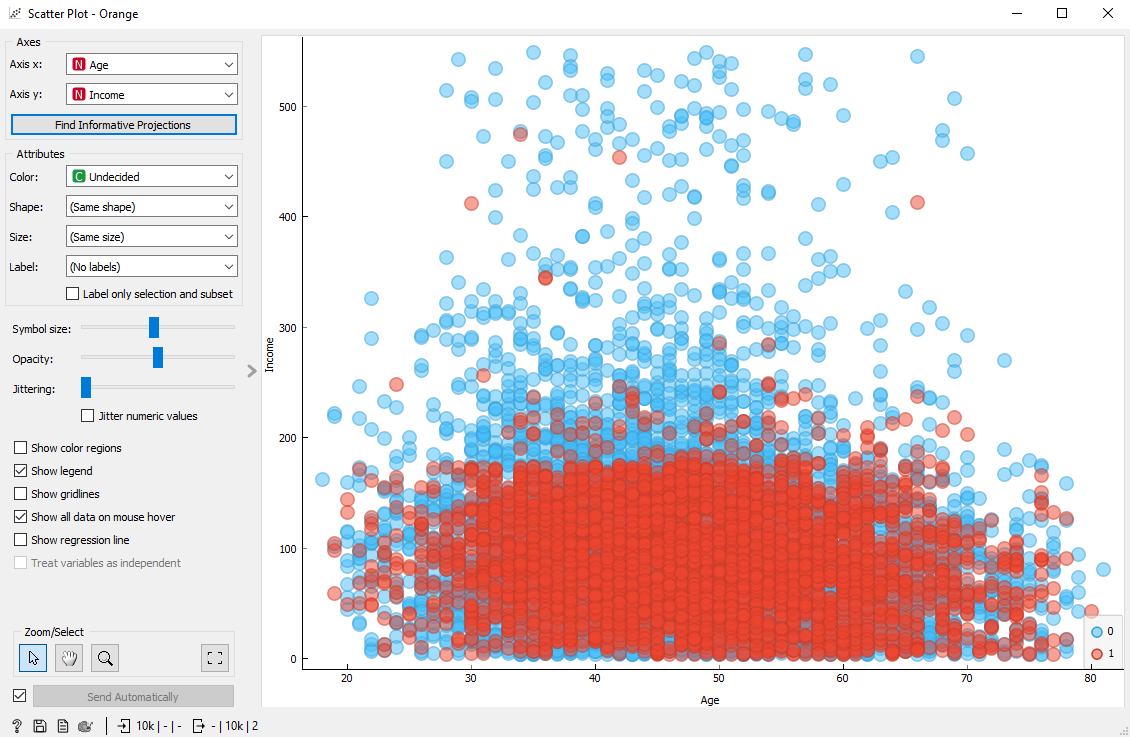
Chris Arancio

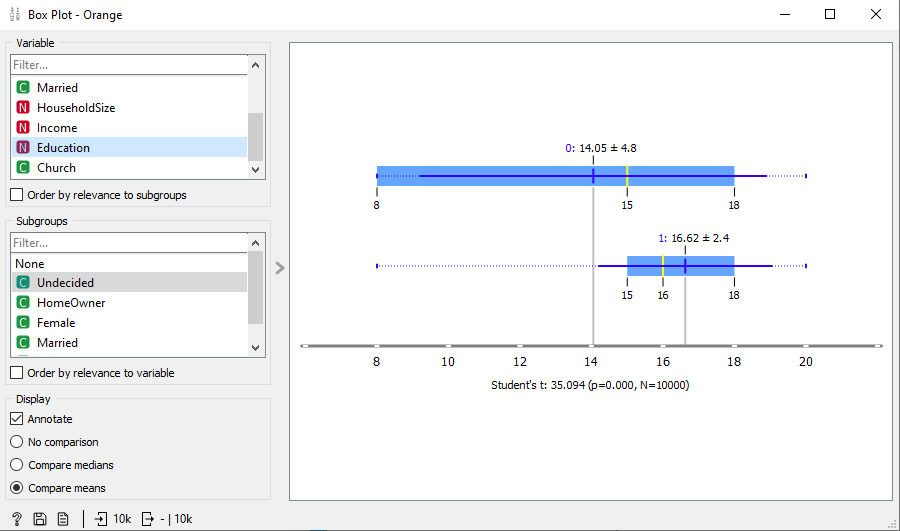
**Modeling and Predicting Voters using Voters.csv Dataset**

