A systematic data science approach towards predictive maintenance application in manufacturing industry

Master's thesis in Product Development

ADARSH VASUDEVAN
XINJIE DUAN

DEPARTMENT OF INDUSTRIAL AND MATERIALS SCIENCE
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

Master's thesis 2021

A systematic data science approach towards predictive maintenance application in manufacturing industry

ADARSH VASUDEVAN
XINJIE DUAN

Department of Industrial and Materials Science
Division of of Product Development
Chalmers University of Technology
Gothenburg, Sweden 2021

A systematic data science approach towards predictive maintenance application in manufacturing industry

ADARSH VASUDEVAN & XINJIE DUAN

Supervisor: Kanika Gandhi, Volvo Group
Ebru Turanoglu Bekar, Chalmers University of Technology

Examiner: Anders Skoogh, Department of Industrial and Materials Science

Master's Thesis 2021
Department of Industrial and Materials Science
Division of of Product Development
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

A systematic data science approach towards predictive maintenance application in manufacturing industry
ADARSH VASUDEVAN & XINJIE DUAN
Department of Industrial and Materials Science
Chalmers University of Technology

Abstract

The use of data and obtaining insights from the data through data science approaches is invading every domain of industrial applications. Today, most of the industries harness the power of data to assess their performance and also to find any possible chance for improvement. This thesis project explores the data driven approach in order to perform a diagnostics and predictive analysis on the industrial alarm data collected from Volvo Group, using a systematic data scientific methodology.

The company is currently facing with an issue of having machine stoppages due to the triggering of alarms, for which the cause is unknown. The machine performs a process of Part A1 and Part A2 used in the engines of the trucks. Therefore, the company is having challenge in finding the root cause of the alarms and also to get insights into correlation of alarms and the product types being processed in the machines, which is Part A1 and Part A2. Therefore, this thesis aims to find the combination of Part A1 arm and Part A2, which produces or triggers more alarms in the specified machine chosen for the study. CRISP-DM methodology is been followed to perform the data driven approach and thereby to answer the research questions put forward in the earlier stages of the thesis. The data is obtained from 2 sources, namely Manufacturing execution system and process data which is then integrated to perform the further analysis in which exploratory data analysis is performed and the list of Part A1 - Part A2 combination which triggers more alarms are been found out. Along with this, the behaviour of problematic alarms are analysed and more insights are obtained and reported. After the exploratory data analysis, the predictive analysis is performed using machine learning models in which the multi class classification models are generated using different machine learning models wherein the model predicted the alarm category based on the selected input variables used in the modelling. The decision tree model is selected based on the accuracy score and its corresponding rules are derived and the rule with higher accuracy is chosen to provide as a decision support to the company. As a practical contribution of this thesis, it is proposed that more historical data can be used for the analysis and also the data from the quality department can be used to explore the possible in-depth root causes. Finally, this thesis introduced the novelty of implementing predictive modelling in industrial alarm problems and hence produced an academic contribution.

Keywords: Exploratory data analysis (EDA), Machine learning, Multi-class classification, Maintenance, Alarm management

Acknowledgements

We are deeply grateful for all the support and assistance that we received all throughout the thesis project. First of all, we would like to thank our industrial supervisor at Volvo Group, Kanika Gandhi, whose expertise was vital in providing directions for the thesis. Her insightful reflections helped us to explore more areas in this thesis. Secondly, we would like to thank our academic supervisor, Ebru Turanoglu Bekar for the tremendous help and guidance which enhanced our confidence to pursue the thesis. Her constant encouragement helped us to maintain our motivation even in tough times. Thirdly, we would like to thank our examiner, Anders Skoogh for providing us with new ideas to implement the research in the thesis. Finally, we would like to thank all our friends and family for their well wishes for the thesis.

Adarsh Vasudevan & Xinjie Duan
Gothenberg, May 2021

Contents

List of Figures

List of Tables

1 Introduction

In this chapter, the background of thesis is introduced explaining why predictive maintenance is an interesting topic. Followed by the aim and the research questions formulated. Further, a problem description was given for the reader to better understanding the problem. Finally, the scope and limitations of the thesis were presented in the end.

1.1 Background

With the new state of the art, data science has experienced an immense influx into various industrial applications on a broad front. Today, we can see how data science is used in health-care, customer service, government, cyber-security, engineering, aerospace, and other industrial applications. Among these, manufacturing has grown in importance to achieve a simple goal: reducing production disturbances. Over the last 100 years, manufacturing has undergone four major industrial revolutions. We are currently going through the fourth Industrial Revolution where data is collected from machines, the environment, and products to get closer and make the right decision at the right time.

The data collection is the part of a predictive approach where the focus has been mainly on real time data due to its benefits in productivity improvement, minimizing unplanned down-times, and increased efficiency of production systems. The application of machine learning techniques is emerging and the current focus is to include more extensive data sets and use a more predictive mind-set under a challenge, which is the interpretation of data to train the algorithms effectively.

The recent researches indicate that the implementation of predictive maintenance practices is widely accepted among the manufacturing industries. Predictive maintenance had gained its acceptance in manufacturing due to its ability to predict the machine failures by implementing machine learning algorithms and thereby reducing the maintenance costs. Machine learning techniques had gained increased acceptance in predictive maintenance and the research trend shows that the paper publications of machine learning in predictive maintenance increased from 0.5 percentage per year in 2009 – 2011 to 11.3 percentage per year in 2013-2018 (Carvalho et al., 2019). The predictions from machine learning thus helps to provide a clear planning for the maintenance team and as the paper (Carvalho et al., 2019) states predictive maintenance had evolved into a novel tool for monitoring and controlling maintenance events.

The project PACA (Predictive Maintenance using Advanced Cluster Analysis), which is an on-going research project in the Department of Industrial and Materials Science at the Chalmers University of Technology, aims to develop predictive algorithms to increase precision and make the data understandable to decision-makers in maintenance department. Within the scope of PACA project, data from real-world cases in the industry will be collected from several streams and machines to identify interesting patterns. The analysed patterns are then compared with historical data to develop an understanding of how the patterns relate to productivity losses through the use of data science and AI techniques. One of the industrial case study involved in this project has been carried out together with Volvo Group, which will be the focus of this thesis project.

1.2 Aim

The aim of the project is to use a systematic data science approach to investigate the relationships between station alarms and the product type being processed by performing exploratory data analysis and devising machine learning models which enables proper planning of maintenance activities through predictive modelling.

1.3 Research Questions

Knowing the aim of the thesis and the requirements of industrial case study at hand, the following research questions are specified for investigation during the thesis:

RQ 1: What product variants or combination of variants trigger more alarms?

RQ 2: What kind of machine learning models can be developed for predictive analysis of alarms which enables to explore the possible in-depth root causes in production and maintenance?

1.4 Problem Description

A good production line is vital for the manufacturing products to ensure the timely demands of the market, thereby constituting towards the growth of the organisation. The output of the production in a factory is aligned towards the proper and uninterrupted functioning of the collective machines involved in the entire production line. Any stoppages or failure of machines in the production line leads to loss in production time and also causes economic adversities for the organisation (Mehmeti, Mehmeti, & Sejdiu, 2018).

Volvo Group is one of the most valuable companies in the world, which is renowned for their capability in manufacturing heavy trucks across the world. Over the years, Volvo Group had pioneered in their state-of-the-art manufacturing of trucks. Volvo Group had addressed the problem of having machine stoppages during the production, for which the cause is unknown. This thesis project explores and investigate the underlying cause of these machine stoppages in the Volvo Group's manufacturing plant . The machines in the manufacturing plant had been equipped with sensors and also various alarms are been incorporated to each machines to determine the irregularities occurring in the production process based on alarm conditions which are categorised as good and bad alarms. The good alarms are the one which does not hinders the production process, while bad alarms is the one which put halt to the production. The case here is to understand the relationship between station alarms and the type of product being processed and to come with an exploratory analysis to find the root cause of the bad alarms occurring in the machines throughout the production line. The possible chances of triggering bad alarms can also be due to the products from the external suppliers or from the outputs of internal processes during the manufacturing operations inside the production plant. The proposed method of analysis is to have a data driven exploration and therefore the data from manufacturing execution system and process data will be used to analyse the cause of bad alarms that hinders the production process.

The production process in this thesis will be a process of Part A1 and Part A2 on Part A. Part A1 is an oscillating lever which is used to convert the radial movement from the Part A2 to linear movement, which facilitates opening and closing of the poppet valves. One end of the Part A1 is connected to the poppet valve and the other end to the lobe in the Part A2. The opening and closing of the poppet valves in Part A are triggered by the corresponding movement of the Part A2's lobe, i.e., the Part A1 opens the poppet valve when the Part A2's lobe raises the other end of the Part A1 and vice-versa. Part A2 is a cylindrical rod which contains lobes of same number as the valves in the engine. The radial movement of the Part A2 enables the opening and closing of the intake and exhaust valves in an engine.

1.5 Scope and limitations

The scope of the project as mentioned before is to do the exploratory data analysis about the correlation and using the existing approach to generate machine learning models in order to predict the occurrence of the alarms. This will enable to identify those alarms that lead to the breakdown of the specific stations, and helps forecasting whether the product variant will cause the alarms or not in the future.

The machines used for study are limited to two selected machines located in the production line in the manufacturing plant. The data sources using here were the higher level data source instead of signal or sensor data as entries including events log data from manufacturing execution system (MES) and process data. They contain all the information related to the production history such as alarm type, alarm count, production machine number, part number, and many more which are useful for data mining and exploration phase. Hence these event log data will be used for the finding the correlations between the alarm ID's and the parts produced in the manufacturing plant.

In the end, the models will be evaluated by using performance matrix and business success criteria, they should be seen as suggestions and decision support for the production line. The deliveries will also be a way to assess if using the event log data is a reliable source to conduct predictive maintenance strategies or not. However, deployment of the application is out of scope for this project.

Due to the circumstance of COVID-19, the project is conducted remotely, so that quality department's data is not possible to get because of confidentiality. Another influence is that the data deliveries were delayed so the report was finished in a short amount of time.

2

Literature Study

This chapter reports the insights obtained from the literature study performed for the thesis project. The literature review chapter is divided into 4 sub-sections namely Industry 4.0 and Digitalization, Artificial intelligence in Manufacturing, Maintenance and Machine learning models used in the thesis. The literature review is done in a top-down approach starting with Industry 4.0 and digitalization, and then followed by the corresponding sections in order to create a good reading experience.

2.1 Industry 4.0 and Digitalization

The modern world is continually subjected to rapid developments in every industrial domain, which are made possible by the invention of novel technologies in various field of engineering. Over the past few centuries, the technology had evolved and the advent of novel technologies had led to different phases of industrial development, which is known as Industrial revolutions (Gustavo Dalmarco, 2019). As time progresses, the customer demands had been increasing and the industries had to keep up with large amount of supplies to ensure the demand, by incorporating new technologies into their system. As a result, there have been 4 major industrial revolutions, namely industry 1.0, industry 2.0, industry 3.0 and the current industrial revolution, which is industry 4.0 (Yong Yin & Li., 2018). The driving force or the technology innovations behind each of the 4 industrial revolutions in chronological order are Mechanization, Electrification, Computerization and Digitalization (Gustavo Dalmarco, 2019). The current industrial revolution (Industry 4.0), is aimed to develop an efficient system to comprehend and adapt to the demand for customised and high quality products, faster delivery times and to bring in automated processes to increase the efficiency of the industry. Manufacturing industry is one of the most impacted areas by the industry 4.0; the recent developments in manufacturing industry had proven its capability in enhancing the production processes and had achieved optimization within the operational processes by adopting industry 4.0 practices. The digitalization of industry is achieved through innovations in various technologies such as internet of things (IoT), Cyber-physical systems, Big Data Analytics, Cloud technologies, Artificial Intelligence, Simulation and modelling, Visualization technologies, automation and industrial robots, Additive manufacturing (Ting Zheng, 2020).

2.2 Artificial Intelligence in Manufacturing

This thesis project explores the usage of artificial intelligence technologies in the industry 4.0, especially in manufacturing domain and therefore the usage of Artificial intelligence in manufacturing will be given more importance and widely discussed. Manufacturing industries are in a mission to adapt to the volatile customer demands and in also prioritising to have an edge over their competitor by reducing the time to market, hence to increase the business value. Therefore, more firms are adopting to change their traditional way of manufacturing in order to increase their productivity by smart factory implementation. Artificial intelligence provides the pathway to implement smart factories, which are capable of adapting to the dynamic conditions and in increasing the productivity of the firm, thereby enhancing the business(David R. Sjödin & Petrovic, 2018). According to IBM, “Artificial Intelligence enables computers and machines to mimic the perception, learning, problem solving and decision making capabilities of the human mind”(Education, 2020). The Artificial intelligence is widely used in the manufacturing field and is contributing towards the development of a novel model of manufacturing, which is termed as intelligent manufacturing. Intelligent Manufacturing enables the integration of all the technologies related to the product such as Manufacturing technology, communication technology, system engineering technology and more with the entire development cycle of the product. The Artificial intelligence when combined with the intelligent manufacturing systems, which are made possible by the Internet of Things (IoT) drives the industry towards intelligent manufacturing. An intelligent manufacturing system is identified as the system which is capable of performing intelligent sensing, being cognitive, provides better decision making, managing the machine-human-material environment etc (Bo hu LI, 2017). It brings in integration with all the connected technologies throughout the product life-cycle.

The intelligent manufacturing generates large amount of data from these connected technologies inside the industry production line and these data are of great value to the decision making. The data generated needs to be assessed using certain methods in order to obtain insights, which can lead to a strong foundation for the necessary actions to be performed in increasing the productivity. Data Science along with the Artificial Intelligence approaches are been widely used to explore the large data set generated in order to gather insights from the same in order to train the machines to act and learn from experience. Data science aims towards obtaining valuable information from the data which are mainly categorised by 3 V’s (Volume, Velocity and Variety). The volume of the data represents the quantity or amount of data generated, velocity of the data represents the speed at which the data is generated and the variety of the data represents the mix of different types of data obtained from the data collection (Bushra, 2020). Data Science provides a better medium for converting the raw data which lacks data quality into analysis ready data which can be further used to obtain insights for decision making.

Artificial Intelligence is then used to achieve or mimic human like actions or decision making to the machines by using the different methods such as machine learning or deep learning. According to IBM “Machine learning focuses on applications that learn from experience and improve their decision making or predictive accuracy over time” (cloud education, 2020). Recent researches had proven that the application of Machine Learning in manufacturing results in many benefits such as maintenance cost reduction, machine fault reduction, enhancement of operator safety, increased production, to name a few. Therefore, the machine learning is of great importance in analysing the data which holds valuable information and knowledge, which can be used to enhance the productivity of the production process in the industry (Zeki Murat Çınar & Safaei, 2020).

2.3 Maintenance

In product realization life cycle, especially under a sustainable manufacturing context, maintenance has started shifting its role from “merely values” or “just repair works” to controlling the product’s condition with regards to the product’s physical and functional life (Takata et al., 2004).

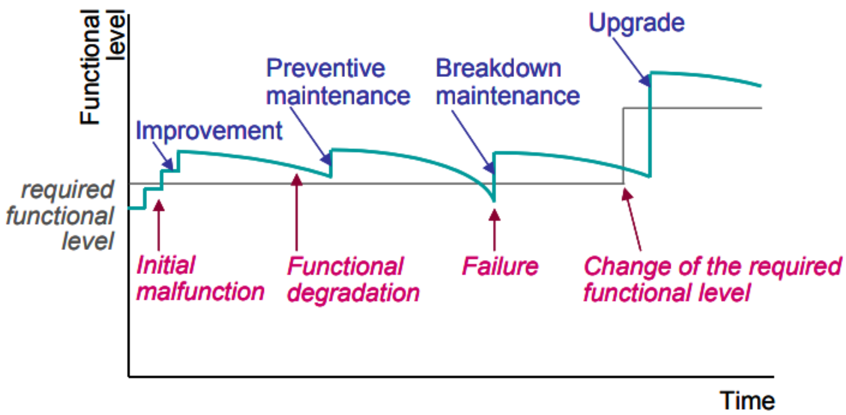


Figure 2.1: Maintenance activities (Takata et al., 2004)

For most companies that have established their industrial system, refrain from making investments in new plants as long as the current works safely and efficiently. Maintenance, as a consequence, has become one of the main impacts contributing to maximizing productivity under the circumstance (Nakagawa, 2006). As (Mobley, 2002) points out, “Depending on the specific industry, maintenance costs can represent between 15 and 60 percent of the cost of goods produced”. Heavy industries spending cost up to 60 percent in this case. Hence, effective maintenance has proven to be essential to many operations, not only for reducing equipment downtime, but also to optimize overall availability or reliability of systems, minimizing costs, and multi-objective optimizations (Ran, Zhou, Lin, Wen, & Deng, 2019).

2.3.1 Maintenance strategies

Different strategies for maintenance have been developed and implemented. Various terms and categories have brought up by different researchers so far. Susto (Susto, Schirru, Pampuri, McLoone, & Beghi, 2014) proposed three main categories, which are run-to-failure, preventive maintenance, and predictive maintenance.

- **Run-to-Failure:** As known as reactive maintenance, or corrective maintenance, is the most straight-forward yet least efficient strategy where the interventions only happened after the failure occurs. Hence, adding unnecessary cost to the maintenance.
- **Preventive maintenance:** Preventive or scheduled maintenance, is a time-based strategy that schedule the maintenance performance periodically. However sometimes with an imprecise frequency which causes the waste of resources.
- **Predictive maintenance:** Sometimes was also referred to condition-based maintenance, anticipate the potential needs to trigger the maintenance actions accurately.

