FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

IoT and Machine Learning for Process Optimization in Agrofood Industry

Gonçalo de Soares Lemos

FOR JURY EVALUATION

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Pedro Alexandre Rodrigues João

Orientadores Externos: Raquel Oliveira, Vasco Pires

July 23, 2019

Resumo

Em ambiente de produção alimentar de elevado volume, é particularmente importante garantir que os produtos comercializados se encontram dentros das normas de qualidade e segurança. Nesse sentido, a sua monitorização afigura-se crucial para garantir a qualidade pretendida. Com alguma frequência, são necessários ajustes e alterações ao processo para que o produto seja adequado para armazenagem e manuseamento e posterior consumo, enquadrado com os requisitos de qualidade. Na maioria dos casos, os problemas que surgem não são causados por um fator isolado. São o resultado de interações entre vários fatores, incluindo a origem e tipo de matéria-prima e configurações dos parâmetros de operação.

Na indústria de produção de farinhas e óleos vegetais, a concentração de solvente comercial no produto acabado é uma característica de qualidade muito importante. Independentemente do produto adquirido pelo cliente, a concentração de solvente é considerada como sendo uma especificação de qualidade, pelo que o desvio dessa característica do respetivo valor tabelado conduz ao não cumprimento das normais internacionais de qualidade e segurança. Uma alteração desta medida representa um problema com consideráveis implicações, pelo facto poder levar à aplicação de sanções relativas ao incumprimento das normas, ou a uma perda de lucro da empresa, pelas perdas de solvente no produto, que se pode refletir em desperdício.

A presente dissertação enquadra-se num projeto de otimização industrial desenvolvido numa empresa industrial de produção de farinhas e óleos vegetais, enquadrado numa opção estratégica de crescimento e expansão através da melhoria operacional com base na transformação digital, enquadrada com as tecnologias emergentes da Indústria 4.0. Um dos desafios considerados foi alcançar a estabilização do processo de extração de solvente, através do desenvolvimento de modelos preditivos que permitam medir em tempo real a concentração de solvente do produto em circulação, podendo ser tomadas ações que garantam a minimização da variabilidade e redução do excesso de solvente do produto. Para compreender a influência das variáveis do processo, é utilizada uma abordagem normalizada e iterativa de *Data Mining*. Este estudo focou-se no equipamento de dessolventização visto ser um equipamento monitorizado e para o qual existe captura de dados, sendo um equipamento possível de produtizar modelos no futuro de forma rápida e escalável. Posteriormente, servirá de base aos outros equipamentos e processos industriais das unidades fabris da empresa.

Com perspetiva de aplicação futura, foram construídos modelos preditivos que estimam a concentração de hexano do produto em circulação em tempo real e também modelos que indicam um intervalo de valores ótimos para cada parâmetro operacional, de acordo com um intervalo de concentração de solvente que se pretende alcançar à saída do processo, tornando possível, numa fase futura de operacionalização, a automatização de alertas com recomendação de ações a efetuar no processo. Projetou-se ainda uma arquitetura baseada em tecnologias SAP, para a implementação das soluções desenvolvidas.

Os resultados obtidos comprovam a eficácia da implementação de modelos preditivos para a otimização do processo de extração de solvente, constituindo uma alavancagem na criação de

valor, permitindo monitorizar em tempo real a concentração de solvente à saída do processo, automatização de mecanismos de alerta e antecipação de problemas no processo, levando a uma melhoria dos índices de qualidade e segurança e uma melhor gestão de informação que suporta a tomada de decisões. A implementação dos modelos com base na arquitetura apresenta um enorme potencial de melhoria de eficiência do processo e permite a escalabilidade da solução nos restantes processos da unidade fabril com uma gestão centralizada dos processos.

Abstract

In a high volume food production environment, it is particularly important to ensure that the products marketed meet the quality and safety standards. In this sense, their monitoring is crucial to ensure the desired quality. Often, adjustments and changes to the process are required in order for the product to be suitable for storage, handling, and subsequent consumption, in line with the quality requirements. In most cases, the problems that arise are not caused by an isolated factor. They are the result of interactions between several factors, including the origin and type of raw material and configurations of the operating parameters.

In the industry of flour and vegetable oils, the concentration of commercial solvent in the finished product is an essential quality characteristic. Regardless of the product purchased by the customer, the solvent concentration is considered to be a quality specification. Therefore, the deviation from this characteristic of the respective tabulated value leads to non-compliance with international quality and safety standards. Any change to this measure presents a problem with considerable implications because it could result in the imposition of penalties for non-compliance with the rules or a loss of profit for the company due to solvent losses in the product, which can be reflected in waste.

This dissertation is part of an industrial optimization project developed in an industrial flour and vegetable oils production company, part of a strategic option for growth and expansion through operational improvement based on digital transformation, in line with emerging technologies of Industry 4.0. One of the challenges considered was to achieve the stabilization of the solvent extraction process, through the development of predictive models that allow to measure in real-time the solvent concentration of the product in circulation, being possible to take actions that guarantee the minimization of the variability and reduction of the solvent excess in the product. To understand the influence of process variables, a standardized and iterative approach to Data Mining is used. This study is focused on the desolventization equipment since it is a monitored equipment for which there is data capture, and an equipment that is likely to be used to produce future models in a fast and scalable way. Subsequently, it will serve as a base for other equipment and industrial processes of the company's plants.

With a view to a future application, predictive models were constructed which estimate the hexane concentration of the product in circulation in real time. Also, models indicate a range of optimal values for each operational parameter, according to a range of solvent concentration that is intended to reach at the exit of the process. In a future phase of operationalization, it makes possible the automation of alerts with a recommendation of actions to be carried out in the process. An architecture based on SAP technologies was also designed to implement the solutions developed.

The results obtained demonstrate the effectiveness of the implementation of predictive models for the optimization of the solvent extraction process, constituting a leverage in the creation of value, allowing real-time monitoring of solvent concentration at the exit of the process, automation of warnings and anticipation mechanisms of problems in the process, leading to an improvement

in quality and safety indexes and better information management that supports decision making. The implementation of the architecture-based models presents massive potential for improving process efficiency and allows the solution to be scalable in the remaining processes of the plant with centralized process management.

Acknowledgments

To Professor Pedro João, from Faculty of Engineering of the University of Porto, for all guidance, and patience. I am grateful for all the teachings that made it possible to carry out this dissertation with confidence, security, and motivation.

To my Deloitte colleagues, especially to Raquel Oliveira and André Lopes, for the quick integration, demand, and support. A great example of competence, determination, and passion that I will keep with me. The success of this dissertation is due in large part to them.

To all my colleagues and professors of the Faculty of Engineering of the University of Porto, who accompanied me in academic life. I am particularly grateful to Professors Américo Azevedo, José Faria and Henriqueta Nóvoa. Although they were not active people in the elaboration of this dissertation, they inspired and guided me on the happy and successful journey that I am today.

To my parents and my brothers, for always believing in me, and for supporting me unconditionally in all the decisions of my life. The best I could ever have had.

Some families are not born under the same roof, a family is not only a mother, son, grandfather, and grandson. They are a support to the weight that I carry, they are a way of breaking the rules. When everyone says go, she says come, when everyone says no, she says yes. They remind me of who I am, who I can become, where I come from, and where I want to live. I want to thank all my friends who, in one way or another, have inspired me to be the person that I am very proud of today.

The definition of success is relative. It depends on the perspective of each person. For me, success is continually improving, it is a person with values and principles that are better defined and rooted with each passing day. To succeed is to be able to share this improvement, this growth, these values and beliefs with the people for whom I have a special affection. It is to see them grow every day, to see them become human beings with a value impossible to quantify. Above all, it is to be able to grow with them. Thank you, my friends.

Gonçalo de Soares Lemos

"Tudo é possível!
Quando temos noção da dimensão do nível,
O complicado torna-se mais acessível."

Mundo Segundo

Contents

List of Figures

List of Tables

Abbreviations and Symbols

AI	Artificial Intelligence
API	Application Programming Interfaces
AWS	Amazon Web Services
CPS	Cyber-Physical System
CRISP-DM	Cross Industry Standard Process for Data Mining
CV	Cross Validation
DC	Dryer-Cooler
DM	Data Mining
DMCi	Digital Manufacturing Cloud for Insights
DSN	Digital Supply Network
DT	Desolventizer-Toaster
DT/DC	Desolventizer-Toaster-Dryer-Cooler
DTTL	Deloitte Touche Tohmatsu Limited
ERP	Enterprise Resource Planning
FEUP	Faculty of Engineering of the University of Porto
IBM	International Business Machines Corporation
INE	*Instituto Nacional de Estatística*
IoT	Internet of Things
kNN	k-Nearest-Neighbours
KPI	Key Performance Indicator
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MII	Manufacturing Integration and Intelligence
PCA	Principal Component Analysis
PCO	Plant Connectivity
RMSE	Root Mean Square Error
SCP	SAP Cloud Platform
SSE	Sum Square Error
SVM	Support-Vector-Machine

Chapter 1

Introduction

This study is part of a curricular dissertation project [1] in a business environment. The intervention proposal is part of the operational improvement of a company [2] in the agrofood industry, whose main activity is the production of flour and vegetable oils.

In this first chapter, the targeted project is presented and contextualized, describing the sector and company under study and clarifying the objectives and areas of research that are intended to be answered. Finally, the approaching methodology adopted as well as the structure adopted to present this dissertation is explained.

1.1 Project Framework and Motivation

The agrofood industry is characterized by a high governmental intervention regarding the level of quality control of the products marketed. The industrial production of flour and vegetable oils considers the monitoring of a set of variables/attributes, process, and product, to ensure the legislated quality specifications [9]. The commercial solvent concentration of the final product is one such critical point.

From a marked point of view, it has become relevant to analyze how the company subject to study under this dissertation can continue to grow sustainably, improving procedures, and work routines every day. Faced with the exponential growth of emerging Industry 4.0 technologies, a digital transformation strategy is outlined, leading the company to increase its quality indexes. The excess of the commercial solvent of the extracted flours was demarcated as being critical. This problem is a representative cost component, so a substantial rise in the monetary loss is expected for the company if the problem is not addressed promptly. It translates directly into a loss of competitiveness, and the company is still affected by the instability of the solvent extraction process.

The present project arises in this scope and environment, and carries out a detailed analysis of the process and defined a standard approaching methodology for Data Mining (DM) problems.

[1] Part of the 5th year of the integrated study cycle leading to the Master's Degree in Electrical Engineering and Computers from the Faculty of Engineering of the University of Porto (FEUP).

[2] For confidentiality reasons, the name of the company in question is not mentioned.

1.2 Deloitte Presentation

Deloitte Touche Tohmatsu Limited (DTTL), commonly known as Deloitte, is a multinational professional services network, founded in 1845 by William Welch Deloitte in London, United Kingdom. Deloitte is one of the Big Four accounting organizations and the largest professional services network in the world by revenue and number of professionals.

Deloitte provides audit, tax, consulting, enterprise risk, and financial advisory services with more than 286.200 professionals worldwide by the end of 2018. In the fiscal year of 2018, the network earned a record of $43.2 billion in aggregate revenues. As of 2017, Deloitte is the 4th largest privately owned company in the United States [10].

One of the services Deloitte member firms offer is in consulting. It consists of assisting clients by providing services in the areas of Enterprise Applications, Technology Integration, Strategy and Operations, Human Capital, and Short-term Outsourcing.

This curricular dissertation project integrates the EA team in the Business Intelligence segment, focused on implementing information systems, such as Enterprise Resource Planning, Customer Relationship Management, Supply Chain Management, Strategic Network Optimization, Enterprise Asset Management or Manufacturing Execution System, for making more effective management decisions. These systems allow businesses to have an information-based management culture, which leads not only to the streamlining and optimization of ongoing activities but also makes the transfer to target-oriented management possible [11].

1.3 Client Presentation

The client in which this dissertation is developed is a great industrial player in the agrofood sector, operating in the oilseeds business segment. The company mainly devote its time to the production of vegetable oils. The client's permisses are strategic located at the entrance of a river in Portugal, which facilitates and enable transport by ships, consequently increasing the company logistical competitiveness.

The raw oils and seeds are obtained in several countries located in different continents, supplying mainly the Portuguese, Spanish, and African markets. Intending to act in an integrated fashion across the entire value chain of this business area, in Portugal, they supply oilseeds for planting. After harvesting, they buy seeds to produce and refine the oils.

1.4 Agrofood Sector in Portugal

The agrofood sector is characterized by a large subsector and business dispersion, constituting, as a whole, a vital part of the European and national economy. In the last few years, the industry has suffered an appreciable evolution. The food industry keeps innovating food features, not only

by adapting the products to the taste of consumers at the time but also working towards to create healthier products. Through differentiation in raw materials, the industry keeps continuously committing on quality of the finished products to make them more competitive [12].

Despite the instability and changes in the economic cycle, the agrofood sector remained solid according to the latest report released by *Crédito y Caución*, the leading operator of domestic and export credit insurance in Portugal. Its study reported that this sector should continue to grow in the coming years, registering an increase of 1.8% in 2017. In Portugal, 11.047 companies are accounted, mainly small and medium size, responsible for a volume of 15,384 million euros and about 108.041 employees. Since 2010, this number has been the highest value of the number of companies and employees for this purpose. According to SISAB, the world's largest fair of Portuguese food and beverage products, this industry occupies a prominent place in the Portuguese economy, representing 16% of the total national industry [13][14][15].

Using the data obtained by Statistics National Institute[3] (INE) 2016, Table 1.1 shows the turnover of each subsector in the food industry. The groups that stand out the most are the slaughter of animals, the preparation and preservation of meat and meat products (19,4%) and the manufacture of bakery products and other flour-based products (14,2%). The curricular dissertation is inserted in the production of oils, animal, and vegetable fats (9,8%) context.

Table 1.1: Turnover distribution by subsector (CAE) in the food industry (adapted from INE [8])

Subsector	**%**
101 - Slaughter of animals, preparation and preservation of meat and meat products	19,4
107 - Manufacture of bakery products and other flour based products	14,2
109 - Manufacture of animal feed	11,5
105 - Dairy industry	11,3
102 - Processing and preserving of fish, crustaceans and molluscs	9,9
104 - Production of oils, animal and vegetable fats	**9,8**
103 - Processing and preserving of fruits and horticultural products	7,2
1083 - Coffee and tea industry	5,9
106 - Processing of cereals and legumes; fabrication of starches and similar products	5,0

At the export level, this sector contributes to the internationalization of the Portuguese economy. According to the Office of Political Planning and General Administration (GPP) of the Ministry of Agriculture and the Sea[4], the average export growth of the agrofood complex was 7.9%, a rate higher than the average annual growth of exports of goods (3.4%). Over the years, as Figure 1.1 illustrates, these data have undergone positive developments, reflecting the importance of the agrofood sector in the Portuguese economy, with growth potential.

[3] *Instituto Nacional de Estatística (INE).*
[4] *Gabinete de Planeamento Político e Administração Geral (GPP) do Ministério da Agricultura e do Mar.*

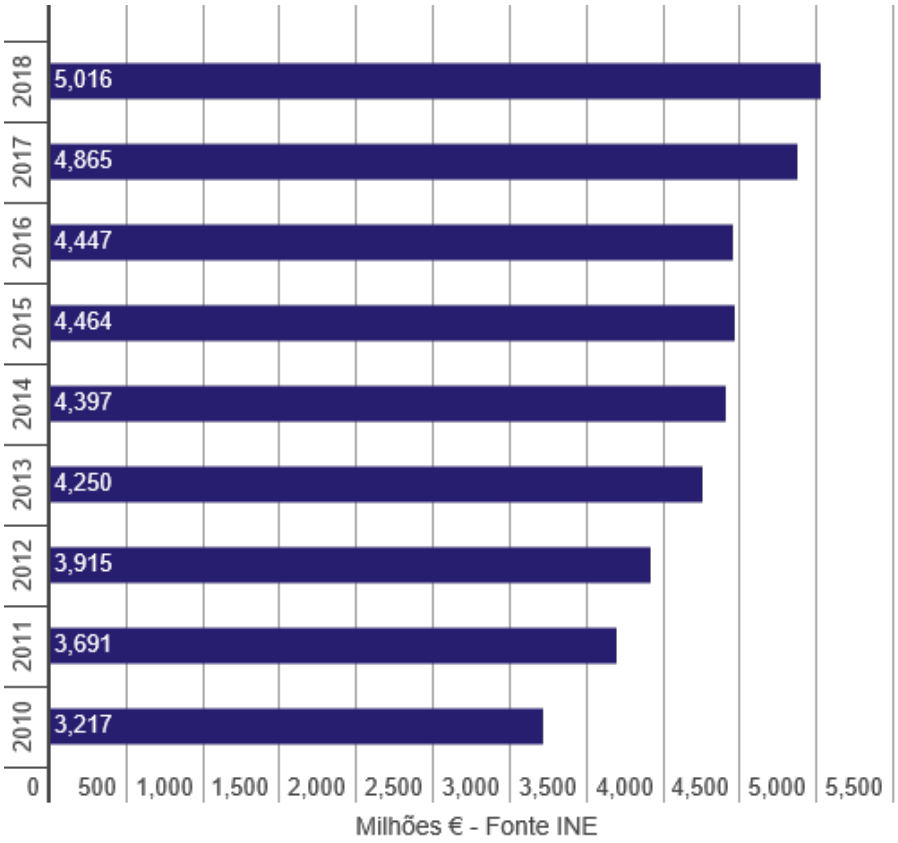

Figure 1.1: Exports evolution 2010-2018 [1]

The agrofood industry in Portugal is an important business area to support and maintain the Portuguese economic growth, with a growing tendency during the next years [16].

1.5 Objectives

Previous work was done with the client to understand the challenges that the business was facing, including an analysis of the the data available for the pilot project were analyzed. In addition, it was also carried out a diagnosis in order to choose the productive process representative of the present study. The desolventization equipment (DT/DC) was elected to be the object of this study for the following reasons:

- It is an equipment where there is a recognized challenge for the business, due to the high levels of hexane at the exit of the equipment;
- There is available real-time captured data of the operating parameters;
- It is a monitored equipment, therefore it is possible to produce the models in the future, quickly and scalable to the other equipment of the different plants of the company.

Predictive models are developed based on the data available with the following main objectives:

- To understand which variables better explain the solvent concentration (hexane) of the product at the output of the equipment;
- To define optimum operating values to meet the legislated quality and safety standards, i.e., below $500 ppm$.

It leads to an improvement in the quality and safety indexes of the final product and real-time monitoring of the solvent concentration of the final product.

1.6 Approaching Methodology

The curricular dissertation project followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. It is a standard process model that describes conventional approaches used by data mining experts, defined by International Business Machines Corporation (IBM), a remarkable American multinational information technology company. Its objective is to define a methodology to address Data Mining problems in industrial projects.

CRISP-DM provides an overview of the life cycle of the project in the scope of Data Mining. It illustrates and describes the phases of a project, their respective tasks, and their relationships. The life cycle of a Data Mining project consists of six phases, as shown in Figure 1.2. The sequence of phases is not fixed. An iteration between each phase is always required. Their outputs determine which phase, or a particular task, shall to be done next. The arrows indicate the dependencies between phases. The outer circle symbolizes the cycle of a Data Mining project.

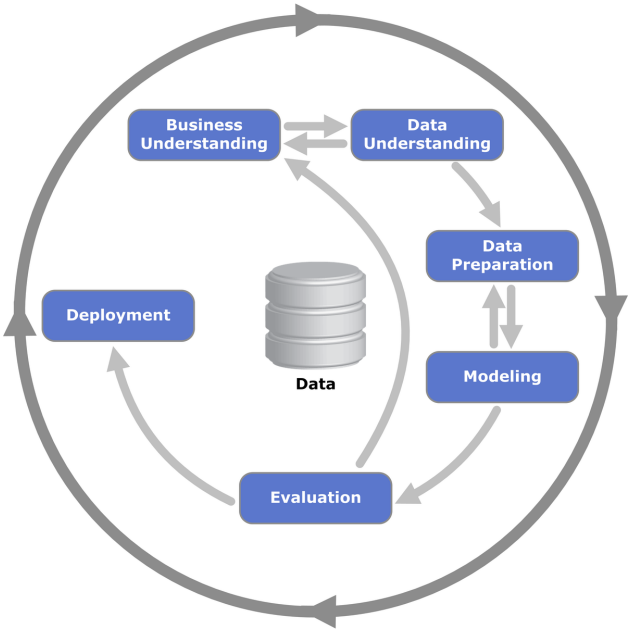

Figure 1.2: Phases of the CRISP-DM reference model [2]

The methodology is divided into six phases. All are important in this process, so it is necessary to describe each of the phases, what is their purpose and outputs:

1. **Business understanding:** this initial phase is focused on understanding the business context, identifying the project objectives and requirements from a business perspective, and defining the Data Mining problem as well as the plan to meet the objectives;

    - *Background:* it provides a basic overview of the project context;
    - *Business objectives:* it describes the goals of the project in business terms. It should also list the objectives that were considered but rejected and the rationale of selection;
    - *Terminology:* it identifies and describes the business terms, in order to allow people to become familiar with the problems addressed by the project.

2. **Data understanding:** the data understanding phase starts with initial data collection and proceeds with activities that enable us to become familiar with the data, identify quality problems, discover first insights into the data, and detect interesting subsets to form hypotheses;

    - *Data description:* it presents each dataset in detail, including a list of tables, description of each field, units, and size of the datasets;
    - *Data exploration:* it analysis the data exploration and its results, covering patterns found, expected or unexpected, and conclusions for data transformation, data cleaning, and any other pre-processing.

3. **Data preparation:** the data preparation phase covers all activities required to construct the final dataset from the initial data. Data preparation tasks are performed multiple times and not in any prescribed order. Tasks include the selection of tables and variables, as well as data cleaning and transformation for modeling tools;

    - *Dataset description:* it covers the rationale of inclusion/exclusion of attributes, actions that were necessary to address data quality issues, and a detailed description of resulted datasets.

4. **Modeling:** in this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. There are several techniques for the same DM problem. Some techniques depend specifically on the form of data. Therefore, going back to the data preparation phase is often necessary;

    - *Select modeling technique:* it selects the modeling technique more appropriate for the study;
    - *Test design:* it presents the procedures to build, test and evaluate the models;
    - *Model building:* it runs the modeling technique on each dataset to build the models;

- *Model assessment:* it put forward the results of testing the models according to the test design.

5. **Evaluation:** at this stage in the project, we have built a model that has high quality from a data analysis perspective. Before proceeding to the deployment of the model, it is important to evaluate and review the steps performed to create it, in order to be sure that the model properly achieves the business objectives. At the end of this phase, a decision on the use of the DM results should be reached;

   - *Assessment of Data Mining results:* it compares the DM results with the business objectives;
   - *List of possible actions:* it makes recommendations regarding the next steps in the project.

6. **Deployment:** it consists of the integration of the project solutions on the shop floor, training the client in order to understand how to use the model and its results properly.

   - *Deployment plan:* it laids down the deployment of the DM results;
   - *Monitoring and maintenance plan:* it identifies how the deployed results are to be maintained;
   - *Final report:* it summarizes the project and its results [2].

Due to time constraints, it was not possible to enter the deployment phase.

1.7 Dissertation Structure

This dissertation is developed over seven chapters, that are design and prepared to provide a more detailed explanation of the whole context and the methodology used for the project.

In this chapter, an introduction to the project, its objectives, and the approach methodology adopted is carried out. It is also made a brief contextualization of Deloitte, the company for which this dissertation was carried out, the agrofood sector in Portugal and the company under study.

Chapter 2 deals with the bibliographic analysis, with a contextualization of the four Industrial Revolutions and respective technologies that emerged, as well as a theoretical framework of the concepts and methodologies that served as support in the development of the entire dissertation.

Chapter 3 focuses on understanding the business, describing the high level of the complete solvent extraction process and a detailed analysis of the DT/DC equipment.

