A Comparative study of Different Classifiers

Introduction:

This report discusses the performance of three classifiers: Decision Tree, Naive Bayes and Random Forest. Later on, we use suitable evaluation metrics to compare the performance of the three classifiers.

Introduction to the Data Set:

The "Bank Marketing Data Set" from the UCI Machine Learning Repository is related with direct marketing campaigns (phone calls) of a Portuguese banking institution.

The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data pre-processing:

From the Bank Marketing Dataset, the columns we are dropping are ‘duration’, ‘default’ and ‘pdays’.

Duration of a call is not known beforehand and once we know the duration; we also know if the final answer is yes or no. Hence, we drop the duration column.
We drop the default column because it has too many 'unknown' values. We drop the 'pdays' column because it has high standard deviation and as less than 5% of the people have been contacted before.

Then, we change the ordinal values manually into numerical values ranging from -1 to 6 while preserving the order. We also changed the remaining categorical values to numerical values using get dummies (which implements One-Hot Encoding). Then we split the data into training and test data.

Building different Models and final performance evaluation:

Then we fit the following four models on both the datasets:

i) Decision Tree Classifier
ii) Naïve Bayes Classifier
iii) Random Forest Classifier

We use the following metrics to compare the performances of the three and conclude which one is more significant in our area of interest.

	Decision Tree Classifier	Naïve Bayes Classifier	Random Forest Classifier
Accuracy	83.44%	86.2%	83.09%
Precision	35.47%	39.69%	35.17%
Recall	58.46%	44.75%	60.45%
F1 Score	44.15	42.07	44.47

In our scenario, when forecasting whether clients will subscribe to the term deposit, we prioritize maximizing recall rather than solely focusing on accuracy due to the dataset's imbalance. Our goal is to minimize the risk of incorrectly categorizing potential subscribers as 'Not likely to subscribe' i.e. to minimize the number of False negatives to ensure we capture all potential subscribers.

Out of the three models we implemented, Random Forest Classifier gave the maximum recall, so we can conclude that Random Forest Classifier performed better than the other two classifiers.