Maintenance Decision Support Procedures Based on Machine Learning

Nooshin Salehabadi

A Thesis

In The Department Of

Mechanical, Industrial and Aerospace Engineering (MIAE)

Presented in Partial Fulfilment of Requirements

For the Degree of

Master of Applied Science (Industrial Engineering)

at Concordia University

Montreal, Quebec, Canada

November 2021

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Nooshin Salehabadi

Entitled: Maintenance Decision Support Procedures Based on Machine Learning

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Industrial Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Onur Kuzgunkaya _____ Chair

Dr. Onur Kuzgunkaya _____ Examiner

Dr. Hossein Hashemi Doulabi _____ Examiner

Dr. Mingyuan Chen _____ Supervisor

Approved by Dr. Sivakumar Narayanswamy _____
Chair of Department or Graduate Program Director

November 25, 2021 Dr. Mourad Debbabi _____
Dean, Gina Cody School of Engineering and Computer Science

Abstract

Maintenance Decision Support Procedures Based on Machine Learning

Nooshin Salehabadi

In a competitive production environment, a manufacturing company must have plans to improve production performance. To improve production performances depends on various issues such as production efficiency and machine availability. Various preventive maintenance procedures have been developed for efficiently maintaining and repairing machines and equipment in a manufacturing system to maximize machine availability and its readiness for production. In recent years, Artificial Intelligence (AI) technology has been applied in developing maintenance procedures in industries utilizing advanced information technology such as the Internet of Things (IoT). This thesis presents a machine learning model to predict machine failures and maintenance requirements for certain industrial machine tools. Machine learning methods enable manufacturing systems to make smart decisions through communications with humans and machines using sensors. A Logistic regression model is developed in this study to predict machine failures for the purposes of avoiding machine breakdowns and improving system performance. The supervised classification method was incorporated in the developed prediction model. The developed model is tested verified with realistic machine maintenance data. Computational experiments are conducted with results analyzed.

Acknowledgments

I would like to thank Dr. Mingyuan Chen, my supervising professor, for his guidance and support in this research project. Dr. Chen provided tremendous guidance and direction in my thesis work, from getting started with reading relevant papers, to the final stages of evaluation and writing my thesis. I also would like to thank the rest of the committee members, everyone who participated in the experiments for the thesis.

My heartfelt appreciation goes to my parents, Mohammad and Fereshteh, who have always supported me in every step of my life. Their constant devotion and encouragement were among the most significant factors that helped me complete this challenging journey. I would like to thank my sister and brother, Nazanin and Omidreza, for being always supportive and helpful.

Last but not least, I would like to express my love and deepest appreciation to my husband, Mohamadhosein, for her endless support, encouragement, understanding, and patience.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

This chapter provides information about the importance of maintenance procedures for system engineering. It introduces different types of maintenance procedures such as planned maintenance, unplanned maintenance, and breakdown maintenance with the cost of maintenance and breakdown in a production considered as important aspects.

1.1 Introduction

Maintenance engineering requires applying engineering concepts to optimize the equipment, procedure, and budgets to achieve better productivity. In general, to achieve better reliability and availability of equipment, maintenance engineers follow certain procedures in maintenance practice. In addition to probability and statistics, a maintenance engineer must have sufficient practical experience and be familiar with the equipment and machines to perform maintenance tasks.

A manufacturing or production system normally requires maintenance engineers to plan and implement effective maintenance programs. The process of equipment and machine maintenance should be monitored to be able to detect various faults that may happen during operation and typically, the purpose of maintenance is to operate and perform under normal conditions without any waste of time and failure (Joel 2009).

1.2 Importance of maintenance

Maintenance planning and scheduling play a vital role in manufacturing. Good maintenance for industrial machines is critical to solving many operational problems. Under normal conditions, a particular operating machine or equipment can be considered an important object in a manufacturing system. Regular maintenance prolongs the life of the equipment and the machines. In many cases, it is required to use a method to detect the faults in a short time period (Kou et al. 2017). On the other hand, careless maintenance negatively decreases the lifetime of the equipment. Monitoring machine conditions to predict the needs of maintenance plays a key role in a good performance and can increase productivity. Advanced technology in monitoring equipment and machine performance for maintenance emerges as technologies are developed (Masani et al. 2019).

1.3 Maintenance Plan

New technologies help manufacturing engineers and managers manage maintenance tasks and train maintenance technicians. Maintenance is necessary because without it the efficiency and availability of equipment and machines will decrease causing system breakdowns. In addition, equipment and machine tools worn out reduce the quality of products and also cause machine failure. Regular maintenance over time considerably reduces the risk of machine failure and regular service, repair, and replacement of worn-out equipment make the entire system more efficient with reduced cost (Mehmeti et al. 2012). The main purpose of regular maintenance is to ensure that the equipment works with 100% efficiency during its lifetime. Normally replacing the components periodically increases manufacturing costs. Based on research conducted by Mehmeti

et al. (2018), machine failure has different reasons. One of the main reasons is improper electrical current supply, which takes 44% of all failures. The next main reason is the age of the machine which takes 22%. It is followed by the carelessness of staff with machines and equipment which takes 17%. Heavy load and other factors are respectively 13% and 4%. According to another study by Van Tung and Yang (2009), maintenance cost is between 15% to 40% of total production cost depending on product types and manufacturing processes. In recent years, advanced technology such as artificial intelligence (AI) and machine learning has been used to improve maintenance methods to prevent machine failure. Maintenance approaches can be categorized into two main groups: planned and unplanned. Both approaches are associated with high costs as they require the replacement or repair of the equipment and machines.

1.4 Different Types of Maintenance Procedure

With the complexity of manufacturing processes increasing and more sophisticated manufacturing machines are widely used in industry, maintenance operations become more important and challenging tasks. Maintenance is critical not only for preventing machine failures but also for solving various manufacturing-related problems. Maintenance is generally categorized as preventive maintenance, corrective maintenance, and predictive maintenance. It depends on the needs of a company to adopt the corresponding maintenance approach (Jimenez-Cortadi et al. 2019). These maintenance approaches are briefly presented below.

1.4.1 Preventive Maintenance

This type of maintenance occurs outside production time such as scheduled maintenance. Despite corrective maintenance, this type of maintenance happens at an appropriate time allocated for maintenance to check the machines and replace the worn-out components to avoid the happening of corrective maintenance (Coro et al. 2018).

1.4.2 Corrective Maintenance

When a fault happens, corrective maintenance is applied to solve the problems caused by the fault. It may happen during production and can cause the manufacturing process to stop. If it happens, production will decrease and cost will increase. In addition, the time needed to perform corrective maintenance is not predictable (Vathoopan et al. 2018).

1.4.3 Predictive Maintenance

Predictive maintenance (PdM) is to predict failures before they occur. Predictive maintenance, if performed properly and correctly, should prevent sudden breakdowns of the machines or the system. Advanced sensoring technology and sensors have critical roles to play in collecting data and measuring parameters related to possible machine failures (Jimenez-Cortadi et al. 2019). Using predictive maintenance may reduce the time of periodical and unnecessary maintenance tasks (Mobley 2002).

Predictive Maintenance (PdM) is to prevent machine failure and to improve machine availability. PdM can be considered as a practical strategy in designing and developing an embedded system to predict the status of a machine to be working or not (Paolanti et al. 2018).

The advent of the Internet of Things (IoT) and different machine learning methods has affected manufacturing systems through the use of sensors to collect a large set of data on machines. Artificial intelligence technology may also help manufacturing system managers and technicians to predict and reduce the frequency of machine failures. One of the aspects of the application of predictive maintenance is to estimate machine maintenance time (Traini et al. 2019).

1.5 Research Objectives and Contributions

The primary objective of this thesis study is to develop a machine learning model to predict the failure of certain manufacturing equipment. This model takes the data from sensors installed on the considered machines to monitor the conditions of an industrial machine and to determine the probability of occurrence of machine failure.

The second objective of this study is to identify the significance of the signal in detecting machine failure. Different levels of significance of the signals indicate various situations of the monitored machine. The third objective of this study is to investigate the accuracy of the predictions made by the model and the level of performance of the developed model.

The main contributions of this thesis research are:

- This research has made an attempt to effectively utilize the Logistic regression model to predict the failures for the purpose of preventive maintenance of the system. The existing research in this area is very limited as shown in the literature.
- The general Logistic regression model is revised with parameters tested to identify the values to be suitable for predicting the failures of the system studied in this research.
- The specific Logistic regression model developed in this research is extensively tested with real data from an engineering system. The effectiveness and advantages of using such a

model for solving problems arising from preventive maintenance practice are demonstrated.

1.6 Organization of the Thesis

The thesis is organized as follows.

- Chapter Two: It provides a review of predictive maintenance procedures and discusses related works that are done in this and related areas. There is a wide range of models and techniques developed for failure prediction using machine learning.
- Chapter Three: This chapter includes the structure of the developed model for failure prediction. Machine learning algorithms are incorporated in the model development considering the type of predictions and the type of target variables in selecting a proper algorithm. The assumptions to ensure proper use of the algorithm are presented.
- Chapter Four: This chapter presents the details of the procedure for predictive maintenance based on the Logistic regression model. Numerical analysis and computational results are also presented.
- Chapter Five: This final chapter summarizes the study and presents conclusions of the research with possible future work in this area.

Chapter 2 Literature Review

Electromechanical machine failure leads to high-cost maintenance and unplanned production breakdown. The main purpose of maintenance engineers is to keep industrial machines in normal and stable conditions. Various fault detection methods are applied to ensure machine performance. Machine fault identification can be determined by different methods. They are categorized as temperature monitoring, lubricant, noise and vibration signature analysis, analyzing instruments, and various signal conditioning. Visual inspection is one type of monitoring that can be useful and practical for a maintenance procedure. The visual inspection is explained as follows and then some of the other identification will be explained (Jayaswal et al. 2008).

2.1 Visual Inspection

The visual inspection normally can be utilized in the industry as it does not need more analysis to keep the system alive. The primary devices can be used for visual inspection such as magnifying glass and a low-power microscope. Other forms of visual monitoring include utilizing dye penetrates, heat sensitive or thermographic paints to determine the cracks of equipment's surface. Typically visualization assists in many terms like wear on surfaces or investigating the condition of lubrication, the appearance of teeth, wear, and overloading the machines.

2.2 Temperature Monitoring

Temperature monitoring includes the operational temperature, the temperature of equipment, and component surfaces. Some of the researchers find that the source of the component's temperature

is the wear of the machine's elements. The temperature monitoring techniques can be done by devices such as resistance thermometers, optical pyrometers, thermography, and thermocouples.

2.3 Vibration Signature Analysis

Vibration signature analysis is the most common and popular machine fault identification compared with other types. Vibration monitoring includes all vibrations that a specific machine produces during operation time. When a machine operates in a normal condition, it has a small and constant vibration. When a fault occurred during the production, so some of the dynamic processes will change the spectrum of the vibration.

2.4 Noise Analysis

Another type of identification is noise analysis. Noise analysis can provide valuable information. The noise signal is measured at the external surface of industrial machines and collects useful information about operational machine conditions. When the machines works under normal condition the noise frequency spectra have a specific shape. Although when a fault occurs the frequency spectra shape will change. Typically the noise frequency spectra do not observe easily and sometimes the noise signal merges with other signals. In the sever merging the different signals, image processing methods could help to detect noise but generally the image processing is a complex approach.

2.5 Maintenance Performance

In a competitive manufacturing world applying new methods and technologies is an enterprises' necessity. Optimization techniques are required for production processes. Maintenance is one of the aspects that could assist companies in optimization processing (Joel 2009). The maintenance procedure needs to ensure that conditions are as follows:

- The equipment works optimally in a normal condition with the lowest possible cost.
- Significantly, the maintenance does not affect the time for delivery to customers.
- The availability and the performance of machines and equipment are reliable.
- Typically the performance of the industrial machines should keep in condition with the fewest breakdowns.
- The maintenance cost should keep at a constant level.
- Equipment's lifetime is prolonged with considering the maintenance and it affects avoiding unnecessary replacements.

In general, the maintenance procedures enhance the performance of equipment. Furthermore, it faces many challenges because of the advent of new technologies. The emergence of the Internet of things and machine learning are found in various research domains such as manufacturing and production, image processing, medicine, autonomics, and other subjects (Masani et al. 2019). Artificial Intelligence evolved the process of maintenance in the modern digital world. Applying machine learning algorithms that learn from experiences is practical and useful to detect faulty equipment. The machine learning algorithms are used to increase the accuracy of the production machine (Shen et al. 2020).

2.6 Predictive Maintenance

The industrial world tends to transform into a technological world of production. Data analytics is an approach to enhance efficiency, cost reduction, increasing safety and production performance (Fernandes et al. 2020). The recent trends in digitalization define the fact that the world faces a huge amount of massive data. In this competitive market, advanced production systems require using a wide range of technology maintenance management, increasing production, and professional technician. Predictive maintenance is one of the aspects to reach the goal (Kanawaday et al. 2017). Predictive maintenance has a key role to decrease maintenance intervals. Typically artificial intelligence and the Internet of Things provide the opportunity to reduce the machine's breakdowns, supply chain improvement, and grow the production process. Developing the machine learning algorithm requires a big amount of machine data. Traditionally visual inspection is one of the monitoring methods to collect the required data. Monitoring automatically is new and necessary to record the dataset (Masani et al. 2019). The steps for predictive maintenance include data acquisition, data processing, and machine decision-making (Jimenez-Cortadi et al. 2019).

2.7 Condition Monitoring

Condition-based maintenance nowadays is commonly used. Maintenance procedure under supervision is a critical factor to avoid unplanned stopping production lines and reach the high rate of production. Condition monitoring is implemented differently. By the usage of sensor technology, recording and collecting useful information is possible. The first objective for collecting required information is to determine which parameters are necessary for analysis and prediction. It is critical to know the detail of information and maintenance time which has been

recorded during the operation of systems. Condition monitoring utilizes two or more parameters that affect machine failure (Hassan et al. 2018).

Traini et al. (2019) studied the maintenance prediction in the milling machine. In this research tool condition monitoring is one of the unique condition monitoring that is applied. The parameters which have been measured in this study are vibration, acoustic emission, and temperature. These parameters can affect the tool wear lifetime. In the milling machine, there is a component called metal cutting that has a significant effect on the quality of the milling process. While this component is not broken down, the machine will work but not with high efficiency so it requires to replace after its lifetime. In this research, useful information is recorded by applying the sensor's processing. Sensors record the measurement of parameters to develop a machine learning model. Machine learning models assist to predict the efficiency and lifetime of cutting metal.

The parameters measured by tool condition monitoring categorize into two groups: a-priori and a-posteriori. The first one is the machine parameters specification and the second one is measured by the sensors during the test.

The research conducted by Uhlmann et al. (2018) investigated the selective laser machine (SLM) and utilized the condition monitoring of the dataset. This study used offline data to investigate the health of the product. Monitoring assists to detect the three types of failure. Three sensors were applied to measure the parameters. Temperature, oxygen percentage, and pressure are the parameters to measure. These three parameters are considered as the independent variables.

Another research explains the predictive maintenance of an industrial machine in a cement plant. Power report is condition monitoring that implemented in this study. Energy is measured by two kinds of convertors that are utilized as inputs for the model. These energy parameters define the performance of a machine (Masani et al. 2019).

