# Diabetes Prediction Project

HarvardX Data Science Professional Certificate: PH125.9x Capstone 2

*Kanin Limniyakul*

4/4/2022

# Contents



# List of Figures

# List of Tables

## Executive Summary

This is the a part of HarvardX professional certificate in Data Science capstone project. The aim of this project is to find the most accurate classification model on diabetes dataset from all female patients age older than 21 of Pima Indian heritage with 768 rows and 9 columns.

The analysis of this project started from perform exploratory data analysis (EDA) examining the correlation of each features, check near zero value and then center and scale the data as pre-processing process. Next, the data is partitioned into 80% training set and 20% test set. Then, the various models have been introduced including Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis,K-Nearest Neighbor, K-Nearest Neighbor -Cross Validation, Classification Tree, Random Forest and Ensemble Model.

The analysis is included in each models, hyper-parameters tuning (where applicable) as well as comparing between training and test set. The accuracy on the test set ranging between 0.7 - 0.75 where the Random Forest model yield the best accuracy of 0.75 in prediction.

## Exploratory Data Analysis

The dataset comprises of 768 rows with 8 feature columns and 1 outcome column. The description on each features are defined as below,

```
## spec_tbl_df [768 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Pregnancies             : num [1:768] 6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : num [1:768] 148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : num [1:768] 72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : num [1:768] 35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : num [1:768] 0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num [1:768] 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num [1:768] 0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : num [1:768] 50 31 32 21 33 30 26 29 53 54 ...
##  $ Outcome                 : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 1 2 2 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Pregnancies = col_double(),
##   ..   Glucose = col_double(),
##   ..   BloodPressure = col_double(),
##   ..   SkinThickness = col_double(),
##   ..   Insulin = col_double(),
##   ..   BMI = col_double(),
##   ..   DiabetesPedigreeFunction = col_double(),
##   ..   Age = col_double(),
##   ..   Outcome = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Table 1: Features and Outcomes Descriptions

| Name | Description |
|---|---|
| Pregnancies | No of times pregnant |
| Glucose | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| BloodPressure | Diastolic blood pressure (mm Hg) |
| SkinThickness | Triceps skin fold thickness (mm) |
| Insulin | 2-Hour serum insulin (mu U/ml) |
| BMI | Body mass index (weight in kg/(height in m)^2) |
| DiabetesPedigreeFunction | Diabetes pedigree function |
| Age | Age (years) |
| Oucome | Factor class (0 or 1), 0= no diabetes, 1 = diabetes |

Table 2: The first 5 rows of the dataset

| Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

There are 268 diabetes cases and 500 non diabetes cases.Thus the data is not considered as unbalanced data set.

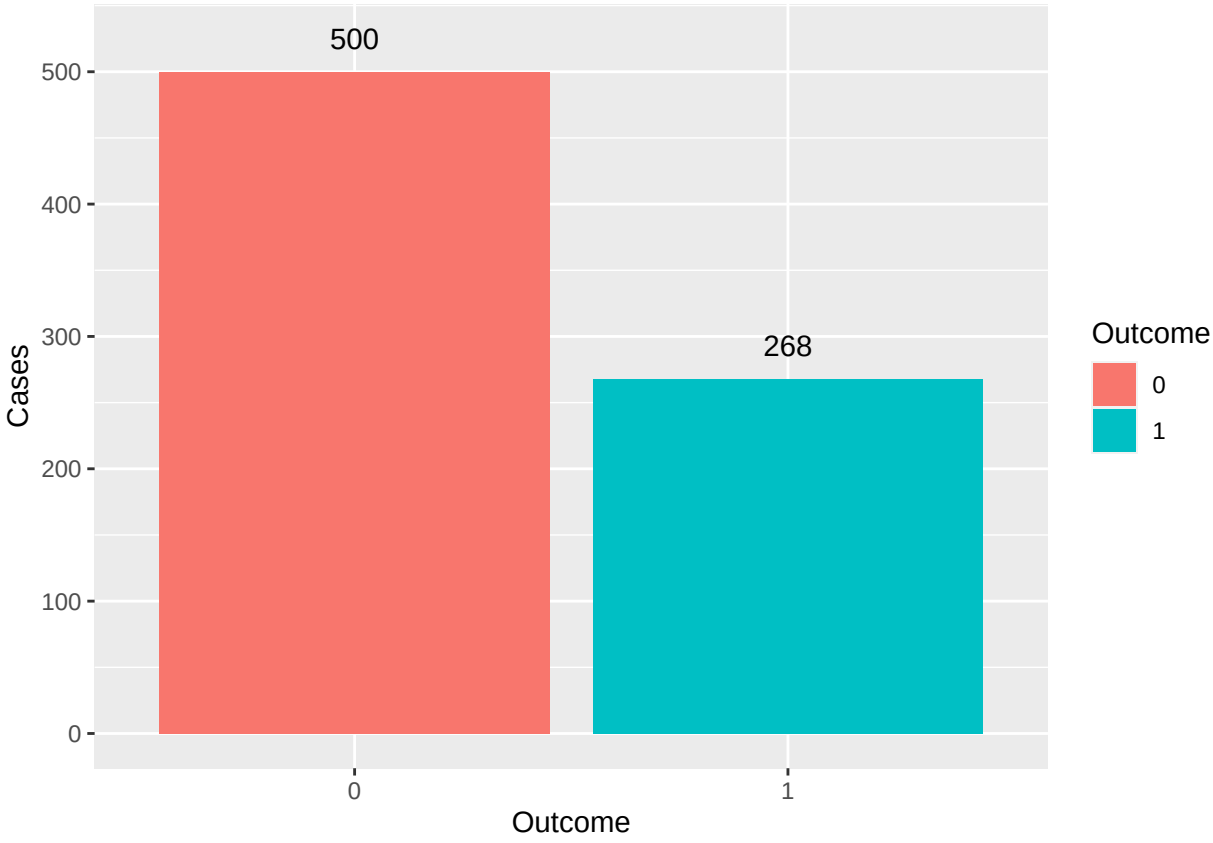



Figure 1: Outcome Summary

## Heatmap Plot

The heatmap below is generated to quickly to show the correlations between all features and the outcome. From the map it's quite clear that Glucose, Pregnancies and BMI are quite positively correlated to Outcome.

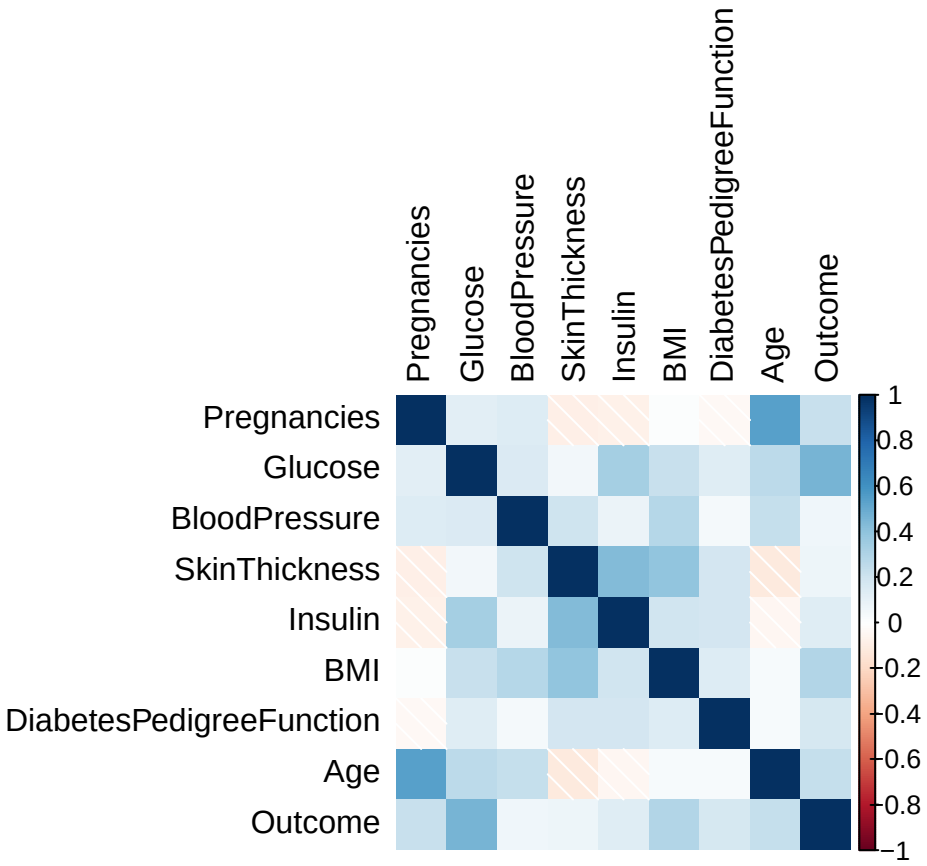



Figure 2: Heatmap showing correlation between features

## Features Plot

The box plots (the feature plots) below show the overview of features distribution according to the Outcome. For diabetes cases, the median of BMI DiabetesPedigreeFunction, Age, Pregnancies, Glucose Blood Pressure and SkinThickness are higher than no diabetes cases. The range of each features are quite different, for example, Insulin range from roughly o to 800 whereas Pregnancies range is only 0 - 15.

