Machine Learning Project

(Predictive maintenance of machines in IOCL Refinery)

Step wise process:

Step-1)

* Importing Required Libraries (sklearn, numpy , pandas, etc..)
* Importing the dataset

step-2)

Data Preprocessing :

* Drop unwanted features
* Remove replace missing values from the dataset
* **Exploratory Data Analysis** (EDA) is an approach to perform initial investigations on data to discover patterns, spot anomalies, test hypothesis and check assumptions with the help of statistics and graphical representations.

STEP -3)

Data Visualisation

* Representing data in graphical form

(bar graph, Pie chart, Box Plot..)

Step-4)

Data Preprocessing for Prediction

* Encoding categorical features
* Splitting test & train data

(The test-train split is used to estimate the performance of machine learning algorithms using training data and validate with test data.)

* Feature Scaling

**Feature scaling** *is a method used to normalize the range of independent variables or features of data and make them closer to each other. Feature scaling is essential for machine learning algorithms that calculate distances between data (Ex- KNN).*

Step-5)

Prediction: Model development

* Here we will use different classifiers like-
* Logistic Regression
* k Nearest Neighbours
* Support Vector Machine
* Random Forest
* naïve Bayes
* Decision tree

Step-6) Finding the best model with highest accuracy and performance

* **Accuracy**- Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.(Accuracy = TP+TN/TP+FP+FN+TN)
* **Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.(Precision = TP/TP+FP)
* **Recall** - Recall is the ratio of correctly predicted positive observations to the all observations in actual class.(Recall = TP/TP+FN)
* **f1-score** - f1-score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
* **Support**- Support is the number of actual occurrences of the class in the specified dataset
* The **Confusion Matrix** is a matrix used to determine the performance of the classification models for a given set of test data. It is used to visualize important predictive analytics like recall, specificity, accuracy, and precision. Confusion matrices are useful because they give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives.
* After testing all the algorithms, the one with highest accuracy will be selected and will be used on the actual dataset