Murphy (2006) studied fault diagnosis in electric drives. This fault detection identifies the faults in switches of an inverter that cause the failure. Fault happens if the switch fails and it may affect the performance of synthesis voltage that is required for the motor terminal. The motor terminal supplies the electric power of a machine.

J-Cortadi et al. (2019) has researched in the field of predictive maintenance on machining processes and machine tools. This study emphasizes at the beginning of the 20th-century, maintenance became a concern not only for solving the machine failure but also for preventing the failure problems. The methods for collecting the dataset divides into two main groups: event data and condition monitoring data. Event data are the data that provides information about what happened for the equipment and which type of maintenance is needed. The condition monitoring provides information about the measurement of the health of physical equipment. There are different types of sensors to collect the data such as gyroscopes, humidity, accelerometers and ultrasonic sensors, and other ones.

Due to the emergence of machine learning, sensors have a significant key to detecting and diagnosing faults. Ensuring the health of sensors to have real and accurate data is an important object to consider (Ahmad 2020). In other studies for device failure, prediction machine learning techniques are applied (Fernandes et al. 2020).

2.8 Sensor's Fault

Sensors have a key role to detect the faults of industrial machines. Although, defective sensors have a negative effect on the prediction. Saeed et al. (2021) studied the faults of wireless sensor networks used for temperature and humidity. It considered six types of common faults that sensors face: drift, bias, spike, erratic, stuck, and data loss. Implementing the sensors in different

environments leads to changing behavior of sensors because of natural conditions or interference of electromagnetic. When the sensors work in abnormal situations it can lead to a decrease in the system reliability, performance of the machine, and safety. Additionally, the output of the sensor may face other issues such as battery defects.

2.9 Data Preparation

The data collected by the sensors require cleansing and some steps to be ready for machine learning models. The advent of technology assists in the improvement of data collection by the sensors and computers which provide an easy way for collection data (Galeano and Pena 2019).

Some analogical sensors which record the data, require an application to prepare accurate data for machine learning models. Appropriate data size and data collection assist to apply an accurate machine learning model (Borgi et.al 2017). The accurate data collection includes useful information to develop a machine learning model. The new technology provides the opportunity to collect the operation and process condition data from the equipment and machines (Dai and Gao 2013).

Preparation data for both methods of classification and regression methods are the same. Training the machine learning models and the validation method for the developed model is the difference between the two methods of classification and regression (Abdelkrim et al. 2019).

2.10 Machine Learning Models

Murat Cinar et al. (2020) studied the applied machine learning models on predictive maintenance with smart manufacturing. This research discussed collecting the operational and process condition

data with smart sensors which are built-in or external sensors. All the data management and data accumulation are done through the Internet of Things. The goal of this research is to minimize downtime and increase the production rate and lifetime. The type of classification or regression machine learning model which is utilized for predictive maintenance depends on the type of collected data. If the predictive maintenance model is not appropriate, it may waste the time and cost. Supervised learning, unsupervised learning, and reinforcement learning are three different groups of machine learning. Unsupervised learning is a method, which there is no label and this algorithm decides to classify the data. Furthermore, there is no feedback from the learning, for example, it is like a piece of news that is got by different sources. In supervised learning, there is labeled and it consists of the regression and classification methods. There are several methods for supervised learning, like Logistic regression, decision tree, support vector, artificial neural networks which are part of classification methods and will explain shortly in this research.

- Artificial neural networks (ANN): ANN method is an intelligent computational method that is developed by considering the human brain (Deepika et al. 2018). In this method, several neurons received the signals and they work together. The neurons receive the signals by the dendrites and transform them by the axons. At the end of axons, there is a synapse as the correction. In the ANN method, there are several layers as input and hidden layer and one output layer. This method is commonly used for complex systems which have not enough information (Zhang et al. 2018).
- Support vector machine (SVM): The SVM method is one of the classification methods with a statistical learning model. This method divides the dataset into two classes and after developing this model it categorizes the data in new classification by the assist of a hyperplane between two categories (Deepika et al. 2018).

- Decision tree: This method is another classification method with decision nodes and this node divides into other branches and nodes (Deepika et al. 2018).
- Random forest: This method is defined as a classification method. This model utilizes the ensemble algorithm that has several decision trees as classifiers to predict (Deepika et al. 2018).
- Logistic regression: This method is used for prediction and considering the binary target value. This model does the prediction based on the probability of occurring an event. There are various metrics to evaluate this method and determine its goodness (Deepika et al. 2018).

This research presented several classification machine learning algorithms for predictive maintenance. These models are successful and based on the data type a model will be developed. Although the survey determine that just 11% of companies are interested in smart manufacturing and applied machine learning models to predict maintenance (Seebo 2019).

2.10.1 Predictive Maintenance in Milling Machine

Traini et al. (2019) presented a model to predict the maintenance of a milling machine. This research utilized condition monitoring to collect the required data to build the model. The milling machine has a rotating tool that includes various cutting edges. Cutting edge affects the job duration, quality, and cost of production. Therefore, the parameters in cutting edge are important to investigate and define the life of the tool. The collection dataset divides into two groups, the first one are the machine parameter like cutting speed, spindle speed, feed rate, depth of cut, hardness, and toughness. The other group includes the data collected by the sensors such as cutting forces, vibration emission, and acoustic emission. The preparation of the data is important and

remove some observations that have missing values. In this research various classification model (Logistic regression, random forest, neural network, and decision tree) is applied. Although the result of the Decision tree model is provided that had an accurate result. The accuracy of developing the Decision tree model was about 96%. The presented accuracy is good and can be used in the manufacturing process (Traini et al. 2019).

2.10.2 Predictive Maintenance in Selective Laser Melting Machine Tool

Uhlmann et al. (2018) studied prediction maintenance for another specific type of machine. In this study selective laser melting machine is considered for prediction. This selective laser-melting machine produces metallic components and complex geometrics. The data to develop the machine learning model is provided by three sensors. Recorded data defines the normal performance of the machine and three faulty conditions that happened. Three different sensors collect the needed information. These three sensors are temperature, oxygen percentage, and pressure. If a failure occurred during the manufacturing process this fault affects the sensors.

The data is collected by 20 sensors which are recorded for 206 manufacturing processes in every second of operation. So this model has three independent variables as mentioned earlier (temperature, pressure, and oxygen). The statistical calculation is used to calculate minimum, maximum, average, median, skewness, mode, and standard deviation for each column. Therefore there is a matrix of seven rows and three columns and there are 206 manufacturing processes. The clustering machine-learning algorithm is developed for this study with the Elbow method. The Elbow method determines the number of clusters for modeling. The number of clusters for this model is three (Uhlmann et al. 2018). Developing the clustering model detects the early faults.

2.10.3 Predicting Maintenance and Monitoring of Industrial Machines

Masani et al. (2019) studied maintenance prediction for industrial machines. In this research, condition monitoring is considered to energy meter parameters to collect the power report. The issue for the machine learning model was about how to prepare the data which is collected by the energy meter parameters. Supervised learning is used for classification analysis. There are different techniques like the Logistic regression model, Bayesian network, K nearest neighbor, SVM, and Decision tree for prediction in this study. A Decision tree with binary classification was the model that is utilized in this study. Power is the dependent variable and the current phase is regarded independent variable or predictor for this research. Power feeds the machines to work well and increases the performance of the machines. The features for this model are measured by the different parameters. The parameters that are regarded for the power are the average voltage and current of phase one, phase two, and phase three. The power formula is calculated by multiplying the voltage and current of the line and with a power factor that has a constant value. The current of the line depends on the load of material on the machine. In the power formula with increasing the I_L the power and efficiency of the machine increase and it assists to reduce the breakdowns of a machine (Masani et al. 2019).

2.10.4 Fault Diagnosis in Electric Drives by Machine Learning

Typically electric motor and power electric-based are the most significant components at industrial machines and equipment. Murphey (2006) studied the fault diagnosis in order to determine and locate some types of faults in an electric drive. Power electronic inverter has an important role to supply the energy in machines. These power suppliers have some switches, if a switch failed, so the process of voltage synthesis will fail. Switch failure means an open or short circuit in a device

and consequently, the motor will fail to generate the torque shaft even if the inverter supplies the voltage to the motor. In this research developing a robust machine learning model helps to detect the faults. For developing the model a set of variables such as speed and torque is sent to a simulation device to generate the signals. The signals will be trained in a neural networks machine learning model.

2.10.5 Dynamic Fault Diagnosis

Zhang et al. (2018) researched a delay dynamic coupled fault diagnosis which was a significant type of diagnosis. This model is based on a probability graph for the diesel engine. The data was collected by the six cylinders in the diesel engine. For each cylinder, there is a pressure sensor to record the information. There are 11 types of faults to detect.

This research mentioned there is some information that is not accurate and useful such as noises. Diagnosis assists to deal with these issues. One model utilized for dynamic fault diagnosis based on a probability graph is the Bayesian network. The advantage of a probabilistic graphical model is a combination of both historical data and prior knowledge. The Bayesian method in this research has some drawbacks and it needs to improve.

The Bayesian method for learning includes two separate parts, parameter learning and structure learning which is more difficult. There are four parameters: fault transition probability, fault appearance probability, fault detection, and false alarm. For the parameter learning part, the maximum likelihood estimation and Bayesian estimation are utilized. The structure learning is applied because maybe there is some dependent relationship in the model that is not checked. Furthermore, structure learning has three methods to apply: constraint-based method, score-based method, and hybrid method. After developing the model in both parameter learning and structure

learning, the accuracy is high. In addition, the results determine that the score-based method is the best method for structure learning (Zhang et al. 2018).

2.10.6 Predictive Maintenance on Real Machining Process

Jimenez-Cortadi et al. (2019) studied increasing a tool life. Applied regression machine learning methods are mentioned to predict the remaining useful life (RUL) and two methods for the classification maintenance prediction. This study is investigating the CNC center which moved on two axes. This study is applied a conservative maintenance method. When the number of pieces that changed in the machine increase up to higher than the threshold, therefore, a maintenance strategy is needed. The data are recorded during 2017 from December to 2019 May which indicates the performance of a machine. For collecting the information different signals were recorded. Although the spindle load is a more critical factor compared with other parameters because it defines the machine efforts.

The Support vector machine (SVM) and K-means classification model are applied. For predicting the RUL to avoid failure, the linear regression model is used. This method considered the time for collecting data and it is investigated that during the time some wear will happen or not (Jimenez-Cortadi et al. 2020).

2.10.7 Machine Learning Model for Predictive Maintenance on Woodworking Machine

Calabrese et al. (2020) researched the failure prediction of a woodworking machine. The data were collected during 24 hours of working this machine by proper sensors. This research explains

different types of maintenance and divides them into three groups of maintenance. Run to failure, predictive maintenance, and preventive maintenance are these three types.

- Run to failure: It is the most common method for maintenance that is applied after a failure occurred.
- Predictive maintenance: It is more practical for manufacturing performance. This method determines the time for maintenance according to the inspection of equipment. The physical parameter defines a component that needs to be repaired or replaced before happening failure. It needs some special sensors to collect the required data such as vibration, temperature, currents, and run time to failure.
- Preventive maintenance: It is a good method because provides the opportunity that to ensure the machines are available and prevent a machine-down failure. In this method, the machine has regular maintenance based on the information (average failure of a machine) recorded during the time. It helps to reduce the cost of production (Calabrese et al. 2020).

There are very few studies about preventive maintenance and it is more practical in the case of medical devices or some electronic equipment (Sipos et al. 2014).

The classification model used three different methods: distributed random forest (DRF), Extreme gradient boosting (EGB), and Gradient boosting machine (GBM). The data was divided into 70% for the train set and 30% for the test set. The result of modeling defines that the gradient boosting model has the highest accuracy compared with the other methods. Although the three models have near accuracy (Calabrese et al. 2020).

The model can be trained by the artificial neural network (ANN), Bayesian networks (BN), Support vector machines (SVM), and Hidden Markov model. The ANN method is most commonly used for the remaining useful life and, it is strong enough for nonlinear simulations. BN is a

statistical acyclic graphic that each node in this graph defines a variable. This variable can take a value of discrete or continuous.

2.10.8 Forecasting of Appliance Failure

Fernandes et al. (2020) studied forecasting malfunctioning devices and mentioned that these failures increase the cost. The goal of this research is to improve efficiency and cost reduction. This study discussed the fault detection on heating, ventilation, and air conditioning (HVAC) systems, especially on heat pumps and boilers. Boilers have a critical role to supply hot water for the customers. Installing the boilers in the customer's house satisfies their need for hot water. As the same for other research collecting the data is the first step for modeling. Necessary information is collected by several heating appliances. The data was provided during 16 months from 1000 different appliances. Each of the boilers has some sensors to collect the required information. Sensors provide information such as temperature, the number of boilers, the start number of requests for heating, and the duration of each cycle. This model predicts failure based on previously collected data for failure conditions. This research predicts failure for boilers up to seven days before a failure occurred. The Neural network model is utilized for this research and the number of layers is three hidden layers and 15, 25, and 50 neurons. The accuracy of the model was not good enough and the data was unbalanced, therefore it was mentioned as future works to increase the accuracy.

2.10.9 Predictive Maintenance System

Kaparthi and Bumblauskas (2020) studied a large agricultural equipment manufacturing company. The main reason for research is to increase the operation time of the machine. In this research, a new strategy for the predictive maintenance of machines and increasing performance and efficiency is proposed. Machine learning techniques assisted to perform and developing a prediction model. The model developed in this research was a Decision tree and the results were good. This model is applicable for other industrial machines and can be generalized. Developing the model is based on the past observation that is utilized as input variables. The input variable can take both discrete and continuous values. Furthermore, there is a discrete target value and the goal is to find the relationship between the inputs and output or target variable.

The ability of a Decision tree to predict unseen or future data is so important and it is practical if the model is generalized. One of the main subjects that reduce generalization is overfitting. In this research ensembles method is applied to reduce the overfitting. Ensembles techniques work like a Random forest, it chooses a random sample of the subset to make the multiple trees and choose the final one. Logistic regression and Random forest were the other two classification models that are developed in this research. The result of the confusion matrix and the accuracy determine that the random forest makes a good prediction compared with the other two models. Developing a Random forest model has an accuracy of about 73% but Logistic regression and Decision tree respectively 63.4% and 63.1% (Kaparthi and Bumblauskas 2020).

These techniques and the accuracy of the model defined that the model was successful and the prediction was useful.

2.10.10 Fault Diagnosis of Power Distribution Lines

Togami et al. (1995) studied fault diagnosis on power distribution lines by developing machine learning models. In this study, the Decision tree method was applied for prediction. This model needs to revise because the sensors can be affected by the noise. This diagnosis is one of the important diagnoses in the field of electronics. Detecting fault in distribution power lines is so important because it supplies the electric power that is the necessary facilities in every country. The model has three lines and three feeders. The fault occurs between the source and the end of the feeder, that fault resistance is between 0-300 ohms. This model aims to determine the faults in distribution power by developing a Decision tree model. This model was developed by utilizing the sensor information which includes the fault between two lines and the fault between line and ground. The machine learning algorithm developed and detected the faults in distribution power.