2.3.2 Predictive maintenance

With the advent of digitalization, a great deal of previous research into maintenance strategies has focused on predictive maintenance which helps the plant proactively improve the availability and reliability of manufacturing systems, extending the useful lifespan of equipment, and enhancing the quality of products (He, Gu, Chen, & Han, 2017). Predictive maintenance, to an extent, averted early interventions or late remediation which respectively wasting resources and the possibilities of causing irreversible failures (Jimenez, Schwartz, Vingerhoeds, Grabot, & Salaün, 2020).

There are three main approaches for predictive maintenance, covered by knowledge-based models, data-driven models, and physics-based models. Extensive research has been carried out for each of these single-model approaches, however, for the complex systems, the hybrid models based on the combinations of these three could improve accuracy for solving the tasks (Jimenez et al., 2020).

- **Knowledge-based models:** A traditional approach based on the knowledge of domain experts so as to build a corresponding PdM system. Knowledge-based models are explicative comparing to other approaches (Jimenez et al., 2020). For instance, data selection criteria, how much data has to be extracted from the stations can be decided after domain analysis.
- **Data-driven models:** Nowadays, massive data has been created and stored in the database. For the purpose of finding the pattern of system behavior, data-driven approach have been widely perceived as a powerful solution (Ran et al., 2019). Mostly useful when solving a complicated problem.
- **Physics-based models:** Perform accurate simulations to identify the system behavior on account of an accurate model of the physical behavior (Jimenez et al., 2020)

2.3.2.1 Predictive maintenance challenges

Some challenges of predictive maintenance involve the initial setup (Sakib & Wuest, 2018), how to guarantee the representativeness of the dataset (Fink, 2020), or how to structure and integrate multisource heterogeneous data (Yan, Meng, Lu, & Li, 2017), etc. Nonetheless, performing predictive maintenance strategy urge for making best maintenance decisions, improving production efficiency and the company competitiveness on account of the extraction of experiences from domain experts based on the available resources from the system, associated with the exploitation of advanced data analysis methods (Kalathas & Papoutsidakis, 2021). This is where exploratory data analysis comes in handy. It is an essential first step in understanding the variation of the data set (Diggle, Chetwynd, & Chetwynd, 2011). The exploratory data analysis enable searching for patterns and trends from huge volume of data, and visualization techniques play an important part in this quest (Skiena, 2017). Hence, exploratory data analysis is also root of our analysis in this thesis work,

2.3.3 Data-driven approach

As a trendy approach under predictive maintenance, the data-driven approach, as the name shows, is based on the large amount of data provided. As a result of the exponential growth of available information and data, data-driven decision-making process has become more crucial. In order to apply predictive maintenance in regards of data-driven approach, the basic prerequisite involved a robust condition monitoring system with effective sensors to capture the data while some of the measurements can be directly acquired from the control system, with a format either in a continuous state or at discrete points of time (Fink, 2020).

(An, Kim, & Choi, 2015) presented a state-of-the-art review stating that the data-driven approach is consisting of the statistical approach and the artificial intelligence approach. Statistics and probability theory are the basis of data-driven models (Mobley, 2002). In 90s, some famous statistical approaches, Markov process and Bayesian method have been created to study the degradation of the components, estimated failure rate (Sakib & Wuest, 2018). Later in the 2010s, artificial intelligence approach has grown rapidly where machine learning is a major part of, showing great potential in fault detection and diagnosis, monitoring current health state of the system, as well as remaining useful life (Zhang, Yang, & Wang, 2019). Providing with the extensive extracted data, a great variety of applications can be achieved. Generally, different approaches have their own advantages. Researches such as (Baptista et al., 2018) have studied the integration of machine learning model approach and statistical model approach, the result shows the combined models outperform the traditional one.

2.3.4 Related work - Data-driven study based on Industrial Alarms and Event Log Data

The previous studies performed on devising maintenance strategies by analysing industrial alarm and event log data were analysed and is discussed below. The Literature study performed on the industrial case study related to alarm and event log data indicates more research has been focused on descriptive and diagnostics approach of alarms and event log data. According to (Gröger, 2018), the descriptive analysis emphasizes transparency and provides description of the historical data set. In other words, the descriptive analysis provides assistance in understanding what has happened or what is the problem, by analysing the existing data. The diagnostics analysis aims in finding the root cause of the problem through performing exploratory analysis. Considering the case of an industrial alarm data, the diagnostics analysis provides details to the operators regarding what had caused or triggered the alarms which resulted in the process deviation (Vásquez, Travé-Massuyès, Subias, Jimenez, & Agudelo, 2016).

According to the paper proposed by the authors (Tidriri, Chatti, Verron, & Tiplica, 2016), fault detection and diagnosis in industries had been actively researched over the recent years. This above mentioned paper studied the performance of model-based and data driven approaches on process fault diagnosis. The data driven approach performs the fault detection and diagnostics as classification tasks and uses both supervised and unsupervised classification methods. The paper discusses the use of Bayesian networks and Artificial neural networks in fault detection and diagnosis. The model based approach for fault detection and diagnosis were also discussed and the combination of different data driven and model based approaches were analysed and compared. Another paper proposed by the authors (Mahadevan & Shah, 2009) discusses the implementation of Support vector machines (SVM) in performing fault detection and diagnosis. The fault detection is performed on the basis of conventional method, where the fault is detected based on the threshold value set. If the value exceeds the threshold, a fault is registered or detected. Whereas the fault diagnosis was done by performing feature selection in order to investigate the variables that are related to the normal and faulty category classification.

The above discussed papers consider an approach towards fault detection and diagnosis without exploring the benefits of integrating alarms into fault detection and diagnosis. According to the paper proposed by the authors (Lucke, Chioua, Grimholt, Hollender, & Thornhill, 2020), alarm management and fault detection diagnosis (FDD) disciplines have close relations but are still separated in an industry. Their study also points out that the alarm data had been rarely used to perform the fault detection and diagnosis in an industry and also mentions the benefits of using alarm data in fault detection and analysis, by analysing the recent works and also by producing a case example from an oil gas separation plant.

The literature paper proposed by (Bezerra et al., 2019) also discussed the importance of analysing the alarm and event log data in an industry and performed a diagnostic analysis on the event log and alarm data from a petrochemical plant. The paper proposed by (Bezerra et al., 2019) further highlights the importance of analysis on the alarm data. The alarms are incorporated in industries to analyse the possible process deviation and malfunctioning of equipment, which when not managed properly affects the overall performance of the production and also results in the decrease in quality of the final products. Hence, investigating the root cause of the alarms is a major step to devise relevant strategies to tackle the process deviations or equipment malfunctioning. The paper proposed by (Bezerra et al., 2019) conducts root cause analysis through performing exploratory data analysis. Exploratory data analysis delivers better insights about the data by providing the overall numerical and graphical summary of the data and in preparing the data for further modelling (Shmueli et al., 2010).

The importance of fault detection and diagnosis in industry has been highly recognised and is undergoing further research. Although the fault detection and diagnosis helps in ensuring plant safety and product throughput (Mahadevan & Shah, 2009), it is important to further expand the analysis into a predictive approach in order to create a futuristic model which helps in predicting the problem based on the modelling performed with available data. The paper proposed by the authors (Bousdekis, Lepenioti, Apostolou, & Mentzas, 2019) discussed the various research performed in the maintenance domain and identified the research gap. This paper proposes various future research recommendations and one of the recommendations suggests in developing generic decision models which represents the decision making process. It also mentions creating generic decision models which are applicable to manufacturing equipment by employing prescriptive analytics, artificial intelligence and machine learning algorithms on the historical maintenance related data sets. Furthermore, the importance of research on developing data driven techniques for constructing decision models is also mentioned by highlighting the current trend of researches which focus on building decision models to enable less human intervention in decision making, with making use of big data analytics.

From the knowledge obtained from the literature study, this thesis project contributes to research on industrial alarm data by implementing a predictive analysis to develop the generic decision model that represents the decision making, which is a research gap in industrial alarm data problems as mentioned by (Bousdekis et al., 2019).

2.4 Machine learning models used in the thesis

As mentioned earlier in Chapter 2.1, machine learning, as a branch of artificial intelligence, has became one of the most powerful tool for predictive maintenance having the ability to handle high dimensional data. Machine learning models had been categorised into 3 models (Jimenez et al., 2020) which are as follows -

- **Supervised Learning :** Supervised learning are type of machine learning model used when the data under study is a labelled data set, i.e., the algorithm train the model based on the input-output pairs. Regression and classification machine learning method falls under supervised learning.

- **Unsupervised Learning :** Unsupervised machine learning model use unlabelled data and the algorithm trains the model to find the or detect the patterns or similarities which was not previously identified. Clustering machine learning method comes under unsupervised learning.

- **Reinforcement Learning :** Reinforcement machine learning model uses the algorithm to train the model based on the experience obtained from performing actions.The reinforcement learning have 3 components namely agent, environment and actions. Thus the reinforcement learning trains the model based on the experience obtained by the agent after performing an action in the given environment.

Since the thesis project aims to find the combination of Part A1 and Part A2 triggering the alarms, the proposed approach is to perform and develop a multi class or multinomial classification machine learning model. Hence, the some of the machine learning classification algorithms will be mentioned and briefly discussed in the next section.

2.4.1 K-Nearest Neighbors (KNN)

K-Nearest Neighbors or KNN is a supervised machine learning algorithm which is used to perform the classification as well as regression on data. As the name suggests, KNN algorithm groups or classify a data point based on the nearest data points, thus, k value became an important parameter to determine training result, i.e., if the k value is set as 7, the algorithm will group the new data point by having a majority voting with the 7 nearest data points and classify the new data points to the label which has majority among the 7 nearest data points. To determine the k value, we need to understand how the distance was being measured. Euclidean distance is the most well known choices to measure the distance between points. A and B here are points represented by feature vectors where m is the dimensionality of the feature space (Hu, Huang, Ke, & Tsai, 2016).

$$dist(A, B) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{m}} \tag{2.1}$$

In Scikit-learn, k value can be set with parameter $n\_neighbors$. Or it's also worthwhile to use *GridSearchCV* function to find the optimum k value. Though KNN is relatively memory inefficient, it is suitable for performing multinomial classification and is relatively easy to implement and interpret (Yildirim, 2020).

2.4.2 Decision Tree

Decision tree is a supervised machine learning algorithm, which is used to execute classification and regression problems. A decision tree represents a tree like structure and is explained by defining its decision nodes and leaf, wherein the nodes represents the place at which the data is split and the leaves represents the final results of the decision, see Fig 2.2. At each node, a certain condition is checked and its corresponding output is recorded and further split is made accordingly until the final split is performed (Gupta, 2017). With that being said, by given a labeled dataset, a tree learns from a sequence of if-else statement about individual features to infer the labels. Compared to rules based systems which is based on manually set thresholds and experience, machine learning algorithm is able to have a general better performance that is combining with rules. They can produces a probability, rather than only binary scores.

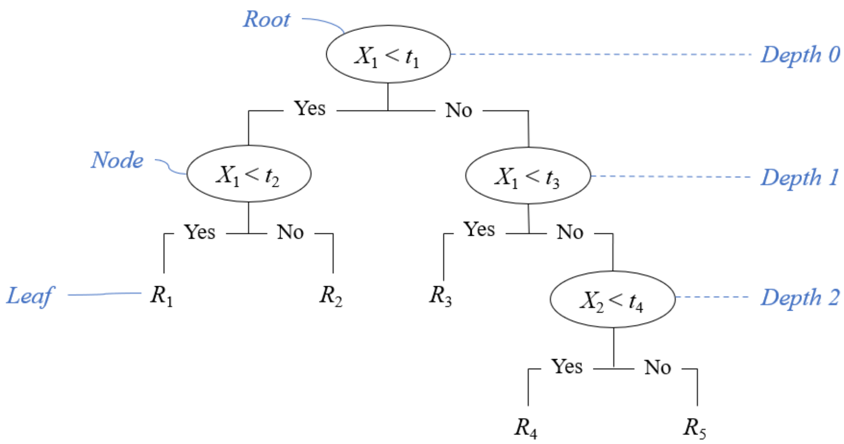


Figure 2.2: Sample binary trees adapted from (Hastie et al., 2009) and (Safavian & Landgrebe, 1991)

Among all the different tree methods, an optimized version of Classification and Regression Tree (CART) was used and the tool for implementing the method is Python with Scikit-learn package. To produce the purest leafs possible, which means bringing the result clearer to a decision, the measurement regarding the impurity of the node can be evaluated by Gini index or Information Gain. For CART, the formula to minimize Gini index is shown as follows:

$$G(p) \;=\; 1 - \sum_{i}^{j} (p_i)^2 \tag{2.2}$$

where p_i represents the probability of a node i being classified as a specific category (Skiena, 2017). Decision tree is suitable for multinomial classification tasks and also is easy to interpret. Apart from this, the performance of decision tree is not affected by the presence of any non-linear relationship between the parameters and is adaptable to both numerical and categorical variables.(Gupta, 2017)

2.4.3 Random Forest

Random Forest is defined as the collection of individual decision trees which forms an ensemble learning method. Random forest is a supervised machine learning model which is used to perform both regression and classification. Though random forest is a collection of individual decision trees, but it is more effective in decision making since random forest predicts the output by considering a majority vote from all the output from individual decision trees (Yiu, 2019). With only decision tree, it is easily to be over-fitted. Whereas, with each decision tree generates its own prediction of the class and the random forest algorithm will take in account of the majority value among the prediction given by decision tree.

2.4.4 XGboost

One of another ensemble method is called gradient boosting, the idea is to build weak models at first, and use the features to build a stronger model. Extreme Gradient Boosting as known as XGBoost, aims at maximize the speed and efficiency, is also based on this framework.

When solving the problem, the data used in real world practice always facing some quality issues, one of them will be how to handle missing values. All the previous algorithms written before does not accept missing values as ready for modelling. While with XGboost, the instance is classified into the default direction when a value is missing (Chen & Guestrin, 2016). Another advantage described in this paper (Chen & Guestrin, 2016) is that normal gradient boosting decision tree usually perform badly in dealing with sparse input. XGboost can, to an extent, solve the sparsity for instance due to one-hot encoding.

3
Methodology

This chapter is comprised of the methodology which was used to perform the project. Since the project deals with exploration and modelling of the data acquired and given from the industry, it was then approached with a systematic way of data mining method used in the industries. The methodology which was used in this project does not explore the data acquisition phase, which was already done by the industry.

The standard cross-industry data mining process, also known as the CRISP-DM process, is used to conduct this study. It provides a standard framework for planning and managing the project and is independent of the technology and the sectors of industry involved in the project (Wirth & Hipp, 2000). CRISP-DM divides the process of data mining into six main phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The CRISP-DM process can be iterative and the phases also develops certain dependency within each other to ensure that the data is relevant and the results of the process aligns with the business objective of the industry defined in business understanding phase of the process. The phases which constitutes CRISP-DM process are described in detail below.

3.1 Business Understanding

Business understanding phase is the core step which determines the success of the data mining project. It includes formulation of the project goal from a business perspective, analysis of objectives and success criteria based on the industrial case study is carried out in this step. These decisions are made with consideration of all stakeholders to understand the business objective and articulate the expectations for the project. The project goal is transformed into a data-driven problem definition, and a preliminary plan is created to achieve the goal.

3.2 Data Understanding

The study will be conducted on the project application process to understand if the available data of the process is sufficient to achieve the business objective or if additional data is required. This phase includes *Collect Initial Data*, *Describe Data*, *Explore Data*, and *Verify Data Quality* (Chapman et al., 2000). Explore data will be discussed in the subsection. Inspection for data quality, which incorporate checking for data completeness, the distribution of data and many more to ensure the relevancy of the data to achieve business objectives. This study allows becoming familiar with the data and gaining initial insight into the data to form a subset of interest for hypothesis testing.

3.2.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a promising data-driven approach to pave industrial alarm and event analysis (Bezerra et al., 2019). It is used to extract hidden pattern and reveals undiscovered relationship without making assumptions (Mao et al., 2015). From here, we will start answering our first research question, exploring the relationship between combinations of products and alarms.

The starting point to implement EDA is using descriptive data summarization and visualization techniques. Descriptive data summarization illustrate basic information of the data by using summary statistics, pairwise correlations, or class breakdowns (Skiena, 2017). Summary statistics covers the basic information like mean, count, or sum. They are basic statistic results but are generally very informative about the data distribution, and it will be quite useful to see the pattern of the alarm's data. There are also other measurements such as central tendency, dispersion, quantiles etc, but will not suit our thesis purpose.

Visualization, on the other hand, is especially important in converting the data to a format that is easy to understand and heighten communication that enables effective action (Subramaniyan et al., 2020). There are extensive tools for visualization. The visualization tools used in this project will be python and Tableau with histograms, box plots, and scatter plots as main visualization techniques.The python programming language is chosen because its easy to use, open source programming language and also its well compatible with all types of files, in this case, the excel file. The python programming is executed in Google colab environment, to ensure proper on-time coordination within the pair. Google colab comes in with all the necessary libraries and it thus helps to work on various libraries without even downloading it to the system. Tableau is chosen to perform data visualization due to its easy user interface and quality visualization. Tableau for desktop is downloaded for this project.

