Kuan-E Chao

CS 7641

Assignment 1: Supervised Learning Report

Classification Problem: which one is the best classifiers?

Which is better? We will look into two different and interesting datasets and compare their prediction rate in training and testing datasets. The higher prediction rate will be consider as the "better" classifiers.

About my datasets:

For the first data set I choose the quality of Red wine as the first data of my testing set. I choose this data because this a continuous variable which makes the prediction less than 50% some of the times, which is interesting. I want to see how these algorithms perform when the outcome is not 50-50 or true-or false situation.

The attributes for this data sets are: (11 dimensions)

fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol

The targets is the quality of the alcohol, which is made by wine experts.

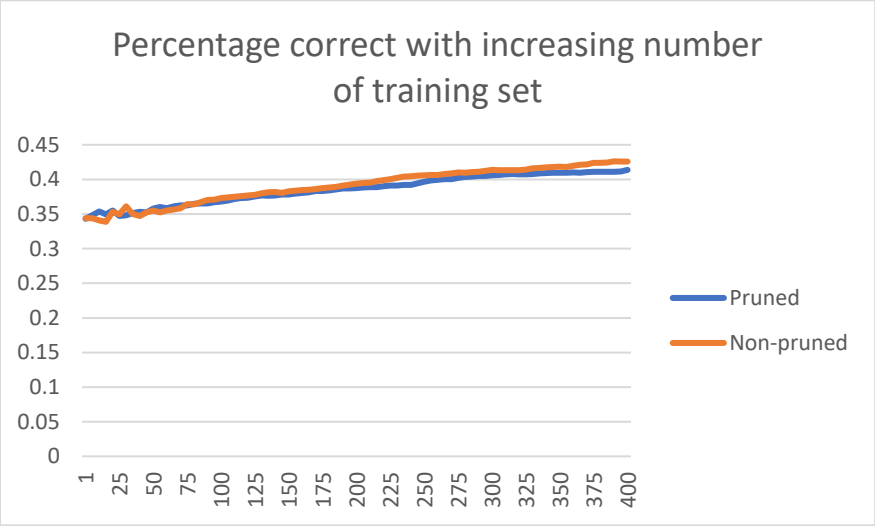
The second dataset I choose is to tell the origin of the wine base on the datasets. For the second data set I choose the wines in general. The purpose of the data set is to find the origin of the wine based on the chemical element of the dataset. I want to see in this kind of small dataset, how relatively accurate does each dataset predicts. It turns out these type of small data sets requires high accuracy, therefore I want to see which algorithm will provide highest accuracy.

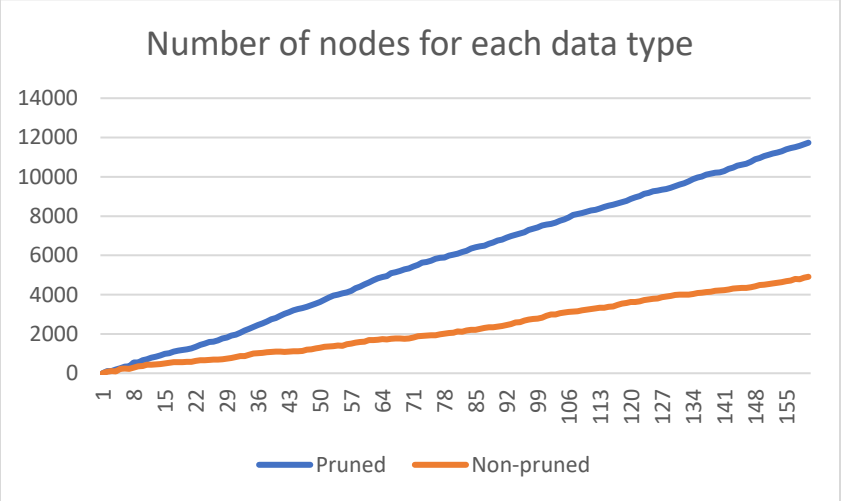
The attributes for this data sets are:

Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline

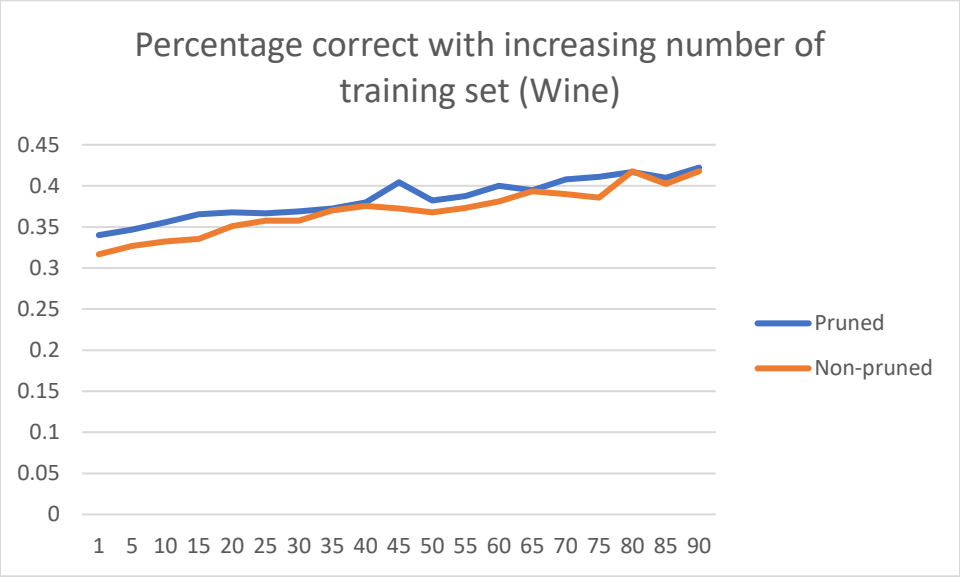
The target is the origin of the alcohol, which labeled 1 2 3, representing different regions.

