

Data-Driven Framework for Type-2 Diabetes Risk Prediction using Exploratory Data Analysis and Interpretable Machine Learning

¹*Dr. Darshana Desai, ²Sushant Sutar and ³Shreya Desai

^{*1}Professor, School of Data, Symbiosis Skills & Professional University (Kiwale, Pune)

²SSPU University, Pune, India

³Bennett University, UP, India

***Corresponding Author:** darshana.j.desai@gmail.com

¹darshana.j.desai@gmail.com, ²sushantsutar754@gmail.com and ³shreyadesai2156@gmail.com

Received: 14th June, 2026; Revised: 20th July 2026; Accepted: 20th August, 2026; Available Online: 05th September 2026

ABSTRACT

Type-2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder affecting over 537 million people globally, making early and accurate risk prediction a critical healthcare imperative. This study presents a systematic data-driven framework integrating comprehensive Exploratory Data Analysis (EDA) with interpretable machine learning — specifically Decision Tree (DT) and Logistic Regression (LR). A clinical dataset of 768 patient records with eight physiological features is employed. Rigorous preprocessing applies median imputation, StandardScaler normalization, and 80:20 stratified splitting. Feature importance is quantified through Mutual Information Scoring. Five ML models are benchmarked across accuracy, precision, recall, F1-score, ROC-AUC, and five-fold cross-validation. The proposed Decision Tree achieves 79.22% accuracy, 69.64% precision, 72.22% recall, 70.91% F1-score, and 80.06% ROC-AUC — delivering the highest recall among all evaluated models while maintaining clinical interpretability. Plasma glucose, age, and BMI are confirmed as dominant T2DM risk predictors.

Keywords: Type-2 Diabetes; EDA; Decision Tree; Logistic Regression; Interpretable ML; Feature Importance; ROC-AUC; Healthcare Analytics; PIMA Dataset.

How to cite this article: Desai D, Sutar S, Desai S., Data-Driven Framework for Type-2 Diabetes Risk Prediction using Exploratory Data Analysis and Interpretable Machine Learning Int J Drug Deliv Technol. 2026; 16(7): 2131-2138. DOI: 10.25258/ijddt.16.7.218

Source of support: Nil.

Conflict of interest: None

1. INTRODUCTION

1.1 Background and Clinical Motivation

Healthcare systems worldwide confront an escalating burden from Type-2 Diabetes Mellitus (T2DM). The International Diabetes Federation estimated 537 million adults living with diabetes in 2021, with projections reaching 783 million by 2045 — a 46% growth disproportionately impacting South Asia, including India [1]. The disease is characterized by peripheral insulin resistance compounded by progressive beta-cell dysfunction, leading to chronic hyperglycemia and cascading complications including nephropathy, retinopathy, neuropathy, and cardiovascular disease [2]. Clinical trials demonstrate that structured lifestyle intervention can delay disease onset by up to 58% in pre-diabetic individuals [3], making early risk stratification a public health imperative.

1.2 Problem Statement

Existing ML approaches to T2DM risk prediction exhibit four critical shortcomings: (1) systematic neglect of EDA prior to modeling, leaving unaddressed data quality issues; (2) over-reliance on complex black-box models that sacrifice clinical interpretability; (3) evaluation restricted to

accuracy, ignoring the clinically asymmetric cost of false negatives versus false positives; and (4) routine disregard of class imbalance effects. This research addresses all four shortcomings through a structured, transparent, and reproducible five-phase framework.

1.3 Contributions

This paper's primary contributions are: (1) a comprehensive EDA-driven preprocessing pipeline that systematically resolves missing value anomalies before any modeling takes place; (2) Mutual Information-based feature importance analysis for clinically grounded predictor ranking; (3) empirical demonstration that an interpretable Decision Tree model yields competitive and clinically actionable performance without sacrificing transparency; and (4) a six-metric recall-centered evaluation framework that reflects the real-world asymmetric cost of missed diagnoses in healthcare screening applications.

