Framework for Process Improvement in Manufacturing of Metal Packaging

Eric Rautenbach

Thesis presented in partial fulfilment
of the requirement for the degree of
Masters of Engineering (Industrial Engineering)
in the Faculty of Engineering at Stellenbosch University

Supervisor: Dr. T. Dirkse van Schalkwyk

April 2022

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: April 2022

ABSTRACT

Due to increased competitiveness in the packaging industry, process improvement is important to give businesses an edge over their competition. This thesis represents a study of the application of machine learning for process improvement in metal can manufacturing. A five step process improvement framework based on the Six Sigma process improvement methodology and the CRISP-DM data science framework was developed. The framework consisted of different steps that included steps used in the Six Sigma process improvement methodologies as well as steps used in data science processes.The five steps were; Define, Understand, Model, Evaluate and Deploy (DUMED). The DUMED framework was used in a case study that predicted the axial load resistance of 2-piece metal food cans during the manufacturing process. The objective is to understand how axial load resistance relates to other factors in the process with the outcome that any changes made in the process will still deliver cans with suitable axial load resistance. A predictive model on axial load resistance will give enhanced capability to control axial load resistance, and will lead to less rejections and therefore less waste. A predictive model on axial load resistance can also supply valuable information on the possible viability for light weighting of material, which will have a decreased cost of raw material as a result and therefore hold financial benefit for the manufacturer. Various data science and machine learning principles were applied during the study related to data understanding, data assessing, data preparation, data modelling and model assessing. The framework was successfully applied in the case study, with the exception of the fifth step, deployment. The deployment phase will be dependent on further improvement of the predictive model. Machine learning was successfully used in the case study to develop a predictive model; the axial load resistance could be predicted within 2.3% of the actual values. The best results were obtained from using feature selected data obtained from a random forest feature selection algorithm that was modelled by using a gradient boost ensemble regression model. Machine learning was successfully applied to a metal package manufacturing line to predict quality characteristics of the final product and possibly bring about process improvement.

OPSOMMING

As gevolg van die toenemende kompetisie in die verpakkings industrie is proses verbetering belangrik om besighede 'n voorsprong oor hulle kompetisie te gee. Hierdie tesis is 'n studie van die gebruik van masjienleer vir proses verbetering in metaal blik vervaardiging. 'n Vyf stap proses verbeterings raamwerk wat gebaseer was op die Ses Sigma proses verbeterings metodologie an die CRISP-DM data wetenskap raamwerk was ontwikkel. Die vyf stappe was; definieer, verstaan, modeleer, evalueer, en ontplooi (DUMED, na aanleiding van die engelse akroniem). Die DUMED raamwerk was gebruik vir 'n gevallestudie wat die aksiale ladings weerstand van 2-stuk metaal kos blikke voorspel gedurende die vervaardigings proses. Verskeie data wetenskap en masjienleer beginsels was toegepas gedurende die studie relevant tot die verstaan van die data, assessering van die data, voorbereiding van die data, modelering van die data en die assessering van die data modelle. Die raamwerk was suksesvol toegepas vir die gevallestudie, behalwe vir die vyfde stap, naamlik die ontplooing. Die ontplooings fase sal afhanklik wees van verdere verbeteringe op die voorspellende data model. Masjienleer was suksesvol gebruik in die gevallestudie om 'n voorspellende model te ontwikkel; die aksiale ladings weerstand kon voorspel word tot binne 2.3% van die werklike waardes. Die beste resultaat was verkry deur die 'gradient boost' masjienleer algoritme toe te pas op 'random forest feature selected' data. Masjienleer was suksesvol toegepas op 'n metaal verpakkings vervaardigings lyn om kwaliteits eienskappe op die finale produk te voorspel en so moontlikke proses verbetering te bewerkstellig.

ACKNOWLEDGEMENTS

I would like to thank;

- Dr. Theuns Dirkse van Schalkwyk for all his guidance and time.
- Nampak for their sponsorship of both my studies as well as the project I used for my case study.
- Marius Biermann, Simon Allsop and Aneé Sieberhagen for their advice and proof reading.
- Jacobus for his patience and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND/OR ACRONYMS

ANOVA	Analysis Of Variance
CI	Continuous Improvement
CRISP-DM	Cross Industry Standard Process for Data Mining
CTC	Critical-To-Cost
CTQ	Critical-To-Quality
CTS	Critical-To Schedule
DoE	Design of Experiment
DMAIC	Define, Measure, Analyze, Improve, and Control
DR	Double Reduced
DRD	drawn and Re-Drawn
DUMED	Define, Understand, Model, Evaluate, and Deploy
DWI	Drawn and Wall-Ironed
ECCS	Electro-coated Chrome Coated Steel
OEE	Overall Equipment Effectiveness
FFCH	Factory Finished Can Height
FMECA	Failure Mode, Effects and Criticality Analysis
FTA	Fault Tree Analysis
IoT	Internet of Things
JIT	just In Time
KPCA	Kernel Principal Component Analysis
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
MAE	Mean Absolute Error
ML	Machine Learning
MRO	Multiple Response Optimization
OA	Orthogonal Arrays
OFAT	One Factor At a Time

PCA	Principal Component Analysis
PDCA	Plan, Do, Check, Act
RANSAC	RANdom SAmple Consensus
RMSE	Root Mean Squared Error
ROI	Return On Investment
R&R	Repeatability and reproducibility
RSM	Response Surface methodology
SBS	Sequential Backward Selection
SFS	Sequential Forward Selection
SIPOC	Supplier, Inputs, Processes, Outputs and Customer
SPC	Statistical Process Control
SR	Single Reduced
SS	Shainin System
SVM	Support Vector Machine
TOC	Theory Of Constraints
TQM	Total Quality Management

CHAPTER 1

INTRODUCTION

This dissertation reports on the development of a framework for process improvement that incorporates data science principles in manufacturing of metal packaging.

1.1 BACKGROUND

According to the McKinsey Global Institute (Manyika et al. 2012) developing economies will drive the demand for manufacturing globally in the near future. The report concludes that the use of analytics and information, together with skilled employment and advancement in machinery will drive growth in the manufacturing sector. Companies that recognize and act on these trends in manufacturing can shape their competitive environment and thrive. The economic environment holds more risk and uncertainty than before and for companies to succeed, foresight must be established by analytics to give companies the confidence to act on new developments. Metal packaging manufacturing industries in a developing continent, such as Africa, will need to develop adequate analytic tools and systems to increase their likelihood to succeed and stay ahead of the competition curve.

The packaging industry is a competitive industry. Metal packaging such as food cans, non-food cans, aerosol cans and beverage cans are constantly evolving to be able to stand out against competitor packaging. Manufacturers of metal packaging must continuously maintain a balance between consumers' needs and manufacturing costs while still ensuring that the metal packaging complies with all specifications and standards.

According to Wuest, Irgens, and Thoben (2014) Big Data has a tremendous amount of potential for manufacturing to predict the best outcome for the manufactured product. The traditional way of modelling cause and effect in manufacturing had reached its limitations due to the multi-dimensionality and complexity of production systems. Cluster analysis as well as supervised machine learning can be used to cope with data with high dimensionality without unreasonable effort. Even in advanced manufacturing systems there are variations in the input materials, products as well as in process parameters. When processes are inter-dependent (which is the case for most manufacturing operations), the variation in one sub-system can seem acceptable, but combined with other intertwined sub-systems in the process the total variability can lead to products that are not of acceptable quality.

A framework, to systematically approach process improvements in metal packaging manufacturing, can be employed when

- down gauging packaging materials. Down gauging is when a thinner metal plate is used to manufacture the metal packaging.
- designing of new shapes or sizes of packaging.

- developing new products for packaging in the packaging materials.
- using of new metal for manufacturing of the packaging metal.
- using new suppliers of packaging materials.
- improving quality on the process and the final product.

An approach to process improvement for the manufacturing of metal containers was introduced by following a framework that was designed to guide and measure the process. Current processes in Nampak Ltd. were used for a case study from a metal packaging manufacturer's point of view. The framework consisted of different steps that included steps used in the Six Sigma process improvement methodologies as well as steps used in data science processes. The framework was described and the probable outcomes were defined. The framework was applied to a case study to predict quality characteristics of 2-piece metal food cans during the manufacturing process. The results obtained were analyzed to determine the sustainability and viability of the changes. Nampak Ltd. would be able to increase its competitive advantage in the metal can manufacturing industry by using predictions from machine learning algorithms to reduce the wall thickness of 2-piece metal cans whilst still maintaining the axial load resistance and panelling pressure resistance of the cans. Reducing the incoming tinplate thickness, allows for reducing the mid-wall thickness of the cans which will reduce the manufacturing cost of the cans and can ultimately lower prices for Nampak's customers.

1.2 RESEARCH PROBLEM

The research problem statement is: How can data science, such as machine learning, be applied to process improvement in the metal packaging manufacturing environment? Applying data science to process improvement can assist Nampak to remain competitive through possible cost savings. Machine learning will be used as a data science tool since the availability of quality related data in manufacturing has much potential to apply machine learning . Huge amounts of data or process variables can make focused quality improvement difficult, but data-driven approaches such as ML can be used to overcome some of the challenges manufacturing faces today (Wuest, Irgens, and Thoben (2014)). Cost savings may be achieved by better process efficiencies, improved quality or by obtaining suitable products from more cost effective raw materials. As part of the process to solve the research problem a framework for process improvement in metal container manufacturing will be developed. This research approach will be to;

- look at a general overview of metal packaging.
- look at process improvement and frameworks as discussed in academic literature.
- elaborate on data science in manufacturing and describe the basic principles of how machine learning can be used.

- design a systematic framework that will be suitable to use for the research problem.
- conduct a case study in order to implement the framework at a real world metal packaging manufacturer.
- use various tools to analyze and interpret results as well as to discuss the meaningfulness of the outcomes.

1.3 RESEARCH OBJECTIVES

The objective of the study was to develop a framework for process improvement for manufacturing of metal packaging, incorporating principles of data science. In order to develop and demonstrate the utilization of the framework the following were proposed:

- A sufficient literature review to be conducted that relates to metal packaging manufacturing, process improvement, framework development and machine learning.
- An appropriate framework to be designed to systematically analyze and solve solutions optimally for different inputs with regards to process or packaging materials improvements.
- To identify and systematically test the critical variables, that form part of the process and/or packaging materials.
- The use of data from a process metal packaging manufacturing process to build predictive models to predict quality characteristics of the metal packaging, such as axial load resistance of a metal food can.
- The evaluation of the predictive capabilities of machine learning models for a process or on a product of metal packaging manufacturing.
- The discussion of possible improvement and deployment strategies of the machine learning models that were developed.

1.4 RATIONALE FOR RESEARCH

Nampak is Africa's largest diversified packaging manufacturer, producing metal, paper and plastic packaging. There are 18 Nampak sites in South Africa and Nampak has operations in 11 other Africa countries. The R&D facility is based in Cape Town and provides innovative solutions and services to the Nampak operations (Nampak 2021). Nampak has various projects running continuously, of different scale and importance. Various of these projects will benefit from following a process improvement framework.

PackagingSA (2018) states that metal packaging constitutes 9.2% (R65 billion) of the South African packaging market and is poised to grow . An advantage of metal packaging is the fact that it is infinitely recyclable. When steel or aluminium cans are recycled the aluminium or steel atoms reconstitute into their original atomic arrangements, therefore completely renewing the material for use again. South Africa currently has 73% recycling rates for metal packaging.

Process improvement has been studied and written about in academic submissions often and general frameworks for their associated processes have been developed. There is a lack of academic research on the development of a process improvement framework related to the metal packaging manufacturing industry in South Africa. Developing a framework that focus on improvements in metal packaging manufacturing will be beneficial for the specific case studies the research will be focused on as well as for any future improvement projects in packaging manufacturing systems.

The specific contributions this study will have is to outline and demonstrate the use of a framework that uses machine learning for process improvement on a 2-piece metal food can manufacturing line. The use of machine learning on such a manufacturing line allows for an understanding of how variables can influence the process and the end product. Traditionally, the technical knowledge and experience of employees were used to design process and product improvement trials. A structured framework that incorporates machine learning has the potential for the understanding of the process at a more complex level as what the previous process improvement methods allowed for.

1.5 RESEARCH APPROACH

The research approach consisted of a comprehensive literature review that focused on variables that can have an effect on the development of a framework for process improvement in metal container manufacturing. The research also looked at the development of a framework to guide and measure the process. The research methodology included the obtaining of samples. Samples were drawn from different production runs.

During the research process a framework were designed that incorporated steps for process improvement, with the aid of machine learning, in packaging systems. In broader terms, the research strategy included the following main steps;

- The formulation of research question and objective. To answer the research question the research methodology as outlined in this section was followed.
- A literature study based on research of topics related to the research question included the following topics;
  - A review of metal packaging manufacturing.
  - A review of process improvement methodologies.

  - A review of framework development with a specific focus on manufacturing.
  - A review of machine learning.
- The design of a framework on process improvement that included data science principles and which could be applied in metal packaging manufacturing.
- The implementation of the designed framework on process improvement in metal packaging manufacturing by
  - applying the framework on a real world case study.
  - analysing and discussing the case studies in light of the applied framework.
- Conclude and recommend improvements and further research.

The proposed research strategy is shown in **Figure 1.1**.

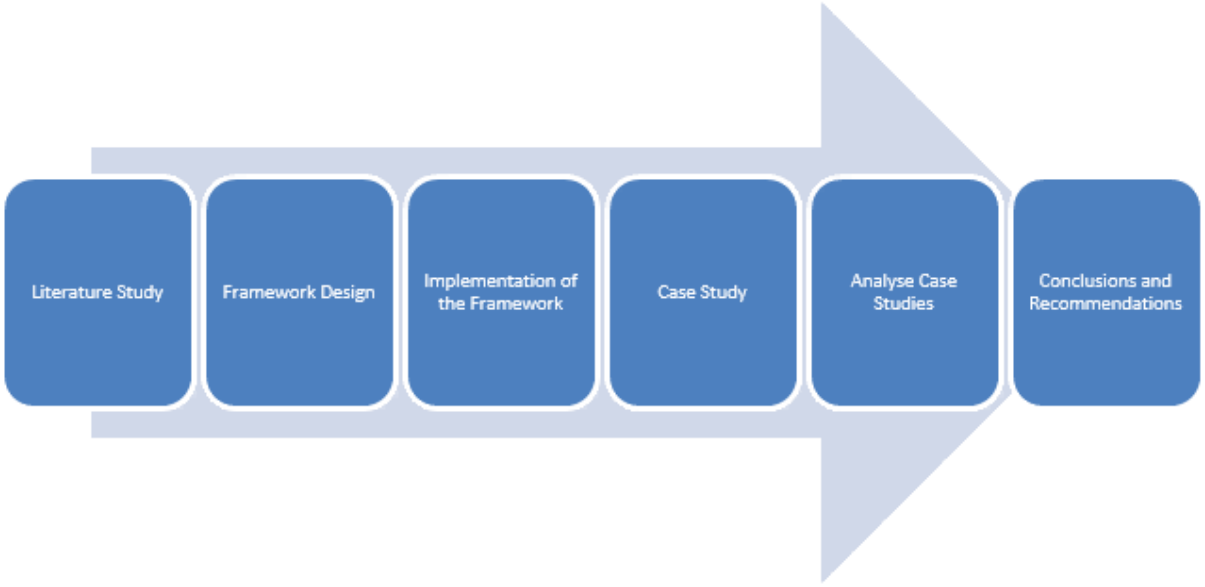

Figure 1.1: Strategy of the research study

1.6 THESIS OUTLINE

The thesis is structured in the following way;

- **Chapter 1: Introduction** - This chapter states the research question that aims to be answered. Further, the objective of the study, the rationale behind the study, as well as the research approach and outline are given.

- **Chapter 2: Literature Study** - In this chapter metal packaging manufacturing is discussed by;

  - Expanding on the uses of metal packaging.

  - Discussing the metals used for metal packaging manufacturing.

  - Describing the various manufacturing processes used in manufacturing of metal packaging.

  - A review of machine learning.

  The second segment in Chapter 2 focus on process improvement in manufacturing. Process improvement methodologies are discussed with much focus on the Six Sigma methodology for process improvement. Some reasons for the failures of process improvement methodologies are also expanded on. The third section in Chapter 2 relates to frameworks used in process improvement. Firstly various frameworks for process improvement that have been developed is discussed. Secondly, process improvement frameworks that use statistical and mathematical principles such as design of experiment and data science are described. Thirdly, the way decisions are made on how and when the steps of some of these frameworks are followed are investigated.

- **Chapter 3: Machine Learning** - In this chapter, the most widely used data science process improvement model; the CRISP-DM model is described. Further in this chapter machine learning is defined and discussed and the use of machine learning in manufacturing is touched upon. The rest of this chapter discusses the cleaning, exploration, understanding and preparation of data for machine learning modelling, as well as on different machine learning regression algorithms.

- **Chapter 4: Framework Design** - In this chapter a framework that incorporated Six Sigma and data science principles is designed. The main steps for this framework is outlined and explained.

- **Chapter 5: Case Study** - In this chapter a case study is described in detail that utilizes the framework that was developed in the previous chapter. The case study focuses on the process improvement in a 2-piece metal food can manufacturing line. The objective and rationale for the case study is described. The process specific to the case study is described in detail. The rest of the chapter focuses on the various steps related to the describing, understanding and preparation of the data that was used for the case study. The chapter also discusses the various regression algorithms that was used for the case study inclusive of linear, penalized, support vector machine and ensemble regression models. Finally the results from all the different ma-

chine learning methods are discussed and how the results could possibly be improved on is also discussed. The possible ways of how to deploy the final steps in the framework, that do not form part of the case study, is discussed.

- **Chapter 6: Conclusion** - In this chapter the final conclusions of the study is presented together with the identification of possible future work.

1.7 SUMMARY

This chapter gave the background to the development of a framework for process improvement that incorporated the use of data science for metal packaging manufacturing. The research problem was presented as a research question. The objectives were listed to how the research problem could be solved. The rationale of the research was given for the development and demonstration of a process improvement framework and how it could be demonstrated in the proposed case study. The research approach was given to how the stated objectives could be met. Finally the thesis outline was summarized.

CHAPTER 2

LITERATURE STUDY

2.1 INTRODUCTION

Chapter 1 gives the research question that this thesis aims to answer. As part of the research approach a literature study is conducted. The first chapter of the literature study focuses on the first part of the research question and the second chapter of the literature study focuses on machine learning. This chapter presents a general review on metal packaging manufacturing, followed by a review on methodologies used in process improvement and framework development.

2.2 METAL PACKAGING MANUFACTURING

2.2.1 Introduction

According to the McKinsey Global Institute (Manyika et al. 2012) the worldwide value of consumer packaging in 2010 was about $395 billion. 51% of this market was for food packaging and 18% was for beverage packaging. The metal cans market was about $60 billion worldwide of which the food metal can market contained 60% of this market and the beverage metal can market contained 40% of this market. By 2012 a total of 26 789 million food cans were manufactured and 92 239 million beverage cans were manufactured worldwide. In their 2019 report the Institute of Packaging SA (Packaging SA 2019) reports that in 2011, 2.5 million tons of packaging was consumed in South Africa. The turnover of the packaging and paper industries was more than R50 billion and it employed 90 000 people directly.

2.2.2 Use of packaging

Robertson (2013) lists the four main functions of packaging as containment, protection, convenience and communication. Packaging must be designed to be physically strong enough against everyday handling. It must be able to withstand damage due to environmental effects and micro and macro organisms. It must also be designed in such a way that it interacts adequately with humans, taking their capabilities and general regulations into account. Packaging allows people to easily consume fresh and uncontaminated products. If it was not for packaging most of the foods we consume today would not have been so easy to come by, since we would have been limited to locally grown and reared food that we would have had to acquire almost daily (Packaging SA 2019).

About 70% of packaging is used as containers for food and beverages. Almost all packaging are paper-, plastic-, metal- or glass-based. Metal packaging probably dates back to the time of Napoleon, but only became mainstream during developments in heat sterilization of foods during canning operations. Aluminium cans were first introduced in the 1950's and are mainly used for beverage cans and to a lesser extent for aerosols (Emblem 2012).

2.2.3 Metal used in packaging

Metal packaging are made of steel, tin (used in tinplate), chromium (used in electro-coated chrome coated steel (ECCS)) and aluminium. Metals are advantageous to use for manufacturing of packaging because

- of its strength, malleability and availability.
- of its low toxicity.
- of its superior barrier properties to gas, moisture and light.
- it can be easily coated and decorated.
- it can be manufactured on high speed lines. (Packaging SA 2019).

Tinplate is a low carbon, mild steel sheet with electroplated tin on both sides of the plate. A tin-iron alloy forms between the iron and tin layers. On the two outside surfaces of the tin there are a thin passivation film and oil film, see **Figure 2.1**.

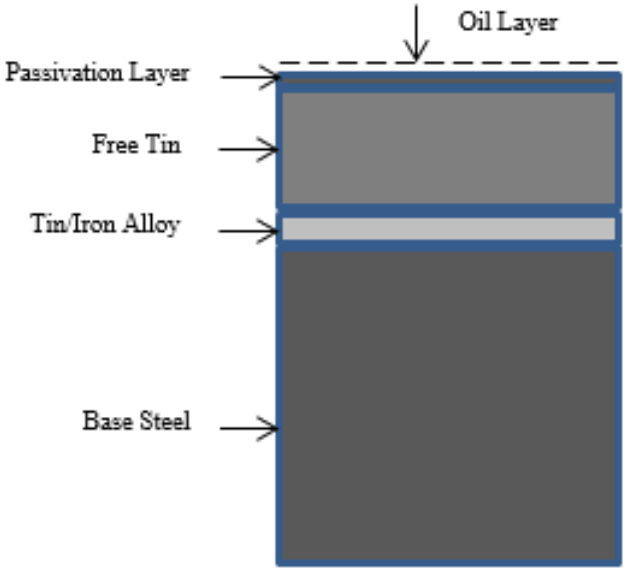

Figure 2.1: Structure of Tinplate

Various criteria can be considered when choosing tinplate: Steel base chemistry, type of annealing, whether it is single or double reduced, thickness and dimensions, tin coating mass, temper and hardness, surface finish, surface treatment and oiling (Marsal 1988).

The steel base gives strength to the tinplate and can be either defined as L-type or MR-type steel. MR-type steel base has fewer limits on metal elements and can be used with less corrosive products (Marsal 1988). L-type steel is similar to MR-type steel, but has lower copper and phosphorous levels. L-type steel base has a low content of metal elements and is preferred to be used with highly corrosive products. For fruit canning an L-type steel base will be used to make tinplate for cans because a high internal corrosion resistance is required. When steel is needed to undergo severe drawing operations, D-type steel can be used. D-type steel has less carbon than the other types used for tinplate. (Robertson 2013).

After the first few steps of the steelmaking process the result is typically a cold rolled sheet of steel with a thickness of less than 0.2mm. Cold rolling is followed by the annealing process to relieve the stresses that had built-up in the steel. Annealing typically takes place at temperatures of between 600°C and 700°C and can be a batch process or a continuous process. Continuous annealing gives strong steel with a fine grain structure and batch annealing gives steel with a coarse grain structure with excellent formability. After annealing, steel can be single reduced (SR) or double reduced (DR). If steel is temper rolled after annealing, it underwent only one cold rolling step and therefore is known as SR steel. If steel is cold rolled after annealing, it underwent two cold rolling steps and is therefore known as DR steel. SR plate has better formability than DR plate, but DR plate is stronger and harder than SR plate and can be used at comparatively thinner gauges (Robertson 2013).

The tin layer is the sacrificial anode in a tin-iron cell that forms when the tinplate comes in contact with the product under de-aerated conditions. Tin thickness is therefore a determining factor for the shelf-life of canned products. Tinplate can have a range of tin coating weights and can have identical coating weights on both sides of the tinplate, known as E tinplate, or tinplate can have differential coating weights on the two sides of the tinplate, known as D tinplate see **Table 2.1** (Marsal 1988).

Table 2.1: Tin Coating Weights for Electrolytic Tinplate

Tinplate	wt/wt on each face of tinplate
E1	$2.8 g.m^{-2}/2.8 g.m^{-2}$
E2	$5.6 g.m^{-2}/5.6 g.m^{-2}$
E3	$8.4 g.m^{-2}/8.4 g.m^{-2}$
E4	$11.2 g.m^{-2}/11.2 g.m^{-2}$
D2/1	$5.6 g.m^{-2}/2.8 g.m^{-2}$
D3/1	$8.4 g.m^{-2}/2.8 g.m^{-2}$
D4/1	$11.2 g.m^{-2}/2.8 g.m^{-2}$
D4/2	$11.2 g.m^{-2}/5.6 g.m^{-2}$

Aluminium is also used widely in packaging. Metal closures, foil packaging and beverage cans and

ends are just a few common examples of aluminium packaging. According to a manufacturer of aluminium; (Hulamin 2021), there are many characteristics of aluminium that makes it suitable for the use of food packaging such as non-toxicity, no odour, stability in a wide range of temperatures and good conductivity of heat. Aluminium, present in bauxite ore, is a very common element on earth. Aluminium oxide is extracted from the ore before it is melted in a high energy consuming smelting process. The melted aluminium is then filtered through ceramic filters before it is casted in ingots. The surfaces of the ingots are scalped to present a smooth surface for the upcoming rolling steps. Rolling takes place in two stages; firstly a hot rolled stage to intermediate thickness, and secondly a cold rolling stage to a final hardness and thickness (Emblem 2012).

According to RecyclingInternational (N.D.) South Africa's waste management industry was worth 2 billion rand in 2019. Steel packaging such as tinplate had a 72% recycling rate and aluminium packaging had a 75% recycling rate in South Africa in 2019. Metal packaging has higher recycling rates than any other form of packaging in South Africa. Metal cans are 100% recyclable and it can be recycled infinitely.

2.2.4 Tinplate metal can manufacture process

Food cans are made of tinplate. Tinplate is either delivered in coils or in stacks of flat sheets. Can bodies are made by slitting the tinplate in rectangular blanks. The blanks are bent into a cylindrical shape with the help of a mandrel. A side seam is formed by electrical resistance welding, resulting in a lap joint (Simal-Gándara 1999). A continuous copper wire electrode is used in welding to produce a 0.4mm to 0.8mm overlap. A good weld will have the same tensile strength as the base plate. The internal weld is enameled by a side-stripe to protect against interaction with the product. Final metal forming of the can consists of flanging and beading. A three-piece can will consist of a cylindrical tinplate body and two ends seamed on either side of the can's open ends (Robertson 2013). Refer to **Figure 2.2** for an illustrated process flow for the manufacturing of 3-piece tinplate cans.

A two-piece can is a seamless drawn can body with an end seamed onto the can's open end. Two-piece tinplate cans can be drawn and wall-ironed (DWI) or drawn and re-drawn (DRD).Two-piece DWI tinplate cans are formed by ironing in a press where the clearance between the tools is less than the thickness of the metal being sent through them. In this process the wall of the metal is thinned by up to 50%, generating a lot of new surface. In the case of tinplate, the surface of the tinplate is severely disrupted as a result of this wall ironing and therefore a coating is needed to prevent metal pick up (Simal-Gándara 1999). DWI cans start off as a flat disc which is formed into a cylindrical cup by a punch drawing it through a die. The cylindrical cup is then passed successively through a series of ironing dies causing the wall thickness to decrease and the body height to increase correspondingly. Refer to **Figure 2.3** for an illustrated process flow for the manufacturing of 2-piece tinplate cans.

Drawn and redrawn cans are cans that went through a multi stage drawing process. The DRD stages are similar to the first step in the DWI can making process. DRD cans start off as a flat disc which is formed into a cylindrical cup by a punch drawing it through a die.

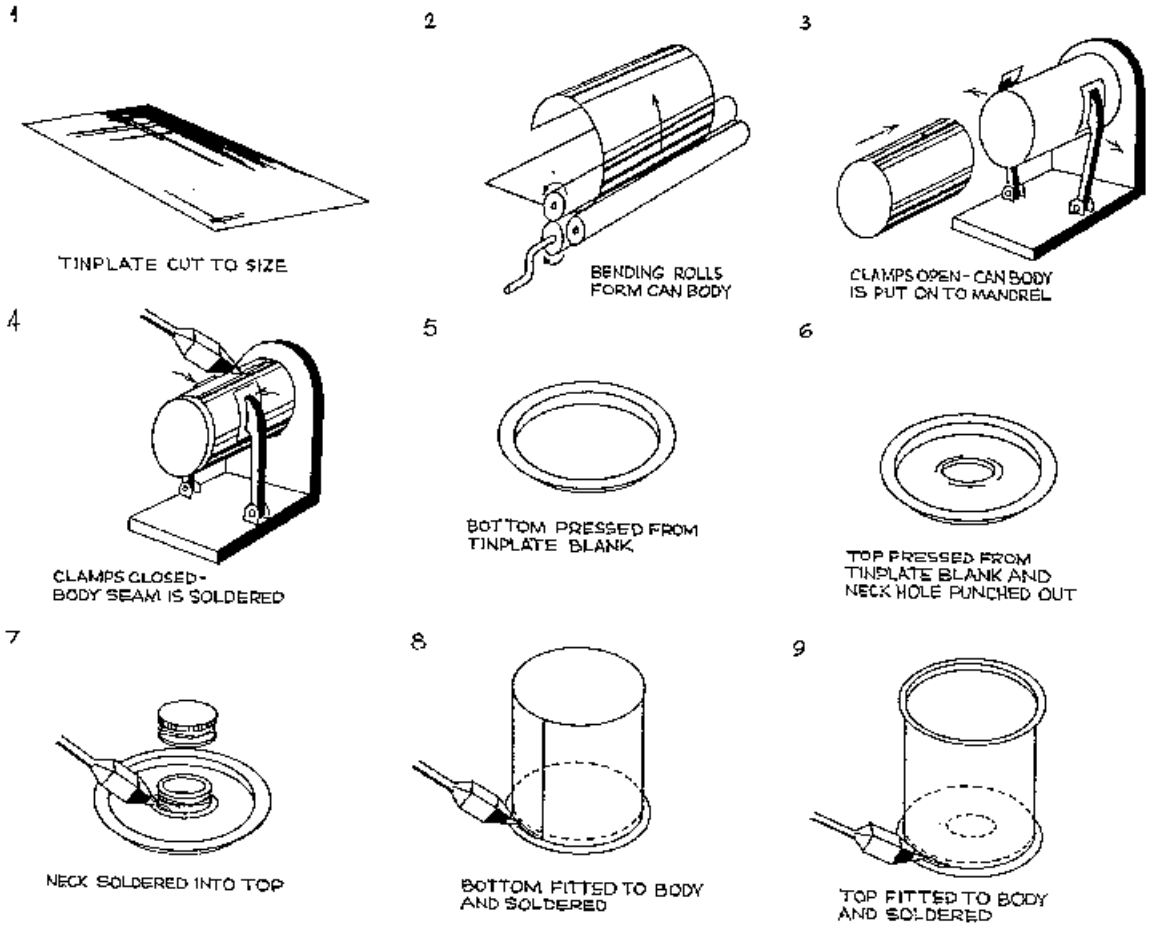

Figure 2.2: 3-piece food can manufacture process flow (Fellows, Axtell, et al. 1993)

Two-piece cans have better integrity since there is no side seam and only one double seam. The single double seam is also better formed and controlled because of the absence of the side seam. There are up to 35% savings in materials by using the DWI cans compared to three-piece cans. It can also be aesthetically more pleasing when a can is printed, since no seam allows for all round decoration (Robertson 2013).

The end of the formed cylinder is flanged to accept the end after filling (Simal-Gándara 1999). Tinplate ends are stamped on presses to a specific design for optimum deformation behaviour. After stamping the outside curl is formed and a sealing compound is applied in the seaming panel of the end. The sealing compound is a synthetic or natural rubber dispersed in a solvent or water. The sealing compound assists in the formation of a hermetic seal in the double seam of the can by providing a gasket between the tinplate layers. Ends can be sanitary or easy open ends (Robertson 2013).

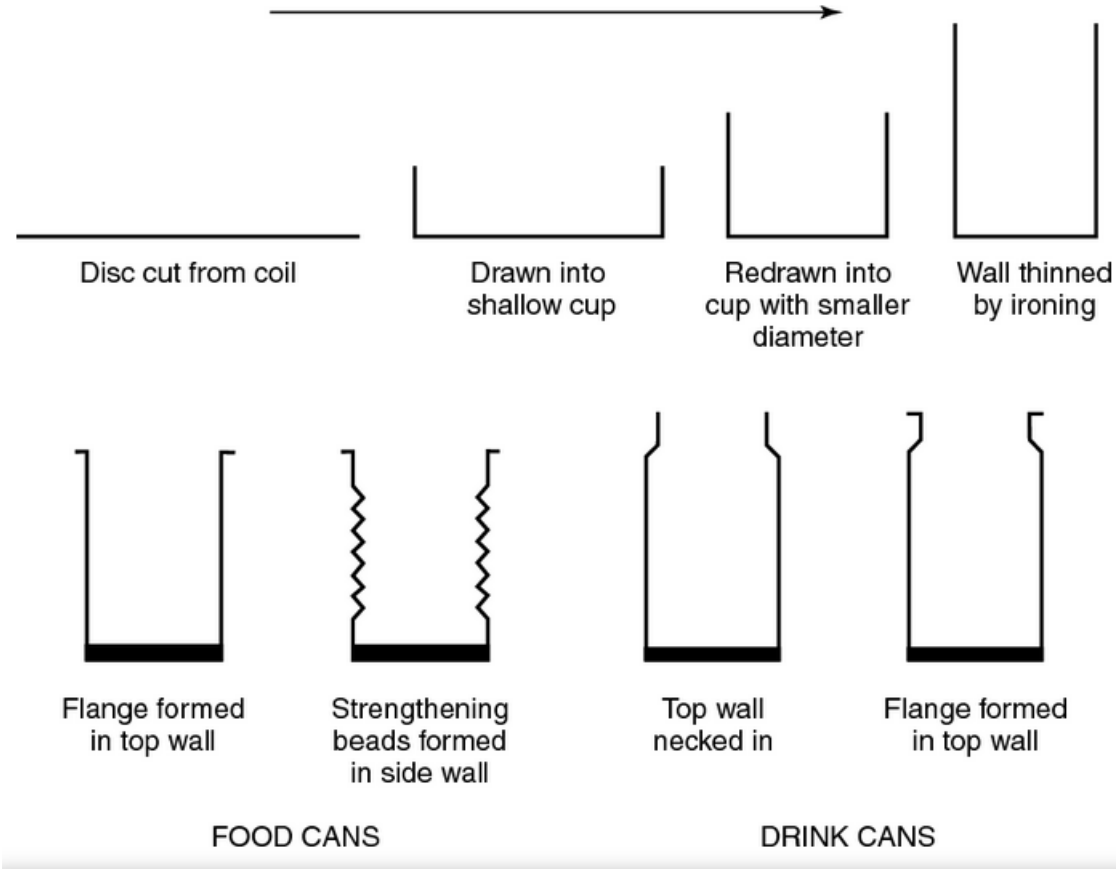

Figure 2.3: 2-piece food can manufacture process flow (Page, Edwards, and May 2006)

2.3 PROCESS IMPROVEMENT

2.3.1 Introduction

The goal of process improvement is to minimize error and waste in your process and to maximize productivity and efficiency. The three broad steps in process improvement methodology are

- to identify the problem,
- to apply methods to overcome the problem, and
- to analyze the outcome (White 2019).

According to Mauri, Garetti, and Gandelli (2010) process improvement can take the form of continuous improvement, or it can be the monitoring of performance parameters to detect the overall equipment effectiveness (OEE). For a process improvement strategy to be successful you have to be able to measure your performance and you have to be able to determine which factors in your process

is critical.

2.3.2 Process improvement methodology

There are various process improvement methodologies. These methodologies normally consist of a number of steps that form part of a general framework. Some of the better known process improvement methodologies are described below;

- Six Sigma is used widely in business and manufacturing. It aims to measure and subsequently eliminate inconsistencies and defects in a process. The generalized framework of Six Sigma follows the DMAIC steps, which stands for define, measure, analyze, improve and control (Shankar 2009).

- Lean manufacturing looks at value streams in a process and aim to identify which steps in a process add value to the final product and which steps in a process does not add value to the final product (Panizzolo et al. 2012).

- Lean Six Sigma, as the name suggests, is a hybrid or combined framework encompassing steps of both the Six Sigma end Lean methodologies. The Lean Six Sigma methodology aims to eliminate defects and waste from a process. According to Snee (2010b), Lean Six Sigma is used for the initial process improvement steps and also to guide the continuous improvement (CI) process.

- Total Quality Management (TQM) holds forth that the principles of quality are entrenched and applied at every level and in every department of an organization (Hackman and Wageman 1995). This methodology follows a systematic approach to achieve goals. The goals are determined by customers' needs. The TQM methodology continually looks for ways to be more effective and competitive and aim that everyone in an organization knows what these goals are and how to work towards achieving them. TQM follows the PDCA (Plan, Do, Check, Act) framework (Isniah, Purba, Debora, et al. 2020).

- Just-in-time works on the premise that a manufacturing concern should only produce what is needed and decrease the stock inventory. According to Cheng and Podolsky (1996), the JIT philosophy is to have only what is needed, in the quantity that it is needed in, exactly when it is needed.

- Theory of Constraints (TOC) is a methodology that aims to identify the biggest constraint on a system, and then to systematically decrease the effect of this constraint until it does not negatively affect the outcome anymore. According to Gupta and Boyd (2008) TOC looks at a system, not as individual sub-sections, but as a chain of interlinked and interdependent sections.

2.3.3 Lean six sigma

Many of the process improvement frameworks can be incorporated into the Six Sigma fold, except maybe for the Lean manufacturing philosophy. The combined Lean Six Sigma approach is a process improvement philosophy that covers a wide range of methodologies and tools. The approach and specific tools that could be used by different manufacturing processes can be different depending on various factors.

According to Mileham (2007), Six Sigma has defect rate as its focus, whereas Lean has process quality as its focus. Both these approaches are valid since a decrease in the number of defects of a process will lead to a better quality process. Most of the problems that are related to the product or the process can be traced back to one or more of four things;

- Design
- Process
- Materials
- Operations

To solve problems related to these four points in a manufacturing company there should be adequate support, resources and capabilities.

The DMAIC framework used in Lean six Sigma is a general framework. With each step of this framework there are numerous approaches and tools that can be used. Not every tool or approach will necessarily be useful for every improvement project. The steps in the DMAIC framework will be discussed broadly in the next few paragraphs.

2.3.3.1 Define

The define step can focuses on

- the problem that needs to be fixed.
- the process that needs to be improved.
- the product that needs to be developed.
- the raw materials that is the most suitable.
- the operation that is the most suitable.

Process improvement projects, that can show some form of direct or indirect financial benefit as driving force behind them, are normally prime candidates to get off the ground in manufacturing environments. Some of the areas where a process improvement project can be initiated, are where (Mileham 2007)

- the defect rates are high.
- the rework rates are high.
- the labour costs are excessive.
- the time constraints are too many.
- the inventories are too high.
- the down times are too high.
- the audit results point to deficiencies or shortcomings.

Various tools can be used in the define stage of the Lean Six Sigma process improvement, a few examples are listed below;

- Affinity diagrams is a form of brainstorming where related ideas are grouped together. Affinity diagrams help to organize large groups of data from surveys or brainstorming sessions into managable pockets in order to find trends (Plain 2007).
- Failure Mode, Effects and Criticality Analysis (FMECA) looks at what the causes of failures are, the impact these failures have on the process or the business, and the possible causes of these failures. Borgovini, Pemberton, and Rossi (1993) describes FMECA as a technique which tries to identify potential failure modes within a process. The effect of such a failure is then rated based on the effect it potentially can have on the process and the safety of the personnel involved.
- Process flow charts are visual representations of the sequential steps and decision points in a process. The basic general system schematic that is used in modelling of a manufacturing process is given by Cameron and Hangos (2001). The model of the process illustrates the processes in a system. The model can represent a single step or unit in a process, or a section in the process or the whole process, see **Figure 2.4**.

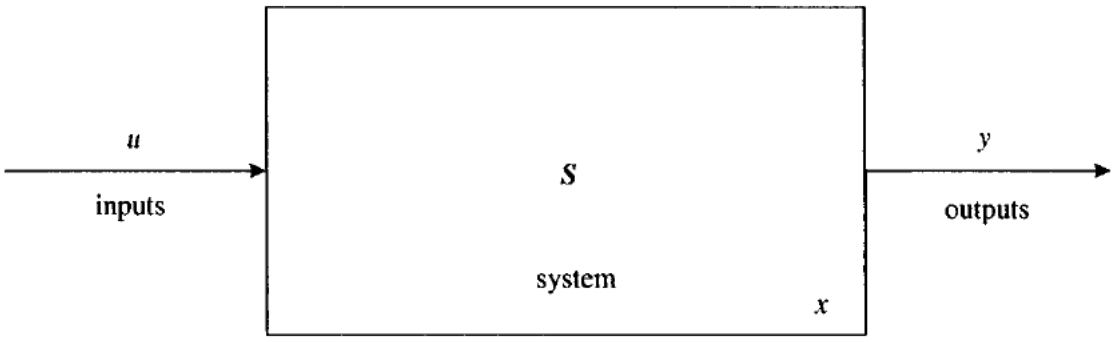

Figure 2.4: General system schematic (Cameron and Hangos 2001)

- Supplier, Inputs, Processes, Outputs and Customer (SIPOC) frameworks are business process maps that identify all the relevant elements of a process improvement project. SIPOC states that any organization is constructed of the 5 elements as captured in the acronym. These 5 departments are all interrelated, and by analyzing each section the whole system can be understood (Cao et al. 2015).

- Project charters are overall plans for the process improvement projects, they stipulate time lines and responsibilities. According to Gijo and Scaria (2014) the project charter shows the ownership and responsibilities of all the stakeholders in the project. The project charter helps everyone involved to understand the process to follow and their necessary involvement.

2.3.3.2 Measure

Keller (2011) states that the objective of the measuring stage of Six Sigma is to measure all the relevant facets of the current process. Measuring should assist to understand the process at a detailed level. In the define stage, a high-level process map is compiled, whereas in the measure stage a much more detailed process map is drawn of the section on which the improvement effort is concentrated. Compiling this detailed process map needs the input of various process personnel that might have good process knowledge and will be able to add valuable insight into the process.

The measurement stage should also define the metrics of the system to be able to do estimations of the process. To be able to have access to data to measure the metrics of a process, there need to be a reliable and accurate measuring system in place. Typical examples of metrics are

- those that are critical-to-quality (CTQ) such as
  - throughput yield,
  - specifications, and

  - process standard deviation.

- those that are critical-to-cost (CTC) related to costs or return on investment (ROI) such as
  - rework figures,
  - scrap figures,
  - stockpiling figures,
  - approval times, and
  - lost orders.
- those that are critical-to-schedule (CTS) such as
  - cycle times,
  - processing times,
  - delivery times,
  - queue times, and
  - downtime.

During the measurement stage, baseline estimations of the process are done to use as a starting point of the project. It is necessary to estimate a baseline of the responses you want to improve on, in a process. Normal curves are used to show when data points are outliers. Statistical process control (SPC) charts can be used to measure the metrics in a system. SPC data should not be used as a proposal to do a process improvement project when the SPC data indicates a special cause for data being out of control. The special cause can be identified and acted upon rather. When a process is in control according to SPC then the process capability index can be used as a metric, when a process is out of control according to SPC, or there is no SPC data available, a process performance index can be used as a metric.

During the measurement stage it is also important to analyze the measurement system to be able to quantify errors associated with measurement itself. The measurement systems need to be measured for accuracy and linearity. Check the repeatability of the measuring system by comparing multiple measurements of the same sample unit. Check the reproducibility by comparing the measurements of the same sample unit by different personnel.

Various tools can be used in the measure stage of Lean Six Sigma process improvement, a few

examples are;

- Process maps which are flow diagrams that graphically show a process with all the inputs, outputs and actions as described in **Section 2.3.3.1**.

- SPC is a form of quality control that uses statistical methods to monitor a process, and to employ the analysis to keep the process in control. SPC measures quality in order to be able to control the quality and ultimately improve on the business success of the manufacturing industry (Oakland 2007).

- Visualization can be achieved by graphs and plots such as histograms, box plots, dot plots, Pareto charts, running charts and pie charts to illustrate some aspects related to the data of a process. Data visualization is the representing of data in a graphical form, that aids in gaining information from the data and understanding the structure of the data (Chen, Härdle, and Unwin 2007)

- Statistics such as median values, mean values, standard deviations, process capability index, process performance index, goodness of fit and confidence intervals are all simple techniques that can be employed to measure a process.

- Repeatability and reproducibility (R&R) studies are techniques that estimates the repeatability and the reproducibility of measurement systems in a process. Repeatability measurements is the analyzing of the variability in measurements of gauges, and reproducibility is the analyzing of the variability in measurements of the operators. R&R studies is important because the ultimate goal for process control is to minimize the variation in a process (Durivage 2015).

2.3.3.3 Analyse

According to Shaffie and Shahbazi (2012) the analyse phase for process improvement only starts once the process has been mapped, all necessary data has been collected and presented and the project has been approved by the relevant stakeholders.

Pyzdek and Keller (2018) list a few objectives of the analyze phase in Lean Six Sigma;

- Identify ways in a process to eliminate the gap between the wanted performance versus the current performance.

- Determine the source of variation that contribute to this difference in performance of the process.

- Determine the drivers that significantly influence the process and those requirements that are CTQ, CTC and CTS.

- Benchmarking.

Various tools can be used in the analyse stage of Lean Six Sigma process improvement, a few examples are listed below;

- Brainstorming is a method used in a group setting to generate ideas, increase efficiency or find solutions (Wilson 2013). Brainstorming is a non-mathematical tool and can be used in various ways to analyze outcomes. Examples of different ways in which brainstorming can be applied are 5-why's, fishbone and cause and effect analysis.

- Value stream analysis is a tool that comes from the Lean methodology where a process is analyzed to determine which steps in a process add value to the process and which steps do not add value to the process. According to Parab and Shirodkar (2019) value stream analysis is an important tool to use when the aim is to decrease lead times or decrease work in progress in a process.

- Design of experiment (DoE) is used when simple comparative analysis between two factors is limited since many factors impact a process. A fundamental approach of utilization of DoE in process improvement in manufacturing systems can consists of three phases (Montgomery 1999).

  - Characterization is the process to discover the specific process variables that are responsible for the variability in the system's output responses. This is a classical approach also known as the Fischer approach. Full factorial and fractional factorial designs are used in characterizing systems. A full factorial DoE is best when you want to test the main effects and interactions between 2 to 4 factors. Normally this approach will be used when there are enough time and resources. Screening experiments are used to determine important factors. A fractional factorial design is best for screening of critical factors, when there are more than 6 factors. This will normally be used when there are unknown factors and you want to identify the effect of the most important factors.

  - Control is the process to obtain a consistent performance from the system. This phase uses orthogonal arrays and is known as the Taguchi method.

  - Optimization is the process of manipulating process variables to levels that will result in the best obtainable set of operating conditions for the system. Response surface methodology (RSM) is a useful method to model and optimize a system. RSM is used when you want to optimize 2 to 4 critical factors which have defined ranges.

- Data Science has various methods and principles that can be applied during the analyse phase of a continuous PI project. Buer, Fragapane, and Strandhagen (2018) proposes a framework for process improvement by data driven methodology. The 5 broad steps in the framework are

collecting, sharing, analysis, optimization and feedback.

2.3.3.4 Improve

According to Shaffie and Shahbazi (2012), at the start of the improvement phase there should be a clear list of process steps that need to be improved. In the improve phase the focus can be on improving the flow and efficiency of the process or to statistically improve measurable responses in your process that were analyzed in the previous step. Pyzdek and Keller (2018) says that the primary objective of the improve phase is to implement the new system. Typical broad steps can be to

- choose the process improvement strategy,
- optimize and define the improved process settings, and
- analyze the effect of the process changes to control and possibly make further improvements.

Various tools can be used in the improve stage of Lean Six Sigma process improvement, a few examples are:

- Tools that identify and eliminate unnecessary steps such as
  - 5S (sort, set in order, shine, standardize, sustain) to prevent excessive movement of materials and people,
  - TOC to identify and eliminate bottlenecks,
  - standardization to eliminate process errors,
  - pull systems which eliminate excess inventory and,
  - queuing theory to smooth out the processing systems and flow.
- Adams et al. (1999) describes how simulation can be used as a tool throughout a continuous improvement process. A way that simulation can be used when there is a detailed model of a system, is to simulate changes to this model and track how these changes may affect the model's outcomes.
- Tools used for risk analysis such as Fault-tree analysis (FTA) and FMECA. Risk analysis has three main steps according to Aven (2015) which are planning, assessment and treatment.

2.3.3.5 Control

According to Shaffie and Shahbazi (2012), in the control phase the emphasis is to ensure that the objective has been achieved after the improvement phase and that there are measures in place to sustain the positive change. Buy-in from all involved is critical to ensure that the improvements will stay in place in future operations. Pyzdek and Keller (2018) states that the main objectives of the control stage is to statistically validate that the improvement phase has reached the objective. Once verification has been performed, a control plan needs to be formalized and documented to sustain the positive changes of the improvement phase.

Various tools can be used in the control stage of Lean Six Sigma process improvement, a few examples are;

- Validation by monitoring the SPC data, monitoring the CTQ factors and running pilot operations with the new process to ensure everything works as predicted.
- Business process control planning is to change anything that needs to be changed on documents and procedures due to the outcomes of the PI. The relevant team members need to ensure necessary changes in
  - policies,
  - standards,
  - procedures,
  - audits,
  - price modules,
  - change engineering,
  - production planning,
  - manpower needs,
  - training needs, and
  - information systems.
- Control charts to monitor the process gains to ensure that these gains are maintained.

2.3.4 Main reasons why process improvement projects fail

According to McLean, Antony, and Dahlgaard (2017) most of Lean Six Sigma continuous improvements projects fail to effect the change envisioned in manufacturing. Failures have been identified to fall under 8 central themes:

- Motives and expectations – The motive should be to address a need in the company, and not to do something that is done by other companies. Expectations can be damped by negativity or people that form part of the process that do not understand everything.

- Organizational culture and environment – Sometimes the existing culture in a company is resistant to change. Sometimes the lack of budget, support and weak systems can hamper successful change implementations.

