CHALMERS
UNIVERSITY OF TECHNOLOGY

A data-driven approach to improve process robustness for bolted critical joints

PRASANNA SRINIVASAN

DEPARTMENT OF INDUSTRIAL AND MATERIALS SCIENCE

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

Master's thesis 2022

A data-driven approach to improve process robustness for bolted critical joints

PRASANNA SRINIVASAN

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

A data-driven approach to improve process robustness for bolted critical joints
PRASANNA SRINIVASAN

Industrial Supervisor: Quality Assurance manager; Director Manufacturing Engineering & Main from the company ABC.
Academic Supervisor: Ebru Turanoglu Bekar, Chalmers University of Technology
Examiner: Anders Skoogh, Chalmers University of Technology

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

A data-driven approach to improve process robustness for bolted critical joints.
PRASANNA SRINIVASAN
Department of Industrial and Materials Science
Chalmers University of Technology

Abstract

In the current world, the companies are inclining towards advanced, rapidly evolving industry 4.0 technologies for meeting dynamic customer demands and producing products at a much higher volume, less cost, and more efficiently. These technologies facilitate communication links between different systems that talk with each other to use intelligent manufacturing techniques, industrial digitalization, cloud systems, Internet of Things (IoT), and autonomous machines to acquire, store and efficiently handle big data. Some companies hold an edge over their competitors due to their extensive knowledge of processing raw data into valuable insights. The exploration and exploitation of the analytics and visualization techniques using Data Science, Data Mining (DM), and Machine Learning (ML) have become the critical enablers for data-related technologies, forming the base of this thesis.

This thesis study aims to develop a data-driven approach using ML and visualization techniques to build a predictive model and a dashboard for improving the robustness of the bolted critical joint process in a truck production assembly line. Additionally, the study will shed some light on the harmful effect of poor data quality. Moreover, the non-robust process was directly and indirectly impacting various departments in manufacturing such as Quality, Production, Audit, Maintenance, Procurement, and internal information technology (IT). Additionally, it became evident that the existing solution by the third-party provider was not used for developing data insights and used only as a reactive solution.

A supervised ML classification model was developed to identify critical parameters influencing the tightening process outcome and predict this outcome by extracting data from different data sources. A visualization solution with a dashboard is designed to direct the attention of all levels of management to reduce the use of manual actions/backup routines. The results of this study, along with discussions, are presented. The ML model results were promising, with high accuracy for the training dataset, test dataset, and cross-validated with live data. The ML model and dashboard can be used as a standard framework for developing sustainable models in the future, provided the data for the model is automated. The limitation of the study is that some of the results can be generalized, whereas others are not. Furthermore, this study's practical and academic contributions are highlighted along with future recommendations.

Keywords: Bolt tightening, Digitalization, Smart manufacturing, Process improvement, Data-driven approach, Data mining, Automotive production systems.

Acknowledgements

I would like to extend my sincere thanks and gratitude to all who are a part of this thesis for their constant support and assitance. With their able guidance, I could come up with valuable results for my study and most importantly gained a lot of exposure with valuable learnings from the experts who are involved in this thesis (industrial as well as academic experts). This master thesis has been conducted full time in a real industry setting at a Swedish manufacturing company, which is the leading global truck manufacturer. I received academic guidance and support from the department of Industrial and Materials science at Chalmers University of technology.

First of all I would like to extend my thanks to employees at ABC for providing me a valuable opportunity for carrying out my master thesis. A special thanks to senior VPS coach from the U.S, who has been the main thesis coordinator, for believing in me by offering this thesis opportunity and for continuously supporting my work throughout the study. I would like to thank the top management team led by VP Operations of the plant; Director Digitalization & IT; Digital transformation manager and Director R&TD at the ABC plant for funding this study.

I am deeply gratified to thank my main industrial supervisors at ABC, Quality assurance manager and Director Manufacturing Engineer & Main. Both of them have been a pillar of support by mentoring and guiding me with their abundant knowledge. I also want to thank them for creating a healthy atmosphere for me to blend in with the employees at ABC. I also want to thank a project quality engineer, a quality requirement engineer and an engineer from maintenance for their continuous technical support and thank internal audit manager, Director Production, Director VPS Quality Environment, manager product quality and a project quality manager operations for their constant feedback. I finally want to extend my thanks to employees at ABC from different departments.

Secondly, I want to sincerely thank my academic supervisor, Ebru Turanoglu Bekar for the consistent and relentless support and guidance offered by her which is quite instrumental for me to reach this point in this thesis. Her very kind gesture, rapid response to mails and her valuable feedback had inspired and pushed me to come out of the challenges.

Thirdly, I would like to thank my examiner, Anders Skoogh for guiding me in the right path and providing me with constructive and valuable feedback by looking at the problem from a different perception. Finally, I would to like to sincerely thank my friends, family for extending their relentless support and continuously wishing me to get the best things in life.

PRASANNA SRINIVASAN, Gothenburg, June 2022

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis.

ABC	Company where study is conducted
AI	Artificial Intelligence
BI	Business Intelligence
CC	Critical characteristics
CRISP-DM	Cross Industry Standard Process for Data Mining
CL	Control Limits
CLA	Classification Learner Application
DM	Data Mining
DDSM	Data Driven Smart Manufacturing
EDA	Exploratory Data Analysis
EOL	End Of Line
FP	False Positive
FN	False Negative
ICT	Information and Communication Technologies
IoT	Internet Of Things
IT	Information Technology
KDD	Knowledge Discovery from database
KPI	Key Performance Index
LSL	Lower Specification Limit
LCL	Lower Control Limit
ML	Machine Learning
MATLAB	MATrix LABoratory
OEM	Original Equipment Manufacturer
p-set	Program set
SC	Special Characteristics
SEMMA	Sample Explore Modify Model and Assess
SPC	Statistical Process Control
SL	Specification Limits
SQL	Structured Query Language
TP	True Positive
TN	True Negative
Q	Quality gate
UCL	Upper Control Limit
USL	Upper Specification Limit
XYZ	ABC's Global Group

Contents

1 Introduction **1**
1.1 Background . . . 1
1.1.1 Company Description . . . 2
1.1.2 Factory Layout . . . 3
1.2 Aim . . . 4
1.3 Research Questions . . . 4
1.4 Problem Description . . . 5
1.4.1 Technical implications and Business implications . . . 5
1.5 CC joint and current state of CC bolt tightening process in ABC . . 6
1.6 Limitations . . . 10

2 Literature Review **11**
2.1 Industry 4.0 Technology Evolution . . . 11
2.2 Data Driven Smart Manufacturing . . . 12
2.3 Digitalization in Automotive Production Systems . . . 13
2.4 Artificial Intelligence (AI), ML in Manufacturing . . . 14
2.5 Inductive Learning in Supervised Technique: Rule-based Learners and Decision Trees . . . 16
2.6 Data Mining Approaches in Manufacturing Applications . . . 17
2.7 CRISP-DM and its Application in Manufacturing . . . 19
2.8 Data Analytics and Visualization Techniques Application in Automotive Industries . . . 20
2.9 Basics of Bolted joints . . . 23
2.10 Effect of Bad quality on Decision making . . . 24

3 Methodology **27**
3.1 Business Understanding . . . 27
3.2 Data Understanding . . . 28
3.2.1 Exploratory Data Analysis . . . 28
3.3 Data Preparation . . . 28
3.4 Modelling . . . 29
3.4.1 Feature Extraction . . . 29
3.5 Evaluation . . . 29
3.6 Deployment . . . 30

4 Results **31**

4.1 Business Understanding 31
4.2 Data Understanding 35
4.2.1 Data Quality Verification Report for EDA in existing BoltsNut system and data from SQL database 38
4.2.2 Data Quality Report for ML model 44
4.2.3 Manual Actions Dashboard 44
4.3 Data Preparation 45
4.4 Modelling 48
4.4.1 Selecting Modeling Technique 48
4.4.2 Generating Test Design 49
4.4.3 Building the model 49
4.4.3.1 Feature Extraction 49
4.4.3.2 Training the model 51
4.4.3.3 Assessing the model 53
4.4.4 Extracting Decision Rules 56
4.4.5 Visualization Solution for the Ok-adjusted Manual Actions Dashboard 56
4.5 Evaluation 57

5 Discussion **59**
5.1 Academic Contribution 62
5.2 Practical Contribution 63

6 Conclusion and Recommendation **65**
6.1 Recommendations for Improvement 66

Bibliography **69**

References **69**

A Appendix A **I**

B Appendix B **III**

C Appendix C **V**

List of Figures

1.1 Representation of the production layout. 4
1.2 Ok status CC bolt tightening process (Torque controlled/Angle monitoring strategy). 8
1.3 Nok status CC bolt tightening process (Torque controlled/Angle monitoring strategy). 8
1.4 CC bolt tightening two step methodology 9

2.1 The four major industrial revolutions (Pop, 2020) 12
2.2 The evolution of manufacturing data in different timelines (Tao, Qi, Liu, & Kusiak, 2018) 13
2.3 AI and ML applications in manufacturing (Rai, Tiwari, Ivanov, & Dolgui, 2021) 15
2.4 Structuring of ML techniques and algorithms (Wuest, Weimer, Irgens, & Thoben, 2016) 15
2.5 Generic tree adapted (Kotsiantis, 2013; Safavian & Landgrebe, 1991) 17
2.6 Overview of tasks and algorithms in DM (Wuest et al., 2016) 18
2.7 The CRISP-DM life cycle (Azevedo & Santos, 2008) 19
2.8 Illustration of a process under statistical control, adapted from (Burgess, 2018) 21
2.9 Illustration of a control chart (left side) and histogram model (right side) for stable process, adapted from (Burgess, 2018). 22
2.10 Illustration of a process not under statistical control, adapted from (Burgess, 2018) 22
23figure.caption.20
2.12 Basics of bolted joints (Chen, Gao, & Guan, 2017) 24

3.1 CRISP-DM phases and tasks, adapted from (Wirth & Hipp, 2000) . . 30

4.1 Data analytics and visualization solution. 32
4.2 Nok trend per day for March 2022. 35
4.3 Nok trend per day for the month of February 2022. 36
4.4 SPC studies for all joints in a week (Ok + Nok joints). 36
4.5 SPC studies for all joints in a week (Only Ok joints). 37
4.6 Illustration of wider-angle distribution in the selected program in the workstation. 37
4.7 Search result for the 2-month duration in BoltsNut. 38
4.8 Search result for last 12-month duration in BoltsNut. 38

4.9 Scatter plot of the CC joints in the selected workstation, timeline - September 2021 to March 2022. 39
4.10 Histogram of the CC joints in the selected workstation, timeline - September 2021 to March 2022. 39
4.11 Visualization of the undefined status data from the historical database. 40
4.12 Data unreliability in BoltsNut Scenario 1 41
4.13 Data unreliability in BoltsNut Scenario 2 41
4.14 Data unreliability in BoltNut Scenario 3 42
4.15 Duplicates found in the results in BoltsNut scenario 4 42
4.16 Tool locked by the open-source protocol in BoltsNut scenario 5 43
4.17 Actual scenario during the tightening process in a smart tool in ABC 43
4.18 Data unreliability in BoltsNut scenario 5 43
4.19 Historical data for Ok Adjusted Manual actions in January 2021 . . . 45
4.20 The current year 2022 data for the Ok adjusted Manual actions, timeline January 2022 to April 2022. 45
4.21 Visualization of tightening sequence for bolt tightening joints 46
4.22 Representation of constructed raw data 46
4.23 Transformed raw data with reduced columns 47
4.24 Data transformed in the form of timetable 47
4.25 Torque with respect to time in a timetable 48
4.26 Visualization of the time-based features of the training dataset. . . . 49
4.27 Visualization of the frequency-based features of the training dataset. . 50
4.28 Ranking of features for the training data through DFD. 50
4.29 Feature table for training dataset (Table dimension: 1529x109). . . . 51
4.30 Feature table for top six features of training data (Table dimension: 1529x7). 52
4.31 Histograms for the selected top six features. 52
4.32 Validation accuracy of the trained model with selected features for the training dataset. 53
4.33 Confusion matrix for the trained model. 53
4.34 Validation results of the exported model with the test dataset. 54
4.35 Confusion matrix for the exported model with validation of testdata. 55
4.36 Confusion matrix for the exported model with cross-validation. 55
4.37 Illustrates the manual action dashboard. 56
4.38 Drill-down options on the dashboard. 57
4.39 Representation of manual actions performed in multiple areas. 57

A.1 Nok trend per day for the month of January 2022. I
A.2 Nok trend per day for the month of December 2021. I
A.3 Nok trend per day for the month of November 2021. II
A.4 Nok trend per day for the month of October 2021. II

B.1 Signal trace for the extracted time-domain features of angle. III
B.2 Signal trace for the extracted time-domain features of speed. III
B.3 Power spectrum of extracted frequency domain features for angle. . . IV
B.4 Power spectrum of extracted frequency domain features for speed. . . IV

C.1 Confusion matrix with percentage for exported model of the training dataset. V
C.2 Testing resuls of the exported model with cross-validation dataset. . . V
C.3 Extracted rules for the exported model using decision trees. VI

List of Tables

4.1 Overall Status of the joint in the specific timeline in the chosen workstation (Illustrative purpose). 35
4.2 Illustrates the data quality error in the undefined status from the historical database. 40

1
Introduction

This chapter provides the background of this thesis, followed by the company description, factory layout, aim, and formulated research questions. Further, a formulated problem description has been provided to enhance the reader's understanding. Finally, the current state of the process is presented, followed by the study's limitations.

1.1 Background

Over the last decade, manufacturing companies have evolved from mechanization, electrification, and computerization to the current popular industry 4.0 digitalization trend. According to Kang et al. (2016) and Pop (2020), the highly dynamic customer demands and global markets have forced the manufacturing companies to adapt to dynamic customer requirements. These customer adaptations have increased manufacturing complexities and brought the need for advanced technologies that could connect different systems. Industry 4.0 has overcome these barriers and proved to be quite handy, which made current manufacturing companies inclined towards the advanced Information and Communication Technology (ICT) such as the Internet of Things (IoT), smart manufacturing, data security, data processing, data storage in cloud systems, big data, autonomous machines, process simulation, system integration vertically and horizontally (Pop, 2020).

As the manufacturing companies have realized the importance of these technologies for their business, there has been a significant transformation from traditional manufacturing systems into intelligent factories by making considerable investments in infrastructure and resources (Ebrahimi, Baboli, & Rother, 2019). Smart manufacturing effectively combines ICT and traditional manufacturing methodologies to support real-time decision-making (Kang et al., 2016). Moreover, the automotive original equipment manufacturers (OEM) globally face numerous challenges as a change in the global markets leads to more product varieties, increased regulation, and stiff competition from other established players with advanced technologies (Peters et al., 2015). Many researchers have opined that digitalization combined with intelligent manufacturing approaches and ICT can overcome these challenges and can boost the efficient production of variants in the automotive domain.

Smart manufacturing, which emerged from industry 4.0, paved the way for analytics in manufacturing industries (Rai et al., 2021). Smart manufacturing effectively translates data acquired into manufacturing intelligence to impact manufacturing activities positively. The massive volume of data collected in real-time by sensors in automotive plants demanded cloud servers and databases for data storage (Tao et al., 2018). The twenty-first century lies on the data deoxyribo-nucleic acid (DNA), where big data and data economy provide essential knowledge, insights, and potential (Cao, 2017). Interestingly, data DNA's understanding lies in data science and analytics. Understandably, there has been a shift from data analysis on a small scale, with less volume of data along with hypothesis testing, to data analytics on the complex, massive volume of data for hypothesis-free knowledge and insights over the years (Cao, 2017).

According to Rai et al. (2021), this cyber-physical environment with digital solutions and advanced technologies aids industries in monitoring, controlling, and optimizing their process efficiently with the use of data. The evolution of industry 4.0 paved the way for intelligent machines that can gather real-time data from machines in an automotive manufacturing environment. The most prevalent predictive models in automotive manufacturing are data-driven approaches based on statistical process control, pattern recognition, and ML algorithms (Matope & Bekker, 2021). However, the main challenge is that these Cyber Physical Systems (CPS) produce a lot of raw, uncleaned data that needs to be structured with data pre-processing techniques before using the data for further analysis (Sha & Zeadally, 2015; Haug, Zachariassen, & Van Liempd, 2011).

The data analytics and visualization study focused on this thesis is performed in a company ABC which is a leading global truck manufacturer. ABC conducts a lot of bolt tightening operations every day for critical and non-critical joints. The non-critical joint uses pneumatic or hydraulic tools, whereas smart tools perform critical joint tightening operations. With industry 4.0 initiatives and smart manufacturing, ABC has invested in a third-party solution to provide smart equipment for data acquisition and storage. Every day, ABC performs a high volume of bolt tightening operations with these smart types of equipment for critical joints. ABC has not developed any insights into the data acquired and stored. Therefore, the main objective of this study is to improve the reliability of the Critical Characteristics (CC) bolt tightening process using ML techniques to develop insights for better decision-making. Additionally, it will focus on improving the reliability of ABC's third-party solution and highlight the importance of having good quality data for improved decision-making.

1.1.1 Company Description

The company in which this study is conducted is a global automotive giant in the heavy-duty sector producing different variants of highly advanced trucks. The company will hereafter be referred to in this study as ABC. ABC belongs to a global group XYZ. The XYZ group is headquartered in Sweden but active in 10 business areas with 12 segments such as trucks, buses, construction equipment, power solutions for marine and industrial applications, financing and services, marine manufacturing, etcetera. The XYZ group is spread globally in more than 190 markets and employs more than 100,000 people. The global vision of the XYZ group is to drive prosperity and become a leading and successful transport and infrastructure solution provider in the world.

Company ABC is an industrial entity within XYZ's group specializing in manufacturing a vast volume of customized trucks with the latest technology for different customer segments. ABC produces innovative, advanced, world-class industrial trucks, employs around 100,000 employees across 32 countries, and covers 190 markets. ABC focuses on achieving customer success through the sustainable operation. ABC's long-term vision is to serve customers with its products, services, and solutions. It has a sustainability goal within 2030 to produce a 100 percent safe, fossil-free, and more productive environment. The first step in achieving the 2030 sustainability goal is to slowly replace all the tightening tools which run on pneumatic, hydraulic to be replaced by smart electric tools.

In the recent few years, the heavy-duty automotive market in which ABC operates has undergone rapid developments in digitalization, smart manufacturing, and electrification. To keep up with the ongoing trends and stay competitive in the market, ABC has initiated many digitalization initiatives. This thesis will focus on one of ABC's manufacturing plants in Sweden. This plant uses a continuous assembly line to produce medium and heavy-duty trucks from scratch. ABC is in line with its steady progress towards the 2030 sustainability goal with its digitalization initiatives, smart manufacturing, automation, smart electrical tightening equipment for CC joints, and its ongoing shift towards electro-mobility. The manufacturing plant production layout is explained in detail in the following section.

1.1.2 Factory Layout

The current factory layout includes one main assembly line supported by three feeder lines and many sub-assembly lines. The main assembly line consists of three main divisions, namely the Base module, final assembly 1 (FA1), and final assembly 2 (FA2). The feeder lines build axles, engines, and cabs and feed them to the mainline (Refer to Figure 1.1). Each division is broken into several small parts that contain individual workstations. The production in the plant is supported by quality, production, safety, audit, maintenance, and procurement departments. To ensure the quality of their products, the production line has quality gates at three different points of the line that is Q6, Q7, and Q8. Q6 is located at the end of the mainline, Q7 is near the lighter adjustment area, and the last quality gate, Q8, is near the heavier adjustments area. The Q8 is the final quality gate that must be cleared before delivering the final product to the customer. Figure 1.1 represents the production layout in the company ABC.

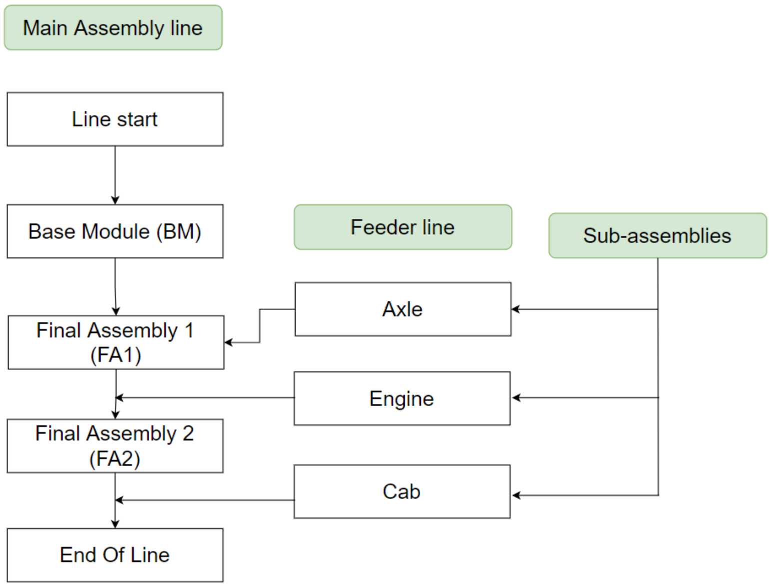


Figure 1.1: Representation of the production layout.

1.2 Aim

The thesis aims to develop a predictive model using ML techniques to predict and improve the reliability of the tightening process for CC joints. Moreover, to identify and demonstrate how data analytics and visualization improve the decision-making process for various departments within manufacturing. This study will also shed some light on poor data quality in the existing smart system and how the developed ML model can be beneficial for the company.

1.3 Research Questions

The following research questions (RQs) are formulated and will be investigated in the following sections to understand the problem at hand better.

RQ1: How can a data-driven approach be used to predict and improve the reliability of the tightening process for critical joints?

RQ2: How can the data collected from different sources be used to improve decision-making by various departments in manufacturing (Quality, Production engineering, Maintenance, Procurement, Audit and IT team)?

1.4 Problem Description

A robust process is deemed critical in mass manufacturing assembly lines. One minor problem could significantly impact another process on the line. The overall equipment efficiency of the line would be maximum for a company if all their systems are well aligned, and the entire unit should function in cohesion. Any unnecessary or unplanned stoppage of the main assembly line results in a substantial monetary loss.

The company presented in this study, ABC, has significantly invested in improving its digital solutions to monitor and control the process with the latest industry 4.0 solutions, i.e., smart tools, systems, and solutions. Although ABC has invested much money to access the third-party provider's intelligent solutions, such as tools and systems, the massive volume of data acquired from the tightening process of CC joints is not used proactively for developing insights or supporting decision making. The company is currently using the software solution 'BoltsNut' reactively after the incidence of the problem. Interestingly, many opined the reason behind their limited usage of the existing system. The other problem, ABC is not fully aware of the consequences of having a non-robust CC tightening process. This study will highlight the impact of poor data quality in their existing system. It will explain why an ML model is required and how it will positively impact the organization. A detailed explanation will be provided in the results chapter of the thesis.

