Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2026;15(3):1445-50

Machine learning approaches for diabetes prediction: A comparative analysis of classification models

Fatma Hilal Yagin

Malatya Turgut Özal University, Faculty of Medicine, Department of Biostatistics, Malatya, Türkiye

Received 13 April 2026; Accepted 17 August 2026
Available online 31 August 2026 with doi: 10.5455/medscience.2026.04.075

Abstract

Early and accurate identification of individuals at risk for type 2 diabetes mellitus (T2DM) is a clinical priority given the global scale of the epidemic. Machine learning (ML) methods offer a data-driven complement to conventional screening approaches; however, systematic comparisons of multiple classifiers under identical experimental conditions remain valuable for guiding model selection in practice. This study evaluated four ML classification algorithms — Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Bagged Classification and Regression Trees (CART) — for T2DM prediction using the Pima Indians Diabetes Database (n = 768). Physiologically implausible zero values were treated as missing and median-imputed within each training fold, class weighting was applied to address the class imbalance between diabetic (34.9%) and non-diabetic (65.1%) participants, and hyperparameters were selected via grid search. Models were evaluated using 10-fold stratified cross-validation and reported as mean ± 95% confidence interval. Bagged CART achieved the highest overall accuracy (0.77 ± 0.04), sensitivity (0.78 ± 0.06), and F1-score (0.70 ± 0.05), while Naïve Bayes attained the highest specificity (0.82 ± 0.06). All models achieved comparable discrimination (AUC-ROC 0.81 – 0.84). SHAP-based feature importance analysis of the Bagged CART model identified plasma glucose concentration as the dominant predictor, followed by body mass index and age. These findings support the potential value of ensemble tree-based methods as a component of non-invasive diabetes risk-screening tools and highlight the continued relevance of glycaemic variables and anthropometric measures in T2DM risk stratification. As this analysis is retrospective and confined to a single, demographically restricted cohort, these results should be regarded as hypothesis-generating rather than evidence of clinical readiness, pending external validation.

Keywords: Type 2 diabetes mellitus, machine learning, classification, ensemble methods

Introduction

Type 2 diabetes mellitus (T2DM) constitutes one of the most pressing non-communicable disease burdens of the twenty-first century. The International Diabetes Federation estimates that approximately 537 million adults were living with diabetes in 2021, a figure projected to reach 783 million by 2045 [1]. The metabolic consequences of uncontrolled T2DM — including cardiovascular disease, chronic kidney disease, retinopathy, and lower-limb amputation — impose a substantial burden not only on affected individuals but also on healthcare systems globally. Crucially, T2DM is often asymptomatic in its early stages, and a substantial proportion of cases therefore remain undiagnosed at any given time [2]. This diagnostic gap underscores the importance of population-level screening and predictive modelling strategies capable of facilitating earlier clinical intervention.

Machine learning (ML) has emerged as a powerful methodological framework for clinical prediction tasks, offering the capacity to detect complex, non-linear relationships among predictor variables without requiring a priori specification of functional forms. A growing body of evidence has applied a wide range of ML algorithms to diabetes prediction, encompassing approaches from simple probabilistic classifiers to sophisticated ensemble methods [3,4]. A systematic review and meta-analysis by De Silva et al. [5] identified 40 ML prediction models for T2DM in community settings and reported a pooled c-index of 0.812, indicating generally good discriminative performance

CITATION
Yagin FH. Machine learning approaches for diabetes prediction: A comparative analysis of classification models. Med Science. 2026;15(3):1445-50.

Corresponding Author: Fatma Hilal Yagin, Malatya Turgut Özal University, Faculty of Medicine, Department of Biostatistics, Malatya, Türkiye
Email: hilal.yagin@gmail.com

DOI: 10.5455/medscience.2026.04.075

across methods. Despite this proliferation, related results across studies are difficult to synthesise owing to variation in datasets, preprocessing decisions, validation strategies, and the performance metrics reported. Direct, head-to-head comparisons of multiple classifiers under identical experimental conditions therefore remain informative, particularly when combined with feature importance analysis that can elucidate the relative clinical significance of predictor variables.

