Assignment: HW1- SL

CS7641

Uma Batta
ubatta3@gatech.edu

gatech id: ubatta3

Introduction

This project aims at exploring some techniques in supervised machine learning models. The five models include

- Decision tree
- Neural networks
- Boosting
- Support Vector Machines
- K-nearest neighbors

The classification problem that we will focus on is to investigate the factors that affect the quality of the red wine. This kind of research is very crucial in all aspects of life which includes winemaking. The factors that have been suggested to influence the quality of wine are plant's environment, Temperature, CO2, pH, Alcohol. The second classification problem aims at classifying the diagnosis using the cell nuclei present in the image of a fine needle aspirate (FNA) of a breast mass. The data contains 32 variables and 569 rows. Similarly, the same algorithm techniques will be applied on the data. The results from the classification problems will be outlined below. Why I chose these are different reaons, wine making was about the small dataset and breast cancer was out of knowledge from personal experience with my aunt.

Exploratory Data Analysis

Red wine dataset:

The following figure shows the classification of the quality of the alcohol.

Fig 1.0: Dist of Alcohol quality

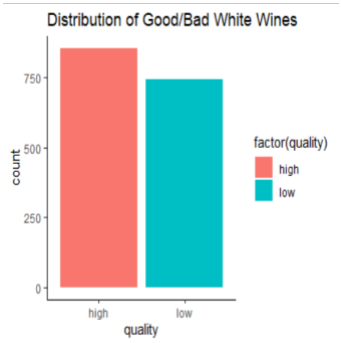


Table 1.1

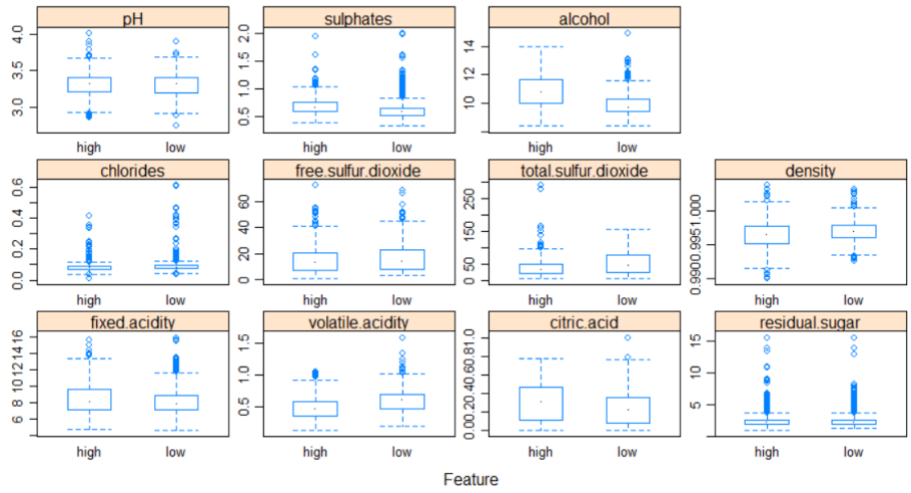
	vars	n	mean	sd	median	trimmed	mad	min	max	range
fixed.acidity	1	1599	8.32	1.74	7.90	8.15	1.48	4.60	15.90	11.30
volatile.acidity	2	1599	0.53	0.18	0.52	0.52	0.18	0.12	1.58	1.46
citric.acid	3	1599	0.27	0.19	0.26	0.26	0.25	0.00	1.00	1.00
residual.sugar	4	1599	2.54	1.41	2.20	2.26	0.44	0.90	15.50	14.60
chlorides	5	1599	0.09	0.05	0.08	0.08	0.01	0.01	0.61	0.60
free.sulfur.dioxide	6	1599	15.87	10.46	14.00	14.58	10.38	1.00	72.00	71.00
total.sulfur.dioxide	7	1599	46.47	32.90	38.00	41.84	26.69	6.00	289.00	283.00
density	8	1599	1.00	0.00	1.00	1.00	0.00	0.99	1.00	0.01
pH	9	1599	3.31	0.15	3.31	3.31	0.15	2.74	4.01	1.27
sulphates	10	1599	0.66	0.17	0.62	0.64	0.12	0.33	2.00	1.67
alcohol	11	1599	10.42	1.07	10.20	10.31	1.04	8.40	14.90	6.50
quality*	12	1599	1.47	0.50	1.00	1.46	0.00	1.00	2.00	1.00

	skew	kurtosis	se
fixed.acidity	0.98	1.12	0.04
volatile.acidity	0.67	1.21	0.00
citric.acid	0.32	-0.79	0.00
residual.sugar	4.53	28.49	0.04
chlorides	5.67	41.53	0.00
free.sulfur.dioxide	1.25	2.01	0.26
total.sulfur.dioxide	1.51	3.79	0.82
density	0.07	0.92	0.00
pH	0.19	0.80	0.00
sulphates	2.42	11.66	0.00
alcohol	0.86	0.19	0.03
quality*	0.14	-1.98	0.01

The figure shows that the 855 observations were high quality while 744 observations were of low

