CHALMERS
UNIVERSITY OF TECHNOLOGY

Implementing data analytics for improved quality in manufacturing: a case study

Master's thesis in Production Engineering

NILS LUNDÉN

DEPARTMENT OF INDUSTRIAL AND MATERIALS SCIENCE

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

Implementing data analytics for improved quality in manufacturing: a case study

NILS LUNDÉN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Industrial and Materials Science
Chalmers University of Technology
Gothenburg, Sweden 2022

Implementing data analytics for improved quality in manufacturing: a case study
NILS LUNDÉN

Supervisors: Ebru Turanoglu Bekar and Martin Dahl, Department of Industrial and Materials Science
Examiner: Anders Skoogh, Department of Industrial and Materials Science

Master's Thesis 2022
Department of Industrial and Materials Science
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Implementing data analytics for improved quality in manufacturing: a case study
NILS LUNDÉN
Department of Industrial and Materials Science
Chalmers University of Technology

Abstract

The possibilities for extracting important information from big data have accelerated over the last few years. More manufacturing companies are realizing the potential of advanced data analytics, such as machine learning, and how this can be used to improve productivity and sustainability for future competitiveness. It is however often not clear how to approach the adoption of data analytics and overcome the necessary challenges before real value can be created using these techniques. This thesis proposes what some of the most relevant challenge areas are to improve the conditions of conducting successful pilot projects for data analytics in manufacturing. A suitable methodology for the implementation of data analytics is also presented, which is evaluated in a case study at a machining department at Volvo. The problem being solved in this case study was related to product quality improvement, where to goal was to derive which influencing factors in a machining process have the highest importance for the quality outcome. A state-of-the-art literature review was performed to evaluate which requirements are necessary efficiently utilize data analytics and which methodologies can be used for the performing use cases. An adapted version of an implementation methodology, the Cross Industry Standard Process for Data Mining (CRISP-DM), was developed to be better suited for the specific challenges in manufacturing. This methodology was thereafter used and evaluated in the case study. A combination of interpretable machine learning models was used in conjunction with association rules to describe which influencing factors are the most relevant for the quality outcome. From the results, it was possible to identify improvement areas that Volvo needs to address for better use of data analytics. Important findings were that the internal knowledge needs to be increased and that certain technical challenges need to be further developed, such as connectivity, system integration, and traceability. These results can help Volvo and other manufacturing companies in a similar development stage to understand how to prioritize efforts for succeeding with the implementation of data analytics. It especially gives insights into how to deal with use cases related to quality improvement and how to increase the interpretability during the implementation of data analytics.

Keywords: Data analytics, machine learning, manufacturing, quality improvement.

Acknowledgements

This thesis has been conducted in collaboration with Volvo Powertrain in Skövde at the Manufacturing Engineering Development department. I was involved in a project together with several other people to investigate how data analytics can be utilized at Volvo's machining department in the future. I would like to thank Volvo and the involved departments for allowing me to be part of this project and conduct my thesis on this interesting topic.

I want to express a special appreciation to my supervisor from Volvo, Carl Cederqvist, who has shown high enthusiasm for the work I have done and motivated me throughout the whole process. We have had many insightful discussions which have helped me to structure my work, and he has also provided support with technical tasks needed to perform the analysis.

I also want to give a sincere thanks to all the other people at Volvo involved in the project who have contributed to setting up the technical requirements, and who also have provided valuable insights for the thesis. Thank you Niklas Habbe, who managed the project and coordinated the work; Jani Nurmi, for the valuable expertise of the investigated machining process; Jonathan Strende, for solving the technical requirements needed for data collection; Stefan Berntsson, for further support for accessing the collected data; and Tony Jacobsson, for the valuable expertise in the measurement system used in the project. My thanks also go to all the other experts involved from Volvo and collaborating partners who have contributed to the project and the thesis in different ways. Without this collaborative contribution from everyone involved, it would not have been possible for me to write this thesis.

I would also like to give my deepest appreciation to my supervisors from Chalmers; Ebru Turanoglu Bekar and Martin Dahl. They have assisted me in every stage of the thesis and helped me forward with suggestions and different ways to handle difficult situations. They have also given me valuable feedback which has helped me to improve on my work.

Nils Lundén, Gothenburg, May 2022

Contents

List of Figures

List of Tables

1

Introduction

In the flourishing era of big data and digitization, new opportunities have arisen in the manufacturing industry. A paradigm shift towards smart manufacturing is being experienced where manufacturing data is utilized in new ways [1]. This is commonly referred to as Industry 4.0, where advanced data analytics plays an important role to achieve more safe, efficient, and sustainable production systems [2]. Considerable amounts of research have been dedicated to developing new solutions and tools to create value from big data in manufacturing [3]. The development of efficient algorithms, hardware, and sensors has created promising capabilities for scaling the usage of data analytics in the industrial sector [4]. The value is created when the data is translated into timely information that can be used to identify faults in production processes, predict optimal maintenance intervals, or optimize scheduling [3] to name some examples. Today, however, only a small proportion of the data being generated in manufacturing is used for this type of information acquisition due to the roadblocks companies need to get past to utilize this technology [5].

According to McKinsey [6], digital transformations within the automotive industry are best approached with an overarching strategy covering several core areas. This strategy should include a roadmap covering all areas that need to be included to reach the organization's goals, and an important part of it is to conduct pilot projects with a test-and-learn approach. Companies should adapt to industry 4.0 and data analytics in a step-wise approach which might not be synchronized between departments, but where each department progresses at a pace suited for their circumstances [7].

Most manufacturing companies today are still a long way from reaping the benefits of big data. Often, they are not aware of how they can use the data they have at hand and use it to improve their processes [8]. Furthermore, they need to be aware of what capability is required for different applications. The field of data analytics is also diverse and requires knowledge and experience to understand which algorithms, methods, and theories work best for certain purposes, which many manufacturing companies lack today [9]. Several standard process models exist which describe how data analytics projects should be approached, but these are not adapted specifically for manufacturing applications and the challenges that are unique in this area [5].

This thesis has been performed in collaboration with Volvo Powertrain in Skövde where the manufacturing of truck engines is taking place. They have started a digital transformation journey, and the machining department is now at a stage where pilot projects are to be conducted. Through these pilot projects, it will be evaluated what problems can be solved within manufacturing operations with current capabilities and identify improvement potential. The long-term goal for the department is to achieve efficient large-scale usage of data analytics to solve a variety of problems, such as reducing quality errors, optimizing tool changes, and using condition-based smart maintenance.

1.1 Aim and research questions

The aim of this project is two-fold. One aim is to give a better understanding of the requirements and to suggest an approach for conducting successful data analytics projects within manufacturing. The other is to show how data analytics can be used for improving quality problems in machining processes, where the understandability of how parameters influence the quality is important. The following research questions will be answered to fulfill these aims:

RQ1: What can be the challenges and requirements for performing data analytics projects in manufacturing?

RQ2: What can be a suitable methodology for performing data analytics projects in manufacturing?

RQ3: How can data analytics be used to create interpretable models which can be used for quality improvements?

The first question will give a clearer view of companies new to data analytics and what is required to use the technology. Volvo's readiness for these requirements will also be evaluated. The second question will lay a foundation for a methodology adapted for manufacturing challenges, which will make it easier to approach pilot projects in a structured and repeatable manner. This methodology should be general enough to be used for diverse applications and can be adapted to company-specific needs during the learning process from pilot projects. The third question is investigated through a case study at Volvo and provides an example of how value can be created through data analytics with the adapted methodology.

1.2 Scope and limitations

The scope of this thesis is to provide companies at the beginning of a transition towards industry 4.0 with directional guidelines for using data analytics. The project is limited to showing the basic requirements and suggesting a suitable approach for achieving a proof of concept (POC) of the technology. Therefore, a complete roadmap is not presented for how to fully integrate data analytics in an organization, and the organizational requirements are not taken into regard. Furthermore,

the project does not focus on how to deploy analytic models and how they are maintained. The focus is on investigating the capability needed to perform successful pilot projects through the development stage.

The target group of the thesis is all manufacturing companies that may benefit from this investigation. The challenges and requirements however are chosen and evaluated considering where Volvo currently is in its digitization journey. The proposed methodology will therefore assume that a starting ground has been reached where computerization and some level of connectivity are already in place. The project is further limited to investigating specifically how data analytics can be used for quality problems in machining processes with dimensional variations in focus. This type of issue is a high impact problem in the machining department at Volvo Powertrain and is therefore of high priority to find new solutions for.

1.3 Thesis outline

This report is divided into six Chapters. The following Chapter 2 contains a literature review starting with a general background to data analytics within manufacturing. This review presents how data analytics is used specifically for product quality problems in machining processes. It then presents relevant challenges and requirements for conducting data analytics projects, and common methodologies for these types of projects.

Chapter 3 presents an adapted version of a general methodology, CRISP-DM. It is described how this methodology can be altered to be more descriptive for execution in manufacturing environments.

Chapter 4 shows how the adapted methodology can be implemented for a use case at Volvo. In the use case, it is investigated how data analytics can be used to reduce the impact of dimensional variations of milled holes in cylinder heads for a machining process. The result of each step in the methodology is presented.

Chapter 5 contains a discussion about the practical contributions of the thesis for Volvo and how the company can proceed with implementing data analytics. It also discusses what theoretical contributions the adapted methodology has, and the implications of the chosen analytics approach to improving quality. Finally, a conclusion of the project is given in Chapter 6.

2

Literature review

In this chapter, a review of state-of-the-art knowledge of the topics is presented. First, a brief background is provided to explain how data analytics can be applied on different levels of complexity and how this can create value in manufacturing. Thereafter, it is described how data analytics can be used for improving product quality with different approaches (part of RQ3). It is also discussed what challenges and requirements are necessary to achieve the capability to conduct data analytics projects in manufacturing (part of RQ1). Lastly, a review of methodologies for conducting these projects is presented (part of RQ2).

2.1 Data analytics in manufacturing

The growing relevance of big data and data analytics is creating new opportunities to achieve smart manufacturing. Data analytics in this report refers to retrieving information from data with the help of statistical tools. The use of data analytics in manufacturing enables pattern recognition in large data sets from the shop floor and previously isolated systems, which can reveal insights for improving processes [10]. The implementation of data analytics in manufacturing operations has accelerated over the last few years. Europe is in the lead where over half of the top manufacturers have implemented at least one AI use case according to a report from Capgemini [11]. They found that the most common use cases are within smart maintenance, quality control, and demand planning. This is mainly due to the clear business value that can be gained in these areas as well as the availability of data and relative ease of implementation they argue. In [12] it is described how a well-implemented data analytics strategy can improve overall equipment efficiency (OEE) in production systems. Quality is improved by reducing process deviations, availability through reduced setup times and predictive maintenance, and performance through optimization of machine parameters to reduce the number of minor stops.

2.1.1 Levels of data analytics

It is common to categorize data analytics within manufacturing into four levels: Descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics according to Dai et al. [1]. They describe that descriptive and diagnostic

analytics can be viewed as reactive approaches while predictive and prescriptive analytics is viewed as proactive approaches. Furthermore, they describe that the implementation becomes more complex with each level, but at the same time, higher business value can be gained by advancing to the higher levels as illustrated in Figure 2.1. In a study performed by Acatech [7] it is claimed that companies adopting a data analytics strategy should aim to achieve these levels by a step-by-step approach. This is since different capabilities will be required to achieve the different levels, and companies should therefore adopt these capabilities at a pace that is suited for their circumstances as described in the report. In this section, an overview is given of the four levels.

Descriptive analytics

The goal of descriptive analytics is to answer "What happened" by providing insights into the status of the machine [13]. The various sensors from machines can capture target values and correlated values for a specific problem, but this data will only provide a partial view of the process and will need analytics to be converted into knowledge that can be used for decision support [14]. Different data mining and statistic methods are used to identify relationships in the data to reveal important characteristics [1]. This data-to-information conversion is necessary to transform the large quantities of process data into perceptible information that is usable for operators and process experts [15]. To create an efficient data-driven decision process with descriptive analytics, the extracted information should be presented clearly through visualization tools such as dashboard applications, and automatically send alerts when interventions are required [13].

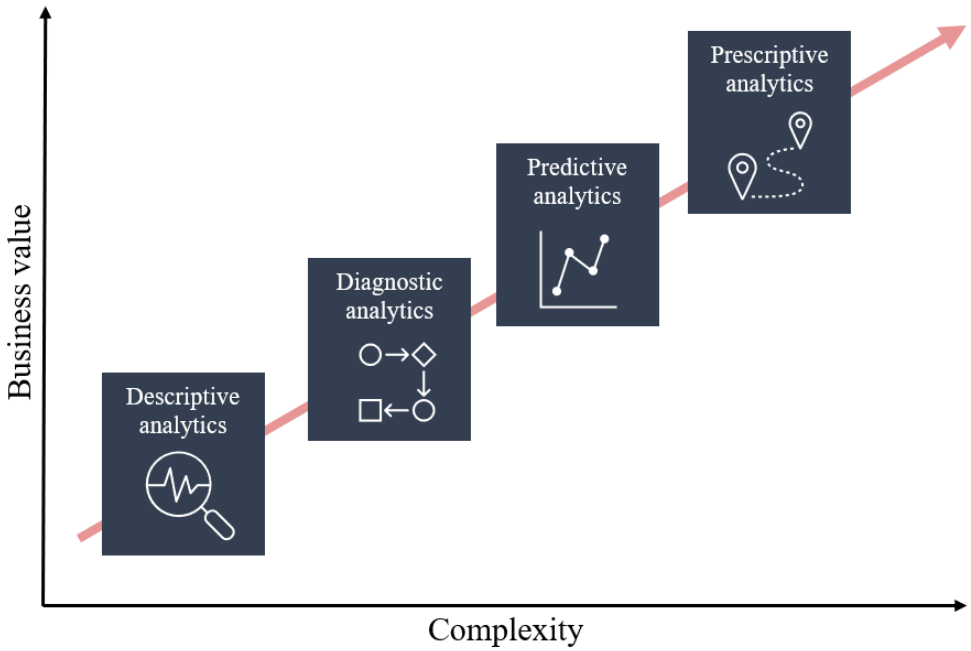

Figure 2.1: The four levels of data analytics and their relation to complexity and business value, inspired by [1]

Diagnostic analytics

On the diagnostic level, attempts are taken to answer why something happened by taking a deeper look at the data to understand the causes of problematic events [1]. This analysis usually builds upon descriptive analytics, but additional data and analytics techniques are required to find the root cause of a problem. Common methods to apply to reveal these causes are generalization, association, sequence pattern mining, and clustering analysis [13]. The possibility to find the most important parameters influencing the performance of the manufacturing system allows for improvement work of the process. Either domain experts can modify the process to reduce the impact of identified important parameters, or optimal ranges for the parameters can be derived to alert when these ranges are about to be exceeded [10].

Predictive analytics

The predictive analytics level aims to provide insights into the future and answer the questions "what is likely to happen" [13]. This is done mainly by analyzing historical data to find trends with supervised and unsupervised machine learning (ML) models that can make predictions of future events [1]. To achieve predictive analytics in manufacturing, advanced prediction tools need to be developed that can transform data into information that can explain uncertainties and assist in making more informed decisions [16]. Forecasting is one of the main issues in manufacturing according to Dogan and Birant [17], and ML models can approximately predict future scenarios with high accuracy. They explain that the prediction ability will vary depending on the ML algorithm and available data, but it has been proven several times in the manufacturing field that robust and high-performing prediction models can be developed.

Prescriptive analytics

Prescriptive analytics is the most advanced level within data analytics and extends the predictive level by also answering what should be done to achieve a desired predicted outcome [1]. To be able to prescribe the optimal set of decisions through the cause-effect relationships in the analytics results, optimization methods are commonly used in conjunction with ML models such as Discrete Choice Modeling, Linear and Non-linear Programming, and Value Analysis [13]. In this way, optimal production parameters can be determined by analyzing the data from already running production processes [18]. The prescribed decisions can either be executed manually or automatically. If done manually, the goal can be to augment the employee's ability to select the optimal parameters for a manufacturing process [19]. Autonomous optimization on the other hand implies that a machine can adjust its parameters on its own to achieve a set goal [20].

2.1.2 Machine learning

As described in the previous section, ML is a common technology to use within advanced data analytics to gain insight into manufacturing. ML is a subset of artificial

intelligence (AI) and offers new possibilities for gaining value from manufacturing data. The traditional way of creating models (before ML) has been to define rules and formulas that are applied to the data to obtain an output. With ML however it is the opposite, the data makes out the starting point, and then algorithms are applied to learn from the data to create the rules [21]. The goal of ML is to detect patterns or regularities that describe relations in large sets of data, which then can support decision making or automatically improve manufacturing systems [9]. ML has never been as easy to apply and promising as today due to recent developments in efficient algorithms, computing hardware, and sensor technology for data collection according to Bauer et al. [4]. They claim that within manufacturing, ML holds great potential due to the high amount of automatable tasks and increase of connected devices. ML is a tool that can increase the understanding of problems in the manufacturing domain which can often be seen as "data rich but knowledge-sparse" [9]. The ML models are commonly classified into supervised learning, unsupervised learning, and reinforcement learning. The division of these three learning methods is depicted in Figure 2.2 where also some examples of common algorithms for each method are shown. A brief introduction to the different learning approaches is given in the remainder of this section.

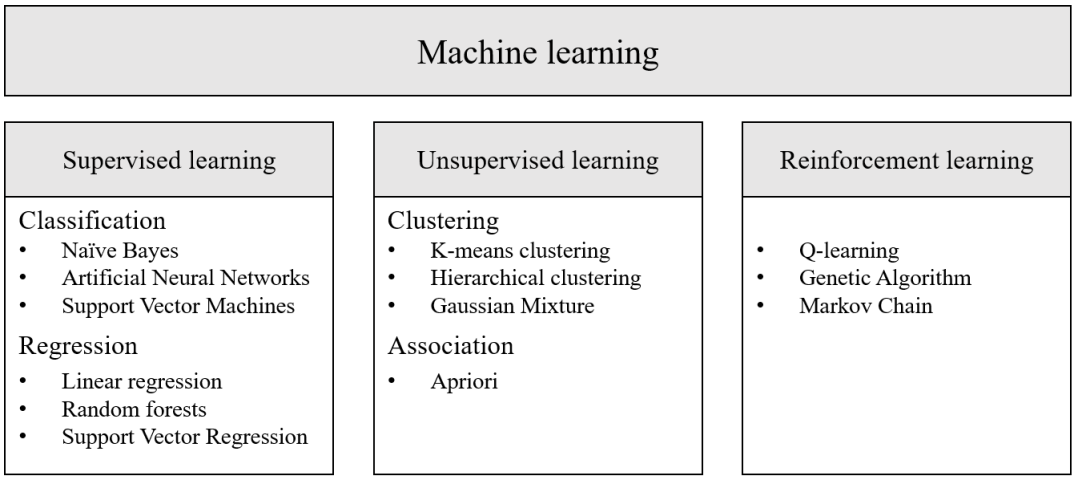

Figure 2.2: Classification of ML methods with examples of common algorithms

Supervised learning

A requisite to using supervised learning is that access to both input data and output data can be obtained, where the output is what the model should be able to predict. Supervised learning is further divided into regression and classification, where a function is developed to map the input data to the output data with a prediction accuracy as high as possible [2]. Regression algorithms aim to develop models which can predict continuous values whereas classification algorithms predict between two or more distinguishing classes. The process of supervised ML starts with a data set which is divided into training data, the data the algorithm uses to create the model, and test data, which is used for evaluating the model [9]. The following common steps after the data collection are to perform data preparation, choosing a model, training the model, evaluating the model, tuning the model parameters, and make

predictions for the final evaluation [22]. In [20], an investigation of commonly applied ML algorithms for machining applications was made. They showed that for example Support Vector Machines (SVM) and Artificial Neural Networks (ANN) have been successful in creating prediction models in manufacturing since the algorithms are good at handling the large quantities of machine processing data.

Unsupervised learning

In contrast to supervised learning, unsupervised learning does not require output data for modeling but is rather used for defining the underlying structure within input data [2]. This might be useful during instances where labeled output data is not available. Unsupervised learning is also a capable tool for creating descriptive models, as it can reveal relations among variables and provide better process understanding [17]. Wuest et al. [9] describe unsupervised learning as a way of creating structure in data when an identified output or feedback is missing. Two common approaches for this which they bring up are clustering, where data is grouped by distinguishing attributes, and association rules, where rules which describe relations in the data are defined.

Reinforcement learning

Reinforcement learning is a training method for ML that rewards desired behaviors and punishes undesired behaviors. This method, therefore, requires direct feedback which enables the model to learn through trial and error when evaluating possible actions. A distinguishing characteristic of reinforcement learning compared to supervised or unsupervised learning is that training information is provided continuously from the environment instead of a historical data set as explained in [9]. The other characteristic is that the model learns by trying which actions generate the best results instead of being told. They further claim however that reinforcement learning has not been widely applied in manufacturing and that there is a lack of examples from successful applications.

In this thesis, supervised and unsupervised learning will be of focus when discussing ML models as a solution to manufacturing quality problems. A description of the algorithms that were used for the case study is provided in Chapter 4, Section 4.2.5.

2.2 Data analytics for improving quality

Manufacturers have over the last decades been able to improve product quality and reduce variability in their machining processes by utilizing tools such as Six Sigma. Quality improvement work has always been of high relevance to reducing the amount of waste and saving costs in production. As described in [10], large variability is unavoidable in complex production systems and a more granular approach is needed to complement existing tools. Data analytics offers an approach for improving quality in processes by turning the raw machining data into useful knowledge. It has

been applied for use cases such as quality prediction of products, quality improvement, defect detection, and classification of defects as described in [17]. They argue that control variables play an important role when addressing these problems. The dominant factors of the workpiece to improve final product quality is dimensional accuracy and surface finish [14]. In this report, the focus is mainly on the problems related to dimensional accuracy in machining processes.

