JOURNAL OF MECHANICS OF CONTINUA AND
MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-21, No.-02, February (2026) pp 1-22 ISSN (Print) 0973-8975

XGBOOST AND COST-SENSITIVE CART FOR IMBALANCED MULTICLASS DIABETES CLASSIFICATION IN IRAQ

Nabila A. Alsharif[1], **Inaam Aboud Hussain**[2], **Loaiy F. Naji**[3]

[1,2,3] Department of Statistics, College of Administration and Economics,
University of Baghdad, Iraq.

Email: nabila_alsharif@coadec.uobaghdad.edu.iq,
Inaam.aboud@coadec.uobaghdad.edu.iq, Loay.F@coadec.uobaghdad.edu.iq

Corresponding Author: **Nabila A. Alsharif**

https://doi.org/10.26782/jmcms.2026.02.00001

(Received: December 11, 2025; Revised: January 23, 2026; Accepted : February 03, 2026)

Abstract

Diabetes imposes a substantial public health burden; according to the International Diabetes Federation, there were about 3.4 million diabetes related deaths worldwide in 2024, and in Iraq, the Federation reports that one in nine adults lives with diabetes in 2024, with 14,683 adult deaths attributable to diabetes and a total diabetes related health expenditure of 2,078 million United States dollars. The dataset analyzed in this study contains 1,000 records collected in 2020 from two Iraqi teaching hospitals and includes multiple clinical and laboratory measurements with three outcome classes, namely Non diabetic, Pre diabetic, and Diabetic, with a low prevalence of the Pre diabetic class and an imbalanced overall class distribution; the data are challenging because they contain many outliers, non homogeneous covariance matrices across classes, exact duplicate rows that were removed before modelling, and linear correlations among certain variables. The study objective was to train and evaluate models that discriminate among the three classes and yield accurate, well calibrated predictions for future cases in similar clinical settings, but the diagnostic properties of the data limited the applicability of classical discriminant functions; therefore two supervised learners were employed: Classification and Regression Trees (CART) and Extreme Gradient Boosting (XGBoost), together with preprocessing that removed exact duplicate rows and excluded VLDL because it is algebraically derived from triglycerides in mmol per liter as VLDL equals triglycerides divided by 2.2, which would introduce redundancy and multicollinearity. On the held-out test set, XGBoost achieved higher Accuracy at 98.18 percent compared with 97.58 percent for CART and higher Balanced Accuracy at 93.84 percent compared with 88.16 percent for CART, indicating that XGBoost provided the strongest overall operating point for this three-class task while CART remains useful when simple and transparent rules are required.

Keywords: Classification, XGBoost, CART, Class imbalance, Diabetes, Pre diabetic

I. Introduction

Diabetes mellitus is a rapidly growing global health problem with large mortality and cost burdens. In 2024, an estimated 589 million adults were living with diabetes worldwide, projections indicate approximately 853 million by 2050, diabetes caused about 3.4 million deaths in 2024, and global health expenditure attributable to diabetes was about one trillion US dollars [V].

Iraq faces a substantial national burden. In 2024, an estimated 2.7 million Iraqi adults were living with diabetes, 47.1 percent were undiagnosed, there were about 14,683 diabetes-attributable deaths, and diabetes-related health expenditure totaled roughly 2,078 million US dollars, or about 778 US dollars per person with diabetes [XI]. Clinical practice guidelines emphasize early detection and accurate assignment to glycemic status because timely intervention can prevent progression and complications; in this study, we apply the diagnostic categories and thresholds specified in the American Diabetes Association Standards of Care [I].

The dataset analyzed here was obtained from Mendeley Data and contains adult records from two tertiary hospitals in central Baghdad: Medical City Teaching Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital. The repository documents the three target classes (Non-diabetic, Prediabetic, Diabetic) and the routine clinical and laboratory predictors used in our work [XIV]. The data are challenging and representative of routine clinical practice: several variables exhibit heavy right tails and outliers; covariance structures differ across the three classes; there is modest collinearity among lipid and renal markers; and exact duplicate records were identified and resolved during cleaning. The classes are strongly imbalanced, and, importantly, the explicit Prediabetic stratum makes these data more informative than binary designs that collapse Prediabetic into Non-diabetic or Diabetic, while also increasing the difficulty of classification and calibration [I].

We compare two complementary approaches for three-class assignments: a cost-sensitive CART model that yields transparent decision rules, and XGBoost as a strong nonlinear baseline. Both models use the same original predictors: nine continuous variables (anthropometric and laboratory measures) and Gender. CART ingests these ten predictors directly with native handling of categorical splits. XGBoost operates on a one-hot encoded design matrix with eleven columns (the same nine continuous variables plus two indicator columns for Gender, no intercept). Performance is evaluated on a held-out test set using Accuracy, Balanced Accuracy, Precision, Recall, F1 and Macro-F1, ROC, and precision–recall behavior, the Brier score for probability accuracy, and Cohen's kappa for chance-corrected agreement; formal definitions are provided in Methods.

II. Literature Review

Sabariah et al. in 2014 [XVI]: In this research, the combined (CART) and (RF) were used to build the classification model that is used in the early detection of diabetes mellitus type II disease. Those methods are selected based on the characteristics of the dataset used in medical records of diabetes mellitus, which consist of complex attributes consisting of several categorical attributes and continuous attributes. This research has tested a different number of trees and numbers of candidate attribute splitters, with the optimal inputs 50 trees and 3 number of attributes splitter, the average accuracy 83.8%. The important attributes of early detection of diabetes mellitus type II are heredity, age, and body mass index.

Nuankaew et al. in 2021 [XIII]: This study proposes a novel prediction method named Average Weighted Objective Distance (AWOD) based on the assumption that the individual has diverse health conditions resulting from different individual factors, a requirement for an effective prediction model. AWOD is a modification of Weighted Objective Distance by applying information gain to reveal significant and insignificant individual factors having different priorities, which are represented by different weights. Two datasets from open sources, Pima Indians Diabetes (Dataset 1) and Mendeley Data for Diabetes (Dataset 2), each containing 392 records, with a 70:30 partition, were studied. The prediction performance for both datasets is compared with the machine learning-based prediction methods, including K-Nearest Neighbors, Support Vector Machines, Random Forest, and Deep Learning. The comparison results showed that the proposed method provided 93.22% and 98.95% accuracy for Dataset 1 and Dataset 2, respectively.

Sahid et al. in 2024[XVII]: They propose a multiclass diabetes mellitus detection and classification approach using an extremely imbalanced Laboratory of Medical City Hospital data dynamics. They also formulate a new dataset that is moderately imbalanced based on the Laboratory of Medical City Hospital data dynamics. To correctly identify the multiclass diabetes mellitus, they employ three machine learning classifiers, namely support vector machine, logistic regression, and k-nearest neighbor. To optimize the classification performance of classifiers, they tune the model by hyperparameter optimization with 10-fold grid search cross-validation. In the case of the original extremely imbalanced dataset with a 70:30 partition and SVM classifier, they achieved a maximum accuracy of 0.964 by using the top 4 features according to the filter method. By using the top 9 features according to wrapper-based sequential feature selection, the KNN provides an accuracy of 93. 5% and 100% for the other performance metrics. For the moderately imbalanced dataset with an 80:20 partition, the SVM classifier achieves a maximum accuracy of 93.8% and 100% for other performance metrics.