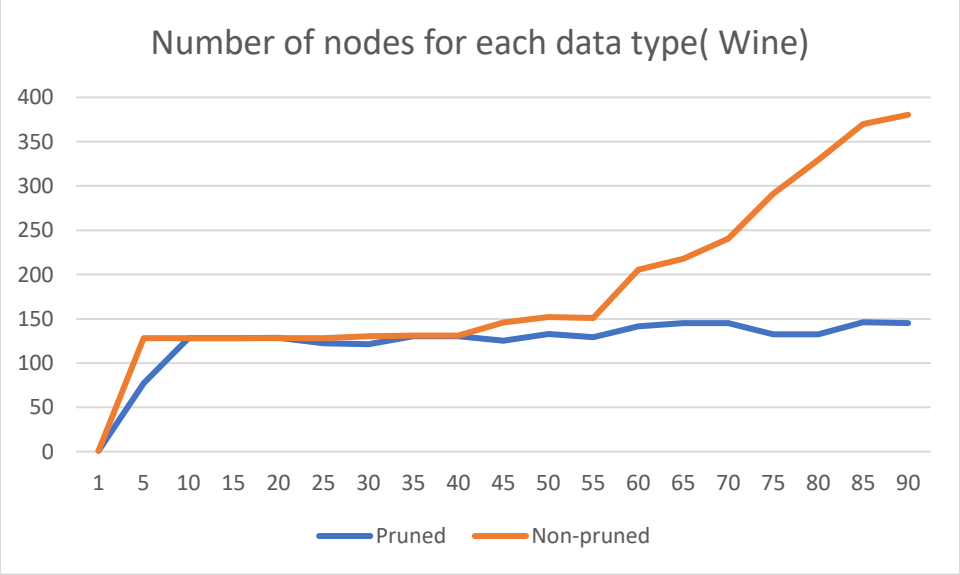
Decision Tree algorithm





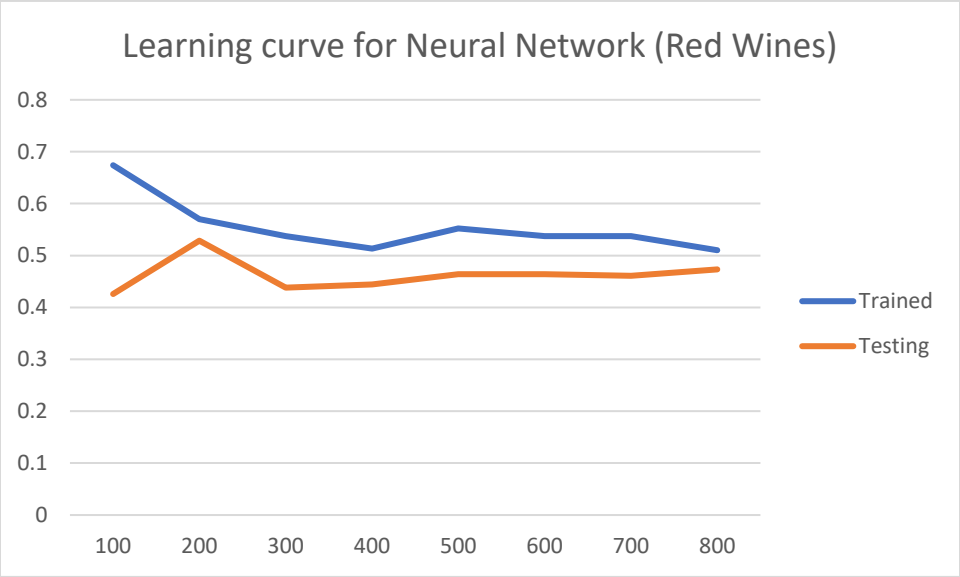
From the decision tree algorithm, we can see that the non-pruned method might have some overfitting because the percentage correct doesn't improve by a lot. There is only 1 percent of difference. However, Non-pruned trees tend to have more than double of number of nodes than the pruned trees.

