ISCTE IUL
Instituto Universitário de Lisboa

Department of Information Science and Technology

A Predictive Maintenance Approach based in Big Data Analysis

João Pedro Gonçalves da Silva

Dissertation submitted as partial fulfilment of the requirements for the degree of Master's in Business Intelligence

Supervisor:
PhD João Carlos Ferreira, Assistant Professor,
ISCTE – Instituto Universitário de Lisboa (ISCTE-IUL)

October, 2019

“I started from the bottom
Now I’m still at the bottom
Never Rise, I fall
Every Season is Autumn”

Hello Peril, 2019

Acknowledgments

É difícil conseguir expressar em apenas duas páginas o quanto estou agradecido a todas as pessoas que estiveram presentes durantes estes últimos dois anos de mestrado. De certa forma, estas páginas não parecem ser suficientes para atingir o seu propósito.

Antes demais, gostaria de agradecer ao meu Orientador, João Carlos Ferreira, que com esta tese me conseguiu desafiar, permitindo-me assim desenvolver as minhas capacidades e consolidar todo o conhecimento que adquiri ao longo deste mestrado. A sua capacidade crítica e apoio foram peças fulcrais para atingir este resultado.

A toda a equipa do INOV e HFA, um grande obrigado pelo interesse e, acima de tudo, pela vontade demonstrada ao longo deste ano no projeto.

Pai e Mãe, parece que o barco partiu e eu não fiquei no porto a ver navios, hum? É difícil expressar o quanto vos agradeço. Um obrigado por estes dois anos seria insuficiente. Sou-vos grato por tudo, por 23 anos e 352 dias a aturarem-me. Se cheguei aqui hoje foi graças à vossa dedicação e sou capaz de atravessar fogo por vós. Obrigado por todos os conselhos e conhecimento que me passaram e passam diariamente, por todas as vezes em que usaram a vossa "bola de cristal" e, mesmo sem dados ou modelos preditivos, conseguiram adivinhar o futuro e aconselhar o caminho certo a seguir. Obrigado por me terem ensinado a ser como sou hoje e por me terem dado o melhor que dispunham para eu poder atingir sempre os meus objetivos.

Xana, gostaria de escrever em francês, mas o meu domínio de tal dialeto ainda consegue ser inferior ao que possuo de Engenharia Civil. Ainda assim, e cingindo-me aquilo que sei, obrigado pelas conversas às seis da manhã depois das habituais noitadas e por seres a única que compreende verdadeiramente o meu sentido de humor, esse que pode facilmente ser comparado ao de uma criança de cinco anos. Mesmo a mais de 1500 km de distância parece que estás cá comigo todos os dias.

Ricardo, depois disto podes ter a certeza de que vamos jogar um FIFA… até te deixo ganhar. Fora de brincadeiras, como é óbvio isso não vai acontecer. Obrigado por estares sempre presente. Vir para Lisboa com 17 anos foi provavelmente um dos passos mais difíceis da minha vida, e saber que tinha cá alguém com quem podia contar e confiar foi o melhor que podia ter acontecido. Obrigado por tudo, sou um sortudo em ter um Irmão como tu na vida. Não me irei esquecer nos próximos natais e aniversários… Ahah

Á minha avó e tio por todo o apoio e preocupação que tiveram comigo, por todo o interesse e motivação para conseguir dar este passo.

Ao meu primo António, a quem eu devo um profundo obrigado por toda a sua preocupação em eu me tornar a melhor pessoa possível. Ensinaste-me uma das lições mais valiosas que tive até hoje.

Ao Duarte, à Pipas, ao Cruz, ao Batalha, ao Sheng, ao Pedro e ao Renato pela compreensão em todos os momentos em que tive que dizer que não e não consegui estar presente. Sei que foi difícil, mas obrigado pela vossa paciência e por terem manifestado todo interesse para que esta fase corresse o melhor possível.

Ao Diogo, uma das pessoas que me acompanhou ao longo desta tese, *Bakayaro Konoyaro*. Obrigado pela oportunidade de podermos trabalhar juntos, sem sombra de dúvidas foram os trabalhos que mais gostei de fazer até hoje. Obrigado por todas as boleias e motivação que me deste, e por estares sempre presente, principalmente nesta fase final do mestrado. És uma pessoa fantástica. Obrigado por termos partilhado este caminho juntos e espero que venham mais anos a partilhar risos e conversas.

A todos os meus colegas e funcionários do ISCTE-IUL, em particular aos seguranças Rui e Edgar pela paciência que tiveram durante estes anos. Foram pessoas que fiquei a conhecer cada vez melhor em cada noite que passei no ISCTE, uma casa que durante seis anos me acolheu de braços abertos.

À professora Elsa Cardoso, por ter despertado o meu interesse por Sistemas Integrados de Apoio à Decisão e por ter sido um dos principais catalisadores que me fez matricular neste mestrado. Resumidamente, obrigado… “*Drops the mic*”

Por último, Mags isto não é suficiente para te agradecer. Sabes que lá no fundo se fosse a escrever tinha suficiente para fazer um livro. Não é simples dizer isto, mas obrigado. Obrigado por estares sempre presente, nos bons e nos maus momentos, e me motivares sempre a continuar mesmo quando tudo corria mal. Esta tese foi uma perfeita sintonia de altos e baixos que me testou em todos os limites e tu estiveste sempre pronta para me ajudar a ver que isto era o desafio que eu tanto procurava. Tenho imenso respeito por ti, pela tua capacidade de nunca desistir e por quereres sempre o melhor para mim.

A todos vós, porque sem vocês eu não estaria nesta posição. Sem vocês eu não tinha conseguido chegar onde cheguei. Obrigado por tudo o que fizeram por mim.

Resumo

Através da evolução dos sistemas de informação (SI), o fluxo de dados atingiu novos limites, permitindo assim às empresas desenvolver diferentes focos e aplicar novas perspetivas nos departamentos fulcrais à sua atividade, tais como produção, logística e, mais especificamente, a manutenção. Esta última componente evolui paralelamente à indústria, evidenciando novos desenvolvimentos em cada iteração da mesma. Particularmente, a quarta revolução industrial destacou-se pela capacidade de conectar máquinas entre si e pela evolução posterior do processo de extração de dados. Assim, surgiu uma nova perspetiva focada na utilização dos dados extraídos para resolução de problemas. Consequentemente, esta inovação fomentou uma redefinição das prioridades nas decisões tomadas relativas à manutenção, dando primazia à compreensão dos dados gerados. Por conseguinte, a correta elaboração de um plano de implementação de manutenção preditiva (MP) destaca-se como um passo fulcral para as empresas. Este plano tem como objetivo permitir uma abordagem mais segura, possibilitando assim alocar os recursos estrategicamente, reduzindo o risco e potenciando a recompensa. Mediante a análise de múltiplas abordagens de MP, é proposto um modelo genérico que reúne um conjunto diretrizes. Este tem intuito de auxiliar os departamentos de manutenção que pretendem compreender a viabilidade da instalação de uma solução de MP na empresa. A fim de perceber a utilidade dos artefactos desenvolvidos, foi realizada uma aplicação prática do modelo numa pequena e média empresa (PME).

Palavras-chave: Indústria 4.0; Manutenção Preditiva; *Big Data*; *Data Mining;*

Abstract

With the evolution of information systems, the data flow escalated into new boundaries, allowing enterprises to further develop their approach to important sectors, such as production, logistic, IT and especially maintenance. This last field accompanied industry developments hand in hand in each of the four iterations. More specifically, the fourth iteration (Industry 4.0) marked the capability to connect machines and further enhance data extraction, which allowed companies to use a new data-driven approach into their specific problems. Nevertheless, with a wider flow of data being generated, understanding data became a priority for maintenance-related decision-making processes. Therefore, the correct elaboration of a roadmap to apply predictive maintenance (PM) is a key step for companies. A roadmap can allow a safe approach, where resources may be placed strategically with a ratio of low risk and high reward. By analysing multiple approaches to PM, a generic model is proposed, which contains an array of guidelines. This combination aims to assist maintenance departments that wish to understand the feasibility of implementing a predictive maintenance solution in their company. To understand the utility of the developed artefact, a practical application was conducted to a production line of HFA, a Portuguese Small and Medium Enterprise.

Keywords: Industry 4.0; Predictive Maintenance; Big Data; Data Mining;

Index

Table of Figures

Table of Tables

Acronyms

AOI - Automated Optical Inspection

BT– Bayesian Networks

BTA – Boosting Tree Algorithm

CART – Standard Classification and Regression Tree

CMT – Condition Monitoring Tool

CPS – Cyber-Physical Systems

CRISP–DM – Cross-Industry Standard Process for Data Mining

DB – Database

DT– Decision Trees

FIA – Fédération Internationale de l'Automobile

FL – Fuzzy Logic

HFA – Henrique Fernando & Alves

HTPR – Highest True Positive Rate

IoT – Internet of Things

LFPR – Lowest False Positive Rate

MNE – Multi National Enterprises

MRMR – Minimum Redundancy Maximum Relevance

NB – Naïve Bayes

NN – Neural Network

NNE – Neural Network Ensemble

OEE – Original Equipment Effectiveness

PM – Predictive Maintenance

P&P – Pick and Place

PCB – Printed Circuit Boards

RCF – Recolha Chão Fábrica

RPM – Rotations per minute

SLM – Selective Laser Melting

SME – Small and Medium-sized Enterprise

SMT – Surface-Mount Technology

SVM – Support Vector Machine

SPI – Solder Paste Inspections

TDT – Televisão Digital Terrestre

USA – United States of America

1 Introduction

This first chapter aims to elucidate the problem and motivation that allowed the development of this dissertation. Consequently, the goals of this work are defined and what contributions are expected to be delivered upon the scientific community regarding the matter in hand. Finally, the outline methodology of this thesis is introduced and described.

1.1 Research Problem and Motivation

Throughout this past couple of years, there has been an immersive revolution in the digitalized world we live in (Marr, 2016). New technologies can perform in standards that were yet to be conceived a decade ago. The performance of processing and storage components breached barriers and still manages to break new ones, upon taking into consideration that perfection is just a term that does not manage to uphold any consideration regarding new solutions. The reason for this statement is mainly due to the fact that new solutions in processing components released today may be considered outdated in time-lapse of two to three years[1] (Manyika et al., 2011). This line of thinking generated shock in multiple areas, being one of them, Industry.

Currently, data is everywhere. “According to IBM, we create around 2.5 quintillion bytes of data on a daily basis, and it is estimated that the size of data will double every 2 years” (Lau, Zhao, Chen, & Guo, 2016). Taking into consideration that in the year 2000 only a quarter of all world’s stored information was digital, it can be stated that currently “the amount of data available is literally exploding” (Marr, 2016, p. 2), since in modern days only less than two per cent of all stored information is nondigital (Cukier & Mayer-Schoenberger, 2013).´

Companies face a fierce market, regardless of the sector of the industry they are positioned. Despite a boost in information systems, companies still face a problem regarding the data generated. Disregarding the fact that more and more information is being collected, the consequent analysis and storage are not performed accordingly which leads to the increased difficulty to extract knowledge (Efthymiou, Papakostas, Mourtzis, & Chryssolouris, 2012). Consequently, this issue generates shockwaves to each department of an enterprise, such as

[1] Moore’s Law, by Gordon Moore, states the number of transistors that can be placed on an integrated circuit doubles approximately every two years. In other words, the amount of computing power that can be purchased for the same amount of money doubles every two years.

production, logistic, maintenance amongst others. Therefore, a change in focus is desperately needed in order to allow more data-driven solutions to aid the process of decision-making.

This vast generation of data changed the mindset of industries to focus on a more data-driven perspective, mostly due to the shockwaves of the fourth industrial revolution. Industry 4.0 can be considered as “an umbrella term made up of the tools which form its structure”(Chesworth, 2018, p. 1). In total three tools underline this umbrella, Cyber-Physical Systems (CPS), Big Data and Internet of Things (IoT). The collaboration of these components triggered specific outputs to cope with this data-driven approach to industry, one of them being PM (Meissner, Ilsen, & Aurich, 2017).

Throughout time maintenance, practices were an outcome of each industrial revolution. They allowed companies to adapt their methodologies to the new technologies/concepts generated. Nevertheless, only in this revolution, industry was given the possibility to further understand equipment operation and how to fine-tune it (Coleman, Damofaran, & Deuel, 2017). By applying the foundations of industry 4.0 such as Data Mining and IoT, users in different sectors of industry broaden their flow of data and therefore the quality of information gathered for the process of decision making.

1.2 Research Aim

Being PM a direct output from the last industrial revolution, it affected directly all sectors of industry, therefore, constituting a goal to achieve for all companies regardless of the sector. Nevertheless, the maturity stage of each sector varies, which leads to limitations for the implementation of this breakthrough. For instance, sectors where automation cannot be implemented such as metal casting face tremendous difficulties regarding the implementation of this data-focused methodology, mostly due to the fact that it resorts in its majority to manual labour.

With this knowledge in hand, the main goal of this thesis is to produce a predictive maintenance model and a set of guidelines that can simplify the process of implementing this methodology regarding the prediction of unscheduled stops in equipment.

Since the application of this methodology can be morphed into different scenarios, for this dissertation, the focus shall remain on establishing a common ground regarding the steps to be performed in the process of implementation. Therefore, the model to be generated must take as basis a set of cases from different scenarios, which consequently allows to produce a more robust model. Since PM relies on the combination of several concepts, a set of guidelines were created to ease the process of this step-by-step execution. These guidelines focus on the relevant steps of the model to which additional attention is required to ensure proper employment.

Because PM is a direct output of Industry 4.0 and applies its foundations, it is still not enough to establish a mandatory presence of these concepts. As established previously, certain sectors of industry may have limitations regarding the transition into this digitalized era. Thus, a possibility presents itself, in the capability of foundations of industry 4.0 not being utterly defined, is it still possible to proceed to the implementation of its direct output?

Overall, this thesis aims to answer the following questions:

- Can Predictive Maintenance have a leading role, despite the main foundations of Industry 4.0 being not being established?
- Is it possible to achieve a generic model for the implementation of Predictive Maintenance in an Industry 4.0 era?

1.3 Research Contribution

Maintenance methodologies have been a direct output in all industrial revolutions allowing companies to adapt to each different face industry took throughout time. In each iteration, it was possible to perceive a different path being taken, which led to the new methodology overlapping the previous one. This evolution throughout time spiralled into a data-driven approach which led companies to focus their resources in data analysis. In the latest industrial revolution, as stated previously, predictive maintenance was regarded as a direct output since it allows companies to understand their equipment through the analysis of the data generated. Since companies adopt a mindset which emphases results, whether these are production-related, sales-related or even both, trespassing the barrier into the “new” industrial front-line is perceived as a daring move. Additionally, the wide variety of sectors available and fierce market competition leads to knowledge being maintained within area being rated a highly valuable attribute. All this leads to a delay in the adaptation of the main pillars of Industry 4.0 and consequently its direct output.

Taking into consideration the relevance of this event in the industry and the role this new data-approach can have in modern industry this aims to contribute with:

- Encourage the adoption of PM methodology through the elaboration of a generalized model to aid the application.
- Simplify the implementation of Predictive Maintenance through a set of guidelines that complement the model generated.

As a result of this thesis a Conference Proceedings named “The aftermath of Industry 4.0 in Small and Medium Enterprises” was created and presented at INTERACT 2019. The aim of this paper is to analyse the effects of the fourth industrial revolution in the Small and Medium Enterprises (J. Silva, Ferreira, & Gonçalves, 2019).

1.4 Research Methodology

In order to understand and address the research questions previously stated in chapter 1.2, the methodology Design Science Research, proposed by Ken Peffers was implemented (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). “Design science…creates and evaluates IT artefacts intended to solve identified organizational problems”. (Hevner, March, Park, & Ram, 2004, p. 77). This methodology is divided into six steps that were applied in the following order:

- **Identification and Motivation** – Previously developed in section 1.1, this step aims at describing the relevance of the data-driven approach and the effect induced by with its introduction through Industry 4.0. To understand how this latest industrial revolution impacted industry PM was identified as a highly relevant output for analysis
- **The Objective of the Solution** – Defined in section 1.2, the main goal of this thesis is to produce a PM model and a set of guidelines that can simplify the process of implementing this methodology regarding the prediction of unscheduled stops in equipment.
- **Design and Development** – In order to understand and properly develop the artefact, foundations are needed to create a stable basis for creation. Therefore, in chapter 3, five Cases from different sectors were analysed, which served as foundations for the development of a model and a set of guidelines in chapter 4.
- **Demonstration and Evaluation** – With the sole purpose of demonstrating the utility of the artefact created, chapter 5 aims to show its implementation in a practical experiment conducted in HFA, a Portuguese SME that produces PCB boards.

- **Communication** – Last but not least, in chapter six, the conclusions obtained are discussed. More specifically, the main knowledge obtained in the process of constructing this dissertation and the future work to be advised is presented.

2 State of the art

This chapter provides a proper theoretical placement regarding the problems and objectives previously exposed. The first section aims to clarify the concept of Industry 4.0 and its main fundamentals, CPS, IoT and Big Data. The second concept to be analysed regards the maintenance activity, more specifically, predictive maintenance, conducted by several companies throughout the world. Last but not least, a review of Data Mining is presented to provide a linkage to the concepts previously reviewed.

2.1 Industry 4.0

With a high volatility market constantly changing its regulations and conditions for operability, companies face new challenges that automatically discard solutions previously conceived and implemented (Kampker, Heimes, Bührer, Lienemann, & Krotil, 2018; Schuh & Blum, 2016). Furthermore, product characteristics and lifecycle are changing by each generation manufactured, which leads to solutions inputted to the market that are far more detail specific. This chain of events is mostly due to the fact that the customer is demanding more and more customization (Hees et al., 2017). The only way for companies to address this encounter is to offer its customers with a vast degree of possibilities. On the one hand, this approach allows increased visibility in terms of marketing, since the customer holds power to customize the desired product. On the other hand, the same approach "forces" the company to change their production processes to small-batch manufacturing (Meissner et al., 2017; Wadhwa, 2012). To ensure a proper result upon facing multiple distinct orders from clients, batch production, induced by product volatility, can be seen as an advantage in a competitive point of view, due to the fact that it requires coordination as a vital foundation. To guarantee that the basic foundations are well established and maintained, companies keep updating their systems and production lines, therefore pledging for a viable position in their respective market (Hees et al., 2017; Mehrabi, Ulsoy, Koren, & Heytler, 2002).

With the sole purpose of facing the challenges stated above a new industrial revolution takes place, Industry 4.0. With multiple definitions emerging, this concept is characterized by Meissner as "conjunction of many technologies -both existing and new - which now work together."(Meissner et al., 2017, p. 165). In another point of view, Özbebek Tunç and Aslan (2020) consider that Industry 4.0 could be defined "by the fact that machines become intelligent and coordinate self-managed production facilities through the internet and inter-machine communication" (Özbebek Tunç & Aslan, 2019, p. 94). For the purpose of this dissertation, the definition proposed by Chesworth was adopted, which considers that Industry 4.0 is the joint effect of CPS's, IoT, Big Data, therefore, creating a decentralized control and advanced connectivity (Chesworth, 2018). Taking into consideration the relevance of these three pillars, an analysis is conducted in the following sections.

2.1.1 Cyber-Physical Systems

A CPS network is a major component regarding the generation and communication of data. It allows hardware to understand the data being generated. In other words, these systems, provided with machine learning algorithms and historical data have the capability to understand and react to possible scenarios due to past behaviours and therefore adapt to situations without human interference (Chesworth, 2018; Jones, Romero, & Wuest, 2018; Rojko, 2017). One of the major advantages of these systems relies on the fact that all data generated is processed and stored, creating a digital library for future use. Applying this methodology to a production line generates a network of machines communicating past experiences, allowing decentralization of power where "machines, products or other elements in the production system can make decisions on their own without any superior control unit." (Meissner et al., 2017, p. 167). This approach is best suitable in this specific scenario where industry is facing the modern problem of great variety through batch production. However, the major disadvantage seen in this scenario regards the possibility of each node of network pursuing individual goals, consequently exposing the structure itself (Chesworth, 2018).

To implement the concept of CPS, machines and the surrounding environment are required to gain further abilities. In other words, it is mandatory to assess the notion of sharing data and knowledge generated within the network itself (Hänel & Felden, 2017). This ability once implemented in equipment restructures the concept of centralized control into a vast network where data is exchanged in real-time "with the goal of identifying, locating, tracking, monitoring and optimizing the production process." (Rojko, 2017, p. 77). Nevertheless, implementing sensors cannot be perceived in the simple idea of generating and collecting data from equipment and environment. This implementation must be seen in the perspective of allowing a two-way communication where machines can both receive and send data which increases the difficulty, since "Today, only some of a manufacturer's sensors and machines are networked and make use of embedded computing" (Rüßmann et al., 2015, p. 6). Consequently, this important factor leads to high dependency towards the remaining pillars of Industry 4.0 (IoT and Big Data), to ensure proper implementation.

2.1.2 Internet of Things

Data extraction was considered as a difficult task mostly due to incapability to extract most of what was desired. To fill this gap, the urge of sensors has been a major role increasingly allowing more data available. With this kind of device, machines can harvest data regarding their physical operations and their neighbouring conditions that allow to understand how different situations unveil regarding machine behaviour (Coleman et al., 2017; Meissner et al., 2017). The data from the devices in question may vary from temperature, power consumption, vibration, amongst others, relying on the scenario of implementation and the weight added to each (Lee, Ardakani, Yang, & Bagheri, 2015). Consequently, the immense quantity of data generated through the sensors is streamed into servers to be properly processed, stored and analysed (Chesworth, 2018; Lee et al., 2015). The analysis is an important step allowing "to increase the knowledge about the use, reliability and efficiency of the equipment/process" (Ferreiro, Konde, Fernández, & Prado, 2016, p. 2), which was previously unavailable.

Impact of IoT can be considered as relevant perspective in terms of its adoption, more specifically in an industry, where it potentially generates "an economic impact of $1.2 trillion to $3.7 trillion per year." (Manyika et al., 2015, p. 8). Companies gain multiple benefits from the implementation of this concept in their production lines, more specifically the creation of knowledge through pattern discovery. One example of this situation was presented by Lee et al. (2015) where considers the "hidden degradation or inefficiency patterns within machines or manufacturing processes can lead to informed and effective maintenance decisions which can avoid costly failures and unplanned downtime." (Lee et al., 2015).

