Joel Kerfoot
V00855134

Assignment 1 – SENG 474

Classification Problem

The second classification problem is classifying an email as spam or not spam. The dataset was obtained from the UCI Machine Learning Repository [1]. The dataset consists of 56 predictor variables and the variable to predict is encoded as 0 (not spam) or 1 (spam). This is an interesting problem as everyone has received spam mail. Most email providers do a good job of sorting out spam from your main inbox and we will look at the ability of various machine learning algorithms at classifying email as spam. I chose this dataset to compare the methods as it contains a large number of samples 4600 compared to the ~300 heart disease samples. It also contains significantly more variables, 56 to 13, compared to the heart disease dataset. I was curious if the larger samples and more variable set would produce more accurate results.

Background Information

Initially the data was split 80% for training and 20% for testing. This was done for both the heart disease and spam data sets. The best attributes were found for each algorithm, decision trees, random forests, and neural networks. These attributes where then used to analyze the effects of changing the split between training and test data.

Performance and Analysis

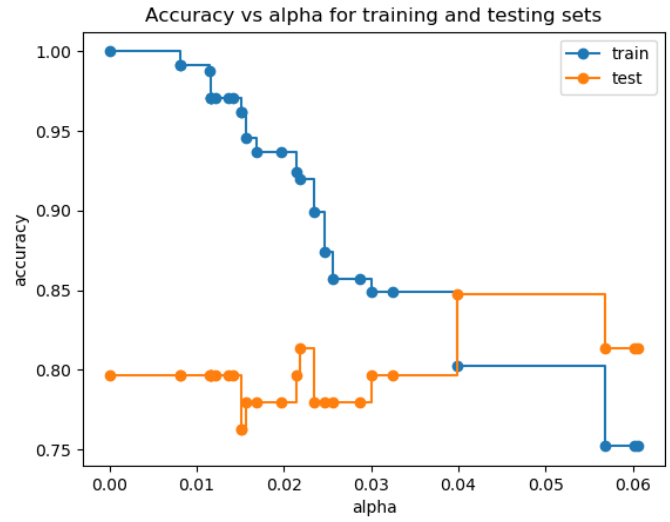
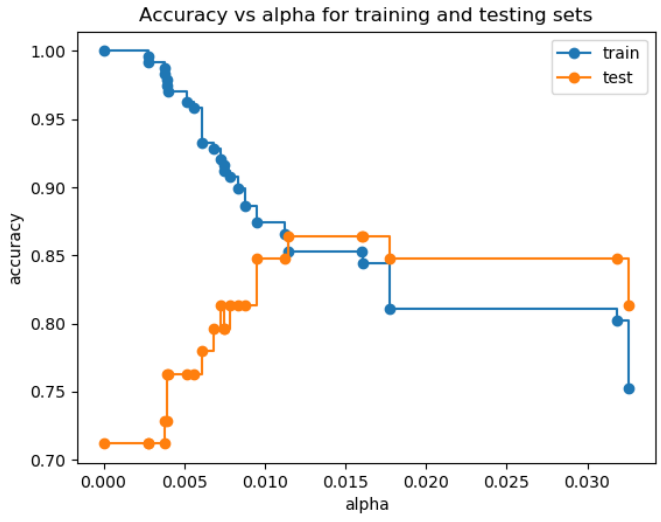
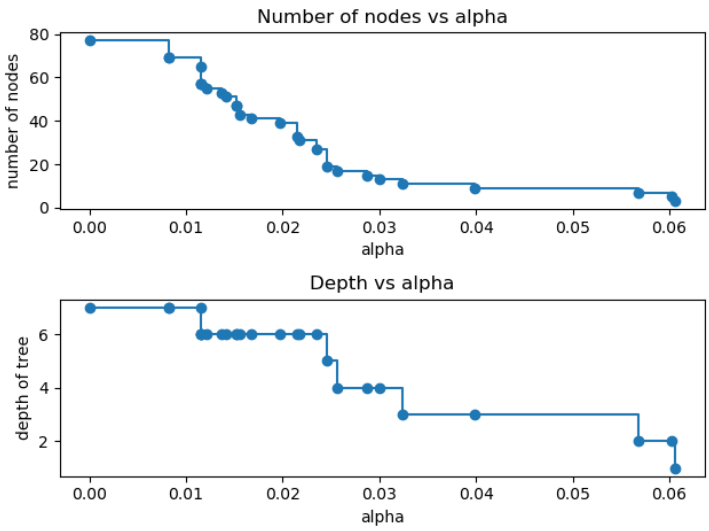
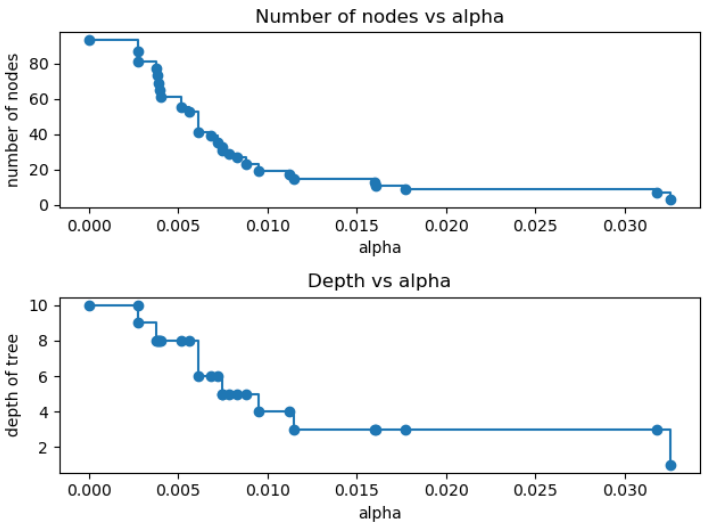
The following sections analyses the performance and results of the heart disease and spam datasets being used to train decision trees, random forests, and neural networks.

