Universidade do Minho
School of Engineering

Pedro Miguel Ferreira Ribeiro

Machine Learning Applied to Companies Management

October, 2022

Universidade do Minho
School of Engineering

Pedro Miguel Ferreira Ribeiro

Machine Learning Applied to Companies Management

Master Thesis

Master in Informatics Engineering

Specialization in Artificial intelligence

Work developed under the supervision of:

Rui Manuel Ribeiro Castro Mendes

October, 2022

STATEMENT OF INTEGRITY

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

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

Barcelos, 31 October 2022
(Place) (Date)

Pedro Miguel Ferreira Ribeiro
(Pedro Miguel Ferreira Ribeiro)

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

License granted to the users of this work

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.10.10) [1].

Acknowledgements

First and foremost, I'd like to thank PRIMAVERA for providing ideal conditions for the development of this project, particularly my supervisors, Orlando Rocha and Miguel Domingues. I'd also like to thank my advisor at the University of Minho, Rui Mendes, for his support and guidance.

To my parents, Elvira and Domingos, and to my sister, Andreia, for all the love, patience, affection, and trust you have placed in me, and for everything you have done and continue to do for me. My love, Inês, for all the love and for always believing in me.

To my old friends and those Bragança gave me, thank you for sharing unforgettable moments and stories with me.

Finally, at the University of Minho, all of its faculty members, for having made these two years a path of full knowledge.

“The true master is an eternal student.” (Yi)

Resumo

Machine Learning Aplicado à Gestão de Empresas

O mais recente progresso reconhecido pela comunidade europeia por Indústria 5.0 atende as evoluções imergentes no mundo da indústria e revela haver uma necessidade de evolução nos sistemas de ERP's (Enterprise Resource Planning). É esperado que estes sistemas que auxiliem na gestão das empresas de uma forma mais dinâmica, assim tornando-se mais autónomos, atendendo a toda a informação aglomerada no sistema.

O PRODUTECH, formalmente conhecido como o "Cluster" português das Tecnologias de Produção, é uma rede estabelecida por empresas de tecnologia de produção capazes de reagir às dificuldades do sector de produção com soluções criativas, adaptáveis, integradas e competitivas. Incluído dentro deste consórcio está a PRIMAVERA, sendo uma empresa portuguesa pioneira no desenvolvimento de soluções de gestão, mais particularmente, sistemas ERP. Estes sistemas constam com elevados volumes de dados, o que poderá levar a operações complexas no tratamento e análise dos mesmos.

Neste relatório de dissertação de mestrado é documentada uma visão de solução para dar resposta a assistência computacional na escolha de fornecedores, bem como a aglomeração dos conhecimentos dos diversos conceitos que envolvem a Indústria, visando a exploração de técnicas de data science e de machine learning.

Esta implementação teve em consideração aspetos importantes na gestão estratégica das empresas atendendo as necessidades das mesmas, sendo possível a adaptação da solução para cada caso em particular.

Palavras-chave: Seleção de fornecedores, Inteligência Artificial, Gestão da cadeia de fornecimento, Machine learning

Abstract

Machine Learning Applied to Companies Management

The most recent advancement recognized by the European community as Industry 5.0 attends the immerging evolutions in the industry world and reveals a need for evolution in the ERP's (Enterprise Resource Planning) systems. It is expected that these systems will assist in the management of companies in a more dynamic way, thus becoming more autonomous, attending to all the information agglomerated in the system.

PRODUTECH, formally known as the "Portuguese Cluster of Manufacturing Technologies", is a network established by production technology companies capable of responding to industry challenges with creative, adaptable, and integrated solutions. PRIMAVERA, a pioneering Portuguese company in the development of management solutions, particularly ERP systems, is included in this consortium. These systems constantly deal with high volumes of data, which can lead to complex operations in their treatment and analysis.

This master's thesis report documents a proposed solution for providing computational assistance in the selection of suppliers, as well as the aggregation of knowledge from various industry concepts, with an emphasis on data science and machine learning techniques.

This implementation took essential aspects into account in the strategic management of businesses, meeting their needs while allowing for the solution to be amended for each individual case.

Keywords: Supplier selection, Supply chain management, Machine learning, Artificial intelligence

Contents

List of Figures xi

List of Tables xiii

List of Listings xiv

Glossary xv

Acronyms xvi

1 Introduction **1**

1.1 Framework 1

1.2 Objectives and predicted phasing 1

1.3 Summary of performed activities 2

1.3.1 Chronology 2

1.4 Primavera Business Software Solutions 3

1.5 Produtech 3

1.6 Report structure 4

2 Literature review **5**

2.1 The evolution of industry 5

2.2 Supply chain 6

2.2.1 Supply chain 7

2.2.2 Exploring Complex Business Challenges 8

2.2.3 Designing Supply Chain Systems 9

2.3 Multiple-criteria decision analysis 10

2.3.1 Theoretical overview 10

2.3.2 Choosing an appropriate MCDA technique 11

2.3.3 Fuzzy Set Theory 13
2.3.4 Fuzzy SAW 14
2.3.5 Fuzzy VIKOR 14
2.3.6 Fuzzy TOPSIS 14
2.3.7 Supplier Criteria 15
2.4 Process model for data exploration 17
2.5 Artificial Intelligence 18
2.6 Machine Learning 18
2.6.1 Learning models types 19
2.6.2 Key terminology 20
2.6.3 Machine learning model construction procedure 21
2.6.4 Machine learning algorithms 23

3 Software Tools review **27**
3.1 Python 27
3.2 Synthetic data generation 28
3.3 Pyspark 28
3.4 Azure Cloud Services - Databricks 29
3.5 Automatic Machine Learning 29
3.5.1 TPOT 30
3.5.2 H2O - AutoML 30
3.5.3 Databricks AutoML 31

4 Implementation **33**
4.1 Exploration environment 33
4.2 Dataset's 33
4.2.1 COLEP 33
4.2.2 v10 34
4.3 Architectural insight concept 36
4.4 Supplier selection criteria 38
4.5 Custom package development 39
4.6 COLEP development 40
4.6.1 Aggregation of data. 41
4.6.2 Feature Enginnering 43
4.7 v10 development 44
4.7.1 Metrics Calculation 44
4.7.2 Custom package - v10 45
4.7.3 Feature Enginnering 45

5 Results 50
5.1 COLEP Machine Learning 50
5.2 v10 Machine Learning 51
5.2.1 Product restriction. 52
5.2.2 Machine Learning results overview. 53
5.2.3 Models results exploration. 54
6 Conclusions and Future Work 57
6.1 Conclusions 57
6.2 Future 58
Bibliography 59

Appendices

Annexes

I Annex 1 66

List of Figures

1 The Bullwhip Effect on the SC. [19] 9

2 Supply chain management (Traditional system) 10

3 Demand-Driven system 10

4 Fuzzy Number Graph 13

5 Phases of CRISP-DM Process Model for Data Mining. [32] 18

6 High-level components of AI. [33] 18

7 Various types of machine learning techniques. [35] 19

8 ROC-AUC Cuver example 21

9 A decision tree learned for this dataset.[44] 23

10 Linear and Logistic Regression. [50] 26

11 Pandas Profiling short example of an overview on one column 28

12 TPOT example Machine Learning pipeline.[61] 30

14 Databricks representation of glass-box approach. [64] 31

13 H20 AutoML Methodology.[63] 31

15 Architectural concept diagram 37

16 Architecture Data Flow 38

17 Total number of records per year present in the orders dataset. 41

18 Total number of records per year present in the dataset of receptions. 41

19 Total null values present in orders dataframe. 41

20 Total null values present in reception dataframe. 42

21 Total null values after joining the dataframes via *pandas.DataFrame.join*. 42

22 Diagram that represents the synthetic delivery data algorithm. 46

23 Goal of the ranking distribution. 48

24 SHAP Feature value in Machine Learning model. 55

25 Correlation overview of XGBoost model. 56

26 Correlation overview of XGBoost model. 56

27 Variable classification survey. 69
28 Supplier classification survey. 70

List of Tables

1 Chronogram of project management activities 2

2 Fuzzy preference scale 13

3 Selection criteria from literature review:**A**:Dickson (1966); **B**:Wind et al. (1968); **C**:Lehmann and O'Shaughnessy(1974); **D**:Perreault and Russ (1976); **E**:Abratt (1986); **F**:Billesbach et al.(1991); **G**: Weber et al. (1991); **H**, Segev et al. (1998); **I**, Min and Galle (1999); **J**, Stavropolous (2000). [21] 16

4 Confusion matrix for binary classification. 21

5 A dataset describing weather conditions and a target variable (Play Golf?)[44] 23

6 Databricks experiment exemple 32

7 Description and example of one order. 34

8 Description and example of receptions of one order reception. 34

9 Description and example of receptions of one order. 36

10 Supplier selection criteria according to projects needs. 39

11 v10 database matching supplier selection variables 44

12 Mapping of normalized value to ranking 47

13 Comparison of algorithms in the ranking of variables 48

14 Test performed to prove the performance of the restrictions applied. 53

15 Machine learning models runs performance. 54

List of Listings

4.1 SQL querry for data extraction 35
4.2 Code using merge_asof. 43
4.3 Percentage remaining delivery time conditions. 45
4.4 Unique items by supplier. 46
4.5 Call function for ranking system 48
4.6 Example of the use of the function Fuzzy TOPSIS. 49
4.7 Restriction on the absolute number of suppliers per item. 49
5.1 Execution of Databricks AutoML. 51
5.2 Example of experiment instigation and data storage. 51

Glossary

Force majeure Irresistible compulsion or coercion. The phrase is used particularly in commercial contracts to describe events possibly affecting the contract and that are completely outside the parties' control. Such events are normally listed in full to ensure their enforceability; they may include acts of God, fires, failure of suppliers or subcontractors to supply the supplier under the agreement, and strikes and other labour disputes that interfere with the supplier's performance of an agreement. An express clause would normally excuse both delay and a total failure to perform the agreement. *(p. 8)*

Haversine The haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes. Important in navigation, it is a special case of a more general formula in spherical trigonometry, the law of haversines, that relates the sides and angles of spherical triangles. *(p. 45)*

SCM SCM encompasses the planning and management of all activities involved in sourcing and procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third-party service providers, and customers. In essence, SC management integrates supply and demand management within and across companies. SCM is an integrating function with primary responsibility for linking major business functions and business processes within and across companies into a cohesive and high-performing business model. It includes all the logistics management activities noted above, as well as manufacturing operations, and it drives coordination of processes and activities with and across marketing, sales, product design, finance, and information technology.[2] *(p. 7)*

Acronyms

AI	Artificial intelligence *(pp. 1, 11, 18, 19)*
AUC	Area under curve *(p. 21)*
AutoML	Automatic Machine Learning *(pp. 29–31, 37, 50, 51, 54, 55, 57)*
CCG	Centro de Computação Gráfica *(pp. 2, 50)*
CCi	Closeness Coefficient index *(pp. 15, 55)*
CLM	Council of Logistics Management *(p. 7)*
CRISP-DM	Cross-Industry Standard Process for Data Mining *(p. 17)*
CSCMP	Council of Supply Chain Management Professionals *(p. 7)*
DL	Deep Learning *(p. 18)*
DM	Decision Maker *(pp. 10, 12–15, 36, 47, 58)*
EFB	Exclusive Feature Bundling *(p. 25)*
ERP	Enterprise resource planning *(pp. 37, 58)*
FNIS	Fuzzy Negative Ideal Solution *(pp. 14, 15)*
FPIS	Fuzzy Positive Ideal Solution *(pp. 14, 15)*
GBDT	Gradient boosting decision tree *(p. 24)*
GOSS	Gradient-based One-Side Sampling *(p. 25)*
IFTT	”if this, then that” *(p. 8)*
LASSO	Least Absolute Selection Shrinkage Operator *(p. 24)*
LightGBM	Light Gradient Boosting Machine *(pp. 25, 31, 32)*

MB	Megabyte *(pp. 34, 35)*
MCDA	Multi-Criteria Decision Analysis *(pp. 10, 11)*
MCDM	Multi-Criteria Decision-Making *(pp. 11, 14)*
ML	Machine Learning *(pp. 1, 2, 5, 18–25, 27–31, 37, 42, 43, 46, 50, 52, 54, 57, 58)*
PBSS	PRIMAVERA Business Software Solutions, S.A. *(pp. 1–3, 33, 37, 50)*
PPS	Products, Processes and Services *(pp. 3, 4)*
PTI	Production Technologies Industry *(p. 3)*
rmse	root-mean-square error *(pp. 20, 50)*
ROC	Receiver operating characteristic curve *(p. 21)*
SAW	Simple Additive Weighting *(p. 14)*
SC	Supply Chain *(pp. 2, 6–9)*
SCM	Supply Chain Management *(pp. 5, 7, 8, 10)*
SHAP	SHapley Additive exPlanations *(pp. 54, 55)*
SS	Supply Selection *(pp. 2, 11, 15)*
SSC	Supplier Selection Criteria *(pp. 11, 38)*
TDIDT	Top-down induction of decision trees *(p. 23)*
TOPSIS	Technique for Order Performance by Similarity to Ideal Solutions *(pp. 14, 36, 39, 47, 48)*
TPOT	Tree-based Pipeline Optimization Tool *(pp. 29, 30, 50)*
UDF	User-defined functutions *(p. 39)*
VIKOR	VIseKriterijumska Optimizacija I Kompromisno Resenje *(p. 14)*
XAI	Explainable AI *(p. 55)*
XGBoost	eXtreme Gradient Boosting *(pp. 24, 25, 31, 32, 50, 53–55, 57)*

1 Introduction

1.1 Framework

This thesis project is proposed by PRIMAVERA Business Software Solutions, S.A. (PBSS), a national reference company active in enterprise solutions development. Since 1993, PBSS computer solutions have facilitated and streamlined recurring processes in the management and conduct of companies.[3]

Within this ecosystem, constant efforts are conducted in the evolution of industries. With the contribution of Experts from research and technology organizations during two virtual workshops held on 2nd and 9th of July 2020, the concept of industry 5.0 was born. Industry 5.0 is set with a vision towards the adoption of a human-centred approach to digital technologies, including Artificial intelligence (AI); reskilling and upskilling of European workers, particularly digital skills; modern, resource-efficient, and sustainable industries and the transition to a circular economy; a globally competitive and world-leading industry, accelerating investment in research and innovation. [4]

PBSS currently uses several in-house developed technology solutions. However, it wants to understand how it can take advantage of the various platforms and Machine Learning (ML) tools currently available.

1.2 Objectives and predicted phasing

The objective of this project is to gather a solid theoretical base that will serve for the development that will aid in the election of the most reasonable suppliers for a specific order. In this way, Table 1 is a Gantt diagram used to illustrate the progress of the different stages of the project. The time intervals designated for each stage describe a milestone in the project's development.

	Months								
Phases	**Oct.**	**Nov.**	**Dec.**	**Jan.**	**Feb.**	**Mar.**	**Apr.**	**May**	**June.**
State-of-the-art assessment	🟩	🟩	🟩	🟩	🟩				
Analysis of the case study and planning of problem-solving strategies		🟩	🟩	🟩	🟩	🟩			
COLEP		▨	▨						
v10				▨	▨	▨			
PRIMAVERA data factory and database analysis				🟩	🟩	🟩			
Development of models and functionalities for supplier selection			🟩	🟩	🟩	🟩	🟩	🟩	
COLEP			▨	▨					
v10					▨	▨	▨	▨	
Writing of the developed work (Thesis report)				🟩	🟩	🟩	🟩	🟩	🟩

Table 1: Chronogram of project management activities

1.3 Summary of performed activities

The following activities were undertaken to complete this report:

The initial phase of this project, which corresponds to phase 1 on Table, started without data relating to the requirement in question, so a theoretical study was carried out on the functioning of the overall industry and scientific articles on multiple subject areas (e.g. Supply Selection (SS), SC, Industry evolution, and needs).

Following the collection of the data, analysis and transformation were performed on the COLEP Database in the second phase, which corresponds to steps 2 and 3 on Table . However, the approach was changed to use the v10 database due to time and information constraints.

In the last phase (step 4), after the data treatment, the ML process was initiated and compared to the intended results. After achieving the desired results, they were analyzed and judged in light of the desired solution, passing through process optimization and restructuring processes.

1.3.1 Chronology

The project development had a duration equal to the protocol period, which was nine months, starting on the 25th of October 2022 until the 30th of June 2022. At the beginning of the process, there was a phase of integration in the intelligence team of PBSS in which he attended the daily meetings of Scrum. On November 9th, 2021, a series of weekly meetings began with the PBSS intelligence department team to monitor the development of the projects of the onboard trainees.

With a weekly recurrence, the first meeting with the Centro de Computação Gráfica (CCG) took place on the 2nd of November 2021 to share project information and requirements.

Regarding the *COLEP* approach, the suppliers and orders dataset was provided in the first month of 2022. This data is of significant importance for the development of this project; Following the receipt of these datasets, the weekly meetings with CCG added elements belonging to the *PRODUTECH* project, allowing for the clarification of common doubts about the data from the company under study.

In the last month of the stage, several meetings were held to present the developed project to the PBSS intelligence team, and information was exchanged to make the project's final adjustments.

1.4 Primavera Business Software Solutions

Headquartered in Braga, PBSS, is engaged in the development and marketing of management solutions and platforms for the integration of business processes in a market that includes small, medium, and large organisations. With the beginning of activity at the end of 1993, it was a pioneer in developing management solutions; currently, it has more than 40 thousand customers spread across nearly 20 countries, being the market leader in many of these countries. Currently, PBSS core business consists of the conception, design, and development of management software solutions and in technical support to customers (commercial partners). A constant effort by the PBSS team to provide management solutions that approach market needs, focusing on two essential points: quality and innovation. The excellence of products and services offered by PBSS is the result of a team of professionals who are highly qualified and motivated. The continuous investment in the permanent evolution of its skills and achievements, anticipating the needs and expectations of customers, companies and the market. The will to surpass itself in pursuing excellence in all its areas of activity, drawing inspiration from the future to innovate. Regular participation in research programs and close collaboration with the academic community. [3]

1.5 Produtech

The mobilizing project *"PRODUTECH sustainable & circular"* started on July 1st, 2020 and is expected to conclude on June 30th, 2023.

This project responds to the Production Technologies Industry (PTI) challenge for the development of a circular and sustainable industry. Speculates the creation of innovative solutions in response to the urgent need for the industry to change in the direction of sustainability and circular economy.[5, 6]

It includes a compelling set of interventions focused on creating tools, applications, processes and strategies supported by the new digitalization paradigms, mainly focuses on production systems and components, zero defects and zero environmental impacts, and symbiotic management of production ecosystems and supply networks in a circular environment. The end result is expected to be the construction of new technologies or high-value services strengthen the international positioning of the production technologies sector, and enhance its qualified integration in global value/innovation chains and knock-on effects for the industry as a whole.

The consortium consists of 26 public and private entities divided into six development points, called Products, Processes and Services (PPS). The 4th is divided into several development activities, of which the prominent members are PBSS and CCG.

This project falls under PPS 4: “Process 4.0 for Intelligent Logistics in the Supply Chain”, and in sub-activity 4.1.6 “Development of decision support tools for intelligent logistics”. This activity has been broken down into the following requirements:

- 6.2: Perform predictive/prescriptive task
  - 6.2.1: Forecast Customer Orders
  - 6.2.2: Predict Production Lead-Time
  - 6.2.3: Forecast Delivery Date
  - 6.2.4: Optimize Demand Plan
  - 6.2.5: Predict Quantity Order to Supplier
  - 6.2.6: Forecast Order Reception Date
  - 6.2.7: Optimize Purchase Plan

This academic research is based on an examination of requirement 6.2.7, focusing solely on analysis and assistance in supplier selection.

1.6 Report structure

This report is structured as follows:

In chapter I, the subject of the dissertation is presented. The evolution of the production line technique industry is highlighted to provide answers to the sector's current situation. Furthermore, the objectives and the methodology adopted for the development of the dissertation are exposed.

In chapter II, state of the art is portrayed, where the information gathered on the subject of study. The concepts involved in the industry and the techniques adopted to obtain effective planning are described with more study.

In chapter, III presents a study of the technologies and tools used in the project's course.

In chapter IV, Implementation testing, architecture and discussion of eventualities and views on the approaches.

In chapter V, exploration of the results obtained and discussion on points of view of chosen approaches to the development process.

In chapter VI, conclusions and recommendations for further development are presented.

2

Literature review

This chapter describes all the literature review related to the Industry 5.0. In order to get an understanding, is necessary to understand the processes involved in this area, thus the SCM ecosystem is described to obtain an overview and understand the framework of the overall project objective. Understanding the challenges facing the business tasks and complexity is an important aspect that also will be described in this chapter. Additionally, technical concepts and explanation of ML models used in the development stage are presented.