Idhom et al. in 2025[XII]: This study aims to address the challenge of predicting customer credit eligibility by employing two machine learning techniques: CART and XGBoost. The research follows a structured methodology, including data acquisition, preprocessing, splitting the data into 80:20 training and testing sets, applying the CART and XGBoost algorithms, and evaluating the models' performance. Through this approach, the study seeks to enhance the accuracy and efficiency of credit approval decisions, helping financial institutions streamline their processes.
The CART method achieved an accuracy rate of 88%, while combining CART with the XGBoost algorithm increased accuracy to 90%.

III. Data source and variable dictionary

This study uses an anonymized dataset obtained from Mendeley Data XIV, comprising adult records from two tertiary hospitals in central Baghdad: Medical City

Teaching Hospital and Al-Kindy Teaching Hospital. The outcome has three classes, non-diabetic (Non), prediabetic (Pre), and diabetic (Diab), with markedly unequal class sizes (Non =103, Pre= 53, Diab =844). The predictors are routine clinical and laboratory measures collected in real-world practice. Table 1 below lists each variable with its full clinical name, unit, and type.

Table 1: Variable dictionary

	Variable (dataset label)	Full name/definition	Unit	Type
1	ID	Record identifier	none	Categorical (identifier)
2	No_Pation	Patient identifier	none	Categorical (identifier)
3	Gender	Sex	none	Categorical
4	AGE	Age	years	Numeric
5	Urea	Serum urea (blood urea nitrogen)	mmol/L	Numeric
6	Cr	Serum creatinine	µmol/L	Numeric
7	HbA1c	Hemoglobin A1c	%	Numeric
8	Chol	Total cholesterol	mmol/L	Numeric
9	TG	Triglycerides	mmol/L	Numeric
10	HDL	High-density lipoprotein cholesterol	mmol/L	Numeric
11	LDL	Low-density lipoprotein cholesterol	mmol/L	Numeric
12	VLDL	Very-low-density lipoprotein cholesterol	mmol/L	Numeric
13	BMI	Body Mass Index	kg/m²	Numeric
14	CLASS	Glycemic class	none	Categorical (outcome; levels: Non, Pre, Diab)

In Table 2, statistical descriptions for numeric variables are provided, with counts, means, standard deviations, medians, interquartile range, minimum, and maximum values.

Table 2: Descriptive statistics for numeric variables

	Variable	n	Mean	SD	Median	IQR	Min	Max
1	AGE	1000	53.528	8.799	55.0	8.0	20.0	79.00
2	Urea	1000	5.125	2.935	4.6	2.0	0.5	38.90
3	Cr	1000	68.943	59.985	60.0	25.0	6.0	800.00
4	HbA1c	1000	8.281	2.534	8.0	3.7	0.9	16.00
5	Chol	1000	4.863	1.302	4.8	1.6	0.0	10.30
6	TG	1000	2.350	1.401	2.0	1.4	0.3	13.80
7	HDL	1000	1.205	0.66	1.1	0.4	0.2	9.90
8	LDL	1000	2.610	1.115	2.5	1.5	0.3	9.90
9	VLDL	1000	1.855	3.664	0.9	0.8	0.1	35.0
10	BMI	1000	29.578	4.962	30.0	7.0	19.0	47.75

Table 3 shows categorical variables for Gender and CLASS, list counts, and percentages for each level.

Table 3: Categorical variables: counts and percentages by level

	Variable	Level	n	Percent
1	Gender	Female	435	43.5
		Male	565	56.5
2	CLASS	Non	103	10.3
		Pre	53	5.3
		Diab	844	84.4

1. Data Preparation

The dataset was cleaned before modeling as follows:

- Gender (case harmonization). We identified inconsistent casing for the female category in the Gender column (values recorded as “F” and “f”, e.g., row 992). These values were standardized to “F”.
- CLASS (whitespace trimming). We detected leading/trailing or internal spaces in several single-character CLASS codes (“N”, “P”, “Y”) affecting rows 103, 997, 998, 999, and 1000. All extraneous spaces were removed to ensure consistent label formatting.
- Removal of direct identifiers. The first column (Patient ID) and the second column (NO_PATION) were removed to eliminate direct identifiers and prevent their unintended use during modeling.
- Removal of the undocumented field. The column “Sugar Level Blood” was dropped because it is not documented in the original data description and offered no demonstrated analytical value; its removal does not affect downstream analyses.
- Deduplication. We removed 174 exact duplicate rows (retaining the first occurrence), yielding a final analytic dataset of (N = 826) observations with class counts Non = 96, Pre = 40, and Diab = 690.
- VLDL exclusion by design. Very-low-density lipoprotein (VLDL) was excluded because it is approximately redundant with triglycerides (commonly approximated as (in mmol/L, (VLDL ≈ TG/2.2)) [VI]; removing it avoids collinearity without loss of information.

After cleaning, we performed a stratified 80/20 split (seed = 7) into training and test sets. Categorical variables were used natively by CART, whereas the boosted model used a one-hot encoded design matrix (no intercept).

IV. Methods

Classification and Regression Trees (CART)

We trained a **cost-sensitive** multiclass CART for {Non, Pre, Diab} using the Gini impurity**.** For a node t with class proportions $p_k(t)$, the impurity is
$G(t) = 1 - \sum_k p_k(t)^2$.

For a candidate split of a parent node with $n$samples into left/right children of sizes $n_L, n_R$and impurities G_L, G_R, the size-weighted child impurity is

$G_{\text{split}} = \frac{n_L}{n} G_L + \frac{n_R}{n} G_R$,
and the chosen split maximizes the impurity reduction
$\Delta G = G_{\text{parent}} - G_{\text{split}}$.

To prioritize clinically important distinctions, we used the misclassification cost matrix:

$$L = \begin{bmatrix} 0 & 1 & 2 \\ 5 & 0 & 3 \\ 2 & 1 & 0 \end{bmatrix} \quad (rows = true\ class;\ columns = predicted).$$

At a leaf with estimated class probabilities $\{p_k\}$, the reported class minimizes the expected loss $\arg\min_j \sum_i p_i L_{ij}$. The tree growth limits were cost–complexity parameter $cp = 0.0005$, minimum leaf size = 5, and maximum depth = 8, using 10-fold cross-validation; the final model was pruned at the complexity parameter with minimum cross-validated error (standard cost–complexity pruning). Leaf class probabilities are the empirical class frequencies [II]. To link the tree structure to the percentages reported under each terminal node in Figure 4, we define the leaf coverage as follows: Let ℓ denote a terminal node (leaf), n_ℓ the number of training observations routed to ℓ, and N_{train} the total number of training observations. The share of the training set in leaf ℓ is

$$s_\ell = \frac{n_\ell}{N_{\text{train}}}, \%_\ell = 100 \times \frac{n_\ell}{N_{\text{train}}}.$$

These %_ℓ values correspond to the percentages printed under each leaf in the CART diagram (Figure 4) and summarize each terminal node's coverage. [X]

Extreme Gradient Boosting (XGBoost)

We fit a gradient-boosted tree ensemble for the same three-class target using the multiclass softmax objective (softmax maps arbitrary class scores to probabilities that lie in (0,1) and sum to 1). Let x be a feature vector, $f_i(x) = \sum_{t=1}^{T} g_{t,i}(x)$ the accumulated score for class i after T rounds, and

$$\pi_i(x) = \frac{\exp\ (f_i(x))}{\sum_{m=1}^{K} \exp\ (f_m(x))},$$

the predicted class probability. Training minimizes the weighted multinomial log-loss

$$\mathcal{L} = -\frac{1}{n} \sum_{j=1}^{n} w_j \sum_{i=1}^{K} \mathbf{1}\{y_j = i\} \log \pi_i(x_j),$$

with labels encoded $y \in \{0,1,2\}$ and observation weights w_j addressing class imbalance. [VII], [IV]. We used inverse-frequency base weights and doubled the Pre-class weight (Pre ×2). Implementation settings matched the saved analysis: tree depth = 4, learning rate $\eta = 0.15$, row subsampling = 0.9, column-by-tree subsampling = 0.9, $T = 400$ boosting rounds, single-thread execution, and a fixed random seed. The predicted class is $\arg\max_i \pi_i(x)$, while the full probability vector $\pi(x)$ is retained for later evaluation (e.g., precision–recall area, calibration, and Brier score).

