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

2024-05-05

Modeling and predicting diabetes

Introduction

Diabetes is among the current chronic condition posing danger globally. Detecting and doing an early intervention is appropriate for for managing the condition. Modeling and predicting the likelihood of one having this chronic condition will be helpful in the medical and healthcare facilities. There exist various approaches for modeling and predicting diabetes condition including but not limited to binary logistic regression analysis. However, in the recent times, machines learning algorithms for classification have proved to be the overall the best approach of modeling and predicting diabetes occurrence cases. In this paper Machine Learning (ML) algorithms for classification are utilized to model and predict the occurrence of this chronic condition based on the characteristic of patients. The secondary data used in this study is the electronic health records obtained from kaggles website (https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset).

Description of the Electronic Data used

Electronic Health Records (EHRs) are the primary source of data for the Diabetes Prediction dataset (Mustafa, 2023). EHRs are digital versions of patient health records that contain information about their medical history, diagnosis, treatment, and outcomes. The data in EHRs is collected and stored by healthcare providers, such as hospitals and clinics, as part of their routine clinical practice (Mustafa, 2023). The variables in this study include gender, age, hypertension, heart disease, smoking history, BMI, HBA1C level, blood glucose level, and the response variable is the occurrence of diabetes.

Objectives

This study is guided by the following objectives * To evaluate the performance of various machine learning algorithms, which include Naive Bayes, k-Nearest Neighbors (kNN), Hierachical clustering and K-Means Clustering

- Assessing the effectiveness of the developed models
- Predicting the occurrence of the diabetes using the best overall model.

Methodology

This study employed the use of secondary data obtained from Kaggle website. The used comprised the demographic information as well as the clinical data of patients. These information include age, gender, BMI, heart disease, blood sugar level, hypertension and diabetes status. The machine learning applied in this study include the following;

- Classification and Regression Tree (CART): CART is a decision tree algorithm that recursively splits the data into subsets based on the value of predictor variables. At each step, it chooses the variable that best splits the data, resulting in a tree-like structure where the leaves represent the predicted outcome.
- Random Forest: Random Forest is an ensemble machine learning method constructed from various decision trees to create one classification and prediction algorithm.
- k-Nearest Neighbors: This algorithm is a non-parametric machine learning algorithm thta classifies an individual based on the k-Nearest neighbors.

- Support Vector Machine (SVM): SVM is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyper-plane that best separates the classes in the feature space. Using this algorithm, three kernels options are always specified, that is Sigmoid, Linear and Polynomial, however, in many instance, linear kernel has always outperformed the sigmoid and polynomial kernel.
- Naive Bayes Classifier: This is based on Bayes' theorem to classify individual, holding the assumption that feature are independent

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

The mean age of the participants was 41.89 years (SD = 22.52), with a range from 0.08 to 80.00 years. On the other hand, prevalence of hypertension was 0.07 (SD = 0.26), while heart disease was reported in 0.04 (SD = 0.19) of the cases. The mean body mass index (BMI) was 27.32 (SD = 6.64), ranging from 10.01 to 95.69. HbA1c levels averaged at 5.53 (SD = 1.07), with values ranging from 3.50 to 9.00. Blood glucose levels had a mean of 138.06 (SD = 40.71), with a wide range from 80.00 to 300.00. The prevalence of diabetes was 0.09 (SD = 0.28), indicating a relatively low frequency in the sample.

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
93185 Female 57.00            0             0           never 33.21         3.5
87812 Female 62.00            0             0         No Info 27.32         6.6
35831   Male 44.00            0             0          former 35.58         4.5
72550 Female 34.00            0             0            ever 27.32         6.5
7832    Male  1.72            0             0         No Info 28.69         6.0
26375 Female 77.00            0             0          former 41.70         6.2
62739   Male 74.00            0             0          former 36.62         6.2
76558   Male 44.00            0             0         No Info 27.32         6.0
66221 Female 18.00            0             0           never 23.81         4.8
91583   Male 38.00            0             0           never 47.17         6.5
      blood_glucose_level diabetes
93185                 140       No
87812                 126      Yes
35831                 145       No
72550                 145       No
7832                  145       No
26375                 155       No
62739                 155       No
76558                  80       No
66221                 160       No
91583                 145       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, 239, 240, 240, 240, 241, ...
Resampling results across tuning parameters:

  cp         Accuracy   Kappa
  0.0000000  0.9600564  0.6025236
  0.2954545  0.9623735  0.6099530
  0.5909091  0.9410647  0.2774304