The Pima Indians Diabetes Database, originally compiled by the National Institute of Diabetes and Digestive and Kidney Diseases [6], has served as a widely used benchmark for evaluating diabetes prediction models. Although it reflects a specific demographic cohort, the dataset's well-characterised clinical variables and publicly accessible status make it a valuable testbed for methodological comparisons. Previous studies employing this dataset have reported varying levels of predictive performance. While Kavakiotis et al. 4 noted in a systematic review that the vast majority of ML classifiers applied to diabetes datasets achieved accuracy above 80%, performance on the Pima Indians Diabetes Database specifically tends to be more modest, reflecting its inherent class imbalance and well-documented data quality limitations.

The present study aimed to conduct a rigorous comparison of four ML classifiers — Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Bagged CART — for T2DM prediction using the Pima Indians Diabetes Database, with consistent preprocessing, stratified train-test splitting, and a comprehensive set of performance metrics. A secondary aim was to identify the most influential predictor variables in the best-performing model through feature importance analysis. Together, these analyses are intended to provide methodologically transparent and reproducible evidence to inform model selection in diabetes prediction research.

Material and Methods

Data Source

This study utilised the Pima Indians Diabetes Database, an open-access dataset originally compiled by the National Institute of Diabetes and Digestive and Kidney Diseases and made publicly available through the UCI Machine Learning Repository and Kaggle (https://www.kaggle.com/code/mragpavank/pima-indians-diabetes-database/input). Since secondary data analysis was performed using open-access data in this study, ethical committee approval is not required. The dataset consists of 768 female participants of Pima Indian heritage aged 21 years and above. All participants were screened for T2DM and classified as diabetic (n = 268; 34.9%) or non-diabetic (n = 500; 65.1%). Eight clinical and anthropometric predictor variables were recorded: number of pregnancies, plasma glucose concentration (2-hour oral glucose tolerance test), diastolic blood pressure, triceps skinfold thickness, serum insulin, body mass index (BMI), diabetes pedigree function, and age.

Statistical Analysis

Descriptive statistics were computed for each predictor variable, stratified by diabetes outcome. Normality of each variable within each group was assessed using D'Agostino and Pearson's omnibus test [7]. All eight variables in both outcome groups violated the assumption of normality ($p < 0.05$), with the sole exception of skin thickness in the non-diabetic group ($p = 0.586$). Given the predominantly non-normal distributions, data are presented as median [interquartile range (IQR)] for all variables. Between-group comparisons were performed using the two-sided Mann-Whitney U test [8], which is appropriate for independent, non-normally distributed continuous data. Statistical significance was defined at $\alpha = 0.05$. Effect sizes are reported as the rank-biserial correlation (r), where $|r| \approx 0.1$, 0.3, and 0.5 conventionally denote small, medium, and large effects, respectively. All analyses were conducted in Python using the SciPy and Pandas libraries.

Data Preprocessing

Physiologically implausible zero values in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI were treated as missing and imputed using median imputation, fitted exclusively on each training fold to prevent data leakage [9]. To address the class imbalance between diabetic (34.9%) and non-diabetic (65.1%) participants, class weights inversely proportional to class frequency were applied during model training (*class_weight='balanced'*) rather than synthetic oversampling [10,11], as this avoids introducing interpolated data points for physiologically constrained variables.

Classification Models

Hyperparameters for each classifier were selected via grid search with 5-fold stratified cross-validation on the training set, optimising for F1-score [12]. The final configurations were: Logistic Regression [13] (L2 penalty, C = 1.0, class-weight balanced); Random Forest [14] (300 trees, max depth = 10, max_features = 'log2', min_samples_leaf = 5, class-weight balanced); Bagged CART [15] (50 base decision trees, max depth = 5, min_samples_leaf = 5, class-weight balanced applied to the base estimator); Naïve Bayes [16] (var_smoothing = 1×10^{-11}).

