TÉCNICO LISBOA

Machine Learning for Predictive Maintenance of Industrial Equipment

Miguel Clemente Correia

Thesis to obtain the Master of Science Degree in

Chemical Engineering

Supervisors: Prof. Ana Paula Vieira Soares Pereira Dias

Prof. João Miguel da Costa Sousa

Examination Committee

Chairperson: Prof. Henrique Anibal Santos de Matos

Supervisors: Prof. João Miguel da Costa Sousa

Members of Committee: Eng. Luís Miguel Mendes Gomes

November 2018

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the *Universidade de Lisboa*.

“The secret of happiness is to see all the marvels of the world, and never to forget the drops of oil on the spoon”

Paulo Coelho – *The Alchemist*

To Inês,

my partner in life.

Acknowledgements

I would like to express my deepest appreciation to all those who provide me with knowledge and experience to perform the present thesis.

An especial gratitude should be given to the manufacturer personnel for incubating the study and granting the opportunity to evolve one of my greatest areas of interest, displaying responsiveness at all times during the development of this joint work.

I am particularly grateful to Prof. Ana Paula Dias and Prof. João Sousa for guiding this work with their enthusiastic encouragement, keeping my progress on schedule.

To Eng. Luís Gomes for the constant availability and valuable insights, whose experience granted the construction of notable perspectives throughout this thesis.

I would also like to extend my thanks to Joaquim Viegas and Prof. Susana Vieira as their insights were of crucial importance.

Finally, last but by no means least, I want to express my deepest gratitude to my family for their unceasing encouragement, support and attention giving me the possibilities to complete my studies. I owe them everything.

Abstract

The manufacturing industry is embarking on a progressive level of maturity concerning digitalization. Maintenance is a fundamental segment where competitive advantage can be acquired, through the implementation of quality and data-driven techniques. In manufacturing and especially in the present use case, quality standards are of major importance as they configure a level of difficulty to the global production process. Manufacturers have available a more advanced technological environment, looking at data as an insight generation tool. Currently, the machines are not prepared to offer an integrated perspective of the system, thus being challenging the identification of the root causes for the majority of the failures. The present thesis reports the work developed to address the requirements stated above. Equipment data acquires in this work, a unique interest while it is manipulated to develop an intelligent predictive model adapted to the industrial context. With this aim, it is proposed a novel method for integrating equipment data suppressing the need for a large number of variables and facing uncertainties in the system data flow. In addition, a framework for visualizing patterns and equipment interdependency is developed. The current study provides the manufacturer and the respective maintenance team an effective tool that can be integrated into the existing system providing insights and increasing the conformity with quality standards and production goals.

Keywords

Data-driven analysis, Machine learning, Failure prediction, MTBF, Manufacturing Industry.

Resumo

A indústria de fabrico está a embarcar num nível avançado de maturidade no que diz respeito à sua digitalização. A Manutenção é um dos segmentos fundamentais para a aquisição de vantagem competitiva, através da implementação de técnicas orientadas pela qualidade e pela informação. Neste tipo de indústria, e especialmente no presente caso de estudo, os padrões de qualidade são parte importante e conferem um incremento no nível de dificuldade do processo global de produção. Os fabricantes têm à sua disposição um ambiente tecnológico mais avançado, considerando os dados como uma fonte de criação de conhecimento. No presente, as máquinas não estão preparadas para oferecer uma visão integrada do sistema, tornando-se desafiador a identificação dos motivos base que levam à maioria das sua falhas. A presente dissertação reporta o trabalho desenvolvido para endereçar os requisitos acima mencionados. Os dados de equipamentos adquirem neste trabalho um interesse único pela sua utilização no desenvolvimento de um modelo preditivo inteligente adaptado ao ambiente industrial. Com este objetivo, é proposto um método original para a integração dos dados de equipamentos, suprimindo a necessidade por um grande número de variáveis e enfrentando incertezas no fluxo de dados do sistema. Em adição, foi desenvolvida uma estrutura de visualização de padrões e interdependência entre equipamentos. Este estudo disponibiliza ao fabricante e à equipa de manutenção uma ferramenta eficaz que pode ser integrada nos sistemas existentes, proporcionando conhecimento e ampliando a conformidade com os padrões de qualidade e os objetivos de produção.

Palavras-chave

Análise orientada por dados, *Machine learning*, Previsão de falhas, TMEF, Indústria de manufatura.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AP	Average Precision is a weighted mean of the precisions obtained at each threshold when evaluating a machine learning model through the Precision – Recall Curve.
AUC	Area Under the Curve is related to both Precision – Recall Curve as well as Receiver Operating Characteristic Curve. It provides an indication on how the model performed and is directly related to the accuracy.
ECS	Equipment Causing Stop is a field in the machine production data, which indicates the equipment responsible for the respective stop. It's left blank when the machine status is *Running*.
ET	Extremely Randomized Trees is inserted in the category of ensemble methods combining several base estimators to one learning algorithm with improved generalizability and robustness.
FDR	False Discovery Rate provides the proportion of discoveries that are in fact false.
FN	False Negative is an outcome of the machine learning model where it incorrectly predicts the negative class.
FP	False Positive is an outcome of the machine learning model where it incorrectly predicts the positive class.
FPR	False Positive Rate is the rate of positive outcomes that are in fact negative.
MTBF	Mean Time Between Failures gives and estimation on the elapsed time between two consecutive failures in a defined system.
PR Curve	Precision – Recall Curve displays a visual method for the balance between precision and recall for different thresholds.
ROC Curve	Receiver Operating Characteristic Curve displays a visual method for the balance between true positive rate and false positive rate for different thresholds.
TN	True Negative is an outcome of the machine learning model where it correctly predicts the negative class.
TP	True Positive is an outcome of the machine learning model where it correctly predicts the positive class.
TPR	True Positive Rate is the rate of positive outcomes that are in fact positive.

Chapter 1

Introduction

Previous to laying out the motivation and problem description, this chapter gives an overview of the current State of the Art concerning different approaches which can be adapted to the problem, bringing out the innovative aspects presented on the thesis. At the end of the chapter, the structure of the main report is also outlined.

1.1 Motivation

The industrial environment is changing and digital technology is the leading actor in what is considered as the emerging paradigm - Industry 4.0. Attached to this concept is the idealization of smart manufacturing which is supported by the integration of “smart technologies” with standard manufacturing devices as sensors and other equipment. One may affirm that the manufacturing is by itself adjusting to human needs and also to the constraints of its supply chain. This fourth revolution can be materialized by attending to cyber-physical systems – a merge between the physical and the digital grades. An example of this systems can be found in the preventive maintenance area, where the condition of a physical equipment and all the associated parameters are reflected in a Digital Twin [1].

A Digital Twin can be seen as a mirror of a physical equipment being composed by the combination of the equipment itself, the virtual components and the cyber-physical data that connects the two grades. The introduction of Digital Twin has occurred initially in aeronautics and astronautics fields for failure prediction, being present within the maintenance context. This concept has attached a large amount of data from production status to operational information and the manipulation of this information provides a wiser equipment maintenance as well as improved process design [2].

The combination of artificial intelligence, big data, streaming analytics and machine learning, provides a powerful tool for the manufacturing industry environment and supports the basic concepts embraced in the current thesis.

The maturity of technology dictates the short number of documents regarding the application of predictive maintenance to the industrial context, the contribution of this research holds the following alignment of achievements:

- A tool for combining asymmetric data sets;
- An apparatus for visualizing patterns and interdependency of machine behaviour;
- An approach to overcome unknown variables;
- A data-driven machine learning classifier adapted to the manufacturing industry.

1.2 Topic Overview

The diversification of challenges faced by industries is leading maintenance job to constantly grow to a more mature state, taking advantage of technological evolution to redefine the strategies followed by maintenance teams.

Corrective maintenance is far known as the “only fault repair” approach and it is on the basis of the maintenance ideology. Despite being the earliest maintenance mode, this strategy it’s deeply common in industries with reduced complexity and conservative culture. Taking into account the majority of industries, preventive maintenance is the endorsed method to address maintenance needs, performing a regular inspection at well-defined periods of time in order to prevent deterioration of equipment and possible loss of product quality generated by faulty components [3].

Bearing in mind a lack of data regarding the system operation and the absence of intelligence with the capability to process this information, it’s acceptable to adopt the previous strategies. However, several companies are realising the true cost of maintenance and the potential they can achieve with a more adapted strategy. For this reason, predictive maintenance is the main hypothesis on this thesis, representing the concept that allows high reliability and enhances economic efficiency [4].

Predictive maintenance itself is not a replacement for the two methods previous presented, it's essential to have a combination between "run to failure" approach and preventive actions. In this blend of strategies, predictive procedures can complement the results by reducing the number of unexpected failures and identifying periods to apply specific maintenance tasks [5]. This late concept is usually adopted as a tool for maintenance management aiming to prevent unscheduled downtime, as a plant optimization tool helping the definition of production procedures and parameters and also as a reliability improvement tool identifying the deviations in the operative specifications triggering operator actions preventing the potential failure or loss of quality in the production [6].

To better understand maintenance methodologies, it's important to have a clear knowledge of the behavioural progression expected in an equipment lifetime. Figure 1 represents a conceptual evolution of failure rate minding the equipment usage, displaying the probability of failure in the three stages of the system. The described interpretation is on the basis of maintenance tasks definition and it's widely used to plan both preventive and predictive procedures.

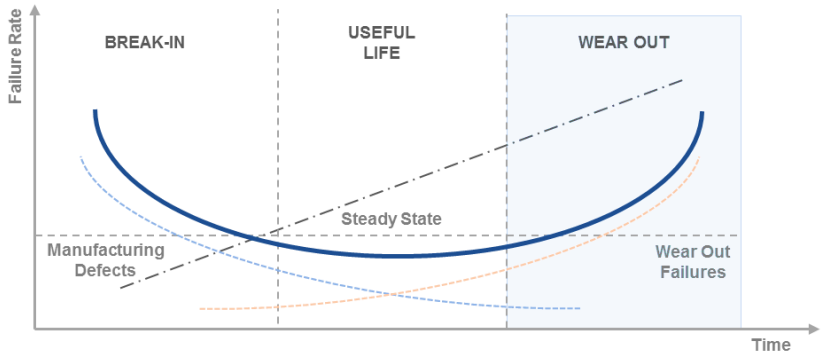

Figure 1 | "Bathtub Curve" - Hypothetical Asset Failure Rate versus Time [7].

Industries can address the previously stated challenges by implementing several monitoring technologies like thermography, vibration monitoring, tribology and ultrasonic analysis or by conducting failure analysis through parameter and historical data processing. A technology-based approach often combines more than one monitoring mechanism, increasing complexity in data processing due to the integration of information from different sources [6].

On the other hand, the technic referred in second place can efficiently adapt to the existing systems, using a principle settled on the equipment behaviour defined as Condition-Based Maintenance. CBM takes into account the progression on the equipment status and deviations from standard parameters to identify potential problems and unexpected failures. This can be a powerful tool when used along with historical failure data, assuming a relation between the conditions that cause an event in the past and the replication of the same conditions in the present. Correlations between equipment usage and component deterioration are well-known examples of potential failure motivators that are feasible to apply CBM theory [8].

Usually, both intrusive and non-intrusive methods are used to complement each other, achieving superior results by combining them with maintenance and operational data to build a device capable of estimate equipment lifetime and plan maintenance unexpected activities [9].

Predictive maintenance models are typically developed by considering machine learning techniques as data mining and machine learning classification. This last task can be divided into two groups: Supervised Learning and Unsupervised Learning. In the first type of problems, it is present an outcome variable to guide the learning process while in unsupervised learning only the features are observed having no measurements of the output variables. In the maintenance environment, the aim of supervised learning is usually to predict a value or a label of one variable such us time to failure or the end of life of a component, and it can be achieved by knowing the values of other variables (production, sensors data and other information). If the variables being predicted are not a value but a category, then the problem is described as classification [10].

Bearing in mind the previous statements, it's essential to understand some of the classifiers that are often used in similar problems. Ensemble methods were used in the current thesis to develop a machine learning model for classification. These methods are learning algorithms that, from a set of classifiers, perform a weighted vote on their classification and then estimate a class to the new entries. In a simple manner, the ensemble classification combines the prediction of several estimators in a given algorithm, thus improving the robustness of a single classifier [11].

Contained in the category of ensemble methods are the forests of randomized trees. In the case of this methodology, a set of classifiers (decision trees) is created by introducing randomness in the classifier construction as detailed in the following illustration.

$$D = \begin{bmatrix} X_{A1} & X_{B1} & \cdots & y_1 \\ \vdots & \vdots & \vdots & \vdots \\ X_{AN} & X_{BN} & \cdots & y_N \end{bmatrix}$$

$$T_1 = \begin{bmatrix} X_{A12} & X_{B12} & \cdots & y_{12} \\ X_{A34} & X_{B34} & \vdots & y_{34} \\ \vdots & \vdots & \vdots & \vdots \\ X_{A72} & X_{B72} & \cdots & y_{72} \end{bmatrix} \quad T_2 = \begin{bmatrix} X_{A24} & X_{B24} & \cdots & y_{24} \\ X_{A38} & X_{B38} & \vdots & y_{38} \\ \vdots & \vdots & \vdots & \vdots \\ X_{A47} & X_{B47} & \cdots & y_{47} \end{bmatrix} \quad T_3 = \begin{bmatrix} X_{A15} & X_{B15} & \cdots & y_{15} \\ X_{A54} & X_{B54} & \vdots & y_{54} \\ \vdots & \vdots & \vdots & \vdots \\ X_{A62} & X_{B62} & \cdots & y_{62} \end{bmatrix} \quad \cdots$$

Figure 2 | Representation of forests of randomized trees in machine learning classification.

In Figure 2, D is the learning data set fed to the algorithm, while T_1, T_2, T_3 are the randomly generated decision trees. X and y are, respectively, the features and the associated output class. In its learning stage, the algorithm adjusts its parameters to learn from the learning data set to classify the output class y.

On this thesis, predictive maintenance converges with condition base monitoring, applying an intelligent system which can identify patterns and predict potential failures based on progression of the equipment status, operating conditions and maintenance team experience. This approach can be adapted to systems without a substantial volume of information and it's also qualified for the initial stage of equipment's lifetime in which the number of failures is higher, usually characterized by the difficulty in identifying potential problems in the quality control inspections.

Therefore, using technics that are non-destructive and adaptable to the installed system, it's possible to take actions in order to bypass the potential failure of an equipment improving one of the most important performance indicators, the Overall Equipment Efficiency [12].

1.3 Thesis Outline

Previous to laying out a detailed description of the system in section 2.1, it is contextualized the production process addressed in this thesis. Chapter 2 then provides a full characterization of machine data utilized throughout this work (section 2.2). The last section of this chapter finally presents the problem constraints, introducing the main objectives endorsed.

The following chapter minds the manipulation of data considering the intervention of experienced personnel when interpreting the information (section 3.1). Then, it presents one of the main approaches of this work by referring to the concept of Sequence Windows (section 3.2.1), the utilization of *Apriori* algorithm (section 3.2.2) and finally the development of a visualization tool for presenting patterns in machine events – the Support Matrix.