Predictors for XGBoost were one-hot encoded without an intercept so that trees split directly on the resulting indicator columns; continuous predictors were used as observed. CART used native categorical splits and therefore did not require one-hot encoding.

Metrics

Let K be the number of classes $\{\text{Non}, \text{Pre}, \text{Diab}\}$ and n test instances, and let the confusion matrix be $M \in \mathbb{N}^{K\times K}$with entry M_{ij} (true i, predicted j); define $\text{TP}_i = M_{ii}$, $\text{FN}_i = \sum_{j\neq i} M_{ij}$, $\text{FP}_i = \sum_{j\neq i} M_{ji}$, and $\text{TN}_i = \sum_{p\neq i} \sum_{q\neq i} M_{pq}$.

Overall Accuracy is $\frac{1}{n}\sum_{i=1}^{K} M_{ii}$;

Recall/Sensitivity for class i is $\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}$;

Specificity is $\text{Specificity}_i = \frac{\text{TN}_i}{\text{TN}_i+\text{FP}_i}$;

Balanced Accuracy averages recall across classes, Balanced Accuracy$=\frac{1}{K}\sum_{i=1}^{K} \text{Recall}_i$ (recommended when class sizes are unequal).

Precision (positive predictive value) is $\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i+\text{FP}_i}$;

The per-class F1 is the harmonic mean $F1_i = \frac{2\,\text{Precision}_i\,\text{Recall}_i}{\text{Precision}_i+\text{Recall}_i}$;

The Macro-F1 is the unweighted mean $\frac{1}{K}\sum_{i=1}^{K} F1_i$.

These definitions and macro-averaging are standard in multiclass evaluation [XVIII]. For ROC analysis, one-vs-rest curves plot $(\text{FPR}_i, \text{TPR}_i)$ where $\text{TPR}_i = \text{Recall}_i$ and $\text{FPR}_i = \frac{\text{FP}_i}{\text{FP}_i+\text{TN}_i}$; the AUC summarizes threshold-free discrimination for a class, and the Hand–Till multiclass AUC generalizes by averaging pairwise class AUCs: $\text{AUC}_{\text{HT}} = \frac{2}{K(K-1)}\sum_{1\leq i<j\leq K} \text{AUC}\,(i \text{ vs } j)$ [VIII]. For precision–recall (PR), curves plot $(\text{Recall}_i, \text{Precision}_i)$; the area under the PR curve (AUPRC) can be more informative than ROC when classes are imbalanced; we report the macro-average over classes [XV]. For probability accuracy, the multiclass Brier score uses the predicted class probabilities and one-hot truth: if $\pi_k(x_j)$ is the predicted probability for instance $j$and class k (e.g., the softmax output in XGBoost), define $y_{jk} \in \{0,1\}$as the indicator of the true class;

then $\text{Brier} = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K}(\pi_k(x_j) - y_{jk})^2$ (lower is better); it is a strictly proper scoring rule. [III]

For chance-corrected agreement, Cohen's Kappa κ uses observed agreement $p_o = \frac{1}{n}\sum_i M_{ii}$ and chance agreement $p_e = \sum_{i=1}^{K}(\frac{\sum_j M_{ij}}{n})(\frac{\sum_j M_{ji}}{n})$, with $\kappa = \frac{p_o-p_e}{1-p_e}$. [IX]

V. Results

In Figure 1, the matrix shows a few strong positive clusters and otherwise modest associations. Urea and creatinine exhibit a very strong positive correlation, consistent with shared renal physiology, indicating potential redundancy. Total cholesterol correlates strongly with LDL, and triglycerides show a moderate positive correlation with total cholesterol and LDL. HDL displays a mild negative association with triglycerides and LDL. HbA1c correlates positively (moderate) with BMI and age, consistent with higher glycemic burden in older and heavier patients. No broad pattern of extreme collinearity is evident beyond the Urea–Cr and Chol–LDL pairs, suggesting that most predictors contribute distinct information to the tree-based models.

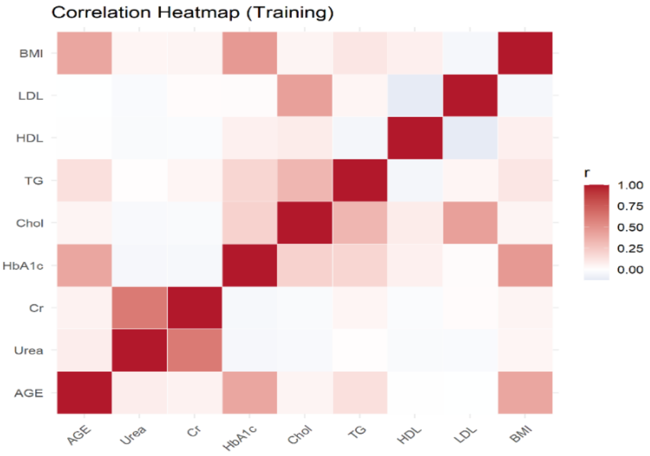


Fig. 1. Correlation heatmap (training set)

In Figure 2, the class-wise distributions demonstrate clear separation on several key variables. HbA1c shows the most distinct shift: Diab has the highest values with limited overlap, Pre is intermediate, and Non is lowest, supporting its role as a primary discriminator. BMI is higher in Diab with Pre slightly above Non, indicating added discriminatory value for adiposity. Triglycerides are elevated and right-skewed in Diab, with Pre between Diab and Non. HDL is lower in Diab, consistent with an adverse lipid profile. LDL and total cholesterol are modestly higher in Diab but retain some overlap across classes. Creatinine and urea tend to be higher in Diab with longer right tails, though distributions overlap. Age is shifted upward in Diab relative to Non and Pre. Overall, these patterns explain why splits on HbA1c, BMI, and lipid measures are effective for assigning patients to Non, Pre, or Diab.

