Vol. 12 (2026): Proceedings of the 2026 International Conference on Life Sciences and Health Engineering (IC-LSHE 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/lshe.20260106

Comparative Study of Early Diabetes Risk Stratification Based on Machine Learning Algorithms

Tianze Zhang*

School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China

**Corresponding author: Tianze Zhang*

Abstract

Objective: This study aims to systematically compare the performance of multiple machine learning models in diabetes risk prediction and identify key risk factors, thereby providing data-driven decision support for early diabetes screening. **Methods:** Using the UCI Pima Indians Diabetes dataset, five models-logistic regression, K-nearest neighbors, support vector machine, decision tree, and random forest-were trained and evaluated. Model performance was comprehensively assessed via metrics including AUC-ROC, precision, and recall, with feature importance analysis employed to elucidate core diabetes risk factors. **Results:** The random forest model demonstrated superior performance across multiple metrics (AUC = 0.8167). Plasma glucose was consistently identified as the strongest predictor, with body mass index (BMI) and age also emerging as significant contributors. **Conclusion:** The random forest model exhibits robust performance and effective capture of feature interactions, making it well-suited for early diabetes prediction with considerable potential for clinical application.

Keywords

diabetes prediction, machine learning, feature importance, random forest, model comparison

1. Introduction

1.1 Diabetes Background

Diabetes, a chronic metabolic disorder affecting populations worldwide, impacted 537 million adults in 2021 according to the International Diabetes Federation (IDF), with projections indicating a rise to 783 million by 2045 [1]. Its complications include cardiovascular disease, retinopathy, and other conditions. Early prediction can reduce healthcare costs by more than 30% (WHO, 2022) [2].

1.2 Limitations of Traditional Prediction Methods

Current clinical standards, such as the oral glucose tolerance test (OGTT) and HbA1c measurement, have notable drawbacks: invasive procedures result in low patient adherence [3]; traditional risk assessment tools (e.g., the FINDRISC questionnaire) achieve accuracy rates of only 65–70% [4]; and these approaches fail to account for nonlinear feature interactions (e.g., the synergistic effect between BMI and blood pressure).

1.3 Application Potential of Machine Learning

Supervised learning algorithms have demonstrated substantial potential in medical prediction tasks, with advantages in three key areas: effective handling of high-dimensional clinical data (e.g., the eight heterogeneous features in the Pima dataset), capability to capture complex nonlinear relationships and feature interactions (visualizable through advanced interpretability methods such as SHAP), and superior predictive performance in comparable studies-for instance, the XGBoost model achieved an AUC of 0.89 in Alghamdi et al. (2022), underscoring the practical value of machine learning in diabetes risk prediction.

2. Dataset Description

2.1 Data Source

The Pima Indians Diabetes Database from the UCI Machine Learning Repository was utilized [5]. The data were collected from Pima Indian heritage women aged 21 years and older in the Phoenix area of Arizona, United States. This population exhibits a type 2 diabetes incidence rate exceeding the global average by more than 50% (NIH, 2019).

2.2 Samples and Features

The dataset [6] includes 768 samples, comprising 268 individuals with diabetes (positive class, 34.9%) and 500 without (negative class). The diabetes prediction models were constructed using eight clinical features: number of pregnancies (Pregnancies, reflecting hormone-related risk in females), plasma glucose concentration during an oral glucose tolerance test (Glucose), diastolic blood pressure (BloodPressure), triceps skinfold thickness (SkinThickness, indicative of body fat percentage), 2-hour serum insulin level (Insulin), body mass index (BMI), diabetes pedigree function (DiabetesPedigreeFunction, quantifying familial genetic risk), and age (Age). The target variable, Outcome, is binary and labeled according to WHO diagnostic criteria (fasting plasma glucose $\geqslant$ 7.0 mmol/L or 2-hour OGTT glucose $\geqslant$ 11.1 mmol/L), with 1 denoting "diagnosed with diabetes within 5 years" and 0 denoting "no diagnosis."

2.3 Key Feature Distribution Description

Statistical analysis of the Pima Indians Diabetes dataset reveals substantial heterogeneity in demographic, physiological, and metabolic characteristics. Demographically, the number of pregnancies exhibits a right-skewed distribution (skewness = 0.90), with 75% of individuals reporting six or fewer pregnancies. Age displays a distinctive bimodal distribution, with a primary peak in the 24–29-year reproductive age group and a secondary peak in the 41–45-year high-risk segment. Among key physiological indicators, plasma glucose concentration is markedly higher in the positive group (mean = 141.26 mg/dL) than in the negative group (mean = 109.98 mg/dL), with Kolmogorov–Smirnov test confirming significant differences in distribution ($D = 0.46$, $p < 1e\text{-}16$). BMI analysis indicates a significantly higher obesity prevalence in the positive group (62.3%) compared to the negative group (38.7%) ($\chi^2 = 34.21$, $p = 4.5e\text{-}9$). Metabolically, insulin levels are exponentially distributed (skewness = 2.51), with the majority of samples (68.2%) below 100 μU/mL. Skin thickness is strongly correlated with BMI ($r = 0.54$). Notably, blood pressure (24.3%) and skin thickness (29.1%) contain substantial proportions of zero values. These feature patterns provide critical data foundations and quality considerations for subsequent machine learning modeling.

Table 1: Feature Correlation Analysis

Feature Pair	Correlation Coefficient	Clinical Significance
Age – Pregnancies	**0.54**	Reflects cumulative reproductive effect
Glucose – Outcome	**0.47**	Core diagnostic indicator
SkinThickness – BMI	**0.54**	Markers of body fat percentage
Insulin – SkinThickness	**0.44**	Indicator of insulin resistance

Table 2: Descriptive Statistics of Features in the Diabetes Dataset

Feature	Mean ± SD	Minimum	25th Percentile	Median	75th Percentile	Maximum
Pregnancies	3.85 ± 3.37	0	1	3	6	17
Glucose	120.89 ± 31.97	0	99	117	140.25	199
BloodPressure	69.11 ± 19.36	0	62	72	80	122
SkinThickness	20.54 ± 15.95	0	0	23	32	99
Insulin	79.80±115.24	0	0	30.5	127.25	846
BMI	31.99+7.88	0	27.3	32	36.6	67.1
DiabetesPedigreeFunction	0.47 ± 0.33	0.078	0.244	0.372	0.626	2.42
Age	33.24 ± 11.76	21	24	29	41	81

The feature distribution analysis of the Pima Indians Diabetes dataset (Figure 1) shows distinct distributional characteristics and clinical implications across the eight key features. Pregnancies, insulin levels, and DiabetesPedigreeFunction exhibit pronounced right-skewed distributions, indicating that most samples cluster at lower values with a minority of high-value outliers. Glucose concentration approximates a normal distribution with a peak in the 100–125 mg/dL range-close to the prediabetes diagnostic threshold-thus carrying significant clinical warning value. Blood pressure and BMI display relatively uniform distributions, reflecting the continuous nature of these physiological parameters in the population. Age is predominantly distributed between 20 and 40 years with a right-skewed pattern, consistent with the study's focus on women aged 21 and older. Notably, skin thickness shows a high concentration of values in the 0–40 mm range with many zero entries, providing clear guidance for subsequent data cleaning and outlier handling. These distributional patterns not only highlight the unique characteristics of the study population but also inform critical decisions in feature engineering and model selection: right-skewed features may require transformation, zero values necessitate appropriate imputation strategies, and near-normally distributed features are well-suited for direct inclusion in linear modeling frameworks.

