Machine Learning Unsupervised Learning Algorithms Analysis

The Datasets

I used the same data sets from the first homework. I wanted to see how supervised and unsupervised learning compared. You may notice that the description of the data sets is very similar to the previous homework.

Student Alcohol Consumption

Description

This data set takes a number of factors and compares it against the amount of alcohol a student consumes. The population of the data set consists of students in one of two schools attributed in the data taking either a Portuguese or Mathematics course. The data was initially separated by subjects in two separate csv files and contained course final grades in the subject. Since my primary interest was to gauge correlation of other variables to alcohol consumption I formatted the data to exclude course grades. I also combined the two files and added an extra variable indicating the course a student was taking. Two factors were used to assess alcohol consumption: workday alcohol consumption and weekend alcohol consumption, both on a scale from 1 to 5. I combined both these factors, and created a categorical variable, High. If total alcohol consumption was greater than or equal to six, alcohol consumption was considered high. There were two reasons for doing this. The first being that a one to ten scale is too broad and would likely have caused large errors in many of the algorithms. The second reason is that establishing a clear boundary ensures that the analysis is more meaningful in explaining what the algorithm predicts.

Why this data set?

This data while only containing 1044 data points had 30 categorical variables, 27 of which I converted to numeric variables and used. I considered this to be a more practical use of machine learning algorithms. A parent field of machine learning is statistics. In statistics calculating the influence of multiple variables to a single variable is achieved through multiple correlation against a linear function. I wanted to see how useful the algorithms are in comparison to traditional statistics.

Poker Hands

Description

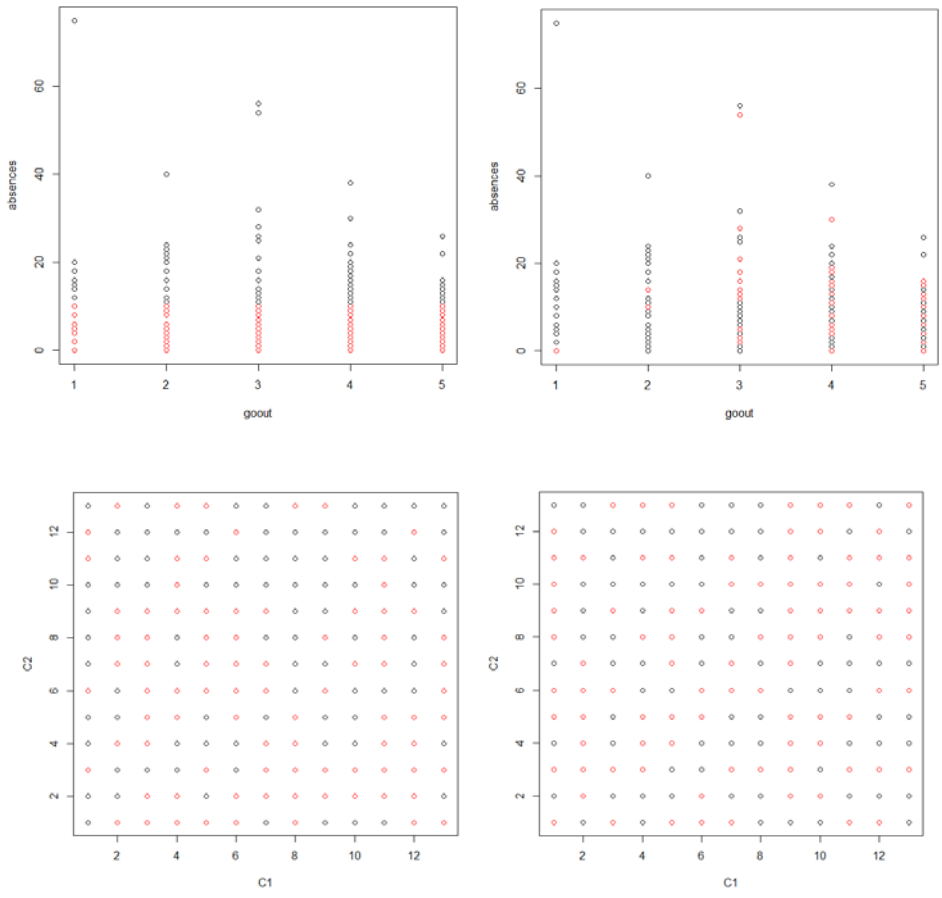
This data set contains values of all five cards and their respective suits matched with the outcome of the hand as a whole. Initially this data set contained the outcome of the hand as a whole ranked from 0-9, 0 being nothing and a 9 being a royal flush. I chose to convert the outcomes to a binary variable called Something. If the hand contained something then Something was a yes, else something was a no. The reason for this was that roughly 50% of the hands in both sets contained nothing and roughly 42% of the hands contained one pair, leaving the remaining 8% to be distributed over the remaining outcomes, a one to one correspondence between these outcomes and

