Web Mining on E-learning

Francisco José Casanova Faria Campos

UMinho | 2022

Universidade do Minho

Escola de Engenharia

Francisco José Casanova Faria Campos

Web Mining on E-learning

June 2022

Universidade do Minho
Escola de Engenharia

Francisco José Casanova Faria Campos

Web Mining on E-learning

Masters Dissertation
Information Systems Engineering and Management
Integrated Master

Work developed with guidance from:
Professor Manuel Filipe Vieira Torres dos Santos
Professor Carlos Filipe da Silva Portela

June 2022

Direitos de autor

Agradecimentos

Findando o desenvolvimento desta dissertação, queria aproveitar para agradecer a todas as pessoas na minha vida que permitiram chegar a este ponto e partilharam esta experiência comigo.

Primeiramente, agradecer ao Professor Doutor Carlos Filipe da Silva Portela, meu coorientador, por toda a ajuda ao longo deste processo e total disponibilidade ao longo destes anos para transmitir conhecimento que levarei para o resto da minha vida.

Ao Professor Doutor Manuel Filipe Santos, meu orientador, por me ter disponibilizado a possibilidade de realização da minha dissertação de mestrado.

Aos meus amigos João, Hugo, Nuno, Francisco, Pedro e todos os meus colegas, pelos momentos inesquecíveis que passamos e acima de tudo, pelas amizades que sei que levarei para a vida.

Um grande obrigado à Universidade do Minho e a todos os seus docentes, que me proporcionaram todas as condições para o meu crescimento e sucesso enquanto estudante.

Um agradecimento muito especial para o meu irmão, o meu pai, os meus avós maternos e toda a minha família que ao longo destes anos foi sendo o meu suporte para que eu conseguisse atingir todos os meus objetivos.

E acima de tudo, o maior obrigado vai para a minha mãe que, apesar de não estar presente entre nós, foi a pessoa que mais acreditou em mim e permitiu que tivesse todas as condições possíveis para lutar e chegar a esta posição.

Declaração de integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio, nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

Atualmente, a utilização das tecnologias de informação é vital no dia a dia das pessoas e, sobretudo, das empresas. Esta utilização e troca de informação gera uma quantidade de dados que podem criar valor para vários setores da sociedade, caso seja possível traduzir a informação neles contida.

Se este problema for adotado ao setor da educação, muitas questões e problemas podem ser resolvidas com o correto tratamento desses dados. A possibilidade de perceber os comportamentos dos alunos e as metodologias de estudo que permitem um maior sucesso escolar é certamente uma oportunidade tentadora e que pode levar à otimização do ensino como é conhecido e permitir formar melhores pessoas e profissionais

Assim sendo, o foco deste projeto centra-se no tratamento destes dados gerados no contexto universitário, a partir da criação de uma solução que consegue receber diferentes tipos de dados, aplicando modelos analíticos, de modo a gerar relatórios e dashboards sobre a realidade dos dados em estudo, e modelos preditivos, de modo a poder prever futuras notas de alunos com base no seu comportamento académico.

Em termos analíticos foi possível comprovar que existe uma relação forte entre um grande nível de assiduidade e participação nas aulas com um bom desempenho académico, que 77% dos alunos foram capazes de tirar uma nota final acima de 15 valores, que todos os alunos estiveram presentes, pelo menos, num terço das aulas, além de provar o sucesso de ferramentas aplicadas previamente como o sistema de cartões e resgate, com 8 alunos, que primeiramente tinham reprovado num momento de avaliação crítico, a terem possibilidade de completar o curso graças a este último mecanismo. Em termos preditivos, o protótipo revelou ser eficaz, principalmente em termos de regressão, com um erro absoluto de 1,10 valores.

Os dados para este projeto foram fornecidos pela empresa IOTech, que os reuniu a partir da utilização dos alunos de diferentes plataformas na unidade curricular “Programação Web” durante o ano letivo 2020/2021. Este trabalho está também inserido no projeto IOScience, levado a cabo pela mesma empresa.

Palavras-Chave: dados estruturados, dados não estruturados, bases de dados NoSQL, *Educational Data Mining,*, *Web Mining*

Abstract

Nowadays, the use of information technologies is vital in people's lives and companies. This use and exchange of information generate a quantity of data that can create value for various sectors of society if possible to translate the information contained therein.

Adopting this problem to the education sector, many issues and problems can be solved with the correct treatment of this data. The possibility of perceiving the students' behaviours and study methodologies that allow greater school success is certainly a tempting opportunity and that can lead to the optimization of teaching as it is known and allow to create better people and professionals.

Therefore, the focus of this project focuses on the processing of these data generated in university context, by creating of a solution that can receive different data types, apply analytical models, in order to generate reports and dashboards about the reality of the data In study, and predictive models, so that it can predict future grades of students based on their academic behaviour.

In analytical terms it was possible to prove that there is a strong relationship between a great level of attendance and participation in classes with a good academic performance, that 77% of the students were able to take a final grade above 15 values, that all students were present at least one-third of the classes, in addition to proving the success of previously applied tools such as the card and rescue system, with 8 students, who had first failed at a critical evaluation moment, to be able to complete the course thanks to the latter mechanism. In predictive terms, the prototype proved to be effective, mainly in terms of regression, with an absolute error of 1.10 values.

The data for this project were provided by the company IOTech, which gathered it from the use of students from different platforms in the "Web Programming" curriculum during the 2020/2021 school year. This work is also inserted in the IOScience project, carried out by the same company.

Keywords: structured data, unstructured data, NoSQL databases, *Educational Data Mining, Web Mining*

Index

Figure Index

Table Index

Acronyms

ACID – Atomicity, Consistency, Isolation, Durability

BASE – Basic Availability, Soft state, Eventual consistency

BigD – Big Data

CRISP-DM – Cross Industry Standard Process for Data Mining

DB – Database

DM – Data Mining

DSR – Design Science Research

DW – Data Warehouse

EDM – Educational Data Mining

ETL – Extract, Transform and Load

NoSQL – Not only Structured Query Language

RDBMS – Relational Database Management System

SQL - Structured Query Language

TM – Text Mining

WCM – Web Content Mining

WM – Web Mining

WSM – Web Structure Mining

WUM – Web Usage Mining

Chapter 1 – Introduction

This first chapter has the objective of introducing this dissertation by giving it context and motivation. Furthermore, the main goals, along with some structuring ones, are presented turning clearer what this project is about. Then the present document structure is presented to give some guidance over it.

1. Context and Motivation

In today's world, the quantity of data produced and available is bigger than at any point in time. This event is called "Big Data". The main importance of Big Data consists in the potential to improve efficiency using a large volume of data of several types. If Big Data is used the right way, organizations can get a better view of their business and be more successful in different areas (Alsghaier et al., 2017). According to Sowmya & Suneetha (2017), "For modern industry, data generated by machines and devices, cloud-based solutions, business management has reached a total volume of more than 1000 Exabytes annually and is expected to increase 20-fold in the next ten years". This, by itself, should not be a problem, however, a large amount of this data is not used properly or lost along the way. To solve this problem, new tools, technologies, and techniques need to be used to transform and analyse this data so that it can help with decision-making and be accessible to users.

This can also be applied to services like education. The existence of data from thousands of students, having similar learning experiences but in very different contexts, gives information that was not possible before, for studying the influence of different contextual factors on learning and learners (Baker,2010). Through data mining, a university could try to predict the percentage of students that are most probable to fail and the factors that contribute to that. The university would use this information to help the students most at risk (Jing Luan, 2006). This is where this project comes in. Exploring this area, solutions can be found for the problem above, using new technologies and tools for data mining and processing.

This work is a part and will follow the methodology of the IOScience project, created by IOTech. IOScience consists of a system that enables a data analysis process, even in offline mode while also having the possibility of incorporating an artificial intelligence module. This

system is able to analyse data even when it is offline, updating them when it comes online. Working as *Data Science as a Service* (DsaaS), this platform allows for individuals or companies to use it by a web or mobile app, making it a resource with the goal of being available to the society in general, solving problems like connectivity and different data types available. This system is composed of six different layers that can work independently and connected with each other (Portela & Fernandes, 2020):

- Data source that allows the system to start;
- API that is responsible for the data processing tasks;
- Data warehouse where the multidimensional model is filled and stored;
- OLAP (Online analytical processing) layer for data query;
- Cache layer where the different data queries are stored and managed;
- Visualization layer where the data is transformed into dashboards and available for the client;

For this dissertation, the data was collected from a real case, the course "Web Programming", from the 5th of October in 2020 to the 29th of January in 2021 and taught by Professor Filipe Portela. According to Portela (2022), this course had the value of 10 ECTS (European Credit Transfer and Accumulation System) and 168 students were registered in it where 90% of them were active participants (attended more than 50% of the classes). These classes were divided in three different formats: Theoretical (T), Theoretical practice (TP), laboratory practices (LP) that used different learning approaches and techniques for each one of them. Since this was with online classes only, the researcher used different platforms for the e-learning process and to collect data from it: IOEduc, IOChat, Zoom, Kahoot! And HackerRank. Google analytics was also used to collect data as well as the students' responses to surveys made during the course. The evaluation for this course was divided in two different types: A project in teams that was made during theoretical practice and laboratory practice classes, and continuous evaluation methods to evaluate professors, students, and the course. This included weekly quizzes, card system and three individual mini tests during the year to evaluate different types of knowledge. Additionally, the students filled three different surveys during the course: one in the beginning, one in the middle and one in the end, about the course, their experiences and expectations. After collecting the data, the researcher was able to draw some results. Firstly, using the surveys, it was obvious the students fear with this new program. In the first survey, in

the beginning of the course, when asked about what they felt about having a full course with online classes, 52,68% of the students answered that they were afraid, even though 73,08% were enrolled because they wanted to learn the subject (web programming). However, during the second survey, in the middle of the course, 98.86% of the students were enjoying online classes. These surveys were also helpful to understand that most of the students enjoyed the gamification approach (85%), while 96,52% considered this method an average or above average teaching methodology. Using Google analytics in the IOEduc platform, the researcher was able to see statistics like more than 220k page views were made, 15k messages were sent in the chat and more than 11k files were downloaded during the course time. It was also possible to check that basically every hour of the day had at least one student online in the platform. Finally, using a word cloud in the end of the course, the researcher was able to collect feedback about what went well and not so well during the learning time. The positives include "dynamic classes" while the negatives were mainly "nothing" and "less material", meaning that some changes might need to be made in the amount of material given in following years. (Portela, 2022). This experiment followed the TechTeach paradigm.

According to Portela (2020), the TechTeach paradigm is a new concept of learning and teaching in higher education that "allows the creation of a B-learning environment and uses gamification to motivate the students to participate in the class". This paradigm includes interactive classes, quizzes and surveys, project-based learning, and others. The aim of this concept is to show "how information systems and new technologies can contribute to the reengineering of processes and digital transformation" (Portela, 2020).

The curiosity for data mining technologies and techniques and the possibility of contributing to an important ongoing project were a big factor on the choice of the realization of this dissertation project. The main objective of this project aims to optimize the education system that exists nowadays and help the digital transition to happen as smoothly as possible for the learner and teacher. The fact that more and more mobile devices are used on a day-to-day basis allows a whole universe of possibilities with the created data. Connecting it to all the existent platforms that are created by the day, the educational sector can only get better in the future, if the right path is taken.

In view of this, this dissertation project emerges as an enticing challenge, taking into account the entire educational area and the importance of the proper functioning of e-learning.

2. Goals

This project has the main objective of answering the following research question: "Is it possible to predict the students grade based on his behaviour in class?"

To answer this question, this project is based on the development of a prototype that can process and analyse education data to predict future students' grades. As concept proof was used the company IOTech, which provided the data to test the output of this dissertation. This data can only be used in this project context, being forbidden its reproduction.

The main goals for this project are:

- Development of a prototype to consult information through different Dashboards about the data that was processed;
- Prediction feature that can call if a student will pass or fail, based on certain parameters;

These main goals can be divided in secondary goals:

- Implement the artifact using ioScience;
- Develop a Multidimensional Model;
- Develop a set of Dashboards;
- Develop mechanisms to handle Unstructured Data;
- Develop mechanisms to clean and select Data;

During the process of writing this dissertation, it was expected to have a fully developed and functional prototype which suited its purpose and that could handle different types of data from different data sources. In a more theoretical level, it was also expected an increase related to the combined knowledge between information systems, programming skills, and data mining.

To achieve these objectives, a literary review was initially carried out, where similar solutions, tools and technologies were identified for the prototype development. Consequently, it was also analysed the different methodologies that could help in the different phases of the project.

Finally, the extraction, transform and load (ETL) process was carried out followed by Data Mining modelling and data analysis.

3. Document structure

This dissertation is structured in six chapters. The first one is this introduction, where it is identified the context and motivation for this work, the main and secondary objectives, and its goals.

The second chapter consists of the literature review that was made for this project. It is composed of the theoretical review of the main topics related to the subject of this work and the state of art of what has been written about this problem. This chapter allows to define some important concepts related to this dissertation and what can this project improve or add in the area.

In the third chapter is presented the methodological approach used during this dissertation, it is presented which methodologies were chosen to support this project (DSR, SCRUM and CRISP-DM) and a brief description of each one.

The fourth chapter consists of the several steps made during the prototype development, since the data collection and transforming part to data warehouse description, finishing with the presentation of data mining results and dashboards and the discussion of its results.

In the end, the fifth chapter gives a general analysis of all the document and its main conclusions, while the final chapter is a list of bibliographic references used along the present document.

Chapter 2 – Literature Review

This chapter serves to introduce several relevant concepts to this dissertation development as present existent works on the field and its actual situation. This way the dissertation theme can be presented along with the state of the art.

1. Search Strategy

This project started with some scientific research on different topics to cover the main things around the chosen theme. Firstly, it was explained the data mining process, its techniques and sub areas like Web Mining and Educational Data Mining. The second part covered the big data concept and its characteristics. The following section explored Structured Query Language (SQL) and Not only Structured Query Language (NoSQL) databases, their differences and the different types of data that exist. Finally, it was made a brief research about the state of art that already exists in this topic.

During this work, several rules and protocols were followed to save time and select the best information possible, and which one was relevant to this work. The main rules were:

- Searching information about topics like data mining and web mining inside education, big data, SQL, and NoSQL databases;
- The search needed to include different services and databases like Google Scholar, ResearchGate, Scopus, ScienceDirect, and others;
- The first search was made on books, followed by scientific work and finally by another type of documents like master's dissertation, journal articles, etc.
- The search included Portuguese and English content;
- The main filter was made by reading the abstract and conclusions to check if the information was relevant;
- The main criteria used to choose the bibliographic references were the author's reputation, and the publishing year, where the documents publish year must be from 2010 ahead, except in a few cases.

2. Data Mining

This section presents the data mining (DM) concept, how can it be described, the main techniques used, and its role in nowadays world.

2.1. What is Data Mining?

According to Sadiku et al. (2015), DM can be described as the process of finding insightful and predictive models and information from large amounts of data. This process combines data analysis (DA) with complex algorithms for processing massive data. Aggarwal (2015) says that "data mining is a broad umbrella term" that is used to represent several aspects like collecting, cleaning, processing, analysing, and gaining useful information from data.

Gorunescu (2011) claims that finding a unique definition for DM is very difficult. In that sense, he has several approaches that would describe DM in the best way possible. According to Gorunescu (2011), data mining is:

- An automatic search of patterns in massive databases, using computational techniques from several fields like statistics, machine learning, and pattern recognition.
- The extraction of implicit, unknown, and potentially valuable information from data.
- The science of taking valuable information from massive sources of data.
- Generating information automatically through identification of patterns and relationships 'hidden' in data.

2.2. The Data Mining Process

The DM process that transforms raw data into information and knowledge can be described in several steps. The first ones are used for data pre-processing, which means that this is the part where the data is adapted and prepared to be used later in the process. The following steps are used to work on the data that was formed to extract the valuable knowledge and information that is available to us (Agarwal, 2013). Every data mining process explained in this dissertation follows the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, explained more in detail later in the Methods and Tools chapter. Figure 1 represents the data flow for this process.

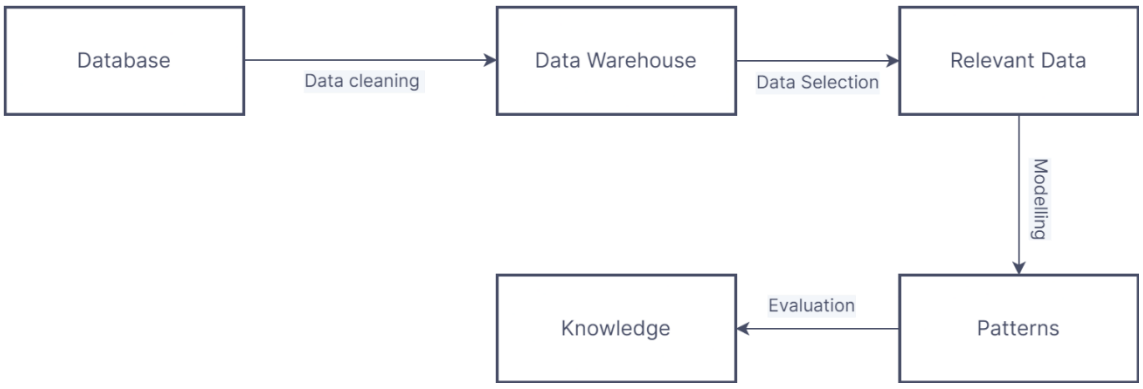

Figure 1 - Steps in Data Mining process (adapted from (Agarwal,2013))

This process can be briefly explained by the data flow. In order to start it, a database with raw data needs to be selected as data source. The data will then be sent into cleaning and integration processes, followed by being stored in a data warehouse (DW). After this, the data goes through a selection process in order to filter the relevant data and make it ready for the modelling process. After using data mining techniques to model it, the result of this modelling (patterns) will be evaluated to confirm if it is trustworthy. In the end, this data will be transformed in knowledge that can help in decision making.

After understanding the process, it is important how to identify patterns from the data that already got cleaned and selected. This is where the DM techniques are valuable for the process.

2.3. Data Mining Techniques

Like it was said in the DM process section, the data mining process includes the use of several tools to identify patterns. However, this task can be used in different ways and each one of them is called a data mining technique. Each technique must be chosen based on the type of business and problem the business faces. In this paper, three techniques are described: classification, regression and clustering.

a) Classification

According to Voznika & Viana (2007), classification can be described as the act of predicting a certain outcome based on a given input. In order to predict the result, the algorithm processes

a training set where is possible to find a set of attributes and the prediction attribute. Figure 2 represents an example of a classification DM process.

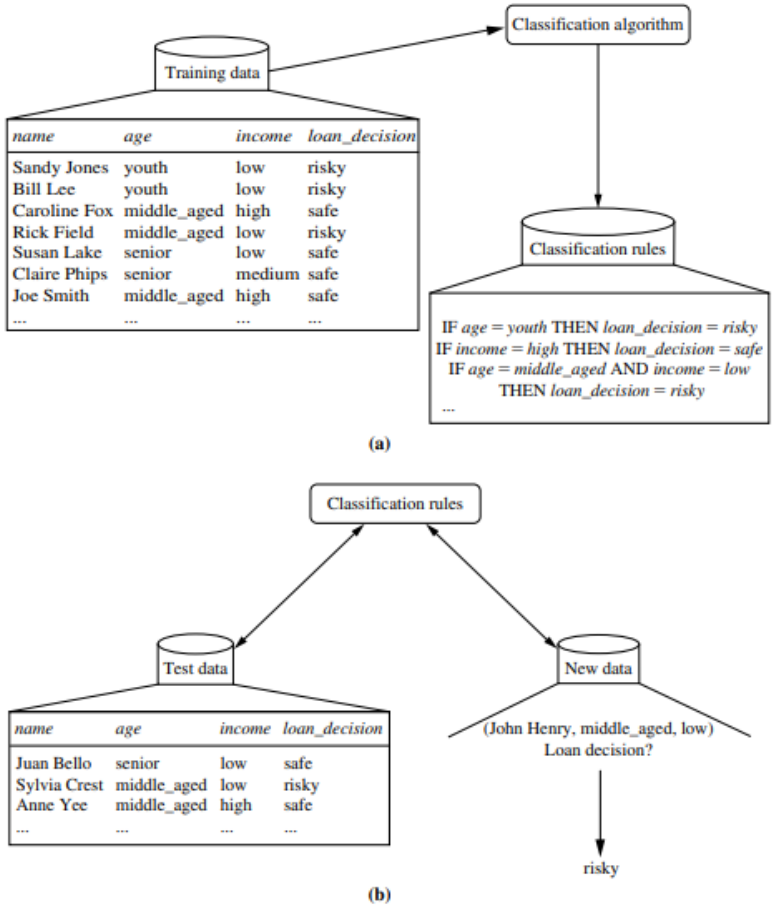

Figure 2 - The data classification process (adapted from (Han et al., 2011))

This technique is divided into two different steps: Firstly, it starts by training the model where a classification algorithm builds the classifier by learning and analysing a given training set (Han et al., 2011). This training set contains a set of attributes, including the prediction attribute. The algorithm will read this set and try to find the connections and relationships between attributes that can help to predict the goal (prediction attribute) later (Voznika & Viana, 2007). In the second phase of the process, the algorithm will face a dataset with data with the attributes that were given in the training set but with different content that was never analysed. The algorithm, based on the first step, will try to predict the prediction attribute.