3.3 Data Preparation

This phase comprises of tasks involved to generate a data set that will be used to build the model. This is an iterative process that is likely to be performed multiple times and not in a prescribed order. In these steps, gap analysis, data transformation, and cleaning are performed to improve the quality of data, before feeding the data for model generation. This phase is important, when it comes to the modelling phase the performance of the model depends on the data prepared in this phase.

3.4 Modelling

The modelling phase comprises of the selection of various modelling techniques and applying it to the model. Here, some assumptions are been made which supports to the selection of the modelling technique used. Once the modelling technique is selected and applied, the model is then tested for analysing the performance and the test design in generated. Model's parameters are tuned to improve the performance of the model. However, there are certain requirements on the dataset to improve the performance. Therefore, back and forth data preparation and modelling is necessary to realize a good model. Finally the model assessment is performed in order to summarize the review of the model, i.e., if there are 2 different models modelled using different techniques, the assessment provides the comparison of performance of these models. After trying and testing in the modelling part, we can now answer the second research question, what kind of machine learning models can be developed for predictive analysis of alarms which enables to explore the possible in-depth root causes in production and maintenance.

3.5 Evaluation

In this phase, the generated model is evaluated with business objectives defined in the business understanding phase. The evaluation helps to summarize the models capability to answer or to satisfy the success criteria of the business. The model will be organized and presented in such a way that the partner can make the best use of it. The main objective at this stage will be to evaluate if any important consideration has been omitted that contributes to the business objective. If the model does not satisfy the business objectives, it is then iterated and a new process will start from the Data understanding to the evaluation of the model, until it meet the business objectives.

3.6 Deployment

The deployment plan is created which is used to install or deploy the model in to the organisation. Continuous monitoring plan is also created to ensure proper inspection on the data mining process in order to ensure error free working of the model. The final part is to generate the report of the whole process, which includes all the necessary up to date results of the project and a final presentation which conveys the whole project to the organisation.

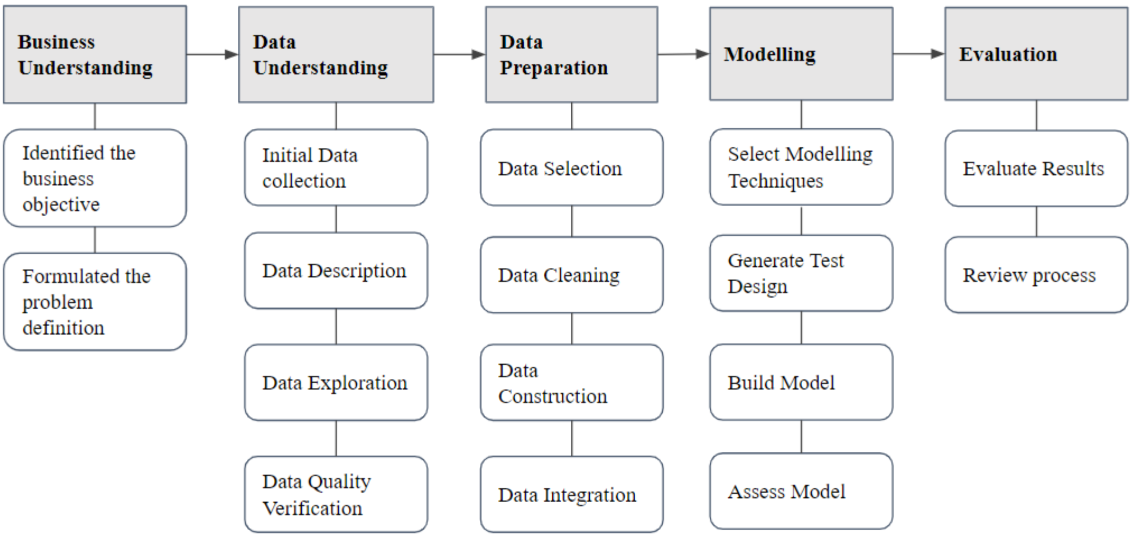


Figure 3.1: Methodology flow chart

4

Results

This chapter is divided into 6 sections illustrating the outcomes and insights generated from different phases of the CRISP-DM methodology. The section 4.1 describes the business understanding phase, which provides insights into the business objective of the project and thereby formulating a problem definition. Section 4.2 describes about the data understanding and discuss the results of data collection,data description, data exploration and data quality verification. Section 4.3 describes the Data preparation phase, where the results of data selection, data cleaning, data integration and data formatting will be discussed. Section 4.4 illustrates the finding obtained from the exploratory data analysis. Section 4.5 and 4.6 discusses the modelling phase and the evaluation phase.

4.1 Business Understanding

Understanding the business problem and defining a business objective is the first step or phase in CRISP-DM. The business objective for this project was understood from the insights obtained from the meetings with the stakeholder (Company Supervisor). The company aims to find the cause which triggers the alarms in the specified machine and thereby devising plans to tackle these alarms which results in reducing the overall production halt duration. Since the machine under study for this project performs a process on Part A1 and Part A2, main focus is to find the Part A1-Part A2 combination which triggers more number of alarms. Therefore, the business objective of the project is to reduce the production halt time by investigating the cause for the occurrence of alarms, through performing an exploratory data analysis. With the business objective been defined, it is now important to state formulate the problem definition, which is to perform a diagnostic approach on the alarms in the specified machines under study and to investigate on the combinations of Part A1-Part A2 which triggers more number of alarms in the machines and to devise a classification machine learning model to predict the category of the alarms.

4.2 Data understanding

The data was collected from two sources, where one is from the process data, while the other is from manufacturing execution system (MES). Process data contains different part number, measurements, the corresponding date and time and other variables. Data from manufacturing execution system mainly provide information for the alarms. Both files are provided in an excel format and was collected during the production process. There are 5 excel data files for process data corresponding to each Part A1 and 2 excel data files for manufacturing execution system data which contains the information regarding all the alarms which had occurred during the time period of which the data was given. The general information regarding the 2 data sets are given in the Table 4.1.

SL.NO	DATA SOURCE	ROWS	COLUMNS
1	Assembly data - 7081	768	50
2	Assembly data - 7281	304464	50
3	Assembly data - 9571	1010490	50
4	Assembly data - 9611	1048573	50
5	Assembly data - 9651	131004	50
6	MES - Machine A	1264	7
7	MES - Machine B	1314	7

Table 4.1: General information of data source

The data is acquired from the two production machine used for our study, namely Machine A and Machine B, where a process of Part A1 and Part A2 takes place. The reason for choosing these two machines is because they have very narrow limits for quality check, if they get any deviation, they cannot go out from the company. The data set collected from the machine A contains 50 features and distinct number of rows for each of the Part A1. The data from the machine A was collected based on the Part A1 number (7081, 7281, 9571, 9611 and 9651), machine B based on (7281, 8211, 8231, 8431, 8771, 9571, 9611 and 9651) were given for study (provided through process data repository), which contains data collected based on each Part A1 respectively. The time period of the given data are with a range of seven months for exploring the features. The Table 4.2 below shows 18 out of 50 the properties of the process data, for the reason that the others are irrelevant to this project from the beginning according to company supervisor. Within the process data, there are different combinations of Part A1 and Part A2, all the possibilities of combinations are listed in Table 4.3. The data from manufacturing execution system contains the information about the alarms, alarm ID, time of alarm e.t.c. The manufacturing execution system data was also provided in excel format and it contains 7 attributes with 1264 and 1314 records of data respectively. The Table 4.4 below shows the properties of data from manufacturing execution system.

SL.NO	Attribute Name	Type
1	Engine Number	int64
2	Customer Number	int64
3	Receive Number	int64
4	Chasis Number	int64
5	Chasis Type	object
6	Cost Center	int64
7	Machine Number	object
8	Part A1 Number	int64
9	Part A2 Number	int64
10	Operatioin Type	int64
11	Position	int64
12	Part A Position	object
13	Measurement Number	int64
14	Measurement	object
15	Measured Value	int64
16	Date	int64
17	Time	int64
18	Result Status	object

Table 4.2: List of attributes in the process dataset

SL.NO	Attribute Name	Type
1	Location	object
2	Start Time	object
3	End Time	object
4	Alarm Duration	object
5	Alarm ID	float64
6	Alarm Text	object
7	Links	object

Table 4.4: List of attributes in manufacturing execution system dataset

The process data sets and the manufacturing execution system data set have a common medium or attribute (date and time), which is further used to develop a connection between the 2 data sets. The date and time helps to create a better understanding between the Part A1's data in the process data and the corresponding alarm data in the manufacturing execution system data . The data acquired has the necessary features to create the prediction model to satisfy the business objectives, which is to find the Part A1-Part A2 combination triggering the alarms.

Initial exploration of the two data sets (process data and manufacturing execution system data) were performed and few insights were obtained. It was obvious that the 2 data sets had to be integrated or merged to perform the required data analysis. The one possible and relevant way to integrate these data sets is to merge these data sets based on the 'Date and Time', such that the new merged data set

Part A1	Part A2	Machine A	Machine B
7081	8003	✓	
	1063	✓	
	0564	✓	
	1068	✓	
	1080	✓	
	0560	✓	
7281	1079	✓	✓
	1070	✓	✓
	8002	✓	✓
8211	0563		✓
	1066		✓
	1069		✓
	1078		✓
	8001		✓
	0561		✓
8231	1081		✓
	1083		✓
8431	1079		✓
	1070		✓
	8002		✓
	0565		✓
8771	1095		✓
	0407		✓
9571	1066	✓	✓
	1069	✓	✓
	1078	✓	✓
9611	1081	✓	✓
	1083	✓	✓
	2475	✓	✓
	5905	✓	✓
9651	1095	✓	✓
	0407	✓	✓

Table 4.3: Combinations of Part A1 and Part A2

will contain all the records in process data and manufacturing execution system data sets corresponding to each matching date and time respectively. Also after further exploration, it was found out that some of the columns or attributes are not necessary or will not make any impact for the data analysis. Out of the 24 and 7 features in process and manufacturing execution system data set respectively, only a few were relevant for the analysis and which would help to analyse and model to attain the business objective.

The data quality was verified and relevant modification were performed on the data in order to improve the quality. The first process of verifying data quality was to check for any missing values in the given data. The missing values, if present, affects the performance of the final prediction model, therefore it is mandatory to treat and fix the missing values in the data set. The alarm data from manufacturing execution system data set contained some missing values, hence these missing values can be deleted from the data using *dropna()* function in python. The missing values were very low when compared to the total records in the data and therefore, deleting the rows containing missing values will be optimum.

Apart from the missing values, the data set also had irrelevant formats for some attributes namely Date and Time attributes. In process data set, the date and timestamp columns were of type numerical (int) instead of being date format and time format, also the Date Time column in manufacturing execution system data was in ISO Date time format. The irrelevant format of these attributes had to be formatted correctly to ensure proper data analysis.

4.3 Data preparation

The initial treatment for two raw data sets were conducted in the first place. For all the attributes in process data-set, engine number, Part A1 number, Part A2 number, position and measured values are the most important characteristics to study from. Date and time are used as a medium for matching with data from manufacturing execution system. On the other hand, process centre, machine number (Machine A, Machine B), and measurement attribute were used for removing the unrelated rows one step forward. This means that the data entries has to be within the selected process centre, the rows value that are not belong to process centre under study were deleted. Depending on what machine was studied, the machine could be chosen as either Machine A or Machine B. Besides, during the production process, the three measured parameters in the raw data are representing moment, angle, and depths respectively. According to the domain expert from the company, only the moment and its corresponding values are the most valuable features to focus on. While measuring moment values, there will be a situation when some position of the specific Part A1 was not measured, hence shown as an NaN, this is valuable to the analysis as well and cannot be in anyway deleted. In order to make the maximum usage of the existing data, other attributes were not the key point for this project and has no use to the analysis and was therefore excluded. After manipulating with all files for process data sets, they were combined together as a whole data set treated as pre-processed process data, see Table 4.5.

Engine Number	Part A1 Number	Part A2 Number	Position	Measured Values	Date Time
17776	9571	1066	10	38	2020-09-04 04:58:35
17776	9571	1066	12	52	2020-09-04 04:58:35
17776	9571	1066	13	38	2020-09-04 04:58:35
10834	7281	1079	10	38	2020-09-04 07:34:19
10834	7281	1079	12	52	2020-09-04 07:34:19
10834	7281	1079	13	NaN	2020-09-04 07:34:19
13782	9611	1083	10	38	2020-09-30 05:35:45
13782	9611	1083	12	NaN	2020-09-30 05:35:45
13782	9611	1083	13	38	2020-09-30 05:35:45

Table 4.5: Sample pre-processed process data

Same for the manufacturing execution data, only the relevant ones: start time, end time, stop duration, and alarm ID were included. The alarm ID has already indicate what is the major part of the problem, hence, the specific reason for why the alarm is happening by using alarm text will not be involved in the analysis.

Since the time format from both data set are different, they were modified to the same. After the both data sets were selected and reconstructed, they were merged together based on when the part measurement is being recorded and when the manufacturing execution system started set off the alarm. This is for matching the the engine that is processing in the machine and the alarm, if it is happening at that time. The best possibilities for merging these is when the time were same, or the measurement is slightly behind the alarm. This is based on the reality that the measurements almost always happened after the alarm were provoked. The Python function *merge_asof* was used in this case for merging where the manufacturing execution data was on the left while the process data was on the right. In case the time difference varies too much, a tolerance of 1 hour were set in the function.

Now a merged data was ready, however the measured moment values were lost while merging. Therefore by implementing pivot table to the pre-processed process data (Table 4.5), vertical moment values were rearranged horizontally, as shown in Table 4.6. Combine it with the previous merged data, a complete data source with matched target values, Alarm ID and their features is ready to go into data exploration and the modelling part.

Engine Number	Part A1 Number	Part A2 Number	10	12	13
17776	9571	1066	38	52	38
10834	7281	1079	38	52	NaN
13782	9611	1066	38	NaN	38

Table 4.6: Sample process data after pivoting

4.4 Exploratory data analysis

Exploratory data analysis is one of the initial steps to reveal the potential patterns through investigations. This part of the result is mainly exploring different features from processed data sets. Consisting of studying the alarms individually using pre-processed manufacturing execution data, and the correlation of alarms with Part A1 and Part A2 utilizing final merged data. Both data sets are having the time range for seven months.

4.4.1 Alarm feature exploration

Figure 4.1 shows the summary statistics of the alarm IDs in machine A, its properties cover the number of alarms that have happened and the total duration the machines have stopped because of the alarms, ranking based on total duration. The figure has provided several messages. They are important for identifying which of the alarms are the most problematic or cause more problems to the production process.

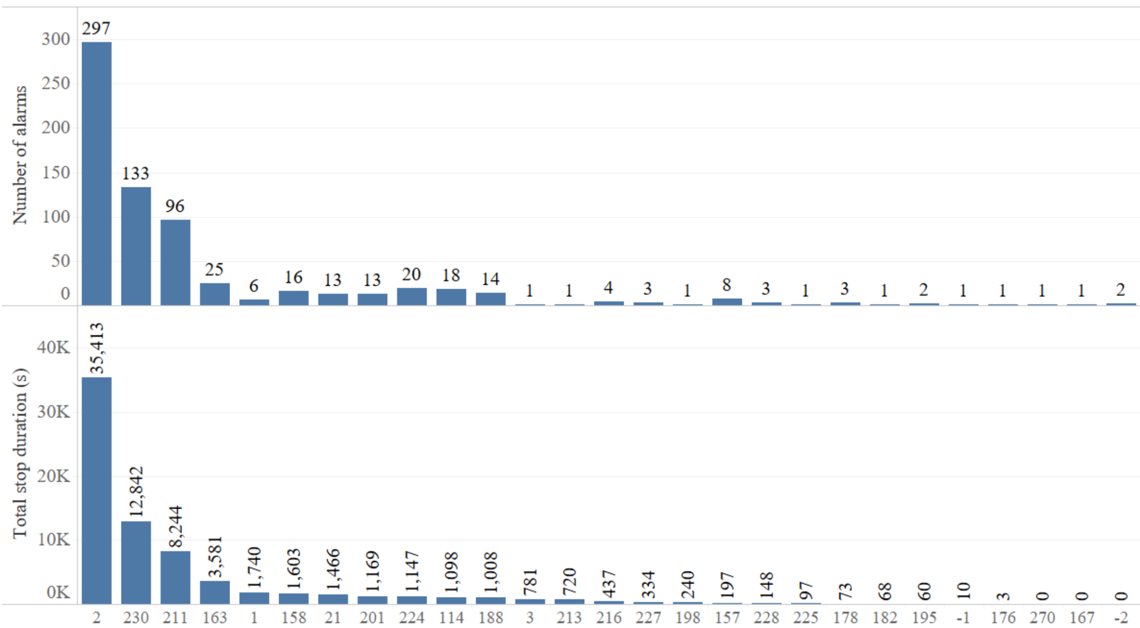


Figure 4.1: Summary statistics of the Alarm IDs in machine A

Alarm ID 270 and 167 and -2 in the far right are the only three good alarms that do not cause any production stop, all the others are bad alarms though having a different level of severity. Alarm ID 2 has the highest number of records as well as the highest sum of stop duration, hence it is one of the alarms that have to be focused on. Follow with alarm ID 230 and 211. Though following a decreasing trend, alarm ID 163 which is ranking the fourth has dramatically dropped in regard to both the counts and the total stop duration. It should be noted that nearly half of the alarms have only appeared once or twice through the entire seven months, and those will be hard to either finding the root cause or to be predicted in the future.