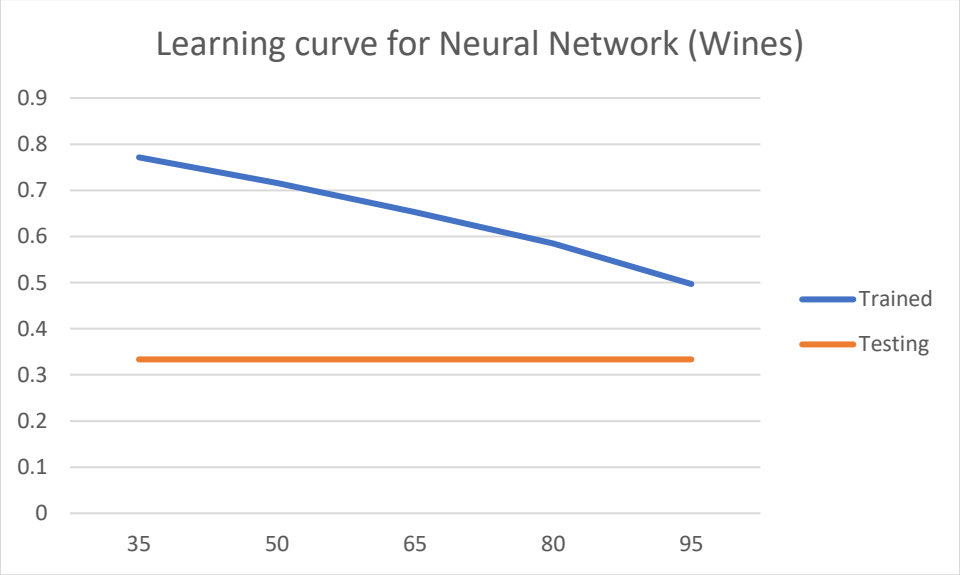




For the wine also follows the same idea. One thing that is noticeable is that the percentage correct is lower than expect. This is probably because this type of data does not work well in the decision tree since most of the attributes are continuous data, which makes the decision tree algorithm hard to distinguish the data.