Model Validation

Model performance was evaluated using 10-fold stratified cross-validation, preserving the original class distribution in each fold, rather than a single train-test split, to obtain stable performance estimates with 95% confidence intervals [17]. Reported metrics include accuracy, sensitivity, specificity, F1-score, area under the receiver operating characteristic curve (AUC-ROC), and area under the precision-recall curve (AUC-PR). Variable importance for the best-performing model was assessed using SHAP (SHapley Additive exPlanations) values [18], computed via TreeExplainer averaged across the base estimators of the Bagged CART ensemble on the held-out test set [19]. All analyses were conducted in Python 3.12, using scikit-learn 1.8.0, SciPy 1.17.1, pandas 3.0.2, NumPy 2.4.4, imbalanced-learn 0.14.2, and SHAP 0.52.0.

DOI: 10.5455/medscience.2026.04.075

Results

Descriptive Statistics and Between-Group Comparisons

Table 1 presents the demographic and clinical characteristics of participants stratified by diabetes status. Normality testing using D'Agostino and Pearson's omnibus test indicated non-normal distributions in at least one group for all variables, with the sole exception of skin thickness in the non-diabetic group (p = 0.586). Significant between-group differences were observed for seven of eight variables; serum insulin was the sole exception (p = 0.066). Diabetic participants exhibited markedly higher plasma glucose concentrations (140.0 (119.0 – 167.0) vs 107.0 (93.0 – 125.0) mg/dL), greater BMI (34.2 (30.8 – 38.8) vs 30.1 (25.4 – 35.3) kg/m²), and older age (36.0 (28.0 – 44.0) vs 27.0 (23.0 – 37.0) years) relative to non-diabetic participants. Number of pregnancies, blood pressure, skin thickness, and diabetes pedigree function were also significantly higher in the diabetic group (Table 1, and Figure 1).

Table 1. Baseline characteristics of participants by diabetes status (n = 768)

Variable	Non-diabetic (n = 500)	Diabetic (n = 268)	U statistic (×10³)	p-value	Effect size
Pregnancies (n)	2.0 (1.0 – 5.0)	4.0 (1.8 – 8.0)	50.98	< 0.001	0.24
Glucose (mg/dL)	107.0 (93.0 – 125.0)	140.0 (119.0 – 167.0)	28.39	< 0.001	0.58
Blood pressure (mmHg)	70.0 (62.0 – 78.0)	74.0 (66.0 – 82.0)	55.41	< 0.001	0.17
Skin thickness (mm)	21.0 (0.0 – 31.0)	27.0 (0.0 – 36.0)	59.81	0.013	0.11
Insulin (µU/mL)	39.0 (0.0 – 105.0)	0.0 (0.0 – 167.2)	61.92	0.066	-
BMI (kg/m²)	30.1 (25.4 – 35.3)	34.2 (30.8 – 38.8)	41.86	< 0.001	0.38
Diabetes pedigree function	0.3 (0.2 – 0.6)	0.4 (0.3 – 0.7)	52.76	< 0.001	0.21
Age (years)	27.0 (23.0 – 37.0)	36.0 (28.0 – 44.0)	41.95	< 0.001	0.37

Data are presented as median (interquartile range); Between-group comparisons were performed using the two-sided Mann-Whitney U test; p-values highlighted in red denote statistical significance (α = 0.05); BMI: body mass index