- Management and leadership – You need capable people to run with the changes, but at the same time everything cannot solely depend only on one person.

- Implementation approach – This is the main reason for improvement projects not to be successful. Sometimes the projects are just ceremonial or partial. Sometimes a quick fix is looked for and not a systematic long-term improvement. The road-maps in organizations are either not sufficient or not followed properly. The improvement strategy should be standard across the organization and not controlled by a single person or a small group.

- Training - Can be inadequate or not capable of delivering the necessary skills. Training should not be too generic with no practicality to it. Training should go hand-in-hand with actual changes that will occur in the process improvement approach.

- Project management – Projects can fail if it does not focus on where the organization can benefit most. People can fail if the scope is too large, if the wrong people are involved, if there are time constraints or if there is not enough commitment. Projects can fail if the team fail to analyze the interaction between processes, or if the data used in the project is not good enough.

- Employee involvement levels – It is difficult to implement comprehensive changes in large organizations because of the resistance to change.

- Feedback and results – There should be mechanisms in place to review the process improvement initiatives. Sometimes failures are not reported at all or data are skewed to make results seam more favourable. Implementations of process improvement findings sometimes do not bring immediate financial benefit but sometimes first need capital input before reaping returns later over time.

Companies implement Lean Six Sigma to improve facets of their processes that relates to efficiency, quality and costs. If the strategy or framework that was used is poor or was not implemented properly

there is a good likelihood that the company will not truly benefit from Lean Six Sigma (Albliwi et al. 2014). Albliwi et al. (2014) lists 34 factors that were identified from literature as failures in the Lean Six Sigma approach. The top factors identified were;

- Lack of top management commitment and involvement
- Lack of training
- Selecting improper projects
- Lack of financial and human resources
- Resistance within the company as a result of the existing culture
- Poor communication

Albliwi et al. (2014) further dissects failures specific to the manufacturing sector, where there were found to be certain factors that were more prevalent that could cause failures.

- Poor data or unavailability of data causes difficulty in analysis.
- The improvement projects are not always aligned with a company strategy or sometimes even the strategies were improper e.g. when a company built their strategy on another similar company's strategy there are increased risk of failures.
- There is a lack of understanding from management of how to implement process improvement projects and the various tools and techniques that can be used in Lean Six Sigma. There is also a lack of understanding of how to start a project, what steps to follow and how to implement findings of such a process improvement project using Lean Six Sigma.

2.4 FRAMEWORKS FOR PROCESS IMPROVEMENT

2.4.1 Frameworks for process improvement in research

Aqlan and Al-Fandi (2018) states that process improvement in many fields are successfully implemented by Six Sigma and Lean principles. Lean philosophy is centered on the minimizing of any waste. Tools used in Lean manufacturing can be Kanban, just-in-time (JIT), standard work and 5S. Standard work is defined as the performing of a process that will deliver the safest, best quality and most efficient result. Six Sigma aims to identify the cause of defects and then to eliminate the defects by the DMAIC process. A disadvantage of Lean is that the methodology does not consider advanced statistical tools required to achieve process capabilities. A disadvantage of Six Sigma is that the methodology does not assist in helping to improve the process flow.

J. Singh, H. Singh, and Pandher (2017) use the DMAIC steps of Six Sigma as the 5 steps of a framework to improve the process of a manufacturing business by decreasing their defect rates. They describe the framework with a case study.

- In the define phase it is determined where the problems are and which problems are the most important to solve. In the case study J. Singh, H. Singh, and Pandher (2017) describe, the total production figures of the various products have been gathered for a specific time period as well the rejection rate.

- In the measure phase it is determined how the process is measured and how it is performing. In the case study the product, with the highest defect rate has been selected and the prevalence of specific defects was calculated.

- In the analyze phase the most common cause of failure are determined. In the case study the causes of the defects were discussed in brainstorming sessions with root cause analysis.

- In the improve phase it is determined how to remove the cause of defects from the process. In the case study the findings from brainstorming have been implemented.

- In the control phase it is determined how to maintain the improvements in the process. In the case study the improvements in the process is validated by measuring defect rate and comparing the before and after values.

J. Singh, H. Singh, and Pandher (2017) illustrate how the steps of Six Sigma can be easily implemented to solve simplistic problems in manufacturing. The case study discussed in J. Singh, H. Singh, and Pandher (2017) only looks at three products over a five week period and determine the product with the highest number of defects, and then rate the defects also from most to least prevalent. Root cause analysis is done and the findings from it are implemented. The implemented actions have shown some improvements in the follow-up data analysis.

Snee (2010a) highlights some critical considerations when monitoring process performance. Manufacturing processes that are capable and stable over time are processes that will consistently produce material that are within specifications. To be in control, a system must produce products that vary within the process control limits, where the mean and 3 times the standard deviation are the variance measured against. A capable process is a process where the the process variance falls within the specifications. Process stability and capability should be tested for each batch, but they also cover longer periods such as monthly or yearly. The difference between control limits and specification limits is that control limits are determined by the process while the specification limits are determined by the product. Steps to follow to monitor process performance and product quality are:

- data is collected from the process,

- data is analyzed,
- from the analysis, process adjustments are made when process goes out of target,
- records are kept of causes for the process to go off-target, and
- process improvements are completed.

Tools used to assess the data are;

- Control chart analysis – since a stable process will only vary between set limits, any special cause variation is caused by something that is out of the ordinary. Such a special cause should be investigated and rectified.
- Process capability analysis.
- Do variance analysis by doing Analysis of variance (ANOVA).

The majority of the steps Snee (2010a) suggest for monitoring process improvement describes the analyze phase of Lean Six Sigma. Once data is collected from the process (the measure phase of Lean Six Sigma), the next three steps concentrate on the analysis of the data. Step 5 is a very broad step which suggests that analyzed data must then be used as basis to complete process improvement.

Mauri, Garetti, and Gandelli (2010) write that process improvement can take the form of continuous improvement or it can be the monitoring of performance parameters to detect the OEE. Mauri, Garetti, and Gandelli (2010) create a generic structure to assess and improve a production system in which a special parameter called the operating system effectiveness (OSE) is introduced. The process improvement methodology consists mainly of three phases that encompass 13 steps. The three phases are:

- Phase 1 - Measuring the current level of performance
- Phase 2 - Use failure mode, effects and criticality analysis (FMECA) to identify what causes inefficiency
- Phase 3 - Decide which tools to use to remove the causes of inefficiency

The steps Mauri, Garetti, and Gandelli (2010) used in the framework for process improvement are summarized in **Figure 2.5**.

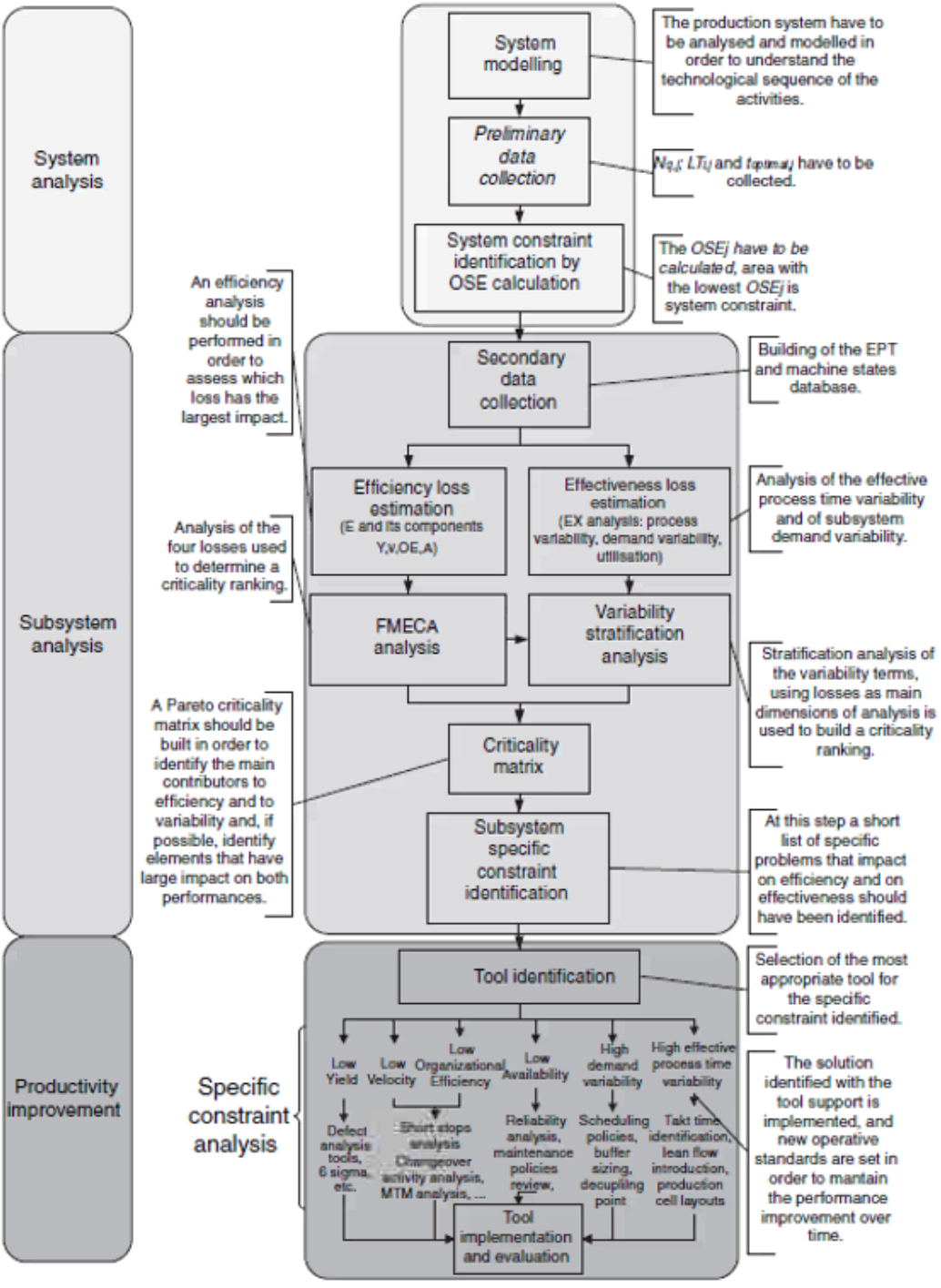

Figure 2.5: Method for process improvement in manufacturing systems (Mauri, Garetti, and Gandelli 2010)

Mauri, Garetti, and Gandelli (2010)'s approach to process improvement is to measure process effi-

ciency in terms of time studies for each subsystem in the overall process. The problematic subsystems are identified using FMECA and remedied by using various tools. This method can be effective to identify where the inefficiencies and ineffectiveness are in a process, but will not be able to identify specific factors and different factor interactions influence on a process. The method uses tools to improve subsystems that are reliant on procedural and management changes such as Lean, Six Sigma, policy reviews, standard operating procedures, time-based activities and plant layouts.

According to Prashar (2016) DoE is a powerful statistical tool to investigate hidden causes for variation in complex industrial systems. Traditional DoE methods are full factorial and fractional factorial designs. DoE has a lot of potential for continuous process improvement but it lacks an implementation framework that stems from specific requirements. This is where Six Sigma's DMAIC is effective. Taguchi introduced the orthogonal arrays (OA), linear graphs and signals to noise (N/S) ratio to reduce variation in the process. The disadvantages of the Taguchi method are that it is focused on optimal process settings but does not scan for critical to quality characteristics. The Shainin system (SS) was developed by Shainin to use observational investigations before experimentation. This approach relies heavily on engineering judgement. Industry is adopting a number of hybrid approaches to process improvements to overcome inherent weaknesses in processes; Prashar (2016) developed an integrated framework for process improvement using Taguchi methods, Shainin System and Six Sigma, see **Figure 2.6**.

Prashar (2016) shows how the effectiveness of the Six Sigma process improvement can be improved by adding aspects of other process improvement tools into a hybrid framework. Prashar also shows the effectiveness of using statistical methods used in OA and DoE in process improvement.

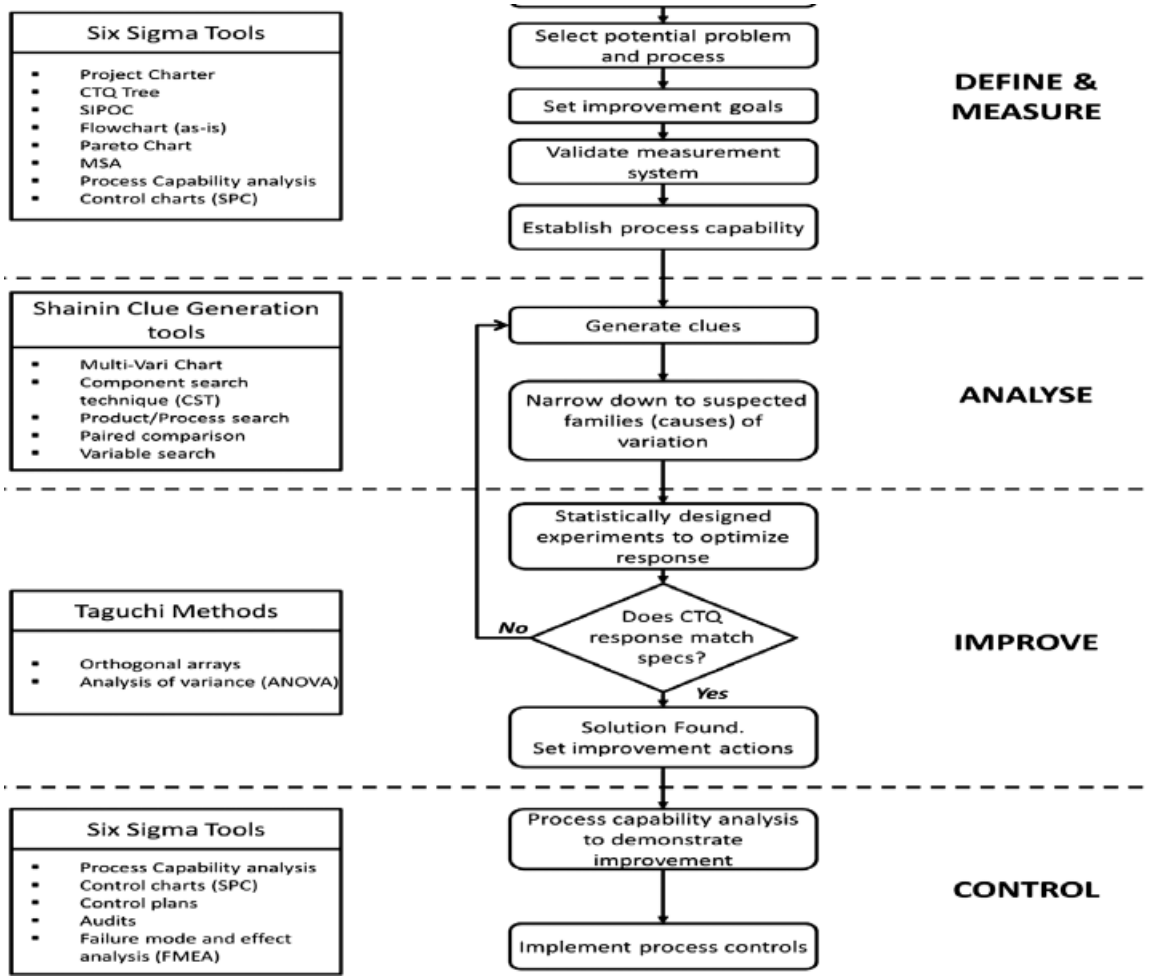

Figure 2.6: Six Sigma / DoE hybrid framework (Prashar 2016)

2.4.2 DoE used in process improvement frameworks

According to Tanco et al. (2007) DoE forms an integral part of the Six Sigma philosophy. Six Sigma uses the DMAIC methodology and experiments will fall under the analysis and improve phases. There are three types of experiments; best guess, one factor at a time (OFAT) and DoE. DoE is the best way to conduct experimentation where there are multiple variables present. The most companies and engineers followed the OFAT approach for experimentation. There is a need for a methodology for companies to use statistical methods to carry out their experiments. DoE has the following advantages over OFAT:

- It requires fewer resources (experiments, time, material, etc.) for the amount of information obtained.
- The estimates of the effect of each factor (variable) on the response are more precise.

- The interactions between factors can be estimated systematically (Interactions are not estimable with OFAT).

- There is experimental information in a larger region of the factor space.

Research done in the UK has shown a gap in knowledge required for applying DoE to solve quality problems (Antony 2001). Reasons for this are:

- A lack of understanding of how to use DoE to solve product and process problems.

- Firefighting, caused by reactive and not proactive process improvement.

- Lack of skills.

- Lack of communication between different departments and different people in the organization.

- Commercial software is not very user friendly to apply DoE.

- A lot of academic literature is all about the statistics and the mathematics and not really about practical use of DoE in manufacturing.

DoE is used for (Al-Ghamdi 2011):

- Screening experiments - To identify the most influential control factors. This approach will be used when there are many factors and the effect of them on the process is not understood well.

- Characterizing - To identify how the important controllable factors should be adjusted to get desirable responses.

- Optimizing - To determine the best set of operating conditions by manipulating the most influential process parameters.

- Dealing with the complexity of interactions in a process - Complexity arises from interconnection in processes and variations in processes. $Y_i = f(X_1, X_2 \cdots X_n) + e$, where Y is a response that is a function of various factors X_i, and e are the noise factors.

- Dealing with variances in a process - Variances are everywhere even though most people would not recognize them or at least understand the impact variance has on a process. DoE recognizes the influential X's to minimize the variance of Y.

- Modelling – DoE can be conducted on an actual process or on a model of the process.

- Improving the process of formal decision making – DoE can be used to formulate alternative

courses of action. DoE can be effectively used in any process where there is some form of repetitive process, which is the case for manufacturing.

Antony (2001) describes a case study that looks at short life of core tubes when subjected to hydraulic fatigue testing. The following main steps were followed in the process improvement framework using the DoE methodology.

- Step 1 – Identify the problem.
- Step 2 – Identify the control factors that affect the outcome and variability of the response by using brainstorming and cause and effect analysis.
- Step 3 – Determine the optimal levels the factors should be set at to maximize the desired response.
- Step 4 – Identify the potential interactions among the factors by using process data and engineering knowledge.
- Step 5 - Choose the experimental design – design an experimental layout which shows all the possible combinations of control factors at their respective levels.
- Step 6 - Run the experiments, but be sure to randomize the runs and to replicate the runs if possible.
- Step 7 - Analyze the experiment data.
  - Compute the main effects, compute the interaction effects and identify the significant factors. Do graphical analysis by doing half normal probability plots.
  - Analyze which factor or interaction effects have variability by calculating the standard deviation as well as the log of the standard deviation for normal distribution.
  - Determine which of the factor effects and interaction effects have a significant impact on the response by doing ANOVA.
- Step 8 - Select the optimal control factor settings – combine maximum response and minimum variability to obtain the best settings for your best response outcome.
- Step 9 - Illustrate the cost reduction process capability where possible.

A further example of where DoE is used for process improvement is described by Teng et al. (2019). Teng et al. (2019) uses DoE in the framework but also uses principal component analysis (PCA) to

reduce the number of variables in the data (see **Figure 2.7**). PCA reduces the magnitude of your data, but still retain as much as possible of the information.

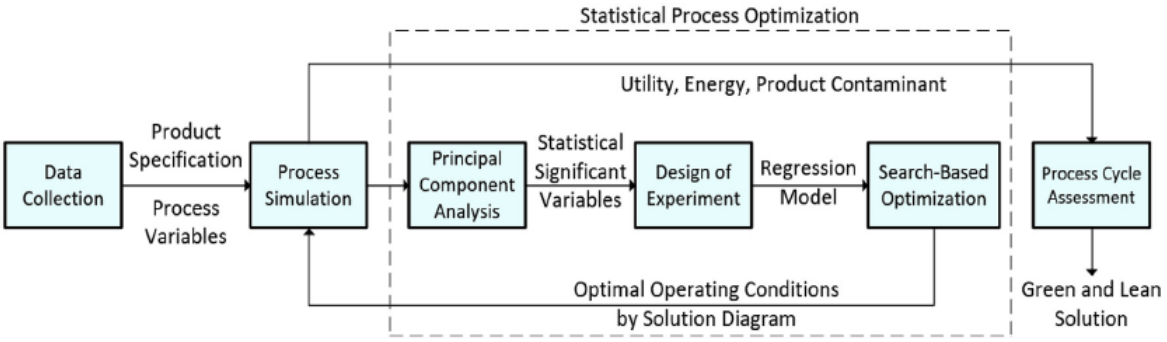

Figure 2.7: Overall process methodology for PCA-aided statistical process optimization (Teng et al. 2019)

According to Bera and Mukherjee (2018) a common problem generally encountered during manufacturing process improvements is to simultaneously optimize different responses in order to determine the best process operating conditions. The simultaneous optimization of multiple quality characteristics is generally referred to as a multiple response optimization (MRO). MRO helps to determine the input settings so that responses are close to their targets with minimum variances. There are two approaches to MRO:

- A response surface framework that consists of sequential steps of response surface methodology (RSM), problem formulation and optimization. According to Bezerra et al. (2008) RSM consists of a group of mathematical and statistical techniques used in relation to data used to develop empirical models during experimental design. The objective of RSM is to model experimental conditions until an optimal condition can be simulated.

- Data mining approaches to determines optimum input-variable conditions for an MRO such as;

  - The 'patient rule induction method' which is a black-box-type approach that can handle high dimensionality and is less sensitive to outliers.

  - An adaptive neuro-fuzzy inference system that has a developed approach that captures nonlinear relationships between factors and response variables.

The main objective of RSM is to optimize the response surface developed by the controllable input variables. In a response surface framework, the flowchart relationship between response variables and input variables is developed using models (see **Figure 2.8**).

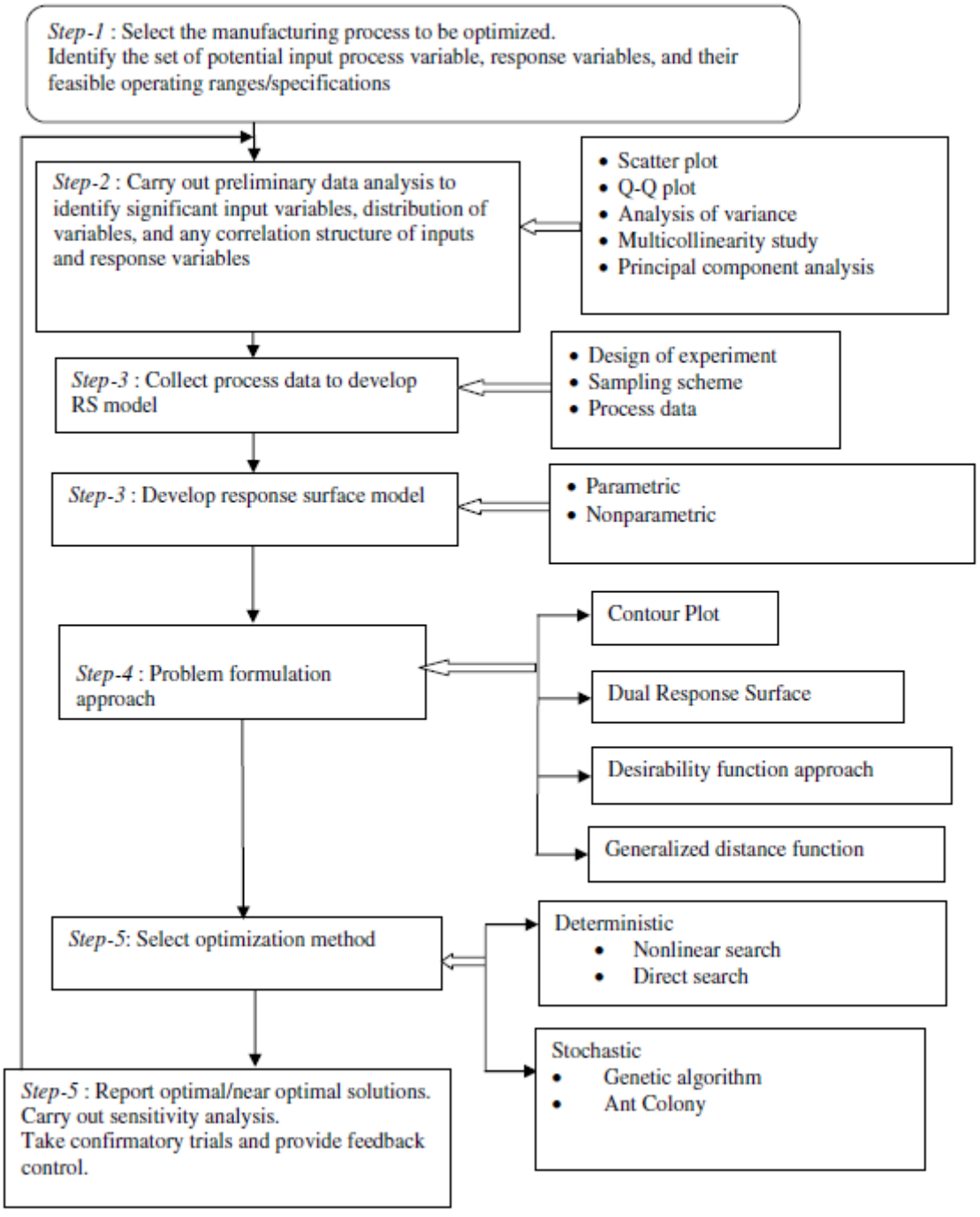

Figure 2.8: A response surface-based solution framework for MRO problems in manufacturing (Bera and Mukherjee 2018)

Bera and Mukherjee (2018) also uses DoE for process improvement but add RSM as part of the

improvement and control phase of process improvement. The response surface model establishes the functional relationship between controllable inputs and response variables. To establish such a relationship an empirical data driven model is needed. In many manufacturing facilities the input variable or response variations are assumed and in many cases the assumptions are unrealistic and therefore the need for data driven mathematical mapping of these relationships are needed.

2.4.3 Data science used in process improvement frameworks

Zwetsloot et al. (2018) states that lean six sigma projects are data driven in most stages of the DMAIC framework. In the last decade, in manufacturing, there are more process metrics available due to many technological developments which also lead to the availability of much larger datasets. Statistical methods such as t-tests and linear regression become less effective if you have larger data sets, therefore data science techniques become more effective to process data. Cross industry standard process for data mining (CRISP-DM) is one of the most used data science frameworks.

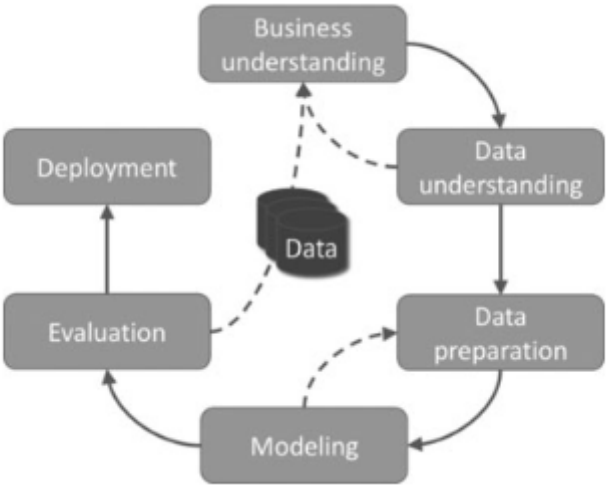

Figure 2.9: CRISP-DM phases and their relations for a data science project (Zwetsloot et al. 2018)

According to (Zwetsloot et al. 2018), the two differences between the traditional DMAIC methodology and a framework which incorporates data science are:

- Data science knowledge or expertise will be needed for projects that incorporate data science.
- The DMAIC is a sequential process and not iterative as which is the case with data science techniques.

Zwelsloot developed a framework that allows the DMAIC and CRISP-DM to run parallel to each other (see **Figure 2.10**).

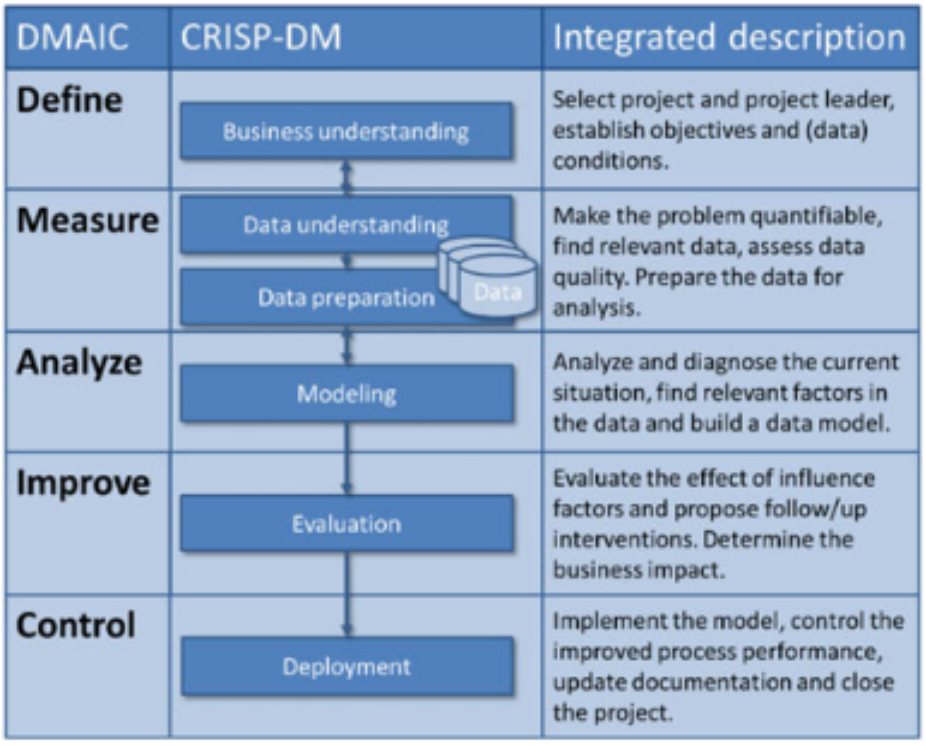

Figure 2.10: Integration of CRISP-DM in the DMAIC roadmap (Zwetsloot et al. 2018)

Khanbabaei et al. (2018) states that large organizations have many different processes which are typically poorly documented and the relationship between these processes are poorly specified. Khanbabaei et al. (2018) developed a framework with the aim to create a process ontology using process flows and data mining to categorize concepts of the organization and show relationships between them. Data mining techniques are used to extract, evaluate and classify patterns in the processes and to identify relationships between these processes. The findings from data mining are then used to improve processes by using the process improvement framework and the tools associated with it. Data mining is employed to select large amounts of data and find patterns in the data that are useful for the organization. There are various data mining techniques of which clustering and decision trees are examples. **Figure 2.11** shows the proposed integrated framework for data mining and process improvement using process ontology as the basis.

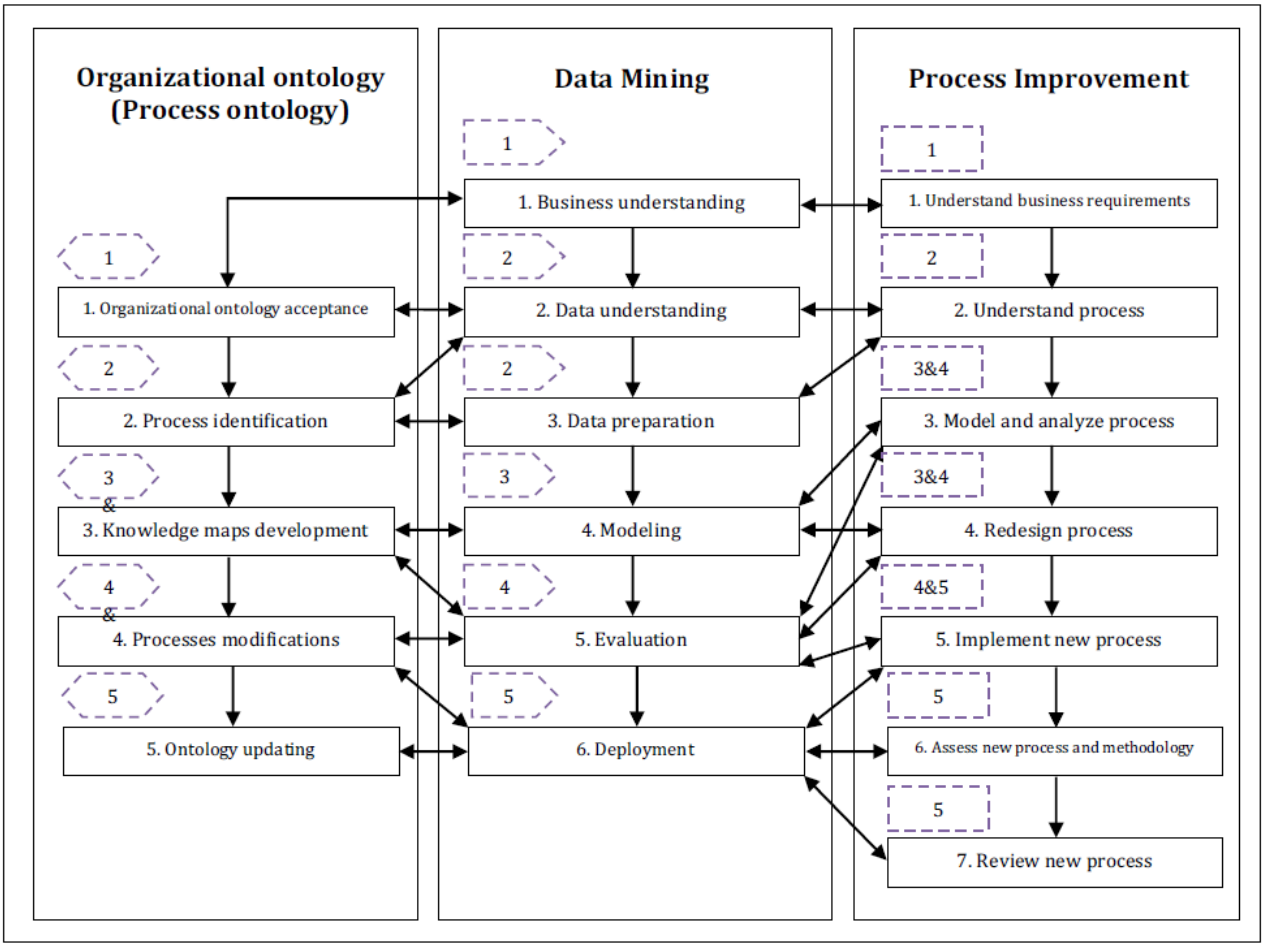

Figure 2.11: Integrated framework of data mining and process improvement based on organizational ontology (Khanbabaei et al. 2018)

From the previous sections it was shown that there are various frameworks developed or hybridized from the Lean Six Sigma methodology of DMAIC for process improvement. It is also clear from literature review that Lean Six Sigma does not always have the desired and sustained effects targeted for process improvement. To be successful in process improvement strategies, an organization needs to have, amongst other factors, the correct approach to reach the process improvement objective.

2.4.4 Decision routes for different process improvement strategies

Aqlan and Al-Fandi (2018) developed a new mechanism of how to prioritize improvement projects and identify the proper problem-solving tools. The framework could be applied over different applications. The framework consists of three phases:

- Phase 1 - Prioritizing workplace area (**Figure 2.12**)

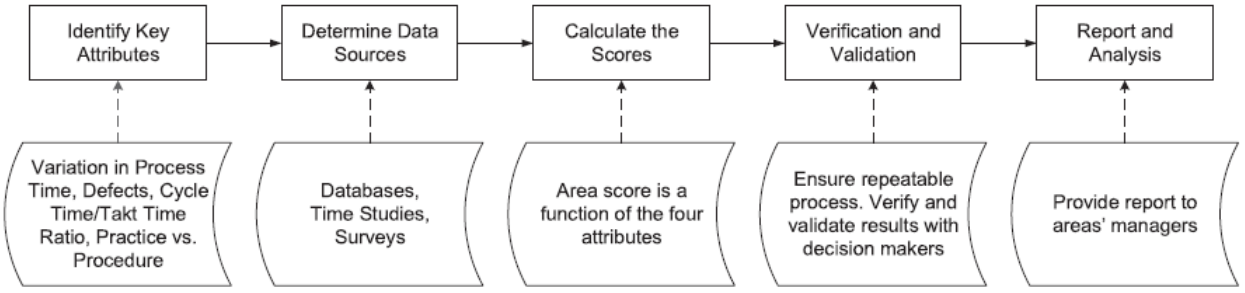

Figure 2.12: Prioritizing of workplace areas (Aqlan and Al-Fandi 2018)

In phase 1 of this approach the objective for improvement is defined. Relevant data is gathered from applicable resources. The data is analyzed to determine the areas in the process that need to be prioritized for improvement. The findings are verified and reported to the decision makers to decide whether to go ahead with the process improvement methodology.

- Phase 2 - Selection of problem solving methodology (see **Figure 2.13**). In phase 1 the problem area was identified. To identify the specific problem associated with the identified area, various process improvement methodologies can be followed. The team needs to decide which route to follow by brainstorming and root cause analysis.

- Phase 3 - Project selection - The process improvement project which is feasible or most valuable or most important is chosen. The maximum total organizational benefit, the minimum difficulty of implementing the selected project, budget constraints, time constraints and various other constraints are calculated and used to choose the process improvement project.

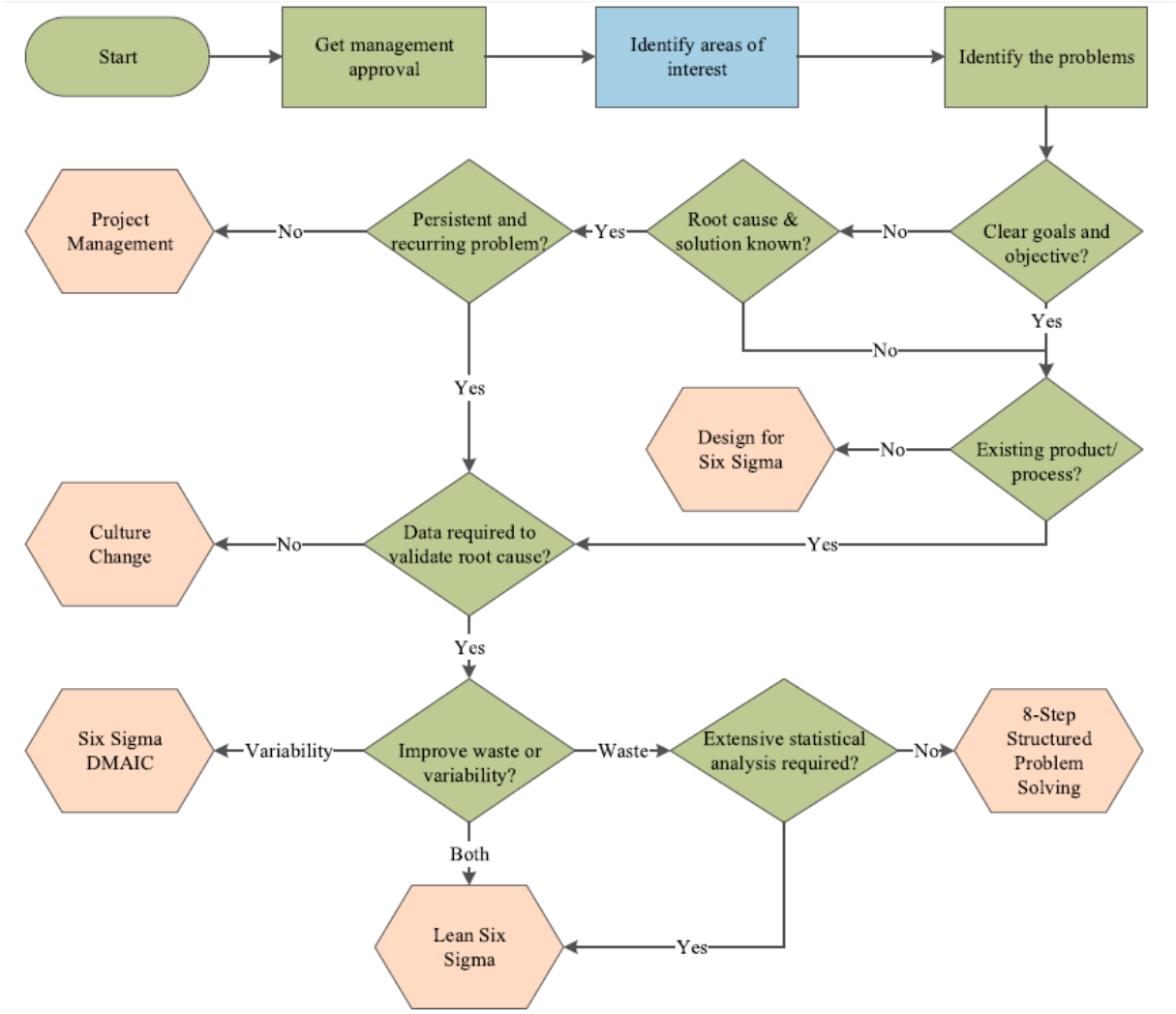

Figure 2.13: : Selection of problem solving methodologies (Aqlan and Al-Fandi 2018)

Aqlan and Al-Fandi (2018) elaborate on the first steps of process improvement methodology; defining the problem areas and initial measuring of process data to aid in decision making. Analysis associated with the framework is to assists in prioritizing and selecting a project for process improvement and is not associated with experimental or improvement data. Defining the objective and understanding clearly what needs to be achieved can lead you into choosing the correct approach for improvement of a process. Where no data and statistics are used in the process the solutions tend to be structural changes or cultural changes. Where data and statistics are needed for process improvement, Lean Six Sigma approaches are generally employed.

Once a decision has been made regarding the main methodology to use for problem solving, there may be more decision pathways within the chosen methodology. If Lean Six Sigma has been chosen because process data analysis is essential to improve the process, it is important to choose an

appropriate design for data measurement and analysis.

Engineers use statistics on a regular basis, and therefore they have to be able to apply it properly. DoE is an efficient method of experimentation to solve quality problems in key processes. In DoE you have to purposefully change certain input factors in a process to understand the changes in the outputs. DoE is part of Six Sigma but generally it has not been applied in a practical way in industry (Tanco et al. 2007).

According to Al-Ghamdi (2011) the type of DoE being employed depends on:

- Number of factors in your process that are being measured. A single factor design is one factor at two or more levels and is for comparative purposes. A factorial design is when two or more factors are varied simultaneously at two or more levels.
- Execution restrictions that might be necessary throughout your process are;
  - Randomizing is applied when there are unknown and uncontrolled noise factors.
  - Analysis of co-variance is applied when there are known and uncontrollable noise factors.
  - Blocking is applied when there are known and controllable noise factors.
- Factor levels can be fixed or random. Fixed levels are when specific levels of each factor are tested. Random levels are when the whole range between a factor's possible levels is of some interest.
- The objective of DoE can be to screen for the factors that have the most influence on the process, or to characterize how factors or interaction of factors affects the efficiency of the process, or to optimize the setting of factors to obtain the best response.
- The approach pioneer determines the design methodology that will be followed:
  - Fisher is a classical design used to study average response and its variation around its target along with the cause of the variation.
  - Taguchi studies the variation around the average response as well as the individual response around their average.
  - Shainin's main objective is to identify the most influential parameters on a process performance.

Figure 2.14 describes the classification of DoE types.

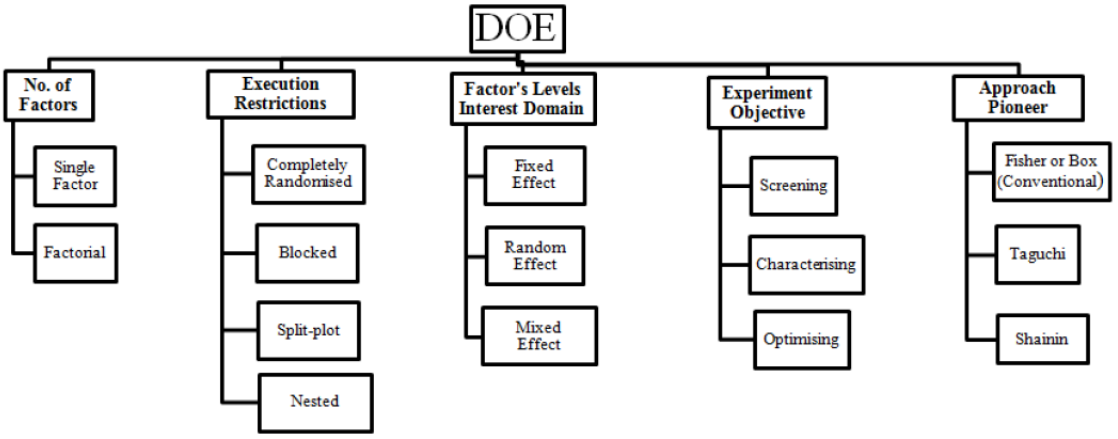

Figure 2.14: : Classification of DoE types (Al-Ghamdi 2011)

Zwetsloot et al. (2018) designed a matrix that assists in choosing the correct route to follow depending on the data volumes and complexity contained in the process or project. Data science forms part of the analysis approach when high volumes and high complexities of data are relevant to process improvement (**Figure 2.15**).

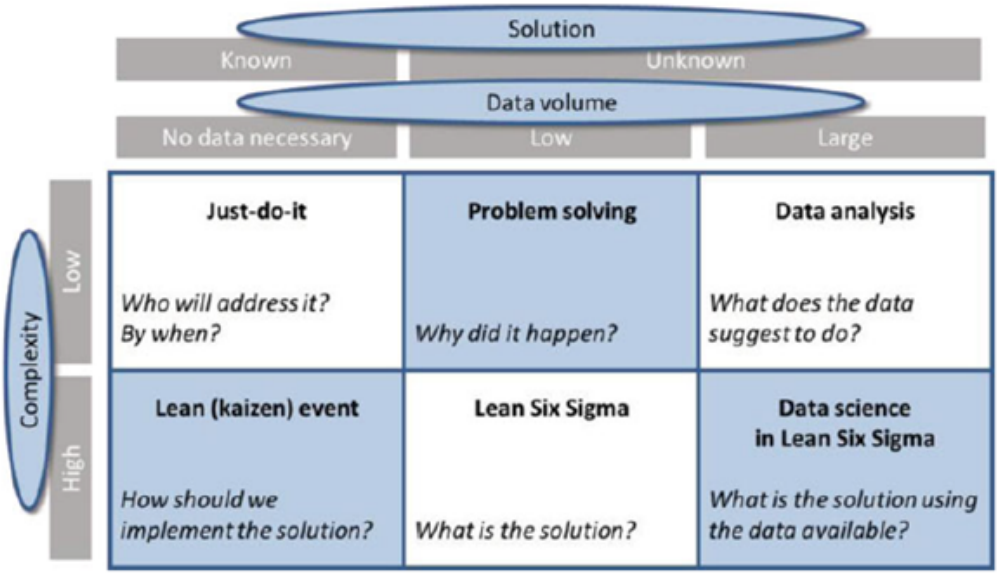

Figure 2.15: : Decision matrix to choose the appropriate process improvement strategy to use (Zwetsloot et al. 2018)

2.5 CONCLUSION

In Chapter 2 the relevant literature related to framework development for process improvement in a metal packaging manufacturing process is reviewed.

In the first part of this chapter literature surrounding metal packaging manufacture is reviewed. The importance and uses of the packaging industry is described. The manufacturing processes related to the two metals generally used in metal packaging, tinplate and aluminium, is described. The process to manufacture tinplate cans are described in detail.

The second part of Chapter 2 reviews various process improvement methodologies. The Lean Six Sigma process is highlighted as the process that contains the most complete approach towards process improvement. In this section the main steps of the Six Sigma DMAIC process is described in detail. Each of the five main steps is defined, their uses highlighted and various tools that can be used for each step is described. Lastly, in this section of Chapter 2, the main reasons for failures in the success of Six Sigma projects are discussed with added focus on reasons for failure in manufacturing improvement projects.

The third part of this chapter reviews various frameworks used in process improvement. Various frameworks are described and reviewed. The frameworks reviewed in this chapter mainly uses facets of the Lean Six Sigma steps and tools such as DoE and data science. The CRISP-DM framework used in data science and how it relates to the DMAIC framework is also described. Lastly in this chapter various decision mechanisms in process improvement is reviewed. The focus is on Six Sigma, DoE and data science and when you might choose which process improvement route to follow.

CHAPTER 3

MACHINE LEARNING

3.1 INTRODUCTION

The previous chapter reviewed frameworks used in process improvement, methods used in process improvement as well as metal packaging manufacturing. Chapter 3 of this thesis reviews machine learning, since data science and specifically machine learning, is used as an important tool in the framework that is developed for process improvement in metal packaging manufacturing.

3.2 CRISP-DM FRAMEWORK

Bekar, Nyqvist, and Skoogh (2020) describes how data science principals can be used in for predictive maintenance in an industrial setting. The methodology that was followed in the case study Bekar, Nyqvist, and Skoogh (2020) was the CRISP-DM method as described in **Figure 2.10**. Also see **Figure 3.1**.

- The business outcome has to show what decisions can be made on the current data streams. After the business understanding phase there should be an understanding of quality problems related to the available data. The following aspects related to the data should be understood:
  - Accessibility of the data
  - Relevancy of the data
  - The different sources of data and how well these sources are synchronized
  - Completeness and correctness of the data
  - Noise associated with the data
  - Uniformity of the data
- The data preparation outcome is to get a final dataset from raw data. Data preparation is an evolutionary type of process that can have multiple steps and iterations, that can include addition of recorded data as well as data cleaning and formatting. The following additional steps should be considered after a final data set has been obtained:
  - Data should be scaled if clustering is going to be used in the modelling and analysis of the data. Scale the data by normalizing or standardizing the data.

  - Data's dimensionality can be reduced by using principal component analysis, which in turn also helps to point out similarities and differences in the data.

  - Find patterns in the data using clustering methods.

- The data modelling outcome is built models and algorithms. The models and algorithms can be developed by using programming languages such as Python. Modelling and data preparation can overlap in back-and-forth steps as data format or requirements are met.

- The data evaluation outcome is the decisions reached and conclusions made from a model outcome that has been built using good and relevant data.

- The deployment outcome is the implementation of the findings in a practical and user-friendly way. From here the cycle can be repeated again.

