Turkish Journal of Engineering
https://dergipark.org.tr/en/pub/tuje
e-ISSN 2587-1366

A predictive machine learning framework for diabetes

Danjuma Maza [*1], **Joshua Olufemi Ojo** [1], **Grace Olubumi Akinlade** [1]

[1] *Obafemi Awolowo University Ile-Ife, Department of Physics and Engineering Physics, Nigeria, mazadd@oauife.edu.ng, jojo@oauife.edu.ng, gakinlade@oauife.edu.ng*

Cite this study: Maza, D., Ojo, J. O., & Akinlade, G. O. (2024). A predictive machine learning framework for diabetes. Turkish Journal of Engineering, 8 (3), 583-592

https://doi.org/10.31127/tuje.1434305

Keywords
Classification
Diabetes
Prediction
Accuracy
Recall

Research Article

Received: 09.02.2024
Revised: 16.04.2024
Accepted: 17.04.2024
Published: 15.07.2024

Abstract

Diabetes, a non-communicable disease, is associated with a condition indicative of too much glucose in the bloodstream. In the year 2022, it was estimated that about 422 million were living with the disease globally. The impact of diabetes on the world economy was estimated at $ 1.31 trillion in the year 2015 and implicated in the death of 5 million adults between the ages of 20 and 79 years globally. If left untreated for an extended time, could result in a host of other health complications. The need for predictive models to supplement the diagnostic process and aid the early detection of diabetes is therefore important. The current study is an effort geared toward developing a machine learning framework for the prediction of diabetes, expected to aid medical practitioners in the early detection of the disease. The dataset used in this investigation was sourced from the Kaggle database. The dataset consists of 100,000 entries, with 8,500 diabetics and 91,500 non-diabetics, indicating an imbalanced dataset. The dataset was modified to achieve a more balanced dataset consisting of 8,500 entries each for the diabetic and non-diabetic classes. Gradient Boosting classifier (GBC), Adaptive Boosting classifier (ADA), and Light Gradient Boosting Machine (LGBM) were the best three performing classifiers after comparing fifteen classifiers. The proposed framework is a stack model consisting of GBC, ADA, and LGBM. The ADA classifier was utilized as the meta-model. This model achieved an average accuracy, area under the curve (AUC), recall, precision, and f1-score of 91.12 ± 0.75 %, 97.83 ± 0.29 %, 92.03 ± 1.55 %, 90.40 ± 1.01 %, and 91.12 ± 0.77 %, respectively. The selling point of the proposed framework is the high recall of 92.03 ± 1.55 %, indicating that the model is sensitive to both the diabetic and the non-diabetic classes.

1. Introduction

An estimated 422 million people globally are believed to be living with diabetes in the year 2022, according to the World Health Organization [1]. The International Diabetes Federation (IDF) projected this figure to rise to 643 million and 783 million by the years 2030 and 2045, respectively [2]. As of 2015, the impact of diabetes on the global economy was estimated to be US $1.31 trillion. Furthermore, in 2015 diabetes was implicated in the death of an estimated 5 million adults aged between 20 and 79 years [3]. Studies have shown that the risk of getting infections is enhanced in people living with diabetes compared to the normal population, with an attendant increase in their morbidity and mortality [3].

Diabetes, most often, is associated with hyperglycemia, a condition indicative of too much sugar (glucose) in the bloodstream. This happens when the body has too little insulin (Type I diabetes) or if the body can't properly utilize the available insulin (Type II diabetes), leading to insulin resistance [4]. If left untreated for an extended period, hyperglycemia can damage the nervous system, blood vessels, tissues, and critical organs [5,6], which may result in a host of complications, namely: renal failure, retinal failure, coma, cardiovascular dysfunction, cerebral vascular dysfunction, peripheral vascular disorders, sexual dysfunction, joint failure, weight loss, ulcer, ocular diseases, kidney failure, and loss of immunity against pathogens [5–9]. Additionally, diabetes has been implicated in adverse pregnancy outcomes, including increased risk of maternal and fetal morbidity and mortality [10]. Furthermore, hyperglycemia has been known to increase the risk for adverse events and outcomes for patients undergoing cancer treatment [11,12].

Conventionally, the diabetes diagnosis process requires that multiple blood sugar tests be taken both before and after a meal. This presents practitioners with a difficult task in the diagnosis of diabetes. However, the diabetes diagnostic regime could be made simpler with the use of computational methods. Over the years, predictive models have been developed to aid in the diagnosis and consequently the treatment of diabetes.

The earlier diabetes predictive models [13–17] were developed based on simple scoring or logistic regression. These models used data derived from population-based studies that included risk factors that can easily be measured, such as age, BMI, waist circumference, physical activity, daily consumption of fruits, vegetables, or berries, history of antihypertensive drug treatment, and history of high blood glucose. In these early models different scores, ranging from 0 to 10 or 0 to 20, were assigned to each risk factor based on the ß-coefficients of a regression model (Equation 1).

$$x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (1)$$

The emergence of new technologies, however, ushered in new techniques and methods that could improve the performance and applicability of these earlier predictive models. To this end, cutting-edge algorithms such as: Machine learning (ML) and Deep Learning (DL) could and have been employed to utilize the massive amount of data being churned out by both research cohorts and medical practitioners [4,18–23]. Machine learning is fast becoming a tool of choice among researchers in the medical field for building predictive models to supplement the diagnostic process and treatment of various diseases [24]. This tool has demonstrated its usefulness in handling large numbers of variables, detecting, and interpreting complex relationships between variables [9,24–27].

This research effort is geared towards the development of a predictive model using machine learning tools implemented using Python programming language. In this study, we propose a stacked model consisting of a gradient boosting classifier (GBC), an adaptive boost classifier (ADA), and a light gradient boosting machine (LGBM), with the ADA boost as the meta_model. The proposed framework is intended to achieve not only high accuracy but also high sensitivity (recall) towards both the diabetic and non-diabetic classes. The proposed model is expected to aid medical practitioners in the clinical diagnosis of diabetes. Furthermore, this model could be useful and helpful to numerous undiagnosed diabetic patients who, very often, are unaware of their condition.

In a study conducted by Maniruzzaman [28] and co-researchers, a framework was developed where missing values and outliers were replaced by the group median and median values, respectively. Using a combination of Radom Forest feature selection and Random Forest classifier, the authors achieved accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and area under the curve of 92.26 %, 95.96 %, 79.72 %, 91.14 %, 91.20 %, and 93.0 %, respectively.

