National College of Ireland

Project Submission Sheet – 2022/2023
Group G

Student Names:	Harshitha Poolakanda Somanna, Jose Geo Vattolly		
Student ID:	x22150366, x22139508		
Programme:	Master of Science in Data Analytics	**Year:**	2023-2024
Module:	Domain Application of Predictive Analytics (MSCDAD_JAN23C_1)		
Lecturer:	John Kelly		
Submission Due Date:	11-08-2023		
Project Title:	Predictive Modelling for Engine Maintenance and Repair Prediction in Automotive Vehicles		
Word Count:	2095		

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 references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Harshitha Poolakanda Somanna, Jose Geo Vattolly

Date: 11-08-2023

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

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

Predictive Modelling for Engine Maintenance and Repair Prediction in Automotive Vehicles

Harshitha Poolakanda Somanna
Student ID: *x22150366*
Email ID: x22150366@student.ncirl.ie
MSC in Data Analytics
National College of Ireland

Jose Geo Vattolly
Student ID: *x22139508*
Email ID: *x22139508@student.ncirl.ie*
MSC in Data Analytics
National College of Ireland

***Abstract*— The main of this analysis is to determine and apply the most effective predictive modeling method for engine fault prediction in automotive vehicles. The goal of this analysis is to forecast when an engine is likely to need maintenance and repair, which could assist mechanics and car owners in foreseeing potential problems before they get out of hand and result in better vehicle maintenance and engine durability. Random forest classification is applied after careful and extensive examination of previous analysis in this area. The model showed a accuracy of 67% after hypertuning with a F1 score of 77% for class 1 and 45% for class 0.**

Keywords—Engine condition, Random Forest, Faulty engine, Accuracy

I. Introduction

Predictive analytics models evaluate past data and discover patterns and trends to forecast future trends. Predictive maintenance in engines plays a critical role as it increases the efficiency and sustainability of vehicles. Traditional methods rely on predetermined schedules which leads to inefficiency, increased cost, and unexpected failures. Thus predictive modeling in engine maintenance is a critical area of research in the automotive industry as early detection helps in providing informed decisions to the vehicle owners and reduces the cost of maintaining them.

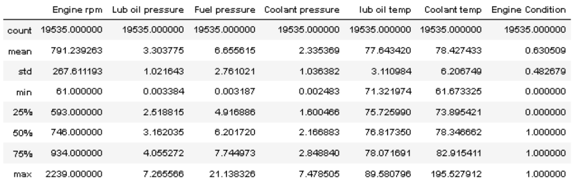
In automotive engineering, the innovation of new things day by day such as turbocharging, direct fuel injection, and hybrid power trains in the engine makes it difficult to monitor the engine condition manually. In this way, it is important to create a model that will be helpful to predict the condition of the engine by comparing the factors such as lube oil pressure, Fuel pressure, coolant pressure, lube oil temperature, and coolant temperature. Predictive modeling in automotive vehicles is important as it works in a complex environment under a variety of situations which causes deterioration in the engines. Early diagnosis of these situations is made possible by predictive modeling. Machine learning, data analytics and artificial intelligence are used in the automotive sector to anticipate engine issues from the sensor data available, thereby increasing reliability, cost savings and sustainability. This can make it easy to understand when the engine needs to be repaired on time in order to overcome engine failure in between the journey. To achieve this, a model has been created to investigate the trends and patterns in the engine dataset to predict the engine's condition based on the given measurements.

