# Assignment 1 – SENG 474

## **Classification Problem**

The second classification problem is classifying banknotes (Euros or American Dollar Bill) as authentic or counterfeit. The dataset was obtained from the UCI Machine Learning Repository [1]. The dataset consists of 4 predictor variables (variance of image, skewness, kurtosis, and entropy) and the variable to predict is encoded as 0 (authentic) or 1 (counterfeit). This is an interesting problem as it can be quite difficult for humans to determine if a bill has been counterfeited. The ability for a machine to accurately detect fraudulent bills would be important for businesses, banks, or individuals to be certain of the authenticity of their money. I chose this dataset to compare the methods as it contains a large number of samples 1372 compared to the ~300 heart disease samples. It also contains only 4 variables to the 13 of the heart disease dataset and I was curious if the larger samples and smaller variable set would produce more accurate results.

## **Background Information**

## **Performance and Analysis**

The following sections analyses the performance and results of the heart disease and banknote 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 banknote authenticity results are shown in figure 2.

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.

Figure 2 – Banknote Authenticity Training And Test Results

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

For the banknote authenticity data set both entropy and gini split criterion were able to achieve approximately 99% accuracy on the test dataset. Unlike the heart disease dataset further pruning the tree reduced the accuracy of both the test and training data. The tree using the entropy however was a significantly smaller tree. It had a depth of 6 and contained 33 nodes compared to the depth 7 and 47 node tree obtained using the gini index. The best decision trees for each split critea can be found in Appendix C – Banknote Entropy Split Tree and Appendix D – Banknote 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.

### **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. For all forests gini was used for split criterion, and and max features as square root of the total number of features. Figure 3 compares the accuracy of the test and training sets against these variables used on the heart disease dataset and Figure 4 shows the results from the banknote dataset.

Figure 3 – Heart Disease Random Forest Training And Test Tesults

|  |  |
| --- | --- |
|  |  |

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.

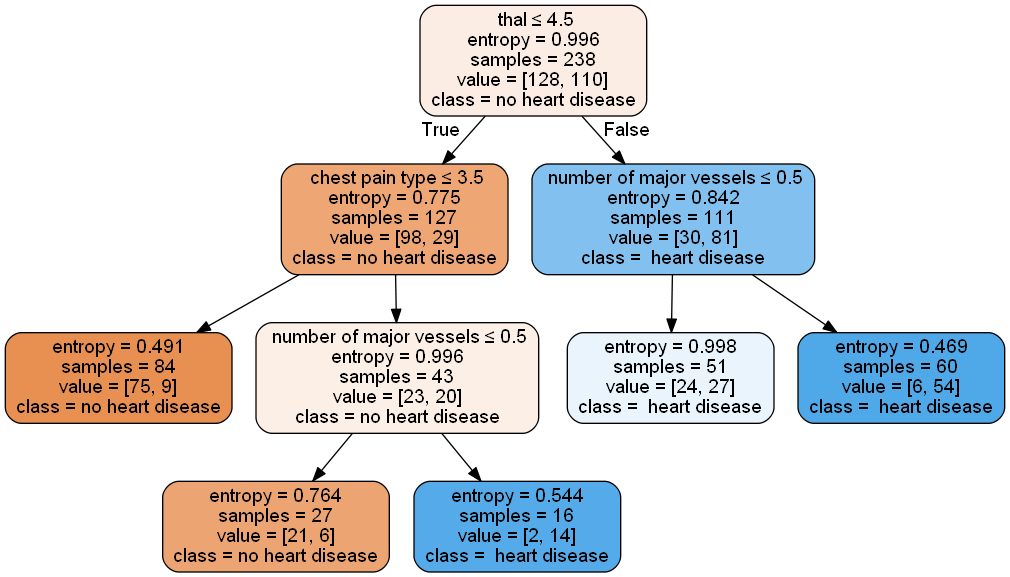
Figure 4 – Banknote Authenticity Training And Test Results

|  |  |
| --- | --- |
|  |  |

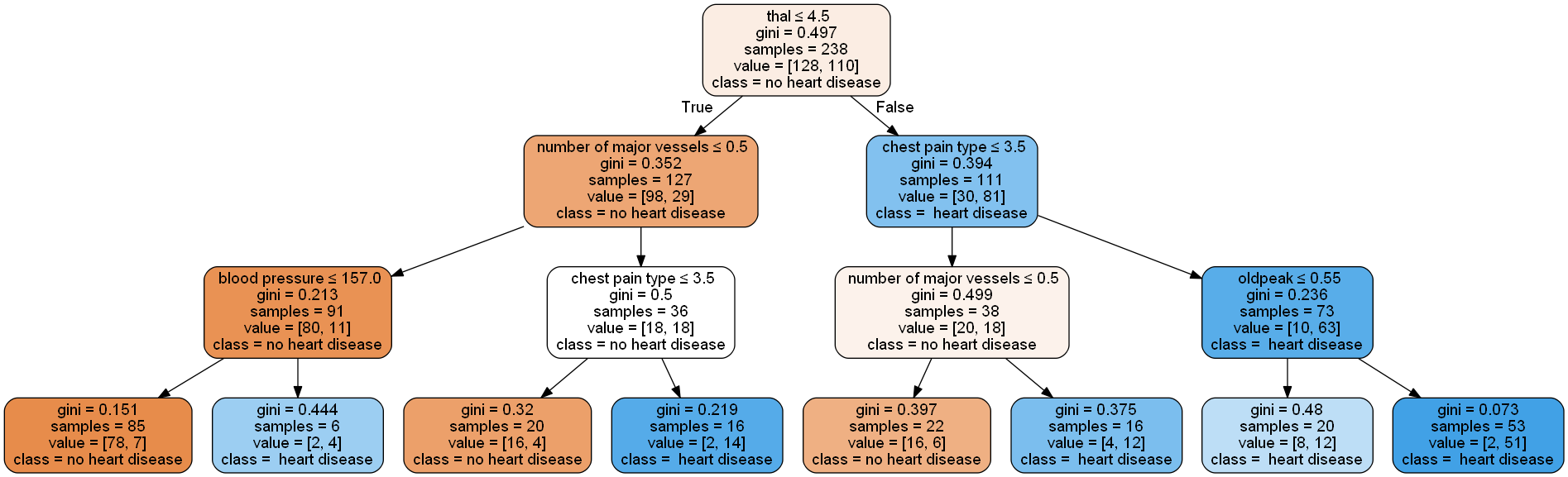
## **References**

|  |  |
| --- | --- |
| 1 | UCI, “Banknote Authentication Data Set” [Online]. Available: <https://archive.ics.uci.edu/ml/machine-learning-databases/00267/>. [Accessed February 1, 2020]. |
| 2 | <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier> |
| 3 | <https://scikit-learn.org/stable/modules/tree.html#bre> |
| 4 | <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html> |

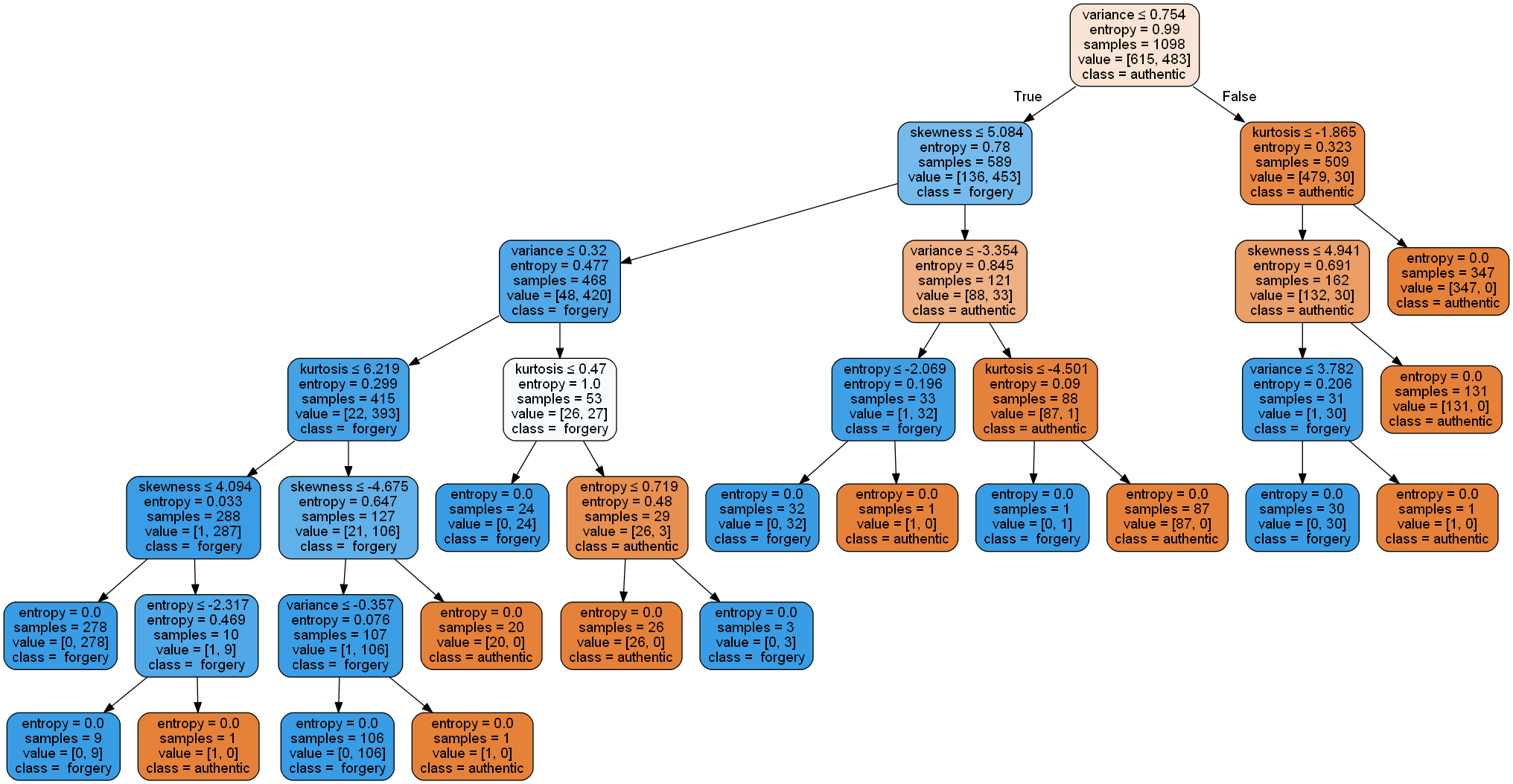
## **Appendix A – Heart Disease Entropy Split Tree**



## **Appendix B – Heart Disease Gini Split Tree**



## **Appendix C –Banknote Entropy Split Tree**

****

## **Appendix D –Banknote Gini Split Tree**