The same approach has been implemented for machine B. In Figure 4.2, it is clearly shown that the defective alarm in machine B is alarm ID 36. Same as alarm ID 2 in machine machine A, it has the longest total stop duration, as well as happened the most frequently among all the alarms in the machine. However, the second-highest in total stop duration, alarm ID 3, has only happened 22 times. This might due to the reason that some of the alarms will caused a long time production halt when it was triggered. For instance, one of the stoppages from alarm ID 3 has lasted for around one and a half hours. Out of all the alarms that happened in machine B, it can be observed that alarm ID 36, 37, and 243 are the most influential ones.

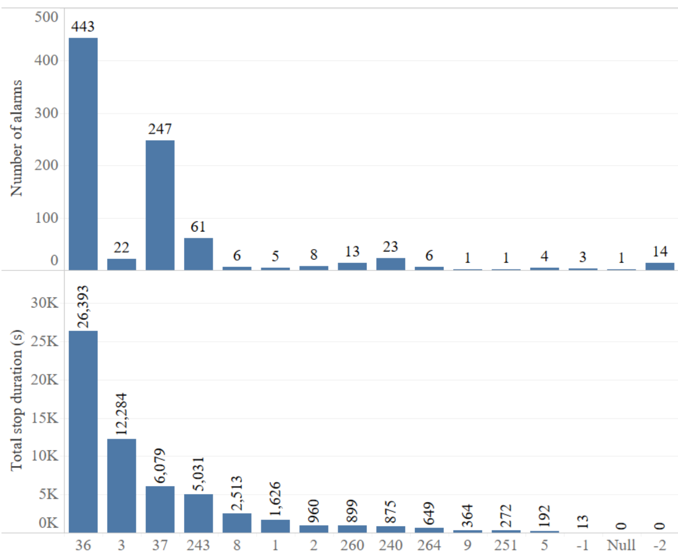


Figure 4.2: Summary statistics of the alarm IDs in machine B

In addition to the basic statistics of the alarm IDs mentioned above, additional analysis on the alarms had also being performed to find out the behaviour of alarms in terms of a monthly basis. The findings of this further analysis are given in the figures below. The line graph shown in Figure 4.3 depicts the distribution of alarms over each month showing the trend of which months are having more alarms compare to the others. It is evident from the graph that, comparing to machine B, the distribution is quite flat for machine machine A. Machine A has recorded the least alarms in month 6, whereas month 1, month 3, together with month 7, witnesses higher records in the alarms. There are possibilities backed by the information from the company where it states the production was slow down due to the spare parts shortage issue, and thus the above result seems reasonable. But not always, if taking the number of products that are being produced into account, the tendency will not always be the same. This will be seen in the following section.

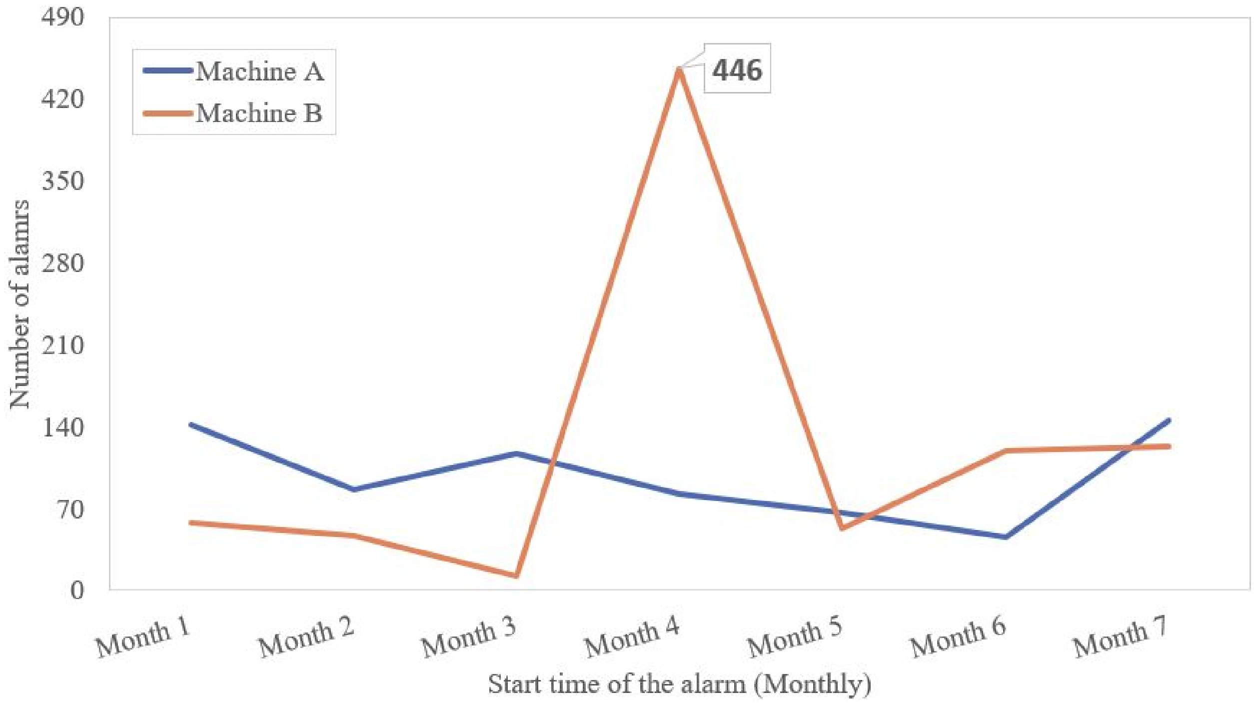


Figure 4.3: Correlation between number of alarms and months

On the contrary, the graph shows more variation in the number of alarms for machine B, and the trend has differed a lot from machine A. The major difference is the steep increase of the alarms in month 4. Since the rise was too severe, with a number of 446 alarms happening, which already triple the size of the second-largest month in month 7. Figure 4.4 was plotted to reveal which alarm ID is actually increasing during month 7. The plot has eliminated the alarm ID with an alarming number of counts smaller than ten considering they were not influential to the root cause. In the plot, the yellow bar shows the number of alarms with corresponding alarm ID while the green bar beside shows the average alarm numbers for each month. Comparing to the average number of alarms which is around 36, alarm ID 36 is having 273 alarms in month 4, being the main issue causing the increase. If checking on Alarm ID 36 particularly in Figure 4.5, it validates that compare to other months, it does happen in month 4 mostly. Alarm ID 37 also exceeds too much relative to the average. A deeper analysis has to be carried out to discover the root cause, however, this will not be realistic with available data.

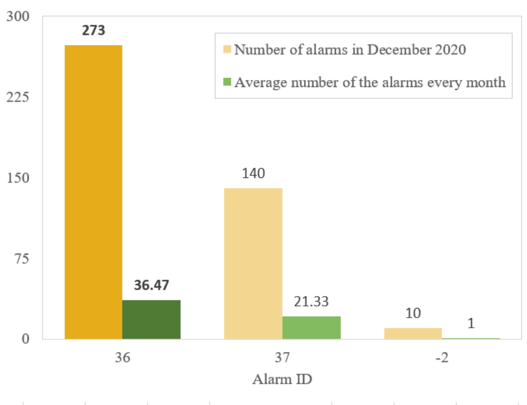


Figure 4.4: Alarm distribution happened in month 4 - Machine B

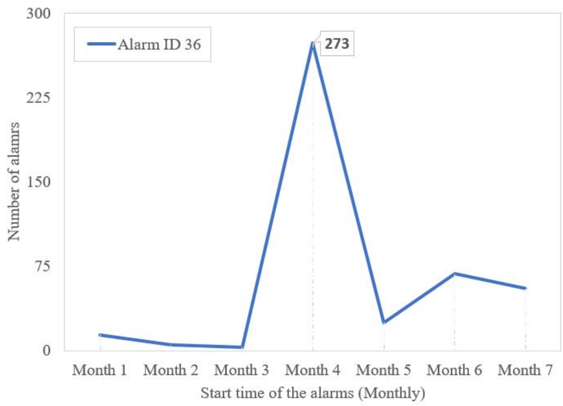


Figure 4.5: Number of alarms through different months for Alarm ID 36 alone

In the next analysis, the hourly distribution of the alarms was made and plotted in the line graph shown in Figure 4.6. The hourly analysis is performed to find out the pattern of alarm distribution during the day across month 1 till month 7, which will help to understand or provide insights about when exactly is the time of the day witnesses more alarms. From the result obtained, it is understood that all the numbers of alarms are ranging from approximately 10 to 60. The lowest was about at clock 20 at night and the highest were about at clock 15 or during midnight for both machines. The reason for this is possibly due to the change in shift of the production workers where the machine is left unattended for a certain period of time till the new shift production worker starts monitoring the machine.

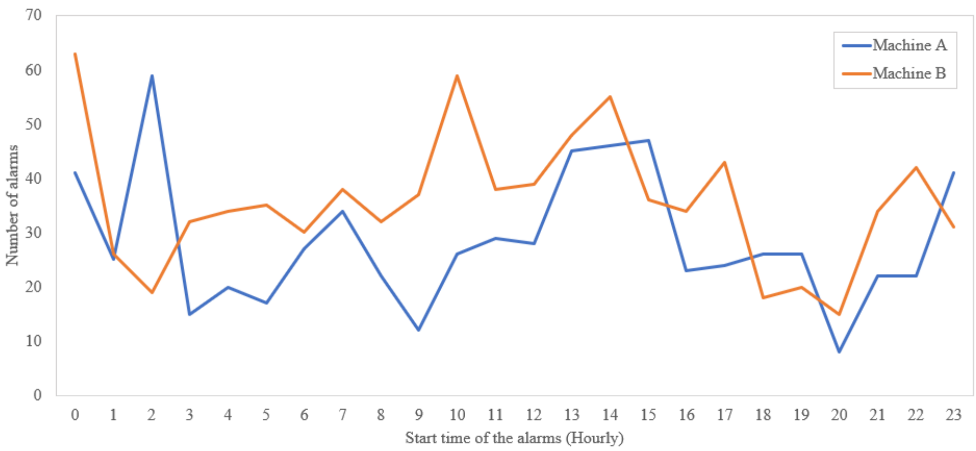


Figure 4.6: Correlation between number of alarms and hours

4.4.2 Correlation of alarms with Part A1 and Part A2

To find out which variants are causing more alarms, a heat map was used to visualize the magnitude of correlation at first. The darker color, the more alarms were generated with that variant. The variants can either be the specific type of Part A1 or the various combination of Part A1 and Part A2. For the reason that alarms generated in the different machine are having different alarm IDs, this process resulted in two maps for each of the machines separately.

Figure 4.7 displays the relationship between the Part A1 and alarm ID in machine A. By seeing simply blue area, the first few alarms IDs 2, 230, and 211 were proven again as the most problematic alarms causing the majority of production halt. On the other hand, Part A1 number 9571 and 9611 are the ones that trigger the alarms mostly. Nonetheless, these two Part A1 are having relatively mass production comparing the others, which might be the reason giving rise to the alarms. Thus, by taking the original number of the amount of production into account, Part A1 9651 is actually causing more alarms than all the others with a percentage of 4.86, the percentage is calculated as shown in Table 4.7, using the number of products came across alarms divided by all the products with specific Part A1. The percentage was added later as the pink area in Figure 4.7.

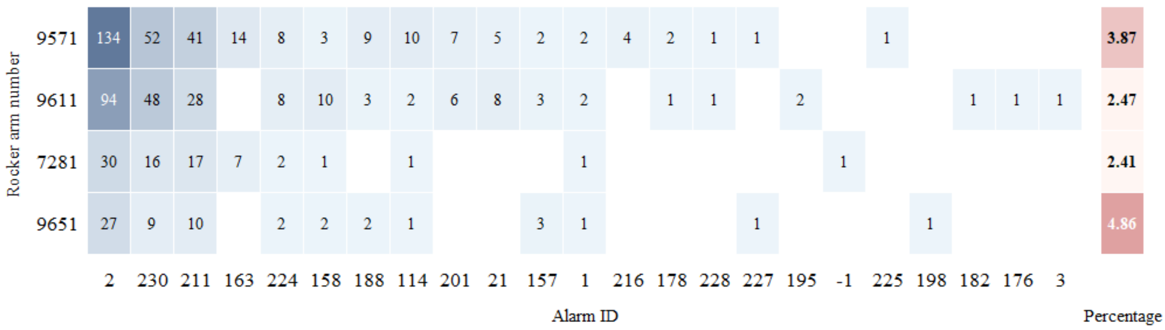


Figure 4.7: Relationship between Alarm ID and Part A1 - Machine A

Part A1 Number	Number of products	Number w/ alarms	Percentage (%)
9611	8864	219	2.47
9571	7652	296	3.87
7281	3160	76	2.41
9651	1214	59	4.86

Table 4.7: Percentage of the Part A1 causing the alarms - Machine A

When combining the Part A1 and Part A2 together, the heat map in Figure 4.8 has illustrated the relationship between Alarm IDs with different combinations in machine A. The combination of Part A1 9571 and Part A2 1066 is having the majority of alarms among all that generated, where Alarm ID 2 is the dominating factor. By calculating the percentage and taking percentage into account, this combination is in fact ranking fourth among the rest. As a matter of fact, Part A1 9651 is ranking the highest again with the same percentage of 4.85, due to it has only one matched Part A2. The second highest percentage is both for Part A1 9571, combined with Part A2 1078 and 1069 with a percentage of 4.17 and 3.97 correspondingly. Part A1 9611 has less alarm in general whereas, for the combinations with Part A1 7281, Part A2 1070 has apparently generated more alarms than Part A2 1079. It is worth mention that for each combination, the Part A2 number is unique, hence, there is no need to discuss the relationship between Alarm ID and Part A2 separately.

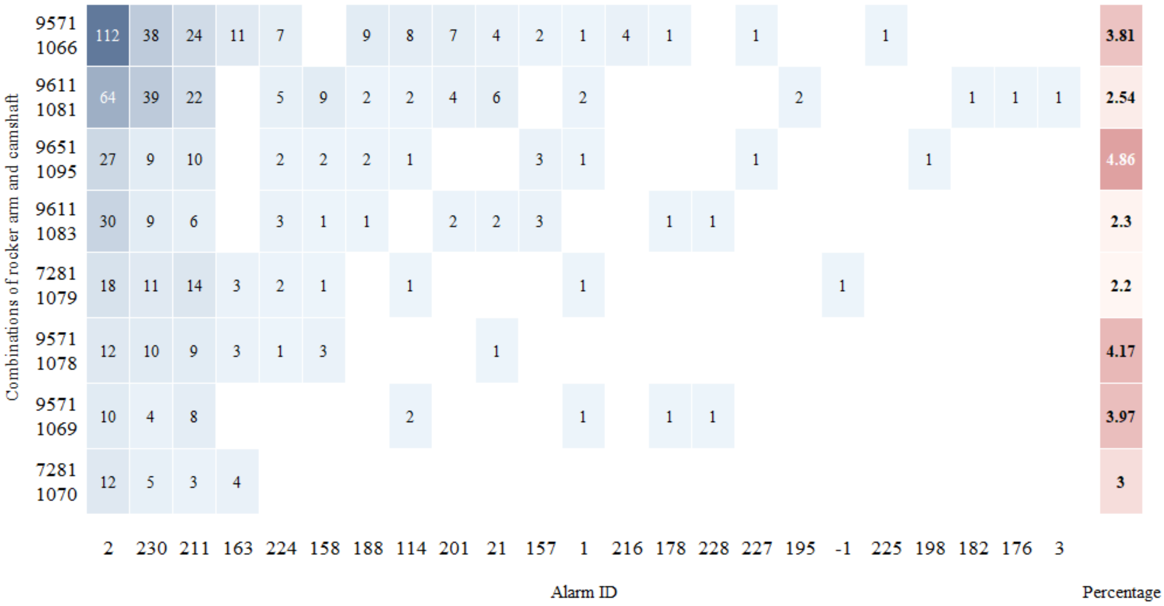


Figure 4.8: Relationship between Alarm ID and Combinations of Part A1 and Part A2 - Machine A

As can be seen from the figures below. Figure 4.9 and Figure 4.10 depicts the relationship between Alarm ID and the Part A1 variants in machine B. Alarm ID 36, 37, 243, and 240 are having most of the alarms and might be the reason for the cause in B. The most interesting aspect in this plot is that Alarm ID 243, 240, and 264 was triggered only when using Part A1 number 7281.