In the end, based on his success, the algorithm can be ranked based on his accuracy. The accuracy of a classifier on a given test set is the percentage of prediction attributes that are correctly classified or predicted by the classifier (Han et al., 2011).

In figure 2, it is represented a case where the goal prediction is the loan decision of a bank based on their clients' information. The attributes are name, age, and income while the goal prediction has two different values, "safe" or "risky". The algorithm will learn from the training set and then make decisions for the test set Han et al., 2011).

According to Han et al. (2011):

- "a) Learning: Training data are analysed by a classification algorithm. Here, the class label attribute is loan decision, and the learned model or classifier is represented in the form of classification rules."
- "(b) Classification: Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples"

Once the results and all class labels have been predicted, some measures are used for assessing how good or how "accurate" the classifier is. Some of these measures are accuracy, confusion matrix and sensitivity. The accuracy of the model can be calculated like this:

$$\frac{\text{Correct Predictions}}{\text{Total predictions}} \times 100$$

This number will be a percentage and will tell us from 0% to 100% how accurate the model is.

According to Han et al. (2011), "The confusion matrix is a useful tool for analysing how well your classifier can recognize tuples of different classes." For this case, some terms need to be explained:

- TP – True Positives
- FP – False Positives
- TN – True Negatives
- FN – False negatives

In figure 3 is possible to see the representation of a confusion matrix.

		Predicted	
		yes	**no**
Actual	**yes**	TP	FN
	no	FP	TN

Figure 3 - Confusion Matrix

In this example, TP and TN represent the labels that were predicted right, while FP and FN represent the wrong labels predicted by the classifier. The goal is to have the positives (TP and TN) the higher number possible while the negatives (FP and FN) stay low.

Lastly, the sensitivity of a classifier is used to measure the fraction of positive patterns that are correctly classified. It is very close to the accuracy measure, but it only focuses on the positive values:

$$\frac{\mathrm{TP}}{\mathrm{TP+FN}} \times 100$$

b) Regression

The regression technique is very similar to the classification technique that was explained just before. Both predict a specified attribute based on the relationships of the other ones. However, in regression, the algorithm will predict missing or unavailable numerical data values instead of class labels (classification) (Han et al., 2011).

For example, if a store wants to predict the future sales of their products based on their sales data over the years, they can both use classification and regression. If classification is the used technique, the results will be in labels like “Increase in sales” or “Decrease in sales” but not the exact number. On the other hand, if the store chooses the regression technique, the results will be a specific number like “2000” or “5689”. Both techniques are very useful, and both have different approaches and areas where they should be used.

The regression method has different ways of calculating the success of the model. In order to do this, it is needed to calculate the RMSE (Root mean squared error) of the model. The formula is:

$$RMSE = \sqrt{[\Sigma(P_i - O_i)^2 / n]}$$

where:

- Σ represents "sum"
- P_i is the predicted value for the i^{th} observation in the dataset
- O_i is the observed value for the i^{th} observation in the dataset
- n is the sample size

Furthermore, the regression technique success can also be calculated using the mean absolute error (MAE) of the model. The formula for this metric is:

$$MAE = (1/n) * \Sigma|O_i - P_i|$$

where:

- Σ represents "sum"
- P_i is the predicted value for the i^{th} observation in the dataset
- O_i is the observed value for the i^{th} observation in the dataset
- n is the sample size

c) Clustering

After covering the classification and regression techniques, is time to talk about another type of process that is also used on a large scale, clustering. Clustering can be defined as the process where data is separated into groups/subparts (Han et al., 2011). Clustering algorithms partition data objects like patterns or entities into a certain number of clusters like groups or categories (Wunsch & Xu, 2008). Each one of these groups is called a cluster and all the elements of a certain cluster have similarities with each other and are different from the elements from other clusters. In figure 4 it is possible to see an example of a cluster created with the k-means method.

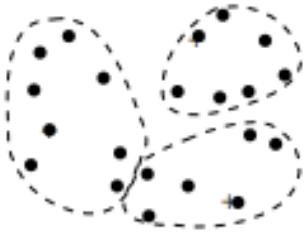

Figure 4 - Clustering of a set of objects using the k-means method (adapted from (Han et al., 2011))

This technique is useful in order to have a better look at the distribution of the data that is being analysed and have a clear separation of groups by the characteristics that will be useful to reach the business goals. For example, clustering is largely used in business intelligence in order to split the customers into groups with different characteristics which helps massively in creating different strategies for different kinds of customers.

According to Wunsch & Xu (2008), the clustering process can be divided into four different steps: Feature selection or extraction, clustering algorithm design or selection, cluster validation, and result interpretation.

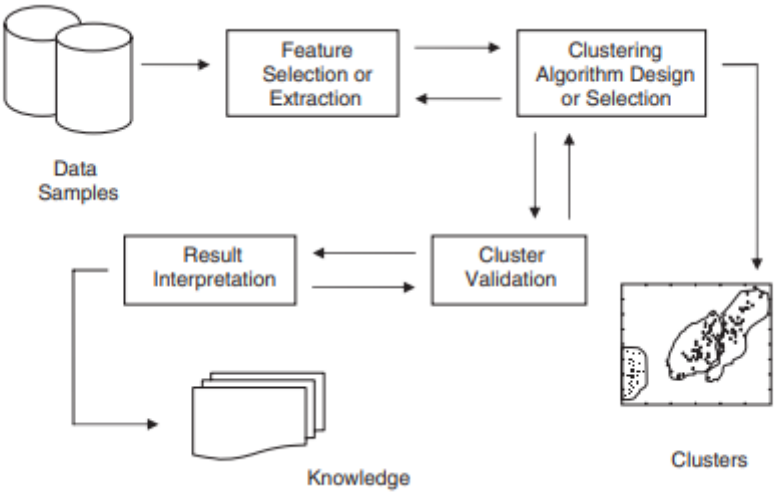

Figure 5 - Clustering procedure (adapted from (Wunsch & Xu, 2008))

Figure 5 represents "the basic process of cluster analysis consists of four steps with a feedback pathway. These steps are closely related to each other and determine the derived clusters." (Wunsch & Xu, 2008). The feedback represents the feedback exchanged between phases, which means that this process can go back and forth, it is not a one-shot thing.

In the first step, the feature selection consists of selecting the most important attributes for analysis in the future. This is highly important because the attributes that are not selected here, will be eliminated from the process so it is important to create some ranking of the attributes in order to order the several attributes by their level of importance. In the feature extraction, the selected attributes are transformed in order to generate new features that will be more useful to the project down the line. The feature extraction process results in a smaller and richer set of attributes. Both processes are very important to the effectiveness of clustering applications and it's important to know their differences.

Following that, there is a selection or design of an appropriate algorithm for the clustering process. There are several algorithms available, and this decision should be made by thinking about which one is more useful in order to achieve the goals of the project. There is no universal clustering algorithm to solve every problem that exists so is crucial that the team carefully investigate the characteristics of the problem in order to select or design an appropriate clustering algorithm.

Once the algorithm has been selected and created the clusters, the process of clustering validation starts. Each algorithm will create different clusters so there must be a way to evaluate correctly the results by creating standard rules and criteria that will be put in place in order to rank these algorithms. This evaluation should create confidence in the team so it must be objective and independent of any favourite algorithm. During this validation, it should be explained every choice between algorithms and in which way the clusters will help to accomplish the business goals.

Finally, the last phase consists of the result interpretation. After the clusters are created and approved by the team, they can start to be interpreted and observed in order to answer the business questions and solve the problems that they are supposed to solve.

Once the DM process and techniques were explained, the next important thing to know is where are these techniques applied. Since data mining is such a large topic, it can be applied to different areas. One of the biggest data mining areas where it's applied is the World Wide Web (WWW). This process is called Web Mining.

2.4. Web Mining

This section presents the Web Mining (WM) concept and how it relates to the data mining topic.

Nowadays, the Web has turned to be the largest information source available on this planet (Kumar,2015). The Web is a huge, explosive, diverse, dynamic, and mostly unstructured data repository, which contains an incredible amount of information (Kumar,2015). The different types of data available must be organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is more important by the day in order to collect and transform this data into useful information.

These data techniques in the web can be classified as WM that, according to Bin & Zhijing (2003), is the process of using data mining techniques to find and collect information from web documents and services. Although Web mining comes from data mining, it is not equivalent to data mining. The complexity of Web data makes the task of Web mining more challenging (Kumar,2015).

The WM process can be decomposed into smaller subtasks/steps that can be seen in figure 6.

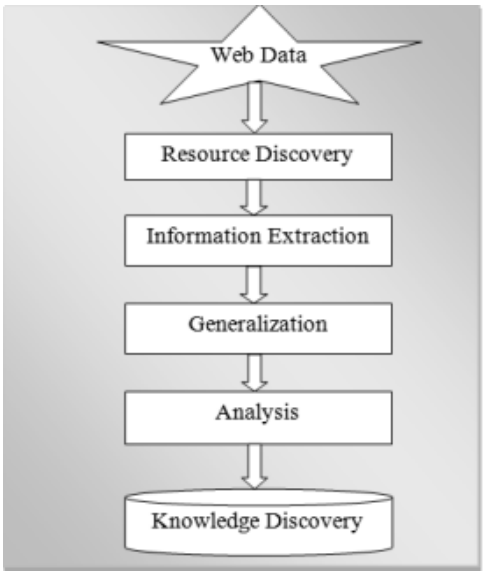

Figure 6 - Steps of Web Mining (adapted from (Kumar,2015))

According to Kumar (2015), resource discovery is the task of collecting the intended information from the Web. Following that, the information extraction consists of automatic selection and pre-processing specific information from the web sources collected before. Once this is done, comes the automatic discovery of patterns in the websites (Generalization). In this step machine learning and traditional data mining techniques are typically used. Finally, the mined patterns are analysed and validated (Analysis).

Web mining is a big process that can be done in different contexts. In that way, the process was divided into three different types of techniques of mining: Web Content Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM). This division is described below in figure 7.

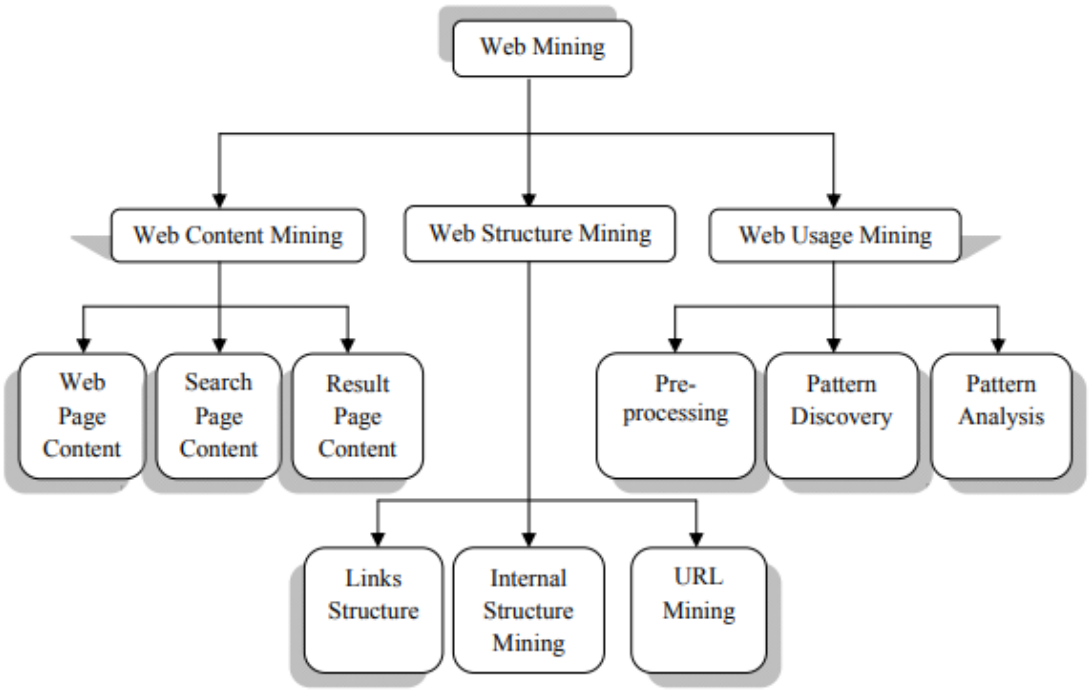

Figure 7 – Classification of Web Mining (adapted from (Sharma et al., 2011))

The three different types of web mining can be described as followed:

a) Web Content Mining

The first technique, WCM, consists of extracting useful information from the content of web documents. This process is done by mining, scanning, and extracting text, videos, graphs, and pictures from web documents. Using this data, it can provide effective and interesting patterns about user needs. When done in texts, this type of mining can be classified as text mining. Text mining performs scanning and mining of the text, images, and groups of web pages according to the content of the input.

b) Web Structure Mining

WSM is the extraction of structural information of the Web. Imagining it as a graph, the web pages are represented as nodes and Hyperlinks represent edges (Sharma et al., 2011). This type of mining can give information on connected pages (by information or direct links), making it an important tool for search engines for example. WSM is also very helpful to users since it allows them to access the desired information through keyword association and content mining (Kumar,2015).

c) Web Usage Mining

Finally, WUM can be defined as the type of web mining activity that involves the automatic discovery of user access patterns from web data. This mining tries to discover the valuable information from the data that is generated from the interactions of the users while surfing on the Web. WUM is especially used to predict user behaviour when it interacts with the web. The analysis of this data can help companies to determine the lifetime value of customers, cross-marketing strategies across products, the effectiveness of promotional campaigns, etc (Bin & Zhijing, 2003). According to Kumar (2015)," Web Mining is Data Mining techniques applied to the WWW". WUM is thy WM type that will be mostly used during this project.

The next section will introduce how these topics can be applied to the educational field.

2.5. Educational Data Mining

This section presents one of the most important concepts for this dissertation, the educational data mining, and how it relates to the data mining topic and the work ahead.

Educational Data Mining (EDM) is an emerging discipline of data mining, focused on developing methods for exploring the increasingly amount of data that come from educational settings and explore it to find out descriptive patterns and predictions that characterize students' behaviour (Peña-Ayala, 2014).

According to Peña-Ayala (2014), "EDM can be applied to assess students' learning performance, to improve the learning process and guide students' learning, to provide feedback and adapt learning recommendations based on students' learning behaviours, to evaluate learning materials and courseware, to detect abnormal learning behaviours and problems, and to achieve a deeper understanding of educational phenomena". This means that EDM can have a great impact and revolutionize the way the education process is conducted nowadays.

EDM does not only impact students/learners but every actor that participates in this process. Teachers, school administrators and even the researchers that conduct the studies can benefit from it. Firstly, the students can have feedback from their work until then and a customized learning experience based on their needs that will try to improve the learning performance. The teachers will be able to understand better each one of their students, how their characteristics influence their performance and how to optimize the learning process to take the best out of them. School administrators will have the data to optimize their intuitional resources, both human and material. Finally, the researchers will be able to understand which techniques are more suitable for each individual situation. Obviously, these goals cannot be accomplished all in the same study so, before starting, it should be defined one main goal (students' performance prediction, behaviour analysis etc.) (Romero & Ventura, 2013).

To achieve the EDM goals, the researchers usually use the traditional techniques, that were explained in a section above, like classification, regression, and clustering. Classification is often used for predicting categorical values (for example "pass/fail"), while regression is also used for prediction purposes but when the predicted variable is a numerical value (for example 0-20 grade) (Peña-Ayala, 2014). Clustering, on the other hand, is used for different purposes. Clustering is a great way of grouping students by their characteristics or behaviours, making it

easier to study specific groups. Text mining can also be used on EDM, especially when the goal is to o analyse discussion boards, forums, chats, Web pages, documents, and other text material.

The next section reflects how the last two topics, Web Mining and Educational Data Mining, create the core for this project and how they relate to each other inside this project context.

3. Project Placement

Both EDM and Web Mining are the central core for this project since it includes properties and characteristics from both areas. Figure 8 gives how these themes relate to the project and how they connect with each other.

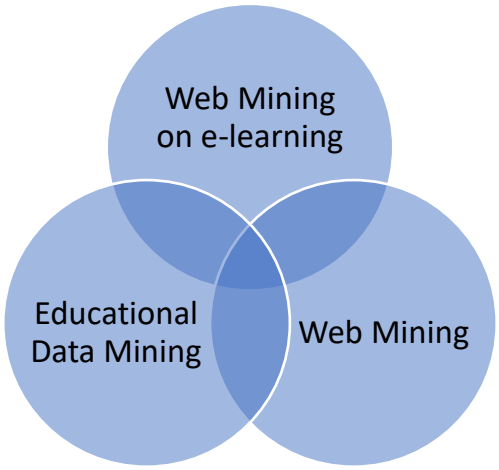

Figure 8 - Project Overview

Firstly, regarding Web Mining, this project used Web Usage Mining in different web platforms (explained in the Context and Motivation section) in order to get the information and analyse the student's behaviour based on their interaction during the classes. This work will help to develop the prediction functionality. Then, EDM brings the educational environment into this project that allows to analyse this data from an educational point of view. It also opens new horizons in terms of data analysis, new parameters to evaluate and new features regarding data generated from the educational sector. Both themes associate with each other since WM techniques are also used on EDM.

4. Big Data

In the Web mining section, it was referred the amount of data that is available on the Web. Extrapolating this to every possible source, the data that is available nowadays is bigger than it has ever been. This event can be classified as "Big Data" (BigD). This subchapter must start with a proper introduction of this concept so, how did it appear and how can it be described?

According to Fan & Bifet (2013), during the last few years, society have witnessed a big increase in our ability to collect data from different sensors, devices, from independent and connected applications. This enormous data collection has surpassed our capability to process and store this information. Therefore, the term "Big Data" can be defined as "information that can't be processed or analysed using traditional processes or tools" (Zikopoulos et al., 2012).

Wilder-James (2012) adds that this data is too big, moves too fast, or does not fit the structure of the database architecture that is in place. To get the value from this data, it needs to be chosen a new way to process it.

Hurwitz et al. (2013) adds a new view and defines BigD as "capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction."

4.1. Big Data Evolution

There are three innate characteristics of BigD known as the "3 V's of Big Data" which help us to better understand the essential elements of big data. These 3 V's are Volume, Velocity, and Variety, shown in figure 9.

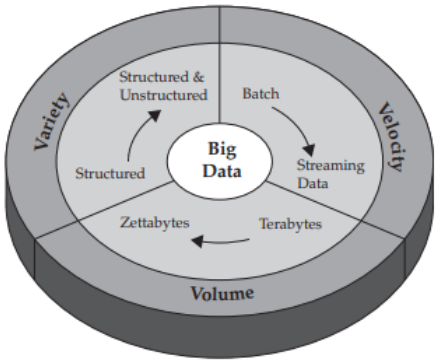

Figure 9 - IBM characterizes Big Data by its volume, velocity, and variety (adapted from (Zikopoulos et.al, 2012))

With time and with the growing of this concept, some authors started to define Big Data as five V's instead of three like it was just described. According to Song & Zhu (2016), the new meanings added are Veracity and Value.

Figure 10 represents the concept of the 5Vs of big data, where the new dimensions added are veracity and value which is at the centre of the diagram and intersects with the other dimensions.

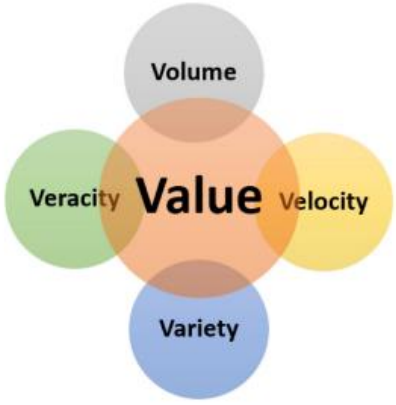

Figure 10 - 5 Vs of Big data (adapted from (Song & Zhu, 2016))

However, this definition did not stop here. Nowadays, the big data concept is so wide, with so many variables that some authors already consider existing 7 different V's. According to Mcnulty (2014), BigD can be characterized by more two characteristics. These "new" V's are visualization and variability, as represented in figure 11.

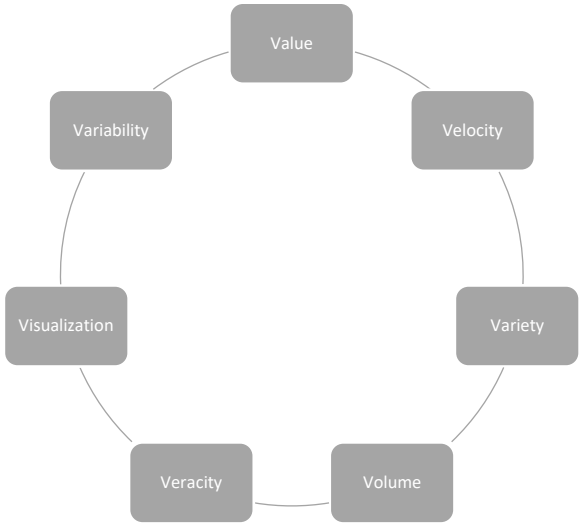

Figure 11 - 7 Vs of Big Data

In the next section, each one of the previous named characteristics will be explained in detail, in order to understand each challenge that exists in the area right now.