Considering implementation without providing a solid foundation risks the idea itself, which generates major holdbacks in its execution. In other words, the generated amount of data from this principle affects surrounding infrastructures and departments, which combined "with the affordability of bandwidth and storage, massive amount of data can be transmitted to give not only a full picture of assets in a single plant, but of an entire production network." (Coleman et al., 2017, p. 4).

Regarding its application, in 1992 the price for a sensor was approximately 22 US dollars, however, the evolution led this price to lower to 1.40 US Dollars in 2014. In contrast, the usage of these types of equipment followed an inverse tendency, therefore, increasing the range of options available to cover the needs of the market. Table 1 shows the multiple options of types of sensors available, depending on the type of data to extract. (Holdowsky, Mahto, Raynor, & Cotteleer, 2014).

Table 1 - Sensor types (Holdowsky et al., 2014)

Sensor type	Sensor description
Acoustic	Acoustic sensors measure sound levels and convert that information into digital or analogue data signals.
Biosensors	Biosensors detect various biological elements such as organisms, tissues, cells, enzymes, antibodies, and nucleic acids.
Chemical	Chemical sensors measure the concentration of chemicals in a system. When subjected to a mix of chemicals, chemical sensors are typically selected for a target type of chemical (for example, a CO2 sensor senses only carbon dioxide).
Flow	Flow sensors detect the rate of fluid flow. They measure the volume (mass flow) or rate (flow velocity) of fluid that has passed through a system in a given period of time.
Force	Force sensors detect whether a physical force is applied and whether the magnitude of the force is beyond a threshold.
Humidity	Humidity sensors detect humidity (amount of water vapour) in the air or a mass. Humidity levels can be measured in various ways: absolute humidity, relative humidity, mass ratio, and so on.
Light	Light sensors detect the presence of light (visible or invisible).
Occupancy and Motion	Occupancy sensors detect the presence of people and animals in a surveillance area, while motion sensors detect movement of people and objects. The difference between the two is that occupancy sensors generate a signal even when a person is stationary, while a motion sensor does not.
Position	A position sensor measures the position of an object; the position measurement can be either in absolute terms (absolute position sensor) or in relative terms (displacement sensor). Position sensors can be linear, angular, or multi-axis.
Pressure	Pressure sensors are related to force sensors and measure the force applied by liquids or gases. The pressure is measured in terms of force per unit area.
Radiation	Radiation sensors detect radiations in the environment. Radiation can be sensed by scintillating or ionization detection.
Temperature	Temperature sensors measure the amount of heat or cold that is present in a system. They can be broad of two types: contact and non-contact. Contact temperature sensors need to be in physical contact with the object being sensed. Non-contact sensors do not need physical contact, as they measure temperature through convection and radiation.
Velocity and acceleration	Velocity (speed of motion) sensors may be linear or angular, indicating how fast an object moves along a straight line or how fast it rotates. Acceleration sensors measure changes in velocity.

An example of a company that correctly implemented IoT to the purpose of improving operations was Rolls-Royce. Besides producing high-end luxurious vehicles, Rolls-Royce is well known worldwide regarding its aeroplane engines and propulsion systems. This company implemented Big Data in three major areas of operations: design, manufacture and after-sales support, yet after-sales support is the one to be highlighted regarding operations. Its engines and propulsion systems were inbuilt with numerous sensors with the main goal to gather data concerning multiple aspects of the product's performance. In case any change is registered then it is reported *in loco* to engineers so that a decision can be made. This introduction of IoT to the company allowed Rolls-Royce to offer a new service model to clients called "Total Care". Basically, provides support to customers per hour of use of its engines (Marr, 2016).

As expected, the harvest and storage of these mass amounts of data introduce the last key component, Big Data. The relation between IoT and Big Data is easily perceivable since in industry the second one is a consequence of the first.

2.1.3 Big Data

The huge amount of data generated on a daily basis led to the creation of the popular term in current days, Big Data. But the question remains, what is Big Data? Like the other concepts, this one has multiple definitions. Kimball and Ross (2013) consider that "Big data is structured, semi-structured, unstructured, and raw data in many different formats" (Kimball & Ross, 2013, p. 527). According to Manyika et al. (2011) "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse" (Manyika et al., 2011, p. 1). For the purpose of this dissertation and to simplify the diversity of definitions, Big Data shall be defined as provided by Kimball and Ross. But what size must a dataset be in order to be considered as Big Data? The answer is very subjective since there is no specific size defined and most part is not possible to be analysed via SQL (Kimball & Ross, 2013). Hence Manyika's statement where "we don't define big data in terms of being larger than a certain number of terabytes" (Manyika et al., 2011, p. 1). Just like technology keeps progressing over time, the same principle is applied to the definition of Big Data (Lau et al., 2016). Most of the definitions give an enormous focus regarding the size of data. Of course, volume is an important attribute, yet it's not the only one, velocity and variety are also important to the definition (Russom, 2011). Based on this, big data can be characterized in three V's: (1) Volume, (2) Velocity and (3) Variety, as shown in Figure 1.

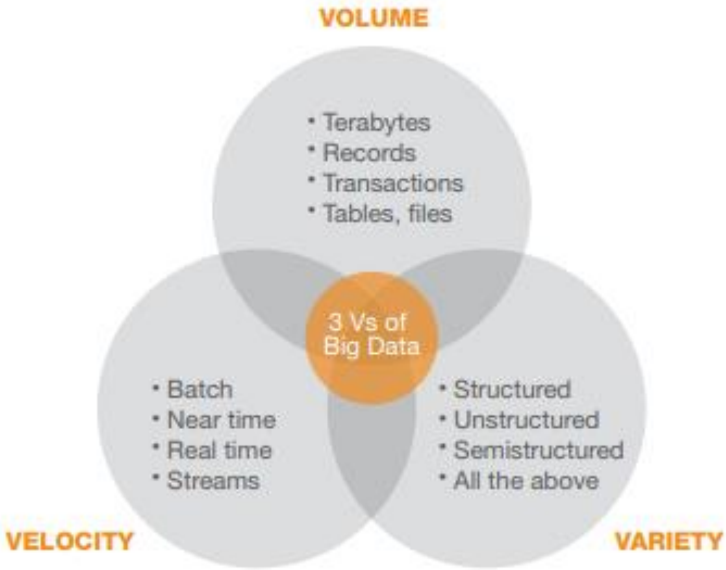

Figure 1 - The three V's of Big Data (Russom, 2011)

The first attribute, **(1) Volume**, refers to the magnitude of data, since "most people define big data in terabytes – sometimes in petabytes" (Russom, 2011, p. 6), but big data can also be defined by counting records, tables, files or even time. Yet, a big volume for a business might not be as big at all in another domain. Defining a specific value for volume in general in big data is not possible, so the rule to have in consideration should be a "rule of thumb" where "when the size of data grows to the extent such that current technologies find it difficult to cope with the storage, retrieval, analysis, and use of the data, we can consider such a volume to be cast as big" (Lau et al., 2016). For example, Wal-Mart was estimated to produce 2.5 petabytes of data in every hour of functioning (Lau et al., 2016).

Data itself can be generated in numerous ways. Traditional analytics methods are formulated to be executed in structured or semi-structured data. Currently, we cannot count with only these two types of data. That is where big data makes the difference regarding **(2) Variety**, with the capability to analyse the same types plus unstructured data. Lau et al. (2016) defines variety as "the inherent heterogeneities of the structures, formats and sources of data."(Lau et al., 2016). For example, User Generated Contents (UGC), like videos, images and audios that are usually seen in social media websites are unstructured data. In 2016 over twenty billion UGC had been uploaded to Instagram. This amount of content that is unstructured can be a huge difficulty for traditional methods to store and analyse (Manyika et al., 2011)

The term **(3) Velocity** regards "the rate at which data are produced and yet the speed at which data should be retrieved and analysed by a big data analytics system" (Lau et al., 2016). It can also be referred to as "the frequency of data generation or the frequency of data delivery"(Russom, 2011, p. 7). From the three V's, velocity is the one that most relates to IoT, since the wide use of smartphones and gadgets with inbuilt sensors led to the increasing rate of data generated. On a basis it is not only associated with IoT, but traditional businesses are also still able to produce huge amounts of data at an incredible rate, such as Wal-Mart (Manyika et al., 2011).

Although, the three V approach is fairly structured nowadays, there are several authors that consider that two more V's should be applied to Big Data, which are: **Veracity** and **Value**. The term Veracity translates to "varying quality and validity of data among a sheer volume of data items" (Lau et al., 2016). Whatever is the business data quality being a major concern. The term Value is defined as "the relatively low value density of a sheer volume of data" (Lau et al., 2016). Having enormous quantities of data is useless if it can't provide proper knowledge of the business. For instance, a huge volume of user log data in electronic commerce website has very low business value itself. Without a proper analysis of the data available, valuable information cannot be obtained, therefore companies should always take into consideration this factor (Manyika et al., 2011).

In conclusion, with the introduction of CPS and IoT in the industry, the amount, variety and speed of data generated heightened into new boundaries. However, methodologies and practices must be implemented to extract viable knowledge from data itself, since "the act of data collection alone does not offer advantages. Instead, to make use of the data and generate additional value, it is important to analyse and process the data in real-time and separate unimportant and important information." (Meissner et al., 2017, p. 166). Multiple cases have been reported in research studies for the joint implementation of Big Data and IoT. One solid example of this application was performed in *Fédération Internationale de l'Automobile* (FIA) Formula 1, a worldwide competition where speed is imperative. In Formula 1, telemetry is used to stream live data from the car to the engineers in the pitlane. It is used by all teams with sole purpose of improving performance on track. Lotus implemented Big Data to the analysis of telemetry since the team was "collecting and analysing a lot of data. We're not talking gigabits or terabytes but petabytes." (Marr, 2016, p. 46).

2.2 Predictive Maintenance

With the amount of data generated, companies must be able to extract viable knowledge. Multiple sectors gain crucial support from these analyses, in particular, maintenance. Maintenance consists of restoring a machine or system into their Original Equipment

Effectiveness (OEE) or the desired status (Park, Moon, Do, & Bae, 2016). Nevertheless, times changed, and maintenance strategies became more complex and increasingly relevant (Fluke Corporation, 2005). According to Efthymiou et al. (2012) "manufacturing systems maintenance is becoming increasingly important, since in many industrial plants, the maintenance costs often exceed 30% of the operating costs and in the context of manufacturing systems lifecycle, maintenance and support, account for as much as 60 to 75% of total lifecycle costs" (Efthymiou et al., 2012, p. 221). When considering these values in terms of budget for companies, maintenance methodologies are taken under special consideration. With the analysis of historical data presented in Table 2 that states the costs incurred for traditional maintenance in the United States of America (USA), it is possible to perceive an upwards trend (Fluke Corporation, 2005).

Table 2 - Traditional Maintenance costs in USA

Year	Maintenance costs ($)
1981	600 Billion $
1991	800 Billion $
2000	1.2 Trillion $

Coleman et al. (2017) state that "poor maintenance strategies can reduce a plant's overall productive capacity between 5 and 20%." (Colemen, Damodaran, ChandramoulinMahesh, & Deuel, 2017, p. 2). Companies resorting to traditional methodologies, such as reactive maintenance or planned maintenance, enter a limbo. On one side of this limbo, they choose to achieve best possible up-time available, disregarding machine life. On the other side, they can address possible scenarios before they happen, therefore not attaining the maximum durability of parts (Qinming, Ming, Chen, Wenyuan, & Chunming, 2018). Each side beholds different advantages and disadvantages, which justify the decision taken by each company. Table 3 shows the differences between these two types of maintenance.

Table 3 - Reactive maintenance vs planned maintenance (Colemen et al., 2017)

	Benefits	**Challenges**
Reactive maintenance	• Maximum utilization of tooling or machine components.	• Potentially greater damage to the machine; • Unplanned downtime; • Higher maintenance costs.
Planned Maintenance	• Less likelihood of broken machinery; • Less unplanned downtime; • More cost-effective than reactive.	• Increased replacement costs over time; • Need for additional spare parts inventory; • Increased planned downtime.

Currently, with spreading of technological instruments to generate and send data, brought by IoT, industry decided to introduce to its methodologies comprehensive maintenance service. The reason for this bold action relies on the fact that users want to ensure high effectiveness and diminish downtimes, which could proportionate considerable losses to their business (Ferreiro et al., 2016; Fluke Corporation, 2005). The re-emergence of predictive maintenance is due to the fact that “only recently have technologies become both seemingly capable and inexpensive enough to make PdM widely accessible.” (Colemen et al., 2017, p. 7). Although PM suffered multiple mutations regarding its definition throughout its lifespan, it is important to clarify its meaning. For Sakib and Wuest (2018) PM aims “to measure and record physical parameters continuously for analysing and comparing data to make maintenance decisions.” (Sakib & Wuest, 2018, p. 268). In other words, and for the purpose of this dissertation PM consists in a maintenance methodology with the sole purpose of knowing in advance actions that can be taken in order to prevent likely scenarios, where machine down-time is expected, and therefore maintaining the desired OEE (Colemen et al., 2017; Fluke Corporation, 2005). Yet, PM doesn’t solely consist of predicting possible scenarios based on historical data. According to many authors, the direct monitoring is also an important part, since the implementation of this “philosophy” can aid companies to optimize total plant operation (Fluke Corporation, 2005; Park et al., 2016; Sakib & Wuest, 2018).

The implementation of PM allows to reach a common ground between the Reactive and Planned Maintenance, obtaining the advantages of both while not compromising the integrity of the assets with possible down-time due to malfunction. Regarding its advantages, from literature analysis, it is possible to state several factors to which companies may take advantage, upon the correct implementation of this methodology, as follows (Colemen et al., 2017; Fluke Corporation, 2005):

- Increased useful life (33-60%);
- The decrease in maintenance expenditures (10-15%);
- Increase in sustainable capacity (15-40%);
- Reduce the time required to plan maintenance (20-50%);
- Increase equipment uptime (10-20%);
- Reduce overall maintenance costs by (5-10%).

The implementation of this methodology can be executed through either standard software provided by third-parties (IBM, SAP, Siemens and Microsoft), or though the development of custom software in-house. As expected, divergencies oppose these two types of application, price being one of with high relevance. This sole factor directly affects the status-quo of the SME's and Multi National Enterprises (MNE). Therefore opting for custom solutions developed within the company can outdraw further benefits, since the methodology is being implemented thoroughly with the sole purpose of achieving the enterprise goals.

Several companies have implemented PM methodologies in their assets/machines. Trenitalia, an Italian train company, due to incidents that occurred planned to integrate predictive maintenance in their locomotives detected that over 1600 locomotives had to be removed from service because of malfunction. Taking into consideration the number involved, delays were expected since a huge portion of actives were being placed to maintenance. To understand and prevent this type of situation, the company placed sensors to control several parameters in almost 1500 locomotives. All the data generated in service was then transferred into the private cloud server for future analysis. The information gathered allowed the company to "decrease downtime by 5-8 per cent and reduce its annual maintenance spend of $1.3 billion by an estimated 8-10% , saving about $100 million per year." (Colemen et al., 2017, p. 11). Consequently, with the correct usage of sensors, this company managed to implement a new methodology, resorting to state of the art components while decreasing costs, which allowed to provide a better service output to their clients (Colemen et al., 2017).

2.3 Data Mining

With the evolution of the current markets, sustainable competitive advantages can no longer be applied which leads companies to desire a decrease of costs while maintaining productivity levels (Kampker et al., 2018; Park et al., 2016). Currently, "engineers and scientists are frequently working to develop models, methods and features to minimize costs, optimize production and increase reliability." (Sakib & Wuest, 2018, p. 268). The reason behind this need, to further develop solutions, mostly stands out due to the fact companies need improved decision support systems, that can extract useful knowledge from all the data available (Hänel & Felden, 2017). Through the usage of sensors, companies were able to gather important data regarding their operations, unfortunately, “the act of data collection alone does not offer advantages.” (Meissner et al., 2017, p. 166). To generate value to a company, data must be harvested, processed and analysed through data mining techniques, which allows to attain “the right information at the right time” (Park et al., 2016, p. 625).

The data mining practices allow the capability to extract information from data which can range from associations, patterns or tendencies that would remain hidden in data without proper analysis. In specific cases, with the usage of classification algorithms and historical data, this practice allows the prediction of future occurrences, such as possible failures or events (Pham & Afify, 2005). With industry 4.0 and PM it is possible to further understand machine operation and status, via the analysis of control parameters, gathered by sensors and log activity automatically generated throughout daily-basis operations (Hänel & Felden, 2017; Meissner et al., 2017).

Multiple sectors of industry are investing in data mining techniques to extract important information, such as Selective Laser Melting (SLM), machine tool linear axes, injection moulding, amongst others. Even though the application can vary into multiple scenarios, it is not possible to state the amount of data that must be generated to perform data mining techniques and obtain viable results. Nevertheless, the time-span associated with the use cases is greater than two years (Park et al., 2016; Schmidt & Wang, 2018; Uhlmann, Pontes, Geisert, & Hohwieler, 2018).

The evolution of this field allowed to generate frameworks to create a common basis to explore data. These elements allow to better understand how to approach specific problems and provide an overall guide of how to implement data mining methodologies from ground-up. A known framework in this area is the Cross-Industry Standard Process for Data Mining (CRISP-DM) that is divided into six steps, as it can be seen in Figure 2 (Jenke, 2018).

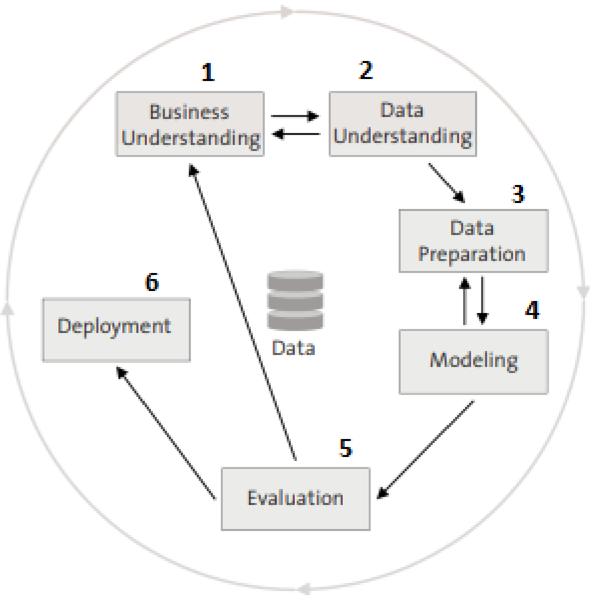

Figure 2 - CRISP-DM Model (Jenke, 2018)

The first step, (1) **Business Understanding,** regards the contextualization of the project. Its goal is to understand the background in which the data is going to be used. All relevant analysis regarding the surrounding aspects of the project is analysed in this step, such as the goals of implementation, the success criteria and the gathering of business knowledge which can aid in the overall project (Jenke, 2018).

The previous step and **(2) Data Understanding** are heavily linked since in this second one data is obtained, and initial analysis is performed. These analyses may lead the user to retract to Business Understanding to clarify features or details which were uncovered. This iteration allows the opportunity to generate a clear view of the project, data and goals (Jenke, 2018).

The **(3) Data Preparation** regards the setup of the datasets where the methodologies and procedures can vary according to the type of needs. Datasets must contain useful and rich data to extract the best results from the models, since they are the sole input for the machine learning methods. The increment of quality can be performed with the removal of outliers and poisonous samples of data, therefore reinforcing the structures of the dataset (Jenke, 2018).

Once the datasets are prepared, follows the **(4) Modeling** phase. This step regards the usage of the content generated in Data Preparation, where the start point regards a choice of models, with "usually, more than one method is available" (Jenke, 2018, p. 72). The choice for more than one model relies on the fact that each model fits differently in the same dataset, resulting in different outcomes. Furthermore, since certain models may have certain necessities, an iteration is generated, which allows the user to regress to the data preparation step and adapt the dataset to the model's needs. Regarding the algorithms to use in data analysis, the choice may vary between several supervised machine learning algorithms, such as Neural Network (NN), Decision Trees (DT), Bayesian Networks (BT), Support Vector Machine (SVM), amongst others. To identify which type of method is most appropriate to the use case in question, it is necessary to perfectly understand the dataset and the objectives. Inadequate data or variables may result in incorrect or inadequate results. Consequently, misinformation is generated, which does not uphold any support for the process of decision making (Jenke, 2018; Schmidt & Wang, 2018). Table 4 presents some of most commonly known algorithms and their features.

Table 4 - Features of common algorithms (Lee et al., 2014)

Algorithm	Usage	Advantage	Disadvantage
NN	- Simulate the structures and functions of neural networks; - Can learn the knowledge by modelling complex relationships between inputs and outputs and find patterns in data.	- For complex systems which involve non-linear behaviour and unstable processes; - Adaptive system.	- No standard method to determine the structure of the network; - Requires enough computational resources.
DT	- Make a decision or classify data item by starting at the root node of the tree and following the assertions down until reaching a terminal node (leaf of tree); - A special form of a rule set, characterized by a hierarchical organization of rules.	- Good visualization, easy interpretation and quick analysis ability for decision making.	- Need high-level experience and knowledge to formulate the tree structure.
SVM	- To project feature space into a higher dimensional space by a kernel function; - To find an optimized separation hyperplane in the projected space to maximize the decision boundary.	- Achieves better decision accuracy in special cases because of the maximized decision boundary. - Efficient for a large dataset and real-time analysis.	- No standard method to choose the kernel function which is the key process for SVM;
BN	- A directed acyclic graph tool to present the structure of conditional interdependency relations and probability distributions between variables in one domain system;	- Reduces the number of parameters to learn a domain structure by marginalizing conditional probability distributions; - Visualizes the dependency links between each pair of variables.	- Leaning an unknown structure can be complex and costly; - Relies on a certain amount of prior knowledge of the domain.

By understanding the model that best fits the dataset in question, the following step is **(5) Evaluation**, which regards the validation of the model in terms of the business needs and criteria established in the first step. Once these are validated, then the model is approved and deployed. Furthermore, this step also links to Business Understanding, in the possibility that criteria were not created accordingly, since “too often, false assumptions require going back to step 1 and to revise the business understanding.” (Jenke, 2018, p. 72).

