Universidade do Minho
Escola de Engenharia

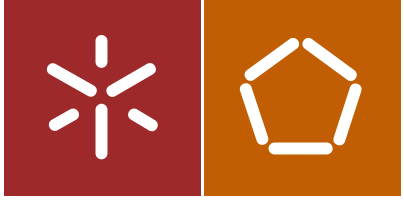
António João Oliveira da Silva

An Intelligent Decision Support System for the Analytical Laboratories of a Chemistry Industry

September 2022

António João Oliveira da Silva An Intelligent Decision Support System for the Analytical Laboratories of a Chemistry Industry

UMinho | 2022



Universidade do Minho
Escola de Engenharia

António João Oliveira da Silva

An Intelligent Decision Support System for the Analytical Laboratories of a Chemistry Industry

Doctorate Thesis
Doctoral Program in Information Systems and Technologies

Work developed under the supervision of:
Paulo Cortez

September 2022

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

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

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

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

License granted to the users of this work

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

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

Acknowledgements

This PhD journey was an incredible challenge, which led me to meet fantastic people who taught me a lot and for whom I would like to dedicate this work.

First, I would like to thank my supervisor, Professor Paulo Cortez, for all his support. Thank you for your perseverance, trust, and encouragement so that I always wanted to learn more and be more demanding with myself. Thank you for the motivation to overcome all the challenges that were encountered in this PhD.

To the fascinating colleagues with whom I shared knowledge and had incredible life experiences during the years of this PhD, both in classes and group projects, as well as in the laboratories in which I worked in the ALGORITMI Center of School of Engineering of Minho University.

To my co-workers at Computer Center Graphics (CCG), namely at EPMQ IT: Engineering Process Maturity and Quality domain, without ever forgetting "my" Machine Learning team, which I witnessed at its conception, thank you for all the patience (which was a lot), and support throughout the years that I worked and did my PhD. To my co-workers at DTx - Digital Transformation CoLab, more particularly to the Software and Information Systems team, that were also always available to me. To the incredible AI4Medimaging team, who make me feel at home, and for the great patience with me mainly in this final stage of the PhD.

The cooperation with organization where this PhD thesis took place was also very important. The organization's specialists provided us the data and their business model knowledge. Finally, your feedback was valuable for the conclusion of this PhD. Thank you for all your availability.

To my friends, all of them, from childhood friends, friends from high school, university, Afonsina, thank you for the comradeship and for helping me to keep focused on my goals. Thank you for all advice and support, which were essential to keep me motivated in the most difficult moments. Finally, I would like to thank to my family, who have always supported and encouraged me during this PhD project, providing everything they could to my success.

STATEMENT OF INTEGRITY

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

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

_____, _____

(Place) (Date)

_____

(António João Oliveira da Silva)

Resumo

Um Sistema Inteligente de Apoio à Decisão para os Laboratórios Analíticos de uma Indústria Química

A Indústria 4.0 representa a quarta revolução industrial e envolve uma implementação que utiliza várias tecnologias de informação para dar suporte à produção, bem como uma monitorização em tempo real dos processos industriais. O tópico de *Business Analytics* é particularmente valioso neste contexto, uma vez que resulta de uma combinação de *Business Intelligence* com Optimização e Previsão. O objectivo é obter conhecimentos orientados por dados que podem ser úteis para ajudar na tomada de decisões sobre processos de produção. Por exemplo, *Business Analytics* pode ser utilizada para analisar dados históricos para ajudar a detectar e prever problemas ou falhas na produção. Outra possibilidade interessante é a previsão de ordens de procura, que pode ajudar no processo de gestão de *stocks*.

Este trabalho de doutoramento é realizado no âmbito de um projecto de Investigação e Desenvolvimento (I&D). O principal objectivo é a investigação e implementação de um Sistema Inteligente de Apoio à Decisão (IDSS em Inglês) que utiliza técnicas de *Business Analytics* (Descritiva, Prescritiva e Preditiva), integrado no conceito de Indústria 4.0 e aplicado a Laboratórios Analíticos de Empresas Químicas. Inicialmente, as necessidades das empresas analisadas foram elicitadas, e posteriormente foram desenvolvidos vários módulos do IDSS com o objectivo de resolver os objectivos das empresas Químicas. O primeiro módulo estudado foi a previsão da chegada de amostras aos Laboratórios Analíticos, utilizando uma ferramenta de *Auto Machine Learning* (AutoML). Em seguida, foi desenvolvido um módulo para prever o consumo de materiais nos laboratórios. Este módulo incluiu três abordagens de previsão diferentes que foram comparadas, uma com um AutoML, outra utilizando a metodologia ARIMA e a última baseada num algoritmo de aprendizagem profunda (*Long Short-Term Memory* em Inglês). Os melhores resultados de previsão foram obtidos através da abordagem AutoML. Finalmente, foi desenvolvido um módulo com métodos prescritivos para atribuir os instrumentos às análises a realizar, bem como o desenvolvimento de *Dashboards* de fácil utilização para o IDSS concebido. O sistema IDSS completo foi avaliado através de questionários e entrevistas abertas com os gestores do Laboratório Analítico. Globalmente, foi obtido um *feedback* positivo.

Palavras-chave: *Business Analytics, Chemical Laboratories, Industry 4.0, Machine Learning, Optimization, Prediction.*

Abstract

An Intelligent Decision Support System for the Analytical Laboratories of a Chemistry Industry

The Industry 4.0 represents the fourth industrial revolution and involves an implementation using several Information Technologies to support production, as well as a real-time monitoring of industrial processes. The topic of Business Analytics is particularly valuable in this context, since it results from a combination of Business Intelligence with Optimization and Forecasting. The objective is to obtain data-driven knowledge that can be useful to help decision making on production processes. For example, Business Analytics can be used to analyze historical data to help detect and predict problems or failures in production. Another interesting possibility is the prediction of demand orders, which can help in the process of stock management.

This PhD work is carried out within the scope of a Research & Development (R&D) project. The main objective is the research and implementation of an Intelligent Decision Support System (IDSS) that uses Business Analytics techniques (Descriptive, Prescriptive and Predictive), integrated within the Industry 4.0 concept and applied to Analytical Laboratories of Chemical companies. Initially, the analyzed company needs were elicitated, and subsequently several IDSS modules were developed aiming to solve the Chemical company goals. The first studied module was the prediction of arrival of samples at the Analytical Laboratories by using an Auto Machine Learning (AutoML) tool. Next, a module was developed for predicting the consumption of materials in the laboratories. This module included three different forecasting approaches that were compared, one with an AutoML, another using the ARIMA methodology and the last based on a deep learning algorithm (Long Short-Term Memory). The best forecasting results were achieved by the AutoML approach. Finally, a module was developed with prescriptive methods to allocate the instruments to the analyses to be performed as well as the development of the friendly user Dashboards for the designed IDSS. The full IDSS system was evaluated by using questionnaires and open interviews with the Analytical Laboratory managers. Overall, a positive feedback was obtained.

Keywords: Business Analytics, Chemical Laboratories, Industry 4.0, Machine Learning, Optimization, Prediction.

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
AL	Analytical Laboratories
ANN	Artificial Neural Networks
AREC	Area of Regression Error Characteristic
AutoML	Automated Machine Learning
BDW	Big Data Warehouse
BI	Behavioral Intention
CNN	Convolutional Neural Networks
CPS	Cyber-Physical Systems
CRISP-DM	Cross-Industry Standard Process for Data Mining
DL	Deep Learning
DM	Data Mining
DNN	Deep Neural Networks
DSRM-IS	Design Science Research Methodology for Information Systems
DSS	Decision Support Systems
DT	Decision Trees
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
FP	Final Product
FS	Feature Selection
GBM	Gradient Boosting Machine
GLM	Generalized Linear Model
GMP	Good Manufacturing Practices

HS Holdout Split

ICT Information and Communication Technologies
IDSS Intelligent Decision Support Systems
IN Intermediate
IoT Internet of Things
IPC In Process Control
IS Information Systems
IT Information Technology

KNN K-Nearest Neighbor

LinR Linear Regression
LogR Logistic Regression
LSTM Long Short-Term Memory

MAE Mean Absolute Error
ML Machine Learning
MSE Mean Squared Error

NMAE Normalized Mean Absolute Error
NN Neural Networks

OUT Output Quality

PEC Perception of External Control
PEOU Perceived Ease of Use
PU Perceived Usefulness

QC Quality Control

R&D Research & Development
REC Regression Error Characteristic
REL Job Relevance
RF Random Forest
RFID Radio-Frequency Identification

RM	Raw Material
RMSE	Root Mean Square Error
RSE	Relative Squared Error
RW	Rolling Window
SE	Stacked Ensemble
SLR	Systematic Literature Review
SVM	Support Vector Machine
TAM	Technology Acceptance Model
TRL	Technology Readiness Level
TSF	Time Series Forecasting
UC	Use Case
WP	Work Package
XGB	XGBoost
XRT	Extremely Randomized Trees

Chapter 1

Introduction

This chapter contains an motivation and contextualization of the problem that leads to this doctoral thesis. Then, the research objectives and methodologies are presented. Finally, the scientific contributions of this thesis are presented, and a description about the structure of this thesis is given.

1.1 Motivation

Business Analytics plays an important role in several businesses. It focuses in the analysis of historical raw data in order to achieve useful and focused insights and a better understanding of the business performance areas (Krishnamoorthi & Mathew, 2018). Business Analytics is the result of combining Business Intelligence techniques with Optimization, Forecasting, Predictive Modeling and Statistical Analysis (Arnott & Pervan, 2014). Business Analytics systems are being applied in the Industry sector, and this, in conjunction with the Industry 4.0 phenomenon, is causing significant changes in this sector.

Nowadays, most of the Industry is facing times of change. This change is being enabled by new techniques and technologies, including sensors and communication devices that generate Big Data and also analytic systems capable of analyzing such data, allowing to produce new insights and knowledge about the productive system. The term Industry 4.0 is used to identify this process. The German Federal Ministry of Education and Research defines the Industry 4.0 concept as: "the flexibility that exists in value-creating networks is increased by the application of cyber physical production systems. This enables machines and plants to adapt their behavior to changing orders and operating conditions through self-optimization and reconfiguration... The main focus is on the ability of the systems to perceive information, to derive findings from it and to change their behavior accordingly, and to store knowledge gained from experience. Intelligent production systems and processes as well as suitable engineering methods and tools will be a key factor to successfully implement distributed and interconnected production facilities in future Smart Factories"(Shrouf et al., 2014).

This PhD was developed within a three-year Research & Development (R&D) project that was funded by a private company and whose main objective relies in the creation of an integrated intelligent system

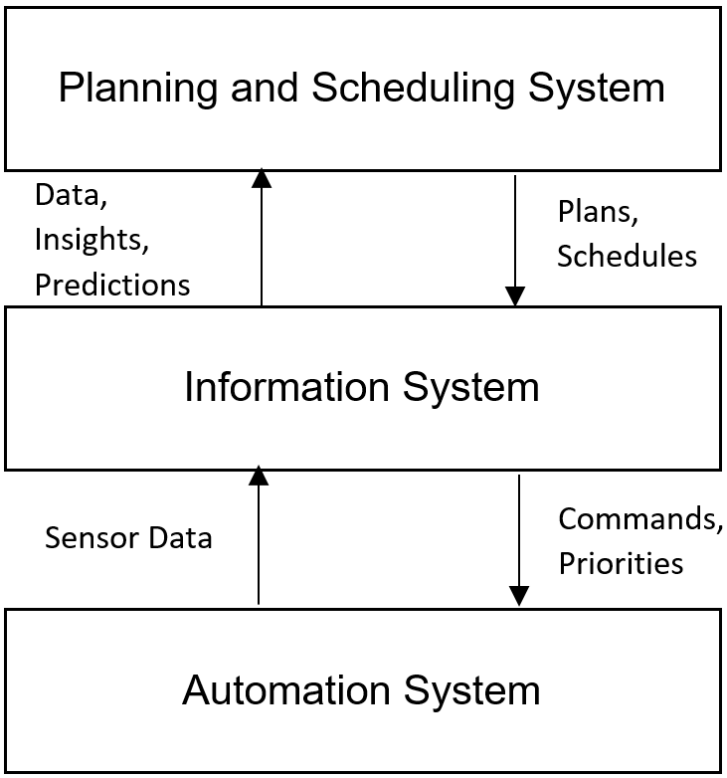


Figure 1: Relations existing between the three R&D project WPs.

based on state-of-the-art technology, under the Industry 4.0 concept, and that can improve the processes and the efficiency of the organization. The organization in question is from the Chemical Industry sector. A key aspect of this R&D project, and that is addressed in this PhD work, is the adaptation of Business Analytics techniques, under the Industry 4.0 context, to Chemical Analytical Laboratories. It should be noted that the full R&D project contains three main Work Package (WP), as shown on Figure 1. The "Planning and Scheduling System" aims to provide an automatic tool for scheduling laboratory tests. The "Automation System" WP2 is more linked with hardware (e.g., robotic arm automation). The goal is to automate some manual laboratory processes. Also, it will generate real-time laboratory data from sensors. The "Information System" (WP3) aims to design and implement the Laboratories Information Systems (IS), being based on a Big Data Warehouse, to collect, storage and process the data. These three main systems are expected to heavily communicate and interact. In particular, the "Automation System" (WP2) will interact with the "Information System" (WP3) by sending the sensor data information. Also, the IS (WP3) will provide the Commands and Priorities for the laboratory machines. The planning and scheduling module (WP1) will provide information about the plans and schedules to the WP3, and the latter will send the data, insights and predictions to the WP1.

This thesis aims to cover the "Intelligence"component of the IS (WP3). It will be focused on the design of a Business Analytics system that is capable of analyzing historical laboratory data, under the Industry 4.0 concept, in order to extract useful knowledge (e.g., predictions, insights) to improve laboratory

processes and management.

1.2 Problem Formulation

Generally, an industrial context may be established in any one of three sectors: primary, secondary and tertiary. The primary sector relates to the transformation and the extraction of Raw Material (RM) from the land or sea (e.g. oil, iron ore, timber and fish). Some examples of industries within this sector are mining, quarrying, fishing, forestry, and farming. These materials are then used in industries from the secondary sector, which may also be called Manufacturing and Industry sector, or production sector, in which RM are transformed into finished goods on a large scale. This sector includes all branches of human activities that transform RM into products or goods, as secondary processing of RM, food Manufacturing, textile Manufacturing and Industry. Finally, the tertiary sector, also known as service sector, includes all branches of human activity whose essence is to provide services, thus contributing to physical/mechanical work, knowledge, financial resources, infrastructure, goods or a combination of those (Kenessey, 1987; Wolfe, 1955).

This doctoral project takes places is a multinational company, founded in Portugal, and active in seven countries worldwide: Portugal, China, Ireland, Switzerland, USA, India and Japan. In this project the focus will be the factory in Portugal. The domain focus, where this company is positioned, is the secondary sector Industry or Manufacturing. With regard to the areas composing this specific context, this company's industrial site is divided in three areas: Warehouse, Production and Laboratories. To have a better understanding about the current state of the organization and the relationship between those areas, Figure 2 presents the current, known as AS-IS, architecture and workflow of materials (RM, Samples) and Information between the different areas and the softwares that they use.

With regard to the Warehouse area, this is the place where the products are received and shipped. The products that are provided by the suppliers are named RM. In the case of Intermediate (IN) Products and Final Product (FP), these arrive in the Warehouse from the Production area. The Production area is where the product manufacturing is performed. This area receives RM from the Warehouse and during the production process In Process Control (IPC) samples are created. The IPC sample is critical to the production process, and production may stop until the samples are approved by the Analytical Laboratories (AL). At the end of the process, IN and FP samples are created and the product packaging is sent to the Warehouse.

With regard to AL, we can identify four main events, namely, planning the arrival of samples, the arrival of samples, weekly planning and scheduling, and testing. Both branches also require continuous support from Quality Control (QC) Laboratories. These Laboratories evaluate products throughout their whole life-cycle, assuring rule compliance to Good Manufacturing Practices (GMP), while complying with Good Laboratory Practices. This regulation is important to assure the pharmaceutical product's safety and control, as well as its continued quality. With the recent company growth as a Contract Manufacturing

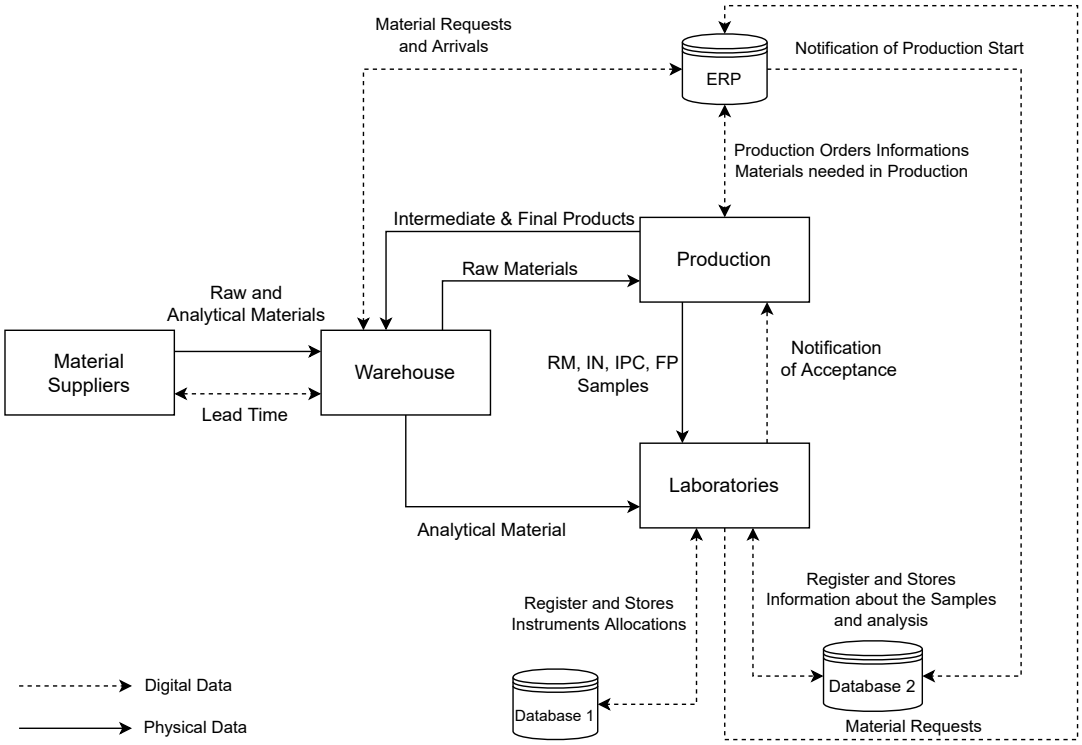


Figure 2: AS-IS Architecture of the Organization.

Organization, the mix of products and laboratory tests has increased laboratory process complexity. Under this context, AL are fundamental to the company and to QC.

In relation to the flow between the three areas (Warehouse, Production and Laboratories) and focusing more specifically on the Laboratories, it is important to point out that the Warehouse sends the physical samples of RM to the Production and here, an employee sends by email to the Laboratories, the names of the RM so that they can collect them. When the analyst is in charge of going to the Manufacturing building to raise the RM sample, it has to enter the aforementioned matter in the laboratory by registering it in "Database 2". Then, since the RM sample is already in the laboratory, the analytical test can be carried out or, if it is not possible to start the analytical test at that moment, the RM is stored locally in the laboratory. Regarding the workflow between AL and the Production area, there are certain moments in production when it is necessary to take samples from the production line and send them to the AL to ensure that the production of the products is up to standards quality requirements for the products. The sample is taken by the Operator of the production machines in certain periods of time defined in the production sheet and then the analyst will collect the samples to analyze them in the AL. These samples are called IPC and have analysis priority over the remaining samples. The IPC Sample data is generated automatically in "Database 2" after the start of a Production Order.

In the AL, the analysts usually do not record the sample arrival in the "Database 2" software, as they usually perform the analysis in the samples and when the analysis is done they write the sample

arrival and the test result at the same time in "Database 2". This happens because in the analysts have to record the sample state and analysis procedure in physical documents named Logbooks. Regarding to the analytical tests that are carried out in the laboratory, if an analysis which is running does not get approval at the end due to deviations from the standard, the process is repeated to try to check where the error occurred. If the error occurs before the injection phase, there is no event registration. In case the analytical test goes wrong and this error occurs after the sample injection, it is communicated to QC, which will check if the problem comes from the components or materials that are being used and the deviation is reported in the Corrective Action and Preventive Action software.

The AL have vast data stored in physical documents, which reduces the productivity during the analytical tests. This happens because the analyst must have to write all the products used and all the conditions of the analytical tests. However the AL use some Information Technology (IT) applications to help the management of the Laboratories. The application used in the Laboratories an Enterprise Resource Planning (ERP) for material requests and to have information about the Production Orders. Regarding the Samples management and tests, the Laboratories uses the "Database 2" to register the samples and the result of the tests performed to the samples and the "Database 1" to view the allocation of the HPLC and GC instruments. For privacy purposes, the names of the softwares and databases used in the AL are anonymized in this document.

The implementation of an Intelligent Decision Support Systems (IDSS) can be potentially useful to create a ground truth of data by integrating the data from the different softwares used in the Laboratories. This would provide for the Analysts, new insights and improve various processes performed in the Laboratories. In this doctoral project, the goal is to improve the functioning of the AL by using Business Analytics techniques in the Industry 4.0 context, aiming to solve the analyzed company needs.

1.3 Objectives

This PhD program, as stated before, was developed within an R&D project that aimed to implement three different WP in AL. In what concerns WP3, where this thesis is inserted, the objective was to know how Business Analytics techniques can improve processes in AL. Based on that, the research question to be answered in this PhD project is: **How can an Intelligent Decision Support System (IDSS) be designed and implemented under the Industry 4.0 concept to create value in the Analytical Laboratories of a Chemical Industry?**

To answer the Research Question defined, we addressed the following intermediates objectives:

- Conduct a Systematic Literature Review (SLR) on the use of Business Analytics techniques in Industry 4.0, to verify what type of techniques have been used in this context, which areas of industry are embracing the concept of Industry 4.0 and what are the open research gaps regarding the use of Business Analytics in Industry 4.0.

- Develop predictive models that have the ability to predict the arrival of samples at the AL. These data-driven models will have to be able to predict the arrival of IPC samples at the Laboratories with high reliability in terms of arrival intervals, such that the analyst has time to prepare the materials in advance to analyze the samples in time, thus avoiding delays in the production process. Prediction models will have to be adaptable over time and be able to choose automatically the best algorithm for each training time interval.

- Create ML models that are able to predict the consumption of materials in the AL based on historical consumption and the tests used. This algorithm will consume the forecasts of arrival of samples to the Laboratories (which contain the information about the analysis that will be performed) and will have to be able to timely forecast the requests of materials to the Warehouse, such that the laboratory does not have a shortage of materials for the analysis to be performed. If this happens, it may lead to delays in the analysis and, consequently, in the production process.

- Develop models that are able to assign the best instrument for each analysis, taking into account the specific analysis, the maximum load of the instrument and the instrument's capabilities, in order to make a more equitable distribution of the instruments to be assigned to the analyses.

- Develop and evaluate an architecture for a IDSS that can be applied to all ALs in the Chemical Industry within an Industry 4.0 context. This architecture will encompass the aforementioned models, which must also comply with the current conditions in the Chemical Industry. The planned architecture contains a set of models that encompass all types of Gartner Analytics (Descriptive, Predictive, and Prescriptive).

1.4 Research Methodology

In this PhD program, which is essentially a research project that involves the development of an artifact - an IDSS system - we use a Design Research, more specifically the Design Science Research Methodology for Information Systems (DSRM-IS), as our research methodology. This methodology consists of a set of techniques and methods with the objective of developing an IS artifact. Figure 3 presents the methodology and in it we can identify its five steps: Awareness of Problem, Suggestion, Development, Evaluation and Conclusion. In this section, each of these steps will be detailed when applied in the development of this PhD.

Within the Chemical Industry, there are AL that are very important for the proper functioning of the industry because it is in these Laboratories that all products used and produced in the organization are analyzed to confirm that they are within the quality standards. However, currently much of the work that is done in the laboratory is manual and the communication that the laboratory makes with other entities is done manually, and delays in these processes can lead to pauses in production, which is not desirable. The first step of this project is the awareness of this issue, where a Systematic Literature Review was

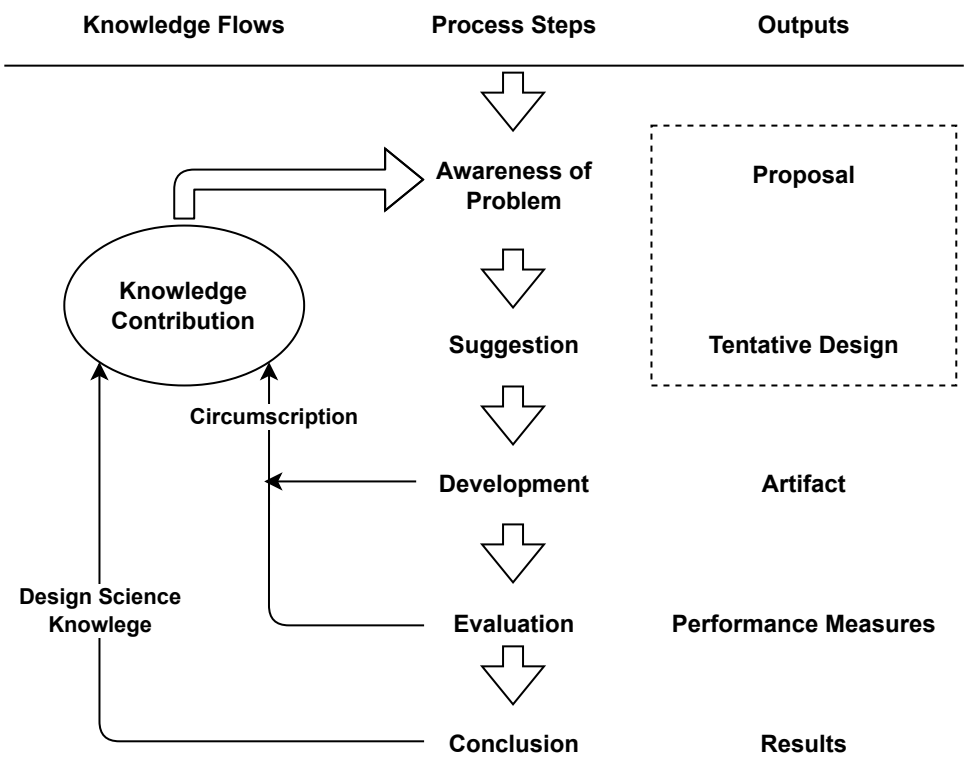


Figure 3: DSRM-IS Model, adapted from Vaishnavi and Kuechler (2004).

performed and concluded that this situation in the AL can be improved with a set of novel techniques, along with new contributions that can be made in terms of business and body of knowledge. Therefore, the problem formulated arises that with the creation of an integrated system that uses different Business Analytics techniques (Descriptive, Predictive and Prescriptive), it will help the decision making process in Laboratories, as well as the streamlining of some of their processes. Furthermore, the exploration and implementation of these techniques can lead to scientific contributions within the subject of Business Analytics. The proposed solution was designed in the next step of the methodology.

The next step of DSRM-IS aims at presenting the proposed solution to the problem formulated, where it will have to be implemented and evaluated in order to increase the body of existing knowledge and contribute to the resolution of the problem mentioned. As far as this PhD project is concerned, the proposed solution consists of an IDSS that uses Descriptive, Predictive and Prescriptive Business Analytics techniques to help decision making in AL.

After the presentation of the proposed solution, the next step in this methodology is related to the development of the proposed solution. It is important to mention that the DSRM-IS methodology is cyclic, which means that the previous phases can be subject to change whenever such change is needed. For the development of this solution, it was divided into three phases, where the first phase was to create a hybrid architecture that uses Automated Machine Learning (AutoML) to predict the arrival of IPC samples to the ALs. The second phase of the development, focused on predicting the consumption of materials in the AL

also using AutoML. Finally, the last phase, was the development of the method of allocating instruments to analyses to be performed and the creation of the IDSS dashboards. To develop this solution, R and Python programming languages were used. The output of this step is our IS artifact: an IDSS for AL in the Chemical Industry.

Once an artifact is created, the DSRM-IS methodology assumes that it is evaluated by considering performance measures. For the predictive algorithms, to evaluate the performance, we used a wide range of performance metrics, namely Mean Absolute Error (MAE), Normalized Mean Absolute Error (NMAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Regression Error Characteristic (REC), Area of Regression Error Characteristic (AREC) and R^2. To evaluate the performance of our algorithms over time, we use a Rolling Window (RW) with a 20-week window to ensure the robustness of our algorithms. To evaluate our fully integrated IDSS, we uses questionnaires with 10 questions from Technology Acceptance Model (TAM) version 3 (Venkatesh & Bala, 2008), together with a matrix comparing the current functionalities (As-Is) with the functionalities provided by our IDSS. This evaluation is complemented with open interviews from the laboratory managers.

When the results of the previous step are at a satisfactory level of acceptance, we move on to the last step of this methodology. In this doctoral project, this level of acceptance would mean that the system developed has a better performance, which is statistically proven when compared to existing methods and if the system reaches all the established objectives. Therefore, the conclusions will be withdrawn, whereby this usually means the publication of scientific publications. This project had as a result of its scientific process three conference papers and one journal article published. These articles are detailed in the next section.

1.5 Contributions

This thesis includes a collection of research paper that were written during the execution of this PhD project with the goal of reaching the research objectives outlined in the previous section.

Section 2.1 presents a SLR regarding the use of Business Analytics techniques in Industry 4.0 covering a selection of 169 papers obtained from six major scientific publication sources from 2010 to March 2020. The selected papers were first classified in three major types, namely, Practical Application, Reviews and Framework Proposal. Then, we analyzed with more detail the practical application studies, which were further divided into the three main categories of the Gartner analytical maturity model: Descriptive Analytics, Predictive Analytics and Prescriptive Analytics. In particular, we characterized the distinct analytics studies in terms of the Industry application and data context used, impact in terms of their Technology Readiness Level (TRL) and selected data modeling method. Our SLR analysis provides a mapping of how data-based Industry 4.0 expert systems are currently used, disclosing also research gaps and future research opportunities. This work resulted in a journal paper:

- A. J. Silva, P. Cortez, C. Pereira, A. Pilastri, **Business Analytics in Industry 4.0: A systematic**

review *Expert Systems*, e12741 (2021) DOI: https://doi.org/10.1111/exsy.12741

Section 3.2 addresses one of the major problems in the AL, the sample arrival, in this specific case, the IPC samples, as those have priority to be analyzed in order to avoid the stop of the production. The forecasting of sample arrival at the Laboratories is crucial for preparing the analytical materials on time, in order to analyze those samples at the Laboratories. To predict the sample arrival, different Cross-Industry Standard Process for Data Mining (CRISP-DM) iterations were performed, each focusing on a different regression approach. An AutoML was adopted during the modeling stage of CRISP-DM. Using recent real-world data from the Chemical organization, it was concluded that a proposed two-stage Machine Learning (ML) model was competitive and provided interesting predictions to support the laboratory management decisions (e.g.,preparation of testing instruments). This work was published in the following conference:

- A. J. Silva, P. Cortez, A. Pilastri, **Chemical Laboratories 4.0: A Two-stage Machine Learning System for Predicting the Arrival of Samples** *Artificial Intelligence Applications and Innovations*, Springer International Publishing, 232-243 (2020) DOI: https://doi.org/10.1007/978-3-030-49186-4_20

Section 3.3 focused also on the improvement of the AL management, more specifically the stock management, where the goal was to predict the material consumption in the AL based on the week plans of samples analysis using ML techniques. Several CRISP-DM iterations were performed and, to reduce the modeling effort, an AutoML was used to select the best ML model. Using real data from the Chemical company and a realistic rolling window evaluation, several ML train and test iterations were executed. The AutoML results were compared with two time series forecasting methods, the ARIMA methodology and a deep learning Long Short-Term Memory (LSTM) model. Overall, competitive results were achieved by the best AutoML models, particularly for the top 10 set of materials. This study resulted in the following conference paper:

- A. J. Silva, P. Cortez **An Automated Machine Learning Approach for Predicting Chemical Laboratory Material Consumption** *Artificial Intelligence Applications and Innovations*, Springer International Publishing, 105-116 (2021) DOI: https://doi.org/10.1007/978-3-030-79150-6_9

Section 3.4 presents an IDSS to enhance the management of AL of a company operating in the Chemical Industry. This IDSS incorporates two predictive ML models, related with the prediction of the arrival of samples at the AL and the consumption of AL materials, which are then used to perform Prescriptive Analytics for AL instrument allocation tasks. The IDSS is also complemented with Descriptive Analytics of instrument similarities regarding the tests performed, for better supporting the AL manager decisions. The IDSS includes interactive dashboards and it was successfully validated by the AL managers using the TAM model 3 and open interviews, which resulted in a positive feedback. This work was accepted and therefore published in the following conference:

- A. J. Silva, P. Cortez **An Industry 4.0 Intelligent Decision Support System for Analytical Laboratories** *Artificial Intelligence Applications and Innovations*, Springer International Publishing, 159-169 (2022) DOI: https://doi.org/10.1007/978-3-031-08337-2_14

1.6 Thesis Organization

This thesis is structured as follows:

- Chapter 1 describes the motivation, problem formulation, research objectives, contributions and PhD organization of this thesis.
- Chapter 2 presents the main background associated with this PhD work. The first section details a SLR study that was performed on the topic of Business Analytics applied within the Industry 4.0 concept. The second section presents additional topics that were not discussed in the SLR but that are relevant for this PhD. Finally, the third section surveys studies that involve the usage of Business Analytics within the Chemical industry domain.
- Chapter 3 presents the proposed methods and conducted experiments that led to the design of the proposed IDSS. In Section 3.2 we present the two-stage model to predict the arrival of samples at the AL. In Section 3.3 we present the model for predicting material consumption in the AL. And in Section 3.4 we detail the full designed IDSS, which includes a prescriptive model for instrument allocation as well as the development of the Dashboards used in the IDSS.
- Finally, Chapter 4 summarizes the main conclusions of this PhD work, discussing some of its main impacts and limitations. Moreover, the last section presents future lines of research.