4.2. Big Data Characteristics

Data Volume

Starting by data volume, this concept can be described by the amount of data that is generated continuously (Krishnan, 2013). The volume (or size) is the characteristic that can label a certain dataset as "Big Data".

According to Zikopoulos et al (2012), the volume of data being stored nowadays is exploding. In 2000, the world had 8000 PB (petabytes) of data stored. In 2012, just the Facebook platform generated more than 10 TB (terabyte) of data every single day. Just to get an idea of today's world, according to Sinha (2021), more than 64.2 ZB (zettabytes) were consumed in 2020. These numbers make sense since devices can track and record anything nowadays, from financial data to medical data or even entertainment data.

This volume presents the biggest challenge to conventional IT structures. Several companies nowadays have large amounts of stored data, perhaps in different forms and types, but not the capacity to process it (Zikopoulos et al., 2012).

Data Velocity

Just like, the volume, the velocity of data has changed over the years. Data velocity is "the increasing rate at which data flows into an organization" (Wilder-James, 2012). This rate has increased very quickly due to the internet and mobile era where people deliver and consume even more products and services, generating a data flow back to the provider.

In order to gain an advantage, organizations need to analyse and process data in near real-time if they want to find insights and value on it. For example, companies like Facebook or Google gather millions of clicks from users every second, amounting to large volumes of data. In order to study consumer behaviour effectively, these companies must analyse, process, and store these data the fastest way possible, especially since the number of users is growing by the day. The volume of data that is processed by these companies made them an example, opening this technology to the rest of the world. The velocity of data produced by clicks on any website today is a prime example of this big data velocity (Krishnan, 2013).

Data Variety

Another characteristic that is key to big data is data variety. This variety brings new challenges to organizations trying to deal with a big volume of data. With the growth in devices used like sensors, smartphones, or social collaboration technologies, data in these organizations has become more complex. This complexity means that this data includes different data types: structured, semi-structured, and unstructured data (Zikopoulos et al., 2012).

This variety increases the problem complexity since the data processing is depending on the existence of proper metadata to inform the information available. This adds pre-requisites for the platform for processing new formats like scalability, distributed, image processing, graph, video, and audio capabilities (Krishnan, 2013).

Data Value and Data Veracity

Data veracity means quality, reliability, and uncertainty in data. Veracity is a challenge that needs to be researched in order to understand its impact on data integration and analytics.

The fifth "V" corresponds to the Data Value. This dimension is the most important one because it includes all the others above and contains the result that makes big data projects valuable. This value means "the discovery of actionable knowledge, high return on investment, increased relevancy to customers or products, or innovations in business operations/processes." (Song & Zhu, 2016).

Data Visualization and Data Variability

Data visualization can be described as the way of representing the data after being processed. Representing data as clear and understandable as possible is one of the big challenges in BigD so, data visualization is a big theme in this area.

Furthermore, Data Variability refers to the data which keeps on changing constantly (Mcnulty, 2014). This area mainly focuses on understanding and interpreting the correct meanings of raw data. Variability differs from variety since it is not about different types of data, but mainly about the meaning of the data and its constant changing.

5. SQL & NoSQL Databases

After understanding the concept of big data and its challenges for society it's easier to explain what a Structured Query Language and Not only Structured Query Language database is, their differences, and which one is more useful in certain contexts.

Firstly, let's understand the different types of data available for processing:

1. **Structured data**: According to Kanimozhi & Venkatesan (2015), structured data can be described as "data included in a relational database system". This data can be managed by technologies like SQL, follows a consistent order, and can be easily searched and accessed. An example of this data is a DB table with rows and columns.
2. **Semi-Structured Data**: Semi-structured data is data that has some structure but lacks data model structure or does not follow a rigid structure. This means that this kind of data does not need a schema definition, it can have markers that separate semantic elements and enforce hierarchies. This data is not stored in a relational DB but has some organizational properties that make it easier to search and analyse. Usually, languages like XML or JSON are used to manage this type of data. In order to convert to structured data, it's necessary to do some data mining work (Kanimozhi & Venkatesan,2015).
3. **Unstructured Data**: The last type of data, and the most complex one, is unstructured data. This data has no identifiable structure making it harder to be used. This data can be non-textual unstructured data (images and videos) or textual unstructured data (emails and instant messages) (Kanimozhi & Venkatesan,2015).

In table 1 is possible to compare the differences between the three different data types.

Table 1 - Data Types Comparison

Property	Structured Data	Semi- Structured Data	Unstructured Data
Format type	Relational Database	XML, JSON	Binary
Flexibility	Schema dependent and less flexible	More flexible than structured data but less flexible than unstructured data	More flexible, no schema
Scalability	Very difficult to scale DB schema	Simpler scaling than structured data	Easier to scale
Robustness	Very robust	Limited Robustness	-

This shows that unstructured data is the most complex and harder to find information from it. Adding to it, 95% of the digital universe is made from this unstructured data (Kanimozhi & Venkatesan,2015). Therefore, it's essential to find tools and techniques that can give some structure to this data in order to get value and information from it.

5.1. SQL Databases

Starting with the Relational Database Management System (RDMS), or simply Structured Query Language databases (SQL DB), these databases are written in, Structured Query Language (SQL) and are a rigid, structured way of storing data. The data is stored in rows and columns where each table needs to have one primary key. According to Kotecha & Joshiyara (2017), "each row represents an entry, and each column sorts a very specific type of information".

In figures 12 and 13, it is possible to see an example of a relational database and its structure. In this case, a table called "Employee" will be used to store the employee's information according to several attributes that characterize them. The "E#" attribute is the key attribute of this table, this is the number that uniquely identifies each one of the employees (Meier & Kaufmann, 2019).

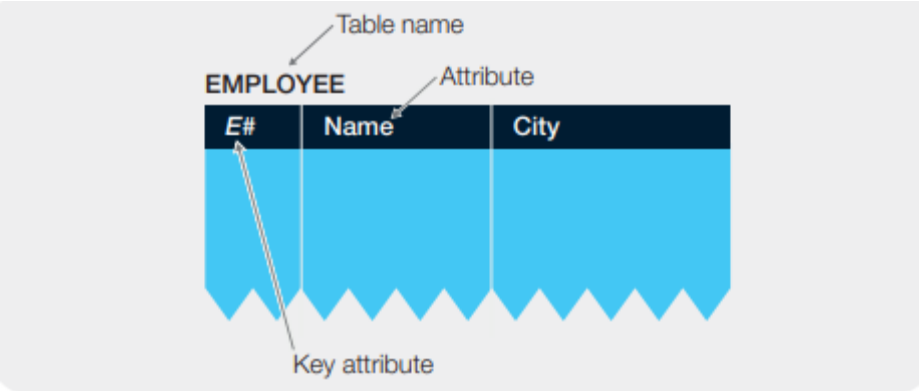

Figure 12 - Table structure for an EMPLOYEE table (adapted from (Meier & Kaufmann, 2019))

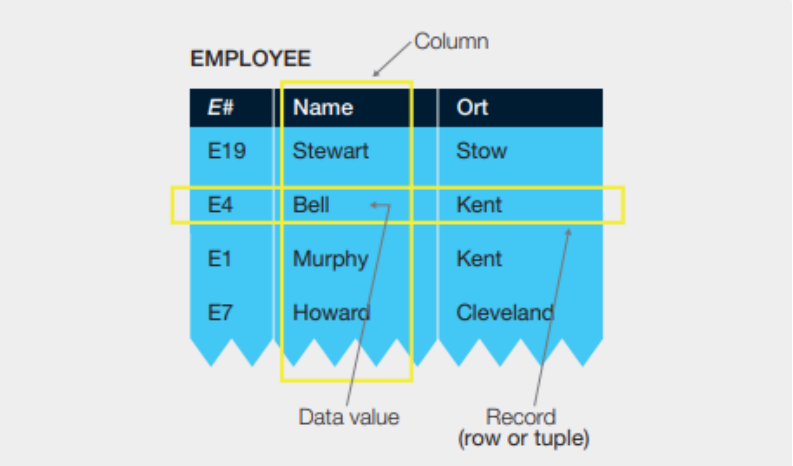

Figure 13 - EMPLOYEE table with manifestations (adapted from (Meier & Kaufmann, 2019))

In figure 13, two records have the same value (Kent). For this reason, the key attribute E# is required to uniquely identify each employee in the table.

SQL is the language used to do operations and selections in relational databases. The result of a selective operation is returned by the database management system as a table, even a blank one if the selection processed has no results found. One single SQL query can process multiple actions within the database management system (Meier & Kaufmann, 2019). Using the same example above, figure 14 can give an example of a SQL operation.

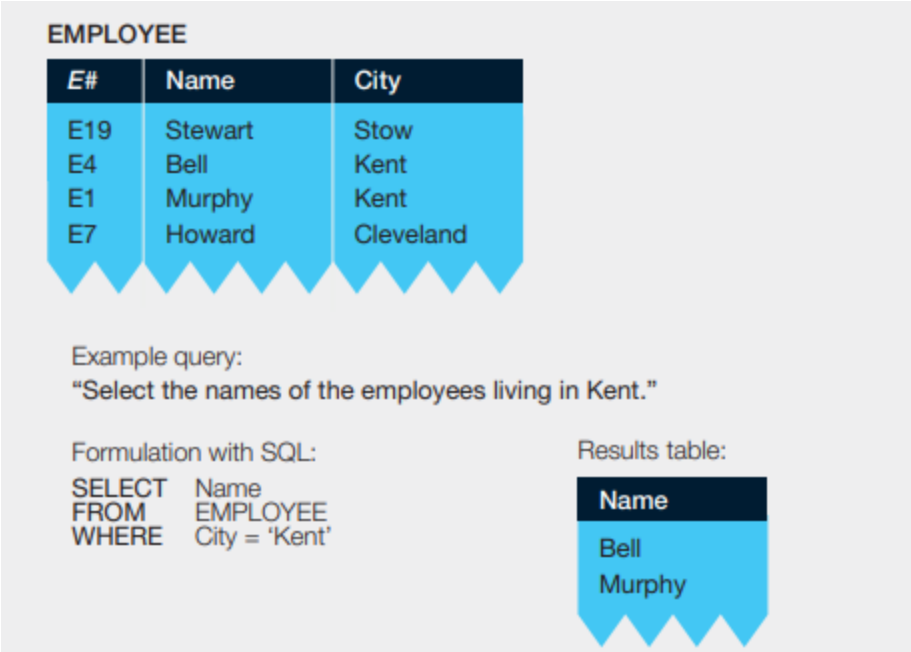

Figure 14 - Formulating a query in SQL (adapted from (Meier & Kaufmann, 2019))

These databases make it possible to perform focused queries obtaining very formatted results which is an advantage (Kanimozhi & Venkatesan, 2015). However, due to its difficulty to lead to data that is not in the required schema and by being of difficult scalability, is not advisable to use this for big data processing. (Kotecha & Joshiyara, 2017). To solve this problem, in 2009, the NoSQL databases appeared.

5.2. NoSQL Databases

NoSQL databases are non-relational databases made for storing and processing unstructured big data which is distributed over many servers (Jumaa, 2018). These databases were created by big companies to solve the problems dealing with large amounts of data that were impossible to store and process in relational databases. It allows to store and retrieve data that is not modelled in tabular relations like typical relational databases. According to Kotecha & Joshiyara (2017), "NoSQL deals with unstructured schema so, less data can be stored in multiple collections and nodes, and it does not require fixed table schemas".

These databases have several benefits compared to the previous ones. Firstly, NoSQL databases are highly and easily scalable. The expansion is done horizontally, this means adding more machines into your pool of resources. In relational databases, this expansion is made

vertically, by increasing server hardware power need to acquire expensive and bigger servers. This makes NoSQL databases maintenance way cheaper than relational ones. Another benefit comes from the support of all data types, instead of only accepting structured data, "NoSQL databases offer flexible and dynamic schema that accepts all key data formats including structured, semi-structured, and unstructured" (Kotecha & Joshiyara, 2017). The last main benefit from NoSQL databases comes in form of performance. These databases support caching in system memory, so it increases data output performance. In relational databases, this needs to be made using outside infrastructures.

NoSQL also differentiates from relational databases in terms of its properties. While relational databases follow the ACID (Atomicity, Consistency, Isolation, Durability) properties, NoSQL databases support the BASE (Basic Availability, Soft state, and Eventual consistency). Databases with ACID properties have strict rules for the data and its consistency. These properties require data to always be in a consistent state, which takes a lot of time for every type of operation. This is the reason why relational databases are not suitable for applications with huge traffic. On the other hand, BASE properties consist of data being available at any time, even though it might not be consistent all the time. Keeping the data in a consistent state is moved to the application and the developer is responsible for this task. Eventual consistency means that the data is not consistent at all times but will be at some time (Mitreva & Kaloyanova, 2013). In figure 15 is possible to see the differences between these two properties.

ACID	BASE
Strong consistency	Weak consistency – stale data OK
Isolation	Availability first
Focus on "commit"	Best effort
Nested transactions	Approximate answers OK
Availability?	Aggressive (optimistic)
Conservative (pessimistic)	Simpler!
Difficult evolution (e.g. schema)	Faster
	Easier evolution

Figure 15 - ACID vs. BASE (adapted from (Strauch et al., 2011))

According to Jumaa (2018), there are four types of NoSQL databases that can be divided into different groups:

- **Key-Values Databases** are the simplest way of data storage. It consists of indexed keys and values. Every object in the database is stored as a field name along with its value of any type. This value can be accessed with a key.
- **Document Databases** it's similar to the previous type but more complex. The value corresponds to a document and each one has its own data, and its own unique key, which is used to retrieve it. This type is flexible, and it allows dynamic data modification, adding or removing content from the document. It's usually used for storing, retrieving, and managing data that's document-oriented but still somewhat structured.
- **Graph Databases** consist of using a flexible graph model to store the data, where objects are represented as nodes and the relationships between them are represented as edges. The nodes are organized according to the relationship between them.
- **Column Databases** stores data tables as columns rather than rows. These columns consist of three elements: name, value, and timestamp. This allows high performance and scalability.

Figure 16 represents three different types of these databases, document, graph, and key-value.

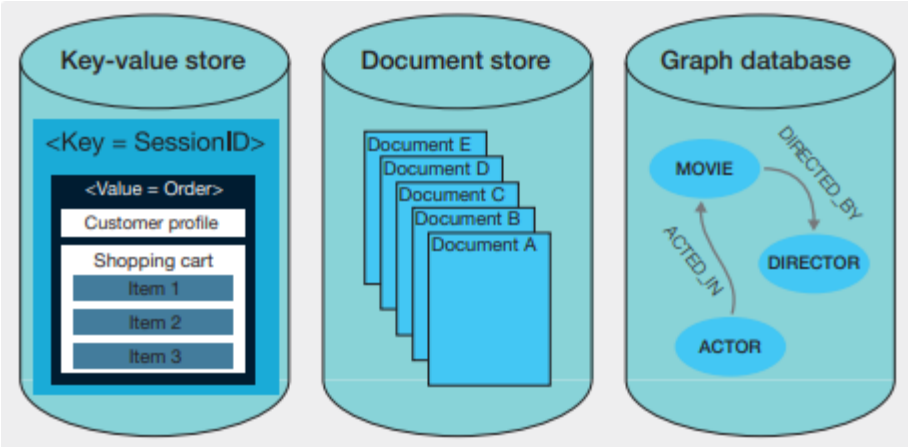

Figure 16 - Three different NoSQL databases (adapted from (Meier & Kaufmann, 2019))

In Table 2 is possible to compare the different types according to their performance, scalability, complexity, and flexibility.

Table 2 - Non-Relational database comparison

	Performance	Scalability	Flexibility	Complexity
Key-Value Store	High	High	High	None
Column Store	High	High	Moderate	Low
Document	High	Variable	High	Low
Graph Database	Variable	Variable	High	High

6. Related Work

Before trying to solve the problem itself, nothing more logic than search about what already was purposed by others and how it approaches a solution for these questions.

This search was mainly made on previous studies made on students' performance, even though there is a large amount of work made in data mining applied to the education sector but in different areas.

In 2008, Paulo Cortez and Alice Silva made a study in Portugal where the goal was to find if it is possible to predict students' performances and discover the factors that affect student achievement. For this matter, the researchers decided to analyse two different subjects, Mathematics and Portuguese with information from 2005-2006 in a secondary school in Alentejo, Portugal. The data was collected from school paper sheets containing the grades and absences, and from questionnaires made to students about their personal and families characteristics. For this study, the researchers chose a large set of variables. Some examples of these variables are age, sex, school, parents' education, family size, travel and study time, grades from previous periods, number of absences, number of failures, if the student had access to internet at home or current health status, just to name a few. The techniques used for this project were both classification and regression. The researchers used binary classification to divide the students into "fail" or "pass" categories and regular classification to label the grades from I (Very good) to V (insufficient). Finally, the students were evaluated from 0% to 100% using regression techniques. For this project, it was used five different data mining algorithms. In order to select the best algorithm, the researchers used the model accuracy in the classification models, and the RMSE in the regression models as performance metrics. In terms of results, when using the binary classification model, in Mathematics the Naïve Bayes algorithm was the most successful (91,9% and 83,8% accuracy) in two of the three tests, while in Portuguese, the Random Forest and Decision Tree algorithms got three models with accuracy above 85%, being the most joint most successful in this case. For the regular classification models, the same pattern was found, Naïve bayes algorithm the most successful in Mathematics (78,5% accuracy) while the Decision Tree and Random Forest algorithms were better in the Portuguese subject. However, this second modelling technique got accuracy percentages a lot lower when compared to the binary classification. Finally, in the regression models, the Random Forest got the lowest RMSE values in Mathematics making it the best algorithm on this subject, while in Portuguese the Random Forest algorithm stood out with the best RMSE values. In the end, the researchers were able to create a successful tool that was able to give information about the relation between absences or previous grades with the final grade. It was also possible to show some other relevant features, school and non-school related, for the main problem.

Romero et al. (2013) give a new view on this problem. The project has similar goals to the previous one but uses different data, instead of previous grades or student characteristics, the

researchers try to predict students' performance based on the activity on a forum in Moodle. This process was more complex since it included creating mechanisms to gather and select the data that was used. For this work, the researchers used a set of variables like the number of messages, words and sentences written, threads created, total time spent, average score per message and degree prestige and centrality. The researchers used tools that were able to rank the messages from invalid to messages that provide very complete or precise information about difficult topics (the highest rated messages). The DM technique used for this work was regular classification and classification via clustering, this last consists of a classifier based on the assumption that each cluster corresponds to a class. Both techniques were used to predict if the student passes or fails according to their activity. The metrics used for this project were accuracy and f-measure (harmonic mean of precision and recall). Analysing the results, it is possible to see that from the classification via clustering algorithms, the EM algorithm was the best performing one. However, three classical classification algorithms (SMO, BayesNet and NaiveBayesSimple) had better performance in most of the analysed tests for this project, even though with very close results. Therefore, it is possible to conclude that both techniques performed similarly. This type of work is very valuable since the project that was developed on this document contains different data sources.

Ai & Laffey (2007), did research in order to understand how Web usage mining can be used in Course Management Systems (CMS). CMS are software applications used frequently to implement e-learning in higher education (Blackboard for example). The main goal of this research is to show how web mining can be helpful to optimize the e-learning process. Some examples of this improvement can be by understanding the learner behaviour, test the effectiveness of these platforms in order to adapt better to the student's activity and provide quantitative feedback to teachers about the outcomes of their activity through text mining of e-mails, chats, forums and others. For this work, the researchers used the platform WebCT. The population for this research was of 748 undergraduate students in a large enrolment blended course (face to face and online) of a research university in the Midwest. The process started by selecting data from the user's profile, variables like student ID, gender, or academic level. The collection was also made with usage data (IP address, page reference, time of access and others) and structure data (hierarchy of web pages). This process was followed by data cleaning, elimination of all the entries of the images as well as entries with HTTP status code 404

"resource not found". Once this step was completed, the data was formatted so that can be used by mining application. The researchers used the classification technique to identify students' performance and use it to help possible students with low grades by giving them specific recommendations to solve this problem. In this study, it was used the decision tree algorithm C4.5. To label the different grades, it was used three different classes, good, medium, and poor. This was made with one week, two week, three week and one month log data. In the end the results gave a growing accuracy with 72,2% in one week data but 73% in one month data. This research shows that web mining is a real possibility for future educational research and that can be a real option to get educational information from CMS. However, further work needs to be made to generate models with higher accuracy for students' performance.

