**ASSIGNMENT 3**

**Assigned October 15, 2024**

**Due October 26, 2024**

**On Canvas, 11:59PM**

**50 points**

**In a comment section at the top of your program list all the team members and the contribution of each team member.  There are three opportunities to submit your program.  However, only the last submitted version will be graded.**

This assignment is asking you to implement ID3 and Naïve Bayes classifiers Churn\_Modelling.csv data set from Assignment #2.

**NOTE:**

* **You may use MATLAB ML toolbox or sklearn to implement the ID3 algorithm.**
* **You may not use any toolbox to implement the Naïve Bayes algorithm.**
* **If you use python, you can use numpy, matplotlib, seaborn, pandas.**
* **For both, you may use train test split from sklearn.**
* **Make sure that the same data are used for training and testing with each model (otherwise comparing the two models does not make sense).**
* **Do not use the F1\_score and accuracy score or any other metrics from sklearn. Implement them by yourselves.**
* **For full points, all plots should be properly labelled (with axis titles, legend, plot title)**
* **DO NOT share your work. You can discuss with your group peers, but do not share your final submission or your code with other groups. No points will be awarded if we find concrete evidence of cheating**

You are to implement the following.

1) **(10 points)** The basic ID3 algorithm.

2)**(10 points)** The Naïve Bayes classifier

For computing the probabilities needed in the Naive Bayes algorithms, you can use a function that computes histograms, to output the frequency of values in a histogram bin.

Then discretize by using the bin center (rounded to the nearest integer) as the discrete value of an attribute.

Run your programs for bin numbers varying from 5 to 20 incrementing by 5.

For each algorithm:

1. **(10 points)**Compute the accuracy for each discretization. Compute *Max\_accuracy*, *min\_accuracy*, and *average accuracy* over the different number of bins.  Plot **on the same figure** (with different colors) these accuracy values as a function of the number of bins.

Some pointers:

* Make 5 train-test splits of the same data (with test size 33% and random split). For each split, compute the TEST accuracies for each bin size ([5:20:5]) and print them out
* Output format:

Text

Description automatically generated

* For the plots: Simply plot the accuracies for each bin size (you will get 4-line plots, plot them on the same figure.
* You will have to do this for decision trees and for the Naïve Bayes.
* Note: The dataset is binary classification of which we do not know if it is linearly separable. Therefore, we cannot anticipate (almost) perfect accuracies.

2. **(5 points)**Compute the F measures (refer to my ROC ppt slides) for each discretization. Plot the F-measures over the different number of bins.

Some Pointers:

* Repeat similar steps, but this time, calculate TEST F1 score
* You do not need to print out the F1 scores
* Plot the F1 scores for all 4 bin sizes on the same plot

3. **(5 points)**Compute the ROC points for each discretization. Plot the ROC curves over the different number of bins.

Some pointers:

* Repeat similar steps, but this time, calculate TEST FPR and TPR to construct the ROC Plot.
* You do not need to print out the FPR and TPR scores
* Plot the ROC curves for all 4 bin sizes on the same plot

4. To further compare the two algorithms, assume one algorithm is the ground truth

Some pointers:

* By this we mean that use the TEST predictions from one model as the y\_true (test) for the other model
* For instance, y\_test\_true (DT) = NB(X\_test) and vice versa

4.1 **(5 points)**Compute the F measures (refer to my ROC ppt slides) for each discretization, when the ID3 is assumed to be the ground truth. Plot the F-measures over the different number of bins.

Some pointers:

* Feed X\_test to ID3, get y\_preds (this will be y\_test\_true for NB)
* Like Q2 (Get F1 score plot for NB)

4.2 **(5 points)**Compute the F measures (refer to my ROC ppt slides) for each discretization, when the Naive Bayes is assumed to be the ground truth. Plot the F-measures over the different number of bins.

Some pointers:

* Feed X\_test to NB, get y\_preds (this will be y\_test\_true for ID3)
* Like Q2 (Get F1 score plot for ID3)

**LIST OF DELIVERABLES:**

1. **Decision Trees:**

* Accuracy scores for each discretization and each bin size (check sample output)
* 1 Accuracy Score plot
* 1 F1 Score plot
* 1 ROC Curve
* 1 F1 Score plot with Naïve Bayes as Ground Truth

1. **Naïve Bayes:**

* Accuracy scores for each discretization and each bin size (check sample output)
* 1 Accuracy Score plot
* 1 F1 Score plot
* 1 ROC Curve
* 1 F1 Score plot with Decision Tree as Ground Truth

**SUBMISSION:**

1. **Report (.pdf):**
   * **Include group member names and for each the contribution to the assignment in percentage.**
   * **Attach/include screenshots of output/plots**
2. **Source code (.zip)**

**NOTE: Please turn in the report and source code as two separate attachments on canvas.**