To be able to reduce dimensional variability, it is important to understand the influencing factors of the problem so relevant data is collected and analyzed. In [23], Shi argues that there are two main reasons for dimensional variations in machining processes: Setup errors and machine errors. The setup errors can be divided into fixture errors and datum errors. Fixture error means that there is an error with the fixture in relation to the working table of the machine, while datum errors occur when there is an error of the workpiece in relation to the fixture [23]. The datum is the surface of the workpiece which has contact with the fixture locators Shi explains, and datum errors can therefore arise in multistage machining processes if the datum surface has been processed in earlier stages. Machine errors are errors arising as the cutting path deviates from the programmed path. This can happen due to geometric and dimensional errors of the kinematic links of the machine, thermal errors, or wear of cutting tools [23]. It has been shown in other studies that machine errors can impact dimensional variations for machined holes due to variations in feedrate and cutting speed [24], cutting time and built-up edge on tools [25], and torque of spindles or axes [26]. In [23], the variational errors are further divided into quasi-static errors and dynamic errors. Quasi-static errors are static or slow varying errors between the cutting tool and the workpiece. Examples of such errors are geometric and kinematic errors, thermal errors, cutting force induced errors, tool wear induced errors, and fixturing errors. Dynamic errors relate to fast-changing errors, which can arise from spindles, machine vibrations, or controllers [23].

Several approaches have been taken to reduce the impact of quality problems such as dimensional variations in manufacturing. Depending on the complexity of the problem being solved and the capability possessed by the company, different methods can be used suited for these needs and constraints. Commonly applied methods found in the literature are process monitoring, root cause analysis, quality predictions, and process optimizations. These will be described in this section.

2.2.1 Process monitoring

Process monitoring mainly relates to descriptive data analytics tasks. The purpose of process monitoring as described in [27] is to monitor the product quality in real-time to discover unusual variability instead of waiting till the end of a process to discover the outcome of the product quality. When unusual variability is detected, manual adjustments are made to the process to hinder these variations from occurring again. They further explain that monitoring charts should be used to display and detect the unusual variability for one or several values. The common characteristic of such a chart is to display values in a sequence plot where a target line is shown together

with upper and lower control limits. The chart should be updated in real-time or close to real-time.

In [28] it is described how statistical process monitoring, sometimes called statistical process control (SPC), can be divided into three different generations. The first generation has been practiced since the 1920s, where variations in product quality have been monitored with the help of statistical distributions, such as the mean and variance of quality variables. The second generation of SPC took traction in the 1980s and can detect faults by not only monitoring quality variables but also process parameters to utilize the information that can be gained from the correlations between process parameters and product quality. The third generation has been developed during the last two decades where the goal is to utilize the information that was not available in the second generation of SPC, such as feature-based monitoring.

Feature-based monitoring is described in [2]. A common approach is to use unsupervised learning methods to extract interpretable features from the input data of the machining process. These features can provide information about the status of the process which would not be observable from single parameters that are directly measurable. They describe that there are many methods for extracting such features, but it is important to consider that derived features are closely related to prior knowledge about the process. A good process monitoring model should not only be able to describe process status, but also allow production personnel to make clear interpretations from the model.

Escobar et al. [29] describe a strategy called Process Monitoring for Quality (PMQ) which is an extension to SPC that builds on feature-based monitoring. They argue that SPC only works when there is a mature understanding of the process, the quality characteristics are observable, and a process to verify the quality criteria is in place. The benefit of PMQ is that it can be applied when there is a lack of understanding of the cause and effect between machine parameters and quality outcome, or when the quality outcome is missing. The PQM framework utilizes statistics and ML to predict the quality outcome through a binary classifier to label the product as "good" or "suspect". Products, where defects are suspected, can then be manually controlled to verify the predicted outcome. This technique allows for more efficient and cost-beneficial quality controls where only suspect products are subjected to manual control. This would be relevant when automatic inspections are not possible, or when the defects are difficult to detect, such as for internal faults in materials [17].

2.2.2 Root cause analysis

A common approach for dealing with quality issues is to try and identify the root cause of the problem so the issue can be fully corrected. Traditionally root cause analysis (RCA) has been performed by experienced personnel at a site to describe the casual relations between process steps and the quality outcome with the help of manual methods as described in [30]. They explain that the problem with this approach however is that this valuable knowledge might sometime be difficult to

transfer between individuals or between sites, and the experts might be biased in their assessments. Data analytics on a diagnostic level offers possibilities to perform objective RCA where models learn the causal relations in a machining process. In [31], automatic RCA is defined as a process where "one can find the true cause of a problem through the analysis of data, and at least part of this analysis does not require human intervention". The use of data-driven automatic RCA might make it possible to detect previously unknown root causes through advanced data analytics. It might also be possible to save costs as human inspectors do not need to spend as much time searching for the root causes [32].

Different approaches can be taken to achieve data-driven RCA depending on the previous knowledge of the causal effects in the machining process. In [31] it is described how RCA is performed when there is a low understanding between parameters and the problem. The idea builds on creating associations between the parameters and the problems which can be derived by feature importance from ML classification models. They explain that parameters with a strong association with the problem are likely to be root causes and can be further examined by experts. The requirement for applying this approach is to have a labeled data set so supervised learning models can be used to learn the relations between input parameters and the labeled quality outcome. It is further explained in [31] that classification models with an interpretive structure should be used to fully understand which features have the highest importance for the problem. The interpretability aspect when applying ML for RCA is elaborated on by Mueller et al. [33]. They make a distinction between white-box models and black-box models, where white-box models' structures can more easily be understood by humans. They mention Decision Tree models as a suitable example for RCA application as the feature importance easily can be extracted from such a model. Black-box models such as SVM or ANN should be avoided in these applications as they are highly non-linear and difficult to interpret [33].

Besides finding potential root causes through feature importance in classification models, it is also common to use association rule mining methods [17] to find relations between the input data and gain more knowledge of how these are connected. In [15], an association rule algorithm called "Apriori" is applied to mine relationships between operation parameters and quality outcomes. The extracted rules were able to show how combinations of values of different input parameters were related to the quality outcome.

It is important to be able to differentiate between correlation and causation when performing RCA, as only finding correlated values might not reveal the true root cause of the problem. The methods explained above are mainly useful for finding the correlations between input and output data and can act as decision support for domain experts to find causation among the parameters. To be able to identify causality in the data only through data analytics, graph-based methods such as Bayesian Networks have been successfully used for RCA in manufacturing environments as these methods can map sequential relations between factors [34, 35, 31]. In [30] it is described how a model can be developed for RCA of quality deviations

with Bayesian Networks. This approach however requires good knowledge of the process and previously recorded data where root causes have been identified. The framework includes a step called "structure learning" which has the aim to "deduce the structure of the network from the independencies and dependencies in the data". This is not an easy task for a complex manufacturing process they explain, even if knowledge from process experts can be used as guidelines and constraints.

2.2.3 Prediction of quality outcome

Being able to have the predictive capability and determine what is going to happen in the future can have high value in quality-related problems. It is not only product quality itself that causes costs and inefficiency as described in [18], but also the production steps needed to finish the product. If the final quality outcome can be predicted early in a manufacturing line, the product can be removed to avoid costly unnecessary production steps, or corrective actions can be taken. The difficulty with early predictions they further explain is that relevant production parameters need to be collected for all machining stages in a manufacturing line, which results in large quantities of data to analyze. Due to the many interrelationships between machines and other units in the manufacturing line, more advanced modeling approaches might need to be applied to account for these [15].

In [36] Peres et al. claim that product dimension variability is one of the most challenging problems in multistage manufacturing systems. The complexity of such systems and the random nature of variations and disturbances make it difficult to guarantee good quality outcomes. Therefore, they propose a method for achieving automated early prediction capability. In a case study performed at a company in the automotive industry, they developed an ML model that predicted the quality outcome at an assembly line. The model took inputs from measurement values collected at an early stage in the line and compared these values to the ideal values from the product's CAD model. This data could be used to determine if the product would have problems with dimensional integrity downstream in the line. A binary classification model was used to label the products as "OK" or "NOK". A domain expert could then use this information to assess whether the product should be discarded or if corrective measures could be taken to avoid quality issues.

2.2.4 Process optimization for improved quality

Process optimization for quality outcomes relates to the prescriptive capability to determine the optimal process parameters to avoid quality problems. As described in Section 2.1.1, prescriptive analytics is usually realized through a combination of predictive ML models and optimization models. The components needed to deploy an optimization model for machining are presented in [37]. First, a monitoring module that collects data and extracts features is needed. Second, an estimation module is needed to predict the state of the machining process. Third, an optimization unit optimizes the machine's parameters for the goal of the case based on the estimated state. Fourth, a CNC interaction module is used to send the optimized parameters to the machine, if the machine should adapt its parameters autonomously.

The optimization module for these methods can use a variety of approaches, such as Newton's method or gradient descending algorithms to find a local optimum as described in [18]. The other approach they describe is to use evolutionary algorithms which leads to global optimums in the search domain. In [38] an ANN model was developed coupled with a genetic algorithm to predict and optimize machining parameters for minimization of surface roughness for a milling machine.

2.3 Challenges with implementation of data analytics

The large amounts of data produced from machines and different manufacturing software can reveal important insights which can lead to increased business value. But there are challenges to overcome to be able to gain value from big data within manufacturing. Big data is often described with five main characteristics referred to as the 5 V's [3]: Volume, velocity, variety, veracity, and value. In [39] these characteristics are described. The first of the 5 V's, volume, refers to the size and amount of data being collected. Velocity refers to the speed at which data is generated and how quickly this data moves. Variety means the diversity of data types and sources. Veracity refers to the quality and accuracy of the data. And lastly as described in [39], value refers to what value the data can provide and what the organization can achieve with collected data.

When manufacturing companies start making use of advanced data analytics in their operations, it is commonly approached step-wise where pilot projects are conducted before trying to scale its usage [6]. Kirmse et al. [40] divide manufacturing factories into greenfield and brownfield. They explain that a greenfield factory is relatively new with few constraints present, while a brownfield factory is older with more constraints and is therefore also more challenging to upgrade. They further describe that it is important to not assume a factory to be greenfield with best case scenarios when investigating how data analytics can be utilized. It is pointed out that normally when factories are built their operations are planned over decades and the manufacturing environment evolves over the years. Therefore, due to high investment costs and amortization times, new investments in machines or manufacturing software are not an option at an early stage [40]. In this project, the challenges which are focused on are those that are mainly relevant in brownfield factories to propose solutions in a modular approach.

Literature shows that certain areas are of especially high relevance to enable efficient use of data analytics in brownfield factories and deal with the challenges associated with the 5 V's. One of these areas is *connectivity and data acquisition*, which enables the collection of large amounts of data and fast enough machine-to-machine communication (volume and velocity). *System integration* is also of importance to be able to collect data from the many sources in manufacturing environments and handling with heterogeneous data, meaning data with a lot of variation in structure and type (variety). By being able to track units in the factory, data from various sources

can easily be associated with specific units, and therefore *traceability* is needed to be able to solve a larger range of complex problems (value). Commonly, collected data from machines are subjected to errors and disturbances, and therefore an evaluation of *data quality* is also needed with appropriate measures so the data can be trusted (veracity). As manufacturing is a complex area with data and behaviors that requires experience to understand, *Internal knowledge and interpretability* is also important to address to ensure that only relevant data is collected for certain problems and that the analytics results are relevant for the business problem (volume and value). The remainder of this section discusses the challenge areas described above and highlights potential solutions in these areas to improve the prerequisites of performing data analytics projects.

2.3.1 Connectivity and data acquisition

Having the possibility to collect large amounts of data from machines and other relevant sources is a key aspect to be able to work efficiently with data analytics and solve a diverse range of problems. This is a common challenge in manufacturing and of high importance as availability, quality, and composition of data will have a strong influence on derived analytics models [9]. To achieve higher connectivity, cyber physical systems (CPS) are commonly discussed which is a system that connects physical objects with virtual objects and thereby creates an information flow from machines to information processing units [41]. Digital thread and digital twin are two other concepts that have been popularized in manufacturing with Industry 4.0, which creates a connection between physical space and cyberspace in manufacturing with the help of CPS and Internet of Thing (IoT) sensory tools [3]. One of the goals of these concepts is to create machines that are "self-aware" [16] with predictive capabilities.

In existing manufacturing factories, machines are normally kept as long as they continue to produce with high enough efficiency and quality, and it is not uncommon with machines that are 50 years or older on shop floors [7]. A modular approach for enabling data analytics needs to be adapted for implementation in such brownfield factories without the need to make significant changes in the legacy system [40]. Due to the complexity of a mix of older and newer machines and systems, no single standard exists to integrate these systems [41]. Machine vendors sometimes use communication and protocols which is only usable for their products [40]. To enable machine-to-machine communication, standardized communication must therefore be established that can integrate and combine different standards from machines so these machines can be kept and provide production data for as long as they remain productive [41]. An example of a standard interface that has been popular during the Industry 4.0 development is OPC-UA [7]. It is an uprising communication protocol that aims to harmonize information flows from the shop floor, where vendor-specific information can be described according to this standard [40]. The technology builds on a client-server architecture, where clients request data and servers respond with that data, which makes it possible to have data transmission between heterogeneous system components [41]. Besides OPC-UA, MQTT is a common protocol to use for

connectivity in manufacturing. In [42] a comparison is made between OPC-UA and MQTT. The main benefit of OPC-UA is that it holds more features than MQTT and can have servers communicate data with object orientation which is useful for a lot of different client applications. They describe that the main difference between the standards is that MQTT builds on a publisher-subscriber model with a broker in between, that distributes data from the publisher to the subscribers. An advantage with MQTT as described in [42] is that the information packages are lightweight compared to OPC-UA, which can make this protocol better suited for networks with limited bandwidth.

Apart from being able to connect the machines, there are also requirements for being able to handle the large volumes of data collected. When advanced analytics methods are used related to AI, large sets of training data are required to create well-performing models. The data needed will differ for different use cases, and it is, therefore, ideal to have access to all available information to increase the chances of finding new unexplored relations in the data during analysis [40]. Commonly, companies have too few samples of collected data to be statistically meaningful when presented to a data analyst, and it is therefore important that companies take a long-term focus and invest in systems that enable the collection of large data sets [10]. With growing complexity in sensor equipment and information technology, the amount of available data is growing as well. To handle the problem with storage and computation of this data it might be necessary to move to big data platforms with technologies such as cloud computing [17]. One of the main advantages of cloud computing, as described in [43], is to enable an agile approach to collecting on-demand data with advanced analytic capabilities. It is also pointed out that cloud storage is scalable, which allows companies to only pay for the capacity they need and where it can be increased or decreased over time.

It is also important that large volumes of data can be collected and transferred from the machines to a database at high enough velocity. Being able to collect data with high velocity makes it possible to collect data with higher frequency. From high-frequency data, it might be possible to discover patterns that otherwise would not be visible. In [8] it is discussed that different data should be collected at different frequency rates depending on the need. Temperatures can be collected at a low sampling rate while for example vibrations generally needs to be collected with sampling rates under one second. They, therefore, argue that appropriate sampling rates are selected to balance the high detail in the data with the ability to store long-term data. Due to limitations in the network, it might be appropriate to use edge computing to increase the velocity of collected data [44, 45, 11]. Unlike cloud computing, edge computing takes place closer to the machine and has therefore the advantage of significantly reducing latency and network bottlenecks [44]. Low latency is a crucial aspect in manufacturing use cases to be able to control processes in real time with advanced models, which would not be possible if the data need to travel to a distant data center [45]. Another benefit of edge computing is that it can be used for older machines that lack OPC-UA or MQTT compatibility as a protocol translator, thus enabling machine-to-machine communication for these machines [45].

Edge and cloud computing are both promising solutions for collecting and analyzing large amounts of data with different advantages. The techniques should be used complementary where edge computing is used for high-velocity data collection and real-time analysis, while cloud computing is used for the collection of large volumes of data from different sources for analysis and modeling purposes [7]. Chen [46] elaborates on this and states that edge computing should be focused on real-time or short-term data analysis and cloud computing focuses on big data analysis by integrating edge computing networks. They mention that another benefit of this arrangement during data acquisition is that data can be stored temporarily on the edge device and therefore less data loss will occur during network failures.

To increase the potential of data analytics in manufacturing it is also important to be able to collect the outcome for the problem being solved, as this enables supervised learning methods. The outcome could for example be if a product is defective or not to a quality problem. This information could be vital to have on a diagnostic level to examine correlations between process parameters and quality outcomes. Sometimes automatic measurement systems are already in place at production lines for producing this data, but for some problems, it might be necessary to find new ways of collecting relevant labeled data. Alexopoulos [47] says that a limitation of data analytics in manufacturing is that manual labeling by humans is often needed to acquire the necessary information for solving specific problems that lack automatic collection. They claim that there is a high risk of receiving errors in the data and that it becomes labor extensive to label the large amounts of data that needs to be collected. It might therefore be necessary to install new sensory equipment to get around these problems and label the data automatically, where different levels of sophisticated techniques will be required. As an example, in [14] it is mentioned that vision-based systems can be used for automatic product surface evaluation to produce labeled data for ML applications. Another approach to collecting labeled data as described in [47] is to create synthetic data sets with simulation software, which has the benefit of creating large data sets in a short time. This approach would however only be suitable for some cases where detailed sensory information is not needed. Wuest [48] argues that unsupervised and supervised learning can be used in combination when there is a low understanding of the data and labels are missing, and that this approach is suited for quality-related problems. It is suggested that cluster analysis is first performed to divide the data into subsets, so data in the same cluster possess similar characteristics. They claim that this technique is suitable when there is no specific class that should be predicted but when instances from machine runs can be divided into natural groups. These groups can thereafter be named and used as labels when applying supervised learning.

2.3.2 System integration

To make it possible to efficiently analyze data from multiple sources, system integration plays an important role. System integration refers to the ability to combine operation technology (OT), meaning the hardware and software used for controlling the manufacturing operations, with information technology (IT) [49]. Except

for the data collected from machines on the shop floor, OT also includes the wide range of software used to plan and control manufacturing such as Manufacturing Executive System (MES), Product Lifecycle Management (PLM) and Enterprise Resource Planning (ERP) [3]. Smart manufacturing cannot be achieved only by traditional manufacturing software as these are developed by different vendors and cannot cooperate, and the sensory data from the shop floor must also be considered [3]. Furthermore, it is a common problem in manufacturing that data originating from different departments such as production, maintenance, and quality is disconnected and siloed [49]. This means that it will not be used for anything outside of its original intended purpose even if valuable insights could be gained by combining and analyzing information from these sources. To create better conditions for data analytics, it is important to use common data by breaking down silos and integrating data both vertically and horizontally, so all users have access to the same data and a "single source of truth" is established [7]. The data should be collected centrally with a clear structure so data analysts easily can access the relevant data and derive insights from the information [10]. There are two main approaches to integrating systems and breaking down silos according to Cui et al. [3]. The first is to use data warehouses, which is a traditional approach to integrating data from various databases into a centralized data store where the data is stored in a structured format. The second approach they mention is to use data lakes which is a newer way of creating centralized storage where data is stored unstructured in its raw format, which can make it easier to store heterogeneous manufacturing data. They also mention however that data management tools should be used in combination with data lakes to avoid creating a "data swamp".

Apart from being stored in multiple sources, a challenge with integrating data is the heterogeneous characteristics of manufacturing data. Various data types are produced on the shop floor such as sensory data, product records, pictures, and logs, which can take a structured, semi-structured, or non-structured form [1]. When using ML, the success of the applied algorithm is highly dependent on a well-structured data set, which can be difficult to obtain in manufacturing environments due to the heterogeneous nature of the data according to Dogan and Birant [17]. To better handle the large volumes of heterogeneous data in manufacturing, it is common for companies to use complete platforms for the integration of IoT units which makes it easier to provide analysts with the information they need in a structured format [7]. Capgemini [11] says that such IoT platforms simplify the effort of data governance which is needed to create clean and structured data sets. They argue that the platform should be feature rich enough to store data, govern access to it, and develop AI applications from it. Taisch et al. [32] elaborate on the ability to store and manage analytic models on the platform and its necessity for large-scale use in manufacturing companies, as it would result in large amounts of custom data sets and models to keep track on. They claim that a platform that can unite all models across lines, departments, and plants with a standardized development approach can significantly reduce the development time of models. In [50] it is described how such a platform can be developed by combining different units. This platform includes a unit for the collection of raw manufacturing data which is structured based on

attributes into a data warehouse. A manufacturing application unit is used to connect manufacturing applications such as MES and PLM to communicate with the platform. A unit consisting of an analytics center is used to make queries from the warehouse and the manufacturing applications to create data sets which then can be used for creating models that are also managed in the analytics center during its whole life cycle. To achieve efficient system integration, it is required to both solve technical problems as well as create an internal company culture that understands the necessity of changing ways to store and manage data to realize the value of data analytics according to Dallemand [49]. He mentions that for solving the technical problems, there are many IoT-platform solutions available on the market specifically developed for integration and analysis of manufacturing data

2.3.3 Traceability

Being able to track products and processes is an important aspect of Industry 4.0 to achieve a detailed overview of operations. Product traceability is used for determining the location of a product while process traceability establishes what type, sequence, and variables from processes have affected a product [51]. Digital traceability is used in several business areas including production and according to Beliatis et al. [52], its main purposes are identification, tracing, and tracking. It is an important part of the connectivity ecosystem they argue, where proper implementation allows for better possibilities to optimize production processes through data analytics. In [51], traceability is described as a "product life cycle and risk management tool", which is needed to create well-functioning smart factories integrated with IoT devices. They bring up as an example that traceability plays a crucial role in predicting and diagnosing problems in product quality. Wang [53] mentions that the importance of traceability for solving quality-related problems comes from the ability to trace data from several machines and systems to defective products. This makes it possible to identify problematic steps in the manufacturing line that might cause these defects.

As was discussed in Section 2.2, it is common to collect data from multiple processes among a manufacturing line when data analytics is used to improve product quality in multistage manufacturing systems (MMS). The final product quality depends on how the product was processed at all stages, and therefore series of sensor measurements from each process step might contain quality-related patterns [54]. In [15] Kao et al. note that quality inspection through data analytics is commonly applied for just single workstations, but in MMS several factors such as equipment, operators, and parameters from several stations can have cumulative and interactive effects on the product quality that is not considered if only data from one station is collected. They, therefore, used a Cascade Quality Prediction Method (CQPM) in a use case to find relationships between stations in a manufacturing line and create a model for these. There are three important relationships to consider in such a model they further explain: Relationships between operations in a station, relationships between different stations, and relationships between operation variables and the final product quality. Different approaches to using CQPM to solve complex

quality-related problems in MMS have been applied with successful results in other studies as well [54, 55, 56]. Traceability is a requirement for enabling these modeling techniques.