Decision Tree

The DecisionTreeClassifier from scikit learn was used to create decision tree classifiers for the data sets [2]. DecisionTreeClassifier uses cost complexity pruning to avoid over-fitting the decision tree [3]. The complexity parameter α was compared against the accuracy of the tree as it was pruned. This was done using both Gini index and entropy as split criterion and the heart disease results are shown in figure 1 and the spam results are shown in figure 2.

Joel Kerfoot
V00855134

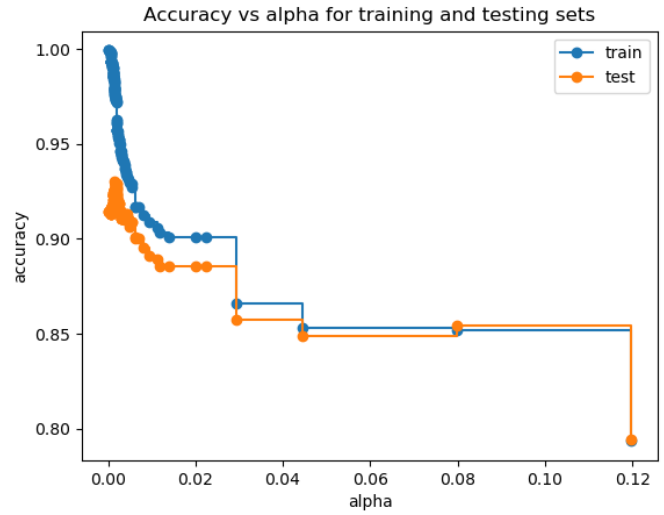
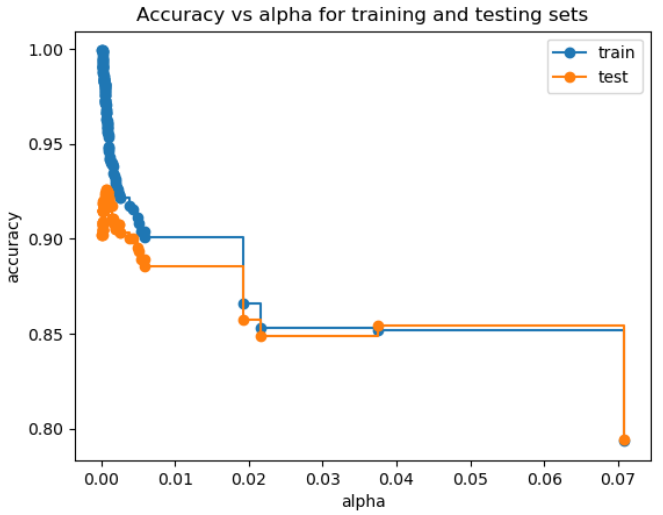
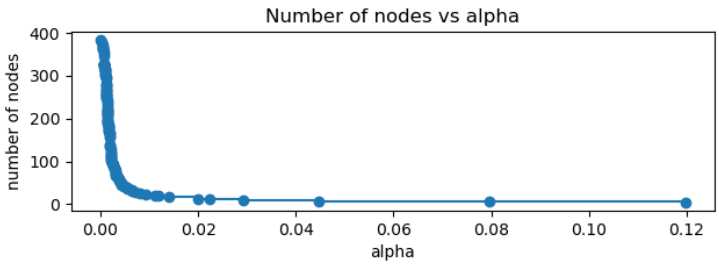
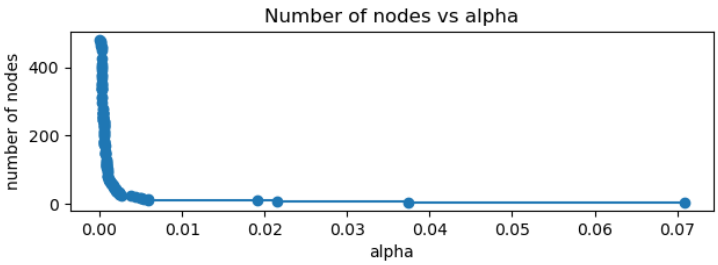
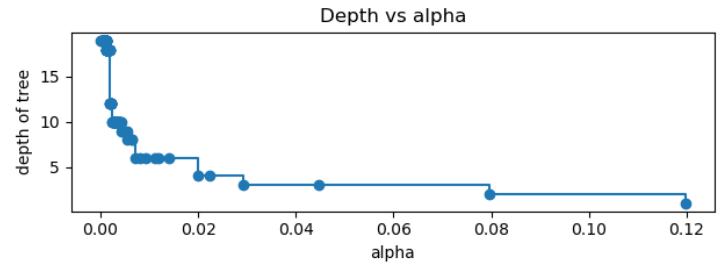
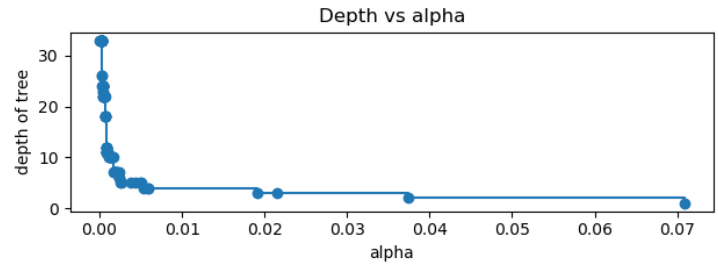
FIGURE 1 – HEART DISEASE TRAINING AND TEST RESULTS