CHAPTER 3 Logistic Regression Model

Prediction models can be used to predict the failure of industrial machines, by applying the machine learning methods in both regression and classification models. Failure prediction plays an essential role for companies to reduce their cost. These models provide opportunities for companies and factories to forecast the situation of devices. Enterprises investigate utilities and collect needed data for future conditions. With this approach, they can identify the factors that affect the production line. Fault diagnosis assists to know the approximate time of maintenance and repair or replacement of equipment. The primary goal of prediction models is avoiding unpredictable costs. Failure production line of a company can lead to property damage by miss of repairing inspection. Providing valuable information is the first step to performing an appropriate prediction model to reach the company's goal.

3.1 Preparation of the Dataset

Data collection is used for many purposes not especially for approaches such as machine learning or data mining. Analysis of the dataset is one of the requirements for learning systems or other goals. The collection of different types of data is important, the dataset can affect the accuracy of models. Typically, the preparation dataset is a time-consuming process. Raw data requires preparation to make a relationship between the data and machine (Abdallah et al. 2017).

Pre-processing the data is the first step in this dissertation, the Logistic regression model is regarded to perform prediction. Statistical features such as minimum, maximum, mean, skewness and, the standard deviation for some independent variables (inputs) were calculated (Uhlmann et

al. 2018). Finding missing values, removing duplicates, checking the types of variables balancing the dataset are regarded as the steps of preparation.

3.1.1 Dummy Variables

Categorical features should be changed to dummy variables. The number of dummy variables is the same as the number of categories for categorical features. The dummy takes the value of one for the presence of the categorical variable and on the contrary, takes the value of zero for the absence of it (Hosmer & Lemeshow 2013). In this dissertation, some categorical features are binary therefore, they take the value of zero or one so they do not need to be considered as dummies. But for one of the features dummy variable is required. This feature in the dataset is a categorical independent variable whose name is Model and has four different values. Therefore, to take the values zero or one it needs the dummy variable as shown below:

Model ϵ {1,2,3,4}

Dummy variables for Model:

Model1= (1,0,0,0), Model2= (0,1,0,0), Model3= (0,0,1,0), Model4= (0,0,0,1).

3.1.2 Splitting the Dataset

For developing the model the dataset will be divided into two groups of sets. Train set and test set for the validation process. There are about 80% of the dataset in the train set and the remaining 20% for the validation process. In this dissertation, the dataset that is used includes 15 features that are used as independent variables. The dataset matrix has 876100 rows that show features measured during 24 hours of each day in one year. So based on the 876100 rows which define the

observation the test set has approximately 175220 records and the 80% including approximately 700880 records for the train set.

The number of rows shows the number of observations, for instance, the row number zero defines the observation zero and the row number n defines the observation number n. Furthermore, the number of columns explains the number of features.

3.1.3 Balancing the Dataset

Imbalanced data is one of the issues encountered when applying a machine learning algorithm and it can be effective on the results of prediction. Python provides a different method to cope with these problems. In an imbalanced dataset called x, there are two subsets of the majority (x_{maj}) and minority (x_{min}). This minority subset cannot be predicted accurately due to the model's desire to train the majority subset. The balancing ratio is defined by the number of the minorities over the majority:

$$r_x = \frac{|x_{min}|}{|x_{maj}|} \tag{3-1}$$

and when the data changes to a balanced dataset we have a new resample data that is called (x_{res}) such that $r_x < r_{res}$.

In the first step, the different methods that are used to deal with this problem are explained shortly. These methods divide into four groups

- Under-sampling method
- Over-sampling method
- Combination of the two methods (over and under sampling)
- Ensembles learning method.

Imbalanced problems exist in different areas: telecommunication, medical diagnosis, fraud detection. Under-sampling and over-sampling are the most common methods for balancing the dataset and each of them has its advantages and disadvantages.

3.1.3.1 Under-sampling

Under-sampling is a method that reduces the samples on the majority subset (x_{maj}) to balance the distribution of the dataset. This method may lose some useful information that can affect the result of prediction. In other words, in the under-sampling method, the majority class reduces to the minority class.

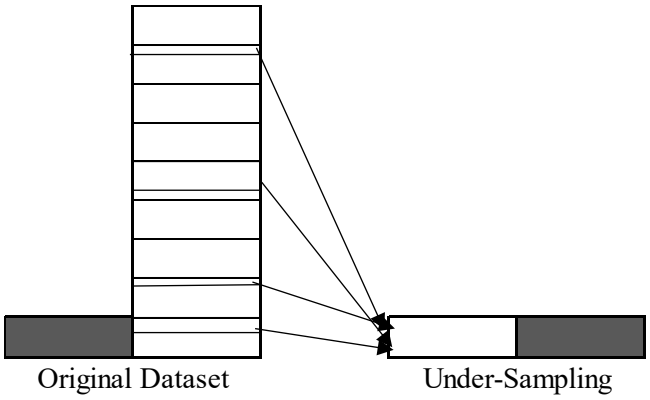

Figure 3.1: Under-sampling process

The under-sampling method divides into two groups and for each one, there are different methods to perform the under-sampling method. The first group is fixed under-sampling. In this group, some methods are performed to reach the suitable balancing ratio (r_{res}). The most common methods are Random Under Sampling, Cluster Centroids, and NearMiss. The second group is cleaning under-sampling: this group does not allow to reach the appropriate balancing ratio and clean the space feature like Tomek Links method. For each of the groups, there are different methods to perform the under-sampling (Lemaitre et al. 2016).

- Random Under Sampling: This method selects the random samples from the majority class to reduce the size of the majority and it will repeat to the appropriate size. Typically it may delete some useful information needed for tarin and it cannot be useful for every situation. Also, the information to train and fit the model may not be sufficient.
- Cluster Centroids: This method generates a new under-sample dataset based on centroids by the clustering method. It works with cluster centroids with the KMeans algorithm. KMeans algorithm chooses the k cluster with the k points that are centroids of each cluster. It will assign each dataset to the closest centroid and then measure it by Euclidian distances.
- NearMiss: This method assists in balancing data by choosing n-neighbors. This method selects n samples from the majority class by considering which has the smallest average distances from the minority class.

3.1.3.2 Over-sampling

Oversampling increases the number of samples on the minority subset (x_{min}) to balance the distribution to train the dataset with sufficient information.

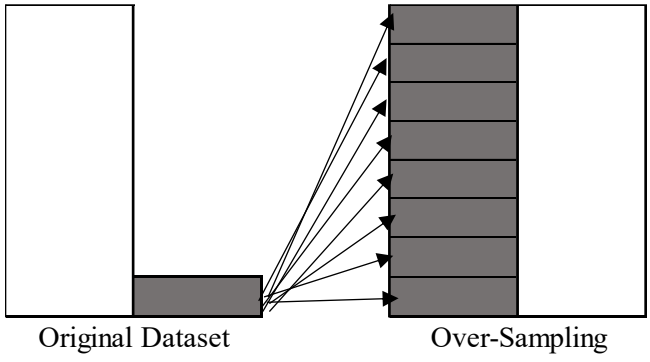

Figure 3.2: Over-sampling process

There are different methods for oversampling the dataset like Random Oversampling, Smote, Boreline smote, and Adasyn.

- Random Oversampling: This method is one of the simplest methods that is used. For balancing the dataset the minority classes replicate randomly. One of the advantages of applying this method is the low risk of missing information and its disadvantage is prone to overfitting. Therefore, it is important to check the overfitting and develop the model by some techniques that avoid this problem.
- Smote: This method is very practical and widely useful because, opposite of the Random Oversampling which has the risk of overfitting, this method does not have this problem. The synthetic Minority Oversampling technique comes to synthesize the new samples from the existing sample. It works by considering the k- nearest neighbor algorithm to generate new synthetic data. To achieve this goal, a line is drawn as the feature vector in the feature space and determines the k nearest neighbor. The next step is calculating the distances between two samples and multiplying the distance by a random number between zero and one to determine new artificial data on a line distance and repeat the process.
- Borderline Smote: In the smote method there are bridges because of the creating new point on the minority class. The Borderline smote method improves this method by dividing the minority class points into two groups of border points and noise points. Border points are the minority class points in which their neighbors are both minority and majority class points. Noise points are the minority points which most of its neighbors are the majority class points. The borderline method desires to synthesize points from the border points and it ignores the noise points.

- Adasyn: Adasyn method is Adaptive Synthetic Sampling, this method generates the synthetic data by considering the data density. Creating the synthetic data is based on the density of the minority class. In other words, the synthetic data are generated more in the area of low density compared with areas that have higher density in the minority class.

In this dissertation, the Adasyn method is applied to balance the dataset.

3.2 Logistic Regression

There is a lot of significant situation to know the relationship between input and output such as regression, in this situation label values (y) are continues and real-valued. On the other hand, there are many situations in which label values (y) are categorical and discrete values. To predict the categorical label values the classification models can be used. Different classification algorithms predict the output, these kinds of algorithms train the model and guess the outputs of test set inputs. Both regression and classification models are important topics in statistics and machine learning (Jurafsky & Martin 2020).

Logistic regression is divided into three different groups of binary Logistic regression, multinomial Logistic regression, and ordinal Logistic regression (Park 2013). In this dissertation because of the type of dependent variable binary Logistic regression is developed.

A binary or dichotomous Logistic regression model is used when predicting the situation has two responses like zero or one, pass or fail, yes or no, good or bad. In this dissertation, if the y variable takes zero value, the industrial machines will work. Otherwise, the industrial machine will be failed. The multivariable analysis is used for modeling. This analysis utilizes multiple variables (predictors) to predict a single outcome (Katz 1999). The multivariable method makes a

relationship between predictors (independent variables) and an outcome (dependent variable). The model expresses the predicted value of outcome by the sum of multiplying each independent variable by the coefficients. The coefficients determine the effect of each independent variable on the outcome variable. Logistic regression (or it called the logit model) investigate the relationship between multiple independent variables and categorical dependent variable and explore the probability of an event by applying the Logistic curve to fit the model. In other words, this model estimates the probability of occurring an event based on fitting the data into a Logistic curve.

In a binary Logistic regression model, the goal is to calculate the conditional probability P(Y=1|$X = x$) as a function of x. Typically the Logistic regression calculates the probability of occurring an event over the probability of not occurring the event and the effect of independent variables will be explained by the odds (Park 2013).

In the first step, it is better to talk about a mathematical model to make the classifier that can help to make the decision. Classifier requires the train of some pair of x as inputs and y as outputs. The sigmoid function is the classifier in this model to make a decision. In Logistic regression when the output is one, it means the output belongs to this class, and zero means it does not belong to this class. With considering this method, we know the probability of a variable belonging to a class or not (Jurafsky and Martin 2020).

As discussed earlier for the binary Logistic regression the probability of getting variable y given an explanatory x is calculated as below:

$$\pi(x) = p(y = 1|X = x) = 1 - p(y = 0|X = x) \quad (3\text{-}2)$$

For calculating π(x), the exponential function $\exp(b + wx) = e^{b+wx}$ is used as below:

$$\pi(x) = \frac{exp(b+wx)}{1+exp(b+wx)} \quad (3\text{-}3)$$

The Logistic regression has a linear form for the logit of the success probability

$$logit(y) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = b + wx \quad (3\text{-}4)$$

The Logistic regression method solves problems by considering the learning from a train set by a vector of weights and a bias term (b) that is called an intercept (Agresti, 2007). Each input (x_i) has a wight of (w_i) which defines its importance (Jurafsky and Martin 2020).

For better understanding, a sigmoid function and its curve are shown below:

$$y = \frac{1}{(1+e^{-z})} \quad (3\text{-}5)$$

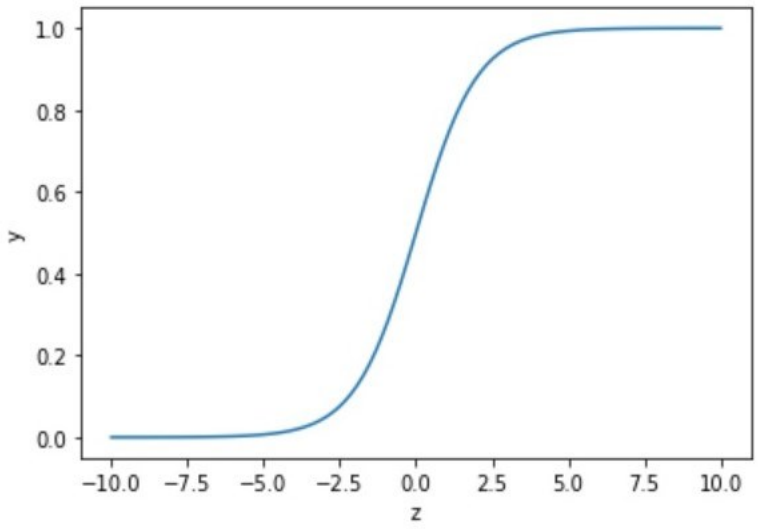

Figure 3.3: Sigmoid function

Figure 3.3 shows the Logistic function or Logistic curve. The sigmoid function uses this instead of the straight line in the regression model.

In the figure, the range of z is between -10 to 10, while the range of y is between 0 and 1 that illustrates the probability. The z passes through the sigmoid function for creating probability. It is the point that is used for the probability because the sigmoid function applied the summation of weights multiplied by x plus b. Therefore, y just takes two values of 0 or 1.

The regression model uses the straight-line because the variable y is continuous and the ranges vary from -∞ to +∞ but Logistic regression estimates the probability of getting an event y by an S-shaped curve that provides the ranges between 0 to1.

After the algorithm learned weights from the train set, in the test set the classifier multiply x by w and add the value of b. The multiple Logistic regression defines as below:

$$logit(y) = ln\left[\frac{p(y = 1|x_1 x_2 \dots x_n)}{1-p(y = 1|x_1 x_2 \dots x_n)}\right] = b + \sum_{i=1}^{i=n}(w_i \times x_i) \quad (3\text{-}6)$$

Both b and w have real values.

For a given x, y takes the value one, if the probability p(y=1) is bigger than 0.5. Otherwise, it takes the value of zero (Jurafsky and Martin 2020).

$$\hat{y} = \begin{cases} 1, & \text{if } p(y = 1|x) \geq 0.5 \\ 0, & \text{Otherwise} \end{cases} \quad (3\text{-}7)$$

The calculation for the Logistic regression to reach the accurate prediction is explained systematically in the following parts.

3.3 Odds Ratio

The odds ratio (OR) is a function that is used to measure two odds relative to different events. Odds is the probability of occurring an event (π) to the probability of not occurring (1- π). The following equation expresses the odds function:

$$Odds = \frac{\pi}{1-\pi} \quad (3\text{-}8)$$

Here π is the probability of success. For example, if the probability of success is 0.75 (π=0.75) the odds ratio is equal to 3 (0.75/0.25=3). The odds is always non-negative with values greater than

one (Agresti, 2007). Therefore, in Logistic regression for having a value between zero and one that indicates the probability the natural logarithm of odds is calculated.

$$logit(y) = ln(Odds) = ln\left[\frac{\pi}{1-\pi}\right] = b + w_1x_1 + \ + w_2x_2 + \cdots + w_nx_n \quad (3\text{-}9)$$

The OR is a measure of association between an event occurring in one group, compared to the odds that occur in another group. For instance, there are two groups of events A and B, corresponding odds of occurring A relative to occurring B is estimated as the following equation:

$$OR = \frac{odds_A}{odds_B} = \frac{[\frac{\pi(A)}{1-\pi(A)}]}{[\frac{\pi(B)}{1-\pi(B)}]} \quad (3\text{-}10)$$

In other words, the OR demonstrates the odds of an outcome (failure of an industry machine) will occur with a determined exposure (such as breakdown of a component) compared to the odds without that exposure. Therefore, the OR is the ratio of odds for x=1 to odds for x=0:

$$OR = \frac{odds_1}{odds_0} = \frac{\left[\frac{\pi(1)}{1-\pi(1)}\right]}{\left[\frac{\pi(0)}{1-\pi(0)}\right]} = \ e^{w_1} \quad (3\text{-}11)$$

The OR function takes the non-negative values, while if the OR is equal to one, it means that the probability being in both groups of x=1 and x=0 is the same. When the OR is greater than one it demonstrates the outcome is most likely to happen when the x is equal to one. On the contrary, for the odds ratio of less than one, it happens when x is equal to zero (Park 2013).

