Data-Driven Models for Predictive Quality in Precision Metalworking

António Fernando de Lacerda Queiroz e Almeida

Dissertação de Mestrado

Orientador na FEUP: Prof. Ricardo Teixeira Sousa
Orientador no INESC TEC: Eng. António Correia Alves

Mestrado Integrado em Engenharia e Gestão Industrial

2020-06-29

Abstract

All industries are in constant pursuit of improvement in quality. With the rapid development of computer science and data mining techniques, the cooperation of these two fields was natural and inevitable, not only to help describe and understand but also to help predict.

Managing non-conformities in production are critical to maintain a high standard of quality, since rejected products originate wastes, that is often not recyclable, of material and energy which have high environmental and economic impact.

Much research has been done in the application of data mining with the goal of predicting quality in industry but, surprisingly, not much as been done specifically in the precision metalworking industry. This is an industry that due to the very specific and sensitive nature of the products manufactured requires focused attention when analysing, since there is a void for predictive tools.

The goal is to predict (before production) non-conformities in order to prevent them and to fulfill in a timely manner customer requisites, while reducing waste. For this, predictive models were elaborated from generic predictors that could be used in any context of the precision metalworking industry.

The results obtained are satisfactory since the best model predicts with 90% accuracy when restricted to a certain machine. Further research is necessary to deepen the predictive capabilities of these predictors, to extract new and to conduct tests in other contexts/companies in order to validate results.

Resumo

Todas as indústrias estão em constante procura de melhoria na qualidade. Com o rápido desenvolvimento das ciências computacionais e técnicas de *data mining*, a cooperação entre estes dois campos é natural e inevitável, não apenas para descrever e compreender, mas também para prever.

As não-conformidades nos produtos constituem um aspeto crítico no que concerne qualidade, na medida em que o produto rejeitado origina desperdícios, frequentemente não recicláveis, a nível da matéria-prima e energia que têm impacto ambiental e económico.

Foi realizada investigação na aplicação de *data mining* de uma forma intensiva com o objetivo de prever qualidade na indústria contudo, surpreendentemente, pouco foi realizado na aplicação à indústria metalomecânica de precisão. Esta indústria, dada a natureza específica e sensível dos produtos fabricados requer atenção focada na sua análise, mostrando que existe um vazio na utilização de ferramentas de previsão.

O objetivo deste trabalho de dissertação é desenvolver uma ferramenta capaz de prever (antes da peça ser produzida) não-conformidades de forma a preveni-las, conseguir atempadamente respeitar os requesitos do cliente, com redução de desperdícios. Para tal foram essencialmente criados modelos de previsão a partir da extração de preditores genéricos passíveis de serem utilizados em qualquer contexto da indústria metalomecânica de precisão.

Os resultados obtidos mostram que os modelos produzidos são satisfatórios conseguindo prever com cerca de 90% de precisão, para alguns casos restritos por máquina. Uma investigação mais profunda será necessária para aprofundar a capacidade preditiva destes preditores, para extrair novos preditores e para realizar testes noutros contextos/empresas para validar resultados.

Agradecimentos

No culminar deste percurso académico, vejo-me na necessidade de expressar o sentimento de agradecimento que se tem vindo a acumular ao longo destes cinco anos.

Quero começar por agradecer à minha família. Se nenhum Homem é uma ilha, o quebrar desse isolamento começa por aqui. Sem a minha família e o seu incessante apoio não teria chegado onde estou e não chegarei onde hei de chegar. Neste pequeno parágrafo fica uma vida e não conseguir distinguir aqui nenhum singular ato ou feito meu que atribuia a eles, só mostra que estão presentes em todos.

De seguida, quero agradecer a quem permitiu e contribuiu para a realização deste projeto. Começo por agradecer aos meus dois orientadores, Ricardo Sousa Teixeira e Engenheiro António Correia Alves. A sua constante disponibilidade e genuíno interesse no meu desenvolvimento permitiram-me ultrapassar todos os desafios que encontrei no decorrer deste projeto. Agradeço também às muitas pessoas, demasiadas para serem nomeadas, professores, colegas e amigos, que deram o seu contributo, pequeno ou grande, para que este projeto se concretizasse.

Finalmente, mas não menos importante, quero agradecer a todos aqueles que transformaram estes últimos cinco anos num percurso, e não num processo. Ficam aqui todos os amigos de peito, todas as Fénix, Dragões, Lobos, Vikings e todos os que vieram antes que assim o permitiram. Ficam aqui as Clavículas, o Baile e todos os que sabem o que é viver uma quarta-feira. Ficam aqui todas as associações e todos os amigos que elas trouxeram, sejam desta Faculdade ou não, deste Porto ou não, deste país ou não. Ficam aqui todos os colegas com quem tive aulas, que me explicaram o que não percebia, que me deram os apontamentos que eu não passei, que me ajudaram na minha parte do trabalho. Neste parágrafo, uma outra vida fica.

Quem lê isto, espero que se tenha revisto nalguma destas palavras. Se esse for o caso, digo mais uma vez.

Obrigado.

"There is no path. Beyond the scope of light, beyond the reach of dark, what could possibly await us?"

unknown

Contents

Acronyms and Symbols

IoT	Internet of Things
IIoT	Industrial Internet of Things
ERP	Enterprise Resource Planning
DRBM	Deep Restricted Boltzmann Machine
SAE	Stack Autoencoder
SVM	Support Vector Machine
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
LASSO	Least Absolute Shrinkage and Selection Operator
MMS	Multistage Manufacturing System
PCA	Principal Components Analysis
KDD	Knowledge Discovery in Databases
ANN	Artifical Neural Networks
UCL	Upper Control Limit
LCL	Lower Control Limit
USL	Upper Specification Limit
LSL	Lower Specification Limit
OP	Order of Production
LSS	Lean Six Sigma
CRISP-DM	Cross-Industry Standard Process for Data Mining
ASUM-DM	Analytics Solution Unified Method
SPC	Statistical Process Control

List of Figures

List of Tables

Chapter 1

Introduction

The world is constantly evolving and is constantly demanding better quality. Whoever fulfills this demand, has a competitive advantage. As a consequence of this, quality, either of product or of service, has never been more important for market participants in all sectors. But perfection is impossible. This being the case, if failure is inevitable, the ability to predict failure must be developed.

Technology is also constantly evolving. In the last few decades, computer science has undergone several breakthroughs. Advances in computational power, storage capacity, and data analysis techniques have provided fascinating insights that would otherwise be impossible in all sectors of enterprise, be it industry, healthcare, and beyond[1].

This project and thesis aim to do just that. To ally the power of computer data analysis to the necessity of the market for further improved quality standards and procedures. The development of computer models that predict the quality of products.

1.1 Context

In this section, the context in which this project was developed is presented.

Quality is of the utmost importance for industry. Quality is a differential factor that be used to obtain a competitive advantage[2]. As industry progressed, the discovery that investing in quality, more specifically in quality control, would not only improve company quality but also improve price and lead time.

Many would say that industry is undergoing a fourth great revolution. This revolution has therefore taken the name *Industry 4.0*. The main aspect of this new industrial revolution is the capacity to incorporate digital solutions into traditional industry.

One great new technology is the Internet of Things (IoT), and more specifically the Industrial Internet of Things (IIoT). Similar to the common Internet, which connects computers all around the world, IoT connects robots and software at a local level, be it in a factory or our home. This permits all robotized instruments and software to communicate with each other, which enables things such as real-time analysis of production and machine learning[3].

IIoT is also a source for an enormous amount of data. As technology progresses, machines and industrial robots are outfitted with more and more sensitive sensors for various parameters. This provides factory managers data analysts with a wealth of data that can be used to further develop industry and control quality.

Another new technology that is in development is the concept of Digital Twin. A Digital Twin is a virtual replica of a physical entity. For example, the Digital Twin of a factory would copy the machines, workers, processes, and movement of the factory and from there can simulate a normal day of operation. Digital Twins can help monitor a factory in real-time. Linked with the concept of IIoT, Digital Twins can receive the inputs of all sensors and is constantly learning and updating from these inputs from the physical counterpart.

More advanced Digital Twin models not only mimic the shop floor, but are also able to simulate the operation and changes made to the shop floor. Digital Twins are useful because, for instance, they can be used to simulate the advantages and disadvantages of making changes to the shop floor, which are usually very costly both in terms of money and time.

There are, however, challenges that must be faced. One such challenge is the standardization of information storage. For example, digital solutions such as enterprise resource planning (ERP) software are already widespread. However, each ERP stores data in a format commonly unique to that ERP only. The transformation of that format to other more conventional formats is usually very difficult. The overcoming of this hurdle could facilitate the sharing of information which in turn, for example, can help with forecasting and the understanding of customer needs, improving customer experience overall.

This project was born from the need of JASIL to improve their product quality, service level, and lead time. Being that all of these are important performance indicators and competitive fronts, the improvement of these measures is a constant concern of any company. This project can help with the improvement of all of these indicators. Knowing what productive factors influence quality will help mitigate potential damage and thus improve overall product quality. With better prediction of quality outcome, it is possible to plan production accordingly so that production orders can be fulfilled on time.

JASIL is a company where quality concerns are of paramount importance. This company produces precision metalworking products for several sectors of industry such as the automotive sector. Therefore, many quality issues are considered. JASIL is also a company that invests heavily in innovation and state of the art solutions to tackle the challenges they face. Taking these two factors into account, JASIL agreed to cooperate in the development of this thesis.

As was mentioned before, JASIL is inserted in the metalworking industry. Metalworking is the process of transforming metal into parts or assemblies of various components. This industry can produce large parts to be used in ships or bridges, or it can produce small and precise components for example for the automotive industry, which is the case of the endorsing company.

1.2 Motivation

This section presents the driving motivations of this dissertation. The motives are related to economic, environmental, energetic and social issues.

Non-conformities have a great impact on the bottom line of a company. Non-conformities both reduce revenue, due to a lesser capacity to supply the market, and increase costs, many of which are oftentimes neglected.

The obvious costs of non-conformities are material costs. However, these are not the only ones. Every time a piece is rejected one more must be produced to replace it. This means labour costs, energy costs, logistics costs (handling the rejected products). Also, the wear of machinery and tools is a cost that is often forgotten, in part because of the small impact on the bottom line. Every product that a machine manufactures wears out the tools used.

From an environmental standpoint, one can argue that waste reduction is imperative. Rejected products may not be recyclable (or eliminated) and may be considered as waste that might represent a potential threat to the environment. As an example, steel is one of the five materials that are responsible for 55% of global CO_2 emissions[4]. Water contamination is also a frequent and severe impact[5]. Frequently this contamination is not even intentional.

All this waste comes mainly from old industry habits i.e. outdated manufacturing practices, and not from necessity. The reduction of these yield losses could result in a decrease of environmental impact factors. However, one must consider that not all waste can be eliminated and not all waste can be recycled. Therefore, prediction of frequency and type of waste complements other efficiency-driven methodologies such as circular economy [6]. Circular economy is a methodology and resource philosophy which has the aim not of reducing waste but of re-purposing and re-integrating waste, either material or energetic, into the productive system.

One waste that can also be overlooked is energetic waste. Electricity makes everything in the modern world happen. Every time an action is taken in a factory energy is wasted. Bearing this in mind, one can argue that the more non-conformities a factory, process, or machine produces the more energy will be wasted since these non-conformities must be replaced.

Society as a whole can benefit from waste reduction. On a more global scale, waste reduction leads to companies having lower production costs. This results in lower prices which means more people have access to better products. Also, customers today are more conscientious of the products they consume and what impact the product has. Customers nowadays may boycott products from companies that do not do their part with mitigating their environmental and social footprint.

And even the employees working in the company can benefit from waste reduction. For example, previous research has shown the link between the implementation of Lean waste reduction methods and employee satisfaction[7].

The context of this project is the development of the field of predictive quality. This is driven by the desire of JASIL to improve performance and maintain competitiveness in the market. Besides the economic motivations of the company, the project also tackles environmental and social

issues through the reduction of waste.

1.3 Work Methodology and Timeline

This section explains the work methodology of the development of the project and presents a timeline of all the phases of development, from literature review to conclusions. A subsection is included that presents the criteria utilized for the retrieval and filtering of the referenced literature.

The elaboration of this project was divided into 4 milestones.

The first milestone consisted of the collection and reviewing of literature to support the project. The main output of this phase was the writing of the introduction and of certain parts of the literature review of the thesis.

The second milestone, firstly, was the characterization of the database. After an initial analysis and preparation of the data, the case study selection and definition of research questions were possible. The main outputs of this were the writing of the literature review and the selection of the appropriate algorithms for the project.

The third milestone is the implementation of the selected algorithms. After being selected, the algorithm models were trained and tested with previously received data. Having this, it was possible to draw conclusions.

The fourth and final milestone is the conclusion of the writing of the thesis. The main output of this phase is the final version of this document.

1.4 Main objectives

One of the main propositions of this work is to further develop the field of predictive quality. The goal is to analyse the influence of certain productive factors on product quality. This will be achieved through the creation of a predictive model that foresees the amount or likelihood of the non-conformities.

This project was developed as a proof of concept but always bearing in mind the possibility of practical industrial application in a future time. The purpose of the resulting software should be to aid decision making in the planning phase of production, taking into account the latest production information available.

1.5 Thesis Outline

This document is divided into 7 chapters, some of which contain subchapters.

Chapter 1 *Introduction* introduces the project, going through the concept, context, motivation, and project timeline.

Chapter 2 comprises the literature review. This chapter goes in-depth on the topics of quality in industry, data-driven methods of analysis, and predictive quality, the culmination of which is the research question of this thesis.

Chapter 3 follows the work methodology used to resolve the challenges of this project, as well as the contextualization of this project with the company. Due to the nature of the project, the contextualization and understanding of the company are intrinsic to the methodology. This chapter goes through the initial steps of data understanding, preparation, and analysis. Then, the description of what models were applied and the performance metrics obtained. Finally, the models are compared.

Chapter 4 discusses the results obtained, how well the objectives for the thesis were achieved, and what are its main strengths and shortcomings. Following this is the proposal of potential ways the work done in this thesis can be used as a stepping stone to further develop the state of the art.

Chapter 2

Literature Review

This chapter explores the current state of the art of quality assessment, the application of data-driven methodologies, and how these two overlap. It culminates with the research question of this thesis and what it proposes to add to the current understanding of the topic.

2.1 Quality in Industry

This section defines the current understanding and practices of quality control in industry. The literature review of this section was done, unless stated otherwise, based on the book *Introduction to Statistical Quality Control* by Douglas C. Montgomery[8].

Quality is not a new concept. This concept has been around for quite some time and, throughout that time, has taken many definitions. The typical definition of quality is that products and services must follow the specifications and requirements of its user, i.e. quality is fitness for use, how well the product can be used by the customer. This definition is lacking as it only takes into account the conformance aspect of the product or service, disregarding the design aspect. A more fitting and inclusive definition is that quality increases when variability decreases.