Finally, the last step in this framework is **(6) Deployment**. In this final procedure, the artefact developed is transmuted from a confined setup into a normal operation scenario. Its usage may vary according the needs of the user which can range from simple dashboards to wide enterprise tracking of data operations (Jenke, 2018).

3 Case Studies

The main goal of this chapter is to illustrate the application of predictive maintenance in Industry 4.0 scenarios through the study of five case studies in different areas of industry. Taking in consideration the goal of this dissertation, this analysis will serve as an input for the development of a PM model. Being industry sectors closed fields where the competitive advantages can easily be attained by competitors, attaining information is of increased difficulty since companines tend to maintain a close loop of data. Therefore, the content analysed in this chapter was obtained from published and reviewed papers, in which the author did not have direct contribution to the development. The content analysed in this chapter derives from the following use cases:

- **Case A**: “Intelligent pattern recognition of a SLM machine process and sensor data” from Eckart Uhlmann, Rodrigo Pastl Pontes, Abdelhakim Laghmouchi and André Bergmann;
- **Case B**: “The prediction and diagnosis of wind turbine faults” from Andrew Kusiak and Wenyan Li;
- **Case C:** “Data analysis and feature selection for predictive maintenance: A case study in the metallurgic industry” from Marta Fernandes, Alda Canito, Verónica Bolón-Canedo, Luís Conceição, Isabel Praça and Goreti Marreiros;
- **Case D:** “Improving rail network velocity: A machine learning approach to predictive maintenance” from Hongfei Li, Dhaivat Parikh, Qing He, Buyue Qian, Zhiguo Li, Dongping Fang and Arun Hampapur;
- **Case E:** “Fault Detection in induction motors based on artificial intelligence” from Vinicius Augusto Diniz Silva and Robson Pederiva.

3.1 Case A - Pattern Recognition in an SLM machine

SLM machines are known due to their capability to use high power lasers to fuse metal powder. These machines perform under the basis of additive manufacturing, otherwise known as stacking layer by layer to achieve the desired product which allows to build workpieces using multiple layers (Uhlmann, Pontes, Laghmouchi, & Bergmann, 2017).

Several researchers have already addressed problems and concerns in terms of assessing data generated by these types of machines since previous analysis showed that manual verifications do not provide the desired output. Disregarding the fact that they can be executed, it does not provide an optimistic ratio in terms of results/time wasted. With the published article the authors aimed to understand them if it was possible to identify patterns in the data from this type of machine and if it was possible to assess the condition of the machine based in control parameters. Through the usage of a CMT (Condition Monitoring Tool) developed by Fraunhofer IPK, the goal was to use the results to predict the behaviour of the machine. Consequently, this could allow to boosting performance measures, therefore, improving the reliability of the components and quality of the output. For this specific case, a total of 16 parameters were monitored. Table 5 displays the parameters and its units. All these regard operations of the machine with substantial influence on either the layer quality or the time to manufacture (Uhlmann et al., 2017).

Table 5 - Monitored parameters for SLM machine (Uhlmann et al., 2017)

Number	Parameter	Unit
1	Platform Temperature	ºC
2	Process Chamber Temperature	ºC
3	Pump Temperature	ºC
4	Process Panel Temperature	ºC
5	Electrical Panel Temperature	ºC
6	Optical Bank Temperature	ºC
7	Collimator Temperature	ºC
8	Environment Temperature	ºC
9	Process Oxygen	%
10	Process Pressure	mBar
11	Filter Conditions	%
12	Total Layer Time	Seconds
13	Layering Time	Seconds
14	Idle Time	Seconds
15	Recounter Motion Time	Seconds
16	Recounter Filling Time	Seconds

Taking into consideration the issues to be tackled, the authors decided to perform the methodology shown in Figure 3. The process starts with data acquisition from the two distinct sources, the machine’s log files and sensors installed. This is followed with process and treatment of data prior to its integration in a database in order to consolidate data integrity and shape for consequent analysis. From the integration a split is performed, where two analysis are conducted: Pattern classification (I) and Data Clustering (II).

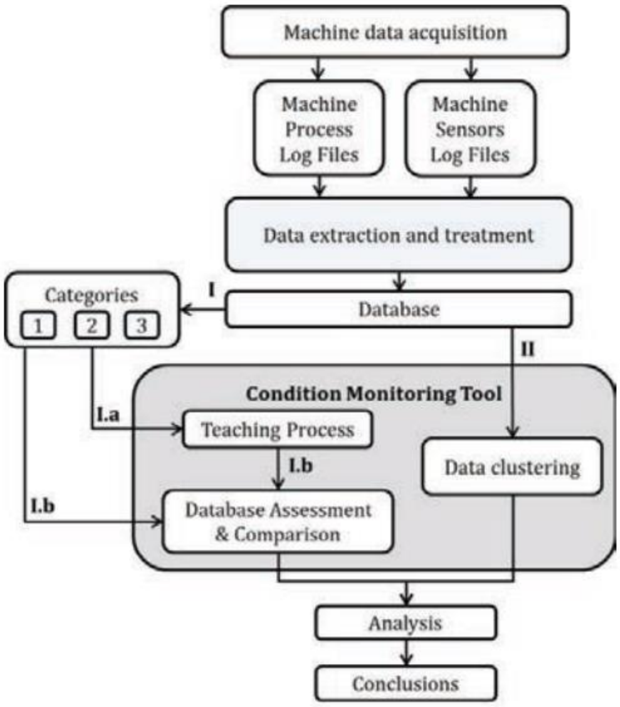

Figure 3 - Methodology (Uhlmann et al., 2017)

In terms of **Pattern Classification (I)**, the machine output was classified into three categories: “Finished perfectly”, “Finished with errors” and “Not finished”. Taking into consideration these categories, a dataset was generated, which contained a total of 90 manufacturing processes of the three categories from a total of 271 originally stored in the database. With this dataset four algorithms were tested with different input parameters to the categories Finished Perfectly and Not Finished: Nearest Neighbour, Naïve Bayes (NB), NN and SVM. The parameters that attained the best results were Process Oxygen (over 75%) and Idle Time (over 90%). Bearing in mind that three categories were available, the authors decided to obtain the average of the algorithms results for the two most suitable variables across the three categories. This resulted in the selection of idle time as the most adequate variable with Bayes Classifier and SVM as the most appropriate algorithms (around 60%).

Data Clustering (II) was used by the authors with the goal of understanding the number of categories to input into the algorithms. By using K-means, it was possible to identify on average 3.5 categories, which is very close to the number of categories defined by the authors in the dataset.

Through the analysis conducted and the methodology executed by the authors (Figure 3), it is possible to conclude that the usage of categorization allows to understand how each feature directly relates into the prediction of machine’s output type. Since machine output can vary, it is understandable that the authors deem worthy of understanding the most viable option for the available scenarios, therefore attaining a lower result.

3.2 Case B - Wind Turbine Prediction

With the need to search for new solutions in terms of harvesting and generating new sources of energy, interest started to grow regarding wind power solutions, more specifically, wind farms, since this solution allows countries to meet the targets of carbon emission. Yet, multiple problems arise for countries and companies that are managing this equipment once they are installed. Since wind turbines are in remote locations and access is difficult, maintenance operations increase difficulty due to their complexity, therefore, transmuting the cost of a bearing from \$5000 to \$250000 with the addition of no power being generated during the operation. All this leads to an increase in priority to attain condition monitoring tools.

To provide a solution for the prediction of faults in wind turbines, the authors proposed a methodology to access faults in a three-level system:

- **Level 1 –** Predict the occurrence of a status/fault;
- **Level 2 –** Determine the category of the status/fault;
- **Level 3 –** Predict the specific fault.

The case study combines data from four turbines in a time-span of three months from two sources: SCADA data and the fault data. The first dataset type contains approximately 25000 records and 60 variables for each turbine. Below follows an example of the variables that were grouped into four categories:

- **Wind-related variables** – wind speed, wind direction, wind intensity and turbulence;
- **Energy-related variables** – power output, blade pitch angle, generator torque, rotor; speed;
- **Vibration related variables** – drive train acceleration and tower acceleration;
- **Temperature related variables** – bearing temperature, nacelle interior temperature.

The fault dataset contains approximately 7000 occurrences with over 350 different status codes. The authors consider a fault as a “status that with a certain probability results in a severe consequence to the wind turbine system” (Kusiak & Li, 2011, p. 17). Nevertheless, faults with *No error* may not lead to any damage. Table 6 shows examples of status codes:

Table 6 - Status Codes (Kusiak & Li, 2011)

Status Code	Status Text	Category
1	Program start PLC	2
2	No Errors	4
3	Manual Stop	4
4	Remote Stop	4
5	Remote Start	4
6	System OK	4
9	Under-Voltage	4
21	Cable Twisting left	4
25	No speed reduction with primary braking	1
28	No speed reduction with a secondary braking	1

The first step performed was the treatment and cleansing of data, such as removing duplicates and incorrect values which led to decrease in quantity of occurrences and status codes. Taking into consideration the vast diversity of fault codes, the authors categorized the fault type into four values in terms of impact on the wind turbine where "categories 1,2 and 3 might adversely impact the wind turbine system and its components. But the status codes in category 4 are not likely" (Kusiak & Li, 2011, p. 19). To integrate both datasets, the same process was performed in the SCADA dataset by establishing status code and categories.

From this point, the following step was to generate a model to predict in a three-layer system, as stated previously in this sub-chapter, with n×5 minutes in advance where "n is the number of timestamps in advance of the status/fault." (Kusiak & Li, 2011, p. 20). Therefore, for each level of prediction, the authors generated a training dataset with two-thirds of occurrences to train the model and a test dataset with the remaining ratio. The model used is available in Figure 4 which features the variables used for each layer of prediction, wind speed and power output. In terms of algorithms, the authors chose a wide range, differing between layers:

- **Level 1** – NN, SVM, Boosting Tree Algorithm (BTA) and Neural Network Ensemble (NNE);
- **Level 2** – NN, Standard Classification and Regression Tree (CART), BTA and SVM;
- **Level 3** – NN, NNE, BTA and SVM.

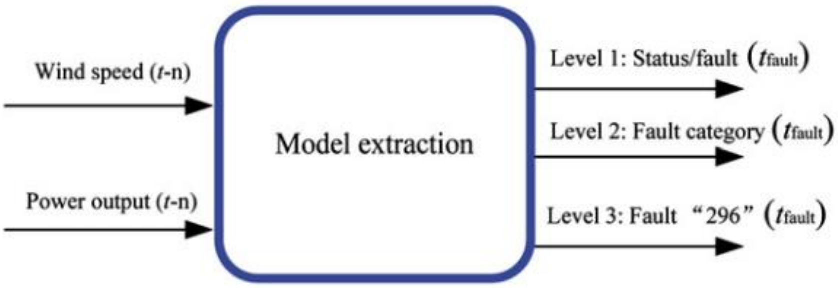

Figure 4 - Model used for three-layer prediction (Kusiak & Li, 2011)

Table 7 shows the results obtained regarding the first level of prediction. All the algorithms implemented achieved an accuracy rate higher than 65%, with special regard to NNE achieving almost 75% of accuracy.

Table 7 - Algorithm's results for the first layer of prediction (Kusiak & Li, 2011)

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
NN	74.71	81.00	68.67
NN Ensemble	**74.56**	**83.67**	**65.81**
BTA	71.27	84.66	59.50
SVM	69.64	59.97	78.92

Regarding Level 2, prediction of the category, all algorithms except CART failed to obtain accuracy's higher than 50% to all the categories, as shown in Table 8. The difference between values is mostly due to the fact that the categories "Normal" and "4" had a higher amount of records than the remaining ones, both for the training dataset and the test dataset.

Table 8 - Algorithm's results for the second layer of prediction (Kusiak & Li, 2011)

Algorithm	Prediction accuracy for normal	Prediction accuracy for category 1	Prediction accuracy for category 2	Prediction accuracy for category 3	Prediction accuracy for category 4
NN	76.66	0.00	0.00	12.00	74.91
BTA	41.00	22.22	83.33	0.00	72.15
CART	**96.08**	**62.50**	**52.94**	**56.00**	**95.20**
SVM	80.88	0.00	0.00	0.00	69.28

For the last level, predicting a specific fault, the algorithms obtained similar results to the ones obtained in level 1, nonetheless, the algorithm with the highest accuracy and sensitivity was BTA.

Table 9 - Algorithm's results for the third layer of prediction (Kusiak & Li, 2011)

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
BTA	**69.81**	**86.67**	**63.16**
NN	72.00	66.67	70.45
NN Ensemble	68.00	82.88	66.67
SVM	70.59	47.06	82.35

Through the analysis conducted, it is possible to see a feature that stands out due to its specificity, which is multi-level prediction. It is duly noted that one of the major limitations of this use case was low data quantity, which can be seen more specifically in second level of prediction where categories 1, 2 and 3 obtained low results. Despite the fact that initial quantities show vast amount of data (7000 instances of status/fault data regarding Turbine 4), by applying pre-process it “led to 1329 instances covering 66 different status codes.”(Kusiak & Li, 2011, p. 18). Therefore, it is possible to conclude that the classification is strongly linked to dataset richness.

3.3 Case C - Predictive Maintenance in Metallurgic Industry

Multiple companies in a wide range of sectors are trying to take the first steps into predictive maintenance to further understand their processes, production and equipment. Nevertheless, the transition can be difficult if companies do not have proper foundations established upper hand. One sector trying to ride this wave is metallurgic. Being a stable area with wide search all over the world, companies in this area are trying to understand how predictive maintenance can establish a new course in the company itself.

Following this line of thought, the authors aimed to assess the early stages of the implementation of predictive maintenance in a disclosed company in the metallurgic sector. The company in question “is specialized in precision parts production and uses raw materials, such as aluminium, steel, bronze and technical plastics, to produce custom parts for industry clients.” (Fernandes et al., 2019, p. 254). In order to ease the comprehension, the company shall be named XPTO.

During the time the study was conducted, XPTO had available the InValue platform that consists of an all-in-one solution addressing big data problems in a predictive maintenance scenario. The system’s architecture is available in Figure 5.

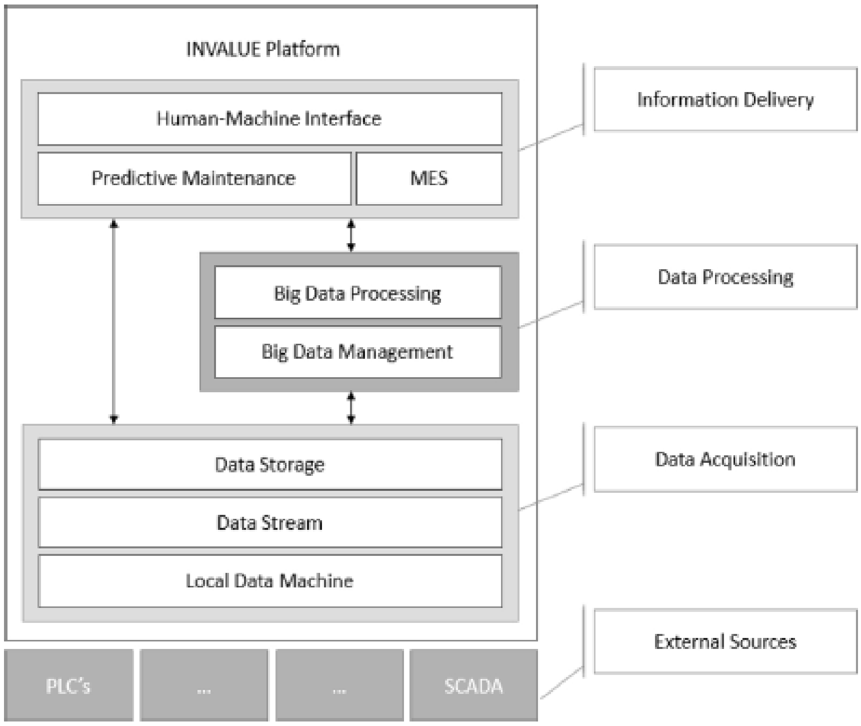

Figure 5 - InValue architecture (Fernandes et al., 2019)

Comprised of three layers, the platform automatically harvests data in the first layer, Data Acquisition. The data obtained is then stored in a storage system that shall serve as input for the second and third layer (Data Processing and Information Delivery). This allows easy access for visualization and processing purposes.

At the time the authors conducted the study, a data acquisition module was installed in one of the machines, a Hass ST-30 lathe to extract machine and sensors data. In total, 48 features (43 from the machine, four from the sensors and one timestamp) were extracted with a timespan of six days resulting in 23124 rows. The features extracted are available in Annex K. Since it's an initial project with implementation to one machine, it was not possible to extract information regarding issues that occurred in the machine, a fact disclosed by the authors which limited their approach available below (Fernandes et al., 2019):

1 Data acquisition;
2 Data analysis and treatment;
3 Feature selection from the dataset;
4 Definition of rules to monitor data generated.

Since the core of this project is to expand to the remaining machinery in the production line, the data acquisition module is the same as the one to be implemented in the remaining machines. Therefore, the ST-30 that served as a pilot did not possess axes that the remaining machines have at disposal. This resulted in the generation of features that did not contain any values, which led the authors to remove the following variables:

- Maximum axis load for axes Y, B, C, U, V, W, T;
- Present machine coordinate position for axis Z;
- Present work coordinate position for axis Z;
- Present tool offset for axes Z, A and B.

With the evaluation of the variables "Tool in Spindle" and "Tool Number in Use", the authors were able to conclude that these do not provide the desired information to the scenario in hand, since the machine does not identify the tool that is used. Instead, it registers the position where it was placed for operation, making " therefore, impossible to identify the operating tool using only the information obtained from the machine." (Fernandes et al., 2019, p. 255). Regarding manufacturing operations, one feature that the authors deemed as high importance was the Present Part-Timer, because it registers the time it took to produce a single object. Furthermore, the timer stops whenever the production of a part is not being conducted, which allows "to discern if a part is being produced or not." (Fernandes et al., 2019, p. 255). Nevertheless, the authors could still not perceive through the data obtained if a part was successfully generated. To understand the success of production it is necessary to cross-examine two variables, **Complete Part-Timer** which specifies the amount of time needed to produce a part and **M30 Parts Counter** that registers how many times the M30 code[2] was generated. In terms of information gathered from components performance, the authors consider spindle's load and spindle's speed as two very important features for analysis, since these are directly related to machine's anomalies. The spindle's load registers the load in the form of energy that is fed into the spindle in use by the machine, whereas the spindle's speed is registered in the form of Rotations per minute (RPM).

In conclusion, the authors were able to assess high importance to three group features:

- Spindle features (specific to machine);
- Coolant fluid and machine coordinates (specific to machine);
- Features regarding the production of parts (general to the entire production line).

[2] Code to signal the end of a cycle.

Taking into consideration these three main groups, the authors performed feature selection to understand, which could provide a better foundation in this specific scenario. The method utilized was the Minimum Redundancy Maximum Relevance (MRMR), which relates to the filter methods of feature selection. Therefore, the authors came up with 15 feature pairs that constitute 10% of the most redundant pairs, available in Table 10. As the authors describe, some relations are expected to deliver high redundancy since the features regard similar roles, as for instance MP30PC1 and MP30PC2 which count "how many times a given operation was executed;" (Fernandes et al., 2019, p. 258). Consequently, this allowed to diminish the number of features from 47 to 32. The selected features are available in Annex P.

Table 10 - 10% most redundant pairs (Fernandes et al., 2019)

Feature 1	Feature 2	Redundancy
MP30PC1	MP30PC2	2.2327
Tool in Spindle	Tool Number in Use	2.1321
Last Complete Part timer	Last Cycle Time	1.9991
Present Machine coordinate Pos A	Present Work coordinate Pos A	1.9444
Tool in Spindle	Present Machine coordinate Pos A	1.8258
Tool in Spindle	Present Work coordinate Pos A	1.8247
Tool Number in Use	Present Machine coordinate Pos A	1.8226
Tool Number in Use	Present Work coordinate Pos A	1.8223
MP30PC1	Total Tool Changes	1.8056
MP30PC2	Total Tool Changes	1.8056
Present Work coordinate Pos X	Present Work coordinate Pos Y	1.1861
Present Machine coordinate Pos X	Present Work coordinate Pos X	1.0914
Last Complete Part Timer	Previous Cycle Time	1.0908
Last Cycle Time	Previous Cycle Time	1.0876
Coolant Level	Total Tool Changes	0.9294

Since no information regarding failures and malfunctions were available, the authors decided to assess the situation in terms of rule-based models. These rules are executed by the InValue system described previously to assess data with a higher response from the system itself. An extract from the rules is available in Annex M for further analysis.

In conclusion, through the analysis of this use case, it is possible to state the relevance of feature selection in a predictive maintenance scenario. The implementation of this procedure allowed to understand valuable relations between features to which the authors did not know and also reduce drastically the dataset in hand from 47 features to 32. Even though valuable information was extracted, the fact that issue-related data was unavailable for extraction led to a limitation of the scope. Nevertheless, this constraint led to the conception of a set of rules which serve as an input for the development of customized learning models.

3.4 Case D - Predictive maintenance in Rail network

With bigger demands in public transportation, faster methods were needed to display a solid and reliable foundation for transporting people and cargo to specific points. Allowing a faster way of transport without compromising the safety of the passengers is a vital step in a long-term perspective. With this line of thought, the authors aim to assess and create machine learning solutions that allowed to predict possible failures in components of rail cars. Consequently, this allows conducting “proactive inspections and repairs, reducing operational equipment failure.” (Li et al., 2014, p. 19). According to the authors the goals of this work are **Alarm Prediction** and **Bad truck and wheel prediction.**

3.4.1 Alarm Prediction

In this scenario, alarms related to failures due to hot bearings are predicted through the usage of multiple detectors. This work allows reducing service interruption and unscheduled train stops. The usage of alarms for failure prediction is a common proactive measure against impacting events. Nevertheless, the implementation of this measure in this specific scenario is considered alarming by the authors, since the L1 alarm is generated only when the failure is about to occur. Consequently, this results in “little time for planning and thus putting the maintenance organization in reactive mode.” (Li et al., 2014, p. 20). One of the examples demonstrated by the authors is the Hot Box Detector (HBD) detecting the temperature of the bearings. If this sensor registers 170º Fahrenheit, the L1 alarm issues a flag to immediately stop the respective train. Following this line of thought the authors approached the situation with a different solution. By applying machine learning methodologies to generate new rules allowing the respective operators could gain more time to act upon the eventuality of the L1 alarm being triggered.

