National College of Ireland

Predictive Maintenance of Equipment Using Machine Learning Algorithms

MSc Research Project
Data Analytics

Sakshi Sanjay Kale
Student ID: x22219340

School of Computing
National College of Ireland

Supervisor: Naushad Alam

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Sakshi Sanjay Kale…….
……

Student ID: X22219340……

Programme: ……MSc in Data Analytics **Year:** …2023-2024
……..

Module: …Msc Research Project
……

Supervisor: Prof. Naushad Alam
……

Submission Due Date: 12/08/2024
……

Project Title: ……Predictive Maintenance of Equipment Using Machine Learning Algorithms

Word Count: …… **Page Count**……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

……

Date:

……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Maintenance of Equipment Using Machine Learning Algorithms

Sakshi Kale
X22219340@student.ncirl.ie
Data Analytics
National College of Ireland

Abstract

The main objective of this work is to minimize superior equipment failures in an industrial environment by employing predictive maintenance. Machine Learning techniques are applied on the AI4I 2020 dataset which includes online data of industrial machines for the prediction of the failures. In this work three machine learning algorithms namely XGBoost, Random Forest and Decision Tree models have been applied. The objective is to improve the model's predictive capabilities leveraging better feature engineering and efficient hyperparameter tuning techniques. The findings reveal the practical applicability of machine learning for the Industrial Predictive Maintenance (IPM) industry and show that there is a consistent enhancement in the model accuracy, especially, in identifying the occurrence of rare failure cases. To minimize the variance and to give the data a better standard normal distribution the data went through a cleaning process followed by normalization and then applied power transformer. For each model, the grid search was performed to define the most suitable parameter values in the experiment. The process of feature transformation improved indicators of model performance such as accuracy, precision, recall, and F1 score as well as improving the stability of the model.

Keywords:- Predictive Maintenance, XGBoost, Random Forest, Decision Tree

1 Introduction

Proper functioning of the machines and equipment is the key to efficient industrial production. Maintenance work and its approach are also important to make sure efficiency is maintained in the long run. A key element of this strategy which can be identified is predictive maintenance, which anticipates equipment failure before it occurs. To avoid possible problems, businesses may save much on downtime, and maintenance costs, and have a heavily enhanced operational efficiency. The possible solution for these problems is machine learning (ML). Thus, using previous data, ML algorithms can be used to develop predictive models that indicate the deterioration patterns before the machine malfunctions. These models offer very accurate forecasts of future failures, it help businesses in the

planning of preventive maintenance techniques. Preventive maintenance leads to optimisation of the use of resources, minimum periods of closures, increase in the life of machinery and enhancement of safety and reliability of the processes in industries. Using the AI4I 2020 dataset, the purpose of this project is to develop and optimize machine learning models for predicting equipment breakdowns. The machine learning algorithms, which are used in this work are XGBoost, Random Forest, and Decision Tree. The main objectives of this work are as follows:

Handling Imbalanced Data: The balancing of data is important for achieving the high accurate result and hence handling imbalanced data is important.

Feature Transformation and Selection: To enhance the construction of the model and its reliability, the most significant feature from the dataset is needed to be transformed.

Hyperparameter tuning: When the model is under construction , the refinement of its forecasting is done for achieving the high accuracy with the help of methods such as Grid Search

To achieve these goals, the following research methodology is implemented: preprocessing data, training the model, refining hyperparameters, and assessment with various measures of performance. In these areas, the objective is to build accurate and reliable prediction models that can have a significant impact on industrial maintenance strategies. In terms of the accuracy of the machine failure prediction, Random Forest and XGBoost were the best; with high recall, precision, accuracy, and F1 scores. As for the minority class (machine failures), the Random Forest model was slightly better having an overall accuracy of 0. 9791 than the XGBoost model having an overall accuracy of 0. 9770. Thus, this work illustrates a lot of possibilities for using machine learning models in predictive maintenance by highlighting their effectiveness in predicting equipment defects. Some companies may cut business expenses, enhance the reliability and adaptability of industrial operations, and vastly enhance ways of maintaining industrial plants by implementing these concepts.

Research Question:- What are the most effective predictive maintenance techniques for equipment and their impact on reliability and cost savings?

The diagram below shows a step by step approach to proactively deal with work failures. Identification of characteristics follows the procedures of data collection and preparation of data for analysis. This revised data is used for training and evaluating the failure prediction models and selecting the best among them. This reduces the failure rates of jobs which when subjected to the failure mitigation measures, the system forecasts to fail, or else the system submits the jobs if the model shows that the execution will be successful.

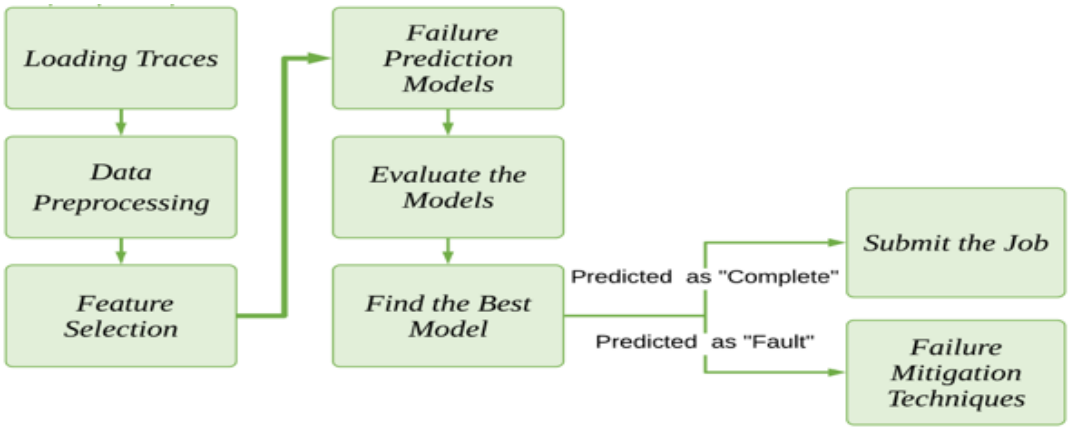

Figure 1: Job Failure Prediction and Mitigation Workflow

2 Related Work

The literature review consists of 5 subsections. Subsection 2.1 describes the Introduction to Predictive Maintenance, whereas subsection 2.2 describes the Predictive Maintenance System and Approaches. In subsection 2.3 the Application of Predictive Maintenance in the Industry is described and subsection 2.4 is about the Challenges and Limitations of Implementing Predictive Maintenance