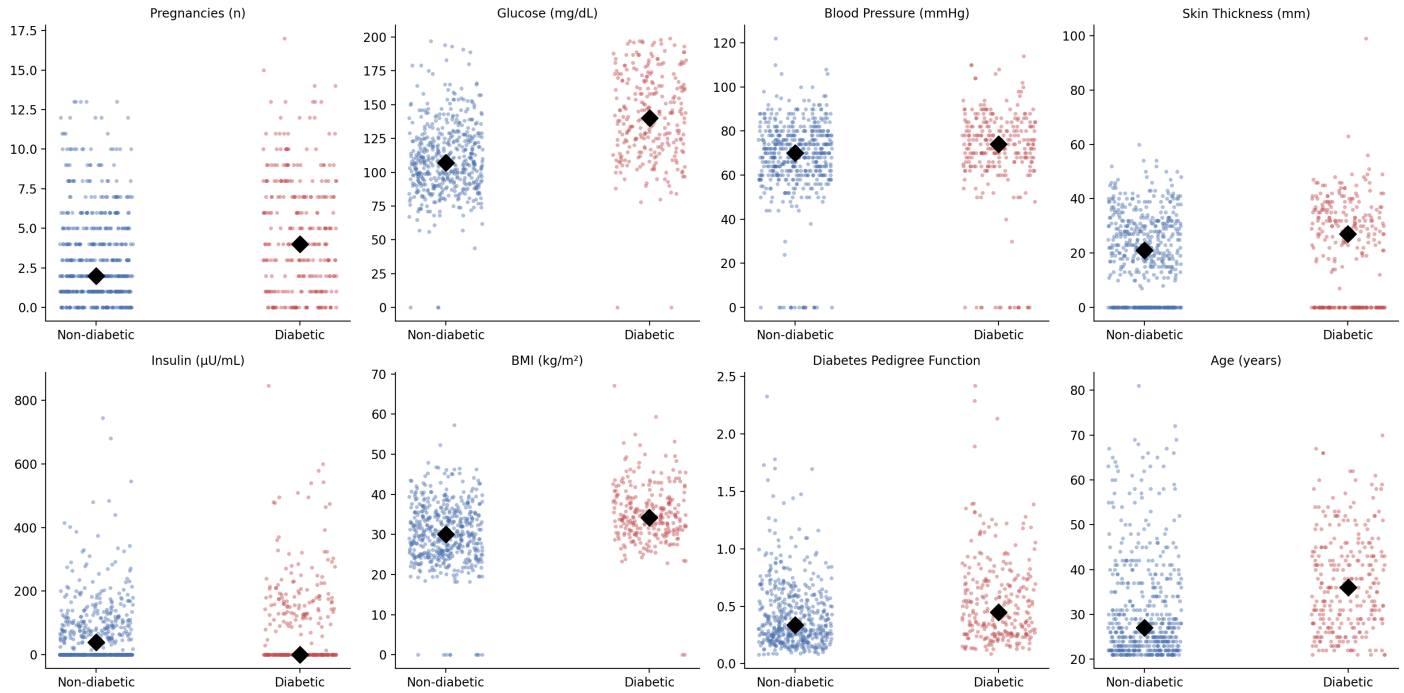

Figure 1. Distribution of each predictor variable by diabetes outcome; Individual data points are jittered horizontally for visibility; black diamonds indicate group medians

DOI: 10.5455/medscience.2026.04.075

Classification Model Performance

Table 2 presents the performance metrics of all four classifiers, evaluated using 10-fold stratified cross-validation with median imputation of physiologically implausible zero values and class-weighted training; hyperparameters were selected via grid search optimising for F1-score. Bagged CART achieved the highest overall accuracy (0.77 ± 0.04), sensitivity (0.78 ± 0.06), and F1-score (0.70 ± 0.05). Random Forest performed comparably (accuracy 0.76 ± 0.03, sensitivity 0.73 ± 0.06, F1-score 0.68 ± 0.05). Logistic Regression achieved similar sensitivity to Random Forest (0.72 ± 0.08) with an accuracy of 0.76 ± 0.05. Naïve Bayes achieved the highest specificity (0.82 ± 0.06) but the lowest sensitivity (0.60 ± 0.08) and F1-score (0.62 ± 0.07) among the four models. All models achieved comparable discrimination on AUC-ROC (0.81 – 0.84) and AUC-PR (0.69 – 0.74). Overall, Bagged CART provided the most favourable balance of accuracy, sensitivity, and F1-score and was selected as the best-performing model for diabetes classification in this dataset; unlike Naïve Bayes, it does not favour specificity at a substantial cost to sensitivity, which is a relevant consideration in a screening context where false negatives carry higher clinical cost.