In a related development, a research effort undertaken by [29] utilized the Pima Indians Diabetes Dataset, which consisted of 768 entries, to develop a predictive model for diabetes. The authors applied feature selection techniques along with five classification algorithms, namely: Support Vector Machine (SVM), Multi-Layer Perception (MLP), Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT), to achieve the highest classification accuracy of 78.7% with MLP.

In another research effort [7], the authors used an artificial neural network that was implemented in the Jacobian Neural Network (JNN) environment, for the prediction of diabetes. The dataset used in the study consists of 1004 entries and 9 features. Their effort resulted in a predictive model for diabetes with 87.3 % accuracy. On the other hand, Hasan and co-researchers [8] proposed a framework for the prediction of diabetes using the Pima Indian Diabetes Dataset. A weighted ensemble of different machine learning classifiers (K-Nearest Neighbors, Decision Tree, Random Forest, Naïve Bayes, Ada Boost, XGBOOST Multi-Layer Perception) was employed. The Area Under the Curve (AUC) was the metric chosen for evaluating the performance of the mode. The ensemble classifier achieved a sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC of 0.789, 0.934, 0.092, 66.234, and 0.950, respectively.

Butt and collaborators [30], proposed a model for early-stage detection and prediction of diabetes. The authors compared three classifiers, Random Forest, Multi-Layer Perception, and Logistic Regression, where the Multi-Layer Perception was the best-performing model with an accuracy of 86.08%. This was improved to 87.26 % by using Deep Learning based Long Short-Term Memory.

In another study conducted by Roy and collaborators [31], a diabetes predictive model was developed, based on an Artificial Neural Network (ANN) algorithm. In the study, median value data imputation was used to handle missing data, while the imbalance in the dataset was handled using a combination of SMOTETomek and random oversampling. This model achieved an accuracy of 98%, however, like other models using the Pima Indian dataset, it was plagued with concerns over the dataset. The dataset consisted of all females aged 21 years and above, with the population limited to Native Americans. Using the Pima Indian Diabetes Dataset, [4] developed a framework to conduct a comparative study based on the effectiveness of a three-category categorization model. In the first category, the authors considered the model's performance with and without data pre-processing. In the second category, using the Recursive Feature Elimination (RFE) feature selection method, the performance of five different algorithms was compared. While in the third category, data augmentation was done employing the SMOTE oversampling method, and model performances were compared with and without.

With dataset collected from the Murtala Mohammed Specialist Hospital, Kano, Nigeria [32] developed a supervised predictive model based on Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Radom Forest (RF), Naïve Baiyes (NB), Gradient Boosting Classifier (GBC). In that study RF achieved the highest accuracy of 88.76 %, while RF and GBC had better receiver operating characteristic curve (ROC) of 86.28 %. Lai and cohorts [33] used a dataset of 13,309 Canadian patients aged between 18 and 90 years. In their research effort, RF was the best-performing model with sensitivity and area under the receiver operating characteristic curve (AROC) of 84% and 73.4 %, respectively.

Abnoosian and co-investigators [34] proposed a multi-classification model framework to predict diabetes in three classes: namely; diabetic, non-diabetic, and pre-diabetics. Their framework achieved accuracy, precision, recall, and F1-score values of 0.9887, 0.9861, 09792, 0.9851, and 0.999, respectively.

2. Method

2.1. Dataset

The dataset used in this study was sourced from the Kaggle database [35]. This dataset donated by Mohammed Mustafa, is a combination of demographic and medical data of patients, including their diabetes status. Pre-processing analysis showed that the dataset consists of 100,000 entries with 41,430 males, 58,552 females, and 18 others, with no missing values; It has eight features, namely: gender, age, hypertension, heart disease, smoking history, body mass index (BMI), haemoglobin A1c (HbA1c) level, and blood glucose level; gender and smoking history were categorical features while the rest of the features were numerical; the smoking history feature has three categories: current, never, and no info. Further pre-processing analysis shows that there were 8500 diabetics and 91500 non-diabetics in the dataset, indicating an imbalance in the dataset. Two approaches were used to handle the imbalance in the dataset before experimentation. First, the dataset was augmented using the SMOTE oversampling method. This resulted in the dataset being augmented to 158100 entries. Secondly, the dataset was reshuffled and the size was reduced to 17,000 entries, with 8500 diabetics and 8500 non-diabetics.

2.2 Data analysis

To determine the inter-relationship between the different features of the dataset, a correlation analysis was carried out on the original unbalanced and the reduced-balanced datasets. Figure 1 and 2 display the correlation heat map of the unbalanced and reduced-balanced datasets, respectively. These heat maps show the correlation coefficients (R-value) between respective features, and also with the target feature (diabetes). These coefficients indicate the strength and direction of the relation. These coefficients indicate the strength and direction of the relation.

For further analysis, features with continuous data fields such as age, BMI, HbA1c level, and blood glucose levels were modified into categorical fields. To this end, the age feature was split into four fields (< 40, 40 – 49, 50 – 60, and > 60 yrs). On the other hand, the BMI feature was split into three fields (< 18.5, 18.5 – 25, 25 – 30, > 30 kg/m^{-2}), the HbA1c level was split into three fields (< 5.7, 5.7 – 6.4, > 6.4 mmol/mol), and the blood glucose level was also split into three fields (< 126, 126 – 200, > 200 µmol/dl).

Figure 3 - 9 show the distribution of diabetes across the various categorical fields of these features. These distributions show the predisposition to diabetes of people across the categorical field groupings, for the respective features.

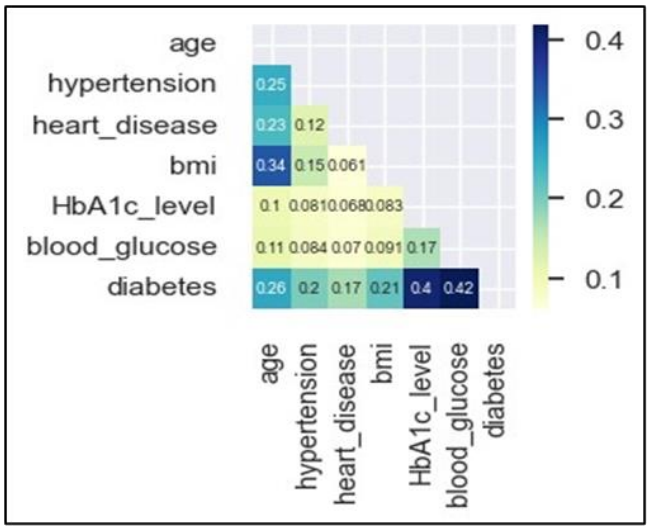

Figure 1. The heat map of the unbalanced dataset.

