Ryerson University

Department of Aerospace Engineering
Faculty of Engineering & Architectural Science

Semester, Year:

Semester, Year:	☑ Fall ☐ Winter ☐ Spring/Summer 2023
Course Code:	AER 850
Course Title:	Introduction to Machine Learning
Section Number:	01
Instructor:	Reza Faieghi
Submission:	☐ Assignment ☐ Lab Report ☑ Project Report ☐ Thesis ☐ Other: ________________
Due Date:	October 15

Project 1 Report

September 11, 2023

Authors/Contributors:	Student Number (XXXX12345):	Signature*:
Kotasthane, Gaurang	500922614	G.K.

**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/academicintegrity/students/ryersons-academic-integrity-policy-policy-60/*

Discussion

Data Visualization

In this dataset, there are three input variables (or features) X, Y, Z and a target ‘Step’ which has 13 classes. The dataset has been visualized in a few different ways to gain a deeper understanding of the behavior and the findings have been summarized below.

1. Pairwise Scatter Plots:

The was to examine the relationships between the features X, Y, Z. Understanding what the dataset represents is important while looking into the relationships between the features. Since X, Y, Z are the coordinates in space, it is very plausible that might not have direct relation with each other.

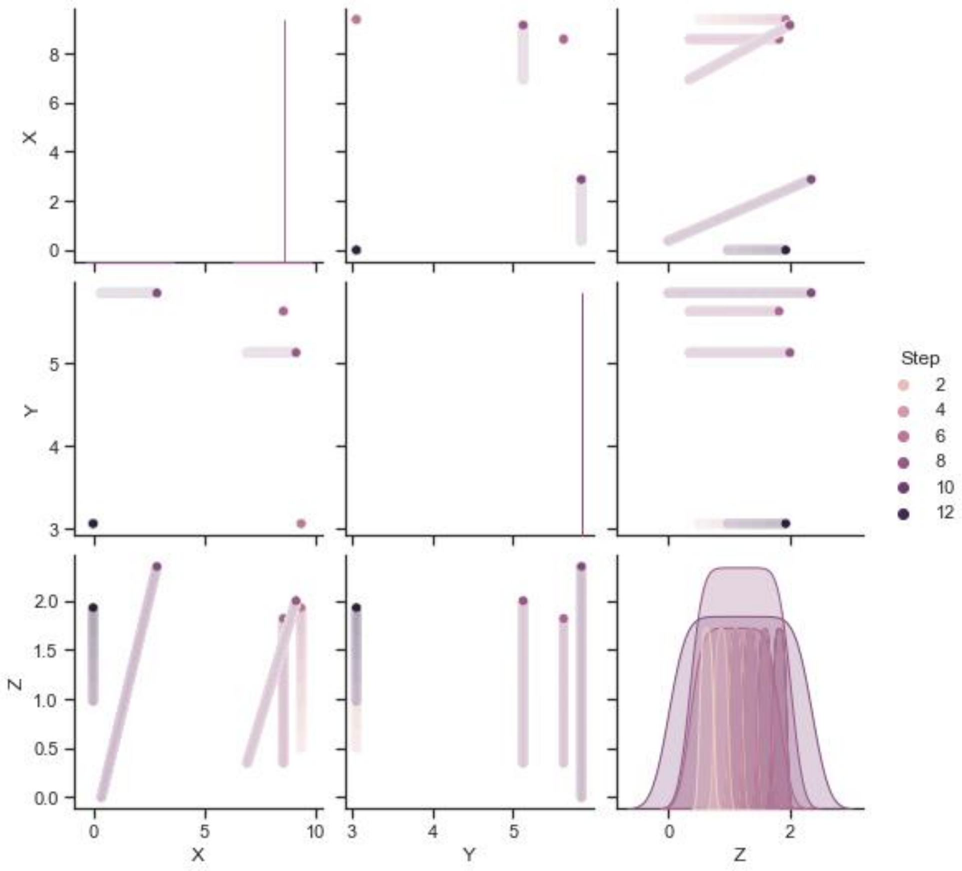


Figure 1. Pairwise Scatter Plot

2. Histogram Pair Plot:

This was to examine the relationship between the features and the target variable. The plots that are particularly interesting are the individual features (X, Y, Z) vs Step. It gives an idea how each feature might be able to predict the target. For example, the X vs Step graph in the top right corner helps us understand that the higher the value of X, the lower is the step class and vice-versa. Similarly looking at Y vs Step, classes 7, 8, 9 have a Y value significantly higher than the rest.