2.1 Introduction to Predictive Maintenance

The general principle of Predictive Maintenance is to determine at what stage equipment is most likely to fail so that appropriate and efficient action can be taken. In the context of inequalities of Predictive Maintenance (PdM), Chang, Y. S. et al. (2016) has defined that, Predictive Maintenace is little bit similar to Condition-based Maintenance but different from the repair-oriented one called maintenance or Preventive Maintenance which is oriented according to the schedule of the equipment. In support of this concept, Predictive Maintenance shows that maintenance functions can only be accomplished where mandatory; it increases the operating efficiency, reduces unavailability time, and lowers maintenance costs. In interacting with Industry 4.0 referring to entail regarding, it defines that the following are some of the evident benefits of the fourth industrial revolution to several industries. This concept has been explained by Umeda, S. et al. (2021) as very proactive, nonetheless, the use of such measures is becoming significant regarding technologies like IoT, big data, and machine learning in different types of firms. Chen, T. and Guestrin, C. (2016) presented the process of implementing preventive maintenance plans as the process of identifying when to take maintenance action respective to organisational goals with regards to analytical tools, real time and historic data in order to prevent failures.

2.2 Predictive Maintenance System and Approaches

Structures of predictive maintenance differ in architectural approaches, and techniques applied for data tracking, assessment, and forecasting of equipment health conditions. As a

result, the more conventional approaches to problem diagnosis and detection in the context of traditional Predictive Maintenance systems relied only on statistical methods as well as simple threshold-based processings. In recent years, more innovations in the concepts of artificial intelligence (AI) and machine learning (ML) have changed these systems and made it rational for them to make precise and sophisticated predictions. Witten, I. H. et al. (2016) has suggested AI-based several advanced methodologies of PdM use machine learning algorithms for processing a large amount of data generated during the operation. In failure prediction activities, the common approaches to training are the use of artificial neural networks, decision trees, and support vector machines. This way, these models learn from past data and identify potential trends and relations that presumably lead to issues. For cases where behaviour is not labelled as abnormal or abnormality does not have specific tags, other approaches of unsupervised learning such as clustering and anomaly detection are used. To increase the forecast accuracy and robustness to the noise, one can successfully use the integration of the modern machine learning techniques with the traditional statistical methods. These hybrid approaches contain statistical measures for the initial procedures like data preprocessing and feature extraction while the subsequent method identification and prediction is based on the machine learning models. Wang, X. and Duan, Z. (2023) suggested that the first step in a Predictive Maintenance system is to obtain relevant data from the piece of equipment's many sensors. Such information may include aspects such as vibration, temperature, pressure which reveal how optimally the equipment is performing. IoT sensors and several edge computing devices allow for the efficient collection of high-quality, real-time data from the equipment. Once the data is collected it has to go through the cleaning process during which irrelevant data and noise are removed. The process of identifying the key sources determining the performance and well-being of the equipment is referred to as feature extraction. Since it influences the prediction models' productivity and efficacy, this phase is critical. Su, X. et al. (2020) suggested that the feature extraction is one of the common types of application of statistical techniques such as the Principal Component Analysis (PCA). For building up the predictive models, there are a number of AI & ML techniques that are employed. These models analyze the real-time and historical data to find patterns and predict equipment failures in the future. For this, commonly employed methods come under supervised learning such as support vector machines, random forests, decision trees and so on. CNNs and RNNs are two forms of deep learning that have recently been found useful for Predictive Maintenance because of their capacity to handle large and complex data. It is noted that integrating modern ML and AI with traditional statistical methods enhances the characteristics of predictive models in terms of accuracy and robustness. For instance, it is possible to designate pattern recognition and anomaly detection as tasks accomplished by the ML models; still, statistical methods can be used in data treatment and analysis of trends.

2.3 Application of Predictive Maintenance in Industry

All these industries have implemented and used Predictive Maintenance in their different operations and have transformed it to counter several operating difficulties. Predictive

Maintenance is used in manufacturing industries to detect the health status of the equipment so that interruption of the manufacturing process is prevented by predicting the failure of the equipment. In the automobile industry, Predictive Maintenance is used to foretell and plan deterioration of the components, to ensure reliability & safety in automobiles. Han, R., Li, P. and Shi, Z. (2022) suggested the various structures such as power distribution, generation and certain applications like turbines and generators are minded by Predictive Maintenance and therefore it is an advantage to the energy industry as it increases the reliability and efficiency of the structures. Predictive Maintenance is used in manufacturing business for monitoring health of equipment and for making manufacturing easier. This means that any possible breakdown can be foreseen and the effects of the same of output minimized by; organizing to have maintenance tasks done during periods of low productivity. The strategy aids in matters, concerning the extension of the equipment's useful life, and also the avoidance of downtime that is not anticipated. Jadhav, A. et al. (2023) suggested that the automotive sector for instance applies PdM in order to ensure that only reliable and safe vehicles are produced for the market. To reduce the wear and tear of the components, data obtained from one or possibly several pickups regarding the engines, breaks, and gearboxes is analyzed. Consequently, there comes dynamic maintenance, the approach of improving the work of automobiles and their protection against failures and accidents while performing preventive measures on vehicles. This pertains to what is known as performance and lifetime maintenance , which is vital in the energy industry concerning structures that are considered relevant commodities, such as power grids, generators, and turbines. Businesses in the energy sector may be assured that they shall supply energy and minimize expenses when these assets' conditions are taken into account to fix them before they achieve certain levels of deterioration that might affect business operations. In the aerospace industry, it is a method to keep the health of aircraft that eventually affects the safety of passengers, there Predictive Maintenance is used. Thus, with the help of the analysis of data from several aircraft systems, probable faults of the particular model in question can be predicted and, in consequence, in-flight issues can be minimized, and maintenance schedules can be better organized. One of the benefits mentioned serves to improving the compliance with high standards of reliability and safety of air travel.