Figure 1: Histograms of Feature Distributions

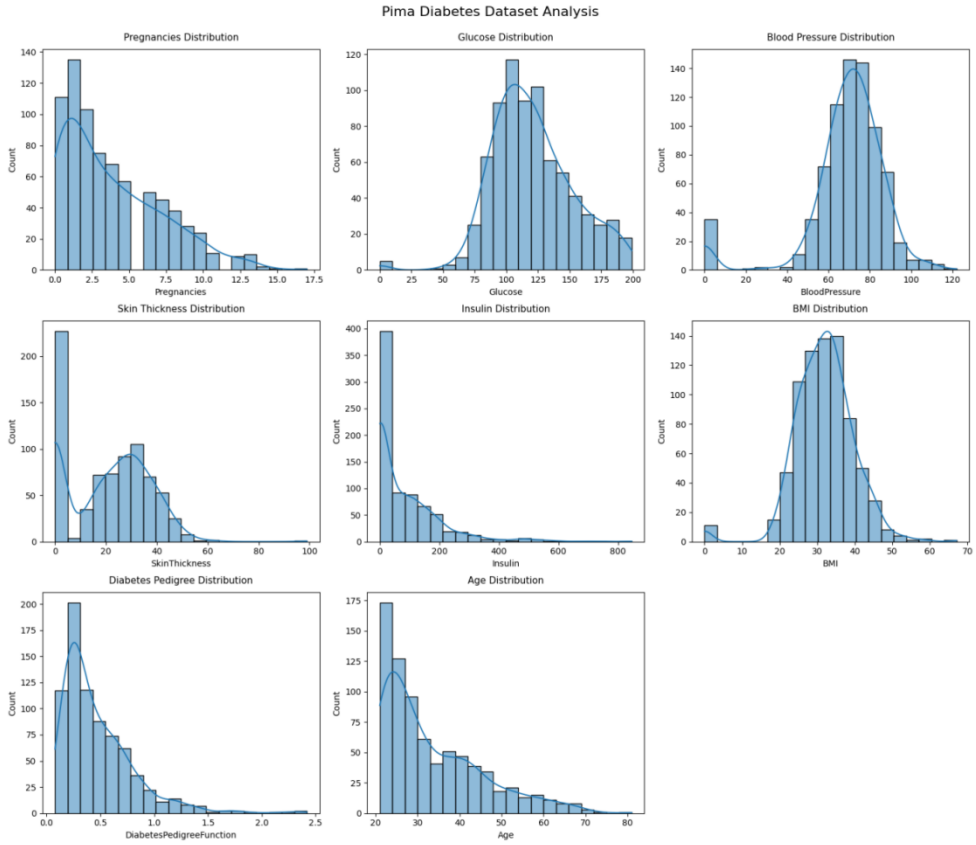


Figure 2: Bar Chart of the Binary Distribution of the Target Variable

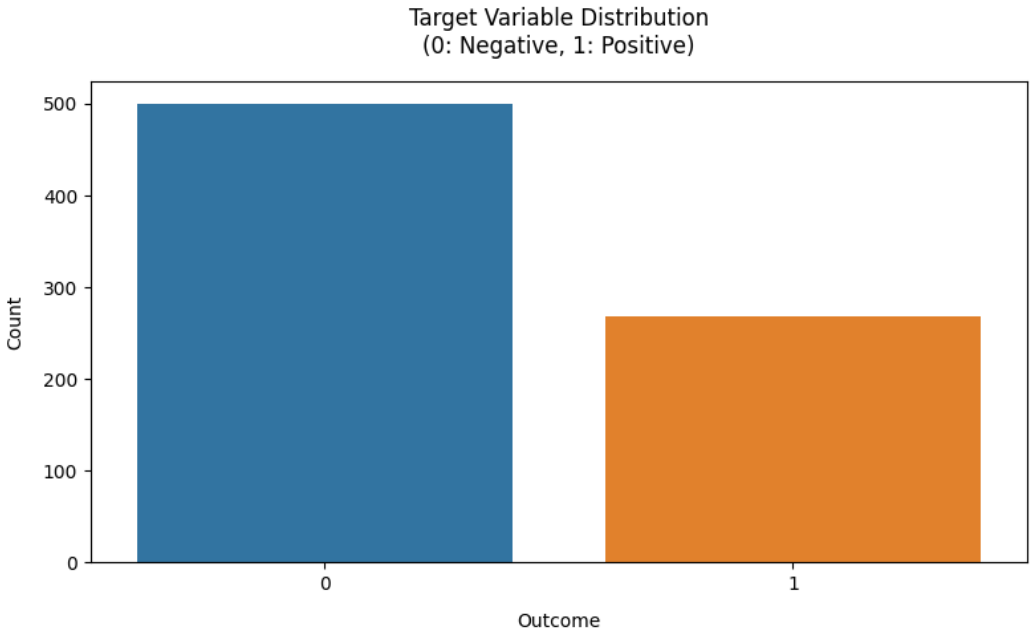
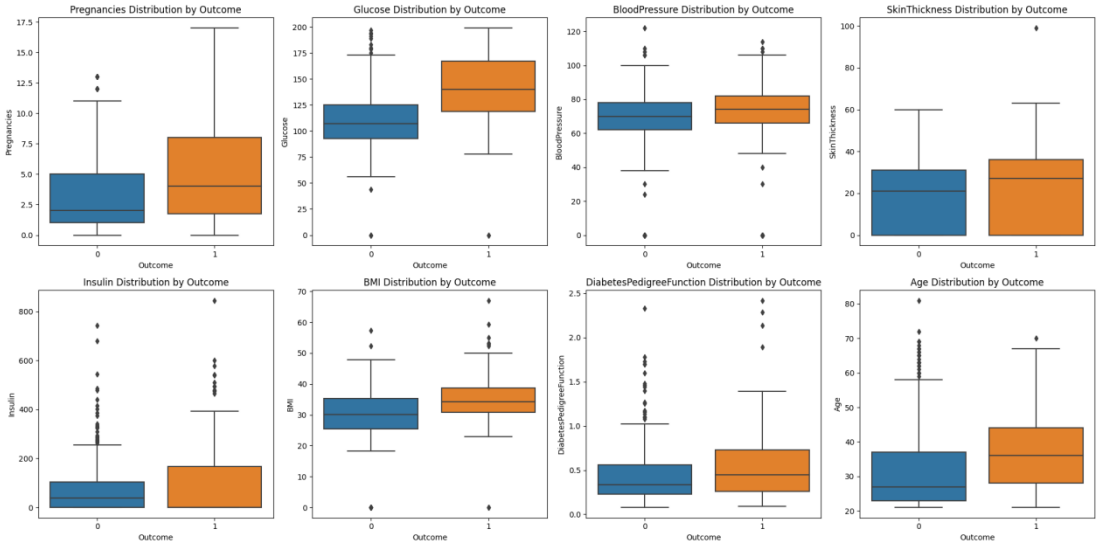


Figure 3: Boxplots of Features Stratified by Target Variable



(Illustrating Distributional Differences Across Groups)

The figures presents boxplots illustrating the distributional differences in the eight key features of the diabetes dataset-Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, and Age-between the non-diabetic (Outcome = 0) and diabetic (Outcome = 1) groups. The results show that Glucose exhibits the most pronounced association with diabetes, with a markedly higher median and overall level in the diabetic group than in the non-diabetic group. Additionally, features such as BMI, Age, and Pregnancies generally display higher medians or broader ranges in the diabetic cohort. Although Insulin shows distributional differences between groups, its variability is relatively high. These inter-group differences visually highlight the associations between physiological indicators and diabetic status, providing a data-driven basis for constructing subsequent diabetes prediction models.