Chapter 2

Background

This chapter presents an SLR and the theoretical concepts that are relevant for this PhD. The need to create a SLR on the use of Business Analytics in Industry 4.0 arose after a bibliographic research on literature reviews in this topic. In effect, it was concluded that there were no surveys that properly addressed the targeted topic. Since the published SLR is more broad (in terms of industry sections), we end this chapter with a section that specifically targets the state of the art works regarding the application of Business Analytics to the Chemical industry domain.

2.1 Business Analytics in Industry 4.0: A Systematic Literature Review

2.1.1 Introduction

In the recent years, several industry sectors are being changed through the adoption of Information and Communication Technologies (ICT). More digital and connected sensors are being added to production systems, generating Big Data that can be processed using analytical systems, allowing to produce new insights and knowledge about the productive processes. Born in Germany in 2011 (BMBF, 2011), the term "Industry 4.0"is widely used to identify this fourth industrial revolution. Indeed, the German Federal Ministry of Education and Research defines the Industry 4.0 concept as *"the flexibility that exists in value-creating networks is increased by the application of cyber physical production systems. This enables machines and plants to adapt their behavior to changing orders and operating conditions through self-optimization and reconfiguration. ... The main focus is on the ability of the systems to perceive information, to derive findings from it and to change their behavior accordingly, and to store knowledge gained from experience. Intelligent production systems and processes as well as suitable engineering methods and tools will be a key factor to successfully implement distributed and interconnected production facilities in future Smart Factories"* (Shrouf et al., 2014).

Business Analytics is a major ICT tool for the Industry 4.0. It focuses in the analysis of historical

raw data in order to achieve useful and focused insights and a better understanding of the business performance areas (Krishnamoorthi & Mathew, 2018). Business Analytics is an expert systems subarea that results from the combination of Business Intelligence techniques with Optimization, Forecasting, Predictive Modeling and Statistical Analysis (Arnott & Pervan, 2014). Business Analytics systems are being increasingly applied in the Industry sector, thus behaving as the data intelligence component of the Industry 4.0. Indeed, Business Analytics can bring new advantages to the organizations such as product and process digitization, the creation of new products, services and solutions, the offering of Big Data Analytics as a service, the breadth of product customization and the mass production of custom products. There are also other potential advantages for industries, such as obtaining larger profit margins and increasing the market share of key business products by gaining valuable insights from customers using Data Analytics (Geissbauer et al., 2016). Industries can gain efficiencies and lower costs by using real-time production line controls via Big Data Analytics. In addition, the Industry 4.0 offers production concepts that are modular, flexible and customer-tailored. Real-time visualization of the production process and variance of the product, as well as the use of data analytics for optimization and augmented reality, have emerged with the context of Industry 4.0. Predictive maintenance is another advantage that arises in this context because it uses forecasting algorithms to optimize the maintenance and repair processes. An increased vertical integration can be obtained by using sensors through the manufacturing execution system, allowing a real-time production planning with the objective of obtaining greater efficiency in terms of machine occupation times. Horizontal integration is another efficiency gain that allows track-and-trace products for better inventory management and improved operating speeds. Other efficiency gains include the digitization and automation of processes for a more efficient use of human resources (Geissbauer et al., 2016).

Given the emergence of this topic, this chapter performs a SLR on the usage of Business Analytics within the Industry 4.0 concept, which a particular focus on practical applications and three main types of analytics (Descriptive, Predictive and Prescriptive). The specific Research Question addressed by this SLR is: ***How and in what areas of the industry are Business Analytics techniques being used within an Industry 4.0 context?*** To answer the Research Question, a total of 169 papers, from 2010 to March 2020, were selected for the review. Then, the practical studies were further analyzed, allowing to identify the specific industrial context where analytics were used (e.g., business goal, data used), the selected modeling method (e.g., analysis of variance, artificial neural networks) and the obtained impact. Thus, the performed SLR characterization summarizes how Industry 4.0 expert systems are being used, also disclosing current research gaps that can be addressed in future research works.

2.1.1.1 Business Analytics

The Business Analytics topic assumes the Big Data age in an extensive manner. It also includes useful data processing decision support methods, namely Optimization, Forecasting, Predictive Modeling, and Statistical Analysis. The goal is to extract useful, often actionable knowledge from historical data based

on advanced Artificial Intelligence (AI) analytics (Arnott & Pervan, 2014; G. Cao et al., 2015; H. Chen et al., 2020; Koch, 2015; Lu, 2019). In 2013, the the famous Gartner Group defined four main types of analytics: Descriptive, Diagnostic, Predictive and Prescriptive.

The Descriptive analysis attempts to answer the question "what happened?". Business Intelligence and Big Data systems (e.g., Data Warehousing) can be used to access the historical data and provide summarization reports, visualizations and dashboards (e.g., pie charts, bar charts, table or generated views). Next, the Diagnostic analysis aims to understand "why did it happen?", using mostly exploratory data analysis techniques via a interaction with the data analyst which is looking for insights. For example, by visualizing drill down/up operations of an online analytical processing tool of a Data Warehousing. Then, the Predictive analysis aims to answer the question "what will happen?". This can be achieved by using Statistical Analysis and Machine Learning (e.g., Classification, Regression, Time-Series Forecasting). Predictive Analytics are being used in diverse application domain areas, such as Marketing (Chi-Hsien & Nagasawa, 2019) and Finance (Swamy & Sarojamma, 2020). The last and most difficult analytic type is termed Prescriptive Analysis and it is related with the question "how can make it happen?". This type of analytics can be achieved by using diverse techniques, including Simulation, What-if scenarios, Machine Learning, Heuristics and Optimization. We note that Diagnostic analytics are often difficult to distinguish from Descriptive ones, since both are assumed to analyze historical data and are often performed simultaneously by the same analysts. Thus, in this SLR, we adopt the same strategy used by Chong and Shi (2015) and Khatri and Samuel (2019), which group all historical analyses (Descriptive or Diagnostic) into a single Descriptive analytics category.

2.1.1.2 Industry 4.0

The Industry 4.0 is defined as the global transformation of the manufacturing industry through the introduction of digitalization and the Internet. The transformations applied imply enormous advances in the design and the manufacturing processes, operations and services of manufacturing products and systems. The term Industry 4.0 was coined in Germany in 2011 and it shares similarities with developments produced in many European countries and that have been labelled differently, as Smart Factories, Smart Industry, Advanced Manufacturing of Internet of Things Internet of Things (IoT) (BMBF, 2011; Tjahjono et al., 2017). The term Industry 4.0 was born in Germany because the German engineers realized that manufacturing had been developed into a new paradigm shift, where products tend to control their own manufacturing process (Lasi et al., 2014). The Industry 4.0 is considered the fourth industrial revolution, which contains a extreme potential impact in the future (Kagermann et al., 2013). Smart Factories use new technologies, such as advanced robotics and Artificial Intelligence (AI), cloud computing, IoT, Data Analytics, Software-as-a-Service and platforms that use algorithms to direct motor vehicles, delivery and ride services, and the embedding of all these elements, and many more, in an interoperable global value chain, shared by many companies from different countries (Geissbauer et al., 2016).

Until recently, the term Industry 4.0 has not yet been conclusively defined, neither are its features.

Nevertheless, there are four main features that typically categorizes the term (Tjahjono et al., 2017): vertical networking of smart production systems; horizontal integration via a new generation of global value chain networks; through-life engineering support across the entire value chain; and acceleration through exponential technologies. This perspective of the analysis is believed to be relevant since there is no complete or concise knowledge of how to implement Industry 4.0 correctly or predict future problems to be prevented in advance. The use of IoT and Cyber-Physical Systems (CPS) on Industry 4.0 made possible the connection between materials, sensors, machines, products, supply chain, and customers, which means these necessary objects are going to exchange information and control actions with each other independently and autonomously. The technologies that support the Industry 4.0 concept are the IoT, CPS, Cloud Computing and Big Data Analytics (Lasinkas, 2017). These concepts are described in the next subsections.

Internet of Things The concept of IoT describes an inter-networking world where various objects inside of that world are embedded with sensors, and other digital devices, so they can be networked in order to be possible to collect and exchange data from them (Xia et al., 2012). IoT-enabled manufacturing features real-time data collection and sharing among various manufacturing resources such as machines, workers, materials, and jobs. Usually, the IoT can provide advanced connectivity of various objects, systems and services, and enable data sharing. IoT is particularly useful for industries (R. Y. Zhong et al., 2017). In the future, it is expected to occur a convergence of IoT-related technologies, such as ubiquitous wireless standards, Data Analytics and Machine Learning (L. D. Xu et al., 2014; R. Y. Zhong et al., 2017). IoT is being applied in other sectors besides Industry, such as in the Healthcare area where IoT is being combined with Machine Learning techniques to predict lung cancer in patients (Pradhan & Chawla, 2020).

The Radio-Frequency Identification (RFID) is an example of a technology that is used in IoT. The manufacturing industry will be affected by this change because RFID is used for identifying various objects in warehouses, distribution centers, production shop floors, logistic companies and disposal/recycle stages (Y.-M. Wang et al., 2010). The identifiers have smart sensing abilities, and they can connect and interact with other objects, which may create a huge amount of data from their movements and behaviors. These objects are given specific applications or logics, so that they can be followed after being equipped with the RFID readers and tags (Guo et al., 2015). RFID can also capture data related to the daily operations so that production management is achieved on a real-time basis (R. Y. Zhong et al., 2017).

Cyber-Physical Systems A CPS involves a various number of methodologies such as cybernetics theory, mechanical engineering and mechatronics, design and process science, manufacturing systems, and computer science. The ability to have highly coordinated and combined relationships between physical objects and their computational elements or services is one of the key elements of a CPS (Tan et al., 2008). Unlike a traditional embedded system, the CPS contains networked interactions that are developed and designed with inputs and outputs, along with their cyberwined services, such as computational capacities and control algorithms. A large number of sensors have crucial roles in a CPS. For example, multiple

sensory devices can be used in a large number of purposes such touch screens, light sensors, and force sensors (R. Y. Zhong et al., 2017).

One example of a real-world project with CPS is the Festo Motion Terminal, which aims to create a standardized platform that makes full use of an intelligent fusion of mechanics, electronics, embedded sensors and control (R. Y. Zhong et al., 2017). However, the typical CPS applications have been reported for using sensor-based communication-enabled autonomous systems, and a various number of wireless sensor networks can supervise aspects such as environmental so that information can be centrally controlled and managed for decision-making (Ali et al., 2015).

Cloud Computing The general term that describes computational services through visual and scalable resources over the Internet is cloud computing (Armbrust et al., 2010; X. Xu, 2012). Cloud computing is interesting for business owners because the advantage of scalability allows organizations to start small and invest in more resources if the service demand goes up (Q. Zhang et al., 2010).

An ideal cloud service must have these five characteristics (Mell & Grance, 2011): on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. The ideal cloud service model is composed of four deployment models (public, private, community, and hybrid) and three delivery models (Software-as-a-Service, Platform-as-a-Service, and Infrastructure-as-a-Service). Cloud computing services are being implemented by all kind of organizations to increase their capacity with a minimum budget investment, as cloud computing does not require investments in new software, incorporate new infrastructures or train new personal (Saxena & Pushkar, 2016).

Despite the benefits of cloud computing, this technology also has challenges, in particular related with privacy and security concerns. Other challenges, such as data management and resource allocation, scalability and communication between clouds, reduce the reliability and efficiency of cloud-based systems (R. Y. Zhong et al., 2017). Because of its relative innovation and increasing development in recent years, a great body of research has been conducted on cloud computing (H. Yang & Tate, 2009).

Big Data Analytics The Big Data trend was mostly motivated by the use of Internet and IoT technologies, which generate vast amounts of data in various industries (Manyika et al., 2011). Big Data stems from various channels such as sensors, devices, networks, transitional applications and social media feeds (Rich, 2012). The Big Data environment has gradually taken shape in the manufacturing sector. Besides the advance of the IoT and the collection of data, there are questions to be resolved, such as how to collect and store the Big Data obtained from real-time sensors which can be processed properly in order to provide the right information for the right question at the right time (J. Lee et al., 2013). Y. Chen et al. (2016) defined Big Data Analytics as the fusion of Big Data and IoT technologies that created opportunities for the development of services for smart environments like smart cities. Nowadays, there are a set of Big Data technologies available to process the large data obtained from the IoT devices which have emerged as a need to process the data collected from different sources in the smart environment.

The Big Data datasets are much larger than the normal datasets, thus they can be too complex for conventional data analytics software (Barton & Court, 2012). As such, it is essential for organizations and manufacturers with vast operational shop-floor data to have advanced analytics techniques for uncovering hidden patterns and unknown correlations between the data, or other things such as market trends, customer preferences and other information useful for the business (R. Y. Zhong et al., 2017). The particular concept of Big Data Warehouse (BDW) emerged due to the studies made about the applications of BDW in Big Data (Krishnan, 2013; Mohanty et al., 2013). Actually, the state-of-the-art refers that the design of BDW should focus on the physical layer and logical layer using two strategies. The first strategy, "lift and shift", is the use of Big Data technologies to solve specific cases and augment the capabilities of traditional and relational Data Warehouses. However, the use of a case driven approach instead of a data modeling approach can lead to possible uncoordinated data silos (Clegg, 2015; Russom, 2014). The second strategy, "rip and replace", is where occurs a replace of the Data Warehouse in favor of Big Data Technologies (Russom, 2014, 2016). In this field, a number of literature reviews were performed; however, they did not focused on the application of Big Data Analytics in Industry 4.0. Duan and Xiong (2015) performed a literature review about the use of Big Data Analytics and Business Analytics, and they concluded that the Big Data concept implies the investment in equipment to capture and store data combined with a Business Analytics approach linked to each business strategy and organizational process, and being aware with the evolving of the state-of-the-art techniques in Big Data. Chong and Shi (2015) studied the use of Big Data Analytics and concluded that these techniques can help the decision-making process, increase the business model understanding, and reveal hidden information to attain competitive advantage.

2.1.2 Related Work

There are several reviews about the implementation of analytical techniques in the Industry 4.0 context. O'Donovan et al. (2015) made a mapping study about the use of Big Data in the manufacturing sector. The research method was performed manually. Chiang et al. (2017) reviewed the recent advances of Big Data in data-driven approaches in five industries inside the manufacturing sector. Similarly to O'Donovan et al. (2015), the study only focused on the use of Descriptive and Diagnostic Analytics techniques. Nikolic et al. (2017) made a review about Predictive Analytics in the Industry 4.0 context. The authors searched and reviewed different types of Predictive Maintenance systems. They provided an overview of the various challenges, existing solutions, and benefits of Predictive Manufacturing systems in Industry 4.0. Qi and Tao (2018) reviewed the state of Big Data and digital twins in the manufacturing sector. This review also included the applications in product design, production planning, manufacturing, and predictive maintenance. On this basis, the similarities and differences between Big Data and digital twins were compared from the general and data perspectives. Uhlmann et al. (2017) made a literature review about the historical development of intelligent production systems in the context of adding value to business models. They focused on techniques such as the use of barcodes, RFID, and wireless sensor nodes to make condition

Table 1: Literature surveys about the topic of Business Analytics in Industry 4.0.

Reference	Industry Sector	Search Method [a]	Descriptive Analytics	Predictive Analytics	Prescriptive Analytics
O'Donovan et al. (2015)	Manufacturing	SLR	X		
Chiang et al. (2017)	Manufacturing	Manual	X		
Nikolic et al. (2017)	Manufacturing	AA		X	
Uhlmann et al. (2017)	Manufacturing	Manual		X	
X. Xu and Hua (2017)	Manufacturing	AA	X	X	
J. Yang et al. (2017)	Manufacturing	AA		X	
Bordeleau et al. (2018)	All Industry	SLR	X		
Sharp et al. (2018)	Manufacturing	AA		X	X
Diez-Olivan et al. (2018)	Manufacturing	Manual	X	X	X
Qi and Tao (2018)	Manufacturing	Manual	X		
Muhuri et al. (2019)	All Industry	AA	X	X	
Bakar et al. (2019)	Manufacturing	Manual			X
This Review	**All Industry**	**SLR**	**X**	**X**	**X**

[a] Automatic Analysis (AA), Systematic Literature Review (SLR)

monitoring and Predictive Maintenance in the availability oriented business model. They also studied, based on practical examples, the organizational prerequisites for an implementation of these techniques in the industry. X. Xu and Hua (2017) summarized and analyzed the current research status for industrial Big Data Analysis in smart factories (both domestic and abroad). Also, they proposed research strategies for Industrial Big Data Analysis, including acquisition schemes, ontology modeling, predictive diagnostic methods based on Deep Neural Networks (DNN) and three-dimensional self-organized reconfiguration mechanism. In the area of Augmented Reality solutions, J. Yang et al. (2017) presented a comprehensive survey of AI in 3D painting to detect defective products in the Industry 4.0 context. The survey only analyzed Predictive Analytics techniques. Bordeleau et al. (2018) also performed a literature review of Business Intelligence in the context of Industry 4.0. The goal was to understand how Business Intelligence and data analysis generate value creation in manufacturing companies. This review only studied Descriptive Analytics. Sharp et al. (2018) presented another literature review about the use and development of Machine Learning in smart manufacturing. We note that this review studied practical cases that used Machine Learning in contexts different to the Industry 4.0 context. They reviewed the articles published between 2007 until 2017, while the Industry 4.0 concept was introduced in the 2010s. Moreover, the authors only analyzed the Diagnostic and Predictive Analytics. More recently, Diez-Olivan et al. (2018) presented a survey of the recent developments in data fusion and Machine Learning for industrial prognosis during the Industry 4.0 context. In the same year, Muhuri et al. (2019) performed a literature review about the growth of the Industry 4.0 in the last years. Bakar et al. (2019) presented a survey regarding the use of Metaheuristics techniques and Robotic Assembly Line Balancing in the Manufacturing industry. This SLR review is more focused on the whole Industry 4.0 concept, and thus it does not detail much the Business Analytics methods.

A summary of the related work is presented in Table 1. None of the reviews analyzed addressed all main Gartner's Analytical levels. In contrast, this SLR contains a stronger focus on the Descriptive, Predictive and Prescriptive analytics, when applied to the context of the Industry 4.0. Moreover, we

Table 2: Summary of the literature search protocol.

Subject	Business Analytics in Industry 4.0
Time period	January 2010 to March 2020
Search Engines	Scopus, ScienceDirect, SpringerLink, IEEE Xplore, Google Scholar, Google Books
Search Criteria	English; Title, abstract and keywords OR All (except full text)
Search Query	"Industry 4.0 + Decision Support Systems", "Industry 4.0 + Business Analytics", "Industry 4.0 + Predictive Analytics", "Industry 4.0 + Machine Learning", "Industry 4.0 + Data Mining", "Industry 4.0 + Text Mining", "Industry 4.0 + Process Mining", "Industry 4.0 + Forecasting", "Industry 4.0 + Metaheuristic"

particularly detail the practical applications, allowing to characterize the main business goals, data usage, modeling methods and obtained impacts. It should also be noted that most surveys consider only the Manufacturing sector, which is where the Industry 4.0 concept is producing a higher impact. Indeed, while this SLR considers all industry sectors, the selected practical research works in this SLR are highly related with the Manufacturing sector (as shown in Section 2.1.4.1).

2.1.3 Literature Review Method

2.1.3.1 Paper Collection

We performed a manual SLR review, similar to what was proposed by Kitchenham et al. (2009). For this literature review, we used several scientific search engines, in order to search for the relevant documents: Google Scholar (https://scholar.google.com/), Google Books (https://books.google.com/), ScienceDirect (https://www.sciencedirect.com/), SpringerLink (https://link.springer.com/), Scopus (https://www.scopus.com/home.uri) and IEEE Xplore (https://ieeexplore.ieee.org/Xplore/home.jsp). The term "Industry 4.0"was coined in 2010. As shown in Figure 4, the Web interest in the term starts from 2010, although the popularity only increases substantially after 2014. Thus, we have retrieved articles that were published since 2010 until March 2020 (when this SLR was executed). Using the listed search engines, we performed several queries, using the combinations of the following keywords: "Industry 4.0", "Decision Support Systems", "Business Analytics", "Predictive Analytics", "Machine Learning", "Data Mining", "Text Mining", "Process Mining", "Forecasting", and "Metaheuristic". Table 2 presents the literature search protocol used during this SLR.

Paper Selection Table 3 presents the distribution numbers of the collected scientific publications for the different search engines used. The paper search queries resulted in a total of 285 articles. All retrieved documents were manually inspected to check their relevance. First, the title and abstract was read. When the abstract was not conclusive, a more in-depth reading of the article was performed, in order to verify if the document fits the SLR goal. The manual inspection filtered 116 papers that were considered irrelevant

Figure 4: Evolution of the interest in the term "Industry 4.0"in Google Trends.

Table 3: Distribution of papers obtained by each database.

Database	**Quantity**
Scopus	202
ScienceDirect	75
SpringerLink	35
IEEE Xplore	60
Google Scholar	35
Google Books	5
Total with duplicates	390
Total without duplicates	285

for the survey, thus resulting in a total of 169 articles that were selected.

2.1.4 Literature Review Analysis

2.1.4.1 Quantitative Analysis

As stated earlier, 169 papers were selected for this literature review. To make a general overview about the papers selected, a quantitative analysis was performed, in which the papers are characterized according to the year of publication and the paper type.

Paper type The papers collected were manually inspected and divided into the three different categories proposed in Öchsner (2013):

- *Practical Application* - These papers describe and discuss real implementation results of a framework, methodology, method or IT in one or more application domain areas;
- *Reviews* - Articles of literature review (such as this SLR), with the main objective of performing a survey of the state-of-the-art on a certain scientific research topic area, possibly identifying research gaps; and
- *Framework Proposal* - The aim is to document the proposal of a new framework developed by the authors. However, these articles do not have a specific application target, thus the authors do not validate the framework in a real-world environment.

Table 4 shows the respective distribution of the selected 169 papers in terms of the three main paper categories. The majority of the selected papers are Practical Application ones (139 papers). There are 11

Table 4: Distribution of the three main paper types.

Paper Type	**Quantity**
Practical Application	139
Reviews	12
Framework Proposal	18
Total	169

papers that were categorized as Reviews and 18 publications categorized as Framework Proposal. Given that this survey is more focused on practical usage of Business Analytics, we will only further detail and analyze the 139 Practical Application studies. The quantitative analysis includes the industry sector, the Gartner Analytic type and year, and finally the paper keyword frequencies.

Table 5: Distribution of the Practical Applications per industry sector.

Industry Sector	**Quantity**
Manufacturing	130
Transportation and Warehousing and Utilities	3
Construction	2
Educational Services, and Health Care and Social Assistance	1
Agriculture, Forestry, Fishing, and Hunting, and Mining	2
Finance and Insurance, and Real Estate, and Rental and Leasing	1
Total	139

Industry sectors of the Practical Applications To describe the Industry sections we adopted the Standard Industrial Classification Bureau, 2017, which includes five main categories listed in Table 5.

The Manufacturing sector is by a large margin the sector with most Industry 4.0 practical applications of Business Analytics, with 130 papers. This happens because the manufacturing sector is a vast sector that includes a relevant number of production processes, widely used by several industries. The manufacturing sector has also high Business Analytics needs. For instance, the shop floor usually has different kinds of machines, which should work efficiently and produce quality products. Thus, Predictive Maintenance and automatic quality inspection/prediction methods, based on data-driven models, can be used to enhance the manufacturing process.

The other industry sectors have much less practical application works. Within the Transportation and Warehousing and Utilities sector, the surveyed papers relate with three practical applications. In the Eolic Energy area, Canizo et al. (2017) presented a data-driven solution deployed in a cloud that used Random Forest (RF) for predicting failures on wind turbines. In the transformation energy field, Bagheri et al. (2018) analyzed the analytical approach to the transformer vibration modeling, using Machine Learning techniques such as Linear Regression (LinR), Model Trees, Support Vector Regression with Gaussian Kernel and Multilayer Perceptron, and also signal techniques to develop prognosis models of transformer operating condition based on vibration signals. Masoudinejad et al. (2018) proposed a set of Support Vector Machine (SVM) algorithms, addressing indoor localization within a warehouse. The Construction sector has two practical applications. J. Lee et al. (2014) made a review about the trend of the manufacturing service transformation in Big Data and proposed a framework for sustainable innovative service. The data used to make the case study came from sensors installed in a bulldozer. They used a Bayesian Belief Network to classify if the engine had some problem or malfunction and used a Fuzzy-Logic based algorithm to predict the remaining useful life of the engine. R. Costa et al. (2017) proposed a system with the aim to create knowledge representations from unstructured data sources used in a construction environment, based on enriched semantic vectors.

Regarding the Educational Services, and Health Care and Social Assistance sector, Bordel and Alcarria (2017) presented a solution to automatically assess the human motivation in Industry 4.0 scenarios with the use of an ambient intelligence infrastructure. Turning to the Agriculture, Forestry, Fishing, and Hunting,

Table 6: Distribution of the Practical Applications for the three Analytics types and year of publication.

Year	Descriptive Analytics	Predictive Analytics	Prescriptive Analytics	Total
2015	2	0	0	2
2016	2	8	2	12
2017	8	17	2	27
2018	9	21	5	35
2019	2	25	10	37
2020	0	9	8	17
Total	23	80	27	130

and Mining sector, Teschemacher and Reinhart (2017) used Ant-Colony Optimization algorithms to enable dynamic milk-run logistics. Also, Dutta et al. (2018) implemented a Machine Learning based interactive architecture for industrial scale prediction for dynamic distribution of water resources across the continent and, at the same time, keeping four corners of Industry 4.0 in place. The algorithms tested were LinR, Bayesian Ridge Regression, Logistic Regression (LogR), Linear Discriminant Analysis, Adaptive Neuro-Fuzzy Inference System, Multi-Layer Perceptron, and Radial Basis Function Network. Finally, within the Finance and Insurance, and Real Estate, and Rental and Leasing sector, Ma and Li (2018) used a Grey Model to predict eight indexes of the tertiary industry.

2.1.4.2 Analytics Type

Table 6 shows the distribution of the selected Practical Application papers in terms of publication year and analytics type. The most common type is the Predictive Analytics level, with 80 applications, followed by the Prescriptive Analytics, with 27 applications, while the Descriptive Analytics were only addressed in 23 applications. The smaller number associated with the Prescriptive and Descriptive Analytics denote an important research gap. The lack of further Prescriptive studies is probably due to two main reasons. Firstly, the Industry 4.0 concept implementation is very recent (just a few years). Most of its initial implementation effort is devoted to setting the right infrastructure to generate and collect data, and Business Analytics can only be applied after collecting enough historical data. Secondly, Prescriptive Analytics are more complex than other types of data analyses (Koch, 2015). As more mature Industry 4.0 applications are implemented, we expect this gap to be reduced. It is also interesting to note that there are more Predictive Application studies than Descriptive ones. This behavior might be explained by the current Machine Learning hype. Also, building a stable and valuable Data Warehousing system, which results in better Descriptive analysis, requires several Extract, Transform, Load (ETL) processes that are often costly, requiring manual effort and time, but that do not tend to translate into novel methodologies or interesting application usages that justify a research publication. Overall, the yearly numbers from Table 6 show a substantial growth in the number of publications starting from 2017: 27 papers in 2017; 35 works in 2018; and 37 research publications in 2019 (the 8 papers from the year of 2020 report only until the

	Term	Frequency
1	industry	74
2	data	63
3	learning	46
4	manufacturing	42
5	maintenance	35
6	machine	34
7	predictive	34
8	big	31
9	smart	23
10	analytics	21

Figure 5: Word cloud of the keywords (left) and top 10 term frequency values (right).

month of March).

Keywords frequencies The last quantitative analysis is obtained by applying a word cloud technique to the 112 application paper keywords. We have selected keywords because these help to index and classify papers, facilitating research queries. The word cloud analysis was performed using R tool with the package `wordcloud`. The word cloud is presented in Figure 5, which also details the top term frequency numeric values. The most frequent term is "Industry", followed by "data", "learning"and "manufacturing". Other terms such as "maintenance", "machine"and "predictive" are also popular, which aligns with Table 6, since most practical applications use Predictive Analytics.

2.1.4.3 Qualitative Analysis

The qualitative analysis was executed by a manual inspection of the selected practical papers. The description of practical cases are divided by the analytics type (Descriptive, Predictive and Prescriptive), using a chronological order. Each practical application is briefly described, including the:

- **Function** – Industry 4.0 function area, which is categorized by the four main functions of the Industry 4.0 architecture presented by Qin et al. (2016): Hardware Connection (HC), focuses on hardware development (e.g., sensor network); Information Discovery (ID), where the raw data is transformed into useful knowledge; Predictive Maintenance (PdM), aiming to anticipate maintenance issues; and Intelligent Production (IP), automating or adapting the production process.

- **Data** – type of industry data used (e.g., generated by a production machine, captured image).

Table 7: Description of the Technology Readiness Levels (TRL).

Phase	Level	Definition
Research	TRL 1	Basic research
	TRL 2	Technology formulation
	TRL 3	Concept validation
Development	TRL 4	Prototype in laboratory environment
	TRL 5	Prototype in relevant environment
	TRL 6	Prototype system tested in relevant environment
Deployment	TRL 7	Demonstration system in operational pre-commercial environment
	TRL 8	First commercial system, ready for operational environment
	TRL 9	Full commercial system with general availability

- **Sector** – addressed industry sector (e.g., aerospacial, automotive).
- **Goal** – brief description of the application goal.
- **Impact** – measured using the TRL scale, from 1 to 9 (Table 7) (ESRTC, 2009).
- **Modeling** – Business Analytics method used to analyse the data.

Table 8: Overview of the Practical Articles that used Descriptive Analytics Techniques

Reference	Func.[1]	Data[2]	Sector[3]	Goal	Impact	Modeling[4]
Neuböck and Schrefl (2015)	ID	Pr	ND	New analysis graphs are proposed for building production insights (e.g., show urgent missing materials).	7	DW, AG
Niño et al. (2015)	IF	MF	CE	Big Data Analytics for pursuing a servitization strategy.	2	DA

[1]Hardware Connection (HC), Information Discovery (ID), Intelligent Production (IP), Predictive Maintenance (PdM)

[2]Car Specification (CS), Grippers (G), Historical (H), Machine (MC), Manufacturing (MF), Production (Pr), Sensor (S), Sparse Data (SD), Temporal Logs (TL)

[3]Additive Manufacturing (AM), Aerospace (As), Automotive (A), Capital Equipment (CE), Chemical Industry (CI), Glass Industry (GI), Not Disclosed (ND), Semiconductor (SC), Spring Manufacturing (SM)

[4] Analysis Graph (AG), Artificial Neural Networks (ANN), Augmented Reality (AR), Back-Propagation Artificial Neural Networks (BPANN), Back-Propagation Neural Networks (BPNN), Browns Double Exponential Smoothing (BDES), Classification Trees (CT), Clustering (CI), Cross-Departmental Data Analytics (CDDA), Data Analysis (DA), Data Warehouse (DW), Decision Trees (DT), Deep Learning (DL), Descriptor Silhouette (DS), Digital Twin (DigT), Failure Mode Metrics (FMM), Fuzzy Logic (FL), Genetic Algorithm (GA), Interpolation Fitting (IF), K-Mean Clustering (KMC), Linear Regression (LinR), Monkey Algorithm (MA), Neural Networks (NN), NeuroEndocrine-Inspired Manufacturing System (NEIMS), Partial Least Square (PLS), Residual Prediction Calculator (RPC), Self Organizing Map (SOM), Simulation (Sim), Standard Silhouette (SS), Two-Stage Clustering (TSC)

Y.-M. Lee et al. (2016)	ID	S	A	Real-time analysis to explore the reasons for abnormality of load rate data of main shaft machine.	5	BPANN, TSC
Tang et al. (2016)	HC	MF	ND	Intelligent architecture for the smart shop floor.	5	NEIMS
Durakbasa et al. (2017)	ID	S	ND	Improve the quality of the manufacturing process.	2	FL
Kirchen et al. (2017)	ID	S	CI	Explore signal data quality.	4	DA
C.-J. Kuo et al. (2017)	IP	S	SM	Explore inexpensive add-on triaxial sensors for the monitoring of machinery.	5	NN
Qin et al. (2017)	ID	MC	AM	Facilitate a better understanding of the energy consumption of digital production processes.	5	LinR, DT, BPNN
Sanz et al. (2017)	ID	S	A	Advanced monitoring of an industrial process that integrates several data sources.	3	BDES
Trunzer et al. (2017)	ID	S	ND	Classify failures in control valves.	4	FMM, GA
Y. Wang et al. (2017)	ID	SD	ND	Methodology to enrich sparse data by fast and frugal reduced models.	3	CI, CT
Zheng and Wu (2017)	ID	Pr	SC	Smart spare parts inventory management system for semiconductors.	5	DA, Sim
Birglen and Schlicht (2018)	ID	G	A	Review the characteristics of pneumatic, parallel, two-finger and industrial grippers.	3	DA
Lenz et al. (2018)	ID	S	ND	Holistic approach for machine data analytics.	2	CDDA
C. Lin and Yang (2018)	HC	S	ND	Intelligent Computing System to connect the different facilities in a logistic center.	6	MA, GA

Mozgova et al. (2018)	ID	S	A	Monitor actual stress state of a structural component and estimate its residual fatigue life.	6	RPC
Ploennigs et al. (2018)	ID, HC	S	ND	Cognitive IoT architecture with scalability and self-learning capabilities.	5	AR
Stürmlinger et al. (2018)	HC	S	ND	Development of a new generation of a manufacturing system.	5	DA
Subakti and Jiang (2018)	ID	MC	ND	Augmented reality system to visualize and interact with machines in smart factories.	7	DL
Tieng et al. (2018)	ID	S	As	Virtual metrology system for sampling.	4	BPNN, PLS, GA, IF
Vathoopan et al. (2018)	HC	H	ND	Corrective maintenance using the digital twin of an automation model.	3	DigT
Kaupp et al. (2019)	IP	TL	GI	Outlier identification to measure the glass quality.	5	NN
Ventura et al. (2019)	ID	S, P	ND	Automatic industrial equipment maintenance system.	6	DS, SS, KMC

Descriptive Analytics Table 8 presents an overview of the practical applications that used Descriptive Analytics techniques. As shown in the table, there is a diversity of Descriptive applications and adopted types of historical analyses. For instance, some studies perform a simple statistical analysis (Birglen & Schlicht, 2018; Lenz et al., 2018; Mozgova et al., 2018; Niño et al., 2015; Sanz et al., 2017; Stürmlinger et al., 2018; Tang et al., 2016; Ventura et al., 2019), while others use more sophisticated outlier detection (Y.-M. Lee et al., 2016; Trunzer et al., 2017) and clustering methods (Y. Wang et al., 2017). Some studies use data warehousing databases and dashboards (Kirchen et al., 2017; Neuböck & Schrefl, 2015; Vathoopan et al., 2018; Zheng & Wu, 2017), and other studies used Neural Networks (Kaupp et al., 2019; C.-J. Kuo et al., 2017; Qin et al., 2017; Subakti & Jiang, 2018; Tieng et al., 2018).