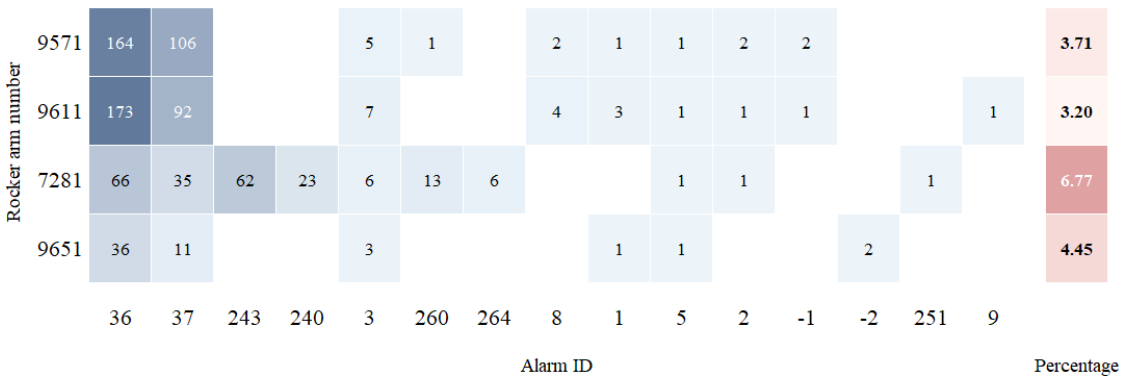


Figure 4.9: Relationship between Alarm ID and Part A1 - Machine B

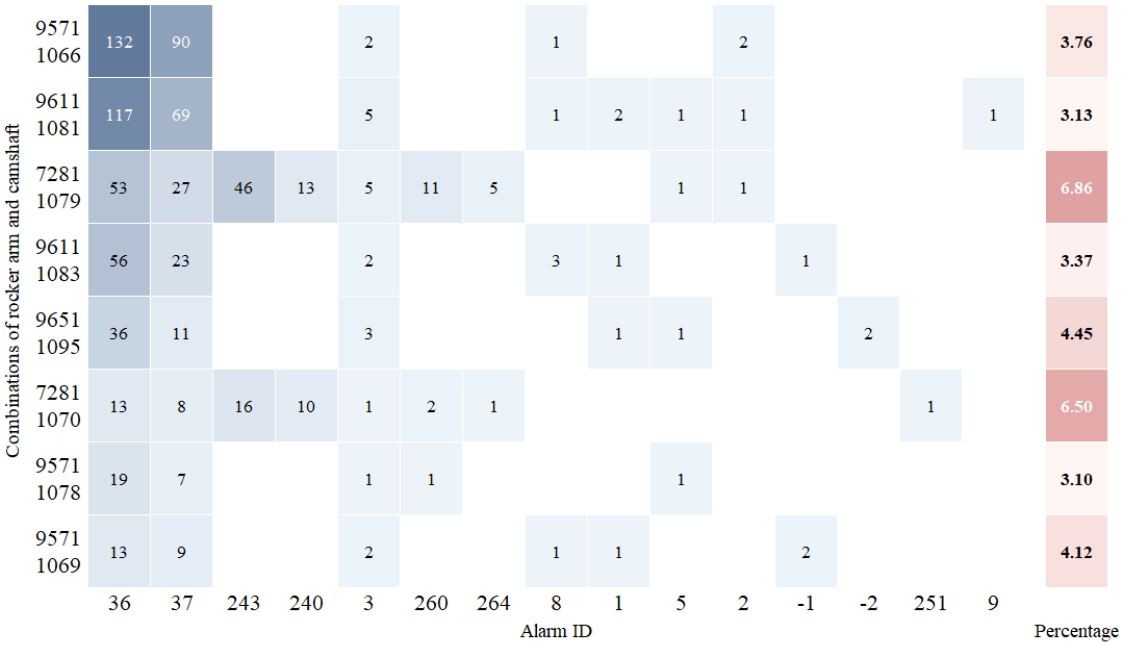


Figure 4.10: Relationship between Alarm ID and Combinations of Part A1 and Part A2 - Machine B

For better visualization of the assessment, a combo chart, Figure 4.11 is made to summarize and comparing the percentage of alarms each combinations has triggered in machine A and B. By observing only the blue bar and red bar, which stands for machine A and machine B respectively, they are having relatively similar alarm's rate. However, distinct difference between both combinations with Part A1 7281 can be observed. For instance, with the lowest value in machine A, combinations of Part A1 9571 and Part A2 1079 is actually having the highest value in machine B as 6.82. Which means out of 100 products having Part A1 7281 and Part A2 1079 going into the machine B, there will be approximately 7 products that had somehow generated the alarms. Combining with the findings previously, Part A1 7281 could be a condemned component for machine B.

Another plot could be interesting to see as if the months were taken apart. We can see from the figure that for each combinations, how the percentage of alarms are changing over each months. Figure 4.12 has the distribution for machine A. One unanticipated finding was that most of the combinations are having a significant high percentage in month 7. Same things happened in machine B. In Figure 4.13, month 7 is again having a high percentage, except that month 4 is having even higher percentage then in month 7. Almost all the combinations are having a 20 percent of products having alarms.

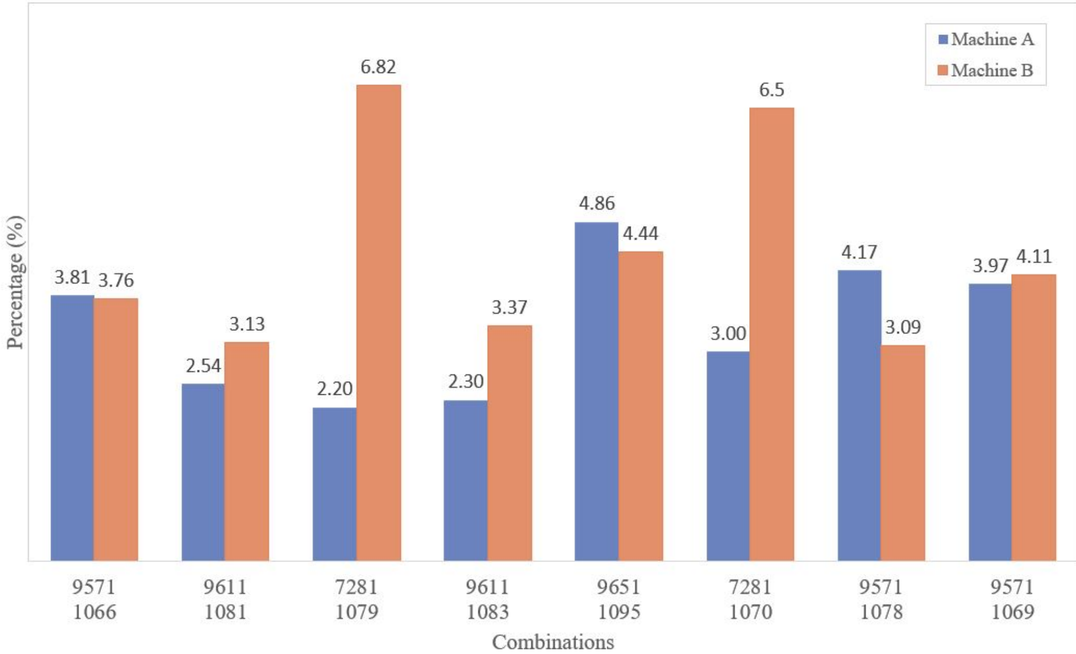


Figure 4.11: Summarize and comparing the percentage of the alarm in both machine

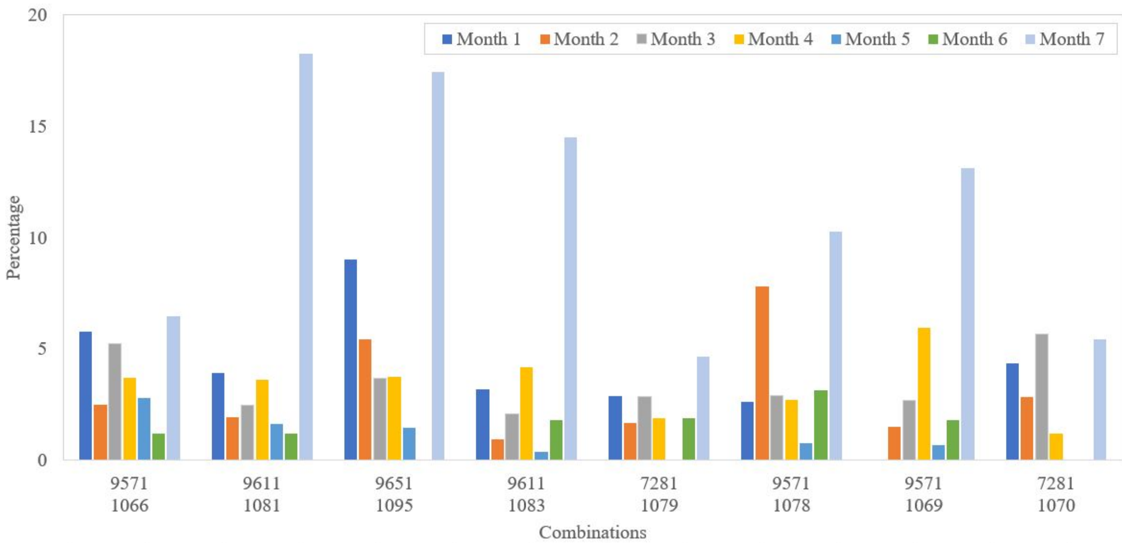


Figure 4.12: The percentage of products causing alarms for different combinations and months - Machine A

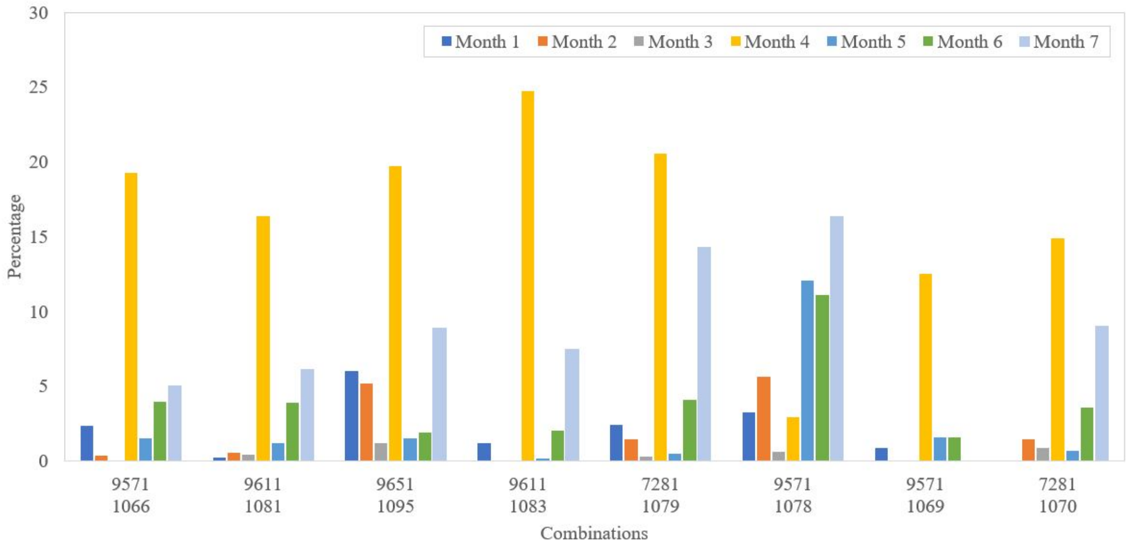


Figure 4.13: The percentage of products causing alarms for different combinations and months - Machine B

As a result of answering RQ1, *What product variants or combination of variants trigger more alarms?* . For machine A, Part A1 9651 and 9571 is clearly having more alarms than others, 9571 without a doubt need more attention since it produces the most of the products. Whereas for machine B, Part A1 7281, having the dominant percentage of the alarms, while the 9651 being the second. Part A2 number is not a critical as in this analysis.

To conclude, these explorations are for revealing the patterns behind, and these will be valuable insights for future root cause analysis. Many things will be worth studying further on, such as, what happened for machine B in month 4, what happened for both machines during month 7, or why Part A1 9651 is having more alarms in machine A, why Part A1 7281 is having more alarms in Machine B? How to reduce the production halt to the maximum?

4.5 Modelling

After performing the exploratory data analysis , the data set is then used for modelling, i.e., to generate machine learning models. The modelling is the 4th phase in a CRISP-DM, the machine learning algorithm comes into act in this phase in order to produce the desired output which satisfies the requirements stated in the business understanding phase. The modelling comprises 4 steps, which are selecting the modelling techniques, generating test design, building model and assessing the model. Each of these steps is further explained in detail in the sections below.

4.5.1 Selecting Modelling Techniques

The selection of modelling type is the first step in the modelling phase and a modelling type is selected to execute the modelling from a range of available modelling types. The important factor to consider in this step is to select the most suitable modelling technique which align with the organisations need and the other factors which influence the selection includes data types and the data mining goals (IBM, 2021b).

- **The data type of the available data set:** The data sets may contain attributes which are of different types, example numerical value or categorical value. Furthermore, the target attribute or variable data type have to be taken into consideration while devising or selecting the modelling techniques. In this project, the target attribute or variable is the alarm ID and it is a categorical variable.

- **Data mining goal:** The type of modelling technique should enable to achieve the desired data mining goals. The data mining goal for this thesis project is to predict the alarm id based on the Part A1- Part A2 combinations and its associated measured values of moment. Therefore, the modelling type is to be selected with consideration of its ability to perform the classification tasks on the available data set. Furthermore, the target class variable alarm id has more than 2 classes, therefore the classification modelling techniques selected should be able to conduct multi-class or multinomial classification.

After considering these 2 factors, the modelling techniques were selected which can be used to perform the multinomial or multi class classification tasks. Hence, the machine learning algorithms such as Decision Tree, Random Forest, K-Nearest Neighbors and XGBoost were selected for the modelling. The different selected models will be used to perform the modelling and the final selection is done based on the performance of each modelling technique.

4.5.2 Generate test design

The second step in the modelling phase is to generate a test design. It is important to generate a procedure which is used to check the model's quality and also to perform the validation of the model. Therefore, a test design is generated in which the data used in the modeling is split into training and testing data sets. The splitting of the data is done in order to evaluate the error rates, which determines the quality of the modelling. Hence for this, the model will be built on using the training data set whereas the model's quality will be tested using the test data set.

The data set used in the thesis is splitted into training and testing sets. The splitting of data is performed using the python library sklearn model selection by importing the train-test-split function. The train-test-split function performs the splitting of the data sets into training and testing sets. The train-test-split function has a range of parameters in which the first two parameters represent the data sets which are selected to use in the modeling. These two parameters consist of an input data set which is the data set with all the independent variables and an output data set which is the data set which contains only the dependent or the target variable (here the target class variable). Apart from this, the test-size parameter allows to define the fraction of data which is to be treated as the test data, as it is a fraction, the test-size value ranges between 0 and 1. For this data set, the test-size is chosen to be 0.30, which means 30 percent of the data set (both input and output) will be considered as the test data set and the model will be tested using these data sets. Whereas 70 percent of the data set will be treated as the training set and the model will be built using this training set.

4.5.3 Building the model

After performing the splitting of data into training set and test set, the training set is now ready to undergo modeling using previously selected lists of modelling techniques. In this step, the models will be created using each of the modelling techniques and the independent variable and dependent variables used in model is shown in the Table 4.8 below. The independent variables chosen are used to train the model in order to perform the prediction on the dependent variable, which is the alarm id. The target alarm id which the model predicts includes alarm id 2, 211 and 230, because from the exploratory data analysis performed it was understood that these 3 alarms are occurring more frequently.

Input Variables / Independent Variables	**Target Variable/ Dependent Variable**
Part A1 Number	Alarm ID
Part A2 Number	
Position	
Measured Values	

Table 4.8: Independent and Dependent Variables of the model

The model building includes several key steps which should be performed to enhance the model performance of which the important one is to investigate into the parameter settings (IBM, 2021a). The parameter settings or hyper parameter tuning is performed to find the optimum values of the parameters which will provide the best performance or accuracy to the generated model. Models produce change in performance when the parameters are varied and one of the important tasks here is to identify which parameter to choose and its corresponding values which will help to generate a better model. Also before generating the model, the data set is modified since the data set used in the modelling have missing values and different strategies are being used to deal with the missing values, which was due to the reason that, the measurement was not been performed at certain positions as mentioned in the table 4.6. The strategies which are used for modelling are shown in the table 4.9 below.

Strategy 1	Impute the missing values with zero
Strategy 2	Impute the missing values with median
Strategy 3	Impute the missing values with mode
Strategy 4	Impute the missing values with the nearest neighbor
Strategy 5	No Imputation - Keeping the missing values

Table 4.9: The Imputation Strategies for missing values in the data set

The modelling techniques mentioned earlier will be used to generate the models using the data obtained from different strategies as mentioned in the table 4.9. The modelling techniques Decision Tree, Random Forest and K-Nearest Neighbors will be used to generate models by using data obtained from strategy 1 to strategy 4, whereas XGBoost model will be generated with the data obtained from strategy 5. This is because only XGBoost can handle the missing values and all the other modelling techniques mentioned here don't work with missing values. Now the models are subjected to parameter tuning in order to identify the best values of the parameters. Since there are many strategies followed, the parameter tuning of models which uses data sets obtained from strategy 2 - Imputation by using median, will be discussed in detail below. Also at the end the total summarised parameter values of models generated is shown in the Table 4.10.

The K-Nearest Neighbors(KNN) algorithm was used first to create the model and its associated hyper parameters were varied and the changes in accuracy were plotted. For KNN, the hyper parameters such as number of nearest neighbors, leaf size were varied and tested for accuracy. As the name suggests, the number of nearest neighbors in KNN represents the number of neighbors which the KNN algorithm considers for its classification tasks, in this case the accuracy is following a decreasing trend as we increase the number of nearest neighbors for the data set under study. The figure 4.14 shows the line graph of accuracy versus number of nearest neighbors for KNN.

After observing the graph in figure 4.14 , it was clear that the accuracy is higher when the number of neighbors is one, but that is not a relevant approach to consider since the data point only compares with one nearest data point while performing the classification, which might also lead to a chance of over-fitting. Therefore, the number of neighbors 4 and 6 produces a better accuracy when compared to all the other values and thus it is considered as the optimum values for generating the KNN model.