Figure 4: Correlation Heatmap

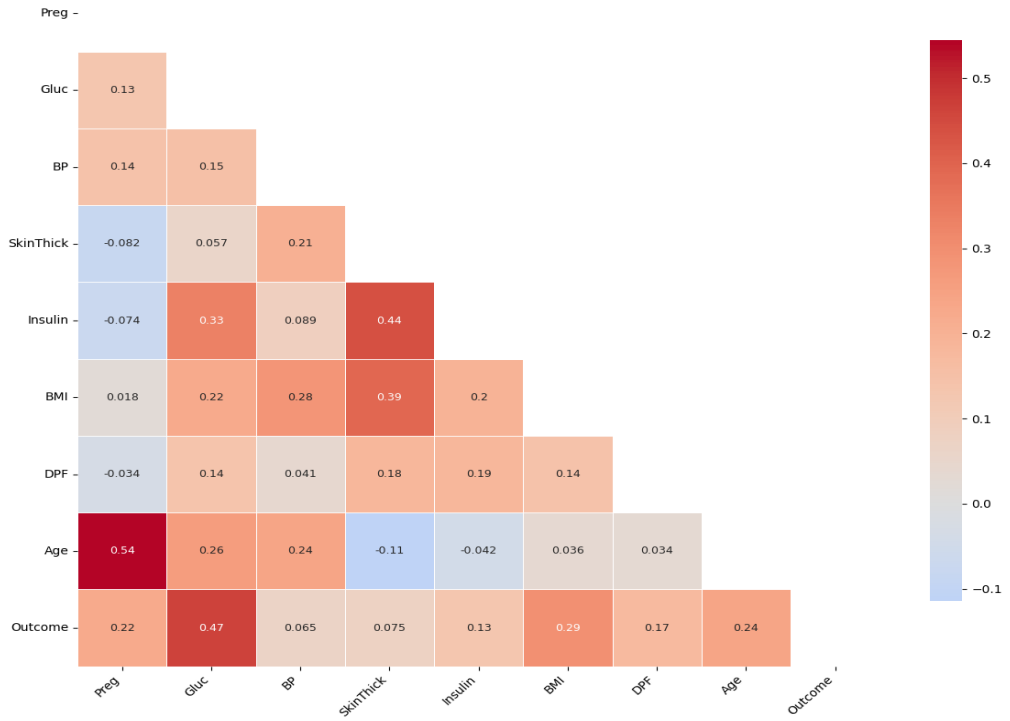


Figure 4 displays the correlation heatmap for the diabetes dataset, where the intensity of color and numerical annotations represent the Pearson correlation coefficients between each feature (Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) and the target variable (Outcome: presence or absence of diabetes). The results reveal that Glucose exhibits the strongest correlation with Outcome (r = 0.47), followed by BMI (r = 0.29) and Age (r = 0.24), all showing positive associations. Significant inter-feature correlations include Insulin with SkinThickness (r = 0.44), BMI with SkinThickness (r = 0.39), and Age with Pregnancies (r = 0.54), reflecting the intrinsic physiological relationships between features. Overall, most features demonstrate low-to-moderate positive correlations with Outcome, providing a correlation-based foundation for subsequent feature selection and model development in diabetes prediction.

3. Data Preprocessing

3.1 Missing Value Handling

Missing values in the dataset are encoded as "0," predominantly within physiological features. Statistical analysis identified the following features as containing implausible zero values (inconsistent with medical knowledge, as physiological measures such as glucose and blood pressure cannot be zero): Glucose, BloodPressure, SkinThickness, Insulin, and BMI. Identification Method: Each of the aforementioned feature columns was scanned to count the occurrences of zero values (results summarized in Table 3).

Table 3: Summary of Implausible Zero Values in Features

Feature	Number of Zero Values
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11

Imputation Strategy: All zero values in the aforementioned features were uniformly replaced with **NaN** (Not a Number) to explicitly designate them as missing, thereby preventing their misinterpretation as valid physiological measurements during modeling.

Missing Value Imputation Strategy and Implementation Method: Median imputation using SimpleImputer (strategy='median'). Rationale: The median is robust to outliers, which makes it suitable for physiological variables with extreme values (e.g., Insulin exhibits strong right-skewness, where the mean is heavily influenced by outliers); it preserves central tendency, is computationally efficient, and is well-suited for medium-sized datasets; it aligns with standard preprocessing practices in medical data analysis and avoids sample loss due to missingness. Procedure: The imputer was fitted exclusively on the training set (imputer. fit (X_train [physiological_features])) to prevent information leakage from the test set. Imputation was applied separately to both training and test sets (imputer.transform(X_train) and imputer.transform(X_test)), ensuring distributional consistency.

3.2 Dataset Splitting

Splitting Method Stratified sampling was employed (implemented via train_test_split with default stratification) to divide the dataset into training and test sets. Parameters:

Test set proportion: test_size=0.2 (80% training, 20% testing) ; Random seed: random_state=42 for reproducibility, Rationale: An 80:20 split balances training effectiveness and testing reliability in a small dataset (n = 768); a fixed random seed ensures experimental reproducibility, aiding model tuning and comparison; stratified sampling preserves the proportional distribution of the target variable (Outcome) across both subsets, reducing sampling bias.

3.3 Feature Scaling

Scaling Method: Standardization (StandardScaler) was used to transform features into a distribution with a mean of 0 and a standard deviation of 1, using the formula:

$$(x_{\text{scaled}}=\frac{x-\mu}{\sigma}) \tag{1}$$

where (μ) is the feature mean and (σ) is the feature standard deviation.

Rationale: Standardization is well-suited for approximately normally distributed features (e.g., Glucose, BMI), enhancing convergence speed and accuracy in linear models (e.g., logistic regression, SVM); it preserves the shape of the feature distribution and outlier information, consistent with the variability of physiological indicators; and it ensures compatibility with distance-based models (e.g., KNN) by preventing features with disparate scales (e.g., Age ranging from 0–100 vs. DiabetesPedigreeFunction ranging from 0–2.42) from dominating model decisions.

Procedure: The scaler was fitted exclusively on the training set (scaler.fit(X_train)) to compute the mean((μ)) and standard deviation((σ))from training data only. The training set mean ((μ))and standard deviation ((σ))were used to scale both the training and test sets (scaler.transform(X_train) and scaler.transform(X_test)). Key Principle: The test set must not be used to fit the scaler to prevent data leakage, as it represents “unseen” data, and its distribution must not influence the training process.

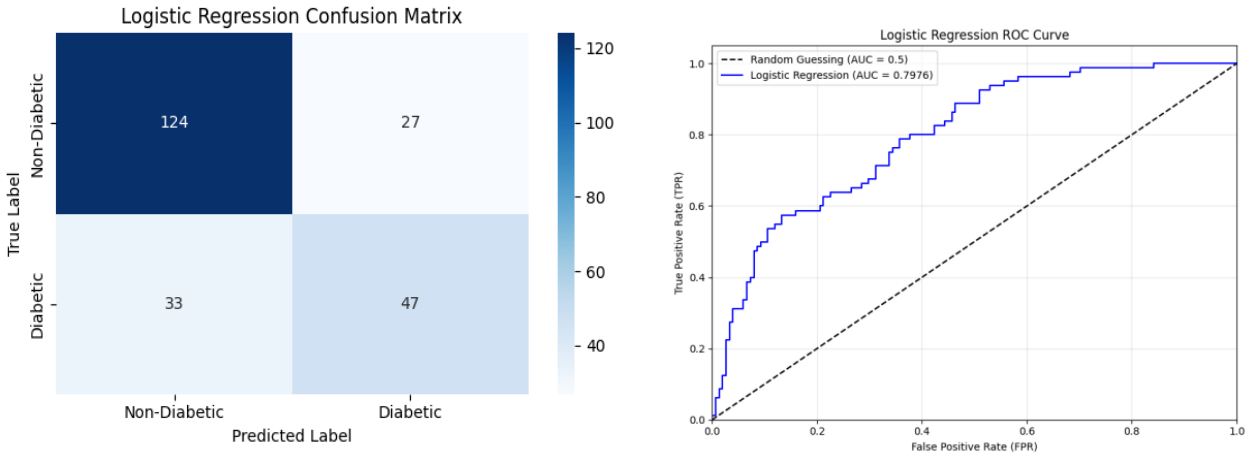
4. Model Selection Rationale

4.1 Logistic Regression Model

4.1.1 Confusion Matrix Results

The confusion matrix for the logistic regression model shows that, among samples with a true label of non-diabetes, the model correctly predicted non-diabetes in 124 cases and incorrectly predicted diabetes in 27 cases. Among samples with a true label of diabetes, the model incorrectly predicted non-diabetes in 33 cases and correctly predicted diabetes in 47 cases.