their percentages appeared logarithmic, meaning that the higher outcomes would rarely occur; every combination above three-of-a-kind occurred less than 1% of the time combined.

Why this data set?

The number of combinations of a deck of hands is 52!. The number of winning poker hand combinations is upwards of three hundred million. I find this data set fascinating because the set winning combinations is very small compared to all possible hands. It is possible to program the rules of poker; I was interested to see how well a machine learning algorithm performs in comparison.

K-Means



Graphical Predictions of K-Means. Student Alcohol Consumption Top, Poker Hands Bottom

K Means

Description

The R algorithm for decision trees uses the Hartigan-Wong algorithm by default. It is also possible to use the MacQueen Algorithm as well as the Lloyd/Forgy Algorithm. The k-Means algorithm seeks to minimize the sum of squares from randomly assigned centers.

I have shown above the graphical predictions of K-Means on two variables from the data set. These graphs are a good visual representation of the differences between true clusters and k-means clusters. A better representation are tables mapping the clusters to true outputs. The tables are reproduced below:

	No	Yes
1	117	46
2	733	148

Student Alcohol Consumption

	No	Yes
1	6232	6201
2	6261	6316

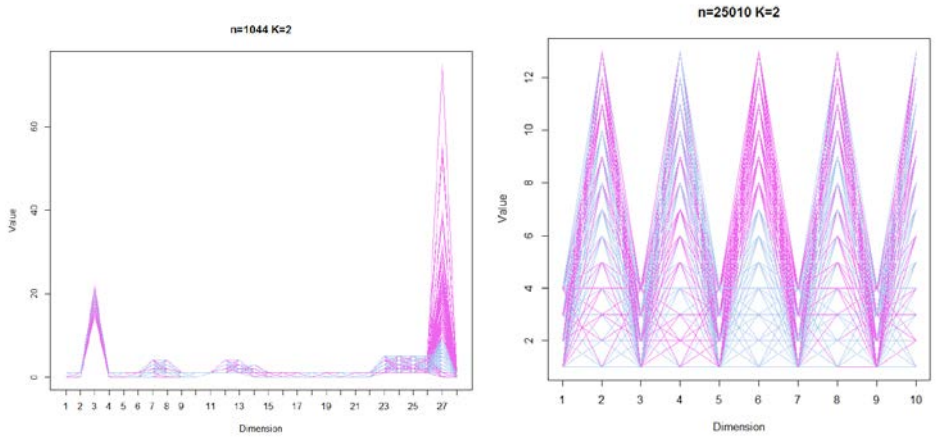
Poker Hands

Analysis

The student alcohol consumption produces a 0.253 error, I found this surprising since the algorithm knows nothing about the actual groupings as is the nature of unsupervised learning. This performed almost as well as a supervised neural network.

In comparison the unpruned error of the Poker Hands tree was 0.497 on average, which still satisfies the criteria for a weak classifier. In fact, it performed about as well as the decision tree algorithm on poker hands.

Expectation Maximization



EM Plots. Student Alcohol Consumption Left. Poker Hands Right.

Description

The R Expectation Maximization Algorithm creates clusters based on the assumption that the variables have a Gaussian distribution.