Predictive Analytics The practical applications that used Predictive Analytics techniques are shown in Table 9. Predictive Analytics involve a set of data-driven models that are typically obtained by applying supervised Machine Learning algorithms.

Table 9: Overview of the Practical Articles that used Predictive Analytics Techniques

Reference	**Func.**[5]	**Data**[6]	**Sector**[7]	**Goal**	**Impact**	**Modeling**[8]
Kohlert and König (2016)	ID, PdM	S	PI	Human-machine-based process monitoring and control for yield optimization in polymer film industry.	6	NN, SVM, KNN, NOV-CLASS
H. Kuo and Faricha (2016)	IP	S	SC	Improve the accuracy of grating displacement offset prediction.	4	ANN
T. Lin et al. (2016)	PdM	S	Sp	Triaxial sensors to aid in machine monitoring to facilitate the transition of data.	5	NN, SVM, KNN, NFM

[5] Hardware Connection (HC), Information Discovery (ID), Intelligent Production (IP), Predictive Maintenance (PdM)

[6] Acoustic (Ac), Car Manufacturing (CM), Car Specification (CS), Chemical (Ch), Chemical Laboratory (ChL), Gas Turbine (GT), Gesture Images (GI), Image (I), Machine (Mc), Machine Center (McC), Material (Ma), Network (N), Pellets Images (PI), Production (Pr), Reference Metadata (RM), Robotic (Rb), Sensor (S), Sheet Material (SM), Simulated Sensor (SimS), Solar Panel (SolP), Steel (St), Text (T), Time Series (TS), Welding Images (WI)

[7] Aerospacial (Ae), Automotive (A), Coil (C), Electronic (El), Energy (En), Food (Fo), Footwear (F), Furniture (Fu), Healthcare (Hc), Naval (Na), Not Disclosed (ND), Oil (O) Petrochemical (Pc), Polymer (Pl), Robotic (Rb), Semiconductor (SC), Spring (Sp), Steel Plate (SP), Transportation (Tr)

[8] Adaptive Neuro-Fuzzy Inference Systems (ANFIS), Analysis of Variances (ANOVA), Artificial Neural Networks (ANN), Association Rules (AsR), Backtracking Search Optimization Algorithm (BSOA) Bagged Decision Trees (BDT), Bagged Trees (BagT), Bagging (Bag), Bayesian Filter (BF) Boosting Trees (BosT), Complex Fuzzy (CF), Conference Trees (CT), Convolutional Neural Networks (CNN), Decision Forest (DF), Decision Jungle (DJ), Decision Trees (DT), Deep Learning (DL), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Discriminant Analysis (DA), Extreme Gradient Boosting (EGB), Extreme Learning Machine Boundary (ELMB), Extremely Randomized Trees (ERT), Fast Nearest Neighbors (FaNN), Feed Forward Neural Network (FeNN), Fog Computing (FC), Fuzzy-Logic (FL), Gaussian Model (GM), Gaussian Noise (GN), Genetic Algorithm (GA), Genetic Programming Based Symbolic Regression (GPBSR), Global Local Outliers in Sub Spaces (GLOSS), Gradient Boosted Regression Trees (GBRT), Gradient Boosted Tree Classifier (GBTC), Gradient Boosting (GB), Gradient Boosting Decision Trees (GBDT), Gradient Boosting Machine (GBM), H20 Deep Learning (h2oDL), Hidden Gama Process-Particle Filter (HGP-PF), Hidden Markov (HM), In Situ Classification System (ISCS), Isolation Forest (IF), Kalman Filter (KF), Kurtosis (K), K-Means (KM), K-Nearest Neighbor (KNN), Linear and Polynomial Fit (LPF), Linear Regression (LinR), Local Outlier Factor (LOF), Logistic Regression (LogR), Map Reduce (MR), Matlab Model Predictive Toolbox (MMPT), Mean and Standard Deviation (MSD), Mean Shift (MS), Microsoft Azure Machine Learning (MAML), Micro-Cluster Continuous Outlier Detection (MCCOD), Model Predictive Controller (MPC), Multiple Regression (MR), Multivariate Adaptive Regression Splines (MARS), Multi-Entity Bayesian Networks Regression (MEBNR), Multi-Layer Regression (MLR), Naive Bayes (NB), Neural Networks (NN), Neuro-Fuzzy Networks (NFN), Noise Impulse Integration (NII), Novelty Classifier (NOVCLASS), Out-of-Bag Error (OBE), Partial Least Squares (PLS), Particle Swarm Optimization (PSO), Principal Component Analysis (PCA), Pure Quadratic Regression (PQR), Quadratic Discriminant Analysis (QDA), Random Forest (RF), Random Support Vector Machine (RSVM) Recursive Partitioning (RP), Regression Trees (RT), Ridge Regression (RR), Rule-Based (RB), Skewness (Sk), Spectral and Agglomerative Clustering (SAC), SRT Model (SRTM), Stochastic Model Predictive Controller (SMPC), Support Vector Machines (SVM) Survival Analysis (SA), Time Series Forecasting (TSF), ZeroR (ZR)

Miškuf and Zolotová (2016)	ID	T	ND	Multi-Class Classifiers and Deep Learning in the Industry 4.0 Context.	4	NN, DF, DJ, LogR, SVM, h2oDL
Saldivar, Goh, Chen, et al. (2016)	ID	CS	A	Developed a Predictive Analytics framework to add mass customization.	3	SOM
Saldivar, Goh, Li, Yu, et al. (2016)	IP	CM	A	Predict the decision-making and customize the intelligent product design.	4	FL, GA
Saldivar, Goh, Li, Chen, et al. (2016)	ID	CS	A	Predictive Analytics framework for the automotive area.	4	GA, KM
Stein et al. (2016)	IP	Pr	A	On-line process monitoring and predictive modeling to optimize the car production process.	2	GLOSS
Albers et al. (2017)	ID	Ac	ND	Evaluate the product quality and tool defects by using an acoustic emission sensor.	6	ANOVA
Borgi et al. (2017)	PdM	Rb	ND	Predictive maintenance of industrial robots using movements power condition-monitoring.	6	MSD, Sk, K
Choi et al. (2017)	ID	RM	ND	Deep Learning to analyze and evaluate the performance of the Deep Learning method.	3	DL
Cicconi et al. (2017)	IP	Mc	F	Modeling and simulation of an induction heating process for aluminum-steel mold.	4	MMPT
Gomes et al. (2017)	IP	S	ND	Ambient Intelligent decision support system for creation of standard work procedures.	2	ANN

Haffner et al. (2017)	IP	WI	A	Automatic welding recognition on cloud computing and single-board computer.	4	MAML
M. He and He (2017)	PdM	Ac	ND	Deep Learning for bearing fault diagnosis.	4	DL, CNN
Z. Li et al. (2017)	ID, PdM	McC	ND	Fault diagnosis and prognosis using data mining to formulate a systematic approach.	4	ANN
S. C. Li et al. (2017)	HC	N	ND	Data mining to detect network intrusions in a Industry 4.0 context.	4	DT, NN, ZR
Park et al. (2017)	IP	Pr	SP	Predictive Manufacturing situation awareness system for enhancing competitiveness in manufacturing.	6	MEBNR
Peralta et al. (2017)	IP	N	ND	Fog computing-based IoT scheme to predict future data measurements.	4	MLR, RT, BDT, ANN, FC
Spendla et al. (2017)	PdM	Mc	A	Hadoop based knowledge discovery platform focused on predictive maintenance for production systems.	4	NN
Sun and Chen (2017)	PdM	S	ND	Low-cost customized wireless data transmission module to predict the remaining useful life of the machines.	6	LPF
Vazan et al. (2017)	PdM	Pr	ND	Data Mining to obtain knowledge of the future behavior in manufacturing systems.	4	CT, DT, BagT, RF, MARS, SVM, KNN, MR, NN
Wan et al. (2017)	PdM	Mc, S, Ma	ND	Big Data solution for active Preventive Maintenance in manufacturing environments.	7	NN
J. Yan et al. (2017)	PdM	Pr	ND	Predict remaining life of a key component of a machining equipment.	5	ANN
Zhou and Yu (2017)	ID	S	ND	FNN and KNN to resolve the incorrect or biased analysis of sensor data.	4	FNN, KNN

Apiletti et al. (2018)	PdM	Mc	ND	Integrated Self-Tuning Engine for Predictive Maintenance, based on big data.	4	RF, LinR, SVM, GBM
Charest et al. (2018)	IP	S	PI	Artificial Intelligence to improve the injection molding process performance.	5	DT, BosT, RF, NB, KNN, FaNN, ANN
Y.-J. Chen and Chien (2018)	IP	Pr	SC	Diffusion model and adjustment mechanism to incorporate domain insights.	2	SRTM
Cisotto and Herzallah (2018)	PdM	GT	Na	Used NNs in a system that support the maintenance function in the decision-making process.	4	NN
Dwaraka and Arunachalam (2018)	ID	Ac	ND	Acoustic emission signals to characterize the spark activity in the Electrical Discharge Machining process.	5	MLR
C. Lin et al. (2018)	PdM	SimS	ND	Learning method with multiple classifier types and diversity for condition-based maintenance in manufacturing industries.	4	MR
Maggipinto et al. (2018)	IP	S	SC	Deep Learning using Computer Vision to model data that have both spacial and time evolution.	4	CNN
Mulrennan et al. (2018)	IP	S	PI	Hybrid Principal Component Analysis RF (PCA-RF) soft sensor model for the inline prediction of tensile properties of polylactide (PLA).	5	DT, Bag, OBE, PCA, RF
Nuzzi et al. (2018)	IP	GI	Rb	Smart hand gesture recognition for collaborative robots with R-CNN object detector to find the hands position.	4	DL

Kiangala and Wang (2018)	PdM	S	PI	Predictive and scheduling maintenance based on the data gathered by the sensors in the conveyors.	5	RB
Kumar et al. (2018)	PdM	S	ND	Health state estimation to facilitate autonomous diagnostics and prognostics models.	5	LinR, PQR
Rendall et al. (2018)	IP	PI	Ch	DNNs in images to predict the pellet shape.	4	ISCS, PLS, DA, RF, DL
Sala et al. (2018)	IP	S	St	Predict the endpoint temperature and chemical concentration of phosphorus, manganese, sulfur and carbon at the basic oxygen furnace.	5	RR, RF, GBRT
Sezer et al. (2018)	IP	S	ND	CPS to predict the parts rejection based on a quality threshold.	6	RP, RT
Straus et al. (2018)	HC, PdM	S	A	Low-Cost Sensors to enable predictive maintenance in old production machines.	6	IF, LOF, KM, MS, SAC, DBSCAN, DT, RF, SVM, LinR, KNN, NB, QDA, NN
Subramaniyan et al. (2018)	IP	Pr	A	Predict throughput bottlenecks in the production line for the future production run.	5	TSF
Susto et al. (2018)	PdM	Mc	SC	Adaptive parameter identification to verify the best trade-off between promptness and low noise sensitivity.	4	SVM, HGP-PF
Tiwari et al. (2018)	IP	Mc	ND	Explored the opportunities in the area of tool wear prediction.	5	KF
Tsai and Chang (2018)	IP	SM	C	Deep Learning application based for coil leveling system.	6	DL

Wu et al. (2018)	ID, PdM	TS	Pc	Visual analytics system to reach automated analytical approaches, and generating results for real-world applications.	4	GM
H. Yan et al. (2018)	PdM	S	ND	Device electrocardiogram and an deep denoising auto-encoder algorithm to predict the remaining useful life of the equipment.	5	DL
Antomarioni et al. (2019)	PdM	Pr	O	Predict component breakages and determine the optimal set of components to repair.	5	AsR
Akhtari et al. (2019)	IP	S	Tr	DNN to detect and classify the load on a power-train.	5	DL
Aydemir and Paynabar (2019)	PdM	I	ND	Deep Learning methods for estimating time-to-failure of and industrial system using its degradation image.	5	DL
Bousdekis et al. (2019)	PdM	S	St	Predictive Maintenance architecture according to RAMI 4.0.	2	NN, DL, HM
Bose et al. (2019)	PdM	Mc	Ae	Anomaly Detection based Power Saving (ADEPOS) scheme using Extreme Learning Machine Boundary through the lifetime of the machine.	4	ELMB
Bruneo and De Vita (2019)	PdM	Mc	Ae	Deep Learning to analyze the history of a system to predict the Remaining Useful Life.	4	DL
Candanedo et al. (2019)	PdM	S	Tr	Predict failures in Air Pressure System in trucks.	4	KNN, NB
Hesser and Markert (2019)	IP	S	ND	ANN to monitor the tool wear in retrofitted CNC milling machines.	5	ANN

W. J. Lee et al. (2019)	PdM	Pr	ND	Predictive Maintenance to monitor two machine tool system elements, the cutting tool, and the spindle motor.	5	SVM, DL
Liulys (2019)	PdM	S	El	Open-source software to develop predictive maintenance applications with basic programming knowledge.	3	GBM, NN
Massaro, Manfredonia, Galiano, and Xhahysa (2019)	IP	Pr	Fu	ANN to predict the product defects in a kitchen manufacturing Industry.	5	ANN
Massaro, Manfredonia, Galiano, Pellicani, et al. (2019)	IP	S	Fo	Predict the humidity during the pasta production.	5	ANN
Martinek and Krammer (2019)	IP	I	El	Machine Learning based prediction methods to optimize the process parameters of pin-in-paste.	5	ANN, ANFIS, GBDT
Packianather et al. (2019)	PdM	ChL	Hc	Three phase methodology to automate quality control in healthcare clinical laboratory.	5	KNN
Pinto and Cerquitelli (2019)	PdM	S	Rb	Predict the fault detection and remaining life estimation of robots.	5	SA, ERT, KNN, CNN
Plehiers et al. (2019)	IP	S	Ch	Framework for chemical production in process-steam cracking to optimize the process control.	4	ANN

Proto et al. (2019)	IP	S	Ch	PREdictive Maintenance service for Industrial procesSES (PREMISES) to predict alarms in slowly-degrading multi-cycle industrial process.	6	GBTC, RF
Rogier and Mohamudally (2019)	IP	SolP	En	NN to predict the conversion of solar energy by a photovoltaic unit.	5	NN
Rosli et al. (2019)	PdM	S	SC	Preventive maintenance for air booster compressor motor failure.	4	ANN, PSO
Rousopoulou et al. (2019)	PdM	Pr	Hc	Predictive analytics for industrial ovens in the healthcare industry.	5	SVM
Sellami et al. (2019)	PdM	Mc	SC	Predict machine failures and presented an algorithm for frequent chronicles extraction.	4	Clasp-CPM
Soto et al. (2019)	PdM	S	ND	IoT Machine Learning and orchestration to failure detection of surface mount devices during production.	4	NN, RF, GB
Naskos et al. (2019)	PdM	Mc	O	Predictive Maintenance with applied unsupervised Machine Learning techniques to detect early oil leaks.	5	MCCOD
Zenisek et al. (2019)	PdM	S	ND	Machine Learning algorithms to detect changing behavior to enhance the maintenance on a microscopic level.	3	RF, SVM, GPBSR
T. Zhang et al. (2019)	IP	Pr	El	Random-SVM (R-SVM) to predict the quality of the TFT-LCD liquid.	4	RSVM
Alasali et al. (2020)	IP	Mc	Tr	Predict the stochastic loads to improve the performance of a low voltage network.	6	MPC, SMPC
Calabrese et al. (2020)	PdM	Mc	Fu	Machine Learning to predict the health status of a woodworking industrial machine.	6	GB, RF, EGB
Q. Cao et al. (2020)	PdM	Pr	ND	Rule-based refinement approach for detect and predict anomalies.	4	RB

Essien and Giannetti (2020)	IP	Mc	ND	Deep Learning model for univariate, multi-step machine speed forecasting in a manufacturing process.	4	DL
Kabugo et al. (2020)	IP	Pr	En	Predict syngast heating value and hot flue gas temperature from data obtained from soft sensors.	5	NN
Karakose and Yaman (2020)	PdM	S	Tr	Fuzzy system-based approach for Predictive Maintenance on electric railways.	4	CF
Kim et al. (2020)	IP	Pr	ND	Predict the state of an unseen camera lens module using semi-supervised regression.	5	DL
Ruiz-Sarmiento et al. (2020)	PdM	Pr	SP	Estimate and predict the gradual degradation of production machines.	5	BF
de Sá et al. (2020)	HC	Pr	ND	Metaheuristics to identify data injection attacks by man-in-the-middle.	4	BSOA, GN, NII

Predictive Analytics are the most used techniques in the practical applications obtained for this SLR. For instance, some studies perform classification techniques (Q. Cao et al., 2020; Kiangala & Wang, 2018; S. C. Li et al., 2017; Miškuf & Zolotová, 2016; Sellami et al., 2019), while other used regression techniques (Calabrese et al., 2020; Charest et al., 2018; Peralta et al., 2017; Rousopoulou et al., 2019). Simple Neural Networks (NN) are used in several research works such as (Cisotto & Herzallah, 2018; Kabugo et al., 2020; Miškuf & Zolotová, 2016; Soto et al., 2019; Spendla et al., 2017). Other studies used more advanced Deep Learning (DL) NN (Choi et al., 2017; Essien & Giannetti, 2020; H. Kuo & Faricha, 2016; W. J. Lee et al., 2019; Maggipinto et al., 2018). Furthermore, some of the surveyed Preditive Analytics used optimization techniques (e.g., Genetic Algorithm, Paticle Swarm Optimization) (Rosli et al., 2019; Saldivar, Goh, Li, Chen, et al., 2016; Saldivar, Goh, Li, Yu, et al., 2016), while other works focused on outliers detection and statistical analysis (Albers et al., 2017; Stein et al., 2016).

Prescriptive Analytics The last table of this SLR (Table 10) presents the practical cases that used Prescriptive Analytics. These types of analytics aims to describe what courses of action may be taken in the future to optimize business processes in order to achieve business objectives. Typically, this is

achieved by associating decision alternatives (or choices) with estimated business outcomes. A diverse set of modeling tools can be used to obtain such analytics, namely optimization and simulation, design experimentation and scenario scheduling (Banerjee et al., 2013; Jugulum, 2016).

The majority of the surveyed studies used optimization techniques. In particular, the most explored method was the Genetic Algorithm (Khayyam et al., 2019; D. Silva et al., 2020). Other authors (Ansari et al., 2019; Brik et al., 2019; Fu et al., 2018; H. Li, 2016; Qu et al., 2016; Tsourma et al., 2018; Uriarte et al., 2018), employed other optimization techniques, such as (Tsourma et al., 2018) that proposed a Task Distribution Engine to automate and optimize the task scheduling and resources assignment procedure in industrial environments. We also found studies that performed Prescriptive Analytics by using predictive models to directly perform actions: DL (Richter et al., 2017); Regression Trees and Nearest Neighbors (Romeo et al., 2018); and SVM combined with Q-Learning (Qu et al., 2016).

Table 10: Overview of the Practical Articles that used Prescriptive Analytics Techniques

Reference	**Func.**[9]	**Data**[10]	**Sector**[11]	**Goal**	**Impact**	**Modeling**[12]
H. Li (2016)	IP	Co	ND	Classification algorithm and Q-learning algorithm to reduce the electricity consumption in an automation system.	4	SVM, QL
Qu et al. (2016)	IP	Pr	ND	Synchronized, station-based flow shop with multi-skill workforce and multiple types of machines.	3	RL, MARL, Op
Klement and Silva (2017)	IP	Pr	Pl	Hybrid approach with List Algorithm and Metaheuristic to optimize planning, assignment, scheduling and lot sizing.	3	LA, SA
Richter et al. (2017)	IP	Mc	El	Optimization techniques for the manufacturers and users of AOI machines.	2	DL
Bányai et al. (2018)	HC	Ge	Tr	Black Hole Optimization for first mile and last mile supply.	6	BHO

[9]Hardware Connection (HC), Information Discovery (ID), Intelligent Production (IP), Predictive Maintenance (PdM)

[10]Conveyor (Co), Geospatial (Ge), Industrial (In), Machine (Mc), Network (N), Production (Pr), Sensor (S)

[11]Automotive (A), Chemical (Ch), Electronic (El), Lean (Le), Mechanical (MC), Not Disclosed (ND), Polymer (Pl), Transportation (Tr)

[12]Artificial Neural Networks (ANN), Black Hole Optimization (BHO), Constrained Optimization (CO), Coyote Optimization Algorithm (COA), Crow Search Algorithm (CSA), Decision Trees (DT), Deep Learning (DL), Fireworks Algorithm (FA), Fog Computing (FC), Genetic Algorithm (GA), Global Cheapest Arc (GCA), Grey Wolf Optimizer (GWO), Guided Local Search (GLS), Iterative Local Search (ILS), K-Nearest Neighbor (KNN), List Algorithm (LA), Memetic Algorithm (MmA), Mixed Integer Linear Programming Model (MILPM), Multi-Agent Reinforcement Learning (MARL), Multiple-layer perceptron neural network (MLPNN), Multi-Objective Optimization (MOO), Neighborhood Component Feature Selection (NCFS), Optimization (Op), Particle Swarm Optimization (PSO), Path Cheapest Arc Savings (PCAS), Prescriptive Maintenance Model (PriMa), Q-Learning (QL), Random Forest (RF), Regression Trees (RT), Reinforcement Learning (RL), Self Organizing Migrating Algorithm (SOMA), Simplified Swarn Optimization (SSO), Simulated Annealing (SA), Simulated Annealing Tabu Search (SATS), Simulation-based Multi-Objective Optimization (SBMOO), Support Vector Machines (SVM), Tabu Search (TbS), Variable Neighborhood Descent Based (VNDB), Variable Neighborhood Search (VNS), Whale Optimization Algorithm (WOA)

Fu et al. (2018)	IP	In	ND	Two-objective stochastic flow-shop deteriorating and learning scheduling problem for advanced intelligent machines.	4	MOO, FA
Romeo et al. (2018)	IP	Mc	EI	Design Support System (DesSS) for the prediction and estimation of machine specification data.	4	DT, RT, KNN, NCFS
Tsourma et al. (2018)	IP	In	ND	Task Distribution Engine to automate and optimize the task scheduling and resources assignment procedure in industrial environments.	5	CO
Uriarte et al. (2018)	IP	Le	ND	Simulation and optimization to improve the lean efficiency, speeding up system improvements and reconfiguration.	2	SBMOO
Ansari et al. (2019)	IP	Mc	MD	Prescriptive Maintenance model for production CPS.	6	PriMa
Brik et al. (2019)	IP	In	ND	Fog computing architecture to deal with system disruption monitoring.	4	FC
Khayyam et al. (2019)	IP	Pr	PI	Genetic Algorithm to predict the stabilization process of a Plyacrylonitrile fiber structure.	5	GA
Leite et al. (2019)	IP	Pr	PI	Optimize the integrated planning and scheduling using Metaheuristic approach.	4	VNDB
Liang et al. (2019)	PdM	S	ND	Memetic Algorithm and Variable Neighborhood Search to improve Predictive Maintenance.	4	MmA, VNS
Negri et al. (2019)	IP	S	ND	Metaheuristics with Digital Twin for scheduling optimizations based on the equipment health predictions.	6	GA
Pane et al. (2019)	IP	Mc	MC	Two reinforcement learning based compensation methods for robot manipulators.	5	RL
Pierezan et al. (2019)	IP	S	En	Coyote Optimization Algorithm to optimize a heavy duty gas turbine used in power generation.	6	COA

Senkerik et al. (2019)	IP	S	Ch	Ensemble of strategies and Metaheuristic for optimization of waste processing batch reactor geometry and control	4	SOMA
Yeh et al. (2019)	IP	N	ND	Optimization techniques to find the cost minimization deployment of a smart factory.	4	SSO
Abdelmaguid (2020)	IP	Pr	ND	Algorithm to obtain optimal solutions for Dynamic Open Shop Scheduling Problem.	4	MILPM
Abdirad et al. (2020)	HC	Ge	A	Two-stage metaheuristic to solve dynamic vehicle routing problem.	4	PCAS, GCA, GLS, SATS
Abdous et al. (2020)	IP	S	A	Design semi-automated assembly lines using Machine Learning and Optimization techniques.	4	ILS
Kharwar et al. (2020)	IP	S	Pl	Particle Swarn Optimization to optimize milling parameters (weight, spindle speed, feed rate and depth of cut).	4	PSO
Y. Li et al. (2020)	IP	Pr	Pl	Hybrid model using Optimization and Machine Learning for production rescheduling.	4	GA, TbS, RF, SVM, MLPNN
Milošević et al., 2020	IP	Pr	ND	Compared three optimization algorithms for intelligent process planning optimization.	4	GWO, WOA, CSA
Rahman et al. (2020)	IP	Pr	ND	PSO for line balancing and automated guided vehicles scheduling for smart assembly systems.	4	PSO
D. Silva et al. (2020)	IP	Pr	ND	Hybrid ANN model and use GA for the multi-objective strength optimization of concrete with fiber.	5	ANN, GA

2.1.5 Discussion

Figure 6 presents the Literature Map resulted from this SLR. This Literature Map contains three different levels of interactions, where the first level is the Analytics Level and the second level contains the components of the different Analytics application levels (Data Visualization, Detect Production Anomalies,

Improve Product Quality, Detect Costumers Needs, Predictive Maintenance and Resources Optimization). The last level presents the different techniques used for each component, as well as some studies that use these techniques. To simplify the visualization, the map only details business analytics techniques that were used in two or more practical cases.

It is clear in Figure 6 that Supervised Learning techniques (Classification and Regression algorithms) are a popular approach of Business Analytics in Industry 4.0, being adopted in all the application types identified in this SLR. Statistical Data Analysis is a technique used mainly for Data Visualization, but it was also used for Predictive Maintenance (Mozgova et al., 2018), to Detect Anomalies in Production (Zheng & Wu, 2017) and to Improve Product Quality (Kirchen et al., 2017). Clustering is a more advanced technique compared to Statistical Data analysis, and is used to find Production Anomalities (Y. Wang et al., 2017), to improve the products quality (T. Lin et al., 2016), to detect costumers needs (Saldivar, Goh, Li, Yu, et al., 2016), and for predictive maintenance (Candanedo et al., 2019). Reinforcement Learning was used mostly for Resources Optimization (Pane et al., 2019; Qu et al., 2016), while Optimization techniques were used for Resources Optimization (Uriarte et al., 2018), to Detect Production Anomalies (Trunzer et al., 2017) and to Improve Product Quality (Khayyam et al., 2019).

Regarding the Supervised Learning techniques, based on Classification and Regression algorithms, it is important to mention the popularity of NN (in their Artificial Neural Networks (ANN), DL, or Convolutional Neural Networks (CNN) forms), in the different Industry 4.0 areas. In effect, the use of NN reaches every area of application studied in this SLR with a total of 38 practical applications retrieved in this study. Moreover, the use of NN is growing over the time, with 4 applications in 2016, 10 in 2017, 7 in 2018, 14 in 2019, and 3 applications in the first months of 2020.

The Literature Map from Figure 6 provides a general overview of the different application areas of Business Analytics in Industry 4.0, where it is clear that the areas of Improve Product Quality, Anomalies Detection and Predictive Maintenance are the most popular. While Business Analytics techniques can also be employed to optimize resources in the Industry or to Detect Costumers Needs, a small number of research application studies have addressed these topics, with 9 applications focused on Resources Optimization and 4 applications in Detect Costumer Needs.

2.1.6 Conclusions and research implications

This section presents the results of this SLR to analyze the evolution and the application of Business Analytics techniques in the Industry 4.0 context. As stated in Section 2.1.1, the Research Question targeted by this SLR research is: ***How and in what areas of the industry are Business Analytics techniques being used in an Industry 4.0 context?*** The papers were surveyed by performing an initial keywords query on scientific search engines. Then, the retrieved papers were manually inspected by performing a careful analysis, to assure that the most relevant studies for this SLR were selected. Next, we have analyzed the selected papers in terms of both quantitative and qualitative elements. The quantitative analysis

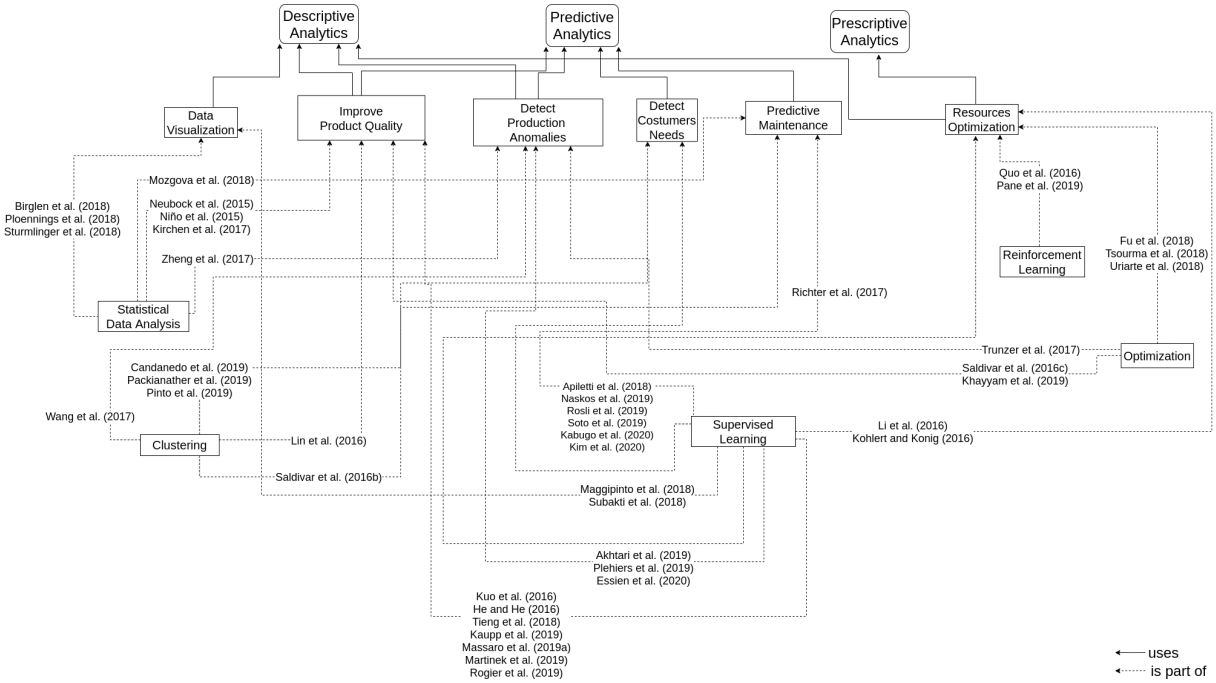


Figure 6: Literature Map.

showed that the most published type of paper is the Practical Application. As for the quantitative analysis, it consisted in a characterization of the Descriptive, Predictive and Prescriptive analytics in terms of what types of applications are implemented in the Industry, what are the techniques used in the practical applications and the impact of the results achieved. Considering the presented SLR we highlight that:

- The application of Business Analytics techniques within the Industry 4.0 concept has grown in recent years and its popularity is still rising (as shown in Figure 4 and Table 6). Thus, there is a research opportunity for publishing more papers regarding Business Analytics applied to the Industry 4.0.

- Manufacturing is the industry sector with the most practical applications (Table 5). One contributing factor for this phenomenon is that there has been a financial support for the adoption of innovative manufacturing techniques (European Commission, 2013). Nevertheless, there is a research opportunity set in terms of addressing other industry sectors, such as Transportation or Construction.

- Within the manufacturing sector, most of the practical applications focused on problems existing in production lines, with different goals, such as detecting faults in production components, defective products, until monitoring the production process and optimization of productive components such as energy consumption and resources allocation.

- Regarding the type of analytics, Descriptive Analytics involved a total of 23 practical applications, Predictive Analytics with 80 application studies and Prescriptive Analytics included 24 research

works. The popularity of Predictive Analytics is being linked with the growing interest in the fields of Machine Learning and Data Science in the decade of 2010 (C. Costa & Santos, 2017).

- Regarding the modeling techniques used, Supervised Learning was the most used approach, with NN being used in 39 applications, RF in 10 applications, SVM in 6 applications, Decision Tree in 5 applications and Rule-Based in 2 applications. Classical Statistical Data Analysis was used in 11 applications, Clustering was addressed in 6 applications, the same number as Optimization techniques, and Reinforcement Learning was employed in 2 applications. Given the current success of the DL field (Goodfellow et al., 2016), it is expected that the number of Industry 4.0 research works that use NN will further increase in the future.

- Practical applications that use Descriptive Analytics are focused on analyzing the data obtained in order to find answers for diverse problems, such as verifying the tool wear through the time or what is the most common cause that leads to the equipment failure.

- The practical applications that used Predictive Analytics were more focused in Predictive Maintenance, such as predict when the equipment will fail, or verify if the equipment is not corresponding in terms of its typical performance.

- Practical applications that use Prescriptive Analytics target more on resources optimization, such as optimize the energy consumption or optimize the resources scheduling. However, the SLR results reveal that there is still a scarce number of research studies that use Prescriptive Analytics techniques within the Industry 4.0. Therefore, there is a huge potential for future research on more Prescriptive Analytics studies since there is a large number of industrial needs that are related with resource optimization and schedulingMoreover, as pointed out by Davenport (2013), these are the analytics "that tell you what to do" and thus hold a higher business value by providing an actionable knowledge for the industry. Thus, in future works, we believe there will be an increase of Prescriptive Analytics applications for the Industry 4.0.