In terms of conditions for execution at the beginning of the project, two challenges needed to be addressed:

- Low false alarm rates: it was required a false alarm rate of no further than 0.014%, due to limited budget which generated constrained resources for maintenance;
- Generate rules that could easily be interpreted by humans to ease the decision process.

By proposing the development of a customized SVM model (that achieves low false alarm rates) and combining data from multiple sensors (to increase the accuracy of prediction), it was, therefore, the authors were able to address the two challenges stated above.

In terms of data, the network contained 800 fully functioning HBD sensors which generated three terabytes of data in a year. Regarding the remaining sensors, Acoustic Bearing Detectors (ABD) and Wheel Impact Load Detector (WILD), the network contained twelve of each, where the WILD sensor registers data regarding wheel impact load and the ABD is responsible for generating tables with the alarm event data. Together, the data of these three sensor types are linked by the bearing ID, which consequently leads to a low volume of data for a short window frame. To solve this situation, the authors applied a varied window frame to each sensor, where HBD was chosen seven and fourteen days of data and WILD and ABD data were collected one and three months respectively.

From the initial analysis conducted to the available data, results allowed to understand that the features from ABD sensors do not generate significant impact since the data collected is noisy. From WILD and HBD measurements, 55 features were extracted to which the authors implemented Principal Component Analysis (PCA) therefore, reducing to 12 features.

In order to understand if the implemented methodology was best suited to the scenario in hand, the authors faced it with a decision tree to compare results in possible scenarios, three or seven days in advance. Table 11 shows the results obtained for the decision tree and Table 12 shows the results obtained for the SVM.

Table 11 - Decision Tree Results (Li et al., 2014)

Scenario	7-7		14-3	
	TPR (%)	FPR (%)	TPR (%)	FPR (%)
Highest True Positive Rate (HTPR)	91.546	6.849	92.568	4.998
Lowest False Positive Rate (LFPR)	61.256	0.976	68.463	1.073

Table 12 - SVM Results (Li et al., 2014)

Scenario	7-7		14-3	
	TPR (%)	FPR (%)	TPR (%)	FPR (%)
(HTPR)	97.585	5.657	99.775	3.966
(LFPR)	7.459	0.000	8.987	0.000
(HTPR) under constraint	38.542	0.014	45.368	0.012

From the results obtained, it is possible to notice that the SVM attained a better ratio regarding True Positive Rate (TPR) and False Positive Rate (FPR) for the two scenarios in common. Since the lowest false positive rate for the decision tree is of almost 1% in both cases, it does not meet the agreed challenge of a false alarm rate lower than 0.014%. The SVM, while succeeding in this challenge only attains a TPR of 38% for seven days and 45% for three days. Overall the results are fairly good, taking into consideration the challenge established in upperhand.

3.4.2 Bad Truck/wheel prediction

In order to predict bad truck/wheel failures, the authors used data from several sources, such as previous failures, sensor data, configurations and condition data. Since the approach is similar to both the truck and the wheel (pattern identification of errors in wheel related features obtained by detectors), the authors proposed a similar approach:

1. Feature Selection
2. Data Aggregation
3. Data Labelling
4. Data Prediction
5. Rule Extraction

In terms of feature selection, the authors applied an ANOVA test which allowed to select 20 features. These features were selected from four sensors, Machine Vision (MV), Optical Geometry Detector (OGD), WILD and Truck Performance Detectors (TPD). The features selected are available in Table 13.

Table 13 - Features selected from sensors (Li et al., 2014)

Detector Type	Attribute
MV	Wheel Flange Height
	Wheel Flange Thickness
	Wheel Rim Thickness
	Wheel Diameter
	Wheel Thread Hollow
	Brake Shoe Upper Thickness
	Brake Shoe Lower Thickness
OGD	Truck hunting peak-to-peak (PTP) measurement
	Truck hunting amplitude
	Truck inter-axle misalignment (IAM)
	Truck rotation measurement
	Truck tracking error
	Truck shift measurement
WILD	Wheel average downward load reading
	Wheel peak downward load reading
	Wheel average lateral load reading

	Wheel peak lateral load reading
	Difference between peak and average downward load reading
	Truck hunting index
TPD	The ratio of later and vertical load

In terms of data available, approximately 500 Gb of data were collected in a time span of almost two years (January 2010 to March 2012). Nevertheless, only a time-lapse of three months was used regarding the prediction phase. Since easy human interpretability is vital to the organization, the authors chose a decision tree, since it allows an easy interpretation of the rules to be attained. With a split of 80%-20% for training and test respectively, the authors achieved a TPR of 97% for both datasets and FPR of 0.23% for the test dataset and 0.20% for the training dataset. The decision tree is available in Figure 6, which allowed to easily attain the rules regarding the last step.

From the analysis, it is possible to state that the integration and coordination of data from “multiple detectors play an important role in the analysis” (Li et al., 2014, p. 25). The reason behind this statement is mostly due to the fact that two core concepts are shown in this use case, Big Data and IoT, which allows the authors to achieve great results, especially in bad truck/wheel prediction. Furthermore, the methodology implemented can serve as a benchmark for the remaining sectors in industry, since “the algorithms described above is more generally applicable to many other industries that use sensor network for equipment health monitoring” (Li et al., 2014, p. 26).

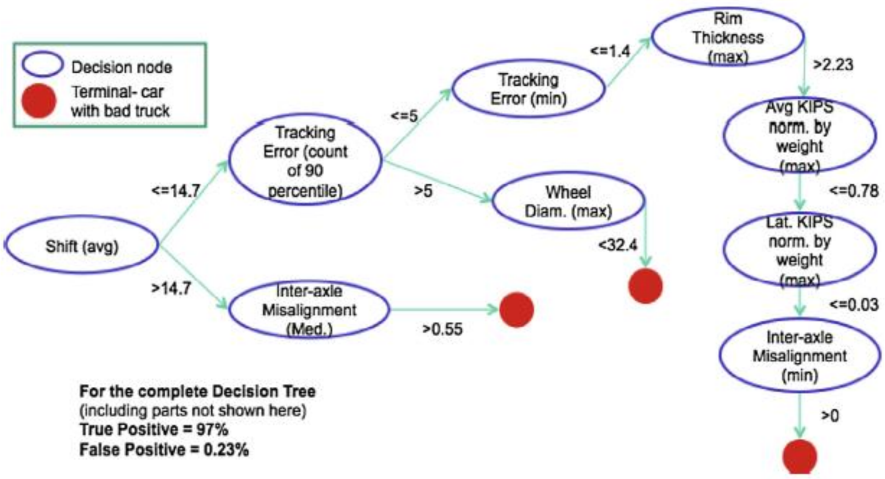

Figure 6 - Neural network obtained and consequent rules (Li et al., 2014)

3.5 Case E - Machine learning in electric motors

Electric motors are an essential component in industry and other sectors. The wide usage of this equipment is mainly due to their simplicity and capability to adapt to a wide variety of scenarios. However, with such dependability comes a cost. When in the presence of failure of this equipment, production can be halted, which leads to downtime of production and increase in costs due to inability to produce the desired products. With this case study, the authors aim to detect and diagnose electrical malfunctions in electric motors. The equipment used for the experiment is available in Figure 7.

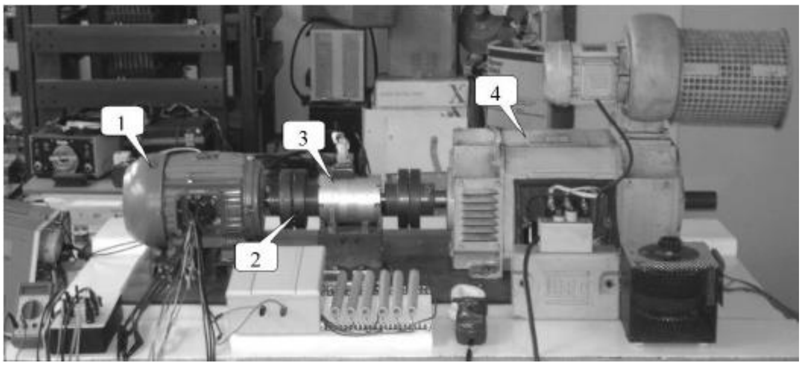

Figure 7 - Three-phase motor (V. Silva & Pederiva, 2013)

The four main components shown are: (1) three-phase motor, (2) flexible couplings, (3) torque meter and a (4) CC generator. In total the authors collected 680 vibration signals from the equipment which couple into six different faults. Table 14 states the distribution of signals regarding the six possible faults and normal behaviour.

Table 14 - Number of patterns per condition (V. Silva & Pederiva, 2013)

Condition	Number of patterns
No-Fault	170
Unbalance	110
Misalignment	110
Mech. Looseness	110
Short Circuit	60
Phase unbalance	60
Broken Bars	60

In total, three machine learning techniques were applied: SVM, ANN and Fuzzy Logic (FL). The input variables chosen for this procedure were the amplitudes of the frequency of the vibration signals, categorized into three possible values: Small, Medium and High. In terms of data, the sets of data regarding mechanical and electrical faults were split and partitioned, resulting in three subsets for training and three for validation. The results obtained are divided into two tables, mechanical and electrical respectively.

Table 15 - Detection Results of mechanical faults (V. Silva & Pederiva, 2013)

Fault	**AI**	**S1 Hit Rate (%)**	**S2 Hit Rate (%)**	**S3 Hit Rate (%)**
No Fault	SVM	96.00	96.00	98.00
	ANN	86.24	84.49	91.29
	FL	100	100	100
Unbalance	SVM	96.00	96.00	92.00
	ANN	90.80	86.97	83.35
	FL	96.00	96.00	88.00
Misalignment	SVM	88.00	88.00	88.00
	ANN	94.08	84.77	81.18
	FL	84.00	84.00	88.00
Mech. Looseness	SVM	80.00	80.00	84.00
	ANN	74.65	70.72	60.43
	FL	72.00	84.00	80.00

From the results shown in Table 15, in the three possible mechanical faults, the algorithms obtained good hit rates. The lowest score recorded by the authors was approximately 60% with the neural network in mechanical looseness fault. Nonetheless, it remains an average of almost 70% taking into consideration the three algorithm's results. Furthermore, the remaining algorithms in the referred. Regarding the singular performance of each algorithm, the SVM obtained the most satisfactory results throughout the four possible categories.

Table 16 - Detection Results of electrical faults (V. Silva & Pederiva, 2013)

Fault	AI	S1 Hit Rate (%)	S2 Hit Rate (%)	S3 Hit Rate (%)
No Fault	SVM	100	100	100
	ANN	95.11	90.93	92.93
	FL	100	100	100
Short circuit	SVM	100	100	100
	ANN	91.22	89.73	83.61
	FL	100	100	100
Phase unbalance	SVM	77.78	100	100
	ANN	94.90	95.47	94.94
	FL	100	100	100
Broken bars	SVM	100	100	100
	ANN	92.60	93.23	91.48
	FL	88.89	83.33	77.78

Regarding electrical faults, Table 16, the results obtained far outperform the ones obtained for the mechanical faults. FL obtained mostly 100% hit rate, except in Broken bars, where all algorithms experienced lower results.

With the analysis conducted, it is possible to notice that data categorization plays a vital role in this use case since it simplifies the predictive maintenance approach for the authors. The usage of three types of frequency allows to understand what type of frequencies better adapt to each specific scenario, therefore allowing to generate a common ground for all types of mechanical and electrical faults.

4 Predictive Maintenance Model in Industry 4.0

In this chapter, the main goal is to establish a generic model of a predictive maintenance model, taking in consideration the use cases described previously. Furthermore, to aid the usage of the model itself this chapter aims to provide also a set of guidelines to ease the process.

4.1 Generic Predictive Maintenance Model

With industry providing multiple fields of action, one factor stands out regarding all sectors: maintenance. Maintenance is an indispensable routine implemented throughout all fields, since it is the only line of action to be taken by companies to maintain their OEE. With the introduction of Industry 4.0, the amount and variety of information available skyrocketed into new thresholds. This leads companies to question what variables and data are mandatory to have in order to implement predictive maintenance.

To ease the process of implementation of predictive maintenance, a model is proposed, based on the use cases reviewed in Chapter 3, for the sake of aiding companies to understand what variables and path they should follow in order to successfully generate a viable solution regarding the implementation of predictive maintenance. Figure 8 constitutes the main roadmap for implementing predictive maintenance, which is divided into two different phases: **(1) Data Extraction** and **(2) Data Prediction**.

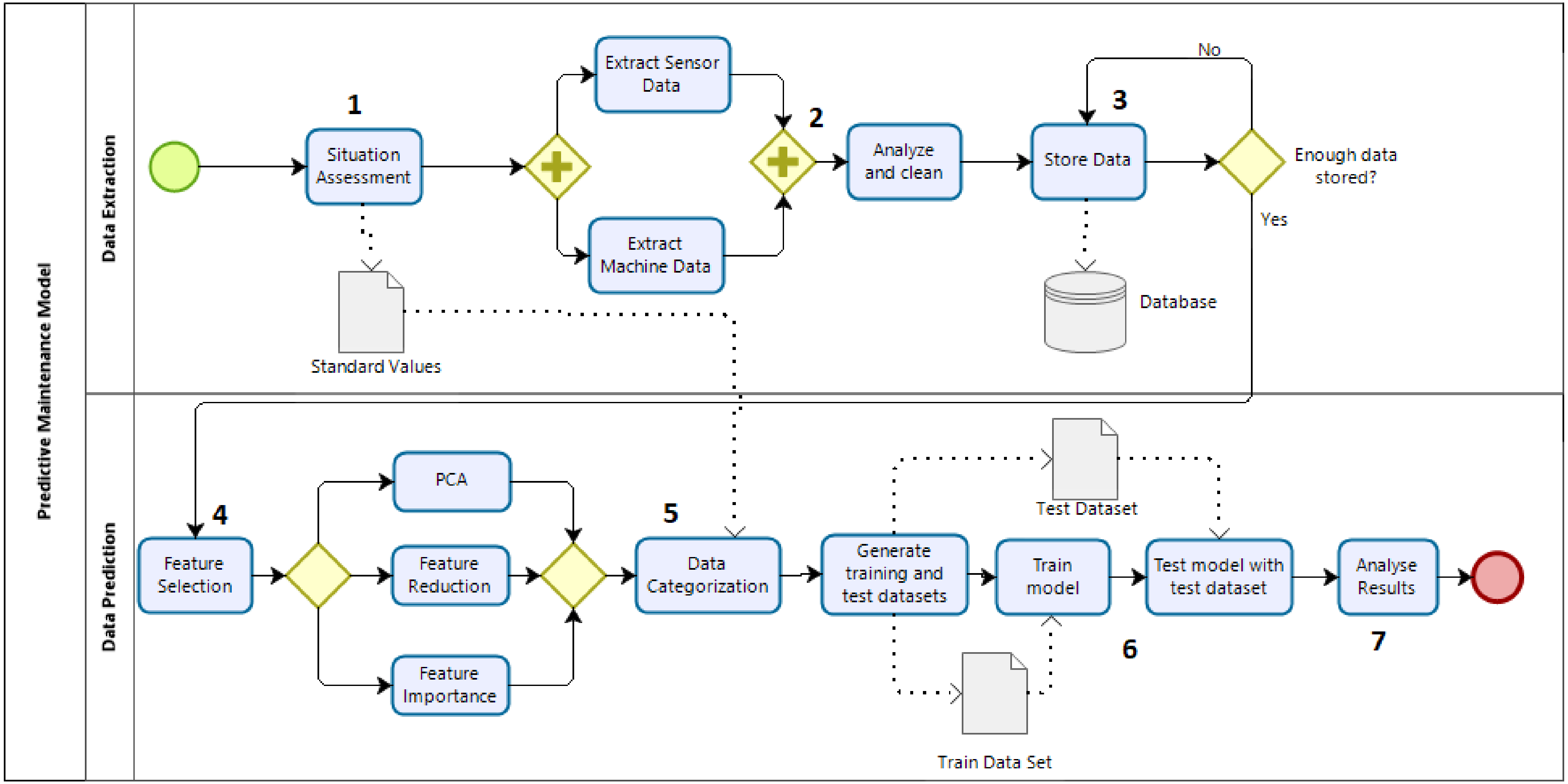

Figure 8 - Predictive Maintenance Roadmap Proposal

4.1.1 Data Extraction

Data Extraction allows users to access and create the main structures required to implement a PM solution. Divided into three major steps, this model starts with the **Situation Assessment (1)**. Assessing the main issues and structures that require further attention deriving from traditional maintenance methodologies is imperative. This step allows the user to understand what the possibilities are, both in terms of data extraction (is the required data available?), but also in terms of risk/reward assessment (if resources are limited). In a possible scenario where not all required data is available to perform analysis, then it is suggested the implementation of sensors in order to reduce the gap in question. The reason behind this step is mostly due to the use cases in section 3.2, 3.3 and 3.4. In these use cases, the authors labelled and organized the data into different categories (use case 3.2) and installed data acquisition modules (use case 3.3).

Once the situation is assessed, it is followed by **Data Extraction and Analysis (2)**. This step has high relevance to the process since it allows the users to understand the data being obtained from the machines and clean data mistakes, therefore preventing illogical errors in the following stages. All use cases rely on this step since it's a common practice regarding data mining methodologies, such as CRISP-DM. Yet, some use cases do not assert practised methodologies, as for instance, Case E. Nevertheless, it is possible to state indirect methodologies such as detecting data patterns for initial data analysis. The remaining ones mention different cleaning techniques, such as extracting duplicates from the original dataset. The removal of incorrect values and duplicates allows to obtain a more coherent dataset for consequent analysis, but also decrease the quantity of data available, as it can be seen in Case B. This specific use case lost a wide variety of data, mostly due to the removal of incorrect and duplicate values.

By conducting data extraction and analysis, it is possible to attain a solid dataset to be used for predictive maintenance methodologies. However, one extraction in most cases is not enough to execute these methodologies, since the array of possibilities is sparse and classification models directly rely on historic data. Therefore, training the model with a high volumetry allows for more accurate results. This premise can be validated through regular operations of extraction, analysis and **Data Storage (3)** in a database or a data lake (depending on quantity of data being generated). Understanding the amount of data needed to perform data analysis is a difficult question to ask since most cases of predictive maintenance require a wide time span of data. One use case that perfectly represents this sentence is Case D, where the authors obtained 500 Gigabytes of data referring to two years and three months. Amount data should differ according the situation to be assessed, and exceptions can be found, for instance in Case C where the authors obtained 23000 rows of data that regard six days of operation. This concludes the first stage of the predictive maintenance model.

4.1.2 Data Prediction

The second stage of the PM model is Data Prediction, which regards an analytic and practical usage of the data stored in the first stage. This part allows the user to directly access and analyse the data to extract viable knowledge that can be used in the process of decision making.

The first step recommended is to perform **Feature Selection (4)**, an important execution in a predictive maintenance scenario, since features/variables selected play a vital role in the outcome results of the models. From the use cases analysed in chapter three, this is considered a vital step used by all the models of PM, where it was applied the following methodologies: PCA and ANOVA (Case D), MRMR (Case C), Feature Importance[3] and try-based system (Case A). With features playing such a vital role in the performance of classification models, it is proposed in Figure 9 a set of common and specific features from the cases reviewed in the previous chapter. Should be taken into consideration that in no regard the list of variables can address the wide variety of sectors in the industry, nevertheless it allows for users of the model to understand what leads they should follow in the scenario of implementation.

Common Features

- Equipment Vibration
- Process Time
- Equipment Temperature
- Engine/motor metrics (rpm, speed, acceleration
- Machine Output and Errors

Specific Features

- Tool in use
- General Metrics regarding equipment
- Position of tools
- Oxygen

Figure 9 - Common and specific feature lists

[3] The choice of features to input into the algorithms in the use case in sub-chapter 3.1 was influenced by pre-existent production knowledge which suggested the usage of a set of variables that were directly related with the generated product.

The list of specific features regards the ones that are not used in more than one use case. They should be considered under strict scenarios where they are available and therefore relevant for analysis, such as the percentage of oxygen used or the type of tool in usage and its position in the process. The list of common features, as the name suggests, regards features that were used in the process throughout the other use cases, such as time-related features, temperatures obtained, engine metrics and the vibration of the equipment.

When feature selection is completed, the following suggested step is **Data Categorization (5)**. This step overall is important in terms of output classification, and more specifically in terms of non-numerical data. The capability to transform non-numerical data into text data allows the user to access a wider array of possibilities. One of the main advantages of categorizing data is the possibility to reduce the flow variety of data that is inputted into the model. Choosing either way leads to different results, both in terms of model results and predicted output, if output is categorized. Therefore, should be considered what the goal of the implementation is and if categorization is a required step. In Cases A and B data categorization was performed to further enhance data analytics capabilities regarding the usage of classification models.

Upon the conclusion of data categorization, it is possible for the user to see the final step of preparation to execute the model, which regards the **Dataset and Model Preparation (6)**. The dataset preparation is executed through the creation of two datasets by data split, training and test respectively, as shown in Figure 10. The first dataset is used to train the model and the second one to test the accuracy of the model. The main difference between these two datasets regards only the output variable because the test dataset does not contain the output to be predicted. In normal scenarios, the split of the original dataset tends to favour a higher percentage of data into the training dataset. The model preparation is performed with the datasets generated. It is imperative to first train the model and only after testing it since the usage of the training dataset allows the user to build the model and prepare it to conditions to which the desired output is established by the predicting variables. The usage of the test dataset allows the user to perceive the performance of the model on unseen data since the possibility remains of the model is being well trained but delivering a poor performance in the test dataset. This model does not state a methodology regarding the choice of the algorithm itself since it can differ regarding the preferences of the users and the desired outcome. In Case D was implemented a decision tree, mostly due to the fact it could allow an easier human interpretation and extract a rule-based system for consequent monitoring.