Different technologies can be used to achieve traceability in manufacturing. These can be divided into two main categories according to Schuitemaker and Xu [51], indirect marking and direct marking of products. Indirect marking is often used for larger products and means that a tracing item is temporally placed on the product they explain, where Radio Frequency Identification (RFID) is among the most common technologies. RFID consists of a transmitter and a receiver, where the transmitter is a wireless tag placed on the product which automatically transmits product data to the receiving tag readers [52] which are placed at several locations along a manufacturing line. Direct marking is usually applied for smaller products and means that the surface of the product is marked with either a numeric, alphabetic, bar, or 2D code with for example laser engraving [51]. These codes can then be read by for example cameras located among the manufacturing line.

2.3.4 Data quality

Data quality relates to the veracity characteristic in the 5V's of big data. Proper quality of collected data is essential to be able to gain important insights through analysis and to be able to trust the obtained result. Inferior data quality might result in inaccurate feedback from models and will undermine the confidence in analytic models, which makes them unusable for data-driven decision-making [7]. In manufacturing, several quality issues need to be addressed in data analytics projects, such as missing values, outliers, noise, and imbalance [3].

Missing values is one of the most common problems in manufacturing data and can depend on machinery problems or measurement errors from certain sensors [17]. The missing values become a challenge during the development of ML models as there will be gaps in the information for many of the collected samples, and therefore approaches for filling these gaps might be necessary for a way that minimizes bias and negative influences as much as possible [9]. A common approach for filling missing values is to replace the missing data with mean values from all instances [15]. If certain parameters have too many missing values, it might be necessary to exclude these parameters from the model to avoid bias [57].

It is also common that data collected in manufacturing is noisy and erroneous due to interference from the environment, malfunctions from sensors, and periodic disturbances in communications [1]. The performance of ML models might be reduced by outliers arising from the noisy data and is therefore often needed to remove such outliers by for example applying smoothing techniques [58]. Preprocessing techniques must be used both during modeling to obtain well-functioning models, and also that noisy data is automatically detected and filtered out after the model has been deployed. Erroneous data points might cause costly disruptions in production otherwise due to false alarms [8].

Imbalanced data is a common problem mainly when developing supervised ML models and means that there is a significant imbalance between the classes in the training data set. In manufacturing, unbalanced data sets are common when dealing with product quality problems, as there will be much more samples collected in a steady state than samples where a quality issue occurred [32]. This becomes problematic as the trained model then will be biased towards predicting the quality outcome as approved due to the overrepresentation of these samples in the training data. To deal with unbalanced data, the first and best solution according to [59] is to collect more data to increase the number of underrepresented samples in the training data set. To balance the data, they suggest that oversampling can be used to remove samples of steady state data while undersampling can be used to increase the number of problematic samples by creating duplicates. Another common balancing technique that was applied for a product quality problem in [17] is a synthetic minority oversampling technique (SMOTE), which creates new synthetic samples of the underrepresented class.

Collected data sets in data analytics projects needs are prepared with preprocessing which often includes data cleaning, normalization, and transformation of the data that decreases the uncertainty due to the problems described above and other data quality issues [60]. It is usually a time-consuming task in the project to achieve a structured data set from heterogeneous manufacturing data where appropriate preprocessing steps are applied. These preprocessing steps will have a critical impact on the result of the model, but there are no standardized approaches for which preprocessing methods to use, and therefore the appropriate techniques need to be determined by trying different options [17].

2.3.5 Internal knowledge and interpretability

Besides having the requisites to collect relevant data and structure it into a usable format, it is also important to have the internal knowledge and competence to analyze the data and retrieve insights from it that can improve business goals in manufacturing. Advanced data analytics in manufacturing is a multi-disciplinary area that requires knowledge both in data science and manufacturing systems as described in [17]. They further describe that a lack of understanding of the collected data might harm the result, and therefore domain experts are needed to clarify the meaning of the collected parameters. In [11] it is described that a talent pool with both a deep understanding of the manufacturing domain and data analytics is required to properly exploit the data and to understand how and which business problems can be solved by data analytics.

A common challenge with data analytics in manufacturing is the acquisition of relevant data [9]. Sometimes only a small part of all the data which can be collected from machine measurements will be relevant for a specific problem, so collecting all data would result in large volumes of useless data that the data analyst would waste time on [17]. It is therefore important to address what data should be collected with the help of manufacturing experts for each use case to reduce the volume of the initial data set.

After data has been collected it is also important to possess the right analytics skills in the company. According to [11] however, the skillsets needed for advanced data analytics are usually not found within manufacturing companies' IT departments, and therefore hiring new staff and upskilling current employees is needed to meet these requirements. A solution commonly discussed to make data analytics more accessible to manufacturing experts is to include more interactive tools as part of the IoT platform [61]. A way to achieve this is to create user-friendly tools developed through third-party collaborations which can be used by manufacturing experts to scale the digital transformation process in a company [61]. In [62] a concept of such a tool is developed to solve a variety of production problems. This concept includes a preprocessing engine that has rules that easily can be applied to an uploaded data set to find the best preprocessing methods for a specific use case. Cui et al. [3] states that a drawback with this approach however is that it would be difficult to develop a "one size fit all" tool that would work for all possible use cases. They assert that due to the large amounts of data with different characteristics in manufacturing environments, it would require a very complex tool to be able to address all the prepossessing and modeling techniques that will be needed. It might therefore contradict the goal of making the analytics process more user-friendly for complex ML solutions where high model performance is required.

Another challenge is the interpretation of the results obtained from a use case, where the output needs to be presented in a format that is suitable for the end user [9]. The real value of data analytics is created first when the insights gained from data can help employees to support their work. It is therefore important to consider that the data should be prepared and presented in a user-friendly manner that supports decision making [7]. In [16] it is described that an important aspect of a data analytics platform is to be able to convey the results in this way. They suggest that the platform should include functionality to present the result through visualization tools and clear statistical values which could be used by for example a machine operator. They further suggest that this information should be made available for management systems such as ERP and MES to achieve transparency of the manufacturing operations which management can use to determine facility-wide OEE.

2.4 Methodologies for data analytics projects

To scale usage of data analytics in an organization, a pilot stage must be reached where its value has been proved and the development process has been made standardized and repeatable [11]. Several methodologies have been suggested in the literature for efficiently integrating the use of data science in a standardized manner within organizations across different domains, such as Knowledge Discovery in Database (KDD) and SEMMA (Acronym for Sample, Explore, Modify, Model, and Assess) [63, 64]. A drawback with these methodologies however is that they lack integration of organizational management [65]. In industrial contexts, it is important to include business understanding and expertise when conducting data analytics

projects as discussed in the previous section. Without proper domain understanding, there is a risk of only detecting trivial or already known patterns which are not relevant to the business goal [66]. For this reason, the cross-industry standard process for data mining (CRISP-DM) has become widely spread for data mining in an industrial context, which is a framework for translating business problems into data analytics goals [67]. In a survey conducted by KDnuggets in 2014 [68] it was shown that CRISP-DM is the most popular method for data analytics projects used by 43% followed by using a self-developed methodology at 28% by the asked companies. Due to the popularity of CRISP-DM and its suitability for industrial applications, this methodology was evaluated more thoroughly in this project to evaluate its applicability in manufacturing.

2.4.1 CRISP-DM

CRISP-DM constitutes six phases that are performed sequentially: Business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The flow between these phases can be seen in Figure 2.3. A detailed description of the method is provided by IBM in [69] and a brief description of the phases is presented in this section.

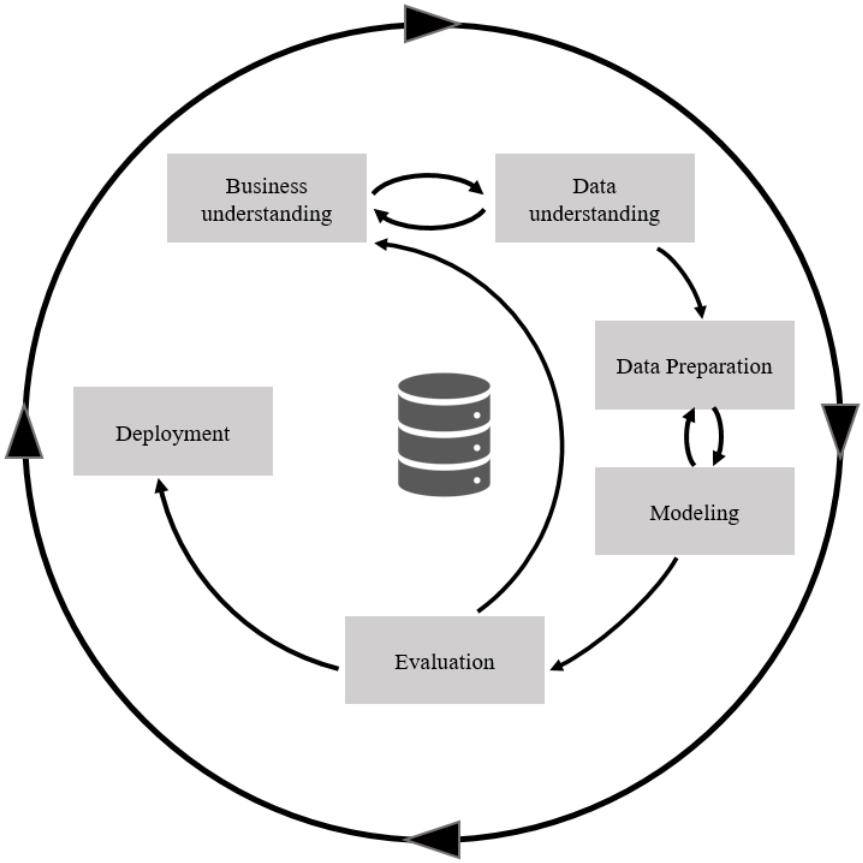

Figure 2.3: Flowchart of the CRISP-DM methodology inspired by [69]

Business understanding

The initial phase, business understanding, is where an understanding of the project objectives from a business perspective is formed and then converted into a data mining problem [70]. The usual steps to perform in this phase according to [69] is to first define the business goals and criteria. Then an assessment of the current situation is made by listing resources, defining requirements, reflecting on potential risks, and evaluating cost versus benefit for the project. Thereafter the data analysis goals and criteria are defined and lastly, a project plan is produced.

Data understanding

The business understanding phase is followed by creating a data understanding, where the aim is to collect and then evaluate data. After the initial collection, a general description of the data is made and then followed by an exploration of data [69]. The exploratory data analysis (EDA) is a task for finding relations and characteristics in the data by using visual methods to discover patterns and test hypotheses [57]. EDA is also used to understand the potential quality issues that might need countermeasures, which are dealt with in the next phase.

Data preparation

Depending on the identified quality issues with the data, appropriate preprocessing steps are performed to convert the data into a usable format for the model algorithm [9]. A common way to perform prepossessing is to follow four steps: Data cleaning, data integration, data transformation, and data reduction [71, 72]. In the data cleaning step, the incomplete, inaccurate, or irrelevant data is identified which is then modified or deleted [57]. Common quality issues in manufacturing that are dealt with in this step and different remedies were described in Section 2.3.4. In the following data integration step, data from different sources are combined into one table that can later be used as input for the model. In the data transformation step, the format of the data is changed by different methods such as normalization [72] or aggregation [71] to make the data better suited for the algorithm being applied. In the last step, the dimension of the data is reduced to avoid overfitting of models (when the model fits too well with the training data) and to extract relevant features.

Modeling

In the modeling phase, the aim is to select appropriate ML algorithms and train models with the features extracted in the previous phase. A common approach for selecting algorithms according to [9] is to first choose from the available data and the type of problem between a supervised, unsupervised, or reinforcement learning approach. Thereafter, the suitability of models is examined by the type, quantity, and structure of the data, and by investigating which models have been successful in similar cases. Several models are chosen this way and after they have been trained, a tuning of model parameters is performed to optimize their performance before further evaluation.

Evaluation

During the evaluation phase, the models that were obtained in the previous phase are evaluated to find out if they fulfill the business criteria specified in the business understanding phase [70]. Certain metrics can be used to give a comparable view of model performances so the best-suited model can be chosen for deployment. In this stage, it is also common to review the process to see if something has been overlooked that could further improve the models before deployment [69]. The phase ends by deciding whether the models perform sufficiently or if the process should be reiterated to fulfill the business criteria.

Deployment

The deployment process aims to use gained insights from the project to implement improvements in the organization. Common steps in this phase are to plan the deployment by creating a strategy for implementing the models in production, planning monitoring and maintenance of the models, and creating a report with an overview of the results [69].

2.4.2 Review of CRISP-DM in manufacturing

The CRISP-DM model can be viewed as a general guideline for planning and documentation of the data mining process. The phases do not have to be followed in strict order and backtracking between phases is often necessary [70]. Furthermore, the generic model often needs to be adapted to the domain where it is executed to consider the domain-specific requirements. Commonly discussed topics when applying CRISP-DM in manufacturing is to involve experts within different fields in the manufacturing domain in different steps of the process, and that higher emphasis needs to be put on data acquisition [64].

Huber et al. [67] mention that several essential tasks need to be executed within the manufacturing domain to acquire the relevant data for the data analytics tasks. They, therefore, propose an extended version of CRISP-DM which considers data acquisition. This was done by adding two phases to the process which they call *technical understanding* and *technical realization*, which are performed between the business understanding and data understanding phase. The goal of the technical understanding task is to create a better understanding of the analyzed system and to define the important physical parameters that need to be measured and collected. This is done by involving experts in the production system with a better understanding of the problem and important data. The technical realization phase then aims to define a plan for how this information should be collected where it is important to consider possible sources of errors and the level of data quality. The process for acquiring relevant data is thoroughly described in [67], but they do not discuss how to reason when it is difficult or costly to realize the collection of data that is interpreted as important.

The challenge with acquiring all the data that might be needed is elaborated on in [5] by Ungermann et al. where they mention that CRISP-DM is not supported by

production-related methods or tools and does not specify how to collect additional important data. They suggest that an ideal data set is obtained with the help of production experts containing all data that might be needed. From this data set, only the data with high utilization and feasibility is collected initially. An iterative process is proposed so that the data set can be extended later, if necessary, with more factors from the ideal set by the installation of new sensors or integration of data from other systems.

Data acquisitions for large-scale usage of data analytics in manufacturing are also discussed by Kampker et al. [73] who agree that it is not a viable solution to collect all data in brownfield manufacturing factories due to costs and limitations. They, therefore, introduce the concept *adaptive data availability* which means that only the data that has been identified as relevant by the production experts is made accessible. To further facilitate large-scale usage of data analytics an approach for prioritizing cases is described in [73]. The prioritization is made by estimating the cost and effort of collecting relevant data for several potential causes, and then this cost and effort is weighted against the estimated benefit from the same case. The cases can then be ranked so that cases with high benefit and low effort are prioritized to maximize the value of the data analytics projects.

Kristoffersen et al. [63] mentions that another important aspect of large scale data analytics is to be able to re-use gained insights and addresses how a standardized method based on CRISP-DM can be made easier to understand with improved repeatability. They argue that the general CRISP-DM method lacks a proper management view and communicates knowledge insights. They also argue that the method is often not fully understood from the business side, and it might therefore be difficult to gain actual business value from the data analytics projects. To deal with these issues they propose the use of *analytic profiles*, which can be seen as a standardized structure to re-use analytic insights. These profiles should describe the best practice for a particular mechanical process or problem and should contain information including the goals, domain knowledge, and data sources that were used in a case. Such profiles can re-use insights from previous cases and thereby make the development process for the following cases more efficient. Furthermore, it is discussed in [63] the importance of involving domain experts not only in the early stages but in the later stages of CRISP-DM as well. During the data preparation phase, the data often changes characteristics and large quantities of data are deleted. They, therefore, suggest that a validation phase is added after the data preparation phase where the domain experts verify that the altered data still is a good representation of the original business problem, where it might be important that extracted features are understandable for the production experts.

Besides involving production experts in different areas, it is also seen in the literature that it is important to include the usage of domain-specific tools in CRISP-DM. For creating a better understanding of the problem and to identify relevant data, it is appropriate to use brainstorming methods commonly used in manufacturing such as Ishikawa diagrams, Failure Mode and Effects Analysis (FMEA), and Five Whys (5W) [67, 5, 73, 74]. Schäfer et al. [65] also argue that production-specific methods

should be used alongside all phases of CRISP-DM and describes how this can be achieved for quality management by combining the popular six sigma methodology, DMAIC, with CRISP-DM. The article thereby gives good insight into how to combine data analytic skills with manufacturing expertise. Their proposed method is however specifically developed for quality management, and it is not shown how the method can be adapted to be repeatable for other problems.

To conclude, CRISP-DM is a general process to perform data analytics projects with a connection to business goals. It has proven its value within manufacturing contexts and is one of the most popular methods. Because of the genericness of the model, several enhancements have been suggested for the model in literature to make it better suited and easier to follow in the manufacturing domain. Some of the discussed improvement areas a descriptive data acquisition phase, how to involve domain experts in different phases, how to involve manufacturing specific tools and methods complementary, and how to increase the repeatability of the process to facilitate large scale usage of data analytics. So far, no comprehensive data analytics process covering all these areas has been developed specifically for usage in the manufacturing domain. This might make it difficult for new or aspiring practitioners of data analytics within the manufacturing industry to make valuable use of this technology.

The methodology developed in this project contributes to the creation of a holistic process that is adapted to the challenges within manufacturing. This methodology is an adaptation of CRISP-DM and integrates several of the suggested improvements found in the literature. An overview of this adapted methodology is described in the following Chapter.

3

Adaptation of CRISP-DM for manufacturing

As was shown in the previous chapter, the manufacturing domain faces unique challenges to start making use of the large quantities of data produced on the shop floor for valuable analytics. There are several methods describing how to conduct data analytics projects in a structured way where CRISP-DM is one of the most popular and has shown to be prominent within manufacturing. In this chapter, a proposed methodology which is an adaptation of CRISP-DM specific for manufacturing is presented. This methodology combines suggested adaptions described in Section 2.4.2 to provide a holistic framework that can be a basis for manufacturing companies to integrate data analytics in their manufacturing operations.

3.1 Enhanced methodology

The goal of the enhanced methodology is to provide clarity into how CRISP-DM can successfully be implemented in manufacturing to have a structured and repeatable process that enables large-scale use of data analytics. The main difference with the enhanced methodology is that higher emphasis is put on re-using previous insights, describing how and what data to collect, and how to achieve higher integration of domain experts in the process. An overview of the methodology can be seen in Figure 3.1. The orange boxes show the alterations that have been made to the original CRISP-DM process in terms of added phases and tasks.

To increase the repeatability, the concept of analytic profiles suggested in [63] is integrated into the process. Each time a case has been completed a new analytic profile is created. The profile should have a standardized structure where information that can be re-used for the following cases is provided. Since large manufacturing companies usually have many different machines from different suppliers, the prerequisites for collecting data will be different depending on the machines involved in an analytics project. This combined with the fact that different data sources will have varying importance depending on what type of problem is being solved makes it important to save insights gained from previous projects so these can be applied to similar projects to come. This will make it easier to repeat the successful aspects of previous cases and avoid previous pitfalls.

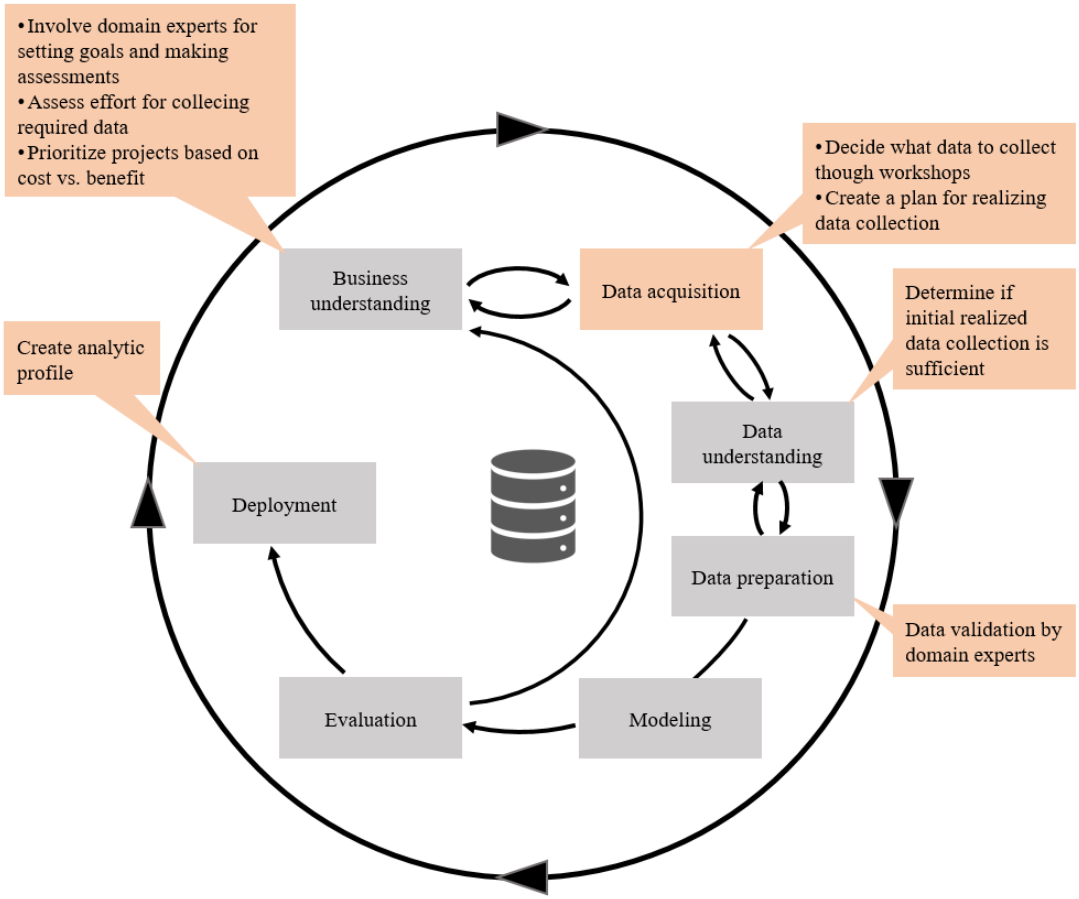

Figure 3.1: Enhanced CRISP-DM for manufacturing with altered phases and added data acquisition phase

Due to the challenges associated with data collection in manufacturing, it is motivated to develop a more detailed data acquisition phase. Therefore, a phase dedicated to this has been added in the enhanced methodology between the business understanding and data understanding phases. This phase constitutes two tasks, technical understanding, and technical realization, in conformity with [67]. The technical understanding phase focuses on deciding what data is necessary to collect while technical realization aims to develop a plan for collecting this data.

