Applied Information Technology And Computer Science Vol. 3 No. 2 (2022) 961-972

AITCS

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/aitcs
e-ISSN :2773-5141

A Model in Predictive Maintenance for a Manufacturing Company

Bethany Muyou Angkaus[1], Azizul Azhar Ramli[1]*

[1]Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, BatuPahat, 86400, MALAYSIA

DOI: https://doi.org/10.30880/aitcs.2022.03.02.058
Received 17 August 2022; Accepted 28 October 2022; Available online 30 November 2022

Abstract: The predictive maintenance of CMMS system by SynapseCore is implemented to carry out maintenance activities and subsequently eases the burden towards maintenance personnel. This research's main objective is to apply linear regression and naïve bayes with parameters consisting of flowrate and different vibrations. The research uses the R programming language in RStudio to train the algorithm and follows the CRISP-DM process methodology. Results would show a graph comparing the predictions for both implemented algorithms including an evaluation of each algorithm's performance. Through this study, the company would get a better insight on the implementation of different algorithms and how they affect the performance and prediction.

Keywords: Machine learning, linear regression, naïve bayes, predictive maintenance

1. Introduction

A model is a mathematical representation of a process or system that will be analysed or automated in a precise way. Predictive maintenance is the frequent monitoring of mechanical condition, operating efficiency and other measures that would provide the information needed to minimize the number and cost of unplanned disruption by machine failures. In addition, it also enhances the productivity, product quality and overall efficiency of manufacturing [1]. Manufacturing is the production of items though labour, machinery, tools and biological or chemical processing in a large scale or to produce other items. A company is an organisation that sells goods or services, which in this case, would be items produced. Machine learning is a part of Artificial Intelligence and is the analysis of computer algorithms which can be taught by training it using sample data. It would either forecast or make decisions without programmed explicitly. Computerised Maintenance Management System (CMMS) is a software that maintains a system of information of the maintenance operations of an organisation that would also improve workflows and provides valuable insights on the operations [2]. Other things CMMS provide includes tracking the motion of spare parts, faster reports by operators and former information needed to develop preventive maintenance schedules[15].

Currently, the company is in the development of creating a CMMS called SynapseCore. It includes four modules, which are human resources, inventory, project management and maintenance. This

*Corresponding author: azizulr@uthm.edu.my

research would be focusing on the maintenance module, where predictive maintenance would be implemented.

The problem statement of this research is that the company is trying to create a balanced workload for the maintenance personnel and has a hard time tracking the predictive maintenance schedule for the machines. With this, a mathematical model is needed to predict the maintenance activities in the company where two models will be compared to determine the best algorithm to be used in predictive maintenance.

The objectives of this research are:

- To apply linear regression in naïve Bayes in predictive maintenance.
- To identify the appropriate model for the given dataset based on linear regression and naïve bayes.
- To compare the performances between linear regression and naïve bayes models.

The reason why this research is conducted is to figure out the best algorithm to be used in predictive maintenance between linear regression and naïve Bayes. Through this research, the company will have a more in-depth realisation on the pros and cons of each algorithm and to finally decide which of the two would be most suitable based on the business needs. Since predictive maintenance is a part of a whole system, which is SynapseCore, the users of this system would also be using predictive maintenance and expect an accuracy of forecasts.

The reason why linear regression and naïve bayes is because linear regression takes a regression approach and naïve Bayes takes a classification approach. The difference in approaches would create a difference in how the prediction would take place. The comparison would determine which is a better method to be used in predictive maintenance in terms of accuracy, complexity, and others.

Thus, this research would compare linear regression and naive Bayes in predictive maintenance. It would predict the amount of time left before the machine would fail to work. This research would compare which models would be most suitable in predictive maintenance in terms of accuracy, effectiveness, speed, and complexity.

This article is organized into five sections. The first part is an introduction describing the context of the project. The second section describes the analysis of the relevant work. The third section contains the methodology which will also be explained. The implementation of research will be described in the fourth section and consequently, a conclusion with instructions for future employment will be described in the last section.