This SLR reviewed research papers published in the last decade (from 2010 to 2020). In the next decade, it is expected that Business Analytics will be more prevalent in the Industry, due to further advances in AI and ML. In particular, as the European Commission plans a future investment of 7.5 Billion EUR in the areas of Advanced Computing and AI (Commission, 2020), several of these funds will be devoted to Industry applications, which surely will be reflected in an increased number of research papers.

2.2 Other Relevant Concepts

In this section we present theoretical concepts that were not addressed in the SLR, but are relevant for this PhD work, as well as a survey of the state-of-the-art related to the aplication of Business Analytics in the Chemical domain and its AL.

2.2.1 Machine Learning (ML)

ML is a branch of AI that enables computer algorithms to learn from experience without explicitly being programmed (Breiman, 2001; Obermeyer & Emanuel, 2016). ML uses computers with the objective of simulating human learning and allows the machines to identify and acquire knowledge from the real world, and improve the performance of tasks with the knowledge obtained (Portugal et al., 2018). Mitchell (1997) defined ML as "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". There are four main types of methods in ML: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning (Portugal et al., 2018).

This PhD thesis is mostly focused on Supervised Learning, where the learning algorithms have labeled training data, meaning that each input example contains a target output. The task is then to learn an implicit mapping function that exists in the training data. Such function should be able to generalize to new situations (Portugal et al., 2018). Therefore, supervised learning plays a key role in predictive analytics. The two main supervised learning types are Classification and Regression (Zhu et al., 2003). Common Classification methods include Decision Trees (DT), K-Nearest Neighbor (KNN), SVM, NN and RF (Larose, 2004). A few examples of pure Regression methods are LinR, Lasso and Elastic nets (Ogutu et al., 2012). SVM, RF and NN can also be applied to regression (Rodriguez-Galiano et al., 2015; Steyerberg et al., 2014). Recently, there has been an increasing interest in the adoption of DL NN, which can be applied to both classification and regression, and that achieved competitive results in several ML challenges (e.g., object recognition from images) (Goodfellow et al., 2016).

Unsupervised learning does not have a target variable defined, it is the algorithm that searches for patterns and structures the information among data variables. One of the most used unsupervised techniques is clustering. A clustering algorithm automatically groups data variables or examples according to a distance function and clustering metric. Another popular unsupervised learning approach is rule association mining (Larose, 2004). These algorithms can be used to obtain a descriptive knowledge.

The semi-supervised approach falls between the supervised and unsupervised approaches. It often assumes that there some labeled data and also unlabeled data. The semi-supervised approach is useful when labeling data is costly to obtain, such as requiring a manual effort (Zhu et al., 2003). Active Learning (e.g., co-training) is an example of a semi-supervised algorithms.

Reinforcement algorithms work by providing rewards or penalties to the result of the algorithm's suggested actions. The algorithms can learn something given by an external feedback from the environment or a thinking entity, in a continuously trial-and-error way. A commonly used reinforcement learning algorithm is Q-Learning. (Portugal et al., 2018; Sutton & Barto, 1998).

2.2.2 Auto Machine Learning (AutoML)

One of the key tasks of ML is identifying a model to use for a particular dataset, the attributes to consider and defining the right choice of its hyperparameters (Feurer, Springenberg, et al., 2015). When performed

manually, the proper ML model selection, often becomes a time-consuming process. Moreover, it is executed by assuming ad-hoc methods (e.g., heuristics).

For non ML experts, it is not trivial to setup a ML model. Thus, these users often adopt default pre-selected methods, as provided by a computation tool. For practitioners, the modeling technique often relies on the use of handmade heuristics or the use of domain experts to exploit the often-large hypothesis space and the trade-off between the various features of models, such as size, speed, and constraints (Y. He et al., 2018). With the increasing number of non-specialists working with ML (Thornton et al., 2013), it is important to enable people with more limited ML knowledge to easily choose and apply-templates.

In response to this problem, the concept of AutoML emerges. According to the definition of Feurer et al. (2019), AutoML addresses the fundamental problems of ML, which are the choice of algorithm to use in a given dataset, whether or not to perform its attributes processing and how to establish all hyperparameters, formalizing AutoML as a Combined Algorithm Selection and Hyperparameter Optimization problem, as defined by (Thornton et al., 2013). Guyon et al. (2015), as part of the ChaLeam AutoML Challenge argued that, AutoML is associated with all aspects of progressive automation of all ML phases (beyond those already available and the choice of a model and hyperparameter optimization), where it includes automation of:

- First Phase - Data loading and formatting;
- Second Phase - Detection and processing of skewed or missing data;
- Third Phase - Selection of learning representation and feature extraction;
- Fourth Phase - Matching problem/algorithm;
- Fifth Phase - Obtaining new data (active learning);
- Sixth Phase - Creation of sized and sized training, validation and testing sets;
- Seventh Phase - Selection of algorithms that meet resource constraints both in training and test progress;
- Eighth Phase - Ability to generate and reuse workflows;
- Ninth Phase - Meta-learning and transfer of learning; and
- Tenth Phase - Explanatory reports.

Other authors have different definitions for AutoML, such as Jin et al. (2019) that defines AutoML as a tool that allows non-ML experts to use deep learning techniques easily.

2.2.3 Intelligent Decision Support Systems (IDSS)

Decision Support Systems (DSS) is a relevant subfield of the IS discipline that aims to assist managerial decisions by using IT (Arnott & Pervan, 2014). Alter (1980) defined that a DSS has three major characteristics. The first characteristic is that DSS must be designed specifically to facilitate the decision process. The second one is that DSS, instead of giving an automate answer, must provide support to the decision process. Finally, DSS must have the ability to give rapid responses in order to adapt the changing needs of the decision makers. Simon's well-known theory of decision making is one of the most accepted models; contains 4 phases (assuming the extended version proposed by Sprague): Intelligence Phase, Design Phase, Choice Phase and Monitoring Phase (Campitelli & Gobet, 2010). It is noteworthy that when adopting a data-oriented DSS, the models obtained through the CRISP-DM methodology can be very useful when applied in one or more phases of Simon's model. This work will follow the CRISP-DM methodology to guide the process of obtaining predictive models.

Over the last years, the DSS topic has evolved, resulting in (slight) distinct decision support approaches, namely (Arnott & Pervan, 2008):

- Personal Decision Support Systems;
- Group Support Systems;
- Negotiation Support Systems;
- Intelligent Decision Support Systems (IDSS);
- Knowledge Management-Based DSS;
- Data Warehousing; and
- Enterprise Reporting and Analysis Systems.

This PhD is particularly focused on Intelligent Decision Support Systems (IDSS), which is a DSS that uses AI (including ML and other methods) to enhance managerial decisions (Gottinger & Weimann, 1992). As any IT, DSS are rapidly changing. After the 2000s, there has been a trend in the usage of data-driven models for DSS (Arnott & Pervan, 2014). Figure 7 shows an interesting evolution of the DSS field over the last decades, from the 1960s to the 2010s. As already mentioned, this PhD focused on the Business Analytics aspect of DSS. To understand the rise of Business Analytics, we must move a few years back, in the late 1990s, when the Data Warehousing and Business Intelligence concepts emerged from the Executive IS concept. It should be noted that most Business Intelligence systems were more focused on accessing historical data (e.g., using data warehousing and dashboards) and lacked true "intelligence"capabilities (Michalewicz et al., 2006). Given the pressure to extract more actionable knowledge, AI techniques (e.g., ML, Modern Optimization) were more applied to perform Business Intelligence tasks, resulting in what Arnott and Pervan (2014) term as Business Analytics.

Figure 7: Arnotts' genealogy about the DSS, adapted from Arnott and Pervan (2014)

In Figure 7 the elements that are highlighted using blue are introduced in this PhD work and do not belong to the original figure proposed by Arnott and Pervan (2014). Indeed, following the growing impact of data on AI systems, we assume the term Modern IDSS to reflect DSS that adopt currently available impacting AI technology (e.g., ML, AutoML). The update is relevant since traditional IDSS, adopted in the mid 1970s and 1980, were mostly expert-driven (e.g., based on explicit knowledge rules that were extracted from the domain experts). In contrast, current IDSS, as targeted in this PhD research, are mostly data-driven, relying on ML and other AI algorithms.

2.2.4 Cross-Industry Standard Process for Data Mining (CRISP-DM)

In this PhD, two methodologies were used to achieve the proposed objectives. For the design of the predictive models, the methodology used was the CRISP-DM. As shown in Figure 8, this methodology describes the Data Mining (DM) process in six phases: business study, data study, data preparation, modeling, evaluation and implementation. The CRISP-DM methodology has several advantages when applied to DM projects, such as: faster speed, lower execution costs, greater security and feasibility (Santos & Azevedo, 2005).

The methodology was developed by Chapman et al. (2000) and it includes the following phases:

- Business Understanding - the first step of the CRISP-DM model. It is at this point that the real business needs are evaluated, that the DM problem is formulated, and that the objectives are defined, including their link to the business goals. Then, this obtained knowledge is converted into a DM problem and a preliminary plan is designed.

- Data Understanding - comprises four main tasks: data collection, data description, data exploitation and quality verification.

- Data Preparation - this phase covers all activities related to the construction of the final set of data, that is, the one that will be used in the modeling tool. It may include the selection of tables, registers and attributes, as well as the transformation and cleaning of the data to be used in the modeling tool.

- Modeling - it is at this stage that the various modeling techniques are selected and there is an adjustment of the parameters, in order to optimize the results. In this selection, it is necessary to consider not only the adequacy of the technique to the DM problem, but also the specific requirements that these techniques have.

- Evaluation - at this stage the usefulness of the models is evaluated. The steps performed to construct the models are verified in order to assess if they meet the business objectives.

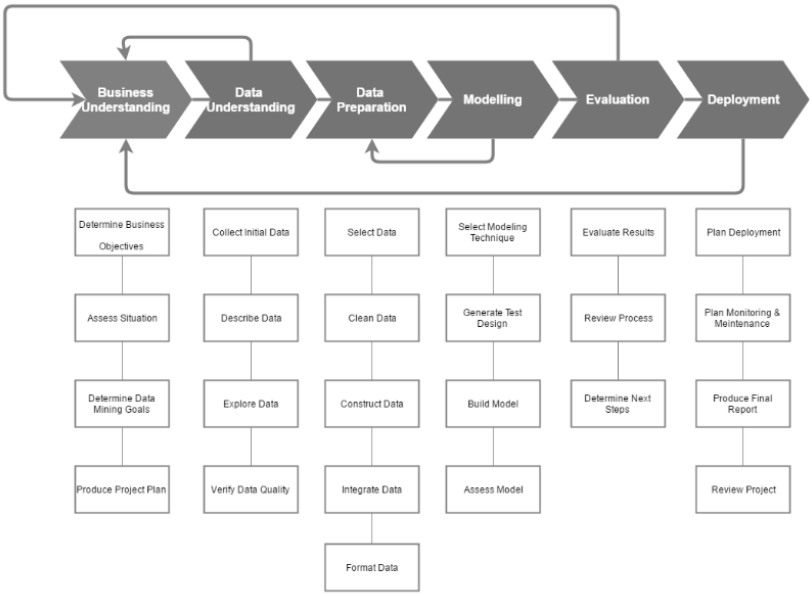


Figure 8: CRISP-DM Phases, adapted from Chapman et al. (2000)

- Deployment - in the last phase, the knowledge acquired from all the previous steps needs to be organized and presented. Monitoring and maintenance is planned, the final report is produced and the project is revised.

2.3 Business Analytics applied to the Chemical Industry

Concerning the sample arrival prediction, and following the Industry 4.0 revolution (Shrouf et al., 2014), many factories now are generating data that can be analyzed by DM and ML techniques in order to support managerial decision-making. Yet, several real-world DM projects tend to fail due to a misalignment between business needs and ML analyses (Deal, 2013). The CRISP-DM is an open standard and robust methodology that was specifically developed to reduce this misalignment and increase the success of DM projects (Wirth & Hipp, 2000). CRISP-DM is a popular methodology. For instance, it has been applied to the Banking (Moro et al., 2011) and Health Care (Caetano et al., 2014) domains. Regarding the analyzed chemical industry, the AL are mostly managed manually, with the usage of IT being more focused on storing the test values rather than the process (Kammergruber et al., 2014; Skobelev et al., 2011).

It should be highlighted that most predictive ML studies in industry are focused on non chemical sectors and target the predictive maintenance task. Examples of ML algorithms that were proposed for such task include: RF (Canizo et al., 2017), NN (Spendla et al., 2017), Gradient Boosting Machines (Liulys, 2019) and SVM (Straus et al., 2018). There are also studies about non maintenance prediction applications, such as: the classification of quality products produced by injection molding processes via

Boosting, RF and NN models (Charest et al., 2018); and estimation of endpoint temperature and chemical concentration of a furnace when producing low-carbon steel using RF and ridge regression algorithms (Sala et al., 2018). All these studies require the selection and configuration of the right ML algorithm, which often depends on the ML expert knowledge and that involves the usage of heuristics or trial-and-error experiments (Gibert et al., 2018).

Regarding the materials consumption predictions, most predictive analytics studies for the chemical sector involve the production processes, rather than AL. For instance, Roe et al. (Wen et al., 2020) used a Fuzzy NN model to perform a predictive control on a solar-thermal chemical processing. Moreover, Longone et al. (Langone et al., 2020) used a LogR to predict production anomalies in a chemical plant that adopted the Industry 4.0 concept. In all these ML predictive studies, expert knowledge and trial-error experiments were used to select and tune the predictive ML algorithms, which is a common ML practice. However, there is a recent ML trend that assumes the usage of AutoML (Ferreira et al., 2020). The main advantage of AutoML is that it alleviates the ML analyst effort, allowing to focus on other aspects of the glsml pipeline process (e.g., data engineering). Moreover, the data is typically spread through different databases what work as information silos (e.g., production, laboratory testing), thus it is difficult to have an easy access to all data under a single version of the truth. By adopting the Industry 4.0 concept, which assumes a better usage of IT, there is a potential gain to optimize the management of the AL.

With respect to the uses of IDSS within the Industry 4.0 concept, there are several studies proposing data-based interactive dashboards. For instance, our survey about the usage of Business Analytics in Industry 4.0 has found several examples of dashboards used to monitoring the production process, as well as verify new insights on the shop floor (Neuböck & Schrefl, 2015; Niño et al., 2015). Moreover, in the automotive industry, data-based dashboards were used to monitor the assembly processes (N. Silva et al., 2021). Also in the manufacturing sector, sensors and IoT data were also integrated into dashboards to monitor the productive process (Mahmoodpour et al., 2018). Concerning the specific chemical industry, we have found one one dashboard example that was proposed to control and monitor the production of a chemical plant (Bellini et al., 2021).

Turning to the incorporation of AI techniques for decision support, there are a few studies that integrate ML results in dashboards. For instance, a few examples are: use NN to improve the energy saving in factories (Kabugo et al., 2020); usage of a RF algorithm and IoT sensors to improve fault diagnosis tasks (Tran et al., 2021); and a predictive maintenance system using a Remaining Useful Life model to estimate the health index of production machines (Chiu et al., 2017). However, regarding the application of IDSS in the AL of chemical industry the research is very scarce. This occurs because the AL are mostly managed manually, where IT is mostly focused on storing the quality values and not the AL processes.

Table 11 summarizes the state-of-the-art works that are more closely related with this PhD thesis. Although there are some cases of application, most of them focus on the chemical processes of some laboratories, for improvement and automation of their chemical processes, with only one case focusing on the improvement of laboratory processes, namely in the prediction of energy consumption in laboratories.

Table 11: Overview of the Practical Articles that used Business Analytics in the Chemical Domain

Reference	**Func.**[13]	**Data**[14]	**Goal**	**Impact**	**Modeling**[15]
Montavon et al. (2013)	IP	L	ANN model that simultaneously predicts multiple electronic ground- and excited-state properties.	5	ANN
Morellos et al. (2016)	IP	L	Used Regression method to predict the soil total nitrogen, organic carbon and moisture.	4	C, LS-SVM, PCR, PLSR
Coley et al. (2018)	IP	L	Used Neural Networks to improve the synthesis planning.	5	NN
Häse et al. (2018)	IP	L	Implemented an Optimization framework for self-driving laboratories.	4	SOOA
Wahab et al. (2020)	ID	H	Artificial Neural Networks to predict energy consumption at the laboratories.	5	ANN
M. Zhong et al. (2020)	IP	L	Integrated Machine Learning Algorithms in a framework to accelerate the discovery of chemical compounds.	4	ND

[13] Hardware Connection (HC), Information Discovery (ID), Intelligent Production (IP), Predictive Maintenance (PdM)

[14] Laboratory (L)

[15] Artificial Neural Networks (ANN), Cubist (C), Least Squares Support Vector Machines (LS-SVM), Neural Networks (NN), Not Disclosed (ND), Principal Component Regression (PCR), Partial Least Squares Regression (PLSR), Single-objective optimization algorithms (SOOA)

Chapter 3

Methods, Experiments and Results

This chapter presents the main methods, experiments and results obtained during this PhD work. The first section presents the framework that was developed and used as a guide to develop our IDSS. The remaining sections introduce the published articles, which are presented following a chronological order (also the same order assumed by the PhD project execution):

- Section 3.2 presents the development of a two-stage ML model to predict the sample arrival at the AL. The associated work was published in the 16th International Conference on Artificial Intelligence Applications and Innovations (A. J. Silva et al., 2020).
- In Section 3.3 the focus is on the prediction of material consumption at the AL. This work was published in the 17th International Conference on Artificial Intelligence Applications and Innovations (A. J. Silva & Cortez, 2021)
- Finally, Section 3.4 presents the instruments allocation module, along with the development of the proposed IDSS that contains several data analytics modules integrated in dashboards. This work was submitted to a scientific conference.

3.1 Adopted Framework

As mentioned earlier, the work at the AL of the analyzed Chemical company is mainly based on the use of physical documentation. Moreover, there are several IT applications and databases that work as silos, with few or none data integration. In this PhD project, we present an IDSS architecture that uses ML algorithms and other data analytics, where the objective is to improve the functioning of the AL as well as the existing workflow between Laboratories, Warehouse and Manufacturing. This IDSS is intended to give a unified view of this workflow, and also help bring new insights to support the work executed at the AL.

To achieve the above goals, we assumed an integrated framework that is illustrated in Figure 9. We take this framework as an instantiation of the DSRM-IS methodology that was followed in this PhD thesis. In addition, we also use the CRISP-DM methodology for the development of the ML models (as shown

in Sections 3.2 and 3.3). The first two components of the adopted framework have a parallel with the first two stages of the CRISP-DM methodology and were essential for the development of the remainder components. In effect, both Business and Data Understanding components were executed when designing the IDSS four main modules, which are:

- **Sample Arrival Prediction** – Based on the Production and Warehouse data, the system will predict the sample arrival at the Laboratories. This applies for the RM, IPC, and FP samples.
- **Materials Consumptions Prediction** – Using the Sample Arrival data (forecasted and historical) in the Laboratories and material requests data, the goal is to predict the Materials Consumptions in the Laboratories in order to guarantee that the Warehouse always have the quantity of the material to be requested.
- **Suggest Instruments Allocation** – Using historical of records the instruments usage, knowledge about the samples that will be arriving at the Laboratories, the respective information regarding the product, sample type and tests to be performed, the goal of this module is to assign the best instrument for the quality analysis.
- **Decision Support Dashboards** – This module joins all the predictions and suggestions created in the previous modules and presents them to the users using friendly Dashboards. This module also generates reports about the activities performed in the laboratory, regarding the sample arrival, material requests and instruments allocations, as well as the historical visualization of samples arrived and tests performed at the laboratory. These Dashboards are useful to find and resolve bottlenecks in the laboratory workflow.

In the next sections of this chapter, the developments related with the last three components of the framework are presented. In Section 3.2 the sample arrival prediction module is detailed. Section 3.3 presents the module for predicting the consumption of materials in the AL. Finaly, Section 3.4 contains the instruments allocation module along with a presentation of the fully designed IDSS.

3.2 Predict Sample Arrival in the Laboratories

In this study, we address a relevant Business Analytics need of a Chemical company, which is adopting a Industry 4.0 transformation. To ensure the quality of the products being manufactured, samples taken from the company production processes need to be tested in Laboratories. The tests assure that the products are compliant with quality standards, allowing their usage by the company clients. Under this context, predicting the arrival of production samples at the laboratory is a key issue, since it helps in the allocation of equipment and human resources. Aiming to solve this task, this study presents a novel two-stage ML prediction system, which was developed during the implementation of a CRISP-DM (Wirth & Hipp, 2000) project that included three iterations, each focusing on a distinct regression strategy. During

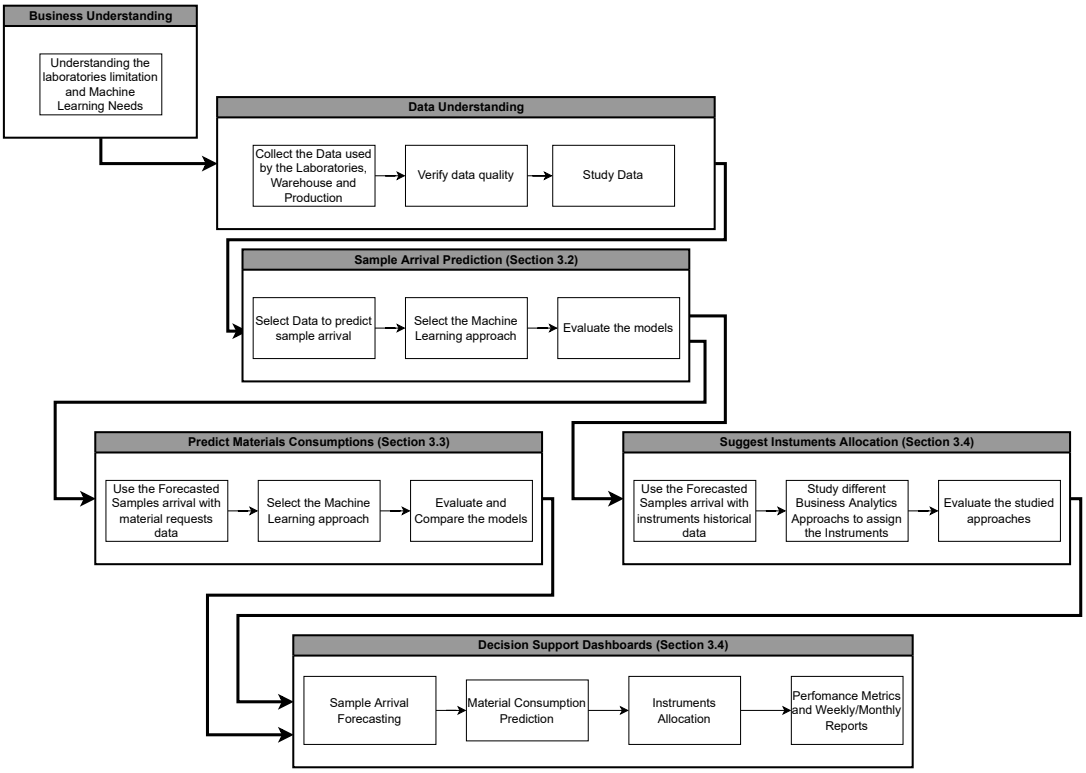


Figure 9: Adopted PhD framework.

the modeling stage of the three CRISP-DM iterations, an AutoML (Feurer, Klein, et al., 2015) procedure was adopted, allowing to compare and configure six state-of-the-art ML algorithms.

3.2.1 Materials and Methods

3.2.1.1 Business Task

This Chemical company produces several products, in batches. During the production-batch execution process, a sequence of samples, called IPC, are selected for quality Laboratory inspection, in order to ensure that the production process is running as expected. In terms of the Chemical Laboratories, the IPC samples have the highest priority, because the production process can not continue without their approval. A fixed amount of IPC samples are selected from each production-batch ($s \in \{1, ..., IPC_{\text{max}}\}$). The production information system registers several attributes related to the IPC sample production, including its initial production time, denoted here as IPC production time PT_s. One by one, the IPC samples arrive at the Laboratory at time LT_s, under irregular intervals that are difficult to be estimated in advance.

The business goal is thus the non-trivial task of predicting of arrival time for each IPC sample at the Chemical Laboratories. Solving this task efficiently allows a better management of the Laboratory equipment and human resources. For instance, some IPC quality tests require a setup time, in which the analysts need to prepare in advance the Laboratory testing instruments. The business goal was

Table 12: Summary of the data attributes.

Input Attributes:		
Name	**Description**	**Range**
day	day of the week when the production-batch started	{1,...,7}
month	month when the production-batch started	{1,...,12}
product	product type (nominal code)	155 levels
version	version of the product (numeric)	{1,...,108}
grade	product grade (nominal, related with the lab tests)	15 levels
stage	product stage (nominal, related with the lab tests)	1,272 levels
batch	batch identification of the product (nominal)	925 levels
s	sequence number of the sample ($s \in \{1, ..., IPC_{\text{max}}\}$)	{1,...,169}
Output Targets:		
Name	**Description**	**Range**
y_1	time lag arrival of two consecutive samples	[0.2,5315.3]
y_2	time lag between PT_s and LT_s	[0.0,3270.0]

addressed as a regression task, under two main target goals. In the first CRISP-DM iteration, we only used Laboratory temporal data and the target goal was defined as predict $y_1 = LT_{s+1} - LTs$, which corresponds to the time lag between the next IPC sample arrival (LT_{s+1}) and the current (already known) Laboratory sample arrival (LT_s). In the second and third CRISP-DM iterations, we explored production temporal data, predicting $y_2 = LT_s - PT_s$, where the Laboratory arrival time can be immediately estimated once the IPC sample starts its production.

3.2.1.2 Data Understanding and Preparation

We used an ETL procedure to merge the relevant data from two main databases related with the production and Laboratory testing information systems, populating an integrated and business oriented data warehouse system. The ETL resulted in a raw file with 226,929 rows and 33 columns regarding all Laboratory samples that were analyzed during a three-year time period. The data warehouse was further filtered in order to contain rows related with IPC samples and with complete values in terms of the input and output attributes (Table 12), leading to a dataset with 26,611 instances. The input variables were manually selected and defined from the filtered raw file using expert domain knowledge, obtained by interacting with the chemistry experts. Due the complexity of the Chemical factory processes and information system integration issues, it was not possible to have access to a more richer set of data features (e.g., which components and machines were used to produce the samples). Thus, the resulting set of 8 inputs is rather small, which makes more challenging the prediction task. Both output targets were computed using a particular time unit, which is not disclosed here due to business privacy issues.

3.2.1.3 Machine Learning Models

In terms of computational environment, we adopted the R tool and its `rminer` package (Cortez, 2014) for data manipulation and ML result evaluation, while the AutoML adopts the H2O implementation (Cook, 2016). The AutoML procedure was configured to select the regression model and its hyperparameters based on the best RMSE computed using a validation set that is obtained by applying an internal 10-fold cross-validation method over the training data. All computational experiments were executed on the same personal computer and each individual ML model was trained up to a maximum running time of 3,600 seconds. Once a ML model is selected, the model was retrained with all training data. As in (Ferreira et al., 2020), the AutoML was configured to include a total of 6 distinct regression algorithms: RF, Extremely Randomized Trees (XRT), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), XGBoost (XGB) and a Stacked Ensemble (SE). The RF is a popular ensemble method that combines a large number of decision trees based on bagging and random selection of input features (Hastie et al., 2009). The XRT algorithm extends the RF approach by randomly selecting the decision thresholds of the tree nodes (Geurts et al., 2006). GLM estimates regression models for exponential distributions (e.g., Gaussian, Poisson, gamma) (Hastie et al., 2009). The GBM algorithm is a based on a generalization of tree boosting, sequentially building regression trees for all data features (Hastie et al., 2009). XGB is another ensemble tree method that uses boosting to enhance the prediction results (T. Chen & Guestrin, 2016). The SE method, also known as stacked regression (Breiman, 1996), combines the predictions of different base learners by using a second-level ML algorithm. The H2O implementation (Cook, 2016) uses the following AutoML setup: RF and XRT – set with the default hyperparameters; GLM - grid search used to set one hyperparameter (*alpha*, a regularization parameter); GBM and XGB – grid search used to tune nine and ten hyperparameters (e.g., number of trees, maximum depth, minimum rows); SE – all five algorithms (RF, XRT, GLM, GBM, XGB) are used as base learners and the individual predictions are weighted by using a second-level GLM learner. For the ML algorithms that require numeric inputs (e.g., GLM), the nominal inputs (e.g., product, grade) are previously transformed by using the standard one-hot encoding, which assigns one boolean input per categorical level. For instance, a categorical feature with three levels ($\{a,b,c\}$) is encoded as: a =(1,0,0), b =(0,1,0) and c =(0,0,1).

A total of three CRISP-DM iterations were executed, aiming to improve the regression results and the potential value of the ML models. The first CRISP-DM iteration targeted the y_1 output, while the second and third CRISP-DM iterations approached y_2, under two variants. The y_1 target is assumes that at least one IPC sample from the production-batch as arrived at the Laboratory. The trained ML model can be used each time new sample arrives, allowing to estimate when the next sample will be delivered ($\widehat{y}_1$). A different perspective is adopted by the y_2 target, since the fitted ML model can be applied to predict the Laboratory sample arrival once an IPC sample production has started. The model employed in the second CRISP-DM iteration uses a simple regression with a single ML model ($\widehat{y}_2$). During the evaluation stage of the second CRISP-DM iteration, we identified that there were some high prediction errors, in particular when predicting the arrival times for the first sample of the production-batch ($s = 1$). In order to check

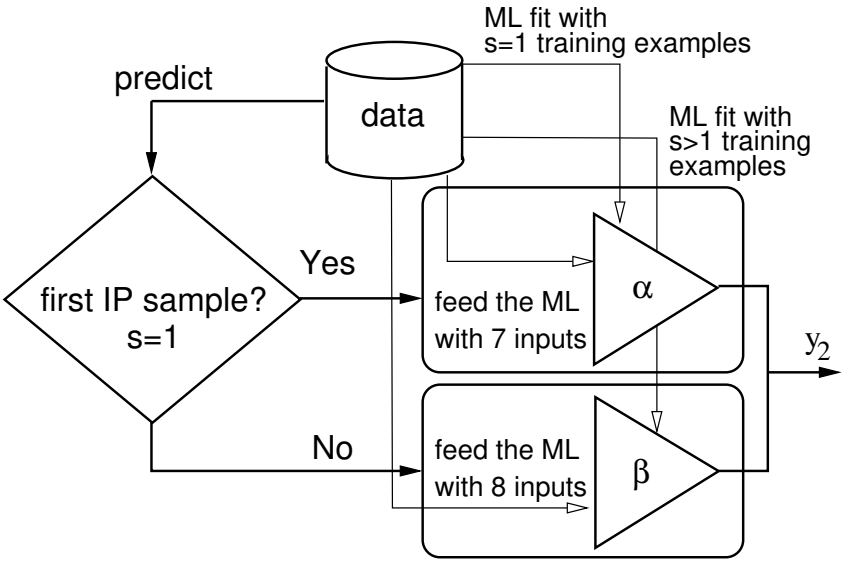


Figure 10: Schematic of the proposed two-stage ML prediction model ($\widehat{y}_{2\alpha\beta}$).

if we could improve these results, a third CRISP-DM iteration was executed, in which we specialize two distinct ML models (α and β). The first ML model (α) is trained using only the first product-batch sample examples ($s = 1$) and thus the fitted model includes only seven input attributes ({day, month, product, version, grade, stage,batch}). The second model (β) is only activated when producing the other product-batch IPC samples ($s > 1$). Similarly to the second CRISP-DM iteration model, this ML model is trained with all eight inputs (including s, the sample sequence number). The proposed two-stage model ($\widehat{y}_{2\alpha\beta}$) is shown in Figure 10.

3.2.1.4 Evaluation

The collected data was divided into three main sets, by using a chronological order. The last 20 weeks of data (total of 5,110 examples) was kept out of the initial ML experiments. The goal is apply this additional unseen data in a more realistic evaluation, provided by a RW validation (Tashman, 2000) that is executed for the best ML regression approach. The remaining and oldest 21,501 examples (not used as test set by the RW) were further divided into training and test sets (holdout split) (Schorfheide & Wolpin, 2012). The time ordered Holdout Split (HS) was used to compare the three distinct main regression approaches (from the CRISP-DM iterations). The training data included the oldest 15,050 examples (around 70%). As for the HS test set, it included 6,451 instances.

Regarding the RW, it was set using a fixed training window with six months of data and a weekly testing of the ML models, in a total of 20 iterations. In the first iteration, at the first Sunday, the ML was trained with the last six months of historical data. Then, the model was used to perform sample arrival predictions for the incoming week (fixed test size of seven days). In the second iteration, executed at the second Sunday, the training window was updated by discarding one week of the oldest data and adding

the previous week examples, allowing to update (retrain) the ML model, which then predicted the next week sample arrival times, and so on.

In this work, we adopt two popular regression error measures: RMSE and MAE. We also use the Acc@T metric, which is more easily understood by the business analysts, since it measures the percentage of examples accurately predicted when assuming an absolute error tolerance of T. A quality regression model should provide low RMSE and MAE values and also a high accuracy for a small T value. The Acc@T concept allows to compare the predictive performance of different regression modes in a single graph, as proposed in (Bi & Bennett, 2003) with the REC curves, which plot in the y-axis the Acc@T for different T values (x-axis). The overall quality (for distinct T values) can be measured by computing the AREC curve when assuming a maximum tolerance of T_{max} (in %).

3.2.2 Results

Table 13 presents the test data errors, in terms of the RMSE error measure, for the HS evaluation and when comparing the two y_2 prediction strategies: $\widehat{y}_2$, executed during the second CRISP-DM iteration; and $\widehat{y}_{2\alpha\beta}$, explored in the third CRISP-DM iteration. The RMSE values confirm that for both prediction strategies, it is more difficult to predict the arrival of the first IPC sample ($s = 1$) than the arrival of the remaining samples ($s > 1$). It is interesting to notice that by specializing a learning model for each of these IPC sample types, as executed in the third CRISP-DM iteration ($\widehat{y}_{2\alpha\beta}$), a substantial error reduction is achieved for both sample types ($s = 1$ and $s > 1$).

