Universidade do Minho
Escola de Engenharia

Rita Maria Araújo Bezerra

Ambient Intelligent and Machine Learning Applications for increasing Efficiency and Skills: A Case Study on Secondary Education on Math

October, 2026

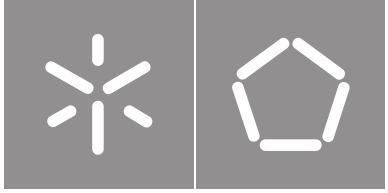
Ambient Intelligent and Machine Learning Applications for increasing Efficiency and Skills: A Case Study on Secondary Education on Math

Rita Maria Araújo Bezerra

UMinho | 2026

REPÚBLICA PORTUGUESA

UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional



Universidade do Minho
Escola de Engenharia

Rita Maria Araújo Bezerra

Ambient Intelligent and Machine Learning Applications for increasing Efficiency and Skills: A Case Study on Secondary Education on Math

Doctoral Thesis

Doctoral Program in Informatics

Work developed under the supervision of:
Paulo Novais

Dalila Durães

October, 2026

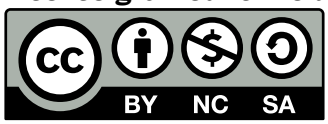
COPYRIGHT AND TERMS OF USE OF THIS WORK BY A THIRD PARTY

This is academic work that can be used by third parties as long as internationally accepted rules and good practices regarding copyright and related rights are respected.

Accordingly, this work may be used under the license provided below.

If the user needs permission to make use of the work under conditions not provided for in the indicated licensing, they should contact the author through the RepositoriUM of Universidade do Minho.

License granted to the users of this work



Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International
CC BY-NC-SA 4.0

`https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en`

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.2.1) [108].

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my thesis supervisors, Dalila Durães and Paulo Novais, for the trust they placed in me from the very beginning and for the opportunity they gave me to undertake this research project. That initial trust, combined with their ongoing support, rigorous guidance and patience throughout the entire process, made this journey an intellectually demanding yet enriching challenge for my academic and professional development.

I am particularly grateful to Professor Paulo Novais for his insightful suggestions, his consistently relevant guidance, and the decisive scientific contribution he made at key stages of this work. To Professor Dalila Durães, I express my sincere gratitude for her constant presence, her unconditional availability and her attentive guidance throughout these nearly five years. Her dedication, rigour and support were crucial to the completion of this thesis.

I would also like to thank the Algoritmi Centre/LASI at the University of Minho for welcoming me and for providing the scientific and technical facilities necessary for the completion of this work. To the entire team at the centre, who contributed to a stimulating and collaborative working environment and who supported me in overcoming the difficulties encountered along the way, I offer my most sincere thanks.

On a personal level, I would like to express my deep gratitude to my husband for his constant encouragement and for the support he has given me in taking on this academic challenge. To my daughters, whom I love so dearly, I am grateful for the understanding, patience and affection they have shown in the face of the demands and inevitable sacrifices associated with this journey. To my parents, for the unconditional support, both personally and as a family, that they have always given me, and to my sister, my friend and constant companion, for the support, strength and motivation she has never ceased to offer me, I extend my very special thanks.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the Universidade do Minho.

_____, _____

(Place) (Date)

_____

(Rita Maria Araújo Bezerra)

DECLARATION ON THE USE OF ARTIFICIAL INTELLIGENCE (AI) TOOLS

I hereby declare having acted this academic work with integrity and confirm that I have not resorted to any form of undue use of AI tools, nor to the falsification or omission of information, sources, data, results or processes in any of the stages leading to its elaboration.

I further declare that any use of AI tools was done in a transparent and responsible manner, being duly noted in the work, and that I critically verified the accuracy of the content produced with their support, assuming full responsibility for what I submit.

I further declare that I have fully acknowledged the Code of Ethics and Conduct of the University of Minho, as well as the applicable standards of academic integrity and data protection.

_____, _____

(Place) (Date)

_____

(Rita Maria Araújo Bezerra)

Resumo

Ambientes Inteligentes e Aplicações de Aprendizagem Automática para aumentar a eficiência e as competências: Um Caso de Estudo sobre o Ensino Secundário na área da Matemática

O desempenho académico no ensino secundário é influenciado por fatores académicos, comportamentais e contextuais. Neste contexto, a integração da Inteligência Artificial, da Aprendizagem Automática e dos Grandes Modelos de Linguagem na educação oferece oportunidades para identificar padrões de desempenho, apoiar a aprendizagem personalizada e facilitar a tomada de decisões pedagógicas baseadas em evidências.

Esta tese propõe uma Arquitetura de Referência para a Educação Inteligente, enquadrada numa perspetiva de Inteligência Ambiental centrada no software e orientada por dados. A Inteligência Ambiental é entendida como um paradigma que integra dados educativos heterogéneos, informação contextual e componentes computacionais inteligentes para apoiar a tomada de decisões educativas. A arquitetura desenvolvida integra modelação preditiva baseada na fusão precoce de dados, seleção de atributos e Inteligência Artificial Explicável, bem como um Tutor de Matemática baseado em Geração Aumentada por Recuperação e num Modelo de Linguagem de Grande Dimensão executado localmente através do Ollama, alinhado com o currículo do 10.º ano.

A validação centrou-se no desempenho preditivo, na explicabilidade e nas perceções de utilidade e usabilidade, envolvendo, no caso do tutor, um professor e 27 alunos do 10.º ano. Os resultados constituem uma validação inicial de prova de conceito e não incluem a medição longitudinal de ganhos formais de aprendizagem.

De um modo geral, a investigação integra modelação preditiva, explicabilidade e tutoria adaptativa numa arquitetura modular, contribuindo para o desenvolvimento de ambientes educativos inteligentes baseados em dados e para a aplicação responsável da Inteligência Artificial em contextos de ensino formal.

Palavras-chave: Aprendizagem automática, Geração Aumentada por Recuperação, Inteligência artificial explicável, Inteligência Ambiente Virtual, Modelação Preditiva, Tutores Inteligentes

Abstract

Ambient Intelligent and Machine Learning Applications for increasing Efficiency and Skills: A Case Study on Secondary Education on Math

Academic performance in secondary education is influenced by academic, behavioural and contextual factors. In this context, the integration of Artificial Intelligence, Machine Learning and Large Language Models into education offers opportunities to identify performance patterns, support personalised learning and facilitate evidence-based pedagogical decision-making.

This thesis proposes a Reference Architecture for Intelligent Education, framed within a software-centred, data-driven Environmental Intelligence perspective. Environmental Intelligence is understood as a paradigm that integrates heterogeneous educational data, contextual information and intelligent computational components to support educational decision-making. The architecture developed incorporates predictive modelling based on early data fusion, attribute selection and Explainable Artificial Intelligence, as well as a Mathematics Tutor based on Retrieval-Augmented Learning and a Large Language Model run locally via Ollama, aligned with the 10th-year curriculum.

The validation focused on predictive performance, explainability, and perceptions of usefulness and usability, involving, in the case of the tutor, a teacher and 27 Year 10 pupils. The results constitute an initial proof-of-concept validation and do not include longitudinal measurement of formal learning gains.

Overall, the research integrates predictive modelling, explainability and adaptive tutoring within a modular architecture, contributing to the development of data-driven intelligent educational environments and to the responsible application of Artificial Intelligence in formal educational contexts.

Keywords: Machine Learning, Predictive Modelling, Retrieval-Augmented Generation, Explainable Artificial Intelligence, Virtual Ambient Intelligence, Intelligent Tutors

Contents

List of Figures **xiii**

List of Tables **xv**

List of Algorithms **xvi**

Acronyms **xvii**

1 Introduction **1**

1.1 Motivation 2

1.2 Hypothesis and Objectives 4

1.3 Methodology 6

1.3.1 Design Science Research (DSR) 6

1.3.2 PRISMA 8

1.4 Structure of this Thesis 9

2 Concepts and Technologies **11**

2.1 Machine Learning 11

2.1.1 Supervised Learning 13

2.1.1.1 Regression 14

2.1.1.2 Classification 15

2.1.2 Unsupervised Learning 18

2.1.3 Semi-supervised Learning 21

2.1.4 Ensemble Learning 21

2.1.4.1 Bagging 22

2.1.4.2 Boosting 23

2.1.5 Neural Networks and Deep Learning 28

2.1.6 Reinforcement Learning 29

2.2 Large Language Models and AI Agents 30
2.2.1 Description of Language Models 32
2.2.2 Chatbot Architecture 36
2.2.2.1 Rule-Based Chatbots 36
2.2.2.2 Retrieval-Based Chatbots 37
2.2.2.3 Generative Chatbots 37
2.2.2.4 Hybrid Chatbots 38
2.2.3 Data Fusion in Large Language Model and AI Agents 41
2.3 Agentic AI 42
2.4 Synthesis of the Conceptual and Technologies 43

3 Literature Review **44**
3.1 State-of the Art in Intelligent Education 44
3.1.1 Analysing the existing systems on the education intelligent 46
3.1.2 Analysing the systems that considered affective learning and the student profile 51
3.1.3 Strategies and tools that use AI to predict student success 55
3.2 Systematic Review in Intelligent Education 59
3.2.1 The use of Retrieval Augmented Generation in educational applications . . 61
3.2.1.1 RAG for Educational Applications: A Systematic Survey 61
3.2.1.2 Artificial Intelligence Generated Content 62
3.2.1.3 Development of Interactive Assistance for Academic Preparation Using Large Models Language 65
3.2.1.4 Evaluating Large Language Model for Student Dropout Prediction 67
3.2.2 Use of LLM and RAG in Assessment 67
3.2.2.1 Enabling Educators to Build RAG-Based AI Chatbots 67
3.2.2.2 Revolutionizing Assessment with RAG and LLM 69
3.2.3 LLM and RAG-powered multi-agents 71
3.2.3.1 Taxonomy of LLM-Powered Multi-Agent Architecture 71
3.2.3.2 LLM Multi-Agent Framework for Intelligent Tutoring 74
3.2.3.3 LLM Agents: Advances in Education 77
3.2.3.4 GPTutor: AI-Powered Intelligent Tutoring 78
3.2.4 Frameworks that uses LLM and RAG 80
3.2.4.1 LAMB: AI Assistants for Learning Management Systems 80
3.2.4.2 MerryQuery: Trustworthy LLM Support for Education 81
3.2.4.3 Explainable LLM Chatbot for Learning Support 85
3.2.4.4 LLMs for Feedback in K-12 Data Science Education 87
3.2.4.5 Embodied AI Digital Teachers 88
3.3 Synthesis of the Literature Review 90

4 Proposed Architecture **92**
4.1 Framework . . . 92
4.1.1 Data Sources . . . 94
4.1.1.1 Student Information System (SIS) . . . 94
4.1.1.2 Questionnaires . . . 95
4.1.1.3 Logs . . . 95
4.1.2 Data Processing . . . 95
4.1.3 Modelling and Validation Pipeline . . . 97
4.1.4 Model Training and Management . . . 98
4.1.5 Model Explainability and Fairness . . . 99
4.1.6 Dashboard and Alerts . . . 100
4.1.7 Monitoring and Retrain . . . 100
4.2 Ambient Intelligent Learning Architecture - Mathematics Tutor . . . 101
4.2.1 Architecture Components and System Agents . . . 103
4.2.2 Step-by-Step Operational Flow . . . 104
4.2.3 Formalisation of the Roles and Responsibilities of Agents . . . 105
4.2.4 Limitations of the current implementation . . . 106
4.2.5 Interaction between components . . . 107
4.2.6 Presentation of answers to the student . . . 109
4.2.7 Auditability and compliance . . . 110
4.2.7.1 Conversational Memory Management Architecture . . . 111
4.2.7.2 Data Privacy and Regulatory Compliance . . . 113
4.2.7.3 Audit Mechanisms and Operational Traceability . . . 113
4.3 Synthesis of the Proposed Architecture . . . 114

5 Analysis of practical results **116**
5.1 Case Study 1: Future of Education on Pre-University . . . 117
5.1.1 Context of the paper . . . 117
5.1.2 Motivation . . . 117
5.1.3 Objectives . . . 118
5.1.4 Results . . . 119
5.1.5 Outcomes . . . 121
5.2 Case Study 2: Forecasting and AI-Powered Exploration of Student choices across different types of Mathematics in Pre-University School . . . 122
5.2.1 Context of the paper . . . 122
5.2.2 Motivation . . . 123
5.2.3 Objectives . . . 123
5.2.4 Results . . . 124

5.2.5 Outcomes . . . 127
5.3 Case Study 3: AI-driven educational transformation in secondary schools: Leveraging data insights for inclusive learning environments . . . 127
5.3.1 Context of the paper . . . 127
5.3.2 Motivation . . . 128
5.3.3 Objectives . . . 128
5.3.4 Results . . . 129
5.3.5 Outcomes . . . 130
5.4 Case Study 4: Data Fusion and Predictive Modeling for Academic Performance Assessment: A Case Study on Grade Variation . . . 131
5.4.1 Context of the paper . . . 131
5.4.2 Motivation . . . 132
5.4.3 Objectives . . . 132
5.4.4 Results . . . 133
5.4.5 Outcomes . . . 138
5.5 Case Study 5: Predictive Analytics in Education: A Comparative Analysis of Machine Learning Models for Predicting Student Performance . . . 139
5.5.1 Context of the paper . . . 139
5.5.2 Motivation . . . 139
5.5.3 Objectives . . . 140
5.5.4 Results . . . 140
5.5.5 Outcomes . . . 141
5.6 Case Study 6: Mitigating Large Language Models Limitations in Learning Environments through Retrieval-Based Knowledge Integration . . . 142
5.6.1 Context of the paper . . . 142
5.6.2 Motivation . . . 143
5.6.3 Objectives . . . 143
5.6.4 Results . . . 144
5.6.5 Outcomes . . . 146
5.7 Conclusions of the Analysis of pratical results . . . 147

6 Conclusion and Future Work **148**
6.1 Contributions . . . 149
6.2 Validation of the Research Hypothesis . . . 154
6.3 Dissemination of Results . . . 155
6.4 Final Remark and Future Work Considerations . . . 156

Bibliography **158**

Appendices

A Publications Composing the Doctor Thesis **172**

A.1 Survey for smart and adaptative education 172

A.2 Future of Education on Pre-university Studies 173

A.3 Forecasting and AI-Powered Exploration of Student Choices Across Differents Types of Mathematics in Pre-university School 174

A.4 AI-Driven Educational Transformation in Secondary Schools: Leveraging Data Insights for Inclusive Learning Environments 175

A.5 Predictive Analytics in Education: A Comparative Analysis of ML Models for Predicting Student Performance 176

A.6 Data Fusion and Predictive Modeling for Academic Performance Assessment: A Case Study on Grade Variation 177

A.7 Mitigating LLM Limitations in Learning Environments through Retrieval-Based Knowledge Integration 178

List of Figures

1 Steps of the DSR method adapted from [166]. 8

2 Conceptual structure of the main approaches and methods in Machine Learning (ML) adapted from [142]. 12

3 Decision Trees (DT) model adapted from [29]. 16

4 Maximising margins in linearly separable data using the Support Vector Machine (SVM) algorithm. Adapted from [111]. 17

5 Block diagram of parallel ensemble learning adapted from [114]. 22

6 Block diagram of sequential ensemble learning adapted from [114]. 22

7 The workflow of RAG adapted from [103]. 40

8 Model adapted from [14] for intelligent education. 46

9 Ambient intelligence model adapted from [143]. 47

10 Smart classroom users "STM triangle model"adapted from [68]. 48

11 Structure of the teaching process in the smart classroom (in three phases and ten steps) adapted from [68]. 49

12 Ambient intelligence system model in a smart classroom based on [141]. 51

13 Model adapted from [96, 141]. 52

14 Evaluation model based in [59]. 56

15 An AI-based evaluation index based on model in [59]. 56

16 Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Flow Diagram 60

17 Process for generating content using artificial intelligence adapted from [179]. 63

18 Architectural model adapted from [95]. 66

19 System Architecture adapted from [160]. 70

20 The main features of autonomous multi-agent systems powered by Large Language Model (LLM) adapted from [64]. 72

21 Taxonomic structure with characteristics adapted from [64]. Each point of view integrates levels of autonomy and alignment. 73

22 The diagram illustrates the aspects and mechanisms of a more specific point of view (b-e) adapted from [64]. 73
23 Illustration of the dynamic methodology of the student adapted from [172]. 76
24 Comprehensive overview of LLM agents for education adapted from [36]. 78
25 Lamb system architecture adapted from [8]. 81
26 Pipeline of LLM-Powered Trustworthy Agents adapted from [161]. 83
27 Chatbot architecture within the Angular web application and its connections between Knowledge Graphs (KG), recommendation algorithm, and the LLM system adapted from [3]. . 86
28 The MAGI system for creating an interactive teacher adapted from [191]. 89

29 Machine Learning-Based Architecture for Intelligent Education. 93
30 Architecture of the proposed ambient intelligent. 102

31 Gender distribution by area. 119
32 Last year subject distribution by gender. 120
33 Early Fusion. 137
34 Beeswarm graphics 142
35 Distribution of Mathematics Levels in the Previous Year. 144
36 Feelings per module of the Mathematics subject. 145
37 Regarding the use of the Maths Tutor. 145
38 Student Motivation and Perception of the Maths Tutor. 146

List of Tables

1 Description of the advantages and disadvantages of different ML algorithms. Adapted from [154]. 19
2 Comparison of LLM Models on Mathematical Reasoning Benchmark adapted from `https://paperswithcode.com/sota/math-word-problem-solving-on-math` . . 35
3 Comparative analysis of chatbot architectures based on characteristics discussed in the literature [55, 82, 98]. 39
4 Characteristics of the different stages in Chinese education. Adapted from [180]. 58
5 Results for Machine Learning Models 120
6 Mean and standard deviation by gender for Mathematics A and MACS in the municipalities of Guimarães and Braga. 124
7 Classification Accuracy Across Machine Learning Models and Feature Scenarios. 125
8 Chi-squared test to study the relationship between the variables grades, internet access and computer access. 129
9 Target Variable Distribution and Random Oversampling Balance in Math_Dataset. 134
10 Model Performance Comparison for Portuguese Language Before and After Feature Selection. 135
11 Model Performance Comparison for Mathematics Before and After Feature Selection. . . 135
12 Per-Class Performance Metrics for the Best Performing Model (XGBoost on MFS_Dataset). 136
13 Confusion Matrix for the Best Performing Model (XGBoost on MFS_Dataset Test Partition). 136
14 Mean and Standard Deviation of Accuracy Values obtained through K-Fold Cross-Validation. 136

List of Algorithms

1 AdaBoost.M1 Algorithm 25
2 Gradient Boosting Algorithm 26

Acronyms

AI Artificial Intelligence *(pp. 1–5, 7, 9, 11, 28, 30, 31, 39, 41, 43–47, 53, 54, 56, 58, 62, 71, 104, 110, 113, 116, 128, 130, 131, 147, 149, 151, 157, 172)*

AIED Artificial Intelligence in Education *(pp. 4, 54, 143, 146, 147)*

AIGC Artificial Intelligence Generated Content *(pp. 62–64)*

AIISE Artificial Intelligence Assisted Interactive Smart Education *(pp. 2, 3, 44)*

ANN Artificial Neural Networks *(pp. 14, 17, 18, 28)*

BN Bayesian Network *(pp. 17, 19, 46)*

CNN Convolutional Neural Networks *(pp. 12, 29)*

DBSCAN Density-based clustering algorithm *(p. 20)*

DQN Deep Q-Network *(p. 30)*

DRL Deep Reinforcement Learning *(p. 30)*

DSR Design Science Research *(pp. 6–8)*

DT Decision Trees *(pp. xiii, 15, 16, 18, 19, 22, 25, 133, 134)*

EDM Educational Data Mining *(pp. 4, 55, 139, 147)*

IoT Internet of Things *(pp. 1, 44, 172)*

ITS Intelligent Tutoring Systems *(pp. 74, 78)*

k-NN K-Nearest Neighbour *(pp. 15–19)*

KG Knowledge Graphs *(pp. xiv, 85–87)*

LLM Large Language Model *(pp. xiii, xiv, 3, 11, 30–33, 35, 36, 39–41, 43, 44, 61, 62, 64, 65, 67–72, 74, 75, 77, 78, 80, 81, 83–89, 105–108, 110–112, 114–116, 142–144, 146, 147, 153, 157, 178)*

LR Logistic Regression *(pp. 15, 18, 19)*

MAE Mean Absolute Error *(pp. 14, 15)*
MDP Markov Decision Process *(p. 29)*
ML Machine Learning *(pp. xiii, xv, 1, 4, 8, 9, 11–15, 17, 19–23, 25, 27–30, 36, 43, 56, 67, 116, 122, 123, 128, 130, 147, 149, 151, 172, 174–177)*
MLP Multi-layer Perceptron *(pp. 12, 16, 28)*
MSE Mean Square Error *(p. 14)*

NB Naïve Bayes *(pp. 15, 17, 18)*
NLP Natural Language Processing *(pp. 30, 32, 36, 37, 90)*

PCA Principal Component Analysis *(p. 20)*
PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses *(pp. xiii, 6, 8, 9, 59, 60, 117, 118)*

RAG Retrieval Augmented Generation *(pp. 38–43, 61–63, 65–69, 71, 76, 78–84, 87–90, 101, 103, 105, 108, 110, 111, 114–116, 143, 144, 146, 147, 153, 178)*
RF Random Forests *(pp. 14, 15, 22, 23, 123, 133, 134, 149)*
RL Reinforcement learning *(pp. 11, 12, 29, 30, 83)*
RLHF Reinforcement Learning from Human Feedback *(pp. 31, 32)*

SHAP Shapley Additive exPlanations *(pp. 92, 99, 140–142, 153, 154)*
SIS Student Information System *(pp. 94, 95)*
SKIT Specialised AI Tutor *(pp. 68, 69)*
SLR Systematic Literature Review *(pp. 8, 60, 118, 172, 173)*
SOA State of the Art *(pp. 8, 144)*
STEAM Science, Technology, Engineering, Arts and Mathematics *(pp. 2, 117, 118, 121–123, 173)*
SVD Singular Value Decomposition *(pp. 20, 21)*
SVM Support Vector Machine *(pp. xiii, 15–19, 118, 120, 123, 125–127, 129, 130, 139–141, 149, 173–176)*
SVR Support Vector Regression *(p. 14)*

TEL Technology-Enhanced Learning *(p. 85)*

WEKA Waikato Environment for Knowledge Analysis *(p. 15)*

1 Introduction

Optimal education is fundamental for the continued development and sustainability of human civilization. Educational systems must not only address the fundamental demands of everyday life but also adapt to the increasing presence of intelligent systems that shape contemporary societies. In this context, education must evolve in response to current societal and market needs, promoting the efficient use of resources while contributing to a more sustainable world.

In the era of digital transformation, characterised by rapid technological advancement and an exponential explosion of information, digital literacy and data literacy have assumed a central role in the development of both individual and organisational competencies [65]. At the same time, the expansion of Machine Learning (ML), coupled with the continuous evolution of information technologies, has established these fields as strategic domains of research and innovation, with a cross-sectoral impact across modern society.

These technological developments have led to the widespread integration of intelligent systems within complex information infrastructures and digital service ecosystems, reinforcing the interconnectivity between humans, machines, and data. Consequently, information technologies are emerging as key enablers of the transition towards more sustainable educational models, promoting more efficient resource management and enhancing the effectiveness and quality of teaching and learning processes.

In particulary, the rapid development of Internet of Things (IoT) and Artificial Intelligence (AI) has made the implementation of intelligent models for Internet-based remote education both technically feasible and increasingly accessible [27]. These advances create new opportunities for the design of innovative educational solutions, enabling more flexible, adaptive, and efficient learning environments for students to acquire and consolidate knowledge.

IoT often described as a core outcome of the third scientific and technological revolution [71], enables the interconnection of physical objects, sensors, cloud infrastructures, and communication devices through standardised protocols. This connectivity supports functionalities such as remote monitoring, control, identification, and real-time data acquisition [14]. When integrated into educational contexts, these capabilities open new possibilities for the development of interactive, adaptive, and student-centred learning environments.

Simultaneously, AI has been increasingly adopted across multiple dimensions of education, including administration, instruction, and learning processes [31]. AI-supported educational models, such as Artificial Intelligence Assisted Interactive Smart Education (AIISE), have demonstrated the potential to enhance student engagement and improve academic performance [189]. These approaches contribute to more dynamic and responsive learning experiences, aligning educational practices with the demands of modern learners.

The ongoing integration of advanced technologies has driven a significant transformation in education, combining traditional pedagogical approaches with digital tools in what is often referred to as educational informatization. This transformation encompasses the use of robotics, e-learning platforms, e-learning combined with Science, Technology, Engineering, Arts and Mathematics (STEAM), and virtual learning environments, all aimed at advancing ambient intelligent education. In this context, improving traditional methods of classroom evaluation has become a critical challenge. AI-based evaluation approaches enable the analysis of teaching and learning behaviours, providing more objective, timely, and efficient assessment outcomes.

The primary objective of AI in education is to enhance learning processes and better understand student behaviour [88]. While AI is often associated with highly advanced computational systems, its application in education extends far beyond this perception. In fact, in the area of education, AI technologies are increasingly embedded in diverse tools and platforms, including intelligent tutoring systems, adaptive learning environments, and educational robots, which support personalised learning experiences from early education onwards.

Furthermore, online education has evolved significantly, moving beyond the simple distribution of learning materials towards the integration of intelligent and adaptive systems capable of analysing both instructor and learner behaviour [31, 38, 84, 133]. These systems dynamically adjust content and instructional strategies to enhance the overall educational experience.

Recent studies also highlight the role of AI in shaping future educational scenarios through data-driven approaches. By considering factors such as student aptitude, responsiveness, and skill sets, AI-based models can optimise both teaching strategies and learning outcomes [53]. Additionally, these systems often incorporate self-assessment mechanisms for educators, supporting continuous improvement in teaching practices. Overall, AI-enhanced educational environments contribute to improved learning outcomes while reducing operational costs.

1.1 Motivation

The sustained development of human civilisation implies an increasing commitment to education, requiring educational systems that are capable of responding effectively to the social, technological, and economic transformations of contemporary society. In this context, schools face the dual challenge of preparing students for the demands of everyday life while adapting to the growing presence of intelligent

systems that increasingly shape modern environments. It is therefore essential that educational institutions promote the efficient use of resources and adopt teaching and learning models aligned with current societal realities and individual student needs.

Recent technological advances have contributed significantly to overcoming several limitations of traditional educational models, fostering the development of computer-assisted and technology-enhanced learning environments. These systems typically integrate smart objects, sensors, mobile devices, cloud computing services, and interconnected infrastructures, enabling functionalities such as remote monitoring, identification, control, and data-driven interaction through digital devices [14]. The incorporation of these technologies into educational contexts opens new opportunities for personalising instruction and collecting and analysing meaningful data about the learning process.

Within this context, the use of intelligent educational systems has demonstrated promising results in improving student engagement and academic performance. According to [189], and within the paradigm of AIISE, AI-mediated interactive learning environments can significantly enhance students' motivation and participation in classroom activities. These findings highlight the potential of Artificial Intelligence (AI) to transform traditional pedagogical models towards more dynamic, adaptive, and student-centred approaches.

At the same time, the evaluation of student learning outcomes remains a critical challenge in education. Traditional assessment methods often suffer from limitations related to subjectivity, lack of scalability, and delayed feedback. In contrast, AI-based assessment approaches enable more objective, efficient, and timely analysis of teaching and learning behaviours, allowing for continuous monitoring of student progress and more informed pedagogical decision-making.

The study presented in [53] further emphasises that effective AI-based educational systems must consider multiple dimensions of the learner, including aptitude, responsiveness, behavioural patterns, and individual skill sets. In such frameworks, students benefit from structured and adaptive pedagogical support, while educators are assisted through intelligent self-assessment and feedback mechanisms. However, despite these advances, the implementation of AI-driven educational environments remains a complex challenge, as learning processes are influenced by a combination of pedagogical, cognitive, and contextual factors that are difficult to model comprehensively.

Moreover, the rapid development and widespread adoption of AI, particularly Large Language Models (LLM) such as ChatGPT, has introduced new opportunities for enhancing educational experiences, but also significant challenges. Among these, the issue of hallucinations, defined as the generation of plausible but incorrect or unverified information, represents a critical concern, especially in domains requiring high levels of conceptual rigor, such as mathematics. These limitations highlight the need for robust, reliable, and context-aware AI systems capable of supporting learning without compromising conceptual accuracy.

In this context, there is a clear need for further research into adaptive, AI-driven educational models that are capable of addressing the diversity of student profiles while effectively integrating pedagogical principles, educational data analysis, and intelligent technologies. This necessity constitutes the main motivation for the research presented in this thesis, which aims to contribute to the development of more

effective, reliable, and student-centred intelligent educational environments.

1.2 Hypothesis and Objectives

This research lies at the intersection of Artificial Intelligence in Education (AIED), ML and Educational Data Mining (EDM), with primary and secondary education within the Portuguese education system serving as its empirical context. Given the growing availability of educational data and advances in predictive analysis techniques, it is pertinent to investigate how these approaches can contribute to improving academic performance, promoting educational equity and supporting evidence-based pedagogical decision-making. Within this framework, this research is based on the assumption that the integration of Artificial Intelligence techniques into educational contexts can play a decisive role in improving teaching and learning processes, on the premise that the application of advanced technological solutions not only optimises the management of educational resources but also promotes more engaging, interactive and pedagogically effective learning environments, with the potential to support more engaging, interactive and pedagogically effective learning environments.

Within this framework, the following research hypotheses are formulated to guide the development, implementation, and empirical validation of the proposed model. These hypotheses address three complementary dimensions of the research: predictive performance, explainability and decision support, and perceived usefulness and usability by teachers.

Predictive models based on Artificial Intelligence are able to identify, with adequate predictive performance, pupils' academic performance and the factors most strongly associated with that performance.

The integration of Explainable Artificial Intelligence mechanisms into the educational environment enables the generation of interpretable and actionable information, capable of supporting pedagogical decision-making and the allocation of resources by teachers.

Teachers perceive the AI-based educational environment as useful and usable in helping to identify pupils' needs and plan pedagogical interventions.

In this research, taking secondary education as a case study, the aim is to design and develop a proof-of-concept methodology that combines intelligent systems adapted to students' needs to support personalised learning, facilitate the identification of learning needs, and provide recommendations for more effective pedagogical interventions.

In this context, the overall aim of this thesis is to investigate and develop methodologies based on AI for the analysis and prediction of academic performance, as well as for supporting learning through adaptive educational systems. At the same time, the research aims to analyse the state of the art in intelligent educational systems, with particular emphasis on the role of ML techniques in supporting personalised learning and enhancing data-driven teaching support and decision-making.

Against this backdrop, we propose the development of an ambient intelligent for secondary education, using mathematics as a case study and focusing on the areas of algebra, functions and geometry, with the aim of promoting personalised learning and improving pupils' academic performance.

Thus, the main objective of this Doctoral Thesis is:

To investigate, design and develop a methodology based on AI techniques for the analysis and prediction of secondary school pupils' academic performance, and for supporting evidence-based pedagogical decision-making, through the integration of technological innovations aimed at optimising educational resources and promoting personalised and engaging learning environments, using Mathematics as a representative case study.

To achieve the main objective of this PhD Thesis, it is necessary to define a list of objectives, which are as follows:

(OB1) Conduct a systematic review of the literature on the application of Artificial Intelligence and Machine Learning in pre-university education, with particular focus on predictive modelling of academic performance, gender equity in STEAM fields, and ambient intelligent and intelligent tutoring systems, thereby establishing the theoretical and methodological foundations of the research.

(OB2) To systematically identify, select, and model the key individual or class group, behavioral data and contextual variables influencing students' academic performance, and to analyse how the integration of emerging technologies in educational settings can support data-driven optimisation of educational resources.

(OB3) To assess and compare the effectiveness of multiple Machine Learning models in the education sector, focusing on their predictive performance in classifying and forecasting student academic outcomes.

(OB4) To design an ambient intelligent based on AI intelligence techniques that integrates heterogeneous educational data sources applying early fusion techniques and feature selection methods to support the development of predictive models for grade variation.

(OB5) To identify and interpret the behavioural factors that most significantly influence academic performance in core subjects, using model explainability techniques and permutation-based importance analysis to translate predictive results into practical teaching recommendations.

(OB6) To evaluate the model's performance through case studies or pilot projects in real-world contexts, using quantitative and/or qualitative metrics.

To address these objectives, the following research questions have been formulated. Each question corresponds to one of the study's core objectives:

R.Q.1: What does the current literature reveal about the application of Artificial Intelligence, Machine Learning, ambient intelligence and intelligent tutoring systems in pre-university education?

R.Q.2: Who is most significantly affected in terms of academic performance, and how can emerging technologies support the data-driven optimisation of educational resources?

R.Q.3: Which Machine Learning models perform best in classifying and forecasting student academic outcomes within the pre-university education sector?

R.Q.4: How can an AI-based ambient intelligent system integrating heterogeneous educational data through early fusion and feature selection techniques support accurate and robust predictive modelling for grade variation?

R.Q.5: When do specific factors most strongly influence performance in core subjects, and how can explainability and permutation-based importance techniques translate these findings into practical teaching recommendations?

R.Q.6: Where does the proposed model demonstrate effective performance when evaluated through a case study or pilot project using quantitative and/or qualitative metrics?

1.3 Methodology

In this thesis, the primary research methodology adopted is Design Science Research (DSR), which provides the conceptual and methodological framework for the development and validation of the proposed artefact. Complementarily, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol was employed specifically to guide the literature review process, ensuring a structured, transparent, and reproducible identification and selection of relevant studies. In the following subsections, each of these approaches is presented and discussed, with the aim of clarifying their distinct roles within the overall research design and their contribution to the development of this work.

1.3.1 Design Science Research (DSR)

We followed the DSR methodology in this doctoral thesis. DSR emphasises two fundamental aspects in its approach, rigour and relevance, and represents a research paradigm that aims to solve complex real-world problems through the design and evaluation of artefacts. This methodology stands out because it is not limited to describing or explaining reality, but proposes innovative solutions in the form of artefacts that expand the frontiers of knowledge and practice.

The DSR emerged in the 1990s, mainly from the work of Hevner, March and other authors, as a response to two obvious gaps in Information Systems research.

The DSR generates scientific knowledge through the design, development and evaluation of innovative, purpose-driven artefacts intended to solve real-world problems [66, 131]. The knowledge resulting from this process can take various forms, including generalisable solutions to classes of problems [175], design

principles that guide the development of artefacts [28] and design theories that explain and underpin the creation and use of these artefacts [80].

DSR thus emerged as a proposal for integrating scientific rigour with practical relevance. This approach complements the positivist (quantitative) and interpretive (qualitative) paradigms by simultaneously emphasising the creation and validation of innovative artefacts that contribute to both theoretical advancement and practical utility.

In this context, DSR is particularly well suited to building our predictive model and our mathematical tutor, as it allows us to solve the specific problem of predicting students' academic results and supports students at risk.

DSR is characterised by three fundamental features that are closely interlinked throughout the research process: the cycle of continuous improvement, the dual contribution (theoretical and practical) and the problem-centred approach.

The classic formulation of DSR is presented by [66], who define fundamental principles for conducting design-oriented research, emphasising the importance of the usefulness, rigour and evaluation of the artefacts developed. Subsequently, [130] propose a process model structured in six phases: problem identification, definition of solution objectives, design and development, demonstration, evaluation and communication, which, although presented sequentially, should be understood as iterative and flexible. This iterative nature allows the research process to evolve progressively, integrating continuous feedback and refining both the understanding of the problem and the effectiveness of the proposed solution.

The application of DSR to research in AI in Education is particularly relevant given the hybrid nature of this field, which combines technological, pedagogical and social dimensions. Through DSR, it is possible to develop AI systems, such as intelligent tutors, adaptive learning platforms or personalisation algorithms, which are designed on the basis of educational theories and simultaneously tested in real-world learning contexts. This link between theory and practice is one of the main advantages of the methodology, enabling not only the empirical validation of the solutions developed, but also the production of transferable scientific knowledge.

The stages of this methodology, shown in Figure 1, are: Problem identification; Definition of solution objectives; Design and development; Demonstration and evaluation; Communication.

In the problem identification phase, the research problem is clearly defined and contextualised, establishing the objectives that guide the investigation and grounding them in existing theoretical foundations. This is followed by the definition of solution objectives, where the requirements and expected outcomes of the proposed solution are specified, taking into account available methodologies and insights from the state of the art.

The design and development phase focuses on the creation of the artefact, informed by analogous solutions and prior research, translating theoretical and methodological considerations into a concrete implementation. Subsequently, the demonstration and evaluation phase aims to validate the artefact in an appropriate context, assessing its effectiveness, relevance, and alignment with the defined objectives, while also contributing to the refinement of design knowledge.

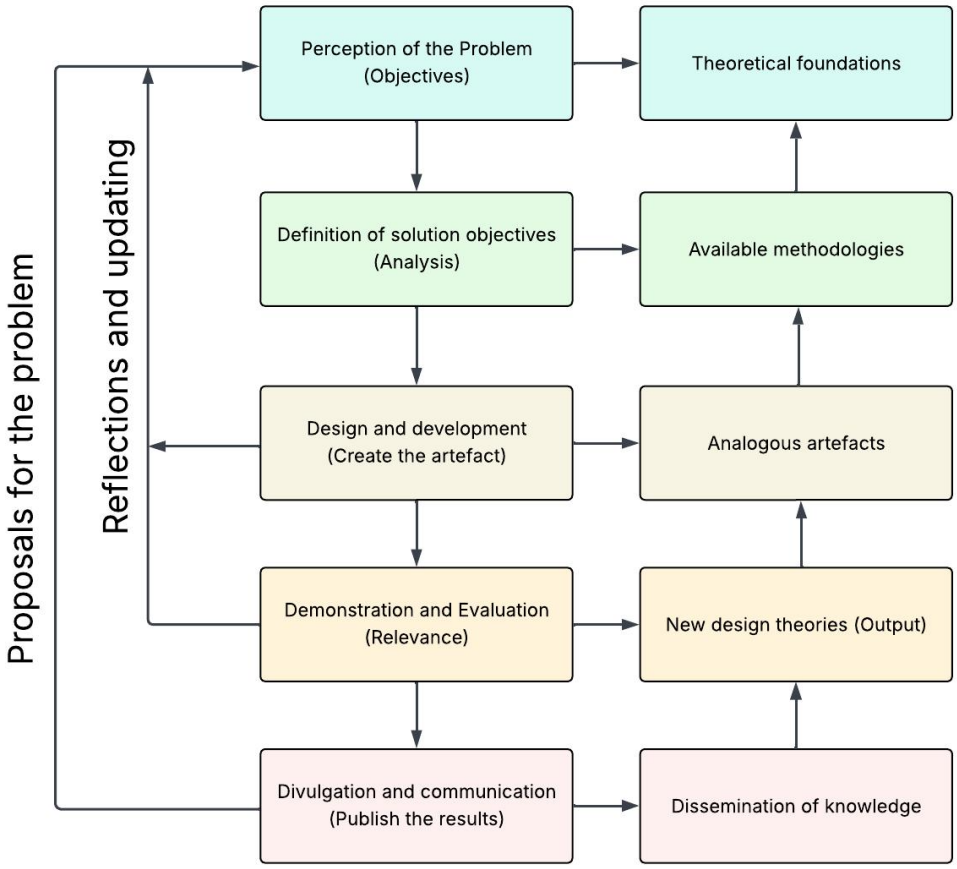


Figure 1: Steps of the DSR method adapted from [166].

Finally, the communication phase involves the dissemination of the research outcomes to both the scientific community and practitioners, ensuring the transfer of knowledge. It is important to note that this process is inherently iterative, incorporating continuous cycles of reflection and refinement that allow for the progressive improvement of both the problem understanding and the proposed solution.

This methodology allowed the thesis to be structured into chapters based on these five stages, framing the development of the maths tutor as a DSR artefact, justifying the iteration between theory (State of the Art (SOA) in ML) and practice (real data from schools) and demonstrating the dual contribution (the digital tutor artefact together with design knowledge, principles for developing predictive systems in education).

In summary, DSR presents itself as a robust and appropriate methodological approach for research into Artificial Intelligence in Education, in that it enables complex problems to be addressed through the development of innovative solutions that are theoretically grounded and empirically validated. Its ability to integrate different dimensions of knowledge, combined with its flexibility and practical orientation, makes it particularly effective in producing relevant scientific and technological contributions in this field.

1.3.2 PRISMA

The Systematic Literature Review (SLR) was based on the PRISMA statement and checklist.

Within the scope of this research, which seeks to integrate AI with the field of education, systematic literature reviews stand out as an effective strategy for synthesising and organising existing scientific knowledge. This approach makes it possible to identify and analyse the key elements of complex, multidimensional systems. In this context, the PRISMA methodology plays an essential role by providing a structured process for conducting systematic reviews, promoting the transparency, reproducibility and reliability of the results achieved [78, 87, 144].

Among the advantages of the PRISMA protocol, its ability to minimise bias through rigorous and well-defined procedures for the identification, screening and selection of studies stands out, which contributes significantly to the robustness of the results [87, 121]. In the field of education, where literacy in AI and the use of AI-based tools are becoming increasingly prominent, the use of PRISMA proves particularly relevant. This methodology enables the identification of gaps in curricula, the systematisation of good practices, and the development of innovative, evidence-based strategies for curricular evolution and restructuring.

1.4 Structure of this Thesis

This doctoral thesis is structured as a compendium of case studies developed throughout the research process, incorporating a presentation of the data and results obtained, as well as the scientific publications produced in the course of this work. In addition, the main conclusions are presented and possible avenues for future research are identified.

As previously mentioned, and with the aim of facilitating the follow-up of the research process, this document is organised into seven chapters.

Chapter 1, the Introduction, presents the framework of the problem under study, proceeding to its characterisation and contextualisation. This chapter also formulates the research hypothesis, defines the objectives guiding the work carried out, and describes the methodology adopted. A brief overview of the thesis structure is also included, to guide the reader.

Chapter 2 is devoted to presenting the concepts and technologies relevant to the field under study, addressing, in particular, classification techniques within the scope of ML, as well as large-scale language models and artificial intelligence agents.

This is followed by Chapter 3, which presents a literature review, systematising the state of the art and identifying research gaps.

The main body of the thesis begins in Chapter 4, which describes the initial architecture of the study, as well as its evolution towards the final architecture, resulting from significant advances in large-scale language models.

Chapter 5 analyses and discusses the results obtained, highlighting the contributions of the work carried out.

It concludes with Chapter 6, which presents the study's conclusions, validates the research hypothesis, summarises the main findings and outlines considerations for future work.

Finally, in Appendice presents the scientific publications that form part of this doctoral thesis, consolidating the main results achieved throughout the research.

2

Concepts and Technologies

The progress of AI in recent decades has been largely driven by the development and diversification of machine learning (ML) techniques, which form the theoretical and methodological foundation for creating systems capable of learning from data. These techniques, ranging from supervised and unsupervised methods to reinforcement learning approaches, form the conceptual core that has enabled the emergence of large-scale models, known as LLM, which play a central role in advancing natural language processing. Furthermore, LLMs represent a paradigm shift from traditional ML, integrating deep neural networks with attention mechanisms that enable the modelling of natural language with previously unattainable levels of coherence and contextualisation. At the same time, the integration of these models into autonomous and adaptive systems has led to the development of so-called AI agents, computational entities capable of perceiving the environment, reasoning, planning and acting in a goal-oriented manner.

It is therefore essential to have a comprehensive understanding of ML techniques, LLM and AI agents in order to analyse the current trajectory of AI and its impact on education, which is characterised by a growing convergence between statistical learning, semantic representation and intelligent autonomy.

2.1 Machine Learning

The ML is a very broad interdisciplinary field that draws on concepts from computer science, statistics, cognitive science, optimization theory involving also a lot of mathematics. There are many fields where ML can be used, however, the most significant of all is data mining.

The field of ML has grown to address the same core problem: extracting patterns and making decisions from data. Different families have emerged because each captures distinct types of relationships between variables (linearity vs. non-linearity), different levels of supervision (labels available or not), and different practical requirements (explainability, scale, efficiency). The history of ML can be seen as a succession of responses to previous limitations: from simple linear models (easy interpretation) to non-linear and distance-based models, to tree-based methods (robustness/partial interpretability) to ensemble methods (increased stability/accuracy) and finally deep neural networks (ability to represent very complex functions) and Reinforcement learning (RL) (learning through interaction).

The Figure 2 summarises, in a visual and hierarchical manner, the contemporary landscape of ML, organising the main families of methods and representative examples for each of them. At the centre is the generic notion of ML, from which four main branches emerge: Classical Learning (divided into supervised approaches, with classification and regression tasks, and unsupervised approaches, which include clustering and dimensionality reduction, as well as pattern discovery techniques such as Apriori/FP-Growth); Ensemble Methods (strategies for combining models, such as bagging, boosting, and stacking); RL (methods based on agents and rewards); and Neural Nets and Deep Learning (from perceptron and Multi-layer Perceptron (MLP)) to deep architectures: Convolutional Neural Networks (CNN) for vision, RNN/LSTM/-GRU for sequences, autoencoders for representation, and GANs for generation). This organisation is not intended to be exhaustive, but rather functional: it highlights conceptual relationships (e.g., how ensembles enhance classical models or deep nets) and differentiates between usage objectives (prediction, clustering, dimensionality reduction, control), methodological trade-offs (complexity and predictive power versus interpretability and computational cost) and suitability for types of data and problems.

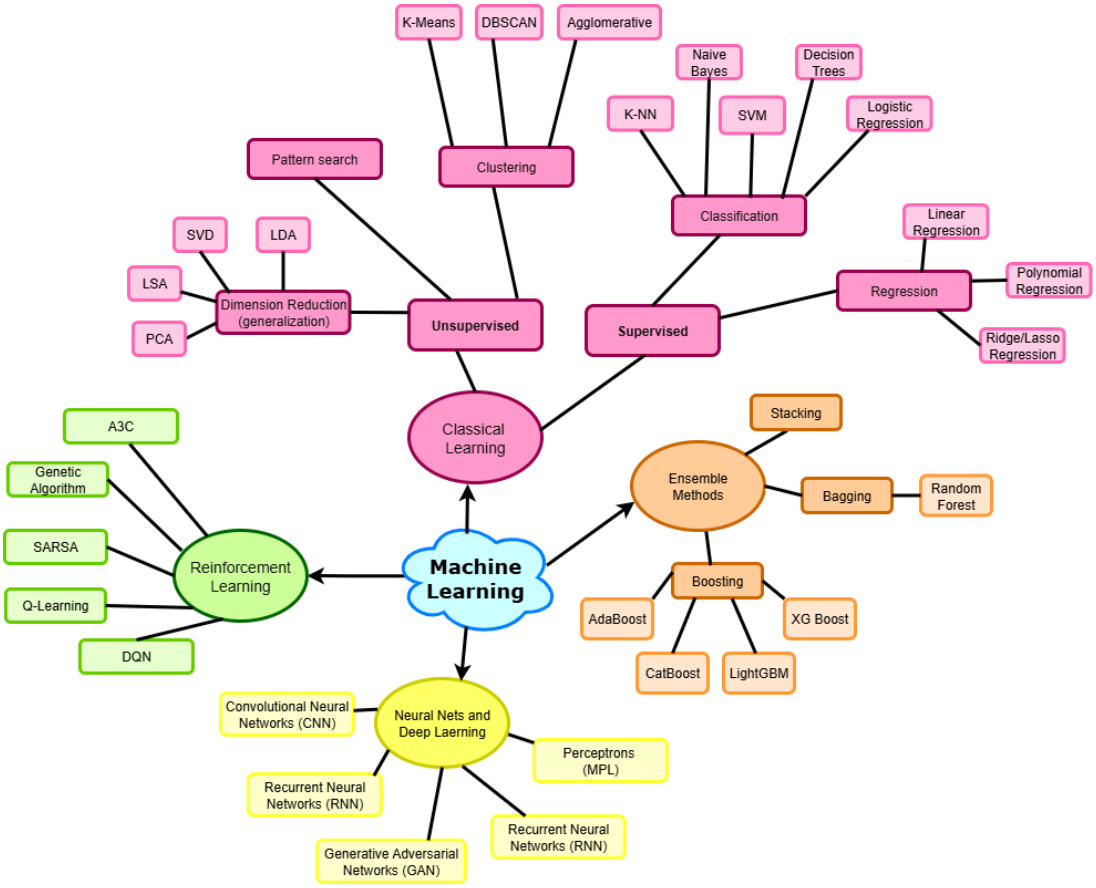


Figure 2: Conceptual structure of the main approaches and methods in ML adapted from [142].

The Figure 2 aims to serve as a reference map to contextualise the methodological choices made, justify the selection of algorithms for the study tasks, and guide the discussion on the evaluation, robustness, and applicability of the artefacts developed.

The ML algorithms basically consume and process data to learn patterns related to individuals, companies, process processes, transactions, events, etc. Since they use data it is important to make a distinction between the various types of data [156]. Data can be of various types, such as: structured, semi-structured, or unstructured [63, 113]. In addition, "metadata" is another type that typically represents data about data.

Structured: Has a well-defined structure, conforming to a data model following a standard order, that is highly organized and easily accessible, and used by an entity or a computer program. Structured data is typically stored in a tabular format. For example, names, dates, addresses, credit card numbers, stock information, geolocation, etc.

Unstructured: On the other hand, unstructured data does not have a predefined format, which makes it difficult to collect, process and analyse. This data mainly comprises text and multimedia content, such as emails, PDF files, audio, video, images, web pages and other business documents.

Semi-structured: Semi-structured data is not stored in a relational database like the structured data mentioned above, but has certain organisational properties that facilitate analysis. Examples are: HTML, XML, JSON documents, NoSQL databases, etc.

Metadata: Not the normal form of data, but "data about data". The main difference between "data" and "metadata" is that data is simply the stuff that can classify, measure, or even document something concerning the properties of an organisation's data. Metadata, on the other hand, describes the relevant information in the data, giving more meaning to users of the data. A basic example of metadata for a document, can be the author, leak size, keywords to dispute the document, etc.

The following is a brief presentation of each of these types of learning technique with the scope of their applicability to solve real world problems [156].

2.1.1 Supervised Learning

Supervised learning is typically the task of ML to learn a function that maps an input to an output based on sample input-output pairs [62]. It uses labelled training data and a collection of training examples to infer a function. Supervised learning is carried out when certain goals are identified to be accomplished from a certain set of inputs [149], i.e., a task driven approach. The most common supervised tasks are "classification" that separates the data, and "regression" that fits the data. For instance, predicting the class label or sentiment of a piece of text, like a tweet or a product review, i.e., text classification, is an example of supervised learning.

A supervised ML algorithm takes input data (features) and outputs labeled data at the input. They are mostly used for classification and regression tasks.

Supervised ML is the construction of algorithms capable of producing patterns using externally supplied instances to predict the fate of new instances. It aims to characterise data based on prior information.

Regression and classification are the two fundamental subfamilies of supervised learning, with distinct but often complementary objectives.

Regression methods are methods whose purpose is to estimate or predict a continuous response variable from a set of predictors, favouring the quantification of functional relationships between variables and inference about the impact of each predictor.

Classification is based on data science problems. Several successful techniques have been proposed to solve these problems, including techniques based on rules, logic, instances or stochastic patterns.

The study published in [154] analyses various algorithms in terms of accuracy, learning speed, complexity and the risk of overfitting measures. The class of each test instance is decided by combining the characteristics and finding patterns common to each class from the training data.

The most significant distinction between classification and regression is that classification predicts distinct class labels, while regression facilitates the prediction of a continuous quantity.

2.1.1.1 Regression

Regression models aim to predict a continuous dependent variable from a set of explanatory variables, capturing patterns and functional relationships in the data. Linear regression is the base model, assuming a linear relationship between the predictors and the response, often complemented by extensions such as polynomial regression, which introduces non-linear terms for greater flexibility. It offers advantages in interpretability and has proven to be a solid theoretical basis.

To mitigate problems of overfitting and multicollinearity, regularised variants are used, such as Ridge Regression (L2 penalty) and Lasso Regression (L1 penalty), which allow control over model complexity and automatic variable selection. In more complex contexts, non-linear models such as regression trees, Random Forests (RF), Gradient Boosting Machines or Support Vector Regression (SVR) are used, which capture non-linear interactions and relationships between variables. In high-dimensional domains or heterogeneous data, approaches based on Artificial Neural Networks (ANN) offer expressive modelling capabilities, learning latent representations of the data. The evaluation of these models uses quantitative metrics such as Mean Square Error (MSE), Mean Absolute Error (MAE) or coefficient of determination (R^2), complemented by cross-validation techniques and residual analysis, ensuring the robustness and generalisation capacity of the model in new contexts [48].

Regression analysis includes several methods of ML that allow to predict a continuous (y) result variable based on the value of one or more (x) predictor variables [62]. The most significant distinction between classification and regression is that classification predicts distinct class labels, while regression facilitates the prediction of a continuous quantity.

2.1.1.2 Classification

Classification is carried out in two phases. First, a classification algorithm is applied to the training data set and then the extracted model is validated with a test data set to measure the model's performance and accuracy.

In the study [154], the algorithm comparison factors taken into account were: accuracy (proportion of correct classifications), speed (calculation time required), degree of comprehensibility (degree of complexity of an algorithm), learning speed (important in a real-time system where a classification rule must be learnt quickly or adjustments must be made). The data used in this study was obtained from KAGGLE, which is an online data competition platform. Characteristics of the data set: Number of observations (N) = 891, Number of attributes (p) = 12, Number of classes (q) = 2.

The data was then pre-processed. Missing values were replaced by the mean for continuous data and the mode for categorical data. If the class value is missing, then the entire observation is omitted. Values outside the interquartile range were considered outliers and therefore removed.

Supervised classification is one of the tasks most frequently performed by intelligent systems. The study in [126] describes various methods of supervised ML classification techniques, compares various supervised learning algorithms, as well as determining the most efficient classification algorithm based on the dataset, the number of instances and the variables (features).There are several ML algorithms, among which the following stand out: Logistic Regression (LR), RF, Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbour (k-NN), NB Classifier and Decision Tree (J48) using the Waikato Environment for Knowledge Analysis (WEKA). To implement the algorithms, the study used a classification dataset of 786 instances, with eight attributes as independent variables and one as a dependent variable. The research shows that the time needed to build a model and the precision (accuracy) is an important factor, however the kappa value and the MAE must also be considered.

Therefore, ML algorithms require precision, accuracy and minimum error in order to have supervised ML.

Classification methods represent one of the main aspects of supervised learning in ML, dedicated to the fundamental task of assigning observations to predefined discrete categories. As computational tools that learn patterns from labelled historical data, these algorithms enable the prediction of classes in new instances based on observable characteristics.

The classifier ecosystem encompasses a variety of approaches, of which I will describe the most widely discussed and those that contributed to the development of this research work.

The Decision Tree (DT) is a tree-structured model that classify instances by recursively partitioning the feature space according to the values of their attributes. Each internal node represents a test on an attribute, each branch corresponds to a possible outcome of the test, and each leaf node represents the predicted class, as illustrated in Figure 3. During classification, an instance traverses the tree from the root node to a leaf node by following the branches defined by the outcomes of successive attribute tests [93]. DT learning is widely used in data mining and ML to build predictive models that map input

observations to target class labels. To improve generalisation and reduce overfitting, decision trees are commonly pruned by removing branches that contribute little to predictive performance, typically based on validation data or statistical criteria [93].

DT are widely appreciated for their interpretability and their ability to capture complex interactions between attributes. As non-parametric models, they make no assumptions about the underlying data distribution and can naturally handle linearly inseparable data. In addition, they can process different types of attributes, including categorical, numerical and textual data, while also accommodating missing values and redundant features. When properly pruned, DT generally exhibit good generalisation performance, are relatively robust to noise and outliers, and provide efficient training and inference with comparatively low computational cost.