2. Related Work

2.1 Machine Learning

Machine learning is a computational method that uses experience to improve performances or to predict accurately [3]. In a way, it is also considered a subset field of Artificial Intelligence since these algorithms are like building blocks in order for computers to react and behave more intelligently [5]. Experience in this case refers to previous information made available to the learner, or in this case the machine, usually in the form of electronic data collected for analysis. For every case, the quality and size of data provided or obtained is important to ensure the success of predictions by the machine.

Machine learning includes designing efficient and accurate forecasting algorithms. The notion of sample complexity is needed to evaluate the sample size for the algorithm to learn various concepts. Generally, the complexity of the classes and size of training sample would guarantee an accurate algorithm [3].

Machine learning be divided into three learning approaches, namely supervised learning, unsupervised learning, and reinforced learning. Supervised learning is when the machine is given labelled data and their predicted outputs. Unsupervised learning would use algorithms to analyse and cluster unlabelled datasets and find hidden patterns without human intervention. Reinforced learning is a behavioural machine learning model that is similar to supervised learning but differs since reinforced learning does not use a training dataset. Instead, it would learn through trial and error [4]. This research would be using the supervised learning approach.

2.2 Supervised Learning

As stated in the section above, this research would use supervised learning, where a set of input variables would be used to predict a response based on the inputs. The two main approaches under supervised learning are regression and classification. Supervised learning is most used to train neural networks and decision trees. For neural networks, classification is used to determine errors in the network and alter it to minimize the error. In decision trees, the classification will be used to determine the attributes that will contribute to solving the classification puzzle. Figure 1 shows the general classification architecture for classification.

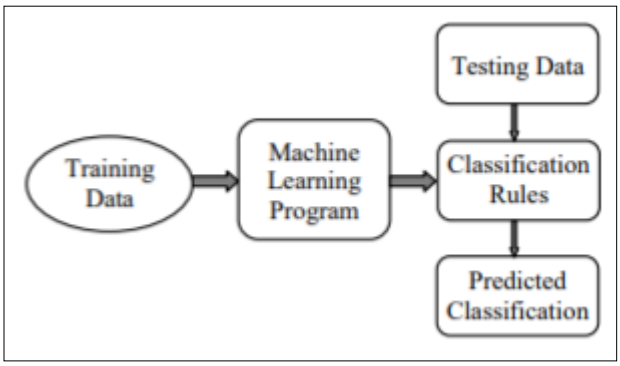

Figure 1: Classification Architecture [5]

Regression is the study of dependence, where it is used to answer questions revolving many applications. It is used for two theories, the first which is usually for forecasting and prediction and the second is used to determine casual relations between dependent and independent variables. This approach would only show the relationship between the dependent variable and the dataset collection of different variables [6].

There are two kinds of regression models, which are simple linear regression and multiple regression. Simple linear regression is a model with one regressor, x, with a response, y, in a straight line. The model is as shown in equation 1.

$$y = \beta_0 + \beta_1 x + \varepsilon \quad Eq.1$$

where the intercept and slope, which are β_0 and β_1 respectively, are unknown constants and ε is a random error element [7]. The error is assumed to be unrelated and thus, the value of one error is not dependable to the value of any other.

However, multiple linear regression will be implemented for this research. Here, more than one regressor variable will be involved. A mathematical model that can describe this would be as stated in Equation 2.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad Eq.2$$

As in simple linear regression, the β is the unknown parameters to be estimated. When $p = 1$, it means that X only has one value as stated in simple linear regression [8].

The reason why linear regression is chosen for this research is because it is a statistical model that can include more than one parameter to be able to predict the results accurately. Another reason is because the company prefers to use linear regression since linear regression is commonly used in cases of predictive maintenance.

Naïve Bayes is an algorithm that falls under the classification approach of supervised learning that is known for its simplicity [9]. It is a form of Bayesian Network Classifier that is based on the Bayes' rule which is as stated in Equation 3.

$$P(y|\boldsymbol{x}) = \frac{P(y)P(\boldsymbol{x}|y)}{P(\boldsymbol{x}|y)} \quad Eq.3$$

along with the assumption that the parameters are independent given the class. For an attribute-value data, this can be shown in mathematical format as shown in Equation 4.

$$P(\boldsymbol{x}|y) = \prod_{i=1}^{n} P(x_i|y) \quad Eq.4$$

The x_i in equation 4 refers to the value of the i^{th} attribute in x and n is the number of attributes. The next equation would show where k is the number of classes and c_i is the i^{th} class.