Chapter 4 moves forward analyzing and exploring the data provided for familiarization and understanding. We also report some questions and concerns that came up during this analysis.

Chapter 5 covers all the activities required to prepare the data for the next phase of model building.

Chapter 6 presents the solutions proposed for solving the problem, as well as the main results and their evaluation in business terms.

Finally, Chapter 7, identifies the main conclusions and suggestions for improvement and implementation in a future perspective.

Chapter 2

State of the Art

This chapter revisits the evolution of technology and the theoretical foundations that served as a basis for the development of this Data Mining project.

2.1 Prior Industrial Revolutions

Before the Industrial Revolution, people lived in an environment where goods and materials were handmade. Late in the 18th century, there was the 1st Industrial Revolution, characterized by power generation and transition to new manufacturing processes, optimizing traditional objectives, like cost, quality, and innovation. It was the most quick and abrupt in history: everything changed. Around 1740, there was a transition from a small scale handmade domestic manufacture to a mass production trend, by the industrialization of the textile industry and significant development of mining to find coal. The first assembly line was set up by Josiah Wedgwood for the mass production of fine pottery. His establishment divided the production into segments, separating the potters according to their assigned tasks and each focused on one aspect of the making of the final product. The separation into specific repetitive tasks is the model for mass production for this Industrial Revolution, which is extremely efficient, allowing a production of a mass number of consumer goods, which meant mass profits for the owners. One of the significant developments of the 1st Industrial Revolution was in transport, with the widespread introduction of inexpensive iron, the rolling mill for making rails, and the development of the high-pressure steam engine, which made possible the existence of steam locomotives.

The 2nd Industrial Revolution, in the early 20th century, was characterized by the industrialization. The electricity and assembly lines made mass manufacturing possible and improved infrastructures. Its specific events can be traced to innovations in manufacturing, with the establishment of an industry based on machine tools, the development of methods for manufacturing interchangeable parts, and the invention of the Bessemer Process to produce steel [17]. Mass production was one of the milestones of this Revolution era. It started with the Bessemer processes, which was the first inexpensive industrial process for the mass production of steel. This

process revolutionized steel manufacture by decreasing its cost, along with significantly increasing the scale and speed of production to this vital raw material, as well as decreasing the labor requirements for steel-making [18].

Late in the 20th century, electronic automation was a key factor for the 3rd Industrial Revolution, also known as the Digital Revolution. It is the shift from mechanical and analog electronic technology to digital electronics, which began with the proliferation of digital computers and digital record keeping that continues to the present day. It also refers to the sweeping changes brought by digital computing and communication technology, which marked the beginning of the information age. The mass production of digital logic circuits and its derived technologies, including the computer, digital cellular phone, and the internet are central topics in this Revolution. Computers and internet connectivity provide better management with access to information and enhanced decision-making capability [19].

At the present stage, we are in the middle of the 4th Industrial Revolution. The convergence of the new technologies, such as Cyber-Physical systems (CPS), Cloud Computing and Internet of Things (IoT) make digital transformation possible, connecting products, customers and the entire supply chain, providing even better management through better visibility, more variability, velocity and volume [20].

In Figure 2.1 is presented a summary of the Industrial Revolutions.

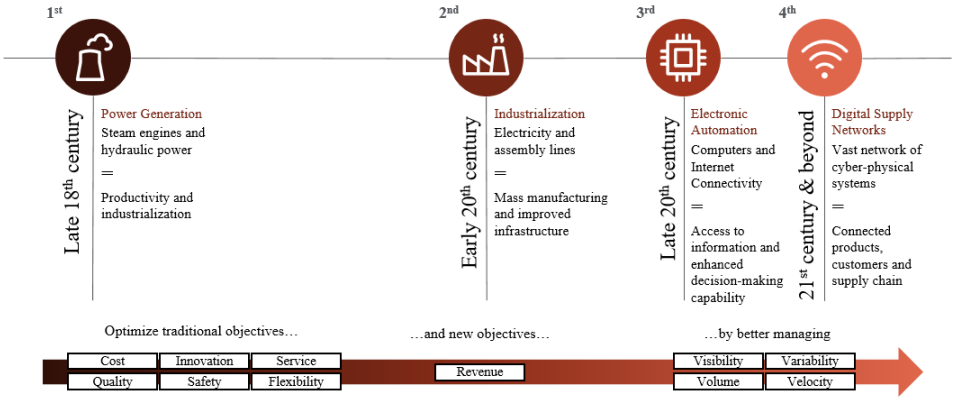

Figure 2.1: Industrial revolutions summary

The term "Revolution" has been employed to describe the aforementioned process, since the changes they caused had a significant and transforming impact, directly affecting how people worked and lived.

2.2 4th Industrial Revolution

After the three major Industrial Revolutions, XXI century is the stage for the 4th Industrial Revolution, commonly known as Industry 4.0. It is characterized by a convergence of technologies that is

blurring the lines between physical and digital fields, collectively referred to as cyber-physical systems. It is marked by emerging technology breakthroughs in several fields, including robotics, artificial intelligence, nanotechnology, quantum computing, biotechnology, IoT, decentralized consensus, fifth-generation wireless technologies (5G), additive manufacturing/3D printing and fully automated vehicles. The 4th Industrial Revolution goes far beyond manufacturing. Connected and smart technologies are transforming how products and services are design, made, used and maintained, transforming organizations, linking each one into a digital reality, transforming the roots of the business in which they operate.

The Digital Transformation era is changing the way business operates and, by extension, the stakes by which they are forced to compete. Organizations must decide how and where they are capitalizing the investments in the new technologies, analyzing and identifying which ones are the best to meet their needs. If they do not fully understand and act through these changes and opportunities, companies risk losing their ground. Business leaders used to traditional linear data and communications will, fundamentally, change the way they lead their businesses with the shift to real-time access to data and intelligence enabled by the new technologies. The integration of many sources and locations of digital information will drive the act of doing business, in an ongoing cycle [20].

Throughout this cycle, there is a continuous and cyclical flow of data and actions between the digital and physical worlds. To achieve this process, appropriate digital new technologies are combined with physical systems, enhancing the success of this process.

What constitutes the real value of Industry 4.0 is the ability to act upon data that has been analyzed, namely the digital-to-physical process. Although Industry 4.0 extends far beyond these limits. Instead of focusing only on manufacturing and production, it mainly focuses on the entire ecosystem - partners, suppliers, customers, workers, and operational considerations.

In short, Industry 4.0 is more than just advance in technologies: it is the manner those technologies are capitalized and converged together, how organizations use to invest and use them to drive operations and grow. In accordance with this perspective, companies should be taking an holistic view of the 4th Industrial Revolution and the ways it changes business. Changes are happening fast, the fastest ever in history, and it is challenging to keep up [21].

Industry 4.0 matters because it can transform an organization's processes and operations, and it is essential to understand why it touches all of us. While it has grown to involve business operations, the workforce, and society, it is rooted in supply chain and manufacturing, which is the core of the personal and business world. The digitization of the whole supply chain enables companies to combine learnings to make better, more holistic decisions.

There are great opportunities provided from fully connected processes. Before Industry 4.0, the processes where linear and operated reactively. Now, companies have the opportunity to digitize and connect all processes, learn to see what is happening in all the value chain, and adjust accordingly in real-time. This leads to smarter decisions, better-designed products and services, more efficient resources, and more exceptional ability to predict future needs.

This digital thread links the entire design and production processes with a seamless strand of data that extends from the initial design concept, the production of the finished part or product and deliver to the final customer. It enables organizations to gain insight into their working systems and facilities, simulate possible scenarios, and understand the impacts of changes, in particular segments of the value chain, on the all network.

The digitization is only possible with the implementation of certain relevant technologies, that work with each other [22].

2.2.1 Internet of Things

Some factors are creating somewhat a perfect world for IoT to penetrate. The internet is becoming widely available, the cost of connecting is decreasing, more devices are created with Wi-Fi capabilities and sensors built in them, technology costs are going down and smartphone penetration sky-rocketed [20].

The IoT concept consists of basically connecting any device to the internet and to each other. It includes everything from cellphones, coffee makers, washing machines, headphones, wearable devices, and almost everything we can think. It also applies to components of machines. The IoT is a vast network of connected "things", which also includes people, and is a relationship between people-people, people-things, and things-things.

The IoT can be applied to a variety of things like mobile, wearable devices, and transportation networks. Smart Cities are prepared to help reducing waste and improve efficiency for things such as energy use as well to improve how we work and live. The reality is that the IoT opes the door for endless opportunities and connections to take place. However, it also raise many challenges, like security, which is a big issue that is often discussed all over the world because of the privacy and data sharing policies adopted by companies. On the other hand, with this technology, companies face massive amounts of data that these devices produce. It is extremely important to figure out a way to store, track, analyze and make sense of this vast amount of data that will be generated, in order to effectively use it in their favor, while acting upon them in real-time [23].

2.2.2 Cloud Computing

Cloud Computing makes computer systems resources available on demand without active management by the user. It usually describes data centers available to many users on the internet. Today, there are large clouds which have functions distributed over multiple locations from central servers. There are enterprise clouds that are limited to a single organization, public clouds available to many organizations or even hybrid cloud, which is a combination of both. The availability of high-capacity networks, low-cost computers and storage devices as well as the widespread adoption of hardware virtualization, service-oriented architecture, autonomic and utility computing has led to growth in cloud computing [20].

The array of available cloud computing is vast, but most fall into one of the following categories:

- **Software as a Service (SaaS):** it provides applications delivered through the browser. One of the most popular SaaS applications is Google's G Suite and Microsoft's Office 365. The majority of enterprise applications, including Oracle and SAP, have adopted the SaaS model. These applications offer extensive configuration options as well as development environments that enable customers to code their modifications and additions;

- **Infrastructure as a Service (IaaS):** it provides storage and computational services on a pay-per-use basis. However, the full array of services offered by all major public cloud providers is huge - scalable databases, developer tools, big data analytics, virtual private networks, machine learning, and application monitoring;

- **Platform as a Service (PaaS):** it provides services and workflows targetting developers, who use shared tools, processes, and APIs to develop, test, and deploy applications. PaaS assures that developers have access to resources, follow the processes, and use only specific services, while operators maintain the infrastructure;

- **Functions as a Service (FaaS):** it delivers serverless computing, adding another layer of abstraction to PaaS. Instead of futzing with virtual servers, containers, and application runtimes, they upload functional blocks of code and set them to trigger given a certain event. A benefit of FaaS applications is that they consume no IaaS resources until an event occurs, reducing pay-per-use fees [24].

Cloud Computing technologies have tremendous benefits, especially sided with other Industry 4.0 technologies, reducing time to market of applications that need to be scaled dynamically. Increasingly, developers are drawn to the cloud by the high variety of advanced services that can be applications can incorporate, from machine learning to IoT connection [20].

2.2.3 ERP Systems

The Enterprise Resource Planning (ERP) is the integrated management of core business processes, often in real-time and mediated by software and technology, used by organizations to collect, store, manage and interpret data from their business activities. It provides a continuously updated view of core business processes using common databases, tracking business resources, such as cash, raw materials, production capacity and the status of business commitments, like orders, purchase orders and payroll.

An ERP system covers functional areas, grouped as ERP modules:

- Finance and accounting;

- Management accounting;

- Human resources;

- Manufacturing;

- Order processing;
- Supply Chain management;
- Project management;
- Customer relationship management;
- Data services.

Implementation of ERP's usually implies significant changes to staff work processes and practices, generally with the help of three types of services - consulting, customization and support - since it requires changes in existing business processes. Poor understanding of needed processes changes before starting the implementation is the main reason for project failure. It is therefore crucial that organizations thoroughly analyze processes before they implement an ERP software, and it is more challenging to implement in decentralized organizations since they often have different processes, data semantics, authorization hierarchies, and decision centers.

An advantage of ERP is that the integration of multiple business processes saves time and expense. Management can make decisions faster and with fewer errors. Data becomes visible across the organization, which benefits multiple tasks, such as sales forecasting, it allows inventory optimization, order tracking from acceptance through fulfillment, and centralizes business data.

The main benefits that companies gain by implementing an ERP system are flexibility and speed when reacting to changes in business processes or on the organization level[25].

2.2.4 Blockchain and Cryptography

Blockchain is a list of transaction records distributed over a network sequentially and permanently. Each block contains a hash of the previous block, along with a timestamp and transaction data. It makes the blockchain inherently resistant to attack or manipulation. It is ideal for recording various types of transactions where data is sensitive or targeted by hackers for unauthorized duplication or other fraudulent activity. Blockchain can be used in business applications like recording contracts, medical records, monetary transactions, and much more.

Cryptography studies the practice of securing private messages so that the intended parties can only read them. It involves encrypting and decrypting information through complex mathematics. It is a core of blockchain technology.

Blockchain technology uses cryptography to protect the identity of users, ensuring transactions are done safely and securing all information and storage of value. Therefore, anyone using blockchain can have complete confidence that once something is recorded on it, it is done so legitimately and in a manner that preserves security [26].

Through blockchains, companies gain a real-time digital ledger of transactions and movements for all participants in their supply chain network, gaining better visibility into procurement, more accurate and reliable data for analytics, and increased trust among all participants. Its underlying logic and processes force data to become synchronized. The blockchain mainly functions as a

layer of supplementing the existing ERP software. If done correctly, blockchain installation slots into the workflow without disruption [20].

2.2.5 Cyber-Physical Systems

A cyber-physical system (CPS) refers to the tight conjoining and coordination between computational and physical resources, integrated with the internet and its users. The physical and software components of a system operate on different spatial and temporal scales, exhibiting multiple and distinct behavioral modalities, interacting with each other in many different ways, depending on the context. Some examples of CPS applications include communication, energy, infrastructure, health care, manufacturing, physical security, robotics, smart grid, transportation, and many others.

Unlike more traditional embedded systems, a fully developed CPS is designed as a network of elements that interact with physical input and output instead of as standalone devices. They are closely tied to concepts of robotics and sensor networks with intelligence mechanisms proper of computational intelligence. The link between computational and physical elements increases the adaptability, efficiency, functionality, reliability, safety, and usability of cyber-physical systems [27].

A challenge in the development of CPS is a large number of differences in the design practice between the various engineering disciplines involved. Additionally, there is no "language" in terms of design practice that is common to all the involved disciplines in CPS. In a marketplace innovating exponentially, it is assumed to be essential for engineers from all fields the need to be able to explore system designs as a team, allocating responsibilities to software implemented in physical elements, and analyzing trade-offs between them [28].

CPS provides the necessary technological basis to facilitate the realization and corresponding automation of large-scale complex systems, such as smart grids, smart buildings, smart transportation, among other application areas. The CPS era requires solutions that will support it at the device, system, infrastructure, and application level. It includes the whole lifecycle from the cradle-to-grave of its CPS components and services. It is a scientific, technical, industrial and social challenge that includes a multi-disciplinary engineering approach and the confluence and sometimes a fusion of heterogeneous communication, information and control/automation technologies [20].

2.2.6 Digital Supply Chain

A supply chain is a network of individuals, activities, and technology involved in production and sale of a product, from delivery of raw materials from the supplier to the manufacturer to its delivery to the end user. Supply chain activities involve a vast number of activities, from the transformation of natural resources, raw materials, and components into a finished product to its delivery to the end customer. In other words, supply chain management includes design, planning implementation, control and monitoring of supply activities to create net value, building a

competitive infrastructure, leveraging logistics, synchronizing supply with demand, and measuring performance worldwide.

The supply chain digitization enables companies to address the customers new requirements and the remaining expectations in efficiency improvement. The power and efficiency of Industry 4.0 technologies manifested in significantly reduced transaction costs for business operations, allowing to gain insight into each little step of operations in the chain, and deeply understanding customer and supplier demand patterns. That is why companies need to redesign their supply chains from a traditional linear approach to a transparent, flexible, and interconnect digital supply network, as shown in Figure 2.2.

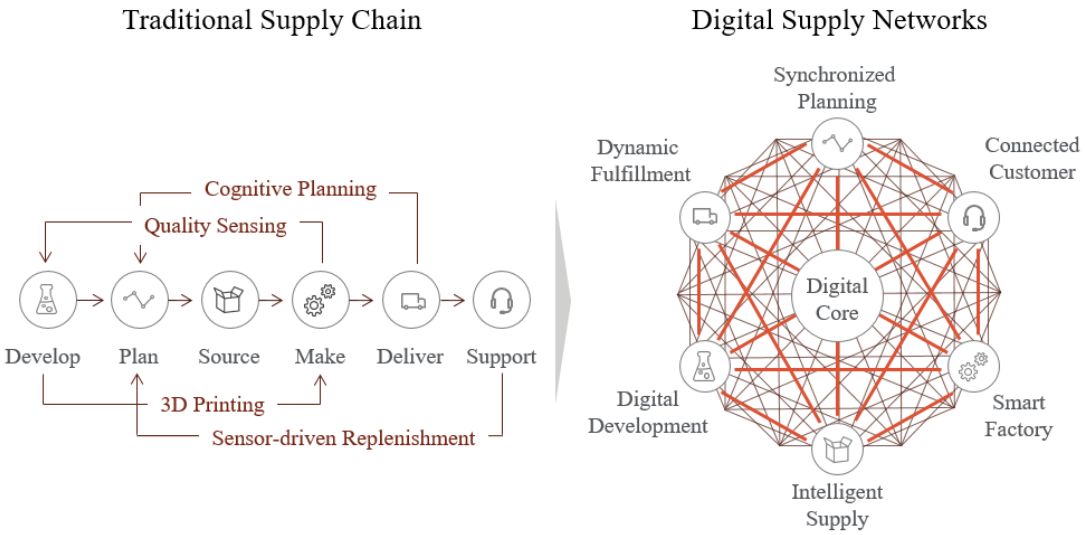

Figure 2.2: Shift from traditional supply chain to digital supply network [3]

The dynamic and integrated supply network overcome the delayed action-reaction process of the linear supply chain, based on better-informed decisions driven by the availability of real-time data, also providing greater transparency, enabling enhanced collaboration across the entire supply network. In Figure 2.2, the connection between every cluster of the chain is visible. The interactions from every point of the network to each node are endless, connecting all areas that were not connected previously [3].

2.3 Data Mining

Data Mining (DM) is the process of discovering patterns in large datasets involving methods intersecting machine learning, statistics, and database systems. The goal is to transform that raw data into a comprehensible structure for further use and facilitating the interpretation [29].

The proliferation and increasing power of computer technology have dramatically increased data collection, storage, and manipulation ability. Datasets had exponentially grown in size and

complexity, which lead to an increase of automation in data processing, aided by a variety of technologies discovered, such as cluster analysis, neural networks and support vector machines. Hidden patterns in large datasets are uncovered with the application of these methods, bridging the gap from applied statistics and artificial intelligence (AI). It provides the mathematical background, to database management by exploiting data to effectively store and index data in the databases in order to execute the discovery and learning algorithms more efficiently.

Before DM algorithms can be used, the dataset must be assembled and large, in order to that data contain the patterns and DM algorithms being able to uncover them, while remaining concise to be mined with an acceptable time limit. It is essential to analyze the multivariate datasets before DM. Then, the set is cleaned by removing the observations containing noise and those with missing data [30].

Data Mining involves six basic classes of tasks:

- **Anomaly detection:** it identifies unusual data records (outlier/change/deviation detection) or data errors, which are interesting and require further investigation;
- **Association rule learning:** it analyses dependency models that search for relationships between variables;
- **Clustering:** it discovers groups and structures in the data that are in some way similar, without the use of known structures in the data;
- **Classification:** it generalizes a known structure to apply to new data;
- **Regression:** attempts to find a function that models the data with the least error, establishing relationships among data;
- **Summarization:** it provides a more compact representation of the dataset. It includes visualization and report generation [31].

2.3.1 Classification and Regression

Classification and regression concepts represent a generality of problems in which DM is currently applied, namely through the creation of models, at the expense of a set of classified examples, to predict for each record that is part of another dataset, belonging to a particular class in classification or value in regression.

For example, an analyst from a logistics company wants to understand why some customers remain loyal to the company while others terminate their contracts. Fundamentally, the analyst intends to predict which customers the company might lose to its competitors. With this goal in mind, the analyst can build a model from the historical data of "loyal customers" and the clients who have left. A good model allows a better understanding of the company's customers, and predict which ones will stay and which will left.

Customer loyalty modeling illustrates the process of defining a study. A study specifies the scope of a DM activity, identifying its purpose and the data to be used. The definition of the

objective does not need to be in a precise way, but can only be broadly defined. At the expense of defining the problem, the DM study is initiated with the formulation of its objective. Based on the example, it is possible to state the following objective: it is intended to understand which factors potentially contribute to the permanence or abandonment of the company's customers. One objective differs significantly from the setting of a specific issue since no correlation between the factors is being assumed. For example, one could have asked the following question: among those where there has been a declining steadily in the utilization of the services in the last six months, how many have ceased to be customers? This question assumes that it exists a relation between the utilization rate and the possible abandonment by the client.

Building such a model requires knowledge about clients who remain loyal and those who have terminated their contract. This type of study is considered supervised learning since the records used in the DM study are classified in one of two perspectives (faithful or lost) [31].

Predictive modeling can be described as a mathematical problem, $y = f(X)$, approaching a mapping function (f) of the input variables (X) for the output variables (y), which is a categorical value in case of a classification problem, or a continuous value in case of a regression model.

Model building is divided into two parts. In the first part, the classification or regression model is built based on a training dataset, where the function $y = f(X)$ is obtained. In the second, the function obtained is evaluated with a test dataset. The training and test datasets are disunited, ideally built based on a large dataset, which is divided proportionally in the training and test sets. For example, we have a dataset that has 100 individuals that belong to class A and 200 that belong to class B. In an ideal world, the dataset is divided in a way so that the training dataset has 70 individuals of class A and 140 of class B, and the remaining belong to the test dataset.

When building a model, sometimes it is necessary to compare individuals with each other. One of the possible approaches is to measure the distance between each in the multidimensional space. To accurately perform the task, it is necessary to normalize the data, which are aimed at resizing the attributes to a range of values, enabling the comparison between the individuals, without any attribute influencing the result solely because it presents a scale of higher values.

There are several data normalization techniques. The difference for each one of them is how they resize the attributes to a range of values. For example, the feature scaling technique is used to bring all values into the range [0, 1], based on the minimum and maximum values of the attribute, as we can see in Equation 2.1 [32].

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.1}$$

During the models' evaluation phase, these assign a class Y to each X of the test dataset. From this result, it is necessary to compare with the real value to evaluate the performance of the developed model. The metrics used to evaluate the models are critical since those influence on how the performance of machine learning algorithms are measured and compared.

Let us say we are solving a classification problem where there are two classes, the Actual classifications and Predicted classifications, both having two classes, positive and negative. The

result of this model can be represented in a confusion matrix, as shown in Table 2.1, evaluating the model performance afterward.

Table 2.1: Confusion matrix

		Actual	
		Positive	**Negative**
Predicted	**Positive**	TP	FN
	Negative	FP	TN

The confusion matrix allows us to visualize all the possible results of the classification model, where:

- **True positives (TP):** it is the number of individuals that belong to the positive class and the model classified them as positive;
- **False positives (FP):** it is the number of individuals that belong to the negative class and the model classified them as positive;
- **False negatives (FN):** it is the number of individuals that belong to the positive class and the model classified them as negative;
- **True negatives (TN):** it is the number of individuals that belong to the negative class, and the model classified them as negative.

There are several methods to evaluate the classification model. The most common, based on the confusion matrix, is the Accuracy and Error Rate, these being the percentage of incorrect classified individuals and represented in Equation 2.2 and 2.3.

$$Accuracy = \frac{TP+TN}{(TP+FP)+(FN+TN)} \tag{2.2}$$

$$Error\ Rate = 1 - Accuracy \tag{2.3}$$

The Accuracy metric is a good measure when the target variable classes in the data are nearly balanced. A balanced set is one who has not a significant predominance of one class relatively to the other.