The development of the intelligent model is detailed in chapter 4. It begins by defining the objectives of the machine learning algorithm and it's then followed by the construction of an evaluation system in section 4.2, allowing to understand the purpose of the model and its results. Previous to describing the modelling of the classifier (section 4.4) it is detailed one of the most essential tasks in this type of problems – the data preparation. In this section, a description of the technologies used is followed by the features construction process and the environment creation.

Chapter 5 delivers the results of the previous engagements divided into two perspectives: The Data Science and the Operational. While the first one describes typical indicators of the performance of machine learning algorithms, the other contextualizes the results from the point of view of the manufacturer.

The results of the approach taken are then discussed in the conclusions chapter (Chapter 6) as well as guidelines to further continuation of the presented work.

Chapter 2

Industrial Environment

In the current chapter, a detailed description of the system concerned in this study is presented considering three main topics. In section 2.1 is engaged the characterization of the industrial system addressed in the present thesis examining closely its environment along with upstream and downstream processes. A full perspective of the available data is then delivered in section 2.2, referring the data extracted from the machines, the contribution of the insights provided by experienced personnel and a detailed analysis on machine failures, relevant to understand the problem attended later in the chapter. Section 2.3 is focused on exploring the thematic of operational downtime motivated by failures occurred in the **target area** as long as the definition of the related circumstances.

2.1 System Description

The current thesis comprises a study of an industrial process for the production of a solid light weighted product subjected to strictly regulated quality standards.

Given the complexity of the process and all its variables, it was considered data of the two machines with the largest impact on overall downtime. From now on they will be referred as Maker and Packer, respectively the machine that constructs the product by assembling all of its components and the one that groups the individual products, also referred as Units, in a pack.

Bearing in mind the mentioned processual area, it is important to contextualize the upstream and downstream environments as they interact directly and indirectly. As Figure 3 suggests, Maker is fed by product's main component and other secondary constituents. The first element, **β**, is primarily processed and experiences several treatment steps in the initial stage of the global process, conceiving from the raw material a highly standardized substance with well-defined characteristics. However, particles that compose this component have variable size, weight and properties. Taking as an example, two particles that have experienced equal conditions of processual treatment can vary in ductility, density, and shape. This is mainly due to feedstock traits and confers on the product a blend of properties that gives its unique characteristics. Nonetheless, some particles may disturb the stationary functionality of the process if they present attributes that conduct to flow blockages, uneven distribution of weight and loss of material during product transportation. On the other hand, secondary constituents are less likely to retain properties that could interfere with machine stability even though existing slight irregularities common to every raw material.

Upstream, is located a sequential set of unitary operations designed to group the constructed packs in larger volumes concluding the global production process. In this case, the interaction on the system is mainly direct as a stoppage on upstream machines can represent an interruption on system flow.

To prevent this direct relation caused by upstream stops, there are buffers placed between the unitary operations, which accumulate a defined amount of product from one machine to the later. Reasonably, in the system Maker-Packer, the flow is not direct and accounts with a large buffer that can sustain about 15 minutes of continuous packer consumption. Considering that both machines are programmed to produce 14000 units per minute, the Buffer is estimated to hold 200000 units at average capacity. In the Buffer, the individual units have some freedom of movement that could enable the loss of material and moisture, minor deformation and accentuation of previous assembling flaws, hence interfering with the proper functioning of the system.

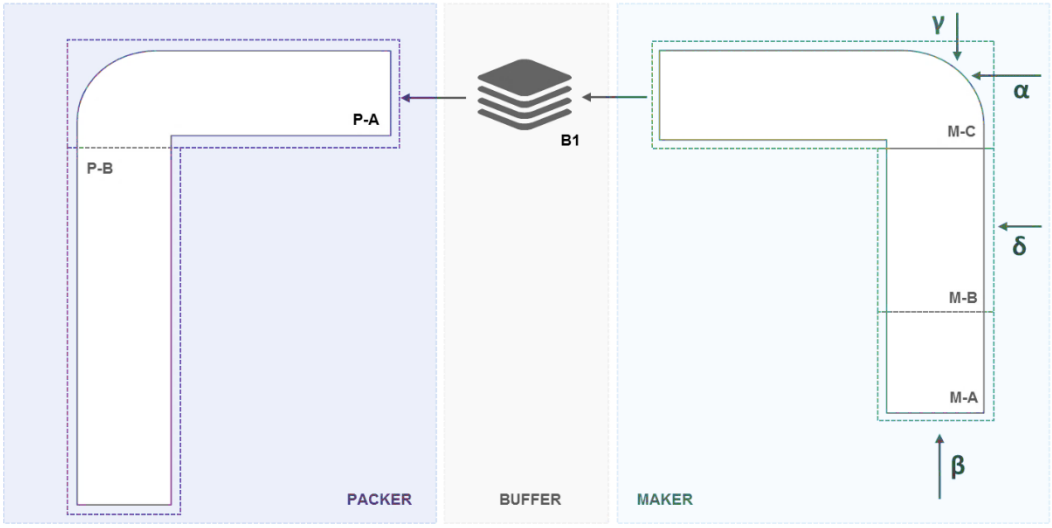

Figure 3 | Process diagram of target section compromising both machines addressed in the study (Maker and Packer) and the intermediate Buffer. Representation of intake areas.

As previously mentioned, Maker is responsible for assembling all of the constituents in order to produce the individual unit. With this aim, Maker can be subdivided into three stages:

M-A. Admission and distribution of feedstock;
M-B. Construction of preliminary product;
M-C. Formation of the final product (**Product A**) and the introduction of finishing steps crucial to grant desired properties.

After the individual unit goes through the buffer it's then forwarded to Packer intake zone where it's reorganized along with other units to fit the package geometry. Later, the group of units is directed to the second stage of the machine designed to build the pack. Accordingly, it is possible to define two stages for this operation:

P-A. Admission and reorganization of a set of units;
P-B. Construction of the pack (**Product B**) and coupling of secondary constituents.

Machines composing the System are current top performers and account with a robust rejecting system based on real-time measurement of well-defined parameters and assessment of product quality. In the case of product's properties doesn't meet quality standards, it will be rejected. This exclusion of non-compliant items could be not immediate for the reason that exists proper areas to make material discharge. Therefore, assuming that a produced good doesn't fulfill the necessary requirements, the machine will consider this product location and determines the moment when it passes through the discharge section in order to prevent more severe situations downstream caused by defective products. It is also essential to bear in mind that machines are programmed to stop if the number of rejected pieces overcome a defined threshold that could vary dependently on the equipment.

Despite the large number of reject motives in Packer, the first machine accounts only with a few parameters that are measured with high precision and stated below:

- Weight;
- Tightness;
- Weight distribution;
- Ventilation;
- Suction resistance.

Deviations on these parameters could result in a defective **Product A**, potentially inducing malfunctions in downstream unitary operations as it will be later addressed in section.2.3.

2.2 Data Characterization

As previously stated, data from two machines is the ground base for the current thesis. For this purpose, it was considered as sample a historical time span of 4 months that represents 100 days of production. Data were retrieved from a framework that integrates sensors from all connected equipment and provides aggregated operational information.

2.2.1 Machine Events Data

For both machines, the same data structure was adopted. A set of 30 variables describe the 14911 and 49550 rows of Maker and Packer, respectively. The fact of the packaging machine having three times more entries will be addressed in section 2.3.3.

Bearing in mind the confidential responsibilities of the information involved in the study, any material that could lead to the manufacturer was omitted. It is nonetheless necessary to characterize the variables adopted in the current thesis.

- **Start Time and End Time:** Considering the data structure that will be attended later in this section, each row defines a time period from Start Time to End Time;
- **Is Stop:** Flag that indicates the moment when the machine stopped;
- **Category:** Broad description that defines machine status;
- **Sub-Category:** Detailed description that defines machine status. For events with high duration this field can be manually introduced by the operator;
- **Equipment Causing Stop:** This variable indicates what equipment is responsible for the machine stop. Possibilities are the machine itself and upstream or downstream equipment;
- **Stop Reason:** Characterization of the actual motive that induced the machine to stop. The reason provided may not be the root cause of the event;

- **Average Speed:** Taking into account the time period, it presents an average speed in Units per minute;
- **Total Product Produced:** Total amount of product made by the machine irrespective of its condition;
- **Rejected Production:** Number of Units rejected given their condition;
- **Good Production:** Result of the difference between Total Product Produced and Rejected Production;

Without treatment, the information used from both machines has to be analysed as two independent datasets as a result of an asymmetrical temporal domain. To confirm this declaration, it can be considered the following conjectures that are a consequence of the way data was generated from the system and are valid for both machines:

- Machine events data are a continuous data set grouped in time periods by machine status;
- Each row represents an alteration of machine status from the precedent row, thus a new event;
- Data is arranged in descending order of time;
- No major maintenance operations were performed on the machines;
- No manipulation has altered the data, thus providing a plain exportation of measured variables.

Addressing the clarification of first and second statements, further demonstration is essential. Taking as an example the starting of production, one row should be generated and it's expected to demonstrate the status *Starting* appended to the Category *Running*. Then, assuming correct functionality in the initial stage, the next row on top of the previous would make a reference to *Normal Run* status. Also inserted in *Running* Category is the following stage – the *Ramp-down.* Whenever the machine identifies a deviation in the normal behaviour of the production line, it initializes the slowdown process until it finally stops. Examples of the presented deviations are, among other possibilities, lack of feedstock or the number of rejected Units superior to a defined threshold.

These three status identified in the machines represent a typical sequence in the data set. *Starting* always follows a stop or the begging of a production, and it is proceeded by *Normal run* after an average time of 30 seconds. Notwithstanding *Ramp-down* status being essential to complete the stop procedure, it isn't necessarily true that it will occur. More severe circumstances usually determine an immediate stop of the machine, independently on the current status (*Starting* or *Normal run*).

The integration of the two data sets constitutes the first challenge engaged in this thesis as represented in the following illustration:

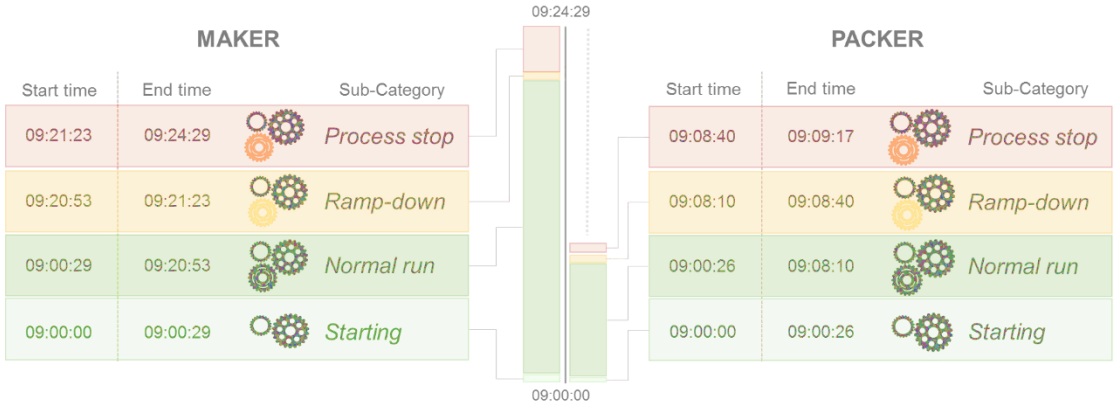

Figure 4 | Transposition of an operational example into the data set structure.

Chronologically, it's not possible to integrate information from both machines into one data set containing all the variables mentioned above. Bearing in mind that each row in the data set characterizes the machine status in a given time span, it's correct to assume that each coloured block represents an entry. Therefore, as Figure 4 suggests, even considering that Maker and Packer start their production simultaneously, the periods of one machine does not match necessarily the time windows of the other. From now on, the presented blocks will be referred as Events as they are the reflection on the data of the events occurred during the operational time.

The presented evidence makes unattainable the calculation of any variables associated with the equipment that precedes the target area. The Buffer is fed by Maker and consumed by Packer, declaration that can be defined by the relation below:

$$Buffer\ Size\ (Units) = Initial\ Size + Maker\ Good\ Production\ - Packer\ Total\ Product\ Produced \quad (1)$$

In the closing stage of each production, Buffer is cleaned out in order to prevent changes in product properties motivated by long exposition to external conditions, such as temperature and humidity. For this reason, the Initial Size of the Buffer is zero at the beginning of each production, summing up the previous equation to the difference between the Units that Maker evaluate as acceptable and every Unit that Packer utilizes to compose the packs, independently on their conditions. However, given the asymmetry of both data sets, the estimations resulted as not adequate to the problem as they didn't provide a viable appreciation on the Buffer size that could define the amount of time needed for a Unit to go from Maker to Packer. It is also essential to take into account that the Average Speed is determined considering only the final and initial speeds and for that reason, information on instant production could not be calculated from this variable.

2.2.2 Human Expertise

A crucial part of the present thesis concerns the expert knowledge of the people whose daily responsibilities comprise direct contact with the system in study. Operators experience is essential to understand raw data from the machine as they have the expertise to transpose the information recorded into concrete day-to-day situations. Evolution in manufacturing with human-centered process automation defined its role in the industrial environment as a decision maker in planning and controlling, aided by technology [13].

In an initial stage, insights from operators, shift supervisors and maintenance personnel provided three valuable resources included in:

- **Problem Definition** (section 2.3)**:** Recognition of important variables and relations within or between machines, allowing to focus on a target area.
- **Aggregation of events** (section 3.1)**:** Aggregation of failure status that represents the same effect. Development of machine's digital twin.
- **Data Integration** (section 3.2)**:** Interpret machine behaviors from data. Validation, identification of anomalies and exclusion of entries.

2.2.3 System Stops Analysis

Notwithstanding the process is the same for different products, the system behaviour is slightly different depending on product constituents and their characteristics. Therefore, the study was centralized in the product that presents higher production, necessarily having more historical data to work on. In addition, its production has the lowest MTBF, a simple indicator that takes into account the total uptime and divides it by the number of stops. For this variety, the value of this indicator referred to the historical period was 19.0 and 6.73 minutes respectively for Maker and Packer. In proportion, the making machine holds an MTBF that is almost three times superior, declaration that promptly justifies the differences existing in the number of rows of each data set. For the same operational time, lower MTBF implies a larger number of stops, thus representing higher variation in machine status that consequently generates more rows. However, only 80% of Packer stops are attributed by the operational system to Packer responsibility, while the remaining percentage is associated to Maker and Downstream machines. This monitoring system integrates all machines and takes into account the stops caused by shortage of feedstock and buffer limitations. As an example, if the making machine stops for a duration such that causes the Buffer between the two machines in study to reach a defined minimum threshold, for security reasons, Packer will also stop, thus imputing this stop to Maker's responsibility. Assuming that any of the subsequent equipment fails, Packer will continue to produce until it finally fulfils downstream buffer capacity, necessarily forcing the machine to cease production. In this case, the operational system accredits the *Equipment Causing Stop* to Downstream Machines.

Bearing in mind the described concept of *Equipment Causing Stop (ECS)*, it was conducted a study on machine failures.