The table 1.1 above shows various values of the red wine. The average fixed acidity of the red wine was 8.32 (SD = 1.71) and the minimum was 4.6 while the maximum acidity of the red wine was 15.9. The average volatile acidity of the red wine was 0.53 (SD = 0.18) and the minimum was 0.12 while the maximum was 1.58. The average citric acidity of the red wine was 0.27 (SD = 0.19) and the min 0.00 while the max was 1.00. The average residual sugar content 2.54 (SD = 1.41) and min was 0.9 while the max was 15.5. The average chloride content is 0.09 (SD = 0.05) and the min was 0.01 while the max is 0.6. The average free sulfur dioxide content is 15.87 (SD = 1.41) and the min is 1 while the max is 71. The average total sulfur dioxide is 46.47 (SD = 32.9) and the min was 6 while the max was 289. The average density is 1 (SD = 0) and the min was 0.99 while the max is 1. The average pH 3.31 (SD = 0.15) and the min pH was 2.74 while the maxpH of the red wine was 4.01. The average alcohol content is 10.42 (SD = 1.07) and the min was 8.4 while the max was 14.9.

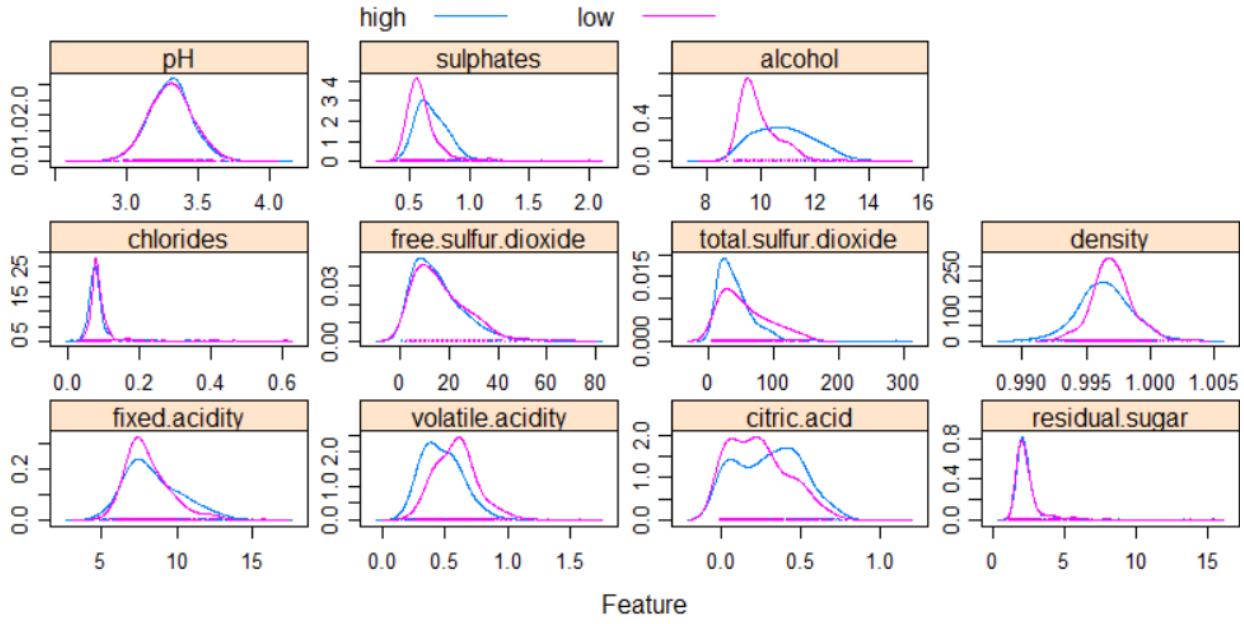
Fig 1.2: Distribution according to the red wine content according to the quality using boxplot



The boxplot above shows that high-quality red wine had the highest pH compared to the low-quality red wine. However, the graph suggests that the median pH for both the high and low-quality red wine was the same. Similarly high-quality highest level of sulfates compared to the low-quality. High-quality had the highest level of alcohol content compared to the low-quality red wine. The low-quality red wine had the highest chloride content compared to the high-quality red wine. The high and the low-quality red wine had an almost equal level of free sulfur dioxide contents. The low-quality red wine had the highest level of total sulfur dioxide content compared to the high-quality red wine. The low and high-quality red wine had almost homogeneous density. The high-quality red wine had a higher fixed acidity content compared to the low-quality red wine. The low-quality red wine had a higher volatile acidity content compared to the high-quality red wine. The high-quality red wine had a higher citric acidity content compared to the low-quality red wine. The high-quality red wine had a higher residual sugar content compared to the low-quality red wine.

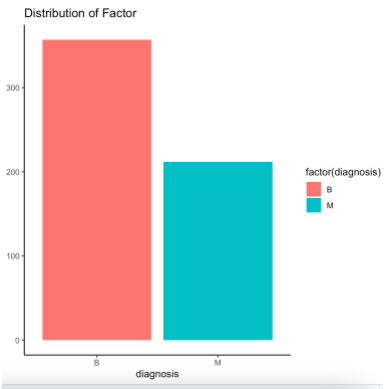
The same information can be seen using the density graph as shown below

Fig 1.3: Distribution according to the red wine content according to the quality using density



Breast Cancer:

The second classification problem aims at classifying the diagnosis using the cell nuclei present in the image of a fine needle aspirate (FNA) of a breast mass. The data contains 32 variables and 569 rows. Based on the data, I tried to analyze the features to help me to predict malignant or benign cancer. For this will use the classification methods to classify the type of cancer as malignant or benign. The diagnosis column gives the data as Malign or Benign as M or B respectively. As Machine Learning algorithms needs numerical values to plot graphs, I have replaced M with 1 and B with 0. Below is the plot of Benign and Malign count.



