CS 7641- Machine Learning

Project 1 – Supervised Learning

Madhu Mohan
MMOHAN8

Selected Datasets

The 2 datasets that I selected were:

1. Adult Data Set from UCI Machine Learning Repository

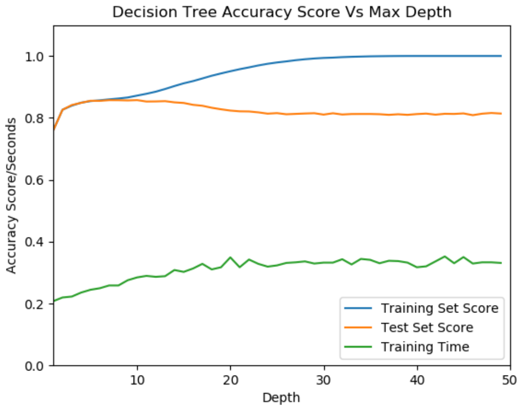
The primary reason I liked this data set was for the real-life implications of what I can find while analyzing this data. Is it really possible to build a model based that can predict poverty? Later on, other things of interest in this data were the combination of continuous and discrete attributes and the volume of this data, mirroring a more real-life example.

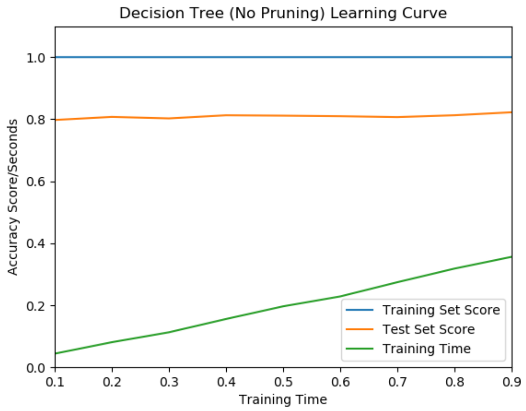
2. Wine Data Set from UCI Machine Learning Repository

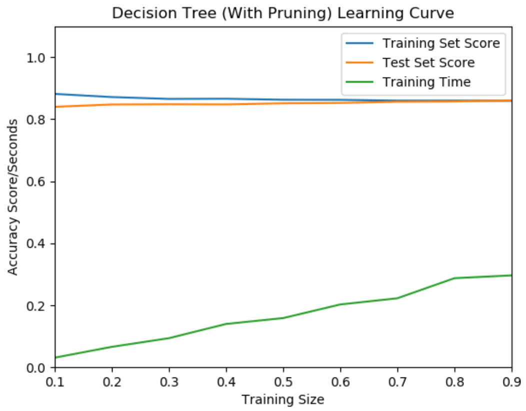
This was an interesting data set for me because I have always been interested in the qualities of a good wine. Also, the simplicity of the data attributes and the small size of the overall data made me think that this will be an interesting data set to test out the models that I build.

Decision Trees

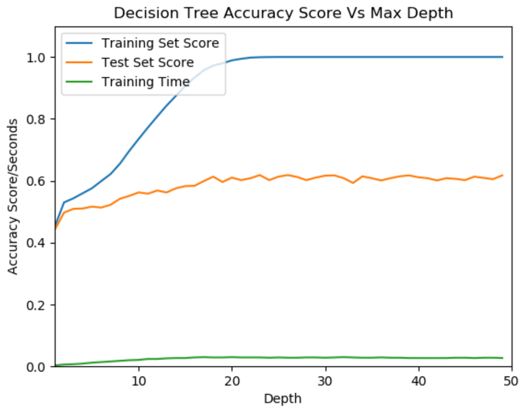
Adult Data:

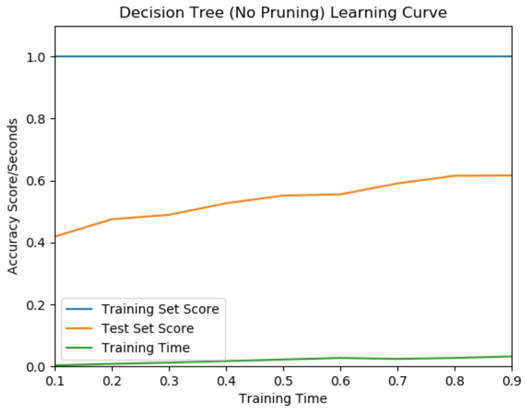


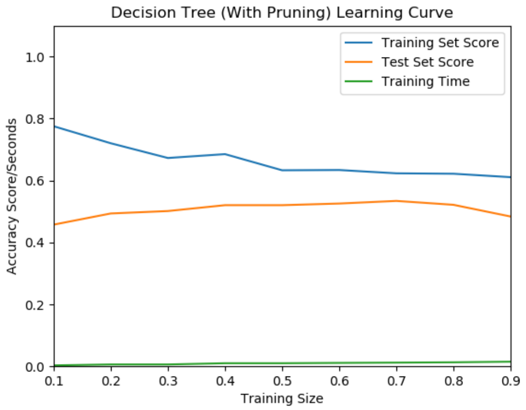




Wine Data:





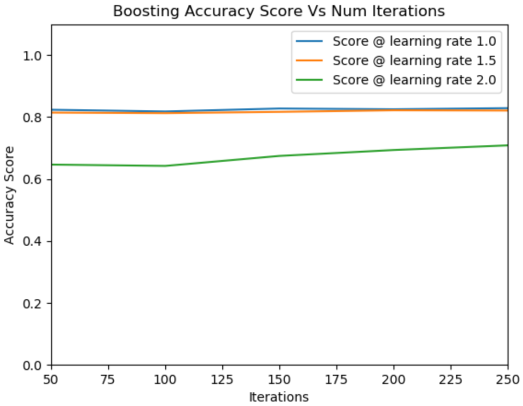


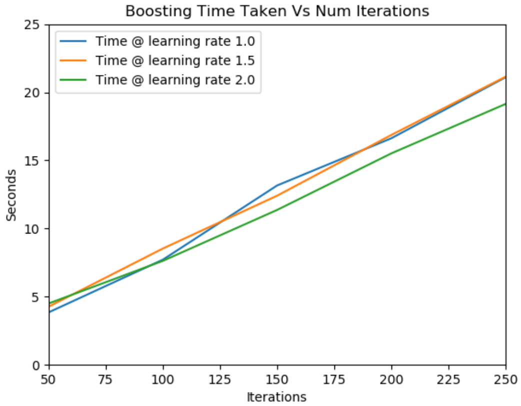
I used SKLearn's DecisionTreeClassifier to implement decision trees. I did pruning by identifying optimal number of depth for the data sets and then limited the depth with the max_depth parameter to the DecisionTree function. When we look at the Decision Tree Accuracy vs Max Depth chart, we can see that the accurach score does not improve beyond 20 for both data sets. I used that as the optimal max_depth for later calculations.

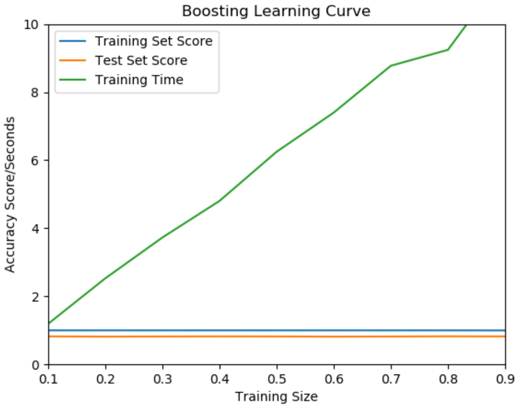
Boosting

I used SKLearn's AdaBoostClassifier to implement Boosting. I experimented with different estimators and learning rates to identify the optimal parameters to use with the algorithm. I found that while the parameters were providing good predictability, the time taken to build the model increased drastically as I increased the number of estimators and learning rates. I identified 100 Iterations and 0.6 learning rate gave me optimal results.