Despite this definition, certain dimensions must be defined to evaluate quality. In 1987, Garvin defined the following 8: performance, reliability, durability, serviceability, aesthetics, features, perceived quality, and conformance to standards. These are adequate in an industrial context. Should a product not respect quality standards expected by the company or the client, the designation of non-conformity is given[9].

To aid in the detection of errors in the various production processes, quality control has several tools that can be used. The following are some of the more common and more powerful tools used in quality control, but there are many more:

Flowcharts help to organize and visualize the various processes of a company. Flowcharts consist of a succession of symbols that represent the phases, intermediate outputs, and the intervening agents of a product or service. Developing process flowcharts usually permits the identification of problems that could not be noticed in everyday operation.

Cause-and-effect diagrams, fishbone diagrams, or *Ishikawa* diagrams are representations of possible causes that result in specific problems or errors. These are a mere visualization that represent all possibilities. Considering these facts, not all factors are at the root of a problem and each possibility must be analysed and tested individually.

2.1.1 Statistical Quality Control

One of the main branches of quality control is based on statistical methods, having three major areas: **statistical process control**, **design of experiments**, and **acceptance sampling**. Statistical process control or SPC seeks to develop workplace culture in which all employees strive for continuous improvement. Statistical process control contains a large set of tools, such as process control charts and Pareto charts, based on well-founded statistical knowledge.

Acceptance sampling consists of testing a batch of outbound product or inbound materials for non-conformities. Should the proportion of non-conformities be lower than a previously established value, the lot is accepted. Otherwise, the lot is not accepted. Acceptance sampling helps reduce the cost and time required for testing an entire batch. Acceptance sampling works best when done frequently and in small batches.

In a designed experiment, inputs are controlled and varied in order to determine their influence on the output. These often result in major process performance and product quality increases. Designed experiments help to understand the impact of controllable variables on the quality of output. The ultimate goal is to understand the value a set of n variables that have influence has to assume in order for quality metrics to assume ideal values. However, designed experiments should also help eliminate the influence of uncontrollable variables.

However, SPC, like any methodology, is not without disadvantages and shortcomings.

For example, SPC requires a lot of field data in order to conduct an adequate analysis of the current situation of the company. This requires a long time frame before the methodology can even be put into practice, much less before the effects produced can be felt. This large data requirement also implies a large initial investment. To be able to measure the necessary dimensions in order to be able to apply SPC, the company must invest in the proper equipment and in the proper training of the personnel.

Also, while a great methodology for analysing and assessing the state and problems of a company or process, SPC is limited in terms of the ability to predict future outcomes. This being the case, SPC can be allied to the predictive capabilities of computer science and data mining techniques (which will be discussed further on).

As was previously stated, JASIL works in precision metalworking. One of the main concerns in terms of quality of this industry is the very rigorous control of dimensions of the pieces produced. For this purpose, among the many existents, two tools exist: histograms and process controls charts.

The application of histograms in SPC is very common due to their very simple yet powerful properties.

In short, histograms summarize data by aggregating all the instances in intervals known as bins. The size of the bins is also important. If the amount of data is high, an insufficient number of bins might result in overgeneralization of the data. If the amount of data is low, an excessive number of bins might result in the histogram having gaps and not being able to properly display the distribution of the dimension in question. In practice, using the square root of the sample size as the number of bins yields good results.

Histograms are a great visualization tool that permit the effective identification of results, such as the capacity of a process (which will be discussed further on), and the communication of these results. Histograms are also crucial in the identification of the potential distribution a process might follow, more specifically if the process follows a normal distribution. Most of the techniques provided by SPC assume a normal distribution. Should a process follow a normal distribution, the application of any statistical method is much more simple.

Another one of the main tools statistical process control has to properly assess and manage irregularities in dimensional quality are control charts.

Statistical process control charts or simply control charts are also a very important tool in quality control overall, not just the metalworking industry. These consist of charts that plot various samples with which it is possible to understand if the process is in control or not. A process is controlled when variation can only be attributed to arbitrary causes.

As Figure 2.1 shows, a control chart is characterized primarily by 3 values: the process average (represented by the green line), the upper control limit (UCL) and the lower control limit (LCL). Every point represents a sample that was taken. These points should be represented in chronological order.

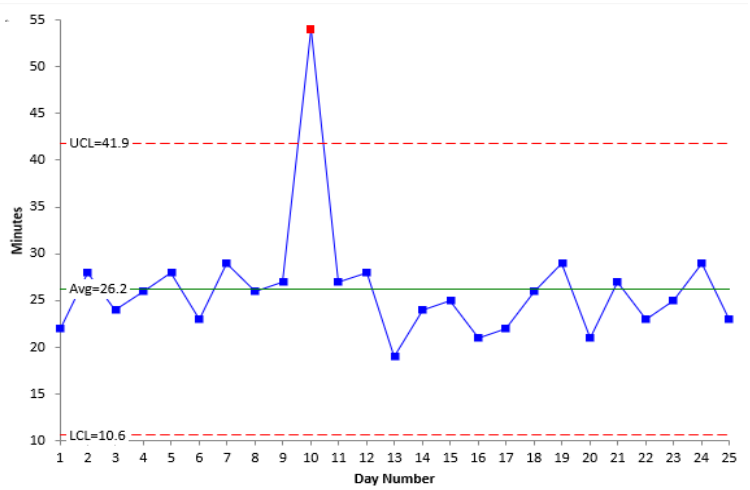

Figure 2.1: Control Chart Example
in "https://www.spcforexcel.com/spc-blog/what-control-chart consulted on 2020-05-10, 12:46"

The UCL and LCL are calculated using the mean value and standard deviation (often represented by the greek letter σ) of the process.

$$UCL = Avg + 3\sigma$$

$$LCL = Avg - 3\sigma$$

Every point in the chart represents a sample that was taken. Should a point be above the UCL or below the LCL the process is said to be out of control. The red dot in Figure 2.1 is an example of this. Whenever a sample goes beyond the control limits, the sample and the conditions of production should be analysed to determine what caused that irregularity.

There are several types of control charts, all of which are more appropriate for a given set of conditions. These control charts can be divided into 2 different categories: whether what is being measured is an attribute (defects counted or measured in discrete intervals) or a variable (measured continuously).

When measuring defects in discrete intervals, control charts can further be divided into 2 other categories: considering the total amount of defects present in the sample or the percentage of defective products in the sample. Of these two categories, the first should be used when the number of occurrences per sample is low.

When considering the total amount of defects, c charts and u charts are implemented, when sample sizes are constant and variable respectively. When considering the percentage of defective pieces, np charts and p charts are used, also when sample sizes are constant and variable respectively.

For the control of variables, 3 different control charts exist: X-MR charts, for when the sample is a single entity; $\overline{X}$-R charts, for when the sample is between 1 and 10 entities; and $\overline{X}$-S charts, for when the sample size is equal or larger than 6. Figure 2.2 summarizes the process of selecting the adequate control chart to use.

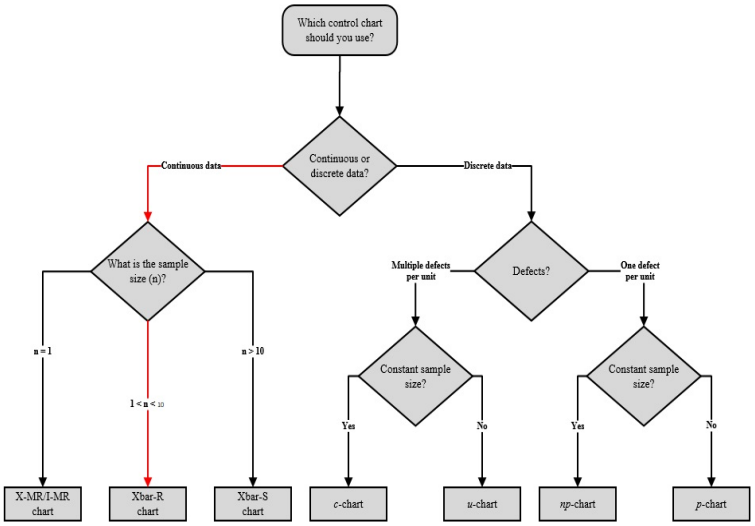

Figure 2.2: Control Chart Selection Flowchart in "https://towardsdatascience.com/quality-control-charts-x-bar-chart-r-chart-and-process-capability-analysis-96caa9d9233e consulted on 2020-05-10, 16:32"

Despite not being represented in the control chart, there are other parameters that are important in the design and elaboration of the control chart. Namely, sample size and sample frequency are two crucial factors. As sample size increases, shifts in the process mean are more easily detected. The same happens with the sampling frequency.

Ideally, large samples would be taken very frequently. However, this would be very costly. Therefore, either small samples are taken very frequently or large samples are taken infrequently, being that industry standard has taken a preference for the former. Small, frequent samples are ideal when production is done in high volumes or shifts in mean can be attributed to several causes.

Two measures are utilized to evaluate the adequacy of adopted sampling policies: average run length (ARL) and average time to signal (ATS).

ARL is the amount of points required, on average, for an out-of-control signal to be produced in a control chart. If the probability of a non-conforming signal occurring is p then $ARL = 1/p$. ARL has suffered much criticism due to having the mean being the same as the standard deviation, i.e. actual ARL will suffer great variations. ATS is the expression of ARL in terms of time instead of frequency. For example, should the ARL of a process be 350 and a sample is taken every hour, the ATS of that process will be 350.

Despite the criticism suffered, ARL and ATS are both useful, and selecting an adequate amount of points for the control chart is vital, in order to minimize the possibility of a non-conformity signal being produced, despite the process being under control.

However, the process remaining within the limits of control does not mean that control has been achieved.

For example, certain trends may be verified such as several consecutive points increasing in magnitude. Patterns such as this (not necessarily increasing in value) are called *runs*. In a process that is in control, runs with a high amount of points are very unlikely to occur. Thus, the presence of a run can indicate that the process is actually out of control.

Runs are not the only indicator that a process is out of control despite respecting control limits. Any pattern, such as cycles, is an indicator that the process is potentially out of control. Ultimately, the plot of a control chart should appear random within the control limits i.e. variation is within expectation and should only be attributed to arbitrary causes.

To achieve this purpose, a guiding set of rules has been established to more easily determine whether a process is controlled or not. Of note that these rules, not only utilize the 3 sigma control limit, but also the 1 and 2 sigma warning limits. The rules are:

1. 1 or more points are outside the control limits.
2. 2 of 3 consecutive points beyond the 2 sigma limits.
3. 4 of 5 consecutive points beyond the 1 sigma limits.
4. A run of 8 points on either side of the center line.
5. 6 consecutive points steadily increasing or decreasing.

6. 15 consecutive points between the upper and lower 1 sigma limits.

7. 14 consecutive points alternating between above and below the center line.

8. 8 consecutive points on either side of the center line, none within the 1 sigma limits.

9. Unusual, nonrandom patterns (cycles, for instance).

10. Points nearing warning or control limits.

These rules are in descending order of importance and in descending order of likelihood that the process is out of control. As such, while the first rule indicates immediately that the process is out of control, the subsequent rules act as more of a warning. In the case that one of these other rules is broken, perhaps more frequent sampling should be implemented, to conclude with more certainty that the process is indeed out of control.

Despite their extreme usefulness, control charts only evaluate the consistency of a process and help to isolate irregular occurrences. Control charts cannot determine whether a product conforms to the required specifications.

The specifications of a process are dependent on the specific requirements a customer has for a product. A product will have the desired mean value for a certain dimension and a dimensional tolerance, forming an interval in which the piece conforms. When transposed to the control chart, these become the upper specification limit (USL) and lower specification limit (LSL).

Knowing the product specifications and the process control limits, the capacity of the process, both short-term (C_{pk}) and long-term (P_{pk}), can be calculated. Considering that the process mean value and specification mean value may not be equivalent we have that:

$$C_{pk} = \frac{MIN(USL - Avg, Avg - LSL)}{3\sigma_{ST}}$$

$$P_{pk} = \frac{MIN(USL - Avg, Avg - LSL)}{3\sigma_{LT}}$$

where σ_{ST} and σ_{LT} are the short-term and long-term variation of the process, respectively. If the process be stable, C_{pk} and P_{pk} will be equal.

But where does the use of six standard deviations originate from? The use of this value is not arbitrary. The use of this value stems from the Six Sigma methodology. The six sigma states that ideally 99.99966% of products should conform to quality standards and offers a variety of tools to achieve this end.

A recent development in industry that has since become common practice is the implementation of the Lean Six Sigma (LSS) methodology. LSS allies the Lean manufacturing methodology with the Six Sigma methodology of variation reduction to reduce all types of waste, either non-conformities in products or excess inventory.

LSS is more than just the powerful set of tools that are provided by this methodology. LSS is a mindset of continuous improvement that affects all members of an enterprise and aims to encourage collaboration between the various functions and sectors of a company to reach a common goal.

2.2 Data-driven Methods

This section makes explicit some keys concepts in the fields of knowledge discovery in databases, data mining and machine learning referring how these tie in with quality control. The literature review for this section was based, unless stated otherwise, on the books *Introduction to Data Mining* by Pang-Ning Tan *et al.*[10] and *Deep Learning* by Ian Goodfellow *et al*[23].

The demand for greater quality control created the necessity for greater data gathering. More factors that could disrupt quality had to be taken into account, so more and more varied data was required. The rapid evolution of computational technology has permitted modern companies to gather and store the necessary information to control their enterprise and their products.

This is something that is being done by companies in all sectors of development and service, such as medicine and science. This practice is not at all reserved for traditional industry. Modern companies store huge amounts of information, sometimes having databases that reach *petabytes* in volume.

As the amounts of data increased, the methods to process this information and gain benefits from it also evolved. Concepts such as *knowledge discovery in databases*, *data mining* and *machine learning* were born from this evolution.

Although data-driven methods can be used in parallel with SPC, these are methodologies that present small connection between them and, therefore, SPC was not used directly in this dissertation but is an important mention in the state-of-the-art and are techniques that are currently used by JASIL.

2.2.1 Knowledge Discovery in Databases

Knowledge discovery in databases (KDD) refers to the method of extracting useful information from a given database. Being that the raw database is the input for this process, KDD is usually divided into 3 stages: data preprocessing, data mining, and data postprocessing.

Data preprocessing refers to the preparation of the database to permit a proper and worthwhile outcome and is the first step to be taken. This involves steps such as the removal of outliers, standardization of data, and the selection of features that are relevant to the proposed problem. Understanding the context of the data is also very important. Meaningless patterns may be discovered if the proper context is not kept in mind.

Data postprocessing is the integration of the acquired knowledge into the current procedures of the company. This phase may involve steps such as visualization of results to permit their interpretation by competent authorities. Results obtained may even still undergo further statistical analysis to eliminate fake or nonsensical data mining conclusions.

2.2.2 Data Mining

Data mining is the intermediate phase of the KDD process. From large amounts of data, the aggregation of individual records into similar groups or the discovery of patterns of influence between certain attributes is possible through the combination of machine learning, database systems, and statistics. With this acquired knowledge, future predictions can be made when the value of influencing factors is known.