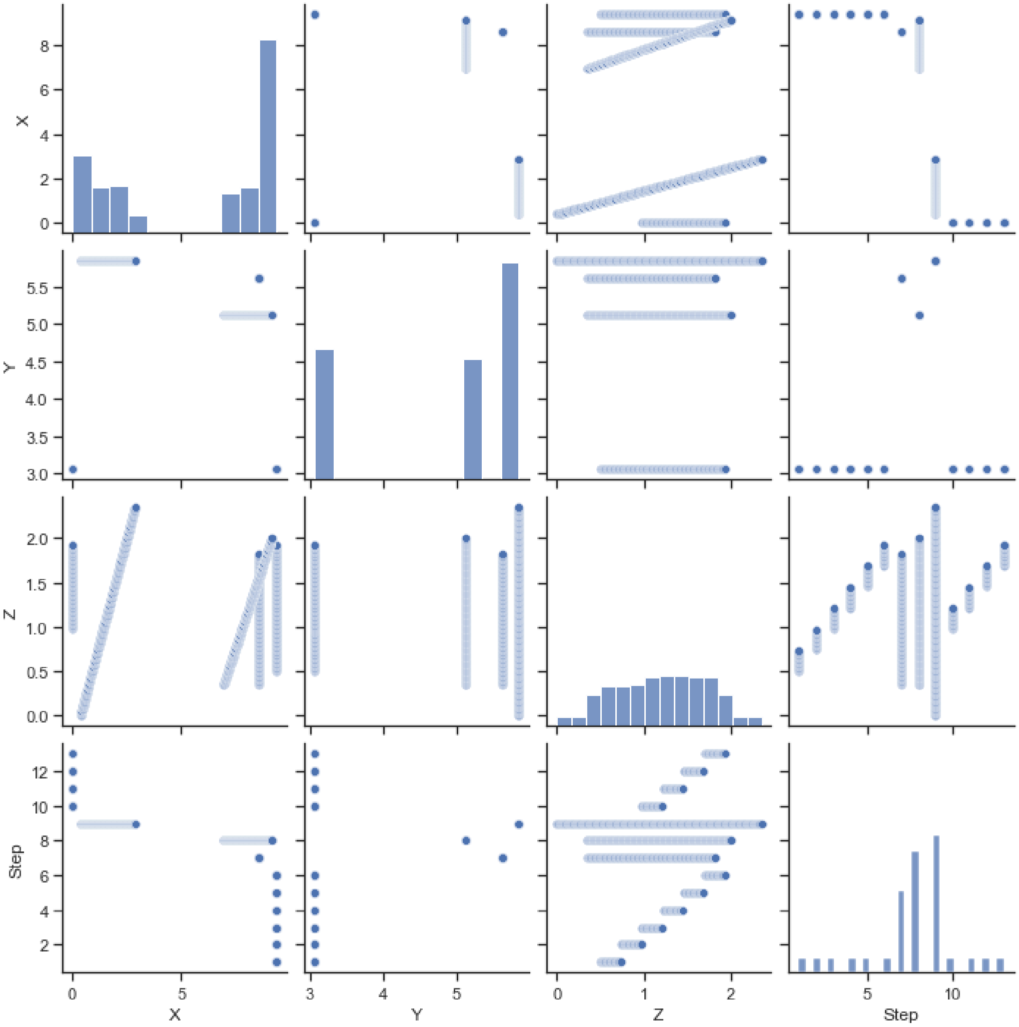


Figure 2. Histogram Pair plot

3. Step Frequency Graph:

This graph helps us visualize the distribution of the training data. The graph clearly shows that the data set has a lot more data for Step class 7, 8 and 9 which is important to understand as it tells us that there is a need to stratify the data when splitting it into training and test data sets.