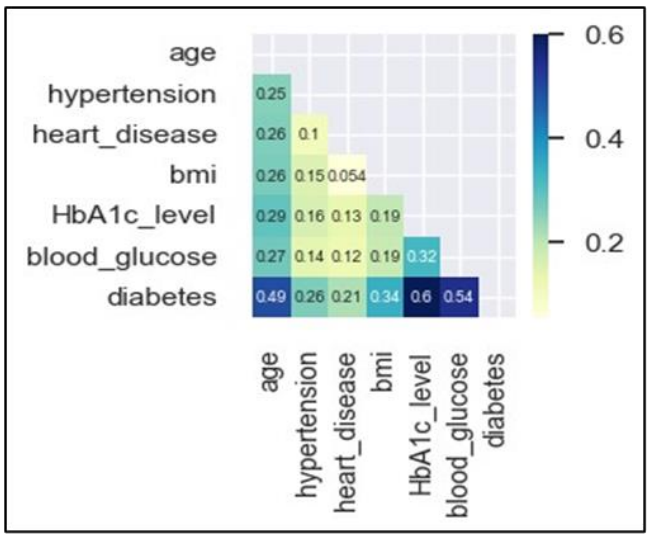

Figure 2. The correlation heat map of the balanced dataset.

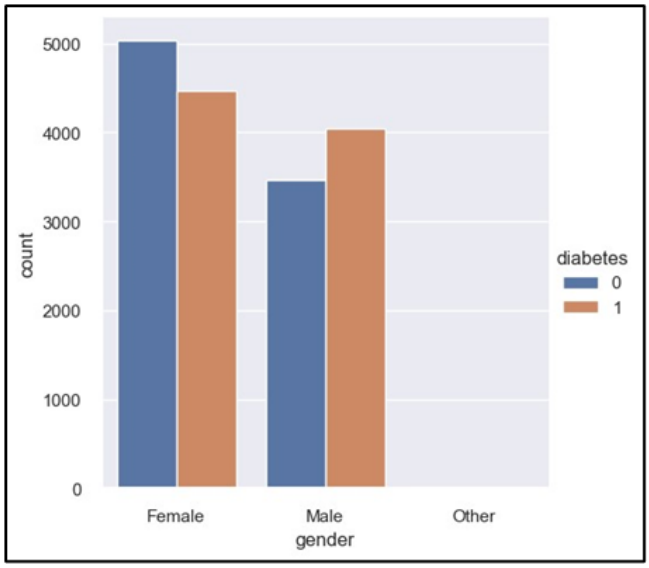

Figure 3. Diabetes distribution across gender groups.

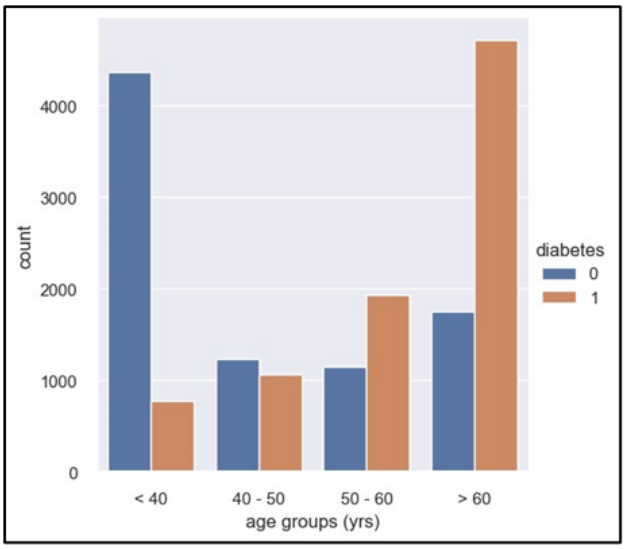

Figure 4. Diabetes distribution across age groups.

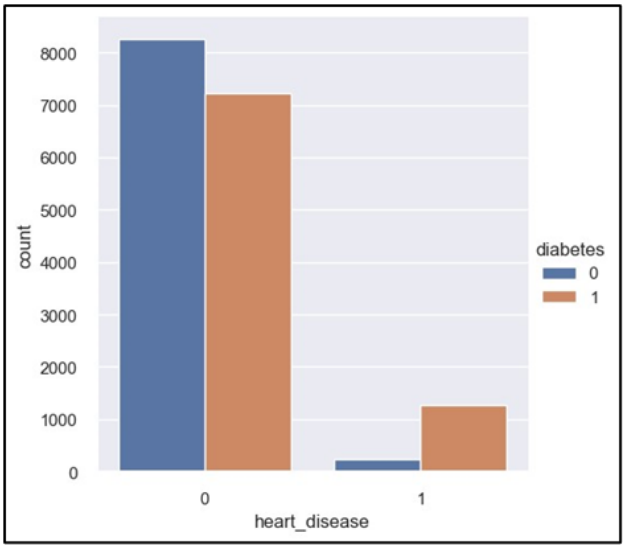

Figure 5. Diabetes distribution across heart disease status (On the x-axis: 0 = no heart disease, 1 = heart disease).

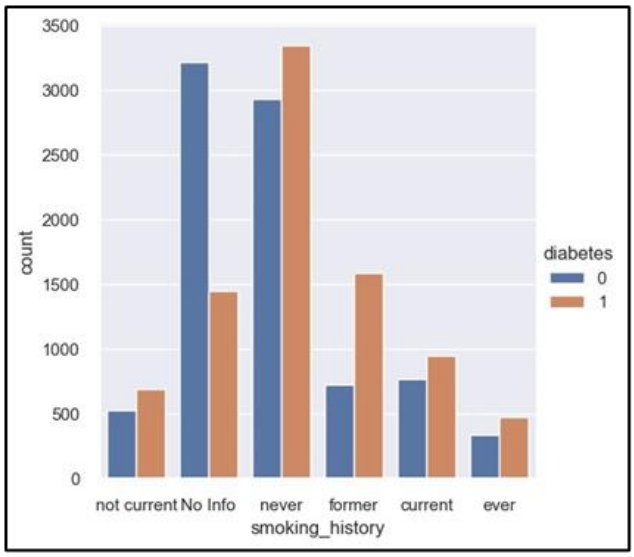

Figure 6: Diabetes distribution according to smoking history (x-axis: not current = not currently smoking, no info = no information, never = never smoked, former = former smoker, current = current smoker, ever = ever smoking).

