REPORT

The main goal of the project is to predict the failure of a machine using various inputs using three algorithms, namely 1. Linear regression 2. Logistic regression and 3. Naïve Bayes algorithm. There are a number of input functions such as air temperature, rotational speed and torque that determines the outcome of 5 failure modes (tool wear failure, heat dissipation failure……random failure). If any one of the failure modes is violated, then it can be said that the machine has undergone a failure.

Datasets: An existing dataset called "AI4I 2020 Predictive Maintenance" is being used for the project consisting of 10000 data points and 14 features or columns. Out of these features, 7 are numeric columns and the remaining 7 are categorical.

Algorithms used: 1. Linear regression 2. Logistic regression and 3. Naïve Bayes algorithm

Data visualization and preprocessing: The first step towards any machine learning problem is to understand the dataset presented and visualize the data graphically and theoretically so that we can have a profound understanding of the problem that we are dealing with and solve premature problems like removing outliers, convert categorical columns to numeric ones, check consistency of the datatypes or check for duplicate values. The techniques mentioned in the previous sentence come under the category of data preprocessing: The process of cleaning and tidying up datasets so that the machine learning model can understand the datasets properly. Data visualization techniques like the bar graph, pie chart and correlation matrix are used in the study to identify how closely each of the features correlated to the output (machine failure) and spot outliers. Dividing the datasets into training and testing sets also comes under the above category.

Model implementation and evaluation: After structuring the datasets, the ML model for all 3 algorithms is trained with training data sets consisting of 7000 data points. For linear regression, we used these training sets to fit the model and predict the outcome of the test sets. The cost function is calculated using mean squared error and the correlation between the datasets and the linear regression model is obtained by using the r2 score followed by finding the coefficients and y intercept. Logistic regression follows the same first step as linear, except the fact that in the evaluation metrics, we use lr.score(which is r2 score by default) and accuracy score(number of predictions/total number of predictions). The training and model accuracy is then obtained, and the obtained test results are visualized using a confusion matrix. Additionally, a classification report that includes the precision, recall, f1 score and support is obtained. Naïve bayes model follows a similar trend as logistic regression.

Result and conclusion: From the results obtained by comparing all the 3 algorithms on the dataset, logistic regression fits the best and the reasons are stated below. It has an accuracy nearing 60% for both the general and training data sets as compared to naïve bayes algorithm which has nearly 30% for both the datasets. Also, for the logistic regression model, the general accuracy is slightly higher than trained accuracy, meaning it can fit unseen data better.