I have shown above the graphical predictions of K-Means on two variables from the data set. These graphs are a good visual representation of the differences between true clusters and k-means clusters. A better representation are tables mapping the clusters to true outputs. The tables are reproduced below:

	No	Yes
1	426	96
2	424	98

Student Alcohol Consumption

	No	Yes
1	6720	6802
2	5773	5715

Poker Hands

Analysis

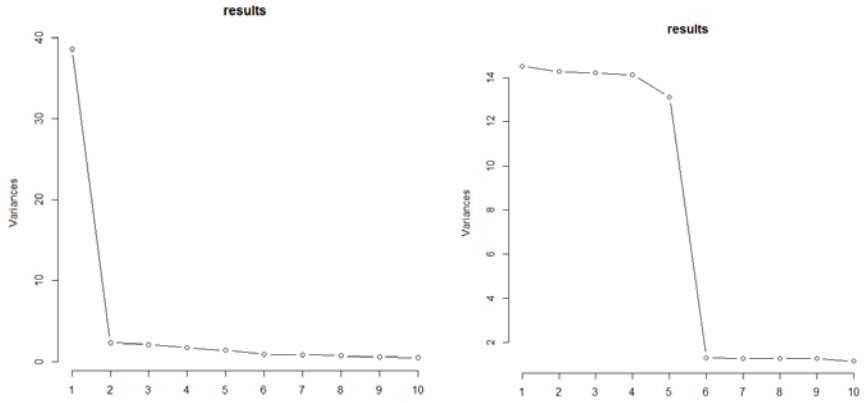
The algorithm produced an unexpectedly large error for student alcohol consumption: 0.498. This indicates that the output does not produce a normal distribution. Of course we know this from our domain knowledge. There are far more “no” results than “yes” for high student alcohol consumption.

The algorithm produced the same amount of error for poker hands as the k-means algorithm: 0.497. The distributions for outputs here are quite even but we know that the outputs are unevenly distributed within the dataset from our knowledge of poker hand composition.

PCA

Description

The Principal Component Analysis Algorithm computes related vectors from given variables in a manner that minimizes variance. As a result, some of these PC vectors can be ignored in later analysis.



Variance mapped against Principal Components. Student Alcohol Consumption Left. Poker Hands Right.

Analysis

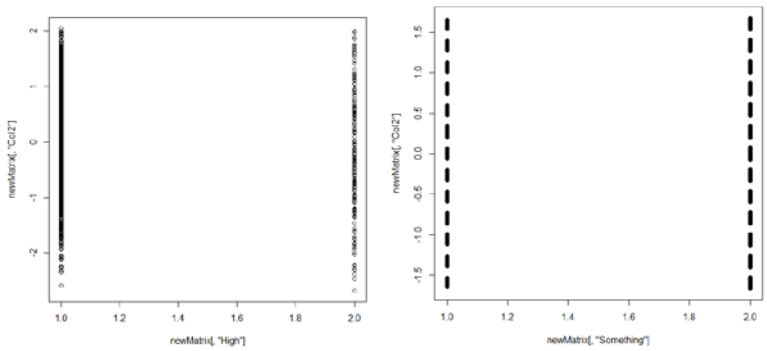
In the Student Alcohol Consumption data set it appeared that roughly 95% of the variance in the dataset was accounted for by the first 10 PC vectors. Effectively, this allowed us to reduce our input set to 10 variables instead of the 27 we originally had.

In the Poker Hands data set the first 5 Principal Components accounted for 90% of the variance in the dataset. Logically this makes sense as our ten components are really representations of five cards.

ICA

Description

Independent Component Analysis predicts the independent components in a dataset and performs a dimension reduction by removing the dependent components. In R's implementation of ICA, you have to list the number of independent components that may exist. I chose this number to be 5 for both datasets.

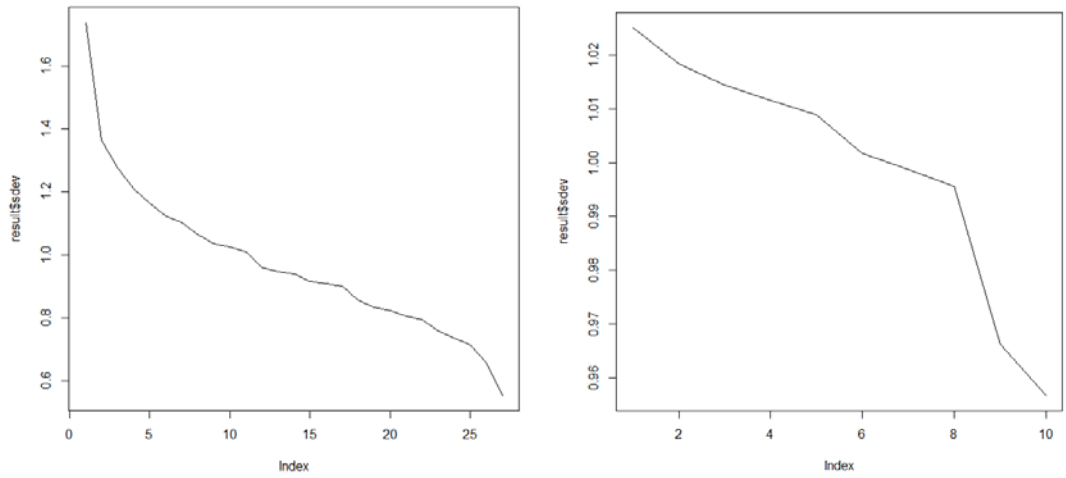


ICA. Student Alcohol Consumption Left. Poker Hands Right.