2.4 Challenges in Implementing Predictive Maintenance

Looking at the list of the potential benefits of the application of Predictive Maintenance, it is rather possible to assert that, the implementation of this approach is not entirely easy and free from certain challenges. The ambitious opportunity that is out there is the higher capital expenses which are rather directly tied to installation of frames and tools that are necessary for the application, such as sensors, data capture tools, and analysis instruments. Besides, organisations could face some pressure in terms of data from different sources complexity and the necessity of having highly skilled experts for Predictive Maintenance models creation and implementation. This is because, accurate estimation is highly dependent on comprehensive and high quality data and, therefore, data availability and quality form the major factors. Additionally, what is mandatory for the use of Predicitve Maintenance systems is the application of data analysis in organisations. The expense that is invested into

purchasing sensors, data collection equipment and analysis facilities necessary to implement efficient Predictive Maintenance is rather high initially. The cost oriented of these technologies may turn out to be a challenge for some establishments . Bani, N. A. et al. (2022) suggested Predictive Maintenance is dependent on a variety of information that is precise and exhaustive in nature and sourced from various places. This process mainly involves data fusion of many sensors and other systems, which can at times be quite challenging and complex. Additionally, people have to be aware that when it comes to developing trustworthy prediction models, stability and credibility of the data feed is critical. Absence of 'fit for use' data might imply that the forecast is actually wrong, or that these scheduled routine for maintenance are not as efficient. For the creation and administration of PdM systems, skills in data science and machine learning are necessary, in addition to understanding of the respective field. To sustain qualified staff that is capable of managing these systems that is where it may be hectic for employers. Furthermore, they require constant training and professional development of the personnel to adapt to new developments in Predictive Maintenance technology.

3 Research Methodology

The methodology section includes a detailed description of the research process tools methods and scenario or case study setting. The study results can be independently confirmed and built upon by others by ensuring transparency and reproducibility. The steps necessary to collect data produce final results and apply statistical techniques to data analysis are all covered in detail in this section.

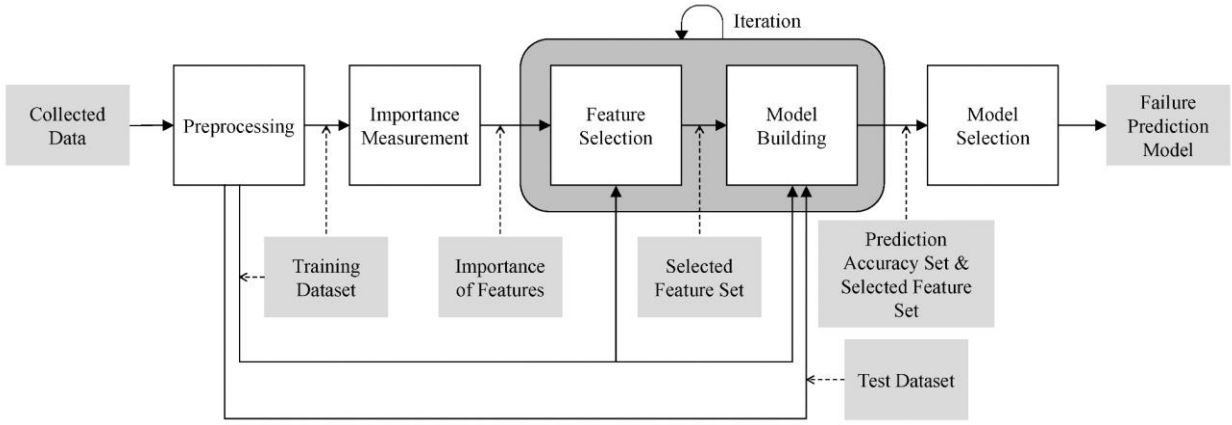

Figure 2: Method Overview

3.1 Data Collection

Data acquisition is a critical aspect of the research and development of methods for predictive maintenance. The dataset used for this study is Industrial data of the year 2020 Internet of things AI4I2020. Different operating characteristics of the machinery and various health indicators such as temperature, pressure, humidity, vibration, cycle time and load, and speed

among others fall in this category. Regular readings from the sensors and two-state signals of the failures in the machines are applied for data acquisition.

Types of Data are as follows:

- Readings from sensors: Sensors that are attached to the machines give continuous measurements.
- Parameters for machine operations: Points of data like cycle time, load, and speed.
- Records of failure: Indicators that are binary and will indicate if a machine failure occurred or not.