1.4.1 Technical implications and Business implications

There are some business and technical implications that could be important to make ABC company aware of them in order to improve the robustness of the current CC joint tightening process. These implications are summarized in the following paragraphs.

Technical implications: ABC lost much time performing non-value-added activities due to the non-robust CC tightening process. These activities created much stress for the operators and indirectly impacted the productivity of the Quality, Production, Audit, Maintenance, and Procurement departments. The technical implication of a non-robust process is listed below:

- Less transparency and traceability of the CC tightening process due to the unreliability of the existing system.

- Unplanned stops in the mainline.

- A lot of non-value-added additional activities.

- Lots of reworks, repairs, and adjustments.

- Increased shopfloor employees' workload and resulted in stress and dissatisfaction.

- Load on existing systems and database servers due to poor data quality.

- No standardised way of working for pre-tightening and tightening of the bolt.

Business implications: The company loses money on not utilizing the installed system properly and thereby this can increase direct and indirect costs on fixing the poor data quality, prevention, detection, and repair. As the failure of CC joints results in impacting the safety of the customers directly, having a non-robust process results in:

- Impacting the vehicle functionality, trajectory, stability, and integrity due to CC joint failure.

- Loss due to stoppage of the mainline.

- Risking customer safety, trust, and loyalty to their brand.

- The reputation of their brand.

- Losing out business with competitors.

- Not utilizing the existing third-party systems efficiently.

- Incurring financial loss due to poor data quality and fixing them.

The company ABC has been experiencing these technical and business implications that impacted its economic, environmental, and social sustainability in the past and present.

1.5 CC joint and current state of CC bolt tightening process in ABC

ABC has some special characteristics, categorized as special characteristics (SC) and critical characteristics (CC), to protect the customers and their products from safety and compliance-related risk. These special characteristics are analyzed during the design phase using design/process failure modes and effects analysis (D/P-FMEA). SC refers to special characteristics that affect compliance with regulations, form, fit, function, and performance. CC are special characteristics that can result in impacting customer safety. The failure target rate for CC joints is zero, and the required process capability as per their internal standard should be greater than or equal to 1.67. CC joint failures create the risk of accident and injuries as they affect vehicle functionality, stability, trajectory, and integrity. Moreover, these CC joint failures can cause sudden vehicle stoppage and thermal events.

As in the current state, the company ABC wants to track real-time process monitoring of CC joints due to their high criticality. As mentioned before in the background, to achieve its digitalization goal, ABC has invested a substantial amount in a third-party solution provider who offers end-to-end support from tooling, software, data storage, and aid in calibration, repair, etc., maintenance of the CC tightening process. This implies that ABC performs CC joints with intelligent electrical equipment and stores data in a database due to its criticality. No data is currently tracked for non-CC joint tightening operations performed by pneumatic or hydraulic equipment.

The outcome of the tightening process is an Ok or Nok joint generally based on pre-set tightening parameters during tightening operation. The p-set (program set) is set for a particular tightening sequence based on batch count with different parameter values assigned to that specific p-set. Although the equipment can measure various tightening parameters like time, torque, angle, force, speed, cycle time, and run-down angle, ABC measures and stores only four parameters: time, torque, angle, and speed.

The tightening strategy used in ABC is torque controlled/angle monitoring strategy. When the final target torque P113 is reached, the tool is programmed in the p-set to shut off irrespective of other parameters. The angle is only for monitoring purposes. The assumption is that no other parameters except torque influence the joint's output status.

An Ok joint is termed an acceptable joint in the plant, and the system indicates green color in the Operator's digital display if the final target torque, P113, is achieved during tightening. In contrast, the Nok joint is termed as a not acceptable joint and needs rework to fix the joint to be Ok. The Nok joint is indicated by red color in the Operator's digital display if the final target torque is not reached during the tightening operation.

Figures 1.2 and 1.3 illustrate for the readers to understand more about the tightening process parameters and the outcome of the CC bolt tightening process, i.e., Ok/Nok status. P112 and P114 correspond to the final torque minimum and maximum range within which it will operate. P122 and P124 represent the final angle minimum and maximum values, respectively. The area between the minimum and maximum values represents the control limits within which those parameters can vary.

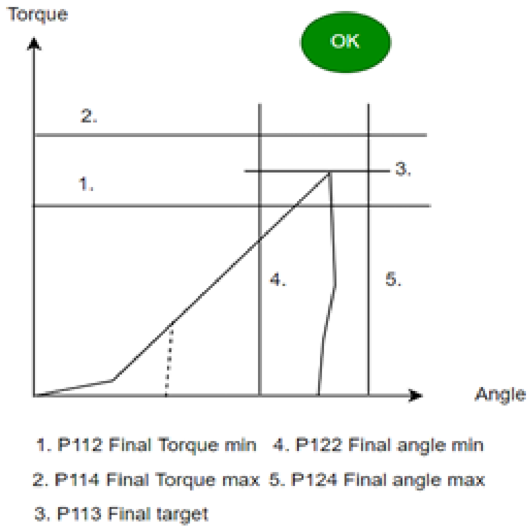


Figure 1.2: Ok status CC bolt tightening process (Torque controlled/Angle monitoring strategy).

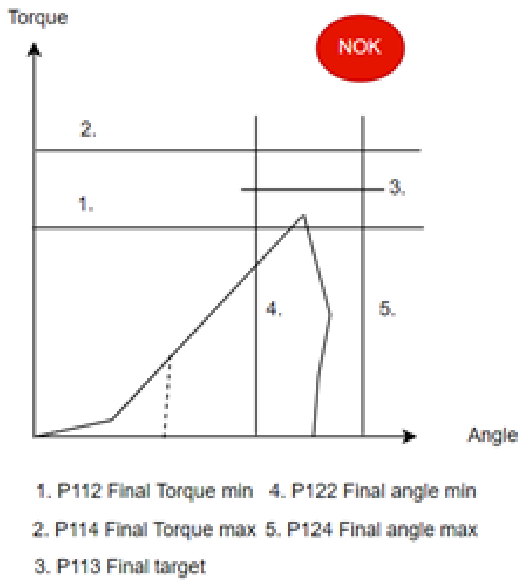


Figure 1.3: Nok status CC bolt tightening process (Torque controlled/Angle monitoring strategy).

Another common type of Nok occurs if the final torque value falls below the final torque minimum, P112, then the status is Nok due to low torque. In another case, if the final torque value goes above the final torque maximum limit, P114, then the status would still be Nok due to high torque.

The bolt tightening for CC joints in ABC is done in two steps. In the first step, 80 percent of the first target torque is achieved, and there will be a slight pause before the final step to counteract short-term relaxation efforts. The two-step is carried out in four stages, i.e., start, rundown, tightening, and stop phase (Refer to Figure 1.4). In the first two stages, the bolt will be pre-tightened up to a point where the actual torque builds on the bolt, and the tightening is completed in the third phase. This is a standard method in ABC for all CC bolted joints.

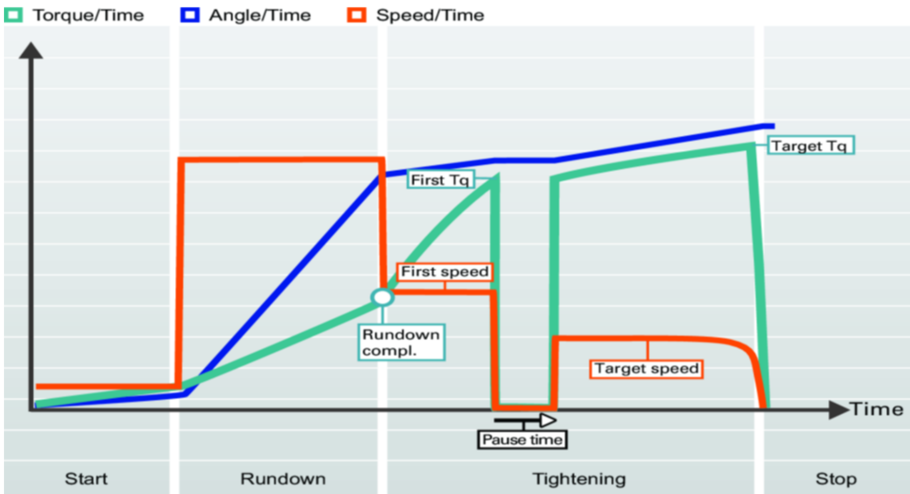


Figure 1.4: CC bolt tightening two step methodology

.

Analyzing, exploring big data, and translating the raw data into meaningful insights is a significant challenge for many companies working with industry 4.0 technologies (Cai & Zhu, 2015). The quality of highly accurate and reliable is a must for doing big data analysis as most of the companies invest a considerable sum of money on detecting, removing, and improving the poor quality of data for developing better insights (Cai & Zhu, 2015; Haug et al., 2011). The Data-Driven Approach (DDA) used in this study will become vital, as the decision can be based on facts with concrete data evidence to support the facts. This study will demonstrate how data analytics and visualization techniques can be used to explore and analyze the past historical data to find trends, solve the problem at present, and predict the outcome or provide recommendations on how to proceed forward to solve the problems in the future successfully.

1.6 Limitations

This study only focussed on tightening operations related to CC joints and did not include non-CC joint tightening operations. Due to the complex interlinked internal systems within the company and the large size of the plant, the study was limited to one workstation. Although, some of the results from this study – data quality errors and improving poor data quality can be generalized to other workstations. However, some of the results cannot be generalized to other workstations as there are a lot of factors involved in the process, i.e., different operations in individual workstations, the difference in tightening material, length of the bolt, size of nut, the impact of external factors - damaged thread, oil, grease on bolt surface or threads and human, equipment errors. As the data for the ML model is fed manually through extensive data pre-processing such as data cleaning, integration, and preparation, the sustainability of the model can be limited.

2

Literature Review

This chapter outlines the theoretical background that supports answering the research questions of this study. The literature review includes topics such as industry 4.0, data-driven smart manufacturing, digitalization in automotive production systems, AI including ML in manufacturing, inductive learning approaches, data mining approaches, CRoss-Industry Standard Process for Data Mining (CRISP-DM) application in manufacturing, data analytics, and visualization application in automotive industries, some background knowledge regarding bolted joints, previous research on bolted joints, and finally effect of bad quality on decision making.

2.1 Industry 4.0 Technology Evolution

The four major industrial revolutions are namely Mechanization (Industry 1.0), Electrification (Industry 2.0), Computerization (Industry 3.0) and Digitalization (Industry 4.0) (Pop, 2020). The evolution of these industrial revolutions over the years can be seen in Figure 2.1. Industry 1.0 is characterized by a mechanical production system with a hydraulic or pneumatic drive. It was followed by the assembly line production system, which facilitates mass production with electrical machines and machinery (industry 2.0). The introduction of production automation through information technology characterized Industry 3.0 with the use of computers, IT systems, and microprocessors (Pop, 2020).

Ebrahimi et al. (2019) state that in the past, companies implemented World Class Manufacturing (WCM) based on industry 3.0 in order to improve their production systems to remain competitive in the global market. However, the increase in customer demand for the customization of products forced manufacturing companies to be more flexible and agile. The major challenge with these previous revolutions was the absence of interconnection of technologies (Pop, 2020). Industry 4.0 has overcome these challenges with cyber-physical systems with mass connectivity and smart devices that collect and store real-time data.

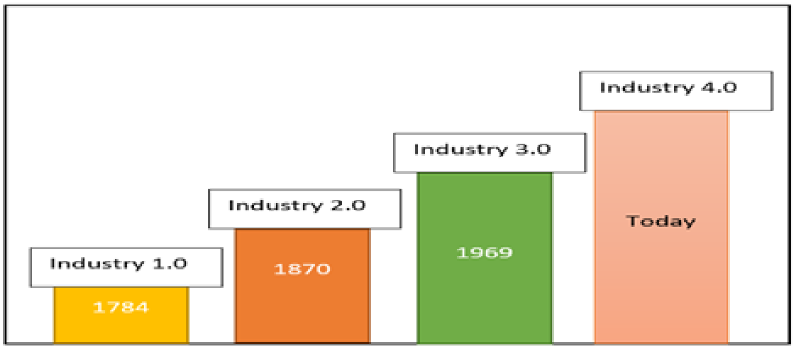


Figure 2.1: The four major industrial revolutions (Pop, 2020)

.

Frank, Dalenogare, and Ayala (2019) state that the term industry 4.0 was a German initiative created in 2011 by the federal government with universities and private companies. Industry 4.0 was a strategic program to develop advanced production systems to improve the productivity and efficiency of their domestic industries. This concept integrates a set of emerging and merged advancements that increase the value of the entire product lifecycle. Ebrahimi et al. (2019) argues that although the evolution of industry 4.0 or digital transformation with its access to extensive data and information has allowed these companies to make decentralized decisions with global vision and global profit optimization. Nevertheless, at the same time, there are numerous challenges of industry 4.0, and its objectives can take a long time to achieve its desired results (Ebrahimi et al., 2019).

In industry 4.0, the activities of the value chain will be carried out with smart approaches with information and communication technologies (Frank et al., 2019). On the other hand, Dalmarco, Ramalho, Barros, and Soares (2019) list the main opportunities and improvements as a result of the implementation of industry 4.0 by various industries in their study. The main challenges faced by the companies without industry 4.0 are low productivity, capacity, human resources, less efficiency in the value chain, and low quality of products and services. These lead to a lack of information security, insufficient knowledge in information, decision making, and a reactive mindset of taking corrective actions once the problem arises. Dalmarco et al. (2019) believe industry 4.0 will increase efficiency, flexibility, productivity, cybersecurity, and quality of products, services, quality decision-making process as it monitors, controls, and predicts the performance of the system in real-time.

2.2 Data Driven Smart Manufacturing

Cao (2017) regards data as the new oil that drives the future of science, technology, and economy. Many countries are trying to unlock the full potential of data DNA, which will promote data science research, innovation, policymaking, industrialization, and economy. Tao et al. (2018) opine that the volume of data collected at manufacturing stands at 1000 EB annually as a result of the latest technological advancements. Shao, Shin, and Jain (2014) states that the manufacturing companies accumulate a large amount of data from the shop floor due to these technological advancements. However, the major challenge is converting the raw data into meaningful insights to facilitate improved decision-making. Over the years, data in manufacturing has undergone major evolutions (Refer to Figure 2.2)

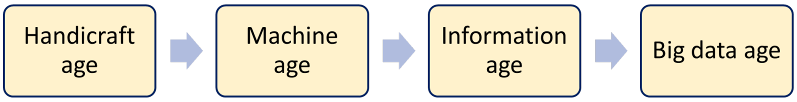


Figure 2.2: The evolution of manufacturing data in different timelines (Tao et al., 2018)

.

According to Tao et al. (2018), the advancements in IT such as IoT, cloud computing, mobile internet, and AI effectively support data acquisition, storage, off-site analysis, and processing and make timely decisions with minimal human involvement. The sensors in the manufacturing plant collect real-time data stored in cloud servers and are made easily accessible through mobile, web-based applications. AI processes the data for timely insights with minimal human involvement is an illustration of Smart manufacturing.

Cao (2017) lists the different types of analytics, namely descriptive, diagnostic, predictive, prescriptive, explicit, implicit, and deep analytics, which are widely used in different fields. Interestingly, data-driven manufacturing acts as a prerequisite for Smart manufacturing. The five stages of Data-Driven Smart Manufacturing (DDSM) are Integration, pre-processing, analysis, visualization, and collection. The different characteristics of DDSM are self-learning and self-adaption, self-regulation, self-execution, self-organization, and customer-centric. The applications of DDSM are in the field of smart equipment maintenance, product quality control, manufacturing process monitoring, material distribution, tracking, smart planning, and smart design applications (Tao et al., 2018).

2.3 Digitalization in Automotive Production Systems

Pop (2020) clarifies the difference between digitization and digitalization. The term digitization refers to transforming manual handwritten documents into an electronic format (Pop, 2020). In contrast, digitalization refers to converting a traditional business into a digital one with intelligent data-based systems which are very smart and autonomous (Pop, 2020). Peters, Chun, and Lanza (2015) list the numerous manufacturing challenges that many global automotive OEMs face, leading to a crisis, making the European market weak, slowing down the progress of the US, and China markets, inconsistency in the Indian market, decreasing market shares in Russian and South American markets. Additionally, the strict and dynamic environmental regulations to stay below 95g CO2 emission have further worsened OEMs. According to Pop (2020), modern technologies have aided in increased connectivity between humans, equipment, and the connected network of systems, creating opportunities for overcoming the challenges mentioned above and in successful applications.

Traub, Vögel, Sax, Streichert, and Härri (2018) indicate that autonomous systems as a part of the digitalization journey are becoming an integral part of current and future solutions for various industries, including automotive. As a result of digitalization, smart connectivity has enabled advancements in automotive systems such as automated driving, AI, drivetrain electrification, connected systems, and seamless, rich system integration to reduce system complexities. Pop (2020) illustrates how a Romanian automotive company could develop a Smart Production System (SPS) based on the digitalization of the production process, quality control, and logistics systems. The induction of SPS resulted in radical changes in digitalizing the existing process to a fully automated process to trace, monitor, and control the process.

ICT technologies have aided cars with autonomous driving and connectivity in all dimensions. In contrast, these had impacted the automotive production system with smart manufacturing/industry 4.0 techniques such as scalable model-mix lines, operational excellence, and assistance system for workers. Although the role of ICT in production and products is vital in the automotive industry, the ICT solution cannot commodify manufacturing knowledge that drives operational efficiency and innovation in various fields (Peters et al., 2015).

2.4 Artificial Intelligence (AI), ML in Manufacturing

The researchers and practitioners believe ML, a rapidly developing field, will solve old and new manufacturing challenges. The field of ML is successfully applied in various process optimization, monitoring, control, and predictive maintenance in various industries. The most exciting developments in ML include data mining (DM), AI, and knowledge discovery from databases (KD). However, the challenging part is the exploitation and adoption of some ML algorithms, theories, and methods in practical situations to solve problems as this field is many diverse (Wuest et al., 2016).

Industries have been vastly successful with the application of AI in operations, productions, and post-productions (Rai et al., 2021). Interestingly, around 40 percent of the potential value of analytics comes from AI and ML techniques, where ML accounts for 3.5 trillion to 5.8 trillion annually (in dollars). The use of analytics in manufacturing with the application of ML techniques introduces computer vision-based inspection and monitoring, fault detection, process improvement and optimization, cloud manufacturing, improved assembly-line efficiency, customer experience, inventory management, minimizing cost, reducing errors, increased visibility and life of assets (Rai et al., 2021) (Refer to Figure 2.3). Furthermore, the ML techniques and algorithms are structured as supervised, unsupervised, and reinforcement ML, which is represented in Figure 2.4.

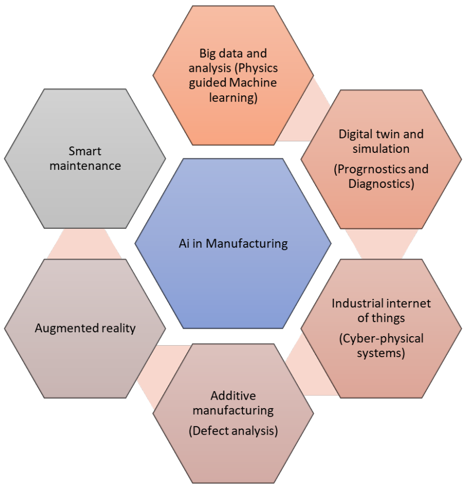


Figure 2.3: AI and ML applications in manufacturing (Rai et al., 2021)
.

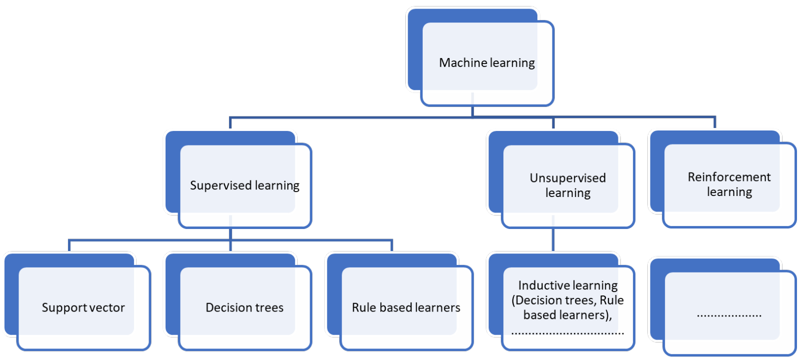


Figure 2.4: Structuring of ML techniques and algorithms (Wuest et al., 2016)
.

According to Kotsiantis, Zaharakis, Pintelas, et al. (2007) in the supervised ML, the instances are given with known labels. Singh, Thakur, and Sharma (2016) define supervised ML model as the construction of algorithms that can produce patterns or hypotheses using externally supplied instances to predict the future outcome. According to Wuest et al. (2016), supervised ML is mainly applied for problems where data is abundantly available, but knowledge of the process is scarce. The general steps in the supervised ML are splitting the data set into training and test data and then pre-processing the data before modeling (Wuest et al., 2016).

In contrast, Kotsiantis (2013) states that unsupervised learning has instances that are unlabeled. According to Wuest et al. (2016) in the unsupervised method, the ML process tries to learn structure in the absence of a pre-defined path and no primary outcomes. In Reinforced Learning (RL), the learner will not be provided with information regarding which actions to take but should discover which action leads to the best output by trying every time (Kotsiantis et al., 2007). Wuest et al. (2016) underlines the differences between supervised ML and reinforcement ML as the problems in RL can be described without ‘good’ and ‘bad’ labels, and there is no need for a knowledgeable supervisor as it works on sequential environment response which emulates how humans learn.

2.5 Inductive Learning in Supervised Technique: Rule-based Learners and Decision Trees

De Mantaras and Armengol (1998) describe inductive learning, which is a part of supervised and unsupervised ML technique as an instance-based or case-based reasoning used to capture general knowledge from examples to solve complex problems. According to Maimon and Rokach (2005), researchers among various fields: statistics, ML, pattern recognition, and data mining believe decision trees are the most popular approach for representing classifiers. Decision trees and Rule-based learners constitute an inductive-based learning approach (Wuest et al., 2016). Inductive methods are classified into supervised/unsupervised learning, single/multiple concept learning, and proportional/relational learners (De Mantaras & Armengol, 1998). The goal is to formulate rules based on facts for effectively predicting unseen facts. The challenge is that this method works very well for already provided or trained examples, but the same cannot be expected from unseen examples. They can be used in two ways: as communicative tools capturing critical information from examples or as a part of learning systems. The user can effectively control the usage of this method by providing familiar examples. As a part of the learning system, the method can generate feedback to another system component from positive or negative examples to complete the desired task (De Mantaras & Armengol, 1998).

