Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Diabetes Risk Prediction Model Using Machine Learning

Boyi Yang *

Department of mathematics, The University of British Columbia, Vancouver, Canada

* Corresponding Author Email: byang25@student.ubc.ca

Abstract. Diabetes is a major global health challenge, contributing to increased mortality and long-term complications worldwide. Early diagnosis and effective risk stratification are critical to reducing the disease burden. This research aims to evaluate and compare the predictive performance of five machine learning (ML) models—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM)—using the Pima Indians Diabetes Dataset. A standardized experimental workflow involving data preprocessing, missing value imputation, feature scaling, model training was applied. Performance metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC) were used to evaluate model outcomes. Among the models tested, Gradient Boosting achieved the highest accuracy (75.97%), whereas Random Forest attained the highest AUC (0.833), indicating its superior classification capability. These results demonstrate that Random Forest model, offers a promising and practical approach for implementing robust diabetes risk prediction tools in clinical or public health contexts.

Keywords: Diabetes prediction; Machine learning; Receiver Operating Characteristic (ROC); The area under the ROC curve (AUC); Accuracy.

1. Introduction

Diabetes is one of the most widespread chronic diseases. According to the World Health Organization (WHO), over 830 million adults were living with diabetes in 2022. In 2021, diabetes and related kidney disease accounted for over 2 million deaths worldwide. Furthermore, elevated blood glucose levels were responsible for approximately 11% of cardiovascular-related deaths. Additionally, more than half of the individuals diagnosed with diabetes did not receive any form of treatment in 2022. In 2021, global health expenditures related to diabetes reached USD 966 billion, and according to estimates by the International Diabetes Federation (IDF), this figure is projected to rise to USD 1.03 trillion by 2030 [1]. These alarming statistics underscore the urgent need for early detection and effective risk prediction strategies as part of global public health efforts [2].

The recent rise of data-driven health technologies has made machine learning (ML) a popular tool for disease risk prediction. ML algorithms are capable of identifying complex, non-linear relationships within large datasets, frequently outperforming traditional statistical models in predictive accuracy. In the context of diabetes, ML techniques have been widely explored and applied for diagnostic support and outcome forecasting. Multiple studies have explored ML-based diabetes prediction. Sisodia and D.S. Sisodia applied Naïve Bayes on the Pima dataset, achieving the accuracy of around 76% [3]. Umm E Laila et al. applied models such as AdaBoost, Bagging, and Random Forest to a dataset comprising 17 variables from the UCI repository for the early prediction of diabetes. Among the models assessed, the Random Forest algorithm showed outstanding performance, achieving the highest classification accuracy of 97% [4]. Additionally, Siva et al. proposed a heterogeneous stacked ensemble model for predicting diabetes, achieving an accuracy of 93.1%, which showed superior performance compared to existing models such as logistic regression (72%), Naïve Bayes (74.4%), and LDA (81%) [5].

Building upon these studies, the present work aims to compare five widely used machine learning models for predicting diabetes risk. Using the Pima Indians Diabetes Dataset, the study implements a structured framework that includes data cleaning, normalization, model training, and comparative evaluation.

The primary contribution of this research lies in its systematic benchmarking of model performance using both accuracy and the area under the ROC curve (AUC) as metrics. Additionally, this paper offers practical recommendations for selecting and applying robust ML models for diabetes risk prediction in clinical and public health settings.

2. Organization of the Text

2.1. Dataset

This study uses the Pima Indians Diabetes Dataset, which originates from the National Institute of Diabetes and Digestive and Kidney Diseases, the dataset includes 768 medical records of Pima Indian women aged 21 and older. The target variable is the diagnosis of diabetes: 268 individuals tested positive and 500 tested negatives. Each record comprises eight numeric attributes along with a binary target variable indicating diabetes presence (1) or absence (0). Table 1 presents Pima Indians Diabetes Dataset.

Table 1. Pima Indians Diabetes Dataset Description

Variable	Description
Preg	Number of times pregnant.
Plas	Plasma glucose concentration at 2 hours in an oral glucose tolerance test (OGTT).
Pres	Diastolic blood pressure (mm Hg).
Skin	Triceps skin fold thickness (mm).
Insu	2-hour serum insulin (μU/ml).
Mass	Body mass index.
Pedi	A function that scores the likelihood of diabetes based on family history.
Age	Age of the patient in years.
Class	Binary variable indicating the presence (1) or absence (0) of diabetes.

Table 2. Pima Indians Diabetes Dataset Sample

No.	preg	plas	pres	skin	insu	mass	pedi	age	class
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1

Table 2 show the pima indians diabetes data sample. To ensure data quality and enhance model performance, a structured preprocessing pipeline was applied. Zero values in the variables 'Plas', 'Pres', 'Skin', 'Insu', and 'Mass' were treated as missing and imputed using the median of the respective feature since they are not biologically plausible. All features were normalized using Min-Max scaling to rescale their values within a 0 to 1 range. This normalization process ensures that no individual feature dominates due to scale differences, especially when using distance-based models like SVM. The dataset was randomly split into training and testing sets, using an 80% to 20% ratio, to provide an unbiased evaluation of model generalizability.

2.2. Methods

Five classification models were selected for this study: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine (SVM). These models were chosen based on their widespread use in medical diagnostics and their ability to handle both linear and non-linear relationships.