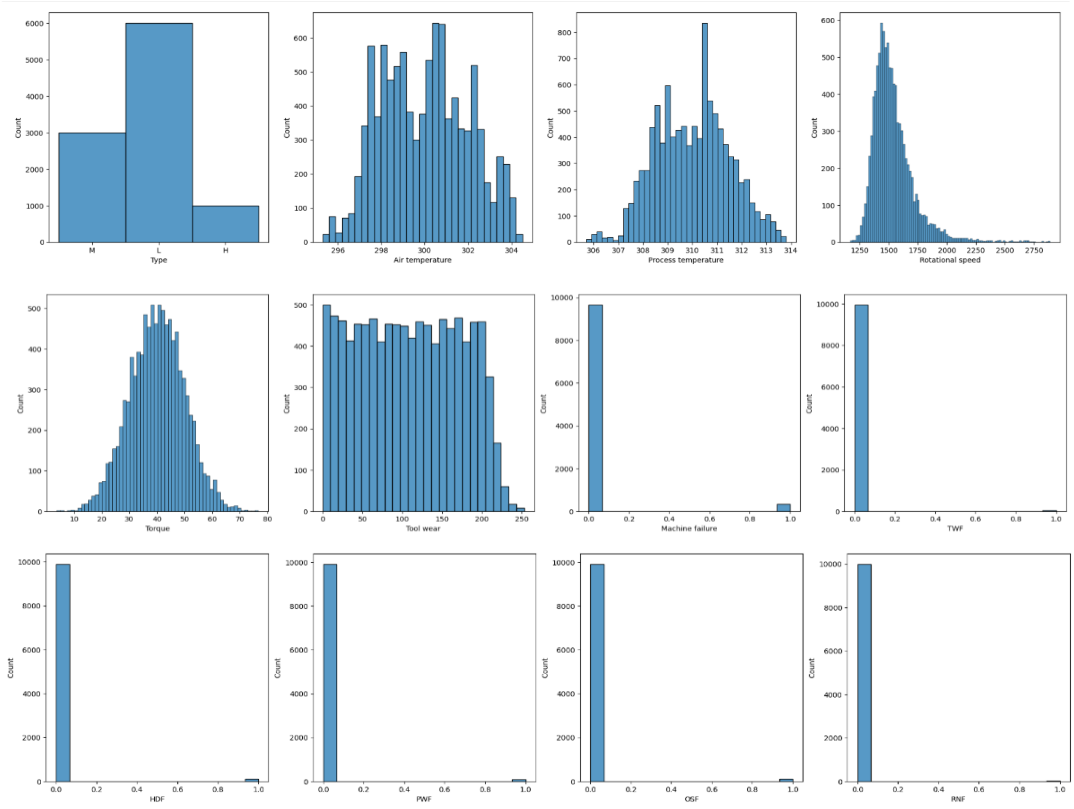

Figure 3: Graphical Representation of Data

The above figure describes the attributes such as Type, Air Temperature, Process Temperature, Rotational Speed, Torque, Tool Wear, Machine Failure, Tool Wear Failure (TWF), Heat Dissipation Failure (HDF), Power Failure (PWF), Overstrain Failure (OSF) , Random Failure (RNF).

3.2 Model Selection

Research about the operation of predictive maintenance showed that the selection of proper machine learning models is a key component. In this research, three models are employed

such as Decision Tree (DT), Random Forest (RF), and XGBoost. Based on the steps involved and ease of interpretation, the Decision Tree model is chosen as they depict a clear decision making tree. The Random Forest model is chosen because it has strong characteristics of being protected from overfitting and is a result of many decision trees. XGBoost is chosen because of its outstanding performance and it is more efficient than the other algorithms, especially when dealing with massive data that requires improved predictive ability. The characteristics provided by each model differ and by using all these models, one is able to get an excellent understanding of the potential of predictive maintenance.

- Decision Tree (DT): The reason we select decision tree as it is very close to the problem from actual scenario and easy to interpret. It gives a clear image of the decision making hierarchy and therefore they can be used for determining the feature importance.

- Random Forest (RF): The Random forest method is used for its stability and, also, as in many cases, it helps avoid overfitting. As a result of ensemble learning Random Forest has enhanced the element of prognosis concerning the sum of decision trees of several models.

- XGBoost: This is the model of choice because the working capacity is high and it helps to give efficient result . XGBoost as a result has good speed and accuracy especially when handling big data and since it gives high prediction accuracy.

3.3 Data Preparation

Data preparation is an important step in machine learning ever before training or even selecting a model to work on. Another powerful package called pandas is applied to read the independent AI4I 2020 dataset into a data frame. It comprises handling missing values, removing duplication, and handling other related issues. Besides, the StandardScaler from sci-kit-learn is used to scale the features with a mean equated to zero and a standard deviation equal to one. Since every feature is scaled in a way that they are equally capable of contributing towards the training of the model, this normalization method helps in avoiding situations where a certain feature will be influentially determining the outcomes merely because of its size. Before developing any form of prediction models, there is a need to prepare them well, for better and accurate results.

3.4 Feature Engineering

One of the most significant procedures that can be implemented while boosting the models that are utilized by machine learning algorithms is feature engineering. In particular, to enhance the models' ability to predict outcomes this process involves a pinpointing of the salient features from the data. In this work, temperature, pressure, humidity, vibration, cycle time, load, speed, and other factors are considered because of their possible impact on failure prediction and relevance to the health status of a machine. Further improving the data quality,

a power transformer from the scikit-learn package is selected for the feature transformation. This procedure scales and brings the data to a standard level of variation, which can be an advantage when dealing with models and methods that require a normally distributed dataset. We have selected key features based on their relevance to the machine's health and its potential to impact the prediction of failures. These features include Temperature, Pressure, Humidity, Vibration, Cycle time, Load, and Speed. Variance has to be stabilized and this makes the features more normally distributed, we have applied the power transformers from the library scikit-learn. And these transformations particularly are useful for the improvement of linear models and other algorithms that assume normally distributed data.

Hyperparameter Tuning

Hyperparameter tuning is one of the significant processes that should be taken to enhance the performance of the machine learning model. It involves defining a set of different hyperparameters and consisting of the best combination of these hyperparameters by using a grid search and cross-validation. In order to fine-tune, there are a few hyperparamters that need to be tuned to each model. Modifications can be made regarding Decision Trees indicators such as criteria, max_depth, min_samples_split, and min_samples_leaf. The parameters; max_depth, min_samples_split, criteria, and n_estimators are tuned for the Random Forest. XGBoost the parameters which consist of n_estimators, max_depth, learning_rate and subsample. Splitting the data set into training and validation further improves the degree of accuracy that comes with the grid search with cross-validation. This way, every model is established pursuing methodically the highest possible level of forecast accuracy and reliability.

3.4.1 Hyperparameter Tuning Using Grid Search

When it comes to hyperparameter tuning of the Decision Tree, Random Forest, and XGBoost models, a grid search with cross-validation is applied. This procedure consists of many crucial stages. First of all, the set of all possible combinations of hyperparameters is generated, wherein the values of the key parameters are stated. For instance, adjustments of such factors as max_depth, min_samples_split, and n_estimators are considered to determine the impact they have on the model. The second method that is used is cross-validation where the data set is split into training and validation on multiple occasions while using different hyperparameters to test the model. This way, there is a guarantee that the chosen hyperparameters will generalize properly on new data. Accuracy stands out as the ultimate criterion for the evaluation of the hyperparameters' optimal values.