2.1 The evolution of industry

Human life has always been based on evolution; all evolutions have historical precedents and exist as a result of advances in science and technology. The first industrial revolution, known as Industry 1.0, began in 1784 with the introduction of mechanized production facilities that used hydraulic energy and vapour energy. Because of the significance of this development for the world, it was necessary to classify it and any subsequent developments. [7]

The application of the concepts of shared labour responsibility and mass production in the 1870s was referred to as the second industrial revolution due to the introduction of electrical energy.

The third evolution occurred after the first programable logic controller was developed in 1969, focusing on automating production through robotics, electronic systems, and the internet. Techniques that were initially developed in the last decades of the *XX* century concluded that the technologies available at the time were not sufficiently advanced and capable of accomplishing the intended goals. However, due to technological advancements in the twenty-first century, these advancements were already technologically capable of accomplishing the goals and achieving an industry-wide evolution. As a result, the industry connected IoT with production techniques; the development of new technologies has been the main driver of the transition to Industry 4.0.

Industry 5.0 is the final historical turning point; because it aims to solve qualities that haven't been defined until now, it can be considered an addition to Industry 4.0 rather than a complete revolution. [8, 9]

It was created due to numerous contributors from funding organizations, research and technology organizations, and organizations throughout Europe participating in two virtual workshops held by the Directorate-General of Research and Innovation on July 2nd and July 9th, 2020.[10]

In the current economic and sociological changes we are living through, European industry is a major force. Industry must take the lead in the digital and green transitions if it wants to continue being the source of wealth.

Industry 5.0 offers a vision of the business that goes beyond the narrow focus on production and efficiency and strengthens the function and value of the industry in society. It leverages new technology to provide prosperity beyond jobs and growth while respecting the planet's production constraints and places the worker's welfare at the centre of the production process.

According to the European Commission, and under its technological framework, there are six key categories for funding Industry 5.0:

1. "Individualized Human-machine-interaction";
2. "Bio-inspired technologies and smart materials";
3. "Digital twins and simulation";
4. "Data transmission, storage, and analysis technologies";
5. "Artificial Intelligence";
6. "Technologies for energy efficiency, renewables, storage and autonomy".

The European Commission also defends that the Major Commission policy efforts currently include the following aspects of Industry 5.0:

- "Improving and retraining European employees' abilities, especially their digital skills";
- "Using artificial intelligence and other digital technologies with a human-centred perspective";
- "Industry that is highly competitive and leading in the globe, accelerating investment in research and innovation"
- "Industries that are contemporary, resource-efficient, and sustainable and the shift to a circular economy";

2.2 Supply chain

The following subsections provide a brief introduction to the concepts of SC and the associated problems and challenges are presented.

2.2.1 Supply chain

After the third industrial revolution (Industry 3.0) and when personal computers entered the business and commercial world, the SC began to take shape. Although the market as we know it today has come a long way, the fundamental act of entrepreneurs purchasing goods from suppliers and selling them to customers has always existed.[11]

Many people try to create definitions for SC, here are some of them:

- "A SC is the alignment of firms that bring products or services to market";[12]
- "A SC consists of all stages involved, directly or indirectly, in fulfilling a customer request. The SC includes not only the manufacturer and suppliers, but also transporters, warehouses, retailers, and customers themselves";[13]

If this describes what a SC is, then it is possible to describe the SCM, including the actions taken to influence a SC's behaviour and the intended results. The corporate and scientific communities strived throughout the several years to come up with a final SCM definition, although it never worked out.

The definition from *Lalonde* says that what a SCM drives is the exchange and flow of information and materials, which is beyond the boundary of the enterprise.[14]

On the other hand, *Martin Christopher* speaks from the point of view of the entire SC, treating the customer in the backward position, and consequently giving more value to the customer at a lower cost, and the supplier in the forward position.[15]

In 2005, the former Council of Logistics Management (CLM) organization has become the Council of Supply Chain Management Professionals (CSCMP) and created a global defenition for the SCM. [16]

Since there is an exponential relationship between the company and the SC, as one grows, so does the other, it is crucial to maintain alignment among the various SC functions. For example, while price negotiations, manufacturing schedules, and logistics management all impact the company, their interdependence can make managing the SC challenging. [17]

For the business to pursue successful agreements, sales, logistics, manufacturing, procurement, and all other departments must be coordinated. Companies that succeed in managing SC are better able to take advantage of value-creation opportunities that their competitors might miss. [17] For example:

- Companies can lower inventories by embracing lean manufacturing.
- By being attentive to customer needs, they can enhance their relationships with clients and increase revenue.
- By working closely with their suppliers, they may obtain the products they require at a fair price when needed.

In most companies today, more than 70% of the costs and 100% of the revenues depend on how the SC is managed. So keeping all of the parts of the SC aligned is key to running any business successfully, so the importance of SCM is quickly rising.

2.2.2 Exploring Complex Business Challenges

Numerous parts are involved in the management of a business, making long-term plans turn out to be impossible; this challenge involves all price swings, natural disasters or even a financial meltdown. In another way, the goal is to focus on a range of scenarios and hence a possible plan to execute when the times comes. Those scenarios are exclusively chosen because they impact the SC; SCM resumes to an ”if this, then that” (IFTT), it becomes a process of sensing and responding to scenarios. The objective is to disperse far away from Force majeure. [11]

A research study of over 100 manufacturers, distributors, and retailers conducted more than ten years ago revealed some commonly employed SC initiatives and tactics. These concepts and methods were condensed into seven guiding principles and presented in a SCM review article [18]:

1. “Segment customers based on the service needs of distinct groups and adapt the SC to serve these segments profitably”.;

2. “Customize the logistics network to the service requirements and profitability of customer segments”.;

3. “Listen to market signals and align demand planning accordingly across the SC, ensuring consistent forecasts and optimal resource allocation”.;

4. “Differentiate product closer to the customer and speed conversation across the SC”.;

5. “Manage sources of supply strategically to reduce the total cost of owning materials and services”.;

6. “Develop a SC-wide technology strategy that supports multiple levels of decision making and gives clear view of the flow of products, ítem services, and information”.;

7. “Adopt channel-spanning performance measures to gauge collective success in reaching the end-user effectively and efficiently”.;

These timeless principles, even though they are more than ten years old, emphasize the necessity for SC leaders to put the customer first. Additionally, they emphasize how crucial it is to coordinate tasks (such as demand planning, sourcing, assembly, delivery, and information exchange) both inside and between businesses.

2.2.3 Designing Supply Chain Systems

Like many other systems, SC comprises various interrelated, unpredictable components, including people, processes, and technological advancements. For the system to function as intended, each of these elements must be appropriately structured and handled.

SC has a set of underlying rules and patterns that are key to understanding how they work, a good example of this is "The Bullwhip Effect" as is illustrated in Figure 1.[19]

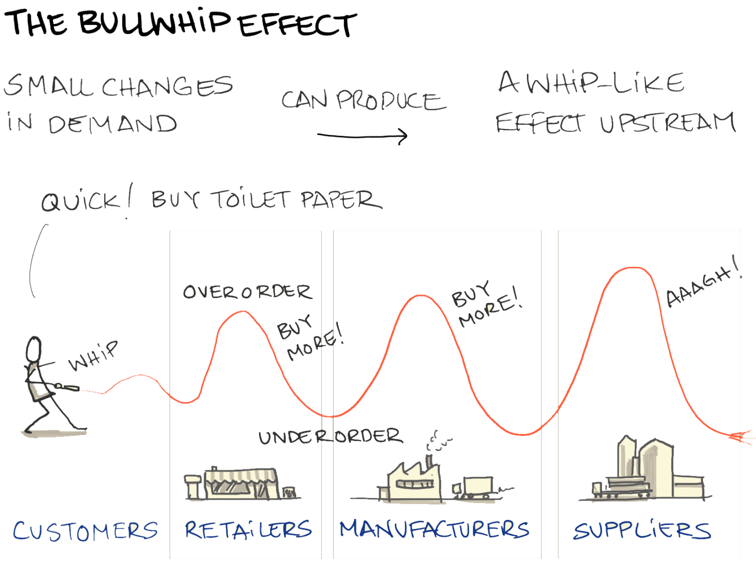


Figure 1: The Bullwhip Effect on the SC. [19]

This effect is a phenomenon that regularly arises in SC systems and is a typical, predictable outcome of everyone in the SC making decisions that seem rational. It can cause inventory peaks and valleys that are more exaggerated as they go upstream from one step to the next in a SC.

It can be challenging to create a model of a SC that illustrates how the many components work together. These models are built on cause-and-effect linkages, which indicate how one department influences another and causes something to occur. Systems models frequently show reinforcing loops, in which a series of events repeatedly occurs and gets stronger each time. Alternately, they can demonstrate balancing loops that have the reverse result, where a succession of events gradually becomes weaker.[11] The flow of information and products across a hypothetical SC can be observed in figures 2 and 3. The objective of this demonstration is to compare the traditional and demand-driven flow.

In a traditional system, each entity operates independently, with its priorities, responsibilities and goals. Typically, processes, data and information are dispersed throughout the entire organization. This essential and straightforward system can be seen in Figure 2.

However, it is necessary to innovate this outdated system that makes management, innovation and organizational evolution impossible. The prime objective in designing an SC is to have the appropriate level of responsiveness to the present; this comes from the requirement for the automated production of processes linked to the use of intelligent systems. It also incorporates a sophisticated and efficient control of the chain, making it absolutely necessary to analyze information. Its illustration can be seen in Figure 3.

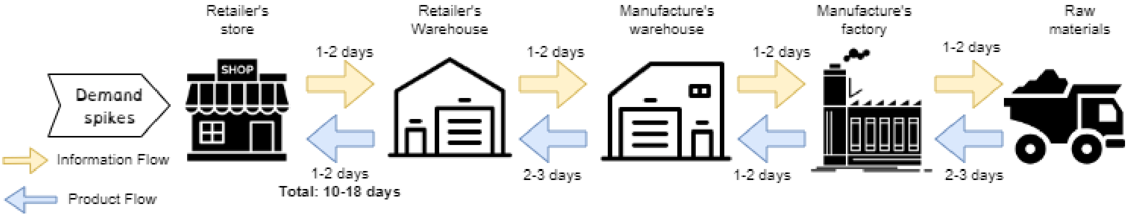


Figure 2: Supply chain management (Traditional system)

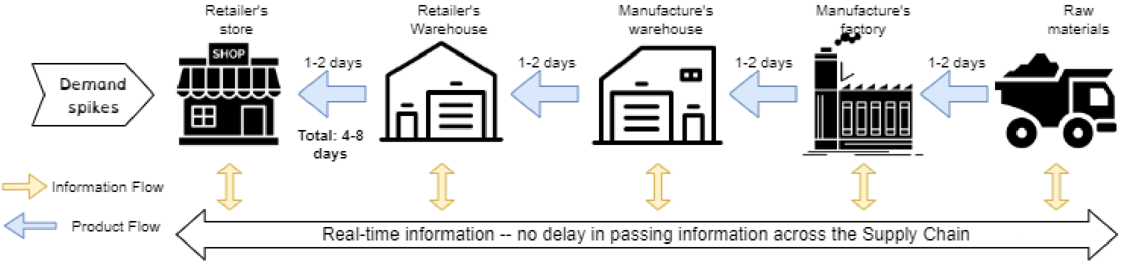


Figure 3: Demand-Driven system

2.3 Multiple-criteria decision analysis

This sub-section explains what the MCDA models are, which ones were chosen for the more detailed exploration according to the intended goal of this project and how their selection was made.

2.3.1 Theoretical overview

The primary objective of the Multi-Criteria Decision Analysis (MCDA) approach is to assist a Decision Maker (DM) in determining an ideal supplier from a large pool of alternatives while considering a wide range of factors that determine whether or not a particular decision variant is suitable. One of the most significant aspects of these mathematical models is that when all possibilities are acceptable and choosing the best one is a subjective decision, the criteria can also rank the quality of the alternatives. Subjectivity in this context refers to the relative weight given to various measures since, for each DM, some criteria are often more important than others.

Additionally, the subjectivity of judgment is impacted by the ambiguity and inaccuracy of data-defining options. [20]

In a SCM, most companies have to allocate and expend a considerable amount of the budget on raw materials (up to 70%) since this is an essential part of a product's production and final cost.[21]

Choosing the right supplier directly affects the final cost of the product and the company's profitability since this process is dependent on an evaluation and selection from a manager, which is a challenging process for the purchasing departments.

Supplier Selection Criteria (SSC) primarily require domain expert assessment and judgment; as discussed before, it's necessary to make trade-offs between countless factors/weights, some of which may conflict with each other;

On a SS, choosing it correctly or incorrectly can drastically impact the output of the action: choosing it effectively reduces purchasing risk, maximises consumer satisfaction and builds good relationships with the stakeholders; on the other hand, choosing it poorly may cause an economic problem and impact on the whole company.

Multi-Criteria Decision-Making (MCDM) or MCDA is one of the techniques that researchers have applied to assist purchasing managers:

- Making the right decisions that will help their company to reach maximum achievement;
- Evaluating and analysing the decision-making process and the techniques used to achieve the company's objectives;

More often, the complexity involving MCDM techniques is a significant factor, which becomes difficult to manage and evaluate; on the other hand, the recent propagation of Big Data and AI induce many researchers to consider it in their applications.

2.3.2 Choosing an appropriate MCDA technique

A group of university faculty members from the University of Szczecin (Szczecin, Poland) and West Pomeranian University of Technology (Szczecin, Poland) published a scientific article comparing all currently existing mathematical models and classifying them using divisions based on those models;[22] Additionally, an online tool was developed to make the entire procedure simple and easy to understand. [23]

There are 50 simple and 6 compound methods included in this article. Through the use of nine filtering parameters, it is possible to obtain one or more processes suitable for the issue intended to be solved.

The first parameter requests if the issue has different weights for the various criteria that will be considered; the possible answers are yes or no.

If the decision-making criteria have weights, it is necessary to specify whether they are qualitative, quantitative, or relative. Every qualitative value is considered ordinal, and quantitative is the cardinal value; when the scale is relative can be of interval or ratio type.

The third parameter is unfolded into four, consisting of the question of whether the criteria suffer from uncertainty and its kind, the type of uncertainty about the data and whether there is personal preference about the data;

If the criteria have associated uncertainty, all other options are unlocked; as would be expected if the opposite is indicated, it is not possible to describe in more detail precisely at which level of uncertainty.

The level of uncertainty may be determined by the uncertainty of the DM's preferences, the input data uncertainty, or both.

If the uncertainty concerns the DM's preferences, the thresholds that will be used in the decision problem are indifference, preference or both.

After analysing and studying all of these filtering methods, it is very vital to ensure that the options are chosen adequately because only then can we obtain the best mathematical model.

The following selections of filtering were used to determine which models may use:

- **Weights**: yes;
- **Weights Type**: Quantity;
- **Weights Scale**: Quantity;
- **Has uncertainty**: yes;
- **Uncertainty Type**: Input;
- **Data uncertainty type**: undefined;
- **Preference uncertainty type**: No Preference uncertainty;
- **Topic**: Ranking and choice;
- **Ranking type**: Complete.

Breaking it down, the model will take into account the weights, quantity as a weight type and as a weight scale; Uncertainty exists; The kind of uncertainty will be inputted by DM's; No preference uncertainty is the type of uncertainty that exists. The topics are selection and ranking, which will result in a complete ranking.

Presented in more detail in the following subsections, the following results were obtained by the selection tool:

- Fuzzy SAW;
- Fuzzy TOPSIS;
- Fuzzy VIKOR;

2.3.3 Fuzzy Set Theory

The fuzzy technique aids the DM in minimizing subjective value judgments and improving applicant selection, using various evaluation criteria. Due to the complexity and variety of suppliers applicants , evaluating them precisely is a challenging procedure; [24]

The theory's fundamental premise is that fuzziness, rather than randomness, is the primary cause of imprecision in many real-world problems. When handling uncertainty in certain circumstances, fuzzy set theory is preferable to probability theory.

With the use of a fuzzy dictionary, DM's assign weights to each skill with its associated ranking fuzzy attributes in a fuzzy suitability Table 2, when the level of appropriateness is known, each linguistic value in this table has a value that is specifically assigned to it.

Term	Fuzzy Number
Very Low	1,1,3
Low	1,3,5
Average	3,5,7
High	5,7,9
Very High	7,9,9

Table 2: Fuzzy preference scale

In Figure 4, it's possible to observe how the fuzzy number in Table 2 is generated, After identifying which ranking is to be assigned, the fuzzy number is calculated through the following procedures: the value is made up of three values (*x1*, *x2*, and *x3*); the values of *x1* and *x3* represent the extremes of the desired ranking(in the case of extremities, which are not listed with a value in the immediate vicinity ('Very low'" and "Very high"" ranks), the value of the rank itself is repeated); the value of *x2* is the value of the ranking itself.

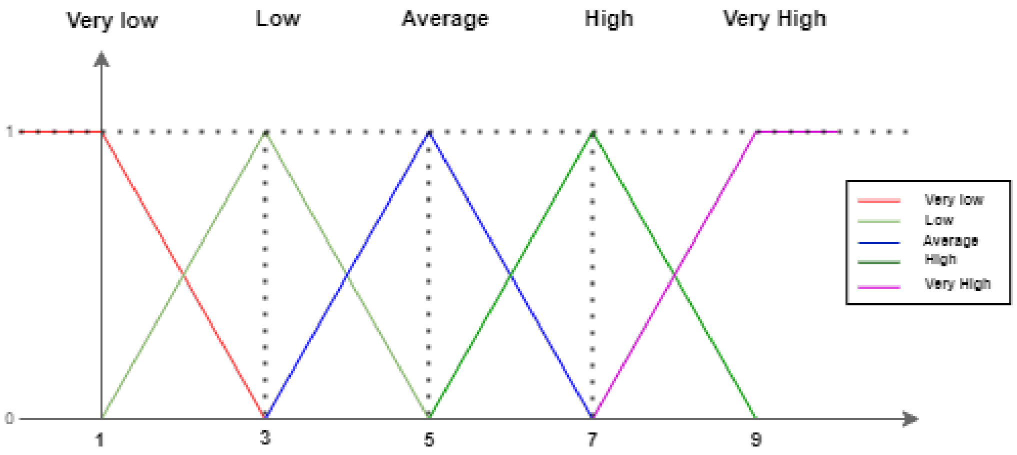


Figure 4: Fuzzy Number Graph

2.3.4 Fuzzy SAW

Composed by the original Simple Additive Weighting (SAW) method component and added to the fuzzy general model. The variants and weights must be evaluated quantitatively. The ratings of the variants in relation to each individual criterion should be proportionally normalized to the rating given to each criterion at its greatest level. In order to aggregate preferences, a criterion's weights must be determined, and a variant's relevance to the criterion must be assessed. Then the total of all such items for a given variant is calculated. [25]

2.3.5 Fuzzy VIKOR

The VIseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) approach entails seeking the closest compromise to the optimal course of action. The worst and best values of each criterion are first established. Utility measures and regret measures for variants with regard to each criterion are calculated based on the values. Next, the minimum and maximum values for each variant, whose combination allows calculating a variant's ranking position. [26]

2.3.6 Fuzzy TOPSIS

One of the most well-known methods for resolving MCDM issues is the Fuzzy Technique for Order Performance by Similarity to Ideal Solutions (TOPSIS), which was put forth by Hwang and Yoon. [27]

This approach is predicated on the idea that the selected alternative should be closest to the Fuzzy Positive Ideal Solution (FPIS), or the solution that minimizes costs and maximizes benefits, and the furthest from the Fuzzy Negative Ideal Solution (FNIS).[28]

Fuzzy TOPSIS explanation can be divided into 8 steps:

Step 1. Assignment rating to the criteria and to the alternatives. Presuming that the DM group has k members. The weight of criterion C_j is indicated, along with the fuzzy rating of the k^{th} decision maker;

The criteria itself is denoted as: $\tilde{x}_{ij}^k = (a_{ij}^k, a_{ij}^k, a_{ij}^k)$, and it weight is denoted as: $\tilde{w}_{ij}^k = (w_{j1}^k, w_{j2}^k, w_{j3}^k)$.

Step 2. Compute the aggregated fuzzy ratings for alternatives and the aggregated fuzzy weights for criteria, can be seen at Table 2

The aggregated fuzzy rating $\tilde{x}_{ij} = (a_{ij}, b_{ij}, c_{ij})$ of i^{th} is obtained as follows:

$$a_{ij} = \min_k \left\{a_{ij}^k\right\}, b_{ij} = \frac{1}{k}\sum_{k=1}^{k} b_{ij}^k, c_{ij} = \max_k \left\{c_{ij}^k\right\} \tag{2.1}$$

The aggregated fuzzy weight $\tilde{w}_j = (a_{j1}, b_{j2}, c_{j3})$ of C^J are calculated by formula:

$$w_{j1} = \min_k \left\{w_{j1}^k\right\}, w_{j2} = \frac{1}{k}\sum_{k=1}^{k} w_{j2}^k, w_{j3} = \max_k \left\{c_{j3}^k\right\} \tag{2.2}$$

Step 3. Compute the normalized fuzzy decision matrix.