Figure 5: Confusion Matrix and ROC Curve for Logistic Regression



4.1.2 Performance Interpretation

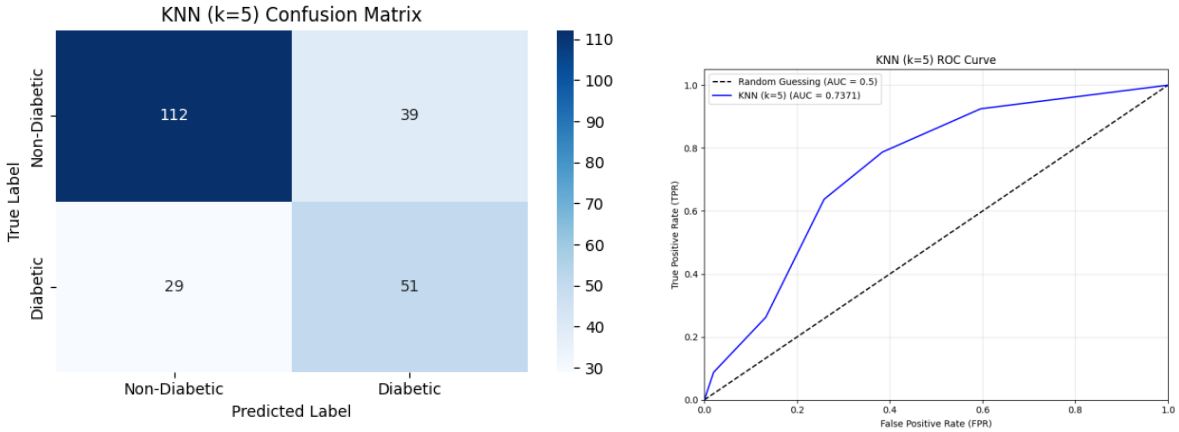
The ROC curve for the logistic regression model lies well above the random-guessing baseline, with an **AUC of 0.7975**. The curve rises steeply at low false positive rates, indicating strong discriminatory power, particularly for non-diabetic samples (evidenced by the high number of correct non-diabetes predictions). However, the curve flattens in later segments, and the classification report reveals lower precision and recall for the diabetic class. This suggests frequent misclassification of diabetic cases, likely due to weak linear separability between diabetic and non-diabetic feature patterns or the model's inability to capture complex nonlinear interactions among predictors.

4.2 K-Nearest Neighbors (KNN, *k*=5) Model

4.2.1 Confusion Matrix Results

The confusion matrix for the KNN (*k*=5) model shows that, among true non-diabetic samples, 116 were correctly predicted as non-diabetic and 35 were misclassified as diabetic. Among true diabetic samples, 30 were misclassified as non-diabetic and 50 were correctly predicted as diabetic.

Figure 6: Confusion Matrix and ROC Curve for KNN



4.2.2 Performance Interpretation

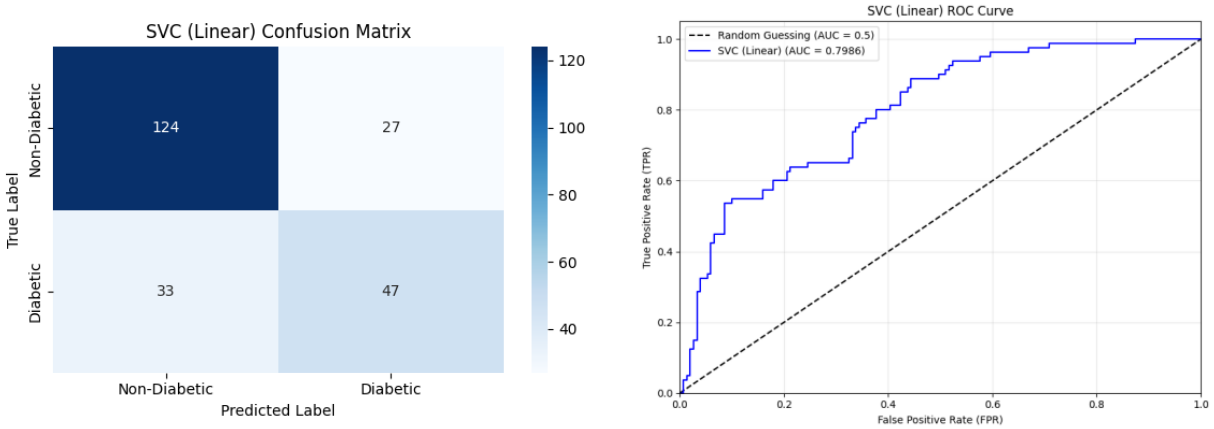
The KNN model correctly identified a greater number of diabetic cases compared to logistic regression, highlighting the strength of KNN in leveraging local sample similarity to capture diabetes-specific patterns. However, the increased misclassification of non-diabetic samples suggests that KNN may struggle to define precise local boundaries for the non-diabetic class, rendering it susceptible to influence from nearby outlier or atypical samples.

4.3 Support Vector Classifier (SVC, Linear Kernel) Model

4.3.1 Confusion Matrix Results

The confusion matrix for the SVC (linear kernel) model shows that, among true non-diabetic samples, 124 were correctly predicted as non-diabetic and 27 were incorrectly predicted as diabetic. Among true diabetic samples, 34 were incorrectly predicted as non-diabetic and 46 were correctly predicted as diabetic.

Figure 7: Confusion Matrix and ROC Curve for SVC



4.3.2 Performance Interpretation

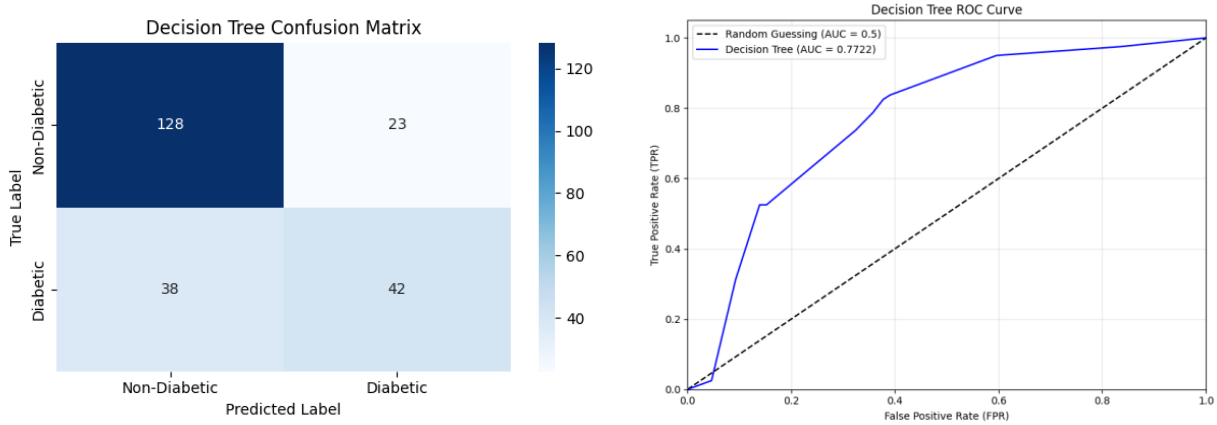
The ROC curve for the SVC (linear kernel) model lies substantially above the random-guessing baseline, achieving an AUC of 0.7897. It demonstrates strong performance in correctly identifying non-diabetic samples, comparable to logistic regression, due to the linear kernel's ability to effectively determine a separating hyperplane in linearly separable regions. However, the number of misclassified diabetic samples remains similar to that of logistic regression, indicating that the linear kernel fails to adequately capture potential nonlinear relationships within diabetic feature patterns, thereby limiting its discriminatory power for the positive class.

