PROJECT-4 REPORT

DATA ANALYTICS - CS40003

TOPIC -4.1

Roll no: 16CS30033 Name: Shaikh Moin Dastagir
Mob no: 7709400624

- Similar steps are done for both the algorithms C4.5 and CART
- Only the training function is different
- First divided the data in 7:3 ratio,trained the classifier using C4.5 algorithm
    - Used the library (RWeka)
    - Predicted the results using the training set and stored the result in predictions
    - Calculated accuracy by checking how many instance correctly classified divided by total no of instants
        - Accuracy =0.8726852
    - Used the library caret to calculate the confusion matrix and calculate F1 score , recall and precision
    - Below are the results for confusion matrix

predictions	not_recom	priority	recommend	spec_prior	very_recom
not_recom	1302	0	0	0	0
priority	0	1007	2	121	98
recommend	0	0	0	0	0
spec_prior	0	274	0	1084	0
very_recom	0	0	0	0	0

- Divided the index of the whole training set in 10 equals parts
- Then stored the indexes of the training data needed for each iteration in ten fold cross validation in index_train variable.
- Applied the algorithm on each dataset generated in each iteration
    - Trained the classifier
    - Predicted the values on test data
    - Calculated confusion matrix for each iteration (printed on the console during runtime)

    - Calculated F1_score,recall,precision and stored it to average it later after all the iterations

- After all the iterations the values calculated are shown on the console

For CART

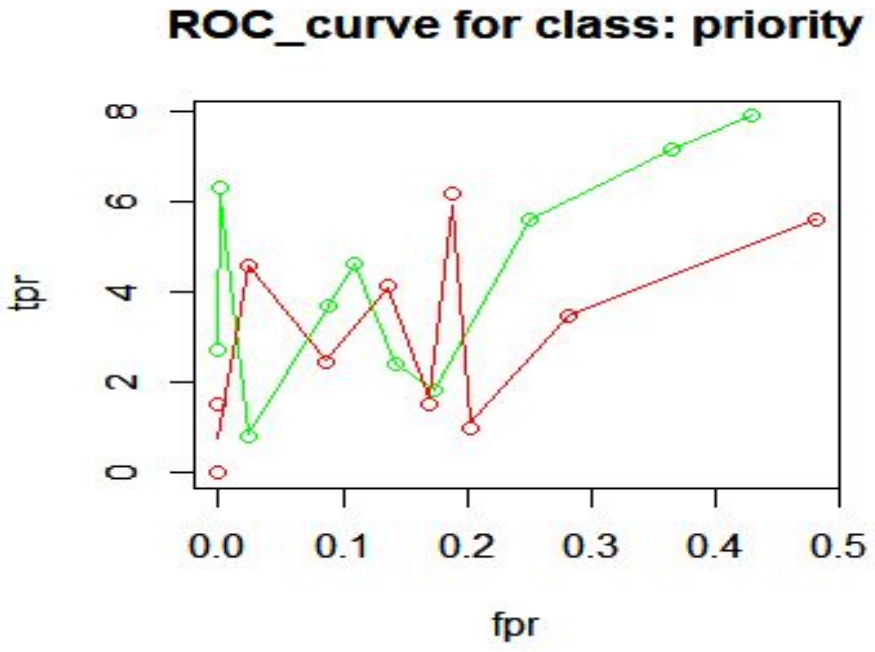
- First divided the data in 7:3 ratio,trained the classifier using Cart algorithm
    - Used the library (rpart)
    - Predicted the results using the training set and stored the result in predictions
    - Calculated accuracy by checking how many instance correctly classified divided by total no of instants
        - Accuracy =0.8726852
    - Used the library caret to calculate the confusion matrix and calculate F1 score , recall and precision
    - Below are the results for confusion matrix
        -

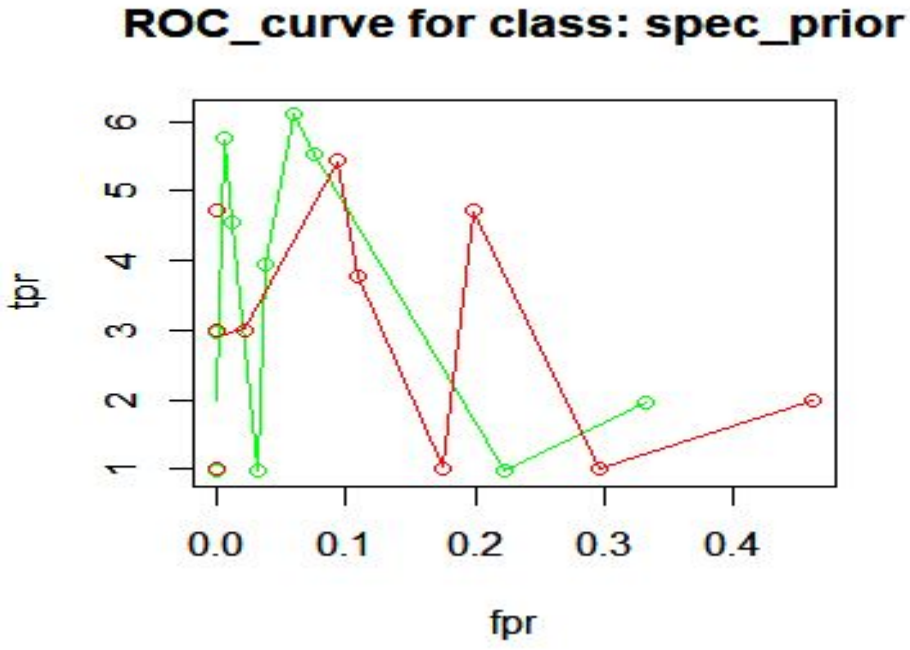
predictions	not_recom	priority	recommend	spec_prior	very_recom
not_recom	1302	0	0	0	0
priority	0	1007	2	121	98
recommend	0	0	0	0	0
spec_prior	0	274	0	1084	0
very_recom	0	0	0	0	0

- Divided the index of the whole training set in 10 equals parts
- Then stored the indexes of the training data needed for each iteration in ten fold cross validation in index_train variable.
- Applied the algorithm on each dataset generated in each iteration
    - Trained the classifier
    - Predicted the values on test data
    - Calculated confusion matrix for each iteration (printed on the console during runtime)
    - Calculated F1_score,recall,precision and stored it to average it later after all the iterations

- After all the iterations the values calculated are shown on the console

ROC curves:
TPR vs FPR





Conclusion : The green graph is from the algorithm C4.5 and the red one from CART
It can be clearly seen that the AUC under the green graph is more than that under the red graph, hence c4.5 is the better classifier for this dataset.