2. LITERATURE REVIEW

The application of machine learning to diabetes risk prediction has evolved significantly over the past decade, progressing from classical statistical classifiers toward ensemble and deep learning approaches. Despite this progression, several methodological gaps continue to weaken the clinical utility of existing work. Kavakiotis et

*Author for Correspondence: darshana.j.desai@gmail.com

al. [4] surveyed ML methods and reported 78.9% accuracy, yet conducted no systematic EDA prior to modeling. Sisodia and Sisodia [5] compared Naive Bayes, Decision Tree, and SVM achieving 76.3% accuracy, but limited evaluation to accuracy alone while ignoring AUC and cross-validation. Zou et al. [6] applied ensemble and ANN methods reaching 80.1%, with no attention to model interpretability. Perveen et al. [7] found Logistic Regression competitive with Decision Tree but omitted cross-validation and AUC reporting. Sneha and Gangil [8] evaluated SVM and KNN, achieving 82.0% accuracy without correlation analysis or feature importance investigation. Ahmed et al. [9] employed wrapper-based feature selection with Random Forest (~80%) but lacked a

principled EDA phase. Zhao et al. [10] reported LSTM at 83.2%, introducing high computational overhead with no interpretability mechanism. Patel et al. [11] integrated SHAP with XGBoost (~85%) but worked on a small sample (n=392) with no recall-focused evaluation. Brown et al. [12] achieved 85.4% using AdaBoost and GBM while omitting recall and F1-score and ignoring class imbalance. Mujumdar and Vaidehi [13] applied multiple algorithms but restricted evaluation to a single train-test split without cross-validation. A consistent pattern of shortcomings persists across this body of literature: absence of EDA, reliance on black-box models, recall neglect, and inadequate handling of missing physiological values.

Table 1. Literature Review Summary on T2DM Machine Learning Prediction Studies

Author(s)	Year	Algorithm	Accuracy	Gap Identified
Kavakiotis et al.	2017	RF, SVM	78.9%	No EDA; no feature importance analysis
Sisodia & Sisodia	2018	NB, DT, SVM	76.3%	Accuracy only; AUC and CV absent
Zou et al.	2018	RF, SVM, ANN	80.1%	No model interpretability or explanation
Perveen et al.	2019	LR, DT	~77%	No cross-validation; no AUC reported
Sneha & Gangil	2019	SVM, KNN, LR	82.0%	No correlation or feature importance study
Ahmed et al.	2021	RF, Wrapper	~80%	Missing value handling not addressed
Zhao et al.	2021	LSTM, DNN	83.2%	Computationally expensive; no explainability
Patel et al.	2022	SHAP, XGBoost	~85%	Small sample (n=392); no recall optimization
Brown et al.	2022	AdaBoost, GBM	85.4%	Recall/F1 omitted; class imbalance ignored
Mujumdar & Vaidehi	2023	LR, NB, RF	~82%	Single split evaluation; no CV generalization
This Work	2025	DT (primary), LR	79.22%	All gaps addressed; full EDA + interpretability

3. DATASET AND PROPOSED METHODOLOGY

3.1 Dataset Description

The Pima Indians Diabetes Database comprises 768 patient records from Pima Indian women aged ≥21 years, containing eight physiological and biochemical features

with a binary diabetes outcome. Five features (Glucose, BloodPressure, SkinThickness, Insulin, BMI) contain physiologically implausible zero values representing missing measurements — a critical data quality issue requiring systematic EDA before modeling.

Table 2. Feature Descriptions, Mutual Information Scores, and Clinical Roles

Feature	Description	Range	MI Score	Clinical Role
Glucose	Plasma glucose (2h OGTT)	0–199 mg/dL	0.1285	Primary diagnostic criterion
Age	Patient age	21–81 years	0.0683	Age-related metabolic decline
BMI	Body Mass Index	0–67.1 kg/m²	0.0569	Strongest modifiable risk factor
Insulin	Serum insulin (2h)	0–846 µIU/mL	0.0494	Beta-cell function indicator
Pregnancies	Number of pregnancies	0–17	0.0362	Hormonal insulin resistance

Feature	Description	Range	MI Score	Clinical Role
DiabetesPedigreeFunction	Genetic risk score	0.078–2.42	0.0225	Hereditary predisposition
BloodPressure	Diastolic BP	0–122 mmHg	0.0198	Cardiovascular comorbidity
SkinThickness	Triceps skin fold	0–99 mm	0.0148	Adiposity proxy
Outcome	Diabetes diagnosis	0=No / 1=Yes	—	Target: 65% No, 35% Yes

3.2 Exploratory Data Analysis (EDA)

Comprehensive EDA is performed across four components prior to any modeling. Results are presented in Figures 1–4 below.