In addition to the techniques mentioned, other metrics can be used to evaluate the constructed classification model, such as Precision and Recall, shown in Equations 2.4 and 2.5, respectively.

$$Precision = \frac{TP}{TP+FP} \tag{2.4}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.5}$$

Precision is the percentage of instances that have been classified as positive, and that belongs to the positive class. On the other hand, Recall is the percentage of instances of positive classes that were well classified [33][34].

In respect to regression, the purpose is to assess how far the forecast has been to the real value. This can be done through several metrics that evaluate the performance of regression models: Sum Square Error (SSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), as shown in Equations 2.6, 2.7, 2.8 and 2.9, respectively.

$$SSE = \sum_{i=1}^{n} (f(x_i) - y_i)^2 \tag{2.6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (f(x_i) - y_i)^2}{n}} \tag{2.7}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - y_i| \tag{2.8}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{f(x_i) - y_i}{f(x_i)} \right| \tag{2.9}$$

In statistics, SSE measures how far off our model's predictions are from the observed values. A value of zero indicates that all predictions are spot on, and a non-zero value indicates errors. The RMSE is the standard deviation of the residuals (predicting errors). It tells how concentrated the data is around the line of best fit. The MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The MAPE measures how accurate a forecast system is. It measures this accuracy as a percentage [32][35][36].

These metrics are crucial to evaluate if the developed model has good results. If it does not, other techniques can also be used to improve the results. The available data may not be enough or be balanced, in case one class is predominant concerning the others, influencing the model by giving more relevance to the abundant class. It is possible to cancel this type of problem using data balancing techniques. Oversampling and undersampling in data analysis adjust the class distribution of the dataset. Both oversampling and undersampling involve biasing the data to select more observations from one class than from another, compensating the imbalance that is either exists in the data or is likely to develop if a purely random sample were taken [37].

It is fundamental to analyze the predictive capacity of the built models for unknown samples. Usually, the data is split into training and testing sets. The training set is used to train the model and the testing set to test the model. Then, the model performance is evaluated based on an error metric to determine its accuracy. However, this method is not very reliable, as the accuracy obtained varies for each distinct test set. There are several methods for the effect, including:

- **K-fold Cross Validation:** it is a technique that splits the dataset into a K number of sections/folds where each fold is used as a testing set and the others K-1 sets used as training. This approach is repeated K times, one for each section, building a model for each of the K sets, evaluating the overall accuracy by averaging the accuracy of all models [38].

- **Bagging:** it is a Machine Learning ensemble meta-algorithm designed to improve the stability and accuracy of algorithms used in statistical classification and regression. It produces several different training sets of the same size with replacement and then builds a model for each one using the same Machine Learning scheme, combining predictions by voting for a nominal target or averaging for a numeric target. It also reduces variance and helps avoid over-fitting. It is very suitable for unstable learning schemes, which means that small changes in training data can make a big change in the model [39].

2.3.2 Clustering

Clustering is the task of grouping a set of data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group (cluster) be similar, or related, to one another and different from the other clusters. The higher the homogeneity within a cluster and the more significant the difference between clusters, the better or more distinct the clustering. It is a method of unsupervised learning, a common technique for statistical data analysis. This method draws references from datasets as a part of input data without labeled responses.

This method of exploratory data analysis proceeds to the grouping of individuals based on the existing information. Associations can be found in the data that were previously not evident. The key idea of using this type of technique is that the formed clusters are interpretable and meaningful.

The selection of variables is a critical point in clustering analysis since it has a significant weight on the output results. It is necessary to, *a priori*, make a good selection of variables that will be used in the development of the models based on the knowledge gathered by the data analyst in a business perspective.

Some different methods can be used to carry out a cluster analysis, which can be classified as follows:

- **Hierarchical methods:** a hierarchical clustering method produces a classification in which small clusters of very similar individuals are nested within larger clusters of less closely-related individuals;

  - Agglomerative algorithms: it generates a classification in a bottom-up manner, in which subjects start in their separate cluster. The two most similar clusters are then combined, and this is done repeatedly until all subjects are in one cluster. In the end, the optimum number of clusters is then chosen out of all cluster solutions;

  - Divisive algorithms: it generates a classification in a top-down manner, in which all subjects start in the same cluster, and the above strategy is applied in reverse until every subject is in a separate cluster.

- **Non-hierarchical methods:** A non-hierarchical method generates a classification by partitioning a dataset, giving a set of non-overlapping groups having no hierarchical relationships between them. A systematic evaluation of all possible partitions is infeasible, and many different heuristics have been described to allow the identification of good, but possibly sub-optimal, partitions;

  - Relocation algorithms: it assign compounds to a user-defined number of seed clusters and then iteratively reassign compounds to see if better clusters result. Such methods are prone to reaching local optima rather than a global optimum, and it is generally not possible to determine when or whether the global optimum solution has been reached;

  - Nearest neighbor algorithms: it assign compounds to the same cluster as some number of their nearest neighbors. User-defined parameters determine how many nearest neighbors need to be considered, and the necessary level of similarity between nearest neighbor lists.

The selection of methods depends on the purpose of the study and the different properties of the methods. The goal is to use different methods and then evaluate and compare the one that best suits the context of the work [40].

In the hierarchical methods, agglomerative algorithms have n groups, each containing a single object. Iteratively, the objects are grouped successively until a group containing all of them is found. Regarding the divisive type, it starts with a group containing all the objects and in each iteration separates them into disjoint groups, until n groups of one object each are obtained. In this type of approach, we proceed to the successive division of the groups. Figure 2.3 illustrates the two methods.

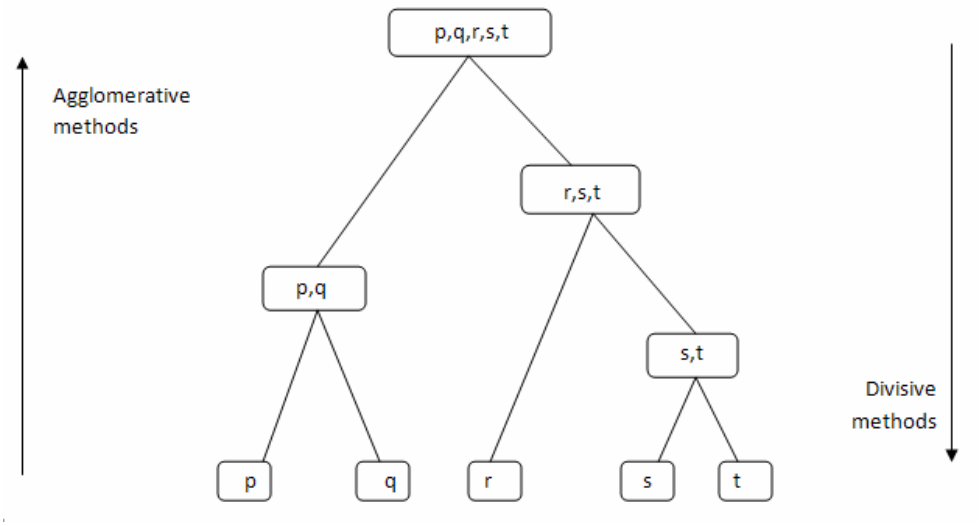

Figure 2.3: Agglomerative and divisive process [4]

Two objects are considered close when their distance is small or when they are similar. However, the great question that arises is how to proceed to aggregation between two objects. There are several approaches, and each provides a different hierarchical method:

- **Single linkage:** the distance between two clusters is defined as the minimum distance between any single data point in the first cluster and any single data point in the second cluster, as shown in Equation 2.10. Based on this definition of distance between clusters, at each stage of the process we combine the two clusters with the smallest single linkage distance;

$$d_{12} = min_{ij}\, d(X_i, Y_j) \tag{2.10}$$

- **Complete linkage:** the distance between two clusters is defined to be the maximum distance between any single data point in the first cluster and any single data point in the second cluster, as shown in Equation 2.11. Based on this definition of distance between clusters, at each stage of the process we combine the two clusters that have the smallest complete linkage distance;

$$d_{12} = max_{ij}\, d(X_i, Y_j) \tag{2.11}$$

- **Average linkage:** the distance between two clusters is defined to be the average distance between data points in the first cluster and data points in the second cluster, as shown in Equation 2.13. Based on this definition of distance between clusters, at each stage of the process we combine the two clusters that have the smallest average linkage distance;

$$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j) \tag{2.12}$$

- **Centroid method:** the distance between two clusters is the distance between the two mean vectors of the clusters, as shown in Equation 2.13. At each stage of the process we combine the two clusters that have the smallest centroid distance.

$$d_{12} = d(\bar{x}, \bar{y}) \tag{2.13}$$

Where:

$X_1, X_2, \ldots, X_k$, are the observations from cluster 1
$Y_1, Y_2, \ldots, Y_k$, are the observations from cluster 2
$d(x, y)$, is the distance between a subject with observation vector x and vector y

The hierarchical representation is usually made through a dendrogram, where it is possible to observe how the different groups of clusters were united or separated through its visualization. We can have the perception of the number of clusters that can be formed, through the height represented in the vertical axis. It should be noted that the grouping process takes into account the distances between objects [40][41].

Regarding non-hierarchical methods, we can have partition algorithms, density based algorithms, overlapping algorithms, or diffusive algorithms. The main differences are in the way the initial aggregation of individuals is made, and in the way, the distances are calculated between the centroids of the clusters and the objects.

The partition K-Means algorithm processes with the thought that new clusters should be formed according to the distance between points and center of clusters. Distance between elements of the dataset and center of clusters also gives error rate of clustering. K-Means algorithm consists of four basic steps:

1. Determination of centers;
2. Assigning points to clusters which are outside of the centers according to the distance between centers and points;
3. Calculation of new centers;
4. Repeating these steps until the decided clusters are obtained.

The second step of the K-Means algorithm involves the exclusive assignment of instances to the closest class, by calculating the Euclidean distance between two instances X and Y which are represented by n continuous attributes, as shown in Equation 2.14.

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{2.14}$$

In the third step, the algorithm uses the attribute values of the instances assigned to a cluster to recalculate the cluster's centroid. We recompute the estimates from only those instances currently assigned to the class.

The biggest problem of the K-Means algorithm is the determination of starting points. If initially a wrong choice is made, many changes will be at the clustering period and in this case for each time. Different clustering results can even be obtained with the same number of iterations. At the same time, if dimensions of data groups are different, the density of data groups will be different, or if there is contrariety in data, the algorithm may not get effective results [42].

2.3.3 Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of variables under consideration by selecting a set of critical variables. This process is divided into feature selection and feature extraction.

- **Feature selection:** it tries to find a subset of the original set of variables to get a smaller subset which can be used to model the problem. It usually involves three methods:
  - Filter;
  - Wrapper;
  - Embedded.
- **Feature extraction:** it reduces the data in a high dimensional space to a lower dimension space, i.e. a space with less number of dimensions. These are the three most common methods used:
  - Principal Component Analysis (PCA);
  - Linear Discriminant Analysis (LDA);
  - Generalized Discriminant Analysis (GDA).

Dimensionality reduction may be both linear or non-linear, depending upon the method used. The prime linear method, called PCA, is a method of selecting the most important variables, in the form of components, from a broad set of variables available in a dataset. It extracts a low dimensional set of features from a high dimensional dataset capturing the most information possible. As we have fewer variables, the more meaningful becomes its visualization.

For example, we have a dataset with a dimension of $n * p$, where $n = 300$ which represents the number of observations and $p = 50$ representing the number of predictors. Since we have a hefty p, there can be $\frac{p(p-1)}{2}$ scatter plots, more than one thousand plots possible to analyze the variable relationship.

It would be an efficient approach to select a subset of p ($p << 50$) predictor, which captures as much information — followed by plotting the observation in the resultant low dimensional space. Figure 2.4 shows the transformation of a high dimensional data (three dimensions) to low dimensional data (two dimensions) using PCA.

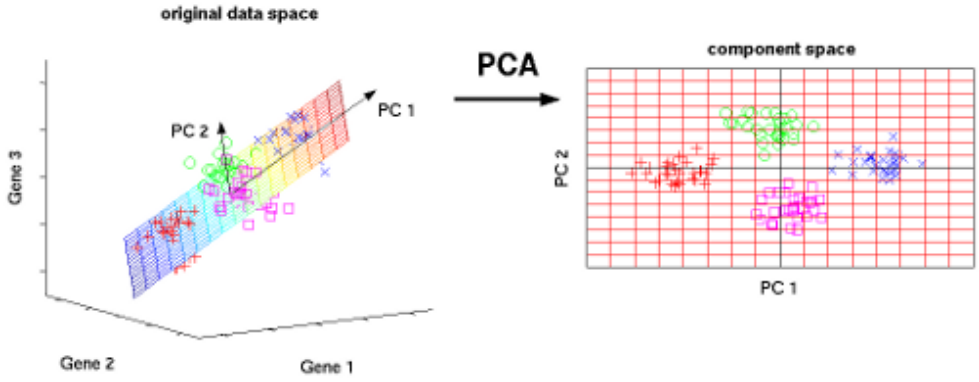

Figure 2.4: PCA transformation [5]

A principal component is defined as a linear combination of the original predictors in a dataset, represented as PC1 and PC2 in Figure 2.4. For example, we have a set of predictors as $X_1, X_2, \ldots, X_p$. The principal component can be written as Equation 2.15:

$$Z_n = A_{1n}X_1 + A_{2n}X_2 + \ldots + A_{pn}X_p \tag{2.15}$$

Where:

Z_n is the n principal component

A_{pn} is the loading vector comprising of the principal component loadings

$X_1, X_2, \ldots, X_p$ are normalized predictors

The first principal component, Z_1, is a linear combination of original predictor variables which captures the maximum variance in the dataset. The direction of higher variability in the data is determined as larger the variability captured in the first component is, the more significant is the information captured. The first principal component has a higher variability. It results in a line that is closest to the data, which results in the minimum sum of squared distance between an observation and the plot line.

We can compute the second principal component, Z_2, which captures the remaining variance in the dataset and is uncorrelated with Z_1. Their directions should be orthogonal, so the correlation between these components should be zero.

All the following principal components follow a similar concept. The remaining variation is captured without a correlation with the previous component. For $n * p$ dimensional data, there are $min(n-1, p)$ constructed principal components. Its directions are identified under an unsupervised way, so that the response variable, Y, is not used to determine the component direction [43] [44].

2.3.4 Summary of the Techniques Used

This section is intended to briefly describe all the techniques used in this dissertation context, and the purpose of its use.

Correlation Analysis

Correlation is a bivariate analysis that measures the strength of the association between two variables and the direction of the relationship. This strength is measured by the value of the correlation coefficient, which varies between $+1$ and -1. A value of ± 1 indicated a perfect degree of association between the two variables. As the correlation coefficient goes to 0, the relationship will be weaker. The direction of the relationship is indicated by the sign, with the "+" sign indicating a positive relationship, and a "–" sign indicating a negative relationship. In statistics, the correlation coefficient is measured by three types of methods: Pearson, Kendall, and Spearman correlations.

Pearson's correlation coefficient, which is the most used, consists of the division of the two variables covariance multiplied by their standard deviations. The correlation coefficient $\rho_{X,Y}$ is defined as in Equation 2.16:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_y} \tag{2.16}$$

Where:

X and Y are two random variables

μ_X and μ_Y are the expected values of X and Y

σ_X and σ_Y are the standard deviations of X and Y

E is the expected value operator[45]

k-Nearest-Neighbours

The k-Nearest-Neighbours (kNN) is a non-parametric algorithm that can be used in different contexts, from problems of classification, regression, or imputation methods. A non-parametric algorithm means that it does not need, *a priori*, to know relevant information about the data, as well as its distribution.

The idea of this algorithm is to find the closest k observations and estimate each point according to k nearest neighbors. The kNN fills the missing values based on the values of other variables in the dataset. This filling is done, taking into account the similarity between the k most similar samples to the missing point. The similarity between the two samples is determined using the Euclidean distance, illustrated in Equation 2.14. This method looks for the most similar k nearest neighbors of the missing point, and replaces it based on the average, for regression, or mode, for classification, of the k nearest neighbors. Below, the main advantages of using kNN imputation:

- It does not require the creation of a predictive model for each missing variable;
- It can handle missing values even if there are categorical and numeric variables;
- It Handles values with multiple missing values fairly well;
- It takes into account the structure of the correlation [46].

Stepwise Regression

Stepwise regression is a technique that aims to select the variables that best explain the variable to be estimated in the regression model. This technique is used for the iterative construction of regression models and uses one of the following approaches:

- **Forward selection:** this type of approach involves starting the regression model with no variables explaining the variable to be estimated. The idea of this technique is to add the variable that has the highest coefficient of determination, also called R^2. R^2 varies between 0 and 1, indicating how much the model can explain the observed values. Iteratively, we add the candidate variable that increases the coefficient of determination R^2. This technique ends when the other variables no longer add improvements to the model, or the improvements are not significant;

- **Backward elimination:** this approach involves starting with all the candidate variables to explain the variable that we want to predict. Iteratively, the removal of each variable is tested through a comparison criterion, in case the best model, that is, the best determination coefficient R^2, then the variable is removed. This process is repeated until there are no further improvements;

- **Bidirectional elimination:** this approach is a combination of the above, testing at each step for variables to be included or excluded [47].

Multiple Linear Regression

The Multiple Linear Regression is a technique in which through k independent variables $X_j(j = 1,...,k)$ we try to explain a dependent variable Y. Independent variables are also called explanatory or regressive variables, once that these are used to explain the variation of the dependent variable Y. The conditions underlying multiple linear regression are analogous to simple linear regression:

- There must be a linear relationship between the outcome variable and the independent variables;

- The difference o the observed and real value, also called residuals, of each variable, must be normally distributed;

- It assumes that the independent variables are not highly correlated with each other;

- The variance of error terms must be similar across the values of the independent variables.

The multiple linear regression model can be described by Equation 2.17. It is necessary to find the values of β, designated as the coefficients that best explain the dependent variable, from the

values of the independent variables. This model describes an hyperplane in the multidimensional space of the regressors X_i.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{2.17}$$

During the construction of the regression model, which tries to explain the dependent variable from the independent variables, it is necessary to choose the coefficients that minimize the error in the estimation of the dependent variable. Initially, coefficients are fixed, unknown values, and throughout the construction of the models' process, these are adjusted so that the error of the estimation of the dependent variable is minimized [48].

Classification and Regression Tree

Decision tree prediction uses a decision tree, as a predictive model, to go from observations to conclusions about the item's target value. The target variable of the tree model can take a discrete or continuous set of values, for classification or regression, respectively.

There are majorly two steps involved in building a tree:

1. It divides the predictor space set of possible feature variables into distinct and non-overlapping regions;

2. For every observation in a region, a prediction is made, which is the mean value in the training set in that particular region.

For step 1, the goal is to find boxes that minimize the RSS shown in Equation 2.18.

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - Y_{R_j})^2 \tag{2.18}$$

The data is split by a top-down approach known as recursive binary splitting. It is top-down as it starts at the top of the tree, at which all observations fall into a single region, and successively splitting the predictor space into two new branches. To perform binary splitting recursively, the predictor and cut point that reduces more significantly the RSS is selected, applying Equations 2.19 and 2.20:

$$R_1(j,s) = X|X_j \leq s \tag{2.19}$$

$$R_2(j,s) = X|X_j > s \tag{2.20}$$

For step 2, we seek the value of j and s that minimize the Equation 2.21:

$$\sum_{x_i \in R_1(j,s)} (y_i - Y_{R_1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - Y_{R_2})^2 \tag{2.21}$$

The purpose of this technique is to use the variable that allows the model to have more expressiveness on the data set. It should also be noted that this does not require any special treatment for

variables, nor is there a need for normalization of the variables [49].

Random Forest

Random forests are an ensemble learning method for classification and regression that operate by building a vast number of decision trees when it is trained, and outputting the mode of the classes or mean prediction of the individual tree for classification and regression trees, respectively.

A bootstrap aggregation or bagging technique is applied when training the random forest models. Given a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging selects a random sample repeatedly with replacement of the training set and fits trees to these samples, for $b = 1, ..., B$:

1. It samples, with replacement, n training examples from X, Y, calling these X_b and Y_b.
2. It trains a classification or regression tree f_b on X_b and Y_b.

When trained, predictions for unseen samples x' are made by averaging the predictions from all the regression trees on x', as shown in Equation 2.22, or by taking the majority vote in the case of classification trees:

$$F = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \tag{2.22}$$

This leads to better model performance because it decreases the model variance without increasing the bias. A single tree is highly sensitive to outliers in the training set, and the average of many trees is not if they are not correlated. Training many trees with a single training set would result in strongly correlated trees. A way to de-correlate the trees is by showing them different training sets with techniques like bootstrap sampling [50].

Support Vector Machine

Support-vector machines (SVM) are a type of supervised learning in which associated learning algorithms are applied to analyze data, and it is used for regression and classification analysis. Given a set of training data, a training algorithm constructs a model that assigns new observations to one category, being a non-probabilistic binary linear classifier. A SVM model represents the points in space, mapped so that their separate categories are divided by an evident gap, which is as wide as possible. New points are then mapped in that space and predicted to belong to a class or value, for classification and regression respectively, based on the side of the gap they fall.

A regression SVM, instead of attempting to classify new unseen variables x' into one of two categories $y' = \pm 1$, predicts a real-valued output for y' so that our training data is of the form shown in Equation 2.25, ans illustrated in Figure 2.5:

$$y_i = w.x_i + b \tag{2.23}$$

Where:

i = 1, ..., n

$y_i \in \Re$
$x \in \Re^D$

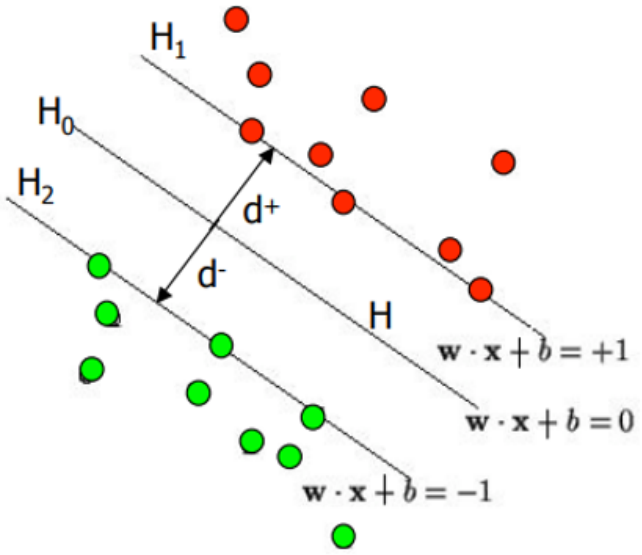

Figure 2.5: SVM regression [6]

The hyperplanes are defined in way that:

$$w.x_i + b \geq +1 \tag{2.24}$$

$$w.x_i + b \leq -1 \tag{2.25}$$

Where the planes H and d are:
$H_1 : w.x_i + b = 1$
$H_2 : w.x_i + b = -1$
H_0 is the median between H_1 and H_2: $w.x_i + b = 0$
$d+$ is the shortest distance to the closest positive point
$d-$ is the shortest distance to the closest negative point

The regression SVM will use a more sophisticated penalty function , not allocating a penalty if the predicted value y_i is less than a distance d away from the actual value t_i, i.e., if $|t_i - y_i| < d$. Referring to Figure 2.5, the region bound by $y_i \pm d$ is called an insensitive tube. The other modification to the penalty function is those output variables that are outside the tube are given one of two slack variable penalties depending on whether they lie above ($d+$), or below ($d-$) the tube.

Using SVM for regression as some advantages:

- It works well with a clear margin of separation;

- It is effective in high dimensional spaces;
- It is effective in cases where the number of dimensions is greater than the number of samples;
- It uses a subset of training points in the decision function, called support vectors, so it is memory efficient [51][52].