To utilize domain expertise later in the process as well, a validation task is also integrated into the process as suggested in [63]. This task is performed at the end of the data preparation phase, where the domain experts will validate that the preprocessed data still makes a good representation of the business goals.

Apart from these additions, the enhanced methodology proposes which people to involve during different phases to utilize domain knowledge. It also shows how use cases can be prioritized to make more valuable use of resources spent on data analytics projects in large-scale manufacturing.

3.2 Altered phases from CRISP-DM

As could be seen in Figure 3.1, the changes from the original CRISP-DM methodology take place in the following phases: Business understanding, data understanding, data preparation, and deployment. In this section, important considerations when applying CRISP-DM in manufacturing are described for these phases. These considerations serve as a complement to the original CRISP-DM methodology. Furthermore, it is described how the added data acquisition phase should be carried out.

3.2.1 Adapted business understanding

During the startup phase of the data analytics project, it is important to involve domain experts with different competencies to set feasible goals and make a reliable assessment of the current situation. Experts from production management are required to understand why there is a need to conduct the project so that business goals can be formulated in terms of production KPIs. Experts with more extensive knowledge of the production system are needed to create a better understanding of how different units in the system interact and to set the system boundaries of the problem being investigated [5]. It is also important to involve machine operators in this stage as they will have the best knowledge of how specific machines works and might have additional hypotheses of possible causes of the problem. Data analysts should also be included at this stage as they must gain a better understanding of the data and estimate the significance of later interventions [64]. The data analysts also have the best understanding of how to formulate feasible data analytics goals. It might also be needed to make tradeoffs where the business goals and data analytics goals are adjusted until they can meet the same target. For example, it is important to make sure that information is presented in a suitable and user-friendly format so it can be used to solve the business goals by the indented end users [7].

When an assessment of the current situation is performed, the resources that are expected to be needed in the project are listed and potential risks are evaluated. In this task, analytic profiles from similar cases should be examined to both gains a better understanding of important resources and what potential pitfalls in previous projects can be avoided. One of the important resources to consider is what data sources that likely needed to solve the problem [69]. Even if a more thorough evaluation of required data is performed in the next phase, it is needed to have some understanding of available data in the business understanding phase to be able to formulate feasible goals [70]. It is also important to consider what people to involve who have the technical competencies to realize the collection of the needed data sources. It might also be necessary to involve people with more specific domain expertise for the problem being solved, such as quality experts when dealing with quality problems.

After the required resources have been listed and an appreciation of the required effort for collecting and analyzing the data has been made, the cost of the project can be estimated. The estimated cost versus the estimated benefits can then be used to

put the value of the case in relation to other cases. Prioritization can be performed as suggested in [73] where cases with low cost and high benefit are performed first. It is, therefore, appropriate to perform several business understanding phases of potential use cases before starting projects to be able to find the most value-adding cases and to build a list of potential use cases that can be prioritized. An important consideration when evaluating the effort of collecting data is that several use cases might use the same data sources, and therefore a scaling effect can be achieved by realizing data that can be used for multiple cases where the implementation cost can be vastly reduced [73].

3.2.2 Data acquisition

The first task in the data acquisition phase is to gain a more profound technical understanding of the problem and the required data. In this phase, the same people should be involved as in the business understanding and it might be possible to loop between these phases back and forth to adjust the goals to the available data. Additional domain experts for specific problems should also be involved in this phase. Structured workshops are executed to set the definitive system boundaries and decide what data is needed, which is illustrated as the first step in the flowchart in Figure 3.2. By utilizing domain expertise in this stage, the risk of conducting complex analyses that only leads to the discovery of obvious or previously known patterns in production data is reduced [66]. It is important to consider during these workshops that the data needed should not be limited only to known relationships, but all data that potentially could have an impact on the investigated problem should ideally be collected to enable new patterns and knowledge to be extracted [7]. A suitable method should be chosen for brainstorming the data, such as Ishikawa diagrams, FMEA or 5Ws. Additional manufacturing-specific tools might also be used in this task if they can help to create a better understanding of the problem. The data used in previous projects that can be gathered from analytic profiles should lay the foundation before the brainstorming. If there are no previous relevant analytic profiles, similar cases can be studied from literature to gain insights into data that might be important [67].

The second task in this phase is to realize the collection of the identified data. As suggested in [5], the list of data obtained from the workshops should be regarded as the ideal data set where a subset is regarded as required data. The required data is the data where there are strong hypotheses that can have significant importance for the investigated problem. The collection of the required data should be realized first and determines if it is possible to proceed with the case or not. Thereafter, as much of the remaining ideal data as feasible should be made available. If there is data from certain sources that require high effort to obtain, a decision needs to be made if this data should be collected or not. The importance of the data should then be estimated and put in relation to the effort, so that data with high collection effort and believed low importance is not collected initially to save costs and time [5]. The data set is thereby reduced to a data set containing the required data and as much of the remaining data that seems appropriate to collect initially based

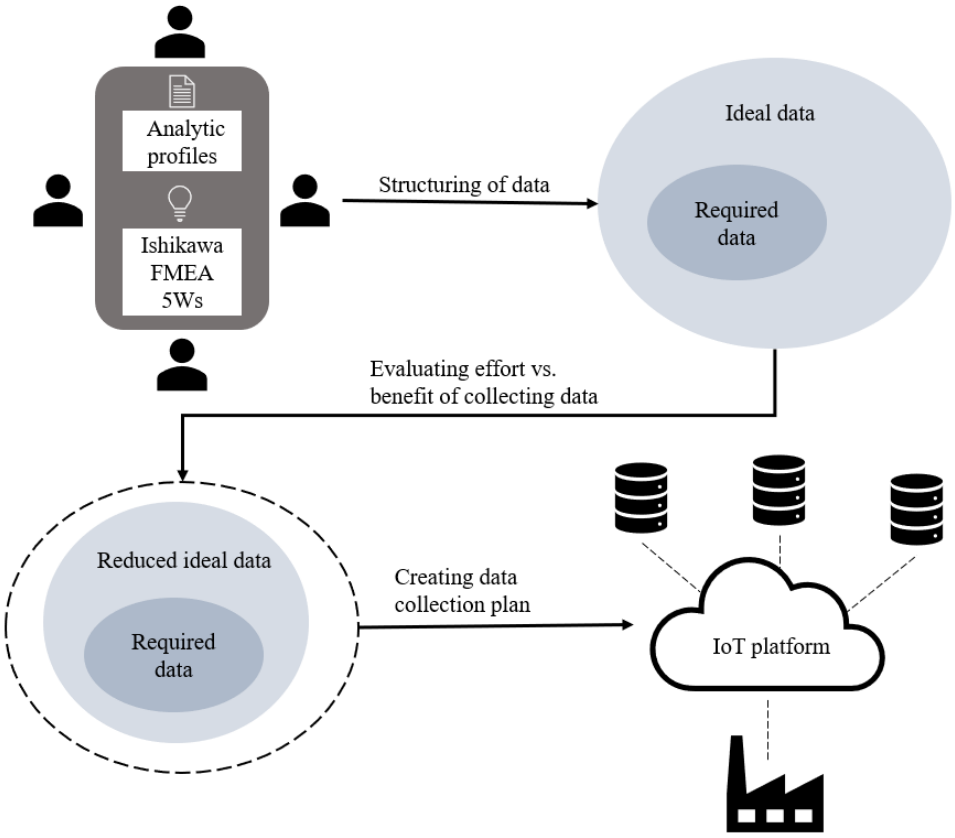

Figure 3.2: Flowchart of data aqcuisition phase starting with a workshop for defining relevant data, followed by obtaining a feasible data set that can be collected

on the evaluation. The different systems that this data belongs to are integrated by connecting the data in these systems so it can be collected in the company's IoT platform. From the IoT platform, the data can then easily be extracted and processed by data analysts.

3.2.3 Adapted data understanding

The EDA performed in the data understanding phase will aid to visualize the importance of initially collected data. This can be done by for example evaluating the correlations between collected factors and the labeled outcome if it is a supervised learning problem. As proposed in [5], if there are low correlations between factors and outcomes this might indicate that the collected data will not be enough to solve the problem with good results. It would then be appropriate to go back to the data acquisition phase and extend the data set being collected with more factors from the ideal data set.

3.2.4 Adapted data preparation

During the data preparation phase, appropriate preprocessing methods are chosen to extract features that can be used for modeling. The analytic profiles can provide

insights into what preprocessing rules are appropriate to use based on what methods have been successful for similarly collected data in previous cases. However, Burdack and Rössle [62] mentions that due to different positioning of sensors and external influences including heat, vibrations, and noise from adjacent machines, each machine will have its own "data fingerprint". By this, they mean that different patterns will be found in the data from many machines even if the machines are of a similar type due to these influencing factors. Therefore, each machine needs unique preprocessing rules. The analytic profiles should thereby mainly serve as a guideline while it is necessary to evaluate different rules iteratively to find the best features to solve the business goals.

In the added validation task, domain experts that will use the result from the project to perform improvement work are involved to assess the suitability of extracted features. This task is important if for example the model is not only used for optimized predictions but where a deeper process understanding also should be achieved. As discussed in Section 2.2.2, a clear understanding of features is often necessary to understand relations between factors in a machine when data analytics is used for root cause analysis. The domain experts would in such a case determine if the features are intuitive enough to make process improvements.

3.2.5 Adapted deployment

When the models have been trained and the evaluation has shown an approved result, the deployment phase is initiated. Besides creating a deployment plan for usage of the obtained results, a new analytic profile should be created that can be used for future projects. The analytics profile can for example be a structured document. It should contain all the information that can be advantageous to use in any following project. Some examples of what can be included in an analytic profile are given in [70]:

- Description of the use case and the goals
- KPIs that were used to measure the performance of the deployed model
- Important domain specific insights for the case
- Data sources and specific data from these sources that were used
- Applied preprocessing methods
- Models that showed the best results
- Lessons learned and improvement potential for future cases

The information in analytic profiles can be adapted over time when a company learns what insights are important to re-use. Adoptions can also be made as to how this information should be documented in the best way to enable more efficient execution of the following projects.

In this chapter, it has been shown what adaptations could be made to CRISP-DM to create a more holistic standardized process for data analytics projects in manufacturing. The goal of this proposed methodology is to remove some uncertainties, improve the quality, and enable more efficient execution of data analytics projects. In the following Chapter, this methodology is applied to a case study at Volvo Trucks which gives a more intuitive example of how this enhanced version of CRISP-DM is carried out.

4

Implementation of adapted CRISP-DM methodology

The purpose of this Chapter is to put the enhanced version of CRISP-DM into practice in a case study at Volvo. This was done to evaluate the effectiveness of the methodology in a manufacturing context, but also to evaluate Volvo's readiness regarding conducting data analytics projects. An introduction is given to the case study, which was about improving process stability for a machining process by reducing dimensional variations for machined holes. The results are then presented for the case study following a step-by-step implementation of the modified methodology.

4.1 Case study: Reducing dimensional variations for machined holes in cylinder heads

The case study was conducted at a production line at Volvo Powertrains's machining department where they manufacture engine components. The targeted process to improve is a multitasking machine in which exhaust intake and outtake holes in cylinder heads are machined. The factory in which the machine resides was built several decades ago where machines have been replaced gradually, but there are still many older machines in production. The machine investigated in this case is relatively old, but still highly productive. The problem with the machine is that the process is unstable which causes different quality-related issues to occur with a relatively high frequency. The most frequent problem is that the diameters of the exhaust holes are machined either too large or too small, resulting in defective products. The machine has caused high costs for the department due to disruptions and the many products that need to be discarded.

The machine is divided into four units, which all perform different operations but runs simultaneously. An overview of the machine process can be seen in Figure 4.1, where the parts of the system has the following purposes (corresponding to the numbering in the figure):

1. A conveyor is transporting cylinder heads from the previous machine on the line into the first unit of the machine

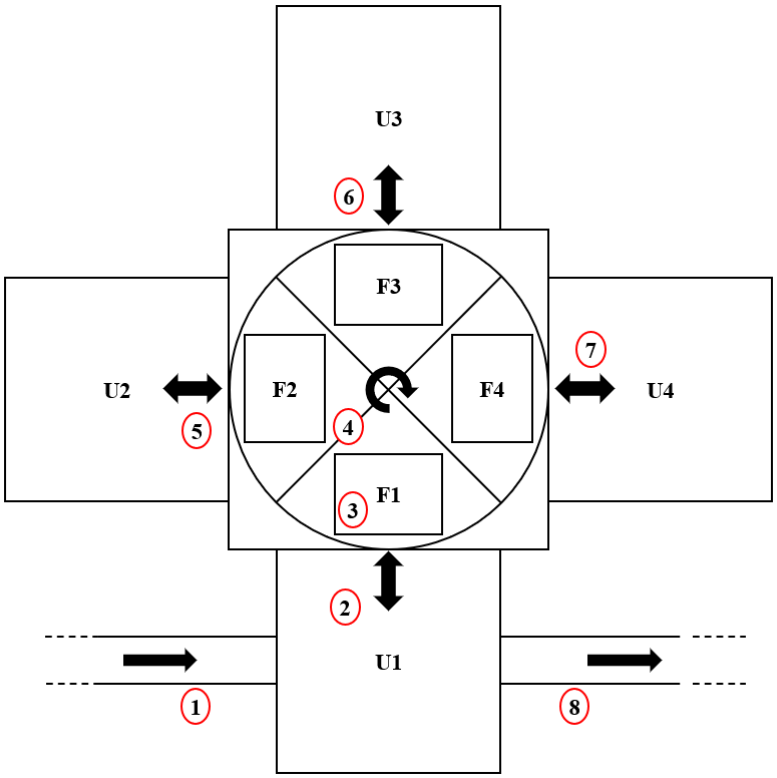

Figure 4.1: Overview of the machine system

2. The first unit is where the loading of products into a fixture in the machine takes place. The loading is performed automatically by a robot

3. There are four identical fixtures in the machine which go into the four different units at the same time

4. A round-table in the center of the machine rotates the cylinder heads placed in the fixtures so they can move between the units

5. In the second unit, the exhaust holes are rough machined by milling tools

6. In the third unit, the exhaust holes are finely machined by reaming and specialized cutting tools

7. In the fourth and last unit, measurements are taken by an automatic measurement system where touch probes are used to measure diameters, positions, and ovality of the holes. These measurements are used to verify if the product quality is approved, which is the case if the measurements fall within the specified tolerance limits

8. After the product has been measured, it is unloaded from the first unit onto the conveyor again, where it continues to the next machine

Conventional methods have been used to try and find the causes of the instability of the process and the quality errors. It has been difficult however to gain an understanding of the relation between the many process parameters and the quality outcome. Therefore, this case study was designed to gain a better understanding of the machining process through data analytics. The remainder of this chapter presents the results obtained from this study.

4.2 Results

The phases have been performed iteratively according to the flowchart which was shown in Figure 3.1. The business goals have been adjusted to be feasible with the constraints that were detected along with the execution of the project. Among the complexity levels of data analytics discussed in Section 2.1.1, it was decided that the case aimed to reach at least the descriptive stage, where the state of the process stability is visualized. Ideally, the case would also reach the diagnostic stage where potential root causes of the problem can be made visible.

4.2.1 Business understanding

A group was formed before the start of the project with a mix of experts to get a broad understanding of the business goals, the constraints, and how to formulate the data analytics goals. This group consisted of:

- One production engineer and project manager
- Two production engineers with data analytics knowledge
- One operator with extensive knowledge of the machine
- Two experts within maintenance and IT

This gave a wide knowledge base which is important as discussed in Section 3.2 to gain an understanding of how data analytics can be used to address the business problem effectively.

Business goals

The main business goal was to improve OEE, which is achieved by improving the quality and availability of the case machine. The quality is improved by increasing the process stability. This also increases availability due to fewer minor stops and less time required for manual process stability control. Since this is a pilot project, a sub goal was also set to better understand how data analytics can be used in the machining department and how the implementation process can be improved.

The criteria that were set to fulfill these goals were to use data analytics to establish solutions for:

- Statistical process monitoring - The operator should be able to keep track of the process stability in real time and know when to intervene
- Root cause analysis - It should be clear which machine parameters have the highest impact on the dimensional variations

The fulfillment of these criteria would aid machine operators to control the process stability. At the same time, it would create a better understanding of the process which can aid production engineers to perform long-term process improvements.

Inventory of resources

An evaluation of available resources and what would be needed was performed to ensure that it would be feasible to perform the project. It investigated what experts should be involved, what data sources are available, and what software and hardware would be needed.

Domain experts

Apart from the experts involved from the start, it would also be necessary to involve other experts in the project. It was mainly identified that more experts would be needed to be able to realize the data acquisition. As the automatic measurement system comes from another vendor than the machine control system, an expert for this system was needed so a separate solution for collecting measurement data could be developed. It was also identified that collecting machine parameters from the machine's numerical control (NC) system could be challenging. Therefore, an expert from the control system vendor had to be involved as well.

Data

An exploration of available data sources was performed to evaluate if the necessary data to fulfill the specified goals were accessible. The primary identified sources to use were process data extracted from the machine PLC and NC system, measurement data from a separate database, and operational data collected from a production monitoring software. With the previously identified experts involved in the project, it was determined that the required authority and knowledge to extract data from these sources were available.

Software and hardware

Certain software is needed to enable connectivity for the machine, store and manage the collected data, and perform the analytics tasks. For this, the software already in place at Volvo which has been used in previous projects was utilized. A brief description of the software is presented here and a description of how it was used for data acquisition is given in Section 4.2.2:

- Kepware - A software that enables connectivity for sensory systems from different machine vendors

- ThingWorx - A data analytics platform used to gather and store selected data in a suitable structure. It can also be used for performing data analytics and deploying models

- Python - The programming language used in this project to enable flexible data analytics with libraries such as scikit-learn

Additional hardware needed to perform the project was a camera to enable traceability. This camera was installed by the machine so it could scan a bar code engraved on the cylinder heads and assign each product a unique product ID.

Requirements and constraints

Apart from the needed resources to carry out the project, it was further evaluated what preconditions are required to perform the project and which constraints were present. The goals had to be adapted after these constraints to ensure that it would be realistic to fulfill these goals under the current circumstances.

Connectivity and data acquisition

The machine investigated in this case is of a relatively old model compared to other machines in the department. It was estimated that there would be fewer machine parameters available to connect compared to more modern machines that had been investigated in previous projects. Furthermore, it had been found in an earlier project related to data collection from the machine that the amount of data that can be extracted simultaneously from the machine is limited. If large volumes of data are gathered from the NC system at the same time, it is a risk that the control system will fail due to overload. Therefore, it was decided that only the most necessary data should be gathered initially, and the collection frequency should be kept low.

Traceability

It is required that traceability is established for the whole machine process. As the machine is divided into four units, the process parameters collected from each unit must be possible to trace to a unique product ID. As described earlier, a camera was set up which scanned the cylinder heads going into the units to make this possible. However, as there were no traceability options available for previous processes at the manufacturing line, the system borders for this case study were restricted to only cover this machine process.

System integration

As the initially identified data sources are located in different systems (such as the machine and measurement system), it is also required to integrate the data from all these systems. The traceability requirement therefore also applies to all systems, as

otherwise, it would not be possible to trace the data from these systems to specific product IDs.

Internal knowledge
As described in Section 3.2, analytic profiles containing information about successful approaches from previous cases can be of good help when deciding what data to collect and choosing an analytics strategy. As no previous projects have been performed for a similar machine or a similar problem at the company, there were limitations to how much insights could be reused from previous projects. The chosen data and analytics approach used in this project was therefore mainly based on estimations from domain experts as well as insights from the literature.

Time
The project had to be finished within four months. During this time, the relevant data had to be identified and a plan for collecting the data created. Thereafter, the historical data set could be collected. This put a limitation on how long it would be possible to collect data for the historical data set, as there also had to be time for performing the analysis.

Project prioritization

After an evaluation has been performed of requirements and constraints, it would be appropriate to compare the case to other potential cases to prioritize the case with the highest value in comparison to cost as discussed in Section 3.2. As this is a pilot project and there were no other use cases in the pipeline ready for initialization, no such prioritization took place for this study. It was deemed however that this case would be able to generate the highest value if a successful solution could be found, as the quality problems associated with this machine generate higher losses than any other machine in the department.

Data analytics goals

The goals of the data analytics process are to use historical data to create predictive models used for visualization and diagnostics of real-time data. For statistical process monitoring, critical parameters should be shown along with derived control intervals. For the root cause analysis, it should be evident which parameters have the highest impact on the quality outcome and how different parameters relate to each other. This information should be possible to use by a domain expert as support to find root causes.

The following criteria were specified for the data analytics goals:

- A relative ranking should be derived from the predictive model to show which parameters have the highest influence on the quality outcome
- Of the parameters with high rankings, the critical interval should also be identified for usage in control charts

- A conceptual control chart should be created which shows targets values for critical parameters as well as upper and lower control limits

- The predictive model used for retrieving this information is evaluated with a variety of metrics and should perform at an adequate level. This is verified by the project manager

4.2.2 Data acquisition

The same experts involved in the business understanding phase were also involved in the data acquisition phase, to achieve a broad overview of relevant data. A combination of brainstorming among the domain experts, inspiration from literature, and previous data analytics projects conducted at Volvo laid the foundation for what data to collect. The data acquisition followed the process described in Section 3.2 and started with workshops to create the initial data set to use for analysis.

Workshop for defining relevant data

The workshops were performed in a structured way and initially relied on brainstorming. The chosen method to perform the brainstorming and to structure the data was an Ishikawa diagram, due to its ease of use and proven effectiveness to detect cause and effect relationships in manufacturing. Ji and Wang [75, 76] investigated methodologies for data analytics for machining processes. They argue that common categories of relevant data in a machining context are data related to the workpiece, machine tool, machining process, machining result, machining time, and human factors. These categories were used in the Ishikawa diagram to group the data. It was considered that a combination of quasi-static data and dynamic data should be collected to provide good possibilities to find casual data for dimensional variations, as was discussed in Section 2.2.

During the workshops, it was decided that for the workpiece data, only the product ID is relevant in this case. Only one type of product is produced on the line and therefore there is no other variation in product properties. Due to the constrained traceability, it was also not possible to evaluate measurement data from previous processes on the line. For the machine tool, tool wear was estimated to have the largest impact on the dimensional variation. The most relevant data possible to collect from the machine in this regard was the number of cycles that have been produced by each tool.

The machining process-related data was the category where most data was deemed relevant. Examples of quasi-static data were temperatures and fixture numbers. Examples of dynamic data were spindle speeds, feedrates, tool positions, torque, forces, and vibrations.

