


RYERSON UNIVERSITY

DEPARTMENT OF AEROSPACE ENGINEERING
FACULTY OF ENGINEERING, ARCHITECTURE AND SCIENCE

Course Code:	AER850
Course Title:	Intro to Machine Learning
Instructor:	Dr. Faeighi
Report:	Project 1
Due Date:	16-10-2023

Section Number:	1
Submission Date:	16-10-2023

Name	Student Number	Signature
Akus Chhabra	xxxx70974	

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.ryerson.ca/senate/current/pol60.pdf.

Table of Contents

1.0 Introduction.....	1
2.0 Results & Discussion.....	1
2.1 Variable Identification.....	1
2.2 Split of Training and Testing Data.....	2
2.3 Data Visualization.....	2
2.4 Correlation Analysis.....	9
2.5 Classification Model Development.....	11
2.6 Model Performance Analysis.....	13
2.7 Model Evaluation.....	15
3.0 Conclusion.....	16

List of Figures

Figure 1:	Plot of the X-coordinates and the step number.	3
Figure 2:	Plot of the Y-coordinates and the step number.	4
Figure 3:	Plot of the Z-coordinates and the step numbers.	5
Figure 4:	Normal distribution of the X-coordinates.	7
Figure 5:	Normal distribution of the Y-coordinates.	8
Figure 6:	Normal distribution of the Z-coordinates.	9
Figure 7:	Heatmap displaying the correlation between each variable.	10
Figure 8:	Pairplots displaying the correlation between each variable.	11
Figure 9:	Confusion matrix for the Support Vector Machine model.	14

List of Tables

Table 1:	Median values of the X, Y, and Z coordinates.	5
Table 2:	Mean values of the X, Y, and Z coordinates.	6
Table 3:	Standard deviation of the X, Y, and Z coordinates.	6
Table 4:	Accuracy, precision, and F1 score for each developed model.	14
Table 5:	Predicted step number for the provided x, y, and z-coordinates.	15

1.0 Introduction

Machine learning is an integral tool in the modeling of data from simple linear data to abstract and complex data. Some of the major applications of machine learning include creating chat bots for streamlining user interface, credit card fraud detection for consumer security, and speech recognition to overcome communication barriers using computers. Despite the expansion of the machine learning field, the underlying fundamentals of grasping machine learning require an understanding of developing models to predict an output based on a set of inputs. The focus of this project is building a model to predict the maintenance step for the inverter of the FlightMax Fill Motion Simulator using its coordinate positions in the x, y, and z directions. The steps are broken down into 13 individual steps where the feature variables are the x, y, and z coordinates and the target variable is the maintenance step number the coordinates align to. An excel sheet with the coordinate and step data was provided to develop, train, and test the models. Ergo, the objective is to visualize and analyze the data, understand the correlations between the variables, develop classification models to predict the step number, analyze the performance of the models, and evaluate the model using a set of test coordinates.

2.0 Results & Discussion

2.1 Variable Identification

In order to build the machine learning models, the variables required must be identified for interpretation, classification, and to ultimately construct models that can predict the desired output. The variables identified as independent variables are the x, y and z coordinates and the target variable - dependent variable - is the maintenance step number. However, it must be noted that the coordinate data is numerical and continuous whereas the step numbers are numerical but discrete. This means that specific models such as linear regression would require the data to be rounded to the nearest whole value in order to finalize the output. Despite the features being continuous, the target variable behaves more as a classification rather than a discrete variable. Ergo, regression and classification models are both valid means of approaching the problem.

The advantage for this model development is that we know there is a finite limit to the number of possible outputs achievable with the model. These finite values are the step numbers from 1 to 13. The simplicity of the data allows for controllable predictions rather than having to explore the infinite space to predict numbers. In the case of a wide range of output values, more data would be required to develop a satisfactory model for making predictions. This process can be inconvenient as it is bounded by the complexity of the sample.

2.2 Split of Training and Testing Data

The splitting of the data into training and testing data is essential to ensure the model is not biased towards the data. This is termed as data snooping bias where there is statistical manipulation of the data to improve the overall results. Although splitting the data may result in output values that do not match the expected result, this is a desired result as it allows for the tuning of the models which can be used to improve the model's predictions. To split the data, sampling was used to create the test data which composes 20% of the total data values. The remaining 80% of the data was used to train the models. The sampling of the data was performed using the `.sample()` command built in python in conjunction with `random_state` set to 1 such that sampling is performed without replacement. This means that the sample data will remain the same every time to ensure there are no discrepancies in the model outputs. The disadvantage of this sampling method is that it may not select sufficient data to model certain steps or it may not include any data for those steps.

2.3 Data Visualization

The process of visualizing the data is important as it provides an understanding of how the data behaves in relation to the target variable. Although the coordinates are 3-dimensional, 3-d scatter plots do not provide any illustrative conclusions in comparison to 2-d scatter plots. Thus, intricate visualization methods can make data analysis cumbersome and complex. Figures 1, 2, and 3 display the relationship between each coordinate and the step number.

Figure 1 displays the step numbers for various x-coordinates. Steps 8 and 9 contain a set of values that are consistent for a set of x-coordinates. The general trend in the data is that as the x-coordinate increases, the step number decreases; however the middle steps do not vary as much with an increase in x-coordinate values. Since there exists an x-coordinate value for multiple step values, the x-coordinate is not a function of the step number.