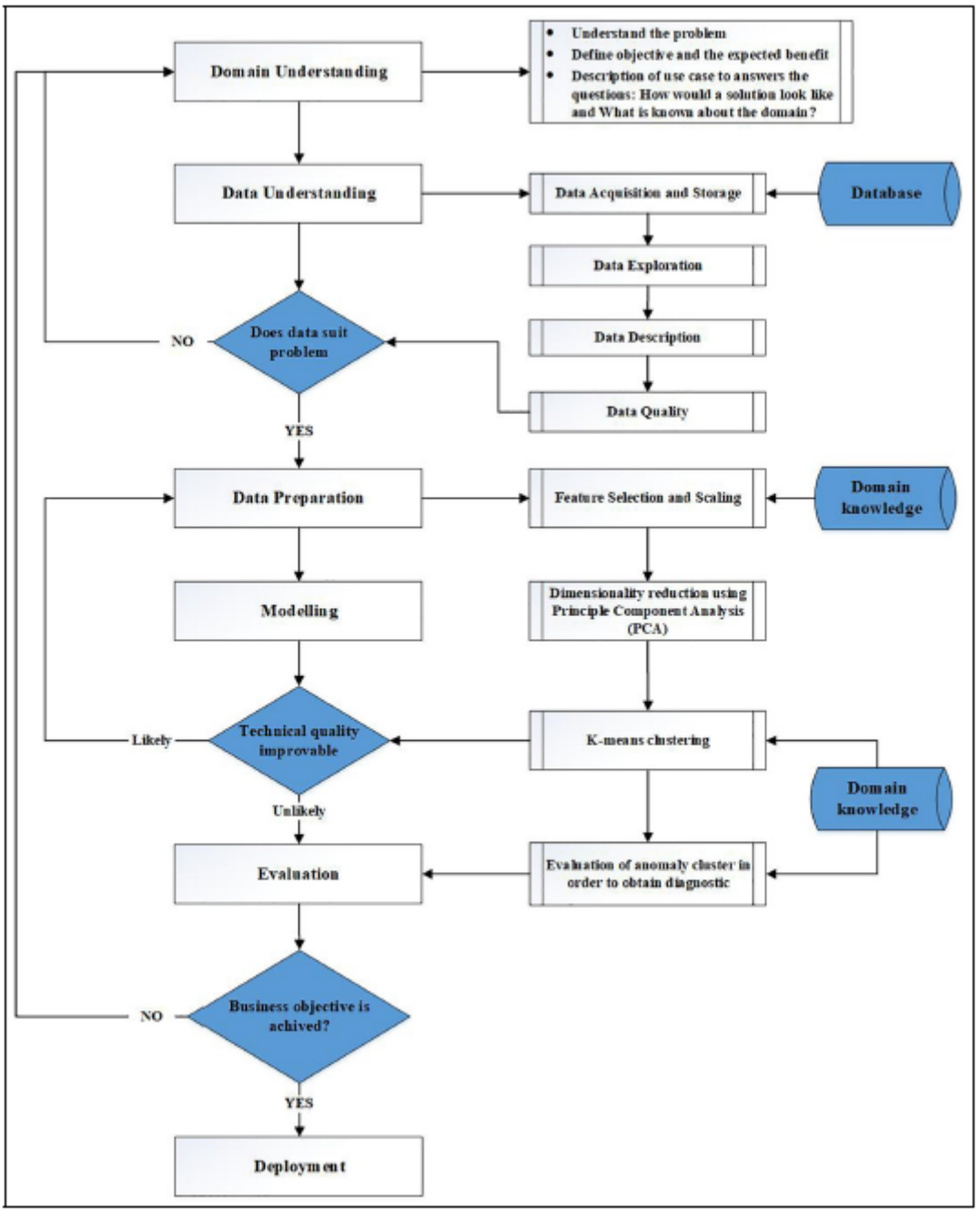

Figure 3.1: : The flow diagram of the formulated approach based on CRISP-DM methodology (Bekar, Nyqvist, and Skoogh 2020)

3.3 MACHINE LEARNING USED IN MANUFACTURING

Wuest, Irgens, and Thoben (2014) state that the availability of quality related data in manufacturing has much potential to apply machine learning and industry 4.0 principles. Huge amounts of data can make focused quality improvement difficult, but data-driven approaches such as ML can be used to overcome some of the challenges manufacturing faces today. ML can be used when there is a great deal of factors in a process, by using a technique called the Support Vector Machine (SVM) technique. In such a case, where SVM is utilized, the need to reduce the dimensionality of the data is less important. Seemingly irrelevant data can be included in these types of models and it can possibly be discovered that this data can be relevant under certain circumstances. The goal of ML is to detect patterns that can describe relations. ML uses existing data to make predictions, therefore data can only become useful if it is analyzed in a proper way. The challenge of ML lies as much in the capturing and availability of relevant data then in the subsequent ML analyses.

Wuest, Irgens, and Thoben (2014) further list a few advantages as well as challenges that can be associated with ML. Advantages of ML applications in manufacturing are that it:

- can be used when data has high dimensionality.
- can be used to undercover previously unknown correlations and relationships.
- is dynamic in the sense that ML uses the current data, but can adapt as time and new data changes.
- can discover relationships and therefore assists with decision making (or can possibly be used in automated responses).

Some of the challenges of ML applications in manufacturing are:

- the obtainment of relevant data.
- that pre-processing of data is needed, and for this, the correct approach must be followed depending on the characteristics of the data.
- that the correct ML algorithm must be used (supervised, unsupervised or reinforcement learning) and the applicability of the model must be assessed.
- that the results must be interpreted correctly in terms of over-fitting, bias and variance.

Wuest, Irgens, and Thoben (2014) also distinguish between reinforcement learning, supervised and unsupervised learning. Unsupervised learning is used when there is no feedback from a knowledgeable source or person. Therefore the algorithm, determines which clusters of data fit together most cohesively, in terms of all the data fed into the model. Reinforcement learning is based on the feed-

back on actions from the environment. **Figure 3.2** gives a summary of the types of machine learning algorithms that can be used in manufacturing. In manufacturing there normally would be knowledgeable persons that could give some feedback on which factors are important and on how these factors are then fed into a training model, therefore neither unsupervised learning, nor reinforcement learning would necessary be used in process improvement. Supervised learning is generally applied in manufacturing due to the fact that there is a lot of data that can be used to solve a problem of which there is not much knowledge. According to Cunningham, Cord, and Delany (2008) Supervised learning is the learning of the relation between a set of input variables (factors) and an output variable (response). This relationship is used to predict the response from unseen data.

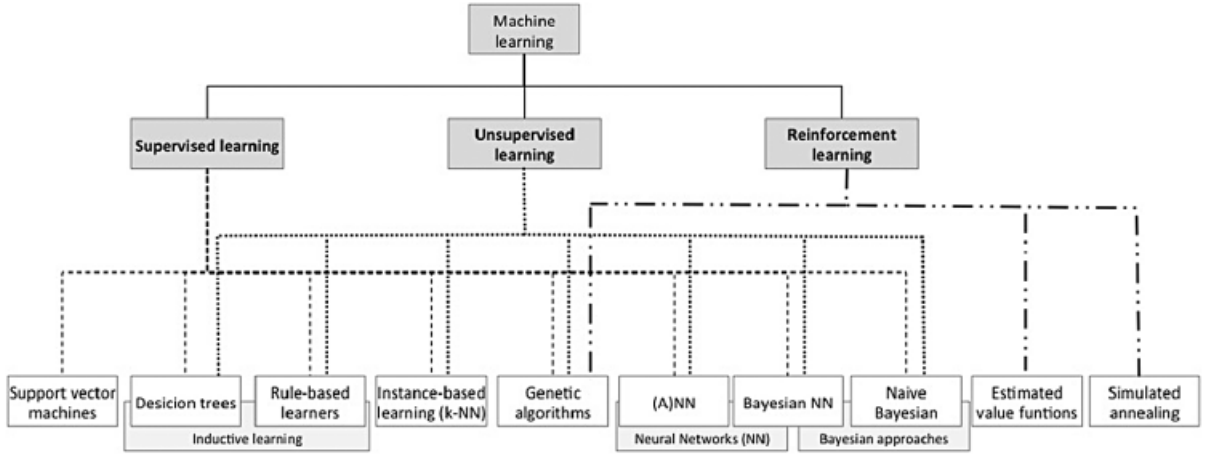

Figure 3.2: : Structuring of ML techniques and algorithms (Wuest, Irgens, and Thoben 2014)

The basic steps to employ supervised ML with data from a process are shown in **Figure 3.3**, and outlined as follow;

- Understand the objective.
- Describe the problem by understanding and describing the data.
- Find solutions for your problem by using ML models and their evaluations.
- Deploy the solution (Bowles 2019).

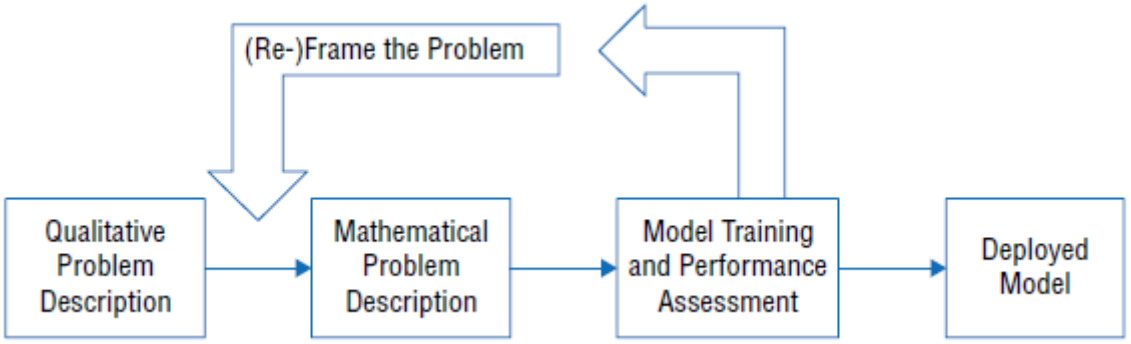

Figure 3.3: : Steps from formulation to performance for a ML framework (Bowles 2019)

3.4 TYPES OF MACHINE LEARNING

ML in data science is used to make informed decisions and typically is used when:

- tasks need to be automated.
- anomalies need to be detected.
- complex analyses need to be performed.
- events need to be predicted (Subasi 2020).

Machine learning uses computational techniques for predictions or to improve performance. Machine learning converts data into information by using various principles of statistics, engineering and computer science. Many fields can use machine learning such as finance, medicine and manufacturing. The machine learning technique is to use data in a computer program known as a model. This model can then be used to give insight and aid in decision making by recognizing characteristic patterns in the data.

The main types of machine learning problems are:

- Classification where the response is predicted as a category.
- Regression where the response is predicted as a real value.
- Ranking where the responses are ranked according to certain factors.
- Clustering where the responses are grouped into homogeneous groups.

- Dimensionality reduction where data is represented in a lower dimension, but without losing too much information.

To solve real-world machine learning problems, machine learning frameworks are used that begin with data collecting and end with valuable information or knowledge. **Figure 3.4** gives an example of a typical machine learning framework.

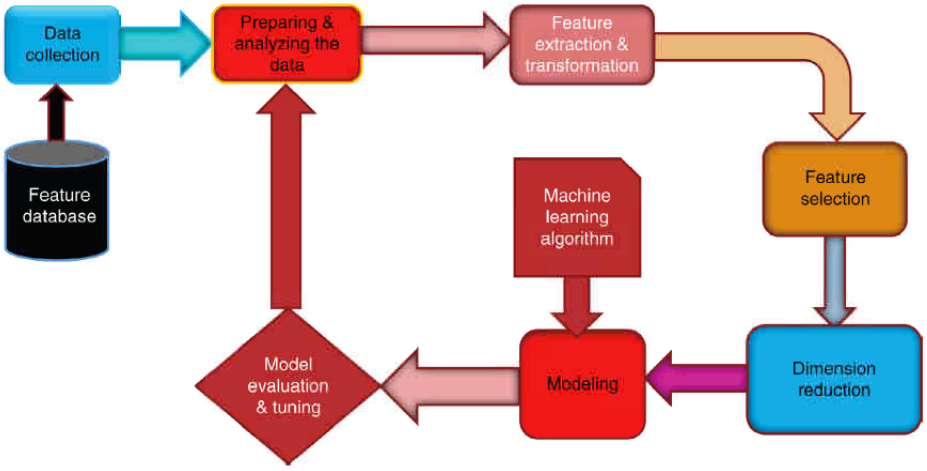

Figure 3.4: : An overview of a typical machine learning framework with the main stages highlighted in their blocks (Subasi 2020)

Machine learning can be either supervised or unsupervised. According to Raschka (2015) Supervised machine learning predictions can be categorical or continuous. Categorical predictions assign unordered labels for new inputs, and continuous predictions assign real numbers on a continuous scale for new inputs by using regression. **Figure 3.5** shows a simple representation of the outcome of a categorical classification model used in supervised ML. **Figure 3.6** shows a simple representation of the outcome of a regression classification model used in supervised ML.

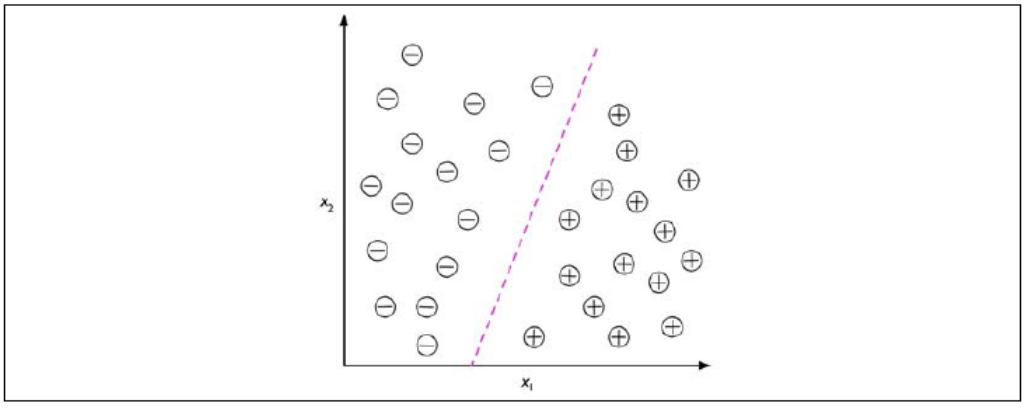

Figure 3.5: :Example of categorical classification (Raschka 2015)

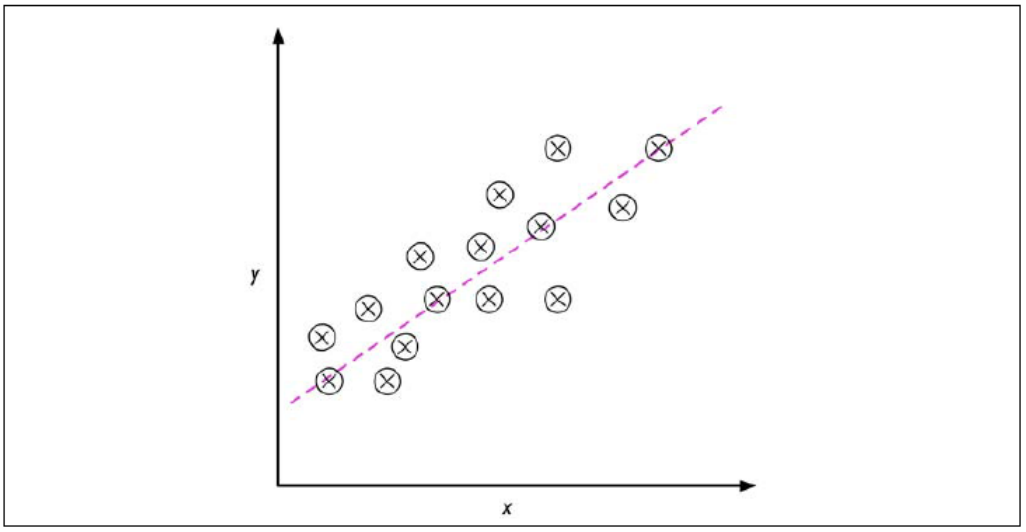

Figure 3.6: : Example of continious regression classification (Raschka 2015)

Unsupervised machine learning attempts to organize unstructured data. Clustering and dimensionality reduction are two methods used in unsupervised ML. **Figure 3.7** shows a simple representation of the outcome of a clustering model used in unsupervised ML. **Figure 3.8** shows a simple representation of the outcome of dimesionality reduction from 3D to 2D in unsupervised ML.

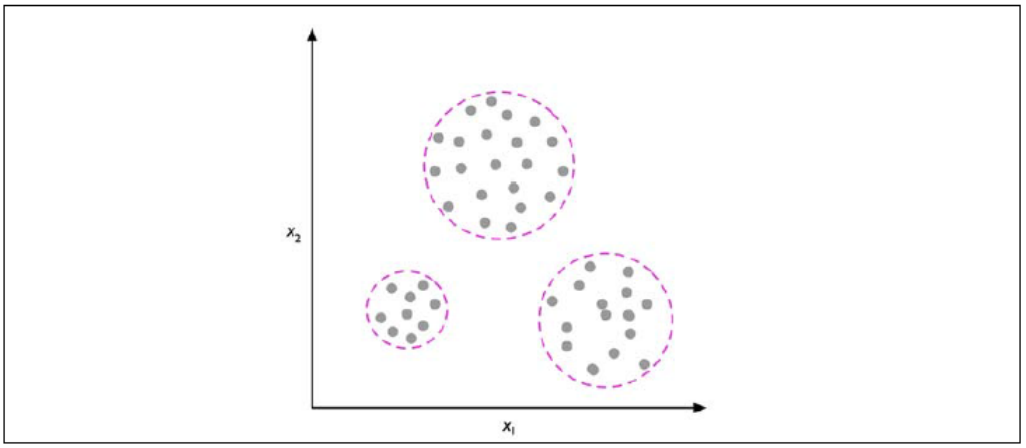

Figure 3.7: :Example of clustering classification (Raschka 2015)

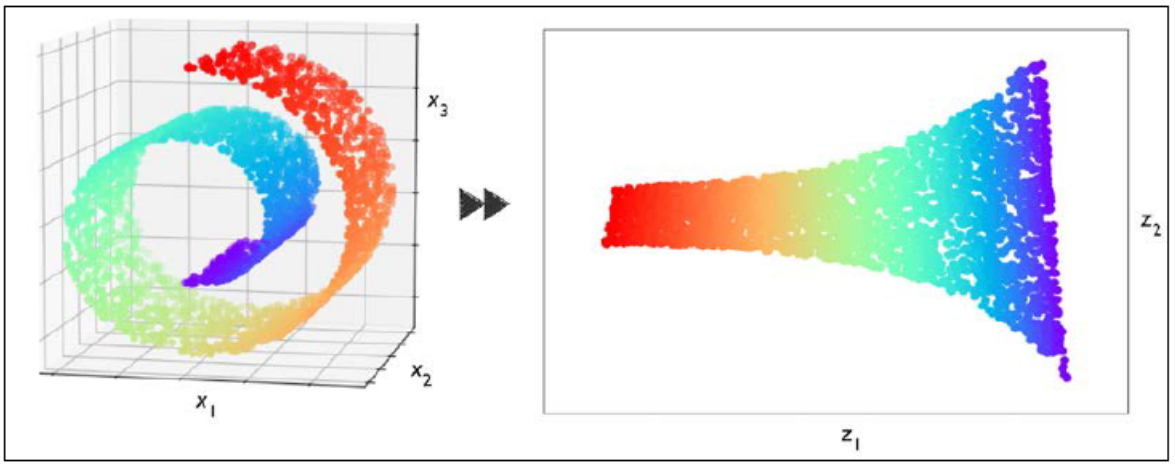

Figure 3.8: : Example of dimensionality reduction classification (Raschka 2015)

According to Deisenroth, Faisal, and Ong (2020) the four pillars of ML is regression, dimensionality reduction, density estimation and classification. The mathematics that form the basis of these ML pillars are the study of vectors and matrices. Numerical data is represented as vectors and represent a table of such data as a matrix.

ML uses an algorithm as predictor to predict outputs that are similar to previous vector inputs and outputs. To show the similarity between vectors, a numerical value representing their similarity is calculated. The construction of similarity and distances between vectors is central to analytic geometry. ML identifies the true underlying signal from data that seems to be a whole lot of noise. ML uses these signals to make predictions and also quantify the uncertainty of these predictions.

To be able for a ML model to make a prediction, these models need to be trained. The training of machine learning models involves the identifying of those factors that increases the performance of the model. The optimization of these ML models requires the use of vector calculus to find the maxima or minima of the functions which need to be optimized.

Subasi (2020) list a few challenges due to the complicated algorithms associated with ML:

- To be able to have a high level of data processing and feature extraction, the data quality must be good.
- Many aspects of data acquisition and pre-processing are very tedious and time consuming.
- There should be enough training data to build a proper ML predictive model.
- Extracting features and reducing data dimensionality can be challenging to do correctly.
- To be able to express clear business objectives can be difficult.
- Over- and under-fitting of data will lead to poor performing models.
- Sometimes models are too complicated because there are too many factors involved.
- Implementation of complex models in a real world situation can be difficult.

3.5 DATA

3.5.1 Cleaning and exploration

3.5.1.1 Data cleaning

For the purpose of using data in a data science programming language, such as Python, data should generally be arranged in a tabular form. The table will consist of the independent variables also known as factors, and the dependent variable, also known as the response. The factors and response will be represented as column headings, and each row in the table will represent a specific instance or sample data.

The reason data should be in a tabular form is to represent the data as vectors and therefore be able to use linear algebra in the algorithms. In many ML algorithms vectors are compared to be able to compute the similarity between vectors. This comparison of vectors requires the construct of a geometry which in turn allows for optimization of such a construct.

According to Deisenroth, Faisal, and Ong (2020), in recent years, ML has been applied to many types of data that do not obviously come in the tabular numerical format such as genomics sequences,

text and image contents of a web-page, and social media graphs.

Data cleaning is the first step of representing the data as vectors in order to construct the predictive function. The predictive function predicts an output when presented with input examples. According to Klosterman (2019) data cleaning can include:

- the determining of the number of columns and rows in the data table.
- the determining of whether factors are categorical or numerical.
- the determining of the frequencies of data categories and the ranges of numerical data.
- The determining of the data integrity by identifying missing data, repeated data or inconsistencies in the data. According to Bowles (2019), the way to handle missing data is either to discard rows that have missing data, but if these discarded rows are a large part of your data tables it might influence your results. Another option to handle missing data is to fill them, which is called imputation. Ways to impute data is to replace the missing values with average values of the relevant attribute, or to use algorithms that replace the missing values with predicted values.

3.5.1.2 Data exploration

According to Raschka (2015) data exploration is the first step to take before ML model training. Data exploration consists mainly of basic statistics on the data and the visualization thereof;

- Basic statistics such as mean values, minimum and maximum values and standard deviations
- Visualization of the data by using tools such as;
  - Histograms visualize the data on a scale grouped together in bins of data as seen in **Figure 3.9**.

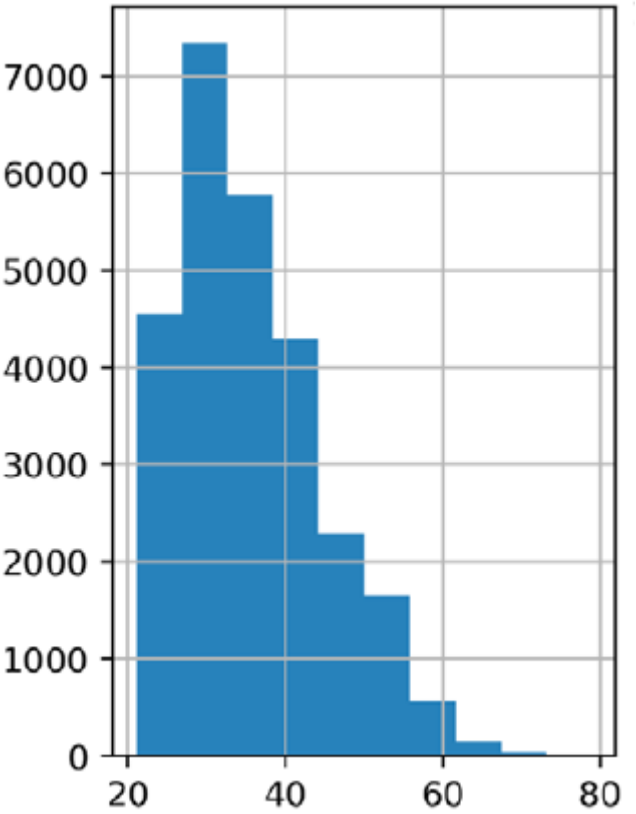

Figure 3.9: : An example of a histogram (Klosterman 2019)

– Quantile-quantile plots assits to visualize the distribution of data and identify outliers in the data. Quantile-quantile plots as seen in **Figure 3.10** are scatter plots of the factor you are testing against the distribution you are testing it against. It is a visual tool and therefore any interpretations from this graph are fairly subjective. What to look for in this graph is whether the graph generally lies in a straight line at a 45° angle which will indicate that the data is distributed normally. Any clear digression on the edges of the line of a few data points might be outliers.

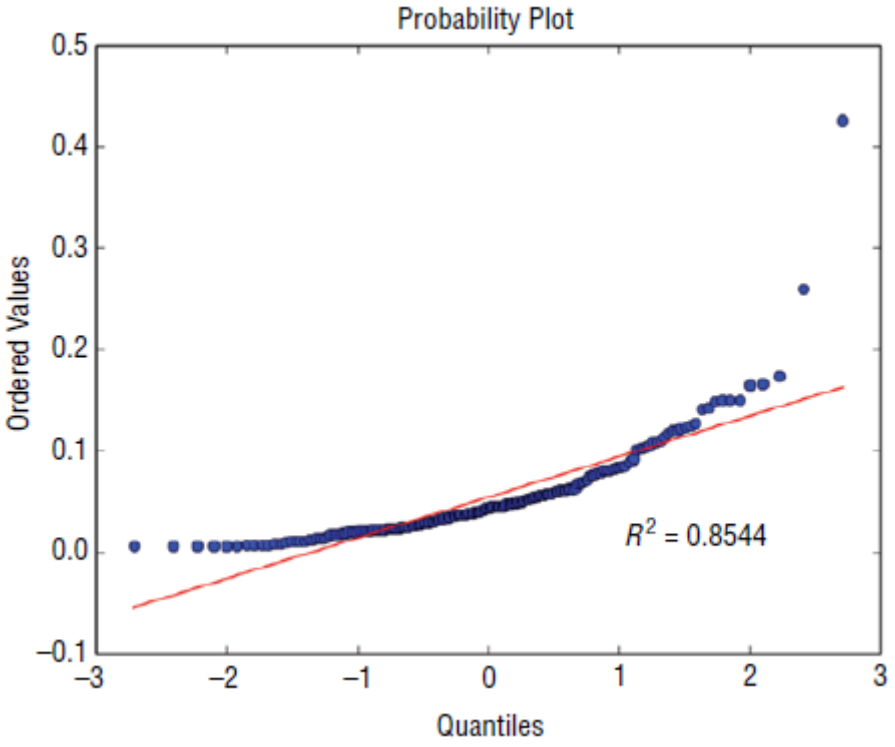

Figure 3.10: : An example of a quantile-quantile plot (Bowles 2019)

Outliers might cause problems during model building as well as when making predictions. Once the model has been trained with the data that contains the outliers, the accuracy of the model can be assessed. If errors are associated with these outlier values, then the outliers can be approached in various ways; Outliers can be removed if they were abnormalities that would not normally be in the data that the model will have to process, or they can be removed if they were a true error, such as wrongly captured data. Outlier data can be separated as a separate class or attribute of your data set and then used in the model. If outliers are a true reflection of the data, they can be better represented in your data by increasing the data volumes, or replicating conditions where these outlier values are produced (Bowles 2019).

– Scatter plots to visualize the correlation between two attributes. Scatter plots, as seen in **Figure 3.11**, plot two variables along two axes as points. The pattern of the points can reveal any correlation between the two factors. The response and factors that will form part of the ML model can be visually represented with these scatter plots.

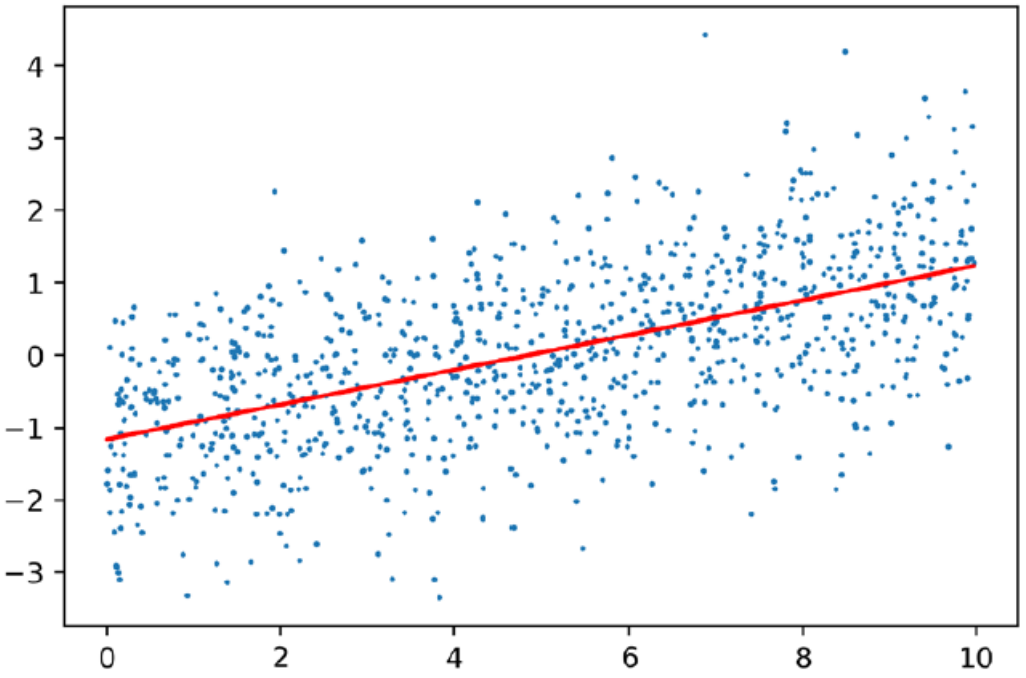

Figure 3.11: : An example of a scatter plot (Klosterman 2019)

– A heat heat map can be used for visualization of correlations between factors when there are a large number of factors in the data set. Correlation heat maps, as seen in **Figure 3.12**, are represented as coloured cell blocks in a 2D matrix. Each cell in the matrix represents the Pearson's correlation of two factors by a colour code where red is a perfect correlation and blue is no correlation. The correlation between two factors or a factor and the response can be quantified by the Pearson's correlation coefficient (r), which is the co-variance between two factors divided by their standard deviations (Raschka 2015)

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

A Pearson's correlation coefficient value of $r = +1$ indicates a perfect positive correlation, $r = -1$ indicates a perfect negative correlation and $r = 0$ indicates no correlation. High correlation between two factors is called multi co-linearity and can influence the predictions the model makes. Multi col-linearity can be eliminated by dimensionality reduction if only one of the co-linear factors is included in your model, or the co-linear factors are combined to form one factor in the ML model. **Section 3.5.2** discusses dimensionality reduction in more detail.

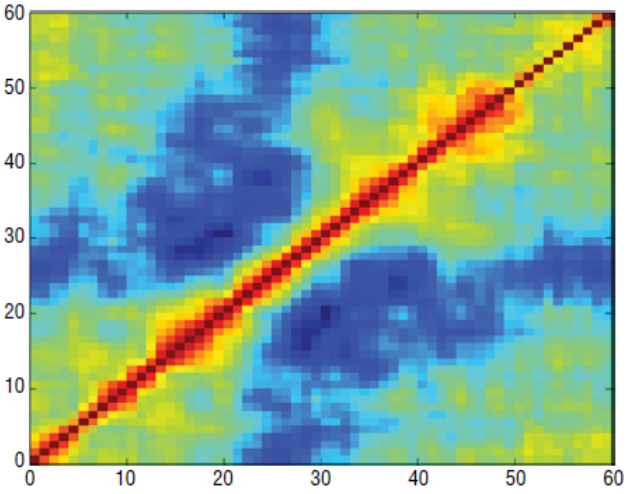

Figure 3.12: : Example of a heat map displaying the correlations between factors in a data set (Bowles 2019)

– Box plots, as seen in **Figure 3.13**, visually show the median value, and the 25th percentile and the 75th percentile, respectively, as a box. So called whiskers are drawn in at levels that are 1.4 times the inter-quartile spacing above and below the box. Data points that extend beyond the whiskers can be described as outliers (Bowles 2019). Box plots is a fast visual method to identify outliers. To be able to represent different factors with different scales on the same graph, the data will need to be scaled. **Section 3.5.2** discusses the scaling of data in more detail.

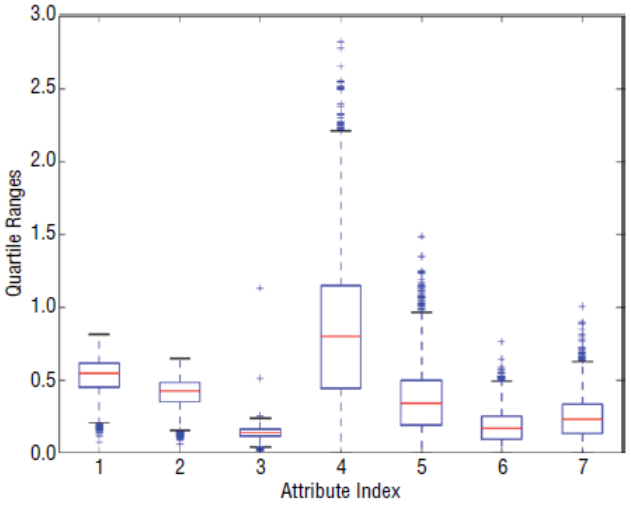

Figure 3.13: : Examples of Boxplots (Bowles 2019)

– Parallel coordinate plots are a useful visual tool to see systematic coordination between attributes as seen in **Figure 3.14**. In parallel coordinate plots vertical bars represent a factor and row values are plotted as series of lines connected across each axis. The parallel coordinate plot is an effective way to visualize high dimensional data-sets because it allows the comparison of many data observations on a set of numerical variables. In a parallel coordinate plot the colour coded lines represent higher to lower values of your response variable. This colour coded line runs through all the factors in your data set and relates how these factors correlate with the response variable (Bowles 2019).

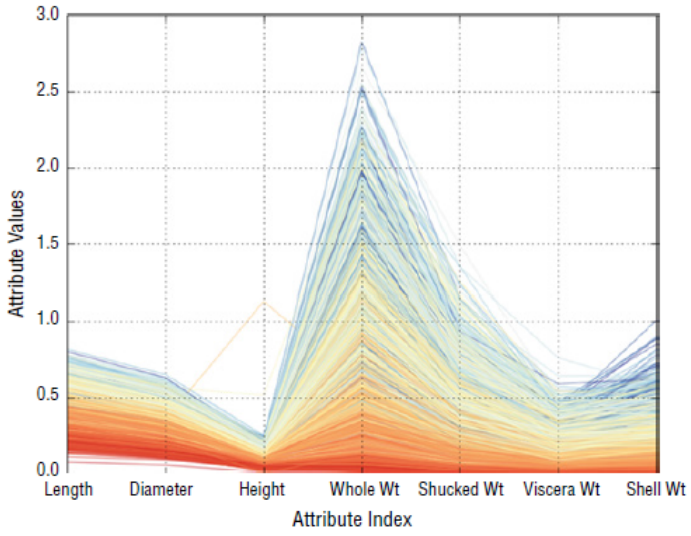

Figure 3.14: : An example of a colour coded parallel coordinates plot (Bowles 2019)

3.5.2 Understanding and preparing data

Understanding of the data mainly comprises assessing the data quality and then preparing the data for ML modelling.

3.5.2.1 Data assessing

Assessing of data will include tasks such as;

- Determining the data types. The data types need to be established since data can be seen as objects, integers, real numbers, ordinal categories, date/time and indexes. The data types need to be changed to what makes sense for each factor's data as well as what the intended algorithms is that will be used in the modelling phase. Categorical factors can be changed to numerical values by using One Hot Encoding in Python programming. One Hot Encoding creates a new column for every category of the categorical factor and assigns a numerical 1 in the column that represent that instance's category and zeros for the rest of the created columns (Klosterman 2019).

- Combine different data tables into one data table that will be used in the modelling phase.
- Perform ANOVA on the categorical features to determine whether the categories of these features are statistically different or similar. The following steps need to be followed when doing a comparative analysis by using ANOVA as described in Montgomery (2017).
  - State the statistical hypothesis of the factor you are analyzing.

    $$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$$

    $H_1 : \mu_i \neq \mu_j$ for at least one pair of i and j and where μ_1 to μ_n are the means of the n levels of factors.
  - Calculate the sample mean values of all the factor levels;

    $$\overline{y} = \frac{\sum_{i=1}^{n} = y_1}{n}$$

  - Decide on the significance level of the analysis by choosing an α value. α is the probability of committing a Type I error, or the chance of rejecting H_0 when H_0 is true.
  - Calculate the total sum of squares;

    $$SS_T = \sum_{i=j}^{a} \sum_{j=1}^{n} (y_{ij} - \overline{y})^2$$

    Where n is the total number of observations and a is the total number of levels for the factor.
  - Calculate the sum of squares between the levels and the factor;

    $$SS_{Treatments} = n \sum_{i=j}^{a} (\overline{y}_i - \overline{y})^2$$

  - Determine the mean square between the levels of the factor;

    $$MS_{Treatments} = \frac{SS_{Treatments}}{(a-1)}$$

    Where $(a-1)$ is the degrees of freedom.

– Determine the sum of squares of the error within the levels of the factor;

$$SS_E = SS_T - SS_{Treatments}$$

– Determine the mean square of the error within the levels of the factor;

$$MS_E = \frac{SS_E}{(N-a)}$$

Where $(N-a)$ is the degrees of freedom, and N is $a * n$.

– Determine the F-test statistic;

$$F_0 = \frac{MS_{Treatments}}{(MS_E)}$$

– Read the $F_{\alpha,\alpha-1,N-1}$ value from the t-distribution table, where $a-1, N-1$ is the degrees of freedom.

– Accept or reject your null hypothesis; Reject

$$F_0$$

if

$$F_0 > F_{\alpha,\alpha-1,N-1}$$

3.5.2.2 Data preparation

Data preparation mainly consists of those tasks that will change the data in such ways to prepare it for modelling of the data. The main tasks associated with data preparation are data scaling and data dimensionality reduction. Scaling of data can be done by normalizing the data or standardizing the data.

According to Raschka (2015) many ML algorithms are scale sensitive, which means the ML models will perform better when all the factors are brought onto the same scale. Forest tree regression and random forest regression are two ML algorithms that are not sensitive to scale and therefore will not need feature scaling. Feature scaling is either done by normalization or by standardization.

Normalization is the re-scaling of data between 0 and 1, known as a min-max scaling:

$$x^{(i)}_{norm} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}}$$

Standardization may be a more practical use of scaling when using ML algorithms. During standardization, the feature column is centred on 0, with standard deviations of 1. In this way a feature column has the same parameters as a normal distribution. Standardization also keeps useful information with regards to outliers:

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

Standardization is only applied once on the training dataset. Those parameters are then used on the test set as well as on new data.

Dimensionality reduction of the data will include tasks such as;

- Feature selection which removes ineffective features in order to make faster and more accurate predictive analytics possible.

- Feature extraction which is the process of identifying the most informative set of factors from the data set that still gives a suitable amount of information. Feature extraction takes a set of factors' dimensions and combines those dimensions to form a new set of dimensions (Subasi 2020).

According to Raschka (2015) if a model performs much better on a training data set than on a test data set it is an indication of over-fitting of the data. Over-fitting occurs when the data fits the parameters of the training data set very closely, but this fit does not generalize to other data for the same parameters. Over-fitting occurs if the model is too complex for the specific training data set. To prevent over-fitting; use more training data, use a simplified ML model, reduce the dimensionality of the data or introduce regularization to penalize complexity. **Figure 3.15** illustrates the under-fitting and the over-fitting of data.

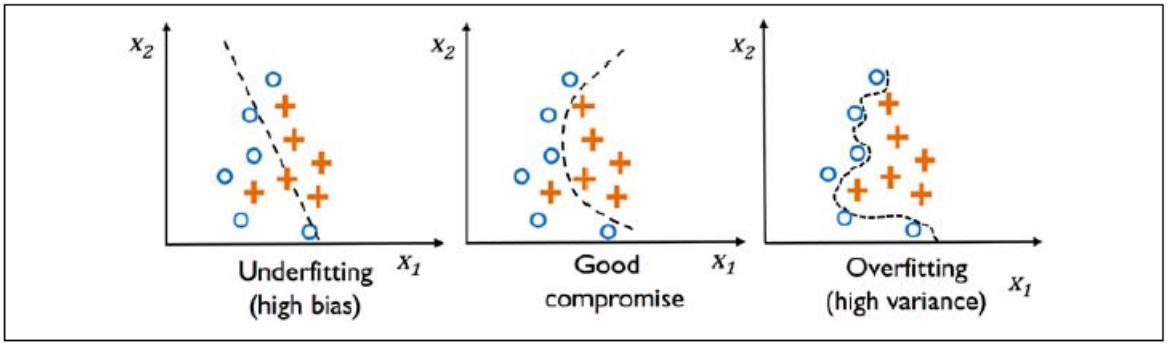

Figure 3.15: : An example of under- and over-fitting of data (Raschka 2015)

One way to reduce dimensionality is with feature selection. Feature selection is, as stated, a form of dimensionality reduction but also aims to decrease the complexity of the model by only selecting

those features that are most relevant to your problem. Feature selection algorithms can be used for feature selection, e.g. sequential forward selection (SFS) or sequential backward selection (SBS). Random forest regression can also be used to determine feature importance. The disadvantage of using random forest is that if two features are highly correlated, the one feature might rank low and the other feature might rank high, but this is only an issue if feature importance to your problem is of importance. If predictions are all that you are interested in then this disadvantage will not matter.

SFS and SBS are classical feature selection algorithms. According to Haq et al. (2019) SFS and SBS are algorithms that eliminates features (factors) from the data table in a sequential manner until a stated criteria has been achieved. The performance of the classifier is measured after each sequence after which the factor that maximizes the criterion is eliminated. Sequential elimination continues up to the point where the best performance can be achieved.

According to (Raschka 2015) feature extraction can be seen as an alternative approach to feature selection. Feature extraction summarizes the content of data by transforming the data onto a new feature space with less dimensionality. Feature extraction is a form of data compression and is important because it makes data storage and analysis easier.

There are various feature extraction methods, including the following:

- PCA is used for unsupervised data.
- Linear discriminant analysis (LDA) is used for supervised data.
- Kernel principal component analysis (KPCA) is used for nonlinear dimensionality reduction.

Both feature selection as well as feature extraction decreases the number of factors in the data set, but with feature selection the features do not change whereas with feature extraction the features are transformed. Feature extraction not only compresses the data, but also takes away some of the data's dimensions, which can make the predictions from that data more accurate.

PCA is a method where new features are built from existing features. PCA is a multivariate statistical analysis that finds transformations of the multivariate data, which gives a smaller and more meaningful data set than the original data set (Subasi 2020). According to Raschka (2015) PCA can, in addition to feature extraction, also be used as exploratory data analysis. PCA identifies patterns in data by looking for correlations in the data. The process of PCA is to look at the highest variance between feature sets and then projecting the data onto a subspace of lower dimensionality as visualized in **Figure 3.16**.

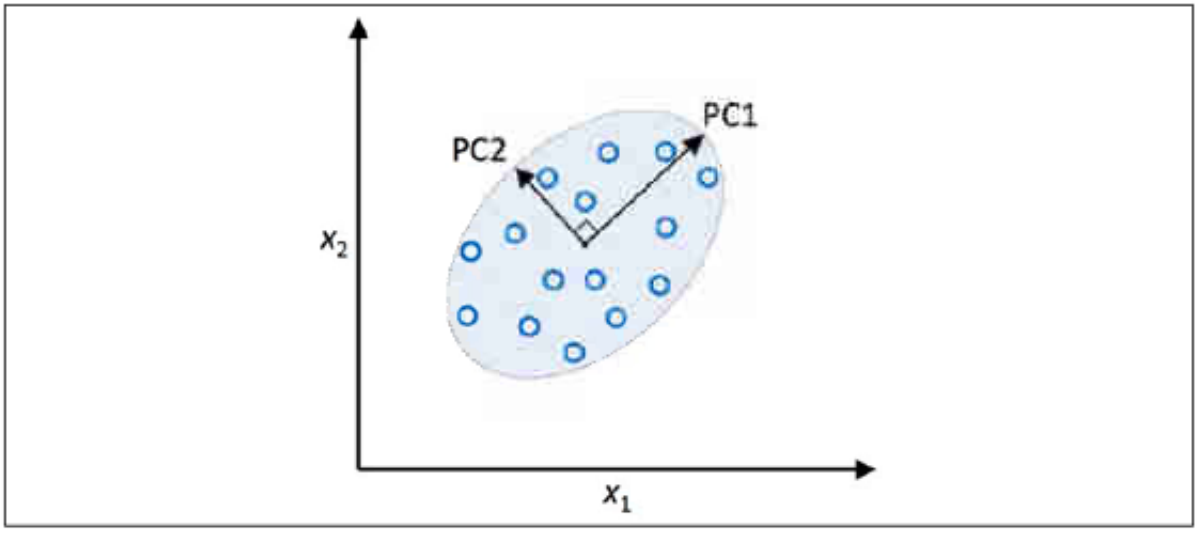

Figure 3.16: : Visual depiction of PCA (Raschka 2015)

Another feature extraction method is LDA, which is a supervised data compression technique. According to Raschka (2015) LDA is also a feature extraction method that can be used to increase the computational efficiency of a model. The goal of LDA is to find sub-spaces in the data where the factors can be separated most efficiently. **Figure 3.17** demonstrates LDA where LD1 is a good linear discriminant and LD2 not. LD1 manages to capture the class-discriminatory information in the data set and LD2 does not. With LDA the data must be normally distributed and co-variance between classes is assumed.

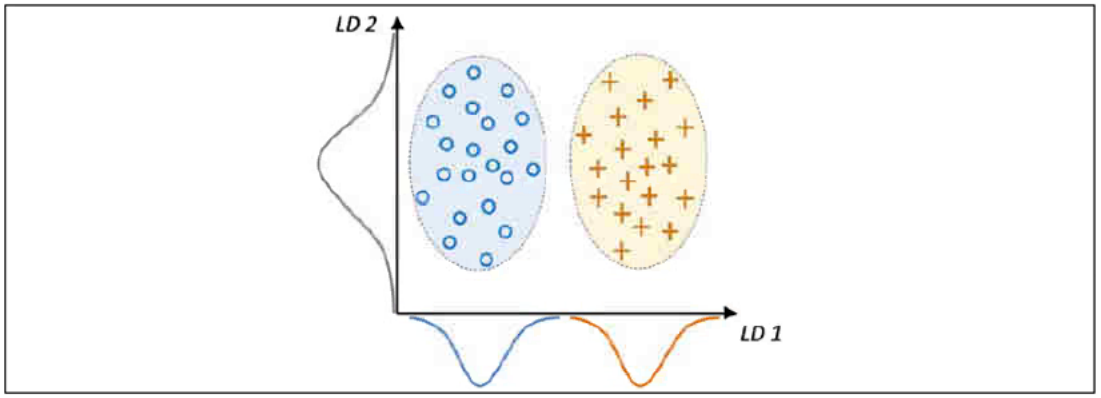

Figure 3.17: : Visual depiction of LDA (Raschka 2015)

The third feature extraction method was KPCA, which is used to transform data that is not linear.

Non-linear data is depicted in in **Figure 3.18**. Data non-linearity is encountered a lot in real world situations. KPCA employs the concept of kernel SVM (Raschka 2015).

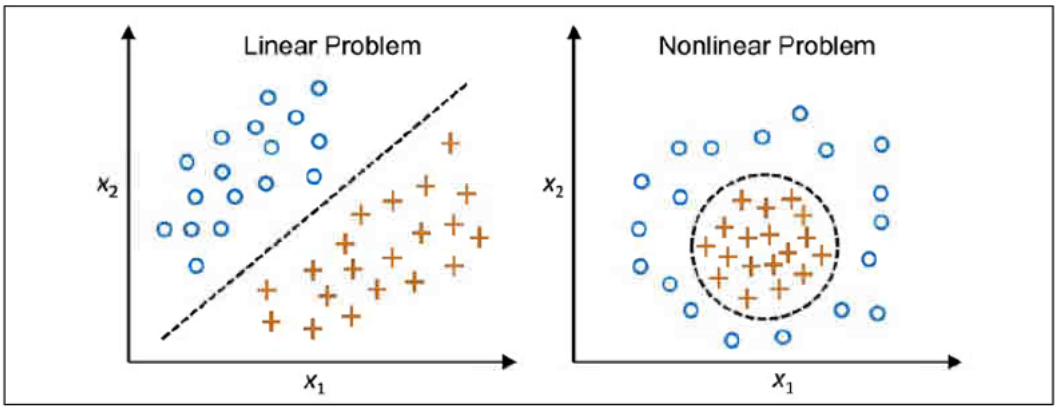

Figure 3.18: Visual depiction of data that can be separated with a linear solution and data that can be separated by a non-linear solution (Raschka 2015)

The data reduction steps, such as discussed in this section, can be run first, followed by the fitting of data in a model, or the two steps can be combined in a pipeline see **Figure 3.19**.