3.4 Assumptions of Logistic Regression

There are some assumptions for choosing a Logistic regression model. It is essential to know the assumptions in order to have a reliable model. These assumptions include independent and dependent variables structure, no multicollinearity, linearity for independent variables and log odds and, sample size. The first assumption is to have a binary or dichotomous dependent variable and a large sample is another important factor. The binary dependent variable is checked and there are two situations of failure or not.

The definition of assumptions in the Logistic regression model is as follows:

1) There is no relationship between the dependent variables in the Logistic regression model. It means that the observations are independent of each other and should not come from repeated measurements.
2) It is not required that independent variables be normally distributed.
3) The sample should be large.
4) There is little or no multicollinearity between the independent variables in Logistic regression. In other words, the correlation between independent variables should not be too high.
5) It requires the linear relationship between independent variables and log odds. However, there is no need for a linear relationship between the independent variables and dependent variables.

As explained above assumptions three, four, and five need to be verified. To check assumptions number three and four there are some methods, following methods explain how to check these assumptions (Schreiber-Gregory 2018).

3.4.1 Sample Size

The Logistic regression model needs a large sample size and calculating the sample size is a bit complicated because of effects on some factors. Statistical power, standard error, number of parameters to estimate sample size are some factors that affect the calculation of sample size (Schreiber-Gregory and M-Jackson 2019). There are various methods for determining the sample size for the Logistic regression model. For the simple Logistic regression model with continuous variables and normal distribution, the formula is provided below (Park 2013):

$$n = \frac{(z_{1-B/2} + z_{1-w})^2}{p_1 \times (1-p_1) \times {w_1}^2} \tag{3-12}$$

For the situation when the variables are binary, there is the following formula:

$$n = \frac{(z_{1-\frac{B}{2}}\sqrt{\frac{p(1-p)}{w}} + z_{1-w}\sqrt{\frac{p_1(1-p_1) + p_2\,(1-p_2)(1-w)}{w}})^2}{(p_1 - p_2)^2 \times (1-w)} \tag{3-13}$$

The dataset that is regarded in this dissertation has continuous and categorical variables, and it can be considered as multiple variables. Therefore, none of the above formulas can be used. To calculate the sample size for this dissertation another formula is needed which can be utilized for both categorical and continuous variables. The following formula is appropriate to determine the sample size for this dataset:

$$n = \frac{10 \times k}{p} \tag{3-14}$$

Where K is the number of independent variables and P is the proportion of the minority class of dependent variables.

3.4.2 Correlation Coefficient

Correlation is a statistical method that helps us to determine the relationship between two or more variables. In statistics, the range of values for correlation coefficients is between -1 to 1. When the correlation coefficient is -1 or +1 there is a strong relationship. When the value is positive there is a positive relationship and on the contrary negative values indicate a negative relationship. Moreover, for the correlation coefficient when the value gets zero there is no relationship between variables. The prediction based on the correlation analysis is more valuable and it is close to reality (Asuero et al. 2006).

There are three various types of correlation measurement in statistics: Pearson correlation, Kendall rank correlation, Spearman correlation. In general, for developing a model highly correlated independent variables are not suggested.

3.4.2.1 Pearson Correlation

The Pearson correlation method is used to measure the degree of relationship between two linear variables. The correlation coefficient is calculated in the following formula:

$$R = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \quad (3\text{-}15)$$

Where R is the Pearson correlation coefficient and n is the number of variables in a dataset.

3.4.2.2 Kendall Rank Correlation

Kendall rank correlation is another method, that is non-parametric that defines the dependency between two variables. This method is a good alternative for the Pearson correlation and when there is a small dataset. To calculate the Kendall rank correlation the following formula is used:

$$T = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (3\text{-}16)$$

Where n_c is the number of concordant, n_d is the number of discordant and n is the sample size.

For calculating the number of concordant the ranking should be in ascending order, concordant pairs define how many larger ranks are under a specific rank.

3.4.2.3 Spearman Rank Correlation

Spearman rank correlation method is non-parametric that is developed for measuring the correlation between two variables. Spearman rank correlation does not assume any special assumption for the type of distribution between the variables. To calculate the spearman correlation the following formula is used:

$$\rho = \frac{1-(6\sum {d_i}^2)}{n(n^2-1)} \quad (3\text{-}17)$$

Where ρ is the Spearman correlation coefficient, d_i is the differences between the rank of corresponding between x_i and y_i and n is the number of variables in the sample.

In this dissertation, the Spearman rank correlation is used because it does not assume any assumptions for linearity or the distribution. This method is reliable to examine the correlation between variables and at another point, it can avoid any multicollinearity problem.

The range for the measured correlation coefficients is described between -1 to +1. The values of -1 and +1 illustrate a strong relationship between paired datasets. The degree of relationship is defined in Table (3-1).

Table 3.1: Degree of correlation

Range	Degree of Correlation
0.00 - 0.19	Very-Weak
0.20 - 0.39	Weak
0.40 - 0.59	Moderate
0.60 - 0.79	Strong
0.80 - 1.00	Very-Strong

3.4.3 The Linear Relationship between Independent Variables and Log Odds

The linear relationship between independent variables and log odds of an event is another assumption that is important to check. It can be validated by drawing the plot in Python or R programming language to determine the linear relationship between the independent variable and the log odds. Also, it can be developed on SPSS software with Box-Tidwell estimation or Python software. In this dissertation, the Python software is applied. In this method, the point is that the continuous independent variable and their logit transformation are applied to meet the assumption. The linearity is investigated by GLM (general linear model) by applying the Hosmer-Lemeshow test (Agresti 2007).

3.4.3.1 Hosmer - Lemeshow Test

Hosmer – Lemeshow test is based on the estimation probabilities. In this test, the predicted probabilities are divided into 10 groups and after that, a Pearson Chi-square statistic is calculated that shows the comparison between the predicted to the real frequencies in a table of two rows and ten columns. The statistic is calculated with the following formula:

$$H = \frac{\sum_{g=1}^{10}(O_g - E_g)^2}{E_g} \tag{3-18}$$

Here, O_g and E_g define the observed events and predicted events. The degree of freedom is 8. The P-values less than 0.05 define the poor fit and P-values near the one indicate the good fit model (Park 2013). To perform this assumption it is necessary to determine which independent variables are continuous and which are categorical. Because for continuous variables, the natural logit should be calculated before performing the test. All the independent variables and the multiplication of each of the continuous independent variables with its logit are added to the Logistic regression model. After developing the model, the P-values are checked only for the multiplications of the continuous independent variables with its logit. If the significance (P-values) is less than 0.05 the independent variable can be removed and the final Logistic regression model develops without this variable because these variables violate the assumption.

3.5 Fitting the Logistic Regression Model

The Logistic regression model uses the equation $ln\left(\frac{\pi}{1-\pi}\right) = b + (w \times x)$ to calculate the intercept and coefficient of each independent variable. The method for calculating these parameters is different from the linear regression model. The Maximum Likelihood Estimation (MLE) function estimates the values for b and w, which maximizes the probability of the observed dataset. The likelihood function defines the probability of observing data as a function of unknown parameters (b, w).

Typically, the likelihood is the probability of predicting seen dependent variable values from the seen independent variable values and it takes the value from zero to one.

$$L = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)^{n-y_i}\right] \quad (3\text{-}19)$$

In the above equation, n is the sample size, y_i is the number of successes and π indicates the probability (Park 2013).

3.6 Evaluating the Classification Models

Typically the classification techniques are trained by a set of data and the objective is to find an optimal classifier to perform a model for predicting. Understanding the quality of the model and its performance needs the evaluation of the model. Therefore, selecting a suitable evaluation method is necessary. There are various evaluation methods to review the machine learning models. In this dissertation, the confusion matrix and ROC curve analysis are used to evaluate the performance of machine learning models. For the binary classification models confusion matrix is the best evaluation technique to find the best optimal solution (Hossin and Sulamian 2015).

3.6.1 Confusion Matrix

The confusion matrix expresses the number of correctly or incorrectly predicted for each class. In the confusion matrix, true positive and true negative respectively define the number of positive and negative that are classified correctly. In contrast, false positive and false negative respectively show the positive and negative outcomes that are classified incorrectly.

The accuracy matrix defines the quality of prediction results for the model by considering the number of true predictions over the total number of instances evaluated. The accuracy is often used as metrics evaluation for classification models in machine learning.

Table (3.2) demonstrates the confusion matrix.

Table 3.2: Confusion Matrix

Confusion Matrix		Predicted Label	
		Positive	Negative
Real Label	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Being easy to be calculated and well understood by users are the advantages for accuracy and confusion matrix.

The other measurements that can be calculated in the confusion matrix are as follow:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{3-20}$$

$$Sensitivity\ (True\ positive\ rate) = \frac{TP}{(FN+TP)} \tag{3-21}$$

$$Specificity\ (True\ negative\ rate) = \frac{TN}{(TN+FP)} \tag{3-22}$$

1) Accuracy: Typically, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
2) Sensitivity: It measures the ratio of true positive over the total number of true positive and false negative or in other words fraction of positive patterns that are correctly classified.
3) Specificity: It measures the ratio of true negative over the total number of true negative and false positive or in other words the fraction of negative patterns that are correctly classified (Hossin and Sulamian 2015).

In this dissertation, the confusion matrix is used to calculate the accuracy.

3.6.2 Alternative for Accuracy

Sensitivity and specificity measures are applied to evaluate the accuracy of the classification model used. The sensitivity illustrates how well the classifier can predict the positive values on the other hand; specificity illustrates how well the classifier can predict the negative values.

$$Sensitivity = \frac{true\ positive}{positive} \tag{3-23}$$

$$Specificity = \frac{true\ negative}{negative} \tag{3-24}$$

True positive and true negative respectively are the number of positive and negative predictions that are true. Positive and negative is the number of positive and negative samples (Biau et al. 2016). Therefore, the accuracy can be defined as below with considering the sensitivity and specificity:

$$Accuracy = Sensitivity \times \frac{positive}{(positive + negative)} + Specificity \times \frac{negative}{(positive + negative)} \tag{3-25}$$

3.6.3 ROC Curve

A receiver operating characteristic curve (ROC) is a plot that depicts the pair of sensitivity and specificity of the prediction with all possible cut-offs points between zero to one (Hastie et al. 2017). The area under this curve (AUC) determines the measuring of fitting the model. This curve presents the sensitivity on the vertical axis versus the horizontal axis, which shows the (1-specificity). In other words, the horizontal axis identifies the false positive rate and the vertical one identifies the true positive. When the cut-off points get the value near zero, all the prediction gets the value near one ($\hat{y}$=1) so the sensitivity gets the value near one and specificity gets the value near zero with considering the relation (1-specificity, sensitivity) so the point gets the value (1,1). On the contrary, when the cut-off point is near one so the sensitivity is near zero and specificity is

near one and the point is near (0, 0). Therefore the curve identifies the points above the diagonal dividing have much better results compared to the points under the dividing line. Therefore, for the points above the diagonal line model prediction works well and the classification is more correct. The point (0,1) is called the perfect classification point (Biau et al. 2016).

3.7 Overall Model Evaluation with Likelihood Ratio Test

The first step needs to assess the relationship between all the independent variables and the dependent variable. To reach this goal the model can be investigated in two different situations including without independent variables and with the independent variables. The first one considers to fit the model with no independent variables (null model) and fit the model with the only constant term so the coefficients or weights (w_i) of the model gets the value of zero because for the null hypothesis there are no independent variables.

$$H_0: w_1 = w_2 = w_3 = \cdots w_k = 0 \tag{3-26}$$

After that, the model is fitted with all independent variables and the constant. For each model, the log-likelihood is calculated. The likelihood ratio test is the difference between the natural logarithm of each log-likelihood multiplied by -2 to determine G. Multiplication of -2 is necessary to assess a distribution that is known for hypothesis investigating (Park 2013). χ^2 is a statistic with the degree of freedom k.

$$G = \chi^2 = -2\, ln\left(\frac{L.L.\ null}{L.L\ Alter}\right) \tag{3-27}$$

The likelihood ratio test is required to investigate the model. The P-value help to indicate the good fit to the data. If the P-value is greater than 0.05 the null hypothesis will be approved on the

contrary if it is less than 0.05 the null hypothesis will be rejected and it defines that the model will fit at least with one independent variable.

3.8 Significance of the Coefficients

After fitting the model and estimating the coefficients, the first essential subject is identifying the significance of variables in the model. Typically providing the formulation and testing of a statistical hypothesis is required to specify that the independent variables are significantly related to the dependent variables. There is a quite general method to test the hypothesis. The approach to test for significance of coefficients can be earned by comparing the observed values for dependent variables and predicted values by performing the determined model in two different conditions one with independent variables and the other without the independent variables. The mathematical function that is used to investigate the predicted values and dependent variables depends much on the problem. Totally if the predicted values with independent variables are more accurate and better than without independent variables so the coefficients are significant. In the Logistic regression model if the value of an independent variable (x_j) changes one unit with considering all other independent variables (predictors) remain constant so the log odds of the dependent variable change (w_j) units. For evaluation of the Logistic regression model, the Wald test and likelihood ratio test are most popular than others (Hosmer and Lemeshow 2013).

3.8.1 Wald statistic Test

Wald test is a method to define the significance of the individual coefficients in the Logistic regression model.

$$wald = \frac{{w_i}^2}{SE(w_i)^2} \quad (3\text{-}28)$$

In the equation (3-28) w_i is the coefficient of variable and SE is the standard error. The Wald test is a ratio of coefficients to the standard error of coefficient. In this test, if the result of the Wald test for a specific independent variable is equal to zero it defines that the variable is not significant and can remove from the model. In contrast, if the result of the Wald test is greater than zero, the variable is considerable and should keep it in the model.

3.8.2 Log-likelihood Ratio Test

In the Logistic regression, model log-likelihood compares the dependent variable values to the predicted values in conditions of with and without independent variables. This test is a measurement to indicate that the independent variable can be effective on the dependent variable (Park 2013). The test calculates as follow:

$$G = -2\ln\left(\frac{L_0}{L_1}\right) = -2[\ln(L_0) - \ln(L_1)] \quad (3\text{-}29)$$

3.9 Overfitting

One of the most significant issues in supervised learning is overfitting. Because of overfitting, the model cannot be generalized perfectly and it does not work as well as on unseen data on the test set. Overfitting will happen based on different reasons such as the limited size of the training set, presence of noise, and complex classifier. Overfitting causes poor performance on the test set. Therefore, it is important to solve this problem and avoid overfitting.

