Bryan Guzman
Homework 3
CSI 431 – Data Mining
ID: 001265918

A note to the grader:

My program allows for some command line arguments that I was using for myself. If no argument is provided, the percentages for scores and cross-validation will be printed by default along with the graphs. I'll provide a list of the commands it takes (it isn't many):

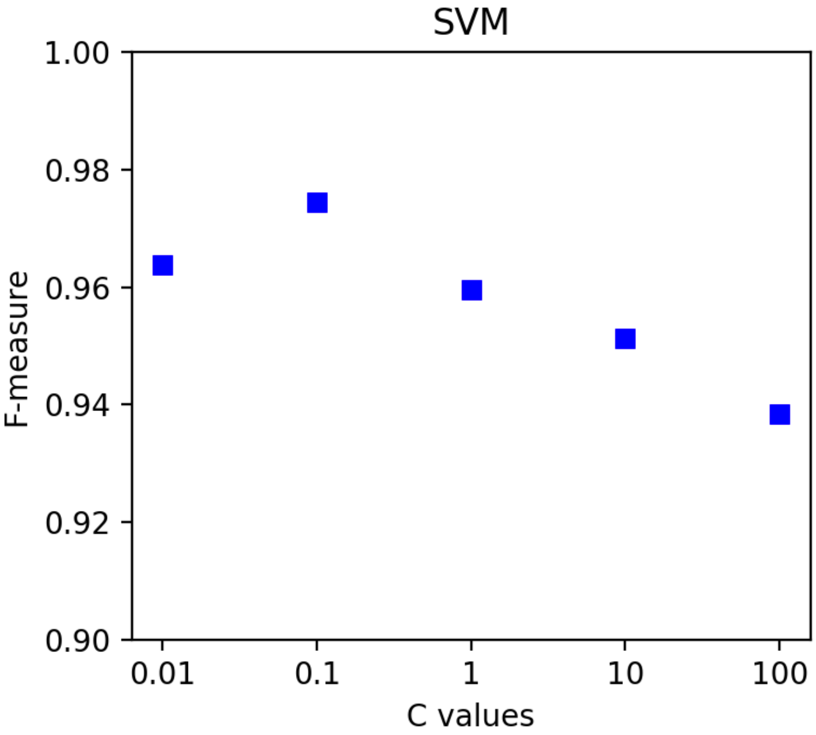
standardized: will run the program with a dataset that is standardized. Does this anyway when no arguments are provided.

raw: will run the program with the raw dataset. No standardization or centering will be ran on the dataset. I mostly did this to see if there would be any difference in the values produced. I also found out how important it is to standardize your dataset for performance.

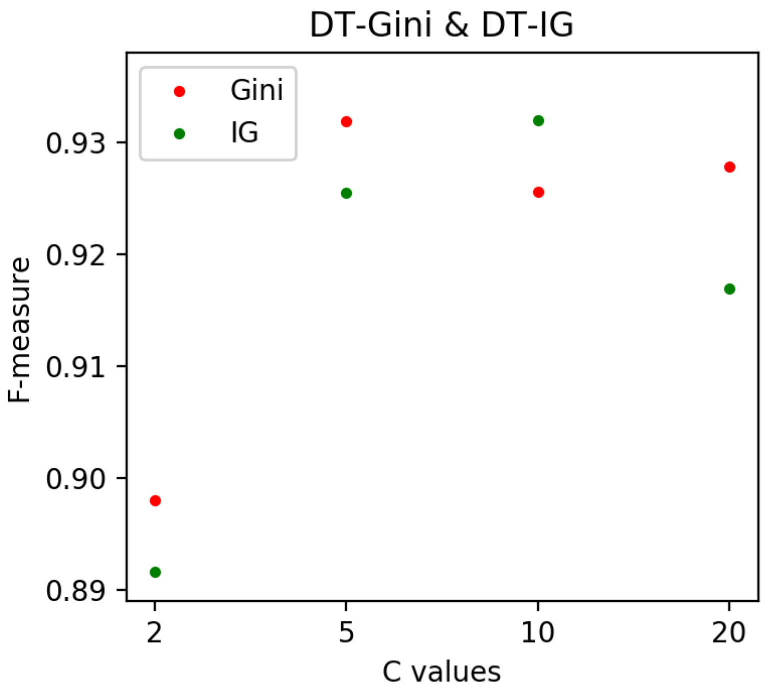
print_pred: will print out the predicated classes for a standardized dataset.

raw print_pred: will print out the predicated classes for a raw dataset.

