IDA PROJECT REPORT

G11 TEAM DETAILS:

NAME	ROLL NUMBER
PRANJAY GUPTA(Lead)	S20200010169
RAHUL RAJ	S20200020296
UTKARSH VAISH	S20200020309
PARTH RAUT	S20200020298
T TEJASWANTH	S20200020306

TITLE:

Decision tree based classifier model building using CART Algorithm on Mobile Price Data.

SUBMISSION DATE:

24 November 2022

PROBLEM STATEMENT:

a) Do proper data pre-processing

b) Build a classifier model based on the CART algorithm.You should divide the data set randomly in 2:1 ratio using any random sampling method and then learn the model using the training data set.

c) Report the performance measure in terms of Confusion matrix, Predictive accuracy, F1-score, Precision and Recall in each case of your verification.Obtain an ROC curve comparing the different classifiers you have built during your model validation.

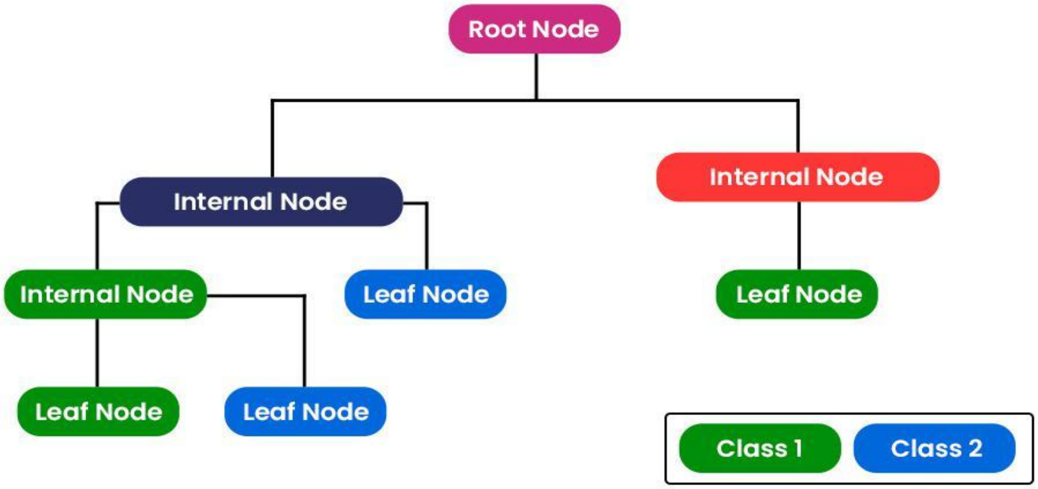
THEORY

- Initially we preprocess the dataset which includes removal of the outliers using the box plot method,then we remove the NA values from the data set and after that we remove the redundant columns using the Correlation Method.

  Correlation Analysis is a statistical method that is used to discover if there is a relationship between two variables/datasets, and how strong that relationship may be.
- After preprocessing of the data we divide the complete dataset into 2:1 ratio in training and testing

data respectively using the sampling method and building a classifier based on CART algorithm.

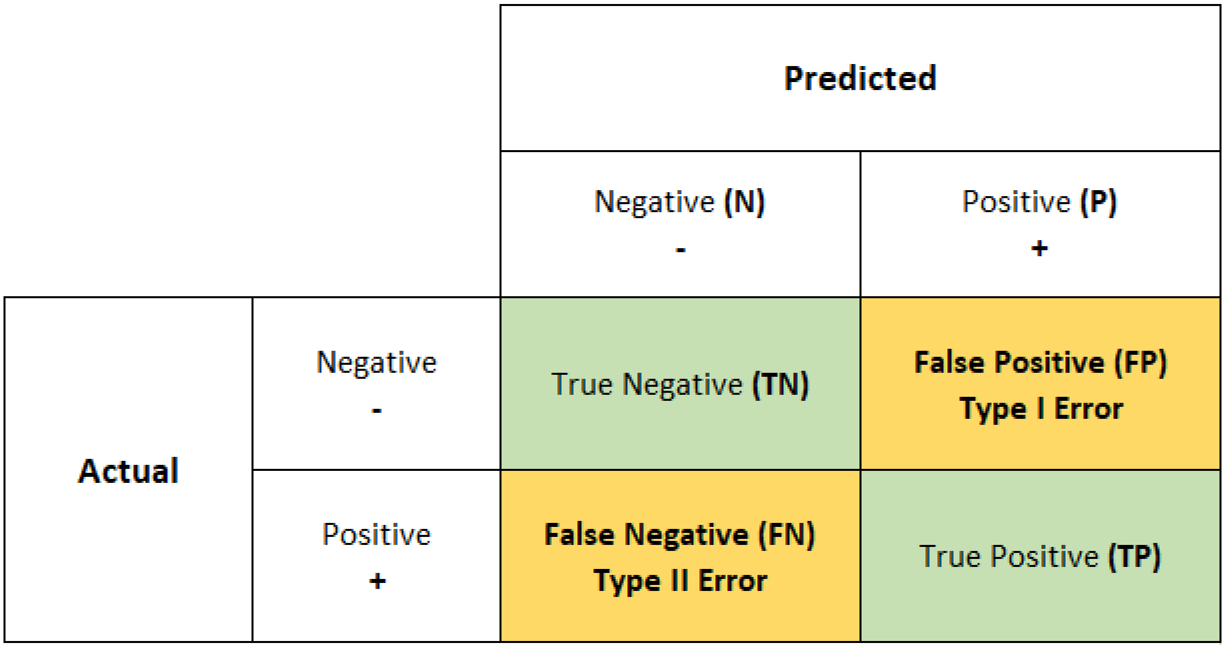
CART(Classification and Regression Tree) - CART is a predictive algorithm used in Machine learning and it explains how the target variable's values can be predicted based on other matters. It is a decision tree where each fork is split into a predictor variable and each node has a prediction for the target variable at the end.