The normalized fuzzy decision matrix is $\tilde{R} = [\tilde{r}_{ij}]$, where:

$$\tilde{r}_{ij} = \left(\frac{a_{ij}}{c_j^*}, \frac{b_{ij}}{c_j^*}, \frac{c_{ij}}{c_j^*} \right) and\ c_j^* = \max_i \{c_{ij}\}\ (benefit\ criteria) \tag{2.3}$$

$$\tilde{r}_{ij} = \left(\frac{a_j^-}{c_{ij}}, \frac{a_j^-}{b_{ij}}, \frac{a_j^-}{a_{ij}} \right)\ and\ c_j^- = \min_i \{a_{ij}\}\ (cost\ criteria) \tag{2.4}$$

Step 4. Compute the weighted normalized fuzzy decision matrix. The weighted normalized fuzzy decision matrix is $\tilde{V} = (\tilde{v}_{ij}),\ where\ \tilde{v_{ij}} = \tilde{r_{ij}} \times w_j$.

Step 5. Compute the FPIS and FNIS and are calculated as follows:

$$A^* = \tilde{v}\overset{*}{1}, \tilde{v}\overset{*}{2}, ..., \tilde{v}\overset{*}{n},\ \text{where}\ \tilde{v}\overset{*}{j}, = \frac{max}{i}{}_v ij3 \tag{2.5}$$

$$A^- = \tilde{v}\overset{-}{1}, \tilde{v}\overset{-}{2}, ..., \tilde{v}\overset{-}{n},\ \text{where}\ \tilde{v}\overset{-}{j}, = \frac{min}{i}{}_v ij1 \tag{2.6}$$

Step 6. Compute the distance from each alternative to the FPIS and to the FNIS.

$$d_i^* = \sum_{j=1}^{n} d(\tilde{v}_{ij}, \tilde{v}_{ij}^*),\ d_i^- = \sum_{j=1}^{n} d(\tilde{v}_{ij}, \tilde{v}_{ij}^-) \tag{2.7}$$

Step 7. Compute the closeness coefficient Closeness Coefficient index (CCi) for each alternative.

$$CC_i = \frac{d^-}{d_i^- + d_i^*} \tag{2.8}$$

Step 8. Rank the alternatives The best alternative is the one with the highest proximity coefficient.

2.3.7 Supplier Criteria

The selection of suppliers is dependent on a method of evaluation made by the DM's, multiple methods have been proposed over the year to the outcome of an effective SS. According to literature, some SS criteria are found to vary in different situations, and experts agree that there isn't a best way to evaluate and select suppliers, and those organizations use a variety of different approaches in their evaluating processes:

The first document mentioning this type of criteria was in 1966 by Dickson. He studied 23 criteria for assessing a supplier's performance, the study contained 300 organizations, and, surprisingly it was observed that the **price** was not the most important factor in SS. [29]

The ability of each supplier to meet the required quality was an extremely important criterion, and reciprocal arrangements had slight importance on the supplier evaluation.

Thereafter, several researchers continued studying the effects of various criteria in the SS process. In the early days, the price was the sole factor determining the suitable supplier; however, the selection attributes have been expanded, and some new ones have been introduced, responding to the growth of new business needs.

Between 1966-1990 many articles were published that continued Dickson's study, and Weber re-examined Dickson's work. 47 of the 74 reviewed articles used multiple criteria as listed by Dickson for the selection process, and some additional criteria were added.[30]

Table 3 shows the selection criteria conceded in multiple articles by multiple researchers from articles published between 1966 and 2000, this literature review was conducted by *Sung Ho Ha*, and *Ramayya Krishnan* [21], and it was obtained from multiple sources, for example, in article performed by Dickson, was conducted a survey that identified factors that buyers considered in awarding contracts to suppliers [29].

Selection criteria	A	B	C	D	E	F	G	H	I	J	Total
Price	✓	✓	✓	✓	✓	✓	✓		✓	✓	9
Quality	✓	✓		✓		✓	✓		✓		6
Delivery	✓	✓	✓	✓		✓	✓		✓		7
Reputation and position in industry	✓	✓	✓		✓		✓				5
Reciprocal arrangements	✓	✓		✓			✓				4
Amount of past business	✓	✓	✓				✓				
Technical capability	✓	✓					✓	✓			
Impression	✓		✓		✓		✓				
Geographical location	✓	✓		✓			✓				
After-sales service	✓		✓		✓		✓				
Training aids	✓		✓				✓				3
Attitude	✓				✓		✓				
Financial position	✓		✓				✓				
Management and organization	✓			✓			✓				
Technical support			✓		✓	✓					
E-commerce capability								✓	✓	✓	
Communication system	✓						✓				2
Response to customer request			✓			✓					
Packaging ability	✓						✓				
Operational controls	✓						✓				
Production facilities and capacity	✓						✓				
Ease-of-use			✓		✓						
Warranties and claims	✓		✓								
Labor relations	✓						✓				
JIT capability						✓		✓			
Maintainability			✓		✓						
Performance history	✓						✓				
Environmentally friendly products									✓		1
Product appearance										✓	
Catalog technology										✓	

Table 3: Selection criteria from literature review:**A**:Dickson (1966); **B**:Wind et al. (1968); **C**:Lehmann and O'Shaughnessy(1974); **D**:Perreault and Russ (1976); **E**:Abratt (1986); **F**:Billesbach et al.(1991); **G**: Weber et al. (1991); **H**, Segev et al. (1998); **I**, Min and Galle (1999); **J**, Stavropolous (2000). [21]

2.4 Process model for data exploration

Data Science is a subject that uses mathematical and analytical models and applications to acquire meaningful insights from data. Processing enormous amounts of data to aid decision-making is one of the top priorities for IT corporations. [31]

Project management and process techniques can help data science initiatives, but if it's strictly adhering to a project methodology may be difficult for data science teams. Process models such as Cross-Industry Standard Process for Data Mining (CRISP-DM) may and should be augmented by agile methodologies.

Published in 1999, CRISP-DM is a data mining process paradigm that is sector agnostic. This methodology became the most widely used method for data mining and data science projects. It is divided into six iterative stages, beginning with business knowledge and ending with implementation [32]:

- **Business Understanding:** Understands the project objectives and needs from a business standpoint. The analyst formalises this information as a data mining issue and creates a preliminary strategy.

- **Data understanding:** Beginning with basic data gathering, the analyst moves on to actions to become acquainted with the data, detect data quality issues, and gain early insights into the data. During this stage, the analyst may also identify relevant subgroups in order to build hypotheses regarding hidden information.

- **Data preparation:** The data preparation step includes all processes required to create the final dataset from the raw data. The analyst assesses, chooses, and employs relevant modelling approaches.

- **Modeling:** The data modelling process consists of selecting a modelling technique, developing a test case, and developing the model.

- **Evaluation:** Evaluation of the results is directly compared with the objectives established in the first phase. The final outcome is the choosing of the champion model(s).

- **Deployment:** In most cases, this will include integrating a code version of the model into an software/operating system. This also includes techniques for scoring or categorising previously unknown data as it emerges. The mechanism should apply the new knowledge to the original business challenge.

It's possible to observe the phases described above in the diagram represented in Figure 5.

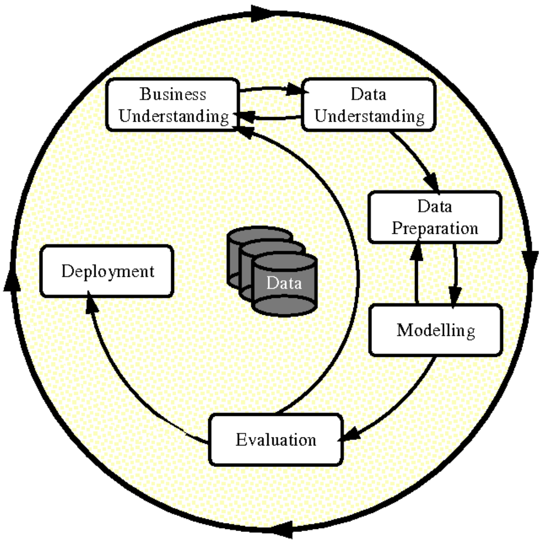


Figure 5: Phases of CRISP-DM Process Model for Data Mining. [32]

2.5 Artificial Intelligence

With a perspective of simplifying the structure of AI, Figure 6 illustrates how these key components connect to one another from a high-level perspective. AI is at the top, and it's possible to divide it into ML and Deep Learning (DL), which are the two primary categories.[33]

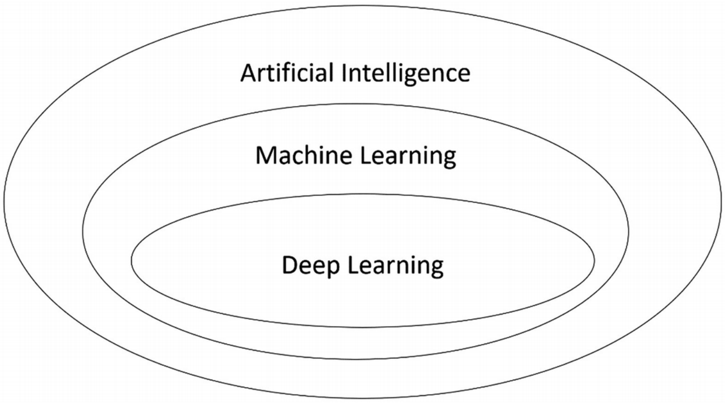


Figure 6: High-level components of AI. [33]

2.6 Machine Learning

This section presents ML concepts and notations, as well as an overview of the available literature on learning techniques.

ML is defined as "The study of computer algorithms capable of learning to improve their performance of a task based on prior experience." [34]

ML techniques are based on pattern recognition, computer science, and statistical inference. These methods work best when pointed directly to data-intensive areas, such as finance, economics, medicine, and others; The main objective is usually to obtain valuable insights and forecasts with an evidence-based perspective. In other words, ML is a subfield of AI that uses algorithms to synthesize the underlying relationships between data and information.

2.6.1 Learning models types

According to Sarker, in ML, there are four types of models, which are illustrated in Figure 7. [35]

- **Supervised learning**: refers to a problem in which a model is used to learn a mapping between input samples and the target variable.
  - **Classification**: it's a process of classification data after learning from a dataset using partern recognition.
  - **Regression**: used when the result consists of a real or continuous value.
- **Unsupervised Learning**: when a model is used to characterise or extract relationships in data.
  - **Clustering**: is an approach for finding and grouping similar data points; it does it by classifying each item, and those belonging to the same category, known as a cluster, are grouped.
  - **Association**: detects latent connections in databases by applying some metric of interest to provide an association rule for new searches. Association rule learning measures degrees of similarity to uncover significant relationships between variables or features in a dataset.
- **Semi-Supervised Learning**: learns from a small number of labelled instances and tries to label many unlabeled data.
  - **Classification**
  - **Clustering**
- **Reinforcement Learning**: a set of challenges in which an agent must learn to operate in a given environment using feedback. The goal is to make effective use of all of the available data.

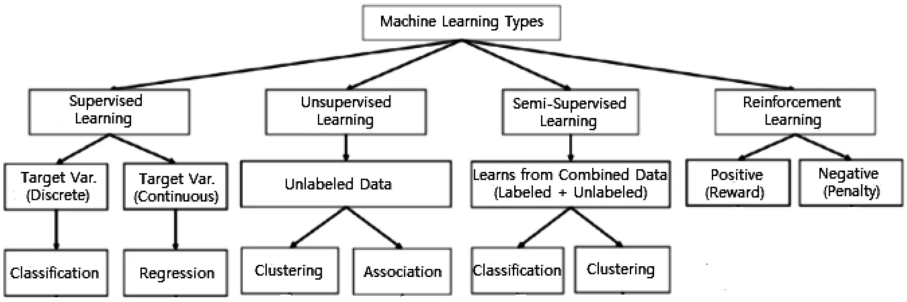


Figure 7: Various types of machine learning techniques. [35]

2.6.2 Key terminology

This section identifies and explains some essential transdisciplinary conceptual terminology concerning ML to enhance comprehension.[36]

Dataset: A collection of data that conform to a schema with no ordering requirements. In a typical dataset, each column represents a feature, and each row represents a member of the dataset.

Canonical data: The Enterprise Integration Patterns propose a canonical data model to minimise dependencies when connecting apps that use diverse data formats. In other words, a component (an application or a service) should communicate with another component using a data format that is independent of the data formats of both components.[37]

Model: A model is a file with a set of methods and mechanics that describe or forecast the dataset. Each model can be customised to meet the specific needs of a given application. Large datasets with multiple predictors and attributes in big data applications are too complicated for a basic parametric model to extract relevant information.

Predict probability: The prediction probability is the estimate of the prediction for all classes, which are ordered by the label of classes.[38]

Accuracy: Accuracy is one metric for evaluating classification models; it's calculated by the fraction of predictions of the model. As follows: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. [1]

Precision: The precision is calculated with the following mathematical expression:

$precision\_score = \frac{TP}{TP+FP}$. [1]

The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. [2]

Recall: The recall is calculated with the following mathematical expression:

$recall\_score = \frac{TP}{TP+FN}$. [1]

The recall is the classifier's ability to find all the positive samples. [2]

F1: The F1 score it can be understood as a harmonic mean of precision and recall, and is calculated with the following mathematical expression: $F1 = \frac{2 \cdot precision \cdot recall}{precision+recall}$. [2]

In the case of many classes and labels, this is the average of the F1 scores for each class, with weighting based on the average parameter.

Root-mean-square error The standard deviation of the residuals is defined as the root-mean-square error (rmse)(prediction errors). Residuals measure how far away data points are from the regression line; rmse is a measure of how to spread out these residuals. In other words, it indicates how robust the data is around the best fit line. [39]

Logistic loss or cross-entropy loss: Log loss is a loss function used in (multinomial) logistic regression and extensions of it, defined as the negative log-likelihood of a logistic model that returns y_pred probabilities for its training data y_true. The log loss is only defined for two or more labels. For

[1] Where TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives.

[2] The best value is 1 and the worst value is 0.

a single sample with true label $y \in 0, 1$ and a probability estimate $p = Pr(y = 1)$, the log loss is: $L_{log(y,p)} = -(ylog(p) + (1 - y)log(1 - p))$

Receiver operating characteristic curve: The Area under curve (AUC) - Receiver operating characteristic curve (ROC) curve is a performance evaluation for classification issues at various threshold levels. ROC is usually a curve, and AUC reflects the degree or measure of separability. It expresses how well the model can discriminate between classes. The higher the AUC, the better the model is in predicting.[40]

Figure 8 is shown as an illustration with a description of the ROC-AUC curve, in this metric, the TPR and FPR are the rates of TP and FP.[1]

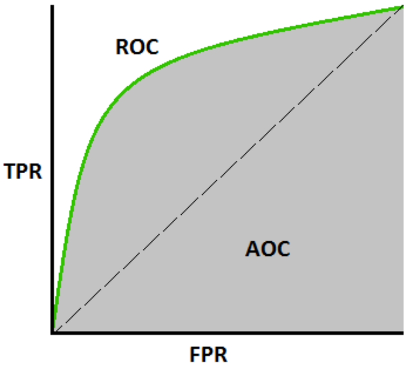


Figure 8: ROC-AUC Cuver example

Confusion matrix: The confusion matrix is a comprehensive measure for solving classification problems. It can be used for both binary classification and multiclass classification problems. Confusion matrices represent the sum of predicted and actual counts. The output "TN" stands for True Negative and represents the number of correctly classified negative examples. Similarly, "TP" stands for True Positive, and it denotes the number of correctly classified positive examples. The abbreviation "FP" refers to the number of actual negative examples classified as positive, while "FN" refers to the number of actual positive examples classified as negative. In Table 4 is represented this matrix. [41]

	Predicted Class	
True Class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Table 4: Confusion matrix for binary classification.

2.6.3 Machine learning model construction procedure

ML contributes to the development of systems that, via experience and training, increase their performance on a specific task. Modern computing equipment is capable of handling massive data quantities and increased complexity.

The process of developing ML algorithms may be decomposed into the following steps [42]:

1. *Collect the data.*
2. *Preprocess the data.*
   - i *Formating.*
   - ii *Cleaning.*
   - iii *Sampling.*
3. *Transform the data.*
4. *Train the algorithm.*
5. *Test the algorithm.*
6. *Apply human reinforcement learning.*

The development of ML begins with identifying all the factors that are crucial to a decision process, because it achieves the highest performance possible when these variables are tuned, it is critical to carefully choose the variables in conceptual phases.

The extraction of characteristics that can be used for model development is a key stage in the automated identification of patterns and correlations from vast datasets. A feature, in general, describes a property obtained from raw data input to give an appropriate representation. Thus, feature extraction aims to preserve discriminatory information and separate factors of variation relevant to the overall learning task. [43]

Preprocessing is divided into three steps: formatting, cleaning, and sampling. Data composition is essential to use it in a functioning state, facilitating data injection. In addition to the correct format, treatment is crucial through eliminating or replacing corrupt or missing values. Data must be sampled at regular or adaptive intervals in such a way that redundancy is minimized while information is not lost for transmission across communication channels.

Data transformation is specific to the algorithm and problem understanding, this can be achieved by scaling, deconstruction, or accumulation of features. Features can be created or transformed to identify relevant components buried in the data or aggregated to merge numerous occurrences into a single feature.

Testing the algorithm is a process that modifies how the process of training the model works based on the test before this step. This learned knowledge or information is saved as a model for future cross-validation and use.

Human interpretation of the created model is critical to its improvement. Combining tests and critical analysis of model results makes the process of improving performance essential.

2.6.4 Machine learning algorithms

Throughout this section, all ML algorithms used in the development process are presented in the following subsections: 2.6.4.1, 2.6.4.2, 2.6.4.3, 2.6.4.4. Algorithms that depart from others that have already been explained, only the particularities are introduced.

2.6.4.1 Decision Tree

Decision trees are one of the oldest and most widely used strategies for learning discriminatory models. A decision tree is a tree-structured classification model that learns from data in a top-down fashion, with an algorithm known as Top-down induction of decision trees (TDIDT), recursive partitioning, or divide-and-conquer learning.[44]

The algorithm selects the best attribute for the root of the tree, splits the set of examples into disjoint sets, and adds corresponding nodes and branches to the tree.

This algorithm analyzes and creates groupings (branches) to distinguish situations that cause a change in the decision-making process. The result is an algorithm that can be converted into a decision tree that contains all of the associated limits and decisions; a exemple of this process can be observed in Figure 9 with the corresponding data from the Table 5.

Outlook	Humidity	Windy	Golf?
rainy	high	false	no
rainy	high	true	no
rainy	normal	false	yes
rainy	normal	true	yes
sunny	high	false	yes
sunny	high	true	no
sunny	normal	false	yes
sunny	normal	true	no
overcast	normal	true	yes
overcast	high	false	yes

Table 5: A dataset describing weather conditions and a target variable (Play Golf?)[44]

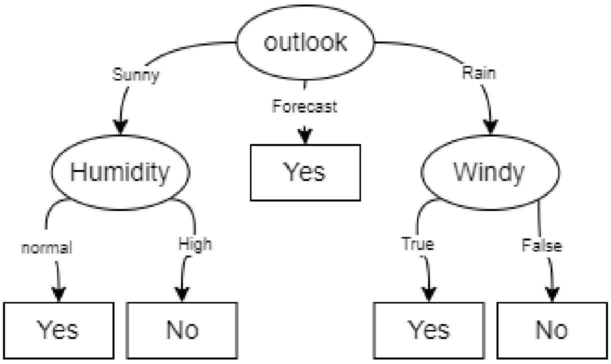


Figure 9: A decision tree learned for this dataset.[44]

2.6.4.2 Random Forest

Random Forests is a technique for ensemble learning and can be used for both classification and regression tasks. It is a mix of the Bagging technique and the random subspace approach, and its basis classifier is decision trees. [45]

The Forest is created by a combination (ensemble) of decision trees. The main idea behind the bagging technique method is to combine learning models to improve overall results. Each tree is built from a bootstrap sample taken from the original dataset; after its construction, the trees aren't pruned, allowing them to be overfitted to their sample of data. To ensure the diversity of classifiers, the selection of which feature to split on is constrained to a random subset.

2.6.4.3 eXtreme Gradient Boosting

Similar to Random Forest, eXtreme Gradient Boosting (XGBoost) is also based on decision trees, it is a gradient boosting ensemble ML. Decision tree-based algorithms are regarded best-in-class for small-to-medium structured/tabular data.[46, 47]

One of the positive aspects of this learning model is that the process is built in a parallelized manner and consists of cache awareness and out-of-core computation, which will make the best use of hardware resources.

Gradient boosting decision tree (GBDT) is an ensemble model of decision trees which are trained in sequence, by fitting the residual errors, it learns the decision tree in each iteration. This means, that every subsequent learner tries to learn the difference between actual output and the weighted sum of predictions until the previous iteration. The errors are minimised using the gradient method.