Implementation and Analysis:

Decision Tree

A decision tree is one of the supervised learning algorithms technique that can solve problems that involve classification and regression. The advantage of the decision tree is that it can use both factor and numeric variables to create its model. It has a root node that represents the entire sample or population. The root nodes are divided into homogeneous sets. Splitting takes place by diving the node into several sub-nodes. The sub-nodes are further split to form the decision node. Terminal nodes are the ones that do not split.

Pruning is the process that involves the removal of sub-nodes of a decision node. (Song & Ying, 2015; Bhargava et al. 2015; Delen, Kuzey & Uyar, 2013; Hssina et al. 2014).

Decision tree classification predicts each observation by the most commonly occurring class of training observations in the area that the observation belongs.

Red wine dataset:

The red wine data set was portioned into the training and testing data set at 80 % and 20 % intervals. The training data set was used to build the decision tree model and testing sets were used to evaluate the performance of the decision tree. The quality variable was used as the target variable while the rest was the response variable.

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  124  48
      low    39  98

               Accuracy : 0.7184
                 95% CI : (0.6647, 0.7679)
    No Information Rate : 0.5275
    P-Value [Acc > NIR] : 5.282e-12

                  Kappa : 0.4334

 Mcnemar's Test P-Value : 0.3911

            Sensitivity : 0.7607
            Specificity : 0.6712
         Pos Pred Value : 0.7209
         Neg Pred Value : 0.7153
             Prevalence : 0.5275
         Detection Rate : 0.4013
   Detection Prevalence : 0.5566
      Balanced Accuracy : 0.7160

       'Positive' Class : high
```

Fig 1.1 – Intial Conf Matrix.

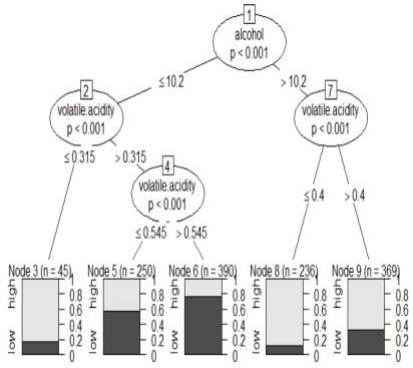


Fig 1.2 Post-Pruning DTree ROC

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  116  45
      low    47 101

               Accuracy : 0.7023
                 95% CI : (0.6479, 0.7527)
    No Information Rate : 0.5275
    P-Value [Acc > NIR] : 2.831e-10

                  Kappa : 0.4031

 Mcnemar's Test P-Value : 0.917

            Sensitivity : 0.7117
            Specificity : 0.6918
         Pos Pred Value : 0.7205
         Neg Pred Value : 0.6824
             Prevalence : 0.5275
         Detection Rate : 0.3754
   Detection Prevalence : 0.5210
      Balanced Accuracy : 0.7017

       'Positive' Class : high
```

Fig 1.3 – Post Pruning Conf mtx

The decision model produced 71.84 % accuracy on the testing data set. The confusion matrix shows that 124 observation which was of high quality were accurately predicted while 39 observations were wrongly predicted. On the other hand, 98 observations that were of low quality were accurately predicted while 38 observations were wrongly predicted.

Single decision tree modelled with high variance and low bias. The confusion matrix against the Training dataset gave around 83% accurancy which is overfitting the model as Testing accuracy was 71.84%. To avoid the overfitting, Pruning was applied to the data and the following decision tree plot was obtained.

The following observations can be made from the decision tree plot

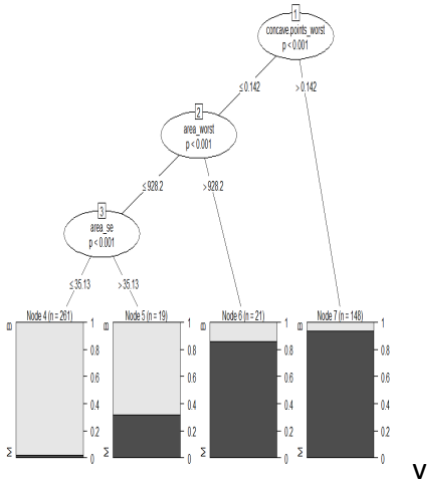
- The alcohol content and volatile acidity were the best predictors of the quality of the red wine.
- The red wine that had an alcohol content that is equal or less than 10.2, volatile acidity equal to or less than 0.315 has 80 % chances of being of high quality.
- The red wine that had an alcohol content that is equal or less than 10.2, volatile acidity equal to or less than 0.545 has about 40 % chances of being of high quality.
- The red wine that had an alcohol content that is greater than 10.2 and volatile acidity greater than 0.4 has about 60 % chances of being of high quality.
- The red wine that had an alcohol content that is greater than 10.2 and volatile acidity equal to or less than 0.4 has about 90 % chances of being of high quality.