$$P(x) = \prod_{i=1}^{k} P(c_i)\, P(\boldsymbol{x}|c_i) \quad Eq.5$$

From this, the first formula in this section can be determined through the normalisation of numerators of the right-hand-side equation [10]. Naïve Bayes is chosen in this research because it differs from the statistical model previously chosen, which is linear regression. Since naïve Bayes is a classifier, the method of prediction would be different compared to linear regression. There should be a difference in results when comparing both algorithms when the results are chosen. Since search is not used, it has low variance, but the cost of this is due to high bias[14].

2.3 Comparison of Existing Algorithms

The implementation of algorithms from the study of existing algorithms in other applications has given an example of the differences of how the algorithms would be implemented. The papers that were studied in this literature review for linear regression are Development of Predictive Maintenance Interface using Multiple Linear Regression and Predictive Maintenance Decision Using Statistical Linear Regression and Kernel Methods. For naïve Bayes, the papers studied are Application of Naïve Bayes Classifier Theorem in Detecting Induction Motor Bearing Failure and Intelligent Predictive Maintenance Model for Rolling Components of a Machine based on Speed and Vibration.

Table 1: Comparison of existing linear regression applications

	Linear regression	
Research paper	Development of Predictive Maintenance Interface using Multiple Linear Regression	Predictive Maintenance Decision Using Statistical Linear Regression and Kernel Methods
Total parameters used	14 parameters	2 parameters
Accuracy	Uses RMSE	Uses RMSE

Table 1 shows the comparisons of proposed algorithms for linear regression, which includes the name of the research paper, the total parameters used in each paper as well as how the accuracy is calculated.

Table 2: Comparison of existing naïve bayes applications

	Naïve bayes	
Research paper	Applications of Naïve Bayes Classifier Theorem in Detective Induction Motor Bearing Failure	Intelligent Predictive Maintenance Model for Rolling Components of a Machine based on Speed and Vibration
Total parameters used	2 parameters	2 parameters
Accuracy	Confusion matrix	Confusion matrix

Table 2 shows the comparison of proposed algorithms for naïve bayes, which includes the name of the research papers, the total parameters used as well as how the accuracy is calculated in each paper. In short, both algorithms have different approaches, complexity, and methods to run. However, it cannot be denied that both methods can be used to predict machine failure. Based on the papers reviewed, the best algorithm to be used is linear regression since the accuracy of the model is higher since more parameters are used to predict the dependant variable.

3. Research Methodology

The research methodology used in this research is Cross-Industry Standard Process for Data Mining (CRISP-DM). The phases conducted in this research would be business understanding, data understanding, data preparation, modelling, evaluation, and finally deployment. These phases will be the foundation of how this research is conducted and ensures that the research will be completed by the timeline provided.

Table 3: CRISP-DM process model and their tasks and outputs

Phase	Task	Output
Business Understanding	• Determine business objectives • Situation assessment • Determine machine learning goal • Produce project plan	• Background • Business objectives • Business success criteria • Machine learning goals • Project plan • Initial assessment of tools and techniques
Data Understanding	• Collect initial data • Describe data • Explore data • Verify data quality	• Data description given by company representative
Data Preparation	• Data set • Select data • Clean data • Construct data • Integrate data • Format data	• Data set description • Derived attributes given by company • Cleaned dataset from company

Table 3: (cont)

Phase	Task	Output
Modelling	• Select modelling technique • Generate test design • Build model • Assess model	• Implemented modelling technique • Modelling assumptions • Models • Model description • Model assessment
Evaluation	• Evaluate results • Review process • Determine next steps	• Assessment of machine learning results • Approved models • List of possible actions decision
Deployment	• Plan deployment • Plan monitoring and maintenance • Produce final report • Review project	• Final report • Final presentation

Table 3 shows the phases used in this experiment, the tasks involved, and the output produced in each phase.

4. Research Results and Discussion

This section discusses about the results of the conducted research.