The machining result category constituted the measurement data that was possible to collect from the measurement system in the fourth machine unit. This data includes the diameters, positions, offsets, and ovality of the exhaust holes.

The relevant machining time parameters were the cycle time and the idle time between cycles. Other time-related data was also determined to be relevant, such as work shift, weekday, and hour of the day. This data might indicate if human factors or other time varying factors could influence the quality.

Selection of data

From the ideal data set created from the workshops, a subset of required data was also formulated with the most important data. As there is a constraint on the velocity of data extraction from the machine, it was decided to be more restrictive in the selection of data. Furthermore, the collection frequency of the dynamic data was restricted to 1 Hz to not overload the control system.

Some machining process-related data were not possible to collect without purchasing extended access from the control system vendor. By a cost versus benefit evaluation, this was not done as the cost and time to access this data would exceed the potential benefits. The effect of this was that for example force and torque data were not possible to collect. However, it was possible to collect load data for spindles and axes, which was estimated to have a high correlation to this data.

Vibration data was also rejected as this information was stored in a separate system where it would be difficult to enable traceability. For the same reason, it was also decided to not use the operational data from the manufacturing tracking system.

Data collection plan

The collection of the final data set was realized in two separate ways as the data resides from two sources. Most of the data could be extracted directly from the machine. Kepware was used as a bridge to enable communication between the sensory systems from different vendors and the analytics platform, ThingWorx. OPC-UA and MQTT were used in conjunction to establish connectivity between the machine and Kepware. In ThingWorx, the data was stored in separate databases structured by the machine units. In these databases, relational data sets were created, meaning that the different process parameters were stored in separate data sets with a key column included in each data set which made it possible to integrate rows from separate data sets later on. The key column consisted of the product-ID, so data from each data set would be possible to trace to specific products. A user interface was also developed in ThingWorx to enable easy extraction from the analytic platform to CSV files which then could be analyzed in Python. As the measurement data resides in a separate system that could not be connected to Kepware, this data was instead stored locally at a computer by the machine and gathered manually.

After this plan had been developed and tested to ensure that both machine and measurement data were collected correctly, a historical data set was collected which was later used in the analytics phases. It was however not possible to collect the measurement data during the same time period as the machine parameter data due to a failure of the measurement machine. The plan was to use the measurement

data as the outcome for the diameter measurements. As this information was not available, data from a log filled by the operators was used instead to be able to see which products received an OK or not OK (NOK) result for the diameter dimension for any of the machined holes. The measurement data was still analyzed, but from an earlier time period before the collection of the other parameters had been realized. This was done to get better insights into how the variations behave and so these insights can be used for the machine in a follow-up project.

4.2.3 Data understanding

After a historical data set had been built up for a couple of weeks, the initial assessment of the characteristics of the data was performed. As was described in the previous section, not all of the data that was considered relevant for the case was collected due to the high effort of obtaining certain data. Some of the data that was considered was also not considered in the historical data set due to obvious quality issues which made it unsuitable to use for analytics. An example of this was parameters for federates of axis interpolations, as this data was gathered with highly inconsistent intervals which would make it unusable for finding consistent patterns between samples.

The data that was included in the historical data set and further evaluated with analytical methods are presented in Table 4.1. The data is heterogeneous as it has been collected with different frequencies and in different formats, varying between string, Boolean or numeric. Most of the data however are numeric, so there was not much effort required to transform the non-numeric data. All the dynamic data (eg. loads, positions, and speeds) has been collected with the same frequency of 1 Hz which also makes it possible to compare values over the same time frame without re-sampling the data. The exception was the temperature data, which could only be collected with one measurement every 30 seconds.

The CSV files of the data were generated and exported from ThingWorx. In total, 102 files were exported with a combined size of 1.02 GB. A total of 5 739 samples were stored in this data, where each sample resembles the data gathered during one machine cycle and can be associated with a specific product ID.

Explorative data analysis

During the EDA, the previously described data were examined to gain insights that could be used for modeling and data preparation. Since it was possible to collect the quality outcome regarding diameter dimensions (quality_result in Table 4.1), it was decided to use this data as target values and use a supervised learning approach to create the models that can fulfill this purpose. There was a total of 0.35% of defective products in the data set, which indicates a significant balancing issue. It was therefore deemed necessary to use a balancing technique for the data preparation phase.

Parameter	Description	Interval	Unit	Type
trace-ID	ID used to identify products	Cycle*	-	String
tool_wear	Number of units produced by a tool. Three tools in Unit 2, six tools in Unit 3	Cycle	Quantity	Numeric
fixture	Fixture ID	Cycle	-	Boolean
sp_temp	Temperature of spindle motors. Two in Unit 2, six in Unit 3	30 sec	$°C$	Numeric
ax_temp	Temperature of x-, y-, and z-axis motors. Unit 2 and Unit 3	30 sec	$°C$	Numeric
sp_load	Load of spindle motor. Two in Unit 2, six in Unit 3	1 sec	%	Numeric
ax_load	Load of x-, y-, and z-axis motors. Unit 2 and Unit 3	1 sec	%	Numeric
sp_act_speed	Actual speed of spindle motor. Two in Unit 2, six in Unit 3	1 sec	RPM	Numeric
sp_com_speed	Commanded speed of spindle motor. Two in Unit 2, six in Unit 3	1 sec	RPM	Numeric
ax_enc_pos	Encoded position of x-, y-, and z-axis. Unit 2 and Unit 3	1 sec	mm	Numeric
ax_lag	(Encoded - desired) position of x-, y-, and z-axis. Unit 2 and Unit 3	1 sec	mm	Numeric
hole_diameter	Difference from the nominal diameter of the machined holes. Inner and outer diameter for six intake and six outtake holes	Cycle	μm	Numeric
hole_x-pos	X-position for all holes	Cycle	mm	Numeric
hole_y-pos	Y-position for all holes	Cycle	mm	Numeric
quality_result	Logged quality outcome, OK or NOK	Cycle	-	String

Table 4.1: Description of collected machine and measurement data that were used for analytics. *Cycle means that the parameter was collected once per cycle

As explained in the *Data collection plan* section, it was not possible to obtain measurement data in the historical data set, but an earlier collected set was still analyzed to gain insight from this data. In Figure 4.2, it is shown how the diameter varies for a certain hole of the cylinder heads. It can be seen that large variations occur for certain products both in a positive and negative direction. For some cycles, the

variations have been so large that the diameter has been outside of the tolerance limit (illustrated with the red lines).

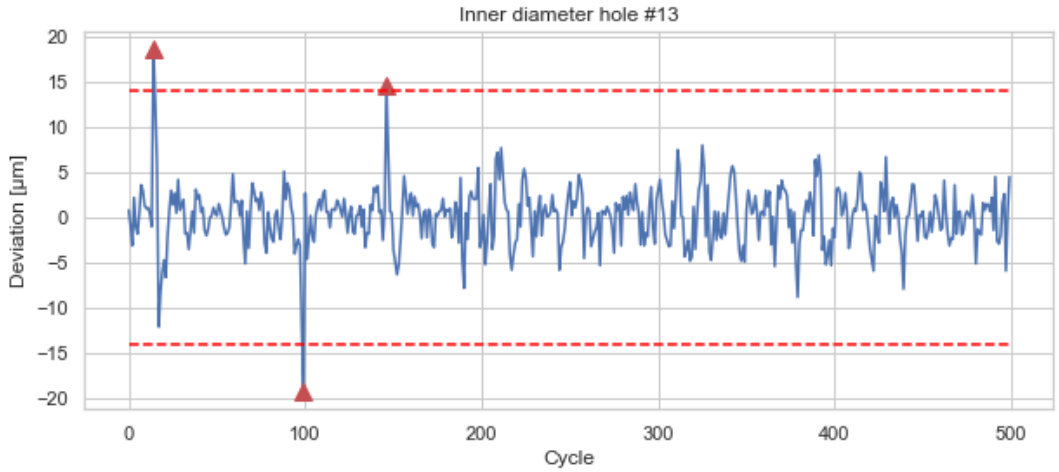

Figure 4.2: Deviation from the nominal value of the inner diameter for one of the holes. For three cycles the result was NOK over 500 cycles

Another discovery that was made from the diameter measurements was that all of the holes have a similar variation pattern, especially for the large variations. This can be seen in Figure 4.3 where the inner diameter of three different holes is shown. When on diameter deviates significantly from the nominal in either a positive or negative direction, the holes of the other diameters follow in the same direction. The same pattern was recognized when comparing all 24 diameters against each other, both for inner and outer diameters.

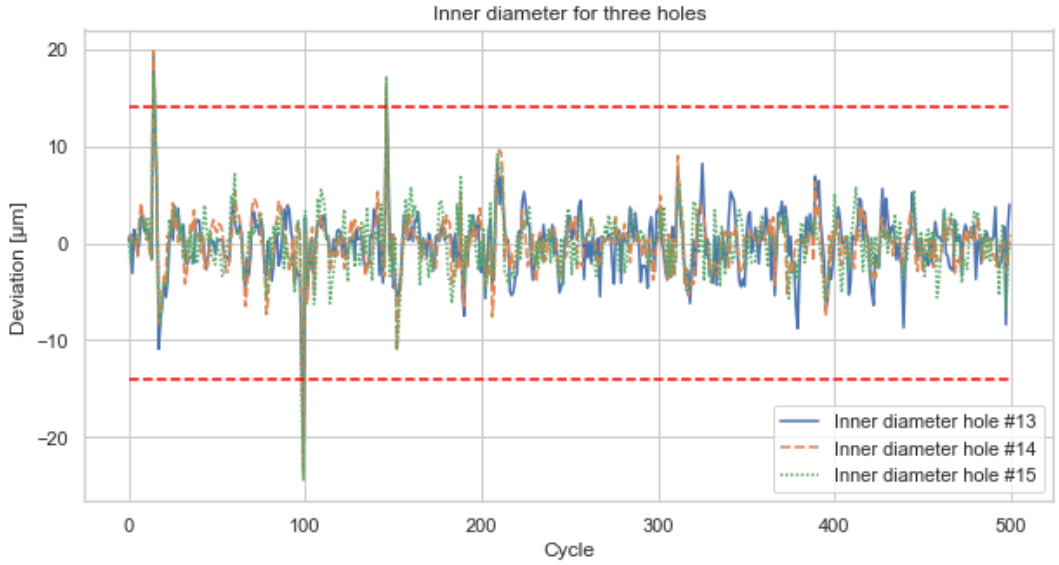

Figure 4.3: Deviation from the nominal value of the inner diameter for three different holes. It can be seen that the deviations follow the same pattern for larger variations

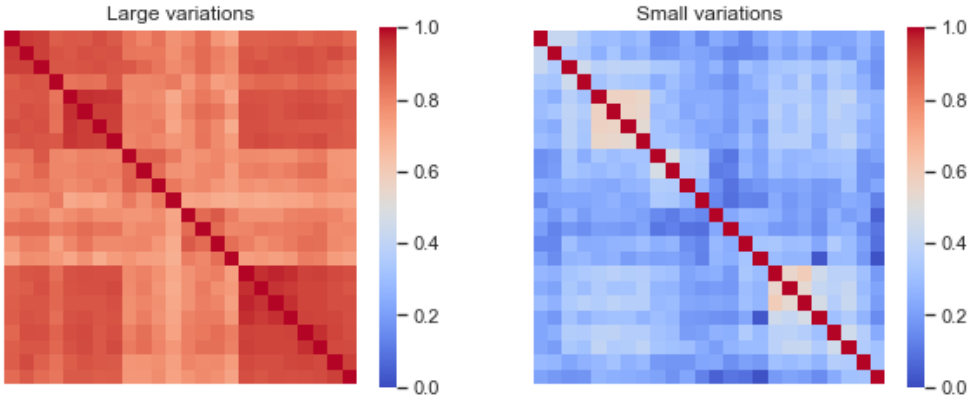

Figure 4.4: Correlation matrix for all hole diameters grouped by cycles with large variations (left) and small variations (right)

A custom threshold was set as $\pm$ 7 μm from the nominal value to group the data by larger and smaller variations. As seen in the correlation matrix in Figure 4.4, there was a high correlation for diameter measurements of the cycles where the result was outside of this threshold. There was a significantly lower correlation for the measurements within the limits of this threshold. This might indicate that there are common causes for the larger variations of the holes. Therefore, it was decided that a clustering technique can be useful to separate cycles where large variations have occurred from small variations. These clusters can then be used as classification labels.

No apparent quality issues could be observed in the measurement data. For the machine parameters, however, some problems that required countermeasures in the data preparation phase could be seen. The largest issue was missing data points, which could be divided into a cyclic level and inter cyclic level. Missing values on the cyclic level mean that no observation has been registered for one or several parameters for certain cycles. The proportion of these samples was relatively low, however, and it was decided that those samples with missing data for some parameters could be removed from the data set without losing too much information. Missing data on inter cyclic level meant that there were gaps in the collected sensor values. The gaps constituted one or several seconds where no measurements had been registered. Figure 4.5 illustrates the normal distribution of the number of measurements collected for load, position, and lag parameters for the x-, y-, and z-axis in machine unit 3. It is mainly due to the varying gaps of missing data in these parameters that make them deviate from the expected numbers of measurements that could be measured during a normal cycle time. To be able to compare the different parameters over the same time frame, it was decided to use a technique for replacing missing data points within the sensor measurements samples. This is further explained in the data preparation section.

Another potential quality problem was noise in the sensor measurements. This was mainly apparent for the load parameters, where an example is shown in Figure 4.6 for the z-axis motor load in unit 3. Large spikes occur at different places for

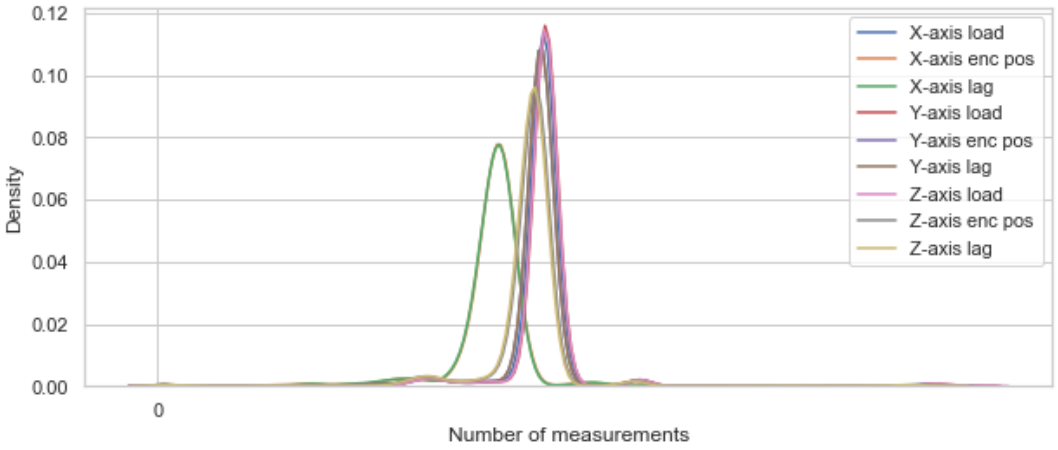

Figure 4.5: Distribution of amount of data measured during cycles for parameters related to the x-, y-, z-axis in unit 3

different cycles. Filtering or smoothing techniques could be applied to reduce the noise, but it was decided to avoid that preprocessing step initially as it was difficult to understand what spikes occurred due to erroneous measurements and what was natural behavior.

Validation of data set

From the insights gained in the EDA, it was decided that most of the collected and evaluated parameters had high enough quality to be used for modeling after certain data preparation techniques had been applied. It was however seen that the parameters for the spindle speeds in unit 3 had too many missing within the sensors measurement samples to be used for analysis. Those parameters were therefore

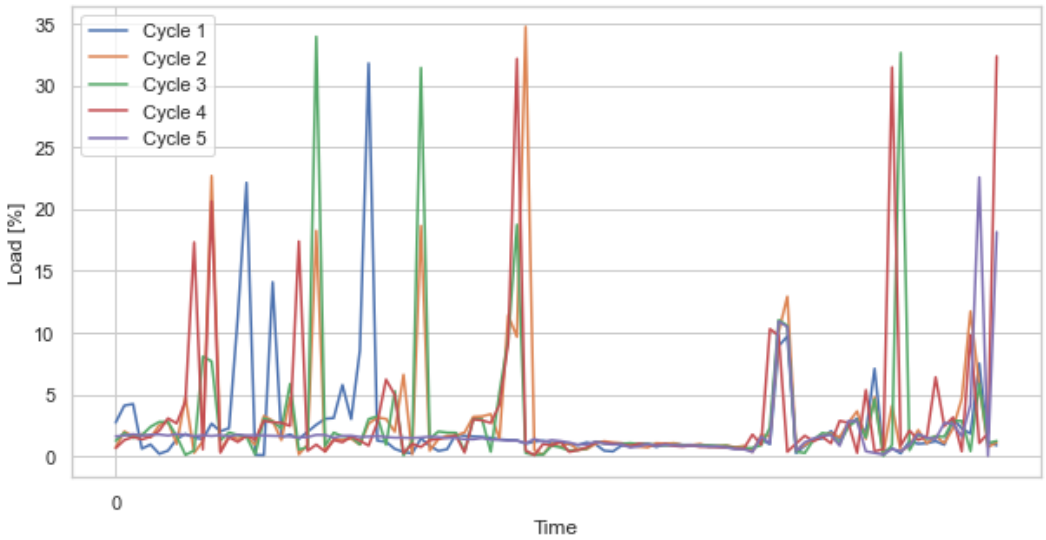

Figure 4.6: Load value for five different cycles for the z-axis motor in unit 3. Some large spikes occur at different times for the cycles

dropped from the data set, all the other parameters were kept. Overall, it was estimated that these remaining parameters would be appropriate to create models with a supervised learning approach.

The largest issue with the data set was the unbalance between the OK and NOK classes. More data could be collected to reduce the impact of lacking NOK samples. It was however decided to move further with the current data set due to the time constraint of the project and collect more data at a later stage if needed.

4.2.4 Data preparation

The preparation of the data into a format suited for modeling followed the process described in [71] which was discussed in Section 2.4.1. The process consists of four different steps: data cleaning, data integration, data transformation, and dimension reduction. This section describes what data preparation methods were performed during each of these steps.

Data cleaning

As was identified during the EDA, the data set contained missing values for certain parameters in some samples, and also some missing measurements within the time series measurements from sensors. The samples with missing parameters were removed from the data set. For the time series measurements, an upper and lower limit of collected measurements was set to remove the outliers with too few or too many measurements collected during a cycle. This lower and upper and limit was decided with the standard deviation method [77] as shown in Equation 4.1 and 4.2:

$$Lower = \mu - 2\sigma \tag{4.1}$$

$$Upper = \mu + 2\sigma \tag{4.2}$$

Here μ is the mean of the number of measurements collected during a cycle for a parameter that is the same as the expected cycle time, and σ is the standard deviation. By choosing to remove all the data two standard deviations from the mean, 5% of the samples were removed.

To deal with the missing values within time series samples, the technique applied in [15] was used. The missing values were thereby replaced by the mean value of the preceding and following values in the time series. After the removal of outlier samples and the replacement of missing values, the parameters had a more even distribution of measurements.

Data integration

Data integration was a straightforward process as all the data files already had been structured into the same format in the analytics platform. This data consisted of relational data sets, one for each parameter, which all contained the product trace number. The integration was thereby performed by merging all these files by trace number into a tabular format. Each row in the table resembles a sample collected for a certain product-ID and contained data collected from both machine units where the machining operations are performed. The target value indicating the quality outcome was also integrated into this table.

Data transformation

As most of the data were in numeric format, there was a low effort to convert the data. The fixture data which was in Boolean format however was hot encoded into 0 and 1, as many ML models cannot operate on categorical data and require all inputs to be numeric [78].

All the parameters were normalized to have a common scale between all parameters. Having different scales between parameters, where for example temperature ranges from 30 - 50 while tool wear ranges from 0 - 900, can cause worse performance for certain ML algorithms and make it difficult to derive feature importance [24]. Min-max normalization was therefore applied which is a commonly used technique for re-scaling data [79]. The data for all parameters were thereby transformed so the minimum value for a parameter is 0 while the maximum value is 1.

It is described in [79] that two important categories of features to consider for data analytics in industrial applications with time series data are timestamp features and statistical features. Timestamp features in this case were extracted from the timestamps which each measurement was associated with. From this, it was possible to derive the cycle time, idle time between cycles, weekday, working shift, and time of the day.

For the time series data (the parameters collected with 1 Hz frequency), a method suggested in [80] was used which is a general approach for extracting statistical features from sensors. This method is divided into four steps: Fragmentation of all the time series data, aggregation of features with statistical values, determining feature importance through ranking, and selection of features. Fragmentation means to split the data into equal time frames, which was done already during the collection of data in this case as only measurements during active cycles were kept for each product.

For the aggregation step, different sizes of time windows over which the data will be aggregated into statistical values can be used. These windows can be set with different sizes depending on what granularity is required for the modeling problem [81]. The method constitutes of a window with a given width that slides through time series data so that a given number of data points at the time are used for extraction of statistical features [79]. The windows can be of equal or varying size,

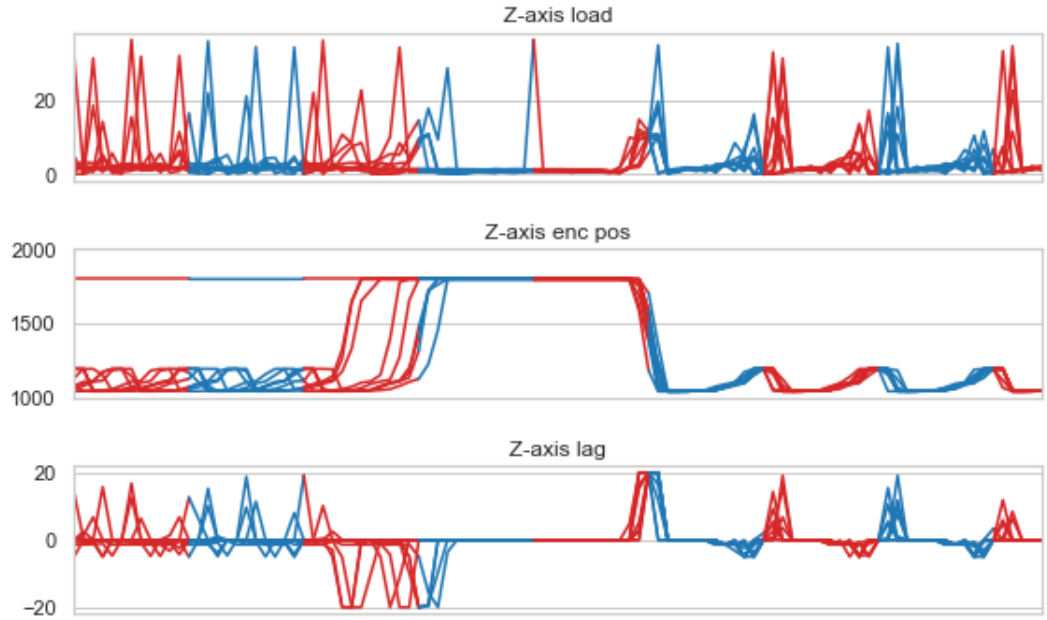

Figure 4.7: Visualization of time windows for parameters belonging to the Z-axis in unit 3. The colors show window divisions

as well as they can be overlapping or non-overlapping [81]. For this case study, fixed width of non-overlapping windows was used for simplicity and to ensure that the features become intuitive. In Figure 4.7 it is shown how different parameters were used for parameters associated with one of the axes of the machine. When deciding what statistical values to extract from the time windows, it is important to receive a balance between meaningful but fragile, as well as robust but less significant features according to Christ et al. [80]. They elaborate on this with the example that a median is a robust value that is not heavily influenced by outliers, while a max value is useful to detect high values but is at the same time more fragile to outliers. To achieve a good mixture of statistical values which describes value ranges as well as variation, the mean, median, maximum value, minimum value, standard variation, variance, and kurtosis were extracted over each window.

