**Week-3 Assignment Report:**

**Introduction:**

This assignment is aimed at understanding and implementing different classification and regression techniques to the dataset provided to us. The model needs to be trained using Logistic Regression and Naïve Bayes Classification, along with the EDA, model performance evaluation, make predictions and analyze the results and draw conclusions. The AI4I 2020 is a synthetic dataset that reflects real predictive maintenance data encountered in industry. It contains 10,000 data points stored as rows with 14 features in columns. Machine failure in the dataset has 5 independent failure modes.

**EDA:**

Exploratory Data Analysis (EDA) is carried out on the dataset to understand the data better by checking for missing or duplicate values in the dataset. Duplicate values are checked and dropped, and the categorical columns are converted to integer values as required. We then drop the UDI and product ID as it isn’t necessary for model training, check for missing values and find out the unique values for each column.

In order to determine how closely each of the features correlated to the output (machine failure) and find outliers, the research uses data visualization techniques like the bar graph, pie chart, and correlation matrix.

**Model Implementation and evaluation:**

The information being used is split into a training set and a testing set as 70:30. The values x train and y train contain 7000 data elements. In the variables x test and Y test, 30% of the training data is reserved. The model is then taught using training data, where predictions are used as y train and input is used as x train. The model is evaluated using the testing data collection after training by providing the input X test. Then, Y pred is used to keep this info. Comparing y test and y pred is used to test the model's accuracy using the mean square error technique. Recall, f1-score, support, and classification report are all computed for model precision assessment. The model's confusion matrix is computed and visualized using matplotlib.

**Conclusion:**

The results show that logistic regression has the highest accuracy among the different regression and classification algorithms used with an accuracy of 60% compared to the accuracy of just 30% of naïve bayes algorithm. Since the general accuracy of logistic regression model is higher than the trained accuracy, it can be concluded that unseen data fits better when this algorithm is used.

Github link: <https://github.com/josephgit10/ENPM808Y_proj1>