Data mining tasks are usually divided into 2 major categories.

The first category is predictive tasks, also frequently referred to as supervised learning. Predictive tasks employ a target variable, a variable assumed to be dependent on many others. Having all independent variables and one dependent variable, predictive tasks attempt to understand the connection between the former and the latter. Predictive tasks can also be subdivided into classification tasks, used when the target variable is discrete, and regression tasks when the target variable is continuous.

The second category is descriptive tasks or unsupervised learning. Contrary to supervised learning and as the name suggests, descriptive tasks attempt to capture patterns and describe relationships between data points. From these tasks, clusters, trends, and anomalies may be discovered. The output of descriptive tasks usually lacks context and therefore usually requires the application of postprocessing techniques to have practical applications.

Some models frequently used in data mining include Decision Trees and Random Forests, Support Vector Machines (SVM), Genetic Algorithms and Artificial Neural Networks (ANN), this last one being closely linked to the concept of *deep learning*.

The development of the field of data mining is currently facing some challenges. Firstly, datasets continue to grow in volume and amount of attributes. This results in impossibly long processing times and currently used algorithms may not even be appropriate to handle the huge number of attributes. Data ownership is also another problem. The necessary data is usually scattered throughout multiple entities. This raises communication speed and security issues that must be addressed. This might also cause incomplete data which can compromise the efficiency of the methods.

2.2.3 Machine Learning

Data mining is closely linked to machine learning. The data mining process usually employs several machine learning algorithms to achieve the desired conclusions.

A machine learning algorithm is an algorithm from which one can extract knowledge when applied to a dataset. Generically, they are characterized by the input data, the defined task to perform, and the ultimate performance measure of that task.

The possible tasks are the aforementioned predictive or descriptive tasks. The task that should be performed helps to decide the algorithm to be applied. Once applied, an algorithm becomes a model.

After selection, the model must be trained and tested. To achieve this, the dataset is typically separated in a training set, that as the name suggests trains the model and gives the model predicting capability, and a test set, with which performance can be measured. Typically the proportion in which they are separated is 70%:30% but can go to 80%:20%, respectively. The more data is allocated to training, the better the model becomes, and the more data is allocated to testing, the more trustworthy the performance metrics are.

Separating the dataset into two is important because the model should be evaluated using data that was never seen before in order to avoid any possible bias. The main purpose of a machine learning algorithm is to perform just as well when new and different information is provided.

Considering these facts, two common challenges that machine learning faces may occur. The problems of *underfitting* and *overfitting*. When a model cannot create a pattern that explains the dataset well enough, this is defined as being underfit. Should a model perform well with given data but cannot adapt well to new information, this is defined as being overfit.

To avoid a model entering any of these two states, the *capacity* of the model can be changed. In simple terms, capacity is the flexibility of a model. If a model is not flexible enough, all datapoints will not be represented. If a model is too flexible, the model will describe the datapoints themselves and not the underlying function that describes them.

2.2.3.1 Hyperparameters

One issue that cannot be neglected when training a model is to choose the optimal hyperparameters for the situation, seeing that there is no universally optimal combination.

The hyperparameters of a model cannot be estimated from the data, even though the optimal combination is highly dependent on the data, and there are no analytical formulas.

The commonly used methods rely on brute-force heuristics in which a multitude of combinations for the hyperparameters is selected either randomly or from a pre-defined set and models are trained and tested with each combination. The models are then evaluated to determine which had the best performance.

2.2.3.2 Model Evaluation Metrics

Different models with different hyperparameters yield different performance metrics. The performance measure is relative to the specific task of the model. Also, there is no universal performance measure. Since this project focuses on binary classification, only the performance metrics for this paradigm are explored.

The output of a binary classifier can be split into 4 different categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

The most generic performance metric for binary classifiers is accuracy, the rate at which the model is correct. However, other performance metrics are required to fully understand the usefulness of the model, as extracting only the accuracy of the model can be misleading.

Considering widely used metrics for this context, there are 4 main performance metrics for a binary classifier:

- **Accuracy**
  Rate at which the prediction of the model is correct.

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

- **Precision**
  Fraction of true positive outputs among all positive outputs.

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

- **Recall**
  Fraction of elements that were given a positive output among all the elements that should have been given a positive output.

$$recall = \frac{TP}{TP + FN}$$

- **F1 Score**
  Also a measure of the accuracy of the model but that gives more importance to false positives and false negatives. Useful when there are high class imbalances.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

2.2.4 Data-driven Project Methodologies

When working on a project that involves data analysis as a whole, not just data mining, the following of the proper phases of project development is essential to ensure the optimal outcome.

Projects such as the one developed in this thesis follow 4 naturally progressive stages. Each stage is progressively more complex but also brings more benefits to the company.

The first stage is the **descriptive stage**. This phase is closely linked to data understanding and consists of the description and summary of the data provided. In this phase, the concept of business understanding also comes into play. For results and conclusions to have any validity, the context in which they are inserted must be understood. The employment of data visualization is very important.

The second stage is the **diagnostic stage**. The diagnostic stage attempts to understand the causes that lead to the current situation determined in the previous step. Confounding factors should be simplified and isolated to determine the source of the problem.

The third stage is the **prediction stage**. In this phase, based on the information gathered from the descriptive and diagnostic phase, predictions on future outcomes are made. Data mining algorithms are usually applied in this phase as a decision support tool.

The fourth and final stage is the **prescriptive stage**. Knowing the current state of the company, what problems the company has and what is most likely to happen in the future, recommendations can be made on what steps to take in order to correct existing problems and better the situation of the company.

However insightful and successful data mining projects all follow them, the stages described previously lack structure and practicality. By themselves, these stages are not enough for an appropriate methodology.

Therefore, the methodology adopted was the Cross-Industry Standard Process for Data Mining or CRISP-DM. Conceived in 1996 and since then applied by a majority of data scientists, CRISP-DM is a universal methodology, applicable to all industries[11].

CRISP-DM follows 6 major phases:

The first phase is the **business understanding**. This phase aims to determine the current situation of the entity for which the project is being developed. This comprises identifying current solutions being applied, understanding the objectives that are hoped to be achieved with the project, and translate those objectives to milestones and outputs in data mining.

The second phase is the **data understanding** phase. Once again, visualization plays a big part in this phase. Data should be collected, as much as possible, described, and summarized. Interesting attributes and problems identified with the data should be outlined to be followed up on in the next phase.

The third phase involves the **preparation of the data**. This phase usually takes the majority of the time. In this phase, relevant data is selected (knowledge acquired during the business understanding phase comes into play here as well). The data acquired is then cleaned (missing values filled, false data corrected or eliminated, are some examples) and predictors are made and inferred with the existing information.

The fourth phase is the **modelling** phase. In this phase, the appropriate model or models are selected and applied. Note that, selecting the model is important, but selecting the appropriate parameters for the model is also key to the success of the project. After the model is built, an assessment of the results is required, either against itself or against other models elaborated. The results should also be compared to the knowledge currently detained, to ascertain whether something new was discovered or not.

The fifth phase is the **evaluation** phase. While some evaluation is done in the previous phase, the evaluation phase is much deeper, as this phase should analyse the process as a whole and think of the entire project and not just the model developed. For example, the evaluation phase should question whether or not the data mining goals set previously were met. This phase should also elaborate on the next steps to take and how to apply the results obtained.

The sixth and final phase is the **deployment** phase. Deployment involves applying the results that were obtained, either with the production of new software or the incorporation of the acquired results into the existing solutions, and monitoring the performance of the application. The project as a whole should also be reviewed and reported on. Figure 2.3 shows the cycle of a project that follows CRIPS-DM.

Although still widely applied, in 2015 IBM created the Analytics Solution Unified Method for Data Mining or ASUM-DM. ASUM-DM is a refined and expanded upon version of CRISP-DM.

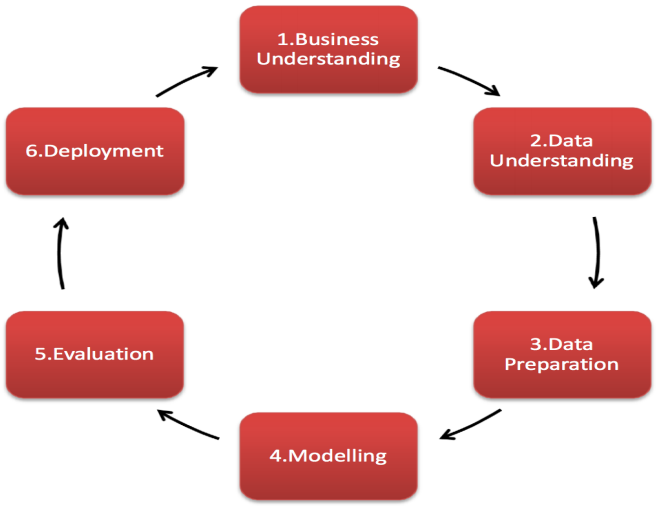

Figure 2.3: CRISP-DM Project Cycle in "https://www.sv-europe.com/crisp-dm-methodology/" consulted on 2020-06-12, 12:3"

Bearing all this in mind, one can state that knowledge discovery in databases and machine learning are invaluable to the development of industry overall. The levels of complexity that have been reached make automated data mining processes of the utmost importance and the knowledge acquired would be inaccessible otherwise.

2.3 Predictive Quality

This section makes explicit some of the recent developments in the field of predictive quality, with a focus on data-driven methods. Although it is a recent field of work and is still in the early stages of development, data-driven models have been previously used with success with the objective of predicting quality in manufacturing.

For example, research was done to compare the performance of linear, non-linear and tree-based models when predicting defect rates at a lot level, concluding that the last one outperformed the rest[12].

2.3.1 The importance of data collection, feature and algorithm selection

Due to the complexity of modern Multistage Manufacturing Systems (MMS), which have several influencing factors such as raw material, operator, machine, and tools used, it is important to understand their influence, for example through Principal Components Analysis (PCA), when considering essentially numeric data[13].

And it is also important that data be collected throughout every stage of production. A model's accuracy is extremely limited by the production processes of which it has no data unless these have a small impact on product quality[14].

2.3.2 Potential Algorithms

The selection of the appropriate algorithms with adequate criteria to solve a data science problem is essential. There is no algorithm that is universally better than the rest. Every problem is unique and the most appropriate algorithms must be selected and even these need to be customized in order for the optimal solution to be achieved.

Deep-learning algorithms Deep Restricted Boltzmann Machine(DRBM) and Stack Autoencoder (SAE) were used to demonstrate their feasibility to predict quality, even though the SAE algorithm is typically used for anomaly detection. It was also shown that deep-learning algorithms outperform shallow-learning algorithms[15].

Autoencoders are a type of artificial neural network. While the concept of an autoencoder has existed for several decades, this algorithm has gained popularity recently in the stacked framework. The applications of this framework vary from quality in manufacturing to social media applications. These algorithms, in the various frameworks that can be assumed, are used for unsupervised learning.

A stacked autoencoder is group to autoencoders stacked in layers. The output of a layer serves as an input to the subsequent layer.

DRBM is also a type of artificial neural network. DRBM has the capacity to, from the inputs received, determine a probabilistic distribution. Depending on the application, this algorithm can be used in both supervised and unsupervised manners.

Similarly, in the field of Quality Assessment but also of Quality Prediction, the Support Vector Machine (SVM) or Support Vector Network has surged in popularity in part due to its capacity to efficiently handle large datasets and the possibility of being adapted to be insensitive to the uncertainty inherent to manufacturing data[16].

SVM has roots in statistical learning methodologies. SVM separates training datapoints into one of two categories. This model physically represents datapoints in a hyperplane with a multitude of dimensions. The predicting capability of the model is based on the mapping of new datapoints and determining on which side of the dividing line they fall[17].

And the combined use of supervised and unsupervised methods of machine learning as also been used successfully before. For predictive quality, supervised and unsupervised methods were used simultaneously in the heat rolling industry[18].

Applications that predict in real-time instead of relying on sampling inspection based on Least Absolute Shrinkage and Selection Operator (LASSO) regression methodology have also been developed and proven successful. LASSO outperformed certain more conventional regression methods, despite still having certain disadvantages, like being prone to overfit the model[19].

Finally, the utilization of genetic algorithms to improve upon the already successful use of neural networks in this topic has also proven to be a promising path for the betterment of this industry[20].

Genetic algorithms are an adaptation to data science and machine learning of Darwin's theory of evolution. Starting with an initial population, the solutions that are more fit to answer the problem are selected and kept while the less fit are removed. Each solution, known as an *individual* is also given a chance to randomly mutate. With the selected solutions, a new generation of solutions is created and so forth until the optimal solution is reached.

Genetic algorithms are a powerful tool and highly flexible due to the fact that the algorithm is not trying to reach an exact solution, but rather an optimal one. However, this flexibility might be seen as a disadvantage in scenarios where precision is imperative[21].

2.4 Research Question

This literature review allows for the conclusion that, while data-driven models for the prediction of quality have been successfully used in the past, close to nothing has been done in the precision metalworking sector.

This thesis serves as a stepping stone to a future practical application that may derive from the investigation conducted. Therefore, this thesis proposes to describe the data in terms of non-conformities and analyse the influence of production factors on quality output. Following this, the performance of a series of preliminary experiments with state of the art algorithms to predict non-conformities when applied to data regarding the metalworking industry.

To summarize, the research questions for this thesis are:

1. Describe the data in terms of non-conformities and potential factors.
2. Understand the data and extract information that may be used to predict non-conformities.
3. Perform preliminary experiments with state of the art algorithms to predict non-conformities.

Chapter 3

Project Methodology

3.1 Overall

As was mentioned before, the source of methods applied in the development of this project is data science and, for that end, JASIL provided an SQL database with production data.

This being the case, there is the need to know how a Structured Query Language (SQL) database operates. A database is a system that collects and allows for the consultation and manipulation of information. Typically, this information is stored in tables with various columns that describe the attributes of a specific entity. Each specific entity has an attribute or combination of attributes that are unique to them and allows the entity to be identified. These tables present relationships between them that allow the crossing of information between the various tables of the database.

SQL, as the name suggests, is a language that has been universally adopted, and that allows the extraction of specific information from the database. SQL is a text-based command language. Since the number of entities of certain databases can range in the billions, SQL provides a non-visual and practical way of accessing, filtering and grouping information without the need of the indexes of the entities.

As previously mentioned, JASIL provided production data without which the elaboration of this project would be impossible. This data was provided in the format of a backup database. The database consists of 7 different tables, which are described below. The information contained in these tables ranges from measurements made to a specific dimension of a specific piece to batch quantity details.

- **Table *Estruturas***
  Contains information on the structures of the orders of production (OP), including operation and machine used.

- **Table *Consumos***
  Details consumption derived from production, both of raw materials and components needed or tools utilized during production.

- **Table *Colaboradores***
  Stores data on the workers of the company, including name and work station.

- **Table *Lotes***
  Catalogs the batches created by registering production during the OP.