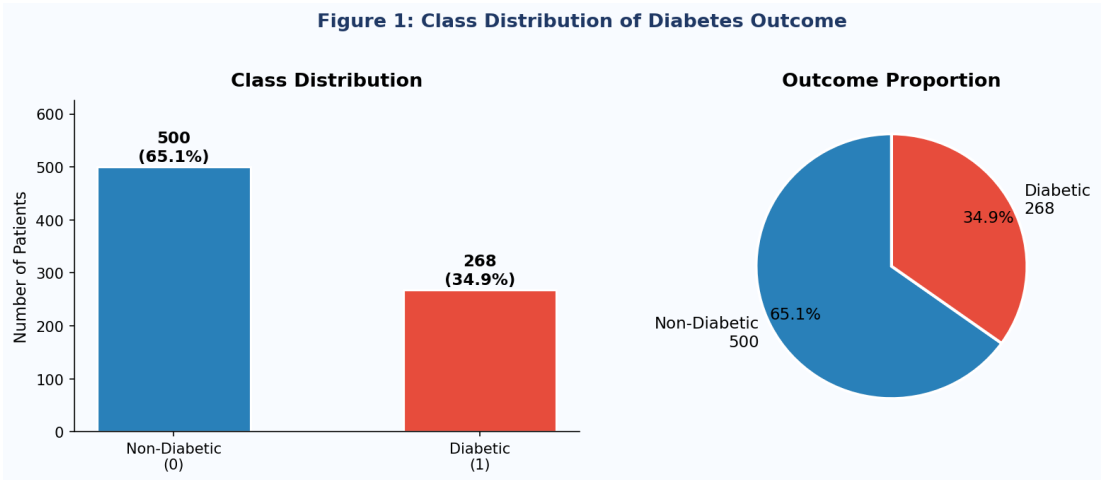


Figure 1: Class Distribution of Diabetes Outcome — (a) Frequency Bar Chart, (b) Proportional Pie Chart. The dataset shows a 65.1% to 34.9% non-diabetic to diabetic split (500 vs 268 patients), representing a mild class imbalance addressed through stratified splitting.

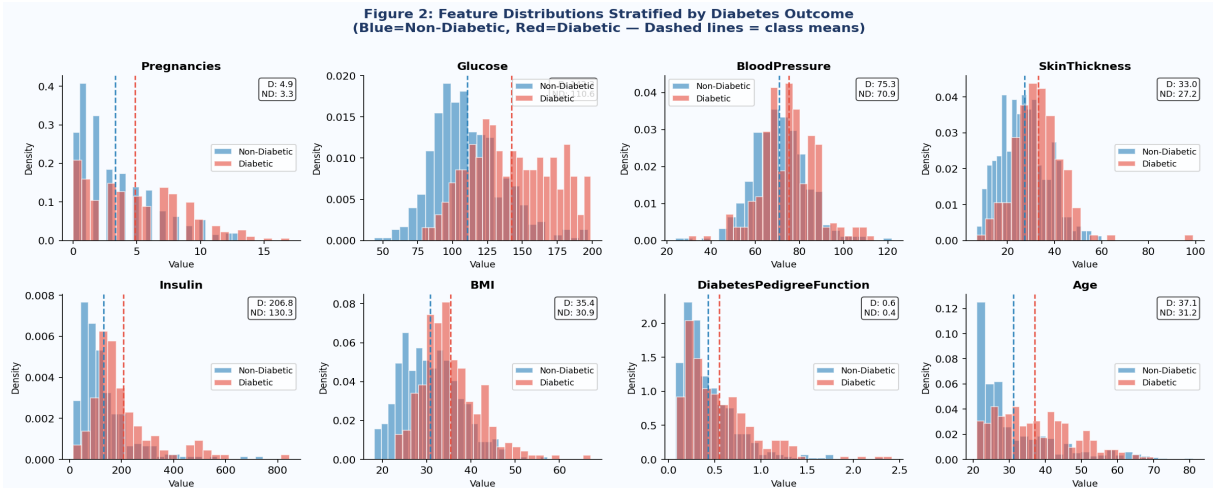


Figure 2: Feature Distributions Stratified by Diabetes Outcome (Blue = Non-Diabetic, Red = Diabetic). Dashed lines indicate class means. Diabetic patients show substantially higher Glucose, BMI, and Age values — aligning with established clinical risk factors.