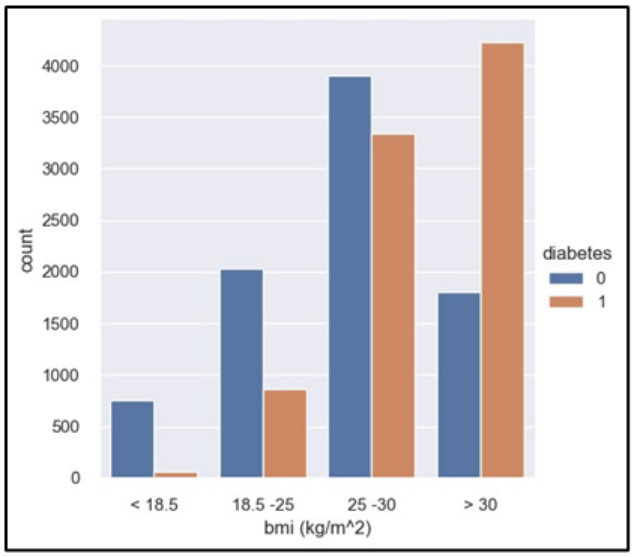

Figure 7. Diabetes distribution across bmi groupings.

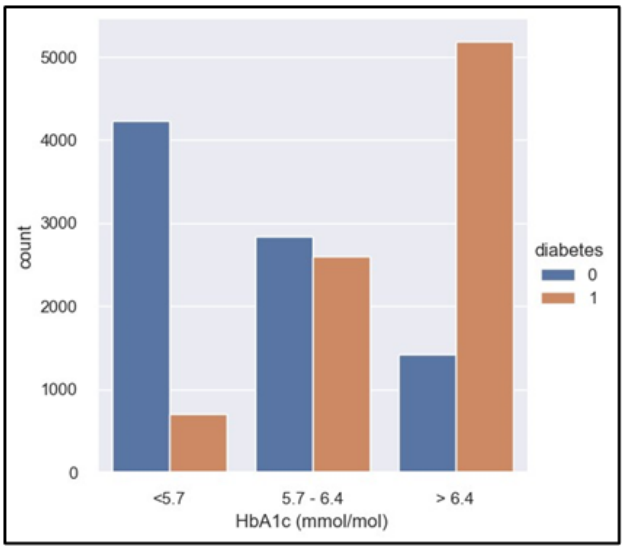

Figure 8. Diabetes distribution across HbA1c groups.

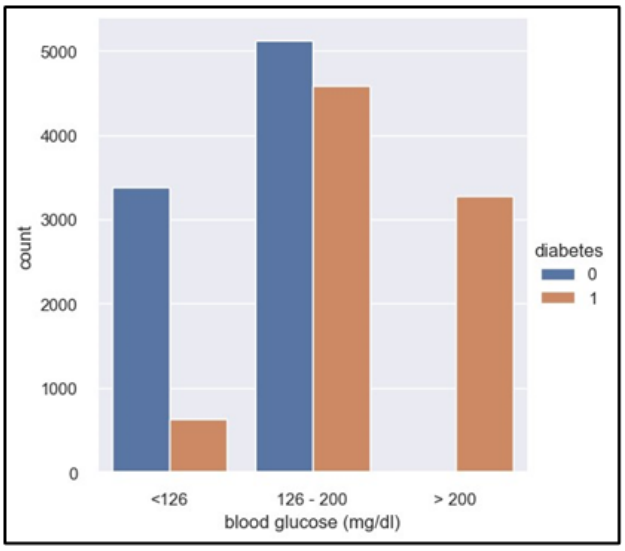

Figure 9. Diabetes distribution according to blood glucose level.

2.3 Experimentation

In this study, the PyCaret (version 3.0.2), a Python library, was employed for the investigation. The PyCaret

library is preloaded with fifteen (15) models (classifiers). These are: gradient boosting classifier (GBC), light gradient boosting machine (LGBM), adaptive boosting classifier (ADA), extreme gradient boosting (xgboost), random forest classifier (RF), extra trees classifier (ET), logistic regression (LR), linear discriminant analysis (LDA), k-nearest neighbors' classifier (KNN), decision tree classifier (DT), SVM-linear kernel (SVM) ridge classifier (RIDGE), dummy classifier (dummy), naïve bayes (NB), and quadratic discriminant analysis (QDA). The dataset was split into training and test sets using the PyCaret setup. The default setting of 0.3 was retained in splitting the dataset, thus, allocating 70 % of the dataset for training the model and 30 % for testing or evaluation. The setup was first initialized with the original, unbalanced dataset of 100,000 entries. The 15 classifiers were compared, where GBC was the best-performing model with an accuracy, AUC, recall, precision, and F1-score of 97.23 %, 97.91 %, 68.37 %, 98.60 %, and 80.74 %, respectively, followed by ADA, and LGBM, with accuracies of 97.22 % and 97.19 %, respectively. The best-performing model, GBC, was further tuned with 10-fold cross-validation to achieve an average accuracy of 97.20 %, with an average recall of 68.25 %. Next, the setup was initialized with the augmented dataset of 158100 entries. This was an attempt to improve on the recall. With the augmented dataset, the LGBM was the best-performing classifier with an average accuracy of 97.15 % and recall of 68.86 %, following a 10-fold cross-validation. There was no significant improvement in the recall with the augmented dataset. Thereafter, the setup was initialized with the reduced-balanced dataset of 17,000 entries, having 8,500 diabetics and 8,500 non-diabetics. With this setup (utilizing the reduced-balanced dataset), the 15 models were again compared. The best four (4) performing models from this setup were GBC, ADA, LGBM, and XGBOOST, achieving accuracies of 91.25 %, 91.21 %, 90.83 %, and 90.71 %, respectively with recalls of 92.30 %, 92.12 %, 91.78 %, and 91.28 %. The recalls with this setup were significantly improved. Table 1 shows the best four (4) classifiers with the different versions of the dataset.

Table 1. Best four models with different versions of the dataset.

Dataset	Model	Accuracy	AUC	Recall	Precision	F1
imbalanced	gbc	0.9723	0.9791	0.6837	0.9860	0.8074
	ada	0.9722	0.9790	0.6929	0.9715	0.8088
	lgbm	0.9719	0.9788	0.6887	0.9734	0.8066
	xgboost	0.9715	0.9779	0.6961	0.9568	0.8058
augmented	lgbm	0.9715	0.9787	0.6886	0.9659	0.8039
	xgboost	0.9709	0.9779	0.6939	0.9501	0.8020
	gbc	0.9683	0.9757	0.7151	0.8907	0.7931
	rf	0.9677	0.9613	0.6983	0.8995	0.7861
reduced-balanced	gbc	0.9125	0.9788	0.9230	0.9042	0.9134
	ada	0.9121	0.9786	0.9212	0.9049	0.9129
	lgbm	0.9083	0.9774	0.9178	0.9008	0.9092
	xgboost	0.9071	0.9761	0.9128	0.9027	0.9077

Table 2. Performance of tuned GBC, ADA, LGBM, xgboost, and NB classifiers.