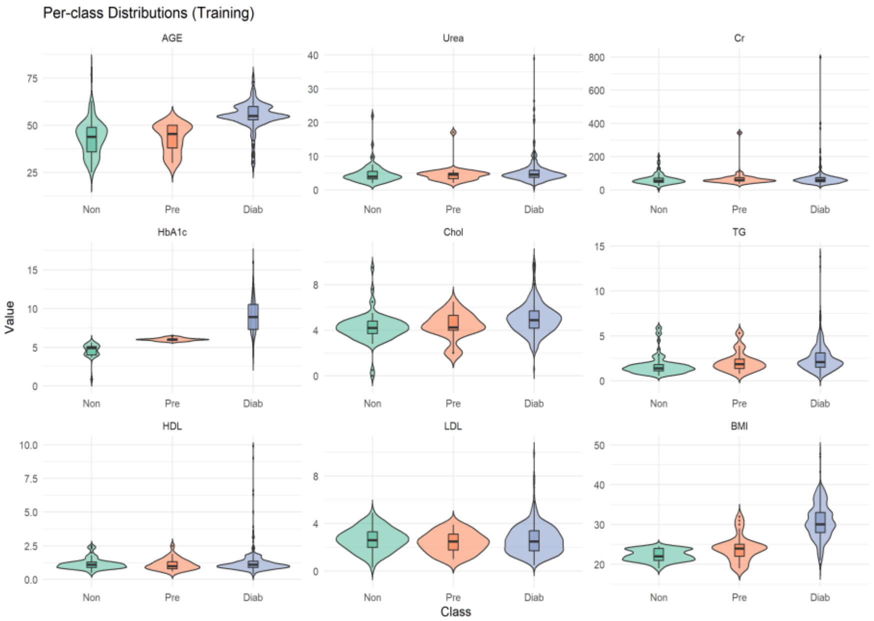


Fig. 2. Per-class distributions

In Figure 3, across classes, most variables deviate from normality chiefly through right-skew and heavy upper tails, most pronounced in Diab. HbA1c shows the clearest class separation: Non is lowest, Pre is tightly clustered near the diagnostic threshold (stepwise pattern from small n), and Diab has a pronounced upper tail, consistent with poorer glycemic control. BMI is right-skewed in all groups, with the heaviest tail in Diab. Triglycerides, creatinine, and urea show strong right-tail departures, again largest in Diab, indicating occasional very high values. LDL and total cholesterol are near-normal in Non/Pre with heavier upper tails in Diab, while HDL shows mild curvature and a few high outliers. These distributional features support the observed tree splits and help explain why probability calibration must be checked; they do not invalidate the tree-based models used for assigning patients to Non, Pre, or Diab.

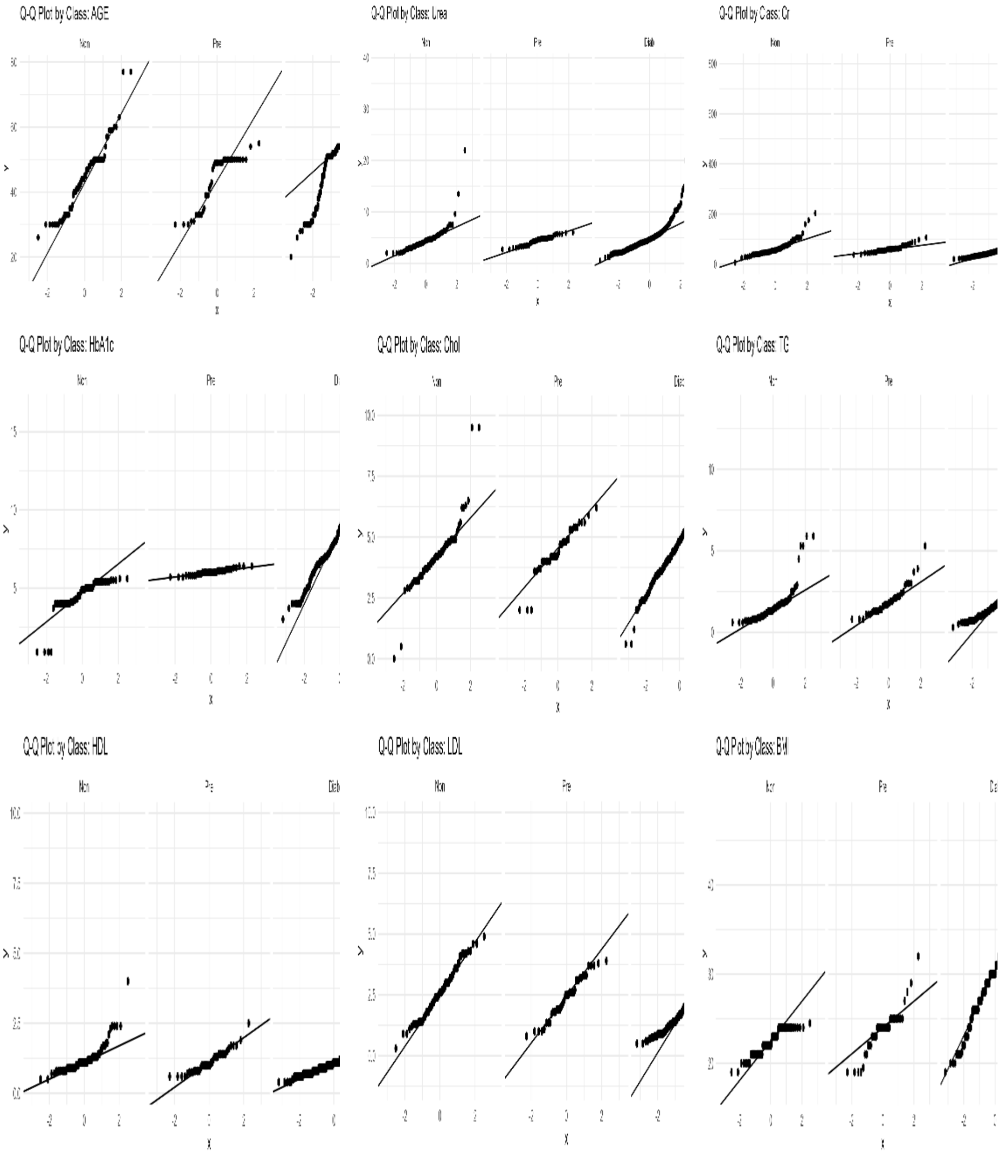


Fig. 3. Q–Q plots by class for the nine continuous predictors (training set)

The pruned tree in Figure 4 ends in eight terminal leaves whose internal leaf numbers are 15, 5, 13, 14, 11, 9, 10, and 6. Each percentage is the share of all training patients that ultimately fall into that leaf after being routed by the splits; the shares are obtained by counting patients per leaf and dividing by the training-set size. The largest leaf is node 15 (75.95%) and corresponds to the simple rule HbA1c ≥ 6.5, producing a near-pure Diab assignment. Within HbA1c < 6.5, node 13 (4.84%) captures the range 5.7 ≤ HbA1c < 6.5 with AGE < 51 and is labeled Pre, whereas node 14 (3.03%) is 5.7 ≤ HbA1c < 6.5 with AGE ≥ 51 and is labeled Diab. In the low-HbA1c branch (HbA1c < 5.7), lipid profile and adiposity refine the decision: node 5 (10.14%) is Chol < 5 and

BMI < 25 (Non); node 6 (1.06%) is Chol < 5 and BMI ≥ 25 (Diab); when Chol ≥ 5, node 11 (2.27%) is TG < 2 with BMI < 24 (Non); node 10 (1.21%) is TG < 2 with BMI ≥ 24 (Diab); and node 9 (1.51%) is TG ≥ 2 (Diab). Thus, the numbers 15, 5, 13, 14, 11, 9, 10, and 6 are the leaf identifiers, and the percentages 75.95, 10.14, 4.84, 3.03, 2.27, 1.51, 1.21, and 1.06 report how much of the training set each rule covers.