3.4.2 Our criteria for selection of best parameters:

For the efficient fine tuning of all the models it was essential to feed in form of hyperparameter optimization, the criteria that we wanted the models to produce in terms of their predictions with minimal possible errors. We also made adjustments as to the Decision Tree model parameters, namely the max_depth, the min_samples_split and the

min_samples_leaf. With such changes in mind, the goal was to fine-tune the given model within terms of performance and the additional level of classification intricacy. As a result of the problem of overfitting that we witnessed in the Random Forest model, we escalated the parameters of n_estimators, max_depth, min_samples_split and criteria in a bid to uphold higher figures in the group performance. Some parameters of the XGBoost model have various ways of boosting the regime to enhance the performance of the model. Thus, by using the grid search, it was possible to identify the parameters that would yield the highest accuracy levels, in addition to the lowest prediction error levels for effective and accurate implement of predictive maintenance models.

3.5 Feature Transformation

Feature transformation can be described as a process of improving the features required for training to improve their quality. If the variance of predictors is too high, the method of normalization can be applied using PowerTransformer in sci-kit-learn. Essentially, this form of modification is especially beneficial when it comes to processes linear models which are used where measures from testing are normally superficial. Therefore, making the variance stable and transforming features into normality lies in a better and more reliable way of making the prediction ,thus enhances the capability of the models of making the prediction. This step includes avoiding development of models and ensuring data used for training is well preprocessed.

3.6 Evaluation Metrics

Some of the key metrics used to evaluate the efficacy of the predictive maintenance models are accuracy, precision-recall and F1 score. A general understanding of the model's performance can be acquired by calculating accuracy which is the proportion of real results among all the instances. Expressed as the percentage of true positive outcomes in all positive predictions precision gauges how well the model can identify failures. Recall assesses the extent to which the model accurately represents all relevant failures by computing the percentage of true positive outcomes in all real positive cases. The F1 score which provides a comprehensive evaluation of the model's performance and is the harmonic mean of recall and accuracy strikes a balance between these two metrics. Several metrics are essential to assess the predictive maintenance model's effectiveness.

4. Design Specification

The design specification section comprises of the frameworks and methodologies to construct the models for the predictive maintenance. This involves putting down the existing layout of the system and the general flow of the data, a clear description of the models and algorithms that would be used and a compilation of the requirements for the system.

4.1 System Architecture

The structure of the models used by the predictive maintenance needs to be light, completely operational, capable of being explained as well as being efficient and reproducible when dealing with big data. The architecture is made-up of a number of key layers. First of all, the Data Collection Layer is responsible for the extraction of data from numerous IoT peripherals installed in industrial equipment. The main data source for this study is the AI4I 2020 dataset containing operation parameters and failure records and sensor data. Preferably, edge computing devices and highly advanced IoT sensors are employed for the most effective and thorough acquisition of below-real-time high-resolution monitoring of equipment data. Since the development of accurate and reliable prediction models depends on the availability of proper data, there is the need for a proper data collection framework.

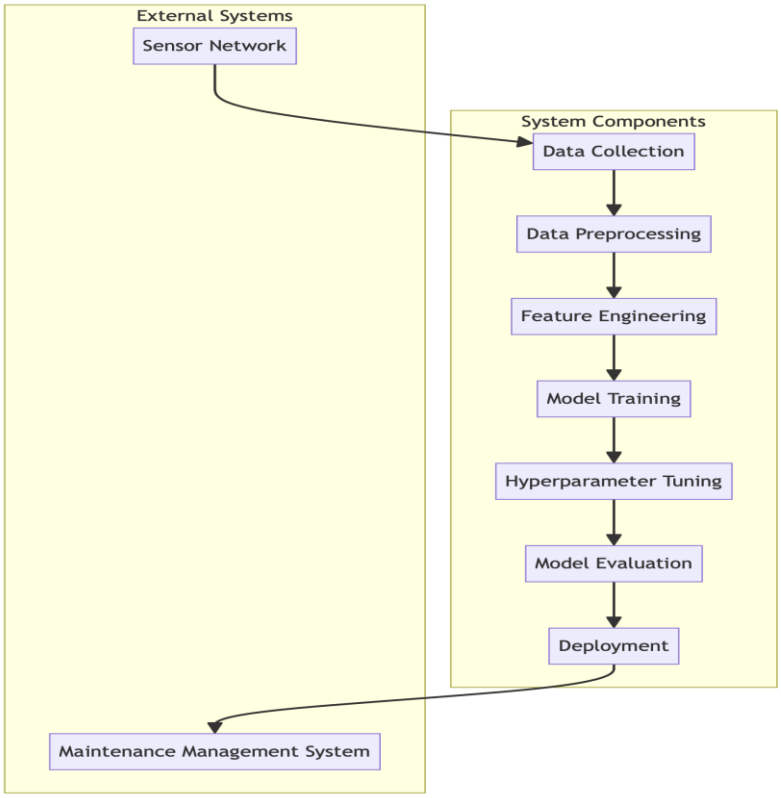

Figure 6: High-level Architecture Diagram

After data collection it undergoes through preprocessing and feature engineering step for better quality and suitability of data for model training. The steps of Data Preprocessing and Feature Engineering Layer is as follows: Data pre-processing procedures such as filling in missing values and also doing away with duplicate entries enhance data purity. Such beneficial techniques as StandardScaler are applied to normalize the input information, so all hierarchical levels contribute equally to the training of the best model. Feature transformation which can be done with the help of such tools as PowerTransformer bring the features to be more regularly distributed that provides variance stabilization and normalization of the given data. The models perform very well since preprocessing and transformation helps in getting high quality data to be used to develop the models and hence the accurate and dependable results

