

Course Code: AER 850

Course Title: Introduction to Machine Learning

Semester: F2024

Instructor: Dr. Reza Faigehi

Submission: Project 1


Due Date: Sunday, October 20th, 2024

Title: Project 1

Section Number: 02

Submission Date: Sunday, October 20th, 2024

Submission By: Arjun Tripathi

Name	Student Number (XXXX99999)	Signature
Arjun Tripathi	XXXX21964	

By signing the above you attest that you have contributed to this submission and confirm that all work you contributed to this submission is your own work. Any suspicion of copying or plagiarism in this work will result in an investigation of Academic Misconduct and may result in a "0" on the work, and "F" in the course, or possibly more severe penalties, as well as a Disciplinary Notice on your academic record under the Student Code of Academic Conduct, which can be found online at www.torontomu.ca/senate/policies/

Table of Contents

Table of Contents.....	2
Introduction.....	3
Results.....	4
Discussion.....	9
Step 2: Data Visualization.....	9
Step 3: Correlation Analysis.....	9
Step 4: Classification Models.....	9
Step 5: Model Performance Analysis.....	10
Step 6: Stacked Model Performance Analysis.....	10
Conclusion.....	11
Appendix.....	12
Appendix A.1 - Project Code:.....	12

Introduction

In the Aerospace industry, Augmented Reality (AR) and Virtual Reality (VR) have become a significant technological development accelerating the research and development of complex system processes. Both AR and VR technologies are incomplete without reliable machine learning algorithms (ML). Through this project various ML models were developed and understood. The objective of the project was to develop a predictive machine learning model that accurately & precisely predicts the maintenance steps of an inverter used in a FlightMax Fill Motion Simulator. For such an application, classification-based ML models were further studied and utilized. These models included, Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine (SVM). A given dataset including 13 unique steps, each with precise X, Y, Z coordinates, was used to train and test these models. This dataset was also visualized and analyzed through plots including a 3D Scatter Plot and BoxPlot, and also the correlation was verified through a correlation heatmap and matrix. Furthermore, each model was evaluated through performance metrics including accuracy scores, precision scores, and F1-scores. Many difficulties were encountered in regards to finding the ideal parameters for each model while also ensuring that there is no overfitting. Cross-validation functions such as GridSearchCV and RandomizedSearchCV were also used to determine the hyperparameters that gave the best results for the model. In the end, the Decision Tree was deemed to be the best classification model. This is due to its outstanding performance metrics. This report will provide further discussion on each step in the machine learning model development process and illustrate the corresponding plots, heatmaps and matrices.

Results

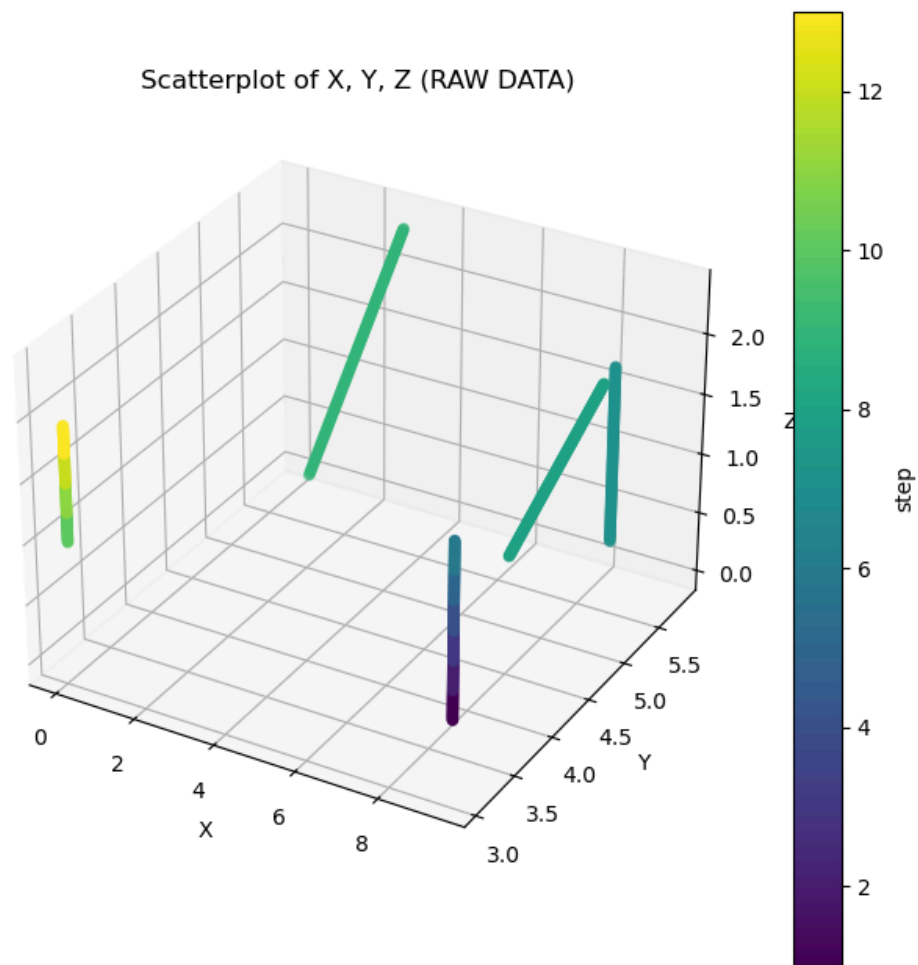


Figure 1.0: 3D Scatterplot of Raw Data

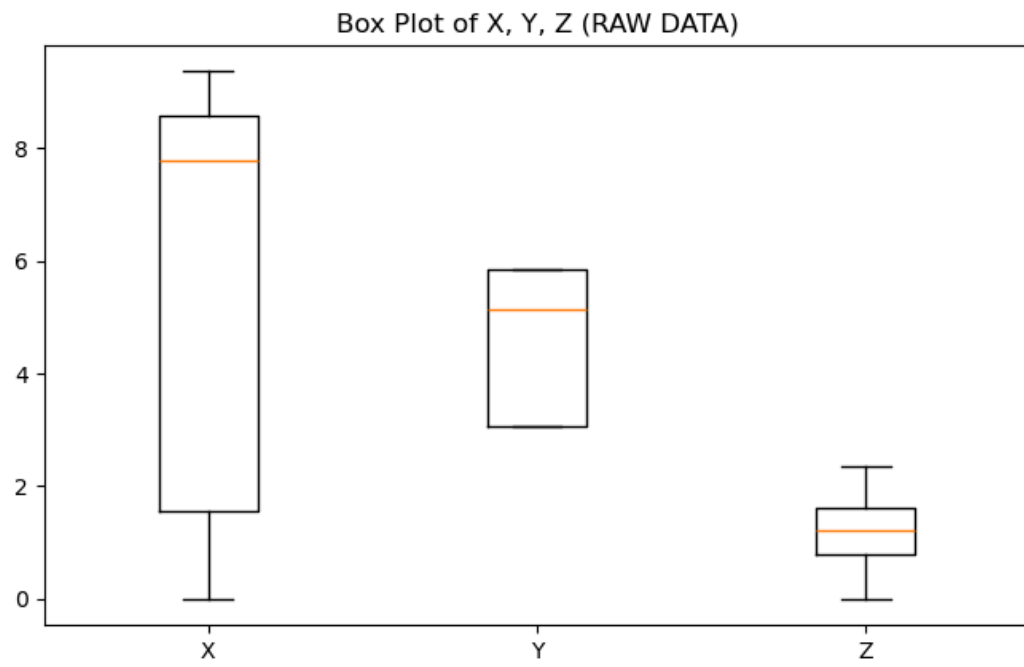


Figure 2.0: Box Plot of Raw Data

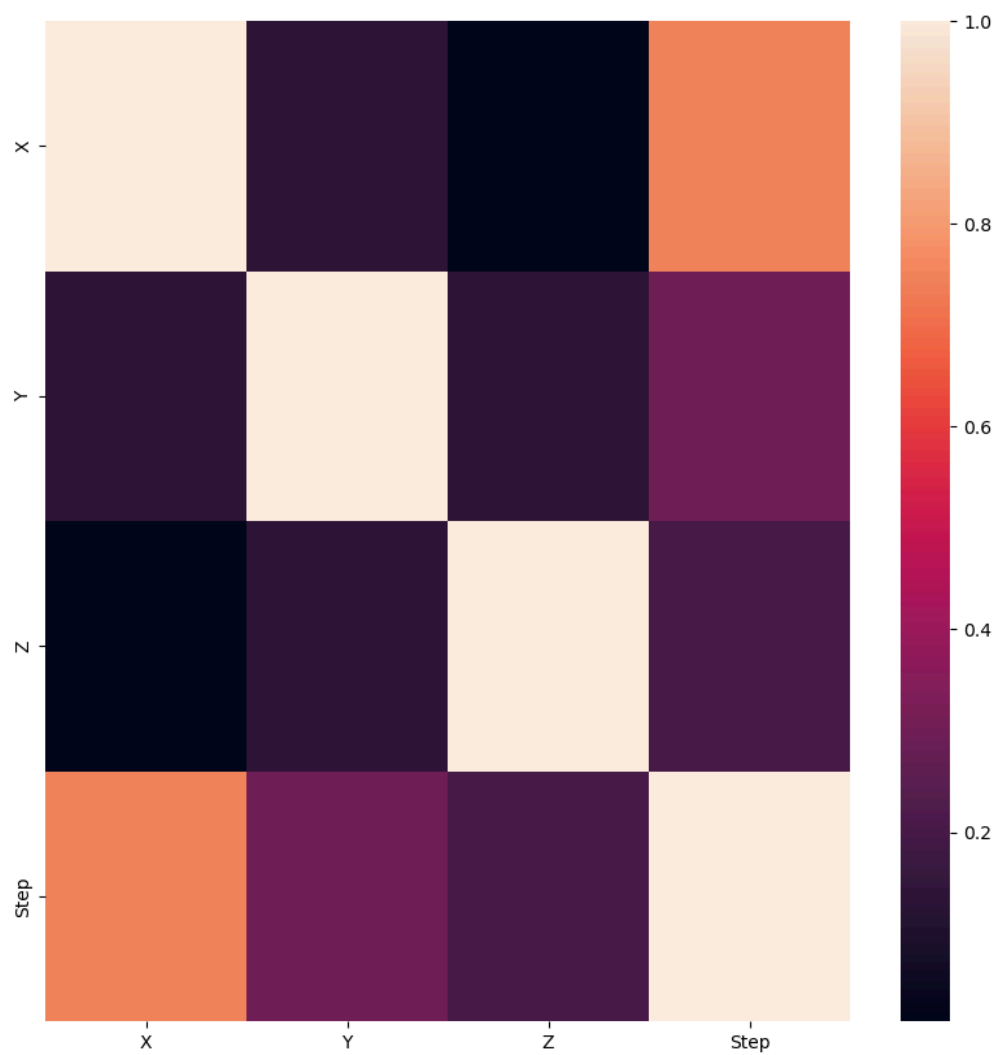


Figure 3.0: Heatmap of Correlation Matrix of Raw Data

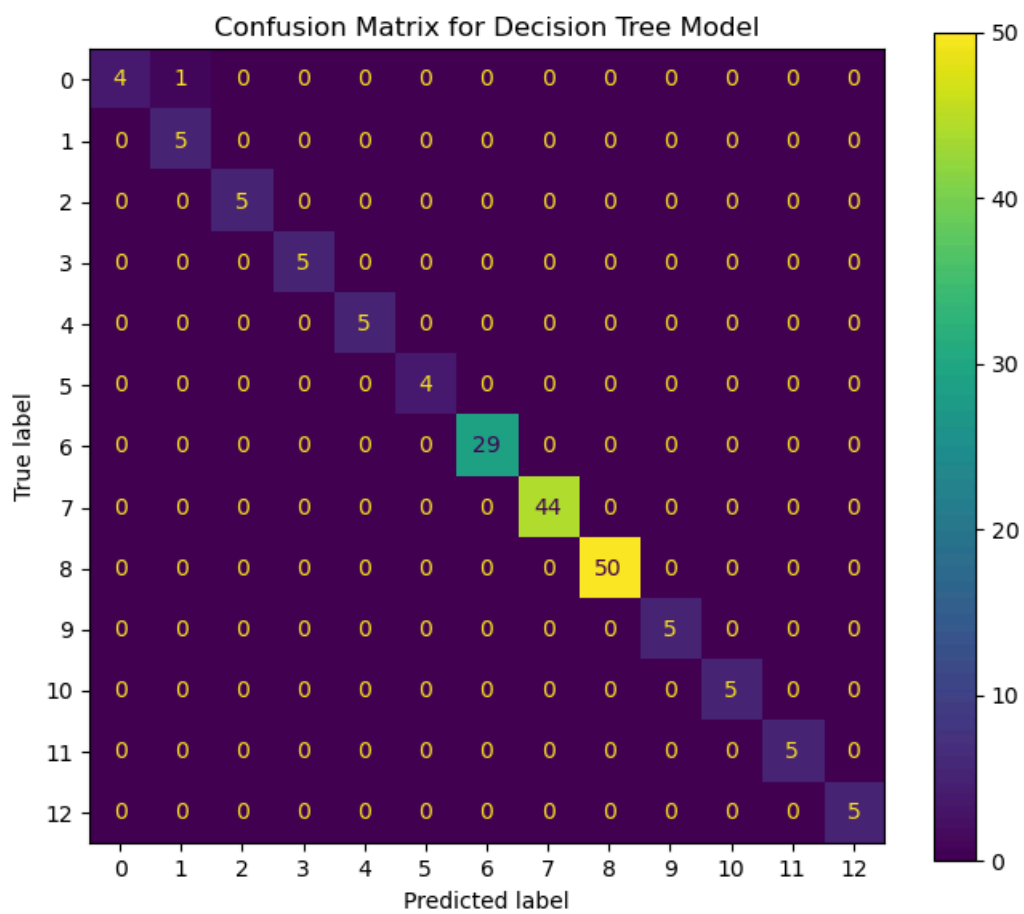


Figure 4.0: Confusion Matrix of Best Model Chosen

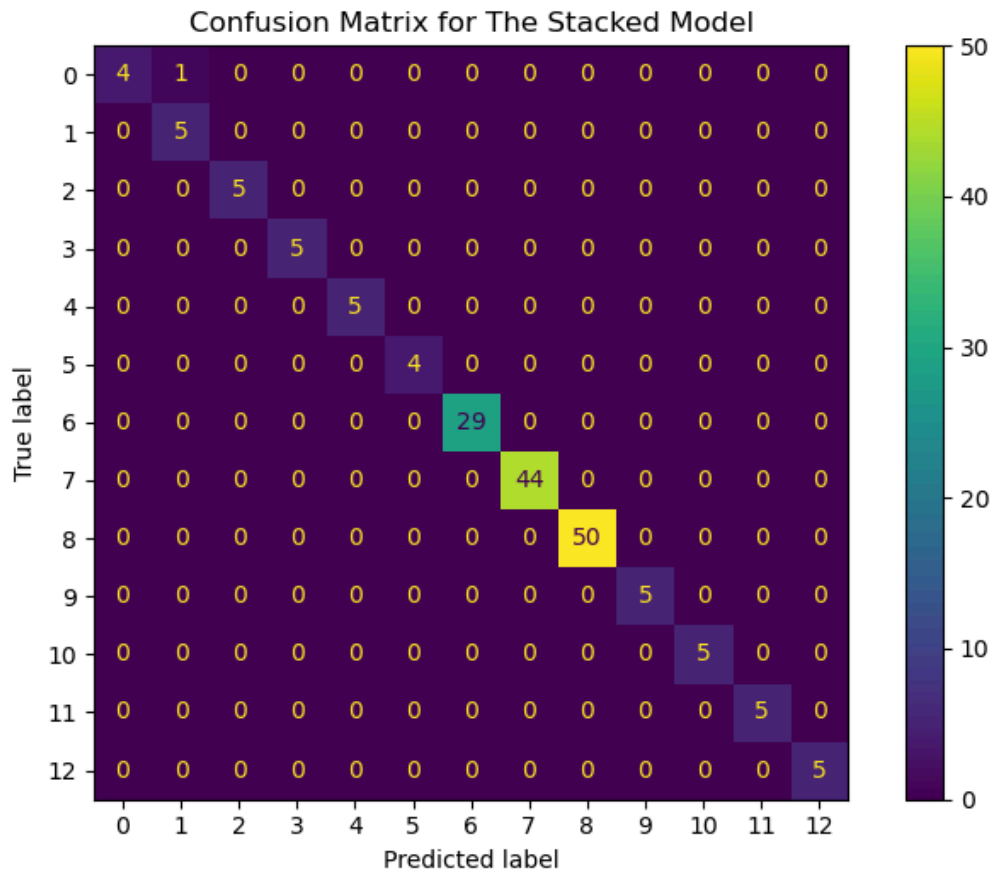


Figure 5.0: Confusion Matrix of Stacked Model

Discussion

The following discussion answers all the questions asked in Steps 2-6 in the project 1 outline posted on D2L.

Step 2: Data Visualization

The raw data that was provided was illustrated through a 3D Scatter Plot and a Box Plot, as seen in Figure 1.0 and 2.0 respectively. Statistical analysis was performed using the `df.describe` function. From the 3D Scatter Plot of the raw data, one can observe that the coordinates in the data are mainly of higher steps in the process such as steps 7 and above. Furthermore, there seems to be a clear relationship between the X,Y,Z values and the corresponding maintenance steps as coordinates linked to certain steps such as steps 7-10 seemed to be grouped in a certain specific space. In regards to the Box Plot, there is a high variability in X values compared to Y & Z as there is a visibly larger interquartile range depicted by its box for X. It can also be seen that there are no extreme outliers in the dataset.

Step 3: Correlation Analysis

The correlation between the feature variables, the coordinates and the target variables, the steps, can be observed through Figure 3.0. These correlations were found through a correlation matrix and depicted through a heatmap. Observing the heatmap, one can see that there is a strong positive correlation between the X coordinate and Steps suggesting that the X coordinate will be significant in the prediction of the step. On the other hand, there are lower correlations between the Y and Z coordinates and Step, indicating that they have a low impact on determining the accurate maintenance step. There were no significant correlations between the coordinates therefore a need for elimination of data was not required.

Step 4: Classification Models

For this project and its objective, 4 classification models were used: logistic regression, random forest, decision tree, and support vector machine (SVM). The dataset, after stratified sampling and standard scaling, was observed to be numerical & categorical, multi-class and have a class imbalance due to some steps having more coordinates than the others. Models had to be chosen keeping these characteristics in mind. Model development was started with logistic regression as the first baseline model. Logistic regression is simple to interpret and does not require many parameters to initialize. Secondly, the random forest model was chosen as it handles multi-featured datasets well, providing more accuracy. It also addresses the issue of overfitting that is present in the regular decision tree model as this includes multiple trees. Thirdly, the decision tree model was used as this is quite simple as well due to its singular structure and minimal preprocessing. Decision tree model is also known to be good at capturing non-linear data, which in this case is helpful as there were multiple coordinates and multiple steps which most likely created a complex relationship. Lastly, the support vector machine (SVM) was developed due to its ability to handle high-dimensional data such as the X,Y,Z

coordinates in the given dataset. Furthermore, the SVM also has the ability to classify based on linear and non-linear kernels which is beneficial as all types of data are processed at once.

Step 5: Model Performance Analysis

Three performance metrics were used in this project to evaluate the developed models, accuracy, precision and F1 score. Accuracy measures the number of correct predictions compared to the total number of predictions. Precision is the measure of the number of true positives in a prediction to the sum of true positives and false positives. In the context of this project, a higher precision in prediction would mean less unnecessary maintenance steps. F1 Score is the mean value of the precision and recall. It is extremely useful in imbalanced datasets such as this one. It aids in measuring how many actual positives the model has predicted. In this use-case, it is quite important to ensure that maintenance is done accurately and efficiently. Therefore, the F1 score would be the metric to prioritize when it comes to model evaluation in this use-case. Based on the F1 score, the decision tree model was the best model amongst four classification models developed. The confusion matrix in Figure 4.0, perfectly shows the evaluation of the decision tree.

Step 6: Stacked Model Performance Analysis

For the stacked model, consisting of the decision tree and logistic regression models, it was observed from Figure 5.0 that there was a little improvement compared to the individual models. This stacked model had limited effectiveness because the individual models were already quite accurate and precise. The logistic regression only captured linear data whereas the decision tree captured non-linear data, resulting in a full evaluation of both types of data. Therefore, when combining both models and harnessing their individual strengths, there was only a minimal improvement.

Conclusion

To conclude, the project aimed to develop a ML model that could accurately and precisely predict maintenance steps of an inverter used in aerospace applications, based on given coordinates. Through data visualization, various ML model research and development and performance metric evaluation, it was concluded that the decision tree algorithm would be the best model to perform such a predictive task. A stacked model was also developed and evaluated however, that showed minimal improvements as the models independently developed were already quite accurate and precise. Overall, this project was a great learning experience, allowing the further exploration of various machine learning algorithms and understanding the machine learning development pipeline.

Appendix

Appendix A.1 - Project Code:

GitHub Link: <https://github.com/Arjunt10/AER850Project1.git>