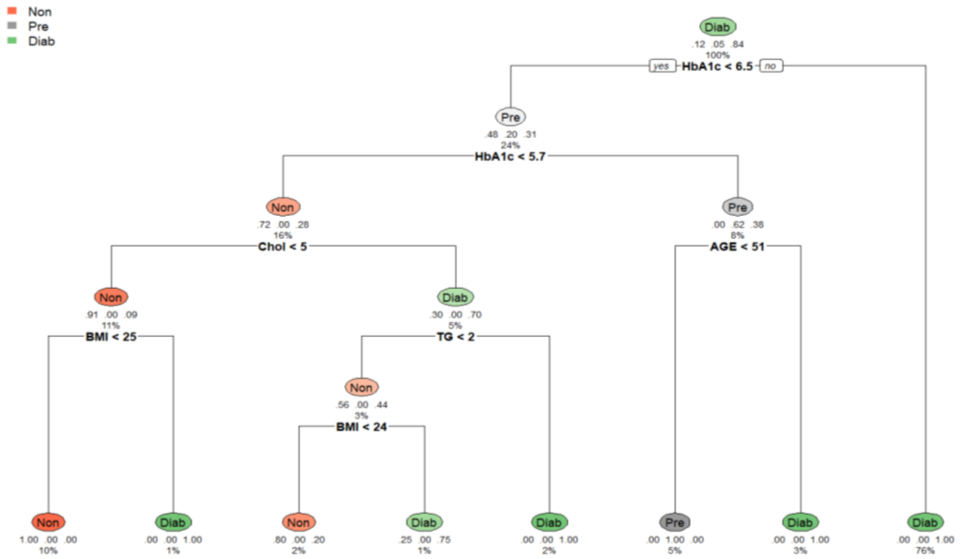
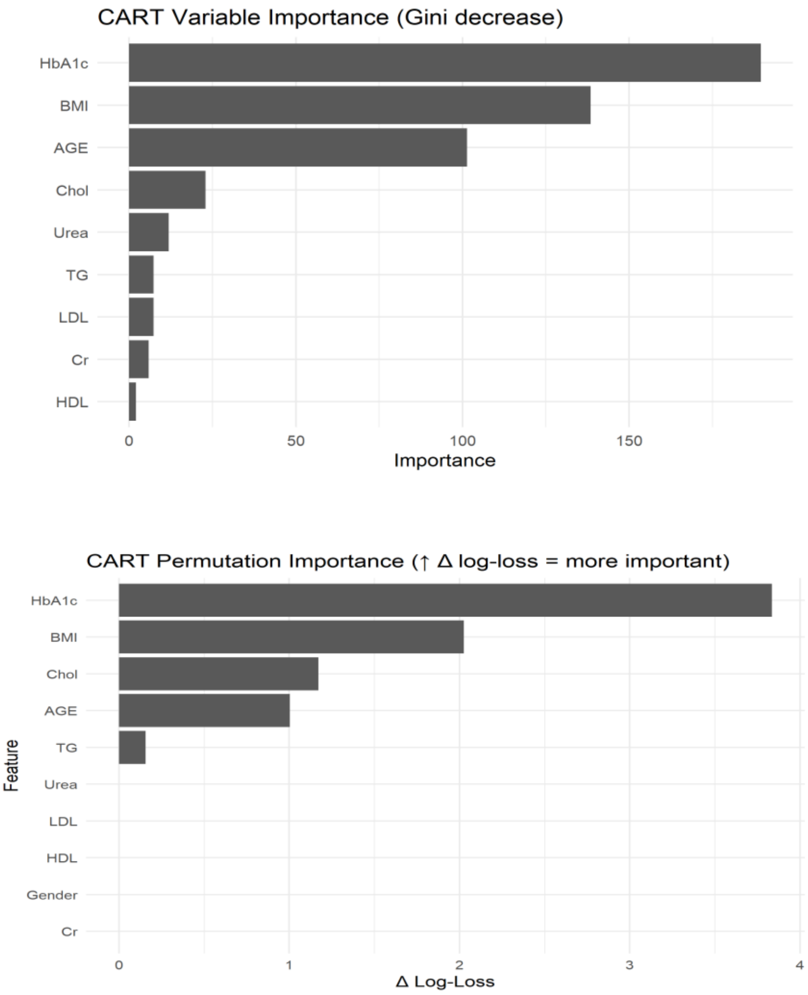


Fig. 4. CART tree (pruned)

Across both views in Figure 5, HbA1c is the dominant driver, BMI is the next most influential, and AGE and Chol contribute at a moderate level; TG and Urea have small effects, while LDL, Cr, and HDL are negligible. The slight reordering (Chol above AGE in the permutation plot but below in the Gini plot) indicates that Chol, though used in fewer or shallower splits, improves the predicted probabilities more than its split frequency alone suggests. The agreement between the two important measures supports a stable interpretation: model decisions are governed primarily by glycemic status, with adiposity as a strong secondary factor and AGE/Chol providing additional refinement.



CART Variable Importance (Gini decrease)

CART Permutation Importance (↑ Δ log-loss = more important)

Fig. 5. Top: CART importance (Gini-based) Bottom: CART permutation importance

The distributions are sharply bimodal, as we see in Figure 6, with most probabilities at 0 or 1 and very little mass in the mid-range, indicating confident, near-deterministic leaves. For P(Diab), the dominant spike at 1.00 reflects the large HbA1c ≥ 6.5 leaf in the tree (Figure 4), which yields confident Diab assignments; the small bar at 0.00 corresponds to clearly non-diabetic cases. For P(Non), probabilities cluster at 0.00 with a small spike at 1.00, matching the small pure Non leaves formed at low HbA1c with favorable lipid/adiposity thresholds. For P(Pre), most probabilities are at 0.00 with a small spike at 1.00, consistent with the narrow Pre leaf ($5.7 \leq$ HbA1c < 6.5 and AGE < 51). Overall, the histograms confirm that CART produces hard, threshold-based decisions aligned with the splits in Figure 4; this explains strong classification accuracy while motivating separate checks of probability calibration.

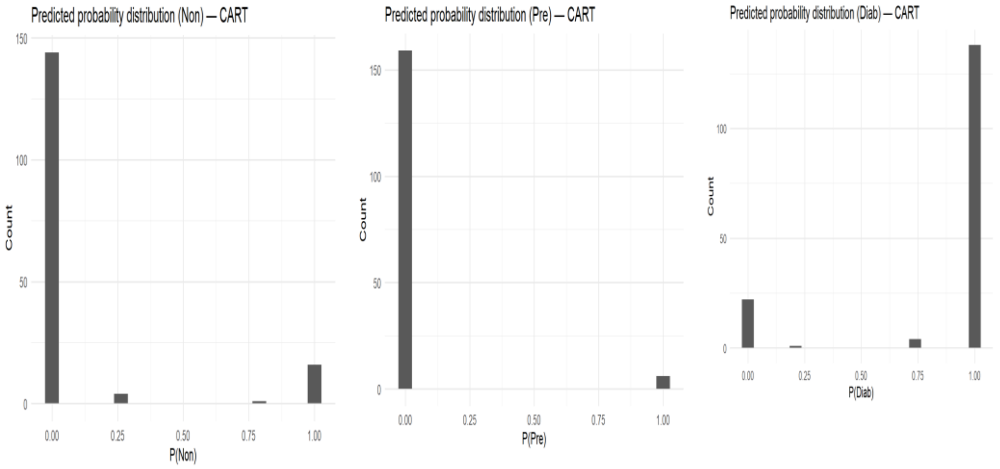


Fig. 6. CART predicted-probability histograms (Non, Pre, Diab)

The three panels in Figure 7 compare binned observed prevalence with mean predicted probability. Because CART assigns sharp, leaf-based probabilities (see Figure 6), only bins near 0 and 1 are populated. For Diab, the high-probability bin lies slightly below the 45° line, indicating mild over-confidence driven by the large HbA1c ≥ 6.5 leaf (Figure 4). For Non, the near-0 and near-1 bins fall close to the diagonal, suggesting good calibration at the extremes; the wider interval around $p \approx 1.0$ reflects limited counts in pure Non leaves. For Pre, the populated low-probability bin sits a little above the diagonal, implying a slight underestimation of Pre risk outside the small Pre leaf. Overall, calibration is reasonable at the extremes but largely undefined in the mid-range, consistent with the concentrated probability histograms in Figure 6 and the threshold structure in Figure 4.

