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  
21950 Female 17 0 0 No Info 24.18 6.6  
68 Female 13 0 0 No Info 28.16 5.0  
57509 Male 68 0 0 never 34.90 6.2  
99073 Male 53 1 0 former 23.01 4.8  
57856 Male 63 0 1 not current 27.32 6.6  
56013 Female 37 0 0 No Info 24.75 5.0  
692 Female 80 0 0 never 27.32 6.0  
64060 Male 43 0 0 No Info 27.32 3.5  
26729 Male 80 0 0 No Info 27.32 4.5  
1619 Female 47 0 0 current 19.97 4.8  
 blood\_glucose\_level diabetes  
21950 85 No  
68 90 No  
57509 85 No  
99073 130 No  
57856 200 No  
56013 145 No  
692 200 Yes  
64060 100 No  
26729 140 No  
1619 158 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: 240, 240, 240, 240, 240, 240, ...   
Resampling results across tuning parameters:  
  
 cp Accuracy Kappa   
 0.0000000 0.9437069 0.4552880  
 0.2291667 0.9477018 0.4697721  
 0.4583333 0.9310280 0.2035499  
  
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

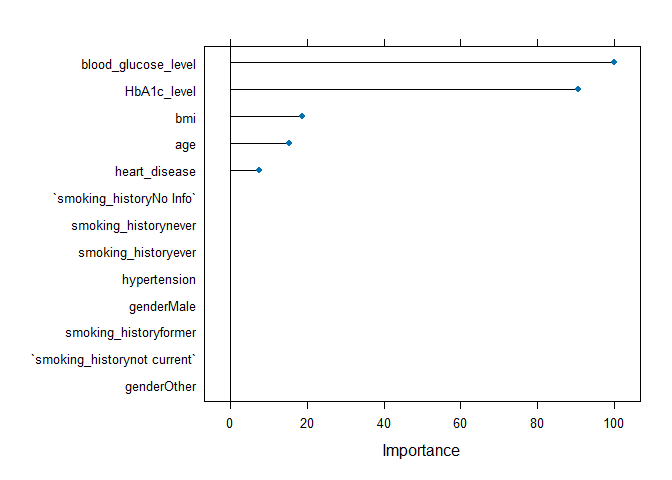
The classification and regression model estimated shows that the model has an accuracy of approximately 95.67%. This shows that the model classifies the respondents correctly into their respective categories as either (0 or 1), 95.67% of the time.

### Variable Importance

rpart variable importance  
  
 Overall  
blood\_glucose\_level 100.000  
HbA1c\_level 90.596  
bmi 18.780  
age 15.263  
heart\_disease 7.401  
smoking\_historyformer 0.000  
smoking\_historynever 0.000  
hypertension 0.000  
smoking\_historyever 0.000  
`smoking\_historyNo Info` 0.000  
genderOther 0.000  
genderMale 0.000  
`smoking\_historynot current` 0.000

The results show the most important and significant variable in the classification and regression trees model developed are HbA1c\_level, bmi, age, and so on. From the results, blood glucose level is 100% important to be in our model, followed by HbA1c\_level with 90.596%, bmi with 18.780%, age with 15.263%, and heart disease with 7.401%. 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, 240, 240, 240, 240, 239, ...   
Resampling results across tuning parameters:  
  
 mtry Accuracy Kappa   
 2 0.9509746 0.5039124  
 7 0.9633034 0.6686704  
 13 0.9606255 0.6554870  
  
Accuracy was used to select the optimal model using the largest value.  
The final value used for the model was mtry = 7.

### Confusion Matrix

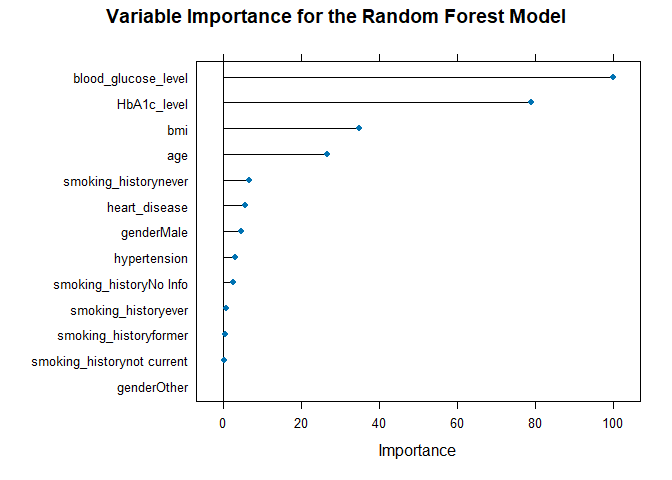
Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 276 0  
 Yes 0 24  
   