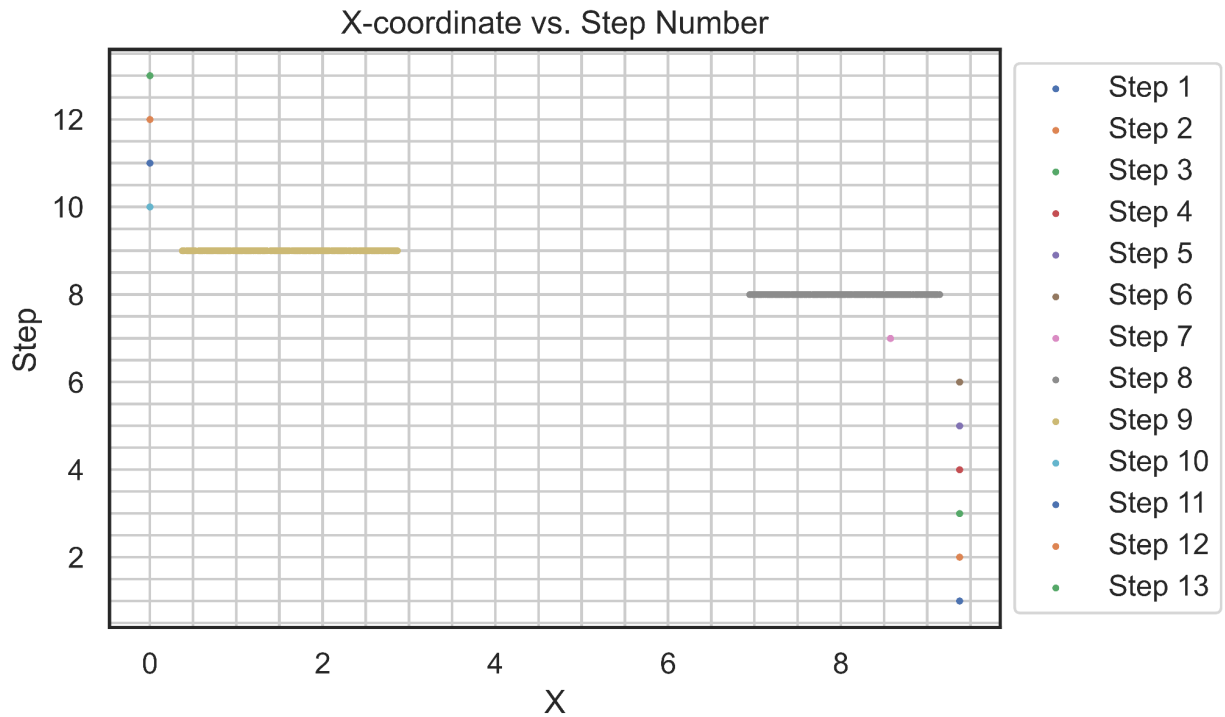


Figure 1: Plot of the X-coordinates and the step number.

Figure 2 displays the change in step number with respect to the y-coordinate value. The data displays a trend of one y-coordinate having multiple step values except for steps 7, 8, and 9. These steps were plotted for y-coordinates above 5. The issue with this graph is that it is difficult to illustrate a relationship between the y-coordinate and the step number as a single y-coordinate can output multiple step numbers. This means that the y-coordinate is not a function of the step number.

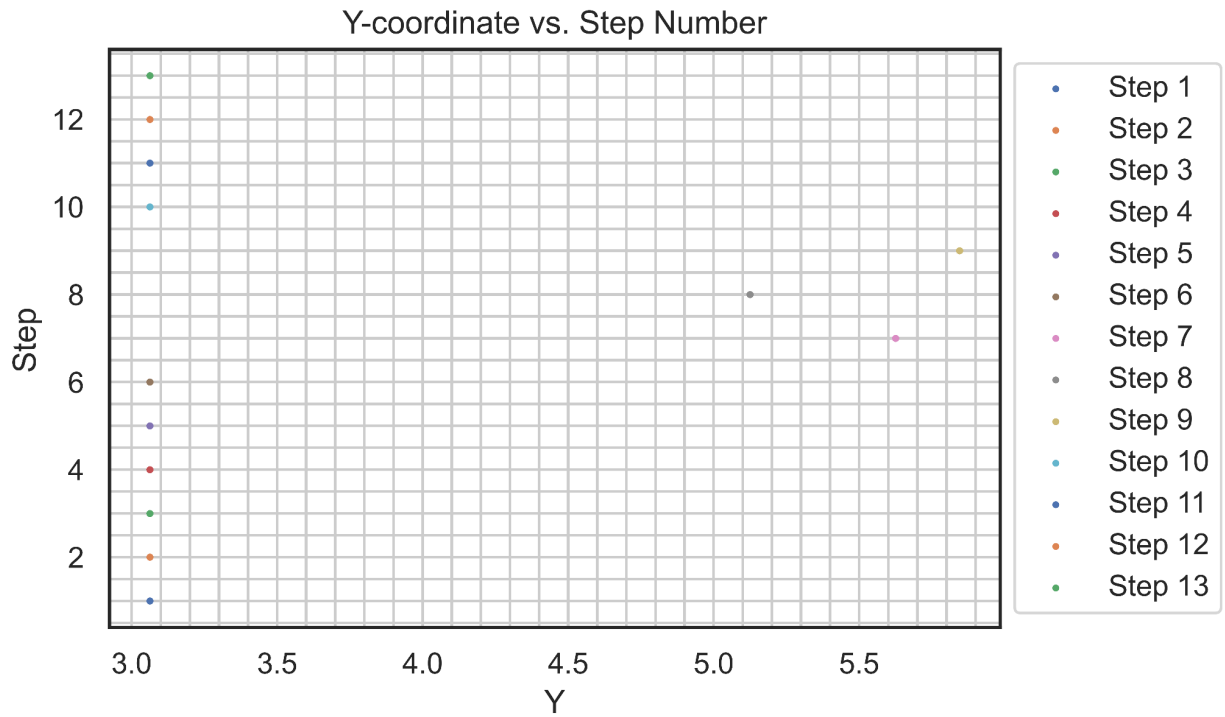


Figure 2: Plot of the Y-coordinates and the step number.