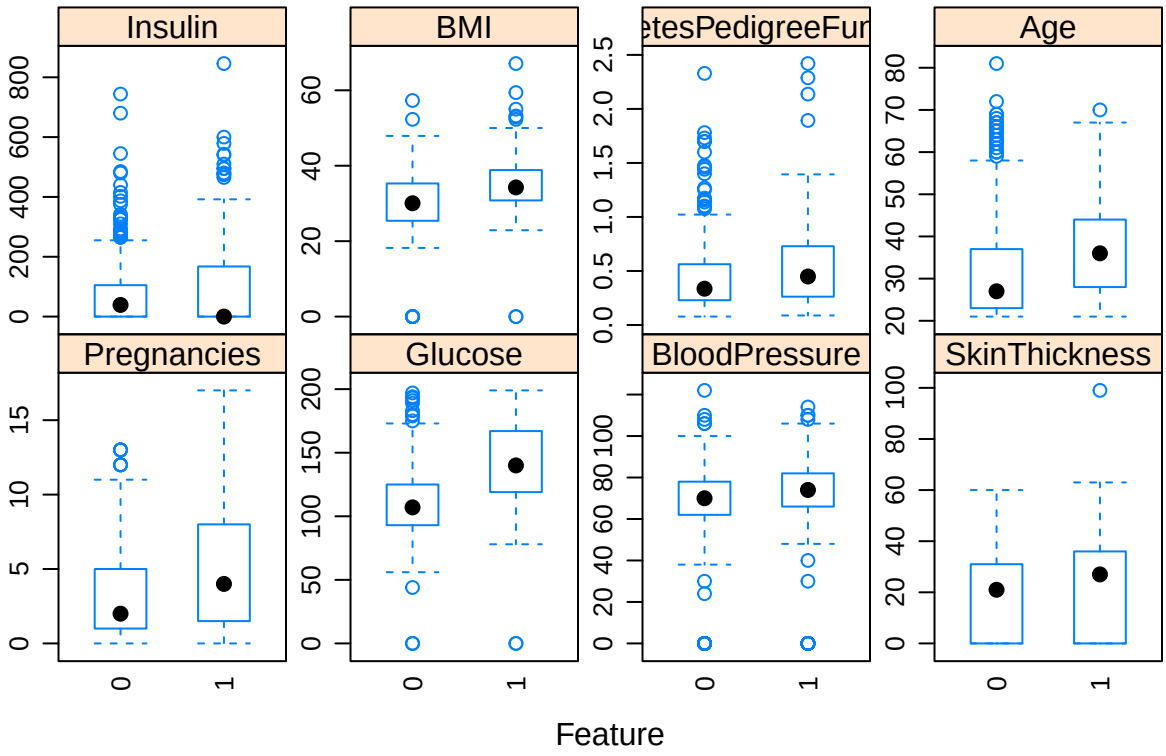



Figure 3: Features box plots

### Near Zero Variance Features

The next thing to explore is to check zero and near zero variance which in this data, there is no zero and near zero value as shown below,

Table 3: Checking Near Zero Variance Predictors

| | freqRatio | percentUnique | zeroVar | nzv |
|---|---|---|---|---|
| Pregnancies | 1.216216 | 2.213542 | FALSE | FALSE |
| Glucose | 1.000000 | 17.708333 | FALSE | FALSE |
| BloodPressure | 1.096154 | 6.119792 | FALSE | FALSE |
| SkinThickness | 7.322581 | 6.640625 | FALSE | FALSE |
| Insulin | 34.000000 | 24.218750 | FALSE | FALSE |
| BMI | 1.083333 | 32.291667 | FALSE | FALSE |
| DiabetesPedigreeFunction | 1.000000 | 67.317708 | FALSE | FALSE |
| Age | 1.142857 | 6.770833 | FALSE | FALSE |

# Pre-Processing data

### Scaling Data

As observed in the previous EDA section, there are a lot of data range variances, so we center the data to be at 0 and scale the data to have the standard deviation as 1 (mean = 0, sd = 1) to improve efficiency and the accuracy of the models.

Table 4: The first six rows of scaled dataset

| Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|
| 0.6395305 | 0.8477713 | 0.1495433 | 0.9066791 | -0.6924393 | 0.2038799 | 0.4681869 | 1.4250667 | 1 |
| -0.8443348 | -1.1226647 | -0.1604412 | 0.5305558 | -0.6924393 | -0.6839762 | -0.3648230 | -0.1905477 | 0 |
| 1.2330766 | 1.9424580 | -0.2637694 | -1.2873733 | -0.6924393 | -1.1025370 | 0.6040037 | -0.1055154 | 1 |
| -0.8443348 | -0.9975577 | -0.1604412 | 0.1544326 | 0.1232213 | -0.4937213 | -0.9201630 | -1.0408711 | 0 |
| -1.1411079 | 0.5037269 | -1.5037073 | 0.9066791 | 0.7653372 | 1.4088275 | 5.4813370 | -0.0204831 | 1 |
| 0.3427574 | -0.1530851 | 0.2528715 | -1.2873733 | -0.6924393 | 0.8108128 | -0.8175458 | -0.2755801 | 0 |

the features density plots after centered and scaled as shown as below,

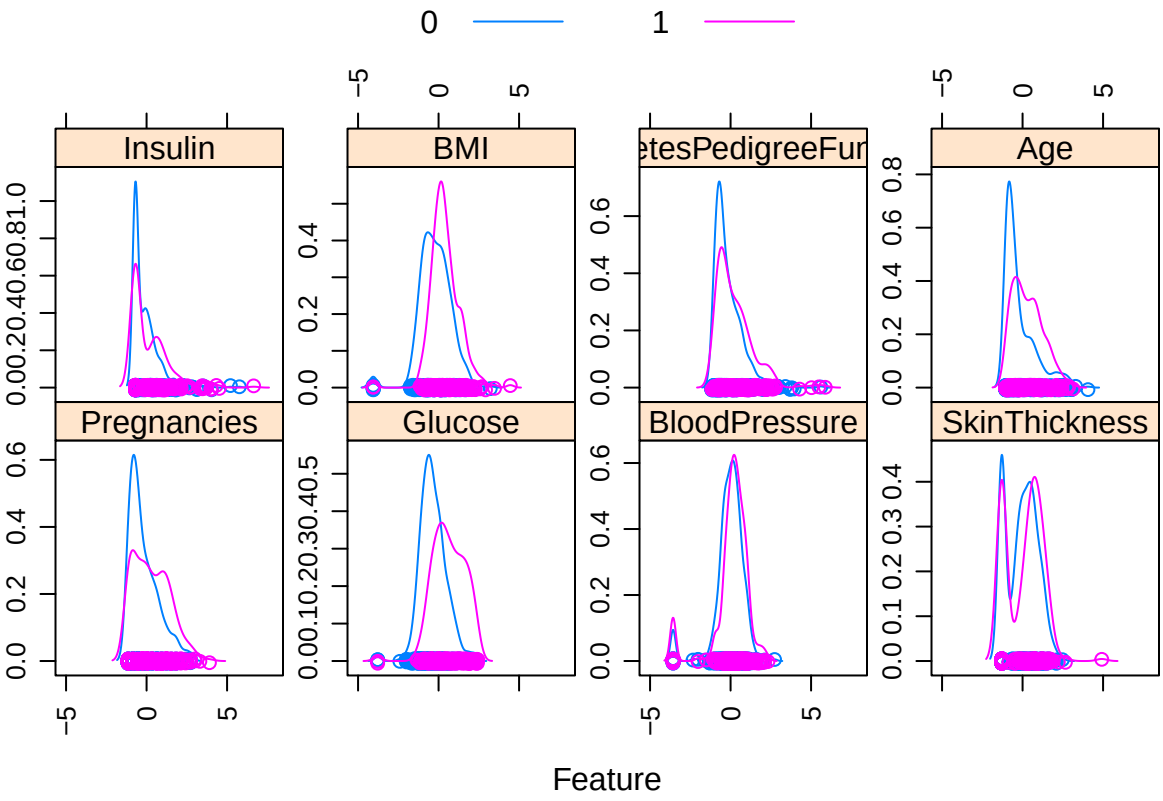



Figure 4: Scaled features density plot

### Data Partitioning

The data set has been splitted to 80% training set and 20% test set. The summary are shown as below.

Table 5: train/test split

| Name | n |
|---|---|
| train | 614 |
| test | 154 |

## Methodology and Analysis

The methodology is to perform various machine learning algorithms including Logistic regression, Linear Discriminant Analysis(LDA), Quadratic Discriminant Analysis(QDA), K-Nearest Neighbor (KNN), K-Nearest Neighbor -Cross Validation (KNN-CV), Classification Tree and Random Forest to see the results of each model then the ensemble model will be conducted. Finally we will choose the best model to predict our data set. The hyper parameters tuning are also conducted on KNN, Classification Tree and Random-forest Model.

### Logistic Regression

Logistic Regression is the first model to train in this data set by using all features as the predictors. Then the insignificant features, namely SkinThickness, Insulin and Age, are removed. The training accuracy is improved from 0.7673724 to 0.7692762. However the test accuracy remains the same at 0.7467532. Note that The training accuracy and test set accuracy are quite close -indicating good fit.

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6249  -0.7385  -0.4246   0.7331   2.9815
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.86274    0.10792  -7.994 1.30e-15 ***
## Pregnancies               0.35824    0.12108   2.959  0.00309 **
## Glucose                   1.15271    0.13194   8.737  < 2e-16 ***
## BloodPressure            -0.31518    0.11933  -2.641  0.00826 **
## SkinThickness            -0.03854    0.12857  -0.300  0.76437
## Insulin                  -0.19366    0.11520  -1.681  0.09276 .
## BMI                       0.70172    0.13527   5.188 2.13e-07 ***
## DiabetesPedigreeFunction  0.34691    0.11313   3.066  0.00217 **
## Age                       0.20922    0.12141   1.723  0.08485 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 793.94  on 613  degrees of freedom
## Residual deviance: 575.86  on 605  degrees of freedom
## AIC: 593.86
##
## Number of Fisher Scoring iterations: 5
```