K-Means

K-means clustering is an unsupervised learning algorithm, which is used when we have data that is not labeled, i.e., data without groups or categories defined. The goal is to find the K groups in data. The algorithm assigns each data point to one of K groups iteratively, based on the features that are provided. Data points are grouped based on feature similarity.

The K-means clustering algorithm iterates to produce a refined final result. The algorithm inputs are the number of clusters K and the dataset. The algorithms start with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the dataset. The algorithm then iterates between two steps:

1. **Data assignment:** each cluster is defined by a centroid. In this step, the square Euclidean distance is calculated to each data point, and it is assigned to the nearest centroid cluster. More formally, if c_i is a set of centroids in the dataset C, then each observation x is assigned to a cluster based on Equation 2.26, where dist is the Euclidean distance:

$$agrmin\, dist(c_i, x)^2 \tag{2.26}$$

2. **Centroid update:** let the set of data point assignments for each i^{th} cluster centroid be S_i. The centroids are recomputed, and it is done by taking the mean of all data points assigned to that centroid's cluster, as shown in Equation 2.27:

$$c_i = \frac{1}{|\, S_i \,|} \sum_{x_i \in S_i} x_i \tag{2.27}$$

The algorithm is iterated between steps one and two until it meets the stop criterion, i.e., no observations change clusters, the sum of the distances is minimized, or the defined maximum number of iterations is reached. This algorithm always converges to a result. The result may be a local optimum, i.e., not the best possible outcome, meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome [53].

2.4 R Language

R is a programming language and free open source software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. It is widely used among

data miners and statisticians for developing statistical software and data analysis. R is popular because it has a vast number of compelling algorithms implemented as third-party libraries called packages. R uses and improves the idioms and metaphors useful for statistical computing, like working in matrices, vectors, and data frames.

Currently, it has excellent online support due to its easy use, accounting with a large number of libraries, where we can find a vast majority of statistical functions, data analysis, Data Mining algorithms, and graphics techniques. The vast majority of these libraries have documentation to support programmers and statisticians, with a description of the particular functions, bibliographic references, and informative articles.

One of the best advantages of working with R is that it allows integration with multiple databases, importing data from CSV, SAS, and Excel files. This technique also allows the reading of data through web services [54].

Chapter 3

Business Understanding

This chapter presents the scope of this study for the five months, defined based on the mapping of the oil extraction process, describing the primary operations involved. Next, the desolventizing process, responsible for a significant part of the total solvent loss in the extraction process is characterized, along with its operating variables. Finally, the objectives are defined.

3.1 Oil Extraction Process Mapping

For a better understanding of the project's context, workshops with the business domain experts were held, in order to know all the stages of the oil extraction process and the desolventizer-toaster-dryer-cooler (DT/DC) operation. In Figure 3.1 is presented a concise flow chart of the soybean extraction process in the factory plant and the transformations that occur.

At the end of the preparation process, a percentage between 60 and 70% of oil is extracted, and a porous net is formed in the solid which facilitates the extraction of the remaining oil with solvent. The second step of oil removal occurs in a solvent extractor by the percolation method, that is, the solvent entering the interior of the seeds dissolves the oil and then transports it outside due to the difference in concentration, forming the miscella (oil plus hexane).

The extractor is a structure where a bed of solids is formed, moving the soybean cake (solids plus oil) and the miscella in opposite directions to achieve a continuous countercurrent extraction. The principle of operation of this equipment is to achieve successive equilibrium between the oil concentration inside and outside the seeds through a series of pumps that move the miscella over the soybean cake, so that the miscella with the highest concentration in oil is used to extract the entering soybean cake and, at the opposite extremity of the extractor, pure hexane washes the lowest concentrated soybean cake. The solvent percolates through the soybean cake or submerges it, allowing the diffusion of the lipids into the liquid phase, resulting in the end product, marc (solids plus hexane), which is transferred to the DT/DC. The miscella extract from the extractor is made in the place of the highest concentration in oil, then is sent to the distillation operation.

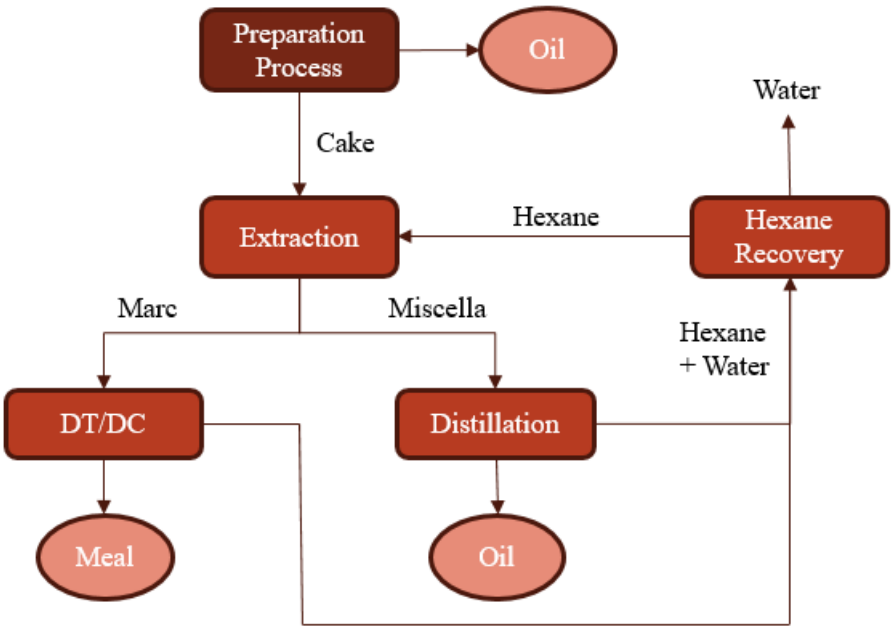

Figure 3.1: Concise flowchart of the extraction process

In the DT/DC solids pass through several plates that make up the equipment. In pre desolventizing plates, occurs most of the removal of the solvent from solids. In the desolventizing zone, the remaining solvent removal and toasting are carried out. In the last plates, the flour is cooled and dried.

The miscella is treated in a vacuum distillation operation to remove the hexane to form the crude oil. The hexane evaporated in the distillation is condensed, being separated from the water in a decanter. The recovered hexane is then reused in the extraction. The condensation gases undergo a final recovery process of hexane by absorption into a mineral oil at low temperature and its desorption at high temperature, before being sent to the atmosphere.

Due to the instability of the DT/DC process, the client faces an operational problem: the variability of the amount of remaining solvent in meals after desolventizing, which has a strong effect on environmental contamination, industrial safety, quality of meals and global economy of the process. This dissertation presents a study of the solvent extraction in the DT/DC.

3.2 DT/DC Equipment

The DT/DC represented in Figure 3.2 is a thermodynamic system in which the equilibrium occurs in a relation of temperature vs. product level vs. steam vs. pressure, evaporating the solvent from the flour as it progresses through the different trays (floors) of the equipment. The initial trays are for pre-desolventization, where the material heats with indirect steam. After this phase, the

desolventization itself is passed where, in addition to the indirect heating, it is a passage of steam straight through the bed to heat and drag hexane. At the end of desolventization, heat treatment is performed. Then drying and cooling takes place. The flour enters the top with about 25 to 35% hexane and should exit with safe values for storage and handling, approximately 500 ppm of residual hexane.

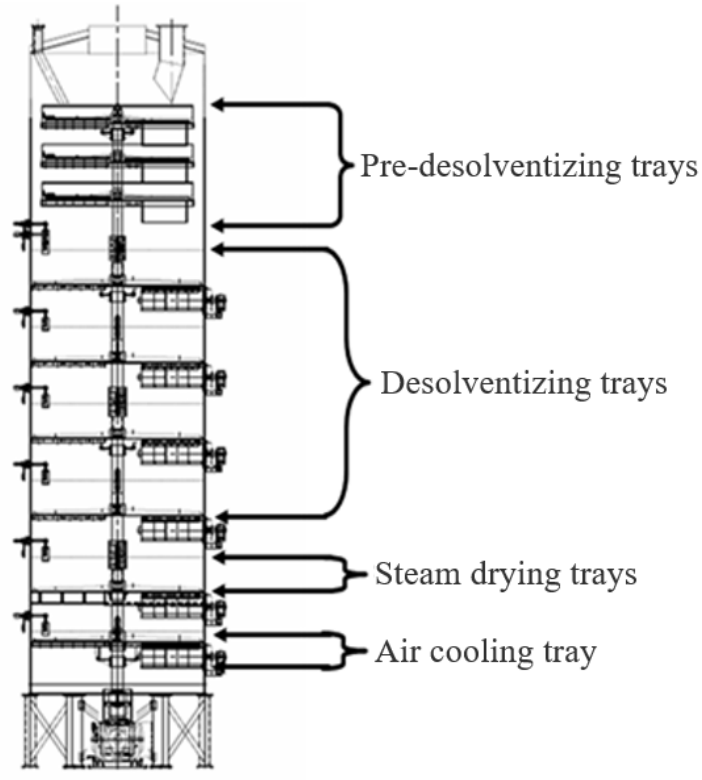

Figure 3.2: DT/DC schematic representation (adapted from [7])

After the solvent extraction process, the marc is in shape of flakes soaked with solvent, typically containing 60 to 65% of its weight in dry solids, 30% of residual solvent, 5% moisture and less than 1% residual oil. The marc usually is at atmospheric pressure and 55 to 60°C in temperature. In many cases, the marc contains anti-nutritional factors that can inhibit digestion. As is, it has no commercial value, is unsafe to transport, and requires further processing in the DT/DC.

In the DT/DC process path, the solvent is removed from the marc and recovered for reuse. Then the flour is toasted to reduce anti-nutritional factors. Next, it is dried to within the trading limit moisture requirements. Finally, the marc is cooled to near ambient temperature to remain flowable during storage and transport. The resultant DT/DC product is referred to as meal. The DT/DC processes are accomplished in a single vessel, where the desolventizer-toaster (DT) trays are in the top half, and dryer-cooler (DC) trays are in the bottom half.

DT/DC is a vertical, cylindrical vessel with a multitude of flat trays, where the marc enters at the top, and the tray supports it. The marc is mixed above each tray and conveyed downward from

tray to tray by agitating sweeps anchored to a central rotating shaft. The heat for increasing marc temperature and evaporate the solvent is supplied by steam introduced directly or indirectly into the marc. The trays of the DT are structural members, designed to hold pressurized steam, and has two different types of trays: pre-desolventizing and desolventizing trays. Next, it enters the DC where the trays are designed with an upper plate, lower plate, and a structural member, designed to distribute low-pressure air vertically into the meal layer supported above. DC has two types of trays: steam drying and air cooling trays. The flow of tasks and materials inside the DT/DC is represented in Figure 3.3.

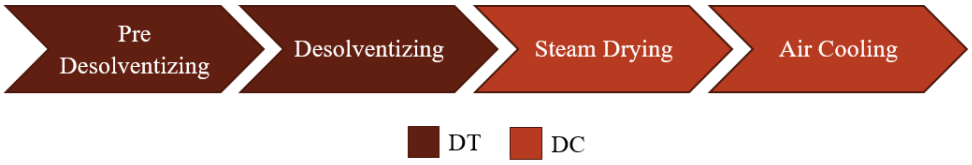

Figure 3.3: Flow of tasks and materials in the DT/DC

The three initial trays, 1st to 3rd floors, are for the pre-desolventizing task. The primary function of these trays is heating the solids through their surface by conduction, by the condensation of indirect steam. The solid material moves through the floors, passing through sluice valves that rotate faster or slower according to the level of solids in the trays. These trays are designed to allow the ascending gases to pass around without contacting the bed of solids.

The five central trays, 4th to 8th floors, are for the desolventizing task. Here there is indirect and direct heating, due to the passage of live steam through the bed. They also have holes that allow the passage of steam, which prevents the material from falling when passing. There are four desolventizing trays in order to obtain the desired residence time. In order to maintain the flow of material between the trays, there are sluice valves. These valves allow the passage of solids according to their level.

The next three trays, 9th to 11th floors, are for steam drying and the last tray for air cooling, giving a total of twelve trays in the equipment. The drying trays are also drilled in order to allow the passage of air through the bed. The air enters the chamber of each floor through a blower, and the exit passes cyclones for the separation of entrained particles. The passage of the material between floors is also done through sluice valves. In the last tray, 12th floor, there is a blower that injects cold air into the chamber, in order to cool the meal before it leaves the DT/DC for storage.

3.2.1 Operation Description

The material from the extractor enters the DTDC from the top, with about 30% hexane, at a temperature of 60°*C*, to the first pre-desolventizing plate, through a sluice valve. In the pre-

desolventizing zone, there is a removal of 15 to 25% of the solvent. The temperature of the solids should vary between 57 and 65°C, while the temperature of the gases varies between 72 and 85°C. As the heat is transferred by conduction, it is necessary to have maximum contact with the plate. Therefore, the bed height should not exceed 300mm.

The material exits the pre desolventizing trays and falls onto the first desolventizing tray. In these central trays, the mixture warms due to two processes: heat transfer by conduction to the solid material on the hot plate (indirect steam) and transfer of heat by convection of the rising steam (direct steam). Most of the heat is transferred into the meal tray by condensation of direct steam, and a deep layer of 300 to 1000mm meal depth is held above the tray. The number of countercurrent steam circulation trays has to guarantee the residence time required to evaporate the solvent, which is 25 to 30 minutes for soybean, the seed used by the client in this industrial process. The bed temperature of solids should be between 100 to 110°C, in order to avoid the auto-inflation temperature of the flour (140°C) and the protein degradation temperature. In the last tray of the desolventizing, the toast occurs. It consists of heat treatment to improve the nutritional quality of the flour as toxins and other anti-nutritional elements in animal feed are destroyed. Here, 75 to 85% of the solvent is removed, exiting the desolventizing trays with 100 to 500ppm of residual hexane.

After the meal exits the DT trays, it is at approximately 100°C and contains 15 to 20% moisture. From there it enters the DC steam drying trays, where the flour is dried with air, by forced convection, that was filtered, *a priori*, to remove dust to obtain the final product with moisture suitable for storage.

After the meal exits the DC steam drying trays, it is around 60°C and contains 12 to 13% moisture. Then it enters the DC air cooling tray, where cold air is injected up through the meal. The meal continues to cool evaporatively and convectively. The cold air exits the top of the meal layer and then exits the sidewall of the DC to a cyclone collector. The meal exits the DC with approximately ambient temperature.

The dry, cold meal is conveyed outside the solvent extraction plant for size reduction and then on to meal storage. It is essential to properly dry and cool the meal to prevent continued evaporative cooling in storage or transport, which will cause reduced flowability, solidification, and bridging of the meal inside storage and transport vessels.

3.2.2 Operating Variables

Several factors influence the DT/DC process. Steam and solid material temperature, the height (level) of the material in the trays, residual hexane in the solid material, and the addition of direct steam are the main factors influencing the process.

The steam temperature is a good indicator of how the process is performing because hexane has a specific evaporation temperature. If it is below the minimum of 72°C it means that little hexane is being evaporated, resulting in a waste of hexane and, possibly, of the final product. On the other hand, if the steam temperature is too high, it may indicate excessive use of direct steam, leading to an unnecessary increase in its consumption.

The addition of direct steam is carried out to regulate and maintain constant the temperature at the top of the DT and the temperature gradient along with the desolventization trays. The amount of steam inject is related to the amount of desolventized meal in the DT. The residual hexane in the solids decreases with increasing vapor density.

Although most desolventizers are sized for optimal residence time, some of the flour may reach the exit conveyor shortly after it has entered. It is due to problems in level control in the trays that can occur in factories that process different seeds with different daily flows. Therefore, the adjustment of the solid levels in the trays is necessary, passing this adjustment by the regulation of the sluice valves discharge speed between them. The amount of hexane in solids at the exit of the desolventization makes it possible to ascertain whether or not there are problems in the process.

There are critical factors that must be supervised, such as solids temperature and level in the trays, floor temperature, pressure, and residence time, controlled by the variables that can be managed, in each floor of the DT/DC:

- Direct steam flow;
- Steam temperatures;
- Driving force of the fans;
- Discharge speed of the sluice valves;

3.3 Business Objectives

Being the DT/DC equipment the object of study defined for this dissertation period, it was possible to establish an approach based on three layers that are considered fundamental for the optimization of the DT/DC process:

1. Monitoring the variables and its trends:
    - Monitoring all process variables over time, allowing to set alerts according to anomalous values[1];
    - Development of a predictive model that allows the estimation of the hexane concentration according to the input variables, segmented by the origin of the seed;
    - Development of forecast models that estimate trends of the variables according to the history of information[1].
2. Input parameters optimization:
    - Defining the optimal values for the input variables according to the hexane concentration set at the parameters of quality and safety ($500 ppm$);

[1]This objective is not considered under the scope of this dissertation.

- With the optimization of the optimal values of the input variables, there is a potential to increase energy efficiency[1].

3. Recommendations, according to the hexane concentration[2]:

   - Segmentation of the data into different groups according to ranges of values for the hexane concentration;
   - Definition of actions according to the input values in the different defined groups, in order to optimize the desolventization process;
   - Development of an analytical model that learns the recommendations and automatically inform the operators of the actions to be carried out[3].

The listed objectives that were not considered under the scope of this dissertation, with a footnote, are due to the short period available.

[1] This objective is not considered under the scope of this dissertation because it is necessary an in-depth study of each operating variable energy expenditure. This study is not compatible with the time available for this dissertation.

[2] This high-level objective was not defined for the scope of this dissertation. However, in the course of its development, this opportunity was identified for the creation of value, having branched the project in this direction.

[3] This objective is not considered under the scope of this dissertation.

Chapter 4

Data Understanding

This chapter presents and describes the data provided by the client. Then, the data is explored in order to become familiar with it, discover first insights and analyzing the data quality to verify if it is appropriate to achieve the objectives defined in the previous Chapter 3.

4.1 Data Presentation

The available information comes from sensors installed throughout the extraction process, measurements of the hexane concentration in the laboratory, and information about the origin and type of the seed flowing.

One hundred twenty-five sensors in the oil extraction process are recorded continuously every 5 seconds, from 16-02-2019 to 31-05-2019, for a total of 1.770.724 records per sensor, representing the operating parameters of the whole extraction process. Through workshops with the client, it is decided to study the parameters of operation of the DT/DC, because its inputs, such as moisture, hexane concentration, and residual oil, are constant. It allows the study to focus only on the desolventization equipment operation. The data is supplied in a daily data file of type DBF (Excel file). In the supervisory system (SCADA), which can be visualized in ANNEX C.1, the operational parameters are shown in order to aid the production in the adjustments that will be necessary to make, according to the current value of the variables.

Concerning laboratory measurement data, they are measured every three days at 7:00, approximately, counting with a total of 51 records from 01-01-2019 to 31-05-2019 with information about the hexane concentration of the extracted flours, along with the origin and type of the seed. It should be noted that these records are offline. They are registered manually on an Excel file in order to analyze them in RStudio.

4.1.1 Sensors Data

The DT/DC operating data is collected continuously, every 5 seconds, through monitoring sensors installed in the oil extraction process. Once the variables of interest are selected, together with the

client, this study ends up with 42 operating parameters to analyze. The variables follow a name nomenclature: "Type.Of.Sensor_what.Is.Monitoring_which.Floor".

- Types of sensors:
  - **SC:** Speed Controller;
  - **LIT:** Level Indicator Transmitter;
  - **TT:** Temperature Transmitter;
  - **PT:** Pressure Transmitter;
  - **IT:** Intensity Transmitter.
- What we can monitor:
  - **fan:** blower;
  - **valv:** sluice valve;
  - **prod:** flour;
  - **prodOut:** flour at the exit of the DT/DC;
  - **gasOut:** vapor at the exit of the DT/DC;
  - **floor:** the floor itself;
  - **bomba:** lubrication pump;
  - **motor:** main engine.
- Floors that are monitored:
  - 1, 4, 5, 6, 7, 8, 9, 10, 11 and 12.

The sensors data that is used in the scope of this dissertation are summarized in Table 4.1. In addition to the 42 operating variables, the date and time are also recorded.

Table 4.1: DT/DC operating parameters summarized

Variable	Directly Controlled?	Floor (X)	Unit	Description	Total
SC_fan_X	Yes	$X \in (9,10,11)$	Hz	Driving force of the blowers	3
SC_valv_X	Yes	$X \in (4,5,6,7,8,12)$	Hz	Sluice valves speed controller	6
LIT_prod_X	Yes	$X \in (4,5,6,7,8,12)$	^{o}gr	Flour level	6
TT_gasOut_X	No	$X \in (1)$	^{o}C	Vapor temperature at the exit of the DT/DC	1
PT_gasOut_X	No	$X \in (1)$	mmH_2O	Vapor pressure at the exit of the DT/DC	1
TT_prod_X	No	$X \in (1,4,5,6,7,8,$ 9, 10, 11, 12)	^{o}C	Flour temperature	10
TT_floor_X	No	$X \in (9,10,11)$	^{o}C	Floor temperature	3
PT_floor_X	No	$X \in (4,5,6,7,8,$ 9, 10, 11, 12)	mmH_2O	Floor pressure	9
TT_prodOut_X	No	$X \in (12)$	^{o}C	Flour temperature at the exit of the DT/DC	1
TT_bomba	No		^{o}C	Lubrication pump temperature	1
IT_motor	No		A	Main engine current intensity	1

It should be noted that only the levels of flour on each floor, the rotational speeds of the blowers, and the sluice valves speed are the only ones controlled directly by human action.

4.1.2 Laboratory Measurements

The hexane concentration data is based on measurements done in the laboratory so the client can control the hexane concentration exiting the DT/DC and adjust the process according to these results. Usually, a sample is collected every three days, at around 7 AM, to analyze in the laboratory. The results are available about eight hours later, making them not reliable to fine-tune the process.

In Table 4.2 is summarized which information is recorded in the laboratory reports, in a total of 51 records.

Table 4.2: Laboratory measurements summarized

Variable	Unit	Description
Hexane concentration	ppm	Extracted flour hexane concentration
Origin	USA or Brasil	Where the soy seeds come from
Type	44 or $47,5$	Type of flour in process

In addition to the three variables in the laboratory reports, the date and time in which measurements were performed are also recorded.

For future analysis, it should be noted that the origin of the seed seems to follow a pattern, as can be seen in ANNEX A.1. Each origin has a more extended circulation period, unlike the seed type, which is changed on a non-regular basis. Furthermore, the client does not have records identifying in each date and time which origin and type of seed are in the process.

4.2 Exploratory Data Analysis

One of the essential processes in a Data Mining study is the analysis performed on the data, as well as its quality because, in this way, it can be identified problems with the data by extracting inferences and finding relations that meet the objectives of the study.

4.2.1 Sensors Data

Before doing any analysis between variables, a basic statistical analysis of the data is performed. In Table 4.3, it can be visualized the minimum, maximum, and mean value for the variables, as well as their standard deviation, for one hour.

Through the analysis of Table 4.3, it can be seen that the variation over an hour is practically zero. Since the DT/DC is a thermodynamic system in equilibrium, it is not expected that the different operating parameters of the process would vary significantly over an hour. This way, it is probably a good idea to aggregate the data by periods of an hour.

Table 4.3: Basic statistics on a sample of operating parameters

Variable	Minimum	Maximum	Mean	Standard Deviation
SC_fan_9	37	37	37,0	0
SC_fan_10	36	36	36,0	0
SC_fan_11	35	35	35	0
TT_prod_1	75	78	76,4	0,68
TT_prod_4	62	63	62,9	0,12
TT_prod_5	105	105	105,0	0
TT_prodOut_12	28	29	28,4	0,50
TT_floor_11	23	24	23,9	0,33

Through workshops with the client, it is decided to remove the temperature of the flour of the 4th floor (*TT_prod_4*). As it can be seen in the precedent Table 4.3, the temperature varies between 62 and 63 degrees Celsius. It happens because the temperature sensor is at the top of the floor and, most of the time, since the level of the flour is usually low, the sensor is never submerged in the product. Therefore, the temperature it indicates its not the real value, which could bias negatively the future analysis. It should be noted that the real temperature value of the flour on this floor is the same as on the 5th, 6th and 7th floors.

