Modeling and Predicting the Occurrence of Diabetes using Machine Learning Algorithm for Classification

2024-05-06

## Results

### Load the Required Libraries

### Load the Data

gender age hypertension heart\_disease smoking\_history bmi HbA1c\_level  
1 Female 80 0 1 never 25.19 6.6  
2 Female 54 0 0 No Info 27.32 6.6  
3 Male 28 0 0 never 27.32 5.7  
4 Female 36 0 0 current 23.45 5.0  
5 Male 76 1 1 current 20.14 4.8  
 blood\_glucose\_level diabetes  
1 140 0  
2 80 0  
3 158 0  
4 155 0  
5 155 0

### Summary Statistics

|  | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 1 | 100000 | 41.89 | 22.52 | 43.00 | 42.00 | 26.69 | 0.08 | 80.00 | 79.92 | -0.05 | -1.00 | 0.07 |
| hypertension | 2 | 100000 | 0.07 | 0.26 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 3.23 | 8.44 | 0.00 |
| heart\_disease | 3 | 100000 | 0.04 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 4.73 | 20.41 | 0.00 |
| bmi | 4 | 100000 | 27.32 | 6.64 | 27.32 | 26.91 | 4.51 | 10.01 | 95.69 | 85.68 | 1.04 | 3.52 | 0.02 |
| HbA1c\_level | 5 | 100000 | 5.53 | 1.07 | 5.80 | 5.57 | 1.19 | 3.50 | 9.00 | 5.50 | -0.07 | 0.22 | 0.00 |
| blood\_glucose\_level | 6 | 100000 | 138.06 | 40.71 | 140.00 | 134.88 | 28.17 | 80.00 | 300.00 | 220.00 | 0.82 | 1.74 | 0.13 |
| diabetes | 7 | 100000 | 0.09 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 2.98 | 6.86 | 0.00 |

# Model Estimation

## Model One: Classification and Regression Tree (CART) Model

gender age hypertension heart\_disease smoking\_history bmi HbA1c\_level  
1 Female 80 0 1 never 25.19 6.6  
2 Female 54 0 0 No Info 27.32 6.6  
3 Male 28 0 0 never 27.32 5.7  
4 Female 36 0 0 current 23.45 5.0  
5 Male 76 1 1 current 20.14 4.8  
6 Female 20 0 0 never 27.32 6.6  
7 Female 44 0 0 never 19.31 6.5  
8 Female 79 0 0 No Info 23.86 5.7  
9 Male 42 0 0 never 33.64 4.8  
10 Female 32 0 0 never 27.32 5.0  
 blood\_glucose\_level diabetes  
1 140 No  
2 80 No  
3 158 No  
4 155 No  
5 155 No  
6 85 No  
7 200 Yes  
8 85 No  
9 145 No  
10 100 No

### Take a sample of 300 observations for easier code execusion

gender age hypertension heart\_disease smoking\_history bmi HbA1c\_level  
41964 Female 39 0 0 never 27.64 4.8  
15241 Male 77 0 1 not current 26.68 6.5  
33702 Female 57 0 0 never 26.14 6.1  
83023 Female 18 0 0 never 19.79 6.6  
80756 Female 59 0 0 never 33.49 5.8  
85374 Male 80 0 0 never 21.14 4.5  
68158 Female 38 0 0 never 32.84 4.0  
59944 Male 51 0 0 former 26.38 5.7  
68536 Female 14 0 0 No Info 25.45 4.8  
17380 Female 80 0 1 ever 23.29 3.5  
 blood\_glucose\_level diabetes  
41964 160 No  
15241 140 No  
33702 130 No  
83023 126 No  
80756 240 Yes  
85374 140 No  
68158 155 No  
59944 140 No  
68536 80 No  
17380 85 No

### Model Summary

CART   
  
300 samples  
 8 predictor  
 2 classes: 'No', 'Yes'   
  
No pre-processing  
Resampling: Cross-Validated (5 fold, repeated 10 times)   
Summary of sample sizes: 241, 240, 239, 240, 240, 241, ...   
Resampling results across tuning parameters:  
  
 cp Accuracy Kappa   
 0.0000000 0.9487114 0.5134720  
 0.2291667 0.9526788 0.5296865  
 0.4583333 0.9353380 0.2666150  
  
Accuracy was used to select the optimal model using the largest value.  
The final value used for the model was cp = 0.2291667.