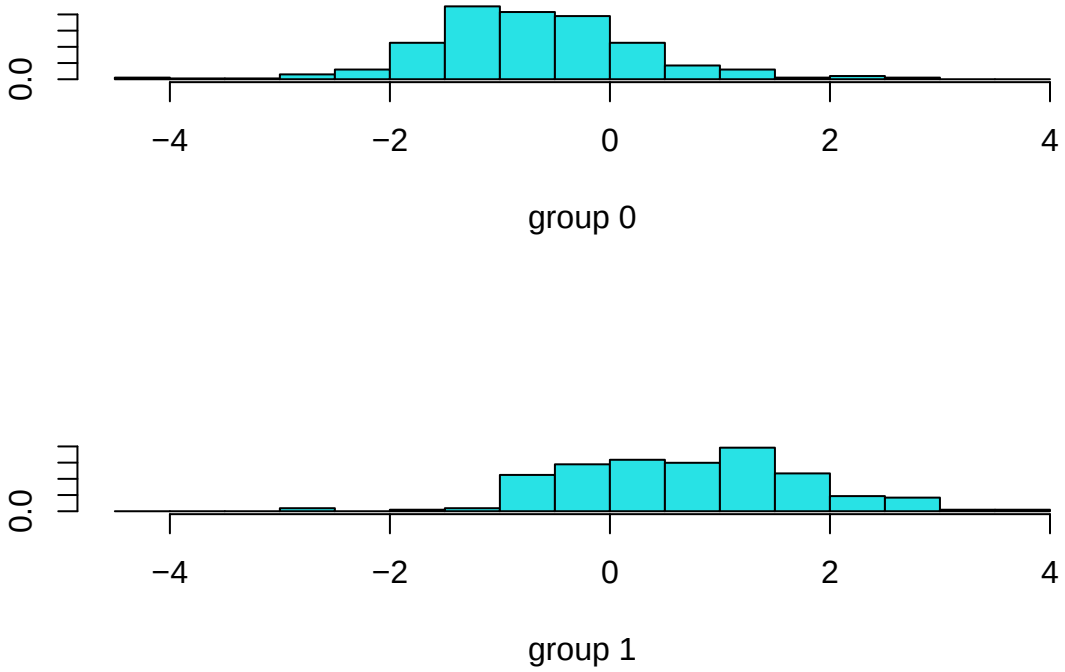
Table 6: Accuracy Table (Logistic Regression)

