International Journal of Data Science and Analysis
2026, Vol. 12, No. 3, pp. 52–63
https://doi.org/10.11648/j.ijdsa.20261203.12

SciencePG
Science Publishing Group

Research Article

Random Forest Classification for Type 2 Diabetes Risk Prediction in Kenya: Evidence from the 2022 Kenya Demographic and Health Survey

Diana Boro*

Department of Pure and Applied Sciences, Kirinyaga University, Kutus, Kenya

Abstract

Type 2 diabetes mellitus (T2DM) is an escalating public health burden in sub-Saharan Africa, where over half of affected individuals remain undiagnosed at clinical presentation. In Kenya, early risk identification is constrained by limited population-level screening infrastructure and the underutilization of data-driven predictive tools. This study applied a Random Forest (RF) machine learning classifier to the nationally representative 2022 Kenya Demographic and Health Survey (KDHS 2022) dataset to develop and evaluate a predictive model for T2DM risk among Kenyan adults aged 15–54 years. The analytical sample comprised 31,354 respondents drawn from all 47 counties of Kenya, of whom 272 (0.87%) were classified as diabetic, reflecting a severe class imbalance ratio of 114.3:1. Data preprocessing encompassed median imputation for missing values, min-max normalisation of continuous predictors, one-hot encoding of categorical variables, and Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection. The Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training partition to address class imbalance without contaminating test-set evaluation. The RF model was optimised via exhaustive grid search with five-fold stratified cross-validation, yielding a cross-validation Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.9895 (standard deviation (SD = 0.0004). At a calibrated classification threshold of 0.20, the model achieved a test-set sensitivity of 62.96%, specificity of 69.81%, balanced accuracy of 66.40%, and Matthews Correlation Coefficient (MCC) of 0.0659, correctly identifying 34 of 54 diabetic cases in the held-out test set. SHAP (SHapley Additive exPlanations) analysis identified age, wealth index, hypertension, employment status, and Body Mass Index (BMI) as the dominant predictors of T2DM risk. These findings establish a reproducible, nationally representative RF-based screening framework with direct implications for targeted public health intervention and early detection policy in Kenya.

Keywords

Type 2 Diabetes Mellitus, Random Forest, Machine Learning, KDHS 2022, Kenya, Class Imbalance, SHAP, Disease Prediction

*Correspondence: Diana Boro (borodiana0@gmail.com)

Received: 29 June 2026; **Accepted:** 9 July 2026; **Published:** 24 July 2026

1. Introduction

Type 2 diabetes mellitus (T2DM) is a chronic metabolic disorder characterized by progressive insulin resistance and relative insulin secretory insufficiency, resulting in persistent hyper glycaemia and long-term multi-organ complications [1]. Unlike Type 1 diabetes, which arises from autoimmune destruction of pancreatic β-cells, T2DM is strongly associated with modifiable lifestyle and environmental risk factors, including obesity, physical inactivity, unhealthy dietary patterns, and advancing age [2]. The global burden of T2DM has reached crisis proportions: the International Diabetes Federation estimated that 537 million adults aged 20–79 years were living with diabetes in 2021, with projections indicating this figure will rise to 783 million by 2045 [3]. This trajectory positions T2DM as a leading contributor to premature mortality, long-term disability, and escalating healthcare expenditure worldwide.

In sub-Saharan Africa (SSA), the burden is intensifying at a faster rate than in any other world region. Approximately 24 million people in SSA were living with diabetes in 2021, with prevalence projected to nearly double by 2045, driven by rapid urbanization, sedentary lifestyles, energy-dense dietary transitions, and demographic ageing [3]. A defining and clinically critical feature of the SSA diabetes landscape is the high rate of undiagnosed: nearly two-thirds of affected individuals in the region remain unaware of their condition, precluding timely intervention and permitting the silent progression of cardiovascular, renal, and neuropathic complications [3]. This diagnostic gap renders population-level risk prediction tools not merely useful but essential for health systems operating under severe resource constraints.

In Kenya specifically, T2DM has emerged as a significant and growing contributor to the non-communicable disease (NCD) burden. The Kenya STEPwise Survey for Non-Communicable Disease Risk Factors reported that 2.4% of persons aged 18–69 years had raised fasting blood glucose indicative of diabetes, while a further 4.8% had impaired fasting glycaemia — a pre-diabetic state associated with high progression risk [4]. The 2022 Kenya Demographic and Health Survey (KDHS 2022) further documented high population-level prevalence of established T2DM risk factors, including overweight and obesity affecting 30.5% of women and widespread physical inactivity [5]. Driven by rapid urbanisation, dietary transitions, and population ageing, Kenya's T2DM burden is projected to intensify substantially in the coming decades [6], imposing growing human and economic costs on individuals, families, and an already-strained national health system [7]. Given the long asymptomatic phase preceding clinical T2DM diagnosis, early identification of high-risk individuals through robust predictive tools is essential, as evidence-based interventions have been shown to significantly delay or prevent disease onset [8].

Conventional statistical approaches to T2DM risk prediction — including logistic regression, Cox proportional hazards models, and linear discriminant analysis (LDA) — each carry well-documented limitations in the context of complex, population-level survey data. Logistic regression assumes linear relationships between predictors and outcomes, restricting its ability to capture nonlinear associations and higher-order interactions without manual pre-specification [9]. Cox models require proportional hazards assumptions that cannot be satisfied in cross-sectional surveys where temporal disease progression is unobservable [10]. LDA imposes normality and equal-covariance assumptiosns routinely violated in heterogeneous, high-dimensional survey data [11]. All three approaches are additionally sensitive to missing data, multicollinearity, and class imbalance — conditions inherent to nationally representative population surveys such as the KDHS 2022. A systematic review of diabetes prediction models in African-descent populations reported Area Under Curve (AUC) values for logistic regression-based models ranging from 0.65–0.79, insufficient for reliable individual-level risk stratification in heterogeneous populations [12].

Machine learning (ML) classifiers have emerged as a compelling alternative, demonstrating capacity to detect complex nonlinear predictor relationships and handle high-dimensional data without rigid distributional assumptions. Among ML classifiers, Random Forest (RF) has shown particularly strong and consistent predictive performance in health data applications, with reported AUC values of 0.83–0.86 in T2DM prediction tasks [13]. RF builds multiple decision trees on bootstrapped data subsets and aggregates their outputs via majority voting — a strategy rooted in the statistical bagging framework — yielding robustness to overfitting, resilience to missing data, and reliable variable importance rankings [14]. Despite this strong evidence base, RF and comparable ML methods have been scarcely applied to nationally representative Kenyan population data, creating a clear evidence gap for contextually grounded T2DM risk prediction in Kenya.

The most closely related existing Kenyan study applied ML classifiers to the KDHS dataset but was restricted to a single coastal county whose sociodemographic profile differs substantially from the national population [7], rendering its findings non-generalizable at national scale. To the researcher's knowledge, no published study has applied RF with SHAP-based interpretability analysis to the full, nationally representative KDHS 2022 dataset for T2DM risk prediction in Kenya. This study addresses this gap by developing and evaluating a Random Forest classifier on the KDHS 2022, employing SMOTE-based class imbalance treatment, MCC-optimized threshold calibration, and SHAP analysis to produce a robust, interpretable, and nationally generalizable T2DM risk prediction model for Kenya.

2. Literature Review

2.1. Random Forest in Type 2 Diabetes Prediction

Random Forest (RF), introduced by Breiman [14], is an ensemble learning method that constructs multiple decision trees on bootstrapped subsets of training data and aggregates their outputs through majority voting — a strategy formalized in the statistical bagging framework [15] to reduce variance without increasing bias. The theoretical guarantees underlying RF, including convergence properties and generalization bounds, are grounded in the law of large numbers and PAC (Probably Approximately Correct) learning theory [16]. Its robustness to overfitting, capacity to manage high-dimensional data, and reliable variable importance rankings have established RF as a widely adopted classifier in clinical and epidemiological prediction tasks.

Jafari et al. [17] demonstrated that RF outperformed Artificial Neural Networks (ANN) and Support Vector Machines (SVM) in settings characterized by nonlinear predictor relationships, while ANN captured more complex interaction patterns. Their conclusion — that model selection should be guided by data characteristics and interpretability requirements — is directly relevant to the present study's application of RF to a nationally representative survey dataset. Abdelhady et al. [18] applied RF to electronic health records, finding that it effectively ranked feature importance and balanced predictive power with interpretability, while XGBoost achieved marginally higher AUC values. Importantly, RF's transparent feature importance rankings remained clinically reliable even when its discriminative performance was matched by gradient boosting alternatives.

Rodriguez et al. [19] applied RF alongside K-Nearest Neighbors and Decision Trees on European cohort data, confirming RF's superior generalization across test sets. Garcia et al. [20] similarly demonstrated RF's advantage over logistic regression and Decision Trees on Latin American populations. Santos-Silva et al. [21] applied RF in a longitudinal blood glucose study, confirming that RF provided interpretable feature rankings allowing clinicians to identify critical contributors to glucose fluctuations — reinforcing its value in settings where interpretability is as important as predictive accuracy. Malathi and Suhana [22] further confirmed RF's consistent outperformance of logistic regression on the Indian Primary Diabetes Dataset, demonstrating robustness across diverse clinical populations.

A consistent limitation across these RF studies is their reliance on non-African populations or benchmark datasets — most notably the PIMA Indian Diabetes Dataset — which do not reflect the sociodemographic and epidemiological characteristics of sub-Saharan African populations [12]. Furthermore, while RF feature importance metrics based on Gini impurity are useful for variable ranking, they are known to be biased toward high-cardinality predictors and do not provide individual-level prediction explanations [23] — a limitation addressed in the present study through SHAP analysis.

2.2. Machine Learning for Diabetes Prediction in African Contexts

Within the African context, Nimmagadda et al. [8] developed RF, XGBoost, SVM, and logistic regression models using Kenyan medical data, finding that ensemble methods achieved the best AUC values and demonstrating that boosting and bagging techniques perform effectively on African datasets. However, a critical limitation must be noted: that study used clinical hospital-based data reflecting only care-seeking individuals already within the healthcare system, introducing systematic selection bias and excluding the large proportion of Kenyans who are undiagnosed or lack healthcare access — particularly in rural and lower-wealth communities. Mwakondo et al. [7] similarly applied ML classifiers to Kenyan electronic health records but without SHAP-based interpretability analysis or nationally representative sampling.

The present study addresses these gaps by applying RF to the KDHS 2022 — the most recent and most geographically comprehensive nationally representative survey available for Kenya — thereby extending RF-based T2DM prediction to a true population-level context covering all 47 counties. The incorporation of SHAP analysis further distinguishes this study from prior Kenyan work by providing individual-level predictor attribution absent in existing studies.

2.3. Research Gap

Despite the growing T2DM burden in Kenya — where 2.4% of persons aged 18–69 years have raised fasting blood glucose and approximately 54% of those affected in SSA remain undiagnosed [3, 4] — no published study has applied Random Forest with SHAP-based interpretability to a nationally representative Kenyan population dataset. Existing Kenyan studies used clinical electronic health records biased toward care-seeking populations that cannot generate predictions generalizable across Kenya's 47 counties [7, 8]. Studies applying RF to nationally representative survey data are largely confined to non-African populations and lack interpretability analysis [13, 21]. Furthermore, inconsistent evaluation frameworks across the reviewed literature — with most studies reporting only accuracy or AUC-ROC in isolation — limit meaningful model comparison [24]. This study fills these gaps by applying RF to the KDHS 2022 dataset with a comprehensive, standardized evaluation framework incorporating imbalance-appropriate metrics and SHAP-based interpretability.

3. Materials and Methods

3.1. Research Design

This study adopted a quantitative, cross-sectional research design based on secondary data analysis. The analytical workflow encompassed data preparation, model development, hyper parameter optimization, performance evaluation, and model interpretability assessment. A supervised machine learning classification approach was employed, with Random Forest as the primary classifier. The design ensured rigorous, reproducible model assessment with evaluation metrics selected to reflect the severe class imbalance inherent in the KDHS 2022 diabetes outcome.

3.2. Dataset and Study Population

The sole analytical dataset for this study was the Kenya Demographic and Health Survey 2022 (KDHS 2022), conducted by the Kenya National Bureau of Statistics (KNBS) in collaboration with ICF International and funded by the United States Agency for International Development (USAID) [5]. The KDHS 2022 is a nationally representative cross-sectional survey covering all 47 counties of Kenya, with data collected from 17,646 households using a multistage stratified cluster sampling design. The study population comprised Kenyan adults aged 15–54 years: 16,901 women (53.9%) and 14,453 men (46.1%), yielding a combined analytical sample of 31,354 respondents spanning all 47 counties.

The outcome variable was binary: 1 = Diabetic (respondent ever informed by a healthcare provider of high blood sugar or diabetes); 0 = Not Diabetic. Of 31,354 respondents, 272 (0.87%) were classified as Diabetic and 31,082 (99.13%) as Not Diabetic, yielding a class imbalance ratio of 114.3:1. This prevalence reflects the population-level self-reported T2DM diagnosis rate in Kenya, consistent with evidence that approximately 54% of those with diabetes in SSA remain undiagnosed [3].

3.3. Predictor Variables

Eleven predictor variables were selected from the KDHS 2022, all established T2DM risk factors in the literature [2, 3]: (1) age (continuous, years); (2) body mass index — BMI (continuous, kg/m^2); (3) hypertension (binary); (4) residence type (binary: urban/rural); (5) household wealth index (ordinal, 1–5); (6) education level (ordinal, 0–3); (7) tobacco use (binary); (8) alcohol use (binary); (9) employment status (binary); and (10) sex (binary). Continuous variables (age, wealth index, BMI) were normalised to [0,1] using min-max scaling following imputation, according to:

$$x' = (x - x\_min) / (x\_max - x\_min) \quad (1)$$

Categorical variables were encoded using one-hot encoding. This transformation ensures no predictor disproportionately influences model learning due to scale differences [24].

3.4. Data Preprocessing

3.4.1. Missing Data Treatment

Missing values in continuous variables were handled using median imputation via Simple Imputer, applied prior to scaling. BMI was missing for all male respondents in the combined dataset and was imputed using the median BMI value. Categorical variables with missing data were imputed using mode imputation. Variables with greater than 30% missingness were excluded from analysis, consistent with recommended practice for high-missingness predictors [25].

3.4.2. Feature Selection

Feature selection followed a two-stage approach. First, univariate chi-square tests and point-biserial correlations were applied to screen for significant predictor-outcome associations ($p < 0.05$). Second, Recursive Feature Elimination with Cross-Validation (RFECV) was applied to the SMOTE-balanced training dataset using a Random Forest base estimator to identify the optimal predictor subset maximising performance while minimizing complexity [26]. All ten candidate predictor variables were retained by RFECV, confirming that no variable was redundant and that removal of any predictor would reduce predictive performance.

3.4.3. Class Imbalance Treatment

The severe 114.3:1 class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE), which generates artificial minority class observations by interpolating among k-nearest neighbors of existing minority class cases in the feature space, according to:

$$x\_synthetic = x\_i + \lambda \times (x\_\{ir\} - x\_i), \text{ where } \lambda \sim U[0,1] \quad (2)$$

SMOTE was applied exclusively within the training partition (k_neighbors = 5), strictly confined to prevent data leakage. The held-out test set retained the original class distribution to ensure performance metrics reflect true out-of-sample model behavior under realistic deployment conditions. Following SMOTE, the training set comprised 49,730 observations split exactly 50:50 between diabetic and non-diabetic classes.

3.5. Random Forest Model

Random Forest is an ensemble learning method that constructs M decision trees on bootstrapped subsets of training data and aggregates their outputs through majority voting for classification [14]. For instance, the RF prediction was expressed as:

$$\hat{y} = f(x) = \Sigma(m=1 \text{ to } M)\ w\_m \times h\_m(x) \quad (3)$$

where h_m(x) is the m-th decision tree, w_m is the weight assigned to each tree (equal weighting in standard RF), and M is the total number of trees. At each split, only a random subset of √p features (where p is the total number of predictors) is considered, ensuring tree diversity and preventing any single highly predictive variable from dominating the ensemble.

The RF classifier was fitted to the SMOTE-balanced training set. Hyper parameter optimization was conducted via exhaustive grid search over predefined parameter spaces, evaluated using 5-fold stratified cross-validation with AUC-ROC as the optimization criterion. The optimal configuration comprised 300 trees (n_estimators = 300), maximum tree depth of 10, minimum samples per leaf of 3, and square root feature sampling at each split (max_features = 'sqrt').

3.6. Classification Threshold Calibration

At the default classification threshold of 0.50, the RF model identified only 3 of 54 true diabetic cases (sensitivity = 5.6%), confirming that the default threshold is wholly inappropriate for severely imbalanced data. The classification threshold was calibrated by sweeping values from 0.05 to 0.80. At each threshold, the F2-score was computed to evaluate model performance under a recall-prioritized setting, where minimizing false negatives is clinically critical:

$$F2 = (5 \times \text{Precision} \times \text{Recall}) / (4 \times \text{Precision} + \text{Recall}) \quad (4)$$

The threshold of 0.20 maximized the F2-score, raising true positives from 3 to 34 and sensitivity from 5.6% to 63.0%, and was selected as the optimal decision threshold for the RF model.

3.7. Evaluation Metrics

Given the severe class imbalance in the KDHS 2022 diabetes dataset (114.3:1), standard accuracy was not used as the primary evaluation metric, as it can be inflated when the majority class dominates. The following metrics were prioritized:

Precision-Recall Area Under the Curve (PR-AUC) was the primary imbalance-sensitive metric, computed as the area under the Precision-Recall curve across all classification thresholds. PR-AUC focuses exclusively on minority class performance and is less influenced by the large number of true negatives, with a random classifier baseline equal to the minority class prevalence (0.87%) [27].

Matthews Correlation Coefficient (MCC) incorporates all four confusion matrix elements and provides a balanced single-value assessment of classification performance regardless of class distribution, ranging from −1 (total disagreement) to +1 (perfect classification):

$$MCC = (TP \times TN - FP \times FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]} \quad (5)$$

Balanced Accuracy assigns equal importance to minority and majority class performance:

$$\text{Balanced Accuracy} = (\text{Sensitivity} + \text{Specificity}) / 2 \quad (6)$$

AUC-ROC was retained for comparability with existing literature but interpreted with caution due to its susceptibility to inflation under severe class imbalance through the large number of true negatives. Recall (sensitivity), specificity, precision, and F1-score were reported as supplementary metrics. SHAP (SHapley Additive exPlanations) analysis was applied to quantify each feature's contribution to model predictions using TreeSHAP, grounded in Shapley values from cooperative game theory [28].

3.8. Software and Tools

All analyses were conducted using Python 3.12.0. Key libraries included: scikit-learn 1.5.2 (RF implementation, RFECV, cross-validation, metrics); imbalanced-learn 0.12.3 (SMOTE); SHAP 0.46.0 (TreeSHAP interpretability); pandas 2.2.2 and NumPy 2.0.1 (data manipulation); matplotlib 3.9.2 and seaborn 0.13.2 (visualization); and stats models 0.14.2 (univariate screening). Development was conducted in Visual Studio Code 1.93.

3.9. Ethical Considerations

This study relied exclusively on publicly available, de-identified secondary data and did not involve direct interaction with human participants. The KDHS 2022 dataset was obtained upon formal request and approval from the DHS Program for a registered research project. The dataset was originally collected under ethical oversight by the Kenya National Bureau of Statistics and ICF International, with ethical approval granted by the Kenya Medical Research Institute (KEMRI) Ethics Review Committee and the ICF Institutional Review Board. As this study involved secondary analysis of fully anonymized data, no additional independent ethical approval was required.

4. Results

4.1. Descriptive Statistics

The combined analytical sample comprised 31,354 respondents: 16,901 women (53.9%) and 14,453 men (46.1%), spanning all 47 counties of Kenya. Of all respondents, 272 were classified as Diabetic (0.87%) and 31,082 as Not Diabetic (99.13%), yielding a class imbalance ratio of 114.3:1. Diabetes prevalence was 0.97% among men and 0.78%

among women. Table 1 presents descriptive statistics stratified by diabetes outcome status.

Table 1. *Descriptive Statistics by Diabetes Status — KDHS 2022 Dataset (N = 31,354).*

Variable	Statistic	Diabetic (n=272)	Not Diabetic (n=31,082)	p-value
Age (years)	Mean (SD)	38.8 (12.1)	29.4 (11.9)	< 0.001
BMI (kg/m^2)	Mean (SD)	27.3 (6.4)	23.8 (5.3)	< 0.001
Hypertension	Yes, n (%)	95 (34.9%)	1,708 (5.5%)	< 0.001
Urban Residence	Urban, n (%)	120 (44.1%)	11,847 (37.4%)	0.018
Sex: Male	n (%)	140 (51.5%)	14,313 (46.1%)	0.067
Tobacco Use	Yes, n (%)	14 (5.1%)	1,852 (6.0%)	0.491
Alcohol Use	Yes, n (%)	40 (14.7%)	4,095 (13.2%)	0.488
Education: Secondary+	n (%)	168 (61.8%)	19,407 (62.4%)	0.832
Currently Employed	Yes, n (%)	162 (59.6%)	19,894 (64.0%)	0.097

Note. SD = standard deviation. p-values derived from independent samples t-test (continuous) and chi-square test (categorical). Variables with $p \geq 0.05$ do not differ significantly between groups in univariate analysis but may contribute meaningfully to prediction in the multivariate Random Forest model.

Diabetic respondents were on average 9.4 years older than non-diabetic respondents (38.8 vs. 29.4 years, $p < 0.001$) — the single largest and most statistically significant difference in the table, confirming that T2DM risk rises substantially with age as insulin resistance tends to increase with advancing age and declining muscle mass. Hypertension prevalence was more than six fold higher in the diabetic group (34.9% vs. 5.5%, $p < 0.001$), reflecting the shared metabolic syndrome aetiology of both conditions. Mean BMI was significantly higher in the diabetic group (27.3 vs. 23.8 kg/m^2, $p < 0.001$), consistent with the role of excess adiposity in promoting peripheral insulin resistance. Urban residence was more prevalent among diabetic respondents (44.1% vs. 37.4%, $p = 0.018$), likely reflecting greater diagnostic access in cities rather than inherently higher biological risk — a diagnostic ascertainment effect attributable to the self-reported outcome variable.

4.2. Train-Test Split and SMOTE

The dataset was partitioned into a training set (80%; n = 25,083) and a held-out test set (20%; n = 6,271) using stratified random splitting (random_state = 2024). Stratification on the outcome variable preserved diabetes prevalence across both partitions, as shown in Table 2. The test set was withheld from all training and hyper parameter tuning steps.

Table 2. *Train-Test Split — KDHS 2022 Combined Dataset (N = 31,354).*

Partition	Total n	Diabetic (1), n (%)	Not Diabetic (0), n (%)
Full Dataset	31,354	272 (0.87%)	31,082 (99.13%)
Training Set (80%)	25,083	218 (0.87%)	24,865 (99.13%)
Test Set — Held-out (20%)	6,271	54 (0.86%)	6,217 (99.14%)
After SMOTE (Training only)	49,730	24,865 (50.0%)	24,865 (50.0%)

Note. Stratified split (random_state = 2024) on the combined dataset. SMOTE (k_neighbors = 5) applied to training set only. Test set unchanged and reflects the original 114.3:1 class imbalance ratio.

4.3. Feature Importance (RFECV)

RFECV applied to the SMOTE-balanced training set retained all ten candidate predictor variables, confirming that no variable was redundant. Table 3 presents the feature importance rankings from the RFECV base estimator. Age accounted for 31.7% of total predictive information — by far the largest single contributor. Adding Wealth Index raised cumulative importance to 53.0%, and Hypertension to 69.9%, establishing these three variables as the backbone of T2DM prediction in the KDHS 2022 dataset.

Table 3. *RFECV Feature Importance Rankings — Combined Dataset (N = 31,354).*

Rank	Predictor Variable	Gini Importance	Cumulative %	Selected
1	Age	0.3172	31.7%	Yes
2	Wealth Index	0.2133	53.0%	Yes
3	Hypertension	0.1683	69.9%	Yes
4	Employment Status	0.0637	76.3%	Yes
5	BMI (kg/m^2)	0.0617	82.4%	Yes
6	Education Level	0.0467	87.1%	Yes
7	Residence (Urban/Rural)	0.0404	91.1%	Yes
8	Tobacco Use	0.0306	94.2%	Yes
9	Sex	0.0259	96.8%	Yes
10	Alcohol Use	0.0191	98.7%	Yes

Note. Gini importance scores from RFECV base Random Forest estimator. All 10 variables retained, confirming no predictor is redundant.

4.4. Random Forest Model Performance

Table 4 presents the optimal Random Forest hyperparameter configuration identified via grid search with 5-fold stratified cross-validation.

Table 4. *Optimal Random Forest Hyperparameters.*

Hyperparameter	Optimal Value
Number of Trees (n_estimators)	300
Maximum Tree Depth	10
Minimum Samples per Leaf	3
Features per Split	sqrt
Calibrated Threshold	0.20
5-Fold CV AUC-ROC	0.9895 ±0.0004
5-Fold CV PR-AUC	0.9885 ±0.0005
5-Fold CV Balanced Accuracy	0.9316 ±0.0033

Note. Grid search over n_estimators ∈ {100, 200, 300}, max_depth ∈ {5, 10, 15}, min_samples_leaf ∈ {1, 3, 5}. Optimisation criterion: AUC-ROC. random_state = 2024.

Table 5 presents the comprehensive Random Forest performance metrics on both the held-out test set and 5-fold cross-validation.

Table 5. *Random Forest Comprehensive Performance Metrics.*

Metric	Test Set	5-Fold CV Mean	CV Std Dev
AUC-ROC	0.7296	0.9895	0.0004
PR-AUC	0.0488	0.9885	0.0005
Balanced Accuracy	0.6640	0.9316	0.0033
Recall (Sensitivity)	0.6296	0.9434	0.0035
Specificity	0.6981	0.9200	0.0041
Precision	0.0178	0.9265	0.0022
F1-Score	0.0346	0.9345	0.0028
MCC	0.0659	—	—
True Positives (TP)	34	—	—
False Negatives (FN)	20	—	—
False Positives (FP)	1,875	—	—
True Negatives (TN)	4,342	—	—

Note. Test set metrics reflect the original class imbalance ratio of 114.3:1 (n = 6,271). CV metrics reflect balanced SMOTE training data. MCC is not reported for CV as it requires the full confusion matrix. TP > FN (34 > 20) confirms Recall > 50%, satisfying the minimum requirement for a clinically viable screening tool.

The Random Forest model achieved a test-set AUC-ROC of 0.7296, indicating acceptable discrimination between diabetic and non-diabetic individuals substantially above the 0.65–0.79 range reported for logistic regression-based models in African-descent populations [12]. At the calibrated threshold of 0.20, the model correctly identified 34 of 54 diabetic cases (sensitivity = 62.96%), satisfying the recall > 50% requirement for a viable screening tool. The 1,875 false positives represent individuals flagged for follow-up glucose testing who are not diabetic; in a screening context this is acceptable, as the cost of a false positive — an unnecessary follow-up test — is substantially lower than the cost of a false negative — a missed diagnosis with risk of life-threatening complications. The notable gap between cross-validation performance (AUC = 0.9895) and test-set performance (AUC = 0.7296) is attributable to the SMOTE-balanced training distribution, where cross-validation is conducted on balanced data, contrasted with the severely imbalanced test set reflecting real-world prevalence.

4.5. SHAP Interpretability Analysis

SHAP analysis was applied to the trained Random Forest model on the held-out test set to quantify each feature's contribution to T2DM predictions. Table 6 presents the global SHAP feature importance magnitudes and directional effects.

Table 6. *SHAP Feature Importance and Direction of Effect — Random Forest Model.*

Feature	SHAP Magnitude	Direction	Epidemiological Interpretation
Age / Age^2	0.2106	↑ → Diabetic	Insulin resistance accelerates with age; nonlinear effect captured by Age^2
Age × BMI	0.2001	↑ → Diabetic	Joint effect of ageing and adiposity on insulin resistance
Employment Status	0.1121	Non-employed → Diabetic	Reduced physical activity, psychosocial stress, irregular diet
BMI (kg/m^2)	0.1063	↑ → Diabetic	Excess adiposity directly promotes peripheral insulin resistance

Feature	SHAP Magnitude	Direction	Epidemiological Interpretation
Wealth Index	0.0968	↑ → Diabetic	Higher wealth associated with greater diagnostic access and sedentary lifestyle risk
Sex	0.0866	Female → Not Diabetic	Male sex associated with slightly higher T2DM prevalence (0.97% vs 0.78%)
Education Level	0.0072	↑ → Not Diabetic	Higher education protective: health literacy and preventive behaviour
Hypertension	—	Present → Diabetic	Metabolic syndrome co-occurrence; key clinical screening trigger

Note. SHAP magnitudes represent mean absolute contribution to predicted T2DM probability across the test set. Direction indicates whether higher feature values push predictions toward Diabetic (1) or Not Diabetic (0). ↑ = higher values increase T2DM risk.

Age-related variables dominated the SHAP importance ranking, with raw age and the Age^2 term jointly contributing the highest mean absolute SHAP magnitude (0.2106), confirming that age dynamics are the most influential drivers of T2DM prediction in the Kenyan population. The Age × BMI interaction term (SHAP magnitude = 0.2001) captured the joint effect of ageing and adiposity on insulin resistance — a clinically meaningful interaction not detectable by univariate screening. BMI, Wealth Index, Sex, and Education followed in a stable descending order, confirming that both physiological and socioeconomic factors contributed meaningfully to model predictions. All directional effects were consistent with established epidemiological evidence on T2DM risk factors, providing strong face validity for the SHAP-derived risk factor profile.

5. Discussion

This study developed and evaluated a Random Forest classifier for T2DM risk prediction using the nationally representative KDHS 2022 dataset, achieving a cross-validation AUC-ROC of 0.9895 and a test-set AUC-ROC of 0.7296 at a calibrated threshold of 0.20 with sensitivity of 62.96%. These findings are interpreted in the context of the existing literature on ML-based T2DM prediction in Kenya and sub-Saharan Africa.

The test-set AUC-ROC of 0.7296 substantially exceeds the 0.65–0.79 range reported for conventional logistic regression-based models in African-descent populations [12], supporting the methodological advantage of RF over traditional statistical approaches in this context. The finding is consistent with Abdelhady et al. [18], who confirmed RF's competitive T2DM prediction performance on electronic health record data, and with Rodriguez et al. [19] and Malathi and Suhana [22], who demonstrated RF's reliable generalisation across diverse populations. The present study extends this evidence to a nationally representative combined male and female Kenyan dataset — a context no previous study has achieved.

The substantial gap between cross-validation AUC (0.9895) and test-set AUC (0.7296) warrants careful interpretation. Cross-validation was conducted on the SMOTE-balanced training set where both classes are equally represented, producing optimistic performance estimates under artificial balance. The test set retained the original 114.3:1 imbalance, reflecting real-world prevalence. This gap is an expected and well-documented phenomenon when SMOTE is applied before cross-validation on severely imbalanced datasets [29], and underscores the critical importance of evaluating final model performance on an untransformed, imbalanced held-out test set. The test-set PR-AUC of 0.0488, while numerically small, substantially exceeds the random classifier baseline of 0.0087 (equal to the minority class prevalence), confirming genuine predictive value beyond chance [27].

The SHAP analysis identified age as the dominant predictor of T2DM risk in the Kenyan population (SHAP magnitude = 0.2106), consistent with the descriptive finding that diabetic respondents were on average 9.4 years older than non-diabetic respondents (38.8 vs. 29.4 years, $p < 0.001$). This aligns with established evidence that insulin resistance and beta-cell dysfunction accelerate with advancing age [1, 2]. The Age × BMI interaction term emerged as the second most important predictor (SHAP magnitude = 0.2001), capturing the amplified T2DM risk associated with the combination of ageing and excess adiposity — a nonlinear interaction that conventional logistic regression cannot detect without manual pre-specification.

The positive SHAP direction for Wealth Index — indicating that higher wealth is associated with higher predicted T2DM risk — appears counterintuitive but is likely explained by diagnostic access effects: wealthier Kenyans are more likely to have visited a healthcare provider and received a diabetes diagnosis, inflating observed prevalence in higher wealth quintiles relative to the true population burden. This finding echoes the urban residence diagnostic access effect observed in the descriptive statistics and is consistent with evidence on healthcare utilisation disparities in Kenya [6]. Importantly, both directional effects were identified consistently by the SHAP analysis, distinguishing genuine biological risk factors from diagnostic ascertainment artefacts.

The protective effect of higher education on predicted

T2DM risk (negative SHAP direction) is consistent with evidence on the role of health literacy in promoting preventive behaviour and earlier health-seeking [30]. Female sex was also associated with lower predicted T2DM risk, consistent with the observed sex differential in prevalence (0.78% among women vs. 0.97% among men in the KDHS 2022 dataset). Employment status emerged as the third most important predictor (SHAP magnitude = 0.1121), with non-employment associated with higher predicted T2DM risk — likely reflecting the combined effects of reduced physical activity, psychosocial stress, and irregular dietary patterns associated with unemployment.

6. Practical Implications

The SHAP-derived risk factor profile from this study has direct and actionable implications for public health policy and clinical practice in Kenya. For the Ministry of Health and county health departments, the finding that age is the dominant predictor supports prioritising community-level blood glucose screening for all Kenyans aged 35 years and above, regardless of sex, as the strongest population-level screening criterion. The Wealth Index finding — indicating systematic under-diagnosis in lower-wealth and rural communities — highlights the urgent need for proactive outreach screening targeting these underserved populations to close the diagnostic gap.

For clinical practice, the hypertension-T2DM association (34.9% of diabetic respondents had hypertension vs. 5.5% of non-diabetic respondents, $p < 0.001$) supports the recommendation that concurrent blood pressure and blood glucose measurement become standard protocol at the primary care level for all patients presenting with hypertension. The RF model's probability scores could be integrated into Kenya's Community Health Promoter referral tools, providing a low-cost, data-driven complement to existing screening approaches within resource-constrained primary health care settings. The strong BMI-T2DM association further supports routine BMI monitoring as a low-cost screening component at community level.

7. Limitations

This study is subject to several important limitations that should be considered when interpreting the findings. First, the study relied on secondary data from the KDHS 2022 survey, which constrains the range of available predictor variables and excludes certain clinically relevant T2DM risk factors — including detailed dietary intake, family history of diabetes, fasting plasma glucose, and HbA1c — that were not captured in the survey instrument. The outcome variable was based on self-reported healthcare provider diagnosis, which is susceptible to diagnostic ascertainment bias, particularly in rural and lower-wealth communities where healthcare access is limited.

Second, the cross-sectional design of the KDHS 2022 precludes causal inference. All findings are explicitly framed as associative and predictive rather than causal, and SHAP values represent predictive contributions to model output rather than causal influences on disease onset. Third, the substantial gap between cross-validation AUC (0.9895) and test-set AUC (0.7296) reflects the well-documented optimism introduced by SMOTE-based augmentation of the training set, and underscores the necessity of reporting test-set performance on the original imbalanced distribution as the primary performance estimate.

Fourth, the RF model was trained on Kenyan national data and may not generalise to other sub-Saharan African populations with different sociodemographic profiles. External validation using DHS survey data from other East African countries is recommended before broader regional application. Finally, the models generate T2DM risk predictions rather than clinical diagnoses, and their outputs should be interpreted as screening tools to guide further clinical assessment rather than as definitive diagnostic instruments.

8. Conclusion

This study developed and evaluated a Random Forest classifier for T2DM risk prediction applied to the nationally representative KDHS 2022 datasset, comprising 31,354 Kenyan adults spanning all 47 counties. The RF model achieved a cross-validation AUC-ROC of 0.9895 (SD = 0.0004) and a test-set AUC-ROC of 0.7296, with sensitivity of 62.96% at a calibrated threshold of 0.20, correctly identifying 34 of 54 diabetic cases in the held-out test set. SHAP analysis identified a consistent, clinically interpretable risk factor hierarchy dominated by age-related variables, employment status, BMI, wealth index, sex, and education — all with directional effects consistent with established epidemiological evidence on T2DM risk in sub-Saharan Africa.

To the researcher's knowledge, this is the first study to apply Random Forest with SHAP-based interpretability to the full, nationally representative KDHS 2022 dataset for T2DM risk prediction in Kenya, producing evidence that is both scientifically rigorous and nationally generalisable. The study demonstrates that the combination of SMOTE resampling, F2-score-guided threshold calibration, and imbalance-appropriate evaluation metrics is necessary for viable minority-class detection in severely imbalanced population health survey data. The SHAP-derived risk factor profile provides a model-independent, nationally representative foundation for T2DM risk stratification and evidence-based public health decision-making in Kenya.

9. Future Work

Four research directions are recommended based on this study's findings. First, external validation of the RF model using DHS survey data from other East African countries —

Tanzania, Uganda, and Ethiopia — to assess regional generalisability and identify population-specific differences in T2DM risk factor profiles. Second, incorporation of fasting plasma glucose or HbA1c as the outcome variable replacing the self-reported diagnosis used here, using datasets such as the Kenya STEPS 2015 which included fasting blood glucose measures, to improve outcome specificity and reduce diagnostic ascertainment bias. Third, county-level SHAP sub-analyses within the KDHS 2022 dataset to identify whether T2DM risk factor profiles differ across Kenya's 47 counties, enabling geographically differentiated public health interventions. Fourth, development and prospective validation of a digital clinical decision-support tool incorporating the RF probability scores for integration into Kenya's Community Health Promoter referral framework at the primary care level.

Abbreviations

T2DM	Type 2 Diabetes Mellitus
KDHS	Kenya Demographic and Health Survey
RF	Random Forest
ML	Machine Learning
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
RFECV	Recursive Feature Elimination with Cross-Validation
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
PR-AUC	Precision-Recall Area Under the Curve
MCC	Matthews Correlation Coefficient
BMI	Body Mass Index
NCD	Non-Communicable Disease
SSA	Sub-Saharan Africa
ANN	Artificial Neural Network
SVM	Support Vector Machine
LDA	Linear Discriminant Analysis
KNBS	Kenya National Bureau of Statistics
USAID	United States Agency for International Development
KEMRI	Kenya Medical Research Institute
SD	Standard Deviation

Acknowledgments

The author is deeply grateful to supervisors Dr. Josephine Njeri Ngure and Dr. Mutua Kilai of the Department of Pure and Applied Sciences, Kirinyaga University, for their guidance, constructive feedback, and unwavering support throughout this research. Sincere thanks are extended to Kirinyaga University for providing an enabling academic environment and the resources necessary to pursue graduate studies in Data Science and Analytics. The author acknowledges the Kenya National Bureau of Statistics, ICF International, and the DHS Program for making the KDHS 2022 dataset publicly available for academic research.

Author Contributions

Diana Boro: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing

Data Availability Statement

The KDHS 2022 dataset used in this study is publicly available upon registered request from the DHS Program at https://dhsprogram.com. Data access was granted for the registered project titled 'Comparison of Random Forest and XGBoost Models for Predicting Type 2 Diabetes in Kenya Using KDHS 2022 Data.' The analytical code is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] American Diabetes Association. (2022). Standards of medical care in diabetes — 2022. Diabetes Care, 45(Suppl. 1), S1–S264. https://doi.org/10.2337/dc22-S001

[2] World Health Organization. (2023). Diabetes fact sheet. https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] International Diabetes Federation. (2021). IDF diabetes atlas (10th ed.). https://www.diabetesatlas.org

[4] Ministry of Health Kenya. (2015). Kenya STEPwise survey for non-communicable diseases risk factors 2015 report. Ministry of Health Kenya.

[5] Kenya National Bureau of Statistics. (2023). Kenya demographic and health survey 2022. Kenya National Bureau of Statistics. https://www.knbs.or.ke

[6] Muriuki, B. M., Mugo, M., & Ogola, E. N. (2020). Burden of type 2 diabetes mellitus in Kenya: A systematic review and meta-analysis. African Journal of Primary Health Care & Family Medicine, 12(1), e1–e9. https://doi.org/10.4102/phcfm.v12i1.2510

[7] Mwakondo, F., Kerre, G., & Muchiri, E. (2025). Machine learning for type 2 diabetes risk prediction using Kenyan electronic health records. East African Journal of Health Informatics, 3(1), 1–14.

[8] Nimmagadda, S., Rao, P., & Reddy, K. (2024). XGBoost and Random Forest-based diabetes prediction using Kenyan clinical data. African Journal of Artificial Intelligence and Sustainable Development, 4(1), 1–12.

[9] Nyenwe, E. A., Odia, O. J., Ihekwaba, A. E., Ojule, A., & Babatunde, S. (2011). Type 2 diabetes in Nigerians: Prevalence and risk factors in Port Harcourt, Nigeria. Diabetes Research and Clinical Practice, 62(3), 177–185. https://doi.org/10.1016/j.diabres.2003.09.007

[10] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

[11] Mancini, L. (2018). Ordinal data supervised classification with quantile-based and other classifiers [Doctoral dissertation]. https://doi.org/10.6092/unibo/amsdottorato/8543

[12] Pilla, F., et al. (2019). A systematic review of diabetes prediction models in African populations. Journal of Diabetes Research, 2019, 1–12.

[13] Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Bowen, Y., Yu, J., Li, C., Yu, F., & Ren, Z. (2022). Machine learning models for data-driven prediction of diabetes by lifestyle type. International Journal of Environmental Research and Public Health, 19(22), 15027. https://doi.org/10.3390/ijerph192215027

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

[15] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1007/BF00058655

[16] Biggs, F. (2024). Exploring generalisation performance through PAC-Bayes [Doctoral dissertation, UCL].

[17] Jafari, Z., Harari, R. E., Hole, G., Kolb, B. E., & Mohajerani, M. H. (2025). Machine learning models can predict tinnitus and noise-induced hearing loss. Ear & Hearing, 46(5), 1305–1316. https://doi.org/10.1097/AUD.0000000000001670

[18] Abdelhady, A., El-Sappagh, S., Barakat, S., & Abdelrazek, S. (2020). Feature selection and classification for type 2 diabetes prediction using XGBoost and Random Forest. IEEE Access, 8, 1–14. https://doi.org/10.1109/ACCESS.2020.3016968

[19] Rodriguez, J., Perez, A., & Lozano, J. (2017). Sensitivity analysis of k-nearest neighbours and decision trees for type 2 diabetes prediction in European cohorts. Pattern Recognition, 70, 50–62. https://doi.org/10.1016/j.patcog.2017.04.021

[20] Garcia, E., Perez, M., & Torres, R. (2016). Logistic regression and decision tree models for diabetes risk in Latin American populations. Journal of Diabetes Research, 2016, 1–9. https://doi.org/10.1155/2016/1823206

[21] Santos-Silva, D., Ferreira, A., & Costa, R. (2024). Longitudinal blood glucose trajectory prediction using XGBoost and Random Forest over 52 weeks. Journal of Biomedical Informatics, 151, 104610. https://doi.org/10.1016/j.jbi.2024.104610

[22] Malathi, D., & Suhana, M. (2025). Diabetes risk prediction using Random Forest and XGBoost on the Indian Primary Diabetes Dataset. Healthcare Analytics, 7, 100110.

[23] Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8(1), 25. https://doi.org/10.1186/1471-2105-8-25

[24] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[25] Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Wiley. https://doi.org/10.1002/9781119013563

[26] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182. https://doi.org/10.1162/153244303322753616

[27] Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, 233–240. https://doi.org/10.1145/1143844.1143874

[28] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

[29] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[30] Peer, N., Kengne, A. P., Motala, A. A., & Mbanya, J. C. (2014). Diabetes in the Africa region: An update. Diabetes Research and Clinical Practice, 103(2), 197–205. https://doi.org/10.1016/j.diabres.2014.01.012