Model		Accuracy	AUC	Recall	Precision	F1-score
gbc	Mean	0.9119	0.9780	0.9212	0.9046	0.9127
	Std	0.0105	0.0034	0.0143	0.0131	0.0105
ada	Mean	0.9115	0.9791	0.9224	0.9030	0.9125
	Std	0.0091	0.0029	0.0116	0.0129	0.0088
lgbm	Mean	0.9097	0.9758	0.9240	0.8985	0.9110
	Std	0.0069	0.0034	0.0133	0.0103	0.0070
xgboost	Mean	0.8974	0.9769	0.9751	0.8441	0.9048
	Std	0.0083	0.0035	0.0036	0.0117	0.0070
naive	Mean	0.8303	0.9069	0.8173	0.8390	0.8280
	Std	0.0137	0.0107	0.0170	0.0133	0.0142

Table 3. Performance of the different stacked models.

		Accuracy	AUC	Recall	Precision	F1-score
Stacked model of tuned gbc, ada, lgbm, xgboost, with untuned nb as meta_model	Mean	0.9067	0.9741	0.9513	0.8736	0.9107
	Std	0.0083	0.0043	0.0101	0.0117	0.0077
Stacked model of tuned gbc, ada, lgbm, with tuned nb as meta_model	Mean	0.9098	0.9724	0.9168	0.9043	0.9104
	Std	0.0080	0.0047	0.0138	0.0101	0.0082
Stacked model of tuned gbc, ada, lgbm, with tuned gbc as meta_model	Mean	0.9112	0.9783	0.9203	0.9040	0.9120
	Std	0.0075	0.0029	0.0155	0.0118	0.0077
Stacked model of tuned gbc, ada, lgbm, with tuned ada as meta_model	Mean	0.9120	0.9779	0.9249	0.9018	0.9131
	Std	0.0091	0.0028	0.0180	0.0089	0.0096
Stacked model of tuned gbc, ada, lgbm, with tuned lgbm as meta_model	Mean	0.9113	0.9782	0.9222	0.9028	0.9123
	Std	0.0088	0.0028	0.0154	0.0117	0.0090

The best 3 performing models were then separately created and tuned (Table 2). These tuned models were then stacked to produce a more robust model. After experimenting with GBC, ADA, LGBM, XGBOOST, and NB

as the meta-model (Table 3), the ADA classifier was adapted as the meta-model, since it achieved a higher recall without sacrificing much accuracy and precision. Although with the NB classifier as the meta-model, the stacked model achieved the highest recall of 95.13 %, accuracy, and precision were sacrificed. The final model therefore was a stack of GBC, ADA, and LGBM, with ADA as the meta-model. Figure 10 shows the framework for the model development.

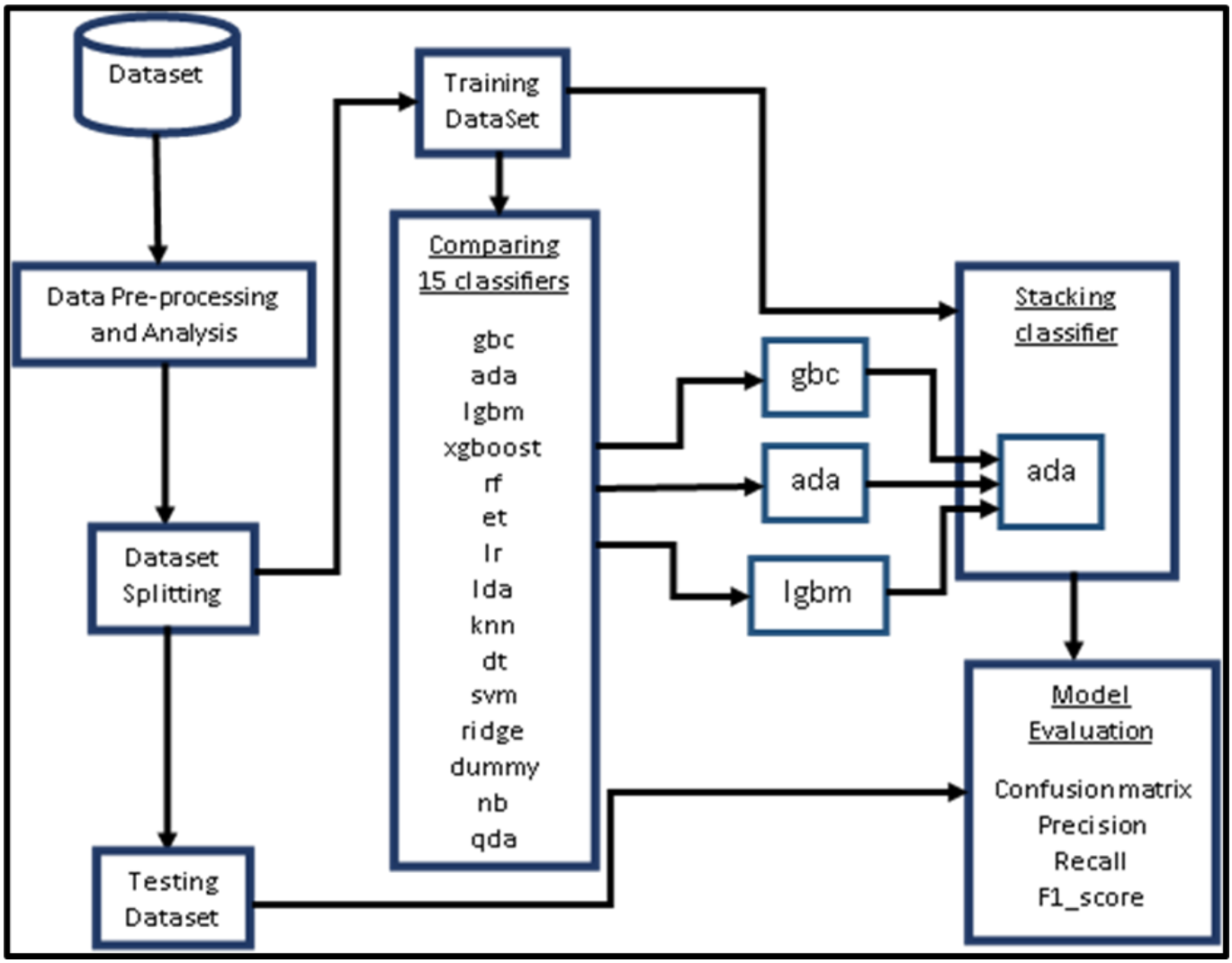

Figure 10. Framework for the model development.