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.2954545.
```

Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  278   9
       Yes   0  13

               Accuracy : 0.97
                 95% CI : (0.9438, 0.9862)
    No Information Rate : 0.9267
    P-Value [Acc > NIR] : 0.001102

                  Kappa : 0.728

 Mcnemar's Test P-Value : 0.007661

            Sensitivity : 0.59091
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 0.96864
             Prevalence : 0.07333
         Detection Rate : 0.04333
   Detection Prevalence : 0.04333
      Balanced Accuracy : 0.79545

       'Positive' Class : Yes
```

The classification and regression model estimated shows that the model has an accuracy of approximately 94.67%. This shows that the model classifies the respondents correctly into their respective categories as either (0 or 1), 94.67% of the time.

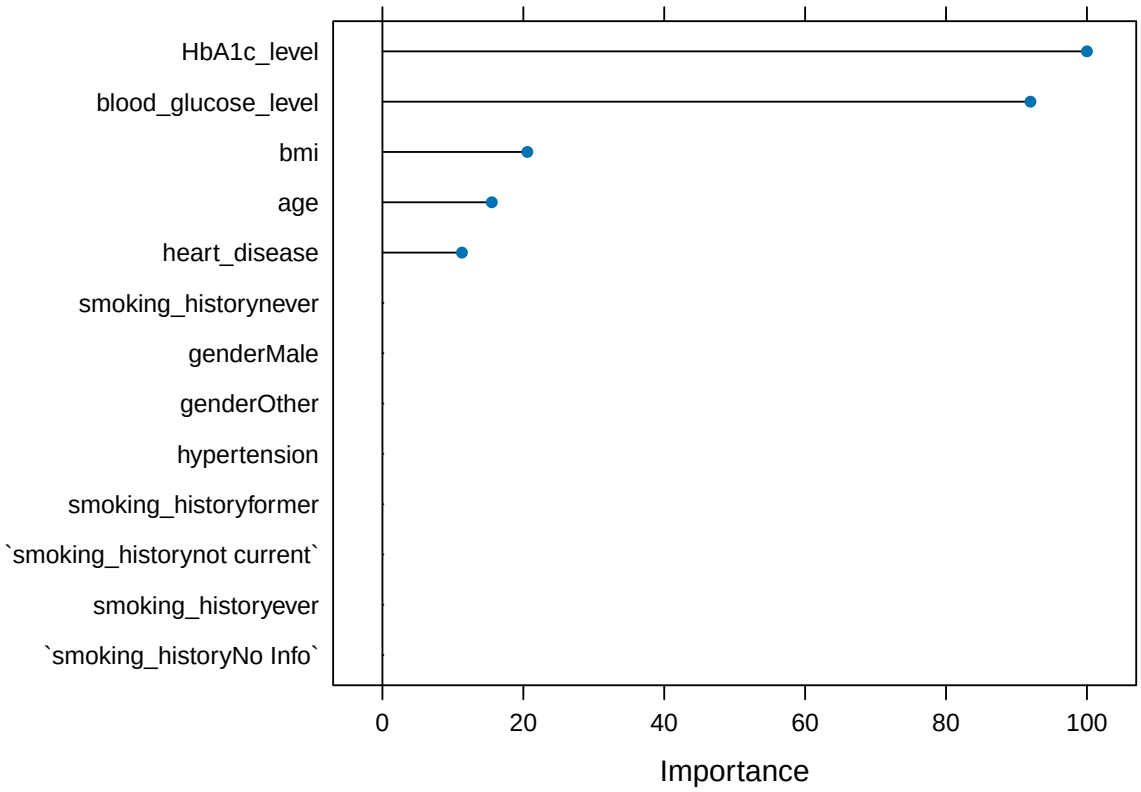
Variable Importance

rpart variable importance

	Overall
HbA1c_level	100.00
blood_glucose_level	91.99
bmi	20.57
age	15.52
heart_disease	11.28
hypertension	0.00
smoking_historynever	0.00
genderOther	0.00
smoking_historyformer	0.00
smoking_historyever	0.00
‘smoking_historyNo Info‘	0.00
‘smoking_historynot current‘	0.00
genderMale	0.00