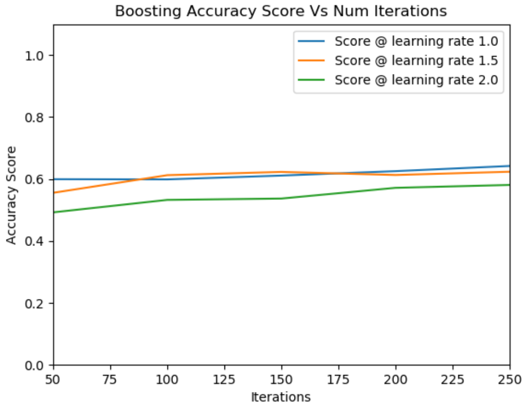
Adult Data:

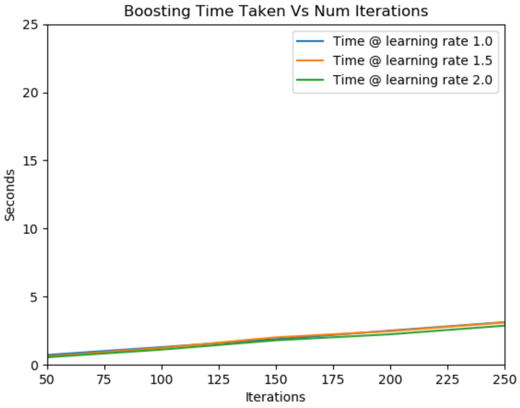


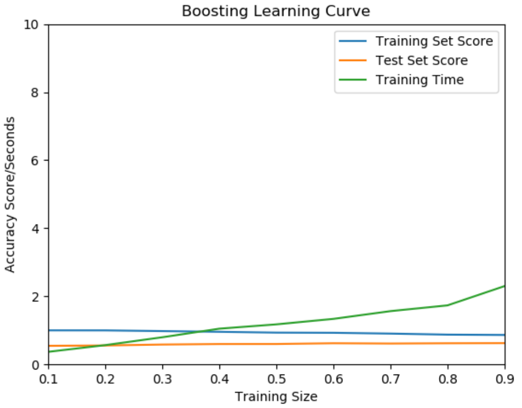




Wine Data:



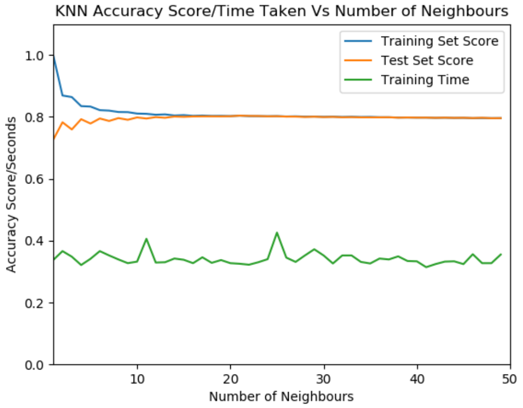


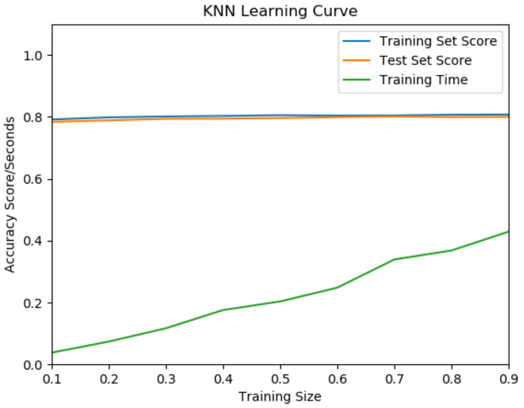


KNN

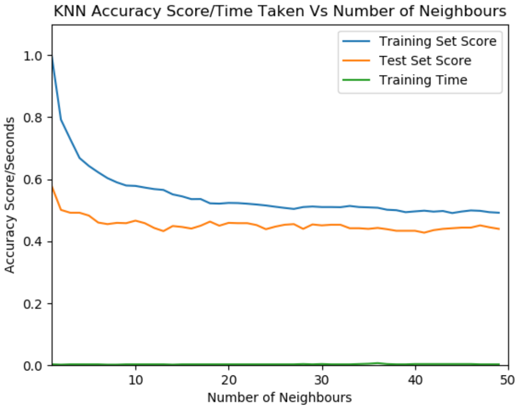
Implemented KNN with SKLearn's KNeighborsClassifier. Experimented with different number of neighbours and found that the optimal results were obtained with k=12.