Quinlan (1990) explains that decision trees can extract descriptive decision-making knowledge from data. Decision tree classification rules employ a top-down, divide-and-conquer strategy that divides objects into smaller and smaller branches or nodes in each step when the tree grows (Refer to Figure 2.5). These smaller subsets are called attributes. The tree-based methods are relevant in providing clarity, conciseness, context-sensitivity, and flexibility. Maimon and Rokach (2005) state that decision tree consists of nodes from a rooted tree. The root has no input branches, but all other nodes that sub-divides from the root have exactly one input. A test or internal node consists of an outgoing edge.

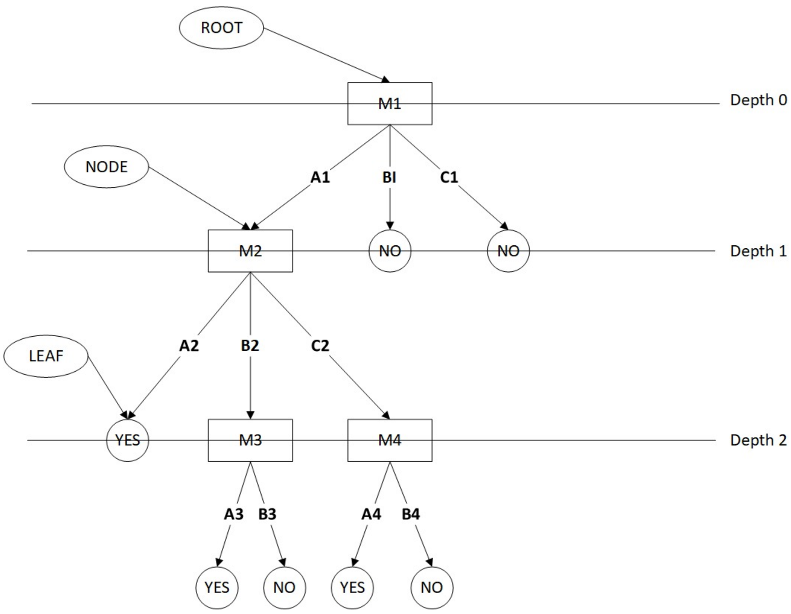


Figure 2.5: Generic tree adapted (Kotsiantis, 2013; Safavian & Landgrebe, 1991)
.

Kotsiantis et al. (2007) state that decision trees are sequential models which combine simple tests that compare a numeric attribute against a threshold value. They tend to provide them an advantage over black-box methods due to their simple nature, comprehensibility and close resemblance of human reasoning. In some cases, if the decision tree is built with more irrelevant features, they tend to provide good accuracy with training data but work poorly with the new data resulting in data over-fitting. Pruning and data pre-processing avoid over-fitting data and effectively control decision tree complexity. Pruning is further divided into pre-pruning and post-pruning. Pre-pruning is used to terminate some of the branches before they develop prematurely. Post-pruning removes some of the developed branches from the generated decision tree to improve the classification accuracy for future predictions. The data pre-processing technique includes feature extraction that helps in finding an optimal number of critical parameters in order to build a simpler decision tree (Kotsiantis et al., 2007).

2.6 Data Mining Approaches in Manufacturing Applications

Wuest et al. (2016) describe the tasks in data mining namely classification, clustering, association rules and feature selection (Refer to Figure 2.6). Data mining is a solution for effectively transforming raw data into useful knowledge. The extracted knowledge is used to find relationships, patterns, model, classify, and to making new predictions (Harding, Shahbaz, & Kusiak, 2006). Wirth and Hipp (2000) believe that the individual or team should have special competence and knowledge in order to carry out data mining projects successfully. Harding et al. (2006) believe that knowledge is the most critical asset for any manufacturing company as it provides them a clear edge over their competitors, thus enabling them to optimize their efficiency and in-house abilities. Although the knowledge exists in different forms, the challenge for the companies is to capture and convert the knowledge into data and information.

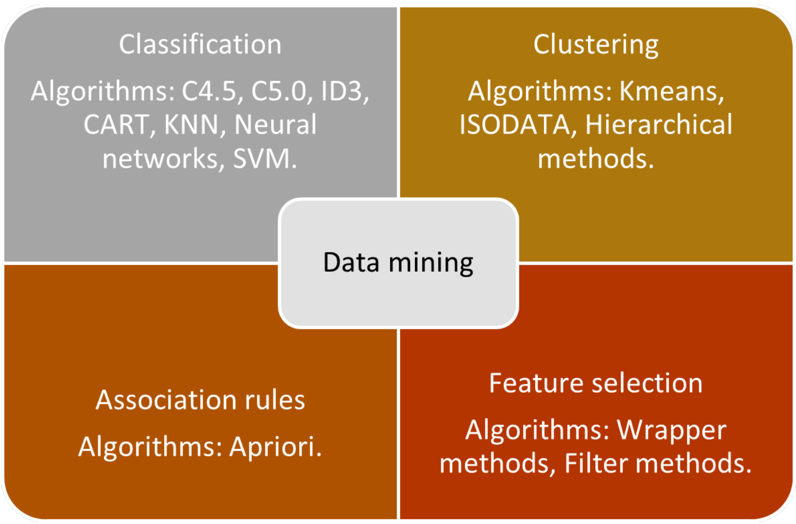


Figure 2.6: Overview of tasks and algorithms in DM (Wuest et al., 2016)

.

Azevedo and Santos (2008) similarly state that CRISP-DM (Cross Industry Standard Process for Data Mining) and SEMMA (Sample Explore, Modify, Model and Assess) are established standards for data mining projects as they follow a set of sequential steps for implementation. Harding et al. (2006) state that data mining makes use of ML algorithms, statistics, AI, and data management. Due to its high significance, data mining is applied in various engineering applications such as banking, finance, retail, marketing, insurance, fraud detection, and science. In manufacturing, it is extensively used for predictive maintenance, fault detection, design, production, quality assurance, scheduling, decision support systems, and controlling manufacturing processes. Azevedo and Santos (2008) compare KDD, SEMMA, and CRISP-DM and finds that SEMMA is similar to KDD in terms of practical implementation. In contrast, KDD cannot be compared straightforwardly with CRISP-DM. Interestingly, Azevedo and Santos (2008) conclude that SEMMA and CRISP-DM are implementations of the KDD process. On the contrary, Shafique and Qaiser (2014) compares these three models and concludes that most researchers follow the KDD model since it is complete and accurate, but CRISP-DM and SEMMA are more company oriented.

2.7 CRISP-DM and its Application in Manufacturing

Azevedo and Santos (2008) states the different steps in CRISP-DM model which is represented in Figure 2.7. Wirth and Hipp (2000) identifies CRISP-DM as a standard model for carrying out data mining projects because this model is not biased towards industry or technology. The CRISP-DM model can efficiently handle large Data Mining (DM) projects as more reliable, repeatable, and manageable.

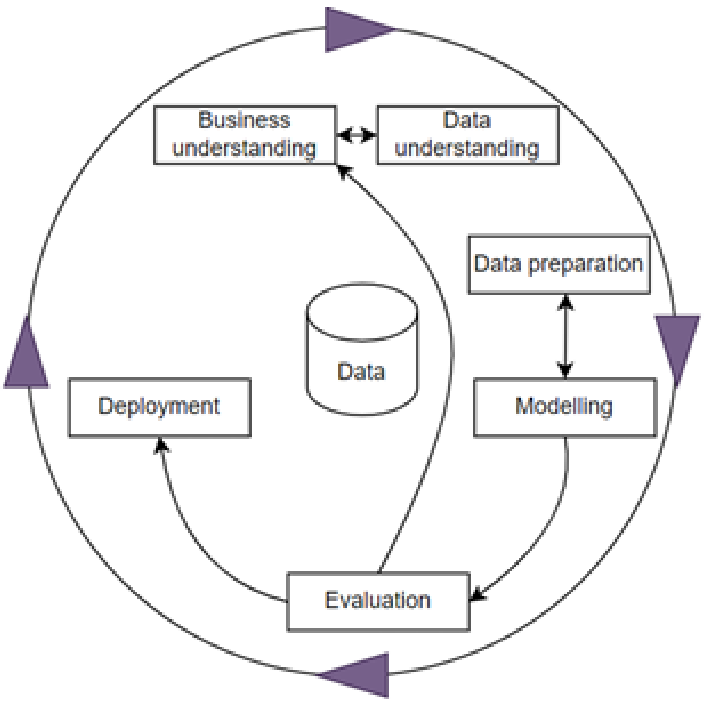


Figure 2.7: The CRISP-DM life cycle (Azevedo & Santos, 2008)

.

Wirth and Hipp (2000) identifies the model to provide an overview of the life cycle of a data mining project that includes the phases, tasks, and corresponding outputs of the project. Customers will have a clear picture of the pitfalls of DM and the expected outcomes. The other advantage of CRISP-DM is that it links different tools and diverse sectors in effectively communicating and documenting the results.

Schröer, Kruse, and Gómez (2021) describe different domains in which CRISP-DM is applied extensively based on use cases. The domains include health, education and research, engineering and manufacturing, government and public sector, management, information technology, food, finance, and entertainment. The application of model is widely recognized in health and education domains (Schröer et al., 2021). Schäfer, Zeiselmair, Becker, and Otten (2018) illustrate how CRISP-DM has been successfully applied in an electronics production industry to develop a predictive model for error forecasting. Schröer et al. (2021) explain the successful application of CRISP-DM in cancer diagnostics using a classification model and in lithium battery production to classify batteries based on the produced quality. Ribeiro et al. (2020) applied CRISP-DM to predict the tear strength of woven fabrics by automated ML techniques. There are some master thesis projects by Lené and Rajashekarappa (2021) and Vasudevan and Duan (2021) that show the implementation of CRISP-DM in manufacturing. The study by Lené and Rajashekarappa (2021) focused on detection of air leakeage in a pneumatic system and the Vasudevan and Duan (2021) study concentrated on reducing industrial alarms by predictive maintenance application.

Due to the simple nature, reliability, and faster implementation time and flexibility, even non-experienced professionals in the field of data mining prefer to use CRISP-DM over SEMMA (Harding et al., 2006). Schröer et al. (2021) describe CRISP-DM as the chosen de-facto standard for applying a process model in most data mining projects because of its structured, reliable, simple, and industry-independent approach. The deployment phase seems to be missing in some companies. This model has some drawbacks and limitations that result in enhancements or adaptions of the CRISP-DM model. Martínez-Plumed et al. (2019) argue that CRISP-DM was the de-facto standard process model for data mining projects and knowledge discovery projects until 2016. The CRISP-DM model is based on a goal-oriented approach, whereas data science tends to be more exploratory (Martínez-Plumed et al., 2019). Since 2016, there has been a drastic shift in trend from goal-oriented process to exploratory process. Although the modeling phase has been automated, the deployment and data wrangling phase has not been automated.

2.8 Data Analytics and Visualization Techniques Application in Automotive Industries

As discussed in the previous sections, advanced technologies have enabled automotive manufacturing companies to acquire and store large amounts of data effectively. The manufacturing data are characterized by high volume, variety, velocity, veracity, and value (Tao et al., 2018). Some of the companies were successful as they could develop valuable insights using data science techniques for improved decision-making using advanced data analytics like ML techniques. This section elaborates on applying some of the relevant data analytics and visualization techniques related to this study for carrying out process improvement projects in the automotive sector.
Statistical Process Control: According to Sousa, Rodrigues, and Nunes (2018), SPC is a method that combines control charts and statistical methods to monitor, control the process variations, and predict the capability of the process. SPC involves a feedback loop that will be used to improve the control parameters. According to Godina, Matias, Azevedo, et al. (2016), SPC consists of a set of tools used for process management and as a strategy for improving capability through the reduction of variability of products, process, materials, deliveries, and equipment. With

SPC, companies can remain competitive by producing quality products at low cost, enabling quality improvement and waste elimination. Sousa et al. (2018) term SPC as a much prevalent technique in automotive mass production companies where it is critical to assess the process capability assessment constantly. SPC carries out process capability assessment. Process capability, according to Burgess (2018) is defined as the performance of the process operating under statistical control. Control limits are the range within which we can regulate the process.

Godina et al. (2016) defines process stability, i.e., a process is said to be stable in statistical control if the probability distribution is constant over time. Control charts assess process stability. If a change is observed in the distribution over time represents the special cause of variation that needs to be identified and eliminated to keep the process under statistical control.

According to Burgess (2018), the quality metrics for process capability and performance are Cp (Process capability), Cpk (Process capability index), Pp (Process Performance), Ppk (Process performance index). The Cp and Pp are of theoretical relevance, and Cpk and Ppk are of practical relevance. Cpk is defined as short-term Key Performance Indicators (KPIs), whereas Ppk is long-term KPIs. Figure 2.8 and Figure 2.9 show a normal distribution of a stable, capable process under statistical control which can be seen in the control chart as all the points control limit and specification limits.

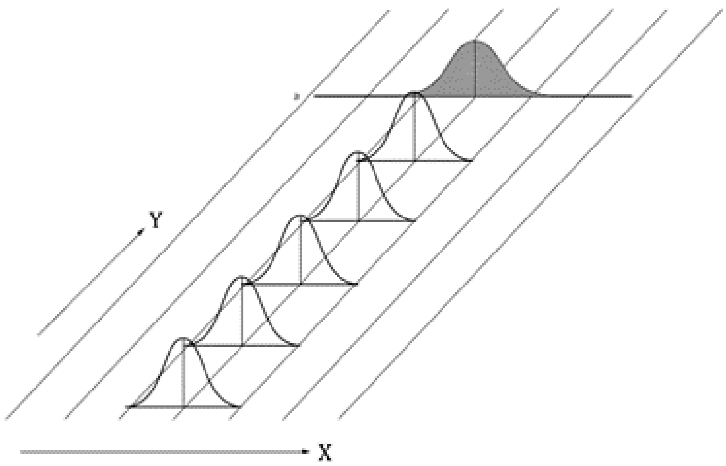


Figure 2.8: Illustration of a process under statistical control, adapted from (Burgess, 2018)

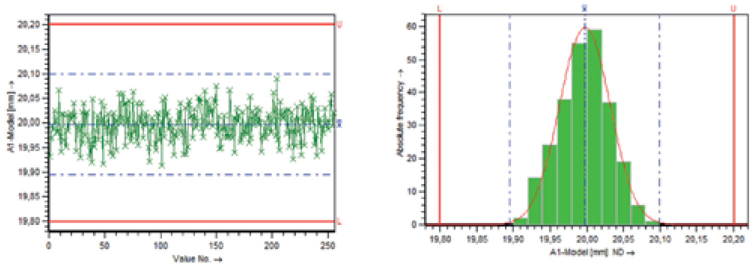


Figure 2.9: Illustration of a control chart (left side) and histogram model (right side) for stable process, adapted from (Burgess, 2018).

On the other hand, Figure 2.10 and Figure 2.11 show that the distribution over time is not constant, i.e., the process is not stable but is still capable of producing most of the products within limits. This process will create defects in the long run. According to Osborne and Overbay (2004), the availability of outliers produces inflated error rates and cause substantial distortions of statistical results.

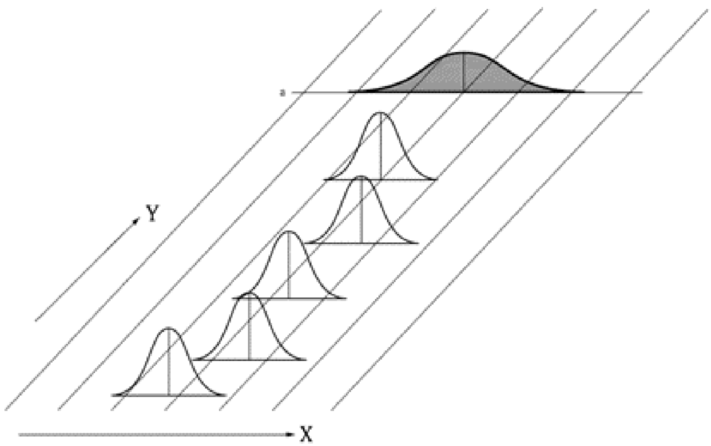


Figure 2.10: Illustration of a process not under statistical control, adapted from (Burgess, 2018)

.

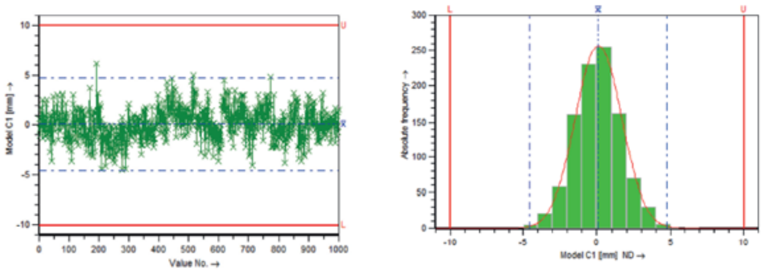


Figure 2.11: Illustration of a control chart (left side) and histogram model (right side) for un-stable process, adapted from Burgess (2018)

.

Cepeda and Lopes (2019) term visual management system as a robust communication channel between departments as it only visualizes relevant information. Visual management is an integral part of lean manufacturing methodology as it highlights the critical problems to be addressed. According to Sedrakyan, Mannens, and Verbert (2019), the objectives of visualization are comparison, relationship, distribution, the trend over time, and composition. The reasons for visualization are exploration, discovery, summary, and present. Smith, Coleman, Bacardit, and Coxon (2019) identify that the automotive sector produces large sets of informative data and emphasizes the need to do analytics and visualization on the data to identify hidden patterns, potential opportunities for developing insights with the data. Sajid et al. (2021) lists the different tools and techniques used for a visualization based on application. With EDA, business intelligence tools like Power BI, and advanced excel, it is possible to find trends in data, detect anomalies or outliers, and find the shape of the distribution, i.e., normal or non-normal distribution and statistical parameters like mean, mode, median, standard deviation, and variance. According to Smith et al. (2019), the data quality is critical for developing insights as missing values, incomplete data, unbalanced data can create deceptive results. Tavera Romero, Ortiz, Khalaf, and Ríos Prado (2021) highlight the need for having harmonization between product and intelligent data technology from industry 4.0 technologies.

The application of Power BI technology in the automotive industry has brought a positive shift at the economic and business levels for improved company decision-making and also supports developing dashboards (Tavera Romero et al., 2021). The dashboard is an intelligent interactive tool that visualizes KPIs related to organizational goals, and strategies (Cepeda & Lopes, 2019). At the same time, the dashboard allows monitoring and distribution of this critical information to all levels of management, thus supporting tasks and enabling decision making.

2.9 Basics of Bolted joints

Bolt joint is widely applied in a variety of industrial machines due to its simple configuration, convenient operation, and fewer cost (Chen et al., 2017). Figure 2.12

represent bolted screw thread. Bolts are essential to machine parts and the most common type of joint in the automotive sector (Monville, 2016).

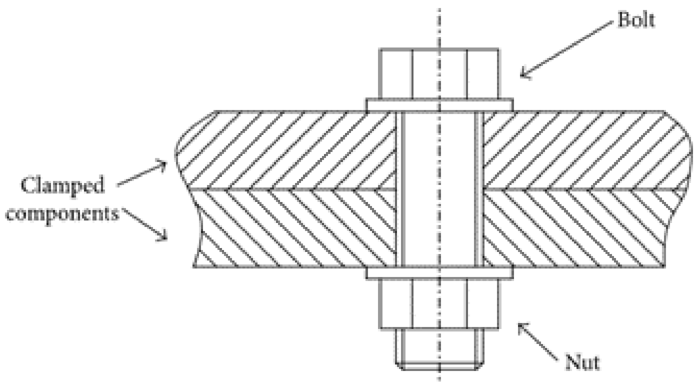


Figure 2.12: Basics of bolted joints (Chen et al., 2017)

.

The preload for the bolt joints should be high for the bolt to create a transmission force between the clamped components (Chen et al., 2017). However, when performing a bolted joint, the load should not be too low or too high and equally distributed among the bolts. If not done correctly, it will create ductile rupture, thus affecting the performance of the assembly operation and, in the long run, causing fatigue failure of bolt (Monville, 2016). The parameters involved in the bolted joint are torque, angle, speed, cycle time, and run-down angle (pre-tightening load) for a particular tightening sequence. Torque is the product of force exerted on the bolt and the corresponding rotation of the bolt in degrees (angle). There is little research on the torque-controlled/angle monitoring strategy for bolted joints. Interestingly, there is one master thesis by CERVANTES and Arman (2018) related to this study performed in a different company. A majority of the research is found on the bolt tightening process by finite element analysis and simulation-based studies (Fukuoka & Takaki, 2003; Piscan, Predincea, & Pop, 2010; J. Kim, Yoon, & Kang, 2007; McCarthy & McCarthy, 2003; Gray & McCarthy, 2010).

2.10 Effect of Bad quality on Decision making

The amount of big global data acquired by industry 4.0, IoT, and cloud is increasing exponentially from GB, TB, PB, EB to ZBs, but the analysis of big data must be done on accurate, and high-quality data for developing valuable insights (Cai & Zhu, 2015). The most critical challenge is to detect and filter erroneous data entering into the CPSs built on constrained resources that could not handle heavy computations and communications during their operation (Sha & Zeadally, 2015). In a real-time decision-making environment, time is a critical factor as decisions need to be within milliseconds, leaving no room for identifying and fixing poor-quality data before making decisions (Rego, 2020). Data from sensors can become corrupt and delayed leading to deadly consequences of the decisions made. Most of the sensors are not smart enough to actively participate in decision-making other than sensing data and recording it.

According to a research and industry report, vast amounts of investments are spent on improving data quality. The poor data impacts the organization's operations (i.e., creating customer dissatisfaction), tactics (impacts decision making and trust), and the organization's overall strategy (Laranjeiro, Soydemir, & Bernardino, 2015). An organization's database data quality problem has accounted for 50 to 80 percent of inaccurate, incomplete data (Strong, Lee, & Wang, 1997). Poor data quality negatively impacts business users, organizational culture, and operational costs. There are direct and indirect costs caused by the low quality of data. Moreover, there are other additional costs such as prevention, detection, and repair costs for improving or assuring the quality of data (Haug et al., 2011). IBM's estimate of poor-quality data annual cost in 2016 in the US region is 3.1 trillion dollars as decision-makers, managers, knowledge workers, and data scientists must accommodate bad data in everyday work, which costs money and time. Improving the data quality provides many benefits beyond reduced costs (Redman, 2016). The quality criteria for good data are being available, usable, reliable, relevant, and in a presentable way (Cai & Zhu, 2015).

3
Methodology

This chapter elaborates on the methodology used to conduct this study. As the project deals with analyzing a lot of quantitative data through exploration and modeling, the CRISP-DM, the ‘de facto’ standard and process model for most industrial data mining projects, is the’ de facto’ standard and process model chosen for this study. The CRISP-DM model used in this study is independent of industry technology and is a simple, reliable, and structured approach well suited for planning and managing the project efficiently (Wirth & Hipp, 2000; Schröer et al., 2021; Martínez-Plumed et al., 2019)).