A. Dataset Descreption

In this research, a publicly available automotive vehicles engine health dataset[8] is used consisting of 19535 records and 7 columns(Fig 1), is carefully examined, processed and modeled using predictive analytics to obtain actionable business results. More importantly, the following problem statements are targeted to be answered in this study by the application of machine learning techniques.

a. *To what extent can predictive modeling be used to predict the engine's condition based on the given measurements.*

b. *To what extent can predictive modeling techniques be used to identify early warning signs of engine failures*



	Engine rpm	Lub oil pressure	Fuel pressure	Coolant pressure	lub oil temp	Coolant temp	Engine Condition
count	19535.000000	19535.000000	19535.000000	19535.000000	19535.000000	19535.000000	19535.000000
mean	791.239263	3.303775	6.655615	2.335369	77.643420	78.427433	0.630509
std	267.611193	1.021643	2.761021	1.036382	3.110984	6.206749	0.482679
min	61.000000	0.003384	0.003187	0.002483	71.321974	61.673325	0.000000
25%	593.000000	2.518815	4.916886	1.600466	75.725990	73.895421	0.000000
50%	746.000000	3.162035	6.201720	2.166883	76.817350	78.346662	1.000000
75%	934.000000	4.055272	7.744973	2.848840	78.071691	82.915411	1.000000
max	2239.000000	7.265566	21.138326	7.478505	89.580796	195.527912	1.000000

Fig 1. Dataset Description for Engine Data

II. Literature Review

Predictive maintenance can forecast problems in the system before they could happen if the equipment and status data is monitored. Machine learning algorithms helps analyze sensor data, logs from the equipments and detect patterns to predict when a component in a machine is likely to fail. Various predictive models using machine learning algorithms are built to analyze patterns and to make informed decision , and in this section, we are going to see how they are used in various applications to make better decision. In this research, a novel based approach where Fault Isolation

Observer(FIO) and Support vector machine(SVM) classifier is used to accurately identify sensor data and to evaluate healthy and faulty data for vehicle suspension systems which is obtained by quarter car test equipmenet[1]. The FIO generates residuals to capture the difference between actual sensor readings and expected values as determined by system modeling so as to distinguish sensor defects from other system fluctuations and disturbances. This is sent as an input to SVM classifier where data is labeled as healthy/ faulty and five-fold cross-validation is used to evaluate the performance of the model. The efficiency of this approach is tested experimentally in cases containing the fault types sine wave, sine sweep, signal loss, and random noise and even in the presence of unaccounted-for sensor bias and untrained fault types, the system shows respectable accuracy in recognizing sensor defects. Neural network methods such as CNN are valuable tools in prediction. To predict the remaining life of the lead-acid batteries, this study employed Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) where the objective is to determine when a battery might cross a failure threshold by analyzing the voltage degradation of batteries over time[2]. The data is preprocessed using the moving average technique to perform a smoothing procedure. The findings are verified on three distinct batteries (Batt A, Batt B, and Batt C) utilizing various input techniques and each method's model patterns, correlations, and forecast accuracy are analyzed. CNN outperformed the MLP method as they resulted in more precise predictions on estimating how long the batteries last, thereby enhancing the battery management systems in electric vehicles.

In this paper[3], the authors examine several machine learning methods and their use in predictive maintenance including decision trees, random forests, logistic regression, and MLP. The SECOM dataset is fitted with a variety of machine learning models, including k-NN, logistic regression, MLP, random forest, and decision trees. Techniques for dimensionality reduction, like as PCA and LDA, are used since there are so many features. The results of the investigation show that false discovery rate logistic regression performs well than the rest. This study[4] investigates on variable selection in battery prognostic case study, synthetic survival analysis resembling real vehicles and misfire detection using a forest based algorithm called Variable Depth Distribution(VDD). The VDD unveils a novel strategy by accessing even distribution of variables among the forest trees and giving an additional understanding of variable significance. VDD outperformed the traditional selection of variable selection methodologies like Variable Importance Measure (VIMP) and Minimal Depth(MD).