Heart Disease – Split Criterion: Entropy	Heart Disease – Split Criterion: Gini
	
	

For the heart disease dataset using the gini index as the spilt criterion yielded the best results of approximately 87% accuracy on the test set which was marginally better than the 85% accuracy achieved using entropy as the split criterion. The best preforming trees for the different split criterion both had a tree depth of 3 while the Gini tree contained 15 nodes while the entropy tree only contained 9. The best decision trees can be found in Appendix A – Heart Disease Entropy Split Tree and Appendix B – Heart Disease Gini Split Tree.

Joel Kerfoot
V00855134

FIGURE 2 – SPAM TRAINING AND TEST RESULTS

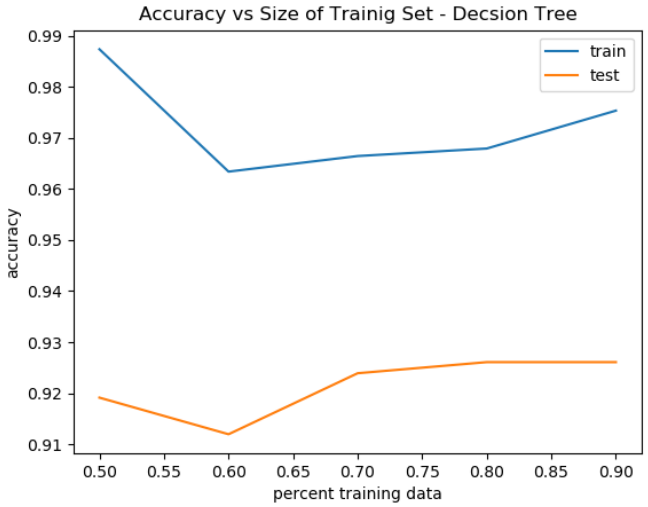
Banknote Authenticity – Split Criterion: Entropy	Banknote Authenticity – Split Criterion: Gini
	
	
	

For the spam dataset both entropy and gini split criterion were able to achieve approximately 93% accuracy on the test dataset. Initial each tree started out with around 400 nodes but at the maximum accuracy for entropy and gini the trees had 185, and 153 nodes respectively. Further pruning either tree quickly reduced the accuracy. The best decision trees for each split criterion can be found in Appendix C – Spam Entropy Split Tree and Appendix D – Spam Gini Split Tree.

For both datasets using entropy as a split criterion produced a smaller tree after pruning. The gini index preformed marginally better on the heart disease data while both split criteria preformed equally well on the simpler and larger banknote dataset.

Using the best attributes for the decision tree different training and test sets were used where the percentage of data used for training varied from 50-90%. For both datasets 80% training and 20% testing yielded the best results, with minimal difference using a 90/10 split instead. The results of the different data set partitions are shown in figure 3.

FIGURE 3 – DECISION TREE RESULTS FROM DATA PARTITIONS

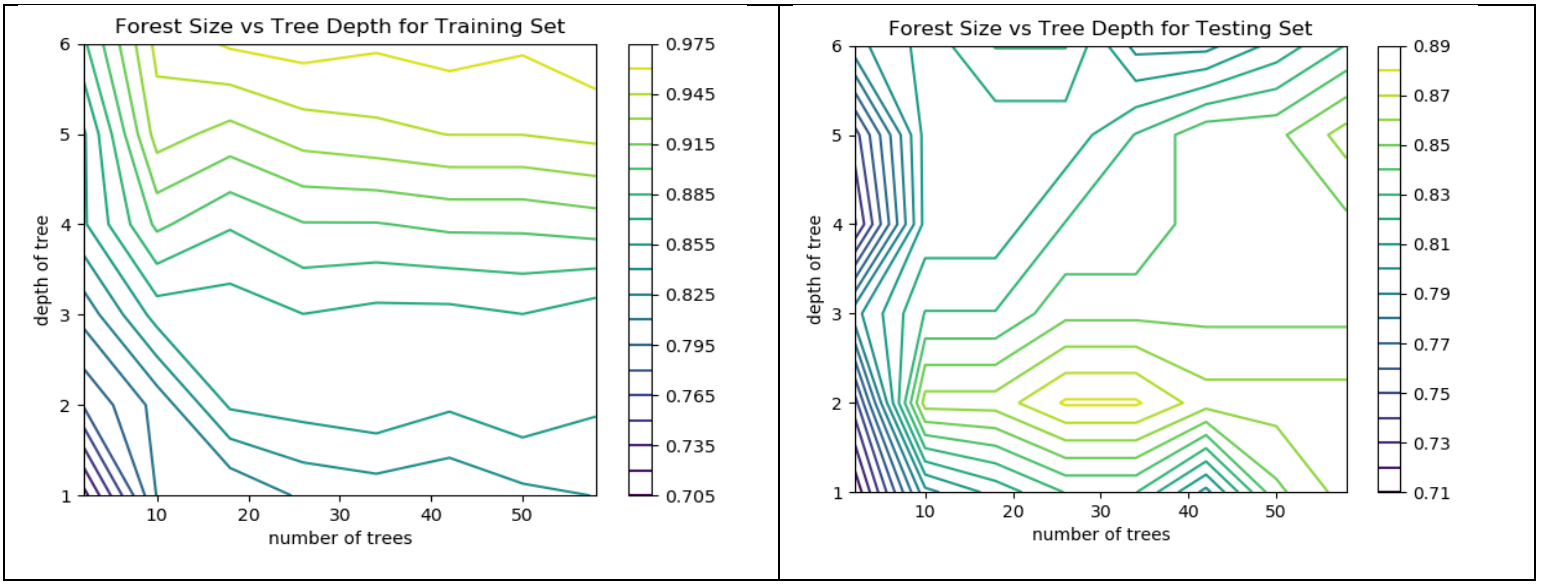
Heart Disease Dataset	Spam Dataset
	

Random Forest

The RandomForestClassifier from scikit learn was used to create decision tree classifiers for the data sets [4]. Several forests were generated by combining different variations of maximum tree depth and maximum size of the forest. These attributes were used because they provided the greatest variation in results. For all forests gini was used for split criterion, and and max features as square root of the total number of features. Figure 4 compares the accuracy of the test and training sets against these variables used on the heart disease dataset and Figure 5 shows the results from the banknote dataset.

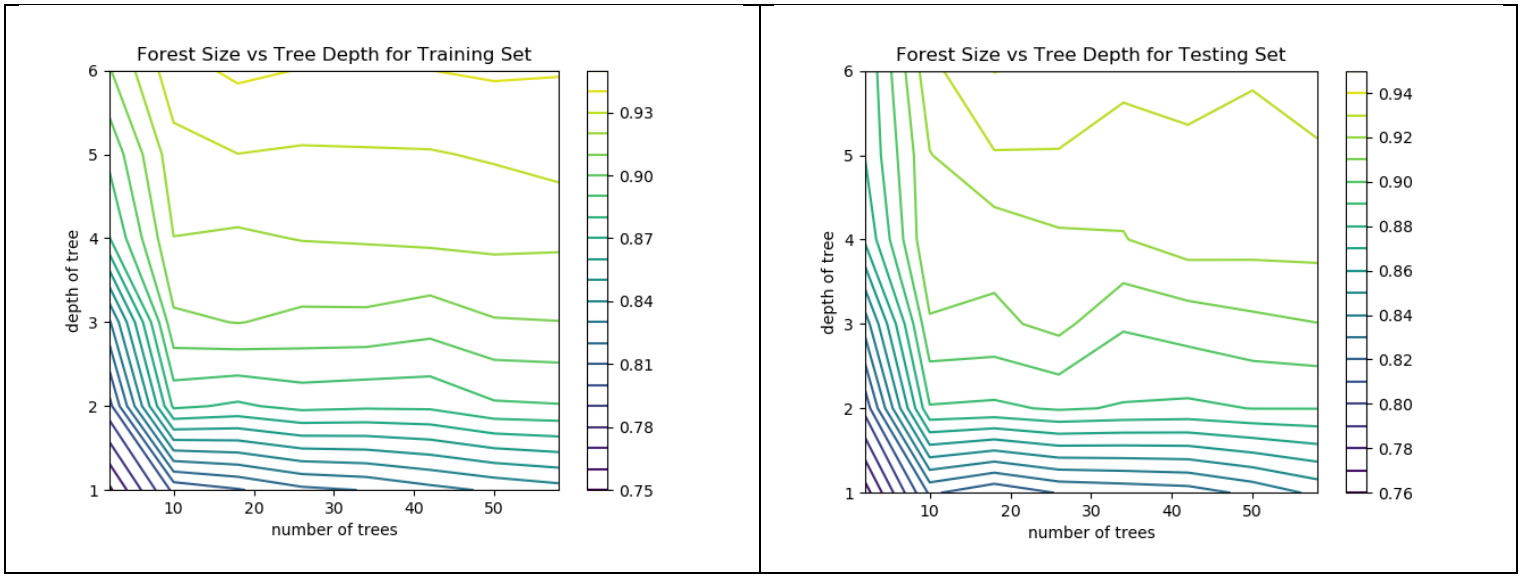
Joel Kerfoot
V00855134

FIGURE 4 – HEART DISEASE RANDOM FOREST TRAINING AND TEST RESULTS



For the heart disease training dataset random forests were able to achieve 97% accuracy with a forest size between 20 and 50 trees where the trees have a max depth of 6. Comparing this to the test data the large trees overfit the data because the same forest only yields roughly 80% accuracy. The forests that preformed the best on the test set were smaller forests of around 30 trees with depth 2. The smaller forests were able to achieve around 88% accuracy on the test data.

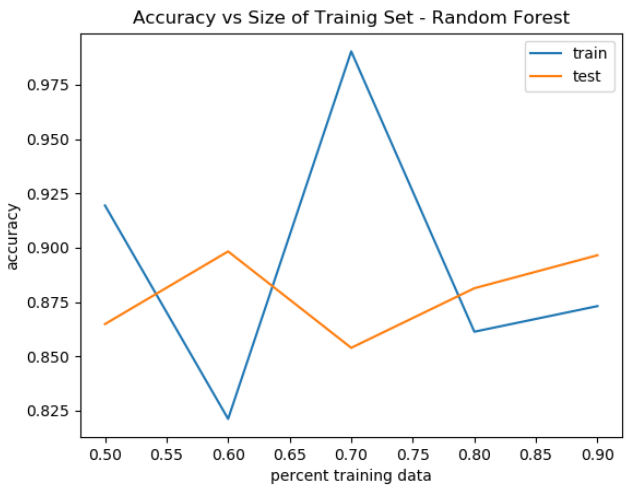
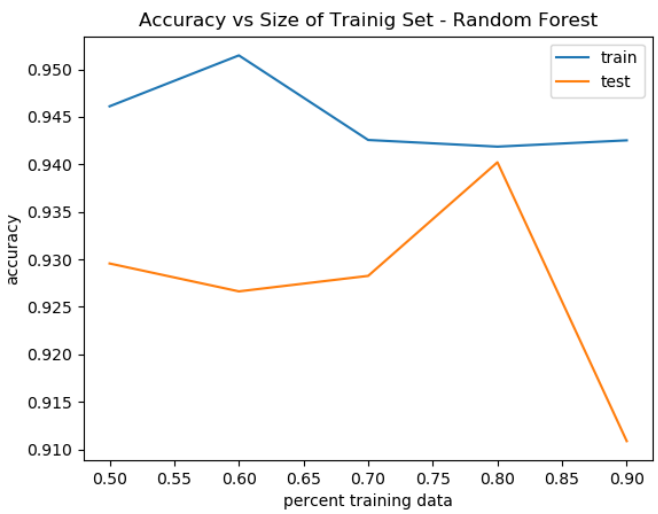
FIGURE 5 – SPAM TRAINING AND TEST RESULTS



The random forest classifier was able to achieve a much higher accuracy on the spam dataset compared to the heart disease one. The depth of the tree provides more accurate results faster than more trees in the forest do. A small forest with trees of depth 5 or 6 performs just as well as a larger forest of similar depth. The best accuracy on the test set was achieved with a forest size of 18 and a max tree depth of 6.

Using the best attributes for the random forest classifier different training and test sets were used where the percentage of data used for training varied from 50-90%. The best test accuracy on the heart disease dataset was achieved when the data was allocated 90/10 in favor of training but performed just as well with a 60/40 allocation. Like the decision tree the spam dataset performed the best with an 80/20 split and otherwise performed at least 1% worse. Figure 6 outlines the results achieved from the different data partitions.

FIGURE 6 – RANDOM FOREST RESULTS FROM DATA PARTITIONS

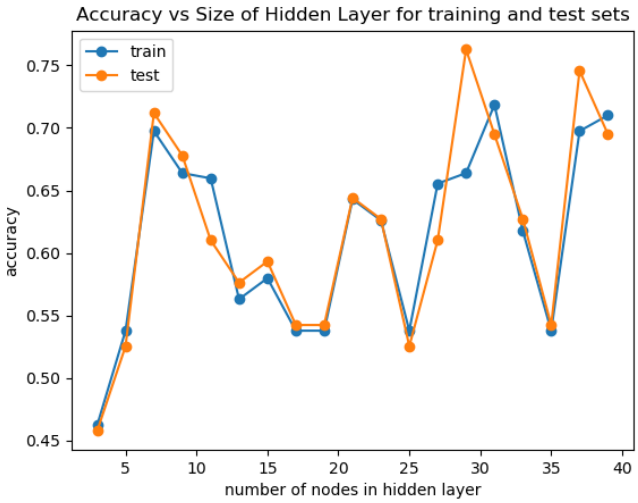
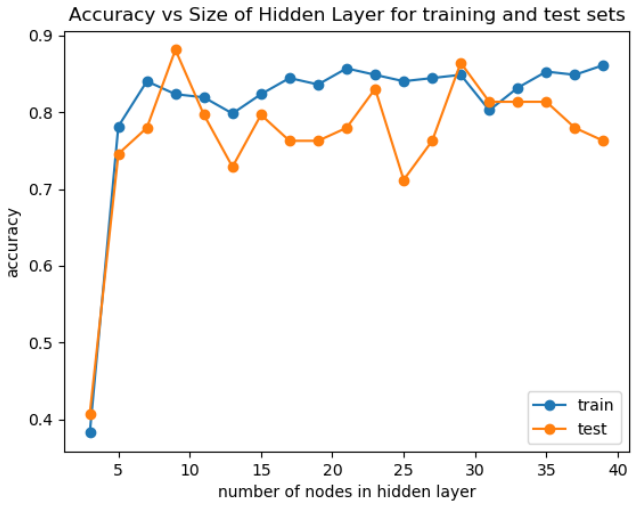
Heart Disease Dataset	Spam Dataset
	

Neural Network

The MLPClassifier from scikit learn was used to create neural network classifiers for the data sets [5]. Several networks were created using the stochastic gradient descent and Adam solver. Both solvers were used with a single hidden lay varying in size from 3 to 40 nodes. Figure

7 compares the accuracy of the test and training sets against these variables used on the heart disease dataset and Figure 8 shows the results from the spam dataset.

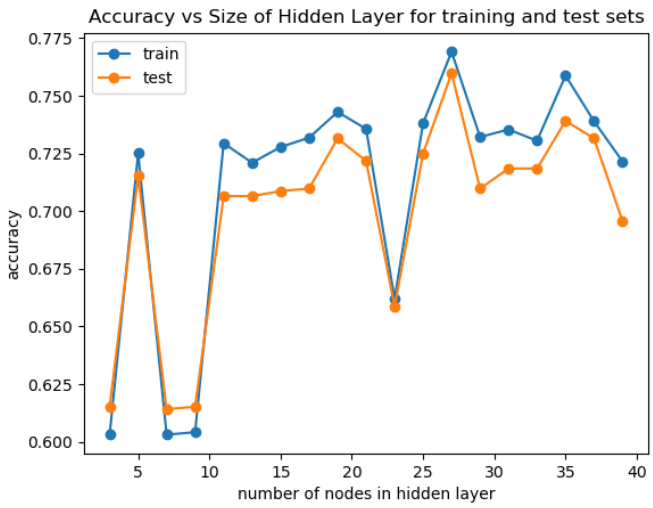
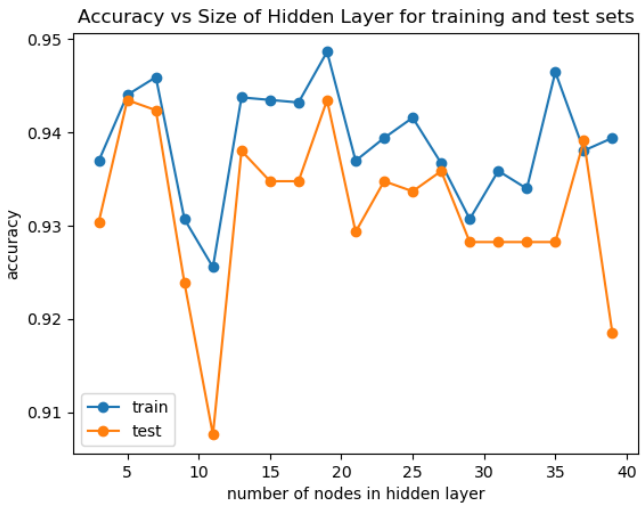
FIGURE 7 – HEART DISEASE NEURAL NETWORK TRAINING AND TEST RESULTS

Solver – SGD	Solver – Adam
	

The stochastic gradient descent did not produce the results I would have expected. The extreme variation with the number of nodes was unexpected. Trying different parameters did not change the large fluctuations so perhaps it's a limitation of the algorithm or caused by the nature of the data. The maximum accuracy of 75% is reached with 29 nodes in the hidden layer. The Adam solver on the other hand produced a smoother accuracy curve over the size of the hidden layer. It reached a peak accuracy of 88% with 9 nodes in the hidden layer.

Joel Kerfoot
V00855134

FIGURE 8 –SPAM NEURAL NETWORK TRAINING AND TEST RESULTS

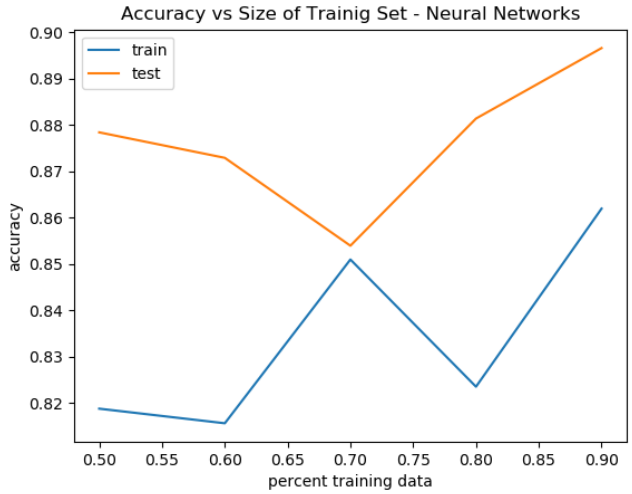
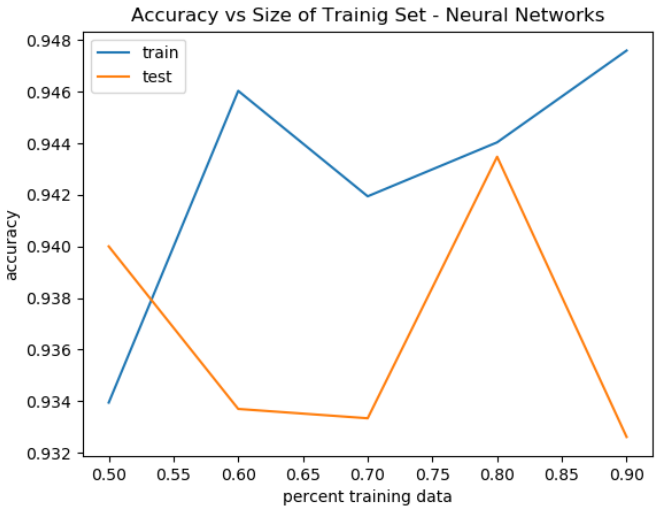
Solver – SGD	Solver – Adam
	

The spam dataset again yields large accuracy fluctuations when using stochastic gradient decent, but Adam solver isn't as smooth as the heart disease one. Like the heart disease networks the Adam solver produces better results. A maximum test accuracy of 94.8% is achieved with 19 nodes in the hidden layer.

Using the best attributes for the neural classifier different training and test sets were used where the percentage of data used for training varied from 50-90%. Again, the heart disease dataset performed the best with a 90/10 split. This could be because of the smaller amount of test data means there is less chance of having a test that will fail. The spam dataset performed the best with an 80/20 split, the same as with the other classifiers. Figure 9 shows the results from the different partions.

Joel Kerfoot
V00855134

FIGURE 9 – NEURAL NETWORK RESULTS FROM DATA PARTITIONS

Heart Disease Dataset	Spam Dataset
	

Conclusion

For both the hear disease and the spam mail dataset the three classifiers: decision tree, random forest, and neural networks all produced similar test accuracies. On the heart disease dataset decision trees, random forests, and neural networks yielded 87, 88, 88 percent test accuracy respectively while the spam data set yielded 93, 94, and 94% test accuracy. The largest variation in test accuracy came not from the method used but using the appropriate parameters for the method being used.

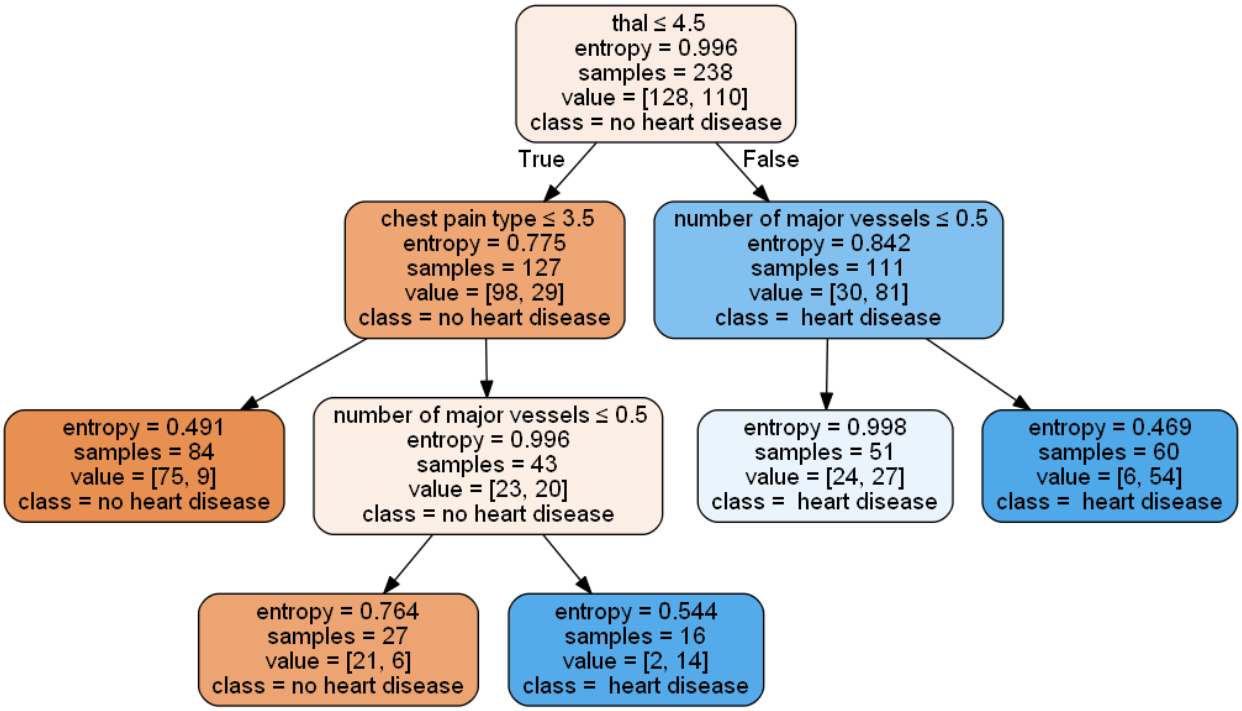
Joel Kerfoot
V00855134

References

1	UCI, “Spam Base” [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/. [Accessed February 1, 2020].
2	UCI, “Decision Tree Classifier” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier/. [Accessed February 1, 2020].
3	UCI, “Minimal Cost-Complexity Pruning” [Online]. Available: https://scikit-learn.org/stable/modules/tree.html#bre. [Accessed February 1, 2020].
4	UCI, “Random Forest Classifier” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed February 1, 2020].
5	UCI, “Neural Network Classifier” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier.[Accessed February 1, 2020].

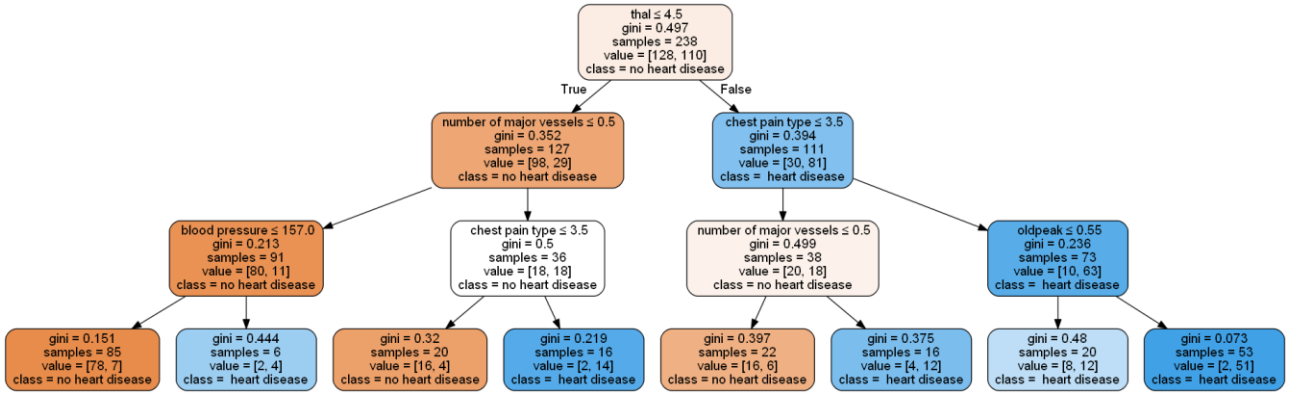
Joel Kerfoot
V00855134

Appendix A – Heart Disease Entropy Split Tree



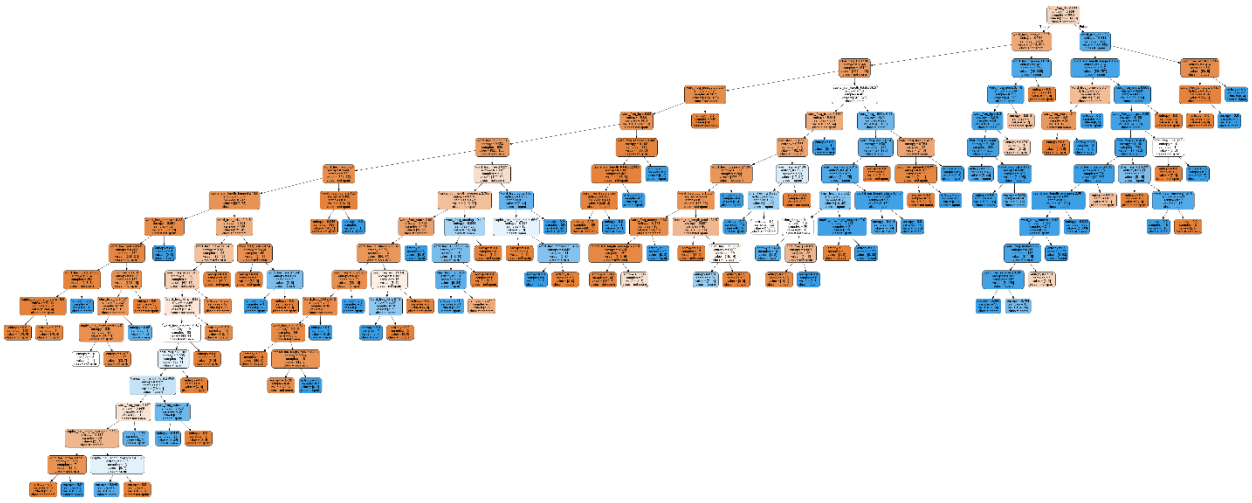
Joel Kerfoot
V00855134

Appendix B – Heart Disease Gini Split Tree



Joel Kerfoot
V00855134

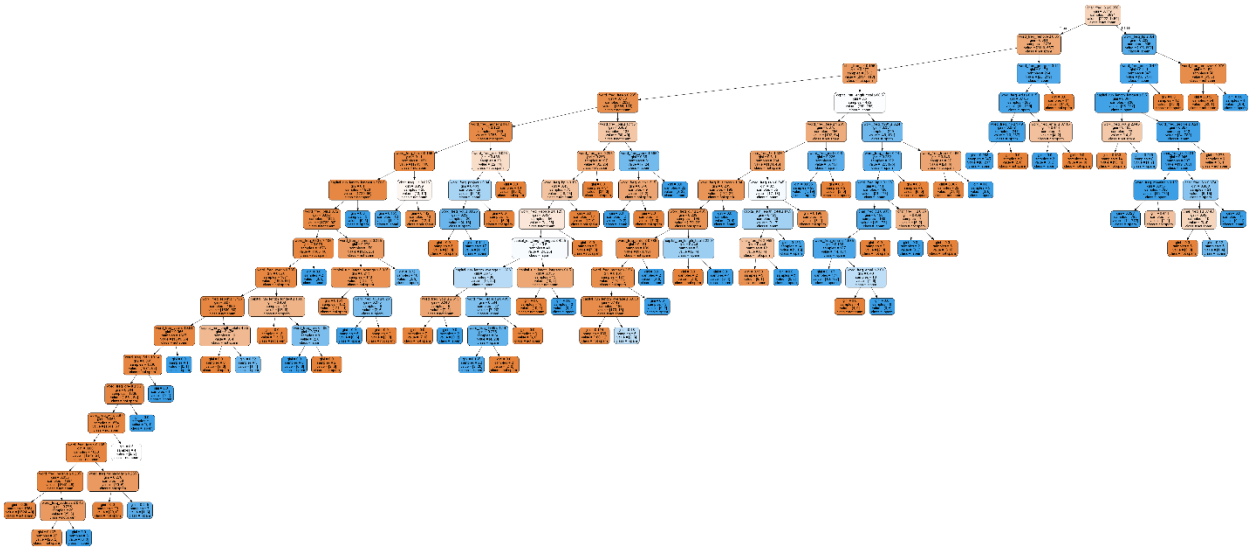
Appendix C – Spam Entropy Split Tree



Viewable online at: https://i.imgur.com/YXZ7TG9.png

Joel Kerfoot
V00855134

Appendix D – Spam Gini Split Tree



Viewable onlone at: https://i.imgur.com/JfxDKCL.png