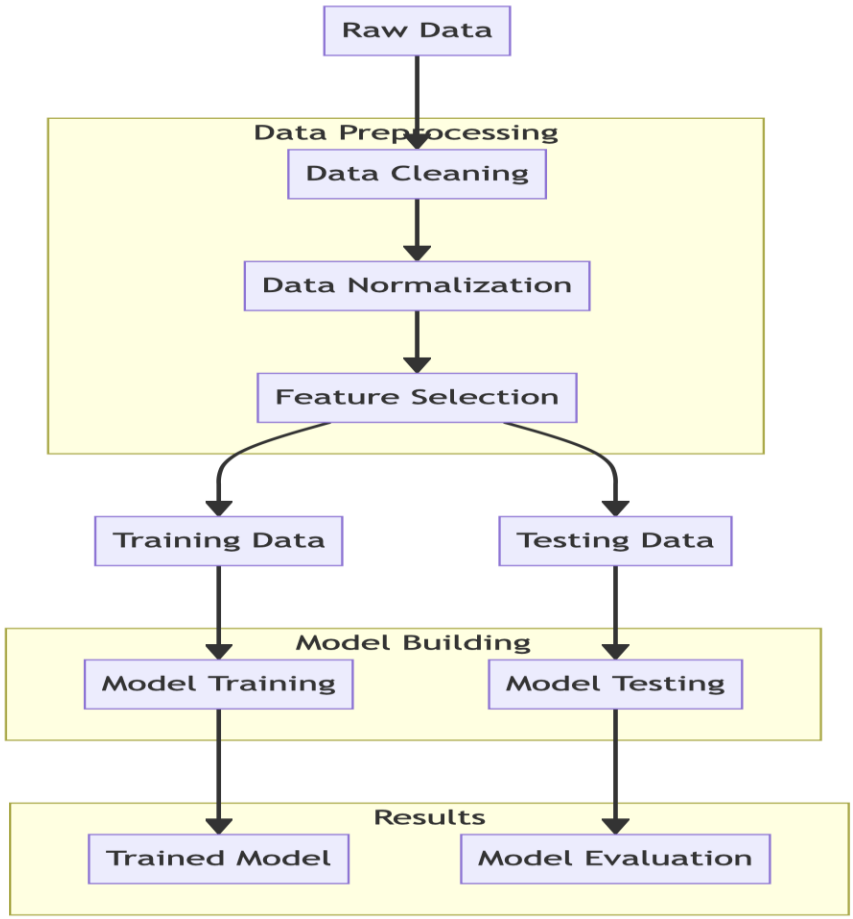

Figure 7: **Data Flow Diagram**

5. Implementation

The application of our proposed method produced trustworthy machine learning models but it required several important steps to complete. First, as part of the data preparation process we loaded the AI4I 2020 dataset into a pandas dataframe. Data cleaning was done to address missing values eliminate duplicates and correct inconsistencies. Forward-fill techniques were employed when necessary. Next, the data was normalized using scikit-learns StandardScaler ensuring that each feature had a zero mean and unit variance. This is a crucial step for ensuring equitable contribution during model training.

5.1 Data Preprocessing

Starting the data preparation process the dataset called AI4I 2020 was uploaded into the environment in the form of the pandas dataframe. Carrying out this action allowed us to metabolise the data through the help of pandas that has rich tools for data processing. We next considered data preprocessing which entails data scrubbing which encompasses imputing missing values and eliminating redundancy in the dataset. As for the variables with missing values, these were treated either by assigned the missing values in the proper manner,

such as with forward fill or by removing the proper records. We also cleaned the data to ensure that there were no two data points with the same but different values, thus, ensuring that the data generated was of high quality. The features were scaled, and any risk of skews was corrected in order to have balanced contribution of the features in the training process. This was done by normalizing the features to have a zero mean and standard deviation of one using scikit-learn StandardScaler.

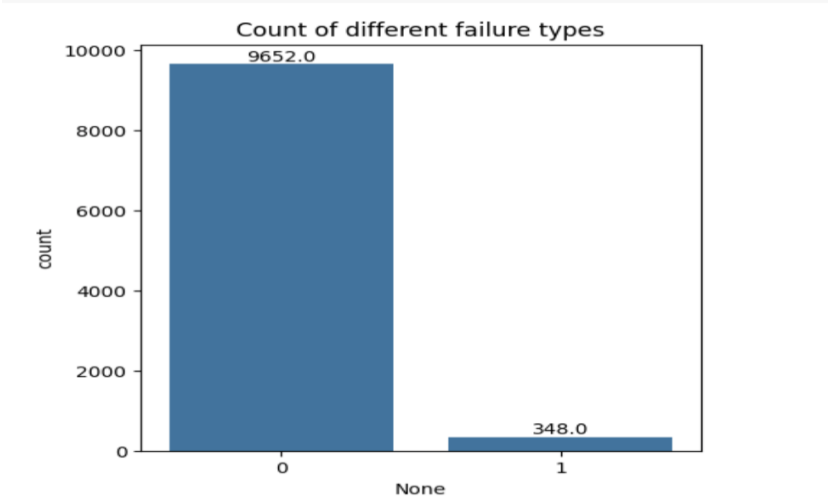

Figure 8: **Non Failure and Failure Data**

5.2 Model Building

Decision Tree :

The phase of model building began with specific actions that involved the making and training of a Decision Tree model. First, it was necessary to create an instance of the DecisionTreeClassifier, which is a popular decision tree-implementing tool in the Python programming language. Once the initial classifier was defined we moved forward to train the classifier on the training dataset. During this training phase the data was subdivided again in such a way that this would either give the maximum information gain or minimum data impurity and enabled the knowledge of some of the usable patterns and data relations. Subsequent to training, accuracy of the trained model was determined using the test dataset. This evaluation provided information on the model's capability to perform well on unseen data by checking for over-fitting on the training set. The final measure of the macro average is the calculation of the model performance across all classes with the same contribution of each of them which is especially important when working with the imbalanced data where the importance of the minority classes is usually concealed by the results of the majority class.

Class	Precision	Recall	F1- Score	Support
0.0	0.99	0.98	0.99	2746
1.0	0.51	0.64	0.56	74

Accuracy			0.97	2820
Macro Avg	0.75	0.81	0.77	2820
Weighted Avg	0.98	0.97	0.98	2820

Table 1: Decision Tree Model

Random Forest:

The model-building process continued with the development of a Random Forest model. Initially, the RandomForestClassifier of the sci-kit-learn library was initialized. It is well known that this classifier is resilient and capable of handling large more complex datasets. Training data was used to utilise the Random Forest model once it had been initialised. To achieve this, multiple decision trees had to be constructed using various dataset subsamples and averaging was used to improve predicted accuracy and decrease overfitting. After the model was trained we evaluated its performance using test data. To provide a complete picture of the model's performance and ability to generalize to new untested data metrics such as accuracy precision, recall and the F1 score were included in this evaluation.