As previously mentioned, Packer's location in the process is highly sensitive to uniformities in feedstock properties. Given that this feedstock is a result of Maker's *Good Production,* it is fundamental to describe the behaviour of the making machine. With this aim, two dimensions were considered:

- **Average rejected units by stop**: This indicator depicts the number of rejected units that a stop by a given *Stop Reason* rejects in average, providing an estimation on the extent of the damage motivated by the failure.
- **Average duration by stop:** It defines in average the amount of time spent by the operators to fix the failure and restore the operation. Can also be considered as an indication of how frequent this failure happens, considering that a given common type of stop is more efficiently handled than an occasional stop, which requires time evaluating peculiar variables.

The indicators above were determined by equations (2) and (3), respectively.

$$Average\ Rejected\ Units\ by\ Stop\ of\ Stop\ Reason\ X\ (Units/stop) = \frac{\sum_{x=0}^{x=n} Rejected\ Units}{n} \quad (2)$$

$$Average\ Duration\ by\ Stop\ of\ Stop\ Reason\ X\ (Units/stop) = \frac{\sum_{x=0}^{x=n} Duration}{n} \quad (3)$$

Considering the universe of stops in the making machine and their relative position, it was expressed in the diagram below three areas related to the stages identified in Figure 3. In brief, it's appropriate to declare that M-C is the section that presents a more complex environment as it has a high number of rejected Units in each stop and includes a vast window of average duration overcoming the challenges imposed by the failures. On the other hand, stoppages inducted in M-B are promptly solved as they are highly frequent and don't require a complicated intervention, often being a common daily operation.

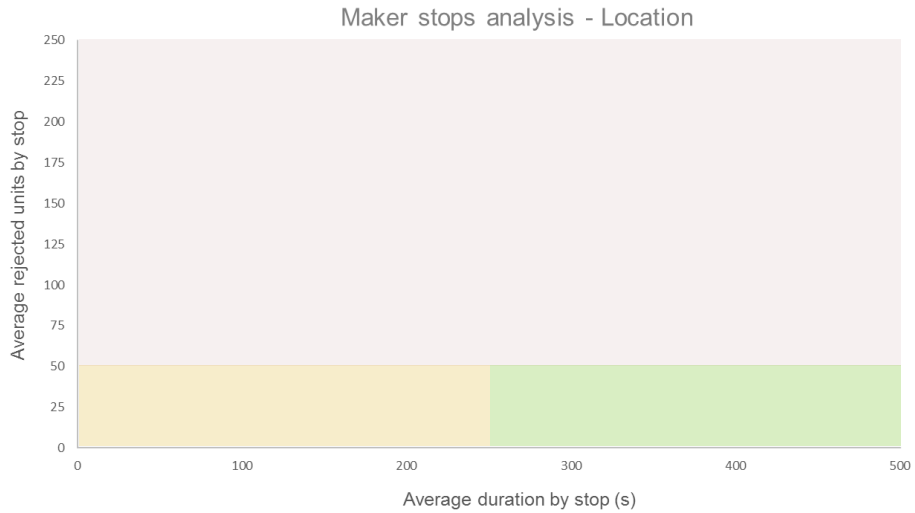

Figure 5 | Characterization of Maker stops given their location. M-C (Red on top); M-B (Yellow on bottom left); M-A (Green on bottom right).

Bearing in mind the previous diagram and the matter addressed on this thesis, a set of eleven Maker *Stop Reasons* is presented from the carried study of all failures. This set is a result of the approach addressed later in section (3.2.3) and is part of the correlation settled between the two machines, established by the principle that a defective Unit produced by the first can later disturb the normal operation of the following equipment. In the representation below, the *Stop Reasons* were identified with a code due to the sensibility of the information given and coloured based on their relative location in the machine.

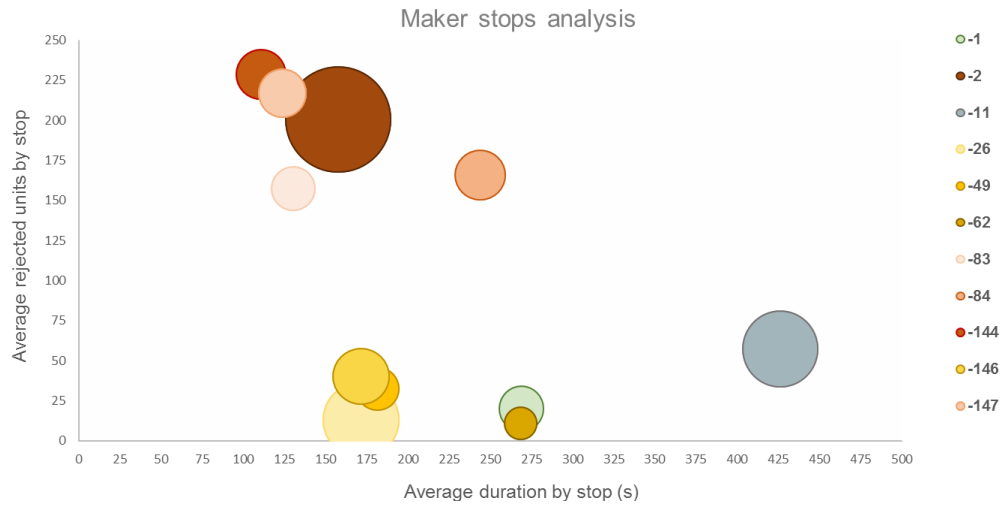

Figure 6 | Bubble representation of Maker stops analysis. The bubble size is given by the frequency of the stop (number of occurrences in the historical data). M-A (green, bottom); M-B (yellow palette, bottom); M-C (red palette, top); External Component (grey, bottom right).

It is important to refer that the codification used to decharacterize machine stops is the same as the presented in this document in order to guarantee the integrity of the analysis taken. On the other hand, the following table is devoted to describing each stop providing information on the location, metrics used in the representation and potential causes and consequences.

Table 1 | Detailed characterization of maker stops.

Code	Avg. Duration by stop	Avg. rejected units by stop	Location
-1	269	20	M-A
-2	158	201	M-C
-11	426	57	M-B (External)
-26	171	13	M-B
-49	182	32	M-B
-62	269	11	M-B
-83	130	157	M-C
-84	244	166	M-C
-144	111	229	M-C
-146	171	40	M-B
-147	124	217	M-C

In the presented study, information from occasional events with excessive impact on the overall indicators was excluded. In this category are included the *Breakdown* and *Long process stop* records as they don't provide a substantial picture on the daily operation and typically induce inaccurate conclusions.

A similar approach was also conducted to examine Packer stops, however, as it will be addressed in the coming section, the conclusions are not applicable to the focus of this thesis.

2.3 Problem Definition

2.3.1 Identification of the target area

Packer is accountable for the global process major downtime. The behaviour is fairly justified if we consider this machine as the concentration of a variety of flaws, minor defects and irregular features of feedstock units. On the first stage of the global process, compromising the treatment of the main component, these peculiarities are easily imperceptible as they reside in a large amount of material. Transposing from the mentioned stage to the studied system, the items involved are about 10^6 times lighter and require high precision handling, thus amplifying the impact of the mentioned variations on the line operation. On the other hand, Packer intake zone represents a bottleneck both literal and figurative to the production process.

Bearing in mind that this section is preceded by the buffer, which proportion has already been specified, it's important to refer that the products are grouped into two trays with a few dozens of units per each, being Packer necessarily a high-performance machine to carry the production rate. In order to guarantee the line's maximum uptime, every unit produced by maker must have standardized properties within the boundaries of the quality parameters. For the above reasons, Maker and Packer compose a complex system with multiple interactions and variables that require continuous study, suitable as the focus of this thesis.

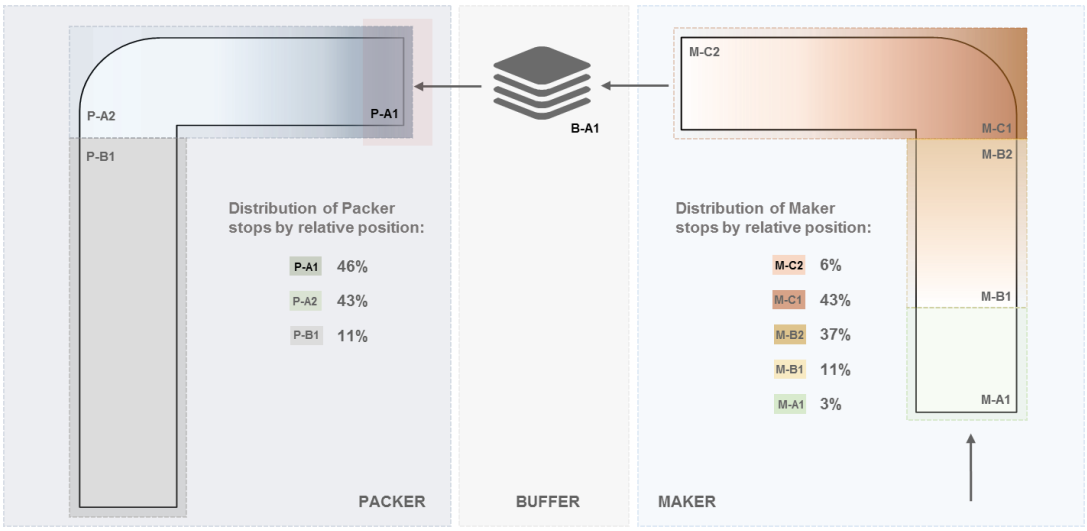

Figure 7 | Process diagram of target section embedded with stop analysis based on relative position. P-A1 is marked in red as the target area.

The preceding diagram supports the distribution of failures in the two machines granting an intuitive perspective of most distressed areas. Notwithstanding M-B and M-C accounting with a large number of stops, Packer's location P-A1 accounts with more than 50% of all system failures. This section presents a peculiar set of characteristics related to machine stoppages as they have a short average duration of 67 seconds, meaning that an operator is able to identify and quickly mitigate the disturbance. In the industrial environment, when an obstacle is easily managed, it is typically comprehended as a daily reality underestimating the impact that can have in the production and in the activity of the operator. Nonetheless, given its high number of occurrences, these particular failures have an extensive repercussion on the global downtime - close to 2 weeks for each year. In addition, as previously mentioned, one stop can generate a stage of *Ramp-down* and necessarily a *Starting* stage. The two phases of the operation are moments of low production rate and perform a total of 9 days without taking advantage of the capability of the machine in its target speed.

Aiming to fully understand what variables can induce a stop in P-A1 is foremost necessary to define the *Stop Reasons* involved. As described above, this section portrays a bottleneck, where hundreds of Units are forwarded to fit in a space for small dozens in a high-speed performance machine. The indicated circumstances along with uniformity of feedstock Units compose the ideal environment for one particular type of complication, the blockages. In this case, it is possible to define a possible situation that can represent the vicissitudes found in the **target area** and provide meaningful information giving context to the declarations:

A defective Unit produced in Maker inverts its position while approaching the **target area** causing the clogging of one of the channels that lead the Units to Packer's intake zone. The clogging restrains the possibility of fulfilling all the required space in the pack and forces the machine to stop and reject the incomplete production. The stoppage can be identified by ten different *Stop Reasons*, however, later in section 3.2 will be addressed the method adopted to merge similar *Stop Reasons* based on operators' insights.

2.3.2 Perception of circumstances

According to the above line of reasoning, P-A1 is an evident choice for the **target area** giving that:

- Failures in that section are frequent;
- Blockage is mostly motivated by uniformities in feedstock Units;
- Operators have a clear knowledge of the actions they should take to restore the normal operation;
- Total downtime is heavily affected by PA-1 blockages.

On the other hand, there are still some variables that should be considered as a barrier to getting a definite picture of the system. The following line-up provides a description of the main challenges addressed considering the focused area:

- Machines' data sets are asymmetric in time;
- Lack of information on sensors data or other variables;
- Unworkability on buffer size estimations;
- High rate of failures near P-A1;
- Defective maker products can also perturbate other sections.

Chapter 3

Data Pre-processing

Data extracted from machines and correspondent equipment is often subject to high variance in the information, composed by an abundance of noise and imperceptible elements. This chapter addresses the extraction of relevant information from both data sets. Concerning further machine learning application, it is not a good practice to bring all the available information, on the opposite, it is crucial to understand and separate significant data in order to improve quality and reliability in the solution. Section 3.1 introduces that process by aggregating events based on their similarity, avoiding ramification of the outputs. Afterward, the conditions are set to begin the process of extraction of the relevant variables, following a mechanism endorsed in the current thesis. The conception of a Support Matrix allows the identification of patterns in the behaviour of both machines and provides a set of Maker events essential to the operationalization of the classification model, the **input events**. This mechanism is assembled using statistical techniques relying on the construction of Lookup Windows and adaptation of *Apriori* algorithm as completely described in section 3.2.

3.1 Aggregation of events

Bearing in mind the knowledge of the operators and maintenance personnel regarding the system itself, it is possible to delineate some ground base assumptions for the analysis of machine data. In the first place, it should be considered that the *Stop Reasons* are registered by the Operational Framework and are a result of the integration of the information from multiple devices that detects or measures physical and mechanical properties.

These devices, can detect the presence of unexpected components, prevent machine damage and loss of product quality. Generally, the *Stop Reason* recorded by the Operational Framework is directly related to the device that flagged the situation and this raises two important points:

1. The *Stop Reason* identified by the system may not be the root cause that motivated the stoppage;
2. More than one *Stop Reason* can be found to be related to the same stop.

Considering the second point, the operators' perspective was crucial to identify which *Stop Reasons* could be linked to the same cause, thus aggregating the data based on similarity, managing the noise induced by multiple variables in the later analysis. Note that the reason provided by the system is the same status that appears in the machine monitor for the operator on his daily activity.

As previously mentioned, the study is focused on stops that are motivated by blockages in Packer's intake zone. However, along with the above line of reasoning and considering the second statement, these clogs can be identified by more than one *Stop Reason*. The **target area** is therefore comparable to a sequential set of nets design to permeate the passage of Units without defects. Examples of this situations are physical deformations and uniformities such as lack of components or their incorrect assembling.

On the other hand, each figurative net is composed by a series of connected devices arranged to identify any alarming behaviour having the possibility to induce a machine stoppage. The following alignment describes the consecutive progress of nets considered in P-A1, taking into account that the codes used to map the *Stop Reasons* adopt the same logic than those referred earlier, employing in this case positive identificatory for Packer stops.

Table 2 | List of *Stop Reasons* recognized in the **target area**.

Code	*Stop Reason*
121	Hopper malfunction
80	Operator visually identifies an improper situation in the hopper
12	Faulty vibration of the equipment
7	Unit not detected in destination channel
4	Machine turret covered, Left
18	Machine turret covered, Right
6	Missing component **β** in track 1
17	Missing component **β** in track 2
9	Missing component **α** in track 1
13	Missing component **α** in track 2

The packaging machine comprises two parallel production tracks acting as twins and designed for the synchronized manufacturing of a pair of Packs.

Blockages in the **target area** are from now on described by code 1000 also referred as the **target stop**. Since any of the *Stop Reasons* labelled in Table 2 was classified by operators as a potential identifier of the blockage, the codes were combined in one, in furtherance of the analysis performed on this thesis.