Table 2. Classification performance (mean ± 95% CI) of four machine learning models

Model	Accuracy	Sensitivity	Specificity	F1-score	AUC-ROC	AUC-PR
Naive bayes	0.75 ± 0.05	0.60 ± 0.08	0.82 ± 0.06	0.62 ± 0.07	0.81 ± 0.04	0.69 ± 0.08
Logistic regression	0.76 ± 0.05	0.72 ± 0.08	0.78 ± 0.06	0.67 ± 0.06	0.84 ± 0.04	0.74 ± 0.07
Random forest	0.76 ± 0.03	0.73 ± 0.06	0.78 ± 0.04	0.68 ± 0.05	0.84 ± 0.03	0.73 ± 0.06
Bagged CART	0.77 ± 0.04	0.78 ± 0.06	0.76 ± 0.05	0.70 ± 0.05	0.84 ± 0.04	0.73 ± 0.06

Bagged CART: Bootstrap Aggregated Classification and Regression Trees, AUC-ROC: Area Under the Curve – Receiver Operating Characteristic, AUC-PR: Area Under the Curve – Precision-Recall

Model explainability results with SHAP

To enhance the interpretability of the best-performing model, SHAP analysis was conducted on the Bagged CART model following median imputation, and the resulting bee swarm plot is presented in Figure 2. Glucose concentration exerted the greatest influence on model predictions, exhibiting a pronounced bidirectional effect: high glucose values were associated with large positive SHAP values, indicating substantially increased predicted diabetes risk, whereas low glucose values corresponded to strongly negative SHAP values. This pattern is consistent with the central role of hyperglycaemia in the pathophysiology and diagnostic criteria of T2DM.

BMI emerged as the second most influential predictor, with elevated BMI values generating positive SHAP contributions, consistent with the established association between adiposity and insulin resistance. Age ranked third, with higher age values associated with increased predicted risk, in line with recognised epidemiological associations between advancing age and T2DM incidence. Diabetes Pedigree Function showed a smaller but consistent positive contribution for higher values.

Insulin, Skin Thickness, Number of Pregnancies, and Blood Pressure demonstrated narrow SHAP value distributions centred near zero, indicating limited and directionally inconsistent contributions to model output. In particular, Insulin, despite its physiological relevance to T2DM, ranked among the least influential predictors, likely reflecting the high proportion of missing/zero-recorded values in this variable and the resulting information loss even after imputation.

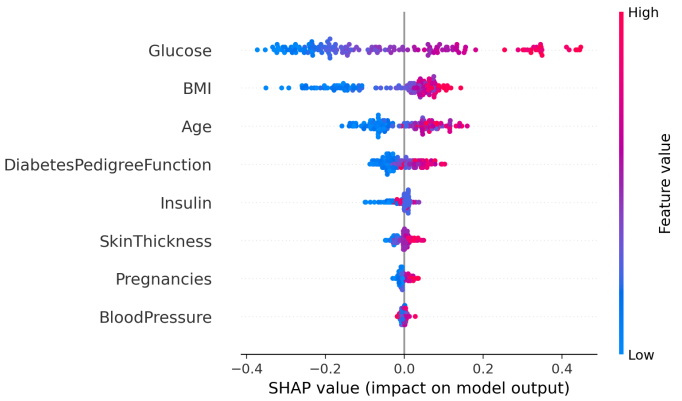

Figure 2. SHAP importance plot derived from the Bagged CART model

Discussion

This study systematically compared four ML classifiers — Naïve Bayes, Logistic Regression, Random Forest, and Bagged CART — for T2DM prediction using the Pima Indians Diabetes Database. Bagged CART emerged as the best-performing model, attaining a cross-validated accuracy of 0.77 ± 0.04, sensitivity of 0.78 ± 0.06, and F1-score of 0.70 ± 0.05. These results align with the growing evidence base supporting ensemble tree-based classifiers in clinical prediction tasks. Deberneh and Kim [20] similarly demonstrated that ensemble ML models achieved superior performance over single classifiers for T2DM prediction in a longitudinal Korean cohort, attributing this advantage to the variance reduction achieved through bootstrap aggregation. Uddin et al. [21] reported that Random Forest attained the highest

DOI: 10.5455/medscience.2026.04.075

classification accuracy (85.06%) among five ML algorithms applied to T2DM prediction from administrative claims data. The accuracy achieved by Bagged CART in the present study is broadly consistent with the pooled discriminative performance (c-index = 0.812) reported in a systematic review and meta-analysis of ML models for T2DM prediction in community settings [5].