Maia et al. (2018), proposed a Web Intelligence System (WIS) in order to study and investigate some education variables like the growth of retention and grades drop in the Portuguese higher education system. For this study, the researchers selected a sample of more than 133 students, that attended a discipline of Web Programming, in the course MiEGSI (master's degree in Information Systems) in University of Minho and it was conducted from September 2017 to February 2018. The researchers used an application that was utilized by students and teachers during the classes called "ioEduc". This application supplied the data that was presented and analysed in this research. Some of the used variables from the application database were gender, name, grades, name of discipline, retention, success, questionnaires, exams, application visits and dates, and others. After cleaning, transforming, and analysing data, the researchers were able to conclude some things about the object in study. Firstly, it was analysed that the total number of application visits was 21349 and the number of distinct students, that access the application, was 149. During these visits, the month with the most application visits was October, while the least visited was February, conclusions that make sense since the distribution of classes follows the same path. It was also possible to confirm that most of the enrolled students were full time and first registration students (only 5 students were part time and only 12 students were repeating the course, in a total of 133). Another important factor was that 89,5% of the students enrolled in this course completed it with success while only 10,5% failed in this task. The grades distribution showed that 41% of the students had a grade of 17 or higher, while 35% had 13 or lower. Finally, when analysing the failed students and their number of absences, it was possible to see a direct correlation between these two factors, from the 14

failed students that went to at least one class, almost every one of them has at least one absence in every week of studies. These results show that absences are one of the biggest factors of retention and the importance of being in classes in order to understand and collect all the information given. This study concluded that Web Intelligence (WI) systems could influence Education, in terms of learning and teaching. This dissertation aims to continue this study, expanding it to data mining techniques and predictive models.

Chapter 3 – Methods and Tools

Considering the complexity associated with the development of a dissertation project, it is recommended to use methodologies or provide a set of principles and good practices for a specific scope. For the realization of this dissertation, 3 methodological approaches were followed, the "Design Science Research" as a scientific methodology, the "CRISP-DM" for the practical component and "SCRUM" as a methodology to manage the project in an agile way. In this section the three selected methodologies will be addressed as well as every tool used for the prototype development ahead.

1. Design Science Research (DSR)

According to Peffers et al. (2007), the DSR incorporates the principles, practices, and procedures necessary to carry out research projects. When using this methodology, is important that the research ends with the creation of a relevant artifact for an unsolved situation and it must be relevant for the business problem defined earlier (Peffers et al., 2007). Figure 17 presents this methodology and its different phases.

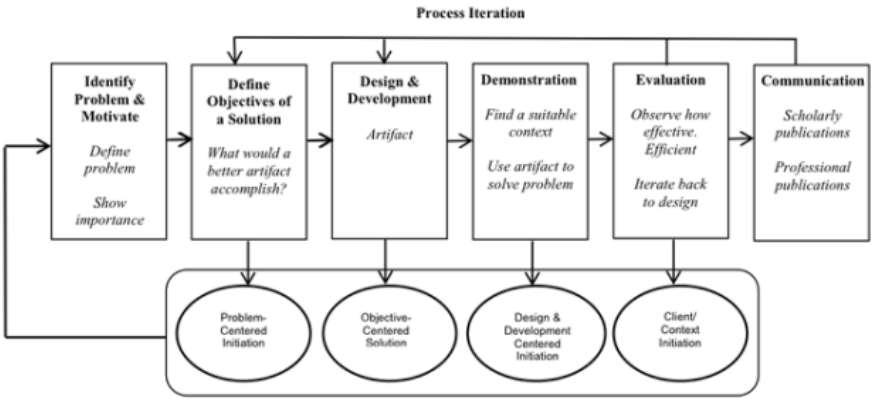

Figure 17 - DSR methodology process model (Adapted from (Ken Peffers, 2007))

Peffers et al. (2007) divides the DSR process into six different activities:

- **Problem Identification and Motivation**: in this first step the goal is to clearly define the specific research problem related to the project and justify the motive for the solution. In order to accomplish this, it is required to research the actual state of the problem, the progress that was made to solve this question, and recognize the importance of this project in the area. In this document, the motivation for this

project is clearly reflected in the first subchapter of the first chapter of the present document where it is explained that the main motivation for this project was to create a tool that can help the education sector and the students themselves to upgrade their methods by using data mining tools and techniques in educational data.

- **Define the objectives for a solution:** Once the problem is identified, the researcher should be able to define achievable goals for the project. These goals can be quantitative or qualitative, being found through the problem definition. In this document, the goals are presented in the second subchapter of the first chapter of the present document and it consists of the development of a prototype that can process and analyse education data in order to help and decision making, through both predictive and analytical features.;
- **Design and development:** This step include the creation of an artifact. This topic also includes the theoretical research made for the development of the artifact, also known as literature review. In this document, it is possible to check the prototype development on the fourth chapter, where it can be found the several steps, from the data collection and preparation to the analytical and predictive modelling phases. The literature review can be found in the second chapter, where different topics were explored, like data mining, big data, SQL & NoSQL databases and also research about work that was already made in the area.
- **Demonstration:** This phase is about testing the functionality of the solution in some proper context. Such demonstration could be achieved through experimentation, simulation, case study, proof, or other appropriate activity. In this document, the demonstration can be found on the fourth and fifth subchapter of the fourth chapter, in the modelling and evalution phase, where a group of dashboards that translate the reality from the course being studied and also data mining models from classification and regression techniques were made using real data from ioEduc in order to predict final grades from students in the future.
- **Evaluation:** In this step, the researcher needs to be able to observe and measure how well the produced artifact supports the solution to the problem. This can only be made by checking if it achieves the predefined goals and if satisfy the necessary

conditions to its validation. In this project, this phase can be found in the sixth subchapter of the prototype development chapter, where it's the results achieved in previous phases are measured and it is explained that the analytical modelling was able to create valuable conclusions that will help in decision making of the people in charge and also a success in terms of data mining modelling, regarding the regression technique.

- **Communication:** The final step consists of the communication of the problem and the importance of the artifact, its usefulness of the and effectiveness for researchers and other audiences. It is considered a common structure for empirical research work, where communication requires knowledge of culture discipline. In this case, the prototype is presented on a scientific article, in this dissertation report and it will be again in the presentation of the dissertation.

2. CRISP-DM

In this section it will be explained the methodology that was followed in the development phase that is included in this project. Both the data mining process and the educational data mining process, that were explained in the Literature Review chapter, follow this methodology and its different phases.

Cross Industry Standard Process for Data Mining (CRISP-DM) can be described as a methodology or as a process model. Since it includes descriptions of the typical phases of a project, the tasks involved with each phase, and an explanation of the relationships between these tasks, it can be named a methodology. However, CRISP-DM provides an overview of the data mining life cycle making it a process model as well.

According to Wirth & Hipp (2000), the life cycle of a data mining project is broken down into six different phases. The order between them is not strict and the arrows indicate only the most important and usual dependencies between phases. Figure 18 represents the life cycle of a project.

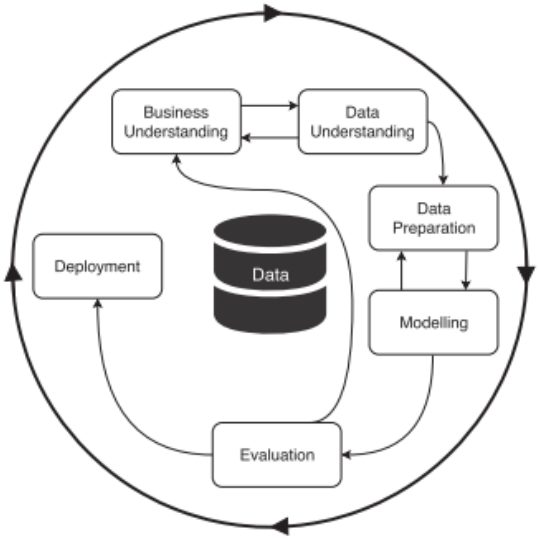

Figure 18 - The CRISP-DM process model of data mining (adapted from (Martínez-Plumed et al., 2019)

The six phases of this methodology can be described as:

- **Business Understanding:** In this phase, the focus is to thoroughly understand, from a business perspective, what the customer really wants to accomplish. The team should be especially careful with the constraints and factors that can be key to the final product. After understanding the main goals and objectives of the customer, the team should be able to translate and convert them to a data mining problem and develop a primary plan to solve it. This phase is extremely important because it can compromise the whole work if not done correctly. In the end, the team should have a project plan ready where all the steps that will be followed during the project are written and specify the tools and techniques that will be used during its time. In this project, the business understanding is the first subchapter of the Prototype Development phase, where it was explained the goals and characteristics of the data that was being used in the project. It was also created a plan in terms of software architecture to explain how each phase was divided.
- **Data Understanding:** This phase starts with the acquisition of the data that will be used in the project (data collection). During this phase, the team starts to get familiar with the data, understand its potential and its problems. After this, there should be a data report with details about the data description, data quality, and

with the correlation between various attributes, how will they connect to accomplish the goals of the project, and if new data sources will be needed to complement this work. The data understanding phase has a close link with the first one (Business Understanding) because, in order to fully determine the goals for the project and determine its plan correctly, the team needs to have some knowledge of the data he is about to work with. In this document, the data understanding phase is the second subchapter of the 4th chapter, started by explaining the data source and how is the data structured. It was followed by a description of every table that was used in the project and their attributes. Once every table was described, it was made a small first exploring of the data, some information that was already possible to get and finally, it was developed an entity and relationships diagram as well as multidimensional model.

- **Data Preparation:** During this phase, several tasks are in place to "create" the final dataset that will be used for modelling. Firstly, the team selects the data that will be used for analysis, this can include the selection of only a part of the attributes (columns) or/and the rows in a table. After the selection of the data, the team will clean some of it, this may involve the selection of clean subsets of the data or estimating some missing data that will in order to achieve the data quality needed for the modelling tool that was chosen before. Once this is done, the team focuses on the data construction and integration, this is the creation of derived attributes (new attributes that are constructed from one or more existing attributes in the same record) and merging data from different tables creating new information that will be needed in future tasks. Finally, in order to make the data completely ready to model, the team should adapt the dataset to the modelling tool details like changing the order of the records or having an identifier field for each record. After every step of this phase is completed, the data is ready for the next one (Modelling). In this dissertation, this phase can be found as the third subchapter of the fourth chapter and it consisted of attribute correction (encoding and filtering) and creation of new attributes for the modelling phase.
- **Modelling:** In this phase, the team starts by choosing the right modelling technique. This can be the same that was chosen during the Business

Understanding phase, or the team could select a new one if the context changed and other options are better for this task. Typically, there are several techniques for the same data mining problem type which means that more than one modelling technique can be used. After the team runs every model, they should be able to rank the models according to the accuracy and generality of the models. There is a close link between the last phase (Data Preparation) and this one. Often, one realizes data problems while modelling or one gets ideas for constructing new data. In the end, there should be a ranking of the models that were run in order to advance to the next phase. In this project, this modelling phase is divided into two different approaches, analytical and predictive. The first consists of the definition of a layer that helps to visualize data from the data warehouse and creation of indicators that will help decision making in the business. On the other hand, the second phase consists of the prediction of both label and continuous variables through different data mining algorithms, with the stipulation of several metrics to evaluate them. This can be found on the fourth subchapter from "Prototype Development".

- **Evaluation:** After selecting the best models in the previous phase, the team should now evaluate if the model meets the business objectives and determine if there is some business reason why this model is deficient. After approving one (or more) models, there needs to be a review of the data mining engagement in order to determine if there is any important factor or business issue that has somehow been overlooked or not been sufficiently considered. Once the models and the review are approved, a decision on the use of the data mining results should be reached. The team decides the future steps, which can be going to the next (and last) phase or going for new projects depending on the available resources and budget available. In this project, this phase is also divided into two different groups (like the modelling) and can be found on the fifth and sixth subchapter of the fourth chapter of the document. Here is when the dashboards that were produced before are shown and the main conclusions that is possible to take from them. In terms of predictive evaluation, the best models are shown in terms of metrics score and a group of conclusions in terms of their performance was made.

- **Deployment:** The creation of the model is generally not the end of the project. In order to present the results to the client in a way that they can easily understand, a deployment strategy should be put in place. This should also include the monitoring and maintenance strategy if the result will be a part of the client's business every day. Depending on the requirements, the deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process. Once the final deployment report is complete and all the deliverables are given to the client, the team should do an inside report that assesses what went right and what went wrong, what was done well, and what needs to be improved. Finally, the deployment phase includes the different considerations and future work that can be found on the fifth chapter of this document.

3. SCRUM

To organize the work that was done, it was decided that following the SCRUM framework would be the best way to do it.

SCRUM is a "framework for developing, delivering, and sustaining complex products" (Schwaber & Sutherland, 2017). With SCRUM the tasks are divided into Sprints. According to Schwaber & Sutherland (2017), a sprint is a time-box of a short amount of time in which a ready, useable, and potentially releasable product Increment is created. Sprints have consistent durations during product development. A new Sprint starts immediately after the conclusion of the previous Sprint.

The tasks needed for the project can be defined by the SCRUM backlog. Schwaber & Sutherland (2007), refer to the product backlog as a list of what's required, ranked in order of value to the customer or business, prioritizing the highest value tasks at the top. The Product Backlog evolves over the lifetime of the project, and items are continuously being moved by priority or even deleted or added.

This methodology is composed of three main components: the team, the events, and the artifacts. Figure 19 gives an overview of the whole SCRUM process.

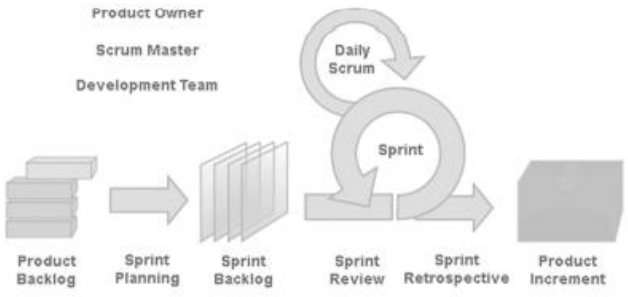

Figure 19 - SCRUM Model

The scrum team is usually composed by three different roles. Firstly, the product owner is the main person responsible for the project that will be held accountable in the end. The product owner is the sole person responsible for managing the Product Backlog. This person's decision must be respected and followed. Then, there are the people responsible for working on the product and delivering a complete one in the end. These people are called the development team and can have different sizes, whatever makes them functional and self-organized. Finally, the last role on the team is the scrum master. This person is responsible for making sure that everyone understands the Scrum process and its rules. This person helps those outside the team understand which of their interactions with the persons from the team are helpful and which aren't.

All the process happens with several events through it. As explained before, the tasks are divided into sprints. However, before the sprint, the team needs to make the sprint planning. This event corresponds to elaborate work to be performed in the sprint ahead. This event sets the sprint goal, what and how will it be done. It consists of a meeting with everyone from the team where the scrum master makes sure everyone understands its purpose. During the work, there is an event that happens daily with the development team to optimize team collaboration and performance by inspecting the work since the last meeting and to make sure that everyone is updated with the work. This event is called daily scrum and consists of a 15-minute meeting every day at the same place and time. Then appears the sprint review and retrospective. This is the process that includes a meeting to conclude the present sprint. All these meetings are supposed to be something informal with the purpose to increase the work quality each sprint.

Finally, the last component of the SCRUM process is the artifacts. Here is included the Product Backlog, that, like was explained before, is a list of requirements, ranked in order of

value to the customer or business, prioritizing the highest value tasks at the top. The Sprint Backlog is an artifact that consists of a list of the several sprints and their list. The Sprint Backlog is a plan with enough detail that changes in progress can be understood in the Daily Scrum (Schwaber,& Sutherland, 2017).

This project was divided into seven different sprints with a duration of two weeks each. The roles are divided in the following way:

- Product Owner – IOTech;
- Development Team – Francisco Campos;
- SCRUM Master – Professor Filipe Portela.

The table 3 represents the product backlog of this project, and as explained previously, it consists of a list composed by the requirements asked by the client, the priority and the effort of each one. The values on these columns go from 1 to 5, on a growing scale, and also by a column stating if the requirement was completed or not.

Table 3 - Product Backlog

ID	Tasks	Priority	Effort	Completed?
1	Data Collection	4	2	Yes
2	Data Analysis and Selection	3	3	Yes
3	Data construction	3	4	Yes
4	Multidimensional Model	3	3	Yes
5	Different data models	4	4	Yes
6	Relevant dashboards	4	3	Yes
7	Functional Prototype	5	5	Yes

The table 4 represents the sprint backlog of this dissertation. Each sprint had an estimated duration of two weeks each. These sprints are connected to the development phase, however this division was not strict since this is an agile project that could have changes in time and tasks during the development process. The column Tasks completed represents the requirements from the product backlog that were completed during each sprint.

Table 4 - Sprint backlog

Work Phase	Sprint	Backlog task	Start Date	End Date
Business understanding	Sprint 1	Data Collection	15/2/22	1/3/22
Data understanding	Sprint 2	Data Analysis and Selection Data Construction	2/3/22	16/3/22
Data preparation	Sprint 3	Data Analysis and Selection Data Construction Multidimensional Model	17/3/22	1/4/22
	Sprint 4	Data Construction	2/4/22	16/4/22
Modelling	Sprint 5	Different Data Models Relevant Dashboards	17/4/22	31/4/22
Evaluation	Sprint 6	Different Data Models Relevant dashboards	1/5/21	15/5/22
Deployment	Sprint 7	Relevant dashboards Functional Prototype	16/5/22	30/5/22

4. Project Tools

This section presents the tools that were used during the practical phase of this dissertation, since the programming languages to the data mining algorithms used for data modelling.

Firstly, as it was said in the first chapter, this project is included in the IOScience project, that consists of a data mining system that is able to create both predictive and analytic models. This system analyses and processes data using the language **Python** which was the

programming language used through this project. This was the chosen language, mainly because of the number of libraries available for it that assist developers in several tasks, such as Pandas for data processing and modelling. For this project, the main library used was the **Pandas** library. According to Chen (2017) "Pandas is an open-source Python library for data analysis". This package allows Python to work with data like a spreadsheet increasing the speed of different operations like data loading, manipulation, merging, and others (Chen,2017).

In order to get the information out of the analysed data, some data mining algorithms were put in place during the modelling phase. Some of these algorithms were:

- **SVM**: According to Nikam (2015), SVM is one of the most used data mining algorithms in the last decade and is typically used for classification and regression problems. The SVM algorithm goal is to determine the location of decision boundaries also known as hyperplane that produce the optimal separation of classes (Nikam, 2015).
- **ANN**: Artificial Neural Networks are a data mining algorithm that is used to approximate functions that can depend on a large number of inputs and are generally unknown. This algorithm is usually used in machine learning and pattern recognition because of their adaptability. (Nikam,2015)
- **NB**: The Naive Bayes Classifier is often used "when the dimensionality of the inputs is high" (Nikam,2015). This algorithm characterizes for assuming the independence of attributes between each other.
- **RF**: The Random Forest algorithm is a statistical method, based on decision trees, for both classification and regression. This algorithm consists of the extraction of multiple slices from the original sample, modelling the decision tree for each slice, with the goal of combining the predictions of the different decision trees and get the average score of the forecasting (Zhang et al., 2021).
- **DT**: Decision Trees are an algorithm that follows a tree structure with nodes and branches. Each one of the nodes represents an attribute from a category to be classified and each branch represents a different value that can fill the node (Mathew et al., 2017). It can be used for classification and regression problems.

This project includes regression and classification modelling, consequently, not all the algorithms were used for both. Table 5 represents which algorithms were used in each of the techniques.

Table 5 - Distribution of Algorithms

	Regression	Classification
SVM	X	X
ANN	X	
NB		X
RF	X	X
DT	X	X

In this table it is possible to understand that the 5 algorithms were divided in the two techniques used. The SVM, RF and DT algorithms were used in both, classification and regression modelling. On the other hand, the ANN algorithm was uniquely used in regression, while the NB was just used on classification modelling tasks.

Chapter 4 – Prototype Development

During this chapter, every step of the prototype development is described in order to reach the final goal. This description is based on the main knowledge extraction process that was used in the project: Data Mining (DM). All the work in this area followed the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology.

1. Business Understanding

Inside this section it is possible to find the main foundations for this project, starting by explaining the environment around it, the business goals that come with it and the first plan in terms of software architecture for the prototype.

1.1. Business goals and environment

The theme for this dissertation was selected and chosen to study the benefits and the impact of e-learning in students and how the usage of new tools and new digital ways of learning could create a better studying environment that is able to help students reach their goals in a better and easier way. This project was conducted in the Web Programming course, part of the Integrated Master's degree in Information Systems Engineering and Management so that it can be tested and used in different contexts if proven beneficial. From this point of view, in order to achieve the proposed goals, the definition of a set of business requirements was distinctly discussed with the different parts that interact in the project and compared with similar work to understand the focal points and the biggest needs during the whole process. Relative to the project plan, it was developed and implemented in accordance with the proposed methodologies and considering the time space scheduled for the development of this dissertation.