A study conducted by [5] used machine learning to create a predictive maintenance model for identifying laser welder equipment failures in continuous galvanizing lines (CGLs) at a Korean steel company. The Long short-term memory (LSTM) model for time series data has been combined with an auto-encoder (AE) for regular data. The model outperformed the current system by 18 hours, and successfully detected problems 27 hours before failure during operational tests, inspite of the precision-recall trade-off showing occasional duplicate predictions. This strategy, which has been verified as a pilot system, shows promise for digital transformation in the steel industry, improving sustainability and safety. Research by [6] compared various machine learning techniques in Predictive maintenance in the context of Industry 4.0. Linear regression, ANN, Gradient boosting, Random Forests, and SVM are the most commonly used method. RF is the most used method and is widely accepted and applied across several industrial equipment with SVM and ANN receiving good attention. To ensure the efficient operation of industrial equipment, this study[7] underlines the value of predictive maintenance in which vibrations using sensors attached to low-voltage motors are analyzed using a two-phased ML approach. Three machine learning algorithms, RF, SVM, and Backpropagation Neural Network (BPNN) are used in the first phase and SVM accurately classified the faulty data and was selected in the second phase to distinguish between normal and abnormal behavior which gave a pleasing but not quite convincing results as the historical data was missing in the second phase. However, this approach was effective as it could prevent accidents and stop before it could occur but suggested to test using other classification methods such as Logistic regression, decision trees and RF for the second phase. In this study, RF was used to predict the Remaining Useful Life(RLU) of wide-type industrial machines. Cross validation was performed on Linear regressor, Decision tree and random forest model where RF was chosen as it had a RMSE of 23.77 and smaller deviation. The more interesting aspect of RF here is that it may display the relative importance of each feature in producing accurate predictions.

After a thorough and detailed review of numerous research in this field, a well-informed choice was made about the methodology and modeling tools to be used. The Random Forest technique was chosen for the scope of this research even though several machine learning techniques, including KNN, neural networks, Support vector machines, logistic regression, linear regression, and decision trees, were taken into consideration. This choice was made based on its proven robust performance and ability to recognize important elements that influence the results of classification jobs. The Random Forest technique was deemed most suitable for the research objectives due to its robustness, interpretability, and capacity to extract meaningful information for the automotive industry.

III. Methodology

The implementation of this methodology follows the KDD approach[10] as illustrated in Fig 2 which includes data selection, pre processing the data, feature selection, building the model and interpretation of the results. Feature importance analysis is used to identify the impact on the model's prediction, resulting in a better model.

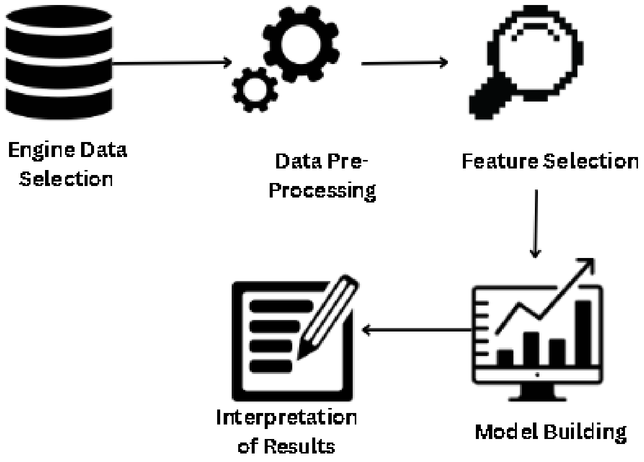


Fig 2. Research methodology

A. Data Pre-Processing

The engine_data.csv file was loaded via python code in jupyter notebook and the attributes were checked for duplicate values and analyzed for multicollinearilty. There were no duplicates found and multicollinearity was very low and hence all the variables were kept for further processing. The *Engine Condition* was converted to categorical variable for future analysis and prediction.

B. Exploratory Data Analysis

The data is checked for missing values, negative values and outliers in the data. The data is normalized using min-max scaling and the normalized data for engine RPM is as shown in Fig2.

