DMML Assignment 1

Om Ambaye BMC202006

Alok Dhar Dubey BMP202002

Task 1:

From the given 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.
Then we fit our Decision Tree Classifier as well as the Naïve Bayes Classifier model to the training data.

We used GaussianNB Classifier because this set has continuous values of attributes such as cons.price.idx, emp.var.rate, etc.

Evaluating our Models:

	Decision Tree Classifier	Naïve Bayes Classifier (Gaussian)
Accuracy	83.44%	86.2%
Precision	35.47%	39.69%
Recall	58.46%	44.75%
F1 Score	44.15	42.07

Naïve Bayes model provides more Accuracy but Decision Tree model gives considerably more Recall score which is important in this case since the number of ‘yes’ is in a minority. So as a whole, Decision Tree model would be preferred in this case.

Time and Memory consumed

	Decision Tree Classifier	Naïve Bayes Classifier (Gaussian)
Time	2.07s	1.53s
Peak Memory Usage	38.84MB	27.49MB

Task 2:

Note that the Budget and Revenue columns in the given Dataset are interchanged. This was evident from checking the data for a few known hit movies where the Budget was given to be more than the revenue in the dataset. So, we interchanged the column heading names of Revenue and Budget. Then, we drop the column 'Movie Name' because it does not provide any information.Then, we change the ordinal values manually into numerical values (1 or 0 based on yes or no). 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. We make a new column hit_or_drop which is our target variable. Then we drop the column Revenue but not the Budget because Budget of a movie is decided at the time when movie is made. So, this is known beforehand.

Then we fit our Decision Tree Classifier as well as the Naïve Bayes Classifier model to the training data.

We used MultinomialNB Classifier because this set has discrete values of attributes such as Genre, Number of Screens, etc.

Evaluating our Models:

	Decision Tree Classifier	Naïve Bayes Classifier (Multinomial)
Accuracy	82.35%	81.18%
Precision	68.69%	60.53%
Recall	60.71%	82.14%
F1 Score	64.45	69.7

Both provide similar Accuracy and Naïve Bayes provides much better Recall Score so Naïve Bayes Classifier works much better in this case.

Time and Memory consumed

	Decision Tree Classifier	Naïve Bayes Classifier (Multinomial)
Time	1.22s	0.66s
Peak Memory Usage	51.72MB	25.30MB