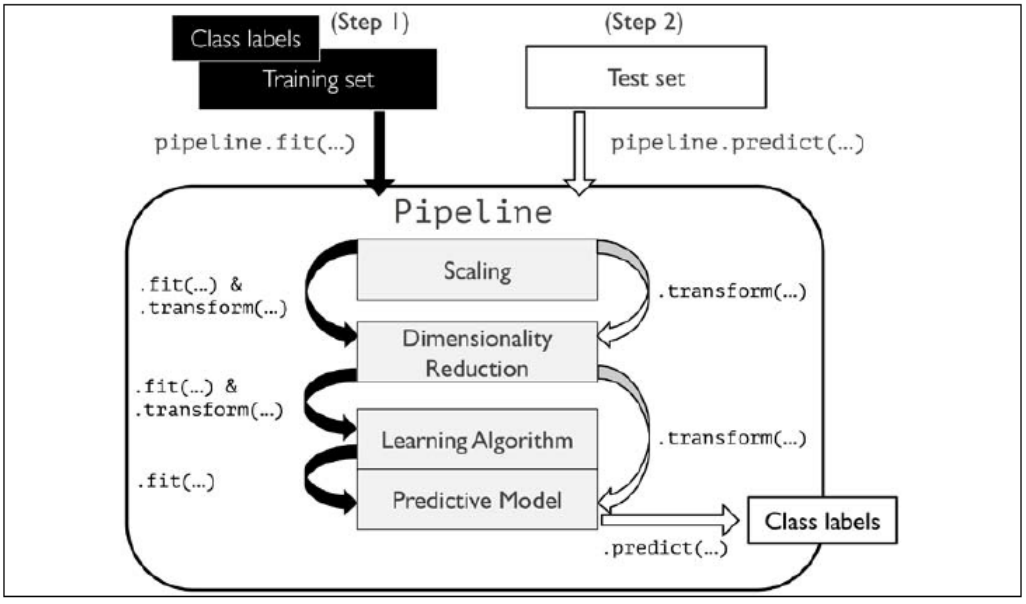

Figure 3.19: Pipeline combining data preparation and modelling steps (Raschka 2015)

3.6 MODELLING

There are various programming languages that can be used as a tool in data science and ML. One of the most widely used programming language is Python, which can be used to perform various functions related to data science and ML. There are various software packages that can be loaded into Python to perform various tasks described in **Section 5.6** as well is this section of the chapter. Some of these software packages are listed below (Klosterman 2019):

- Pandas is used to load, clean and explore data.
- Numpy arranges the data table into matrixes that can be used for mathematical calculations.
- Scipy is used for linear regression and programming.
- StatsModels is used for statistical analysis and also for time series analysis.
- Matplotlib and Seaborn is used for data visualization.
- Scikit-Learn is used for predictive analytics with ML algorithms.
- TensorFlow, Keras, and PyTorch are all used for deep learning.

Klosterman (2019) says that to decide which ML model must be used for every unique objective, it is important to determine how appropriate the model will be to reach the desired outcome. The desired outcome should answer a business question that will ultimately be beneficial to the business, usually in monetary terms.

According to Bowles (2019) the goal of building a predictive model is to obtain the best performance from that model to predict the response you are interested in. The more complicated the problem is and the wider the data sets are the more complicated model will need to be employed.

In simple terms a model will attempt to utilize a set of predictors $(x_{11} \cdots x_{mn})$ to predict a target $(y_{1 \cdots m})$;

$$y_i \sim pred(x_i)$$

A good performing model will generate an y_i that is acceptably close to the true y value. For regression problems, the mean absolute error (MAE) and the root mean squared error (RMSE) are used to measure the performance of a ML model.

MAE calculates the average amount that each predicted value differs from the real value. RMSE calculates the average standard deviations of the residuals of the model, where residuals are the size of the error.

The MAE can be calculated as follow:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Where y_i is the prediction, x_i is the true value and n is the number of samples.

The RSME can be calculated as follow:

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

If the model is a classification model, then the mis-classification error of the model can be used as a performance measurement. A good way to test the accuracy of a classification model, or a logistic regression model, is with a confusion matrix. A confusion matrix is a matrix that displays the number of the true positive, true negative , false positive , and false negative predictions of a classifier, **Figure 3.20**.

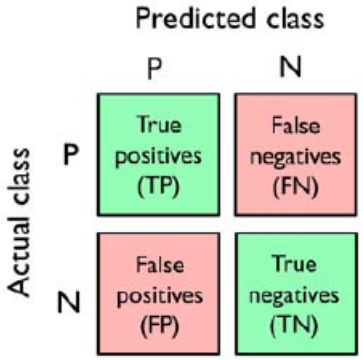

Figure 3.20: Example of a confusion matrix (Raschka 2015)

The accuracy of a ML model is measured on a simulated performance of that model, by testing the model on a set of data that does not include the data that was used to train the model. To do this, data is split into a training set and a test set. The training set will be the the majority of the data and is used to fit the model. The test set is the remaining data and is used to determine the performance of the model without including any of the data that was used to fit the model (Bowles 2019). If data sets are very big, the training data set can be up to 99% of all the data, since 1% of a data table with 1 000 000 rows will still be 10 000 rows of data to test the predictions on.

3.6.1 Linear regression

3.6.1.1 Simple linear regression

According to Subasi (2020) simple linear regression models in ML are very useful because they are relatively simple and stable. Simple linear models are not prone to over-fitting, but is susceptible towards under-fitting of data. Simple linear regression models are suitable when there are not large amounts of data.

Simple linear regression is regression between one factor and one response. the equation of such a model will be that of a straight line;

$$y = w_0 + w_1 x$$

Where w_0 is the weight of the dependent variable or response variable and w_1 is the weight of the independent variable or factor. The linear regression model will aim to predict new responses from the relationship of w_0 and w_1. The straight line is called the regression line and the distance off each true data point to the regression line is the error in prediction for that point.

3.6.1.2 Multiple linear regression

A multiple linear regression model can be represented as a column vector of responses (the dependent variable),

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

and as a matrix of factors (the independent variables),

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

and as a column vector of model coefficients,

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$$

where the prediction of $y_i = x_{i1} * \beta_1 + x_{i2} * \beta_2 + \cdots + x_{im} * \beta_m + \beta_0$ (Bowles 2019).

3.6.1.3 Penalized Linear Regression

Penalized linear regression has been derived from the ordinary least squares regression method, but with the aim to solve the problem of over-fitting of data. Penalized linear regression algorithms systematically reduce the complexity of the data by reducing the dimensions of the data. Penalized linear regression is specifically good to use when the data has very high degrees of freedom (Bowles 2019).

Penalized linear regression aims to find the coefficient values ($\beta's$) that give the smallest error. Ridge regression is an example of penalized regression which applies a penalty (λ) to all the coefficients in the regression model.

The concept behind penalized linear regression is to introduce additional bias to to all the factors in the model. This additional bias will penalize those factor weights ($w's$) that are most extreme (Raschka 2015);

$$\frac{\lambda}{2}\|W\|^2 = \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2$$

The Ridge regression model then solves a minimization problem for various $\lambda's$. Some of the other penalized linear regression algorithms are the LASSO regression, the least angle regression and the glmnet regression methods.

3.6.2 Ensemble regression

According to Bowles (2019), ensemble methods for regression performs better than single methods for regression when the ensemble methods have a measure of independence between them. The ensemble method uses a single ML algorithm as base learner and a second tier of algorithm that control the inputs of data into the base learner algorithm in such a way that the models that are generated are all somewhat independent. The upper level algorithms are known as either bagging, boosting or random forest and the base learners are normally binary decision trees.

Subasi (2020) says that ensemble methods are normally the most effective in ML, but they are also more complex and that leads to higher computational costs associated with the ensemble methods.

According to Klosterman (2019) some advantages of ensemble regression methods are:

- that they can handle complex data sets.
- that there is no need for scaling of the data since the decision tree node splitting algorithm considers each factor as a separate entity not related to the other nodes.
- that they can describe non-linear data sets.

3.6.2.1 Bagging methods

Binary decision trees can be over-fitted when the tree depth has too many levels because that will cause too many decision nodes. Binary decision trees are also prone to high variance in performance. This high variance in performance can be overcome by combining many of these binary decision tree models by a bagging algorithm. Bagging takes various randomly generated sets of data from the training data set and train the base learner on these sets. In a regression problem, the bagging model aggregates the outcomes from all the base learner models, and in a classification problem the bagging model determines the probabilities for each class (Bowles 2019). Random forest regression is a form of a bagging regression method, but subsets are drawn randomly from the attributes of the training data set.

3.6.2.2 Boosting methods

Gradient boosting trains each binary decision tree in an ensemble on different labels in the data table. The gradient boosting model refines the predictions of the response by recalculating the residuals after each new step in the ensemble regression algorithm.

Subasi (2020) explains that boosting works on the principal of identifying the mis-classified instances during the first step of classification, and giving these instances a higher weight during the next steps of the algorithm.

3.7 CONCLUSION

In Chapter 3 of this thesis machine learning is reviewed. The CRISP-DM framework used in data science projects is described. Ways in which the steps in the CRISP-DM framework can be approached during its application is discussed shortly. The possible use of ML in manufacturing and how the use of ML fits into the CRISP-DM model is described. ML is described in more detail by showing the need for ML and listing and reviewing the various types of ML.

The steps related to data and modelling when using ML in process improvement are described. Data cleaning, data exploration, data assessing and data preparation are described in detail. Various tools and the statistics involved with these data related steps in ML is reviewed. Some of the ML algorithms that can be used in predictive analytics is described, inclusive of their general uses and the basic statistical principles related to them.

CHAPTER 4

FRAMEWORK DESIGN

4.1 INTRODUCTION

In this chapter, a general framework for process improvement in metal packaging manufacturing will be discussed. The framework will include steps used in the general DMAIC approach from Six Sigma, as well as the CRISPM-DM framework used in data science.

Process improvement can be executed in different ways in the metal packaging manufacturing industry. There is no one-fits-all solution in manufacturing. The type of problem and the manufacturing environment are important factors to consider what type of process improvement framework should be applied. The general framework that this chapter expands on will be focused on process improvement with the aid of machine learning.

4.2 GENERAL FRAMEWORK

4.2.1 Basic framework following the lean six sigma DMAIC and CRISP-DM approach

Figure 4.1 summarizes the main steps in the framework for process improvement in metal manufacturing. The framework consists of five main phases; Define, Understand, Model, Evaluate and Deploy (DUMED). The DUMED framework is a new framework and is based on the DMAIC and CRISP-DM frameworks.

Figure 4.1: : Basic framework for process improvement in metal packaging manufacturing using data science - DUMED framework

Each of the five steps can have different sub-steps and each sub-step can utilize different tools. Using the basic DUMED framework in **Figure 4.1** as guideline and combining various steps used in the referenced frameworks from the literature study the basic framework with associated sub steps has been compiled as seen in **Table 4.1**.

Table 4.1: DUMED framework for process improvement in metal packaging manufacturing in relation with DMAIC and CRISP-DM

Primary Stage	Primary Stage Sub-Divided
Define	Determine the Objective
	Rationale to solve the Problem
	Collect and Describe the Data
Understand	Understand the Process
	Asses Data Quality
	Prepare Data for Analysis
Model	Build Various Data Models
	Assess and Improve the Data Models
Evaluate	Evaluate the Models' Outcomes
	Determine Business Improvements
Deploy	Implement the New Process
	Control the New Process

4.2.2 Steps in DUMED framework

The implementation of the DUMED process improvement framework as described in the previous sections can be triggered by marketing, sales, technical, quality, production research and development or management departments. Process improvement frameworks can be used:

- To solve a problem of known origin
- To solve a problem of unknown origin
- To test the suitability of different raw materials
- To test the suitability of a new piece of equipment
- To change a process
- To develop a new product
- To continuously improve a process

Exactly how the steps in the process improvement are implemented, and which tools and methods are used may largely depend on the objective of the process improvement.

4.2.2.1 Objective

The first step of the DUMED process improvement framework is to state the objective for process improvement. According to Aqlan and Al-Fandi (2018), process improvement in a manufacturing environment has the objective of improving the quality of your product or to reduce waste, both of these outcomes will result in an improved business process.

The objective will consist of a short description on which part or aspect of the process that will be improved, and how this improvement might be accomplished. From the objective it should be clear which attributes in the process needs to be controlled, improved, or predicted. These attributes is the responses of your process improvement project. Responses of the project should be outputs that are measurable. A response is normally a quality characteristic and can be variables such as temperature, length and time; or attributes such as pass/fail, yes/no and good/bad. Variables will need fewer samples to be statistically significant when compared to attributes as responses.

4.2.2.2 Quantify the problem

Once the objective of a project is understood, a high level business case for the project can be outlined. The business case may include the possible financial advantages the process improvement project will generate. With reference to a short list of problem solving methods a problem can be quantified as follow to justify the business case for the solving of the problem (Mauri, Garetti, and Gandelli 2010):

- If your objective is to improve quality – set a goal for and quantify the increased yield
- If your objective is to improve speed – set a goal for and quantify the increased velocity performance
- If your objective is to improve procedures – set a goal for and quantify the increased organizational efficiency
- If your objective is to decrease downtime by improved maintenance – set a goal for and quantify the decrease in downtime and maintenance setups.
- If your objective is to improve bottlenecks – set a goal for and quantify the throughput variability reductions and improved scheduling.
- If your objective is to improve utilization – set a goal for and quantify the improved line usage

4.2.2.3 Data collection and description

To determine the data conditions it will be necessary to collect data and do initial data description. According to Subasi (2020) data collection will entail the extraction of available data from historical records, websites, instrument historical log files, surveys, experiments, simulations, or any data from

any other sources of data. It may also be necessary to collect data by initiating a data capturing regime if some or all of the necessary data is not available for collection.

If a measurement regime is set up during the data collection phase of the process improvement project, it should be clearly defined so that everyone knows what to measure, when to measure, where to measure and who is doing the measurement.

Once the data collection has been accomplished it will be necessary to describe the data:

- Describe the sources from where the data was obtained.
- Expand on the data attributes such as variables, units and volume of data.
- Perform basic statistics on the data.

Data collection and description can follow the same tasks as in phase 2 of the CRISP-DM model;

- Gather the data:
  - List the type of data that is needed, the time ranges you need the data for and the format the data should be in.
  - Determine if the specified data is available, and decide how the unavailability of data will be addressed.
  - Define the sources the data will be obtained from, and specify within those sources where the data can be obtained.
  - Obtain the data by importing the data into the platform being used for data analysis.
- Describe the data:
  - Describe the different formats of the data.
  - Describe the suitability of the data for the end goal.
- Explore the data:
  - Do basic statistical analysis on the data.
  - Check the normality of the data.
  - Describe any quality issues with the data.

- Verify the data quality;
  - Make sure the required data exists.
  - Make sure the required data is accessible.
  - Make sure the required data does not have many missing values or incorrect values.
- Select the data:
  - Exclude all data that is not in the correct format, or that has technical issues.
  - Exclude all data that is not relevant to the process improvement goal.
  - Exclude all data that is of poor quality.

4.2.2.4 Understand the process

To understand the significance of the data in a process, it is necessary to understand the process from which the data was collected. The process can be described with process flows that indicate the various steps and sub-processes in your process, and how these steps interrelate to each other and to the complete process. Important factors in the process can be researched and described in more detail. The process flow gives a good overview of the steps in the process which are being looked at in the form of a flowchart. The basic general system schematic that is used in the representation of a manufacturing process is given by **Figure 2.4** in **Section 2.3.3.1** (Cameron and Hangos 2001). The process flow illustrates the processes in a system. The model can represent a single step or unit in a process, or a section in the process or the whole process.

Mauri, Garetti, and Gandelli (2010) suggests that the system that needs to be improved must be modelled. The model should distinguish between active and passive processes and, if possible, superimpose the sequential steps of the process onto a layout of the factory floor. Active steps are steps that add value to the product such as physical changes or tests. Passive steps are the steps that do not add value such as buffering or queuing.

4.2.2.5 Assess data quality

Once the raw data has been collected and it is understood how this data relates to the process the data can be explored an assessed. From (Subasi 2020):

- Firstly, the data quality should be assessed by looking for any errors in the data like missing values, inconsistent values and erroneous data.
- Secondly, ensure the data-set is in a format that can be used for the modelling phase e.g. some

algorithms cannot work when there are strings in the data. Some algorithms will not distinguish between categorical numbers and numbers as a range in magnitude. Data-sets that needs to be combined should be combined on a mutual attribute or on time stamp values for uniformity.

- Thirdly, once obvious errors and missing values have been addressed and your data-set is in the shape and format it needs to be, explore the data by visualizing the data attributes to find relationships and associations in the data.

According to Snee (2010a) data are analyzed by visualization with tools such as control charts, process capability indices, ANOVA, time plots, boxplots and histograms.

4.2.2.6 Prepare data

According to Subasi (2020) in this stage of the process the data is prepared for the algorithms they will be subjected to:

- New attributes can be calculated from the existing attributes in the data table by performing mathematical calculations on existing attributes, or by dividing a categorical attribute into new attributes for each category.

- Data is also scaled or normalized, dependent on which ML algorithms the data will be subjected to.

- Data reduction can also be accomplished by feature selection and dimension reduction. Feature selection is the process of selecting only a subset of attributes for the modelling phase. Dimension reduction is the removal of unimportant information from a vector and only continuing with the most distinctive information of the vector.

4.2.2.7 Model data

During the modelling stage (Subasi 2020), the data from the previous stages are fed to ML algorithms to extract meaningful information from the data. Good practice will be to create various iterative models and select the model that gives the best results. The ML algorithms that will be used in model building is dependent on factors such as the type of data, the desired outcomes and constraints from your process data as well as constraints from the algorithms. Model building essentially combine the attributes from your data table with ML techniques to obtain the most useful information from your input data to gain insight into the process or make predictions.

4.2.2.8 Assess model

Models are assessed by firstly determining and comparing the accuracies of all the different models. Accuracy can be expressed and calculated in different manners, but generally the information gained from the model is compared to real process outcomes. The accuracy of the model's outcome will

determine whether a model is acceptable for the process improvement goal or whether the model needs to be first improved to get more accurate results. The model can be improved by fine tuning some parameters in the ML algorithms, or the model can be improved by improving the amount, accuracy or relevancy of the data that is used in the model (Subasi 2020).

4.2.2.9 Make conclusions from model results

Evaluate the conclusions made from the ML model and the possible effect these learnings can have on the process. Points that can be evaluated after the ML models have been completed and the most effective model has been chosen:

- Evaluate the reproducibility of the model
- Evaluate how easily findings from the model can be implemented in the process
- Evaluate how the final models can possible be extended or evolved further
- Evaluate the limitations and constraints of the model

4.2.2.10 Evaluate the viability for implementation

Once the model is found to be advantageous for process improvement from a technical and practical point of view, the financial cost of implementation should be evaluated. Considerations should be given to the costs for data extraction, data storage and resources used for computational steps during the running of the model. Costs associated with manpower and expertize can also have an effect on the financial evaluation of such a project. Finally costs related to the software engineering facet of implementing such a project should also be considered.

4.2.2.11 Evaluate the business improvements

Once the viability of implementation of process changes have been established the business advantages related to these changes can be determined.

4.2.2.12 Deployment

Implement and control the process changes. Since the process improvement pathways that use data science and ML as tools are iterative, the process can be reviewed and repeated as necessary in order to continuously improve the outcomes. Implement process controls that involve empirical verification of responses to sustain the improvements. Use tools such as SPC, audits, FMEA. Added control of processes can be achieved by assigning rolls, tasks, frequencies and methods to the adjusted process.

4.3 CONCLUSION

In this chapter of the thesis, a framework was developed. The DUMED framework was based on the DMAIC steps that is used for the Six Sigma process improvement methodology as well as the CRISP-DM steps that is used in data science. The DUMED framework consisted of five main phases, that incorporates 12 steps. DUMED stands for define, understand, model, evaluate and deploy. The 12 steps of the DUMED framework was discussed and explanations was given to when to apply each step and possible what these steps would entail. In the next chapter, Chapter 5, the steps that were followed in the case study is discussed in detail.

CHAPTER 5

CASE STUDY

5.1 INTRODUCTION

In this chapter of the thesis the case study is described in detail. The case study demonstrates the application of the DUMED framework that was described in the previous chapter. The DUMED framework is one of the objectives that was stated to answer the research question of how data science principles such as ML can be used for process improvement in metal packaging manufacturing. The case study demonstrates the use of this framework in process improvement of a 2-piece metal food can manufacturing line. Each of the 12 steps in the DUMED framework is described comprehensively in relation to what was done for the case study or how the fulfillment of these steps could still be accomplished in the case study.

5.2 BACKGROUND

A process improvement project was initiated at a manufacturing plant of Nampak Ltd. Nampak is Africa's biggest packaging company and has various manufacturing sites across South Africa as well as into the rest of Africa. The plant where the process improvement project was initiated was the 2-piece tinplate food can manufacturing site in Epping, Cape Town.

In the initial phase of the project the aim of the project was to use data from the manufacturing process to control quality parameters, such as factory finished can height (FFCH) better. The reasoning behind this project was to consistently produce cans within the quality specifications and to improve on the process capabilities of finished can quality parameters. Previous audit findings did suggest that process capabilities should be improved for some of the finished can quality parameters.

After consultations with several of the technical, quality and production personnel, a data collection regime was initiated. Measurements related to a wide range of factors were recorded over a period of time. The factors numbered almost a hundred and included quality data on the product throughout the process, data related to pressures and temperatures of the manufacturing instruments and tools as well as basic raw material data.

The data was collated into a data table and the data was cleaned to prepare it for further analysis. Further analysis mainly consisted of data science methods related to machine learning to build a predictive model of the process.

Predictions were made after applying the random forest regression algorithm on the input data. The results showed the factors that have the biggest impact on the outcome we wanted to predict as well as the error associated with that prediction.

The outcomes after this phase of the project were:

- The regression model used to predict the FFCH deviation from nominal has an average error of 0.029mm; this means on average the model predicts the FFCH 0.029mm from the true value per can. The graph below shows the actual vs. predicted values.

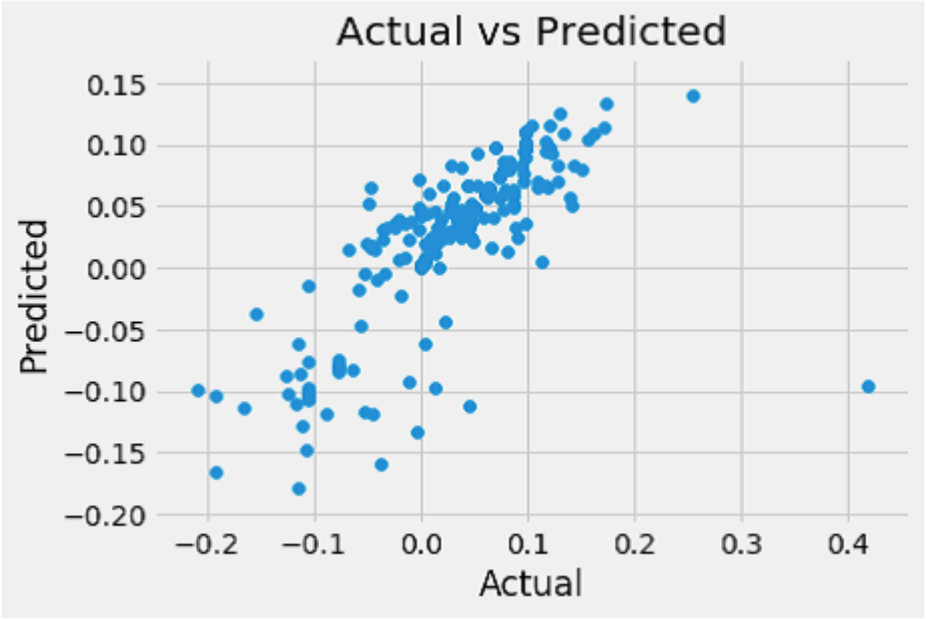

Figure 5.1: : Predicted versus actual deviation from nominal FFCH of manufactured 2-piece tinplate cans

- The factors that were the most influential to predict the FFCH deviation from nominal were listed from highest to lowest. Factors related to quality outcomes throughout the first and last parts of the process as well as the raw materials were the factors that had the biggest impact on the response.

After presenting the results from the project's analysis a brainstorming session was held between various technical persons employed at the R&D section and the manufacturing section of Nampak Ltd. The results did show potential to be used for predictive analytics in the manufacturing process, but it was decided to change the objective of the project. To predict and ultimately improve the control of the FFCH of the manufactured cans will assist in audit results, but not necessarily in better yields since the cans generally are manufactured within the specification standards for FFCH and the plant does not experience high volumes of customer complaints related to FFCH.

The first phase of the project, as described in the above paragraphs, acted as a proof of concept to the manufacturing plant for process improvement with the aid of ML and predictive analytics. After

the brainstorming session it was decided that a good way to use this model will be to predict quality characteristics of the 2-piece metal cans such as the axial load resistance and panelling resistance of current cans. It is generally understood that the bead depths of the cans and the material properties, such as gauge and supplier, probably affect the the axial load resistance and panelling resistance of the manufactured cans. What is unknown is to what extend these factors and other unknown factors can influence the quality characteristics, such as the axial load resistance and panelling resistance. The ability to predict the axial load resistance and panelling resistance from those factors that has the biggest effects on them allows for better quality control and ultimately less waste and as a direct result financial benefits.

A second potential outcome of such a predictive model is the systematic decreasing of the gauge of the can walls, but still retain acceptable axial load resistance and panelling resistance. This will be possible if the predictive models are updated and trained continuously with data from different gauges of material. Ultimately the possibility of using thinner material can result in large savings for the manufacturing plant.

Building such a ML model, and practically demonstrating the usefulness of such a model in the production environment, will be the goal of phase 2 of this project. Phase 2 will involve developing a whole pipeline from data acquisition to model outputs in order to practically demonstrate how we can use ML models in the production process. If successful, a phase three of the project will include the implementation that may include the software engineering aspects associated with the deployment of such a project.

This case study is a description of phase 2 of the process improvement project on a 2-piece metal can manufacturing line. The steps of the process improvement project using principles of data science, as described in **Figure 4.1** and **Table 4.1**, will be described in the rest of this case study chapter. The case study demonstrates the use of the framework described in **Section 4.2** for the prediction of axial load resistance of 2-piece metal food cans. Refer to **Appendix A** for a similar approach towards prediction of the panelling pressure resistance of 2-piece metal food cans.

5.3 OBJECTIVE

The objective of the process improvement project, as described in this case study, is to predict the axial load resistance of metal 2-piece food cans in a manufacturing process. The advantage of understanding the relationship between important attributes in the process and the axial load resistance of the manufactured food can relates to possible financial benefits as well as better control of quality parameters. As part of the objective, the aim is to understand how axial load resistance relates to other factors in the process with the outcome that any changes made in the process will still deliver cans with suitable axial load resistance. A predictive model on axial load resistance will not only give enhanced capability to control quality parameters throughout the process and in the final manufactured product, but can also supply valuable information on the possible viability for light weighting of material. The outcome of the expanded model will be to control the important factors in the manufac-

turing process to be able to keep the axial load resistance of the 2-piece food cans within specifications across different tinplate thickness gauges.

The axial load resistance is measured twice per shift on 16 manufactured food cans. The measurements are performed on a gauge that stores the results and automatically send the results to the Datalyzer program. The gauge is called the Trac Gauge and this gauge performs various dimensional and strength measurements on the 2-piece food cans throughout the manufacturing process. Datalyzer is a quality tool where various quality measurements are stored. Datalyzer not only provide an interface where historical quality data can be accessed, but also is capable of providing quality information such as process capabilities.

The axial load resistance is a dependent variable and the output is given as real numbers. The axial load resistance is measured in Newtons, with a specified range of 165N to 235N. Generally the axial load resistance is within these specifications and generally there are no problems in terms of process capabilities for this quality measurement. The objective is therefore not focused on improving axial load resistance, but rather to understand how axial load resistance relates to other factors in the process with the aim that any changes made in process improvement will still deliver cans with suitable axial load resistance.

The specified range for axial load resistance for a 2-piece metal food can is 70N. If the range is represented as a six sigma range or six standard deviations within the normal distribution with a specified alpha value, then a standard deviation will be 11.67N for such a normal distribution as illustrated in **Figure 5.2**.

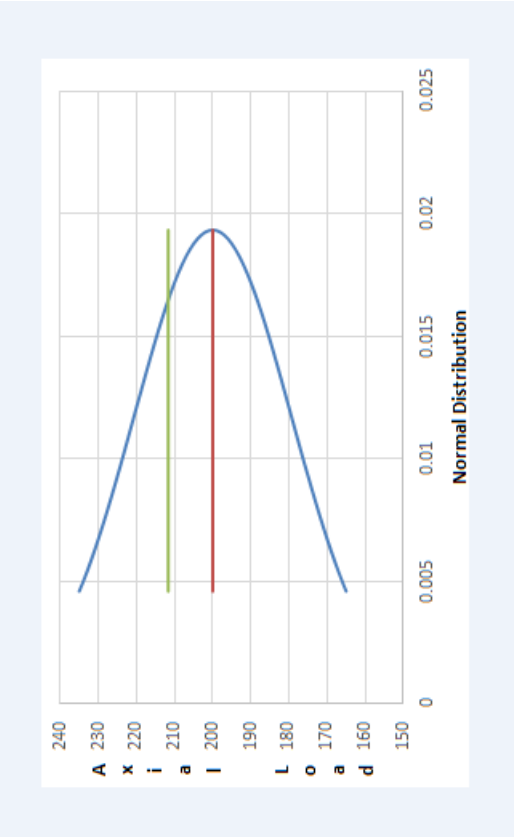

Figure 5.2: : Standard deviation of the normal distribution of axial load resistance of 2-piece metal cans.

5.4 RATIONALE

In this case study the complex relationships between various factors in the process and the response variable of axial load resistance of 2-piece food cans will be described in a model. Knowledge of how factors in the process influence the response can supply valuable information on the possible viability for light weighting of material. A further benefit of being able to understand how factors in the process influence the response is enhanced capability to control quality parameters throughout the process and in the final manufactured product.

The ability to better understand the relationship between the axial load resistance, and various factors in the manufacturing process for 2-piece metal food cans, can have a financial benefit for the manufacturer. The possible financial benefit can be derived from light weighting of the metal cans by decreasing the incoming tinplate thickness gauge, therefore saving on raw materials. Currently the incoming material is tinplate sheet wound in a coil, the coils weigh about 9 metric tonnes each. The thickness gauge of the tinplate is 0.29mm currently. The potential savings of a decrease of 0.01mm thickness gauge per metric tonne is about 30 US dollars. On an average day there are more than a 1

000 000 cans produced. For every cent saved on manufacturing costs per can, the savings per day will translate to about R10 000.

To consider light weighting of 2-piece food cans it is important to understand the complex relationship between factors related to the raw material, factors related to various steps in the manufacturing process as well as uncontrollable factors. Light weighting of food cans can only be considered if all the quality specifications of the food cans can still be adhered to.

After completion of this case study, there will be a predictive model that can predict the axial load resistance for 2-piece metal food cans. In further work, beyond the scope of this case study, the model will be expanded to include data from down-gauged tinplate. The outcome of the expanded model will be to control the important factors in the manufacturing process to be able to keep the axial load resistance of the 2-piece food cans within specifications across different tinplate thickness gauges.

A further consideration that will fall outside of the scope of this case study is to determine whether light weighting can be achieved with our current process, or if it will be necessary to change some of the equipment. If light weighting cannot be achieved with the current equipment, the equipment or tooling that will need to be changed will be the beader rail. The approximate cost of replacing the beader rail will be about 35 000 British pounds. If a spare rail is included in the purchase the investment will add up to 70 000 British pounds.

5.5 MANUFACTURING PROCESS FLOW

The metal can manufacture process is described in **Chapter 2, section 2.2.4**. The 2 -piece metal can manufacturing process is a DWI process of which the process flow is displayed in **Figure 5.4**. There are various benefits in using 2-piece food cans instead of 3-piece food cans:

- There is no side seam or bottom seam on the cans. This results in a more visually appealing can, but more importantly, the 2-piece metal can is a can with less risk to food contamination through weaknesses on seams due to much less seam distance per can. **Figure 5.3** shows a section of a 2-piece can with a cross-section of the double seam at the top.

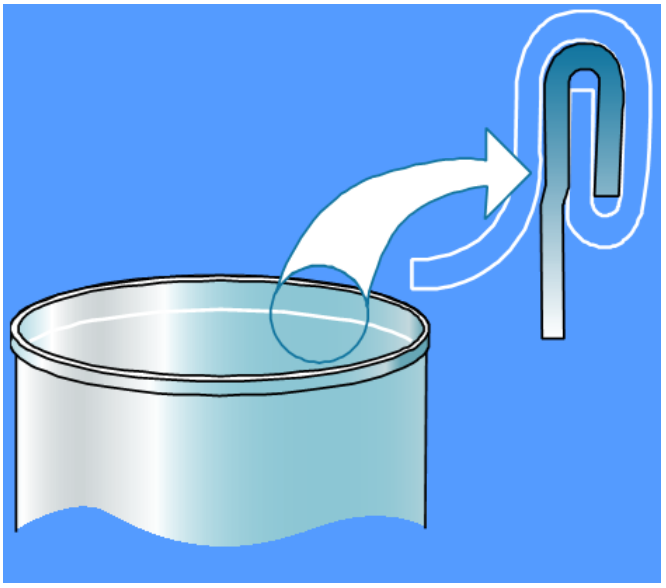

Figure 5.3: : 2-piece can with cross-section of the double seam

- Lacquer adhesion is excellent on these cans due to passivisation of the metal surface.
- Cost savings due to less metal usage in the manufacturing of 2-piece cans when compared to 3-piece cans.

For this case study, data was collected on three sub-sections of the 2-piece food can manufacturing; the front end, the flanger and the beader. The factors, on which data was collected as part of the process improvement project, were largely determined by the findings from the preparatory work that was done and that was described in **Section 5.2**.

The front end sub-section process flow is illustrated in **Figure 5.5**. The figure flows from right to left starting at the raw materials. Tinplate in coils are brought from the stores as per scheduled job numbers. A coil is loaded onto a coil handling station, from where the coil is continuously fed through a lubrication station. During lubrication very small amounts of oil is applied to the tinplate surface to assist with the ironing process later on in the process. From lubrication the tinplate is fed through the cupper press, where disks are cut out of the tinplate and those disks are formed into cups. The cups are then fed, via conveyor belts, into 1 of 5 body-makers, where the tinplate cups' walls are ironed out to form cylinders with one closed end as illustrated in **Figure 5.6**. Each body-maker has 3 trimmer heads which trim the cans to within a specified trimmed can height.

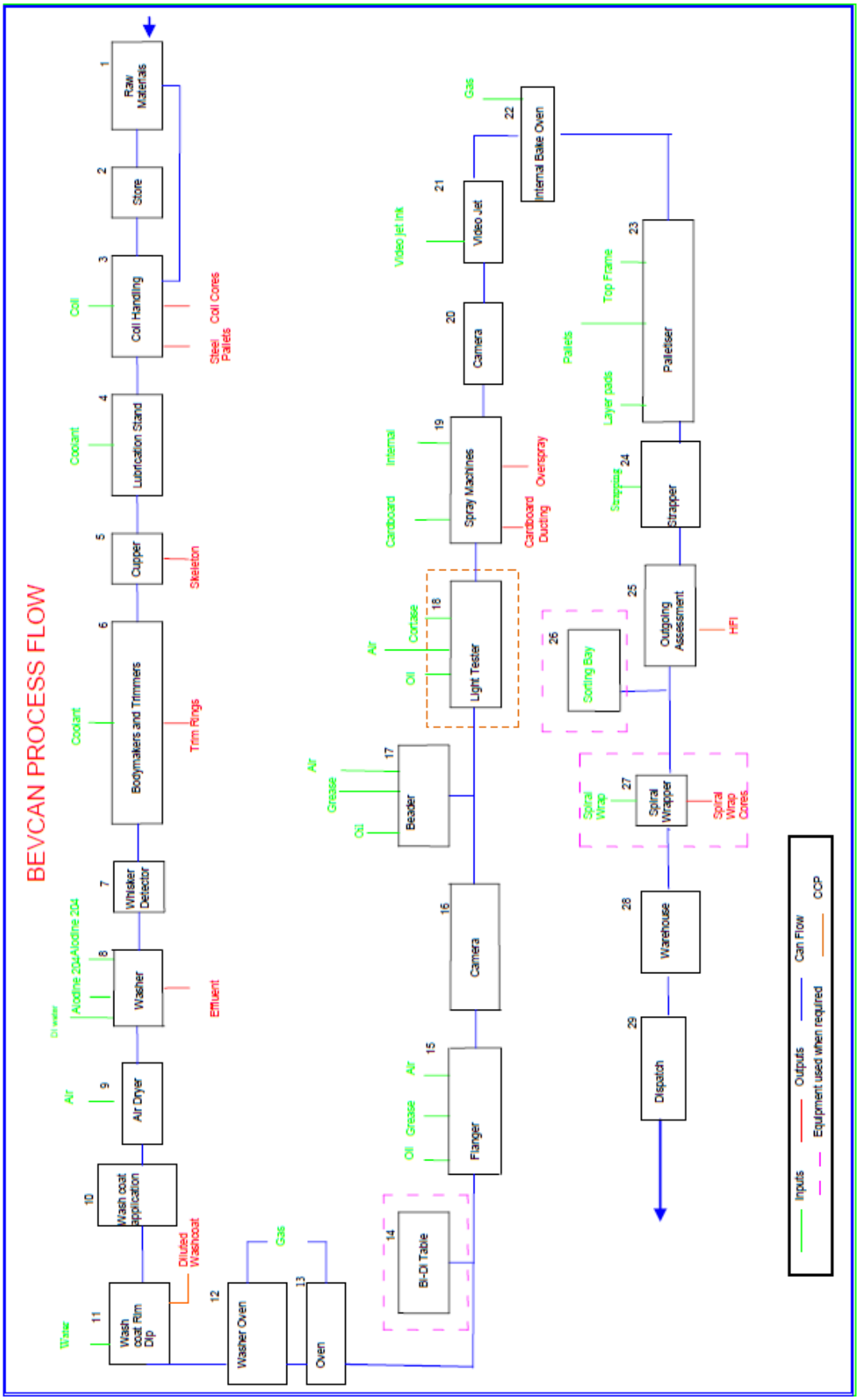

Figure 5.4: : 2-Piece metal food can process flow

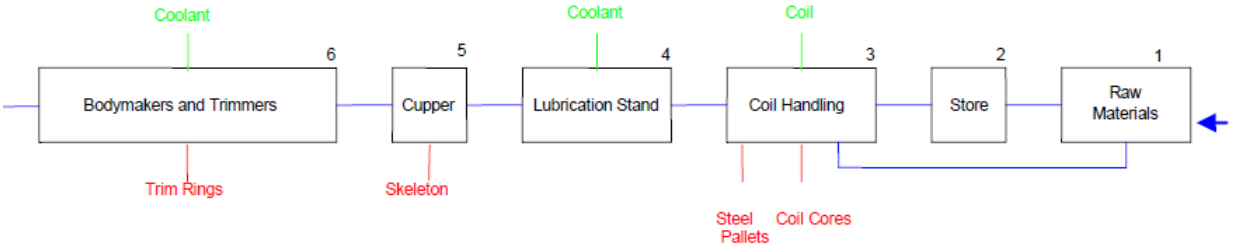

Figure 5.5: : 2-Piece metal can front end process flow

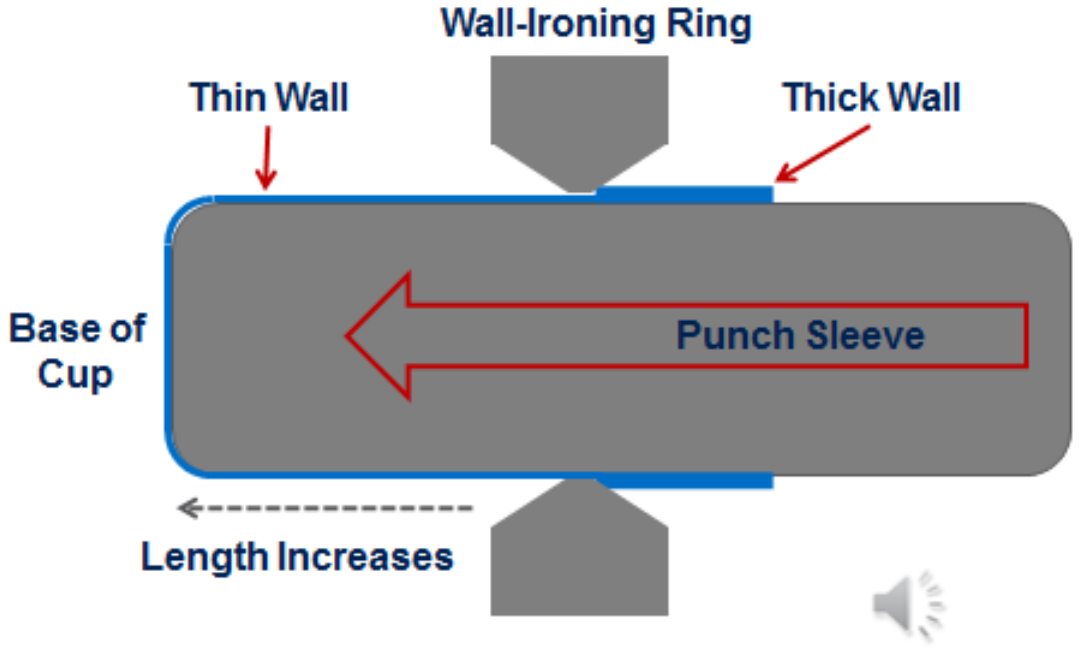

Figure 5.6: : Wall ironing of a 2-piece metal can body at a front end body-maker

The factors that were used for the case study in the front end sub-section of the process were the raw material supplier, the body-maker number, the trimming head number, as well as the quality data from the front end can measurements which include front end trimmed can height, front end mid-wall thickness, front end top-wall thickness and panel depth.

The flanger sub-section process flow is illustrated in **Figure 5.7**. After the trimmed cans exit the front end section of the manufacturing process the trimmed cans go through a washer and a dryer. The dried cans are then coated on the stand rim of the can and send through a curing oven. The rim coated and cured cans are then flanged. The flange on the open end of the can is necessary to be able to seam

and end onto the can after it has been filled with product at the customer. A suitable flange length is needed to ensure proper double seams are formed when the end is seamed onto the can body.

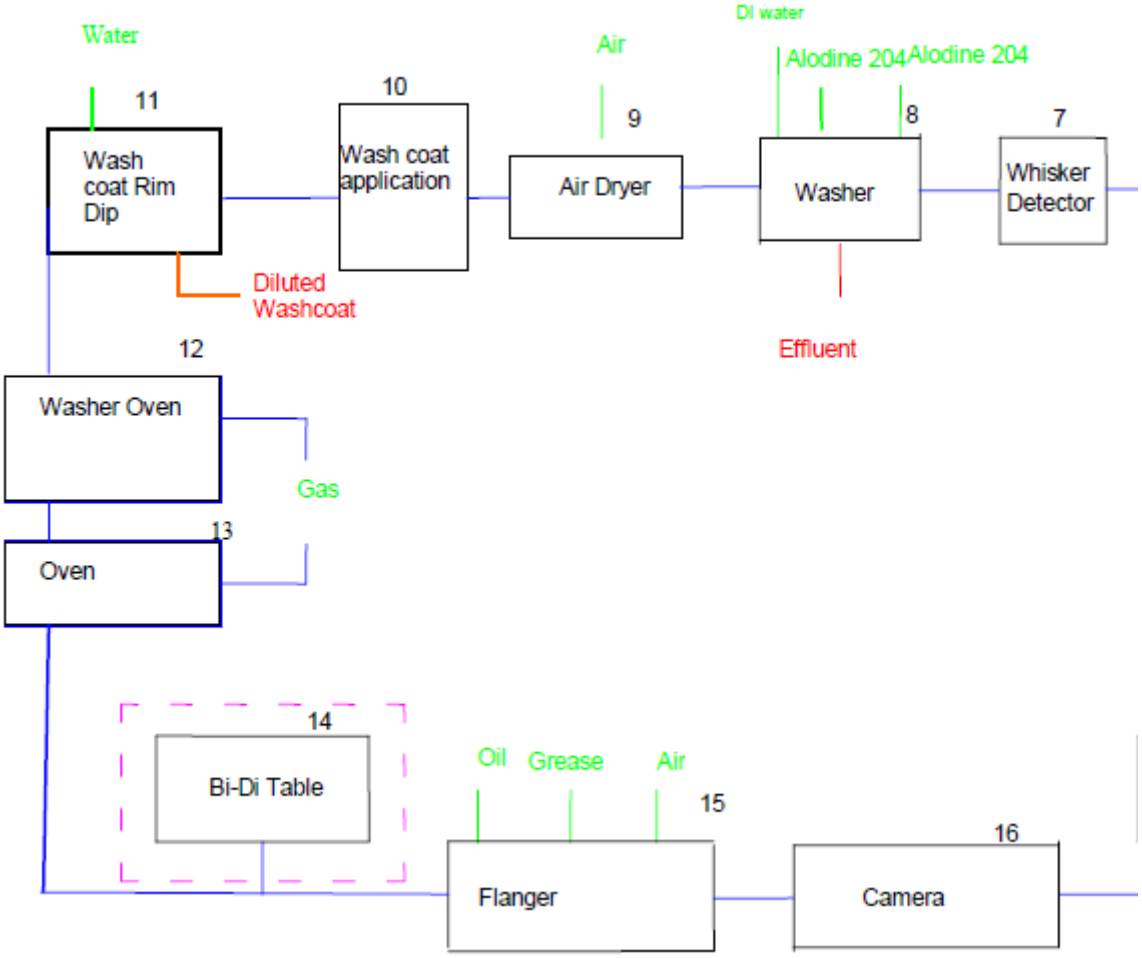

Figure 5.7: : 2-Piece metal can flanger process flow

During flanging the flanger heads spin separately whilst forcing the edge on the open side of the can into the required shape as illustrated in **Figure 5.8**. There are 10 different flanger heads in the flanger sub-section.

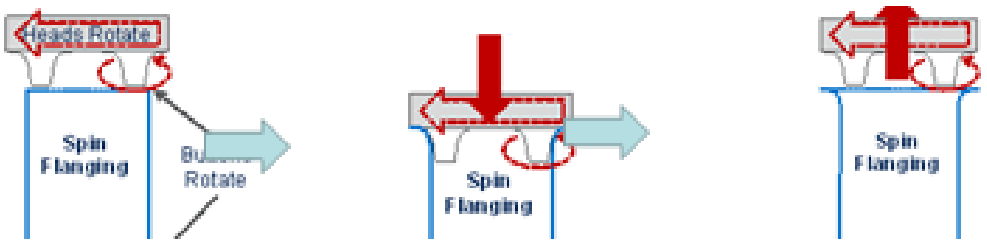

Figure 5.8: : Flanger illustration of flange formation on metal food cans

The factors that were used for the case study in the flanger sub-section of the process were the flanger head number as well as the quality data from the flanger can measurements which include flanger can height and flange width.

The beader sub-section process flow is illustrated in **Figure 5.9**. After the flanger, the cans are video inspected and any can with trim inside or with a dented body will be ejected from the conveyor into a scrap basket. From the camera inspection section the cans are transported via a conveyor belt to the beader. The cans are beaded and a roll-bead is also created. The roll-bead increases the strength of the lower wall of the can and also assists in the can handling. After the beader, the cans are spray-coated with an epoxy coating, inspected, marked for tracing and identification, cured and packed onto pallets with an automatised packing system.

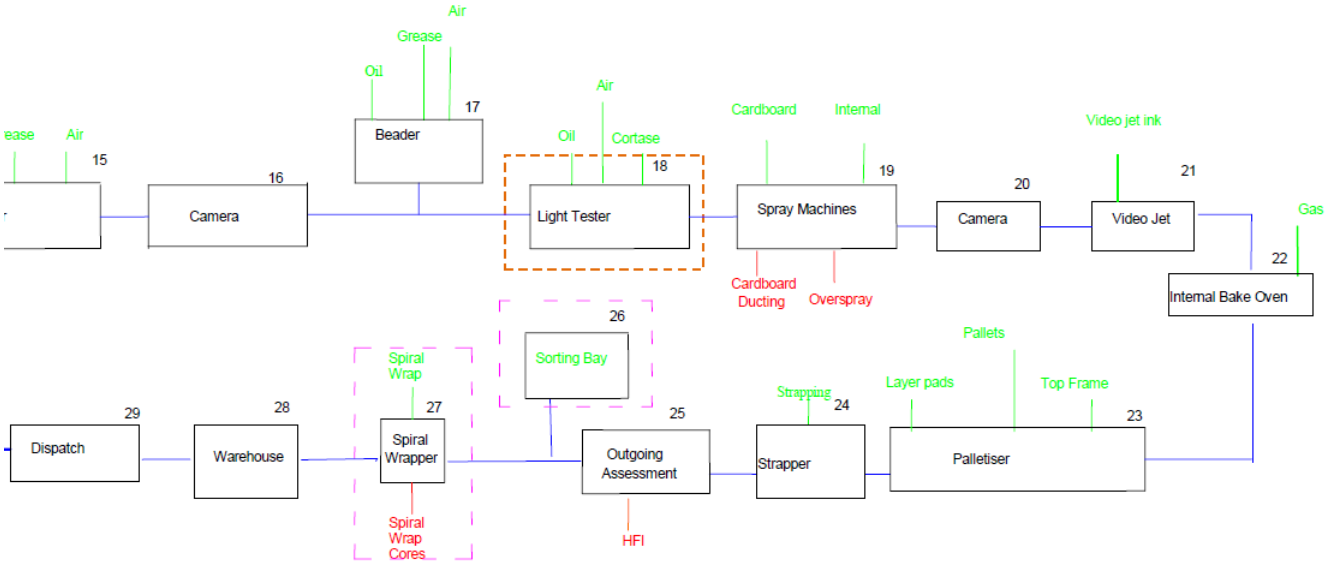

Figure 5.9: : 2-Piece metal can beader process flow

At the beader, the beading spindles enter the cans and roll them against a beading rail. The beads on the rail coincide with the hollows on the mandrels and generate the profile as shown in **Figure**

5.10. Beads strengthen the can's panelling pressure resistance, which helps to prevent the can from collapsing inwards, during food processing stages, once filled. Other advantages of beaded cans is that cans are more dent resistant and thinner tinplate can be used to manufacture the cans. There are 16 different beader mandrels in the beader sub-section.