Next, a simple visual analysis is performed on each variable through histograms and boxplots to get a feel of the variables. Based on the distribution shape, it is possible to get an idea of what to expect when distinct variables concerning each other are visualized, as well as identify possible outliers. In Figure 4.1 it can be visualized the histogram and boxplot of the variables *SC_fan_9* and *SC_valv_4*. Although not represented, this visual analysis is performed for each of the variables under study.

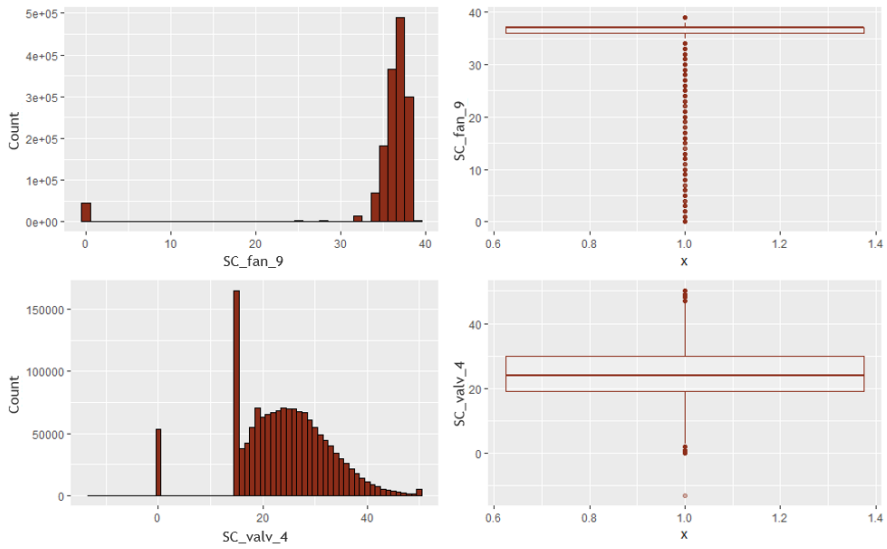

Figure 4.1: *SC_fan_9* and *SC_valv_4* histograms and boxplots

Through the analysis of these pairs of plots, it can be concluded that all variables have a significant amount of outliers. Their removal or replacement must be clarified with the client, as they may be typical operating values. Furthermore, most of the operation parameters have a high number of zeros, which may be standard operating parameters. If it indicates a shutdown in production or sensor failures, it means that they can be treated as outliers.

In order to visualize the zeros found in all the operating variables, it is drawn a temporal diagram for each variable. In Figure 4.2 is represented the temporal diagram for a sample of variables. Although not represented, these diagrams were performed for each of the variables under study.

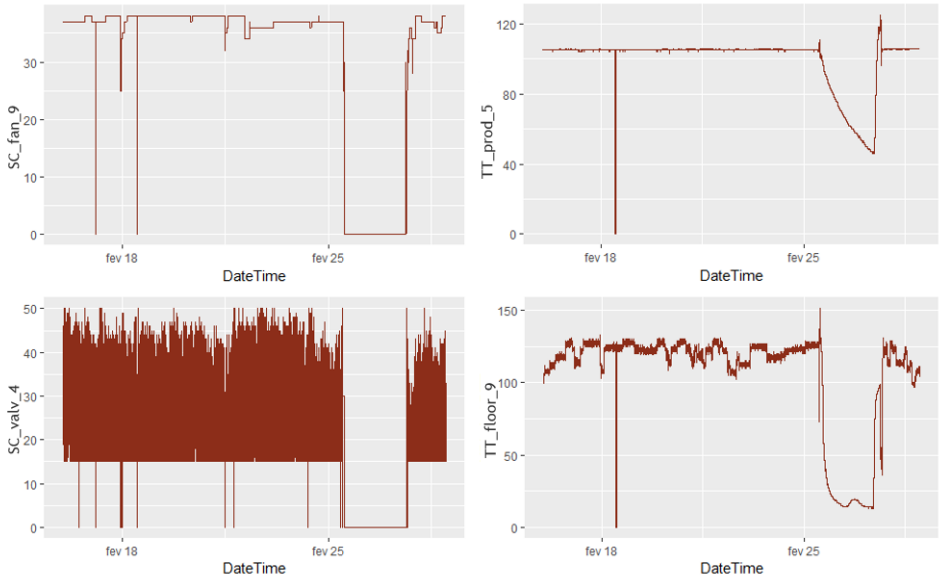

Figure 4.2: *SC_fan_9*, *TT_prod_5*, *SC_valv_4*, and *TT_floor_9* temporal diagrams

These temporal diagrams allow to see if there is any correlation between the zeros and the temporal period. For better visualization, these graphs only illustrate the February month. After analyzing these plots, it can be seen that the blowers of the 9^{th}, 10^{th}, and 11^{th} floors have stopped working, so as the sluice valves of all the floors, in the same periods. In addition, the flour temperatures of each floor are reducing gradually, which means that the direct and indirect steam was switched off. These unusual situations have to be clarified with the client.

As already mentioned in this dissertation, the DT/DC is a thermodynamic system in equilibrium, so it is expected that some variables are highly correlated with each other. To get some initial guidance about which operating parameters are correlated or not, a correlation analysis, described in Section 2.3.4, is performed. Highly correlated variables can be explained one by another, so including both in the analysis does not add value to the study. The correlation table is illustrated in ANNEX B.1. Through the analysis of the correlation table, it can be seen that practically all the variables are highly correlated with each other. Since outliers highly bias correlation tests, this

correlation analysis might be misleading. Therefore, the best option is to clarify and perform an outlier treatment before performing a correlation analysis.

4.2.2 Laboratory Measurements

Concerning laboratory measurements, it is essential to note that the first 13 records, from 01-01-2019 to 15-02-2019, will not be used as an object of this study because they lack their operational data which only exist from 16-02-2019. It can be concluded that the laboratory measurements are a poor sample compared to the operating parameters: 38 VS 1.770.724 records, respectively, illustrated as a percentage in Figure 4.3. From now on, it must be kept in mind that in order to merge the laboratory measurements with the operative variables and enrich the dataset, it is necessary to estimate the missing values for the laboratory measurements. A basic statistical analysis of the hexane concentration measures can be seen in Table 4.4.

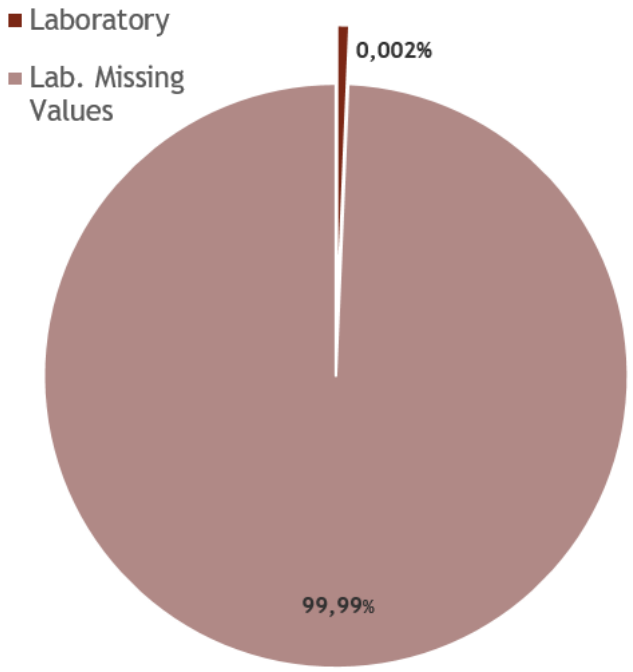

Figure 4.3: Laboratory results available vs. missing values

Table 4.4: Basic statistics on hexane concentration measurements

Variable	Minimum	Maximum	Mean	Standard Deviation
Hexane Concentration	348,6	1638,0	941,2	326,5

In the laboratory reports, in addition to the hexane concentration of extracted flour, one has the origin and type of seed. There is a need to clarify with the client if the parameters of operation

of the DT/DC vary for each origin and type of seed. Through workshops with the client, it can be concluded that the parameters of operation have slightly different values for some variables. Therefore, it is probably a good idea to split the primary dataset by origin and type to not skew the results of the future analyses. It should be noted that the flours of distinct origins and types are never mixed in the process.

Chapter 5

Data Preparation

This chapter reports the data preparation and selection techniques used in this dissertation. Firstly, the operating data is aggregated by periods of an hour, which is followed by the outlier treatment. Then, a correlation analysis is performed to perceive if there are highly correlated variables. Next, the missing values of the hexane concentration are estimated, followed by the imputation of the missing values for the origin of the seed, in order to divide the primary dataset. Finally, the variables that better explain the hexane concentration of the extracted flour to be used for the development of the models are selected.

5.1 Data Aggregation

Through workshops with the client, it is concluded that being the DT/DC a thermodynamic system in equilibrium, there is not, in fact, a significant variation of the operating parameters over an hour. Therefore, aggregating the data by the hour will benefit future analysis, because it is expected to reduce the dimensions of our dataset drastically.

When aggregating the data, it is essential to explore and decide which is the best metric to perform that aggregation. The metrics considered were the last record, mean, median, and mode of the respective hour. As it can be observed in Section 4.2.1, there is a significant amount of outliers in the dataset, namely the zeros, so aggregating the data by the mean values is not the most effective approach, since that metric is extremely sensitive to outliers. As a result of the existence of a considerable amount of zeros, it can be concluded that aggregating by the mode is not a practical approach too, once it will get the value of zero for a considerable amount of periods. Based on these facts, it is decided to use the median to aggregate the data, because it takes into consideration all the values from the period of an hour and it does not matter whether the extreme point is far away or near the other observations, as long as the central value is unchanged, unlike the last record which ignores every variation from the first record to the penultimate. The Figure 5.1, shows a drastic reduction of the observations resulting from the data aggregation.

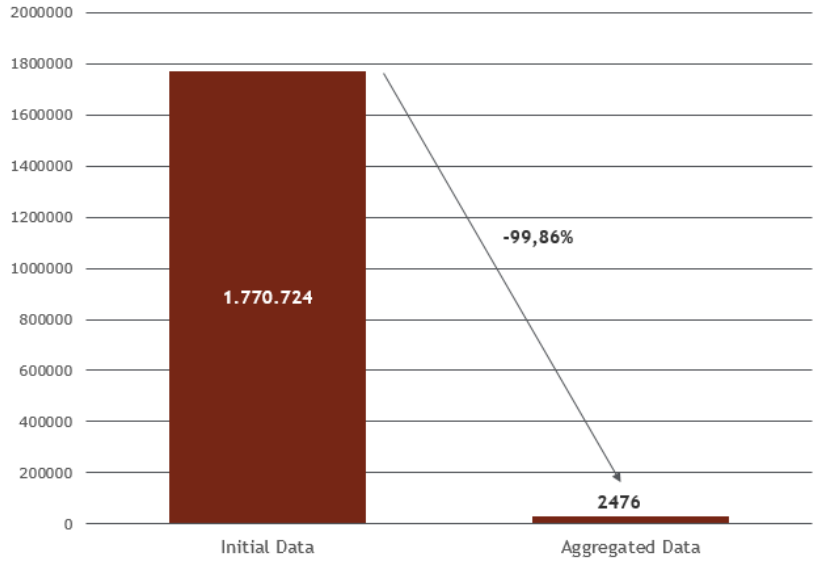

Figure 5.1: Number of observations reduced resulting from data aggregation

From now on, it is used the aggregated dataset for the analyzes, which counts with 2476 observations, a reduction of $99,86\%$ compared to the 1.770.724 initial observations. This reduction not only optimizes the analysis in terms of CPU processing but also reduces the number of missing values of the laboratory measurements significantly, since there are now 38 measurements for 2476 values needed, instead of the initial 1.770.724.

5.2 Outlier Treatment

In the context of Data Mining, statistical outliers are individual readings of data that are distant from the rest of the observations. It is essential to understand their occurrence, in the business context of the study, to be able to deal with them effectively, since they might disproportionately bias the results of a statistical analysis of the dataset as a whole.

The outlier treatment should be done carefully and shall not be confused or mistaken with the specific activity of the DT/DC, since it may indicate a different property. Therefore, a study needs to be made together with the client before an outlier is discarded.

Through workshops with the client, it is verified what is the meaning of the zeros in the DT/DC process context. After a rigorous analysis, it is confirmed that the zeros in the DT/DC were due to a production stop at the factory. Therefore, the elimination of these values from the dataset is the best approach to take, because these bias our results negatively.

The periods removed where:

- From 25-02-2019 12:00 to 27-02-2019 16:00;
- From 14-04-2019 7:00 to 14-04-2019 10:00;

- From 23-04-2019 5:00 to 23-04-2019 8:00;
- From 06-05-2019 8:00 to 06-05-2019 16:00;
- From 14-05-2019 3:00 to 14-05-2019 7:00;
- From 17-05-2019 5:00 to 18-05-2019 18:00;
- From 20-05-2019 12:00 to 20-05-2019 15:00;
- From 24-05-2019 12:00 to 24-05-2019 15:00.

The samples removed in this study results in a reduction of 174 observations which adds up to a total of 2325 observations per variable.

After this treatment, one plot again the temporal diagrams for each variable to check if the outliers are effectively removed. In Figure 5.2, it can be observed that the variable *PT_floor_8* is an horizontal line. Therefore, it does not have any value for this study, so it is removed from the dataset.

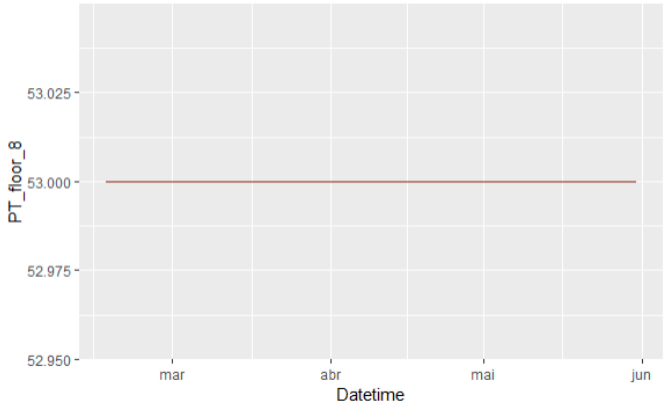

Figure 5.2: Temporal diagram of the *PT_floor_8* variable

From now on, as a consequence of the permanent removal of the outliers from the dataset, there is no chance to have this study negatively bias by them.

5.3 Correlation Analysis

As it was demonstrated in Section 4.2.1, correlation tests are highly biased by outliers. Now that the dataset is cleaned, one shall return to analyze the correlation between the variables given the following assumptions:

- Analyze the highly correlated variables ($\geq 0,85$);
- Remove the dependent variable, i.e., without direct human control;

- If two independent variables are highly correlated, one shall not remove any.

Through the analysis of the correlation table, presented in ANNEX B.2, one can clearly verify the influence of the outliers in this study, comparing with the correlation table in ANNEX B.1, referred in Section 4.2.1. Here, there were identified correlations, summarized in Table 5.1:

Table 5.1: Kept and removed correlated variables

Kept	**Removed**	**Correlation (%)**
SC_fan_10, SC_fan_11		88
SC_fan_10	PT_floor_10	97
SC_fan_11	PT_floor_11	94
PT_gasOut_1	PT_floor_4	97
TT_prod_12	TT_prodOut_12	91
TT_prod_12	TT_prod_11	86
PT_floor_6	PT_floor_5	93
PT_floor_6	PT_floor_7	95

It was expected that the blowers from the floors 10 and 11 would be highly correlated with the respective pressures of those floors because the blowers control those pressures. Since the blowers are controlled directly by human actions, and the pressures are a consequence of these actions, those pressures are removed from the analysis. Note that *SC_fan_10* and *SC_fan_11* are highly correlated, but due to the fact that these two variables are directly controlled by human action, it is decided not to remove none.

The correlation between the pressure of the vapors at the exit of the DT/DC and the pressure of floor 4 is a specific case. The pre-desolventizing floors, from 1 to 4, are not directly controlled by sensors. Since those floors are in equilibrium and operating doing the same tasks, they are controlled by the information on floor 1 and 4. Although these parameters are not directly controlled by human action, *PT_gasOut_1* is a good indicator of how the process is performing, as already referred to in Chapter 3. Therefore, it is decided to remove the variable *PT_floor_4*.

For the variable indicating the temperature of the flour in floor 12, *TT_prod_12*, the high correlation with the temperature of the flour exiting the DT/DC and the temperature in the previous floor is expected, based on the knowledge in the context of the business and the operation, because the flour goes sequentially from floor 11 to 12 to the exit. It is decided to keep the variable *TT_prod_12*, due to the fact that its the variable in the middle and can be better controlled by human action by injecting more or less cold air, with individual temperature variation.

Finally, the values of the pressure on floor six are highly correlated with floors 5 and 7. The pressure is not controlled individually for each floor, but the whole equipment. The values differ based on other variables such as the level of the flour in each tray. These are the first three trays of desolventizing tasks, which means they operate approximately with the same parameters. The value of the pressures on each floor is not directly controlled by human action, however, the criteria used to keep the variable *PT_floor_6* is because it is correlated with both floors 5 and 7, allowing us to remove one more variable from the dataset.

From here on, the variables removed are not contemplated in this study.

5.4 Data Construction

Having selected the variables of interest, it is decided to merge the laboratory data with the operational data in the same dataset, facilitating future analyzes. Next, there is a need to estimate the missing values from the hexane concentration to enrich and complete the dataset. Finally, it is estimated at each day and time there is the transition between the two possible origins of the seed, in order to have two subsets of data, according to each origin, since they have slightly different operational parameters.

5.4.1 Hexane Concentration Missing Values Estimation

To solve the missing values problem referred to in Section 4.2.2, the kNN method is applied as an estimation method to replace those missing values, which fills them based on similarity. It should be noted that the kNN method uses the other variables in the dataset to replace the missing, i.e., it is sensitive to the local structure of the data, thus there is a need to enter the date as another variable, because it is a time series, and account shall be taken of the evolution of each variable.

The study of the use of different k values for the kNN method was carried out. The mean and standard deviation is calculated to perceive the impact that the different k values have on the substitution of the missing values.

The use of the different k values is compared in Table 5.2, in order to verify that each approach has very close mean values, but significantly different standard deviation values. It effectively quantifies the amount of variation for the set of values, indicating if the observations tend to be close to the mean, or if they are spread out over a wide range of values for low and high standard deviation values, respectively.

Table 5.2: Comparison between different k parameters for the kNN method

	k = 3	**k = 5**	**k = 7**	**k = 9**
Mean	954,2	959,7	950,5	949,5
Standard Deviation	277,3	239,5	198,0	179,0

Ideally, in a general case, one wants the lower standard deviation. In the context of this analysis, and comparing with the Table 4.4 from the previous Section 4.2.2, it is decided to substitute the missing values with $k = 3$, since it has the closest value of standard deviation when compared to the value of the actual laboratory data provided, which means that the dispersion of the estimated values is congruent with the dispersion of the actual measurement values.

5.4.2 Origin Missing Values Imputation

As concluded in Section 4.2.2, the origin and type of the seed in the DT/DC process influences the values of some operational parameters. However, the client does not have records that accurately identify the origin and type of seed in the process. Therefore, and taking into account the evidence regarding the origin and type of the seed referred to in Section 4.1.2, the origin follows a regular pattern, and the type does not. In this way, it is decided not to use the type of the seed in this analysis, once it cannot be accurately estimated what type is in process, considering the scarce sample of available laboratory measurements.

To estimate the origin of missing values, the kNN method is applied. The resulting dataset is then analyzed in order to adjust exactly at which hour of the day the seed is swapped, with the respective error associated. This results in the following periods for each seed origin:

- USA:
  - 16-02-2019 00:00 to 20-03-2019 16:00;
  - 08-04-2019 09:00 to 08-05-2019 15:00.
- Brasil:
  - 20-03-2019 17:00 to 08-04-2019 08:00;
  - 08-05-2019 16:00 to 31-05-2019 23:00.

With the seed origin estimated, it is subset the primary dataset into a subset for each origin, which results in a subset with 1399 observations for the USA and a subset with 926 observations for Brasil.

5.5 Variable Selection

At this stage, having the datasets with the hexane concentration estimated, the variables that better explain the hexane concentration of the extracted flours are selected. In order to get initial guidance about which variables are highly correlated with the output concentration, it is performed a correlation analysis for the main dataset and the USA and Brasil subset. The correlation matrices can be found in ANNEX B.3, B.4 and B.5, respectively. The hexane concentration variable is named "saida".

By analyzing those correlation matrices, it can be concluded that there is not any variable in each dataset that is highly correlated with the hexane concentration. This result was expected, because, as it is already known, the DT/DC is a thermodynamic system in equilibrium. Therefore, it would not make sense to have a single variable adequately explaining the hexane concentration from the extracted flours.

Then, it is applied the stepwise regression technique described in Section 2.3.4. Here, the stepwise regression technique is used for the forward and backward selection methods. The resulting number of variables for each method and datasets are summarized in Table 5.3.

Table 5.3: Number of variables resulting from stepwise regression methods

Dataset	Nº of Observations	Stepwise Method	Nº of Variables
Main	2325	Forward	27
		Backward	25
USA	1399	Forward	30
		Backward	28
Brasil	926	Forward	22
		Backward	20

Ideally, it would be appropriate to remove more variables using backward selection to simplify the datasets and future analysis. However, it may not be the best option. As it was previously observed in the preceding correlation matrices, almost all the variables have a correlation coefficient approximately equal to zero, which indicate that the more parameters this study has, the better. Also, there is a risk associated when applying each method of the Stepwise Regression technique, since at this stage the regression model may still not know specific patterns of each variable given the reduced size of our datasets.

As it can be seen in the resulting number of variables for each dataset and method, the selected variables differ significantly between datasets, especially between the USA and Brasil. Although the seeds of different origins have some slightly distinct operating parameters, the seeds are the same, of soybean. Therefore, it is surprising to note the significant difference in the selection of variables between each of these. Note that the number of observations in the USA dataset is approximately 50% higher than the Brasil dataset, which corroborates the observation we made before.

5.6 Datasets Systematization

At the end of this data preparation phase, it is possible to present the datasets summarized in Table 5.4, which will be used in the next chapter for the modeling phase.

Table 5.4: Summary of available datasets

Dataset	Origin		Stepwise	
	USA	Brasil	Forward	Backward
A (baseline)				
B			X	
C				X
D	X			
E	X		X	
F	X			X
G		X		
H		X	X	
I		X		X

It is important to highlight, for the next chapter, that the division of data by origin may be skewed since it is not known exactly when the seed exchange occurred. Besides, these same subsets of data have a significantly smaller number of observations as compared to the baseline dataset (A). This is likely to negatively influence the selection of variables in the Stepwise technique, hence a particular attention will be given when analyzing their results going forward.

Chapter 6

Modeling and Evaluation

This chapter describes the process of choosing, building, and evaluating predictive models. First, it is chosen the most appropriate technique for the study. Before constructing the models, it is defined the procedure to test the quality and validity of the models. Then, the results of the different techniques are compared. Then, the technique that presents the most effective results for the context of the study is chosen. Predictive models are then constructed, based on the chosen technique in various scenarios. Again, the quality and validity test procedures are defined for each of the models, which are evaluated and compared based on this procedure. Finally, having chosen the most appropriate model to the context of the study, a clustering analysis is performed, allowing to segment the data into different interesting groups in the business scope.

6.1 Select Modeling Technique

Initially, several techniques were considered to construct predictive models. The following four techniques were chosen:

- Multiple Linear Regression (MLR) - baseline;
- Support Vector Machine (SVM);
- Regression Tree (RT);
- Random Forest (RF).

