Binary Classification with Supervised Learning-

INTRODUCTION-

This report gives the details about the performance comparison of four supervised learning algorithms on the given dataset in Python using the Scikit-learn library. The comparison between the algorithms is done using the computational performance metric – training time and predictive performance metrics – accuracy of the model and the F-measure score also known as F1 score. The metrics are calculated on the evaluation set of data generated using the stratified splits. The dataset consists of 4601 instances and 58 attributes out of which 1 attribute is the target attribute. Initially in this project, the dataset is splitted into 4:1. The model is then trained and then the instances are classified using the four supervised learning algorithms. The algorithms chosen are - Gaussian Naive Bayes, Decision Tree Classifier, Random Forest and Logistic Regression. Once the instances of the classifier models are created and the results are generated.

II. Analysis

In the exploratory analysis we get to know that Feature X1 to X55 contains float values while X56 and X57 have int values, and X58 is the output feature or dependent variable. There are some max values which serve as outliers in the X55,X56,X57 feature columns.Then we get to know that there are no missing values in the dataset. Whereas most of the columns have average 80-85% of zeroes in that. While X34 is highly correlated with the X32 so we should drop any of it.
And 2376 values are identified as class'0' whereas 1534 values are from class '1'.

Then for the feature reduction process we have done LDA. For splitting with the LDA in case for checking variance in the dataset and then we see the ratio of explained variance.
And then we have trained it with the Logistic Regression,DT and Random Forest models. Logistic Regression slightly performed better after the new feature set, but the Tree classifiers have done poorly on that. Given below the table for the Logistic Regression algorithm

Metric	Before LDA	After LDA
Accuracy	90.4	91.43
Training Time	1.32	0.01
F1 score	87	88

We will see some feature selection methods for automatically reducing the features according to the variance.Shape of the X_train and reduced variance train set is the same so no feature is reduced. Thus we will be using X-train as is.

VarianceThreshold is a simple baseline approach to feature selection. It removes all features whose variance doesn't meet some threshold. By default, it removes all zero-variance features.

III Results

Hence after training the models on the evaluation dataset we get to know the training time,accuracy and F1 score of different models. We plot the ROC graph to show the comparison between the models. And by the tree classifiers we get to know that X51 and X6 have highest feature importance in the prediction.

Metric	Naive Bayes	Decision Tree	Logistic Regression	Random Forest
Accuracy	82.74	91.3	90.41	95.2
Training Time	0.009s	0.065s	1.42 s	0.38s
F1	81.3	89.1	87.8	96

For the sake of speed the naive bayes outperforms all the other models. But the accuracy,F1 and ROC characteristics are better for Random Forest and DT.Hence we will go with the Random Forest for testing it with a hold out set we have reserved. And then we scored it against a test set and appended the predictions.