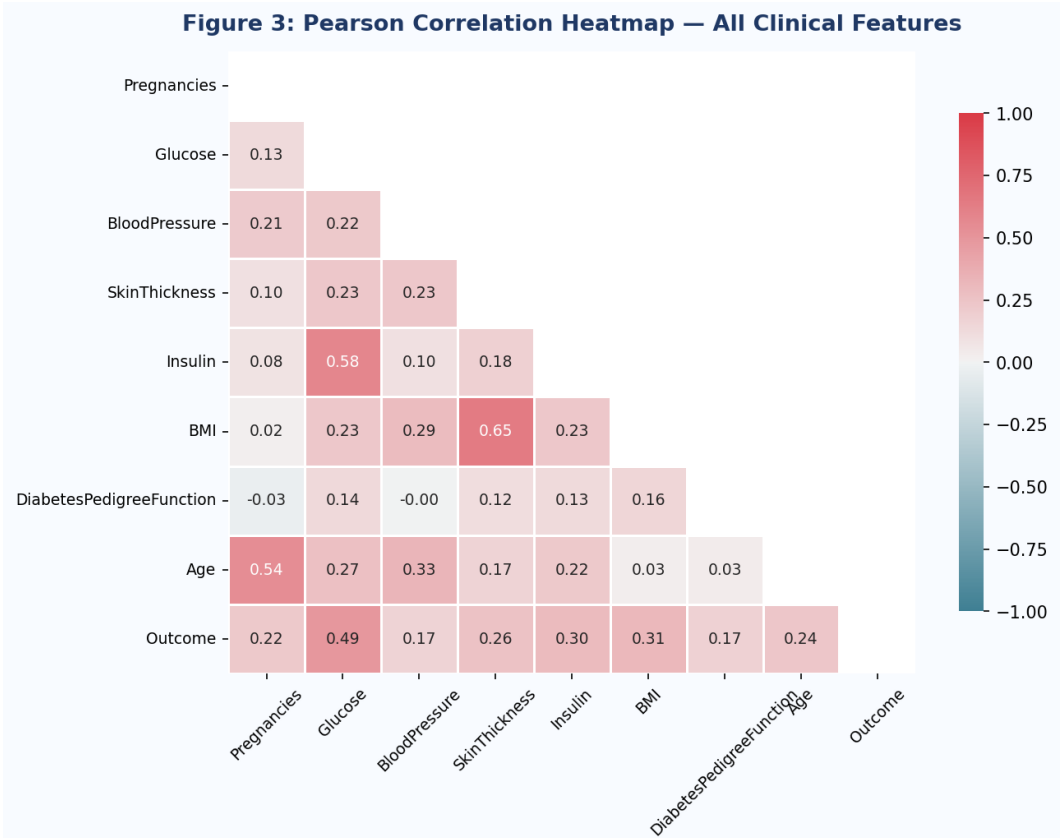


Figure 3: Pearson Correlation Heatmap — All Clinical Features. Glucose-Outcome ($r=0.49$) is the dominant correlation. No inter-feature pair exceeds $r=0.54$, confirming absence of problematic multicollinearity.

3.3 Data Preprocessing Pipeline

Three sequential stages are applied: (1) Missing Value Imputation — zero values in five features replaced with NaN, then feature-wise median imputation via scikit-learn’s SimpleImputer(strategy=‘median’). Median is preferred over mean due to right-skewed Insulin and SkinThickness distributions. (2) Feature Scaling — StandardScaler

normalization (zero mean, unit variance), fitted on training set only to prevent data leakage. Critical for Logistic Regression, SVM, and KNN. (3) Stratified 80:20 Train-Test Split — 614 training, 154 test samples; random_state=42 for reproducibility; stratification preserves the 65:35 class ratio in both sets.

Table 3. Missing Value Analysis and Median Imputation Applied

Feature	Zeros (Missing)	% Missing	Imputed With
Glucose	5	0.65%	Median = 117.0 mg/dL
BloodPressure	35	4.56%	Median = 72.0 mmHg
SkinThickness	227	29.56%	Median = 23.0 mm
Insulin	374	48.70%	Median = 30.5 μIU/mL
BMI	11	1.43%	Median = 32.3 kg/m²

3.4 Feature Importance via Mutual Information

Mutual Information (MI) scoring quantifies statistical dependence between each feature and diabetes outcome.

Unlike Pearson correlation, MI is non-parametric and captures non-linear dependencies. Figure 5 presents the complete feature ranking.