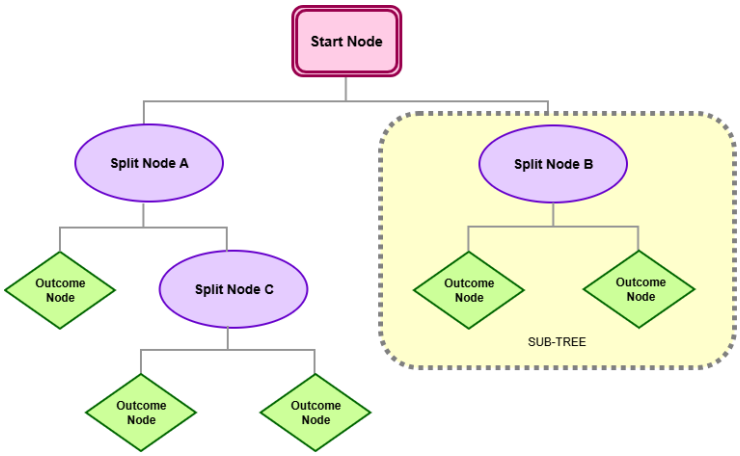


Figure 3: DT model adapted from [29].

The SVM can be used for binary/multi-class classifications. SVM models are closely related to classical MLP [167]. The SVM model revolves around the notion of a "margin", one side of a hyperplane that separates two classes of data. Maximising the margin and consequently creating the greatest possible distance between the separating hyperplane and the separating instances and the instances on either side of it has been proven to reduce an upper bound on the expected generalisation error [93].

It's a complex algorithm, but it can provide high accuracy. With an appropriate channel, the model can work well even when the data is not linearly separated by features. They are based on the concept of maximising the minimum distance between the hyperplane and the nearest sampling point [9]. Accuracy and performance are independent of the amount of data but vary with the number of training cycles, unlike k-NN. Especially popular in text classification problems where very high dimensional spaces are the norm.

In addition to performing linear classification, SVM can efficiently perform non-linear classification using what is known as a kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Basically, they draw margins between the classes, as shown in the Figure 4. The edges are drawn in such a way that the distance between the edge and the classes is maximum and therefore minimises the classification error [111].

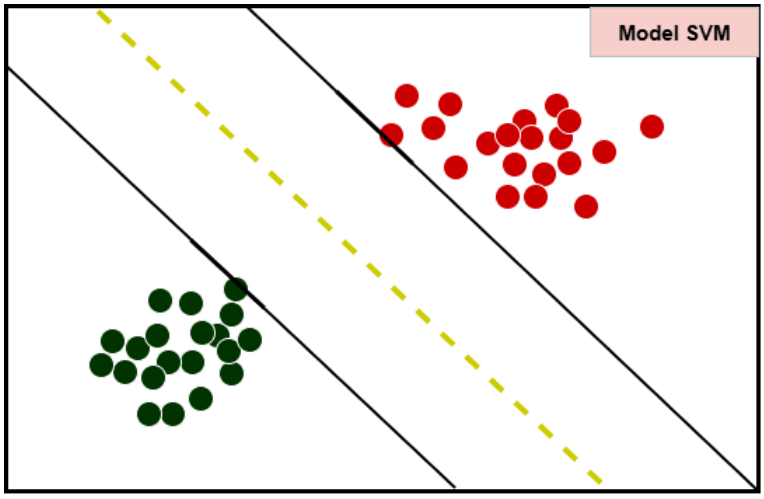


Figure 4: Maximising margins in linearly separable data using the SVM algorithm. Adapted from [111].

k-NN algorithm is simple but computationally exhaustive. It is a non-parametric classification algorithm. It assigns an unlabelled sample point the class of the closest point in a set of previously labelled points [107]. The rule is independent of the joint distribution of the sample points and their classifications. It is suitable for multimodal classes, as well as for applications where the object may have many labels. It is a simple lazy learning method, but it has a lower efficiency. The model's performance depends on the good selection of the value of k, however there is no single way to choose this value, other than by expensive computational techniques such as cross validation. Performance also varies according to the dimension, since all the data has to be revisited [72, 134].

Among supervised ML methods, the NB classifier stands out. It is a probabilistic model based on Bayes' theorem, which estimates the probability of an instance belonging to a given class based on the likelihood of its predictor variables. The adjective naive comes from the assumption of conditional independence between variables given the class label, which is a simplification that is rarely true in practice, but which gives the model high computational efficiency and good generalisation capacity even in high-dimensional data sets. Naive Bayes calculates the posterior probability of each class from the prior probabilities and conditional distributions of the attributes, assigning the new observation to the class with the highest estimated value. There are several versions of this algorithm, adapted to the nature of the data: Gaussian Naive Bayes for continuous attributes (assuming normal distribution), Multinomial Naive Bayes for count variables (frequent in text classification), and Bernoulli Naive Bayes for binary data.

Bayesian Network (BN) are very simple, consisting of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to the observed nodes), with a strong assumption of independence between the child nodes in the context of parenthood [90].

The independence model (Naive Bayes) is based on the estimate [93]. Bayesian classifiers are generally less accurate than other more sophisticated learning algorithms (such as ANN).

NB requires less computing time for training. Unlike neural networks or SVM, there are no free parameters to be grouped, which greatly simplifies this model. It cannot be applied when it is necessary to

take into account the interactions between several characteristics. Despite its simplicity and assumption of independence, Naive Bayes has shown competitive results in several tasks, especially in text classification, spam detection, sentiment analysis, and medical diagnosis, and is appreciated for its robustness, training speed, and probabilistic interpretability.

LR is a linear model for binary classification. Statistical model in which the logistic curve is fitted to the data set. This technique is applied when the dependent or target variable is dichotomous. It gives a good probabilistic interpretation and the model can be updated with new data.

This is a classification function that uses class to construct and uses a single multinomial LR model with a single estimator [12]. LR usually indicates where the boundary between the classes exists, it also states that the probabilities of the classes depend on the distance from the boundary. It moves to the extremes (0 and 1) more quickly when the data set is larger. Because it involves probabilities, it is considered to be more than a classifier. It makes stronger and more detailed predictions, and can be adjusted in a different way; however, these strong predictions can be wrong.

LR is an approach to prediction, like ordinary least squares regression. LR is one of the most widely used tools in applied statistics and discrete data analysis.

The Table 1 below presents a brief comparison between the advantages, disadvantages, and areas of applicability of the different models.

Generally speaking, SVM and neural networks tend to perform much better when dealing with multidimensions and continuous features. On the other hand, logic-based systems tend to perform better when dealing with discrete/categorical features. For neural network models and SVM, a high number of data is required to achieve their maximum prediction accuracy, while NB may need a relatively small dataset.

There is a general consensus that k-NN is very sensitive to irrelevant features. Furthermore, the presence of irrelevant features can make neural network training very inefficient, and even impractical. Most decision tree algorithms fail to perform well on problems that require diagonal partitioning. ANN and SVM performances are better when multicollinearity is present and there is a non-linear relationship between the input and output features.

NB requires little storage space during the training and classification phases: the minimum required is the memory needed to store the prior and conditional probabilities.

Finally, DT and NB generally have different operational profiles; when one is very accurate, the other is not, and vice versa. On the contrary, DT and rule classifiers have a similar operational profile. SVM and ANN also have a similar operational profile. No learning algorithm can uniformly outperform other algorithms on all data sets.

2.1.2 Unsupervised Learning

Unsupervised learning analyses unlabelled data sets without the need for human intervention. It is used to group data into different segments based on input characteristics. It is widely used to extract generative characteristics, identify significant trends and structures, group results, and for exploratory purposes.

Table 1: Description of the advantages and disadvantages of different ML algorithms. Adapted from [154].

Algorithm	Advantages	Disadvantages	Application Areas
DT	As a non-parametric model, it can handle non-linear relationships and heterogeneous data types; it is easy to interpret, robust to noise, and requires relatively low computational cost	Prone to overfitting, especially with deep trees; small variations in data may lead to different tree structures; limited ability to model complex relationships without ensemble methods	Implantable devices, welding quality, medication analysis, remote sensing
SVM	High classification accuracy, effective in high-dimensional spaces, and flexible kernel functions allow modelling of non-linear relationships	Computationally intensive for large datasets; sensitive to parameter selection and kernel choice; limited interpretability	Text classification, bioinformatics, image recognition
k-NN	Simple and intuitive method; effective for multimodal class distributions; no explicit training phase	Computationally expensive at prediction time; sensitive to the choice of k and distance metric; affected by noise and irrelevant features	Pattern recognition, recommendation systems, computational geometry
BN	Captures probabilistic dependencies among variables; handles uncertainty explicitly; relatively efficient training for structured data	Performance degrades in high-dimensional spaces; requires prior knowledge or structure learning; sensitive to data sparsity	Document classification, medical diagnosis, decision support systems
LR	Provides probabilistic outputs; easy to interpret; efficient to train; can incorporate interaction and polynomial terms	Assumes linearity in the log-odds; sensitive to multicollinearity; limited capacity to model complex non-linear relationships without feature engineering	Risk modelling, medical studies, social sciences, classification tasks

Unsupervised methods seek to extract the structure of the data without resorting to labels: natural clusters (clustering), frequent rules in transactions (pattern mining), and low-dimensional representations (dimension reduction). In contrast to supervised learning, the goal is to describe and discover patterns (segments, anomalies, latent topics) that inform hypotheses, support decision-making, or compress features for subsequent tasks. The most common approaches to unsupervised learning include techniques such as clustering, density estimation, representation learning, dimensionality reduction, association rule mining and anomaly detection, amongst others.

For example, imagine the task of organising a group of 100 people into five distinct groups. Rather than using a predefined classification, the system looks for patterns within the data itself. To do this, it can draw on information such as personal interests, hobbies or social connections between individuals, automatically forming groups based on natural similarities between them.

For example, to organise a group of 100 people into five distinct groups. Rather than using a predefined classification, the system looks for patterns within the data itself. To do this, it can draw on information such as personal interests, hobbies or social connections between individuals, automatically forming groups based on natural similarities between them.

Cluster analysis, also known as clustering, is an unsupervised ML technique for identifying and grouping related data points in large datasets without concern for the specific outcome. Used in different areas such as cybersecurity, e-Commerce, mobile data processing, health analytics, user modeling and behavioral analytics.

In short, clustering offers a powerful lens for revealing order in seemingly disconnected data, making it an essential tool for exploratory analysis in domains as diverse as education, biology, and marketing.

Partitional clustering, whose most prominent example is the K-means algorithm, is a clustering approach whose main objective is to partition the data space into K groups in order to minimise the sum of the squared distances of each point to the centroid of its cluster.

Formally, this objective translates into minimising the following expression [185]:

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{1}$$

where μ_k represents the average of the points in cluster C_k, functioning as the centroid or prototype of that group.

However, K-means also has important limitations: its sensitivity to centroid initialisation can lead to suboptimal solutions, a problem that the K-means++ algorithm seeks to mitigate through more intelligent and dispersed initialisation. Furthermore, the algorithm is sensitive to the presence of noise and outliers, assumes that clusters are spherical and of approximately equal size, and relies heavily on the use of Euclidean distance.

Density-based clustering algorithm (DBSCAN), which identifies clusters based on point density; and Hierarchical Clustering, which constructs a tree of clusters representing different levels of granularity, Agglomerative clustering, which represents a hierarchical clustering model.

There are also other types of clustering methods including partitioning methods, density-based methods, hierarchical-based methods, grid-based methods, model-based methods, constraint-based methods, mean-shift clustering, GMM clustering (Gaussian Mixture Models).

Dimensionality reduction, which is an unsupervised learning technique, is important because it leads to better human interpretations, lower computational costs, and avoids overweight and redundancy by simplifying models. The examples of dimensionality reduction are: Feature selection; Feature extraction.

The most common methods include the variance threshold, Pearson's correlation, ANOVA, the chi-squared test, recursive feature elimination, RFE, and model-based selection. In addition, dimensionality reduction techniques, such as Principal Component Analysis (PCA), t-SNE and Singular Value Decomposition (SVD), are widely used. PCA allows the number of variables to be reduced by creating independent

components, albeit with lower interpretability; t-SNE enables the representation of high-dimensional data in lower-dimensional spaces; and SVD decomposes matrices into simpler components, facilitating efficient processing [156].

2.1.3 Semi-supervised Learning

Semi-supervised learning can be defined as a combination of the supervised and unsupervised methods, in the sense that it works with both labelled and unlabelled data [62, 149]. In the real world, labelled data may be rare in various contexts, and unlabeled data are numerous, where semi-supervised learning is useful [120]. The ultimate goal of a semi-supervised learning model is to provide a better result for prediction than that produced using only the labelled data from the model. This type of learning is widely used in machine translation, fraud detection, data labelling and text classification.

2.1.4 Ensemble Learning

Ensemble learning techniques have achieved state-of-the-art performance in several ML applications by combining the predictions of two or more base models. The study focuses on widely used ensemble algorithms, including random forest, adaptive boosting (AdaBoost), gradient boosting, extremum gradient bossting (XGBoost), light gradient boosting machine (LightGBM) and categorical boosting (CatBoost). An attempt is made to concisely cover their mathematical and algorithmic representations, which according to the authors there is little literature on and which would be beneficial to researchers and ML professionals.

Research has shown that conventional ML algorithms tend to underperform when trained with unbalanced datasets. Researchers have therefore frequently turned to new and improved learning approaches, such as ensemble learning and deep learning.

Ensemble learning methods train multiple base learners and combine their predictions to obtain better performance and generalisation capacity than individual base learners [115]. It's important to recognise that ML models have limitations and also make mistakes. However, ensemble learning aims to improve classification performance by taking advantage of the strengths of various base models. However, some limitations of ML algorithms include: the fact that they result in models with high variance, high bias and low accuracy [118, 159]. Several studies have shown that ensemble models often achieve greater accuracy than individual ML models [114]. These models limit the variance and bias errors associated with individual ML models. For example, bagging reduces variance without increasing the bias, while boosting reduces bias [7, 23, 41]. In general, the ensemble classifiers are more robust and perform better than individual ensemble learners.

The practical and theoretical motivation is to reduce generalisation error through aggregation: different models error in different ways, and the appropriate combination of their outputs can reduce both variance and bias, increasing robustness and stability. In tabular problems and many applied problems, ensemble methods (specifically boosting and bagging) often lead benchmarks.

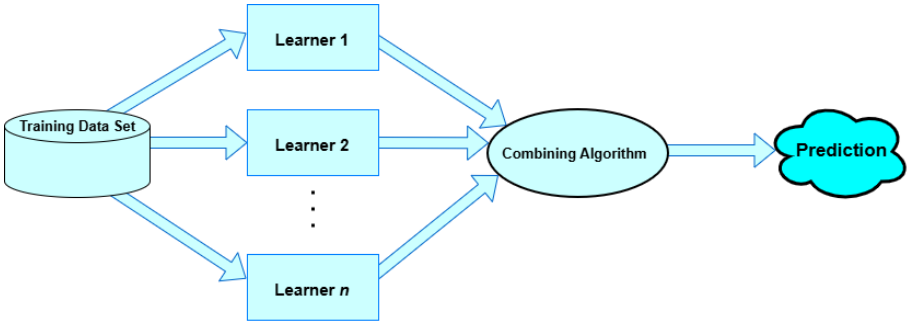


Figure 5: Block diagram of parallel ensemble learning adapted from [114].

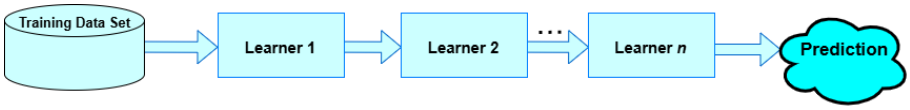


Figure 6: Block diagram of sequential ensemble learning adapted from [114].

Ensemble learning methods are broadly categorized into boosting, bagging, and stacking [115]. Gradient boosting, XGBoost, and AdaBoost are examples of boosting algorithms, while random forest and Extra Trees classifier are well-known bagging algorithms. Examples of the stacking framework include super ensemble and blending techniques [47, 125, 145].

Figures 5 and 6 show block diagrams describing parallel and sequential ensemble learning. Parallel ensembles can also be categorized as homogeneous or heterogeneous, depending on the homogeneity of the basic learners. Homogeneous ensembles are made up of models built using the same ML algorithm, while heterogeneous ensembles comprise models from different algorithms [21, 102, 146].

2.1.4.1 Bagging

The Bagging (Bootstrap Aggregating) method is an ensemble learning technique based on a fundamental principle: generating diversity by creating multiple subsets of training data using bootstrap sampling, that is, sampling with replacement from the original set. For each of these subsets, a base model is trained independently. The final prediction is obtained by aggregating the individual predictions: by averaging, in the case of regression problems, or by majority vote, in the case of classification.

The main effect of bagging is to reduce the variance of the final model by "smoothing out" fluctuations or instabilities that may arise due to the particularities of a single training sample. This approach is especially effective when applied to unstable base models, such as DT, whose structures can vary significantly with small changes in the training data.

The flagship algorithm that extends and popularises this idea is Random Forest (RF). This technique not only applies bagging, but also introduces an additional layer of randomness: in each split of each tree, only a random subset of features (feature subspacing) is considered when searching for the best splitting criterion. This mechanism further increases the diversity among the trees in the forest, reducing

the correlation between them and, consequently, improving the robustness and generalisation capacity of the ensemble.

Among the most valuable properties of RF are its robustness to outliers and noise, the existence of few critical hyperparameters and the relative ease of interpretability it offers through feature importance, allowing analysts to identify which variables contributed most to the predictions.

2.1.4.2 Boosting

Boosting is a ML technique capable of converting weak learners into a strong classifier. It is a type of ensemble meta-algorithm used to reduce bias and variance. From this perspective, a weak learner is a classifier that performs a little better than random guessing, while strong learners are those that achieve good accuracy [159], and are the core of the construction of boosting ensemble algorithms. The boosting algorithm was first discussed in a 1990 article by Schapire [150]. His work had a significant impact on ML and statistics, which led to the development of several boosting algorithms, including AdaBoost [50] and XGBoost [33].

The idea of boosting is to iteratively apply the base learning algorithm to adjusted versions of the input data [115]. In particular, boosting techniques use the input data to train a weak learner, calculate the weak learner's predictions, select misclassified training samples and train the subsequent weak learner with an adjusted training set that includes the misclassified instances from the previous training round [116]. The iterative learning process continues until a predefined number of baseline learners is obtained and the baseline learners are weighted together [100]. Boosting focuses more on reduction and therefore reinforces base learners with high bias and low variance, such as decision stumps (a decision tree with an internal node).

However, this iterative learning approach makes boosting unsuitable for learning noisy data because the weight assigned to the noisy samples is usually much higher than the weights assigned to the other samples, thus forcing the algorithm to focus excessively on the noisy samples, which results in overfitting. Despite this, boosting-based ensemble methods are among the most successful algorithms in ML [81, 116].

1. **AdaBoost**

   This algorithm was developed in 1995 by Freund and Schapire [50] and remains the most robust ML algorithm. The most successful implementation of AdaBoost is AdaBoost M1, used for binary classification tasks [94]. The different base learners are added sequentially and weighted to obtain the strong classifier [170]. In each iteration, the AdaBoost algorithm assigns weights to each instance of the training set [51].

   AdaBoost (Adaptive Boosting) is the initial exponent of this lineage, introducing a fundamental principle: iterative adaptation through dynamic reweighting of training instances. The algorithm operates by reassigning higher weights to observations misclassified by previous models, thus

forcing subsequent learners to progressively focus on the most difficult cases. Its algorithmic simplicity is supported by a solid theoretical foundation in margin theory, which provides a formal explanation for its generalisation ability, demonstrating how boosting effectively maximises the classification margin of instances.

Given m labeled training instances $S = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_m, y_m)\}$, where y_i is the target label of sample x_i and $y_i \in Y = \{-1, +1\}$, the weight D_1 of sample x_i and the weight update D_{t+1} are calculated as follows [114]:

$$D_1(i) = \frac{1}{n}, \quad i = 1, 2, \ldots, m \tag{2}$$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \exp(-\alpha_t y_i h_t(x_i)), \quad i = 1, 2, \ldots, \tag{3}$$

where $h_t(x)$ is the base classifier, $t = 1, \ldots, T$ is the number of iterations, Z_t represents a normalization factor, while α_t is the weight of the $h_t(x)$ classifier. The weight α_t measures the importance of the classifier $h_t(x)$ in obtaining the classifier's final prediction. Instances that are wrongly predicted in $h_t(x)$ are given higher weights in the $t + 1$ training round. In addition, Z_t is selected so that D_{t+1} is a distribution. Z_t and α_t are obtained using:

$$Z_t = \sum_{t+1}^{n} D_t(i) \exp(-\alpha_t y_i h_t(x_i)) \tag{4}$$

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1 - \epsilon_t}{\epsilon_t}\right) \tag{5}$$

where t represents the error rate of the classifier, and is obtained using:

$$\epsilon_t = P[h_t(x_i) \neq y_i] = \sum_{i=n}^{n} D_i(i) I[h_t(x_i) \neq y_i] \tag{6}$$

When a number of iterations have been completed, the final strong classifier is calculated using:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{7}$$

The AdaBoost algorithm, summarised in Algorithm 1 ([114]), is easy to implement with little need to adjust its [176] hyperparameters. Furthermore, AdaBoost is flexible and can use a variety of algorithms as a basis for learning. However, a limitation of AdaBoost is that it is sensitive to noisy data and outliers due to its iterative learning approach, causing overfitting.

Algorithm 1 AdaBoost.M1 Algorithm

Input: Training data $S = \{(x_1, y_1), \ldots, (x_2, y_2), \ldots, (x_m, y_m)\}$
The base algorithm L
The number of iterations T

Procedure:
Initialize the weight D_1 of sample x_i using (2)
for $t = 1, \ldots, T$ **do**
1) Train the base classifier $h_t(x)$ by minimizing ϵ_t
2) Calculate the weight α_t of the classifier using (5)
3) Update the sample weights using (3)
end for

Output: Apply (7) to combine the predictions of the base classifiers to obtain the final strong classifier $H(x)$.

2. **GardientBoosting**

A fundamental conceptual evolution emerged with Gradient Boosting Machines (GBM), which reformulated the problem of boosting within the framework of numerical optimisation in function space. In this perspective, each base model (typically a limited-depth decision tree) is viewed as a step in a functional gradient descent, being trained to approximate the (negative) gradient of a loss function defined in relation to the current ensemble predictions. In concrete terms, each new tree is constructed to predict the residuals of the existing composite model, allowing for sequential and targeted minimisation of error.

Gradient boosting is a ML algorithm that uses the boosting technique to create strong ensembles, where DT are used as the base learner to produce a robust ensemble classifier, and is also called a gradient boosted decision tree (GBDT). The gradient boosting technique was first introduced by Breiman [23], who noted that boosting can be represented as an optimisation technique on a suitable loss function. Later, an extended version of the algorithm was developed by Friedman [52]. The learning process of this algorithm consists of sequentially training new models to obtain a robust classifier. The central idea is to develop base learners that are highly correlated with the negative gradient of the loss function related to the entire ensemble [123].

Given a training set $S = \{x_i, y_i\}_1^N$ the *gradient boost* technique aims to find an approximation $\hat{F}(x)$ of the function $F^*(x)$ that maps the predictive variables x into their response variables y, minimising the function $L(y, F(x))$. GBDT creates an additive approximation through a weighted sum of functions:

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x) \tag{8}$$

Where ρ_m represents the weight of the function $m^{th}(h_m(x))$. The algorithm performs the approximation iteratively. However, a constant approximation of is obtained using:

$$F_0(x) = 2 \min_{\alpha} \sum_{i=1}^{N} L(y_i, \alpha) \tag{9}$$

The goal is to minimise:

$$(\rho_m h_m(x)) = 2 \min_{\rho,h} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i)) \tag{10}$$

Algorithm 2 Gradient Boosting Algorithm

Input: Training data $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$
A differentiable loss function $L(y, F(x))$
The number of iterations T

Procedure:
1) Initialize the model with a constant value using:

$$F_0 = 2 \min_{\alpha} \sum_{i=1}^{N} L(y_i, \alpha)$$

2) **for** $m = 1, \ldots, M$ **do**
(i) Calculate the pseudo-residuals:

$$r_{mi} = \left[\frac{\delta L(y_i, F(x))}{\delta F(x)} \right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, N$$

(ii) Train a base learner using the dataset $D = \{(x_i, r_{mi})\}_{i=1}^{N}$
(iii) Perform line search to find ρ_m and $h_m(x)$:

$$(\rho_m h_m(x)) = 2 \min_{\rho,h} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \rho h(x_i))$$

(iv) Update the model:

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x)$$

end for

Output: Return the final model $F_m(x)$.

The gradient boosting procedure is summarised in Algorithm 2 ([114]). The main advantage of gradient boosting is that, like other boosting algorithms, it can learn complex patterns from the

input data, since it is trained to correct the errors of the previous model. However, a model built with this algorithm can overfit and model noise if the input data is noisy [123, 187]. This algorithm is great for applications with small data sets [77].

The maturation of this paradigm has given rise to a generation of advanced implementations that combine sound theoretical principles with high-performance computational optimisations.

3. **XGBoost**

XGBoost (Extreme Gradient Boosting) stands out for introducing L1 and L2 regularisation terms directly into the objective function, explicitly controlling model complexity and mitigating overfitting. At the same time, it implements efficient algorithms for identifying optimal split points using the histogram method, native support for missing values, and extensive parallelisation.

The XGBoost algorithm is a decision tree-based ensemble that uses the gradient boosting structure. It is a scalable and highly accurate algorithm used for classification and regression applications. XGBoost has recently dominated the field of applied ML and has won several Kaggle competitions. It was developed in 2016 by Chen and Guestrin [33], with several advances compared to the conventional gradient boosting algorithm. Unlike gradient boosting, the XGBoost loss function contains a regularisation term that prevents overfitting [102]:

$$L_M(F(x_i)) = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{m=1}^{M} \Omega(h_m) \tag{11}$$

where $F(x_i)$ represents the prediction in the i-th instance in the M-th iteration, $L(*)$ represents a loss function that calculates the differences between the predicted class and the actual class of the target variable. Meanwhile, (h_m) represents the regularisation term, and is formulated as:

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{12}$$

where γ represents the complexity parameter, and controls the minimum loss reduction gain required to split an internal node. Assigning a high value to γ leads to simpler trees. Meanwhile, T represents the number of leaves in the tree, λ is a penalty parameter and ω denotes the output of the leaf nodes. A second-order Taylor approximation of the objective function is employed in XGBoost and Equation 11 is transformed to obtain:

$$L_M \approx \sum_{i=1}^{n} \left[g_i f_m(x_i) + \frac{1}{2} h_i f_m^2(x_i) \right] + \Omega(h_m) \tag{13}$$

where g_i and h_i represent the first and second derivatives of the loss function. Assuming that I_j represents the samples at leaf node j, the final loss value is calculated by adding up the loss values of the various leaf nodes. Thus, the objection function is represented as:

$$L_M = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \tag{14}$$

One of the advantages of using the XGBooost algorithm is that it requires minimal feature engineering, such as data normalisation and feature scaling, since the algorithm can handle these situations. In addition, it is able to deal with missing values. This algorithm can be used to better understand input features and contribute to feature selection. XGBoost is faster than most ML algorithms, can handle large data sets and is not prone to overfitting. In addition, it often outperforms other ML algorithms, however, due to its high number of hyperparameters, it becomes limited.

2.1.5 Neural Networks and Deep Learning

ANN emerged from simple, interpretable models that sought to reproduce, in a rudimentary way, the behaviour of a biological neuron. Among these early models, the perceptron,introduced in the 1950s, stands out, it was considered the first practical architecture capable of performing linear classifications using a set of adjustable weights. This elementary model demonstrated, from an early stage, both its potential and its limitations, notably its inability to represent composite non-linear functions.

These limitations were overcome with the generalisation of the perceptron to multilayer architectures and the rediscovery of the backpropagation algorithm in the 1980s. This advance enabled networks with hidden layers to learn non-linear hierarchical representations, allowing architectures of the MLP type to perform arbitrary approximations of complex functions. Despite their theoretical soundness, by the end of the 20th century these networks still faced significant challenges, including optimisation difficulties, problems associated with the saturation of activation functions such as the sigmoid and hyperbolic tangent, as well as limitations in available data and large-scale computational capacity.

The MLP can be described as a classifier in which the network weights are determined by solving a quadratic programming problem with linear constraints, in contrast to the non-convex, unconstrained minimisation problem characteristic of traditional neural network training [12].

In turn, the perceptron algorithm is used to learn from a set of training instances, running iteratively over this set until a correct prediction vector is obtained for all training examples. The decision rule thus obtained is subsequently applied to the classification of new data, namely in the test set [93].

In this context, the development of neural networks must be understood within the broader framework of AI, where machine learning emerges as an approach centred on the automatic learning of patterns from data, in contrast to the explicit programming of rules. According to [34], ML constitutes a subset of AI that enables the construction of models capable of inferring relationships between variables and generalising to new data, with deep learning being an extension of these approaches based on deep neural networks.

ANN, particularly in the form of MLP, represent one of the main methodologies within this paradigm, characterised by their layered structur and by their ability to iteratively transform data using non-linear

activation functions. These architectures typically operate in a feedforward mode, in which information flows unidirectionally between layers, being progressively transformed until a final output is produced. The introduction of multiple hidden layers enables the learning of hierarchical representations, in which simple features are combined into more abstract and informative descriptions. The natural evolution of these approaches led to the development of deep learning, characterised by the use of networks with a large number of hidden layers, capable of automatically learning complex representations directly from the data. Among the most significant architectures are CNN, which are particularly effective in computer vision tasks, as they preserve the spatial structure of the data and enable the automatic extraction of relevant features through convolution operations [34].

Despite the progress made, these approaches still present significant challenges, notably the need for large volumes of training data, high computational demands and the limited interpretability of the models, which are often described as 'black boxes'. Furthermore, issues such as overfitting and dependence on the quality of training data continue to pose obstacles to the generalisation of models in real-world contexts [34].

In this context, the development of neural networks and deep learning has followed a clear chronological progression: from the linear perceptron to backpropagation-trained multilayer perceptrons, through specialised architectures for image and sequence data, to generative and attention-based paradigms, culminating today in large-scale models capable of learning transferable and highly expressive representations. In parallel, new issues relating to computational efficiency, interpretability and ethical implications have been shaping the future directions of research.

2.1.6 Reinforcement Learning

RL is a distinct paradigm within ML, based on the idea of learning through interaction. While supervised learning requires labelled data and unsupervised learning exploits hidden structure in the data, RL learns through experience, maximising a reward obtained by interacting with a dynamic environment.

The concept was inspired by behavioural psychology theories, namely Skinner's operant conditioning: the agent learns behaviours that increase rewards and avoids actions that lead to punishments. Mathematically, RL is formalised as a Markov Decision Process (MDP).

RL is a type of ML algorithm that allows software and machines to automatically evaluate the optimal behaviour in a given context or environment to improve its efficiency [83]. This type of learning is based on reward or penalty. It is a powerful tool for training AI models that can help increase automation or optimize the operational efficiency of sophisticated systems such as robotics, autonomous driving tasks production and supply chain logistics, however, it is not preferable to use it to solve basic or simple problems. The problem to be solved in RL is defined as a MDP [137], i.e., all about sequential decision making. An RL problem typically includes four elements such as Agent, Environment, Rewards, and Policy. Q-learning, Deep Q Network, Monte Carlo Control, SARSA (State–Action–Reward–State–Action), etc. are some examples of model-free algorithms [83].

Existing data is used by ML algorithms to develop predictive models and automate various time-consuming tasks. The ML techniques take input parameters and expected reference output data and generate logic which is then deployed into production.

When deep neural networks are combined with RL, the field of Deep RL emerges, responsible for remarkable advances in the last decade (AlphaGo, OpenAI Gym, robotics). The network acts as an approximator of the value function or policy, allowing it to handle continuous and high-dimensional state spaces (e.g., images, sensors) [10].

This family of Deep Reinforcement Learning (DRL) algorithms includes Deep Q-Network (DQN), Genetic Algorithms, and A3C (Asynchronous Advantage Actor-Critic), one of the most effective DRL methods. Its applications in various fields, mainly in games and simulations, have achieved historic feats, with systems such as AlphaGo surpassing world champions in Go, and agents dominating a range of classic Atari games, or exhibiting sophisticated tactics in challenging environments such as StarCraft II, where long-term planning and real-time adaptation are crucial [119, 169].

In robotics, RL enables motion control and adaptive object manipulation, where robots learn, through trial and error, complex motor skills that would be extremely difficult to program explicitly.

In the education sector, its potential is particularly promising through the development of intelligent tutoring agents. These systems can learn and adapt their teaching strategies in real time, based on rewards derived from student engagement and performance, thus personalising the learning path to maximise understanding and knowledge retention [132].

In healthcare, RL is applied to the personalisation of treatments, modelling sequences of therapeutic interventions as a Markovian decision process to identify optimal regimens for each patient [158]. Finally, in the financial sector, it is used to develop adaptive trading strategies, where agents learn to execute orders in a dynamic and competitive market, maximising returns over time [105].

2.2 Large Language Models and AI Agents

Advances in LLM have significantly reshaped the field of AI, particularly within Natural Language Processing (NLP), enabling new paradigms for human–machine interaction and intelligent information processing. Built upon transformer architectures [168], these models rely on self-attention mechanisms to capture complex semantic and contextual relationships within textual data, and are trained on large-scale corpora through a combination of unsupervised pre-training and supervised fine-tuning. This design has proven highly effective in producing coherent, contextually relevant, and linguistically fluent outputs across a wide range of tasks [117, 128].

The progression from early transformer-based models to large-scale foundation models reflects a fundamental shift in AI research. GPT-1 introduced the effectiveness of transfer learning through unsupervised pre-training followed by task-specific adaptation [139]. GPT-2 extended this paradigm by demonstrating strong generative capabilities in open-ended text generation, highlighting both the scalability of

transformer architectures and the emerging risks associated with their deployment [140]. GPT-3 further expanded model capacity to 175 billion parameters, introducing strong few-shot and zero-shot learning capabilities that allowed the model to perform a broad range of tasks with minimal or no task-specific training data [24]. This evolution contributed to the consolidation of the concept of foundation models, characterised by large-scale pre-training and broad adaptability to downstream tasks [22].

Within this continuum, ChatGPT represents a significant milestone in the operationalisation of LLM for conversational AI. Built on GPT-3.5 and later enhanced versions, ChatGPT incorporates Reinforcement Learning from Human Feedback (RLHF), combining supervised fine-tuning with preference-based optimisation to better align model outputs with human expectations [127]. This alignment process has been fundamental in improving response quality, reducing harmful outputs, and enabling more natural and interactive dialogue systems. Subsequent developments, including GPT-4 and more recent multimodal models, have further extended these capabilities by improving reasoning, robustness, and enabling the processing of multiple modalities such as text and images [147].

From a broader scientific perspective, these systems should be interpreted as statistical models of language rather than autonomous agents, generating outputs based on learned probability distributions over large datasets. Their emergence reflects not only a technical evolution but also a broader socio-technical transformation, in which natural language becomes a universal interface for computation, knowledge access, and decision support. This development raises important challenges related to bias, transparency, accountability, and the epistemic reliability of machine-generated content [17].

In parallel, recent extensions of LLM have focused on increasing their functional capabilities beyond static text generation, including tool use, planning, and interaction with external systems [151]. These developments highlight a transition towards more agent-like architectures capable of integrating reasoning with action, which is particularly relevant for complex applied domains.

In educational contexts, these advances have led to a wide range of applications, including intelligent tutoring systems, personalised learning environments, automated feedback generation, and educational content creation [86]. LLM-based chatbots can function as virtual tutors, providing real-time explanations, adaptive support, and interactive learning assistance [104, 136]. They have also been applied in automated assessment tasks, such as essay evaluation and feedback generation, as well as in the production of quizzes and summarisation of educational materials [74]. These capabilities have contributed to the increasing integration of LLM into AI-driven educational ecosystems.

However, despite these advances, several limitations remain when deploying LLM in real-world educational environments. A central issue is hallucination, where models generate plausible but factually incorrect or misleading information due to their probabilistic nature [75]. Additionally, their reliance on static training data limits their ability to reflect up-to-date curricular content and scientific knowledge [112], while concerns related to bias, lack of transparency, and academic integrity remain critical in educational deployment scenarios.

2.2.1 Description of Language Models

With the emergence of deep learning, systematically introduced through deep neural networks, it has become possible to learn hierarchical representations directly from the data, reducing the reliance on feature engineering and enabling significant advances in computer vision, speech recognition and NLP [97].

From this paradigm emerged the family of LLM, based on Transformer architectures [168], which allow for the modelling of long-range dependencies in text and the scaling of training to large volumes of data.

Trained on massive volumes of textual data, these models demonstrate remarkable capabilities in natural language generation, question answering, summarisation, translation and the performance of various linguistic tasks with high performance [22, 24, 39]. Their effectiveness stems, to a large extent, from learning deep and contextualised representations of language, enabling generalisation across multiple tasks without the need for specialised architectures for each problem.

One of the most significant developments associated with these models is the pre-training followed by fine-tuning paradigm, in which a model is initially trained on large text corpora to learn general linguistic patterns, and is subsequently adapted to specific tasks [39, 139]. More recently, approaches based on instruction tuning and RLHF have further enhanced these systems' ability to align with human objectives and specific usage contexts [127]. These strategies have contributed to substantial performance improvements across multiple tasks and consolidated LLM as foundational models for general-purpose applications.

In an educational context, the potential of these models has been attracting growing interest. For primary and secondary school pupils, LLM can serve as learning support tools across various subjects, including mathematics, physics, languages and literature, by generating explanations, exercises, contextualised problems and adaptive quizzes. These capabilities can contribute to a deeper understanding of the content, promoting retention, contextualisation of knowledge and the development of critical thinking [86, 106]. Furthermore, by providing step-by-step solutions, exploratory questions and support for problem-solving, these models can foster analytical skills, creativity and independent learning.

From an educational perspective, these systems can also support teachers in preparing teaching materials, generating summaries and outlines of complex texts, as well as in designing more interactive and personalised activities. This type of support has the potential to reduce time spent on repetitive tasks, allowing teachers to devote more effort to creative, student-centred pedagogical strategies [186].

In terms of assessment, LLM have also been explored as tools to support marking and formative feedback, particularly in written tasks such as essays, reports or research papers. Although they do not replace human assessment, they can support semi-automated marking processes, identify strengths and weaknesses in students' work, and contribute to more personalised and rapid feedback. Some studies also point to potential in detecting plagiarism or AI-assisted writing patterns, although these applications remain the subject of scientific and ethical debate [37].

Recent developments in LLM have not been limited to natural language generation and conversational interaction, but have progressively extended to more sophisticated capabilities in reasoning, problem-solving and symbolic manipulation. In particular, a new generation of models geared towards reasoning-intensive tasks has emerged, capable of combining linguistic skills with logical inference, code generation and the solving of complex mathematical problems [24, 173]. This trend has been reinforced by techniques such as chain-of-thought prompting, instruction tuning and integration with external tools, which have significantly enhanced the performance of models on tasks traditionally associated with formal reasoning [151, 173, 183].

In this context, some models stand out particularly in mathematical and scientific problems. GPT-4 has demonstrated significant advances in mathematical reasoning, symbolic problem-solving and code generation for computational mathematics, with impressive results on benchmarks such as GSM8K, MATH and HumanEval [5]. Similarly, models from the Gemini family, developed by Google DeepMind, have shown high performance in multimodal reasoning and advanced mathematics tasks. In the case of the Minerva model, specifically trained for mathematics and the exact sciences, it has been demonstrated that scaling up combined with specialised corpora can substantially improve the ability to produce step-by-step solutions and formal proofs [99].

In the open-source domain, highly competitive architectures have also emerged. The Qwen2.5 family, developed by Alibaba Cloud, stands out for its strong performance in mathematical reasoning, with the Qwen2.5-72B-Instruct variant achieving competitive results on benchmarks such as GSM8K, MMLU and HumanEval [181]. Similarly, mixture-of-experts models such as Mixtral, from Mistral AI [76], and compact reasoning-oriented models such as Phi-4 [1], highlight a growing trend towards specific optimisation for logical-mathematical tasks, demonstrating that competitive performance does not always depend exclusively on scale.

A comparison between models such as Phi-4 [1] and Qwen2.5 [13] reveals significant structural and operational differences that determine their application in different contexts. Phi-4, developed by Microsoft, has approximately 14 billion parameters, whilst Qwen2.5, developed by Alibaba Cloud, reaches higher scales, notably in the 72 billion parameter variant, which gives it greater representational capacity [1, 13].

With regard to performance, both models demonstrate competitive results on several widely used benchmarks, including MMLU, GPQA, GSM8K and HumanEval, as reported in their respective technical papers [1, 13]. It is important to note, however, that these evaluations are conducted independently and with distinct experimental configurations; there are not necessarily direct and uniform comparisons between the two models across all scenarios. Thus, any differences in performance should be interpreted with caution.

Both models also share the feature of being open-source, which facilitates their analysis, adaptation and implementation by the scientific community. However, they differ significantly in terms of context capacity: Qwen2.5 supports up to 128,000 tokens, allowing for the processing and generation of lengthy documents, whilst Phi-4 typically operates with smaller context windows, in the tens of thousands of tokens

[1, 13].

So, Phi-4 demonstrates that more compact models can achieve high levels of performance, especially when optimised for efficiency and reasoning, whilst Qwen2.5, with greater parametric capacity and support for extended contexts, positions itself as a robust solution for applications requiring the processing of large volumes of textual information. The choice between these models should therefore be guided by the specific requirements of the application, taking into account the nature of the tasks, the scale of the data and the available computational resources.

In addition, the open-source model ecosystem also includes Gemma 2, developed by Google, which is available in versions with 9 billion and 27 billion parameters, thereby broadening the range of options for researchers and developers who prioritise accessibility, flexibility and customisation when developing solutions based on language models [165].

When comparing the performance of different large language models (LLMs) in mathematical problem solving, distinct capability profiles emerge, largely shaped by their architectural design choices, training objectives, and domain-specific optimisation strategies.

The Gemma 2 model, developed by Google, is primarily designed as an efficient and accessible open-weight model for general-purpose language understanding and generation [165]. Although it demonstrates solid performance across a wide range of reasoning and conversational tasks, it is not explicitly optimised for advanced mathematical problem solving. As a result, its performance in complex mathematical benchmarks tends to be more limited when compared to models specifically trained or tuned for mathematical reasoning, positioning it as a suitable solution for general-purpose applications rather than specialised analytical domains.

In contrast, Llama 3.1 [58], developed by Meta, exhibits strong capabilities in both programming and mathematical reasoning tasks, particularly in its larger-scale variants. These models benefit from increased parameter capacity and extensive training on code and technical corpora, which contributes to their competitive performance on benchmarks such as HumanEval and MATH [58]. The largest variants, including those exceeding hundreds of billions of parameters, demonstrate enhanced ability to model complex symbolic relationships and generate accurate multi-step solutions, making them well-suited for technically demanding applications.

From a different perspective, the Phi-4 model, developed by Microsoft, was designed with a strong emphasis on reasoning efficiency and high-quality training data [1]. Despite its relatively compact scale of approximately 14 billion parameters, it achieves competitive performance in reasoning-intensive benchmarks, including tasks related to mathematical problem solving. This result reflects a significant trend in recent research: a model's performance is no longer determined solely by the number of parameters it contains, but also by other key factors such as how the data is curated, the training approach used, and the quality of the architectural design and optimisation. Consequently, Phi-4 challenges the conventional assumption that larger models are inherently superior, demonstrating that smaller, well-optimised models can achieve comparable results in specific domains.

Given the focus on performance in mathematical reasoning tasks, Table 2 presents a comparison

of the main language models, including Gemini 2.0 Flash, Qwen2.5 and GPT-4 Turbo, based on results reported in the MATH benchmark, as well as on relevant characteristics such as parameter scale and the use of additional training data.

Table 2: Comparison of LLM Models on Mathematical Reasoning Benchmark adapted from `https://paperswithcode.com/sota/math-word-problem-solving-on-math`

Rank	Model	Accuracy	Parameters	Extra Training Data	Paper	Year
1	**Gemini 2.0 Flash Experimental**	89.7	–	–	–	2024
2	**Qwen2.5-Math-72B-Instruct** (TIR.Greedy)	88.1	72	✓	Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement	2024
3	**GPT-4 Turbo** (MATH, w/ code, voting)	87.920	–	–	MACM: Utilizing a Multi-Agent System for Condition Mining in Solving Complex Mathematical Problems	2024
4	**Qwen2.5-Math-72B-Instruct** (CoT.Greedy)	85.9	72	✓	Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement	2024
5	**Qwen2.5-Math-7B-Instruct** (TIR.Greedy)	85.2	7	✓	Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement	2024

Benchmarks apply different metrics to evaluate the performance of LLM. Here are some of the most common ones:

- **Accuracy** or precision calculates the percentage of correct predictions.
- **Recall**, also known as sensitivity rate, quantifies the number of true positives, or actual correct predictions.
- The **F1 score** combines precision and recall into a single metric. It assigns equal weights to both measures to balance false positives and false negatives. The F1 score ranges from 0 to 1, with 1 indicating excellent recall and precision.
- **Exact match** measures the proportion of predictions that exactly match the correct answers, which is an important criterion for tasks such as translation and question answering.
- **Perplexity** evaluates a model's ability to correctly predict the next word in a sequence. The lower the perplexity, the better the model understands the task.
- **Bilingual Evaluation Understudy (BLEU)** evaluates machine translation by calculating the correspondence of n-grams (sequences of n adjacent text symbols) between the translation generated by the LLM and the human reference translation.

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** evaluates text summarisation and has different variants. ROUGE-N, for example, performs calculations similar to BLEU for summaries, while ROUGE-L identifies the longest common subsequence between the generated summary and the one produced by a human.

Benchmarks can fall into one or more categories. The one we used to test the best model was GSM8K, as it tests the mathematical reasoning skills of an LLM. It contains a corpus of 8 500 elementary school math problems. The solutions are collected in natural language rather than mathematical expressions. AI evaluators are trained to evaluate the solutions generated by the model.

2.2.2 Chatbot Architecture

Chatbots are conversational systems designed to interact with users through natural language, leveraging advances in NLP, ML, and computational linguistics. Over the past decades, their architectures have evolved significantly, reflecting broader developments in artificial intelligence. The architecture of a chatbot determines how user input is interpreted, processed, and transformed into responses. In the literature, chatbot systems are commonly categorised into three main paradigms: rule-based, retrieval-based, and generative models [82, 153]. More recently, hybrid approaches have emerged, aiming to combine the strengths of multiple paradigms, particularly in terms of accuracy, flexibility, and robustness [98].

Each of these paradigms presents distinct characteristics, advantages, and limitations, making them suitable for different application contexts.

In order to provide a structured understanding of chatbot systems, it is essential to examine the main architectural paradigms that underpin their design. These paradigms differ fundamentally in how they process user input, manage knowledge, and generate responses, reflecting distinct trade-offs between control, flexibility, and computational complexity. Accordingly, the following subsections present a detailed analysis of the principal approaches, namely rule-based, retrieval-based, generative, and hybrid architectures, highlighting their operational mechanisms, advantages, limitations, and typical application domains.

2.2.2.1 Rule-Based Chatbots

Rule-based chatbots represent the earliest and most deterministic form of conversational systems. Their architecture relies on predefined rules, such as decision trees, regular expressions, and keyword matching, to guide interactions. User inputs are mapped to predefined responses through explicit “if–then–else” logic, typically supported by structured knowledge bases containing frequently asked questions and scripted dialogues [153].

The operation of these systems is straightforward: input text is analysed for specific patterns or keywords, which are then associated with predefined responses. This deterministic behaviour ensures high

predictability and transparency, making rule-based chatbots particularly suitable for applications requiring auditability and strict control.

However, this simplicity also imposes significant limitations. These systems lack the ability to generalise beyond predefined rules, struggle with linguistic variability, and cannot adapt dynamically to new inputs. Furthermore, they require continuous manual maintenance to remain effective.

A classical example is ELIZA [174], which simulated a psychotherapist using pattern matching and substitution rules. Despite its historical importance, it illustrates the inherent limitations of purely rule-based approaches.

Rule-based chatbots are therefore most appropriate for simple applications, such as menu-driven systems, FAQ automation, and basic customer support scenarios.

2.2.2.2 Retrieval-Based Chatbots

Retrieval-based chatbots represent an evolution of rule-based systems, incorporating more advanced NLP techniques and supervised learning methods. Instead of generating responses, these systems select the most appropriate answer from a predefined repository based on the user's input [177].

Their architecture typically involves several stages. First, user input is transformed into a vector representation using embedding techniques. This representation is then compared with indexed candidate responses using similarity metrics. Traditional approaches rely on methods such as TF-IDF or BM25, while more recent systems employ neural embeddings derived from models such as BERT [39] or Sentence Transformers, often combined with vector search frameworks such as FAISS.

The main advantage of retrieval-based systems lies in their reliability: since responses are pre-defined and validated, they ensure grammatical correctness and factual consistency. Additionally, they offer greater control over generated content and are computationally more efficient than generative models.

However, their limitations are evident. These systems cannot generate novel responses, are constrained by the size and quality of their response database, and may produce repetitive outputs in extended interactions. Consequently, they are best suited for applications requiring high precision and consistency, such as technical support systems, recommendation engines, and enterprise knowledge assistants.

2.2.2.3 Generative Chatbots

Generative chatbots represent the current state-of-the-art in conversational AI, leveraging deep neural networks to generate responses dynamically. Unlike retrieval-based systems, they produce novel outputs token by token, based on learned statistical patterns in large corpora [24].

These systems are typically based on Transformer architectures [168], which include components such as encoders, decoders, and attention mechanisms that enable contextual understanding and sequence modelling. Prominent examples include models such as GPT-5, and BERT.

The key advantage of generative chatbots is their flexibility. They can handle a wide range of topics, generate context-aware responses, and support complex, multi-turn conversations. Furthermore, they are

capable of performing diverse tasks such as summarisation, translation, reasoning, and content generation.

Despite these strengths, generative models present important challenges. They may produce factually incorrect or fabricated information (commonly referred to as "hallucinations"), exhibit unpredictable behaviour, and require substantial computational resources for both training and inference [17]. Additionally, controlling their outputs remains a non-trivial problem.

These systems are particularly suitable for general-purpose assistants, educational tools, and creative applications.

2.2.2.4 Hybrid Chatbots

Hybrid chatbots have emerged to address the limitations of individual paradigms by combining multiple approaches within a unified architecture. These systems integrate rule-based, retrieval-based, and generative components, often coordinated by an orchestration layer that dynamically selects the most appropriate response strategy.

A prominent example is the Retrieval Augmented Generation (RAG) framework [98], which combines a retrieval module with a generative model. In this approach, relevant documents are first retrieved from a knowledge base and then used as contextual input for response generation, improving factual accuracy while maintaining linguistic fluency.

Hybrid architectures may include several components, such as intent classifiers, rule engines, retrieval modules, generative models, and routing mechanisms. Different integration strategies can be employed, including pipeline (cascade), ensemble, and context-augmented generation approaches.

The main advantage of hybrid systems lies in their ability to balance accuracy and flexibility. They reduce hallucinations by grounding responses in external knowledge, optimise computational efficiency, and adapt to diverse interaction scenarios. However, these benefits come at the cost of increased architectural complexity, higher implementation effort, and potential latency due to multi-stage processing.

Hybrid chatbots are particularly well-suited for enterprise applications, advanced technical support, educational systems, and domains requiring both factual reliability and natural interaction, such as healthcare and legal services.

Comparison of different chatbot architectures

When comparing these four architectures, clear patterns emerge. In terms of flexibility, rule-based systems are the most limited, while generative and hybrid systems offer high flexibility. In terms of control, rule-based systems offer total control, retrieval systems offer high control, generative systems offer low control, and hybrid systems maintain medium to high control. Accuracy is high in rule-based systems within their restricted domain, high in retrieval systems, variable in generative systems, and consistently high in hybrid systems. The naturalness of responses increases progressively from rule-based to generative systems, with hybrids maintaining high naturalness. Computational cost is minimal for rules, low for

retrieval, high for generation, and medium for hybrids. This information can be more easily analysed in Table 3.

Table 3: Comparative analysis of chatbot architectures based on characteristics discussed in the literature [55, 82, 98].

Characteristic	**Rule-Based**	**Retrieval**	**Generative**	**Hybrid**
Flexibility	Low	Medium	High	High
Control	Total	High	Low	Medium-High
Accuracy	High (restricted domain)	High	Variable	High
Naturalness	Low	Medium	High	High
Computational Cost	Very Low	Low	High	Medium
Maintenance	High	Medium	Low	Medium-High
Scalability	Low	Medium	High	High
Error Risk	Low	Low	Medium-High	Low-Medium

Current trends in chatbot development show a clear dominance of hybrid models. Most modern commercial chatbots adopt hybrid architectures, especially using RAG, to balance response quality with control and reliability. There is also an evolution from simple chatbots to more sophisticated 'agents' that can perform concrete actions, use external tools, and maintain long-term goals through multiple interactions. Personalisation has become a priority, with systems increasingly able to tailor responses to specific users and maintain relevant context across multiple sessions. Finally, multimodality is gaining importance, with the integration of text, voice, image, and video into a single coherent conversational architecture.

RAG is an advanced AI paradigm that integrates information retrieval (IR) with generative models to improve response accuracy and reliability [98]. Traditional IR systems efficiently locate relevant documents but cannot generate new content, while LLM generate fluent text but rely solely on pre-trained knowledge. RAG bridges this gap through a two-step process: first, it retrieves relevant documents from external databases, and then leverages them to guide text generation [85]. As illustrated in Figure 7, the RAG workflow consists of three main phases: indexing, where documents are stored in a database; retrieval, where relevant documents are obtained based on user queries; and generation, where an LLM synthesises responses using the input request, combining the user query with relevant documents. This reinforced retrieval mechanism allows RAG to incorporate up-to-date knowledge, which in turn improves factual accuracy, contextual relevance, and adaptability.

As a result, RAG is particularly effective for knowledge-intensive tasks that require accurate and reliable information [73].

The RAG includes three phases:

Phase 1: Indexing aims to transform raw text into a structured and searchable format, enabling efficient and accurate retrieval of relevant information in the RAG framework. This process consists of

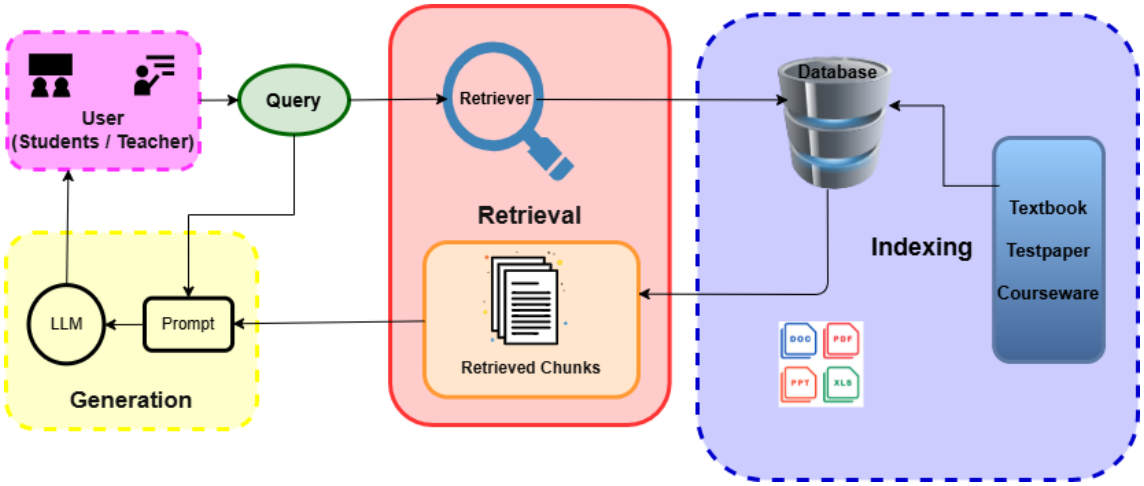


Figure 7: The workflow of RAG adapted from [103].

several fundamental steps, including data pre-processing, text fragmentation, vectorisation, and index storage [56].