The trade-off between sensitivity and specificity observed across classifiers warrants careful clinical interpretation. Naïve Bayes achieved the highest specificity (0.82 ± 0.06) but the lowest sensitivity (0.60 ± 0.08), whereas Bagged CART achieved the highest sensitivity (0.78 ± 0.06) at a moderately lower specificity (0.76 ± 0.05). In a clinical screening context, sensitivity is often accorded greater weight, as false negatives failing to identify a truly diabetic individual carry more serious clinical consequences than false positives. Kumar et al. [22] demonstrated, using the same Pima Indians Diabetes Database, that class imbalance substantially influences classifier performance and that class-rebalancing strategies can improve sensitivity; Nnamoko and Korkontzelos [23] reported comparable gains from selective oversampling. Consistent with these findings, applying class-weighted training in the present study (rather than the unweighted models used in earlier work on this dataset) substantially raised sensitivity in the tree-based models — most notably for Bagged CART, from approximately 0.61 without class weighting to 0.78 with class weighting — at a corresponding reduction in specificity, confirming that explicit handling of class imbalance is a meaningful and necessary methodological step rather than an optional refinement.

SHAP analysis identified Glucose as the overwhelmingly dominant predictor in the Bagged CART model, consistent with the established pathophysiology of T2DM. High glucose values were strongly associated with large positive SHAP values, indicating a substantial increase in predicted diabetes risk, whereas low glucose values produced strongly negative SHAP values. This bidirectional pattern underscores the central diagnostic role of hyperglycaemia, converging with findings by Seah et al. [24], who demonstrated that glycaemic variables alone achieved good discrimination (AUC ≥ 0.81) across multi-ethnic cohorts.

BMI ranked second in SHAP-based importance, with higher BMI values generating positive contributions, consistent with its well-established epidemiological association with T2DM risk [2]. Age ranked third, showing a predominantly positive contribution for higher values. Insulin, Skin Thickness, Number of Pregnancies, and Blood Pressure demonstrated narrow, near-zero SHAP distributions, indicating limited independent contribution to model output once Glucose, BMI, and Age were accounted for. The comparatively low importance of Insulin, despite its physiological centrality to T2DM, likely reflects the substantial proportion of missing/zero-recorded values for this variable in the dataset, which persists as an information limitation even after median imputation.

Several limitations of the present study should be acknowledged. First, the Pima Indians Diabetes Database is restricted to female participants of Pima Indian heritage, a population with a particularly high genetic predisposition to T2DM; the generalisability of these findings to broader demographic groups cannot be assumed without external validation on an independent, more heterogeneous cohort. Second, although zero values for glucose, BMI, blood pressure, skin thickness, and insulin were treated as missing and median-imputed in the present analysis, more sophisticated imputation strategies (e.g., multiple imputation, KNN-based imputation) may yield further improvement, particularly for insulin, which retained low predictive importance even after imputation. Third, class-weighted training rather than synthetic oversampling (e.g., SMOTE) was used to address class imbalance; while both approaches produced comparable gains in sensitivity in supplementary analyses, a systematic comparison of resampling strategies was beyond the scope of this study.

In light of these findings, Bagged CART may warrant further investigation as a candidate model for non-invasive T2DM risk stratification, particularly where specificity is prioritised, such as resource-limited or secondary referral contexts. For applications demanding higher Sensitivity, such as initial community-level screening, LR may represent an alternative worth exploring, given its competitive sensitivity and interpretable coefficient structure; the latter property is valued in regulatory and clinical decision-support contexts [25]. These interpretations are drawn from a single retrospective dataset without external or prospective validation; any statement of clinical suitability would require confirmation in independent, demographically diverse cohorts and prospective evaluation before these models could be considered for actual screening use [26].