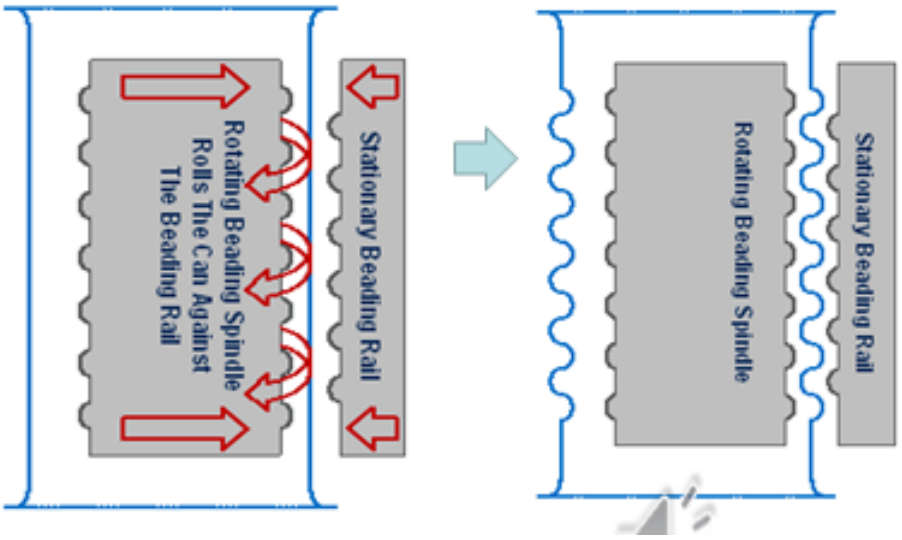

Figure 5.10: : Beader illustartion of beading of metal food cans

The factors that were used for the case study in the beader sub-section of the process were the beader mandrel number as well as the quality data from the beaded can measurements which include factory finished can height, beaded can flange width, roll-bead diameter, roll-bead position, bead symmetry and bead depths.

5.6 DATA

Refer to **Appendix B** for the Python code used throughout this section and the next section of the case study.

5.6.1 Data description

Data used for this case study consisted of quality data, raw material supplier and instrument identification numbers. The quality data used for the analysis are all physical measurements of dimensionality or strength during the three sub sections of the manufacturing process. **Section 5.2** looked at many factors of the 2-piece metal can manufacturing process. The feature selection algorithm employed during this proof of concept found that the quality data together with the raw material data had the biggest influence on the finished can height. Since the quality data is fairly easily obtainable and can

be transformed into an usable format, the data analysis used the quality data as major contributor of the data used in the machine learning models. Future iterations of the predictive analytics process can, if necessary, delve into more depth of the factors that may influence the quality outcomes e.g. ambient temperature, machine oil temperature, line speeds or chemical analyses of tinplate coils.

Except for the quality data that can be extracted from the Trac gauge, the only other data that was used for the data analysis was the categorical data of raw material suppliers as well as the different instrument identification numbers. The raw material supplier data can be obtained from log sheets as well as from a QlikView page. The instrument identifications, such as beader mandrel number, can also be obtained from the Trac gauge data.

Data was extracted from the Trac Gauge in a .dat format. From the .dat format the data was transferred to a text file and imported into a Microsoft Excel file. Data from 31 March 2021 to 08 July 2021 was saved into Excel in four data tables.

The first data table consisted of front end data, which were data from measurements on the cans after it exited the front end of the manufacturing process. The trimmed can heights, top wall thickness, mid wall thickness and the panel depth of the cans were measured. The cans were sampled from the different body-makers and trimmer heads. The data table consisted of 6298 row entries. Some row entries that had no entered values were deleted, leaving 5958 row entries. There were a few days where a different size can was run through the manufacturing process. The standard can is a 73mm by 110mm can size, but for the period from 21 April 2021 to 26 April 2021 jam cans were produced, which have a size of 73mm by 97mm. Since there are different specifications for some of these two cans' attributes, the jam can data was also deleted from the data table, leaving 5403 row entries.

The second data table consists of flanger data, which were data from measurements on the cans after it exited the flanger of the manufacturing process. The flanged can heights and the flange widths of the cans were measured. The cans were sampled from the different flangers. The data table consisted of 2446 row entries. Some row entries that had no entered values were deleted, leaving 2354 row entries. After deleting the jam can data from the data table, it left 2151 row entries.

The third data table consisted of beader data, which were data from measurements on the cans after it exited the beader of the manufacturing process. The beaded can heights, beader can flange widths, roll bead diameter, roll bead position, bead symmetry, bead depths and the axial load resistance of the cans were measured. The cans were sampled from the different beaders. The data table consisted of 3487 row entries. Some row entries that had no entered values were deleted, leaving 3379 row entries. After deleting the jam can data from the data table, it left 3363 row entries.

The fourth data table also consisted of beader data, but in this table the panelling pressure resistance of the cans were measured instead of the axial load resistance. The data table consisted of 3584 row entries. Some row entries that had no entered values were deleted, leaving 3344 row entries. After deleting the jam can data from the data table, it still left 3344 row entries.

Various data exploration methods were executed on the data of the four data tables. Things that were explored and changed were:

- Some of the data tables had many unnamed columns, which were also empty, these were deleted.
- The data was searched for any possible duplicate values, none were found.
- The data attributes' data types were searched. Some categorical attributes were integers because they were seen as ordered numbers e.g. body-maker numbers 1 to 5, but 5 is not larger than 1 in this case. Numbers 1 to 5 identifies which body-maker is referred to and therefore the data types were changed to objects for these attributes.
- Data was described by calculating the count, mean, standard deviation, minimum, 25% data interval, 50% data interval, 75% data interval and the maximum. This description of the data gives an idea of the spread of data, and if there might be any outliers. From **Figure 5.11**, **Figure 5.12**, and **Figure 5.13** it is indicative that there are some values that might be outliers, specifically the maximum values of some the measured factors. See **Appendix A.2.1** and **Figure A.1** for the beader descriptive statistical table of the beader data set that includes the panelling pressure resistance. The reason for the two sets of data over the same time period for the beader sub-section of the manufacturing process is that both axial load resistance as well as panelling pressure resistance are destructive tests and therefore it is not possible to perform both tests on the same beaded can.

	FE_Can_height_average	FE_Can_height_range	FE_top_wall_thickness_average	FE_top_wall_thickness_range	FE_midwall_thickness_average	FE_midwall_thickness_range	Panel_depth
count	5403.000000	5403.000000	5403.000000	5403.000000	5403.000000	5403.000000	5403.000000
mean	112.258125	0.046073	0.173616	0.011898	0.128962	0.002430	3.978706
std	0.023544	0.026094	0.002122	0.006347	0.002102	0.001972	0.067653
min	111.971500	0.000000	0.166500	0.000000	0.120250	0.000000	3.543000
25%	112.244500	0.028000	0.172000	0.007000	0.127500	0.002000	3.934000
50%	112.257750	0.041000	0.173500	0.011000	0.129250	0.002000	3.980000
75%	112.272250	0.059000	0.175250	0.015000	0.130500	0.003000	4.028000
max	112.894250	0.287000	0.187500	0.050000	0.157500	0.118000	4.134000

Figure 5.11: : Basic Statistics table for front end data of 2-piece metal food can manufacturing process

	Flanger_flange_width_range	Flanger_flange_width_average	Flanged_can_height_range	Flanged_can_height_average
count	2151.000000	2151.000000	2151.000000	2151.000000
mean	0.077522	2.531732	0.068748	110.661011
std	0.031857	0.035993	0.038021	0.086076
min	0.006000	2.369750	0.001000	109.427000
25%	0.055000	2.510000	0.041000	110.642333
50%	0.075000	2.534250	0.063000	110.663333
75%	0.097000	2.555500	0.091000	110.684667
max	0.234000	2.637000	0.414000	110.959667

Figure 5.12: : Basic Statistics table for flanger data of 2-piece metal food can manufacturing process

Quantile-quantile plots were drawn for each of the factors to determine the normalcy of the data distribution and to visualize outliers. See **Figure 5.14** for quantile-quantile plots of front end can heights, mid-wall and top-wall thicknesses as well as panel depths.

Figure 5.15 show the quantile-quantile plots of flanged can heights as well as flanged can flange widths.

Figure 5.16 show the quantile-quantile plots of beaded can heights, flange widths, roll bead diameters, roll bead positions, bead depths as well as axial load resistance. A similar set of data exists for the beader, but which replaces the axial load resistance with panelling pressure resistance. See **Appendix A.2.1** and **Figure A.2** for the quantile-quantile graphs of the beader data set that includes the panelling pressure resistance.

	Beaded_can_flange_width_range	Beaded_can_flange_width_average	Beaded_can_height_range	Beaded_can_height_average	Roll_bead_diameter	Beaded_can_roll_bead_position_range	Beaded_can_roll_bead_position_average
count	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000
mean	0.062481	2.531511	0.165577	109.493406	74.224121	0.221090	5.174550
std	0.030756	0.035171	0.061503	0.086234	0.066599	0.157206	0.128379
min	0.005000	2.366750	0.003000	109.203333	73.943000	0.000000	4.663000
25%	0.041000	2.511000	0.105000	109.436000	74.179000	0.105000	5.096000
50%	0.059000	2.532750	0.159000	109.496667	74.225000	0.209000	5.180000
75%	0.080000	2.554750	0.219000	109.550333	74.269000	0.318000	5.255000
max	0.469000	2.796000	0.534000	109.800333	74.437000	4.561000	7.296500

Bead-Symmetry	Axial_load	Bead1_depth	Bead2_depth	Bead3_depth	Bead4_depth	Bead5_depth	Bead6_depth	Bead7_depth	Bead8_depth	Bead9_depth	Bead10_depth	Bead11_depth	Bead12_depth	Bead13_depth	Bead14_depth
3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000
-0.030014	198.813369	0.496104	0.523233	0.533154	0.529083	0.531955	0.535368	0.570457	0.573358	0.582038	0.591391	0.598964	0.606949	0.613058	0.613321
0.018969	12.903369	0.015767	0.025061	0.025306	0.025275	0.026072	0.026616	0.026598	0.027429	0.027858	0.027756	0.028064	0.028533	0.028168	0.028336
-0.099000	163.680000	0.436500	0.435500	0.446500	0.446000	0.447000	0.434500	0.490500	0.493000	0.504000	0.512500	0.521500	0.526500	0.530000	0.530500
-0.042000	189.560000	0.489500	0.513500	0.523000	0.518500	0.520000	0.521000	0.557500	0.559000	0.567000	0.575500	0.583000	0.590500	0.597500	0.597000
-0.029000	200.900000	0.496000	0.526500	0.536000	0.531000	0.533500	0.537500	0.569500	0.571000	0.579000	0.588000	0.596000	0.604000	0.609500	0.611000
-0.017000	207.650000	0.506500	0.537500	0.546500	0.542500	0.545500	0.549500	0.582000	0.586000	0.596500	0.606000	0.613500	0.622500	0.626000	0.627000
0.032000	372.220000	0.537500	0.574000	0.587500	0.586000	0.594000	0.606500	0.643000	0.650500	0.663000	0.671000	0.678500	0.687500	0.693000	0.692000

Bead15_depth	Bead16_depth	Bead17_depth	Bead18_depth	Bead19_depth	Bead_depth_range	Bead_depth_average
3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000	3313.000000
0.613990	0.625631	0.644031	0.640651	0.555200	0.149775	0.577796
0.027247	0.027733	0.026758	0.029456	0.045496	0.017685	0.026016
0.521000	0.540000	0.559000	0.550500	0.429000	0.101500	0.496579
0.599000	0.612000	0.629000	0.623500	0.528000	0.137500	0.565632
0.612000	0.624500	0.642500	0.639000	0.552000	0.146500	0.576947
0.626500	0.639500	0.658000	0.656500	0.580500	0.159000	0.588789
0.694000	0.700500	0.716000	0.723500	0.701000	0.212000	0.635269

Figure 5.13: : Basic Statistics table for beader data including axial load resistance data of 2-piece metal food can manufacturing process

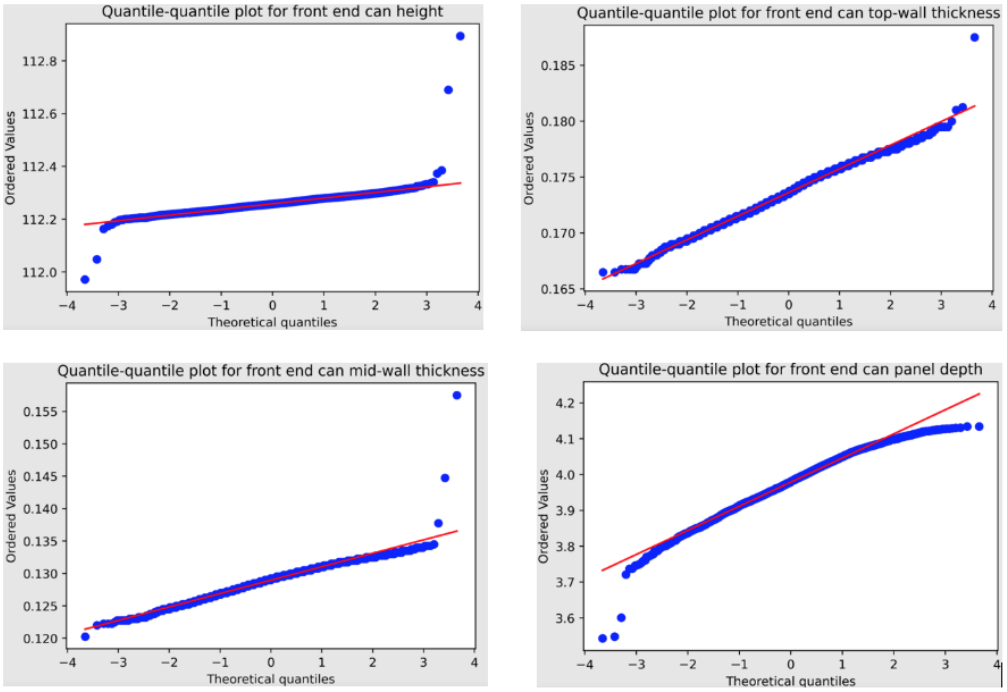

Figure 5.14: : Quentile-quentile plots for the front end of 2-piece metal food cans

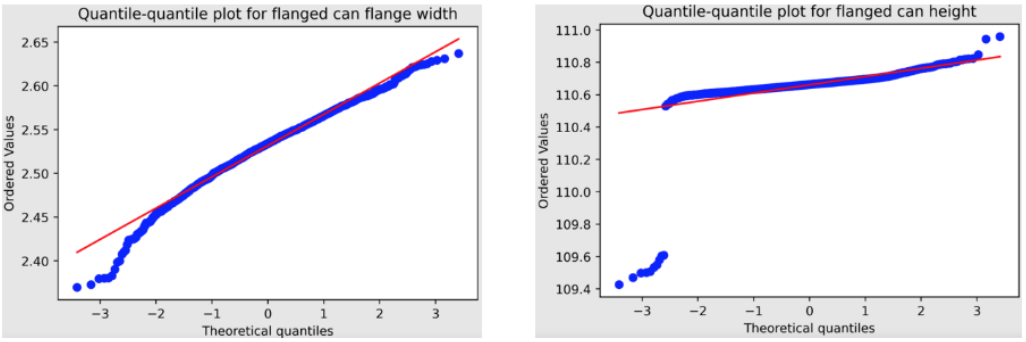

Figure 5.15: : Quentile-quentile plots for the flanger of 2-piece metal food cans

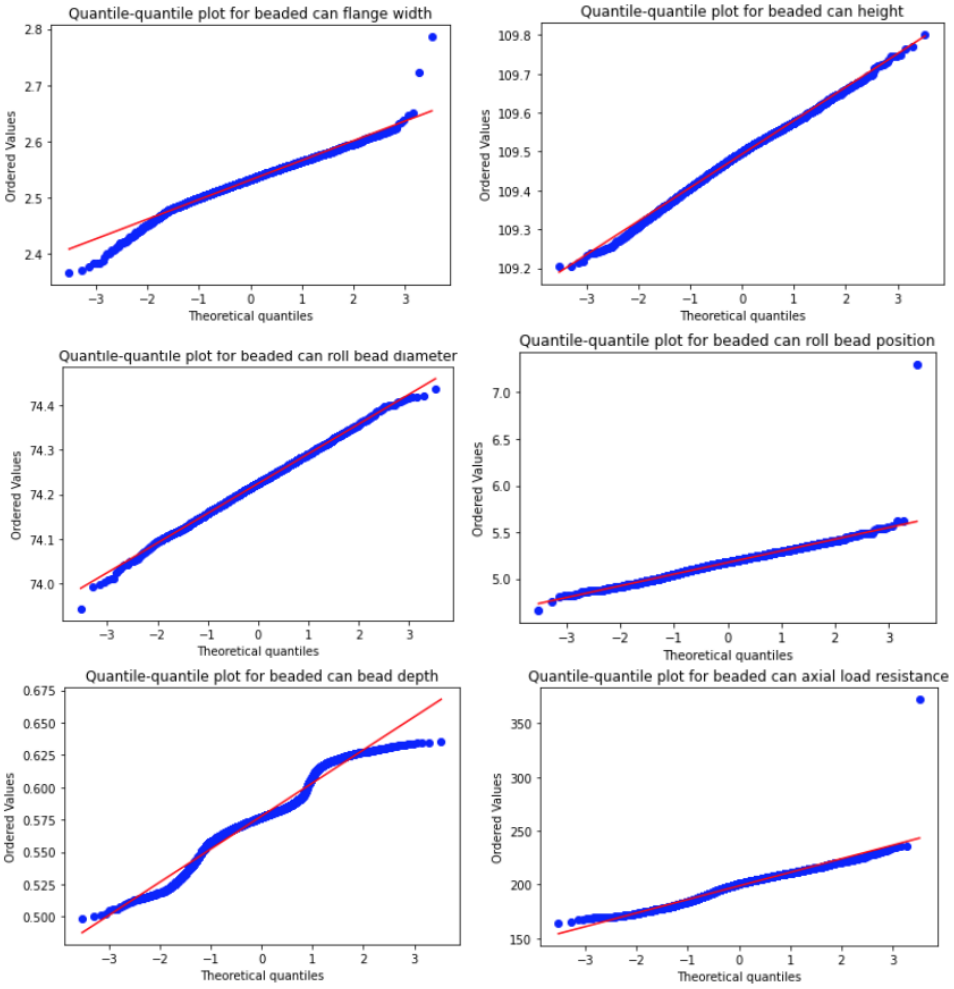

Figure 5.16: : Quentile-quentile plots for the beader of 2-piece metal food cans

From the graphs it can be seen that the data, for all the factors, was normally distributed, but there are some evidence of outliers. The outliers for the various factors were not across the same dates and times, therefore the outliers could not be traced to a specific instance where wrong data was captured. The different outliers did however cause large ranges for most of the factors. Many of these single measured values were too high or too low to be realistic and therefore it would make sense to treat the outliers as data errors.

5.6.2 Data assessing

Assessing of data included the identifying of missing data and the identifying of the data types, combining data tables where possible, searching for correlations in the data, visualizing the data and assessing the categorical data.

5.6.2.1 Data types and missing data

According to (Raschka 2015), the amount and quality of data that is fed into a ML algorithm is very important for the model to be able to learn from the data and give useful outputs. For various reasons there may be missing data. Missing data cannot be left as blanks as the ML algorithms generally cannot handle missing values and will respond with error messages, or the outcome of the model will be inaccurate. Some methods to handle missing data is to remove the data or to impute the missing values. Removing too many data points can affect the accuracy of the predictive model and decrease the reliability of the ML model. Imputation of data can include replacing missing values with the mean values for that factor, or the median value or the most frequent value. **Section 5.6.1** of the case study describes that there were not much missing data, and that the missing data was deleted and not imputed.

Data types can be categorical or numerical. Categorical data is data that is separated by different classes. The categorical data can be either nominal or ordinal. Ordinal data is data in categories that can be ordered from a smaller number to a bigger number with the bigger number also representing a bigger value. Nominal categories do not subscribe a ranking in the magnitude of the different categories. Ordinal features should be represented as integers in a ML model. Nominal integers cannot be represented as numbers in a single feature column and therefore one-hot encoding is applied. In one-hot encoding, dummy variables are created for each nominal class variable and binary values are used to represent the presence of a class, where 1 indicates the presence of that variable in the new dummy variable column and 0 indicates the opposite. A disadvantage of introducing one hot encoding is that it can cause multicollinearity in the dataset which can cause problems for some ML methods (Raschka 2015)

In the case study, the factors that have different categories are the various instrument identification numbers e.g. body-maker number 1 to 5, trimmer head 1 to 3, flanger 1 to 10 and beader 1 to 16. Refer to **Figure 5.17** where the beader column in the dataframe, which was coded as objects and not numbers, has been expanded to 16 different columns. Each beader number from 1 to 16 now is a separate column consisting of 0 and 1 integers and 1 represents the beader in that column and 0 represents the absence of the beader in that column. The one hot encoded dataframe was concatulated with the original dataframe to replace the original beader column with the 16 new beader columns.

Beader_1.0	Beader_2.0	Beader_3.0	Beader_4.0	Beader_5.0
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

Figure 5.17: : Snippet of dataframe that includes one hot encoded columns of the beader categorical variable

5.6.2.2 Combination of data tables

As described in **Section 5.2**, there are 3 sub-sections in the manufacturing process flow of 2-piece metal food cans. Sub-section 1 and 2 both were represented with a single data table and sub-section 3 was represented with two data tables. The two data tables of sub-section 3, the beader, were similar in dates, times and all the factors, except for one factor. One data table had axial load resistance and one data table had panel pressure resistance as factor. The reason that there were two data tables, was that both axial load resistance as well as panel pressure resistance were destructive tests, and therefore it was impossible to do both measurements on the same beaded can. These two tables could be appended due to similar dates and times these measurements were performed and logged into the data tables. Although the axial load resistance and panelling pressure resistance are not directly relatable, there still might be strong correlations between these two response factors due to measurements made on cans that were produced at the same time and under the same conditions. In this case study the panelling pressure resistance data was appended to the axial load resistance data table. The advantage of appending the two tables is that axial load resistance and panelling pressure resistance can be captured into one data table. The disadvantage is that the panelling pressure resistance, as a response, will not be as accurately predicted as the axial load resistance, since the data table consists of measurements done on the same cans as was done for axial load resistance. A solution is to develop two models for each response variable. Any further phases in the conclusion of the project on which this case study is based can include two separate predictive models for the two response variables.

The flanger and front end data tables also had similar dates and times for measurements, as the beader, but the factors are different. The difficulty in appending these two tables with the beader table was the impossibility to directly relate the flanger head number and the body-maker number to the beader mandrell number.

When a can is manufactured, it first goes through one of six body-makers. The body-maker num-

ber is recorded during the front end quality measurements. In the second section, cans go through a flanger and the flange is formed via 1 of 10 flanger heads. The flanger head is recorded during the measurements, but at this stage the body-maker number is not captured during the flanger measurements. The same applies for the beader quality measurements at the third section, where beads are formed via 1 of 16 bead mandrells, but neither the body-maker number, nor the flanger head number is related to these cans when they are measured.

Due to the reasons described in the above paragraph, data tables for sub-section 1 and 2 cannot be directly related to the beader. One solution was to only use beader data and discard data tables for sub-section 1 and 2. The disadvantage for this solution is that valuable and relevant information for the predictive analytics could be lost. A second solution, and the method that was used, was to calculate the average value of each flanger and front end measurement instance and append these average values to the beader data table. An example of how this was done, was to calculate the average values of flanger head 1 to 10 factors for each measuremnt instance and then append this average value to beaders 1 to 16 for the same measurement instance. In this way the ML model might be able to relate all the factors in the flanger and front end section to the beader section, with the exception of the categorical factors: flanger head number, front end body-maker number and front end trimmer head number.

5.6.2.3 Correlations in the data

The data of the combined data table consisted of 47 factors and 3179 measured instances. Of the 47 factors, the first two columns were the date and the time factors. There were a further 3 categorical factors; the team, the raw material supplier as well as the beader mandrel number. The other 42 factors all were numerical factors. Of those 42 numerical factors, two were considered the response variables namely the axial load resistance and the panelling pressure resistance. The response variable that the model should predict in this case study is the axial load resistance. A correlation table of the 42 numerical factors were drawn as can be seen in **Figure 5.18**.

The correlation between factors can be visualized with heat maps, scatter plots as well as parallel coordinate plots.

- A correlation heat map was drawn for the numerical factors in the process to visualize the correlation between these factors as seen in **Figure 5.19** . A correlation heat map is a visual representation of the correlation table in **Figure 5.18**. In the correlation heat map the blocks closer to yellow show good positive correlation and blocks closer to dark green show negative correlation. The lighter the shade of yellow or green the weaker the correlation.

	FE_Can_height_aver	FE_Can_height_range	FE_top_wall_thickness_average	FE_top_wall_thickness_range	FE_midwall_thickness_average	F
FE_Can_height_average	1.0	0.06374548020304460	-0.19482086577763200	-0.02192571266805650	-0.427460009943429	
FE_Can_height_range	0.0637454802030446	1.0	0.06103727496691220	-0.022941955472955300	0.015223322943858300	
FE_top_wall_thickness_average	-0.194820865777632	0.06103727496691220	1.0	0.1837049699447400	0.6228937478581610	
FE_top_wall_thickness_range	-0.021925712668056	-0.022941955472955300	0.1837049699447400	1.0	0.05605898086916130	
FE_midwall_thickness_average	-0.427460009943429	0.015223322943858300	0.6228937478581610	0.05605898086916130	1.0	
FE_midwall_thickness_range	-0.071843111661207	0.06701026269284010	-0.043523820279404200	0.007699914288618210	0.20896831403042700	
Panel_depth	0.1460459738918710	0.20410183262717	0.05194847761452110	-0.25240042356336700	-0.11895239628729600	
Flanger_flange_width_range	-0.066485663282206	0.005773281411851930	-0.005531981694339790	-0.04129767060361450	-0.03850767787585450	
Flanger -flange_width_average	0.0260203015901235	-0.16923385586642300	-0.04304934515710390	0.132682139079199	-0.20104076568612300	
Flanged_can_height_range	-0.062628225119814	0.08494306288703370	-0.01530886328874280	-0.07350964914496230	0.10195454258468200	
Flanged_can_height_average	0.1442039707912880	0.0021815400334240600	-0.026180494136897700	0.009742122483747010	-0.09397662536323720	
Beaded_can_flange_width_range	-0.054488981090113	0.020958898936228200	0.01807636125191470	-0.02245501058826020	0.028264709653532400	
Beaded_can_flange_width_average	-0.034673953480719	-0.08234533697144310	-0.012344220816887100	0.056773098151158300	-0.12546902278973600	
Beaded_can_height_range	0.0336979076744208	-0.022705816895776000	-0.03324003685122240	0.04524572708071660	-0.033771931013989100	
Beaded_can_height_average	-0.124859937449246	0.14394410710520100	0.11466936737615900	0.04772931068419610	0.17560569113380800	
Roll_bead_diameter	-0.078769673684763	0.08402451795299000	0.1575739440252960	-0.16334490286244700	0.15172608930150000	
Beaded_can_roll_bead_position_range	-0.029339208869177	0.030614270215087300	0.012321265165640500	-0.016513924504925500	0.01787969849365010	
Beaded_can_roll_bead_position_average	0.0055933806146534	0.12895233175705900	0.07863067098911940	-0.057705769681430800	0.013200250042180800	
Bead-Symmetry	0.0471441428564458	-0.059733394782512 10	-0.07410419664209080	-0.02710148201011120	-0.022635924753713200	
Axial_load	-0.349866687399897	0.11863421323687000	0.26701047167915400	-0.005367040335515840	0.47386729283647100	
Panel_resistance	0.0569761152066673	-0.16136615371030100	-0.08938879172458020	0.04445939843377070	-0.055551310351965400	
Bead1_depth	0.1179743243169280	-0.0853361264886566	-0.08246602284687410	0.020662455645331100	-0.10491327846241100	
Bead2_depth	0.1626429527902420	-0.11561214903598500	-0.13343629010931500	0.04406927672488750	-0.18237409664636500	
Bead3_depth	0.1826845340723220	-0.12329296914348800	-0.1512830304175030	0.047488205076531400	-0.20366497186791900	
Bead4_depth	0.2013061869528890	-0.1288488874663210	-0.1644837249509450	0.047434958472038400	-0.22332210421910400	
Bead5_depth	0.2071545981812230	-0.1314957232345940	-0.1659668354258650	0.05238695070971060	-0.22580612054700800	
Bead6_depth	0.2051942486710840	-0.14196387215291900	-0.17062916726320900	0.04048155405341190	-0.22703357863986100	
Bead7_depth	0.2465398892578390	-0.1647488152971510	-0.19258318852846100	0.030595845824415100	-0.2680969303495460	
Bead8_depth	0.2510421206739740	-0.17102322990640700	-0.19450441344846300	0.024736340413981900	-0.26985966552240400	
Bead9_depth	0.2447775056661670	-0.16962837627024	-0.19021528871533900	0.020043007790670900	-0.2603166476955520	
Bead10_depth	0.2434476331735350	-0.16904619858732000	-0.1886253176815640	0.016413850216049300	-0.2583099926927040	
Bead11_depth	0.2461754389712320	-0.17057878664382700	-0.18685594800690900	0.011466047646525800	-0.25514596781390000	
Bead12_depth	0.2375614949356950	-0.16577634231402600	-0.18011245810736100	0.010883995186709000	-0.24275611571935900	
Bead13_depth	0.2336865111716430	-0.16432984181022400	-0.17982197891291600	0.007412977910496320	-0.24081097560615100	
Bead14_depth	0.2330886165509700	-0.17018497012612500	-0.18045542656068600	0.010123698615148700	-0.24252410355246300	
Bead15_depth	0.2267617379542170	-0.1643795648022310	-0.1778327294138670	0.0179600898392253	-0.2305597597385770	
Bead16_depth	0.2109434232480210	-0.15183924445305100	-0.16325224054109300	0.011781853633765700	-0.21234850125226700	
Bead17_depth	0.2205032503016610	-0.15244711123348700	-0.17377160856467700	0.008645122468579930	-0.22860909723010900	
Bead18_depth	0.2457007980108280	-0.16092247331839500	-0.1904460838353290	0.030402352420379200	-0.2637319954103990	
Bead19_depth	0.2480792859940880	-0.14973190740141500	-0.18278497208978200	0.06764484443958010	-0.2764026308943090	
Bead_depth_range	0.2554477694498260	-0.16411734611246800	-0.20414936282058300	-0.0008936472551003390	-0.2831272503942060	
Bead_depth_average	0.2373246531821960	-0.16167967407697500	-0.18485729715466100	0.0303865986847958	-0.25244384927253300	

Figure 5.18: : Extract of correlation table of the final data frame used for model building in 2-piece metal food can case study

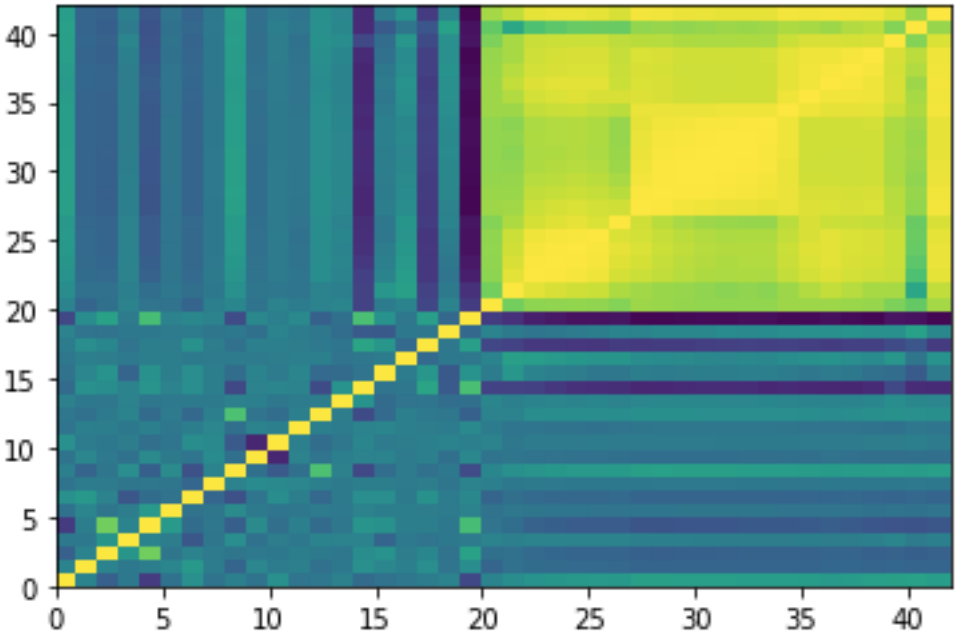

Figure 5.19: : Correlation heat map for the process factors

The diagonal line of yellow blocks evident in **Figure 5.19** represents the factors' correlations with themselves which is a perfect 1. The other factors that display good correlations with each other as well as with the two responses are all the bead depth factors and is visually represented by the mostly yellow block in the top right quadrant of the correlation heat map. Generally what could be seen from the correlation heat map as well as the correlation table are;

- The two response factors of axial load resistance and panelling pressure resistance were inversely correlated to each other; the higher axial load resistance was, the lower panelling pressure resistance was.

- There were strong relationships between all the individual bead depths as well as average bead depth with the response factors; the higher the bead depths the higher the panelling pressure resistance and the lower the axial load resistance were.

- Beaded can height, also known as the factory finished can height, also was correlated to the response factors; the higher the beaded can height the higher the axial load resistance and the lower the panelling resistance were. Beaded can height was also related to bead depths in a similar way that the response factors were, and therefore it was likely that the response factors are influenced by only bead depths or can heights.

- Other correlations between the response factors were; axial load resistance was positively

correlated to front end mid-wall thickness, and panelling pressure resistance was negatively correlated with roll bead positions. Again it can be seen that roll bead position was also related to bead depths in a similar way that panelling pressure resistance was, and therefore it was likely that the panelling pressure was influenced by only bead depths or roll bead position.

- Other correlations of interest were: front end mid-wall thickness was positively correlated to front end top-wall thickness and negatively correlated to front end can height. The flange width on the flanged cans and the flange width on the beaded cans were also positively correlated.

- All the bead depths from 1 to 19 showed positive correlations with each other. Some of these correlations were very high at almost 100 correlation e.g. beads 7 to 15 has shown stronger correlations with each other than with beads 1 to 6.

- Scatter plots were drawn to visually show the correlation between the average bead depth and respectively the panelling pressure resistance and the axial load resistance as seen in **Figure 5.20**.

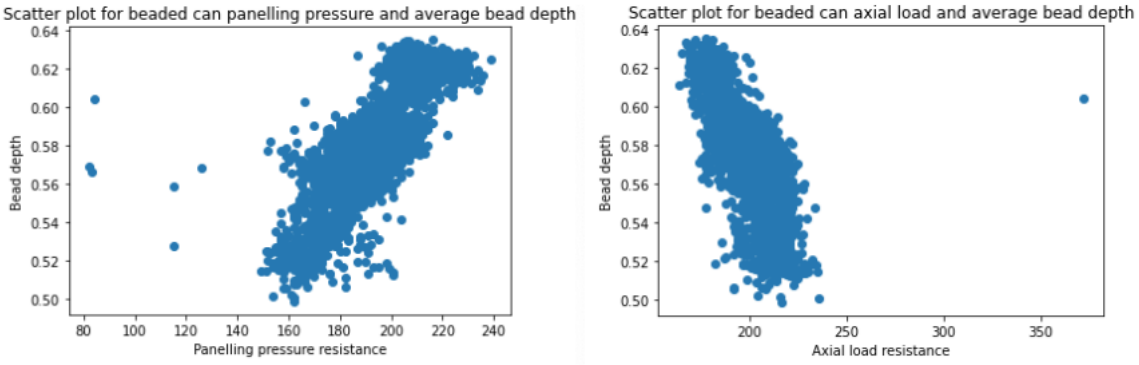

Figure 5.20: : Scatter plots for panelling pressure resistance and axial load resistance respectively against the average bead depth of beaded food cans

The Pearson's correlation for these two factors with average bead depth is 0.760080 and -0.703971 respectively, which is indicative of a significant statistical positive correlation between bead depth and panelling pressure resistance of 2-piece metal food cans and a significant negative correlation between bead depth and axial load resistance of 2-piece food cans. Refer to **Appendix A.2.2** and **Figure A.3** for the scatter plot between the two response variables.

- A parallel coordinate plot has been drawn to visualize coordination between factors which will be difficult to see from the other visual tools (see **Figure 5.21**).

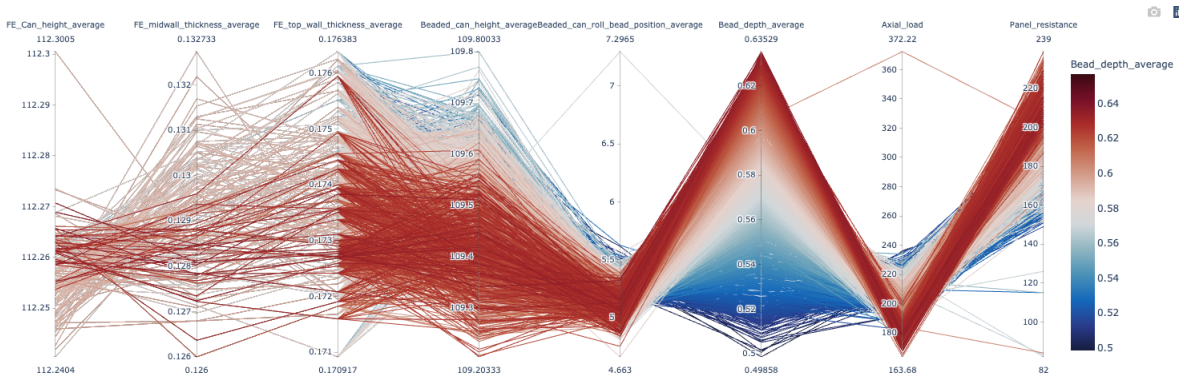

Figure 5.21: : Parallel plot of average bead depths versus some other process factors

The parallel coordinates plot vary lines from shades of blue to shades of red that respectively show lower to higher values for average bead depths of the beaded cans and how each instance relates to the other factors shown in the graph. From **Figure 5.21** it can be seen that there was a relation between bead depth averages and panelling pressure resistance, axial load resistance, roll bead position and beaded can height average. The parallel coordinate graph also showed poorer correlations between average bead depth and the front end top-wall thickness, front-end mid-wall thickness and front end can height, but it was evident from the graph that there were slight trends that can be picked up there as well.

5.6.2.4 Visualization of the data

Evident from the quantile-quantile plots, the correlation plots, the parallel coordinate plots as well as the statistical summary of the data, was that there were outliers. The spread of the data can be visualized by using boxplots. A boxplot shows the distribution of data. The orange line in the box shows the median value. The box represents all the data from the 25^{th} to the 75^{th} percentile and is known as the inter-quartile range. The two whiskers on either side of the box are 1.5 inter-quartile ranges beyond the box. Any points outside the whiskers can be described as the outliers.

As can be seen in **Figure 5.22** there were outliers in both the axial load resistance data as well as the panelling pressure resistance.

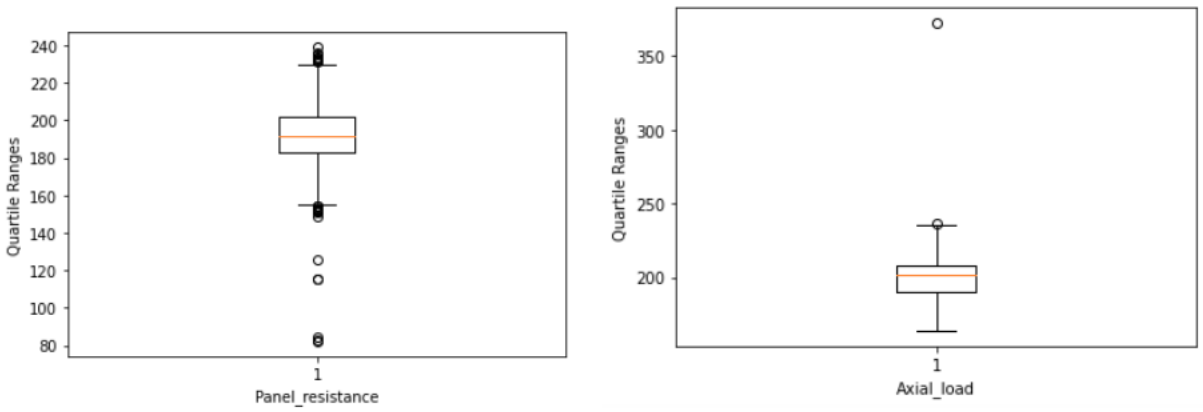

Figure 5.22: : Boxplots of axial load resistance and panelling pressure resistance

Outliers were evident in many of the factors represented in the data collected from the process as can be seen in **Figure 5.23**.

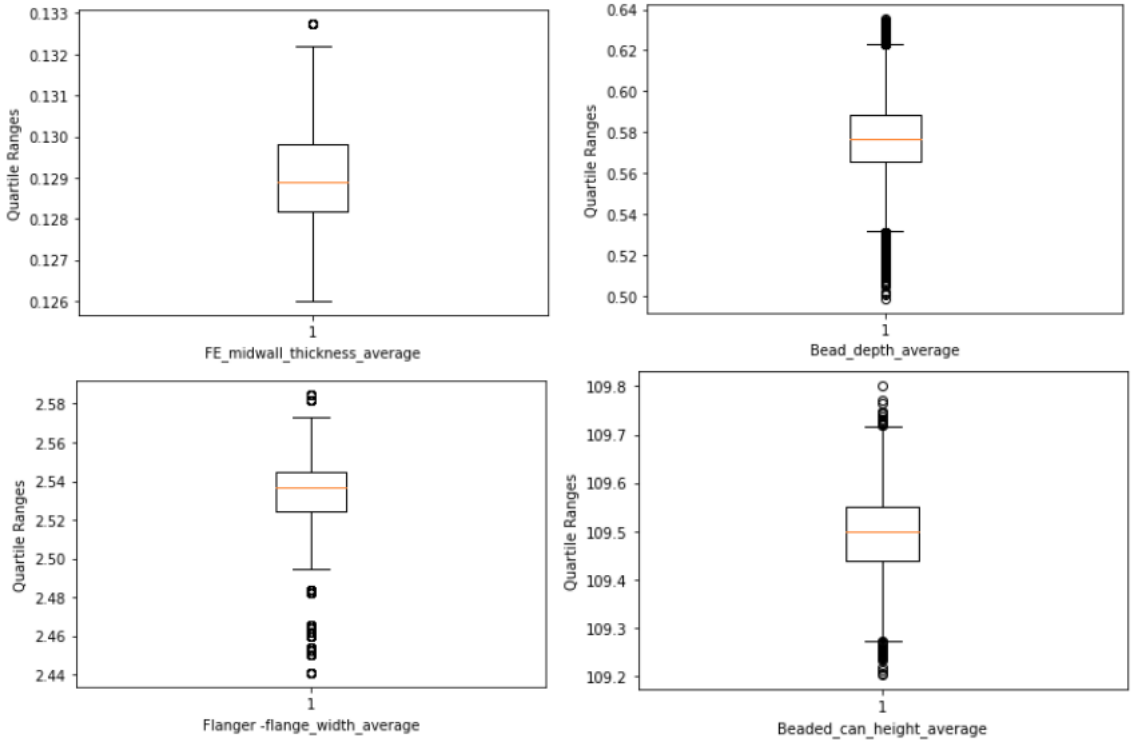

Figure 5.23: : Selected boxplots from the 2-piece metal can manufacturing process

5.6.2.5 Categorical data

ANOVA can be used to test whether categorical variables are similar. A categorical factor will have 2 or more different categories or groups. The one-way ANOVA is performed by taking random samples from each group. The mean of each group is calculated and the variation of the means within the group is compared to the mean among the groups. The hypothesis for similarity is then tested based on the F-score and the P-value, see **Section 3.5.2.1** for a mathematical description of ANOVA. Assumptions that must be satisfied when doing an ANOVA is that the data from all the sample sets are independent and that the data points for all the sets are normally distributed. The normality was already checked and satisfied with the drawing of the quantile-quantile plots as described in **Section 5.6.1**.

ANOVA uses a null hypothesis and an alternate hypothesis. The Null hypothesis in ANOVA is valid when all the sample means in the group are equal, or they don't have any significant difference, thus they can be considered as a part of a larger set of the population. On the other hand, the alternate hypothesis is valid when at least one of the sample means is different from the rest of the sample means.

ANOVA was performed on beader mandrel numbers, production teams and raw material suppliers as categories. The numerical factors used for the ANOVAs were the response factor of axial load resistance. Random samples of 500 instances were drawn and the steps involved in ANOVA were performed on the data and the hypotheses were accepted or rejected. The null hypotheses for axial load resistance was rejected for both the beader mandrels as well as the raw material suppliers. This indicated that

- all the different beader mandrels did not give statistically similar responses, therefore the null hypothesis was rejected.
- all the different raw material suppliers did not give statistically similar responses, therefore the null hypothesis was rejected.
- all the different production teams did give statistically similar responses, therefore the null hypothesis was accepted.

The outputs of the Python code for the categorical variables related to the axial load resistance can be seen in **Figure 5.24**.

ANOVA for beader mandrels 1 to 16 in terms of axial load resistance

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	12039	15	802.597	5.51354	1.09134e-10	1.85131
Within Groups	103353	710	145.568			
Total	115392	725	159.162			

```
Axial load resistance ANOVA
Approach 1: The p-value approach to hypothesis testing in the decision rule
F-score is: 5.5135429177703825  and p value is: 1.0913381309762826e-10
Null Hypothesis is rejected.
--------------------------------------------------------------------------
Approach 2: The critical value approach to hypothesis testing in the decision
rule
F-score is: 5.5135429177703825  and critical value is: 1.8513071271050896
Null Hypothesis is rejected.
```

ANOVA for production teams in terms of axial load resistance

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1160.25	3	386.749	2.44443	0.0628882	3.13421
Within Groups	114232	722	158.216			
Total	115392	725	159.162			

```
Axial load resistance ANOVA
Approach 1: The p-value approach to hypothesis testing in the decision rule
F-score is: 2.444434677744234  and p value is: 0.06288820126650396
Failed to reject the null hypothesis.
--------------------------------------------------------------------------
Approach 2: The critical value approach to hypothesis testing in the decision
rule
F-score is: 2.444434677744234  and critical value is: 3.1342146531749218
Failed to reject the null hypothesis.
```

ANOVA for raw material supplier in terms of axial load resistance

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	22968.7	2	11484.3	89.8381	1.11022e-16	3.70776
Within Groups	92423.8	723	127.834			
Total	115392	725	159.162			

```
Axial pressure resistance ANOVA
Approach 1: The p-value approach to hypothesis testing in the decision rule
F-score is: 89.83806430822877  and p value is: 1.1102230246251565e-16
Null Hypothesis is rejected.
--------------------------------------------------------------------------
Approach 2: The critical value approach to hypothesis testing in the decision
rule
F-score is: 89.83806430822877  and critical value is: 3.707764981631528
Null Hypothesis is rejected.
```

Figure 5.24: : ANOVA results for beader mandrel numbers, raw material suppliers and production teams in relation to axial load resistance

The null hypotheses were rejected for the raw material suppliers as well as the beader mandrels,

which suggested that these two categorical values should be incorporated as numerical categories in the data frame that will be used for predictive modelling.

5.6.3 Prepare data

Data preparation is the process of preparing data for algorithms that will be used in the ML models. Data preparation for the case study comprised of scaling data and reducing the data dimensionality with feature selection and feature extraction.

5.6.3.1 Scaling of data

Before the data was scaled the outliers were removed from the data table. Some of the outliers did not seem to be realistic outcomes for the process and therefore can probably be attributed to some or other form of error in capturing the data. Outliers were specifically removed from the axial load resistance response factor by replacing the top half percent and bottom half percent of the data values with the mean values of the axial load resistance data. Outliers were also removed from the following process factors; front end mid-wall thickness, flanged can height, beaded can flange width, beaded can height and beaded can roll bead position. Outliers were identified in these process factors from the descriptive statistics and quantile-quantile plots and boxplots.

According to (Subasi 2020) data needs to be normalized or standardized when using some algorithms in the modelling phase of the process improvement framework. Normalization is the re-scaling of data between two values such as -1 and 1. Standardization is the re-scaling of data around zero to form a normal distribution around the zero point. See **Section 3.5.2.2** for an expanded description on scaling. As part of the algorithms used in the ML modelling described later in this chapter, data used in this was standardized.

5.6.3.2 Feature selection

Both SFS as well as random forests were used for future selection on the case study data.

Python was used to do SFS on the case study data. The data table consisted of 66 factor columns. One factor was the date/time column and two more factors were the response factors, which were removed from the data used in the SFS algorithm. The remaining 63 factors were used to determine the performance of the axial load resistance as response variable. The algorithm was set to continue until the top 20 factors have been selected. **Figure 5.25** shows the performance of the top 20 factors to predict the axial load resistance of a 2-piece metal food can manufacturing line using SFS.