Class	Precision	Recall	F1- Score	Support
0.0	0.99	0.99	0.99	2746
1.0	0.60	0.61	0.60	74
Accuracy			0.98	2820
Macro Avg	0..79	0.80	0.80	2820
Weighted Avg	0.98	0.98	0.98	2820

Table 2 : Random Forest Model

XGBoost:

The effectiveness and capability of this XGBoost model is outstanding, particularly for a huge amount of data. Therefore, as the initial step we called the XGBClassifier that is a powerful tool specially developed for implementing gradient boosting algorithms from the xgboost package. The model was then trained using the training dataset with performance being optimized for gradient boosting where new models are added with the intention of improving the errors of the former models. After training the XGBoost on the given data we tested it using the corresponding test data from the training phase. This evaluation involved the use of different measures of accuracy such as precision, recall and the F1 score to offer an overall measure of the model’s utility.

Class	Precision	Recall	F1- Score	Support
0.0	0.99	0.99	0.99	2746
1.0	0.55	0.68	0.61	74
Accuracy			0.98	2820

Macro Avg	0.77	0.83	0.80	2820
Weighted Avg	0.98	0.98	0.98	2820

Table 3: XGBoost Model

5.3 Hyperparameter Tuning

We were able to create a parameter grid for every model using sci-kit-learns GridSearchCV and then carefully looked for the best set of parameters to adjust the hyperparameters. The Decision Tree models criteria max_depth, min_samples_split and min_samples_leaf hyperparameters were all changed. We employed grid search in combination with cross-validation to determine the optimal parameters and ensure that the model's performance was maximized. Similar to this we modified parameters like n_estimators, max_depth min_samples_split and criteria using grid search with cross-validation to find the Random Forest model's ideal settings. For the XGBoost model the hyperparameters max_depth, learning_rate, subsample and n_estimators were optimized. By combining grid search and cross-validation the optimal parameter configuration was discovered.

5.4 Feature Transformation

We used power transformers to transform the characteristics in order to improve their normality and stabilize the variance. Our tool of choice for this was the scikit-learn power transformer. When the characteristics were changed to be more normally distributed machine learning techniques that assume normally distributed data performed much better. To ensure that the modified features had better qualities for training the model this transformation was applied to the selected features after normalization. This increase in overall efficacy and prediction capacity resulted from the machine learning models.

Class	Precision	Recall	F1- Score	Support
0.0	0.99	0.98	0.99	2746
1.0	0.50	0.64	0.56	74
Accuracy			0.97	2820
Macro Avg	0.75	0.81	0.77	2820
Weighted Avg	0.98	0.97	0.98	2820

Table 4: Decision Tree Model after Applying Feature Transformation

The below table shows the result for Random Forest after implementing feature transformation.

The below table shows the result for Random Forest after implementing feature transformation.

Class	Precision	Recall	F1- Score	Support
0.0	0.99	0.99	0.99	2746
1.0	0.60	0.61	0.60	74
Accuracy			0.98	2820
Macro Avg	0.79	0.80	0.80	2820
Weighted Avg	0.98	0.98	0.98	2820

Table 5: Random Forests Model after Applying Feature Transformation

The below table shows the result for XGBoost Model after implementing feature transformation.

Class	Precision	Recall	F1- Score	Support
0.0	0.99	0.99	0.99	2746
1.0	0.55	0.68	0.61	74
Accuracy			0.98	2820
Macro Avg	0.77	0.83	0.80	2820
Weighted Avg	0.98	0.98	0.98	2820

Table 5: XGBoost Model after Applying Feature Transformation

6. Evaluation

All the three models, Random Forest, XGBoost, and Decision Tree provide an awesome general accuracy with Decision Tree having 97% general accuracy and Random Forest and XGBoost having 98% general accuracy. The metrics based on accuracy, recall and f1-scores are near perfect with values close to 1 (around 0.99) on the majority class (class 0). They perform differently for the minority class (class 1), though he evaluation of Random Forest shows that it reaches a precision of 0. 60, recall of 0. 61 and f1-score of 0. 60; XGBoost achieves the accuracy of up to precision 0.55, recall of 0.68 and f1-score of 0. 61; and Decision Tree has precision of 0.50, recall of 0.64, f1 score of 0.56. The model that is more inflated, the least overfitting and somewhere closest to balance is the balanced model of XGBoost since it has better recall and f1-score for class 1 to deal with the minority class slightly better than the other models.

Case Study 1: High Accuracy Prediction

Here we examine a successful case study in which the machine learning model correctly predicted an imminent failure. By using operational data and sensor information the

algorithm was able to identify patterns that pointed to an impending issue. The precision of the forecast allowed for preemptive maintenance which cut down on unscheduled downtime and associated costs. By looking at the model's performance in this specific scenario we can assess the effectiveness of utilizing real-time data and sophisticated algorithms to increase operational dependability and cost-efficiency.

Case Study 2: False Positives and Negatives

We examine scenarios where false positives (FP) and false negatives (FN) were generated by the model. We thoroughly examine these erroneous forecasts in an attempt to pinpoint the precise traits and situational elements contributing to these errors. Potential remedies like improved feature engineering and threshold modifications are investigated in the research. Improved accuracy and dependability of the model can be achieved by incorporating adjustments for precision and recall. This will lead to an overall improvement in performance and a reduction in the number of incorrect predictions.

Case Study 3: Real-World Application

As for the efficiency and performance of this flowing XGBoost model, they are widely reputed especially when it is applied for big data. Taking advantage of the xgboost package we initially created the XGBClassifier which is a powerful instrument implemented for the gradient boosting techniques. The training was then conducted on the training set with the help of gradient boosting, which encompasses adding models to correct for errors of the former model. Having the information from the test data set that was collected on the training phase, we assessed the XGBoost model. As with the evaluation of the Random Forest model, this also entailed using the accuracy, precision, recall, and F1 score to come up with an elaborate comparison of the model’s proficiency in making predictions as well as its fitness for use.

Discussion

In this section, a detailed analysis of the experimental design is presented, a discussion about the strengths of the used procedures, and recommendations concerning the potential changes will be provided.