4.1 Experiment Design

Designing the experiment phases of the experiment is very important to make the experiment successful. When the experiment fails, the flowchart can help determine which part needs to be changed or re-done. Figure 2 shows the flowchart of the experiment phases. The dataset was given by the company, which will then be pre-processed to interpret the data and have a clearer view of what data is available or missing. Then, the data would need to be normalised. From there, two methods of machine learning will be used, with Method 1 using linear regression and Method 2 using Naïve Bayes. The dataset will be using an 8:2 ratio with 80% of the dataset used to train the dataset and 20% used to test the data. Then, the prediction for both methods will be evaluated, especially in terms of accuracy.

Figure 2 shows the flowchart of how the research is conducted from start to finish.

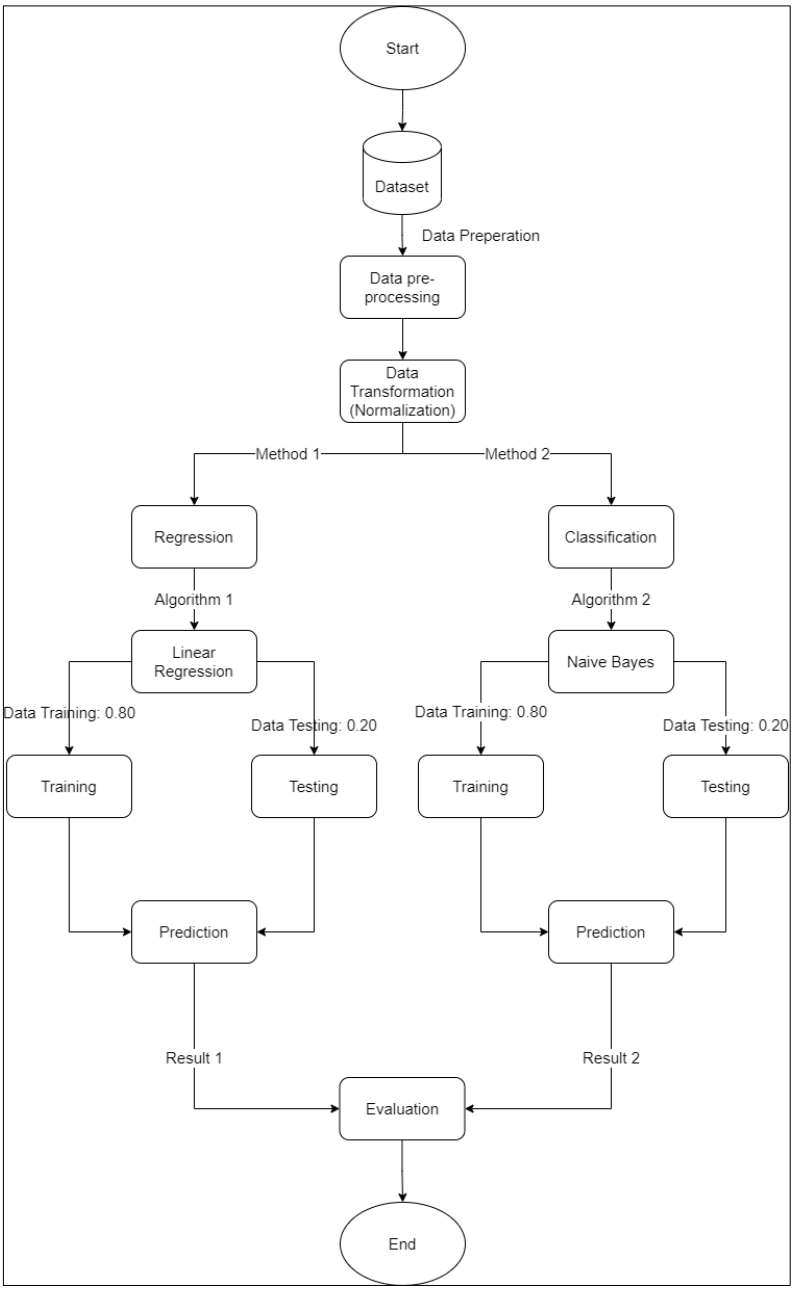

Figure 2: Flowchart of the experiment phases

For this research, RStudio is used alongside the programming language R when conducting this experiment. Python and R are frequently used in machine learning, but in this experiment R language is used.

Table 4: Data Information

No	Dataset	Number of data	Number of the attribute
1	Sheet1	527,041	4

Table 4 shows the data information of the dataset.