- **Table *PIE***
  Registers plans of inspection and quality control. This includes information on dimensional tolerance and the associated blueprint of the piece in question.

- **Table *Registos***
  Contains all registries of production. In this table, the information on whether a batch was approved or not can be consulted.

- **Table *Medicoes***
  Stores all measurements made during plans of inspection. The value measured, the limits of tolerance and the resulting decision from these parameters is kept in this table.

Each table has a corresponding primary key. However, due to the nature of the ERP used by the company, the original database does not contain foreign keys and does not work in cascade i.e. there is no explicit relationship between the tables. Nevertheless, there are equivalent columns in each table with which the posterior creation of foreign keys was possible. A more thorough description of all the columns of each table can be found in appendix A.

The final product of the relationships between tables can be seen in Figure 3.1.

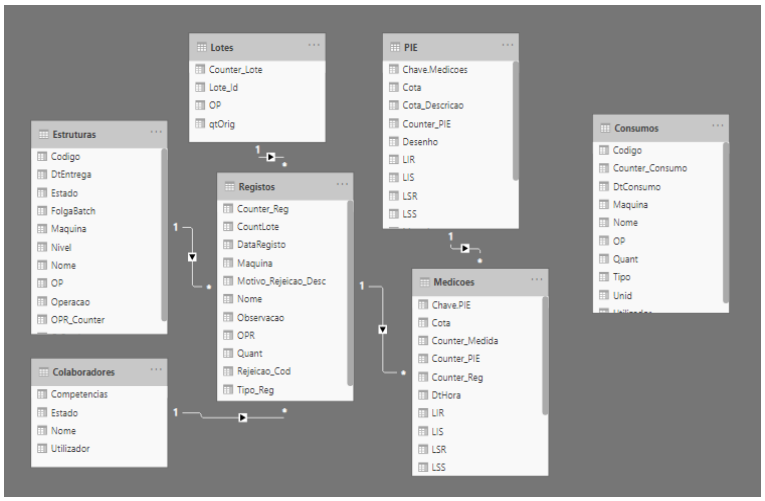

Figure 3.1: UML Diagram

This was a description of the database at first glance. The entire database was not used for data mining purposes and the database presented problems (such as the lack of connection with Table *Consumos*), all of which are explored in the section dedicated to data understanding.

3.2 Company Overview

Business understanding is the first phase in any data mining project and the description of the company is a part of this phase. This being the case, it was decided to include the description of the company in the methodology chapter, as this is intrinsic to the development of the project.

This information was obtained in part through the analysis of the database provided, but also much of it came with the courtesy of JASIL.

JASIL is a precision metalworking company that has been in the market for over 70 years. Despite this, the company is modern due to its constant investment in new technology and its bet on product development and customer and quality focus. JASIL operates in 4 main areas of production: forging, turning, milling, and grinding. Being that most of their products are custom-made, JASIL works essentially in a make-to-order paradigm, keeping in stock only intermediate and standard components.

JASIL has stored in database the production of 5348 different references of products, ranging from standard products to custom-made products. All of these products were aggregated into families and the result can be seen in Figure 3.2.

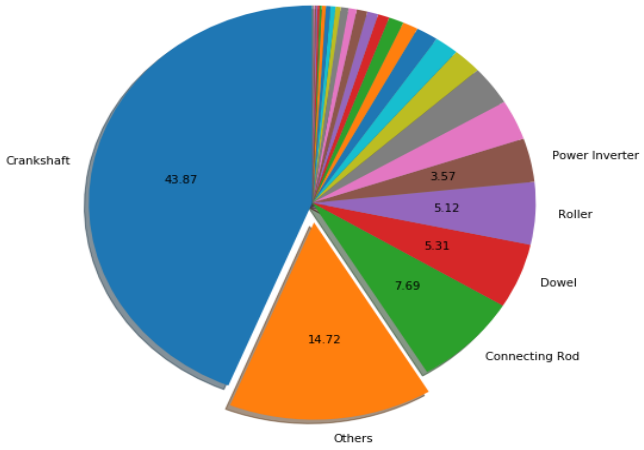

Figure 3.2: Relative Distribution of Product Families

As Figure 3.2 shows, the bulk of the production is in the family of crankshafts. Note that each family includes the products themselves and other components, such as screws and pins, required to produce that same product. The *Others* category aggregates the minority classes of the products.

JASIL currently has 93 employees. These employees are separated in their respective working sectors. The following bar chart illustrates how the employees are distributed throughout the various sectors (like accountants and commercial employees).

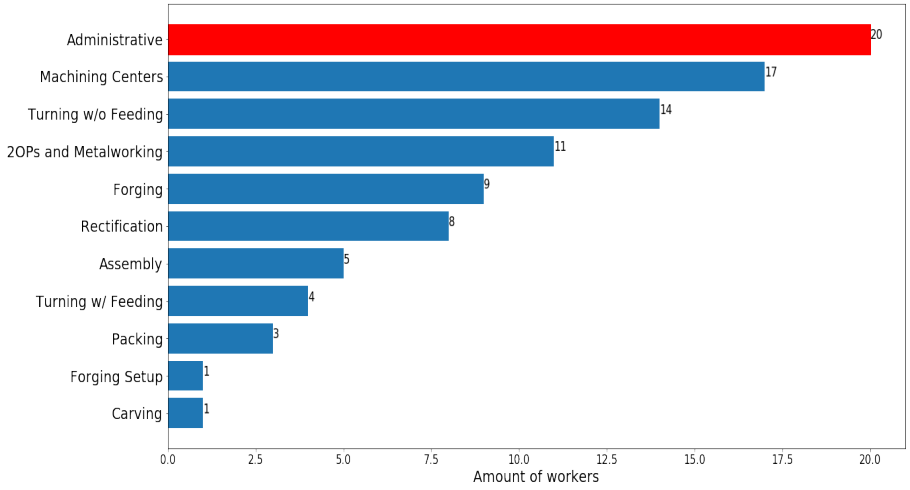

Figure 3.3: Worker Distribution by Sector

As the graph shows, the company is divided into 10 different sectors. Outlined in red is the administrative sector, the only sector which does not operate directly on the shop floor.

Two aspects should be taken into account. First, there are certain operations, such as polishing, that do not have a dedicated sector. Such operations are internal to each sector. Second, in each sector and between sectors there are varying degrees of competence. Although the company keeps track of this fact, there is no record in database of the level of expertise of each employee. Future plans include the implementation of these records.

JASIL holds a vast array of machines that are used in the manufacturing of their products. In regards to different references, JASIL has 3 cutting machines, 4 ovens for heating, 7 pressing machines for forging, 9 turning machines with feeder and 8 without feeder, 8 machining centers, 14 machines dedicated to rectification, 3 polishing machines, 4 stations of manual operations and 12 machines dedicated to the production of gears.

Tools play an extremely important part in performing an operation seeing that the tool is what comes into contact with the raw material and components, and not the machine itself. Thus the quality and state of the tool is a defining factor for the quality of the final product. JASIL has 748 different tools in database used in daily operations. These were, like the products, split into families based on function during production. In the *Others* family, minority classes such as oils, paints, and specific tools whose amount was negligible are aggregated.

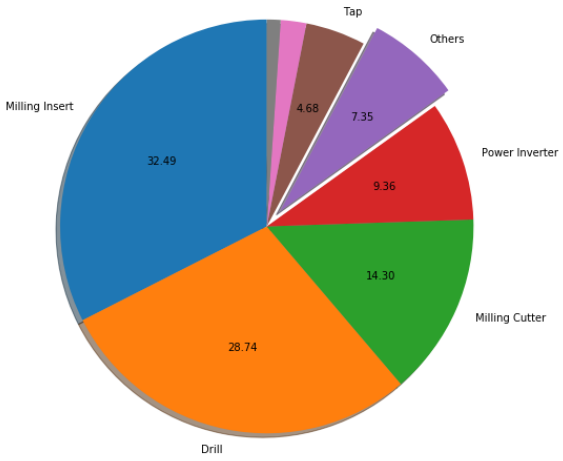

Figure 3.4: Relative Distribution of Tool Families

In February 2011, JASIL was certified with the ISO standard NP EN 9001 – Quality Management Systems – by SGS-ICS. Paired with their management style committed to continuous improvement, we get the confirmation of their commitment to quality[22].

The approach JASIL has taken to quality control is an individual sampling approach. Instead of taking a sample of n units in a given time interval, a single unit is taken from a batch, of which the size is typically 25 units, but can vary under certain circumstances. After sampling, the unit is analysed. Should the unit be up to standard, the entire batch is accepted as being up to standard. Otherwise, the batch is segregated and every single unit is inspected by the quality department.

Currently, JASIL has no data mining or derived from data mining solutions incorporated. Production is controlled using the ERP *Vanguarda*. This ERP collects and stores in a database all of the information regarding production, personnel and, other business aspects to which access was not given. JASIL also utilizes spreadsheet software like *Excel* as an internal solution to budgeting and accounting necessities.

3.3 Data Understanding and Analysis

This section comprises the phase of understanding, preparation and preliminary analysis of the data for the project.

3.3.1 Preliminary Database Analysis

One of the main steps in data understanding is to have proper knowledge of the context of the problem.

The importance of this step cannot be stressed enough. One of the major downfalls of data science is arriving at decontextualized conclusions i.e. conclusions that make sense mathematically but that have no practical value. Therefore, having the proper context is essential.

The first task for the appropriate contextualization of the problem is the understanding of the industry the company operates in, more specifically, understand the machines used, processes executed and pieces produced. This was done through the study of a preliminary diagnosis of the company.

To have the problem and context always present throughout the entire project, mind maps were produced. A mind map or a concept map is a layout of several concepts and how they connect. Mind maps maintain information organized and permit to view concepts both in detail and in a broader scope. Two mind maps were produced. To yield these maps, the software *CmapTools* was utilized. CmapTools permits easy elaboration and editing of concept maps. Both mind maps can be consulted in appendix B.

One of the tools utilized in the visual analysis of the data was *PowerBI*. PowerBI is a powerful tool that joins database management, information manipulation, and strong visualization tools in a single software. The main advantage of PowerBI is how intuitive and easy to use the software is.

However, PowerBI lacks the ability to plot more advanced and useful charts, such as advanced histograms and Pareto charts. This being the case, this software was used initially to quickly acquire a basic understanding of the structure and certain problems of the database, but was limited to a simple analysis.

Therefore, after an initial analysis and consideration of the data provided, all data analysis was done resorting to the Python programming language. All graphs that were elaborated in PowerBI and that merited being presented were reproduced again utilizing the plotting capabilities of Python for the sake of consistency.

A crucial step of the project was the identification and localization of relevant information within the database. Being that the project stems from a desire for quality control and improvement, identifying information regarding non-conformities was key.

Having this in mind, the Tables *Estruturas, Medicoes* and *Registos* were identified as being the most important. These 3 tables contain production information for a given batch, such as machine used, operation executed, product manufactured; but also contains information regarding quality, such as measurements taken and whether or not the product was up to specifications.

As was mentioned previously, the database as a whole presents problems and inconsistencies that hindered the progress and potential of the project. The most noteworthy of which are the problems with the Tables *Consumos* and *PIE*.

Firstly, Table *Consumos* lacked any connection with the remaining tables. The logical explanation for this is that the necessary tables and information were deleted when the database was prepared and delivered by the company. This table contains information regarding tools and materials used which could have been useful for data mining purposes.

Secondly, Table *PIE*, which directly relates to Table *Medicoes*, only related with 160 entries. While the table itself presents no problems and even contained information that could have been

useful, although not crucial, the lack of a meaningful amount of related entries ultimately rendered the table useless.

Also, the database was highly unbalanced in the proportion of conforming and non-conforming instances. This problem was tackled further on.

Having understood the database as a whole, the next step was to define the factors that were important from a data analysis standpoint.

Also, keeping in mind the concepts outlined in the mind maps produced beforehand, the following factors were considered relevant to further develop the project:

- Machine used
- Product manufactured
  - Product family
  - Product dimension and dimension limits
- Operation executed
- Measurements taken
- Final acceptance or rejection of the piece
- Motive for rejection, should that be the case

Note that the dimension limits are not necessarily measurable limits. Some represent variables, can be measured using the appropriate tool and the result can take any continuous value. However, the results of certain operations are binary in nature, representing a feature of the piece, and therefore the value introduced and limits are also binary.

3.3.2 Non-Conformity Analysis

Seeing as the main goal of the project is the improvement of quality through the reduction of non-conformities, analysing the current state of non-conformities is a crucial step.

Overall since 2005, throughout all of the product lines and operations, JASIL presents an average non-conformance rate of 4.21%, as is illustrated in Figure 3.5. This indicates that, despite a commitment to quality, Six Sigma standards are not being met. An analysis of the rate of non-conformance throughout the years, illustrated below in Figure 3.6, shows that, despite the existence of outlier years, this rate has stayed somewhat constant.

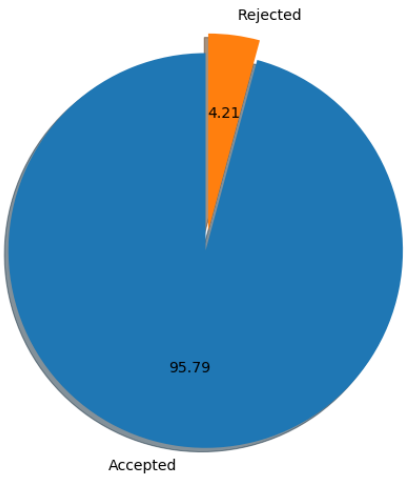

Figure 3.5: Overall Distribution of Conformance

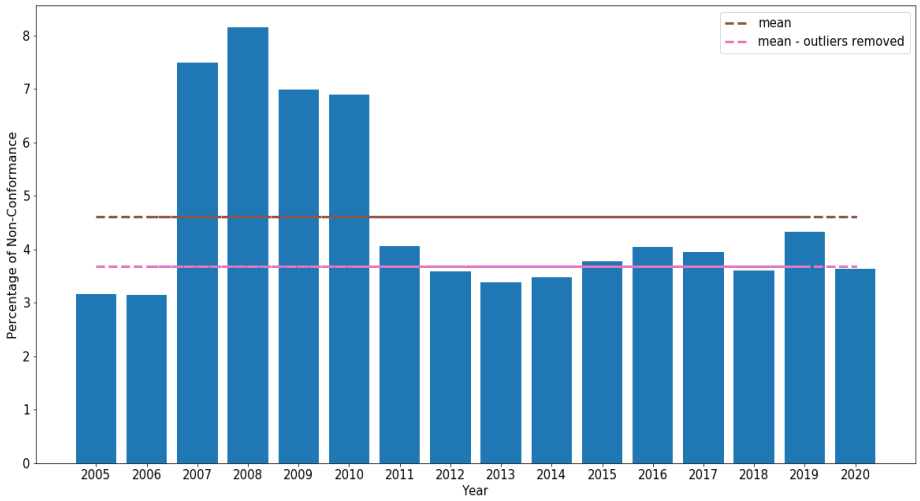

Figure 3.6: Rate of Non-Conformance by Year

Having done this, Pareto charts were elaborated to better understand which factors and areas were producing the largest number of non-conformities.

