DATA MINING AND MACHINE LEARNING ASSIGNMENT-I
Srijit Saha-MDS202150

Project report:
The dataset worked upon is ”The bank Marketing Dataset” from the UCL Machine learning Repository, consisting of 41187 rows and 21 columns,corresponding to the attributes of the dataset and one class vector.
The corresponding features are:
'age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed'
Class category – ‘y’
The class category refers to the last column which specifies whether the loan has been approved for the particular individual or not.
Data filtering
We have removed all those rows featuring an “unknown” value in any one or more attribute of a row.
Moreover, we have dropped the column “default” in the dataset as it contains redundant values.
Feature Selection:
Use of KbestFeatures has been made to select the K-best (K =6) features to fit our data to the model.,using mutual information again procedure.
One Hot Encoding has been used to encode the data,owing to the cardinal nature of the categorical attributes.

Using these K-features,the data was split into training and testing data in a ratio of 7:3.
For each model, a measure of accuracy, precision, and recall was chosen on which we could judge the working of each model through accuracy metrics.
Further, we have applied some hyperparametric tuning for each model so as to optimise their performance based on the accuracy metric feedback.
We have tried to maximise the precision and recall values without hampering the accuracy of the model too much.

A summary of the metrics has been given below:

	Accuracy	Precision	Recall	Running Time
Decision tree	91.057	61.54	55.56	0.057
Naïve Bayes	87.135	45.49	71.76	0.030

Random Forest	91.109	61.75	57.17	1.533

Observations:

The dataset is highly imbalanced as can be seen from the exploratory data analysis.

In terms of accuracy, both decision trees and random forest gives more or less the same values ,while the Naïve Bayes method gives us a slightly lower value.

Even though the precision value of the naïve bayes is lesser than both of the decision tree and the random forest approaches, its recall value is superior as compared to those of both the random forest and the decision tree methods.

The running time of the Random forest is significantly larger than the other procedures owing to greater computation involved in its algorithm.