2.4 Evaluation metrics

The outcomes of predictions from the model developed were assessed using the following metrics: Confusion matrix, Accuracy, AUC, Recall, Precision, and F1-score. These are respectively defined as follows:

2.4.1 Confusion matrix

This is a matrix (Figure 11) used for the evaluation of a model's overall performance, highlighting, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions of a classification model [36 - 38]. It compares the actual target values with the values predicted by the classification model. It is a matrix summarizing the number of correct and incorrect predictions by the classifier. A confusion matrix forms the basis for the calculation of other performance metrics such as accuracy, precision, recall, and F1-score.

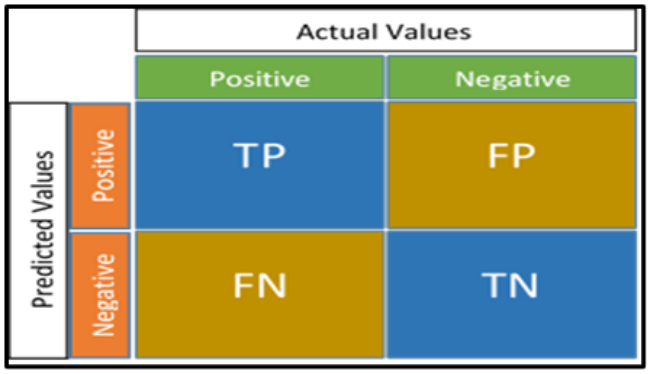

Figure 11. Confusion matrix highlighting the positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions of a classification model.

2.4.2 Accuracy

The accuracy of a classification model is the ratio of

correct predictions to the total number of predictions (Equation 2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

It is a measure of how often a model makes correct predictions. The accuracy metric, however, is not suited for imbalanced data. For instance, if the model predicts that all the predicted values are in the majority class, the accuracy will be high but the model itself is not accurate.

2.4.3 Precision

The precision of a classification model is the ratio of the total number of correctly predicted positive classes to the total number of predicted positive classes (Equation 3).

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

In a nutshell, it gives us a measure of how many predictions are actually positive out of all the positive predictions.

2.4.4 Recall

Recall, also referred to as the sensitivity of a classification model is a measure of actual observations that are correctly predicted (Equation 4). It is thus a measure of how many positive classes are predicted as positive, given as:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

Recall is an important metric in the present study where actual positive cases mustn't go undetected.

2.4.5 F1-Score

F1-Score is the harmonic mean of the precision and recall of a classification model. It is defined in Equation 5.

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (5)$$

F1-Score maintains the balance between Recall and Precision of a classification model.

3. Results and discussion

Figure 1 and 2 show that with the unbalanced dataset, all the features have a weak but positive correlation with one another and with diabetes. However, with the reduced-balanced dataset, all the features have a weak positive correlation with one another, except for HbA1c/diabetes and blood glucose level/diabetes which have R-values of 0.6 and 0.54, respectively. Pre-modeling analysis of the reduced-balanced dataset Figures 3 to 9 show the predisposition to diabetes as follows: People above 50 years of age; individuals with heart disease; BMI greater than 30 kg/m^2; HbA1c level above 6.4 mmol/mol; and blood glucose level above 200 mg/dl; respectively are predisposed to diabetes.

The final model in this study is a stack model consisting of GBC, ADA, and LGBM with ADA as the meta-model. Figure 12 depicts the confusion matrix of the final model based on the reduced-balanced dataset of 17,000 entries. On the other hand, Figure 13 shows the classification report, while Figure 14 and 15 show the class prediction error, and precision-recall curve, respectively, of the stack classifier. The confusion matrix (Figure 12) shows that out of the 5,100 entries in the test set, the diabetic and non-diabetic classes have 2,550 entries each. Of the 2,550 diabetics, the model predicted 2,374 correctly and only 176 incorrectly. On the other hand, 2,290 of the 2,550 non-diabetics were correctly predicted by the model while 260 were incorrectly predicted. This gives an overall accuracy of 91.45 % for the model. A comparison of the performance of the model developed in this study with other studies is shown in Table 4.

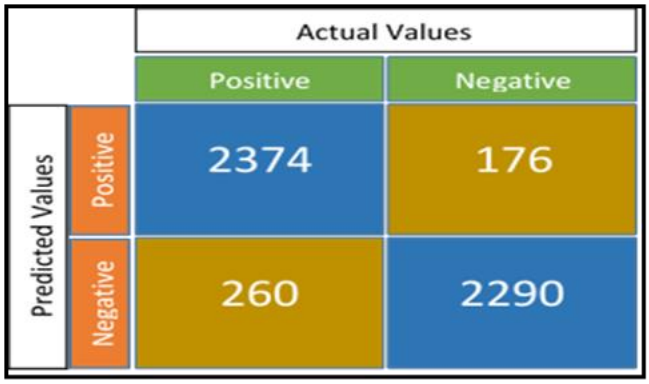

Figure 12. The confusion matrix of the stacked classifier used for predicting diabetes.

Table 4. Comparison of present study with similar studies.

Accuracy (%)	AUC (%)	Recall (%)	Precision (%)	Reference
91.12	97.83	92.03	90.40	Present study
92.26	93.0	95.26	79.72	28
78.7	-	-	-	29
87.3	-	-	-	7
-	95.0	78.9	93.4	8
87.26	-	-	-	30
98.17		97.00	99.00	31
88.76	-	-	-	32
-	-	73.4	-	33
98.87	-	97.92	98.61	34

Furthermore, the classification report (Figure 13) shows that the model has a sensitivty (ability to correctly predict positive cases) of 92.9 % and specificity (ability to predict true negatives) of 92.6 %. This high values for both sensitivity and specificity are a pointer to the model's great potential and efficacy to assist medical practioners in early diagnosis and treatment of diabetic patients. It is important to stress that the sensitivity (recall) of a classification model is an important metric that shows the sensitivity of the model, particularly towards the positive class. A high accuracy, or precision alone does not speak to the sensitivity or otherwise of a classification model. The values of both the sensitivity

and specificity of the model shows that the model is "sensitive" to both diabetics and non-diabetics. For a disease like diabetes where early detection is crucial, a predictive model designed for the detection of the disease in patients must be sensitive to the positive class in particular. It is not enough for a predictive model of this nature to have high accuracy and precision; it must also have high sensitivity towards the positive class. This ensures that positive cases are not easily missed which will aid in the early diagnosis and treatment of patients. The authors believe that a recall of 92.9 % for the positive class is a good starting point while seeking improvements.