Initially, 11 windows of equal sizes were created for each parameter. The extraction of the above-mentioned statistical values resulted in 2 156 statistical features. When combined with the features extracted from the static data and the time stamps, a total of 2 198 features were obtained. As this is a high number of features and many ML models are sensitive to overfitting, the next step was to rank and select the most important features following the methodology described in [80].

Dimension reduction

To reduce the dimension of the predictors (the extracted features), the analysis of variance (ANOVA) method was used as this is a well-suited method for ranking features in data sets with numerical predictors and categorical response variables [82]. ANOVA tests the relationship between each predictor and the response vari-

ables by evaluating if there is equal variance between groups of the variables as described in [82]. They explain that if there is equal variance between such groups, it means that a feature will not have any impact on the response variable and is therefore assigned a low ranking. In this way, it is possible to select only the most valuable features which receive the highest rank for modeling. Different amounts of top-ranked features were used during modeling in this case study to find a balance between overfitting and underfitting of the models.

During the EDA, it was noticed that the diameter measurements varied with a high correlation for all holes for the larger variations. It is therefore reasonable to reduce the dimension of the 24 diameter measurements into a binary classification problem with larger and smaller variations. This follows the approach discussed in 2.3.1 to combine unsupervised learning to create clusters for the outcome data where the cluster is the label, and then use supervised classification methods to predict these labels [48]. To create these clusters, the anomaly detection algorithm Isolation Forest was used. This algorithm is described in [83]. They describe that an ensemble of decision trees is created first by training the algorithm with a data set of N features (in this case the 24 diameter measurements). Thereafter they describe, data points are traversed through the created isolation forests where it is clustered as an anomaly or not. The scatter plot in Figure 4.8 illustrates these clusters when two measurements are compared against each other. It shows that the algorithm has divided the measurements revolving close to the nominal diameter measurement has been divided into the normal cluster, while the measurements deviating significantly in either positive or negative direction have been clustered as anomalies. The same pattern could be identified when all the 24 measurements were compared against each other. The distribution of samples for each cluster is shown in Figure 4.8. The anomalies represent 5.8% of the total amount of samples, which indicates that balancing might still be needed for certain ML algorithms that are sensitive to unbalanced data sets. These clusters can be re-created and used as target values for further iterations when the measurement data can be included in the historical data set. This data was as earlier explained not used for the iteration presented in this report, but this approach for creating classification labels can be evaluated for future iterations.

Data validation by domain experts

As was described in Section 3.2, it is important to involve domain experts during the data preparation phase to ensure that the selected features still make a good representation of the original business problem. A validation was therefore performed before moving to the modeling phase. As interpretability is important in this case to understand relationships between machine parameters and the quality outcome, it must be possible to understand what the extracted features mean. It was concluded that the statistical representations of parameters over time windows might not be an intuitive way of presenting the result to the end user. However, it was decided that this approach can give initial insights into which parameters and during which time windows have the greatest impact on the quality. These parameters can then

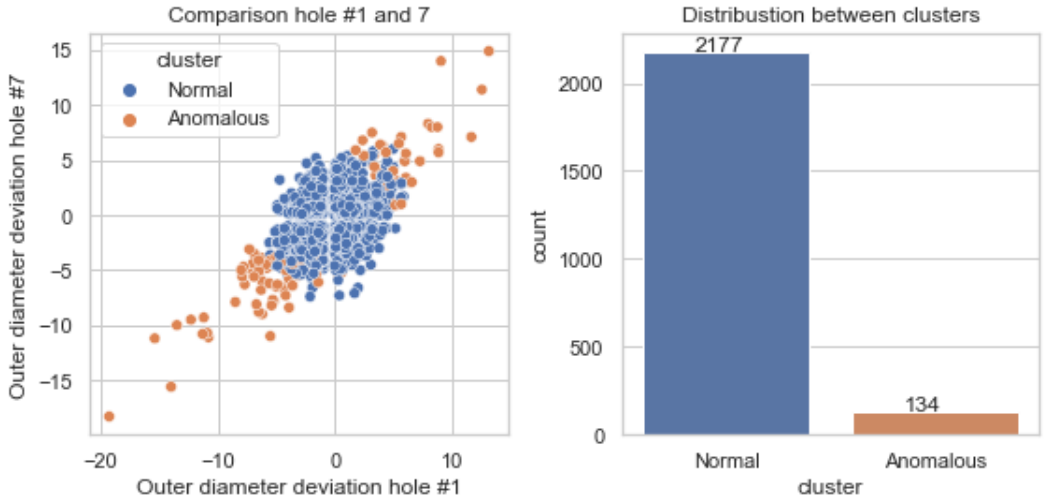

Figure 4.8: Visualization of diameter measurement deviation of two holes (left) and the distribution of samples is each cluster (right)

be further analyzed to create a more intuitive presentation of the information which can be used for process monitoring or finding root causes.

4.2.5 Modeling

The modeling phase was conducted according to the recommended steps in CRISP-DM. In this section, the selection of ML is presented along with the test design and evaluation of the models. It is also shown how association rules were created from the most relevant features derived from the models.

Model selection

As interpretability is important for this case study, this was used as one of the main criteria during the selection of models. As was described in Section 2.2.2, tree-based models have proven to be efficient for finding root causes in manufacturing problems due to their interpretable structure and ability to show importance from input parameters. In [84] some common examples of tree-based models are given, such as Decision Tree, Random Forest, and XGBoost. Another model they describe which is not tree-based but still advantageous for its easy structure and ability to solve binary classification problems is Logistic Regression. These four models were trained to find which algorithm has the best ability to find the patterns that can be used to predict which cycles receive a NOK quality result. The most significant features for predicting the outcome were thereafter identified from each model. As was also described in Section 2.2.2 association rules can be used to gain further understanding of how relevant features relate to each other to gain deeper insights into the underlying problems of quality defects. Similar to the case described in [15], the Apriori algorithm was used to derive rules for how the parameters influence the quality outcome. A brief description of each of the used models is given in the remainder of this section.

Decision Tree

A decision tree has a tree-like structure that starts with a root node at the top from which it splits into other nodes based on certain conditions, as explained by Gupta [85]. The nodes are connected by so-called branches, the author further explains, and at the end of the structure, a leaf node is reached where a decision is made for the classification problem. The advantage of this model is that the feature importance is clear, and it is easy to see how the model makes decisions [85]. The disadvantages the author brings up are that the model is sensitive to overfitting as this might create overly complex trees and that the tree might become biased if the data set is unbalanced. This motivated the use of ANOVA to reduce the dimension of the feature set and to use a balancing method before training the model.

Random Forest

The random forest model consists of an ensemble of decision trees where each tree makes its prediction, as described by the study proposed by Yiu [86]. The concept of random forests is the wisdom of crowds the author mentions, meaning that the model's prediction will be based on which class receives the most votes by the many decision trees. An advantage with random forests is that they are created from many subsets of data, and since the output is based on an averaged majority ranking, random forests are not sensitive to overfitting [87]. ANOVA was still used to remove redundant features for the random forest model, but a higher number of features were used initially compared to the decision tree model.

XGBoost

The XGBoost model is in similarity to random forest also an ensemble of decision trees but works differently [88]. XGBoost stands for "extreme gradient boosting", which means that new models are created to predict the residuals which previous models have made, and then these are added together to make a final prediction as explained in [88]. The main benefits with XGBoost they bring up are that it has been proven to be very accurate on tabular data sets and that the model is fast to train.

Logistic Regression

Unlike the previously described models, the logistic regression model is not tree-based. As described by IBM [89], logistic regression estimates the probability of an event occurring by applying a logit transformation to the odds, which is called log odds. The log odds are associated with each of the features, and they explain that the model is easy to interpret as these log odds can show what importance certain features have. This model is also sensitive to overfitting [89] which makes feature selection important before training.

Apriori

Apriori is an association rules analysis technique that aims to uncover how different features in a data set are related to each other. This algorithm is explained in [90] where it is mentioned that there are different ways to measure the association. One

of these is support, which they explain shows the fraction of samples that contain a value of a certain feature, as shown in Equation 4.3 for a feature I.

$$support(I) = \frac{Number\ of\ samples\ containing\ I}{Total\ number\ of\ samples} \tag{4.3}$$

Another important measure is confidence [90], which shows the likelihood that a set of features Y occurs in the same sample as a set of features X occurs. This can be expressed as a rule, $X \longrightarrow Y$, meaning that if X occurs then also Y will occur, and the confidence for this rule is calculated according to Equation 4.4:

$$confidence(X \longrightarrow Y) = \frac{Number\ of\ samples\ containing\ X\ and\ Y}{Number\ of\ samples\ containing\ X} \tag{4.4}$$

As the goal in this case study is to find how machine parameters affect the quality, the same approach as in [15] was used where X was set to the machine parameters and Y to the quality outcome. A shortcoming of Apriori is that the set of candidate features to derive rules from can become extremely large [90]. Therefore, only the features with which were ranked as the most significant were considered when applying the algorithm. Apriori also only takes categorical features, so the numerical features had to be divided into categorical ranges instead. Five uniform ranges were created for each numeric feature X according to Table 4.2.

Category	Range
Very high	$80\% \leq X$
High	$60\% \leq X \leq 80\%$
Medium	$40\% \leq X \leq 60\%$
Low	$20\% \leq X \leq 40\%$
Very low	$X \leq 20\%$

Table 4.2: Applied categorization of continuous data for the Apriori algorithm

Generation of test design

A train-test split approach was used to divide the data set into training data used for training the models, and test data used for validating the models. It was divided so 75% consisted of training data and 25% of test data. Due to the imbalance between the classes, it was ensured that the same distributions between the classes were held in both the training and test data. The data set only contained a total of 15 samples where a NOK quality outcome had been registered. This meant that among these 15 samples with a NOK label, 11 were used for training, and 4 were used for validation.

To improve the balance in the training data set, SMOTE was applied which was discussed in Section 2.3.4 as a common approach to improve balance in classification

problems. The SMOTE algorithm creates new synthetic samples with the same label as for the minority class with feature values like those of the existing samples [91]. To avoid creating too many synthetic samples and introducing bias in the training data, it is usually appropriate to combine SMOTE with an undersampling technique to achieve class balance [91]. Therefore, the training set was balanced by first increasing the ratio of the minority class with SMOTE, and then reducing the ratio of the majority class by random undersampling to achieve the desired ratio.

When training the models, different configurations were tried to find the best settings for each model. The adjusted parameters were the number of aggregation windows for the time series data (11, 8, or 5), the ratio between the minority and majority class, and the number of highest ranked features from the ANOVA result.

To evaluate ML models, accuracy is a common metric to use which shows the proportion of correct predictions the algorithm makes. As the validation data set is extremely unbalanced in this case, it would be unappropriated however to rely on accuracy as the models can achieve high accuracy by only predicting that the outcome will be OK. To give a more nuanced view of the model performance, a confusion matrix can be created which visualizes the proportion of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) among the predictions [92]. From these values, other metrics can be calculated such as precision, which measures the model's exactness; recall, which measures the model's completeness; and the F1-score, which is a weighted average of precision and recall [92]. The metrics are calculated according to Equation 4.5, 4.6, and 4.7:

$$Precision = \frac{TP}{TP + FP} \tag{4.5}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.6}$$

$$F1\text{-}score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.7}$$

These metrics were used to better determine the models' abilities to differentiate OK cycles from NOK cycles. Thereby it was easier to evaluate if the feature ranking from a certain model would be robust by investigating the balance among the different metrics.

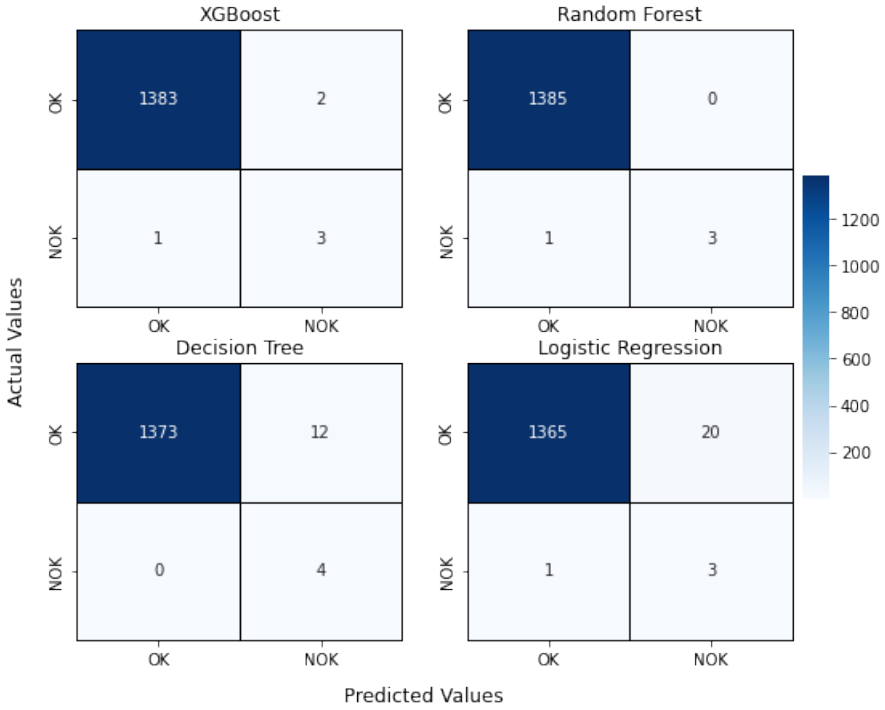

Figure 4.9: Confusion matrices for the four classification models

Assessment of models

After training the models with the previously described configurations, the configuration that achieved the best result was chosen for each algorithm. The result of the models was sorted so the configuration which generated the highest recall during testing was selected. This was done to ensure that the model was able to identify as many of the NOK cycles as possible, which indicates that the model has not become biased to only predicting OK outcomes. Precision and F1-score were evaluated secondly to evaluate how exact each model was with its predictions. The confusion matrices for each of the four classification models can be seen in Figure 4.9. All of the models were able to classify all or most of the four NOK samples in the validation data set. The decision tree and logistic regression model had a slightly higher extent of FP classifications than the random forest and XGBoost model.

The used configurations and the calculated metrics can be observed in Table 4.3. It shows how many aggregation windows were used, what ratio was used between the classes in the training set, and how many of the top features from the ANOVA ranking was used for each model. For most of the models, five aggregation windows for the time series parameters seemed to generate more accurate models. A balancing ratio of around 1:5 among the NOK and OK classes was the most successful ratio. It could also be seen that all the models performed better with a significant feature reduction to around 50 of the top ANOVA features. Even the random forest and XGBoost models which are less sensitive to overfitting benefited from this feature selection. The result shows that the decision tree model predicted all the NOK labels correctly, while the other three models made three out of four correct predictions of

Algorithm	No. Win.	Ratio	Features	Accuracy	Precision	Recall	F1
Random Forest	5	3:10	50	0.99	1.00	0.75	0.86
Decison Tree	5	1:5	50	0.99	0.25	1.00	0.40
Logistic Regression	5	1:5	40	0.98	0.13	0.75	0.22
XGBoost	8	1:5	50	0.99	0.60	0.75	0.67

Table 4.3: The best configuration used for training of each classification model together with the obtained evaluation metrics

Features	Feature ranking			
	Random Forest	Decison Tree	Logistic Regression	XGBoost
T1303_SP2	1	2	2	9
Idle_time	2	5	1	3
SP34_temp	3	-	-	4
SP31_temp	-	1	-	-
T1306_1465	6	3	-	-
T1303_427	-	4	-	1
T1304_SP4	7	-	3	-
Fixture_3	-	-	-	2

Table 4.4: Feature ranking for each model

the NOK class. Due to the higher number of FP classifications for the decision tree and logistic regression models, however, the random forest and XGBoost models received a higher precision and F1 score.

As all models performed sufficiently well according to the evaluation metrics, the feature importance was extracted for each of the models. The top three ranked features were selected for each of the models and are presented in Table 4.4. The tool wear parameters had a high presence among the top ranked features for each of the models (the parameters with prefix T130X), and the temperatures of spindles. The idle time and the third fixture also showed to be important features for classifying the outcome. These rankings appear to be reasonable given the historical data set. It could be seen for example that one of the tool wear parameters, T1303_SP2, which was ranked high among all models indeed had very high tool wear for most of the cycles with a NOK quality outcome (over 95% for most cycles).

Rule no.	Parmaters		Result	Support	Confidence
1	T1303_SP2_VH, T1304_SP4_H	$\Longrightarrow$	NOK	22%	100%
2	Idle_time_VH	$\Longrightarrow$	NOK	16%	100%
3	T1306_427_L, T1303_SP2_VH, Idle_time_VL	$\Longrightarrow$	NOK	16%	100%
4	T1304_SP4_H, SP34_temp_H	$\Longrightarrow$	NOK	16%	78%
5	Fixture_3	$\Longrightarrow$	NOK	16%	44%

Table 4.5: Association rules obtained with the Apriori algorithm where five of the 78 rules have been highlighted

Association rule mining

The eight features presented in Table 4.4 were further evaluated with the Apriori algorithm to derive association rules. The goal was to find rules which show which parameter ranges have a high risk of causing quality defects. The original data set was downsampled with random undersampling to create a 1:3 ratio of samples with NOK and OK labels. This was done to ensure that the support of the NOK samples would not become too small, as this would cause a very large data set of rules that in this case would be redundant. With these settings, a total of 78 rules were created by the algorithm which shows how a parameter or combinations of parameters relate to the quality outcome. Some of the rules can be seen in Table 4.5. Several of the rules are very similar to each other, so the table mainly presents the rules with high confidence and support, as well as rules that seemed to reveal interesting insights. The full list of extracted association rules can be seen in Appendix A. The suffix of the parameters in the table indicates the numeric range: Very high (VH), high (H), medium (M), low (L), or very low (VL).

The rules can be interpreted so that the parameter combinations on the left in the table lead to the outcome on the right. The first rule for example can be read "if T1303_SP2 is very high and T1304_SP4 is high, then the quality outcome will be NOK". The support shows how many times a rule appeared in the data set and the confidence reveals how many times the rule appeared in relation to have many times the parameter appeared. For the 5th rule, for example, the support was 16% as this rule was true for 7 out of the 45 samples in the data set (7/45 = 16%). The confidence of the 5th rule was 44% as fixture 3 had been used for clamping the workpiece in 16 of the samples (7/16 = 44%). It is important to evaluate both the

support and confidence of a rule to determine if the rule can reveal usable insights [90]. A certain threshold can be decided for what support and confidence must be met. It is also important however to also use domain expertise to decide if a rule only reveals trivial knowledge, or if the rule has been created due to flaws in the data set.

4.2.6 Evaluation

In this section, it is discussed if the results are shown in the previous section fulfill the specified criteria from the business understanding phase. The next steps for continuing with the case are also presented.

Evaluation of results

The first business criteria for this case study is to create a statistical process monitoring solution to keep track of the process stability. The second criterion is to find which parameters have the highest impact on the product quality so this can be used as decision support for root cause analysis. The methods used during the described data analytics phases showed promising results to be able to fulfill these criteria. With the used ML models, it was possible to rank features extracted from the machine parameters (the first data analytics criteria). These rankings seemed reasonable considering the collected data set. It could be seen for example that a tool wear parameter, T1303_SP2, which received a high rank among all models indeed was in a high range for all of the NOK samples. The idle time also showed as important, which is reasonable as 7 out of the 15 NOK cycles were idle for over 1 000 seconds before the cycle started.

None of the time series features were represented among the top three features from the models, but still showed some importance for the predictions. For example, the decision tree ranked the variance in the third time window for the load of SP3 in unit three as the 6th most important feature. A visualization of this parameter can be seen in Figure 4.10 where the OK cycles are colored blue and the NOK cycles are colored red. It can be seen that large variations occur for many of the NOK cycles in the third time window, and it is, therefore, reasonable that the variance could be used to distinguish these cycles.

By the help of association rules, it was possible to gain further insights into how specific ranges of the parameters influence the product quality (the second data analytics criteria). These rules can be helpful to create a solution for process monitoring that keeps track of when combinations of features enter critical ranges.