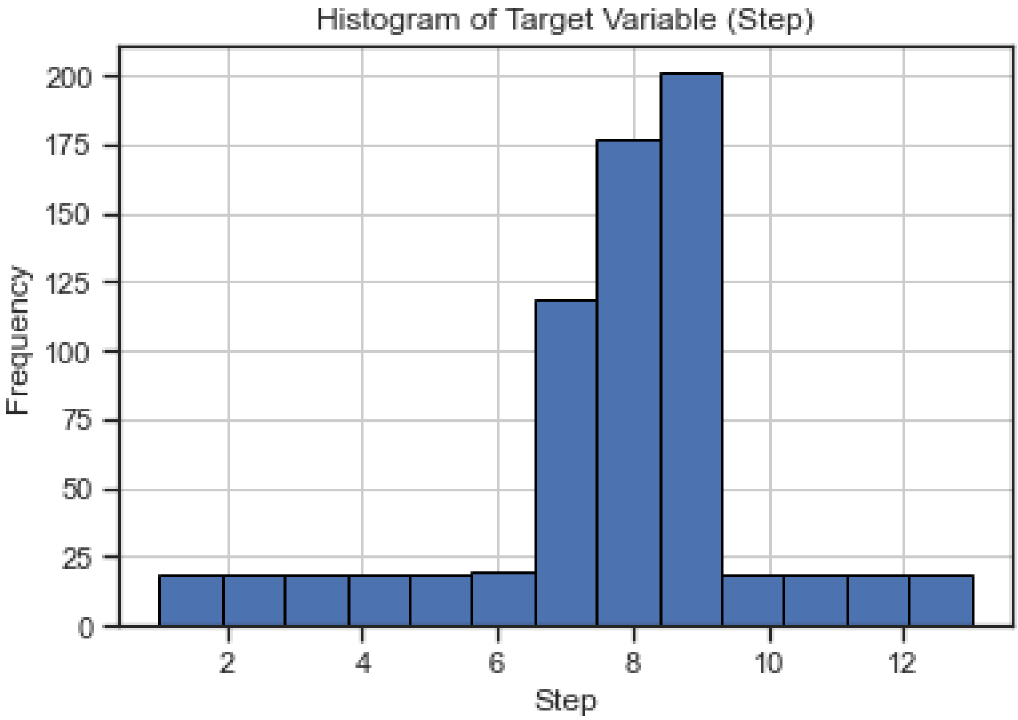


Figure 3. Step Frequency Graph

Correlation Analysis

In the section, a correlation plot between the features (X, Y, Z) and the target variable (Step) was developed.

X vs Step: -0.75

The negative correlation coefficient of -0.75 indicates that there is a strong inverse relationship. This means that as the value of the feature X increases, the Step variable tends to decrease. Conversely, as X decreases, the Step variable tends to increase. This was also indicated in the top right graph of figure 2.

Y vs Step: 0.29:

The positive correlation coefficient of 0.29 between feature Y and target step is a mild correlation but no significant inference can be made from this value. This value could also be affected by the 3 Y data points that are higher in value associated with Step class 7, 8, and 9.

Z vs Step: 0.2

The positive correlation coefficient of 0.2 between feature Z and the target variable Step indicates there is very low correlation between the variable and the target. Although it might help with prediction of the target class, its impact isn't as prevalent as with feature X and Y.