The company provided the quantitative data analyzed in the study as historical databases and live internal systems were used. This study does not examine how the company acquires, collects, and stores the data but instead focuses on developing insights from the data acquired. The CRISP-DM model includes six phases: Business understanding, Data understanding, Data preparation, Modelling, Evaluation, and Deployment. The phases are flexible (the steps need not be sequentially followed). They can be iterative to ensure the business objectives and the project’s success criteria are met. The different phases of CRISP-DM are elaborated on below. Figure 3.1 explains the different phases and tasks of CRISP-DM.

3.1 Business Understanding

This is the primary and most critical step in determining the success of the data mining projects. This step involves identifying data mining project goals and objectives, formulating a clear problem description, developing success criteria, and a project plan with a timeline to achieve the goals. The initial project goals identified and analyzed will be presented to the stakeholders to collect feedback from the top management before making decisions. Once finalized, these project goals are transformed into well-formulated research questions, and a project plan is developed as a guide to achieving the desired results within the timeline. According to Wirth and Hipp (2000), there is a close link between business understanding and data understanding. Formulating a data mining problem and project plan requires at least a limited knowledge of the data. In this study, observations on the shop floor through Gemba walks and qualitative interviews with industry experts are performed to understand the technical and business implications of the problem.

3.2 Data Understanding

This includes initial data collection, description, data exploration, and quality verification. This step involves understanding the data by exploring the data to find patterns or trends or to identify exciting subsets to form new hypotheses with the data. In some cases, Statistical analysis might be carried out to gather concrete facts. The data quality is verified by checking the completeness, distribution of data, usability, availability, timeliness, and relevance concerning the problem at hand (Cai & Zhu, 2015). The source of study for this thesis work provided by the company is a replica of the original SQL live database that contains historical data. Initially, an SPC study was conducted using process capability analysis and verifying the control charts. It was followed by exploratory data analysis (EDA) and identifying data quality errors in their existing system through a data quality report.

3.2.1 Exploratory Data Analysis

EDA is defined as the numerical detective work or graphical detective work where a researcher examines the data without making assumptions to identify what the data tries to convey concerning the problem considered (Martinez, Martinez, & Solka, 2017). EDA is a graphical technique where the data is visualized, plotted, and manipulated without any preconceived notions to assess the quality of the data (Critical Data, 2016). EDA is an iterative process where the analysts constantly ask questions, and correspondingly a design will be created to extract knowledge from the data (Mao et al., 2015). EDA is used to search for patterns or trends in the given data set where visualization plays a significant part in the visual analysis (Skiena, 2017). The visualization tools used in this study are Minitab, Microsoft Power BI, and MATLAB. Minitab was used for data analysis SPC studies; Power BI was used for developing visualization plots and dashboards, and MATLAB was used for advanced data analytics as well as for visualizations.

3.3 Data Preparation

This phase includes pre-processing raw unstructured data into structured data to feed into the model for further analysis. This phase contains different steps such as data selection, data cleaning, data construction, and data integration. The article by Schröer et al. (2021), state that the inclusion and exclusion criteria should be defined for data selection effectively. Data reduction, filtering, and extraction of features would be carried out during data pre-processing. This phase is critical as the model's performance heavily relies on the quality of data prepared in this phase. The SQL database did not contain values for the individual CC bolt tightening sequences for this study. These values were downloaded manually from their existing live 'BoltsNut' system, and the data was prepared using pre-processing techniques to be suitable for creating the model. The prepared data was in the form of a timetable that represents other factors such as torque, angle, and speed with time.

It was followed by the model building covered in the next phase.

3.4 Modelling

This phase involves selecting the modeling technique, generating a test design, building the model, and assessing the built model. In practical, 70 percent of the data is regarded as training data, 20 percent data for evaluation in case of bias, and finally, 10 percent data for testing (Wuest et al., 2016). For building the ML model in this study, the prepared dataset is split into 70 percent training data and 30 percent test data. It was followed by feature extraction.

3.4.1 Feature Extraction

After splitting the data into training and test datasets, the training data's time and frequency domain features were extracted, sorted, and ranked by the Diagnostic Feature Designer (DFD) application in MATLAB. In this study, the top-ranked six features were exported to the Classification Learner Application (CLA) in MATLAB for further analysis.

After extracting the top six features by DFD for the training data set, these steps are repeated for the test data. Using CLA, the top six extracted features of the training data were trained with all the ML algorithms in this study. Once the model is introduced, the model with the highest validation accuracy for training data is selected, and the corresponding confusion matrix will be analyzed in the following steps. The confusion matrix for the model chosen (highest validation accuracy for training data) is investigated. If the model predicted class vs. actual values is close, then the model is exported. The exported model is validated by importing the test data and checking the validation accuracy of the test data. The model will be further cross-validated with new data set.

3.5 Evaluation

This phase evaluates results and reviews the process. The built model is considered with the project objectives and success criteria. This phase aims to assess if any other things need to be added to improve the model's value for the company. The entire steps can be repeated if the model does not meet our project objectives or success criteria. A new process will originate from the Data understanding phase until the desired results are achieved. In this study, the results were evaluated and presented to the company. Further actions were defined in the next stage.

3.6 Deployment

This phase involves creating a deployment plan to monitor and maintain the results and document the reports for future work. In this study, the deployment phase is not carried out as the database provided by the company is not a live database. The recommendations for the company were provided on how to extend this study to be sustainable for the company.

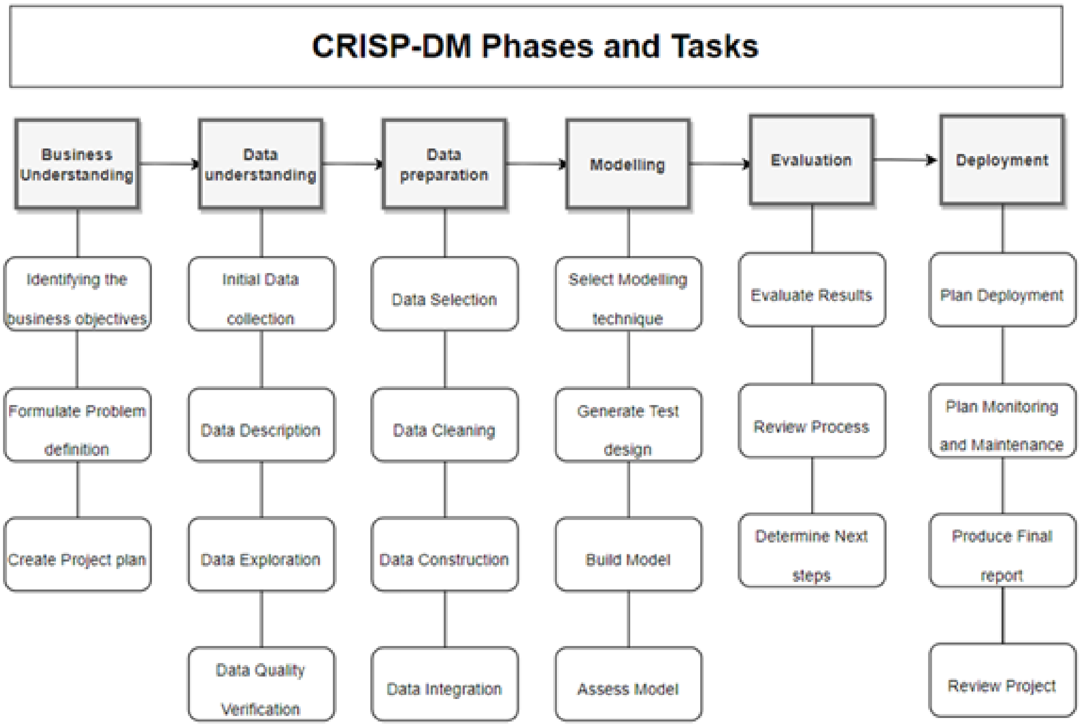


Figure 3.1: CRISP-DM phases and tasks, adapted from (Wirth & Hipp, 2000)
.

4

Results

This chapter will explain the findings and results in the six phases mentioned above of CRISP-DM in detail. The first two stages business understanding and data understanding rely heavily on identifying and highlighting the effect of poor quality of data in their existing BoltsNut system and explaining why a ML model must be integrated or have this as an individual solution for solving the problem. The other phases focus on ML model building, testing, and validation of the built model. The last section of results include a dashboard for the manual actions/backup routines followed when non-CC joint equipment powered by pneumatic is used to tighten the CC bolt, and the values are verified by a manual torque wrench occasionally. The benefit of the ML model results and how this dashboard can be helpful to the company are elaborated on.

4.1 Business Understanding

As understanding the business goals and defining the problem objective is the initial step in CRISP-DM methodology, the technical and business implications were understood from the insights obtained by performing visual observations on the shop floor by Gemba, conducting qualitative interviews from industry experts at all levels of management to have a deeper understanding of the problem. The Gemba and interviews aim to understand more about the impact of the non-robust CC bolt tightening process and the reason for the limited usage of their existing BoltsNut system. From the Gemba and interviews, it was evident that the company was using the current BoltsNut system as a reactive measure only when a problem occurs. Additionally, some employees in the company were unaware of the negative consequences they were facing due to the current non-robust process. The reasons for their limited usage of their existing system were numerous. From their business objectives, it was evident that an analytics and a visualization solution were needed to solve some of their current issues (Refer to Figure 4.1).

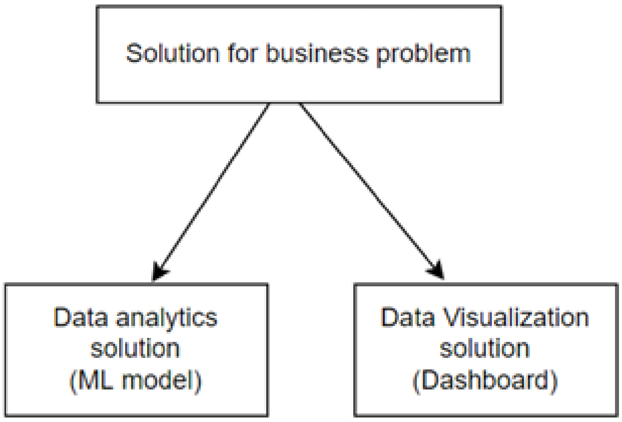


Figure 4.1: Data analytics and visualization solution.

The ML model will be the advanced analytics solution, and the Ok-adjusted Manual actions dashboard will be the visualization solution. The Ok-adjusted is the operation in which the operators and team leaders will carry out a backup routine for the tightening operation using non-CC equipment and occasionally verifying the torque values with a manual torque wrench if they cannot complete the joint with the smart tool. The success criteria are based on technical and business implications, explained in section 1.4.1. Some figures illustrated in this section may contain blurred information due to their confidentiality.

Success criteria: To develop a robust process for improved traceability of the CC tightening process with an ML model and a dashboard to efficiently utilize resources and reduce the load on their internal servers.

Responses from **interviews and Gemba observations** are listed below:

- Not many are trained in the BoltsNut solution.

- The search process in the BoltsNut solution was consuming a lot of time and did not yield desired results.

- There is no authorized person for this solution. Only a few individuals from various departments know this solution, but they don't use it often.

- Some mentioned the unreliability of data.

- There is no proper setup for parameters related to the tightening process. Need to do on a trial-and-error basis after the occurrence of the problem.

The high number of Nok joints was not considered critical by some of the employees because of the following reasons:

- Not impacting First Time Through (FTT).
- No major claims or campaigns due to CC joint failure.
- Nok does not stop the main production line.

FTT is a primary quality KPI in the company ABC, defined as the 100 percent error-free outcome of the process at the first occurrence. In the plant, FTT was not measured at the individual workstation level but instead measured on an overall plant level when the truck reaches the EOL, i.e., Quality gate Q6. However, if these Nok joints are rectified at the workstation level before reaching the EOL, the Nok joints won't impact FTT. The statements from all levels of management are provided below. The smaller number of claims and campaigns was a factor in not taking this problem seriously. Additionally, the operators were found to be overriding the system; if they could not be able to complete the CC joint before the truck left the workstation, the sensor would send a trigger to stop the main production line. These were from different industry experts regarding the practical problems.

'We are not measuring residual torque regularly in quality gates' – Manager.

'We are losing a lot of time in doing non-value-added activities. There was an accident due to CC joint failure resulting from manual actions' – Manager.

'These special tools used for CC joints cost four times higher than normal non-CC ones, but they are not handled effectively' – Procurement.

'There is a lot of stress and added workload to rectify a Nok joint' – Operator.

'It is a time-consuming process as the operator calls for Andon and I need to come and do the backup routine procedure if the tool, the system does not function' – Team leader.

'Although we have smart equipment in place, manual actions/backup procedures' tightening has often become the norm' – Production Engineer.

'The greater number of attempts made by an operator using the smart tool reduces the tool's lifespan in the long run' – Procurement team and an Engineer.

'We are not storing the final torque results which we collect during CC audits' – Internal Auditor.

'We could not stick to the plan due to limited resources and diverse CC joint operations in many plant areas' – Audit team.

Technical implications: By looking at the statements of industry experts, ABC was losing a lot of time in performing non-value-added activities resulting from the non-robust CC tightening process. These activities created a lot of stress for the operators and indirectly impacted the productivity of the Quality, Production, Audit, Maintenance, and Procurement departments. The results of the current non-robust process for the study's end-users, i.e., Quality, Production, Audit, Maintenance, and Procurement team, provide the readers a better understanding of the challenges faced by each team.

Quality team - Loses time in deviation case creation and implements quality control measures.

Production team - A lot of time was lost for rework and performing backup routines in case of tool malfunction.

Audit team - They could not stick to their control plan for inspecting CC joints due to problems in diverse areas. The values of CC joints noted during audits are not digitalized, making it difficult to find any trends over time.

Maintenance team - More frequent service and maintenance of equipment were done.

Procurement team - Due to errors caused by humans i.e., releasing the tool's trigger ealry or late, with not proper handling, incorrect orientation of tool or as a result of system i.e., due to malfunction of controller, smart tools, internal systems or display screen, there will be Nok joints. The more significant number of attempts by the operators to perform the correct tightening operation which is to produce a Ok joint creates load on tools and impacts tool's EOL in the long run.

Business implications: There were some instances of accidents that had happened due to CC joint failure. These were 'x' campaigns in which around 'n' chassis were affected due to CC joint failures, and the company had lost a sum of 'xk' SEK. These failures affect ABC financially and affect the image of the company and result in negative brand value. The other two campaigns were due to CC joint failure, but there was not much-supporting data. Although the number of campaigns related to CC joint failure in the last few years is less, these can drastically impact the company's reputation. Moreover, ABC has invested a lot in purchasing 'm' number of licenses for using the smart tools and BoltsNut, but they are not utilizing them effectively. The values of 'x,' 'n,' 'Xk,' and 'm' were anonymized due to confidentiality.

4.2 Data Understanding

The data was collected from two sources: one is from the SQL replica (not a live) database provided by the company, and the other is from the existing BoltsNut live database. The historical SQL database was explored to develop insights into the data. The input for the ML model was taken from the BoltsNut live database as the historical database does not contain graph values for individual CC joint tightening sequences.

Since the plant was large with many workstations and several interlinked complex systems, this study focuses on analyzing one workstation in the Cabtrim area. The joint performed in the workstation is a CC joint. The selected workstation was producing dashboards for the truck driver's cabin. The main reason behind choosing this workstation is that it shows a more significant number of Nok joints percentage than Ok joints percentage for the CC bolt tightening joints performed (Refer to Table 4.1). The trend was the same when conducting an EDA for last year and this year. Interestingly, when looking at the trend per day status in that workstation, it was noted that there had existed a trend where Nok joint percentage had been greater than 50 percent daily for the past six months. Figures 4.2 and 4.3 were made for illustrative purposes to show the Nok trend percentage for the months March and February 2022 using Power BI. In Appendix A, the trend of Nok percentage for the other months are provided (October 2021 to January 2022). Furthermore, after looking at the results, the reliability of the data in the BoltsNut system was investigated and analysed in the further sections due to the reason that smart tool is designed to function correctly.

Table 4.1: Overall Status of the joint in the specific timeline in the chosen workstation (Illustrative purpose).

Timeline	Ok-percentage	Nok-percentage	Total joints
Last year(2021), 12months	45.65	54.35	14830
This year(2022), January to April	46.12	53.88	5913

Figure 4.2: Nok trend per day for March 2022.

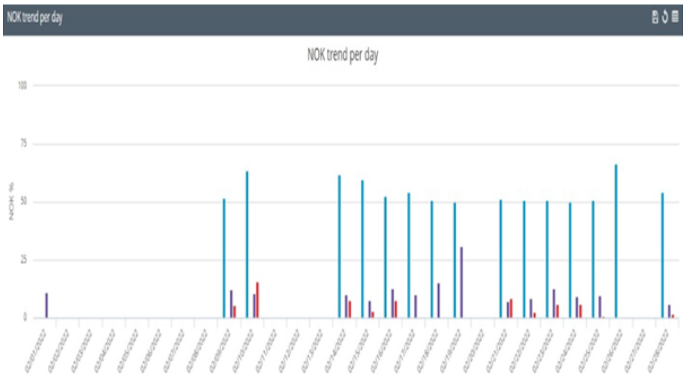


Figure 4.3: Nok trend per day for the month of February 2022.

Furthermore, looking at the SQL historical database, out of the 18090 joints performed at that workstation, the Nok joint percentage was 51.64 percent. As the Nok joint percentage in the chosen workstation was greater than 50 percent, a process capability analysis was carried out with one-week data to understand more about the robustness of process in the selected workstation. SPC and control charts are used for analyzing process capability and stability. A Normality test was also conducted to determine whether the sample data is normally distributed. The results were developed from Minitab, which are provided in Figure 4.4 and 4.5. The process capability analysis was performed for Ok + Nok joints as shown in Figure 4.4, and it was conducted only with the Ok joint, which is represented in Figure 4.5. The outcome of these SPC tests will be further analyzed and discussed in the discussion section.

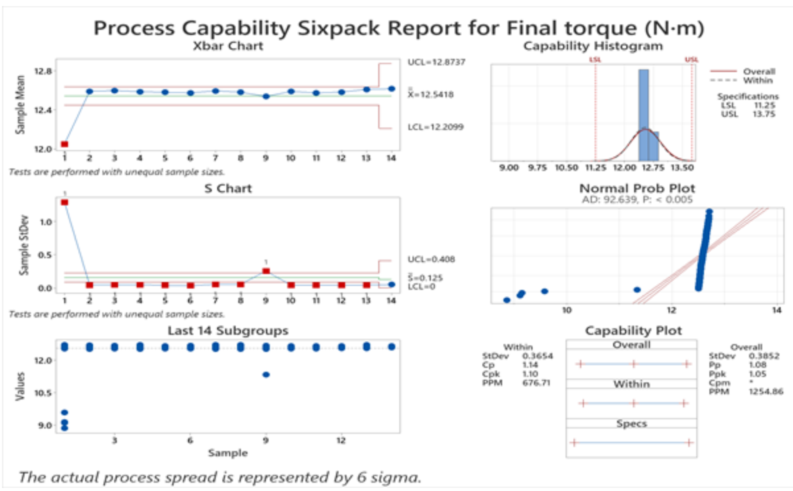


Figure 4.4: SPC studies for all joints in a week (Ok + Nok joints).

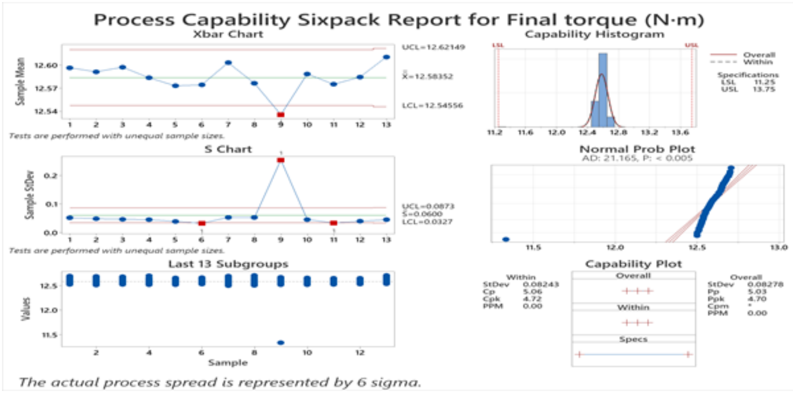


Figure 4.5: SPC studies for all joints in a week (Only Ok joints).

Wider Angle distribution

From the exploratory analysis of historical SQL data for the joints performed in this workstation, it was evident that the Torque values have a smaller limit varying within the control limits from 11.25 to 13.75 Nm, and the target torque set in the p-set is 12.5 Nm. As in the case of angle, there is a large bracket set for angle values ranging from 0 degrees to 10000 degrees. There is no control on other parameters like time and speed as the study has a fixed torque-controlled angle monitoring strategy. Some random distribution values have been chosen to represent the broader distribution of angle values from 0 to 10000 degrees (Refer to Figure 4.6).

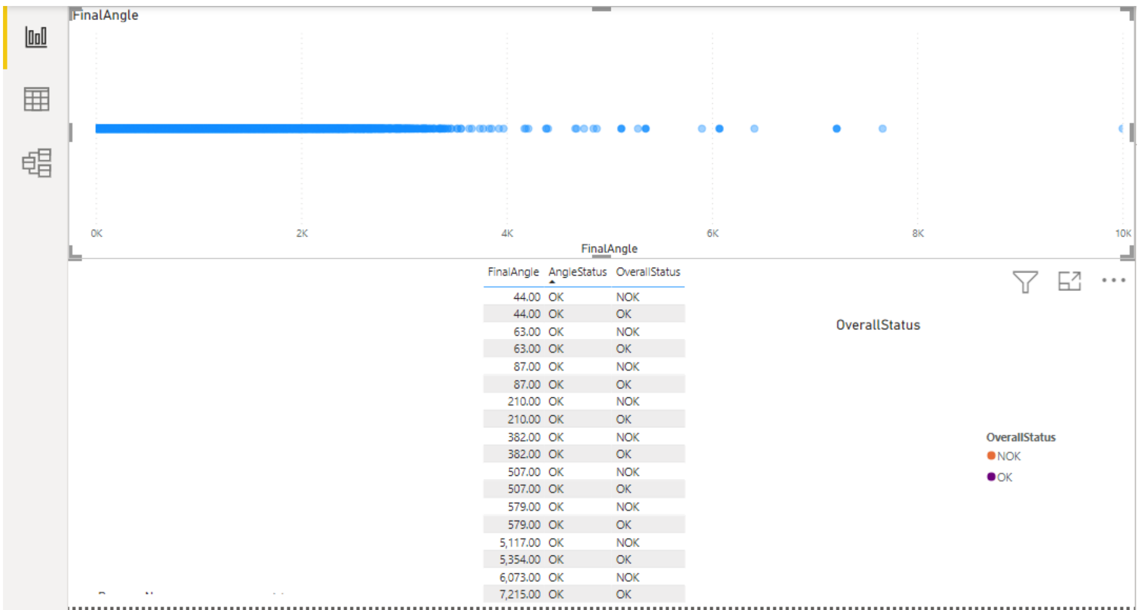


Figure 4.6: Illustration of wider-angle distribution in the selected program in the workstation.