Besides the fact that this aids in identifying the areas which require the most immediate intervention, this analysis also permitted the isolation of factors that had a sufficient number of cases should a singular analysis ever be implemented.

The first one to elaborate should be the one detailing the reasons that demanded the rejection. Due to the style of quality control that the company utilizes, the fact that the majority of the pieces

are rejected in final testing was expected. However, this withholds important information, since this type of rejection does not discriminate the flaw that the piece had. This information could be useful to better understand what underlying problems exist.

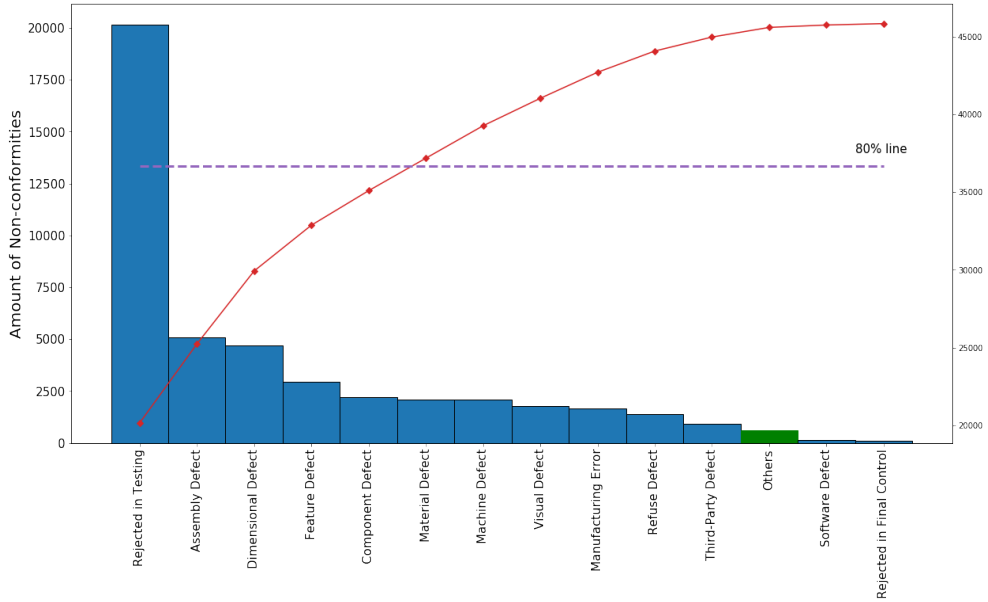

Figure 3.7: Non-Conformity Pareto chart by Reason for Rejection

Following this, Pareto charts were elaborated for the main objects of study: product families, operations, and machines, respectively. Certain aggregations were made in an *Others* class, marked in green. Note that, for readability purposes, the absolute and cumulative parts of the Pareto charts were written in different axes.

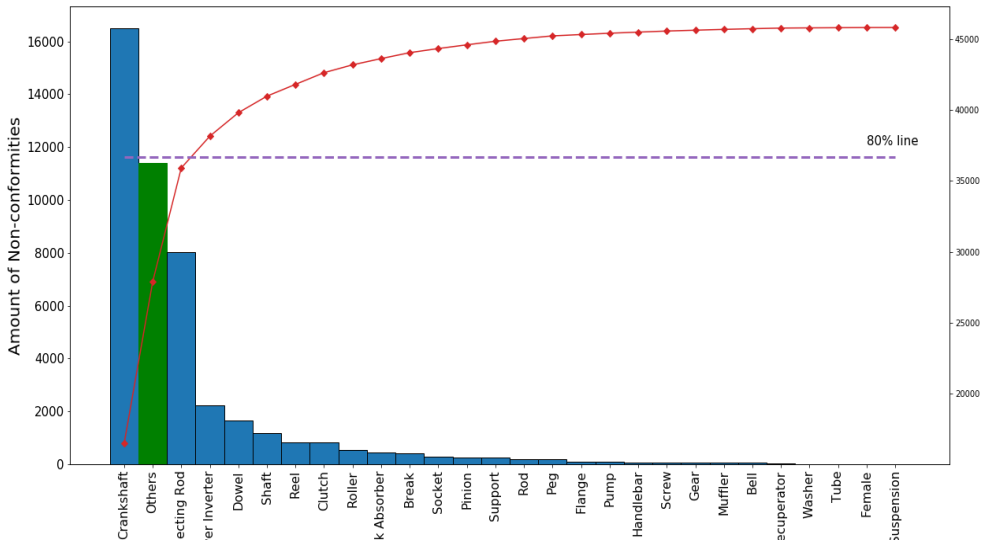

Figure 3.8: Non-Conformity Pareto chart by Product Family

Figure 3.8 visually illustrates the fact that JASIL produces mostly custom-made products since the minority classes aggregated in the *Others* class make up the second class with the greatest volume of non-conformity.

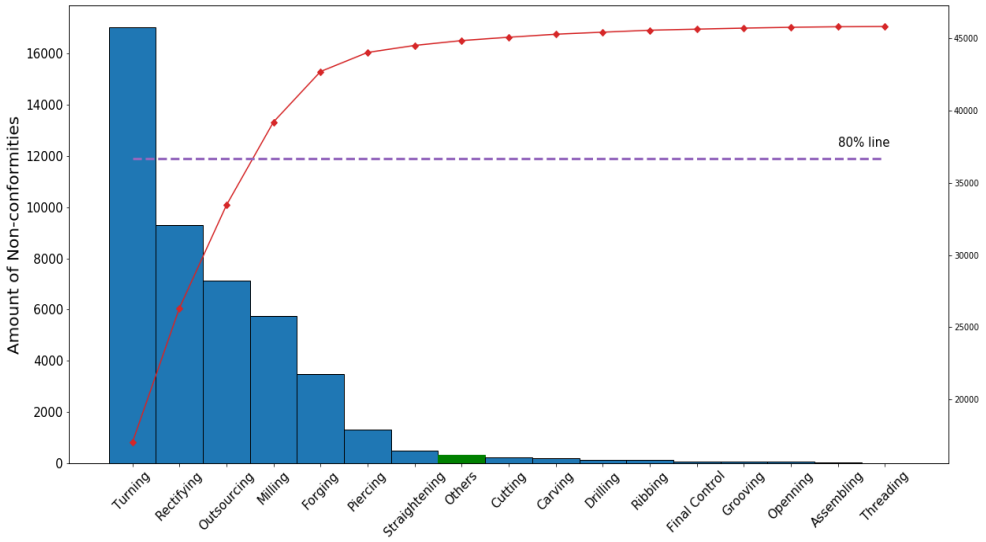

Figure 3.9: Non-Conformity Pareto chart by Operation

In the specific case of the Operations chart, the class *Outsourcing*, whilst being the third class with the most cases, is an irrelevant class, since outsourcing operations cannot be attributed to the company.

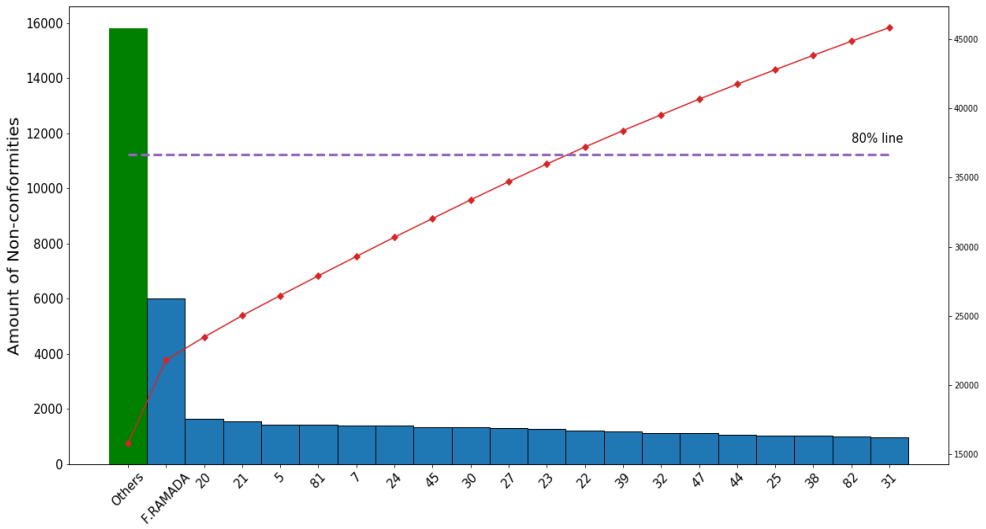

Figure 3.10: Non-Conformity Pareto chart by Machine

The charts show us that, more than the other 2, the machine used is a factor that more greatly limits the case being studied, being that production is spread throughout the entirety of the machines available. All of the descriptive tables containing the values of the Pareto charts can be consulted in appendix C.

3.3.3 Data Cleaning

Since the majority of the variables present in the database are categorical, the process of data cleaning was relatively simple.

Part of the process lay in the grouping of the products into the appropriate families and the identification and clustering of identical operations that were designated with different names. This presented a challenge, particularly with the product references, since some products were listed in different languages in the database.

There exist a multitude of mistakes in the columns containing information on dimensional limits of products, which are of the few numerical variables present in the database. Some instances in the database also presented impossible dates of entry (entries that occurred in the year 3000, for example).

Since the specifications of a product need to be exact, techniques such as averaging and smoothing could not be applied. However, with information and help provided from the company, the errors were successfully corrected.

3.4 Model Development and Results

This section describes the modelling phase of the project, beginning with the selection of models and the creation of potential predictors, finishing with the selection of the model with the best performance, and the evaluation of the results obtained. Due to the often iterative nature of this type of project, the application of the models is accompanied by the explanation for the selection of that particular model.

As was shown in Figure 3.5 the database is highly unbalanced in the proportion of non-conformities. To properly elaborate models, the dataset was balanced in order to have an equal number of conforming and non-conforming instances, sampling randomly all other factors. This approach avoids significant losses of data and the insertion of artefacts. SMOTE is an alternative method for balancing the data that was not used in this context due to its complexity[24].

Unless stated otherwise, all of the models were balanced to have the same amount of conforming and non-conforming instances and the dataset used was split into 70% for training purposes and 30% for testing purposes.

3.4.1 Elaboration of an Elementary Case

With the knowledge obtained from the phase of data understanding, the decision was made to isolate an elementary case to be studied.

An elementary case is a defining set of variables that define and constrain a part of the system. The purpose of the segregation of this case was to simplify the application of algorithms and remove confounding effects.

The elementary case should fully define a case of production but should also have a sufficient amount of instances of non-conformity occurring in the database so that the results obtained have validity. The minimum amount of instances of non-conformity the case should have was defined as 200.

Knowing the existing variables, the decision was made that the elementary case would be defined by the process executed, the machine utilized, the product being manufactured, and a certain dimension of the mentioned product. However, the filtering of the database did not provide an elementary case with a sufficient number of instances.

An alternative was proposed which consists of substituting the product and dimension with the upper and lower rejection limits, following the idea that an operation can be described with the limits that are imposed. Following this rationale, the following 3 elementary cases, presented in Table 3.1, were discovered by crossing the defined variables and selecting the cases with the minimum amount of instances previously defined.

Operation	Machine	Upper Limit	Lower Limit
Turning	Machine 27	1	1
Turning	Machine 22	1	1
Turning	Machine 30	1	1

Table 3.1: Identified Elementary Cases

Two issues should be noted. Firstly, only the machine variable changes. All other variables remain constant throughout the 3 cases. Secondly, the upper and lower limits both take the value of "1". This occurrence indicates that the operation is defining a feature of the piece.

This last point raises concerns due to the fact that features differ between products but are represented with the same limits in the database. This being the case, there exists the possibility of for there to be confounding of characteristics, even though the similarity is assured since the operation performed is the same. Knowing this, the decision was made to elaborate broad models and only afterwards apply restrictions on a case-by-case basis.

3.4.2 Apriori Algorithm

The first algorithm that was applied was the Apriori algorithm.

The Apriori algorithm is a data mining algorithm used to discover association rules between items in large relational databases, first proposed in 1994. The algorithm searches the database for combinations of items that occur frequently and infers rules of consequence between the mentioned items[25].

The main advantage of this algorithm is the relatively high speed of application to large, and often still untreated, databases. The association rules retrieved can then be used to create predictors to be used in more advanced models. However useful this model may be, the capacity for prediction is very limited, being that this model considers all items as categorical and all the items are decontextualized i.e. should an item appear in columns that have different meanings, the model will treat them as if they are the same.

This model was applied to the totality of the database with the goal of retrieving association rules that had as consequence the as occurrence of a conformity or non-conformity. The only conditions imposed were the restriction to the elementary cases that were defined previously and the removal of columns that could not provide rules with meaning e.g. counters.

The application of this model did not yield any useful rules. For example, one of the rules retrieved was {25.0000} -> {Approved} (the first bracket representing the antecedent and the second the consequent), which has no actual meaning. This is due, in part, to the fact that the database is highly unbalanced in the proportion of instances of non-conformities contained. Since this is not a model whose primary purpose is prediction, balancing techniques that are valid for other models, cannot be applied in this case.

3.4.3 Creation of Predictors

Before any prediction algorithms could be applied, predictors must be extracted or inferred upon from the relevant factors selected.

Besides the operation executed, in the database there were no direct predictors i.e. none of the factors had generic or scientific meaning. Therefore, predictors had to be synthesized from the existing data.

The created predictors were the following:

- **Sensitivity of the Process**
  The name *Sensitivity* was given to difference between the upper and lower rejection limits of a certain operation on a given dimension of a product, expressed in millimeters. This predictor was created with the hypothesis that a higher lower sensitivity (lower difference between the limits) would result in a higher rate of non-conformance.

  - **Relative Sensitivity of the Process**
    As the name suggests, this is a variation of the previous predictor, calculated by dividing the Sensitivity by the average of the limits. This predictor was created to account for cases with the same Sensitivity and overall larger or smaller limits and to test the influence this might have.

- **Dimension Count**
  An adimensional integer, as a way to estimate the complexity of a product, the amount of dimensions that are operated upon and controlled was taken into account. Empirically, this

makes sense as a predictor of non-conformance, since the higher the number of operations executed, the higher the risk of non-conformance.

- **Product Family**
  The products themselves offer little predictive capability, in part due to the amount existent, in part due to the specificity of each one. However, since the project is focused on the precision metalworking industry, the family to which the product belongs is a potential differentiating factor with predictive capability.

- **Machine Non-Conformance Rate**
  The rate of non-conformance of a machine (calculated using the amount of accepted and rejected pieces) acts as an estimator of the performance ability of the machine.

  - **Dynamic Machine Non-Conformance Rate**
    While the predictor from which this one derives is constant, the dynamic non-conformance rate assumes that the non-conformance rate of a machine can change over time, for example due to wear. This makes the model more capable of adapting should there be any re-training in the future.