Figure 3 provides a plot of the step number with respect to the z-coordinate. The trend displays an increase in step number with z-coordinates up to step 6, however steps 7 to 9 span most of the entire y-coordinate domain. Although the z-coordinates for steps 10 to 13 increase, the z-coordinates for these steps shift back, overlapping the data from steps 3 to 6. Therefore, the step numbers are not a function of the z-coordinates.

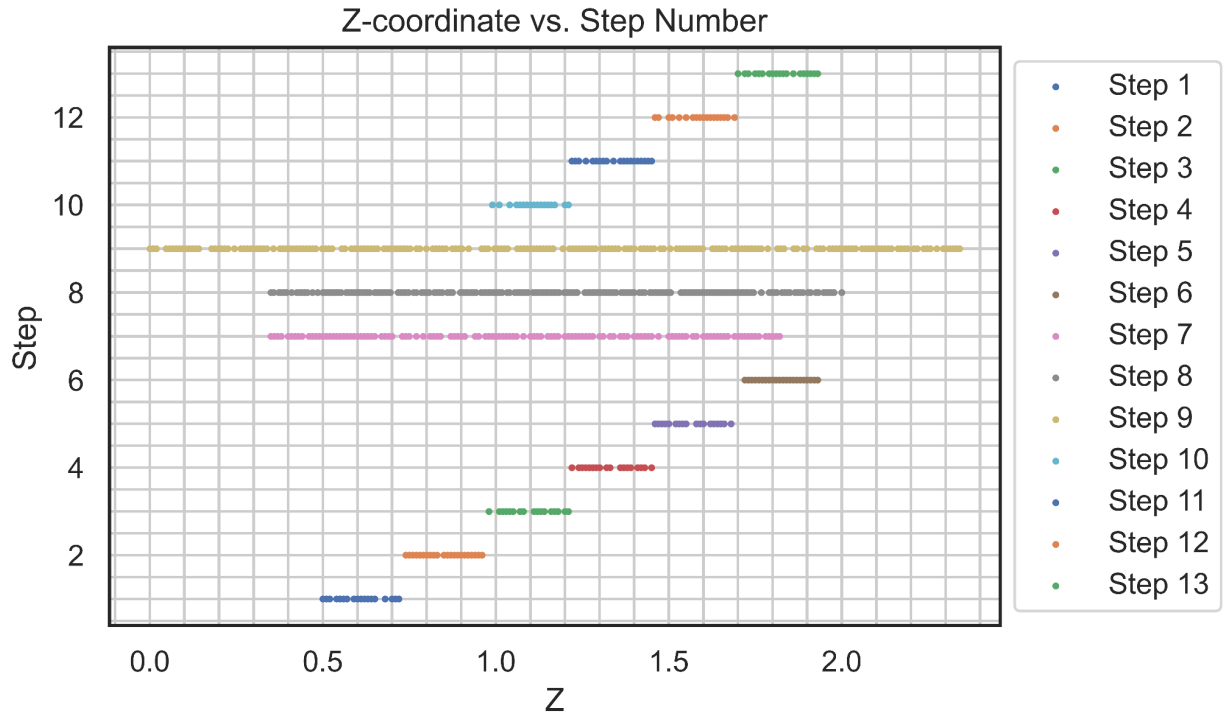


Figure 3: Plot of the Z-coordinates and the step numbers.

The data shown in table 1 represents the median values for each coordinate. The median number is the midpoint of a distribution of data. The importance of the median value is that it provides the middle value for the data set and shows how the data is distributed. Based on this table, the x-coordinates appear to have the highest median value indicating a more distributed set of values, followed by the y and z-coordinates.

Table 1: Median values of the X, Y, and Z coordinates.

Variable	Median Value
X	7.83
Y	5.125
Z	1.22

Table 2 displays the mean values for each coordinate. The mean represents an average of the data. Similar to median values, mean values are a measure of center however this value incorporates outliers in the data. The x-coordinate displays the highest mean value followed by the y and z-coordinates. This confirms that the x-coordinates have the most distribution.

Table 2: Mean values of the X, Y, and Z coordinates.

Variable	Mean Value
μ_X	5.644
μ_Y	4.853
μ_Z	1.196

Table 3 displays the standard deviation for the x, y, and z-coordinates. The standard deviation provides a measure of variation in the data. It determines an average of how far the data is from the mean value. The x-coordinate displays the highest standard deviation followed by the y, and z-coordinates verifying the x-coordinates hold the highest distribution among all the coordinates.

Table 3: Standard deviation of the X, Y, and Z coordinates.