In addition, a new parameterization element called *"max_depth"* is introduced, causing the algorithm to stop prunning trees on the first criterion and instead begin pruning trees backward. This "depth-first" method dramatically enhances computing performance.

Coming to algorithm enhancements, their are three new introductions:

- **Regularization:** To prevent overfitting, it penalizes more complicated models using both Least Absolute Selection Shrinkage Operator (LASSO) (L1) and Ridge regressions (L2) regularization.

- **Sparsity-aware:** It is important to make the algorithm aware of the sparsity pattern in the data. XGBoost handles all sparsity patterns consistently, it is exploited to make computation complexity proportional to the number of non-missing elements in the input.

- **Weighted Quantile Sketch:** Proposing possible split points is a key stage in the approximation method. Typically, feature percentiles are employed to ensure that candidates are distributed uniformly across the data.

2.6.4.4 Light Gradient Boosting Machine

In 2016, Microsoft created Light Gradient Boosting Machine (LightGBM), an open-source distributed gradient boosting algorithm for ML applications. It is based on a decision tree algorithm and emphasizes performance and scalability.[48] Although the similarities to XGBoost, LightGBM employs a revolutionary approach of Gradient-based One-Side Sampling (GOSS) to filter out data instances to discover a split value, whereas XGBoost uses a pre-sorted algorithm and a Histogram-based algorithm to compute the optimal split. Discarding data instances with modest gradients is a simple concept. However, by doing so, the data distribution will be altered, which will reduce the trained model's accuracy.

Researchers suggest a novel technique termed GOSS to get around this issue: GOSS down samples instances based on gradients, examples with modest slopes have received adequate training (low training error), while those with huge gradients have not. A simplistic method of downsampling would be to ignore cases with tiny angles instead of just concentrating on examples with significant rises, but this would change the data distribution. In essence, GOSS performs random sampling on instances with tiny gradients while keeping hold of examples with big gradients.

Exclusive Feature Bundling (EFB) is also utilized in this new algorithm; although there are often many features in practical applications, the feature space is typically very small, allowing us to build a virtually lossless method to minimize the number of useful features.

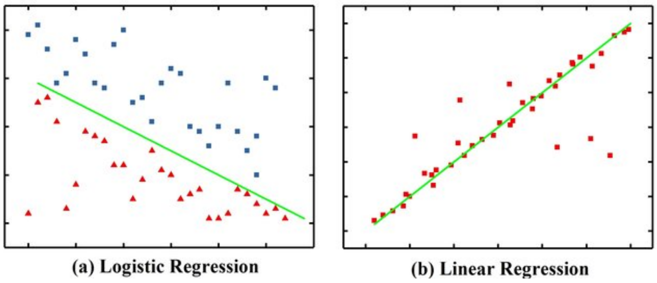
It's possible to combine such unique characteristics with confidence. To achieve this, we create an effective algorithm by converting the optimal bundling problem into a graph colouring problem (by treating features as vertices and adding edges for each pair of features if they are not mutually exclusive) and solving the resulting problem using a greedy algorithm with a fixed approximation ratio.

2.6.4.5 Logistic Regression

The logit is the natural logarithm of an odds ratio, and it's the central mathematical concept underlying logistic regression. The dependent variable in a logistic regression has only two categories, generally, the event's occurrence is coded as one and its absence as 0, but it can be changed because the codification alters the signal of the coefficients and thus their substantive interpretation. [49]

Linear regression is a method for finding the line that best fits the data. [50] If the data is from the actual world, there is probably no equation of that type that fits every data point precisely. However, if the problem is devised by a metric that indicates how distant each data point is from a straight line, it's possible to minimize the error and arrive at the best feasible equation. Linear regression typically does this by calculating the distance between each point and a line and then setting the values to minimize that distance metric.

However, the logarithmic regression is similar to the linear regression, although the values of the independent variable must be introduced, and the measure used to determine how well the independent variable adjusts to the predicted value is far more sophisticated. The data do not need to be adjusted to a straight-line model. The comparison between these two models and be found in Figure 10.



(a) Logistic Regression

(b) Linear Regression

Figure 10: Linear and Logistic Regression. [50]

3

Software Tools review

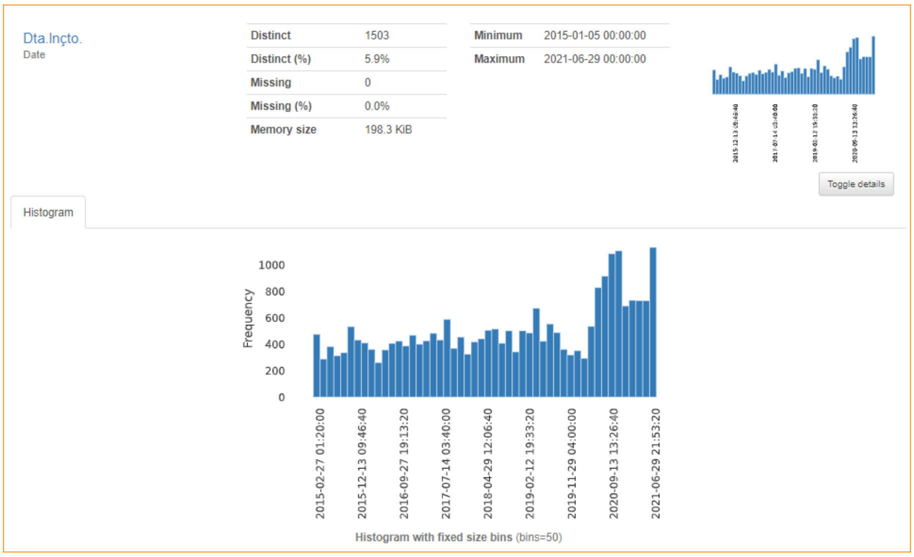
The tools used in the development process are discussed in this chapter.

3.1 Python

Since Python first appeared in 1991, it has become one of the most popular interpreted programming languages, along with Perl and Ruby. In 2005, after the release of the Django framework for web development, the language adoption skyrocketed. In the last ten years, Python has become one of the most important languages for data science, ML and also for developing computer applications in general. [51]

A tool known as Pandas was used for the data analysis and processing procedure. This framework started as a standalone project in 2008 and became open-source the following year. Currently, the maintenance and development of new versions are produced by various organizations and with contributions from the community. This technology came to help simplify data manipulation methods in the financial market area, where other technologies were not up to dealing with scenarios with a correlation between two similar data sets and the constant change in their values. Today the Pandas framework is an indispensable component in areas such as Data manipulation, Data analysis and Data science.[52]

Data exploration and analysis are extensive processes due to the power of impact on the outcome of the models; As a result, additional tools that complement this framework have been developed, one of which is the Pandas Profiling. This open-source tool generates standard statistical information automatically, which was used to quickly assess information and understand the distribution of the data, by default it contains an overview of all matters, statistical analytics and Interactions between each “variable”; a short example of the overview output from pandas profile is demonstrated in Figure 11. [53]



Dta.Inçto.
Date

Distinct	1503
Distinct (%)	5.9%
Missing	0
Missing (%)	0.0%
Memory size	198.3 KiB

Minimum	2015-01-05 00:00:00
Maximum	2021-06-29 00:00:00

Figure 11: Pandas Profiling short example of an overview on one column

3.2 Synthetic data generation

Synthetic data refers to information created artificially rather than data encountered in real-world occurrences. [54]

This data is generated algorithmically and is used as a replacement for production or operational data test to validate mathematical models and, increasingly, to train ML algorithms. The advantages of employing synthetic data include minimizing limits when using sensitive or regulated data, adapting data needs to specific conditions that cannot be obtained with actual data.

Inconsistencies while attempting to reproduce the complexity inherent in the original dataset and the inability to completely substitute authentic data, as correct actual data is still necessary to make meaningful synthetic instances of the information, are drawbacks.

During the development process, two frameworks were used to generate synthetic data: the *FAKER* [55] library and the *dbldatagen* [56] within the databricks environment.

3.3 Pyspark

In 2009, Apache Hadoop dominated the market related to Big Data because it was the first open-source system that made it possible to parallelize data in a cluster with one or more nodes. PySpark provides an interface for Apache Spark in Python, which is one of the Apache Hadoop ecosystem. It not only enables you to create Spark applications using Python APIs, but it also gives you access to the PySpark shell, which enables interactive data analysis in a distributed setting. The majority of Spark's features, including Spark SQL, DataFrame, Streaming, MLlib (Machine Learning), and Spark Core, are supported by PySpark. As the only framework that allows Big Data analytics in distributed processing and ML in the same engine, Apache Spark has become a worldwide success.[57]

3.4 Azure Cloud Services - Databricks

Microsoft Azure, formerly known as Windows Azure, is Microsoft's public cloud computing platform. It provides a range of cloud services, including computing, analytics, storage and networking. Users can pick and choose from these services to develop and scale new applications or run existing applications in the public cloud.[58]

One of these services is Azure Databricks, a data analytics service platform optimized for the Microsoft Azure cloud services platform. This technology offers three environments for developing data-intensive applications: Databricks SQL, Databricks Data Science & Engineering, and Databricks ML.

Databricks SQL provides an easy-to-use platform for analysts who want to run SQL queries on their data lake, create multiple visualization types to explore query results from different perspectives, and build and share dashboards.

Databricks Data Science & Engineering provides an interactive workspace that enables collaboration between data engineers, data scientists, and ML engineers. For a big data pipeline, the data (raw or structured) is ingested into Azure through Azure Data Factory in batches or streamed in near real-time using Apache Kafka, Event Hub, or IoT Hub. This data lands in a data lake for long-term persisted storage in Azure Blob Storage or Azure Data Lake Storage. As part of the analytics workflow, use Azure Databricks to read data from multiple data sources and turn it into breakthrough insights using Spark.

Databricks ML is an integrated end-to-end ML environment incorporating managed services for experiment tracking, model training, feature development and management, and feature and model serving.

3.5 Automatic Machine Learning

There has been significant progress in the creation of user-friendly ML software (e.g. Auto-PyTorch[59], Auto-sklearn[60], Tree-based Pipeline Optimization Tool (TPOT)[61], H2O Automatic Machine Learning (AutoML)[62]), focusing on the development of unified interfaces for a wide range of ML techniques. [62]

Although these tools have made it easier for non-experts to train ML models, there is still a good amount of knowledge necessary to produce cutting-edge results. AutoML technologies provide a simple interface for training a large number of models (or a powerful single model), and can be useful for both novice and professional ML practitioners.

Simplifying ML model training and tuning by providing a single function to replace a process that would normally require many lines of code frees up the practitioner to focus on other aspects of the data science pipeline, such as data preprocessing, feature engineering, and model deployment.

This section will offer tools that are being explored throughout the development process, such as TPOT and H2O AutoML tools are related to exploration in the AutoML approach.

3.5.1 TPOT

The TPOT library is considered a data Science assistant. Technically, it is a Python tool that automatically optimizes ML pipelines using genetic programming, it is under active development and is built on top of scikit-learn. [61] TPOT will intelligently explore thousands of potential pipelines to find the best one, automating the most time-consuming aspect of ML.

The data cleaning process is entirely the user's responsibility; afterwards, the data is presented to the TPOT, which performs a feature selection, a feature preprocessing, and a feature construction, all of which intercommunicate to pass information with each other; all of these data are introduced into the Model selection and are later optimized using its parameters. It is possible to see the process diagram in Figure 12.

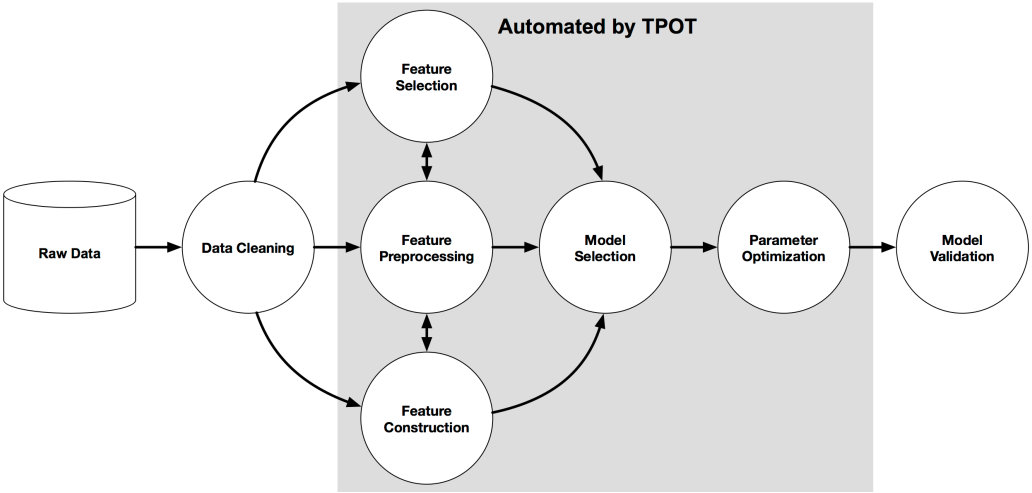


Figure 12: TPOT example Machine Learning pipeline.[61]

3.5.2 H2O - AutoML

The H2O AutoML algorithm was first released in June 2017, and it is a straightforward automatic ML technique included in the H2O framework that creates high-quality models appropriate for deployment in a corporate context. H2O AutoML allows supervised training of regression, binary and multi-class classification models on tabular datasets. The ability of H2O models to provide fast predictions is one of its advantages. H2O AutoML can be used in many languages (R, Python, Java, Scala). It's also useable through *Flow*, a web interface that makes autonomous ML more accessible. H2O provides several model explainability methods that may be used for AutoML objects (groups of models) and individual models. With a single function call, explanations may be created automatically, giving a convenient interface for exploring and explaining AutoML models.

In Figure 13 is possible to observe the methodology suggested by H2O.

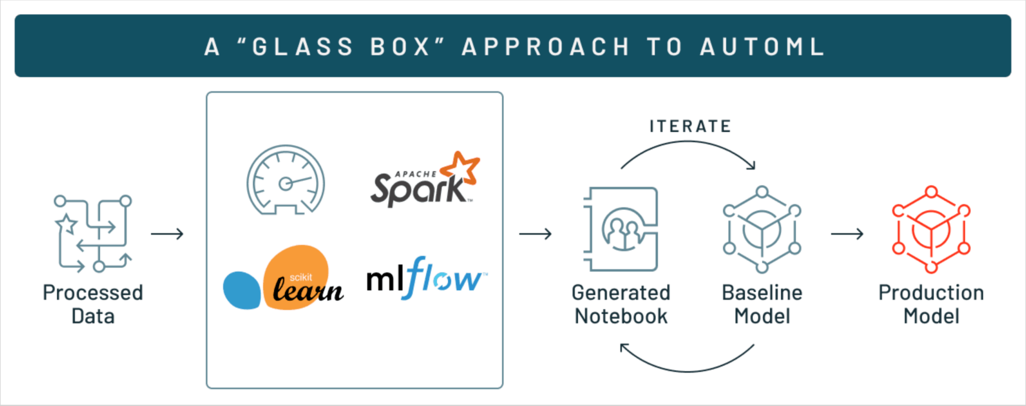


Figure 14: Databricks representation of glass-box approach. [64]

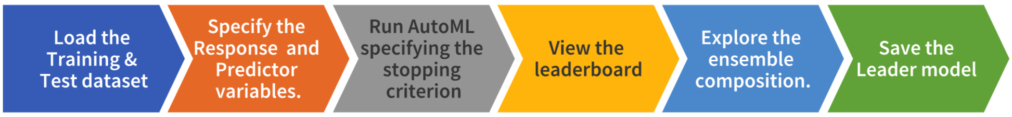

Load the Training & Test dataset
Specify the Response and Predictor variables.
Run AutoML specifying the stopping criterion
View the leaderboard
Explore the ensemble composition.
Save the Leader model

Figure 13: H20 AutoML Methodology.[63]

3.5.3 Databricks AutoML

Databricks AutoML is an automatic framework that generates ML templates according to user specifications. There are multiple tools on the market that already do this process but Databricks approach to this problem is rather unique. The core of the concept is an open approach called glass-box; the aim of this different concept is to give the users the opportunity to visualize and examine the generated code, the Figure 14 as a graphical representation of this approach. [64]

Scikit-learn decision trees, random forests, and logistic regression models can be used to produce classifications challenges. Similar to decision trees and random forests, linear regression with stochastic gradient descent can produce models for regression issues. Models for XGBoost and LightGBM can be created for either of these types of issues. Prophet and Auto-ARIMA are two models that can be used in forecasting.

The workflow is divided by runs and by experiments; each experiment can contain one or more runs; a table with the following columns/information is created for each experiment: Start Time, Duration, Run Name, User, Source, Version, Models, Metrics, Parameters, Tag; Table 6 shows an example of some runs in a classification experiment. To start an experiment, a dataset present in the *Azure Feature Store* is selected, and the target variable of prediction is chosen. The type of problem (regression, classification, or forecasting) and the maximum running time must be selected.

After the experiment, the user receives a notebook for data exploration and one notebook for each model created, Table 6 represents the final outcome of a experiment, in which lists all of the models that were tried during the experiment along with the associated details organized by model accuracy criteria.

	Start Time	2 days ago	2 days ago
	Duration	4.1 min	
	Run Name	XGBoost	LightGBM
	User	*email*	*email*
	Source	Notebook: XGBoost	Notebook:LightGBM
	Version	-	-
	Models	sklearn	sklearn
Metrics	traning_accuracy_score	0.917	0.353
	traning_f1_score	0.893	0.305
	training_log_loss	0.555	0.305
	(...)	(...)	(...)
Parameters	Classifier	XGBClassifier(...)	LGBMClassifier(...)
	(...)	(...)	(...)
Tags	estimator_class	sklearn.pipeline	sklearn.pipeline
	estimator_name	Pipeline	Pipeline
	(...)	(...)	(...)

Table 6: Databricks experiment exemple

4

Implementation

This chapter details the solution and methodology used to build the models. The section 4.1 describes the environment and the version of the tools used. The section 4.2 section describes the data from the different databases. The section 4.3 summarizes the implemented solution.

The sections 4.5, 4.6 and 4.7 describe the development points and techniques used.

4.1 Exploration environment

For the development of the technologies addressed in this first part of the project, a laptop was provided with the following characteristics: i) Intel Core i7–4600U@2.10Ghz; ii) 16GB RAM; iii) Windows 10 Enterprise Build 19042.

The internal development infrastructure of PBSS is centralized in *Microsoft Azure*, the company's team provided a cluster inside *Microsoft Databricks* composed of a master node with 28GB RAM and 8 CPU cores and one to two workers with 14GB RAM and 4 CPU cores. For environment replication purposes, the software versions used were: Databricks 0.14.3; Spark 3.1.2; Scala 2.12; MlFlow1.20.2

4.2 Dataset's

Two datasets from different sources were used as the project evolved; the first, presented in sub-section 4.2.1, shows the beginning of the statistical exploration of the COLEP database. The PBSS management software database is described in the sub-section 4.2.2.

Due to confidentiality concerns, it is not possible to identify all the fields in this report, with the exception of the fields used in the procedures. Also, all data has been anonymized beforehand.

4.2.1 COLEP

During the course of the *PRODUTECH* project, several files were provided by COLEP containing real records of the production lines, monthly stock of items, orders and among others; for the development of the project, two files in "*.xlsx*" format relating to orders of materials to suppliers were particularly important.

The first file contains information regarding the quantities of material ordered; the second file contains one or more records of the receipt of the order (the order can be delivered by one or more suppliers and can be divided into one or more deliveries).

The document related to supplier orders contains 17 columns, Table 7 provides a record that fits this structure (but does not belong to the datasets provided, for the reasons already indicated).

Column Name	Code identifier variable	Description	Sample data
Ordem Compra	purchase_doc	Purchase Order identifier	1234567890
Material	material	Material identifier	11222
Fornecedor	supplier	Supplier identifier	11–12345
Qtd Pedida	qty_ordered	Quantity requested	250
Qtd Recebida	qty_received	Quantity delivered	250
Dt Remessa	date_delivery_expect	Date the order expected to be received	09/01/2020
Prazo	max_delivery_time	Maximum time to deliver	11222
DT Ordem Compra	date_place_order	Date the order was issued	02/01/2020
UMP	ump	Unit scale, e.g. "MIL" Thousands	e.g: MIL, FOL, KG

Table 7: Description and example of one order.

The last document, which relates to the receipt of orders from suppliers, has 18 columns; Table 8 offers a record that conforms to this format.

Column Name	Code identifier variable	Description	Sample data		
Material	material	Material identifier	11–12345	11–12345	11–12345
Quantity	quantity	Quantity	149,50	91,00	9,50
UMP	ump	Unit scale, e.g. "MIL" Thousands	MIL	MIL	MIL
Value	value	Price	96.13	32.79	32
DatDocumen	date_place_order		09/01/2020	09/01/2020	09/01/2020
Dta.Inçto.	date_delivery	Date the order was received	03/01/2020	05/01/2020	07/01/2020
Doc.compra	purchase_doc	Purchase Order identifier	1234567890	1234567890	1234567890
Supplier	supplier	Supplier identifier	11222	11222	11222

Table 8: Description and example of receptions of one order reception.