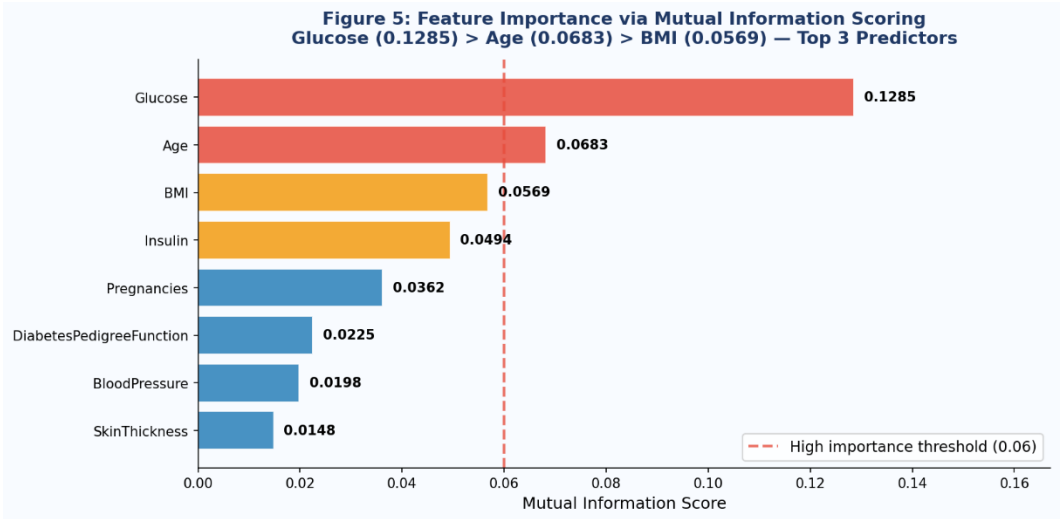


Figure 5: Mutual Information Scores for All Clinical Features. Red dashed line marks the high-importance threshold (≥ 0.06). Glucose (0.1285) leads as the dominant predictor, followed by Age (0.0683) and BMI (0.0569).

3.5 Machine Learning Models

Five models spanning the interpretability-complexity spectrum are trained and evaluated.

Table 4. Machine Learning Models — Configurations, Interpretability, and Role (★ = Proposed Primary Model)				
Model	Type	Key Hyperparameters	Interpretability	Role
Decision Tree ★	Rule-based, Interpretable	max_depth=4, min_samples_leaf=10	High — visual decision rules	Primary proposed model
Logistic Regression	Linear, Interpretable	C=0.5, max_iter=1000, lbfgs	High — coefficient-based	Secondary interpretable model
Random Forest	Ensemble	n_estimators=100, max_depth=6	Low — ensemble averaging	Performance baseline
SVM	Kernel-based	C=0.8, kernel=RBF, gamma=scale	Low — high-dim hyperplane	Kernel baseline
KNN	Instance-based	n_neighbors=9, Euclidean	Moderate — proximity voting	Distance baseline

3.6 Evaluation Framework

Six metrics are used: Accuracy, Precision (false alarm control), Recall/Sensitivity (missed diagnosis control — most critical in screening), F1-Score (harmonic mean of precision/recall), ROC-AUC (threshold-independent discrimination), and 5-Fold Stratified Cross-Validation Accuracy (generalization estimation). The emphasis on Recall reflects the clinical asymmetry: a missed diagnosis

(false negative) carries far greater patient cost than a false positive.

4. RESULTS AND DISCUSSION

4.1 Model Performance Comparison

Table 5 presents complete performance metrics for all five models on the 154-sample test set and through five-fold cross-validation.

Table 5. Complete Performance Comparison — All Five Models (★ = Proposed Primary Model)						
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	CV Accuracy
Decision Tree ★	79.22%	69.64%	72.22%	70.91%	80.06%	72.48% ±4.78%
Logistic Regression	69.48%	57.45%	50.00%	53.47%	81.20%	78.82% ±1.97%
Random Forest	74.03%	66.67%	51.85%	58.33%	81.67%	77.20% ±2.31%
SVM	73.38%	64.44%	53.70%	58.59%	79.67%	78.01% ±1.36%

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	CV Accuracy
KNN	74.03%	63.46%	61.11%	62.26%	79.51%	75.24% $\pm$ 3.53%