4.2.1 Data Quality Verification Report for EDA in existing BoltsNut system and data from SQL database

This section contains data quality verification report for the EDA performed. The data for the EDA was primarily obtained from two sources namely the SQL database and from the live BoltsNut system. The data from the SQL database was queried into Power BI for further analysis. According to Cai and Zhu (2015), the different parameters of the data quality are availability, usability, reliability, relevance and presentability. The data quality examined from the SQL database as well as from the live BoltsNut system are mentioned below. The data quality such as availability, usability, reliability and relevance are explained in this section.

Availability: The historical SQL database was having two years of data available for analysis from January 2020 to February 2022. The data was available in the live existing BoltsNut system as well, but the search results were limited to a certain number, which showed the same result irrespective of the timeline. For instance, the pooled results were generated within two to four days in some workstations, so if we search for one day, one week or one month, or one year, irrespective of the search timeline, it was displaying the same results.

Figures 4.7 and 4.8 visualize the outcome of the search results done for different periods, but their internal system showed the same outcome for both the cases. This demonstrates that the limited search results provide a wrong perception for users who want to get insights from the BoltsNut system.

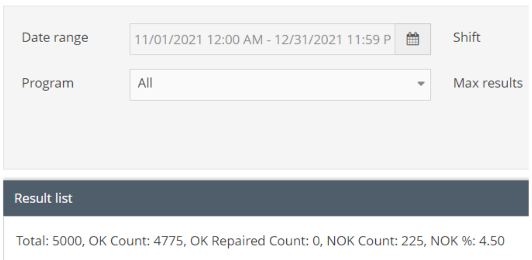


Figure 4.7: Search result for the 2-month duration in BoltsNut.

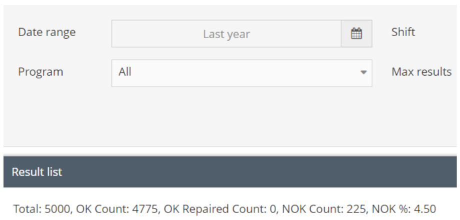


Figure 4.8: Search result for last 12-month duration in BoltsNut.

Usability: Some of the data acquired was deemed fit to be usable, but some data had a lot of missing values. The data was not complete, so it was hard to analyze the data or derive any insights from their existing live system. With the historical database, the missing values impacted the data analysis.

Reliability: The data from the existing BoltsNut system is unreliable due to poor data quality. The historical database also supports this erratic nature due to the poor quality of the data. There were different scenarios or cases where the negative effect of poor data quality is visualized below.

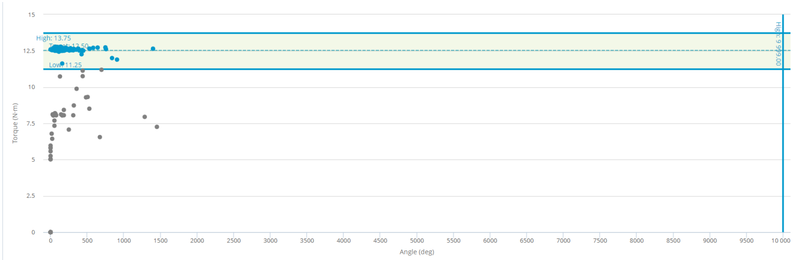


Figure 4.9: Scatter plot of the CC joints in the selected workstation, timeline - September 2021 to March 2022.

For the past six months, it was noted that in total, 7014 joints in that workstation, 3356 are OK. The remaining 3658 are Nok which gives a 52.15 Nok percentage (Refer to Figure 4.9). In that case, 3658 points should lie outside the SL as the joints did not reach the minimum target torque value of 11.25 Nm. Still, when plotted in the histogram, it was surprising to see only 87 points lie outside the SL as outliers, and the remaining points are missing (Refer to Figure 4.10).

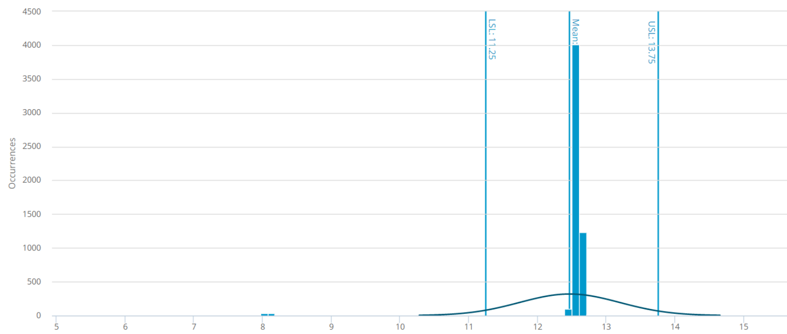


Figure 4.10: Histogram of the CC joints in the selected workstation, timeline - September 2021 to March 2022.

.

There were different scenarios where the data was unreliable and incomplete, which led to many other errors and data quality problems described below. Figure 4.11 and Table 4.2 visualize the errors in data due to missing values and undefined status from the SQL database. Out of 28 million joint results in the historical database, around 3.83 million results fall under this undefined category. These Not a Number (Nan) values or blank values should be discarded from the data to gain valuable decision-making insights. This 3.83 million consists of different scenarios which need to be fixed in the existing system.

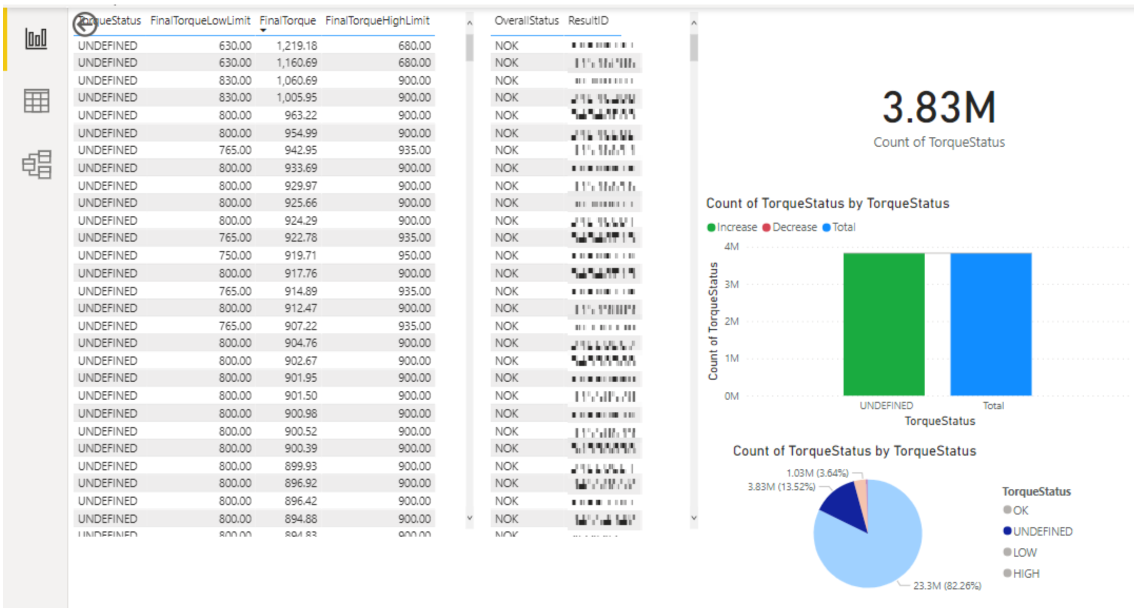


Figure 4.11: Visualization of the undefined status data from the historical database.

.

Table 4.2: Illustrates the data quality error in the undefined status from the historical database.

Torque-status	Angle-status	Final-torque	Overall-status	Count
Undefined	Undefined	Blank	Nok	51,591
Undefined	Undefined	Blank	Ok	21,840
Undefined	Undefined	Value-present	NOk	132,341
Undefined	Undefined	Value-present	Ok	3,337,333
Low	Ok	Blank	Nok	36291
High	Ok	Blank	Nok	29

Scenario 1: BoltsNut records Nok, although target torque is achieved. BoltsNut records status as Nok, although the target torque is achieved and falls within the specification limits. It can be noted that the status in the BoltsNut system is Nok is indicated in red, although the desired final torque is achieved (Refer to Figure 4.12). Another problem is the duplicates, as it recorded multiple entries for the same product or chassis number.

NOK	[illegible]	[illegible]	[illegible]	05/18/2022 05:26:45 PM	12.55 (11.25 - 13.75) N·m	OK	169.65 (0.00 - 9,999.00) deg	OK
NOK	[illegible]	[illegible]	[illegible]	05/11/2022 07:20:00 AM	12.56 (11.25 - 13.75) N·m	OK	78.89 (0.00 - 9,999.00) deg	OK
NOK	[illegible]	[illegible]	[illegible]	05/20/2022 02:38:01 PM	12.58 (11.25 - 13.75) N·m	OK	99.59 (0.00 - 9,999.00) deg	OK
NOK	[illegible]	[illegible]	[illegible]	05/12/2022 07:54:31 PM	12.59 (11.25 - 13.75) N·m	OK	25.31 (0.00 - 9,999.00) deg	OK
NOK	[illegible]	[illegible]	[illegible]	05/18/2022 01:40:53 PM	12.60 (11.25 - 13.75) N·m	OK	20.40 (0.00 - 9,999.00) deg	OK
NOK	[illegible]	[illegible]	[illegible]	05/09/2022 03:24:26 PM	12.66 (11.25 - 13.75) N·m	OK	114.63 (0.00 - 9,999.00) deg	OK

Figure 4.12: Data unreliability in BoltsNut Scenario 1

.

Scenario 2: BoltsNut records Nok status and shows torque status as Low, although the target torque value is blank. In some cases, the BoltNut automatically records the torque status as Low, although no values are found on the final target torque values (Refer to Figure 4.13). This was due to the poor quality of data.

Status	Identifier 1	Unit name	Program n...	Result time	Final torq...	Torque sta...	Final angle	Angle status
NOK	[illegible]	[illegible]	[illegible]	12/29/2021 05:13:05 PM	-	Low	-	OK
NOK	[illegible]	[illegible]	[illegible]	12/27/2021 11:35:31 PM	-	Low	-	OK
NOK	[illegible]	[illegible]	[illegible]	12/27/2021 11:35:30 PM	-	Low	-	OK
NOK	[illegible]	[illegible]	[illegible]	12/23/2021 02:09:01 PM	-	Low	-	OK
NOK	[illegible]	[illegible]	[illegible]	12/23/2021 02:08:55 PM	-	Low	-	OK
NOK	[illegible]	[illegible]	[illegible]	12/23/2021 01:33:14 PM	-	Low	-	OK
NOK	[illegible]	[illegible]	[illegible]	12/23/2021 01:32:28 PM	-	Low	-	OK
NOK	[illegible]	[illegible]	[illegible]	12/23/2021 12:35:53 PM	-	Low	-	OK

Figure 4.13: Data unreliability in BoltsNut Scenario 2

.

Scenario 3: BoltsNut records Ok status, although the target torque values are blank. This is another scenario where the BoltNut records the overall status of the joint as Ok, although the target torque values are blank (Refer to Figure 4.14).

	Status	Identifier 1	Unit name	Program n...	Result time	Final torq...	Torque sta...	Final angle	Angle status
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:25:15 AM	-	OK	-	OK
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:25:09 AM	-	OK	-	OK
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:25:04 AM	-	OK	-	OK
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:24:59 AM	-	OK	-	OK
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:24:52 AM	-	OK	-	OK
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:24:47 AM	-	OK	-	OK
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:24:42 AM	-	OK	-	OK
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:24:37 AM	-	OK	-	OK
▸	OK	[illegible]	[illegible]	[illegible]	12/30/2021 11:24:31 AM	-	OK	-	OK

Figure 4.14: Data unreliability in BoltNut Scenario 3

.

Scenario 4: BoltsNut records Nok once or twice for every chassis before the actual joint is performed. It can be noted that in the result time and the overall status before the actual joint is performed in every chassis, there exists Nok status twice for blank target torque values followed by the actual status of the joint with target torque values (Refer to Figure 4.15). There are some duplicates for the same chassis number as well.

Status	Identifier 1	Unit name	Program name	Result time	Final torque	Torque status	Final angle	Angle status
OK	A20001	[illegible]	[illegible]	05/20/2022 03:32:41 PM	12.61 (11.25 - 13.75) N·m	OK	116.13 (0.00 - 9,999.00) deg	OK
NOK	A20001	[illegible]	[illegible]	05/20/2022 03:30:43 PM	-	OK	-	OK
NOK	A20001	[illegible]	[illegible]	05/20/2022 03:30:43 PM	-	OK	-	OK
OK	A20002	[illegible]	[illegible]	05/20/2022 03:24:15 PM	12.57 (11.25 - 13.75) N·m	OK	104.66 (0.00 - 9,999.00) deg	OK
NOK	A20002	[illegible]	[illegible]	05/20/2022 03:17:51 PM	-	OK	-	OK
NOK	A20002	[illegible]	[illegible]	05/20/2022 03:17:51 PM	-	OK	-	OK
OK	A20003	[illegible]	[illegible]	05/20/2022 03:16:55 PM	12.58 (11.25 - 13.75) N·m	OK	101.02 (0.00 - 9,999.00) deg	OK
NOK	A20003	[illegible]	[illegible]	05/20/2022 03:14:55 PM	-	OK	-	OK
NOK	A20003	[illegible]	[illegible]	05/20/2022 03:14:55 PM	-	OK	-	OK
OK	A20004	[illegible]	[illegible]	05/20/2022 03:08:27 PM	12.58 (11.25 - 13.75) N·m	OK	141.77 (0.00 - 9,999.00) deg	OK
NOK	A20004	[illegible]	[illegible]	05/20/2022 03:08:09 PM	-	OK	-	OK
NOK	A20004	[illegible]	[illegible]	05/20/2022 03:08:09 PM	-	OK	-	OK

Figure 4.15: Duplicates found in the results in BoltsNut scenario 4

.

Scenario 5: A Special case was found in this workstation and other workstations.

In this specific case, the BoltsNut system recorded the Nok joint even before receiving any instructions from the controller. In general, the internal main control system sends instructions to the controller. The controller converts this instruction to the p-set and sends it to the smart tool, which performs the tightening operation based on the p-set (Refer to Figure 4.16). For the actual scenario, refer to Figures 4.17 and 4.18, which represent the special case of scenario 5.

▸	OK	AA302335	[illegible]	[illegible]	04/26/2022 12:59:40 PM	12.62 (11.25 - 13.75) N·m	OK	129.94 (0.00 - 9,999.00) deg	OK
▸	NOK	AA302335	[illegible]	[illegible]	04/26/2022 12:53:41 PM	-	OK	-	OK
▸	NOK	-	[illegible]	[illegible]	04/26/2022 12:53:40 PM	-	OK	-	OK

Figure 4.16: Tool locked by the open-source protocol in BoltsNut scenario 5
.

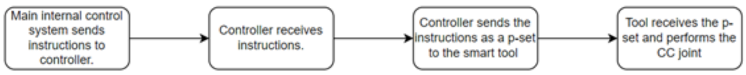


Figure 4.17: Actual scenario during the tightening process in a smart tool in ABC
.

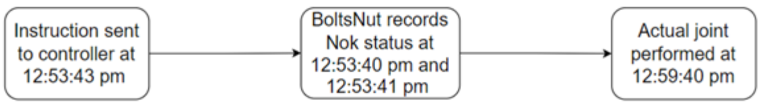


Figure 4.18: Data unreliability in BoltsNut scenario 5
.

The root cause was done for this specific case where it was found that the error message displayed ‘Tool sequence abort, tool locked by open-source protocol’. This recorded multiple entries of Nok joint on the selected workstation and in other workstations. It was identified that the problems exists in the communication channel between the company’s internal control system and the smart tool’s controller which automatically sends wrong p-set in between to produce the Nok results even before the instruction reaches the tool. This was creating Nok results due to the improper setup of the program in the controller, smart tool and internal system.

Relevance: Most of the blank values or errors were not relevant to the actual functioning of the existing system.

4.2.2 Data Quality Report for ML model

As mentioned before, the input data for the ML model consists of different parameter values such as time, torque, angle, and speed for the individual tightening sequence. The data was acquired for the three-month timeline from Feb 10th to April 10th, 2022, but the process was time-consuming as they needed to be manually downloaded from the live BoltsNut database.

Availability: The input values are acquired from the live BoltsNut database as the historical database does not contain parameter values for individual tightening sequences.

Usability: The raw data acquired from the live database was unstructured, and it required pre/processing steps to be applied to be helpful for the ML model.

Reliability: Although the BoltsNut system was having many un-reliability issues due to the poor quality of data, the ML model input data was found to be reliable. Some blank values in the feature table were cleaned using is missing () command in MATLAB before further processing.

Relevance: The raw data had a lot of irrelevant columns that were pre-processed before further processing.

4.2.3 Manual Actions Dashboard

As discussed in the business understanding, many operators and team leaders were performing manual actions/backup routines due to many reasons that the management is not fully aware of. To highlight the criticality of the issue to the management and understand if there is a trend or any pattern with the data, historical data from last year, January 2021 was considered, and four-month data in the current year from January 2022 to April 2022 was chosen for the analysis. This was done to understand whether there was any change in the previous year compared to the current year.

Figure 4.19 represents the number of Ok adjusted manual actions carried out in January 2021. The Cabtrim area seems to have more manual actions, consisting of a significant share of 78.98 percent in the pie chart. In contrast, the other areas had a relatively tiny section.

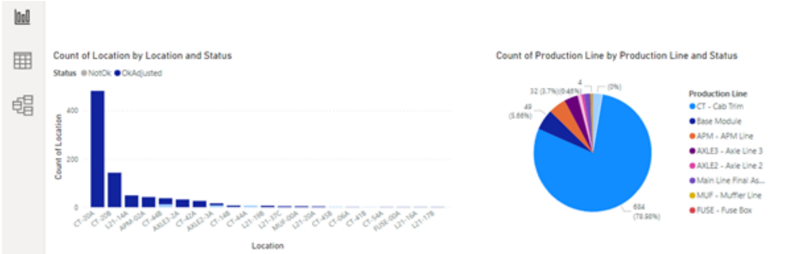


Figure 4.19: Historical data for Ok Adjusted Manual actions in January 2021

.

Figure 4.20 as given below represents the visualization for the current year - four-month timeline. Furthermore, by comparing the previous year and current year, a trend with the data can be found. The Cabtrim area emerged as the significant contributor with 59.22 percent, which indicate that a lot of manual actions are performed in this area while the other areas in the plant had least contributions.

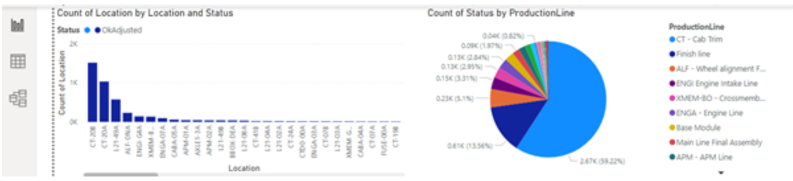


Figure 4.20: The current year 2022 data for the Ok adjusted Manual actions, timeline January 2022 to April 2022.

It can be noted that two stations in Cabtrim CT-20B and CT-20A held the first two positions in both the years (Refer to Figures 4.19 and 4.20). The dashboard template for the plant will be further explained in the modeling section of the results.

4.3 Data Preparation

The data preparation stage was critical as the efficiency of the output model depends on the quality of the input data fed into the model. The data preparation was conducted by performing many pre-processing techniques on the raw data. As the focus of this study was limited to a particular CC workstation in the Cabtrim area, the data selected for building the model was acquired for a three-month timeline from February 10th to April 10th, 2022, with 2184 individual samples initially chosen for this study. Figure 4.21 illustrates the tightening sequence for three-bolt tightening joints.

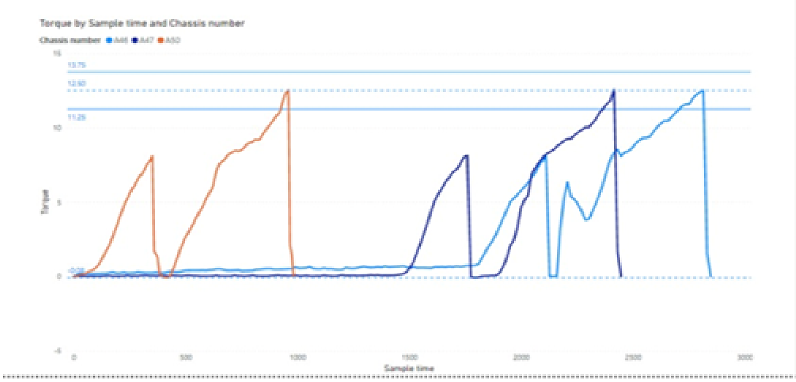


Figure 4.21: Visualization of tightening sequence for bolt tightening joints

.

Similarly, for the 2184 samples considered, the graph values for each tightening sequence were downloaded manually from the live database. As each tightening sequence detail must be downloaded manually from the live BoltsNut database, the first step was to integrate and construct a common file by merging these individual tightening sequence results. The individual excel sheets with tightening sequences were merged into a common file with other columns using queries in advanced excel. Following the data construction, the data was cleaned by removing irrelevant columns from the file. The raw data contained around 18 columns, out of which four were selected to prepare the model (Refer to Figure 4.22). The other 14 columns were removed due to missing values and non-relevance to the model.

Excel file name	Sample time (ms)	Torque (N·m)	Angle (deg)	Speed (rpm)	Result ID	Torque limit low (N·m)	Torque limit high (N·m)	Angle limit low (deg)	Angle limit high (deg)	Result time	Unit name	Program name	Max torque (N·m)	Max angle (deg)	Final time	Final Torque (N·m)	Final angle (deg)

Figure 4.22: Representation of constructed raw data

.

Further, the data was cleaned and reduced, an additional column, ‘Status code,’ was constructed along with four columns (Sample time, torque, angle, and speed). The Status code represents the joint status outcome, which is indicated as 0 for Nok status and 1 for Ok status (Refer to Figure 4.23).