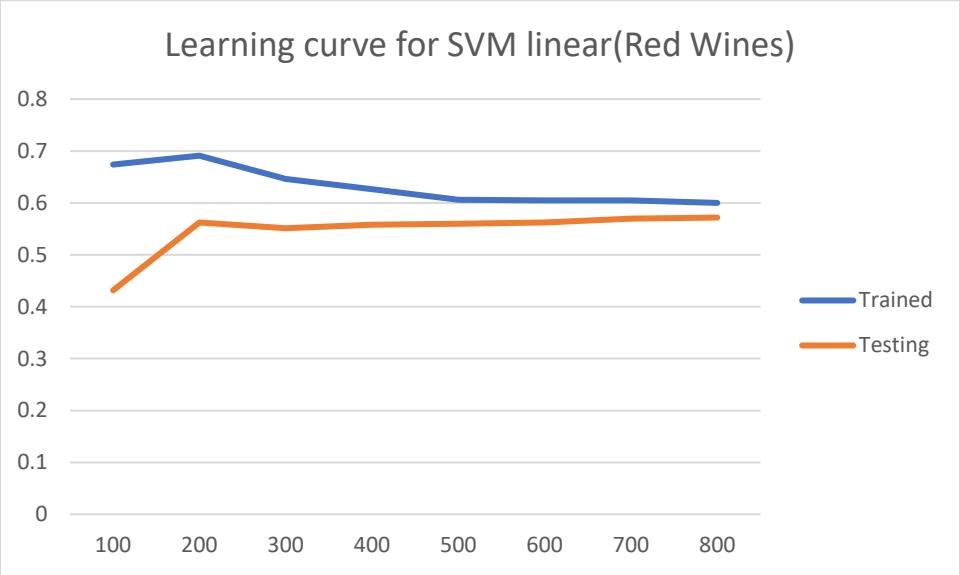
Neural Network

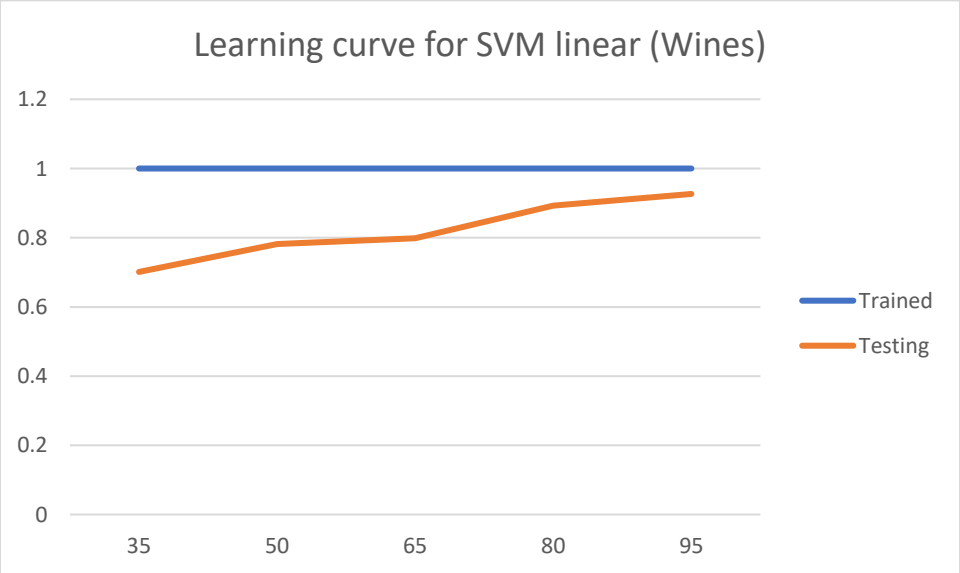




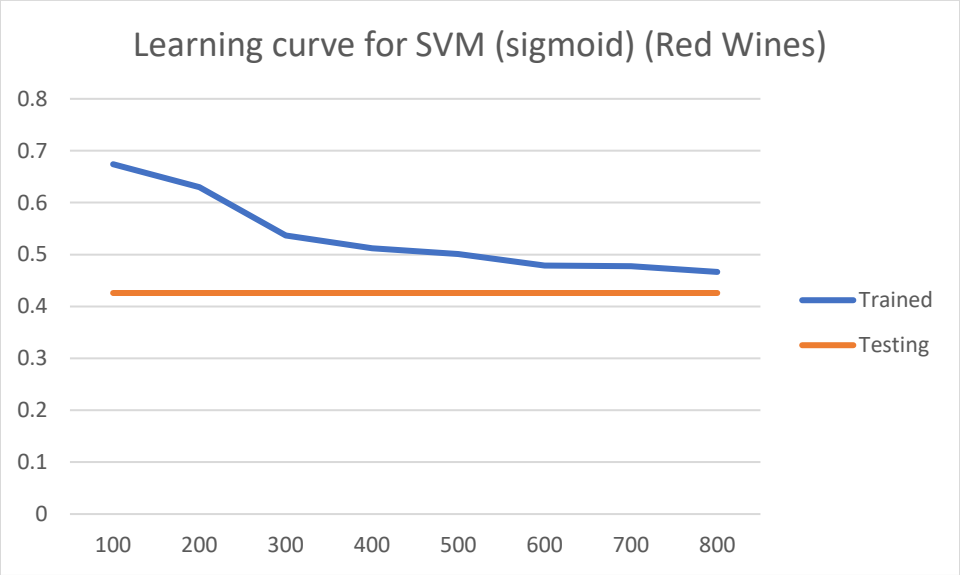
For the Neural network, we can see the testing accuracy is about 45% accurate, this is because Neural network provide weight on several categories which would make the prediction very consistent somewhat independent to number of testing samples.

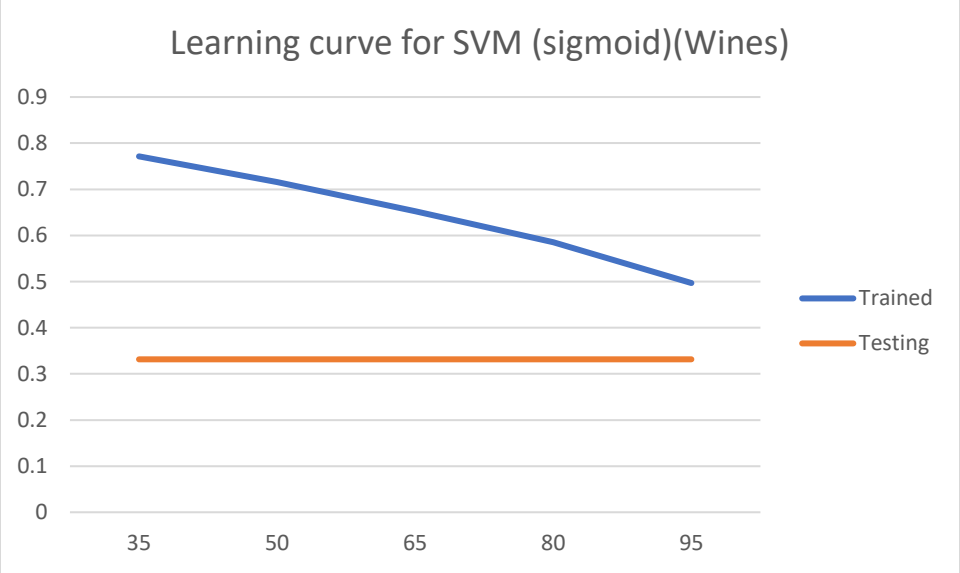
Support Vector Machine





For the support vector machine, I realized that even though the training data has relatively low accuracy rate compare to other classification algorithms, the actual predicted rate is actually the highest among all algorithms. We can also see that the trained data prediction goes lower and the predicted data increases its value as number of training data increases, which is the same as the prediction.