The used methods show a promising capability to reveal new important insights into the quality problem. The data set used to derive important features and rules was however highly unbalanced and only contained a small number of NOK cycles. Several of the NOK samples were furthermore collected close in time to each other. This might have caused bias in the result as these samples for example had similar tool wear which thereby showed in the result, even if the tool wear might not have

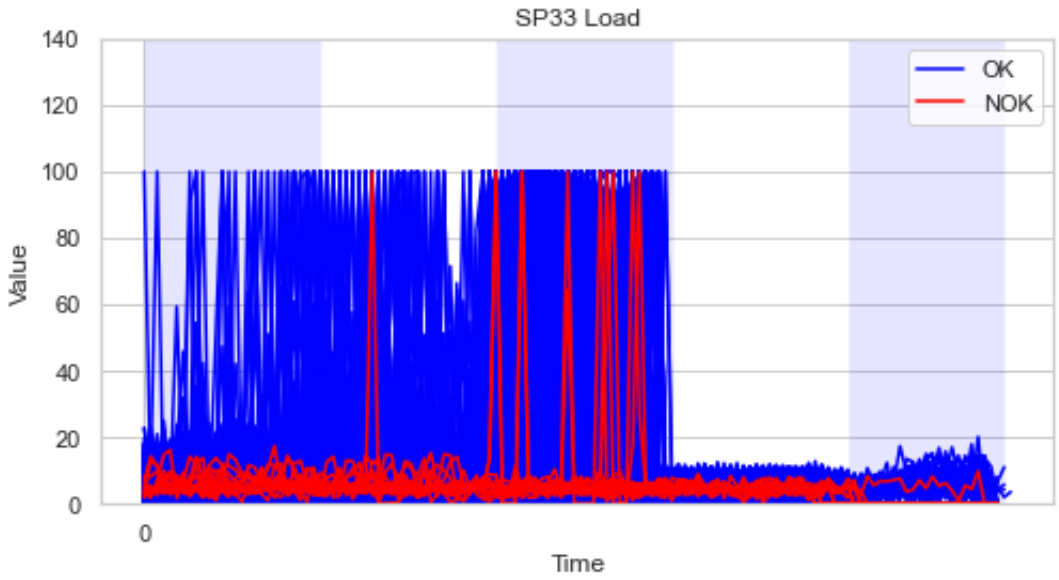

Figure 4.10: Load of spindle 3 in unit 3. Large variations for the NOK cycles have occurred in the third time window

been the real cause for the problems. Possibly, different results would have been obtained if more samples of NOK cycles had been collected. Therefore, this use case should not be concluded as finished at this stage, as further steps are required before a proper validation of the results can be performed.

Next steps

More iterations of the adapted CRISP-DM process should be performed until a satisfactory and robust result has been obtained before moving to the deployment phase. The first step should be to collect a larger set of historical data, as this would give a more robust analytics result. When more data has been collected, the same preprocessing and modeling techniques presented during this first iteration should be applied once again to evaluate how the results are impacted. Thereafter a conceptual interface for the process monitoring solution can be created.

Different preprocessing and modeling techniques could also be experimented with to determine if this will yield better results. Instead of using manually logged quality outcomes to label the samples, it should be evaluated if labeling by the clusters of the measurement data presented in the EDA can improve the models' performances. Using this approach could make the classes more distinct from each other, as many of the OK cycles had unusual large variations as well.

5

Discussion

In this thesis, a review of state-of-the-art literature and a conducted case study have resulted in further insights into how data analytics can be used in manufacturing. These insights can be expressed as theoretical and practical contributions. The theoretical contributions aim to help fill the gaps in research for the investigated topics, while the practical contributions are mainly useful for Volvo and its implementation of data analytics. In this Chapter, these theoretical and practical contributions are described.

5.1 Theoretical contributions

The theoretical contribution from this thesis was firstly the evaluation of how CRISP-DM can be adapted to better suit the implementation of data analytics in manufacturing. Secondly, it was to evaluate how data analytics can be used to improve quality in cases where high model interpretability is important and to demonstrate this approach in a case study. In this section, both these types of contributions are described. A critical reflection on the used methods is also provided which can be considered for making improvements in future research.

5.1.1 Adapted CRISP-DM methodology

There exist several methodologies to carry out data analytics projects. The second research question of this thesis was to find *what can be a suitable methodology for performing data analytics projects in manufacturing?*. It could be seen in the literature that one of the most popular standard methodologies to use in manufacturing environments is CRISP-DM. This methodology is well suited for manufacturing problems as it puts high emphasis on connecting business goals with data analytics goals. This can make the analytics results more aligned with was is needed to solve complex manufacturing problems. CRISP-DM is however a general process, and it has been shown in several studies that the methodology lacks a level of detail to show how it can be applied efficiently in manufacturing contexts [67, 5, 73, 63, 65]. In this thesis, several suggested improvements for facilitating the use of CRISP-DM in manufacturing were identified in the literature review. These improvements were thereafter combined to create a holistic and easy-to-follow version of CRISP-DM adapted for manufacturing. This methodology was evaluated in a case study at

Volvo for improving quality in a machining operation. The machining department at Volvo is in a beginner phase for applying data analytics and needs a clear methodology that describes how a data analytics project should be performed. The aim is that this adapted methodology can be applied by Volvo and other manufacturing companies in a similar position to fill the knowledge gaps on how to apply data analytics efficiently for different kinds of problems. Several insights could be gained regarding the effectiveness of methodology through the case study.

CRISP-DM provided a clear step-by-step approach which made it easier for all involved experts to understand the different phases of a data analytics project. The added elements to the methodology, such as the data acquisition phase, provided more clear guidelines for how to reason when deciding which data was necessary to collect for the problem. It could be seen however that the effectiveness of conducting a data analytics project in manufacturing is highly impacted by the team's knowledge and experience on the topic. Several experts within different areas were involved in the use case, but there was a general low knowledge of data analytics among the team. This made it difficult to formulate feasible goals during the business understanding phase and to create concrete goals for data analytics. Thereby it was necessary to revisit the goals and adjust them several times along the way of the project. The lack of practical experience in data analytics projects also had a significant impact. The initial assessment of the required effort to obtain data from the identified sources was highly misjudged. This caused delays in the project and made it necessary to postpone the planned deadline.

One takeaway from the project methodology is the importance of increasing the theoretical knowledge of data analytics concepts among all team members. Forming efficient teams with the required multidisciplinary knowledge is described as a key aspect of succeeding with data analytics in manufacturing in [11]. This will make it easier to set goals and to communicate in a team with multidisciplinary knowledge. It is also important to perform more pilot projects to build experience as this will improve the effectiveness of the projects over time and make it easier to do accurate assessments. Analytic profiles should be created after each project has been finished, which can also help to reuse insights and reduce the time required to perform data analytics projects.

5.1.2 Interpretable models for improving quality

The case study in this thesis aimed to use the adopted methodology described in the previous section to improve the quality of a machining operation. It was important in this case that it could be well understood how certain parameters influence the quality so that the information can be used to make process improvements in the future. This relates to the third research question, *how can data analytics be used to create interpretable models which can be used for quality improvements?*. In the literature review, it was shown how descriptive and diagnostic models have been developed for quality problems before. A common approach is to use supervised models with high interpretability [31], such as tree-based models [33], to find correlations between machine parameters and the quality outcome. An issue with this

approach however is that only showing correlation might make it difficult to interpret the information and find the causality among the parameters, which is required to identify the root cause. Some studies have therefore suggested the use of association rule mining to better visualize how combinations of parameters affect the quality outcome [15, 17]. In this thesis, a combination of feature ranking with interpretable ML models and association rules was used. This approach was demonstrated in the use case where its effectiveness was also evaluated. As was discussed in Section 2.2.2, Bayesian Networks is also a common alternative to deriving root causes in manufacturing. This method was not seen as applicable for this use case however as there was too much information lacking about the process and the influencing factors to be able to create such as model.

Important steps in the used analytics approach presented are to first create intuitive features from the collected data, where it is important to consult domain experts who will use the model to ensure that the features can be well understood. Thereafter, training of ML models can be performed and evaluated, which is then used to rank the features by importance. Analysis with the help of domain experts should be used to evaluate if the ranking of the features seems reasonable. It could then be useful to create visualizations of the data to more thoroughly investigate if anomalous patterns can be spotted for certain features in specific time intervals of a machining cycle, as was illustrated in Figure 4.10. The last step is to choose the most important features and create association rules, for which the Apriori algorithm was used.

A problem that was noticed when using Apriori on an unbalanced data set was that the support measurement was very low for the minority class. When using Apriori, a support threshold needs to be set to make the algorithm find all the candidate parameters and combinations of parameters above this threshold that can then be used to derive rules. Using the original unbalanced data set resulted in unmanageable large volumes of rules when the support threshold was set just above the support of the minority class. Undersampling was therefore applied to improve the balance between the classes which prevented this issue. It is difficult however with this solution to know how the balancing ratio should be set to find the most meaningful rules. If the number of samples from the majority class is reduced significantly, the created rules might become misleading. It could then show that certain conditions will cause quality defects with high confidence, even if these conditions could be active without causing quality defects for a much larger proportion of the cycles where no quality defect occurred. It is therefore recommended to experiment with the underbalancing ratio, and to use reasoning with the help of domain experts to sort out which rules provide important insights.

5.2 Practical contribution

The first research question was *what can be the challenges and requirements for performing data analytics projects in manufacturing?*. In Section 2.3 some key challenge areas identified in literature were explained. It is important to work with all these challenging areas to improve the conditions of being able to gain value from big data

in manufacturing. Through the case study, Volvo's readiness for implementing data analytics was evaluated in terms of these areas which are described in this section. It is also explained how the result of the case study is meaningful for Volvo and what the recommended future directions are for the company to improve the conditions for succeeding with the following pilot projects.

5.2.1 Evaluation of Volvo's readiness for data analytics

Important challenge areas to efficiently use data analytics for quality problems that were identified in this thesis are connectivity and data acquisition, system integration, traceability, data quality, and internal knowledge. Through the case study, it was evaluated how Volvo's current preconditions are within these areas.

Regarding *connectivity and data acquisition*, a working strategy has been implemented where parameters from different machines can be connected to a single communication bridge, Kepware. To be able to establish communication between machines and Kepware however, some requirements need to be in place. A standard communication protocol like OPC-UA or MQTT is be used to enable this communication. Since the factory is old and has been developed over many centuries, there is a mixture of new and old machines. Some of the systems in these machines might lack support for a standard communication protocol, and some can have outdated control systems which cause other issues. This was a problem in this case study, which required a special solution to be established with the help of the control system vendor to enable connectivity for the NC parameters in the machine. For this reason, the effort of collecting required machine parameter data was much higher than anticipated. By performing more projects and gaining more experience, however, it will become easier to identify which machines require more effort for data acquisition. The use of project prioritization as suggested in the extended CRISP-DM methodology can be used to focus on the use cases where high value can be gained in relation to this effort.

For *system integration*, the machining department at Volvo has created good conditions as they have invested in a central data analytics platform, ThingWorx, which can integrate data on demand from various systems. If all the systems have connectivity capability towards Kepware, data from the systems can be collected through ThingWorx. Currently, however, ThingWorx is only capable to collect shop floor data from the most common control systems. The development could thereby be made by enabling system integration and data sharing between more departments other than production, such as maintenance and quality. Quality data was collected in the case study from the measurement machine to be integrated into ThingWorx, but due to the failure of the measurement machine this data could not be collected during the same period as the remaining data. Some data files could however be gathered outside of the analytics platform before this failure, which was analyzed to show how this data can be used for further iterations of the analytics process. If for example maintenance data or data from the manufacturing tracking software could have been used, it would have been possible to also evaluate machine component wear and examine the impact of certain machine alarms. Continued work should be

performed to enable data from a larger variety of sources to be collected through the analytics platform. This will provide better possibilities to detect how quality issues might be connected to information residing from other sources than the production data.

The procedure of collecting and managing data with the data analytics platform worked well in the case study. The process of pre-structuring the data for all machine parameters in a standard format as described in Section 4.2.2 made it possible to convert the data into the desired format for analytics with low effort. In this case study, the data files were extracted from the platform and the analysis was performed with Python. The reason for this was to achieve the high flexibility that open-source programming languish can offer to be able to perform a wide range of data preparation and modeling tasks. It is however also possible to perform analytics directly through ThingWorx. The benefit of this could be that the data can be collected, analyzed, and deployed through the same platform. This could be beneficial when using data analytics at a large scale as it could speed up the process. Every analytics process is unique which often requires a lot of flexibility to try different data preparation and modeling techniques before a good solution is found. It could therefore be difficult to integrate tools where all these techniques can be used in an analytics platform as discussed in [3]. The complexity of data analytics projects varies, however, and certain problems might be very similar regarding which data is used and how this data should preparade and modeled. It could therefore be possible to develop tools for faster deployment of models for such problems directly in the platform, while more complex problems requiring higher flexibility could be handled with a similar approach as in the case study.

Traceability was discussed in Section 2.3.3 to often be an important prerequisite for solving quality-related problems in manufacturing with data analytics. For this case study, a temporary solution to achieve traceability was achieved by setting up a camera that scans the barcodes of the products to identify product IDs. This solution provided traceability for the data estimated to be most crucial in the case study, which was the machining data from the different units and the quality data. This camera was relatively expensive and might not be a feasible solution for enabling traceability throughout the whole factory or between factories. As quality issues in multi-stage manufacturing systems often can reside from previous processes which have affected the product [54, 55, 56], a higher level of traceability is often needed to be able to capture this data. It was for example discussed in the case study that the material hardness could have a high impact on the resulting quality outcome. This data was not obtainable however as it would have required traceability to the foundry where the material hardness is measured.

The *data quality* is also an important aspect to consider since high data quality is needed to be able to gain valuable insights. One of the main issues with the data identified in the case study was the unbalance between the classes (Ok or NOK quality outcome). It is therefore important to ensure that enough data is collected to be able to obtain robust analytics results. The data which was collected had otherwise generally high quality. The main issue was missing values. This

issue was more severe for certain parameters but could be dealt with through data preparation for most of the collected samples. A constraint in the case study was that measurements for parameters could not be collected with a higher frequency than 1 Hz, as this could risk overloading the control system. This could have harmed the quality of for example the load parameters, as higher granularity could have been achieved in the time series data with higher sampling rates.

Internal knowledge and interpretability is crucial to set up effective teams and succeed with data analytics projects, as was also discussed in Section 5.1.2. There are many new areas where new knowledge needs to be acquired to improve the efficient implementation of data analytics projects. A general understanding of at least the basics of data analytics as well as the production system is necessary to efficiently share knowledge between domain experts and data analysts in the team. Besides this general knowledge, it is also required to develop new competencies for how to extract data from machines in a way that has not been done before, how to efficiently store and manage this data, and how to ultimately analyze and extract information from the data. This case study has provided new knowledge in all these areas.

5.2.2 Implications from case study result

The motivation behind the case study was to reduce the number of defects caused by dimensional variations in a machining process which has caused high costs for Volvo. It was decided that appropriate solutions to achieve this could be with statistical process monitoring and also by finding patterns that can be used to derive root causes. These solutions relate to the descriptive and diagnostic levels of data analytics presented in Section 2.1.1. This was determined to be a feasible approach for improving the process concerning the current preconditions at Volvo. The more advanced levels, predictive and prescriptive analysis, could be evaluated as possible solutions to further improve the process with help of data analytics in the future and achieve higher business value. Solutions relating to proactive prediction of the quality outcome (described in Section 2.2.3) or optimization in regard to quality (described in Section 2.2.4) could then be evaluated. As described in [7], it is wise to implement solutions on the different levels step by step to ensure that the complexity does not exceed the company's capability.

With the used analytics approach it was possible to find rules showing which parameters have high importance for the quality. It was also possible to find the critical ranges of the relevant parameters and see how different parameters relate to each other. This approach should therefore be possible to use for later creating a process monitoring solution where operators can keep track of the process stability. The diagnostic capability can also aid in finding previously unknown root causes so improvements can be made to the process. The derived rules showed that the wear of certain tools has a large impact as well on the temperatures of spindles. It was also shown that there is a higher risk to receive quality errors if the machine has been idle for a long period. It could also be seen fixture 3 had a higher chance of causing defects than the other fixtures, but this rule had relatively low confidence. It should be noted however that due to the few samples of NOK samples collected, the high

oversampling rate used for balancing might have caused bias in the data set. For a more robust evaluation of which parameters are most significant for the quality, more data should be collected as was described in Section 4.2.6.

Besides the collection of more data, additional further actions are required before the deployment stage can be entered. The first step should be to perform the same analytics steps already applied once again to evaluate how the result changes with a higher volume of data. Depending on the results, it then needs to be decided if more parameters need to be accessed or if better results can be achieved by trying different data preparation strategies. It should be evaluated if the clustering method of the measurement data with the Isolation Forest algorithm which was presented in the case study could generate better results. The target value used to obtain these results was extracted from a manually recorded log which showed which products received too large or too small diameters. As discussed in [47], such logs can be prone to errors, where for example not all outcomes are labeled or where some outcomes have been mislabeled.

5.2.3 Recommended future directions

The insights gained from performing the case study resulted in some recommendations for how the machining department at Volvo can continue working with the implementation of data analytics. First and foremost, it is necessary to increase the knowledge of data analytics so efficient teams can be created that perform more pilot projects. Upskilling of personnel, dedicated hiring, or third-party collaborations is some way to achieve this increase of knowledge [11].

It is also recommended to evaluate the usage of edge devices which was discussed in Section 2.3.1. A benefit of edge devices is that they can enable the extraction of data with a higher velocity. This can create better possibilities for data extraction from the older machines and avoid certain constraints which were present in this case study. They also reduce the bandwidth requirements of the network as data can be stored and transferred in packages instead of having a continuous flow of data between several machines and the analytics platform. Edge devices can be invested in and used only for the machines where they are needed, and therefore allows for a modular upgrading approach.

Continued work should also be performed with the integration of data from more departments and manufacturing software to ThingWorx. This will enable a wider range of problems to be solved as it will be possible to access all data that can be of interest for a certain problem for analysis. Data silos can be broken down one by one with prioritization to the data sources which are most necessary to access to enable more potential use cases.

An increased level of traceability should also be sought as this will be of high importance to retrieve necessary data when working with quality problems. Temporary traceability solutions can be set up similarly to the one in the case study. For scaling

the usage of data analytics, it would however be necessary to implement a solution that enables a higher level of traceability in the factory and between factories.

The way of managing data in ThingWorx and enabling convenient extraction for a data analytics software of choice proved efficient in the case study. The possibility of performing analytics directly in ThingWorx should however also be evaluated. This might be more convenient for faster implementation of certain use cases which are similar to each other, and where less analysis flexibility is needed than other analytics software can provide.

Different kinds of pilot projects should be performed to learn for which use cases data analytics can create the highest business value. It is recommended to start with the use cases which require lower investments and where it is known that required data sources can be accessed with low effort, following the use case prioritization explained in Section 3.2. The proposed methodology in this thesis can be used as a foundation for performing more projects, where it can be refined over time with continuous improvement to enable more efficient implementation. When a standardized implementation process with a proven ability to create business value has been achieved, it will be possible to move on with implementing data analytics on a larger scale.

6

Conclusion

As the innovation in techniques for creating business value from big data progresses, the relevance to start making use of advanced data analytics in manufacturing increases. There are large amounts of data being produced in companies today which are not utilized beyond its original intended purpose, even if important insights can be gained from the combination of all this data. There are several challenges to overcome before this can be realized. Many manufacturing companies today need to gain a better understanding of these challenges to be able to adopt data analytics for improving their operations. This thesis aimed to improve the understanding of the challenges and requirements that need to be overcome to achieve this. It was also to provide a suitable methodology for how data analytics can be implemented efficiently. This increased understanding can help manufacturing companies in a beginner stage of adopting data analytics to improve their conditions for successful pilot projects. The pilot projects serve as an important learning path to better understand which areas need to be improved and which use cases can bring the most value for the companies. When the value of data analytics can be proven and a process for performing projects has been standardized, it is possible to scale the usage of data analytics in the company.

In the thesis, a common methodology for performing data analytics projects, CRISP-DM, was adapted to be better suited for manufacturing specific problems. This methodology was evaluated in a case study performed at a machining department at Volvo. Through the case study, Volvo's readiness for implementing data analytics was evaluated. It was concluded that increased knowledge of data analytics among the personnel would improve the efficiency of following the proposed methodology. It was also recommended that certain technical developments regarding connectivity, system integration, and traceability be performed to achieve better conditions for large-scale data analytics usage.

The proposed methodology is applicable for a variety of use cases, but in this thesis, the main focus has been on how it can be used to solve quality-related problems. In the performed case study, it was of high importance to understand how different machine parameters and other influencing factors impact the quality of the investigated machining process. The chosen approach was therefore to use interpretable ML models in combination with association rule mining to derive important parameters and show how these relate to each other and the quality output. The results

showed that the used approach can be used to present information in an intuitive way which can enable domain experts to use the information to perform process improvements. However, due to a large imbalance in the historical data set, more data should be collected so more robust analytics results can be obtained.

In summary, this thesis has provided an example that Volvo and other companies can use to prioritize necessary actions that need to be taken to improve readiness for data analytics. The proposed methodology can constitute a basis for a standardized implementation process that can be further adapted for a specific company's needs. The case study provides an example of how the methodology is used in practice and shows an approach to handling quality problems where a high level of interpretability is required. Inspiration can be taken from this example to develop a complete approach for performing data analytics projects. This can in turn improve the chances of succeeding with pilot projects and help these companies on the way towards utilizing data analytics on a larger scale.

Bibliography

[1] Hong-Ning Dai et al. "Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies". In: *Enterprise Information Systems* 14.9-10 (2020), pp. 1279–1303.

[2] Chao Shang and Fengqi You. "Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era". In: *Engineering* 5.6 (2019), pp. 1010–1016.

[3] Yesheng Cui, Sami Kara, and Ka C Chan. "Manufacturing big data ecosystem: A systematic literature review". In: *Robotics and computer-integrated Manufacturing* 62 (2020), p. 101861.

[4] Harald Bauer et al. "Smartening up with artificial intelligence". In: *Hg. v. McKinsey & Company* (2017).

[5] Florian Ungermann et al. "Data analytics for manufacturing systems–a data-driven approach for process optimization". In: *Procedia CIRP* 81 (2019), pp. 369–374.

[6] Ondrej Burkacky et al. *A blueprint for successful digital transformations for automotive suppliers*. 2018.

[7] Günther Schuh et al. *Industrie 4.0 Maturity Index: Die digitale Transformation von Unternehmen gestalten*. Herbert Utz Verlag, 2017.

[8] Andrew Kusiak. "Smart manufacturing must embrace big data". In: *Nature* 544.7648 (2017), pp. 23–25.

[9] Thorsten Wuest et al. "Machine learning in manufacturing: advantages, challenges, and applications". In: *Production & Manufacturing Research* 4.1 (2016), pp. 23–45.

[10] Eric Auschitzky, Markus Hammer, and Agesan Rajagopaul. "How big data can improve manufacturing". In: *McKinsey & Company* 822 (2014).

[11] Capgemini. *Scaling AI in Manufacturing Operations: A Practitioners' Perspective*. Brochure. 2019. URL: `https://www.capgemini.com/research/scaling-ai-in-manufacturing-operations/`.