Table 5: Data Split

No	Dataset	Training (80%)		Testing (20%)	
		Linear Regression	Naïve Bayes	Linear Regression	Naïve Bayes
1	Sheet1	221,374	206,283	36,479	51,570

Table 5 shows the data split when training the model. The data is based on clean data that will be used to evaluate the model.

4.2 Parameter and Testing Methods

This section discusses about the parameters used during the experiment using linear regression and naïve bayes. Parameters, or sometimes called attributes, represents the characteristic of a data object [12]. In this research, the parameter that is used is flowrate. Testing methods refers to the evaluation of the models. For this project, different testing methods are used to evaluate the accuracy of the different models. For linear regression, the root mean square error (RMSE) is used and for naïve bayes, confusion matrix is used [11].

In linear regression, the RMSE is used to determine the accuracy of the model. The formula to determine the RMSE is as stated in Equation 6.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i-1}^{m}(X_i - Y_i)^2} \quad Eq.6$$

where X_i is the predicted i-th value and Y_i is the actual value i-th value [11].

For naïve bayes, confusion matrix is used to calculate the accuracy of the model. An example of how a confusion matrix looks like is shown in Table 6. The 'true' or 'false' values determines if a class is correctly predicted or not while the 'positive' or 'negative' shows the prediction of the class of whether the predicted class is healthy or unhealthy [13].

Table 6: Confusion matrix [13]

	Actual value	
Predicted value	True negative (TN)	False positive (FP)
	False negative (FN)	True positive (TP)

Table 6 shows the format of the confusion matrix and what each row and column represent. The formula in equation 7 shows how the accuracy of the model is obtained.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad Eq.7$$

The accuracy is obtained from the confusion matrix as shown in table 5.

4.3 Prediction and Evaluation

This section explains about the prediction results as well as the evaluation of each model.

4.3.1 Linear regression

This sub-section discusses about the predictions and evaluations on the linear regression model. From the model, the testing would produce a graph consisting of the plotted points as well as a linear line of the model created. The graph shows the relationship of the pressure and the flowrate.

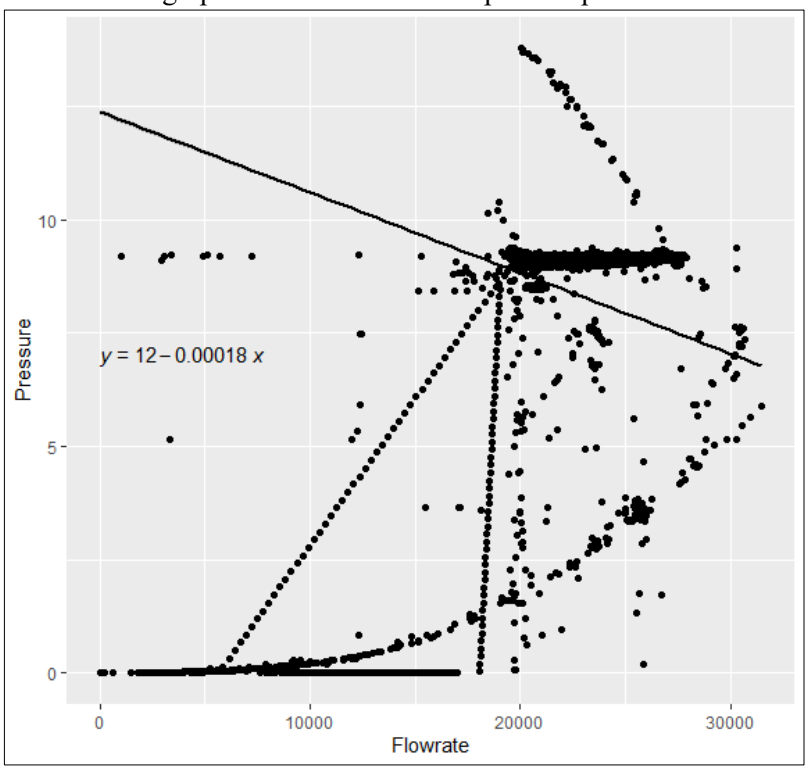

Figure 3: Scatterplot with linear line and equation

Figure 3 shows the scatterplot of pressure against flowrate, including the linear line and the equation of the linear line in the figure. The linear line has an equation as seen in Equation 8. From this equation, y represents the pressure while x represents the flowrate. The variable x also represents the dependent variable and y represents the independent variable.