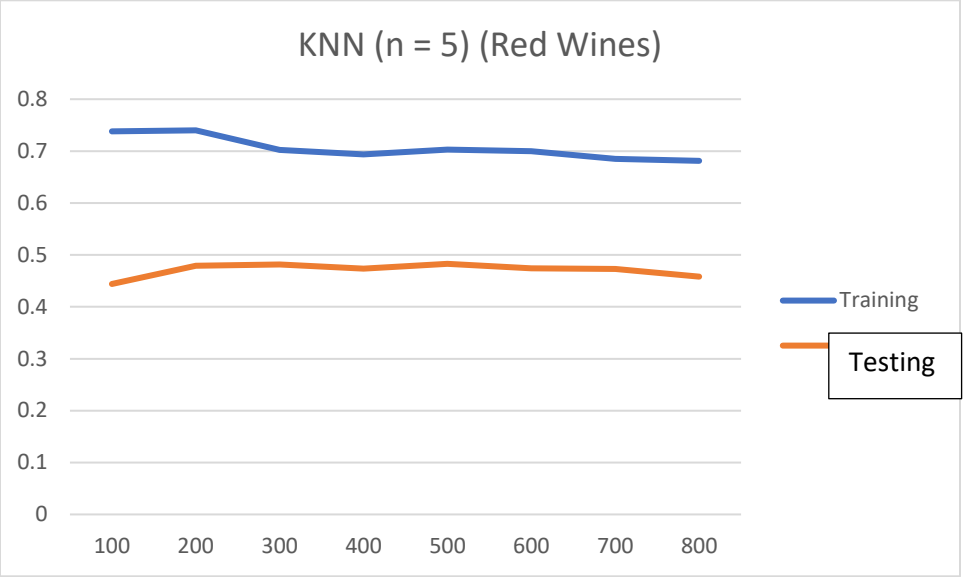


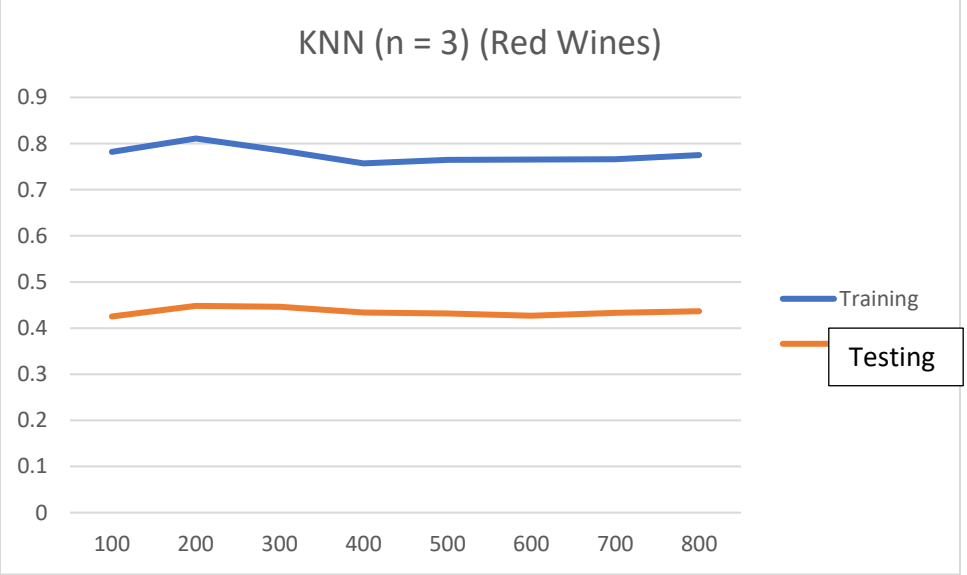


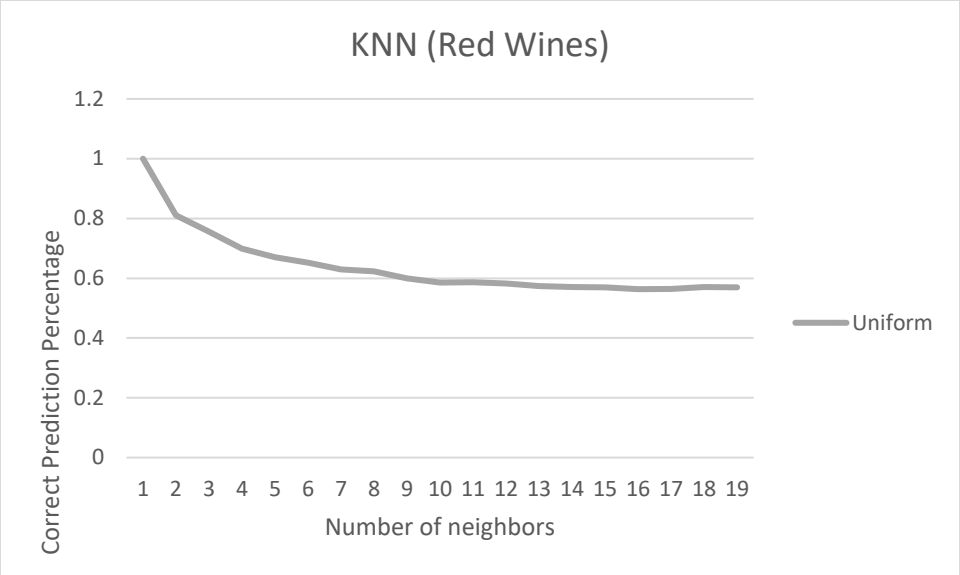
I tried both sigmoid and rbf as the parameter, I realized that both of these gives pretty low prediction accuracy for both of these data sets. One good reason would be the dataset would be enough if we use linear kernel instead of sigmoid and rbf. Therefore, linear kernel produce far

better results than the sigmoid kernel. I still have to mention that linear kernel in the SVM has the best result overall in the entire classifier in the supervised learning we have learned so far. With such a high training and testing percentage for both datasets, I would highly consider SVM as the main classifier if I am going to start on new data set in the real world or my future jobs. However, in the future, I do want to do some experiment in the future to see under what circumstance would rbf or sigmoid produce linear result compare to linear kernels.

KNN (Nearest Neighbors)