Variable	Standard Deviation
σ_X	5.644
σ_Y	4.853
σ_Z	1.196

Figures 4 to 6 display the normal distribution for the x, y, and z-coordinates. The normal distribution, also known as a probability density function, displays the measures of central tendency along with the probability of data appearing in the dataset. The distribution is symmetrical about the mean value as it is normalized and the addition of standard deviations classifies the probability of data occurring. For instance, one standard deviation from the mean means there is a probability of 34.1% of the data existing between those values. This distribution is reflective of how dispersed the data is relative to the mean. The x-coordinate distribution displays the largest data distribution in comparison to the y-coordinate and z-coordinate distributions.

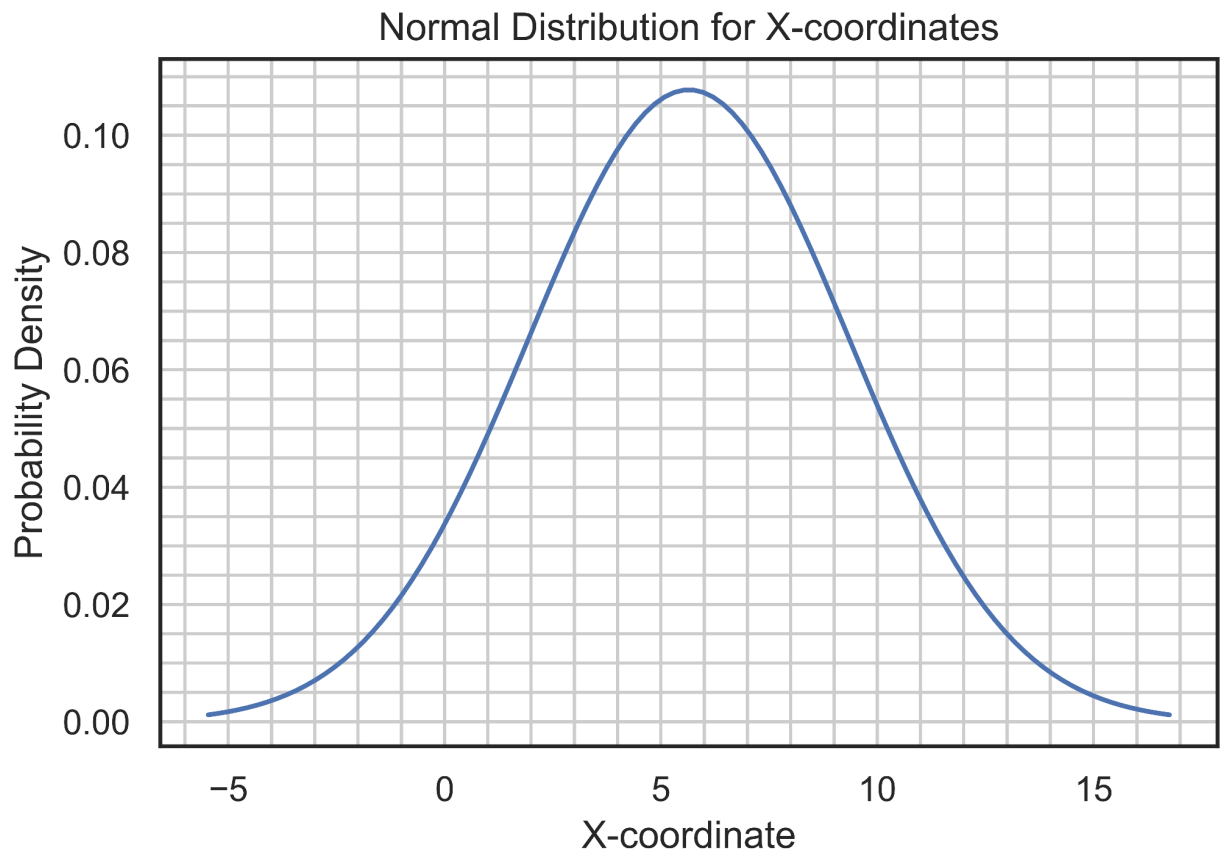


Figure 4: Normal distribution of the X-coordinates.