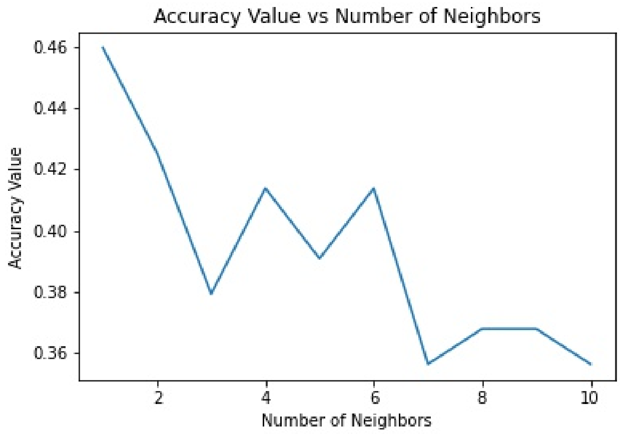


Figure 4.14: Relationship between Accuracy Value and Number of neighbors

Decision tree model generation has also been performed and the parameters have been varied to investigate the accuracy values corresponding to the varying parameter values. The parameters such as maximum depth, Minimum sample splits have been varied to investigate the best fit value for each of the parameters. The maximum depth parameter in a decision tree is the longest path between the root node and the leaf node. The maximum depth, when given a high value can lead to overfitting and neither it can be of a less value which will cause the model to under-fit. Therefore, a good or optimum value should be found out such that the model produces a decent or good accuracy. Hence, the values of max depth parameters had been varied and its associated accuracy were obtained and plotted using matplotlib library. The figure 4.15 depicts the graph of Accuracy versus Maximum depth parameter of a decision tree model.From the figure 4.15 it is observed that, maximum depth of 8 produces a higher accuracy when compared to other values and hence it can be considered as the optimum value to be used while building the model.

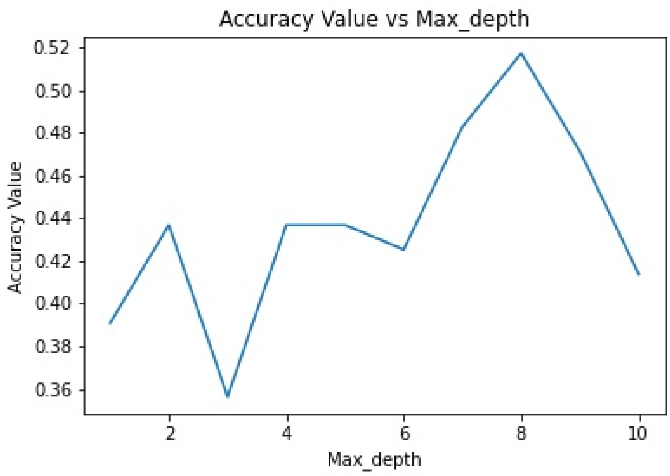


Figure 4.15: Relationship between Accuracy Value and Maximum depth - Decision tree

The other parameter which was varied is the minimum sample split, which is the minimum number of samples which are needed to perform a split in the internal node, for example if the minimum sample split is defined as 3 and a node have only 2 samples, then the node will fail to undergo splitting since it only have 2 samples, which is less than the min number of samples split defined. Higher values of minimum sample split might result in over-fitting and very high values will lead to under fitting as well. Therefore an optimum value is chosen which provides better model accuracy. The figure 4.16 depicts the relationship between accuracy and minimum sample split and it is clear from the graph that minimum sample split of value 6 is having better model accuracy. Hence the optimum value for minimum sample split is considered as 6 for the model building.

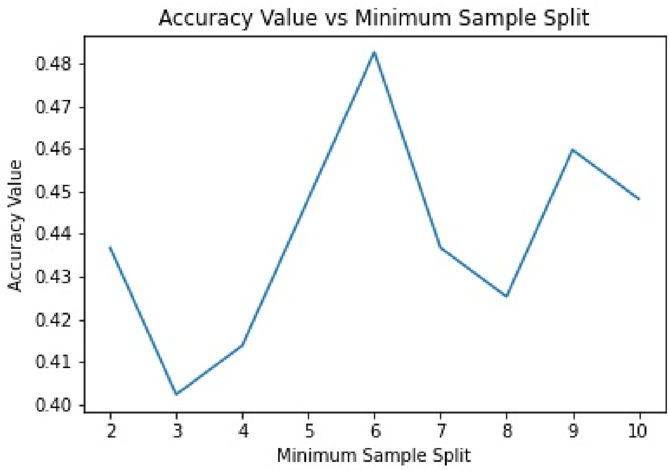


Figure 4.16: Relationship between Accuracy Value and Minimum samples split - Decision tree

The parameters were also varied for random forest and the list of parameters which were varied includes maximum depth, minimum sample splits and number of estimators. Since random forest and decision tree are almost similar, with random forest being the collection of many decision tree, most of the parameters are the same as that of the decision tree. The figure 4.17 depicts the relationship between accuracy and maximum depth of the random forest model. From the figure 4.17 it is noted that the maximum depth of value 4 performs better in accuracy when compared to other values.Therefore, maximum depth is assigned the value 4. The figure 4.18 illustrates the relationship between accuracy and minimum sample split values and it is observed that the minimum sample split of values 3 and 4 has a better accuracy score and hence the optimum minimum sample split values are 3 and 4.

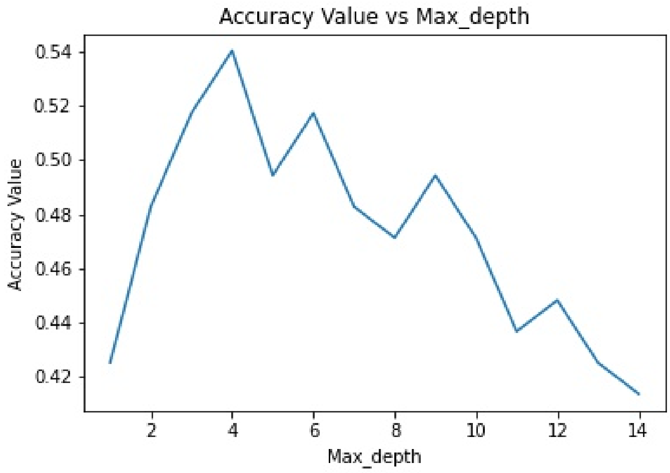


Figure 4.17: Relationship between Accuracy Value and Maximum depth - Random Forest

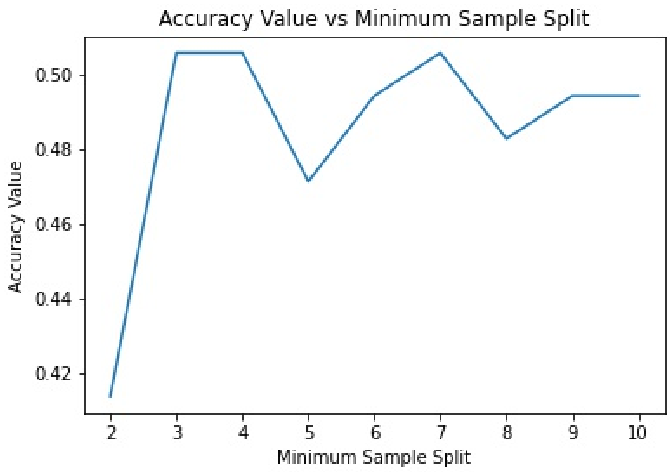


Figure 4.18: Relationship between Accuracy Value and Minimum samples split - Random Forest

As mentioned previously that the random forest is a collection decision tree acting together to perform the prediction, the number of estimators is one of the parameters which is to be varied and tested for accuracy. The number of estimators represents the total number of individual decision trees participating in the prediction process. The figure 4.19 shows the relationship between accuracy and number of estimators. From the graph obtained in figure 4.19 the number of estimators of value 18 produces a good accuracy and hence it is considered as the best value for the number of estimators parameter for random forest.

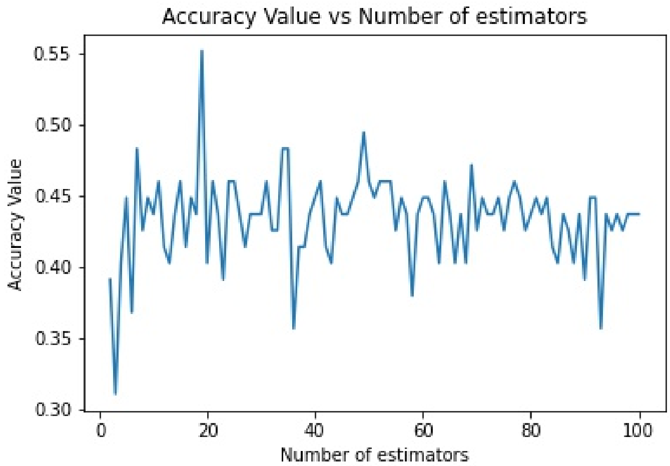


Figure 4.19: Relationship between Accuracy Value and Number of estimators - Random Forest

The data set used in the earlier modelling was imputed , wherein the NaN or missing value were replaced using the strategy 2, which is the imputation of missing values with the median. But, from the suggestion given by the domain expert from Volvo Group, which is to try performing the modelling on the data set which is not imputed or in other terms the unmanipulated data set. Therefore, the modelling had been again performed on the data set which contains missing values and hence XGBoost is used to generate the model using unmanipulated data according to strategy 5. According to the paper (Chen & Guestrin, 2016), XGBoost algorithm is highly efficient in dealing with missing values and the research has proven that XGboost shows good performance with the data sets containing missing values. Apart from this, application of XGBoost in a real case scenario as mentioned in the research article (Rusdah & Murfi, 2020) was also analysed to validate the effectiveness of XGBoost in handling missing values. The research article (Rusdah & Murfi, 2020) describes the risk level classification problem in insurance industries, where classification is a multi-class classification problem due to the many number of risk categories. Since it is related to insurance data sets, there were many missing values in the data set which is to be used for the modelling and uses XGboost algorithm to evaluate the performance by comparing the model performance done on both imputed data sets and the data set which is not imputed. The final result states that the accuracy level of the XGBoost model performed using the data sets without imputation produces a similar score when compared to the XGBoost accuracy score with the imputed data set. Hence, the XGBoost is chosen for our modelling, by using the data set with missing values or the data set which is not imputed.

The XGBoost algorithm is then used to build the model and the parameter tuning is performed to identify the best values for the parameters, which provides more accuracy. Firstly, the Maximum depth parameter is varied and tested for accuracy and the resulting graph which shows the relationship between accuracy value and max depth values is shown in figure 4.20 . Hence, the maximum depth of value 3 produces a good accuracy when compared to other values which are been tested.

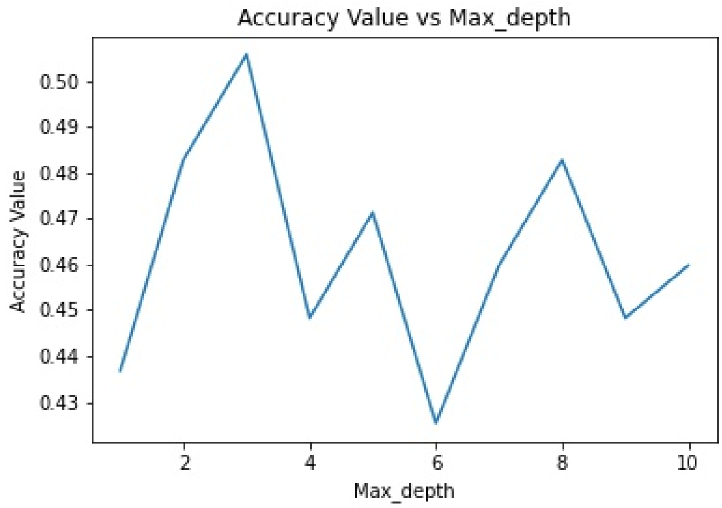


Figure 4.20: Relationship between Accuracy Value and Maximum depth - XG-Boost

Secondly, the learning rate has also been varied and tested for accuracy. The learning rate is the parameter which controls the weight adjustment of the new tree generated with respect to the gradient of loss (Zulkifli, 2018). The smaller the learning rate, the longer the time it will take to find the local minima where the loss is the minimum. The learning rate has been varied from 0.05 to 1 and the corresponding accuracy value obtained had been plotted in the line graph shown in figure 4.21. From the graph, it is clear that the learning rate of value 0.2 has a better accuracy than the rest of the compared values and hence the optimum learning rate if chosen to be as 0.2.

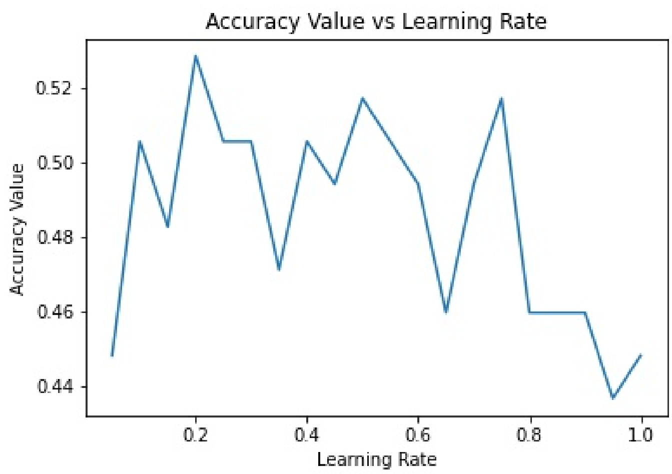


Figure 4.21: Relationship between Accuracy Value and Learning rate - XGBoost

Thirdly, the number of estimators parameters had also been varied and tested for accuracy. The figure 4.22 shows the relationship between the accuracy value and the number of estimators for the XGBoost model. The number of estimators of value

100 have better accuracy and hence it is chosen as the best or optimum value for the number of estimators.

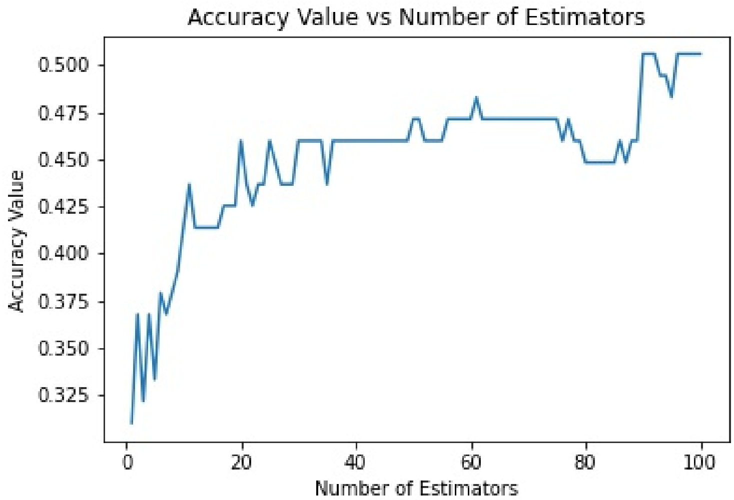


Figure 4.22: Relationship between Accuracy Value and Number of Estimators - XGBoost

The models are built using the data from the previously mentioned strategies and the optimum values for the parameters for each models corresponding to each strategy were obtained by performing the parameter tuning on all the generated models. Therefore, the summarized results of the optimum parameter values are shown in the table 4.10 below. Now, as the parameter tuning is successfully performed and the best parameter values of each model are obtained, it is then used to generated model using these parameter values and further assessed for performance in the next section.

Model	Parameter	Strategy 1	Strategy2	Strategy 3	Strategy 4	Strategy 5
K-Nearest Neighbors (KNN)	n_neighbors	5	4	4	2	X
Decision Tree	Max_depth	7,8	8	8	6	X
	Min_Sample_Split	3,9	6	6	8	X
Random Forest	Max_depth	6	4	5	4	X
	Min_Sample_Split	9	3,4	7	5	X
	n_estimators	28	18	100	18	X
XGBoost	Max_depth	X	X	X	X	3
	Learning_Rate	X	X	X	X	0.2
	n_estimators	X	X	X	X	100

Table 4.10: The best parameter values of each model corresponding to different strategy

4.5.4 Assessing the Model

Assessing the model is the last step in the modelling phase and in this step, each model which has been generated in the previous step will be compared to each other to facilitate the selection of final model. Since the predictive modeling was performed to predict the alarm occurrence and to classify these alarms into their respective categories, accuracy is chosen for comparing the model. The machine stops or comes to halt for every occurrence of the alarm irrespective of the alarm category or alarm id. Hence the percentage of correct predictions of occurrence of alarms is very important for this thesis project. Therefore the accuracy of these different generated models corresponding to each strategies were analysed and compared to find the best model, i.e., the model with higher accuracy value. The table 4.11 below shows the accuracy value obtained for each models corresponding to each strategy.

Model	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
K-Nearest Neighbors (KNN)	41.37%	41.38%	43.67%	45.98%	X
Decision Tree	52.87%	54.02%	54.02%	47.12%	X
Random Forest	51.72%	49.42%	50.05%	43.67%	X
XGBoost	X	X	X	X	52.87%

Table 4.11: Accuracy values of each model with respect to different strategy

From the table 4.11 , it is clear that the Decision tree model is dominating and has comparatively higher accuracy when compared to other models. But, decision tree model accuracy obtained using strategy 2 and strategy 3 are having same value for accuracy, which 54.02 percentage. Hence, further assessment is performed on these two models to select the final model. In this further assessment, the 2 decision tree models mentioned above will be compared among each other based on the false negative rates, which is calculated from the confusion matrix generated for each of the 2 decision tree models. False negatives is defined as the outcomes when the negative class is wrongly predicted, for example in this project, when an outcome is predicted as alarm id 211 but it was actually alarm id 2, is an example of false negative, which can be also considered as a false alarm prediction. False negative rate is defined as the ratio of false negatives to the sum of false negatives and true positives. Hence comparing these 2 models on false negative rates will give a better picture on deciding the model. Confusion matrix is generated to understand the performance of the classification model. Confusion matrix helps to summarise the correct predictions of labels by generating a matrix True label as the rows and the predicted label as the columns. The associated number in each cell provides the count of each label's prediction. Confusion matrix generated for the decision tree models with strategy 2 and strategy 3 is shown in figure 4.23 and figure 4.24 respectively.