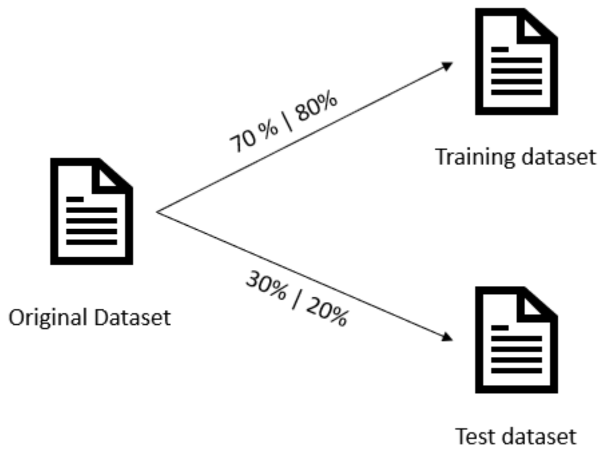

Figure 10 - Splitting original dataset into training and test dataset

The last step of the model is **Result Analysis (7)**. This is performed by analysing the parameters obtained with the execution of the models in the test dataset. Nevertheless, specifications may be defined *apriori* which therefore should be taken in consideration upon analysing results, as used in case D. The usual parameter to analyse is accuracy, which regards the ratio of correct cases in total cases, being an easy to understand and widely used parameter.

4.2 Guidelines roadmap

The implementation of the model can be considered as one of the most difficult steps in the process due to the fact that differences arise outside and inside sectors of industry. In other words, since not all companies produce the same product with the same equipment, different problems affect different company's production. Therefore, understanding the scenario and how to approach is vital to provide solid foundations for the implementation of predictive maintenance. With this line of thought in mind, the guideline roadmap exposed and described in this sub-chapter aims to provide an easy approach to the artefact generated in this chapter. Just like the model proposed, the following set of guidelines, that takes into consideration the methodologies and steps taken in each case reviewed, follows the same structure to ease the interpretation of the user in the process of implementation.

- **<u>Data Extraction</u>**
  - **I. Understanding the problem**
    - What are the main equipment/components to monitor? (A, B, C, D, E)
    - Are the events of the machine being registered? (B, E)
    - Is regular operation halted regularly? (D)
  - **II. Main structures for implementing PM**
- **<u>Data Prediction</u>**

I. **Feature Selection**

- Taking into consideration the issues to monitor and predict, what are the main features related to each? (A, B, C, D, E)
- Are important parameters missing which could be directly correlated to the issues? (C)
- Are there features available that should be discarded from the process? (A, C, D, E)

II. **Preparing Dataset**

- Can data categorization aid in the process of data prediction? (A, B, E)

III. **Result Analysis**

- Are there features available that should be discarded from the process? (A, C, D, E)

5 HFA Business Case

The present chapter corresponds to the demonstration and evaluation of the adopted methodology, a practical application of the knowledge and artefacts developed previously is conducted to the HFA company. It comprehends an in-depth analysis of the use case and project and the consequent application of the model and prediction algorithms.

5.1 Company Description

Henrique Fernando & Alves (HFA) is a Portuguese Small and Medium-sized Enterprise (SME), based in Agueda, that focuses on the assembly and testing of IT components. Created in 1995, the company established a goal to permanently enhance their equipment and technology to easily adapt to clients' needs, therefore, allowing the production of highly personalized products. Currently, it holds a solid position in four main sectors of industry: (1) Telecommunications, (2) Transport Industry, (3) General Electronics and (4) Repair/Maintenance.

In the sector of telecommunications, it is possible to see the strong influence of this company in the market. Through the usage of innovative methodologies and components, HFA clutches the assembly of a wide range of equipment currently being used in telecommunications. From network management systems and "Televisão Digital Terrestre" (TDT) to fibre optic cables for data transmission, it is possible to supply their client's a complete set of options. Furthermore, a repair service is provided in order to further aid their clients.

In terms of Transport Industry, the grasp of HFA allows the production a wide range of systems that can manage traffic information and public services, such as: automatic ticket sale system, passenger counting system and stop light management systems. Furthermore, it also provides highly customizable solutions for private clients, such as fleet management systems.

General Electronics is regarded as a massive sector of industry with a worldwide market. This domain can range from Telemetry, ATM systems and in-house equipment, or in other words, electronic equipment that can be used on a daily basis. With this line of thought, HFA plays an important role in this segment, mostly regarding the creation of systems, such as control access, video/audio for home cinema, illumination through LED technology and distribution of fuel. Above all products, one stands out, the production of automatic payment terminals used worldwide.

In terms of production, this company contains six advanced Surface-Mount Technology (SMT) lines that provide a fast approach to a wide variety of production requests from clients. Through the usage of automatic printers, solder paste inspections (SPI) and Pick and Place (P&P) machines, HFA can provide a ratio of 300.000 components produced per hour while maintaining a high level of quality. Currently, the company is in compliance with ISO 9001:2015 for quality management systems, IATF 16949:2016 regarding the automotive industry for technical specification and the IPC-A-610 and the IPC-711/21.

5.2 Production Layout

The current layout contains six production lines in a sequential H-Pattern that focus on the production of electronic components for telecommunications, with each varying regarding its equipment. The general production process layout is composed of the following:

Figure 11 - Production line six automatic feeder

1) **Automatic Feeder** – Machine responsible for feeding automatically, blank Printed Circuit Boards (PCB) to the printer. The cartridge is manually substituted once the current one is empty. The machine itself generates data that is not being extracted to the database. Nevertheless, no issues are registered that may cause downtime to the machine itself or the remaining ones;

Figure 12 - Production line six automatic printer

2) **Automatic Printer** – The automatic printer places soldering paste in the PCB through a widespread roller. This eases the process and allows to maintain constant flows of solder paste in each board. The paste used is refilled manually by an operator once the tray is below the required point. The average amount of data generated per day for the printer is approximately 300 rows, which constitutes 9300 rows per month approximately;

Figure 13 - Production line six SPI

3) **3D offline Solder Paste Inspection (SPI)** – It is responsible for addressing the output of the previous equipment. In other words, for each PCB, this machine checks the amount of solder paste applied. If the value is between the limits, then it is approved, otherwise, the PCB is rejected. In case it is rejected, the part is reutilized if possible, by extracting the paste deposited on the PCB and reinserting in the production line. The SPI does not generate production line downtime. In terms of data this machine generates 2200 rows per day approximately which constitutes 70000 rows per month;

Figure 14 - Production Line six Pick and Place

4) **Pick and Place (P&P)**– This machine regards the last step to complete the physical structure and components of the PCB. In other words, it is responsible for placing the components in the gaps where the solder paste is placed. The components are attained from cartridges manually placed bellow the machine. In terms of data, each machine generates per month over five million rows approximately;

Figure 15 - Production line six Reflow Oven

5) **Reflow Oven** – Responsible for culminating the tasks executed in production. In this checkpoint, the boards are placed in high temperatures to finish the production process and achieve the expected result, a finished PCB. This machine is vital to the process since each generated PCB must go through this oven within 30 minutes after the paste is laid. If this action does not take place it not possible to guarantee the physical integrity of the board;

Figure 16 - Production line six AOI

6) **Automated Optical Inspection (AOI)** – Is the last equipment in the production line. This equipment aims to analyse the product generated in the previous machines through direct comparison with a model previously loaded. With the usage of sensors and cameras, this machine analyses the PCB in question and compares the layout and component position with the generic successful output. If no discrepancies are found

and the PCB matches the model, then the process is terminated, and the board is ready to be subjected to the following stages.

In terms of data flow, not all data generated by the production lines are being stored. With recent updates in the production lines and heterogeneity of equipment, the processing and storage of data is still a difficult task to tackle. The current layout of information systems implemented in HFA regarding production is available in Figure 17.

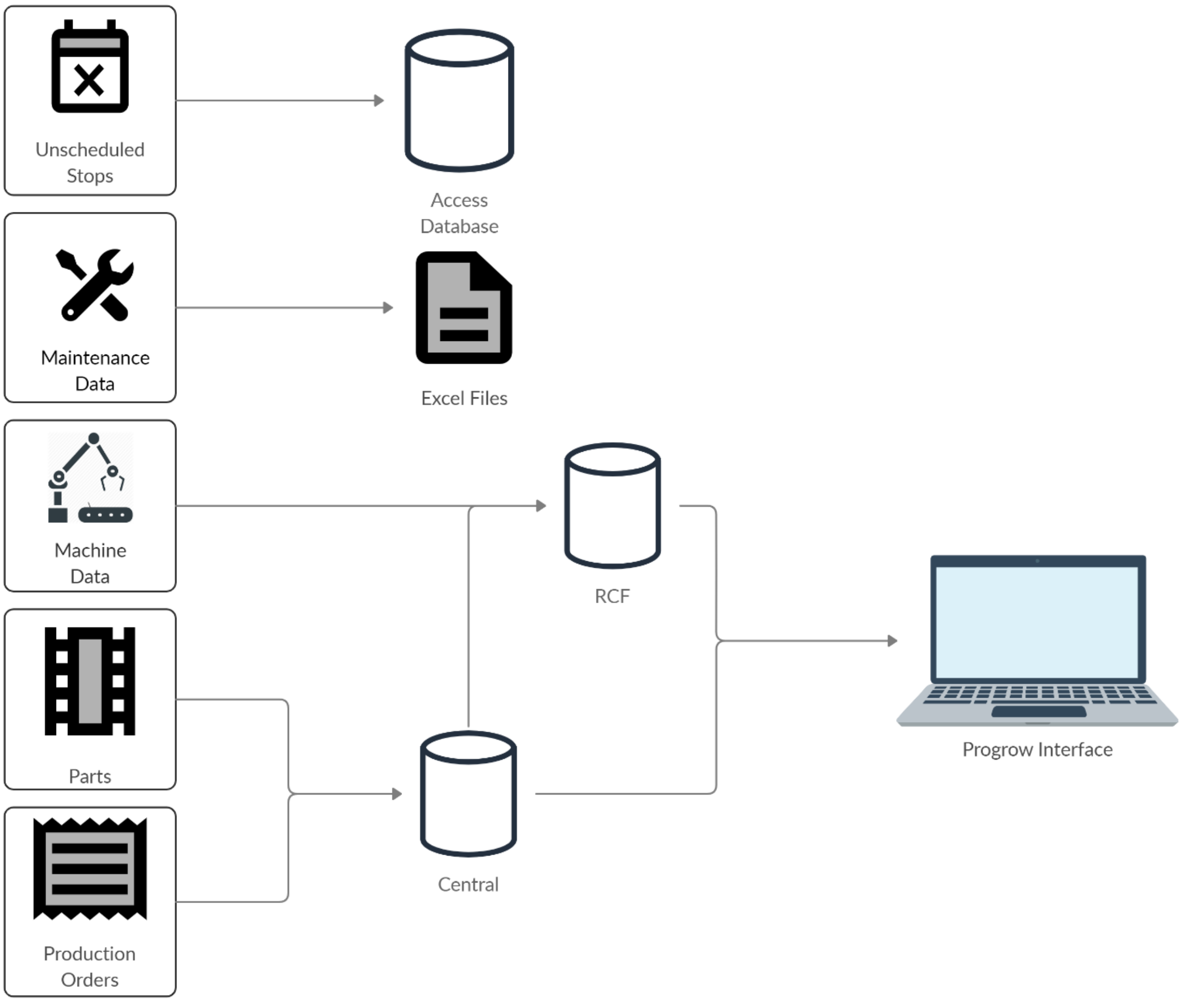

Figure 17 - Data flow of HFA production

In terms of data storage, multiple destinations are being used by HFA. The unscheduled stops in the production lines are registered in loco by the operators, which are then processed and stored in a custom Database (DB), Access Database. Regarding maintenance plans, executions and reports, these are registered in Excel files without being processed and stored in a database. The Central DB can be considered as the bridge between sales, production and logistic departments holding data regarding their daily basis of operations. The “Recolha Chão Fábrica” (RCF) is the main component of production lines data storage.

The “Recolha Chão Fábrica” (RCF) is the main component of production lines data storage where all the three types data regarding production are stored, which are: automatically generated and sent by machines, manually extracted by operators and logistic data. Across the execution of this dissertation, not all data generated by the machines in production lines were being processed and stored, whereas most were extracted and stored in files. Taking into consideration the amount of data available in these two last DB’s, HFA installed an interface to ease the access to relevant data, therefore decreasing overhead risk and mitigating possible bottlenecks. This interface, named Progrow, is mostly used by management, quality, R&D and logistics departments.

5.3 Project Origin

The project Demo Digital 4.0 was promoted by INOV and co-financed by “COMPETE 2020”. With the goal of further enhancing the potential of Industry 4.0 by providing important support to Portuguese SME in the industrial sector, this project started November 2017 with a duration of 24 months and segmented in four parts: (1) Reconversion of equipment for network integration, (2) Cyber-physical production systems, (3) Indoor/outdoor positioning and (4) Management and anticipation of failures in production line equipment. This last part was the sub-project for which HFA applied to further develop its production line, by developing a maintenance management solution. By implementing non-invasive sensors in equipment, it would be possible to further retain viable data for analysis, therefore, reducing the number of unscheduled stops, increasing efficiency and production line up-time.

Though an initial meeting with HFA and a tour of the facilities, production line six was selected, since it contained the newest and most advanced equipment in the factory while also maintaining the highest downtime from the eight lines. Below follows an in-depth analysis of this production line.

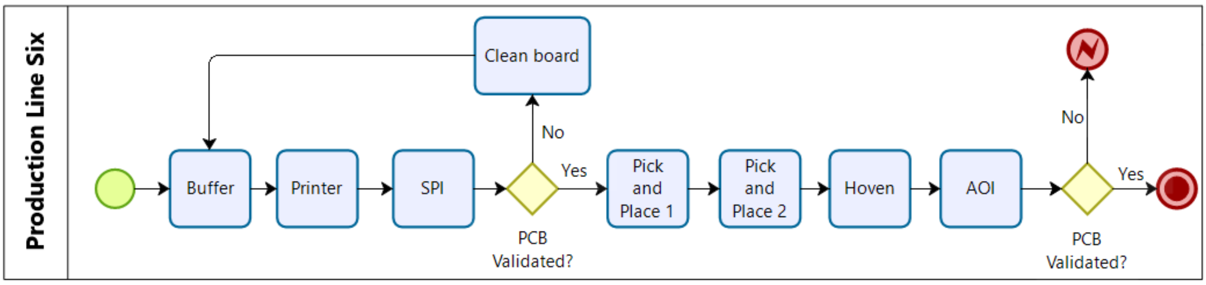

Figure 18 - Production line six layout

The production process starts with an automatic buffer that is responsible for supplying the printer with blank PCB boards. Once the content of this machine is extinguished, it is replaced manually by an operator surveying the production line. An example of a finished PCB is available in Figure 19.

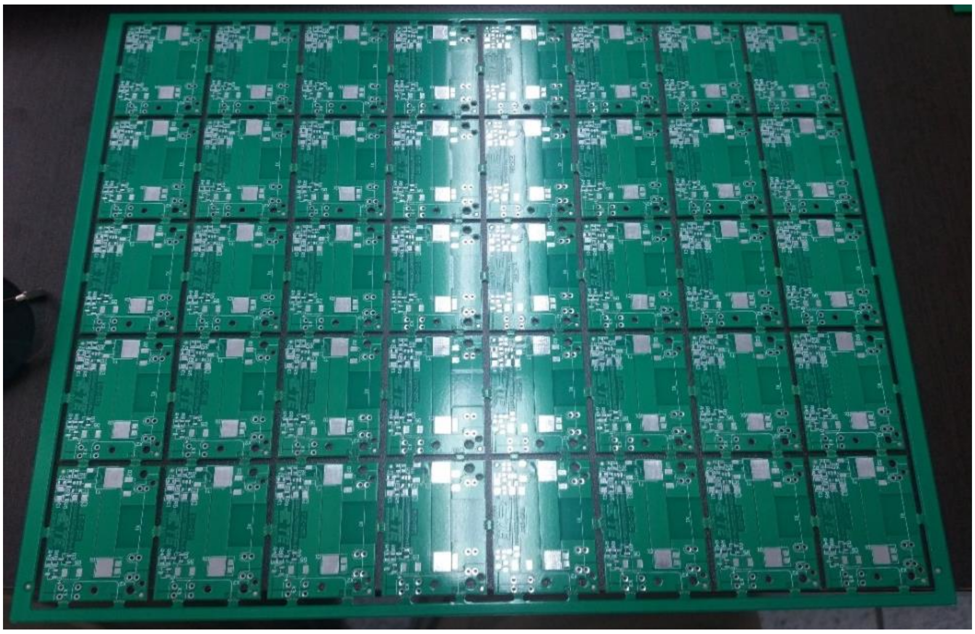

Figure 19 - Example of a finished Set of 40 PCB's

The production process follows with the automatic printer, named Versaprint S1, that applies solder paste to the blank boards (grey areas in the PCB). Before PCB production starts, a metal tray with the layout of the model to be produced is placed in the printer. With this tray, only the specific parts of the board where is expected are filled with the paste. To guarantee if the previous step was performed correctly, an SPI is used to detect the amount of solder paste used for each board. If a PCB does not place within the defined parameters set into the setup, its content (solder paste) is removed manually by an operator and the board is placed at the start of production. Once the board is approved by the system and the operator, then it proceeds automatically to the pick and place machines, Yamaha YSM20-2. These machines are responsible for the implementation of the components in the mid-developed boards, with an automated operation of placing the required components in the board. Consequently, the boards are then forwarded to the reflow oven to finalize the production process and perform a final validation with the AOI.

The production line six contains a total of seven machines to produce a board. Regarding the interviews conducted at HFA, the equipment itself and the goal of the project, only the two pick and place machines, the SPI and the printer were chosen for the scope of the project. The reason behind this decision is mostly since the main unscheduled stops registered are due to the printer and the pick and place machines. Consequently, these events directly affect customer orders and production efficiency which HFA considers a vital threat to business.

5.4 Model & Guidelines application

Bearing in mind the Information Systems layout in HFA and the analysis conducted to the production line, a concept solution was generated for the implementation of PM, which is available in Figure 20. This solution allows the execution of the artefact generated in chapter 4 with an in-depth analysis, which is the aim of this section. Regarding the guidelines generated, these were applied in a semi-structured interview conducted, which its content is available in Annex J.

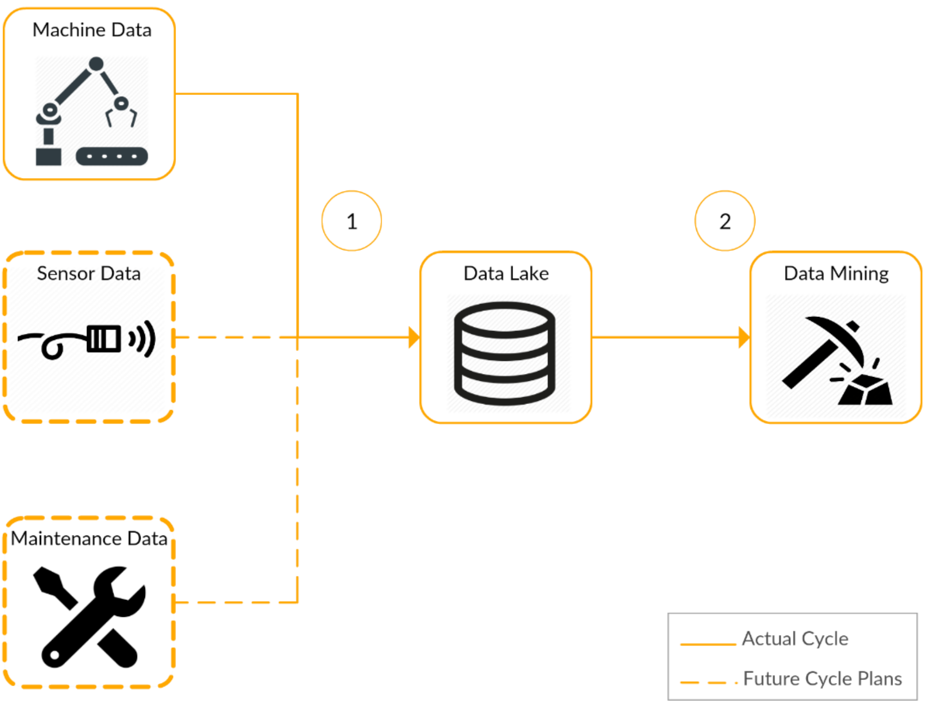

Figure 20 - Proposed Solution for HFA

The solution to be implemented, just like the artefact generated, is composed of two major steps: Data Extraction and Data Prediction. The first step can be compared to the first stage of the artefact, where data is extracted, analysed and stored correctly to generate a dataset for prediction. In this step the modules maintenance data and sensor data are in dotted lines, due to the fact that these weren't available to explore, still HFA plans to implement sensors and process maintenance data collected from daily operations. Consequent to this operation follow the Data Prediction step which is similar to the second stage of the artefact, where all the aspects related to the preparation of data are executed culminating in the training and execution of the model with a final result analysis. To execute an overlap of this proposed solution with the artefact developed, a step-by-step analysis is conducted in this section.

5.4.1 Situation Assessment

The production line chosen for this practical implementation does not contain sensors installed to retrieve control data. Through meetings, it was possible to understand future plans to implement sensors in the printer machine regarding the solder paste used. Therefore, all the data available for extraction resulted from machine operation. In terms of possibilities available and not forgetting the time limitations, it was agreed that the data extraction would have a time span of one month. The choice for a higher focus regarding the pick and place machines solely rested on the fact that these two were the only ones that generated data regarding failures and unscheduled stops. Besides the fact that the SPI placed in production line six does not register failures, it also does not allow data extraction while the machine is operating. This solely constitutes a major issue, since the production line works continuously.

5.4.2 Data Extraction and Analysis

From initial analysis and the interview conducted in the HFA Headquarters (Annex J), it was inferred that the four machines generate in a total of nine types of log files available in Table 17. Each of these Log files is generated under different circumstances and with different timespans. In other words, since these are automatically generated and updated in each machine, they must be extracted within a week from its creation to safeguard the data generated. The first data extraction contained data regarding the previous week of operations. To allow the creation of a continuous dataset for analysis, it was established to perform weekly extractions, preferentially at the beginning of the week (Mondays), to streamline the process of communication and data transfer in the project. This methodology could not produce more substantial advantages regarding the SPI machine since it does not allow for data extraction while the machine is in operations.