3.2 Data Integration

When data integration is referred, it is usually associated with the combination of data from more than one source aiming to provide a full perspective of the domains, facilitating the processing and interpretation of the information. In this particular case, the strategy of integration of both data sets plays a central role in the development of the solution.

Engaging the consequences of blockages identified as **target stops** and considering their occurrence is in the intake zone of the machine, it is fundamental to examine preceding variables as potential motivators. The integration of two asymmetric data sets is the primary challenge addressed in this thesis and it is settled on three consecutive concepts:

1. Sequence Windows;
2. *Apriori* Algorithm;
3. Support Matrix.

The mechanism developed on this thesis provides an innovative approach to reproduce insights from machine data adapted to overcome conditions of lack of information on independent systems.

So far it has been described the system and related variables, the available information and also the problem itself and its constraint. From now on, the approach exploited to address this thematic will be the focal point.

3.2.1 Sequence Windows

Pattern Discovery is embedded in the definition of Data Mining, as it is the detection of similar structures on large data sets. The largest is the data set, more likely it is of having a high content of data distortion and uninteresting patterns, becoming impractical to apply simple statistics on data and consequently paving the way to data science through machine learning and other advanced computer techniques. Notwithstanding, deciding whether the patterns found are or not relevant and pertinent to the circumstances, should take into account the operational context and knowledge from experienced people whose function is directly associated with the subject [14].

Aiming to find similar patterns in the available data sets, modifications had to be made on the schema and structure of each one to fit them together. The first and more immediate adjustment is certainly directed to the limitation of asymmetry in time of both data sets. As previously mentioned, information of each machine is independent and it is a result of the evolution in the operational status. Accordingly, Maker has events in different time spans than Packer, what makes inconceivable the aggregation of two dimensions: *Start-time* and *End-time.* To conveniently associate this information only the first one was considered. However, this assumption originates a particular disadvantage: without a time span, the association of variables is not possible. For instance, only considering the *Start-time* the declaration “At 09:00:29 was produced 4000 Units of which 500 were rejected”, is no longer coherent because in this case the concept of instantaneous production it’s not achievable. On the other hand, it is valid to assume for example that the machine stopped at 09:21:23 with the *Stop Reason ID = -2*. This limiting opportunity makes it attainable to identify patterns in the sequence of Maker events and even in the interdependency between equipment that comprise the system. Bearing in mind the former perspective, data from the two data sets was aggregated and organized in descendant order of *Start-time*, being the latest event the first row of the compiled data set.

Further modifications minded to addressing one evidence common in the machine information which has a direct relation with its operation. Whenever the connected devices identify an abnormal behaviour, the stopping process begins and it differs from failure to failure. This process starts with the deceleration followed by a sequence of stages until it reaches the final programmed condition. For the purpose of this analysis, only the first stopping stage was considered valid, representing the actual *Start-time* of the event. In addition, and along with best practices considered in Data Mining, the resulting data set still contained non-vital information to pattern identification. Therefore, only the stop status was considered excluding all *Running* events from both machines. This resulted in a two-column data set (*Start-time* and *Stop Reason ID)* which comprises 15260 rows.

That said, it is appropriate to assume that there is a sequential set of events with different *Stop Reasons* forming a chronological sequence that illustrates the behaviour of both machines regarding their stops.

Figure 8 | Illustration of the application of Sequence Windows technique. The colours represent the relative location of the events based on the set previously applied.

Intending to identify patterns in the sequence of events, the methodology described by *Vilalta et al* [15] references a similar approach as the one illustrated by Figure 8. The **target stops** define the beginning of the **Lookup Window**, a well-defined time span proposed to determine what events occurred from the *Start-time* of the target event until the edge of the window. In the developed concept, the proposed methodology was adapted to engage the thematic of equipment interdependency.

As depicted in the representation above, the objective was to set a window with a fixed size immediately before a **target stop**, with the ability to identify the preceding Maker events during the given time span. This purpose was accomplished by developing a computational algorithm proceeding in three steps:

1. Importation of the combined data set;
2. Iterate in each **target stop** creating the window and returning the events contained in the time span (Loop Strategy);
3. Exportation of a list of maker events. Each window originated a row of occurred events.

The generated list contains crucial information to identify patterns minding the dependency of the two machines, as it expresses every event that happened in Maker in a well-defined period preceding Packer **target stops**, composing the input for the next topic.

3.2.2 *Apriori* Algorithm

Finding patterns in a processed data set can benefit from statistical techniques, which can be implemented and improved with the use of algorithms. Therefore, the *Apriori* algorithm was considered suitable to the problem. This method is widely adopted to perform temporal data mining and is known for identifying association rules between frequent events and also providing statistical indicators as the support value [16].

Support is directly related to the frequency that two events occur in the data set and can be expressed by the following relationship:

$$s\ (A\ =>\ B) =\ P\ (AB) = \frac{N(AB)}{|D|} \tag{4}$$

Where $N(AB)$ represents the number of times that A and B are simultaneously present in a transaction being $|D|$ the universe of all transactions present in the data set [17].

To better understand the fundaments of *Apriori* it is essential to recognize the following concepts:

- Transaction: Set of events associated with a time. Each row in the input data set is a transaction;
- Item: Base unit of the transaction. Each event is an item;
- Support (**s**): Defined as the percentage of transactions that contain a given item;
- Rule: The rule $-1 => 1000$ with $\boldsymbol{s} = 10\%$ refers that the event with ID -1 occurs before the **target event** in 10% of the transactions presented in the data set.

For instance, considering the illustration depicted in Figure 8, four rules and respective supports can be defined, one for each *Stop Reason*:

- $s\ (-1\ =>\ 1000) =\ \frac{1}{3} \approx 33\%$
- $s\ (-2\ =>\ 1000) =\ \frac{2}{3} \approx 67\%$
- $s\ (-26\ =>\ 1000) =\ \frac{2}{3} \approx 67\%$
- $s\ (-83\ =>\ 1000) =\ \frac{1}{3} \approx 33\%$

Considering that the input data set was originated from the application of the method previously described, it can become undeniable that all rules must have in common one item, the ID 1000, nevertheless, this is a misconception. *Apriori* algorithm is prepared to identify rules between all items, which transposed to the studied case means that it is also optimized to analyse patterns in Maker events. As an example, the rule $(-1\ =>\ -2)$ could be as well an output of this technique, providing interesting information on Maker behaviour. Despite the fact that this information could assist to perceive correlations between failures in the making machine, it could not be considered as fundamental to address the focus of the current thesis, therefore, further work will be mentioned accordingly.

3.2.3 Support Matrix

Differently from the exemplification presented above, the data set accounts with a large number of items, consequently meaning a broad amount of possibilities. Considering the 56 *Stop Reason* that can induce Maker to a stop, there is the same number of rules that can be defined with the employment of the *Apriori* algorithm.

Nonetheless, it is important to be aware that Maker and Packer are separate machines and often present failures that are not related to each other, knowing moreover that the only way that Packer can induce a stop in Maker is if the Buffer's capacity is in a state that requires the first one to cease production. Therefore, it becomes evident that not all the derived rules are relevant to the analysis and it is essential to consider two meaningful aspects:

- A threshold of minimum support, **s**, should be defined with the purpose of obtaining only rules that are sufficiently supported by this statistical indicator.
- Insights from experienced personnel play also an important role deciphering each rule with an effective support. This action provides fundament to the rule itself, complementing the output of the algorithm with the transcription from the operational situation.

The **Support Matrix** is the combination of the two concepts elucidated in the present chapter (Sequence Windows and *Apriori* Algorithm) and it represents one of the innovative aspects conceived in the formulation of the current thesis. It is designed to address the adversities stated earlier on the integration of two asymmetric large data sets and the analysis of the interdependence existing between equipment.

Depicted in Figure 9, the matrix is composed by two axes (size of Lookup Window on the top and ID of Maker Event on the left side) and it provides an intuitive perspective of the rules between a pair of items, the **target event** and the Maker Event. The stronger the correlation, highest is the support and darker the colour set on the respective field. The interpretation is done by considering the concept of Lookup Window detailed earlier and it follows the logic exemplified by the following outcomes:

- The *Stop Reason ID* -2 occurs in 30% of the transactions generated with a Lookup Window of 16 minutes;
- Support associated with *Stop Reason ID* -1 only becomes significant when the Lookup Window is greater than or equal to 17 minutes;
- This set of ten types of Maker Events presents the failures that can occur in Maker and have a significant impact in Packer's functionality given variable time spans from 1 to 20 minutes

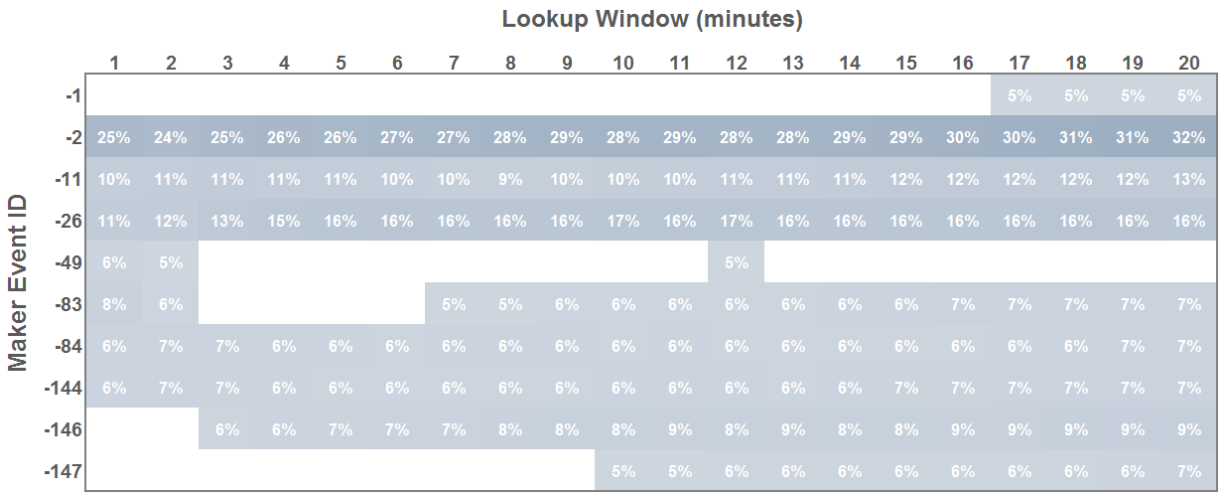

Figure 9 | Support Matrix resultant of the mechanism developed for the integration of both data sets.

Note that minimum threshold was set to a support of 5%, meaning that the set of Maker Events comprises only the ones with support greater than this threshold in at least one of the Lookup Windows.

Support Matrix is, for this reason, a powerful tool in the identification of patterns and correlations between events. One of its main advantages is the versatility shown dealing with incompatible data sets, only requiring one common variable related to time. It provides insightful information on equipment behaviour, creating relations among events, pointing out the statistical strength of the relation.

In the current hypothesis, Support Matrix has another decisive purpose. Its output, the set of Maker Events with a potential association with Packer's **target stops**, plays an essential role in the following section, which comprises the conception of the intelligent model developed to engage the problem. The performance of the model is highly influenced by the process of variables and features selection, important to build faster and more cost-effective predictors facilitating data understanding and reducing training times. One of the earliest stages in the construction of a predictive model is the selection of the input variables. In this stage it's important to visualize what data is unnecessary and has the potential to induce noise, selecting only a subset of features based on relevance and considering their redundancy [18].

In the present section, it has been described an approach to address the selection of relevant Maker Events with the purpose of using this intel to feed the predictive model. From now on, this set of *Stop Reason IDs* will be referred to as **input events**, as they are the ground base to the selection of related features and the development of the solution.

Chapter 4

Model Development

The present document approaches the construction of a machine learning algorithm, referred also as the model, aiming to address the problem of blockage in the intake zone of the packaging machine. The development of the model followed the usual practices of machine learning referenced in the literature. It is initialized by the purpose endorsed by the model (section 4.1), followed by the development of an evaluation system carried to fit the results in the operational context (section 4.2). One of the most important steps minding the development of a machine learning model is the preparation of the data that will serve as an input for the algorithm. Section 4.3 approaches the technologies used and the transformations required to process the data set. Later, the modelling of the classifier is addressed by taking into account the needs in the manufacturing industry environment. In machine learning techniques, the main goal is to retrieve knowledge from data, being fundamental to understand how the algorithm deals with its feedstock. This is the major challenge when implementing machine learning models and is addressed in the current chapter as a way to generate insight on the interdependency of both machines, Maker and Packer.

4.1 Projected Purpose

From operations to process, the application of machine learning systems in the manufacturing industry has been a reality. The integration of this intelligent systems capacitates the manufacturer to understand patterns and generate insights from large data sets [19]. As previously described, supervised learning is related to the prediction of a target function. This form of machine learning is often used in the manufacturing environment as the majority of problems can be divided into two categories:

- Predicting a Quantity: Automated techniques are used to estimate a production quantity by correlating a multiplicity of variables and information of sensors and other devices.
- Predicting a Category: Classification of categories address the most common manufacturing issues where the estimator predicts a defined class label based on the values of the related variables.

Also frequent when referring to the manufacturing data source is the noisy data resultant for example from limitations of measuring instruments and also human error typing logs into a computer. This circumstance, if not resolved, has the ability to limit the achievable accuracy of the mechanisms development, therefore, one distinctive attribute that should be considered when evaluating the model, is its robustness [20].

Developing a machine learning model should invariably begin by the definition of the goals and capabilities that are expected from it. In the current thesis, the mechanism must be capable of addressing a set of situations widely known in the manufacturing environment:

- Handle noisy data comprising outliers;
- Process large data sets from more than one source;
- Generate perceptible insights;
- Provide a positive impact on production and its daily activity;
- Possibility of scaling the model to ensure near real-time processing.

Bearing in mind the analysis taken on the previous chapter and the points stated above, it becomes evident that the problem minded in the current thesis can effectively be handled by applying an estimator for Predicting a Category, also referred as a **classifier**. Accordingly, the projected purpose will be the central thematic of the following line of reasoning.

The **target stops** are a constant reality in the daily operation of the studied line. Nonetheless, in these particular failures, the job of searching for the defective location and resetting the normal behaviour of the packaging machine is done in an effective way. As the operators are familiarized with these blockages, it only takes about one minute to solve them. On the other hand, the impact that the **target stops** have on overall downtime makes inevitable to attend them. It's not only the production that is affected by the number of times these blockages occur, but also the operators' performance should be considered. Looking from the operators' point of view, the blockages occur randomly and with high frequency, interrupting any task that is being performed. Therefore, the model developed must consider this perspective and also taking into account the set of situations mentioned above.

Bearing this in mind, it becomes evident that the goal of the machine learning model was to provide an estimation of when a **target stop** would occur. Transposing from the operation to the model, it means that the conceived **classifier** has the ability to predict an event based on the information from its variables. In the operation, the same trained classifier should be capable of process real-time data and provide a warning when a blockage is predicted.