Table 13: Test data holdout results for $s = 1$ and $s > 1$ IPC sample arrival (best values in **bold**).

	RMSE	
Approach	$s = 1$	$s > 1$
$\widehat{y}_2$	209.9	188.9
$\widehat{y}_{2\alpha\beta}$	**124.8**	**41.3**

The full comparison of the aggregated HS results, assuming all IPC samples, is shown in Table 14, which contains: the evaluation method used (**Eval.**); the best model selected using the AutoML procedure (**Model**); and several predictive performance measures. The AREC was computed by using a maximum tolerance of $T_{\text{max}=16}$ time units. All performance measures confirm that the best predictive model was achieved by $\widehat{y}_{2\alpha\beta}$, while $\widehat{y}_1$ obtained better results than $\widehat{y}_2$. When compared with $\widehat{y}_1$, $\widehat{y}_{2\alpha\beta}$ achieved a substantial predictive improvement: RMSE – reduction of 46.8 points; MAE – difference of 14.1 points; and AREC – increase of 10 percentage points. As for the ML algorithms, the AutoML selected GBM and SE as the best performing models when using the 10-fold internal cross-validation (applied over training data). The $\widehat{y}_{2\alpha\beta}$ uses GBM for predicting the arrival times of the $s = 1$ samples and SE for the other ones.

Figure 11 complements the HS results by showing the respective REC curves for the three main regression approaches. The plot confirms that for most of the low tolerance range (x-axis), $\widehat{y}_{2\alpha\beta}$ provides

a higher classification accuracy, resulting in an overall higher AREC. Indeed, the proposed two-stage ML model can predict correctly 37%, 59% and 70% of the samples for low tolerance values of $T = 1$, $T = 2$ and $T = 4$, a value that increases to 85% when the tolerance is increased to $T = 16$ time units.

Table 14: Test data results (best HO values in **bold)**.

					Acc@T					
Approach	**Eval.**	**Model**	**RMSE**	**MAE**	**AREC**	**T=1**	**T=2**	**T =4**	**T =8**	**T =16**
$\widehat{y}_1$		GBM	98.0	27.0	61%	28%	45%	56%	66%	76%
$\widehat{y}_2$	HO	SE	190.3	112.1	6%	1%	1%	3%	5%	12%
$\widehat{y}_{2\alpha\beta}$		α:GBM;β:SE	**51.2**	**12.9**	**71**%	**37**%	**59**%	**70**%	**77**%	**84**%
$\widehat{y}_{2\alpha\beta}$	RW	α:GBM;β:SE	37.5	11.4	71%	38%	56%	69%	76%	85%

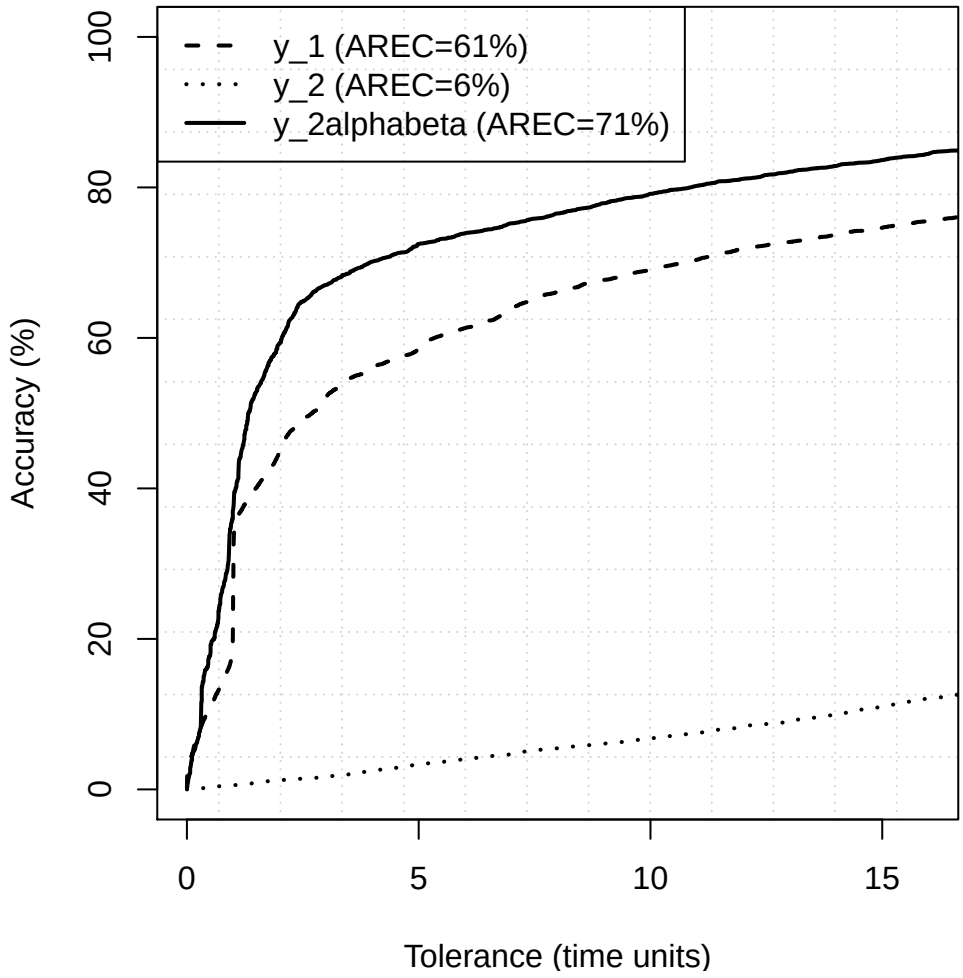


Figure 11: Holdout REC curves for the three regression approaches.

To estimate how the selected model ($\widehat{y}_{2\alpha\beta}$) would behave in a real environment setting, we tested it under a RW evaluation. Figure 12 presents the scheme of a RW. The results for all 20 week iterations are

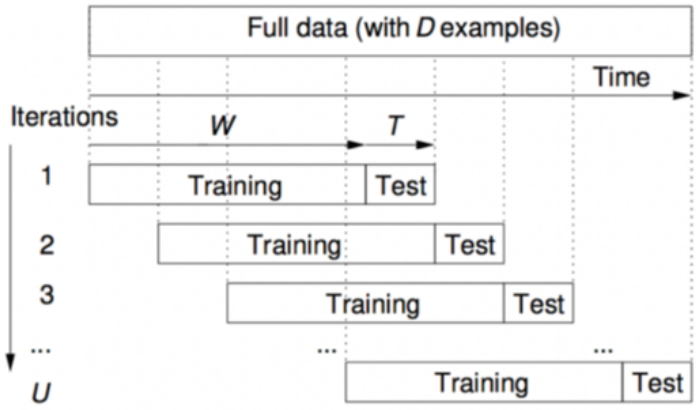


Figure 12: Schematic of the Rolling Window (RW) evaluation.

shown in terms of the last row of Table 14 and show consistency when compared with the HS evaluation. In effect, the same glsarec value is achieved (71%), while the RMSE and MAE values are slightly lower (RMSE of 37.5 and MAE of 11.4).

This is an interesting result, since the RW evaluation used more recent test data, not seen when comparing the HS results. The obtained results were presented to the business domain experts, which considered them very positive, encouraging the incorporation of the two-stage prediction model into a friendly dashboard that included several business indicators to support the Laboratory management decisions. To facilitate the visualization, the dashboard was designed to provide different granularity levels (hourly, daily or monthly) for the sample arrival prediction. For demonstrative purposes, Figure 13 plots the real and predicted values when assuming a daily aggregation of the IPC sample arrival for a particular Chemical Laboratory and for the entire RW testing time period. Due to business privacy issues, the scale of the y-axis is omitted from the graph. Figure 13 shows that the predictions are very close to the real values, denoting a high quality fit of the prediction model.

3.3 Predict Material Consumption in the Laboratories

3.3.1 Introduction

During the production process, selected samples are sent to be tested at the AL, which is responsible for assuring that the products are compliant with quality standards. The analysis of a sample at the AL requires diverse instrumental analyses, each consuming one or more materials (e.g., Acetone, Dichloromethane, Ethanol, Methanol). Under this context, predicting the amount of materials needed for the quality tests is crucial to support a AL stock management, preventing quality inspection delays which would prejudice production. In section Section 3.2, we have adopted a ML approach to successfully predict the arrival times of samples at the AL. By using this predictive approach, the Chemical organization can now perform weekly plans of the expected instrumental AL usage. Under this context, and having in account that

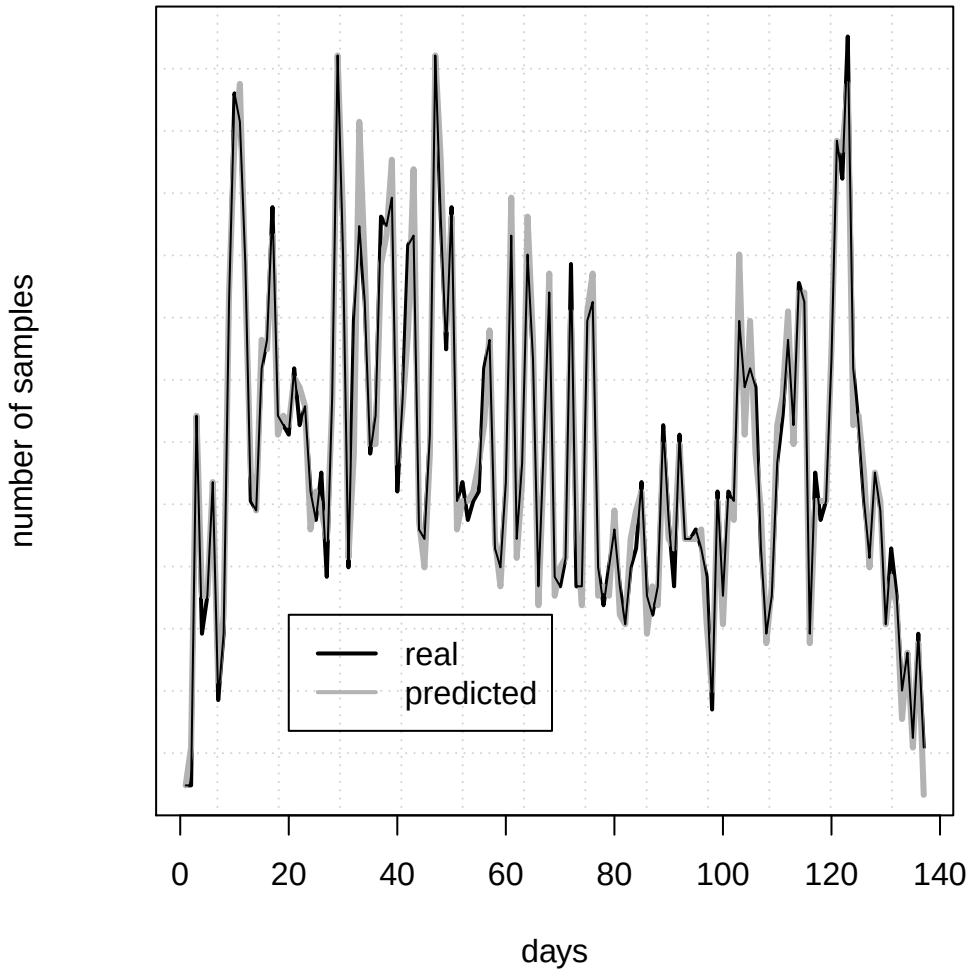


Figure 13: Daily sample arrival values and $\hat{y}_{2\alpha\beta}$ predictions for the rolling window test data.

the last section describes the first module of our IDSS, this section proposes a ML approach to predict the weekly consumption of AL materials based on the expected instrument usage. The approach was developed using the CRISP-DM methodology (Wirth & Hipp, 2000). Similarly to the work conducted in (A. J. Silva et al., 2020), to better focus on feature engineering (data preparation phase of CRISP-DM), we adopt an AutoML (Ferreira et al., 2020), which is executed during the modeling CRISP-DM phase and that allows to automatically select and tune the hyperparameters of the predictive ML models. Using real-word data, collected from a Chemical company, we executed several CRISP-DM iterations, exploring three main input variable selection strategies and two sets of AL materials (top 10 and all consumed materials). The experimentation adopts a realistic RW evaluation scheme, which simulates several train and test modeling updates through time. For benchmark purposes, the proposed ML approach is compared with two time series forecasting methods: the known ARIMA methodology (Box & Pierce, 1970) and a deep learning LSTM (Paszke et al., 2019).

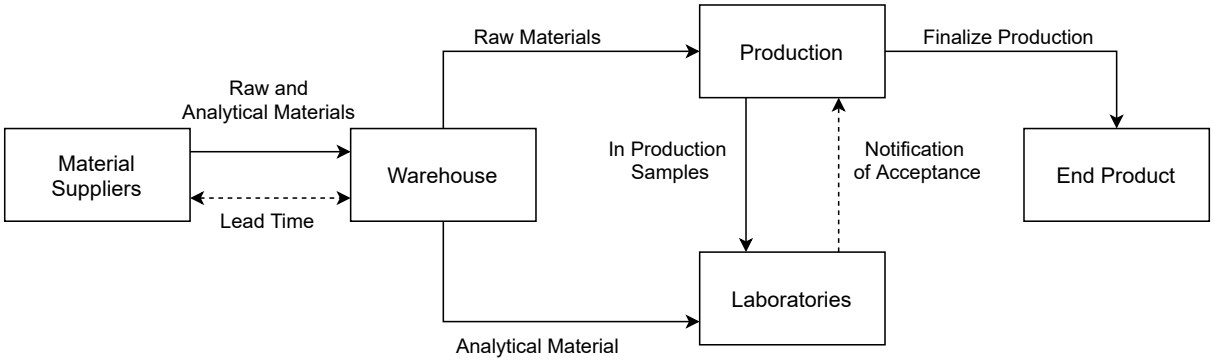


Figure 14: Workflow of materials and production transactions.

3.3.2 Problem Formulation

Figure 14 presents the flow of main transactions that occur between three main sections of the analyzed Chemical company: Warehouse, Production and AL. The Warehouse is responsible for storing and managing the different materials that are provided by the suppliers and that are needed by the company. (e.g., raw production materials). In this work, we focus on analytical materials, which are used in the AL. The Production line is where the production process is performed. A production of a certain product starts when there is a production order for that product on that specific date. A production order contains the several informational elements: the product to be produced, the quantity in batches to be produced, the raw materials to be used and the start and end dates. The dates are added to the database when the production ordered ends. During the production period, several production samples, called IPC, are sent to the AL for quality assessment. If quality is below the client requirements, then the production line will have to perform adjustments, in order improve the expected quality of the product. Thus, the AL are a critical element of the production process, with delays in AL testing resulting in production stops and delays in the execution of new production orders.

At the AL, the quality tests use several instrumental analyses that require analytical materials, in order to guarantee the feasibility of the tests. When there is an AL shortage of materials, they are ordered from the Warehouse, using the ERP production system. In some cases, there is a low stock of the analytical materials in the Warehouse, which needs to produce supplier orders that take time, thus producing AL quality testing delays. As stated earlier, in Section 3.2, we have adopted an AutoML approach to predict the arrival of IPC samples at the AL. Using such predictions, the company information system is capable of producing accurate week plans of AL instrumental needs. In this section, the ML goal is to use the AL tests (or plans) as the inputs of a regression model, aiming to predict a particular analytical material consumption. Let $\mathbf{X}$ denote a data matrix $N \times Q$ with the elements $x_{i,j}$, each representing the number of quality tests of type j that were executed (or are planned) for a particular week i, where N is the total number of weeks and Q is the total number of distinct quality tests. Let $\mathbf{Y}$ denote a matrix $N \times M$ with the elements $y_{i,m}$, each representing the quantity of consumed material of type $m \in \mathcal{M}$ for the week i, where $\mathcal{M} = \{1, 2, ..., M\}$ denotes a selection set with M distinct analytical materials. Another relevant

business concept is the AL total weekly consumption quantity ($T_{\mathcal{M}}$), computed as $T_{\mathcal{M}} = \sum_{m=1}^{M} y_{i,m}$. The total consumption quantity is useful for resizing the AL Warehouse.

The business goal is to estimate the w weekly quantity $\hat{y}_{w,m}$ based on the quality tests that use the m material:

$$\hat{y}_{w,m} = \quad f(x_{w,k_1}, ..., x_{w,k_K}) \tag{3.1}$$

where $\{k_1, ..., k_K\}$ denotes the set of Laboratory tests that are used as inputs and f is the data-driven function that will be learned using the AutoML approach. In this work, each m material consumption prediction requires the training of a different ML model. Moreover, the $\{k_1, ..., k_K\}$ input tests are dependent of the adopted feature selection strategy (Section 3.3.3.2). Once the distinct ML predictive models are built, the AL total weekly consumption quantity for selection $\mathcal{M}$ can be computed as: $\hat{T}_{\mathcal{M}} = \sum_{m=1}^{M} \hat{y}_{w,m}$.

3.3.3 Materials and Methods

3.3.3.1 Data

The data used in this study was retrieved by executing an ETL process, which extracted data records from two main databases related with the production and AL units. The resulting dataset includes a total of $N = 177$ weeks of data, from January 2016 to May 2019. In total, the input X matrix includes a total of $Q = 30$ distinct quality tests, thus with 177×30 elements. Some of the analyzed input tests have a strong correlation, while other variables often include a large portion of zero values. In Section 3.3.3.2, we will use these properties to design feature selection strategies. As for the target Y matrix, it includes a total of $M = 26$ analytical materials (e.g., Acetone, Ethanol, Methanol) After consulting the company experts, we explore two main sets of prediction targets: top 10 - with the $M = 10$ highest consumed materials ($\mathcal{M} = \{1, ..., 10\}$); and all – with all $M = 26$ materials ($\mathcal{M} = \{1, ..., 26\}$). Due to commercial privacy concerns, we do not disclose further details about the specific analyzed variables.

3.3.3.2 Prediction Methods

We adopted the `R` computational tool and its `rminer` package (Cortez, 2014) for data manipulation and computation of the ML regression metrics. The AutoML is based on the `H2O` open-source tool (https://www.h2o.ai/products/h2o-automl/) (Cook, 2016). The `auto.arima` from the `forecast R` package was used to automate and fit the ARIMA models (Box & Pierce, 1970; R. Hyndman et al., 2020; R. J. Hyndman & Khandakar, 2008). Finally, the LSTM model was implemented using the `PyTorch` Python module (Paszke et al., 2019).

The AutoML models were configured to select the best regression model and its hyperparameters for each targeted m material. The selection is based on the best RMSE computed using a validation set that is obtained by applying an internal 10-fold cross-validation method over the training data. All computational experiments were executed on the same personal computer and each individual ML model was trained up to a maximum running time of 3,600 seconds. Once a ML model is selected, the model was retrained

with all training data. As in Ferreira et al. (2020), the AutoML was configured to include a total of 6 distinct regression algorithms: RF, XRT, GLM, GBM, XGB and a SE.

The input matrix **X** includes several variables that are either correlated with other variables or contain a large number of zero values. In order to improve the AutoML results, we explore three main input Feature Selection (FS) strategies, that were applied to the training data: ALL - with all $Q = 30$ inputs, executed during the first CRISP-DM iteration; FS1 – all variables with a correlation higher than 60% or with more than 90% of zeros are removed (resulting in $Q = 15$), executed during the second CRISP-DM iteration; and FS2 - all variables with a correlation higher than 90% or with more than 90% of zeros are removed (leading to $Q = 19$), executed during the first CRISP-DM iteration.

For comparison purposes, we also consider two main time series forecasting methods, each using only the $y_{i,m}$ past observations ($i \in \{1, ..., m-1\}$) to predict $\hat{y}_{w,m}$ at week w: ARIMA and LSTM. The ARIMA is automatically build using the `forecast` R package, while the LSTM assumes a default parametrization with one input node (first time lag, y_{t-1}, where t is the current time), one hidden layer with 100 hidden nodes and hyperbolic tangent activation function, one output node (current observation, y_t), the Adam optimizer, MSE loss function and 150 training epochs.

3.3.3.3 Evaluation

We adopted a RW evaluation scheme (Oliveira et al., 2017; Tashman, 2000), which simulates a realistic execution of the AutoML models by performing several training and test updates through time (Figure 12). With this scheme, the initial training set with a fixed size of W time periods is used to generate the training models and execute a one week ahead prediction ($T = 1$). Then, the W data is updated by discarding the oldest week observations and adding one subsequent week of data. A new prediction model is built, allowing to issue a new prediction, and so on. In total, the RW results in $U = N - W$ training and testing updates. In this work, we have set $W = 147$, which allows to obtain $U = 30$ RW iterations. In order to reduce the computational effort, since we conduct a large number of ML experiments (e.g., we target $M = 26$ distinct outputs), the AutoML model and hyperparameter selection is only executed once for each m material, using the training data from the first RW iteration. Once the ML is selected, it is retrained for each RW iteration.

As for the regression metrics, using the $U = 30$ test predictions, we compute five measures (Hastie et al., 2009; Oliveira et al., 2017): MAE, NMAE, RMSE, Relative Squared Error (RSE), and the coefficient of determination (R^2). The lower the MAE, NMAE and RMSE values the better are the predictions. The NMAE measure is computed as $\frac{MAE}{\max(y_{i,m}) - \min(y_{i,m})}$, where $y_{i,m}$ denotes the target variable for material m. When compared with MAE, the NMAE metric presents two main advantages (Oliveira et al., 2017): it is more easy to interpret, since it expresses the error as a percentage of the full target scale (y); it is scale independent, which is useful for the analytical consumption data given that we handle different materials and thus distinct consumption scales. The RMSE measure is particularly important in this domain, since it is more sensitive to extreme values when compared with MAE. Thus, a lower RMSE should be aligned

with a better upper or lower peak prediction, which is more useful to assist the stock management of the consumed AL materials. The RSE is computed as $\frac{SSE_{\hat{y}_{i,m}}}{SSE_{\overline{y}_{i,m}}}$, where SSE denotes the sum of squared errors and $\overline{y}_{i,m}$ the average of the target variable on the test data. The RSE is similar to the RMSE measure in the sense that it is also more sensitive to extreme errors. The advantage is that RSE is scale independent. While the RSE values can be also presented as percentages (such as NMAE), the RSE values are more difficult to interpret by end users, since it only expresses how good are the predictions when compared with the average target values. As for R^2, it measures the goodness of fit. The higher value, the better is the alignment between consecutive changes in the predicted and real values, with the perfect regression model producing a maximum of R^2=1.

Since we target a large number of individual models (up to $M = 26$), the value of each forecasting approach is globally measured by considering the predictive measures applied to total quantity consumption target for a particular $\mathcal{M}$ selection. For instance, the RW MAE is computed as $MAE = \sum_{u_1}^{U} |T_{\mathcal{M}} - \hat{T}_{\mathcal{M}}|/U$, where u is a RW iteration and $\hat{T}_{\mathcal{M}}$ is the predicted total quantity consumption.

3.3.4 Results and Discussion

Table 15 summarizes the obtained RW predictive results for the total quantity consumption and $\mathcal{M}$ selection of materials. For instance, the upper left value of 193.0 corresponds to the MAE average when considering all $m \in \mathcal{M}, \mathcal{M} = \{1, 2, ..., 10\}$ highest consumed analytical materials of the top 10 selection set. The results from Table 15 confirm that different CRISP-DM iterations produced improved predictions, with the FS2 feature selection strategy obtaining the best AutoML results for all regression metrics. As for the time series forecasting baselines, the ARIMA methodology outperformed the LSTM neural network approach. Overall, the AutoML FS2 method produces the best predictions for the top 10 selection (for all regression measures) and the best RMSE, RSE and R^2 values for the all selection ($M = 26$). As explained in Section 3.3.3.3, for the improving stock management of the analytical materials, the squared error measures (RMSE and RSE) are more important than absolute error ones (MAE and NMAE). Regarding the optimized ML models, the AutoML procedure selected only three of the six considered regression algorithms: GLM, GBM and RF.

Table 15: Summary of the RW predictive results (best values in **bold**).

	Top 10 ($M = 10$)					**All** ($M = 26$)				
Method	**MAE**	**NMAE**	**RMSE**	**RSE**	**R^2**	**MAE**	**NMAE**	**RMSE**	**RSE**	**R^2**
AutoML ALL	193.0	6.30	338.0	51.9	0.49	80.58	2.63	205.6	42.4	0.58
AutoML FS1	203.7	6.66	347.4	54.9	0.46	83.86	2.74	209.6	44.1	0.56
AutoML FS2	**187.7**	**6.13**	**330.2**	**49.5**	**0.51**	78.89	2.58	**200.8**	**40.5**	**0.60**
ARIMA	189.1	6.18	349.0	55.3	0.47	**76.92**	**2.51**	210.4	44.5	0.57
LSTM	230.0	7.52	367.7	61.4	0.41	90.67	2.96	219.1	48.2	0.53

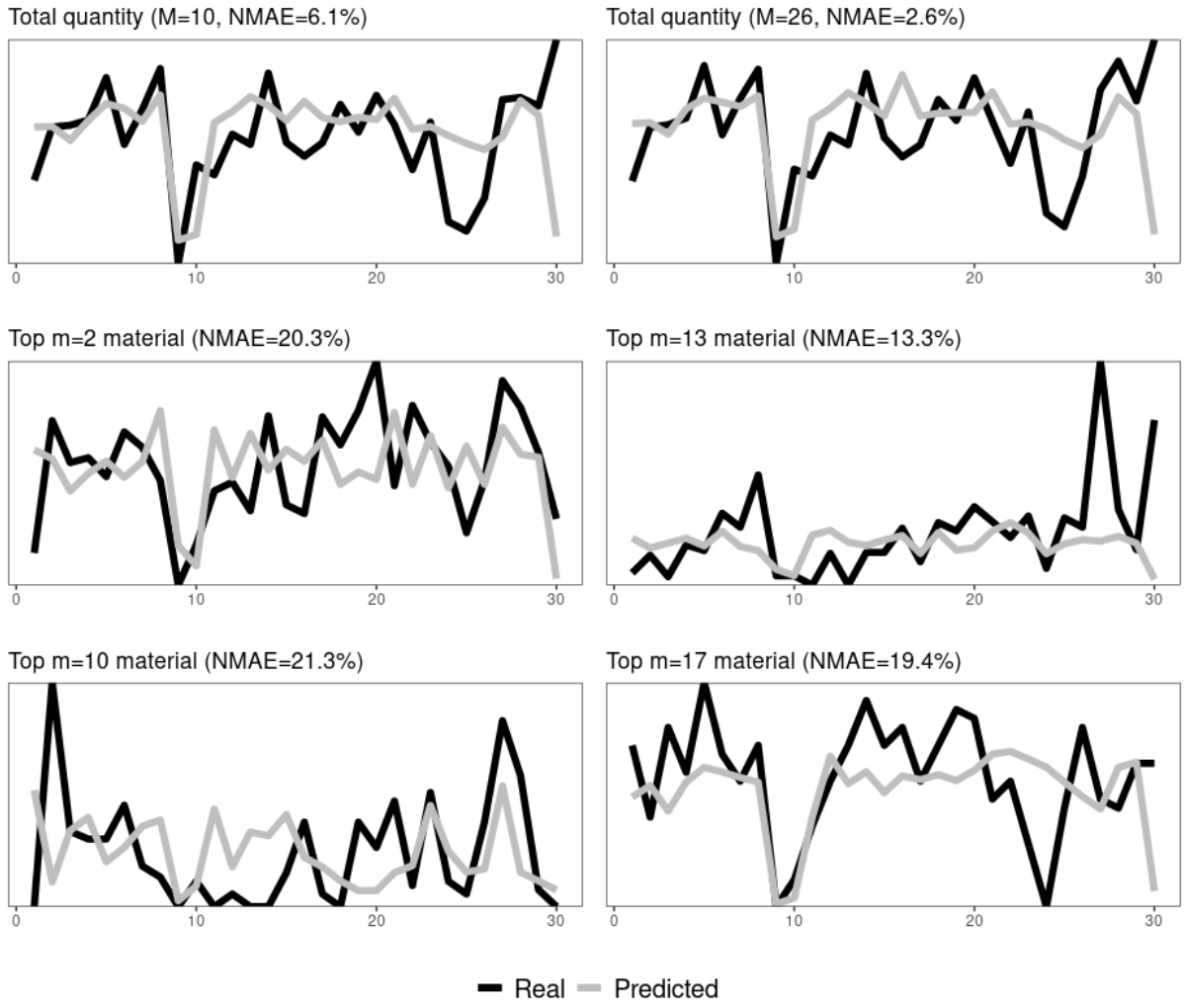


Figure 15: RW predictive results for AutoML FS2 method (x-axis denotes the considered week, from March 2019 to May 2019; y-axis shows the analytical material consumption).

For demonstration purposes, Figure 15 shows the RW predictions for the selected AutoML FS2 method, which provided the lowest squared errors and highest coefficient of determination values. Due to business privacy issues, the scale values of the y-axis are omitted from the plots. In the plots, we also present in brackets the NMAE errors, since these are more easy to be interpreted by the Chemical experts. The top two graphs show the results when predicting the total consumption (top 10 or all), while the middle and bottom graphs denote the prediction results for four individual materials ($m \in \{2, 10, 13, 17\}$). Overall, the real and predictive curves are very close and the prediction models are capable of correctly identifying several high and low consumption peaks, thus confirming that high quality predictions were obtained by the AutoML FS2 method.

The obtained results were shown to the chemical company experts, which highlighted the total quantity results, which can be used to resize the AL Warehouse. Moreover, the chemical experts considered that individual material predictions are interesting, such as for $m = 2$ and $m = 17$ from Figure 15, which have

a strong potential to improve the stock management of these materials.

3.4 An IDSS for Analytical Laboratories within the Industry 4.0 context

3.4.1 Introduction

In this section, we propose an IDSS that is based on Descriptive, Predictive and Prescriptive Analytics, aiming to assist the managerial decisions of AL from a Chemical Industry that is being transformed through the Industry 4.0 concept.

In previous works, we have proposed ML solutions to assist some partial AL tasks: predict the arrival time of IPC samples at the quality testing laboratories (Section 3.2); and estimate the AL materials consumption based on weekly plans of AL sample analyses (Section 3.3). In this section, we present the full IDSS that integrates both Predictive analytics, supporting the allocation of AL instruments (Prescriptive Analytics). The IDSS is also complemented with Descriptive Analytics executed over AL historical records, allowing the AL managers to better identify similarities among instruments. Prior to the Industry 4.0 transformation, the relevant digital records were spread in distinct databases, located in different departments (production and the AL), making the AL manager decisions more difficult. The proposed IDSS integrates all relevant data records into a single data repository, while also providing the business analytics results in terms of an interactive visual tool, based on dashboards. A IDSS prototype was deployed in the chemical company and then evaluated by the AL managers by using the TAM 3 (Venkatesh & Bala, 2008) and open interviews.

3.4.2 Materials and Methods

3.4.2.1 Problem Formulation

As stated earlier, the compay is from the chemical sector and it includes three main areas: Warehouse, Production and AL. The Warehouse is where the raw materials are received. It is also the destination of the products produced before being shipped to the customers. The Production area is where the chemical products are manufactured. Finally, the AL are responsible for testing all products and raw materials, checking if they meet the required quality standards. Before adopting an Industry 4.0 transformation, the entire communication process between these three areas was mainly manual and there was no real-time monitoring of the industrial processes, often leading to delays in the preparation of production materials or in the analyzes performed by the AL. These delays strongly affected deadlines for production plans.

Concerning the AL, these involve human analysts, instruments and several types of samples, namely RM, IPC and FP, that need to be analyzed, i.e., allocated into one or more analytical instruments. In

particular, IPC samples are are a priority because if they are not analyzed in a timely manner, the production process may stop. Each instrument allocation requires time and manual effort, to prepare and conduct the analysis and then collect the obtained results. There is an IS that records all quality test data, but such IT is mostly focused on the testing measurements and not on the AL processes. Thus, the management of the AL (e.g., human resource and instrument allocation planning, sample prioritization, prior preparation of instruments), assumes a strong manual effort, which is difficult due to the lack of a real-time data communication with the Warehouse and Production areas.

3.4.2.2 Proposed IDSS

To solve the previous mentioned AL management issues, and benefiting from the Industry 4.0 transformation performed at the company, we propose an IDSS that incorporates Descriptive, Predictive and Prescriptive Analytics. The proposed IDSS architecture is depicted in Figure 16. It includes two main layers. The Big Data layer is responsible for extracting and processing data from the different databases used in the organization. Indeed, the IDSS consumes the data from the different areas and applications from the organization (e.g., Warehouse, Production, AL), resulting as the ground truth data repository for the AL. The processed data is then fed into the Data Analytics layer, which incorporates Descriptive, Predictive and Prescriptive Analytics for AL management.

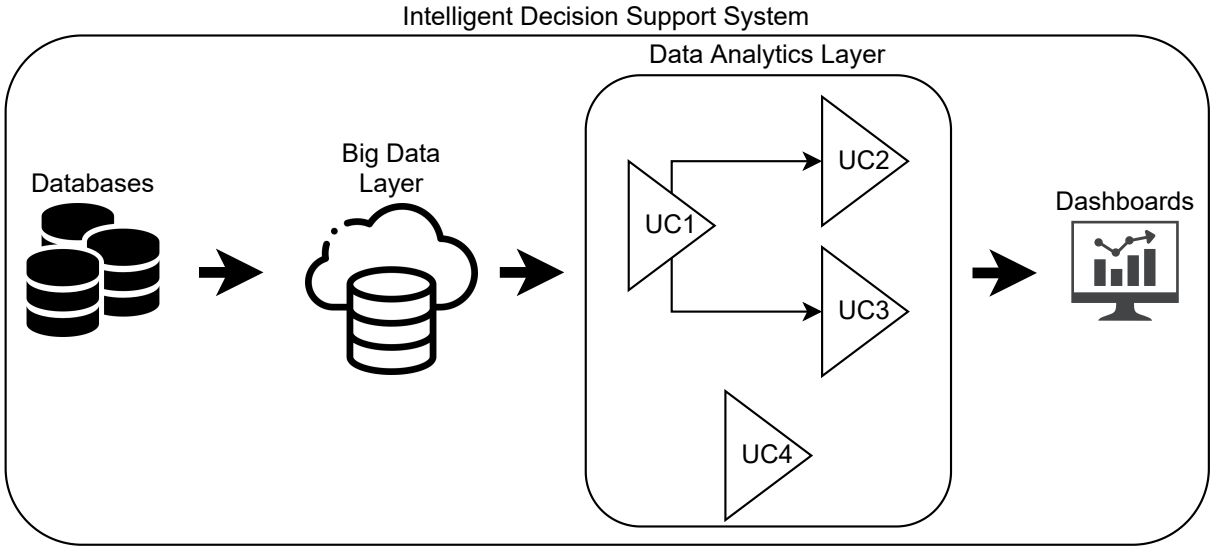


Figure 16: Proposed Architecture.