$$y = 12 - 0.00018x \quad Eq.8$$

From the graph and equation obtained, we are able to determine the relationship between pressure and flowrate, which is the higher the flowrate, the lower the value of pressure. An example of how this equation can be used is if the value of flowrate is 20,000 Nm^3/h and is inserted into the equation, the pressure will have a value of 8.4 Pascals. The health of the machine can then be roughly be determined from the pressure.

The accuracy of the model can be determined through a few ways but in this case, the accuracy will be determined by dividing the RMSE with the mean of the dependant variable and multiplying it by 100. In this case, the RMSE obtained is 1.037. This means that the average distance between the observed data values and the predicted data values is 1.037.

4.3.2 Naïve bayes

This sub-section discusses about the evaluations and predictions for naïve bayes. Figure 4 shows the confusion matrix as well as the accuracy of the model. The 0 in the confusion matrix represents false values and the 1 value represents the true values.

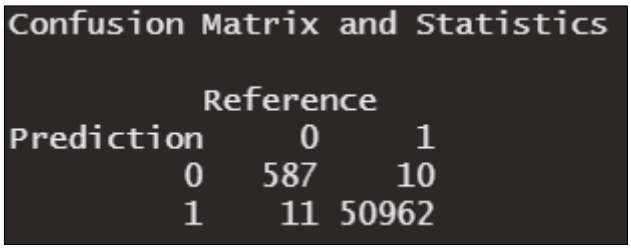

Figure 4: Confusion matrix for Naïve bayes

From Figure 4, it is shown that the model has correctly predicted that the machine is in healthy range 50962 times and has incorrectly predicted that the machine is in healthy range 10 times. However, the model has correctly predicted that the model is in unhealthy range 587 times and incorrectly predicted that the machine was unhealthy 11 times. Using equation 7, the accuracy of this model is 98.84%, making this model very accurate to predict machine health with low false positive and false negative values.

5. Conclusion

To conclude, this section would summarise the research conducted for a model in predictive maintenance for a company. The algorithms used were linear regression and naïve bayes, where linear regression would predict the value of pressure while naïve bayes would predict the health of the machine. Both algorithms use flowrate as a parameter.

From the research, linear regression shows an RMSE value of 1.037 while naïve bayes shows an accuracy of 98.84%. Linear regression uses an RMSE approach to calculate the accuracy of the model while naïve bayes uses confusion matrix.

The two models would predict a different type of output, where linear regression would predict the value of pressure while naïve bayes would predict whether the machine is in healthy or unhealthy range. Thus, for predictive maintenance, it would be more useful for maintenance personnel to know the value of pressure instead of a label. With the predicted value given from the model, it would be more helpful to maintenance personnel to prepare for when the machine should be undergoing maintenance. Naïve bayes is not as suitable because there would be times where the predictions may be inaccurate which might cause workers to call for maintenance when the machine is still relatively healthy. Therefore, linear regression is a more suitable model to be used in predictive maintenance.

To improve on this research, it is suggested that the research is conducted with a dataset with more attributes so that parameter setting can be implemented and compared to discuss which attributes would be most effective in influencing the predictor. An additional improvement that can be made would be predict the specific location of where in the machine is needed for maintenance.

Acknowledgment

The authors would like to thank the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM) and Synapse Innovation Sdn Bhd through Sepadan RE-SIP (vot M073) for its support.