1.2. Software Architecture

During the creation of a robust technological architecture, it's key to be aware of the business requirements and the technologies that can be used to fulfil them. This process usually includes understanding the environment where the prototype is being made as well as the previous strategy that was decided with the client. At this stage, the main objective is to develop a simple framework that allows the integration of various technologies from different sources and with different functionalities.

Figure 20 represents the architecture of this protype. It starts by collecting the data from the "ioEduc" platform. This data was stored in MySQL and MongoDB databases. Once the data is collected, it starts the process of data extracting, transformation and load (ETL) to different dimensions and facts tables that make the data ready and available to analyse and apply machine learning algorithms. Every step of the way was made using the Visual Studio Code characteristics that made it the best solution for it, as well as the Python coding language. Once this process is completed, the results are represented in a simple manner for every user to understand it. The visualization section was also made using Visual Studio Code with python libraries like Seaborn and Matplotlib.

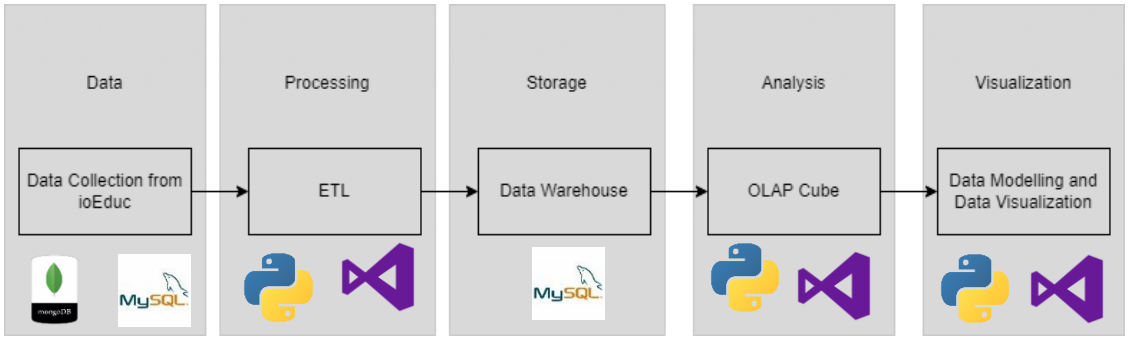

Figure 20 - Technical Architecture

2. Data Understanding

The understanding the data section represents, for the development of the prototype, a great importance, because it provides the project the necessary capacity to correspond to the objectives set. The data used in the development of the prototype was generated by the use of the "IoEduc" application in the classroom, by students and teachers of the discipline of Web Programming, in the 2020/2021 school year (more precisely in the first semester ranging from September 2020 to February 2021). This data includes the evaluation process as well as the day-to-day records, class by class. Some of this data is structured in tables, within a database (relational) called "webitclo_ioeduc". This database has 57 tables and contains almost all the data. However, not all of these tables were used, only the ones that are the most relevant for this project and for this theme. The selected tables made a total of 11 different tables (19% of the total tables) and were the following:

- assessments;
- dates;
- evaluations;
- goals;
- logins;
- moments;
- penalties;
- redeems;
- users;
- classes;
- user_classes;

Each of the attributes from this tables will be described below, during this section.

Besides these tables, it exists another part of the data about the quizzes that were made each class (records of each question and answer), that are stored in json files in a MongoDB database (non-Relational database) that contain 3 different collections where it will be used only the file "ioquiz_answers", described in the section 3.1, in this chapter.

The table 6 represents the attributes of the table "assessments", where it's stored the evaluation given by group members to each other during the work year, that were selected for this project.

Table 6 - Table "assessments" description

Attribute	Type	Description
Id	VARCHAR	Assessment identifier
Assessment_author	VARCHAR	Assessment author identifier
Assessment_target	VARCHAR	Assessment target identifier
Moment_id	VARCHAR	Evaluation moment identifier
Group_grade	FLOAT	Grade if the author and target group
Target_grade	INTEGER	Grade given by the author to the target

Attribute	Type	Description
Not_work	FLOAT	Option if the target made no work at all

Table 7 shows the attributes related to the table "dates" where all the information about dates during the year is stored.

Table 7 - Table "dates" description

Attribute	Type	Description
Id	INTEGER	Date identifier
Db_date	DATE	Date in date format
Year	INTEGER	Year of the date
Month	INTEGER	Month of the date
Day	INTEGER	Day of the date
week	INTEGER	Week of the date

Table 8 represents the attributes in the table "evaluations". This table contains all the information about evaluation moments during the year.

Table 8 - Table "evaluations" description

Attribute	Type	Description
Id	VARCHAR	Record identifier
Moment_id	VARCHAR	Evaluation moment identifier
User_id	VARCHAR	User identifier
classification	FLOAT	Grade achieved

Table 9 describes the attributes in the table "goals", which contains the records for benefits given to students during classes.

Table 9 - Table "goals" description

Attribute	Type	Description

Attribute	Type	Description
Id	VARCHAR	Record identifier
User_id	VARCHAR	User identifier
Class_id	VARCHAR	Class identifier

In table 10 is shown the attributes of the table "logins", where all the information about every login in the platform is stored.

Table 10 - Table "logins" description

Attribute	Type	Description
Id	VARCHAR	Record identifier
User_id	VARCHAR	User identifier

Table 11 represents the attributes of table "moments", which contains all the information about the different evaluation moments, that were used in the following sections.

Table 11 - Table "moments" description

Attribute	Type	Description
Id	VARCHAR	Moment Identifier
Name	VARCHAR	Name of the moment
Is_redeemable	INTEGER	If this moment is redeemable

Table 12 describes the attributes in the table "penalties", which contains the records for penalties given to students during classes.

Table 12 - Table "penalties" description

Attribute	Type	Description
Id	VARCHAR	Record identifier
User_id	VARCHAR	User identifier
Class_id	VARCHAR	Class identifier

In table 13 is shown the attributes of the table "redeems" that contains the records of the students that asked for the rescue option.

Table 13 - Table "redeems" description

Attribute	Type	Description
Id	VARCHAR	Record identifier
User_id	VARCHAR	User identifier
Moment_id	VARCHAR	Evaluation moment identifier

Table 14 represents the attributes of the table "users", that contains all the information about the users of the platform, that were used in this project.

Table 14 - Table "users" description

Attribute	Type	Description
Id	VARCHAR	User identifier
Regime	VARCHAR	Regime of the student
Student_type	VARCHAR	Statute of the student

.

Table 15 describes the attributes in the table "classes", which contains the information about every class during the semester

Table 15 - Table "classes" description

Attribute	Type	Description
Class_id	VARCHAR	Class identifier
Class_type	VARCHAR	Type of class
Subject	VARCHAR	Subject of the class
Date_ID	VARCHAR	Date of the class
Year_studies	VARCHAR	Year of studies
University	VARCHAR	University where the class is given

In table 16 is shown the attributes of the table “user_classes” that contains the records of the students that were present in each class.

Table 16 - Table "user_classes" description

Attribute	Type	Description
Userclasses_id	VARCHAR	Record identifier
User_id	VARCHAR	User Identifier
Class_id	VARCHAR	Class identifier

2.1. Data Exploring

Exploring the data, although introductory, is an important part in the development of data understanding, since it makes possible to produce conclusions that will help during the development of the prototype.

In this first analysis it was possible to make considerations at the student’s profile, their use of the "IoEduc" application, their behaviour in classes etc.

Firstly, it was made a study on how a total of 166 students were divided in terms of statute. Figure 21 represents, as expected, a vast majority in full-time students (E), which means students that have studying as their solo occupation, with 149 (92%) as their statute. It was also possible to know that 2 students had a physical disability statute (EPD) while 1 was the class delegate (DLG) and another one a Student Association Member (EDAEE). On the other hand, 13 (7%) students were student-workers (TE), a statute given to students that are studying and working at the same time.

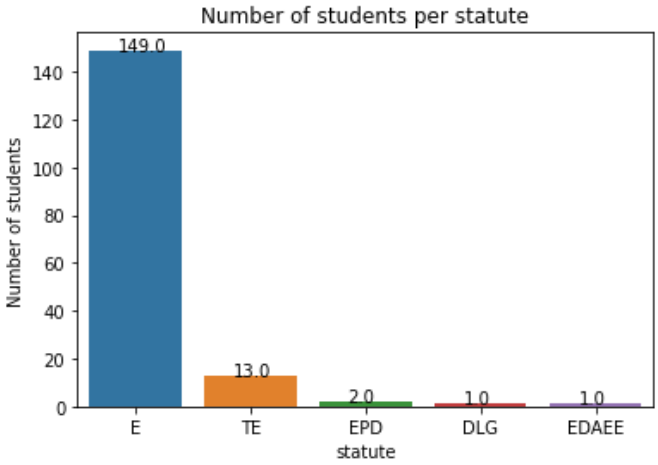

Figure 21 - Number of students per statute

In figures 22 and 23, it is possible to analyse the impact of the card system in class. The card system is a system where the teacher gives white (positive) or yellow (negative) cards to the students. When the limit number of positive or negative cards was achieved, consequences would happen. In this case, 2 positive cards would give a blue card that translates into an extra point in the group project. On the other hand, 3 negative cards (or penalties) result in a red card that translates the student having a grade of 0 in the group project. In this case, 6 students received one positive card and 5 students received the full 2 positive cards which resulted in a blue card. In terms of penalties, 7 students received a single warning, 1 student 2 warnings and finally, 3 students received the limit of 3 warnings what resulted in a red card and consequent failure in group project.

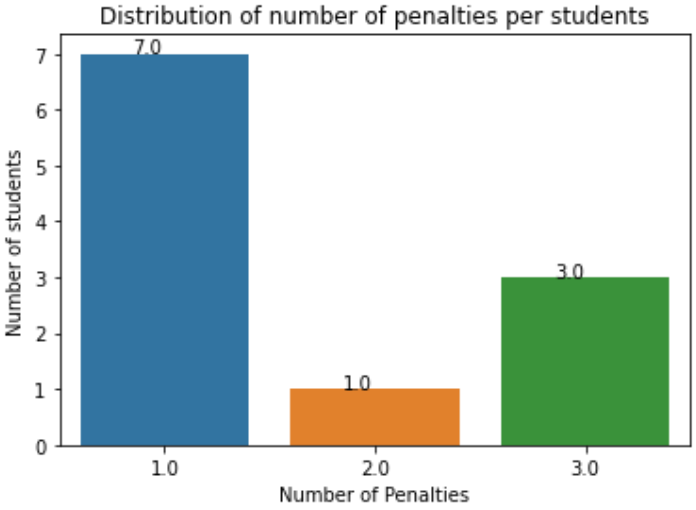

Figure 22 - Distribution of penalties per students

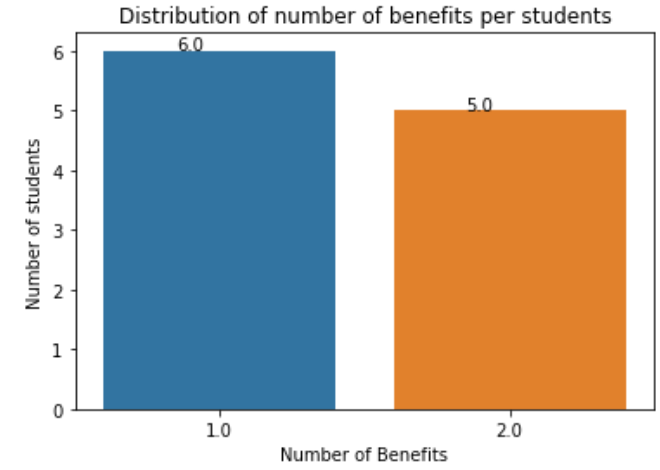

Figure 23 - Distribution of number of benefits per students

The other system that was in place was the rescue system. The rescue system was available to failed students (grade = −1) on the MT2 moment who thought they deserved more. The professor analysed the situation, and, in the case of acceptance, they allowed them to continue with a penalty of fifteen per cent (15%) in the final MT grade. Figure 24 represents the 11 students (6%) who used this option.

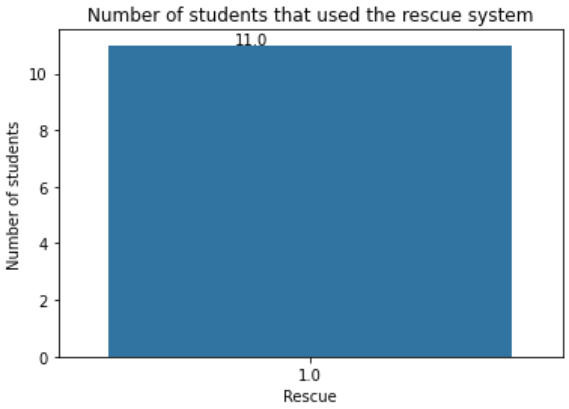

Figure 24 - Students that used the rescue system

2.2. Multidimensional Model

The multidimensional scheme represents how the data warehouse is designed. Figure 25 represents the entity and relationships diagram (ERD) with the designations of the 11 tables (19% of the total number of tables), the relationship between them and their attributes.

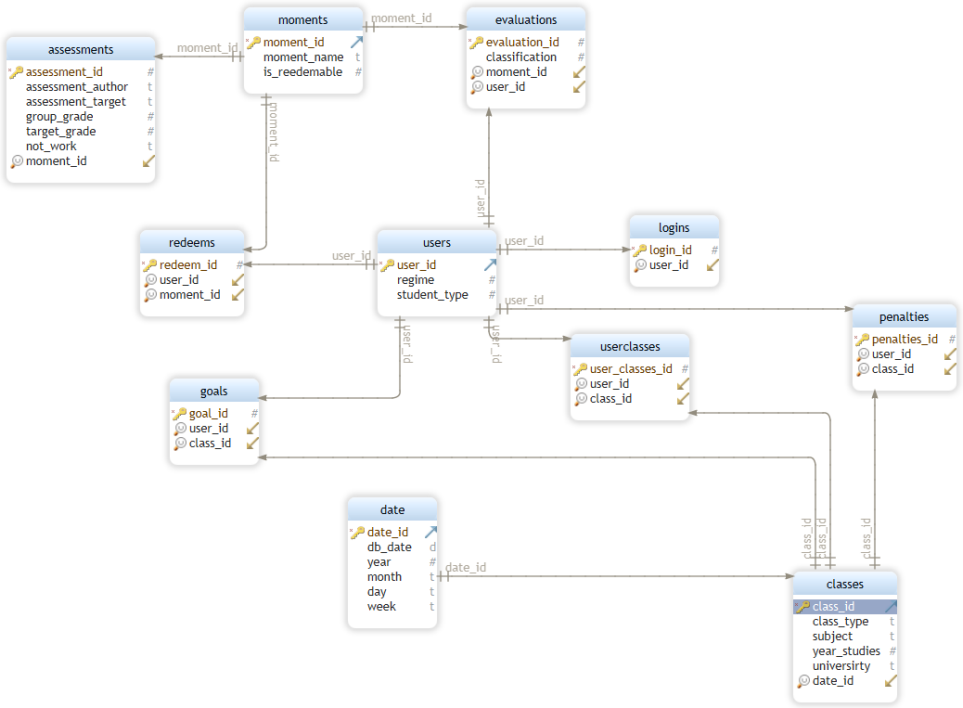

Figure 25 - ERD Model

After building the ERD model, a data warehouse creation was needed, where it was possible to organize the data in the best way possible to accomplish the analytical and predictive goals. Firstly, it was set a group of business requirements that guaranteed that every area of the targeted data was included. Once this phase was set, it was defined that 5 different dimensions were needed for this project:

- Time – Dimension that allows to control every record temporally;
- Course – Where all the data about classes will be stored;
- Assessment – Where every assessment will be stored-,
- Student – Where student personal data will be stored;
- Moment – Where information about evaluation moments are stored;

After deciding the dimensions that would be a part of the data warehouse, it was also decided that this model would have two facts table:

- Attendancefacts – Table that receives the students and classes primary keys and stores every attendance related fact, like logins, presences, number of benefits and penalties.
- Evaluationfacts – Table that receives the student, moment and assessment primary keys and that will store every evaluation related fact, the grade in each evaluation moment, every assessment made to the students or if it applied to the rescue system.

Figure 26 represents the multidimensional model that resulted from this process.

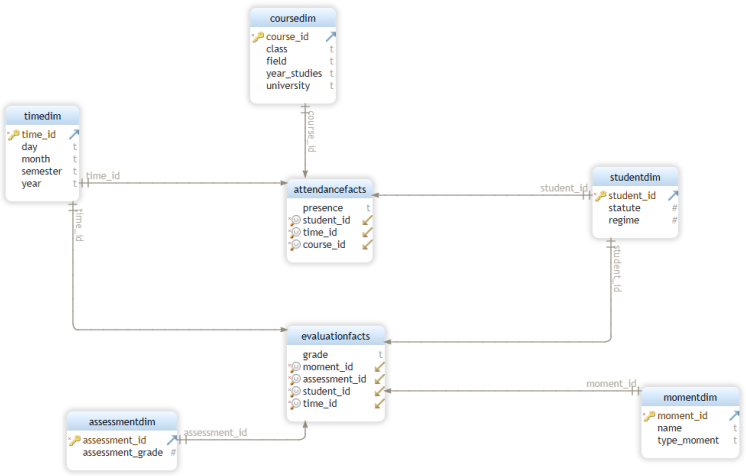

Figure 26 - Multidimensional Model

Analysing figure 26, it is possible to see that the data warehouse is designed in a snowflake model that contains two fact tables and five dimensions. The fact tables, contain the foreign keys, which come from the dimension tables. Some new attributes were calculated for future tasks, but in order to make the model easier to read, this information is not represented here but will be available in section 3.2.

2.3. Fact Tables and Dimensions

Table 17 represents the attributes of the facts table "AttendanceFacts". This table includes the foreign keys of three different dimensions, user, course, and time. This dimension stores all

the information about a class, their presences, if a student received a benefit or a penalty, as well as the number of logins of each student.

Table 17 - Facts Table "AttendanceFacts"

Attribute	Description
Presence	Check if the user was present in the class
Time_id	Date of the class
Course_id	Course identifier
User_id	User identifier

Table 18 describes the attributes of the facts table "evaluationfacts". This table includes the information about every evaluation moment made by students, each grade in each moment and the assessments made by the working group to the student in case.

Table 18 - Facts Table "evaluationfacts"

Attribute	Description
Assessment_id	Assessment given to the student by his group
Moment_id	Evaluation moment of the record
Time_id	Date when the evaluation was made
User_id	Student that is being evaluated
grade	Grade given to the student

Table 19 shows the attributes for each dimension and their data type. Each dimension has a primary key (represented by "PK" nomenclature) that will create the bridge to a given fact table.

Table 19 - Dimensions

Dimension	Attributes	Data Type	Description	Dimension Type
UsersDim	User_id (PK)	INTEGER	Dimension that stores information about the students, their id, statute, and regime	Conformed Dimension
	statute	VARCHAR		
	regime	VARCHAR		

Dimension	Attributes	Data Type	Description	Dimension Type
MomentsDim	Moment_id (PK)	INTEGER	Dimension that stores the evaluation moment id, its name, and its type (quiz, MT, project)	Slowly Changing Dimension
	Moment_name	VARCHAR		
	Moment_type	VARCHAR		
TimeDim	Time_id (PK)	INTEGER	Dimension that stores the information about the date of each record, from day to month, semester, and year.	Conformed Dimension
	Day	INTEGER		
	Month	INTEGER		
	Semester	INTEGER		
	Year	INTEGER		
AssessmentDIm	Assessment_id(PK)	INTEGER	Dimension that stores information about assessments made, the id and the grade given.	Slowly Changing Dimension
	Assessment_grade	FLOAT		
CourseDim	Course_id	INTEGER	Dimension that stores information about each class, it's field of studies, year of studies and university where it was given.	Slowly Changing Dimension
	Class	VARCHAR		
	Field	VARCHAR		
	Year_studies	VARCHAR		
	University	VARCHAR		

3. Data Preparation

Data preparation includes the construction and integration of data for the next phase. For this, different type of processes was made for the correction and creation of attributes that suited the goal of this project. All the data was processed and transformed in order to answer the key performance indicators of this project.

3.1. Non-Structured Data

Firstly, it was needed to understand, from the quizzes made by the students, the theme that the students are more comfortable with and the ones that need to be better worked in the following years. For this, it was needed to analyse the database that contains these files (MongoDB). This information is stored on a JSON file called "ioquiz_answers". Table 20 represents the different attributes in this file where every answer given by every student in every quiz is stored.

Table 20 - IoQuiz file

Attributes	Type	Description
_id	STRING	Record Identifier
quiz_id	NUMBER	Quiz Identifier
Subject_id	NUMBER	Subject Identifier
User_id	NUMBER	User Identifier
Answers	OBJECT	Answer given information
Title	STRING	Title of the quiz
Description	STRING	Quiz description
Meta_tags	STRING	Tags for the quiz
Difficulty	NUMBER	Difficulty of the quiz
Questions	OBJECT	Questions information
timestamp	STRING	Date when the quiz was made

In order to get the quizzes that got the most answers given, it was made a projection to filter the title of the quizzes on each record, as well as their time. Following it by filtering to the dates that are being analysed.