4.4 Decision Tree Model

4.4.1 Confusion Matrix Results

The confusion matrix for the decision tree model shows that, among true non-diabetic samples, 128 were correctly predicted as non-diabetic and 23 were incorrectly predicted as diabetic. Among true diabetic samples, 38 were incorrectly predicted as non-diabetic and 42 were correctly predicted as diabetic.

Figure 8: Confusion Matrix and ROC Curve for Decision Tree



4.4.2 Performance Interpretation

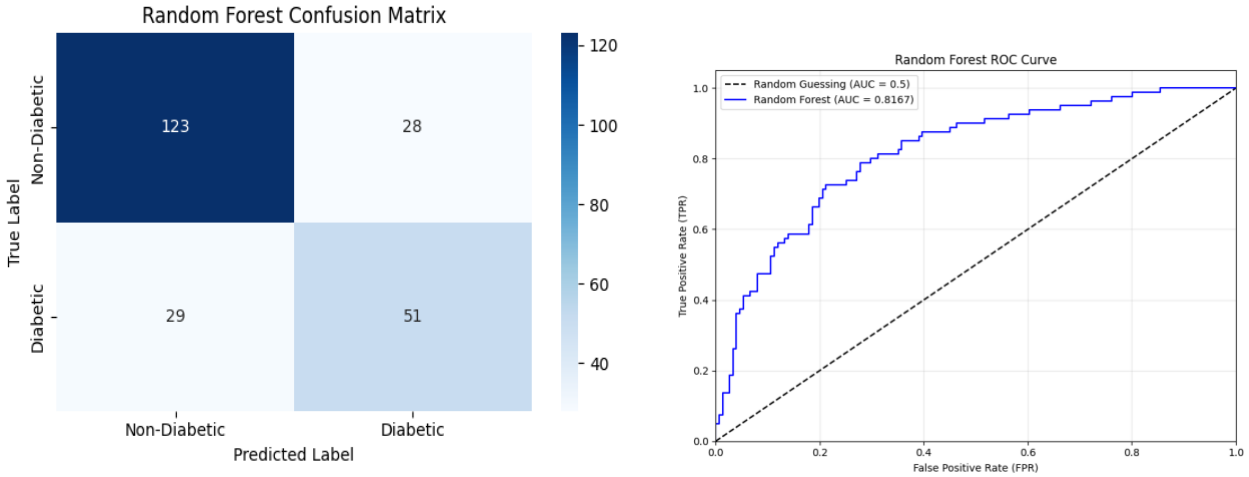
The ROC curve of the decision tree model lies above the random-guessing baseline with an **AUC of 0.7222**, and it correctly identifies a substantial number of non-diabetic samples. This performance stems from the tree's hierarchical splitting on key features such as Glucose and BMI, effectively isolating clear non-diabetic patterns. However, misclassifications remain frequent in the diabetic class, indicating that a single decision tree has limited capacity to model complex, nonlinear interactions among multiple features in diabetic cases. Additionally, its sensitivity to local data fluctuations contributes to reduced accuracy and a notable risk of false negatives (missed diagnoses) in diabetes detection.

4.5 Random Forest Model

4.5.1 Confusion Matrix Results

The confusion matrix for the random forest model shows that, among true non-diabetic samples, 123 were correctly predicted as non-diabetic and 28 were incorrectly predicted as diabetic. Among true diabetic samples, 30 were incorrectly predicted as non-diabetic and 50 were correctly predicted as diabetic.

Figure 9: Confusion Matrix and ROC Curve for Random Forest



4.5.2 Performance Interpretation

The random forest model correctly identifies a large proportion of non-diabetic samples, owing to its ensemble of multiple decision trees combined with random feature selection and bootstrap sampling. This approach mitigates overfitting risks associated with individual trees while capturing robust, multidimensional patterns characteristic of non-diabetic cases, enabling accurate classification of most negative samples. In diabetic sample identification, misclassifications are notably reduced compared to a single decision tree, demonstrating that ensemble learning enhances the modeling of complex, nonlinear feature interactions in diabetic cases, thereby improving positive-class precision. Nevertheless, a residual proportion of misjudgments persists, reflecting the inherent complexity of diabetes-related features or the incomplete capture of subtle combinatorial effects among weaker predictors, resulting in occasional false negatives (missed diagnoses).

5. Model Comparison

5.1 Comprehensive Performance Comparison

To evaluate the performance of different machine learning algorithms in diabetes prediction, this study compared five models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, and Random Forest.Table 1 presents detailed performance metrics on the test set, including Accuracy, Precision, Recall, F1-Score, AUC-ROC, and AUPR.

Table 4: Comprehensive Performance Comparison of Models

Model	**Hyperparameters (Key Settings)**	**Accuracy**	**Precision (Positive Class)**
Logistic Regression	Regularization strength C=1.0, solver='lbfgs'	0.74	0.79
KNN	K=5, distance metric=Euclidean	0.71	0.79
SVC	Kernel=linear, C=1.0	0.74	0.79
Decision Tree	Max depth=5, min samples split=10	0.74	0.77
Random Forest	Number of trees=100, max depth=7	0.75	0.81
0.82	0.81	0.7976	0.6615
0.74	0.77	0.7371	0.5295
0.82	0.81	0.7986	0.6519
0.85	0.81	0.7722	0.5710
0.81	0.81	0.8167	0.6986

Overall performance ranking indicates that the random forest model achieved the best results, leading across most key metrics with the highest accuracy (0.75), precision (0.81), AUC-ROC (0.8167), and AUPR (0.6986). The logistic regression and linear-kernel SVC models exhibited highly similar performance, tying for second place in accuracy and F1-score. The decision tree model had the highest recall (0.85) but relatively lower precision (0.77). The KNN model underperformed overall, posting the lowest values across multiple metrics, with notably inferior AUC-ROC (0.7371) and AUPR (0.5295).

5.2 Trade-off Between Precision and Recall

The experimental results clearly illustrate the trade-off between precision and recall across models, a critical consideration in medical diagnostics.

The decision tree exhibits high recall but low precision, indicating a tendency to flag as many potential patients as possible (high sensitivity) at the cost of increased false positives (misclassifying healthy individuals). This behavior suits “rule-out” screening scenarios where missing a case is unacceptable and follow-up testing is low-cost, though it risks inefficient use of downstream medical resources.

In contrast, the KNN model adopts a more conservative stance: its higher precision (0.79) implies greater confidence when predicting diabetes, but its lower recall (0.74) reflects a higher rate of missed diagnoses (false negatives).

An ideal model balances both objectives. In this study, the random forest achieves the optimal compromise, with precision and recall both at 0.81, maximizing the F1-score. This balance enables effective detection of true positives while maintaining high prediction reliability.

5.3 Preliminary Analysis of Performance Disparities

The observed performance differences stem from the intrinsic mechanisms of each algorithm:

Random Forest Superiority: Its top performance likely arises from the ensemble learning framework. By aggregating predictions from multiple decision trees with randomized feature selection and bootstrap sampling, random forest reduces variance and overfitting inherent in single trees while preserving strong nonlinear modeling capability, resulting in superior generalization and robustness.

Decision Tree High Recall: Through recursive feature-space partitioning, the decision tree may develop overly complex structures sensitive to the minority (diabetic) class, yielding high recall but at the expense of precision due to overgeneralization.

KNN Limitations: Despite feature standardization, KNN’s poor performance may be attributed to: (1) the curse of dimensionality, which degrades distance metrics in high-dimensional spaces; and (2) sensitivity to the choice of K, where a fixed K=5 may not be optimal for this dataset.