The developed tool includes two predictive models that were previously studied. Both models are based on an AutoML procedure but fed with different input attributes and training data. The proposed IDSS includes an extension of the first predictive model, termed here Use Case (UC) 1 (UC1), successfully tested for estimating the arrival of IPC samples at the ALs (Section 3.2). In this proposed IDSS, the model is adapted to perform predictions for all types of AL samples (the studied IPC and also the RM and FP). It should be noted that the predictions for the RM and FP samples ran only in one part of the hybrid

model, as there is no identification of the first sample of each batch. Since the results are still very much in an embryonic state, these were not considered in the sample arrival prediction study. However, the results are attached to this thesis in Annex 1. It should be noted that each sample arrived at the AL is associated with a fixed set of quality tests to be executed. The IDSS also integrates a second predictive model (UC2) that estimates the weekly consumption of AL materials (Section 3.3). This second predictive model requires, as input, a weekly plan of quality tests to be performed, which is built in advance by adopting the UC1 predictive model. The IDSS also includes Prescriptive Analytics (UC3), which is based on sample arrival estimates (UC1) and historical records regarding previous instrument allocations, allowing to provide suggestions of future instrument allocation. Finally, the IDSS also includes Descriptive Analytics set in terms of historical associations of instruments to quality tests (UC4), allowing to identify instrument similarities. All analytics are incorporated into friendly user dashboards.

Regarding the UC3, to issue recommendations of AL instruments allocation, we use a statistical approach that considers the UC1 predictions (tests to be executed) and that are matched with historical records of instrument allocation. For each required test, we assume as the "best" analytical instrument, the one currently available that has been mostly used for executing such test. An instrument is considered available if the its scheduled weekly allocation is lower than 70% (a value that was defined by the AL experts). Once an instrument is allocated, the IDSS is refreshed, with the allocation records being updated.

Finally, the UC4 is based on an $I \times T$ matrix computed using historical records and that measures the total number of tests ($t \in T$) executed by an instrument ($i \in I$). Then, the known Pearson correlation is used to compute the association between two rows of the matrix (i.e, two instruments). In our dashboards, the correlation matrix (Ferré, 2009) is shown as a colored heatmap, where more similar instruments are signaled by a stronger red color.

3.4.2.3 Evaluation

The proposed IDSS was developed by a research team that included both AI and Chemical company experts but not the direct AL managers. Thus, to properly evaluate the IDSS, we adopted the TAM 3 (Venkatesh & Bala, 2008), allowing to define a questionnaire that contains 10 questions and that was answered by the AL managers after experimenting the proposed tool. The questionnaire assumes the following TAM 3 constructs: Perceived Usefulness (PU), Perceived Ease of Use (PEOU), Perception of External Control (PEC), Job Relevance (REL), Output Quality (OUT), and Behavioral Intention (BI). Each question included a 5-point likert scale option for each answer, ranging from 1 (extremely disagree) to 5 (extremely agree). These questionnaires were complemented by a direct feedback from the AL managers, obtained by using open interviews in which the manager freely provided their opinions about the proposed IDSS. Furthermore, we also map the capabilities of the proposed IDSS tool, which are compared with the currently available AL informational processes (denoted as "As-Is") (Darwish, 2011).

3.4.3 Results

3.4.3.1 Developed IDSS Prototype

The designed IDSS was written using the R language, with the ML solutions being developed using specific R R Development Core Team, 2008 packages, namely `rminer` Cortez, 2014, `H2O AutoML` Aiello et al., 2016, `forecast` R. Hyndman et al., 2020; R. J. Hyndman and Khandakar, 2008 and `shiny` (Chang et al., 2021). The IDSS was fed with real-world data from the analyzed chemical company, collected from January 2016 to May 2019 and that results from a merge of the different databases adopted by the organization.

The user interface was developed using `shiny` and it includes three main dashboards to present the Descriptive (UC4), Predictive (UC1 and UC2) and Prescriptive (UC3) Analytics. The first dashboard presents: the expected arrival of samples and quality tests to be carried out in the current week (UC1); the expected raw material consumption (UC2); the history of quality analyzes carried out in the previous week; and an overview of the historical arrival of samples to the laboratory in the last year. The second dashboard shows the current allocation of AL instruments and suggestions on the best instrument to be used for each planned test (UC3). Finally, the last dashboard contains the correlation heatmaps based on the $I \times T$ association matrix (UC4).

The first dashboard is presented in Figure 17 and it contains three components. The first one is the top bar that shows warnings about issues that could occur during the current week. This includes information about how many instruments have an expected occupation above 50%, the number of analyzes without any instrument usage history, as well as the progress of test analyzes for the current day (in Figure 17, this value is set at 0%). The second middle component includes three tables, presenting: the daily sample arrival (UC1) predictions (left table); how many analyzes are planned to be carried out on the current day (middle table); and the predicted weekly AL material consumption (UC2, right table). The third bottom component has two graphs. The first plot (bottom left) shows the number of samples that arrived at the laboratories every week by type (IPC, RM, FP), while the second graph (bottom right) displays the number of analyzes performed per week by sample type.

The selection of the IDSS top menu tab allows the access to the second dashboard (Figure 18). The top left component “Analysis to be performed in this week” allows to select a quality test, refreshing the middle barplot graphs that show the instruments that are used for that specific test and sample (left) or just for that specific test (without sample specification, right plot). At the same time, the table on the top right presents the UC3 results as the suggested instrument to be assigned to that specific test analysis, along with the load work for the same instruments for that week. Finally, the bottom left table contains the information about the tests that have no historical records of instrument usage.

The last and third dashboard is presented in two figures and it is related with the UC4 Descriptive Analytics. Figure 19 displays the correlation tables for a given instrument divided by two groups of instrument machines: HPLC (left table) and GC (right table). The top buttons (“Chosse HPLC/GC”) allows the user to select one instrument from the displayed list. Once the instrument is selected, a table is displayed,

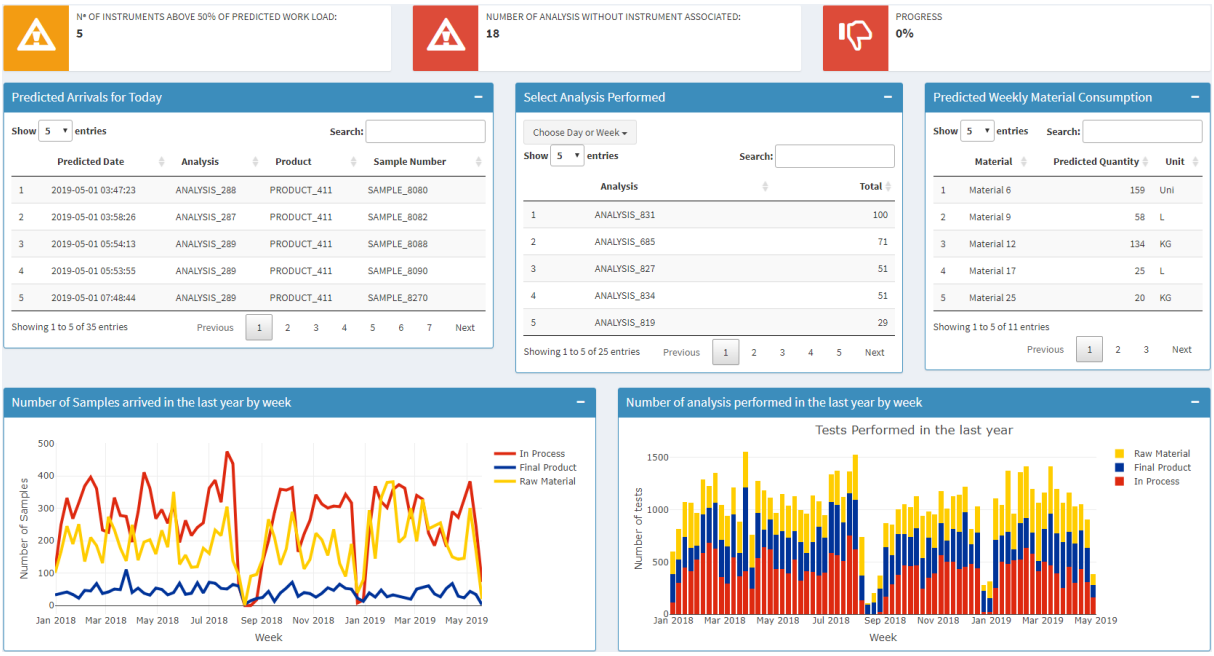


Figure 17: Example of the first IDSS dashboard.

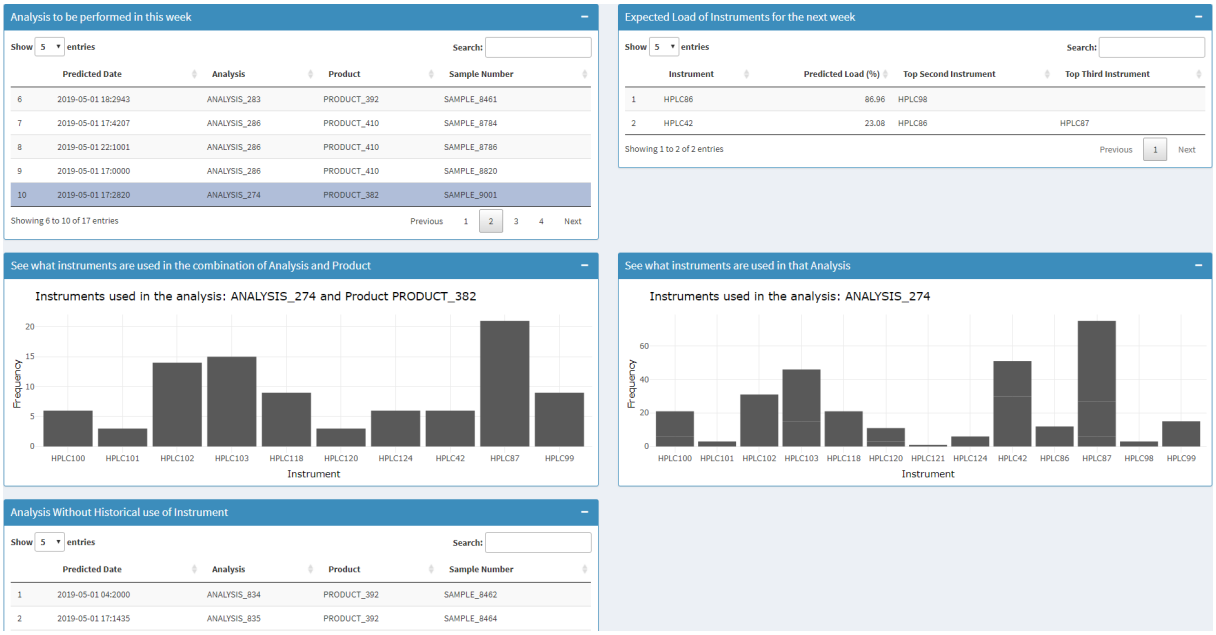


Figure 18: Example of the second dashboard.

sorting in a descending order the correlation values of most similar instruments. The third column on the tables shows the most used test analysis for each instrument. The bottom part of the third dashboard is presented in Figure 20, which shows the instrument correlation heatmaps for each group of instruments. The heatmap provides easy visualization of the most correlated HPLC and GC instruments.

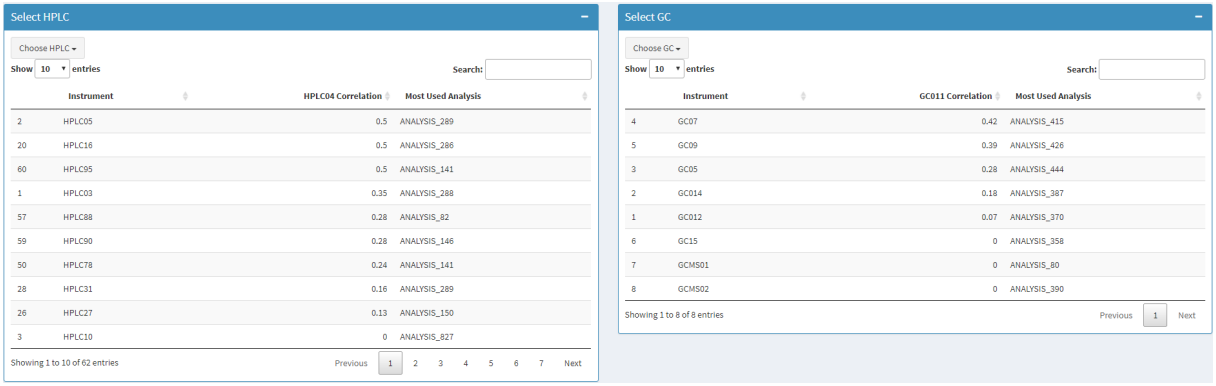


Figure 19: Example of the third dashboard (instruments correlation).

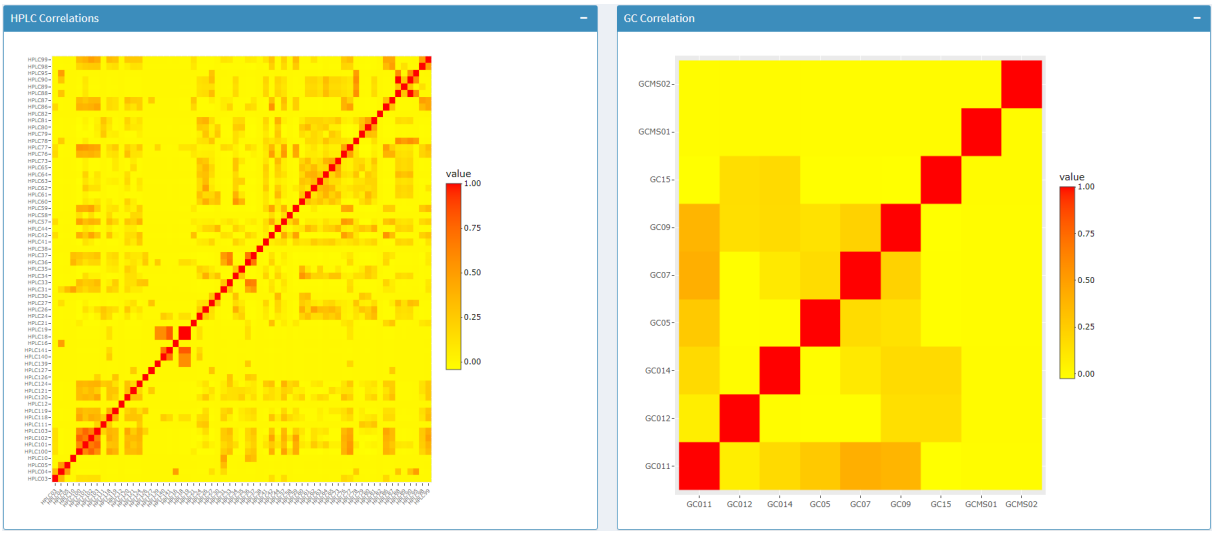


Figure 20: Example of the third dashboard (instruments heatmap).

3.4.3.2 Evaluation

The designed TAM 3 questionnaire is shown in Table 16. The obtained results are presented in Table 17, where each value corresponds to the average of two laboratory managers. We note that these managers correspond to IT AL staff from the analyzed chemical company and that were not directly involved in the presented research. The average responses are between 3.5 (70%) and 4 (80%), which means that laboratory managers had a positive acceptance of our IDSS. The most positive answers were related with the Perceived Usefulness (PU1 and PU2), Job Relevance (REL2) and Behavioral Intention (BI1). After obtaining the questionnaire responses, we have performed individual interviews, where the AL managers provided more specific feedback about the proposed IDSS. Regarding the first IDSS dashboard, both managers agreed that the information provided was simple and objective, being valuable to help the analysts to prepare the materials and the laboratory before the sample arrival. Turning to the second dashboard, related with the instruments load, they found it interesting but signaled the lack of information about new

Table 16: The adopted TAM 3 questionnaire.

Construct	Items	Question
Perceived Usefulness (PU)	PU1	Using the Dashboards improves my performance in my job.
	PU2	The Dashboards are (potentially) useful in my job.
Perceived Ease of Use (PEOU)	PEOU1	I find the Dashboard interface to be easy to use.
	PEOU2	It's easy to get the information that I want from the Dashboards.
Perceptions of External Control (PEC)	PEC1	I have the knowledge to use the Dashboards.
Job Relevance (REL)	REL1	In my job, the usage of the Dashboards is important.
	REL2	The use of the Dashboards is pertinent to my various job-related tasks.
Output Quality (OUT)	OUT1	The quality of the output I get from the Dashboards is high.
	OUT2	I have no difficulty telling others about the results of using the Dashboards.
Behavioral Intention (BI)	BI1	Assuming I had access to the Dashboard, I intend to use it.

Table 17: The TAM 3 questionnaire results (average of two responses).

PU1	PU2	PEOU1	PEOU2	PEC1	REL1	REL2	OUT1	OUT2	BI1
4	4	3.5	3.5	3.5	3.5	4	3.5	3.5	4

instruments and analyses. As for the third dashboard, it was considered helpful, particularly the correlation heatmap, which can be useful to identify new groups of instruments. However, such identification needs to be complemented by human domain knowledge, since there are instruments within the same group that can have different capabilities (e.g., refractive-index or infra-red). The AL managers also considered the dashboard useful to check if there a overlap between groups of instruments and if new groups of instruments could be defined. Overall, the AL managers concluded that the proposed IDSS (including its three dashboards), is valuable for planning the analyzes to be carried out on the samples, to improve the instrument allocation and to know how many analyzes will be carried out. Table 18 summarizes the main features introduced by the proposed IDSS, which substantially enhance the capabilities currently available at the AL (As-Is).

3.5 Summary

In the first published work, the sample arrival prediction, presented in Section 3.2, we adressed the non-trivial task of predicting the arrival of IPC samples at Chemical Laboratories for quality testing. To solve this task, we implemented the CRISP-DM methodology under three iterations, each focusing on a different

Table 18: Comparison between the current AL (As-Is) and proposed IDSS informational processes.

Capabilities	**As-Is**	**IDSS**
Historical overview of samples arrived	✓	✓
Historical overview of analysis performed	✓	✓
Sample arrival prediction		✓(UC1)
Weekly estimates of materials consumption		✓(UC2)
Expected instruments load		✓(UC3)
Suggested allocation of instruments		✓(UC3)
Information of analysis without instruments	✓	✓
Visualization of instrument similarities		✓(UC4)

regression approach. During the data understanding and preparation CRISP-DM stages, we collected recent data from a chemical company, resulting in 26,611 sample arrival examples related with a three-year time period. As for the modeling stage of CRISP-DM, we employed an AutoML procedure, to automatically select and configure the best model when exploring six state-of-the-art ML algorithms. Several experiments were held. Using a time ordered HS, we compared the three main regression approaches: $\widehat{y_1}$ - predict the time lag between the arrival of two consecutive samples (y_1), executed in the first CRISP-DM iteration; $\widehat{y_2}$ - predict the time lag between starting the production of the sample and its arrival to the laboratory (y_2), explored in the second CRISP-DM iteration; and $\widehat{y}_{2\alpha\beta}$ - a two-stage ML model to predict y_2, developed in the third CRISP-DM iteration. For all predictive performance measures, the best results were achieved at the two-stage ML model, which obtained interesting results (e.g., it can accurately predict 70% of the examples under a tolerance of $T = 4$ time units). The selected two-stage ML model ($\widehat{y}_{2\alpha\beta}$) was further evaluated using a realistic RW procedure, which considered 20 weeks of unseen data. A similar predictive performance was achieved, when compared with the HS results, showing that the proposed two-stage ML model is robust for the analyzed chemical company.

The second work, detailed in Section 3.3, addresses a relevant business goal of a chemical company that is being transformed under the Industry 4.0. In particular, a ML approach was conducted, aiming to predict the needs of materials (e.g., Acetone, Ethanol) used in their AL. The ML project was conducted using the CRISP-DM methodology. At the data understanding CRISP-DM stage, we collected 177 weeks of data, from January 2016 to May 2019, involving a total of 30 quality tests and up to 26 consumed AL materials. It should be noted that the chemical company is currently capable of producing weekly quality test usage plans with a good accuracy. Thus, the regression goal is to model AL material consumption as a function of the conducted quality tests. Using the collected data, we have developed large set of regression models (total of M =26 models), which were analyzed in terms of two major sets of material selections: top 10 most consumed materials (M=10) and all materials (M=26). To reduce the ML analyst effort, we have employed an AutoML procedure during the CRISP-DM modeling stage, which allows to automatically select the best among six different regression algorithms. A total of three CRISP-DM iterations were executed, each exploring a different FS method. For comparison purposes, we also considered two time

series forecasting methods: ARIMA and a LSTM NN. Several computational experiments were executed, by considering a realistic RW procedure that simulated 30 training and testing iterations through time. The best overall results were achieved by the AutoML FS2 method (corresponding to the third CRISP-DM iteration), which obtained a total quantity NMAE of 6.1% (top 10 selection) and 2.6% (all materials). The predictive results were shown to the AL managers, which provided a positive feedback.

Finally, in Section 3.4 we present an IDSS that was developed for the AL of a chemical company that is being transformed under the Industry 4.0 concept. The proposed IDSS includes two main layers: Big Data – responsible for extracting and processing data from different data sources, leading to a single and updated AL data repository; and Data Analytics – which includes Descriptive, Predictive and Prescriptive Analytics that aim to enhance the managerial decisions performed by AL managers. Using recent data from a real-world chemical company, in Sections 3.2 and 3.3 we have proposed two Predictive Analytics (IPC sample arrival prediction – UC1 and weekly AL materials consumption – UC2). The Data Analytics layer includes these analytics, extending the arrival prediction capabilities to all AL sample types (e.g., RM and FP). Moreover, it includes a novel Prescriptive method (UC3) for suggesting instrument allocations for quality tests based on historical records and the sample arrival predictions (UC1). Finally, it includes Descriptive Analytics regarding laboratory instrument similarities (UC4). A IDSS prototype was developed, which integrated all proposed analytics in three main interactive dashboards and used data collected from January 2016 to May 2019. The prototype was evaluated by two AL managers that were not directly involved in the IDSS design by adopting TAM 3 questionnaires and open interviews. Overall, a very positive feedback was obtained. In particular, the proposed IDSS was considered valuable to better prepare and assign instruments to samples, as well as to better estimate the ammount of quality tests that will be carried out.

Chapter 4

Conclusions

This chapter presents the conclusions of this doctoral thesis. Initially, a summary of the entire PhD work is presented, going through the definition of the research project, the SLR performed, as well as the summary of the work done throughout the project. The main results obtained are discussed. Finally, several future research directions are disclosed.

4.1 Overview

A major transformation is currently occurring due to the concept of Industry 4.0, also referred to as the fourth industrial revolution. Advances in IT, such as smart and cheaper sensors, IoT, Big Data and Business Analytics, are resulting in more integrated cyber-physical systems that can improve the production process. Business Analytics is a modern trend, defined after the 2010s and includes various Forecasting and Optimization techniques that can be used to analyze historical data and provide useful, often actionable, insights to support management decisions. These techniques applied in light of the concept of Industry 4.0, can potentially bring new insights and improvements to the productive processes.

This PhD program was inserted within a R&D project funded by a private company from the chemical sector, where the objective was to develop an intelligent system based on state-of-the-art technologies within the concept of Industry 4.0, to improve its processes and efficiency. This R&D project was divided into three different WP, with this PhD being inserted within the WP3, which aimed the design and development of an IS for AL based on a Big Data Warehouse system to collect and process all data. This PhD is specifically focused in the "intelligence" part of the project, where the objective of the project would be the integration of an intelligent system that uses Business Analytics techniques that is able to analyze the historical data of the Laboratory, within the context of Industry 4.0, with the aim of extracting knowledge to improve the management of the Laboratory.

An initial SLR was executed, allowing to verify that there are currently no DSS in AL, the research gap that was carried out in this work. Thus, the main objective of this thesis was to create an IDSS that would use Business Analytics techniques (Descriptive, Predictive and Prescriptive) to improve the Laboratory

management processes. In particular, we target tasks that were considered a priority for the analyzed Chemical company AL: the prediction of sample arrival, the prediction of material consumption and the allocation of the best available instrument for the analysis.

For sample arrival prediction, a two-stage ML model was proposed where one of the model predicts the arrival of the first samples of each production, and there is a second half that predicts the arrival of the remainder samples of the same production. This two-stage model emerged during the third iteration of CRISP-DM, where it was verified that it could be advantageous to divide the samples into two groups. This option proved to be the best approach based on the empirical results that were obtained. For the ML task, an AutoML methodology was used to speed up the process of selecting the best algorithms in the model. And for the model evaluation, to ensure that the AutoML platform can adapt over time, a 20-week RW was used.

Regarding the forecast of material consumption at the AL, the samples arrival predictions were used as input along with the historical data of material orders from the Laboratories to the Warehouses. After processing the data, three approaches were used to forecast material consumption in the Laboratories. Of these three approaches, one used regression techniques and the others used time series forecasting techniques. In the regression approach, the AutoML platform was used, and in the Time Series Forecasting (TSF) approaches an ARIMA methodology and deep learning LSTM were tested. The models were evaluated with a RW of 20 weeks, and it was concluded that for the top 10 materials, the AutoML regression was the best approach. The predictive results were considered as positive by the domain experts.

Next, we targeted the instrument allocation functionality and also the development of the full IDSS itself (with all the previously mentioned functionalities). The allocation of instruments is a type of prescriptive analytics. It is defined based on the conditions of the allocation of instruments to the analysis to be performed and the load of analysis already assigned to the instruments. Moreover, the ML models for predicting the arrival of samples and predicting the consumption of materials were also incorporated into the IDSS. In terms of its interface, the IDSS included three main Dashboards. These were evaluated by the AL managers by using TAM 3 based questionnaires and open interviews. The feedback from the Laboratory management staff was positive and they intend to adopt the proposed IDSS in their AL.

4.2 Discussion

Overall, the proposed IDSS offers a set of valuable functionalities that current AL do not have. These features bring new improvements to the AL management process. For example, sample arrival forecasting allows analysts to prepare material in advance to analyze IPC samples in a timely manner. And previously there was no certainty of when a sample would arrive for quality analysis. As for the prediction of consumption of materials, it allows ordering materials in a timely manner to the Warehouse, such that there is no shortage of materials during the sample analysis process. In addition, the IDSS offers new capabilities regarding instruments, such as suggesting the best instrument available for each analysis to

be performed, taking into account the expected load of analysis to be performed by each instrument and the specificity of the analysis to be performed (if it is specific to the product or not). Finally, the IDSS offers several data visualization techniques. For instance, in this case of instruments, we can check the similarity level of the instruments based on a heatmap.

While interesting results were achieved, we found a series of limitations throughout this doctoral project. Initially this project was to be applied to all the subsidiaries of this company. However, by choice of the organization that funded this project, the application was only limited to one of the companies. Moreover, although the system is being applied in the AL, initially it was planned with a higher integration level, such that is could be used both in production, in Warehouses or in the AL. Yet, the Chemical organization opted to focus only on the AL, which limited the scope of application of the proposed IDSS.

During the development of the proposed IDSS we encountered some limitations as in the evaluation of our IDSS when using TAM questionnaires. Indeed, we had a small number of responses due to the low number of Laboratory managers that were available in the company where this project was targeted. Finally, during the development of this IDSS we were able to predict with more accurate results the arrival of IPC type samples to the AL. We executed several initial attempts to predict other types of samples (e.g., RM and FP), but the obtained performance was considered lower and thus more research is needed. One aspect that limited the potential of the predictive results was the reduced access to input features that could influence the targeted outputs.

4.3 Future Work

In terms of future work, there are several interesting possibilities. The first one is the deployment of the proposed IDSS in the ALs of the analyzed Chemical company. This consists of the last step of the CRISP-DM methodology, and it would be important to monitor in a real environment the proposed IDSS in order to confirm its robustness or if it is necessary to make some adjustments over time. Furthermore, as previously mentioned, we intend to enlarge the research studies on the prediction of the arrival of the RM and FP samples, in order to complement our IDSS with this feature. Finally, we intend to implement a system similar to the one developed in this PhD in other industries aiming to check how general is the proposed approach, what adaptations does it require and also if it can provide value to other companies.

Bibliography

Abdelmaguid, T. (2020). Bi-objective dynamic multiprocessor open shop scheduling: An exact algorithm. *Algorithms*, *13*(3). https://doi.org/10.3390/a13030074 (cit. on p. 38)

Abdirad, M., Krishnan, K., & Gupta, D. (2020). A two-stage metaheuristic algorithm for the dynamic vehicle routing problem in industry 4.0 approach. *Journal of Management Analytics*, *8*(1), 69–83. https://doi.org/10.1080/23270012.2020.1811166 (cit. on p. 38)

Abdous, M.-A., Delorme, X., & Battini, D. (2020). Cobotic assembly line design problem with ergonomics. In L. M. Camarinha-Matos, H. Afsarmanesh, & A. Ortiz (Eds.), *Boosting collaborative networks 4.0* (pp. 573–582). Springer International Publishing. (Cit. on p. 38).

Aiello, S., Eckstrand, E., Fu, A., Landry, M., & Aboyoun, P. (2016). Machine learning with r and h2o. *H2O booklet* (cit. on p. 68).

Akhtari, S., Pickhardt, F., Pau, D., Di Pietro, A., & Tomarchio, G. Intelligent embedded load detection at the edge on industry 4.0 powertrains applications. In: *5th International Forum on Research and Technologies for Society and Industry: Innovation to Shape the Future, RTSI 2019 - Proceedings*. Institute of Electrical; Electronics Engineers Inc., 2019, 427–430. https://doi.org/10.1109/RTSI.2019.8895598 (cit. on p. 32).

Alasali, F., Haben, S., Foudeh, H., & Holderbaum, W. (2020). A comparative study of optimal energy management strategies for energy storage with stochastic loads. *Energies*, *13*, 2596. https://doi.org/10.3390/en13102596 (cit. on p. 34)

Albers, A., Stürmlinger, T., Wantzen, K., Bartosz, Gladysz, & Münke, F. (2017). Prediction of the product quality of turned parts by real-time acoustic emission indicators [Manufacturing Systems 4.0 - Proceedings of the 50th CIRP Conference on Manufacturing Systems]. *Procedia CIRP*, *63*, 348–353. https://doi.org/10.1016/j.procir.2017.03.173 (cit. on pp. 28, 35)

Ali, S., Qaisar, S. B., Saeed, H., Khan, M. F., Naeem, M., & Anpalagan, A. (2015). Network challenges for cyber physical systems with tiny wireless devices: A case study on reliable pipeline condition monitoring. *Sensors*, *15*(4), 7172–7205. https://doi.org/10.3390/s150407172 (cit. on p. 15)

Alter, S. (1980). *Decision support systems: Current practice and continuing challenges*. Addison-Wesley Pub. (Cit. on p. 44).

Ansari, F., Glawar, R., & Nemeth, T. (2019). Prima: A prescriptive maintenance model for cyber-physical production systems. *International Journal of Computer Integrated Manufacturing*, *32*(4-5), 482–503. https://doi.org/10.1080/0951192X.2019.1571236 (cit. on pp. 36, 37)

Antomarioni, S., Pisacane, O., Potena, D., Bevilacqua, M., Ciarapica, F. E., & Diamantini, C. (2019). A predictive association rule-based maintenance policy to minimize the probability of breakages: Application to an oil refinery. *The International Journal of Advanced Manufacturing Technology*, *105*. https://doi.org/10.1007/s00170-019-03822-y (cit. on p. 32)

Apiletti, D., Barberis, C., Cerquitelli, T., Macii, A., Macii, E., Poncino, M., & Ventura, F. (2018). Istep, an integrated self-tuning engine for predictive maintenance in industry 4.0. *IEEE Int. Conf. on Parallel Distributed Processing with Applications, Ubiquitous Computing Communications, Big Data Cloud Computing, Social Computing Networking, Sustainable Computing Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 924–931. https://doi.org/10.1109/BDCloud.2018.00136 (cit. on p. 30)

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. *Commun. ACM*, *53*(4), 50–58. https://doi.org/10.1145/1721654.1721672 (cit. on p. 15)

Arnott, D., & Pervan, G. (2008). Eight key issues for the decision support systems discipline. *Decision Support Systems*, *44*(3), 657–672. https://doi.org/10.1016/j.dss.2007.09.003 (cit. on p. 44)

Arnott, D., & Pervan, G. (2014). A critical analysis of decision support systems research revisited: The rise of design science. *Journal of Information Technology*, *29*(4), 269–293. https://doi.org/10.1057/jit.2014.16 (cit. on pp. 1, 12, 13, 44–46)

Aydemir, G., & Paynabar, K. (2019). Image-based prognostics using deep learning approach. *IEEE Transactions on Industrial Informatics*, *16*(9), 5956–5964 (cit. on p. 32).