There are three reasons for overfitting, 1) the dataset has been used for the training set has noise and the training set includes not cleaned data, sometimes it will happen when the size of the dataset

is too small. When a dataset includes noise, noises have a chance to learn in the training set and can be used in prediction. 2) some models have complex classifiers that help the model to have overfitting. 3) the dataset has a high variance, at these kinds of models the performance for the training set is very well but for the test, set has an error. This model cannot be generalized due to the poor performance of unseen data.

There are various methods to reduce overfitting like regularization, cross-validation, ensemble models (Ying 2019).

3.9.1 Regularization

Logistic regression is a powerful tool to analyze the data and prediction. This model includes various regularizations such as l1 (Lasso Regression), l2 (Ridge Regression). The l1 and l2 are the methods known as the shrink method because this method shrinks the coefficient in the regression. When a dataset has a hundred variables the regularization changes to less variable by removing the idle variables. Reducing the variant of the model is preferable to avoid overfitting.

Regularization helps a model to know how many of the independent variables are practical and which of them are more significant than the others. The algorithm to find the intercept (b) and coefficients (w) need to minimize the squared prediction errors that it defines by the below formula:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} w_j x_j)^2 \quad (3\text{-}30)$$

Lasso regularization (l1) adds another term to this equation that defines the summation of magnitudes of all coefficients. This term plays a role as a penalty that the amount of this depends on the total magnitudes of all coefficients and there is another parameter named lambda that adjusts the size of the penalty. The following equation illustrates the lasso regularization:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} w_j x_j)^2 + \lambda \sum_{j=1}^{p} ||w_j|| \quad (3\text{-}31)$$

The Ridge regularization (l2) is another method that is like the Lasso method and it has the same formula but the difference just is on the term of penalty. Ridge uses the summation of coefficients squared to determine the penalty. The following equation defines the Ridge regularization formula:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_j w_j)^2 + \lambda \sum_{j=1}^{p} {w_j}^2 \quad (3\text{-}32)$$

Adding a new feature helps the model to decrease the first term by predicting accurately, although it causes to increase in the second part. Typically, there is a balancing to avoid increased variance of the model that can improve the model for unseen data (Qin and Lou 2019).

In this dissertation, the Ridge regularization (l2) is used to avoid overfitting.

3.9.2 Cross-Validation

Cross-validation is a significant method to avoid overfitting for training and testing set in modeling artificial intelligence. This method provides an opportunity to resample data to prevent overfitting. The easiest way is to build a model based on the data but it is unable to be generalized to a new model for unseen data. This method considers a model with n observation, (x_i, y_i), with i=1…n. The data is split up into two groups training set and test set. The training set is used to learn the model and a test set is used to evaluate the model by label values (Berrar 2018).

3.9.2.1 K-fold Cross-validation

K-fold Cross-validation is one of the ways to avoid overfitting. In k-fold cross-validation, the data set is split up into k subsets that the sizes are equal. The model is using the k-1 subset for the training set and one for the test set that is used for evaluating the model. This process will be repeated for k times. The average of k times performance training set on the k test set is the cross-

validation performance. In this method, all the n observations will use in the training and test set (Berrar 2018).

In this dissertation 5-fold, cross-validation is considered. This method divides the dataset into five folders and one of the folders is a test set and the remaining are train sets. In each step, one folder is considered as a test set, and the remaining four folders are the train set. Totally in this method the all dataset is used to train and test and this is the advantage of using the method. The accuracy result defines the average of five iteration results to understand better the standard deviation is calculated.

The below figures illustrate how the test set will change during each iteration.

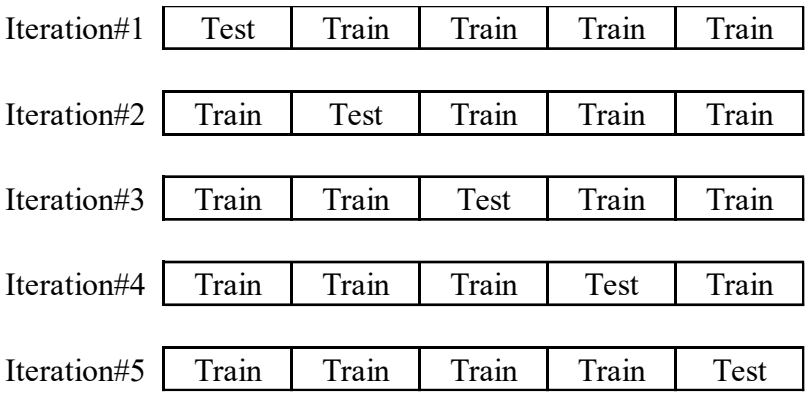

Figure 3.4: 5-Fold cross-validation

3.9.3 Ensembles Models

Ensembles methods assist the model to reduce the risk of overfitting. Typically this method is commonly used for gradient boosted and random forest models. The random forest model uses a bagging technic. Random forest model in machine learning that uses a combination decision tree. Therefore by applying this method for predicting, it aggregates the prediction of each decision tree

that is generated by bootstrapping and features. Bootstrapping chooses some data points from a sample with replacement. Therefore each decision tree uses a different dataset for the train set. It avoids that model focusing too much on a specific dataset or in other words features values (Ghojogh and Crowley 2019).

3.10 Summary

In this chapter, the detail of the Logistic regression model is presented. The multivariable Logistic regression is developed to predict. All the assumption to perform Logistic regression is investigated. It assessed all the tests for assumption and fitting the model. Typically to choose a classification model in machine learning investigating the dependent variable is important. The dependent variable has to be dichotomous or categorical and it is checked in this dissertation. Logistic regression is a practical algorithm that uses in prediction. All the assumptions before developing the model are checked. In this chapter, the Logistic regression model from the basic concept was introduced and analyzed step by step. Furthermore, the techniques that are used and the process of Logistic regression is defined. The maximum likelihood test and Hosmer - Lemeshow test were investigated. The validation of the model, overfitting, and some techniques to avoid overfitting is discussed.

CHAPTER 4 Developing the Logistic Regression Model

4.1 Introduction

The framework of this dissertation is based on the data collected from the conditions of a machine by sensors. It consists of five different excel files that each of which defines a situation. All the machine's components or equipment meet the failure so maintenance plays a vital role in the industry.

Failure history is one of the files that define the failure of a machine. Maintenance history is the other part of data, this part defines the maintenance of a specific machine by repair or a replacement of the component. Features of a machine are another important part, the size of engines, age, make, and model are significant. The data was collected every day for 24 hours in the year 2015. There are 876100 observations. Rotation, voltage, pressure, and vibration are the four important parameters recorded by the specific sensors, these four factors have been averaged hourly. There are 100 types of machines and the information was collected for all of them. For each kind of machine, five types of error codes can happen and four types of considerable components require repair or replacement. If a component for one of the machines is replaced, it will be recorded in collected data. Replacing the component will happen under two scenarios:

1) The technician will replace it during predictable maintenance when visiting the machine based on regularly scheduled maintenance.

2) A component breaks down therefore a technician will replace it based on unpredictable maintenance.

Based on collected information the health of a component or equipment can be predicted to avoid failure by replacing the faulty component.

Python software is selected to develop the model. Developing the model assist to determine in a specific situation which industrial machine will be failed or not.

4.2 Dataset Preparation

To prepare the dataset it was needed to combine different types of information provided in separated excel files. Before modeling, preparation of the dataset is one of the most important necessities. Data preparation is necessary to feed the pure dataset to model and get accurate results. Performing the process is not completely automated and it needs various strategies to do. In this dissertation preparing the dataset because of the combination of different excel files was a bit time-consuming. It requires determining the categorical and numeric variables. One of the important steps is related to the duplicates in the dataset, for some hours of a day the information provided was repeated and it needs to remove from the dataset. It needs to check the dates and machine-ID and after that with considering the same dates remove some rows that were repeated. The data has 876100 rows and 19 columns.

The information of collected data is divided into some groups that are presented in Table (4.1).

Table 4. 1: Groups of the dataset

Category of variables	Variables	Groups	Description
Date and Time of measurement	DateTime	Date format and time format	Time of collectin data of 100 machines
Machine specifications	Machine-ID	Categorical	Number of machines that we collected data based on it
	Age	Numeric	Number of years that machine is working
	Model	Categorical	Type of machine
Measurement Parameters	Voltage	Numeric	The voltage of a machine that measured in a specific time
	Rotation	Numeric	Rotation of a machine that measured in a specific time
	Pressure	Numeric	The pressure of the machine that measured in a specific time
	Vibration	Numeric	The vibration of machine that measured in a specific time
Errors	Error 1	Categorical	Types of errors that will happen for each machine in a specific time
	Error 2		
	Error 3		
	Error 4		
	Error 5		
Components	Component 1	Categorical	Type of component will break down in a specific time
	Component 2		
	Component 3		
	Component 4		

The first step that was the most important part of the preparation dataset is missing values. This part is significant because in coding the nan values can not be read. For instance, In the dataset that is collected, some of the rows had missing values of errors or components failure. In order to solve this issue, the zero value was replaced by those missing values.

4.2.1 Date and Time of Measurement

The date and time in the dataset represent when the different variables are collected. The variable is collected hourly during a year. Considering the method, date and time, and also the machine ID does not have any effect on the prediction. Therefore, these variables will be ignored. To develop the Logistic regression model some assumptions are needed to check.

4.2.2 Calculating the Size

Firstly the sample size will check in the Logistic regression method. The size of the sample is one of the significant subjects. So with considering the formula, it is provided in Chapter Three, the size of the sample is calculated. Equation (4.2) presents the size of the data.

$$n = \frac{10 \times k}{p} \tag{4-1}$$

$$n = \frac{10 * 15}{720/876100} = 182521 \tag{4-2}$$

For the dataset with fifteen features, the minimum number of observations which is required is about 182521. The dataset that is utilized has 876100 observations and it can be good to apply the Logistic regression model. The dependent variable has a binary value and it is another important subject.

4.2.3 Statistical Description

The goal of statistical description is to present a summary of the variables in the dataset and their characteristics. Typically statistical description explains a quantitive analysis for the numeric variables in the dataset. The quantitive analysis for the dataset in this dissertation is presented in Table (4.2).

Table 4. 2: Descriptive Statistics

	Mean	Max	Min	Std	Skew
Volt	170.78	255.12	97.33	15.51	0.09
Rotate	446.61	695.02	138.43	52.67	-0.14
Pressure	100.86	185.95	51.24	11.05	0.40
Vibration	40.39	76.79	14.88	5.37	0.25
Age	11.33	20.00	0.00	5.83	-0.25

Other features are used for modeling such as four types of components and five types of errors are categorical variables and it is not necessary for statistical description. These variable gets the values of zero or one to define the situation. When the variables get zero it means that the component works and no such error will occur, on the contrary when the variables get one it defines that the component does not work or the error happened.

4.2.4 Outliers

Finding outliers and removing these outliers is another important part. Figure (4.1) defines the plot of outliers regard to the continuous variables.

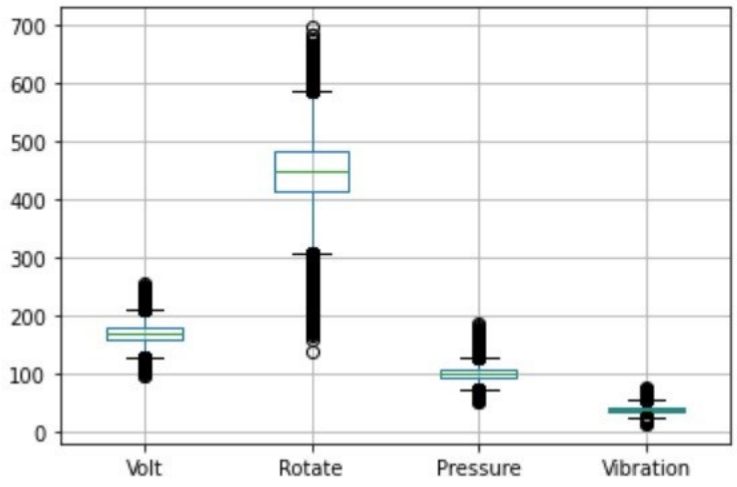

Figure 4. 1: Outliers

4.3 Parameters

There are four parameters consisting of the hourly average of voltage, rotation, pressure, and vibration which are collected from 100 industrial machines. These parameters are continuous and provided by applying sensors. For each machine parameters: voltage, pressure, vibration, and rotation are measured. Figure (4.2) presents these parameters distributed normally. Based on the assumptions for the Logistic regression model, the normal distribution is not required.

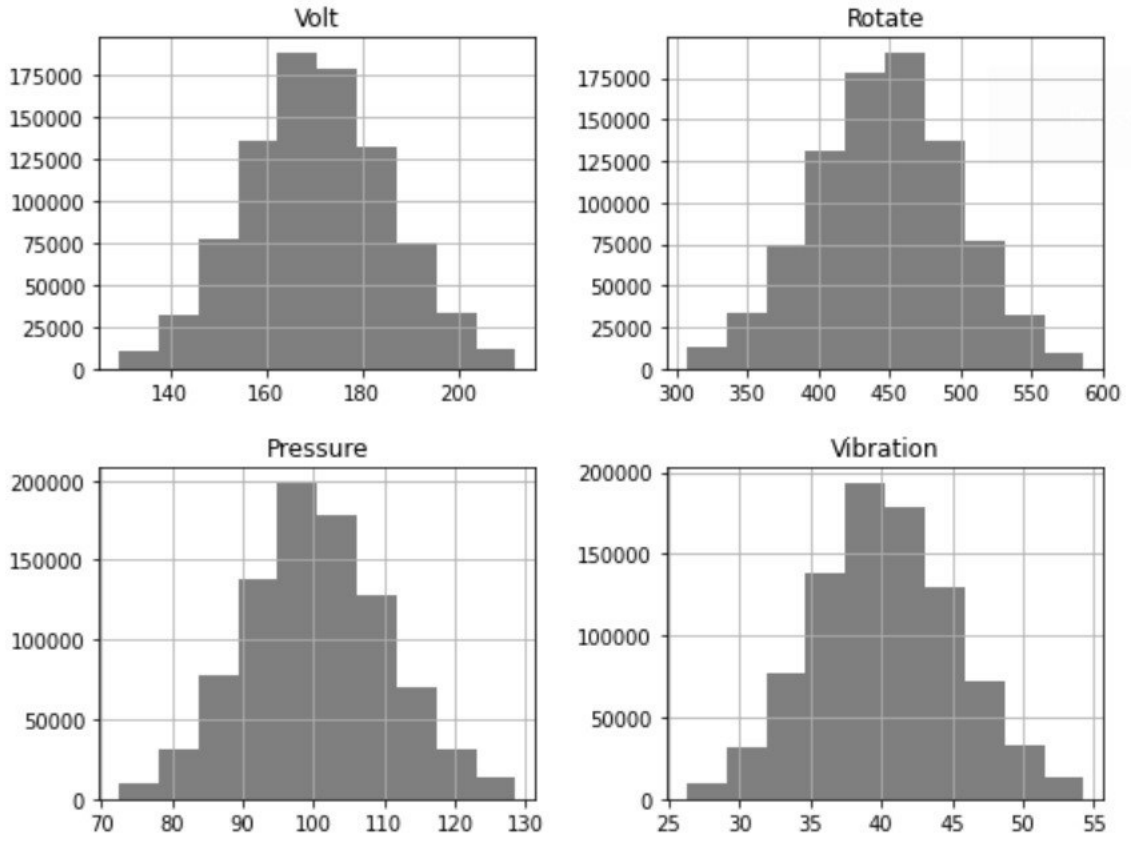

Figure 4. 2: Continuous variables distribution

4.4 Age of Machines

The specification of each machine is different. There are 100 machines and for each of them, there are four types of models. Furthermore, each machine has an age that defines the number of years the machine is working and ranges of them are between zero (the machine is new) to 20. Figure (4.3) determines a large number of machines are at age fourteen.