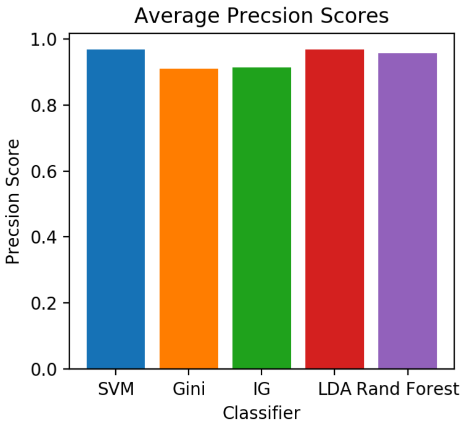
1.) In the dataset provided to us, when the C value is greater than or equal to 1 the less accurate our predictions will be according to 10-fold cross validation. When our C value is set to 1, an f-measure of 95.96% which is still really good, but when C is set to 0.1 the best f-measure, 97.44% is produced. Observing that when C is set equal to 0.01 an f-measure score of 96.38% is produced, it can be assumed that a C value of less than 1 will suite our dataset best which will produce a larger margin.

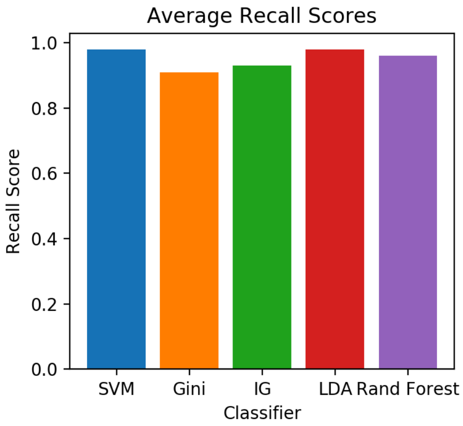


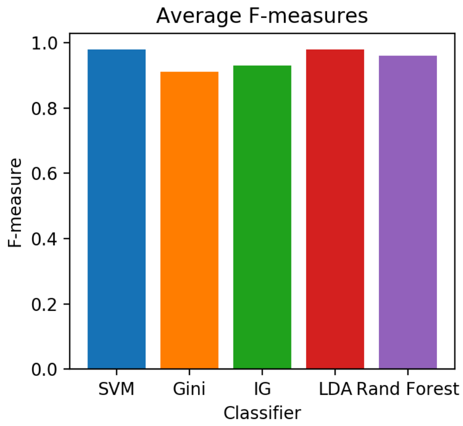
2.) For this dataset, it seems that the best tree size for it is between 5 and 10 leaf nodes. When leaf nodes are set to 5 an f-measure score of 93.19% for Gini and 92.55% for information gain and at 10 leaf nodes the f-measure score is 92.55% and 93.19% for gini and information gain are produced. These are the highest scores achieved for gini and information gain. When leaf nodes were set to 20, the f-measure for Gini was higher than the f-measure for 10 leaf nodes. 20 leaf nodes produced an f-measure of 92.78%. For this dataset, setting the leaf nodes equal to 2 should be avoided given that these produced f-measure scores that were much lower than any other f-measure. Both gini and information gain f-measure fell to 89.80% and 89.16% respectively. For this dataset I set the leaf node value for gini to 5 and information gain to 10 since at those values for max leaf nodes, gini and information gain produced the highest f-measure and would provide the most accurate classification for the data.



3.) Each score that was plotted to the bar graph were averages of scores over the whole dataset. From these bar plots, SVM and LDA are the two best classifiers for this dataset. SVM achieved a precision score of 96.88%, a recall score of 98.00% and an f-measure score of 97.99%. All three of LDA's scores were identical with SVM and either one would be suitable for this dataset at this size. The third best classifier was the decision tree using gini and the decision tree using information gain. The decision trees scores were as follows, precision for information gain and gini are 91.33% and 92.12%, the recall scores are both 93.00% and the f-measure scores are 92.94% and 92.98 respectively for all values. The decision tree using gini scored higher in all scores except the recall score where both had an equal score. Gini gets a slightly higher score than information gain in f-measure meaning that overall it is slightly more accurate, but there should not be any noticeable differences unless the dataset is scaled up to a large size. When including the random forest classifier, it becomes the third best classifier and pushes gini and information gain to the 4th and 5th best classifiers for this data set. The precision score for Random Forest Classifier is 95.04%, recall score is 95.00% and f-measure is 95.01%. The winner here is SVM and LDA, scoring close to perfect scores across the board. While SVM and LDA may maintain a higher accuracy over the other classifiers, when the dataset is scaled up, it can't be said that they will always be the best. If we are only talking about this dataset and it will remain unchanged, SVM and LDA can be classified as the "winners", however, if the dataset was going to be scaled up, the winning classifier could not be determined since it depends on the type of data added.


Average Precsion Scores


Average Recall Scores


Average F-measures