The model evaluation was also obtained after pruning and the following result was obtained. The model accuracy was 70.23 %. The miscalculation error of the decision tree was 29.77 %.

Breast Cancer dataset:

The following decision tree was obtained after pruning

```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 283  14
         M   8 154

               Accuracy : 0.9521
                 95% CI : (0.9283, 0.9697)
    No Information Rate : 0.634
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8959

 Mcnemar's Test P-Value : 0.2864

            Sensitivity : 0.9725
            Specificity : 0.9167
         Pos Pred Value : 0.9529
         Neg Pred Value : 0.9506
             Prevalence : 0.6340
         Detection Rate : 0.6166
   Detection Prevalence : 0.6471
      Balanced Accuracy : 0.9446

       'Positive' Class : B
```



v

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 63  2
         M  3 42

               Accuracy : 0.9545
                 95% CI : (0.8971, 0.9851)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9057

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9545
            Specificity : 0.9545
         Pos Pred Value : 0.9692
         Neg Pred Value : 0.9333
             Prevalence : 0.6000
         Detection Rate : 0.5727
   Detection Prevalence : 0.5909
      Balanced Accuracy : 0.9545

       'Positive' Class : B
```

Intial CM Dtree after pruning CM post pruning

The diagram above shows that the concave point's worst, area worst, area_se were the best predictors of the diagnosis. The figure also shows that

- When the Concave points worst is less or equal to 0.142, the area worst is equal or less than 928.2 and the area_se is equal or less than 35.13 then there is 99 % chances of the image being diagnosed as benign
- When the Concave points worst is less or equal to 0.142, the area worst is equal or less than 928.2 and the area_se is greater than 35.13 then there is 70 % chances of the image being diagnosed as benign
- When the Concave points worst is less or equal to 0.142, the area worst is greater than 928.2 then there is 20 % chances of the image being diagnosed as benign
- When the Concave points worst is greater than 0.142, then there is 90 % chances of the image being diagnosed as malignant

Model validation of testing set

The model showed that 63 out of 65 observations that were diagnosed as benign was correctly predicted and 42 out of 44 observations that was diagnosed as benign was correctly predicted. The model accuracy was 95.45 %. The model accuracy for the training sets was 94.99 %. The classification error for the model was obtained to be 4.5 %.

Neutral Network

The neural network model is compared to the human nervous system in that it is characterized by an activation function that interconnects information processing units that transform input into an output.

The neural network has layers that pass information. The first layer receives the data which is processed and passed to the hidden layers. The hidden layers passed the information to the last layers that produce the output. (Acharya et al. 2017; Li et al. 2017; Kia et al. 2017; Biancofiore et al. 2017).

The neutral model was applied to the data and the model produced the following graph. The neural network showed that 2 hidden layers were used. The black lines show the weights while the blue lines show the bias term. The neural network also produced the best predictors of the quality of the red wine as shown in the figure below. It shows that the best predictor of the quality of the red wine was the chlorides level, followed by alcohol content and lastly the pH of the red wine. The model accuracy was obtained using the testing data and the Table 2.3 output was obtained. The neural network model produced 76.05 % accuracy on the testing data set. The confusion matrix shows that 123 observation which was of high quality were accurately predicted while 40 observations were wrongly predicted. On the other hand, 112 observations that were of low quality were accurately predicted while 34 observations were wrongly predicted.

Fig 2.1: Neutral Network graph

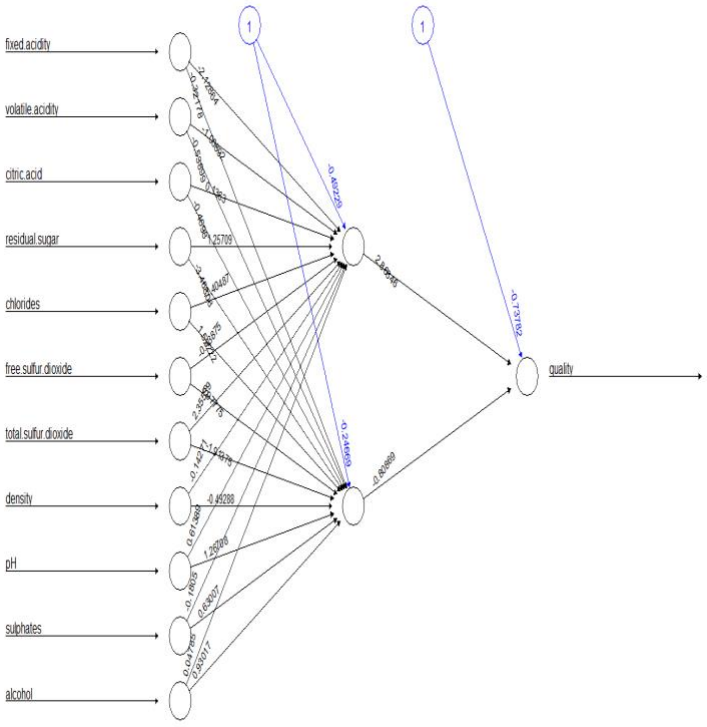


Fig 2.4: Variable importance

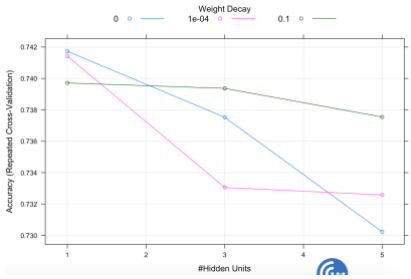
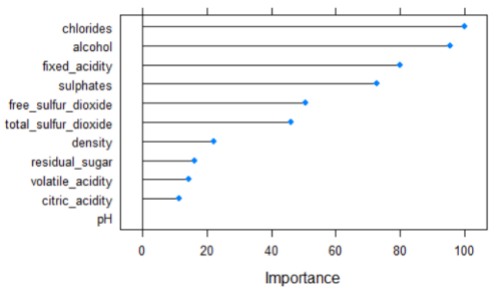


Table 2.5: Neural network model evaluation



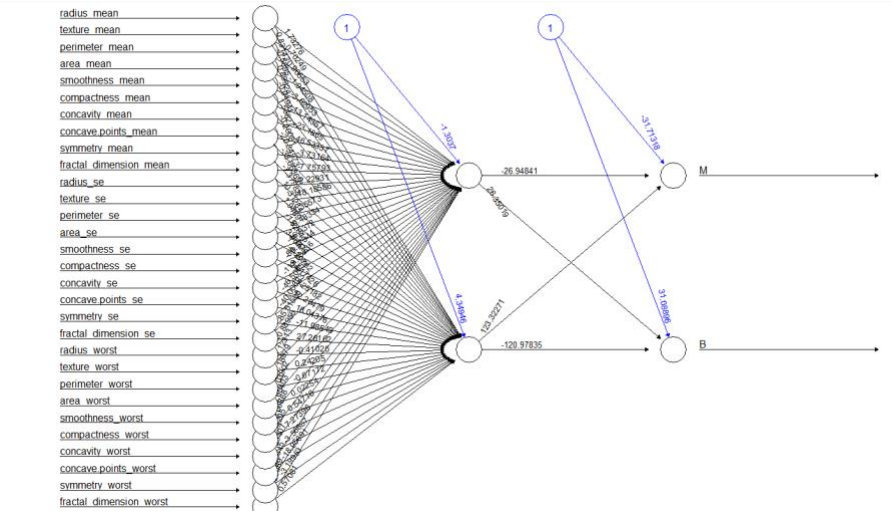

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  123  34
      low    40 112

               Accuracy : 0.7605
                 95% CI : (0.7089, 0.807)
    No Information Rate : 0.5275
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5206

 Mcnemar's Test P-Value : 0.5611

            Sensitivity : 0.7546
            Specificity : 0.7671
         Pos Pred Value : 0.7834
         Neg Pred Value : 0.7368
             Prevalence : 0.5275
         Detection Rate : 0.3981
   Detection Prevalence : 0.5081
      Balanced Accuracy : 0.7609

       'Positive' Class : high
```