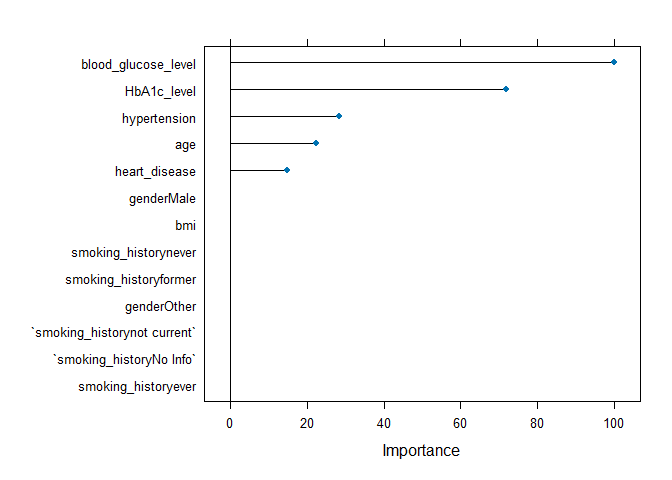
### Confusion Matrix

Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 276 13  
 Yes 0 11  
   
 Accuracy : 0.9567   
 95% CI : (0.927, 0.9767)  
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.0084864   
   
 Kappa : 0.6089   
   
 Mcnemar's Test P-Value : 0.0008741   
   
 Sensitivity : 0.45833   
 Specificity : 1.00000   
 Pos Pred Value : 1.00000   
 Neg Pred Value : 0.95502   
 Prevalence : 0.08000   
 Detection Rate : 0.03667   
 Detection Prevalence : 0.03667   
 Balanced Accuracy : 0.72917   
   
 'Positive' Class : Yes

### Variable Importance

rpart variable importance  
  
 Overall  
blood\_glucose\_level 100.00  
HbA1c\_level 71.98  
hypertension 28.45  
age 22.35  
heart\_disease 14.66  
genderOther 0.00  
`smoking\_historynot current` 0.00  
smoking\_historynever 0.00  
smoking\_historyever 0.00  
genderMale 0.00  
smoking\_historyformer 0.00  
`smoking\_historyNo Info` 0.00  
bmi 0.00

### Plot the Variable Importance



## Model Two: Random Forest

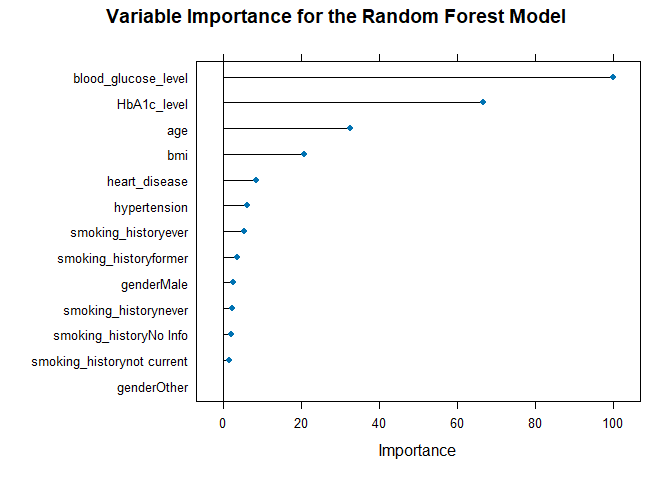
Random Forest   
  
300 samples  
 8 predictor  
 2 classes: 'No', 'Yes'   
  
No pre-processing  
Resampling: Cross-Validated (5 fold, repeated 10 times)   
Summary of sample sizes: 241, 240, 239, 240, 240, 241, ...   
Resampling results across tuning parameters:  
  
 mtry Accuracy Kappa   
 2 0.9470225 0.4504402  
 7 0.9610406 0.6657376  
 13 0.9560069 0.6506096  
  
Accuracy was used to select the optimal model using the largest value.  
The final value used for the model was mtry = 7.