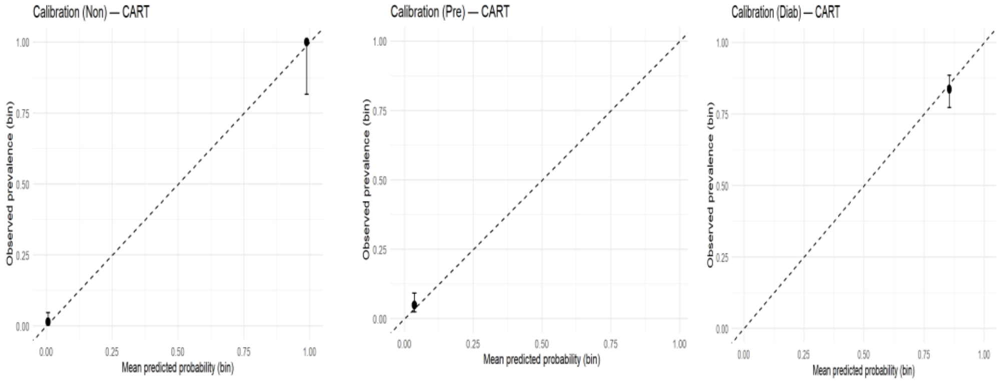
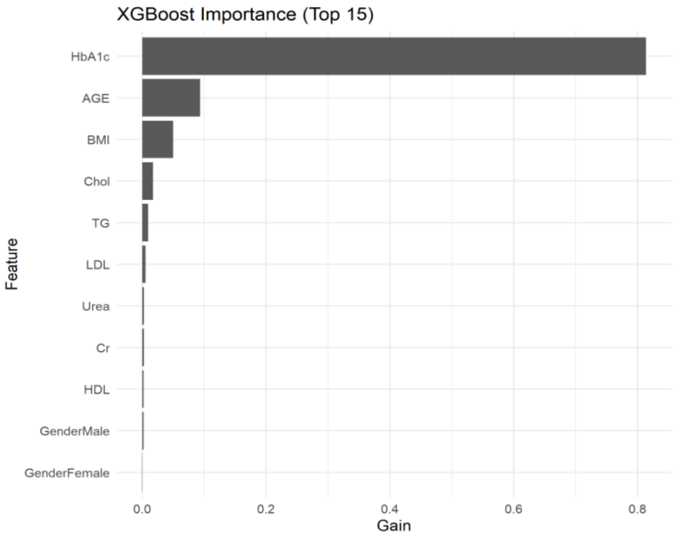


Fig. 7. Calibration by class (CART: Non, Pre, Diab).

Both panels in Figure 8 agree that HbA1c overwhelmingly drives the boosted model. In the gain plot (Top), AGE ranks second and BMI third, with small but non-zero contributions from Chol and TG; LDL, Cr, HDL, and the one-hot gender indicators contribute nothing. In the permutation plot (Bottom), HbA1c remains dominant, but BMI moves into a clear second place, while Chol and TG follow, and AGE drops to a modest effect, showing that, although AGE often appears in splits, BMI and Chol improve class probabilities more when perturbed. This pattern is consistent with Figure 5 for CART (HbA1c first, BMI next, AGE/Chol supporting), reinforcing a coherent story across methods: glycemic status is the primary signal, adiposity provides strong secondary discrimination, and lipid components (especially Chol, then TG) refine probabilities, whereas the remaining variables add little.



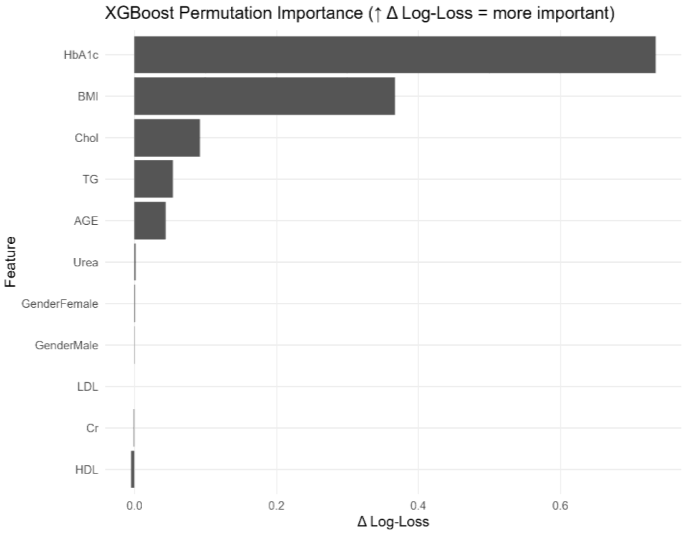


Fig. 8. Top: XGBoost importance (Top-15 by Gain) Bottom: XGBoost permutation importance

The distributions in Figure 9 are strongly bimodal, with most mass near 0 or 1, indicating high separability; compared with CART (Figure 6), XGBoost shows a slightly wider mid-range spread, which is typical for a softmax ensemble. The Diab panel has a dominant spike at 1.00, reflecting many confidently diabetic cases and aligning with HbA1c's leading influence in the importance plots (Figure 8). The Non and Pre panels concentrate near 0 with smaller spikes near 1.00, showing fewer regions that strongly support those classes. Overall, the histograms suggest confident assignments while motivating a check of probability calibration in the mid-probability range.

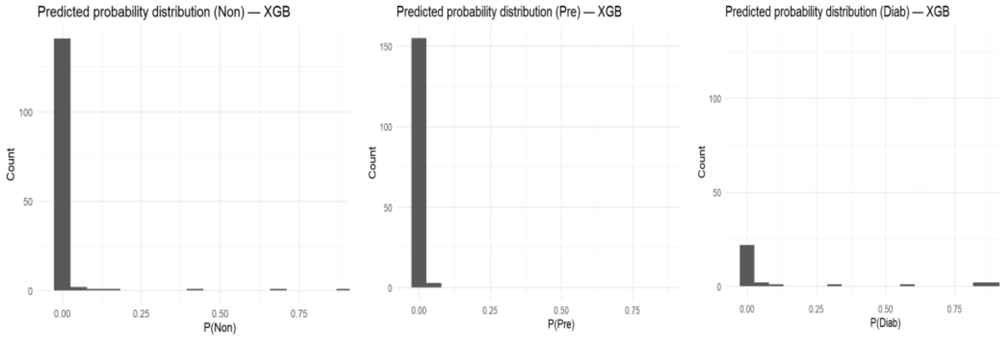


Fig. 9. XGBoost predicted-probability histograms (Non, Pre, Diab)

In Figure 10, using one-vs-rest evaluation, the curves hug the upper-left corner with AUCs of 0.996 (Diab), 1.000 (Non), and 0.992 (Pre), indicating near-perfect discrimination. The perfect AUC for Non likely reflects a small, very separable subset in the test split, which agrees with the strongly bimodal probability histograms in Figure 9 and the dominance of HbA1c in the XGBoost importance profiles (Figure 8). Practically, thresholds can be set to prioritize sensitivity or specificity without large losses in the other. Because Non and Pre have limited test counts, it is advisable to report confidence intervals and also consider precision–recall summaries for completeness.

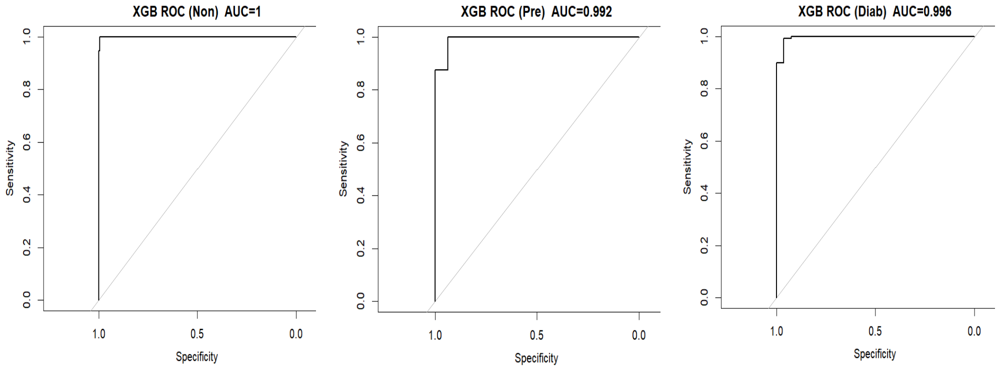


Fig. 10. ROC curves (XGBoost: Non, Pre, Diab)

In Figure 11, the Diab curve (blue) stays near precision ≈1.0 across almost the full recall range, indicating near-perfect positive predictive performance, consistent with the spike at P(Diab)=1.00 in Figure 9 and the very high ROC AUC in Figure 10. The Non curve (red) is also close to the top edge with only a slight drop at extreme recall, matching its near-perfect ROC and the bimodal probability distribution. In contrast, the Pre curve (green) shows markedly lower precision at high recall and greater variability, reflecting the small Pre sample and weaker separability noted in the histograms. Overall, PR confirms that XGBoost is extremely reliable for identifying Diab (and strong for Non), while Pre remains the challenging class; threshold selection should therefore prioritize recall for Pre if missing PreDiabetes is costly.

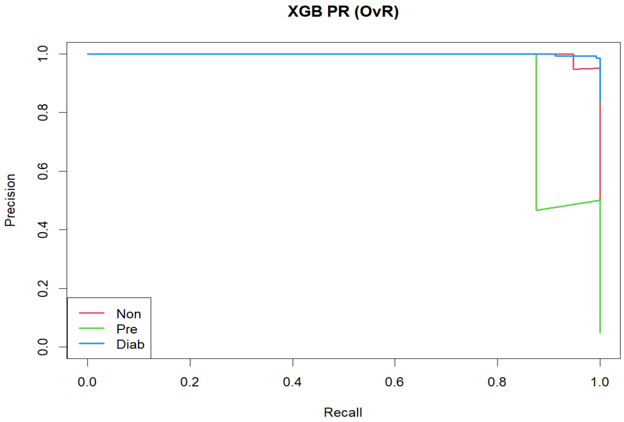


Fig. 11. Precision–Recall curves (by class)

Figures 7 (CART calibration) and 12 (XGB calibration) show that both models are well calibrated at the extremes (near 0 and near 1), but they differ in the mid-probability range. CART produces almost no mid-range probabilities for the Pre class, so its calibration there is effectively degenerate, while Non and Diab sit close to the identity line at the ends. XGB, by contrast, yields meaningful bins across the range for all classes: for Diab and Non, the high-probability bins lie near the identity line, and the mid bin (≈0.45) has a wider interval but a mean close to the line; for Pre, the mid bin is also near the line, indicating useful probabilistic calibration. Figure 13 confirms this: both models align at ≈0 and ≈1, but only XGB provides a credible mid-probability bin, whereas CART collapses to 0/1. This is consistent with the probability histograms (Figures 6 and 9), where CART concentrates mass at the extremes, and XGB still allows informative mid-range probabilities. In short, when calibrated probabilities across the full range matter, especially for Pre, XGB is preferable; if decisions rely only on extreme probabilities, both models are adequate.

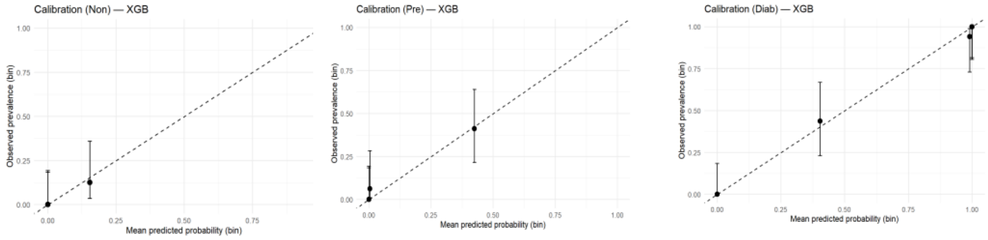


Fig. 12. Calibration by class (XGBoost: Non, Pre, Diab)

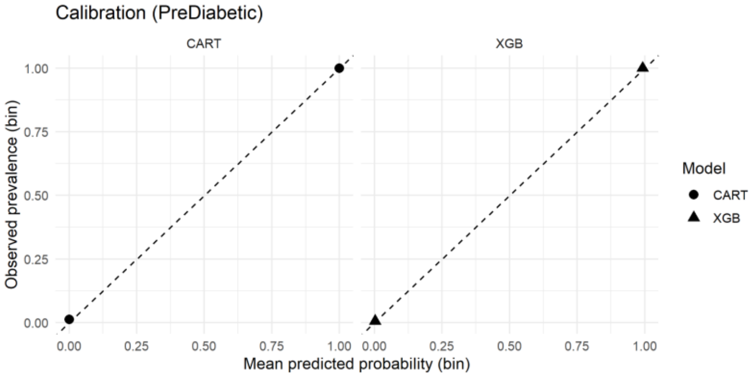


Fig. 13. Joint calibration (PreDiabetic: CART vs XGBoost).

VI. Evaluation of the Models

Across figures 1–13, the evidence is consistent: glycemic status dominates discrimination, adiposity is a strong secondary signal, and lipids refine decisions. The correlation heatmap and per-class distributions (Figures 1–2) show limited problematic collinearity (notably Urea–Cr and Chol–LDL) and clear class shifts for HbA1c and BMI. Distributional diagnostics (Figure 3) reveal right-skew and heavy tails, especially in Diab, supporting the use of tree-based models. The learned CART rules (Figure 4) translate these patterns into simple thresholds, while variable-importance profiles for CART and XGB (Figures 5 and 8) concur that HbA1c is the primary driver, BMI next, with AGE/Chol providing additional refinement and other variables minor. Probability histograms confirm the different probability behaviors: CART concentrates mass at 0/1 (Figure 6), whereas XGB remains sharply bimodal but allows a slightly wider mid-range (Figure 9). Calibration mirrors this: CART is well aligned at the extremes but sparse in the mid-range (Figure 7), while XGB is close to the identity line across bins for all classes (Figure 12); their direct comparison for the Pre class (Figure 13) highlights XGB's advantage when mid-probabilities are needed.

Discrimination metrics favor XGB. One-vs-rest ROC curves (Figure 10) are near the upper-left corner for every class (AUC ≈ 0.99–1.00), indicating excellent ranking; precision–recall (Figure 11) shows Diab and Non maintaining precision near 1.0 across high recall, with Pre remaining the hardest class but still improved relative to CART (in line with Figures 6 and 9). Taken together, the importances (Figures 5 and 8), probability shapes (Figures 6 and 9), and calibration (Figures 7, 12–13) form a coherent picture: both models separate Diab extremely well; XGB additionally yields more reliable probabilities across the range, which is valuable for thresholding decisions and risk communication.

The summary tables confirm the graphical findings. In Table 4 (test set), XGB surpasses CART on all five criteria: higher Accuracy, Balanced Accuracy, Macro-F1, and Macro-AUPRC, and a lower Brier score, indicating better overall correctness, more balanced recognition of minority classes, superior ranking quality, and more faithful probabilities.

Table 4. Test metrics summary (Accuracy, Balanced Accuracy, Macro-F1, Macro-AUPRC, Brier)

Model	Accuracy	Balanced Accuracy	MacroF1	Macro AUPRC	Brier
CART	0.9758	0.8816	0.9291	0.9214	0.0399
XGBoost	0.9818	0.9384	0.9566	0.9750	0.0226

Table 5 shows where the gains occur: XGB raises sensitivity for Non and Pre while keeping Diab sensitivity very high and preserves excellent specificity; this aligns with the ROC/PR advantages and the improved calibration bins. Therefore, for the study goal of assigning patients to Non, Pre, or Diab, XGB is the preferred model on this dataset. CART remains useful for transparent clinical rules, but XGB offers a stronger operating point when both accuracy and well-calibrated probabilities, especially for Pre, are required.

Table 5: Test metrics summary by class

	CART			XGBoost		
	NonDiabetic	**PreDiabetic**	**Diabetic**	**NonDiabetic**	**PreDiabetic**	**Diabetic**
Sensitivity	0.8947	0.75000	1.0000	0.9474	0.87500	0.9928
Specificity	1.0000	1.00000	0.8519	0.9932	1.00000	0.9259
Pos Pred Value	1.0000	1.00000	0.9718	0.9474	1.00000	0.9856
Neg Pred Value	0.9865	0.98742	1.0000	0.9932	0.99367	0.9615
Prevalence	0.1152	0.04848	0.8364	0.1152	0.04848	0.8364
Detection Rate	0.1030	0.03636	0.8364	0.1091	0.04242	0.8303
Detection Prevalence	0.1030	0.03636	0.8606	0.1152	0.04242	0.8424
Balanced Accuracy	0.9474	0.87500	0.9259	0.9703	0.93750	0.9593
Kappa	0.9091			0.9351		

VII. Conclusion

This study assessed a cost sensitive CART and a multiclass XGBoost on a three class diabetes outcome drawn from two Iraqi hospitals, where the data present multiple real world difficulties including frequent outliers, skewed distributions with heavy upper tails, non homogeneous covariance across classes, and pronounced imbalance in class sizes; across the evidence summarized in the Results with thirteen figures and in Tables 4 and 5, XGBoost provided stronger overall performance together with more informative probability estimates for the minority classes, while CART delivered transparent rule based decisions that remain attractive when simplicity and interpretability are prioritized, therefore the boosted model is recommended as the default analytical choice for this dataset and the tree remains a practical alternative in settings that require fast, human readable rules. The patterns observed across correlation structure, per class distributions, and fitted model behavior form a coherent narrative for the reader in which variables reflecting glycemic burden dominate discrimination, adiposity adds a secondary signal, and selected lipid measures provide further refinement, which explains the very high separability of the Diabetic class and the improved recognition of the Non diabetic and Pre diabetic classes under the boosted model; the probability histograms show that the tree concentrates mass near zero and one which favors crisp decisions but limits the mid-range, whereas the boosted model produces a smoother distribution of probabilities that supports threshold tuning and communication of graded risk, and because class sizes are strongly imbalanced the text emphasizes macro averaged summaries to prevent the majority class from dominating aggregates while precision and recall for the minority classes are interpreted alongside specificity so that operating points can be selected according to clinical priorities. The dataset originates from two hospitals in Iraq, which introduces mild site level heterogeneity in measurement practice and case mix and provides a realistic setting for multiclass clinical prediction under distributional stress; the boosted model's advantages across accuracy, balanced accuracy, macro F1, macro AUPRC, and the Brier score align with its more informative probabilities in the mid-range and suggest resilience to modest between site variability, although external validation and periodic recalibration remain necessary to account for temporal shifts in prevalence and workflow, and subgroup checks should be reported to ensure that gains are consistent across demographic and clinical strata. Limitations include the retrospective design in a single city with two hospitals, the small absolute size of the Pre diabetic test subset, which increases uncertainty in its precision and recall behavior, and reliance on routinely available predictors rather than longitudinal or specialized markers that could further sharpen discrimination; therefore, future work should test the models on external cohorts, update or recalibrate probabilities to local prevalence, add decision analytic evaluation that aligns threshold choices with clinical costs and benefits, and include fairness oriented reporting that examines stability of performance across clinically meaningful subgroups. Post hoc sensitivity analysis, motivated by clinical redundancy among lipid and renal markers (LDL can be estimated from Chol/HDL/TG; creatinine covaries with Urea), re-estimated performance after excluding LDL and creatinine. CART was unchanged (Accuracy 0.9758; Balanced Accuracy 0.8816), whereas XGBoost improved modestly (Accuracy 0.9818 → 0.9879; Balanced Accuracy 0.9384 → 0.9559). These changes do not alter our conclusions: the boosted model remains preferred and removing LDL/creatinine preserves, or slightly enhances, discrimination while yielding a more parsimonious predictor set.

Conflict of Interest:

There was no relevant conflict of interest regarding this paper.

References

I. American Diabetes Association Professional Practice Committee. 2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes, 2024. Diabetes Care. 2024;47(Suppl 1):S20–S42. doi:10.2337/dc24-S002.

II. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Monterey, CA: Wadsworth. (Reprinted by Chapman & Hall/CRC, Taylor & Francis, 2017). 10.1201/9781315139470

III. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

IV. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–794. doi:10.1145/2939672.2939785.

V. Duncan BB, et al. IDF Diabetes Atlas 11th edition 2025: global prevalence and key metrics. Nephrology Dialysis Transplantation. 2025. doi:10.1093/ndt/gfaf177.

VI. Friedewald, W. T., Levy, R. I., & Fredrickson, D. S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry, 18*(6), 499–502. 10.1093/clinchem/18.6.499

VII. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. **10.1214/aos/1013203451**

VIII. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. 10.1016/j.patrec.2005.10.010

IX. Guo, R., Liu, J., Yang, X., Wang, X., Xu, Z., & Wu, K. (2025). Enhance health evidence quality in classification tasks. *Digital Health*, 11, 20552076251314097. 10.1177/20552076251314097

X. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. 10.1007/978-0-387-84858-7

XI. International Diabetes Federation (IDF). IDF Diabetes Atlas, 11th edition (global facts and projections). Available at: https://diabetesatlas.org/data-by-location/global/ and Global Factsheet PDF (2025).

XII. Idhom, M., Fauzi, A., Muhaimin, A., & Caesarendra, W. (2025). Evaluation of CART and XGBoost Methods on Customer Loan Risk Prediction Based on Consumer Behavior. *TEM Journal*, 14(3), 2624–2630. 10.18421/TEM143-64

XIII. Nuankaew, P.; Chaising, S.; Temdee, P. (2021). Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction. IEEE Access, 9, 137015–137028. 10.1109/ACCESS.2021.3117269

XIV. Rashid, Ahlam (2020), “Diabetes Dataset”, Mendeley Data, V1, 10.17632/wj9rwkp9c2.1

XV. Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14, 6086. 10.1038/s41598-024-56706-x

XVI. Sabariah, M. K.; Hanifa, A.; Sa'adah, S. (2014). Early detection of Type II Diabetes Mellitus with Random Forest and Classification and Regression Tree (CART). In: Proceedings of the 2014 International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), 238–242. IEEE. 10.1109/ICAICTA.2014.7005947

XVII. Sahid, M. A.; Babar, M. U. H.; Uddin, M. P. (2024). Predictive modeling of multi-class diabetes mellitus using machine learning and filtering Iraqi diabetes data dynamics. PLOS ONE, 19(5), e0300785. 10.1371/journal.pone.0300785

XVIII. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. 10.1016/j.ipm.2009.03.002