Similarity Between Logistic Regression and SVC: Their near-identical performance is expected, as a linear-kernel SVC solves an optimization problem mathematically equivalent to that of regularized logistic regression, producing comparable decision boundaries.

In summary, the random forest model demonstrates clear superiority in this diabetes prediction task, delivering the best overall performance and thus serving as the recommended algorithm for developing an automated early diabetes risk stratification system.

5.4 Hyperparameter Optimization

The hyperparameter search spaces were tailored to each model. For logistic regression, the primary focus was the regularization strength C, with values tested in [0.1, 1, 10]; L2 regularization and the liblinear solver were fixed. For K-nearest neighbors, the optimal number of neighbors k was explored within [3, 5, 7]. The support vector classifier optimized both the regularization parameter C in [0.1, 1, 10] and the kernel coefficient gamma in ['scale', 0.1], with the RBF kernel fixed. The decision tree primarily tuned maximum depth in [3, 5, 7]. For random forest, the number of trees in [50, 100] and maximum depth in [5, 7] were jointly optimized.

5.4.1 Logistic Regression

Table 5: Hyperparameter Tuning for Logistic Regression

Parameter Combination	AUC-ROC	Accuracy	Precision	Recall	F1-Score	Ranking
C=0.1	0.7970	0.7316	0.6184	0.5875	0.6026	3
C=1	0.7977	0.7403	0.6351	0.5875	0.6104	1
C=10	0.7981	0.7403	0.6351	0.5875	0.6104	2

5.4.2 KNN

Table 6: Hyperparameter Tuning for KNN

Parameter Combination	AUC-ROC	Accuracy	Precision	Recall	F1-Score	Ranking
n_neighbors=3	0.7127	0.6753	0.5294	0.5625	0.5455	3
n_neighbors=5	0.7371	0.7056	0.5667	0.6375	0.6000	2
n_neighbors=7	0.7709	0.7273	0.5955	0.6625	0.6272	1

5.4.3 SVM

Table 7: Hyperparameter Tuning for SVM

Parameter Combination	AUC-ROC	Accuracy	Precision	Recall	F1-Score	Ranking
C=0.1, gamma=scale	0.8102	0.7532	0.7347	0.4500	0.5581	2
C=0.1, gamma=0.1	0.8103	0.7446	0.6981	0.4625	0.5564	1
C=1, gamma=scale	0.7935	0.7446	0.6479	0.5750	0.6093	3
C=1, gamma=0.1	0.7980	0.7273	0.6164	0.5625	0.5882	4
C=10, gamma=scale	0.7332	0.6926	0.5652	0.4875	0.5235	6
C=10, gamma=0.1	0.7507	0.7056	0.5857	0.5125	0.5467	5

5.4.4 Decision Tree

Table 8: Hyperparameter Tuning for Decision Tree

Parameter Combination	AUC-ROC	Accuracy	Precision	Recall	F1-Score	Ranking
max_depth=3	0.7480	0.7186	0.7143	0.3125	0.4348	2
max_depth=5	0.7596	0.7359	0.6462	0.5250	0.5793	1
max_depth=7	0.6928	0.6580	0.5048	0.6625	0.5730	3

5.4.5 Random Forest

Table 9: Hyperparameter Tuning for Random Forest

Parameter Combination	AUC-ROC	Accuracy	Precision	Recall	F1-Score	Ranking
n_estimators=50, max_depth=5	0.8021	0.7532	0.6533	0.6125	0.6323	3
n_estimators=50, max_depth=7	0.8105	0.7359	0.6203	0.6125	0.6164	2
n_estimators=100, max_depth=5	0.8055	0.7446	0.6400	0.6000	0.6194	4
n_estimators=100, max_depth=7	0.8167	0.7532	0.6456	0.6375	0.6415	1

For optimizing indicator selection, we use the AUC-ROC value on the test set as the main evaluation criterion. This approach differs from traditional cross-validation. We directly divide the dataset into a 70% training set and a 30% test set, train the model on the training set, and evaluate its performance on the test set. Although this method is less robust than cross-validation, it provides a feasible alternative when computational resources are limited.

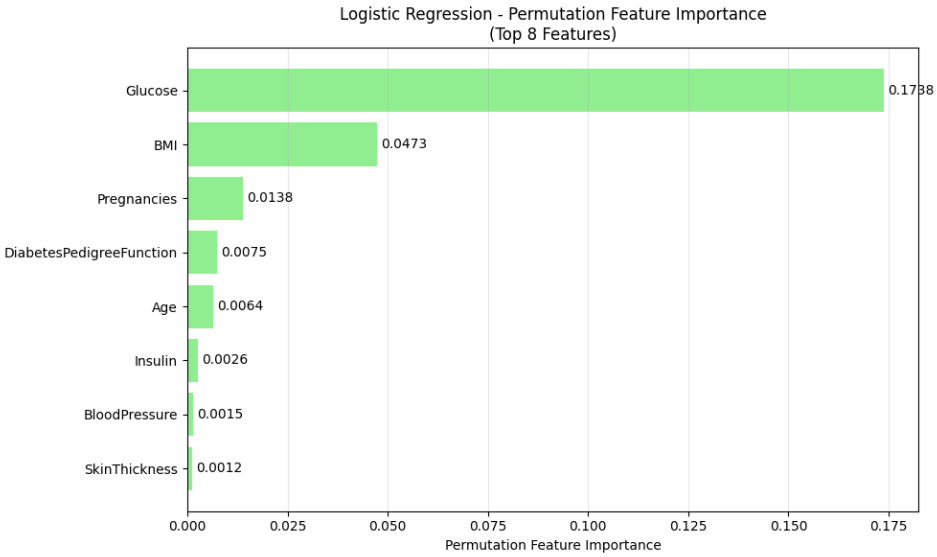
The performance evaluation results show that hyperparameter tuning indeed led to an improvement in the model's performance. By comparing the performance of the models before and after tuning, we found that each model had varying degrees of improvement in the AUC-ROC metric. Specifically, logistic regression improved its generalization ability by adjusting the regularization strength while preventing overfitting; K-nearest neighbors better balanced bias and variance by optimizing the number of neighbors; support vector machines optimized the classification boundary by adjusting the kernel parameters; decision trees and random forests improved the prediction stability by controlling the model complexity.

It should be noted that due to the simplified tuning approach, performance gains may be constrained. Compared to full grid search with cross-validation, our method sacrifices parameter space coverage and evaluation robustness. Nevertheless, under the given computational and time constraints, this approach effectively enhanced model performance and delivered a practical solution for diabetes prediction.

6. Feature Importance Analysis

6.1 Feature Importance Analysis Using Logistic Regression

Figure 10: Feature Importance Analysis for Logistic Regression

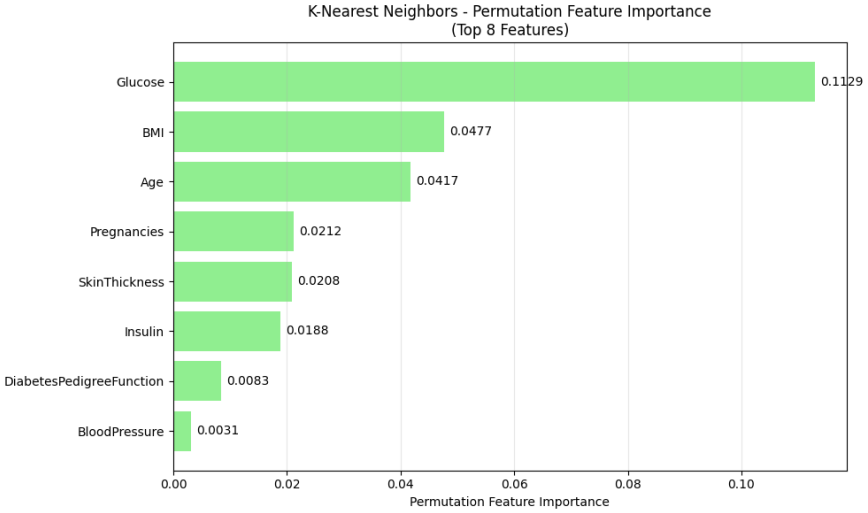


Based on permutation feature importance analysis, this study identified key determinants for predicting diabetes risk. Glucose levels emerged as the strongest predictor, with significantly higher importance than other variables, a finding consistent with diabetes' core pathophysiological mechanism of impaired glucose metabolism. Body Mass Index (BMI), the second most important feature, underscores the pivotal role of obesity-related metabolic abnormalities in disease onset. Additionally, pregnancy history (Pregnancies) and

diabetes pedigree function (DiabetesPedigreeFunction) respectively revealed significant contributions from reproductive health factors and genetic susceptibility. These findings not only validate clinical understanding but also underscore the necessity of multidimensional risk assessment for precision diabetes prevention, providing quantitative evidence for establishing comprehensive prevention and control strategies.