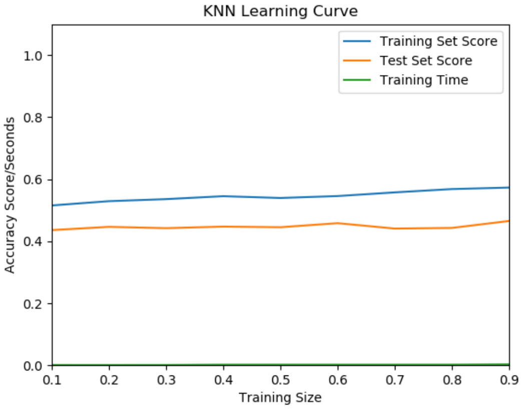
Adult Data:





Wine Data:

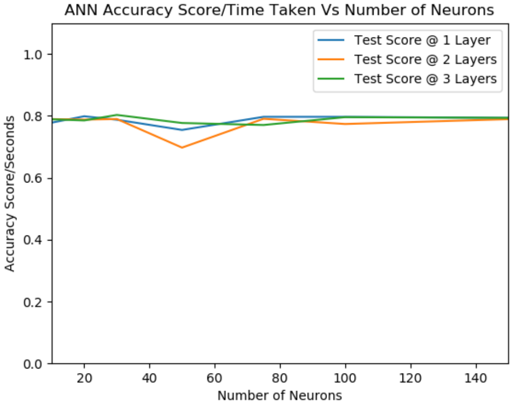


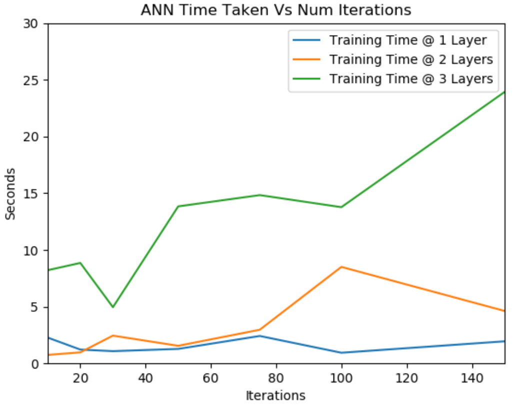


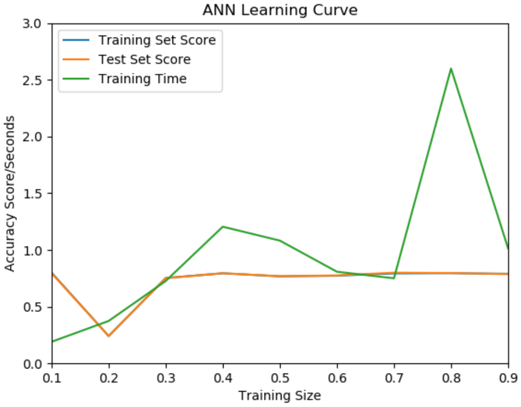
Neural Networks

I implemented Neural networks with SKlearn's MLPClassifier. Experimented with various hidden layer sizes. I tested the model with 1, 2 and 3 layers with different number of neurons. I found that a 2 layer with 30 neurons gave me good accuracy score within an acceptable time frame.

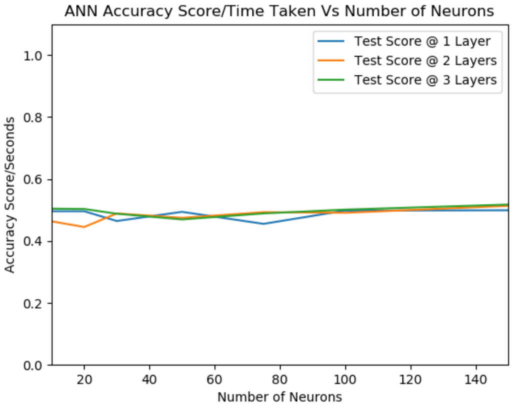
Adult Data:

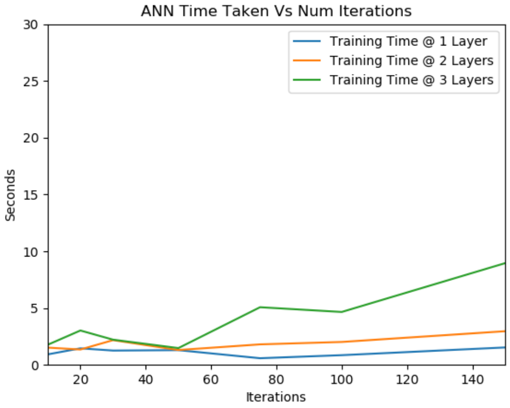


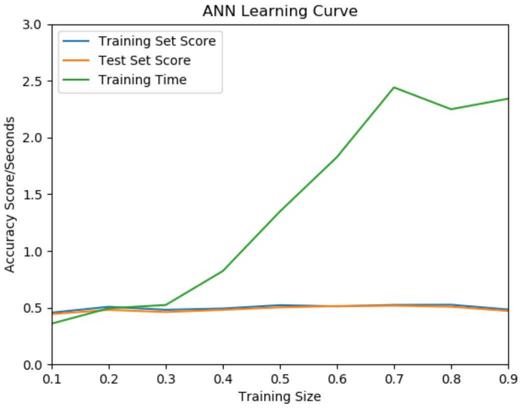




Wine Data:



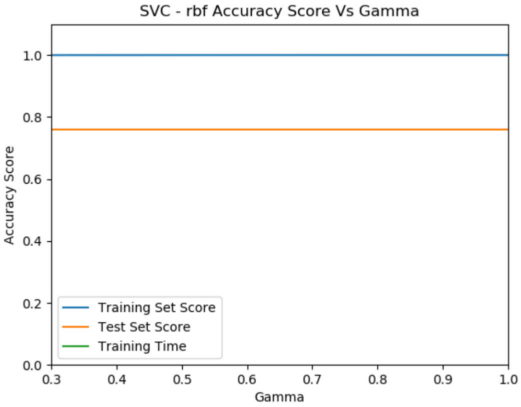




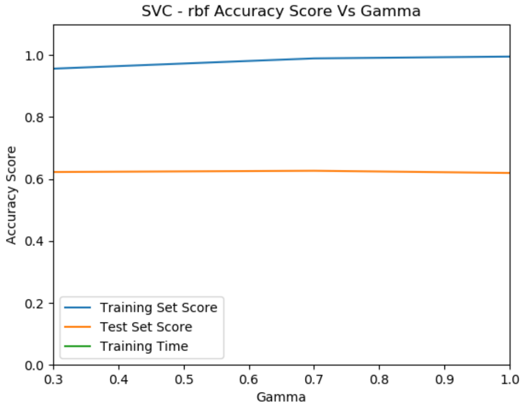
SVC

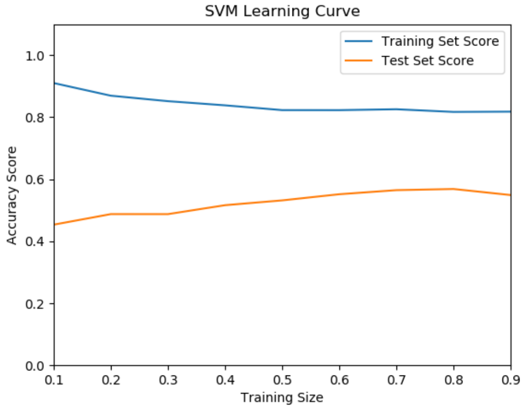
Implemented SVC with SKLearn's svm.SVC. Experimented with different Gamma values. I did not find much difference in the performance of the model or with the time taken to build the model with varying Gamma values.

Adult Data:



Wine Data:


SVC - rbf Accuracy Score Vs Gamma
Accuracy Score
Gamma
0.0
0.2
0.4
0.6
0.8
1.0
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Training Set Score
Test Set Score
Training Time


SVM Learning Curve
Accuracy Score
Training Size
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training Set Score
Test Set Score