3.2. Structured Data

For structured data, firstly it was needed the carry out a treatment to inconsistencies that were encountered in the tables. These inconsistencies were missing data and uncoded data. In table 21 is represented the attributes that had something missing, the problem and the solution found for it.

Table 21 - Inconsistencies Treatment

Attribute	Inconsistency	Solution	Affected records
User_id	User_id was not encoded	Turn the id into a hexadecimal number	166 (100%)
grade	Some grades were null or did not exist	Filtered out users that had null values in an evaluation moment.	13 (8%)

In addition to the treatment of the inconsistencies found and in order to match the multidimensional model presented before, 4 new attributes were created. In table 22 each of the 4 new attributes created is described.

Table 22 - New Attributes created

Attribute	Example	Action
semester	1	Column with the semester when the class is hapenning.
Moment_type	"MT"	Column with the type of evaluation moment
Year_studies	3	Column with indication of the year from the degree when the course is being taken.
University	Universidade do Minho	Column with the name of the

Attribute	Example	Action
		university where the course is being taken

4. Modelling

Education is a key aspect in the daily life of the population, making so important to understand how it functions and the behaviours related to it. In this sense, and according to the project goals, it was developed a prototype that will support decision-making by those involved in the educational system. Given the needs and in order to implement decision models in the prototype, this prototype has two modelling approaches: analytical and predictive. The first consists of the creation of an OLAP layer where it's possible to create dashboards to help decision making based on past experience. The second phase is where data mining algorithms were integrated into it, in order to create a predictive functionality that could help to predict future grades. That said, the goals for this phase consist of defining and evaluating dashboards that translate the reality of the course through different perspectives and then, models capable of predict the student's final grade (and subsequent approval or not to the course)

4.1. Analytical Modelling

The data processing and development of the data warehouse has provided the possibility for the creation of a visualization environment, with the aim of helping the understanding and analysis of the data by the user of the prototype. This environment is based on the data that comes from the fact's tables, dimensions and metrics that are contained in the OLAP cube.

Having said this and, in order to meet the different needs, a group of indicators was designed that facilitated the development of the viewing environment and that aim to answer for the main questions asked by the interested parts in this project. The table 23 describes the developed indicators.

Table 23 - Indicators

Indicator	Example	Action
Total number of goals	2	Sum of benefits given to a student.

Indicator	Example	Action
Total number of penalties	2	Sum of penalties given to a student
Redeem	1	Indication if the student applied to the rescue system or not
Total assessment grade	6	Sum of all the grades given to the student by his group during the evaluation moments
Total Logins	396	Sum of all logins made by student
Total Presences	31	Sum of all presences in classes by student
Total number of quizzes answered	11	Number of total quizzes answered by student
Final grade	15.130250	Final grade by student, calculated using a formula: 25% x grade_MTs + 15% x grade_quizzes + 60% x grade_project

4.2. Predictive Modelling

In this section, the predictive modelling phase will be explained, where two different approaches to the problem were made:

- **Regression**: In order to predict the student final grade, regression algorithms will be used since the variable in study is a continuous variable, through the creation of different scenarios with different features.

- **Classification**: Predict if the student final grade is inside a class, a certain range of grades, with classification algorithms through different scenarios with different features.

The features chosen for both classification and regression targets were:

- Statute (S) – Statue of the student;
- Presences (P) – Number of presences in classes by the student;
- Logins (L) – Number of logins made by the student in a platform;
- N_goals (NG) – Total number of benefits that a student got in classes;
- N_reedems (NR) – If the student applied to the rescue system;
- N_penalties (NP) – Total number of penalties that a student got in classes;
- N_quizzes_answered (NQ) – Number of quizzes answered by the student;
- Sum_grade (SG) – Sum of the grade given by the student group members in assessments during the year;

For both problems, 3 different scenarios were chosen from the list of features, described on table 24:

Table 24 - Modelling Scenarios

Scenario	Description	Features
A	Participation, gets the features from student participation in classes and platform;	NQ, P, L, NG, NP
B	Evaluations, gets the features from evaluation moments;	S, NR, NQ, SG
C	Combines the features from scenario A and B	S, NQ, P, L, NG, NR, NP, SG

Starting with the regression modelling, the first step in this process was to define a target and the features that will be used to calculate it. The target to predict is the **finalgrade**, a continuous variable that goes from 0 to 20 and represents the final grade of students.

The target distribution, in percentage, is represented in figure 27.

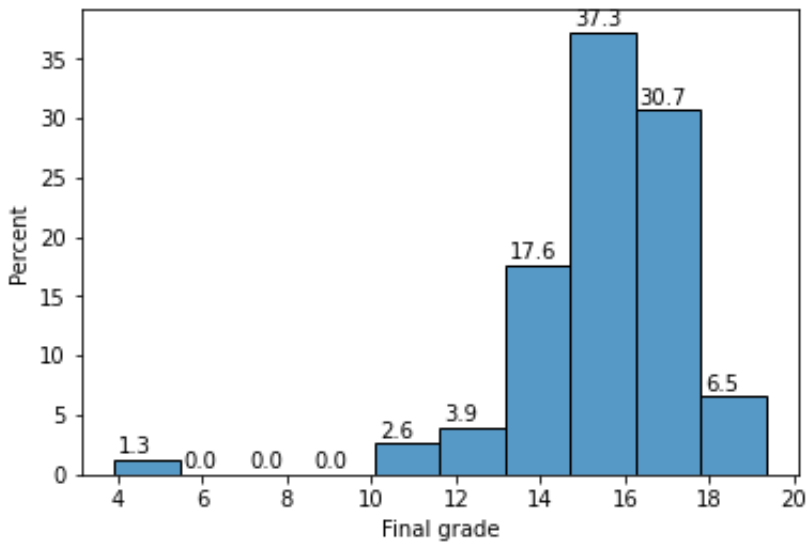

Figure 27 - Target distribution

Before starting the process, it was defined 4 different algorithms that will be used to predict our target, all of them described on the Tools section, and 3 different metrics, described here, that will make it possible to evaluate the success of the prediction, represented on table 26.

Table 25 - Algorithms and metrics used in regression

Algorithm	Metrics
Support Vector Machine (SVM)	• Mean Absolute Error (MAE) • Mean Squared Error (MSE) • Root Mean Squared Error (RMSE)
Artificial Neural Network (ANN)	
Random Forest Regressor (RF)	
Decision Tree Regressor (DT)	

In order to help to boost the performance of this prediction, it was used a cross validation technique, in this case **k-fold cross-validation**. K-fold cross-validation consists of dividing the dataset in different parts (folds), with the same number of observations. Then, one of these folds is a validator, and the rest is used as training sets. This iterates several times until every fold was a validator once. In the end, the mean values for the model performance in all iterations are given.

In this case, the process started by creating a variable X with all the features that were chosen, and a variable Y with the target. This process is followed by a data normalization of the features represented by the variable X, using the function MinMaxScaler. This process aims to

put every feature value inside a defined range (like 0-1) so that they all have the same range. After normalizing the data, a variable model is created with the intended algorithm. Finally, it was applied the function **cross_val_score** that has 5 different arguments: X (features), Y(target), model, 5 (as the number of folds that was used for cross validation) and a score argument with the metric that was calculated (MAE, MSE, RMSE).

After the regression modelling, the process followed to the classification modelling. In this case, the goal was to predict two different class labels in terms of grades. The targets for this is the column **higher15** and **lower15**. The column higher15 has a value of 1 if the grade is between 15 and 20, 0 otherwise. On the other hand, lower15 has value of 1 if the grade is between 10 and 15, and 0 otherwise. These targets were chosen in order to separate the positive results into different categories, from satisfactory [10-15[to very good [15-20] results.

The distribution for the target higher15 is represented on figure 28, with the distribution for the other class (lower15) being the opposite values. In this figure it is possible to see that the number of grades inside the class [15-20] is of 67,3%, about 2/3 of the data, while the other 32,7% have a grade of under 15.

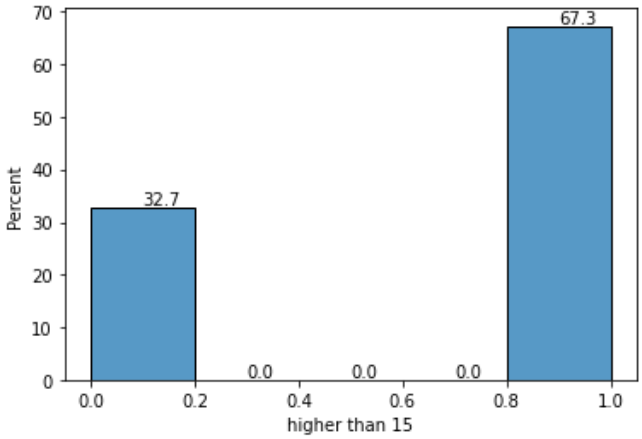

Figure 28 - Target higher15 distribution

After analysing these values, it was decided that this distribution is not perfect, due to the big number of grades higher than 15. On that note, adding to these targets, was also created targets for the classes [16-20] and [10-16[, labelled as higher16 and lower16, since the distribution of values is more balanced and can be used to compare with the results of the first classes. The distribution for these new targets is represented on figure 29. Analysing it, is

possible to see a bigger balance (56,9%/43,1%) between values, creating more distributed datasets to analyse.

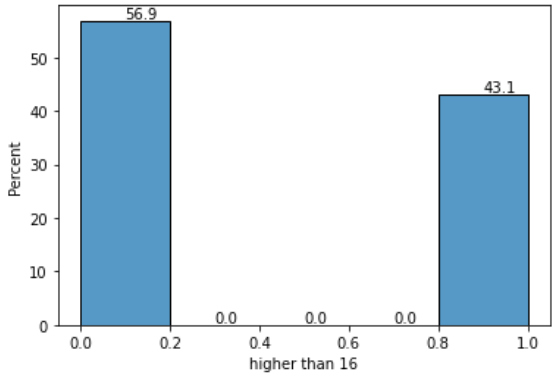

Figure 29 - Target higher16 distribution

In order to complete this process, it was defined 4 different algorithms that will be used to predict our target, all of them described on the Tools section, and 2 different metrics, explained here, that will make it possible to evaluate the success of the prediction, represented on table 27.

Table 26 - Algorithms and metrics used in classification

Algorithm	Metrics
Support Vector Machine (SVM)	• Accuracy (A) • Sensitivity (S)
Naïve Bayes Classifier (NB)	
Random Forest Classifier (RF)	
Decision Tree Classifier (DT)	

Repeating what was made in the regression modelling, this time it was also made a cross-validation technique. However, this time the data was also stratified before being split in different folds. This process means that in each fold created, the ratio between the target classes is the same in each fold as it is in the full dataset. This helps in creating a better and more real sample of the data for better predictions. This technique is called **Stratified K-fold cross-validation.**

This classification modelling started by creating a variable X with all the features that were chosen, and a variable Y with the target. Like the previous one, this process is followed by a data normalization of the features represented by the variable X, using the function MinMaxScaler. After normalizing the data, a variable model is created with the intended algorithm. Once this was

done, it was used the function stratifiedkfold, that split the data into 5 different folds with a correct divison in terms of ratio between target classes, like said previously. In each of these splits, it was created a training and testing set, and applied the modelling algorithm, predicting the target variable and appending the accuracy and sensitivity score of each iteration into a list. In the end, it was calculated the average value for each of the metrics and displayed as result.

5. Evaluation

In the next step, the results of the analytical and data mining models will be presented using dashboards and reports made during the process that can help in decision making and also graphics that represent the values for different data mining models applied in the previous phase.

5.1. Analytic Results

This chapter presents the most relevant reports prepared using the different attributes created before, in the data preparation section. These have as main objective an easier analysis of the data, providing a simple and effective way for everyone that is trying to study the problem in discussion.

First, it is important to understand how the number of presences in classes changed during the semester. For this, it was calculated the total presences in classes per week, this includes Theoretical (T), Theoretical practice (TP) and laboratory practices (LP) classes. In figure 30 it is possible to see that the first few weeks (3,4 and 5) were when the students most attended classes, with over 420 attendances registered in all of them, when the average number of attendances per week is 350. Since 3 different classes happen per week (T, TP and LP), it is important to notice that the maximum of presences per week that was possible to get was 459, 153x3 with 153 being the number of students. The first 2 weeks had also a low number of presences since the first week had just T classes, and the second week had just T and TP classes, with the LP classes only starting on week 3, as it can be seen through the big rise from week 2 to week 3. After this, it happened a drop during the semester, going up again in the second to last week. This makes sense because the end is when the results are discussed and shown, making sense that more people attended during this period. The final weeks (from 14 to 16) were occupied with exams which means that no classes were given in that period.

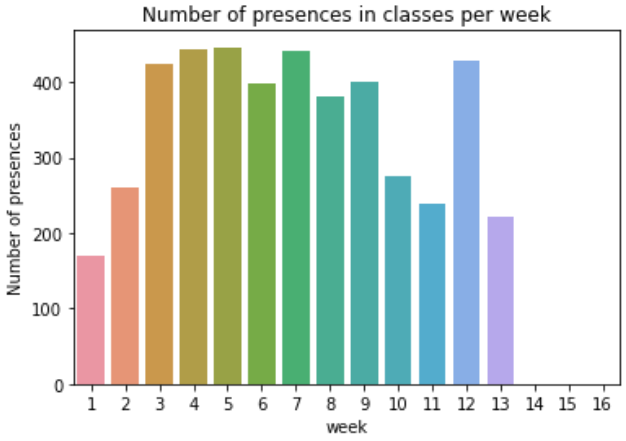

Figure 30 - Number of presences in classes per week

Doing the same analysis, figure 31 shows us the distribution of logins in the platform per week. The number of logins follows the same path as the number of presences with a big number of records (over 3000 when the average number of logins per week is 2371) in weeks 3,4 and 5 with a decrease in number after that. The first week and last weeks are also following the same pattern with a low number of logins on the platform, that can be explained by the start of the semester, with students arriving late, and the end of the semester when all the grades are already out and no work left, while the weeks with the highest number of logins are the ones where the evaluation moments were scheduled.

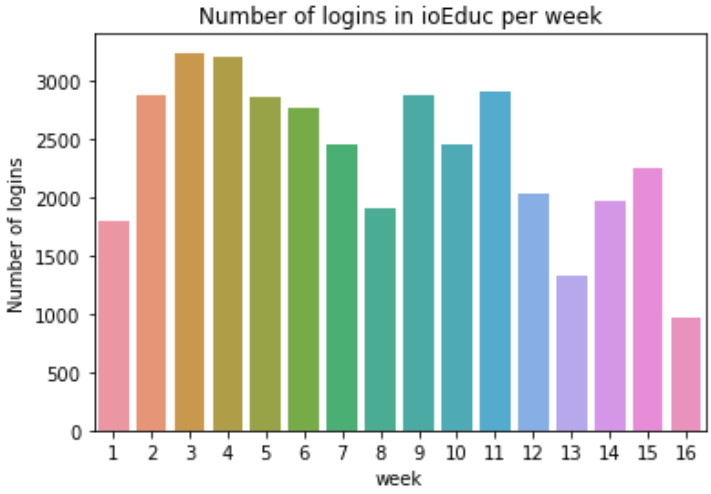

Figure 31 - Number of logins in the platform per week

After this analysis, it is important to understand if the number of logins and presences are co-related with each other. Figure 32 represents this correlation.

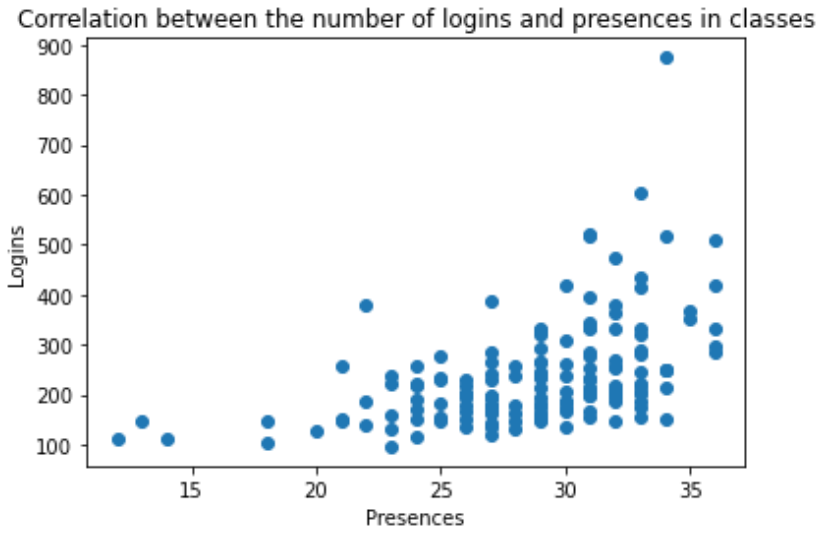

Figure 32 - Correlation between presences and logins

As it is possible to see, most of the students with a large number of logins in the platform, are also students with a high attendance rate to the classes. Adding to that, there is a big percentage of students that attended more than 50% of the classes, which also shows that the efforts made to improve attendance in classes is working. The maximum number of attendances is 36 which represents 100% of attendance rate, while the minimum is 12, which represents one third of the classes attended, meaning that every student attended, at least, 1 in 3 classes.

After evaluating the correlation between these two variables, it's important to understand how each of them affect the final grade of students, which is the most important aspect of this project. In figure 33 and figure 34 it is possible to check if there is any correlation between attendance and participation with the final grade.

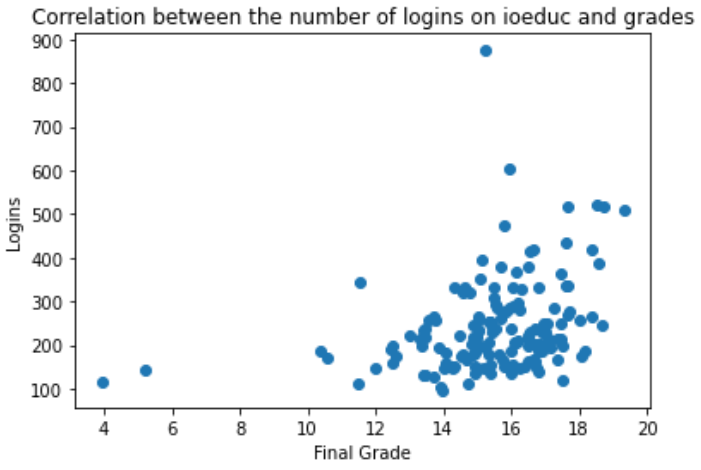

Figure 33 - Correlation between logins and grades

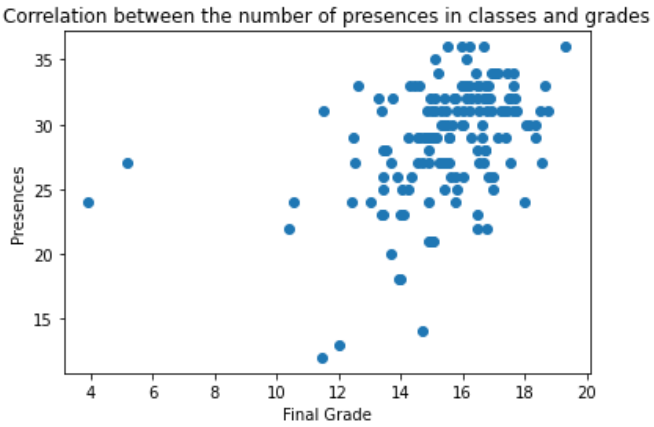

Figure 34 - Correlation between presences and grades

Starting with figure 33, it shows that good grades are usually related to a high number of logins, but especially shows that bad grades (or failed students) are related to a low number of platform interactions, which makes sense since it shows lack of interest or work in the course.

Likewise, figure 34 represents an even bigger relation between grades and attendance. The highest grade had the biggest attendance rate, and it is also noticeably the big number of high grades and with a grade over 16. However, in this case, all failed students had a decent attendance rate which means that this factor is not really related to the failing aspect but more about the excellence in terms of grades.

Adding to this, it was made a study splitting the grades into 4 different groups:

- A for the students with grades in the top 25%
- B for students with grades between the top 25% to 50%
- C for students with grades between the top 50% and 75%
- D for the students with the worst 25% grades

With this division, it was possible to investigate the relation between these 4 groups and the number of presences in classes of each group, represented on figure 35.