The first step in indexing is data pre-processing, which aims to extract clean, structured text from raw data sources. Teaching materials, such as textbooks [57], study programmes [163] and exercise sheets [124], often exist in various formats, including PDFs, PowerPoints, and Word documents [35]. Finally, the processed text and its corresponding vector representations can be organised efficiently for retrieval. This is achieved through the storage of indexes, where each piece of text is assigned a unique identifier, such as a hash or document ID, and stored alongside its vector representation. To support fast and scalable retrieval, specialised vector databases such as FAISS [79], Weaviate2 and Pinecone3 are typically used. These databases enable approximate nearest neighbour searches, allowing for fast retrieval of semantically similar information.

Phase 2: The retrieval phase is responsible for identifying and retrieving the most relevant pieces of text from the indexed database to enrich the contextual knowledge of the linguistic model. The quality and efficiency of retrieval play a crucial role in determining the accuracy and relevance of the generated response. Retrieval methods can generally be classified into sparse retrieval, dense retrieval, and hybrid approaches.

Phase 3: The RAG generation phase is responsible for synthesising responses, integrating the retrieved information with the LLM. The main objective of this phase is to ensure that the generated result is accurate, coherent, and contextually relevant. This section addresses the main aspects of the generation process, including the processing of retrieved information and fusion techniques, as well as strategies for optimising LLM.

2.2.3 Data Fusion in Large Language Model and AI Agents

The evolution of RAG approaches discussed above demonstrates the growing need to combine multiple sources of information in a coherent and contextualised manner. In this context, data fusion emerges as a natural extension of this integrative logic, playing a crucial role in building AI agents capable of reasoning based on diverse and heterogeneous information.

Data fusion is a central process in the design of contemporary intelligent systems, allowing information from multiple sources, modalities, or levels of abstraction to be integrated with the aim of improving the accuracy, consistency, and inferential capacity of models. In general terms, Data Fusion can be defined as the set of techniques that combine heterogeneous data to produce a unified and informed representation of the phenomenon being analysed [60]. In the context of modern AI, and particularly LLM and intelligent agents, this integration plays a decisive role in creating more robust, contextually sensitive and explainable systems.

Historically, data fusion emerged in fields such as robotics, surveillance and multimodal perception, where it was necessary to combine information from different sensors to construct a more complete and coherent interpretation of the environment.

With the advancement of language models and AI agents, this concept has expanded to include forms of semantic and contextual fusion. In other words, it is no longer just a matter of integrating sensory signals, but also of combining linguistic information, facts and context within a single representation model.

In this sense, LLM-based systems have evolved from simple text processors into multimodal reasoning engines, capable of integrating and relating text, images, code, audio and external knowledge bases in a unified manner [22].

Fusion can occur at different levels of processing, traditionally categorised into three: data-level fusion (where sources are combined before feature extraction); feature-level fusion (in which intermediate representations (embeddings) from different modalities are integrated into a common vector space); decision-level fusion (in which multiple models or agents combine their predictions to reach a consensus decision).

A paradigmatic example of data fusion in the context of LLM is RAG framework, originally proposed by [98]. In this architecture, an information retrieval component identifies relevant documents from external knowledge sources, while a generative model, typically based on Transformer architectures [168], produces the final response conditioned on the retrieved context. This process can be understood as a form of dynamic data fusion, in which parametric knowledge (encoded in model weights) is combined with non-parametric, externally stored information. Such integration enables models to leverage both the generalisation capabilities of neural networks and the factual grounding provided by curated knowledge bases [89].

Beyond improving response accuracy and completeness, data fusion plays a crucial role in enhancing the interpretability and trustworthiness of artificial intelligence systems. By explicitly incorporating external sources into the inference process, it becomes possible to trace the origin of generated information and

establish links between evidence and output. This is particularly relevant in high-stakes domains such as education, healthcare, and public policy, where transparency and accountability are essential [17]. In conversational systems, data fusion mechanisms help ensure that generated responses are not solely driven by statistical language patterns, but are instead grounded in verifiable information retrieved from multiple sources.

More broadly, data fusion can be viewed as a foundational element in the evolution of large-scale AI systems, bridging the gap between statistical learning and symbolic reasoning paradigms. Hybrid architectures that integrate retrieval, reasoning, and generation components reflect a shift towards more structured and modular systems, capable of combining heterogeneous information sources in a coherent manner [55]. This trend aligns with ongoing efforts to develop AI systems that exhibit more robust reasoning capabilities and improved factual consistency.

In this context, the study of data fusion not only complements the RAG paradigm but also provides a conceptual framework for understanding the next generation of multimodal and context-aware models. These systems increasingly incorporate diverse data modalities and knowledge representations, moving towards architectures that approximate aspects of human cognition, particularly in terms of integrating and synthesising distributed information [22].

2.3 Agentic AI

Agentic AI refers to AI systems that can pursue complex goals autonomously, plan and execute multi-step actions, and adapt over time with limited human oversight. Recent research spans conceptual definitions, system architectures, LLM-based agents, and domain applications like healthcare [4].

Agentic AI systems operate over long horizons, learn context, make decisions, and carry out complex task chains with minimal supervision, combining autonomy and adaptiveness [4, 11, 101, 171]. hey initiate goal-directed actions, make context-sensitive decisions, and adapt based on feedback rather than just reacting to single prompts[11, 101, 122].

Traditional AI agents, typically single entities, tool-using, linear “sense-plan-act,” suited to well-scoped tasks like customer support or summarization [46, 70, 148]. Agentic AI systems, multi-agent collaboration, dynamic task decomposition, persistent memory, coordinated autonomy, and system-level intelligence [46, 148]. Generative AI, focuses on producing content; Agentic AI adds planning, tool use, and long-term goal pursuit [4, 46, 122, 148].

Agentic AI architectures are consistently described as modular systems combining perception, persistent memory, advanced reasoning/planning, tool use, and orchestration. Types range from simple single agents to complex multi-agent, decentralized, and hybrid paradigms, each suited to different task and risk profiles. The field is converging on layered, orchestrated designs with strong emphasis on safety, governance, and interoperability [178].

2.4 Synthesis of the Conceptual and Technologies

Chapter 2 sets out the conceptual and technological framework underpinning the research, highlighting the evolution of AI from classical ML techniques to more recent approaches based on deep learning and LLM. The main learning paradigms, supervised, unsupervised, semi-supervised, ensemble learning and reinforcement learning, are analysed, highlighting their contributions, limitations and suitability for different types of problems and data. The chapter also addresses the role of deep neural networks as a catalyst for this evolution, culminating in transformer-based models, which introduce a paradigm shift in the way language is modelled, enabling advanced capabilities in generation, comprehension and reasoning.

In an educational context, the potential of LLM as tools to support learning is discussed, particularly in terms of content generation, personalised teaching and the automation of assessment processes. However, significant limitations are also identified, such as the occurrence of hallucinations, reliance on static knowledge and ethical challenges associated with their use. To mitigate these limitations, the RAG paradigm is introduced, which combines generative models with mechanisms for retrieving external information, thereby promoting greater reliability and timeliness of responses. Additionally, different chatbot architectures are analysed, ranging from rule-based approaches to generative and hybrid models, highlighting the trend towards more flexible, contextual and agent-oriented systems.

Against this backdrop, the following chapter presents a systematic review of the literature, focusing on the application of these technologies in the field of education. In particular, it analyses existing intelligent systems, approaches centred on the student profile and affective learning, as well as the use of LLM and RAG in the contexts of assessment, tutoring and content generation.

Finally, Agentic AI refers to a new generation of AI systems capable of autonomously pursuing complex goals, planning and executing multi-step actions, adapting to changing contexts, and learning from feedback with limited human supervision. Unlike traditional AI agents, which typically follow linear workflows and are designed for narrowly defined tasks, Agentic AI incorporates advanced reasoning, persistent memory, tool usage, and dynamic decision-making, often through coordinated multi-agent architectures. While generative AI primarily focuses on content creation, Agentic AI extends these capabilities by enabling long-term goal achievement, task orchestration, and autonomous problem-solving. Current research highlights modular and layered architectures that integrate perception, memory, planning, and orchestration, with increasing emphasis on safety, governance, interoperability, and real-world applications such as healthcare.

This chapter thus enables the work carried out to be situated within the state of the art, identifying gaps and opportunities that underpin the research proposal.

3

Literature Review

The literature review in this thesis is structured in two complementary parts. The first part focuses on the theoretical framework and the evolution of the main concepts related to the subject under study, enabling the establishment of a solid knowledge base and the identification of the most relevant developments in the field. The second part focuses on the most recent literature, with particular attention to the rapid evolution of LLM and their growing integration into advanced computational systems. This update is particularly necessary given the significant impact these models have had on the reformulation of traditional approaches, justifying their incorporation into the proposed system.

3.1 State-of the Art in Intelligent Education

Technology has become a necessary tool for teachers and educators, allowing for more effective lesson preparation, creation of an active repository of student work, and more individualized learning. The use of intelligent systems allows the learning process to be simpler for the student and the teaching process to be more efficient for the teacher.

The development of the IoT and AI can help solve some problems of the traditional teaching and computer-aided system teaching model, and at the same time it makes teaching more motivating. The aim of this study is to compile the most recent studies in this research area and to make a comparative study of the different existing models and applications that make use of this technology to promote more meaningful learning.

Technological developments are filling some gaps in traditional education and improving the computer-assisted learning system, while making learning more exciting. The technology relates objects, sensors, terminals, cloud computing, as well as other equipment, according to pre-established protocols, and subsequently performs remote identification on cell phones or computers [14], supervision, control and positioning, among others. In [189], and based on AIISE, it is believed that intelligent education, that which results from interactive education assisted by AI, changes students' enthusiasm for classroom interaction and makes this teaching model improve students' performance. Improving the traditional standard of

classroom teaching evaluation has become a priority in the field of education. AI-based evaluation methods, the use of intelligent technology in analyzing classroom teaching behaviors can bring more objective, timely and efficient results. The introduction of AI in education aims to improve learning processes and observe students' behavior. The study conducted in [53] takes into consideration parameters such as: student nature, aptitude level, responsiveness and skill sets. In this model students are taught through a pedagogical approach, there is a self-assessment support for educators.

The AI-based classroom entails many challenges, as there are several factors that influence learning. This section surveys smart and intelligent systems for education.

Analysts in education have valued the use of technology in learning, creating intelligent environments aimed at improving learning, and developing virtual boot camps to improve mental health [138]. Such learning environments aim to facilitate students' perception of reality by extending it with computing components such as video, audio, and contextual information generated by immersive augmented reality, IAR technology [26].

In [25] it is mentioned that it is possible to analyze patterns and correlations of biometric data (physiological and neurocognitive) collected from student interaction patterns. By linking the biometric or neurognitive values obtained from the experiments, to the students' interest and the consistency of the human resources and their combinations, it is possible to evaluate and draw conclusions [110].

The model proposed in [25] requires:

1. The management of a variety of data from different sources (text, image, sound);

2. Biometric and neurocognitive monitoring to humanistically measure student performance.

3. Consider the unique condition of the humanities, where there is no single correct answer to a problem, and in how it is possible to address this complexity in computer science.

One should take advantage of non-contact sensors, affective computing, and AI, to establish correlations in student productions. Undoubtedly, prolonged monitoring of emotional states and concentration are important in education and mental health. In [141, 157, 184], the concept of the smart classroom. Not only is it supposed to incorporate smart/mobile devices into the learning experience, but also to equip the physical environment of the classroom with better technological objects. In [14, 59, 164], they reinforce the idea of the importance of individual student learning taking into account their characteristics, by building models that use AI to detect and collect information from students. Through this collection and using different models they try to discover a recommendation system, which is based on the preference of the students, the quality of the materials and the correspondence between the resources and the students. However, this is still an area under exploration. The experimental results obtained in [59] show that the two modules of the model are applicable to the calculation of indicators with different characteristics. According to the performance analysis, the proposed comprehensive model combined with the refined index system proved to be effective for the evaluation of AI-based teaching in the classroom.

In order to identify the state-of-the-art, several questions were posed. Are the existing educational systems intelligent and adaptive? Do these systems take into account affective learning and the student profile? What strategies and tools have already been applied to predict student success using AI? The following subsections aim to answer these questions.

3.1.1 Analysing the existing systems on the education intelligent

In [14] the development of a software system for determining learning styles based on AI is proposed.

Learning styles are used to make a content adaptive. The proposed system is ideated to include more learning theories, for example, Honey and Mumford [67] and VARK's [49]; and, more AI methods, such as Fuzzy Logic, Algorithmic Genetics, BN and Hidden Markov Models.

Learning environments can be at different levels of education. Variations in the data may originate from different courses, such as arts, science and engineering; or due to e-learning environments with text, with multimedia, with aptitude tests or with voice-based interfaces. The selection of the learning theory model for a given environment will be made easier the more learner attributes the model contains.

The following Figure 8 shows the interaction between virtual student and teacher in a cloud environment and how learning styles are determined using the most appropriate model and AI technique.

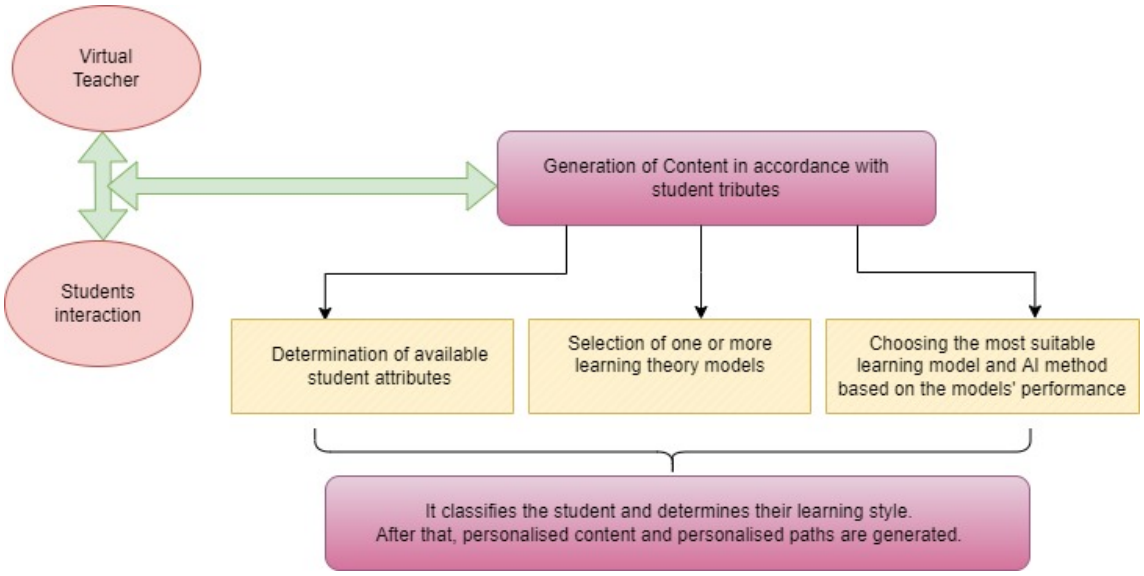


Figure 8: Model adapted from [14] for intelligent education.

In order to adjust the pedagogical learning process to the individual needs of students and teachers in the intelligent environment, it is necessary, according to the study [157], to implement elements that analyze the learning context and adjust the individual student's learning process.

The ambient intelligence model developed in [143] places the user at the center of the system (Figure 9). The user communicates with the ambient intelligence system and the intelligence environment. The ambient intelligence system processes the data received from the environment. The intelligent environment collects and communicates information about the user's presence through multiple sensing

modalities, including sensor readings, RFID technology, facial and voice recognition systems, GPS, and related technologies. The ambient intelligence system then processes this information at two levels: an operational level, which handles basic data acquisition and handling, and an intelligent level, which focuses on higher-level interpretation and decision-making.

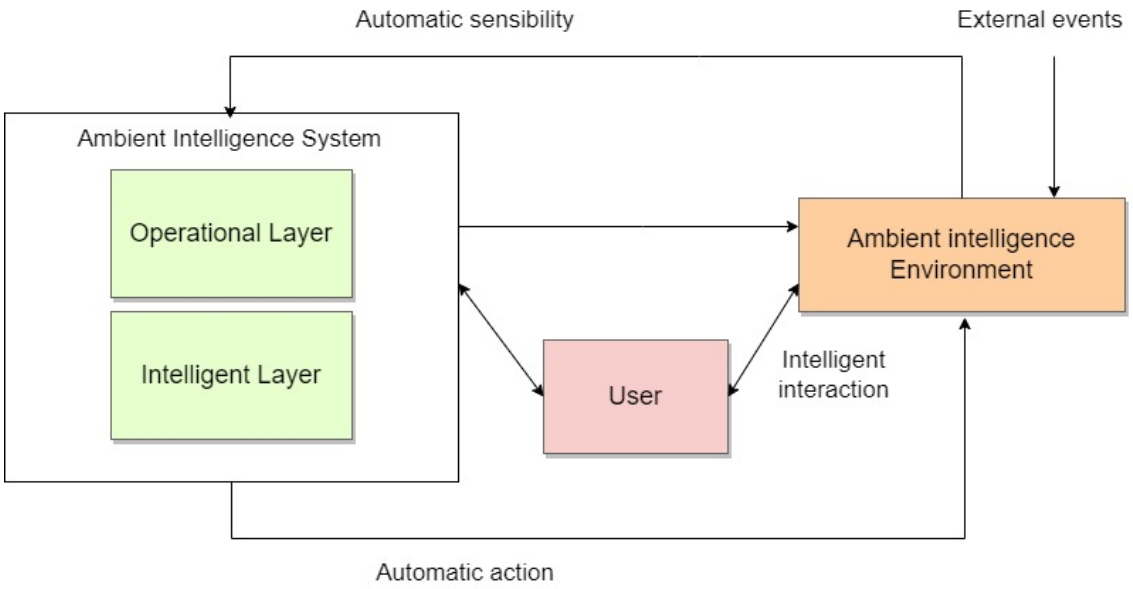


Figure 9: Ambient intelligence model adapted from [143].

These are two examples of educational systems that use AI to adapt learning strategies to the student's style. However, they are not very clarifying as to the factors for grading the student.

The big data technology [68] is seen as key to promote education reform. A modernisation of education is inevitable through the creation of smart classrooms, where the computerisation of education converges on teacher-student interaction and smart education converges on smart thinking of students. Therefore, it is essential to explore the ideals of education for students based on this technology, through building networks, applying social media, developing software, relying on communication tools, and building smart classrooms. communication tools, and building smart classrooms. To this end, we aim to build a smart classroom technology platform to establish teaching objectives and plans, implement interactive teaching, and innovate teaching strategies. Smart classroom users have two behavioural roles, one is the student (Student) and the other is the teacher (Teacher). In addition, students interact with learning resources, contents, platforms and other media in a given learning environment to complete learning the teaching activities. The data used for analysis are obtained by the interactive behaviours. In order to obtain a more correct analysis the data generated, collected and aggregated in the smart classroom, the "triangular STM model" shown in Figure 10 was used to describe the users and their teaching behaviours in the smart classroom.

In the "STM Triangular Model" of smart classroom users, the three vertices represent the student (S), the teacher (T) and the media (M) (which can be resources, content and/or platforms) and structural

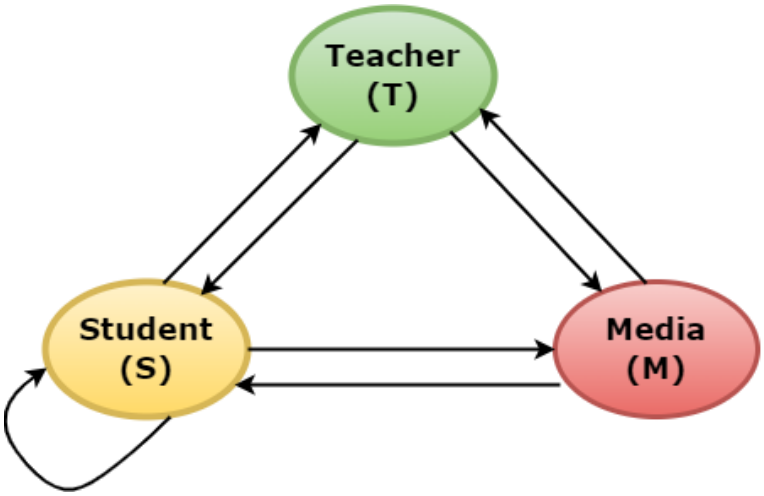


Figure 10: Smart classroom users "STM triangle model" adapted from [68].

relationships also appear. The relationship between the vertices and the edges of the vertices in this model represent the interaction between people and media.

STRUCTURE OF THIS TEACHING PROCESS IN THE INTELLIGENT CLASSROOM

Smart classrooms provide favourable conditions for the integration and union of two important components, 'teaching' and 'learning'. Based on the smart classroom IT platform, the interaction and collaboration between teachers and students will be strengthened, so that "teaching" and "learning" can penetrate and merge into one. The teaching model adopted integrates "three phases and ten stages" of teaching. The so-called "three-stage classroom" is a three-stage closed-loop classroom consisting of "before class, during class and after class". The ten stages of teaching for learning include, for example, real-time detection, summarising and and improvement, homework, micro-class guidance, reflection and evaluation, constitute a complete process of wisdom teaching, unity teaching and learning and sustainable development (as shown in Figure 11).

Below is a brief description of the three phases:

- **BEFORE CLASS:**

  At the lesson preparation stage in the smart classroom, teachers can rely on the mastery of students' acquired knowledge through the teaching information platform to make an objective and scientific analysis and form an accurate academic analysis based on the big data analysis of student performance. Based on the academic situation of each class, the teaching objectives, important and difficult teaching points are defined, and alternative plans that may arise from the teaching process itself are defined. The e-learning platform can collect learning data, such as the length of time learners watch the video, the time they take to do the exercises, and put relevant resources and explanations of errors in the exercises for learners, so that they can learn the

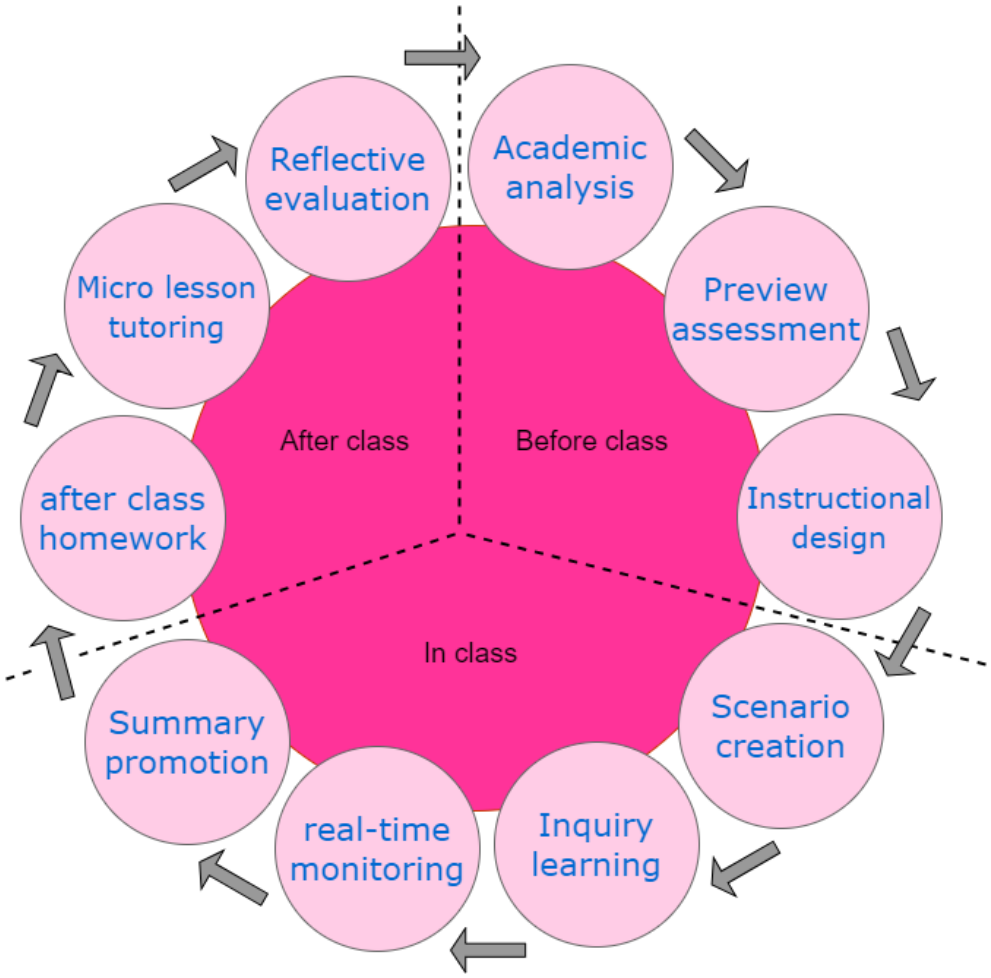


Figure 11: Structure of the teaching process in the smart classroom (in three phases and ten steps) adapted from [68].

knowledge more comprehensively and understand their own learning situation. And generate an individual and academic class learning situation. Teachers can check the completion of students' tasks at any time, quickly understand existing knowledge reserves and accurately identify difficult problems for students either at individual or group level [32], localise teaching objectives and key difficulties according to students' reality and needs, and then propose targeted teaching models and strategies.

- **IN CLASS:**

  In the classroom teaching phase, the traditional lesson becomes monotonous and due to time constraints and the process itself, it becomes impossible for all students to participate in questions and answers, so teachers cannot ensure that every student is integrated into the classroom. However, through smart classrooms, teachers can utilise a variety of information software or platforms to promote teacher-student interaction and student-student interaction. For example, in the introduction part of the classroom, the teacher can publish content and teaching tasks that he/she

wants to assess in tests or through work completed by students. At this stage, when teachers publish interactive tasks through the information platform, students can interact in real time via the Internet and publish their own opinions on the platform. Teachers can give an immediate assessment to the student by instantly providing the answer. Finally, teachers can use the voting function and vote on the degree of mastery of knowledge points learnt during the lesson, which is convenient for students to self-assess and diagnose learning effects, as well as, students can initiate assessments of pedagogical effects to encourage teachers to reflect deeply and improve teaching strategies. The characteristic of this model is that it is built on the basis of classroom teaching process and outcome data, centred on the development of teachers and students [182], which provides an accurate, comprehensive and objective view, fundamental for regular evaluation activities.

- **AFTER CLASS:**

  In the after-class period, teachers can publish homework on the information platform, and set the corresponding time based on the knowledge reflected by the student during the class, then they can put more demanding tasks structured in a guided way, making students autonomous in their learning choices. Next, teachers can make richer resources available on the platform after learning in class and maximise students' interest. Finally, based on the intelligence of the information platform, help in revising assignments. It not only reduces teachers' workload after class, but also allows students to understand the quality of their learning tasks faster. With the help of intelligent terminals such as QQ group, WeChat, and online platforms, students' behaviours are sent to parents, and parents' opinions and suggestions are conveyed to the school in a timely manner. Through the joint efforts of the three parties, it will be possible to find a set of effective ways of tutoring, and achieve it in the shortest time. The effect will be more optimised.

To answer the question "Are the existing educational systems intelligent and adaptive?", the systems analysed are a little intelligent and adaptive, though to some degrees. In terms of adaptivity, [14] uses AI to determine learning styles and make content adaptive to each learner, while [157] adjusts the pedagogical process to the individual needs of students and teachers by analysing the learning context. In terms of intelligence, the ambient intelligence model of [143] places the user at the centre of the system and processes environmental data through sensors, RFID, and facial and voice recognition, while [68] applies big data to model the interactions between students, teachers and media in the smart classroom. These works confirm that existing educational systems sense the learning context and tailor the experience to the individual learner. However, several remain limited in that they do not clarify the factors used to grade the student. The smart classroom of Hu [68] is both intelligent and adaptive, organising teaching into "three phases" (before, during and after class). Before class, it forms an academic analysis from big data on student performance to set targeted objectives for each class's needs [32]; during class, it enables real-time interaction and immediate assessment of each student's mastery; and after class, it tailors

homework and resources to the knowledge each student demonstrated [182]. The system thus senses the learning context and adapts to the individual learner, though it does not fully clarify the factors used to grade the student.

3.1.2 Analysing the systems that considered affective learning and the student profile

Considering various research, the authors [96] developed a learning model in intelligent learning environments that use intelligent environment elements to optimize the learning process. The model provides the student with the appropriate learning strategy that corresponds to their estimated fatigue level. The students' fatigue is estimated based on their involvement in the activities they attended during the day. Upon entering the smart classroom, the student activates the smart classroom system which recognizes and identifies him/her. Information about the student is then sent to the ambient intelligence system which extracts information from the database about the student's daily schedule and other academic activities. Based on this information, the student's fatigue level is estimated. The data on the student's fatigue level adapts the learning process by choosing the appropriate learning strategy for the class.

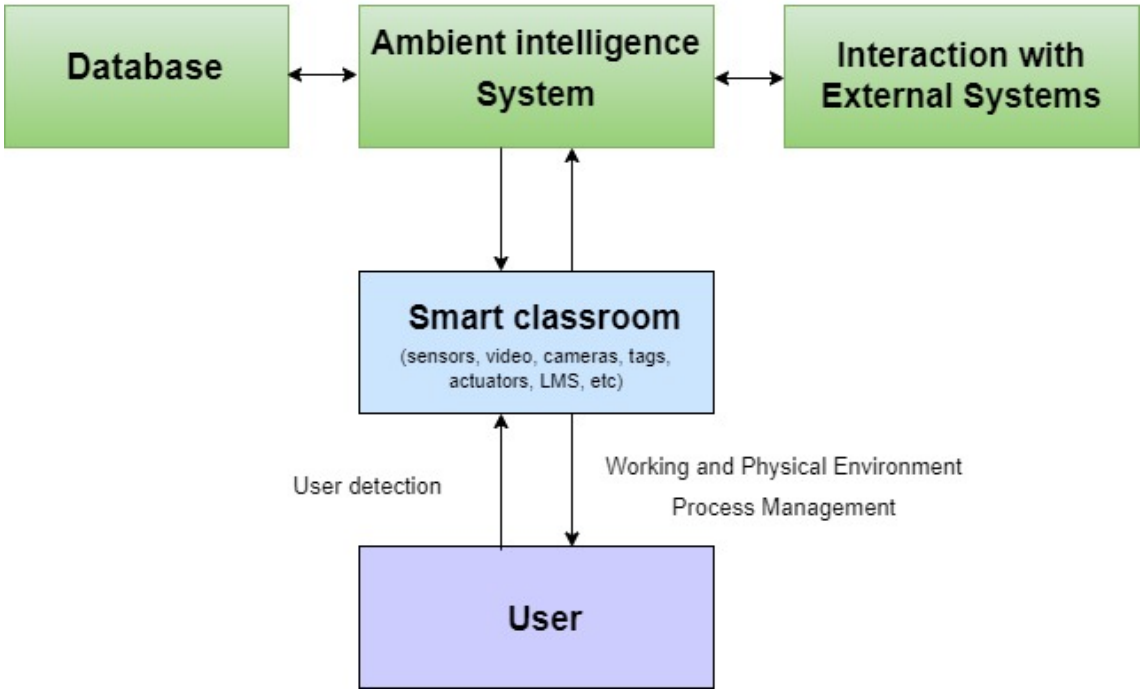


Figure 12: Ambient intelligence system model in a smart classroom based on [141].

The model consists of the following five parts (Figure 12):

- User
- Smart classroom

- Environmental intelligence system
- Data base
- External system integration block.

This model subsequently used in a study that examined the effects of using ambient intelligence models to the smart classroom on students' success in achieving goal-appropriate learning. In [96], the student is detected by the ambient intelligence system through the student's QR code and subsequently uses the information obtained, namely their daily curricular activities, to assess their level of fatigue that day.

Based on this assessment, the ambient intelligence system determines and assigns learning strategies to each individual student (Figure 13). At the end of the lesson, students took a test that measured their success in achieving the learning objectives.

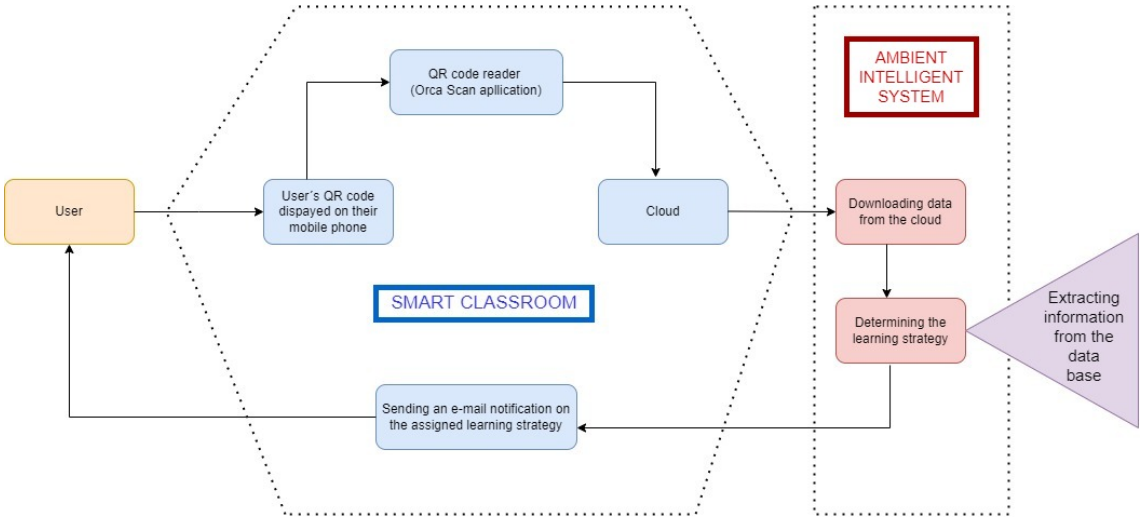


Figure 13: Model adapted from [96, 141].

In the pedagogical evaluation system proposed in [59], firstly, and according to the observation record coding rules in FIAS (Flanders interaction analysis system), the interactive behaviors in the lecture are presented and coded every three seconds, and eight types of datasets are obtained as input base (including teacher movements, teacher emotion, teacher speech text, teacher volume and speed, teacher label, student movement, student emotion, student label). Then, according to traditional classroom evaluation scales, a set of classroom evaluation index systems suitable for AI was designed.

The programme, proposed in [68], is based on the key tasks in the pre-classroom teaching phase, and are the prerequisite and basis for the organisation and implementation of teaching. Intelligent classroom teaching should reflect the fundamental goal of teaching: to teach students as the main ideal, to determine appropriate teaching objectives and to realise which teaching scenarios and cloths for learning so that students actively construct the meaning of knowledge. At the same time, teaching concepts, traditional

and smart classroom teaching methods are utilised and reflected in the teaching model throughout the entire teaching process. The specific design is based on the following:

Pre-selection of curriculum objectives and teaching and learning objectives. Firstly, this new curriculum plan encompasses three three-dimensional objectives: knowledge and skill, process and methods, attitudes and emotional values and should be achieved in the course of curriculum learning. Secondly, each goal of the new curriculum is seen as the goal of teaching and learning. In this conception, teachers should design specific objectives, students' learning objectives and teachers' teaching objectives based on the characteristics of the teaching content of each lesson and the application of the smart classroom information environment, and integrate the three objectives into the smart classroom appropriately. The fundamental concept of the smart classroom is "all students as the center". Therefore, if the teacher does not have a comprehensive understanding of the students in the pre-classroom teaching preparation, every link of the teaching process will deviate from the students' learning needs and the law of growth. The work will be in vain. The intelligent classroom information technology platform provides the student file format and management and maintenance functions, and can statistically analyze the proportion of various useful situations, give constructive teaching suggestions, and provide a basis for formulating an effective teaching program.

Student characteristics files usually include name, student ID, age, gender, personality, cognitive ability, expectations for learning, hobbies, specialties, life and economic experience, cultural and social background, etc. and can be based on student characteristics found in Management files archived in tabular form. Generally, personality refers to the behavior shown by students and can be identified as: active, calm, easy to express, lazy, etc.; cognitive ability refers to the ability of students to process knowledge through their own thoughts, and the words strong, average and weak are generally used; learning expectancy refers to the state of the student's desire to learn which currently contains the students' motivation to learn, which can be described in terms of interest, usefulness, academic performance, teachers' evaluation, and parents' evaluation. Student learning files usually include name, student ID, age, gender, learning content, previous assessment, classroom assessment, homework score, mastery level analysis, etc., and can be archived and managed in the form of student learning files to study the content of this lesson, check students' previous mastery of relevant basic knowledge.

The design principle of the intelligent teaching management system based on facial recognition described in the study [188] is the collaboration between AI and teachers. The camera captures the video information of the student once in 30s and sends it to the server where the algorithm is located. Screen sharing includes the acquisition of screen images, encoding, decoding and transmission. Processes such as visualisation and viewing belong to the domain of streaming media information processing and transmission. To detect faces in images and apply recognition technology to classroom attendance and student status monitoring, neural networks are used to improve the quality of student learning and form a collaborative decision-making model. Face registration technology is applied to the field of education, helping teachers in training students and integrating education management and AI [188]. The two-way

combination can help improve the quality of students' classrooms, assist parents and schools in real-time monitoring of the quality of students' education, and track students in class [188]. In face-to-face teaching, teachers cannot supervise and guide students on site, and the teaching quality of the various live streaming platforms is uneven and the definition of teaching quality is not transparent. Therefore, the quality of students' lessons cannot be truly guaranteed. For this reason, all universities attach great importance to classroom teaching. It is difficult to teach individual pupils, to adjust teaching plans and methods in a timely manner, and to effectively promote the improvement of teaching quality.

Many colleges and universities have started to try to apply various computer programmes to encourage students to attend classes, but this mechanism cannot fundamentally solve the serious problems of inefficiency in the classroom [188], such as students' unwillingness to go to class and their lack of interest in learning the content.

The study in [141] proposes an intelligent classroom management model, builds an intelligent classroom management system, realises the intelligent classroom management system, realises the classroom entry function through face detection, and evaluates the relevant learning emotions of students through facial expression recognition. The results of testing the system and analysing related data show that the system is conducive to teachers' timely and objective assessment of students' individual learning performance. The study points out that the poor quality of teaching is caused by insufficient monitoring of online learning. In the current application process, the face recognition rate is high, which can effectively play the role of classroom supervision. At the same time, the system also has shortcomings. It is not able to collect dynamic video stream data and analyse the static data of the face image alone.

The application of AI in education has been the subject of academic research for over 30 years [61]. Combining AI with learning sciences (such as education, psychology, neuroscience, linguistics, sociology and anthropology) aims to facilitate the development of effective adaptive learning environments and to create various flexible and inclusive tools. Currently, there are several new challenges in the field of education technology in the era of smart phones, tablets, cloud computing, big data, etc., whose current research questions focus on concepts such as personalised learning, mobile learning, educational games, collaborative learning on social networks, MOOC, augmented reality, etc. Therefore, in order to respond to these new challenges in the field of education, several fields of research utilising AI have emerged and have emerged over time to improve teaching and digital technologies.

As well as driving many 'smart' technologies, AIED aims to be a powerful way to explore in detail what is sometimes called the 'black box of learning'. This helps to better understand how learning actually takes place; to analyse the influence of the learner's socio-economic factors, physical context and technology or to study the nature of knowledge and its representation [188]. Determining the most appropriate way to learn and the most effective teaching interaction styles help the learner in their learning, especially when they are used at the right times.

This article [61] reviews the work already done in the area of AI application in education. The earliest AIED stands out for the development of intelligent systems based on modelling elements of learning, such as knowledge of teachers and students, profile of teachers and students through their respective

attributes. Nowadays we have a large amount of varied data on education with the development of Big Data. This has paved the way for new research work, including the field of EDM, where we seek to develop systems that are designed from the exploitation of educational data. In the same context of Big Data, there has also been the emergence of research on the issues of collecting, measuring, analysing and reporting educational data. Thus, we are in the context of developing decision support tools in a learning context.

To answer the question "Do these systems take into account affective learning and the student profile?", the reviewed literature indicates that recent intelligent educational systems increasingly incorporate both affective learning and comprehensive student profiling to personalise the learning process. Several studies integrate learners' emotional states, such as fatigue, engagement and facial expressions, to dynamically adapt learning strategies and improve the learning experience [59, 96, 188]. In addition, these systems build detailed student profiles by combining demographic, cognitive, behavioural and academic information, including learning objectives, prior knowledge, cognitive abilities, motivation, personality traits and academic performance [68]. This information enables intelligent learning environments to provide personalised learning experiences, support teachers' pedagogical decisions and promote adaptive teaching strategies tailored to each learner's individual needs. Overall, the literature demonstrates that modern AI-based educational systems increasingly consider both affective factors and student profiles as fundamental components of adaptive and personalised learning environments.

3.1.3 Strategies and tools that use AI to predict student success

The presented system (Figure 14) includes two evaluation modules, one for statistical modelling (based on AHP - analytic hierarchy process - and entropy weight method) and another for ensemble learning based on AdaBoost.

It also allows completing the mapping between the input data and the index system. By inputting the relevant characteristics of teachers' and students' movements and emotions, the evaluation results are calculated to complete the indicator corresponding to the mapping relation. In the end, based on the results obtained in the performance test, the model that fits the indicators is selected to build a comprehensive classroom assessment teaching.

The issues around emotions and learning are not new, although it has only recently been driven by technology. Affective learning has been a focus of revised systems for more meaningful learning as it can aid decision making and enhance creativity.

According to the universal indicators of LICC and the existing traditional scales, they chose to divide the index system, as shown in Figure 15, into three dimensions: teacher-related indicators, student-related indicators and indicators of standards in the classroom. Through video viewing in the classroom and other means of observation, teaching and research staff evaluate the classroom performance of students and teachers according to the scale of traditional indicators, which are taken as label indicators.

The educational services based on recommendation and search services that have been developed, and which are based on students' preferences, the quality of materials and the correspondence between

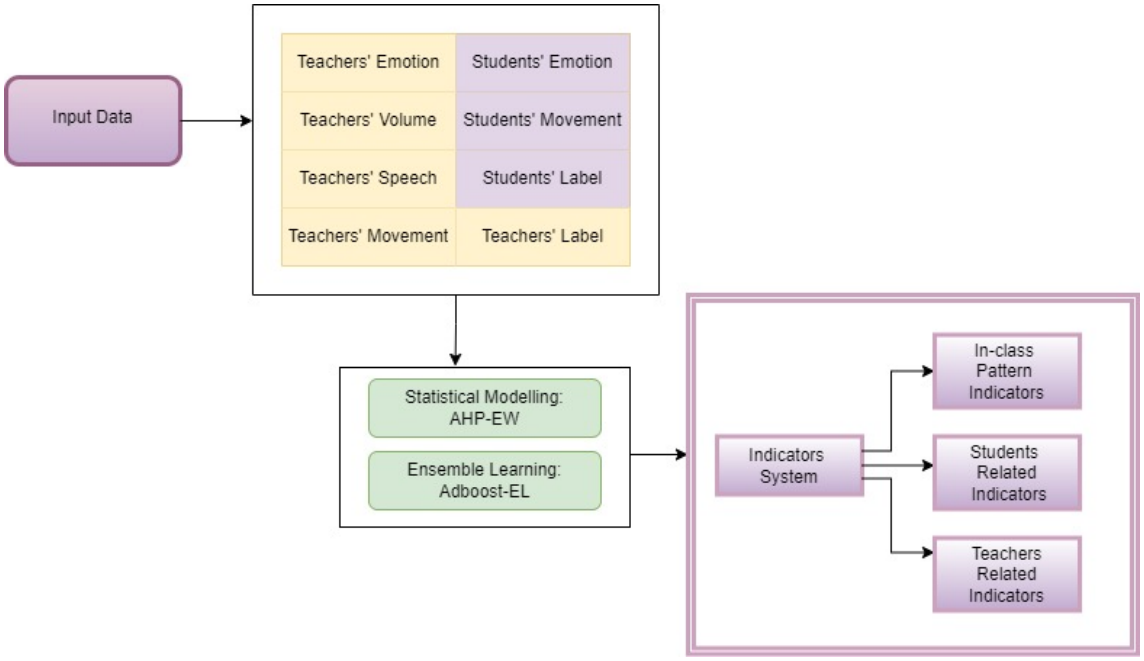


Figure 14: Evaluation model based in [59].

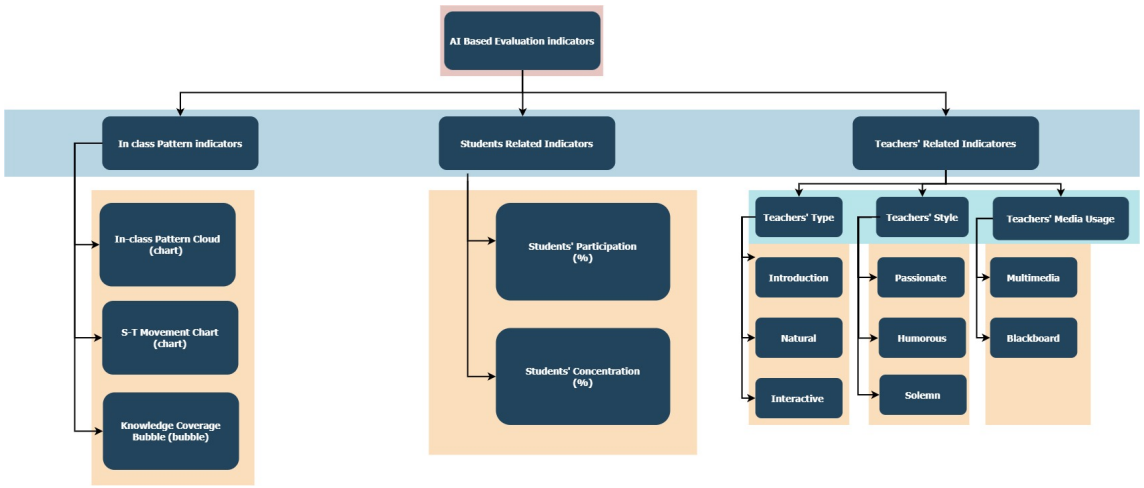


Figure 15: An AI-based evaluation index based on model in [59].

resources and students, are still of poor quality [164]. The software prototype proposed in [164], based on work related to data science, provides, through recommendation, personalized educational content aimed at helping students master their target skills. An example of such a system referred to in this study, is eDoer, which helps students define their learning goals and receive a learning path towards their goals.

The study carried out in [180] proposes to build an intelligent platform that is based on computational intelligence technology, data mining technology and ML technology. The system consists of providing students and teachers with personalised teaching and learning systems. They use AI technology to intelligently analyse what students have learned, build a map of students' knowledge, provide students with personalised learning content and learning solutions, support adaptive learning and implement intelligent recommendations for learning content. By personalising the virtual learner's image to help the learner, it

allows the learner to focus, complete the learning task within the allotted time and optimise the learning process. The virtual companion praises when the learner completes the learning task and supervises and encourages when they do not, so that the learner feels integrated and actively completes the learning task without pressure from teachers and parents. This model comprises three steps:

1. Pre-class teachers prepare lessons through smart teaching platforms where they share lesson plans with teachers across the country in real time, send preview materials to students through the teaching platform, create a personal study space, interact immediately with students, and provide an immediate adjustment and improvement in teaching design. During the lesson, teachers can connect to the teaching platform through various mobile terminals to interact with each other in real time. Teachers can solve the problems of different students, so that all students can participate in classroom communication and truly return everyone's participation. After class, students can complete their homework on the platform, and the platform itself corrects the homework and gives them feedback in real time.

2. The smart learning platform provides different preview materials and tasks of varying degrees of difficulty for each learner, based on collected data and advanced algorithms, according to each learner's learning abilities, subject mastery and effort. Course content is dynamically adjusted according to the learner's progress to achieve adaptive learning. Personalised learning prepares different exam papers for different learners, yet the platform itself generates automatic corrections and learning assessment reports, personalises progress assessment and guides learners' effort.

3. The platform also monitors through comparative analysis. From the teachers' point of view, it analyses how teachers with different teaching ages and backgrounds have different control over the teaching process and the effects of teaching. For the teacher who scores higher in the teaching evaluation, you can analyse in detail their teaching process. For students with poor grades, you can use the survey data to find out when he/she became demotivated, if he/she was not willing to study from start to finish or if he/she had difficulties during the learning process, what caused him/her to withdraw and clearly understand when the student's learning attitude changed. And you can also observe if the learner's attitude changes after receiving the learning alert.

The development of education in China can be divided into three periods: the traditional education phase, the digital education phase and the smart education phase, whose characteristics are reflected in the following Table 4:

Traditional teaching is dominated by teachers. At this stage, whether pupils, teachers or schools, the use of information technology is extremely limited and the teaching activity is done by one person and is mainly carried out autonomously by the teacher.

Digital education is centred on schools. Schools build digital infrastructures and use the the Internet to achieve data interconnection between pupils, classes, schools and regions and to share high-quality educational resources.

Table 4: Characteristics of the different stages in Chinese education. Adapted from [180].

Basic features	**Traditional education**	**Digital education**	**Intelligent education**
Center point	Teacher	School	Student
Learning method	Class and group learning	Digital campus and databases	Personalized learning supported by big data
Information sharing	Group teaching, individual learning	Construction and sharing of school-based resources	Private and personalised services
Technical foundation	Computer + resource library	Internet, data centres	Cloud computing, big data, artificial intelligence
Educational features	Artificial	Ergonomic	Human-centred
Teaching decision	Personal decision	Supported by key data points	Comprehensive support from cloud-based big data

Smart learning is developed on the basis of web-based learning. It is learner-centred and uses cloud computing, big data, AI and other information technologies to record the chain of learner behavioural data in real time. Smart education is personalised, learner-centred teaching that can stimulate learners' potential and enable their overall development. At this stage of education, AI has played an irreplaceable role and is a powerful assistant for teachers and students.

This study concluded that, teachers recognise the importance of AI. 52.1% of teachers consider that AI is very useful for teachers' professional development, 32.3% of teachers consider that AI is more useful for their professional development, and 12.5% of teachers feel more comfortable with AI. The recognition of intelligence is average, and only 3.1% of teachers consider that AI is not useful for their professional development. Overall, as an inevitable trend, AI will develop and grow in education and teaching.

To answer the question "What strategies and tools have already been applied to predict student success using AI?", the reviewed literature shows that a wide range of Artificial Intelligence techniques have been employed to support personalised learning, monitor students' performance and facilitate educational decision-making. Existing approaches integrate machine learning algorithms, data mining, recommendation systems, ensemble learning techniques such as AdaBoost, learning analytics and intelligent evaluation models to analyse students' behavioural, emotional and academic data [59, 164, 180]. These systems exploit multiple data sources, including classroom interactions, learning progress, assessment results, emotional states and learning behaviours, to generate personalised recommendations, dynamically adapt learning content and identify learners who may require additional support. Furthermore, intelligent educational platforms combine cloud computing, big data and AI technologies to continuously monitor learners' progress, evaluate changes in motivation and engagement, and provide teachers with decision-support information for timely pedagogical interventions [180]. Overall, the literature demonstrates that current

AI-based educational systems increasingly rely on predictive analytics, recommendation systems, machine learning and learning analytics to improve learning outcomes and support the prediction of student success through personalised and data-driven educational strategies.

However, although the reviewed studies demonstrate the growing use of AI in adaptive learning and educational decision support, relatively few provide a comprehensive comparison of different machine learning models for predicting student success under the same experimental conditions. Consequently, the following chapters address this research gap by evaluating and comparing several state-of-the-art machine learning algorithms for student success prediction. The experimental analysis aims to identify the most effective predictive models and to provide a clearer and more objective answer to this research question, assessing their predictive performance and suitability for supporting educational decision-making.

3.2 Systematic Review in Intelligent Education

The second part of the literature review was conducted using a systematic review methodology based on the PRISMA statement. The adoption of this methodology is justified by its widespread acceptance within the scientific community, including in the fields of engineering and computer science, as well as by its ability to ensure transparency, reproducibility and rigour in the literature selection process.

In this context, the main steps defined by the PRISMA protocol were followed, namely: the formulation of research questions and the identification of relevant keywords; the construction of the search string for querying scientific databases; the definition of inclusion and exclusion criteria with the aim of refining and filtering relevant studies; the critical analysis of the selected articles; and, finally, presentation and discussion of the results obtained.

This research was conducted in June 2025, using the SCOPUS database (`www.scopus.com`) as the main source of scientific literature. The research process began with the definition of an initial set of keywords, from which the structured search strategy was developed.

The first set of keywords entered was:

```
( TITLE-ABS-KEY ( ( ( llm and agentic ai ) or rag ) AND TITLE-ABS-KEY ( education or classroom or school ) AND PUBYEAR > 2022 AND PUBYEAR < 2026 AND ( LIMIT-TO ( LANGUAGE , "English" ) ) AND ( LIMIT-TO ( SUBJAREA , "COMP" ) )
```

We obtained a total of 142 documents. In the second search iteration, additional filtering criteria were introduced, namely the inclusion of exact keywords such as “Students” and “Teaching”, which made the search more restrictive by focusing specifically on studies directly related to teaching and learning contexts. This adjustment reduced the number of retrieved documents while increasing their relevance to the scope of the study.

(TITLE-ABS-KEY ((llm and agenticAI) or RAG) AND TITLE-ABS-KEY (education or classroom or school) AND PUBYEAR > 2022 AND PUBYEAR < 2026 AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (EXACTKEYWORD , "Students") OR LIMIT-TO (EXACTKEYWORD , "Teaching"))

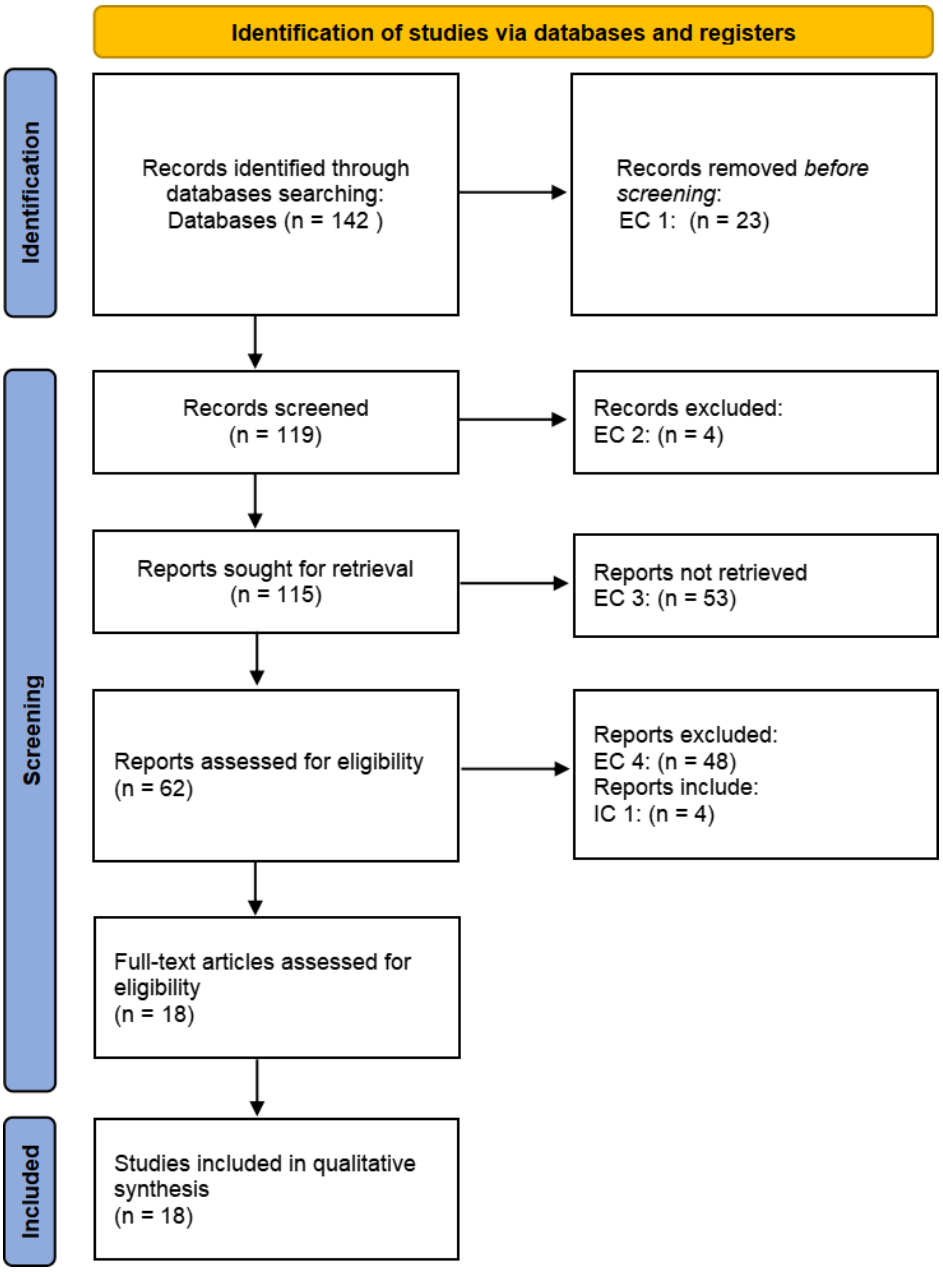


Figure 16: PRISMA Flow Diagram

The PRISMA flow diagram shown in Figure 16 was used to support the analysis of the SLR, beginning with the study identification phase, based on the articles retrieved from the selected databases and the respective searches conducted. The inclusion and exclusion criteria applied in the selection of articles are described below and are also illustrated in the PRISMA flowchart shown in Figure 16.