### Confusion Matrix

Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 276 0  
 Yes 0 24  
   
 Accuracy : 1   
 95% CI : (0.9878, 1)   
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.00000000001369  
   
 Kappa : 1   
   
 Mcnemar's Test P-Value : NA   
   
 Sensitivity : 1.00   
 Specificity : 1.00   
 Pos Pred Value : 1.00   
 Neg Pred Value : 1.00   
 Prevalence : 0.08   
 Detection Rate : 0.08   
 Detection Prevalence : 0.08   
 Balanced Accuracy : 1.00   
   
 'Positive' Class : Yes

### Obtain variable importance

rf variable importance  
  
 Overall  
blood\_glucose\_level 100.000  
HbA1c\_level 66.716  
age 32.644  
bmi 20.844  
heart\_disease 8.551  
hypertension 6.125  
smoking\_historyever 5.350  
smoking\_historyformer 3.531  
genderMale 2.564  
smoking\_historynever 2.215  
smoking\_historyNo Info 2.087  
smoking\_historynot current 1.489  
genderOther 0.000



## Model Three: k-Nearest Neighbors

### View the Final Model

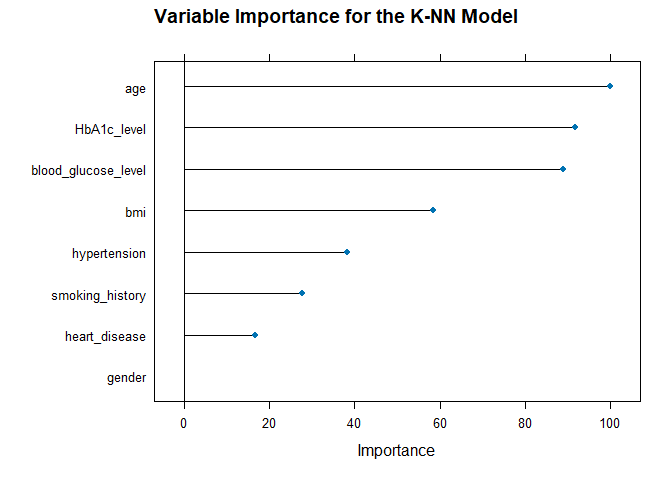
k-Nearest Neighbors   
  
300 samples  
 8 predictor  
 2 classes: 'No', 'Yes'   
  
Pre-processing: centered (13), scaled (13)   
Resampling: Cross-Validated (5 fold, repeated 10 times)   
Summary of sample sizes: 241, 240, 239, 240, 240, 241, ...   
Resampling results across tuning parameters:  
  
 k Accuracy Kappa   
 5 0.9340103 0.2845934  
 7 0.9363440 0.3020008  
 9 0.9326937 0.2466591  
 11 0.9290210 0.1662212  
 13 0.9273542 0.1381262  
  
Accuracy was used to select the optimal model using the largest value.  
The final value used for the model was k = 7.

### Classification Accuracy

Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 275 18  
 Yes 1 6  
   
 Accuracy : 0.9367   
 95% CI : (0.9029, 0.9614)  
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.1694788   
   
 Kappa : 0.3641   
   
 Mcnemar's Test P-Value : 0.0002419   
   
 Sensitivity : 0.25000   
 Specificity : 0.99638   
 Pos Pred Value : 0.85714   
 Neg Pred Value : 0.93857   
 Prevalence : 0.08000   
 Detection Rate : 0.02000   
 Detection Prevalence : 0.02333   
 Balanced Accuracy : 0.62319   
   
 'Positive' Class : Yes

## Variable Importance

ROC curve variable importance  
  
 Importance  
age 100.00  
HbA1c\_level 91.65  
blood\_glucose\_level 88.87  
bmi 58.51  
hypertension 38.30  
smoking\_history 27.73  
heart\_disease 16.65  
gender 0.00



## Model Four: Naive Bayes

### View the Model

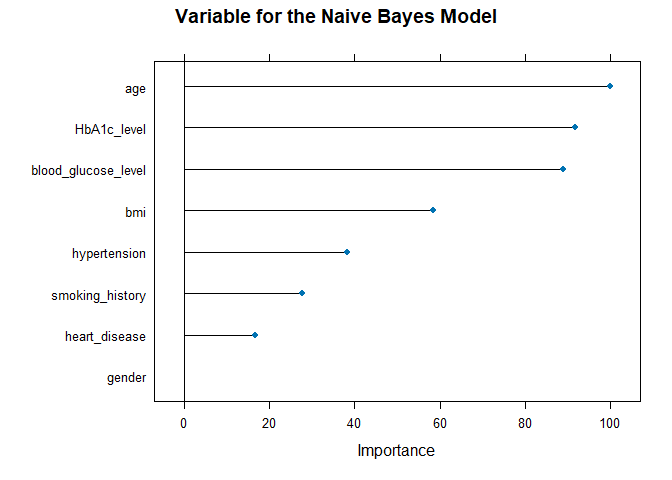
Naive Bayes   
  
300 samples  
 8 predictor  
 2 classes: 'No', 'Yes'   
  
Pre-processing: centered (13), scaled (13)   
Resampling: Cross-Validated (5 fold, repeated 10 times)   
Summary of sample sizes: 241, 240, 239, 240, 240, 241, ...   
Resampling results across tuning parameters:  
  
 usekernel Accuracy Kappa   
 FALSE 0.6518562 0.2228126  
 TRUE 0.9436779 0.4067932  
  
Tuning parameter 'laplace' was held constant at a value of 0  
Tuning  
 parameter 'adjust' was held constant at a value of 1  
Accuracy was used to select the optimal model using the largest value.  
The final values used for the model were laplace = 0, usekernel = TRUE  
 and adjust = 1.

### Prediction and Classification Accuracy

Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 276 16  
 Yes 0 8  
   
 Accuracy : 0.9467   
 95% CI : (0.9148, 0.9692)  
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.0492508   
   
 Kappa : 0.4792   
   
 Mcnemar's Test P-Value : 0.0001768   
   
 Sensitivity : 0.33333   
 Specificity : 1.00000   
 Pos Pred Value : 1.00000   
 Neg Pred Value : 0.94521   
 Prevalence : 0.08000   
 Detection Rate : 0.02667   
 Detection Prevalence : 0.02667   
 Balanced Accuracy : 0.66667   
   
 'Positive' Class : Yes

### Variable Importance

ROC curve variable importance  
  
 Importance  
age 100.00  
HbA1c\_level 91.65  
blood\_glucose\_level 88.87  
bmi 58.51  
hypertension 38.30  
smoking\_history 27.73  
heart\_disease 16.65  
gender 0.00



## Model Five: Support Vector Machine (SVM)

### View the Model

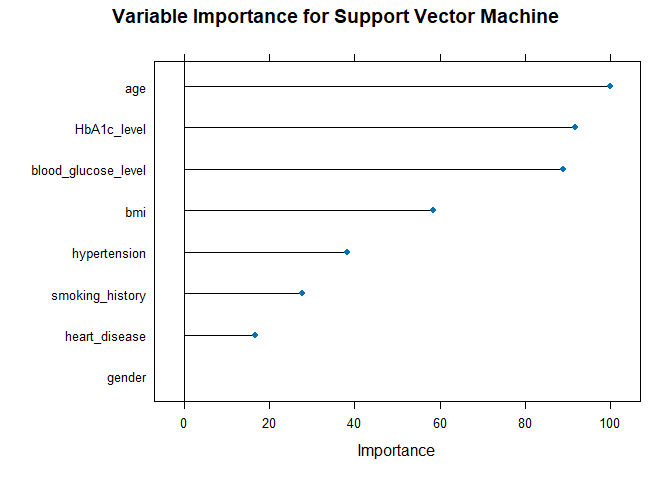
Support Vector Machines with Linear Kernel   
  
300 samples  
 8 predictor  
 2 classes: 'No', 'Yes'   
  
Pre-processing: centered (13), scaled (13)   
Resampling: Cross-Validated (5 fold, repeated 10 times)   
Summary of sample sizes: 241, 240, 239, 240, 240, 241, ...   
Resampling results:  
  
 Accuracy Kappa   
 0.9433946 0.5337302  
  
Tuning parameter 'C' was held constant at a value of 1

### Prediction and Classification Accuracy

Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 273 10  
 Yes 3 14  
   
 Accuracy : 0.9567   
 95% CI : (0.927, 0.9767)  
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.008486   
   
 Kappa : 0.6604   
   
 Mcnemar's Test P-Value : 0.096092   
   
 Sensitivity : 0.58333   
 Specificity : 0.98913   
 Pos Pred Value : 0.82353   
 Neg Pred Value : 0.96466   
 Prevalence : 0.08000   
 Detection Rate : 0.04667   
 Detection Prevalence : 0.05667   
 Balanced Accuracy : 0.78623   
   
 'Positive' Class : Yes

### Variable Importance

ROC curve variable importance  
  
 Importance  
age 100.00  
HbA1c\_level 91.65  
blood\_glucose\_level 88.87  
bmi 58.51  
hypertension 38.30  
smoking\_history 27.73  
heart\_disease 16.65  
gender 0.00



## Reference

Mustafa, M. (2023). Diabetes prediction dataset. Kaggle.com. <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>