It was considered the hypothesis of using neural networks, which is a supervised learning technique built on a large number of simple elements. It is a widely used technique because it outperforms virtually all other prediction algorithms. It was decided not to construct a predictive model based on the neural networks, since one of its requirements is a significantly larger dataset than for the remaining techniques, in the order of million observations.

The basis of MLR is to assess whether the dependent hexane concentration variable can be predicted from the set of independent operational variables. This is the simplest of the models,

working as our baseline model. Unlike the MLR technique, the SVM is more effective in high dimensional spaces and with a clear margin of separation. The RT technique is selected because it does not require any assumptions of linearity in data and its focus on the relationship among various events, replicating the natural course of events, remaining robust with little scope for errors. Following the rationale used to select the RT model, the RF model is selected, since it is an ensemble-learning algorithm that aggregates the outputs of multiple regression trees, it is expected to generate more accurate results.

6.1.1 Test Design

Before building a model, it is necessary to define a procedure to train, test, and evaluate it. In Figure 6.1 it is shown the procedure adopted. A machine learning model aims to make good predictions on new, previously unseen data. In order to achieve that, it is needed to divide the data preparation phase outputted datasets into train and test data. The analytical model is trained with the train set and tested with the test set, which is the unseen data, i.e., the data that the model does not know. It is essential to make sure that the test set is large enough to yield statistically meaningful results and is representative of the dataset as a whole. Finally, the model performance is evaluated based on validation metrics.

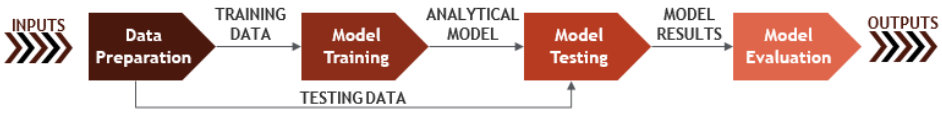

Figure 6.1: Procedure to train, test and evaluate analytical models

It was decided to use datasets A, D, and G from Table 5.4, because in this phase the objective is to perceive the behavior of the models with the complete dataset and segmented by origin. The addition of other factors, such as the selection of variables, which differ from each dataset, can influence the results. In comparing the performance of the models, it would not be clear which factor is influencing their composition, with the risk of misleading the study. Therefore, it was decided at this stage not to use the datasets with the variables selected by the Stepwise Regression technique.

The datasets chosen are randomly split at 75% for training and 25% for the test, making sure the test set is large enough to represent the dataset as a whole, and the training set is large enough not to hide essential patterns that the analytical model needs to learn. Note that, since the dataset A has more observations than the other datasets chosen, it could be split it in a way to have less percentage in the test set. Although to effectively compare each model, it is not wise to vary the test design for each model.

The resulting models are tested with the test data and evaluated based on the following chosen validation metrics:

- **Coefficient of determination (R^2):** it is the proportion of the variance in the dependent variable that is predictable from the independent variable. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model;
- **Root mean square error (RMSE):** it is the square root of the average of squared errors. The effect of each error is proportional to the size of the squared error. Thus larger errors have a disproportionately large effect and are sensitive to outliers. It serves to aggregate the magnitudes of the errors in prediction for various times into a single measure of predictive power;
- **Mean absolute error (MAE):** it is a measure of the difference between two continuous variables, being the absolute vertical distance between each point and the identity line. Each error contributes to MAE in proportion to the absolute value of the error, which is not true for RMSE;
- **Mean absolute percentage error (MAPE):** it is a measure of prediction accuracy, usually expressing it as a percentage. It is commonly used as a loss function for regression problems and in model evaluation, because of its very intuitive interpretation in terms of relative error.

With this procedure applied, the results of each dataset for each technique can be analyzed and compared. Then, the models that better fit the study are chosen.

6.1.2 Evaluation of the Different Techniques

In order to compare the different techniques, the generate test design procedure is followed. In Table 6.1 it can be found the individual results of each validation metric chosen for each technique and dataset used.

Table 6.1: Validation metrics applied to each model for each dataset

Dataset	Technique	Metric			
		R^2	RMSE	MAE	MAPE
A	**MLR**	0,304	233	181	24,0%
	SVM	0,651	166	106	14,2%
	RT	0,583	180	124	15,6%
	RF	0,801	128	80	10,2%
D	**MLR**	0,268	253	192	24,5%
	SVM	0,589	185	120	14,8%
	RT	0,594	185	126	14,9%
	RF	0,769	140	84	9,9%
G	**MLR**	0,299	199	140	17,6%
	SVM	0,627	145	85	10,3%
	RT	0,448	180	120	14,0%
	RF	0,766	118	76	9,0%

Through the analysis of the precedent table, it is possible to verify that the baseline model (MLR) is weakest in all cases. Not only is the associated error untenable, but also the low R^2 value indicates that there is no strong correlation between the independent variables and the dependent variable. On the other hand, the RF model shows significantly lower errors for all the datasets. Furthermore, the R^2 values are considerably higher than for the other techniques, being unanimous that it is the most appropriate model to use from now on.

The results of each technique for the complete dataset (A) and USA dataset (D) are very close to each other, unlike the results for the Brasil dataset (G). Although the RF technique applied to the G dataset presents a MAPE smaller than the one applied to the dataset A and D, the R^2 has a lower value, which leads to believe that it is not advisable to use the dataset G, because it indicates that the model does not effectively replicate the results. It is mainly because the number of observations in the Brasil dataset is significantly lower than the others, missing some crucial patterns that the model does not know, corroborated by the error associated by estimating the origin in Section 5.4.2 to divide the dataset into Brasil and USA.

The similarity of results between the complete and USA datasets is interesting. Although MAPE is slightly lower for the USA dataset, the R^2 is smaller, and the RMSE and MAE are higher, indicating that a lower portion of the data explains the output variable, and the predicted value has more variation compared to the expected value. It is important to note that the similarity between these results indicates that, although some operating parameters are distinct, this difference is small, which is expected once the seed in the process is the same. It is decided not to use the segmented origin datasets until accurate records are available regarding the origin of the seed in process, so as not to bias the next analyzes negatively.

6.2 Construction of the Random Forest Model

The technique selected for the construction of the predictive model is the Random Forest, which, through the combination of the operation data predicts what will be the hexane concentration of the extracted flours. There is a need to develop different models for various scenarios, based on the same predictive technique, in order to compare the results of the different approaches, aligning them with the business context. It allows drawing conclusions about data quality, as well as how the data should be prepared in the data preparation phase, optimizing their effectiveness in the predictions. It is done by constructing models for each of the datasets A, B, and C following different procedures for training, testing, and evaluating the resulting models.

6.2.1 Random Forest Test Design

For each dataset, the procedure referred to in Section 6.1.1 is applied. However, since this analysis is performed with the complete datasets, which have a more significant number of observations, it is decided to divide the datasets into 80% for training and 20% for testing. Besides, the K-fold Cross Validation technique is also applied in order to better adjust the RF error. This technique divides the data into k sets of the initial data and trains the models with the *k-1* sets, using the

other set that is missing for testing. This approach is made k times, always selecting a different set for testing. A model is built for each test set, and the accuracy of each model developed is used to evaluate the overall accuracy, averaging all the models. It should be noted that the random split into a percentage of training and testing data is essential since the models need to learn all the events and patterns, and also need to have those cases in the testing set, which validate those events without taking into account the temporal space.

In addition to the procedure described previously, two further are outlined. The second procedure starts by randomly choosing 50% of the observations from the complete dataset, where the K-fold Cross Validation technique is applied, and the results of the validation metrics are recorded. This process is repeated iteratively, increasing the number of observations collected from the complete dataset by 10% to 100% of the observations, allowing to see if the errors tend to decrease with the increment of our dataset.

The third procedure has the purpose of simulating the application of the model in a real industrial context. For this, the test set is set to be the last 15% observations of the primary dataset, and the training data is generated following a non-random temporal approach, starting with the first 45% observations, being the respective model tested with the test set. This process is repeated iteratively, in which the sample used for the training set is incremented by 10% at each iteration, up to 85% of our total observations. Note that this increment is represented by the 10% temporally following the previous observations. It allows understanding the behavior of the model with the increment of data from a temporal perspective, an approach that is more approximate than the previous ones to the real industrial context.

6.2.2 Assess the Models Results

The results of applying the first procedure to the datasets A, B, and C, which are the datasets with all observations, with the variables resulting from Stepwise Regression by the forward and backward method respectively, are shown in Table 6.2. For the K-fold Cross Validation technique, a $K = 5$ is chosen, since it divides the dataset into five subsets, 20% for each, and trains the model with four of these subsets, representing 80% for the training set, and tests in the remaining 20% of data, consistent with the precedent baseline data division at 80% for training and 20% for test.

By analyzing the results of Table 6.2, it is concluded that using the Stepwise Regression technique in any of the modes is not the best approach. As can be seen, for the same datasets, where the only difference is the number of variables used to train and test the models, the error tends to increase with the decrease in the number of variables, confirming that due to the reduced size of our dataset, there are variables removed by Stepwise that are important for explaining the hexane concentration of the extracted flour. Also, as it was already seen in Section 5.6, the larger the number of observations in the dataset, the more variables are selected by the Stepwise Regression technique. Then, it is decided not to use the datasets with the variables selected by the Stepwise technique in the two following procedures.

Table 6.2: Results from applying the first procedure to each dataset

Dataset	5-fold CV	Metric			
		R^2	**RMSE**	**MAE**	**MAPE**
A		0,797	129	79	10,4%
	X	0,805	125	77	10,1%
B		0,698	156	97	13,1%
	X	0,705	152	93	12,6%
C		0,701	155	98	13,2%
	X	0,701	153	94	12,7%

Note that the K-fold Cross Validation technique reduces the error. This is due to the fact that the model is trained and tested over the whole dataset, iteratively divided into 80% for training and 20% for test, not forgetting that the five test sets of the different iterations are disjoint, in which their results are the closest to reality, concluding that the K-fold Cross Validation technique is an excellent procedure to evaluate the models results.

The results of applying the second procedure to the dataset A are shown in Table 6.3. Again, a $K = 5$ is chosen for the K-fold Cross Validation technique, consistent with the study that has been doing so far.

Table 6.3: Results from applying the second procedure to the dataset A

Dataset A % of Observations	Metric			
	R^2	**RMSE**	**MAE**	**MAPE**
50	0,340	245	184	20,3%
60	0,374	241	176	19,5%
70	0,405	233	166	17,1%
80	0,548	196	131	14,4%
90	0,701	151	97	12,6%
100	0,805	125	77	10,1%

Analyzing these results, it is possible to conclude that the error of our predictive models tends to diminish gradually with the increase of the number of observations in our dataset. To corroborate this conclusion, the third procedure to train, test, and validate the models is applied for the dataset A, this approach being the closest to the reality. The respective results are shown in Table 6.4. Note that, in this approach, it is not possible to validate the models with the K-fold Cross Validation technique.

Table 6.4: Results from applying the third procedure to the dataset A

Dataset A % of Training Observations	Metric			
	R^2	RMSE	MAE	MAPE
45	0,145	287	223	24,3%
55	0,343	259	196	21,9%
65	0,334	259	197	22,1%
75	0,329	252	186	20,3%
85	0,454	224	165	19,6%

Although in this approach, representative of a more realistic industrial context, the error does not decrease as significantly as in the previous one, it is notorious the gradual decrease of the error of each model, as the training sample is increased. As the model is trained according to the arrival of data on a time basis, it is natural to be unaware of specific behavioral patterns of the operating variables and their hexane concentration at the exit of the DT/DC, as opposed to the second procedure in which the data is split randomly ending up contemplating a higher number of behavioral patterns of the variables.

It should be noted that when the model is perfect, i.e., when it presents an error approximately equal to zero, it will not be necessary to train it more, since it effectively predicts the hexane concentration of the extracted flours when new operating data arrives.

6.3 Clustering Analysis

Following our analysis, the need arises to divide the data into different groups according to ranges of hexane concentration at the exit of the DT/DC, applying clustering techniques for this purpose. This analysis is vital to try to understand the behavior of the operative variables through each hexane concentration range.

Table 6.5 shows the groups that were manually formed according to the hexane concentration of the extracted flours and the number of observations in each of these groups. It should be noted that, during the period of this dissertation, only 4,56% of the flours extracted are within the quality and safety standards, and the highest concentration is between 900 and 1100ppm (37,98%), far from complying with the same standards.

Table 6.5: Groups formed according to the extracted flours hexane concentration

Group	Hexane Concentration Range (ppm)	Nº of Observations	%
1	≤ 500	106	4,56
2	]500, 700]	285	12,26
3	]700, 900]	490	21,08
4	]900, 1100]	883	37,98
5	> 1100	561	24,13

Having the data divided into different groups of hexane concentration ranges, it is possible to analyze their trend in each group. In Table 6.6, all the interesting variables in this analysis with their minimum, mean, and maximum values in each group are illustrated. In ANNEX A.2, it can be seen the complete table with all the variables under study.

Table 6.6: Minimum, mean and maximum values of each variable in each group

Variables	**$\leq$ 500**			**]500, 700]**			**]700, 900]**			**]900, 1100]**			**>1100**		
	Min.	**Mean**	**Max.**	**Min.**	**Mean**	**Max.**	**Min.**	**Mean**	**Max.**	**Min.**	**Mean**	**Max.**	**Min.**	**Mean**	**Max.**
SC_fan_9	36	37,0	38	32	36,7	38	25	36,4	38	26	36,3	38	28	36,5	39
SC_fan_10	34	37,0	38	29	36,7	38	24	35,9	38	26	36,3	38	0	36,1	38
SC_fan_11	34	36,7	38	28	36,4	39	25	35,6	38	26	36,0	39	25	36,2	39
SC_valv_4	18	23,8	29	15	23,7	32	15	24,9	50	15	24,7	42	15	23,7	50
SC_valv_6	16	17,2	23	16	17,3	27	16	18,5	30	16	18,8	31	0	18,7	29
SC_valv_8	12	14,0	15	11	13,7	16	11	13,9	17	0	13,7	18	0	14,5	20
LIT_prod_4	36	42,3	49	34	39,9	53	34	42,4	54	34	42,0	54	28	39,3	55
LIT_prod_6	36	48,3	54	32	46,3	57	23	45,5	59	23	44,7	58	21	41,9	54
LIT_prod_7	43	56,9	59	21	54,6	60	24	51,0	63	21	50,9	64	21	47,1	62
LIT_prod_8	32	53,8	57	24	50,8	57	24	47,8	62	19	48,0	62	23	44,6	58
TT_gasOut_1	73	73,6	76	73	73,4	77	73	73,4	76	73	73,6	80	73	73,8	87
PT_gasOut_1	-92	-59,0	-20	-91	-58,8	-18	-92	-52,2	-10	-91	-49,3	-7	-90	-48,0	-10
IT_motor	330	364,2	393	311	355,4	393	311	353,8	393	268	347,3	393	289	339,7	393
PT_floor_6	133	208,1	296	44	173,9	337	14	182,5	357	9	186,9	448	27	197,2	399
PT_floor_12	347	409,1	448	200	397,1	447	196	378,0	436	197	375,4	445	154	392,9	444
TT_bomba	53	57,8	63	52	57,9	64	49	58,1	65	48	58,9	67	43	58,5	68

The analysis of this table allows to draw the following sets of evidence for each variable shown:

- ***SC_fan_9*:** when the minimum values are below $36Hz$, a high hexane concentration is always indicated;
- ***SC_fan_10*** and ***SC_fan_11*:** when the minimum values are below $34Hz$, a high hexane concentration is always indicated;
- ***SC_valv_4*:** when the maximum values are above $29Hz$, always indicate an hexane concentration that is inconsistent with the quality and safety standards;
- ***SC_valv_6*** and ***SC_valv_8*:** their values tend to be higher for higher hexane concentrations;
- ***LIT_prod_4*:** the range of minimum and maximum values tend to be higher with the increase of the hexane concentration;
- ***LIT_prod_6***, ***LIT_prod_7*** and ***LIT_prod_8*:** the minimum and mean values tend to decrease with increasing hexane concentration;
- ***TT_gasOut_1*** and ***PT_gasOut_1*:** the maximum values tend to be higher for higher hexane concentrations;
- ***IT_motor*:** the minimum and mean values are indirectly proportional to the hexane concentration, tending to decrease with increasing concentration;
- ***PT_floor_6*** and ***PT_floor_12*:** the minimum values tend to be lower for higher hexane concentrations;

- ***TT_bomba*:** the minimum and maximum values tend to have a wider range of values for higher hexane concentrations.

In order to have a visual notion of the tendency of each variable for the different groups, graphs are drawn for all the variables, and four of them can be visualized in Figure 6.2, in which in the ordinates axis is the dependent variable (hexane concentration at the exit of DT/DC) and in the abscissa axis the independent variables. It should be noted that this approach was performed for all variables under study. This visualization was based on the K-means clustering technique described in Section 2.3.4, with the input parameter $K = 5$ clusters, consistent with the five hexane concentration ranges manually formed. The rectangles illustrate the values of the observations that are not within the quality and safety standards.

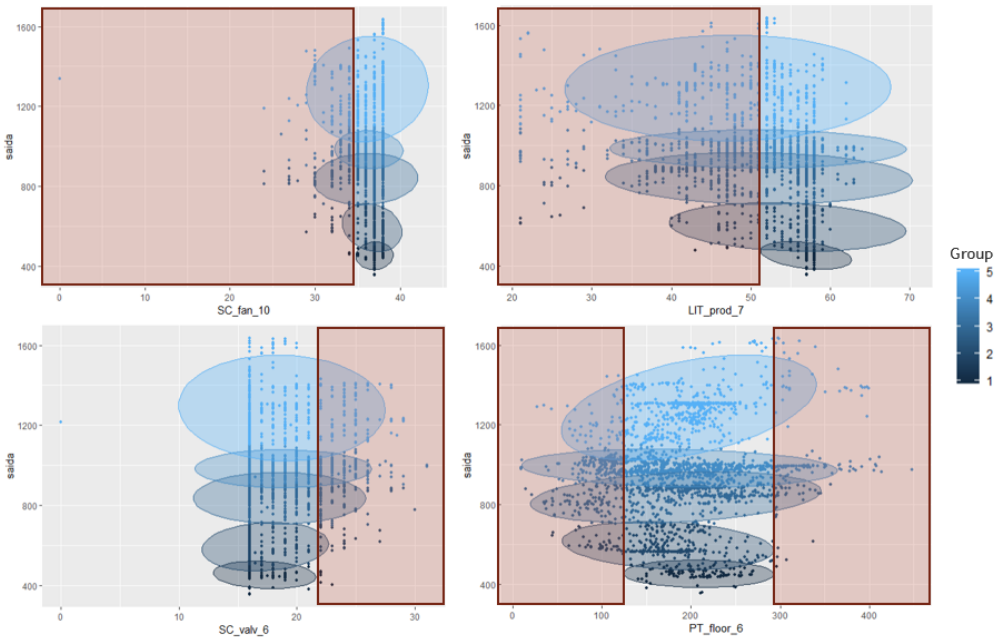

Figure 6.2: Visual analysis of the hexane concentration clusters for each operational variable

This analysis is essential to define actions based on the supervision of each operative variable. For example, when the *LIT_prod_7* variable illustrated in Figure 6.2, which is a variable controlled directly by the factory operator, is below the minimum value for the hexane concentration group that meets the quality and safety standards ($43^{o}gr$), it means that some action has to be taken, and an alert is issued in the supervision system that will be implemented in the factory, noting that it is necessary to increase the level of flour on the 7^{th} floor of the DT/DC. This line of reasoning extrapolates to the remaining operational variables.

Chapter 7

Conclusions and Future Work

In the agrofood industry, there is a high level of demand and external imposition regarding the quality control procedures, together with a high degree of dependence of the operators to control the production, inciting a high variability in the solvent extraction process. As a result of an unstable extraction, the company under study recognized as a recurring problem the excess hexane concentration in the extracted flours due to the high impact on the quality of the final product and safety of the factory and employees which operate there.

The project developed arises from an intervention to optimize the desolventization process, minimizing the hexane concentration of the extracted flours in order to control the process in real time, with the primary objective being the fulfillment of the legislated quality and safety standards. The plot objectives are related to identifying all factors with an influence on hexane concentration at the exit of the process, monitoring the trend of these variables. It allows the definition of alerts according to exceptional values, and the development of predictive models that predict the hexane concentration at the exit of the process according to the input variables. It allows the definition of optimum values for the input operational variables for an hexane concentration fixed within the quality and safety standards. Also, to make recommendations according to the current value of the hexane concentration, by segmenting data into different groups according to ranges of values for this concentration, defining actions for the different groups. It was designed the focus for the production of a product that satisfies the company in all parameters it designs while reducing the waste resulting from the process.

Throughout the project, the analysis of the process, clarification of doubts and evaluation of results in the scope of the business was carried out in work sessions, with the presence of the company production team in the solvent extraction process. The purpose of these events was to incorporate the know-how of the production team regarding the process, to certify more excellent suitability of the objectives and solutions to the organization's strategy. The sessions allowed the sharing of ideas and expertise, enhancing the effectiveness of the proposed solutions. It is considered that the participation and involvement of the production team were central to the business understanding, clarification of doubts, and assumed assumptions and resolution of constraints that arose in the course of this study.

7.1 Main Conclusions

The initial period of the project was to understand the business, the desolventization process and all the variables inherent to the process, being this a thermodynamic system in which the equilibrium occurs in a relation between temperatures, levels, steam, and pressure.

In the next step, the operational data and laboratory measurements were collected, and multivariate and exploratory analysis was made, allowing to perceive the quality of these. The operating data are recorded computationally every 5 seconds, while for the laboratory measurements result from the collection of a periodic sample at 7 AM in most cases, to analyze in the laboratory where its results are available approximately 8 hours later. The late laboratory results do not allow real-time process tuning. Fundamental statistical analysis was performed on the available data in which it was realized that the data would need to be treated and prepared for the construction phase of the models since they have numerous outliers resulting from production stops at the factory. It was also found that operational data had practically no change over an hour, and it was concluded that the best approach was to aggregate them from hour to hour according to the median of their values in that period.

After the data aggregation and outlier treatment, a correlation analysis was performed between the operational variables in which it was decided to remove from the study the highly correlated variables that do not have direct human acting. Next, there was a need to estimate the laboratory measurements missing values due to the limited number of available measurements (about 1.5% after aggregation). There is the same percentage of missing values for the origin and type of seed in circulation in the process. Knowing that the operational parameters are different according to the origin and type of seed, it is concluded that it is essential to record precisely what origin and type of seed are circulating in the process in order not to bias the results of the predictive models negatively. Also, given the scarcity of laboratory measurements, it is considered necessary to collect samples for laboratory analysis on a more frequent basis and recording precisely the date and time in which the sample is collected. After the data were enriched with the estimation of the hexane concentration missing values, the variables that best explain this dependent variable were studied. It was concluded that this is not the best approach given the small size of data available for an equipment in which there is a complex balance between several operative variables, and more data is needed.

In the modeling, validation, and evaluation phase, several approaches were tested with the different datasets resulting from the data preparation phase, in which it was concluded that the Random Forest technique is the most suitable for the project under study. The results of this predictive model for the complete dataset are optimal, given the current performance of the process being studied. One approach was to verify that the model errors decreased with increasing data, and it was concluded that more data is needed for the model to yield perfect results, i.e., an error approximately equal to 0%. It is worth noting the importance of continuously recording the origin and type of seed in the process, allows the construction of models optimized for each case. The results of the final model were discussed with the client, and it was concluded that, although not

perfect, the results are excellent given the enormous current variability of the DT/DC process. The predictive model results made possible the discovery and definition of optimum values for the operational variables according to the hexane concentration that is expected at the output of the equipment.