Appendix A

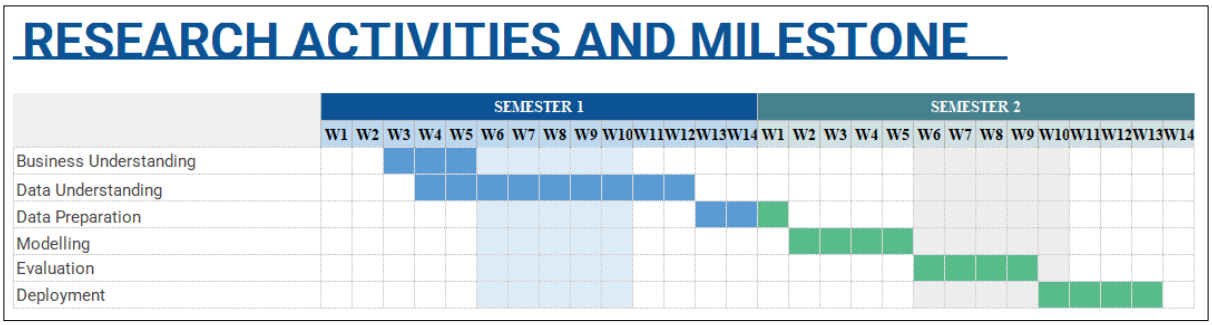

Figure 5: Gantt chart of the research

	Pressure	Flowrate	Vibration X	Vibration Y	Vibration Z
2	-0.04	2526.691162	0.509111205	0.620500038	-0.29055
3	-0.04	2532.03833	0.509218076	0.620500038	-0.29055
4	-0.04	2537.385498	0.509324948	0.620500038	-0.29055
5	-0.04	2542.732666	0.509431819	0.620500038	-0.29056
6	-0.04	2548.07959	0.50953869	0.620500038	-0.29056
7	-0.04	2553.426758	0.509645561	0.620500038	-0.29056
8	-0.04	2558.773926	0.509752432	0.620500038	-0.29056
9	-0.04	2564.121094	0.509859303	0.620500038	-0.29056
10	-0.04	2569.468018	0.509966145	0.620500038	-0.29056
11	-0.04	2574.815186	0.510073016	0.620500038	-0.29056
12	-0.04	2580.162354	0.510179887	0.620500038	-0.29056
13	-0.04	2585.509521	0.510286758	0.620500038	-0.29057
14	-0.04	2590.856689	0.510393629	0.620500038	-0.29057
15	-0.04	2596.203613	0.5105005	0.620500038	-0.29057

Figure 6: Dataset Sample

References

[1] R. K. Mobley, *An introduction to predictive maintenance*. Amsterdam: Butterworth-Heinemann, 2002. [E-book] Available: Google Books.

[2] Dudley, S., 2021. *What is CMMS? Absolutely everything you need to know*. [online] IBM Business Operations Blog. Available at: https://www.ibm.com/blogs/internet-of-things/iot-history-cmms/

[3] Rastegari, A., & Mobin, M. (2016, January). Maintenance decision making, supported by computerized maintenance management system. In 2016 annual reliability and maintainability symposium (RAMS) (pp. 1-8). IEEE.

[4] Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018. [E-book] Available: Google Books.

[5] "What is Machine Learning?", 15-July-2020. [Online]. Available: www.ibm.com/cloud/learn/machine-learning.

[6] I. Muhammad and Y. Zhu, "Supervised Machine Learning Approaches: A Survey," *ICTACT Journal on Soft Computing 5.3,* 2015.

[7] D. Maulud and A. Adnan M., "A Review on Linear Regression Comprehensive in Machine Learning," *Journal of Applied Science and Technology Trends 1.4,* 2020.

[8] D. C. Montgomery, E. A. Peck, G. G. Vining and G. G. Vining, "*Introduction to linear regression analysis"*, John Wiley & Sons, Incorporated, 2012. [E-book] Available: Google Books

[9] S. Weisberg, *"Applied linear regression"* Vol. 528. Johm Wiley & Sons, 2015. [E-book] Available: Google Books

[10] A. Bifet, *"Adaptive stream mining: Pattern learning and mining from evolving data streams"* IOS Press, Incorporated, 2010. [E-book] Available: Google Books

[11] Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. *Encyclopedia of machine learning*, *15*, 713-714.

[12] G. I. Webb, E. Keogh, and R. Miillulainen, *"Encyclopedia of machine learning",* 2010. [E-book] Available: Google Books

[13] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

[14] Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques, Waltham, MA. *Morgan Kaufman Publishers*, *10*, 978-1.

[15] Salmi, N., & Rustam, Z. (2019, June). Naïve Bayes classifier models for predicting the colon cancer. In *IOP Conference Series: Materials Science and Engineering* (Vol. 546, No. 5, p. 052068). IOP Publishing.