4.2 Evaluation System

After laying out the main goals of the model, it's essential to define the evaluation grid, that can answer to one of the most significant questions: How it should be measured the performance of the model?

This is not a straight answer and usually has two perspectives that must be considered: the operational and the mathematical. The first one accounts with metrics such us, the impact of the model in the production and the complexity involved in scaling up the algorithm. From the data science point of view, there are frequent metrics that perform a complete evaluation of the machine learning algorithms. Both perspectives are essential and could not remain without each other, a situation that certainly implicates an inaccurate assessment.

Present in the operational context, it were defined the following metrics:

- Impact on MTBF: This indicator is widely used in the manufacturing industry, acting as an appropriate way of quantitatively measuring the model performance;
- Scalability: The algorithms utilized must allow the possibility of being scaled up to fit the needs of the operation;
- Degree of operationalization: The machine learning mechanism should be transposable to the production environment.

Prior to defining the attributes for evaluating the model based on the second point of view, it is important to briefly introduce the quadrant that sums up the results of the model, also known as the confusion matrix.

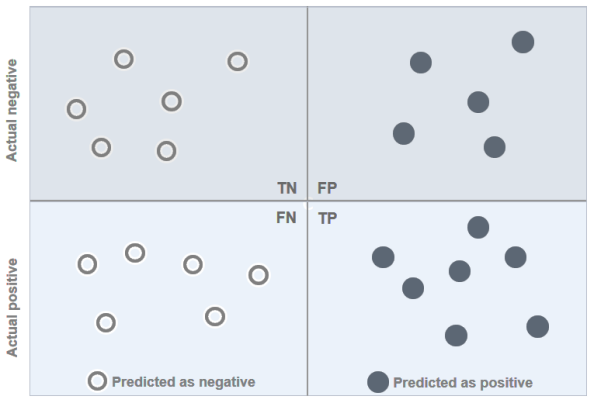

Figure 10 | Diagram of the confusion matrix used to display machine learning outputs.

From the above illustration, a **classifier** can have the following outputs:

- True Positive (TP): Positive prediction that is in fact positive;
- False Positive (FP): Positive prediction that is actually negative;
- True Negative (TN): Negative prediction that is in fact negative;
- False Negative (FN): Negative prediction that is actually positive;

Concerning the data science perspective and taking into consideration the confusion matrix illustrated in Figure 10, the selected metrics are detailed below:

- Sensitivity or True Positive Rate (TPR):

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

  Measures the proportion of actual positives that are correctly predicted from the total amount of positive instances.

- Precision:

$$precision = \frac{TP}{TP + FP} \tag{6}$$

  Measures the proportion of positive predictions that are actually correct.

- Accuracy:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

  Generally, indicates the fraction of right predictions, both positive and negative.

- F1 Score (F1):

$$F1\ Score = \frac{2TP}{2TP + FP + FN} \tag{8}$$

  Represents the harmonic mean of precision and sensitivity.

- False Discovery Rate (FDR):

$$FDR = \frac{FP}{FP + TP} = 1 - precision \tag{9}$$

  Measures the proportion of positive predictions (discoveries) that are actually incorrect.

This last indicator is extensively adopted when the classification of positives predictions must reflect a high yield of actual positives. Besides expressing the probability of error, FDR is frequently associated with other costs indicators, as minimizing this metric means building a cost-effective model [21].

For instance, in medical diagnosis, FDR should be one of the most important measures used to evaluate the model's performance. In this environment, it is only possible to rely on the algorithms if their classifications represent a lower FDR, otherwise, one may tell the patient he presents a positive result for a particular disease though not being actually a positive.

In the manufacturing industry, most of the decisions involve time and cost consuming tasks, therefore, FDR expresses an effective metric to evaluate the performance of the model, serving both perspectives: operational and mathematical. Naturally, one may affirm that every problem has its peculiarities and the response given must address different goals. For this reason and considering that it's not feasible to develop a model that can outstand all existing metrics, it's essential to define the optimization strategy that will be followed when training the model.

Taking into consideration the context this study is inserted, a balance between TPR and FDR was the aim defined for the model construction. Transposing to operational conditions, it is important that the algorithm displays the capacity for estimating a satisfactory amount of **target stops**, providing minimum false discoveries, known in manufacturing ambience as "false alarms".

4.3 Data Preparation

The preparation of the data fed to the **classifier** plays a central role in the model development. Later of being processed, data should be restructured to fit the input framework. The following alignment describes the path of data from the moment of its processing until the one that it is used to train, test and validate the algorithm.

4.3.1 Adopted Technologies

Machine learning can be sustained in several technologies being the most important aspects of the construction of the learning strategy and the selection of the input variable. However, some frameworks allow a wider range of possibilities and simplify these two processes. On the current thesis, the utilization of *Microsoft Office Excel* and *Python* were the main partners on the model development.

The first one is a powerful tool that allows the visualization of the data, facilitating processes such as the identification of outliers. On the other hand, *Python* accounts with a wide range of machine learning algorithms, what makes this programming language the preferred to engage in the data science thematic. Taking into consideration that the **classifier** should be effectively evaluated, S*cikit-learn* was the toolbox chosen to perform and test the machine learning algorithms [22].

4.3.2 Feedstock Data

Given the described steps for the development of the model, emerges another fundamental problem concerning the selection of variables. However, there isn't a systematic way of performing this process that responds to every situation, thus requiring a customized perception of the relevant data [23].

Considering the data sets of both making and packaging machines, one can define the following alignment of attributes describing each Maker failure:

- Average Speed;
- Total Product Produced;
- Rejected Production;
- Good Production.

Two rules that are on the basis of features selection affirms that the considered variables must not be redundant and should have meaning to the classification. Constructing a **classifier** is highly influenced by the set of features adopted, and should aim to realize the highest generalization performance and fastest classification [24]. This final variable, "Good Production", is a result of the difference between the "Total Product Produced" and the "Rejected Production" and it was not considered for the reason stated above.

Given that it is a common issue in machine learning, the situation of dealing with a large number of features tends to be the most important part of the development. Notwithstanding the reduced number of features in this particular case, there was a high amount of possible *Stop Reasons* that could be related. For this reason, a special attention was directed to minimize the number of Maker failures accountable for the input data.

As detailed in Section 3.2, a Support Matrix was developed considering patterns in the behaviour of both machines and their interaction. This apparatus provided an output essential for the definition of feedstock data, as it declares the Maker *Stop Reasons* that are potentially relevant to the current study. Each one of these *Stop Reasons* is able to be described by the previous enumerated features as well as other derived variables that will be later described in the current section.

The algorithms employed in machine learning practices can be described as a computerized learner capable of mapping input variables (**X**) of a given function (**f**) to an output class (**y**) as represented by the relation below:

$$y = f(X) \tag{10}$$

A typical classification problem usually requires the data to be labelled in a binary form. Considering the present use case, the classifier should correctly identify a Packer blockage in the **target area** from all set of events, therefore data labelling was performed by attributing 0 or 1 to the set of events (**y**).

Each occurrence is described by more than one attribute (**X**) as detailed in the following table:

Table 3 | Structure exemplification of the feedstock data.

Label	*Attribute A*	*Attribute B*	*Attribute C*
$\mathbf{0}\ (\boldsymbol{y_1})$	X_{A1}	X_{B1}	X_{C1}
$\mathbf{1}\ (\boldsymbol{y_2})$	X_{A2}	X_{B2}	X_{C2}

Table 3 provides a representation of the input data set for the machine learning model, divided into two groups: Labelled Events (**y**) and correspondent Attributes (**X**). In simple terms, this model is designed for correctly classifying **y**, given the related attributes **X**. The events were binarized considering that the **target stops** are the positive results and the universe of all other stops are the null perspective, as it follows:

Table 4 | Identification of labelled events.

Stop Reason ID	Label	Universe
$\boldsymbol{y = 1000}$	**1**	**Target stops**
$\boldsymbol{y > 0 \ \wedge y \neq 1000}$	**0**	Remaining Packer stops

Taking into account the information stated in Table 4, one may say the defined structure only lacks attributes for relating the labelled events. These attributes were defined by considering that Labelled Events are a result of the past performance of Maker for the reason that, as the current hypothesis suggests, Packer's behaviour is directly influenced by the production of the making machine. Bearing the previous statement in mind and the possible attributes earlier described, the associated variables were determined by the utilization of *Excel* and will be further defined.

Support Matrix served two purposes: avoid noise and redundancy in the inputted data and providing possibilities for generating variables from the attributes already stated. For instance, consider the following declarations:

1. "Maker rejected 1500 Units in the last ten minutes";
2. "Maker rejected 500 Units in the last ten minutes due to *Stop Reason ID -2"*.

Both statements are considered valid and offer vital information to be consumed by the model. For this reason, two additional variables were generated for each Maker *Stop Reason* resultant from the Support Matrix: the number of failures motivated by a *Stop Reason* and the respective number of Units rejected. On the other hand, one may affirm that general information on both machines is as well important and should be taken into account. Therefore, it was considered the following alignment of general variables, given their solid rationalization:

- **Packer – Average speed before failure:** The machine is designed to operate at the constant velocity of 14000 Units per minute. However, given the number of stops in the packaging machine, there are stages of acceleration and deacceleration which signify a great portion of the total operating time. Different stops can also be characterized by different rates of production meaning that this input is valid, as the **target stop** has its own speed profile, essential to the estimator.
- **Packer – Total Product Produced before failure:** In a similar way as the attribute stated above, the Total Product Produced indicates how the respective production was performing by the moment that the failure occurred. Considering the Packer's Labelled Events, two situations can happen: a blockage in the **target area** or any other failure. In this hypothesis, the first situation is likely to be motivated by defective Units produced by the earlier equipment, therefore, there is a higher chance of having a larger number of Packs produced than with other types of stops. For these later items to occur, it is likely that Packer is not performing well, thus accounting with a lower production number.
- **Packer – Rejected Production before failure:** Another general metric that can express if the machine is or not performing as efficiently as the normal operation, is the sum of the rejected number of Units in the packaging machine. This attribute is also justified due to one reason stated earlier: the equipment rejects defective Units until a defined threshold, behind this threshold it causes the machine to stop.
- **Maker – Total number of rejected Units:** Given the location of the Buffer and the level of uncertainty this equipment originates, defining general variables based on Maker's behaviour could turn out to be not effective. For this reason, only the number of rejected cigarettes was considered accurate for the feedstock data.

Note that the last attribute listed is similar to the one mentioned before: the number of Units rejected by each Maker *Stop Reason*. Notwithstanding, the first one provides a general perspective while the last feature is specific for the related Maker failures given by the Support Matrix. This approach ensures, that the relation between Maker and Packer is not independent of the behaviour of the earlier equipment.

4.3.3 Time Dimensionality

One may affirm that in the machine learning environment, the universe of possibilities to address a particular problem is quite significant. However, in the majority of the situations, a fundamental unit is lapsed – the time. Without exception, in every engineering problem, time is a reality that cannot be suppressed and often defines the designed approach to engage it. In some machine learning applications, time is becoming an important variable as well, from medical to cyber-security applications time should be incorporated in the modelling process in order to provide results that are close to reality and to the problem constraints. From the beginning to the end, this variable is an absolute entity, present in the concept of the Sequence Windows, the construction of the Support Matrix and as further addressed in this chapter, in the model development.

The variables presented in the previous chapter doesn't have meaning without the dimension of time as it wouldn't matter much if we provide to the algorithm the value of those variables in the instant of the stop. For this reason, notwithstanding the importance of selecting the variables, is as well essential to reshaping the data to fit in the context of the problem. A constant of time is therefore necessary when defining the variables inputted to the algorithm. This process was iterative, starting with the selection of a fixed time windows and ending with the conception of Lookup Tiers.

At average capacity, Buffer holds 15 minutes of production, meaning that, when leaving Maker, a Unit takes 15 minutes to get to the **target area**. For this reason, the first approach taken was to consider a fixed window of 20 minutes starting in the instant of a Packer stop to calculate the attributes. For example, consider a given failure in packaging machine with label $\boldsymbol{y} = 0$ occurring exactly at 16:00:00. Using the fixed window, the number of failures in Maker motivated by a *Stop Reason* and the respective number of Units rejected will be determined from 15:40:00 until the stop time in the packaging machine. This defined time span was settled in order to ensure that if a given failure occurs in Maker and it produces defective Units, at most, in the 20 minutes later the characteristics should be contained in the variable of the Packer stop.

The results of this approach weren't satisfactory due to the variability of situations that can occur in a 20 minutes window and can affect the performance of the algorithm, as the following examples:

- Packer ceases production and the defective Unit takes longer than 20 minutes to go through the Buffer: In this case, the calculated attributes miss the relevant parameters and provides the algorithm an erroneous description of the events;
- Maker ceases production and the defective Unit takes a shorter time than 20 minutes to go through the Buffer: The packaging machine, as it was presented, accounts with a considerably low MTBF (approximately 6.7 minutes) meaning that, in the defined time span, can occur three different Packer stops. Therefore, the behaviour of the making machine will be present in the attributes of each of these stops, considered as output variables (**y**) to the algorithm.

The current hypothesis has a complex challenge due to the lack of information regarding the Buffer. This equipment performs the transportation of Units and its length is variable depending on the rate of production of both machines. Accordingly, it is not accurate to define a fixed time window and determine the attributes taking that into account. The proposed approach to address this situation is based on the estimation of the Buffer size, aiming to provide a closer contact to the operational reality.

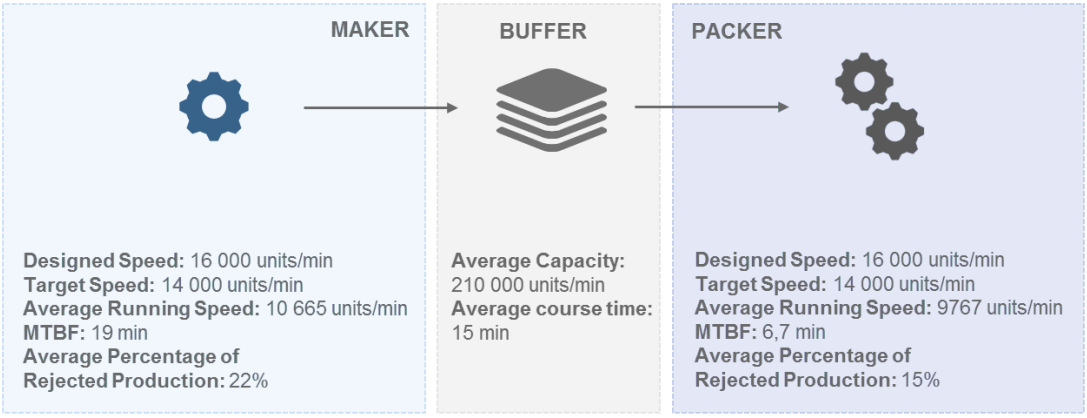

Figure 11 | Characterization of the system and respective equipment related to their capacity, production rate and efficiency.

As previously stated, the exact determination of Buffer size results in non-coherent results due to the asymmetry in time of the information of both machines present in the study. However, instead of estimating this value based on the difference between the Packer consumption rate and Maker Good Production rate, another approach was developed considering the downtime of the equipment. In the case of Maker presenting a higher downtime than the consecutive machine, the Buffer will drain out faster as Packer is consuming more Units than what are being produced by the subsequent machine. Relying on the earlier declaration, the logic of Lookup Tier was built based on the estimation of the Buffer's size.