Analysis

In both data sets we effectively reduced the dimensions to 5. This is based on an assumption that 5 dimensions can accurately account for most of the data in the data set. When we test our ICA against our neural network we will know for sure.

Randomized Projections



Graph of standard deviation over random projections. Student Alcohol Consumption Left. Poker Hands Right.

Description

The Randomized Projections Algorithm produces randomized projection vectors over our data set. I graphed the standard deviations of the projections.

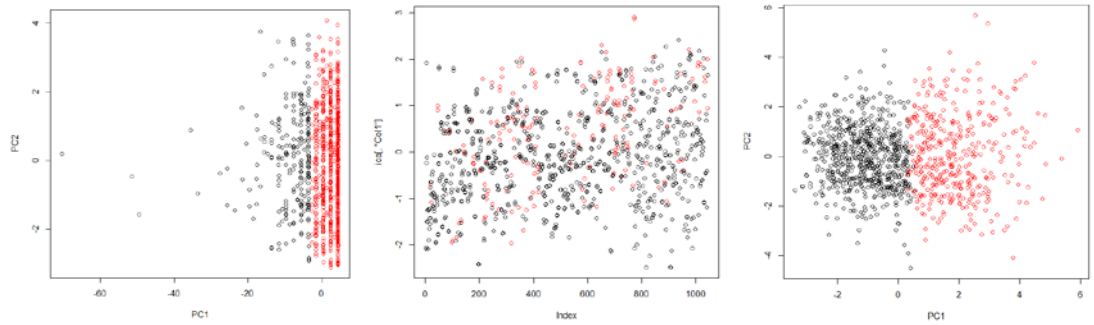
Analysis

Since these were random we were unable to make any reductions but perhaps the random projections would yield better results than the PCA or ICA.

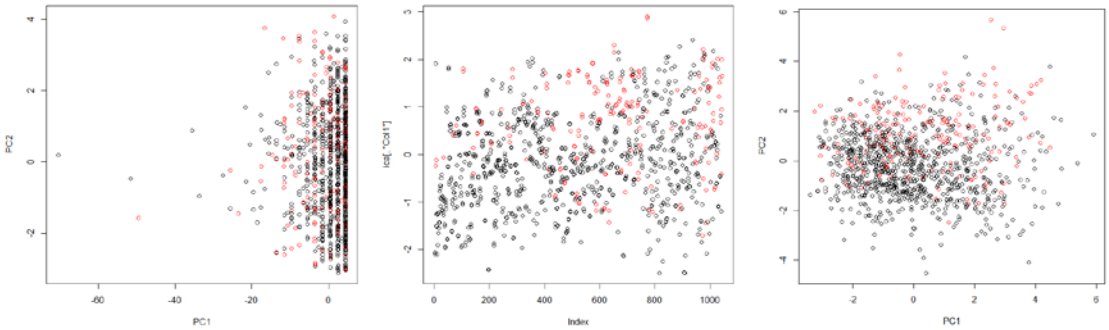
Clustering with Dimensionality

Description

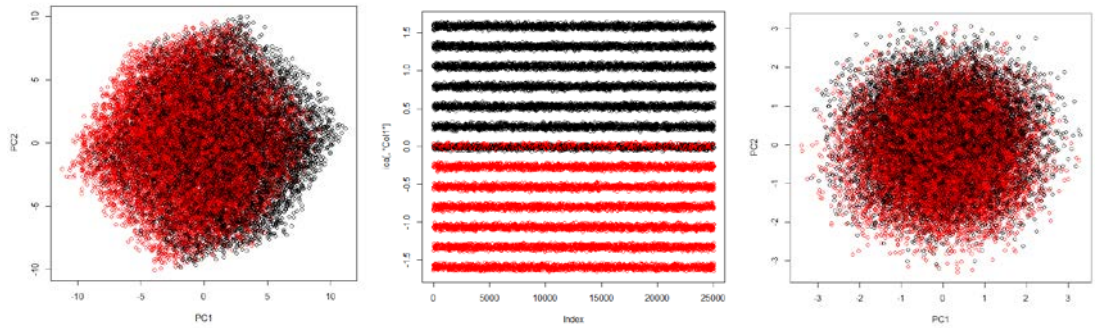
Here we took the clustering algorithms and the dimensionality algorithms and combined them to create a better predictor for our data sets.