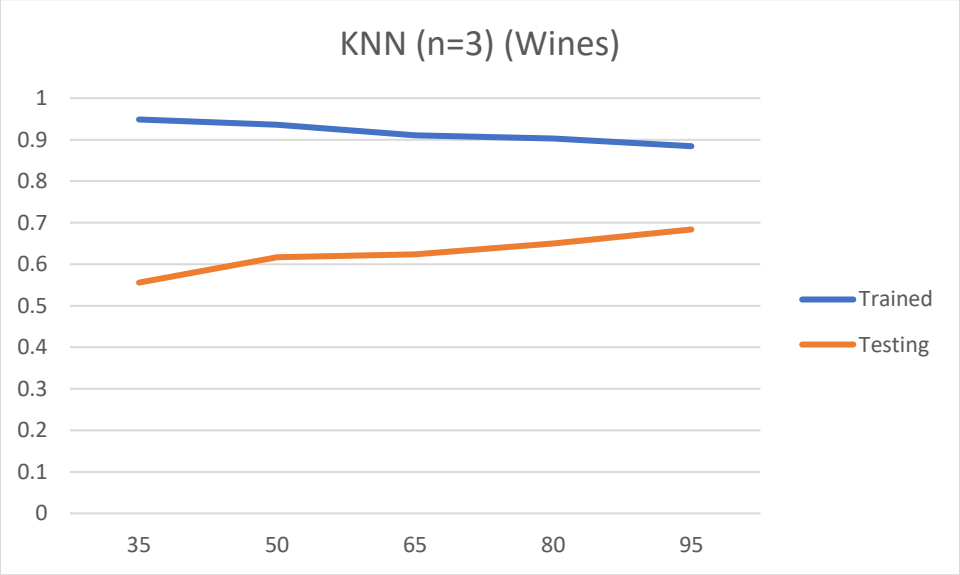


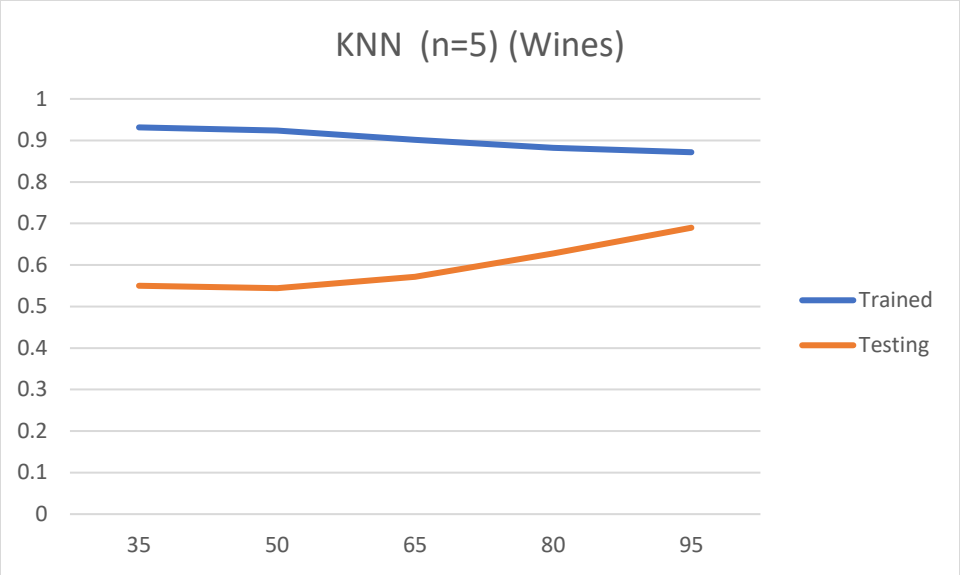


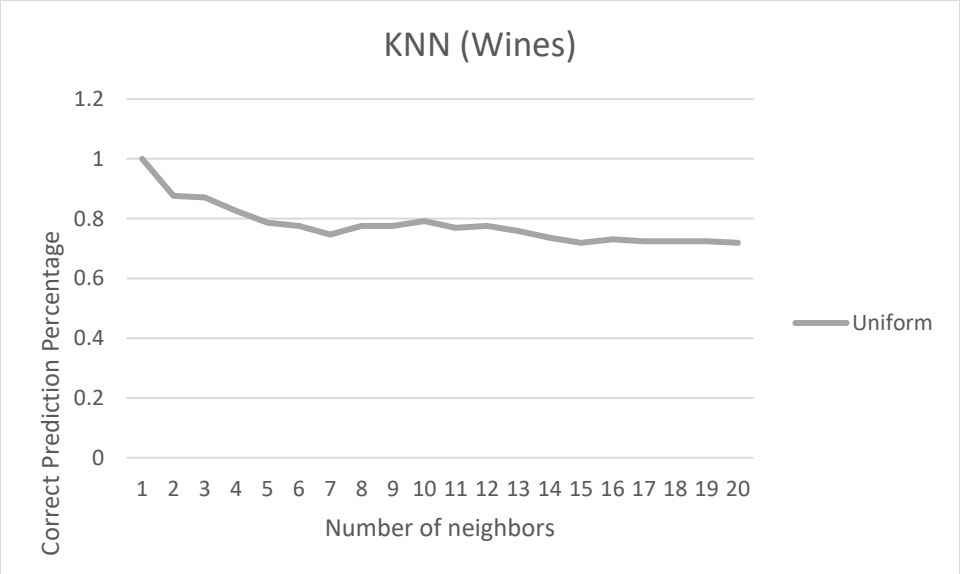


I first run the learning curve by number of neighbors and I realized that the learning curve does not changed by a significant amount, therefore I decided to graph the learning curve of the data versus the number of neighbors. I found out that the prediction percentage decreases as number of neighbors increase. This is reasonable because of the data will be less accurate if we use more neighbors, which might caused overfitting. In general, this dataset has high prediction rate in both training and testing sets when k is lower, which make sense in this particular type of data since the source is made from wine experts, therefore it is quite easy to cause overfitting because data around the boundaries are hard to justify.