For each output variable (**y**), in a time span of 20 minutes preceding that stop, the downtime of both machines was determined. Then, the following relation was applied:

$$Downtime\ Packer\ -\ Downtime\ Maker \tag{11}$$

As it is evident, one may declare that higher the calculated difference, the larger is the Buffer size and, therefore, the longer it will take for a Unit to course through this equipment. For this purpose, it was created the concept of Lookup Tier – a define time span that starts in $t + x_s$ minutes from the Packer stop (t) and ends in $t + x_e$.

Bearing this in mind, the conception of tiers of analysis is described by Table 5.

Table 5 | Lookup Tiers resultant from the estimation of the size of the Buffer.

Tier	Downtime Difference (minutes)		Lookup Tier Boundaries (minutes)	
	Minimum	Maximum	Start (x_s)	End (x_e)
1	< -30	- 30	0	5
2	-30	-20	5	10
3	-20	-10	10	15
4	-10	10	15	20
5	10	20	20	30
6	20	30	30	45
7	30	> 30	45	60

The presented tiers are a result of several iterations taken to find the best estimation of this unknown variable. Considering the information stated in the table above, the Lookup Tiers were defined with time spans of 5 to 15 minutes, getting larger as the difference rises, ensuring a wider window in the tiers with greater uncertainty, the ones where the Buffer's size is substantial. The conception of the Lookup Tiers is depicted by Figure 12.

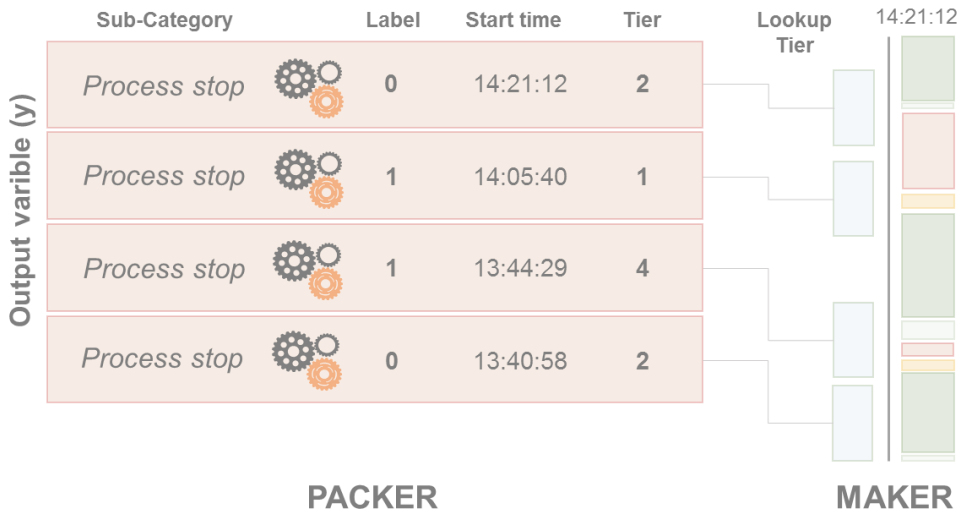

Figure 12 | Conception of Lookup Tiers based on the determination of Downtime Difference for each output variable.

As illustrated in the previous figure, for every output variable (**y**) a Tier is determined given the calculated Downtime Difference – *expression* (11). In each tier, there is a computation of variables taking into account the respective time span. For instance, at 14:21:12 an output variable labelled "0" occurred. The difference in the downtime of packaging and making machines was such that the respective Tier is "2". Therefore, all the variables regarding Maker's behavior will be determined from 14:11:12 until 14:16:12. By conceiving this perspective the time that a defective Unit spent in its path through the Buffer is estimated, aiming to eliminate the influence of this equipment.

On the other hand, the attributes associated with Packer's performance were computed considering the moment that immediately preceded the stop represented by the output variable. The calculations implemented to each variable will be further detailed in the current chapter.

Table 6 | Subset of the input data and auxiliary columns for calculating the Lookup Tiers.

	Auxiliary			***General Attributes***				***Maker Stops Attributes***		
Label	*Aux X*	*Aux Y*	*Aux Z*	*Attribute A*	*Attribute B*	*Attribute C*	*Attribute D*	*Attribute E*	*Attribute F*	***…***
0	8	15	20	4000	1920	40	169	0	0	…
1	-14	10	15	10500	880	0	709	2	48	…
⋮	4	15	20	933	1920	60	95	1	15	…

Table 6 provides an example of the feedstock data with three additional auxiliary columns to produce the attributes. The label on the left represents the set of output variables (**y**) described by the general attributes as well as the specific attributes for the *Stop Reasons* identifying the Maker stops that proceed on the right. As previously stated, the algorithm takes a set of variables as an input and adjusts a function to best classify the outputs. For this reason, it is important to ensure that these variables are the more relevant as possible. The columns that contain this set of attributes are decoded in the following alignment as well as the auxiliary components:

1. **Aux X – Downtime Difference** (minutes)**:**

   Represented by *expression* (*11)*, the downtime difference assists in the definition of the Lookup Tiers for each of the output variables.

2. **Aux Y – Minimum Boundary**(minutes)**:**

   Determined by the respective tier, it defines the start of the time span.

3. **Aux Z – Maximum Boundary** (minutes)**:**

   Determined by the respective tier, it defines the end of the time span.

4. **Attribute A – Average speed** (Units/minute)**:**

   Refers to the average speed of the packaging machine at the moment before the failure. It was considered that whenever the machines initialise the status of *Ramp-down*, a problem in the machine has already been identified, therefore, the average speed selected concerns the speed in the status of *Normal Run*.

5. **Attribute B – Total Product Produced** (Units)**:**

   Indication of the total number of Units consumed by Packer from the moment of the previous failure until the stop identified by the respective output variable.

6. **Attribute C – Rejected Production** (Units)**:**

   Indication of the total number of rejected Units in the packaging machine from the moment of the previous failure until the stop identified by the respective output variable.

7. **Attribute D – Total number of rejected units** (Units)**:**

   Represents the number of Units rejected by Maker in the period comprised by the Lookup Tier associated with the failure of the packing machine.

8. **Attribute E – Number of occurrences by Maker stop:**
   Counts the number of occurrences of a given Maker *Stop Reason* in the period comprised by the Lookup Tier associated with the failure of the packing machine. This attribute is calculated separately for each Maker *Stop Reason* identified by the Support Matrix. The total number of variables is the same as the total number of *Stop Reasons* (11).

9. **Attribute F – Rejected production by Maker stop** (Units)**:**
   Represents the number of Units rejected by Maker in the period comprised by the Lookup Tier associated with the failure in the packing machine. This attribute is calculated separately for each Maker *Stop Reason* identified by the Support Matrix. The total number of variables is the same as the total number of *Stop Reasons* (11).

Two considerations were also applied to the feedstock data aiming to provide to the algorithm a closer perspective of the line's operation. In the first place, it should be considered that a defective Unit could motivate a Packer stop in the **target area**. However, if the irregularity is not very substantial it is possible that it causes another type of blockage downstream of the mentioned area. These blockages were aggregated by the manufacturer following the same line of reason as the one used to combine all *Stop Reasons* associated with the **target stop**. Bearing this in mind, one should consider that Maker is not likely to produce only one defective Unit at a time, but a batch of variable size, ones showing more intensive imperfections than others. For this reason, it is also essential to provide information on this blockages as there could be a potential relation with the **target stops**. In the case of these *Stop Reasons*, two variables were generated per each, conserving the same logic of attributes E and F, excluding in the case the logic of the Lookup Tier, giving that they refer to the packaging machine.

In addition, another action should be performed in the input data regarding the value of the MTBF of the second equipment. As it was stated, Packer has a high number of failures, while a great part is exclusively associated with issues in the machine itself. Therefore, considering that in a given moment it takes place a stop outside the of the area studied ($\boldsymbol{y = 0}$) and 2 minutes later a **target stop** ($\boldsymbol{y = 1}$) occurs.

If we consider the calculation of the variables, it is likely that they present the same set of variables or a really close one, thus inducing the algorithm to make a misclassification of the first stop. As these stops ($\boldsymbol{y = 0}$) are not the goal engaged in the current problem, more importance was provided to the job of limiting the FDR value. This indicator presupposes an outstanding classification of the True Positives ($\boldsymbol{y = 1}$). To address this obstacle, it was excluded from the feedstock data, entries that were not separated by a 15 minutes time window (the average duration of Buffer course).

Accordingly to the best practices defined in the application of artificial intelligence for similar problems, it is essential to standardize the data that is inputted to the algorithm, for the reason that machine learning estimators are likely to be inaccurate if the variables are not standard normally distributed data. In the ideal case, each individual feature should present a Gaussian distribution with zero mean and unit variance. The process of data standardization is made with *Sci-kit learn processing standard scaler*.

4.3.4 Environment Creation

Bearing in mind the definition of a typical machine learning problem, one may declare that two subsets of the feedstock data are required – the training data and the testing data. Both datasets were originated from the same distribution in a random way using the *train_test_split* from *sci-kit learn*, a simple and efficient tool for data mining and data analysis.

While the primary dataset is essential for the algorithm to learn and adjust to the objective function, the second is used to evaluate the performance and tune the model. By looking at the results of implementing the estimator to the testing set, it is possible to adjust its hyper-parameters in order to provide a more accurate performance.

Attention must be taken when splitting the data into several subsets as the content of each one should be representative of the universe. When dealing with a large amount of data, three subsets can be created for training, testing and validating. The last two data sets are both used for quality insurance, however, both are important and should be applied when possible. The testing set not only evaluates estimator's performance but also serves as a reference for tuning the model. On the other hand, validation data is not used to build the model, providing an unbiased sense of model effectiveness [25].

Figure 13 minds the distribution of output variables contained in the splitting of the global universe of data. The gradient of colours depicts the flow of data in the modelling process, starting by the data fed to the algorithm (training data), then the one necessary to test and adjust hyper-parameters (testing data) and later the data used to evaluate the performance of the machine learning model (validating data).

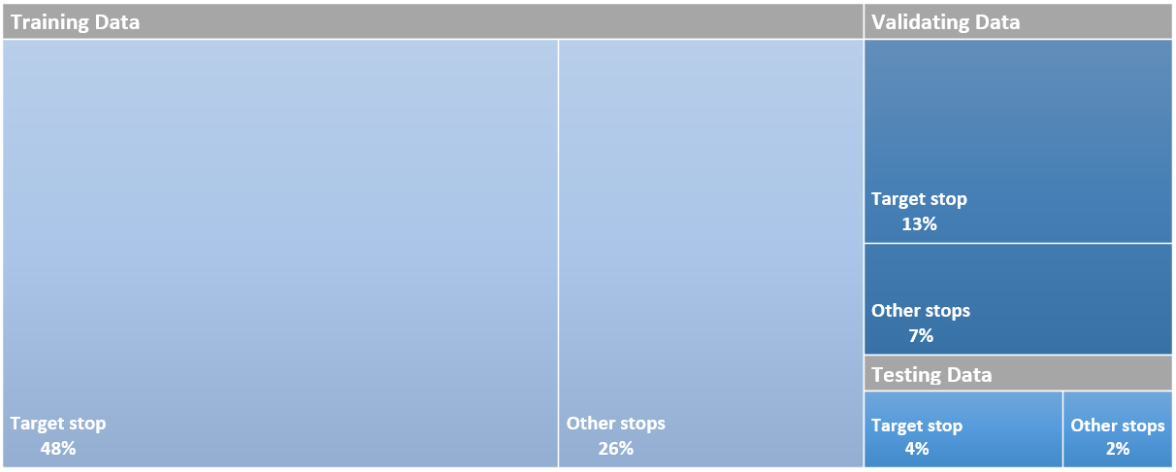

Figure 13 | Distribution of the global data in training, testing and validating data sets.

4.4 Classifier Modelling

Depending on the specific problem, one may affirm that there is usually a right estimator to address its necessities. These estimators are part of a wide library of algorithms and typically suite four purposes: regression, dimensionality reduction, clustering and classification. *Sci-kit learn* provides a decision path in order to define what algorithm is adequate to the problem. Some classifiers such as *SGD* and *kernel approximation* are developed to perform better in the case when the number of samples is greater than one hundred thousand. Below this threshold, other algorithms are frequently used, from *SVC* to *KNeighbors* as well as *Ensemble Classifiers* [26]. The configuration of the decision path followed in the selection of the right classifier goes as follows in the figure below.

Figure 14 | Decision path followed to the selection of the classifier.

As earlier described in the first chapter of the current thesis, *Ensemble Classifiers* holds a group of predictive algorithms developed with the same goal, provide improved generalizability and robustness to the estimator. Present in the context of this group is the *Extremely Randomized Trees* classifier, which is descendant of the widely known category of *Random Forests*. However, in the case of the first classifier (ET), the nodes are split by choosing cut-points fully at random reducing the variance induced in *Random Forests.* In addition, ET uses the all learning sample to grow the trees instead of a bootstrap replica [27].

Table 7 | Parameters defined for the Extremely Randomized Trees classifier.

Parameter	Description	Value
$n\_estimators$	The number of trees in the forest	100
$criterion$	Function measuring the quality of a split	Gini
max_depth	Maximum depth of the trees	100
min_sample_split	Minimum number of samples necessary to split a node	3
$min\_samples\_leaf$	Minimum number of samples in a leaf node	1

The table above provides detail on the parameters tuned for the testing data set, other parameters for ET classifier that are not present in Table 7 were left as default.

Chapter 5

Evaluation of Results

In the previous chapter, it was endorsed the evaluation strategy. Composed by a set of metrics, the approach used considers two different perspectives: Data Science (section 5.1) and Operational (section 5.2). The point of view described in the first concept takes into account the standard evaluation practices in a machine learning problem. On the other hand, the second perspective, groups the results and contextualizes them in the manufacturing industry environment, completing the global evaluation process. An effort was made to offer on both perspectives a detailed description and examination that could allow to entirely understand the relation between the algorithm behaviour and the respective outputs.

It is important to bear in mind that the evaluation results are referred to the employment of the validation data set. This subset of the global data was not used in the construction of the model and, for this reason, is the appropriate structure of information to test the performance of the algorithm.

5.1 Data Science Perspective

Machine learning estimators are often complex and require a great effort in order to understand how the prediction is made and what is behind the applied algorithm. On the other hand, in the Data Science environment, there is a set of performance indicators having a wide range of applications. In the current hypothesis, it is essential to generate insights on the behaviour of the machines and for this reason the metrics chosen to evaluate the performance of the algorithm have to ensure the effective transposition to the operational context.

To fully understand the evaluation strategy, the concept of probability threshold must be taken into account. In a simple binary classification (1 or 0) the classifier estimates the probability of the output variable being one of the classes considering the respective features. By default, if the probability of attributing class 1 is greater than 50 % ($P(\boldsymbol{y} = 1) > 0.5$), then the algorithm associates the output variable to the respective class.