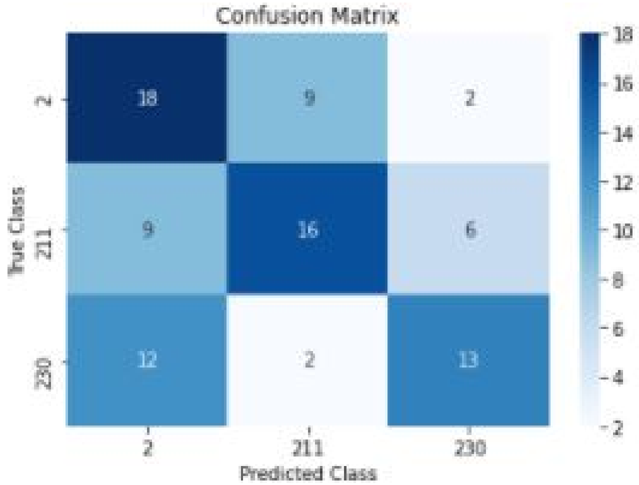


Figure 4.23: Confusion Matrix of Decision Tree model - Strategy 2

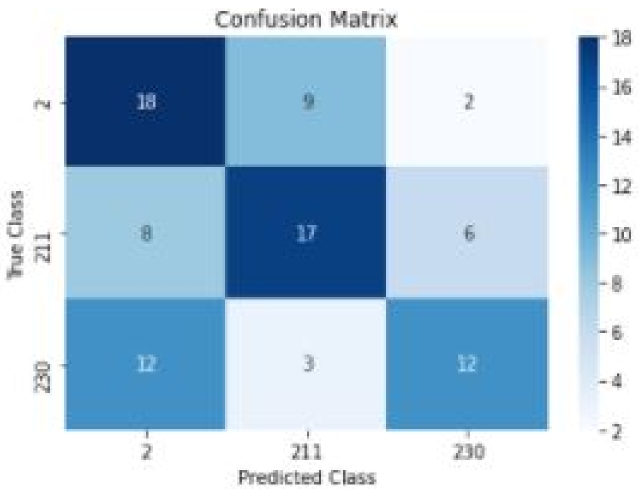


Figure 4.24: Confusion Matrix of Decision Tree model - Strategy 3

Now the confusion matrices obtained above is summarized for each prediction class, which is the 3 alarm id's namely 2, 211 and 230. The summary of the confusion matrix provides details about the counts of false negatives and true positives, which is required to perform the calculation for false negative rate.Therefore, the summary of the confusion matrices obtained is shown in the table 4.12.

Decision Tree Model - Strategy 2				
Class	**True Positive**	**False Positive**	**True Negative**	**False Negative**
Alarm id - 2	18	21	37	11
Alarm id - 211	16	11	45	14
Alarm id - 230	13	8	52	14
Decision Tree Model - Strategy 3				
Class	**True Positive**	**False Positive**	**True Negative**	**False Negative**
Alarm id - 2	18	20	38	11
Alarm id - 211	17	10	44	14
Alarm id - 230	12	8	54	15

Table 4.12: The summary of the confusion matrix obtained for decision tree model

From the summary obtained from the table 4.12, the false negative rate is calculated for each of the 3 prediction class considering both the strategies. The figure 4.25 depicts the comparison of false negative rates for each strategy.

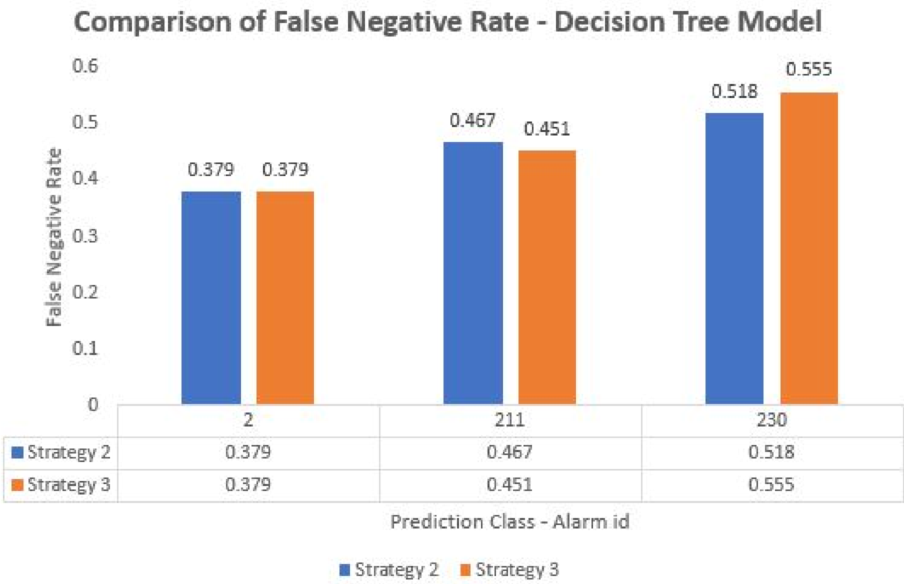


Figure 4.25: Comparison of False Negative Rate - Decision Tree Model

From the figure 4.25, it is observed that both the strategies have similar false negative rate for the prediction class - alarm id 2, where as strategy 2 have slightly higher false negative rate for alarm id 211 and strategy 3 have a higher false negative rate value for the alarm id 230. Considering this, strategy 2 have slightly lower overall false negative rate (considering false negative rate of all the alarm id) and hence the decision tree model with strategy 2 is chosen to be the final model. The decision tree also has the capability to provide the decision rules which serves as a medium for decision making, which will be discussed in the next section. The RQ 2 is answered here,i.e., by selecting the machine learning model to explore the possible in-depth root causes in production and maintenance.

Since the decision tree model with strategy 2 is been chosen based on the false negative rates as mentioned above, more metrics regarding the final model had been calculated and is presented in the table 4.13 below. This is done to get additional understanding of the model regarding the recall, precision and f1-score. For example, considering the alarm id 2, precision is explained as when it predicts alarm id 2, how often is this actually alarm id 2.Recall can be explained as when it is actually alarm id 2, how often is it predicted as alarm id 2. Precision is calculated as the ratio of true positive to the sum of true positive and false positive, whereas recall is calculated as the ratio of true positive to the sum of true positive and false negative. Alarm id has the lowest precision when compared to others and alarm id 230 has the lowest recall of the all. F1-score is not emphasised because, f1-score is used when dealing with unbalanced data sets, whereas the data set used in this model is a balanced one.

Class / Alarm ID	Precision	Recall	f1-score
2	0.46	0.62	0.53
211	0.59	0.52	0.55
230	0.62	0.48	0.54

Table 4.13: Precision, Recall and f1-score of the final model

4.5.5 Decision rules

The extract rules from the decision trees are one of the important outcomes from the modelling phase. It can helps understanding in a visual format how the trees propagate, and facilitate the domain expert to identify what is most crucial features for identifying the alarm IDs. What is interesting about the rules is that instead of Part A1 or Part A2, criteria are mostly based on the moment value in certain position. This might due to the moment values are really crucial in the sense of identifying alarm IDs. There are other possibilities that the preprocessed data after one-hot encoding has became too sparse.

The plot of the decision tree generated was shown in Appendix B. However, not all the rules are reliable, and the probability from model for each rules were not enough to determine either. Therefore, each rules were tested with the test data separately, checking the accuracy of the prediction for each of the rules. If the accuracy were higher than 50 percent, then they were kept as a reliably rule. Below are the sample rules, MP here are representing position number. If Part A1 number or Part A2 number are equal to one, means that they have this component as a criteria. The full version can be found in Appendix C. We can tell from the rules that, the alarms are most likely to happen when Part A1 is 9571, or Part A2 is 1070, 1081 and 1078.

if (MP 32 <= 52) and (MP 62 <= 52.5) and (MP 43 <= 38.5) and (MP 60 <= 38.5) and (RANR_9571 = 1) and (MP 52 <= 52.5) **then** Alarm ID: 2

if (MP 32 <= 52.5) and (MP 62 <= 52.5) and (MP 43 > 38.5) and (MP 52 <= 52.5) **then** Alarm ID: 2

4.6 Evaluation

The final model selected in the previous phase is then evaluated further to check whether the model aligns with the business requirements or objectives of this project. The business objective as mentioned before is to find out the Part A1 - Part A2 combination which triggers more alarms and to devise a model to predict the alarm id . The Part A1 - Part A2 combination which triggers more alarms had been found out from the diagnostic analysis, in which the exploratory data analysis is used to obtain the findings. Apart from that , a predictive model has also been generated by Decision tree using strategy 2 wherein it performs the classification process and predicts the alarm id type based on the input variables chosen as mentioned before in the report. The Decision tree model obtained using strategy 2 has an accuracy of 54 percentage, which is considered as a good value in industrial scenario. The domain expert from Volvo Group supported this by stating that any value of accuracy above 50 percentage is considered as good start for industrial problems.

5 Discussion

This thesis project studied and performed the research on data driven approach on industrial alarm data and demonstrates the diagnostics and predictive approach by using exploratory data analysis and machine learning algorithms respectively. The thesis project was successful in providing new findings from the industrial data set given for the study and also explored a novel approach of implementing predictive modelling in industrial alarm problems. The first and second part of this section discusses about the academic and industrial or practical contribution from this thesis project is explained in detail. This is further followed by the discussion of various challenges faced during the project and lastly the discussion regarding recommendation for improvements.

5.1 Methodology

The method used in the thesis is CRISP-DM, there are other data mining processes such as Knowledge Discovery Databases (KDD) and Sample, Explore, Modify, Model, Assess (SEMMA). They were all similar to each others with different level of completeness (Shafique & Qaiser, 2014). Choosing CRISP-DM as the methodology was determined in the very beginning. Considering the fact that the limitation of accessing quality data, the scope as been revised as the project progressed. It could be better with implementing KDD model instead since KDD is more torwards extracting the hidden knowledge with its own data exploration step (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), while in CRISP-DM, explore data is only a small part under data preparation. However, the objectives has been achieved based on modified version of CRISP-DM.

5.2 Academic contribution

From the literature survey performed in the earlier chapter, it was evident that the industrial alarm data problems mostly followed a descriptive and diagnostics approach and haven't been explored much in terms of implementing predictive modelling. This thesis project takes one step ahead by implementing a predictive approach on industrial alarm data problems by using machine learning methods. The descriptive and diagnostics analysis is performed in the data understanding and exploratory data analysis phase, whereas the predictive modelling is developed by generating machine learning models with the data under study to produce the predictions based on the given input features as mentioned in the modelling phase of the

CRISP-DM process. The motivation of implementing novel approach in industrial alarm cases, which is to develop a predictive model is obtained from the literature paper (Bousdekis et al., 2019), wherein the author identified the research gap and proposes the research recommendations of implementing generic decision models by employing machine learning algorithms. The predictive modelling solution also aligns with the research recommendation as mentioned in (Bousdekis et al., 2019), to enable less human intervention in decision making by using the data driven approaches. The predictive modelling generated in this project will serve as a medium to reduce the human dependency in detecting the alarms in the industrial platforms. Furthermore, as proposed by the authors (Vogl, Weiss, & Helu, 2019), the production is undergoing a customisation approach in order to withstand the increasing competition and is aiming to eliminate the unplanned production halt or stoppage as a part of this. This paper also highlighted the importance of self maintenance, where the maintenance is done by the system itself by performing prognostics or predictive modelling of the failure or possible chances of process deviations which are indicated by the alarms. Therefore, the predictive modelling generated in this thesis helps the organisation to implement prognostics in their manufacturing in order to avoid the unplanned production stoppages. Finally, the implementation of machine learning models are also been performed in this project by obtaining motivation from the paper proposed by authors (Ayvaz & Alpay, 2021). The machine learning model is capable of making prediction from the real time data generated from the machines and thus helping to predict or classify the alarms instantly.

5.3 Practical contribution

Prior studies have noted the importance of diagnostic insights in facilitating the planning of maintenance activities (Subramaniyan et al., 2020). The outcome from the first phase in this project were provided in a way to show the diagnostic result of the production halt, explored the relationship between the alarm ID and product types. This will help the industrial company side to better visualize the correlation and effectively communicate the findings from this exploration. For instance which are the most problematic alarm IDs, which combinations of Part A1 and Part A2 will be more easily to trigger the alarms. Most importantly, it is a decision support for further discovering the root cause, finding the specific reason behind the superficial symptom.

The next level, in the context of predicting the alarm ID, only the traditional methods were implemented in this project. Considering the black box format that is generated by neural networks, and time limitation for this project. Decision tree is the one of the best strategies with a relatively high accuracy and practitioner friendly. The ultimate decision rule were delivered so that no previous knowledge needed from data science. The decision rule which was generated from the decision tree model, provides a better insights into the model's predictions. The decision rules generated in the thesis is of simple and understandable format, such that a normal machine operator will be able to read through the rules and understand the statements mentioned in each rule with ease. The decision rules generated thus acts as the results of this modelling, i.e., it helps to identify the problematic Part A1, Part A2 and also the corresponding position of measurement which triggers the specific alarm id and thereby helps to initiate further investigation into the background data regarding these Part A1 and Part A2. The background data can possibly include the data from quality which when used or explored, can result in finding the root causes. For example, the root causes for the triggering of alarms may be connected with the human errors and thus finding the root cause can help the company to devise a predictive maintenance strategy by connecting these results and analysing this with the maintenance data.

5.4 Challenges

While doing the exploratory data analysis, unlike sensor data which has many features like tendency, skewness. Data from the manufacturing execution system has fewer features to explore making it harder to find the patterns and further increase the difficulty to do machine learning activities.

The major problem we faced was in pre-processing the data. The occurrence of each alarm IDs are different which leads for causing the imbalance of the data. When doing machine learning, it will be difficult to consider all the alarm IDs into account. Thus, only the three alarm IDs happened most frequently was taken into consideration, by doing down-sampling to balance the classes. In this case, leading too few samples (288 in total) for training as well as testing. Another challenge is not all the positions were being measured for different products, leaving some of the moment values being NaN, which cannot be neglected. However by using CART, decision tree classifier from Sklearn does not handle missing values. Algorithms like XGboost will be able to manage, however no clear rules which indicating which lable the rules belong to, can be extracted directly from it. Hence, different imputation methods for decision tree were tested and compared.

Since the thesis project was performed during the time of covid-19, we weren't able to visit the manufacturing plant and thereby didn't get a chance to see the machines and process in real, which would have helped us to explore more from our perspective. Despite having some delay in the getting data from the company, online meetings with the domain expert at the company had provided us all the required information.

5.5 Recommendations for improvement

These findings maybe somewhat limited by the machines that were studied in this project and the features that were chosen to analyse. They were in this case selected based on the expert experience. In future investigations, it might be interesting to have more data for both root cause analysis and predictive modelling.

5.5.1 Using more historical data

According to the article (Intermedia, 2019) the machine learning model's performance follows an increasing trend as the quantity of data increases. Hence, more data implies a better chance of generating a better machine learning model. In the case of this thesis, the amount of data obtained after performing data cleaning, data formatting and data merging was very less when compared to the quantity of initial data. Therefore, the model developed in this project can be used to analyse the historical data according to the company's demand.

5.5.2 Exploring data regarding quality

The data regarding the quality or quality department's data was not available for this thesis project due to confidentiality reasons and hence one suggestion from our side is to include the data regarding quality and further analyse using the same procedure developed in this thesis project. The company can make use of the quality department data to find out more about the root causes, for example, the data regarding the suppliers of materials can be used to track down the suppliers which are providing low quality materials. The suppliers details and material can be added to the input variables and further used in the model. Since the Part A1 - Part A2 combination which triggers more alarms had already been found, now it can be used to track down the suppliers and the materials which are used for manufacturing the Part A1 and Part A2.

6 Conclusion

This thesis has conveyed the result as decision support to predictive maintenance so as to planning the maintenance activity ahead of time. One of the major character studied in the thesis is the alarms. The industrial alarms serve as medium to address the process deviation or equipment malfunctioning and therefore it is important to devise a proper alarm management in order to achieve a good overall production performance. As the modern industries are aiming to reduce human intervention by developing technologies which are capable of decision making, this thesis project also proposed a novel approach of implementing predictive modelling in industrial alarm problems.

The thesis uses a data scientific approach to perform the diagnostic and predictive modelling by following the CRISP-DM methodology. The diagnostics approach was done by performing exploratory data analysis and the results of the exploratory data analysis provided the insights into the Part A1 - Part A2 combination which triggers more alarms. The various data visualization performed as a part of exploratory analysis further gave insights into behaviour of the alarms occurring in the machine under study. Since the problematic Part A1 - Part A2 combination were found from the diagnostics, it was then used for the predictive modelling, wherein the decision tree machine learning algorithm was used to classify the alarm id based on the inputs or independent variables used for modelling. The decision rules were produced from the machine learning models as well. Since they were not all useful, the useful ones with a higher accuracy were extracted for the domain experts to understand easily where could go wrong under the production in advance. This analysis undertaken here, has extended our knowledge of the possibility using manufacturing execution system data to do modelling, and the reliability of using decision rules as decision support.