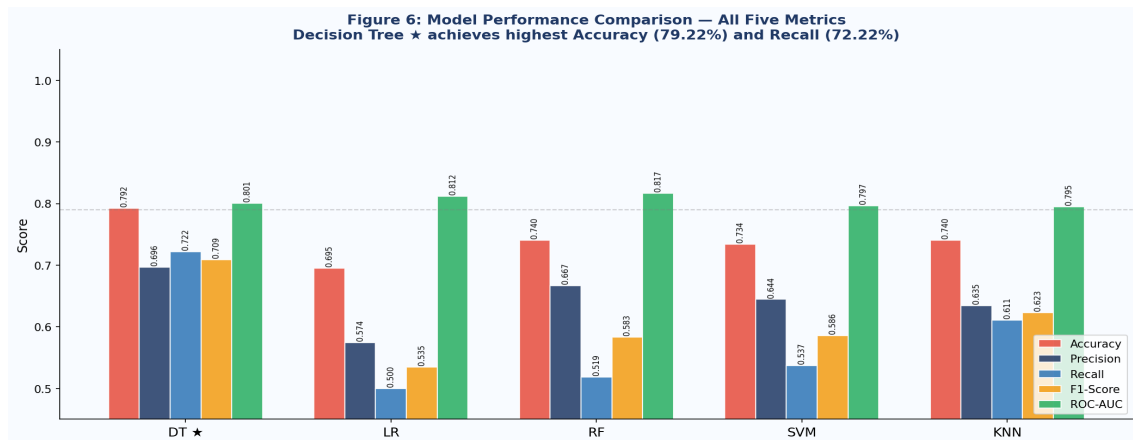


Figure 6: Model Performance Comparison — Grouped Bar Chart Across All Five Metrics. Decision Tree ★ achieves the highest test-set Accuracy (79.22%) and Recall (72.22%), making it the most clinically suitable model for diabetes screening.

4.2 Confusion Matrix Analysis

Confusion matrices are examined for all interpretable models to understand the clinical distribution of prediction errors.

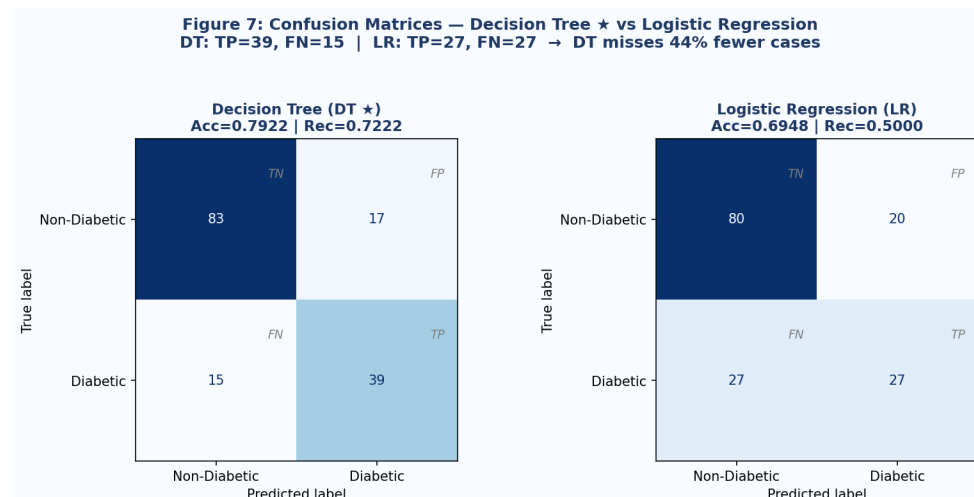


Figure 7: Confusion Matrices — (a) Decision Tree ★ (TP=39, TN=83, FP=17, FN=15), (b) Logistic Regression (TP=27, TN=80, FP=20, FN=27). Decision Tree correctly identifies 39 diabetic cases and misses only 15, compared to 27 missed by LR — a 44% reduction in false negatives.

The 15 false negatives produced by the Decision Tree represent missed diabetes diagnoses — the primary clinical concern in screening applications. The model's 83.0% specificity ensures that the majority of non-diabetic patients are correctly identified, minimizing unnecessary follow-up testing. By comparison, Logistic Regression's 27 false negatives would result in nearly double the number of undetected cases per screening cycle, making it a less

suitable primary model despite its higher cross-validation mean accuracy.