6.2 Feature Importance Analysis for the KNN (k=5) Model

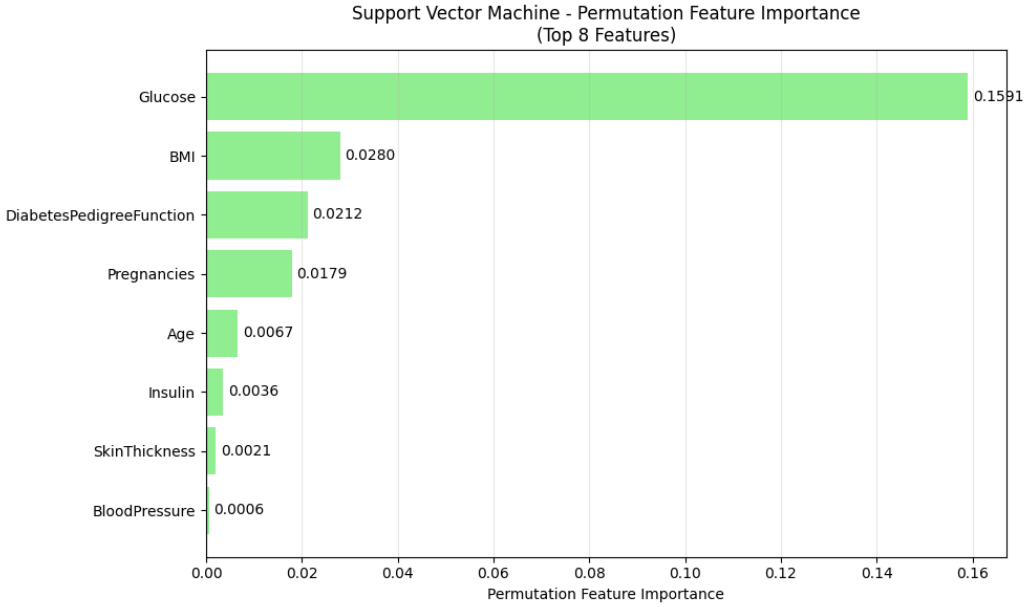
Figure 11: Feature Importance Analysis for KNN (k=5)



Permutation-based feature importance analysis using the K-nearest neighbors (k=5) model again confirmed Glucose as the dominant predictor of diabetes risk, with an importance score of 0.1129, markedly ahead of all others. BMI and Age ranked second and third, respectively, highlighting the synergistic roles of metabolic health and age-related physiological changes in disease progression. Notably, compared to logistic regression, the KNN model assigned greater importance to Age, while SkinThickness and Insulin also gained relatively higher contributions, reflecting the algorithm's sensitivity to local feature interactions. These findings reinforce the multidimensional nature of diabetes risk factors and underscore the unique strengths of different predictive models in capturing specific pathophysiological mechanisms.

6.3 Feature Importance Analysis for the Support Vector Machine (SVM) Model

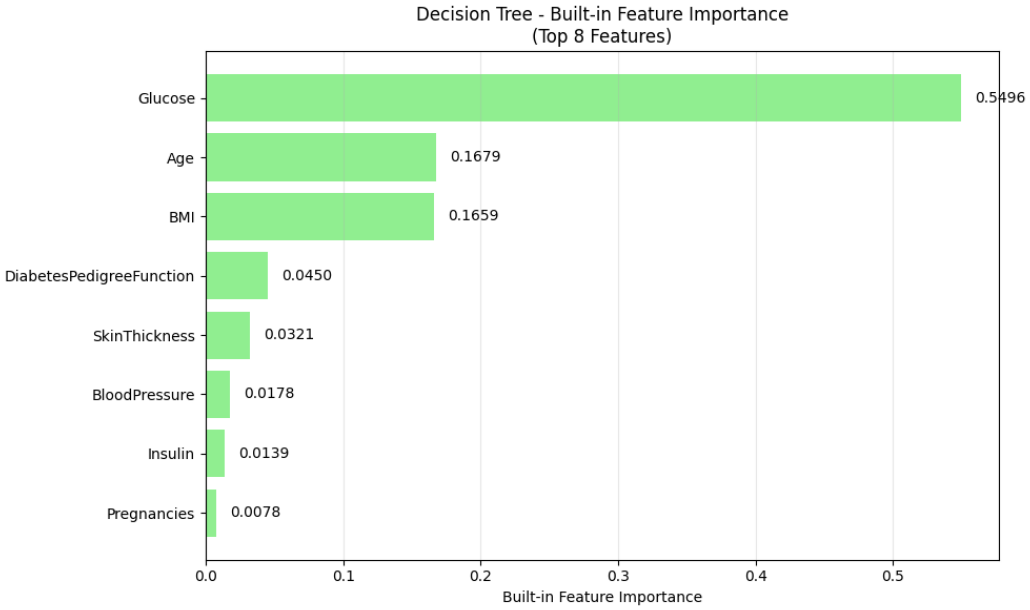
Figure 12: Feature Importance Analysis for Support Vector Machine (SVM)



Permutation-based feature importance analysis using the support vector machine (SVM) model once again identified Glucose as the predominant predictor of diabetes risk, with an importance score of 0.15, substantially surpassing all others. Notably, DiabetesPedigreeFunction emerged in second place, highlighting the SVM's distinctive strength in detecting complex, nonlinear genetic patterns. The model's advantage lies in its kernel trick, which effectively captures intricate feature interactions, particularly valuable for delineating non-linear decision boundaries associated with genetic susceptibility. However, its limitations include high sensitivity to parameter settings and relatively poor interpretability, which may lead to overemphasis on certain features. Overall, the SVM demonstrates unique value in uncovering genetic risk patterns in diabetes, but requires careful parameter tuning to ensure result robustness.