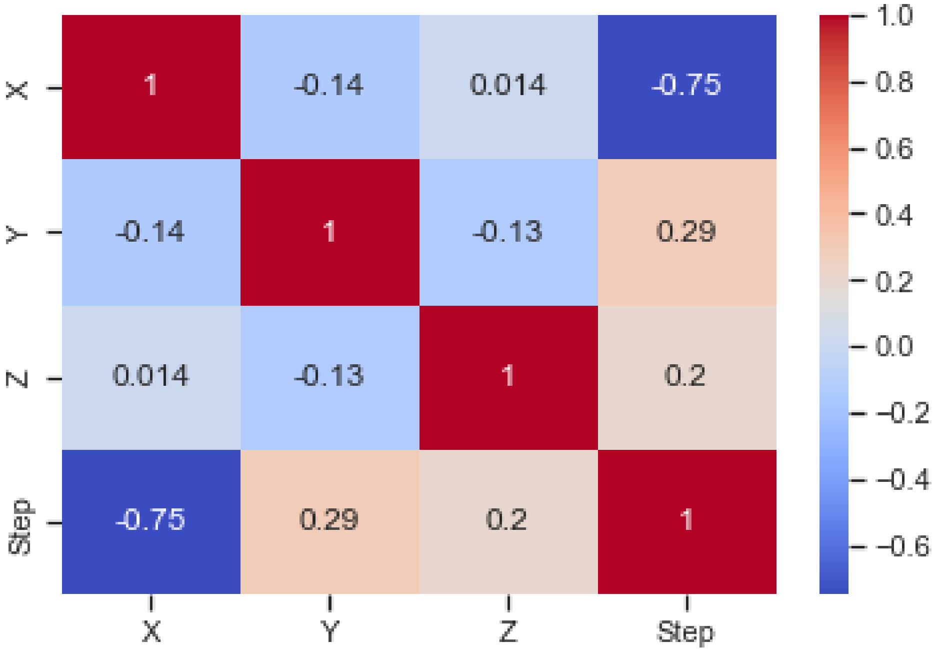


Figure 4: Correlation Plot

Classification Model Development/ Engineering

The top three models chosen were logistic regression, support vector machines and Random Forest. For all three models the best parameters were chosen using grid search cross validation method. The selection of the models was based on their suitability for the dataset. Linear Regression was not chosen because it is more suitable for a continuous distribution of data. Furthermore, the target is a multiclass variable where the linear regression model would perform poorly, especially because it is a classification problem. Here's some key takeaways from each model and the derived best parameters using grid search cross validation.

1.) Support Vector machine:

SVM was specifically developed to solve classification problems, it has a powerful algorithm to create hyperplanes between complex data to segregate into various categories. I selected SVM as one of the models because of its capability to handle both linear and non-linear relationships in the data (as we previously saw in the data visualization section of this report). The results from Grid Search Cross Validation indicate that the best parameters for the given dataset are C=10 and kernel: linear.

2.) Logistic Regression:

Logistic Regression was chosen as one of the models because it is widely chosen for classification problems. Although it is interesting to note that it is usually implemented for binary classification problems rather than multiclass classification and so I thought it would be interesting to see how logistic regression would perform on multiclass classification problem. The best parameters according to grid search cross validation are C=10 and solver = lbfgs

3.) Random Forest:

Random forest is a learning method that combines multiple decision trees to improve classification accuracy. It is renowned for its robustness and generalization. The best parameters included having 10 estimators.

Model Performance Analysis

Model	Precision	Accuracy	F1 Score
Support Vector Machines	99.4%	99.4%	99.4%
Logistic Regression	98.9%	99.4%	98.9%
Random Forest	100%	100%	100%

1.) Accuracy:
It is a metric that measures the proportion of correctly predicted instances (both true positives and true negatives). It ranges from 0-1 or from all incorrect to all correct. While intuitively accuracy might be the best metric for evaluating the performance of a machine learning model, it may be inaccurate when evaluating the performance on an imbalanced dataset.