 Accuracy : 1   
 95% CI : (0.9878, 1)   
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.00000000001369  
   
 Kappa : 1   
   
 Mcnemar's Test P-Value : NA   
   
 Sensitivity : 1.00   
 Specificity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

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  
blood\_glucose\_level 100.0000  
HbA1c\_level 78.9797  
bmi 34.8986  
age 26.7547  
smoking\_historynever 6.6807  
heart\_disease 5.7516  
genderMale 4.6643  
hypertension 3.0159  
smoking\_historyNo Info 2.6892  
smoking\_historyever 0.7317  
smoking\_historyformer 0.4313  
smoking\_historynot current 0.2903  
genderOther 0.0000

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, blood glucose level comprises 100% followed by HbA1c\_level with 78.89%, bmi with 34.89%, age with 26.7548% 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, 239, 241, 240, 240, ...   
Resampling results across tuning parameters:  
  
 k Accuracy Kappa   
 5 0.9256781 0.13483757  
 7 0.9250059 0.10262066  
 9 0.9250282 0.09983629  
 11 0.9233613 0.07273579  
 13 0.9220223 0.03911334  
  
Accuracy was used to select the optimal model using the largest value.  
The final value used for the model was k = 5.

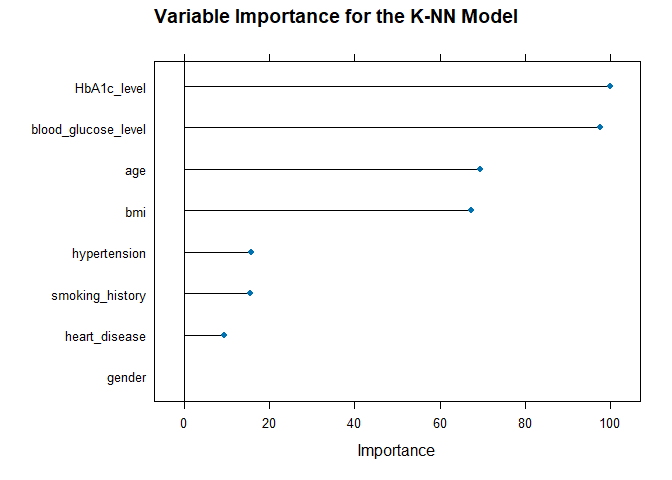
### Classification Accuracy

Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 275 17  
 Yes 1 7  
   
 Accuracy : 0.94   
 95% CI : (0.9068, 0.9641)  
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.118134   
   
 Kappa : 0.4141   
   
 Mcnemar's Test P-Value : 0.000407   
   
 Sensitivity : 0.29167   
 Specificity : 0.99638   
 Pos Pred Value : 0.87500   
 Neg Pred Value : 0.94178   
 Prevalence : 0.08000   
 Detection Rate : 0.02333   
 Detection Prevalence : 0.02667   
 Balanced Accuracy : 0.64402   
   
 '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 94% implying that the model correctly predict and classify patients in their correct categories 94% 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 97.508  
age 69.416  
bmi 67.277  
hypertension 15.877  
smoking\_history 15.634  
heart\_disease 9.526  
gender 0.000

The percentage contribution of each variable to the occurrence of the diabetes is as shown above with HbA1c\_level having 100%, followed by blood glucose level with 97.508%, age with 69.416% 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: 241, 240, 240, 239, 240, 240, ...   
Resampling results across tuning parameters:  
  
 usekernel Accuracy Kappa   
 FALSE 0.7716806 0.3059229  
 TRUE 0.9326830 0.2267197  
  