To be noted, on Table 7 it displays one order, and in Table 8 it shows not one, but threes receptions to that order, one order can have one or multiple deliveries.

The orders dataset has a total size of 12 Megabyte (MB) containing approximately 7380 records ranging from 08/01/2020 to 30/12/2020 and the order reception file has 54MB contains approximately 25400 records, with dates ranging from 2013/11/11 to 2021/06/18.

4.2.2 v10

The database in v10 is of considerable size, due to its extent there was a process of exploration and learning how it is built, after exploration, multiple columns were pointed out as useful data to use in data exploration, but only the table *CabecCompras* which contains the purchases headers and the table *LinhasCompras* which contains the purchases made by the company were used.

To obtain the data the following SQL code line introduced in Listing 4.1 was executed, in order to discard the lines in which there is no associated product, a condition was imposed not to count these cases.

Listing 4.1: SQL querry for data extraction

```sql
use DATABASE_X
SELECT LC.*,  CC.*
FROM LinhasCompras AS LC
LEFT JOIN CabecCompras AS CC ON LC.IdCabecCompras = CC.Id
WHERE LC.Artigo IS NOT NULL
```

The query results in a total of 27589 lines and has a total size of 27MB ranging from 18-08-2017 to 17-02-2020.

Table 9 represents the outcut of the query in Listing 4.1, this is constituted by 3 columns: the original name present in the v10 database, the name portrayed in the developed code and the data type that it represents;

Code identifier variable	Type	Database Variavel Identifier
item	String	Artigo
entity	String	Entidade
quantity	Float	Quantidade
unit	String	Unidade
price_unit	Float	PrecUnit
total_tax	Float	TotalIva
others_total	Float	TotalOutros
discount_total	Float	TotalDesc
fiscal_space	Integer	EspacoFiscal
tax_regime	Integer	IVA
total_doc	Float	TotalDocumento
payment_method	String	payment
date_introduction	String	DataIntroducao
date_due	String	DataVencimento
date_delivery	String	DataEntrega
load_date	String	DataCarga
load_localization	String	LocalCarga
load_country	String	PaisCarga
delivery_date_hour	String	DataHoraDescarga
delivery_localization	String	LocalDescarga
delivery_country	String	PaisEntrega
delivery_postal	String	CodPostalLocalidadeEntrega
expedition_method	Float	ModoExp

Table 9: Description and example of receptions of one order.

4.3 Architectural insight concept

Two diagrams that attempt to explain the general idea of the architecture and the flow of data during the execution of the code were developed to provide a general understanding of how the implementation of the entire ecosystem to be created is anticipated.

It is possible to see two starting points in the Figure 15 (points 1 and 2); the first point describes the origin of the data and its processing: the data that don't need any kind of processing to go directly to *Azure Data Lake Storage*, while the remaining data must go through processing;

The second point refers to the suppliers and the criteria that the company values the most; these values come from three possible sources: DM's in the form of filling in forms illustrated in Figure 27 and Figure 28; calculated classification from values present in *Azure Data Lake Storage*; or in the case of classification criteria from the scientific literature.

Following the treatment of the data, the values go through a process of metric creation in which the goal is to simplify the data so that the difference between values is more perceptible. In the area of supplier classification, the values go via Fuzzy TOPSIS to be mathematically classified.

The Azure databrick AutoML was used to create ML models. This simplification of the ML model creation process is more valuable due to data generalization. The goal of this architecture is not to respond to a specific problem but rather to respond to a generic problem, allowing, for example, the use of this solution by any PBSS client or another Enterprise resource planning (ERP) Software.

In general, this architecture does not focus on problem solutions but rather on the development of processes that result in a solution to the problem. The generality of this solution is one of the strongest points that allow users to choose which procedures to use as well as which priorities to emphasize.

Figure 16 shows one of the architectural diagrams that make the most sense in a solution of this type; this type of diagram shows the flow of data; to do so, it is necessary to identify which are the canonical data, and which are the treatments that will reveal the final result of the operation to the user.

In this case, the process starts by extracting a specific set of data from the software's database; after treatment and classification of the providers, we obtained a set of data to be introduced into AutoML; these data are composed by: item identification, supplier identification, order quantity, order total price, date of purchase, deadline of delivery, and, the suppliers classification obtained through a pre-processing stage.

The predicted consequence will be the probability of each supplier being chosen for the purchase of a specific product, suggesting that each source has a possibility of being chosen.

All of this architecture was developed with users with a significant history of orders in mind, given that for the ML model to have a prediction of which provider is best, this information must be provided in the database.

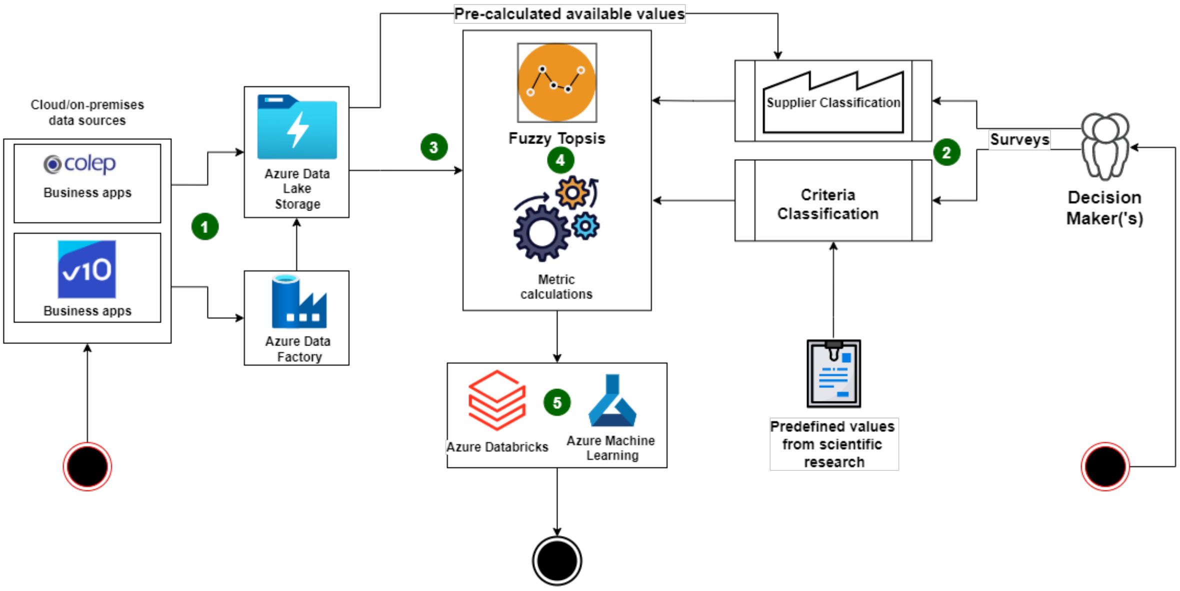
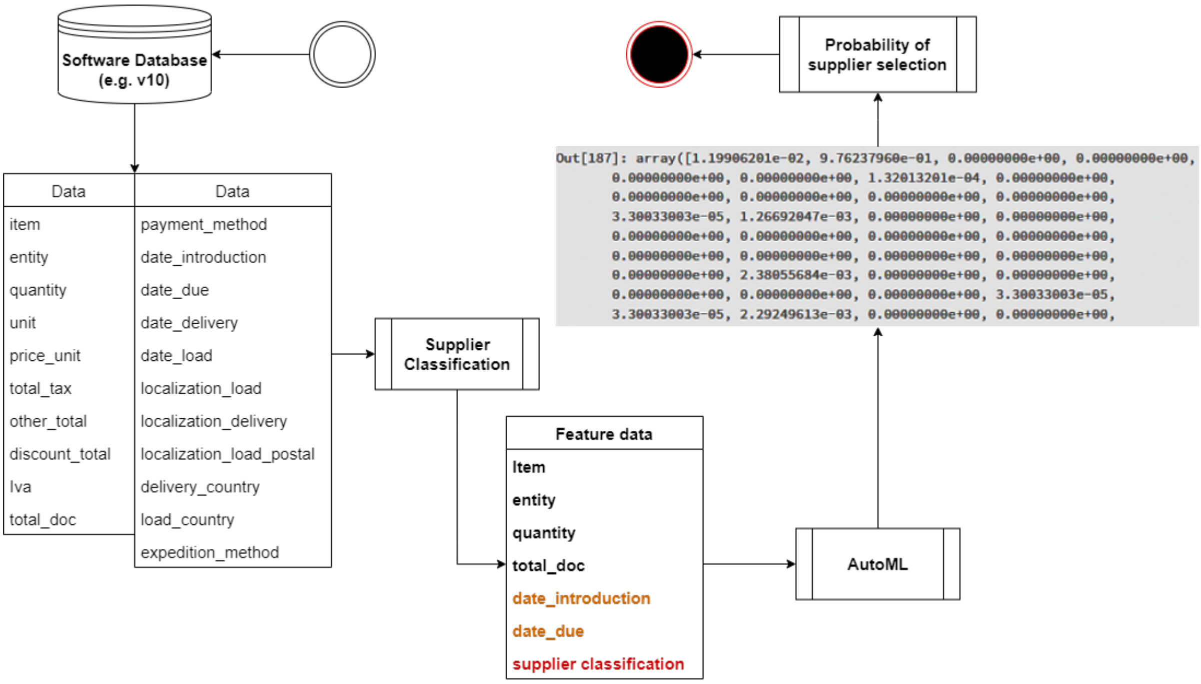


Figure 15: Architectural concept diagram



```
Out[187]: array([1.19906201e-02, 9.76237960e-01, 0.00000000e+00, 0.00000000e+00,
       0.00000000e+00, 0.00000000e+00, 1.32013201e-04, 0.00000000e+00,
       0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
       3.30033003e-05, 1.26692047e-03, 0.00000000e+00, 0.00000000e+00,
       0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
       0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
       0.00000000e+00, 2.38055684e-03, 0.00000000e+00, 0.00000000e+00,
       0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 3.30033003e-05,
       3.30033003e-05, 2.29249613e-03, 0.00000000e+00, 0.00000000e+00,
```

Figure 16: Architecture Data Flow

4.4 Supplier selection criteria

After assessing the literature presented in chapter **??**, a group of multiple criteria variables was compiled to pursue to best satisfy the problem's description; this compilation can be observed in Table 10.

Each SSC will have a weight associated with it, which will be either a pre-defined weight or a weight introduced by one or more decision-makers.

The information from Table 3 was used to calculate the predetermined weights based on the frequency with which each criterion appears in scientific articles. This ranges from very low to very high.

For the manually introduced weights, surveys were developed and can be seen in Figure 27 and Figure 28 this manual classification will be in the same range as the last instance.

This solution was chosen for its versatility; there is already a pre-calculated value calculation based on the scientific community, and if the company needs to customise or adapt this approach to the needs that most value the same, it can do so.

Variable group	Variable
Service Perfomance	Product Price
	Tariff and Taxes
Supplier profile	Delivery Compliance/Performance (Lead Time)
	Financial Position/Situation
	Product Quality/Reliability
	Technological Capability
	Conformance to specification
	Service/Relationship
	Market reputation
Service Perfomance	Flexibility and responsiveness
	Total logistics management cost
	Geographical location
Supplier profile	ISO 14001 (Environmental factors)
	Production Capability
Service Perfomance	Customer response/communication
	Reaction to demand change in time
Supplier Profile	Facility and infrastructure
	Innovation
	Information sharing
Supplier profile	ISO 9001 (Quality assessment technique)
Risk	Political stability and foreign policies
Service performance	Stable delivery of goods
Quality	After-sale/Warranty
Supplier Profile	E-transaction Capability
Risk	Exchange rates and economic position

Table 10: Supplier selection criteria according to projects needs.

4.5 Custom package development

The infrastructure of Azure Databricks necessitated the creation of a custom package containing the functions required for project development to guarantee that all cluster workers had access to the same User-defined functutions (UDF) functions; otherwise, if the code were developed directly on the notebook, the cluster workers would not have had access to these operations, resulting in an execution error.

The four classes that compose this package are:

- **Fuzzy TOPSIS**: All of the mathematical calculations for the Fuzzy TOPSIS operations were developed, and in response to the necessity, with the help of *Databricks datagen*, a function was created to generate synthetic data following the pre-determined execution specifications. The developed algorithm not only has a dynamic component that can compute many decision criteria according to their weight, but it also can keep track of multiple decision-makers.

- **Functions**: This class includes data science-related methods such as ordinal encoding and data normalization, which are used as needed throughout the code.

- **Feature store**: To run an experiment in AutoML, it is necessary to store the intended data in the feature store; as a result, two functions (one for reading and one for writing) were performed; In addition to the necessity of these functions for AutoML, it is important to store the same data in the Feature Store so that this pre-processing will not have to be repeated every time, for instance, the cluster disconnects. This is because the data transformation requires several processing steps that take time and processing power.

There is additionally a "v10" class, which is shown in the sub-section 4.7.2. This class will be utilized exclusively in creating features for the v10 solution.

4.6 COLEP development

The data analysis focused on understanding which data was present in the database suitable for the problem. A review and exploration were carried out on those to understand their condition.

The data exploration was conducted simultaneously with the de-assessment based on that exploration, which enabled the development and futher transformation of the data for efficient and more direct understanding; This behaviour is known in the field of data science as "feature engineering" or "feature extraction".

After the analysis of the two datasets, the following major points can be highlighted:

- **"value":** contains a total of 2.5% of their values at 0;

- **"max_delivery_time":** This field has values at 0, and others miscalculated, meaning the need to re-calculate these values due to their importance;

- **'date_delivery":** about 25% of this field even presented as "0000/00/00", which may be considered an error;

It has been confirmed that no orders were received before the order issue, so it can be verified that the data is correctly placed in the time frame.

As mentioned in subsection 4.2.1, despite the interception of dates existing in both datasets, not all data in the orders file can be related to the respective delivery date, in Figure 17 we can see the total number of records on the dates present in the orders file, and in Figure 18 it appears with the same but for the receipts file.

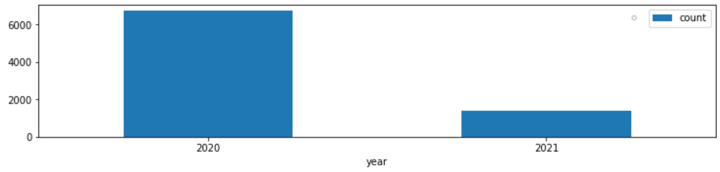


Figure 17: Total number of records per year present in the orders dataset.

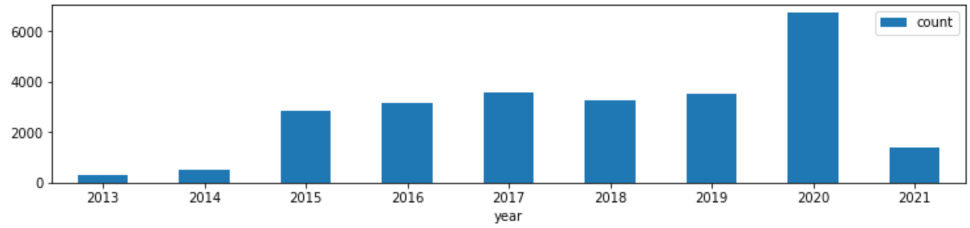


Figure 18: Total number of records per year present in the dataset of receptions.

4.6.1 Aggregation of data.

To achieve the final dataset, it was necessary to join the order and reception data, the first and second attempts were a simple merge (*pandas.DataFrame.merge(how='left' (...))*) with different columns that triggered the merge in that order, and lastly, the use of a robust tool (*pandas.merge_asof*) was used.

Attempt 1. *pandas.DataFrame.merge(right=orders, on=["purchase_doc", "material"], how="left")*

Using an index or a key column, join columns with other DataFrames. Giving a list makes it possible to combine several DataFrame objects together quickly. In Figure 19 and Figure 20, it's possible to observe the total of null values present in order and reception dataframe.

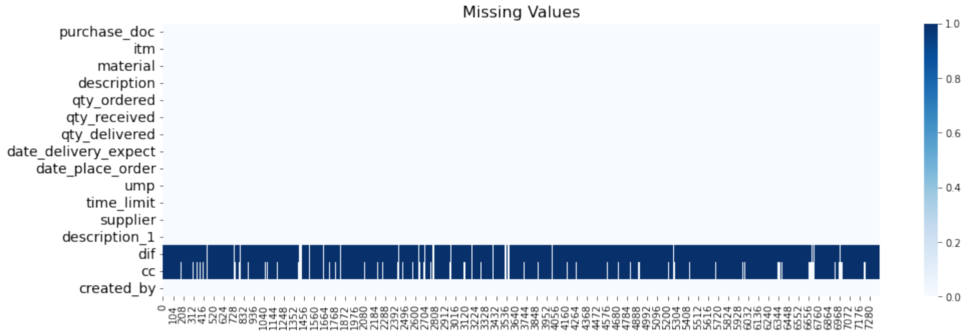


Figure 19: Total null values present in orders dataframe.

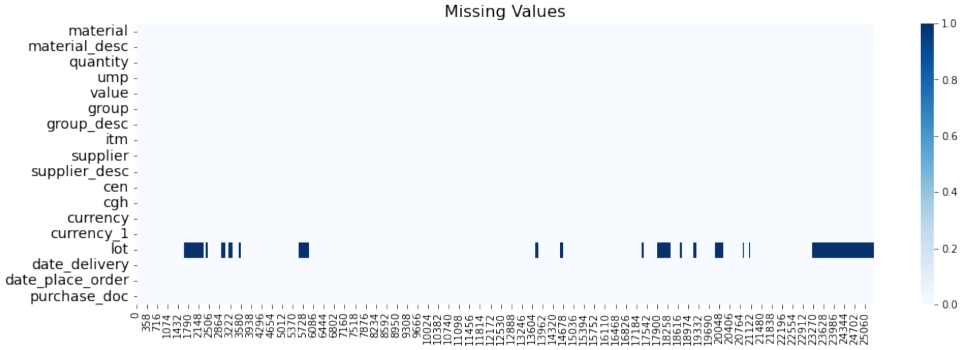


Figure 20: Total null values present in reception dataframe.

On the other hand, Figure 21 shows the total of null values on the resulting merged dataframe.

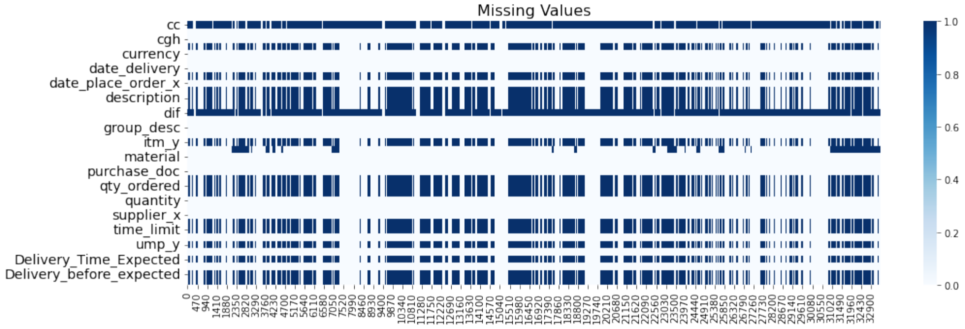


Figure 21: Total null values after joining the dataframes via *pandas.DataFrame.join*.

When looking at the presented figures, it is essential to note that the number of data with null values increased significantly. When looking directly at the results, it was straightforward to conclude that the results were not what was expected, which led to the exclusion of this method.

Attempt 2. *pandas.DataFrame.merge(left= orders, on=[“purchase_doc”, “material”, “date_place_order”, “supplier”], how=”left”)*

There was an attempt to increase the number of variables linking the two datasets, and on some of these variables, data treatment methods were used. The ordinal encoding method was used in the “material” and “supplier” fields. When working with categorical data for ML algorithms, encoding is a necessary pre-processing step; ordinal encoding is utilized for categorical variables that have a natural rank order.

After analysing the results, this method also encountered the same issue as the initial attempt; hence, this attempt was also dropped.

Attempt 3. *pandas.merge_asof*

This robust function is very similar to left-join, with the exception of matching on the closest key rather than equal keys. Also, the key must be used to order both DataFrames.

There are three different types of options for the link method between the keys when the function is parameterized: the last row in the right DataFrame whose “on” key is less than or equal to the left’s key is chosen by a “backward” search; the first row in the right DataFrame whose “on” key is greater than or equal to the left's key is chosen by a "forward" search; and the row in the right DataFrame whose "on" key is closest to the left's key by an absolute distance is chosen by a "nearest".