And then we examined the wine datasets from the wine origin datasets:





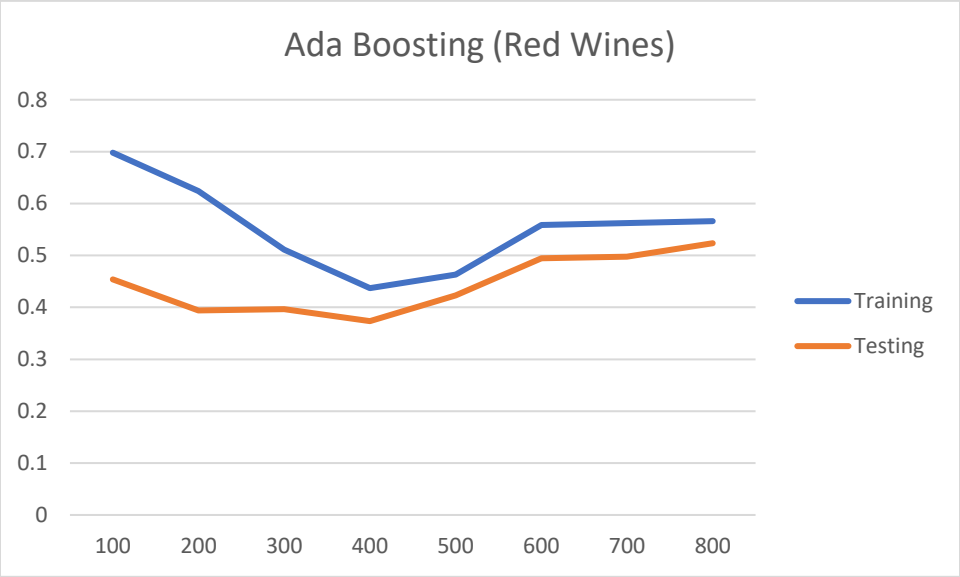


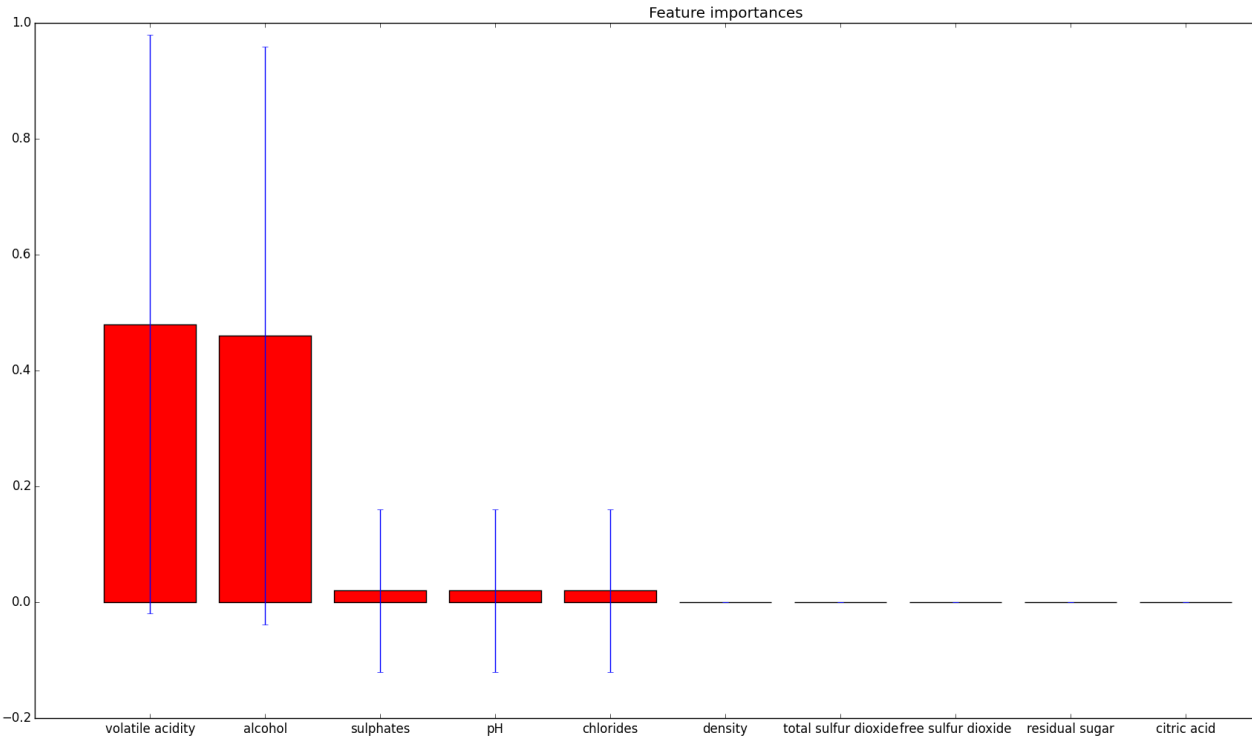
It turns out that no matter which datasets I used, when I use the nearest neighbors for the red both wine quality and wine data, I realized that the more neighbors I used turns out to be less convincing. The reason why this is true is because using more neighbors will possible cause overfitting which will make the training data has low accuracy. In general, KNN has relatively low frequency compare to other data, this is because of for this type of alcohol data, it is very hard to get the data from its neighbor.