In order to refine the set of publications identified, explicit inclusion and exclusion criteria were established. Accordingly, documents were excluded whenever they met any of the following criteria:

EC 1: The articles were not related to education

EC 2: The articles were not related to the subjects Large Language Model (LLM) or Retrieval-Augmented Generation (RAG) or Agentic AI

EC 3: The articles were not related to students or teaching

EC 4: After analysing the abstract, the excluded articles dealt with specific cases of higher education or were too generic or not free

IC 1: Important articles for building architecture

Initially, 142 records were identified through scientific databases, of which 23 were subsequently removed based on preliminary exclusion criteria (EC1), such as duplicates or obvious irrelevance to the topic. During the screening phase, the remaining 119 records were analysed based on their titles and abstracts, leading to the exclusion of a further 4 studies (EC2). Subsequently, 115 articles were considered for full-text review, of which 53 could not be retrieved (EC3), reducing the pool for eligibility assessment. At this stage, 62 studies were analysed in depth, resulting in the exclusion of 48 articles for failing to meet the defined criteria (EC4), and the preliminary inclusion of 4 studies (IC1). Subsequently, 18 full-text articles were assessed for final eligibility, culminating in the inclusion of 18 studies in the qualitative synthesis. This structured process ensures that the selected articles are relevant, consistent with the research objectives, and based on explicit inclusion and exclusion criteria.

Once the selected articles had been obtained, they were systematically analysed and subsequently categorised into four thematic areas: The use of Retrieval Augmented Generation in educational applications; Use of LLM and RAG in assessment; LLM and RAG-powered multi-agents and frameworks that uses LLM and RAG. The papers included under each of these themes were organised into specific subsections, with the aim of structuring and facilitating the reading of this section of the literature review.

3.2.1 The use of Retrieval Augmented Generation in educational applications

3.2.1.1 RAG for Educational Applications: A Systematic Survey

The authors of the article [103] conduct a systematic review of the use of RAG in educational applications. This technique combines LLM with external information retrieval, ensuring greater accuracy, updating, and relevance of AI systems in education.

The challenge of LLM in education is that they can generate incorrect information (hallucinations), their knowledge is static, and they have difficulty adapting to individual needs. The solution is to use RAG, which organises external sources (books, articles) into vector databases, retrieves relevant information for the user to consult, and finally, the LLM synthesises the response using the retrieved data.

RAG improves factual accuracy, enables dynamic knowledge updates, and increases transparency (source citations).

This powerful synergy can be applied in various areas of education. For example, it supports the development of interactive learning systems and educational chatbots (such as MoodleBot and ChatEd), as well as intelligent tutoring systems (such as KG-RAG for medical training). It is also widely used in content generation, including the automatic creation of multiple-choice questions, lesson plans and summaries. Furthermore, it enables automated assessment, providing personalised feedback on assignments and supporting the marking of essays. Finally, it extends to broader educational ecosystems, where it can assist in policy-making processes and help predict school dropout rates.

The main challenges are combating hallucinations, which requires human verification mechanisms, updating knowledge bases (dynamic sources), optimising scalability models, and incorporating images, videos, and code.

RAG emerges as a promising solution to make educational AI more reliable and adaptable, but it requires advances in reliability, efficiency, and multimodal integration to reach its full potential.

The authors present recommendations for future research, highlighting the need to: develop robust mechanisms for detecting hallucinations in AI-generated responses; implement dynamic educational repositories capable of automatically updating themselves with recent information; optimise computational costs, ensuring the feasibility of large-scale implementation.

These measures aim to improve the reliability and efficiency of RAG-based systems in the educational context.

3.2.1.2 Artificial Intelligence Generated Content

The study carried out in the article [179], conducted by researchers at Zhalantun Vocational College in Inner Mongolia, China, investigates how Artificial Intelligence Generated Content (AIGC) can revolutionise digital textbooks in vocational education. The main focus is on the application of RAG technology to improve the interactivity and educational effectiveness of digital teaching materials.

In general, the study analyses how AIGC and LLM are transforming the creation of educational content, especially in the context of professional education. The focus is on combining digital textbooks with a RAG-based Question and Answer (Q & A) system to provide personalised, accurate answers tailored to courses and students.

The research begins by exploring the development of AIGC technology, highlighting how it has transformed content creation from human producers to AI systems. LLM are identified as central elements, having two basic functions: self-learning and output generation.

Current models use deep learning neural networks that simulate the functioning of the human nervous system. Therefore, model adjustments must be made based on real application scenarios to perform the

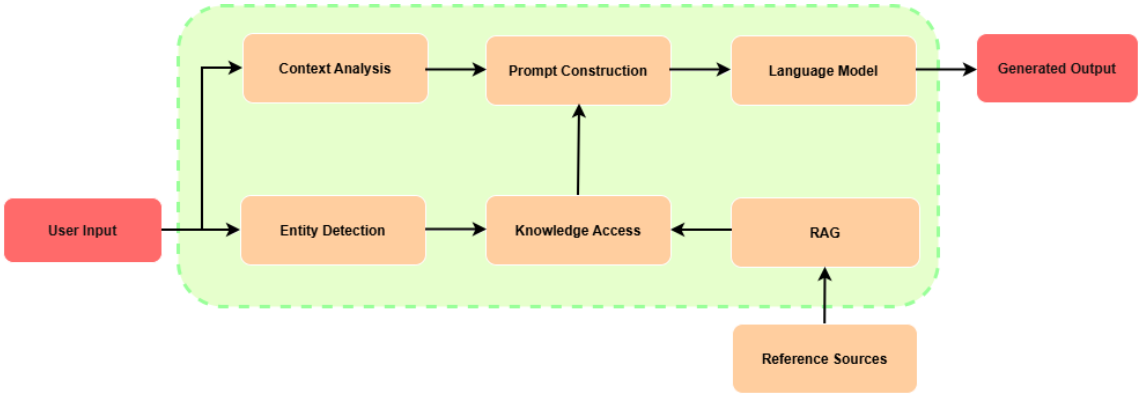


Figure 17: Process for generating content using artificial intelligence adapted from [179].

generative artificial intelligence function with specific functions adapted to the characteristics of vocational education courses. Once the model has been adjusted, the process of using AIGC by users, as shown in Figure 17, consists of triggering the model's output function through 'questions' or 'commands' and issuing it according to the process based on natural language knowledge and knowledge indexing.

Although these models can process large training datasets, they face limitations in producing content specific to professional domains, requiring adjustments based on specific professional education application scenarios.

The study presents different theories about digital textbooks, highlighting the "complex body theory" that defines these materials as a new category of textbook, constructing a complex "Content + Software Platform + Terminal + Tools" and creating a digital learning environment for students.

The main risks identified in the transition from traditional books to digital format are:

- ✓ Content fragmentation: non-linear access can disrupt the pedagogical sequence.
- ✓ State of tension: excessive stimuli can negatively affect the student.
- ✓ Preference for printed material: due to familiarity and ease of use.
- ✓ Reduced legibility: dispersion of content across multiple resources reduces focus and interest.

To mitigate these risks, the study recommends that digital textbooks be:

- ☞ Interactive and customised for specific courses and students;
- ☞ Developed with iterative, flexible, and collaborative training;
- ☞ Designed to maintain consistency between the book's organisation and the logic of connection with the teaching content.

The core of the research is the RAG-based knowledge retrieval system, which combines school textbooks and cooperative company manuals as a knowledge base to generate answers to students' questions, enabling:

1. Building the knowledge base:
    - Text recognition and segmentation.
    - Cleaning of irrelevant data.
    - Vectorisation (embeddings) and storage in a vector database.
2. Knowledge recovery:
    - Vectorisation of the student's question.
    - Search for most similar blocks.
    - Merging of knowledge retrieved with the query.
    - Generation of response by LLM.

The system was designed with four main prompt components:

- General Q & A scenario prompts;
- Question categorisation configuration prompts;
- Knowledge base reference templates;
- Internet search results templates.

This model offers several advantages:

✓ Precise personalisation: Responses tailored to the course context, maintaining accuracy and completeness.

✓ Educational flexibility: Enables decentralised, non-linear, personalised and autonomous learning.

✓ Improved interactivity: Educational chatbots activate points of knowledge and offer multiple perspectives on a problem.

The integration of AIGC into digital textbooks for professional education represents a paradigm shift in the way we learn, ensuring accuracy and relevance in responses, making answers to student queries always relevant, accurate, and personalised, with a low threshold for use, high customisation, and interpretable data, and better preparing students for the demands of the modern job market.

3.2.1.3 Development of Interactive Assistance for Academic Preparation Using Large Models Language

The field of education has seen remarkable improvements in both efficiency and capabilities with the incorporation of innovative and cutting-edge technologies to enhance the experience of both teachers and students. When discussing the latest technologies, it is impossible not to mention LLM. The authors of the article [95] aim to build a conversational platform application, integrating the use of LLM executed locally using Ollama and LangChain, to be used by students for assisted learning. The model also uses augmented retrieval generation techniques to create a large, up-to-date, and dynamic language model. To promote transparency, cutting-edge open-source technologies are preferred over closed-source technologies such as OpenAI's ChatGPT.

The work developed in [95] consists of an interactive academic assistance platform based on locally executed LLM, using Retrieval-Augmented Generation (RAG) techniques to ensure accuracy, dynamic knowledge updating, and data privacy.

The work arises from the need to overcome the limitations of proprietary models, particularly issues related to:

- Data privacy: Closed models raise concerns about the security of sensitive data [6, 40, 155].
- Third-party dependency: The use of external APIs creates dependencies that can affect institutional autonomy.
- Customisation limitations: Proprietary models offer less flexibility for adaptation to specific needs.

The proposed solution consists of using open-source LLM (Llama3), avoiding dependence on external APIs and ensuring privacy, and RAG to combine text generation with up-to-date information retrieval.

The tools used are:

- **Ollama:** For local execution of LLM.
- **LangChain:** Framework for integrating LLM, vector databases, and APIs.
- **FAISS (Facebook AI Similarity Search):** For semantic search in vector embeddings.
- **Flask:** For creating the communication API between the frontend and LLM.

It also uses vector databases to store embeddings of academic texts for RAG.
The architecture shown in the Figure 18 consists of:

1. Local LLM Deployment: Llama3-7B running via Ollama, adapted for modest hardware.
2. Dynamic Knowledge Base: Academic texts converted into vector embeddings and stored in a vector database (e.g., Milvus).

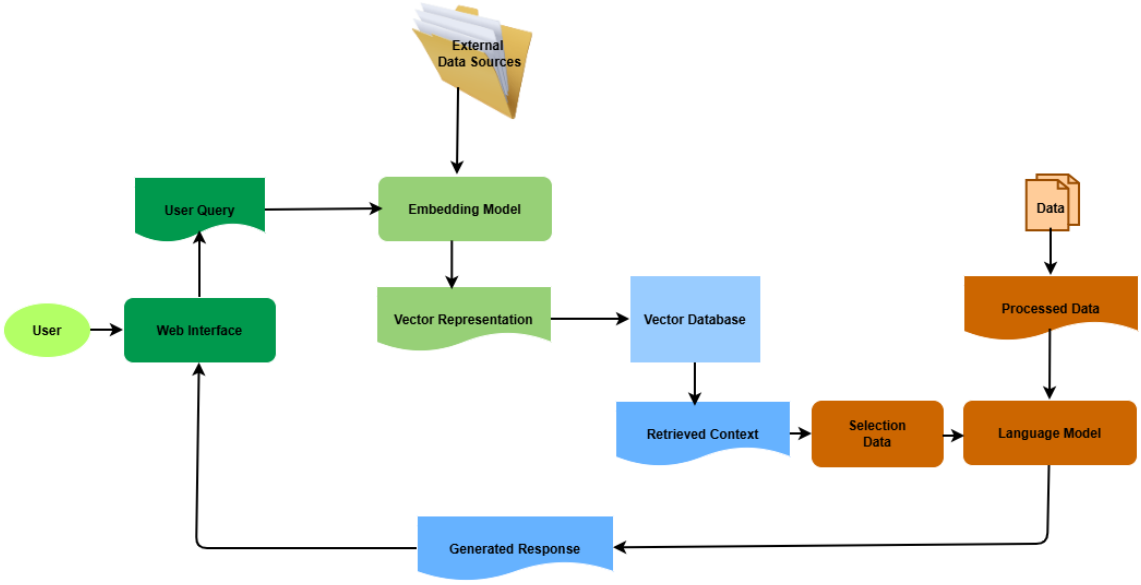


Figure 18: Architectural model adapted from [95].

3. Conversational API: Developed with LangChain + Flask to process queries, retrieve relevant data via RAG, and generate responses.

4. Web Interface: Interactive frontend for students to submit questions and receive contextualised responses.

Using this platform, it was observed that Llama 3 achieved better performance than other open-source models, such as Gemma, across benchmark evaluations including MMLU, which assesses multitask language understanding, and IFEval, which measures a model's ability to follow instructions. The limitation found was that the platform still had lower accuracy than ChatGPT, but this was offset by its privacy and customisation features. It was possible to increase the efficiency of RAG, as there was a reduction in hallucinations when retrieving updated information from the knowledge base, enabling accurate answers to complex academic questions.

The authors point out some advantages such as improvements in privacy, as sensitive data is not shared with third parties, cost reduction (the local solution is economically viable in the long term) and increased flexibility, as this platform is adaptable to different academic disciplines.

One of the main challenges is the limited availability of computational resources, which restricts the deployment of larger models, such as Llama 3-70B. Another important limitation is the need to keep the knowledge base continuously updated to ensure that the information provided remains accurate and relevant. To address these challenges, future research could explore more cost-effective hardware solutions, such as affordable high-performance GPUs, extend support for multimodal data (including images and mathematical expressions) in technical disciplines, and develop mechanisms for automatically updating the underlying knowledge base.

3.2.1.4 Evaluating Large Language Model for Student Dropout Prediction

This study [45] investigates the innovative use of LLM to predict student dropout rates in higher education, overcoming the limitations of traditional ML methods. The authors propose an integrated approach that combines RAG and Few-Shot Learning (FSL) to improve prediction accuracy, using rich and adaptable contextual data.

When an advisor or teacher enters a student's name or ID into the system, the query is transformed into semantic vectors. RAG retrieves relevant information from institutional databases, such as academic history, class attendance, and socioeconomic indicators, and integrates it into a contextualised prompt. The LLM processes this data and generates not only a prediction about the risk of dropout but also interpretable insights into the underlying reasons.

To ensure the accuracy and relevance of predictions, the study employs prompt engineering, a fundamental technique that guides LLM to focus on risk-indicative elements, using carefully selected examples ('shots') to improve model generalisation even with limited datasets [135]. In addition, textual analysis and semantic vectorisation are used to interpret complex data, complementing traditional variables such as socioeconomic and academic factors [2].

Although the article does not explicitly detail the sample size, it highlights the importance of careful selection of historical dropout data, with an emphasis on the quality of pre-processing and the selection of relevant features, such as the "studentship" feature, which improves prediction especially in the early stages of the semester [15]. The method was validated through comparative evaluations, demonstrating that LLM outperform conventional methodologies in predictive accuracy and textual analysis of complex datasets [162].

Despite advances, issues such as data quality and the need for greater model interpretability remain challenges. Future research should focus on refining these techniques, expanding data sources, and integrating real-time information to make predictions even more accurate and actionable.

The integration of LLM into student attrition prediction represents a significant leap forward in the ability of educational institutions to proactively identify and support at-risk students. By combining analytical precision with interpretive richness, this approach opens up new possibilities for creating more inclusive and effective educational environments where every student has the opportunity to succeed.

3.2.2 Use of LLM and RAG in Assessment

3.2.2.1 Enabling Educators to Build RAG-Based AI Chatbots

The study [129] follows a Design-Based Research approach (Reimann, 2010), which combines iterative development with continuous evaluation in real educational contexts. The authors implemented cycles of:

- Prototyping: Creation of a RAG system based on low-code templates.

- Classroom testing: Use of the system by students and educators in different disciplines (e.g., computer science, artificial intelligence).
- Feedback collection: Through questionnaires (Likert scales and open-ended questions) and analysis of interactions with the bots.
- Iteration: Technical and pedagogical adjustments based on the results.

The authors identify several limitations of current AI-based educational tools. In particular, they argue that general-purpose chatbots, such as ChatGPT, are not always well suited for educational settings, as they may generate inaccurate responses, deviate from the intended curriculum, or be susceptible to attacks such as prompt injection. They also highlight the strong dependence on commercial platforms, which often involve high operational costs and offer limited flexibility for adaptation to specific educational needs. To address these challenges, the authors propose Specialised AI Tutor (SKIT), a framework that integrates several complementary technologies, including:

The proposed solution, SKIT, uses various technologies, such as:

✎ Framework: LangChain + Flowise (drag-and-drop interface to create RAG flows without code).

✎ LLM: Llama 3.1 (open-source, hosted locally or in a secure cloud).

✎ Vector Database: Stores embeddings of educational materials (e.g., slides, PDFs).

✎ Web Interface: SKIT (simplified frontend for students).

In this solution, RAG works in three phases:

1. **Indexing:** Educational texts are divided into chunks and converted into embeddings.
2. **Retrieval:** The system searches for relevant excerpts in the vector database using semantic similarity.
3. **Generation:** The LLM (e.g., Llama3) generates responses exclusively from the information retrieved from the knowledge base, thereby reducing the likelihood of hallucinations.

The proposal presented in this article consists of a low-code solution, using tools such as LangChain and Flowise, so that educators with moderate technical knowledge can create and adapt educational chatbots.

It uses LangChain as a framework to orchestrate LLM with RAG flows and Flowise: a low-code interface that simplifies the creation of flows with retrieval and generation. The prototype created is called SKIT (Spezialisierter KI Tutor) and is hosted on local HPC servers to ensure privacy. Initial tests with Llama (local version) and GPT-3.5 demonstrated feasibility.

The methodology used in [91] was inspired by the Will, Skill, Tool, Pedagogy model in which each of these dimensions encompasses, respectively:

- Will: motivating teachers to experiment.
- Skill: basic training in generative AI and RAG.
- Tool: template in Flowise.
- Pedagogy: practical use examples, such as bots for definition queries (knowledge bots) and algorithm explanations (explain bots).

The solution was tested with educators and students. The results obtained were as follows:

1. Tests with Educators: Results from a workshop conducted with 23 teachers indicated that 86% considered SKIT easy to adapt to their teaching subjects, whereas 91% expressed a preference for SKIT over commercial tools, including OpenAI's Custom GPT.
2. Tests with Students: The solution is robust because 0% success in prompt injection attacks (attempts to divert the bot from the topic) and the accuracy is high because 98% of responses were objectively correct (minimal errors in calculations).

The study aims to empower educators to create specialised bots with RAG, ensuring: control over content (avoiding hallucinations and thematic deviations); robustness against attacks (e.g., prompt injection); pedagogical customisation (e.g., bots for review, tutoring or explanation of concepts). It offers a practical and innovative solution to the challenges of AI in education, balancing pedagogical control, ease of use, and security. It differs from others by focusing on the educator as an active agent in the creation of AI tools, uniquely combining RAG + low-code to democratise access and highlight effective measures in real environments.

3.2.2.2 Revolutionizing Assessment with RAG and LLM

The article [160] explores how generative AI technologies, especially RAG combined with LLM, can revolutionise the educational assessment process by providing automated, impartial, and near real-time feedback. The authors highlight the traditional challenges of manual assessment, such as excessive time, teacher fatigue, and the possibility of human error or bias. The article proposes using AI to automate and improve this process, ensuring greater efficiency, objectivity, and personalised feedback for students, alleviating teachers' workloads, and making the assessment process fairer, more scalable, and more informative by integrating data from various sources (student responses, attendance, practical tasks, etc.).

The architecture has seven main stages and combines various data sources into a cohesive system:

1. **Input Sources:** Digitised responses (via OCR), digital books, Student Management System (SMS) data such as attendance, grades, assignments.

2. **Embedding Models:** Transform text (and other formats) into vectors for efficient processing.
3. **Vector DB:** Stores embedding vectors, enabling similarity-based searches.
4. **Query via LLM:** Uses embeddings to generate contextualised feedback and automatic grades.
5. **Indexing and Query:** Using techniques such as HNSW or IVF, facilitating efficient search for relevant data.
6. **Assessment via LLM:** The model evaluates responses considering context, correct answers, and student history.
7. **Persistence:** Results (grades, feedback) are returned to the SMS.

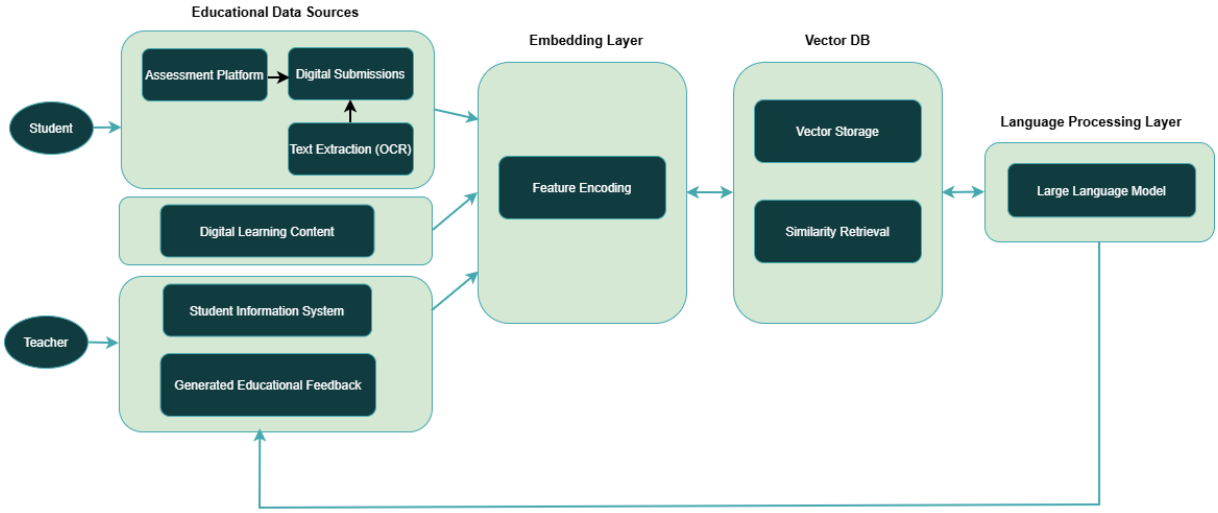


Figure 19: System Architecture adapted from [160].

The system, as can be seen in the Figure 19, uses contextual embeddings, which are more appropriate than simple word embeddings. The processed data is stored in Vector DBs such as Pinecone or Chroma, allowing for quick similarity searches without retraining the model.

The system compares students' answers to previously indexed content and answer keys. It uses personalised feedback based on the student's individual history (previous grades, performance in activities, etc.).

Tests have shown that the system significantly reduces assessment time from 8 minutes per test (manual assessment) to 2-3 minutes, with the possibility of human review in less than 5 minutes. In addition, the system offers personalised feedback, identifying errors and suggesting improvements based on the student's individual performance.

However, the authors also point out limitations, such as the possibility of the system assigning high scores to incorrect answers that contain relevant keywords, due to semantic proximity in the vector

database. To mitigate this problem, they recommend combining the efficiency of AI with human supervision, ensuring fair and accurate assessments.

In conclusion, the article demonstrates that the integration of RAG and LLM can transform educational assessment, making it faster, more objective, and tailored to the individual needs of students. This approach not only alleviates the workload of educators but also improves the quality of feedback, contributing to a more effective learning process. The research paves the way for future applications of AI in education, highlighting the importance of balancing automation with human oversight to ensure reliable results.

3.2.3 LLM and RAG-powered multi-agents

3.2.3.1 Taxonomy of LLM-Powered Multi-Agent Architecture

The LLM have revolutionised the field of AI. However, when faced with complex interconnected tasks that require deep and interactive thinking, LLM have revealed their weaknesses. Autonomous multi-agent systems powered by LLM represent a strategic response to these challenges. Although these architectures have a promising future, it is necessary to find the right balance between different levels of autonomy and alignment. The objective of the article [64] is, therefore, to define a comprehensive multidimensional architecture.

Although LLM are highly skilled at handling large amounts of textual information and have rapid processing and pattern recognition capabilities, they still have great difficulty in maintaining consistent logic in extended chains of reasoning. LLM therefore encounter difficulties when dealing with more complex and interconnected tasks [92, 173].

These limitations of individual LLM have led to the exploration of more sophisticated and flexible AI architectures, including multi-agent systems that divide complex tasks into subtasks, distributed among specialised agents, promoting cognitive synergy. These systems address user-requested objectives through a divide-and-conquer strategy, breaking them down into smaller, more manageable tasks.

Taxonomies represent structured classification schemes used to categorise objects in a hierarchical manner according to specific criteria.

The taxonomy is organised into three dimensions:

1. Levels of Autonomy (x-axis):
    - (Static): Predefined rules, no adaptation.
    - (Adaptive): Adjustable behaviour within a fixed framework.
    - (Self-organising): Dynamic agents, capable of learning and adapting in real time.
2. Levels of Alignment (y-axis):
    - (Integrated): Built-in, immutable alignment mechanisms.
    - (User-Guided): Customisation prior to execution.

- (Real-Time Responsive): Adjustments during operation with human feedback.

3. Architectural Viewpoints (z-axis):

- Task Management: Decomposition, orchestration, and synthesis of tasks.
- Agent Composition: Types, roles, memory, and agent network.
- Multi-Agent Coordination: Communication protocols and action management.
- Interaction with Context: Use of external resources (data, tools).

Below, Figure 20 we briefly describe the main architectural features of autonomous multi-agent systems powered by LLM.

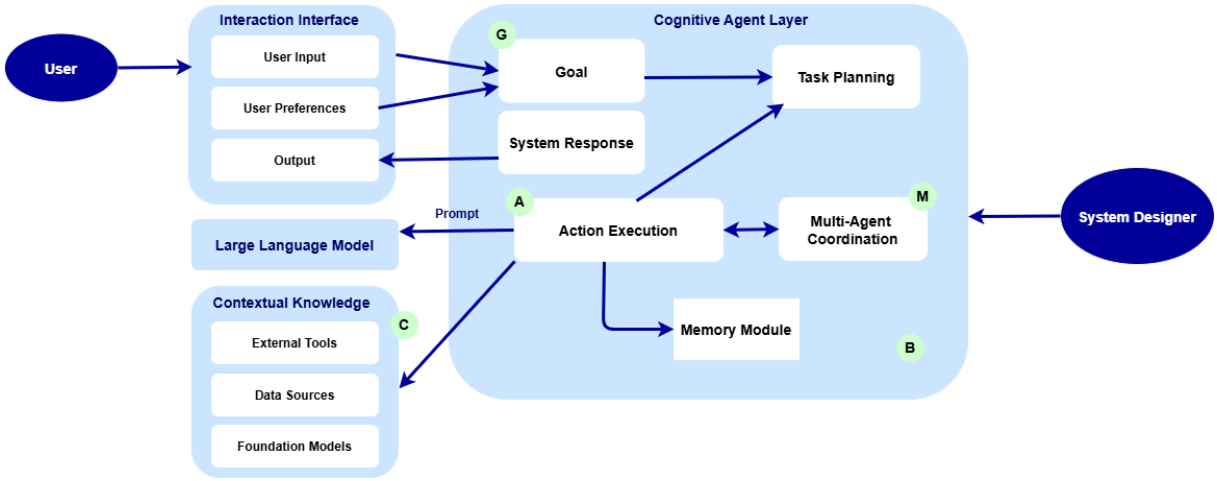


Figure 20: The main features of autonomous multi-agent systems powered by LLM adapted from [64].

✓ Goal-oriented task management (G): Divides complex tasks into manageable subtasks, distributing them among specialised agents and orchestrating the synthesis of results.

✓ Intelligent agents with LLM (A): Each agent has skills, a defined role and its own memory, using LLM for reasoning, planning, accessing contextual resources and communicating with other agents.

✓ Multi-agent collaboration (M): Agents exchange messages to delegate, assist and evaluate, combining strengths in a "cognitive synergy".

✓ Interaction with context (C): They use external resources (tools, data, specialised models) to expand capabilities and perform complex tasks.

✓ Balance between autonomy and alignment (B): The aim is to harmonise the strategic self-organisation of agents with human values and intentions, balancing operational freedom and governance rules, a critical factor for the effectiveness of the system.

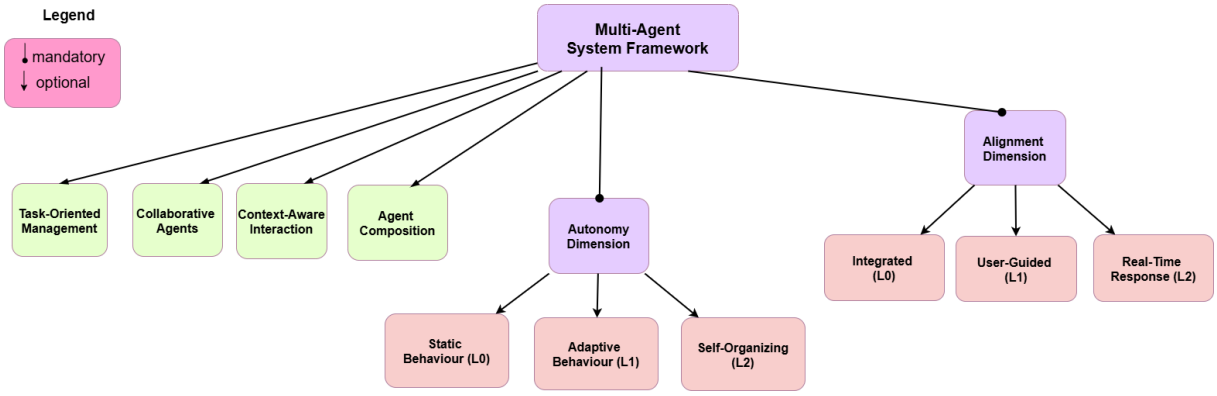


Figure 21: Taxonomic structure with characteristics adapted from [64]. Each point of view integrates levels of autonomy and alignment.

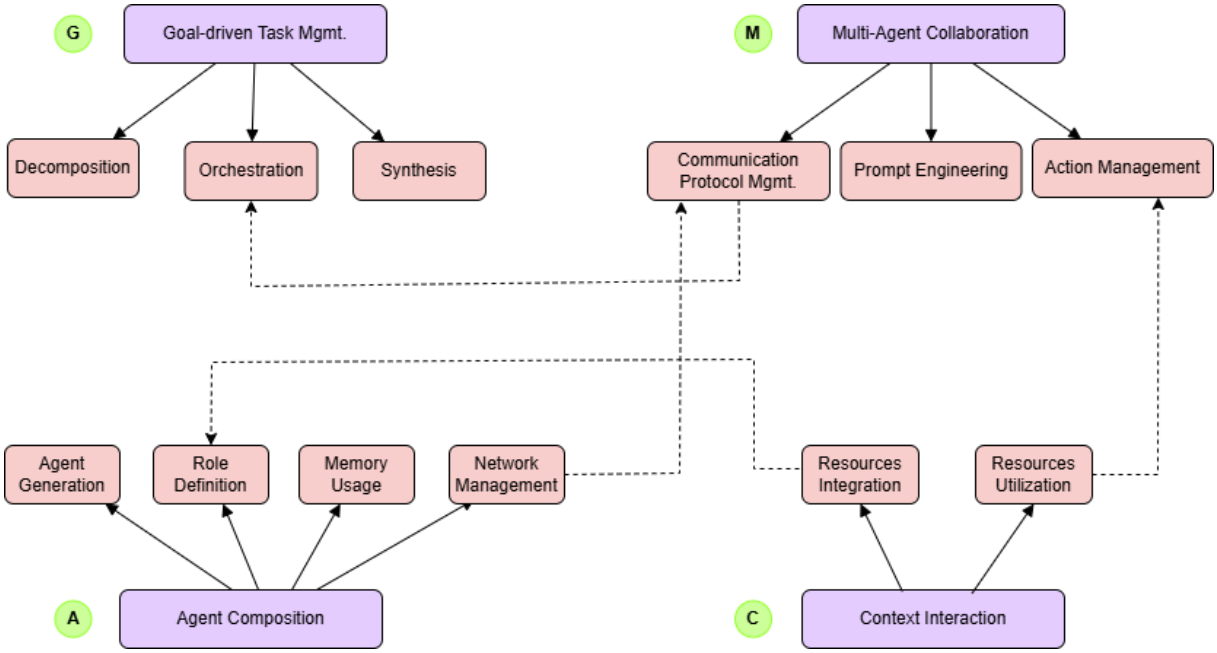


Figure 22: The diagram illustrates the aspects and mechanisms of a more specific point of view (b-e) adapted from [64].

Figures 21 and 22 presents an overview of the characteristics of the proposed taxonomy, structured using a feature diagram [16, 152].

Figure 21 shows the general taxonomic structure, based on a specific viewpoint, where each of the four viewpoints combines different levels of autonomy and alignment.

In Figure 22, this structure is refined by specific aspects of the viewpoint and their interdependencies in terms of requirements-oriented dependencies (adapts-to). These dependencies presuppose a system with high autonomy, in which the capabilities of an aspect evolve according to the needs of the aspect to which they are linked. Finally, each aspect can also be evaluated according to levels of autonomy and alignment, producing a more detailed and accurate taxonomic classification.

3.2.3.2 LLM Multi-Agent Framework for Intelligent Tutoring

Intelligent Tutoring Systems (ITS) have revolutionised education by offering personalised learning experiences. However, as goal-oriented learning, which emphasises the efficient achievement of specific goals, becomes increasingly important in professional contexts, existing ITS often struggle to offer this type of goal-oriented learning experience.

This article [172] proposes GenMentor, a multi-agent framework with LLM designed for goal-oriented learning: it identifies skill gaps, dynamically models the student's profile, schedules adaptive learning paths, and generates personalised content with an explore draft integrate cycle, showing clear gains in skill mapping accuracy, path quality, and content personalisation.

Traditional ITS or LLM chatbots tend to be reactive or follow static, non-goal-oriented curricula (e.g., professional goals). However, GenMentor focuses on proactively guiding students towards a specific goal (e.g., mastering the skills needed for a job or task).

To address the complex subtasks mentioned above, we distribute responsibilities across multiple LLM agents, allowing them to collaboratively manage different tasks. This system begins by accurately assessing the skill gap between the student's current skills and the desired goal.

The proposed system begins with a rigorous assessment of the skills gap between the student's current level and the defined learning objectives. Once this discrepancy has been identified, a personalised learning pathway is generated using an evolutionary optimisation method, with the aim of simultaneously maximising the efficiency of the learning process and the student's level of engagement. In addition, the system selects and generates content tailored to specific objectives, ensuring that it remains up to date and suited to individual needs, thereby promoting a focus on the most relevant materials. Throughout the process, the student's profile is continuously updated based on their interactions, enabling dynamic adaptation that reflects their constantly evolving progress, preferences and needs.

Each of these stages is analysed in detail below, with the aim of identifying their components, describing the underlying mechanisms and highlighting their respective contribution to the overall functioning of the proposed system.

1. **Skill Gap Identification**

   This is where the skills gap is identified and the bridge between the student's cognitive state and the desired objectives is built. Accurate mapping of these objectives requires the LLM to understand the nuances of the objective, given its often abstract and high-level nature. However, direct requests from the LLM can produce irrelevant or incomplete skills, and for this reason, a set of customised data of objectives for skills was constructed with Chain of Thought (CoT) [173] to fine-tune the LLM, improving the alignment of the objectives of the identified gaps.

2. **Adaptive Learner Modeling**

   Understanding the student's status is the central foundation of Intelligent Tutoring Systems (ITS), enabling personalised learning through content adaptation and targeted feedback. In this model,

the potential of LLMcomprehensive profile of the student (U), structured in three fundamental dimensions:

- **Cognitive State** It monitors knowledge acquisition by tracking progress and assessing skill mastery. This dimension is represented by:
  - Set of skills already mastered by the student;
  - Skills pending acquisition;
  - Quantitative metrics of progress and level of mastery.

  After each session, the system updates this status based on test results and explicit feedback from the student, identifying gaps to guide future learning activities.
- **Learning Preferences:** Captures individual differences in knowledge absorption, including:
  - Preferred content styles (e.g., concise summaries vs. detailed explanations);
  - Favourite types of activities (e.g., passive reading vs. interactive exercises).
    This data allows for dynamic adaptation of teaching methods, optimising engagement and comprehension.
- **Behavioural Standards:** Analyses interaction data to identify trends that impact engagement, such as:
  - Frequency of system use;
  - Temporal distribution of sessions (e.g., irregular or prolonged sessions).
    Atypical patterns (e.g., sporadic use or excessive time per session) trigger proactive interventions, such as difficulty adjustment or motivational messages, to realign the student with their goals.

The LLM acts as the procedural core of this model, synthesising the three axes to generate:

- Real-time updated profiles;
- Hyper-personalised recommendations;
- Alerts for human intervention when necessary.

This holistic approach aims to transcend traditional static models, creating an adaptive learning ecosystem that evolves in sync with the student's cognitive needs, preferences, and behaviours. The post-analysis of each session ensures a continuous cycle of improvement, where identified deficits become focal points for the next educational steps.

An illustrative example is presented in Figure 23.

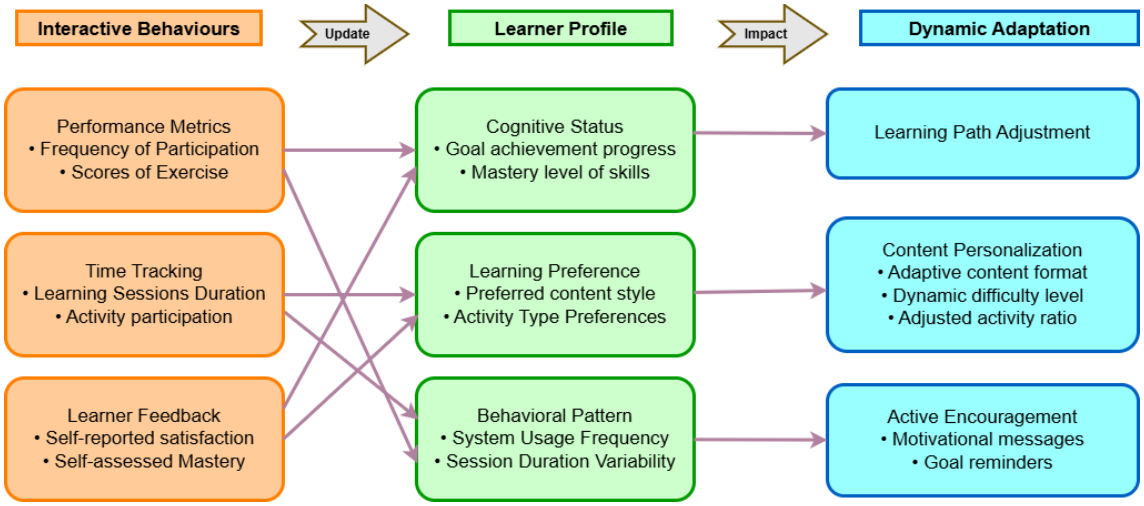


Figure 23: Illustration of the dynamic methodology of the student adapted from [172].

3. Personalized Resource Delivery

To effectively address identified skill gaps, GenMentor employs a personalised and adaptive content delivery method that dynamically aligns resources with each learner's unique profile and progress.

For each learning session l_i, the content creator C_i adapts the content using an exploration-elaboration-integration mechanism to improve the personalisation and targeting of resources. During the learning process, a learner simulator that utilises the dynamic learner profile U_t is used to mimic learner feedback on learning resources for refinement, enabling adaptation of the learning path and content.

The results indicate that GenMentor consistently outperforms the baseline approaches. Compared with existing methods, it achieves better alignment between learning objectives and skills, reflected in higher recall, precision, and goal alignment scores. In addition, the system generates more progressive and engaging learning pathways while producing high-quality, personalised educational content through an exploration–RAG–draft–refinement pipeline supported by a learner simulator. The successful implementation of the system, together with positive feedback from education professionals, further demonstrates its potential for practical use in real-world educational settings.

Despite the progress made, there are still some limitations and risks to consider. The system relies on source data, such as job advertisements, which means that any errors or inconsistencies in these sources may be reflected in the skills mapping. Furthermore, whilst the trainee simulator is useful, it remains only an approximation and does not replace direct feedback from real people. Even with mitigation strategies, isolated cases of 'hallucinations' may still occur, so human validation remains necessary. There is also a need to pay attention to computational costs and privacy issues associated with the storage of profiles and records. Finally, the apprentice representation model could still be improved, as it received slightly lower ratings in terms of performance.

To maximise impact and reduce risks, it is recommended that GenMentor be adopted in professional contexts with active human supervision, ensuring curation and verification of critical content. It is important to monitor key metrics in production, such as accuracy in skills mapping, acceptance of learning paths, content quality, and effective learner progress, and combine the simulator with human feedback to improve personalisation. Finally, a robust data governance policy should be implemented, ensuring anonymity, retention management, and informed consent from users.

Overall, with its adaptive design and learner-centred approach, GenMentor demonstrates the transformative potential of LLM in advancing personalised, goal-oriented education.

3.2.3.3 LLM Agents: Advances in Education

The article [36] provides a systematic review of LLM-based agents applied to education and proposes a taxonomy by tasks that can be categorised into two broad classes: (1) Pedagogical Agents, which focus on automating complex pedagogical tasks to support teachers and students; and (2) Domain-Specific Educational Agents, which are adapted to specialised domains such as science teaching, language learning, and professional development.

First, memory allows LLM agents to retain both students' study habits and the short-term context of real-time interactions, improving contextual understanding and ensuring personalised learning experiences across various educational tasks [190].

Secondly, the use of tools allows LLM agents to access external resources, perform complex calculations, and obtain information in real time, enabling them to automate complex educational tasks such as grading, knowledge retrieval, and adaptive content generation, thereby overcoming limited interactivity and increasing engagement [56].

Thirdly, planning supports structured learning by breaking down complex topics, predicting optimal learning paths, and dynamically adjusting teaching strategies, allowing agents to autonomously guide students through personalised learning experiences [69].

By integrating these capabilities, LLM agents not only improve understanding, engagement, and personalisation, but also automate complex educational workflows, making learning more adaptive, efficient, and scalable.

Education, a domain essentially focused on individualised learning and problem solving, is one of the ideal scenarios for implementing LLM agents. To significantly improve learning, LLM agents must integrate several key capabilities, including memory modules, tool usage, and planning, as illustrated in Figure 24.

The memory module allows the agent to retain an understanding of fundamental concepts, follow up on key points, track the learner's previous interactions and progress, and store personalised profiles to provide accurate responses.

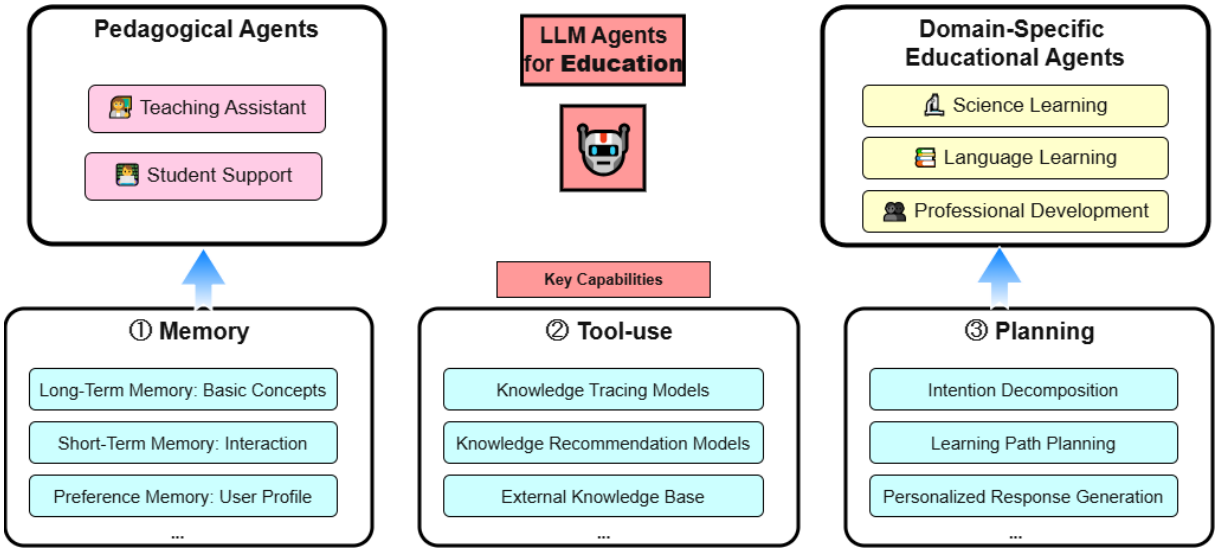


Figure 24: Comprehensive overview of LLM agents for education adapted from [36].

The tool usage module allows the agent to access specialised educational tools, such as knowledge tracking and recommendation models, as well as external resources such as search engines, thereby expanding their functional capabilities and ensuring up-to-date and accurate information.

3.2.3.4 GPTutor: AI-Powered Intelligent Tutoring

The article [109] presents GPTutor, an ITS based on generative artificial intelligence (GenAI), developed to offer an interactive and knowledge-based learning experience. It was implemented to support university students through contextualised responses based on class materials, using an approach known as RAG.

The system integrates a RAG pipeline to minimise typical generative AI problems, such as hallucinations (incorrect responses) and lack of context fidelity. It is also indexed at two levels: it divides learning materials into large and small chunks (pieces of text), optimising information retrieval and response generation. Another feature is that each response includes links to relevant pages of class materials and suggestions for follow-up questions.

Thus, the heart of GPTutor is the RAG system, which consists of three main steps:

- **Indexing:** lesson materials are pre-processed into 'chunks' (blocks of text) with a hierarchical structure:
    - Parent chunks: larger, preserve context.
    - Child chunks: smaller, optimise retrieval accuracy.
    - Chunks are stored as vectors in a vector store, with metadata useful for filtering (e.g. page number).

- **Retrieval:** when the student asks a question:
    - The system detects the intention, reformulates the question (for clearer context), generates a vector, and retrieves the most relevant chunks based on vector similarity.
    - The small-to-big retrieval mechanism uses child chunks for accuracy and parent chunks for additional context.
- **Generation:** with the retrieved data:
    - The language model generates a coherent response, with reference links to the materials and follow-up questions.

GPTutor incorporates the four classic modules of an ITS:

- **Expert module:** where learning materials (slides, notes) are stored.
- **Tutoring module:** uses RAG to respond to students' questions adaptively.
- **Student diagnostic module:** allows the teacher to monitor students' progress and engagement.
- **User interface:** provides natural interaction via chatbot with answers, links, and suggested questions.

The results of the pilot implementation are particularly encouraging. A study involving 129 students on Software Engineering courses found that those who interacted most intensively with the virtual tutor, exchanging more than 10 messages, achieved significantly higher academic performance in their final exams. On average, they scored around 10 percentage points higher than their peers with a lower level of engagement.

This link between use of the platform and academic success suggests that GPTutor does not merely function as a passive reference tool, but can play a significant role in facilitating more active and consistent learning.

Despite the promising results, the researchers identified areas for improvement. Some students, particularly at the postgraduate level, expressed a desire for the system to be capable of generating more in-depth and original analyses, going beyond merely reproducing the content of the study materials.

After testing the model, the authors pointed out some limitations, such as: excessive repetition of class content; lack of deeper discussion, i.e., limitation in the system's ability to promote analytical or critical discussions; and rigid interaction in some scenarios, particularly when questions were not aligned with the materials.

Looking ahead, the GPTutor team plans to expand the platform to other subjects, incorporate gamification elements to increase student engagement, and develop features that allow for more diverse representations of information, including diagrams and data visualisations. These innovations have the potential to radically transform the landscape of digital education, creating virtual tutors that combine academic rigour with the flexibility and accessibility of modern technology.

3.2.4 Frameworks that uses LLM and RAG

3.2.4.1 LAMB: AI Assistants for Learning Management Systems

LAMB [8] is an open-source modular framework for creating learning assistants that combines rapid prompt engineering, RAG, and knowledge base construction from educational content (including video). Development followed an iterative minimum viable product approach, and initial validations indicate that LAMB assistants can improve pedagogical practices, increase student engagement, and personalise learning experiences. Notable features include ease of use, scalability, security, and interoperability with existing LMS. It was implemented in a case study ('Macroeconomics Study Coach') and validated in real scenarios with positive results. A qualitative and quantitative assessment of the use of the assistant by students and teachers was carried out. The LAMB framework, developed in Node.js with SQLite storage and deployment via Docker, integrates LLM models such as GPT-4, Llama3, and Mistral with vector databases (ChromaDB) to offer personalised support on LMS platforms, including Moodle, Blackboard, and Canvas. The objectives of this framework consist of three pillars: Facilitate the creation of educational AI assistants through an intuitive framework, eliminating the need for programming knowledge; ensure seamless integration with LMS platforms (Moodle, Blackboard, Canvas) for scalable deployment; validate the effectiveness of the framework in real educational environments, ensuring practical applicability.

The system relies on two main external components:

1. AI infrastructure (commercial providers or self-hosted models such as Llama3/Mixtral);
2. institutional ecosystem (LMS such as Moodle, Canvas, Sakai, Blackboard).

At the centre is the LAMB Core, which orchestrates the user interface, AI infrastructure, knowledge base components, and LTI integration with the LMS (Figure 25).

Summary of LAMB CORE's main components:

- **Prompt engineering:** creates, tests, and parameterises prompts for RAG and augmentation plugins; allows the assistant to be published via LTI with an Assistant key.
- **Content Library Manager:** encompasses materials (transcripts, documents, web pages), fragments them into chunks, generates embeddings, and builds the Knowledge Base.
- **REST API:** access points compatible with the OpenAI call structure; requires an Assistant Key to authorise use.
- **Pipeline Manager (PM):** executes assistant pipelines based on prompt configuration, integrating retrieved chunks, augmentation plugins, and LLM layer.
- **RAG Engine:** retrieves evidence from the base for more accurate and contextual responses.
- **Augmentation plugins:** customisable extensions that shape the assistant's behaviour.

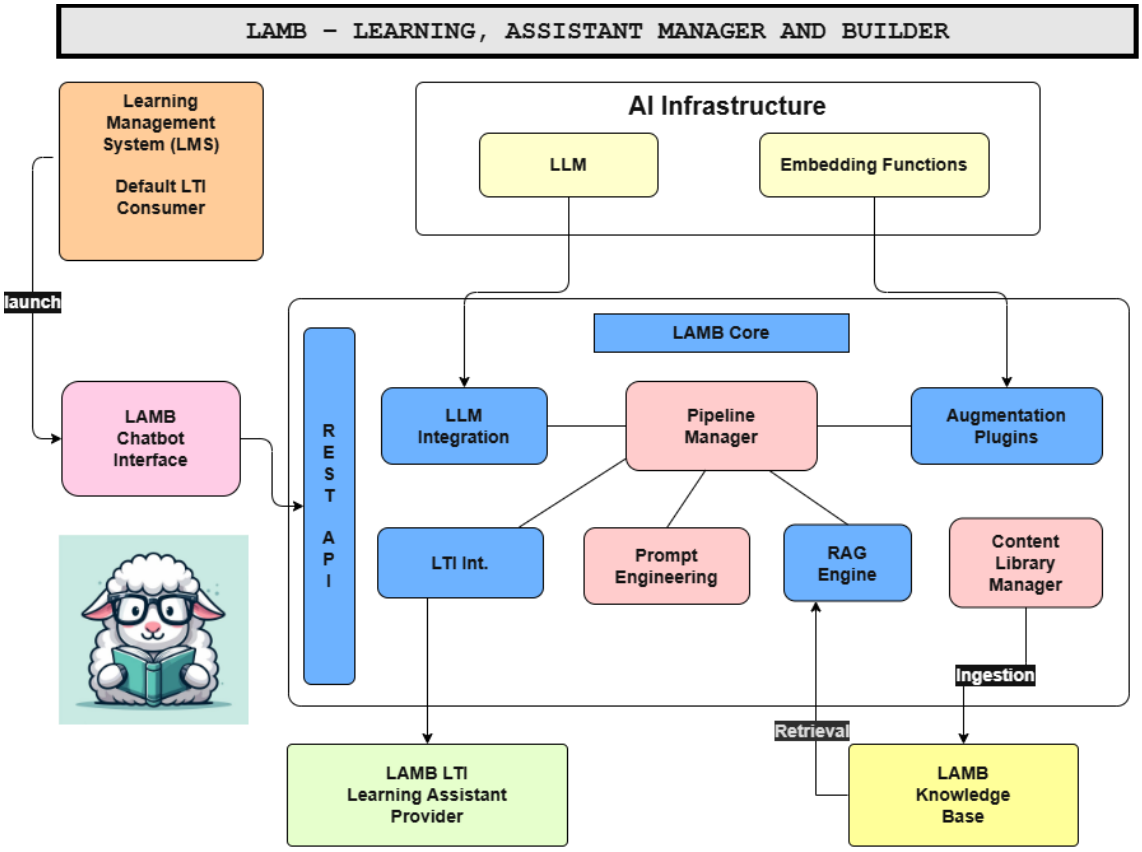


Figure 25: Lamb system architecture adapted from [8].

- **LLM integration layer:** interface with multiple LLM/providers.
- **LTI integration:** publishes each assistant as an LTI tool; allows notes, reports, and feedback to be sent to the LMS.

The Knowledge Base stores chunks and metadata (embeddings) to support RAG. LTI integration ensures authentication, context, roles, and the communication of notes/feedback between assistants and the campus LMS.

The architecture is based on Docker containers and microservices to facilitate deployment, maintenance, isolated updates, scalability, and security.

In summary, LAMB is a modular, integrable, community-oriented platform that facilitates the responsible and scalable incorporation of AI into educational environments.

3.2.4.2 MerryQuery: Trustworthy LLM Support for Education

The authors of the article [161] propose an AI-based educational platform, MerryQuery, powered by LLM, designed to reconcile pedagogical utility with safeguards against academic misconduct, misinformation,

and overconfidence. By relying on RAG and multimodal processing, it offers answers anchored in course materials, transparency (citations), and an interaction flow that mimics step-by-step human tutoring. Thus, RAG provides course-specific answers based on materials provided by teachers and citations from sources to reinforce transparency and trust. Teachers have access to logs and controls over the granularity of responses, making it an education-oriented alternative to generic chatbots.

Key features:

1. Contextual and course-specific answers, supported by RAG.

2. Guided dialogues and step-by-step answer structure (tutor style), inspired by human tutoring (Guided Learning Dialogue Design), promoting critical thinking and autonomous problem solving.

3. Quotes from sources and interaction logs to increase trust and verifiability.

4. Tools for teachers:

    - **Chat History Monitoring:** MerryQuery's conversation logging system stores each student-AI interaction in MongoDB with a timestamp. These records are accessible through the instructor interface, allowing them to review student queries.
    - **Usage Statistics:** The platform also displays real-time statistics for each course. This can help instructors monitor the frequency of conversations initiated in the courses they offer.

5. Adjustment of response granularity: Practice Mode (provides hints without revealing the complete solution) and Review Mode (provides complete solutions and detailed explanations).