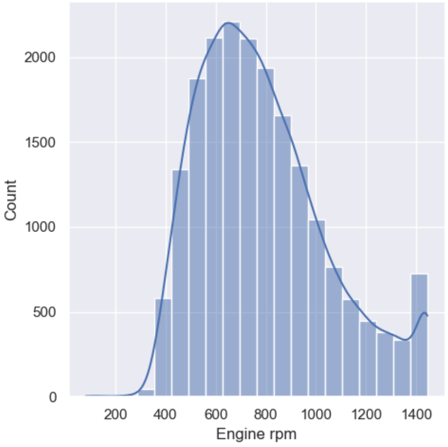


Fig 2. Normalized data for Engine RPM

C. Feature Selection

Principal component Analysis(PCA) is a dimensionality reduction technique[10] and is used to check for important components that are required for our prediction by reducing the number of less significant attributes. The analysis shows that all attributes have significance and hence all the variables were retained for further analysis. The scree plot as shown in Fig 3 determines the PCA components .

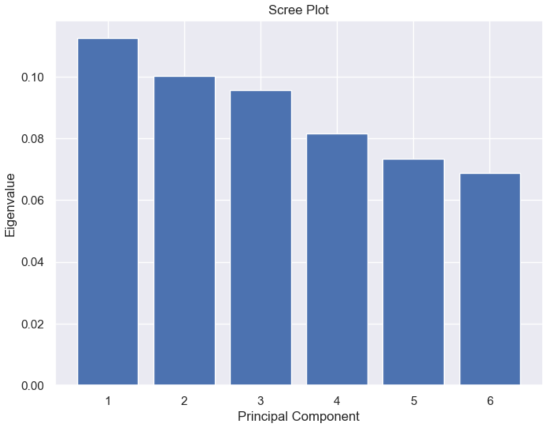


Fig 3. Scree Plot for engine_data Variables

D. Classification Modelling

Random forest classifier is implemented where the parameters are hypertuned with parameters like max_depth, max_features, min_samples_leaf, min_samples_split and n_estimators to get the optimum results.

IV. Results and COnclusion

A. Qualitative Findings

Accuracy and F1 score were taken as our evaluation metric to access the implemented random forest classifier model. F1 score is a useful evaluation metric in predicting the engine condition as healthy and unhealthy as the data is balanced and also the precision and recall measures are considered equally. This model had a training accuracy of 76.2% and test accuracy was 67.18% after hypertuning the parameters. Class 1 where the engine condition is healthy received an F1 score of 77% and class 0 where the engine condition is faulty received a F1 score of 45%.

Additionally, a confusion matrix and ROC curve was obtained to check the actual and predicted scores as shown in Fig 4. The confusion matrix evaluates the performance by taking the count of true positive, true negatives, false positive and false negative values.

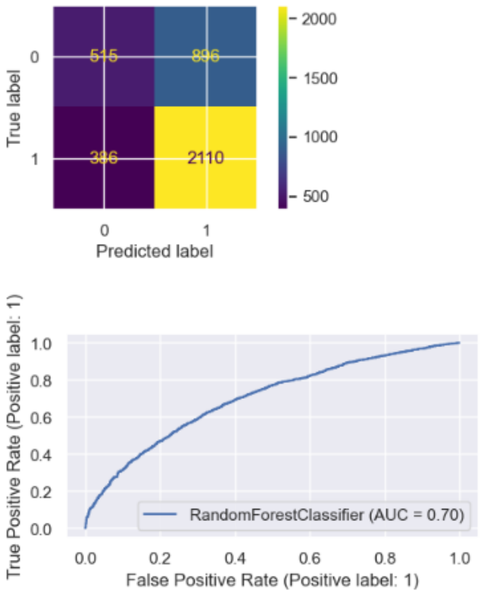


Fig 4. Confusion Matrix and ROC Curve

Feature importance was generated by the tree based random forest model as shown in Fig 5. It indicates which explanatory variables has the highest impact in predicting the faulty and healthy engines aiding in engine maintenance and repair prediction.

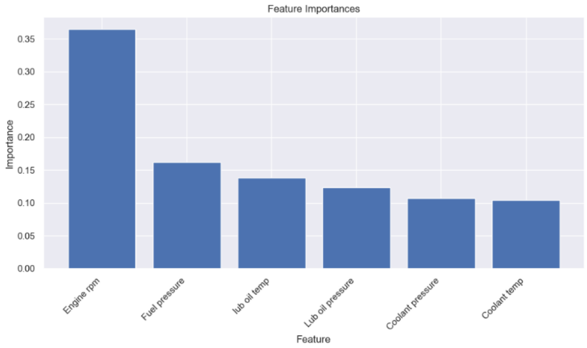


Fig 5. Feature Importance

B. *Qualitative Interpretation of the Findings*

The Quantitative findings from the model building process using Random Forest classifier was analyzed to interpret the business value and gain actionable insights to improve the sustainability of engines by predicting the condition of the engines which has practical applications in preventive maintenance and ensuring the proper functioning of automotive vehicles.

Looking at the feature importance shown in Fig 5, it can be inferred that Engine RPM is an important indicator in maintaining engine health followed by Fuel pressure, Lub oil temperature and Lub oil pressure. On the other hand, coolant pressure and coolant temperature has a relatively low impact on the engine health. From this qualitative inference, we can answer the problem statement in our study. It can be seen that engine RPM, Fuel pressure, Lub oil temperature and Lub oil pressure plays a major role in the health condition of the engine in automotive vehicles and therefore, more focus can be given towards these attributes in maintaining the health conditions of the engine which will help in engine maintenance and repair prediction.

References

[1] K. Jeong, S. B. Choi and H. Choi, "Sensor Fault Detection and Isolation Using a Support Vector Machine for Vehicle Suspension Systems," in IEEE Transactions on Vehicular Technology, vol. 69, no. 4, pp. 3852-3863, April 2020, doi: 10.1109/TVT.2020.2977353.

[2] Z. H. Wang, Horng Hendrick, G. J. Horng, H. T. Wu, and G. J. Jong, "A prediction method for voltage and lifetime of lead–acid battery by using machine learning," Energy Exploration & Exploitation, vol. 38, no. 1, pp. 310-329, 2020.

[3] A. Chazhoor, Y. Mounika, M.V.R. Sarobin, M.V. Sanjana, and R. Yasashvini, "Predictive maintenance using machine learning based classification models," in *IOP Conference Series: Materials Science and Engineering*, vol. 954, no. 1, p. 012001, IOP Publishing, October 2020.

[4] S. Voronov, D. Jung, and E. Frisk, "A forest-based algorithm for selecting informative variables using Variable Depth Distribution," Engineering Applications of Artificial Intelligence, vol. 97, p. 104073, 2021.

[5] J. S. Choi, S. W. Choi, and E. B. Lee, "Modeling of Predictive Maintenance Systems for Laser-Welders in Continuous Galvanizing Lines Based on Machine Learning with Welder Control Data," Sustainability, vol. 15, no. 9, p. 7676, 2023.

[6] Z. M. Çınar, A. Abdussalam Nuhu, Q. Zeeshan, O. Korhan, M. Asmael, and B. Safaei, "Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0," Sustainability, vol. 12, no. 19, p. 8211, 2020.

[7] M. Nikfar, J. Bitencourt, and K. Mykoniatis, "A two-phase machine learning approach for predictive maintenance of low voltage industrial motors," in *Procedia Computer Science*, vol. 200, pp. 111-120, 2022.

[8] https://www.kaggle.com/code/umzi01/automotive-vehicles-engine-health/notebook

[9] R. D. Drennan, “Principal components analysis,” 2009.

[10] D. Proti´c, “Review of kdd cup’99,nsl-kdd and kyoto 2006 + datasets,” *Vojnotehnicki glasnik*, vol.66,2018.