Table 17 - Machine's Files Descriptions

Machine	Name	Description	Periodicity[4]	Annex
P&P	Pcb_Log	File registers components implemented in each board	600 Files/Day	D
	Cart_Log	Registers carts ID's and positions	1 file/day	E
	Setup_Log	Comprises information of the parts used in PCB's	1 file/day	F
	Error_Log	Registers errors of the machine	1 file/day	G
	Lot_Log	Registers metrics of each setup	1 File/Board Type	H
	LotParts_Log	Registers the parts and time they were used for each board type	1 File/Board Type	I
SPI	Area_Result	Registers metrics regarding the solder paste inspected in PCB's	1 File/Extraction	A
	Board_Result	Registers evaluations and false detections of evaluated in PCB's	1 File/Extraction	B
	Category_List	Provides an overview of all types of boards analysed	1 File/Extraction	C

The files extracted from the printer contained data regarding the production of the machine in the format of ".txt". From the initial analysis of the data, it was possible to conclude that the machine did not register failures, therefore not being described in Table 17. Regarding the pick and place machines and the SPI, the files obtained were generated in the format ".csv". The cleaning and preparation of data were performed with a script in C# for automated file processing. Figure 21 presents the process generated to analyse the files. This process inputted the data file obtained and generated two files at the end of the script, an output file and a quarantine file.

[4] Numbers shown in column can vary.

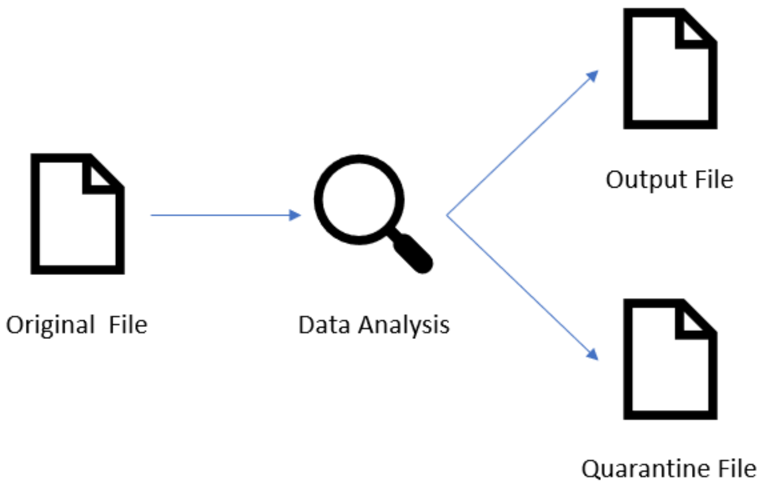

Figure 21 - File Analysis and output generated

Since the files maintained the same structure, any content deviation generated a trigger in the script. Consequently, it led to the entire row of data being sent to quarantine, therefore safeguarding the output file and the database for possible corruption due to unexpected data insertion. The quarantine file would be analysed at the end of the script executions to understand what data was generated incorrectly by the machine.

5.4.3 Data Storage

From the process executed in the previous sub-chapter, the content available in the output file is imported into a database for analysis. With data being retrieved from the machines in a weekly basis (exceptions occur depending on the file type), it was established to perform data integration in every Tuesday in order to decrease possible overheads of data input. Since performing the remaining steps requires data to be available, this cycle of data extraction, analysis and integration was iterated four times to provide a reliable source for analysis.

The process of data integration resembled Extract Transform Load (ETL) methodologies. The reason behind this statement relies on the fact that this methodology allows the system to perform a second analysis of the data regarding its content. In other words, by analysing the data in the output files before inserting in the database it is possible to understand incorrect data type insertion and therefore prevent possible errors. Therefore, quarantine tables were added in the database to hold incorrect values or rows from the files. Once the data is inserted correctly in the database, the last step is to assess the quarantine tables and understand why the rows were redirected.

5.4.4 Feature Selection

Once data is stored properly, and a timeframe is generated with enough delta for analysis, the following step is to breakdown the features by understanding which can provide a better result in terms of algorithm predictions. In this step, two datasets are analysed, from the two pick and place machines. The analysis conducted to the files allowed to achieve linkage between the ERR_LOG and LOT_LOG, through the timestamps. This relation further enhanced the dataset by understanding what type of board was being produced in each error registered.

From the original 54 features obtained through the linkage of the two datasets of P&P1 and 2 (dataset is available in Annex L), modifications were executed to discard features that did not uphold vital weight into the prediction stage. The first step executed was to understand if features matched to the ones described in chapter 4. Taking into consideration Figure 9, most of the common features are not available in the dataset, since the only ones available are **process time related** and **machine errors**. Features such as temperature, vibration and engine metrics were not available, since the types of equipment do not contain sensors to attain such information and engine metrics are not registered.

The second step regards the discard of irrelevant features. The features shown in Table 18 were discarded from both machine's data since they did not register information that could be used for prediction. Also, the feature OPERATOR_CALL_TIME was discarded only in the dataset of P&P2 since it only contains NA values. Further 18 features were discarded in both datasets because they regarded the split of START_DATE_TIME, SETUP_DATE_TIME and FINISH_DATE_TIME. The reason behind this action is due to the fact these features can uphold a viable weight in the Data Visualization stage. With this operation, the number of features available changed from 54 into 27 in P&P1 and 26 in P&P2.

Table 18 - Features initially discarded from datasets

Feature	Reason
ERROR_DETAIL	It did not provide useful information for prediction.
ERROR_CONTENT	The feature was discarded from dataset since it only provides a reason why error occurred.
OPERATOR_ID	Only the user is operating the machine. If admin is in logged in it only means that data was being extracted from the machine.
BOARD_COUNT_MAX	Value is always 0.
NG_BLOCKS	
NOZZLE_ERROR	
MOUNT_TABLE	Value is always “Table A”.
START_DATE_TIME	Features were used to generate a new feature, DateDiff, which regards the number of hours the set was at the function.
SETUP_DATE_TIME	
FINISH_DATE_TIME	
OPERATOR_CALL_TIME[5]	The feature only does not contain data, only NA’s.

In terms of feature selection, the methodology implemented was Feature Reduction. By analysing correlations between features, it was possible to understand which have a higher correlation. Nevertheless, since this methodology cannot be performed with non-numerical data, the features BOARD_NAME, DATE_TIME and ERROR_NUM were discarded in both datasets for **this specific stage of the process** which led to 24 features selected in P&P1 and 23 in P&P2 for feature reduction. Through the usage of this method it was possible to discard 15 features from P&P1 dataset due to their high correlation:

- PRODUCED_BOARDS
- MOUNT_CT_MAX
- TRANSFER_CT_AVE
- MOUNT_CT_AVE
- PARTS_CONSUMPTION
- MOUNT_CT_MIN
- WORKING_RATIO
- MOUNTED_BLOCKS
- NO_PARTS_ERROR
- RECOVERY_TIME
- OTHER_ERROR
- MOUNT_RATE
- MARKREC_CT_MIN
- DATE_DIFF
- TRANSFER_ERROR

Regarding the dataset of P&P2, nine features available below were discarded:

- TRANSFER_CT_AVE
- PRODUCED_BOARDS
- MOUNT_CT_AVE
- MARKREC_CT_AVE

[5] Feature was only discarded in Pick and Place 2.

- PICK_UP_ERROR
- MARKREC_CT_MIN
- MARK_VISION_ERROR
- OTHER_ERROR

5.4.5 Data Categorization

One feature was categorized for both datasets, BOARD_NAME, which regards the type of board being produced to be input into the model. Being in string format, it was categorized into a scale from 1 to 5 where each number represents a different board type produced.

5.4.6 Dataset and Model preparation

The dataset preparation is critical to the model execution since it holds the two most important elements, the training dataset and the test dataset. The original does not contain a vast amount of records, therefore a higher focus must be given into the algorithm’s training. In other words, the split ratio of the original dataset must favour the training dataset without compromising the final test. With the sizes being low for prediction, the first preparation executed was to discard all the errors types where the frequency was lower than two. Since these error types are of low frequency, it increases the difficulty for the models to understand its triggers. With a wide array of features available for input in both datasets, sub-datasets were generated for each machine. These contained four features each, three mutable and one immutable. The last operation performed was a data split with a ratio of 70%/30%. This split allows to obtain a considerable dataset to train the algorithm without compromising the test. Furthermore, from the test datasets, the feature ERROR_NUM was extracted, so it is possible to verify the accuracy of the model. With performance varying regarding the chosen algorithm, for this practical experiment three algorithms were chosen which allow to assess different possibilities: NB, SVM and Adaboost. The choice regarding these three algorithms relies on the fact that: NB allows to attain good conversion with low amounts of training data low training time; SVM due to its adaptability resulted from the usage of kernel allows to attain good results. Last but not least, the Adaboost was used due to its capabilities to boost results, though the usage of decision trees.

5.4.7 Result Analysis

From the execution of the algorithms, it was possible to extract the results available in Table 19 which shows the three best datasets and their results in each algorithm for the two P&P machines. Full results are available in Annex N (P&P1) and Annex O (P&P2).

Table 19 - Algorithm's results for P&P1 and 2

Machine	Features	Algorithm	Accuracy (%)
Pick and Place 1	**D1** TRANSFER_CT_MAX TRANSFER_CT_MIN STANDBY_CT_AVE	NB	20.67
		SVM	44.13
		Adaboost	43.02
	D4 PICK_UP_ERROR PARTS_VISION_ERROR MARK_VISION_ERROR	NB	13.41
		SVM	44.69
		Adaboost	46.93
	D7 OPERATOR_CALL_TIME MARKREC_CT_MAX MARKREC_CT_AVE	NB	07.82
		SVM	49.72
		Adaboost	50.84
Pick and Place 2	**D1** WORKING_RATIO MOUNT_RATE MOUNTED_BLOCKS	NB	20.00
		SVM	44.65
		Adaboost	44.65
	D5 TRANSFER_CT_MAX TRANSFER_CT_MIN STANDBY_CT_AVE	NB	26.98
		SVM	40.97
		Adaboost	43.72
	D10 RECOVERY_TIME NO_PARTS_ERROR MOUNT_CT_MAX	NB	35.81
		SVM	41.86
		Adaboost	41.86

Results obtained fluctuate regarding each dataset and model as expected since no sub-dataset is the same. Through the analysis it is possible to see that the algorithm with the lowest results overall was NB with 7% in P&P1 and 5% in P&P2. %). Regarding the SVM and Adaboost, the results obtained are similar. More specifically, for the datasets D1 and D10 in P&P2 the two models obtained the same accuracy. Nevertheless, it is possible to see that the best result was obtained through the usage of Adaboost in P&P1 (51% accuracy in D7) and P&P2 (44% accuracy in D1). Furthermore, the analsys of other metrics such as Kappa also suggest low adaptability of these algorithms, where results are inbetween 0.05 and 0.30, which are very low results, as it can be seen in Annex N and O. Taking into consideration these two metrics it is possible to understand a viable correlation between the accuracies obtained and their respective Kappas.

5.5 Continuous Information Lifecycle

Since the implementation of PM allows to produce a viable output of information regarding machine operation, the possibility arises to generate a an exemplary concept capable of creating a self-improvement loop. Therefore the concept solution provided in Figure 20 was revised in order to comprehend such features, being the result available in Figure 22.

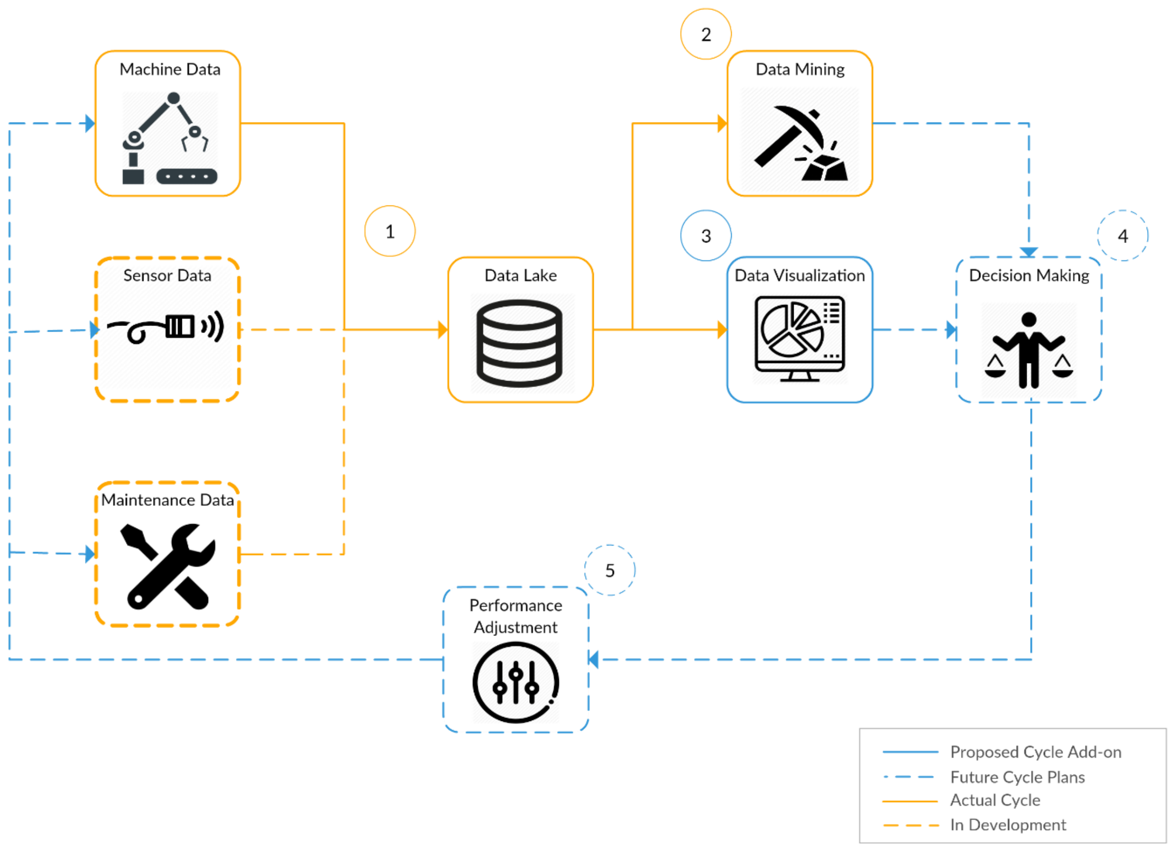

Figure 22 - Continuous Information Lifecycle for HFA

This concept contains in total 5 steps which allow to generate a self-improvement lifecycle opportunity for HFA. This is mostly due to the insertion of the steps **Data visualization (3)**, **Decision Making (4)** and **Performance Adjustment (5)**. Since the steps one and two have been previously described, they are not featured below.

With the execution of step one resulting in the storage of machine-related data, this serves as an input for steps two and three. Taking into consideration the amount of data generated, simplicity allows to provide proper knowledge, which can ease the decision-making process. This results in the production of **Data Visualization (3)** techniques which can provide an unbiased input regarding machine operation throughout customizable timespans. The choice for this module underlines the fact that information can be simplified and comprehended into a short output which can provide a different perspective on the choice of decision making.

The outputs generated in steps two and three consequently serve as a vital input in the consequent **Decision Making (4)** process. This is the major turning point in this loop. In other words, by providing the correct elements to the stakeholders, dashboards, error predictions and patterns, it is possible to understand the status of the production line. Consequently, by allowing this information to be directly embedded to the problem in hand, it is possible for the shareholders to suggest viable strategies without compromising future results.

The last step of this life cycle, **Performance Adjustment (5)**, allows to finetune this system by applying a digital to physical action, where the usage of digital information gathered is translated into physical action, which can carry according to issue or gravity. Consequently, this leads to new data being generated which allows to repeat the cycle and therefore further comprehend the machines in the production line.

As it can be seen in Figure 22, two major parts are perceivable, the actual cycle and Future Cycle Plans. The actual cycle was introduced previously, to which the artefact was implemented, therefore not constituting an element to review in this section. Regarding the **Future Cycle Plans**, **this part is suggested but not implemented, mostly due to time restrictions** which limit the possibility for proper implementation. With the goal of easing implementation, the Data Visualization step is analysed in-depth below with the suggestion of dashboards generated through the data obtained.

5.5.1 Dashboards

The capability to attain information on demand from production lines can constitute a major advantage in the act of deciding strategies. This section aims to provide dashboards generated with the data available for the two P&P machines. For the SPI, a dashboard was not designed, since not enough data and features were available.

With only two machines suffering unscheduled stops (P&P machines), these dashboards gave prominence on providing information regarding error distribution and important parameters of the production of PCB's. In Figure 23, it is possible to see the assembled dashboard of P&P1. With identical machines outputting the same attributes, the structure of the dashboard used for the P&P1 was replicated for the second machine. The dashboard of P&P2 is available in Figure 24.

The structure of these dashboards aims to provide a simple and clear view of main features to track in production line six operations. Therefore, the upper part of the dashboard focuses on the error distribution regarding production hours and weekdays. These visualization elements allow to understand the fluctuations in production and therefore ease board elements to understand the major gaps to pursue towards increasing efficiency. The lower part of the dashboard aims to show the major indicators of production such as Transfer time (Transfer_CT), Mount time (Mount_CT), board recognition time (Markrec_CT), amongst others. Since board types can have a direct influence on registered errors, this dashboard allows the user to customize the visualization elements according to specific dates and board types being produced. Table 20 provides an overview of the features used for these dashboards and their meaning.

Table 20 - Features used in P&P dashboards

Feature	Dashboard Feature	Meaning
Transfer_CT	• Tempo de Transferencia	Amount of time registered to transfer a PCB
Mount_CT	• Tempo de montagem	Time taken to mount a PCB
MarkRec_CT	• Tempo de reconhecimento	Recognition time of a PCB by the machine
Mount_Rate	• Taxa de montagem	Rate of production (PCB's produced successfully / all PCB's)
Produced Boards	• PCB's produzidas	Quantity of boards produced by the P&P machine
Working Ratio	• Taxa de Produção	Ratio of work delivered by the machine (Time producing / time on)
Error_Num	• Top Erros • Erros por dia de semana • Erros por Hora	Type of error registered by the P&P machine
Date	• Interacts with all fields	Typical feature to store a data timeline for the user
Board_Name	• Tipo de Board	Register the board type producedby the P&P Machine