Listing 4.2: Code using merge_asof.

```
In [1]: dataframe = pd.merge_asof(right=recp.reset_index(),
            left=ords.reset_index(),
            right_on=['date_delivery'],
            left_on=['date_delivery_expect'],
            direction='nearest',
            by=['purchase_doc', 'supplier','material', 'date_place_order'],
            suffixes=('', '_recp')
            )
```

After verifying the results obtained through the 4.2 code, it is possible to affirm that this is the best approach to correctly join the two dataframes.

4.6.2 Feature Enginnering

A new column "delivered" was created that explains the order status, if the delivery date is after the expected delivery date, the column value is assumed to be "0", since the deadline for delivery has not been met on time. However, if the delivery is before or on the expected delivery day this value is summarized as "1"

The number of days in transport and the number of days that the delivery has passed (positive values) or is in anticipation (negative values) of the expected delivery date were calculated.

A ranking by the provider was created using a very common technique in ML forecasting techniques; this technique involves grouping a section of data based on the number of days. In this instance, 30 days were determined to be the best value to provide further explanation for the frequency of orders by a provider. In addition to the ranking that is determined by the data available in the "delivered" column, the following columns were created:

- "delivery_efficacy": Develivery effecacy;
- "quantity_avg" & "quantity_std" : average and standard derivation of delivery quantity;
- "avg_days_failed": average number of days the order was past due;
- "avg_DeliveryTime" : average delivery time;
- "avg_Price" : average pricing;

Due to the significance of the magnitude of the materials, a collection of regular expressions that grouped the requests in the appropriate lists of primary materials was explored; however, this approach required more information from COLEP and more implementations to cover all the cases study. So this exploration was held it.

4.7 v10 development

The exploration and cleaning of the data followed the implementation described in section 4.6 and were implemented in two stages, the first using a pandas approach in a local environment and the second using PySpark in an Azure Databricks environment. Both methods attempted to achieve the same impact and purpose using various techniques based on the capabilities and best practices of the framework in question.

In this way, the final notebook was modified to allow it to run in both locations; all required is to specify which environment it is intended to run in.

4.7.1 Metrics Calculation

Through the variables necessary to calculate the classification of the suppliers, it was possible to make a mapping between the classification variables and the columns of the database, as shown in Table 11.

Variable group	Variable	Is the software or the company using this?	Variable Mapping
Service Perfomance	Product Price	X	price_unit
	Flexibility and responsiveness		
	Total logistics management cost	X	others_total
	Geographical location	X	distance metric
	Customer response/communication		
	Reaction to demand change in time		
	Stable delivery of goods		
	Tariff and Taxes	X	total_tax
Supplier Profile	Delivery Compliance/Performance (Lead Time)	X	date_due date_delivery
	Financial Position/Situation		
	Product Quality/Reliability		
	Technological Capability		
	Conformance to specification		
	Service/Relationship		
	Market reputation		
	ISO 14001 (Environmental factors)		
	Production Capability		
	Facility and infrastructure		
	Innovation		
	Information sharing		
	ISO 9001 (Quality assessment technique)		
	E-transaction Capability	X	payment
Quality	After-sale/Warranty		
Risk	Exchange rates and economic position	X	IVA
	Political stability and foreign policies		

Table 11: v10 database matching supplier selection variables

In the context of simplifying the calculations of this classification, there was an effort to develop small algorithms that translate the values into metrics in the columns where this treatment is realistic and possible;

4.7.2 Custom package - v10

Given that the geographic location columns for the pickup and delivery locations are already present in the v10 database, it was necessary to develop functions that calculated the distance between these two points to make the automated entry of the "Geographic location" decision-making criterion feasible;

To address this issue, three functions were developed.

The first uses the Haversine formula to calculate the distance between the locations; if the outcome of this calculation comes to a failure, the second function uses a database to convert the locations into coordinates, guaranteeing that as long as data exists, there will always be a distance between the sites.

To determine whether the delivery deadline was met, it was necessary to calculate the difference between the purchase date and the delivery date. In cases where the delivery deadline has been omitted or was not specified, it is possible to add days to the purchase date to make the delivery date a set of synthetic data that can be used to provide the desired outcome.

To aid this procedure, a helper function was developed to ensure that the DateTime fields are in the format required for their use in other functions. In this way, the final automatic procedure responds to the decision-making criterion "Delivery Compliance/Performance (Lead Time)".

4.7.3 Feature Enginnering

The notebook implementations on feature engineering consist of multiple features:

Features 1. Missing delivery date - synthetic data generated

One of the issues with this perticular raw data was the variable "date_delivered", which had almost 93% of its values missing. In order to have a sense of the supplier performance, it was necessary to calculate this value mathematically:

The conditions in Listing 4.3 make use of the variable "mtr_perc_remaining_time", which was pre-calculated by dividing the entire delivery time (from the day of order issuance to the day of delivery) and the remaining delivery time (from the day the order was sent to the day it was delivered).

Listing 4.3: Percentage remaining delivery time conditions.

```
In [1]:  cond1 = v10.mtr_perc_remaining_time<=0.8
         cond2 = (v10.mtr_perc_remaining_time>0.8)&(v10.mtr_perc_remaining_time<1)
         cond3 = v10.mtr_perc_remaining_time >= 1
```

The delivery date is calculated synthetically using the *faker* package (explained in subsection 3.2); for the values that are checked in the first condition, a date between the day of load and the due date is generated; in the second condition, 15% of the remaining time is added, allowing you to obtain values that reveal deliveries within and outside of time; and in the third case, 50% of the remaining time is added. Figure 22 aims to explain this algorithm in a diagram.

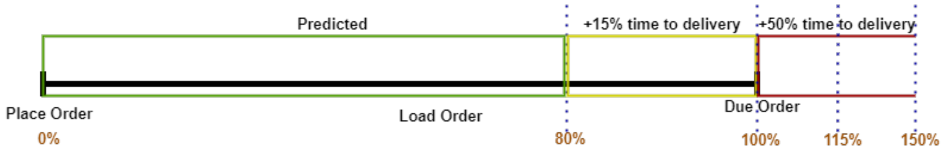


Figure 22: Diagram that represents the synthetic delivery data algorithm.

Features 2. Delivery performance calculations

By comparing the due date with the new scheduled delivery date, the result is saved in a variable called “mtr_deliveried” and it is then determined whether the order was delivered on time; if so, the value is equal to 1, otherwise it is equal to 0; with this value, you can determine the supplier’s delivery performance.

Features 3. Features by 30 days

Similar to the COLEP notebook, in v10, the same grouping of data was made for the same number of days (30 days); this way, the following metrics were calculated: average delivery time, average remaining delivery time, standard deviation quantity, sum and total variable “delivered” (the sum reveals the number of times the delivery was past the estimated time) and the mean of quantity for each order.

Features 4. Product description overview

Some products present in the database have different descriptions with the same unique product ID, this type of inconsistency reveals a problem in the construction of ML models; it is necessary to clean these values. The first option would be a regular expression that detected these inconsistencies and disregarded these values, but after an analysis of the absolute amount that these values are present in the dataset, it was concluded that the best alternative would be to disregard the values that contained more than three descriptions for each item since this value was significantly low. The analysis of this problem can be observed in Listing 4.4.

In this case, products identified as 0 are considered generic products; this type of identification is typically used to represent products that lack a specific need.

When the user finds themselves in an irregular situation where the entry of an article has no bearing on the business but only serves to enter the system (such as an offer of samples), they recognise this product as a generic product. Following the discovery of this article and the conclusion that its significance for the resolution of this issue is extremely low, it was decided to disregard this category of products; in this case, all articles classified as 0 were removed.

Listing 4.4: Unique items by supplier.

```
In [1]:

    Total of Unique items:            1065
    Number of descriptions per item >1:  100
    Number of descriptions per item <2:  965
```

```
A total item with different descriptions: 1363
+-----------------------+-----+
description             | item|
+-----------------------+-----+
Envio Amostras A        |  0 |
Envio Amostras B        |  0 |
Envio Amostras C        |  0 |
Envio Amostras D        |  0 |
+-----------------------+-----+
```

Features 5. Automated variable ranking system.

In order to use the fuzzy TOPSIS method in this architecture, it is necessary for the DM('s) to fulfil a form that qualifies the supplier according to a number of criteria (Figure 28). In an effort to automate this process, a mechanism was developed that will classify the supplier on the same scale as the fuzzy TOPSIS.

To use this function, the dataframe, the ranking column or columns, the type of weight (beneficial or cost), and whether to disregard the values with zero must be entered as entry parameters.

The ranking's outcome follows the following order within the range of values [1,2,3,4,5], with value "1" denoting a "Very Low" ranking and value "5" a "Very High" ranking. If the chosen method is deemed to be "beneficial" the resulting order will reverse, becoming [5,4,3,2,1].

The first calculation to perform is the normalization of the data against the grouping of the data for "item" and "entity", which was done using the "min-max normalization" method: $x = \frac{x - x_{min}}{x_{max} - x_{min}}$; this normalization will always keep the values of the column valued between 0 and 1.

After the data has been normalized, the ranking is assigned according to the Table 12; this mapping aims to arrange the data according to the distance between the worst value and the best value possible. This value distribution aims to aggregate all values closest to the average as "Average" so that the more positive and negative values can be highlighted. The goal is shown in Figure 23.

Ranking		Value
Cost	Beneficial	
1	5	[0 , 0.025[
2	4	[0.025,0.15[
3	3	[0.15,0.50[
4	2	[0.50,0.985[
5	1	[0.985,1]

Table 12: Mapping of normalized value to ranking

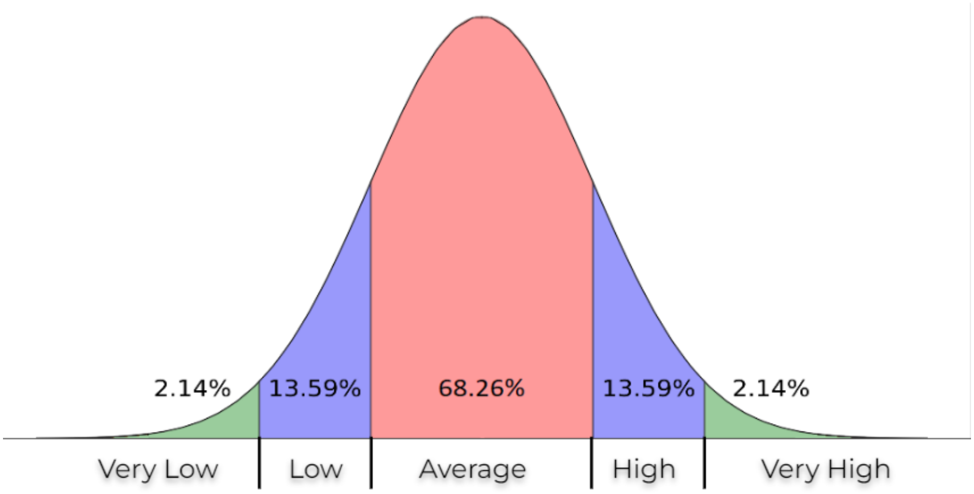


Figure 23: Goal of the ranking distribution.

The listing 4.5 provides an example of this procedure being used in practice.

Listing 4.5: Call function for ranking system

```
In [1]: ranking_sys(df,['mtr_delivery_distance','others_total'],'beneficial',False)
        ranking_sys(df,['price_unit'],'beneficial',True)
```

It is worth mentioning that there are functions in libraries that can be used to accomplish similar tasks, such as the cumulative distribution function (*cume_dis*) and the relative ranking function (*percent_rank*). Although the results were not congruent with the input values, the built algorithm revealed better results, as shown in Table 13.

item	entity	**Mean price**	**cume_dist**	**percent_rank**	**Developed algorithm**
5	2	0.72	3	5	5
	1	0.72	3	5	5
	4	0.72	3	3	5
	3	0.73	3	3	4
	31	0.79	2	3	3
	45	0.89	2	1	3
	18	1.31	1	1	2
	5	1.32	1	1	1
	55	1.32	1	1	1

Table 13: Comparison of algorithms in the ranking of variables

Features 6. Fuzzy TOPSIS.

To use the Fuzzy TOPSIS function, which is included in the developed custom package, it is necessary to set the following class variables to their default values: “COLS”, “COST_CRITERIA” and “COLS_WEIGHTAGE”. In this particular case, it is also necessary to correct the rankings’ values for the variables “FUZZIFIER”, since the default values are Very Low to Very High and due to the ranking values coming from the function being 1 to 5.

This implementation can be observed in Listing 4.6.

Listing 4.6: Example of the use of the function Fuzzy TOPSIS.

```
In [1]:
fz = TrabalhoMiguelRibeiro.FuzzyTopsis()
for itemrow in df_fuzzy.select('item').distinct().collect():
  fz.FUZZIFIER = {1:(1,1,3), 2:(1,3,5), 3:(3,5,7), 4:(5,7,9), 5:(7,9,9)}

  fz.COLS = ['delivery_distance','others_total','price_unit']
  #fz.COST_CRITERIA => 1: COST; 0: BENEFICIAL
  fz.COST_CRITERIA ={'delivery_distance':0,'others_total':0,'price_unit':0}
  fz.COLS_WEIGHTAGE ={'delivery_distance':1,'others_total':3,'price_unit':5}
```

Features 7. Minimum supplier per item.

This proposed solution understands the contents in the database and attempts to perceive the recommended supplier based on the company's priorities. As a result, certain constraints must be applied; after several tests, it was determined that the model performs best when the number of suppliers for each item exceeds three. To put it another way, every item with a history of purchases from two or fewer suppliers is unsuitable for use in the model; this group accounts for 10% of the dataset, in listing 4.7 is set this restriction.

Listing 4.7: Restriction on the absolute number of suppliers per item.

```
In [1]:
restrict = collect_set.select('item','collect_set(entity)').filter
  (f.size('collect_set(entity)') >2)
df = df.join(restrict , df.item ==  restrict.item,"left_anti")
```

5

Results

The results and approaches that were tried in relation to the application of the models will be described in this chapter's chapter 5. Following this, Section 5.1 describes the experiments conducted using the COLEP data base. Section 5.2 exhibits the procedure used to approach the v10 data base and its results.

5.1 COLEP Machine Learning

The solution to the COLEP dataset was the first problem to be addressed; after data analysis and the development of techniques for its treatment, approaches in AutoML were explored.

This process began by identifying various AutoML tools for creating models in ML through research and suggestions from members present at meetings (people from CCG, PBSS, and the project's coordinator). Several tools were suggested and explored, including TPOT, h2o AutoML and Databricks AutoML.

This exploration began with the use of TPOT, and after following the installation procedures outlined in the documentation, a prototype was built solely to determine the potential range of this tool. The best model produced by this method utilised the XGBoost algorithm and achieved outstanding results. Through the use of 10 fold cross-validation, this model achieved an accuracy score of 80%. [1]

Given that the use of this tool was only intended for familiarisation, integration, and potential analysis with tools of this type, h20 exploration began without further exploration on TPOT.

After a quick adaptation to the h20 guide, it was relatively simple to begin the automatic model construction process. Surprisingly, the XGBoost algorithm was utilised, and it achieved an accuracy score of 58% through 5 folds cross-validation, with a mean per class error of 0.19 and a rmse of 0.054.

As mentioned above, following this exploratory phase to become more accustomed to such tools and gain experience in the field of AutoML, there was a period of reflection and comparison between the tools and what they brought to the table.

Practically speaking, Databricks AutoML glass-box approach was the selling point for further exploration and more dedication. Apart from its construction, this tool was already pre-configured and did not

[1] Achieved with the following hyperparameters: learning_rate=0.5, max_depth=2, min_child_weight=1, n_estimators=100, and subsample=0.8.

require any additional installation or configuration, which is another reason for its use in the entire project context.

Concerning the COLEP dataset, due to the data quality and the lack of explanation from the data provider to some parameters, this approach has been placed on hold until new information is obtained.

Following this exploration, the goal shifted to the v10 dataset; however, the main objective will be the development of techniques and approaches that can be applied to the COLEP dataset, in other words, not being entirely focused on the v10 solution, but rather on the development of techniques and methods that reveal useful for both solutions.

5.2 v10 Machine Learning

For the development on v10 database, only the Databricks AutoML framework was used, starting on the code shown in the list 5.1, that is used to launch the AutoML experiment, the parameters strictly required for execution are: the problem type, the data, and the column that is intended for the forecast.

Listing 5.1: Execution of Databricks AutoML.

In [1]:

```
import databricks
summary = databricks.automl.classify(dataframe, target_col=``entity'')
```

The databricks AutoML classification system is divided into four operations: Load Data, Pre-Processors, Training - Validation Split, and Train classification model. The first section begins by setting the “Run” to keep track of the same. The second section begins by loading data for the MlFlow client’s object, which is loaded from the feature store in parquet format. These two processing steps can be seen in Listing 5.2.

Listing 5.2: Example of experiment instigation and data storage.

In [1]:

```
import databricks.automl_runtime
from mlflow.tracking import MlflowClient
import os, uuid, shutil, mlflow
import pandas as pd

# Use MLflow to track experiments
mlflow.set_experiment(``experiment_name'')
target_col =``entity"

# Create temp directory to download input data from MLflow
input_temp_dir = os.path.join(os.environ[``SPARK_LOCAL_DIRS''],
    str(uuid.uuid4())[:8])
os.makedirs(input_temp_dir)

# Download the artefact and read it into a pandas DataFrame
input_client = MlflowClient()
input_data_path = input_client.download_artifacts(run_id, data,
     input_temp_dir)
df_loaded = pd.read_parquet(os.path.join(artifact_uri,training_data))

shutil.rmtree(input_temp_dir) # Delete the temp data
```

Using the scikit-learn library's preprocessing module, the "Run" begins by applying a set of typical ML processes, such as Numerical columns, Feature Standardization e resample rare classes; These are some of the most commonly used methods in the classification process. These pre-processing steps are critical because they allow for changes in data composition without compromising the data's true meaning.

In the case of ordinary values, feature standardization is used, which converts these values into unique numerical values, making comprehension more accessible for the ML model. The last algorithm used is known as Resample uncommon classes; it identifies all classes with a low number of records and applies a resample; this generates new data that is very similar to the existing data, ensuring that the dataset contains at least four records for each class.

Following the pre-processing, the model training begins with the creation of the "pipeline" of "sklearn. pipeline", with the tuning values that are applied based upon the other "runs" already executed in the experiment. Following this training, the AutoLog method from "mlflow.sklearn.autolog" is engaged, which generates a collection of constraints from various entry points, metrics, parameters, and models; resulting in a collection of values that can describe performance in classification. These performance values are calculated using the "estimator" method; for classification models, the metrics "precision score," "recall score," "f1 score," and "accuracy score" are quantified, but only when the probabilistic method ("predict proba") is used, the metrics "log loss" and "roc auc score" are added.

5.2.1 Product restriction.

Given the limitations described in step 7 of subsection 4.7.3, four experiments were carried out sequentially, applying one restriction after another. The results of the best models obtained in each experiment are shown in Table 14.

The first restriction applied was the exclusion of item 0 (identified as a generic product); this was followed by a restriction on the total number of suppliers registered under each product; and the two restrictions were then combined.

Best model	Random Forest		XGBoost		Random Forest		XGBoost	
Restriction	-		item 0 removed		n° suppliers/item >2		n° suppliers/item >2 item 0 removed	
	Train	Validation	Train	Validation	Train	Validation	Train	Validation
accuracy_Score	88.6%	83.1%	90.6%	85.9%	97.2%	87.1%	96.2%	93.3%
precision_score	82.2%	77.5%	85.5%	81.2%	97.1%	86.5%	93.9%	91%
recall_score	88.6%	83.1%	90.6%	85.9%	96.2%	87.1%	96.2%	93.1%
f1_score	85.1%	79.9%	87.8%	83.3%	97%	85.6%	94.8%	92%
roc_auc_score	97.1%	96.1%	97.5%	96.5%	100%	99.6%	99.8%	99.4%
log_loss	1.13	1.17	0.64	0.84	0.52	0.79	0.24	0.36

Table 14: Test performed to prove the performance of the restrictions applied.

It is possible to see from the Table 14 that the cross-validation evaluation metrics improved with each applied restriction, reaching incredibly optimistic values in the final limitation, particularly for the Log loss evaluation measure.

It's important to note that the second constraint produced more promising results than the third. Still, this restriction is crucial since it stops the model's training from considering data viewed as illegitimate.

5.2.2 Machine Learning results overview.

The results of the best models created through this entire process can be seen in Table 15, this table is order by the Validation accuracy score. Since the absolute number of rows presented to run this experiment is less than twenty thousand and more than a thousand rows, the cross-validation used three folds validation. The experiment that was conducted comes into contact with all of the pre-processing and analysis were done as previously described. It is important to note that this experiment includes all of the previously mentioned operations and restrictions, such as the elimination of the product "0" and the minimum number of registered suppliers per product.

This classification experiment was started in code, as described in the Listing 5.1, and had 3 hours of time limit to execute. The execution finished with a total of 5 types of algorithms being explored, including: **XGBoost**, **lightgbm**, **randomforest**, **decision tree**, and **logistic regression**.