The project has achieved results not only aimed at meeting quality and safety standards but also in terms of productivity, radically changing the way the production team analyzes, evaluates, and interacts with the process. The production of this tool on the shop-floor, collecting operating data constantly, significantly improves the entire operation. It allows us the inclusion of a system that monitors the hexane concentration of the flour at the output of the DT/DC in real-time and its parameters. Also, their trends, automating alert mechanisms, and recommending actions based on these trends, contributing to the maximization of productivity and optimization of the process.

The use of analytical models presents an enormous potential to improve the efficiency and effectiveness of the processes, as it allows the scalability of the solution and centrally managing the production processes. Integrating this solution translates into a wide range of benefits that would not otherwise be possible, such as the anticipation of problems in the processes. Also, higher accuracy in the hexane quantification of the extracted flours, improving the quality and safety indexes, higher efficiency of production processes, and better information management and support for decision making.

The core of this solution, which is built on the convergence of the various emerging technologies of Industry 4.0, transcends quality. It has an undeniable impact on the organization as it will revolutionize the entire process culture and work routine, and it is scalable to the other process and the company's plants. It will significantly alter the way workers, processes and machines interact to create efficiencies, new sources of innovation, and new insights that were not available previously, supporting informed decision making, increasing productivity and agility in problem-solving.

7.2 Future Work

In order to materialize the potential demonstrated by this project and to guarantee the stability of the solution, it is imperative to continuously record the origin and type of seed in circulation in the solvent extraction process, together with a more frequent measurement of the hexane concentration in the laboratory with the respective date and hour, in order to optimize and guarantee the reliability of predictive models predictions. This continuous record will allow the development of forecast models that estimate the trend of each variable in real-time, according to the information history.

In this project, it was possible to define a range of optimal values for the input variables of our predictive model according to the hexane concentration set in the legislated quality and safety standards. By optimizing these values, there is a potential for increasing energy efficiency. For this, a thorough study must be carried out in order to perceive what variables directly affect energy expenditure.

During this dissertation project lifecycle, it was possible to define actions according to the concentration of hexane by segmenting the data into different groups according to ranges of hexane concentration values, to optimize the process. When the predictive models implemented are optimized, it will be possible to develop an analytical model that will learn these recommendations and automatically inform the operators of the actions to be taken on the process.

The developed solutions will be implemented through SAP tools on the shop-floor since the client landscape is SAP. Figure 7.1 shows the designed architecture in which the developed solutions will be implemented.

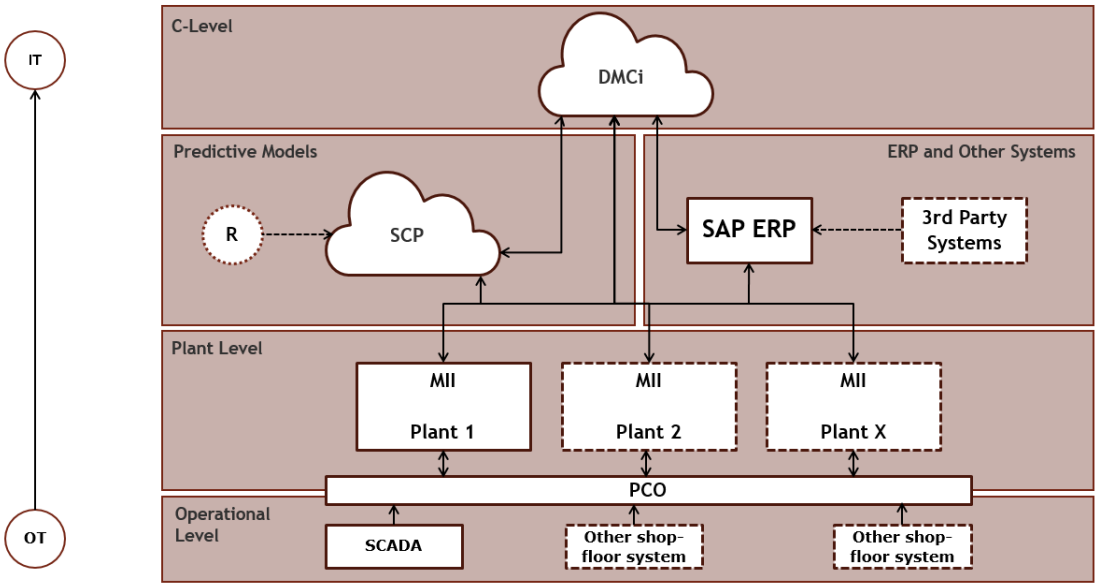

Figure 7.1: Proposed architecture for the developed solutions deployment

The plant's SCADA supervision system at the operational level will connect directly to the SAP Plant Connectivity (PCO) tool that enables the exchange of data between the SAP system and the operational data sources. Two of the advantages this tool presents are to be able to communicate with different software of the different machines on the shop-floor. Also, be configured in one or more locations, allowing the solution to be scalable to the different processes of the different factories of the company, establishing the connection with the plant level. At this level, the SAP Manufacturing Integration and Intelligence (MII) tool is implemented, which is an application that synchronizes manufacturing operations with back-office business processes and standardize data. It functions as a bidirectional data hub between the operational applications, providing analytics and workflow tools to identify problems in the production process, and improving its performance, bringing more transparency to the production processes. The integration component uses Web standards to link SAP Enterprise Resource Planning (ERP) and related business applications with the shop-floor applications in a consistent user interface. It is intended to provide real-time analytics of operations through intelligence, using dashboards and visualization tools to

show alerts and key performance indicators (KPI). Operation managers, business executives, plant employees, and suppliers can use these features to make short-term production decisions.

The SAP Cloud Platform (SCP) integrates data and business processes with a bidirectional connection to the MII. It is a cloud computing platform that enables the creation of new applications or extending existing ones, connecting to cloud-based systems running SAP or third-party software like R. The predictive models developed in R will be integrated into the SCP. MII sends the operational inputs to the SCP, and its outputs are sent back to it, enabling the visualization of real-time information in the MII interface.

In parallel, the SAP ERP incorporates the critical business functions of the company through a bidirectional connection with the MII. The SAP ERP consists of several modules, including plant maintenance, human resources, and financial accounting. The SAP ERP tool allows an easier global integration even with 3rd party systems, providing real-time information, reducing the possibility of errors, creating a more efficient work environment for employees. The client already uses the SAP ERP tool for corporate functions such as accounting, human resources, asset maintenance, and production. One of the advantages is that the user interface is entirely customizable, allowing end users to dictate the operational structure of the tool.

Finally, the SAP Digital Manufacturing Cloud for Insights (DMCi) is a centralized, cloud-based, data-driven manufacturing performance management tool. It enables key stakeholders of manufacturing operations at the chief's level (C-level) to make tactical and strategical decisions that help the company achieve better manufacturing performance. In this centralized tool, the data is collected locally in each MII, providing flexibility in the configuration of new KPI and offering real-time, end-to-end visibility above the manufacturing operations. This configurable solution empowers the top management with intuitive analytics to assess global and plant level performance, improved visibility, and reporting consistency and faster insights to action, achieving global transparency across multiple diverse systems.

Overall, all the objectives outlined for this project were successfully achieved, so in the future, given the potential for scalability of the tool, it will be possible to replicate the solution, by digitally transforming the remaining industrial processes. The approach used proved to be adequate to the optimization of the solvent extraction process and the improvement of the quality indexes.

Digital transformation in companies presents numerous challenges, so it is essential to be aware and open-minded for this shift from operations to a connected and intelligent environment between people, machines, and processes. The implementation of the project resented an effect of resistance to change, so this new culture and routine of workers will be one of the most significant challenges raised by this project. Although this resistance was not possible to quantify, its effect was perceived and reflected in all interactions with the operators. The implementation of these intelligent systems support the operations, whose objective is to increase the human capacity and to provide information to operators that otherwise would not be possible, such as the identification of trends, predictive models, and optimal operational parameters. The company's top management has always shown openness and desire to have this type of project, always aware of these challenges. For this reason, they started this pilot project, in the scope of this dissertation, in order to

guarantee success before climbing to the remaining processes and manufacturing plants.

Appendix A

Annex A

A.1 Laboratory measurements records available

Table A.1: Laboratory measurements record 1/2

Date / Time	Seed Origin	Seed Type	Hexane Concentration (ppm)
3-1-19 7:00	USA	44	235
8-1-19 7:00	USA	47,5	351
10-1-19 7:00	USA	44	645
15-1-19 7:00	USA	44	880
17-1-19 7:00	USA	47,5	720
22-1-19 7:00	USA	44	631
24-1-19 7:00	USA	47,5	692
31-1-19 7:00	USA	47,5	697
5-2-19 7:00	USA	44	982
7-2-19 7:00	USA	47,5	1021
12-2-19 7:00	USA	47,5	998
14-2-19 7:00	USA	47,5	885
19-2-19 7:00	USA	47,5	1310
21-2-19 7:00	USA	47,5	845
28-2-19 7:00	USA	44	434
7-3-19 7:00	USA	47,5	563
12-3-19 7:00	USA	47,5	975
14-3-19 7:00	USA	44	802
19-3-19 7:00	USA	47,5	661
21-3-19 7:00	Brasil	47,5	563
26-3-19 7:00	Brasil	44	1002
28-3-19 7:00	Brasil	44	881
2-4-19 7:00	Brasil	47,5	995
4-4-19 7:00	Brasil	44	349
9-4-19 19:00	USA	44	961
11-4-19 7:00	USA	44	1234
16-4-19 7:00	USA	44	1181
16-4-19 13:00	USA	44	869
17-4-19 13:00	USA	44	710
29-4-19 7:00	USA	47,5	719
30-4-19 7:00	USA	44	798

Table A.2: Laboratory measurements record 2/2

Date / Time	**Seed Origin**	**Seed Type**	**Hexane Concentration (ppm)**
2-5-19 7:00	USA	44	1024
3-5-19 7:00	USA	44	1078
6-5-19 7:00	USA	44	1428
6-5-19 18:00	USA	44	1102
7-5-19 7:00	USA	47,5	1386
9-5-19 7:00	Brasil	44	1638
9-5-19 20:00	Brasil	47,5	970
10-5-19 7:00	Brasil	44	1222
10-5-19 13:00	Brasil	44	1133
10-5-19 19:00	Brasil	44	1249
14-5-19 7:00	Brasil	44	1638
15-5-19 7:00	Brasil	44	1411,7
15-5-19 17:30	Brasil	44	917
16-5-19 7:00	Brasil	47,5	890
21-5-19 7:00	Brasil	44	1005
23-5-19 7:00	Brasil	44	1025
27-5-19 17:30	Brasil	44	531,5
29-5-19 7:00	Brasil	47,5	742,8
30-5-19 7:00	Brasil	47,5	428,1
31-5-19 7:00	Brasil	44	408,3

A.2 Minimum, mean and maximum values of each variable in each group

Table A.3: Minimum, mean and maximum values of each variable for each hexane concentration range

Variables	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
SC_fan_9	36	37,0	38	32	36,7	38	25	36,4	38	26	36,3	38	28	36,5	39
SC_fan_10	34	37,0	38	29	36,7	38	24	35,9	38	26	36,3	38	0	36,1	38
SC_fan_11	34	36,7	38	28	36,4	39	25	35,6	38	26	36,0	39	25	36,2	39
SC_valv_4	18	23,8	29	15	23,7	32	15	24,9	50	15	24,7	42	15	23,7	50
SC_valv_5	15	24,0	44	15	22,7	50	15	27,0	50	15	26,0	50	14	23,2	50
SC_valv_6	16	17,2	23	16	17,3	27	16	18,5	30	16	18,8	31	0	18,7	29
SC_valv_7	0	2,7	6	0	2,6	20	0	2,4	8	0	2,2	20	-6	2,9	19
SC_valv_8	12	14,0	15	11	13,7	16	11	13,9	17	0	13,7	18	0	14,5	20
SC_valv_12	13	14,8	18	13	15,1	21	13	15,5	22	-2	15,6	22	-1	15,5	22
LIT_prod_4	36	42,3	49	34	39,9	53	34	42,4	54	34	42,0	54	28	39,3	55
LIT_prod_5	37	45,5	50	31	41,0	60	26	44,1	75	25	43,9	73	22	41,6	67
LIT_prod_6	36	48,3	54	32	46,3	57	23	45,5	59	23	44,7	58	21	41,9	54
LIT_prod_7	43	56,9	59	21	54,6	60	24	51,0	63	21	50,9	64	21	47,1	62
LIT_prod_8	32	53,8	57	24	50,8	57	24	47,8	62	19	48,0	62	23	44,6	58
LIT_prod_12	21	27,1	36	19	26,2	41	20	27,8	39,5	17	28,8	46	18	29,8	49
TT_gasOut_1	73	73,6	76	73	73,4	77	73	73,4	76	73	73,6	80	73	73,8	87
PT_gasOut_1	-92	-59,0	-20	-91	-58,8	-18	-92	-52,2	-10	-91	-49,3	-7	-90	-48,0	-10
IT_motor	330	364,2	393	311	355,4	393	311	353,8	393	268	347,3	393	289	339,7	393
TT_prod_1	74	77,3	80	74	76,7	81	72	76,7	80	73	77,2	83	74	77,4	90
TT_prod_5	105	105,7	106	104	105,4	106	105	105,4	106	104	105,4	107	104	105,4	107
TT_prod_6	103	105,4	107	102	104,1	107	101	104,5	107	102	105,0	108	102	105,3	107
TT_prod_7	106	106,5	107	105	106,2	116	105	106,2	108	105	106,3	117	105	106,5	117
TT_prod_8	106	106,8	107	105	106,4	108	105	106,5	108	105	106,5	108	104	106,7	111
TT_prod_9	43	56,2	63	42	54,0	62	42	52,3	63	41	51,9	68	42	54,9	70
TT_prod_10	39	50,1	55	41	49,6	54	41	49,0	55	42	48,9	59	42	50,0	58
TT_prod_11	31	48,8	54	31	48,1	54	30	44,7	54	32	45,6	61	28	47,7	57
TT_prod_12	25	36,1	40	26	35,4	41	23	33,3	45	26	33,4	44	24	35,3	46
TT_floor_9	94	115,6	127	77	114,3	128	76	113,6	128	77	112,5	128	93	116,5	129
TT_floor_10	25	105,3	128	25	106,9	132	27	106,8	131	24	100,5	132	25	100,4	132
TT_floor_11	20	86,4	109	20	82,4	113	19	51,1	112	19	49,9	115	20	75,7	113
PT_floor_6	133	208,1	296	44	173,9	337	14	182,5	357	9	186,9	448	27	197,2	399
PT_floor_9	-6	1,8	13	-24	1,2	24	-44	2,0	24	-31	1,6	25	-26	1,9	26
PT_floor_12	347	409,1	448	200	397,1	447	196	378,0	436	197	375,4	445	154	392,9	444
TT_bomba	53	57,8	63	52	57,9	64	49	58,1	65	48	58,9	67	43	58,5	68