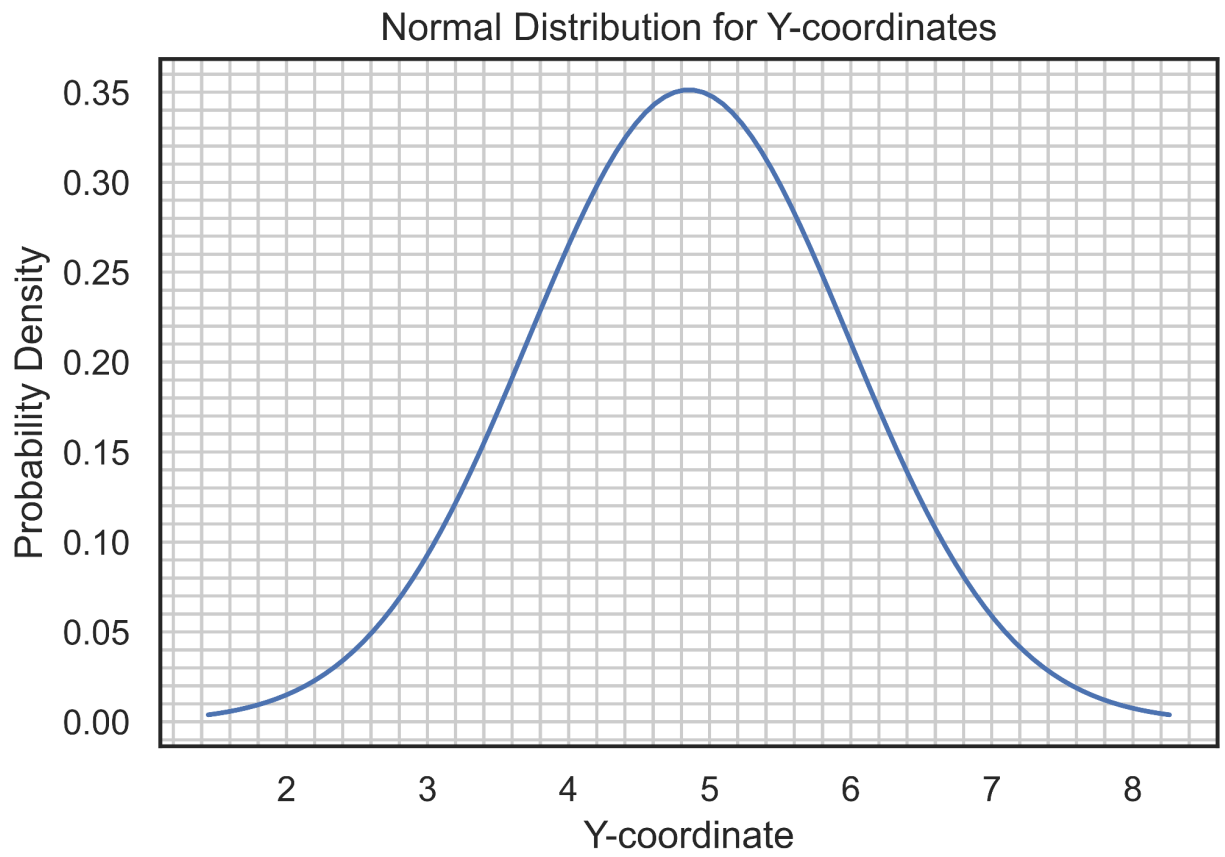


Figure 5: Normal distribution of the Y-coordinates.

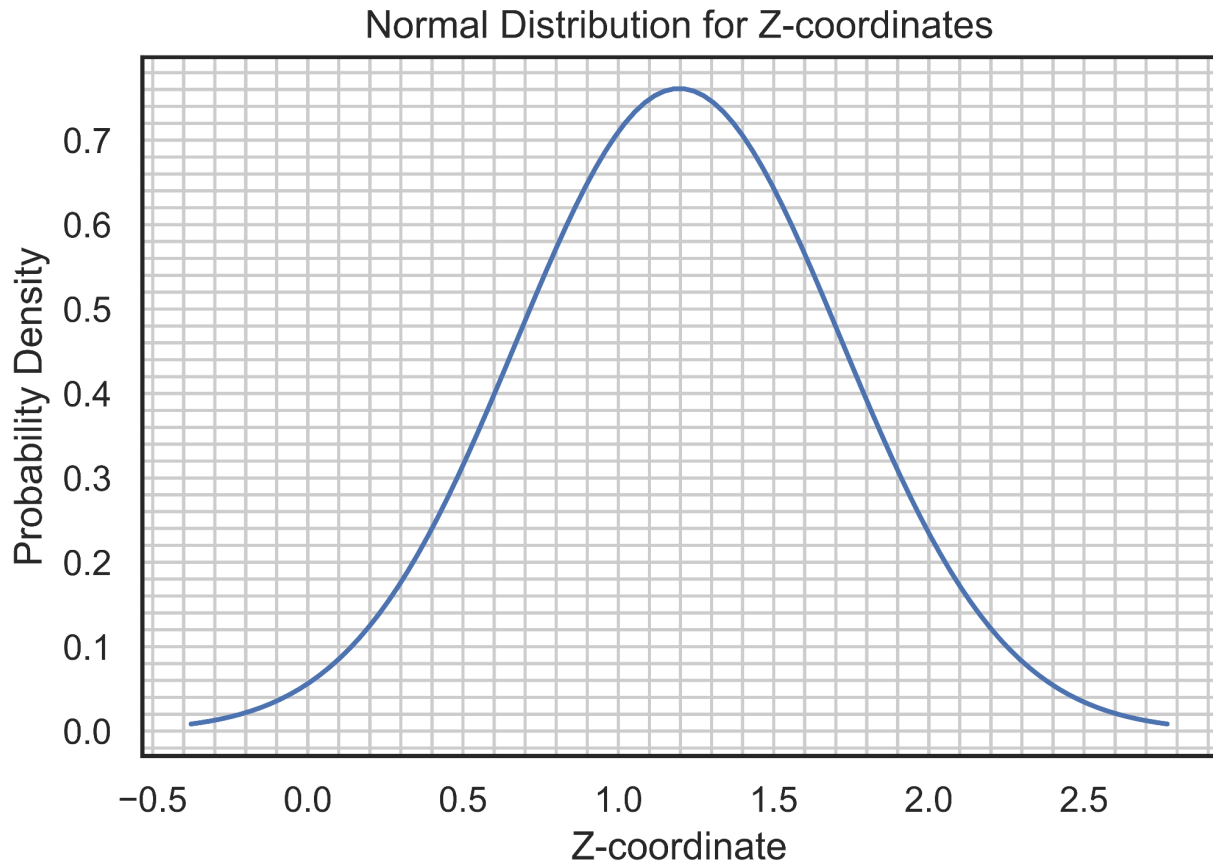


Figure 6: Normal distribution of the Z-coordinates.

2.4 Correlation Analysis

A correlation analysis between variables provides insight on how the variables are interconnected based on the evaluation of the Pearson's coefficient, p . It defines a statistical measure that determines how linear a relationship is. In other words, it is a measurement used for determining the linearity between data sets. The output of this coefficient is a value between -1 and 1 where the lower bound and upper bound represent a collinear relationship. Generally, most variable correlations satisfy a value between these boundaries however there are instances where these values land on the boundary. In such cases, the variable must be removed from the model entirely as eliminating independent variables reduces the complexity of the model as a value of ± 1 indicates collinearity. This means there is a robust relationship between the feature and target variables, making the model redundant. The closer the coefficient is to 1 or -1, the more confidence there is in the linearity between the datasets. The purpose of the model is to identify the underlying relationship between the variables and not an existing one.

Figure 7 displays the correlation plot between the x, y, z, and step variables for the sample. The independent variables indicated display a correlation coefficient with the step variable. The x-coordinate has a correlation coefficient of -0.75 with the step variable, the y-coordinate has a correlation coefficient of 0.3, and the z-coordinate has a correlation coefficient of 0.2. All three variables indicate a value between the boundaries displaying variation in correlations that are not perfectly collinear. Therefore none of the variables can be eliminated. However this does not mean the model will make accurate predictions. The y and z-coordinates display coefficients close to 0 which suggest very little correlation to the step variable. As the data is scattered, the predictions will not match the expected result. Thus, it is the job of the machine learning model to discover the trends in the data and create a fit that represents the target variable based on the feature variables. It must be noted that the model implemented influences the predictions as some models provide better predictions than others. Moreover, the coefficients for these variables do not display a confident correlation with the target variable so additional fine tuning may be required to the implemented machine learning models.



Figure 7: Heatmap displaying the correlation between each variable.

The pairplots shown in figure 8 display the graphical relationship between each variable. Although most of the plots between the independent and target variables have been discussed, this plot provides a general overview of the relationship between all variables, which includes a line of best fit to display the general trend in the data. The additional benefit is the inclusion of histogram plots that display the distribution of an individual variable with respect to the range of values that compose the variable. The histograms display that there is an uneven distribution of the training data which will result in the developed models incorporating bias into its training. Figure 8 can be used as a replacement for figures 1 to 3 however it can be limiting in terms of

the readability of the data and is therefore used to provide a general overview of the data.

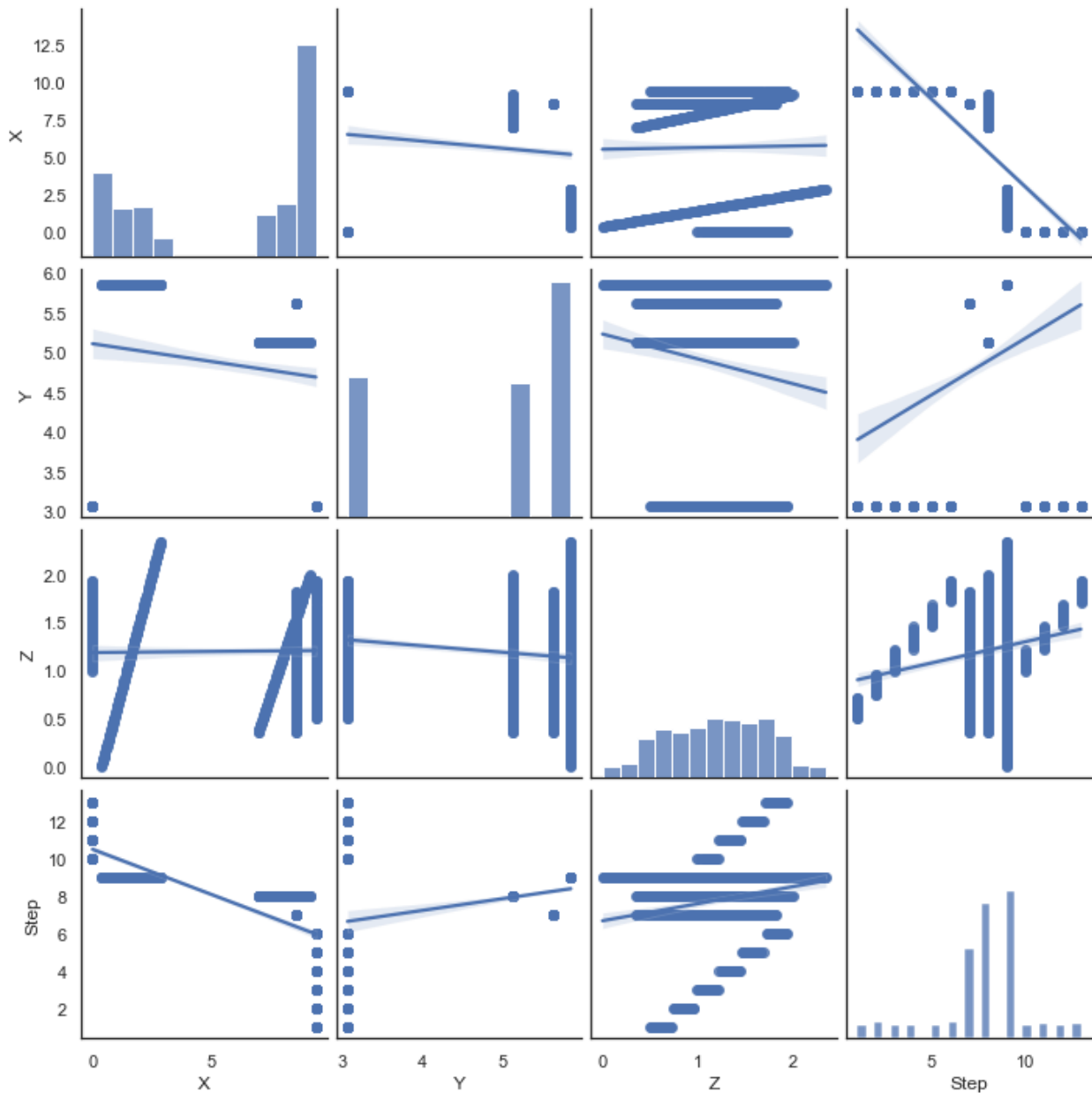


Figure 8: Pairplots displaying the correlation between each variable.

2.5 Classification Model Development

The primary step to developing a machine learning model for the data set involves selecting models that will be implemented. The selected models were the linear regression, decision tree, and support vector machines. These models were developed and trained using the scikit learn library such that the data was fit and used to evaluate the testing data.

Linear regression is a supervised learning model that can make predictions for a target variable based on a set of continuous independent variables. It produces a list of coefficients that represent each independent variable such that when multiplied by a set of inputs, it will predict the desired target variable. The simplest form of this regression is a single independent and dependent variable relationship often referred to as univariate. The linear regression model in this case applies multiple independent variables - multivariate - to predict the output through a linear quantitative relationship. Subsequently, the hyperparameters for each model were tested using the grid search cross-validation method to determine the optimal parameters. Grid search cross-validation determines the optimal hyperparameters that improve the machine learning model predictions. Although each model had a different set of hyperparameters, the methodology for determining the ideal ones is identical. Hyperparameters represent the configuration parameters for a machine learning model that can be modified to provide refined results. The technique used by the linear regression model is the sum of least squares where the objective is to reduce the height of all the data points and the regressed line. The disadvantage of this model is that it is extremely limited for cases where the data is scattered throughout. Variables with correlation coefficients close to 0 forces the model to be rigorous in its predictions, producing high variation error. The model assumes that there is some form of linearity in the data which is limiting when there is little to no correlation between the feature and target variables. Another shortcoming of this model is that the model accounts for all data, including outliers. Outliers hold high leverage such that they are capable of shifting the regressed line towards them by altering the slope of the line. With large datasets, this may not be an issue as one point cannot create a major shift in the slope. Therefore, the data can be predicted more accurately using other machine learning models.

Decision trees are supervised machine learning models that are applicable for continuous and classification type data. The tree is composed of a multitude of nodes that make decisions for a feature. The subsequent nodes represent the outcome of a node and is further divided into leaf nodes that represent a continuous value or classification. The objective of this model is to use the decision tree to predict the target variable in conjunction with the data features to make decisions. The advantage of this model is that it provides a visualization of how the tree is assembled based on the decisions made on each node and the outcome resulting from each leaf node. This makes it simpler to understand the decision making process behind the machine learning model. In addition, the time for training the data depends on the amount of data such that the time required increases logarithmically with the amount of data. In other words, the cost of operating the model for large datasets is exorbitant. Although the model offers an array of benefits, it also comes with several consequences. These include overfitting where over-complicated trees are developed such that they do not fit the data appropriately. This can be controlled by the maximum depth hyperparameter which is responsible for controlling how complex the tree should become to fit the data. In general, the higher the value of maximum depth, the more complicated the tree becomes. An additional parameter that influences overfitting is the minimum number of splits. This parameter controls the number of samples required to split an internal node. A smaller value corresponds to an intricate tree whereas higher values correspond to simpler trees that are resistant to perturbations.

Support vector machines (SVM) is a supervised machine learning model that is used for classification and regression data analysis. Using classifications, the model determines a hyperplane that distinguishes the data into classes all the while ensuring sufficient separation between the classes. The margin of separation is determined by the distance between the points of proximity that define the hyperplane. These points establish the support vectors that bound the hyperplane. The advantage of this model is that it is able to accurately model multi-dimensional sets of data such as the current problem where there are three dimensions of feature variables that compose a single target variable. Another advantage is that the kernel function for the SVM can be specified as the decision function. Kernel functions perform pattern analysis using classifications to solve both linear and non-linear problems. On the contrary, SVM can suffer from overfitting for cases where the number of features exceeds the total number of samples for the dataset.

2.6 Model Performance Analysis

The accuracy, precision, and F1 score reflect the performance of the models based on a set criteria. The accuracy score is a measure for how many of the predicted values match the expected answers from the test data. The precision score is a ratio of true positives and the sum of true positives and false positives that determines the ability of the classifier to not label a positive sample as a false positive. The F1 score is a measure of the harmonic mean of precision and recall that determines the accuracy of the model. Precision measures the number of positive predictions that were correct and recall measures the valid number of positive class samples that were recognized by the applied model. Generally, this metric is only valid if the classes for each dataset contain the same number of samples. Based on the three metrics, F1 score is the ideal metric for evaluating the models as it accounts for the imbalanced labels or uneven distribution of data. This means that the amount of data for each class of target variable is not equal, resulting in skewness in the training data.

Table 4 describes the scores for the accuracy, precision, and F1 scores implemented from the scikit library. Based on the results, the support vector machine model displays the best results as these scores are greater than the linear regression and decision tree models.

Table 4: Accuracy, precision, and F1 score for each developed model.

ML Model	Accuracy Score	Precision Score	F1 Score
----------	----------------	-----------------	----------

Linear Regression	0.13953	0.20682	0.13100
Decision Tree	0.99418	0.99491	0.99341
Support Vector Machine (SVM)	0.99418	0.99612	0.99450

A confusion matrix displays an n by n matrix that evaluates the performance of the classification ability of a machine learning algorithm. The diagonal represents the number of points that the predicted label was able to successfully match to the expected label. If the diagonal values are higher, it means that the predictions made by the model successfully match the expected labels. In addition, data beyond the diagonal indicates incorrectly classified data. Figure 9 displays the confusion matrix for the support vector machine model such that the labels 6, 7, and 8 display the highest number of correct predictions. It must be noted that the labels represent steps 7, 8, and 9 as the confusion matrix begins its indexing at 0. There is only one prediction that was misclassified out of all the test samples indicating a highly accurate model. In other words, out of the 172 sample cases, 171 of the cases were predicted correctly.

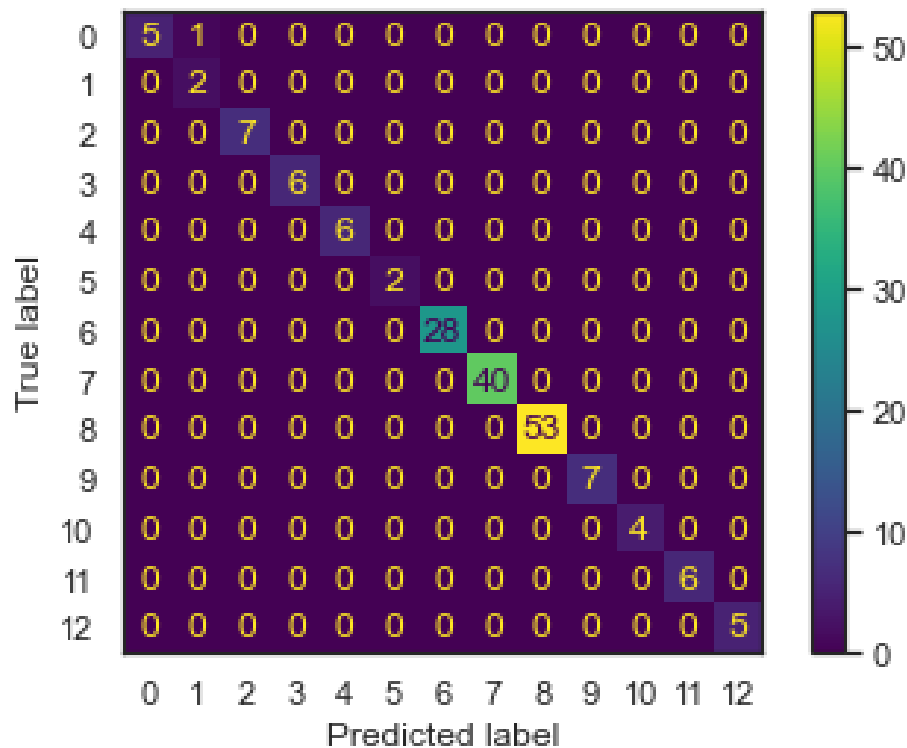


Figure 9: Confusion matrix for the Support Vector Machine model.

2.7 Model Evaluation

Using the joblib library, the machine learning model was saved and loaded. The advantage of the joblib library is that it provides a platform to pipeline python operations for streamlining performance and improving the reliability of machine learning models. The support vector machine model was evaluated using the provided test cases where the provided data was in the form matching the inputs of the SVM model. Table 5 displays the predictions made by the SVM model for each set of coordinate inputs provided.

Table 5: Predicted step number for the provided x, y, and z-coordinates.

[X, Y, Z] Coordinates	Step Number
[9.375, 3.0625, 1.51]	5
[6.995, 5.125, 0.3875]	8
[0, 3.0625, 1.93]	13
[9.4 ,3, 1.8]	6
[9.4, 3, 1.3]	4

3.0 Conclusion

To summarize, the purpose of this project was to develop machine learning models to predict the maintenance step number based on the feature and target variables. The feature variables were identified to be the x, y, and z-coordinates and the target variable was the step number. The provided data was separated into training and testing data using an 80% to 20% ratio respectively. The data was visually and statistically analyzed to determine the trends in the training data and the relationships between the variables. Furthermore, a plot of the correlation coefficients was developed to understand how the feature variables correlate with the target variable. The coefficients were verified to ensure there were no collinear relationships that would make the machine learning development process redundant. It was determined that the x-step correlation was -0.75, the y-step correlation was 0.3, and the z-step correlation was 0.2. Linear regression, decision tree, and support vector machine models were developed and evaluated based on the accuracy, precision, and F1 score. The support vector machine was selected based on its F1 score of 0.99450 which was the highest of all three models. In addition, a confusion matrix was generated for the SVM model that displayed a number of predictions that were correct and incorrect. A total of 171 out of 172 values were predicted correctly, indicating the high accuracy of the SVM model. Lastly, a sample set of coordinates were provided to evaluate the performance of the models.