# **Part 1:**

Visualization #1: This scatterplot maps age on the x-axis, income on the y-axis, and colors the points by whether they are undecided. It appears that the vast majority of undecided voters have incomes under 200k, no matter the age.



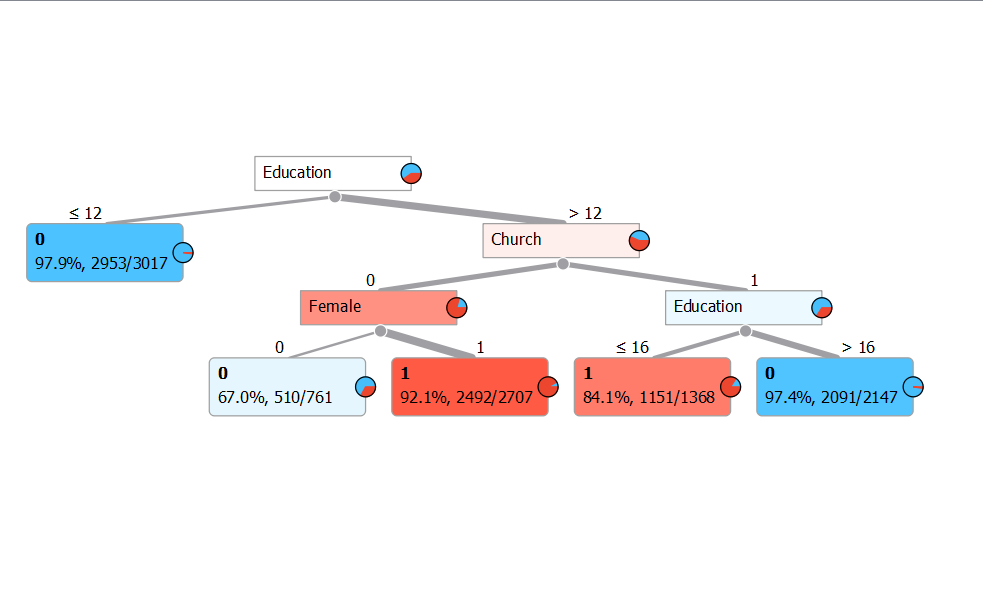
Visualization #2: This boxplot maps years of education and splits it based on those who are decided and undecided. It appears that undecided voters have more years of education on average.



Short Summary: The dataset has 10,000 datapoints and nine different variables that are measured. There are four numeric variables including Age, HouseholdSize, Income, and Education. There are five categorical variables including Undecided, HomeOwner, Female, Married, and Church. While doing an exploratory data analysis, I noticed three trends. First, that the vast majority of undecided voters have incomes under 200k, no matter the age. Second, that undecided voters have more years of education on average. And third, that the majority of undecided voters (69%) do not regularly attend church services. There were slight differences when comparing undecided voters with the Female, Married, and HouseholdSize variables, but I didn’t think that they were significant enough to be conclusive one way or the other.