Sample time	Torque (N·m)	Angle (deg)	Speed (rpm)	Status code
0	0	0.067121243	0	1
8	-0.001298893	0.469848699	9.788534132	1
16	0.033771212	1.342424853	13.98362019	1
24	0.050656817	2.701630017	18.76135709	1
32	0.10391142	4.396441394	22.89817806	1
40	0.03766789	6.191934635	25.79977925	1
48	0.042863461	8.205571915	28.49162613	1
56	0.058450174	10.77295945	32.06244343	1
64	0.081830243	13.97799878	36.4011113	1

Figure 4.23: Transformed raw data with reduced columns

.

Further, the data was transformed into a timetable, i.e., all the other parameters with respect to time (Refer to Figure 4.24). This type of input form was required to convert the numerical data into a signal form from which we can further extract different time-based and frequency-based features in MATLAB. The time series or timetable consists of a sequence taken at successive equally spaced points in time.

tv_ensemble_table ×

2184×4 table

	1	2	3	4	5
	Torque	Angle	Speed	StatusCode	
1	133×1 timetable	133×1 timetable	133×1 timetable	0	
2	152×1 timetable	152×1 timetable	152×1 timetable	0	
3	153×1 timetable	153×1 timetable	153×1 timetable	0	
4	118×1 timetable	118×1 timetable	118×1 timetable	0	
5	118×1 timetable	118×1 timetable	118×1 timetable	0	
6	128×1 timetable	128×1 timetable	128×1 timetable	0	
7	186×1 timetable	186×1 timetable	186×1 timetable	0	
8	169×1 timetable	169×1 timetable	169×1 timetable	0	
9	120×1 timetable	120×1 timetable	120×1 timetable	0	

Figure 4.24: Data transformed in the form of timetable

.

Each row represents each tightening sequence with torque, angle, and speed with respect to time, and the Status code represents the outcome of the joint. For instance, the first column Torque 133x1 timetable indicates the torque values at successive equally spaced points of time for that joint from start time 0 sec to end time of the joint (Refer to Figure 4.25).

tv_ensemble_table × tv_ensemble_table.('Torque'){1,:} ×

133×1 timetable

	Time	1 Torque	2	3
1	0 sec	0.0312		
2	8 sec	0.0468		
3	16 sec	0.1818		
4	24 sec	0.2416		
5	32 sec	0.3923		
6	40 sec	0.5079		
7	48 sec	0.5377		
8	56 sec	0.5598		
9	64 sec	0.5689		

Figure 4.25: Torque with respect to time in a timetable

.

4.4 Modelling

The ML model aims to understand the process's behavior and predict the outcome of the tightening process. As per the initial assumption in the study, torque is the main influencing factor that impacts the result of the process, as Ok and Nok were based on the final target torque values. Therefore, the factors influencing the output will be analyzed through the ML model to understand whether other parameters like angle and speed affect the outcome other than the main parameter torque. Further, the result of the Ok-adjusted manual action dashboard will be presented after presenting the results of the ML model.

4.4.1 Selecting Modeling Technique

The selection of modeling technique is based on the type of the available data set and the data mining goal. In this study, the method used for the ML model is the Classification method under the Supervised learning type. Supervised ML is chosen for this study as the data is labeled, and the model has inputs and output parameters. The input parameters for the model are torque, angle, and speed as the independent variable. The output parameter is the status code, the dependent variable, or the condition variable. The inputs and outputs are labeled, and the result represents two classes, i.e., Ok and Nok, so a classification method was chosen for building the ML model.
In this study, no specific modeling technique was chosen initially. The model was trained with all the modeling techniques, and the final selection was made based on the performance of the respective model. Additionally, the decision trees model was chosen to extract the rule-based system for predicting the model's outcome.

4.4.2 Generating Test Design

The input data with 2184 samples were split into two parts, constituting 70 percent training dataset and 30 percent test dataset for generating test design. 1529 samples were chosen for the training dataset, and the remaining 655 samples constituted the test-data set. The data was split to evaluate the error rates that impact the modeling quality. Another reason is that the model will be built based on the training dataset and tested with test data to validate the built model.

4.4.3 Building the model

The first step in this stage is to extract different features from the training data. After building the model, the model needs to be trained with various features. The final step is to do a hyperparameter tuning to find the optimum values of the parameters that will provide better validation accuracy and enhanced performance for the generated model.

4.4.3.1 Feature Extraction

Using the DFD application in MATLAB, the time-based and frequency-based signal features were extracted from tracing the data in the form of a signal. For instance, Figures 4.26 and 4.27 show the signal trace and power spectrum for the extracted features from the torque signal of the training dataset. Similarly, the different features for other parameters such as angle and speed are illustrated in the Appendix B.

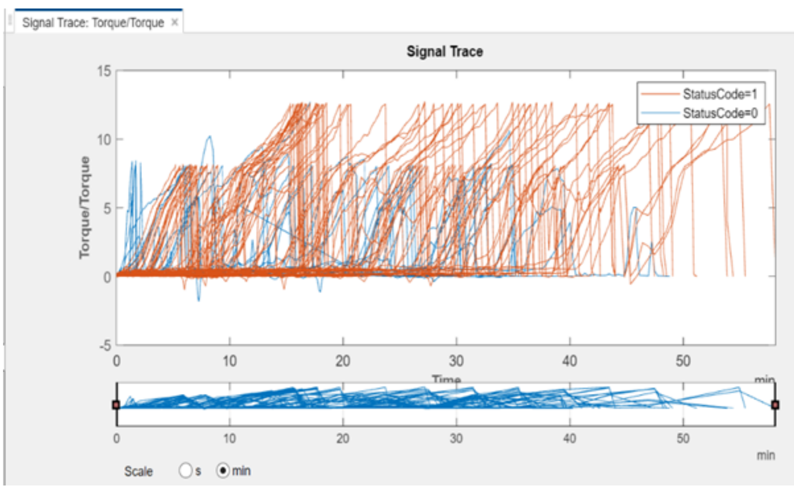


Figure 4.26: Visualization of the time-based features of the training dataset.

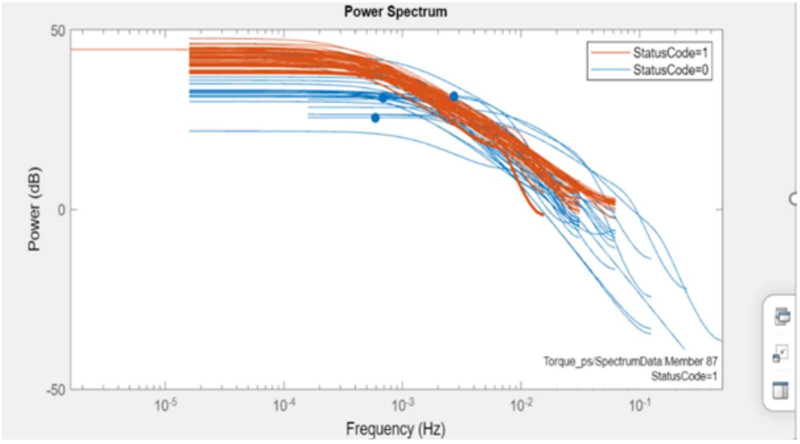


Figure 4.27: Visualization of the frequency-based features of the training dataset.

The number of total extracted time and frequency domain features were 109 from the three signals (torque, angle, and speed) based on time and frequency. These features were exported to the MATLAB workspace as a feature table for the training dataset. These features in the feature table need to be sorted and prioritized based on their importance. The sorting and ranking of the features are carried out using the DFD application, and Figure 4.28 represents the top-ranked features. The values of these parameters are calculated by parametric testing t-test. The t-test is a statistical parametric test used to determine the significant difference between the means of two groups (T. K. Kim, 2015).

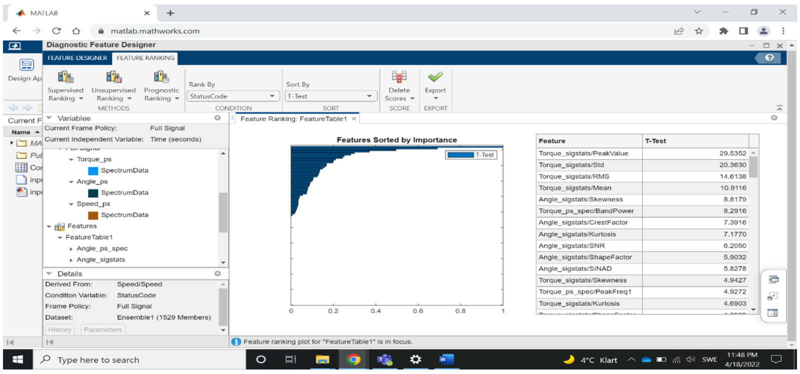
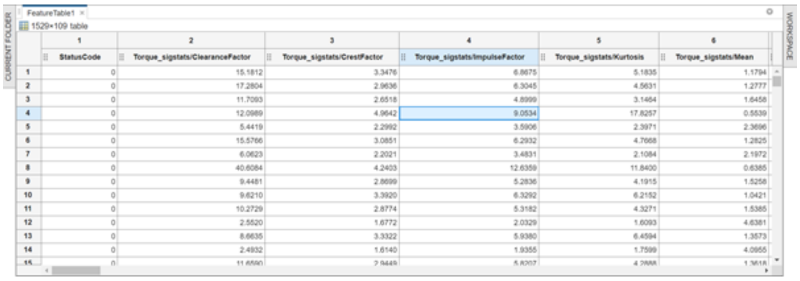


Figure 4.28: Ranking of features for the training data through DFD.

The top-ranked ten features from Figure 4.28 were based on torque and angle features. Interestingly, the torque-related features were expected to be the dominant ones as per the study's assumption that torque is the only factor influencing the output of the joint. The study's finding also correlates with the same assumption. However, another interesting point to note here is that angle-related features occupy five spots in the top ten features indicating that the angle parameter plays a significant role in determining the output. The top ten features are given below:

Torque/sigstats/Peakvalue,
Torque/sigstats/Std,
Torque/sigstats/RMS,
Torque/sigstats/Mean,
Angle/sigstats/Skewness,
Torque/ps/spec/Bandpower,
Angle/sigstats/Crestfactor,
Angle/sigstats/Kurtosis,
Angle/sigstats/SNR and
Angle/sigstats/Shapefactor.

Some of the basic definitions of these features are described below. The feature table for training data is listed below (Refer to Figure 4.29).


FeatureTable1
1529×109 table

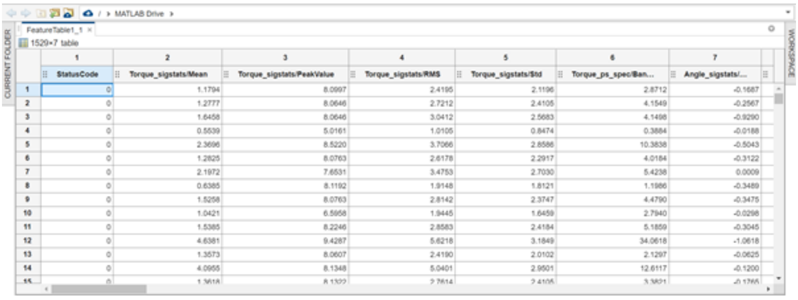
	1	2	3	4	5	6
	StatusCode	Torque_sigstats/ClearanceFactor	Torque_sigstats/CrestFactor	Torque_sigstats/ImpulseFactor	Torque_sigstats/Kurtosis	Torque_sigstats/Mean
1	0	15.1812	3.3476	6.8675	5.1835	1.1794
2	0	17.2804	2.9636	6.3045	4.5631	1.2777
3	0	11.7093	2.6518	4.8999	3.1464	1.6458
4	0	12.0989	4.9642	9.0534	17.8257	0.5539
5	0	5.4419	2.2992	3.5906	2.3971	2.3696
6	0	15.5766	3.0851	6.2932	4.7668	1.2825
7	0	6.0623	2.2021	3.4831	2.1084	2.1972
8	0	40.6084	4.2403	12.6359	11.8400	0.6385
9	0	9.4481	2.8699	5.2836	4.1915	1.5258
10	0	9.6210	3.3920	6.3292	6.2152	1.0421
11	0	10.2729	2.8774	5.3182	4.3271	1.5385
12	0	2.5520	1.6772	2.0329	1.6093	4.6381
13	0	8.6635	3.3322	5.9380	6.4594	1.3573
14	0	2.4932	1.6140	1.9355	1.7599	4.0955
15	0	11.6590	2.9449	5.8207	4.2888	1.3618

Figure 4.29: Feature table for training dataset (Table dimension: 1529x109).

Mean is the sum of observed variables in a data divided by the number of observations (Isotalo, 2001). Standard Deviation (SD) describes the variability of the data about the mean within a given sample (Nerurkar, 2008). RMS is defined as the quadratic mean (Nerurkar, 2008). The peak value is the highest point in a signal (Nerurkar, 2008). Band Power is used to estimate the power of the signal (Nerurkar, 2008). According to Caesarendra and Tjahjowidodo (2017), Impulse factor is the ratio of the maximum absolute value to the mean of the absolute value and the crest factor is termed the ratio of standard deviation to RMS. Skewness is defined as the asymmetry behavior of vibration signals through its probability density function (PDF). Kurtosis symbolizes the peak value of the PDF. Skewness measures the symmetry of the distribution, whereas Kurtosis measures the heaviness of the distribution tails (Caesarendra & Tjahjowidodo, 2017).

4.4.3.2 Training the model

For this study, the features with t-test score higher than 8 are considered significant. Therefore, out of the above features, the top six features were found to be having t-test scores higher than 8. Moreover, these six features were selected due to their clear distinction of classes i.e., StatusCode 0 and StatusCode 1 in the histograms. A hyperparamter tuning for the selected features will be performed. This step can be repeated several times by selecting additional features until we get a model with higher accuracy and better performance. The selected top six features are given below: Torque/sigstats/Peakvalue, Torque/sigstats/Std, Torque/sigstats/RMS, Torque/sigstats/Mean, Angle/sigstats/Skewness, Torque/ps/spec/Bandpower. The feature table represented in Figure 4.30 contains six input features and one output feature representing the status code.


MATLAB Drive

FeatureTable1_1

1529×7 table

	1	2	3	4	5	6	7
	StatusCode	Torque_sigstats/Mean	Torque_sigstats/PeakValue	Torque_sigstats/RMS	Torque_sigstats/Std	Torque_ps_spec/Ban...	Angle_sigstats/...
1	0	1.1794	8.0997	2.4195	2.1196	2.8712	-0.1687
2	0	1.2777	8.0646	2.7212	2.4105	4.1549	-0.2567
3	0	1.6458	8.0646	3.0412	2.5683	4.1498	-0.9290
4	0	0.5539	5.0161	1.0105	0.8474	0.3884	-0.0188
5	0	2.3696	8.5220	3.7066	2.8586	10.3838	-0.5043
6	0	1.2825	8.0763	2.6178	2.2917	4.0184	-0.3122
7	0	2.1972	7.6531	3.4753	2.7030	5.4238	0.0009
8	0	0.6385	8.1192	1.9148	1.8121	1.1986	-0.3489
9	0	1.5258	8.0763	2.8142	2.3747	4.4790	-0.3475
10	0	1.0421	6.5958	1.9445	1.6459	2.7940	-0.0298
11	0	1.5385	8.2246	2.8583	2.4184	5.1859	-0.3045
12	0	4.6381	9.4287	5.6218	3.1849	34.0618	-1.0618
13	0	1.3573	8.0607	2.4190	2.0102	2.1297	-0.0625
14	0	4.0955	8.1348	5.0401	2.9501	12.6117	-0.1200
15	0	1.3618	8.1322	2.7614	2.4105	3.3821	-0.1765

Figure 4.30: Feature table for top six features of training data (Table dimension: 1529x7).

The histograms of these features are visualized below (Refer to Figure 4.31). The StatusCode 0 for the Nok joint is indicated in blue, and the StatusCode 1 for the Ok joint is marked in orange.

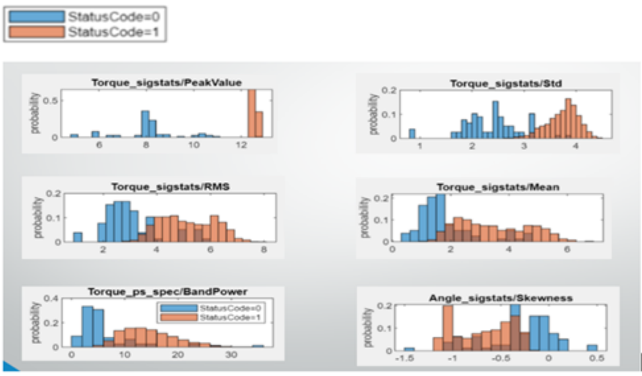


Figure 4.31: Histograms for the selected top six features.

The null values from the feature table are removed. The selected six features were trained in the CLA using all the 32 ML algorithms such as different types of K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision trees, Random forest, Regression, etc. Once the models are trained, the validation accuracy for each model is evaluated in this step. The steps for feature table extraction for the training data set should be repeated for extracting the feature table of the test data set.

4.4.3.3 Assessing the model

The last step in the modeling phase is assessing the model. The model generated from the previous step can be compared among themselves to facilitate the final selection of the model. The validation results of the training data for the selected features will be assessed and analyzed in this step. The best final model should be chosen by looking at the validation accuracy and the confusion matrix. The chosen model will be tested with a 30 percent test dataset to check the model's performance. The results of the trained model with the training data-set is represented in Figure 4.32 and the corresponding confusion matrix can be seen in Figure 4.33.

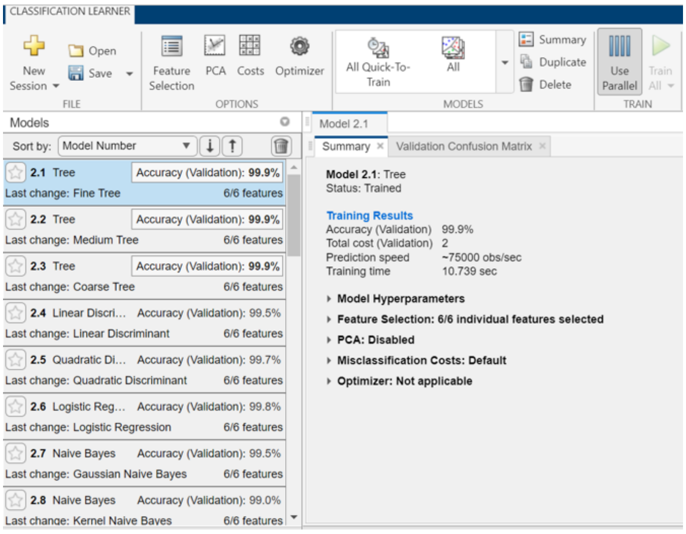


Figure 4.32: Validation accuracy of the trained model with selected features for the training dataset.

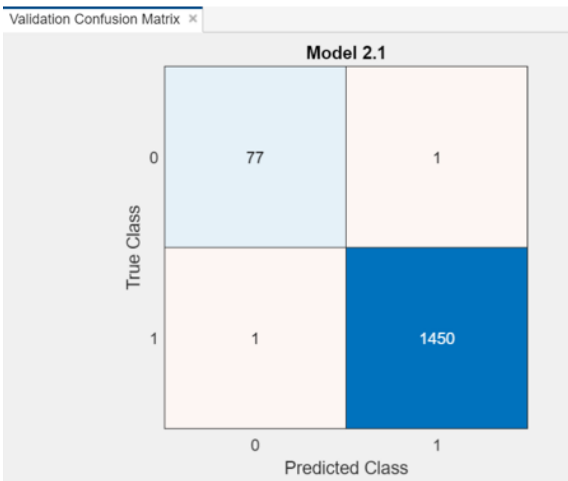


Figure 4.33: Confusion matrix for the trained model.

From Figure 4.32, it is evident that the model 2.1 Fine tree emerged as the model with the highest validation accuracy percentage (99.9 percent). This model was chosen to be the final selected model in this study. By looking at the confusion matrix for the model 2.1, it can be determined that the model could accurately predict the outcome of the process due to the close resemblance of precited and true class (Refer to Figure 4.33). The confusion matrix has four classes, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The confusion matrix with percentages is provided in the Appendix C. TP is when the model correctly interprets and predicts a positive class as the actual positive class. TN is when the model predicts a negative class to be an actual negative class. FP is when the model incorrectly identifies the predicted class as the actual class. FN is when the model's predicted class fails to determine the actual class (Schwenke & Schering, 2014). Lesser the false negative and false positive rates will lead to a better efficient model performance (Concha et al., 2014)

Our model results indicate that the TP value for the model is 99.9 percent, the TN is 98.7 percent, FP is 0.1 percent, and FP is 1.3 percent (Refer to Figure C.1 in Appendix C). Furthermore, the 30 percent test dataset will be fed as input to the exported model for validation with the test data. While testing the trained model with the test data, the feature table for test dataset needs to have the same set of six features chosen for the training dataset (Torque/sigstats/Peakvalue, Torque/sigstats/Std, Torque/sigstats/RMS, Torque/sigstats/Mean, Angle/sigstats/Skewness, Torque/ps/spec/Bandpower). The validation accuracy of the test data is evaluated using the predict function (predictFcn) with the exported model. Figure 4.34 shows a higher validation accuracy for the training as well as test data which indicates that the exported model is performing well with the test data. The confusion matrix for the exported model with validation is provided as shown in the Figure 4.35. The TP and TN values in Figure 4.35 seems to be higher which indicates the higher accuracy and better performance of the exported model.

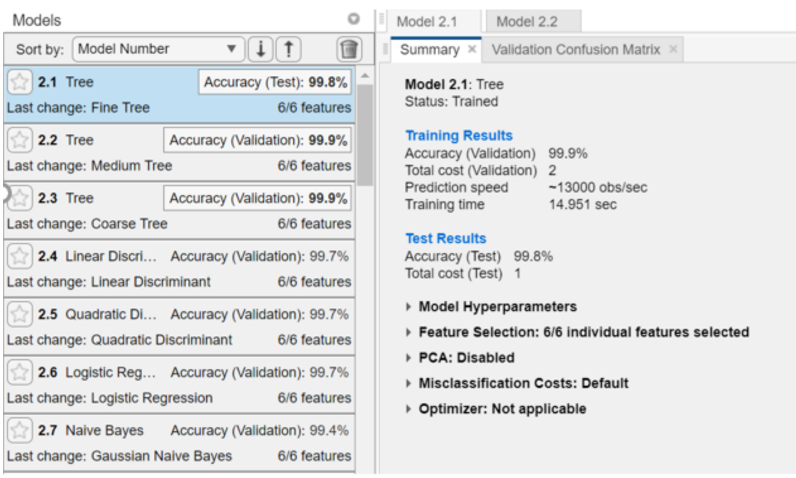


Figure 4.34: Validation results of the exported model with the test dataset.

