ENPM808Y - FUNDAMENTALS OF AI AND MACHINE LEARNING

INTRODUCTION: The problem is to implement and understand the different types of regression and classification algorithms in Machine learning such as Linear Regression, Logistic Regression, and Naive Bayes. The dataset given to perform these algorithms are from the University of California's, Irvine machine learning repository. This AI4I 2020 Predictive Maintenance Dataset is a synthetic dataset that reflects real predictive maintenance data encountered in the industry. This dataset consists of 10,000 data points and 14 features for each of them. The features are UID, Product ID, Air Temperature, Process Temperature, Rotational Speed, Torque, Tool wear, Machine Failure, Tool Wear Failure, Heat Dissipation Failure, Power Failure, Overstrain Failure, and Random Failures. If one of the above-mentioned failures is true, the machine failure label is set to 1.

TO DO: Perform EDA and pre-process the dataset given, build a model using all three different algorithms mentioned above, and evaluate the model's performance using the accuracy of the predictions made by the model. Finally, compare the models and conclude which algorithm is best suited for having better performance.

METHOD FOLLOWED: Import required libraries such as pandas, sci-kit-learn, seaborn, numpy, and matplotlib. Then import the given dataset and read it. Then perform Exploratory Data Analysis by plotting the data points for the feature "Type" using a pie chart or histogram. Then plot the correlation matrix and visualize it using a heatmap. Then split the given dataset into training and testing sets. In the code, it will be mentioned as X_train, X_test, y_train, and y_test.

Linear Regression: Build a model and fit the dataset to the model and using the model we predict the values of y_test for unseen data which is then used for evaluating the metrics. Evaluation of metrics is done using mean squared error and R^2 score.

Logistic Regression and Naive Bayes: Build a model and fit the dataset to the model. Using the model we predict the values of y_test for unseen data which is then used for evaluating the metrics. Evaluation metrics are calculated by comparing the test data and predicted values for the data and it is shown using training accuracy and model accuracy.

CONCLUSION:

Based on the results shown by the algorithms where linear regression has R^2 score of 0.001081 which is not a good score and of the rest two Logistic regression and Naive Bayes, Logistic regression's model has given a higher percentage of training accuracy - 59.46% (~60%). Therefore, Logistic Regression has better performance than the other two algorithms. Having a huge feature list, the Naïve bayes model has lower accuracy because the likelihood would be distributed and the features are independent in this dataset which is also a reason for its low accuracy.

GitHub Repo Link - https://github.com/sandipsharan/ENPM808Y-Homework2