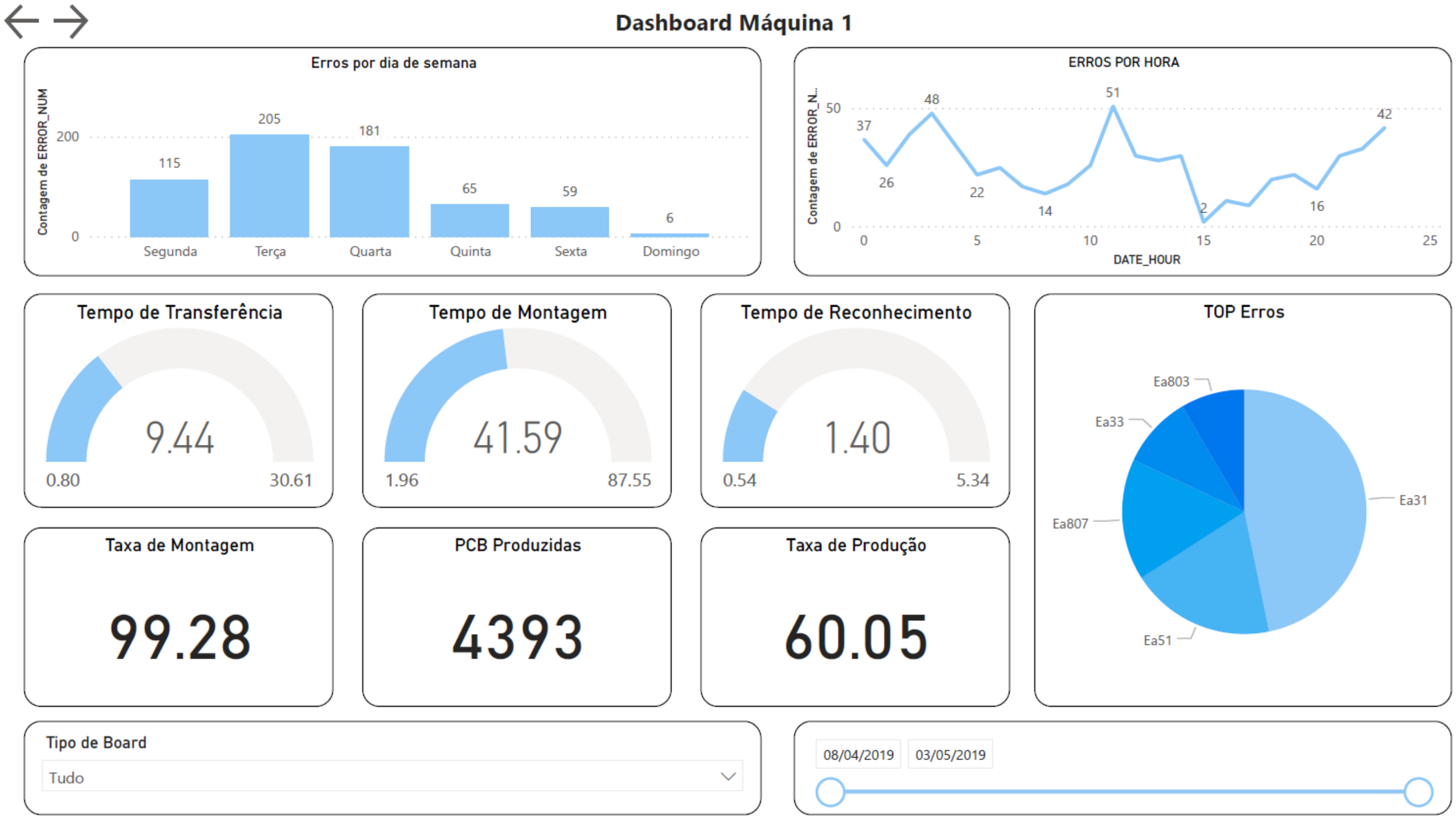

Figure 23 - Dashboard Pick and Place 1

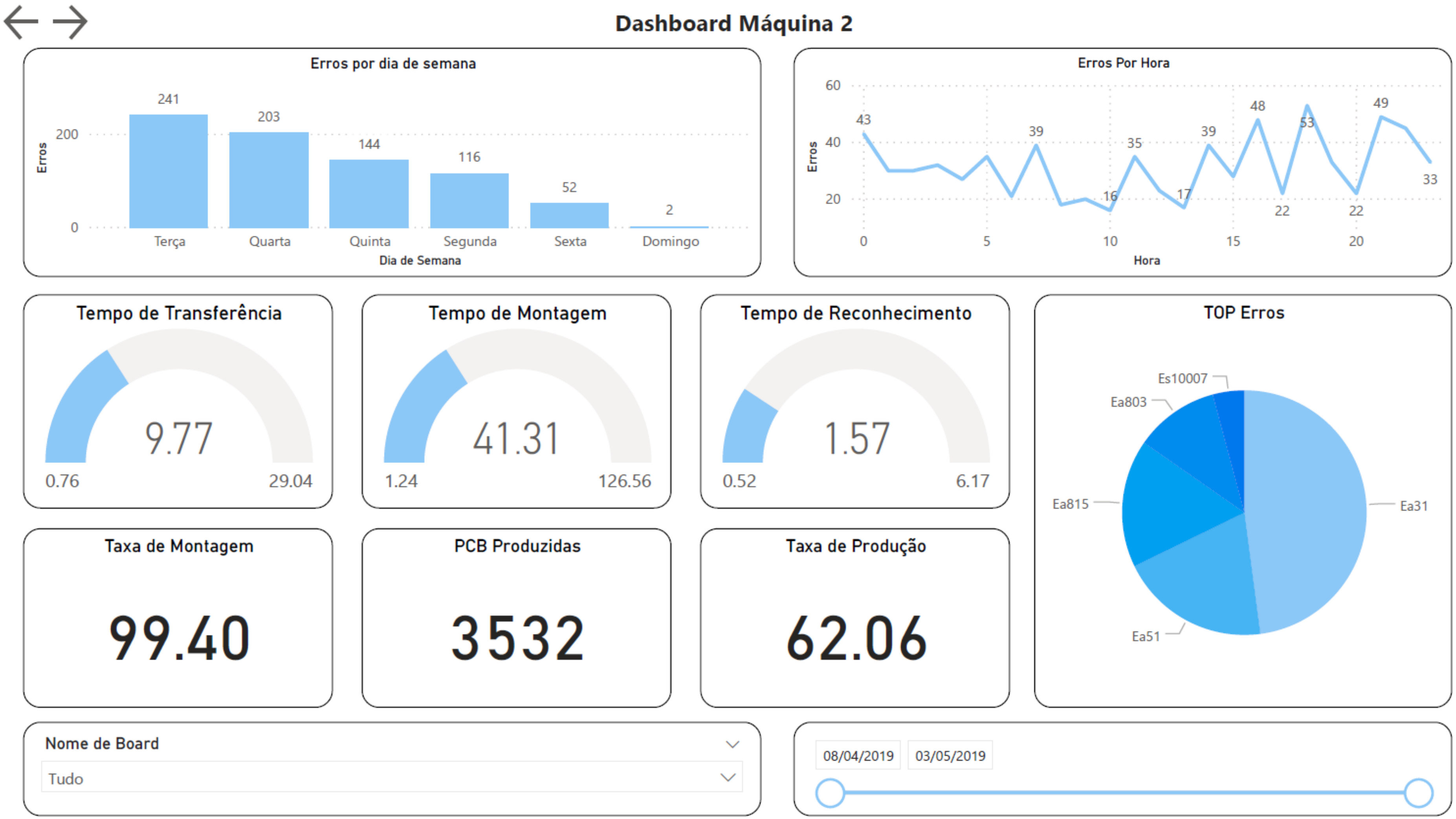

Figure 24 - Dashboard Pick and Place 2

6 Conclusions

This chapter is divided into two main sections: (1) Contributions and (2) Limitations and Future Work. Contributions deliver an overview of the work conducted throughout the previous chapters and aim to provide a concise answer to the research questions established in Chapter 1. The second section of this chapter delivers an analysis of the limitations that occurred throughout this dissertation that had a direct influence on the work and the consequent output. Furthermore, in light of the work developed, guidelines regarding future work are established with the purpose of solidifying the work developed.

6.1 Contributions

The developments performed in the industry allowed the introduction of wide mass sensor usage and CPS's. These, combined with proper information systems, allow companies to generate a continuous data pipeline to further understand their equipment. This joint venture perfectly defines the concept of the fourth industrial revolution, which established a turning point in industry to focus into a solidified data-driven approach. Completed with the re-introduction of PM as a maintenance methodology, this new path allows to extract knowledge from machine behaviour through the analysis of the pipeline previously described. With this approach, enterprises can extract higher quality information which establishes unprecedented decision-making capabilities. Nevertheless, the implementation of this methodology requires multi-factor coordination and a new mindset. Considering the wide spectrum of sectors in industry, this thesis aims to analyse this (new) methodology, with a specification in the prediction of unscheduled stops in equipment. Additionally a paper entitled "The aftermath of industry 4.0 in Small and Medium Enterprises" was created and submitted to the International Conference on Human-Computer Interaction, "INTERACT". The aim with this output was to analyse the effects of Industry 4.0 revolution in the Small and Medium Enterprises (J. Silva et al., 2019).

To establish a theoretical placement, the second chapter of this dissertation comprises a state of the art regarding the three most important concepts of this thesis: (1) Industry 4.0 and a walkthrough of its three main pillars (Big Data, IoT and CPS), (2) Predictive Maintenance and (3) Data Mining. Following, five PM cases, from different industry sectors, were presented and analysed to understand the possibility to establish a common path. From this analysis, it was possible to build an auxiliary model to be applied in a scenario of PM implementation. To complement this model, a set of guidelines were outlined, with regard to the processes and methodologies used in each case.

To show the utility of the artefact generated, a practical application was conducted to a Portuguese SME named Henrique, Fernando & Alves which focuses on the assembly and testing of IT components. With the objective of assessing the possibilities of implementing a PM methodology in this company's production line, a meeting and an interview were held having has as foundation the set of guidelines previously stated. Through the analysis of the information gathered, it was feasible to propose a solution of PM to the board, being comprised of the following elements: Data Harvest and Data Mining. Furthermore, to facilitate results interpretation, a set of dashboards were generated. Even though the presented steps stand-alone do not allow to extract practical results, the concepts of Decision Making and Performance Adjustment were introduced, with the purpose of being integrated, therefore allowing to generate a cycle of self-improvement.

With the implementation of the developed artefact, it was possible to understand the importance of the concepts related to this methodology. As described previously, Industry 4.0 is characterized mostly with the joint venture of its three main pillars, which mutually relate and generate a dependency. Through the literature review, it was possible to understand that this relation is not balanced, since the relevance of each to the implementation of PM derives. With the work conducted in the practical application of the model, it was possible to test the veracity of the previous statements. Since HFA does not contain any of the main pillars in its full, it consequently limits the application of the preferred methodology. Considering the first question of this thesis (Can Predictive Maintenance have a leading role, despite the main cornerstones of Industry 4.0 not being established?), it is possible to confirm that for the specific use case to which the artefact was applied, this methodology cannot have a leading role.This conclusion obtained mostly derives from the fact that the application of the artefact generated held as basis a set of unstable pillars which led to obtain low prediction results. Since the application of this methodology envolves a more stable basis of work, applying possible comparison of results simply was not possible.

The artefact creation took as foundations five study cases from distinct areas. As exposed in the previous question, the HFA case presented flaws in the structure of the three main pillars of Industry 4.0. Nevertheless, through the implementation of the model, it was possible to identify similarities between the features shown in the artefact and the ones identified for the case. Taking into consideration the second question of this thesis (Is it possible to achieve a generic model for the implementation of Predictive Maintenance in an Industry 4.0 era?) and the fact that the model was only applied to HFA, it is therefore not possible to conclude its veracity regarding the polyvalence of its application. The underlining of the artefact developed though can outdraw positive results, once applied to other sectors, yet the HFA case deemed insufficient to cover such mark. Taking into consideration the second question of this thesis (Is it possible to achieve a generic model for the implementation of Predictive Maintenance in an Industry 4.0 era?) and the execution of the model in the use case it is possible to affirm its executability. Yet, no expectations were generated with the execution of the model. Furthermore, implementation was not applied in the production line which does not allow to understand the consequences of the adoption of this methodology. Nonetheless, the underlining of the artefact developed though can outdraw positive results, once applied to other sectors, yet the HFA case deemed insufficient to cover such mark.

To contemplate the knowledge previously established, it is possible to state that historical and current data is vital regarding the prediction of events. Data and predictive models share a similar relation such as fuel to a car, or students to a university. In every single case it represents a crucial role in the consequent execution. However, this data-driven methodology introduced in Industry 4.0 should follow a gradual path in order to assure a correct layering of the pillars that underline this revolution. This gradual overlook should be considered upon its implementation since companies tend to maintain a mindset which is focused on results. This leads to a resilient state where a change of mindset is not established, mostly due to the fact that immediate results are not available. Therefore, this adaptation should be perceived as a ground-up project.

6.2 Limitations and Future Work

The work developed throughout this dissertation suffered multiple limitations being these characterized as functional and technical. From the functional side and considered highly relevant, was a change of the project manager in client-side at mid-project, which led to a reassessment of the objectives and a change in the data types extracted from the machines. Consequently, this led to a snowball effect due to the necessity of restructure and re-implementation of priorities.

From the technical side, the lack of control parameters, which can be obtained through the usage of sensors, directly affected the execution of the project, more specifically regarding the printer machine. Since this equipment solely depends on solder paste used in the production of a PCB and the machine itself did not register any errors, the only path that could be available would be the monitoring of such parameters and their effects in the PCB's. In terms of data extractions one machine which was directly affected was the SPI, mostly due to the impossibility to extract data while in operation. This constituted a critical limitation, since HFA machines operate on a 24/7 basis and the only time frame to perform extraction is when production is halted. Besides the fact that no metadata was available in the duration of the project, historical data was also unavailable since the files used were being extracted for the first time.

The artefact developed in this thesis takes as basis a set of use cases covering a wide array of industry sectors, with the goal of generating a common ground for the implementation of PM. By applying the model to a case where the main pillars of Industry 4.0 were not correctly established led to the incapability to confirm the model's veracity. Thus, this model should be applied in other areas of industry, more specifically those which are more Industry 4.0 ready.

The application of the artefact was performed strictly. Nevertheless, issues appeared regarding feature selection. This is mostly due to the fact that HFA machines did not possess sensors to directly monitor control parameters. This led to an inadequacy of the model regarding a case study where two of three main pillars of industry 4.0 are not properly defined. The results obtained can be improved with the usage of control parameters such as temperature, vibration, amongst others. Furthermore, this application can also allow a breakthrough regarding the implementation of PM for the printer due to the capability to assess solder paste condition. Through this monitoring and implementation, it is possible to maximize the process of replenishing the solder paste deposit without compromising this material's properties.

References

Chesworth, D. (2018). *Industry 4 . 0 Techniques as a Maintenance Strategy (A Review Paper)*. https://doi.org/10.13140/RG.2.2.18116.32644

Coleman, C., Damofaran, S., & Deuel, E. (2017). Predictive maintenance and the smart factory. In *Deloitte University Press*.

Colemen, C., Damodaran, S., ChandramoulinMahesh, & Deuel, E. (2017). Making maintenance smarter. In *Deloitte University Press*.

Cukier, K., & Mayer-Schoenberger, V. (2013). The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs*, *92*(3), 28–40. https://doi.org/10.2307/23526834

Efthymiou, K., Papakostas, N., Mourtzis, D., & Chryssolouris, G. (2012). On a predictive maintenance platform for production systems. *Procedia CIRP*, *3*(1), 221–226. https://doi.org/10.1016/j.procir.2012.07.039

Fernandes, M., Canito, A., Bolón-Canedo, V., Conceição, L., Praça, I., & Marreiros, G. (2019). Data analysis and feature selection for predictive maintenance: A case-study in the metallurgic industry. *International Journal of Information Management*, *46*(431), 252–262. https://doi.org/10.1016/j.ijinfomgt.2018.10.006

Ferreiro, S., Konde, E., Fernández, S., & Prado, A. (2016). INDUSTRY 4.0 : Predictive Intelligent Maintenance for Production Equipment. In *European Conference of the Prognostics and Health Management Society*.

Fluke Corporation. (2005). *The basics of predictive / preventive maintenance*.

Hänel, T., & Felden, C. (2017). Design and Evaluation of an Analytical Framework to Analyze and Control Production Processes. *Procedia CIRP*, *62*, 141–146. https://doi.org/10.1016/J.PROCIR.2016.06.052

Hees, A., Bayerl, C., Van Vuuren, B., Schutte, C. S. L., Braunreuther, S., & Reinhart, G. (2017). A Production Planning Method to Optimally Exploit the Potential of Reconfigurable Manufacturing Systems. *Procedia CIRP*, *62*, 181–186. https://doi.org/10.1016/J.PROCIR.2016.06.001

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information

Systems Research. In *Design Science in IS Research MIS Quarterly* (Vol. 28).

Holdowsky, J., Mahto, M., Raynor, M., & Cotteleer, M. (2014). Inside the Internet of Things (IoT). *Deloitte University Press*, *71*(2), 15.

Jenke, R. (2018). Successful data applications : a cross-industry approach for conceptual planning. *Journal of Business Chemistry*, *15*(2), 71–77. https://doi.org/10.17879/38129709602

Jones, A. T., Romero, D., & Wuest, T. (2018). Modeling agents as joint cognitive systems in smart manufacturing systems. *Manufacturing Letters*, *17*, 6–8. https://doi.org/10.1016/j.mfglet.2018.06.002

Kampker, A., Heimes, H., Bührer, U., Lienemann, C., & Krotil, S. (2018). Enabling Data Analytics in Large Scale Manufacturing. *Procedia Manufacturing*, *24*, 120–127. https://doi.org/10.1016/J.PROMFG.2018.06.017

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit : the definitive guide to dimensional modeling.* (3º Edition). Wiley.

Kusiak, A., & Li, W. (2011). The prediction and diagnosis of wind turbine faults. *Renewable Energy*, *36*(1), 16–23. https://doi.org/10.1016/j.renene.2010.05.014

Lau, R. Y. K., Zhao, J. L., Chen, G., & Guo, X. (2016, December 1). Big data commerce. *Information and Management*, *53*(8), 929–933. https://doi.org/10.1016/j.im.2016.07.008

Lee, J., Ardakani, H. D., Yang, S., & Bagheri, B. (2015). Industrial Big Data Analytics and Cyber-physical Systems for Future Maintenance & Service Innovation. *Procedia CIRP*, *38*, 3–7. https://doi.org/10.1016/J.PROCIR.2015.08.026

Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical Systems and Signal Processing*, *42*(1–2), 314–334. https://doi.org/10.1016/J.YMSSP.2013.06.004

Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., & Hampapur, A. (2014). Improving rail network velocity: A machine learning approach to predictive maintenance. *Transportation Research Part C: Emerging Technologies*, *45*, 17–26. https://doi.org/10.1016/J.TRC.2014.04.013

Manyika, J., Chui, M., Bisson, P., Dobbs, R., Bughin, J., & Aharon, D. (2015). The Internet of Things : Mapping the Value Beyond the Hype. In *Mckinsey global institute*.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., … McKinsey Global Institute. (2011). *Big Data : The next frontier for innovation, competition, and Productivity.* McKinsey Global Institute.

Marr, B. (2016). *Big data in practice : how 45 successful companies used big data analytics to deliver extraordinary*. Wiley.

Mehrabi, M. G., Ulsoy, A. G., Koren, Y., & Heytler, P. (2002). Trends and perspectives in flexible and reconfigurable manufacturing systems. *Journal of Intelligent Manufacturing*, *13*(2), 135–146. https://doi.org/10.1023/A:1014536330551

Meissner, H., Ilsen, R., & Aurich, J. C. (2017). Analysis of Control Architectures in the Context of Industry 4.0. *Procedia CIRP*, *62*, 165–169. https://doi.org/10.1016/j.procir.2016.06.113

Özbebek Tunç, A., & Aslan, P. (2019). *Business Management and Communication Perspectives in Industry 4.0* (A. Özbebek Tunç & P. Aslan, eds.). https://doi.org/10.4018/978-1-5225-9416-1

Park, C., Moon, D., Do, N., & Bae, S. M. (2016). A predictive maintenance approach based on real-time internal parameter monitoring. *The International Journal of Advanced Manufacturing Technology*, *85*(1–4), 623–632. https://doi.org/10.1007/s00170-015-7981-6

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pham, D. T., & Afify, A. A. (2005). Machine-learning techniques and their applications in manufacturing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, *219*(5), 395–412. https://doi.org/10.1243/095440505X32274

Qinming, L., Ming, D., Chen, F. F., Wenyuan, L., & Chunming, Y. (2018). Single-machine-based joint optimization of predictive maintenance planning and production scheduling. *Robotics and Computer-Integrated Manufacturing*, *51*(September 2018), 238–247.

https://doi.org/10.1016/j.rcim.2018.09.007

Rojko, A. (2017). Industry 4.0 Concept: Background and Overview. *International Journal of Interactive Mobile Technologies*, *11*(5), 77–90. https://doi.org/10.1111/j.1600-065X.1972.tb00044.x

Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., & Harnisch, M. (2015). *Industry 4.0 The Future of Productivity and Growth in Manufacturing Industries*.

Russom, P. (2011). Big data analytics - TDWI Best Practices Report. In *TDWI Best Practices Report, Fourth Quarter*.

Sakib, N., & Wuest, T. (2018). Challenges and Opportunities of Condition-based Predictive Maintenance: A Review. *Procedia CIRP*, *78*, 267–272. https://doi.org/10.1016/j.procir.2018.08.318

Schmidt, B., & Wang, L. (2018). Predictive Maintenance of Machine Tool Linear Axes: A Case from Manufacturing Industry. *Procedia Manufacturing*, *17*, 118–125. https://doi.org/10.1016/J.PROMFG.2018.10.022

Schuh, G., & Blum, M. (2016). Design of a data structure for the order processing as a basis for data analytics methods. *2016 Portland International Conference on Management of Engineering and Technology (PICMET)*, 2164–2169. https://doi.org/10.1109/PICMET.2016.7806715

Silva, J., Ferreira, J. C., & Gonçalves, F. (2019). The “aftermath“ of Industry 4.0 in Small and Medium Enterprises. *INTERACT*, 5.

Silva, V., & Pederiva, R. (2013). Fault detection in induction motors based on artificial intelligence. *International Conference on Surveillance*, 1–15.

Uhlmann, E., Pontes, R. P., Geisert, C., & Hohwieler, E. (2018). Cluster identification of sensor data for predictive maintenance in a Selective Laser Melting machine tool. *Procedia Manufacturing*, *24*, 60–65. https://doi.org/10.1016/J.PROMFG.2018.06.009

Uhlmann, E., Pontes, R. P., Laghmouchi, A., & Bergmann, A. (2017). Intelligent Pattern Recognition of a SLM Machine Process and Sensor Data. *Procedia CIRP*, *62*, 464–469. https://doi.org/10.1016/j.procir.2016.06.060

Wadhwa, R. S. (2012). Flexibility in manufacturing automation: A living lab case study of

Norwegian metalcasting SMEs. *Journal of Manufacturing Systems*, *31*(4), 444–454. https://doi.org/10.1016/J.JMSY.2012.07.008

Annex

Annex A

Table 21 - SPI_Traceability_AreaResultList

Variables	Example
Area No.	1
Panel No.	1
Group Name	G-0
Package Name	DefaultPackage
Parts Name	DefaultParts
Circuit Name	DefaultCircuit
Machine judge	Volume too large (warning)
Manual judge	Untested
Area (mm2)	1
Area (%)	116
Gap area ratio (%)	0
Center Gravity Gap (mm)	0.01
Volume (mm3)	0.106
Volume (%)	128
Average height (mm)	0.106
Average height (%)	110
Cross-section (mm2)	0.538
Cross-section (%)	97
Projection (mm2)	0.002
Projection (%)	0
Peak height (mm)	0.16647
Peak height (%)	33

Annex B

Table 22 - SPI_Traceability_BoardResultList

Variável	Exemplo de ouput
Board Index	1
Lot name	0
Lane	Normal
Target inspection machine	10432
Board ID (Side A)	1827177861
Board ID (Side B)	-
Panel No.	-
Insp sequence scale	1531
Board No.	1531
Machine judge	NG
Manual judge	OK
Defect count	0
False Detection Count	107
False Detection Rate	1.91
Insp Start DateTime	10/16/2018 11:43:14 PM
Insp End DateTime	10/16/2018 11:43:27 PM
Operator ID	spi
Manual Operator ID	-
Solder Name	-
Solder Rod	-
Squeegee Name	-
Squeegee Front Rear	-
Metal mask Name	-
Backup jig	-
Board Name	-
Board Lot	-

Annex C

Table 23 - SPI_Traceability_CategoryContainer

Variável	Exemplo de ouput
Group No.	1
Group Name	G-0
Defect area count/Total area count	0 / 2513

Annex D

Table 24 – P&P PcbLog

Variável	Exemplo de ouput
Machine Serial	Y41612
Machine Model	YSM20
Machine Name	MACHINE_NAME
Board Name	57513283_V5_29_BOT_L
Sequential No	198
Board ID	NULL
Date Time	07/04/2019 23:59
Mount Table	Table B
Block Num	0
Not Mounted	0
Mount Num	10
Silk Name	R387
Parts Num	45
Head Num	10
Parts Name	PTI1700075724
Parts Comment	75R_5%_0402
Library Name	NULL
Feeder Type	8mm1005cmp
Set Num	115
Parts ID	PTI1700075724
Parts Lot ID	NULL
Reel ID	DUMMY_02E6FF063C34EF00000DA8000
Feeder ID	ZSY-008-0081285A
Item17	NULL
Item18	NULL
Item19	NULL
Item20	NULL
Item21	NULL
Item22	NULL

Item23	NULL
Item24	NULL
Item25	NULL
Item26	NULL
Item27	NULL
Item28	NULL
Item29	NULL
Item30	NULL
Bincode	NULL