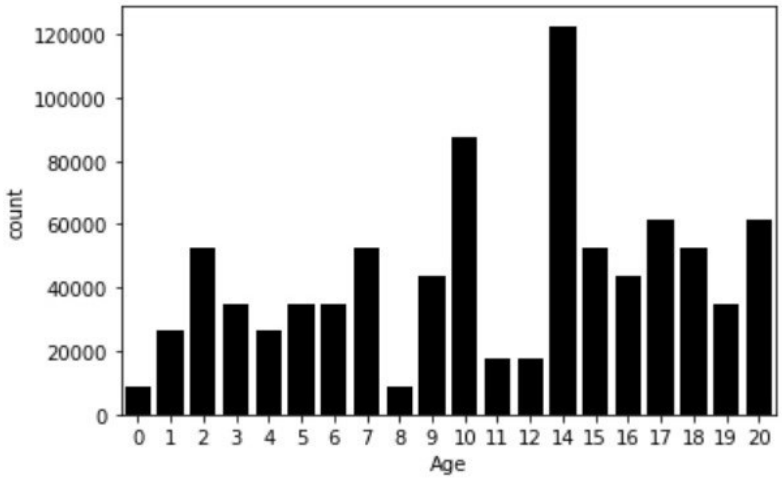

Figure 4. 3: Distribution of the age of machines

4.5 Machines specification

Typically there are four different types of machines. Figure (4.4) defines how many machines are there for each type. Machine model three has the highest number of machines and machine model one has the lowest.

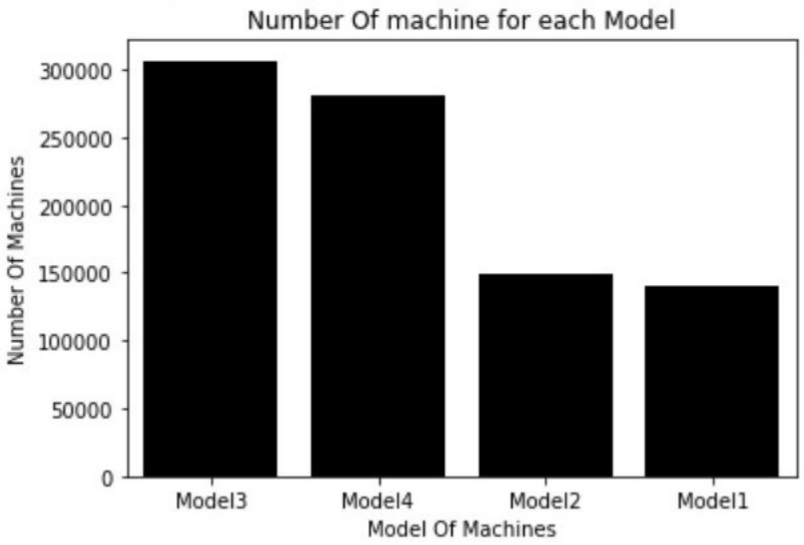

Figure 4. 4: The model of machines

4.6 Components

Four different types of components cause the failure of a machine. This information is collected when a component of a machine is replaced. Figure (4.5) presents how many times each component will replace.

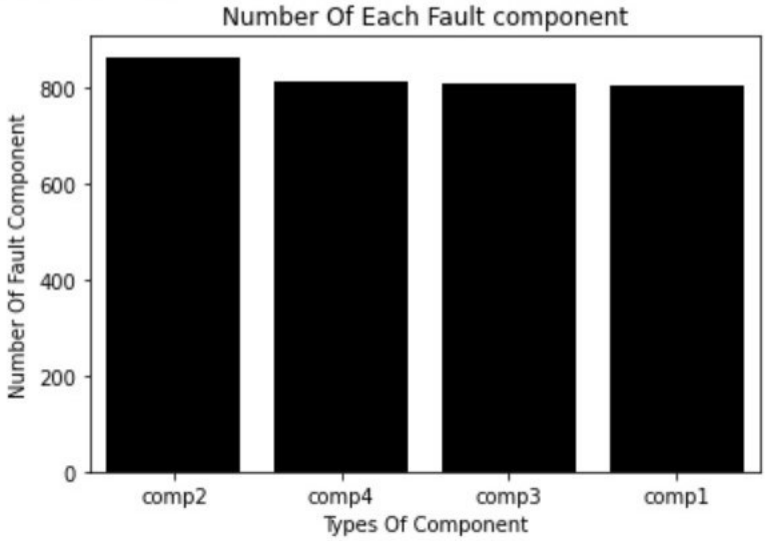

Figure 4. 5: The failure's component

4.7 Errors

For each type of machine, there are five different types of errors that the machine faces. Each of them causes the failure of a component and consequently causes the failure of a machine. The machines have faced these errors during operations. These errors do not shut down the machines and consequently the production line. They are not assumed as a failure and these are collected hourly.

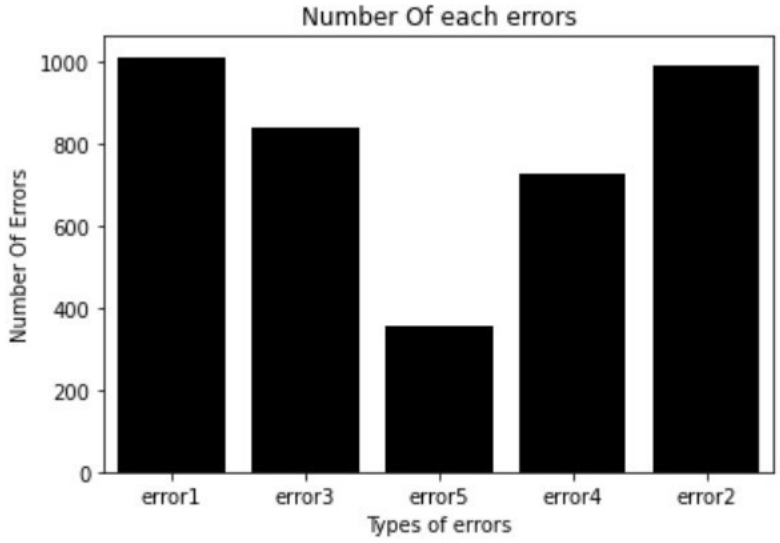

Figure 4. 6: The number of each type of errors

4.8 Dependant Parameter (Target)

The dependent or target variable in this dataset is the variable named Failed. It defines the situation of a machine that may fail or works and it depends on the independent variable or features. Table (4.3) shows the number of situations the machine may be failed and the number of situations that it works correctly.

Table 4. 3: The number of each machine's class

Situation of Machines	Group	Number of Situations
Not Failure	No	875379
Failure	Yes	721

Considering the information it can be understood the dataset is not balanced and the number of events that determines the failure for an industrial machine is less than not failure situations.

4.9 Developing the Model

The model will develop in some steps to reach the improved model.

4.9.1 Step 0

Results of developing the model with raw data that are imbalanced are presented in Table (4.4).

Table 4. 4: Confusion matrix with raw data

Confusion Matrix		Predicted Label		Percent correct predicted
		No	Yes	
Real Label	No	175020	60	99.96%
	Yes	85	55	39.29%
Overall Percentage				69.63%

Where:

No: Not Failed

Yes: Failed

$$Accuracy = \frac{175020+55}{175020+55+85+60} * 100 = 99.86\% \quad (4\text{-}3)$$

$$Sensitivity = \frac{175020}{175020+85} * 100 = 99.95\% \quad (4\text{-}4)$$

$$Specificity = \frac{55}{55+60} * 100 = 47.82\% \quad (4\text{-}5)$$

$$logit(y) = ln\left[\frac{p(y=1|x_1x_2 \dots x_n)}{1-p(y=1|x_1x_2 \dots x_n)}\right] = b + w_1 \times x_{vol} + w_2 \times x_{Rot} + w_3 \times x_{Pre} + w_4 \times x_{Vib} + w_5 \times x_{Age} + w_6 \times x_{E1} + w_7 \times x_{E2} + w_8 \times x_{E3} + w_9 \times x_{E4} + w_{10} \times x_{E5} + w_{11} \times x_{C1} + w_{12} \times x_{C2} + w_{13} \times x_{C3} + w_{14} \times x_{C4} + w_{15} \times x_{D1} + + w_{16} \times x_{D2} + w_{17} \times x_{D3} + w_{18} \times x_{D4} = -14.111 + (0.019 \times x_{vol}) + (-0.011 \times x_{Rot}) + (0.022 \times x_{Pre}) + (0.105 \times x_{Vib}) + (0.056 \times x_{Age}) + (-0.135 \times x_{E1}) + (-0.235 \times x_{E2}) + (0.211 \times x_{E3}) + (-0.343 \times x_{E4}) + (-0.133 \times x_{E5}) + (4.884 \times x_{C1}) + (5.533 \times x_{C2}) + (4.264 \times x_{C3}) + (4.74 \times x_{C4}) + (0.577 \times x_{D1}) + (0.475 \times x_{D2}) + (-0.577 \times x_{D3}) + (-0.58 \times x_{D4}) \quad (4\text{-}6)$$

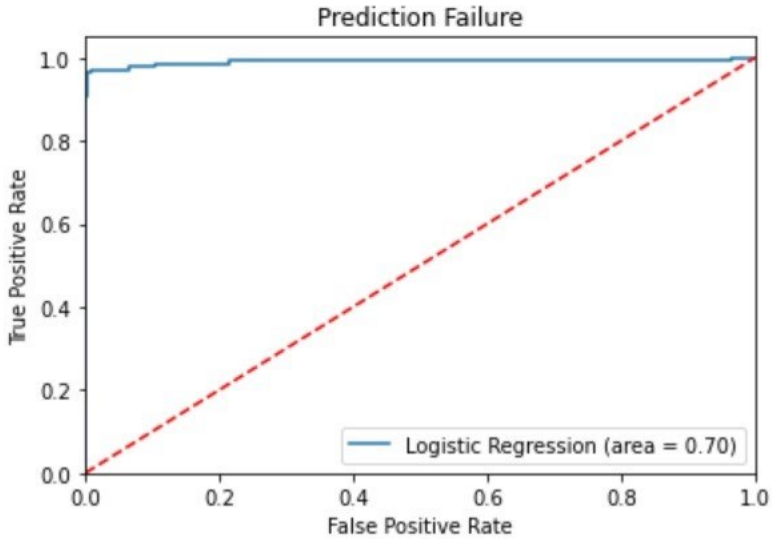

Figure 4. 7: ROC Curve with raw data

K-fold

Accuracy: 99.92 %

Standard Deviation: 0.00 %

4.9.2 Balancing the Data

In this step balancing the dataset is required to have high accuracy. Balancing the dataset was discussed in Chapter Three. The Adasyn method is one of the best oversampling methods which is utilized in this dissertation. Failure machines compared with not failure is too less and it assists the model to desire the not failure and it works as a bias model. Therefore balancing the dataset is necessary.

4.10 Checking Correlation coefficients

Based on the assumptions for the Logistic regression model, it does not need to check the distribution which is normal or not. The other one that does not require checking is about the observation, which does not need to have a relationship between them. The correlation among independent variables is needed to check. For checking this assumption the Spearman method is performed. Based on this method the calculation for the Spearman correlation coefficients is done and the results are presented in Table (4.5).

Table 4. 5: Correlation coefficient of the Spearman method

Var	ID	Vol	Rot	Pre	Vib	Age	E1	E2	E3	E4	E5	C1	C2	C3	C4	Mdl
ID	1	-0.0007	0.0003	0.0047	-0.0009	0.1056	0.0009	-0.0005	0.0009	0.0010	0.0003	0.0007	-0.0010	0.0007	-0.0017	0.0154
Vol		1	-0.0013	0.0021	0.0024	-0.0004	0.0070	0.0011	0.0018	0.0008	0.0011	0.0078	-0.0009	0.0011	0.0016	0.0004
Rot			1	0.0001	-0.0028	-0.0012	-0.0014	-0.0094	-0.0128	-0.0010	-0.0032	0.0013	-0.0110	0.0003	-0.0026	0.0010
Pre				1	0.0013	0.0032	0.0008	-0.0005	0.0001	0.0073	-0.0004	0.0014	0.0014	0.0072	0.0015	-0.0146
Vib					1	0.0136	0.0001	-0.0014	-0.0001	-0.0002	0.0140	0.0029	0.0007	-0.0002	0.0097	-0.0044
Age						1	-0.0036	0.0005	-0.0006	0.0007	0.0087	0.0008	0.0006	0.0001	-0.0007	-0.2045
E1							1	0.0109	0.0120	0.0037	0.0110	0.0014	-0.0010	-0.0010	0.0014	0.0005
E2								1	0.2785	0.0037	0.0213	0.0003	-0.0010	-0.0010	0.0002	0.0007
E3									1	0.0055	0.0250	-0.0009	0.0003	-0.0009	-0.0009	0.0002
E4										1	0.0073	0.0006	-0.0009	0.0006	-0.0008	-0.0059
E5											1	0.0014	0.0013	-0.0006	-0.0006	-0.0027
C1												1	0.1663	0.1595	0.1560	0.0009
C2													1	0.1712	0.1854	-0.0007
C3														1	0.1540	0.0005
C4															1	-0.0008
Mdl																1

The result of the Spearman correlation coefficient defines that there is no strong correlation between independent variables therefore there is no need to remove any of the variables because of avoiding consequences such as multicollinearity or overfitting.

4.10.1 Step 1

In this step, the linear relationship between independent variables and log odds will check. This assumption unfortunately is not checked in some cases but this dissertation performs to improve the model. To check the linearity, GLM (general linear model) model will use to investigate. Therefore from five continuous variables, four of them need to remove. Volt, Rotation, Vibration, and Pressure are the independent variables that would remove from the dataset. So with a new variable, the Logistic regression model is developed and Table (4.6) presents the prediction and accuracy.