Working of CART Algorithm

- Finally, we measure the performance report in terms of confusion matrix, predictive accuracy and F1-score.Then we make a ROC curve by comparing the different classifiers.

Confusion Matrix - A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.



		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negative (TN)	False Positive (FP) Type I Error
	Positive +	False Negative (FN) Type II Error	True Positive (TP)

confusion matrix

F1 score - The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. It is primarily used to compare the performance of two classifiers. Suppose that classifier A has a higher recall, and classifier B has higher precision. In this case, the F1-scores for both the classifiers can be used to determine which one produces better results.

Predictive Accuracy - The predictive accuracy describes whether the predicted values match the actual values of the target field within the incertitude due to statistical fluctuations and noise in the input data values.

EXPERIMENTAL RESULTS

We have worked on the Mobile Price Dataset.

Preprocessing

- As the problem is based on multiclass classification we divided the dataset into four parts as there are four classes given in the data set.
- Then for each part we remove the outliers using the box plot method and combine all the parts together
- After that we remove the NA values but as there were no NA values so not any data is removed.
- Now an exciting outcome came during correlation analysis,feature RAM was highly correlated(0.92) with the price range that is the target and no redundant features were found.
- We avoided min-max normalization as it was affecting the accuracy.

Training the DATA using CART algorithm:

- We randomly split the data set into 2:1 ratio into training and testing data respectively, then we use the **rpart** library to train and plot the model.

Measuring the Performance:

- As we are splitting the data set randomly we will get different results in every instance.

Predictive Accuracy - 83.30%

Confusion matrix (class wise) -

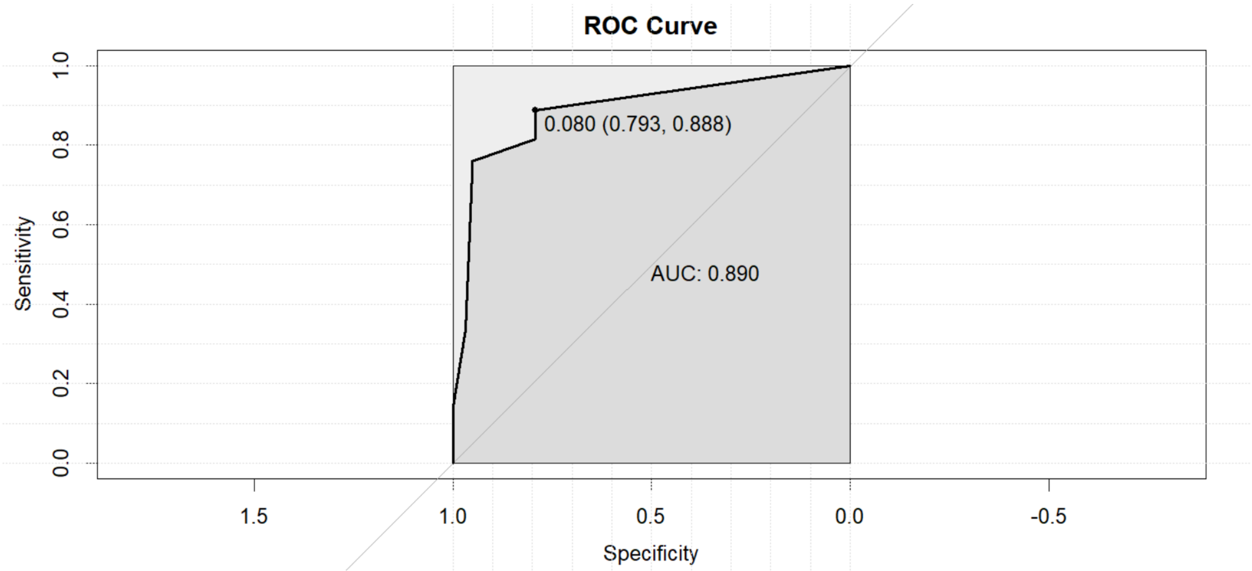
	0	1	2	3
0	156	21	0	0
1	8	91	8	0
2	0	13	98	19
3	0	0	22	109

Precision - 0.95, 0.73, 0.77, 0.85

Recall - 0.88, 0.85, 0.75, 0.83

F1 Scores - 0.91, 0.78, 0.76, 0.84

ROC curve -


ROC Curve
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.5
1.0
0.5
0.0
-0.5
0.080 (0.793, 0.888)
AUC: 0.890