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  544 143
      low   148 455

               Accuracy : 0.7744
                 95% CI : (0.7506, 0.797)
    No Information Rate : 0.5364
    P-Value [Acc > NIR] : <2e-16
```

To start, the neural network performs but as learning rate increases, the performance suffers from high bias. Have used the grid tuning for improving the performance. The grid search parameter tuning helped to increase the permance. The new improved performance was 77.5%

Set B – Breast cancer:

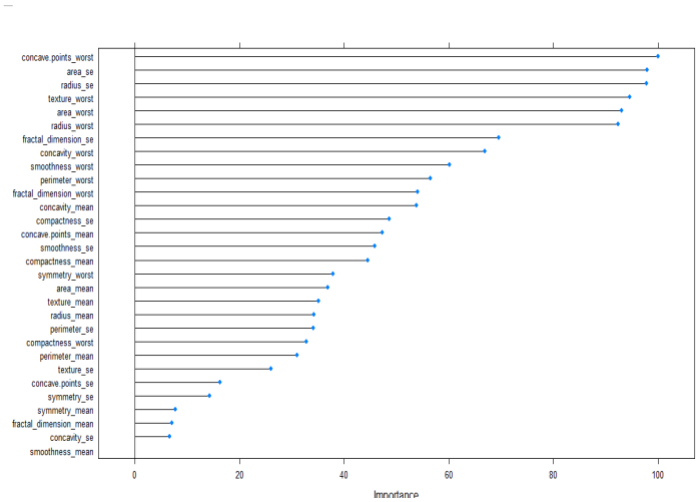
The neural network was conducted using 2 hidden layers. The following graph was obtained



Best predictors of the diagnosis

Model evaluation

The model showed that 63 out of 66 observations that were diagnosed as benign was correctly predicted and 44 out of 44 observations that was diagnosed as benign was correctly predicted. The model accuracy was 97.27 %. The model accuracy for the training sets was 99.13 %.



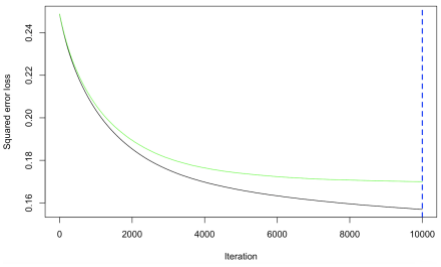

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 63  0
         M  3 44

               Accuracy : 0.9727
                 95% CI : (0.9224, 0.9943)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9438

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 0.9545
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9362
             Prevalence : 0.6000
         Detection Rate : 0.5727
   Detection Prevalence : 0.5727
      Balanced Accuracy : 0.9773

       'Positive' Class : B
```

The diagram above shows that concave points worst was the best predictor of diagnosis, followed by area_se, radius_se. The worst predictor was smoothness mean. The Accuracy of prediction was 97%, so didn't use the hyperparameter tuning.