Utmost importance, the data scientific approach performed in this thesis project is flexible such that the company can follow the same visualization techniques and modelling methods to analyse vast amount of historical data as per their demand.

References

An, D., Kim, N. H., & Choi, J.-H. (2015). Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. *Reliability Engineering & System Safety*, *133*, 223–236.

Ayvaz, S., & Alpay, K. (2021). Predictive maintenance system for production lines in manufacturing: A machine learning approach using iot data in real-time. *Expert Systems with Applications*, *173*, 114598.

Baptista, M., Sankararaman, S., de Medeiros, I. P., Nascimento Jr, C., Prendinger, H., & Henriques, E. M. (2018). Forecasting fault events for predictive maintenance using data-driven techniques and arma modeling. *Computers & Industrial Engineering*, *115*, 41–53.

Bezerra, A., Silva, I., Guedes, L. A., Silva, D., Leitão, G., & Saito, K. (2019). Extracting value from industrial alarms and events: A data-driven approach based on exploratory data analysis. *Sensors*, *19*(12), 2772.

Bo hu LI, W. t. Y. X. b. L. C. w. Y., Bao cun HOU. (2017). Applications of artificial intelligence in intelligent manufacturing: a review. *Frontiers of Information Technology and Electronic Engineering*.

Bousdekis, A., Lepenioti, K., Apostolou, D., & Mentzas, G. (2019). Decision making in predictive maintenance: Literature review and research agenda for industry 4.0. *IFAC-PapersOnLine*, *52*(13), 607–612.

Bushra, A. (2020, July). *Defining data science, machine learning and artificial intelligence @ONLINE.* Retrieved from `https://towardsdatascience.com/defining-data-science-machine-learning-and-artificial-intelligence-95f42a60b57`

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. d. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, *137*, 106024.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., ... others (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, *9*, 13.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

cloud education, I. (2020, July). *What is machine learning @ONLINE.* Retrieved from `https://www.ibm.com/cloud/learn/machine-learning`

David R. Sjödin, M. L., Vinit Parida, & Petrovic, A. (2018). Smart factory implementation and process innovation. *Research-technology Management*.

Diggle, P. J., Chetwynd, A. G., & Chetwynd, A. (2011). *Statistics and scientific method: an introduction for students and researchers.* Oxford University Press.

Education, I. C. (2020, June). *Artificial intelligence @ONLINE.* Retrieved from `https://www.ibm.com/cloud/learn/what-is-artificial-intelligence`

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37–37.

Fink, O. (2020). Data-driven intelligent predictive maintenance of industrial assets. In *Women in industrial and systems engineering* (pp. 589–605). Springer.

Gröger, C. (2018). Building an industry 4.0 analytics platform. *Datenbank-Spektrum*, *18*(1), 5–14.

Gupta, P. (2017, May). *Decision tree in machine learning @ONLINE.* Retrieved from `https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052`

Gustavo Dalmarco, A. C. B. A. L. S., Filipa R. Ramalho. (2019). Providing i4.0 technologies: The case of a production technology cluster. *Journal of High Technology Management Research.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

He, Y., Gu, C., Chen, Z., & Han, X. (2017). Integrated predictive maintenance strategy for manufacturing systems by combining quality control and mission reliability analysis. *International Journal of Production Research*, *55*(19), 5841–5862.

Hu, L.-Y., Huang, M.-W., Ke, S.-W., & Tsai, C.-F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, *5*(1), 1–9.

IBM. (2021a, April). *Building the models @ONLINE.* Retrieved from `https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=modeling-building-models`

IBM. (2021b, April). *Generating a test design @ONLINE.* Retrieved from `https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=modeling-selecting-techniques`

Intermedia. (2019, October). *The importance of data in machine learning model @ONLINE.* Retrieved from `https://www.intermediait.com/blog/2019/10/20/the-importance-of-data-in-machine-learning-models-is-there-a-magic-number-that-determines-the-amount-of-training-data-required/`

Jimenez, J. J. M., Schwartz, S., Vingerhoeds, R., Grabot, B., & Salaün, M. (2020). Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics. *Journal of Manufacturing Systems*, *56*, 539–557.

Kalathas, I., & Papoutsidakis, M. (2021). Predictive maintenance using machine learning and data mining: A pioneer method implemented to greek railways. *Designs*, *5*(1), 5.

Lucke, M., Chioua, M., Grimholt, C., Hollender, M., & Thornhill, N. F. (2020).

Integration of alarm design in fault detection and diagnosis through alarm-range normalization. *Control Engineering Practice*, *98*, 104388.

Mahadevan, S., & Shah, S. L. (2009). Fault detection and diagnosis in process data using one-class support vector machines. *Journal of process control*, *19*(10), 1627–1639.

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

Mehmeti, X., Mehmeti, B., & Sejdiu, R. (2018). The equipment maintenance management in manufacturing enterprises. *IFAC-PapersOnLine*, *51*(30), 800–802.

Mobley, R. K. (2002). *An introduction to predictive maintenance.* Elsevier.

Nakagawa, T. (2006). *Maintenance theory of reliability.* Springer Science & Business Media.

Ran, Y., Zhou, X., Lin, P., Wen, Y., & Deng, R. (2019). A survey of predictive maintenance: Systems, purposes and approaches. *arXiv preprint arXiv:1912.07383*.

Rusdah, D. A., & Murfi, H. (2020). Xgboost in handling missing values for life insurance risk prediction. *SN Applied Sciences*, *2*(8), 1–10.

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

Sakib, N., & Wuest, T. (2018). Challenges and opportunities of condition-based predictive maintenance: a review. *Procedia CIRP*, *78*, 267–272.

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.

Shmueli, G., et al. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

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

Subramaniyan, M., Skoogh, A., Muhammad, A. S., Bokrantz, J., Johansson, B., & Roser, C. (2020). A data-driven approach to diagnosing throughput bottlenecks from a maintenance perspective. *Computers & Industrial Engineering*, *150*, 106851.

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2014). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, *11*(3), 812–820.

Takata, S., Kirnura, F., van Houten, F. J., Westkamper, E., Shpitalni, M., Ceglarek, D., & Lee, J. (2004). Maintenance: changing role in life cycle management. *CIRP annals*, *53*(2), 643–655.

Tidriri, K., Chatti, N., Verron, S., & Tiplica, T. (2016). Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control*, *42*, 63–81.

Ting Zheng, A. B. M. P., Marco Ardolino. (2020). The applications of industry 4.0 technologies in manufacturing context: a systematic literature review. *International Journal of Production Research*.

Vásquez, J. W., Travé-Massuyès, L., Subias, A., Jimenez, F., & Agudelo, C. (2016).

Alarm management based on diagnosis. *IFAC-PapersOnLine*, *49*(5), 126–131.

Vogl, G. W., Weiss, B. A., & Helu, M. (2019). A review of diagnostic and prognostic capabilities and best practices for manufacturing. *Journal of Intelligent Manufacturing*, *30*(1), 79–95.

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).

Yan, J., Meng, Y., Lu, L., & Li, L. (2017). Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance. *IEEE Access*, *5*, 23484–23491.

Yildirim, S. (2020, February). *K-nearest neighbors knn explained @ONLINE.* Retrieved from `https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3`

Yiu, T. (2019, June). *Understanding random forest @ONLINE.* Retrieved from `https://towardsdatascience.com/understanding-random-forest-58381e0602d2`

Yong Yin, K. E. S., & Li., D. (2018). The evolution of production systems from industry 2.0 through industry 4.0. *Journal of High Technology Management Research.*

Zeki Murat Çınar, Q. Z. O. K. M. A., Abubakar Abdussalam Nuhu, & Safaei, B. (2020). Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. *Sustainability Journal.*

Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: a survey. *IEEE Systems Journal*, *13*(3), 2213–2227.

Zulkifli, H. (2018, January). *Understanding learning rates and how to improve performance in deep learning @ONLINE.* Retrieved from `https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059c1c10`

A

Exploratory analysis

Combinations	Number of origin	Number w/ alarms	Percentage (%)
9611 1081	6303	160	2.54
9571 1066	6037	230	3.81
9611 1083	2560	59	2.30
7281 1079	2360	52	2.20
9651 1095	1214	59	4.86
9571 1078	935	39	4.17
7281 1070	800	24	3.00
9571 1069	680	27	3.97

Table A.1: Percentage of the combinations of Part A1 and Part A2 that causing the alarms - Machine A

Part A1	Number of origin	Number w/ alarms	Percentage (%)
9571	7651	284	3.71
9611	8855	283	3.20
7281	3160	214	6.77
9651	1214	54	4.45

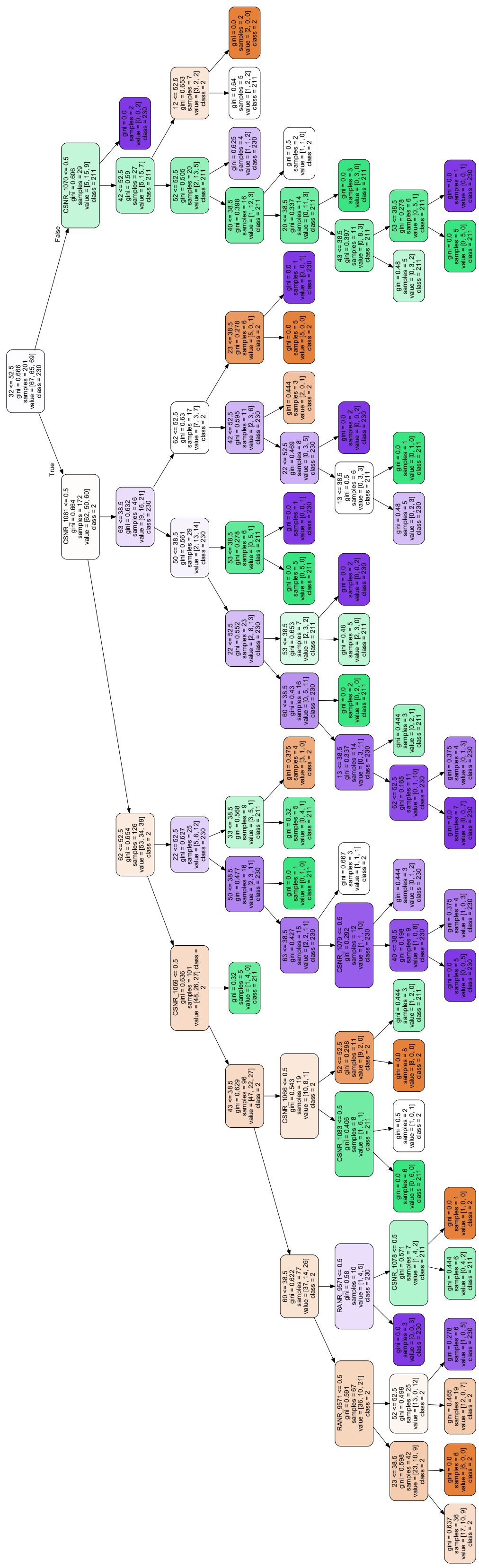
Table A.2: Percentage of the Part A1 that causing the alarms - Machine B

Combinations	Number of origin	Number w/ alarms	Percentage (%)
9611 1081	6300	197	3.13
9571 1066	6036	227	3.76
9611 1083	2554	86	3.37
7281 1079	2360	162	6.86
9651 1095	1214	54	4.45
9571 1078	935	29	3.10
7281 1070	800	52	6.50
9571 1069	680	28	4.12

Table A.3: Percentage of the combinations of Part A1 and Part A2 that causing the alarms - Machine B

B

Decision tree



C

Decision rules

NO.	Rules	Observed	True	Accuracy (%)
1	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 <= 52.5) **and** (Moment Position 43 <= 38.5) **and** (Moment Position 60 <= 38.5) **and** (RANR_9571 = 1) **and** (Moment Position 52 <= 52.5) then **Alarm ID:** 2 **(**proba: 63.16**%)** | based on 19 samples	11	7	63.6
2	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 <= 52.5) **and** (Moment Position 43 > 38.5) **and** (Moment Position 52 <= 52.5) then **Alarm ID:** 2 **(**proba: 100.0**%)** | based on 8 samples	2	1	50.0
3	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 <= 52.5) **and** (Moment Position 43 <= 38.5) **and** (Moment Position 60 <= 38.5) **and** (Moment Position 23 > 38.5) then **Alarm ID:** 2 **(**proba: 100.0**%)** | based on 6 samples	1	1	100.0
4	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 <= 52.5) **and** (Moment Position 43 <= 38.5) **and** (Moment Position 60 <= 38.5) **and** (RANR_9571 = 1) **and** (Moment Position 52 > 52.5) then **Alarm ID:** 230 **(**proba: 83.33**%)** | based on 6 samples	1	1	100.0
5	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 <= 52.5) **and** (Moment Position 43 <= 38.5) **and** (Moment Position 60 > 38.5) **and** (RANR_9571 = 1) then **Alarm ID:** 211 **(**proba: 66.67**%)** | based on 6 samples	3	3	100.0
6	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 <= 52.5) **and** (Moment Position 43 > 38.5) then **Alarm ID:** 211 **(**proba: 100.0**%)** | based on 6 samples	3	3	100.0
7	**if** (Moment Position 32 <= 52.5) **and** (CSNR_1081 = 1) **and** (Moment Position 63 <= 38.5) **and** (Moment Position 50 <= 38.5) **and** (Moment Position 22 > 52.5) **and** (Moment Position 53 <= 38.5) then **Alarm ID:** 211 **(**proba: 60.0**%)** | based on 5 samples	2	2	100.0
8	**if** (Moment Position 32 <= 52.5) **and** (CSNR_1081 = 1) **and** (Moment Position 63 <= 38.5) **and** (Moment Position 50 > 38.5) **and** (Moment Position 53 <= 38.5) then **Alarm ID:** 211 **(**proba: 100.0**%)** | based on 5 samples	1	1	100.0
9	**if** (Moment Position 32 <= 52.5) **and** (CSNR_1081 = 1) **and** (Moment Position 63 > 38.5) **and** (Moment Position 62 <= 52.5) **and** (Moment Position 42 <= 52.5) **and** (Moment Position 22 <= 52.5) **and** (Moment Position 13 <= 38.5) then **Alarm ID:** 230 **(**proba: 60.0**%)** | based on 5 samples	4	4	100.0
10	**if** (Moment Position 32 > 52.5) **and** (Moment Position 42 <= 52.5) **and** (Moment Position 52 <= 52.5) **and** (Moment Position 40 <= 38.5) **and** (Moment Position 20 <= 38.5) **and** (Moment Position 43 > 38.5) **and** (Moment Position 53 <= 38.5) then **Alarm ID:** 211 **(**proba: 100.0**%)** | based on 5 samples	1	1	100.0
11	**if** (Moment Position 32 > 52.5) **and** (Moment Position 42 > 52.5) **and** (Moment Position 12 <= 52.5) then **Alarm ID:** 211 **(**proba: 40.0**%)** | based on 5 samples	3	2	66.7
12	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 > 52.5) **and** (Moment Position 22 > 52.5) **and** (Moment Position 33 > 38.5) then **Alarm ID:** 2 **(**proba: 75.0**%)** | based on 4 samples	2	2	100.0
13	**if** (Moment Position 32 <= 52.5) **and** (CSNR_1081 = 1) **and** (Moment Position 63 <= 38.5) **and** (Moment Position 50 <= 38.5) **and** (Moment Position 22 <= 52.5) **and** (Moment Position 60 <= 38.5) **and** (Moment Position 13 <= 38.5) **and** (Moment Position 62 > 52.5) then **Alarm ID:** 230 **(**proba: 75.0**%)** | based on 4 samples	2	2	100.0
14	**if** (Moment Position 32 > 52.5) **and** (Moment Position 42 <= 52.5) **and** (Moment Position 52 <= 52.5) **and** (Moment Position 40 <= 38.5) **and** (Moment Position 20 > 38.5) then **Alarm ID:** 211 **(**proba: 100.0**%)** | based on 3 samples	2	2	100.0
15	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 > 52.5) **and** (Moment Position 22 <= 52.5) **and** (Moment Position 50 <= 38.5) **and** (Moment Position 63 > 38.5) then **Alarm ID:** 2 **(**proba: 33.33**%)** | based on 3 samples	2	1	50.0
16	**if** (Moment Position 32 > 52.5) **and** (CSNR_1070 = 1) then **Alarm ID:** 230 **(**proba: 100.0**%)** | based on 2 samples	1	1	100.0
17	**if** (Moment Position 32 <= 52.5) **and** (Moment Position 62 <= 52.5) **and** (Moment Position 43 <= 38.5) **and** (Moment Position 60 > 38.5) **and** (RANR_9571 = 1) **and** (CSNR_1078 = 1) then **Alarm ID:** 2 **(**proba: 100.0**%)** | based on 1 samples	1	1	100.0

Note:
1. Moment Position 32 > 52.5 represents Position 32 with a moment value larger than 52.5.
2. CSNR_1081 = 1 means that Part A2 1081 is one of the criteria for the specific rules, same for RANR_9571 (Part A1 9571). Values equal to 0 were neglect in advance.
3. The probability stated with the rules represent the probability generate from the model while on the right hand side, the accuracy is generated from the test data calculated by true class divided by test samples.
4. The rules with an accuracy below 50% were deleted.

DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY