Risk Prediction and Stratification of Diabetes Mellitus Using Data Mining Algorithms Among Patients with Chronic Diseases

Salma S. Suleiman[1*], Olawoyin O. Ishaq[2], Musa M. Uba[3], Umar A. Huguma[4], Nafisat Yusuf[5], Abubakar F. Kabara[6], Fatima U. Muhammad[7], Hamisu A. Hamisu[8]

[1,2,3,4,5,6,7,8]Department of Statistics, Aliko Dangote University of Science and Technology, Wudil, Kano, Nigeria.

*Correspondence: salmasagagi@gmail.com

Abstract

Diabetes mellitus is an increasing public-health burden in Nigeria, while early identification and risk prioritisation remain difficult in resource-constrained clinical settings. This study developed and evaluated data-mining algorithms for diabetes prediction among patients with chronic diseases attending Aminu Kano Teaching Hospital (AKTH), Kano State. A retrospective analytical design was adopted using 557 anonymised patient records containing demographic, anthropometric, laboratory, and chronic-disease indicators. Missing values were assessed and imputed, variables were transformed as required, and exploratory and inferential analyses were performed. Five supervised classifiers, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbours, were trained with stratified data partitioning and ten-fold cross-validation. Model performance was evaluated using accuracy, Cohen's kappa, sensitivity, specificity, precision, F1-score, and receiver operating characteristic area under the curve (ROC-AUC). Diabetes prevalence was 80.6% among 545 patients with known outcome status. Kidney disease (χ^2=37.722, p<0.001), family history (χ^2=5.752, p=0.016), BMI status (χ^2=8.867, p=0.031), and glucose status (χ^2=248.950, p<0.001) were significantly associated with diabetes, whereas hypertension and heart disease were not. Decision Tree achieved the highest accuracy (82.1%), while Naïve Bayes yielded the highest ROC-AUC (0.749) and sensitivity (94.7%), although specificity was low (12.9%). Random Forest importance ranked kidney disease, systolic blood pressure, family history, age, diastolic blood pressure, and BMI as the leading predictors. The Naïve Bayes risk framework classified 91.36% of test patients as high risk and placed 124 of 131 diabetic patients in that group. The findings show that locally trained data-mining models can support initial diabetes screening and prioritisation at AKTH; however, class imbalance, low specificity, calibration, and external validation must be addressed before clinical deployment.

Keywords: diabetes mellitus; data mining; machine learning; risk stratification; Naïve Bayes; Random Forest; Nigeria

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterised by persistent hyperglycaemia arising from impaired insulin secretion, insulin action, or both. Its complications impose substantial clinical and economic burdens through cardiovascular, renal, neurological, and other outcomes (Muhammad *et al*., 2020). The International Diabetes Federation and World Health Organisation continue to identify diabetes as a major non-communicable disease concern, particularly in low- and middle-income countries where diagnosis may occur late and routine screening resources are constrained (International Diabetes Federation, 2025; World Health Organisation, 2024).

Nigeria is experiencing a rising burden of diabetes associated with demographic change, urbanisation, dietary transition, physical inactivity, obesity, and longer survival with chronic disease (Uba *et al*., 2019). Evidence from Nigeria indicates substantial variation in prevalence and risk across settings, emphasising the need for locally grounded prediction tools rather than direct reliance on models developed in other populations (Uloko *et al*., 2018; Muhammad *et al*., 2020; Ukoba *et al*., 2025).

Machine-learning and data-mining methods can learn patterns from routinely collected clinical records and support targeted screening. Logistic Regression provides an interpretable statistical baseline, while Decision Trees, Random Forests, Naïve Bayes, and K-Nearest Neighbours offer different mechanisms for representing non-linearity, interactions, probability, and local similarity (Breiman, 2001; Han *et al*., 2012; James *et al*., 2021). Nevertheless, predictive performance can be misleading when the disease class is dominant; therefore, accuracy should be interpreted alongside sensitivity, specificity, F1-score, and ROC-AUC.

This study aimed to develop and compare five predictive algorithms for diabetes mellitus among patients with chronic diseases attending AKTH. It further examined associated risk factors, identified influential predictors, and developed probability-based risk strata for clinical prioritisation.

2. Materials and Methods

2.1. Study design, setting, and data source

A retrospective analytical design was used. Secondary, anonymised clinical data were obtained from patient records at Aminu Kano Teaching Hospital, a major tertiary referral facility serving Kano State and neighbouring states in northern Nigeria. The source dataset contained 557 patient records. Diabetes status was available for 545 records, while the independent test set used for model evaluation contained 162 observations.

2.2. Study variables

The binary outcome was diabetes mellitus status (diabetic or non-diabetic). Candidate predictors comprised age, sex, glucose and glucose status, weight, height, body mass index (BMI) and BMI category, systolic and diastolic blood pressure, hypertension, kidney disease, heart disease, and family history of diabetes.

2.3. Data preprocessing and exploratory analysis

Data quality assessment included standardisation of variable names and coding, duplicate checking, conversion of numeric and categorical fields, evaluation of implausible observations, and missing-value analysis. Missing values were concentrated in anthropometric and laboratory measurements. Appropriate imputation was applied before modelling. Descriptive statistics summarised continuous variables, while prevalence was estimated using frequency and percentage distributions.

2.4. Inferential analysis

Pearson chi-square tests of independence assessed associations between diabetes status and hypertension, kidney disease, heart disease, family history, BMI status, and glucose status. Statistical significance was assessed at α=0.05.

2.5. Model development and validation

Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and K-Nearest Neighbours models were fitted using the same preprocessed predictors. Stratification preserved the diabetic/non-diabetic distribution during data partitioning, and ten-fold cross-validation supported hyperparameter optimisation and model selection. Final evaluation was performed on independent test data.

2.6. Performance evaluation and risk categories

Performance was assessed using accuracy, Cohen's kappa, sensitivity, specificity, precision, F1-score, and ROC-AUC. Naïve Bayes was selected for risk stratification because it produced the highest ROC-AUC. Predicted probabilities below 0.30 were categorised as low risk, values from 0.30 to 0.69 as moderate risk, and values of at least 0.70 as high risk.

3. Results and Discussions

3.1. Missingness and clinical characteristics

Height, BMI, and BMI status had the most missing observations (136 each), followed by weight (130) and glucose/glucose status (87 each). After preprocessing, the analytic variables were complete. Participants had a mean age of 52.1±10.6 years, mean glucose of 8.77±1.87 mmol/L, mean BMI of 27.8±3.33 kg/m², mean systolic blood pressure of 137.0±14.7 mmHg, and mean diastolic blood pressure of 87.6±8.67 mmHg.

Table 1. Descriptive statistics of continuous variables after preprocessing

Variable	Mean	SD	Minimum	Median	Maximum
Age (years)	52.10	10.6	19.00	53.00	92.00
Glucose (mmol/L)	8.77	1.87	4.90	8.70	12.50
Weight (kg)	78.30	8.56	48.00	78.00	101.00
Height (m)	1.68	0.06	1.49	1.68	1.80
BMI (kg/m²)	27.80	3.33	17.80	27.80	38.00

Diastolic BP (mmHg)	87.60	8.67	60.00	88.00	120.00
Systolic BP (mmHg)	137.0	14.7	60.00	139.00	240.00

3.2. Diabetes prevalence

Among 545 patients with documented diabetes status, 439 (80.6%) were diabetic, and 106 (19.4%) were non-diabetic. This high prevalence is consistent with the tertiary chronic-disease setting, but it also creates class imbalance that can inflate accuracy and sensitivity while weakening specificity.

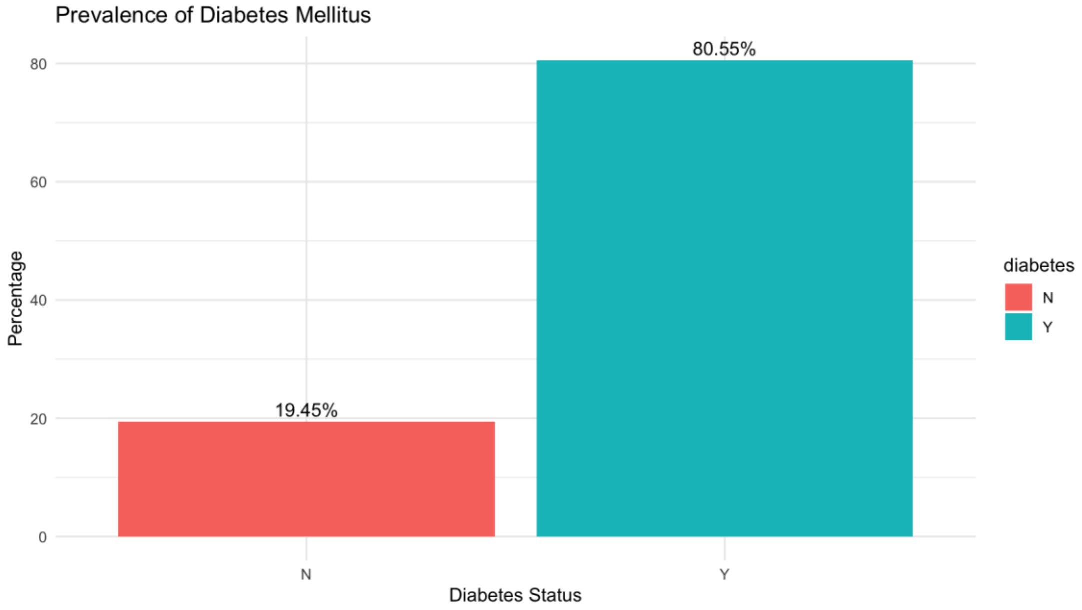

Figure 1. Prevalence of diabetes mellitus among AKTH patients

3.3. Associations with diabetes mellitus

Kidney disease, family history, BMI status, and glucose status were statistically associated with diabetes. Hypertension and heart disease were not significant in this clinical cohort (Table 2). Glucose status produced the strongest association and provides an internal consistency check because glycaemic status is directly related to diabetes diagnosis.

Table 2. Pearson chi-square tests of selected risk factors and diabetes mellitus

Risk factor	**χ^2**	**df**	**p-value**	**Decision at 5%**
Hypertension	1.496	1	0.221	Not significant
Kidney disease	37.722	1	<0.001	Significant
Heart disease	0.017	1	0.895	Not significant
Family history	5.752	1	0.016	Significant
BMI status	8.867	3	0.031	Significant
Glucose status	248.950	2	<0.001	Significant

3.4. Comparative model performance

The Decision Tree achieved the highest accuracy (0.821), kappa (0.245), sensitivity (0.962), precision (0.840), and F1-score (0.897). However, its ROC-AUC was 0.350, indicating poor probability ranking or a possible positive-class orientation issue. Naïve Bayes achieved the highest ROC-AUC (0.749), followed by Random Forest (0.704), and was therefore selected for risk stratification. All models showed low specificity, reflecting the strong outcome imbalance.

Table 3. Test-set performance of the five predictive models

Model	Accuracy	Kappa	Sensitivity	Specificity	Precision	F1	ROC-AUC
Logistic Regression	.790	.070	.954	.097	.817	.880	.679
Decision Tree	.821	.245	.962	.226	.840	.897	.350
Random Forest	.809	.218	.947	.226	.838	.889	.704
Naïve Bayes	.790	.100	.947	.129	.821	.879	.749
KNN	.784	.026	.954	.065	.812	.877	.642

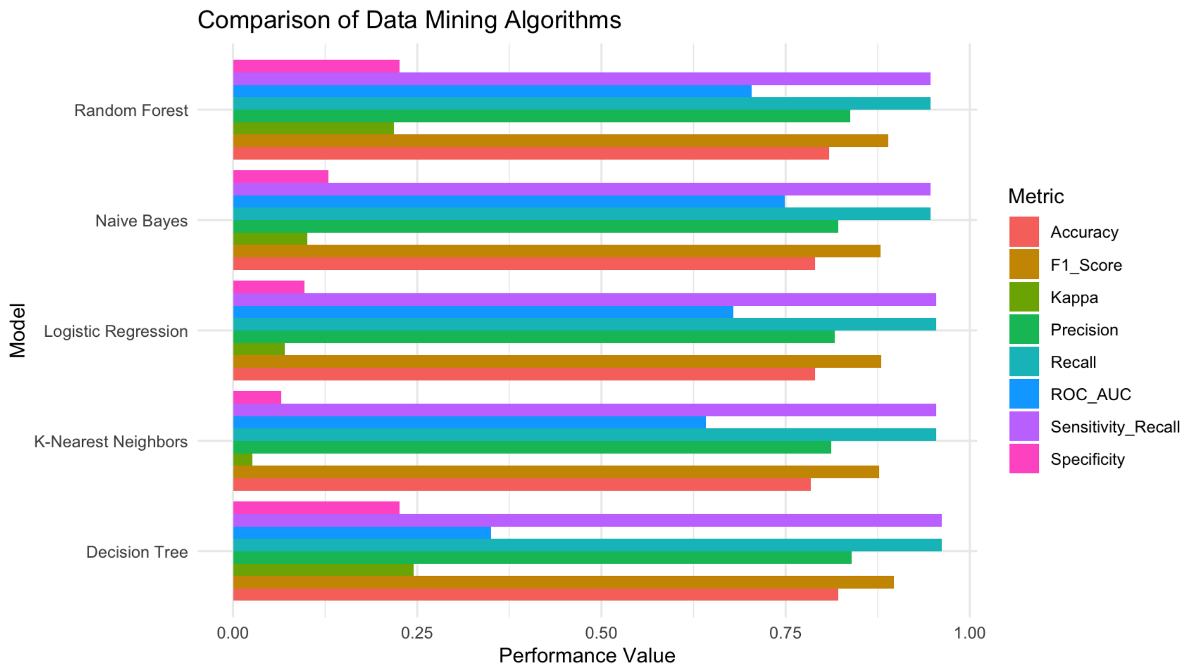

Figure 2. Comparative performance metrics of the predictive models

3.5. Naïve Bayes confusion matrix

Naïve Bayes correctly classified 124 diabetic patients and four non-diabetic patients. It missed seven diabetic patients and incorrectly flagged 27 non-diabetic patients. Thus, the model had strong case detection but weak exclusion of non-cases, making it more appropriate for preliminary screening than stand-alone diagnosis.

Table 4. Confusion matrix for the Naïve Bayes model

Predicted status	Actual diabetic	Actual non-diabetic	Total
Predicted diabetic	124 (TP)	27 (FP)	151
Predicted non-diabetic	7 (FN)	4 (TN)	11
Total	131	31	162

3.6. Variable importance

Random-Forest importance identified kidney disease as the strongest predictor (scaled score=100), followed by systolic blood pressure (51.15), family history (39.16), age (34.44), diastolic blood pressure (30.66), and BMI (22.55). Heart disease, hypertension, and sex contributed comparatively little in this cohort.

Table 5. Random-Forest variable importance

Rank	Predictor	Importance score
1	Kidney disease	100.00
2	Systolic blood pressure	51.15
3	Family history	39.16
4	Age	34.44
5	Diastolic blood pressure	30.66
6	Body mass index	22.55
7	Heart disease	8.08
8	Hypertension	0.82
9	Sex	0.00

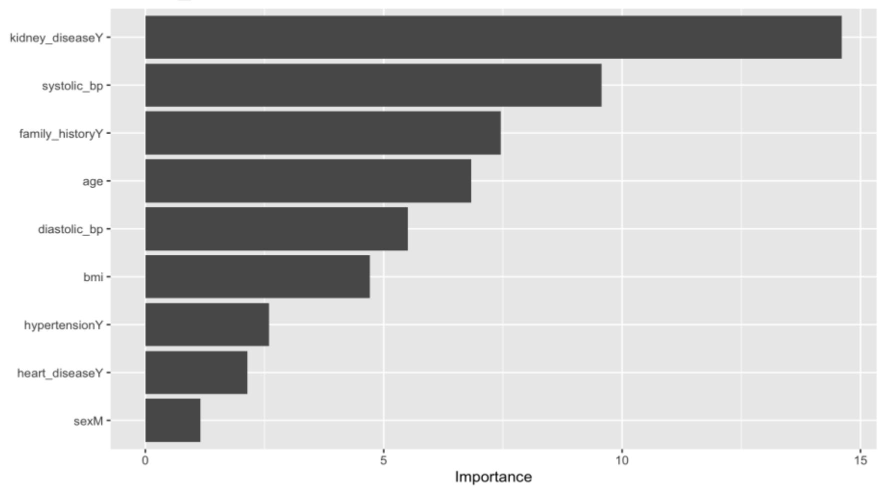

Figure 3. Random-Forest variable-importance plot

3.7. Risk stratification

Naïve Bayes categorised 148 (91.36%) test patients as high risk, six (3.70%) as moderate risk, and eight (4.94%) as low risk. Of 131 actual diabetic patients, 124 (94.7%) were placed in the high-risk group. However, the diabetic proportion was higher in the low-risk group than in the moderate-risk group, reflecting very small subgroup sizes and possible probability-calibration or threshold instability.

Table 6. Cross-classification of predicted risk groups and actual diabetes status

Risk group	Diabetic, n (%)	Non-diabetic, n (%)	Total
Low	5 (62.50)	3 (37.50)	8
Moderate	2 (33.33)	4 (66.67)	6
High	124 (83.78)	24 (16.22)	148
Total	131 (80.86)	31 (19.14)	162

3.8. Discussion

This study demonstrates the feasibility of using routine clinical data to support diabetes prediction in a Nigerian tertiary hospital. The observed prevalence of 80.6% is much higher than population-based estimates because the study population comprised patients attending a tertiary facility for chronic-disease management. Consequently, this figure should not be generalised as a community prevalence estimate. The same selection mechanism explains the pronounced class imbalance and is central to interpreting model performance.