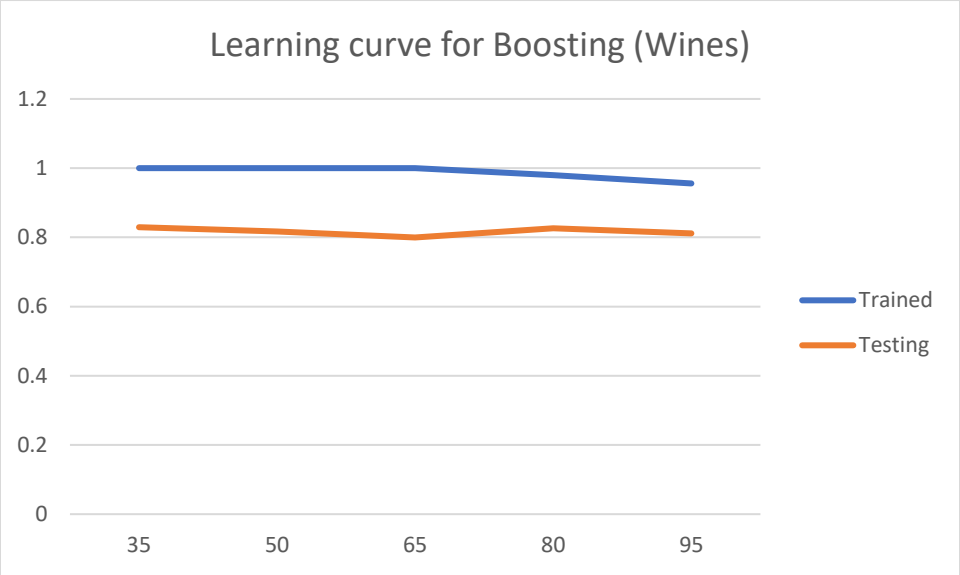
For the wine classification data, I realized that the nearest neighbors is really close to 100% when we use small amount of neighbors. I also believe that there is overfitting since the number of data is relatively small, therefore using more neighbors would cause overfitting.

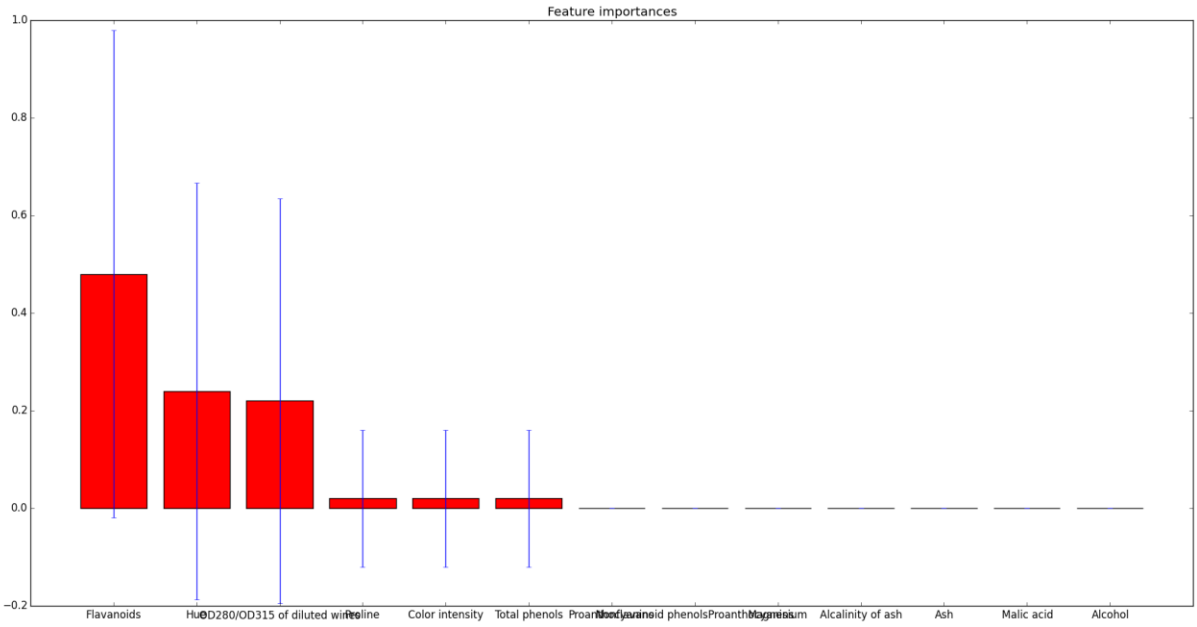
Boosting (I chose Ada Boosting)

The reason I choose Ada Boosting is because the weight of each learner is learned by whether it predicts a sample correctly or not. I think the weight for category is important in this type of data(wine), therefore I choose Ada boosting.









The Ada boosting appears to have high accuracy on both training data sets and testing data sets. The reason why this would be happened is because Ada boosting will make the categories that is not important weight less values, which will make the prediction more accurate. From Ada boosting we can also see which columns are the important factor of the predictions, which is really important. We can see that in general only 3 or 4 factors have significant weight.

Conclusion

There are many things that can be taken away from this assignment. First thing we can see is that these algorithms have good scores on the testing datasets. For the Wine datasets, since the number of the datasets are small, I kind of expect high percentage of predictions being accurate. For the red wine quality datasets, we can see that the predictions are low since the datasets are huge and the quality score are made by wine experts. We can see that these supervised learning are very good at identifying the datasets that is systematic, such as origin of the alcohol. However, we can also see that it is relatively weak on identifying the human behaviors, such as the quality scores for the alcohol rated by the wine experts.

We can also see that SVM and boosting are good classifiers for these datasets. The important reason would be the datasets focus a lot on the continuous variables, and we can see that these two classifiers are good for classifying continuous datasets because it has high testing scores. I believe that if I can do this project again, I will choose datasets with larger amount of data and see how those the standard error of classifiers performed in the large data sets.

Citations:

Scikitlearn. Neural network, Adaboosing, Nearest neighbors, Support Vector Machines

Numpy, Pandas

Allen Madse, Decision Tree.

UC Irvine's Machine Learning datasets:

Wine:

Original Owners:

Forina, M. et al, PARVUS -

An Extendible Package for Data Exploration, Classification and Correlation.

Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno,

16147 Genoa, Italy.

Donor:

Stefan Aeberhard, email: stefan '@' coral.cs.jcu.edu.au

Wine Quality:

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez

A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal

@2009