Annex E

Table 25 – P&P CartLog

Variável	Exemplo de ouput
Machine Serial	Y41612
Machine Model	YSM20
Machine Name	MACHINE_NAME
Date Time	08/04/2019 01:24:38
Operator_ID	_DEFAULT_OPERATOR
Operation	1
Cart/Magazine Type	NULL
Cart/Magazine Set Pos	Front Right Side
Cart/Magazine ID	CTZ1$32$1732813$12

Annex F

Table 26 – P&P SetupLog

Variável	Exemplo de ouput
Machine Serial	Y41612
Machine Model	YSM20
Machine Name	MACHINE_NAME
Date Time	08/04/2019 00:45
Operator ID	_DEFAULT_OPERATOR
Board Name	NULL
Operation	0
Set Num	24
Feeder Type	8mmTape
Feeder ID	ZSY-008-0081271A
Parts Name	PTI1700071369
Parts Comment	10K_1%_0402
Parts ID	PTI1700071369
Parts Lot ID	NULL
Reel ID	DUMMY_02E69809AFF42600000EF8000
Parts Remain Count	1248
Parts Num	44
Old Parts Name	NULL
Old Parts Comment	NULL
Old Parts ID	NULL
Old Parts Lot ID	NULL
Old Reel ID	NULL
Old Parts Remain Count	0
Old Parts Num	0
Remarks	Null

Annex G

Table 27 – P&P ErrorLog

Variável	**Exemplo de ouput**
Machine Serial	Y41612
Machine Model	YSM20
Machine Name	MACHINE_NAME
Date Time	08/04/2019 01:24:38
Operator ID	Default Operator
Error Num	O30100
Error Contents	Feeding not completed.
Error Detail	/TableA/SetNums21

Annex H

Table 28 - P&P LotLog

Variável	Exemplo de ouput
Machine Serial	Y41612
Machine Model	YSM20
Machine Name	MACHINE_NAME
Board Name	57513283_V5_29_BOT_L
Start DateTime	07/04/2019 20:19
Setup DateTime	07/04/2019 20:20
Finish DateTime	08/04/2019 05:17
Board Count Max	0
Produced Boards	648
Working Ratio	98.6
Mount Rate	100
NG Blocks	0
Mounted Blocks	648
Parts Consumption	164658
Transfer CT Max	15.7
Transfer CT Min	7.13
Transfer CT Ave	8.37
Standby CT Ave	25.7
Pick Up Error	60
Parts Vision Error	6
Mark Vision Error	4
Transfer Error	0
Other Error	4
Operator Call Time	316.9
Recovery Time	152.24
Nozzle Error	0
No Parts Error	0
Mount Table	Table B
Mount CT Max	43.07
Mount CT Min	27.18
Mount CT Ave	27.46

MarkRec CT Max	4.17
MarkRec CT Min	0.99
MarkRec CT Ave	1.07

Annex I

Table 29 – P&P LotPartsLog

Variável	Exemplo de ouput
Machine Serial	Y41612
Machine Model	YSM20
Machine Name	MACHINE_NAME
Board Name	57513283_V5_29_BOT_L
Parts Num	1
Parts Name	PTI1700134728
Parts Comment	1K65_1%_0402
Parts ID	PTI1700134728
Parts Lot ID	-
Reel ID	DUMMY_02C3E2069C7DEA00000DC4000
Start Date Time	07/04/2019 20:20
Finish Date Time	08/04/2019 05:17
Library Name	
Feeder Type	8mm1005cmp
Feeder ID	ZSY-008-0097185A
Parts Consumption	648
Mount Rate	100
Mount Table	Table A
Set Num	6
Pick Error Counter	0
Pick Error Rate	0
Vision Error Counter	0
Vision Error Rate	0
Nozzle Error Counter	0
Nozzle Error Rate	0
No Parts Counter	0

Annex J

In this Annex is available the answers of the interview conducted to Luís Moreira, responsible for Research and Development unit at HFA.

1. Quantas linhas de produção existem na fábrica?

Relacionadas com o projeto existem 2 linhas, as linhas 6 e 4. No total na fábrica existem 8 linhas de produção. As linhas 1,2,3,4 e 6 são iguais.

2. Que tipos de manutenção estão a ser praticas nas linhas de produção?

Corretivas. Quando existe avaria, ocorre intervenção da manutenção. Existe muito pouca manutenção preventiva. Maioritariamente são limpezas implementadas tendo em conta experiências anteriores.

3. Em média quantas paragens existem em cada linha?

É complicado de responder, existem muitas paragens a nível diário.

4. As manutenções das máquinas da linha 6 são efetuadas de quanto em quanto tempo?

Em termos de manutenção preventiva é efetuada uma limpeza todas as semanas. De 15 em 15 dias ocorre a troca de peças consumíveis (tubos, peças, óleo) e limpeza de filtros de ar. Isto para as *pick and place*. Para as outras é igual. Na *printer*, ocorre a troca do rolo no final de cada turno (3 vezes por dia). O rolo é limpo automaticamente pela máquina e depois é substituído.

5. Quais os principais motivos/razões para implementação de PM?

Porque existem bastantes avarias. Existem avarias que provocam paragens grandes e estas linhas fulcrais. No caso de se parar uma linha durante a semana pode levar a que seja necessário repor o tempo perdido no sábado. Considera-se isto como o cliché das empresas que leva a que seja um dos maiores motivos para querermos implementar PM

6. As avarias que foram referidas na pergunta 5 estão a ser atualmente registadas e armazenadas?

Sim, atualmente quando ocorre um erro que envolve paragem de produção, todos os detalhes são registados.

7. Que importância tem a PM dentro da HFA? Que impacto (positivo ou negativo pode acarretar para a empresa)

Positivamente, pode-se ganhar a velocidade com que se consegue tratar um problema e também a capacidade de perceber problemas antes de aparecerem através do estabelecimento de medidas proactivas para a garantir o máximo de durabilidade de cada peça sem pôr em causa a performance da máquina.

8. Já existiu algum projeto de passagem para manutenção preditiva ou instalação de sensores?

Este projeto com o INOV foi o primeiro passo que a HFA decidiu dar relativamente à passagem para a Manutenção preditiva

9. Existe uma concordância entre a gestão de topo e o setor de manutenção relativamente à implementação de PM? De que forma?

A administração apoia este tipo de desenvolvimento. Por outro lado, prazos a cumprir levam a que estes projetos não sejam o foco principal da empresa.

10. Como é efetuada a extração de dados das máquinas por parte da HFA?

A nível de dados, as máquinas geram e enviam os dados automaticamente para o servidor, apesar de haver diferenças entre eles (dados). Estamos a retirá-los, mas não estamos a analisá-los. Dai este projeto (Demo Digital) para também começarmos a perceber como compreender quais os dados mais relevantes para nós.

Depois temos dados que retiramos ao longo do dia, como a velocidade da máquina (tempo de ciclo para fazer uma placa), paragens e qualidade (taxa de sucesso). Dependendo que é pretendido, retira-se os dados em questão, mas ainda não estão a ser analisados os todos os dados com o objetivo de determinar padrões das máquinas.

11. Quais os fatores que considera mais relevantes na implementação de PM na linha 6?

Por exemplo a alteração do padrão da velocidade da máquina, porque se por algum motivo não sofreu input e reduziu a velocidade, alguma coisa se passa. Estou a falar na velocidade porque no fundo é o “sumo” da máquina ou no nível de vezes que ela rejeita um componente. Isto para todas as máquinas. Como é obvio o mais importante é perceber qual a que baixou.

Temperatura é um facto. Temperatura da máquina e ambiente e no caso da *printer* a temperatura da pasta que influencia diretamente a soldadura. Mas a temperatura da máquina é extremamente importante na *printer*. Existe especial atenção na *printer*.

Taxa de rejeição e rotações do motor são também duas variáveis bastante relevantes para este processo.

12. Já existe algum repositório central de dados recolhidos das máquinas, manutenção e outras informações?

As linhas 4 e 6 estão a ser retiradas automaticamente. As restantes linhas pretende-se seguir o mesmo caminho, com a exceção de uma das linhas que é mais antiga e como tal é bastante difícil conseguir extrair dados, quer de forma mais automática, quer de forma manual.

Toda a informação que é extraída das máquinas é depositada num servidor. Não existe tratamento de dados, a informação é toda colocada em cru.

Os dados da manutenção são guardados noutro local que não é o servidor das máquinas. São todos guardados sobre o formato de EXCEL e PDF.

Estas 3 fontes de informação são digitais.

13. Existem sensores implementados nas linhas de produção?

Algumas máquinas têm sensores de temperatura outras não. Uma das *printers* já tem informação de temperatura, mas ainda não é possível extrair. A máquina em questão é da linha 3. Neste momento existe um projeto com intuito de instalar sensores na fábrica e máquinas. Os sensores foram feitos por nós para recolher informação relativamente a temperatura e humidade. Para lá das máquinas, a sala onde é armazenada a massa usada para a *printer* tem temperatura controlada. Nem todas as máquinas têm porque o sensor é um extra relativamente ao produto base e como tal não foi possível arranjar sensores para todas as máquinas no ato de compra.

14. Tendo em conta o sistema de produção instalado quais as principais vertentes/variáveis a ter em conta para cada linha?

Todas as variáveis indicadas anteriormente podem pôr em causa a qualidade do produto.

15. Quais são os próximos passos da HFA relativamente à passagem para PM?

Primeiro, mudar a mentalidade de todos os colaboradores da HFA. No final do dia o principal pensamento dos colaboradores é produzir e manter níveis de produção com qualidade. Mas se for possível parar uma hora por semana para realizar ações de manutenção é preferível de modo a ter que evitar parar 1 ou 2 turnos por mês.

O segundo passo consiste na velocidade de atuação. A máquina está a demonstrar um problema e por base a velocidade de resposta aos indícios tende a demorar ao ponto de o problema ocorrer.

16. Todas as linhas de produção contêm SPI's para validação das placas?

Só as linhas 4 e 6 têm validação. A linha 2 vai passar a ter SPI e existe uma terceira neste momento em formato *stand-alone* utilizada para testar qualidade de amostras.

17. Porque é que não se optou pela utilização de buffers nos intervalos de cada máquina?

Primeiro iria requer mais espaço e mão de obra, visto que no formato atual, apenas é necessário um operador para controlar a linha toda. Em segundo ponto é necessário armazenar placas para depois serem utilizados. Visto que uma placa depois de impressa na *printer* deve passar no forno até 30 min depois.

18. Uniformização de números de série de cada placa é um fator relevante para a aplicação de PM?

Cada placa tem um número de série único, mas infelizmente não é possível para as máquinas lerem os números de séries.

19. Existe interação humana direta nas linhas de produção? De que forma?

Não, apenas existe nos casos em que se evidenciam problemas com a linha.

Annex K

Element	Description	Origin
Serial Number	Machine Serial Number	Machine Protocol
Control Software Version	Version of the machine's software	Machine Protocol
Machine Model Number	Machine's model number	Machine Protocol
Tool Changes (total)	Number of times a tool was changed since the machine was first powered on	Machine Protocol
Tool Number in Use	Turret station number currently in use	Machine Protocol
Dry Run	Indicates if the machine is running a program without producing a part	Machine Protocol
Power-On Time (total)	Time since the machine was powered on	Machine Protocol
Motion Time (total)	Time the machine is in motion	Machine Protocol
Last Cycle Time	Last production cycle time	Machine Protocol
Previous Cycle Time	Previous production cycle time	Machine Protocol
M30 Parts Counter #1	Counts the number of times a program completes.	Machine Protocol
M30 Parts Counter #2	Counts the number of times a program completes.	Machine Protocol
Maximum axis loads for X, Y, Z, A, B, C, U, V, W, T	Maximum load an axis has achieved since the machine was powered on	Machine Protocol
Coolant Level	Cutting emulsion level	Machine Protocol
Spindle load with Haas vector drive	Spindle load	Machine Protocol
Present part timer	Effective production time for the part currently in production	Machine Protocol
Last complete part timer	Effective production time for the part previously completed	Machine Protocol
Tool in spindle	Turret station number currently in use	Machine Protocol
Spindle RPM	Spindle rotation speed	Machine Protocol
Present machine coordinate position X, Y, Z, A, B	Current machine position for axes X, Y, Z, A, B	Machine Protocol
Present work coordinate position X, Y, Z, A, B	Position of the part at the start of production in axes X, Y, Z, A, B	Machine Protocol
Present Tool offset X, Y, Z, A, B	Distance of the tool relative to the origin in axes X, Y, Z, A, B	Machine Protocol
Machine Vibration X, Y, Z	Vibration during the cutting process on axes X, Y, Z	Sensor
Noise	Noise inside the machine	Sensor

Figure 25 - Features extracted from Case Study 3

Annex L

Table 30 - Joint Dataset (ERR_LOG and LOT_LOG)

Variável	Exemplo de ouput
Board Name	57513283_V5_29_BOT_L
Start DateTime	07/04/2019 20:19
Setup DateTime	07/04/2019 20:20
Finish DateTime	08/04/2019 05:17
Board Count Max	0
Produced Boards	648
Working Ratio	98.6
Mount Rate	100
NG Blocks	0
Mounted Blocks	648
Parts Consumption	164658
Transfer CT Max	15.7
Transfer CT Min	7.13
Transfer CT Ave	8.37
Standby CT Ave	25.7
Pick Up Error	60
Parts Vision Error	6
Mark Vision Error	4
Transfer Error	0
Other Error	4
Operator Call Time	316.9
Recovery Time	152.24
Nozzle Error	0
No Parts Error	0
Mount Table	Table B
Mount CT Max	43.07
Mount CT Min	27.18
Mount CT Ave	27.46

MarkRec CT Max	4.17
MarkRec CT Min	0.99
MarkRec CT Ave	1.07
Date Time	08/04/2019 01:24:38
Operator ID	Default Operator
Error Num	O30100
Error Contents	Feeding not completed.
Error Detail	/TableA/SetNums21

Annex M

Rule r1. $\forall P_{wp_i} \in Wp,\ i \in \{1,\ 2,\ \ldots,n\}$, *whenever* $lcpt < pt_i$ *then* $pt_i = lcpt$

Let us now define a rule that triggers an alert message anytime a part takes longer to produce than expected:

Rule r2. $\forall P_{wp_i} \in Wp,\ i \in \{1,\ 2,\ \ldots,n\}$, *whenever* $lcpt > pt_i$ *then send message* φ_i

$$= \left\{ \begin{matrix} id_{\varphi_j},\ m_i,\ [\textit{machine operator},\ \textit{production manager}],\ \text{"The current part took longer to} \\ \text{produce than expected. "},\ [\textit{last complete part timer},\ pt_i],\ \textit{Alert} \end{matrix} \right\}$$

The following rule causes the relevant people to be notified whenever the production of a part doesn't end successfully.

Rule r3. $\forall$ *time instant t*, *whenever* $ppt_t = 0$ *and* $M30_t = M30_{t-1}$ *then send message* φ_i

$$= \left\{ \begin{matrix} id_{\varphi_j},\ m_i,\ [\textit{machine operator},\ \textit{production manager}],\ \text{"The part being produced wasn't} \\ \text{successfully concluded. "},\ [\textit{present part timer},\ M30\ \textit{parts counter}\ ],\ \textit{Notification} \end{matrix} \right\}$$

The next two rules start and stop a timer every time the spindle load is greater or smaller than its maximum rating, respectively:

Rule r4. $\forall m_i \in M,\ i \in \{1,\ 2,\ \ldots,n\}$, *whenever* $spindle_{load} > sr_{mi}$ *and* $spindle_{timer} = 0$ *then start* $spindle_{timer}$

Rule r5. $\forall m_i \in M,\ i \in \{1,\ 2,\ \ldots,n\}$, *whenever* $spindle_{load} \leq sr_{m_i}$ *and* $spindle_{timer} > 0$ *then stop* $spindle_{timer}$

Let us now define a rule that triggers an alert message when the spindle load exceeds the maximum rating:

Rule r6. $\forall m_i \in M,\ i \in \{1,\ 2,\ \ldots,n\}$, *whenever* $spindle_{load} > sr_{m_i}$ *then send message* $\varphi_i = \left\{ \begin{matrix} id_{\varphi_j},\ m_i,\ [\textit{machine operator}],\ \text{"The spindle is working at a higher rating than the maximum} \\ \text{recommended value. "},\ [spindle_{load},\ sr_{m_i}],\ \textit{Alert} \end{matrix} \right\}$

The following rule triggers an alarm message if the spindle load exceeds the maximum rating up to 150% by more than 30 min:

Rule r7. $\forall m_i \in M,\ i \in \{1,\ 2,\ \ldots,n\}$, *whenever* $spindle_{load} > sr_{m_i}$ *and* $spindle_{load} \leq sr_{m_i} \times 1.5$ *and* $spindle_{timer} \geq 1800$ *then send message* φ_i

$$= \left\{ \begin{matrix} id_{\varphi_j},\ m_i,\ [\textit{machine operator}],\ \text{"The spindle has exceeded the maximum rating} \\ \text{and has been working at this rate for more than 30min. It must be brought down} \\ \text{to normal levels. "},\ [spindle_{load},\ sr_{m_i}],\ \textit{Alarm} \end{matrix} \right\}$$

The following rule triggers an alarm message if the spindle load exceeds the maximum rating between 150% and 200% by more than 3 min:

Figure 26 - Rules Generated in Case C (Fernandes et al., 2019)

Annex N

Table 31 - Prediction Results P&P1

Features	Algorithm	Accuracy (%)	Kappa
D1 TRANSFER_CT_MAX TRANSFER_CT_MIN STANDBY_CT_AVE	NB	0.1061	0.062
	SVM	0.4469	0.2305
	Adaboost	0.4358	-
D2 TRANSFER_CT_MIN STANDBY_CT_AVE PICK_UP_ERROR	NB	0.2067	0.0201
	SVM	0.4413	0.186
	Adaboost	0.4302	-
D3 STANDBY_CT_AVE PICK_UP_ERROR PARTS_VISION_ERROR	NB	0.1117	0.037
	SVM	0.4134	0.1149
	Adaboost	0.4190	-
D4 PICK_UP_ERROR PARTS_VISION_ERROR MARK_VISION_ERROR	NB	0.1341	0.3302
	SVM	0.4469	0.1465
	Adaboost	0.4693	-
D5 PARTS_VISION_ERROR MARK_VISION_ERROR OPERATOR_CALL_TIME	NB	0.1173	0.697
	SVM	0.3799	0.1678
	Adaboost	0.4022	-
D6 MARK_VISION_ERROR OPERATOR_CALL_TIME MARKREC_CT_MAX	NB	0.0950	0.153
	SVM	0.4134	0.1023
	Adaboost	0.4134	-
D7 OPERATOR_CALL_TIME MARKREC_CT_MAX MARKREC_CT_AVE	NB	0.0782	0.471
	SVM	0.4972	0.111
	Adaboost	0.5084	-
D8 MARKREC_CT_MAX MARKREC_CT_AVE BOARD_NAME_1	NB	0.1899	0.046
	SVM	0.4134	0.098
	Adaboost	0.4190	-

Annex O

Table 32 - Prediction Results P&P2

Features	Algorithm	Accuracy (%)	Kappa
D1 Working_Ratio Mount Rate Mounted blocks	NB	0.2000	0.0868
	SVM	0.4465	0.213
	Adaboost	0.4465	-
D2 MOUNT_RATE MOUNTED_BLOCKS PARTS_CONSUMPTION	NB	0.0512	0.091
	SVM	0.0409	0.197
	Adaboost	0.4140	-
D3 MOUNTED_BLOCKS PARTS_CONSUMPTION TRANSFER_CT_MAX	NB	0.0465	0.096
	SVM	0.3814	0.151
	Adaboost	0.3767	-
D4 PICK_UP_ERROR PARTS_VISION_ERROR MARK_VISION_ERROR	NB	0.1023	0.061
	SVM	0.3814	0.207
	Adaboost	0.4233	-
D5 PARTS_CONSUMPTION TRANSFER_CT_MAX TRANSFER_CT_MIN	NB	0.2698	0.087
	SVM	0.4097	0.137
	Adaboost	0.4372	-
D6 TRANSFER_CT_MIN STANDBY_CT_AVE PARTS_VISION_ERROR	NB	0.2326	0.066
	SVM	0.3628	0.153
	Adaboost	0.4093	-
D7 STANDBY_CT_AVE PARTS_VISION_ERROR TRANSFER_ERROR	NB	0.0884	0.050
	SVM	0.3814	0.260
	Adaboost	0.3953	-
D8 PARTS_VISION_ERROR TRANSFER_ERROR RECOVERY_TIME	NB	0.0698	0.137
	SVM	0.3860	0.214
	Adaboost	0.3907	-
D9 TRANSFER_ERROR RECOVERY_TIME NO_PARTS_ERROR	NB	0.0512	0.035
	SVM	0.4093	0.154
	Adaboost	0.4140	-
D10 RECOVERY_TIME NO_PARTS_ERROR MOUNT_CT_MAX	NB	0.3581	0.222
	SVM	0.4186	0.169
	Adaboost	0.4186	-
D11 NO_PARTS_ERROR MOUNT_CT_MAX MOUNT_CT_MIN	NB	0.1907	0.063
	SVM	0.3814	0.114
	Adaboost	0.3814	-
D12	NB	0.2000	0.127

MOUNT_CT_MAX MOUNT_CT_MIN MARKREC_CT_MAX	SVM	0.3349	0.124
	Adaboost	0.3349	-
D13 MOUNT_CT_MIN MARKREC_CT_MAX BOARD_NAME_1	NB	0.2186	0.143
	SVM	0.3395	0.101
	Adaboost	0.4000	-

Annex P

Serial Number	Dry Run	Last complete part timer	Maximum axis loads for X, Z, A
Control Software Version	Power-On Time (total)	Spindle RPM	Present machine coordinate position X, Y, A, B
Machine Model Number	Motion Time (total)	Coolant Level	Present work coordinate position X, Y, A, B
Tool Changes (total)	M30 Parts Counter #1	Spindle load with Haas vector drive	Present Tool offset X, Y
Tool Number in Use	M30 Parts Counter #2	Present part timer	Machine Vibration X, Y, Z
Noise			

Figure 27 - Selected Features case C