Bagheri, M., Zollanvari, A., & Nezhivenko, S. (2018). Transformer fault condition prognosis using vibration signals over cloud environment. *IEEE Access*, *6*, 9862–9874. https://doi.org/10.1109/ACCESS.2018.2809436 (cit. on p. 21)

Bakar, N., Ramli, M., Sin, T., & Masran, H. (2019). A review on robotic assembly line balancing and metaheuristic in manufacturing industry. *AIP Conference Proceedings*, *2138*. https://doi.org/10.1063/1.5121084 (cit. on p. 17)

Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: Hyped up aspirations or true potential. *Vikalpa*, *38*(4), 1–12. https://doi.org/10.1177/0256090920130401 (cit. on p. 36)

Bányai, T., Illés, B., & Bányai, Á. (2018). Smart scheduling: An integrated first mile and last mile supply approach. *Complexity*, *2018*. https://doi.org/10.1155/2018/5180156 (cit. on p. 36)

Barton, D., & Court, D. (2012). Making advanced analytics work for you. *Harvard business review*, *90*, 78–83, 128 (cit. on p. 16).

Bellini, P., Cenni, D., Mitolo, N., Nesi, P., Pantaleo, G., & Soderi, M. (2021). High level control of chemical plant by industry 4.0 solutions. *Journal of Industrial Information Integration*, 100276. https://doi.org/https://doi.org/10.1016/j.jii.2021.100276 (cit. on p. 48)

Bi, J., & Bennett, K. P. (2003). Regression error characteristic curves. In T. Fawcett & N. Mishra (Eds.), *Machine learning, proceedings of the twentieth international conference (ICML 2003), august 21-24, 2003, washington, dc, USA* (pp. 43–50). AAAI Press. http://www.aaai.org/Library/ICML/2003/icml03-009.php. (Cit. on p. 56)

Birglen, L., & Schlicht, T. (2018). A statistical review of industrial robotic grippers. *Robotics and Computer-Integrated Manufacturing*, *49*, 88–97. https://doi.org/10.1016/j.rcim.2017.05.007 (cit. on pp. 25, 26)

BMBF. (2011). Industrie 4.0 - bmbf [Acessed: 2019-12-29]. https://www.bmbf.de/de/zukunftsprojekt-industrie-4-0-848.html. (Cit. on pp. 11, 13)

Bordel, B., & Alcarria, R. (2017). Assessment of human motivation through analysis of physiological and emotional signals in industry 4.0 scenarios. *Journal of Ambient Intelligence and Humanized Computing*, *9*, 1–21. https://doi.org/10.1007/s12652-017-0664-4 (cit. on p. 21)

Bordeleau, F.-E., Mosconi, E., & Santa-Eulalia, L. A. (2018). Business intelligence in industry 4.0: State of the art and research opportunities. *Proceedings of the 51st Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2018.495 (cit. on p. 17)

Borgi, T., Hidri, A., Neef, B., & Mohamed Saber, N. (2017). Data analytics for predictive maintenance of industrial robots. *2017 International Conference on Advanced Systems and Electric Technologies (IC ASET)*, 412–417. https://doi.org/10.1109/ASET.2017.7983729 (cit. on p. 28)

Bose, S. K., Kar, B., Roy, M., Gopalakrishnan, P. K., & Basu, A. (2019). Adepos: Anomaly detection based power saving for predictive maintenance using edge computing. *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, 597–602. https://doi.org/10.1145/3287624.3287716 (cit. on p. 32)

Bousdekis, A., Lepenioti, K., Ntalaperas, D., Vergeti, D., Apostolou, D., & Boursinos, V. (2019). A rami 4.0 view of predictive maintenance: Software architecture, platform and case study in steel industry. In H. A. Proper & J. Stirna (Eds.), *Advanced information systems engineering workshops* (pp. 95–106). Springer International Publishing. https://doi.org/10.1007/978-3-030-20948-3_9. (Cit. on p. 32)

Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, *65*(332), 1509–1526. http://www.jstor.org/stable/2284333 (cit. on pp. 59, 61)

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*(1), 49–64. https://doi.org/10.1007/BF00117832 (cit. on p. 54)

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726 (cit. on p. 42)

Brik, B., Bettayeb, B., Sahnoun, M., & Duval, F. Towards predicting system disruption in industry 4.0: Machine learning-based approach (S. E., Ed.). In: ed. by E., S. *151*. Elsevier B.V., 2019, 667–674. https://doi.org/10.1016/j.procs.2019.04.089 (cit. on pp. 36, 37).

Bruneo, D., & De Vita, F. (2019). On the Use of LSTM Networks for Predictive Maintenance in Smart Industries. *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, 241–248 (cit. on p. 32).

Bureau, U. C. (2017). New york city housing and vacancy survey (nychvs). Retrieved October 31, 2018, from https://www.census.gov/programs-surveys/nychvs/technical-documentation/code-lists/industry.html. (Cit. on p. 21)

Caetano, N., Cortez, P., & Laureano, R. M. S. (2014). Using data mining for prediction of hospital length of stay: An application of the CRISP-DM methodology. In J. Cordeiro, S. Hammoudi, L. A. Maciaszek, O. Camp, & J. Filipe (Eds.), *Enterprise information systems - 16th international conference, ICEIS 2014, lisbon, portugal, april 27-30, 2014, revised selected papers* (pp. 149–166). Springer. https://doi.org/10.1007/978-3-319-22348-3_9. (Cit. on p. 47)

Calabrese, M., Cimmino, M., Fiume, F., Manfrin, M., Romeo, L., Ceccacci, S., Paolanti, M., Toscano, G., Ciandrini, G., Carrotta, A., Mengoni, M., Frontoni, E., & Kapetis, D. (2020). Sophia: An event-based iot and machine learning architecture for predictive maintenance in industry 4.0. *Information*, *11*, 202 (cit. on pp. 34, 35).

Campitelli, G., & Gobet, F. (2010). Herbert simon's decision-making approach: Investigation of cognitive processes in experts. *Review of General Psychology*, *14*(4), 354–364. https://doi.org/10.1037/a0021256 (cit. on p. 44)

Candanedo, I., González, S., De la Prieta, F., & Arrieta, A. (2019). Maspi: A multi agent system for prediction in industry 4.0 environment. *Advances in Intelligent Systems and Computing*, *771*, 197–206. https://doi.org/10.1007/978-3-319-94120-2_19 (cit. on pp. 32, 39)

Canizo, M., Onieva, E., Conde, A., Charramendieta, S., & Trujillo, S. (2017). Real-time predictive maintenance for wind turbines using big data frameworks. *2017 IEEE International Conference on Prognostics and Health Management, ICPHM 2017, Dallas, TX, USA, June 19-21, 2017*, 70–77. https://doi.org/10.1109/ICPHM.2017.7998308 (cit. on pp. 21, 47)

Cao, G., Duan, Y., & Li, G. (2015). Linking business analytics to decision making effectiveness: A path model analysis. *IEEE Transactions on Engineering Management*, *62*(3), 384–395. https://doi.org/10.1109/TEM.2015.2441875 (cit. on p. 13)

Cao, Q., Zanni-Merk, C., Samet, A., De Bertrand de Beuvron, F., & Reich, C. (2020). Using rule quality measures for rule base refinement in knowledge-based predictive maintenance systems. *Cybernetics and Systems*, 1–16. https://doi.org/10.1080/01969722.2019.1705550 (cit. on pp. 34, 35)

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web application framework for r* [R package version 1.7.1]. https://CRAN.R-project.org/package=shiny. (Cit. on p. 68)

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-dm 1.0 step-by-step data mining guide*. SPSS. (Cit. on pp. 46, 47).

Charest, M., Finn, R., & Dubay, R. (2018). Integration of artificial intelligence in an injection molding process for on-line process parameter adjustment. *2018 Annual IEEE International Systems Conference, SysCon 2018, Vancouver, BC, Canada, April 23-26, 2018*, 1–6. https://doi.org/10.1109/SYSCON.2018.8369500 (cit. on pp. 30, 35, 48)

Chen, H., Li, L., & Chen, Y. (2020). Explore success factors that impact artificial intelligence adoption on telecom industry in china. *Journal of Management Analytics*, *8*(1), 36–68. https://doi.org/10.1080/23270012.2020.1852895 (cit. on p. 13)

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, ca, usa, august 13-17, 2016* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785. (Cit. on p. 54)

Chen, Y.-J., & Chien, C.-F. (2018). An empirical study of demand forecasting of non-volatile memory for smart production of semiconductor manufacturing. *International Journal of Production Research*, *56*(13), 4629–4643. https://doi.org/10.1080/00207543.2017.1421783 (cit. on p. 30)

Chen, Y., Chen, H., Gorkhali, A., Lu, Y., Ma, Y., & Li, L. (2016). Big data analytics and big data science: A survey. *Journal of Management Analytics*, *3*(1), 1–42. https://doi.org/10.1080/23270012.2016.1141332 (cit. on p. 15)

Chiang, L., Lu, B., & Castillo, I. (2017). Big data analytics in chemical engineering. *Annual Review of Chemical and Biomolecular Engineering*, *8*, 63–85. https://doi.org/10.1146/annurev-chembioeng-060816-101555 (cit. on pp. 16, 17)

Chi-Hsien, K., & Nagasawa, S. (2019). Applying machine learning to market analysis: Knowing your luxury consumer. *Journal of Management Analytics*, *6*(4), 404–419. https://doi.org/10.1080/23270012.2019.1692254 (cit. on p. 13)

Chiu, Y.-C., Cheng, F.-T., & Huang, H.-C. (2017). Developing a factory-wide intelligent predictive maintenance system based on industry 4.0. *Journal of the Chinese Institute of Engineers*, *40*(7), 562–571. https://doi.org/10.1080/02533839.2017.1362357 (cit. on p. 48)

Choi, W., Kim, J., Kim, S., & Kim, J. (2017). A study of reference metadata classification with deep learning. *International Conference on Information and Communication Technology Convergence (ICTC)*, 144–146. https://doi.org/10.1109/ICTC.2017.8190961 (cit. on pp. 28, 35)

Chong, D., & Shi, H. (2015). Big data analytics: A literature review. *Journal of Management Analytics*, *2*(3), 175–201. https://doi.org/10.1080/23270012.2015.1082449 (cit. on pp. 13, 16)

Cicconi, P., Russo, A. C., Germani, M., Prist, M., Pallotta, E., & Monteriù, A. (2017). Cyber-physical system integration for industry 4.0: Modelling and simulation of an induction heating process for aluminium-steel molds in footwear soles manufacturing. *IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI)*, 1–6. https://doi.org/10.1109/RTSI.2017.8065972 (cit. on p. 28)

Cisotto, S., & Herzallah, R. (2018). Performance prediction using neural network and confidence intervals: A gas turbine application. *2018 IEEE International Conference on Big Data (Big Data)*, 2151–2159. https://doi.org/10.1109/BigData.2018.8621919 (cit. on pp. 30, 35)

Clegg, D. (2015). Evolving data warehouse and bi architectures: The big data challenge. *Business Intelligence Journal*, *20*(1), 19–24 (cit. on p. 16).

Coley, C. W., Green, W. H., & Jensen, K. F. (2018). Machine learning in computer-aided synthesis planning. *Accounts of chemical research*, *51*(5), 1281–1289 (cit. on p. 49).

Commission, E. (2020). Digital Europe Programme: A proposed €7.5 billion of funding for 2021-2027. Retrieved January 18, 2021, from https://ec.europa.eu/digital-single-market/en/news/digital-europe-programme-proposed-eu75-billion-funding-2021-2027. (Cit. on p. 41)

Cook, D. (2016). *Practical machine learning with h2o: Powerful, scalable techniques for deep learning and ai*. O'Reilly Media. (Cit. on pp. 54, 61).

Cortez, P. (2014). *Modern optimization with r*. Springer. (Cit. on pp. 54, 61, 68).

Costa, C., & Santos, M. Y. (2017). The data scientist profile and its representativeness in the european e-competence framework and the skills framework for the information age. *International Journal of Information Management*, *37*(6), 726–734. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2017.07.010 (cit. on p. 41)

Costa, R., Figueiras, P., Jardim-Gonçalves, R., Ramos-Filho, J., & Lima, C. (2017). Semantic enrichment of product data supported by machine learning techniques. *2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 1472–1479. https://doi.org/10.1109/ICE.2017.8280056 (cit. on p. 21)

Darwish, A. (2011). *Business process mapping: A guide to best practice*. Writescope Publishers. (Cit. on p. 67).

Davenport, T. H. (2013). Analytics 3.0. *Harvard business review*, *91*(12), 64–72 (cit. on p. 41).

Deal, J. (2013). The ten most common data mining business mistakes. https://www.elderresearch.com/most-common-data-science-business-mistakes. (Cit. on p. 47)

de Sá, A., Casimiro, A., Machado, R., & Carmo, L. (2020). Identification of data injection attacks in networked control systems using noise impulse integration. *Sensors (Switzerland)*, *20*(3). https://doi.org/10.3390/s20030792 (cit. on p. 35)

Diez-Olivan, A., Del Ser, J., Galar, D., & Sierra, B. (2018). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0. *Information Fusion*, *50*, 92–111. https://doi.org/10.1016/j.inffus.2018.10.005 (cit. on p. 17)

Duan, L., & Xiong, Y. (2015). Big data analytics and business analytics. *Journal of Management Analytics*, *2*(1), 1–21. https://doi.org/10.1080/23270012.2015.1020891 (cit. on p. 16)

Durakbasa, N., Bauer, J., & Poszvek, G. (2017). Advanced metrology and intelligent quality automation for industry 4.0-based precision manufacturing systems. *Solid State Phenomena*, *261*, 432–439. https://doi.org/10.4028/www.scientific.net/SSP.261.432 (cit. on p. 25)

Dutta, R., Mueller, H., & Liang, D. (2018). An interactive architecture for industrial scale prediction: Industry 4.0 adaptation of machine learning. *Annual IEEE International Systems Conference (SysCon)*, 1–5. https://doi.org/10.1109/SYSCON.2018.8369547 (cit. on p. 22)

Dwaraka, R., & Arunachalam, N. (2018). Investigation on non-invasive process monitoring of die sinking edm using acoustic emission signals [46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA]. *Procedia Manufacturing*, *26*, 1471–1482. https://doi.org/10.1016/j.promfg.2018.07.094 (cit. on p. 30)

ESRTC. (2009). *Technology readiness levels: Handbook for space applications*. European Space Research; Technology Centre (ESRTC). https://books.google.pt/books?id=zzYKngEACAAJ. (Cit. on p. 24)

Essien, A., & Giannetti, C. (2020). A deep learning model for smart manufacturing using convolutional lstm neural network autoencoders. *IEEE Transactions on Industrial Informatics*, *PP*, 1–1. https://doi.org/10.1109/TII.2020.2967556 (cit. on p. 35)

European Commission. (2013). *Multi-annual roadmap for the contractual PPP under Horizon 2020*. (Cit. on p. 40).

Ferré, J. (2009). 3.02 - regression diagnostics. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive chemometrics* (pp. 33–89). Elsevier. https://doi.org/https://doi.org/10.1016/B978-044452701-1.00076-4. (Cit. on p. 67)

Ferreira, L., Pilastri, A., Martins, C., Santos, P., & Cortez, P. (2020). An automated and distributed machine learning framework for telecommunications risk management. In A. P. Rocha, L. Steels, & H. J. van den Herik (Eds.), *Proceedings of the 12th international conference on agents and artificial intelligence, ICAART 2020, volume 2, valletta, malta, february 22-24, 2020* (pp. 99–107). SCITEPRESS. https://doi.org/10.5220/0008952800990107. (Cit. on pp. 48, 54, 59, 62)

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28: Annual conference on neural information processing systems 2015, december 7-12, 2015, montreal, quebec, canada* (pp. 2962–2970). (Cit. on p. 52).

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2019). Auto-sklearn: Efficient and robust automated machine learning. *Automated machine learning: Methods, systems, challenges* (pp. 113–134). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_6. (Cit. on p. 43)

Feurer, M., Springenberg, J. T., & Hutter, F. (2015). Initializing bayesian hyperparameter optimization via meta-learning. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 1128–1135 (cit. on p. 42).

Fu, Y., Ding, J., Wang, H., & Wang, J. (2018). Two-objective stochastic flow-shop scheduling with deteriorating and learning effect in industry 4.0-based manufacturing system. *Applied Soft Computing*, *68*, 847–855. https://doi.org/10.1016/j.asoc.2017.12.009 (cit. on pp. 36, 37)

Geissbauer, R., Vedso, J., & Schrauf, S. (2016). Industry 4.0: Building the digital enterprise. *Retrieved from PwC Website: https://www. pwc. com/gx/en/industries/industries-4.0/landing-page/industry-4.0-building-your-digital-enterprise-april-2016. pdf* (cit. on pp. 12, 13).

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1 (cit. on p. 54)

Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S. H., Rodríguez-Roda, I., & Holmes, G. (2018). Which method to use? an assessment of data mining methods in environmental data science. *Environmental modelling & software*, *110*, 3–27 (cit. on p. 48).

Gomes, M., Silva, F., Ferraz, F., Silva, A., Analide, C., & Novais, P. (2017). Developing an ambient intelligent-based decision support system for production and control planning. *Advances in Intelligent Systems and Computing*, *557*, 984–994. https://doi.org/10.1007/978-3-319-53480-0_97 (cit. on p. 28)

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning*. MIT press Cambridge. (Cit. on pp. 41, 42).

Gottinger, H. W., & Weimann, P. (1992). Intelligent decision support systems. *Decision Support Systems*, *8*(4), 317–332. https://doi.org/10.1016/0167-9236(92)90053-R (cit. on p. 44)

Guo, Z., Ngai, E., Yang, C., & Liang, X. (2015). An RFID-based intelligent decision support system architecture for production monitoring and scheduling in a distributed manufacturing environment. *International Journal of Production Economics*, *159*, 16–28. https://doi.org/j.ijpe.2014.09.004 (cit. on p. 14)

Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Tin Kam Ho, Macià, N., Ray, B., Saeed, M., Statnikov, A., & Viegas, E. (2015). Design of the 2015 chalearn automl challenge. *2015 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2015.7280767 (cit. on p. 43)

Haffner, O., Kucera, E., Kozák, Š., & Stark, E. (2017). Proposal of system for automatic weld evaluation. *21st International Conference on Process Control (PC)*, 440–445. https://doi.org/10.1109/PC.2017.7976254 (cit. on p. 29)

Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2018). Chimera: Enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chemical science*, *9*(39), 7642–7655 (cit. on p. 49).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer. (Cit. on pp. 54, 62).

He, M., & He, D. (2017). Deep learning based approach for bearing fault diagnosis. *IEEE Transactions on Industry Applications*, *53*(3), 3057–3065. https://doi.org/10.1109/TIA.2017.2661250 (cit. on p. 29)

He, Y., Lin, J., Liu, Z., Wang, H., Li, L.-J., & Han, S. (2018). Amc: Automl for model compression and acceleration on mobile devices. *Proceedings of the European Conference on Computer Vision (ECCV)*, 784–800 (cit. on p. 43).

Hesser, D. F., & Markert, B. (2019). Tool wear monitoring of a retrofitted cnc milling machine using artificial neural networks. *Manufacturing Letters*, *19*, 1–4. https://doi.org/10.1016/j.mfglet.2018.11.001 (cit. on p. 32)

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2020). *forecast: Forecasting functions for time series and linear models* [R package version 8.13]. https://pkg.robjhyndman.com/forecast/. (Cit. on pp. 61, 68)

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22. https://www.jstatsoft.org/article/view/v027i03 (cit. on pp. 61, 68)

Jin, H., Song, Q., & Hu, X. (2019). Auto-keras: An efficient neural architecture search system. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1946–1956. https://doi.org/10.1145/3292500.3330648 (cit. on p. 43)

Jugulum, R. (2016). Importance of data quality for analytics. In P. Sampaio & P. Saraiva (Eds.), *Quality in the 21st Century: Perspectives from ASQ Feigenbaum Medal Winners* (pp. 23–31). Springer International Publishing. https://doi.org/10.1007/978-3-319-21332-3_2. (Cit. on p. 36)

Kabugo, J. C., Jämsä-Jounela, S.-L., Schiemann, R., & Binder, C. (2020). Industry 4.0 based process data analytics platform: A waste-to-energy plant case study. *International Journal of Electrical Power & Energy Systems*, *115*, 105508. https://doi.org/10.1016/j.ijepes.2019.105508 (cit. on pp. 35, 48)

Kagermann, H., Helbig, J., Hellinger, A., & Wahlster, W. (2013). *Recommendations for implementing the strategic initiative industrie 4.0: Securing the future of german manufacturing industry ; final report of the industrie 4.0 working group*. Forschungsunion. (Cit. on p. 13).

Kammergruber, R., Robold, S., Karliç, J., & Durner, J. (2014). The future of the laboratory information system–what are the requirements for a powerful system for a laboratory data management? *Clinical Chemistry and Laboratory Medicine (CCLM)*, *52*(11), 225–230 (cit. on p. 47).

Karakose, M., & Yaman, O. (2020). Complex fuzzy system based predictive maintenance approach in railways. *IEEE Transactions on Industrial Informatics*, *16*(9), 6023–6032 (cit. on p. 35).

Kaupp, L., Beez, U., Hülsmann, J., & Humm, B. G. (2019). Outlier detection in temporal spatial log data using autoencoder for industry 4.0. In J. Macintyre, L. Iliadis, I. Maglogiannis, & C. Jayne (Eds.), *Engineering applications of neural networks* (pp. 55–65). Springer International Publishing. https://doi.org/10.1007/978-3-030-20257-6_5. (Cit. on p. 26)

Kenessey, Z. (1987). The primary, secondary, tertiary and quaternary sectors of the economy. *Review of Income and Wealth*, *33*(4), 359–385 (cit. on p. 3).

Kharwar, P., Verma, R., Mandal, N., & Mondal, A. (2020). Swarm intelligence integrated approach for experimental investigation in milling of multiwall carbon nanotube/polymer nanocomposites. *Archive of Mechanical Engineering*, *67*(3), 353–376. https://doi.org/10.24425/ame.2020.131698 (cit. on p. 38)

Khatri, V., & Samuel, B. M. (2019). Analytics for managerial work. *Commun. ACM*, *62*(4), 100. https://doi.org/10.1145/3274277 (cit. on p. 13)

Khayyam, H., Jazar, R., Nunna, S., Golkarnarenji, G., Badii, K., Fakhrhoseini, S., Kumar, S., & Naebe, M. (2019). Pan precursor fabrication, applications and thermal stabilization process in carbon fiber production: Experimental and mathematical modelling. *Progress in Materials Science*, *107*, 100575. https://doi.org/10.1016/j.pmatsci.2019.100575 (cit. on pp. 36, 37, 39)

Kiangala, K., & Wang, Z. (2018). Initiating predictive maintenance for a conveyor motor in a bottling plant using industry 4.0 concepts. *International Journal of Advanced Manufacturing Technology*, *97*(9-12), 3251–3271. https://doi.org/10.1007/s00170-018-2093-8 (cit. on pp. 31, 35)

Kim, S., Lee, Y., Adhi Tama, B., & Lee, S. (2020). Reliability-enhanced camera lens module classification using semi-supervised regression method. *Applied Sciences*, *10*, 3832. https://doi.org/10.3390/app10113832 (cit. on p. 35)

Kirchen, I., Schütz, D., Folmer, J., & Vogel-Heuser, B. (2017). Metrics for the evaluation of data quality of signal data in industrial processes. *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, 819–826. https://doi.org/10.1109/INDIN.2017.8104878 (cit. on pp. 25, 26, 39)

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, *51*(1), 7–15 (cit. on p. 18).

Klement, N., & Silva, C. (2017). A generic decision support tool to planning and assignment problems: Industrial applications and industry 4.0. In B. Sokolov, D. Ivanov, & A. Dolgui (Eds.), *Scheduling in industry 4.0 and cloud manufacturing* (pp. 167–192). Springer International Publishing. https://doi.org/10.1007/978-3-030-43177-8_9. (Cit. on p. 36)

Koch, R. (2015). From business intelligence to predictive analytics. *Strategic Finance*, *96*, 56–57 (cit. on pp. 13, 22).

Kohlert, M., & König, A. (2016). Advanced multi-sensory process data analysis and on-line evaluation by innovative human-machine-based process monitoring and control for yield optimization in polymer film industry. *Technisches Messen*, *83*(9), 474–483. https://doi.org/10.1515/teme-2015-0120 (cit. on p. 27)

Krishnamoorthi, S., & Mathew, S. K. (2018). Business analytics and business value: A comparative case study. *Information & Management*, *55*(5), 643–666. https://doi.org/10.1016/j.im.2018.01.005 (cit. on pp. 1, 12)

Krishnan, K. (2013). *Data warehousing in the age of big data* (1st). Morgan Kaufmann Publishers Inc. https://doi.org/10.1016/C2012-0-02737-8. (Cit. on p. 16)

Kumar, A., Chinnam, R. B., & Tseng, F. (2018). An HMM and polynomial regression based approach for remaining useful life and health state estimation of cutting tools. *Computers & Industrial Engineering*. https://doi.org/10.1016/j.cie.2018.05.017 (cit. on p. 31)

Kuo, C.-J., Ting, K.-C., Chen, Y.-C., Yang, D.-L., & Chen, H.-M. (2017). Automatic machine status prediction in the era of industry 4.0: Case study of machines in a spring factory. *Journal of Systems Architecture*, *81*, 44–53. https://doi.org/10.1016/j.sysarc.2017.10.007 (cit. on pp. 25, 26)

Kuo, H., & Faricha, A. (2016). Artificial neural network for diffraction based overlay measurement. *IEEE Access*, *4*, 7479–7486. https://doi.org/10.1109/ACCESS.2016.2618350 (cit. on pp. 27, 35)

Langone, R., Cuzzocrea, A., & Skantzos, N. (2020). Interpretable anomaly prediction: Predicting anomalous behavior in industry 4.0 settings via regularized logistic regression tools. *Data & Knowledge Engineering*, *130*, 101850. https://doi.org/https://doi.org/10.1016/j.datak.2020.101850 (cit. on p. 48)

Larose, D. T. (2004). *Discovering knowledge in data: An introduction to data mining*. Wiley-Interscience. (Cit. on p. 42).

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering*, *6*, 239–242. https://doi.org/10.1007/s12599-014-0334-4 (cit. on p. 13)

Lasinkas, J. (2017). Industry 4.0: Penetrating digital technologies reshape global manufacturing sector. Retrieved June 25, 2018, from https://blog.euromonitor.com/2017/01/industry-4-0-penetrating-digital-technologies-reshape-global-manufacturing-sector.html. (Cit. on p. 14)

Lee, J., Lapira, E., Bagheri, B., & Kao, H. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. *Manufacturing Letters*, *1*(1), 38–41. https://doi.org/10.1016/j.mfglet.2013.09.005 (cit. on p. 15)

Lee, J., Kao, H.-A., & Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment [Product Services Systems and Value Creation. Proceedings of the 6th CIRP Conference on Industrial Product-Service Systems]. *Procedia CIRP*, *16*, 3–8. https://doi.org/10.1016/j.procir.2014.02.001 (cit. on p. 21)

Lee, W. J., Wu, H., Yun, H., Kim, H., Jun, M., & Sutheralnd, J. (2019). Predictive maintenance of machine tool systems using artificial intelligence techniques applied to machine condition data. *Procedia CIRP*, *80*, 506–511. https://doi.org/10.1016/j.procir.2018.12.019 (cit. on pp. 33, 35)

Lee, Y.-M., Lin, W. .-., Li, M.-H., Xiangqian, Z., & Li, J.-Y. (2016). Research into real-time analysis and exploration of influences on load rate of main shaft of machine of case companies with industry 4.0 technology. *2016 International Conference on Fuzzy Theory and Its Applications (iFuzzy)*, 1–7. https://doi.org/10.1109/iFUZZY.2016.8004968 (cit. on pp. 25, 26)

Leite, M., Pinto, T., & Alves, C. (2019). A real-time optimization algorithm for the integrated planning and scheduling problem towards the context of industry 4.0. *FME Transactions*, *47*(4), 775–781. https://doi.org/10.5937/fmet1904775L (cit. on p. 37)

Lenz, J., Wuest, T., & Westkämper, E. (2018). Holistic approach to machine tool data analytics [Special Issue on Smart Manufacturing]. *Journal of Manufacturing Systems*, *48*, 180–191. https://doi.org/10.1016/j.jmsy.2018.03.003 (cit. on pp. 25, 26)

Li, H. (2016). An approach to improve flexible manufacturing systems with machine learning algorithms. *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, 54–59. https://doi.org/10.1109/IECON.2016.7793838 (cit. on p. 36)

Li, S. C., Huang, Y., Tai, B. C., & Lin, C. T. (2017). Using data mining methods to detect simulated intrusions on a modbus network. *IEEE 7th International Symposium on Cloud and Service Computing (SC2)*, 143–148. https://doi.org/10.1109/SC2.2017.29 (cit. on pp. 29, 35)

Li, Y., Carabelli, S., Fadda, E., Manerba, D., Tadei, R., & Terzo, O. (2020). Machine learning and optimization for production rescheduling in industry 4.0. *International Journal of Advanced Manufacturing Technology*, *110*(9-10), 2445–2463. https://doi.org/10.1007/s00170-020-05850-5 (cit. on p. 38)

Li, Z., Wang, Y., & Wang, K.-S. (2017). Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario. *Advances in Manufacturing*, *5*(4), 377–387. https://doi.org/10.1007/s40436-017-0203-8 (cit. on p. 29)

Liang, Y., Kuo, C., & Lin, C. (2019). A hybrid memetic algorithm for simultaneously selecting features and instances in big industrial iot data for predictive maintenance. *IEEE 17th International Conference on Industrial Informatics (INDIN)*, *1*, 1266–1270. https://doi.org/10.1109/INDIN41052.2019.8972199 (cit. on p. 37)

Lin, C., Shu, L., Deng, D., Yeh, T., Chen, Y., & Hsieh, H. (2018). A mapreduce-based ensemble learning method with multiple classifier types and diversity for condition-based maintenance with concept drifts. *IEEE Cloud Computing*, 1–1. https://doi.org/10.1109/MCC.2017.455160123 (cit. on p. 30)

Lin, C., & Yang, J. (2018). Cost-efficient deployment of fog computing systems at logistics centers in industry 4.0. *IEEE Transactions on Industrial Informatics*, *14*(10), 4603–4611. https://doi.org/10.1109/TII.2018.2827920 (cit. on p. 25)

Lin, T., Chen, Y., Yang, D., & Chen, Y. (2016). New method for industry 4.0 machine status prediction - a case study with the machine of a spring factory. *International Computer Symposium (ICS)*, 322–326. https://doi.org/10.1109/ICS.2016.0071 (cit. on pp. 27, 39)

Liulys, K. (2019). Machine learning application in predictive maintenance. *2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 1–4 (cit. on pp. 33, 47).

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. *Journal of Management Analytics*, *6*(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365 (cit. on p. 13)

Ma, C., & Li, G. (2018). Prediction and Analysis of Tertiary Industry in Financial Center of Xinjiang under ´The Belt and Road Initiative´. *10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 143–146. https://doi.org/10.1109/ICMTMA.2018.00041 (cit. on p. 22)

Maggipinto, M., Terzi, M., Masiero, C., Beghi, A., & Susto, G. A. (2018). A computer vision-inspired deep learning architecture for virtual metrology modeling with 2-dimensional data. *IEEE Transactions on*

Semiconductor Manufacturing, *31*(3), 376–384. https://doi.org/10.1109/TSM.2018.2849206 (cit. on pp. 30, 35)

Mahmoodpour, M., Lobov, A., Lanz, M., Mäkelä, P., & Rundas, N. (2018). Role-based visualization of industrial iot-based systems. *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, 1–8. https://doi.org/10.1109/MESA.2018.8449183 (cit. on p. 48)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity* (tech. rep.). McKinsey & Company. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation. (Cit. on p. 15)

Martinek, P., & Krammer, O. (2019). Analysing machine learning techniques for predicting the hole-filling in pin-in-paste technology. *Computers and Industrial Engineering*, *136*, 187–194. https://doi.org/10.1016/j.cie.2019.07.033 (cit. on p. 33)

Masoudinejad, M., Venkatapathy, A. K. R., Tondorf, D., Heinrich, D., Falkenberg, R., & Buschhoff, M. (2018). Machine learning based indoor localisation using environmental data in phynetlab warehouse. *Smart SysTech 2018, European Conference on Smart Objects, Systems and Technologies*, 1–8 (cit. on p. 21).

Massaro, A., Manfredonia, I., Galiano, A., Pellicani, L., & Birardi, V. Sensing and quality monitoring facilities designed for pasta industry including traceability, image vision and predictive maintenance. In: Institute of Electrical; Electronics Engineers Inc., 2019, 68–72. https://doi.org/10.1109/METROI4.2019.8792912 (cit. on p. 33).

Massaro, A., Manfredonia, I., Galiano, A., & Xhahysa, B. (2019). Advanced process defect monitoring model and prediction improvement by artificial neural network in kitchen manufacturing industry: A case of study. *IEEE International Workshop on Metrology for Industry 4.0 and IoT, MetroInd 4.0 and IoT 2019 - Proceedings*, 64–67. https://doi.org/10.1109/METROI4.2019.8792872 (cit. on p. 33)

Mell, P. M., & Grance, T. (2011). *SP 800-145. The NIST Definition of Cloud Computing* (tech. rep.). National Institute of Standards & Technology. Gaithersburg, MD, United States. (Cit. on p. 15).

Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2006). *Adaptive business intelligence*. Springer. (Cit. on p. 44).

Milošević, M., Durdev, M., Lukić, D., Antić, A., & Ungureanu, N. (2020). Intelligent process planning for smart factory and smart manufacturing. *Lecture Notes in Mechanical Engineering*, 205–214. https://doi.org/10.1007/978-3-030-46212-3_14 (cit. on p. 38)

Miškuf, M., & Zolotová, I. (2016). Comparison between multi-class classifiers and deep learning with focus on industry 4.0. *Cybernetics Informatics (K I)*, 1–5. https://doi.org/10.1109/CYBERI.2016.7438633 (cit. on pp. 28, 35)

Mitchell, T. (1997). *Machine learning*. McGraw-Hill. (Cit. on p. 42).

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *Big data imperatives: Enterprise big data warehouse, bi implementations and analytics* (1st). Apress. (Cit. on p. 16).

Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R., & Anatole von Lilienfeld, O. (2013). Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, *15*(9), 095003. https://doi.org/10.1088/1367-2630/15/9/095003 (cit. on p. 49)

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., & Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using vis-nir spectroscopy [Proximal Soil Sensing – Sensing Soil Condition and Functions]. *Biosystems Engineering*, *152*, 104–116. https://doi.org/https://doi.org/10.1016/j.biosystemseng.2016.04.018 (cit. on p. 49)

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. *Proceedings of European Simulation and Modelling Conference-ESM'2011*, 117–121 (cit. on p. 47).

Mozgova, I., Yanchevskyi, I., Gerasymenko, M., & Lachmayer, R. (2018). Mobile automated diagnostics of stress state and residual life prediction for a component under intensive random dynamic loads [4th International Conference on System-Integrated Intelligence: Intelligent, Flexible and Connected Systems in Products and Production]. *Procedia Manufacturing*, *24*, 210–215. https://doi.org/10.1016/j.promfg.2018.06.037 (cit. on pp. 26, 39)

Muhuri, P., Shukla, A., & Abraham, A. (2019). Industry 4.0: A bibliometric analysis and detailed overview. *Engineering Applications of Artificial Intelligence*, *78*, 218–235. https://doi.org/10.1016/j.engappai.2018.11.007 (cit. on p. 17)

Mulrennan, K., Donovan, J., Creedon, L., Rogers, I., Lyons, J. G., & McAfee, M. (2018). A soft sensor for prediction of mechanical properties of extruded pla sheet using an instrumented slit die and machine learning algorithms. *Polymer Testing*, *69*, 462–469. https://doi.org/10.1016/j.polymertesting.2018.06.002 (cit. on p. 30)

Naskos, A., Gounaris, A., Metaxa, I., & Köchling, D. (2019). Detecting anomalous behavior towards predictive maintenance. In H. A. Proper & J. Stirna (Eds.), *Advanced information systems engineering workshops* (pp. 73–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-20948-3_7. (Cit. on p. 34)

Negri, E., Ardakani, H., Cattaneo, L., Singh, J., MacChi, M., & Lee, J. (2019). A digital twin-based scheduling framework including equipment health index and genetic algorithms. *IFAC–PapersOnLine*, *52*(10), 43–48. https://doi.org/10.1016/j.ifacol.2019.10.024 (cit. on p. 37)

Neuböck, T., & Schrefl, M. (2015). Modelling knowledge about data analysis processes in manufacturing. *IFAC-PapersOnLine*, *48*(3), 277–282. https://doi.org/10.1016/j.ifacol.2015.06.094 (cit. on pp. 24, 26, 48)

Nikolic, B., Ignjatic, J., Suzic, N., Stevanov, B., & Rikalovic, A. (2017). Predictive manufacturing systems in industry 4.0: Trends, benefits and challenges. *28TH DAAAM International Symposium on Intelligent Manufacturing and Automation*, 796–802. https://doi.org/10.2507/28th.daaam.proceedings.112 (cit. on pp. 16, 17)

Niño, M., Blanco, J. M., & Illarramendi, A. (2015). Business understanding, challenges and issues of big data analytics for the servitization of a capital equipment manufacturer. *IEEE International Conference on Big Data (Big Data)*, 1368–1377. https://doi.org/10.1109/BigData.2015.7363897 (cit. on pp. 24, 26, 48)

Nuzzi, C., Pasinetti, S., Lancini, M., Docchio, F., & Sansoni, G. (2018). Deep Learning Based Machine Vision: First Steps Towards a Hand Gesture Recognition Set Up for Collaborative Robots. *Workshop on Metrology for Industry 4.0 and IoT*, 28–33. https://doi.org/10.1109/METROI4.2018.8439044 (cit. on p. 30)

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine [PMID: 27682033]. *New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181 (cit. on p. 42)

Öchsner, A. (2013). Types of scientific publications. *Introduction to scientific publishing: Backgrounds, concepts, strategies* (pp. 9–21). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38646-6_3. (Cit. on p. 20)

O'Donovan, P., Leahy, K., Bruton, K., & O'Sullivan, D. T. J. (2015). Big data in manufacturing: A systematic mapping study. *Journal of Big Data*, *2*(1), 20. https://doi.org/10.1186/s40537-015-0028-x (cit. on pp. 16, 17)

Ogutu, J., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC proceedings*, *6 Suppl 2*, S10. https://doi.org/10.1186/1753-6561-6-S2-S10 (cit. on p. 42)

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Syst. Appl.*, *73*, 125–144. https://doi.org/10.1016/j.eswa.2016.12.036 (cit. on p. 62)

Packianather, M. S., Munizaga, N. L., Zouwail, S., & Saunders, M. (2019). Development of soft computing tools and iot for improving the performance assessment of analysers in a clinical laboratory. *14th Annual Conference System of Systems Engineering (SoSE)*, 158–163. https://doi.org/10.1109/SYSOSE.2019.8753830 (cit. on p. 33)

Pane, Y., Nageshrao, S., Kober, J., & Babuska, R. (2019). Reinforcement learning based compensation methods for robot manipulators. *Engineering Applications of Artificial Intelligence*, *78*, 236–247. https://doi.org/10.1016/j.engappai.2018.11.006 (cit. on pp. 37, 39)

Park, C. Y., Laskey, K. B., Salim, S., & Lee, J. Y. (2017). Predictive situation awareness model for smart manufacturing. *20th International Conference on Information Fusion (Fusion)*, 1–8. https://doi.org/10.23919/ICIF.2017.8009849 (cit. on p. 29)

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf. (Cit. on pp. 59, 61)

Peralta, G., Iglesias-Urkia, M., Barcelo, M., Gomez, R., Moran, A., & Bilbao, J. (2017). Fog computing based efficient IoT scheme for the Industry 4.0. *IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM)*, 1–6. https://doi.org/10.1109/ECMSM.2017.7945879 (cit. on pp. 29, 35)

Pierezan, J., Maidl, G., Massashi Yamao, E., dos Santos Coelho, L., & Cocco Mariani, V. (2019). Cultural coyote optimization algorithm applied to a heavy duty gas turbine operation. *Energy Conversion and Management*, *199*. https://doi.org/10.1016/j.enconman.2019.111932 (cit. on p. 37)

Pinto, R., & Cerquitelli, T. (2019). Robot fault detection and remaining life estimation for predictive maintenance (S. E., Ed.). *Procedia Computer Science*, *151*, 709–716. https://doi.org/10.1016/j.procs.2019.04.094 (cit. on p. 33)

Plehiers, P., Symoens, S., Amghizar, I., Marin, G., Stevens, C., & Van Geem, K. (2019). Artificial intelligence in steam cracking modeling: A deep learning algorithm for detailed effluent prediction. *Engineering*. https://doi.org/10.1016/j.eng.2019.02.013 (cit. on p. 33)

Ploennigs, J., Ba, A., & Barry, M. (2018). Materializing the promises of cognitive iot: How cognitive buildings are shaping the way. *IEEE Internet of Things Journal*, *5*(4), 2367–2374. https://doi.org/10.1109/JIOT.2017.2755376 (cit. on p. 26)

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, *97*, 205–227. https://doi.org/10.1016/j.eswa.2017.12.020 (cit. on p. 42)

Pradhan, K., & Chawla, P. (2020). Medical internet of things using machine learning algorithms for lung cancer detection. *Journal of Management Analytics*, *7*(4), 591–623. https://doi.org/10.1080/23270012.2020.1811789 (cit. on p. 14)

Proto, S., Ventura, F., Apiletti, D., Cerquitelli, T., Baralis, E., Macii, E., & Macii, A. (2019). Premises, a scalable data-driven service to predict alarms in slowly-degrading multi-cycle industrial processes. *IEEE International Congress on Big Data, BigData Congress 2019 - Part of the 2019 IEEE World Congress on Services*, 139–143. https://doi.org/10.1109/BigDataCongress.2019.00032 (cit. on p. 34)

Qi, Q., & Tao, F. (2018). Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access*, *6*, 3585–3593. https://doi.org/10.1109/ACCESS.2018.2793265 (cit. on pp. 16, 17)

Qin, J., Liu, Y., & Grosvenor, R. (2017). Data analytics for energy consumption of digital manufacturing systems using internet of things method. *13th IEEE Conference on Automation Science and Engineering (CASE)*, 482–487. https://doi.org/10.1109/COASE.2017.8256150 (cit. on pp. 25, 26)

Qin, J., Liu, Y., & Grosvenor, R. (2016). A categorical framework of manufacturing for industry 4.0 and beyond. *Procedia Cirp*, *52*, 173–178 (cit. on p. 23).

Qu, S., Wang, J., Govil, S., & Leckie, J. O. (2016). Optimized adaptive scheduling of a manufacturing process system with multi-skill workforce and multiple machine types: An ontology-based, multi-agent reinforcement learning approach [Factories of the Future in the digital environment - Proceedings of the 49th CIRP Conference on Manufacturing Systems]. *Procedia CIRP*, *57*, 55–60. https://doi.org/10.1016/j.procir.2016.11.011 (cit. on pp. 36, 39)

R Development Core Team. (2008). *R: A language and environment for statistical computing* [ISBN 3-900051-07-0]. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org. (Cit. on p. 68)

Rahman, H., Janardhanan, M., & Nielsen, P. (2020). An integrated approach for line balancing and agv scheduling towards smart assembly systems. *Assembly Automation*, *40*(2), 219–234. https://doi.org/10.1108/AA-03-2019-0057 (cit. on p. 38)

Rendall, R., Castillo, I., Lu, B., Colegrove, B., Broadway, M., Chiang, L. H., & Reis, M. S. (2018). Image-based manufacturing analytics: Improving the accuracy of an industrial pellet classification system using deep neural networks. *Chemometrics and Intelligent Laboratory Systems*, *180*, 26–35. https://doi.org/10.1016/j.chemolab.2018.07.001 (cit. on p. 31)

Rich, S. (2012). Big Data Is a 'New Natural Resource' IBM Says. Retrieved January 7, 2018, from http://www.govtech.com/policy-management/Big-Data-Is-a-New-Natural-Resource-IBM-Says.html. (Cit. on p. 15)

Richter, J., Streitferdt, D., & Rozova, E. (2017). On the development of intelligent optical inspections. *IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, 1–6. https://doi.org/10.1109/CCWC.2017.7868455 (cit. on p. 36)

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, *71*, 804–818. https://doi.org/https://doi.org/10.1016/j.oregeorev.2015.01.001 (cit. on p. 42)

Rogier, J., & Mohamudally, N. (2019). Forecasting photovoltaic power generation via an IoT network using nonlinear autoregressive neural network (S. E., Ed.). *Procedia Computer Science*, *151*, 643–650. https://doi.org/10.1016/j.procs.2019.04.086 (cit. on p. 34)

Romeo, L., Paolanti, M., Bocchini, G., Loncarski, J., & Frontoni, E. (2018). An Innovative Design Support System for Industry 4.0 Based on Machine Learning Approaches. *5th International Symposium on Environment-Friendly Energies and Applications (EFEA)*, 1–6. https://doi.org/10.1109/EFEA.2018.8617089 (cit. on pp. 36, 37)

Rosli, N., Ain Burhani, N., & Ibrahim, R. (2019). Predictive Maintenance of Air Booster Compressor (ABC) Motor Failure using Artificial Neural Network trained by Particle Swarm Optimization. *IEEE Student Conference on Research and Development (SCOReD)*, 11–16. https://doi.org/10.1109/SCORED.2019.8896330 (cit. on pp. 34, 35)

Rousopoulou, V., Nizamis, A., Giugliano, L., Haigh, P., Martins, L., Ioannidis, D., & Tzovaras, D. (2019). Data analytics towards predictive maintenance for industrial ovens. In H. A. Proper & J. Stirna (Eds.), *Advanced information systems engineering workshops* (pp. 83–94). Springer International Publishing. https://doi.org/10.1007/978-3-030-20948-3_8. (Cit. on pp. 34, 35)

Ruiz-Sarmiento, J., Monroy, J., Moreno, F.-A., Galindo, C., Bonelo, J., & González-Jiménez, J. (2020). A predictive model for the maintenance of industrial machinery in the context of industry 4.0. *Engineering Applications of Artificial Intelligence*, *87*, 103289. https://doi.org/10.1016/j.engappai.2019.103289 (cit. on p. 35)

Russom, P. (2014). Evolving data warehouse architectures in the age of big data. *The Data Warehousing Institute (TDWI)* (cit. on p. 16).

Russom, P. (2016). Data warehouse modernization in the age of big data analytics. *The Data Warehousing Institute (TDWI)* (cit. on p. 16).

Sala, D. A., Jalalvand, A., Deyne, A. V. Y., & Mannens, E. (2018). Multivariate time series for data-driven endpoint prediction in the basic oxygen furnace. In M. A. Wani, M. M. Kantardzic, M. S. Mouchaweh, J. Gama, & E. Lughofer (Eds.), *17th IEEE international conference on machine learning and applications, ICMLA 2018, orlando, fl, usa, december 17-20, 2018* (pp. 1419–1426). IEEE. https://doi.org/10.1109/ICMLA.2018.00231. (Cit. on pp. 31, 48)

Saldivar, A. A. F., Goh, C., Chen, W., & Li, Y. (2016). Self-organizing tool for smart design with predictive customer needs and wants to realize industry 4.0. *2016 IEEE Congress on Evolutionary Computation (CEC)*, 5317–5324. https://doi.org/10.1109/CEC.2016.7748366 (cit. on p. 28)

Saldivar, A. A. F., Goh, C., Li, Y., Chen, Y., & Yu, H. (2016). Identifying smart design attributes for industry 4.0 customization using a clustering genetic algorithm. *22nd International Conference on Automation and Computing (ICAC)*, 408–414. https://doi.org/10.1109/IConAC.2016.7604954 (cit. on pp. 28, 35)

Saldivar, A. A. F., Goh, C., Li, Y., Yu, H., & Chen, Y. (2016). Attribute identification and predictive customisation using fuzzy clustering and genetic search for industry 4.0 environments. *10th International Conference on Software, Knowledge, Information Management Applications (SKIMA)*, 79–86. https://doi.org/10.1109/SKIMA.2016.7916201 (cit. on pp. 28, 35, 39)

Santos, M. F., & Azevedo, C. S. (2005). *Preâmbulo [a]"data mining: Descoberta de conhecimento em bases de dados"*. FCA-Editora de Informática, Lda. (Cit. on p. 46).

Sanz, E., Matey, J. L., Blesa, J., & Puig, V. (2017). Advanced monitoring of an industrial process integrating several sources of information through a data warehouse. *4th International Conference on Control, Decision and Information Technologies (CoDIT)*, 0521–0526. https://doi.org/10.1109/CoDIT.2017.8102646 (cit. on pp. 25, 26)

Saxena, V. K., & Pushkar, S. (2016). Cloud computing challenges and implementations. *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2583–2588. https://doi.org/10.1109/ICEEOT.2016.7755159 (cit. on p. 15)

Schorfheide, F., & Wolpin, K. I. (2012). On the use of holdout samples for model selection. *American Economic Review*, *102*(3), 477–81 (cit. on p. 55).

Sellami, C., Miranda, C., Samet, A., Bach Tobji, M., & de Beuvron, F. (2019). On mining frequent chronicles for machine failure prediction. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-019-01492-x (cit. on pp. 34, 35)

Senkerik, R., Kadavy, T., Viktorin, A., & Pluhacek, M. Ensemble of strategies and perturbation parameter based soma for constrained technological design optimization problem. In: *IEEE Congress on Evolutionary Computation, CEC 2019 - Proceedings*. 2019, 2872–2877. https://doi.org/10.1109/CEC.2019.8790047 (cit. on p. 38).

Sezer, E., Romero, D., Guedea, F., Macchi, M., & Emmanouilidis, C. (2018). An Industry 4.0-Enabled Low Cost Predictive Maintenance Approach for SMEs. *IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 1–8. https://doi.org/10.1109/ICE.2018.8436307 (cit. on p. 31)

Sharp, M., Ak, R., & Hedberg, T. (2018). A survey of the advancing use and development of machine learning in smart manufacturing [Special Issue on Smart Manufacturing]. *Journal of Manufacturing Systems*, *48*, 170–179. https://doi.org/10.1016/j.jmsy.2018.02.004 (cit. on p. 17)

Shrouf, F., Ordieres, J., & Miragliotta, G. (2014). Smart factories in industry 4.0: A review of the concept and of energy management approached in production based on the internet of things paradigm. *IEEE International Conference on Industrial Engineering and Engineering Management*, 697–701. https://doi.org/10.1109/IEEM.2014.7058728 (cit. on pp. 1, 11, 47)

Silva, A. J., & Cortez, P. (2021). An automated machine learning approach for predicting chemical laboratory material consumption. In I. Maglogiannis, J. MacIntyre, & L. Iliadis (Eds.), *Artificial intelligence applications and innovations - 17th IFIP WG 12.5 international conference, AIAI 2021, hersonissos, crete, greece, june 25-27, 2021, proceedings* (pp. 105–116). Springer. https://doi.org/10.1007/978-3-030-79150-6_9. (Cit. on p. 50)

Silva, A. J., Cortez, P., & Pilastri, A. (2020). Chemical laboratories 4.0: A two-stage machine learning system for predicting the arrival of samples. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial intelligence applications and innovations - 16th IFIP WG 12.5 international conference, AIAI 2020, neos marmaras, greece, june 5-7, 2020, proceedings, part II* (pp. 232–243). Springer. https://doi.org/10.1007/978-3-030-49186-4_20. (Cit. on pp. 50, 59)

Silva, D., Jesus, K., Villaverde, B., & Adina, E. (2020). Hybrid Artificial Neural Network and Genetic Algorithm Model for Multi-Objective Strength Optimization of Concrete with Surkhi and Buntal Fiber. *Proceedings of the 12th International Conference on Computer and Automation Engineering*, 47–51. https://doi.org/10.1145/3384613.3384617 (cit. on pp. 36, 38)

Silva, N., Barros, J., Santos, M. Y., Costa, C., Cortez, P., Carvalho, M. S., & Gonçalves, J. N. C. (2021). Advancing logistics 4.0 with the implementation of a big data warehouse: A demonstration case for the automotive industry. *Electronics*, *10*(18). https://doi.org/10.3390/electronics10182221 (cit. on p. 48)

Skobelev, D., Zaytseva, T., Kozlov, A., Perepelitsa, V., & Makarova, A. (2011). Laboratory information management systems in the work of the analytic laboratory. *Measurement Techniques*, *53*(10), 1182–1189 (cit. on p. 47).

Soto, J. A. C., Tavakolizadeh, F., & Gyulai, D. (2019). An online machine learning framework for early detection of product failures in an industry 4.0 context. *International Journal of Computer Integrated Manufacturing*, *32*(4-5), 452–465. https://doi.org/10.1080/0951192X.2019.1571238 (cit. on pp. 34, 35)

Spendla, L., Kebisek, M., Tanuska, P., & Hrcka, L. (2017). Concept of predictive maintenance of production systems in accordance with industry 4.0. *IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 000405–000410. https://doi.org/10.1109/SAMI.2017.7880343 (cit. on pp. 29, 35, 47)

Stein, B. V., Leeuwen, M. V., Wang, H., Purr, S., Kreissl, S., Meinhardt, J., & Bäck, T. (2016). Towards Data Driven Process Control in Manufacturing Car Body Parts. *International Conference on Computational Science and Computational Intelligence (CSCI)*, 459–462. https://doi.org/10.1109/CSCI.2016.0093 (cit. on pp. 28, 35)

Steyerberg, E., van der Ploeg, T., & Van Calster, B. (2014). Risk prediction with machine learning and regression methods. *Biometrical journal. Biometrische Zeitschrift*, *56*. https://doi.org/10.1002/bimj.201300297 (cit. on p. 42)

Straus, P., Schmitz, M., Wostmann, R., & Deuse, J. (2018). Enabling of Predictive Maintenance in the Brownfield through Low-Cost Sensors, an IIoT-Architecture and Machine Learning. *IEEE International Conference on Big Data (Big Data)*, 1474–1483. https://doi.org/10.1109/BigData.2018.8622076 (cit. on pp. 31, 47)

Stürmlinger, T., Haar, C., Pandtle, J., & Niemeyer, V. (2018). Development of a wear model of a manufacturing system based on external smart production data on the example of a spring coiling machine [51st CIRP Conference on Manufacturing Systems]. *Procedia CIRP*, *72*, 232–236. https://doi.org/10.1016/j.procir.2018.03.260 (cit. on p. 26)

Subakti, H., & Jiang, J. (2018). Indoor Augmented Reality Using Deep Learning for Industry 4.0 Smart Factories. *IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, *02*, 63–68. https://doi.org/10.1109/COMPSAC.2018.10204 (cit. on p. 26)

Subramaniyan, M., Skoogh, A., Salomonsson, H., Bangalore, P., & Bokrantz, J. (2018). A data-driven algorithm to predict throughput bottlenecks in a production system based on active periods of the machines. *Computers & Industrial Engineering*. https://doi.org/10.1016/j.cie.2018.04.024 (cit. on p. 31)

Sun, I.-C., & Chen, K.-S. (2017). Development of signal transmission and reduction modules for status monitoring and prediction of machine tools. *56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 711–716. https://doi.org/10.23919/SICE.2017.8105459 (cit. on p. 29)

Susto, G. A., Schirru, A., Pampuri, S., Beghi, A., & DeNicolao, G. (2018). A hidden-gamma model-based filtering and prediction approach for monotonic health factors in manufacturing. *Control Engineering Practice*, *74*, 84–94. https://doi.org/10.1016/j.conengprac.2018.02.011 (cit. on p. 31)

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (1st). MIT Press. (Cit. on p. 42).

Swamy, A. K., & Sarojamma, B. (2020). Bank transaction data modeling by optimized hybrid machine learning merged with arima. *Journal of Management Analytics*, *7*(4), 624–648. https://doi.org/10.1080/23270012.2020.1726217 (cit. on p. 13)

Tan, Y., Goddard, S., & Pérez, L. C. (2008). A prototype architecture for cyber-physical systems. *SIGBED Rev.*, *5*(1), 26:1–26:2. https://doi.org/10.1145/1366283.1366309 (cit. on p. 14)

Tang, D., Zheng, K., Zhang, H., Sang, Z., Zhang, Z., Xu, C., Espinosa-Oviedo, J. A., Vargas-Solar, G., & Zechinelli-Martini, J.-L. (2016). Using Autonomous Intelligence to Build a Smart Shop Floor [The 9th International Conference on Digital Enterprise Technology - Intelligent Manufacturing in the Knowledge Economy Era]. *Procedia CIRP*, *56*, 354–359. https://doi.org/10.1016/j.procir.2016.10.039 (cit. on pp. 25, 26)

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review [The M3-Competition]. *International Journal of Forecasting*, *16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0 (cit. on pp. 55, 62)

Teschemacher, U., & Reinhart, G. (2017). Ant colony optimization algorithms to enable dynamic milkrun logistics [Manufacturing Systems 4.0 – Proceedings of the 50th CIRP Conference on Manufacturing Systems]. *Procedia CIRP*, *63*, 762–767. https://doi.org/10.1016/j.procir.2017.03.125 (cit. on p. 22)

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 847–855. https://doi.org/10.1145/2487575.2487629 (cit. on p. 43)

Tieng, H., Tsai, T., Chen, C., Yang, H., Huang, J., & Cheng, F. (2018). Automatic virtual metrology and deformation fusion scheme for engine-case manufacturing. *IEEE Robotics and Automation Letters*, *3*(2), 934–941. https://doi.org/10.1109/LRA.2018.2792690 (cit. on p. 26)

Tiwari, K., Shaik, A., & N, A. (2018). Tool wear prediction in end milling of ti-6al-4v through kalman filter based fusion of texture features and cutting forces [46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA]. *Procedia Manufacturing*, *26*, 1459–1470. https://doi.org/10.1016/j.promfg.2018.07.095 (cit. on p. 31)

Tjahjono, B., Esplugues, C., Ares, E., & Pelaez, G. (2017). What does industry 4.0 mean to supply chain? [Manufacturing Engineering Society International Conference 2017, MESIC 2017, 28-30 June 2017, Vigo (Pontevedra), Spain]. *Procedia Manufacturing*, *13*, 1175–1182. https://doi.org/10.1016/j.promfg.2017.09.191 (cit. on pp. 13, 14)

Tran, M.-Q., Elsisi, M., Mahmoud, K., Liu, M.-K., Lehtonen, M., & Darwish, M. M. F. (2021). Experimental setup for online fault diagnosis of induction machines via promising iot and machine learning: Towards industry 4.0 empowerment. *IEEE Access*, *9*, 115429–115441. https://doi.org/10.1109/ACCESS.2021.3105297 (cit. on p. 48)

Trunzer, E., Weiß, I., Folmer, J., Schrüfer, C., Vogel-Heuser, B., Erben, S., Unland, S., & Vermum, C. (2017). Failure mode classification for control valves for supporting data-driven fault detection. *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2346–2350. https://doi.org/10.1109/IEEM.2017.8290311 (cit. on pp. 25, 26, 39)

Tsai, S., & Chang, J. J. (2018). Parametric study and design of deep learning on leveling system for smart manufacturing. *IEEE International Conference on Smart Manufacturing, Industrial Logistics Engineering (SMILE)*, 48–52. https://doi.org/10.1109/SMILE.2018.8353980 (cit. on p. 31)

Tsourma, M., Zikos, S., Drosou, A., & Tzovaras, D. (2018). Online task distribution simulation in smart factories. *2nd International Symposium on Small-scale Intelligent Manufacturing Systems (SIMS)*, 1–6. https://doi.org/10.1109/SIMS.2018.8355301 (cit. on pp. 36, 37)

Uhlmann, E., Hohwieler, E., & Geisert, C. (2017). Intelligent production systems in the era of Industrie 4.0–changing mindsets and business models. *Journal of Machine Engineering*, *17* (cit. on pp. 16, 17).

Uriarte, A. G., Ng, A. H., & Moris, M. U. (2018). Supporting the lean journey with simulation and optimization in the context of industry 4.0 [Proceedings of the 8th Swedish Production Symposium (SPS 2018)]. *Procedia Manufacturing*, *25*, 586–593. https://doi.org/10.1016/j.promfg.2018.06.097 (cit. on pp. 36, 37, 39)

Vaishnavi, V., & Kuechler, B. (2004). Design science research in information systems. *Association for Information Systems* (cit. on p. 7).

Vathoopan, M., Johny, M., Zoitl, A., & Knoll, A. (2018). Modular fault ascription and corrective maintenance using a digital twin [16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018]. *IFAC-PapersOnLine*, *51*(11), 1041–1046. https://doi.org/10.1016/j.ifacol.2018.08.470 (cit. on p. 26)

Vazan, P., Janikova, D., Tanuska, P., Kebisek, M., & Cervenanska, Z. (2017). Using data mining methods for manufacturing process control. *IFAC-PapersOnLine*, *50*(1), 6178–6183. https://doi.org/10.1016/j.ifacol.2017.08.986 (cit. on p. 29)

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decis. Sci.*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x (cit. on pp. 8, 65, 67)

Ventura, F., Proto, S., Apiletti, D., Cerquitelli, T., Panicucci, S., Baralis, E., Macii, E., & Macii, A. (2019). A New Unsupervised Predictive-Model Self-Assessment Approach That SCALEs. *IEEE International Congress on Big Data*, 144–148. https://doi.org/10.1109/BigDataCongress.2019.00033 (cit. on p. 26)

Wahab, N., mat yasin, Z., Salim, N., & Ab Aziz, N. F. (2020). Artificial neural network based technique for energy management prediction. *Indonesian Journal of Electrical Engineering and Computer Science*, *17*, 94. https://doi.org/10.11591/ijeecs.v17.i1.pp94-101 (cit. on p. 49)

Wan, J., Tang, S., Li, D., Wang, S., Liu, C., Abbas, H., & Vasilakos, A. V. (2017). A manufacturing big data solution for active preventive maintenance. *IEEE Transactions on Industrial Informatics*, *13*(4), 2039–2047. https://doi.org/10.1109/TII.2017.2670505 (cit. on p. 29)

Wang, Y., Tercan, H., Thiele, T., Meisen, T., Jeschke, S., & Schulz, W. (2017). Advanced data enrichment and data analysis in manufacturing industry by an example of laser drilling process. *ITU Kaleidoscope: Challenges for a Data-Driven Society (ITU K)*, 1–5. https://doi.org/10.23919/ITU-WT.2017.8246990 (cit. on pp. 25, 26, 39)

Wang, Y.-M., Wang, Y.-S., & Yang, Y.-F. (2010). Understanding the determinants of rfid adoption in the manufacturing industry. *Technological Forecasting and Social Change*, *77*(5), 803–815. https://doi.org/10.1016/j.techfore.2010.03.006 (cit. on p. 14)

Wen, Z., Xie, L., Fan, Q., & Feng, H. (2020). Long term electric load forecasting based on ts-type recurrent fuzzy neural network model. *Electric Power Systems Research*, *179*, 106106. https://doi.org/https://doi.org/10.1016/j.epsr.2019.106106 (cit. on p. 48)

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 29–39 (cit. on pp. 47, 51, 59).

Wolfe, M. (1955). The concept of economic sectors. *The Quarterly Journal of Economics*, *69*(3), 402–420 (cit. on p. 3).

Wu, W., Zheng, Y., Chen, K., Wang, X., & Cao, N. (2018). A Visual Analytics Approach for Equipment Condition Monitoring in Smart Factories of Process Industry. *IEEE Pacific Visualization Symposium (PacificVis)*, 140–149. https://doi.org/10.1109/PacificVis.2018.00026 (cit. on p. 32)

Xia, F., Yang, L. T., Wang, L., & Vinel, A. (2012). Internet of things. *Int. J. Commun. Syst.*, *25*(9), 1101–1102. https://doi.org/10.1002/dac.2417 (cit. on p. 14)

Xu, L. D., He, W., & Li, S. (2014). Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, *10*(4), 2233–2243. https://doi.org/10.1109/TII.2014.2300753 (cit. on p. 14)

Xu, X., & Hua, Q. (2017). Industrial big data analysis in smart factory: Current status and research strategies. *IEEE Access*, *5*, 17543–17551. https://doi.org/10.1109/ACCESS.2017.2741105 (cit. on p. 17)

Xu, X. (2012). From cloud computing to cloud manufacturing. *Robotics and Computer-Integrated Manufacturing*, *28*(1), 75–86. https://doi.org/10.1016/j.rcim.2011.07.002 (cit. on p. 15)

Yan, H., Wan, J., Zhang, C., Tang, S., Hua, Q., & Wang, Z. (2018). Industrial big data analytics for prediction of remaining useful life based on deep learning. *IEEE Access*, *6*, 17190–17197. https://doi.org/10.1109/ACCESS.2018.2809681 (cit. on p. 32)

Yan, J., Meng, Y., Lu, L., & Li, L. (2017). Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance. *IEEE Access*, *5*, 23484–23491. https://doi.org/10.1109/ACCESS.2017.2765544 (cit. on p. 29)

Yang, H., & Tate, M. (2009). Where are we at with Cloud Computing?: A Descriptive Literature Review. *ACIS 2009 Proceedings - 20th Australasian Conference on Information Systems*, 807–819 (cit. on p. 15).

Yang, J., Chen, Y., Huang, W., & Li, Y. (2017). Survey on artificial intelligence for additive manufacturing. *23rd International Conference on Automation and Computing (ICAC)*, 1–6. https://doi.org/10.23919/IConAC.2017.8082053 (cit. on p. 17)

Yeh, W., Lai, C., & Tsai, J. (2019). Simplified swarm optimization for optimal deployment of fog computing system of industry 4.0 smart factory. *Journal of Physics: Conference Series*, *1411*(1). https://doi.org/10.1088/1742-6596/1411/1/012005 (cit. on p. 38)

Zenisek, J., Wolfartsberger, J., Sievi, C., & Affenzeller, M. (2019). Modeling sensor networks for predictive maintenance. In C. Debruyne, H. Panetto, W. Guédria, P. Bollen, I. Ciuciu, & R. Meersman (Eds.), *On the Move to Meaningful Internet Systems: OTM 2018 Workshops* (pp. 184–188). Springer International Publishing. https://doi.org/10.1007/978-3-030-11683-5_20. (Cit. on p. 34)

Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: State-of-the-art and research challenges. *Journal of Internet Services and Applications*, *1*(1), 7–18. https://doi.org/10.1007/s13174-010-0007-6 (cit. on p. 15)

Zhang, T., Feng, Y., & Hao, B. Industrial intelligent forecast of tft-lcd based on r-svm. In: Institute of Electrical; Electronics Engineers Inc., 2019, 25–30. https://doi.org/10.1109/ICIAICT.2019.8784848 (cit. on p. 34).

Zheng, M., & Wu, K. (2017). Smart spare parts management systems in semiconductor manufacturing. *Industrial Management & Data Systems*, *117*(4), 754–763. https://doi.org/10.1108/IMDS-06-2016-0242 (cit. on pp. 25, 26, 39)

Zhong, M., Tran, K., Min, Y., Wang, C., Wang, Z., Dinh, C.-T., De Luna, P., Yu, Z., Rasouli, A. S., Brodersen, P., et al. (2020). Accelerated discovery of co2 electrocatalysts using active machine learning. *Nature*, *581*(7807), 178–183 (cit. on p. 49).

Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent manufacturing in the context of industry 4.0: A review. *Engineering*, *3*(5), 616–630. https://doi.org/10.1016/J.ENG.2017.05.015 (cit. on pp. 14–16)

Zhou, H., & Yu, K. (2017). Imbalanced data classification for defective product prediction based on industrial wireless sensor network. *Sixth International Conference on Future Generation Communication Technologies (FGCT)*, 1–6. https://doi.org/10.1109/FGCT.2017.8103728 (cit. on p. 29)

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, 912–919 (cit. on p. 42).

Annex I

Annex 1 RM and FP first experimental results

Table 19: Test data results from the first experimental test to predict the arrival of FP and RM samples.

Sample Type	**RMSE**	**MAE**	T**=1**	T**=2**	T **=4**	T **=8**	T **=24**	T **=48**
RM	174.25	77.74	0.16%	0.58%	1.09%	6.54%	33.58%	67.35%
FP	139.18	62.10	6.24%	11.01%	23.49%	44.04%	68.89%	76.33%