The results shows the most important and significant variable in the classification and regression trees model developed are HbA1c_level, bmi, blood glucose level, and so on. From the results, age is 100% important to be in our model, followed by HbA1c_level with 91.72%, bmi with 91.58%, blood glucose level with 64.97%, and hypertension with 47.19%. The remaining variable have no significance contribution to be in our model. Consider the plot below.

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: 240, 241, 240, 239, 240, 239, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.9610170  0.5944618
   7    0.9803306  0.8297985
  13    0.9766692  0.8093613

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 7.
```

Confusion Matrix

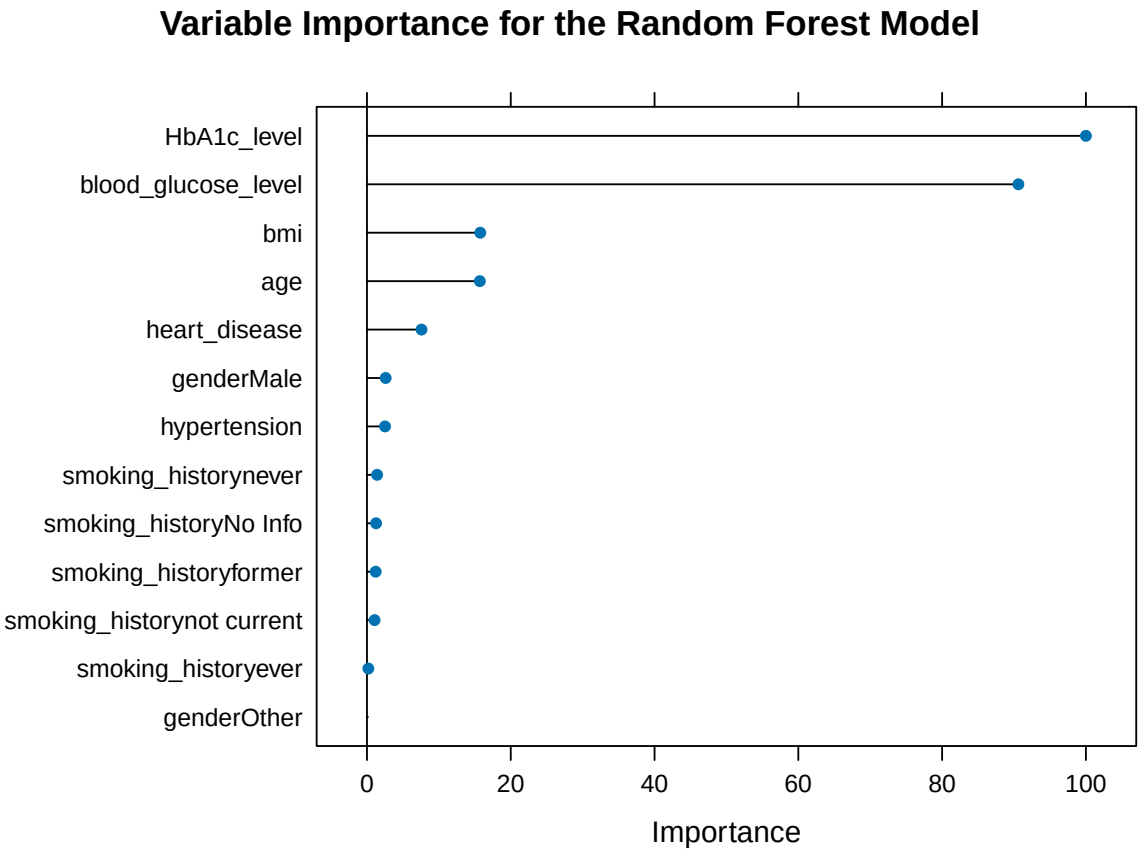
```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  278   0
       Yes   0  22