# **Part 2:**

Classification Tree:

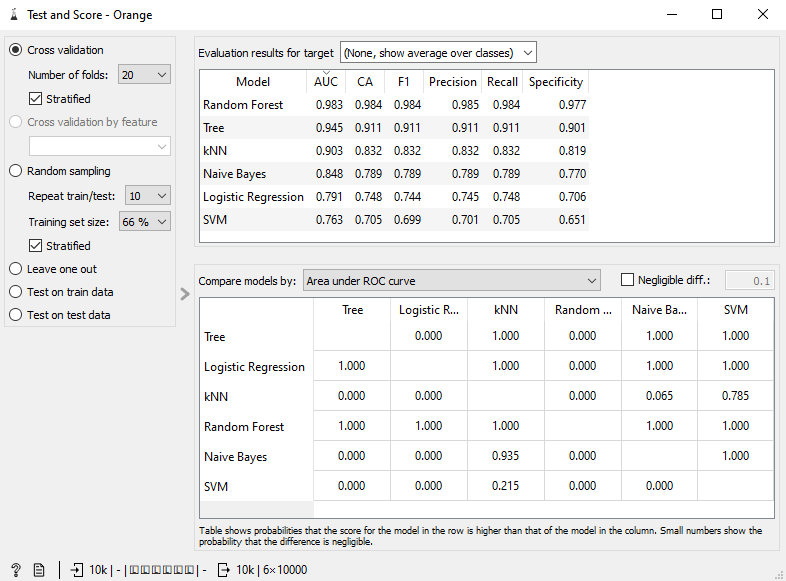


Classifications:

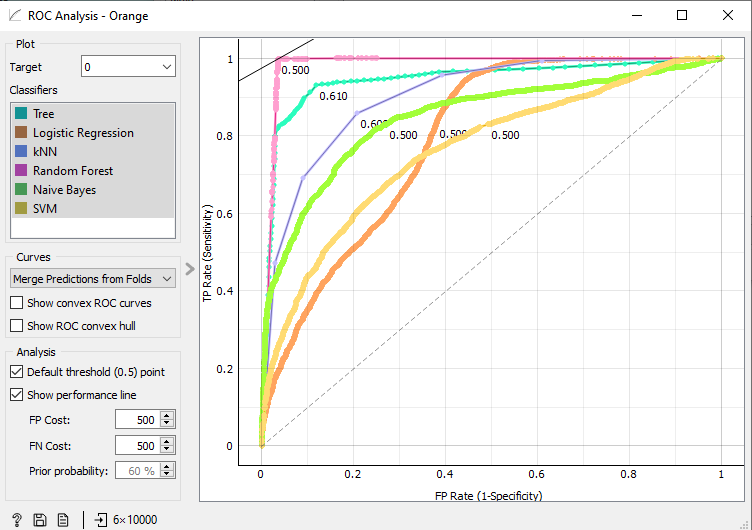
* A female 22-year-old voter that is not married that is not a homeowner that has 16 years of education and does not regularly attend religious services whose household size is 1 and annual income is 125,000.
  + **Answer:** This tree would classify this person as 1/Undecided. It would go right, left, and then right to end up in the 1 leaf with a 2492/2707 distribution of 1/Undecided.
* A non-female 40-year-old voter that is married that is a homeowner that has 12 years of education and does not regularly attend religious services whose household size is 2 and annual income is 65,000.
  + **Answer:** This tree would classify this person as 0/Decided since they have <=12 years of educations. This is the first left leaf and a distribution of 2953/3017 of 0/Decided.

# **Part 3:**

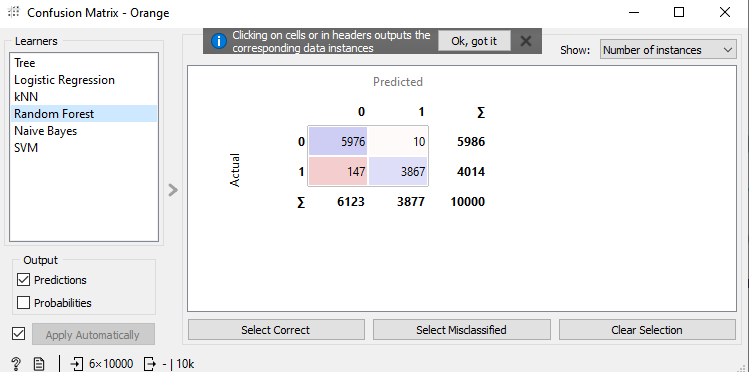
Test and Score of Classification Models: When sorted by AUC, it appears that the Random Forest performs the best. In fact, across all of the metrics shown, the random forest model scored the best with values closest to one.



ROC Analysis: The Random Forest again has the best value and has closest shape to the ideal “upside-down L.”



Confusion Matrix: Again, the Random Forest has the best results in the confusion matrix. It only misclassified 157 instances, when all of the other models misclassified several hundreds.



Changing Settings of KNN, Classification Tree, and Random Forest:

* KNN: It appears that the highest AUC scores appear between k=4 and k=6. Coincidentally, the default of k=5 has some of the best scores across the board. However, if you value slightly higher precision over specificity, then k=4 would be the best choice. Interestingly, k=6 is tied for the best AUC with k=5, but it also has worse scores across the board in the other metrics.



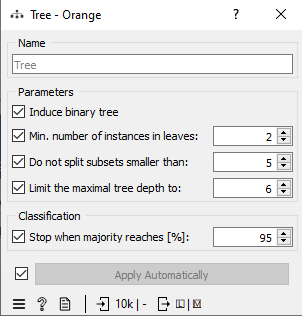
k=4:



k=6



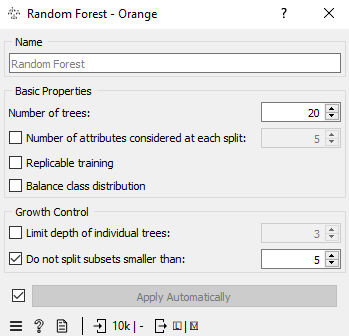
* Classification Tree: It appears that changing the maximal tree depth to 6 increase the metrics in the Test and Score across the board. It also appears that the scores do not increase when increasing the maximal tree depth past 6. Changing the other settings to do seem to increase the scores in Test and Score by very much. To note, increasing the classification majority to 100% improves the scores by .001, but also makes the tree much bigger and complicated. I think that changing the maximal tree depth to 6 is the best change to optimize the classification tree.







* Random Forest: For the random forest, I tried changing the number of trees. The results for 100, 200, and 1000 trees is the same as just 20 trees. The increases in the Test and Score metrics is also minimal (+0.001 in most cases) when compared to the default 10 trees. Therefore, I would recommend using 20 trees, but 10 trees does about the same job.



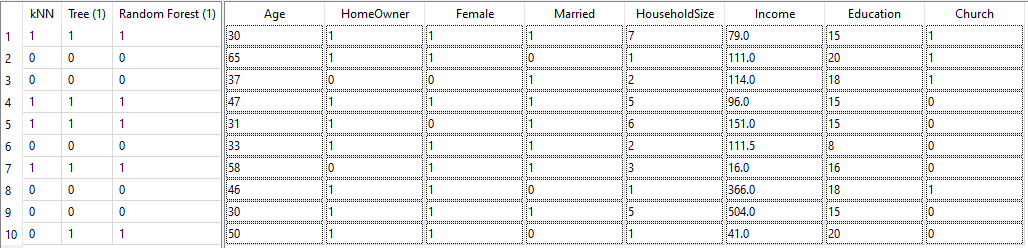




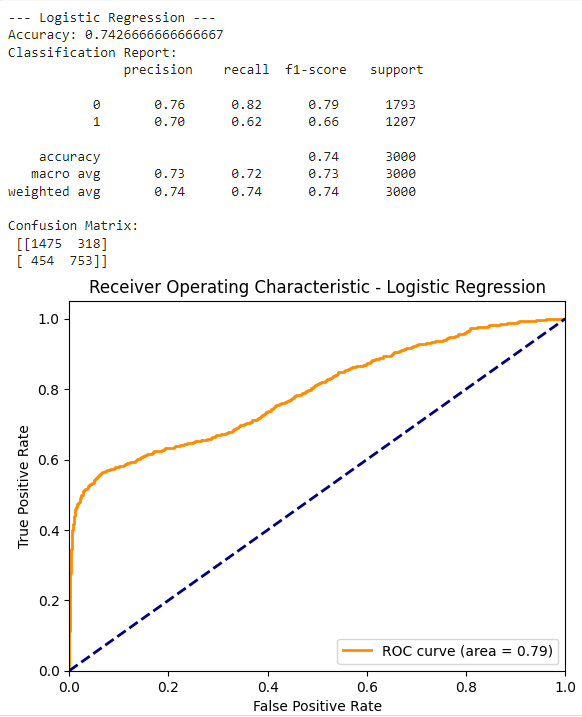
## **Part 4:**

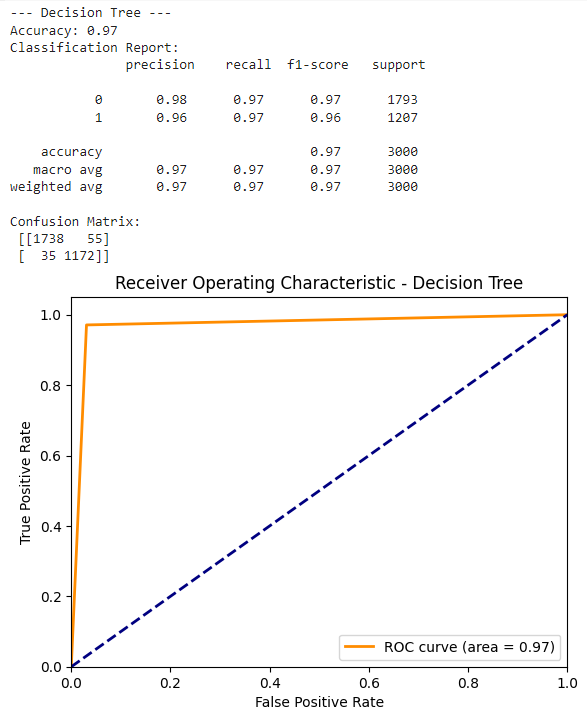
3 Classification Models to Predict: I recommend the highest AUC’s (and other metrics as well): Random Forest, Classification Tree, and KNN.

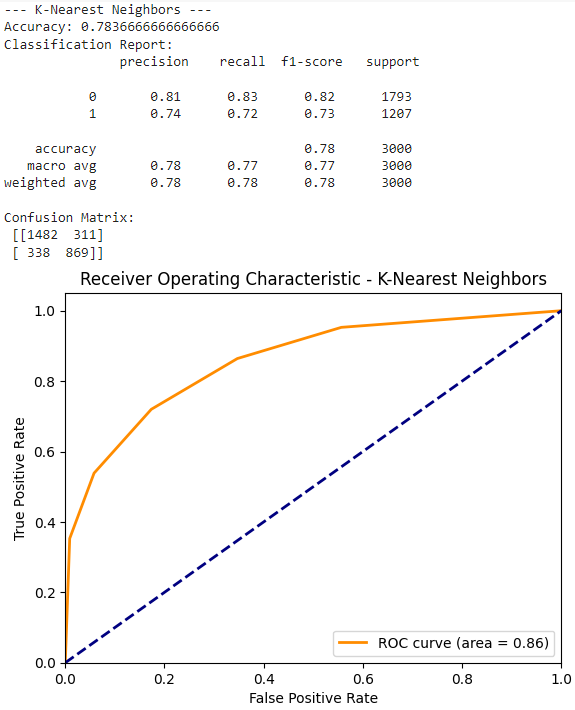
Predictions: The models performed very well. In fact, they all agreed and were correct for 9/10 entries. On the last voter, the KNN model predicted the wrong value, so I would have to recommend either the Tree or Random Forest as the best models. While the Random Forest performs slightly better than the Classification Tree (with adjusted settings) according to the Test and Score metrics, the Tree provides a clear visual of how the model decides whether or not a voter will be undecided. My final recommendation would be the Tree because it performed just as well in our test and allows for a clear diagram of how it is making its decisions. The tree would also be able to provide insights into what factors cause a voter to be undecided.

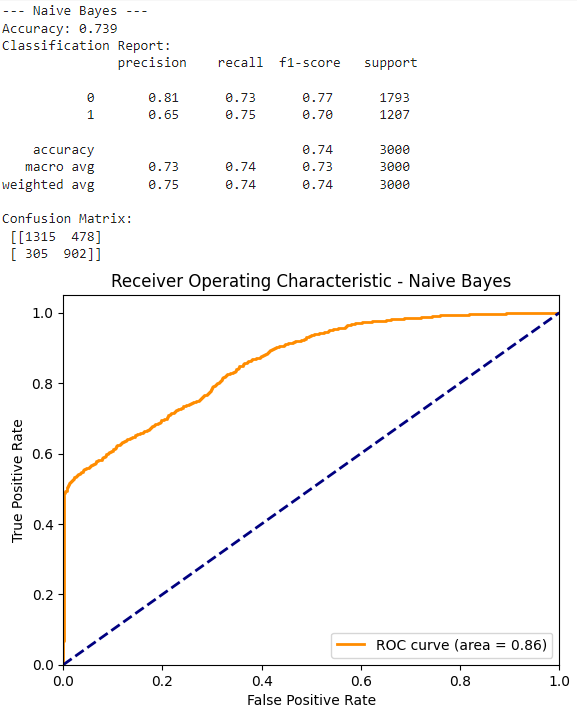


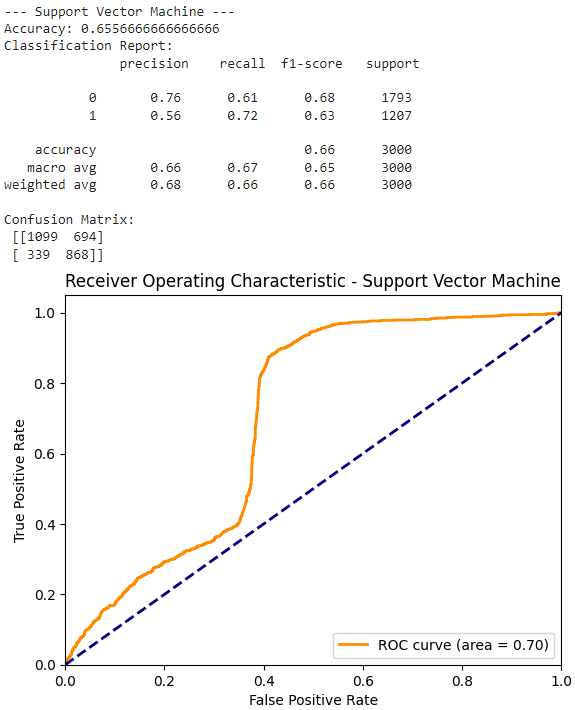
# **Part 5:**

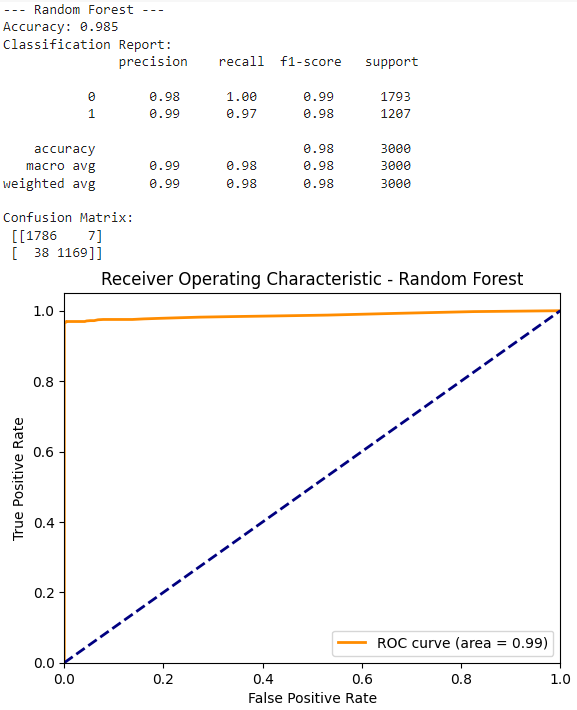


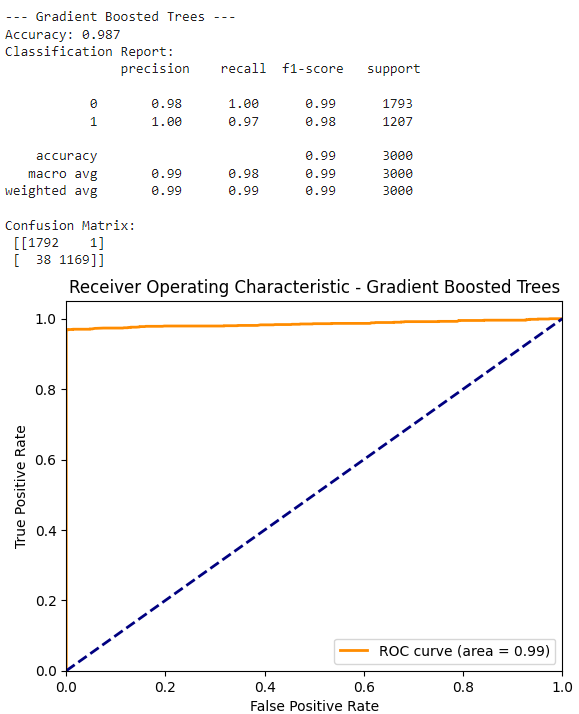


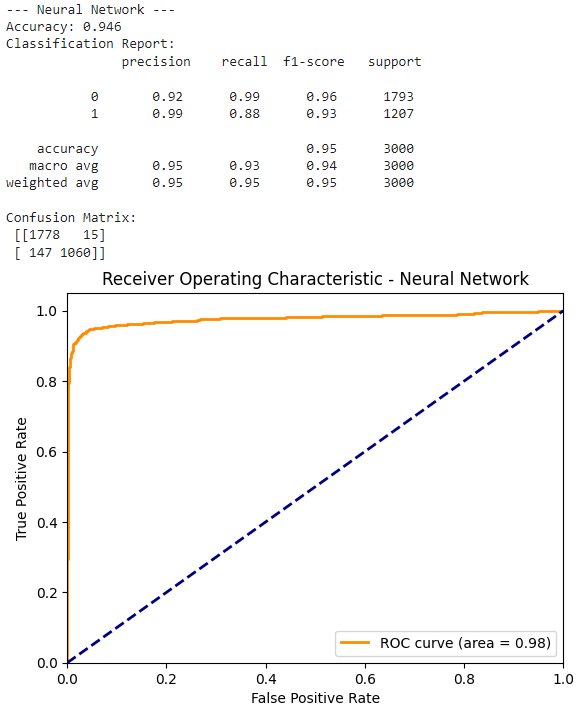


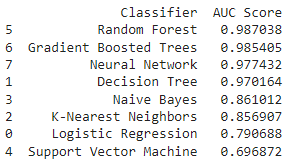












**Final Summary**:

If I add the Gradient Boosted Trees and Neural Network models to Orange, the new top three models in Orange and Python become Random Forest, Neural Network, and Gradient Boosting. The Random Forest performed the best overall and appears to be the best model for this dataset. It also has the benefit that it can also be visualized in a Pythagorean Forest, if desired.