Boosting

Boosting classification algorithm is a method that consists of multiple subsets of the training data set and build several independent tree models. It later finds the best performing predictive model. The way it creates its model is unique in such that each successive tree is grown using the information obtained from the tree which was previously grown (James et al. 2014; Ke et al. 2017). In this manner, it minimizes the error obtained from the previous models. There are different variety of boosting (Natekin & Knoll, 2013; Guelman, 2012; Xi et al. 2017; Hassan & Bhuiyan, 2017). They include *Adaboost*, *gradient boosting,* and *stochastic gradient boosting*. In this project, we will use stochastic gradient boosting. This will be made possible by using the xgboost package in R. *Redwine and Breast cancer datasets training accuracy increases initially and then converges to 1. The CV score gradually increases for set1 and attains max where set2 increases and then becomes flat based on the outliers in the data as shown in the box plot. The* model was created using the training data and the model prediction was evaluated testing data sets.

Red wine



```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  134  31
      low    29 115

               Accuracy : 0.8058
                 95% CI : (0.7572, 0.8484)
    No Information Rate : 0.5275
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6102
```

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  133  33
      low    30 113

               Accuracy : 0.7961
                 95% CI : (0.7468, 0.8396)
    No Information Rate : 0.5275
    P-Value [Acc > NIR] : <2e-16
```

Fig 3.3 Boosting CM Redwine Post tune

The boosting classification model produced 80.58 % accuracy on the testing data set. The confusion matrix shows that 134 observation which was of high quality were accurately predicted while 29 observations were wrongly predicted. On the other hand, 115 observations that were of low quality were accurately predicted while 31 observations were wrongly predicted. Root mean square of the model was at highest and with each iteration it becomes almost near to zero as the model improves.

HyperGrid parameter method was used for tuning. But post tuning, the performance didn't improve but decreased as shown to 79.61% accuracy.

Breast Cancer

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 64  4
         M  2 40

               Accuracy : 0.9455
                 95% CI : (0.8851, 0.9797)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8855

 Mcnemar's Test P-Value : 0.6831

            Sensitivity : 0.9697
            Specificity : 0.9091
         Pos Pred Value : 0.9412
         Neg Pred Value : 0.9524
             Prevalence : 0.6000
         Detection Rate : 0.5818
   Detection Prevalence : 0.6182
      Balanced Accuracy : 0.9394

       'Positive' Class : B
```

```
Confusion Matrix and Statistics

          Reference
Prediction high low
      high  133  33
      low    30 113

               Accuracy : 0.7961
                 95% CI : (0.7468, 0.8396)
    No Information Rate : 0.5275
    P-Value [Acc > NIR] : <2e-16
```

Fig 3.4 Boosting CM Breast Cancer.

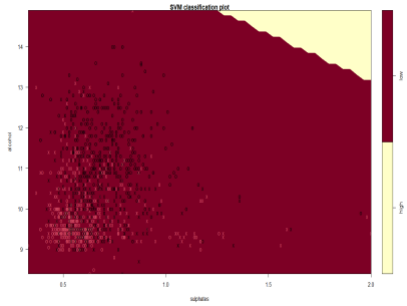
The model showed that 64 out of 66 observations that were diagnosed as benign was correctly predicted and 40 out of 44 observations that was diagnosed as benign was correctly predicted. The model accuracy was 94.55 %. The model accuracy for the training sets was 100 %.

Support Vector Machines

SVM is a supervised machine learning algorithm under the classification algorithm. This means that it can classify data into different classes. It used a hyperplane that acts as a decision boundary between various classes. It is used to generate multiple hyperplanes which divide the data into segments and each segment is homogeneous. It classifies the non-linear using kernels. The kernel transforms data into different dimensions that have a clear diving margin. Later it draws a hyperplane between the classes of data (Tehrany et al. 2015; Deo, Wen, & Qi, 2016; Karimi et al. 2016; Yeganeh et al. 2012).

In this project, the kernel-mode was build using different kernels i.e. linear, radial etc.

Wineset:



```
          Reference
Prediction high low
     high  125  39
     low    38 107

               Accuracy : 0.7508
                 95% CI : (0.6987, 0.798)
    No Information Rate : 0.5275
    P-Value [Acc > NIR] : 5.579e-16

                  Kappa : 0.4999

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7669
            Specificity : 0.7329
         Pos Pred Value : 0.7622
         Neg Pred Value : 0.7379
             Prevalence : 0.5275
         Detection Rate : 0.4045
   Detection Prevalence : 0.5307
      Balanced Accuracy : 0.7499

       'Positive' Class : high
```

The linear model produced a model with 74.43 % accuracy on the testing set. The model build using radial kernel produced a model that was 70.55 % accurate on testing data and 95.74 % accurate on the training set. The model build using a polynomial kernel produced a model that was 53.07 % accurate on the testing set and 49.22 % accurate on the training set. The model build using the sigmoid kernel produced a model that was 52.75 % accurate on the testing set and 53.64 % accurate on the training set. The following figure was obtained while using a linear kernel

The model shows that a high level of sulfates and alcohol produced high-quality red wine. Also the alcohol and sulfate content was one of the best predictors of the quality of red wine.