Conclusion

This study demonstrates that Bagged CART achieves the most favourable overall performance profile for T2DM prediction on the Pima Indians Diabetes Database, outperforming Random Forest, Logistic Regression, and Naïve Bayes in terms of accuracy, sensitivity, and F1-score under a 10-fold cross-validated, class-weighted evaluation framework. Naïve Bayes achieved the highest specificity but at a substantial cost to sensitivity, highlighting a clinically meaningful trade-off between classifier approaches that should inform model selection based on screening context. SHAP-based feature importance analysis confirms that plasma glucose concentration, BMI, and age are the predominant predictors of T2DM status in this dataset, consistent with established clinical evidence. These results contribute to the growing evidence base supporting ensemble ML methods in diabetes prediction and underscore the importance of transparent, methodologically rigorous evaluations — including explicit treatment of class imbalance and missing data — to guide model selection and improve the real-world applicability of clinical prediction models.

Ethical Approval
Because our research was a secondary analysis of open-access data, no ethics committee permission was necessary.

Patient Consent
Since this study involves a secondary analysis of a publicly available (open access) and anonymized dataset, patient consent is not required.

Conflict of Interest
The authors confirm the absence of any conflicts associated with this work.

Funding
The authors declare that they have received no financial support for the study.

References

1. International Diabetes Federation. IDF Diabetes Atlas. 10th ed. International Diabetes Federation, Brussels, 2021.
2. Saeedi P, Petersohn I, Salpea P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes research and clinical practice. 2019;157:107843.
3. Abdulazeem H, Whitelaw S, Schauberger G, Klug SJ. A systematic review of clinical health conditions predicted by machine learning diagnostic and prognostic models trained or validated using real-world primary health care data. PLoS One. 2023;18:e0274276.
4. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017;15:104-16.
5. De Silva K, Lee WK, Forbes A, et al. Use and performance of machine learning models for type 2 diabetes prediction in community settings: a systematic review and meta-analysis. Int J Med Inform. 2020;143:104268.
6. Smith JW, Everhart JE, Dickson WC, et al. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. 12th Annual Symposium on Computer Application in Medical Care, 6-9 November 1988. Washington, DC, USA, 261-5.
7. D'Agostino RB, Pearson ES. Tests for departure from normality. Empirical results for the distributions of b^2 and $\sqrt{b}^1$. Biometrika. 1973;60:613-22.
8. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947;18:50-60.
9. Ali AA, Galal GR, Hassan HS. Diabetes prediction on Pima Indian dataset using machine learning techniques. International Journal of Environmental Sciences. 2025;11:529-50.
10. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21:1263-84.
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57.
12. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-30.
13. Bisong E. Logistic regression. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Apress, Berkeley, CA, 2019;243-50.
14. Breiman L. Random forests. Machine learning. 2001;45:5-32.
15. Zhang H. A novel student achievement prediction model based on bagging-CART machine learning algorithm. Concurrency and Computation: Practice and Experience. 2024;36:e8151.
16. Yang FJ. An implementation of naive Bayes classifier. 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 12–14 December 2018. Las Vegas, USA, 301-6.
17. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. 14th International Joint Conference on Artificial Intelligence (IJCAI-95), 20-25 August 1995. Montreal, Canada, 1137-45.
18. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), 4-9 December 2017. Long Beach, CA, USA, 4765-74.
19. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56-67.
20. Deberneh HM, Kim I. Prediction of type 2 diabetes based on machine learning algorithm. Int J Environ Res Public Health. 2021;18:3317.
21. Uddin S, Imam T, Hossain ME, et al. Intelligent type 2 diabetes risk prediction from administrative claim data. Informatics for Health and Social Care. 2022;47:243-57.
22. Kumar V, Lalotra GS, Sasikala P, et al. Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. Healthcare (Basel). 2022;10:1293.
23. Nnamoko N, Korkontzelos I. Efficient treatment of outliers and class imbalance for diabetes prediction. Artif Intell Med. 2020;104:101815.
24. Seah JYH, Yao J, Hong Y, et al. Risk prediction models for type 2 diabetes using either fasting plasma glucose or HbA1c in Chinese, Malay, and Indians: results from three multi-ethnic Singapore cohorts. Diabetes Res Clin Pract. 2023;203:110878.
25. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Wiley, Hoboken, NJ, 2013.
26. ElSeddawy AI, Karim FK, Hussein AM, Khafaga DS. Predictive analysis of diabetes-risk with class imbalance. Comput Intell Neurosci. 2022;2022:3078025.