[12] Francesc Bonada et al. "AI for improving the overall equipment efficiency in manufacturing industry". In: *New Trends in the Use of Artificial Intelligence for the Industry 4.0*. IntechOpen, 2020.

[13] Amine Belhadi et al. "Understanding big data analytics for manufacturing processes: insights from literature review and multiple case studies". In: *Computers & Industrial Engineering* 137 (2019), p. 106099.

[14] Ke Xu et al. “Advanced data collection and analysis in data-driven manufacturing process”. In: *Chinese Journal of Mechanical Engineering* 33.1 (2020), pp. 1–21.

[15] Hung-An Kao et al. “Quality prediction modeling for multistage manufacturing based on classification and association rule mining”. In: *MATEC Web of Conferences*. Vol. 123. EDP Sciences. 2017, p. 00029.

[16] Jay Lee et al. “Recent advances and trends in predictive manufacturing systems in big data environment”. In: *Manufacturing letters* 1.1 (2013), pp. 38–41.

[17] Alican Dogan and Derya Birant. “Machine learning and data mining in manufacturing”. In: *Expert Systems with Applications* 166 (2021), p. 114060.

[18] Dorina Weichert et al. “A review of machine learning for the optimization of production processes”. In: *The International Journal of Advanced Manufacturing Technology* 104.5 (2019), pp. 1889–1902.

[19] Rahul Rai et al. *Machine learning in manufacturing and industry 4.0 applications*. 2021.

[20] Dong-Hyeon Kim et al. “Smart machining process using machine learning: A review and perspective on machining industry”. In: *International Journal of Precision Engineering and Manufacturing-Green Technology* 5.4 (2018), pp. 555–568.

[21] Keith Shaw. *From POC to production: Keys to scaling machine learning*. 2020. URL: `https://www.cio.com/article/193905/from-poc-to-production-keys-to-scaling-machine-learning.html` (visited on 04/22/2022).

[22] M Mayo. “Frameworks for approaching the machine learning process”. In: *KDnuggets* (2018).

[23] Jianjun Shi. *Stream of variation modeling and analysis for multistage manufacturing processes*. CRC press, 2006.

[24] Turgay Kivak, Kasım Habali, and Ulvi ŞEKER. “The effect of cutting paramaters on the hole quality and tool wear during the drilling of Inconel 718”. In: *Gazi University Journal of Science* 25.2 (2012), pp. 533–540.

[25] Sandvik Coromant. *Troubleshooting for reaming*. URL: `https://www.sandvik.coromant.com/en-gb/knowledge/reaming/pages/troubleshooting.aspx` (visited on 04/22/2022).

[26] Sebastian Schorr et al. “Quality Prediction of Reamed Bores Based on Process Data and Machine Learning Algorithm: A Contribution to a More Sustainable Manufacturing”. In: *Procedia Manufacturing* 43 (2020), pp. 519–526.

[27] Kevin Dunn. “Process improvement using data”. In: *Experimentation for Improvement. Hamilton, Ontario, Canada. Creative Commons Attribution-ShareAlike* 4 (2019), pp. 325–404.

[28] Q Peter He and Jin Wang. “Statistical process monitoring as a big data analytics tool for smart manufacturing”. In: *Journal of Process Control* 67 (2018), pp. 35–43.

[29] Carlos A Escobar et al. “Process-monitoring-for-quality—applications”. In: *Manufacturing letters* 16 (2018), pp. 14–17.

[30] Anna Lokrantz, Emil Gustavsson, and Mats Jirstrand. "Root cause analysis of failures and quality deviations in manufacturing using machine learning". In: *Procedia Cirp* 72 (2018), pp. 1057–1062.

[31] Vera L Miguéis, José L Borges, et al. "Automatic root cause analysis in manufacturing: an overview & conceptualization". In: *Journal of Intelligent Manufacturing* (2022), pp. 1–18.

[32] M Taisch et al. "World manufacturing report 2020: manufacturing in the age of artificial intelligence". In: (2020).

[33] Tobias Mueller et al. "Automated root cause analysis of non-conformities with machine learning algorithms". In: *Journal of Machine Engineering* 18 (2018).

[34] Prasanth Lade, Rumi Ghosh, and Soundar Srinivasan. "Manufacturing analytics and industrial internet of things". In: *IEEE Intelligent Systems* 32.3 (2017), pp. 74–79.

[35] Carlos A Escobar, Megan E McGovern, and Ruben Morales-Menendez. "Quality 4.0: a review of big data challenges in manufacturing". In: *Journal of Intelligent Manufacturing* 32.8 (2021), pp. 2319–2334.

[36] Ricardo Silva Peres et al. "Multistage quality control using machine learning in the automotive industry". In: *IEEE Access* 7 (2019), pp. 79908–79916.

[37] Ricardo Coppel et al. "Adaptive control optimization in micro-milling of hardened steels—evaluation of optimization approaches". In: *The International Journal of Advanced Manufacturing Technology* 84.9 (2016), pp. 2219–2238.

[38] Girish Kant and Kuldip Singh Sangwan. "Predictive modelling and optimization of machining parameters to minimize surface roughness using artificial neural network coupled with genetic algorithm". In: *Procedia Cirp* 31 (2015), pp. 453–458.

[39] Alexander S. Gillis. *5 V's of big data*. 2021. URL: `https://www.techtarget.com/searchdatamanagement/definition/5-Vs-of-big-data` (visited on 04/06/2022).

[40] Andreas Kirmse et al. "An Architecture for Efficient Integration and Harmonization of Heterogeneous, Distributed Data Sources Enabling Big Data Analytics." In: *ICEIS (1)*. 2018, pp. 175–182.

[41] Miriam Schleipen et al. "OPC UA & Industrie 4.0-enabling technology with high diversity and variability". In: *Procedia Cirp* 57 (2016), pp. 315–320.

[42] outlier automation. *What's the Difference Between OPC UA and MQTT?* 2020. URL: `https://www.outlierautomation.com/blog/202020/12/16/whats-the-difference-between-opc-ua-and-mqtt` (visited on 04/06/2022).

[43] Jasmine Chan. *Cloud Computing in Manufacturing*. 2021. URL: `https://tulip.co/blog/cloud-computing-in-manufacturing/` (visited on 04/06/2022).

[44] Sachchidanand Singh. "Optimize cloud computations using edge computing". In: *2017 International Conference on Big Data, IoT and Data Science (BID)*. IEEE. 2017, pp. 49–53.

[45] outlier automation. *Practical Edge Computing for Manufacturing*. 2021. URL: `https://www.outlierautomation.com/blog/2021/02/18/practical-edge-computing-for-manufacturing` (visited on 04/06/2022).

[46] Baotong Chen et al. “Edge computing in IoT-based manufacturing”. In: *IEEE Communications Magazine* 56.9 (2018), pp. 103–109.

[47] Kosmas Alexopoulos, Nikolaos Nikolakis, and George Chryssolouris. “Digital twin-driven supervised machine learning for the development of artificial intelligence applications in manufacturing”. In: *International Journal of Computer Integrated Manufacturing* 33.5 (2020), pp. 429–439.

[48] Thorsten Wuest, Christopher Irgens, and Klaus-Dieter Thoben. “An approach to monitoring quality in manufacturing using supervised machine learning on product state data”. In: *Journal of Intelligent Manufacturing* 25.5 (2014), pp. 1167–1180.

[49] Julien Dallemand. *HOW DATA INTEGRATION IS TEARING DOWN DATA SILOS FOR INDUSTRY 4.0*. URL: `https://blog.datumize.com/how-data-integration-is-tearing-down-data-silos-for-industry-4.0` (visited on 04/10/2022).

[50] Jungyub Woo et al. “Developing a big data analytics platform for manufacturing systems: architecture, method, and implementation”. In: *The International Journal of Advanced Manufacturing Technology* 99.9 (2018), pp. 2193–2217.

[51] Reuben Schuitemaker and Xun Xu. “Product traceability in manufacturing: A technical review”. In: *Procedia CIRP* 93 (2020), pp. 700–705.

[52] Michail J Beliatis et al. “Next generation industrial IoT digitalization for traceability in metal manufacturing industry: A case study of industry 4.0”. In: *Electronics* 10.5 (2021), p. 628.

[53] Jing-Doo Wang. “A novel approach to improve quality control by comparing the tagged sequences of product traceability”. In: *MATEC Web of Conferences*. Vol. 201. EDP Sciences. 2018, p. 05002.

[54] Daniel Lieber et al. “Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning”. In: *Procedia Cirp* 7 (2013), pp. 193–198.

[55] Fahmi Arif, Nanna Suryana, and Burairah Hussin. “Cascade quality prediction method using multiple PCA+ ID3 for multi-stage manufacturing system”. In: *Ieri Procedia* 4 (2013), pp. 201–207.

[56] Fahmi Arif, Nanna Suryana, and Burairah Hussin. “A data mining approach for developing quality prediction model in multi-stage manufacturing”. In: *International Journal of Computer Applications* 69.22 (2013).

[57] Xinyan Ou et al. “First time quality diagnostics and improvement through data analysis: A study of a crankshaft line”. In: *Procedia Manufacturing* 49 (2020), pp. 2–8.

[58] Muhammad Syafrudin et al. “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 (2018), p. 2946.

[59] Jason Brownlee. “Tactics to combat imbalanced classes in your machine learning dataset”. In: *Machine Learning Mastery* 19 (2015).

[60] Jungyub Woo, Seung-Jun Shin, and Wonchul Seo. “Developing a big data analytics platform for increasing sustainability performance in machining operations”. In: *Proceedings of 26th International Conference on Flexible Automation and Intelligent Manufacturing, Seoul, Republic of Korea*. 2016.

[61] Allison Buenemann. *Scaling machine learning for the manufacturing masses*. 2021. URL: https://www.controleng.com/articles/scaling-machine-learning-for-the-manufacturing-masses/ (visited on 04/10/2022).

[62] Marina Burdack and Manfred Rössle. "A concept of an interactive web-based machine learning tool for individual machine and production monitoring". In: *Intelligent Decision Technologies 2019*. Springer, 2019, pp. 183–193.

[63] Eivind Kristoffersen et al. "Exploring the relationship between data science and circular economy: an enhanced CRISP-DM process model". In: *Conference on e-Business, e-Services and e-Society*. Springer. 2019, pp. 177–189.

[64] Fernando Martínez-Plumed et al. "CRISP-DM twenty years later: From data mining processes to data science trajectories". In: *IEEE Transactions on Knowledge and Data Engineering* (2019).

[65] Franziska Schäfer et al. "Synthesizing CRISP-DM and quality management: a data mining approach for production processes". In: *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*. IEEE. 2018, pp. 190–195.

[66] Asma Ladj et al. "A knowledge-based Digital Shadow for machining industry in a Digital Twin perspective". In: *Journal of Manufacturing Systems* 58 (2021), pp. 168–179.

[67] Steffen Huber et al. "DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model". In: *Procedia Cirp* 79 (2019), pp. 403–408.

[68] KDnuggets. *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. 2014. URL: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html (visited on 03/03/2022).

[69] IBM. *Introduction to CRISP-DM*. URL: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-introduction-crisp-dm (visited on 03/03/2022).

[70] Rüdiger Wirth and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. Manchester. 2000, pp. 29–40.

[71] Sadhvi Anunaya. *Data Preprocessing in Data Mining -A Hands On Guide*. 2021. URL: https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/ (visited on 03/09/2022).

[72] Pragati Baheti. *A Simple Guide to Data Preprocessing in Machine Learning*. 2022. URL: https://www.v7labs.com/blog/data-preprocessing-guide (visited on 03/09/2022).

[73] Achim Kampker et al. "Enabling data analytics in large scale manufacturing". In: *Procedia Manufacturing* 24 (2018), pp. 120–127.

[74] Satyabrata Pradhan et al. "A Bayesian network based approach for root-cause-analysis in manufacturing process". In: *2007 International Conference on Computational Intelligence and Security (CIS 2007)*. IEEE. 2007, pp. 10–14.

[75] Wei Ji and Lihui Wang. "Big data analytics based optimisation for enriched process planning: a methodology". In: *Procedia CIRP* 63 (2017), pp. 161–166.
[76] Wei Ji and Lihui Wang. "Big data analytics based fault prediction for shop floor scheduling". In: *Journal of Manufacturing Systems* 43 (2017), pp. 187–194.
[77] Jason Brownlee. *How to remove outliers for machine learning*. 2020.
[78] Jason Brownlee. "Why one-hot encode data in machine learning". In: *Machine Learning Mastery* (2017), pp. 1–46.
[79] Evgeniy Latyshev. "Sensor Data Preprocessing, Feature Engineering and Equipment Remaining Lifetime Forecasting for Predictive Maintenance." In: *DAMDID/RCDL*. 2018, pp. 226–231.
[80] Maximilian Christ, Andreas W Kempa-Liehr, and Michael Feindt. "Distributed and parallel time series feature extraction for industrial big data applications". In: *arXiv preprint arXiv:1610.07717* (2016).
[81] Analytics Vidhya. *Feature Engineering in IoT Age – How to deal with IoT data and create features for machine learning?* 2017. URL: https://www.analyticsvidhya.com/blog/2017/04/feature-engineering-in-iot-age-how-to-deal-with-iot-data-and-create-features-for-machine-learning/ (visited on 05/26/2022).
[82] Sampath Kumar Gajawada. "ANOVA for Feature Selection in Machine Learning". In: *Towards Data Science* (2019).
[83] Analytics Vidhya. *Anomaly detection using Isolation Forest – A Complete Guide*. 2021. URL: https://www.analyticsvidhya.com/blog/2021/07/anomaly-detection-using-isolation-forest-a-complete-guide/ (visited on 05/26/2022).
[84] Jason Brownlee. "How to calculate feature importance with python". In: *Machine Learning Mastery. https://machinelearningmastery. com/calculate-feature-importance-with-python* (2020).
[85] Prashant Gupta. "Decision trees in machine learning". In: *Towards Data Science* 17 (2017).
[86] Tony Yiu. "Understanding random forest". In: *Towards data science* 1 (2019), pp. 1–11.
[87] Analytics Vidhya. *Understanding Random Forest*. 2021. URL: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ (visited on 05/26/2022).
[88] Jason Brownlee. "XGBoost with Python". In: *Machine Learning Mastery* (2019).
[89] IBM. *What is logistic regression?* URL: https://www.ibm.com/topics/logistic-regression (visited on 05/26/2022).
[90] Chonyy. *Apriori: Association Rule Mining In-depth Explanation and Python Implementation*. 2020. URL: https://towardsdatascience.com/apriori-association-rule-mining-explanation-and-python-implementation-290b42afdfc6 (visited on 05/26/2022).
[91] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.
[92] Analytics Vidhya. *10 Techniques to deal with Imbalanced Classes in Machine Learning*. 2020. URL: https://www.analyticsvidhya.com/blog/2020/07/

10-techniques-to-deal-with-class-imbalance-in-machine-learning/ (visited on 05/26/2022).

A

Association rules

In this appendix, the full set of the derived association rules where the consequent was a NOK quality outcome can be seen. The support, confidence, and lift can also be seen for each rule.

antecedents	consequents	support	confidence	lift
EH3_T1304_Sp4_15_High', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
EH3_T1304_Sp4_15_High', 'idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
idle_time_Very_high'	NOK	0,2	1,0	3,0
EH2_T1306_427_Low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
EH2_T1306_427_Low', 'idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
EH3_T1303_Sp_2_15_Very_high', 'EH2_T1306_427_Low', 'EH2_T1306_1465_Low'	NOK	0,2	1,0	3,0
EH2_T1306_1465_Low', 'EH3_T1304_Sp4_15_High', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
EH2_T1306_427_Low', 'EH3_T1304_Sp4_15_High', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
EH3_T1303_Sp_2_15_Very_high', 'EH2_T1306_427_Low', 'idle_time_Very_low', 'EH2_T1306_1465_Low'	NOK	0,2	1,0	3,0
EH2_T1306_1465_Low', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
EH2_T1306_427_Low', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	1,0	3,0
EH3_T1303_Sp_2_15_Very_high', 'EH2_T1306_427_Low', 'EH3_T1304_Sp4_15_High', 'EH2_T1306_1465_Low'	NOK	0,2	1,0	3,0
idle_time_Very_low', 'EH2_T1306_427_Low', 'EH2_T1306_1465_Low', 'EH3_T1303_Sp_2_15_Very_high', 'EH3_T1304_Sp4_15_High'	NOK	0,2	1,0	3,0
EH2_T1306_1465_Very_low', 'idle_time_Very_high'	NOK	0,1	1,0	3,0
idle_time_Very_high', 'EH2_T1306_427_Very_low'	NOK	0,1	1,0	3,0
EH2_T1306_1465_Very_low', 'idle_time_Very_high', 'EH2_T1306_427_Very_low'	NOK	0,1	1,0	3,0
Fixture_3', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,1	1,0	3,0
EH3_T1303_Sp_2_15_Very_high', 'SP34_TEMP_High', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
EH2_T1306_427_Low', 'SP34_TEMP_High', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,1	1,0	3,0
Fixture_3', 'SP34_TEMP_High', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'idle_time_Very_low', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
EH2_T1306_427_Low', 'SP34_TEMP_High', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0

EH3_T1303_Sp_2_15_Very_high', 'SP34_TEMP_High', 'idle_time_Very_low', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
EH2_T1306_427_Low', 'SP34_TEMP_High', 'idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,1	1,0	3,0
EH2_T1306_427_Low', 'EH3_T1303_Sp_2_15_Very_high', 'SP34_TEMP_High', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
EH3_T1303_Sp_2_15_Very_high', 'SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
EH2_T1306_427_Low', 'SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,1	1,0	3,0
EH2_T1306_427_Low', 'SP34_TEMP_High', 'idle_time_Very_low', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
EH2_T1306_427_Low', 'SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'EH2_T1306_1465_Low'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'idle_time_Very_low', 'EH2_T1306_427_Low', 'EH2_T1306_1465_Low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'idle_time_Very_low', 'EH2_T1306_1465_Low', 'EH3_T1303_Sp_2_15_Very_high', 'EH3_T1304_Sp4_15_High'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'idle_time_Very_low', 'EH2_T1306_427_Low', 'EH3_T1303_Sp_2_15_Very_high', 'EH3_T1304_Sp4_15_High'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'EH2_T1306_427_Low', 'EH2_T1306_1465_Low', 'EH3_T1303_Sp_2_15_Very_high', 'EH3_T1304_Sp4_15_High'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'idle_time_Very_low', 'EH2_T1306_427_Low', 'EH2_T1306_1465_Low', 'EH3_T1304_Sp4_15_High'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'idle_time_Very_low', 'EH2_T1306_427_Low', 'EH2_T1306_1465_Low', 'EH3_T1303_Sp_2_15_Very_high', 'EH3_T1304_Sp4_15_High'	NOK	0,1	1,0	3,0
SP34_TEMP_High', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	0,9	2,7
EH2_T1306_1465_Low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	0,9	2,6
EH2_T1306_427_Low', 'EH2_T1306_1465_Low'	NOK	0,2	0,9	2,6
EH2_T1306_1465_Low', 'EH3_T1304_Sp4_15_High'	NOK	0,2	0,9	2,6
EH2_T1306_1465_Low', 'idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	0,9	2,6

EH2_T1306_427_Low', 'idle_time_Very_low', 'EH2_T1306_1465_Low'	NOK	0,2	0,9	2,6
EH2_T1306_1465_Low', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low'	NOK	0,2	0,9	2,6
EH2_T1306_427_Low', 'EH3_T1304_Sp4_15_High', 'EH2_T1306_1465_Low'	NOK	0,2	0,9	2,6
EH2_T1306_427_Low', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low', 'EH2_T1306_1465_Low'	NOK	0,2	0,9	2,6
SP34_TEMP_High', 'EH2_T1306_1465_Low'	NOK	0,1	0,8	2,5
SP34_TEMP_High', 'idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,1	0,8	2,5
EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	0,8	2,4
SP34_TEMP_High', 'EH3_T1304_Sp4_15_High'	NOK	0,2	0,8	2,3
EH2_T1306_1465_Very_low', 'EH2_T1306_427_Very_low'	NOK	0,2	0,8	2,3
idle_time_Very_low', 'EH3_T1303_Sp_2_15_Very_high'	NOK	0,2	0,7	2,2
SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low'	NOK	0,1	0,7	2,1
EH2_T1306_427_Low', 'SP34_TEMP_High', 'EH3_T1304_Sp4_15_High'	NOK	0,1	0,7	2,1
EH2_T1306_427_Low', 'SP34_TEMP_High', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low'	NOK	0,1	0,7	2,1
EH2_T1306_427_Low', 'EH3_T1304_Sp4_15_High'	NOK	0,2	0,7	2,1
EH2_T1306_427_Low', 'EH3_T1304_Sp4_15_High', 'idle_time_Very_low'	NOK	0,2	0,7	2,1
EH3_T1304_Sp4_15_High'	NOK	0,2	0,6	1,9
EH2_T1306_427_Low', 'SP34_TEMP_High'	NOK	0,1	0,6	1,9
EH2_T1306_427_Low', 'SP34_TEMP_High', 'idle_time_Very_low'	NOK	0,1	0,6	1,9
EH2_T1306_1465_Low', 'idle_time_Very_low'	NOK	0,2	0,6	1,8
EH3_T1304_Sp4_15_High', 'idle_time_Very_low'	NOK	0,2	0,6	1,7
Fixture_3', 'SP34_TEMP_High'	NOK	0,1	0,5	1,6
EH2_T1306_1465_Very_low'	NOK	0,2	0,5	1,6
EH2_T1306_1465_Low'	NOK	0,2	0,5	1,6
EH2_T1306_427_Low'	NOK	0,2	0,5	1,5
EH2_T1306_427_Low', 'idle_time_Very_low'	NOK	0,2	0,5	1,5
SP31_TEMP_High', 'SP34_TEMP_High'	NOK	0,1	0,5	1,5
SP34_TEMP_High'	NOK	0,2	0,5	1,4
SP31_TEMP_High'	NOK	0,1	0,5	1,4
EH2_T1306_427_Very_low'	NOK	0,2	0,4	1,3
Fixture_3'	NOK	0,2	0,4	1,3
SP31_TEMP_Very_high', 'EH2_T1306_1465_Low'	NOK	0,1	0,4	1,3
SP34_TEMP_High', 'idle_time_Very_low'	NOK	0,1	0,4	1,2
idle_time_Very_low'	NOK	0,2	0,2	0,7
SP31_TEMP_Very_high'	NOK	0,1	0,2	0,5

DEPARTMENT OF INDUSTRIAL AND MATERIALS SCIENCE
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY