Supervised Learning Report

Datasets:

1. Banknotes

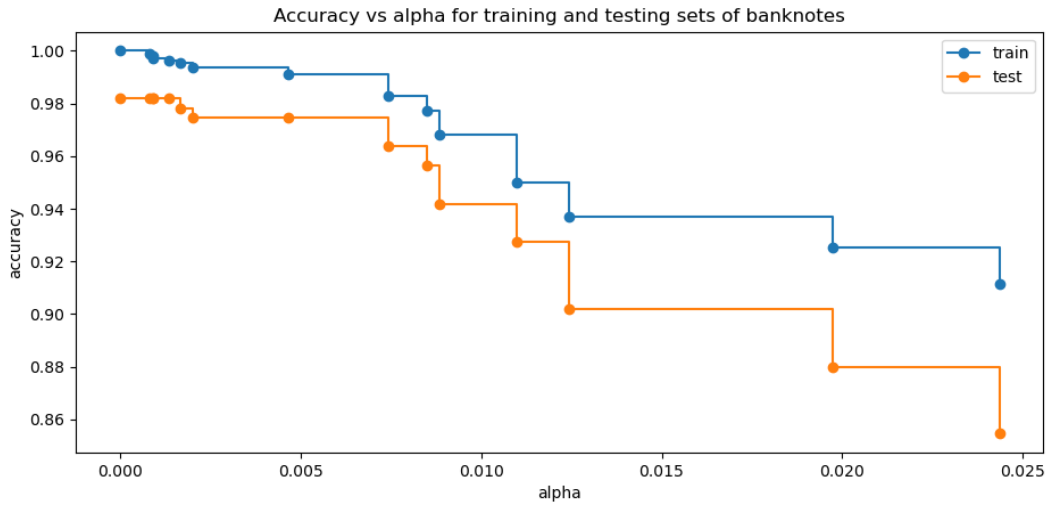
This is a binary classification dataset. Each instance consists of various measured attributes from a photo of a banknote. These attributes are different transforms of the image of each banknote. The classification of each instance is authentic or inauthentic, represented by the values 0 and 1, respectively.

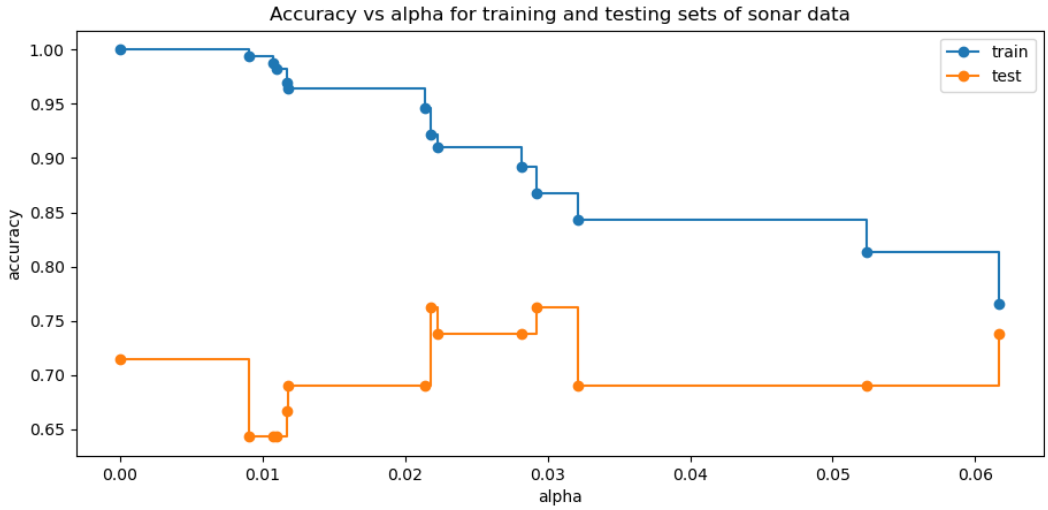
2. Sonar Dataset

This is also a binary classification dataset. The dataset addresses the problem of differentiating between rocks and mines using sonar data. Each instance consists of 60 sonar measurements from angles around the targets. The ‘mines’ used for testing were actually just metal cylinders. Each measurement consists of a value between 0 and 1 showing how strongly the sonar signal returned from the measured angles. The classification values are binary, ‘R’ for a rock and ‘M’ for a mine.

Both datasets are unbalanced with respect to their classifications. The sonar dataset has 111 observations of mines and 97 observations of rocks. The banknote dataset has 762 authentic banknotes and 610 inauthentic banknotes. The difference between the classes is not especially large, so for this study no preprocessing will be performed to address this issue. Additionally, accuracy will be used as the metric to evaluate model performance due to this assumption that the difference in sizes is not significant. Other metrics such as precision, recall, or the F1 score could be applied in the future if it was decided this difference was significant.

Decision Trees:



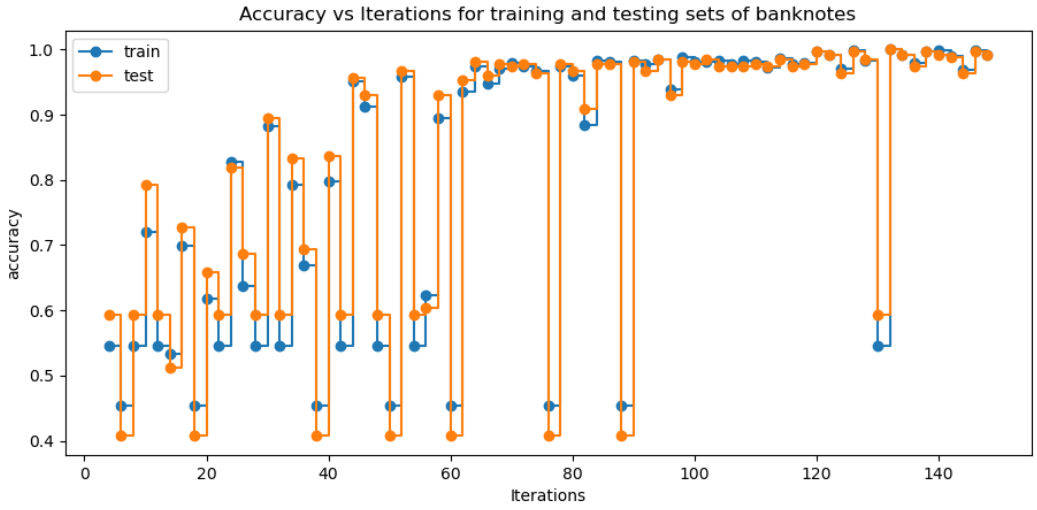


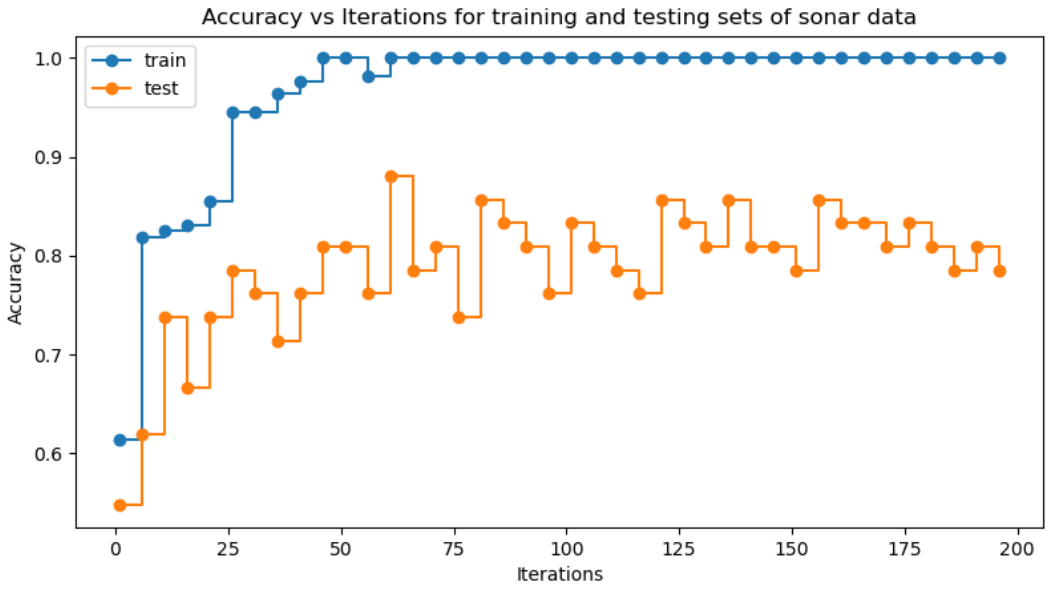
Both datasets were evaluated with respect to the alpha value given to the decision tree model. The alpha value constrains the maximum complexity of the tree, and thus the maximum amount of time spent training and testing using said trees.

The banknote data set achieved near perfect accuracy in both training and testing at low values of alpha. This implies the data set is almost noiseless, a pattern that will be repeated in the below models. Also supporting the lack of noise in the banknote data set is how the accuracy for the testing and the training sets both fall in lockstep as the alpha value increases. The similar decreases imply that errors made fitting to the training set are perfectly replicated by the testing set, which occurs when the dataset is relatively noise free.

In contrast, the sonar data indicates the data possesses a much higher degree of noise. At lower values of alpha the model attains near perfect accuracy for the training data but much lower accuracy for the testing data, demonstrating that the model over fits to the training data at these alpha values. This is further proven when looking at progressively larger values of alpha, for the accuracy with respect to the testing set stays constant at around 70% while the testing accuracy decreases. The test values for the alpha value stopped around 0.06, beyond which the accuracy for both testing and training values stayed approximately constant, indicating that 70% is roughly the accuracy achieved by a few simple splits of the data.

Neural Networks:

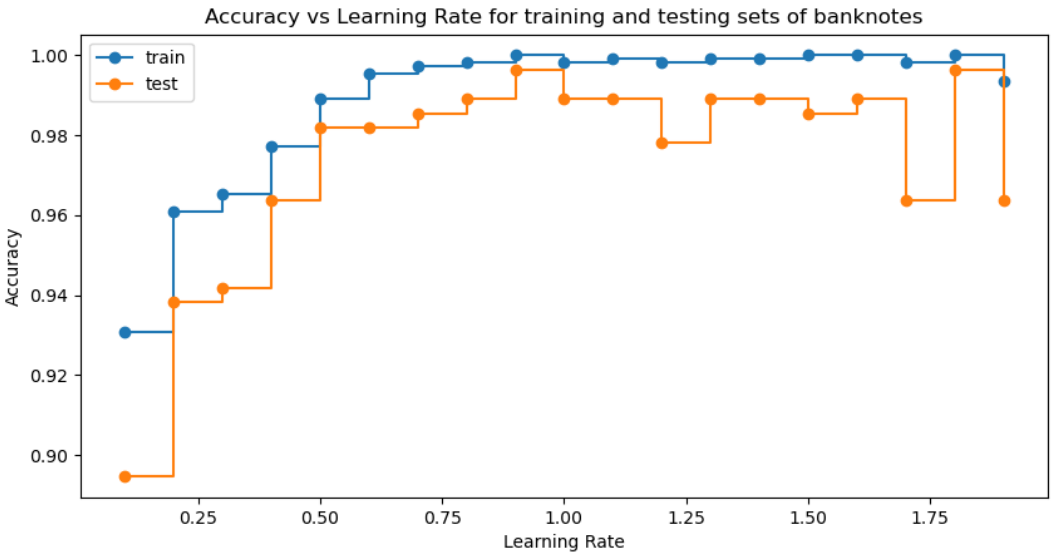


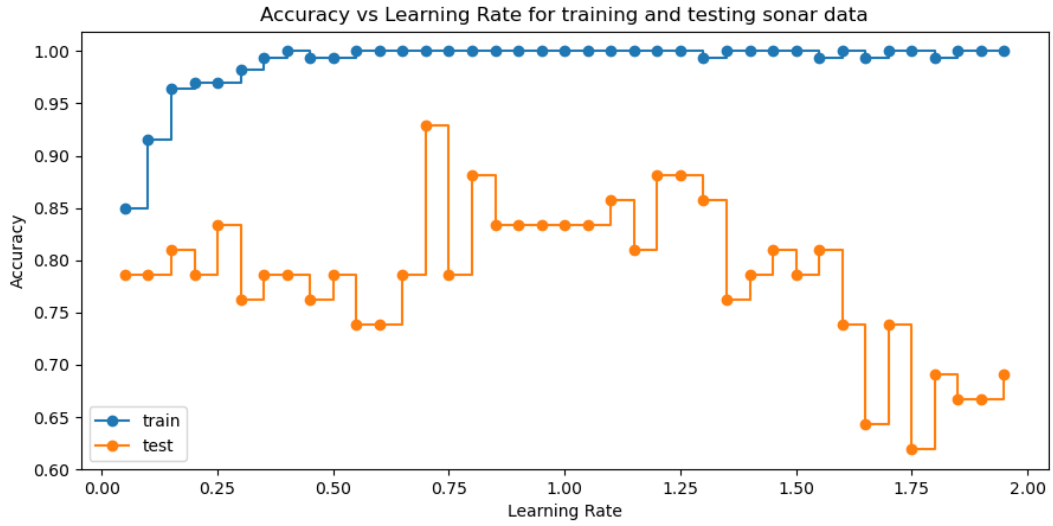


The neural network for the banknotes was designed with three hidden layers with four perceptrons each to mirror the 4 features of the banknotes dataset. As with the previous decision trees, the model rapidly reaches its near perfect accuracy, steady into 100% accuracy for both training and testing at around 60 iterations of the neural network. Again, the similar accuracy for testing and training data further indicates a lack of noise in the data. The occasional large shifts downward to testing and training accuracy of around .4 for testing and .45 for training implies that there is one specific perceptron responsible for these shifts, observed at 20, 40, and 160 iterations, among others. While difficult to say with certainty due to the black box nature of neural networks, those swings indicate the one specific feature controlling that perceptron can be used to very cleanly divide the dataset between classes.

The sonar data neural network model also shows similar issues to the decision tree model. The training accuracy achieves 100% accuracy after about 50 iterations, while the testing accuracy then stays constant at around 80%. This is a 10% improvement of accuracy on the decision trees, implying relations between the features of the sonar data set that are difficult for the decision tree model to capture. Beyond 50 iterations the variations in the testing data are only the result of the different ways the perceptrons over fit to the noise in the testing dataset.

Boosting with decision trees:

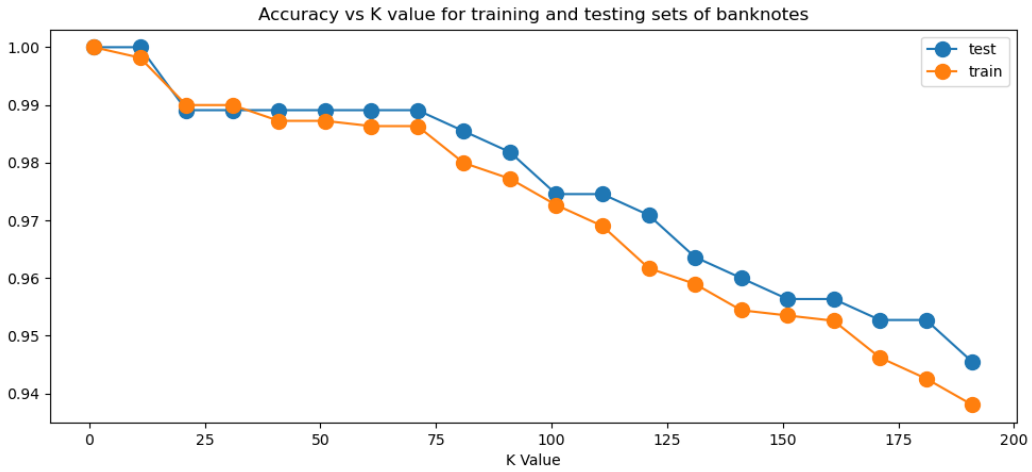


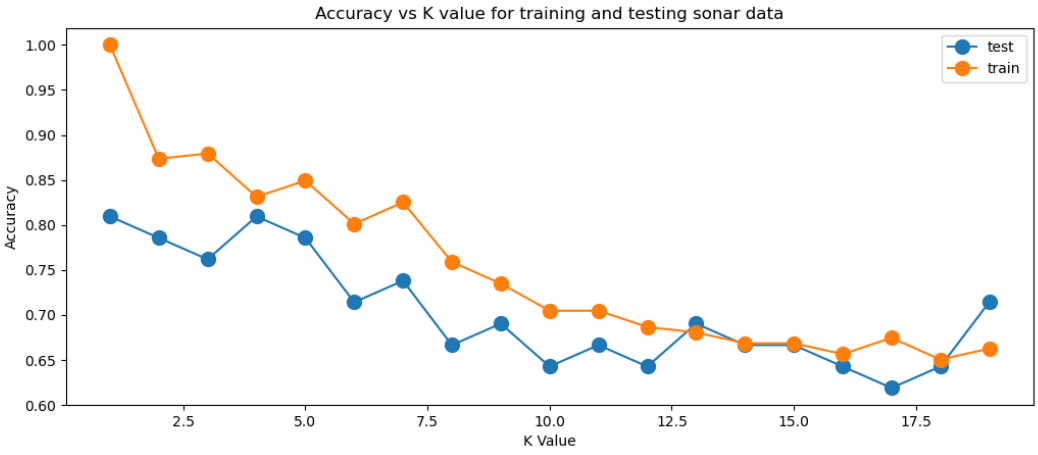


The boosting performed here was done using the same decision trees as before with an alpha value of 0, so no pruning was enforced for the decision trees. The value altered was the learning rate of the tree, which influences how quickly the boosting model settles on a final model, how slowly it settles on those answers, and how fine-tuned the final model will likely be. Learning rates greater than 1.0, generally not used due to the large decrease in model accuracy, were used here to reaffirm that these large value only result in less accurate models, as seen in the above graphs which reach approximately optimum values around 1.0.

There is little extra to say about these models not previously stated. For banknotes the model rapidly reaches near perfect accuracy for both testing and training, while for sonar data the model rapidly over fits to the noisy training data while testing accuracy stays constant around 80%.

K-Nearest Neighbors:





The nearest neighbor algorithm was calculated with Euclidean distance and the distance function for choosing neighbors. The number of neighbors used was varied over a range specific to each dataset, so a larger range was required for the banknotes dataset due to how robust it was, barely losing any accuracy even when using 50 neighbors.

Using multiple neighbors is generally a technique for filtering out noise, so in a relatively noiseless dataset like the banknotes maximum accuracy is achieved with only 1 neighbor and additional neighbors only result in decreased accuracy. The lack of noise results in all neighbors for a particular observation normally being the same classification, so the increased number of neighbors only brings in increasingly larger number of observations with a different classification. The sonar data set, which has higher levels of noise, keeps roughly constant testing accuracy of 80% through the first 5 neighbors while the training accuracy falls. From there on both training and testing accuracy fall as larger numbers of neighbors are used.

The relatively small size of the sonar data set, containing about 100 observations, contributes to its quick decline due to the lack of neighbors within the dataset. However, the 80% testing accuracy achieved is similar to all previous model, so additional observation would likely only slow down the rate at which accuracy decreases with the number of neighbors used.

Support Vector Machines:

The banknotes were almost linearly separable, so a basic linear kernel for the applied SVM produced near perfect results, shown below:

```
Accuracy: 0.9927272727272727

Precision: 1.0

Recall: 0.984

Confusion Matrix:

 [[150   0]

 [  2 123]]
```

There is not much to add to this, except that it provides further evidence of the near linear seperability of the counterfeit banknotes from the real ones. The model only misclassifies two observations, likely a result of the limited noise in the dataset.

The sonar data, in contrast, showed a range of accuracies with different kernels and model attributes. Both polynomial and radial basis function (RBF) kernels were tested with varying exponents and gamma values, respectively.

Kernel	Exponent or Gamma Value	Training Score	Testing Score
poly	1	0.843373	0.666667
poly	2	0.891566	0.809524
poly	3	0.945783	0.857143
rbf	0.01	0.554217	0.595238
rbf	0.5	0.951807	0.833333
rbf	1.0	0.981928	0.904762
rbf	1.5	1.000000	0.904762

The polynomial kernel showed basic improvement with higher exponents, however nothing that has not been observed before in terms of accuracy. But, the RBF kernel demonstrates the highest levels of accuracy so far, achieving 90% testing accuracy multiple times. This implies that the sonar dataset features are distributed in a radial manner, which would explain the relatively lower accuracy of all previous models, which work tend to combine and compare those feature in a relatively linear manner.

A good future project would be to transform the sonar data in a radial manner and then apply the other models to see if this improvement in testing accuracy is consistent.