Furthermore, the prediction error plot (Figure 14) shows that fewer diabetic cases were falsely classified as non-diabetic than non-diabetic cases that were falsely classified as diabetic. This again highlights the sensitivity of the model to the diabetic class. Further still, Figure 15, precision-recall curve, shows that the AUC (the blue shaded area) is large, which is a mark of a good model.

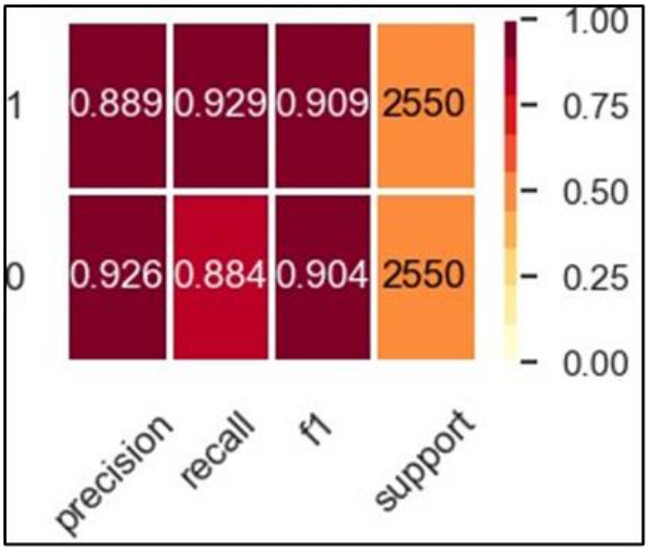

Figure 13. Classification report for the stacked classifier of GBC, ADA, and LGBM with ADA as the meta-model.

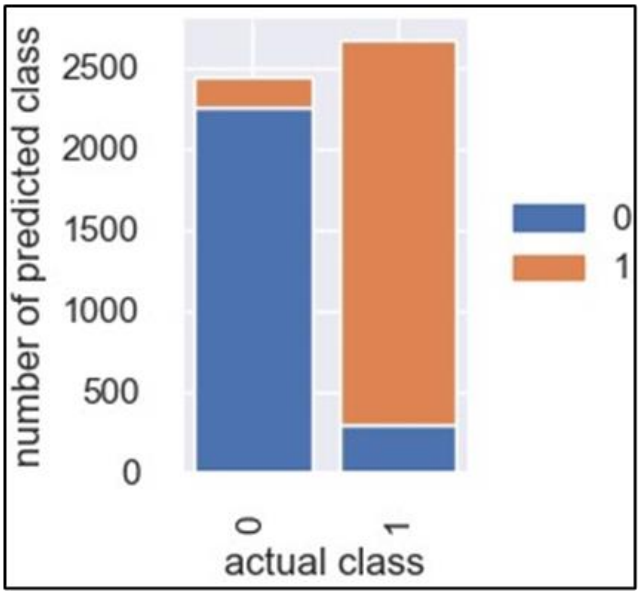

Figure 14. Class prediction error of the stacked classifier.

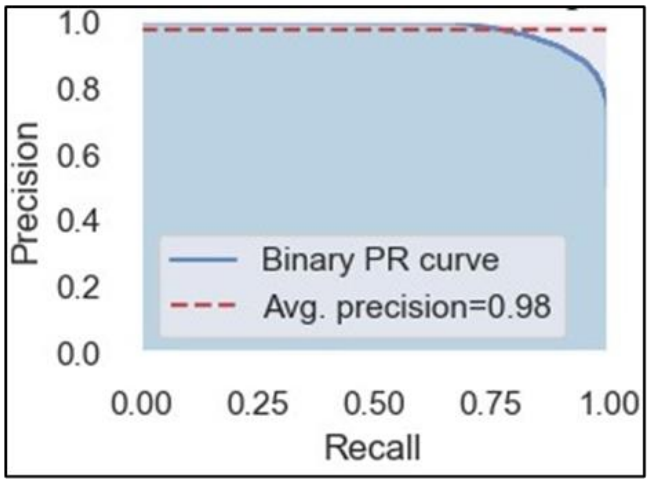

Figure 15. Precision-Recall curves of the stacked classifier.

4. Conclusion

In this study, a model was developed for classifying diabetes in humans, using a modified diabetes dataset sourced from the Kaggle database. The dataset was modified to have a balanced dataset which ensures a good recall for the model. The model is intended to aid medical practitioners in the diagnostic process of diabetes. The model developed in this study was a stack model consisting of GBC, ADA, and LGBM, with the ADA classifier as the meta-model. This stack model achieved an average accuracy, AUC, recall, precision, and F1-score of 91.12 ± 0.75 %, 97.83 ± 0.29 %, 92.03 ± 1.55 %, 90.40 ± 1.01 %, and 91.12 ± 0.77 %, respectively. The current model achieved a relatively high recall (sensitivity) without sacrificing accuracy and precision. Of particular note is the high AUC. The high AUC, high recall, and precision of the model highlight its potential clinical value and efficacy in assisting medical practitioners in diagnosing diabetes. While its high AUC highlights its overall good performance, its high recall means that very few positive cases will be wrongly classified as negative (low false negative rate), ensuring that positive cases are not easily misdiagnosed, which in turn will enhance early detection and treatment. Equally important is the high precision of the model, indicative of its low false positive rate, meaning that non-diabetics will not be easily misdiagnosed as diabetics. For a predictive model of this nature, the model must be sensitive to the positive class, ensuring that positive cases are easily captured.

Acknowledgement

The authors acknowledge the support and encouragement of ENVIRON LAB., Department of Physics and Engineering Physics Obafemi Awolowo University, Ile-Ife, Nigeria.