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

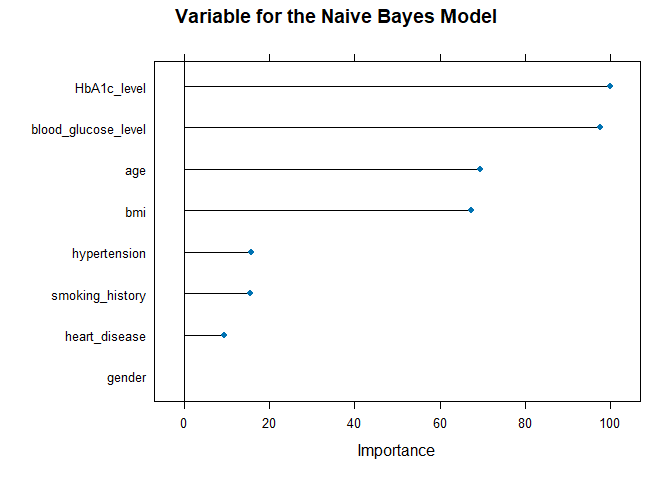
Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 276 17  
 Yes 0 7  
   
 Accuracy : 0.9433   
 95% CI : (0.9108, 0.9666)  
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.0783503   
   
 Kappa : 0.4311   
   
 Mcnemar's Test P-Value : 0.0001042   
   
 Sensitivity : 0.29167   
 Specificity : 1.00000   
 Pos Pred Value : 1.00000   
 Neg Pred Value : 0.94198   
 Prevalence : 0.08000   
 Detection Rate : 0.02333   
 Detection Prevalence : 0.02333   
 Balanced Accuracy : 0.64583   
   
 '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, 94.33% 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 97.508  
age 69.416  
bmi 67.277  
hypertension 15.877  
smoking\_history 15.634  
heart\_disease 9.526  
gender 0.000

Naive bayes give resuls simialar to that of k-NN. The percentage contribution of each variable to the occurrence of the diabetes is as shown above with HbA1c\_level having 100%, followed by blood glucose level with 97.508%, age with 57.0% 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: 240, 239, 240, 240, 241, 240, ...   
Resampling results:  
  
 Accuracy Kappa   
 0.9533402 0.5847142  
  
Tuning parameter 'C' was held constant at a value of 1

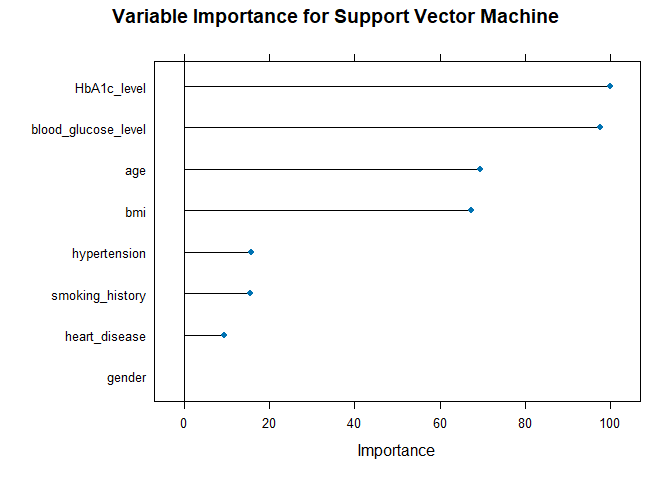
### Prediction and Classification Accuracy

Confusion Matrix and Statistics  
  
 Reference  
Prediction No Yes  
 No 276 10  
 Yes 0 14  
   
 Accuracy : 0.9667   
 95% CI : (0.9396, 0.9839)  
 No Information Rate : 0.92   
 P-Value [Acc > NIR] : 0.0007545   
   
 Kappa : 0.7204   
   
 Mcnemar's Test P-Value : 0.0044265   
   
 Sensitivity : 0.58333   
 Specificity : 1.00000   
 Pos Pred Value : 1.00000   
 Neg Pred Value : 0.96503   
 Prevalence : 0.08000   
 Detection Rate : 0.04667   
 Detection Prevalence : 0.04667   
 Balanced Accuracy : 0.79167   
   
 '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 96.67%. These results implies that the algorithm classifies patients in the correct categories (0 or 1) 96.67% of the time with only 3.33% chances of making a mis-classification error.

### Variable Importance

ROC curve variable importance  
  
 Importance  
HbA1c\_level 100.000  
blood\_glucose\_level 97.508  
age 69.416  
bmi 67.277  
hypertension 15.877  
smoking\_history 15.634  
heart\_disease 9.526  
gender 0.000



HbA1c\_level has 100% in our model followed by blood glucose level with 97.508%, age 69.416%, bmi with 67.277% and hypertension with 15.877% 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>