All **lightgbm** algorithm models were completed in 50 minutes, which is 29 per cent of the total execution time, and reached the lowest accuracy levels both during training and validation.

The **XGBoost** model algorithm comes in at 59% of the execution time (1 hour and 45 minutes), with the 5 best results in the accuracy in training and validation metrics. The **decision tree** models never achieved an accuracy greater than 50% and had a processing time of 5% (10 minutes). The two fastest algorithms were **random-forest** and **logistic regression**, which took together 6% (15 minutes). Here, the random forest stands out due to its accuracy in training between 80% and 86 % and in validation between 78% and 74%; meanwhile, the random forest yielded results similar to the **decision tree** and only managed to reach the 40% mark in both performance metrics.

Following an analysis of these results, it was determined that the **XGBoost** algorithm produced the best results and proceeded to a more detailed analysis of the same.

Multiple models are developed in an effort to achieve the best model possible; This approach in Databricks AutoML is possibly carried out through comparisons with existing models; The tool starts the process by creating various models using various algorithms, analyses the results using cross-validation, and then executes the next model training session based on those results. The process ends when the time limit is reached or the system recognises that there are no more potential optimisations.

In this example, it is possible to see that different models were developed for the various algorithms. It is possible to recognize that some of these algorithms had more generations after this alteration of hyper-parameters while others had significantly fewer generations.

Algorithm	Duration	Training Accuracy Score	Validation Accuracy Score	Training Precision Score	Validation Precision Score	Training Log Loss	Validation Log Loss	Training F1 Score	Validation F1 Score	Training Recall Score	Validation Recall Score	Training Roc Auc Score	Validation Roc Auc Score
xgboost	16m	0.96	0.93	0.94	0.91	0.24	0.36	0.95	0.92	0.96	0.93	1.00	0.99
	10m 30s	0.93	0.91	0.89	0.87	0.41	0.52	0.91	0.89	0.93	0.91	0.98	0.97
	18m 35s	0.89	0.87	0.87	0.85	1.13	1.19	0.87	0.85	0.89	0.87	1.00	0.99
	8m 50s	0.92	0.87	0.87	0.83	0.55	0.73	0.89	0.85	0.92	0.87	0.96	0.95
	17m 5s	0.87	0.86	0.84	0.82	0.92	0.96	0.85	0.83	0.87	0.86	0.99	0.99
Random Forest	3m	0.86	0.79	0.85	0.77	1.02	1.20	0.84	0.75	0.86	0.79	1.00	0.99
	2m	0.81	0.74	0.79	0.73	1.18	1.37	0.77	0.70	0.81	0.74	1.00	0.99
xgboost	19m	0.72	0.71	0.67	0.65	2.95	2.96	0.68	0.67	0.72	0.71	0.98	0.97
Decision Tree	2m 15s	0.51	0.51	0.48	0.47	1.84	2.03	0.47	0.47	0.51	0.51	0.88	0.87
	2m 35s	0.47	0.47	0.42	0.41	2.10	2.20	0.42	0.42	0.47	0.47	0.84	0.84
xgboost	15m 20s	0.42	0.41	0.28	0.27	3.95	3.95	0.31	0.30	0.42	0.41	0.83	0.83
Decision Tree	1m 15s	0.41	0.41	0.26	0.26	2.16	2.29	0.30	0.30	0.41	0.41	0.91	0.90
Logistic Regression	1m 20s	0.42	0.41	0.34	0.30	2.27	2.36	0.33	0.32	0.42	0.41	0.94	0.92
	2m 40s	0.42	0.41	0.34	0.30	2.27	2.37	0.33	0.32	0.42	0.41	0.94	0.92
Decision Tree	1m 20s	0.38	0.38	0.21	0.21	2.61	2.66	0.25	0.25	0.38	0.38	0.75	0.74
Logistic Regression	2m 10s	0.37	0.36	0.29	0.25	2.57	2.62	0.26	0.26	0.37	0.36	0.91	0.89
Decision Tree	2m 20s	0.34	0.33	0.18	0.18	2.58	2.62	0.23	0.22	0.34	0.33	0.85	0.84
lightgbm	10m 20s	0.35	0.33	0.32	0.31	21.92	22.56	0.30	0.29	0.35	0.33	0.71	0.70
	31m 55s	0.23	0.23	0.05	0.05	3.45	3.45	0.09	0.09	0.23	0.23	0.60	0.60
	7m 30s	0.23	0.22	0.28	0.29	3.28	3.28	0.21	0.20	0.23	0.22	0.71	0.71

Table 15: Machine learning models runs performance.

5.2.3 Models results exploration.

Following the presentation of these results, the application of Explainable AI techniques was explored using a variety of tools. The purpose of using these tools emerges from the need for a more detailed analysis of the created model; the goal is to understand the significance and weight of the variables used on the created ML model. These techniques are typically used because of their impact on analysis and understanding of how to expand/improve the previously created model, so initiating a process of reformulation and re-construction based on the results of the analysis.

SHapley Additive exPlanations (SHAP) is a theoretic technique that may be used to explain the output of any ML model. It ties optimal credit allocation to local explanations by employing game theory's traditional Shapley values and their related extensions. [65]

This technology is already pre-installed in the Databricks AutoML environment; however, without further reasoning, it was not possible to use it directly from the development environment; so, installation and configuration of this tool were required. During this procedure, some errors occurred that were not anticipated during the learning process through the documentation, resulting in the removal of DateTime data types so that the tool could be used as intended.

The Figure 24 shows the importance of variables on an AutoML model created without DateTime variables (previously removed due to a limitation of the tool). According to this figure, it is possible to identify variables whose importance in the model's construction is irrelevant or non-existent, such as the quantity and total of the document (total of the order), relative to the variable that is considered the most important (CCi), which includes a small portion that has no bearing on the model and the remaining that has a significant bearing on the model. When it comes to the item field, which represents the unique identifier of the item in question, it is very understandable that this is the most important value, as this is what categorises the items. The model developed to run this experience on SHAP has a precision, f1 and accuracy score of 83%, a log loss of 64.5%, and a Roc auc of 0.996.

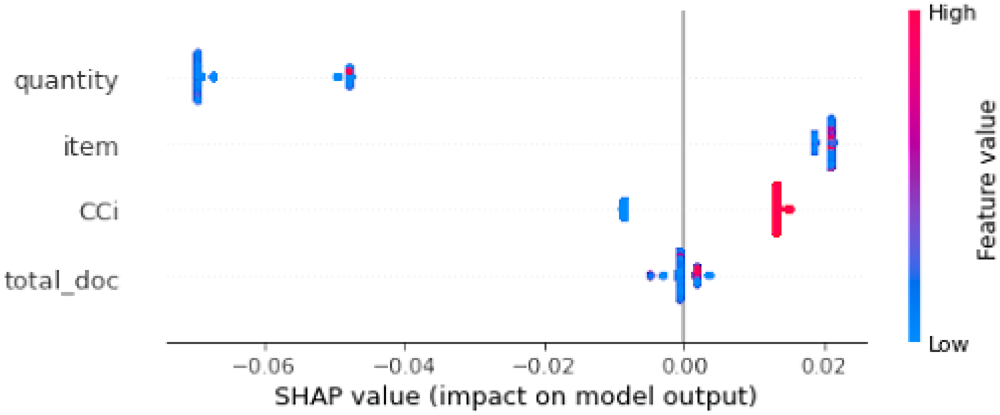


Figure 24: SHAP Feature value in Machine Learning model.

Due to the limitations of DateTime data types, the Explainable AI (XAI) tool was investigated. With this tool, it was possible to gain more significance in relation to the best XGBoost model's output.[66] During the validation process, it was possible to see which of the articles were correctly classified. It was possible to see that 52 products did not choose the correct supplier since the number of registries present was just three. The same thing happened in another 17, although the number of registries varied from 4 to 7 on the basis of data. The lowest number of records in the database with acceptable accuracy was the item 93, which had 53% with 20 records in the database. Through this analysis, it is possible to conclude that the minimum number of registries on the basis of data is an important study that should be taken into account in future versions of the model.

The figures 25 and 26 represent the correlations between the variables introduced in the model's construction. There is a correlation between the data variations since they are generally quite close in time, and the difference between these values also refers to very comparable values. There is also a small correlation between the item and the CCi. Because this value is calculated using three variables, the item, the performance, and the respective supplier, it is natural that the value for each item fluctuates.

The total of the order and the quantity values are typically correlative, since if one increases, the other increases as well, causing the ratio to change due to the price per unit of the respective item.

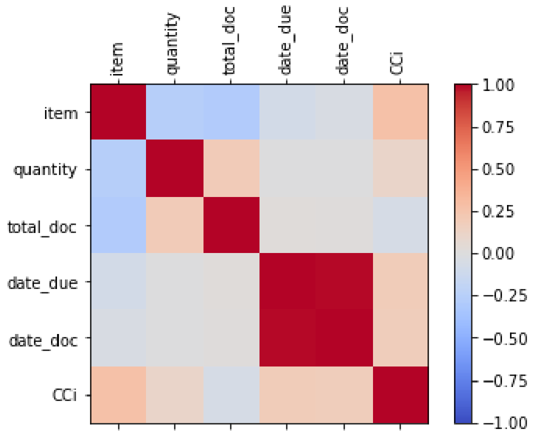


Figure 25: Correlation overview of XGBoost model.

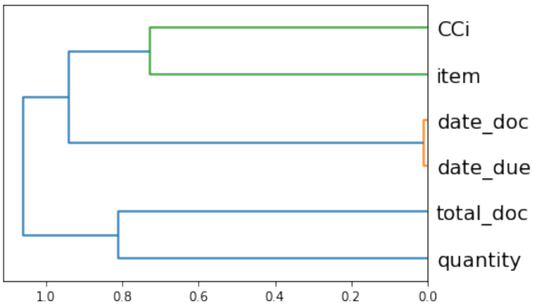


Figure 26: Correlation overview of XGBoost model.

6

Conclusions and Future Work

The current chapter 6 brings the document to a close by summarising the findings of the research. The main challenges and explanations during development are described in section 6.1. Finally, section 6.2 minimises some pertinent topics in order to support the future development.

6.1 Conclusions

The main goal of this work was to develop a ML model to predict which supplier or suppliers are the most appropriate for a particular order. After an intensive study and research on technical concepts and methodology already applied in this area, it was possible to outline a set of processes capable of achieving the desired goal.

Initially, changing the case study during process development was a major setback. However, in hindsight, this decision revealed essential and crucial aspects of the project: due to the changes, a solution that had previously only been intended to be used in a particular context has now been transformed into one that can address the needs of numerous businesses. These processes are a combination of scientific methodologies and mathematical methods, which have led to the current prototype, which was only related to one problem in one solution, to have the possibility of being contextualized in multiple case studies.

Through the use of AutoML from Databricks the whole process of creating ML prediction models was a great development breakthrough, due to its ease of creation and analysis through its glass-box approach. The approach on the COLEP case study revealed promising results, however, these cannot be accounted for due to lack of knowledge on some parameters and due to the quality and quantity of the data. On the opposite, in the software v10 approach, the results were quite positive. The AutoML results allowed to obtain several models with the use of XGBoost that revealed an accuracy higher than 87% in training and 86% in validation, reaching the best possible case of 96% in training and 93% in validation. Although this validation is one of the simplest validation approaches in ML it is not possible not to acknowledge these results.

The search for dynamic processes was the most promising aspect; the ability to give the user a decision-making criterion based on the company's priorities revealed itself to be very important in the adaptation of businesses to this process. To that end, the development of formulas was a critical step in making it possible to track these business priorities. To avoid a lengthy process, the development of mechanisms that automatically complete this process was time consuming, but it makes the entire process easier and more dynamic for the user.

The end result of all the processes and models created shows a promising ability to satisfy business needs, according to the implementation structure of the processes, it becomes possible to put this solution into production in the ERP software supporting the growth and evolution of the company.

6.2 Future

Several enhancements are possible in the future. The most challenging involves increasing the capacity of variables that the automatic supplier classification system is limited to, by adding date type field detection and others, largely dependent on the process's technical capabilities. Furthermore, the anonymous exchange of information between businesses within the ERP would be an interesting case study to investigate, allowing the creation of new processes or the modification of existing ones.

Theoretically, the values present in variable and supplier classification surveys are the closest to reality; however, in some cases, theory does not accurately represent reality (even when efforts have been made to ensure model adaptation in accordance with business priorities). A process of consulting with DM professionals would be a good starting point for comparing the described processes to current reality.

Regarding ML models, an exploration of advanced model evaluation techniques would be essential; an example of starting this exploration would be tools that use the cross-entropy loss to explain the predictions of probabilistic classifiers.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. url: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on pp. iii, xv).

[2] Council of Supply Chain Management Professionals (CSCMP). *SCM Definitions and Glossary of Terms*. 2013. url: https://cscmp.org/CSCMP/Educate/SCM_Definitions_and_Glossary_of_Terms.aspx (cit. on p. xv).

[3] P. BSS. *PRIMAVERA*. url: https://pt.primaverabss.com/en/primavera/ (cit. on pp. 1, 3).

[4] Directorate-General For Research And Innovation. *Industry 5.0*. url: https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en (cit. on p. 1).

[5] Compete2020. *Newsletter - PRODUTECH4S&C*. url: https://www.compete2020.gov.pt/newsletter/detalhe/46102-PRODUTECH4SC-Entrevista-NunoAraujo-NL26032021 (cit. on p. 3).

[6] *PRODUTECH 4 S&C — Português*. url: http://mobilizadores.produtech.org/pt/produtech_4_s_c (cit. on p. 3).

[7] J. Thangaraj and R. Lakshmi Narayanan. "INDUSTRY 1.0 TO 4.0: THE EVOLUTION OF SMART FACTORIES". In: (Oct. 2018). url: https://www.researchgate.net/publication/330336790_INDUSTRY_10_TO_40_THE_EVOLUTION_OF_SMART_FACTORIES (cit. on p. 5).

[8] D.-G. for Research {and} Innovation (European Commission) et al. *Industry 5.0, a transformative vision for Europe: governing systemic transformations towards a sustainable industry*. Publications Office of the European Union, 2022. isbn: 978-92-76-43352-1. doi: doi/10.2777/17322. url: https://data.europa.eu/doi/10.2777/17322 (cit. on p. 5).

[9] K. A. Demir, G. Döven, and B. Sezen. “Industry 5.0 and Human-Robot Co-working”. In: *Procedia Computer Science* 158 (Jan. 2019). issn: 18770509. doi: 10.1016/j.procs.2019.09.104. url: https://linkinghub.elsevier.com/retrieve/pii/S1877050919312748 (cit. on p. 5).

[10] D.-G. for Research {and} Innovation (European Commission) et al. *Industry 5.0 : towards a sustainable, human-centric and resilient European industry*. Publications Office of the European Union, 2021. isbn: 978-92-76-25308-2. doi: doi/10.2777/308407. url: https://data.europa.eu/doi/10.2777/308407 (cit. on p. 6).

[11] D. Stanton. *Supply Chain Management For Dummies*. John Wiley & Sons, 2017. isbn: 978-1-119-67701-7. url: https://www.oreilly.com/library/view/supply-chain-management/9781119410195/ (cit. on pp. 7–9).

[12] D. M. Lambert, J. R. Stock, and L. M. Ellram. *Fundamentals of Logistics Management*. Irwin/McGraw-Hill, 1998. isbn: 978-0-07-115752-0. url: https://books.google.pt/books?id=K8xXAAAAYAAJ (cit. on p. 7).

[13] S. Chopra and P. Meindl. *Supply Chain Management: Strategy, Planning, and Operation, eBook, Global Edition*. Pearson Education, 2015. isbn: 9781292093574 (cit. on p. 7).

[14] Bernard J. ”Bud” Lalonde. “Supply Chain Management: Myth or Reality?” In: Supply Chain Management Review (Mar. 21, 1997). issn: 1521-9747. url: https://www.scmr.com/article/supply_chain_management_myth_or_reality (cit. on p. 7).

[15] M. Christopher. *Logistics & Supply Chain Management*. second. Pearson UK, 2003. isbn: 978-1-292-08382-7. url: https://www.oreilly.com/library/view/supply-chain-management/9781119410195/ (cit. on p. 7).

[16] Material Handling & Logistics. *Council of Logistics Management to become Council of Supply Chain Management Professionals*. 2004. url: https://www.mhlnews.com/global-supply-chain/article/22040540/council-of-logistics-management-to-become-council-of-supply-chain-management-professionals (cit. on p. 7).

[17] Council of Supply Chain Management Professionals (CSCMP). *Supply Chain Management Concepts*. url: https://cscmp.org/CSCMP/Develop/Starting_Your_Career/Supply_Chain_Management_Concepts.aspx#Seven%20Principles%20of%20SCM (cit. on p. 7).

[18] D. Anderson, F. Britt, and D. Favre. “The Seven Principles of Supply Chain Management”. In: *Supply Chain Manage Rev Spring* (Jan. 1997). url: https://www.semanticscholar.org/paper/The-Seven-Principles-of-Supply-Chain-Management-Anderson-Britt/7bd16434f5253650e3dce2a2a1975a6faeff7eb8 (cit. on p. 8).

[19] *The Bullwhip Effect*. Sketchplanations. url: https://sketchplanations.vercel.app/the-bullwhip-effect (cit. on p. 9).

[20] B. Roy. “Paradigms and Challenges”. In: *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, 2005, pp. 3–24. isbn: 978-0-387-23081-8. doi: 10.1007/0-387-23081-5_1. url: https://doi.org/10.1007/0-387-23081-5_1 (cit. on p. 10).

[21] S. H. Ha and R. Krishnan. “A hybrid approach to supplier selection for the maintenance of a competitive supply chain”. In: *Expert Systems with Applications* 34.2 (Feb. 1, 2008), pp. 1303–1311. issn: 0957-4174. doi: https://doi.org/10.1016/j.eswa.2006.12.008. url: https://www.sciencedirect.com/science/article/pii/S0957417406004180 (cit. on pp. 10, 16).

[22] J. W▯tróbski et al. “Generalised framework for multi-criteria method selection”. In: *Omega* 86 (2019), pp. 107–124. issn: 0305-0483. doi: https://doi.org/10.1016/j.omega.2018.07.004. url: https://www.sciencedirect.com/science/article/pii/S0305048317308563 (cit. on p. 11).

[23] *MCDA Method Selection Tool*. url: https://mcda.it/ (cit. on p. 11).

[24] C. Wiley. “Recruitment Research Revisited: Effective Recruiting Methods According To Employment Outcomes”. In: *Journal of Applied Business Research (JABR)* 8 (Oct. 18, 2011), p. 74. doi: 10.19030/jabr.v8i2.6167. url: https://www.researchgate.net/publication/295882093_Recruitment_Research_Revisited_Effective_Recruiting_Methods_According_To_Employment_Outcomes (cit. on p. 13).

[25] M. Modarres. “Fuzzy Simple Additive Weighting Method by Preference Ratio”. In: *Intelligent Automation & Soft Computing* 11 (2005). doi: 10.1080/10642907.2005.10642907. url: https://www.researchgate.net/publication/240748637_Fuzzy_Simple_Additive_Weighting_Method_by_Preference_Ratio (cit. on p. 14).

[26] S. Bag. “Fuzzy VIKOR approach for selection of big data analyst in procurement management”. In: *Journal of Transport and Supply Chain Management* 10 (Jan. 29, 2016). doi: 10.4102/jtscm.v10i1.230. url: https://www.researchgate.net/publication/305716205_Fuzzy_VIKOR_approach_for_selection_of_big_data_analyst_in_procurement_management (cit. on p. 14).

[27] K. Yoon and C. L. Hwang. *Multiple attribute decision making: an introduction*. Sage Publications, 1995. isbn: 978-0-8039-5486-1. url: https://www.researchgate.net/publication/258023856_Multiple_Attribute_Decision_Making_An_Introduction_Quantitative_Applications_in_the_Social_Sciences (cit. on p. 14).

[28] N. Sorin, S. Dzitac, and I. Dzitac. "Fuzzy TOPSIS: A General View". In: *Procedia Computer Science* 91 (Jan. 1, 2016), pp. 823–831. issn: 1877-0509. doi: 10.1016/j.procs.2016.07.088. url: https://www.sciencedirect.com/science/article/pii/S187705091631273X (cit. on p. 14).

[29] G. W. Dickson. "An Analysis Of Vendor Selection Systems And Decisions". In: *Journal of Purchasing* 2.1 (1966), pp. 5–17. issn: 1745-493X. doi: https://doi.org/10.1111/j.1745-493X.1966.tb00818.x. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-493X.1966.tb00818.x (cit. on pp. 15, 16).

[30] C. A. Weber, J. R. Current, and W. Benton. "Vendor selection criteria and methods". In: *European Journal of Operational Research* 50.1 (1991), pp. 2–18. issn: 0377-2217. doi: https://doi.org/10.1016/0377-2217(91)90033-R. url: https://www.sciencedirect.com/science/article/pii/037722179190033R (cit. on p. 16).