6. Multimodal support (documents with images/tables) without data loss or misinterpretation.

7. Memory mechanisms inspired by human tutoring (Guided Learning Dialogue Design), promoting critical thinking and autonomous problem solving.

8. Multimodal Data Processing: MerryQuery's current pipeline processes complex PDFs containing text, tables, and images. First, it repairs the PDFs using a custom function that applies garbage collection, compression, and cleaning techniques to ensure that the files are correctly formatted. Next, the system incorporates the extracted text and OCR-processed image data into a vectorised database to be retrieved through the RAG model.

9. Memory Mechanism for Coherent Conversations: Since the OpenAI API has not been available, MerryQuery implements a custom memory mechanism to maintain context in a conversation. This mechanism identifies and retains relevant past interactions, ensuring a coherent conversation flow.

Additional features are being developed, such as automatic classification, RL from human feedback (RLHF), a comprehensive logging system for researchers, and the generation of personalised teaching and learning strategies.

MerryQuery combines relevant content retrieval RAG with LLM to produce informed responses; integrates multimodal processing to preserve information in images and tables; and exposes controls and logs that enable pedagogical alignment and institutional oversight. From a technical standpoint, it integrates the OpenAI API (GPT-4o), LangChain, MongoDB, and a React.js front-end. It is fully implemented and in the refinement phase, with plans to add features such as automatic assessment, RLHF, and customised generation of teaching/learning strategies.

This platform improves confidence in responses (through citations), promotes critical thinking among students, increases alignment between the tool and pedagogical objectives, and enables the handling of complex resources without compromising information fidelity.

MerryQuery is a solution focused on transforming LLM into responsible educational tools: it combines AI techniques (RAG, multimodality) with transparency and pedagogical control mechanisms to reduce risks (misconduct, misinformation, overconfidence) while enhancing learning.

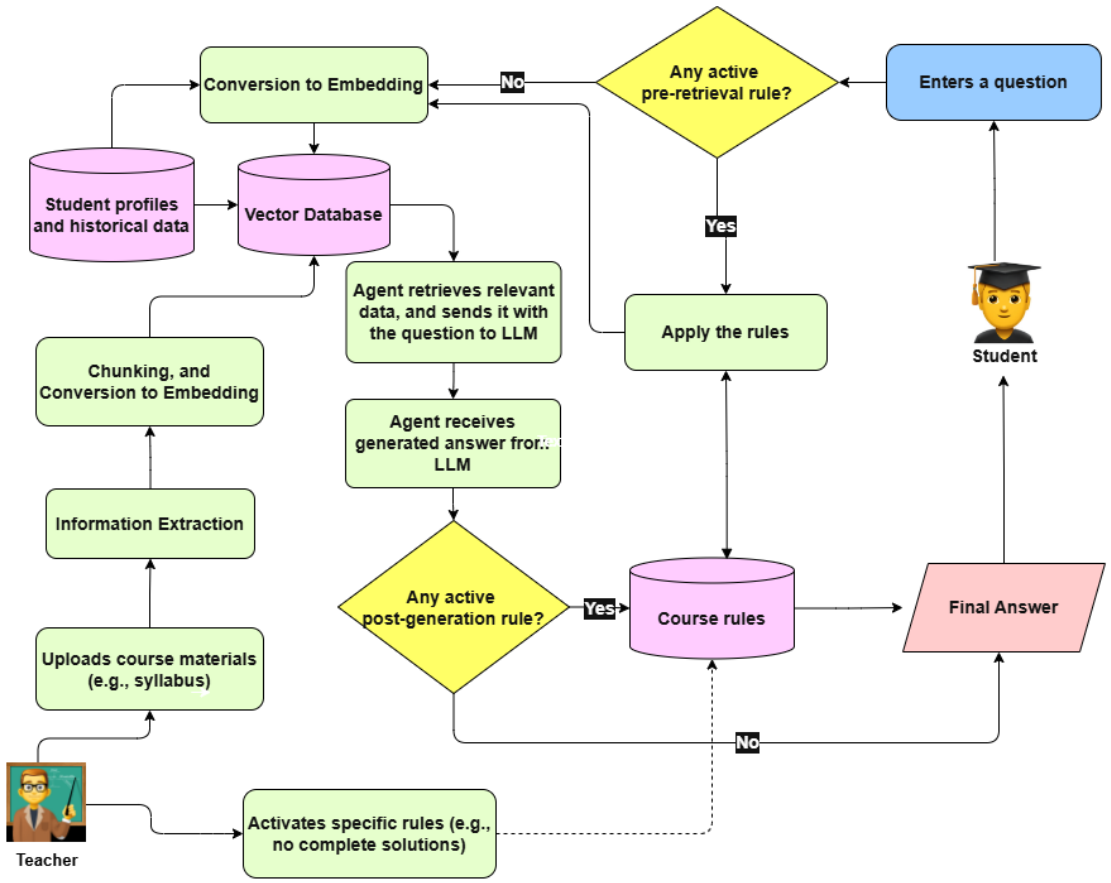


Figure 26: Pipeline of LLM-Powered Trustworthy Agents adapted from [161].

Figure 26 shows a high-level pipeline of MerryQuery. The system uses OpenAI's API for natural

language generation and LangChain to implement RAG. MongoDB stores interaction logs, document embedding, and course data, and React.js creates a reactive and easy-to-use front-end interface.

Summarising the main flow of the pipeline in the figure:

1. The teacher uploads course materials (syllabus, PDFs, images).
2. The system pre-processes the files (repairs, cleaning) and fragments (chunking) to create embeddings.
3. These embeddings are stored in a vector database (embeddings + metadata).
4. When a student asks a question, pre-retrieval rules are applied (e.g., blocking complete answers) and the RAG agent retrieves relevant extracts.
5. The question and excerpts are sent to the LLM (via OpenAI API — GPT-4o) using LangChain for RAG orchestration.
6. Post-generation rules are applied (e.g., filters, citation annotation) and the final result is delivered to the student.
7. Interaction records (chat, timestamp, statistics) are stored in MongoDB for teacher access and analysis.

How MerryQuery attempts to mitigate risks:

- Quotes linked to answers (RAG) allow for verification.
- Teacher controls over granularity reduce incentives for cheating (academic misconduct).
- Logs allow for auditing and detection of patterns of misuse.
- Remaining limitations and risks: If the source documents contain errors, the system may retrieve and justify incorrect information (RAG does not correct bad sources).
- OCR fails on complex images — may introduce noise.
- Privacy and compliance (storing conversations/embeddings) require clear policies, anonymisation, and consent.

MerryQuery is a clear example of a responsible approach to integrating LLM into an educational context. The solution combines RAG mechanisms, which enable responses to be based on factual information, with pedagogical controls that ensure human supervision, as well as a multimodal pipeline that preserves more complex content. This entire architecture is supported by a practical implementation, based on GPT-4, LangChain, MongoDB and React, which is already in the testing phase.

To ensure safe development, it is essential to continue investing in the quality of the sources used, in data privacy protection and in the empirical validation of the tool's educational impact.

3.2.4.3 Explainable LLM Chatbot for Learning Support

Chatbots offer the potential to engage students in conversation, similar to a discussion with a peer or mentor. However, despite advances in generative AI (GenAI) and LLM, chatbots are not yet capable of replacing human mentors.

The availability of an experienced peer or mentor is, however, one of the main challenges in online learning, especially with the vast amount of online learning resources available. Consequently, research on recommendation systems, explainability, and automated feedback systems has attracted growing interest in the field of Technology-Enhanced Learning (TEL) and decision support in education.

The article [3] proposes an LLM-based chatbot to explain learning recommendations to students, combining Knowledge Graphs (KG) to contextualise responses and mitigate risks such as hallucinations and deviations. The system aims to partially replace human mentors in online education environments, promoting conversational explanations (two-way dialogues) that help students understand and personalise their learning paths.

In this research, four strategies were adopted to achieve the goal of utilising the capabilities of LLM while avoiding their susceptibility to errors, biases, and hallucinations:

1. Task Limitation: the chatbot only responds to questions within a predefined scope (e.g., explanations about learning recommendations) to avoid irrelevant or incorrect responses to questions outside the educational context.

2. Dialogue with Intent Confirmation: the chatbot rephrases the user's question to confirm that it has understood correctly. If the user does not validate, the chatbot requests a reformulation or recommends human contact in case of repeated failure.

3. Contextualisation with KG: the LLM prompt is enriched with structured information extracted from the KG, including relationships between learning materials, pedagogical rules, and conversation history. In addition, human experts define standards for the responses, such as their expected length and the types of information permitted.

4. Integration with Human Mentors: group chat (student, mentor, and chatbot interact when the bot reaches its limits) and if the chatbot cannot respond, it redirects the student to a mentor.

The Figure 27 shows the architecture of this system.

The main components are:

- Chatbot: References dialogues, applies contextualisation strategies, and connects to LLM (GPT-4) and KG.
- Knowledge Graph: Provides data processed by experts (e.g., relationships between courses, thematic hierarchies).
- Human mentors: Enter the scene via group chat for complex or out-of-scope questions.

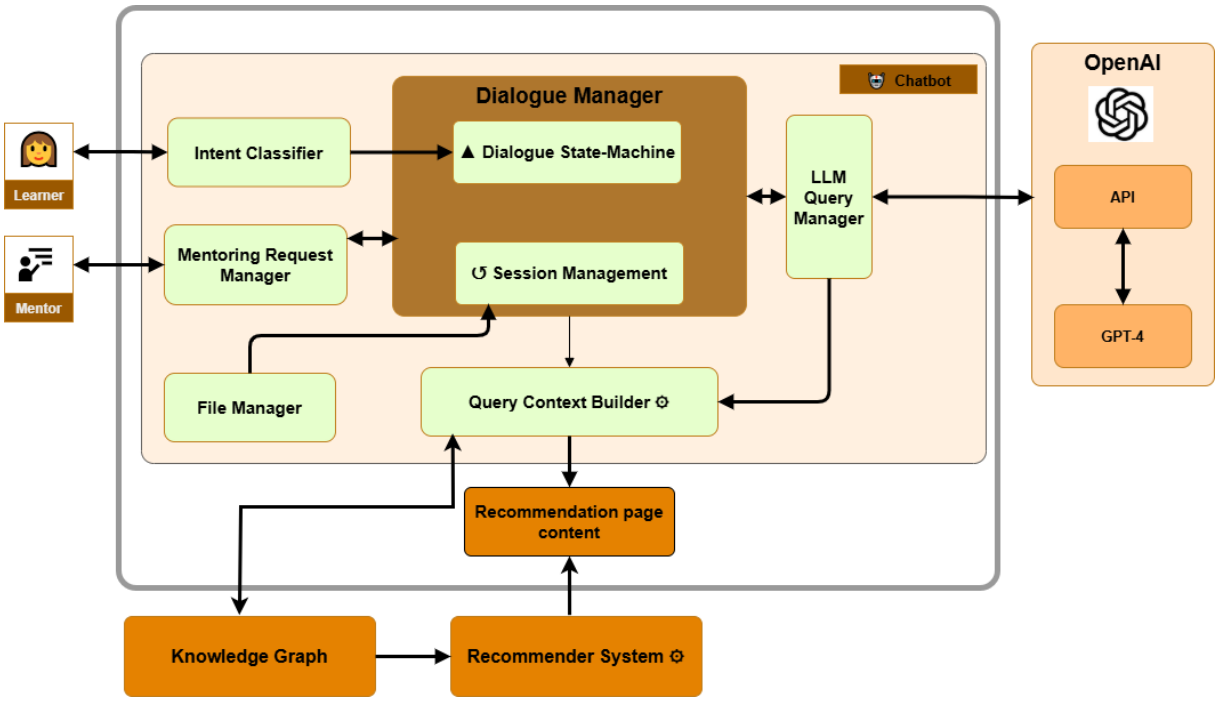


Figure 27: Chatbot architecture within the Angular web application and its connections between KG, recommendation algorithm, and the LLM system adapted from [3].

The chatbot is the central element of this system, functioning as an intelligent mediator between the user, the LLM and the KG database. Its core consists of a dialogue manager that orchestrates the flow of information and applies rigorous strategies to ensure accurate and contextualised responses.

To ensure that the chatbot operates within safe and relevant limits, an intent classifier has been implemented that defines a restricted set of permitted actions. These include explaining the reason behind a recommendation, detailing the content of a web page, describing the benefits of learning material, or clarifying relationships between content in the KG. If a user's question deviates from these predefined topics, the system redirects them to rephrase the question or, after three unsuccessful attempts, suggests contacting a human mentor.

Interaction is managed through a state machine that guides the dialogue. The workflow begins with the student asking a question. The intent classifier validates the category. When a question is ambiguous, the chatbot enters a confirmation state, asking the user to rephrase or clarify their question. If the misunderstanding persists, the system activates a resource state, connecting the student to a real mentor

through a group chat session. In this session, mentor and student can communicate directly, while the chatbot remains available to assist with contextualised responses, identified by the @ symbol.

This approach combines the efficiency of LLM with human curation of KG and direct supervision by mentors, creating a balance between intelligent automation and pedagogical safety. The result is a system capable of offering rich, personalised explanations while remaining within well-defined ethical and functional limits.

This approach to chatbots has proven to be a viable tool for conversational explanations in education, balancing LLM autonomy with human control (via KG and mentors). The approach offers a model for reducing the risks of LLM in critical environments, such as education, while maintaining the benefits of dynamic interaction.

3.2.4.4 LLMs for Feedback in K-12 Data Science Education

The article [54] presents DSRAG, an automatic feedback generation framework for teaching data science in K-12, combining cloud-based LLM with Retrieval-Augmented Generation (RAG) to provide personalised, understandable, and curriculum-aligned responses.

The issue is that data science is traditionally taught at advanced levels, and adapting it to K-12 faces challenges such as complex vocabulary, lack of robust assessment mechanisms for open-ended tasks and programming, and insufficient pedagogical alignment in LLM outputs.

In general, the contributions of this article can be summarised as follows:

- We propose a RAG-based prompting framework, called DSRAG, with cloud-based LLM to provide automated personalised feedback on primary and secondary school students' tasks.
- We built DSRAG with Langchain [30] to access an external database linked to the primary and secondary school data science curriculum.
- We evaluated the proposed framework through case studies on its effectiveness in meeting specific primary school needs.

The architecture comprises two main components:

1. **Data Handler —** prepares and organises educational materials (class transcripts, notes, source code, open-ended questions), segments them into chunks, generates embeddings, and stores them in vector stores for quick retrieval.
2. **Query Processor —** processes student questions, retrieves relevant information via RAG, and constructs prompts adapted to the cognitive level of K-12 to guide the LLM in generating step-by-step explanations.

It also integrates LangChain for data management and retrieval, uses OpenAI GPT-4 as a language model, and ChromaDB as a vector base.

The framework uses explicit prompt engineering to simplify language, uses everyday metaphors and examples, and explains code errors in an educational manner.

Comparisons showed that while 'pure' GPT-4 responded with technical terms ('runtime error,' 'dimensions and data type'), DSRAG translated this into simple analogies (e.g., comparing "y_test" to a 'test template'), maintaining technical accuracy but making the explanation accessible. DSRAG proved effective in providing scalable, personalised feedback in K-12, facilitating the understanding of complex concepts in data science.

Future improvements include testing other LLM and embedding models, applying fine-tuning to local models for greater accuracy, and supporting multimedia data (e.g., text-to-video) to enrich educational interaction.

3.2.4.5 Embodied AI Digital Teachers

The evolution of chatbots based on LLM is radically transforming the educational landscape, offering interactive and personalised responses that not only complement but have the potential to replace traditional teaching methods. It is in this innovative context that the authors of the article [191] have created MAGI, a pioneering system that combines cutting-edge technologies to create immersive and highly functional digital teachers, specially designed for educational environments in the metaverse.

At the heart of MAGI lies the synergistic integration of several advanced technologies. The system leverages the computational power of the Llama 3 8B model, trained on NVIDIA A5000 GPUs, combined with RAG techniques to ensure that responses are not only accurate but also contextually relevant. This combination allows the system to function as a digital tutor with deep knowledge and dynamic adaptability.

One of MAGI's most innovative features is its ability to create realistic 3D avatars that take the educational experience to new heights. Through advanced models such as wav2vec 2.0 and the implementation of acceleration loss, the system can produce extremely smooth facial animations that are perfectly synchronised with the generated audio. Real-time rendering, optimised through Pytorch3D, clearly outperforms other tools such as Pyrender and Blender Python in terms of speed and computational efficiency, ensuring a smooth and uninterrupted experience.

The MAGI interface has been carefully designed to be intuitive and functional, organised into three main sections: a chat input area, a digital teacher display area, and an avatar selection panel offering multiple customisation options. This simple yet powerful structure allows users to interact naturally with the system, focusing on learning without technological distractions.

The diagram in the Figure 28 illustrates the three-stage system that MAGI uses to create an interactive digital teacher, combining language processing, audio generation, and 3D animation:

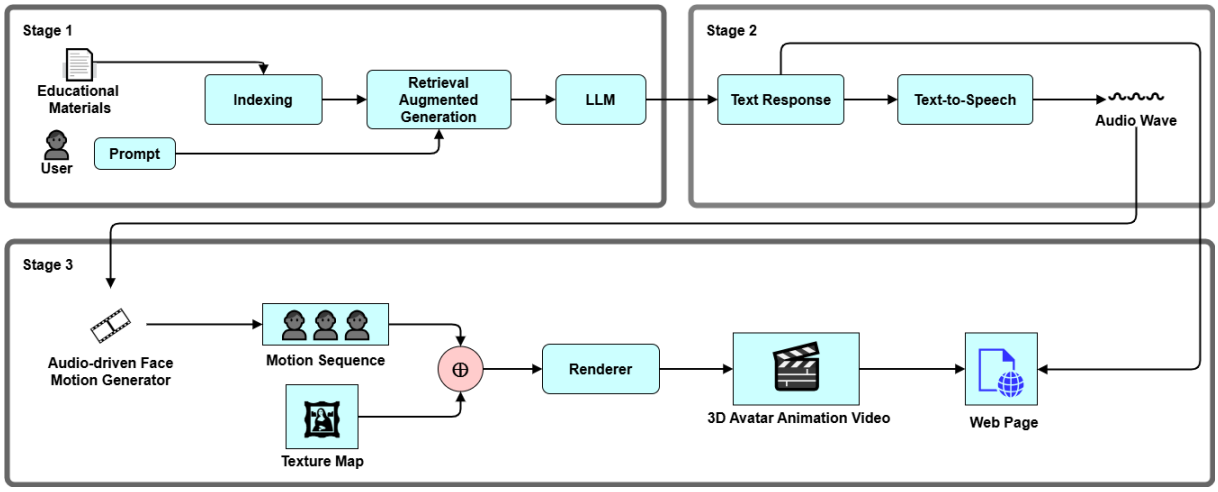


Figure 28: The MAGI system for creating an interactive teacher adapted from [191].

- **Stage 1 – Generating the Educational Response:**
  - Input: The user (student) submits a question.
  - Indexing: Educational materials are indexed for reference.
  - RAG: The LLM (e.g., Llama 3) generates an accurate response, enriched with data retrieved from the materials.
  - Output: A textual response is produced.
- **Stage 2 – Audio Conversion:**
  - The text response is converted into speech using Text-to-Speech (TTS), generating an audio waveform.
- **Stage 3 – 3D Avatar Animation:**
  - Facial Motion Generation: Audio is processed by a model (Audio-driven Face Motion Generator) to create realistic motion sequences (facial expressions, lip synchronisation).
  - Rendering: The Renderer, such as PyTorch3D, transforms the motion data into an animated video of the 3D avatar, applying texture maps to enhance visual detail and realism.
  - Final Output: The animated avatar is displayed on a web page, simulating a natural teacher-student interaction.

The educational advantages of MAGI are numerous and significant. Firstly, the system promotes greater accessibility to quality education by eliminating geographical and economic barriers that traditionally limit access to good tutors. Secondly, it offers an unprecedented level of personalisation, automatically adapting to each student's learning pace and style through dynamic and contextualised interactions. Finally, MAGI greatly increases student engagement thanks to the immersive and visually rich nature of its interface, which transforms learning into a captivating and memorable experience.

Looking ahead, the MAGI team plans to further expand the system's capabilities. Improvements to the user interface, the addition of multi-language support, and the implementation of various interaction modes to accommodate different learning preferences are planned. At the same time, special attention is being paid to the ethical and security issues that arise with the use of digital avatars in metaverse environments, ensuring that the system is not only technologically advanced but also responsible and safe for all users.

In short, MAGI represents a quantum leap in how we think about digital education. By intelligently combining the latest advances in AI, NLP, and 3D graphics, this system promises to transform education, making it more accessible, interactive, and effective for students around the world. The future of learning is being rewritten, and MAGI is at the forefront of this educational revolution.

3.3 Synthesis of the Literature Review

This chapter has helped to consolidate the methodological and technological framework underpinning the development of the proposed solution, highlighting the potential of approaches based on Artificial Intelligence, Machine Learning, LLM and RAG in an educational context. Various predictive models, strategies for personalising learning, and intelligent architectures designed to support decision-making and improve students' academic performance were analysed. In addition, the challenges associated with the use of these technologies were discussed, particularly in terms of interpretability, reliability of responses, data privacy and the need for more contextualised and pedagogically oriented systems.

On this basis, a conceptual overview of the developed solution was also presented, based on a specialised multi-agent approach capable of delivering interactions that are more precise, contextualised and tailored to students' needs. The integration of RAG mechanisms, local language models and explainability techniques aims to address the limitations identified in the literature, whilst promoting greater transparency, security and analytical robustness.

The decision to prioritise a RAG-based approach over fine-tuning strategies was driven by the specific constraints and requirements of educational environments. Whilst fine-tuning requires retraining the model using domain-specific datasets, a process that is computationally demanding, less transparent and more difficult to maintain over time, RAG allows for the direct integration of curriculum documents, teacher-generated materials and continuously updated teaching resources, without modifying the parameters of the underlying model. This feature is particularly relevant in educational contexts, where curriculum content frequently evolves and pedagogical control, response traceability and interpretability are essential requirements. Furthermore, the adoption of a modular architecture geared towards RAG has contributed to the development of a more flexible and sustainable solution, capable of adapting to real-world school contexts with limited computational resources and heterogeneous pedagogical needs.

In this context, the following chapter presents the proposed architecture for the developed system, describing in detail its functional components, the mechanisms for interaction between agents, the data

management and processing procedures, as well as the technological infrastructure underpinning the solution. The aim is to demonstrate how the designed architecture puts the theoretical and methodological principles discussed previously into practice, ensuring their applicability in real educational contexts.

4

Proposed Architecture

4.1 Framework

Based on the theoretical foundations established in the first part of the literature review, the architecture shown in Figure 29 was developed as a comprehensive Reference Architecture for Machine Learning-Based Intelligent Education. It defines the proposed end-to-end conceptual and operational structure for predicting students' academic performance and supporting data-driven pedagogical decision-making.

The architecture presented in this chapter should be understood as a Reference Architecture that is partially instantiated through the components implemented and evaluated during this PhD research. Not all architectural components represent implemented software components; some define architectural principles, supporting functions, or extensions envisaged for future operational deployment. The implemented components are explicitly identified throughout the chapter.

The proposed architecture integrates heterogeneous educational data, including demographic information, academic records, socioeconomic background, behavioural and attitudinal data, and digital learning resources, into a cohesive data-processing and modelling workflow. Data processing and early fusion techniques are used to aggregate heterogeneous information into a unified feature space, while class balancing methods are applied where required to support model training.

A feature selection phase is incorporated into the architecture to reduce redundancy and retain variables relevant to the prediction tasks. The modelling and validation pipeline includes the comparison of multiple machine learning algorithms, including SVM, XGBoost, Random Forest, AdaBoost, and Gradient Boost, together with hyperparameter optimisation and 5-fold cross-validation. These procedures provide a structured basis for evaluating predictive performance and model robustness across the experimental studies.

The implemented components of the architecture were empirically developed and validated across the five case studies and include data integration and processing, early fusion, feature selection, class balancing, model training and management, and model explainability. In particular, explainability mechanisms based on feature importance analysis and Shapley Additive exPlanations (SHAP) enable the interpretation of model decisions and the identification of relevant factors associated with academic performance.

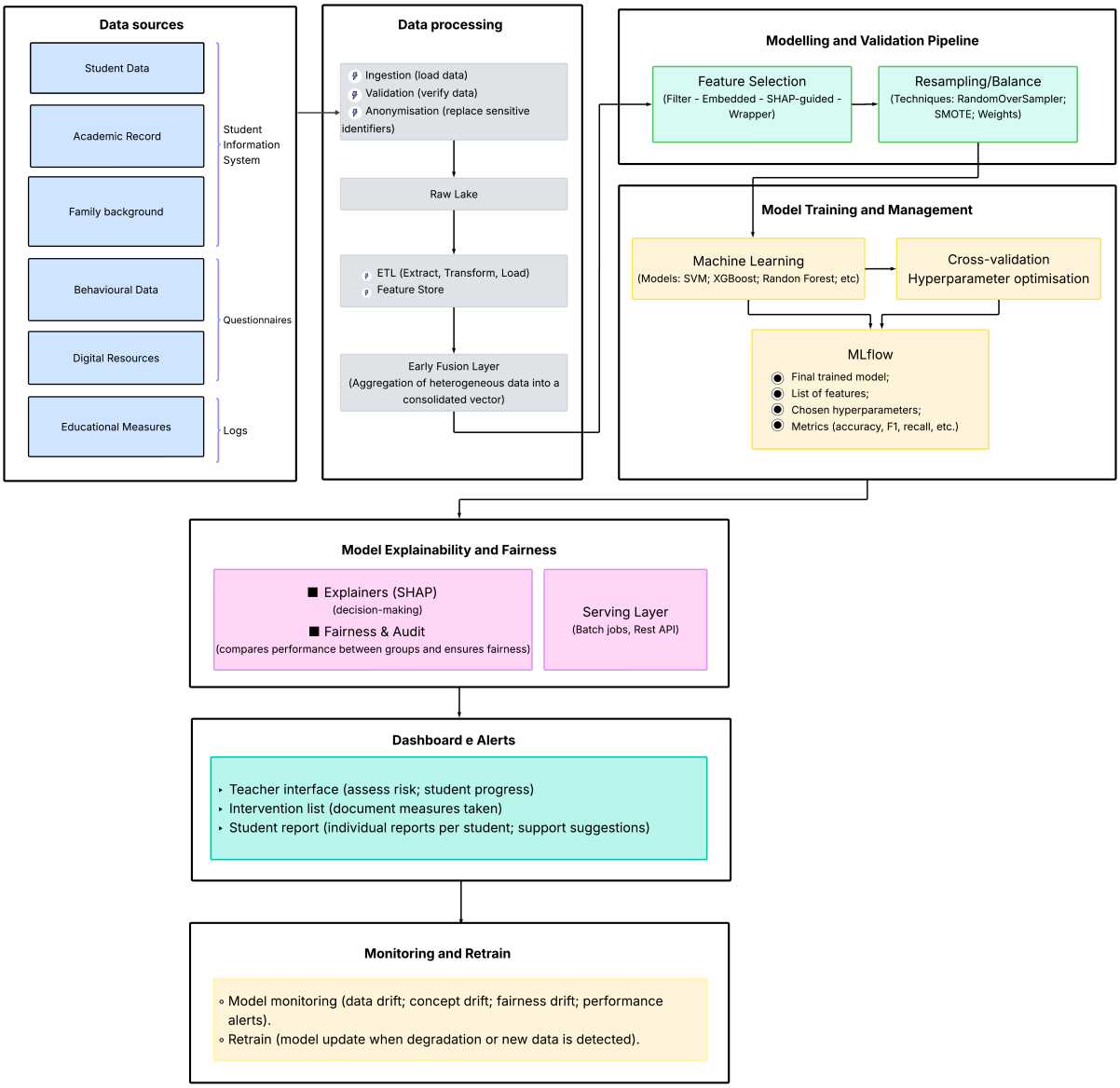


Figure 29: Machine Learning-Based Architecture for Intelligent Education.

The reference architecture also defines components that extend beyond the scope of the implemented empirical pipeline and are envisaged for future operational deployment. These include automated data-processing infrastructure, model serving and governance mechanisms, fairness and audit procedures, interactive dashboards and alerts, and continuous monitoring and retraining capabilities. These components are therefore presented as architectural provisions rather than as implemented and empirically validated functionalities.

The architecture is consequently intended not only to support predictive modelling, but also to provide a structured basis for transparency, explainability, auditability, and evidence-based decision-making in educational contexts. Its empirical support derives from the five case studies conducted during this PhD research, while the complete operationalisation of the reference architecture remains a subject for future development.

Following the definition of the proposed architecture, shown in Figure 29, a detailed description of its components is provided. This analysis clarifies the functional role of each module and the relationships between the implemented components and those envisaged as part of the complete reference architecture, covering the processes of data integration, processing, modelling, interpretation, and, where applicable, monitoring and deployment.

4.1.1 Data Sources

The Data Sources component represents the entry point of the proposed academic performance prediction architecture. Following the evidence identified in the literature, the architecture adopts a multimodal approach that combines institutional, behavioural, contextual, and digital information sources. The objective is to capture different dimensions of the student learning process, recognising that academic performance is influenced by a combination of personal, educational, family, and engagement-related factors.

As illustrated in Figure 29, the data sources are organised into three main groups: Student Information System (SIS), Questionnaires, and Logs. Each source contributes complementary information that can be integrated into a unified representation of the student profile.

4.1.1.1 Student Information System (SIS)

The SIS constitutes the primary institutional source of information and serves as the official repository of administrative and academic records. Data collected from the SIS are generally structured, reliable, and regularly updated, making them particularly suitable for predictive analytics applications in education. This component includes three subcategories:

1. **Student Data** include demographic and personal information such as age, gender, enrolment status, and other sociodemographic characteristics maintained by the institution. These variables provide contextual information about the student population and are commonly considered in educational data mining studies.

2. **Academic Record** comprises historical academic information, including previous grades, course enrolments, attendance records, progression history, completed subjects, and other indicators of academic trajectory. The literature consistently identifies prior academic performance as one of the most relevant sources of information for modelling future educational outcomes.

3. **Family Background** encompasses information related to the student's socio-family context, such as parental education, occupation, household characteristics, and other socioeconomic indicators when available. These variables provide additional context regarding the learning environment and support structures surrounding the student.

4.1.1.2 Questionnaires

Questionnaires provide a complementary source of information by capturing dimensions that are typically unavailable in institutional databases. These instruments allow the collection of self-reported information regarding students' attitudes, behaviours, perceptions, and learning conditions.

This component includes two subcategories:

1. **Behavioural Data** refer to information associated with study habits, learning strategies, self-regulation, motivation, classroom participation, persistence, and other indicators of student engagement. These variables are frequently highlighted in the literature as important factors influencing learning processes and academic achievement.

2. **Digital Resources** describe students' access to technological infrastructure that supports learning activities, including computers, internet connectivity, educational software, and other digital tools. The availability and use of these resources are increasingly considered relevant factors in modern educational environments.

4.1.1.3 Logs

The Logs component captures data generated through students' interactions with digital learning environments and educational platforms. Unlike questionnaire-based information, these records are automatically collected and provide objective observations of student activity.

Within the proposed architecture, this source is represented through the Educational Measures category, which includes information such as access to learning platforms, participation in online activities, assignment submissions, interaction with digital resources, and involvement in complementary educational initiatives. These records provide temporal and behavioural indicators that can support the identification of engagement patterns throughout the learning process.

By combining information from the SIS, questionnaires, and digital logs, the proposed architecture seeks to provide a more complete understanding of the factors that influence student performance. It incorporates information related to students' backgrounds, academic behaviours, levels of engagement, and learning context, recognising that academic success is shaped by multiple interconnected dimensions. This comprehensive view of the student serves as a strong foundation for the subsequent stages of data preparation, feature extraction, and predictive modelling.

4.1.2 Data Processing

The Data Processing component acts as the bridge between the different data sources and the predictive modelling stages. Its main purpose is to transform heterogeneous raw data into a structured, reliable, and consistent representation that can be used by machine learning algorithms and analytical processes.

As illustrated in Figure 29, this module receives information from multiple sources, including institutional records, questionnaires, and digital logs. Given the diversity of formats and data structures involved, a series of preprocessing operations is required to ensure data quality, integrity, and compliance with ethical and regulatory requirements.

The process begins with the **Ingestion** stage, where data is collected and integrated from the different sources available within the educational ecosystem. Since information may originate from databases, spreadsheets, learning platforms, or other digital systems, this stage is responsible for establishing a common entry point for all incoming data. Following ingestion, a **Validation** process is performed to assess data quality. This step includes verifying data completeness, consistency, and correctness, helping to identify potential issues such as missing values, duplicated records, or invalid entries before further processing takes place. To ensure privacy and compliance with data protection regulations, the data then undergoes an **Anonymisation** stage. During this process, personally identifiable information is removed or replaced by anonymous identifiers, allowing analytical and predictive tasks to be performed while protecting student confidentiality.

Once validated and anonymised, the data is stored in the **Raw Lake**. This repository preserves the original version of the collected information and provides a reliable foundation for future processing activities. Maintaining this intermediate storage layer supports data traceability, auditing, reproducibility, and the possibility of reprocessing data whenever required.

The next stage consists of the **ETL (Extract, Transform, Load)** process, which prepares the data for analytical use. During extraction, relevant information is retrieved from the raw repository. The transformation phase includes operations such as data cleaning, handling missing values, standardising formats, encoding categorical variables, normalising numerical attributes, and creating derived features when appropriate. Finally, the processed data is loaded into a structured environment suitable for subsequent analysis and modelling tasks. The resulting variables are managed through a **Feature Store**, a centralised repository that stores and organises the features generated during preprocessing. This component promotes consistency between training and deployment environments while supporting feature reuse, version control, and reproducibility across different modelling experiments.

To fully exploit the information available from multiple data sources, the architecture incorporates an **Early Fusion Layer**. This layer is responsible for integrating heterogeneous data into a unified representation before the modelling stage. By combining demographic, academic, behavioural, and digital information within a consolidated feature space, the system can leverage complementary perspectives of the student learning process while maintaining a coherent representation for subsequent analysis.

Overall, the Data Processing component provides a structured workflow that transforms raw educational data into a high-quality analytical dataset. By ensuring data quality, privacy protection, traceability, and multimodal integration, this module establishes the necessary foundation for feature selection, predictive modelling, and decision-support processes in later stages of the architecture.

4.1.3 Modelling and Validation Pipeline

The Modelling and Validation Pipeline is responsible for preparing the processed data for machine learning model development. Building upon the integrated dataset produced by the Early Fusion Layer, this component aims to improve data quality, reduce unnecessary complexity, and ensure that the information provided to the learning algorithms is both relevant and representative of the prediction problem.

As shown in Figure 29, this module is organised into two complementary stages: Feature Selection and Resampling/Balance. Together, these stages operate within a strict experimental protocol that guarantees methodological rigor, temporal isolation, and statistical validity.

The first stage, Feature Selection, focuses on identifying the variables that provide the most useful information for predicting academic outcomes. Since the integrated dataset may contain a large number of features originating from different data sources, some variables may be redundant, irrelevant, or provide limited predictive value. Feature selection therefore helps reduce the dimensionality of the dataset while retaining the information most relevant to the learning task. In the proposed architecture, this process may incorporate statistical, model-based, and explainability-oriented approaches to support the identification of meaningful features and improve model interpretability.

The second stage, Resampling/Balance, addresses the challenge of class imbalance, which is frequently encountered in educational prediction tasks. In many scenarios, students belonging to specific categories of interest, such as those experiencing a decrease in academic performance, may represent a smaller proportion of the available data. Without appropriate balancing strategies, machine learning models may become biased towards the majority class and fail to adequately recognise underrepresented groups. This component therefore incorporates mechanisms to create a more balanced training dataset, enabling models to learn more effectively from all classes represented in the data.

Experimental Protocol and Pipeline Execution Order

To preserve the statistical integrity of the predictive models and prevent data leakage, the modelling and validation pipeline follows a strict sequential execution order. The integrated dataset is first partitioned into training and testing sets, using an 80/20 hold-out strategy and a 5-fold cross-validation scheme within the training data. This initial partition ensures that the final testing data remain completely unseen throughout model development.

Following data partitioning, Feature Selection is performed exclusively within the training data of each cross-validation fold. Procedures such as correlation-based filtering and feature-importance scoring are therefore computed without access to the corresponding validation or testing data. This prevents information from the evaluation partitions from influencing the selection of predictors.

The third step consists of Class Resampling/Balance. When required, techniques such as RandomOverSampler or SMOTE are applied only to the training portion of each cross-validation fold. The validation and testing partitions remain untouched and preserve their original class distributions. This procedure

ensures that model performance is evaluated under conditions that more closely reflect the real distribution of students and avoids artificially inflating predictive performance through information leakage.

Finally, the selected and balanced training data are used to train the machine learning algorithms, such as Support Vector Machine (SVM), XGBoost, and Random Forest (RF). The resulting models are subsequently evaluated on the corresponding unseen validation folds and, ultimately, on the independent test set. Consequently, all data-dependent preprocessing operations are contained within the training boundaries of each fold, while the evaluation data remain isolated until prediction. This execution order provides a reliable estimate of model generalisation performance and ensures that the reported results reflect the actual predictive capability of the models rather than information inadvertently derived from the evaluation data.

The two main stages of the pipeline therefore operate sequentially within a controlled experimental framework: Feature Selection first refines the information space by retaining the most relevant variables, while Resampling/Balance addresses class distribution within the training data. By combining these procedures with strict partitioning and temporal and statistical isolation, the pipeline establishes the conditions necessary for developing predictive models that can capture meaningful patterns in educational data while maintaining robustness, interpretability, and methodological validity.

4.1.4 Model Training and Management

The Model Training and Management component is responsible for transforming the prepared dataset into predictive models capable of supporting educational analytics and decision-making. Building upon the outputs generated by the Modelling and Validation Pipeline, this component encompasses the processes required to train, evaluate, optimise, and manage machine learning models throughout their lifecycle.

As illustrated in Figure 29, this module is organised into three complementary stages: Machine Learning, Cross-Validation and Hyperparameter Optimisation, and MLflow. Together, these stages ensure that predictive models are developed using rigorous methodological practices while remaining traceable, reproducible, and suitable for future deployment.

The first stage, **Machine Learning**, focuses on the training of predictive models using the features generated and refined in the previous stages of the architecture. Different machine learning algorithms may be explored depending on the characteristics of the educational data and the objectives of the prediction task. The purpose of this stage is to identify patterns and relationships within the data that can support the prediction of academic outcomes and other educational indicators.

The second stage, **Cross-Validation and Hyperparameter Optimisation**, is responsible for assessing and improving model performance. Cross-validation provides a systematic approach for evaluating how well a model generalises to unseen data, reducing the risk of overfitting and providing a more reliable estimate of predictive performance. In parallel, hyperparameter optimisation seeks to identify the most appropriate model configurations, contributing to the development of more accurate and robust predictive solutions.

The final stage, **MLflow**, acts as the experiment tracking and model management layer of the architecture. Its role is to document and organise the artefacts generated during model development, including trained models, feature configurations, hyperparameter settings, and evaluation results. This centralised repository supports reproducibility, facilitates model comparison, and provides a structured foundation for the governance and maintenance of predictive models over time.

Together, these three stages establish a structured framework for developing and managing machine learning models within the proposed architecture. By combining model training, rigorous validation procedures, and systematic lifecycle management, this component provides a reliable foundation for the deployment and continuous improvement of educational predictive analytics solutions.

4.1.5 Model Explainability and Fairness

The Model Explainability and Fairness component aims to ensure that the predictions generated by the system are not only accurate, but also transparent, understandable, and aligned with ethical principles. In educational contexts, where predictive models may support decisions that affect students' learning pathways, understanding how predictions are produced is particularly important.

As illustrated in Figure 29, this component is organised into three complementary elements: SHAP Explainers, Fairness & Audit, and the Serving Layer.

The **SHAP Explainers** provide insights into the factors that contribute to model predictions. By identifying the relative importance of different features, this component helps educators and decision-makers better understand the reasoning behind model outputs. These explanations can support the interpretation of predictions and contribute to more informed educational interventions.

The **Fairness & Audit** component focuses on assessing whether the predictive models behave consistently across different groups of students. Its purpose is to support the identification of potential biases and to ensure that the system operates in a fair and responsible manner. By monitoring model behaviour and performance across different contexts, this component contributes to the development of trustworthy educational analytics solutions.

Finally, the **Serving Layer** acts as the operational interface between the explainability and fairness mechanisms and the end users of the system. It enables the delivery of model predictions, explanations, and audit information through services that can be integrated into educational platforms and decision-support tools.

Together, these elements promote transparency, accountability, and trust in the proposed architecture, ensuring that predictive models can support educational decision-making in a responsible and interpretable manner.

4.1.6 Dashboard and Alerts

The Dashboard and Alerts component represents the interaction layer between the predictive system and its end users. Its primary purpose is to transform model outputs into accessible and actionable information that can support educational decision-making and student monitoring.

As illustrated in Figure 29, this component is organised into three complementary elements: the Teacher Interface, the Intervention Checklist, and the Student Report. Together, these elements provide different perspectives on student performance and facilitate the use of predictive insights within educational settings.

The **Teacher Interface** serves as the main access point for educators, presenting predictions, risk indicators, and relevant student information through an intuitive and user-friendly environment. By providing a consolidated view of student performance and engagement, this interface supports the identification of students who may require additional attention and enables continuous monitoring of their academic progress.

The **Intervention Checklist** supports the documentation and follow-up of educational support measures. This component allows interventions to be recorded and monitored over time, creating a structured framework for tracking the actions taken to support students and evaluating their progression.

The **Student Report** provides an individualised view of each student by combining predictive results, relevant indicators, and explanatory information generated by the system. These reports are designed to support communication among different educational stakeholders and facilitate a better understanding of the factors associated with student performance.

Complementing these functionalities, the alert mechanism supports the timely identification of relevant changes in student status. Overall, the Dashboard and Alerts component transforms predictive outputs into actionable information, supporting student monitoring and informed educational decision-making.

4.1.7 Monitoring and Retrain

The Monitoring and Retraining component is responsible for ensuring that the predictive system remains reliable and effective over time. Since educational environments are constantly evolving, changes in student populations, learning practices, institutional policies, or external conditions may affect the performance of predictive models. This component addresses these challenges by continuously monitoring the system and supporting model updates when necessary.

As illustrated in Figure 29, this module is organised into three complementary elements: Monitoring, Alerts, and Retraining.

The **Monitoring** component continuously evaluates the behaviour and performance of the predictive models after deployment. Its purpose is to identify changes in the data or in model performance that may reduce the reliability of predictions over time. In addition to performance monitoring, this process

can also support the continuous assessment of fairness and model consistency across different student groups.

The **Alerts** component provides a mechanism for notifying administrators or relevant stakeholders when predefined conditions are detected. These notifications may be triggered by significant changes in model performance, data characteristics, or other indicators that require attention. By providing timely feedback, the system supports proactive maintenance and oversight.

The **Retraining** component is responsible for updating predictive models when monitoring activities indicate that performance has degraded or that the educational context has significantly changed. By incorporating more recent data into the modelling process, retraining helps ensure that the system remains aligned with current educational realities and continues to provide reliable predictions.

Together, these elements establish a continuous improvement cycle that supports the long-term sustainability of the proposed architecture. Through ongoing monitoring, timely alerts, and periodic model updates, the system can adapt to changing educational contexts while maintaining the quality and trustworthiness of its predictions.

4.2 Ambient Intelligent Learning Architecture - Mathematics Tutor

In an ever-evolving educational context, where personalisation and precision are increasingly important, the design of intelligent learning environments offers new opportunities to support students throughout their learning journey. In this context, an Ambient Intelligent Learning Architecture for Mathematics is proposed, combining artificial intelligence, multi-agent systems, and retrieval-augmented generation techniques to create a personalised and adaptive learning experience.

This proposal focuses on the development of educational web application designed specifically to support mathematics learning. The architecture is built upon a multi-agent system in which intelligent agents simulate dynamic pedagogical interactions with students. These agents are designed to provide guidances, explanations, feedback, and learning support tailored to different mathematical domains, promoting a more interactive and engaging educational experience.

This intelligent ecosystem includes specialised agents dedicated to **Functions**, **Algebra**, and **Geometry**, allowing the system to provide domain-specific support while maintaining a coherent and integrated learning experience.

Through a modular and scalable architecture, the web application allows for fluid and intuitive interaction, where students and educators benefit from pedagogically sound, contextualised and evidence-based answers, retrieved in real time through RAG. Each step of the process, from submitting a question to delivering a synthesised and auditable response tailored to the user's profile, is designed to promote clarity, confidence, and engagement.

With features such as step-by-step explanations, confidence scores, citable sources, and real-time feedback mechanisms, this application not only automates teaching but also elevates it, combining the best of natural language technology with robust educational principles. All this while ensuring transparency, privacy, and continuous improvement through quality metrics and detailed audit logs.

This is our starting point for designing this web application and reimagining the future of AI-assisted education: a web application that does not merely respond, but understands; that does not simply inform, but actively teaches.

To achieve this vision, the proposed solution is centred on a specialised multi-agent approach, where students interact through dialogue with the agent most appropriate to their specific learning context or subject area. Each agent acts as a domain specialist, enabling more accurate, contextualised, and pedagogically oriented interactions, while significantly reducing the occurrence of hallucinations within its area of expertise. The overall architecture supporting this approach is presented in Figure 30.

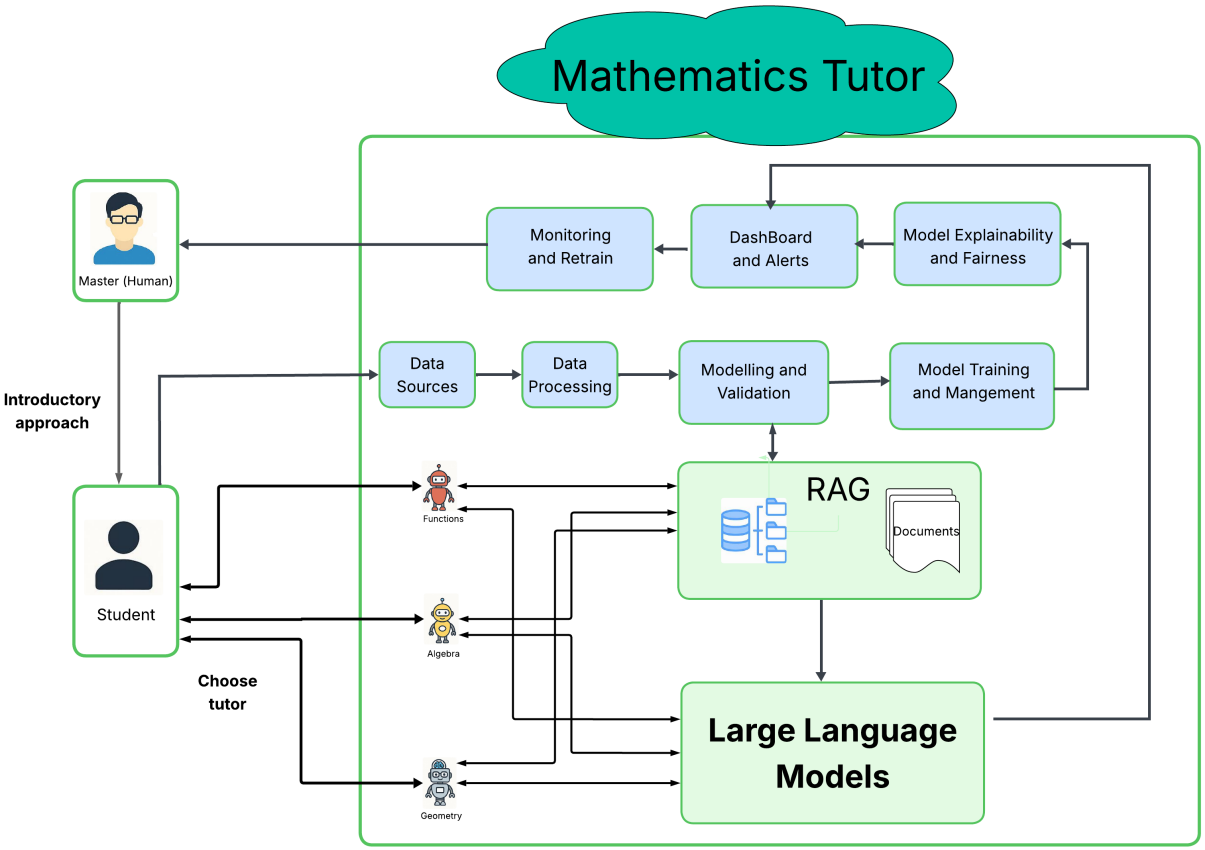


Figure 30: Architecture of the proposed ambient intelligent.

To implement this architecture, we began by defining the system's functional structure, which consists of a front-end developed in Streamlit, a LangChain-based back-end with RAG agents, a local language model running via Ollama, and a Chroma vector database. Below is a detailed analysis of how the agents work, including the interaction and orchestration mechanisms, memory management processes, and the procedures adopted to ensure privacy, traceability and auditability.

4.2.1 Architecture Components and System Agents

The Mathematics Tutor constitutes one of the main implemented instantiations of the proposed Reference Architecture. Its current prototype implementation is organised into three main layers: User Interface, Application Logic, and Knowledge Management. These layers correspond to the functional components implemented and evaluated in the prototype, while other components of the Reference Architecture remain at the conceptual or future-deployment level. The three layers comprise a set of compatible components that operate in a straightforward and accessible manner and that, together, support a pedagogical experience geared towards autonomous study.

The first layer is the User Interface, implemented using Streamlit, which provides the environment through which students interact with the tutor. The interface is organised into different tabs, allowing students to navigate between structured educational content and conversational interaction mechanisms. The frontend also manages the session state, ensuring the persistence of user choices throughout navigation, and integrates the selection of the thematic context, namely Algebra, Functions, or Geometry.

In chat mode, an explicit limit of five interactions per module is applied at the user interface level. This rule is intended to maintain pedagogical focus, encourage autonomous exploration of the structured materials, and prevent conversational drift. Each time a student selects a new study module, the session counter is reset and the conversation history is cleared. This mechanism ensures that the interaction remains focused on the selected mathematical domain.

The Application Logic layer incorporates specialised RAG agents, namely AgentAlgebra, AgentFunctions, and AgentGeometry, each responsible for managing a specific mathematical domain. This layer is responsible for determining which agent should be invoked for each interaction and supervises communication between the user interface, the context-specific RAG agents, the language models, and any symbolic or numerical verification modules, such as an optional Computer Algebra System (CAS). This mechanism ensures that each request is consistently routed according to the system's management rules, avoiding decontextualised requests to the language model.

The context-based RAG agents form the core of the tutor's disciplinary expertise. Each subject area, Algebra, Functions, or Geometry, has an independent pipeline, respectively AgentRAG_Algebra, AgentRAG_Functions, and AgentRAG_Geometry. These agents operate independently and maintain separate conversation memories. Each agent integrates a ChromaDB-based retriever and uses the embedding model embeddinggemma to convert both students' questions and relevant excerpts from the instructional manuals into vector representations. These representations enable the retrieval of the most relevant contextual information from the corresponding knowledge base.

The retrieved excerpts are subsequently provided as contextual information to the language generation component. The system uses the `gemma3n` model to generate the pedagogical responses, with the model executed entirely locally through the Ollama platform. The RAG agents use prompts structured according to grounding principles, requiring the generated responses to rely on the retrieved excerpts. This approach helps minimise the risk of hallucination and supports conceptual and curricular consistency.

The language processing and generation components therefore perform complementary functions within the architecture. AMALIA is focused on processing the Portuguese language and supports the linguistic component of the interaction, while `gemma3n` is responsible for generating the educational responses based on the contextual information retrieved by the RAG pipeline. The system consequently combines Portuguese-language processing, local retrieval and vectorisation mechanisms, and local language generation to provide context-aware educational support in Portuguese. The use of local models through Ollama also reduces dependence on external cloud-based AI services.

The vector knowledge base is implemented using ChromaDB and is persistent and separated by context. Each directory contains documents in `.txt` format that constitute the corpus supporting the RAG process. The `embeddinggemma` model is used to transform the instructional content and user queries into vector representations, which are then used to identify semantically relevant information. The retrieval process uses (k=3) as the default value, balancing the amount of contextual evidence provided to the language model with response conciseness. Whenever the indexes need to be rebuilt, the system automatically processes the documents associated with each context.

An additional mechanism relevant to interaction stability is ChatMemory, which maintains session memory for each agent. This memory stores conversation turns and associated metadata, allowing the agent to preserve discursive coherence within the defined interaction limit. When the student changes modules, the existing memory is cleared and replaced by a new instance, preventing topic contamination and maintaining the focus of the new interaction context.

The system also incorporates several auxiliary subsystems, including document loaders, index-building modules, a reconstruction process triggered at application startup, optional tools for symbolic or numerical verification, and logging and auditing mechanisms to support future usage and performance analysis. These components provide the supporting infrastructure required for maintaining the knowledge bases and monitoring the operation of the tutor.

Finally, the entire architecture is designed to operate in a local or edge execution environment, prioritising privacy, institutional control, and reduced dependence on external services. The `gemma3n` language model and the `embeddinggemma` embedding model are executed locally through Ollama, while ChromaDB provides the persistent vector knowledge base. This local configuration reduces the need to transmit educational interactions and instructional content to third-party AI services and therefore supports a Privacy-by-Design approach. Although local execution does not, by itself, guarantee GDPR compliance, it facilitates the implementation of relevant data-protection principles within the system architecture.

4.2.2 Step-by-Step Operational Flow

The following sequence describes the operational flow implemented in the current Mathematics Tutor prototype:

(1) the student initially selects one of the three available mathematical contexts;

(2) within each context, they choose a specific study module;

(3) the system then presents five interaction options - “What you need to know”, “Step by step”, “Solved example”, “Practise” and “Chat”;

(4) for the chat option, the system activates the corresponding RAG agent, which retrieves relevant contextual information from the vector database before generating a response;

(5) each chat session is limited to five interactions to maintain pedagogical focus. This flow ensures consistent pedagogical progression while allowing flexibility in content exploration. The five-interaction constraint is enforced at the Streamlit user-interface level and does not constitute an intrinsic limitation of the underlying LLM or RAG pipeline.

4.2.3 Formalisation of the Roles and Responsibilities of Agents

The components described in this section correspond to the implemented software components of the prototype. The formalisation below describes the architecture and operational mechanisms of these implemented components, including the context-specific RAG agents, the retrieval service, the vector database, and the locally deployed language model.

In the implemented prototype, agents are the components responsible for processing dialogic interactions within each thematic context available in the Mathematics Tutor.

For each context C_i ($C_i \in \{Algebra, Functions, Geometry\}$) there is a specialised agent, indicated by A_i, which operates as a specialised unit for retrieval, contextualisation and generation of pedagogical responses. Formally, each agent A_i implements a pipeline function that receives a query from the user q and the conversational state S — represented by the ChatMemory instance — and produces a textual response R:

$$A_i(\text{q,S}) \longrightarrow \text{R}$$

Each agent A_i is instantiated from the base class AgentRAGContext[i], which inherits from AgentBase and specialises generic behaviour through the implementation of the abstract method _get_base_runnable(). This inheritance-oriented architecture ensures structural consistency between agents from different contexts, facilitating system maintenance and extensibility.

The RAG agent associated with the context performs a well-defined set of operational responsibilities. First, through ServiceRAG, the agent indexes educational documents in a vector space managed by the Chroma library, using embeddings generated by the embeddinggemma:latest model. During the retrieval phase, the retriever selects the k=3 documents that are semantically closest to the user’s query, applying a cosine similarity search in the embedding space. The retrieved documents are then processed by the static function _format_docs(), which concatenates the textual content (page_content) into a unified string, devoid of structural metadata but preserving the semantic integrity of the educational material. This simplified formatting is a design decision that prioritises contextual clarity over granular traceability at the LLM interface level.