| Method | Accuracy |
|---|---|
| Training Accuracy | 0.7679269 |
| Training Accuracy-Update | 0.7772782 |
| Test Accuracy | 0.7467532 |
| Test Accuracy-Update | 0.7467532 |

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 86 25
##          1 14 29
##
##                Accuracy : 0.7468
##                  95% CI : (0.6705, 0.8133)
##     No Information Rate : 0.6494
##     P-Value [Acc > NIR] : 0.006192
##
##                   Kappa : 0.4166
##
##  Mcnemar's Test P-Value : 0.109315
##
##             Sensitivity : 0.8600
##             Specificity : 0.5370
##          Pos Pred Value : 0.7748
##          Neg Pred Value : 0.6744
##              Prevalence : 0.6494
##          Detection Rate : 0.5584
##    Detection Prevalence : 0.7208
##       Balanced Accuracy : 0.6985
##
##        'Positive' Class : 0
##
```

## Linear Discriminat Analysis (LDA)

LDA is the second model to train in this data set by using all features as the predictors. The graph below represents the histogram of LDA classifier based on the LDA coefficients. The overlapping between 2 graphs represents that the model could not separate two classes completely- the training accuracy is 0.786645.




The graph below showing the classified test data, class 1 is non-diabetes and class 2 is diabetes with colored by the actual test data. the mixed color(red and black) in each class represents the incorrect classifications.

The test set accuracy is 0.7402597.

Table 7: Accuracy Table (LDA)

| Method | Accuracy |
|---|---|
| Training Accuracy | 0.7866450 |
| Test Accuracy | 0.7402597 |

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 86 26
##          1 14 28
##
##                Accuracy : 0.7403
##                  95% CI : (0.6635, 0.8075)
##     No Information Rate : 0.6494
##     P-Value [Acc > NIR] : 0.01009
##
##                   Kappa : 0.3989
##
##  Mcnemar's Test P-Value : 0.08199
```