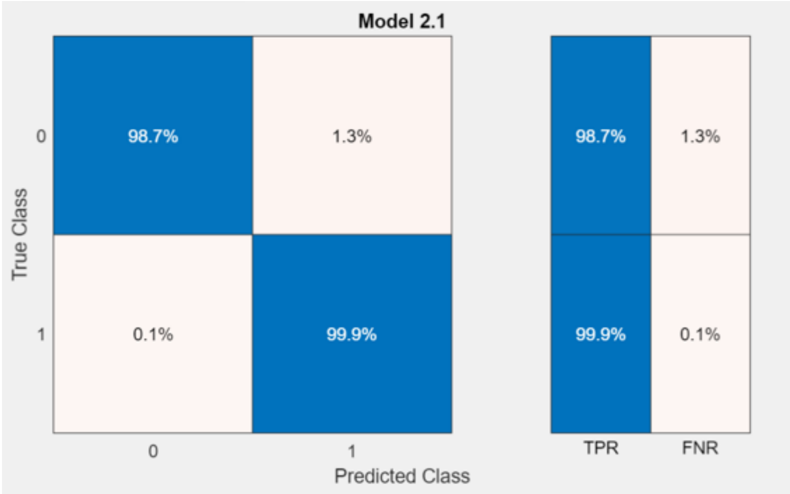


Figure 4.35: Confusion matrix for the exported model with validation of testdata.

Further a cross-validation for the exported model with the latest data was conducted with a small sample set. In this case, 50 joints completed during April 2022 were selected. Similarly, using the predict function, the exported model could predict the validation accuracy of cross-validated data (Refer to Figure C.2 in Appendix C) and the corresponding confusion matrix is evaluated (Refer to Figure 4.36).

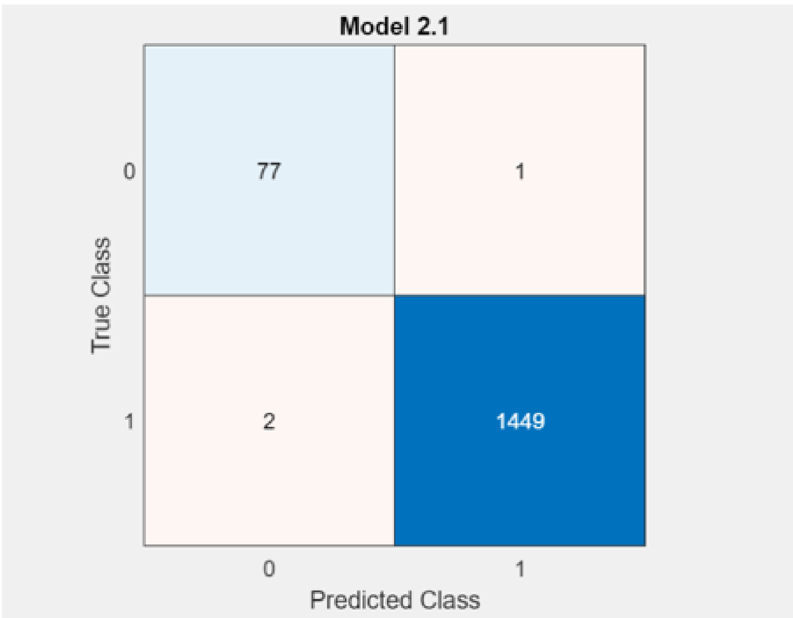


Figure 4.36: Confusion matrix for the exported model with cross-validation.

4.4.4 Extracting Decision Rules

The extraction of rules from the exported model using the decision trees is a critical outcome of the modeling phase. It can be understood how the decisions are made by looking at the decision trees' branches, nodes, and leaves. The plot for the decision trees is provided in Figure C.3 in the appendix C section. The three rules that are extracted from the model using decision trees are listed below:

Rule 1: If (Torque/sigstats/Peakvalue < 11.1405, Status = 0 (NOK)).

Rule 2: If (Torque/sigstats/Peakvalue >= 11.1405 and Angle/sigstats/Skewness < -0.100413, Status = 1 (OK)),

Rule 3: If (Torque/sigstats/Peakvalue >= 11.1405 and Angle/sigstats/Skewness>=-0.100413, Status = 0 (NOK)).

It can be understood from the rules that how the process output status changes as 0 (Nok) and 1 (Ok) based on the Torque/sigstats/Peakvalue and Angle/sigstats/Skewness). These rules demonstrate that angle is also a critical parameter other than torque in determining the output of the tightening process.

4.4.5 Visualization Solution for the Ok-adjusted Manual Actions Dashboard

The visualization solution for the business problem in this study is presented in the form of a dashboard that would help different teams within manufacturing. The dashboard regards the manual actions performed by the operators and team leaders as a part of the backup procedure if the joint cannot be achieved with the smart tool. There have been several instances where it is observed that operators use non-CC equipment for tightening instead of smart tools. The dashboard is illustrated in Figure 4.37.

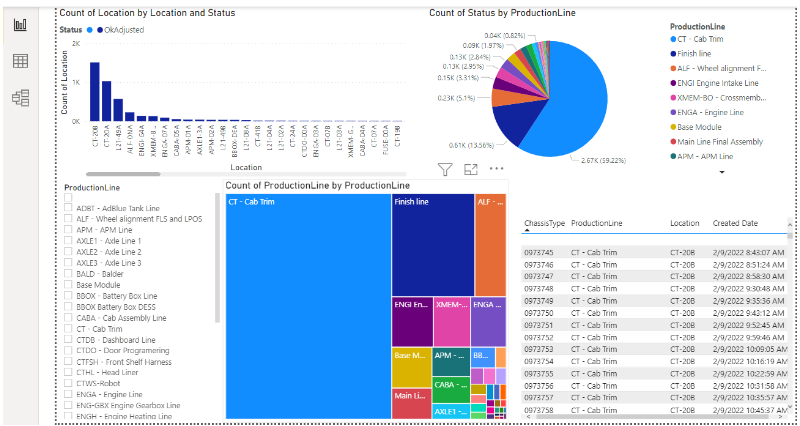


Figure 4.37: Illustrates the manual action dashboard.

The dashboard has been built with slicers that can drill down in a selected area, and all the visualizations will show result concerning the selected area in the slicer. For instance, in Figure 4.38, cab trim is the area chosen in the slicer, and the results will be visualized for cab trim. The above Figure 4.39 has multiple selections on the slicer, which shows combined results for those areas selected in the slicer.

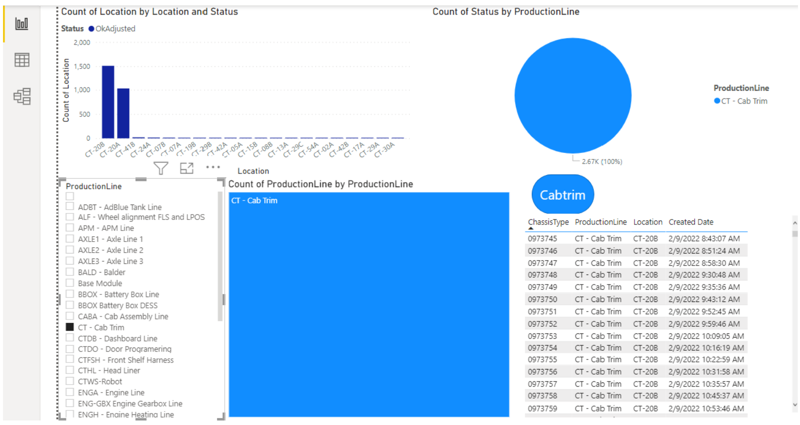


Figure 4.38: Drill-down options on the dashboard.

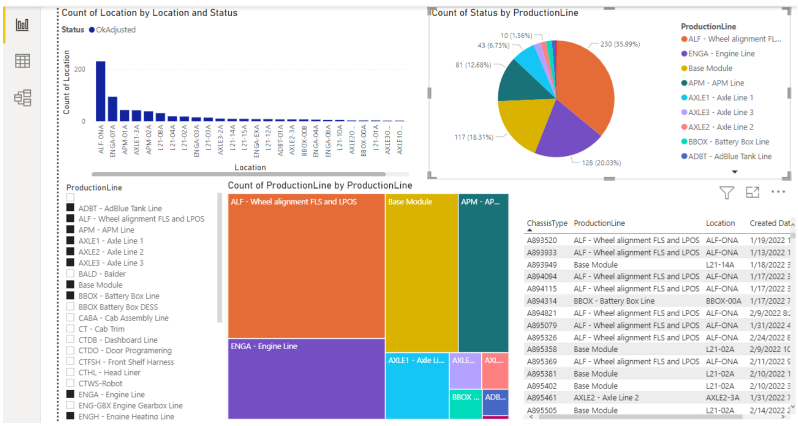


Figure 4.39: Representation of manual actions performed in multiple areas.

4.5 Evaluation

The diagnostic analysis provided some great insights regarding the model. The top features extracted gave an account of the influencing factors impacting the output of the tightening process. Importantly, the results obtained from DFD highlighted the presence of angle features within the top ten and oppose the initial assumption that Torque is the only influencing factor influencing the output. It states that angle also plays a significant role in impacting the output. The CLA provided insights on different ML modeling techniques and their corresponding confusion matrices. The results obtained from CLA is followed by further evaluation of the exported model.

The Fine tree model which is the selected final model have the highest validation accuracy of 99.9 percent for the training dataset. The exported model's confusion matrix also reflected the trend with higher TP, TN values and relatively low FP and FN percentages as shown in the Figure C.1 in appendix C. The higher the validation accuracy of the model for training data reflects the quality of the information provided by the selected six features. The selected top six features had distinct classes of status code 0 and status code 1 (Refer to Figure 4.31). At the same time, the model has a 99.8 percent accuracy for the test dataset. This clearly states that the designed model performs much better as this can be verified with higher validation accuracy for training and the test data. The model results tend to solve the business problems listed in this study and predicts the outcome of the bolt tightening process efficiently.

Further, cross-validation was conducted on company's request to verify the model's efficiency and behavior by testing the exported model with the latest data. The accuracy of the exported model with cross-validated dataset was found to be 100 percent, indicates that the model accurately predicts the outcome of all the samples with respect to the true class. The decision rules were then extracted from the exported model using decision trees to predict the classification based on status code 0 and 1. Three rules were extracted to classify status code 0 (Nok) and status code 1 (Ok).

The focus of dashboard is to critically divert the attention of all levels of management towards an area where more number of manual actions are carried out. The dashboard provides a high-level view of the overall plant as well as a detailed view of the individual workstation. The slicer option in the dashboard helps different departments to drill down to the particular workstation of their choice. As the data for completed joints with manual actions using non-CC equipment are not stored anywhere. More the number of manual actions leads to lesser traceability in the process. Additionally, more the number of manual actions also implies that the expensive smart tools are not used efficiently.

5

Discussion

Many industries face a practical challenge in converting the insights from the results section into competent business action. This chapter focuses on actual results and tries to connect with the literature summarized in Chapter 2. Moreover, the focus will be on answering the research questions of this study to facilitate a deeper understanding of the obtained results. It will be followed by the academic and practical contributions of this study.

A process is said to be stable if the distribution is constant over time; capable if the process operates under statistical control, and normal if the data is symmetrically distributed (Godina et al., 2016; Burgess, 2018; Taghizadegan, 2010). The results of Figure 4.4, which includes all joints (Ok + Nok), indicated that the current CC bolt tightening process seems to be normally distributed as the p-value was less than 0.005. However, the process was unstable due to outliers, i.e., Nok data in this case, but the process was still capable enough with a Cp value greater than 1. Although the process seems capable, it was not operating close to the company's target value, Cpk, which should be greater than or equal to 1.65 according to their standard. Outliers can have adverse effects on statistical analysis and generally reduce the power of statistical tests (Osborne & Overbay, 2004). The results were blowing up due to outliers, drastically impacting the SPC test results. It was found that the outliers in this workstation were because of human errors, and thereby, they were eliminated from further analysis. Figure 4.5 reflects the results for only Ok joints without outliers. Interestingly, the process was normal, highly stable, and capable of operating close to the target, seen in much higher Cp and Cpk values. The process performed very well without outliers; the study proceeded with further EDA.

Big data, characterized by high volume, variety, velocity, veracity, and value acquired by industry 4.0, IoT, and cloud, increases exponentially to ZBs (Cai & Zhu, 2015). Still, big data analysis must be done on accurate and high-quality data to develop valuable insights (Cai & Zhu, 2015). The study also focused on exploring the data quality and highlighting the impact of poor data quality in the company's existing third-party solution BoltsNut system. The poor quality of data negatively impacts the organization on different levels and affects business users, organizational culture, and operational costs (Laranjeiro et al., 2015; Haug et al., 2011). The EDA described in the results section illustrated several case scenarios where the poor data quality led to wrong insights, thus making the existing BoltsNut solution unreliable. Poor data quality generates direct, indirect, and additional costs for detecting, repairing, and preventing data quality-related problems (Haug et al., 2011). Due to poor data quality, the different scenarios discussed accounted for 3.83 million wrong results out of 28 million in the SQL historical database (Refer to Figure 4.11.). This will heavily impact the organization's operational operations, tactics, strategies, and overall decision-making (Laranjeiro et al., 2015)

The first research question **(RQ1)** will be addressed in this paragraph. Data DNA has become the new oil that drives research, innovation in science, technology, policymaking, industrialization, and economy with advanced analytics in data mining, a core discipline of data science (Cao, 2017). The result of the study applied data mining through EDA and advanced analytics using ML to put forward great insights regarding the current state of the process and highlight the issues due to the poor quality of data in their existing BoltsNut system. The use of advanced analytics such as AI/ML in manufacturing has introduced computer vision-based inspection, monitoring, fault detection, process improvement, optimization, prediction, cloud manufacturing, minimizing cost, reducing errors, increased customer experience, line efficiency, visibility, and life of assets with very minimal human involvement (Rai et al., 2021; Tao et al., 2018). Moreover, the ML model developed using MATLAB in this study provided great insights regarding the factors influencing the output and predicted the outcome of the tightening process. The ML model could help the company investigate other factors such as angle, speed, and time instead of having torque as the main influencing parameter. It will facilitate them to move towards smaller brackets for angles. Determining the control limits for the angle parameter was initially planned after finding that the angle significantly impacts the outcome. However, from Figure 4.6 (angle distribution over the program), it was evident that the angle has a much wider bracket from 0 degrees to 10000 degrees. It was challenging to set a control limit as the results show an overall status as Ok and Nok for joints with lower angles and higher angles. Henceforth, it was decided that a separate in-depth study would be required to determine the angle brackets.

Visual management is an integral part of lean manufacturing methodology as it highlights the critical problems to be addressed (Cepeda & Lopes, 2019). This study captured analytics techniques and provided a visualization solution to highlight the critical issues to all levels of management. Some visualization tools such as EDA, advanced excel, and business intelligence tools (Microsoft Power BI, Tableau, Qliksense, Oracle BI, SAP business, etc.) can be used to find possible trends, detect outliers, and statistical parameters (Sajid et al., 2021). This study used Microsoft Power BI to develop visualizations, reports, dashboards, and advanced excel for visualizations, managing queries related to data integration. The dashboard is a powerful, intelligent interactive tool that monitors and distributes critical information to all levels of management, thereby supporting tasks and enabling decision making (Cepeda & Lopes, 2019). The manual actions dashboard developed in this study will help direct all levels of management's attention to where a more significant number of manual actions were performed. The dashboard will allow the management to focus on problem-solving in these highlighted areas and enhance the use of smart tools at those locations. The process can be traced, monitored, and controlled effectively. As a result, the use of a data-driven approach with analytics and visualization techniques will help improve the visibility, traceability, and reliability of the CC bolt tightening process, which answers the first research question of this thesis.

The second research question **(RQ2)** focus will be on addressing the benefits of these techniques mentioned above concerning decision-making in different departments in manufacturing, namely the quality, production engineering, maintenance, procurement, and the audit team who are the end customers of this study. Addressing and solving the upstream problem due to non-robust to a robust process helps reduce the problem downstream's complexity by eliminating errors earlier in the process. Smart manufacturing improves line efficiency using analytics and visualization techniques and facilitates better insights for improved decision-making (Cao, 2017; Tao et al., 2018).

In the case of the quality team, they were losing much time creating quality deviation cases and executing quality control measures in case of a manual action/backup routine. The solution helps to improve the tightening process robustness, resulting in less case creation and reducing the number of manual actions through the dashboard, giving the quality team ample time to focus on other value-adding activities.

The production team loses much time due to Nok joints on rework and carrying out manual actions in case of a smart tool malfunction. Additionally, it stresses operators and team leaders physically and mentally. The operator calls for Andon when they cannot complete the joint before the truck moves to the next workstation. The mainline is designed to stop automatically if the CC joint is not completed within the specific workstation, resulting in production loss for the whole plant. To avoid the line stoppage, the operators override the system to carry out the joint with a backup routine or call for Andon. The team leader arrives at the specific workstation to override the system and perform a backup routine. Using the ML model and dashboard, a robust process helps the production team better quality joints, thereby reducing the rework and the count of manual actions carried out by non-CC equipment with no traceability.

For the audit team, due to the non-robust process, they could not stick to their control plan of inspecting CC joints on the partial audits (audits done on the mainline) because CC joints are performed in diverse areas at the plant. During their regular partial audits, the audit team found cases where the joints were manually tightened above the upper limit. In contrast, in another case, the auditors found the bolts loose. Moreover, there is no clear distinction between the joints tightened by smart equipment and non-CC equipment through manual actions. All the CC joints are marked in green after completion of the joint, which makes it difficult for the audit team to provide more focus on the manual tightened ones. Some reported CC joint failure cases due to manual actions had impacted 'n' number of chassis and cost the company 'xk' SEK. This solution will help them prioritize their efforts on a specific workstation where their resources can be allocated efficiently. A suggestion was provided for the company is to have different color marking for joints performed through manual actions. There is no storage of audit results, making it challenging to look for trends over time.

For maintenance, the non-robust process leads to more frequent maintenance and service of the equipment due to a more significant number of reworks due to Nok joints.

Due to the non-robust process, the load on the smart electric tool increases for procurement. The smart electric tightening tools are four times costlier than the regular non-CC tools. Therefore, the greater the number of attempts to perform the tightening operation creates a load on tools and impacts the tool's end of life in the long run. The engineering department (production and purchase) had invested considerable money in implementing BoltsNut at the plant. However, the existing solution is not fully utilized due to low data quality and unreliability.

Additionally, the load on the server is increased drastically for the IT department. The 3.8 million wrong results mentioned in the different scenarios heavily increase the database load, electricity, and infrastructure cost and reduce the speed of the connected systems. Moreover, the company has many internal systems where the data from the smart tool is stored in the central control system of the plant, from which the pooled results will be transferred to another system that keeps the tightening results for ten years. In BoltsNut scenario five as shown in Figure 4.16 and Figure 4.18, the smart tool records Nok joint even before the controller receives the instruction to perform the joint. At times, it records multiple entries of empty Nok results, which are then stored in the BoltsNut system, database, and in the company's internal central control system server, and finally gets stored in the internal historical server where these are held for ten years. These dummy values resulting from poor quality of data stored in these systems increase the server load and reduce the server's efficiency. These Nok results were found to be occurring due to the poor setup of the program in the controller, smart tool, and internal system. Therefore, vast improvements need to be carried out in examining the setup of these programs/functions in the controller and internal systems while installing the smart tool.

5.1 Academic Contribution

The literature studies on the tightening process of bolted critical joints are limited. In fact, there is no specific literature on the tightening process for CC bolted joints. This study can contribute immensely to the world of academia for researchers, students, and industry experts who want to learn more about the tightening process of CC bolted joints in an industrial setting. The results from this thesis can provide practical insights into applying different analytics and visualization techniques to make the process robust. It can also provide information regarding the importance of outliers and whether they should be considered or not while performing a statistical analysis. It also demonstrated how to apply the ML techniques using a CRISP-DM methodology to build a model and predict the outcome. The dimensions of good quality were presented, and the harmful effects of poor data quality were highlighted.

5.2 Practical Contribution

Although it can be noted in theories that many industries have applied AI/ML successfully, the knowledge of these techniques is very scarce in the ABC company where the study was conducted. The study has introduced these techniques to the ABC by providing the potential knowledge of why companies should consider and use a data-driven approach. Moreover, the ML model developed by the CRISP-DM method also demonstrated the different steps to apply the model in the future if the company wants to extend the solution to other workstations. In a real-time decision-making environment, time is a critical factor as decisions need to be made within milliseconds, leaving no room for identifying and fixing the poor quality of data before making decisions (Rego, 2020). The study also unlocked the existing and potential problems regarding poor data quality. It stressed why companies should not believe in mere numbers or data when working with industry 4.0; instead, they should verify data quality. The essential dimensions of good data quality were discussed. Interestingly, the ML model results outline the parameters to be considered for the outcome of the tightening process. This knowledge can help the company conduct extensive studies on these parameters to set a shorter control limit for angle. Furthermore, the study provided details in the discussion part on how this can be beneficial for the different departments in the company – Quality, Production, Audit, Maintenance, Procurement, and finally, the IT team. Lastly, the study contributes to the importance of using data analytics and visualization techniques to explore, analyze, and improve the insights from the raw data into meaningful, valuable insights to help in the proactive planning process.

The ML model's results using the CRISP-DM approach can seem attractive, but further extended research with possible recommendations in the conclusion and recommendation chapter should be carried out before implementing the model in a real-world setting. As this is a supervised ML model, the output is based on the inputs used to train the model. The challenge with this ML technique is that the model's performance won't be the same if the model has been fed with un-trained inputs. Therefore, as mentioned before, several recommendations should be made to train the model with a different range of inputs. However, if we exclude the poor data, the process seems to be robust, but the poor data occurs because the tool is not used in the way that the software supports or is not correctly set up in the plant. A vast improvement should be made for adequately setting up the software and tool to avoid data quality problems in the future.

6 Conclusion and Recommendation

As most companies are inclining towards industry 4.0 systems and smart manufacturing systems to remain competitive in the market, they should be cautious on how to handle the big data they capture through advanced technologies and IT systems. Their challenge is translating the raw data into valuable insights for improved decision-making. The fast-paced digital environment demands rapid decision-making skills without human interventions. These quick decisions can be made if good, accurate, and reliable data are available. At the same time, poor data quality affects the company's reputation and business by negatively impacting customer experience, business strategies, and tactics.