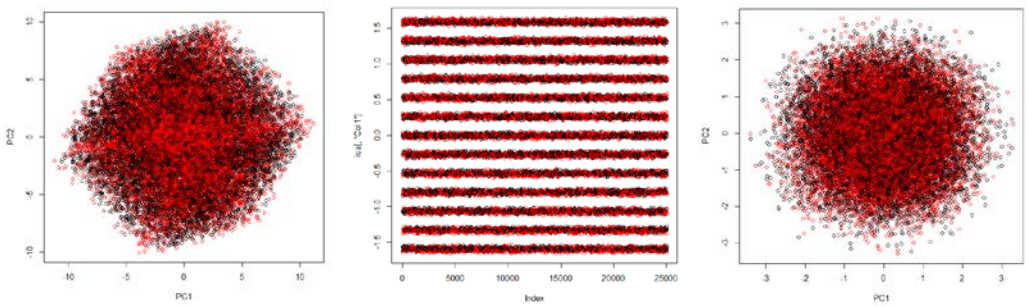
Predicted PCA, ICA, Random Projections Student Alcohol Consumptions K-Means Above.



Actual PCA, ICA, Random Projections Student Alcohol Consumption Means Above.



Predicted PCA, ICA, Random Projections Poker Hands K-Means Above.



Actual PCA, ICA, Random Projections Poker Hands Means Above.

Analysis

For brevity I will not include the graphs for Expectation Maximization, but they are included in the Graphs folder.

For the Student Alcohol Consumption set PCA K Means had a 0.249 Error, ICA K Means had a 0.208 Error, and Randomized Projection K Means had a 0.305 Error. In contrast PCA EM had a 0.450 Error, ICA EM had a 0.307 Error, and Randomized Projection had a 0.256 Error.

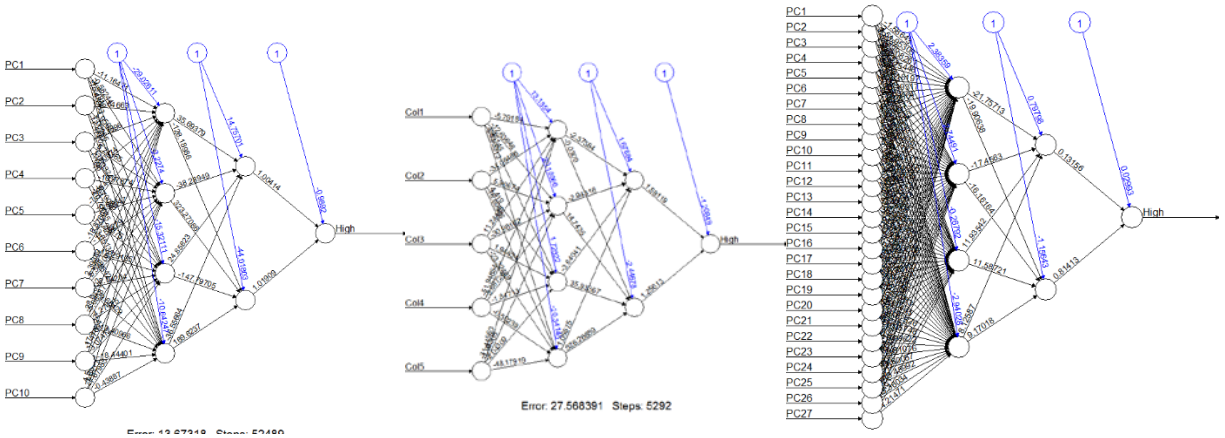
For the Poker Hands set PCA K Means had a 0.499 Error, ICA K Means had a 0.500 Error, and Randomized Projection K Means had a 0.499 Error. In contrast PCA EM had a 0.499 Error, ICA EM had a 0.497 Error, and Randomized Projection had a 0.497 Error.