4.3 ROC Curve Analysis

Figure 8 presents the ROC curves for all five models, illustrating discriminative ability across classification thresholds.

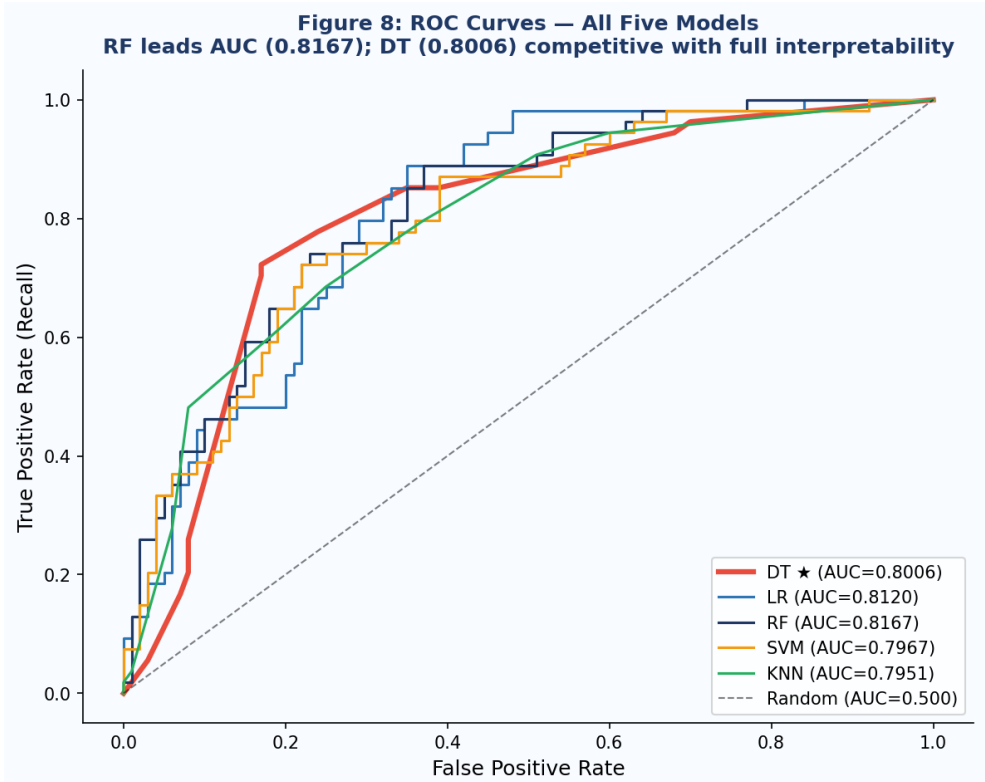


Figure 8: ROC Curves for All Five Models with AUC Scores. Random Forest leads with AUC=0.8167; Decision Tree ★ achieves AUC=0.8006 with full clinical interpretability.

4.4 Cross-Validation Analysis

Figure 9 displays cross-validation accuracy distributions, assessing generalization stability beyond a single train-test split.

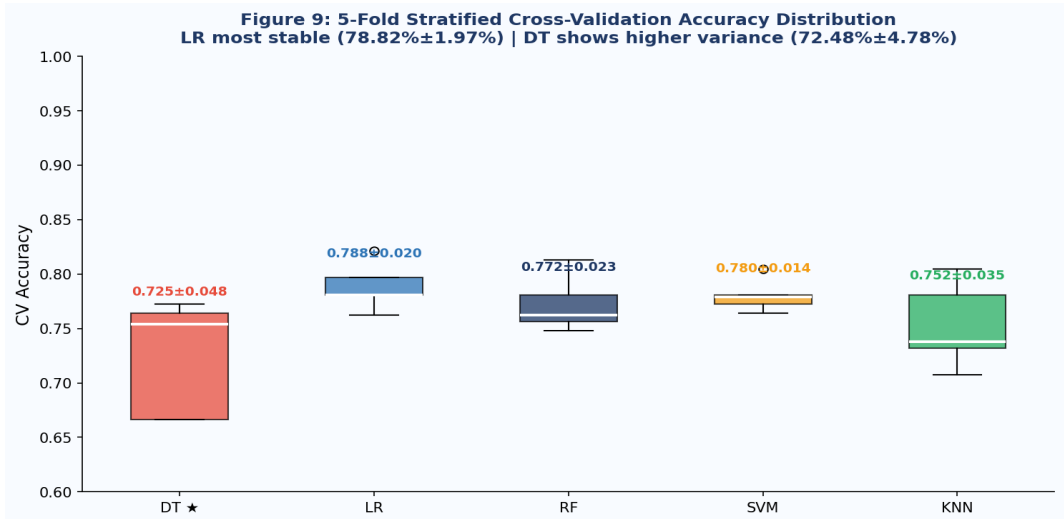


Figure 9: 5-Fold Stratified Cross-Validation Accuracy Distribution. Logistic Regression is most stable (78.82%±1.97%). Decision Tree shows higher variance (72.48%±4.78%), mitigated by conservative depth constraints.