The fine-tune showed that the radial kernel produced the best model. The following table shows the model performance on the testing data. The SVM classification model produced 75.08 % accuracy on the testing data set. The confusion matrix shows that 125 observation which was of high quality were accurately predicted while 38 observations were wrongly predicted. On the other hand, 107 observations that were of low quality were accurately predicted while 39 observations were wrongly predicted.

Breast Cancer:

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 66  5
         M  0 39

               Accuracy : 0.9545
                 95% CI : (0.8971, 0.9851)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9035

 Mcnemar's Test P-Value : 0.07364

            Sensitivity : 1.0000
            Specificity : 0.8864
         Pos Pred Value : 0.9296
         Neg Pred Value : 1.0000
             Prevalence : 0.6000
         Detection Rate : 0.6000
   Detection Prevalence : 0.6455
      Balanced Accuracy : 0.9432

       'Positive' Class : B
```

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 66  1
         M  0 43

               Accuracy : 0.9909
                 95% CI : (0.9504, 0.9998)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.981

 Mcnemar's Test P-Value : 1

            Sensitivity : 1.0000
            Specificity : 0.9773
         Pos Pred Value : 0.9851
         Neg Pred Value : 1.0000
             Prevalence : 0.6000
         Detection Rate : 0.6000
   Detection Prevalence : 0.6091
      Balanced Accuracy : 0.9886

       'Positive' Class : B
```

```
Confusion Matrix and Statistics

          Reference
Prediction  B  M
         B 66 44
         M  0  0

               Accuracy : 0.6
                 95% CI : (0.5022, 0.6922)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.5413

                  Kappa : 0

 Mcnemar's Test P-Value : 9.022e-11

            Sensitivity : 1.0
            Specificity : 0.0
         Pos Pred Value : 0.6
         Neg Pred Value : NaN
             Prevalence : 0.6
         Detection Rate : 0.6
   Detection Prevalence : 1.0
      Balanced Accuracy : 0.5

       'Positive' Class : B
```

Linear Kernel Radial Kernel Polynomial

Fine tuning was conducted to determine the kernel that would provide the best accuracy and the model accuracy was obtained from the fine-tuned model. The fine-tuned model showed that radial kernel produced the best accurate model.

The model showed that 66 out of 66 observations that were diagnosed as benign was correctly predicted and 43 out of 44 observations that was diagnosed as benign was correctly predicted. The model accuracy was 99.09 %.

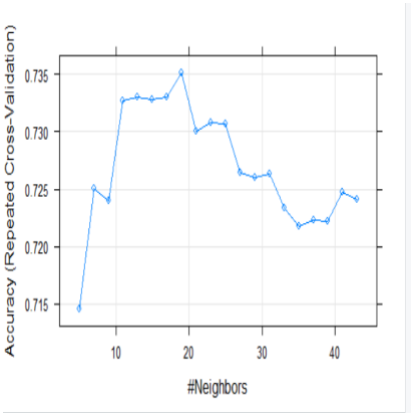
KNN-neighbors model

KNN is a classification algorithm that groups homogenous items. The algorithm segregates unlabeled data into groups defined by k neighbors where k is the number of patterns. The value of K is selected when creating the model. The value of K that is selected plays a major role in determining the results of the algorithm (Cai et al. 2016; Zhang et al. 2013; Yu et al. 2013). A larger value has its advantages which include variance reduction due to the noise data. Conversely, it tends to ignore smaller patterns. Different k-values has been selected for our model, however, the best value of K that was obtained after hyper parametrization was k=19.

Winedata set

At the higher value of k, training scores drops continuously with test scores increasing a bit making it underfit curve and where was lower values of curve was making it viceversa and making the values of overfit. So, the classifier is run thru the gridsearch. After tuning, The training scores and test scores increased with increase of sample set. As we increase the more traning data, the curve might suffer from high variaince a bit.

Fig 2.4: Choosing the best value of K.




```
          Reference
Prediction high low
     high  127  49
     low    36  97

               Accuracy : 0.7249
                 95% CI : (0.6715, 0.774)
    No Information Rate : 0.5275
    P-Value [Acc > NIR] : 9.633e-13

                  Kappa : 0.4456

 Mcnemar's Test P-Value : 0.1931

            Sensitivity : 0.7791
            Specificity : 0.6644
         Pos Pred Value : 0.7216
         Neg Pred Value : 0.7293
             Prevalence : 0.5275
         Detection Rate : 0.4110
   Detection Prevalence : 0.5696
      Balanced Accuracy : 0.7218

       'Positive' Class : high
```

The figure above shows that k=19 produced the highest accuracy. The model was applied to the testing set and the following table was obtained.

The KNN neighbors' classification model produced 72.49 % accuracy on the testing data set. The confusion matrix shows that 127 observation which was of high quality were accurately predicted while 36 observations were wrongly predicted. On the other hand, 97 observations that were of low quality were accurately predicted while 49 observations were wrongly predicted.