               Accuracy : 1
                 95% CI : (0.9878, 1)
    No Information Rate : 0.9267
    P-Value [Acc > NIR] : 0.0000000001194

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.00000
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 1.00000
             Prevalence : 0.07333
         Detection Rate : 0.07333
   Detection Prevalence : 0.07333
      Balanced Accuracy : 1.00000

       'Positive' Class : Yes
```

The random forest model developed shows that the model has an accuracy of approximately 100%. This shows that the model classifies the respondents correctly into their respective categories as either (0 or 1), 100% of the time.

Obtain variable importance

```
rf variable importance

                          Overall
HbA1c_level               100.000
blood_glucose_level        90.607
bmi                        15.762
age                        15.699
heart_disease               7.583
genderMale                  2.595
hypertension                2.510
smoking_historynever        1.408
smoking_historyNo Info      1.267
smoking_historyformer       1.210
smoking_historynot current  1.059
smoking_historyever         0.196
genderOther                 0.000
```

This algorithm give slightly different results from what we saw earlier. All the variable in the algorithm have some level of importance being in our model. For instance, HbA1c_level comprise 100% followed by blood glucose level with 89.59%, bmi with 54.95%, age with 32.48% and so on. Consider the plot below to aid in the visualization



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: 240, 240, 240, 239, 241, 239, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.9360738  0.23056807
   7  0.9327179  0.15013649
   9  0.9330399  0.13192531
  11  0.9317120  0.10261947
```

```
  13  0.9293949  0.05582848

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

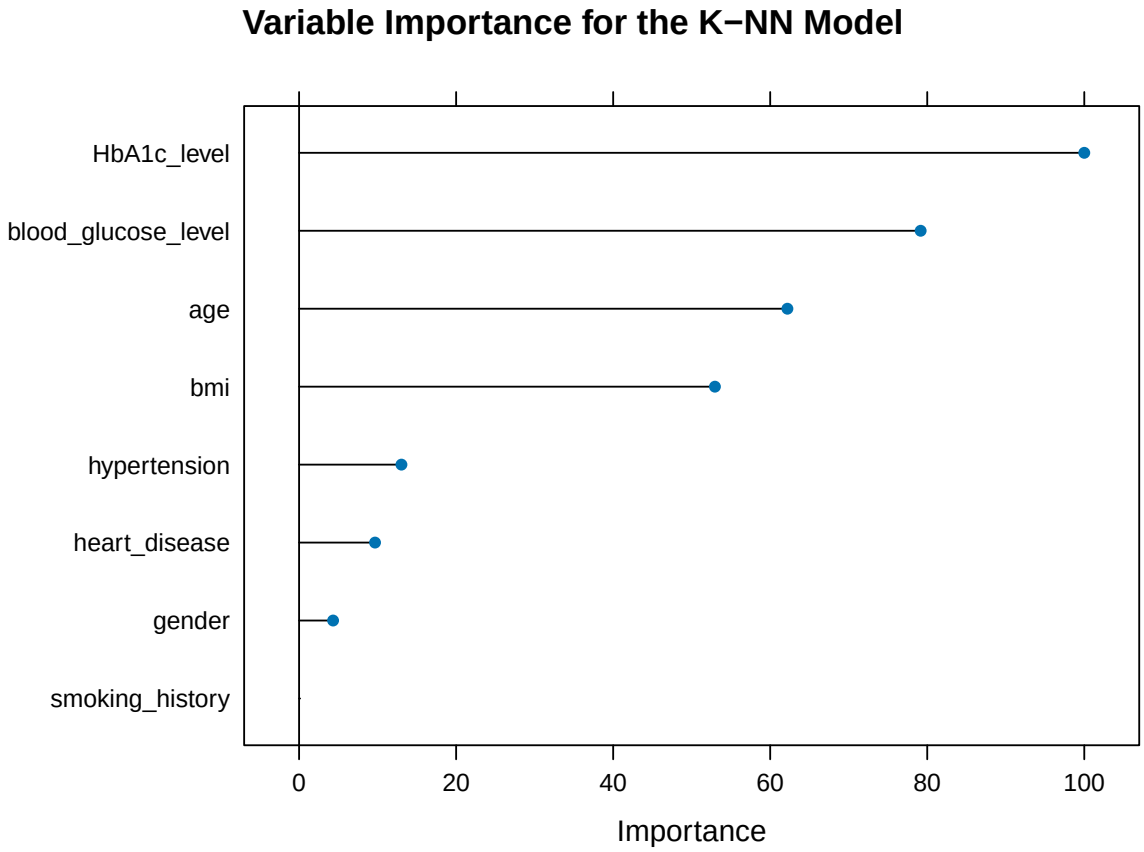
Classification Accuracy

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  278  13
       Yes   0   9

               Accuracy : 0.9567
                 95% CI : (0.927, 0.9767)
    No Information Rate : 0.9267
    P-Value [Acc > NIR] : 0.0236765

                  Kappa : 0.562

 Mcnemar's Test P-Value : 0.0008741

            Sensitivity : 0.40909
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 0.95533
             Prevalence : 0.07333
         Detection Rate : 0.03000
   Detection Prevalence : 0.03000
      Balanced Accuracy : 0.70455

       'Positive' Class : Yes
```

