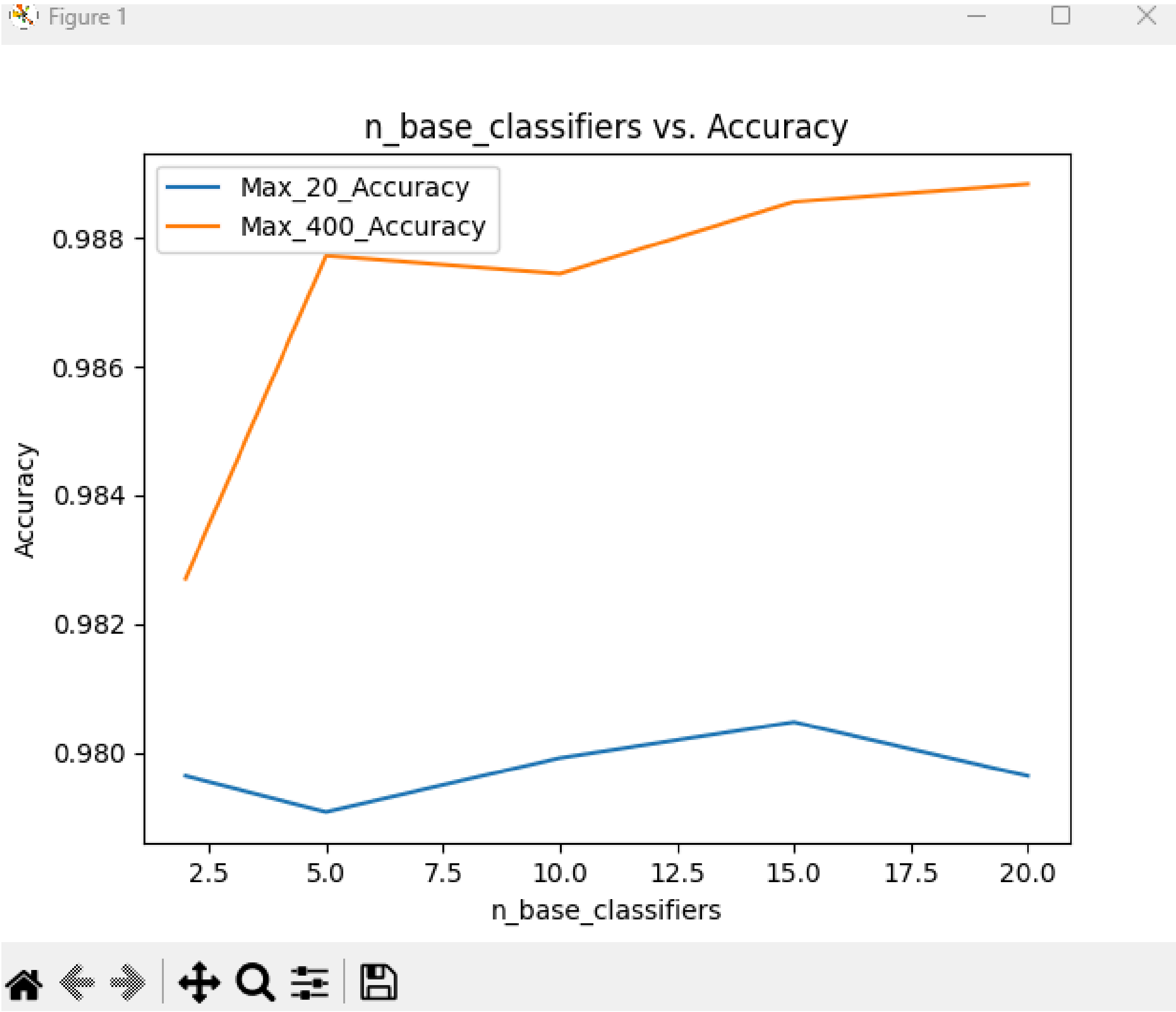
Jeffrey Contreras
Eric Campillo 6242754
David Heller

CAP5771 - Project #2

1. What did every member of the team do for this project? Try to be as specific as possible. The work needs to be balanced between the members of the team.
    a. Jeffrey Contreras - Worked on most of the predict_DT_9 and predict_DT_bagging_9 code.
    b. Eric Campillo - Worked on all parts of the code and on typing the document. Worked especially on the output of validation and test results of recall, precision, F1 Measure, and accuracy for each.
    c. David Heller - Worked on most of the predict_DT_randomforest_9 code.
2. Include the statistics of the data and the training/validation/test sets that you used.
    a. Statistics of the data included in Dataset_train.csv: Total number of instances: 22400 and the total number of features: 30.
    Statistics of the data included in Dataset_test.csv: Total number of instances: 5600 and the total number of features: 30.
    b. Statistics of the training sets included 17920 instances. Data was split as 80% training.
    c. Statistics of the validation sets included 4480 instances. Data was split as 20% validation.
    d. Statistics of the test sets included 5600 instances.
3. Include the plots #1-#3 generated. Comment on each plot. What do you observe?
    a. Plot 1:

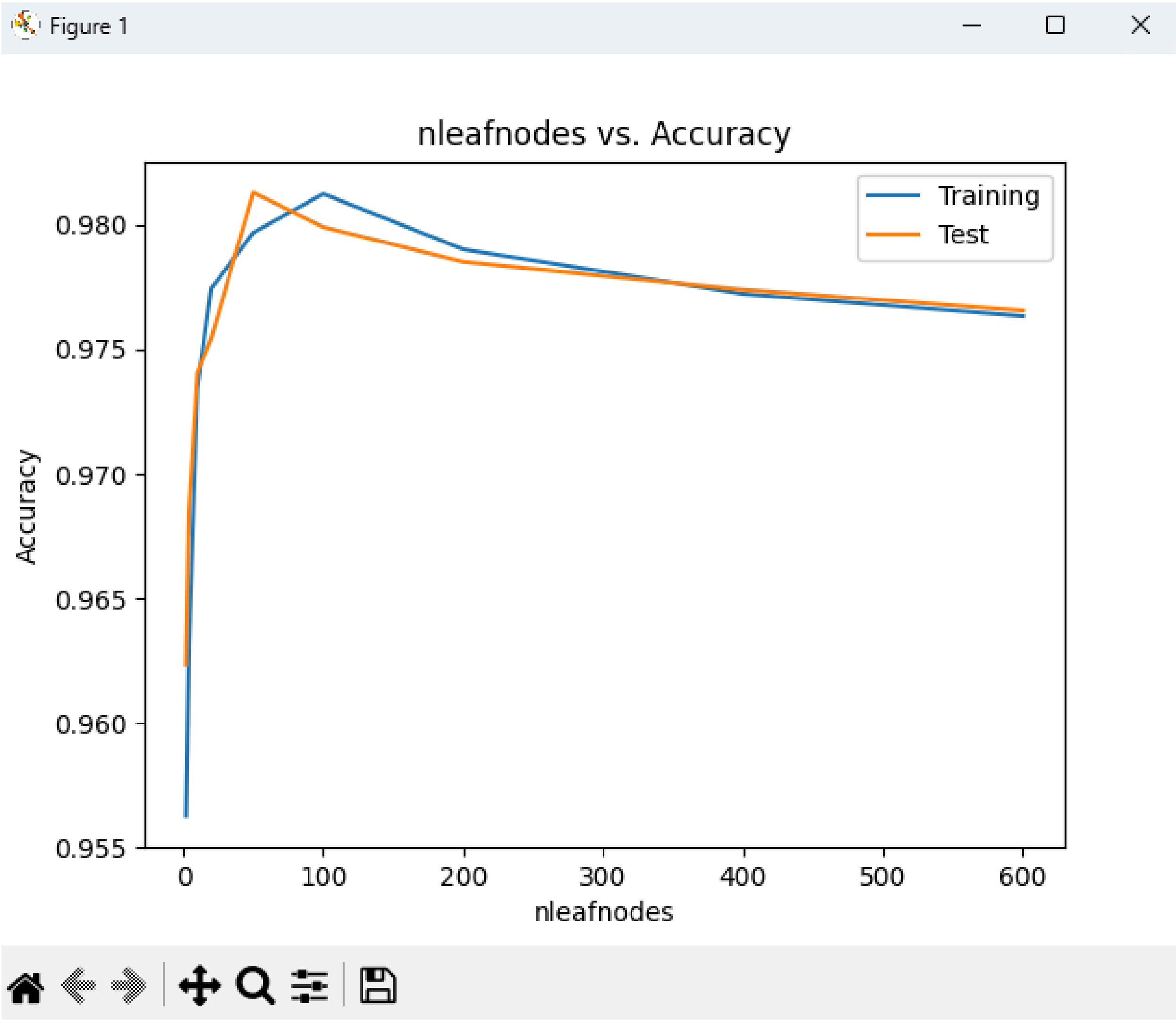


b.

i. Comment 1: It appears that the Max_400_Accuracy displays an accuracy that is higher than the Max_20_Accuracy. This would mean that the model is most accurate during the Max_400_Accuracy and it has a correlation there.