Breast Cancer Dataset

The KNN model showed that 7 clusters produced the best accurate model. Therefore, cluster 7 was used to create the model. As seen in the fig, test scores initially was flat and then increased with the number of training data and reaching training score which might overfit as we have more data. The model showed

that 66 out of 66 observations that were diagnosed as benign was correctly predicted and 41 out of 44 observations that was diagnosed as benign was correctly predicted. The model accuracy was 97.27 %. The accuracy level for the training data was 97.6 %.

```
          Reference
Prediction  B  M
         B 66  3
         M  0 41

               Accuracy : 0.9727
                 95% CI : (0.9224, 0.9943)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9425

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 1.0000
            Specificity : 0.9318
         Pos Pred Value : 0.9565
         Neg Pred Value : 1.0000
             Prevalence : 0.6000
         Detection Rate : 0.6000
   Detection Prevalence : 0.6273
      Balanced Accuracy : 0.9659

       'Positive' Class : B
```

Conclusion:

Method	Time taken for Wine DS	Time Taken for Cancer DS
Decision	21.466	12.45
Neural Network	38.75	15.10
Boosting	48.584	14.12
KNN	13.213	11.14
SVM	118.63	16.19

The table show the time taken for the execution of each of the implementation on each of the data set. KNN is the method which took much lesser time in both the classification datasets. The decision tree showed that the best predictors of the red wine were chlorides and volatile acidity were the best predictors of the quality of red wine. The model accuracy of the decision tree on the testing sets was 70.23 %. The neural network showed that the chlorides and alcohol content was the best predictor of the quality of the red wine. The model accuracy of the decision tree on the testing sets was 76.05 %. The model accuracy of the boosting classification model 80 % on the testing sets was 80.53 %. The model accuracy of the SVM classification model 80 % on the testing sets was 75.08 %. The model accuracy of the KNN-neighbors classification model 80 % on the testing sets was 72.49%. The best model was, therefore, the boosting classification model.

On the second classification to predict the diagnosis, the decision tree produced a model that was 94.45 % accurate. Boosting model produced 94.55 % accurate prediction. Neural network produced 97.27 % accurate prediction, SVM produced 99.09 % accurate prediction while KNN-Neighbor produced 97.27 % accurate prediction. Therefore, the best was SVM.

References:

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., Gertych, A., & San Tan, R. (2017). A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, *89*, 389-396.

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(6).

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., ... & Di Carlo, P. (2017). Recursive neural network model for analysis and forecast of PM10 and PM2. 5. *Atmospheric Pollution Research*, *8*(4), 652-659.

Cai, P., Wang, Y., Lu, G., Chen, P., Ding, C., & Sun, J. (2016). A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. *Transportation Research Part C: Emerging Technologies*, *62*, 21-34.

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert systems with applications*, *40*(10), 3970-3983.

Deo, R. C., Wen, X., & Qi, F. (2016). A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Applied Energy*, *168*, 568-593.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659-3667.

Hassan, A. R., & Bhuiyan, M. I. H. (2017). An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting. *Neurocomputing*, *219*, 76-87.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13-19.

James, P. A., Oparil, S., Carter, B. L., Cushman, W. C., Dennison-Himmelfarb, C., Handler, J., ... & Smith, S. C. (2014). 2014 evidence-based guideline for the management of high blood pressure in adults: report from the panel members appointed to the Eighth Joint National Committee (JNC 8). *Jama*, *311*(5), 507-520.

Karimi, F., Sultana, S., Babakan, A. S., & Suthaharan, S. (2019). An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems*, *75*, 61-75.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154).

Kia, M. B., Pirasteh, S., Pradhan, B., Mahmud, A. R., Sulaiman, W. N. A., & Moradi, A. (2012). An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia. *Environmental Earth Sciences*, *67*(1), 251-264.

Li, X., Zhao, L., Wei, L., Yang, M. H., Wu, F., Zhuang, Y., ... & Wang, J. (2016). Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE transactions on image processing*, *25*(8), 3919-3930.

Liguori, L., Albanese, D., Crescitelli, A., Di Matteo, M., & Russo, P. (2019). Impact of dealcoholization on quality properties in red wine at various alcohol content levels. *Journal of Food Science and Technology*, *56*(8), 3707-3720.
Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.
Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena*, *125*, 91-101.

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., & Wei, D. Q. (2017). PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm. *Journal of theoretical biology*, *417*, 1-7.
Yeganeh, B., Motlagh, M. S. P., Rashidi, Y., & Kamalan, H. (2012). Prediction of CO concentrations based on a hybrid Partial Least Square and Support Vector Machine model. *Atmospheric Environment*, *55*, 357-365.
Yu, B., Song, X., Guan, F., Yang, Z., & Yao, B. (2016). k-Nearest neighbor model for multiple-time-step prediction of short-term traffic condition. *Journal of Transportation Engineering*, *142*(6), 04016018.

Zhang, L., Liu, Q., Yang, W., Wei, N., & Dong, D. (2013). An improved k-nearest neighbor model for short-term traffic flow prediction. *Procedia-Social and Behavioral Sciences*, *96*, 653-662.