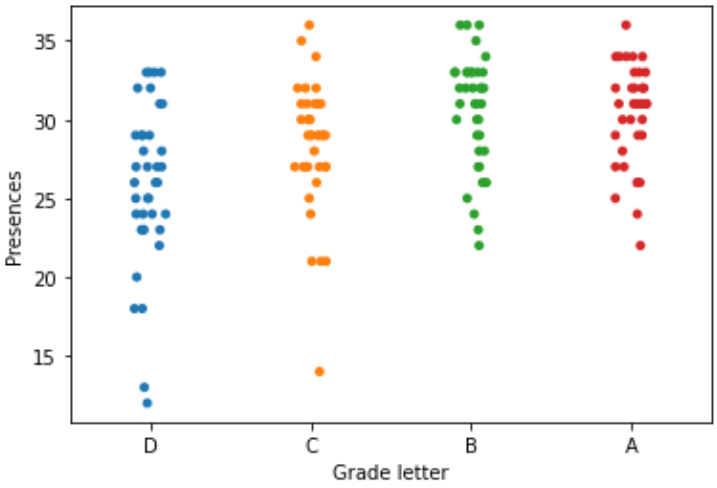

Figure 35 - Presences per group of students

Analysing the figure, it is possible to confirm the assumptions made previously, with the groups A and B, with the highest grades, having a big concentration on the top with an higher number of presences. On the other hand, especially group D, had no students with the maximum attendance and had several students in inferior part of the graphic, with a lower number of attendances, compared to the average

To conclude the analysis between these three variables, it was made a graphic where all three of them would be shown. That's represented on figure 36. In this figure it is possible to see that, as said before, a big concentration on the top with high attendance rate and login count, while most of the students in the having the colour yellow, meaning a grade over 17,5. On the other hand, in the bottom of the graphic (low participation and attendance rate) it is possible to see darker colours like dark green and blue that mean a lower final grade.

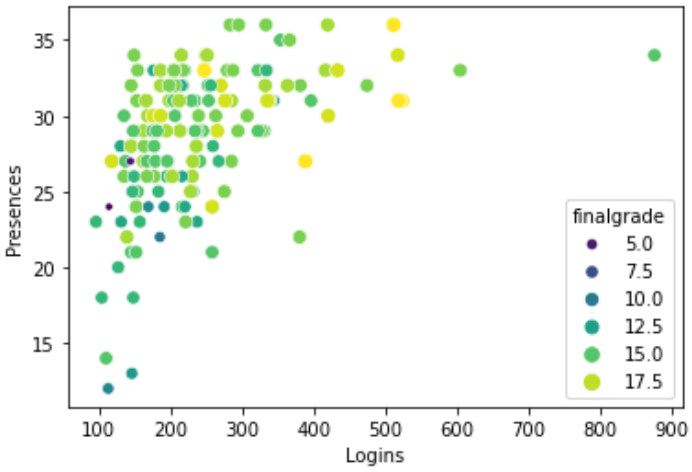

Figure 36 - Correlation between logins, presences, and grades

Going deeper in the evaluation sector, it was time to understand how each phase was being held by students, the ones with better results and in which ones it is possible to improve. In order to reach the final grade, three different types of evaluation are made:

- Quizzes: In each class, the students receive a key that it enables their access to the platform in order to test their knowledge of the content that was given at the class. The final grade for this component is calculated with the average of the grade in each quiz and in how many of them has the student participated. It weighs 15% of the final grade.
- MTs: During the year, there are three different mini tests that work as checkpoints to test the students' knowledge and work through the semester. The MT2 is the most important one, if the students fail in this one, it does not go through to the rest of the course or can apply to the rescues system, explained earlier. It weighs 25% of the final grade.
- Group Project: The biggest component of the three. The students are distributed in groups and have several checkpoints during the semester to show the work done until there. In the end, there is a final presentation as well as a final grade to each group member. It weighs 60% of the final grade.

Starting by analysing the quizzes, it is important to know the frequency that these quizzes were completed. Figure 37 represents the distribution of students per number of quizzes completed.

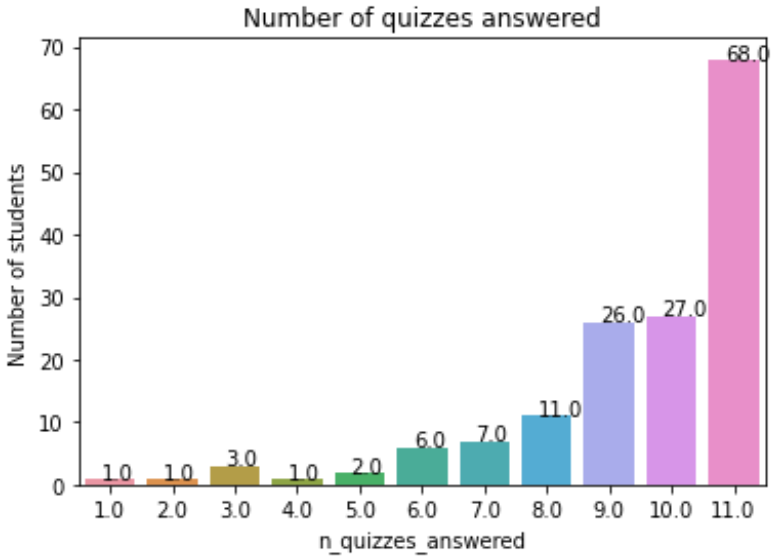

Figure 37 - Distribution of students per number of quizzes answered

Through this graphic, it is possible to see that 68 (44%) have completed 100% of the quizzes (11) and most of the students (79%) have, at least, 9 quizzes completed. It also corroborates the high attendance rates that were possible to check previously, which means that the students are showing interest and commitment to the course. Checking the quiz data, it is possible to rank the quizzes by the number of answers given and ordered by their title, shown in figure 38.

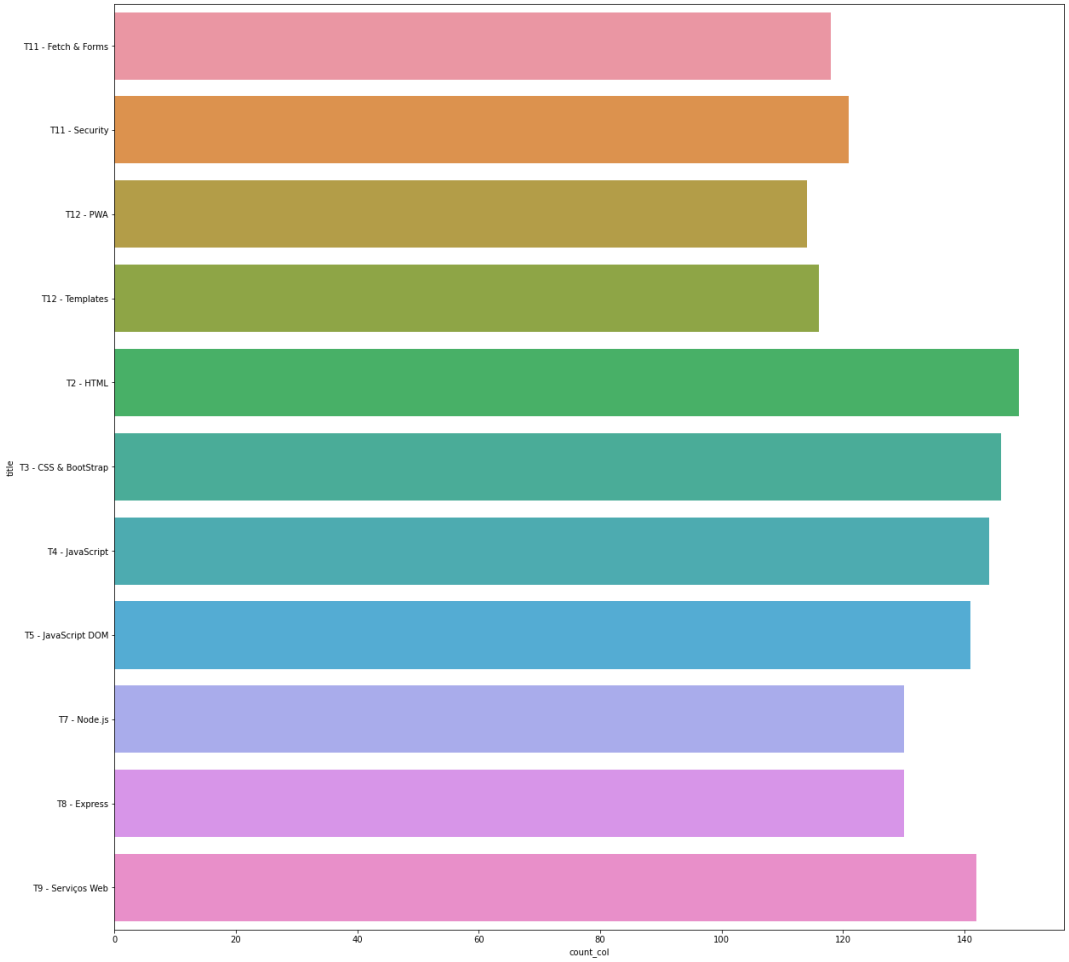

Figure 38 - Number of students participating per quiz

Once again, this graphic shows that every quiz had, at least, 114 students (75%) participating in it, while the best performing quiz (T2 – HTML) had a total of 149 students (97%).

Calculating the average grade in the quizzes component, the results shown in figure 39 are achieved.

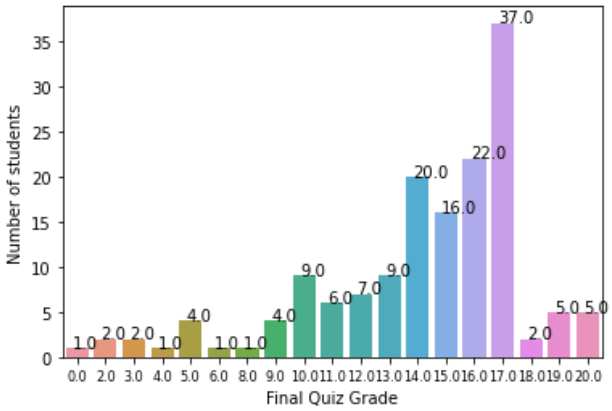

Figure 39 - Distribution of grades in quizzes

This data shows that most of the students (52%) achieve, at least, a grade of 15, which means they have achieved very good grade through this type of evaluation technique, reflecting that the topics approached in class have been explained in the right way, producing good results.

Comparing the relation between the quiz grade and the final grade, it is possible to get the graphic in figure 40.

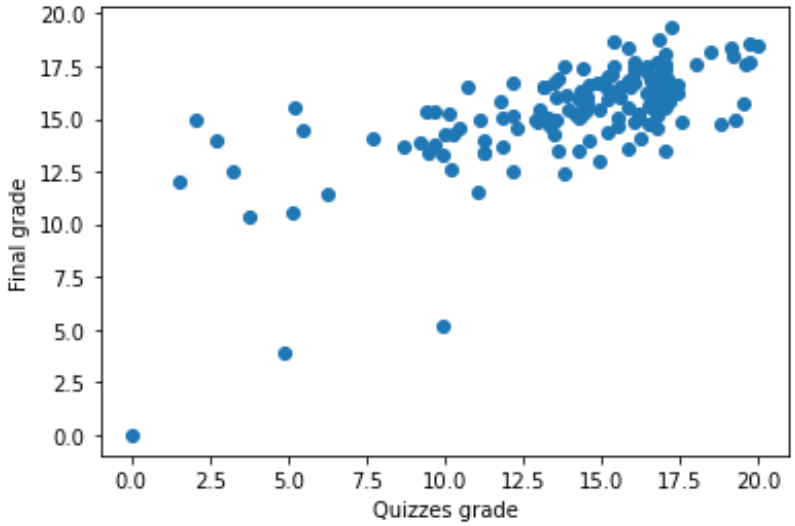

Figure 40 - Correlation between quiz and final grades

This figure represents the tight relation between the two variables. It is possible to see that higher grades in quizzes are directly related to high grade in the course.

Changing the evaluation component, it is time to check how the performance was in terms of MTs. The average grade can be found in figure 41:

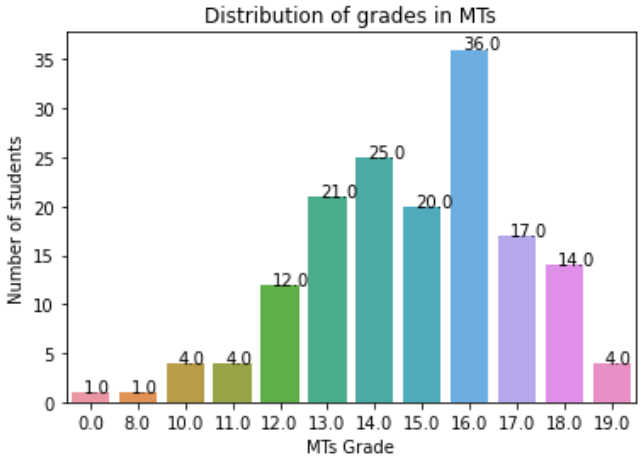

Figure 41 - Distribution of grades in MTs

The results from this analysis can be interpreted as favourable as well, since only 2 students failed during this component, while 91 students (53%) had a very good grade (<15). Figure 42 represents the correlation between these grades and the final grades of the course.

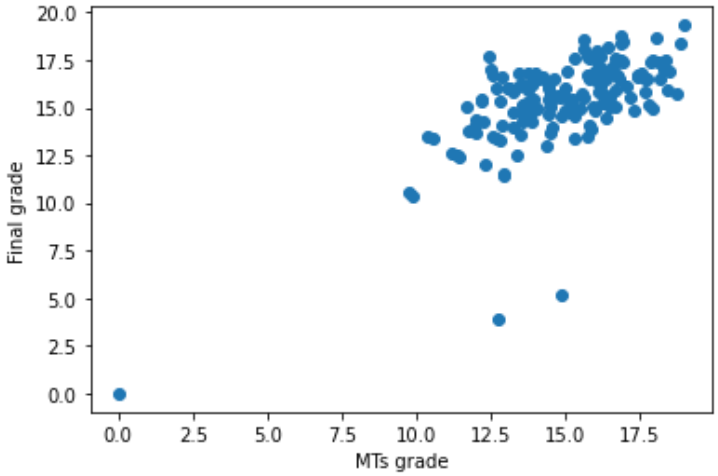

Figure 42 - Correlation between MT grades and final grades

This graphic describes, as expected, a big correlation between high grades in MTs and high grades in the end of the course.

From the 11 students that failed the MT2 and applied for the rescue system, shown in section 2.1., three of them still ended up with a negative grade in MT2, failing the course, like shown in figure 43. On the other hand, 8 students were able to go through by using this system.

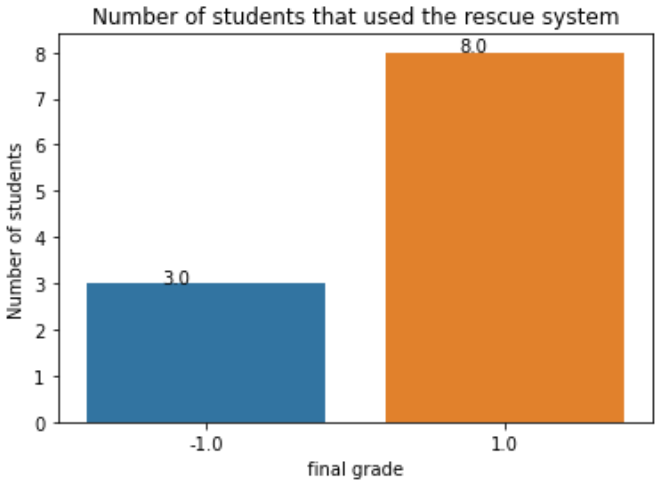

Figure 43 - Students that failed MT2

From the students that were able to being saved, it was made a study on their final grade. Figure 44 reflects the final grade from these 8 students.

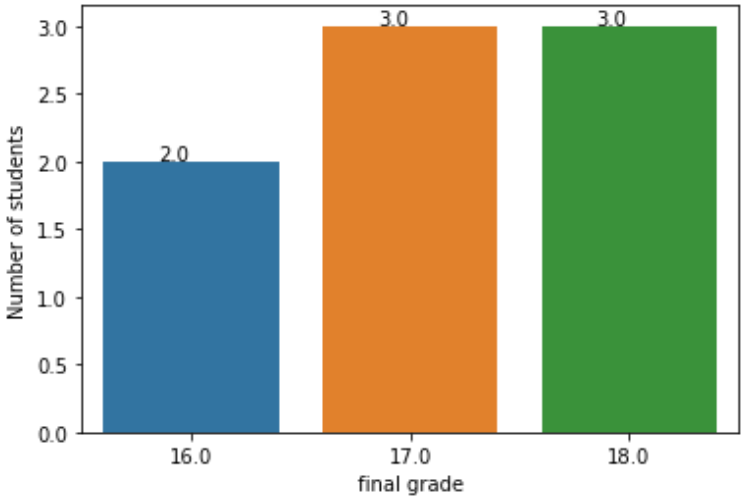

Figure 44 - Grades of rescue system students

These results show that even though the students had some difficulty in MT2, they were able to bounce back and achieve, at least, a grade of 16, corroborating the success of this system.

This evaluation component (MTs) can be understood as successful, as well as the rescue system, since it saved 8 students from failing the class and allowed failed students to have a second chance, if they were close to the minimal grade.

Lastly, it is important to check how the biggest evaluation component, the group project, was in terms of grades. Figure 45 describes the distribution of grades given in it:

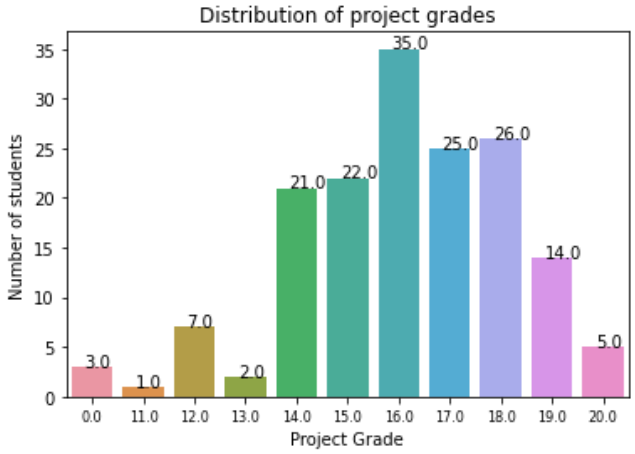

Figure 45 - Distribution of project grades

This graphic shows that the group project had a very good representation in terms of grades with only 3 students, the same that got 3 penalties during the year, had a grade of 0 in the group project consequently. Most of the students had a very good grade (>15) with 5 of them getting the biggest grade possible, 20. The results show that the checkpoints created to help students during the project was a big help and it made sure that almost no group was left alone in terms of work management.

Finally, applying the formula to get the final grade, the final distribution was like the graphic on figure 46.

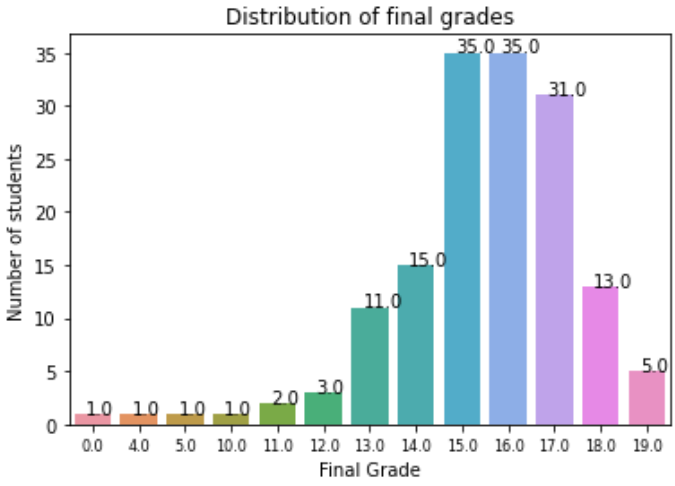

Figure 46 - Distribution of final grades

Reading the data, it is possible to conclude that only 3 students failed the course (the same that got the 3 penalties), while the rest was able to complete it. During the year that is being studied, 119 students (77%) ended up achieving a very good grade, over 15.

5.2. Predictive Modelling Results

Firstly, for the regression models, this prototype includes 12 different models, using 4 algorithms (explained in the modelling section) and creating 3 models per algorithm, 1 for each scenario. The average value for each metric in the best modelling algorithm was the following, represented in table 27

Table 27 - Regression modelling results

Scenario	MAE	MSE	RMSE	Algorithm
A	1,14	2,84	1,61	SVM
B	1,16	2,90	1,79	SVM
C	**1,11**	**2,69**	**1,58**	SVM

These values were accomplished by the SVM algorithm, which was the one with the best results in every scenario. The results show that the scenario C was the one with better results, achieving a MAE value of 1,11, while also achieving a MSE and RMSE value of 2,69 and 1,58.

Regarding the classification models, this prototype includes 48 different models (uses 4 different targets), using 4 algorithms (explained in the modelling section), creating 12 models per algorithm (3 models per target, one per scenario). The best average values for each scenario can be found on table 28.