- **Mean Time Between Non-Conformance and Frequency of Non-Conformance of a Machine**
  Two similar predictors being that both provide an interval for non-conformance. Also predictors that define the machine used, these were included in order to provide versatility to the model and to account for periods were production is stopped or rushed. The first of these was calculated assuming that the machine operates 16 hours per day and 5 days per week.

- **Number of Production in a Machine**
  A measure of how long the machine has been operating continuously, from here on out referred to as the number of production. An integer that indicates how many products have passed through a given machine in a given day. The reasoning behind this predictor is to account for stress related factors that may occur and accumulate throughout the day, such as overheating. While specific information on these factors is not available, the assumption can be made that, should they be verified, their effect is detrimental. This predictor assumes the machine is turned on and off on a daily basis.

- **Batch Quantity**
  How many items were included in the sampled batch. This predictor was included to test the dependency of non-conformities to batch size.

3.4.4 One Dimensional Analysis

While some predictors are straightforward in nature and the influence exerted on the outcome can be assumed as true, such as the rate of non-conformance of a machine, the same assumption cannot be made for other predictors such as the number of dimensions or the number of production.

This being the case, a one-dimensional analysis was first conducted to certain predictors to preemptively understand their potential connection with non-conformance.

The process of analysis was the same for all variables. The factor was isolated and for each value of the factor, the rate of non-conformance was calculated. Considering, scatter plots were elaborated and simple linear regressions were performed to add statistical significance to the plot, should there be any suspicion of any underlying relationship.

- **Number of Dimensions**

Before isolating this factor, the expectation was that as the number of dimensions increases, the rate of non-conformance would also increase due to the increased complexity of the piece and the larger number of operations required.

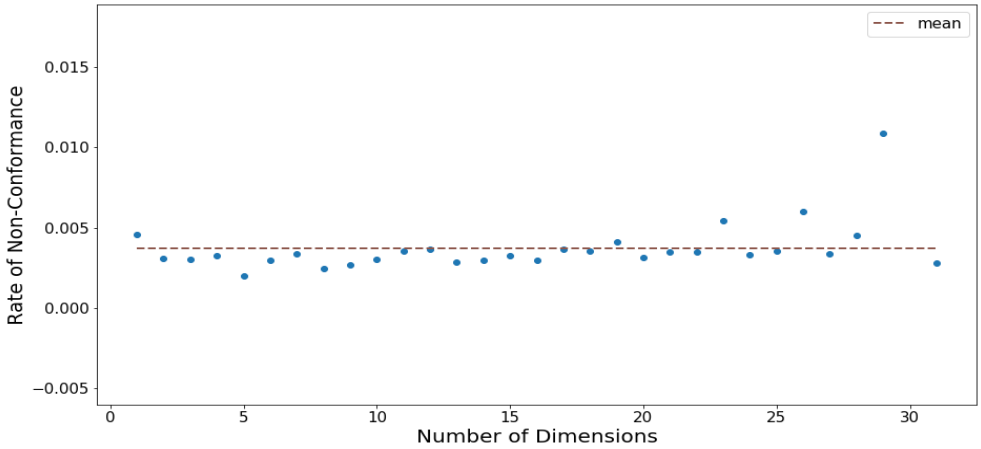

Figure 3.11: Non-Conformance Rate by Number of Dimensions

As seen in Figure 3.11 however, the rate of non-conformance appears to remain constant as the number of dimensions increases, fitting with mean of all values. Despite this, the value of R^2 (the value that measures how well the model fits the data points) is only 0.1944. A good model would be expected to have a R^2 score of around 0.9.

- **Number of Production**

The expectation for this factor was also that the rate of non-conformance would increase as the number of production increases, due to the accumulation of stress factors in a machine.

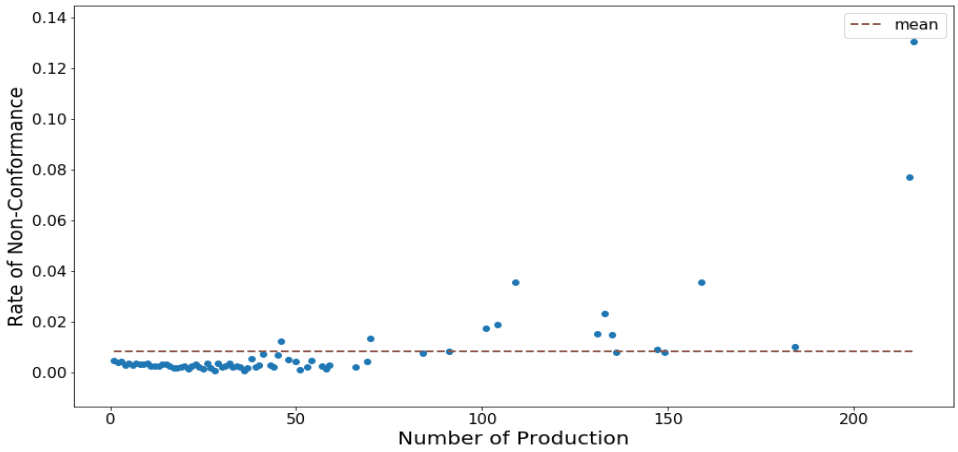

Figure 3.12: Non-Conformance Rate by Number of Production - Overall

Figure 3.12 evidences that the influence of the number of production is not linear (therefore no linear regression was applied) and the rate of non-conformance appears to increase significantly as the number of production reaches high enough values.

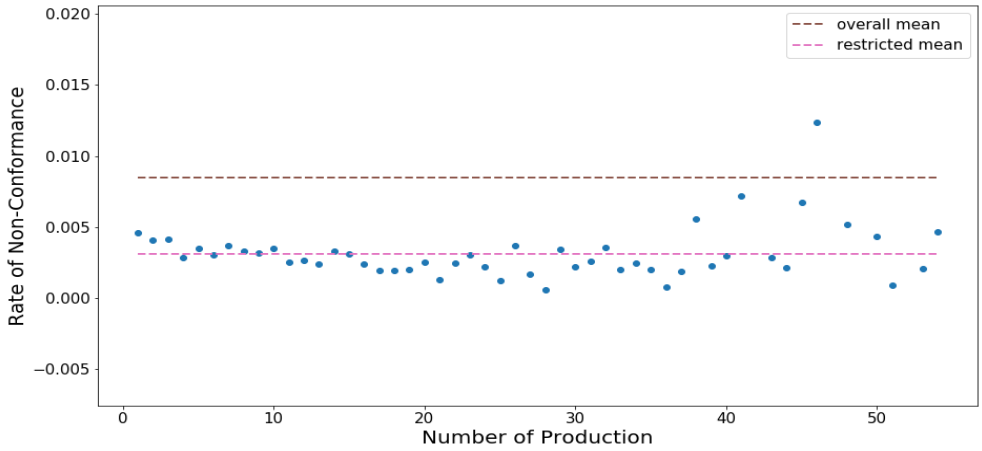

Figure 3.13: Non-Conformance Rate by Number of Production - Filtered

However, when restricted to values below 50, the rate of non-conformance appears to remain constant, as is shown in Figure 3.13. This can hint at the possibility that spreading production throughout all the available machines has an influence on lead time, but also quality. Still, the numbers are not consistent enough to safely make this conclusion.

- **Sensitivity**

Following the Number of Production, a plot for the rate of non-conformance considering the Sensitivity of the process was elaborated.

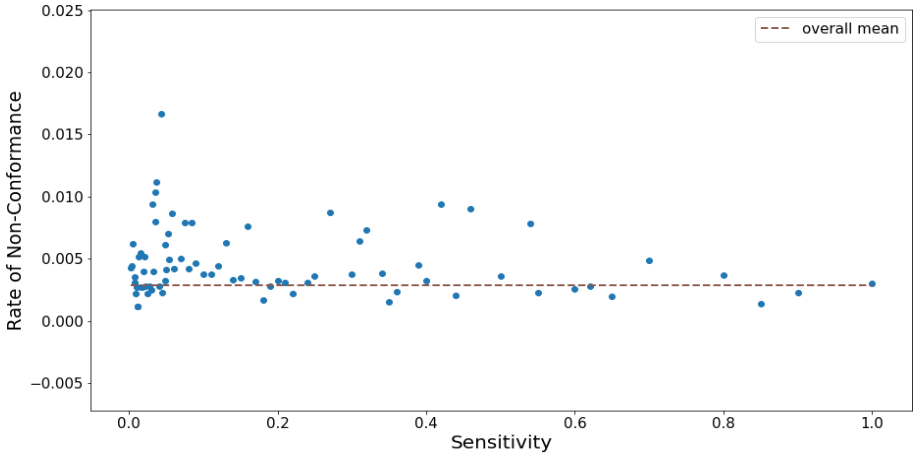

Figure 3.14: Non-Conformance Rate by Sensitivity

As was mentioned before, the expectation was that as the value of the Sensitivity increased, the rate of non-conformance would decrease. Given that there exist outliers with extreme values of Sensitivity, a representative sample was used to elaborate the plot by limiting the maximum Sensitivity to 1 millimeter.

Figure 3.14 shows that no direct connection can be made between these 2 variables. The scatter plot is seemingly random.

3.4.5 Random Forest Algorithm

The first model to be applied with predictive capability was the Random Forest model.

This model was chosen due to the simplicity of applicability and customization and capability of easy adaptation to large quantities of variables.

The most important hyperparameters for the Random Forest model are the number of estimators and the maximum depth of each tree and should always be customized. However, in this application, other hyperparameters were also iterated upon. The following are the values for the hyperparameters of the optimal model, followed by the meaning of each one.

- **Number of Estimators:** 800
  Number of decision trees generated by the model.

- **Maximum Depth:** 50
  Maximum number of splits performed in each tree.
  Originally, the optimal value of this parameter was 100. This was manually reduced in order to prevent the model from overfitting.

- **Minimum Number of Samples per Split:** 5
  The minimum number of samples required for a split to occur in an internal node.

- **Minimum Number of Samples in Leaf:** 1
  The minimum number of samples required to be in a leaf node.

- **Maximum Number of Features:** *Sqrt*
  Sqrt is an abbreviation of square root. The maximum number of features considered when looking for the best split will be the square root of the total number of features.

Also, bootstrap was not used in the development of this model, meaning that the entire dataset was considered when building each tree.

Having determined the optimal hyperparameters, 10 iterations of the model were created using different subsets of the same size. This was done due to the small amount of data that was being used for training and testing purposes and with the goal of giving further validity to the metrics retrieved from the model. This is an alternative approach to the 10-fold cross validation. After this iterative process, the mean of the metrics was retrieved and is presented in Table 3.2.

Accuracy	Precision	Recall	F1 Score
83.83%	0.8207	0.8634	0.8415

Table 3.2: Random Forest Performance Metrics - Overall

TN = 2404	FP = 551
FN = 399	TP = 2522

Table 3.3: Random Forest Confusion Matrix - Overall

Being a binary classifier, an accuracy value of 84% is somewhat lacking. Feature selection was used to improve the performance of the model, resorting to automated techniques for feature selection. Feature selection was done by removing the factors to which the model attributed negligible importance. The Relief algorithm was a potential alternative[26]. The model removed the binary variables regarding the product family and the binary variables regarding the operation.

Accuracy	Precision	Recall	F1 Score
85.47%	0.8403	0.8737	0.8567

Table 3.4: Random Forest Performance Metrics - Features Selected

TN = 2470	FP = 485
FN = 369	TP = 2552

Table 3.5: Random Forest Confusion Matrix - Features Selected

As an alternative to automated feature selection, with the hope of further improving these results, manual restrictions were applied. Firstly, 5 models were elaborated, restricting to 5 different

operations. The same predictors were used, except those that were rendered constant by this restriction. For the sake of consistency, these models used the same hyperparameters as the general model. The metrics extracted for each of the restricted models are written in Table 3.6.

	Accuracy	Precision	Recall	F1 Score
General Model	85.47%	0.8403	0.8737	0.8567
Turning	87.08%	**0.8561**	0.8920	0.8737
Milling	**87.15%**	0.8469	**0.9056**	**0.8751**
Rectifying	82.39%	0.8129	0.8481	0.8296
Forging	81.98%	0.7993	0.8595	0.8273
Piercing	81.60%	0.8102	0.8409	0.8232

Table 3.6: Random Forest Performance Metrics - Restricted by Operation

As is evidenced in Table 3.6, some of the models restricted by operation outperform the general model while others underperform the general model. Following this, hoping to further improve the results, the model was further restricted to both a single machine and a single operation. As a representative sample, 5 machines that executed turning were isolated and the model was retrained with this restriction.

	Accuracy	Precision	Recall	F1 Score
General Model	85.47%	0.8403	0.8737	0.8567
Machine 1	**89.70%**	0.8695	**0.9407**	**0.9036**
Machine 2	89.66%	0.8800	0.9273	0.9027
Machine 3	89.21%	**0.8804**	0.9196	0.8993
Machine 4	87.64%	0.8635	0.9035	0.8828
Machine 5	87.64%	0.8580	0.9055	0.8808

Table 3.7: Random Forest Performance Metrics - Restricted by Operation and Machine

As Table 3.11 evidences, all 5 models outperformed the general model, being that the model for machine 1 performed the best.

3.4.6 SVM

Following the Random Forest algorithm, an SVM was developed in a classification variant. This model was chosen in part due to the previous research done on predictive quality and also because of the expected excelling performance of the SVM on binary classification problems. For the SVM model, since the manual restrictions resulted in the optimal performance, no automated techniques of feature selection were applied.

For this model, the following 3 hyperparameters were defined using techniques of automated search.

- **C:** 10
  Regularization parameter.
- **Kernel:** *rbf*
  Rbf is an abbreviation of Radial Basis Function. Defines the function that will generate new features.
- **Gamma:** 1
  Parameter that regulates the influence of the new features created.

As with the Random Forest model, 10 iterations of the model were created using different subsets of the same size in order to validate the results obtained in Tables 3.8 and 3.9.

Accuracy	Precision	Recall	F1 Score
80.04%	0.8716	0.7018	0.7775

Table 3.8: SVM Performance Metrics

TN = 2653	FP = 302
FN = 871	TP = 2050

Table 3.9: SVM Confusion Matrix

Following this, also as with the Random Forest model, restrictions were made to the SVM model, first to the operations executed and then to both operation and machine.

	Accuracy	Precision	Recall	F1 Score
General Model	80.04%	0.8716	0.7018	0.7775
Turning	84.11%	0.9001	**0.7678**	0.8287
Milling	**84.79%**	**0.9128**	0.7676	**0.8339**
Rectifying	78.44%	0.8730	0.6671	0.7516
Forging	75.19%	0.8373	0.6200	0.7081
Piercing	77.04%	0.8539	0.6525	0.7355

Table 3.10: SVM Performance Metrics - Restricted by Operation

	Accuracy	Precision	Recall	F1 Score
General Model	80.04%	0.8716	0.7018	0.7775
Machine 1	86.13%	0.8696	**0.8584**	**0.8638**
Machine 2	**86.27%**	0.8979	0.8313	0.8623
Machine 3	84.38%	**0.9024**	0.7895	0.8400
Machine 4	84.68%	0.8990	0.7951	0.8422
Machine 5	84.53%	0.8954	0.7876	0.8366

Table 3.11: SVM Performance Metrics - Restricted by Operation and Machine

3.4.7 Neural Network Algorithm

Finally, the last model to be applied was a Neural Network. The Neural Network was chosen as this model provides the most versatility in terms of being able to adapt to future retraining and inputting new variables.