However, it is important to contextualize the performance of the algorithm given its application, because in some cases, the machine learning method could be stricter or softer when attributing a class depending on the threshold, respectively if its value is higher than 0.5 or lower. In most applications, the performance indicators test the algorithm considering a varying threshold from 0 to 1.

5.1.1 Confusion Matrix

The most immediate indicator is, without a doubt, the confusion matrix, as it provides an intuitive visualization of how the algorithm performed. Based on the already stated concepts of TP, FP, TN and FN, the confusion matrix compares the classification done by the algorithm. In other words, the confusion matrix displays the differences between the true and predicted classes. For the reasons mentioned in the previous chapter, the ET should be evaluated by implementing the validation data set, as this structure of information is representative of the universe and it wasn't used to tune the model by adjusting its parameters.

Figure 15 depicts the results of this application to the validation data set and illustrates the performance of the model using a blue colour gradient. In an ideal case, the bottom left and the top right quadrant would be white, meaning a zero value of false positives and false negatives. The following matrix was provided by considering a threshold that maximizes the algorithm accuracy. As previously mentioned the accuracy is the indicator that takes into account all of the components of the confusion matrix, justifying its utilization on the development of this representation.

The defined threshold is represented by the following expression:

$$P(\boldsymbol{y} = 1) > 0.7 \tag{12}$$

Considering this threshold, that makes harder the classification of a **target stop**, the respective accuracy was about 77%.

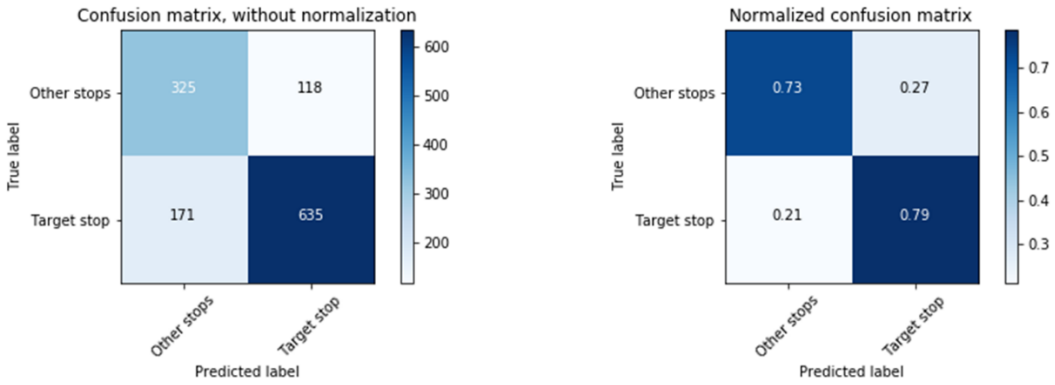

Figure 15 | Normalized and non-normalized confusion matrix referent to the validation data set.

From the above confusion matrix, the indicators referenced in the Evaluation System (section 4.2) given by expressions (5) to (9) were determined and are listed in the table below:

Table 8 | Metrics retrieved from the confusion matrix regarding the class of the target stops.

Metric (Class = 1)	Value (%)
Sensitivity	78.8
Precision	84.3
Accuracy	76.9
F1 Score	81.4
FDR	15.6

Comparing to other cases where machine learning is applying, the values stated above, offers a sense that the model has a considerable positive performance. On the other hand, this statement must be completed by analysing the further indicators.

5.1.2 ROC Curve

The Receiver Operating Characteristic curve provides a visualization of the classifier performance and it is most widely used to select a suitable operating point or decision threshold. The area under the ROC curve (AUC) is frequently used as an indicator of the performance for machine learning algorithms. This area depicts the ability of the algorithm to distinguish each class, thus a higher AUC means a better effectiveness when attributing the respective labels, also known as separability. One of the advantages of using the ROC Curve to evaluate one model's performance is that this indicator does not depend on any decision threshold, comparing the evolution of both TPR and FPR from a range of thresholds between 0 and 1 [28].

In Figure 16, the respective ROC curve is depicted for the algorithm developed. The diagonal line that split the chart represents the curve for a scenario of zero discrimination capacity (AUC of 0.5), meaning providing a reference to visually evaluate the distinguish ability of the algorithm tested.

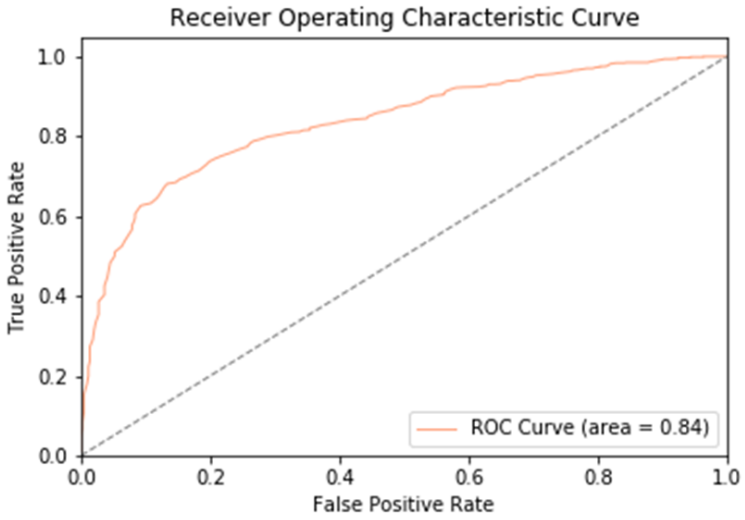

Figure 16 | Representation of ROC Curve and respective AUC to evaluate the performance of the algorithm when addressing the validating data set.

The value of the AUC referring to the model validation was of 0.84, which can be translated as 84% of probability to effectively distinguish between classes. To provide a comparison and contextualize the current value, the range of AUC indicator in the prediction of heart disease data ranges from 70% to 80% [28]. Bearing this in mind and even the shape of the curve, it becomes evident that this indicator references the algorithm with a valuable distinguish capacity.

5.1.3 Precision – Recall Curve

Aiming to evaluate a machine learning algorithm by simply use the accuracy metric can be misleading. A full extent of indicators should be considered in order to test the algorithm robustness under a wider range of circumstances. Precision – Recall Curve is often used to address the influence of imperfections in data such as skew and unbalance. Algorithms that optimize and prosper in the ROC Curve analysis are not guaranteed to perform well under the evaluation of Precision – Recall Curve. On this last indicator, Recall is plotted in *x-axis* while Precision is defined by the *y-axis*. By definition, Recall is the same as TPR. On the other hand, Precision indicated the proportion of positive samples that were correctly labelled [29].

Depicted in Figure 17 is the Precision – Recall Curve (PR Curve) provided for the developed model. Similar to AUC, the average precision indicator (AP) sums up the result of the Curve and thus the indicator performance. AP is a weighted mean of the precisions achieved at each threshold.

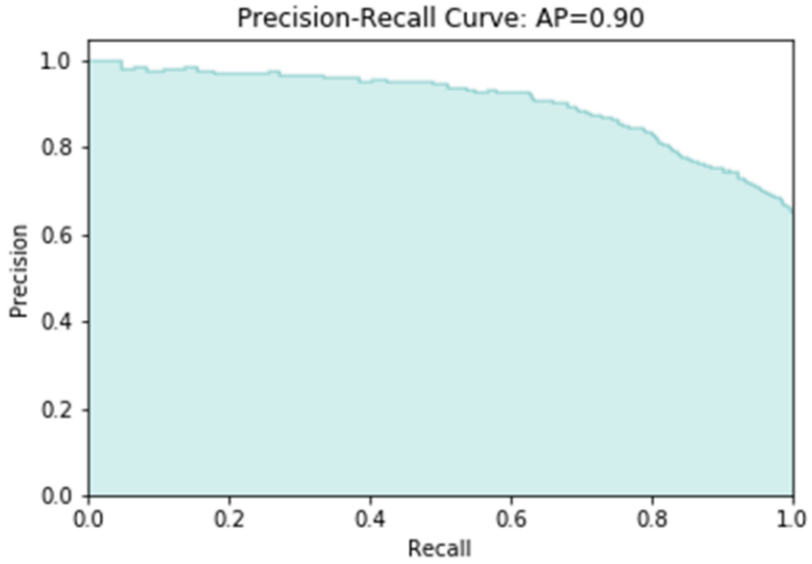

Figure 17 | Representation of Precision – Recall Curve and respective AP to evaluate the performance of the algorithm when addressing the validating data set.

The area under the PR Curve is as well an important representation of the performance of the algorithm, the higher the area, most effective is the classification. However, in PR AUC, the true negatives are not considered, as they are not part of either Precision or Recall.

By interpreting the representation above, it is clear the presence of a substantial area under the PR Curve. This area shows how well the algorithm performed on the validation data set, minding its true positives. In the current thesis, the PR Curve is a fundamental metric as its more important to analyse the effectiveness of the algorithm when correctly classifying the positive class, as it represents the identification of the **target stops**. An average precision (AP) value of 90% detailed in *expression (13)* provides a solid score as this the metric confronts the true positives with the false positives, represented by *expression (6)*.

$$AP = \sum_{n} (R_n - R_{n-1}).P_n \tag{13}$$

5.1.4 TPR – FDR Curve

The evaluating metrics stated above are general and widely applied in the context of machine learning. Nonetheless, it is fundamental to address the specific problem and provide a clear insight from the Data Science perspective on how the model performed in the operational environment. In the manufacturing industry, it is crucial to be effective and simple when addressing a problem.

For this reason, the *False Discovery Rate* is a powerful indicator as it represents the proportion of false positives identified in a real context. Depending on the environment, the number of FDR must be adjusted and optimized to address the problem needs. For instance, having 15 false discoveries (also known as false alarms) in a total of 100 discoveries it might be or not too costly.

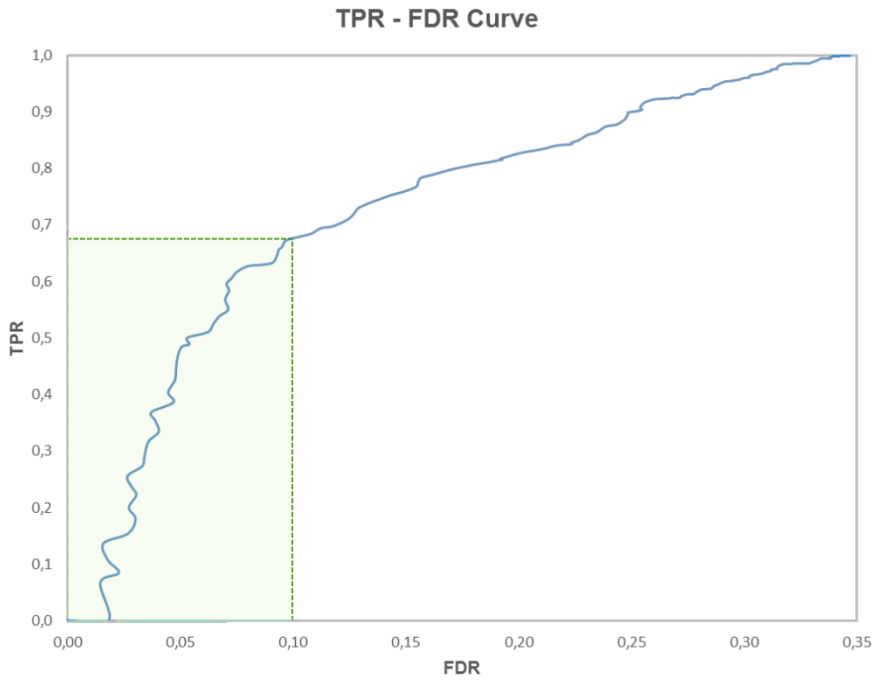

Figure 18 | Representation of TPR – FDR Curve to evaluate the performance of the algorithm when addressing the validating data set. Identification of ideal operating area.

Aiming to analyse the effectiveness of the algorithm in the operational environment, it is essential to bear in mind the TPR – FDR Curve depicted in Figure 18. This curve calculates the respective values of TPR (*y-axis*) and FDR (*x-axis*) for a sequential set of thresholds between 0 and 1. The relevancy of the current analysis minds the fact of having a higher FDR might be costly or implying considerable effort depending on the situation. Therefore, a balance between the two indicators must be taken by identifying the optimal operation point in the TPR – FDR Curve. In the manufacturing industry, this optimal point is often located in the green area illustrated in the representation above. This area comprises all thresholds that offer a maximum FDR value of 10%.

Transposing to the context approached in the present thesis, a value of 10% in FDR consequently refers 68% of TPR. In a real operational environment, an alarm that buzzes every time the algorithm classifies a positive value would have anticipated 68% of all **target stops** providing within the total number of buzzes a 10% rate of false alarms. Bearing this in mind, the TPR – FDR optimal point should be adjusted considering a cost – benefit analysis. In similar problems the FDR is adjusted to be lower than 10%, implying necessarily a penalty in the true positive rate. Comparing the performance of the developed model, with other similar ones in the manufacturing industry the represented curve shows an exceptional performance considering the inputs available and the problem constraints.

5.2 Operational Perspective

The results of any solution are not completely described unless some contextuality and relation to the problem is provided. Bearing in mind the need stated above and the operational metrics earlier describe, the following alignment presents the effects of the algorithm predictions considering a threshold of 0.78 correspondent of a 10% rate of false discoveries.

5.2.1 Reduction of target stops

One of the most immediate results is the number of the **target stops** correctly predicted. Associated with this threshold there is a TPR of 68% meaning the proportion of effective failures classified as true positives. However, it is essential to consider that the prediction is merely indicative providing only the knowledge of when the stop is expected to occur. In order to reduce the blockages in the **target area,** is necessary an action from the operators, which can be not successful in a portion of the times.

The figure below provides an illustration of the application of the developed estimator to the historical data. It was built by considering that 68% of the blockages are correctly predicted in a random way. This estimation was made by randomly discarding the downtime involved in 68% of the **target stops** while excluding from the global downtime the effects of breakdowns, planned stops, and the long process stops.

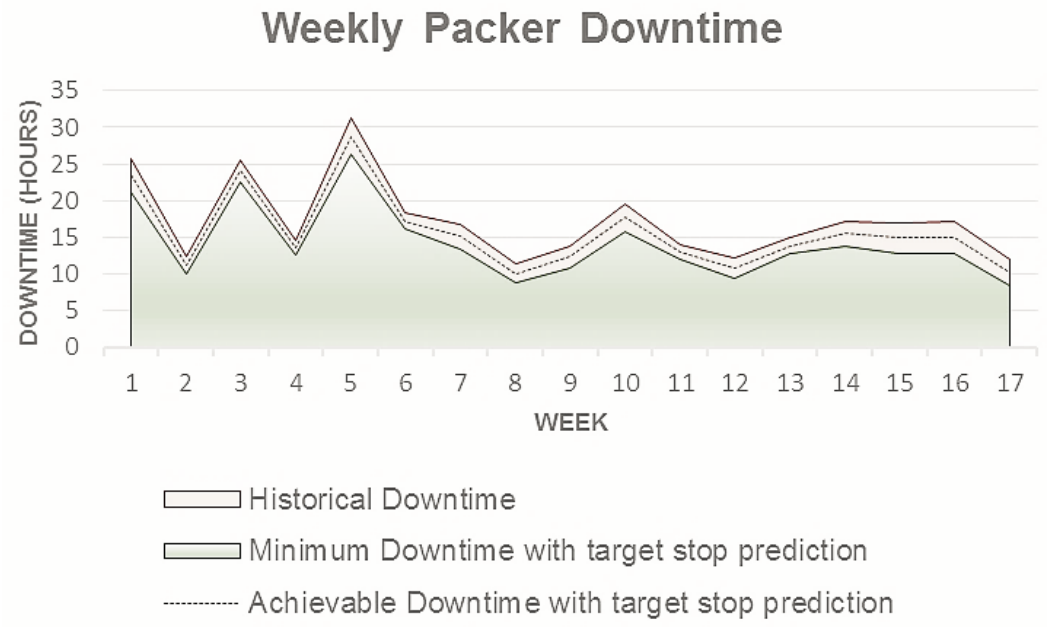

Figure 19 | Illustration of the weekly Packer downtime. Comparison between the historical downtime and the minimum downtime by randomly predicting 68% of the target stops.