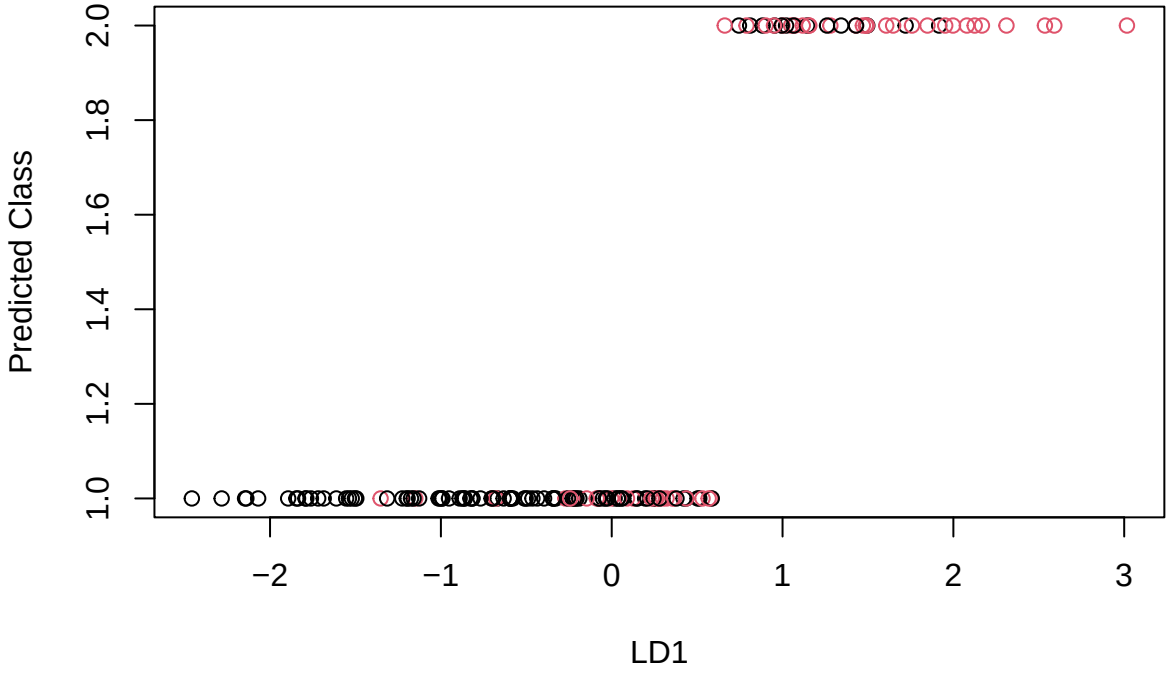



Figure 5: LDA Classification on the test set - colored by the actual outcome from the tes set

```
##
##             Sensitivity : 0.8600
##             Specificity : 0.5185
##          Pos Pred Value : 0.7679
##          Neg Pred Value : 0.6667
##              Prevalence : 0.6494
##          Detection Rate : 0.5584
##    Detection Prevalence : 0.7273
##       Balanced Accuracy : 0.6893
##
##        'Positive' Class : 0
##
```

## Quadratic Discriminat Analysis (QDA)

QDA is the third model to train in this data set by using all features as the predictors. From the plot showing QDA classification results based on the posterior probability colored by the test data. Observing that the colors are more mixed than LDA plot which indicates poorer accuracy (0.7012987) that LDA method.

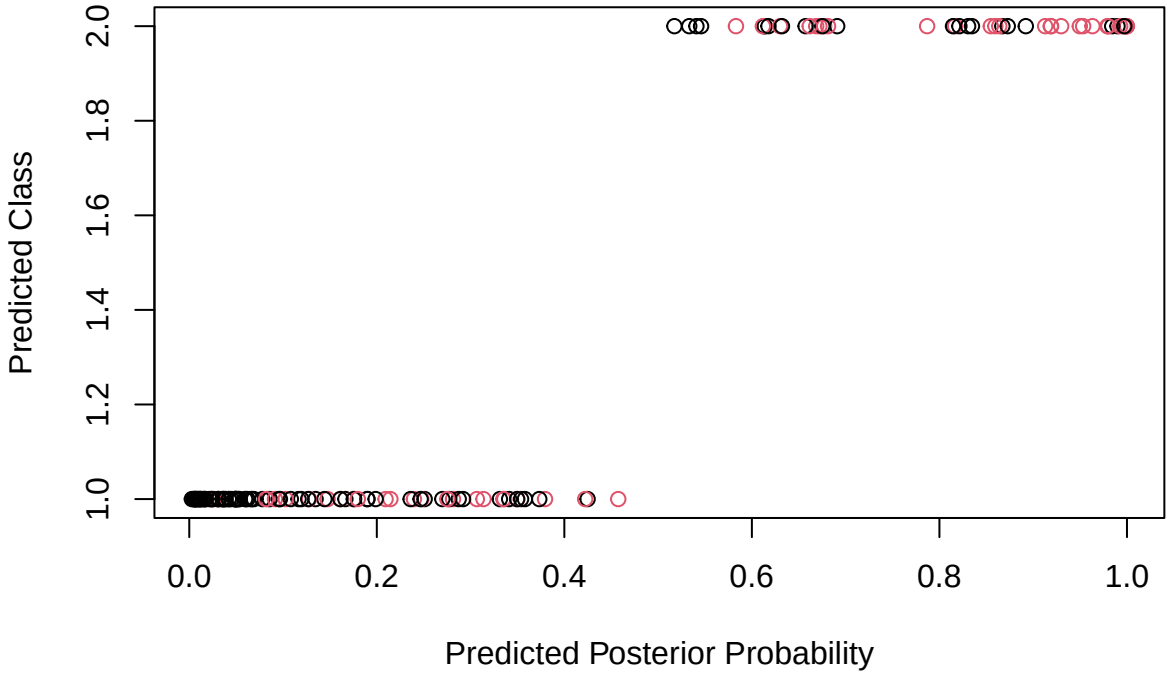



Figure 6: QDA Classification on the test set - colored by the actual outcome from the tes set

Table 8: Accuracy Table (QDA)

| Method | Accuracy |
|---|---|
| Training Accuracy | 0.7768730 |
| Test Accuracy | 0.7012987 |