Based on the retrieved and formatted material, the agent dynamically constructs a contextualized prompt using a ChatPromptTemplate that structures three essential components: (i) system message, which defines the agent's pedagogical role ("10th grade math tutor"), specifies the domain of expertise (e.g., "Numbers, Operations, and Algebra") and establishes critical behavioral constraints, namely, the strict and exclusive use of the retrieved context, the prohibition of inventing information, and the obligation to declare epistemic limitations when the retrieved material is insufficient; (ii) conversational history, injected by the historical_chat parameter, allowing the model to access the conversational history available within the current session for resolving anaphoric references and maintaining thematic coherence; (iii) the student's current question, formatted as a human message.

The language model, in this implementation, gemma3n:latest served locally by the Ollama engine through ServiceLLM, assumes the role of an educational transformer, producing explanations in natural language that are pedagogically appropriate for the age level of the recipients. Text generation is governed by guardrail instructions encoded in the system message, which determine: (i) transparency in explanations, prioritizing clarity over brevity; (ii) logical structuring of reasoning, making inferential transitions explicit; (iii) use of mathematically rigorous but accessible terminology; (iv) absence of extrapolations not based on the documentary context provided. The LLM does not have direct access to the metadata of the retrieved documents (identifiers, titles, similarity scores), operating exclusively on the concatenated text produced by _format_docs().

Interaction coordination is ensured by the interface layer implemented in Streamlit (app.py), which orchestrates the control flow between module selection, static content presentation, and conversational agent invocation. This layer determines when to initialize a new agent (by selecting a different context), when to clear the conversation history (by selecting a new module within the same context via handle_module_selection()), and when to invoke the agent's run() method to process a new query. In the current implementation, there is no centralized "Controller" entity that mediates complex routing decisions or invokes external verification modules. Instead, control logic is distributed between Streamlit's event handlers and the methods of the individual agents.

4.2.4 Limitations of the current implementation

The current prototype does not implement a formal verification component (Verifier) based on computer algebra systems (CAS), such as SymPy, or numerical sampling mechanisms. Consequently, mathematical correctness is not formally guaranteed and depends on several factors, including: (i) the quality of the educational documents indexed in the RAG database; (ii) the mathematical capabilities of the locally deployed `gemma3n` model; and (iii) the prompt constraints that restrict generation to the retrieved context. These constraints are intended to mitigate hallucinations but cannot eliminate logical errors or incorrect symbolic manipulations that may originate from the source documents or result from incorrect inferences by the LLM. The absence of formal verification is therefore a recognised limitation of the current prototype.

Addressing this limitation would require the integration of CAS-based libraries capable of parsing mathematical expressions, performing symbolic or numerical verification, and comparing generated solutions with mathematically valid reference results.

The frontend developed in Streamlit provides the interface between the student and the digital tutor through three main components: (i) a navigation sidebar, allowing selection between the three thematic contexts; (ii) a structured content area, featuring four navigable tabs (conceptual fundamentals, algorithmic procedures, solved examples, and practice exercises), whose static content is defined in `APP_STRUCTURE["hook_questions"]`; and (iii) a chat interface, positioned below the content tabs, displaying the conversation history with visual differentiation between user messages and assistant responses. The interface does not display metadata about the retrieved evidence, such as document titles or similarity scores, nor formal verification status, as these functionalities are not implemented in the current prototype.

The five-interaction constraint per module is operationalised directly at the Streamlit user-interface level through session-state conditionals (`st.session_state.turn_counts`), rather than within the underlying agent pipeline. This design choice separates UI-level session management from the core LLM execution chain, allowing the interface to enforce the pedagogical interaction constraint while keeping the agent backend modular. The constraint is therefore a prototype-level design decision and does not constitute an intrinsic limitation of the underlying RAG or LLM components.

4.2.5 Interaction between components

The interaction between the implemented components follows a deterministic sequential protocol coordinated by the Streamlit interface and the context-specific RAG agents. When the student selects a study module, the Streamlit interface invokes `handle_module_selection()`, which clears the agent's conversation history through `agent.clear_history()` and resets the relevant session state. The first interaction is automatically generated by the system using a predefined prompt ("Explain the topic '[module_name]' to me in a simple and concise way, focusing on the main concepts"), establishing a pedagogical entry point for the dialogue. Subsequent interactions follow the sequence: (i) capture of user input through `st.chat_input()`; (ii) addition of the user message to the conversation history through `memory.add_user_message()`; (iii) invocation of `agent.run(user_input)`, which triggers the retrieve → format → prompt → generate pipeline; (iv) addition of the assistant's response to the conversation history through `memory.add_ai_message()`; and (v) re-rendering of the interface to display the new message.

Within the `run()` method, the operational sequence is implemented as follows:

1. Preparation of chain input: Construction of a dictionary chain_input = { "question": user_input, "chat_history": self.memory.get_history() }, which aggregates the current query and the complete dialog context.

2. Document retrieval: The retriever component (obtained via rag_service.get_retriever()) performs a vector search in the Chroma store, returning a list of Document objects sorted by decreasing similarity, truncated to the top k=3 ranked documents.

3. Context formatting: The RunnableParallel defined in _get_base_runnable() processes three branches in parallel: (a) itemgetter("question") | retriever | _format_docs — pipeline that extracts the question, retrieves documents, and formats them; (b) itemgetter("question") — passthrough of the original question; (c) itemgetter("chat_history") — passthrough of the history. The outputs are aggregated in a dictionary that feeds the prompt template.

4. Prompt construction: ChatPromptTemplate concretely instantiates the system message (injecting the formatted context via context), the conversation history, and the human question, producing a message structure compatible with the LLM API.

5. Response generation: The LLM processes the message sequence and generates token-by-token text through sampling (temperature and top-p configured by default in the Ollama model). The output is a natural language text string.

6. Response parsing: StrOutputParser() extracts the textual response from the LLM output structure, ensuring compatibility with the format expected by the interface. Memory update: The generated response is added to the conversation history, completing the interaction cycle.

When the retriever does not identify relevant documents, the `_format_docs()` function returns the string "No relevant context was found for your question", which is subsequently injected into the `context` field of the prompt. This condition activates the fallback behaviour defined in the system message, instructing the LLM to explicitly acknowledge the absence of supporting material. The resulting response follows the predefined fallback message: "I did not find information on this topic in my documents, but I can try to help based on what we have already discussed." This mechanism preserves conversational continuity while making explicit that the requested information was not found in the indexed documentary context.

Each specialised RAG agent assumes distinct thematic responsibilities within the Mathematics Tutor. `AgentRAGContext1`, instantiated for "Numbers, Operations, and Algebra", manages queries related to solving linear and quadratic equations, manipulating inequalities, applying exponent laws, solving 2×2 linear systems, and determining parameters for equation equivalence. `AgentRAGContext2`, associated with "Functions", processes questions concerning related functions, calculation of images and pre-images, determination of zeros and intersections with axes, calculation of slopes and ordinates at the origin, and modelling of functional relationships. `AgentRAGContext3`, dedicated to "Geometry", addresses calculations of areas and volumes of solids, application of triangle similarity, recognition of vector equalities, and metric relations in geometric figures.

All agents share a common architecture derived from the abstract class AgentBase, which encapsulates: (i) reference to ServiceLLM, which provides access to the gemma3n:latest model; (ii) reference to ServiceRAG, which provides the retriever configured with k=3 and the persistent Chroma vector store; (iii) ChatMemory instance for managing conversation history; (iv) _compile_chain() method that builds the executable_chain by invoking _get_base_runnable(); (v) run() method that orchestrates the complete query processing flow. This structural uniformity facilitates code maintenance, allows reuse of common logic, and simplifies the addition of new thematic contexts by specializing the base class and defining contextualized prompts and appropriate document bases.

The separation of responsibilities between ServiceLLM, ServiceEmbeddings, ServiceRAG, and the agents themselves reflects principles of component-oriented software engineering, namely separation of concerns, state encapsulation, and dependency inversion. The services encapsulate initialization logic, connection management, and validation of external resource availability (Ollama server), exposing uniform interfaces (get_llm(), get_embeddings(), get_retriever()) and readiness verification methods (is_ready()). Agents depend on these abstractions, not on concrete implementations, allowing for backend replacement (e.g., transition from Ollama to OpenAI or Anthropic API) through localized changes in service classes, without requiring modifications to agent logic.

4.2.6 Presentation of answers to the student

The system's responses are structured according to a standardised model that integrates fundamental principles of instructional design designed to support clarity of presentation, methodological transparency, and pedagogical continuity, adapted to the cognitive needs of secondary school students.

The presentation architecture is divided into two complementary levels: static structured content and dynamic conversational assistance. The content of each thematic module is organised into four components that can be navigated through a tabbed interface: (1) 'What you need to know' summarises the conceptual foundations of the topic, presenting definitions, properties, and relationships between concepts through text formatted in Markdown with mathematical notation; (2) 'Step by Step' operationalises resolution procedures by breaking them down into numbered sequential phases, explaining the logic underlying each algorithmic transition; (3) 'Solved Example' instantiates the application of concepts through a standardised format that integrates contextualised statements, explanation of the resolution strategy, detailed development with intermediate justifications, and validation of the result obtained; (4) 'Practise' provides a set of predefined exercises with increasing levels of complexity, whose complete solutions remain hidden by default and are revealed through explicit user interaction via an expandable component (expander), thus promoting independent attempts before consulting the model solution.

Conversational interactions within the chat interface allow students to clarify doubts in natural language. In the current prototype, each module is subject to a five-interaction constraint enforced by the Streamlit frontend, as described in Sections 4.2.2 and 4.2.4. When a new module or thematic context is selected, the system invokes `clear_history()`, resetting the conversation state and ensuring that the

new topic is addressed with a fresh conversational context.

To support the evidential grounding of the information presented, the system implements a RAG pipeline that anchors the tutor's responses to a documentary knowledge base specific to Context 1 (Numbers, Operations, and Algebra). When a student submits a question through the chat interface, the processing flow comprises four main stages: (i) the query is converted into a vector representation using the `embeddinggemma:latest` embedding model, operating in a 768-dimensional embedding space; (ii) the retriever performs a cosine-similarity search in the ChromaDB vector store and retrieves the $k = 3$ most relevant documents according to their semantic similarity to the query; (iii) the retrieved documents are concatenated using the `_format_docs()` function, which serialises their textual content into a unified context while excluding structural metadata; and (iv) the resulting context is injected into the prompt of the `gemma3n:latest` language model through a `ChatPromptTemplate` that incorporates the system message, the available conversation history, and the current question. The LLM then generates a response based on the retrieved contextual information and the instructions defined in the system prompt.

The system prompt defines explicit constraints intended to guide the generation process and reduce the risk of unsupported responses. These constraints instruct the LLM to: (1) use the retrieved context as the primary and explicit source of information for the response; (2) avoid fabricating information or introducing unsupported content; and (3) explicitly acknowledge when the retrieved material is insufficient to answer the student's question. When no relevant context is retrieved, the system uses the predefined fallback message: 'I did not find information on this topic in my documents, but I can try to help based on what we have already discussed.' These prompt-level constraints are intended to mitigate hallucinations and improve transparency regarding the evidential basis of the generated responses; however, they do not constitute a formal guarantee of factual or mathematical correctness.

4.2.7 Auditability and compliance

The design of educational systems based on AI raises ethical and regulatory concerns that transcend purely technical considerations, requiring robust mechanisms for conversational memory management, safeguarding student privacy, and auditing system operations. The implementation developed integrates these three dimensions through deliberate architectural decisions that aim to balance pedagogical effectiveness, regulatory compliance, and operational transparency.

From the point of view of technical auditability and operational traceability, the system maintains a structured log of events through the Python logging module, configured at the INFO level and formatted with timestamps, component identifiers, severity levels, and descriptive messages. Logged events include: (i) successful or failed initialization of essential services (LLM via ServiceLLM, embeddings via ServiceEmbeddings, RAG pipeline via ServiceRAG); (ii) loading, processing, and counting of knowledge base documents; (iii) ChromaDB vector store reconstruction operations, including removal of previous persistent indexes to ensure synchronization with updated versions of source documents; (iv) invocations

of the agent's run() method, capturing the first 100 characters of user input and logging successes or exceptions in chain execution; (v) conversation history management operations, namely message additions (add_user_message, add_ai_message) and cleanups via clear_history().

Conversational history management is operationalized through the ChatMemory class, which encapsulates an ordered list of HumanMessage and AIMessage objects from the LangChain library, preserving both the textual content and the typology metadata of each message. This history is injected into the prompt through the historic_chat parameter, allowing the language model to access the complete dialogical context when processing each new query. The persistence of the history is limited to the duration of the web session (managed by Streamlit's st.session_state); there is no serialization to disk between sessions, a design decision that prioritizes the privacy of students' conversational data but implies the loss of history in case of application restart.

The vector store, implemented through the Chroma library and the ChromaDB client with local persistence settings (is_persistent=True, persist_directory=self.persist_directory), stores the embeddings of the educational documents in a specific directory (.chroma_db_context_1_algebra). In a strategy to ensure timeliness, the system removes and completely rebuilds the vector store each time the application is initialized (shutil.rmtree followed by Chroma.from_documents), ensuring absolute synchronization between the contents of the source files in the ./docs_rag_context_1_algebra directory and the indexed vectors. This approach, although computationally less efficient than incremental update mechanisms, eliminates the risk of desynchronization between document versions and offers deterministic guarantees of index correctness, critical properties in an educational context where curriculum consistency is mandatory.

The separation between static structured content (presented in tabs) and dynamic conversational assistance (mediated by RAG) is a deliberate architectural decision based on instructional design principles: static content provides structured cognitive scaffolding, suitable for systematic learning and reference consultation, while conversational assistance offers just-in-time support for clarifying specific questions, personalising explanations, and adapting to the individual pace of the student. This duality allows the system to accommodate different learning styles and needs, from students who prefer autonomous exploration of structured material to those who benefit from dialogic interaction to build understanding.

4.2.7.1 Conversational Memory Management Architecture

Dialog history management is a critical component for maintaining conversational coherence in LLM-based tutorial systems. The implementation adopts a hierarchical memory architecture that distinguishes three levels of temporal persistence: active session memory, module memory, and no inter-session persistence. The ChatMemory class operationalises conversational history management through the sequential storage of typed HumanMessage and AIMessage objects, structures provided by the LangChain library that encapsulate both textual content and sender-receiver typology metadata. This type-based approach allows the language model to clearly distinguish between student contributions and system responses when

processing the dialogic context, an essential property for maintaining conversational coherence in multi-turn dialogues. The underlying data structure consists of a chronologically ordered list (self.messages: List[HumanMessage | AIMessage]), which preserves the temporal sequentiality of interactions and allows faithful reconstruction of the dialogic flow.

The conversational memory lifecycle is intrinsically linked to the concept of study modules, defined by the hook questions structured in APP_STRUCTURE. When a student selects a new module, for example, moving from “How to solve linear equations” to “How to solve inequalities”, the handle_module_selection() method explicitly invokes agent.clear_history(), which in turn triggers ChatMemory.clear(), resetting the message list to an empty state. This memory compartmentalisation strategy is based on two pedagogical principles: (i) prevention of contextual interference, preventing concepts from one mathematical domain from contaminating reasoning about different domains; (ii) reduction of cognitive load, allowing the system and the student to focus exclusively on the current topic without dialogical distractions from previous contexts.

The injection of conversational history into the language model prompt is performed through the historic_chat structural parameter in ChatPromptTemplate, which dynamically integrates the messages accumulated in each invocation of the processing chain.

The prompt architecture of the tutoring system, which strategically positions the dialogue history between system instructions and the student’s current query, provides the LLM with contextual capabilities critical for pedagogical effectiveness. This structure allows for accurate anaphoric resolution, correctly interpreting implicit references to previously discussed concepts, as demonstrated in the elucidation of questions containing terms such as ‘this’ or ‘that method.’

Simultaneously, it ensures thematic coherence by maintaining the dialogic focus on the established topic, preventing conversational drifts that would fragment the learning process. Pedagogically, the mechanism supports progressive scaffolding, allowing the tutor to construct sequential explanations that gradually increase conceptual complexity based on already shared knowledge.

Finally, the architecture enables the detection of biases through contextual analysis of subsequent questions, identifying contradictions that reveal misunderstandings of previously covered concepts. Collectively, these capabilities transform the tutorial interaction into a cohesive and contextually sensitive dialogical process, fundamental to an adaptive and effective learning experience.

The session memory, managed by the Streamlit framework’s st.session_state mechanism, maintains the conversational histories of all mathematical contexts (st.session_state.chat_histories = context_id: [] for context_id in APP_STRUCTURE) throughout the active web session. This architectural decision allows students to move between different thematic contexts (Algebra, Functions, Geometry) and, when returning to a previously visited context, resume the dialogue exactly where they left off, provided they have not selected a new module within that context. This intra-session persistence balances cognitive continuity with modular compartmentalisation.

4.2.7.2 Data Privacy and Regulatory Compliance

The absence of inter-session persistence is a design decision guided by principles of privacy-by-design and compliance with the General Data Protection Regulation (GDPR). The system does not implement any mechanism for serialising conversation history to disk, nor does it maintain user databases or authentication mechanisms that allow identities to be associated with sessions. This architecture offers intrinsic privacy guarantees through three fundamental properties: technical anonymity by construction, automatic time limitation (closing the browser or timing out the web session automatically triggers the irretrievable destruction of the entire conversation history), and the impossibility of ex-post re-identification (even in scenarios where the computing infrastructure hosting the system is compromised, there is no persistent historical data that would allow past interactions to be reconstructed or user identities to be inferred).

The fully local operation of the technology stack, with language models (gemma3n:latest) and embeddings (embeddinggemma:latest) running through the Ollama server on the user's machine or on a controlled institutional infrastructure, further reinforces privacy guarantees by eliminating dependencies on third-party cloud services. This local architecture prevents the transmission of conversational data to external infrastructures, mitigating risks of unauthorised data subprocessing and ensuring that the educational institution maintains absolute technical and legal control over the data processed by the system.

Nevertheless, it is imperative to recognise that this non-persistence strategy introduces significant pedagogical trade-offs: the system cannot (i) provide longitudinal progress dashboards to students or teachers; (ii) identify recurring patterns of difficulty for adaptive personalisation of interventions; (iii) generate aggregate usage reports to evaluate the effectiveness of the system. The implementation of analytical functionalities would require explicit consent mechanisms, robust anonymisation (ideally satisfying differential privacy criteria) and clear designation of legal bases for processing, constituting future developments that exceed the scope of this prototype.

4.2.7.3 Audit Mechanisms and Operational Traceability

Auditability is an essential requirement for AI systems applied in educational contexts, allowing verification of compliance with pedagogical requirements, detection of operational anomalies, and transparency regarding system behaviour. The implementation incorporates a structured logging system based on Python's logging module, configured to capture critical events throughout the system's lifecycle. The logging configuration sets the INFO level as the severity threshold (logging.basicConfig(level=logging.INFO)), ensuring that relevant informational events are recorded without overloading the logs with low-level debugging messages (DEBUG level). The structured output format (format="%(asctime)s - %(name)s - %(levelname)s - %(message)s") integrates four essential components: (i) timestamp with second precision, allowing chronological reconstruction of events; (ii) logger identifier (module or class name), facilitating filtering and aggregation by component; (iii) severity level (INFO, WARNING, ERROR, CRITICAL), allowing prioritisation of analysis; (iv) contextualised descriptive message.

The audited events are divided into seven functional categories:

1. Service Initialisation: Logging of connection attempts to Ollama models ('Initialising LLM Ollama: model="gemma3n:latest", url="http://localhost:11434"'), documenting success ('LLM service initialised successfully') or failure with full stack trace (logger.error(..., exc_info=True)). This log allows for the diagnosis of connectivity issues, model unavailability, or configuration errors during system startup.

2. RAG Document Management: Documentation of upload operations ('Uploading documents from: "./docs_rag_context_1_algebra"'), counting of processed documents, and reconstruction of the vector store ('Existing vector store removed to ensure update', 'Creating a new Vector Store'). These logs allow auditing the integrity of the knowledge base, verifying that educational documents have been effectively indexed, and identifying failures in the document ingestion pipeline.

3. Chain Compilation: Logging the LangChain construction process ('Compiling chain...', 'Stateless chain compiled'), documenting the computational structure that processes each interaction. Compilation failures indicate compatibility issues between components or invalid prompt template configurations.

4. Agent invocations: Capture of each call to the run() method with sampling of the first 100 characters of user input ('To be executed with input: "{user_input[:100]}..."'), invocation timestamp, and result (success or exception). These records form the core of the audit trail, allowing interaction sequences to be reconstructed, response times to be measured, and usage patterns to be identified.

5. Memory Management: Explicit documentation of history cleanup operations ('Chat memory for "agent_context_1" has been cleared'), allowing verification that modular compartmentalisation policies are being correctly applied.

6. Errors and Exceptions: Capture complete stack traces for any failures in critical components (logger.error(..., exc_info=True)), providing detailed diagnostic information about the nature and location of errors. The distinction between ERROR (recoverable failures) and CRITICAL (failures that prevent system operation) levels allows maintenance interventions to be prioritised.

7. System Artifact Creation: Documentation of the creation of folders for RAG documents ('RAG document folder guaranteed') and histories (although not persisted, the folder is prepared), and population of default documents when the system is first initialised ('Populated with default documents,' 'Default documents successfully created').

4.3 Synthesis of the Proposed Architecture

This chapter has served to present and consolidate the two architectures proposed within the scope of this research, highlighting their technical, pedagogical and methodological foundations for the development of

intelligent learning support systems. The first architecture focused on the analytical and predictive component, integrating different sources of educational data, data preparation and modelling processes, machine learning algorithms, interpretability mechanisms, and monitoring and decision support components. This architecture demonstrated the feasibility of using predictive models to identify patterns associated with academic performance and school risk, contributing to a data-driven educational approach.

The second architecture focused on the development of an intelligent tutoring system based on LLM and RAG, structured according to a multi-agent approach specialised by mathematical domain. The integration of contextual agents, specific document repositories, augmented retrieval mechanisms and locally executed language models enabled the construction of a pedagogical ecosystem geared towards more precise, contextualised and transparent interactions. In parallel, mechanisms for conversational memory, auditability, operational traceability and compliance with privacy and data protection principles were detailed, ensuring greater robustness and reliability of the proposed solution.

The chapter also highlighted the importance of integrating structured educational content with dynamic conversational support, thereby promoting more adaptive, personalised and student-centred learning approaches. Despite the demonstrated potential, technical and pedagogical limitations associated with the developed architectures were also discussed, particularly regarding scalability, contextual persistence, formal validation of generated responses and dependence on specific models, identifying concrete opportunities for future research.

In this context, the architectures presented lay the conceptual and technological foundations for intelligent educational systems capable of complementing traditional teaching practices through personalised, interpretable and scalable support.

The following chapter focuses on analysing the practical outcomes of the research, presenting various case studies that enable an assessment of the applicability of the solutions developed in real educational contexts, as well as the performance and contribution of the proposed approaches to supporting learning and educational decision-making.

5

Analysis of practical results

tudents' skills can be improved through the use of technology, provided that social and personal competencies are not neglected in the educational process. In this regard, secondary education should foster self-awareness, empathy, teamwork and leadership skills, enabling young people to become more productive members of society and contribute to a better world. Within this context, educational data assumes a central role in supporting evidence-based pedagogical strategies and understanding the multiple factors that shape students' academic development. Accordingly, I collected data from secondary school students to analyse their academic performance and subject choices prior to entering higher education, and subsequently applied machine learning algorithms to obtain predictive results.

The empirical work presented in this chapter is structured around six case studies, which I designed collectively to provide a progressive and complementary validation of the central research hypothesis underpinning this doctoral thesis, namely, that the evidence-based integration of AI, ML, Learning Analytics and intelligent educational systems can contribute significantly to improving pedagogical decision-making, predicting academic performance and supporting personalised learning processes in pre-university educational contexts. Rather than treating these studies as isolated experimental exercises, I designed them as interconnected research phases, with each addressing a specific and well-defined dimension of the broader educational problem while informing and enriching the design of subsequent investigations.

The series of case studies reflects a deliberate trajectory of methodological and conceptual progression. In the initial studies, I focused on identifying and analysing the socio-demographic, academic and contextual factors associated with pupils' performance and educational trajectories, using predictive and explanatory machine learning models based on real institutional data. As the research progressed, I broadened the scope to encompass the fusion of multimodal educational data sources, the application of model interpretability mechanisms, and the incorporation of non-academic variables, including motivational, self-regulatory and family-related dimensions. This enabled me to develop a more holistic and nuanced understanding of the factors associated with academic achievement.

In the final case study, I extended this analytical and predictive perspective to the domain of active educational intervention through the design, implementation and evaluation of a curriculum-aligned intelligent tutoring system for secondary school Mathematics, supported by LLM and RAG architectures. Taken

together, these six studies form a coherent and cumulative research pathway, through which I progressed from educational data analysis and performance prediction towards the conception of practical, explainable and pedagogically grounded intelligent support systems, evaluated in real school environments within the Portuguese educational system.

5.1 Case Study 1: Future of Education on Pre-University

This case study investigates students' academic performance and subject choices before entering higher education, with particular attention to gender and participation in STEAM fields. I analysed data collected from secondary school students to examine relationships between academic results and contextual factors, including gender, socio-economic background, and parents' education. Machine learning techniques were also applied to explore the predictive performance of different models in the analysed educational dataset. This case study originated from the research reported in the published paper [19], which explored the application of machine learning techniques to educational data to investigate factors associated with students' academic pathways and performance.

5.1.1 Context of the paper

The study 'Future of Education on Pre-university Studies' combines two fronts: (i) a systematic review of the literature on female presence in engineering/STEAM (PRISMA methodology, SCOPUS search) and (ii) an empirical analysis (data mining) of a set of school data from a Portuguese secondary school to explore patterns of subject/discipline choice and grade prediction, with a focus on the role of gender.

The study is part of the debate on the low representation of women in STEAM courses, a global phenomenon with social, economic and educational implications.

5.1.2 Motivation

One of the biggest challenges for secondary schools is to provide their students with a more practical education that goes far beyond traditional academic subjects, develops critical thinking, fosters creativity, and stimulates communication and teamwork. Some of the reasons given in [25] for women dropping out of engineering courses are essentially related to gender and differences in self-assessed competences.

Efforts to sensitise students to these gender issues relate to parental motivation, teacher incentives and short classes for children aimed at reducing gender disparities.

I based this study on two complementary needs. The central motivation was to understand the factors that perpetuate this gender disparity, which is identified as a global problem by multiple studies cited in the chapter, including international examples and statistics and structural explanations such as stereotypes, academic culture, social expectations, work-life balance difficulties and harassment.

The other motivation focused on the need to create strategies to mitigate this under-representation of women in STEAM/engineering fields, highlighting the need for a more holistic pre-university education that values not only academic skills, but also social, digital and ethical skills. I considered the integration of technology into education as a means of enhancing skills development, without neglecting the importance of abilities such as critical thinking, teamwork and leadership.

5.1.3 Objectives

The study seeks to answer three research questions (RQs):

- **RQ1)** What factors have been identified for the low representation of women in STEAM courses?
- **RQ2)** What can be done to increase women's attendance?
- **RQ3)** What makes students choose a STEAM course?

From a very early stage, I was concerned with the inclusion of women in higher education and in areas that had, until then, been mostly sought after by men, such as engineering.

For the SLR, I followed the PRISMA methodology and analysed 27 studies from the SCOPUS database, focusing on gender, pre-university education and engineering. I applied exclusion criteria to remove duplicates, inaccessible studies, off-topic studies, and studies focused only on minorities or low-income groups.

For the case study and data mining component, I used a database obtained from the management system of a school in the municipality of Guimarães, comprising 625 students aged 14–20. The variables included year/cycle, subjects and grades per term (scale 0–20, with grades ≥ 10 considered positive), age, gender, household composition, parents' education and professions, among others. Data preprocessing included the removal of constant columns and identifiers, elimination of duplicate records and missing values, encoding of categorical variables, and selection of Mathematics grades for the prediction task. The prediction target was the final Mathematics grade, while the preceding Mathematics grades were used as predictive variables. Target-class imbalance was addressed using RandomOverSampler.

The experimental protocol evaluated five machine learning algorithms: SVM, XGBoost, Random Forest, AdaBoost, and Gradient Boost. Hyperparameters were optimised using GridSearch, and model performance was evaluated using 5-fold cross-validation. A second experimental configuration was created by removing the gender variable from the predictor set, allowing the predictive performance obtained with and without gender to be compared under the same modelling procedure.

The experimental protocol also considered the potential risk of data leakage during preprocessing and model validation. In particular, the placement of the class-balancing procedure based on RandomOverSampler relative to the cross-validation procedure is relevant, since applying oversampling before the data are divided into training and validation partitions may introduce information from the validation data into

the training process. Therefore, the use of RandomOverSampler and its position within the validation pipeline should be explicitly documented when interpreting the reported cross-validation results.

5.1.4 Results

In the analysis I carried out, the distribution of students by field of study revealed distinct patterns of gender representation. The field of "Science and Technology" was the most balanced, with 256 students, of whom 128 were women and 128 were men. In contrast, the field of "Socioeconomic Sciences" had a slight male majority, with 54.3% of students being male. The fields of "Arts" and "Humanities and Social Sciences" were predominantly female, with women representing 59.7% and 55.3% of the students, respectively (Figure 31).

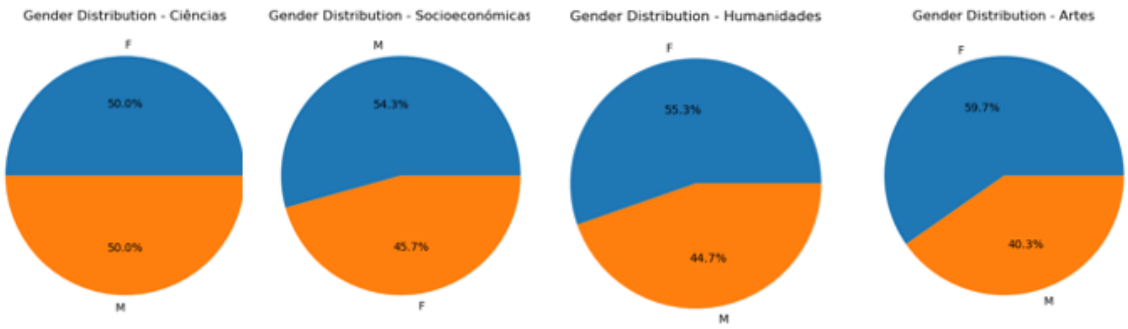


Figure 31: Gender distribution by area.

In terms of academic performance, female students had higher average grades across most of the areas analysed. For example, in the area of "Science and Technology", the average grade for female students was approximately 16.05 points, compared with 15.38 points for male students. A similar pattern was observed in subjects such as Chemistry and Physics, where female students also achieved higher average grades. These results indicate that, within the analysed dataset, female students generally presented higher average academic performance across the areas considered.

One relevant aspect I identified concerned the choice of optional subjects in the final year of schooling, Year 12. In the area of 'Science and Technology', I identified three main combinations: 'Computer Applications B + Chemistry' (73.9%), "Computer Applications B + Biology" (25%) and "Computer Applications B" alone (1.1%). Of the total number of students who chose Computer Applications B, 51 were men and 41 were women. Among the 68 students enrolled in Chemistry, 40 were men, while in Biology, of the 22 enrolments, 12 were women. These results (Figure 32) indicate differences in subject selection by gender within the analysed school population, with a higher proportion of female students among those enrolled in Biology.

In the specific case of Mathematics A, although there was a greater number of male students, with a difference of 12 students, female students achieved an average grade almost one point higher than male

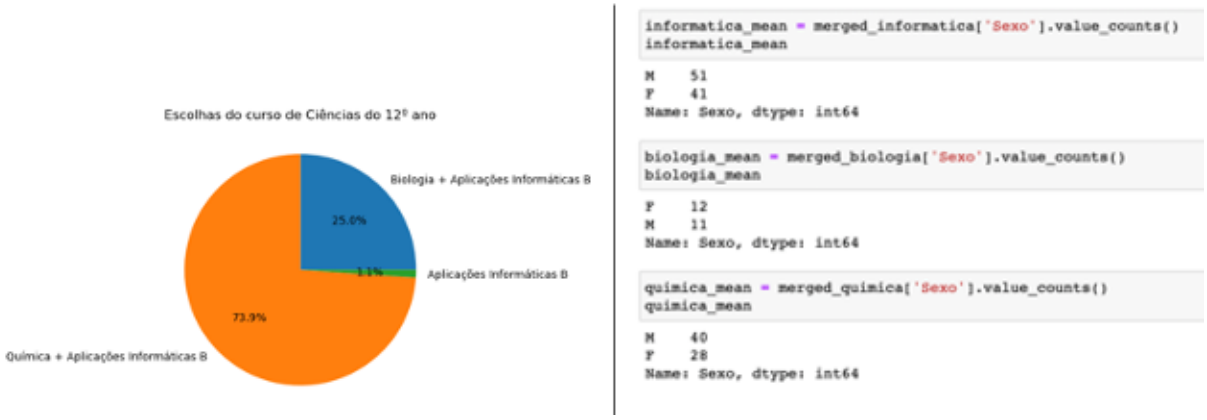


Figure 32: Last year subject distribution by gender.

students. This result is consistent with the differences in average academic performance observed across several areas of the analysed dataset.

Regarding the results of the machine learning models, Table 5 shows that the SVM model performed best on the complete dataset, with an accuracy of approximately 0.87, an F1 score of 0.86, and a cross-validation score of 0.7961. XGBoost and Random Forest achieved accuracies of approximately 0.85, while Gradient Boost reached 0.83. AdaBoost showed substantially lower performance, with an accuracy of 0.19 and an F1 score of 0.11. The difference between its precision (0.78) and F1 score (0.11) indicates that its classification performance was uneven across the evaluation metrics and should therefore be interpreted with caution (Table 5).

Table 5: Results for Machine Learning Models

Dataset	Model	Accuracy	Precision	F1 Score	Recall	CV
Complete Dataset	SVM	**0.87**	0.87	0.86	0.87	0.7961
	XGBoost	0.85	0.85	0.84	0.85	0.8349
	Random Forest	0.85	0.84	0.84	0.85	0.8438
	Adaboost	0.19	0.78	0.11	0.19	0.2991
	Gradient Boost	0.83	0.83	0.83	0.83	0.8199
Dataset With No Gender	SVM	**0.86**	0.87	0.86	0.86	0.7871
	XGBoost	0.85	0.85	0.84	0.85	0.8512
	Random Forest	0.85	0.84	0.84	0.85	0.8498
	Adaboost	0.19	0.78	0.11	0.19	0.2991
	Gradient Boost	0.83	0.82	0.82	0.83	0.8258

When the gender variable was removed from the dataset, the results remained virtually unchanged: the accuracy of SVM decreased slightly from 0.87 to 0.86, and the XGBoost and Random Forest models remained around 0.85 (Table 5). These results indicate that, within the specific dataset and prediction task analysed, the inclusion of gender had a limited effect on the predictive performance of the evaluated models. This finding should not be interpreted as evidence that gender is irrelevant to educational outcomes; rather, it indicates that its contribution to this particular prediction task was limited when compared with

the other variables included in the dataset.

5.1.5 Outcomes

Based on the results I obtained, I concluded that pre-university education should adopt a more holistic approach, integrating technical, digital and socio-emotional skills in a balanced way, in order to prepare students for a constantly changing and technologically mediated world. Although the under-representation of women in engineering fields is a widely documented global phenomenon, the case study I analysed, focusing on a Portuguese school in Guimarães, revealed that, at the local level, female students not only consistently achieved better academic results, but that the gender gap was less pronounced than in university or international contexts. However, I also identified a subtle but relevant pattern: female students tended to choose subjects such as Biology over options more directly related to engineering, such as Computing or Chemistry, suggesting that cultural and social factors may continue to influence educational choices even in contexts of apparent statistical parity.

In terms of practical implications, I identified the importance of intentional, evidence-based educational interventions. Summer programmes with hands-on components, project-based learning, peer or professional mentoring, early awareness-raising initiatives, and teacher training aimed at addressing gender stereotypes are identified as promising strategies for increasing women's interest and retention in STEAM fields, particularly in engineering.

From a critical perspective, I reflected on the results of the machine learning models used. The fact that removing the gender variable did not significantly affect predictive performance suggested that, within this dataset, historical academic variables, such as previous grades, captured most of the information relevant to predicting performance in Mathematics A. This does not imply that gender is irrelevant in educational or social contexts; rather, it indicates that its direct contribution to this specific predictive task was limited. Nevertheless, I interpreted this finding with caution, given the contextual nature of the dataset and the complexity of the factors influencing academic achievement.

Despite the promising findings, I identified several limitations that should be considered when interpreting the results. The study was conducted using data collected from a single educational institution, which may limit the external validity and generalisability of the findings. In addition, the use of oversampling techniques to address class imbalance may introduce unintended biases if not carefully controlled. The cross-sectional nature of the dataset also prevents the analysis of changes in students' aspirations, performance, and educational trajectories over time. Finally, the variability observed in the performance of certain algorithms suggests that model suitability may be influenced by the specific characteristics of the educational context.

For future research, I propose longitudinal studies that follow students throughout their educational journey, enabling a deeper understanding of the factors influencing academic pathways over time. I also propose combining quantitative predictive models with qualitative methods, such as interviews and focus groups, to provide richer insights into the personal, social, and contextual factors underlying students'

educational choices. Additional research should also evaluate the effectiveness of interventions such as mentoring programmes, summer schools, and project-based learning activities. From a machine learning perspective, I propose that future studies prioritise external validation, transparent reporting of data pre-processing procedures, and enhanced model interpretability to ensure reliable and ethically responsible educational applications.

5.2 Case Study 2: Forecasting and AI-Powered Exploration of Student choices across different types of Mathematics in Pre-University School

In today's digital age, it is imperative that pre-university education evolves in ways that strengthen students' critical thinking, teamwork, and social skills. Rather than focusing exclusively on knowledge transmission, educational systems should aim to equip students with the competencies needed to navigate an increasingly complex and technology-driven world. Within this context, I conducted this case study, based on the research published in [18], to examine the choices made by pre-university students regarding different mathematics pathways. I analysed how factors such as gender, parents' occupation, and previous academic performance influenced students' future higher education choices. Using data collected from two schools located in different municipalities and educational contexts, I applied a range of machine learning techniques to analyse these relationships. My analysis revealed that gender appears to influence the selection of certain mathematics pathways. However, I also found that students' educational choices are shaped by a broader combination of academic, personal, and contextual factors.

5.2.1 Context of the paper

The [18], I investigated how secondary school students choose between Mathematics A and Mathematics Applied to Social Sciences (MACS), and to what extent variables such as gender, age and parental characteristics influence grades and future trajectories, including propensity for STEAM areas. I combined descriptive analysis with machine learning (ML) models and compared two public schools in northern Portugal, Guimarães (n=625; 14–20 years old) and Braga (n=621; 14–20 years old).

In Braga, I considered three mathematics streams (A, MACS, and B), whereas in Guimarães there was no Mathematics B. The databases included subjects and grades (0–20; ≥ 10 positive), age, gender, household composition, and parents' education and profession/field. I focused the empirical analysis on Mathematics A and MACS.

Conceptually, I based the study on literature concerning self-efficacy, motivation and gender inequalities in STEAM. I considered that pre-university education should go beyond the transmission of content by developing critical thinking, creativity, communication and teamwork, supported by technology, to better prepare students for higher education and the job market.

5.2.2 Motivation

Two concerns guide my study:

1. Preparing students for the challenges of the 21st century requires developing transversal competencies and ensuring success in Mathematics, a gateway subject for many STEAM programmes and a key factor in access to, and achievement in, higher education.

2. Understanding the factors behind inequalities in choice and performance (e.g., gender and family context) is important for informing educational guidance and targeted interventions that support vocational decisions better aligned with students' potential and interests.

5.2.3 Objectives

The study aims to answer three research questions (RQs):

- **RQ1:** Does the choice of mathematics type by secondary school students depend on their gender?

- **RQ2:** Is the student's age associated with their mathematics grade?

- **RQ3:** Are parents' qualifications associated with students' mathematics grades?

In parallel, I aimed to predict grades in Mathematics A and MACS using ML techniques to (i) examine the contribution of sensitive (gender) and contextual (parental characteristics) variables to predictive performance and (ii) identify patterns that could inform educational guidance and pedagogical decision-making.

The data used in this study came from two secondary schools located in the municipalities of Guimarães and Braga. After an initial preparation phase, I checked for missing or duplicate values and removed constant columns, such as those referring to the school year, and personal identifiers in order to preserve the anonymity of the students. I coded categorical variables numerically and applied an over-sampling technique to balance the distribution of classes, resulting in a more balanced dataset for model training.

The experimental protocol comprised model training and validation following the preprocessing procedures described above. Five classification algorithms were evaluated: SVM, XGBoost, RF, AdaBoost, and Gradient Boost. The models were evaluated separately for Mathematics A and MACS, considering three feature scenarios: the complete dataset, a dataset excluding gender, and a dataset excluding parental occupation. Hyperparameter optimisation was performed using GridSearch, and model performance was assessed using cross-validation.

5.2.4 Results

With regard to the distribution by gender and by mathematics route, I found that, in Mathematics A, the numbers were fairly balanced: in Guimarães, there were 187 girls and 197 boys, while in Braga there were 268 girls and 252 boys. In the subject of Mathematics Applied to Social Sciences (MACS), I found a more pronounced female presence, with 65 girls and 49 boys in Guimarães, and 45 girls and 36 boys in Braga. Overall, I observed that Mathematics A was mainly associated with Science and Technology and Socioeconomic Science courses, while MACS was typically chosen by Humanities and Social Science students.

The distribution of marks showed that, in Mathematics A, the maximum mark was 20 in both schools, with the minimum mark being 7 in Guimarães and 3 in Braga, and the mode being 10. In MACS, the range varied between 8 and 19 in Guimarães and between 5 and 20 in Braga, with the school in Guimarães showing a slightly higher overall profile of results in this subject.

With regard to the research questions, I obtained the following results:

RQ1 — Influence of gender on choice and performance

My analysis of the data showed that, although enrolment in Mathematics A was relatively balanced between genders, female students had higher average grades than male students in both schools. For example, in Guimarães, the average grade for female students in Mathematics A was 13.35, compared to 12.45 for male students; in Braga, the corresponding values were 14.25 and 13.05 (Table 6). IIn MACS, the difference was more pronounced, particularly in Braga, where female students had an average grade of 12.76 compared with 9.86 for male students. Among the 20 highest-performing students, approximately 80% to 85% were girls. These results indicate that, within the analysed datasets, female students achieved higher average grades in both mathematics routes, despite the relatively balanced gender distribution in Mathematics A.

Table 6: Mean and standard deviation by gender for Mathematics A and MACS in the municipalities of Guimarães and Braga.

School	Subject	Gender	Mean	Standard Deviation
Guimarães	*Mathematics A*	Female	13.353	3.185
		Male	12.447	3.139
	MACS	Female	13.738	3.270
		Male	12.939	2.478
Braga	*Mathematics A*	Female	14.254	3.847
		Male	13.052	4.042
	MACS	Female	12.756	4.173
		Male	9.861	3.634

RQ2 — Impact of age on performance

My results indicates differences in academic performance across age groups. In Mathematics A, the strongest results were concentrated among students aged between 15 and 17 years, corresponding broadly to the expected age range for these school years. In MACS, the pattern was less consistent: in

Guimarães, the highest grades were mainly observed among students aged 15 to 16, whereas the Braga dataset showed a more dispersed age distribution. These results indicate that the relationship between age and mathematics performance varied according to the mathematics route and school context analysed.

RQ3 — Parental context and performance

Based on the literature reviewed, parents' education and profession may interact with other contextual factors, such as socioeconomic conditions and parental involvement, rather than acting as isolated determinants of students' academic performance. In the empirical study, I found that the parents of students in Braga had, on average, higher levels of education than those in Guimarães. The distribution of mathematics pathways also differed between the two school populations, with Mathematics A being associated with educational pathways oriented towards areas such as Engineering, Economics, and Management. However, the results obtained in this study do not establish a direct causal relationship between parental educational characteristics and students' choice of Mathematics A. Furthermore, the observed differences in parental educational characteristics were not accompanied by clear differences in the overall mathematics grades between the two school populations.

To further investigate these relationships, I tested five classification algorithms, SVM, XGBoost, Random Forest, AdaBoost, and Gradient Boost, under three different feature scenarios: the complete set of variables, a dataset excluding the gender variable, and a dataset excluding parental occupation. I evaluated these configurations separately for Mathematics A and MACS in the Guimarães and Braga datasets. Table 7 summarises the resulting classification accuracy and provides a systematic comparison of the predictive performance obtained across the different educational contexts and feature subsets.

Machine Learning Classification Performance

Table 7: Classification Accuracy Across Machine Learning Models and Feature Scenarios.

School	Subject	Feature Scenario	SVM	XGBoost	Random Forest	AdaBoost	Gradient Boost
Guimarães $N = 625$	*Math A*	Full Dataset	**0.87**	0.85	0.84	0.18	0.86
		Excl. Gender	0.86	0.84	0.84	0.18	0.85
		Excl. Parental Occ.	0.85	0.83	0.83	0.19	0.84
	MACS	Full Dataset	**0.81**	0.79	0.78	0.19	0.80
		Excl. Gender	0.78	0.76	0.76	0.18	0.77
		Excl. Parental Occ.	0.75	0.74	0.73	0.19	0.74
Braga $N = 621$	*Math A*	Full Dataset	**0.87**	0.86	0.85	0.19	**0.87**
		Excl. Gender	0.86	0.85	0.84	0.19	0.86
		Excl. Parental Occ.	0.86	0.84	0.84	0.18	0.85
	MACS	Full Dataset	**0.89**	0.87	0.86	0.19	**0.89**
		Excl. Gender	0.87	0.85	0.84	0.18	0.86
		Excl. Parental Occ.	0.82	0.81	0.80	0.19	0.81

Table 7 presents the classification accuracy obtained by the different machine learning models across the two schools, the two mathematics subjects, and the different feature scenarios. Overall, I observed that SVM, XGBoost, Random Forest, and Gradient Boost achieved comparatively high and consistent

accuracy values, whereas AdaBoost presented substantially lower accuracy, ranging from approximately 0.18 to 0.19 across all scenarios.

AdaBoost presented substantially lower accuracy than the other evaluated algorithms, with values ranging from approximately (≈ 0.18–0.19) across all scenarios. The results reported in this study do not provide sufficient evidence to establish the specific cause of this behaviour. Therefore, the AdaBoost results are interpreted descriptively, without attributing the observed performance to a particular implementation, hyperparameter effect, or property of the student dataset.

I found that the predictive performance varied according to both the mathematics route and the feature scenario. In Mathematics A, the models generally exhibited high and relatively stable performance, with accuracy ranging from 0.83 to 0.87 across the evaluated configurations. In Guimarães, SVM achieved the highest accuracy with the complete dataset (0.87), followed by Gradient Boost (0.86), while XGBoost and Random Forest achieved 0.85 and 0.84, respectively. When I excluded gender, the accuracy of SVM decreased slightly to 0.86, while the remaining models also showed small reductions. When I excluded parental occupation, SVM decreased to 0.85, and the other models followed a similar pattern.

I observed a comparable pattern in Braga, where SVM and Gradient Boost both achieved 0.87 using the complete dataset. When I removed gender, both models decreased slightly to 0.86, while the exclusion of parental occupation resulted in accuracies of 0.86 for SVM and 0.85 for Gradient Boost. These results indicate that, for Mathematics A, the removal of individual demographic or parental variables produced relatively limited changes in predictive performance.

In contrast, I observed a more pronounced effect of feature removal in MACS. In the Guimarães dataset, SVM achieved an accuracy of 0.81 with the complete dataset, decreasing to 0.78 when I excluded gender and to 0.75 when I excluded parental occupation. The remaining models showed a similar downward trend, with XGBoost decreasing from 0.79 to 0.76 and then to 0.74, while Random Forest decreased from 0.78 to 0.76 and subsequently to 0.73.

In the Braga MACS dataset, I obtained the highest predictive performance observed in this case study. Using the complete dataset, both SVM and Gradient Boost achieved an accuracy of 0.89, followed by XGBoost at 0.87 and Random Forest at 0.86. When I excluded gender, the accuracy of SVM decreased to 0.87, while Gradient Boost decreased to 0.86. A more substantial reduction occurred when I excluded parental occupation: SVM decreased to 0.82, XGBoost to 0.81, Random Forest to 0.80, and Gradient Boost to 0.81.

Overall, I found that the exclusion of gender produced relatively small reductions in predictive accuracy, particularly for Mathematics A, although its effect was more visible in MACS. I also observed that the exclusion of parental occupation was associated with a more pronounced reduction in predictive accuracy on predictive performance in MACS, especially in the Braga dataset. These results reinforced my observation that the contribution of demographic and family-related variables to prediction was context-dependent and varied according to the mathematics route and school population analysed. Therefore, the findings support the need to consider the educational context and the characteristics of the prediction task when interpreting the contribution of individual variables.

5.2.5 Outcomes

Based on the findings I obtained, I concluded that gender, age, and family background were associated with both students' mathematics pathway choices (Mathematics A versus MACS) and their academic performance. I found that female students generally achieved higher results across both subjects and were more strongly represented among the highest-performing students in MACS. I also found that students who were at the expected age for their school year tended to perform better academically.

From a predictive perspective, I found that the machine learning models demonstrated strong performance, particularly SVM, XGBoost, and Random Forest. I observed that, while the exclusion of variables such as gender or parental occupation influenced model accuracy, the magnitude and direction of these effects varied across subjects and school contexts, suggesting that their predictive contribution was context-dependent.

Overall, my findings highlight the potential of these results to support educational guidance, promote gender equity, and inform pedagogical strategies that foster transversal skills and more informed academic and career choices at the end of secondary education.

5.3 Case Study 3: AI-driven educational transformation in secondary schools: Leveraging data insights for inclusive learning environments

There is a growing trend towards teaching tailored to students' characteristics and the rapid development of models for predicting student failure. This transformation has been significantly influenced by the integration of machine learning techniques, which have played a crucial role in leveraging data to improve educational practices. Within this context, I based this case study on the research published in [42], in which I explored the application of machine learning methods to educational data to better understand and predict student outcomes.

5.3.1 Context of the paper

I based this case study on the research presented in [42], through which I examined the ethical considerations and challenges associated with implementing AI systems in educational environments, particularly regarding equity, transparency and data privacy. I considered both this ethical perspective and the empirical analysis I conducted using data from a public school in the municipality of Braga. The dataset comprised 621 students distributed across Mathematics A ($n = 520$), Mathematics B ($n = 20$) and MACS ($n = 81$). It included academic variables, namely grades from three assessment periods on a scale from 0 to 20, with grades of 10 or above considered positive, as well as contextual variables such

as gender, age, place of birth, residence, access to a computer and the Internet, school transport, nationality, and parents' education and employment, among others. I analysed each mathematics subject separately and subsequently compared the results, combining machine learning models with chi-square tests to examine relationships between variables.

5.3.2 Motivation

I focused on persistent gaps in mathematics learning and on the need to personalise educational support through tailored learning paths, adaptive assistance and targeted interventions. Within this context, I examined how AI and ML can support the personalisation of teaching, the prediction of academic outcomes and the identification of students at risk. I also considered their potential to promote more inclusive and equitable educational environments by accounting for socioeconomic and behavioural variables, while respecting ethical principles such as transparency, privacy and non-discrimination. In addition, I considered the ambivalent perceptions of AI among students: while some express concerns about its use, others welcome the opportunities it provides for feedback and access to additional resources. I therefore considered responsible, student-centred adoption to be an important aspect of applying these technologies to education.

5.3.3 Objectives

I organised my investigation around five research questions (RQ):

- **RQ1)** For what purposes are ML techniques used in education (diagnosis, assessment, prediction, dropout prevention)?
- **RQ2)** Are the predictive results identical across different mathematics pathways (A, B, MACS)?
- **RQ3)** What individual differences should be considered for personalisation (e.g., age, gender, family context, digital access)?
- **RQ4)** Which ML algorithms provide the best predictive performance for student academic outcomes?
- **RQ5)** Are there relationships between categorical variables (e.g., Internet/computer access versus grades) according to the chi-square test?

In parallel, I aimed to predict students' grades in Mathematics A and MACS using machine learning techniques. Through this analysis, I sought to examine the contribution of sensitive and contextual variables to predictive performance and to identify findings that could inform educational guidance and targeted pedagogical support.

5.3.4 Results

I began by examining the sample and identifying the characteristics of each mathematics pathway. I found that Mathematics A had the most balanced gender distribution, with 268 female students and 252 male students. Mathematics B showed a strong female predominance, with 18 girls and only 2 boys, while MACS included 45 female students and 36 male students. I also observed that students in all three pathways were generally within the expected age range for their educational level, and that Mathematics A had the greatest diversity of nationalities. When I examined access to digital resources, I observed a general tendency for students with access to both the Internet and a computer to present higher academic performance.

I then applied the chi-square test to investigate the relationship between final grades and access to digital resources. In Mathematics A, I identified statistically significant associations between grades and Internet access ($\chi^2 = 36.11$; $p = 0.0044$), and between grades and computer access ($\chi^2 = 39.54$; $p = 0.0015$). However, I found that these associations did not reach statistical significance in Mathematics B or MACS. The p-values ranged from 0.28 to 0.31 for Mathematics B and from 0.17 to 0.19 for MACS (Table 8). Based on these results, I concluded that the relationship between access to digital resources and academic performance was statistically supported in Mathematics A, whereas the results for the other two pathways did not provide sufficient evidence of a similar association.

Table 8: Chi-squared test to study the relationship between the variables grades, internet access and computer access.

Chi-squared test statistic			
Variables	**Maths Type**	**Test Statistic**	**p-value**
Grades and internet access	Mat A	36.1124	0.0044
	Mat B	11.5751	0.3145
	Macs	19.6334	0.1865
Grades and computer access	Mat A	39.5379	0.0015
	Mat B	12.0139	0.2841
	Macs	20.1096	0.1678

I also examined academic performance by gender. The graphs showed that female students generally achieved higher average grades and were more strongly represented at the upper end of the grade distribution. This pattern was particularly evident in MACS.

To evaluate predictive performance, I tested five machine learning algorithms: SVM, Random Forest, XGBoost, AdaBoost and Gradient Boost. I tuned the hyperparameters using GridSearch and evaluated the models through cross-validation. To investigate the contribution of individual factors, I compared the complete set of variables with alternative configurations in which I excluded, separately, gender, computer access, Internet access and public transport use.

For Mathematics A, I obtained high predictive accuracy across several models. SVM and Gradient Boost achieved values close to 0.87–0.88, while XGBoost and Random Forest ranged from 0.84 to 0.87.

In contrast, AdaBoost achieved substantially lower accuracy, at approximately 0.18. I found that removing variables such as gender, computer access or Internet access produced only small variations in predictive performance, generally between 0.01 and 0.05. These results indicate that the variables removed in these scenarios made a relatively limited contribution to the predictive performance of the evaluated models, suggesting that other variables, including previous-term grades and other academic characteristics, retained substantial predictive information.

For Mathematics B, I obtained the highest accuracy values, approximately 0.78, with XGBoost and Gradient Boost, while SVM achieved 0.67. I observed only modest changes after removing individual variables, with some configurations excluding gender or Internet access maintaining an accuracy of approximately 0.78.

For MACS, I obtained baseline accuracy values of approximately 0.89 with SVM and Gradient Boost. After removing gender, XGBoost achieved an accuracy of 0.94, which was the highest value reported in the study. AdaBoost performed less effectively, with an accuracy of approximately 0.43. Excluding computer or Internet access had virtually no effect on accuracy, while removing public transport use produced only marginal changes. These results indicate that the contribution of the analysed variables to predictive performance varied across mathematics pathways and algorithms. In particular, the increase observed after removing gender from the MACS model indicates that the exclusion of a variable does not necessarily reduce predictive accuracy and should therefore be interpreted in relation to the specific dataset and prediction task.

5.3.5 Outcomes

Based on the results, I identified several findings concerning the use of AI and ML to support personalisation, adaptive assistance and targeted interventions in secondary education. I found evidence of a statistically significant association between access to digital resources and academic performance in Mathematics A, but I did not observe equivalent evidence in Mathematics B or MACS. I also observed a consistent advantage for female students in average grades and representation among the highest-performing students, particularly in MACS.