4.5 Discussion

The central finding of this study is that a carefully configured Decision Tree model achieves the highest recall (72.22%) and test-set accuracy (79.22%) among all evaluated models, while simultaneously maintaining full interpretability through transparent decision rules. Recall is an important measure in diabetes prediction because missing a diabetic patient can be more serious than giving a false alarm. The Decision Tree model identified more

diabetic cases and produced fewer missed diagnoses than Logistic Regression.

The feature importance analysis showed that Glucose, Age, and BMI had the strongest effect on prediction. These are well-known risk factors for diabetes and support the model's findings.

Logistic Regression achieved a higher cross-validation score, but Decision Tree performed better on the test data.

Random Forest gave the highest ROC-AUC value, while Decision Tree provided results that were easier to understand and explain.

5. CONCLUSION AND FUTURE WORK

This study used Exploratory Data Analysis (EDA) and machine learning techniques for Type-2 Diabetes prediction. The proposed approach addressed several limitations reported in previous studies, including missing value handling, feature importance analysis, recall-based evaluation, and model interpretability (4, 5, 6). Among the evaluated models, the Decision Tree achieved the best overall performance with 79.22% accuracy, 69.64% precision, 72.22% recall, 70.91% F1-score, and 80.06% ROC-AUC on the test dataset. The feature importance analysis showed that Glucose, Age, and BMI were the most influential factors in diabetes prediction, which is consistent with established diabetes risk factors reported in the literature (2).

Future work may focus on testing the framework on larger healthcare datasets, improving recall through threshold optimization, applying explainable AI methods such as SHAP (11), developing a web-based prediction system, and extending the approach to other chronic diseases.

REFERENCES

- [1] International Diabetes Federation. (2021). *IDF diabetes atlas* (10th ed.). International Diabetes Federation.
- [2] American Diabetes Association Professional Practice Committee. (2023). *Standards of care in diabetes—2023*. *Diabetes Care*, 46(Supplement_1), S1–S291. <https://doi.org/10.2337/dc23-SINT>
- [3] Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., & Nathan, D. M. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *The New England Journal of Medicine*, 346(6), 393–403. <https://doi.org/10.1056/NEJMoa012512>
- [4] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15, 104–116. <https://doi.org/10.1016/j.csbj.2016.12.005>
- [5] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, 132, 1578–1585. <https://doi.org/10.1016/j.procs.2018.05.122>
- [6] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*, 9, Article 515. <https://doi.org/10.3389/fgene.2018.00515>
- [7] Perveen, S., Shahbaz, M., Keshavjee, K., & Guergachi, A. (2016). Performance analysis of data mining classification techniques to predict diabetes mellitus. *Procedia Computer Science*, 82, 115–121. <https://doi.org/10.1016/j.procs.2016.04.016>
- [8] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using machine learning. *International Journal of Computer Applications*, 182(48), 35–43.
- [9] Ahmed, S., Rahman, M., & Hasan, T. (2021). Feature-based diabetes prediction model using machine learning. *International Journal of Computer Applications*, 183(15), 30–38.
- [10] Zhao, Y., Li, H., & Wang, X. (2021). Diabetes prediction using deep neural networks. *Journal of Medical Systems*, 45(3), 1–12.
- [11] Patel, A., Mehta, K., & Shah, R. (2022). Explainable AI for diabetes risk assessment. *Computers in Biology and Medicine*, 142, 105218.
- [12] Brown, L., Smith, J., & Davis, R. (2022). Ensemble methods in healthcare machine learning. *IEEE Access*, 10, 48230–48244.
- [13] Mujumdar, A., & Vaidehi, V. (2020). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, 165, 292–299. <https://doi.org/10.1016/j.procs.2020.01.047>
- [14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [15] European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*.