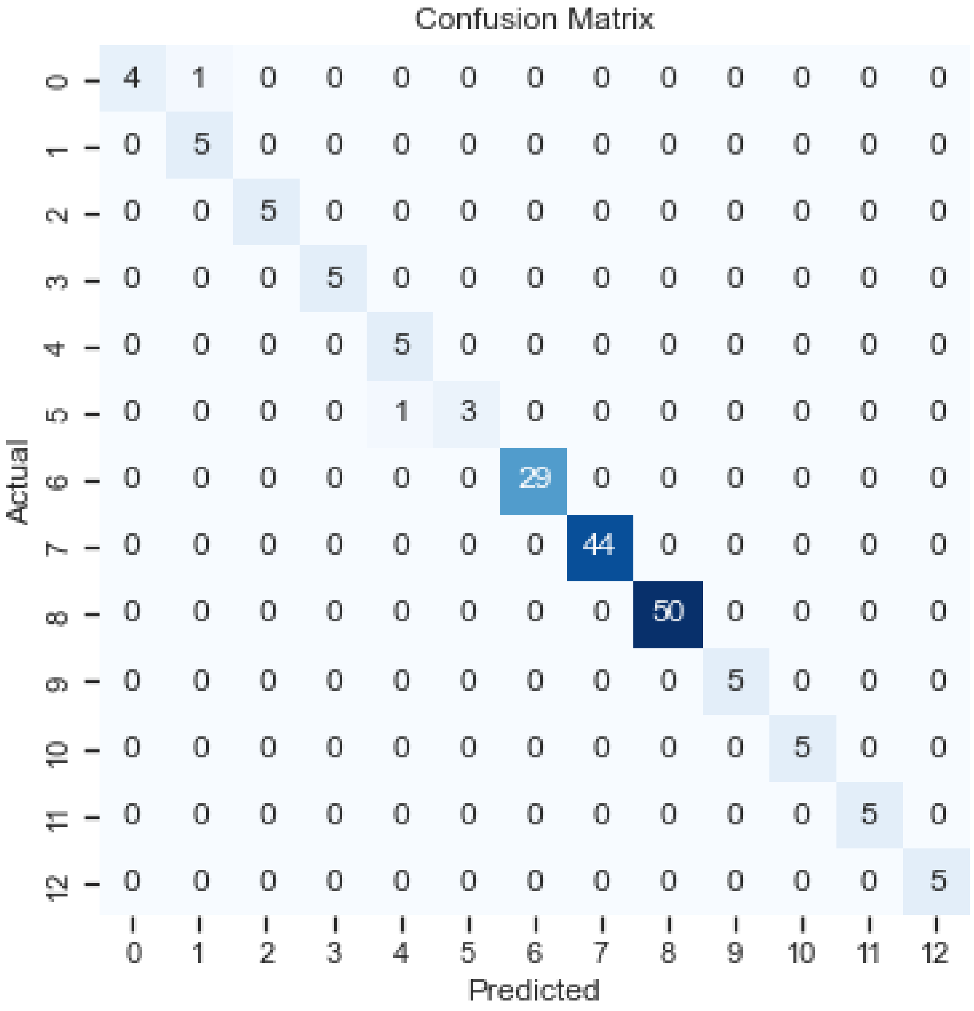
2.) Precision:
It is a metric that calculated the proportion of true positive prediction out of all the positive predictions made by the model. So basically, it tests how often the model does not make false positive predictions.

3.) F1 score:
This metric is a combination of inputs from precision and recall. This is especially an important metric for a lot of reasons, for example, if a model predicts everything as true positive or true negative, it's precision score might be high but that does not mean that the model is performing well because, we want our model to correctly identify the right things but at the same time correctly identify the wrong things and F1 is a sweet spot between precision and recall.

Based on the scores calculated above, **Random Forest** model was selected and its performance was evaluated on the test dataset and the results are tabulated below

Selected Model	Precision	Accuracy	F1 Score
Random Forest	98.8%	99%	98.8%

Further a confusion matrix was created by using the trained random forest model on the test dataset.



	0	1	2	3	4	5	6	7	8	9	10	11	12
0	4	1	0	0	0	0	0	0	0	0	0	0	0
1	0	5	0	0	0	0	0	0	0	0	0	0	0
2	0	0	5	0	0	0	0	0	0	0	0	0	0
3	0	0	0	5	0	0	0	0	0	0	0	0	0
4	0	0	0	0	5	0	0	0	0	0	0	0	0
5	0	0	0	0	1	3	0	0	0	0	0	0	0
6	0	0	0	0	0	0	29	0	0	0	0	0	0
7	0	0	0	0	0	0	0	44	0	0	0	0	0
8	0	0	0	0	0	0	0	0	50	0	0	0	0
9	0	0	0	0	0	0	0	0	0	5	0	0	0
10	0	0	0	0	0	0	0	0	0	0	5	0	0
11	0	0	0	0	0	0	0	0	0	0	0	5	0
12	0	0	0	0	0	0	0	0	0	0	0	0	5

Figure 5: Confusion matrix