Table 28 - Classification modelling results

Target	Scenario	Accuracy	Sensitivity	Algorithm
Higher15 [15-20]	A	77,74%	77,49%	SVM
	B	**80,37%**	**80,46%**	SVM
	C	78,41%	78,83%	SVM

Target	Scenario	Accuracy	Sensitivity	Algorithm
Lower15 [10-15]	A	**76,45%**	72,62%	SVM
	B	73,20%	**80,00%**	NB
	C	75,16%	66,95%	SVM
Higher16 [16-20]	A	**73,89%**	**70,48%**	SVM
	B	70,02%	68,78%	SVM
	C	71,94%	70,47%	RF
Lower16 [10-16]	A	**71,96%**	**74,44%**	SVM
	B	70,67%	74,15%	SVM
	C	70,02%	72,70%	SVM

Reading the results, it is possible to separate into two different groups, one for each target. Firstly, starting with the higher15 target, who represents the class of grades between 15 and 20, it is possible to see that the best scenario in terms of accuracy and sensitivity is the scenario B, with both values reaching the 80% mark. The best results for every scenario in this target were achieved by the SVM algorithm. Regarding with the lower15 target, no model from any scenario was able to reach the results in the other target, with 80% in both accuracy and sensitivity. The scenario with the best results in terms of correct predicted values was the scenario A, with a value of 76,45% in terms of accuracy, while the scenario B was the one with the highest value in sensitivity, 80,00%. In this case, the NB algorithm had the best model for scenario B, while the SVM achieved the best results for scenarios A and C.

Regarding the target higher16, the best scenario in terms of both metrics calculated, was the scenario A with the value of 73,89% for accuracy and 70,84% for sensitivity. For scenario A and B, the SVM algorithm was the one that got the best results. On the other hand, when modelling the scenario C, the RF algorithm was able to get the best values for both metrics. Changing to target lower16, the scenario A was also the one with the best numbers, with a value of 71,96% for accuracy and 74,44% for sensitivity. In this case, for every scenario, the SVM algorithm ended up being the best in every iteration.

Attachment II includes the modelling results from each algorithm individually, made during this phase.

Table 29 represents an example of a confusion matrix done when using the SVC algorithm when predicting the target higher15:

Table 29 - Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	23	6
Actual 1	27	97

6. Discussion

In this section, the results obtained in both modelling phases will be discussed and addressed.

6.1. Analytical Modelling

Considering the whole theme of the dissertation and in a critical perspective, the dashboards developed allowed to reach the following conclusions:

- Attendance in classes and participation in the platform is directly related to better grades;
- 77% of students were capable of getting a grade over 15;
- The attendance in classes is bigger in the beginning of the semester and in the end, during the evaluation period;

- The card system created an impact in the final grade. 3 students received the 3 penalties and a red card consequently, failing the course, while 5 students got 2 benefits and a blue card, that allowed them to get an extra point in the project;
- The rescue system had a total of 11 students applying and saved 73% of them (8), allowing them to finish the course;
- Every student attended, at least, 1 class in 3;
- Every quiz made during the semester had, at least, a 75% participation rate, and 44% of the students participated in every quiz, of a total number of 11.

6.2. Predictive modelling

The elaborated models were evaluated considering different metrics, for regression it was used the following:

- Mean absolute error (MAE);
- Mean squared error (MSE);
- Root mean squared error (RMSE);

All of the metrics above are error-based metrics, which means that the lowest the value, the better the prediction. Regarding the mean absolute error (MAE), the result achieved means that the average error in the prediction, compared to the real result, was, at best, 1,10 grade values. In real terms, this means that a grade of 17 can be predicted from 15,9 to 18,1 by this model. On the other hand, the mean squared error (MSE) and root mean squared error (MSE), measure the average and root squared difference between the observed and predicted values and had the best value as 2,69 and 1,58, respectively.

On the other hand, for classification, two following metrics were used:

- Accuracy;
- Sensitivity;

The accuracy evaluates the predictive capability of a model to predict the correct values, while sensitivity refers to the number of cases in which positive values were correctly identified in the model. In this case, the best value for accuracy was 80,37% for the higher15 target, and the best value for sensitivity was 80,46%, also for the same target. Regarding the group of targets higher16 and lower16, none was able to get more than 75% as their best result in terms of

accuracy or sensitivity. For this metrics, it was set a threshold of 80% in terms of accuracy and 85% in terms of sensitivity.

With this in mind, it was possible to reach the following facts:

- The values accomplished for MAE and MSE show that the modelling phase was successful, with an absolute error of just 1,10 (5%) in a scale from 1 to 20;
- For the target higher15, the model was able to achieve an accuracy higher than 80%, however, it was not successful to achieve a sensitivity value of 85%;
- The target lower15 was not successful, failing to hit both of the values set for accuracy and sensitivity;
- Targets higher16 and lower16 had a better distribution of values, however neither of them was able to get satisfactory results, with >5% difference from the thresholds defined for both accuracy and sensitivity;
- SVM was the best performing model in both regression and classification modelling;

Concluding, the presented models are able to achieve a good result and predict successfully the regression target, however, it still has predictive weaknesses in classification, having a semi-successful result for the class [15-20] but very low scores for the rest of the targets defined.

Chapter 6 - Conclusion

This chapter presents the conclusion of the work carried out, from the results perspective and contributions obtained. In addition, a section talking about future work to be carried out is also presented.

1. Final Considerations

With the purpose of answering the question, which served as a motivational basis for the development of the project, "Is it possible to predict the students grade based on his behaviour in class?", a case was developed of study on the evaluations and evaluative method, used in the discipline of Web programming, of the Integrated Master's course in Information Systems Engineering and Management regarding the academic year 2020/2021. Answering the question, it is possible predict a student's grade based on his behaviour, as it was shown during the predictive modelling phase present in this document, with very good results in terms of regression, achieving a value of 1,10 in terms of absolute error and 2,69 when talking about squared error. The development of this case study presented two well-defined phases:

- Literature review on the presented topics;
- Development of a prototype that could achieve the goals that were selected with the use of ETL (extraction, transformation and loading) processes, Data Warehouse construction and Data Mining techniques applied to the selected data.

The literature review allowed to cement the knowledge that had already been addressed during the academic path and learn new skills and insight that helped in the development phase. Table 31 presents the objectives that were previously outlined for this dissertation, as well as the results that made it possible to accomplish them.

Table 30 - Goals and obtained results

Goals	Results
Develop mechanisms that can handle both structured and unstructured data	Analysis with data from different sources
Develop a multidimensional model	Multidimensional model
Develop a set of Dashboards;	Dashboards created with the processed data
DM Prototype	Architecture

Goals	Results
	Data mining models and reports about their score

Starting for the handling of both unstructured and structured data, in order to deal with the first one, it was used collections from MongoDB and a series of projections and filtering to retrieve just the information that was important to the project. After structuring this data, it was possible to convert to data frames and joining with the data that was already transformed and clean (structured data).

The multidimensional model served as support for the whole extracting, transforming and loading (ETL) phase, by giving helping in terms of data cleaning and transformation. This model had a total of 2 fact tables and 5 dimensions, which represented a clear image of the project.

After completing the data handling part, it was created a bunch of dashboards in order to help the analysis of the course performance, the factors that most impact grades and which ones is possible to upgrade.

Finally, it was created a group of data mining models and metrics to evaluate them, in order to measure the predictive capabilities of the prototype. It started by using regression techniques, to predict the students' final grades, where it was used MAE, RMSE and MSE as evaluation metrics. The process was followed using classification methods, to predict how many students had a very good grade (over 15 and over 16). In this case, in order to measure the performance, it was set a goal of over 80% accuracy and 85% sensitivity.

2. Contributions

The developed artifact ensures help in the decision-making process by the responsible actors inserted in the educational environment, based on the analysis and evolution of the information taken from the data that was provided. It includes both an analytical and predictive model, with a group of dashboards and reports created graphically represent the reality of the investigated case, as well as predictive models which allow these actors to predict the grades from future students, as well as the rate of them that can reach over 15 values in the end of the course, giving them a tool to measure the final grades according to the students effort and participation in the different course activities during the semester.

In a global perspective, this dissertation can help not only the actors in the education sector but also the scientific community, since it replicates a real case of study in terms of web mining applied to teaching and the results that were achieved using different types of strategies. With the background created in the TechTeach methodology, this dissertation can help to corroborate some of the things that are explained in it with real evidence and results.

3. Future Work

Concluding this document, it is possible to say that the goals that were set for this project were accomplished with success. However, the future for this project consists of the continuation of data flowing, that can improve the created models by giving more data to test, and to replicate a better version of the reality with a higher number of students. The goal is also to extend past the Web Programming course for other areas of work and schools, in order to improve the education sector as a whole. Adding to this project, a group of scientific articles will be made to solidify the work done in this document.

In terms of the developed work, it has been successfully completed and can be adaptable to other areas.

Chapter 7 - References

Agarwal, S. (2013). Data mining: Data mining concepts and techniques. *In 2013 International Conference on Machine Intelligence and Research Advancement* (pp. 203-207). IEEE. https://doi:10.1109/ICMIRA.2013.45

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer. https://doi.org/10.1007/978-3-319-14142-8

Ai, J., & Laffey, J. (2007). Web mining as a tool for understanding online learning. *MERLOT Journal of Online Learning and Teaching*, 3(2), 160-169. https://jolt.merlot.org/vol3no2/ai.pdf

Alsghaier, H., Akour, M., Shehabat, I., & Aldiabat, S. (2017). The importance of Big Data Analytics in business: A Case study. *American Journal of Software Engineering and Applications*, *6*(4), 111-115. https://doi10.11648/j.ajsea.20170604.12

Baker, R. S. J. D. (2010). Data mining for education. *International encyclopedia of education*, *7*(3), 112-118. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.696.436&rep=rep1&type=pdf

Bin, W., & Zhijing, L. (2003). Web mining research. In *Proceedings Fifth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2003* (pp. 84-89). IEEE. https://DOI:10.1109/ICCIMA.2003.1238105

Chen, D. Y. (2017). *Pandas for everyone: Python data analysis*. Addison-Wesley Professional. https://www.oreilly.com/library/view/pandas-for-everyone/9780134547046/

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. http://hdl.handle.net/1822/8024

Fan, W., & Bifet, A. (2013). Mining big data: current status and forecast to the future. *ACM SIGKDD explorations newsletter*, *14*(2), 1-5. https://doi.org/10.1145/2481244.2481246

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media. https://doi:10.1007/978-3-642-19721-5

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big Data for Dummies*. John Wiley & Sons, Inc. https://jan.newmarch.name/IoT/BigData/Big%20Data%20For%20Dummies.pdf

Jumaa, A. K. (2018). Secured Data Conversion, Migration, Processing and Retrieval between SQL Database and NoSQL Big Data. *DIYALA Journal for pure sciences*, *14*(4), 68-87. http://dx.doi.org/10.24237/djps.1404.449A

Kanimozhi, K. V., & Venkatesan, M. (2015). Unstructured Data Analysis - A Survey. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(3), 223-225. https://DOI:10.17148/IJARCCE.2015.4354

Kotecha, B. H., & Joshiyara, H. (2017). A Survey of Non-Relational Databases with Big Data. *International Journal on Recent and Innovation Trends in Computing and Communication*,*5*(11),143-148. https://www.academia.edu/36852552/A_Survey_of_Non_Relational_Databases_with_Big_Data?auto=download

Kumar, S. N. (2015). World towards advance web mining: A review. *American Journal of Systems and Software*, *3*(2), 44-61. https://DOI:10.12691/ajss-3-2-3

Luan, J. (2004). Data mining applications in higher education. *SPSS Executive*, *7*. http://www.insol.lt/media/collateral/modeling/education.pdf

Maia, A., Portela, F., & Santos, M. F. (2018). Web Intelligence in Higher Education: A Study on the Usage of Business Intelligence Techniques in Education. *2018 6th International Conference on Future Internet of Things and Cloud Workshops*

https://doi.org/10.1109/W-FiCloud.2018.00034

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., ... & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.

https://doi.org/10.1109/TKDE.2019.2962680

Mathew, V., Toby, T., Singh, V., Rao, B. M., and Kumar, M. G. (2017). Prediction of remaining useful lifetime (rul) of turbofan engine using machine learning. In 2017 IEEE International Conference on Circuits and Systems (ICCS), pages 306–311

https://doi.org/10.1109/ICCS1.2017.8326010

McNulty, H. (2014, May 22). *Understanding Big Data: The Seven V's.* Dataconomy.

https://dataconomy.com/2014/05/seven-vs-big-data/

Meier, A., & Kaufmann, M. (2019). *SQL & NoSQL databases*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-24549-8

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental Journal of Computer Science and Technology*

http://www.computerscijournal.org/?p=1592

Peffers, K., Tuunanen, T., Rothenberger, M. A., & S. Chatterjee. (2007). A design science research methodology for information systems research. *Journal of management information systems*, v. 24

https://doi.org/10.2753/MIS0742-1222240302

Peña-Ayala, A. (2014). Educational data mining. *Studies in Computational Intelligence*. Springer.

https://doi.org/10.1007/978-3-319-02738-8

Portela, F. (2020). Techteach—an innovative method to increase the students engagement at classrooms. *Information (Switzerland),* 1-32, 11(10).

https://doi.org/10.3390/info11100483

Portela, F. (2022). Towards an Engaging and Gamified Online Learning Environment-A Real CaseStudy. *Information*, 27-28, 13(2).

https://doi.org/10.3390/info13020080

Portela, F., & Fernandes, G. (2020). Provisional Patent

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458-472. https://doi.org/10.1016/j.compedu.2013.06.009

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*

https://doi.org/10.1002/widm.1075

Sadiku, M. N., Shadare, A. E., & Musa, S. M. (2015). Data mining: a brief introduction. *European Scientific Journal*, *11*(21), 509-513. https://eujournal.org/index.php/esj/article/view/6017

Schwaber, K. & Sutherland, J. (2007). *The Scrum Papers: Nuts, Bolts, and Origins of an Agile Process.* Scrum Inc. https://www.qagile.pl/wp-content/uploads/2018/11/scrum-papers.pdf

Schwaber, K. & Sutherland, J. (2017). *The Scrum Guide™, The Definitive Guide to Scrum: The Rules of the Game.* https://scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf

Sharma, K., Shrivastava, G., & Kumar, V. (2011). Web mining: Today and tomorrow. In *2011 3rd International Conference on Electronics Computer Technology* (Vol. 1) (pp. 399-403). IEEE. https://DOI:10.1109/ICECTECH.2011.5941631

Sinha, D. (2021, June 25). *Top 10 Big Data Statistics You Must Know in 2021.* Analytics Insight. www.analyticsinsight.net/top-10-big-data-statistics-you-must-know-in-2021/

Song, I. Y., & Zhu, Y. (2016). Big data and data science: what should we teach?. *Expert Systems*, *33*(4), 364-373. https://doi.org/10.1111/exsy.12130

Sowmya, R., & Suneetha, K. R. (2017). Data mining with big data. In *2017 11th International Conference on Intelligent Systems and Control (ISCO)* (pp. 246-250). IEEE. https://doi:10.1109/ISCO.2017.7855990.

Strauch, C., Sites, U. L. S., & Kriha, W. (2011). NoSQL databases. *Lecture Notes, Stuttgart Media University*, *20*, 24. https://bigb.es/lectures/2014/15.5.pdf

Voznika, F., & Viana, L. (2007). Data mining classification. *Washington: University of Washington*. https://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of*

knowledge discovery and data mining (Vol. 1) (pp. 29-39). Springer-Verlag. http://www.cs.unibo.it/~montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Wilder-James, E. (2012). *Planning for Big Data: a CIO's handbook to the changing data landscape.* O'Reilly Media, Inc.

Wunsch, D. & Xu, R. (2008). Clustering. *IEEE Press Series on Computational Intelligence* https://DOI:10.1002/9780470382776

Zhang, W., Wu, C., Zhong, H., Li, Y., and Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on bayesian optimization. Geoscience Frontiers, 12(1):469–477.

https://doi.org/10.1016/j.gsf.2020.03.007

Zikopoulos, P. C., Eaton, C., Deroos, D., Seutsch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. Mc Graw-Hil. https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/I111025E.pdf

Attachment 1 – Risks Table

Table 31 represents the risk analysis that was realized for this project. This is an important section because is possible to plan actions to avoid problems and their impact on the final product if the risks are well documented.

For each of the identified risks was estimated its probability (P), its impact (I), and its severity (S). The probability and impact columns follow a scale from 1(low probability/low impact) to 5 (high probability/high impact) while the severity follows a scale from 1 to 25 (probability x impact), with the same logic, higher the number, higher the severity. This table also contains the consequence of this risk and a possible mitigation action. The last column refers if the risk happened or not during the project.

Table 31 - Risks Table

Risk	P	I	S	Mitigation Action	Risk Happened?
Lack of experience in the area	4	4	16	Read articles and books related to the theme. Constant communication with mentors to solve possible problems.	Yes, mitigation action solved it.
High complexity of the project	3	4	12	Study new technologies and tools that can help to solve new problems.	Yes, mitigation action solved it
Bad project planning	1	5	5	Prioritize activities and be careful with possible delays that can cause problems ahead	No
Lack of comprehension of the data.	2	4	8	Talk with the project mentors and data provider to solve possible questions about it	Yes, mitigation action solved it

Risk	P	I	S	Mitigation Action	Risk Happened?
Loss of files	1	3	3	Doing regular backups and storing in different places	No
Lack of communication with the mentors	1	4	4	Define weekly meetings and a communication platform	No
Bad writing related to using nonmaternal language.	2	3	6	Search and develop the grammar and vocabulary of the new language	No
Difficulty using new data mining tools	3	4	12	Search for documentation and available tutorials about the tool	No
Malfunctioning of machines or devices	1	3	3	Restore files using backups. Use backup machines and devices. Adjust the work plan to minimize the delay	No

Attachment 2 – Algorithm Results Table

Table 32 represents the values for each target during the predictive modelling for the SVM model.

Table 32 - SVM Model Classification

Target	Scenario	Accuracy	Sensitivity
Higher15 [15-20]	A	77,74%	77,49%
	B	80,37%	80,46%
	C	78,41%	78,83%
Lower15 [10-15]	A	76,45%	72,62%
	B	72,20%	69,16%
	C	75,16%	66,95%
Higher16 [16-20]	A	73,89%	70,48%
	B	70,02%	68,78%
	C	71,97%	69,27%
Lower16 [10-16]	A	71,96%	74,44%
	B	70,67%	74,15%
	C	70,02%	72,70%

Table 33 shows the values for each classification target during the predictive modelling for the DT model.

Table 33 - RF Model Classification

Target	Scenario	Accuracy	Sensitivity
Higher15 [15-20]	A	73,28%	80,33%
	B	73,25%	80,31%
	C	75,01%	78,83%
Lower15 [10-15]	A	69,31%	54,52%
	B	67,87%	49,81%
	C	73,16%	59,77%
Higher16 [16-20]	A	66,02%	61,27%
	B	68,02%	65,04%
	C	67,94%	68,63%
Lower16 [10-16]	A	62,15%	65,34%
	B	68,60%	70,85%
	C	68,69%	70,23%

Table 34 shows the values for each classification target during the predictive modelling for the DT model.

Table 34 - DT Model Classification

Target	Scenario	Accuracy	Sensitivity	Algorithm
Higher15 [15-20]	A	66,04%	80,42%	SVM
	B	67,38%	79,29%	SVM
	C	69,89%	79,34%	SVM
Lower15 [10-15]	A	64,69%	45,27%	SVM
	B	71,81%	58,62%	NB
	C	64,66%	42,34%	SVM
Higher16 [16-20]	A	65,25%	50,84%	SVM
	B	66,69%	64,11%	SVM
	C	64,12%	59,40%	RF
Lower16 [10-16]	A	60,88%	66,52%	SVM
	B	63,38%	68,19%	SVM
	C	60,19%	64,28%	SVM

In Table 34 it is possible to analyse the values for each classification target during the predictive modelling for the NB model.

Table 35 - NB Model Classification

Target	Scenario	Accuracy	Sensitivity	Algorithm
Higher15 [15-20]	A	71,22%	70,64%	SVM
	B	71,90%	71,49%	SVM
	C	75,16%	74,89%	SVM
Lower15 [10-15]	A	69,33%	56,67%	SVM
	B	73,20%	80,00%	NB
	C	72,58%	69,33%	SVM
Higher16 [16-20]	A	64,06%	70,03%	SVM
	B	56,17%	50,00%	SVM
	C	64,02%	69,33%	RF
Lower16 [10-16]	A	62,09%	59,62%	SVM
	B	54,90%	55,55%	SVM
	C	60,77%	59,47%	SVM

Regarding the regression modelling, table 36 represents the values for the RF algorithm. The SVM algorithm will not be shown here, since it's represented on the predictive modelling section, in the Evaluation subchapter.

Table 36 - RF Model Regression

Scenario	MAE	MSE	RMSE
A	1,26	2,75	1,64
B	1,24	2,93	1,74
C	1,27	2,65	1,62

Table 37 represents the values for the ANN algorithm during the modelling phase.

Table 37 - ANN Model Regression

Scenario	MAE	MSE	RMSE
A	6,01	36,38	5,18
B	4,47	20,07	3,64
C	7,96	63,25	7,51

Table 38 shows the values for the DT algorithm during the modelling phase.

Table 38 - DT Model Regression

Scenario	MAE	MSE	RMSE
A	1,54	4,02	1,97
B	1,43	3,33	1,82

Scenario	MAE	MSE	RMSE
C	1,59	4,48	2,08