Appendix B

Annex B

B.1 Data understanding correlation matrix

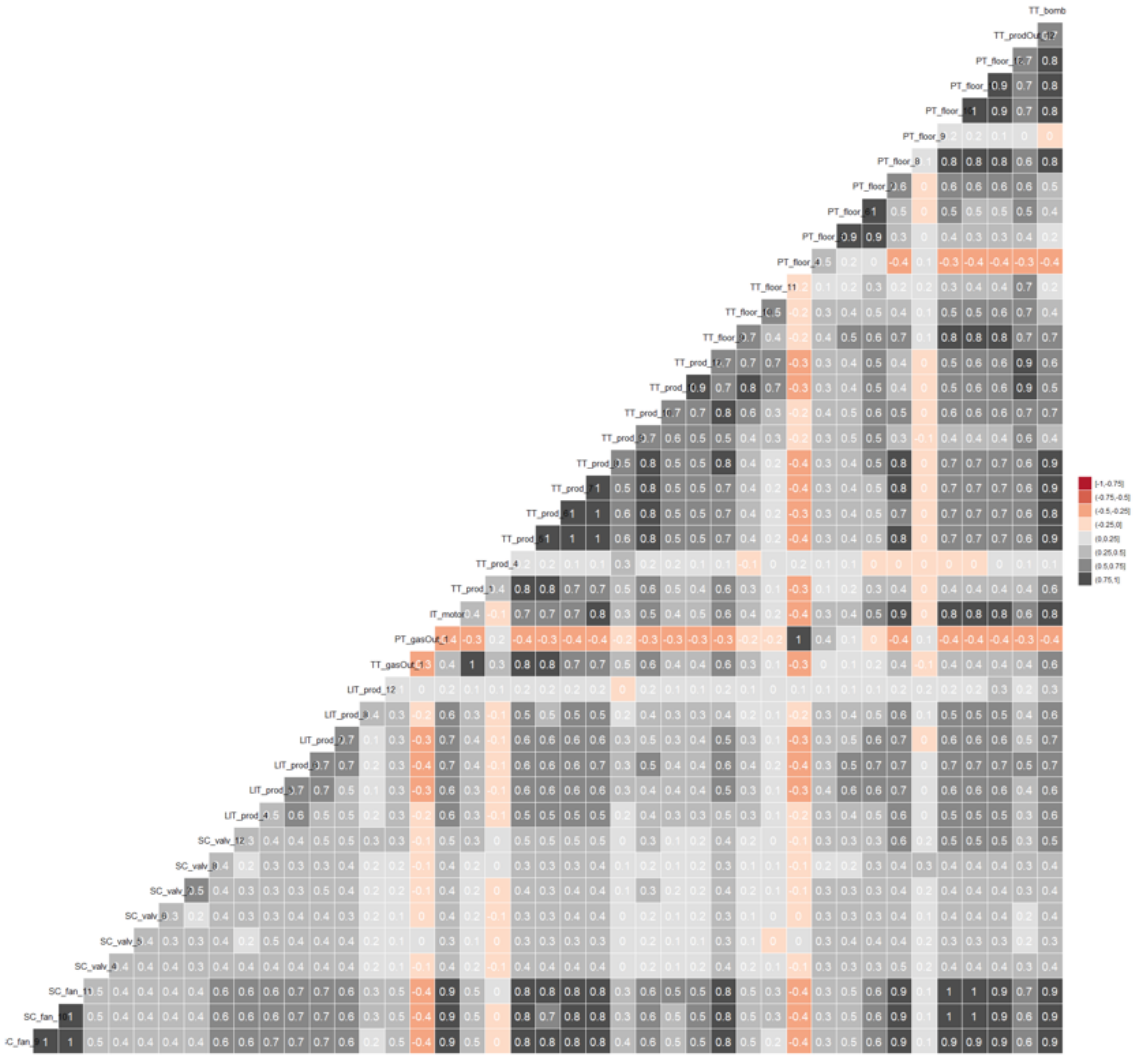

Figure B.1: Data Understanding Correlation Matrix

B.2 Data preparation correlation matrix

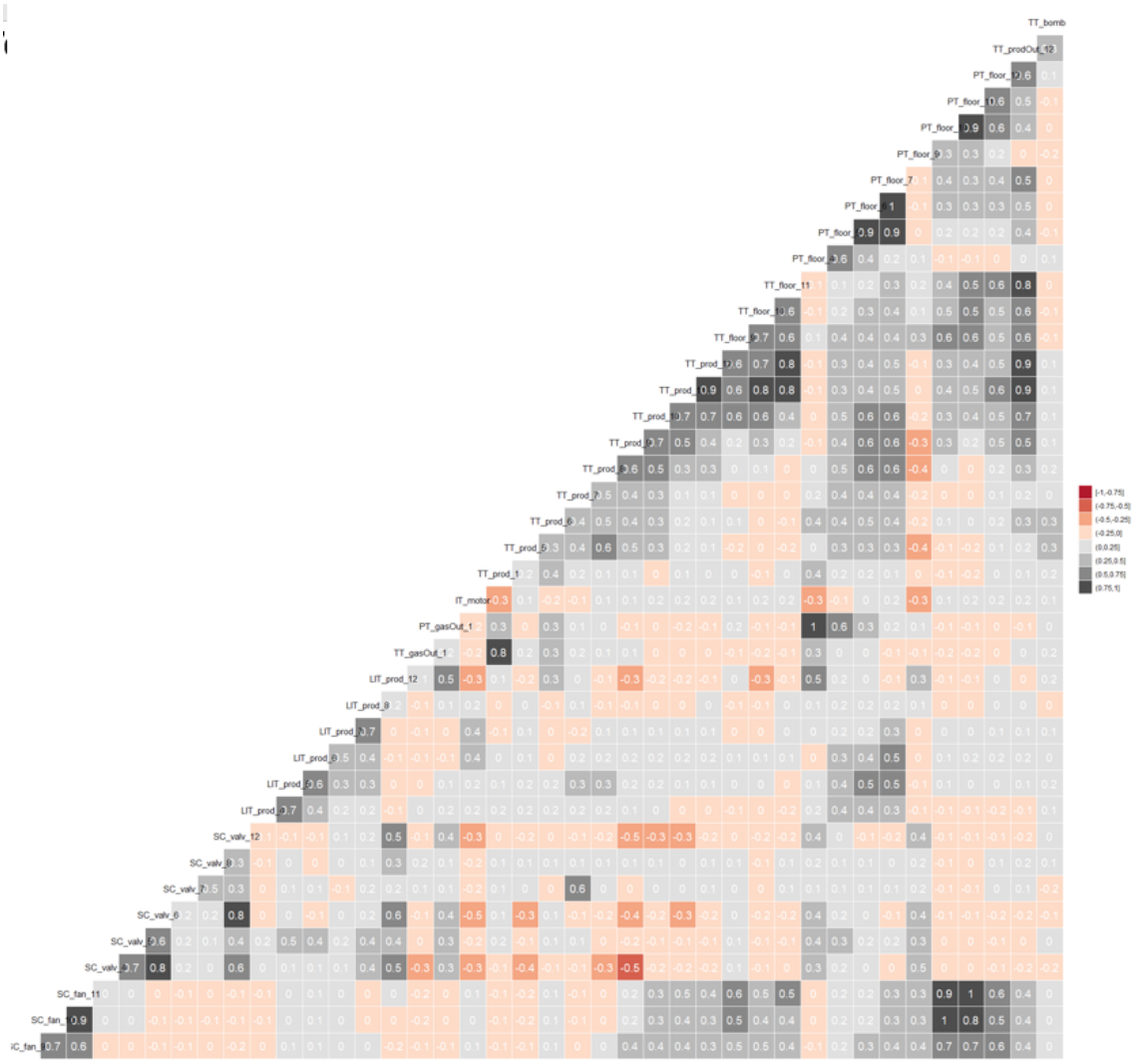

Figure B.2: Data Preparation Correlation Matrix

B.3 Main dataset correlation matrix

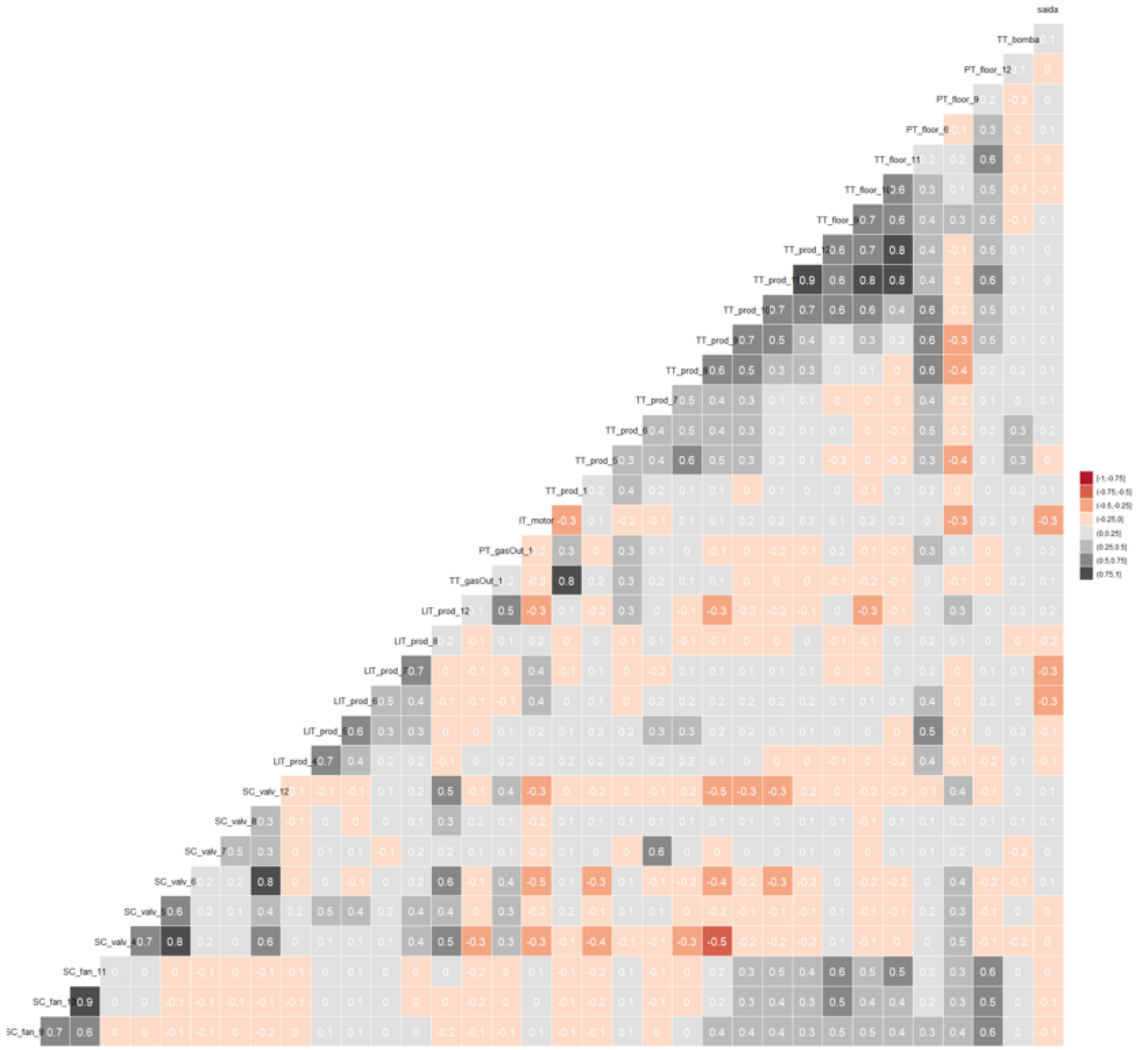

Figure B.3: Main dataset correlation matrix

B.4 USA dataset correlation matrix

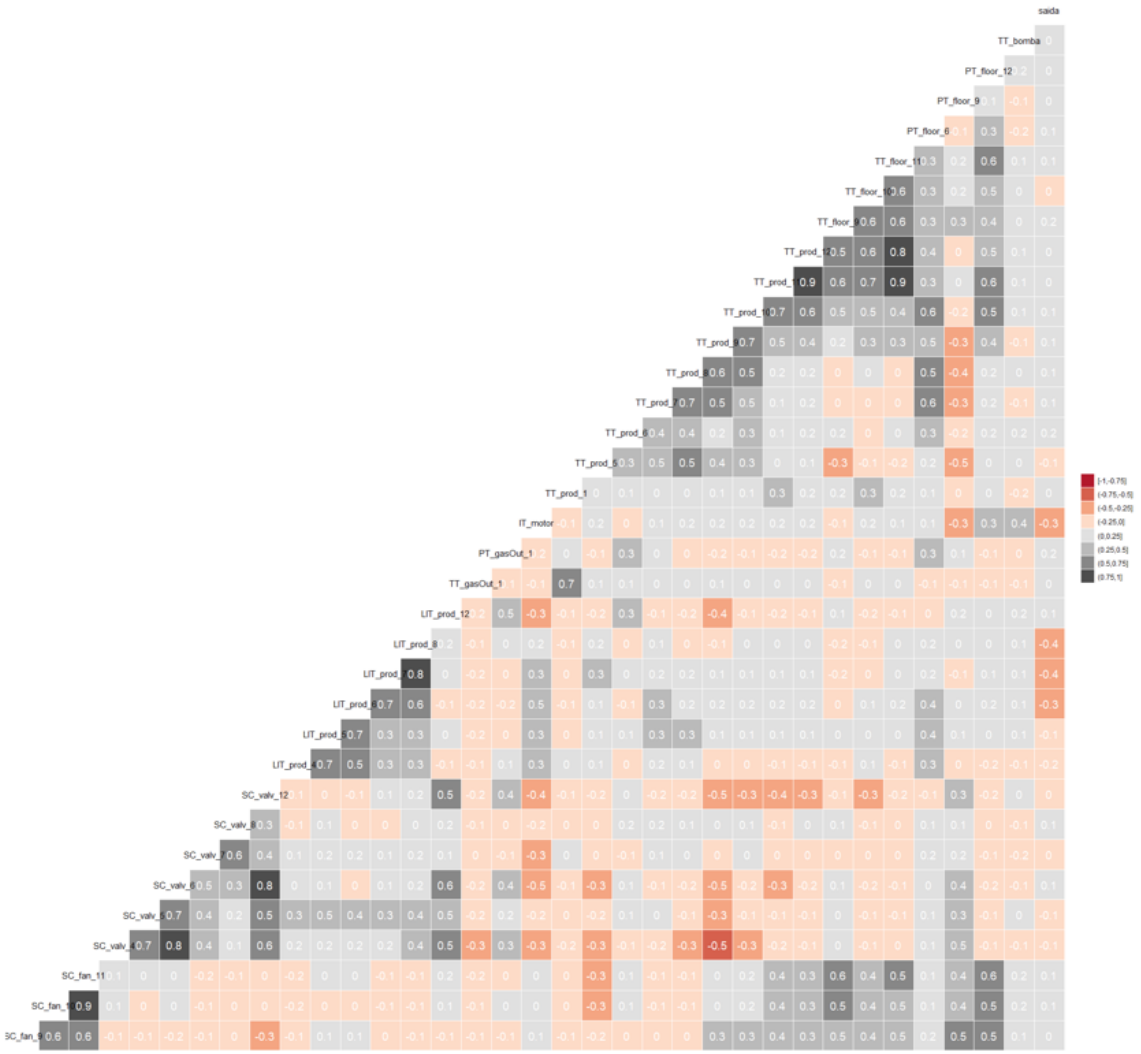

Figure B.4: USA subset correlation matrix

B.5 Brasil dataset correlation matrix

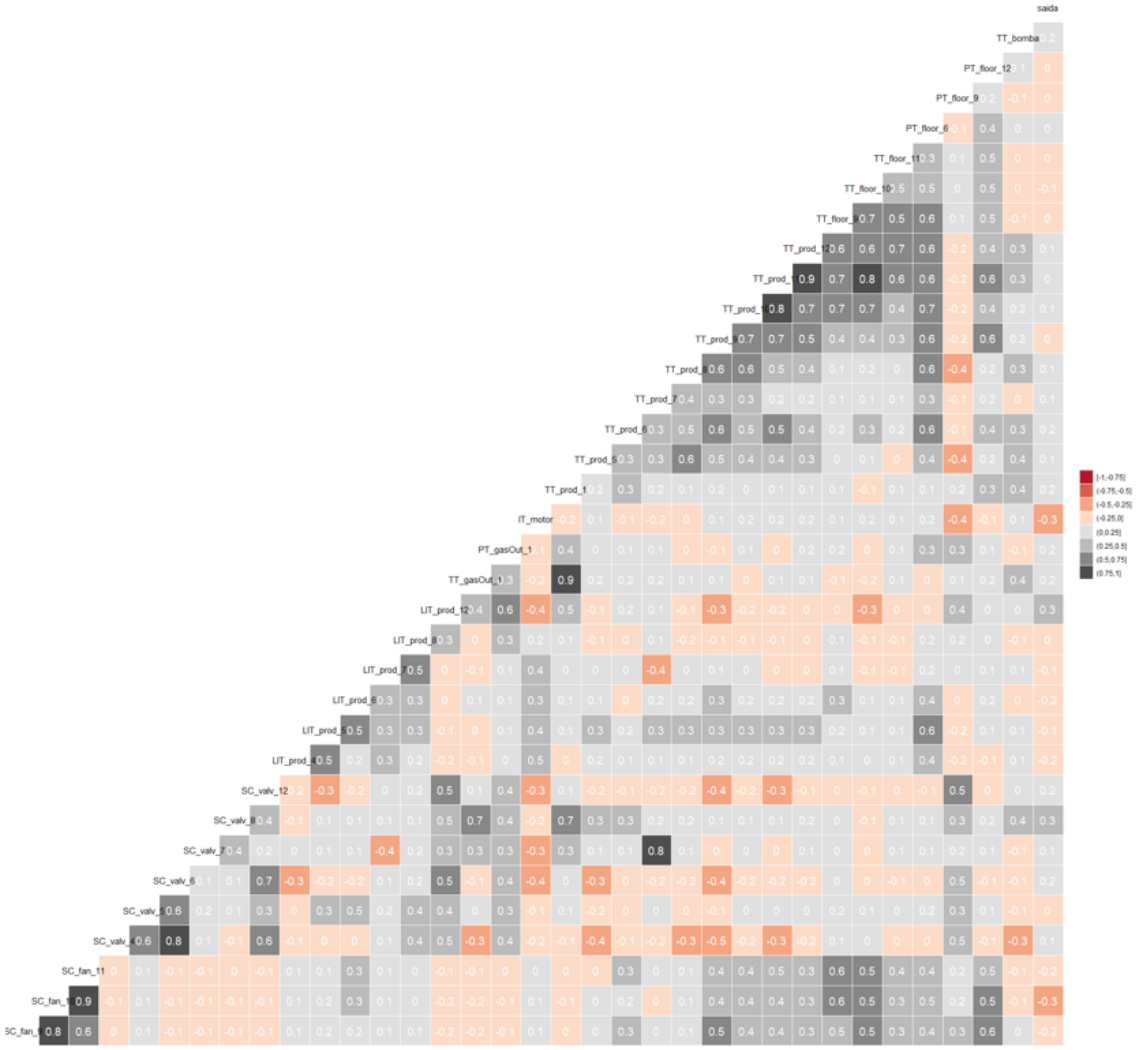

Figure B.5: Brasil subset correlation matrix

Appendix C

Annex C

C.1 Client DT/DC supervisory system

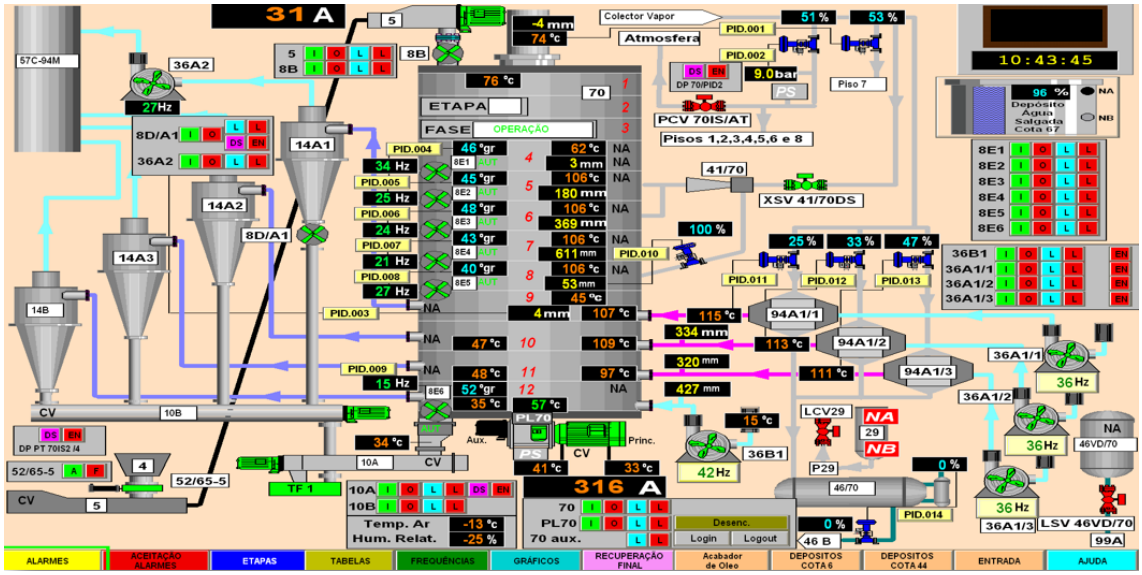

Figure C.1: Client DT/DC Supervisory System

square

References

[1] FIPA. *Evolução das Exportações*. 2018. Available at https://www.fipa.pt/estatisticas/exportacoes-e-importacoes-industria-alimentar, accessed the last time on June 22, 2019.

[2] Randy Kerber Thomas Khabaza Thomas Reinartz Colin Shearer Rudiger Wirth Pete Chapman, Julian Cinton. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM Consortium, August 2000.

[3] Mike Deloso Tim Gaus, Ken Olsen. Synchronizing the digital supply network. May 2018. Available in https://www2.deloitte.com/insights/us/en/focus/industry-4-0/artificial-intelligence-supply-chain-planning.html.

[4] jana Suklan Nusa Erman, Alex Korosec. Performance of selected agglomerative hierarchical clustering metods. January 2015. Available in https://www.researchgate.net/publication/276332381_PERFORMANCE_OF_SELECTED_AGGLOMERATIVE_HIERARCHICAL_CLUSTERING_METHODS.

[5] Matthias Scholz. Pca - principal component analysis. Available in http://www.nlpca.org/pca_principal_component_analysis.html, accessed the last time on June 22, 2019.

[6] R. Berwick. An idiot's guide to support vector machines (svms). Available in http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf, accessed the last time on June 22, 2019.

[7] Timothy G. Kemper. Meal desolventizing, toasting, drying and cooling, 2018. Available in http://lipidlibrary.aocs.org/edible-oil-processing/meal-desolventizing-toasting-drying-and-cooling.

[8] INE. *Ficha Tecido Empresarial - Indústrias Alimentares e de Bebidas*, accessed the last time on june 22, 2019. 2017.

[9] Peter Zuurbier Jacques Trienekens. *Quality and Safety Standards in the Food Industry, Developments and Challenges*. February 2007. Available at https://www.researchgate.net/publication/222216414_Quality_and_Safety_Standards_in_the_Food_Industry_Developments_And_Challenges.

[10] Forbes. America's largest private companies. Available in https://www.forbes.com/largest-private-companies/list/, accessed the last time on June 22, 2019.

[11] Deloitte, 2019. Available in https://www2.deloitte.com/pt/pt.html#, accessed the last time on June 22, 2019.

[12] ENEI. *Diagnóstico de Apoio às Jornadas de Reflexão Estratégica*. 2014. Available at https://www.fct.pt/esp_inteligente/docs/AgroAlimentar_ENEI_Aveiro.pdf, accessed the last time on June 22, 2019.

[13] Ana Cordeiro de Sá. *Crédito y Caución prevê crescimento de 1,8% em 2017 no agro-alimentar português*. January 2017. Available at http://agriculturaemar.com/credito-y-caucion-preve-crescimento-18-2017-no-agro-alimentar-portugues/, accessed the last time on June 22, 2019.

[14] FIPA. *Dados Macroeconómicos*. 2016. Available at http://www.fipa.pt/estatisticas/dados-macroeconomicos-industria-alimentar, accessed the last time on June 22, 2019.

[15] SISAB. *Setor Agro-alimentar*. 2017. Available at https://www.sisab.pt/_sisab/website/pt/setor_aa.html, accessed the last time on June 22, 2019.

[16] Compete 2020. *O COMPETE 2020 alavancou 113 milhões de euros de investimento no setor agro-alimentar, o 2.º maior empregador em Portugal*. September 2016. Available at http://www.poci-compete2020.pt/destaques/detalhe/Setor_agroalimentar_COMPETE2020, accessed the last time on June 22, 2019.

[17] James Hull. The second industrial revolution: The history of a concept. *Storia Della Storiografia*, 1999.

[18] Bessemer steel and its effect on the world. *Scientific American*, 78(13):198–198, 1898.

[19] Jeremy Rifkin. *The Third Industrial Revolution: How Lateral Power is Transforming Energy, the Economy nd the World*. St. Martin's Press, First edition, 2011.

[20] Klaus Schwab. *The Fourth Industrial Revolution*. Crown Publishing Group, 2017.

[21] Brenna Sniderman Mark Cotteleer. Forces of change: Industry 4.0. December 2017. Available at https://www2.deloitte.com/insights/us/en/focus/industry-4-0/overview.html, accessed the last time on June 22, 2019.

[22] Louis Rassey John Nanry, Subu Narayanan. Digitizing the value chain. March 2015. Available at https://www.mckinsey.com/business-functions/operations/our-insights/digitizing-the-value-chain, accessed the last time on June 22, 2019.

[23] Prem Prakash Jayaraman Arkady Zaslavsky. The internet of things: Discovery in the internet of things. October 2015. Available in https://www.researchgate.net/publication/283661119_The_internet_of_things_Discovery_in_the_internet_of_things.

[24] John Salomon. Understanding different types of cloud computing and their benefits, 2018. Available in https://www.chargebee.com/blog/understanding-types-cloud-computing/, accessed the last time on June 22, 2019.

[25] Kamran Meer. *Best Practices in ERP Software Applications: Accounting, Supply Chain Planning, Procurement, Inventory*. iUniverse, First edition, 2005.

[26] Valentin Kalinov. Cryptography blockchain - part 1. December 2017. Available in https://blockchainhub.net/blog/blog/cryptography-blockchain-part-1/, accessed the last time on June 22, 2019.

[27] Edward A. Lee. Cyber-physical systems - are computing foundations adequate? October 2006. Available in https://ptolemy.berkeley.edu/publications/papers/06/CPSPositionPaper/Lee_CPS_PositionPaper.pdf.

[28] Stamatis Karnouskos Paulo Leitão, Armando Walter Colombo. Industrial automation on cyber-physical systems technologies: Prototype implementations and challenges. April 2015. Available in https://bibliotecadigital.ipb.pt/bitstream/10198/15390/1/21.pdf.

[29] Usama Fayyad Johannes Gehrke Jiawei Han Shinichi Morishita Gregory Piatetsky-Shapiro Wei Wang Soumen Chakrabarti, Martin Ester. Data mining curriculum: A proposal. April 2006. Available in https://www.kdd.org/exploration_files/CURMay06.pdf, accessed the last time on June 22, 2019.

[30] Mehmed Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition: Concepts, Models, Methods, and Algorithms*. Wiley-IEEE Press, Second edition, 2011.

[31] Padhraic Smyth Usama Fayyad, Gregory Piatetsky-Shapiro. From data mining to knowledge discovery in databases. January 1997. Available in https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf, accessed the last time on June 22, 2019.

[32] Stephen Nabareseh. Predictive analytics: a data mining technique in customer churn management for decision making. February 2017.

[33] Mohammed Sunasra. Performance metrics for classification problems in machine learning. November 2017. Available in https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b accessedthelasttimeonJune22,2019.

[34] Balamurali M. Classification and regression evaluation metrics — part 1. August 2018. Available in https://medium.com/@balamurali_m/classification-and-regression-evaluation-metrics-part-1-17e6efbe3bf4, accessed the last time on June 22, 2019.

[35] Alvira Swalin. Choosing the right metric for evaluating machine learning models — part 1. April 2018. Available in https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4, accessed the last time on June 22, 2019.

[36] George Athanasopoulos Rob J. Hyndman. Forecasting principles and practive: Evaluating forecast accuracy, April 2018. Available in https://otexts.com/fpp2/accuracy.html, accessed the last time on June 22, 2019.

[37] Lawrence O. Hall W. Philip Kegelmeyer Nitesh V. Chawla, Kevin W. Bowyer. Smote: Synthetic minority over-sampling technique, February 2006. Available in https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume16/chawla02a-html/chawla2002.html, accessed the last time on June 22, 2019.

[38] Jason Brownlee. A gentle introduction to k-fold cross-validation. May 2018. Available in https://machinelearningmastery.com/k-fold-cross-validation/, accessed the last time on June 22, 2019.

[39] Jason Brownlee. Bagging and random forest ensemble algorithms for machine learning. April 2016. Available in https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/, accessed the last time on June 22, 2019.

[40] Anuj Karpatne Vipin Kumar Pang-Ning Tan, Michael Steinbach. *Introduction to Data Mining*. Pearson, Second edition, 2018.

[41] Stephen C. Johnson. Hierarchical clustering schemes. January 1967. Available in https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html, accessed the last time on June 22, 2019.

[42] Israël C. Lerman. Two methods of non-hierarchical clustering. March 2016. Available in https://link.springer.com/chapter/10.1007/978-1-4471-6793-8_2, accessed the last time on June 22, 2019.

[43] Fran¸cois Poulet Edwige Fangseu Badjio. Dimension reduction for visual data mining. January 2005. Available in http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.487.5266&rep=rep1&type=pdf, accessed the last time on June 22, 2019.

[44] Michel Verleysen John A. Lee. *Nonlinear Dimensionality Reduction*. Springer, First edition, 2007.

[45] Leona S. Aiken Patricia Cohen, Stephen G. West. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Psychology Press, Second edition, 2014.

[46] Onel Harrison. Machine learning basics with the k-nearest neighbors algorithm. September 2018. Available in https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761, accessed the last time on June 22, 2019.

[47] Peter Flom. Stopping stepwise: Why stepwise selection is bad and what you should use instead. September 2018. Available in https://towardsdatascience.com/stopping-stepwise-why-stepwise-selection-is-bad-and-what-you-should-use-instead-
accessed the last time on June 22, 2019.

[48] Will Kenton. Multiple linear regression – mlr definition. April 2019. Available in https://www.investopedia.com/terms/m/mlr.asp, accessed the last time on June 22, 2019.

[49] Luis Fernando Rainho Alves Torgo. Inductive learning of tree-based regression models. September 1999. Available in http://www.dcc.fc.up.pt/~ltorgo/PhD/.

[50] Raul Eulogio. Introduction to random forests. August 2017. Available in https://www.datascience.com/resources/notebooks/random-forest-intro, accessed the last time on June 22, 2019.

[51] Lujing Chen. Introduction to random forests. January 2019. Available in SupportVectorMachine\protect\begingroup\

immediate\write\@unused\def\MessageBreak˙\let\protect\
edefYourcommandwasignored.\MessageBreakTypeI<command><return>
toreplaceitwithanothercommand,\MessageBreakor<return>
tocontinuewithoutit.\errhelp\let\def\MessageBreak˙(inputenc)
\def\errmessagePackageinputencError:UnicodecharâĂŁ(U+
200A)\MessageBreaknotsetupforusewithLaTeX.
˙˙Seetheinputencpackagedocumentationforexplanation.
˙TypeH<return>forimmediatehelp\endgroup\T1\textemdash\protect\
begingroup\immediate\write\@unused\def\MessageBreak˙\let\
protect\edefYourcommandwasignored.\MessageBreakTypeI<command>
<return>toreplaceitwithanothercommand,\MessageBreakor<return>
tocontinuewithoutit.\errhelp\let\def\MessageBreak˙(inputenc)
\def\errmessagePackageinputencError:UnicodecharâĂŁ(U+
200A)\MessageBreaknotsetupforusewithLaTeX.
˙˙Seetheinputencpackagedocumentationforexplanation.
˙TypeH<return>forimmediatehelp\endgroupSimplyExplained, accessed the last time on June 22, 2019.

[52] Ajay Yadav. Support vector machines(svm). October 2018. Available in https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589, accessed the last time on June 22, 2019.

[53] Seth Rogers Stefan Schroedl Kiri Wagstaff, Claire Cardie. Constrained k-means clustering with background knowledge. September 2001. Available in https://web.cse.msu.edu/~cse802/notes/ConstrainedKmeans.pdf.

[54] The R Foundation. What is r? Available in https://www.r-project.org/about.html, accessed the last time on June 22, 2019.