Clearly Poker Hands did not improve, but on average Student Alcohol Consumption did better at predicting with dimension reduction than without. We can still consider Poker Hands set's error as an achievement because we accomplished nearly the same error with less dimensions.

Clustering with Neural Networks

Description

Here we used the Student Alcohol Consumption Data Set and used the results from the clustering algorithms to create neural networks.



Neural Networks Produced from PCA, ICA, and Randomized Projections respectively.

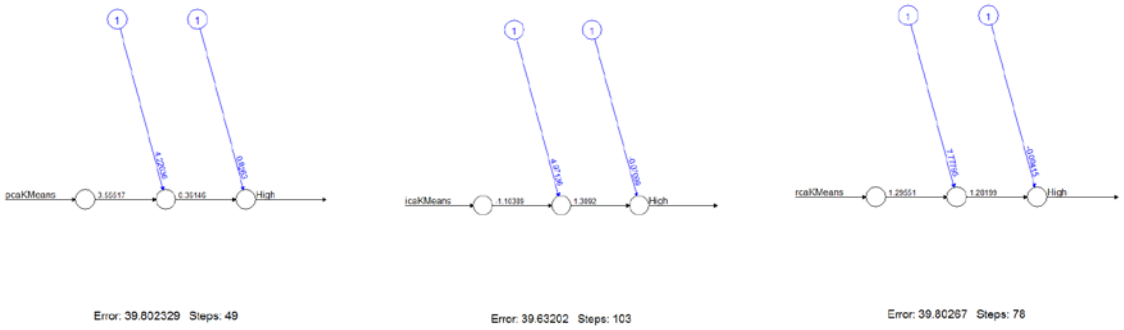
Analysis

For PCA our neural network produced a 0.224 error, for ICA it produced a 0.199 error, and for Randomized Projections it produced a 0.186 error. It might appear that Random Projections do better simply because there are more components, but by that logic PCA should have done better than ICA. It is the case here that Random Projections are just better at predicting data that isn't linearly separable.

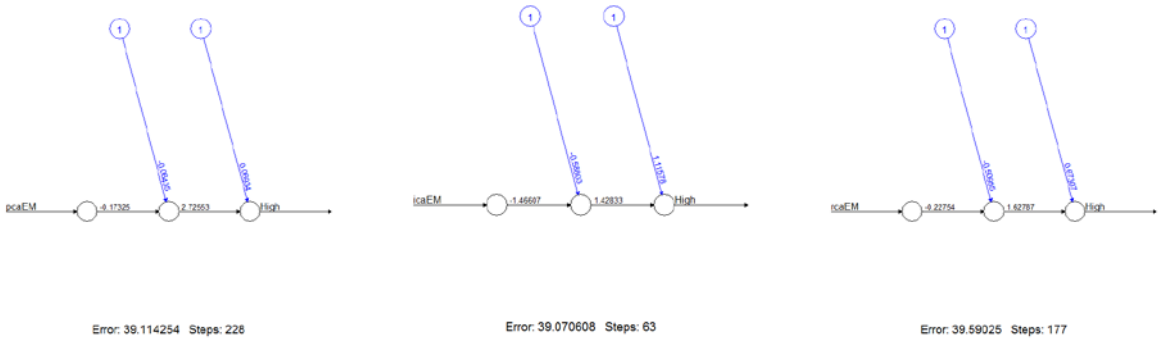
Neural Networks with Clustering and Dimensionality

Description

Here we used the Student Alcohol Consumption Data Set, Clustered and gave them Dimensionality and created Neural Networks.



Neural Networks of K-Means on PCA, ICA, and Random Projections of the Student Alcohol Consumption data set.



Neural Networks of Expectation Maximization on PCA, ICA, and Random Projections of the Student Alcohol Consumption data set

Analysis

If you look at the Neural Networks graphed above, you realize that the look pretty much the same. This shouldn't be a surprise as the EM and K-Means algorithms both produce a single vector that lists the clusters of each row given. This was then mapped against the value of High alcohol consumption. The

results for all of these were equally similar, in fact they were exactly the same. All 6 neural networks produced the same error: 0.1839. This is not very surprising because the neural network was essentially doing a linear mapping of inputs to outputs. Often the output and inputs were the same, as the inputs were simply the predictions of the outputs.