The significant relationships involving kidney disease, family history, BMI status, and glucose status are clinically plausible. Diabetes and renal impairment are closely interconnected, while familial predisposition and excess adiposity are established risk indicators. The absence of significant associations for hypertension and heart disease does not establish that these factors are clinically irrelevant; rather, it may reflect the case mix, high comorbidity prevalence, shared risk pathways, coding quality, and limited between-group contrast in this hospital cohort.

Model comparison illustrates why accuracy alone is insufficient. The Decision Tree obtained the highest accuracy, yet its ROC-AUC of 0.350 suggests that its probability outputs did not rank patients reliably or that the positive-class direction may have been reversed during ROC calculation. This result should be rechecked before publication. By contrast, Naïve Bayes achieved the best discrimination and high sensitivity, but its specificity was only 12.9%. A screening tool may reasonably prioritise sensitivity to minimise missed cases, but very low specificity generates many false alarms, increases confirmatory testing, and limits stand-alone clinical utility.

The dominance of kidney disease in Random-Forest importance is noteworthy but should not be interpreted causally. Variable importance is affected by predictor distributions, correlation, coding, and model structure. Moreover, kidney disease may partly represent a diabetes complication rather than a pre-diagnostic risk factor. For a genuinely prospective early-warning system, predictor timing must be clearly defined, and features recorded after diabetes onset should be excluded to avoid target leakage.

The risk framework captured most diabetic patients in the high-risk category, supporting triage utility. Nevertheless, 91.36% of all test observations were labelled high risk, and the ordering of diabetic proportions across low and moderate groups was not monotonic. These findings indicate that calibration, threshold selection, and decision-curve analysis are needed before deployment. External validation across hospitals and geographic regions is also essential.

4. Conclusion

Data-mining algorithms can support diabetes screening and patient prioritisation using routine AKTH clinical data. Naïve Bayes provided the strongest ROC-AUC and high sensitivity, while Random Forest identified kidney disease, blood pressure, family history, age, and BMI as influential predictors. The proposed risk framework concentrated most diabetic patients in the high-risk group. However, the model's low specificity, strong class imbalance, potential probability-calibration limitations, and single-centre retrospective design mean that it should complement clinical assessment and laboratory confirmation rather than replace them.

References

American Diabetes Association. (2024). Standards of care in diabetes—2024. *Diabetes Care, 47*(Suppl. 1), S1–S350.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Ewwiekpaefe, A. E., & Abdulkadir, N. (2023). A predictive model for diabetes mellitus using machine learning techniques: A study in Nigeria. *African Journal of Information Systems, 15*(1), 1–21.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

International Diabetes Federation. (2025). *IDF diabetes atlas* (11th ed.). International Diabetes Federation.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

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.

Muhammad, L. J., Al-Shourbaji, I., Haruna, A. A., Mohammed, I. A., Ahmad, A., & Jibrin, M. B. (2020). Predictive supervised machine learning models for diabetes mellitus. *SN Computer Science, 1*(5), 240. https://doi.org/10.1007/s42979-020-00250-8

Odukoya, O., Nwaneri, S., Odeniyi, I., Akodu, B., Oluwole, E., Olorunfemi, G., Popoola, O., & Osuntoki, A. (2022). Development and comparison of three data models for predicting diabetes mellitus using risk factors in a Nigerian population. *Healthcare Informatics Research, 28*(1), 58–67. https://doi.org/10.4258/hir.2022.28.1.58

Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science, 82*, 115–121.

Uba, M. M., Jiadong, R., Sohail, M. N., Irshad, M., & Yu, K. (2019). Data mining process for predicting diabetes mellitus based model about other chronic diseases: A case study of northwestern Nigeria. *Healthcare Technology Letters, 6*(4), 98–102. https://doi.org/10.1049/htl.2018.5111

Ukoba, O., Joseph, U. O., Charles, O. L., Opirite, P.-K. B., Omamuyovwi, A. E., & Matilda, O.-N. (2025). A systematic review of machine learning methods for diabetes mellitus prediction and classification in Nigeria. *International Journal of Community Medicine and Public Health, 12*(6). https://doi.org/10.18203/2394-6040.ijcmph20251734

Uloko, A. E., Musa, B. M., Ramalan, M. A., Gezawa, I. D., Puepet, F. H., Uloko, A. T., Sada, K. B., & Borodo, M. M. (2018). Prevalence and risk factors for diabetes mellitus in Nigeria: A systematic review and meta-analysis. *Diabetes Therapy, 9*(3), 1307–1316.

World Health Organisation. (2024). *Diabetes*. World Health Organisation.