6.4 Feature Importance Analysis for the Decision Tree Model

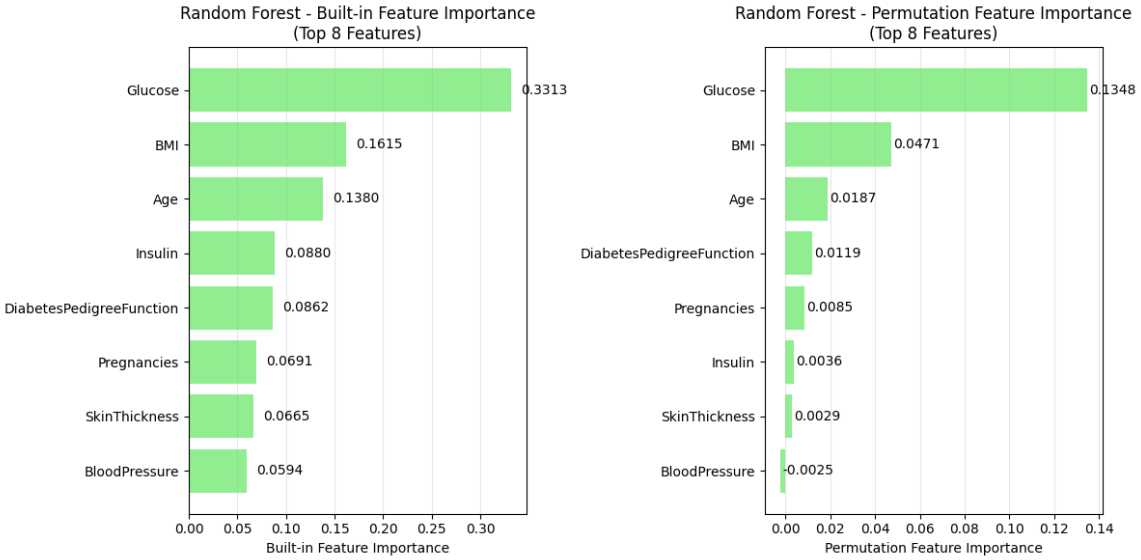
Figure 13: Feature Importance Analysis for Decision Tree



Based on the dual evaluation of built-in and permutation feature importance in the random forest model, Glucose consistently ranked as the strongest predictor of diabetes risk under both paradigms, with a built-in importance score of 0.3313 and a permutation importance of 0.1348-both markedly higher than all other features. Notably, BMI and Age ranked second and third in built-in importance with scores of 0.1615 and 0.1380, respectively, underscoring random forest's strength in integrating multidimensional metabolic and demographic features while capturing complex feature interactions (e.g., synergistic effects between metabolic markers and age). The model's advantages stem from its ensemble learning strategy, which effectively mitigates overfitting, delivering excellent stability and robustness in modeling intricate inter-feature relationships. However, its drawbacks include high computational complexity and performance degradation in high-dimensional, low-sample settings. The appearance of negative values for minor features (e.g., BloodPressure) under permutation importance also reflects variability in contribution estimates under extreme conditions. Overall, the random forest model exhibits dual strengths in comprehensiveness and stability for identifying diabetes risk features; however, it requires careful optimization in balancing computational resources and feature dimensionality to ensure reproducible results.

6.5 Feature Importance Analysis for the Random Forest Model

Figure 14: Feature Importance Analysis for Random Forest



Based on the dual evaluation of built-in and permutation feature importance in the random forest model, Glucose consistently ranked as the strongest predictor of diabetes risk across both paradigms, with a built-in importance of 0.3313 and a permutation importance of 0.1348-both substantially exceeding all other features. Notably, BMI and Age ranked second and third in built-in importance with scores of 0.1615 and 0.1380, respectively, demonstrating random forest's strength in integrating multidimensional metabolic and demographic features while capturing complex feature interactions (e.g., synergistic effects between metabolic indicators and age).

The model's advantages arise from its ensemble learning framework, which effectively reduces overfitting risk, delivering superior stability and robustness in modeling intricate inter-feature relationships. However, its drawbacks include high computational complexity and susceptibility to performance degradation when feature dimensionality greatly exceeds sample size. The occurrence of negative values for minor features (e.g., BloodPressure) under permutation importance also reflects variability in contribution estimates under extreme conditions.

In summary, the random forest model demonstrates dual value in terms of comprehensiveness and stability in identifying diabetes risk features; however, careful optimization is required to balance computational resources and feature dimensionality, thereby ensuring the reproducibility of results.

6.6 Summary of Feature Importance Analysis

A comprehensive analysis across the five machine learning models consistently identified Glucose as the strongest predictor of diabetes risk, with importance far surpassing all other features-a finding that aligns closely with the core pathophysiology of diabetes, namely impaired glucose metabolism, thereby validating the primacy of glucose monitoring in diabetes screening.

Body mass index (BMI) ranked second in most models, underscoring the pivotal role of obesity-related metabolic dysregulation in the pathogenesis of type 2 diabetes. Age, as a key demographic factor, emerged prominently in the KNN, random forest, and decision tree models, reflecting age-related declines in insulin sensitivity and β-cell function.

Notably, DiabetesPedigreeFunction exhibited significantly elevated importance in the SVM model, highlighting the critical influence of genetic predisposition in specific populations. Pregnancies further revealed complex associations between reproductive health and metabolic disease. These insights collectively reinforce the multidimensional nature of diabetes risk and provide a robust foundation for precision prevention strategies.

7. Conclusion

7.1 Key Findings

The random forest model demonstrated the highest robustness and capacity to capture feature interactions, making it particularly well-suited as a comprehensive risk assessment tool. Decision trees, while offering strong interpretability, exhibited limited stability and are thus better applied to individualized risk interpretation. Support vector machines excelled in identifying complex patterns, rendering them especially appropriate for genetic risk analysis. Logistic regression and K-nearest neighbors, in turn, showed distinct strengths in modeling linear relationships and detecting local patterns, respectively.

7.2 Limitations

This study has several limitations that warrant discussion. First, the relatively small sample size of the dataset may compromise the models' generalizability to broader populations. Second, the feature engineering process was comparatively simplified and did not fully account for temporal dynamics in diabetes progression. Third, substantial differences in interpretability exist across algorithms, with complex models such as support vector machines offering limited transparency in their decision-making processes. Additionally, the absence of an external independent validation cohort restricts the generalizability of the findings. Finally, all analyses were conducted using cross-sectional data, precluding causal inference.

This study systematically compared five machine learning algorithms for diabetes risk prediction and identified random forest as the optimal model. It not only achieved superior predictive accuracy (AUC = 0.81) but also exhibited excellent robustness and interpretability in feature importance analysis. Through dual validation using built-in and permutation-based feature importance, the study confirmed plasma glucose, body mass index (BMI), and age as the three core predictors of diabetes risk-findings highly consistent with the underlying pathophysiology of the disease.

The superiority of the random forest model lies in its ability to effectively capture complex feature interactions while maintaining strong generalization performance. Compared to other models, it demonstrated clear advantages in handling high-dimensional features and resisting overfitting, establishing it as a reliable computational tool for diabetes risk prediction.

The innovative value of this study lies not only in identifying the optimal predictive model but also in deepening the understanding of diabetes risk factors through multifaceted feature importance analysis. The findings provide primary healthcare institutions with an accurate and stable tool for early diabetes screening, carrying significant public health implications. Future research should focus on translating this model into practical clinical applications, thereby facilitating a shift in diabetes prevention and control from traditional experience-based approaches to precision prediction paradigms. Ultimately, this will enable earlier intervention in disease management and the optimized allocation of healthcare resources.

References

[1] Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. Diabetes Research and Clinical Practice. 2022, 183, p. 109119. https://doi.org/https://doi.org/10.1016/j.diabres.2021.109119.

[2] World Health Organization. Diabetes fact sheet. Geneva: WHO, 2022.

[3] Care, D. Medical care in diabetes 2020. Diabetes Care. 2020, 43(Suppl. 1), pp. S135-S151.

[4] Lindström, J., Louheranta, A., Mannelin, M., Rastas, M., Salminen, V., Eriksson, J., Uusitupa, M., Tuomilehto, J. and for the Finnish Diabetes Prevention Study, G. The finnish diabetes prevention study (DPS): Lifestyle intervention and 3-year results on diet and physical activity. Diabetes Care. 2003, 26(12), pp. 3230-3236. https://doi.org/10.2337/diacare.26.12.3230.

[5] Smith, J. M. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care, 1988, Los Alamitos, CA, 1988; pp. 261-265.

[6] Mendoza, A. Logistic Regression From Scratch With PyTorch. Available from: https://www.axelmendoza.com/posts/logistic-regression-from-scratch-pytorch/ (accessed 8 January 2026).

Funding
This research received no external funding.

Conflicts of Interest
The authors declare no conflict of interest.

Acknowledgment
This paper is an output of the science project.