Table 4. 6: Confusion matrix with checking linearity

Confusion Matrix		Predicted Label		Percent correct predicted
		No	Yes	
Real Label	No	174766	314	99.82
	Yes	5	135	96.43
Overall Percentage				98.13

$$Accuracy = \frac{(174766+135)}{(174766+135+5+314)} = 99.86\% \quad (4\text{-}11)$$

$$Sensitivity = \frac{174766}{174766+5} = 99.99\% \quad (4\text{-}12)$$

$$Specificity = \frac{135}{135+314} = 30.07\% \quad (4\text{-}13)$$

$$logit(y) = ln\left[\frac{p(y=1|x_1x_2 \dots x_n)}{1-p(y=1|x_1x_2 \dots x_n)}\right] = b + \sum_{i=1}^{i=n} w_i \times x_i = b + \left(w_1 \times x_{Age}\right) + (w_2 \times x_{E1}) +$$

$$(w_3 \times x_{E2}) + (w_4 \times x_{E4}) + (w_5 \times x_{E5}) + (w_6 \times x_{C1}) + (w_7 \times x_{C2}) + (w_8 \times x_{C3}) +$$

$$(w_9 \times x_{C4}) + (w_{10} \times x_{D1}) + (w_{11} \times x_{D2}) + (w_{12} \times x_{D3}) + (w_{13\times} x_{D4}) = -1.311 +$$

$$\left(0.011 \times x_{Age}\right) + (-4.002 \times x_{E1}) + (-3.879 \times x_{E2}) + (-3.232 \times x_{E3}) +$$

$$(-3.502 \times x_{E4}) + (-1.701 \times x_{E5}) + (8.479 \times x_{C1}) + (8.927 \times x_{C2}) + (7.084 \times$$

$$x_{C3}) + (8.329 \times x_{C4}) + (-0.326 \times x_{D1}) + (-0.384 \times x_{D2}) + (1.237 \times x_{D3}) +$$

$$(-2.135 \times x_{D4}) \qquad (4\text{-}14)$$

K-fold:

Accuracy: 99.91 %

Standard Deviation: 0.00 %

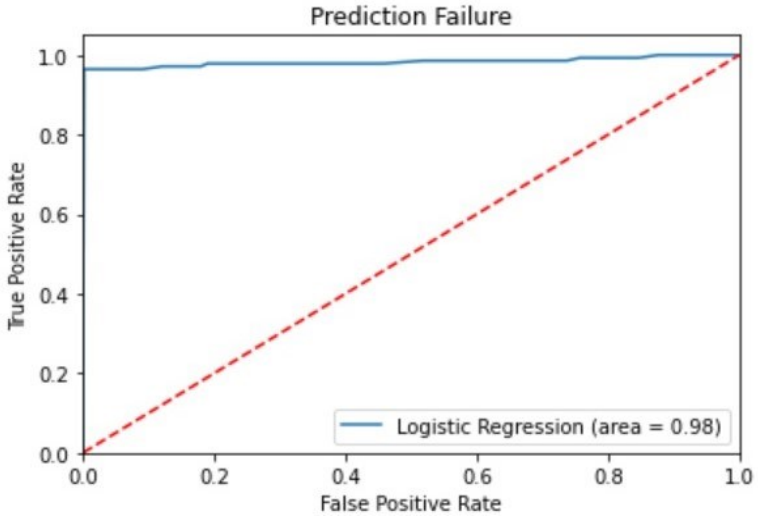

Figure 4. 8: ROC Curve with checking linearity

4.10.2 Step 2

In this step after checking the assumption the logit regression model will develop with the library of stats model, Table (4.7) presents the result of developing the model:

Table 4. 7: Result of logit model

Variable	Coef	P-Values
constant	-3.7504	0.807
X_{Age}	0.0107	0
X_{E1}	-10.9281	0.098
X_{E2}	-22.1812	0.992
X_{E3}	-5.786	0.002
X_{E4}	-23.5938	0.901
X_{E5}	-14.2119	0.946
X_{C1}	8.4712	0
X_{C2}	8.935	0
X_{C3}	7.0969	0
X_{C4}	8.3593	0
X_{Mdl1}	-10.3874	0.772
X_{Mdl2}	-10.4429	0.771
X_{Mdl3}	-11.2892	0.753
X_{Mdl4}	-12.1846	0.734

Considering the result of performing the Logistic regression model some independent variables can remove because the P-values are greater than 0.05 and they do not reject the null hypothesis. Independent variables $x_{E1}, x_{E2}, x_{E4}, x_{E5}, x_{Mdl1}, x_{Mdl2}, x_{Mdl3}, x_{Mdl4}$ will remove from the dataset and the model can be developed with new predictors. The independent variables from x_{E1} to x_{E5} except for x_{E3} are the four types of errors that will happen for industrial machines. Furthermore, x_{Mdl1} to x_{Mdl4} are the dummy variables for the Model-independent variable. Table (4.8) presents the result of the proposed model with the new dataset.

Table 4. 8: Confusion matrix of improved model

Confusion Matrix		Predicted Label		Percent correct predicted
		No	Yes	
Real Label	No	174766	314	99.82
	Yes	5	135	96.43
Overall Percentage				98.13

$$Accuracy = \frac{(174766+135)}{(174766+135+5+314)} \times 100 = 99.86\% \quad (4\text{-}15)$$

$$Sensitivity = \frac{174766}{(174766+5)} \times 100 = 99.99\% \quad (4\text{-}16)$$

$$Specificity = \frac{135}{(135+314)} \times 100 = 30.07\% \quad (4\text{-}17)$$

$$logit(y) = ln\left[\frac{p(y=1|x_1x_2 \dots x_n)}{1-p(y=1|x_1x_2 \dots x_n)}\right] = b + \sum_{i=1}^{i=n} w_i \times x_i = b + \left(w_1 \times x_{Age}\right) + (w_2 \times x_{E3}) + (w_3 \times x_{C1}) + (w_4 \times x_{C2}) + (w_5 \times x_{C3}) + (w_6 \times x_{C4}) = -3.588 + \left(0.017 \times x_{Age}\right) + (-5.727 \times x_{E3}) + (8.502 \times x_{C1}) + (9.116 \times x_{C2}) + (7.072 \times x_{C3}) + (8.288 \times x_{C4}) \quad (4\text{-}18)$$

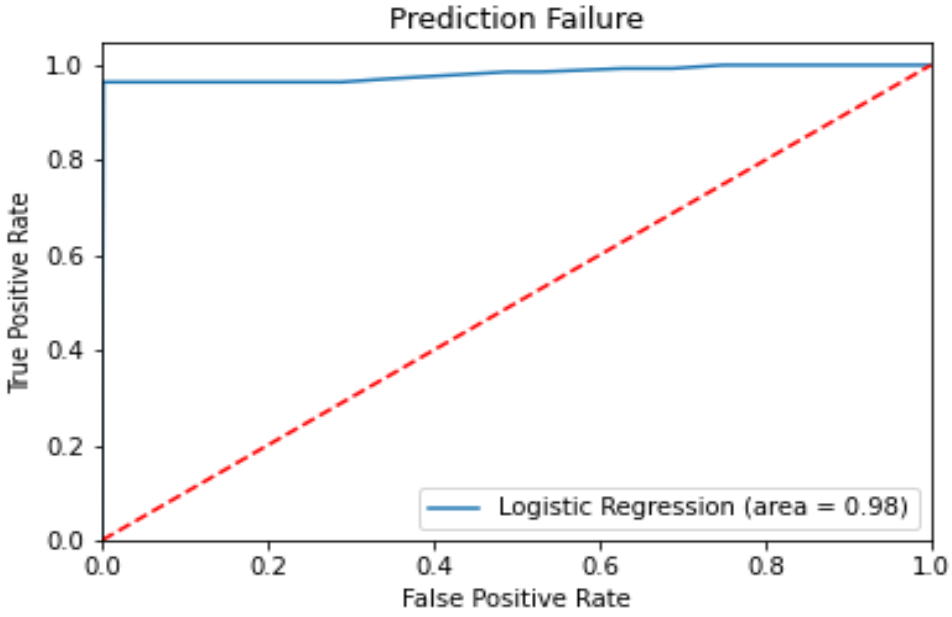

Figure 4. 9: ROC Curve of the improved model

K-fold

Accuracy: 99.91 %

Standard Deviation: 0.00 %

4.11 Summary

In this Chapter, the first step was data preparation. Preparation data is done in some steps such as removing duplicates, filling missing values, finding and removing outliers. The dataset was divided into some groups which the type of them was explained. Some of the independent variables were categorical and some of them were continuous. Dataset had categorical data and it needs encoding, therefore for the Model variable, there are four dummy variables. The plots of all the independent variables are depicted to get some details about the variables. The statistics description has been presented. The model developed with raw data and the result defines that model will not work accurately for predicting the situation that the machine will fail (developing at step0). The plot based on the target value defines that balancing for the dataset is required. The data was balanced with oversampling Adasyn method and the result defines the predicting as more accurate. After that, the first required assumption related to multicollinearity was checked. For developing step one the other assumption related to the linear relationship between independent variables and log odds was utilized and after performing this assumption some variables were removed from the dataset because they violate this assumption. Therefore the model developed with the new data which the accuracy did not change. After that, the P-values were defined, and considering the method, some variables could be removed and only six variables remained. Performing the model with six independent variables defines that the accuracy is still high although the calculation decreased and it had more advantages. Therefore there is no need to collect huge

massive values as the dataset. Decreasing the number of independent variables is more affordable for enterprises to collect the dataset.

CHAPTER 5 Conclusion and Future Research

In Chapter Four, the model is developed and the result of the confusion matrix is provided by the performance of the Logistic regression model. Considering the result of the confusion matrix, it is understood the model works well and is accurate. About 80% of the dataset is trained and the remaining 20% is used for the test set and validate the model.

In a manufacturing system, a breakdown is so expensive and it forces heavy costs for the enterprises. Some issues related to the failure of a machine define failure of a machine leads the idle time for workers during working hours. Furthermore, an industrial machine failure affects other parts and stops other lines that have a dependency on each other.

5.1 Conclusion

Predictive maintenance strategies decrease the costs. Some surveys defined that the cost for a breakdown is more than the maintenance cost. By applying the prediction machine learning algorithm the downtime will decrease. In this dissertation, the failure of components is considered and it assists to be aware of the failure of a machine. Developing a machine learning algorithm is more affordable compared with failing a machine and the cost of a breakdown. As defined earlier in Chapter Four the model developed in some steps. The result of the confusion matrix with the raw dataset (without balancing) defines the situations when a machine fails does not work accurately. With balancing the dataset and performing the model the percentage of prediction for machine failure improved by about 57% (it increase from 39.29% to 96.43%). Moreover, checking the assumption improved the model by decreasing the calculations. Firstly by checking the linearity log odds, the number of variables decreases so the calculation decrease about 28.57%.

Furthermore, it defines that the model will work with fewer variables and it is very useful for the manufacturers. Secondly, after improving the model finally the independent variables will decrease and the calculation decrease up to 40%. In industrial machines, the efficiency of modeling is much dependent on accuracy to predict the failure situation. Decreasing the number of independent variables and consequently decreasing the calculation is more affordable for the companies. Collecting the needed information is hard and some sensors are expensive. It has a benefit for enterprises to decrease the independent variables and present a simple model with the minimum number of variables and high accuracy.

In this dissertation, the model that is developed can be generalized and developed for other real datasets. This model provided all the necessary assumptions to check. It can be developed on different datasets and get acceptable results. The Logistic regression model was developed to predict the failure of a machine to avoid the risk of failure with an accuracy of 99.86% that is excellent. Furthermore, the area under the ROC curve was 0.98 that depicts the high reliability of the developed model. Also, the k-fold validation and regularization are the strategies that were applied to avoid overfitting.

5.2 Limitation of this Research

There is some limitation to developing the model. Collecting the effective parameters and applying the sensors is a significant subject. Furthermore, as discussed earlier the imbalanced dataset has negative effects on the prediction. Unbalanced data may have high accuracy, although it does not work accurately on predicting the situation which belongs to the minority dataset.

5.3 Recommendation for Future Works

Additional research is done to assist in developing the prediction model. Machine learning has various algorithms divided into three different groups. In this dissertation, supervised learning is utilized which divides into two groups of regression and classification. Classification has some algorithms for prediction and for this thesis the Logistic regression model is chosen. It is suggested that the other classification method be checked to reach a good prediction and consequently the high accuracy. Other areas of research for future works can consider as follow:

- The Logistic regression model was developed to predict the failure of a machine to avoid the risk of failure. Considering the other predictors like the speed of production per second or other behavior of a machine can be useful. This model is practical for other types of independent variables that will add.
- The model can be developed by adding some parameters for the independent variables. In this dissertation, the provided information is about the situation of components that are faulty or not. To avoid the risk of failure it might be a professional model to collect the data about components. Therefore, the model decreases the risk of failure.
- The model can be developed by other machine learning algorithms, Neural networks, and Deep learning. The target values can be categorical instead of taking the binary values. It can divide into some groups to investigate the situation and avoid failure. For example in this thesis target value is binary and it has two situations of failure or not. Although it can be developed by categorical dependent variable and regrading situation like low, medium, and high or without risk of failure. The XGboost model algorithm is practical for such categorical variables.

- The results of this model can be regarded for the maintenance of machines. In addition, results can assist in the cost analysis for the enterprises to compare the cost of maintenance and breakdowns.
- The model can be improved by collecting the data including the lifetime of components and equipment. The regression model can be developed for the lifetime prediction of a machine or components.

References

1. Ahmad, Z., Rai, A., Maliuk, A.S., & Kim, J.M. (2020). Discriminant feature extraction for centrifugal pump fault diagnosis, 8, 165512-185528.
2. Asuero, A.G., Sayago, A., Gonzalez, A.G. (2006). The correlation coefficient: An overview, 36, 41-59.
3. Agresti, A. (2007). An Introduction to Categorical Data Analysis, 3rd Edition, Wiley, Hoboken, New Jersey.
4. Abdelkrim, C., Saleh Meridjet, M., Boutasseta, N., & Boulanouar, L. (2019). Detection and classification of bearing faults in industrial geared motors using temporal features and adaptive neuro-fuzzy inference system. Journal Heliyon. DOI: 10.1016/j.heliyon.2019.e02046.
5. Berrar, D. (2018). Cross-validation. Encyclopedia of bioinformatics and computational Biology, 1, 542–545.
6. Borgi, T., Hidri, A., Neef, B., & Naceur, M.S. (2017). Data analytics for predictive maintenance of industrial robots. International Conference on Advanced Systems and Electric Technologies (IC_ASET), 412–417.
7. Biau, G., & Scornet, E. (2016). A random forest guided tour. An Official Journal of the Spanish Society of Statistics and Operations Research, 25, 197–227.
8. Coro, A., Abasolo, M., Aguirrebeitia, J., Lopez de Lacalle, L. (2019). Inspection scheduling based on reliability updating of gas turbine welded structures. Advances in Mechanical Engineering, 11, 1–20.
9. Calabrese, M., et al. (2020). Sophia: An event-based IoT and machine learning architecture for predictive maintenance in industry 4.0. Information, 11. DOI:10.3390/info11040202.

10. Dai, X., & Gao, Z (2013). From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis, 9, 2226–2238.
11. Deepika, J., Senthil, T., Rajan, C., & Surendar, A. (2018). Machine learning algorithms: a background artifact. International Journal of Engineering & Technology, 7, 143-149.
12. Fernandes, S. et al. (2020). Forecasting Appliances Failures: A Machine-Learning Approach to Predictive Maintenance. Information, 11, 208. DOI:10.3390/info11040208.
13. Ghojogh, B., Crowley, M. (2019). The Theory Behind Overfitting, Cross-Validation, Regularization, Bagging, and Boosting: Tutorial.
14. Galeano, P., Pena, D. (2019). Data Science, Big Data and Statistics. DOI: 10.1007/s11749-019-00651-9. https://www.researchgate.net/publication/332276688.
15. Hassan, S., Tahir, M.M., Badshah, S., Hussain, A., & Anjum, N.A. (2018). Classification of rigid rotor faults using time domain features extracted from multiple vibration sensors. 23, 43-52.
16. Hossin, M., Sulamian, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP), 5, 2. DOI: 10.5121/ijdkp.2015.5201.
17. Hosmer, D., W., Lemeshow, S., (2013). Applied Logistic Regression. Second Edition, Wiley.
18. Hastie, T., Tibshirani, R., Friedman, J. (2017). The elements of statistical learning: Data mining, Inference, and Prediction. Springer, London.
19. Joel, L., The Handbook of Maintenance Management (2009). Second Edition, Industrial Press.