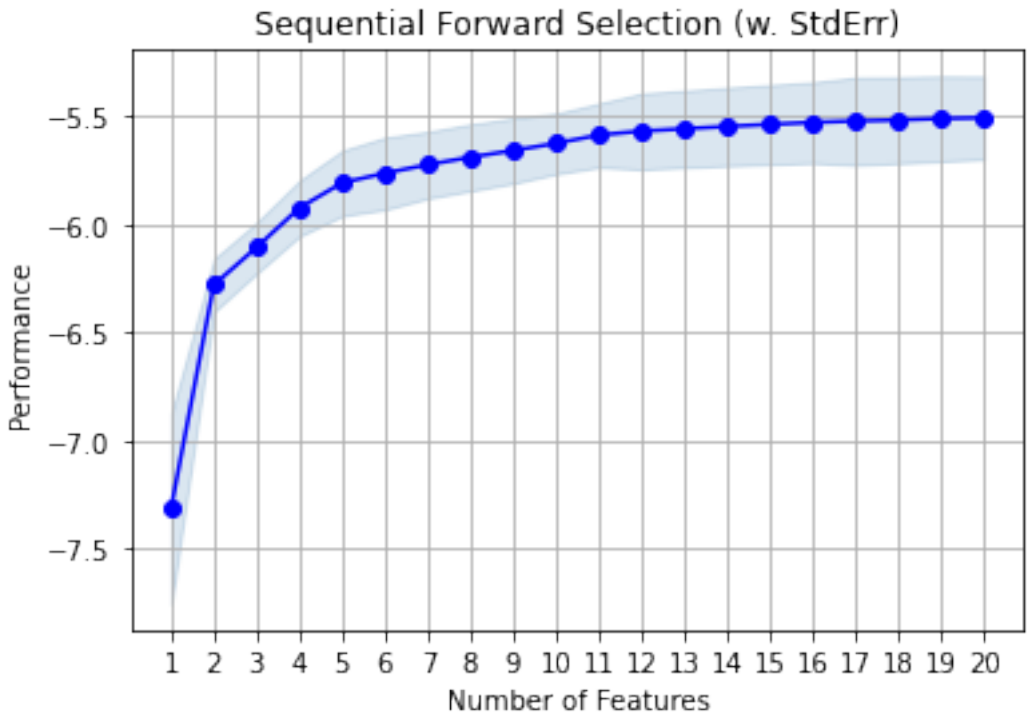

Figure 5.25: : Performance of the top 20 factors to predict the axial load resistance of a 2-piece metal food can manufacturing line using SFS

From **Figure 5.25** it can be seen that the standard error stabilises around 10 or 11 factors which is an indication that the model that will be used to do predictive analytics may be almost as accurate with those 10 factors as with all 63 factors. The top 10 factors that would maximize the performance of a predictive model to predict axial load resistance according to the SFS algorithm were:

- Bead depth of bead 18 on the beaded can
- Front end mid-wall thickness
- Bead depth of bead 8 on the beaded can
- Mandrel head number 9 at the beader
- The flange width on the cans after the flanger
- Mandrel head number 11 at the beader
- Bead depth of bead 12 on the beaded can

- The roll bead position on the beaded can
- The roll bead diameter on the beaded can
- The factory finished can height of the beaded can

From the SFS results for the axial load resistance it seems that the most important factors that will maximize the performance of a predictive model will be the raw materials used, mid-wall thickness and panel depth on the can after it exits the front end, flange width on the flanged can after it exits the flanger, the beader mandrel head used during beading of the cans and the bead depths after the cans exit the beader.

The second feature selection algorithm that was used for the case study was random forest regression. According to Schonlau and Zou (2020) random forest regression are better than linear regression to do predictive analytics. Linear regression models are easy to interpret but random forest regression is much more flexible due to its adaptability towards non-linearity, which makes it more suitable for predictions.

A decision tree divides a given data set into two groups according to a stated condition until a specific criterion has been reached, see **Figure 5.26** as an illustrative example.

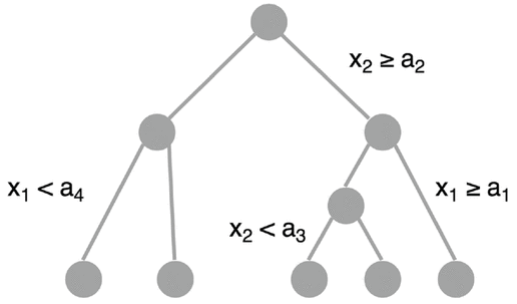

Figure 5.26: : An example of a decision tree (Schonlau and Zou 2020)

Decision trees can be used for classification as well as regression predictions, but tend to be prone to over-fitting which causes generally poor predictive capabilities. To increase the predictive capabilities of decision trees, the predictions from many trees can be averaged out in an ensemble tree-based algorithm called a random forest regression.

Python was used to do random forest regression on the case study data. Just as was the case with SFS, 63 factors were used to determine the performance of the axial load resistance as response variable. **Figure 5.27** shows a snippet of an example of a random forest regression tree to determine the most important factors to predict axial load resistance.

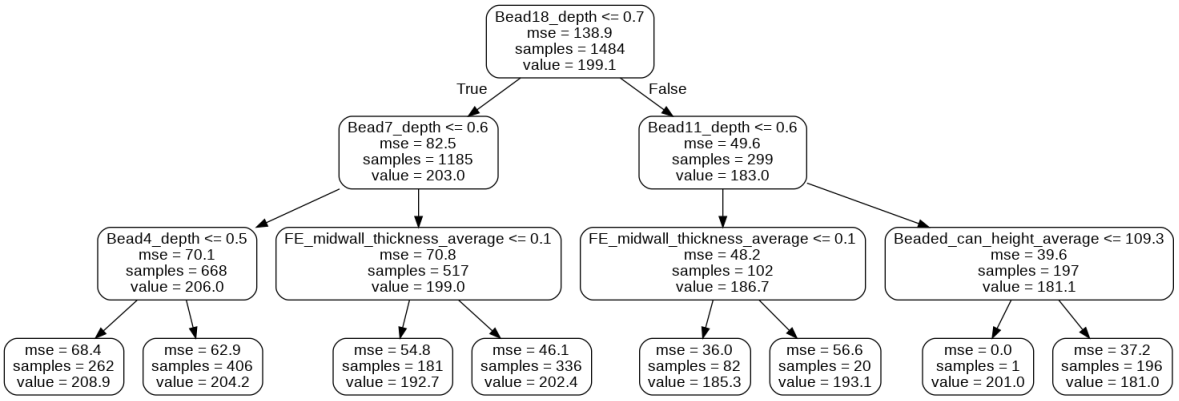

Figure 5.27: : An example of a section of a random forest regression tree to determine the most important factors to predict axial load resistance

The top 10 factors that would maximize the performance of a predictive model to predict axial load resistance according to the random forest algorithm and their relative importance can be seen in **Figure 5.28**

```
Variable: Bead18_depth          Importance: 0.25
Variable: Bead_depth_average    Importance: 0.16
Variable: FE_midwall_thickness_average Importance: 0.09
Variable: Bead7_depth           Importance: 0.09
Variable: Bead14_depth          Importance: 0.04
Variable: Panel_depth           Importance: 0.02
Variable: Flanger_flange_width_average Importance: 0.02
Variable: Beaded_can_height_average Importance: 0.02
Variable: Bead8_depth           Importance: 0.02
Variable: FE_Can_height_average Importance: 0.01
```

Figure 5.28: : The top 10 factors that would maximize the performance of a predictive model to predict axial load resistance according to the random forest algorithm

From **Figure 5.28** it can be seen that only the top 9 most important factors have an importance of more than 1%.

From the random forest regression results for the axial load resistance it seems that the most important factors that will maximize the performance of a predictive model will be the trimmed can height, mid-wall thickness and panel depth on the can after it exits the front end, flange width on the flanged can after it exits the flanger and the bead depths after the cans exit the beader.

5.6.3.3 Feature extraction

Python was used to perform the feature extractions methods of PCA and LDA on the case study data. PCA is affected by scale in the data and therefore data needs to be standardized before applying of the PCA. PCA ends up with a principal components data table that has low correlation between the principal components because variance has been removed from the data. See **Figure 5.29** for the first 20 rows of the PCA data table. The PCA data table was extracted from the case study data table by means of a PCA algorithm. The 8 extracted principal components contain about two-thirds of the information contained in the whole data set as calculated by the explained variance algorithm in Python.

PCA Table

	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5	principal component 6	principal component 7	principal component 8	Axial_load
0	0.781107107	-0.142430463	0.034547223	-0.119493978	0.297117164	0.134556621	-0.055134094	0.15641027	209.42
1	0.781986912	-0.143455131	0.033971913	-0.051589435	-0.084161506	0.131220501	-0.431936316	0.292433741	191.67
2	0.780803692	-0.142520674	0.030924845	0.155020268	0.029819771	0.506130734	-0.039857828	0.151413931	210.74
3	0.786091007	-0.14294457	0.034336735	-0.170580247	-0.361774292	0.236729149	-0.369022394	-0.242088227	210.39
4	0.778423286	-0.142284412	0.032235991	0.420064545	0.212780328	-0.037426972	-0.062744297	-0.066486914	205.18
5	0.791854124	-0.140699159	0.030269665	0.034058694	0.167213669	-0.077544995	-0.054970467	-0.190551883	220.74
6	0.78921121	-0.141212816	0.032389127	-0.072844515	-0.576980789	0.252245977	0.384184025	-0.057282916	210.51
7	0.796255847	-0.139924801	0.031462906	0.021031224	0.433716237	0.324636229	-0.15163594	-0.288209996	219.46
8	0.782428148	-0.141549112	0.029537242	-0.475768552	0.292799722	-0.047348888	0.271652438	-0.233460416	203.47
9	0.777193507	-0.140936917	0.031665083	0.314166849	0.140406968	0.0311799	0.492453161	0.132175109	206.24
10	0.785781592	-0.141027192	0.032333231	-0.496765552	0.090177992	-0.15920248	-0.046670706	0.266241083	200.38
11	0.778207572	-0.141931012	0.031589165	0.118763691	-0.020844039	-0.390562675	-0.146207159	0.321669747	206.52
12	0.786684127	-0.141129994	0.029603934	0.123901133	-0.088077726	-0.458283128	-0.262307738	-0.162406257	209.14
13	0.783930859	-0.141755983	0.031512169	0.122441678	-0.085281983	-0.008422052	0.10095054	0.526756468	204.94
14	0.789725855	-0.140753261	0.032660258	0.173318652	0.017473108	-0.243773719	0.291318508	-0.137445334	215.63
15	0.789823895	-0.140770378	0.033597332	-0.057712779	-0.246096304	-0.137868784	-0.028218493	-0.323671832	215.09
16	0.607371876	-0.812151079	-0.399755105	-0.217398397	0.265235452	0.133133289	-0.074381135	0.143592479	207.85
17	0.600130899	-0.812823317	-0.400830693	-0.002621392	-0.057309746	0.115565984	-0.441209724	0.272908645	203.85
18	0.599085269	-0.812807941	-0.398026162	0.135971926	0.010773836	0.529269556	-0.02318931	0.147662757	198.27
19	0.606608744	-0.811427997	-0.400338733	-0.250963743	-0.377911155	0.218127474	-0.389261342	-0.250696098	200.73
20	0.594550566	-0.81115036	-0.401783193	0.487905112	0.228308427	-0.057245349	-0.056719238	-0.085385752	194.32

Figure 5.29: : Data table with 8 principal components extracted from the case study date set in manufacturing of 2-piece metal food cans

PCA is useful to extract features from the original data and to compress the data when the ML predictive model are going to be sensitive to correlations in your data. If the ML predictive models are ensemble methods such as random forest regression or XGBoost regression, correlations in data will not be an issue.

The axial load resistance variable in the case study data was categorized from numerical values to three categories; high axial load resistance, mid axial load resistance and low axial load resistance. The 63 numerical factors in the case study were then run through an LDA algorithm in Python using scikit-learn and the axial load resistance categories were plotted against the top two linear discriminants, see **Figure 5.30**. **Figure 5.30** shows LDA analysis could discriminate between the high (blue), green (mid), orange (low) axial load resistance. The three groups are distinguishable, but still shows some overlaps.

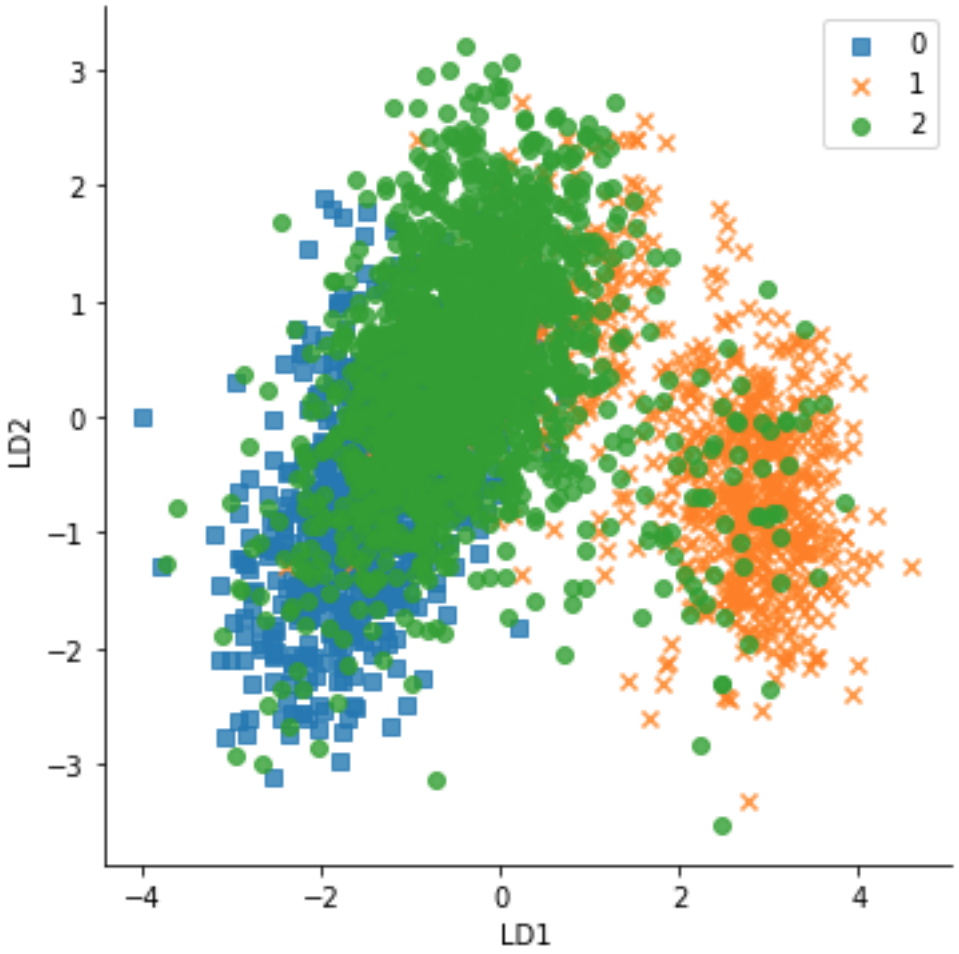

Figure 5.30: LDA of case study data for manufacturing of 2-piece metal food can manufacturing

5.7 PREDICTIVE MODEL BUILDING AND ACCURACY EVALUATION

For this case study, regression models were developed. Regression models are models that aim to predict a response as a numerical value on a continuous scale. Regression models can be used in industry to understand relationships within a process, evaluate trends in a process or to predict an outcome in the process (Raschka 2015). Various regression models have been developed of which some will be deployed for this case study. The following regression models have been attempted:

- Linear regression.
  - Simple linear regression.
  - Multiple linear regression.

- Penalized linear regression.
  - RANSAC regression
  - Lasso regression
  - Bayesian Ridge regression.
- Support vector machines regression.
- Decision trees.
  - Simple decision tree regression.
  - Random forest regression.
- Boost regression.
  - AdaBoost regression.
  - Gradient boost regression.

All the regression models have been evaluated for accuracy.

5.7.1 Linear regression

5.7.1.1 Simple linear regression

Simple linear regression is the most basic type of linear regression. Simple linear regression draws the correlation between the response, or dependent variable, and a single factor, or independent variable, from the data set. A simple linear regression model can be depicted as the best fitting straight line through the training data. See **Figure 5.31** and **Section 3.6.1.1** for more information on simple linear regression.

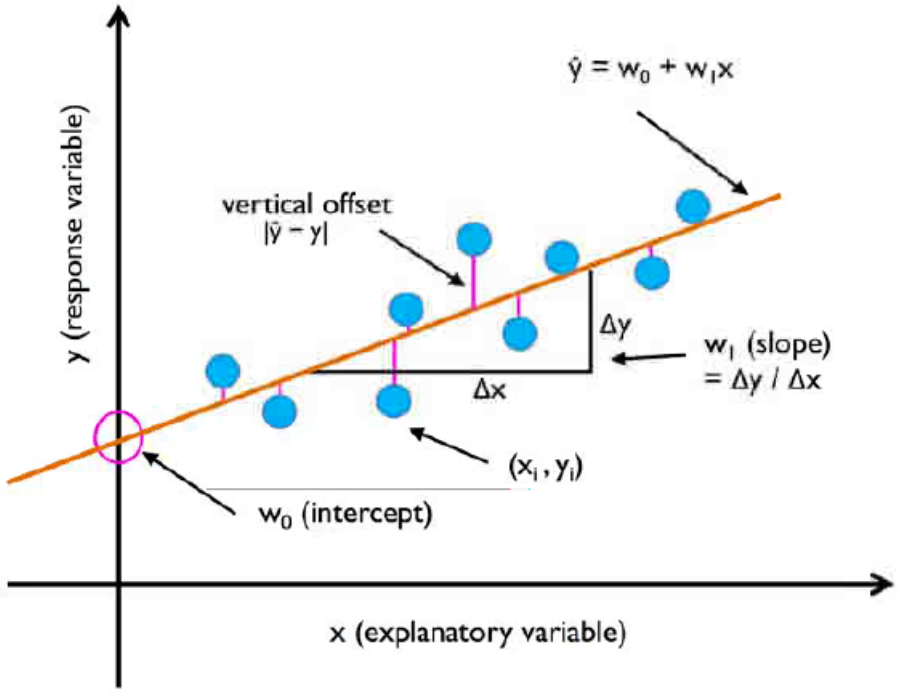

Figure 5.31: Visual representation of a linear regression model (Raschka 2015)

In **Section 5.6.3.2** the most important features were determined by random forest feature selection. The different bead depths were generally the most important factors and therefore the average bead depth was used as the independent variable to predict axial load resistance. **Figure 5.32** shows a graph of the scatter plot of axial load resistance vs. average bead depths.

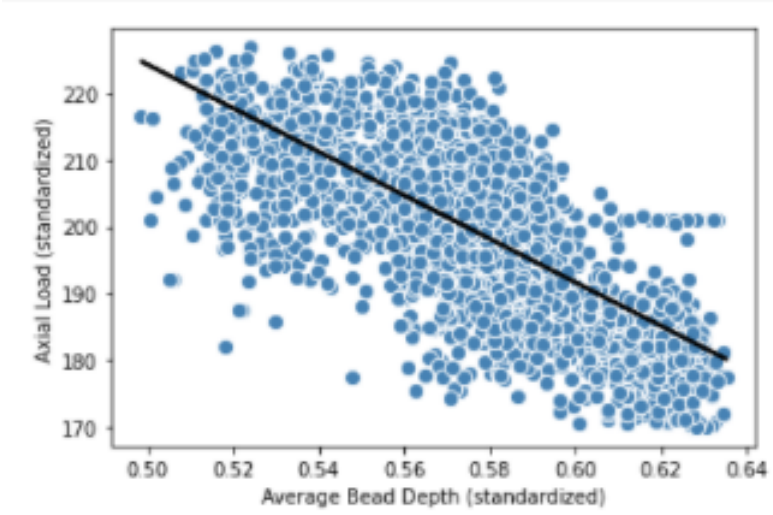

Figure 5.32: Scatter plot of axial load resistance vs. average bead depth of 2-piece metal food cans

The data was split between a test set and a train set at a ratio of 80% / 20%. The training data was used to train the linear regression model and the output was a regressor intercept (w_0) of 386.87533578833893 and a regressor coefficient (w_1) of -325.22496362. The linear regression model was used to predict the axial loads of the test data set, see **Figure 5.33** for some of the predicted values.

	Actual	Predicted
0	201.30	201.197557
1	203.42	201.770980
2	218.32	205.108815
3	205.33	201.711070
4	210.01	200.778188
...	...	...
631	174.69	183.421446
632	203.07	198.356118
633	208.80	198.244857
634	203.77	194.196662
635	203.97	201.925034

Figure 5.33: Linear regression model's predictions of axial load resistance of 2-piece metal food cans

The accuracy of the linear regression model was determined by calculating the mean absolute error (MAE) as well as the root mean squared error (RMSE) of the predicted axial load resistance when compared to the actual axial load resistance.

The MAE was 6.8161940081813155 and the RMSE was 8.446413963663227 for the linear regression model.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first component was used to predict the axial load resistance by means of a linear regression model, the MAE was 6.612780912137475 and the RMSE was 8.240349335477788. The MAE is smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

The PCA data accomplishes accuracy scores similar, but slightly better, than the feature selection data for the linear regression model.

5.7.1.2 Multiple linear regression

Most real world industrial problems will be much too complex to solve with a simple linear regression model. Multiple linear regression replaces one factor with as many that are needed,

$$y = w_0 x_0 + w_1 x_1 + \cdots + w_m x_m$$

where m = the number of factors.

When the multiple linear regression model only use 2 factors to predict the response the mode can be represented as **Figure 5.34** shows.

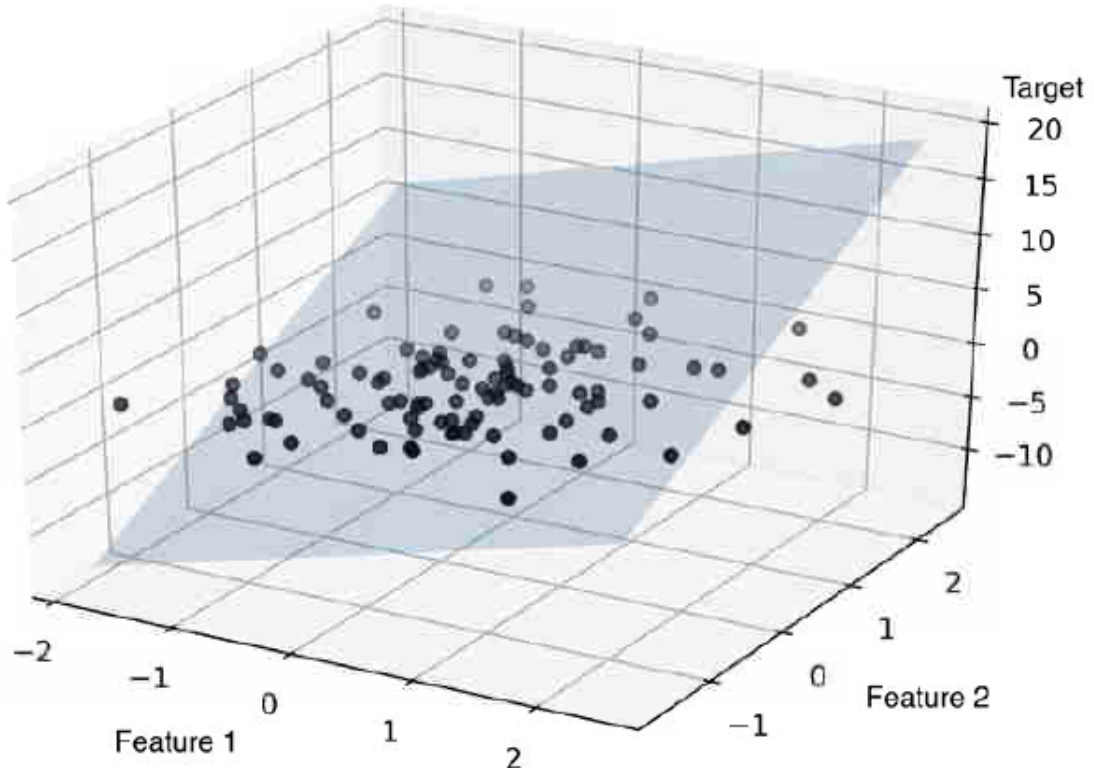

Figure 5.34: Visual representation of a 2-factor multiple linear regression model (Raschka 2015)

Visualization of a regression model beyond two factors are very difficult. See **Section 3.6.1.2** for more information on multiple linear regression.

A multiple linear regression model was built to predict axial load resistance with average bead depth, front end mid-wall thickness, flange width as well as material supplier as independent variables. The material supplier was added as whether Nippon was the supplier or not (if not Nippon, it has to be Arcerol Mittal since there were only two suppliers as part of the catured data). **Figure 5.35** shows the

coefficients (w_1, w_2, w_3, w_4) for the multiple linear regression model to predict axial load resistance in a 2-piece metal food can manufacturing line.

	Coefficient
FE_midwall_thickness_average	2968.801487
Flanger_flange_width_average	-46.835904
Bead_depth_average	-268.867610
Nippon	2.533576

Figure 5.35: Coefficients for the multiple linear regression model to predict axial load resistance of a 2-piece metal food can

The multiple linear regression model was used to predict the axial loads of the test data set, see **Figure 5.36** for some of the predicted values.

	Actual	Predicted
1093	201.30	204.186819
641	203.42	203.431531
1554	218.32	200.541773
575	205.33	199.226346
117	210.01	206.146445
...	...	...
2289	174.69	182.594093
529	203.07	197.149184
1292	208.80	205.932957
900	203.77	201.047649
1648	203.97	202.346856

Figure 5.36: Multiple linear regression model's predictions of axial load resistance of 2-piece metal food cans

The accuracy of the multiple regression model was determined by calculating the mean absolute error (MAE) as well as the root mean squared error (RMSE) of the predicted axial load resistance when compared to the actual axial load resistance.

The MAE was 5.719408990796563 and the RMSE was 7.2522258832943685 for the linear regression model. Both the MAE as well as the RMSE showed improvements for the multiple linear regression model when compared to the linear regression model.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first four components was used to predict the axial load resistance by means of a multiple linear regression model The MAE was 5.653828690227282 and the RMSE was 7.099070162436459. The MAE is smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

The PCA data accomplishes accuracy scores similar, but slightly better, then the feature selection data for the multiple linear regression model.

5.7.2 Penalized linear regression

Examples of ML models that use penalized linear regression algorithms are RANSAC regression, Lasso regression and Bayesian ridge regression.

5.7.2.1 RANSAC regression

Linear regression methods can be affected by outliers. In **section 5.6.3.1** some of the outliers were removed from the data, but the effect of the outlying data on our current compacted data sets are not known. Random sample consensus (RANSAC) regression can be used on data to negate the effect that outlying data points may have on predictions (Raschka 2015). An algoritm such as RANSAC can be used in a process that is prone to outliers, or where an outlier can negatively affect the accuracy of a predictive models.

The data table described in **Section 5.6.3.2**, which was developed from random forest feature regression was used to demonstrate RANSAC. Outliers were identified as those data points that were more than 20 units from the regression line as depicted in **Figure 5.37**.

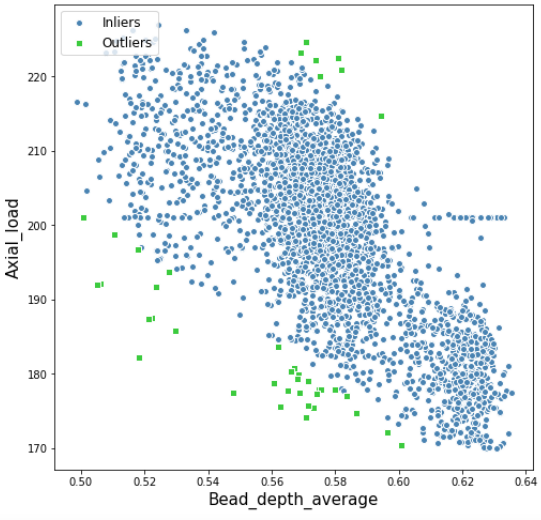

Figure 5.37: RANSAC regression showing outliers in green and inliers in blue

The intercept of the RANSAC regression was 393.731, and the slope was -336.792. The linear regression model had an intercept (w_0) of 386.875 and a coefficient (w_1) of -325.225, which was not that different from the RANSAC regression model. Outliers were removed from the original data set and since the removal of the remaining outlying data points did not change the outcome of the regression model much, there was probably no need to use RANSAC as part of this case study's predictive analytical model.

5.7.2.2 LASSO regression

Least absolute shrinkage and selection operator (LASSO) regression and ridge regression are examples of penalized regression. Both these regression methods apply a penalty to the regression coefficients to control over-fitting. Ridge regression places a penalty on the magnitude of the sum of squares of the coefficients and LASSO regression places a penalty on the amount of coefficients (Dangeti 2017).

LASSO regression has been used on the data table described in **Section 5.6.3.2** which was developed from random forest feature selection. Lasso regression demonstrates how the data can still be compacted further and still be representative in a regression model by setting the most unnecessary variables' coefficients to zero.

A LASSO regression model was built to predict axial load resistance with all the factors from the feature selection data table as independent variables. The Lasso regression model was used to predict the axial loads of the test data set, see **Figure 5.38** for some of the predicted values and **Figure 5.39** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	201.30	205.145894
641	203.42	200.828643
1554	218.32	205.492429
575	205.33	198.313783
117	210.01	205.000527
...	...	...
2289	174.69	183.192676
529	203.07	196.978611
1292	208.80	202.417623
900	203.77	200.885597
1648	203.97	203.399307

Figure 5.38: LASSO regression model's predictions of axial load resistance of 2-piece metal food cans

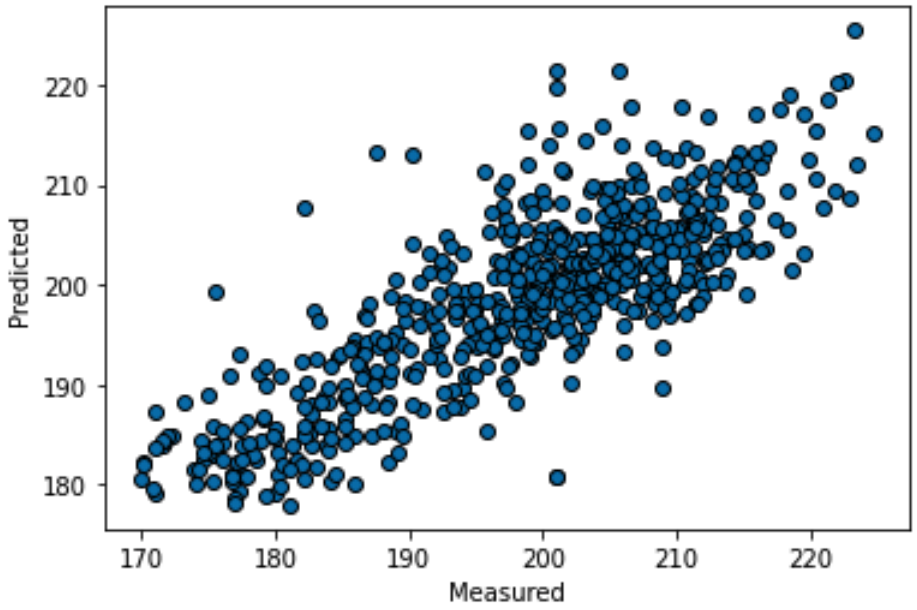

Figure 5.39: Graph depicting the measured values vs. the actual values of a LASSO regression model for the axial load resistance of 2-piece metal food cans

The MAE was 5.442785424670416 and the RMSE was 6.934690410442931 for the Lasso regression model. Both the MAE as well as the RMSE showed improvements for the LASSO regression

model when compared to both the linear regression models. The MAE is smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of a LASSO regression model, the MAE was 8.455854475785497 and the RMSE was 10.187157364219175.

The PCA data accomplishes accuracy scores that were less accurate than the feature selection data for both the linear regression and multiple linear regression model. The PCA accuracy scores were also not as accurate as was accomplished for the LASSO regression when the feature selection data was used.

5.7.2.3 Bayesian ridge regression

Bayesian ridge regression has been used on the data table described in **Section 5.6.3.2** which was developed from random forest feature selection. Bayesian ridge regression is a form of linear regression where the probability distribution of the data is used rather than individual points. Predictions are estimated not as a single value but are rather a representation of a probability distribution for that prediction.

A Bayesian ridge regression model was built to predict axial load resistance with all the factors from the feature selection data table as independent variables. The Bayesian ridge regression model was used to predict the axial loads of the test data set, see **Figure 5.40** for some of the predicted values vs. the actual values.

	Actual	Predicted
1093	201.30	206.030239
641	203.42	201.353077
1554	218.32	207.753722
575	205.33	198.460748
117	210.01	205.156091
...	...	...
2289	174.69	182.965179
529	203.07	197.504439
1292	208.80	201.138665
900	203.77	200.292831
1648	203.97	203.313504

Figure 5.40: Bayesian ridge regression model's predictions of axial load resistance of 2-piece metal food cans

The MAE was 5.353472187632018 and the RMSE was 6.847221857102071 for the Bayesian ridge regression model. Both the MAE as well as the RMSE were similar for the Bayesian ridge regression model when compared to the LASSO regression model. The MAE is smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 principal components were used to predict the axial load resistance by means of a Bayesian ridge regression model, the MAE was 8.45845371387966 and the RMSE was 10.187025779450783. The PCA data accomplishes accuracy scores that were similar for the Bayesian ridge regression model when compared to the LASSO regression model.

5.7.3 Support vector machines regression

SVM is better known as a classification regression model. SVM attempts to map hyper-planes between groups or classes of data in multidimensional spaces. SVM regression uses the same principle as SVM, but instead of mapping the data into categories, it finds a regression function for that mapped data.

An SVM regression model was built to predict axial load resistance with all the factors from the feature selection data table as independent variables. The SVM regression model was used to predict the axial loads of the test data set, see **Figure 5.41** for some of the predicted values and **Figure 5.42** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	201.30	207.069970
641	203.42	206.749090
1554	218.32	219.168983
575	205.33	196.913093
117	210.01	204.659744
...	...	...
983	210.03	207.862036
910	204.22	204.836165
1311	213.15	208.370681
215	204.04	202.551289
2870	195.65	197.338341

Figure 5.41: SVM regression model's predictions of axial load resistance of 2-piece metal food cans

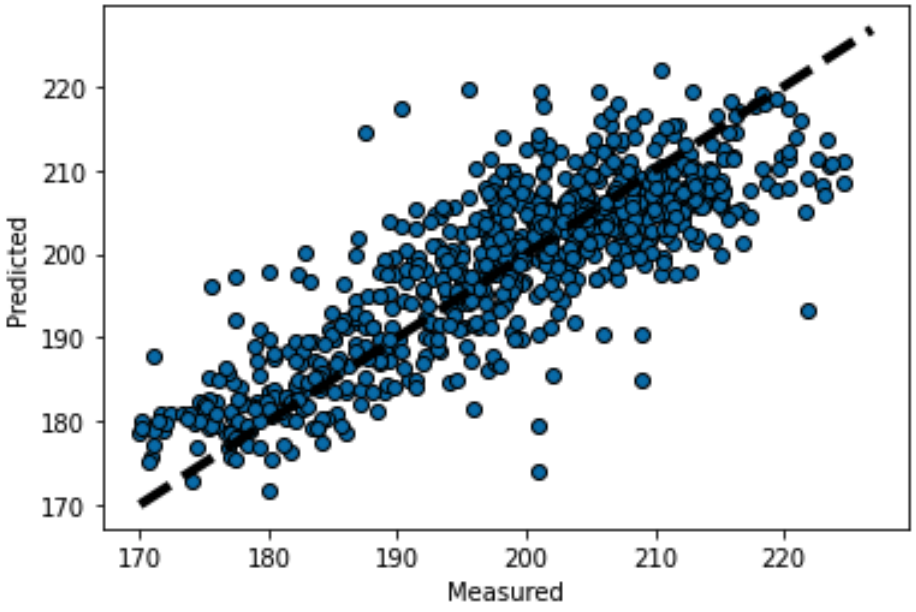

Figure 5.42: Graph depicting the measured values vs. the actual values of a SVM regression model for the axial load resistance of 2-piece metal food cans

The MAE was 5.231672218430161 and the RMSE was 6.780831566497392 for the SVM regression model. Both the MAE as well as the RMSE showed improvements for the SVM regression model when compared to the linear regression as well as penalized regression models. The MAE is

smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of a SVM regression model, the MAE was 6.713128889067907 and the RMSE was 8.50215938479019. The PCA data accomplishes accuracy scores that were less accurate than the feature selection data for the SVM regression model.

5.7.4 Decision tree regression

Decision trees breaks data sets up in subsets that become smaller and smaller until a path of tree nodes lead to a leaf node as predicted outcome.

5.7.4.1 Simple decision tree regression

A simple decision tree regression model was built to predict axial load resistance with all the factors from the feature selection data table as independent variables. The decision tree regression model was used to predict the axial loads of the test data set, see **Figure 5.43** for some of the predicted values and **Figure 5.44** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	201.30	204.13
641	203.42	204.13
1554	218.32	218.83
575	205.33	196.42
117	210.01	208.22
...	...	...
983	210.03	209.28
910	204.22	203.96
1311	213.15	211.13
215	204.04	207.74
2870	195.65	201.32

Figure 5.43: Simple decision tree regression model's predictions of axial load resistance of 2-piece metal food cans

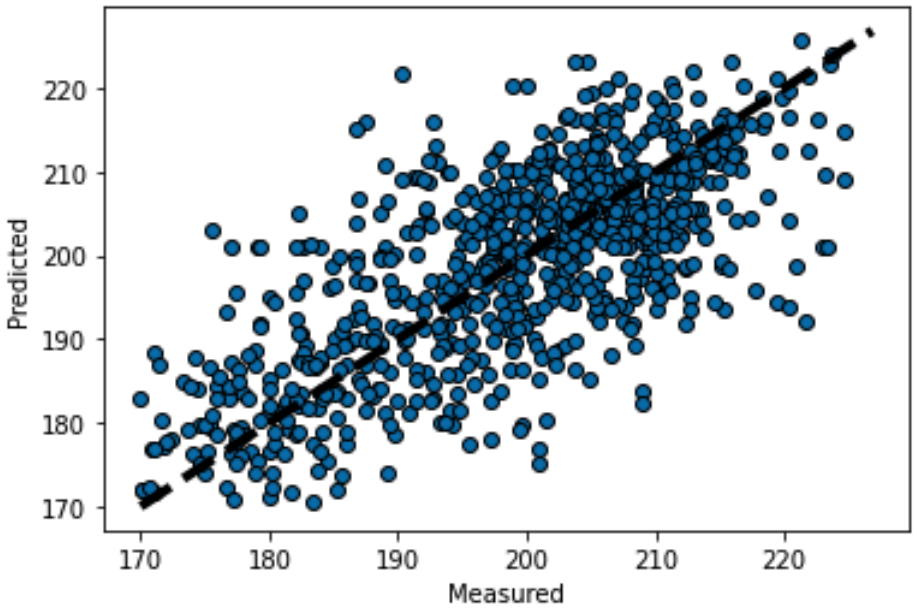

Figure 5.44: Graph depicting the measured values vs. the actual values of a simple decision tree regression model for the axial load resistance of 2-piece metal food cans

The MAE was 7.030943396226416 and the RMSE was 9.063330485911642 for the simple decision tree regression model. Both the MAE as well as the RMSE were not better for the simple decision tree regression model when compared to the simple linear regression models. The MAE is smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of a decision tree regression model, the MAE was 8.623031446540882 and the RMSE was 11.144850727391958. The PCA data accomplishes accuracy scores that were less accurate then the feature selection data for the SVM regression model.

5.7.4.2 Random forest regression

The random forest regression model is an example of an ensemble regression model. An ensemble regression model is a method that runs many predictive models and then combining these models by averaging out the predictions of these models (Bowles 2019). Bagging is when each of these separate predictive models randomly uses different portions of the training data to ensure the uniqueness of each model.

A random forest regression model was built to predict axial load resistance with all the factors from the feature selection data table as independent variables. The random forest regression model

was used to predict the axial loads of the test data set, see **Figure 5.45** for some of the predicted values and **Figure 5.46** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	201.30	205.126269
641	203.42	207.668887
1554	218.32	212.179068
575	205.33	200.518062
117	210.01	205.419440
...	...	...
983	210.03	209.944236
910	204.22	205.541570
1311	213.15	208.441929
215	204.04	203.913055
2870	195.65	197.790841

Figure 5.45: Random forest regression model's predictions of axial load resistance of 2-piece metal food cans

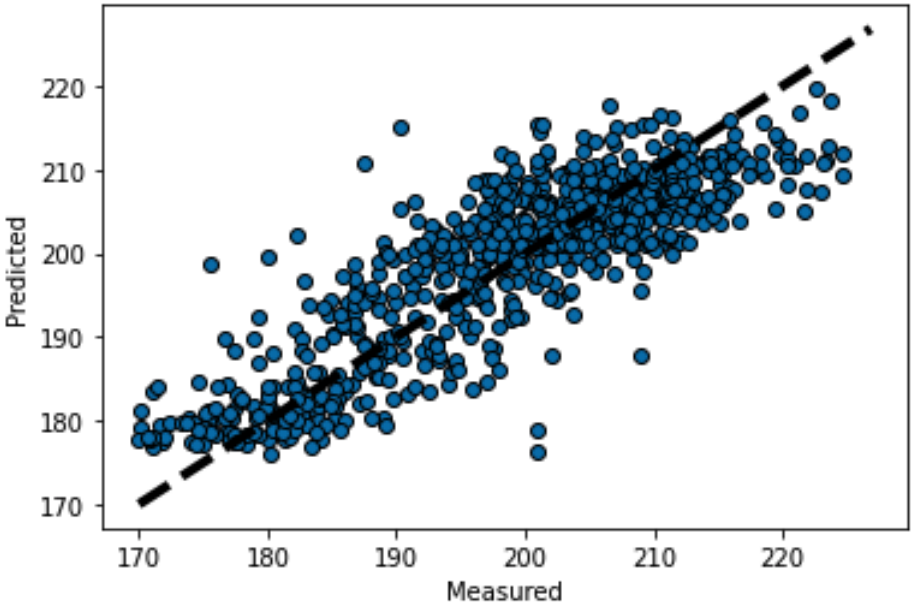

Figure 5.46: Graph depicting the measured values vs. the actual values of a random forest regression model for the axial load resistance of 2-piece metal food cans

The MAE was 4.851662347085858 and the RMSE was 6.199777158860995 for the random forest regression model. Both the MAE as well as the RMSE were better for the random forest regression model when compared to the SVM regression model as well as the other regression models described previously in this chapter. The MAE is smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of a random forest regression model, the MAE was 6.181527793192358 and the RMSE was 7.882959580158262. The PCA data accomplished accuracy scores that were less accurate then the feature selection data for the random forest regression model.

5.7.5 Boosted regression

Boosted regression is an ensemble regression method which repeatedly fit many decision trees in a sequential manner to ultimately obtain the most accurate solution.

5.7.5.1 Adaboost regression

An Adaboost regression model was built to predict axial load resistance with all the factors from the feature selection data table as independent variables. The Adaboost regression model was used to predict the axial load resistance of the test data set, see **Figure 5.47** for some of the predicted values and **Figure 5.48** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	201.30	207.560587
641	203.42	207.457607
1554	218.32	208.959706
575	205.33	201.000000
117	210.01	207.949796
...	...	...
983	210.03	208.635396
910	204.22	206.667475
1311	213.15	208.930822
215	204.04	204.506869
2870	195.65	198.960106

Figure 5.47: Adaboost regression model's predictions of axial load resistance of 2-piece metal food cans

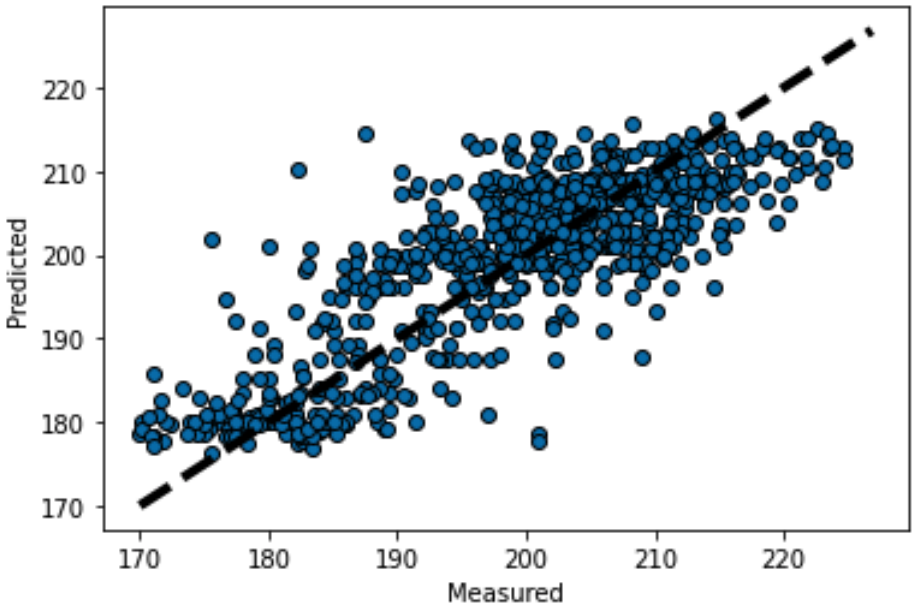

Figure 5.48: Graph depicting the measured values vs. the actual values of an Adaboost regression model for the axial load resistance of 2-piece metal food cans

The MAE was 5.332833876376473 and the RMSE was 6.86427645732993 for the Adaboost regression model. Both the MAE as well as the RMSE were not as good for the Adaboost regression model when compared to the random forest regression models. The MAE is smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of an Adaboost regression model, the MAE was 7.018457695087433 and the RMSE was 8.865729847330092. The PCA data accomplishes accuracy scores that were less accurate than the feature selection data for the Adaboost regression model.

5.7.5.2 Gradient boost regression

A gradient boost regression model was built to predict axial load resistance with all the factors from the feature selection data table as independent variables. The gradient boost regression model was used to predict the axial loads of the test data set, see **Figure 5.49** for some of the predicted values and **Figure 5.50** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	201.30	204.436654
641	203.42	206.383867
1554	218.32	215.440496
575	205.33	201.453417
117	210.01	204.567384
...	...	...
983	210.03	209.290457
910	204.22	203.808756
1311	213.15	208.782881
215	204.04	204.062563
2870	195.65	198.251129

Figure 5.49: Gradient boost regression model's predictions of axial load resistance of 2-piece metal food cans

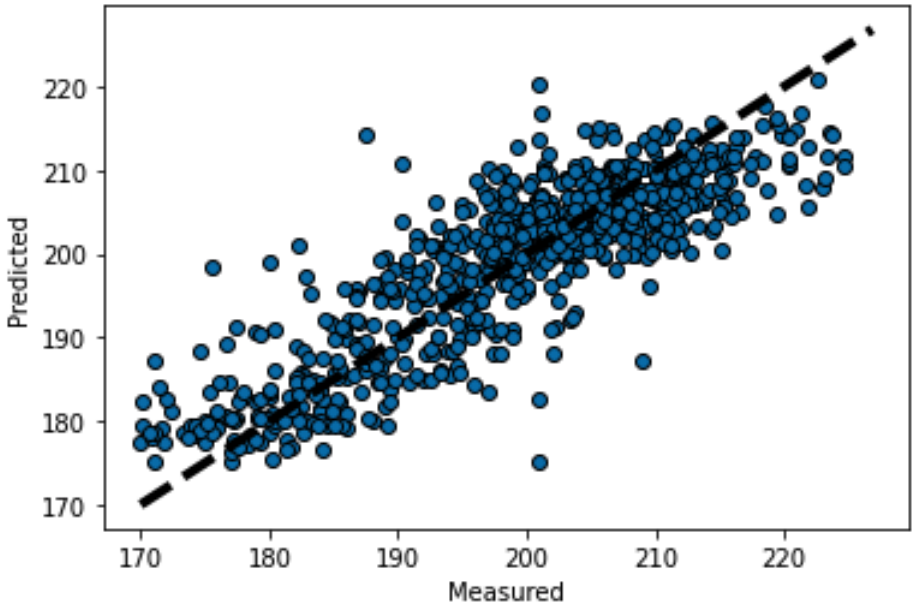

Figure 5.50: Graph depicting the measured values vs. the actual values of a gradient boost regression model for the axial load resistance of 2-piece metal food cans

The MAE was 4.647176229088174 and the RMSE was 6.016645830209083 for the gradient boost regression model. Both the MAE as well as the RMSE gave the best accuracy results for the gradient

boost regression model when compared to the random forest regression models and all the other models featured in this chapter. The MAE is smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2**.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of an gradient boost boost regression model, the MAE was 6.205096872359629 and the RMSE was 7.922144987449707. The PCA data accomplishes accuracy scores that were less accurate than the feature selection data for the gradient boost regression model.

5.8 EVALUATION AND DISCUSSION OF MODEL OUTCOMES

5.8.1 Insight from visual and statistical data analysis

In **Section 5.6** and **Section 5.7** available data from the process for manufacturing of 2-piece metal food cans have been assessed and used to build various predictive models for their axial load resistance as well as their panelling pressure resistance.

Data was obtained from various sub-sections of the process and combined in a data table. The data table had 3179 rows and 47 columns. The data consisted of 2 response factors, 40 numerical factors and 3 categorical factors. Various data visualisation tools and statistical tools were employed to show relationships between factors and between factors and responses. The following correlations were evident:

- The two response factors, panelling pressure resistance and axial load, resistance were inversely correlated.
- Bead depths were strongly correlated with each other, but also positively correlated with the panelling pressure resistance and inversely correlated with the axial load resistance.
- The factory finished can height was negatively correlated with bead depths, positively correlated with axial load resistance and inversely correlated with panelling pressure resistance.
- The mid-wall thickness of the cans was positively correlated with axial load resistance.
- The panel depth was inversely correlated to panelling pressure resistance.
- ANOVA results indicated a statistical difference between raw material suppliers of the tinplate from which the metal cans were manufactured in terms of the axial load resistance of the cans that were produced from the raw materials.

See **Figure 5.51** and **Figure 5.52** for a visual representation of some of the measured factors referred

to in this section.

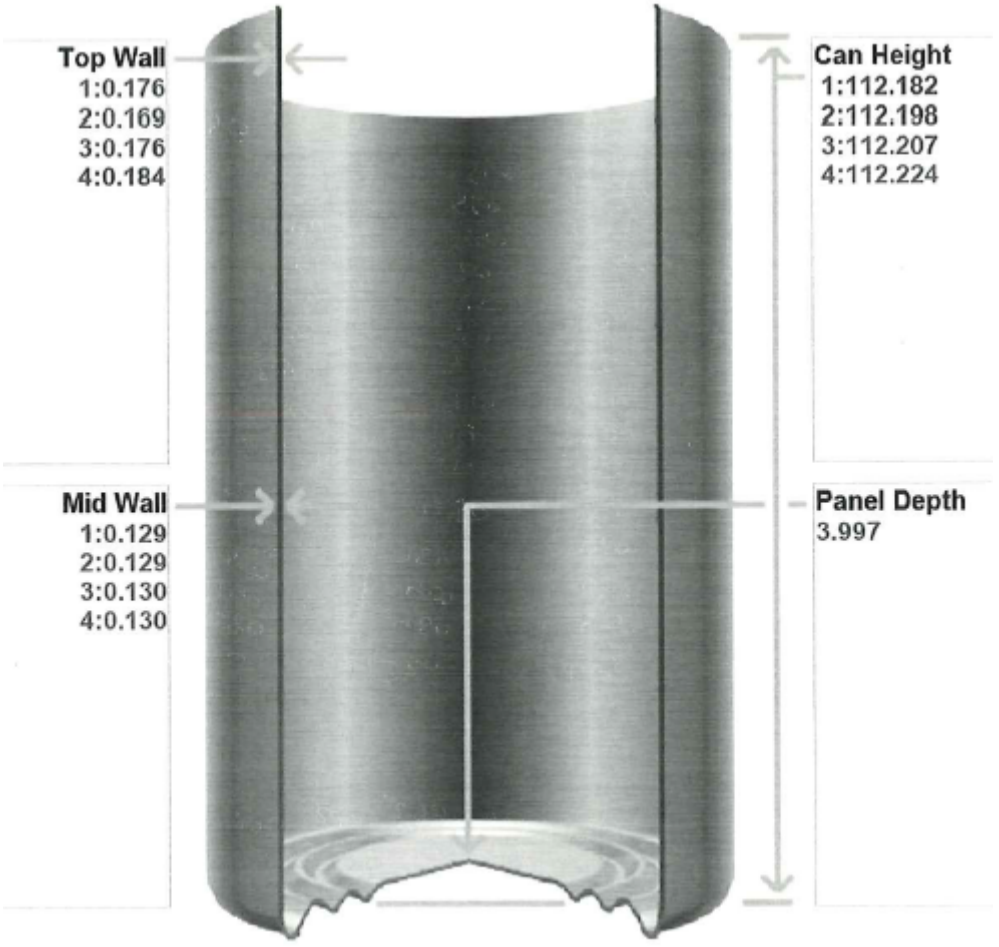

Figure 5.51: Depiction of measurements on a 2-piece metal food can after the front end of the manufacturing process

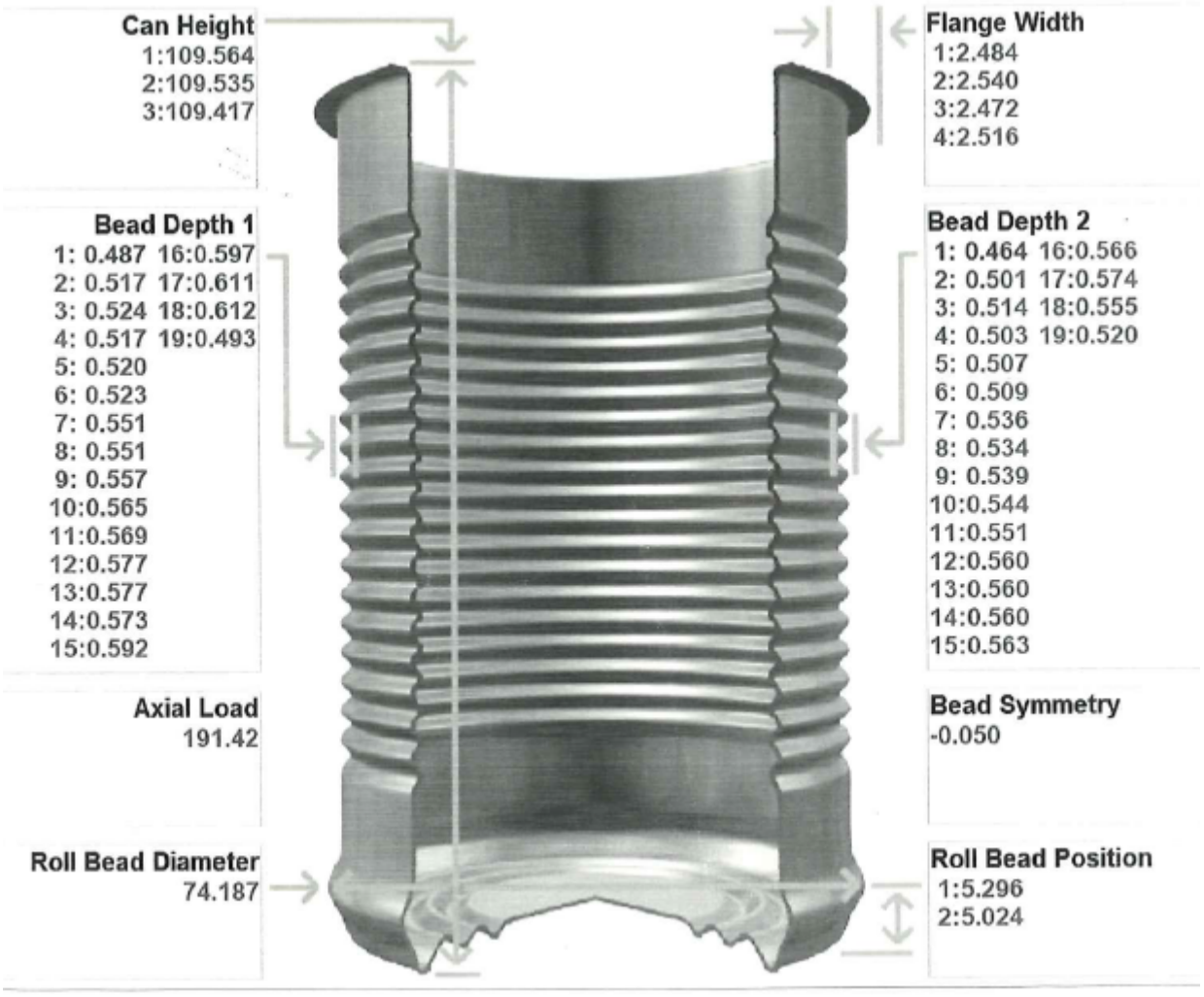

Figure 5.52: Depiction of measurements on a 2-piece metal food can after the beading process of manufacturing

From the initial visual and statistical analysis of the data, it seems that the factors that were most likely to influence the axial load resistance and the panelling pressure resistance were bead depths, factory finished can heights, mid-wall thickness of the cans, panel depths of the cans as well as the raw material supplier.

The visual and statistical data analyses were performed with the aid of Python, but do not include the applications of ML algorithms. The visual and statistical analysis is an important step to understand relationships in the data better and to visualize the data before ML is employed. **Section 5.8.2** and **Section 5.8.3** describes insights from feature selection and extraction as well as insights gained from the regression models. Both these methods were performed by using Python by applying ML

algorithms.

5.8.2 Insight from feature selection and extraction

Feature selection of the process data was done using SFS feature selection as well as random forest feature selection. Feature selection was done on the data table that transformed the categorical factors into numerical factors such as the raw material suppliers, the beader mandrel heads and the production teams. The data table consisted of 63 numerical factors as well as the response of either axial load resistance or panelling pressure resistance. The following factors were determined to be the most influential in predicting the outcome of axial load resistance or panelling pressure resistance:

- The bead depths.
- The mid-wall thickness.
- The beader mandrel heads.
- The flange widths.
- The roll bead position and diameter.
- The factory finished can height.
- The raw material supplier. There were three suppliers that were used during the time of data capture during the case study, they were Nippon, BAO Steel and Arcerol Mittal. No measurements related to the raw materials, such as chemical analyses, or yield strength were used as factors for the ML models, but only the categorical factor of supplier were added. The Suppliers were not compared to see which supplier gives the best response, but the supplier was simply added as a factor in a model to predict the response of axial load resistance.

From the list above it can be seen that those factors that were correlated to the response variables of axial load resistance and panel pressure resistance also, generally, were selected as the most important features in feature selection. The random forest feature selection algorithm indicates the importance of factors with a percentage value out of 100%, and the results showed that only the top 4 or 5 factors had more than 5% influence to predict the response. These top predictors were limited to bead depths and mid-wall thickness of the cans. Besides the average bead depth as a factor, there were specific beads that were more important predictors. Bead number 18 and 7 were the two most significant beads to predict the axial load resistance of the 2-piece food cans, and bead number 15, 16 and 17 were the most significant to predict panelling pressure resistance of the 2-piece food cans.

Feature extraction by means of LDA and PCA was performed on the data. With feature extraction, specific factors are not highlighted as the most important, but rather new principal factors are built

from the existing factors. These new factors are simply known as principal components or linear discriminants.

5.8.3 Insight from regression models

Various regression models were used on the feature selected and feature extracted data, and their accuracies were evaluated by means of MAE and RMSE. The following tables summarize the accuracies of the predictive ML models used in this case study for axial load resistance from either the feature selected data or the feature extracted data.

Refer to **Table 5.1** for a summary of the MAE and the RMSE of the regression models used to predict the axial load resistance with data from the main factors as selected by random forest feature selection.

Refer to **Table 5.2** for a summary of the MAE and the RMSE of the regression models used to predict the axial load resistance with data from the main factors as extracted by PCA.

Table 5.1: ML model accuracies to predict the axial load resistance of 2-piece metal food cans with the random forest selected features

Regression Model	MAE	RMSE
Simple Linear Regression	6.8	8.4
Multiple Linear Regression	5.7	7.3
LASSO Regression	5.4	6.9
Bayesian Ridge Regression	5.4	6.8
Support Vector Machine Regression	5.2	6.8
Simple Decision Tree Regression	7.0	9.1
Random Forest Regression	4.9	6.2
AdaBoost Regression	5.3	6.9
Gradient Boost Regression	4.6	6.0

Table 5.2: ML model accuracies to predict the axial load resistance of 2-piece metal food cans with the PCA extracted features

Regression Model	MAE	RMSE
Simple Linear Regression	6.6	8.2
Multiple Linear Regression	5.7	7.1
LASSO Regression	8.5	10.2
Bayesian Ridge Regression	8.5	10.2
Support Vector Machine Regression	6.7	8.5
Simple Decision Tree Regression	8.6	11.1
Random Forest Regression	6.2	7.9
AdaBoost Regression	7.0	8.9
Gradient Boost Regression	6.2	7.9

From the summarized accuracy tables the following can be seen:

- The regression models that used the feature selected data were generally more accurate than the regression models that used the feature extracted data.
- The ensemble regression models were generally more accurate than the penalized regression models and the linear regression models.
- Referring to **Appendix A.3** the models built to predict the axial load resistance, were generally more accurate than the models built to predict panelling pressure resistance.
- The best regression model in terms of predicting a response was the gradient boost regression model using the feature selected data from the random forest feature selection algorithm.
- The MAE was smaller as a standard deviation of the normal curve of the specification range of axial load resistance as illustrated in **Figure 5.2** for all the regression models that were attempted.
- The most inaccurate regression model was still lower than a standard deviation on the normal distribution of the specification range for axial load resistance of 2-piece metal food cans. For this reason it might be possible to consider ease of implementation of a model together with accuracy of the model when choosing which model to use.

5.8.4 Ways to improve the regression models

Improvement of the regression models will be manifested in improvements in the accuracy scores. The current best model can be further improved in various possible ways outlined as follows:

- The data volume could be increased.
- The various sub-sections all had separate data tables which could not be directly associated with each other i.e. the mid-wall thickness measured at the front end of the process and the axial load resistance measured at the end of the process will never be on the same can. The three sub section data tables were combined by using average values after each test occasion of the front end cans and the flanger cans with the actual values of the beaded cans. The model can possibly become more accurate if the same cans' measurements are used throughout the process from front end to beader.
- The bead depths are clearly a very important predictor for the model's responses of axial load and panelling pressure resistance. The bead depth is a dimensional measurement on the food can, but factors on the beading instrumentation is not built into the predictive model. Process understanding and control could be improved if data on factors related to the beader e.g. settings could be included in the model.
- Changing some parameters related to the various models for optimum performance can improve the predictive accuracy of those models.
- Updating the model whenever there is a change in the process which has not been included in the latest model e.g. a new raw material supplier, or a planned process change that would not fall within the normal running of the process.

5.8.5 Possible business improvements

The rationale for implementing a framework for process improvement using ML was to be able to predict quality characteristics of 2-piece metal food cans. Many quality or process problems in manufacturing is solved by an OFAT approach. OFAT, not only takes many resources and time, the solution it offers cannot necessarily be used for the next problem. The development of a framework where the product quality can be improved continuously in an iterative way by improving the ML predictive model, not only can save a lot of time and effort for future problem solving, but can also be used as a warning to predict when quality parameters are drifting too far away from nominal values.

The best current ML model has been proven to be gradient boost regression with random forest feature selected factors. The case study was approved by the manufacturing plant for 2-piece metal food cans because the potential for continuous process improvement using ML has been understood. The true effect of the continuous improvement framework using Six Sigma and data science principals will only be better understood after a metric, such as quality defects or process efficiency, is compared

before and after deployment. The deployment of such a ML model has not been completed as the model will still have to undergo some refining and deployment will require software engineering, which still need to be procured.

A model has been established that is able to predict the response variable of axial load resistance. The next step to reach the business objective of reducing the incoming tinplate thickness which will reduce the manufacturing cost of the cans and ultimately lower prices for Nampak's customers can be to built a trial suppliers' down gauged coil into the current model. The updated model will be able to predict whether the axial load resistance of the thinner material will still be acceptable.

5.9 DEPLOYMENT

Once the ML model has been developed up to the point where the predictions it makes are acceptably accurate and implementation was found to be financially viable, the model can be deployed. Deployment of the process improvement process is to integrate the ML model into the manufacturing environment. To integrate the model into a manufacturing environment the following will be needed from software engineering:

- Data will have to be accessed from the various manufacturing measuring devices. These measuring devices are the devices that measure the factors and response variables that form part of the ML algorithm.

- Data will have to be prepared to be in the correct format, scale and dimension to be used by the ML model.

- The ML algorithm will have to make a prediction that will have to be communicated to relevant people involved in the manufacturing process.

Heymann and Boza (n.d.) provides a guideline for the selection of suitable options for the deployment of ML models in manufacturing. Heymann and Boza (n.d.) states that the application of ML models to predict quality characteristics during the manufacturing process has a lot of potential. The deployment of such a ML model can be challenging due to technological and organizational difficulties. Five parameters are listed as can be seen in **Figure 5.53**.

Parameters	Values	
Prediction Approach	By batch	In real time
Consuming Application	Web app	Native app
Model Serving	Embedded	Separate
Learning Method	Offline	Online
Hosting Solution	On-premises	Cloud

Figure 5.53: : Deployment design of predictive machine Learning models in manufacturing (Heymann and Boza n.d.)

A batch approach to prediction from the ML model is when the model is based on the data that has been used to develop the model. A real time prediction approach is when predictions are made from the ML model with live data, therefore as new production data is fed into the ML model, the response is predicted immediately. For the purpose of process improvement in manufacturing of 2-piece metal food cans a real time prediction approach will be most suitable.

Syafrudin et al. (2018) proposes a real-time monitoring system that utilizes IoT (internet of Things) based sensors in an automotive industry. IoT sensors can be placed at strategic points in the manufacturing plant. The measured data from these sensors are transmitted to a data processor where data is processed. From the processor the data is stored and the processed data is sent through the ML model to make predictions. The outcome of the model is then communicated to relevant persons to act on if needed, see **Figure 5.54**.

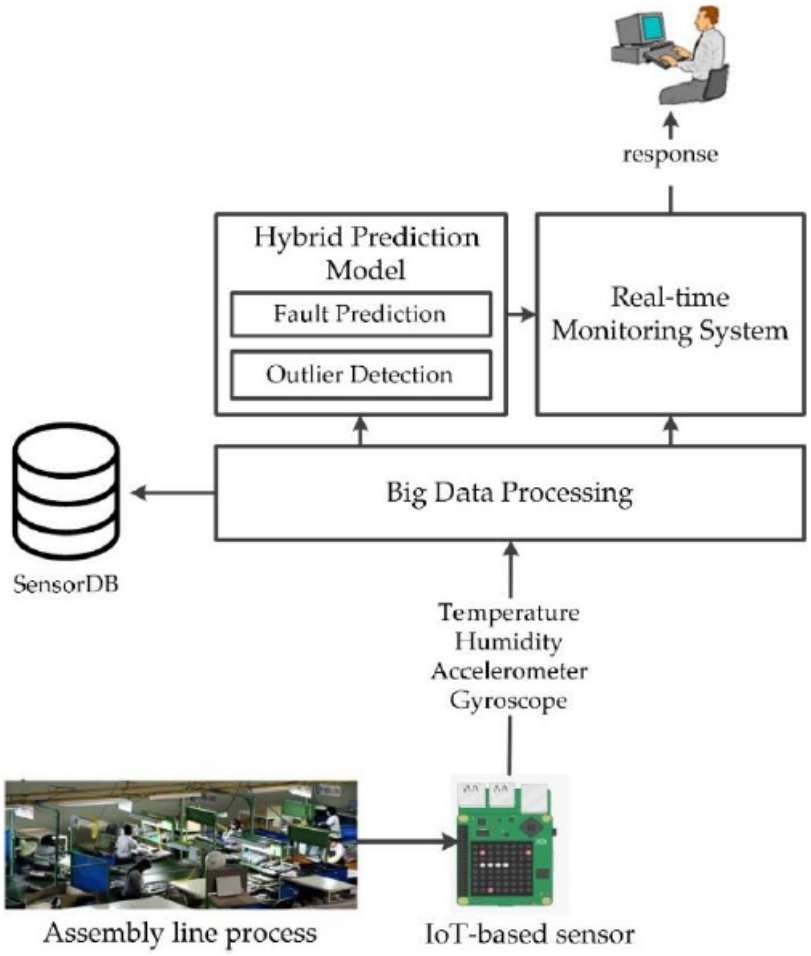

Figure 5.54: : Real-time monitoring in a manufacturing assembly line (Syafrudin et al. 2018)

Web based apps are apps that are accessed from the web, whereas native apps are apps that were developed for a specific platform. Native apps should first be installed by to be able to be used (Heymann and Boza n.d.).

The basic steps needed to build a web based ML application is to build the predictive ML model, create the web app by using a tool such as Flask or Django, and finally deploying the app on the web using a platform such as Herokum or the Google App Engine.

The ML model can be embedded in the application or can be separate. According to Ajani, Imoize, and Atayero (2021), there are trends in the merger of ML methods and embedded computing for ML applications. Ajani, Imoize, and Atayero (2021) summarizes embedded computing architectures for ML optimization, see **Figure 5.55**.

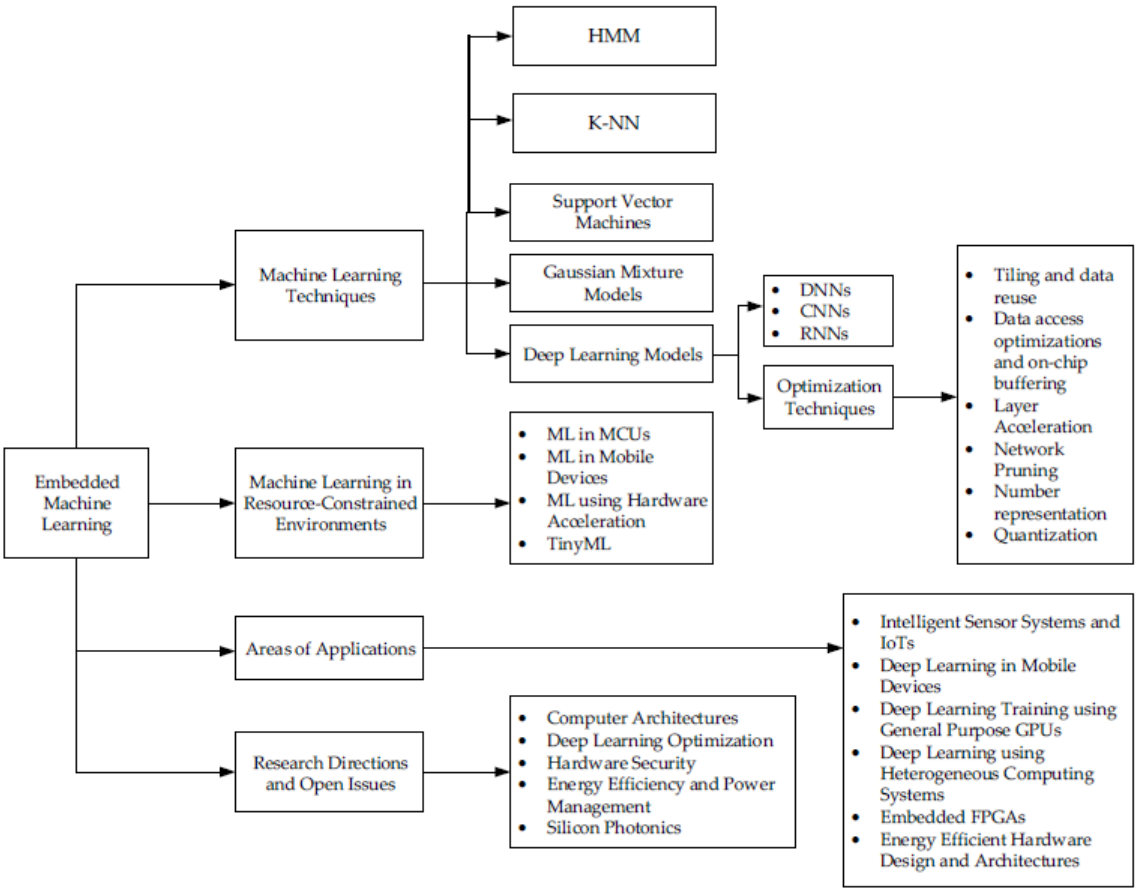

Figure 5.55: : Machine learning methods and embedded computing for machine learning applications (Ajani, Imoize, and Atayero 2021)

During the development stage of the ML modelling, the training of the model occurs offline. Once the ML model is deployed, the ML model can be trained online, in order to update the model predictions on a continuous basis. The online learning of a ML model can form part of an automated ML pipeline that uses tools such as Kubernetes or Docker.

Hosting of the model and data can be either on servers of the company that employs the model, or cloud hosting services such as Azure or Amazon can be used. Aspects to consider when deciding whether hosting will be done on the cloud or not, is security, costs, volume of data and the complexity of the model.

5.10 CONCLUSION

In this chapter of the thesis the case study was described in detail. The background to the development in the manufacturing plant that led up to the point where this case study's description starts was given. The case study was a validation of the application of ML to process improvement in the manufacturing of metal packaging.

The chapter described the 12 steps of the DUMED framework as applied to the case study. The objective of the project was established as the prediction of axial load resistance of 2-piece metal food cans during the manufacturing process. The rationale for this objective was the potential to establish a predictive model that will enable the manufacturer to maintain strength characteristics of the manufactured cans even when process variables such as bead depths and material thicknesses were changed. Such a model has the potential of financial benefits due to better quality control of the product as well as due to cost savings in material down gauging. To understand the process the manufacturing process flow for all the sub-sections of 2-piece can manufacturing was described in detail.

The next section in the chapter described everything related to the data that was used. Data was described; where the data was acquired from, how the data was cleaned and organized, and the basic statistics related to the collected data were given. Data was assessed by identifying the data types and searching for missing data. Data tables where combined into a single data table that was used in the following data steps as described in this case study. Further assessing of the data included finding the correlations in the data, performing ANOVA on the categorical data and visualizing the data in various manners. Finally the data was prepared for machine learning modelling. The preparation of the data entailed the scaling of data by standardization, as well as reduction of the dimensionality of the data by feature selection and feature extraction methods.

The predictive modelling of the data was described in the next section of the case study. Firstly, various data models that used different machine learning algorithms were developed. Both the feature selected data and the feature extracted data were run through these models. Each of these models were then assessed by using MAE and RMSE.

The results from the predictive models were evaluated. The best predictive model was the model that used the feature selected data that was run through a gradient boost ML algorithm. Various ways in which predictive capabilities of the model could be improved were discussed. Possible business improvement was discussed, but since the final deployment phase of the case study has not been completed yet, it was not possible to give definitive business improvements.

Finally, in this chapter, the deployment phase was discussed. Further improvements of the predictive model, as well as a more complete understanding of business improvements will form part of the next phase of the case study before the deployment phase will be completed.

CHAPTER 6

CONCLUSION

6.1 INTRODUCTION

Due to increased competitiveness in manufacturing generally and specifically in the packaging industry, process improvement is important to give businesses an edge over their competition. This dissertation specifically viewed the improvement of the manufacturing process of metal packaging by developing a framework for process improvement incorporating principals of data science. The contributions this study has, is to outline and demonstrate the use of a framework that uses machine learning for process improvement on a 2-piece metal food can manufacturing line. The use of machine learning on such a manufacturing line allows for an understanding of how variables can influence the process and the end product. In the case study the knowledge of how factors in the process influence the response can supply valuable information on the possible viability for light weighting of material. A further benefit of being able to understand how factors in the process influence the response is enhanced capability to control quality parameters throughout the process and in the final manufactured product.

6.2 METHODOLOGY

This thesis aimed to answer the research question of how ML can be applied to process improvement in metal packaging manufacturing. The objectives were listed to how the research problem could be solved which included a literature review, the design of a framework and the use of data from a metal packaging manufacturing in a case study. The rationale of the research was given for the development and demonstration of a process improvement framework at Nampak, Africa's largest packaging company. The research approach was given to how the stated objectives could be met.

The literature review chapters of Chapter 2 and Chapter 3, focused on metal packaging, process improvement, framework development and ML. Packaging and specifically metal packaging was reviewed, that included the uses of packaging, the manufacturing processes related to metals generally used in metal packaging, and the description of the process to manufacture tinplate cans. Process improvement methodologies were reviewed. The Lean Six Sigma process was discussed in detail as the process that contains the most complete approach towards process improvement. The main steps of the Six Sigma DMAIC process were defined, their uses highlighted and various tools that can be used for each step were described. Failures in the application of Six Sigma improvement projects in manufacturing were discussed as well. Various frameworks that used Six Sigma, DoE and data science were described and reviewed. The CRISP-DM framework used in data science and how it relates to the DMAIC framework was also described. Various decision mechanisms in process improvement was reviewed.

A review of ML was presented in a separate chapter. The ways in which the steps in the CRISP-DM

framework can be executed during projects were reviewed. ML was described in detail by showing the need for ML and listing and reviewing the various types of ML. The steps related to data and modelling when using ML in process improvement was described. Data cleaning, data exploration, data assessing and data preparation was described in detail. Various tools and the statistics involved with these data related steps in ML was reviewed. Some of the ML algorithms that can be used in predictive analytics was described.

After the literature reviews a framework was developed. The framework consisted of 5 main phases that incorporated 12 steps. The 5 phases were; Define, Understand, Model, Evaluate and Deploy (DUMED). The DUMED framework was based on the DMAIC steps that were used for the Six Sigma process improvement methodology as well as the CRISP-DM steps that were used in data science. The 12 steps of the DUMED framework were discussed and explanations were given to how to apply each step.

The 12 steps of the DUMED framework were applied to a real world case study:

- The objective of the project was established as the prediction of axial load resistance of 2-piece metal food cans during the manufacturing process.

- The rationale for this objective was the potential to establish a predictive model that will enable the manufacturer to maintain strength characteristics of the manufactured cans even when process variables such as bead depths and material thicknesses were changed. Such a model has the potential of financial benefits due to better quality control of the product as well as due to cost savings in material down gauging.

- The manufacturing process flow for all the sub-sections of 2-piece can manufacturing was described in detail.

- Data was described by establishing where the data was acquired from, how the data was cleaned and organized, and the basic statistics related to the collected data were given.

- Data was assessed by identifying the data types and showing missing data. Data tables where combined into a single data table. Further assessing of the data included showing correlations in the data by scatter plots, heat maps, correlation tables, performing ANOVA on the categorical data and visualizing the data by box plots and parallel coordinate plots.

- Data was prepared for machine learning modelling. The preparation of the data included the scaling of data by standardization. The reduction of the dimensionality of the data by feature selection and feature extraction methods is employed before modelling , but machine learning algorithms are used for feature selection and feature extraction. Feature selection methods used included SFS and random forest feature selection. Feature extraction methods used included LDA and PCA.

- Various data models that used different machine learning algorithms were developed. Both the feature selected data and the feature extracted data were run through these models. The regression algorithms used were, linear regression, penalized regression, support vector machines, decision tree regression and ensemble regression methods.

- Each of these regression models were then assessed by using MAE and RMSE. The results from the predictive models were evaluated to determine which model gave the most accurate predictions.

- The outcomes of the predictive models were evaluated. Various ways in which predictive capabilities of the model could be improved were discussed as well.

- Possible business improvements were discussed. The final deployment phase of the case study has not been completed yet, therefore it was not possible to give definitive business improvements.

- The deployment phase did not form part of this case study. Further improvements of the predictive model, as well as a more complete understanding of business improvements will form part of the next phase of the case study before the deployment phase will be completed.

6.3 RESULTS

From research in process improvement methods and framework development a framework for process improvement in metal packaging manufacturing was developed combining the Six Sigma and CRISP-DM philosophies. The main steps in this framework were; define, understand, model, evaluate, and deploy (DUMED).

The outcome of the case study can be described as follows:

- The framework steps were successfully applied in the case study, with the exception of the deployment phase. The deployment phase will be dependent on further improvement of the predictive model as well as the manufacturing plant's decision whether the project should continue.

- ML were successfully used in the case study, including data preparation and modelling. Axial load resistance as a response variable could be predicted within 2.3% of the actual values on average. The best results were obtained from using feature selected data obtained from a random forest feature selection algorithm. This data was modelled by using a gradient boost ensemble regression model.

6.4 BENEFITS OF THE CASE STUDY PROJECT FOR THE MANUFACTURING PLANT

The case study successfully demonstrated the implementing of a framework for process improvement using ML to predict quality characteristics of 2-piece metal food cans. The benefit of the development of a framework is the continuous improvement of product quality iteratively. In this case study the complex relationships between various factors in the process and the response variable of axial load resistance of 2-piece food cans were described with predictive ML models. The benefits of such a model in the manufacturing process are:

- The knowledge of how factors in the process influence the axial load resistance, which can supply valuable information on the possible viability for down gauging of materials.

- Enhanced capability to control quality parameters throughout the process and in the final manufactured product. As an example; the bead depth of a 2-piece metal food can has a strong influence on the axial load resistance. The ML model can allow an operator to first predict the associated axial load resistance with an adjustment in bead depth before actually changing the bead depth settings.

In both the stated benefits listed above the ability to better understand the relationship between the axial load resistance, and various factors in the manufacturing process for 2-piece metal food cans, can have a financial benefit for the manufacturer.

The true effect of the continuous improvement framework using Six Sigma and data science principals will only be better understood after a metric, such as quality defects or process efficiency, is compared before and after deployment. The deployment of such a ML model will be completed in the next phase of the project.

6.5 RECOMMENDATIONS AND FUTURE WORK

Recommendations for the improvement and completion of the case study project are:

- Even though the framework was successfully executed and it was demonstrated that quality characteristics in the manufacturing process of 2-piece metal food cans can be predicted, any further action will have a cost aspects associated with it. The current outcomes from the framework for process improvement should be discussed amid the stakeholders and decisions should be made to what the future actions with regards to the case study project should be.

- In the case study the evaluation phase of the framework has been completed. If it is decided that the project should continue, since the process is iterative, some steps in the framework can be revisited. Aspects related to data collection, data preparation, data modelling can be adjusted to gain better predictive results for responses.

- Once the predictive model is accepted by all the stakeholders, the model should be deployed. Deployment can be approached in various ways, but the result of such a deployment should be an output screen that can predict the desired quality characteristics of the product as a live feed on which relevant personnel can act as appropriate.

Possible opportunities for further work on this research include the following:

- Deployment of the process improvement project the case study describes.
- Extending the data assessing, data preparation, and data modelling to other datasets in the 2-piece metal food can manufacturing line at Nampak.
- Applying the DUMED framework on other suitable process improvement projects in metal packaging manufacturing.
- Doing trial work with different thickness gauges of material or new tinplate suppliers, or trial work related to process settings. Data from these trials could be used in the training of future ML models to enhance predictive capabilities of these predictive models.

6.6 CONCLUDING SUMMARY

The framework for process improvement in the manufacturing of metal packaging can be used to predict, and therefore also improve product quality. Aspects of the framework are iterative, which makes adjustments to improve predictive results of ongoing projects possible. Results pertaining to the case study could possibly be further improved with adjustments to data volumes and factors, as well as to adjustments in ML regression algorithms.

A final summary of the results was the successful development of a framework for process improvement using ML to predict quality characteristics of 2-piece metal food cans.

References

Adams, M. et al. (1999). "Simulation as a tool for continuous process improvement". In: *Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future-Volume 1*, pp. 766–773.

Ajani, T. S., Imoize, A. L., and Atayero, A. A. (2021). "An Overview of Machine Learning within Embedded and Mobile Devices–Optimizations and Applications". In: *Sensors* 21.13, p. 4412.

Albliwi, S. et al. (2014). "Critical failure factors of Lean Six Sigma: a systematic literature review". In: *International Journal of Quality & Reliability Management*.

Antony, J. (2001). "Improving the manufacturing process quality using design of experiments: a case study". In: *International Journal of Operations & Production Management*.

Aqlan, F. and Al-Fandi, L. (2018). "Prioritizing process improvement initiatives in manufacturing environments". In: *International Journal of Production Economics* 196, pp. 261–268.

Aven, T. (2015). *Risk analysis*. John Wiley & Sons.

Bekar, E. T., Nyqvist, P., and Skoogh, A. (2020). "An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study". In: *Advances in Mechanical Engineering* 12.5, p. 1687814020919207.

Bera, S. and Mukherjee, I. (2018). "A solution framework for response surface-based multiple quality characteristics optimization". In: *International Journal of Reliability, Quality and Safety Engineering* 25.05, p. 1850025.

Bezerra, M. A. et al. (2008). "Response surface methodology (RSM) as a tool for optimization in analytical chemistry". In: *Talanta* 76.5, pp. 965–977.

Borgovini, R., Pemberton, S., and Rossi, M. (1993). *Failure mode, effects, and criticality analysis (FMECA)*. Tech. rep. RELIABILITY ANALYSIS CENTER GRIFFISS AFB NY.

Bowles, M. (2019). *Machine Learning with Spark and Python: Essential Techniques for Predictive Analytics*. John Wiley & Sons.

Buer, S.-V., Fragapane, G. I., and Strandhagen, J. O. (2018). "The data-driven process improvement cycle: Using digitalization for continuous improvement". In: *IFAC-PapersOnLine* 51.11, pp. 1035–1040.

Cameron, I. T. and Hangos, K. (2001). *Process modelling and model analysis*. Elsevier.

Cao, Y. et al. (2015). "Constructing the integrated strategic performance indicator system for manufacturing companies". In: *International Journal of Production Research* 53.13, pp. 4102–4116.

Chen, C.-h., Härdle, W. K., and Unwin, A. (2007). *Handbook of data visualization*. Springer Science & Business Media.

Cheng, T. and Podolsky, S. (1996). *Just-in-time manufacturing: an introduction*. Springer Science & Business Media.

Cunningham, P., Cord, M., and Delany, S. J. (2008). "Supervised learning". In: *Machine learning techniques for multimedia*. Springer, pp. 21–49.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Deisenroth, M. P., Faisal, A. A., and Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

Durivage, M. A. (2015). *Practical Attribute and Variable Measurement Systems Analysis (MSA): A Guide for Conducting Gage R&R Studies and Test Method Validations*. Quality Press.

Emblem, A. (2012). *Packaging technology: Fundamentals, materials and processes*. Elsevier.
Fellows, P., Axtell, B., et al. (1993). *Appropriate food packaging*. TOOL publications.
Al-Ghamdi, K. A. (2011). "Improving the practice of experimental design in manufacturing engineering." PhD thesis. University of Birmingham.
Gijo, E. and Scaria, J. (2014). "Process improvement through Six Sigma with Beta correction: a case study of manufacturing company". In: *The International Journal of Advanced Manufacturing Technology* 71.1, pp. 717–730.
Gupta, M. C. and Boyd, L. H. (2008). "Theory of constraints: a theory for operations management". In: *International Journal of Operations & Production Management*.
Hackman, J. R. and Wageman, R. (1995). "Total quality management: Empirical, conceptual, and practical issues". In: *Administrative science quarterly*, pp. 309–342.
Haq, A. U. et al. (2019). "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection". In: *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. IEEE, pp. 1–4.
Heymann, H. and Boza, A. (n.d.). "Concept for Deployment Design of Machine Learning Models in Production". In: *XXV Congreso de Ingenieri a de Organizacion*, p. 93.
Hulamin (2021). *Aluminium Food Packaging*. https://www.hulamin.com/aluminium-food-packaging.
Isniah, S., Purba, H. H., Debora, F., et al. (2020). "Plan do check action (PDCA) method: literature review and research issues". In: *Jurnal Sistem dan Manajemen Industri* 4.1, pp. 72–81.
Keller, P. (2011). *Six Sigma Demystified®*. McGraw-Hill Education.
Khanbabaei, M. et al. (2018). "Developing an integrated framework for using data mining techniques and ontology concepts for process improvement". In: *Journal of Systems and Software* 137, pp. 78–95.
Klosterman, S. (2019). *Data Science Projects with Python: A Case Study Approach to Successful Data Science Projects Using Python, Pandas, and Scikit-Learn*. eng. Birmingham: Packt Publishing, Limited. ISBN: 9781838551025.
Manyika, J. et al. (2012). *Manufacturing the future: The next era of global growth and innovation*. URL: http://https://www.mckinsey.com/business-functions/operations/our-insights/the-future-of-manufacturing/ (visited on 09/18/2021).
Marsal, P. (Nov. 1988). *The Can and its Uses. Part 1*.
Mauri, F., Garetti, M., and Gandelli, A. (2010). "A structured approach to process improvement in manufacturing systems". In: *Production Planning & Control* 21.7, pp. 695–717.
McLean, R. S., Antony, J., and Dahlgaard, J. J. (2017). "Failure of Continuous Improvement initiatives in manufacturing environments: a systematic review of the evidence". In: *Total Quality Management & Business Excellence* 28.3-4, pp. 219–237.
Mileham, T. (2007). "Essentials of lean six sigma". In: *Proceedings of the Institution of Mechanical Engineers* 221.B8, p. 1375.
Montgomery, D. C. (1999). "Experimental design for product and process design and development". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 48.2, pp. 159–177.
— (2017). *Design and analysis of experiments*. John wiley & sons.

Nampak (2021). *About Nampak*. URL: http://www.nampak.com/About/Overview (visited on 09/19/2021).

Oakland, J. S. (2007). *Statistical process control*. Routledge.

Packaging SA, T. I. of (N.D. 2019). *A handbook of packaging technology*.

PackagingSA (June 2018). *SA Metal Packaging Industry Poised for Growth*. URL: http://www.packagingsa.co.za/2018/06/18/sa-metal-packaging-industry-poised-for-growth/ (visited on 09/19/2021).

Page, B., Edwards, M., and May, N. (2006). "5 Metal cans". In.

Panizzolo, R. et al. (2012). "Lean manufacturing in developing countries: evidence from Indian SMEs". In: *Production Planning & Control* 23.10-11, pp. 769–788.

Parab, P. A. and Shirodkar, V. A. (2019). "Value stream mapping: A case study of lock industry". In: *AIP Conference Proceedings*. Vol. 2148. 1. AIP Publishing LLC, p. 030041.

Plain, C. (2007). "Build an affinity for KJ method". In: *Quality Progress* 40.3, p. 88.

Prashar, A. (2016). "A conceptual hybrid framework for industrial process improvement: integrating Taguchi methods, Shainin System and Six Sigma". In: *Production Planning & Control* 27.16, pp. 1389–1404.

Pyzdek, T. and Keller, P. (2018). *The Six Sigma Handbook. 5-th Ed*.

Raschka, S. (2015). *Python machine learning*. Packt publishing ltd.

RecyclingInternational (N.D.). *Latest Recycling Stats for South Africa*. https://recyclinginternationa com/business/latest-recycling-stats-for-south-africa/27327/.

Robertson, G. (2013). "Legislative and Safety Aspects of food packaging". In: *Food Packaging: Principles and Practice. 3ª edição, CRC PRESS. Boca Raton*.

Schonlau, M. and Zou, R. Y. (2020). "The random forest algorithm for statistical learning". In: *The Stata Journal* 20.1, pp. 3–29.

Shaffie, S. and Shahbazi, S. (2012). *McGraw-hill 36-hour course: Lean six sigma*. McGraw-Hill Education.

Shankar, R. (2009). *Process improvement using six sigma: a DMAIC guide*. Quality Press.

Simal-Gándara, J. (1999). "Selection of can coatings for different applications". In: *Food reviews international* 15.1, pp. 121–137.

Singh, J., Singh, H., and Pandher, R. P. S. (2017). "Role of DMAIC Approach in Manufacturing Unit: A Case Study." In: *IUP Journal of Operations Management* 16.4.

Snee, R. D. (2010a). "Crucial considerations in monitoring process performance and product quality". In: *Pharmaceutical Technology* 34.10, pp. 38–40.

— (2010b). "Lean Six Sigma–getting better all the time". In: *International Journal of Lean Six Sigma*.

Subasi, A. (2020). *Practical Machine Learning for Data Analysis Using Python*. Academic Press.

Syafrudin, M. et al. (2018). "Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing". In: *Sensors* 18.9, p. 2946.

Tanco, M. et al. (2007). "Manufacturing industries need Design of Experiments (DoE)." In: *World Congress on Engineering*. Vol. 20, pp. 1108–1113.

Teng, S. Y. et al. (2019). “Principal component analysis-aided statistical process optimisation (PASPO) for process improvement in industrial refineries”. In: *Journal of Cleaner Production* 225, pp. 359–375.

White, S. K. (Aug. 2019). *What is process improvement? A business methodology for efficiency and productivity*. `https://www.cio.com/article/3433946/what-is-process-improvement-a-business-methodology-for-efficiency-and-productivity.html`.

Wilson, C. (2013). *Brainstorming and beyond: a user-centered design method*. Newnes.

Wuest, T., Irgens, C., and Thoben, K.-D. (2014). “An approach to monitoring quality in manufacturing using supervised machine learning on product state data”. In: *Journal of Intelligent Manufacturing* 25.5, pp. 1167–1180.

Zwetsloot, I. M. et al. (2018). “Lean Six Sigma meets data science: Integrating two approaches based on three case studies”. In: *Quality Engineering* 30.3, pp. 419–431.

APPENDIX A

CASE STUDY FOR PREDICTING PANELLING PRESSURE RESISTANCE

A.1 INTRODUCTION

The panelling pressure resistance is a dependent variable and the output is given as real numbers. The panelling pressure resistance is measured in kPa, with a specified range of 155kPa to 205kPa. Generally the panelling pressure resistance are within these specifications and generally there are no problems in terms of process capabilities for this quality measurement. The objective is therefore not focused on improving this response, but rather to understand how this response relate to other factors in the process with the aim to that any changes made in process improvement will still deliver cans with suitable panelling pressure resistance.

	Beader_flange_width_range	Beader_flange_width_average	Beaded_can_height_range	Beaded_can_height_average	Roll_bead_diameter	Roll_bead_position_range	Roll_bead_position_average	Bead-Symmetry	Panel_resistance
count	3311.000000	3311.00000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000
mean	0.062510	2.53165	0.161422	109.492454	74.224074	0.221008	5.173802	-0.030826	191.782845
std	0.030094	0.03513	0.079723	0.086636	0.066730	0.137655	0.122082	0.019312	15.026350
min	0.004000	2.36600	0.003000	109.166667	73.913000	0.000000	4.671500	-0.106000	82.000000
25%	0.041000	2.51000	0.103000	109.434667	74.179000	0.111000	5.093500	-0.043000	183.000000
50%	0.058000	2.53375	0.156000	109.497000	74.224000	0.208000	5.181500	-0.029000	192.000000
75%	0.081000	2.55525	0.212000	109.548167	74.269000	0.317000	5.259500	-0.019000	202.000000
max	0.307000	2.65125	0.517000	109.799667	74.450000	0.747000	5.596500	0.030000	239.000000

Bead1_depth	Bead2_depth	Bead3_depth	Bead4_depth	Bead5_depth	Bead6_depth	Bead7_depth	Bead8_depth	Bead9_depth	Bead10_depth	Bead11_depth	Bead12_depth	Bead13_depth	Bead14_depth	Bead15_depth	Bead16_depth
3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000	3311.000000
0.496279	0.523080	0.533011	0.526927	0.531795	0.535359	0.569965	0.572781	0.581376	0.590718	0.598366	0.606396	0.612526	0.613097	0.613918	0.625231
0.015891	0.025121	0.025315	0.025272	0.025999	0.026212	0.026449	0.027134	0.027592	0.027516	0.027727	0.026178	0.027796	0.026249	0.027197	0.027640
0.433500	0.439000	0.451500	0.449500	0.444500	0.431000	0.491000	0.497500	0.506000	0.512000	0.521500	0.524000	0.530000	0.531000	0.535500	0.537000
0.489500	0.514000	0.523000	0.518000	0.520000	0.521000	0.557000	0.558500	0.566500	0.575000	0.583000	0.590000	0.597000	0.596500	0.598500	0.611000
0.496000	0.526500	0.535500	0.531000	0.533000	0.537000	0.569000	0.570500	0.579000	0.588000	0.595500	0.603500	0.609500	0.611500	0.612000	0.624000
0.506000	0.537000	0.546500	0.542000	0.545000	0.549500	0.581500	0.585000	0.594500	0.604750	0.612000	0.621500	0.627500	0.627000	0.625500	0.638500
0.537000	0.576000	0.588000	0.585000	0.592500	0.607000	0.642500	0.648000	0.659500	0.671000	0.679000	0.688000	0.694500	0.690500	0.694000	0.710500

Bead17_depth	Bead18_depth	Bead19_depth	Bead_depth_range	Bead_depth_avergae
3311.000000	3311.000000	3311.000000	3311.000000	3311.000000
0.644103	0.641485	0.555665	0.149876	0.577583
0.026429	0.029194	0.045076	0.017634	0.025837
0.562000	0.556000	0.430500	0.098500	0.500816
0.629000	0.624500	0.526000	0.137500	0.565750
0.643500	0.639500	0.553500	0.146500	0.577000
0.658000	0.657000	0.581000	0.159000	0.588092
0.721000	0.720500	0.700000	0.214500	0.638684

Figure A.1: : Basic Statistics table for beader data including panel pressure resistance data of 2-piece metal food can manufacturing process

A.2 DATA PRE-PROCESSING

A.2.1 Data pre-processing

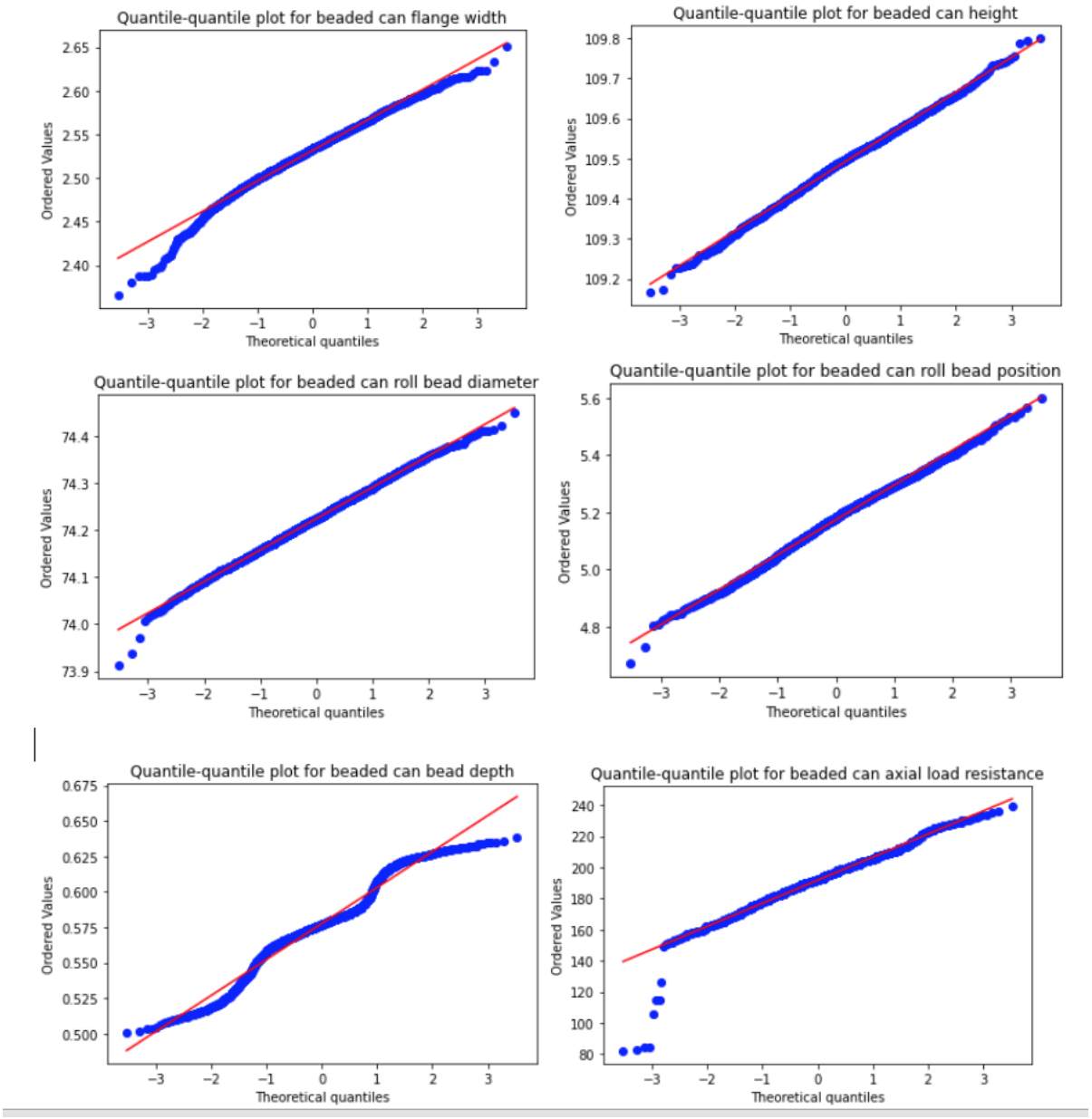

Figure A.2: : Quentile-quentile plots for the beader of 2-piece metal food cans

A.2.2 Data assessing

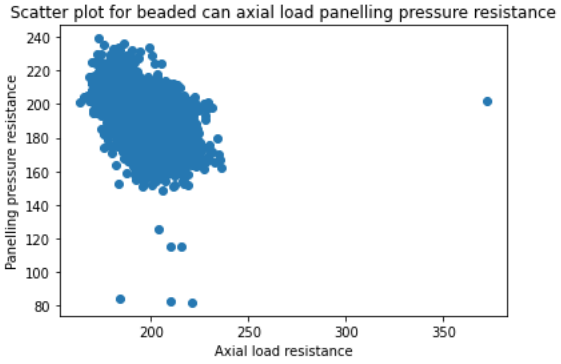

Figure A.3: : Scatter plot for beaded can axial load panelling pressure resistance

ANOVA was performed on beader mandrel numbers, production teams and raw material suppliers as categories. The numerical factors used for the ANOVAs were the response factor of panelling pressure resistance. random samples of 500 instances were drawn and the steps involved in ANOVA were performed on the data and the hypotheses were accepted or rejected. The null hypotheses for axial load resistance was rejected for both the beader mandrels as well as the raw material suppliers. This indicated that

- all the different beader mandrels did not give statistically similar responses, therefore the null hypothesis was rejected.
- all the different raw material suppliers did not give statistically similar responses, therefore the null hypothesis was rejected.
- all the different production teams did give statistically similar responses, therefore the null hypothesis was accepted.

The outputs of the Python code for the categorical variables related to the panelling pressure resistance can be seen in **Figure A.4**.

ANOVA for beader mandrels 1 to 16 in terms of paneling pressure resistance

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	92732.8	15	6182.18	35.5515	1.11022e-16	1.84158
Within Groups	256320	1474	173.894			
Total	349052	1489	234.421			

```
Paneling pressure resistance ANOVA
Approach 1: The p-value approach to hypothesis testing in the decision rule
F-score is: 35.55145740348227  and p value is: 1.1102230246251565e-16
Null Hypothesis is rejected.
------------------------------------------------------------------------
Approach 2: The critical value approach to hypothesis testing in the decision
rule
F-score is: 35.55145740348227  and critical value is: 1.8415774338540256
Null Hypothesis is rejected.
```

ANOVA for production teams in terms of paneling pressure resistance

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	3638.94	3	1212.98	5.21244	0.001395	3.1249
Within Groups	345805	1486	232.709			
Total	349444	1489	234.684			

```
Paneling pressure resistance ANOVA
Approach 1: The p-value approach to hypothesis testing in the decision rule
F-score is: 5.212440761180212  and p value is: 0.0013950011173864407
Null Hypothesis is rejected.
------------------------------------------------------------------------
Approach 2: The critical value approach to hypothesis testing in the decision
rule
F-score is: 5.212440761180212  and critical value is: 3.1249029889623077
Null Hypothesis is rejected.
```

ANOVA for raw material supplier in terms of paneling pressure resistance

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	7551.36	2	3775.68	16.4216	8.8267e-08	3.69805
Within Groups	341893	1487	229.921			
Total	349444	1489	234.684			

```
Paneling pressure resistance ANOVA
Approach 1: The p-value approach to hypothesis testing in the decision rule
F-score is: 16.421624118063022  and p value is: 8.82670072588354e-08
Null Hypothesis is rejected.
------------------------------------------------------------------------
Approach 2: The critical value approach to hypothesis testing in the decision
rule
F-score is: 16.421624118063022  and critical value is: 3.698045805614752
Null Hypothesis is rejected.
```

Figure A.4: : ANOVA results for beader mandrel numbers, raw material suppliers and production teams in relation to panelling resistance

The null hypotheses were rejected for the raw material suppliers as well as the beader mandrels, which suggested that these two categorical values should be incorporated as numerical categories in

the data frame that will be used for predictive modelling.

A.2.3 Data preparation

Figure A.5 shows the performance of the top 20 factors to predict the panelling pressure resistance of a 2-piece metal food can manufacturing line using SFS.

The top 10 factors that would maximize the performance of a predictive model to predict panelling pressure resistance according to the SFS algorithm were:

- Bead depth average of the 19 beads on a beaded can
- Front end mid-wall thickness
- Nippon as raw material supplier of the tinplate
- The panel depth of the cans after the front end
- Mandrel head number 1 at the beader
- Mandrel head number 6 at the beader
- Mandrel head number 2 at the beader
- Mandrel head number 10 at the beader
- The flange width on the cans after the flanger
- Mandrel head number 7 at the beader

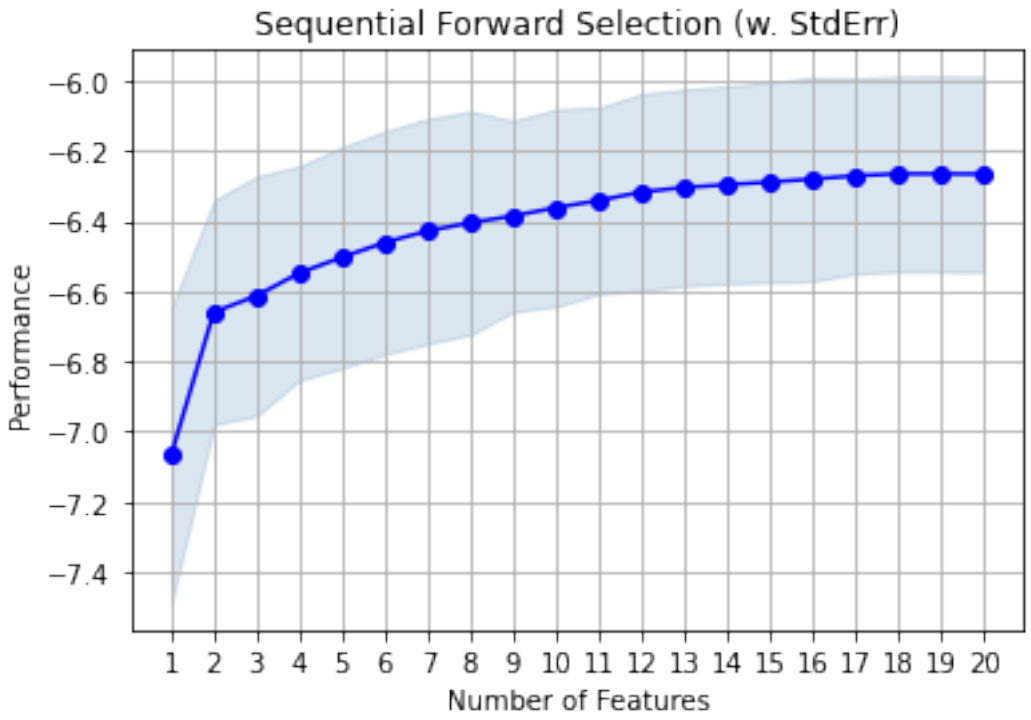

Figure A.5: : Performance of the top 20 factors to predict the panelling pressure resistance of a 2-piece metal food can manufacturing line using SFS

Figure A.6 shows a snippet of an example of a random forest regression tree to determine the most important factors to predict panelling pressure resistance of a 2-piece metal food can during manufacturing.

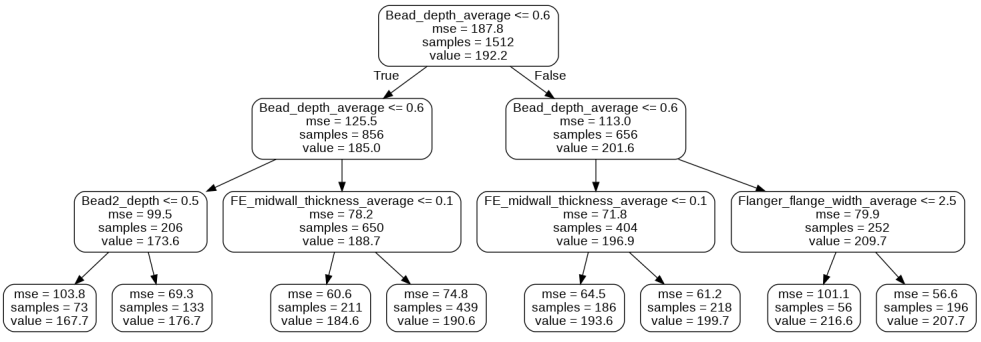

Figure A.6: : An example of a section of a random forest regression tree to determine the most important factors to predict panel pressure resistance

The top 10 factors that would maximize the performance of a predictive model to predict panelling

pressure resistance according to the random forest algorithm and their relative importance can be seen in **Figure A.7**

```
Variable: Bead_depth_average   Importance: 0.32
Variable: Bead15_depth         Importance: 0.13
Variable: Bead16_depth         Importance: 0.06
Variable: Bead7_depth          Importance: 0.05
Variable: Bead17_depth         Importance: 0.05
Variable: FE_midwall_thickness_average Importance: 0.03
Variable: Flanger_flange_width_average Importance: 0.03
Variable: FE_Can_height_average Importance: 0.02
Variable: Bead14_depth         Importance: 0.02
Variable: FE_Can_height_range  Importance: 0.01
```

Figure A.7: : The top 10 factors that would maximize the performance of a predictive model to predict panelling pressure resistance according to the random forest algorithm

From **Figure A.7** it can be seen that, as was the case for the axial load resistance, only the top 9 most important factors have an importance of more than 1%.

A.2.4 Regression

A.2.4.1 Linear regression to predict panelling pressure

in **Section 5.6.3.2** the most important features were determined by random forest feature selection. The different bead depths were generally the most important factors and therefore the average bead depth was used as the dependent variable to predict panelling pressure resistance. **Figure A.8** shows a graph of the scatter plot of panelling pressure resistance vs. average bead depths.

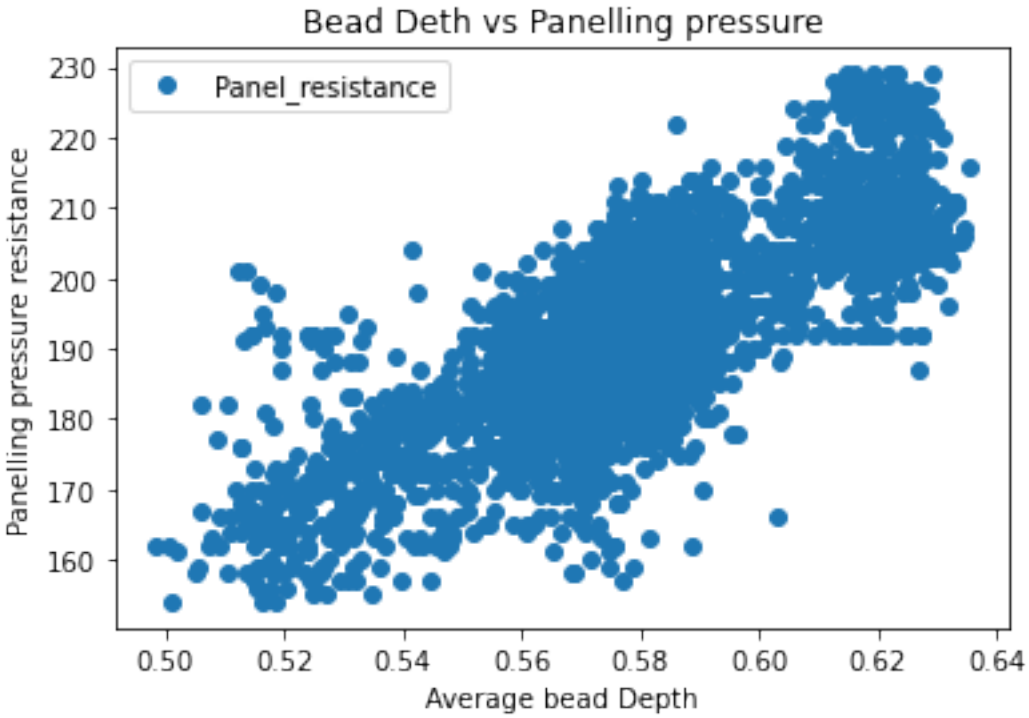

Figure A.8: Scatter plot of panelling pressure resistance vs. average bead depth of 2-piece metal food cans

The data was split between a test set and a train set at a ratio of 80% / 20%. The training data was used to train the linear regression model and the output was a regressor intercept (w_0) of -48.84316990035731 and a regressor coefficient (w_1) of 417.12182232. The linear regression model was used to predict the panelling pressures of the test data set, see **Figure A.9** for some of the predicted values.

	Actual	Predicted
0	189	189.300460
1	179	188.565008
2	189	184.284021
3	190	188.641846
4	195	189.838327
...	...	...
631	205	212.099460
632	196	192.944788
633	199	193.087487
634	201	198.279556
635	193	188.367425

Figure A.9: Linear regression model's predictions of panelling pressure resistance of 2-piece metal food cans

The accuracy of the linear regression model was determined by calculating the mean absolute error (MAE) as well as the root mean squared error (RMSE) of the predicted axial load resistance when compared to the actual axial load Resistance.

The MAE was 6.957170909402181 and the RMSE was 8.799220359011098 for the linear regression model.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first component was used to predict the panelling pressure resistance by means of a linear regression model The MAE was 7.125483350031001 and the RMSE was 8.240349335477788.

The PCA data accomplished accuracy scores that were slightly poorer for linear regression when compared to the data from the feature selection data table.

A.2.4.2 Multiple linear regression to predict panelling pressure

A multiple linear regression model was built to predict panelling pressure resistance with average bead depth, front end mid-wall thickness, flange width as well as material supplier as independent variables. **Figure A.10** shows the coefficients (w_1, w_2, w_3, w_4) for the multiple linear regression model to predict panelling pressure resistance in a 2-piece metal food can manufacturing line.

	Coefficient
FE_midwall_thickness_average	1795.988399
Flanger_flange_width_average	-51.161666
Bead_depth_average	453.770086
Nippon	1.190141

Figure A.10: Coefficients for the multiple linear regression model to predict panelling pressure resistance of a 2-piece metal food can

The multiple linear regression model was used to predict the panelling pressures of the test data set, see **Figure A.11** for some of the predicted values.

	Actual	Predicted
1093	189	190.566362
641	179	189.639090
1554	189	181.223196
575	190	187.185869
117	195	194.539028
...	...	...
2289	205	211.615719
529	196	192.555257
1292	199	197.403500
900	201	202.697933
1648	193	188.635564

Figure A.11: Multiple regression model's predictions of panelling pressure resistance of 2-piece metal food cans

The accuracy of the multiple regression model was determined by calculating the mean absolute error (MAE) as well as the root mean squared error (RMSE) of the predicted panelling pressure resistance when compared to the actual axial load Resistance.

The MAE was 6.354400732891439 and the RMSE was 8.942662462123007 for the linear regression model. Both the MAE as well as the RMSE showed improvements for the multiple linear regression model when compared to the linear regression model.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first four components were used to predict the panelling pressure resistance by means of a multiple linear regression model The MAE was 6.53379984006022 and the RMSE was 8.629284536419815.

The PCA data accomplished accuracy scores that were slightly better for multiple linear regression when compared to the data from the feature selection data table.

A.2.4.3 LASSO regression to predict panelling pressure

A LASSO regression model was built to predict axial load resistance with all the factors from the feature selection data table as described in **Section 5.6.3.2** as independent variables. The LASSO regression model was used to predict the panelling pressure resistance of the test data set, see **Figure A.12** for some of the predicted values and **Figure A.13** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	189	192.043853
641	179	188.287742
1554	189	180.975213
575	190	186.939810
117	195	194.162389
...	...	...
2289	205	209.132530
529	196	193.288172
1292	199	198.278100
900	201	202.298880
1648	193	189.156526

Figure A.12: LASSO regression model's predictions of panelling pressure resistance of 2-piece metal food cans

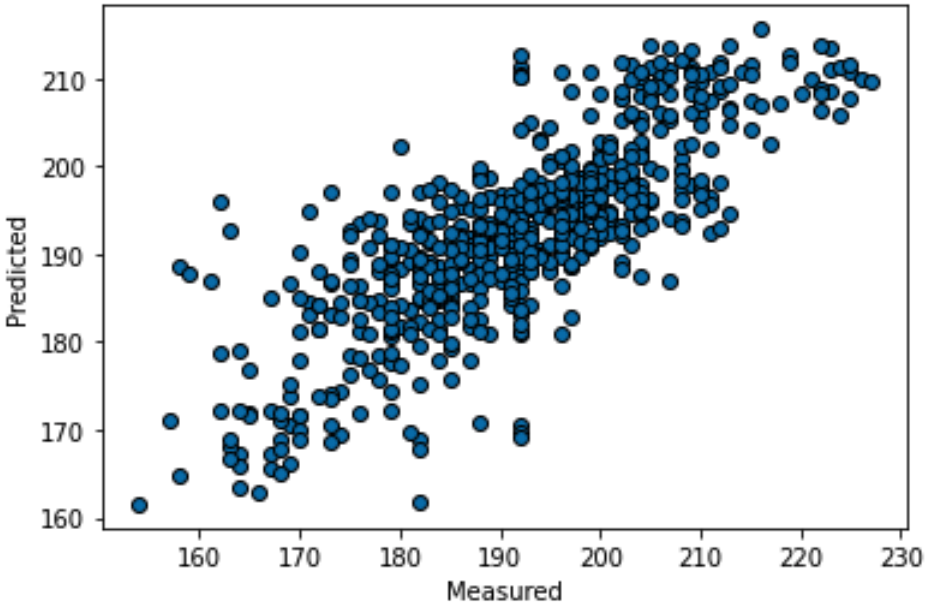

Figure A.13: Graph depicting the measured values vs. the actual values of a LASSO regression model for the panelling pressure resistance of 2-piece metal food cans

The MAE was 6.172260183091495 and the RMSE was 8.147936174863549 for the LASSO regression model. Both the MAE as well as the RMSE showed improvements for the LASSO regression model when compared to both the linear regression models.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the panelling pressure resistance by means of a LASSO regression model, the MAE was 8.95169665680212 and the RMSE was 11.217824383670393.

The PCA data accomplishes accuracy scores that were less accurate when compared with the feature selection data for both the linear regression and multiple linear regression model. The PCA accuracy scores were also less accurate as was accomplished for the LASSO regression when the feature selection data was used.

A.2.4.4 Bayesian regression to predict panelling pressure

A Bayesian ridge regression model was built to predict panelling pressure resistance with all the factors from the feature selection data table as described in **Section 5.6.3.2** as independent variables. The Bayesian ridge regression model was used to predict the panelling pressure resistance of the test data set, see **Figure A.14** for some of the predicted values vs. the actual values.

	Actual	Predicted
1093	189	191.993792
641	179	187.905409
1554	189	180.872695
575	190	186.791045
117	195	194.449857
...	...	...
2289	205	209.195675
529	196	193.145236
1292	199	198.295450
900	201	202.567829
1648	193	189.378436

Figure A.14: Bayesian ridge regression model's predictions of panelling pressure resistance of 2-piece metal food cans

The MAE was 6.162459793188117 and the RMSE was 8.129983182143098 for the Bayesian ridge regression model. Both the MAE as well as the RMSE were similar for the Bayesian ridge regression model when compared to the LASSO regression model.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 principal components were used to predict the panelling pressure resistance by means of a Bayesian ridge regression model, the MAE was 8.956691303856438 and the RMSE was 11.223675026074288. The PCA data accomplishes accuracy scores that were similar for the Bayesian ridge regression model when compared to the LASSO regression model.

A.2.4.5 SVM regression to predict panelling pressure

An SVM regression model was built to predict panelling pressure resistance with all the factors from the feature selection data table as described in **Section 5.6.3.2** as independent variables. The SVM regression model was used to predict the panelling pressure resistance of the test data set, see **Figure A.15** for some of the predicted values and **Figure A.16** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	189	194.461701
641	179	187.765941
1554	189	186.176877
575	190	185.894990
117	195	198.235668
...	...	...
983	168	193.191336
910	206	204.593821
1311	188	203.084458
215	196	191.009655
2870	180	183.731710

Figure A.15: SVM regression model's predictions of panelling pressure resistance of 2-piece metal food cans

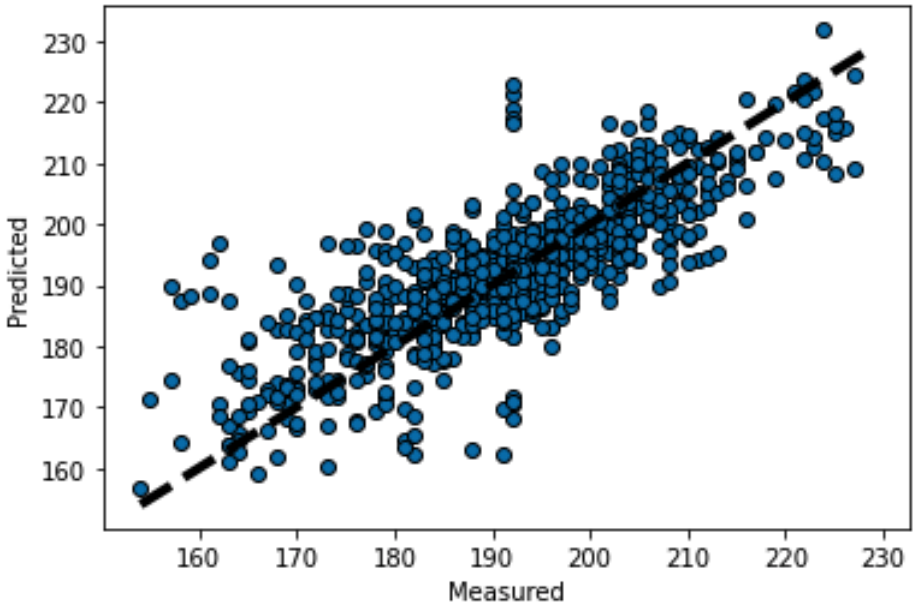

Figure A.16: Graph depicting the measured values vs. the actual values of a SVM regression model for the panelling pressure resistance of 2-piece metal food cans

The MAE was 6.064632993832985 and the RMSE was 8.175375732097228 for the SVM regression model. Both the MAE as well as the RMSE showed improvements for the SVM regression model

when compared to the linear regression as well as penalized regression models.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the panelling pressure resistance by means of a SVM regression model, the MAE was 7.1739485668706715 and the RMSE was 9.525022815748459. The PCA data accomplishes accuracy scores that were less accurate then the feature selection data for the SVM regression model.

A.2.4.6 Decision tree regression to predict panelling pressure

A simple decision tree regression model was built to predict panelling pressure resistance with all the factors from the feature selection data table as independent variables. The decision tree regression model was used to predict the panelling pressures of the test data set, see **Figure A.17** for some of the predicted values and **Figure A.18** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	189	193.0
641	179	195.0
1554	189	165.0
575	190	185.0
117	195	207.0
...	...	...
983	168	178.0
910	206	186.0
1311	188	204.0
215	196	185.0
2870	180	168.0

Figure A.17: Simple decision tree regression model's predictions of panelling pressure resistance of 2-piece metal food cans

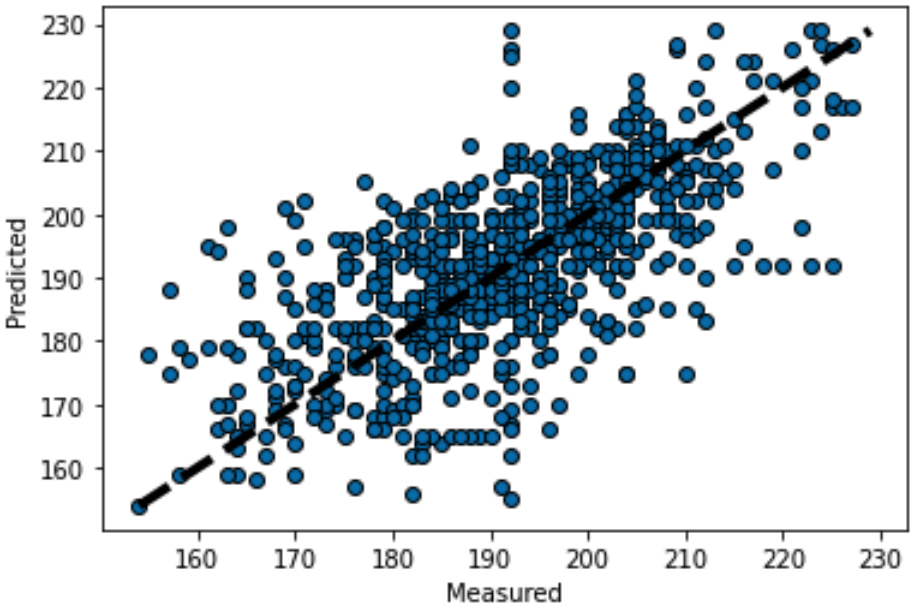

Figure A.18: Graph depicting the measured values vs. the actual values of a simple decision tree regression model for the panelling pressure resistance of 2-piece metal food cans

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of a decision tree regression model, the MAE was 8.964779874213836 and the RMSE was 11.571761334272258. The PCA data accomplishes accuracy scores that were less accurate then the feature selection data for the simple linear regression model.

A random forest regression model was built to predict panelling pressure resistance with all the factors from the feature selection data table as independent variables. See **Figure A.19** for some of the predicted values and **Figure A.20** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	189	189.000000
641	179	190.000000
1554	189	184.192319
575	190	187.000000
117	195	196.000000
...	...	...
983	168	188.000000
910	206	202.000000
1311	188	201.000000
215	196	186.000000
2870	180	189.000000

Figure A.19: Random forest regression model's predictions of panelling pressure resistance of 2-piece metal food cans

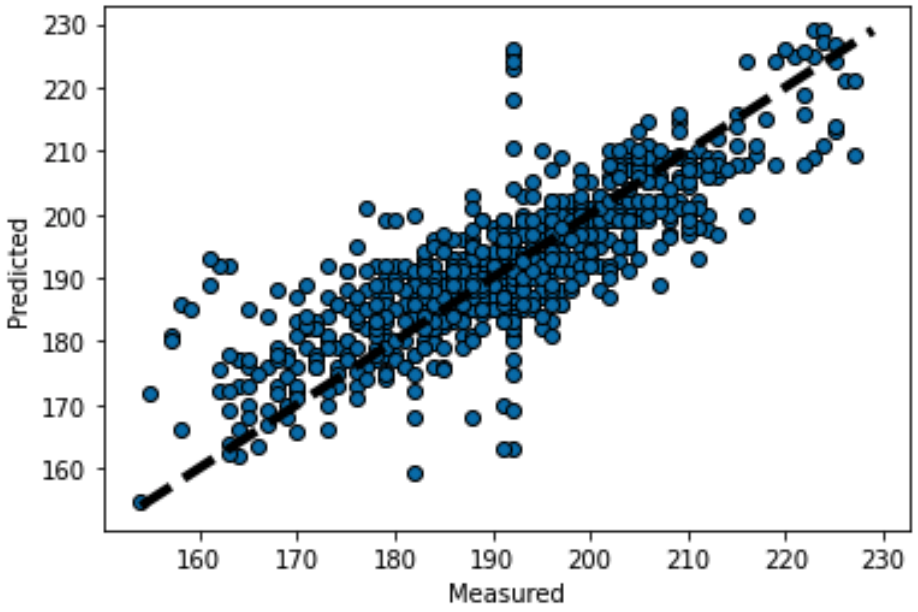

Figure A.20: Graph depicting the measured values vs. the actual values of a random forest regression model for the panelling pressure resistance of 2-piece metal food cans

The MAE was 5.717769072398814 and the RMSE was 7.835425256577493 for the random forest regression model. Both the MAE as well as the RMSE were better for the random forest regression

model when compared to the SVM regression model as well as the other regression models described previously in this Appendix.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the panelling pressure resistance by means of a random forest regression model, the MAE was 6.7978506126138685 and the RMSE was 9.02939112998814. The PCA data accomplishes accuracy scores that were less accurate then the feature selection data for the random forest regression model.

A.2.4.7 Boosting regression to predict panelling pressure

An Adaboost regression model was built to predict panelling pressure resistance with all the factors from the feature selection data table as independent variables. The Adaboost regression model was used to predict the panelling pressures of the test data set, see **Figure A.21** for some of the predicted values and **Figure A.22** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	189	184.0
641	179	182.0
1554	189	181.0
575	190	183.0
117	195	183.0
...	...	...
983	168	183.0
910	206	197.0
1311	188	193.0
215	196	181.0
2870	180	182.0

Figure A.21: Adaboost regression model's predictions of panelling pressure resistance of 2-piece metal food cans

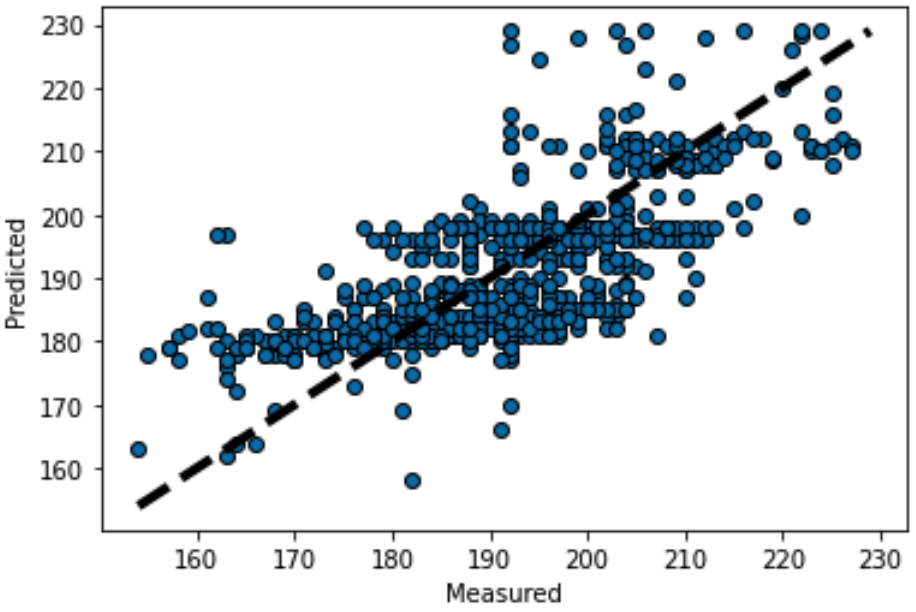

Figure A.22: Graph depicting the measured values vs. the actual values of an Adaboost regression model for the panelling pressure resistance of 2-piece metal food cans

The MAE was 7.705320179648336 and the RMSE was 9.655951641774411 for the Adaboost regression model. Both the MAE as well as the RMSE were not as good for the Adaboost regression model when compared to the random forest regression models.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of an Adaboost regression model, the MAE was 8.692522464454703 and the RMSE was 10.936918733291135. The PCA data accomplishes accuracy scores that were less accurate then the feature selection data for the Adaboost regression model.

A gradient boost regression model was built to predict panelling pressure resistance with all the factors from the feature selection data table as independent variables. The gradient boost regression model was used to predict the panelling pressure of the test data set, see **Figure A.23** for some of the predicted values and **Figure A.24** for a visual representation of the predicted values vs. the actual values.

	Actual	Predicted
1093	189	191.0
641	179	189.0
1554	189	181.0
575	190	187.0
117	195	196.0
...	...	...
983	168	185.0
910	206	203.0
1311	188	201.0
215	196	188.0
2870	180	185.0

Figure A.23: Gradient boost regression model's predictions of panel pressure resistance of 2-piece metal food cans

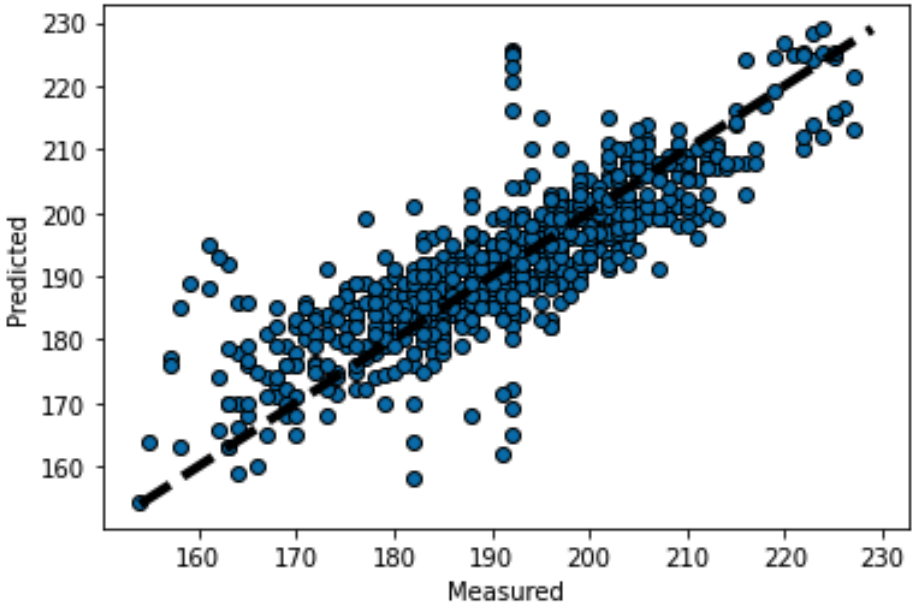

Figure A.24: Graph depicting the measured values vs. the actual values of a gradient boost regression model for the panelling pressure resistance of 2-piece metal food cans

The MAE was 5.3916932722462345 and the RMSE was 7.53399859800151 for the gradient boost regression model. Both the MAE as well as the RMSE gave the best accuracy results for the gradient

boost regression model when compared to the random forest regression models and all the other models featured in this chapter.

In **Section 5.6.3.3** the most important features were determined by PCA. When the first 8 components were used to predict the axial load resistance by means of an gradient boost boost regression model, the MAE was 6.610877840515401 and the RMSE was 8.799394842215872. The PCA data accomplishes accuracy scores that were less accurate then the feature selection data for the gradient boost regression model.

A.3 REGRESSION MODEL COMPARISONS

Various regression models were used on the feature selected and feature extracted data, and their accuracies were evaluated by means of MAE and RMSE. The following tables summarize the accuracies of the predictive ML models used in this case study for panelling pressure resistance from either the feature selected data or the feature extracted data.

Refer to **Table A.1** for a summary of the MAE and the RMSE of the regression models used to predict the panelling pressure resistance with data from the main factors as selected by random forest feature selection.

Table A.1: ML model accuracy to predict the panelling pressure resistance of 2-piece metal food cans with the random forest selected features

Regression Model	MAE	RMSE
Simple Linear Regression	7.0	8.8
Multiple Linear Regression	6.4	8.9
LASSO Regression	6.2	8.1
Bayesian Ridge Regression	6.2	8.1
Support Vector Machine Regression	6.1	8.2
Simple Decision Tree Regression		
Random Forest Regression	5.7	7.8
AdaBoost Regression	7.7	9.7
Gradient Boost Regression	5.4	7.5

Refer to **Table A.2** for a summary of the MAE and the RMSE of the regression models used to predict the panelling pressure resistance with data from the main factors as extracted by PCA.

Table A.2: ML model accuracies to predict the panelling pressure resistance of 2-piece metal food cans with the PCA extracted features

Regression Model	**MAE**	**RMSE**
Simple Linear Regression	7.1	8.2
Multiple Linear Regression	6.5	8.6
LASSO Regression	9.0	11.2
Bayesian Ridge Regression	9.0	11.2
Support Vector Machine Regression	7.2	9.5
Simple Decision Tree Regression	9.0	11.6
Random Forest Regression	6.8	9.0
AdaBoost Regression	8.7	10.9
Gradient Boost Regression	6.6	8.8

APPENDIX B

EXAMPLES OF PYTHON CODE USED FOR CASE STUDY

B.1 DATA DESCRIPTION

```
#import data file
import io
import pandas as pd
df = pd.read_csv(io.BytesIO(data_to_load['Beader Axial 04-08.csv']))
#show columns in data table
df.columns
#remove unnamed columns
df = df.loc[:, ~df.columns.str.contains('^Unnamed')]
df.columns
df.head()
#show data types
df.info()
#describe data by simple statistical calculations
df.describe()
#change beader values from numerical to an object
df['Beader'] = df['Beader'].astype('object')
df.info()
#draw quantile-quantile plots to show distribution of data and
visualize outliers
import matplotlib.pyplot as plt #import plotting package
#render plotting automatically
%matplotlib inline
import matplotlib as mpl #additional plotting functionality
import scipy.stats as stats
df_Beaded_can_flange_width = df['Beaded_can_flange_width_average']
stats.probplot(df_Beaded_can_flange_width, dist="norm", plot=plt)
plt.title("Quantile-quantile plot for beaded can flange width")
plt.show()
df_Beaded_can_height = df['Beaded_can_height_average']
stats.probplot(df_Beaded_can_height, dist="norm", plot=plt)
plt.title("Quantile-quantile plot for beaded can height")
plt.show()
df_Roll_bead_diameter = df['Roll_bead_diameter']
stats.probplot(df_Roll_bead_diameter, dist="norm", plot=plt)
plt.title("Quantile-quantile plot for beaded can roll bead diameter")
plt.show()
df_Beaded_can_roll_bead_position =
df['Beaded_can_roll_bead_position_average']
stats.probplot(df_Beaded_can_roll_bead_position, dist="norm",
plot=plt)
plt.title("Quantile-quantile plot for beaded can roll bead position")
plt.show()
df_Bead_depth_average = df['Bead_depth_average']
stats.probplot(df_Bead_depth_average, dist="norm", plot=plt)
plt.title("Quantile-quantile plot for beaded can bead depth")
plt.show()
df_Axial_load = df['Axial_load']
stats.probplot(df_Axial_load, dist="norm", plot=plt)
plt.title("Quantile-quantile plot for beaded can axial load
resistance")
plt.show()
```

Figure B.1: Python code associated with the description of data

B.2 DATA ASSESSING

```
#one-hot-encoding
y = pd.get_dummies(df.Beader, prefix='Beader')
print(y.head())
# from here you can merge it onto your main DF
# use pd.concat to join the new columns with your original dataframe
df = pd.concat([df,pd.get_dummies(df['Beader'],
prefix='Beader')],axis=1)
# now drop the original 'beader' column (you don't need it anymore)
df.drop(['Beader'],axis=1, inplace=True)
df.head()
#draw a correlation table from all factors in the dataframe
df.corr()
#visualize correlations using heatmap
import seaborn as sns
import pandas as pd
from pandas import DataFrame
import matplotlib.pyplot as plt
#calculate correlations between real-valued attributes
corMat = DataFrame(df.corr())
plt.pcolor(corMat)
plt.show()
#scatter plots
x = df['Axial_load']
y = df['Bead_depth_average']
plt.scatter(x, y)
plt.title("Scatter plot for beaded can axial load and average bead
depth")
plt.xlabel("Axial load resistance")
plt.ylabel("Bead depth")
plt.show()
#create a seperate dataframe for a parallel coordinate plot
df_parallel_coordinate_plot =
df[['FE_Can_height_average','FE_midwall_thickness_average',
'FE_top_wall_thickness_average','Beaded_can_height_average','Beaded_ca
n_roll_bead_position_average','Bead_depth_average','Axial_load','Panel
_resistance']]
#draw a parallel coordinate plot
import plotly.express as px
fig = px.parallel_coordinates(df_parallel_coordinate_plot,
color="Bead_depth_average",
color_continuous_scale=px.colors.diverging.balance,
color_continuous_midpoint=0.577651)
fig.show()
#box plots
x = df['Axial_load']
plt.boxplot(x)
plt.xlabel("Axial_load")
plt.ylabel(("Quartile Ranges"))
plt.show()
```

Figure B.2: Python code associated with the assessing of data related to data types, combination of data, correlations in the data and data visualization

```
#draw random sample for 500 measurements
import random
unique_id = list(df['Axial_load'].unique())
random.seed(30) # set a seed so that every time we will extract same sample
sample_id = random.sample(unique_id, 500)
sample_df = df[df['Axial_load'].isin(sample_id)].reset_index(drop=True)
# two variables of interest
sample_df = sample_df[['Beader', 'Axial_load']]
groups = sample_df.groupby('Beader').count().reset_index()
groups
# calculate ratio of the largest to the smallest sample standard deviation
ratio = sample_df.groupby('Beader').std().max() / sample_df.groupby('Beader').std().min()
ratio
# Create ANOVA backbone table
data = [['Between Groups', '', '', '', '', '', ''], ['Within Groups', '', '', '', '', '', ''],
['Total', '', '', '', '', '', '']]
anova_table = pd.DataFrame(data, columns = ['Source of Variation', 'SS', 'df', 'MS', 'F', 'P-
value', 'F crit'])
anova_table.set_index('Source of Variation', inplace = True)
# calculate SSTR and update anova table
x_bar = sample_df['Axial_load'].mean()
SSTR = sample_df.groupby('Beader').count() * (sample_df.groupby('Beader').mean() - x_bar)**2
anova_table['SS']['Between Groups'] = SSTR['Axial_load'].sum()
# calculate SSE and update anova table
SSE = (sample_df.groupby('Beader').count() - 1) * sample_df.groupby('Beader').std()**2
anova_table['SS']['Within Groups'] = SSE['Axial_load'].sum()
# calculate SSTR and update anova table
SSTR = SSTR['Axial_load'].sum() + SSE['Axial_load'].sum()
anova_table['SS']['Total'] = SSTR
# update degree of freedom
anova_table['df']['Between Groups'] = sample_df['Beader'].nunique() - 1
anova_table['df']['Within Groups'] = sample_df.shape[0] - sample_df['Beader'].nunique()
anova_table['df']['Total'] = sample_df.shape[0] - 1
# calculate MS
anova_table['MS'] = anova_table['SS'] / anova_table['df']
# calculate F
F = anova_table['MS']['Between Groups'] / anova_table['MS']['Within Groups']
anova_table['F']['Between Groups'] = F
# p-value
anova_table['P-value']['Between Groups'] = 1 - stats.f.cdf(F, anova_table['df']['Between
Groups'], anova_table['df']['Within Groups'])
# F critical
alpha = 0.05
# possible types "right-tailed, left-tailed, two-tailed"
tail_hypothesis_type = "two-tailed"
if tail_hypothesis_type == "two-tailed":
    alpha /= 2
anova_table['F crit']['Between Groups'] = stats.f.ppf(1-alpha, anova_table['df']['Between
Groups'], anova_table['df']['Within Groups'])
# Final ANOVA Table
anova_table
# The p-value approach
print ("Paneling pressure resistance ANOVA")
print("Approach 1: The p-value approach to hypothesis testing in the decision rule")
conclusion = "Failed to reject the null hypothesis."
if anova_table['P-value']['Between Groups'] <= alpha:
    conclusion = "Null Hypothesis is rejected."
print("F-score is:", anova_table['F']['Between Groups'], " and p value is:", anova_table['P-
value']['Between Groups'])
print(conclusion)
# The critical value approach
print("\n----------------------------------------------------------------")
print("Approach 2: The critical value approach to hypothesis testing in the decision rule")
conclusion = "Failed to reject the null hypothesis."
if anova_table['F']['Between Groups'] > anova_table['F crit']['Between Groups']:
    conclusion = "Null Hypothesis is rejected."
print("F-score is:", anova_table['F']['Between Groups'], " and critical value is:",
anova_table['F crit']['Between Groups'])
print(conclusion)
```

Figure B.3: Python code associated with the assessing of data related to ANOVA

B.3 DATA PREPARATION

```
# Remove Outliers
import numpy as np
print("Axial Load Skewness")
print(df['Axial_load'].skew())
df['Axial_load'].describe()
df['Axial_load'] = np.where(df['Axial_load'] > 227, 201,
df['Axial_load'])
df['Axial_load'] = np.where(df['Axial_load'] < 170, 201,
df['Axial_load'])
# Combine date and time columns into one datetime column
df['Datetime'] =  pd.to_datetime(df['Datetime'],  format='%d/%m/%Y
%H:%M')
# Perform SFS
from mlxtend.feature_selection import SequentialFeatureSelector
from sklearn.neighbors import KNeighborsClassifier
X = df.drop(['Datetime', 'Axial_load', 'Panel_resistance'], axis=1)
y = df['Axial_load']
knn = KNeighborsClassifier(n_neighbors=4)
import matplotlib.pyplot as plt
from mlxtend.plotting import plot_sequential_feature_selection as
plot_sfs
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
sfs1 = SFS(lr,
           k_features=20,
           forward=True,
           floating=False,
           scoring='neg_mean_absolute_error',
           cv=10)
sfs1 = sfs1.fit(X, y)
fig = plot_sfs(sfs1.get_metric_dict(), kind='std_err')
plt.title('Sequential Forward Selection (w. StdErr)')
plt.grid()
plt.show()
#Show the column nimbers of which features were selected
feat_cols = list(sfs1.k_feature_idx_)
print(feat_cols)
```

Figure B.4: Python code associated with the preparing of data related to feature selection by using SFS

```
#drop columns
df1=df.drop(['Panel_resistance', 'Datetime'], axis = 1)
# Use numpy to convert to arrays
import numpy as np
#  the values we want to predict
predict = np.array(df1['Axial_load'])
# Remove the labels from df
# axis 1 refers to the columns
df1= df1.drop('Axial_load', axis = 1)
# Saving feature names for later use
df1_list = list(df1.columns)
# Convert to numpy array
df1 = np.array(df1)
# Using Skicit-learn to split data into training and testing sets
from sklearn.model_selection import train_test_split
# Split the data into training and testing sets
train_df1, test_df1, train_predict, test_predict = train_test_split(df1,
predict, test_size = 0.25, random_state = 42)
print('Training data2 Shape:', train_df1.shape)
print('Training predict Shape:', train_predict.shape)
print('Testing data2 Shape:', test_df1.shape)
print('Testing predict Shape:', test_predict.shape)
# Import the model we are using
from sklearn.ensemble import RandomForestRegressor
# Instantiate model with 1000 decision trees
rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)
# Train the model on training data
rf.fit(train_df1, train_predict);
predict_pred = rf.predict(test_df1)
# Import tools needed for visualization
from sklearn.tree import export_graphviz
import pydot
# Pull out one tree from the forest
tree = rf.estimators_[5]
# Import tools needed for visualization
from sklearn.tree import export_graphviz
import pydot
# Pull out one tree from the forest
tree = rf.estimators_[5]
# Export the image to a dot file
export_graphviz(tree, out_file = 'tree.dot', feature_names = df1_list,
rounded = True, precision = 1)
# Use dot file to create a graph
(graph, ) = pydot.graph_from_dot_file('tree.dot')
# Write graph to a png file
graph.write_png('tree.png')
# Get numerical feature importances
importances = list(rf.feature_importances_)
# List of tuples with variable and importance
feature_importances = [(feature, round(importance, 2)) for feature,
importance in zip(df1_list, importances)]
# Sort the feature importances by most important first
feature_importances = sorted(feature_importances, key = lambda x: x[1],
reverse = True)
# Print out the feature and importances
[print('Variable: {:20} Importance: {}'.format(*pair)) for pair in
feature_importances];
```

Figure B.5: Python code associated with the preparing of data related to feature selection by using random forest

```
#LDA
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
lda = LinearDiscriminantAnalysis(n_components=2)
X_train_lda = lda.fit_transform(X_train_std, y)
data=pd.DataFrame(X_train_lda)
data['class']=y
data.columns=["LD1","LD2","class"]
data.head()
markers = ['s', 'x','o']
colors = ['r', 'b','g']
sns.lmplot(x="LD1", y="LD2", data=data, hue='class',
markers=markers,fit_reg=False,legend=False)
plt.legend(loc='upper right')
plt.show()

#PCA
from sklearn.preprocessing import StandardScaler
# Separating out the features
x = df1.iloc[:,1:63].values
# Separating out the target
y = df1.loc[:,['Axial_load']].values
# Standardizing the features
x = StandardScaler().fit_transform(x)
from sklearn.decomposition import PCA
pca = PCA(n_components=8)
principalComponents = pca.fit_transform(x)
principalDf = pd.DataFrame(data = principalComponents
             , columns = ['principal component 1', 'principal component
2', 'principal component 3', 'principal component 4', 'principal component
5', 'principal component 6', 'principal component 7', 'principal component
8'])
principalDf.head()
```

Figure B.6: Python code associated with the preparing of data related to feature extraction by using LDA and PCA

B.4 DATA MODELLING

```
import matplotlib.pyplot as plt
from sklearn.compose import TransformedTargetRegressor
from sklearn.preprocessing import QuantileTransformer
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import mean_absolute_error
X = df[['FE_midwall_thickness_average', 'Bead_depth_average',
       'Nippon']]
y = df['Axial_load']
from sklearn.preprocessing import StandardScaler
# Standardizing the features
X = StandardScaler().fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 0)
# Create linear regression object
# Fit regression model
params = {'n_estimators': 2000, 'max_depth': 4, 'min_samples_split': 2,
'learning_rate': 0.01, 'loss': 'ls'}
regr = GradientBoostingRegressor(**params)
# Train the model using the training sets
regr.fit(X_train, y_train)
# Make predictions using the testing set
y_pred = regr.predict(X_test)
print('R^2 score without Transformation: {0:.2f}'.format(regr.score(X_test, y_test)))
# Explained variance score: 1 is perfect prediction
print("R^2 = %0.5f" % r2_score(y_test, y_pred))
# The mean absolute error
print("MAE = %5.3f" % mean_absolute_error(y_test, y_pred))
# The mean squared error
print("MSE = %5.3f" % mean_squared_error(y_test, y_pred))
# Plot outputs
fig, ax = plt.subplots()
ax.scatter(y_test, y_pred, edgecolors = (0, 0, 0))
ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw = 4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.title('Without Transformation')
plt.show()
transformer = QuantileTransformer(output_distribution = 'normal')
regressor = GradientBoostingRegressor(**params)
regr = TransformedTargetRegressor(regressor = regressor,
transformer = transformer)
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 0)
regr.fit(X_train, y_train)
y_pred = regr.predict(X_test)
print('R^2 score with Transformation: {0:.2f}'.format(regr.score(X_test,
y_test)))
# Explained variance score: 1 is perfect prediction
print("R^2 = %0.5f" % r2_score(y_test, y_pred))
# The mean absolute error
print("MAE = %5.3f" % mean_absolute_error(y_test, y_pred))
# The mean squared error
print("MSE = %5.3f" % mean_squared_error(y_test, y_pred))
# Plot outputs
fig, ax = plt.subplots()
ax.scatter(y_test, y_pred, edgecolors = (0, 0, 0))
ax.plot([y.min(), y.max()], [y.min(), y.max()], 'k--', lw = 4)
ax.set_xlabel('Measured')
ax.set_ylabel('Predicted')
plt.title('With Transformation')
plt.show()
y_pred = regr.predict(X_test)
GBoostDT = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
GBoostDT
import numpy as np
from sklearn import metrics
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

Figure B.7: Python code associated with the modelling of data using gradient boost regression