However, it should be considered that the algorithm doesn't reduce the downtime by itself as it requires the action of the operators anticipating the blockages. It was considered that in 50% of the cases the experienced personnel would be able to effectively by-pass the problem and it was represented (grey dashed line) the achievable downtime managed with this circumstance.

It is important to retain from Figure 19, that in the first 6 weeks of production the global downtime value was not stabilized and that the model was developed and tuned by with data from all the historical. As not part of the ambit of the current thesis, another iteration should be taken considering only the stable months of production, as they are more representative of the operational contexts and the remain periods could be considered as outliers.

Bearing in the mind the declaration stated above, from the all universe of stops considered in the historical data (total of 11211 stops) it is estimated an achievability of avoiding 3225 **target stops** (applying 50% to the TPR for this class that has a total of 9486 stops in the data set). Therefore, the calculations regarding the achievable MTBF are contemplated in the following alignment and the table summarizing operational data of the packaging machine.

Table 9 | Summary of stats data from Packer in the historical period considered.

Uptime (hours)	1257
Total number of stops	11211
Historical MTBF (minutes)	6.73
Historical target stops	9486
Target stops predicted (threshold = 0.78)	6450
Achievable Uptime	1317
Achievable target stops	3225

10. Historical MTBF:

$$Historical\ MTBF = \frac{Uptime}{Total\ number\ of\ stops} = 6.73\ minutes$$

11. Achievable MTBF:

$$Achievable\ MTBF = \frac{Achievable\ Uptime}{Total\ number\ of\ stops - Achievable\ target\ stops} = 9.90\ minutes$$

12. Increase in MTBF:

$$Increase\ in\ MTBF = \frac{Achievable\ MTBF - Historical\ MTBF}{Historical\ MTBF} = 47\%$$

As previously stated, a balance must be made to find the ideal conditions which represent an optimal cost – benefit proportion aiming to achieve the model's full potential.

Chapter 6

Conclusions

The present thesis studies the behaviour and interdependency existing between to machines in a given operational line of a manufacturer. The line is located in the secondary and last part of the global production process and, for this reason, the accumulation of imperfections and unwanted peculiarities which might have a prejudicial effect on the machines and consequently on the production. It was approached the adoption of machine learning techniques aiming to effectively predict the type of failure with a larger impact on line's global downtime, the **target stops**.

Knowledge of the experienced personnel was an irrefutable reality providing insights on equipment behaviour and its reflection on the data, as well as defining the focus of this study by identifying all the constraints involved. This cooperation was essential to generate the set of features used as input for the machine learning algorithm.

6.1 Discussion

To address the machine learning problem engaged in the current thesis, two main points were defined as targets, which are detailed in the following alignment:

1. Provide insights on machine behaviour that are understandable and can be transposed to operational environment;
2. Predict **target stops** based on the production data of the making machine.

Regarding the first goal stated above, it was attained in the stage of data processing and feature selection by developing algorithms capable of interpret the data and structure the information in a simple and integrated format. Based on statistical theories, it was conceived the Support Matrix tool. Promoted by *Apriori* algorithm and integrated with the logic of the Lookup Windows, this framework provided a set of outputs fundamental to recognize patterns in the operation of the system.

Furthermore, this visualization tool can be adapted to fit many other similar jobs in the manufacturing industry. The assumptions that are embedded in the idealization of the Support Matrix are the existence of a list of transactions (occurrences) related with one specific item and the concept of time to create the list itself. This relation can be expressed by a determined sequence of events or status and will be processed in order to identify patterns and logical progressions. In the use case addressed in this thesis, the support matrix provided the set of Maker *Stop Reasons* with stronger relation to Packer's **target stops** (support above the threshold of 5%). Each stop reason and respective support can be easily interpreted by the experienced personnel, aiming to understand the extent of the relations between Maker's production and Packer's comportment. In addition, the Support Matrix achieved other of the main challenges engaged – the combination of two asymmetric data sets.

The referenced and other challenges accomplished on the study are listed below:

1. Combine asymmetric data sets;
2. Retrieve variables from Maker that are relevant to the classification;
3. Produce a solid range of features from the few available variables;
4. Properly estimate the influence of Buffer by considering its size;
5. Determine the right classifier for an effective prediction of the **target stops**.

The third of the hindrances stated above concerns the fact that the data provided to the algorithm must somehow reflect the interdependency existent in the system. Regarding the presented circumstances, the contribution of the operators and maintenance personnel was essential to define the features that were more representative of the operational reality.

A level of uncertainty was introduced by the lack of information regarding the Buffer existent immediately upstream the packaging machine. This equipment transports the produced Units from Maker to Packer and dilutes the effect of defective Units. As it is unattainable the calculation of Buffer's instant dimensions, it is not possible to determine the amount of time spent by a Unit during this course. To overcome the stated obstacle, the concept of Lookup Tiers was introduced in the data preparation stage. Time is seen as a fundamental dimension to be considered in problems of this nature and, in the current use case, represents a distinctive factor in feature generation. The Lookup Tiers estimate the delayed time of a Unit during the Buffer's course and were used to calculate the variables for each input, considering the difference between the downtime of the packaging and making machines.

Data was randomly split into three subsets, containing 74%, 6% and 20% of the global input data set for training, testing and validating, respectively. Further, an ensemble classifier was selected as an appropriate estimator to handle problems of this category. The classifier that best fitted this purpose was the *Extremely Randomized Trees*, proposed by *Geurts et al* [27].

Testing data was applied to tune the model's hyper-parameters, while the validation data set was used to evaluate its performance. This evaluation was carried out by inducing two perspectives, the Data Science point of view and the Operational context. In the earlier approach, a confusion matrix was represented as well as its dependent indicators were determined. The accuracy for a probabilistic threshold of 0.7 was of 77% while the F1 Score rounded 81%.

Also endorsed in the Data Science perspective the validation data set was tested under the rating of the ROC Curve. The visual aspect of the curve allowed to understand the level of separability of the classifier while the correspondent area under the curve (AUC) reached 84%. Another typical indicator of the performance of an algorithm used in machine learning problems is the Precision – Recall Curve. Similarly to the first curve, the PR representation offers a visual insight of how well the model performed, however, in this case only the effectiveness to the positive class is evaluated, from both Precision and Recall. The Average Precision that followed up this indicator marked the 90%. Both curves characterize the algorithm for its robustness and generalization.

Justified by the context of the problem and the requirements for a low rate of false discoveries, it was depicted in the TPR – FDR Curve. Given its constraints, the aim of this representation was to select an optimal probabilistic threshold that could minimize the FDR while having a high rate of true positives. The threshold of 0.78 allows a FDR lower than 10%, maintaining the TPR at 68%.

From an Operational perspective, the reduction of **target stops** is the central objective. However, it is important to note that the algorithm only indicates the probability of the occurrence of these blockages and could not have a direct impact on machine performance. The aim is to alarm the operator so he can prevent the failure. Therefore, the analysis of the effectiveness of the algorithm in the industrial context could only be attained by assuming that only 50% of the identified stops could be avoided by the operator. The model can successfully process a large amount of information, therefore, the scalability of this approach is considered adequate to the problem needs.

The value of MTBF is one of the most acceptable indicators for measuring machine performance in the manufacturing environment. Considering only the achievable reduction of **target stops** an increase in 47% was determined for the historical MTBF of Packer.

The importance of preparing the data by taking into account the problem context and by receiving insights from the experienced personnel is one of the most considerable outputs regarding the application of machine learning in an operational environment.

6.2 Future Work

The approach followed end-to-end, aimed to provide a complete knowledge of the capabilities of artificial intelligence in the industry context. Nonetheless, enhancements should always be considered.

Considering that the classifier estimative is supported by the data the is used in the learning stage, it is fundamental to provide the most valuable information as possible, maximizing the generalization of the algorithm. Therefore, as part of further developments, data regarding the properties of Maker's production can be a worthwhile asset. This information minds the quality of the Units being produced and serve two essential purposes:

1. Increases the algorithm performance by maximizing the number of effective classifications and reducing the number of false alarms;
2. Allows the identification of the root cause of the **target stops** by understanding which parameters have a higher impact on the product's quality.

Further analysis should be carried in order to identify possible changes that could benefit the work environment. With this aim, it should be considered daily standards such as the operators' location, for providing the best line of sight to the most impacting failures - the blockages in the **target area**.

In the author's perspective, the step further should be taken by integrating online data with the concept of the Internet of Things applied to machine learning. In this reality, the machine data is retrieved in real time and immediately processed by an optimized algorithm. Use cases in the manufacturing industry have successfully implemented online integration of machines using artificial intelligence to predict and anticipate behaviours.

The adoption of the suggested approach and the interpretation of its results are vital to prove the integrity of this work and contributes to a wider knowledge concerning the predictive maintenance groundwork.

References

[1] H. Lasi, P. Fettke, H. G. Kemper, T. Feld, and M. Hoffmann, “Industry 4.0,” *Bus. Inf. Syst. Eng.*, vol. 6, no. 4, pp. 239–242, 2014.

[2] F. Tao, J. Cheng, Q. Qi, M. Zhang, H. Zhang, and F. Sui, “Digital twin-driven product design, manufacturing and service with big data,” *Int. J. Adv. Manuf. Technol.*, vol. 94, no. 9–12, pp. 3563–3576, 2018.

[3] H. Wang, X. Ye, and M. Yin, “Study on Predictive Maintenance Strategy,” *Int. J. u- e- Serv. Sci. Technol.*, vol. 9, no. 4, pp. 295–300, 2016.

[4] M. Ben-Daya, S. O. Duffuaa, J. Knezevic, D. Ait-Kadi, and A. Raouf, “Reliability Centered Maintenance,” in *Handbook of Maintenance Management and Engineering*, 2009, pp. 397–416.

[5] Y. He, C. Gu, Z. Chen, and X. Han, “Integrated predictive maintenance strategy for manufacturing systems by combining quality control and mission reliability analysis,” *Int. J. Prod. Res.*, vol. 55, no. 19, pp. 5841–5862, 2017.

[6] R. K. Mobley, “Impact of Maintenance,” in *An Introduction to Predictive Maintenance (Second Edition)*, 2002, pp. 1–22.

[7] S. M. Rezvanizaniani, J. Dempsey, and J. Lee, “An effective predictive maintenance approach based on historical maintenance data using a probabilistic risk assessment: PHM14 data challenge,” *Int. J. Progn. Heal. Manag.*, vol. 5, no. 2, 2014.

[8] R. C. M. Yam, P. W. Tse, L. Li, and P. Tu, “Intelligent predictive decision support system for condition-based maintenance,” *Int. J. Adv. Manuf. Technol.*, vol. 17, no. 5, pp. 383–391, 2001.

[9] M. W. Gary R. Garrow, Charles P. Newton, III, Patrick E. Weir, P. West II David, “Performing predictive maintenance on equipment - Patent No. US 6,738,748 B2,” 2009.

[10] T. A. Marques, S. T. Buckland, D. L. Borchers, E. Rexstad, and L. Thomas, “International Encyclopedia of Statistical Science- Distance Sampling,” pp. 398–400, 2011.

[11] T. G. Dietterich, “Ensemble Methods in Machine Learning,” in *Proceedings of the First International Workshop on Multiple Classifier Systems*, 2000, pp. 1–15.

[12] L. Swanson, “Linking maintenance strategies to performance,” *Int. J. Prod. Econ.*, vol. 70, no. 3, pp. 237–244, 2001.

[13] J. Nelles, S. Kuz, A. Mertens, and C. M. Schlick, “Human-centered design of assistance systems for production planning and control: The role of the human in Industry 4.0,” *Proc. IEEE Int. Conf. Ind. Technol.*, vol. 2016–May, pp. 2099–2104, 2016.

[14] D. M. Hand and H. Smyth, “Principles of Data Mining,” in *Principles of Data Mining*, vol. 30, no. 7, 2001, pp. 621–622.

[15] R. Vilalta and Sheng Ma, “Predicting rare events in temporal domains,” *2002 IEEE Int. Conf. Data Mining, 2002. Proceedings.*, pp. 474–481, 2002.

[16] J. Nakamura, “Predicting Time to Failure of Industrial Machines with Temporal Data Mining,” *Masters Sci. - Univ. Washingt.*, 2007.

[17] H. Road and S. Jose, “Fast Algorithms for Mining Association Rules,” in *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994, pp. 487–499.

[18] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 1157–1182, 2003.

[19] T. Wuest, D. Weimer, C. Irgens, and K. D. Thoben, “Machine learning in manufacturing: Advantages, challenges, and applications,” *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, 2016.

[20] D. T. Pham and A. A. Afify, “Machine-learning techniques and their applications in manufacturing,” *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.*, vol. 219, no. 5, pp. 395–412, 2005.

[21] C. Scott, G. Bellala, and R. Willett, “The false discovery rate for statistical pattern recognition,” *Electron. J. Stat.*, vol. 3, pp. 651–677, 2009.

[22] F. Pedregosa, R. Weiss, and M. Brucher, “Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2012.

[23] N. Kwak and C. H. Choi, “Input feature selection for classification problems,” *IEEE Trans. Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.

[24] R. G. Casey and G. Nagy, “Developing Classifiers,” in *Advances in Pattern Recognition*, vol. 224, no. 4, 1971, pp. 56–71.

[25] M. Kuhn and K. Johnson, “Over-Fitting and Model Tuning,” in *Applied Predictive Modeling*, 2013, pp. 27–60.

[26] R. Konieczny and R. Idczak, “Supervised Machine Learning: A Review of Classification Techniques,” *Hyperfine Interact.*, vol. 237, no. 1, pp. 1–8, 2016.

[27] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.

[28] B. Lu, Q. Li, and W. Y. Liu, “Dynamic structure of the Sarcin/Ricin domain in rat 28S ribosomal RNA investigated by hybridization with oligodeoxynucleotide,” *Biol. Chem.*, vol. 378, no. 7, pp. 697–699, 1997.

[29] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” *Proc. 23rd Int. Conf. Mach. Learn. - ICML ’06*, pp. 233–240, 2006.