k-Nearest Neighbors performed slightly poor as compared to the classification and regression tree and k-nearest neighbors as well. From the above algorithm, the classification and prediction accuracy is approximately 92.67% implying that the model correctly predict and classify patients in their correct categories 95.33% of the time. The algorithm has a higher mis-classification error than that of random forest and CART model.

```
ROC curve variable importance

                     Importance
HbA1c_level             100.000
blood_glucose_level      79.162
age                      62.192
bmi                      52.959
hypertension             13.032
heart_disease             9.671
gender                    4.329
smoking_history           0.000
```

The percentage contribution of the each variable to the occurrence of the diabetes is as shown above with HbA1c_level having 100%, followed by age with 65.52%, blood glucose level with 57.07% and so. This is alos indicated in the plot below



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: 239, 240, 239, 241, 241, 240, ...
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.8630505  0.4370990
   TRUE      0.9377565  0.2187392

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

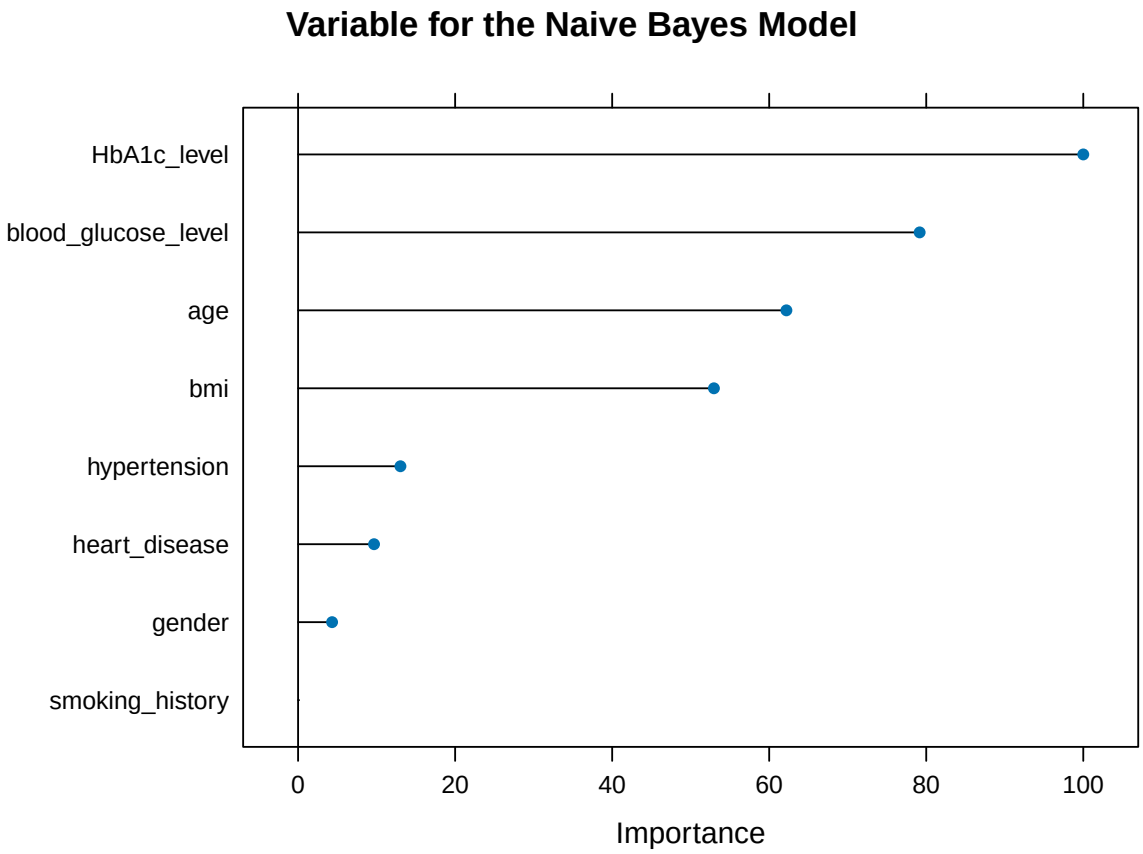
```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  278  16
       Yes   0   6

               Accuracy : 0.9467
                 95% CI : (0.9148, 0.9692)
    No Information Rate : 0.9267
    P-Value [Acc > NIR] : 0.1080120

                  Kappa : 0.41

 Mcnemar's Test P-Value : 0.0001768

            Sensitivity : 0.27273
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 0.94558
             Prevalence : 0.07333
         Detection Rate : 0.02000
   Detection Prevalence : 0.02000
      Balanced Accuracy : 0.63636

       'Positive' Class : Yes
```

Similar to the k-nearest neighbors, naive bayes performs slightly poor in the classification and prediction of the occurrence of diabetes. From the model above,naive bayes correctly predict and classify patients in their respective categories as either having diabetes or not having, 91.00% of the time, which lower as compared to random forest and CART model.

Variable Importance

```
ROC curve variable importance

                    Importance
HbA1c_level            100.000
blood_glucose_level     79.162
age                     62.192
bmi                     52.959
hypertension            13.032
heart_disease            9.671
gender                   4.329
smoking_history          0.000
```

Naive bayes give resuls simialar to that of k-NN. The percentage contribution of the each variable to the occurrence of the diabetes is as shown above with HbA1c_level having 100%, followed by age with 65.52%, blood glucose level with 57.07% and so. These results can be visualized as shown below



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: 239, 240, 241, 240, 240, 241, ...
Resampling results:

  Accuracy   Kappa
  0.9684235  0.7529907

Tuning parameter 'C' was held constant at a value of 1
```

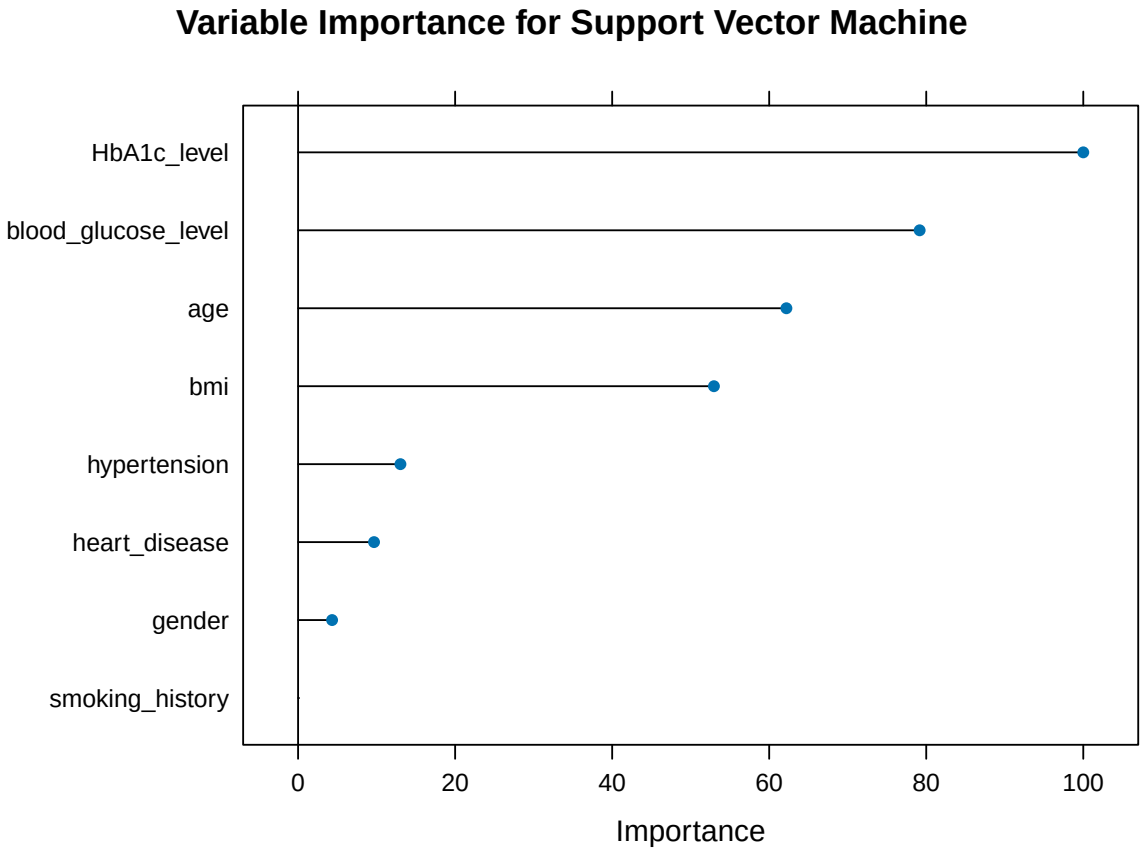
Prediction and Classification Accuracy

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  276   2
       Yes   2  20

               Accuracy : 0.9867
                 95% CI : (0.9662, 0.9964)
    No Information Rate : 0.9267
    P-Value [Acc > NIR] : 0.000001849

                  Kappa : 0.9019

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.90909
            Specificity : 0.99281
         Pos Pred Value : 0.90909
         Neg Pred Value : 0.99281
             Prevalence : 0.07333
         Detection Rate : 0.06667
   Detection Prevalence : 0.07333
      Balanced Accuracy : 0.95095

       'Positive' Class : Yes
```

Support vector machine performed than all the other three models except random, with provides 100% prediction and classification accuracy. The support vector machine model estimated above has a classification and prediction of 97.00%. These results implies that the algorithm classifies patients in the correct categories (0 or 1) 97.00% of the time with only 3% chances of making a mis-classification error.

Variable Importance

```
ROC curve variable importance

                    Importance
HbA1c_level            100.000
blood_glucose_level     79.162
age                     62.192
bmi                     52.959
hypertension            13.032
heart_disease            9.671
gender                   4.329
smoking_history          0.000
```



HbA1c_level has 100% in our model followed by age with 65.52%, blood glucose level with 57.07% and bmi with 39.19% and so on. From the results above, in either model, smoking has significant importance in predicting and occurrence of diabetes.

Compare the various machine learning models

In comparing the five models, Random Forest outperforms the others with the highest accuracy of 100% and perfect agreement (Kappa = 1.000). It also achieves perfect sensitivity and specificity, indicating its exceptional ability to classify both positive and negative instances accurately. Support Vector Machine (SVM) follows closely with an accuracy of 97.00% and the highest Kappa value of 0.8258, suggesting substantial agreement beyond chance. However, SVM's sensitivity is lower compared to Random Forest. k-Nearest Neighbors and Classification and Regression Trees also perform well but have lower sensitivity values compared to Random Forest and SVM. Naïve Bayes exhibits the lowest performance overall with the lowest accuracy, Kappa, and sensitivity values among the models.

Reference

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