This study's outcome can demonstrate the effectiveness and successful application of the data-driven approach using CRISP-DM methodology in the truck manufacturing company ABC to facilitate the company's advanced planning and make effective decisions with data. The study's outcome consists of an analytic solution (in the form of an ML model), a visualization solution (in the form of a Dashboard), and significantly highlighted the poor data quality problems in the existing BoltsNut system through EDA and helped in improving the data quality thereby improving the process robustness. Implementing these take-aways and the recommended steps explained in the next section can help the company have a robust bolted critical tightening process and enhance the decision-making of different departments in manufacturing. The takeaway from the study is listed:

- **Data-Driven approach increased the robustness of the process.**

The ML model could predict the outcome of the process with high validation accuracy for the training, test, and cross-validated data. It also highlights the importance of the angle parameter in influencing the outcome of the process other than Torque. The live visualization solution will help direct the attention of all levels of management to the workstation with a greater number of manual actions. This solution will eventually reduce the number of manual actions, helping to utilize the smart tools effectively. The results led to increased knowledge, visibility, reliability, traceability, monitoring, and control of the CC bolted tightening process.

- **Data-science applications for process improvements.**

The ML knowledge was very scarce in the ABC company when the study was initiated. However, after the promising results, the company now slowly understands these studies' potential and plans to develop similar analyses in the future. Furthermore, these data analytics and visualization techniques provide ample knowledge about the process's current state and help identify and solve the problems.

- **Manufacturing companies working with Smart manufacturing technologies require vast improvements on the parameter setup of smart tools.**

Although the tightening process for bolted critical joints seems non-robust initially, it was evident from the study results that the process appears to be in control and stable as the poor data quality did blow up the Nok counts and percentages drastically. Furthermore, this indicates that the tool is not used in the way that the software supports and highlights the limited knowledge of the setup of parameters for the smart tool in the ABC company. It highlights the need for significant improvements. Therefore, significant improvements need to be taken by the company to fix and prevent these data quality problems in the future by improving the knowledge, communication, and setup of these parameters while setting up the smart tools, controllers, and internal systems.

6.1 Recommendations for Improvement

- Communicate with the third-party solution provider to improve their BoltsNut system to make it accurate and reliable for decision making.

- Check with the internal IT team and third-party solution provider to solve the fifth scenario discussed in the results section regarding the tool sequence aborted by the open-source protocol problem. They are both problem owners in this case.

- To remove process disturbances caused by poor data quality, which will increase the load on internal IT systems before moving towards fully electric tools in 2030.

- Provide different color marking for the joints performed through manual actions with non-CC equipment and digitalize the partial audit results to look for trends over time.

- Recommend the company to have more parameters to be considered for further analysis. Currently, the company records only four parameters through their BoltsNut system - Torque, Time, Angle, and Speed.

- Recommend the technical team to set control limits for angle.

- Check the model output by testing more samples with set control limits for different parameters in different workstations to understand the output since each workstation's material and operation differ.

- To run several test cases in the calibration workbench under different scenarios using damaged threads, oiled threads, greased threads, long and short bolts, nuts, and extra washers. The reason is greased, damaged bolts will not create resistance, resulting in a higher angle for the joint.

- To try and build the ML model with input values starting from the first target torque to nullify the effect of variance caused by operators during pre-tightening of bolt (referred to as rundown). It will be interesting to check for the accuracy and performance of the model.

- To construct an overall bird's eye view or plant level dashboard to understand where to start our analysis initially.

- To conduct the capability study for the process and the machines and compare their results. In this study, the capability of the process was carried out in the selected workstation by excluding the outliers.

- To conduct a Measurement System Analysis (MSA) to analyze the errors due to humans or instruments by gage repeatability and reproducibility (Gage R & R) method.

References

Azevedo, A., & Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview. *IADS-DM*.

Burgess, C. (2018). Process stability, performance, and capability; what is the difference? *Pharmaceutical Freeze Drying Technology*, 40.

Caesarendra, W., & Tjahjowidodo, T. (2017). A review of feature extraction methods in vibration-based condition monitoring and its application for degradation trend estimation of low-speed slew bearing. *Machines*, *5*(4), 21.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, *14*.

Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1–42.

Cepeda, T., & Lopes, I. d. S. (2019). Support methodology for product quality assurance: a case study in a company of the automotive industry. *Procedia Manufacturing*, *38*, 957–964.

CERVANTES, C., & Arman, E. N. (2018). *Quality assurance through smart angle monitoring: Improvement of tightening in scania's cab as-sembly and implementation into an industry 4.0 based system.*

Chen, Y., Gao, Q., & Guan, Z. (2017). Self-loosening failure analysis of bolt joints under vibration considering the tightening process. *Shock and Vibration*, *2017*.

Concha, A., Mills, D. S., Feugier, A., Zulch, H., Guest, C., Harris, R., & Pike, T. W. (2014). Using sniffing behavior to differentiate true negative from false negative responses in trained scent-detection dogs. *Chemical senses*, *39*(9), 749–754.

Critical Data, M. (2016). *Secondary analysis of electronic health records.* Springer Nature.

Dalmarco, G., Ramalho, F. R., Barros, A. C., & Soares, A. L. (2019). Providing industry 4.0 technologies: The case of a production technology cluster. *The journal of high technology management research*, *30*(2), 100355.

De Mantaras, R. L., & Armengol, E. (1998). Machine learning from examples: Inductive and lazy methods. *Data & Knowledge Engineering*, *25*(1-2), 99–123.

Ebrahimi, M., Baboli, A., & Rother, E. (2019). The evolution of world class manufacturing toward industry 4.0: A case study in the automotive industry. *Ifac-Papersonline*, *52*(10), 188–194.

Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal* *of Production Economics*, *210*, 15–26.

Fukuoka, T., & Takaki, T. (2003). Elastic plastic finite element analysis of bolted joint during tightening process. *J. Mech. Des.*, *125*(4), 823–830.

Godina, R., Matias, J. C., Azevedo, S. G., et al. (2016). Quality improvement with statistical process control in the automotive industry. *International Journal of Industrial Engineering and Management*, *7*(1), 1–8.

Gray, P., & McCarthy, C. (2010). A global bolted joint model for finite element analysis of load distributions in multi-bolt composite joints. *Composites Part B: Engineering*, *41*(4), 317–325.

Harding, J. A., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: a review.

Haug, A., Zachariassen, F., & Van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM)*, *4*(2), 168–193.

Isotalo, J. (2001). Basics of statistics. *Finland: University of tampere*.

Kang, H. S., Lee, J. Y., Choi, S., Kim, H., Park, J. H., Son, J. Y., … Noh, S. D. (2016). Smart manufacturing: Past research, present findings, and future directions. *International journal of precision engineering and manufacturing-green technology*, *3*(1), 111–128.

Kim, J., Yoon, J.-C., & Kang, B.-S. (2007). Finite element analysis and modeling of structure with bolted joints. *Applied mathematical modelling*, *31*(5), 895–911.

Kim, T. K. (2015). T test as a parametric statistic. *Korean journal of anesthesiology*, *68*(6), 540.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283.

Kotsiantis, S. B., Zaharakis, I., Pintelas, P., et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3–24.

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A survey on data quality: classifying poor data. In *2015 ieee 21st pacific rim international symposium on dependable computing (prdc)* (pp. 179–188).

Lené, J., & Rajashekarappa, M. (2021). A data-driven approach to detect air leakage in a pneumatic system.

Maimon, O., & Rokach, L. (2005). Data mining and knowledge discovery handbook.

Mao, Y., et al. (2015). Data visualization in exploratory data analysis: An overview of methods and technologies.

Martinez, W. L., Martinez, A. R., & Solka, J. L. (2017). *Exploratory data analysis with matlab®*. Chapman and Hall/CRC.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., … Flach, P. A. (2019). Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.

Matope, S., & Bekker, E. L. (2021). The requirements for predictive maintenance strategy and the data required for successful implementation, for the case of nutrunners.

McCarthy, M., & McCarthy, C. (2003). Finite element analysis of effects of clearance on single shear composite bolted joints. *Plastics, rubber and composites*, *32*(2), 65–70.

Monville, J.-M. (2016). Optimal tightening process of bolted joints. *International Journal for Simulation and Multidisciplinary Design Optimization*, *7*, A4.

Nerurkar, R. (2008). Basics of statistics for postgraduates. *Indian journal of dermatology, venereology and leprology*, *74*(6), 691–695.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), 6.

Peters, S., Chun, J.-H., & Lanza, G. (2015). Digitalization of automotive industry–scenarios for future manufacturing.

Piscan, I., Predincea, N., & Pop, N. (2010). Finite element analysis of bolted joint. *Proceedings in Manufacturing Systems*, *5*(3), 167–172.

Pop, L. D. (2020). Digitalization of the system of data analysis and collection in an automotive company. *Procedia Manufacturing*, *46*, 238–243.

Quinlan, J. R. (1990). Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, *20*(2), 339–346.

Rai, R., Tiwari, M. K., Ivanov, D., & Dolgui, A. (2021). *Machine learning in manufacturing and industry 4.0 applications* (Vol. 59) (No. 16). Taylor & Francis.

Redman, T. C. (2016). Bad data costs the us dollar 3 trillion per year. *Harvard Business Review*, *22*, 11–18.

Rego, A. L. (2020). *A framework to determine the impact of poor data quality on the reliability of iot sensor-based real-time decision-making* (Unpublished doctoral dissertation). University of Arkansas at Little Rock.

Ribeiro, R., Pilastri, A., Moura, C., Rodrigues, F., Rocha, R., & Cortez, P. (2020). Predicting the tear strength of woven fabrics via automated machine learning: an application of the crisp-dm methodology.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660–674.

Sajid, S., Haleem, A., Bahl, S., Javaid, M., Goyal, T., & Mittal, M. (2021). Data science applications for predictive maintenance and materials science in context to industry 4.0. *Materials Today: Proceedings*, *45*, 4898–4905.

Schäfer, F., Zeiselmair, C., Becker, J., & Otten, H. (2018). 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)* (pp. 190–195).

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, *181*, 526–534.

Schwenke, C., & Schering, A. (2014). True positives, true negatives, false positives, false negatives. *Wiley StatsRef: Statistics Reference Online*.

Sedrakyan, G., Mannens, E., & Verbert, K. (2019). Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. *Journal of Computer Languages*, *50*, 19–38.

Sha, K., & Zeadally, S. (2015). Data quality challenges in cyber-physical systems. *Journal of Data and Information Quality (JDIQ)*, *6*(2-3), 1–4.

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222.

Shao, G., Shin, S.-J., & Jain, S. (2014). Data analytics using simulation for smart manufacturing. In *Proceedings of the winter simulation conference 2014* (pp. 2192–2203).

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (indiacom)* (pp. 1310–1315).

Skiena, S. S. (2017). *The data science design manual.* Springer.

Smith, W. S., Coleman, S., Bacardit, J., & Coxon, S. (2019). Insight from data analytics with an automotive aftermarket sme. *Quality and Reliability Engineering International*, *35*(5), 1396–1407.

Sousa, S., Rodrigues, N., & Nunes, E. (2018). Evolution of process capability in a manufacturing process. *Journal of Management Analytics*, *5*(2), 95–115.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.

Taghizadegan, S. (2010). *Essentials of lean six sigma.* Elsevier.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, *48*, 157–169.

Tavera Romero, C. A., Ortiz, J. H., Khalaf, O. I., & Ríos Prado, A. (2021). Business intelligence: business evolution after industry 4.0. *Sustainability*, *13*(18), 10026.

Traub, M., Vögel, H.-J., Sax, E., Streichert, T., & Härri, J. (2018). Digitalization in automotive and industrial systems. In *2018 design, automation & test in europe conference & exhibition (date)* (pp. 1203–1204).

Vasudevan, A., & Duan, X. (2021). A systematic data science approach towards predictive maintenance application in manufacturing industry.

Wirth, R., & Hipp, J. (2000). 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, pp. 29–40).

Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, *4*(1), 23–45.

A

Appendix A

The following figures show the illustrations of trend of Nok percentage for four months (September 2021 to January 2022).

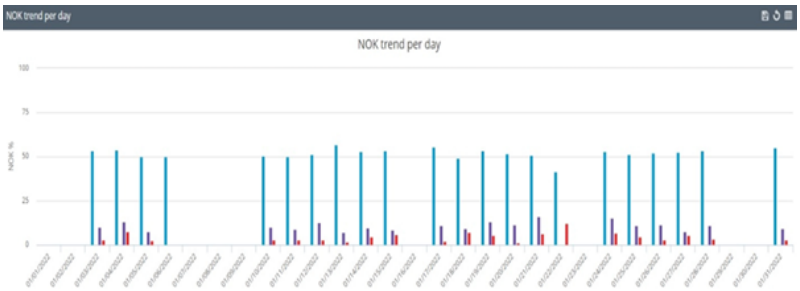


Figure A.1: Nok trend per day for the month of January 2022.

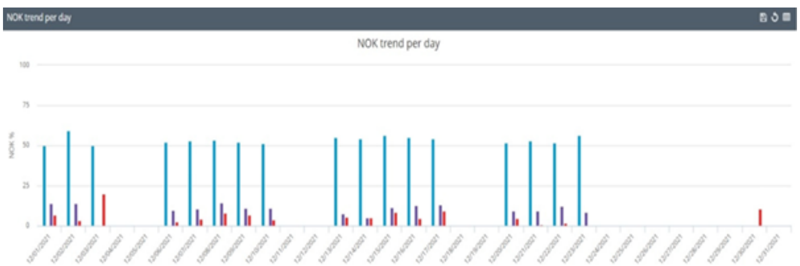


Figure A.2: Nok trend per day for the month of December 2021.

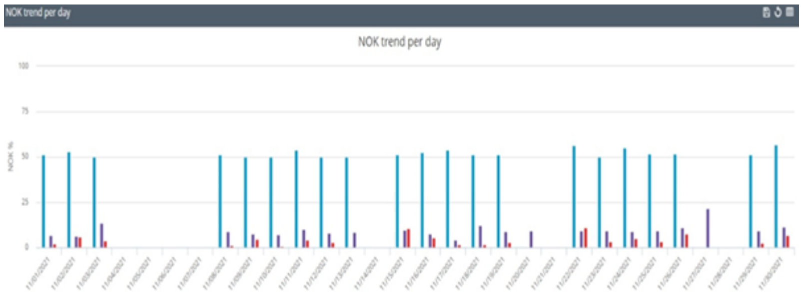


Figure A.3: Nok trend per day for the month of November 2021.

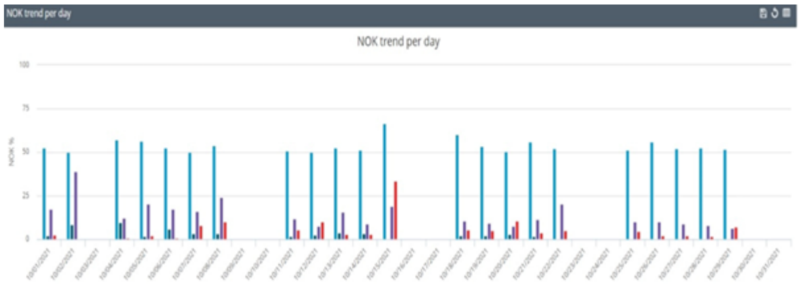


Figure A.4: Nok trend per day for the month of October 2021.

B

Appendix B

The following figures represent the time-domain and frequency-domain features of angle and speed paramters.

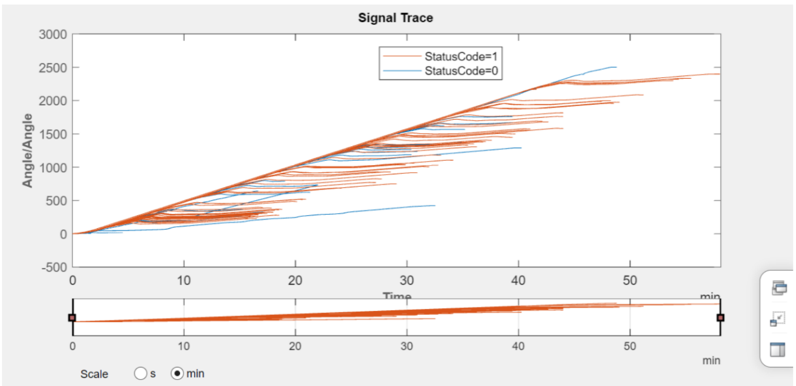


Figure B.1: Signal trace for the extracted time-domain features of angle.

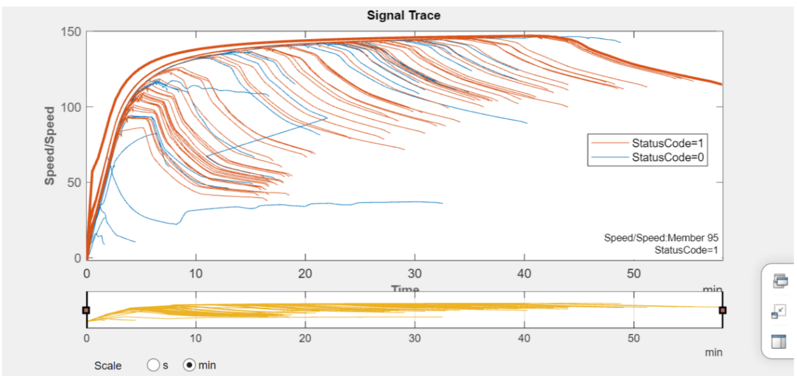


Figure B.2: Signal trace for the extracted time-domain features of speed.

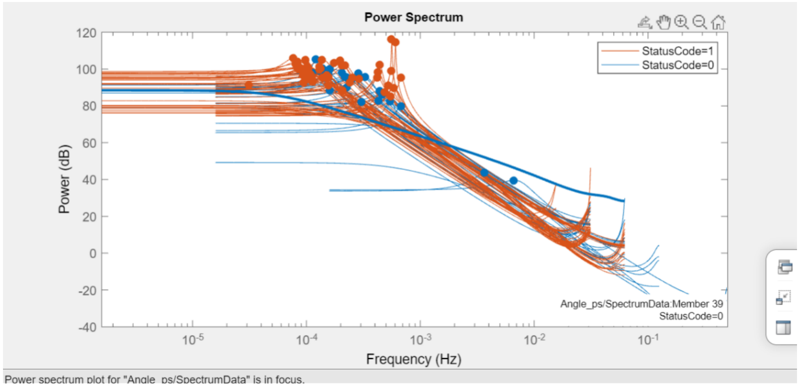


Figure B.3: Power spectrum of extracted frequency domain features for angle.

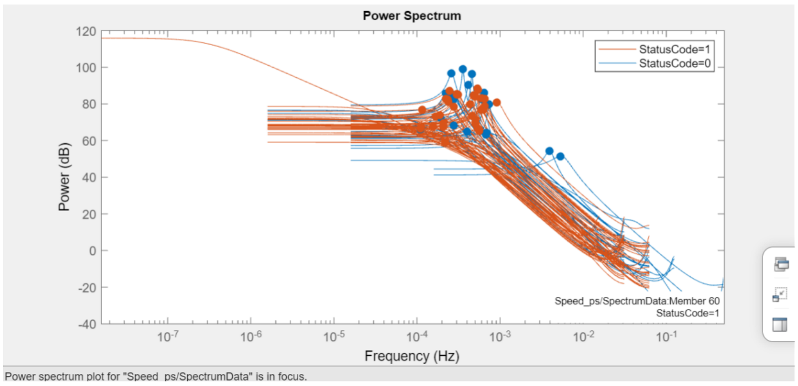


Figure B.4: Power spectrum of extracted frequency domain features for speed.

C

Appendix C

The figures illustrated in this appendix are from the results section representing the confusion matrix, cross-validation test results followed by extracted rules from decision trees.

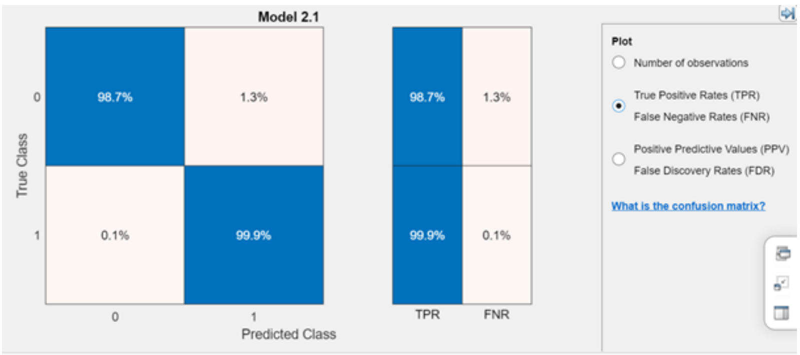


Figure C.1: Confusion matrix with percentage for exported model of the training dataset.

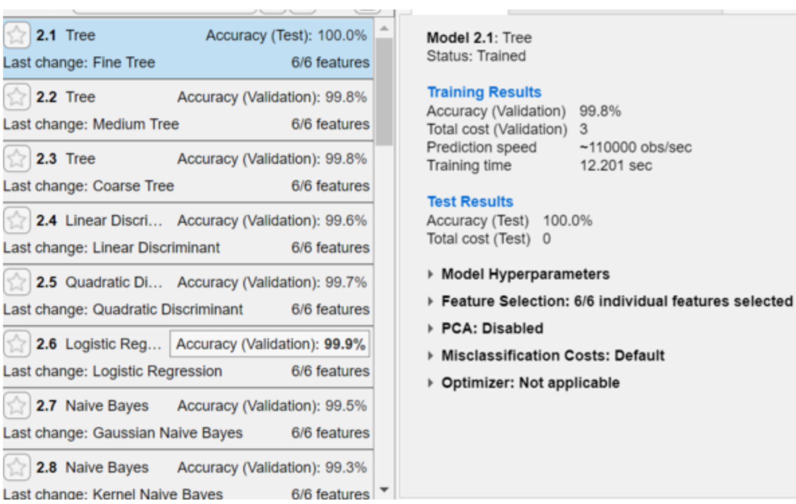


Figure C.2: Testing resuls of the exported model with cross-validation dataset.

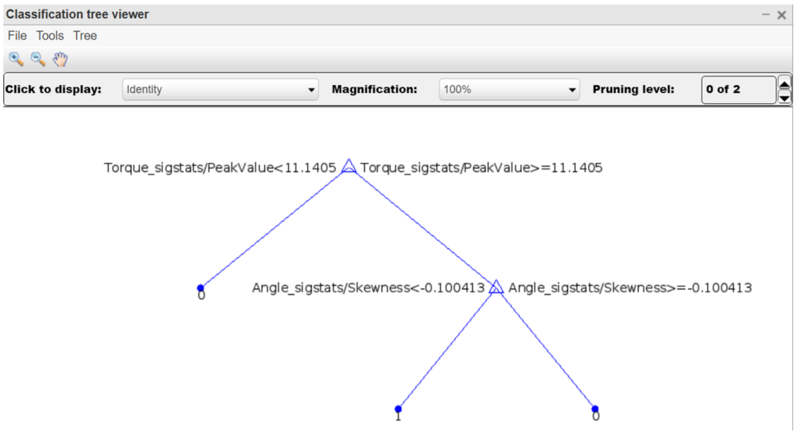


Figure C.3: Extracted rules for the exported model using decision trees.

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

CHALMERS
UNIVERSITY OF TECHNOLOGY