20. Jimenez-Cortadi, A., Irigoien, I., Boto, F., Sierra, B., & Rodriguez, G. (2020). Predictive maintenance on the machining process and machine tool. Applied sciences 10, 224. DOI: 10.3390/app10010224.
21. Jayaswal, P., Wadhwani, A.K., & B.Mulchandani, K. (2008). Machine fault signature analysis. Hindawi Publishing Corporation International Journal of Rotating Machinery, Article ID 583982. DOI:10.1155/2008/583982.
22. Jurafsky, D., & H. Martin, J. (2020). Speech and Language Processing. Copyright © 2020. All rights reserved. Draft of December 30, 2020.
23. Kanawaday, A., Sane, A. (2017). Machine learning for predictive maintenance of industrial machines using IoT sensor data. 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), 87-90.
24. Kaparthi, S., & Bumblauskas, D. (2020). Reliability paper designing predictive maintenance systems using decision tree-based machine learning techniques. International Journal of Quality & Reliability Management, 37, 659-686.
25. Katz, M. H. (1999). Multivariable analysis: A practical guide for clinicians. 3rd Edition, Cambridge University Press.
26. Kou, Y., Cui, G., Fan, J., Chen, X., Li, W. (2017). Machine learning-based models for fault detection in automatic meter reading systems. International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 684-689. DOI: 10.1109/SPAC.2017.8304362.
27. Lu Murphey, Y., & Abul Masrur, M. (2006). Model-Based Fault Diagnosis in Electric Drives Using Machine Learning. IEEE/ASME TRANSACTIONS ON MECHATRONICS, 11. Authorized licensed use limited to Concordia University Library.

28. Lemaitre, G., Nogueira, F., K. Aridas, C. (2016). Imbalanced-learn: A Python Toolbox to tackle the Curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 7, 1-5.
29. Masani, K. I., Oza, P., & Agrawal, S. (2019). Predictive maintenance and monitoring of industrial machine using machine learning. Scalable Computing: Practice and Experience 20, 663–667.
30. Mehmeti, Xh., Mehmeti, B., & Sejdiu, Rr. (2018). The equipment maintenance management in manufacturing enterprises. IFAC PapersOnLine 51-30, 800–802.
31. Mobley, R.K (2002). An Introduction to Predictive Maintenance, Second Edition, Butterworth-Heinemann is an imprint of Elsevier Science.
32. Murat Cinar, Z., et al. (2020). Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. Journal of Sustainability, 12, 8211. DOI:10.3390/su12198211.
33. Paolanti, M., Romeo, L., Felicetti, A., Mancini, A., Frontoni, E., & Loncarski, J. (2018). Machine learning approach for predictive maintenance in industry 4.0. 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA), Oulu, Finland, 1-6.
34. Park, H. A. (2013). An introduction to Logistic regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. Journal of Korean Academy of Nursing, 43, 154-164.
35. Pyle, D. (2007). Data Preparation for Data Mining. Morgan Kaufmann Publishers Inc., San Francisco, CA.

36. Qin, J., Lou, Y. (2019). L1−2 Regularized Logistic regression. Conference Paper, 779-783. DOI: 10.1109/IEEECONF44664.2019.9048830.

37. Saeed, U., Lee , Y.D., Ullah Jan, S., & Koo, I. (2021). Context-Aware fault diagnostic scheme towards sensor faults utilizing machine learning sensors, 21, 617. DOI: 10.3390/s21020617.

38. Shen, Y., & khorasani, Kh. (2020). Hybrid multi-mode machine learning-based fault diagnosis strategies with application to aircraft gas turbine engines, 130, 126-142.

39. S.Abdallah, Z., Du, L., I.Webb, G. (2017). Data preparation, C Sammut and G I Webb (Eds) Encyclopedia of Machine Learning and Data Mining. DOI: 10.1007/978-1-4899-7687-1.

40. Schreiber-Gregory, D., M-Jackson, H. (2018). Logistic and linear regression assumption: Violation Recognition and Control. Paper 130-2018.

41. Seebo (2019). Why predictive maintenance is driving industry 4.0: The Definitive Guide, 1–13. Available online: https://files.solidworks.com/partners/pdfs/why-predictive-maintenance-is-driving-industry-4.0405.

42. Sipos, R., Fradkin, D., Moerchen, F., & Wang, Z. (2014). Log-based predictive maintenance. 20th ACM SIGKDD International Conference on Knowledge and discovery and data, 1867–1876. DOI: 10.1145/2623330.2623340.

43. Traini, E., Bruno, G., D'Antonio, G., & Lombardi, F. (2019). Machine learning framework for predictive maintenance in milling. IFAC (International Federation of Automatic Control) PapersOnLine 52-13, 177–182.

44. Togami, M., Abe, N., Kitahashi, T., & Ogawa, H. (1995). On the application of a machine learning technique to fault diagnosis of power distribution lines. Authorized licensed use limited to Concordia University Library.
45. Uhlmann, E., Pastl Pontes, R., Geisert, C., & Hohwieler (2018). Cluster identification of sensor data for predictive maintenance in a Selective Laser Melting machine tool. 4th International Conference on System-Integrated Intelligence, 24, 60-65. Online available at www.sciencedirect.com.
46. Van Tung, T., & Yang, B.-S. (2009). Machine fault diagnosis and prognosis. International Journal of Fluid Machinery and Systems, 2, 61-70.
47. Vathoopan, M., Johny, M., Zoitl, A., & Knoll, A. (2018). Modular fault ascription and corrective maintenance using a digital twin. IFAC-PapersOnLine, 51, 1041–1046.
48. Wuest, T., Weimer, D., Irgens, C., & Thoben, K.D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. Online Journal of Production & Manufacturing Research, 4, 23-45.
49. Ying, X. (2019). An Overview of Overfitting and its Solutions, IOP Conf. Series: Journal of Physics: Conf. Series 1168. DOI: 10.1088/1742-6596/1168/2/022022.
50. Zhang, S., Luo, X., Yang, Y., Wang, L., & Zhang, X. (2018). Optimization of a dynamic fault diagnosis model based on machine learning, 6, 65065-65977.

Appendices

Appendix A

Final Developed Code in Python:

Import Libraries

```
import numpy as np
import pandas as pd
import seaborn as sn
import matplotlib.pyplot as plt
import imblearn
import statsmodels.api as sm
from decimal import Decimal, ROUND_DOWN
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import ADASYN
from statsmodels.tools import add_constant as add_constant
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, accuracy_score, roc_curve, r
oc_auc_score
from sklearn.model_selection import cross_val_score
```

Data Preparation

```
#Print the number of rows and columns
dataset=pd.read_csv("data.csv")
print("The data has {} number of rows and {} of columns".format(dataset.sh
ape[0], dataset.shape[1]))
dataset
#Information about dataset
display(dataset.describe())
dataset.info()
display(dataset.dtypes.value_counts())
dataset
#Remove the duplicates
dataset=dataset.drop_duplicates(subset=['Datetime', 'MachineID'])
print(dataset)
#Seperate numerical variables and categorical variables
Num_vars= dataset.columns[dataset.dtypes!=object]
Cat_vars= dataset.columns[dataset.dtypes==object]
dataset
#Finding the missing values
dataset[Num_vars].isnull().sum()
#Filling missing values
dataset=dataset.fillna(0)
dataset[Num_vars].isnull().sum()
#Finding Outliers
```

```
Num_vars= ['Volt','Rotate','Pressure','Vibration']
dataset.boxplot(Num_vars)
#Filling the outliers
for i in ['Volt','Rotate','Pressure','Vibration']:
  Q75,Q25=np.percentile(dataset.loc[:,i],[75,25])
  Dif= Q75-Q25
  Max= Q75+(1.5*Dif)
  Min= Q25-(1.5*Dif)
  dataset.loc[dataset[i]<Min,i]=np.nan
  dataset.loc[dataset[i]>Max,i]=np.nan
Median= dataset.median()
dataset.fillna(Median, inplace=True)
```

Data Analysis

```
#Plot to show the distribution of continuous variables
XCont=dataset.iloc[ : , 3:7]
XCont.hist(alpha=0.5, figsize=(9, 7), color='black')
plt.show()
#Plot to show the types of machine models
Class_dataset = sn.countplot(dataset['Model'], color='black')
plt.title('Number Of machine for each Model')
plt.xlabel('Model Of Machines')
plt.ylabel('Number Of Machines')
#Plot to show the types of Components
df1=pd.read_csv('df1.csv')
Class_dataset = sn.countplot(df1['comp'], color='black')
plt.title('Number Of Each Fault component')
plt.xlabel('Types Of Component')
plt.ylabel('Number Of Fault Component')
#Plot to show Age of machines
Class_dataset = sn.countplot(dataset['Age'], color='black')
Class_dataset.set_xticklabels(['0','1','2','3','4','5','6','7','8','9','10
','11','12','14','15','16','17','18','19','20' ])
plt.show()
#Plot to show the types of errors
df2=pd.read_csv('df2.csv')
Class_dataset = sn.countplot(df2['errorID'], color='black')
plt.title('Number Of each errors')
plt.xlabel('Types of errors')
plt.ylabel('Number Of Errors')
plt.show()
#Encoding varibales y
x=dataset.iloc[ : , [3,4,5,6,7,8,9,10,11,12,13,14,15,16,17]].values
y=dataset.iloc[ : , 18].values
LE=LabelEncoder()
```

```
y=LE.fit_transform (y)
print(y)
#Separate classes 0 and 1 for failure or not
Class0 = dataset[dataset['Failed'] == 0]
Class1 = dataset[dataset['Failed'] == 1]# print the shape of the class
print('Failed 0:', Class0.shape)
print('Failed 1:', Class1.shape)
#Correlation Coefficient Spearman Method
dataset=dataset.iloc[ : , 0:18]
Spearman=dataset.corr(method="spearman")
Spearman.to_excel(r'frompythonSpearman.xlsx', sheet_name='Sheet1')
#Calculation for logit of continuos variables
ln=np.log
print(dataset[ 'Volt'])
TrVolt=[]
for i in dataset['Volt']:
  TrVolt.append(round(i*ln(i),4))
TrVolt
dataset["TrVolt"]=TrVolt
TrRot=[]
for i in dataset['Rotate']:
  TrRot.append(round(i*ln(i),4))
TrRot
dataset["TrRot"]=TrRot
TrPre=[]
for i in dataset['Pressure']:
  TrPre.append(round(i*ln(i),4))
TrPre
dataset["TrPre"]=TrPre
TrVib=[]
for i in dataset['Vibration']:
  TrVib.append(round(i*ln(i),4))
TrVib
dataset["TrVib"]=TrVib
TrAge=[]
for i in dataset['Age']:
  if i==0:
    i=0.01
  TrAge.append(round(i*ln(i),4))
TrAge
dataset["TrAge"]=TrAge
#Encoding the Model (Making Dummy varibles for model)
Model1=[]
for i in dataset['Model']:
  if i==1:
```

```
    Model1.append(1)
  else:
    Model1.append(0)
dataset['Model1']=Model1
Model2=[]
for i in dataset['Model']:
  if i==2:
    Model2.append(1)
  else:
    Model2.append(0)
dataset['Model2']=Model2
Model3=[]
for i in dataset['Model']:
  if i==3:
    Model3.append(1)
  else:
    Model3.append(0)
dataset['Model3']=Model3
Model4=[]
for i in dataset['Model']:
  if i==4:
    Model4.append(1)
  else:
    Model4.append(0)
dataset['Model4']=Model4
dataset
#Developing GLM model to investigate the linearity
model = sm.GLM.from_formula("Failed ~ Volt+Rotate+Vibration+Pressure+Age+E
1+E2+E3+E4+E5+C1+C2+C3+C4+Model1+Model2+Model3+Model4+TrVolt+TrVib+TrRot+T
rPre+TrAge", family = sm.families.Binomial(), data=dataset)
Result = model.fit()
Result.summary()
#Export the result of developing GLM
Results_summary = Result.summary()
Results_as_html = Results_summary.tables[1].as_html()
Final=pd.read_html(Results_as_html, header=0, index_col=0)[0]
Final.to_excel(r'Assumption.xlsx', sheet_name='Sheet1')
#Developing Logistic Regression model
#Define the new varibales
x=dataset.iloc[ : , [7,8,9,10,11,12,13,14,15,16,24,25,26,27]].values
y=dataset.iloc[ : , 18].values
print(x)
print(y)
#Split the data
```

```
x_train, x_test, y_train, y_test= train_test_split (x,y, test_size=0.2, ra
ndom_state=1)
print(x_train)
print(x_test)
#Balancing the data
adasyn = ADASYN()
xo_train, yo_train = adasyn.fit_resample(x_train, y_train)
#Add constant
xo_train_constant = add_constant(xo_train)
xo_train_constant
#Train the Dataset
LogReg = sm.Logit(yo_train, xo_train_constant).fit()
print(LogReg.summary())
#Explore the results
Results_LR = LogReg.summary()
Results_as_html = Results_LR.tables[1].as_html()
Final=pd.read_html(Results_as_html, header=0, index_col=0)[0]
Final.to_excel(r'frompython1.xlsx', sheet_name='Sheet1')
```

Developing Prediction Model with new variables (scikit learn)

```
#Define the variables
x=dataset.iloc[ : , [7,10,13,14,15,16]].values
y=dataset.iloc[ : , 18].values
print(x)
print(y)
#Split the data
x_train, x_test, y_train, y_test= train_test_split (x,y, test_size=0.2, ra
ndom_state=1)
#Balancing Over-Sampling
adasyn = ADASYN()
xo_train, yo_train = adasyn.fit_resample(x_train, y_train)
#Training
classifier=LogisticRegression(solver='lbfgs', max_iter=5000)
classifier.fit(xo_train, yo_train)
#Predicting Result
y_pred= classifier.predict(x_test)
print(np.concatenate((y_pred.reshape(len(y_pred),1), y_test.reshape(len(y_
test),1)),1))
```

ROC Curve and Accuracy

```
#ROC Curve
logit_roc_auc = roc_auc_score(y_test, y_pred)
FPR, TPR, thresholds = roc_curve(y_test, classifier.predict_proba(x_test)[
:,1])
plt.figure()
```

```
plt.plot(FPR, TPR, label='Logistic Regression (area = %0.2f)' % logit_roc_
auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Fault Predition')
plt.legend(loc="lower right")
plt.savefig('ROC-Curve')
plt.show()
#Accuracy and Confusion matrix
CM= confusion_matrix(y_test, y_pred)
print(CM)
accuracy_score(y_test, y_pred)
print('ROCAUC score:',roc_auc_score(y_test, y_pred))
print('Accuracy score:',accuracy_score(y_test, y_pred))
np.set_printoptions(precision=3)
print(classifier.coef_)
print(classifier.intercept_)
```

K-fold Validation

```
#Applying 5-Fold cross validation
Accuracies = cross_val_score(classifier, x, y, cv = 5)
print("Accuracy: {:.2f} %".format(Accuracies.mean()*100))
print("Standard Deviation: {:.2f} %".format(Accuracies.std()*100))
```