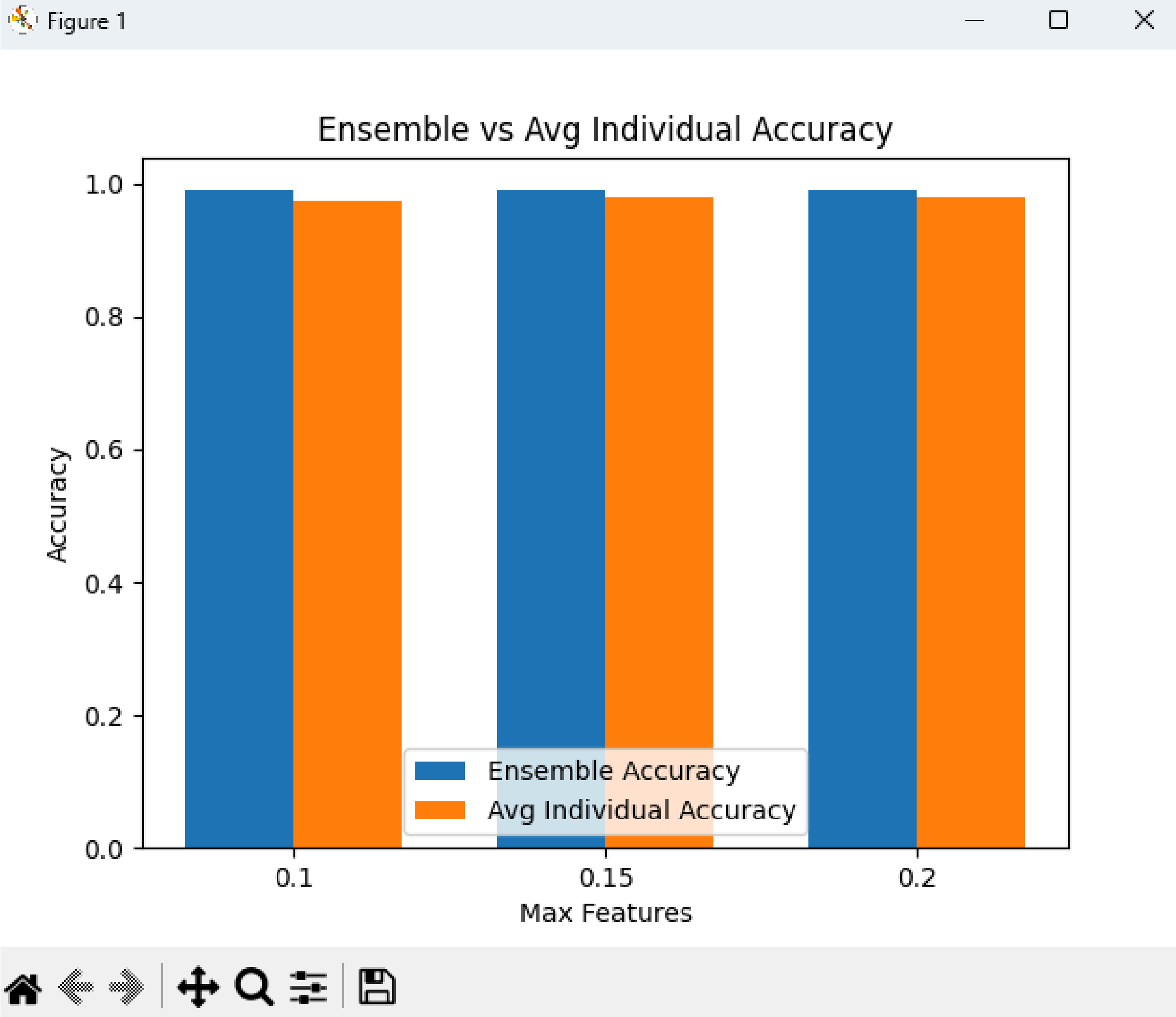
c. Plot 2:

d.



i. Comment 2: In the plot above, the training and test set are both very similar in accuracy until the point where the leaf nodes reach around 20. Afterwards the training set is the most accurate until the test set becomes more accurate and then it flips back to the training set being more accurate for most of the plot until the leaf nodes reach 350 where the test set becomes the most accurate again.

e. Plot 3:



f.

   i. Comment 3: In the pictured bar graph above, the ensemble accuracy is always higher than the Average Individual Accuracy meaning Ensemble Accuracy is the most accurate during these metrics with the Max Features vs Accuracy.

4. Include the tables described in step5. Bold or highlight the model that performs the best for each metric.

   a. DT Bagging
      i. Training: 17920 instances
      ii. Test: 5600 instances
      iii. Recall: 0.9681818181818181
      iv. Precision: 0.9941176470588236
      v. Accuracy: 0.9888392857142857
      vi. F1-Measure: 0.976878612716763

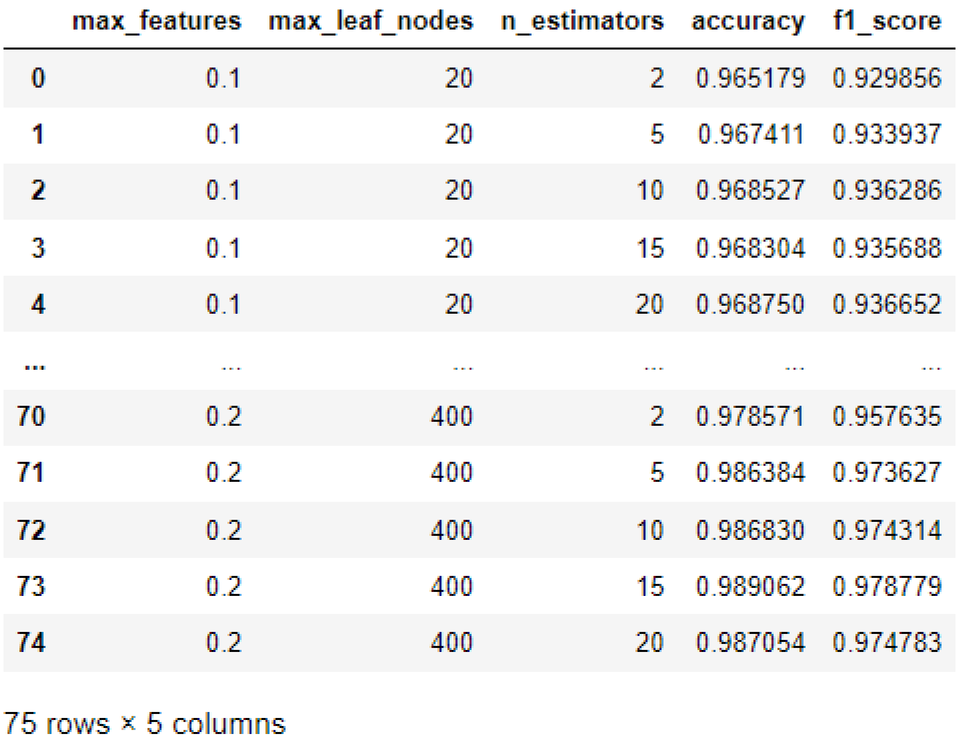
b. DT
  i. Training: 17920 instances
  ii. Test: 5600 instances
  iii. Recall: 0.9715909090909091
  iv. Precision: 0.9392935982339956
  v. Accuracy: 0.9821428571428571
  vi. F1-Measure: 0.9633867276887872

c. DT random forest
  i. Training: 17920 instances
  ii. Test: 5600 instances
  iii. Recall: 0.9628571428571429
  iv. Precision: 0.9963045084996305
  v. Accuracy: 0.9898214285714285
  vi. F1-Measure: 0.979295314202688

5. For each dataset, create a table with the validation and test accuracy, as well as the other test metrics of the best performing model for each classification method. Bold or highlight the model that performs the best for each metric. Comment on the results achieved.

a. **Best Model: Random Forest**
b. **max_features 0.150000**
c. **max_leaf_nodes 400.000000**
d. **n_estimators 15.000000**
e. **accuracy 0.990402**
f. **f1_score 0.981393**
g. **precision 0.996485**

The Random Forest achieved an accuracy of 99.04%, indicating it correctly classified the vast majority of transactions. Moreover, it scored an F1 score of 98.14%, which suggests a well-balanced precision and recall, crucial for the context of fraud detection where the cost of false negatives and false positives can be high. Additionally, a precision of 99.65% means that nearly every prediction of fraud made by the model is correct, minimizing false alarms which is vital for practical applications. These metrics collectively justify the selection of the Random Forest as the best model for detecting fraudulent transactions from the dataset.



	max_features	max_leaf_nodes	n_estimators	accuracy	f1_score
0	0.1	20	2	0.965179	0.929856
1	0.1	20	5	0.967411	0.933937
2	0.1	20	10	0.968527	0.936286
3	0.1	20	15	0.968304	0.935688
4	0.1	20	20	0.968750	0.936652
...	...	...	...	...	...
70	0.2	400	2	0.978571	0.957635
71	0.2	400	5	0.986384	0.973627
72	0.2	400	10	0.986830	0.974314
73	0.2	400	15	0.989062	0.978779
74	0.2	400	20	0.987054	0.974783

75 rows × 5 columns