Regarding predictive modelling, I found that SVM, Gradient Boost and XGBoost achieved comparatively high accuracy in several configurations, whereas AdaBoost produced substantially lower results. I also observed that excluding gender did not consistently reduce predictive performance and, in the case of XGBoost for MACS, was associated with an increase in accuracy. Based on these findings, I concluded that the contribution of gender to prediction depended on the mathematics pathway and the algorithm used, and that its removal did not necessarily impair model performance when academic history and other characteristics were available.

Finally, I considered the contextual limitations of these findings, particularly their reliance on data from a single urban school. I therefore identified the expansion of the dataset, further refinement and validation of predictive models, and the development of ethically grounded precision education practices

as important directions for future work.

5.4 Case Study 4: Data Fusion and Predictive Modeling for Academic Performance Assessment: A Case Study on Grade Variation

5.4.1 Context of the paper

I based the case study presented in this thesis on the research published in the paper [44]. In this study, I investigated the application of AI and Data Fusion techniques to predict variations in students' academic performance, particularly changes in Mathematics grades between the first and second school terms. I aimed to support the early identification of students at risk of academic difficulties and to enable timely pedagogical interventions. Within the fields of learning analytics and multimodal data fusion in education, I adopted a case study approach in a school located in the district of Braga, combining administrative records with questionnaire data collected from students.

The study published in the article [44] was based on an initial sample of 845 students in the 2nd and 3rd cycles of a school in the district of Braga, Portugal. After performing data cleaning, including the treatment of missing values and the elimination of redundant or irrelevant columns for the prediction task, I obtained a final dataset consisting of 843 records and 119 attributes, prior to the feature selection stage. I integrated multiple data sources, combining strictly academic information, such as students' Mathematics grades in the first and second school terms, with key sociodemographic data, including gender, age and school cycle. To further characterise each student's profile, I incorporated information about parents, namely their level of education and marital status. In addition, I collected data through a detailed questionnaire completed by the students, which provided information on a range of psychosocial and behavioural dimensions, including intrinsic and extrinsic motivation, short- and long-term self-regulation skills, perceived support from teachers, peers, and family, and students' involvement and integration in school life.

Within the predictive modelling pipeline, I formally defined the target variable as the variation in students' academic performance in core subjects between consecutive evaluation periods, specifically between the first term (T_1) and the second term (T_2). Rather than predicting the exact grade obtained in the second term, I formulated the task as a multiclass classification problem aimed at identifying the direction of grade variation. I therefore defined the target variable with a cardinality of three ($K = 3$), represented through a ternary encoding scheme:

- **Class 0 — Decrease/Worsened ($y = 0$):** students whose academic grade decreased from T_1 to T_2, such that $\text{Grade}T_2 < \text{Grade}T_1$.

- **Class 1 — Stability/Equal ($y = 1$):** students whose academic grade remained unchanged between the two evaluation periods, such that $\text{Grade}T_2 = \text{Grade}T_1$.

- **Class 2 — Increase/Improved ($y = 2$):** students whose academic grade increased from T_1 to T_2, such that $\text{Grade}T_2 > \text{Grade}T_1$.

I used this formulation to enable the predictive models not only to identify students experiencing academic deterioration, but also to distinguish between stable and improving performance, thereby providing a basis for earlier and more targeted pedagogical intervention.

5.4.2 Motivation

In an educational context marked by the need to predict and mitigate school failure, I focused on the practical problem of identifying students at risk of academic failure in a timely manner. My central motivation was to enable educational institutions to intervene proactively by implementing corrective measures such as personalised tutoring, adjustments to teaching methodologies and the provision of additional academic support before difficulties become entrenched and irreversible. To address this challenge, I proposed an innovative technical-methodological solution focused on exploiting the fusion of multimodal data.

Specifically, I examined whether combining multiple sources of information through an early fusion approach could improve the accuracy of academic performance prediction models. I integrated traditional administrative data, such as grades and socio-economic indicators, with psychosocial information collected through questionnaires addressing aspects such as motivation, self-regulation, and perceived support. I considered that a more comprehensive representation of the student profile could provide greater predictive power than models developed from each data source independently. I therefore considered the potential implications of this approach for the development of adaptive educational systems capable of delivering personalised interventions and learning experiences in a timely manner.

Ultimately, I considered that this refined predictive capability could support the formulation of more effective and targeted school policies, contributing to the prevention of school dropout and failure and promoting a more successful educational path for students.

5.4.3 Objectives

My main objective was to develop and implement a set of methodological procedures for aggregating and integrating heterogeneous data from multiple sources, namely administrative school information, detailed psychosocial questionnaires, and data concerning students' family context. The technical approach focused on the implementation of an early fusion strategy, which consists of concatenating features from different sources into a single, consolidated input vector, thereby creating a unified representation of the students' profiles.

Based on this integrated dataset, I developed predictive models to estimate variations in students' grades between the first and second terms (T_1 and T_2) in Portuguese and Mathematics, with Mathematics used as the main example for more detailed analysis. I systematically evaluated the predictive performance obtained from the fused dataset and compared it with models based on individual data sources. In parallel, I evaluated several machine learning algorithms, including XGBoost, DT, and RF, to identify suitable approaches for the prediction task. Model performance was assessed using Accuracy, *F1-Score*, Recall, and Precision, providing a comprehensive evaluation of the predictive performance of the different modelling approaches and of the potential contribution of heterogeneous data integration.

To ensure that the predictive task reflected a realistic early-warning scenario, I conducted a data leakage audit across the 119 original attributes before model training. Predictor variables were restricted to information available at or before the end of the first term (T_1), including first-term grades, baseline demographic information obtained from the Student Information System (SIS), and initial attitudinal and socioeconomic questionnaire data. Variables derived from second-term (T_2) assessments, attendance records, or end-of-year evaluations were excluded from the predictor set because their availability would introduce information from the future into the prediction process.

I also enforced isolation between training and validation data throughout the modelling pipeline. Data-dependent preprocessing operations, including normalization, correlation-based feature selection, and oversampling, were performed exclusively using the training data within each cross-validation fold. This procedure was intended to prevent information from the validation partitions from influencing model development and to provide a more reliable estimate of predictive performance and generalisation within the early-warning prediction framework.

5.4.4 Results

I first examined the predictive performance obtained before and after feature selection. The comparison between the complete datasets Math_Dataset and Port_Dataset and their corresponding selected-feature datasets (MFS_Dataset and PFS_Dataset) were used to assess the effect of reducing the feature space on the evaluated prediction tasks.

The Mathematics prediction task was formulated as a three-class classification problem based on the variation between students' first and second term grades. Class 0 represents a decrease in academic performance, Class 1 represents stable performance, and Class 2 represents an improvement in academic performance. The original Math_Dataset contained 75 instances in Class 0, 635 in Class 1, and 133 in Class 2, indicating a pronounced class imbalance. This imbalance was particularly relevant to the early-warning objective because students experiencing a decline in academic performance constitute an important group for timely pedagogical support.

Target Variable Distribution and Class Balancing

The prediction task was formulated as a multiclass classification problem with a target variable of cardinality $K = 3$, representing the variation in students' grades between the first term (T_1) and the second term (T_2). The ternary target variable ($y \in 0, 1, 2$) was defined as follows:

- **Class 0 (Worsened / At-Risk):** $\text{Grade}T_2 - \text{Grade}T_1 < 0$ ($y = 0$).
- **Class 1 (Equal / Stable):** $\text{Grade}T_2 - \text{Grade}T_1 = 0$ ($y = 1$).
- **Class 2 (Improved):** $\text{Grade}T_2 - \text{Grade}T_1 > 0$ ($y = 2$).

As detailed in Table 9,I found that the original merged dataset Math_Dataset, with $N = 843$, exhibited a pronounced class imbalance. The stable majority category (Class 1) accounted for 75.3% of the instances ($n = 635$), whereas the critical minority category, corresponding to students whose performance worsened (Class 0), represented only 8.9% ($n = 75$). Class 2, corresponding to an improvement in academic performance, accounted for the remaining 15.8% of instances ($n = 133$).

To mitigate the potential bias of the predictive models towards the majority class, RandomOverSampler was applied within the training data of each cross-validation fold. The procedure increased the representation of the minority classes to match the majority class, resulting in 635 instances per class in the balanced training partitions. Consequently, each balanced training partition contained 1905 instances.

Table 9: Target Variable Distribution and Random Oversampling Balance in Math_Dataset.

Target Class (y)	**Description/Grade Variation (Δ)**	**Original Frequency**	**Original %**	**Balanced Frequency**
Class 0	**Decrease/Worsened** ($\Delta < 0$)	**75**	**8.9%**	**635**
Class 1	**Equal/Stable** ($\Delta = 0$)	635	75.3%	635
Class 2	**Increase/Improved** ($\Delta > 0$)	133	15.8%	635
Total		**843**	**100.0%**	**1,905**

To evaluate the effect of feature selection on the prediction of Portuguese language performance, I first compared the results obtained using the original `Port_Dataset` with those obtained after feature selection using the `PFS_Dataset`. Table 10 presents this comparison across the XGBoost, DT, and RF algorithms.

The results show a clear improvement in predictive performance after feature selection. For the original Port_Dataset, XGBoost achieved an accuracy of 0.849, while Decision Tree and Random Forest achieved 0.812 and 0.845, respectively. After feature selection, the performance of all three models increased substantially. XGBoost achieved the highest overall performance, with an accuracy of 0.966, an F1-score of 0.965, a recall of 0.966, and a precision of 0.967. These results indicate that the selected-feature configuration achieved higher predictive performance for the evaluated Portuguese language prediction task.

Based on the aggregate metrics, XGBoost trained on the PFS_Dataset was selected for further class-level analysis. Aggregate performance metrics, however, do not fully describe how predictions are distributed across the individual target classes. Therefore, the analysis was complemented with a confusion matrix and per-class performance metrics for the selected model.

Table 10: Model Performance Comparison for Portuguese Language Before and After Feature Selection.

Dataset	Algorithm	Accuracy	F1-Score	Recall	Precision
Port_Dataset	XGBoost	0.849	0.835	0.849	0.840
Port_Dataset	Decision Tree	0.812	0.798	0.812	0.805
Port_Dataset	Random Forest	0.845	0.830	0.845	0.838
PFS_Dataset	XGBoost	0.966	0.965	0.966	0.967
PFS_Dataset	Decision Tree	0.915	0.910	0.915	0.920
PFS_Dataset	Random Forest	0.952	0.950	0.952	0.955

For the Mathematics prediction task, the XGBoost model trained on the MFS_Dataset achieved the highest aggregate performance among the configurations evaluated in the study, with an accuracy of 0.958, an F1-score of 0.957, a recall of 0.958, and a precision of 0.960. The corresponding Decision Tree and Random Forest models achieved accuracies of 0.908 and 0.942, respectively. These results indicate a substantial difference in predictive performance between the complete and selected-feature configurations. However, this improvement should be interpreted in conjunction with the complete modelling pipeline, including feature selection, class balancing, and validation procedures.

Table 11: Model Performance Comparison for Mathematics Before and After Feature Selection.

Dataset	Algorithm	Accuracy	F1-Score	Recall	Precision
Math_Dataset	XGBoost	0.775	0.712	0.775	0.699
Math_Dataset	Decision Tree	0.763	0.671	0.763	0.599
Math_Dataset	Random Forest	0.775	0.677	0.775	0.601
MFS_Dataset	XGBoost	0.958	0.957	0.958	0.960
MFS_Dataset	Decision Tree	0.908	0.904	0.908	0.916
MFS_Dataset	Random Forest	0.942	0.941	0.942	0.943

To further examine the behaviour of the selected Mathematics model, I analysed its predictions at the class level. Table 11 presents the precision, recall, and F1-score for each target class, while the confusion matrix provides the corresponding distribution of correct and incorrect predictions.

Per-Class Performance Disaggregation

To evaluate the model's granular effectiveness across all performance trajectories and address potential bias towards the majority class, I disaggregated the performance metrics of the top-performing model (**XGBoost on MFS_Dataset**) across the three individual target classes ($y \in \{0, 1, 2\}$), as presented in Table 12.

Table 12: Per-Class Performance Metrics for the Best Performing Model (XGBoost on MFS_Dataset).

Class	Outcome Category	Precision	Recall	F1-Score	Support (Test Set)
Class 0	**Worsened (At-Risk)**	**0.95**	**0.96**	**0.95**	**15**
Class 1	Equal (Stable)	0.96	0.95	0.96	127
Class 2	Improved	0.97	0.96	0.96	27
Macro Avg / Total		**0.960**	**0.958**	**0.957**	**169**

Confusion Matrix Analysis

The confusion matrix for the XGBoost model evaluated on the MFS_Dataset test partition ($N = 169$) shows that 14 of the 15 Class 0 instances, 121 of the 127 Class 1 instances, and 26 of the 27 Class 2 instances were correctly classified. The errors were distributed across the three classes, with five Class 1 instances classified as Class 2. These results provide additional detail on the class-level behaviour of the selected model and complement the aggregate metrics reported for the Mathematics prediction task.

Table 13: Confusion Matrix for the Best Performing Model (XGBoost on MFS_Dataset Test Partition).

Actual \ Predicted	Class 0 (Worsened)	Class 1 (Equal)	Class 2 (Improved)
Class 0 (Worsened)	**14**	1	0
Class 1 (Equal)	1	**121**	5
Class 2 (Improved)	0	1	**26**

The aggregate results were further assessed using K-Fold cross-validation. As shown in Table 14, the MFS_Dataset achieved an average accuracy of 0.940 with a standard deviation of 0.006, while the PFS_Dataset achieved an average accuracy of 0.966 with a standard deviation of 0.009. In contrast, the Math_Dataset, without feature selection, achieved an average accuracy of 0.745 with a standard deviation of 0.030, and the Port_Dataset achieved an average accuracy of 0.849 with a standard deviation of 0.034.

The relatively small standard deviations observed for the selected-feature datasets indicate comparatively low variability in accuracy across the validation folds. These results provide evidence of consistency across the evaluated folds, but they should not be interpreted, in isolation, as evidence that the models are free from overfitting or that they will generalise to other schools or populations.

Table 14: Mean and Standard Deviation of Accuracy Values obtained through K-Fold Cross-Validation.

Datasets	Average Accuracy Values	Standard Deviation
Port_Dataset	0.849	0.034
PFS_Dataset	0.966	0.009
Math_Dataset	0.745	0.030
MFS_Dataset	0.940	0.006

The results obtained from the selected-feature datasets were interpreted in the context of the complete predictive modelling pipeline. As illustrated in Figure 33, the early fusion approach was used to integrate heterogeneous data sources into a unified representation of the student profile. The original fused dataset

contained 845 student records and 123 attributes and, following data cleaning and preprocessing, resulted in 843 records and 119 features. Feature selection was subsequently used to reduce the dimensionality of this representation and retain a smaller set of features for the predictive models.

For the Mathematics prediction task, the target variable was created by comparing students' first and second term grades. Because the resulting classes were imbalanced, RandomOverSampler was applied to the training data within each cross-validation fold. The balancing procedure resulted in 635 instances per class in the balanced training partitions, corresponding to 1905 instances in total.

Correlation-based feature selection was then used to reduce the feature space. The resulting selected-feature configuration contained 20 features, which were subsequently used in the predictive modelling process.

T

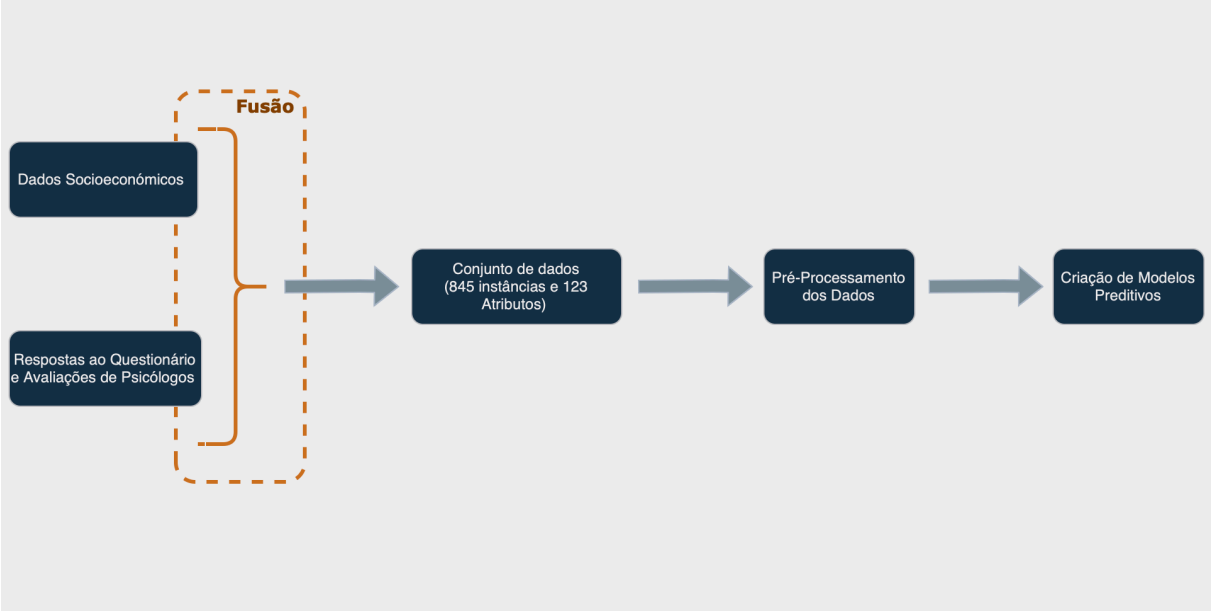


Figure 33: Early Fusion.

Overall, the selected-feature configurations were associated with higher predictive performance than the corresponding complete-feature configurations. For Mathematics, the XGBoost model trained on the MFS_Dataset achieved an accuracy of 95.8%, while the cross-validation analysis yielded an average accuracy of 94.0% with a standard deviation of 0.006. The class-level metrics and confusion matrix complemented these aggregate results by providing a more detailed assessment of the model's behaviour across the three possible academic trajectories.

The results therefore provide evidence of consistent predictive performance across the evaluated cross-validation folds. However, the observed differences between the complete and selected-feature configurations should be interpreted in the context of the full modelling pipeline, as the results do not isolate the individual contribution of early data fusion, feature selection, or class balancing.

5.4.5 Outcomes

Based on my analysis, I concluded that the data fusion technique, namely early fusion, was useful for the task investigated and contributed to the development of competitive predictive models for forecasting variations in student grades. I also found that selecting a small but relevant set of features substantially improved the model evaluation metrics, with the best-performing model achieving an accuracy of 95.8%.

However, several limitations and areas requiring further scrutiny were identified. First, consistency in the definition and interpretation of the target variable is essential. The source code defines Class 0 as representing students whose performance worsened, Class 1 as representing students whose performance remained unchanged, and Class 2 as representing students whose performance improved. This convention should be maintained consistently throughout the study to avoid ambiguity in the interpretation of the prediction results.

Data leakage was also considered an important methodological concern. A data leakage audit was conducted to restrict the predictors to information available at or before the end of the first term. Variables dependent on the second term, attendance information, or other information that would only become available after the prediction point were excluded from the predictor set. Nevertheless, the final list of selected features should be reviewed carefully to ensure that no feature indirectly incorporates information from the prediction period.

The use of RandomOverSampler also represents a methodological consideration. Although the technique was used to address the imbalance between the three target classes, oversampling duplicates observations from the minority classes and may increase the risk of overfitting if repeated instances are incorporated into the modelling process. The cross-validation results showed relatively consistent accuracy across the evaluated folds; however, these results alone do not establish the absence of overfitting. Future work could investigate alternative class-balancing approaches, such as SMOTE or related methods, and evaluate the complete modelling pipeline using an independent external dataset.

The feature selection procedure based on correlation with the target variable also presents limitations. Although the selected-feature configuration was associated with substantially higher predictive performance in the evaluated experiments, correlation-based selection may fail to identify variables whose relationships with the target are non-linear. Consequently, potentially informative features may have been excluded. Future studies could compare this approach with complementary feature selection or feature importance methods, such as Recursive Feature Elimination (RFE), model-based feature importance, L1 regularisation, or mutual information. The robustness of the selected set of 20 features should also be assessed using data from other samples or institutions.

The analysis of the confusion matrix and class-specific metrics provided additional information beyond overall accuracy. This analysis was particularly relevant for Class 0, corresponding to students whose academic performance worsened, because aggregate accuracy alone may not adequately represent the model's ability to identify individual outcome classes. The class-level results therefore complemented the aggregate evaluation by providing a more detailed assessment of the model's predictive behaviour across

the three academic trajectories.

Finally, the generalisation of the results remains a limitation. The study was conducted using data from a single school in the district of Braga. Therefore, further validation of the complete modelling pipeline, from data integration and feature selection to the final predictive model, using data from other schools and regions is necessary before extending the findings to broader educational contexts. Differences in socioeconomic, demographic, and educational characteristics across populations may affect the predictive performance of the developed models.

5.5 Case Study 5: Predictive Analytics in Education: A Comparative Analysis of Machine Learning Models for Predicting Student Performance

5.5.1 Context of the paper

I based this case study on the research published in the paper [43]. Situated within the fields of EDM and Learning Analytics (LA), I explored the application of predictive analytics and machine learning techniques to forecast students' academic performance in core subjects, namely Portuguese and Mathematics. I conducted the research in a school in the municipality of Braga, involving students from grades 5 to 9. I collected data through school records and a 64-item questionnaire, adapted from McInerney's School Motivation Inventory and administered via a chatbot in the classroom environment.

My practical objective was to generate actionable information that could support the anticipation of negative results and inform pedagogical interventions. The final sample consisted of 53 students, as parental authorisations limited the number of participants, with grades coded on a scale from 1 to 5. I performed pre-processing procedures, including category conversion and the creation of numerical variables, applied the RandomOverSampler technique to address class imbalance, divided the data into training and testing sets using an 80/20 split, optimised the model hyperparameters using GridSearch, and validated the models through 5-fold cross-validation.

5.5.2 Motivation

I considered two complementary needs in this case study:

(i) **practical —** identifying at-risk students early on to enable preventive actions (tutoring, individualised support, family/school intervention) and thus improve retention and success;

(ii) **methodological —** assessing the extent to which machine learning models (SVM, ensembles, etc.) can predict grades based on non-academic variables (motivation, family support, participation, tutoring, student perceptions), and which characteristics are most influential.

I also examined the value of explainability techniques, namely SHAP and permutation importance, for transforming predictive results into information that could support pedagogical recommendations. I therefore considered the motivation for this analysis to be both operational, by supporting school decision-making, and scientific, by comparing different models and interpreting the factors associated with their predictions.

5.5.3 Objectives

I defined the objectives of the work as follows:

1) Compare the performance of various classification algorithms (SVM, XGBoost, AdaBoost, Gradient Boost, Random Forest) in predicting Portuguese and Mathematics grades;

2) To evaluate the generalisation gain using cross-validation and GridSearch to optimise hyperparameters;

3) To identify the variables (survey characteristics and family/school context) that most influence predictions, using interpretability techniques (SHAP and permutation feature importance);

4) Propose how the results can be translated into warning signs/local policies for early intervention. These objectives are pursued under the practical constraint of a reduced data set and without using previous grades as predictors (the authors point out that the predictions were based on responses to the questionnaire).

5.5.4 Results

I obtained clear numerical performance metrics and complemented them with an analysis of feature importance and graphical interpretations using techniques such as SHAP and beeswarm graphs. For the prediction of Portuguese grades, I found that the SVM model achieved the best performance, with an accuracy of 0.93, a precision of approximately 0.94, an F1 score of 0.93, a recall of 0.93, and cross-validation (CV) close to 0.92. In comparison, XGBoost achieved an accuracy of 0.89, Random Forest 0.81, Gradient Boost 0.77, and AdaBoost 0.48. For Mathematics, I found that SVM also achieved the highest performance, although with slightly lower values, recording an accuracy of 0.79, a precision of 0.85, an F1 score of 0.78, a recall of 0.79, and a CV of 0.83. The remaining models achieved lower accuracies: XGBoost and Random Forest both reached 0.71, Gradient Boost 0.66, and AdaBoost 0.58.

I interpreted these results in light of the methodological choices adopted in the analysis, which influenced the modelling process. I worked with a relatively small database ($n = 53$), which led me to implement measures to mitigate some of its inherent limitations. To address class imbalance, I applied the RandomOverSampler technique, which replicates instances from minority classes. I also used 5-fold cross-validation to estimate the robustness of the models and GridSearch to optimise the selection of

hyperparameters. Although I obtained promising predictive results, I recognised that the limited sample size represents an important methodological consideration. In this context, I considered the increased risk of overfitting, whereby the models may learn dataset-specific patterns that do not fully reflect broader population trends, potentially reducing their generalisability to other educational settings or future cohorts of students.

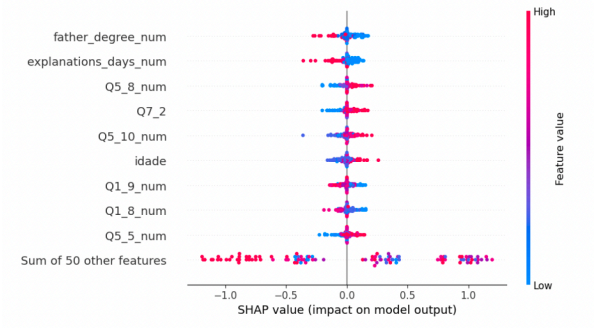
I analysed feature importance using two complementary techniques, SHAP and Permutation Feature Importance. I found partial agreement between the two approaches, while SHAP provided richer information for interpreting the predictions. In particular, I found SHAP to be more stable between runs and useful for visualising not only the magnitude but also the direction of each variable's effect on the prediction. For the prediction of Portuguese grades, I identified as the most influential variables the aggregate sum of characteristics, the father's level of education, the frequency of explanations or tutoring, self-regulation strategies and perseverance, such as the ability to persist in difficult tasks, the habit of reviewing work after completion, the student's age, and indicators of motivation and need for recognition. For the prediction of Mathematics grades, I identified family support with schoolwork, the perception that adults at school listen to students, autonomy and behavioural control, persistence in long-term projects, active participation in class, and the quality of relationships with peers and teachers as the most influential characteristics.

I also identified several consistent and semantically relevant patterns in the results. I found that parental education showed a positive association with better student grades; in the visual representation of SHAP, higher parental education levels were associated with positive SHAP values, indicating a contribution to the prediction of higher grades. I also observed that a higher number of tutoring sessions appeared to be associated with lower grades in some analyses. I interpreted this result cautiously, considering that this variable may function more as a marker of prior difficulties than as a direct cause of lower performance, since students experiencing greater challenges may be more likely to seek additional support. Finally, I found that healthy study behaviours, such as systematic revision of work, persistence in the face of obstacles, and active participation in the classroom, showed a positive and consistent relationship with higher grades. Based on these findings, I considered metacognitive and attitudinal skills to be relevant factors in understanding academic performance (Figure 34).

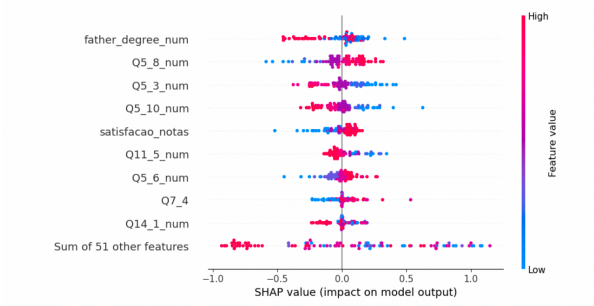
5.5.5 Outcomes

In summary, I found that machine learning models, particularly SVM, provided competitive predictive performance for Portuguese and Mathematics, reaching accuracies of 93% and 79%, respectively, in the dataset analysed. I obtained these results using exclusively non-academic data collected through a questionnaire focused on motivational, socio-relational, and family-context dimensions. I therefore identified useful predictive patterns in these dimensions and identified characteristics such as parental education, frequency of support classes, self-regulation ability, and level of class participation that could potentially support targeted pedagogical interventions.

From a methodological perspective, I found the combined use of model interpretability techniques,



(a) Math



(b) Portuguese

Figure 34: Beeswarm graphics

namely SHAP and permutation importance, to be valuable. I observed that SHAP provided greater consistency and allowed me to examine the direction of each variable's effect, while the permutation-based method exhibited greater variability between runs. However, I interpreted these findings in light of important limitations, including the small sample size of only 53 students, the use of oversampling, which, although necessary to address class imbalance, may have contributed to higher performance estimates, and the fact that the data originated from a single school context. I therefore considered these factors important constraints on the generalisability of the results. Based on this analysis, I identified the need to expand the sample and perform robust external validation of the models before considering their operational implementation in early warning systems in school environments.

5.6 Case Study 6: Mitigating Large Language Models Limitations in Learning Environments through Retrieval-Based Knowledge Integration

5.6.1 Context of the paper

I based this case study on the research published in [20], which is situated within the rapid expansion of LLM applications in education. In this context, I considered the potential of these technologies to support teaching and learning processes through intelligent tutoring, content generation, assessment support, and

formative feedback, while also taking into account the challenges associated with their adoption in formal educational settings. These challenges include hallucinations, static knowledge bases, limited curriculum alignment, ethical concerns, and questions regarding pedagogical reliability.

I positioned this case study at the intersection of AIED and RAG architectures, exploring a technical and pedagogical response to the intrinsic limitations of generalist LLM. I focused specifically on secondary education and on the subject of Mathematics, an area in which conceptual accuracy and logical coherence are particularly important.

Additionally, I addressed a gap identified in the literature concerning the limited number of contextualised empirical studies analysing the application of RAG-based systems in formal school environments, particularly in specific national contexts such as Portugal.

5.6.2 Motivation

I based the motivation for this case study on the observation that LLM, as currently conceived, were not specifically designed for formal educational contexts, which may introduce pedagogical risks when they are used as tutors or learning assistants. In particular, I considered the generation of plausible but incorrect answers, the inability to adhere consistently to official curricula, and the absence of dedicated verification mechanisms as important challenges when applying these technologies to structured subjects such as Mathematics.

At the same time, I considered the growing pressure on teachers to personalise teaching, provide continuous feedback, and support student autonomy in contexts characterised by limited resources and heterogeneous classrooms. In this context, I explored the potential of intelligent tutoring systems to support these needs, provided that they are pedagogically controlled, transparent, and aligned with curricular objectives.

I also considered ethical and institutional aspects, particularly the dependence on cloud services, the transfer of potentially sensitive data, and the opacity of commercial models, which raise concerns regarding privacy, technological sovereignty, and legal compliance, particularly with the GDPR. Based on these considerations, I developed a local, auditable, and replicable solution as a way of addressing these requirements in public educational institutions.

5.6.3 Objectives

I defined the overall objective of this work as the design, implementation, and preliminary evaluation of an intelligent tutoring system for teaching Mathematics at the secondary level, based on the integration of LLM with the RAG approach. Specifically, I developed a curriculum-centred tutoring architecture organised by mathematical domains, namely Algebra, Functions, and Geometry, ensuring alignment with the established programme content. At the same time, I sought to mitigate limitations inherent to LLM, such

as hallucinations and possible knowledge obsolescence, through the use of a RAG system supported by controlled and validated knowledge bases.

Additionally, I designed the system to support user privacy, institutional autonomy, and system transparency by using a fully local infrastructure based on open-source software.

At the pedagogical level, I evaluated students' perceptions of the system's usefulness, its ability to promote autonomy and motivation, and its contribution to learning support. Finally, I sought to contribute empirical evidence to the SOA by providing data on the application of RAG-augmented LLM in Mathematics teaching in a real classroom context.

5.6.4 Results

I conducted the empirical evaluation with 27 tenth-grade students at a Portuguese public school and found an overall positive acceptance of the Mathematics Tutor based on LLM and RAG.

I first characterised the students' previous academic performance in Mathematics. I found that the students generally performed well in the previous school year, with a predominance of level 4 (Figure 35). I considered this profile relevant when interpreting the students' perceptions of the intelligent tutor, as it indicates a group with relatively consolidated conceptual foundations.

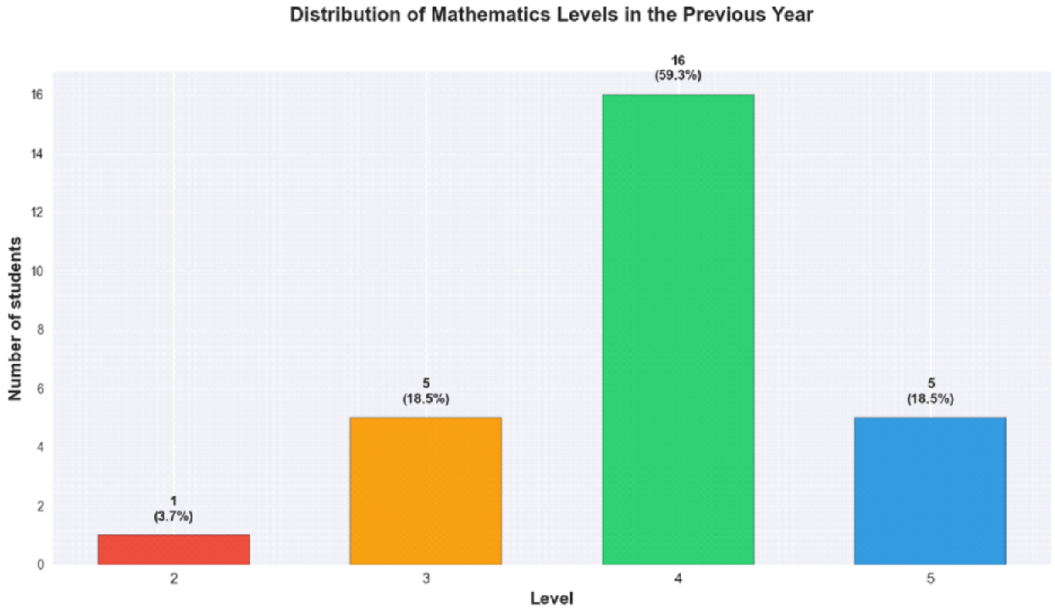


Figure 35: Distribution of Mathematics Levels in the Previous Year.

I then analysed the familiarity reported by the students regarding the different modules of the course. I observed a consistent correspondence between their reported familiarity and their previous academic performance (Figure 36), which I considered relevant when interpreting the perceptions collected through the subsequent instruments.

I also analysed the quantitative results and found a strong perception among students of the educational potential of the resource, reflected in high average values on the agreement scales. I found that the

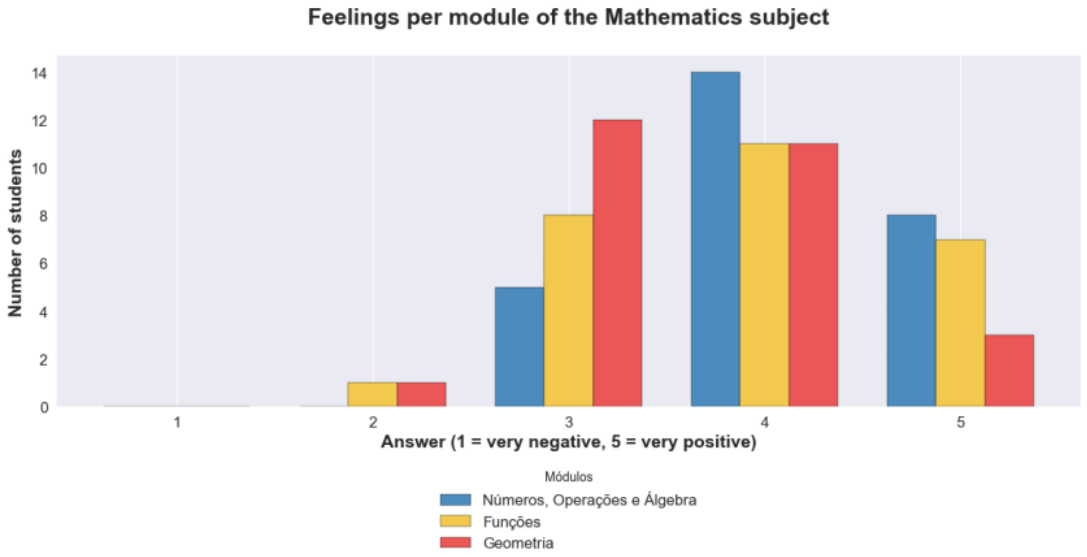


Figure 36: Feelings per module of the Mathematics subject.

system was perceived as contributing to the understanding of content and conceptual consolidation and was particularly valued as a tool for reviewing content and clarifying doubts (Figure 37).

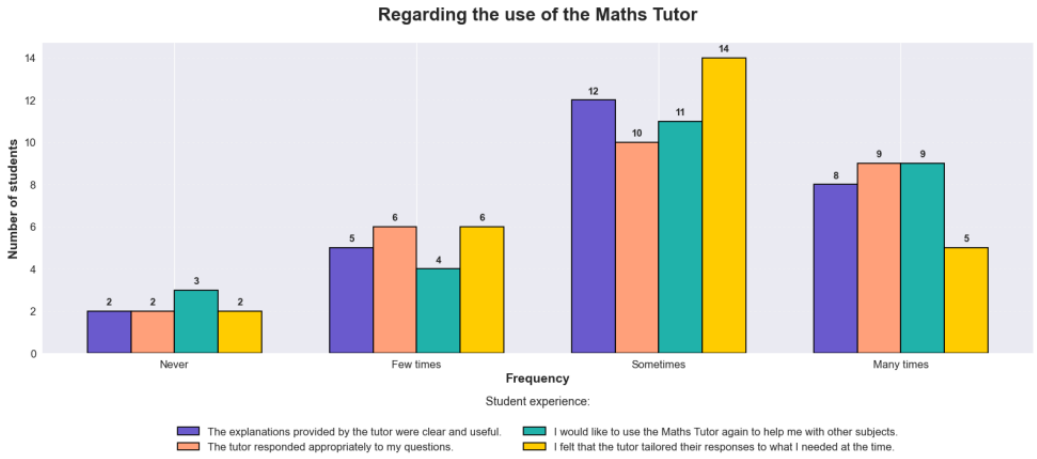


Figure 37: Regarding the use of the Maths Tutor.

I further analysed students' perceptions of autonomy (Figure 38) and found a moderate increase, suggesting that the tutor may support independent study processes without replacing structured pedagogical practices. Regarding confidence and mathematical self-efficacy, I observed a positive but limited effect. Based on this result, I considered that more substantial changes in these psychological dimensions would require more prolonged and curriculum-integrated interventions.

In addition, I conducted a qualitative analysis of the students' feedback and identified a marked appreciation for the clarity of the step-by-step explanations and the fidelity to the official programme.

I also identified criticism mainly related to technical aspects, particularly the latency observed in some system responses. In addition, I identified suggestions for improvement concerning the expansion of the

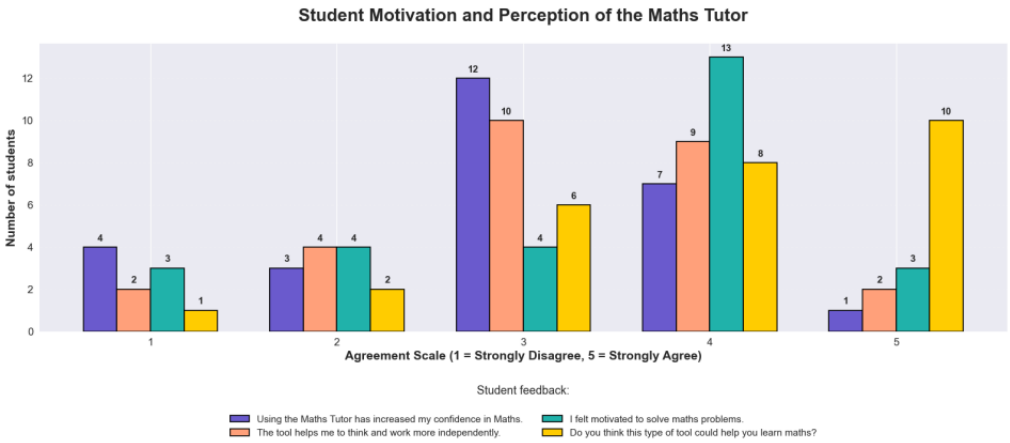


Figure 38: Student Motivation and Perception of the Maths Tutor.

content scope and the adaptation of the pedagogical discourse to different levels of student proficiency.

5.6.5 Outcomes

Based on the results obtained, I found that the integration of LLM with the Retrieval-Augmented Generation approach, when designed and constrained by pedagogical principles, represents a promising strategy for the development of intelligent tutors in secondary school Mathematics education. I demonstrated, within the scope of this proof-of-concept evaluation, the feasibility of implementing a tutor based on RAG and curriculum-oriented knowledge bases to provide contextualised support while reducing reliance on the general knowledge of the underlying LLM. I also incorporated pedagogical alignment, data privacy, and system auditability through the use of a local and modular infrastructure.

I found that the solution was positively perceived by the participating students, particularly for conceptual clarification, content review, and support for independent study. However, I interpret these findings as evidence of perceived usefulness and acceptance rather than as evidence of measured longitudinal learning gains, since the evaluation did not include formal pre- and post-intervention achievement measures or a control group.

I also identified important limitations in the current implementation. In particular, the system does not include formal mathematical verification mechanisms, such as integration with computer algebra systems, which could provide additional guarantees regarding the accuracy of mathematical reasoning. I also identified the need for longitudinal research to evaluate how sustained use of the system may influence students' autonomy, confidence, and self-efficacy over time.

In summary, I established a conceptual, technical, and empirical foundation for the future development of intelligent tutors that are ethical, reliable, and integrated into curricular dynamics. Through this case study, I contributed to the field of AIED by examining the practical application of RAG-augmented LLM in a formal educational context and by providing initial empirical evidence regarding students' perceptions of such a system.

5.7 Conclusions of the Analysis of pratical results

The six case studies presented in this chapter form the empirical core of my research, through which I progressively and complementarily examined the potential of AI and ML as tools for analysis, prediction, and support for pedagogical decision-making in pre-university education.

In the first two studies, I identified statistically significant influences of variables such as gender, age, and family background, both on the choice of academic pathways in Mathematics and on academic performance. I also found that models such as SVM, XGBoost, and Random Forest achieved high predictive performance, with accuracy levels ranging from 0.83 to 0.87.

In the third study, I extended this analysis by incorporating variables related to access to digital resources. I identified an association between technological availability and performance in Mathematics A and observed the robustness of ensemble models across heterogeneous educational contexts.

In the fourth study, I introduced early multimodal data fusion techniques and examined the integration of administrative, family, and psychosocial information. I found that combining these heterogeneous data sources with rigorous attribute selection processes resulted in substantial gains in predictive performance, with accuracy reaching 95.8

Finally, in the sixth study, I moved beyond the strictly predictive paradigm towards the development of intelligent learning support systems. I designed and evaluated an LLM-based tutor using a RAG architecture, aligned with the curriculum and guided by pedagogical principles. I found a positive reception among the participating students and obtained initial evidence supporting the viability of this approach as a response to some of the structural limitations of generalist language models in formal educational contexts.

Taken together, these studies allowed me to establish a coherent research trajectory, progressing from the identification of structural factors associated with academic performance, through predictive modelling and explanatory analysis, to the design of technological solutions aimed at directly supporting learning. Through this research, I obtained evidence supporting the applicability of the proposed methodologies within the Portuguese educational context, while also identifying significant limitations, namely the reliance on contextualised samples, the need for external validation of the models, and the ethical challenges associated with the processing of sensitive data. I consider these limitations important starting points for future research.

The scientific dissemination of this work, as well as the contribution of my research to advancing the state of the art in the fields of AIED, EDM, and Learning Analytics, are presented in the following chapter, where I systematise the articles published in conferences and indexed scientific journals resulting from this research.

6

Conclusion and Future Work

Throughout the work carried out in this doctoral thesis, the aim was to address key challenges associated with the use of Artificial Intelligence, Machine Learning, and the analysis of educational data, with the aim of improving pupils' academic outcomes and promoting more effective, inclusive, and personalised learning experiences. Based on the premise that intelligent technologies can support pedagogical decision-making and contribute to a deeper understanding of the factors influencing academic success, a series of studies was designed to progressively explore different dimensions of this research problem.

Taken together, the studies underpinning this research consistently demonstrate the transformative potential of Artificial Intelligence and educational data analysis in reshaping teaching and learning practices. The results demonstrate that the collection, integration, and analysis of data from multiple sources, combined with advanced modelling and machine learning techniques, enable not only the identification and understanding of patterns of academic performance, but also the anticipation of at-risk situations, the support of early interventions, and the promotion of more informed and equitable pedagogical strategies. These contributions emerge progressively throughout the various case studies conducted, evolving from predominantly analytical and predictive approaches towards the development and validation of intelligent educational support systems, designed to meet the specific needs of students and educational contexts.

At the same time, the findings reinforce the strategic role of educational institutions in fulfilling their social responsibility, particularly by promoting learning environments that are more inclusive, equitable, and aligned with the demands of contemporary society. In this regard, this thesis contributes to the consolidation of a line of research that combines scientific rigour, practical applicability, and ethical sensitivity, demonstrating that technology, when developed in a transparent, explainable, and user-centred manner, can be a valuable tool for improving the quality of education.

This chapter presents a critical analysis of the main contributions of this thesis in light of the objectives initially defined and the research hypothesis formulated at the outset of the work. To this end, it discusses how the results obtained across the various studies have served to validate the proposed hypothesis, highlighting its theoretical, methodological, and practical implications. In addition, a summary is provided of the scientific dissemination activities carried out during the course of the research. The chapter concludes with a final reflection on the work undertaken, highlighting its limitations, the potential

impact of the contributions presented, and the main opportunities for future research.

6.1 Contributions

Overall, the findings presented throughout this thesis provide evidence that the integration of Artificial Intelligence techniques in educational contexts can support more data-informed, personalised, and proactive approaches to learning support. Through the development and analysis of predictive models, intelligent learning environments, and decision-support mechanisms, this research demonstrates the potential of AI-based solutions to assist students and educators in analysing learning needs, monitoring academic progression, and supporting educational decision-making.

Using Mathematics as a representative case study, the research illustrates how AI-based methodologies can support the analysis and prediction of academic performance and grade variation by integrating information from different dimensions of students' educational contexts. The results further indicate the potential value of combining predictive modelling with explainability and decision-support mechanisms to provide information that may complement existing educational practices.

In this context, the findings contribute to addressing the central research hypotheses by providing evidence concerning predictive performance, model explainability, the potential to support educational decision-making, and the perceived usefulness and usability of AI-based learning environments. While the results demonstrate the potential of these approaches within the specific datasets and experimental conditions analysed, further research is required to assess their generalisability and their impact in broader and longitudinal educational settings.

In light of the research hypothesis and research questions presented in Section 1.2, the main contributions of this thesis to these dimensions of analysis are presented below.

R.Q.1: What does the current literature reveal about the application of Artificial Intelligence, Machine Learning, ambient intelligence and intelligent tutoring systems in pre-university education?

Current literature shows that the application of AI, ML, Ambient Intelligence (AmI) and Intelligent Tutoring Systems (ITS) in pre-university education is evolving from approaches centred on the analysis and prediction of academic performance towards solutions geared towards personalised learning and intelligent student support. The case studies analysed demonstrate that these technologies have significant potential to support educational decision-making, identify students at risk of failure and promote more adaptive and inclusive learning environments. A significant part of the literature reviewed focuses on the use of Machine Learning algorithms for predicting academic performance. Case Studies 1, 2, 3, 4, and 5 demonstrate that algorithms such as SVM, XGBoost, and RF can achieve good predictive performance under the specific datasets and experimental conditions analysed. These studies also indicate that academic, family, socio-economic, and behavioural variables can contribute to the prediction and analysis of

students' academic performance. Such information may therefore provide useful support for identifying patterns in students' academic trajectories and for informing evidence-based educational decision-making. However, these results should not be interpreted as establishing direct causal relationships between individual variables and academic outcomes or as demonstrating the effectiveness of interventions based on these predictions.

The literature also highlights a growing interest in personalised learning and in the development of more inclusive educational environments. In this context, Case Study 3 illustrates how AI-based approaches can support the analysis of students' individual characteristics and contextual factors, while Case Study 4 explores the integration of multiple data sources through data fusion techniques to support the prediction of academic performance and grade variation. These findings indicate that combining heterogeneous information sources can provide complementary information for predictive modelling, although the contribution of individual data sources depends on the specific prediction task and experimental context.

With regard to Ambient Intelligence, the studies analysed do not present explicit implementations of fully ubiquitous intelligent educational environments. However, Case Studies 3, 4, and 5 incorporate principles associated with this area, including contextual data collection, the consideration of individual student characteristics, adaptive support, and the use of intelligent systems to assist educational decision-making. These results indicate an alignment between some of the principles explored in this thesis and the broader objectives of Ambient Intelligence in education, while recognising that the work presented does not constitute a full implementation of an ubiquitous AmI infrastructure.

R.Q.2: Who is most significantly affected in terms of academic performance, and how can emerging technologies support the data-driven optimisation of educational resources?

The literature reviewed shows that academic performance is influenced by a range of demographic, family, socio-economic and behavioural factors. Case Studies 1 and 2 highlight that variables such as gender, age and family background affect both academic outcomes and educational choices. In turn, Case Studies 3 and 5 highlight the relevance of factors related to access to digital resources, family context, motivation, self-regulation, and school engagement when analysing students' academic performance. These findings indicate that differences in academic performance should be considered in relation to multiple dimensions of students' educational and social contexts rather than being attributed to a single factor.

Case Study 4 further demonstrates that integrating academic, family, and psychosocial information can provide complementary information for predictive modelling of grade variation. The results therefore support the use of heterogeneous data to analyse patterns associated with changes in academic performance. However, the studies do not establish that these factors independently cause academic outcomes, nor do they provide sufficient evidence to define a general profile of students who are most vulnerable to academic failure across different educational contexts.

With regard to the optimisation of educational resources, Case Studies 1 to 5 demonstrate that AI, ML and Learning Analytics techniques can support the analysis and prediction of academic performance and grade variation. Such predictive information may contribute to evidence-based educational decision-making by helping educators identify patterns that warrant further attention and by providing additional information for the planning of educational support. However, the studies presented in this thesis do not directly evaluate whether the use of these technologies results in more efficient allocation of educational resources.

Furthermore, Case Study 3 illustrates the potential of intelligent systems to support personalised learning by considering individual and contextual characteristics. Case Study 6 further explores the use of Intelligent Tutoring Systems based on Large Language Models and Retrieval-Augmented Generation, demonstrating their potential to provide context-specific support and complement existing educational practices. These results suggest that emerging technologies can provide additional mechanisms for personalised and data-informed learning support, although their effectiveness and impact require further validation.

In summary, the findings indicate that students' academic performance is associated with multiple interacting dimensions, including academic, demographic, family, socio-economic, behavioural, and contextual factors. Case Studies 1, 2, 3, 4, and 5 demonstrate how these dimensions can be incorporated into predictive and analytical models, while Case Studies 3, 4, 5, and 6 illustrate different ways in which emerging technologies can support data-informed educational decision-making and personalised learning support. The extent to which these approaches can contribute to the optimisation of educational resources and improve educational outcomes remains an area for further investigation, particularly through broader and longitudinal studies.

R.Q.3: Which Machine Learning models perform best in classifying and forecasting student academic outcomes within the pre-university education sector?

Analysis of the case studies shows that the machine learning models that perform best in classifying and predicting academic outcomes in pre-university education are the Support Vector Machine (SVM), XGBoost and Random Forest. Case Studies 1, 2 and 3 consistently identify these algorithms as among the most effective in predicting students' academic performance and educational choices.

However, the best overall results were obtained in Case Studies 4 and 5. In Case Study 4, the XGBoost model, combined with data fusion techniques and feature selection, achieved an accuracy of nearly 95.8%, demonstrating high predictive power in identifying variations in academic performance. In turn, in Case Study 5, the SVM performed best in predicting grades in Portuguese and Mathematics, achieving accuracy rates of 93% and 79%, respectively, outperforming models such as Random Forest, Gradient Boost and AdaBoost.

In summary, the empirical evidence from the case studies suggests that SVM and XGBoost are the models that demonstrate the greatest robustness and performance in pre-university education

contexts, whilst Random Forest stands out as a consistent and interpretable alternative. The results also demonstrate that the performance of the models depends not only on the algorithm used, but also on data quality, feature selection and the integration of multiple information sources.

R.Q.4: How can an AI-based ambient intelligent system integrating heterogeneous educational data through early fusion and feature selection techniques support the predictive modelling of grade variation?

The results of Case Studies 4 and 5 provide evidence that integrating heterogeneous educational data through data fusion and feature selection can support the development of predictive models for variations in students' grades. The integration of academic, socio-economic, psychosocial and behavioural data provides a broader representation of students' educational context than the use of isolated data sources. In this context, the combination of heterogeneous information sources can contribute to identifying patterns associated with academic performance and grade variation.

The evidence obtained indicates that data fusion, combined with feature selection, was associated with improved predictive performance in the evaluated configurations. In particular, Case Study 4 achieved an accuracy of 95.8% with XGBoost using the selected feature configuration for Mathematics grade variation. This result indicates that the integration and selection of information from different dimensions of the students' context can provide useful predictive information within the specific dataset and experimental conditions analysed. However, the results do not establish that early fusion alone caused the observed improvement, nor do they demonstrate generalisation to other educational contexts.

Furthermore, Case Studies 4 and 5 indicate that non-academic variables can provide complementary information for analysing students' academic performance and interpreting predictive results. These findings support the potential of AI-based educational systems to complement predictive modelling with information that may be relevant to educational decision-making. However, the extent to which such systems can support interventions, optimise educational resources, or improve students' outcomes requires further validation in broader and longitudinal educational settings.

R.Q.5: When do specific factors most strongly influence performance in core subjects, and how can explainability and permutation-based importance techniques translate these findings into practical teaching recommendations?

Case Studies 1, 2, 4 and 5 demonstrate that the influence of factors associated with academic performance varies depending on the subject, the educational context and the characteristics of the students. Consistently, prior academic factors prove to be the strongest predictors of future performance, but non-academic variables, such as socio-economic background, parents' educational attainment, study habits, self-regulation, motivation and family support, take on increased importance in explaining variations in performance in core subjects such as Mathematics and Portuguese.

The results of Case Studies 4 and 5 indicate that incorporating non-academic variables can affect

predictive performance, although the magnitude and direction of this effect depend on the specific dataset, subject, and model configuration. These findings support the view that academic performance can be analysed through multiple academic and contextual dimensions. However, the results do not establish that individual non-academic factors directly cause differences in academic performance, nor do they by themselves demonstrate the effectiveness of interventions based on these factors.

Furthermore, Case Study 5 demonstrates that explainability techniques, including SHAP and permutation-based methods such as Permutation Feature Importance, can be used to analyse the contribution of variables to model predictions. These techniques provide additional information about which variables are most relevant to the predictions produced by the evaluated models and can therefore improve the interpretability of predictive results.

In this context, explainability methods can support the translation of model outputs into information that may be relevant to teachers and educational decision-makers. They can help identify patterns that warrant further consideration and provide an additional evidence source when analysing students' learning contexts. However, the results of this thesis do not establish a direct or automatic correspondence between feature importance and specific teaching interventions. The development and evaluation of concrete educational recommendations based on these explanations require further validation involving educational practitioners and longitudinal studies.

R.Q.6: Where does the proposed model demonstrate effective performance when evaluated through a case study or pilot project using quantitative and/or qualitative metrics?

The case studies presented in this thesis provide evidence of the performance and applicability of the proposed AI-based approaches in pre-university educational contexts. Case Studies 4 and 5 evaluated predictive models using academic, behavioural, socio-economic, and psychosocial data. Under the specific datasets and experimental conditions analysed, the evaluated models achieved high predictive performance, with the best result reaching an accuracy of 95.8% in Case Study 4. These results demonstrate the ability of the evaluated models to predict academic performance and grade variation within the analysed datasets. They do not, however, establish that the models can reliably identify students at risk of failure across different educational contexts.

Case Study 6 complements the predictive modelling studies through a pilot evaluation of an intelligent Mathematics tutor based on a LLM and RAG. The evaluation combined quantitative indicators related to system usage and user perception with qualitative feedback from students and teachers. The results provide evidence of the perceived usefulness and usability of the tutor within the context of the pilot study and indicate its potential to complement existing learning and teaching practices.