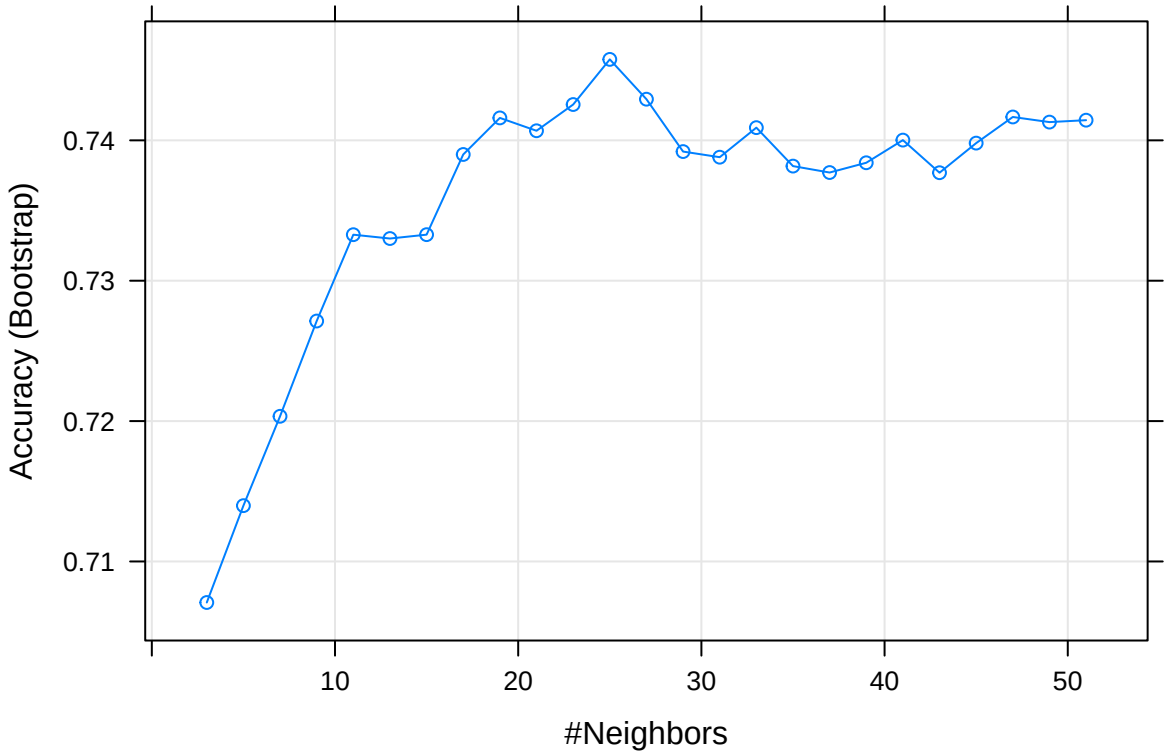
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 79 25
##          1 21 29
##
##                Accuracy : 0.7013
##                  95% CI : (0.6224, 0.7723)
##     No Information Rate : 0.6494
##     P-Value [Acc > NIR] : 0.1016
##
##                   Kappa : 0.3327
##
##  Mcnemar's Test P-Value : 0.6583
##
##             Sensitivity : 0.7900
##             Specificity : 0.5370
##          Pos Pred Value : 0.7596
##          Neg Pred Value : 0.5800
##              Prevalence : 0.6494
##          Detection Rate : 0.5130
##    Detection Prevalence : 0.6753
##       Balanced Accuracy : 0.6635
##
##        'Positive' Class : 0
##
```

## K-Nearest Neighbor Model(KNN)

Firstly, we estimated the best k parameters for model tuning by vary k from 3 to 51 with the incremental of 2.

Table 9: k best tune

| | k |
|---|---|
| 12 | 25 |




KNN model's training and test accuracy can be find as below table.

Table 10: Accuracy Table Comparison (KNN)

| Method | Accuracy |
|---|---|
| Training Accuracy | 0.7457690 |
| Test Accuracy | 0.7207792 |

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 90 33
##          1 10 21
##
##                Accuracy : 0.7208
##                  95% CI : (0.6429, 0.79)
##     No Information Rate : 0.6494
##     P-Value [Acc > NIR] : 0.0363744
##
##                   Kappa : 0.3203
##
##  Mcnemar's Test P-Value : 0.0007937
##
##             Sensitivity : 0.9000
##             Specificity : 0.3889
```

```
##          Pos Pred Value : 0.7317
##          Neg Pred Value : 0.6774
##              Prevalence : 0.6494
##          Detection Rate : 0.5844
##    Detection Prevalence : 0.7987
##       Balanced Accuracy : 0.6444
##
##        'Positive' Class : 0
##
```

### Cross-validated KNN 10-fold, 10%

We will use the cross-validation technique with KNN by setting 10-fold with 10% of the data set, then finding the best k, the accuracy summary is ahown as below table.

Table 11: k_cv best tune

| | k |
|---|---|
| 12 | 25 |

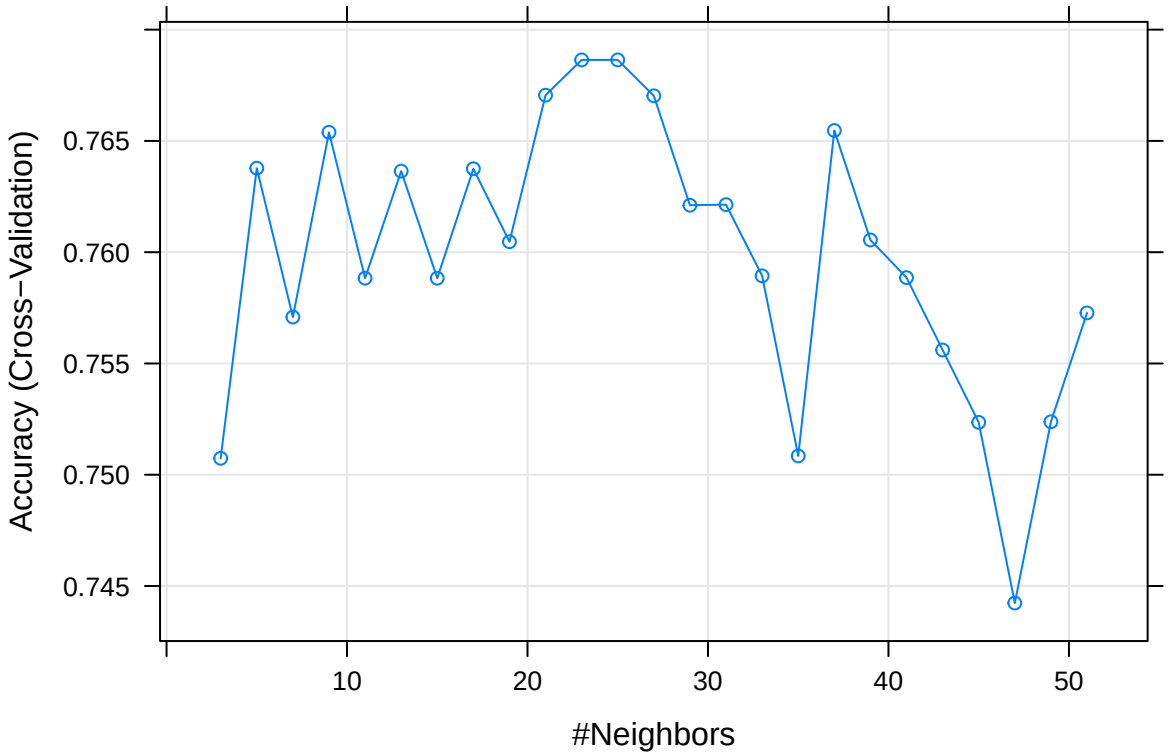


Table 12: Accuracy Table Comparison (KNN_Cross Validation)

| Method | Accuracy |
|---|---|
| Training Accuracy | 0.7686409 |
| Test Accuracy | 0.7207792 |

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 90 33
##          1 10 21
##
##                Accuracy : 0.7208
##                  95% CI : (0.6429, 0.79)
##     No Information Rate : 0.6494
##     P-Value [Acc > NIR] : 0.0363744
##
##                   Kappa : 0.3203
##
##  Mcnemar's Test P-Value : 0.0007937
##
##             Sensitivity : 0.9000
##             Specificity : 0.3889
##          Pos Pred Value : 0.7317
##          Neg Pred Value : 0.6774
##              Prevalence : 0.6494
##          Detection Rate : 0.5844
##    Detection Prevalence : 0.7987
##       Balanced Accuracy : 0.6444
##
##        'Positive' Class : 0
##
```

## Classification Tree Model

Tho model is trained with the complexity parameter(cp) from 0 to 0.06 with an incremental of 0.002, the best cp of 0.022. From the tree model.

Table 13: Best complexity parameter tuning

| | cp |
|---|---|
| 12 | 0.022 |

There are two nodes on glucose level and BMI that classify the outcomes. The accuracy on the test set is 0.7337662.

