**1. Train a decision tree classifier on ALL of the data and set the depth limit at 5. Use the toString method to print out the learned tree. Look at the words. Do they make sense? Do any of them stand out? Include the tree along with 2-3 sentences describing what you see.**

The words make sense, since they mostly seem to be either names of wines or qualities (grape type, color, etc.) associated with the wines. Some that don't make sense are ones like "any" and "a".

(chardonnay

(pinot

(blanc

(zinfandel

(merlot

predict=1.0

predict=7.0)

(sauvignon

predict=10.0

predict=15.0))

(sauvignon

(a

predict=16.0

predict=17.0)

(young

predict=0.0

predict=15.0)))

(grigio

(gris

(brut

predict=2.0

predict=6.0)

predict=3.0)

(valley

predict=4.0

predict=15.0)))

(pinot

(bubbles

(brut

(sparkling

predict=5.0

predict=6.0)

predict=6.0)

predict=6.0)

(syrah

(any

(aromatics

predict=6.0

predict=5.0)

predict=4.0)

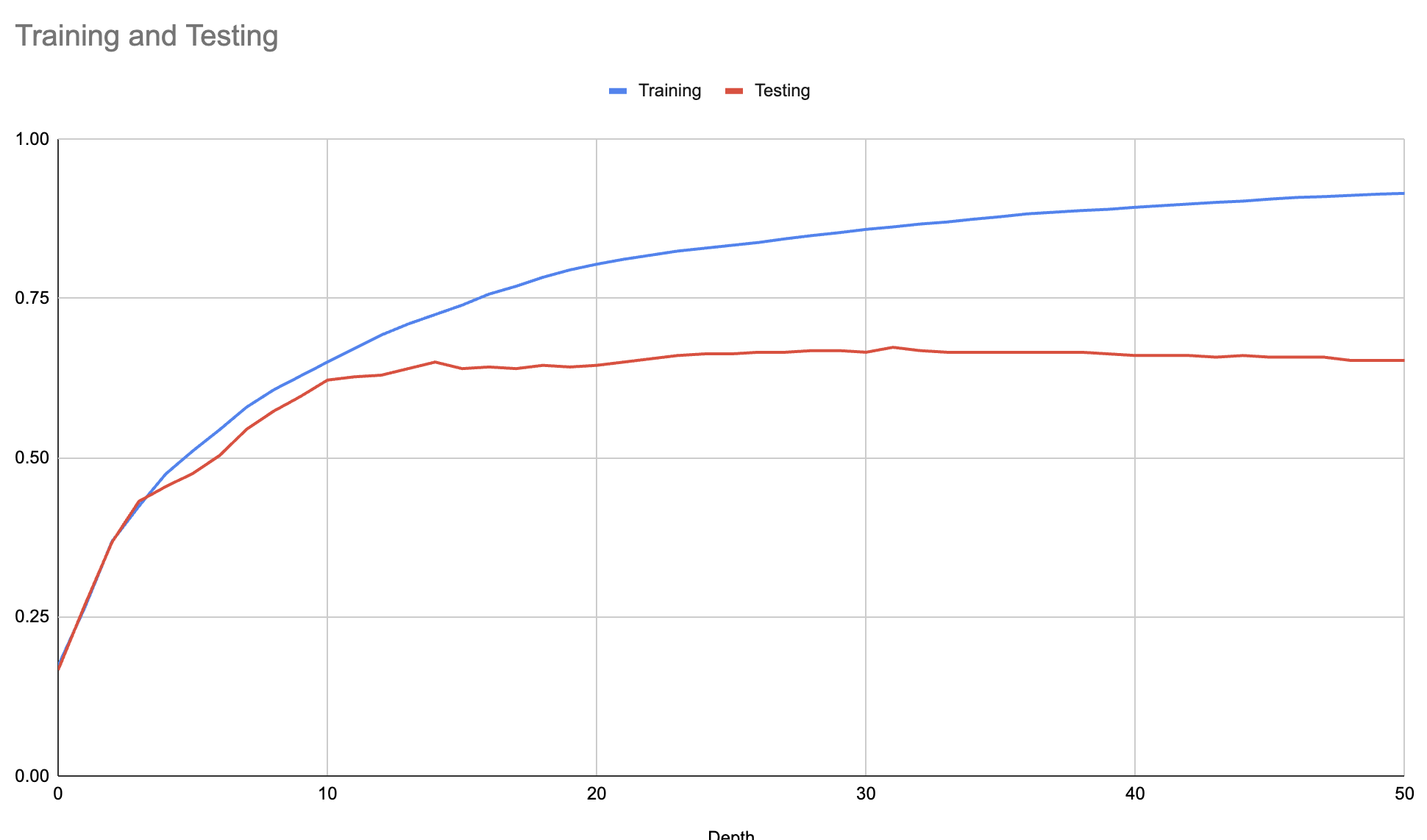
predict=2.0)))

**2. If you just predicted the majority class, what would the accuracy be? This is often a reasonable baseline to measure against (hopefully we can beat this!).**

Cabernet-Sauvignon appears 335 times. There are 1945 examples. 335/1945 = 0.172. So the accuracy would be 17.2%.

**3. On a random 80/20 split of the data, train the decision tree classifier and evaluate both the training and testing accuracy for depth limits ranging from 0 to 50 (for consistency, all tests should be run on exactly the SAME split). Include the output of your results in either a graph or table. What is the best depth to use? Do you see evidence of overfitting?**

The best depth to use is 31 because this gave the highest testing accuracy of 0.673. There is evidence of overfitting because the training accuracy is less than this for all higher depths, and the testing accuracy is less than this for all depths greater than 31.



**4. On a 10-fold cross validation of the wine data set (use the same 10-fold split for all three), calculate accuracies for the following:**

**• OVA with decision trees sized 1, 2 and 3.**

**• AVA with decision trees sized 1, 2 and 3.**

**• Multiclass decision tree (i.e. just by itself) with your best limit found above.**

**Put these all in a spreadsheet or other format (you’ll have 70 numbers, 10 for each experiment). Run a t-test to validate which approach is best. What is the best approach? Is this surprising? Include your table of results and a short write-up describing your results.**

T-test answers:

We performed t-tests on two pairs of results: one with similar means and another with significantly different means. For the pair with close means, the t-test yielded a value of 0.1215748891,

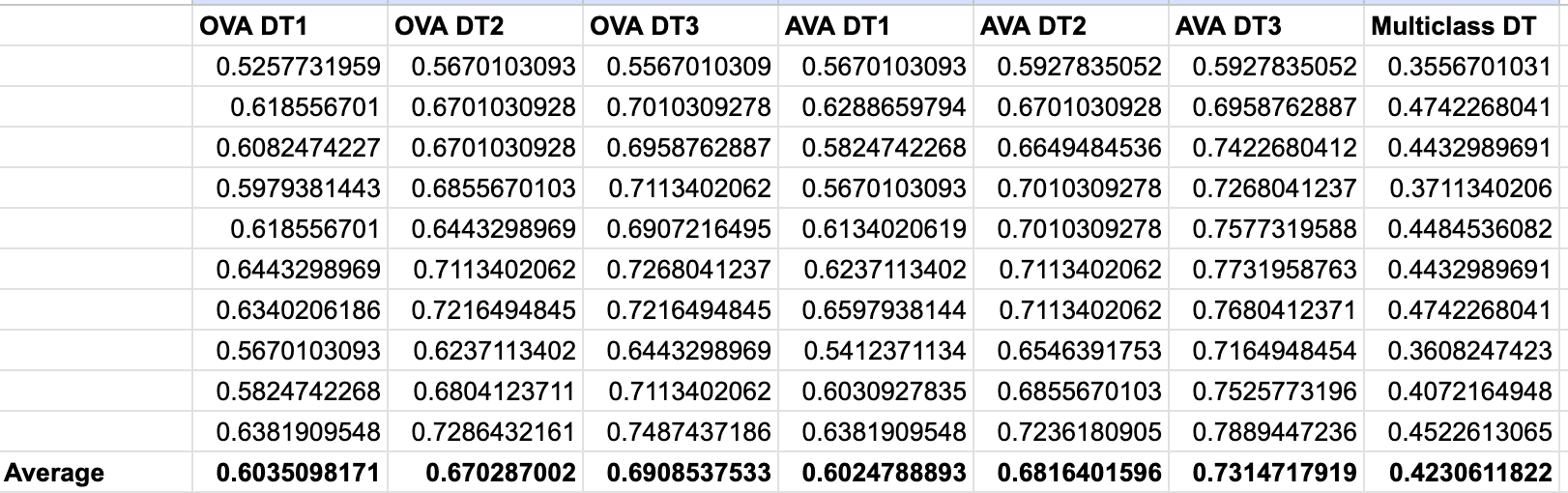
exceeding the critical value of 0.05, indicating significant differences in the means. In contrast, the pair with significantly different means had a t-test value of 0.0000002113022496,

below 0.05, suggesting non-significant differences. This suggests that the best approach is using AVA with decision trees of depth 2, as it is a more complex algorithm which results in more

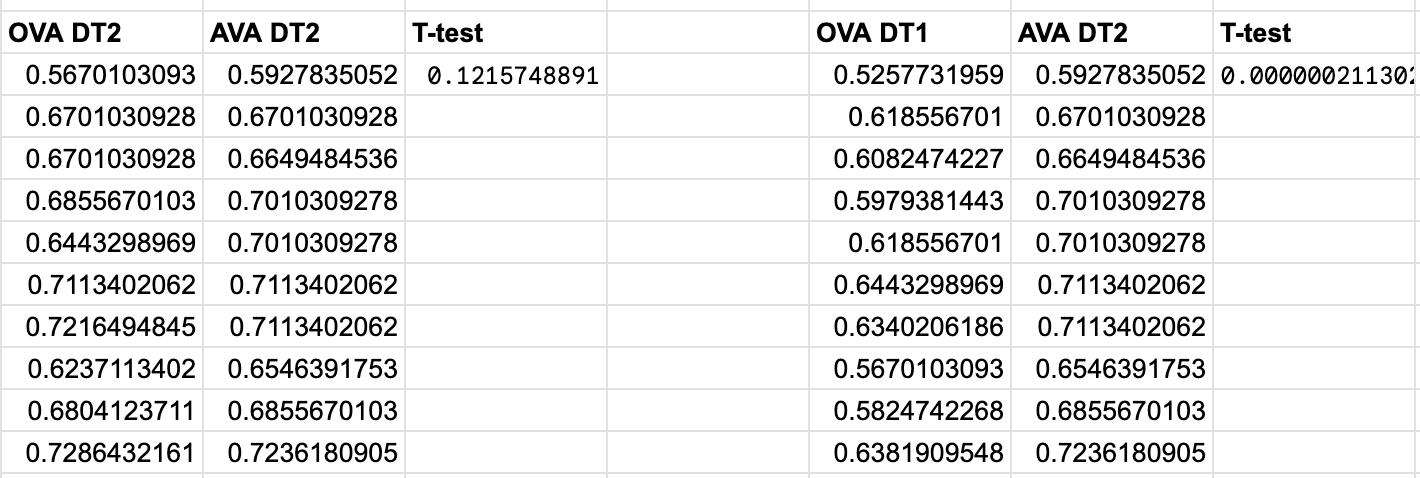
classifiers, which typically leads to better performance. This is not that surprising since even though AVA with decision trees of depth 3 had the same accuracy as the AVA with decision

trees of depth 2, it was not significantly better than OVA with decision trees of depth 1, which had the lowest accuracy (our hypothesis is that as you increase the depth of the decision

tree, for AVA, past 3, you'll start to see overfitting). Thus, for the best approach, it's better to go with AVA with decision trees of depth 2.



Here are the 2 t-tests we performed:



**5. Time both OVA and AVA on some reasonable test of the wine data set and measure both training time and testing time (separately). You can use the ClassifierTimer class if you’d like or just do it yourself. Do the timings make sense? Include your results and 2-3 sentences describing how you generated the timings and explaining the results.**

We tested OVA and AVA with 10-fold cross validation with a decision tree with depth 3 on the same dataset.

OVA took around 5 mins

AVA took around 3.5 mins

OVA would have 20 classifiers, while AVA would have 190. Thus, it is surprising that AVA took less time to test than OVA. This is probably because AVA has more classifiers, but each classifier is smaller (trained on much fewer examples) than the ones in OVA.

**6. On this data set, what would you say is the best approach to use? Briefly justify your answer.**

On this particular data set, it's safest to go with AVA with a decision tree of depth 3. OVA DT3 has the highest accuracy - 0.7314717919 - of all the results we calculated. Moreover, for AVA with DT1 through DT3, we see the accuracy increasing. Based on our timing tests, AVA was also faster than OVA with a depth of 3.

**7. Train the OVA classifier with decision trees sized 3 on all of the data. What is the tree for the zinfandel classifier (again, use the toString method to print it out)? Does this make sense? What words are indicative of the class? What words are not indicative of the class? Include a short summary describing your analysis of the tree.**

(zinfandel

(zin

(amador

predict=-1.0

predict=1.0)

predict=1.0)

(sauvignon

(grenache

predict=1.0

predict=-1.0)

predict=-1.0))

The "zin" branch makes sense because that is a synonym for zinfandel. The "amador" branch makes sense because that is a region in California known for zinfandel. The "sauvignon" and "grenache" branches do not make sense because they are different types of wine.