A Neural Network was applied only to the overall case. This was done in an attempt to reach the same performance values of the various specific cases with the overall case. Since the predictors are all numeric and binary in nature, the architecture of the network is fully connected, strictly linear, with 1 input layer, 4 hidden layers, and 1 output layer. Also, due to the nature of this algorithm, the variables were standardized, a process that was not required with the previous models.

The model was trained through 15000 iterations of the dataset, using a batch size of 1024, using the Adam optimizer. The values of the loss function (a cost function that is minimized during the training of the model) are represented below in Figure 3.15. The value of the loss function is trending towards 0.1. The code utilized to model and train the Neural Network can be consulted in D.

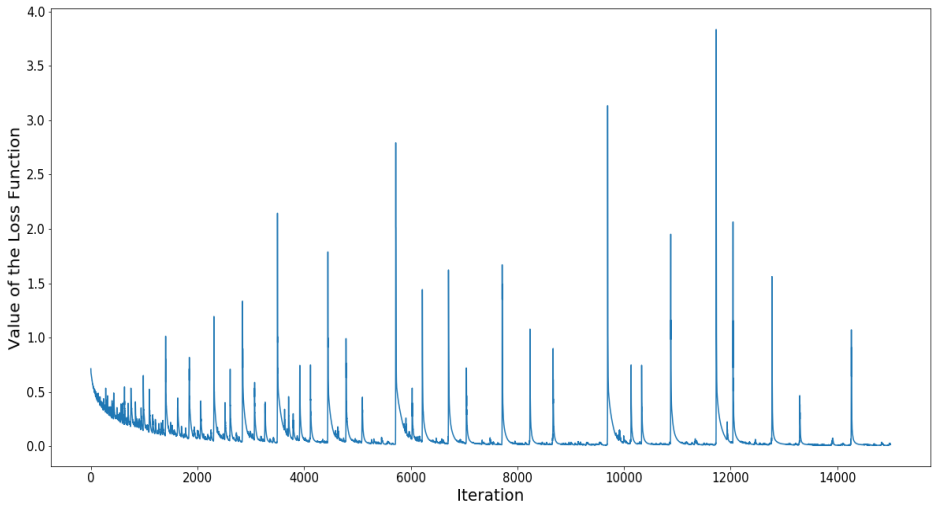

Figure 3.15: Loss Function Throughout Iterations

The metrics and confusion matrix extracted from the testing of the model are represented in Tables 3.12 and 3.13 respectively.

Accuracy	Precision	Recall	F1 Score
77.25%	0.7506	0.8121	0.7801

Table 3.12: Neural Network Performance Metrics

TN = 2167	FP = 788
FN = 549	TP = 2372

Table 3.13: Neural Network Confusion Matrix

3.5 Results Evaluation

To effectively compare the three models produced, the performance metrics for all three were grouped in a table. First the models for the overall scenario will be compared, followed by the models that were restricted to certain operations and machines.

	Accuracy	Precision	Recall	F1 Score
Random Forest	**85.47%**	0.8403	**0.8737**	**0.8567**
SVM	80.04%	**0.8716**	0.7018	0.7775
Neural Network	77.25%	0.7506	0.8121	0.7801

Table 3.14: Performance Metrics Comparison - Overall Models

From the metrics extracted, the model that performs the best is the Random Forest model. This confirms the expectations of the research done previously. According to the research, in a classification problem, the Random Forest model should outperform both the SVM model and the Neural Network [27].

Machine	Model	Accuracy	Precision	Recall	F1 Score
Machine 1	Random Forest	**89.70%**	0.8695	**0.9407**	**0.9036**
	SVM	86.13%	0.8696	0.8584	0.8638
Machine 2	Random Forest	89.66%	0.8800	0.9273	0.9027
	SVM	86.27%	0.8979	0.8313	0.8623
Machine 3	Random Forest	89.21%	0.8804	0.9196	0.8993
	SVM	84.38%	**0.9024**	0.7895	0.8400
Machine 4	Random Forest	87.64%	0.8635	0.9035	0.8828
	SVM	84.68%	0.8990	0.7951	0.8422
Machine 5	Random Forest	87.64%	0.8580	0.9055	0.8808
	SVM	84.53%	0.8954	0.7876	0.8366

Table 3.15: Performance Metrics Comparison - Restricted Models

Table 3.15 presents the values for the Random Forest and the SVM restricted models. Previously, the comparison of the overall version with the restricted version of each model showed that all the restricted versions outperformed the overall version. Table 3.15 shows that, like the overall model, the Random Forest algorithm provided the best performing model.

Also, special importance should be given to the recall of the models. The recall of a model decreases as the amount of false negatives increases. In context, a false negative is a product that was classified as being conforming to specification but was actually non-conforming. This could result in the shipping of defective products, so a high value of recall is important. The Random Forest models prioritize recall over precision which goes in accordance to what was desired.

Chapter 4

Conclusions and Future Work

This chapter is divided into two parts. The first part is comprised of the discussion of the results obtained with the applied methodology. The second part elaborates on plans and possibilities for future work to be built upon what was done previously.

4.1 Conclusions

This thesis aimed to understand the impact of production factors on quality output, measured by the rate of non-conformance. To achieve this, state of the art machine learning algorithms were employed to test the predictive capabilities and potential of the predictors created. The results obtained have the ultimate goal of being employed in the context of production planning in the future.

Firstly the data regarding non-conformity was described, mainly resorting to Pareto charts, to understand visually what and how productive factors were impacting non-conformity. The acquisition and organization of this information permitted and facilitated the next steps of the project by helping to understand the potential factors of influence and how these could be converted to variables to be used.

For the processing of the relevant factors and elaboration of the predictors, firstly the Apriori algorithm was applied to the totality of the database. This was done with the hope that the association rules retrieved would prove useful in understanding underlying relationships. However, this process did not provide useful results. Posterior, predictors were elaborated by defining generic characteristics of each factor that could, in future research, be replicated due to their universality.

Three different algorithms were employed: the Random Forest algorithm, the SVM algorithm, and a fully connected, linear Neural Network. All three were applied resorting to all of the relevant production information, creating models that could control the entirety of production. The first two were employed in scenarios that were restricted by machine used and operation executed with the hypothesis that model performance would improve.

The comparison of the performance metrics retrieved from the overall models indicates that the Random Forest model has the best performance, correctly predicting 85.47% of the cases, and the

Neural Network performs the worst, correctly predicting 77.25% of the cases. The performance metrics of the models built with restricted scenarios lead to the conclusion that restricted models perform better than general models, the best model correctly predicting 89.70% of the cases. The Random Forest models also present preference for recall over precision, which is contextually beneficial, since this avoids the shipping of defective products.

These values retrieved attest to the validity and predictive potential of the predictors used in the development of the models. The models themselves can be used in future implementations with satisfying results and should be done on a machine and operation basis to achieve optimal results in a production planning context.

4.2 Future Work

The results obtained with the development of this thesis met the expectations that were set in the beginning. This being said, there is still potential for development.

As was mentioned previously, this project was developed with the intention of applying the results obtained in the context of production planning. To further validate these results in a practical setting, experiments should be conducted where the model should adapt to a plan of production and predict the outcome of non-conformity, evaluating the performance of the model once production is concluded.

Furthermore, there are still aspects that can still be explored with the hope of improving the performance of the models. For example, two factors that have not been taken into account are the tools used when executing operations and the material used for the manufacturing of the products. The influence of these factors may prove to be important and potentially further develop the performance of the models.

Chapter 5

Bibliography

[1] Shneiderman, Ben, Catherine Plaisant, and Bradford W. Hesse. “Improving Healthcare with Interactive Visualization.” *Computer* 46, no. 5 (2013): 58–66.

[2] Porter, Michael E. *Competitive Advantage: Creating and Sustaining Superior Performance.* New York: Free Press, 2004.

[3] Veneri, Giacomo, and Antonio Capasso. *Hands-on Industrial Internet of Things: Create a Powerful Industrial IoT Infrastructure Using Industry 4.0.* Birmingham U.K.: Packt Publishing Ltd., 2018.

[4] Alwood, Julian, J.M. Cullen. *Sustainable Materials - With Both Eyes Open.* UIT Cambridge, Cambridge: p. 51-54. 2012.

[5] Casper, S.t., A. Mehra, M.e. Farago, and R.a. Gill. “Contamination of Surface Soils, River Water and Sediments by Trace Metals from Copper Processing Industry in the Churnet River Valley, Staffordshire, UK.” *Environmental Geochemistry and Health* 26, no. 1 (2004): 59–67.

[6] Lacy, Peter, and Jakob Rutqvist. *Waste to Wealth: the Circular Economy Advantage.* Basingstoke, Hampshire: Palgrave Macmillan, 2015.

[7] Silva, Cristovão, Paulo Vaz, Luís Miguel D. Ferreira. *The impact of Lean Manufacturing on environmental and social sustainability: a study using a concept mapping approach.* 2013.

[8] Montgomery, Douglas C. *Introduction to Statistical Quality Control*. Hoboken: J. Wiley and Sons, 2013.

[9] Garvin, David A. "Competing on the eight dimensions of quality" *Harvard Business Review.* 1987.

[10] Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to Data Mining.* Harlow: Pearson, 2014.

[11] Chapman, Peter, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas Reinartz, C. Russell H. Shearer and Robert Wirth. *CRISP-DM 1.0: Step-by-step data mining guide.* 2000.

[12] Kim, Aekyung, Kyuhyup Oh, Hoonseok Park, Jae-Yoon Jung. "Comparison of quality prediction algorithms in manufacturing process." *ICIC Express Letters* 11, (2017): 1127-1132.

[13] Kao, Hung-An, Yan-Shou Hsieh, Cheng-Hui Chen, and Jay Lee. “Quality Prediction Modeling for Multistage Manufacturing Based on Classification and Association Rule Mining.” *MATEC Web of Conferences* 123 (2017): 29.

[14] Weiss, Sholom M., Amit Dhurandhar, Robert J. Baseman, Brian F. White, Ronald Logan, Jonathan K. Winslow, and Daniel Poindexter. “Continuous Prediction of Manufacturing Performance throughout the Production Lifecycle.” *Journal of Intelligent Manufacturing* 27, no. 4 (November 2014): 751–63.

[15] Bai, Yun, Zhenzhong Sun, Jun Deng, Lin Li, Jianyu Long, and Chuan Li. “Manufacturing Quality Prediction Using Intelligent Learning Approaches: A Comparative Study.” *Sustainability* 10, no. 2 (2017): 85.

[16] Rostami, Hamidey, Jean-Yves Dantan, and Lazhar Homri. “Review of Data Mining Applications for Quality Assessment in Manufacturing Industry: Support Vector Machines.” *International Journal of Metrology and Quality Engineering* 6, no. 4 (2015): 401.

[17] Cortes, Corinna, and Vladimir Vapnik. “Support-Vector Networks.” *Machine Learning* 20, no. 3 (1995): 273–97.

[18] Lieber, Daniel, Marco Stolpe, Benedikt Konrad, Jochen Deuse, and Katharina Morik. “Quality Prediction in Interlinked Manufacturing Processes Based on Supervised and Unsupervised Machine Learning.” Procedia CIRP 7 (2013): 193–98.

[19] Melhem, Mariam, Bouchra Ananou, Mohand Djeziri, Mustapha Ouladsine and Jacques Pinaton. "Product’s Quality Prediction with respect to equipments data." *IFAC-PapersOnLine* 48 (2015): 78-84.

[20] Li, Xia, Yiru Dai, and Jin Cheng. “Research On Neural Network Quality Prediction Model Based On Genetic Algorithm.” *IOP Conference Series: Earth and Environmental Science* 267 (August 2019): 042026.

[21] Thede, Scott. An introduction to genetic algorithms. *Journal of Computing Sciences in Colleges* 20, (2004).

[22] Jasil. 2020. *Jasil*. https://www.jasil.com

[23] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning.* Cambridge, MA: The MIT Press, 2017.

[24] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research* 16 (2002): 321-357.

[25] Agrawak, Rakesh and Ramakrishnan Srikant. "Fast Algorithms for Mining Association Rules." *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499*, (1994).

[26] Kira, Kenji and Rendell, Larry. " The Feature Selection Problem: Traditional Methods and a New Algorithm" *AAAI-92 Proceedings, pages 129-134*, (1992).

[27] Fernández-Delgado, Eva Cernadas, Senén Barro and Dinani Amorim. "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research* 15, no. 90 (2014): 3133-3181.

Appendix A

Database Diagram and Content Description

The following tables describe the tables contained in the database that was provided. Of note that certain terms are in Portuguese and may even contain grammatical errors.

The *Key* column indicates if the column is part of the primary key (with a *P*) of the table or part of a foreign key (with an *F*) of another table. Should the column be neither, the space will be blank.

The *Column Name* column indicates the name of the column.

The *Type* column indicates the type of variable of that column.

The *Column Description* column describes what that column represents practically and may contain other useful information.