```
## [1] 0.7337662
```

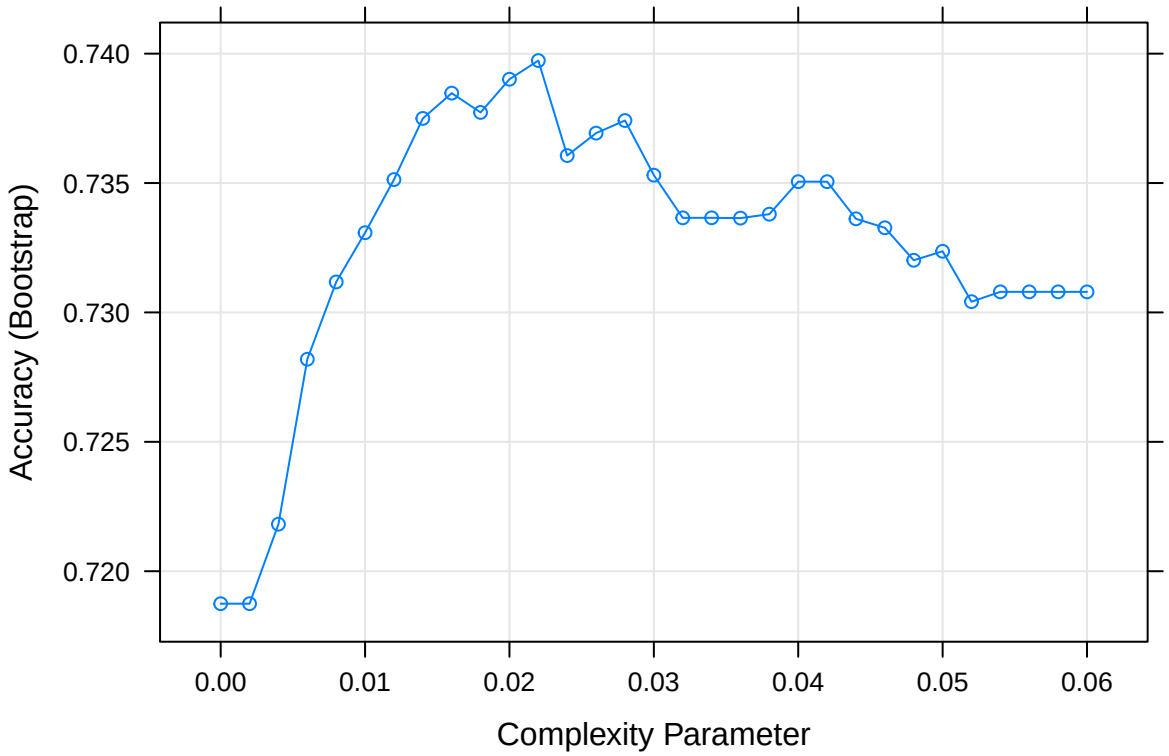



Figure 7: Complexity Parameter Tuning


Figure 8: Classification Tree

Table 14: Accuracy Table Comparison (Classification Tree)

| Method | Accuracy |
|---|---|
| Training Accuracy | 0.7397325 |
| Test Accuracy | 0.7337662 |

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 80 21
##          1 20 33
##
##                Accuracy : 0.7338
##                  95% CI : (0.6566, 0.8017)
##     No Information Rate : 0.6494
##     P-Value [Acc > NIR] : 0.01593
##
##                   Kappa : 0.4129
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.8000
##             Specificity : 0.6111
##          Pos Pred Value : 0.7921
##          Neg Pred Value : 0.6226
##              Prevalence : 0.6494
##          Detection Rate : 0.5195
##    Detection Prevalence : 0.6558
##       Balanced Accuracy : 0.7056
##
##        'Positive' Class : 0
##
```

## Random Forest

Random Forest is tuned by 500 number of trees with randomly selected features from 1 to 8.

Table 15: Best mtry RF model

| | mtry |
|---|---|
| 2 | 2 |

Glucose and BMI are the most important predictors as shown in below figure.

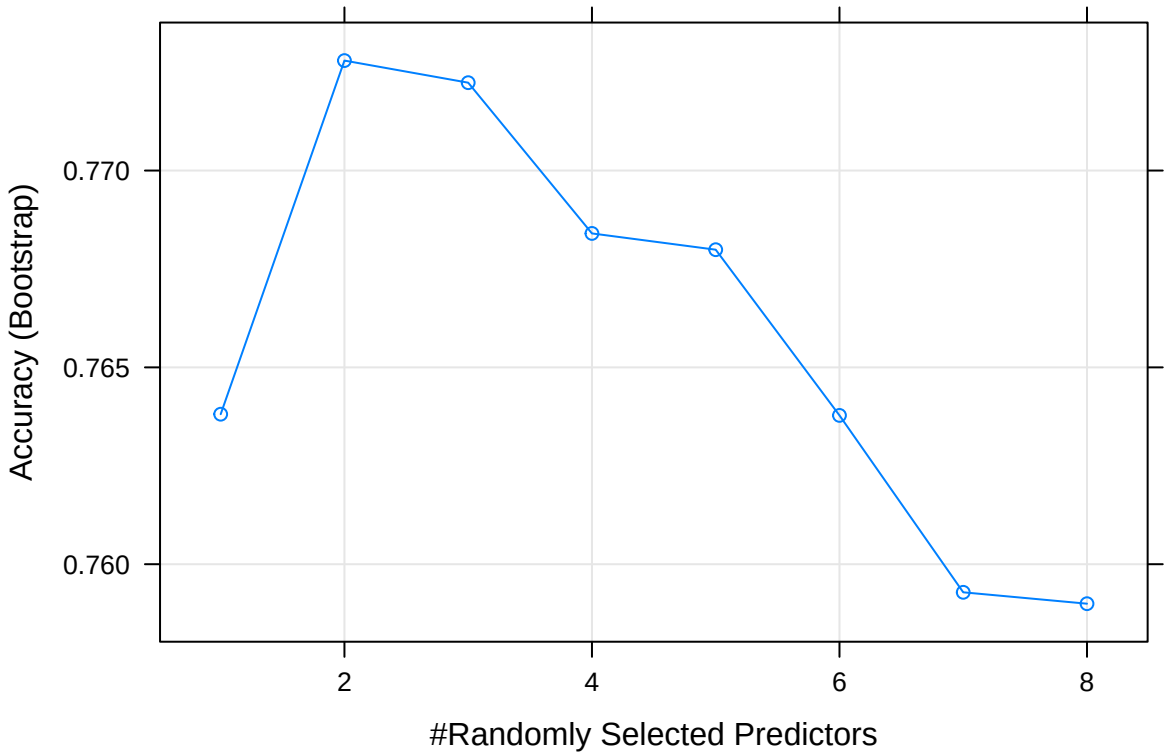
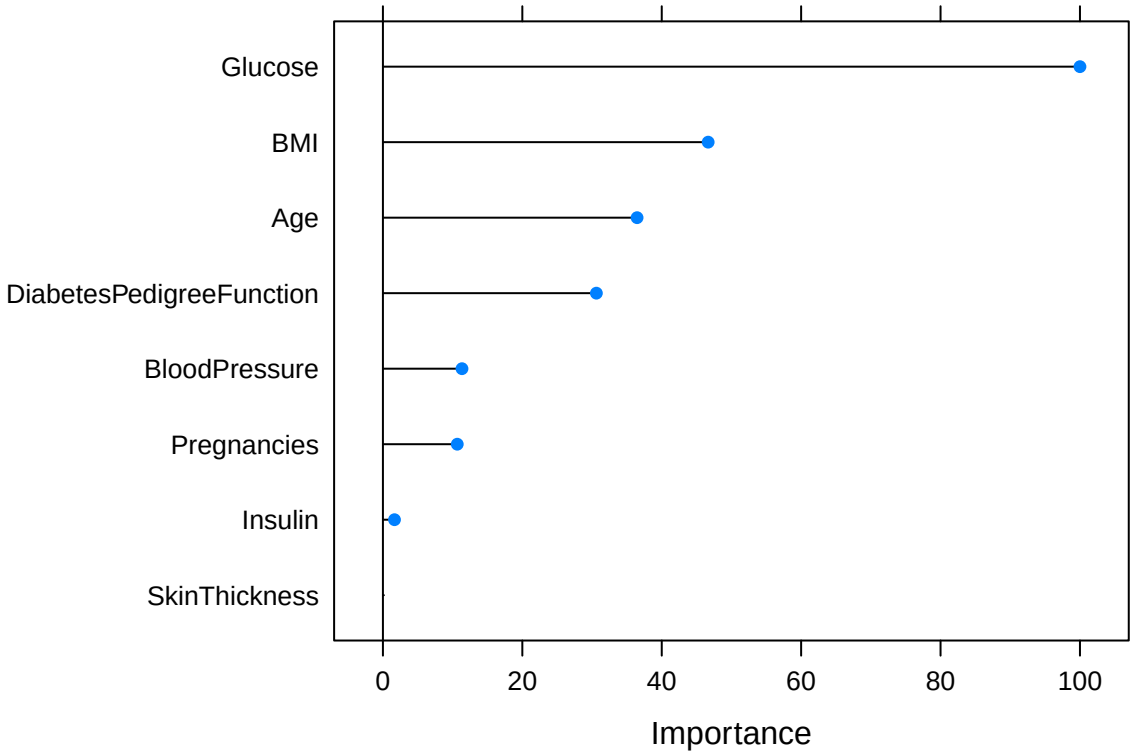



Figure 9: Tuning Parameter k




The accuracy on the test set is 0.7532468.

Table 16: Accuracy Table Comparison (Random Forest)

| Method | Accuracy |
|---|---|
| Training Accuracy | 0.7727935 |
| Test Accuracy | 0.7532468 |

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 88 26
##          1 12 28
##
##                Accuracy : 0.7532
##                  95% CI : (0.6774, 0.8191)
##     No Information Rate : 0.6494
##     P-Value [Acc > NIR] : 0.003683
##
##                   Kappa : 0.4238
##
##  Mcnemar's Test P-Value : 0.034955
##
##             Sensitivity : 0.8800
##             Specificity : 0.5185
```

```
##          Pos Pred Value : 0.7719
##          Neg Pred Value : 0.7000
##              Prevalence : 0.6494
##          Detection Rate : 0.5714
##    Detection Prevalence : 0.7403
##       Balanced Accuracy : 0.6993
##
##        'Positive' Class : 0
##
```

## Ensemble Model

The last step is to ensemble the models by majority vote (>50%) of each outcome.Since each model yields very close results (the wrong predictions happened almost all models) so the accuracy of the ensemble model is 0.7402597, not much improved from the best model. Random Forest is the most accurate model with 0.7532468 accuracy.

Table 17: Ensemble Model

| logistic | KNN | KNN_CV | Classification_Tree | Random_Forest | LDA | QDA | ensemble_prediction | test_outcome |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

Table 18: Accuracy Table Comparison (All Models)

| Model | Accuracy |
|---|---|
| Logistic Regression | 0.7467532 |
| LDA | 0.7402597 |
| QDA | 0.7012987 |
| KNN | 0.7207792 |
| KNN_CV | 0.7207792 |
| Classification Tree | 0.7337662 |
| Random Forest | 0.7532468 |
| Ensemble | 0.7402597 |

# Conclusion

From the implemented models, the accuracy (diabetes prediction) can be arranged from high to low as QDA, KNN, KNN_CV, Classification Tree, LDA, Ensemble, Logistic Regression and Random Forest with range 0.7012987 to 0.7532468. Since each model predicts quite in the same way, hence the ensemble model improved for every models except random forest which yield the highest accuracy of 0.75.

There are many more machine learning algorithms to be explored for future works that can improve the accuracy of the prediction.

# Reference

Dataset Source : https://www.kaggle.com/mathchi/diabetes-data-set