Taken together, the case studies provide complementary evidence concerning predictive performance, system use, and perceived usefulness across different components of the proposed AI-based educational environment. The findings therefore support the applicability of the proposed approaches within the specific educational contexts and experimental conditions investigated. However, the evidence does not

constitute a comprehensive validation of effectiveness across educational settings, nor does it establish a direct impact on students' academic outcomes. Further evaluation with larger and more diverse populations, longitudinal designs, and controlled educational studies would be required to assess the generalisability and educational impact of the proposed approaches.

6.2 Validation of the Research Hypothesis

To assess the extent to which the proposed approach addresses the research objectives, the empirical evidence gathered across the six case studies was systematically analysed in relation to the three research hypotheses defined in Section 1.2. The results provide evidence concerning predictive performance, model explainability and decision support, and the perceived usefulness and usability of the proposed AI-based educational environment. These dimensions are considered complementary, as they address different aspects of the proposed approach, from the identification of academic performance patterns to the interpretation of predictive results and their practical use in an educational setting.

Validation of Hypothesis 1 (Predictive Performance):

The results obtained across Case Studies 1 to 5 provide empirical evidence supporting Hypothesis 1. The case studies explored the application of machine learning algorithms to educational data in order to identify patterns associated with students' academic performance and potential academic risk. In Case Study 4, the combination of early data fusion and feature selection resulted in predictive performance reaching 95.8% accuracy for Mathematics grade variation. The selected features originated from the initially fused datasets, including students' characteristics, grades, and questionnaire responses, supporting the relevance of combining heterogeneous educational information for predictive modelling. In Case Study 5, predictive performance reached 93% when using non-academic behavioural metrics. Taken together, these results indicate that machine learning models can provide useful predictive information regarding academic performance and can identify variables associated with different academic outcomes.

Validation of Hypothesis 2 (Explainability and Decision Support):

Case Studies 4 and 5 provide empirical evidence relevant to Hypothesis 2 through the application of Explainable Artificial Intelligence techniques, namely SHAP and Permutation Feature Importance. These techniques were used to complement predictive performance with an interpretation of the contribution of individual features to the model outputs. Rather than treating the predictive models as purely black-box systems, the explainability analysis provides information about the variables associated with the generated predictions. This information can support the interpretation of academic risk factors and provide educators with more transparent evidence to consider when analysing students' academic situations and planning pedagogical interventions. The results therefore support the role of explainability as a complementary component of predictive modelling and as a potential mechanism for informing educational decision-making.

Validation of Hypothesis 3 (Perceived Usefulness and Usability):

Hypothesis 3 was evaluated through Case Study 6, which involved a formative assessment activity with 27 students from the 10th grade and a pilot evaluation with a Mathematics teacher. The study collected perceptions concerning the Mathematics Tutor, including its usefulness, usability, clarity, contribution to learning support, and potential for integration into teaching practices. The teacher pilot provided an additional perspective on the pedagogical adequacy and practical usefulness of the environment, while the student questionnaire captured the users' perceptions following interaction with the tutor. This evidence provides an empirical basis for assessing the perceived acceptance, clarity, usability, and pedagogical usefulness of the proposed AI-based Mathematics Tutor. However, these results should be interpreted as evidence of perceived utility and usability rather than as evidence of measured improvements in students' academic achievement.

Boundary Conditions and Scope of the Validation:

It is important to emphasize that while this thesis validated predictive accuracy, feature explainability, and user perception of utility, the longitudinal measurement of direct academic learning gains (e.g., standardized pre/post achievement tests with a control group) was beyond the scope of this initial proof-of-concept and remains formally designated for Future Work. Consequently, the results should not be interpreted as demonstrating that the proposed AI-based approach directly improves students' grades or learning outcomes. Instead, the empirical evidence establishes the predictive, interpretative, and perceived-utility dimensions of the proposed framework, while direct measurement of learning gains remains an important direction for future longitudinal evaluation.

Overall, the six case studies provide complementary evidence for the three research hypotheses. The predictive studies demonstrate the potential of machine learning and educational data fusion for identifying academic performance patterns and associated factors; the Explainable Artificial Intelligence analyses provide interpretable information that can complement these predictions; and the formative assessment and pilot evaluation provide evidence regarding the perceived usefulness and usability of the Mathematics Tutor. Together, these findings support the proposed framework while maintaining a clear distinction between predictive capability, explainability, perceived educational utility, and the direct measurement of learning gains.

6.3 Dissemination of Results

The dissemination of research findings constitutes a fundamental dimension of scientific inquiry, ensuring that the knowledge produced reaches and informs both the academic community and broader society. In this context, the present section outlines the dissemination strategy adopted throughout this doctoral research, encompassing peer-reviewed publications, presentations at international scientific conferences, and other forms of academic and professional communication, with the dual purpose of maximising the visibility and impact of the results obtained and contributing to the advancement of the field.

In this regard, the research conducted within this doctoral thesis has resulted in a set of scientific publications that directly reflect the empirical work developed across the case studies presented in the previous chapter, as detailed below:

- "Survey for smart and adaptative education" published at Proceedings of the 2022 5th International Conference on Education Technology Management.
- "Future of Education on Pre-University" published at the International Conference on Interactive Collaborative Learning 2023.
- "Forecasting and AI-Powered Exploration of Student choices across differents types of Mathematics in Pre-University School" published at the International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning 2024.
- "AI-driven educational transformation in secondary schools: Leveraging data insights for inclusive learning environments" published at the IEEE Global Engineering Education Conference (EDUCON) 2024.
- "Predictive Analytics in Education: A Comparative Analysis of Machine Learning Models for Predicting Student Performance" published at the EPIA Conference on Artificial Intelligence 2024.
- "Data Fusion and Predictive Modeling for Academic Performance Assessment: A Case Study on Grade Variation" from the Conference WorldCist 2025.
- "Mitigating LLM Limitations in Learning Environments through Retrieval-Based Knowledge Integration" published at the CAI 2026 IEEE Conference on Artificial Intelligence 2026.

It also includes participation in:

- Participation in Advanced School on Artificial Intelligence conference (EAIA 2024), held in Viana do Castelo, Portugal, from 2 to 3 September 2024.
- Participation in the development of an educational platform within the scope of the AMALIA project, including contributions to the logical and functional design of the platform, particularly regarding the teacher–class–student interaction framework. This work also involved the development of an intelligent tutoring component capable of adapting conversational interactions according to the student's personality traits, aiming to provide a more personalized and effective learning experience.

6.4 Final Remark and Future Work Considerations

In conclusion, the present doctoral research contributes to the growing body of knowledge at the intersection of Artificial Intelligence, Machine Learning and Educational Data Mining by exploring how data-driven

and intelligent technologies can support learning processes and academic success in pre-university education. Through the six Case studies developed throughout this thesis, it was possible to demonstrate not only the scientific relevance of predictive modelling, multimodal data fusion and intelligent tutoring systems, but also their potential practical value for educators, school leaders and policy-makers committed to improving educational quality and promoting more equitable learning opportunities.

Although this research was conducted within specific institutional and geographic contexts, and some limitations remain, namely those related to sample size, contextual variability and the use of self-reported psychosocial data, the findings obtained provide a consistent and meaningful basis for future developments in this field. Future research may benefit from extending the proposed models to broader educational settings, incorporating longitudinal approaches capable of capturing the evolution of student behaviour and performance over time, and further refining AI-based tutoring systems through larger pedagogical implementations and collaborative design processes involving teachers and students.

Additionally, the ethical implications of AI adoption in education, particularly regarding privacy, transparency, explainability and algorithmic fairness, should remain a central concern in subsequent investigations. Although the proposed architecture demonstrates the potential of AI, LLM and predictive analytics to support personalised education, technological progress should not overshadow the human dimension of teaching. Future educational systems should therefore be designed not only to maximise predictive performance and personalisation but also to preserve students' autonomy, critical thinking, privacy and responsible decision-making. In this context, AI should be regarded as a technology that complements rather than replaces teachers, who remain essential as pedagogical, ethical and social mediators throughout the educational process.

Ultimately, it is hoped that the work presented in this thesis may contribute, even modestly, to the development of more personalised, inclusive and evidence-informed educational systems, where emerging technologies are used not merely as tools for prediction, but as instruments to empower students, support teachers and strengthen educational communities. Such systems should promote not only academic achievement but also the development of critical thinking, informed judgement and responsible participation in an increasingly digital society. In this way, technological innovation can remain firmly aligned with the fundamental purpose of education: fostering human development through trustworthy, explainable and human-centred AI.

Bibliography

[1] M. I. Abdin et al. “Phi-4 Technical Report”. In: *arXiv preprint arXiv:2412.08905* (2024) (cit. on pp. 33, 34).

[2] G. S. Abu-Oda and A. M. El-Halees. “Data mining in higher education: university student dropout case study”. In: *International Journal of Data Mining & Knowledge Management Process* 5.1 (2015), p. 15 (cit. on p. 67).

[3] H. Abu-Rasheed et al. “Supporting student decisions on learning recommendations: An llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring”. In: *arXiv preprint arXiv:2401.08517* (2024) (cit. on pp. 85, 86).

[4] D. B. Acharya, K. Kuppan, and B. Divya. “Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey”. In: *IEEe Access* 13 (2025), pp. 18912–18936 (cit. on p. 42).

[5] J. Achiam et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023) (cit. on p. 33).

[6] S. Alagarsamy, V. Nagarajan, and M. Y. Devi. “OMIBC: optimal modified identity-based cryptography for signcryption and private key extraction using fuzzy model”. In: *Wireless Networks* 30.4 (2024), pp. 2159–2172 (cit. on p. 65).

[7] S. Alelyani. “Stable bagging feature selection on medical data”. In: *Journal of Big Data* 8.1 (2021), pp. 1–18 (cit. on p. 21).

[8] M. Alier et al. “LAMB: An open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems”. In: *Computer Standards & Interfaces* 92 (2025), p. 103940 (cit. on pp. 80, 81).

[9] M. Aly. “Survey on multiclass classification methods”. In: *Neural Netw* 19.1-9 (2005), p. 2 (cit. on p. 16).

[10] K. Arulkumaran et al. “A brief survey of deep reinforcement learning”. In: *arXiv preprint arXiv:1708.05866* (2017) (cit. on p. 30).

[11] Y. Awasthi. “Agentic AI Redefined: A New Paradigm in Artificial Intelligence”. In: *Journal of Software Engineering and Simulation* (2025) (cit. on p. 42).

[12] T. O. Ayodele. “Types of machine learning algorithms”. In: *New advances in machine learning* 3 (2010), pp. 19–48 (cit. on pp. 18, 28).

[13] J. Bai et al. “Qwen Technical Report”. In: *arXiv preprint arXiv:2309.16609* (2023) (cit. on pp. 33, 34).

[14] R. Bajaj and V. Sharma. “Smart Education with artificial intelligence based determination of learning styles”. In: *Procedia computer science* 132 (2018), pp. 834–842 (cit. on pp. 1, 3, 44–46, 50).

[15] D. Bañeres et al. “An early warning system to detect at-risk students in online higher education”. In: *Applied Sciences* 10.13 (2020), p. 4427 (cit. on p. 67).

[16] D. Batory. “Feature models, grammars, and propositional formulas”. In: *International Conference on Software Product Lines*. Springer. 2005, pp. 7–20 (cit. on p. 73).

[17] E. M. Bender et al. “On the dangers of stochastic parrots: Can language models be too big?” In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 610–623 (cit. on pp. 31, 38, 42).

[18] R. Bezerra, D. Durães, and P. Novais. “Forecasting and AI-Powered Exploration of Student Choices Across Differents Types of Mathematics in Pre-university School”. In: *International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning*. Springer. 2024, pp. 5–16 (cit. on p. 122).

[19] R. Bezerra et al. “Future of Education on Pre-university Studies”. In: *International Conference on Interactive Collaborative Learning*. Springer. 2023, pp. 278–289 (cit. on p. 117).

[20] R. Bezerra et al. “Mitigating LLM Limitations in Learning Environments through Retrieval-Based Knowledge Integration”. In: *2026 IEEE Conference on Artificial Intelligence (CAI)*. IEEE. 2026, pp. 1342–1347 (cit. on p. 142).

[21] Y. Bian et al. “Ensemble pruning based on objection maximization with a general distributed framework”. In: *IEEE transactions on neural networks and learning systems* 31.9 (2019), pp. 3766–3774 (cit. on p. 22).

[22] R. Bommasani. “On the opportunities and risks of foundation models”. In: *arXiv preprint arXiv:2108.07258* (2021) (cit. on pp. 31, 32, 41, 42).

[23] L. Breiman. “Arcing classifiers”. In: *Annals of Statistics* 26 (1996), pp. 123–40 (cit. on pp. 21, 25).

[24] T. Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901 (cit. on pp. 31–33, 37).

[25] M. Cebral-Loureda and C. Torres-Huitzil. “Neural Deep Learning Models for Learning Analytics in a Digital Humanities Laboratory”. In: *2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop*. IEEE. 2021, pp. 1–8 (cit. on pp. 45, 117).

[26] L. Cen et al. “Augmented immersive reality (AIR) for improved learning performance: a quantitative evaluation”. In: *IEEE Transactions on Learning Technologies* 13.2 (2019), pp. 283–296 (cit. on p. 45).

[27] R. Chacón et al. “Development of IoT applications in civil engineering classrooms using mobile devices”. In: *Computer Applications in Engineering Education* 26.5 (2018), pp. 1769–1781 (cit. on p. 1).

[28] L. Chandra, S. Seidel, and S. Gregor. “Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions”. In: *2015 48th hawaii international conference on system sciences*. IEEE. 2015, pp. 4039–4048 (cit. on p. 7).

[29] B. Charbuty and A. Abdulazeez. “Classification based on decision tree algorithm for machine learning”. In: *Journal of applied science and technology trends* 2.01 (2021), pp. 20–28 (cit. on p. 16).

[30] H. Chase. “Langchain, october 2022”. In: *URL https://github. com/langchain-ai/langchain* (2022) (cit. on p. 87).

[31] M. Chassignol et al. “Artificial Intelligence trends in education: a narrative overview”. In: *Procedia Computer Science* 136 (2018), pp. 16–24 (cit. on p. 2).

[32] J. Chen et al. “Design and research on big data’s smart ideological and political classroom in colleges and universities in the new era”. In: *DEStech Transactions on Materials Science Engineering, no. ameme* (2020) (cit. on pp. 49, 50).

[33] T. Chen and C. Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794 (cit. on pp. 23, 27).

[34] R. Y. Choi et al. “Introduction to machine learning, neural networks, and deep learning”. In: *Translational vision science & technology* 9.2 (2020), pp. 14–14 (cit. on pp. 28, 29).

[35] N. Chondamrongkul, G. Hristov, and P. Temdee. “Addressing Technical Challenges in Large Language Model-Driven Educational Software System”. In: *IEEE Access* (2025) (cit. on p. 40).

[36] Z. Chu et al. “Llm agents for education: Advances and applications”. In: *arXiv preprint arXiv:2503.11733* (2025) (cit. on pp. 77, 78).

[37] D. R. Cotton, P. A. Cotton, and J. R. Shipway. “Chatting and cheating: Ensuring academic integrity in the era of ChatGPT”. In: *Innovations in education and teaching international* 61.2 (2024), pp. 228–239 (cit. on p. 32).

[38] V. Devedžić. “Web intelligence and artificial intelligence in education”. In: *Journal of Educational Technology & Society* 7.4 (2004), pp. 29–39 (cit. on p. 2).

[39] J. Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186 (cit. on pp. 32, 37).

[40] P. Dharanyadevi et al. “Context aware SOA based discovery in mobile environment by bi-matchmaking technique”. In: *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*. IEEE. 2011, pp. 618–623 (cit. on p. 65).

[41] S. Doroudi. “The bias-variance tradeoff: How data science can inform educational debates”. In: *AERA open* 6.4 (2020), p. 2332858420977208 (cit. on p. 21).

[42] D. Duraes, R. Bezerra, and P. Novais. “AI-driven educational transformation in secondary schools: Leveraging data insights for inclusive learning environments”. In: *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE. 2024, pp. 1–9 (cit. on p. 127).

[43] D. Durães et al. “Predictive Analytics in Education: A Comparative Analysis of Machine Learning Models for Predicting Student Performance”. In: *EPIA Conference on Artificial Intelligence*. Springer. 2024, pp. 141–153 (cit. on p. 139).

[44] D. Durães et al. “Data Fusion and Predictive Modeling for Academic Performance Assessment: A Case Study on Grade Variation”. In: *World Conference on Information Systems and Technologies*. Springer. 2025, pp. 21–32 (cit. on p. 131).

[45] U. Durrani et al. “Assessing the Effectiveness of Large Language Models in Predicting Student Dropout Rates”. In: *2024 International Conference on Artificial Intelligence, Metaverse and Cybersecurity (ICAMAC)*. IEEE. 2024, pp. 1–6 (cit. on p. 67).

[46] Y. K. Dwivedi et al. “Agentic AI Systems: What It Is and Isn’t”. In: *Global Business and Organizational Excellence* 45.3 (2026), pp. 253–263 (cit. on p. 42).

[47] K. Fawagreh, M. M. Gaber, and E. Elyan. “Random forests: from early developments to recent advancements”. In: *Systems Science & Control Engineering: An Open Access Journal* 2.1 (2014), pp. 602–609 (cit. on p. 22).

[48] M. Fernández-Delgado et al. “An extensive experimental survey of regression methods”. In: *Neural Networks* 111 (2019), pp. 11–34 (cit. on p. 14).

[49] N. Fleming. “Teaching and Learning Styles: VARK Strategies. Honolulu Community College”. In: *of Publication* (2001) (cit. on p. 46).

[50] Y. Freund, R. Schapire, and N. Abe. “A short introduction to boosting”. In: *Journal-Japanese Society For Artificial Intelligence* 14.771-780 (1999), p. 1612 (cit. on p. 23).

[51] Y. Freund and R. E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139 (cit. on p. 23).

[52] J. H. Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232 (cit. on p. 25).

[53] K. Fu. “[Retracted] A Research on the Realization Algorithm of Internet of Things Function for Smart Education”. In: *Computational Intelligence and Neuroscience* 2022.1 (2022), p. 1330190 (cit. on pp. 2, 3, 45).

[54] S. C. E. Fung, M. F. Wong, and C. W. Tan. “Automatic Feedback Generation on K-12 Students’ Data Science Education by Prompting Cloud-based Large Language Models”. In: *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 2024, pp. 255–258 (cit. on p. 87).

[55] J. Gao, M. Galley, and L. Li. “Neural approaches to conversational AI”. In: *The 41st international ACM SIGIR conference on research & development in information retrieval*. 2018, pp. 1371–1374 (cit. on pp. 39, 42).

[56] Y. Gao et al. “Retrieval-augmented generation for large language models: A survey”. In: *arXiv preprint arXiv:2312.10997* 2.1 (2023) (cit. on pp. 40, 77).

[57] E. J. Gong et al. “The Potential Clinical Utility of the Customized Large Language Model in Gastroenterology: A Pilot Study”. In: *Bioengineering* 12.1 (2024), p. 1 (cit. on p. 40).

[58] A. Grattafiori et al. “The llama 3 herd of models”. In: *arXiv preprint arXiv:2407.21783* (2024) (cit. on p. 34).

[59] J. Guo et al. “An AI-application-oriented in-class teaching evaluation model by using statistical modeling and ensemble learning”. In: *Sensors* 21.1 (2021), p. 241 (cit. on pp. 45, 52, 55, 56, 58).

[60] D. L. Hall and J. Llinas. “An introduction to multisensor data fusion”. In: *Proceedings of the IEEE* 85.1 (2002), pp. 6–23 (cit. on p. 41).

[61] O. Hamal et al. “Artificial intelligent in education”. In: *Sustainability* 14.5 (2022), p. 2862 (cit. on p. 54).

[62] J. Han. *M. Kamber in J. Pei, Data mining: Concepts and techniques: Concepts and techniques, 3. izd.* 2011 (cit. on pp. 13, 14, 21).

[63] J. Han, J. Pei, and H. Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022 (cit. on p. 13).

[64] T. Händler. “A Taxonomy for Autonomous LLM-Powered Multi-Agent Architectures.” In: *KMIS*. 2023, pp. 85–98 (cit. on pp. 71–73).

[65] K. Hannah et al. *OECD reviews of evaluation and assessment in education: Student assessment in Turkey*. OECD Publishing, 2019 (cit. on p. 1).

[66] A. R. Hevner et al. "Design science in information systems research1". In: *MIS quarterly* 28.1 (2004), pp. 75–106 (cit. on pp. 6, 7).

[67] P. Honey and A. Mumford. "Styles of learning". In: *Gower Handb. Manag. Dev* 101 (1994), pp. 101–111 (cit. on p. 46).

[68] Z. Hu, F. Zhao, and X. Zhao. "Research on the Design of Smart Classroom Mode from the Perspective of Big Data". In: *2021 4th International Conference on E-Business, Information Management and Computer Science*. 2021, pp. 102–110 (cit. on pp. 47–50, 52, 55).

[69] X. Huang et al. "Understanding the planning of LLM agents: A survey". In: *arXiv preprint arXiv:2402.02716* (2024) (cit. on p. 77).

[70] L. Hughes et al. "AI agents and agentic systems: A multi-expert analysis". In: *Journal of Computer Information Systems* 65.4 (2025), pp. 489–517 (cit. on p. 42).

[71] H. Iqbal et al. "Smart educational tools and learning management systems: supportive framework". In: *International Journal on Interactive Design and Manufacturing (IJIDeM)* 14.4 (2020), pp. 1179–1193 (cit. on p. 1).

[72] M. J. Islam et al. "Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers". In: *2007 international conference on convergence information technology (ICCIT 2007)*. IEEE. 2007, pp. 1541–1546 (cit. on p. 17).

[73] G. Izacard and E. Grave. "Leveraging passage retrieval with generative models for open domain question answering". In: *arXiv preprint arXiv:2007.01282* (2020) (cit. on p. 39).

[74] J. Jeon and S. Lee. "Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT". In: *Education and Information Technologies* 28.12 (2023), pp. 15873–15892 (cit. on p. 31).

[75] Z. Ji et al. "Survey of hallucination in natural language generation". In: *ACM computing surveys* 55.12 (2023), pp. 1–38 (cit. on p. 31).

[76] A. Q. Jiang et al. "Mixtral of experts". In: *arXiv preprint arXiv:2401.04088* (2024) (cit. on p. 33).

[77] J. Jiang et al. "Boosting tree-assisted multitask deep learning for small scientific datasets". In: *Journal of chemical information and modeling* 60.3 (2020), pp. 1235–1244 (cit. on p. 27).

[78] B. T. Johnson and E. A. Hennessy. "Systematic reviews and meta-analyses in the health sciences: Best practice methods for research syntheses". In: *Social science & medicine* 233 (2019), pp. 237–251 (cit. on p. 9).

[79] J. Johnson, M. Douze, and H. Jégou. "Billion-scale similarity search with GPUs". In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547 (cit. on p. 40).

[80] D. Jones and S. Gregor. “The anatomy of a design theory”. In: *Journal of the Association for Information Systems* 8.5 (2007), p. 1 (cit. on p. 7).

[81] J. R. B. Junior and M. do Carmo Nicoletti. “An iterative boosting-based ensemble for streaming data classification”. In: *Information Fusion* 45 (2019), pp. 66–78 (cit. on p. 23).

[82] D. Jurafsky and J. H. Martin. “Speech and Language Processing: An Introduction to Natural 890 Language Processing”. In: *Computational Linguistics, and Speech Recognition with Language Models. 3rd* 891 (2024), p. 892 (cit. on pp. 36, 39).

[83] L. P. Kaelbling, M. L. Littman, and A. W. Moore. “Reinforcement learning: A survey”. In: *Journal of artificial intelligence research* 4 (1996), pp. 237–285 (cit. on p. 29).

[84] H. T. Kahraman, S. Sagiroglu, and I. Colak. “Development of adaptive and intelligent web-based educational systems”. In: *2010 4th international conference on application of information and communication technologies*. IEEE. 2010, pp. 1–5 (cit. on p. 2).

[85] V. Karpukhin et al. “Dense Passage Retrieval for Open-Domain Question Answering.” In: *EMNLP (1)*. 2020, pp. 6769–6781 (cit. on p. 39).

[86] E. Kasneci et al. “ChatGPT for good? On opportunities and challenges of large language models for education”. In: *Learning and individual differences* 103 (2023), p. 102274 (cit. on pp. 31, 32).

[87] S. E. Kelly, D. Moher, and T. J. Clifford. “Quality of conduct and reporting in rapid reviews: an exploration of compliance with PRISMA and AMSTAR guidelines”. In: *Systematic reviews* 5.1 (2016), p. 79 (cit. on p. 9).

[88] I. A. Khan. “Electronic learning management system: Relevance, challenges and preparedness”. In: *J. Emerg. Technol. Innov. Res* 7.5 (2020), pp. 471–480 (cit. on p. 2).

[89] U. Khandelwal et al. “Generalization through memorization: Nearest neighbor language models”. In: *arXiv preprint arXiv:1911.00172* (2019) (cit. on p. 41).

[90] G. Kneebone. “Probability and the Weighing of Evidence. By IJ Good.(London: Charles Griffin and Company. 1950. Pp. viii+ 119. Price 16s.)” In: *Philosophy* 26.97 (1951), pp. 163–164 (cit. on p. 17).

[91] G. Knezek and R. Christensen. “The Will, Skill, Tool Model of Technology Integration: Adding Pedagogy as a New Model Construct.” In: *International Association for Development of the Information Society* (2015) (cit. on p. 68).

[92] T. Kojima et al. “Large language models are zero-shot reasoners”. In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213 (cit. on p. 71).

[93] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, et al. “Supervised machine learning: A review of classification techniques”. In: *Emerging artificial intelligence applications in computer engineering* 160.1 (2007), pp. 3–24 (cit. on pp. 15–17, 28).

[94] M. Kuhn, K. Johnson, et al. *Applied predictive modeling*. Vol. 26. Springer, 2013 (cit. on p. 23).

[95] P. Kumar, M. Haresh, and V. Hayagreevan. "Development of Interactive Assistance for Academic Preparation Using Large Models Language". In: *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*. IEEE. 2025, pp. 265–269 (cit. on pp. 65, 66).

[96] K. Kumpas-Lenk, E. Eisenschmidt, and A. Veispak. "Does the design of learning outcomes matter from students' perspective?" In: *Studies in Educational Evaluation* 59 (2018), pp. 179–186 (cit. on pp. 51, 52, 55).

[97] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444 (cit. on p. 32).

[98] P. Lewis et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474 (cit. on pp. 36, 38, 39, 41).

[99] A. Lewkowycz et al. "Solving quantitative reasoning problems with language models". In: *Advances in neural information processing systems* 35 (2022), pp. 3843–3857 (cit. on p. 33).

[100] Q.-F. Li and Z.-M. Song. "High-performance concrete strength prediction based on ensemble learning". In: *Construction and Building Materials* 324 (2022), p. 126694 (cit. on p. 23).

[101] X. Li. "A review of prominent paradigms for llm-based agents: Tool use, planning (including rag), and feedback learning". In: *Proceedings of the 31st international conference on computational linguistics*. 2025, pp. 9760–9779 (cit. on p. 42).

[102] Y. Li and W. Chen. "A comparative performance assessment of ensemble learning for credit scoring". In: *Mathematics* 8.10 (2020), p. 1756 (cit. on pp. 22, 27).

[103] Z. Li et al. "Retrieval-augmented generation for educational application: A systematic survey". In: *Computers and Education: Artificial Intelligence* (2025), p. 100417 (cit. on pp. 40, 61).

[104] H.-Y. Liang et al. "Effect of an AI-based chatbot on students' learning performance in alternate reality game-based museum learning". In: *British Journal of Educational Technology* 55.5 (2024), pp. 2315–2338 (cit. on p. 31).

[105] Y. Liu et al. "Adaptive quantitative trading: An imitative deep reinforcement learning approach". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 02. 2020, pp. 2128–2135 (cit. on p. 30).

[106] C. K. Lo. "What is the impact of ChatGPT on education? A rapid review of the literature". In: *Education sciences* 13.4 (2023), p. 410 (cit. on p. 32).

[107] A. C. Lorena et al. "Comparing machine learning classifiers in potential distribution modelling". In: *Expert Systems with Applications* 38.5 (2011), pp. 5268–5275 (cit. on p. 17).

[108] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. url: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. ii).

[109] R. W. C. Lui et al. "GPTutor: A Generative AI-powered Intelligent Tutoring System to Support Interactive Learning with Knowledge-Grounded Question Answering". In: *2024 International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*. IEEE. 2024, pp. 702–707 (cit. on p. 78).

[110] S. Maghsudi et al. "Personalized education in the artificial intelligence era: what to expect next". In: *IEEE Signal Processing Magazine* 38.3 (2021), pp. 37–50 (cit. on p. 45).

[111] B. Mahesh. "Machine learning algorithms-a review". In: *International Journal of Science and Research (IJSR).[Internet]* 9.1 (2020), pp. 381–386 (cit. on pp. 16, 17).

[112] A. Mallen et al. "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories". In: *arXiv preprint arXiv:2212.10511* (2022) (cit. on p. 31).

[113] A. McCallum. "Information extraction: Distilling structured data from unstructured text". In: *Queue* 3.9 (2005), pp. 48–57 (cit. on p. 13).

[114] I. D. Mienye and Y. Sun. "A survey of ensemble learning: Concepts, algorithms, applications, and prospects". In: *IEEE Access* 10 (2022), pp. 99129–99149 (cit. on pp. 21, 22, 24, 26).

[115] I. D. Mienye, Y. Sun, and Z. Wang. "Improved predictive sparse decomposition method with densenet for prediction of lung cancer". In: *Int. J. Comput* 1 (2020), pp. 533–541 (cit. on pp. 21–23).

[116] I. D. Mienye et al. "Enhanced prediction of chronic kidney disease using feature selection and boosted classifiers". In: *International Conference on Intelligent Systems Design and Applications*. Springer. 2021, pp. 527–537 (cit. on p. 23).

[117] B. Min et al. "Recent advances in natural language processing via large pre-trained language models: A survey". In: *ACM Computing Surveys* 56.2 (2023), pp. 1–40 (cit. on p. 30).

[118] S. Mishra et al. "Improving the accuracy of ensemble machine learning classification models using a novel bit-fusion algorithm for healthcare AI systems". In: *Frontiers in Public Health* 10 (2022), p. 858282 (cit. on p. 21).

[119] V. Mnih et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533 (cit. on p. 30).

[120] M. Mohammed, M. B. Khan, and E. B. M. Bashier. *Machine learning: algorithms and applications*. Crc Press, 2016 (cit. on p. 21).

[121] D. Moher et al. "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement". In: *Bmj* 339 (2009) (cit. on p. 9).

[122] S. Murugesan. “The rise of agentic AI: implications, concerns, and the path forward”. In: *IEEE Intelligent Systems* 40.2 (2025), pp. 8–14 (cit. on p. 42).

[123] A. Natekin and A. Knoll. “Gradient boosting machines, a tutorial”. In: *Frontiers in neurorobotics* 7 (2013), p. 21 (cit. on pp. 25, 27).

[124] A. T. Neumann et al. “An llm-driven chatbot in higher education for databases and information systems”. In: *IEEE Transactions on Education* (2024) (cit. on p. 40).

[125] I. K. Nti, A. F. Adekoya, and B. A. Weyori. “A comprehensive evaluation of ensemble learning for stock-market prediction”. In: *Journal of Big Data* 7.1 (2020), pp. 1–40 (cit. on p. 22).

[126] F. Osisanwo et al. “Supervised machine learning algorithms: classification and comparison”. In: *International Journal of Computer Trends and Technology (IJCTT)* 48.3 (2017), pp. 128–138 (cit. on p. 15).

[127] L. Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744 (cit. on pp. 31, 32).

[128] M. E. Ozturk et al. “Distributed BERT pre-training & fine-tuning with intel optimized TensorFlow on Intel Xeon scalable processors”. In: *Science* 11.5 (2017), pp. 746–761 (cit. on p. 30).

[129] B. Pampel, S. Martin, and U. Padó. “Regaining Control: Enabling Educators to Build Specialized AI Chat Bots with Retrieval Augmented Generation”. In: (2025) (cit. on p. 67).

[130] K. Peffers, T. Tuunanen, and B. Niehaves. *Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research*. 2018 (cit. on p. 7).

[131] K. Peffers et al. “A design science research methodology for information systems research”. In: *Journal of management information systems* 24.3 (2007), pp. 45–77 (cit. on p. 6).

[132] H. Peng, S. Ma, and J. M. Spector. “Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment”. In: *Smart learning environments* 6.1 (2019), pp. 1–14 (cit. on p. 30).

[133] R. Peredo et al. “Intelligent Web-based education system for adaptive learning”. In: *Expert Systems with Applications* 38.12 (2011), pp. 14690–14702 (cit. on p. 2).

[134] F. Pernkopf. “Bayesian network classifiers versus selective k-NN classifier”. In: *Pattern recognition* 38.1 (2005), pp. 1–10 (cit. on p. 17).

[135] M. Phan, A. De Caigny, and K. Coussement. “A decision support framework to incorporate textual data for early student dropout prediction in higher education”. In: *Decision Support Systems* 168 (2023), p. 113940 (cit. on p. 67).

[136] E. Prihar et al. “Comparing different approaches to generating mathematics explanations using large language models”. In: *International Conference on Artificial Intelligence in Education*. Springer. 2023, pp. 290–295 (cit. on p. 31).

[137] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014 (cit. on p. 29).

[138] S. A. Qureshi et al. “Improving depression level estimation by concurrently learning emotion intensity”. In: *IEEE Computational Intelligence Magazine* 15.3 (2020), pp. 47–59 (cit. on p. 45).

[139] A. Radford et al. “Improving language understanding by generative pre-training”. In: (2018) (cit. on pp. 30, 32).

[140] A. Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9 (cit. on p. 31).

[141] V. Radosavljevic, S. Radosavljevic, and G. Jelic. “Ambient intelligence-based smart classroom model”. In: *Interactive Learning Environments* 30.2 (2022), pp. 307–321 (cit. on pp. 45, 51, 52, 54).

[142] M. J. Rahaman. “A comprehensive review to understand the definitions, advantages, disadvantages and applications of machine learning algorithms”. In: *Int J Comput Appl* 186.31 (2024), pp. 43–47 (cit. on p. 12).

[143] C. Ramos, J. C. Augusto, and D. Shapiro. “Ambient intelligence—the next step for artificial intelligence”. In: *IEEE Intelligent Systems* 23.2 (2008), pp. 15–18 (cit. on pp. 46, 47, 50).

[144] M. L. Rethlefsen et al. “PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews”. In: *Systematic reviews* 10.1 (2021), p. 39 (cit. on p. 9).

[145] M. H. D. M. Ribeiro and L. dos Santos Coelho. “Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series”. In: *Applied soft computing* 86 (2020), p. 105837 (cit. on p. 22).

[146] M. Sabzevari, G. Martínez-Muñoz, and A. Suárez. “Building heterogeneous ensembles by pooling homogeneous ensembles”. In: *International Journal of Machine Learning and Cybernetics* (2022), pp. 1–8 (cit. on p. 22).

[147] K. Sanderson. “GPT-4 is here: what scientists think”. In: *Nature* 615.7954 (2023), p. 773 (cit. on p. 31).

[148] R. Sapkota, K. I. Roumeliotis, and M. Karkee. “Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges”. In: *Information Fusion* (2025), p. 103599 (cit. on p. 42).

[149] I. H. Sarker et al. “Cybersecurity data science: an overview from machine learning perspective”. In: *Journal of Big data* 7 (2020), pp. 1–29 (cit. on pp. 13, 21).

[150] R. E. Schapire. “The strength of weak learnability”. In: *Machine learning* 5 (1990), pp. 197–227 (cit. on p. 23).

[151] T. Schick et al. “Toolformer: Language models can teach themselves to use tools”. In: *Advances in neural information processing systems* 36 (2023), pp. 68539–68551 (cit. on pp. 31, 33).

[152] P.-Y. Schobbens et al. “Generic semantics of feature diagrams”. In: *Computer networks* 51.2 (2007), pp. 456–479 (cit. on p. 73).

[153] B. A. Shawar and E. Atwell. “Chatbots: are they really useful?” In: *Journal for Language Technology and Computational Linguistics* 22.1 (2007), pp. 29–49 (cit. on p. 36).

[154] A. Singh, N. Thakur, and A. Sharma. “A review of supervised machine learning algorithms”. In: *2016 3rd international conference on computing for sustainable global development (INDIACom)*. Ieee. 2016, pp. 1310–1315 (cit. on pp. 14, 15, 19).

[155] P. Sneharika et al. “A Comprehensive Analysis on Traffic Prediction Methods for Real-World Deployment: Challenges, Cause and Scope”. In: *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*. IEEE. 2024, pp. 119–124 (cit. on p. 65).

[156] A. A. Soofi and A. Awan. “Classification techniques in machine learning: applications and issues”. In: *J. Basic Appl. Sci* 13 (2017), pp. 459–465 (cit. on pp. 13, 21).

[157] H. Stefanidi et al. “The LECTOR Podium. An Innovative Teacher Workstation for the Intelligent Classroom of the Future”. In: *2020 12th International Conference on Education Technology and Computers*. 2020, pp. 126–132 (cit. on pp. 45, 46, 50).

[158] L. N. Steimle, D. L. Kaufman, and B. T. Denton. “Multi-model Markov decision processes”. In: *Iise Transactions* 53.10 (2021), pp. 1124–1139 (cit. on p. 30).

[159] Y. Sun et al. “Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction”. In: *Applied Artificial Intelligence* 35.4 (2021), pp. 290–303 (cit. on pp. 21, 23).

[160] K. Sundar et al. “Revolutionizing Assessment: AI-Powered Evaluation with RAG and LLM Technologies”. In: *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*. IEEE. 2024, pp. 43–48 (cit. on pp. 69, 70).

[161] B. Tabarsi et al. “MerryQuery: A Trustworthy LLM-Powered Tool Providing Personalized Support for Educators and Students”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 28. 2025, pp. 29700–29702 (cit. on pp. 81, 83).

[162] G. Al-Tameemi et al. “A hybrid machine learning approach for predicting student performance using multi-class educational datasets”. In: *Procedia Computer Science* 238 (2024), pp. 888–895 (cit. on p. 67).

[163] K. Taneja et al. “Jill watson: A virtual teaching assistant powered by chatgpt”. In: *International Conference on Artificial Intelligence in Education*. Springer. 2024, pp. 324–337 (cit. on p. 40).

[164] M. Tavakoli et al. “An AI-based open recommender system for personalized labor market driven education”. In: *Advanced Engineering Informatics* 52 (2022), p. 101508 (cit. on pp. 45, 56, 58).

[165] G. Team et al. “Gemma 2: Improving open language models at a practical size”. In: *arXiv preprint arXiv:2408.00118* (2024) (cit. on p. 34).

[166] T. Tuunanen, R. Winter, and J. v. Brocke. “Dealing with complexity in design science research: A methodology using design echelons”. In: *MIS quarterly* 48.2 (2024), pp. 427–458 (cit. on p. 8).

[167] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999 (cit. on p. 16).

[168] A. Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017) (cit. on pp. 30, 32, 37, 41).

[169] O. Vinyals et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning”. In: *nature* 575.7782 (2019), pp. 350–354 (cit. on p. 30).

[170] F. Wang et al. “Feature learning viewpoint of AdaBoost and a new algorithm”. In: *IEEE Access* 7 (2019), pp. 149890–149899 (cit. on p. 23).

[171] L. Wang et al. “A survey on large language model based autonomous agents”. In: *Frontiers of Computer Science* 18.6 (2024), p. 186345 (cit. on p. 42).

[172] T. Wang et al. “LLM-powered Multi-agent Framework for Goal-oriented Learning in Intelligent Tutoring System”. In: *arXiv preprint arXiv:2501.15749* (2025) (cit. on pp. 74, 76).

[173] J. Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837 (cit. on pp. 33, 71, 74).

[174] J. Weizenbaum. “ELIZA—a computer program for the study of natural language communication between man and machine”. In: *Communications of the ACM* 9.1 (1966), pp. 36–45 (cit. on p. 37).

[175] R. Winter. “Design science research in Europe”. In: *European Journal of Information Systems* 17.5 (2008), pp. 470–475 (cit. on p. 6).

[176] P. Wu and H. Zhao. “Some analysis and research of the AdaBoost algorithm”. In: *International Conference on Intelligent Computing and Information Science*. Springer. 2011, pp. 1–5 (cit. on p. 24).

[177] Y. Wu et al. “A sequential matching framework for multi-turn response selection in retrieval-based chatbots”. In: *Computational Linguistics* 45.1 (2019), pp. 163–197 (cit. on p. 37).

[178] B. Xu. “AI Agent Systems: Architectures, Applications, and Evaluation”. In: *arXiv preprint arXiv:2601.01743* (2026) (cit. on p. 42).

[179] Y. Xu and X. Jian. “Generative Artificial Intelligence Empowers the Research of Digital Textbooks in Vocational Education”. In: *Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education*. 2024, pp. 802–806 (cit. on pp. 62, 63).

[180] Y. Xue and Y. Wang. “Artificial intelligence for education and teaching”. In: *Wireless Communications and Mobile Computing* 2022 (2022), pp. 1–10 (cit. on pp. 56, 58).

[181] A. Yang et al. “Qwen3 technical report”. In: *arXiv preprint arXiv:2505.09388* (2025) (cit. on p. 33).

[182] J. Yang et al. “Evaluation of smart classroom from the perspective of infusing technology into pedagogy”. In: *Smart Learning Environments* 5.1 (2018), pp. 1–11 (cit. on pp. 50, 51).

[183] S. Yao et al. “React: Synergizing reasoning and acting in language models”. In: *The eleventh international conference on learning representations*. 2022 (cit. on p. 33).

[184] A. R. Yuliantoputri, W. Muhamad, and S. Suhardi. “Smart classroom services system design based on services computing system”. In: *2019 International Conference on ICT for Smart Society (ICISS)*. Vol. 7. IEEE. 2019, pp. 1–6 (cit. on p. 45).

[185] K. R. Žalik. “An efficient k-means clustering algorithm”. In: *Pattern recognition letters* 29.9 (2008), pp. 1385–1391 (cit. on p. 20).

[186] O. Zawacki-Richter et al. “Systematic review of research on artificial intelligence applications in higher education–where are the educators?” In: *International journal of educational technology in higher education* 16.1 (2019), p. 39 (cit. on p. 32).

[187] B. Zhang et al. “Health data driven on continuous blood pressure prediction based on gradient boosting decision tree algorithm”. In: *IEEE Access* 7 (2019), pp. 32423–32433 (cit. on p. 27).

[188] N. Zhang and J. Xie. “Construction Of an Auxiliary System for Smart Online Courses Based on Shared Screen Technology and Face Verification”. In: *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE. 2022, pp. 1256–1259 (cit. on pp. 53–55).

[189] Y. Zhang et al. “Interactive Smart Educational System Using AI for Students in the Higher Education Platform.” In: *Journal of Multiple-Valued Logic & Soft Computing* 36 (2021) (cit. on pp. 2, 3, 44).

[190] Z. Zhang et al. “A survey on the memory mechanism of large language model based agents”. In: *arXiv preprint arXiv:2404.13501* (2024) (cit. on p. 77).

[191] Z. Zhao et al. “Embodied AI-guided interactive digital teachers for education”. In: *SIGGRAPH Asia 2024 Educator’s Forum*. 2024, pp. 1–8 (cit. on pp. 88, 89).

A

Publications Composing the Doctor Thesis

A.1 Survey for smart and adaptative education

Authors: Bezerra, R. M.; Durães, D.; Novais, P. N.

Author contributions: The author was responsible for the literature review, conceptualization, methodology, analysis, interpretation of results, discussion, and writing of the original manuscript.

Title: Survey for smart and adaptative education.

Date: December, 2022

Conference In Proceedings of the 2022: 5th International Conference on Education Technology Management.

Pages: 237-242

Summary: This article presents a SLR (2018-2022) on the applications of AI, IoT and smart environments in the educational context. The analysis was guided by three questions: the existence of intelligent and adaptive systems, the integration of affective and student profiles, and strategies already used to predict academic success.

The results show that the studies found are distributed among: intelligent adaptive learning, smart classrooms, the use of AI and ML in education, neural networks, and algorithms to support smart environments. A clear trend is the integration of sensors and IoT-connected devices in the classroom, capable of monitoring student attendance, emotions, and behavior. Among the most common techniques are smart tutors, personalized content recommendations, gamification strategies, and ML models to support performance assessment and prediction. Some studies also explore biometric and neurocognitive data to infer attention and emotional states. Despite the opportunities identified, many of these systems remain in the prototype phase. Limitations persist regarding the quality of recommendations, the lack of large-scale empirical validation, and the ethical integration of sensitive data. The conclusion highlights that the future of AI in education depends on more applied research.

A.2 Future of Education on Pre-university Studies

Authors: Bezerra, R; Lacerda, B.; Durães, D.; Novais, P.
Author contributions: The author contributed to the literature review, formulation of the research questions, interpretation of results, discussion, and writing of the original manuscript.
Title: Future of Education on Pre-university Studies.
Date: September, 2023
Conference: In International Conference on Interactive Collaborative Learning.
Book title: Towards a Hybrid, Flexible and Socially Engaged Higher Education
Volume: 900
Pages: 278-289.
Publisher: Springer Nature Switzerland.
ISBN: 978-3-031-52666-4
978-3-031-52667-1
URL: `https://link.springer.com/10.1007/978-3-031-52667-1_28`
DOI: `https://doi.org/10.1007/978-3-031-52667-1_28`
Summary: This article presents a SLR and an empirical study examining determinants of academic performance and major choices in high school, with an emphasis on gender, socioeconomic background, and parental education. The empirical analysis uses a dataset from a Guimarães school with 625 students (aged 14–20), including grades by subject, course options/electives, and family and demographic variables.
The results show consistent patterns: the "science and technology" areas are gender balanced, while "arts" and "humanities" have more female students and "socioeconomic" more male students. Overall, female students achieve higher GPAs than male students in almost all areas and subjects analyzed. In the study of electives, a female preference for Biology and a male preference for Chemistry/Computer Science were observed.
In the methodology section, the article describes the data preprocessing and compared several classifiers: SVM achieved the best accuracy (0.87); AdaBoost performed very poorly. Removing the "gender" attribute barely changed the metrics, suggesting that, in this set, the direct predictive contribution of gender is small, although observable differences persist.
The conclusions advocate for more practical and inclusive pre-university education and point to interventions to increase female representation in STEAM. The article highlights limitations (single sample, need for large-scale validation) and recommends future research on subject choice trajectories and factors explaining female underrepresentation in engineering.

A.3 Forecasting and AI-Powered Exploration of Student Choices Across Differents Types of Mathematics in Pre-university School

Authors: Bezerra, R; Durães, D.; Novais, P.

Author contributions: The author was responsible for the conceptualization, methodology, analysis, interpretation of results, discussion, and writing of the original manuscript.

Title: Forecasting and AI-Powered Exploration of Student Choices Across Differents Types of Mathematics in Pre-university School

Date: June 2024

Conference: MIS4TEL2024

Book title: International Conference in Methodologies and intelligent Systems for Techhnology Enhanced Learning

Pages: 5–16

Publisher: Springer Nature Switzerland

ISBN: 978-3-031-85560-3

URL: `https://link.springer.com/chapter/10.1007/978-3-031-85561-0_1#citeas`

DOI: `https://doi.org/10.1007/978-3-031-85561-0_1`

Summary: The article investigates the factors that influence student choice and performance in different areas of mathematics in high school, combining statistical analysis and ML techniques in two school groups (Guimarães and Braga), with approximately 625 students in each dataset. The study includes academic, demographic, and contextual variables.

The central findings found that gender is associated with patterns of choice and performance—female students, on average, score higher than male students in several subjects and dominate the top spots in the MACS, while the presence of mathematics options varies by course. Age and appropriate grade level also correlate with better grades. In predictive terms, classical models for tabular data were tested: SVM showed the best performance in some configurations; AdaBoost performed very poorly in the evaluated sets. Excluding the gender variable slightly reduced accuracy, indicating that gender contributes to the prediction, but is not the sole determinant. Parental education/occupation showed a limited direct effect on grades, although it is related to the choice of mathematics streams (schools with more educated parents have more students in Math A).

The study concluded that the results support the importance of including sociodemographic and contextual variables in predictive models, highlight gender segregation in curriculum choices, and suggest interventions. They also highlight the need for larger studies to validate the generalizability of the results.

A.4 AI-Driven Educational Transformation in Secondary Schools: Leveraging Data Insights for Inclusive Learning Environments

Authors: Durães, D.; Bezerra, R; Novais, P.

Author contributions: The author contributed to the writing of the manuscript, presentation and analysis of results, and discussion.

Title: AI-Driven Educational Transformation in Secondary Schools: Leveraging Data Insights for Inclusive Learning Environments.

Date: may 2024

Conference: 2024 IEEE Global Engineering Education Conference (EDUCON).

Book title: 2024 IEEE Global Engineering Education Conference (EDUCON).

Pages: 1-9.

Publisher: IEEE.

ISBN: 979-8-3503-9402-3

URL: `https://ieeexplore.ieee.org/document/10578670`

DOI: `10.1109/EDUCON60312.2024.10578910`

Summary: This paper explores the potential of AI in personalizing secondary education by analyzing school data from 621 students at a school in Braga, divided into three Mathematics branches (Math A: 520; Math B: 20; MACS: 81). The paper combined academic information (grades from the three periods) with sociodemographic and contextual variables (gender, age, parental education, computer/internet access, transportation, among others). It applied statistical tests and compared several ML models, adjusting hyperparameters with GridSearch and using cross-validation. The results show that the ensemble models (XGBoost/AdaBoost) achieved the best accuracy in several scenarios, although the SVM also proved robust in some configurations. The chi-square test revealed a significant association between grades and internet/computer access in Math A and a consistent pattern of better performance among female students. It was also found that sets composed solely of non-academic variables performed poorly, indicating that multiple sources of information are necessary for reliable predictions. The authors highlight limitations and advocate for governance, interpretability, and validation on a broader basis before using these models for pedagogical interventions—proposing, as next steps, greater integration of deep learning techniques, interpretable methods, and use oriented toward personalized education.

A.5 Predictive Analytics in Education: A Comparative Analysis of ML Models for Predicting Student Performance

Authors: Durães, D.; Lacerda, B.; Bezerra, R; Novais, P.

Author contributions: The author contributed to the methodology and writing of the original manuscript.

Title: Predictive Analytics in Education: A Comparative Analysis of ML Models for Predicting Student Performance.

Date: setembro 2024

Conference: EPIA Conference on Artificial Intelligence.

Book title: Progress in Artificial Intelligence

Volume: 14967

Pages: 141-153.

Publisher: Springer Nature Switzerland.

ISBN: 978-3-031-73496-0 and 978-3-031-73497-7

URL: `https://link.springer.com/10.1007/978-3-031-73497-7_12`

DOI: `https://doi.org/10.1007/978-3-031-73497-7_12`

Summary: This article analyzes the application of ML models to predict grades for 5th- to 9th-grade students in Portuguese and Math, based on a 64-question questionnaire administered via chatbot. The sample included 53 students (aged 10-14), whose grades were classified into five levels. Preprocessing involved response encoding, data balancing, splitting into training and test sets (80/20), fine-tuning hyperparameters with GridSearch, and five-fold cross-validation. Five models were tested (SVM, XGBoost, AdaBoost, Gradient Boost, and Random Forest). The SVM model achieved the best performance (93% accuracy in predicting grades in Portuguese and 79% in Math) and demonstrated potential for early identification of students at risk of failing. To interpret the most determining factors, they used techniques such as SHAP and permutation feature importance, revealing that parental education, especially that of the father, the frequency of tutoring, family and teacher support, study habits such as reviewing work and persisting with projects, as well as satisfaction with grades, had the greatest influence on the predictions.

While the results are promising for supporting educational interventions and guiding monitoring policies, the article highlights limitations such as the small sample size and the need for oversampling, recommending future studies with larger datasets to validate and expand the conclusions.

A.6 Data Fusion and Predictive Modeling for Academic Performance Assessment: A Case Study on Grade Variation

Authors: Durães, D.; Bezerra, R; Novais, P.

Author contributions: The author contributed to the writing of the manuscript and interpretation of the results.

Title: Data Fusion and Predictive Modeling for Academic Performance Assessment: A Case Study on Grade Variation.

Date: march 2025

Conference: WorldCist 2025.

Book title: Emerging Trends in Informatio Systems and Technologies.

Volume: 1514

Pages: 21-32.

Publisher: Springer Nature Switzerland.

ISBN: 978-3-031-97118-1 and 978-3-031-97119-8

URL: `https://link.springer.com/chapter/10.1007/978-3-031-97119-8_2`

DOI: `https://doi.org/10.1007/978-3-031-97119-8_22`

Summary: This article describes a case study that applies data fusion and ML models to predict variation in student grades between the first and second semesters, using academic and sociodemographic data, as well as responses to a self-assessment and academic support questionnaire collected from 845 primary school students in the Braga region. The combined dataset resulted in approximately 843 instances and 119 attributes. A ternary encoding was created for grade increase, stability, or decrease (improvement = 2, equal = 1, worsening = 0). A strong class imbalance was identified, which was corrected by random oversampling. After cleaning, coding, and removing outliers, they used correlation-based variable selection to reduce the feature space to 20 relevant features (combining grades, questionnaire responses, and personal variables). They compared several algorithms, adjusting hyperparameters and evaluating them with cross-validation. The results show that early fusion followed by feature selection significantly improved prediction: in the reduced dataset (MFS_dataset), XGBoost achieved 95.8% accuracy and Random Forest achieved 94.2%, while in the complete datasets, accuracies were substantially lower (e.g., 77.5% for XGBoost in Math_dataset). The average K-Fold metrics corroborate the robustness and low variance in the selected versions. The article concludes that multimodal fusion is effective for predicting grade variations.

A.7 Mitigating LLM Limitations in Learning Environments through Retrieval-Based Knowledge Integration

Authors: Bezerra, R; Adelino; Marcondes F.; Durães, D.; Novais, P.

Author contributions: The author was responsible for the conceptualization, methodology, analysis, interpretation of results, discussion, and writing of the original manuscript. The development of the Mathematics Tutor was carried out with the assistance of the co-authors.

Title: Mitigating LLM Limitations in Learning Environments through Retrieval-Based Knowledge Integration.

Date: 2026

Conference: IEEECAI2026.

Book title: 2026 IEEE Conference on Artificial Intelligence (CAI).

Pages: 1342–1347.

Publisher: IEEE.

ISBN: 979-8-3315-6039-3 and 979-8-3315-6040-9

URL: `https://ieeexplore.ieee.org/document/11536367`

DOI: `10.1109/CAI68641.2026.11536367`

Summary: The article examines the limitations of LLM in education, namely hallucinations, static knowledge, and lack of curriculum control, proposing the integration of RAG to increase the accuracy, timeliness, and reliability of responses. The development and preliminary evaluation of an intelligent mathematics tutor for secondary education is presented, organized by domains (Algebra, Functions, and Geometry) and based on a modular architecture with RAG agents. The system is structured in three layers (interface, application logic, and knowledge management) and operates entirely on local infrastructure (Ollama and ChromaDB), ensuring privacy and compliance with the GDPR. The RAG pipeline restricts the generation of responses to previously indexed curriculum content, mitigating responses outside the program scope.

The empirical evaluation involved 27 students from 10th grade from a public school in Braga, with pre- and post-intervention questionnaires. The results indicate an overall positive perception, with a significant impact on content comprehension and consolidation, a moderate increase in autonomy and motivation, and a less significant effect on self-confidence. Most students recognized the potential of intelligent tutors as a complementary resource to traditional teaching.

The scientific contribution lies in the proposal of a curriculum-driven, auditable, and replicable architecture, offering empirical evidence of the application of LLM with RAG in mathematics teaching. Among the limitations identified are the absence of a formal mathematical verification module and the small sample size. As future work, we propose the integration of formal symbolic validation mechanisms, incremental updating of the knowledge base, and expansion of the empirical study.