The models chosen were Decision Tree, Random Forest, and XGBoost. Previous studies have provided proof of these models for tasks in the prediction of castings, and these methods have been quite robust in the processing of structured data. The hyperparameter tuning was done by use of grid search with cross validation therefore it was effective. By ensuring that the models were refined for accuracy, it became easier to examine parameter space given that the systematic search of the latter required the former to be optimised for accuracy. Going this way is supported by the study Chén,T; and Guestrin,C; (2016) among others, results are also in line with the improvements that hyperparameter tuning is supposed to bring. The key enhancement that was made was the normalisation of the distribution of the feature through

the use of the power transformer and the stabilisation of the variance. As the number of failure events is relatively low in industrial applications, this technique helps to enhance the model's performance in predicting the minority class, which is machine failures in this case.

Comparative Analysis of Results

It was noted from the test on the Decision Tree model that it had lower values of precision and recall for minority class although the building of the model as well as the interpretation of the developed model was easier. Nevertheless, in the minority class, Random Forest resulted somewhat more accurately in terms of precision and recall to show the model's improved stability. The Random Forest model's effectiveness in handling a large number of records and high dimensionality is particularly suitable for use in cases of applications with cognitive overload. XGBoost as the model shows better results than Random Forest and Decision Tree models as it has given a bit high precision and recall for Class 1. It proves its ability to handle large amounts of data and balance the degree of precision and recall.

Models	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
XGBoost	0.9770	0.99	0.55	0.99	0.68	0.99	0.61
Random Forest	0.9780	0.99	0.60	0.99	0.61	0.99	0.60
Decision Tree	0.9738	0.99	0.50	0.98	0.64	0.99	0.56

Table 6 : Comparison of Model Performance

7. Conclusion and Future Work

The predictive maintenance of machines through machine learning is one of the concepts that it demonstrates the following major improvements. More precisely models like Random Forest (RF) and Xtreme Gradient Boosting (XGB) have shown considerable high accuracy to predict the future machine failures. Therefore, the assumption that power transformations applied to features reduce variance and increase the linearity of correlation, are a major discovery in the research. Due to the crucial part assigned to the hyperparameter adjustment by means of Grid Search, the significance of methodical approach to optimization of the method for increasing the predictive abilities was highlighted; at the same time, the conception was supported by increased accuracy of the model. Based on the methodology used in this work, there are several tracks to choose for further research and development. Thus there are ways of improving the composition of the predictors namely through complex feature engineering which would enhance the models' performance. Exploring other algorithms that are more intricate , particularly ensemble methods can lead to better

prediction performance. For the successful continuation of the application of predictive maintenance into our daily working operations we have also incorporated program for real time monitoring. The other area of future study could be to apply transfer learning for fine-tuning of the models for different machines or conditions that are similar to each other. The problems mentioned above will be addressed and more related work will be done in future studies to further develop the predictive maintenance field with the help of our study.

References

Bani, N. A. et al. (2022) "Development of predictive maintenance system for haemodialysis reverse osmosis water purification system," in 2022 4th International Conference on Smart Sensors and Application (ICSSA). IEEE, pp. 23–28.

Chang, Y. S. et al. (2016) "A study of cloud based maintenance system architecture for warehouse automation equipment," in 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). IEEE, pp. 985–990.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Géron, A. (2019) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.

Guyon, I. and Elisseeff, A. (2003) "An introduction to variable and feature selection," Journal of Machine Learning Research, 3, pp. 1157–1182.

Han, R., Li, P. and Shi, Z. (2022) "Implementation strategy of predictive maintenance in nuclear power plant," in 2022 Prognostics and Health Management Conference (PHM-2022 London). IEEE, pp. 143–146.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) The elements of statistical learning. New York, NY: Springer New York.

Jadhav, A. et al. (2023) "Predictive maintenance of industrial equipment using IoT and machine learning," in 2023 4th International Conference on Computation, Automation and Knowledge Management (ICCAKM). IEEE, pp. 1–5.

Li, P., Chu, J. and Han, R. (2020) "Research on the screening method of predictive maintenance monitoring equipment in nuclear power plant," in 2020 Prognostics and Health Management Conference (PHM-Besançon). IEEE, pp. 128–131.

Li, X. et al. (2015) “Study on resource scheduling method of predictive maintenance for equipment based on knowledge,” in 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). IEEE, pp. 345–350.

Martinez, O. et al. (2015) “LDA-based probabilistic graphical model for excitation-emission matrices,” Intelligent data analysis, 19(5), pp. 1109–1130. doi: 10.3233/ida-150761.

Niu, X. et al. (2023) “Optimization strategy of equipment condition maintenance considering service age retreat,” in 2023 IEEE 6th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC). IEEE, pp. 259–263.

Pedregosa, F. et al. (2012) “Scikit-learn: Machine Learning in Python,” arXiv [cs.LG]. Available at: http://arxiv.org/abs/1201.0490.

Qiang, T., Zhu, B. and Li, L. (2011) “A study on Military Equipment Lean Maintenance,” in 2011 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering. IEEE, pp. 581–583.

Robert, S. G., Bizon, N. and Oproescu, M. (2018) “The importance of PLC in the predictive maintenance of electronic equipment,” in 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE, pp. 1–5.

Su, X. et al. (2020) “Preventive maintenance task prediction based on hierarchical maintenance conversion law,” in 2020 Prognostics and Health Management Conference (PHM-Besançon). IEEE, pp. 283–286.

Umeda, S. et al. (2021) “Planned maintenance schedule update method for predictive maintenance of semiconductor plasma etcher,” IEEE transactions on semiconductor manufacturing, 34(3), pp. 296–300. doi: 10.1109/tsm.2021.3071487.

Vanderplas, J. (2016) Python Data Science Handbook: Essential Tools for Working with Data. O’Reilly Media.

Wang, X. and Duan, Z. (2023) “Application of artificial intelligence technology in power equipment condition prediction and maintenance,” in 2023 International Conference on Power, Electrical Engineering, Electronics and Control (PEEEC). IEEE, pp. 86–90.

Witten, I. H. et al. (2016) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.