[31] C. Schröer, F. Kruse, and J. M. Gómez. "A Systematic Literature Review on Applying CRISP-DM Process Model". In: *Procedia Computer Science* 181 (2021), pp. 526–534. issn: 18770509. doi: 10.1016/j.procs.2021.01.199. url: https://linkinghub.elsevier.com/retrieve/pii/S1877050921002416 (cit. on p. 17).

[32] R. Wirth and J. Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *Procedia Computer Science* 181 (2021), pp. 526–534. issn: 18770509. doi: 10.1016/j.procs.2021.01.199 (cit. on pp. 17, 18).

[33] T. Taulli. *Artificial intelligence basics: a non-technical introduction*. New York: Apress, 2019. isbn: 978-1-4842-5027-3. doi: 10.1007/978-1-4842-5028-0 (cit. on p. 18).

[34] E. Mjolsness and D. Decoste. *Machine Learning for Science: State of the Art and Future Prospects*. Vol. 293. Oct. 2001, pp. 2051–5. doi: 10.1126/science.293.5537.2051. url: https://www.researchgate.net/publication/11789794_Machine_Learning_for_Science_State_of_the_Art_and_Future_Prospects (cit. on p. 18).

[35] I. H. Sarker. "Machine Learning: Algorithms, Real-World Applications and Research Directions". In: *SN Computer Science* 2 (Mar. 2021), p. 160. issn: 2661-8907. doi: 10.1007/s42979-021-00592-x (cit. on p. 19).

[36] *API reference - module-sklearn.metrics*. url: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics (cit. on p. 20).

[37] G. Hohpe and B. Woolf. *Enterprise Integration Patterns : Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley Professional, 2003. isbn: 0321200683. url: https://learning.oreilly.com/library/view/enterprise-integration-patterns/0321200683/ (cit. on p. 20).

[38] en. url: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression.predict_proba (cit. on p. 20).

[39] W. Wang and Y. Lu. "Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model". In: *IOP Conference Series: Materials Science and Engineering* 324 (Mar. 2018), p. 012049. issn: 1757-8981,1757-899X. doi: 10.1088/1757-899X/324/1/012049 (cit. on p. 20).

[40] T. Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (June 1, 2006), pp. 861–874. issn: 0167-8655. doi: https://doi.org/10.1016/j.patrec.2005.10.010 (cit. on p. 21).

[41] J. Xu, Y. Zhang, and D. Miao. "Three-way confusion matrix for classification: A measure driven view". In: *Information Sciences* 507 (2020), pp. 772–794. issn: 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.06.064 (cit. on p. 21).

[42] A. Jung. *Machine Learning: The Basics*. Springer Nature, 2022. isbn: 978-981-16-8192-9. doi: 10.1007/978-981-16-8193-6. url: https://link.springer.com/10.1007/978-981-16-8193-6 (cit. on p. 21).

[43] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. Adaptive computation and machine learning. The MIT Press, 2016. 775 pp. isbn: 978-0-262-03561-3. url: https://doi.org/10.1007/s10710-017-9314-z (cit. on p. 22).

[44] J. Fürnkranz. "Decision Tree". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Springer US, 2010. isbn: 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_204. url: https://doi.org/10.1007/978-0-387-30164-8_204 (cit. on p. 23).

[45] C. Sammut and G. I. Webb. "Random Forests". In: *Encyclopedia of Machine Learning and Data Mining*. Springer US, 2017, pp. 1054–1054. isbn: 978-1-4899-7687-1. doi: 10.1007/978-1-4899-7687-1_695. url: https://doi.org/10.1007/978-1-4899-7687-1_695 (cit. on p. 24).

[46] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System". In: ACM, 2016, pp. 785–794. isbn: 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. url: https://xgboost.readthedocs.io/en/latest/tutorials/model.html (cit. on p. 24).

[47] June 2016. url: https://www.youtube.com/watch?v=Vly8xGnNiWs&feature=emb_title (cit. on p. 24).

[48] G. Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems 30 (NIP 2017)*. Dec. 2017. url: https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/ (cit. on p. 25).

[49] A. A. T. Fernandes et al. “Read this paper if you want to learn logistic regression”. In: *Revista de Sociologia e Política* 28 (2020), p. 6. issn: 1678-9873, 0104-4478. doi: 10.1590/1678-987320287406en. url: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-44782020000200207&tlng=en (cit. on p. 25).

[50] X. Fang et al. “Regression Analysis With Differential Privacy Preserving”. In: *IEEE Access* 7 (2019), pp. 129353–129361. issn: 2169-3536. doi: 10.1109/ACCESS.2019.2940714. url: https://ieeexplore.ieee.org/document/8835087/ (cit. on pp. 25, 26).

[51] W. McKinney. *Python for Data Analysis*. Third edition. O'Reilly Media, Inc., 2022. isbn: 978-1-09-810403-0. url: https://learning.oreilly.com/library/view/python-for-data/9781098104023/ (cit. on p. 27).

[52] D. Y. Chen. *Pandas for Everyone: Python Data Analysis*. 1nd. Addison-Wesley Professional, Dec. 30, 2017. isbn: 978-0-13-454704-6. url: https://www.oreilly.com/library/view/pandas-for-everyone/9780134547046/ (cit. on p. 27).

[53] YData. *Pandas Profiling*. url: https://pandas-profiling.ydata.ai/docs/master/index.html (cit. on p. 27).

[54] B. Horvath. “Synthetic Data for Deep Learning”. In: *Quantitative Finance* 22 (2022), pp. 423–425. doi: 10.1080/14697688.2022.2048062. url: https://doi.org/10.1080/14697688.2022.2048062 (cit. on p. 28).

[55] joke2k. *Faker*. Version 15.1.1. url: https://github.com/joke2k/faker (cit. on p. 28).

[56] Databricks. *Databricks Labs Data Generator*. url: https://databrickslabs.github.io/dbldatagen/public_docs/index.html (cit. on p. 28).

[57] J. S. Damji et al. *Learning Spark*. 2. edition. O'Reilly Media, Inc., 2020. isbn: 978-1-4920-5004-9. url: https://learning.oreilly.com/library/view/learning-spark-2nd/9781492050032/ (cit. on p. 28).

[58] Microsoft. *What is Azure Databricks?* url: https://learn.microsoft.com/en-us/azure/databricks/introduction/ (cit. on p. 29).

[59] L. Zimmer, M. Lindauer, and F. Hutter. “Auto-PyTorch Tabular: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 3079–3090. doi: https://doi.org/10.48550/arXiv.2006.13799. url: https://github.com/automl/Auto-PyTorch (cit. on p. 29).

[60] M. Feurer et al. “Auto-Sklearn”. In: (2020). doi: https://doi.org/10.48550/arXiv.2007.04074. url: https://github.com/automl/auto-sklearn (cit. on p. 29).

[61] R. S. Olson et al. “Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science”. In: ACM, 2016, pp. 485–492. isbn: 978-1-4503-4206-3. doi: 10.1145/2908812.2908918. url: https://epistasislab.github.io/tpot/ (cit. on pp. 29, 30).

[62] E. LeDell. “H2O AutoML: Scalable Automatic Machine Learning”. In: 2020. url: https://www.semanticscholar.org/paper/H2O-AutoML%3A-Scalable-Automatic-Machine-Learning-LeDell/22cba8f244258e0bba7ff4bb70c4e5b5ac3e2382 (cit. on p. 29).

[63] H. Parul Pandey. *A Deep Dive into H2O's AutoML*. url: https://h2o.ai/blog/a-deep-dive-into-h2os-automl/ (cit. on p. 31).

[64] *Databricks AutoML - Automated Machine Learning*. Databricks. url: https://www.databricks.com/product/automl (cit. on p. 31).

[65] S. Lundberg and S.-I. Lee. “A Unified Approach to Interpreting Model Predictions”. In: (2017). doi: 10.48550/ARXIV.1705.07874. url: https://arxiv.org/abs/1705.07874 (cit. on p. 54).

[66] D. Gunning et al. “XAI—Explainable artificial intelligence”. In: *Science Robotics* 4 (Dec. 18, 2019). issn: 2470-9476. doi: 10.1126/scirobotics.aay7120. url: https://www.science.org/doi/10.1126/scirobotics.aay7120 (cit. on p. 55).

I

Annex 1

Universidade do Minho
Escola de Engenharia

Pedido de Admissão à Dissertação/Projeto

Plano de Trabalho

Ano Letivo: **2021/2022**	**Nome: Pedro Miguel Ferreira Ribeiro** **Número: pg42848** **Título** (em Português)**: Machine Learning Aplicado à Gestão de Empresas** **Título** (em Inglês)**: Machine Learning Applied to Companies Management**

Enquadramento e Motivação (150 - 200 palavras)

A PRIMAVERA tem desenvolvido vários projetos ao abrigo da iniciativa “ERP Inteligente” que visa a construção de soluções progressivamente mais autónomas, que permitam aos seus clientes focarem-se essencialmente na estratégia das suas empresas, dedicando menos energia à gestão operacional do negócio.

A evolução tecnológica da Inteligência Artificial (AI) pode ser explorada no sentido de tornar as nossas soluções mais inteligentes, analisando dados de histórico para antecipar necessidades e prever o futuro. Por exemplo, no horizonte temporal do médio prazo é necessário identificar as necessidades das encomendas de clientes, tirando o máximo partido dos recursos instalados, para prever as necessidades de materiais e dar início às ordens de fabrico, otimizando assim os processos de compra e stocks. Tudo isto se resumo ao planeamento concorrente e colaborativo da cadeia de abastecimento em que, para melhorar a produtividade, as organizações adotam sistemas de Material Requirements Planning (MRP), ou mais recentemente de Manufacturing Resource Planning (MRP II), que envolvem a gestão dos materiais, dos processos de produção, dos recursos financeiros e dos recursos de produção (máquinas e pessoas).

Com o desenvolvimento industrial e tecnológico mais recente, sobretudo das plataformas de AI e dos algoritmos de Machine Learning (ML), surge agora um novo impulso para a modernização destas áreas mais tradicionais da gestão das organizações.

Objetivos e Resultados Esperados (150 - 200 palavras)

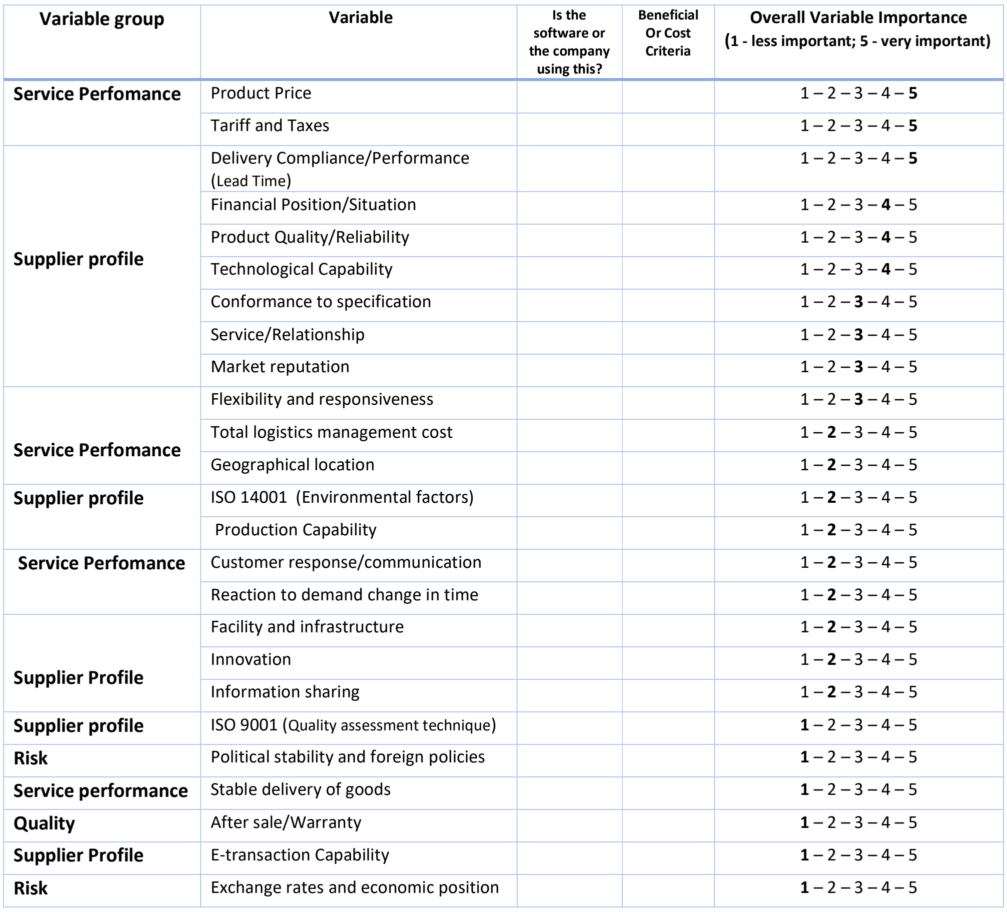
O objetivo deste projeto é explorar tecnologias de AI e modelos de ML para o planeamento concorrente e colaborativo da cadeia de abastecimento.
O resultado do projeto deve incluir uma proposta de modelos válidos.

Calendarização

- [2 meses] Estado da arte e casos de estudo;
- [2 meses] Análise do caso de estudo e planeamento de estratégias para resolução do problema;
- [1 meses] Análise da Fábrica de Dados da Primavera;
- [3 meses] Desenvolvimento de modelos e funcionalidades para otimização de cadeias de abastecimento;
- [1 meses] Revisão, melhoria e descrição escrita do trabalho desenvolvido.

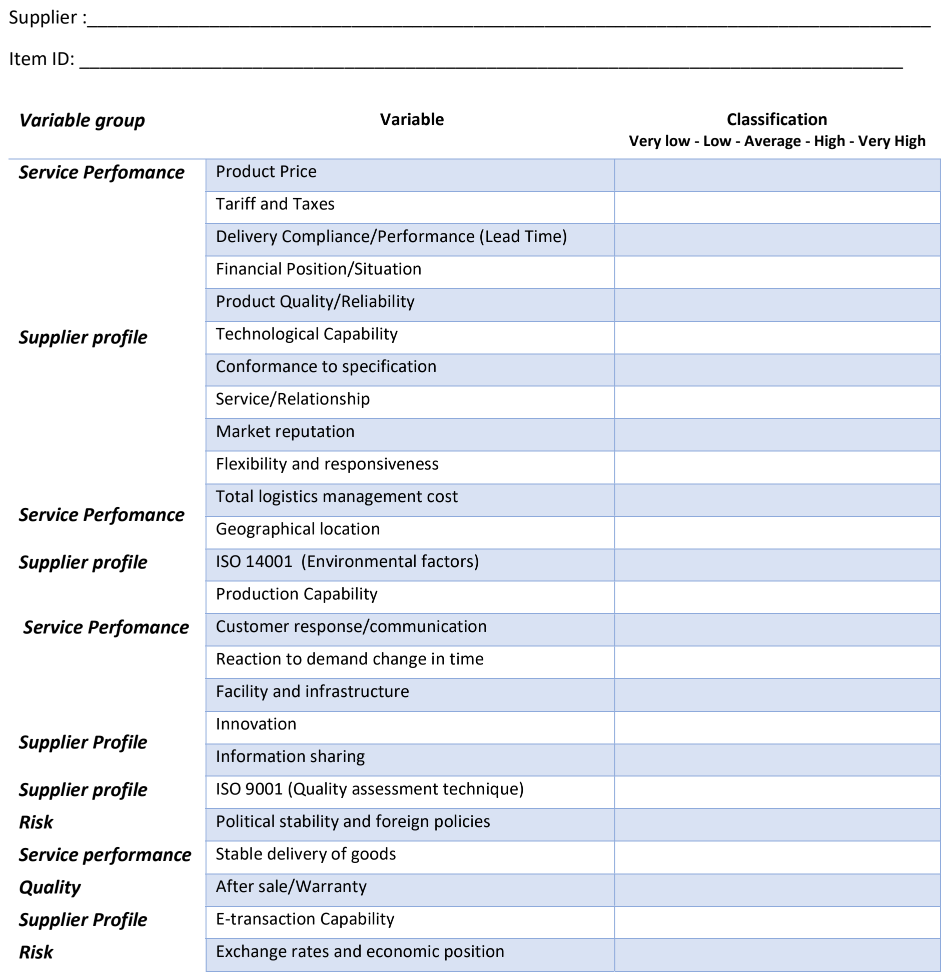
Referências Bibliográficas (5 - 10 referências)

- Supply Chain Management Concepts. (n.d.). From https://cscmp.org/CSCMP/Develop/Starting_Your_Career/Supply_Chain_Management_Concepts.aspx#Seven Principles of SCM
- Industry 5.0 | European Commission. (n.d.). From https://ec.europa.eu/info/research-and-innovation/research-area/industrial-research-and-innovation/industry-50_en
- Industry 5.0: Towards more sustainable, resilient and human-centric industry | European Commission. (n.d.). From https://ec.europa.eu/info/news/industry-50-towards-more-sustainable-resilient-and-human-centric-industry-2021-jan-07_en
- Databricks - The Data and AI Company. (n.d.). From https://databricks.com/
- X. V. Pham, A. Maag, S. Senthilananthan and M. Bhuiyan, "Predictive analysis of the supply chain management using Machine learning approaches: Review and Taxonomy," 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), 2020, pp. 1-9, doi: 10.1109/CITISIA50690.2020.9371842.
- R. Harikrishnakumar, A. Dand, S. Nannapaneni and K. Krishnan, "Supervised Machine Learning Approach for Effective Supplier Classification," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 2019, pp. 240-245, doi: 10.1109/ICMLA.2019.00045.



Variable group	Variable	Is the software or the company using this?	Beneficial Or Cost Criteria	Overall Variable Importance (1 - less important; 5 - very important)
Service Perfomance	Product Price			1 – 2 – 3 – 4 – **5**
	Tariff and Taxes			1 – 2 – 3 – 4 – **5**
Supplier profile	Delivery Compliance/Performance (Lead Time)			1 – 2 – 3 – 4 – **5**
	Financial Position/Situation			1 – 2 – 3 – **4** – 5
	Product Quality/Reliability			1 – 2 – 3 – **4** – 5
	Technological Capability			1 – 2 – 3 – **4** – 5
	Conformance to specification			1 – 2 – **3** – 4 – 5
	Service/Relationship			1 – 2 – **3** – 4 – 5
	Market reputation			1 – 2 – **3** – 4 – 5
Service Perfomance	Flexibility and responsiveness			1 – 2 – **3** – 4 – 5
	Total logistics management cost			1 – **2** – 3 – 4 – 5
	Geographical location			1 – **2** – 3 – 4 – 5
Supplier profile	ISO 14001 (Environmental factors)			1 – **2** – 3 – 4 – 5
	Production Capability			1 – **2** – 3 – 4 – 5
Service Perfomance	Customer response/communication			1 – **2** – 3 – 4 – 5
	Reaction to demand change in time			1 – **2** – 3 – 4 – 5
Supplier Profile	Facility and infrastructure			1 – **2** – 3 – 4 – 5
	Innovation			1 – **2** – 3 – 4 – 5
	Information sharing			1 – **2** – 3 – 4 – 5
Supplier profile	ISO 9001 (Quality assessment technique)			**1** – 2 – 3 – 4 – 5
Risk	Political stability and foreign policies			**1** – 2 – 3 – 4 – 5
Service performance	Stable delivery of goods			**1** – 2 – 3 – 4 – 5
Quality	After sale/Warranty			**1** – 2 – 3 – 4 – 5
Supplier Profile	E-transaction Capability			**1** – 2 – 3 – 4 – 5
Risk	Exchange rates and economic position			**1** – 2 – 3 – 4 – 5

Figure 27: Variable classification survey.



Supplier :____

Item ID: ____

Variable group	**Variable**	**Classification** **Very low - Low - Average - High - Very High**
Service Perfomance	Product Price	
	Tariff and Taxes	
	Delivery Compliance/Performance (Lead Time)	
	Financial Position/Situation	
	Product Quality/Reliability	
Supplier profile	Technological Capability	
	Conformance to specification	
	Service/Relationship	
	Market reputation	
	Flexibility and responsiveness	
Service Perfomance	Total logistics management cost	
	Geographical location	
Supplier profile	ISO 14001 (Environmental factors)	
	Production Capability	
Service Perfomance	Customer response/communication	
	Reaction to demand change in time	
	Facility and infrastructure	
Supplier Profile	Innovation	
	Information sharing	
Supplier profile	ISO 9001 (Quality assessment technique)	
Risk	Political stability and foreign policies	
Service performance	Stable delivery of goods	
Quality	After sale/Warranty	
Supplier Profile	E-transaction Capability	
Risk	Exchange rates and economic position	

Figure 28: Supplier classification survey.