Estruturas			
Key	Column Name	Type	Column Description
	OP	Integer	ID of the OP
	DtEntrega	Date	Estimated delivery date for the OP
	Estado	Integer	Current state of the OP
	Codigo	Integer	Code of the reference to produce
	Nome	Text	Name of the reference to produce
P	OPR_Counter	Integer	ID of the operation
	Operacao	Text	Name of the operation
	Nivel	Integer	Level of the operation within the OP
	Maquina	Text	Machine used
	QtBatch	Integer	Maximum amount in the control sample
	FolgaBatch	Integer	Tolerance to the amount in *QtBatch*

Table A.1: Structures Table Description

Consumos			
Key	Column Name	Type	Column Description
P	Counter_Consumo	Integer	ID of the Consumption
	OP	Integer	ID of the OP
	Maquina	Text	Machine used
	Tipo	Text	Type of consumption
	Codigo	Integer	ID of the consumption
	Nome	Text	Name of the consumption
	Quant		Quantity consumed
	Unid	Text	Unit of the consumption
	Utilizador	Integer	ID of the user
	DtConsumo	Date	Date and time of the consumption

Table A.2: Consumptions Table Description

Colaboradores			
Key	Column Name	Type	Column Description
P	Utilizador	Integer	ID of the user
	Nome	Text	Name of the user
	Estado	Text	Describes whether the user is active or inactive
	Competencias	Text	Section where the user is stationed

Table A.3: Personnel Table Description

Lotes			
Key	Column Name	Type	Column Description
P	Counter_Lote	Integer	ID of the production lot
	OP	Integer	ID of the OP
	Lote_Id	Integer	Number of the production lot (similar to the ID)
	qtOrig		original amount of the production lot

Table A.4: Batches Table Description

PIE (Plano de Inspeção e Ensaio)			
Key	Column Name	Type	Column Description
P	Counter_PIE	Integer	ID of the inspection plan
	PIE	Text	Name of the inspection plan
	Desenho	Text	Blueprint used for control
P	Prod	Integer	Numeric reference of the product
P	Operacao	Text	Operation executed
	Metodo	Text	Method of measurement used
P	Cota	Integer	ID of the dimension to be measured
	Cota_Descricao	Integer	Description of the dimension (specific to blueprint)
	LIR	Numeric Continuous	Inferior rejection limit
	LIS	Numeric Continuous	Inferior safety limit
	LSS	Numeric Continuous	Superior safety limit
	LSR	Numeric Continuous	Superior rejection limit

Table A.5: Planes of Inspection Table Description

Registos			
Key	Column Name	Type	Column Description
P	Counter_Reg	Integer	ID of the registry
F	OPR	Integer	ID of the operation undergone
	Maquina	Text	Machine used
	Quant		Amount registered
	Tipo_Reg	Text	Type of registry (approved/rejected)
	DataRegisto	Date	Date and time of registry
	Utilizador	Integer	ID of the user
	Nome	Text	Name of the user
	Rejeicao_Cod	Integer	ID of the reason for rejection
	Motivo_Rejeicao_Desc	Text	description of the reason for rejection (if any)
	Observacao	Text	observations made
F	CountLote	Integer	ID of the associated production lot

Table A.6: Registries Table Description

Medicoes			
Key	Column Name	Type	Column Description
P	Counter_Medida	Integer	ID of the measurement
F	Counter_Reg	Integer	ID of the registry
F	Prod	Integer	ID of the product
F	Operacao	Integer	ID of the operation
F	Counter_PIE	Integer	ID of the plan of inspection
F	Cota	Integer	ID of the dimension to be measured
	LIR	Numeric Continuous	Inferior rejection limit
	LIS	Numeric Continuous	Inferior safety limit
	LSS	Numeric Continuous	Superior safety limit
	LSR	Numeric Continuous	Superior rejection limit
	Valor_Introduzido	Numeric Continuous	Value of the measurement
	Result	Integer	Decision made (based on measurement)(1)
	Utiliz	Integer	ID of the user
	DtHora	Date	Date and time of measurement

Table A.7: Measurements Table Description

(1) 0 - approved; 1 - approved but outside safety limits; 2 - not approved

Appendix B

Concept Maps

The following are two still images of the mind maps produced during the data contextualization phase of the project.

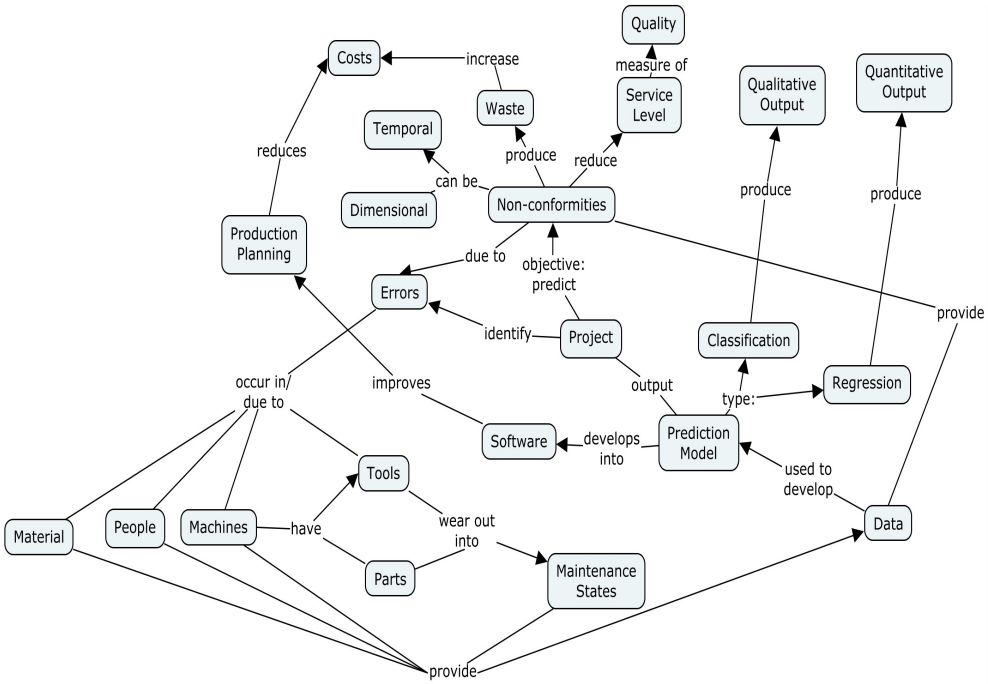

Figure B.1: Mind Map 1

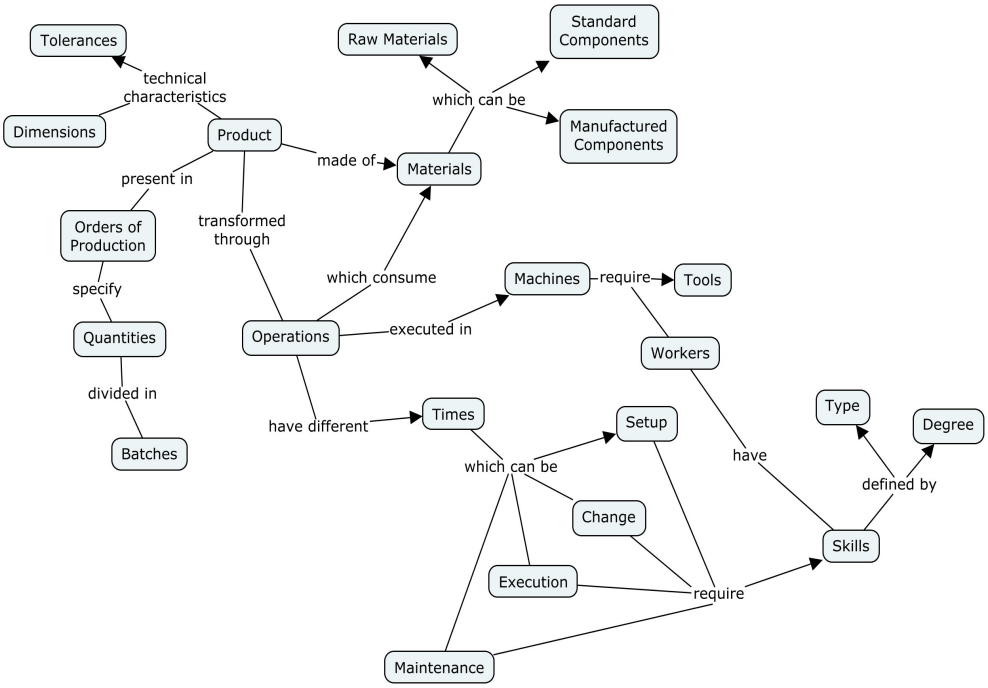

Figure B.2: Mind Map 2

Appendix C

Pareto Charts Descriptive Tables

This appendix contains tables with the values depicted in the form of Pareto charts in subsection 3.3.2 of chapter 3.

Reason	Absolute	Cumulative Absolute	Percentage	Cumulative Percentage
Rejected in Testing	20160	20160	44.0	44.0
Assembly Defect	5073	25233	11.07	55.07
Dimensional Defect	4690	29923	10.24	65.31
Feature Defect	2941	32864	6.42	71.73
Component Defect	2218	35082	4.84	76.57
Material Defect	2088	37170	4.56	81.13
Machine Defect	2084	39254	4.55	85.68
Visual Defect	1766	41020	3.85	89.53
Manufacturing Error	1672	42692	3.65	93.18
Refuse Defect	1366	44058	2.98	96.16
Third-Party Defect	899	44957	1.96	98.12
Others	620	45577	1.35	99.47
Software Defect	154	45731	0.34	99.81
Rejected in Final Control	91	45822	0.19	100

Table C.1: Non-Conformity Pareto chart by Reason for Rejection Table

Product Family	Absolute	Cumulative Absolute	Percentage	Cumulative Percentage
Crankshaft	16486	16486	35.98	35.98
Others	11411	27897	24.9	60.88
Connecting Rod	8019	35916	17.5	78.38
Power Inverter	2242	38158	4.89	83.27
Dowel	1655	39813	3.61	86.88
Shaft	1168	40981	2.55	89.43
Reel	832	41813	1.82	91.25
Clutch	824	42637	1.8	93.05
Roller	554	43191	1.21	94.26
Shock Absorber	440	43631	0.96	95.22
Break	418	44049	0.91	96.13
Socket	290	44339	0.63	96.76
Pinion	263	44602	0.57	97.33
Support	252	44854	0.55	97.88
Rod	191	45045	0.42	98.3
Peg	188	45233	0.41	98.71
Flange	93	45326	0.2	98.91
Pump	93	45419	0.2	99.11
Handlebar	79	45498	0.17	99.28
Screw	76	45574	0.17	99.45
Gear	60	45634	0.13	99.58
Muffler	56	45690	0.12	99.7
Bell	49	45739	0.11	99.81
Recuperator	43	45782	0.09	99.9
Washer	15	45797	0.03	99.93
Tube	12	45809	0.03	99.96
Female	10	45819	0.03	99.99
Suspension	3	45822	0.01	100

Table C.2: Non-Conformity Pareto chart by Product Family Table

Operation	Absolute	Cumulative Absolute	Percentage	Cumulative Percentage
Turning	17018	17018	37.14	37.14
Rectifying	9308	26326	20.31	57.45
Outsourcing	7127	33453	15.55	73.0
Milling	5758	39211	12.57	85.57
Forging	3494	42705	7.63	93.2
Piercing	1321	44026	2.88	96.08
Straightening	485	44511	1.06	97.14
Others	337	44848	0.74	97.88
Cutting	231	45079	0.5	98.38
Carving	209	45288	0.46	98.84
Drilling	142	45430	0.31	99.15
Ribbing	136	45566	0.3	99.45
Final Control	74	45640	0.16	99.61
Grooving	71	45711	0.15	99.76
Openning	57	45768	0.12	99.88
Assembling	38	45806	0.08	99.96
Threading	16	45822	0.04	100

Table C.3: Non-Conformity Pareto chart by Operation Table

Machine	Absolute	Cumulative Absolute	Percentage	Cumulative Percentage
Others	15801	15801	34.48	34.48
F.RAMADA	6012	21813	13.12	47.6
20	1662	23475	3.63	51.23
21	1555	25030	3.39	54.62
5	1448	26478	3.16	57.78
81	1421	27899	3.1	60.88
7	1400	29299	3.06	63.94
24	1390	30689	3.03	66.97
45	1350	32039	2.95	69.92
30	1343	33382	2.93	72.85
27	1313	34695	2.87	75.72
23	1276	35971	2.78	78.5
22	1224	37195	2.67	81.17
39	1185	38380	2.59	83.76
32	1137	39517	2.48	86.24
47	1132	40649	2.47	88.71
44	1082	41731	2.36	91.07
25	1044	42775	2.28	93.35
38	1043	43818	2.28	95.63
82	1014	44832	2.21	97.84
31	990	45822	2.16	100.0

Table C.4: Non-Conformity Pareto chart by Machine Table

Appendix D

Code for Modelling the Neural Network

This appendix contains the code used to model and train the Neural Network implemented during the project. The appendix is divided into 3 blocks, one used to define the Neural Network, one to define important parameters and one to train the Neural Network, in that order. Any text in blue within the code starting with a # symbol is a comment.

Block 1

```
# Begins the definition of the Neural Network.
class Net2(nn.Module):

    # Function to define the architecture of the Neural Network.
    def __init__(self):
        super(Net2, self).__init__()

        # Defines one layer of the Neural Network. The first number is the
        number of input nodes and the second is the number of output nodes.
        self.fc1 = nn.Linear(55, 120)
        # Normalizes the batch between layers.
        self.bn1 = nn.BatchNorm1d(num_features=120)
        self.fc2 = nn.Linear(120, 160)
        self.bn2 = nn.BatchNorm1d(num_features=160)
        self.fc3 = nn.Linear(160, 200)
        self.bn3 = nn.BatchNorm1d(num_features=200)
        self.fc4 = nn.Linear(200, 100)
        self.bn4 = nn.BatchNorm1d(num_features=100)
        self.fc5 = nn.Linear(100, 30)
        self.bn5 = nn.BatchNorm1d(num_features=30)
```

```
        self.fc6 = nn.Linear(30, 2)

    def forward(self, x):

        # Applies the rectified linear unit function.
        x = F.relu(self.bn1(self.fc1(x)))
        x = F.relu(self.bn2(self.fc2(x)))
        x = F.relu(self.bn3(self.fc3(x)))
        x = F.relu(self.bn4(self.fc4(x)))
        x = F.relu(self.bn5(self.fc5(x)))
        x = self.fc6(x)
        return x

# Creates an object with the Neural Network architecture that was defined.
net = Net2().to(device)
```

Block 2

```
    # Defines the optimizer; "lr" refers to the learning rate.
    optimizer = optim.Adam(net.parameters(), lr=0.001)

    # Defines the number of iterations, size of the training dataset and batch
    size, respectively.
    n_epochs = 15000
    numb_of_total_data_instances = 13710
    batch_size = 8192

    # Reshapes the training data to be used by the model.
    all_data = torch.tensor(X_train.to_numpy()).float().to(device)
    all_data_target = torch.tensor(y_train.to_numpy()).float().to(device)
    all_data_target = all_data_target.unsqueeze(1)
```

Block 3

```
# Defines the loss function to be used.
loss_criteria = nn.CrossEntropyLoss()

# Stores the progressive value of the loss function.
losses = []

# Initiates training cycle.
for i in range(n_epochs):

    # Stores batch losses.
    batch_losses = []

    # Cycle that applies the batch to the Neural Network.
    for batch_idx in range(1, ceil(numb_of_total_data_instances/batch_size)):

        batch_out = net(all_data[(batch_idx-1)*batch_size:batch_idx*
        batch_size, :])
        batch_target = all_data_target[(batch_idx-1)*batch_size:batch_idx*
        batch_size, :]

        batch_target_2v = torch.zeros(batch_size, 2).to(device)
        batch_target_2v[range(batch_target_2v.shape[0]), batch_target.cpu().T.long()]
        = 1

        # Computes loss value.
        loss = loss_criteria(batch_out, torch.max(batch_target_2v, 1)[1])
        batch_losses.append(loss)

        # Resets the Neural Network gradients.
        net.zero_grad()

        # Computs the change in loss for all parameters.
        loss.backward()
```

```
        # Updates parameters based on current gradients.
        optimizer.step()

    # Stores the average value of the loss function for that iteration.
    losses.append(sum(batch_losses)/len(batch_losses))

# Plots the graph showing the evolution of the loss function.
plt.plot(losses)
```