Author contributions

Danjuma Maza: Study conception, Design, Methodology, Data Analysis, Experimentation, Writing-Original draft preparation. **Joshua Olufemi Ojo:** Writing-Reviewing and Editing. **Grace Olubumi Akinlade:** Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. WHO. (2023). Diabetes, Diabetes Report. https://www.who.int/health-topics/diabetes#tab=tab_1
2. IDF (2021). Facts & figures. https://idf.org/about-diabetes/diabetes-facts-figures/
3. Woldaregay, A. Z., Årsand, E., Botsis, T., Albers, D., Mamykina, L., & Hartvigsen, G. (2019). Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes. Journal of medical Internet research, 21(5), e11030. https://doi.org/10.2196/11030
4. Sabitha, E., & Durgadevi, M. (2022). Improving the diabetes Diagnosis prediction rate using data preprocessing, data augmentation and recursive feature elimination method. International Journal of Advanced Computer Science and Applications, 13(9), 921-930. https://doi.org/10.14569/IJACSA.2022.01309107
5. Choubey, S., Agrahari, S., Shaw, A., Dhar, S., Sarma, R. R., Singh, S. K., Das, P., & Saha, B. (2023). Diabetes Prediction Using ML. International Journal for Research in Applied Science and Engineering Technology, 11(6), 4209-4212. https://doi.org/10.22214/ijraset.2023.54415
6. Marcovecchio, M. L. (2017). Complications of acute and chronic hyperglycemia. US Endocrinol, 13(1), 17-21. https://doi.org/10.17925/USE.2017.13.01.17
7. El_Jerjawi, N. S., & Abu-Naser, S. S. (2018). Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology, 121, 54-64. http://dx.doi.org/10.14257/ijast.2018.121.05
8. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531. https://doi.org/10.1109/ACCESS.2020.2989857
9. Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications, 36(4), 8610-8615. https://doi.org/10.1016/j.eswa.2008.10.032
10. Bashir, M., Naem, E., Taha, F., Konje, J. C., & Abou-Samra, A. B. (2019). Outcomes of type 1 diabetes mellitus in pregnancy; effect of excessive gestational weight gain and hyperglycaemia on fetal growth. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 13(1), 84-88. https://doi.org/10.1016/j.dsx.2018.08.030
11. Hammer, M., Storey, S., Hershey, D. S., Brady, V. J., Davis, E., Mandolfo, N., Bryant, A. L., & Olausson, J. (2019). Hyperglycemia and Cancer: A State-of-the-Science Review. Oncology Nursing Forum, 46(4), 459-472. https://doi.org/10.1188/19.ONF.459-472
12. Storey, S., Von Ah, D., & Hammer, M. (2017). Measurement of hyperglycemia and impact on the health outcomes in people with cancer: challenges and opportunities. Oncology Nursing Forum, 44(4), E141. https://doi.org/10.1188/17.ONF.E141-E151
13. Griffin, S. J., Little, P. S., Hales, C. N., Kinmonth, A. L., & Wareham, N. J. (2000). Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. Diabetes/metabolism Research and Reviews, 16(3), 164-171. https://doi.org/10.1002/1520-7560(200005/06)16:3<164::AID-DMRR103>3.0.CO;2-R
14. Park, P. J., Griffin, S. J., Sargeant, L., & Wareham, N. J. (2002). The performance of a risk score in predicting undiagnosed hyperglycemia. Diabetes Care, 25(6), 984-988. https://doi.org/10.2337/diacare.25.6.984
15. Lindstrom, J., & Tuomilehto, J. (2003). The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care, 26(3), 725-731. https://doi.org/10.2337/diacare.26.3.725
16. Heikes, K. E., Eddy, D. M., Arondekar, B., & Schlessinger, L. (2008). Diabetes risk calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes. Diabetes Care, 31(5), 1040-1045. https://doi.org/10.2337/dc07-1150
17. Stern, M. P., Williams, K., & Haffner, S. M. (2002). Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test?. Annals of Internal Medicine, 136(8), 575-581. https://doi.org/10.7326/0003-4819-136-8-200204160-00006
18. Kodama, S., Fujihara, K., Horikawa, C., Kitazawa, M., Iwanaga, M., Kato, K., ... & Sone, H. (2022). Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis. Journal of Diabetes Investigation, 13(5), 900-908. https://doi.org/10.1111/jdi.13736
19. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005
20. Nai-Arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132-142. https://doi.org/10.1016/j.procs.2015.10.014
21. Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220, 106773. https://doi.org/10.1016/j.cmpb.2022.106773
22. Singh, A., Halgamuge, M. N., & Lakshmiganthan, R. (2017). Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms. International Journal of Advanced Computer Science and Applications, 8(12), 1-10.
23. Tejedor, M., Woldaregay, A. Z., & Godtliebsen, F. (2020). Reinforcement learning application in diabetes blood glucose control: A systematic review. https://doi.org/10.1016/j.artmed.2020.101836
24. Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and

perspective. Artificial Intelligence in Medicine, 23(1), 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X
25. Asfaw, T. A. (2019). Prediction of diabetes mellitus using machine learning techniques. International Journal of Computer Engineering and Technology, 10(4), 145-148. https://doi.org/10.34218/ijcet.10.4.2019.004
26. Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10, 1-7. https://doi.org/10.1186/1472-6947-10-16
27. MacMahon, H., Naidich, D. P., Goo, J. M., Lee, K. S., Leung, A. N., Mayo, J. R., ... & Bankier, A. A. (2017). Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology, 284(1), 228-243. https://doi.org/10.1148/radiol.2017161659
28. Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of Medical Systems, 42, 92. https://doi.org/10.1007/s10916-018-0940-7
29. Ahuja, R., Sharma, S. C., & Ali, M. (2019). A diabetic disease prediction model based on classification algorithms. Annals of Emerging Technologies in Computing (AETiC), 3(3), 44-52. https://doi.org/10.33166/AETiC.2019.03.005
30. Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. Journal of Healthcare Engineering, 2021(1), 9930985. https://doi.org/10.1155/2021/9930985
31. Roy, K., Ahmad, M., Waqar, K., Priyaah, K., Nebhen, J., Alshamrani, S. S., ... & Ali, I. (2021). An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values. Complexity, 2021(1), 9953314. https://doi.org/10.1155/2021/9953314
32. Muhammad, L. J., Algehyne, E. A., & Usman, S. S. (2020). Predictive supervised machine learning models for diabetes mellitus. SN Computer Science, 1(5), 240. https://doi.org/10.1007/s42979-020-00250-8
33. Lai, H., Huang, H., Keshavjee, K., Guergachi, A., & Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders, 19, 1-9. https://doi.org/10.1186/s12902-019-0436-6
34. Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics, 24(1), 337. https://doi.org/10.1186/s12859-023-05465-z
35. Mustafa, M. (2023). A Comprehensive Dataset for Predicting Diabetes with Medical & Demographic Data. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
36. Morris, A., & Misra, H. (2002). Confusion matrix based posterior probabilities correction.
37. Allen, G. D., & Goldsby, D. (2014). Confusion theory and assessment. International Journal of Innovative Science, Engineering & Technology, 1(10), 436-443.
38. Tharwat, A. (2021). Classification assessment methods. Applied Computing and Informatics, 17(1), 168-192. https://doi.org/10.1016/j.aci.2018.08.003