Logistic regression is a widely used supervised learning algorithm for predicting categorical outcomes, particularly in binary classification tasks. Unlike linear regression, which models continuous values, logistic regression estimates the probability of an event occurring by applying the logit transformation (sigmoid function) to a linear combination of predictor variables. Diabetes prediction, which involves distinguishing between the presence and absence of the disease, is a typical binary classification problem, making logistic regression highly suitable for this application [6].

Decision Tree is a non-parametric supervised learning technique employed for classification and regression problems, where the model infers a set of hierarchical, rule-based decisions from the input features to predict the value of a target variable. Unlike many algorithms that require intensive data preprocessing, decision trees operate on raw data, support multi-output problems, and offer high interpretability through their transparent if-then-else logic structure. Owing to their simplicity, ability to handle both numerical and categorical variables, and robustness to partial violations of modeling assumptions, decision trees are a practical choice for diabetes prediction, particularly in clinical contexts where interpretability is critical [7].

Random Forest is an ensemble learning algorithm that constructs a multitude of decision trees using bootstrap samples from the training data and introduces feature-level randomness when determining node splits. This dual-randomization approach mitigates the high variance typically associated with individual decision trees by averaging their probabilistic predictions, resulting in improved generalization and reduced overfitting. Due to its robustness, ability to handle high-dimensional and noisy data, and effectiveness in capturing complex interactions among features, Random Forest is highly suitable for diabetes prediction tasks in diverse clinical and population health datasets [8].

Gradient Boosting is a powerful ensemble technique that builds models sequentially, where each new model corrects the residual errors of the previous ones, typically using decision trees as base learners. It optimizes performance through gradient descent by minimizing a specified loss function at each stage. Given the multifactorial nature of diabetes, gradient boosting is well-suited for capturing intricate patterns and interactions among risk variables in predictive modeling [9].

Support Vector Machine (SVM) is supervised learning model that identifies an optimal hyperplane to separate data into classes by maximizing the margin between support vectors, the critical data points nearest to the decision boundary. Due to their effectiveness in high-dimensional spaces and their ability to model non-linear relationships using kernel functions, SVMs are widely used in classification problems with complex data structures. In diabetes prediction, where patient data may include numerous interdependent biomedical variables, SVMs offer robust generalization performance and are particularly useful in cases with limited training samples relative to the number of features [10].

3. Experiment

Table 3. Logistic Regression Report

	precision	Recall	F1-score
0	0.80	0.83	0.81
1	0.67	0.62	0.64

Table 4. Decision Tree Report

	Precision	Recall	F1-score
0	0.78	0.77	0.78
1	0.60	0.62	0.61

Table 5. Random Forest Report

	Precision	Recall	F1-score
0	0.80	0.79	0.80
1	0.63	0.65	0.64

Table 6. SVM Report

	Precision	Recall	F1-score
0	0.78	0.84	0.81
1	0.67	0.58	0.62

Table 7. Gradient Boosting Report

	precision	Recall	F1-score
0	0.82	0.80	0.81
1	0.66	0.69	0.67

An analysis of the classification reports presented in Table 3, 4, 5, 6, and 7 reveals consistent patterns in model performance across the five machine learning algorithms. All models demonstrate higher precision, recall, and F1-scores for class 0 (non-diabetic) than for class 1 (diabetic), indicating that they are generally more effective at correctly identifying individuals without diabetes.

Among the models evaluated, Gradient Boosting exhibits the most balanced performance. It achieves a precision of 0.66, recall of 0.69, and F1-score of 0.67 for class 1, indicating a strong ability to correctly detect diabetic cases. Random Forest follows closely, with a class 1 F1-score of 0.64 and slightly better performance in class 0. Logistic Regression also performs reasonably well, but its recall of 0.62 for class 1 suggests a risk of false negatives, which can be critical in clinical contexts where undetected diabetes cases may delay treatment. In contrast, the Decision Tree and Support Vector Machine models demonstrate lower F1-scores for class 1, particularly with recall values of 0.62 and 0.58 respectively, suggesting that these models are less reliable for detecting true positive cases of diabetes.

Table 8. Five model summaries

Model	Accuracy	AUC
Logistic Regression	0.753247	0.822957
Decision Tree	0.714286	0.692929
Random Forest	0.740260	0.833425
SVM	0.746753	0.807163
Gradient Boosting	0.759740	0.819100

Table 8 presents a comparative summary of five machine learning models evaluated based on accuracy and the area under the ROC curve (AUC), providing a comprehensive view of both overall classification performance and discriminative ability. The results indicate that Gradient Boosting achieves the highest accuracy (0.7597), suggesting its effectiveness in correctly classifying the majority of both diabetic and non-diabetic instances. However, Random Forest surpasses all other models in AUC (0.8334), demonstrating superior capability in distinguishing between the two classes across varying decision thresholds.

Logistic Regression performs solidly with an accuracy of 0.7532 and AUC of 0.8230, making it a strong baseline model with the added benefit of interpretability. Support Vector Machine yields slightly lower accuracy (0.7468) and AUC (0.8072), suggesting moderate performance. In contrast,

Decision Tree records the lowest scores in both metrics (accuracy = 0.7143, AUC = 0.6929), reaffirming its limited generalizability and greater susceptibility to overfitting.

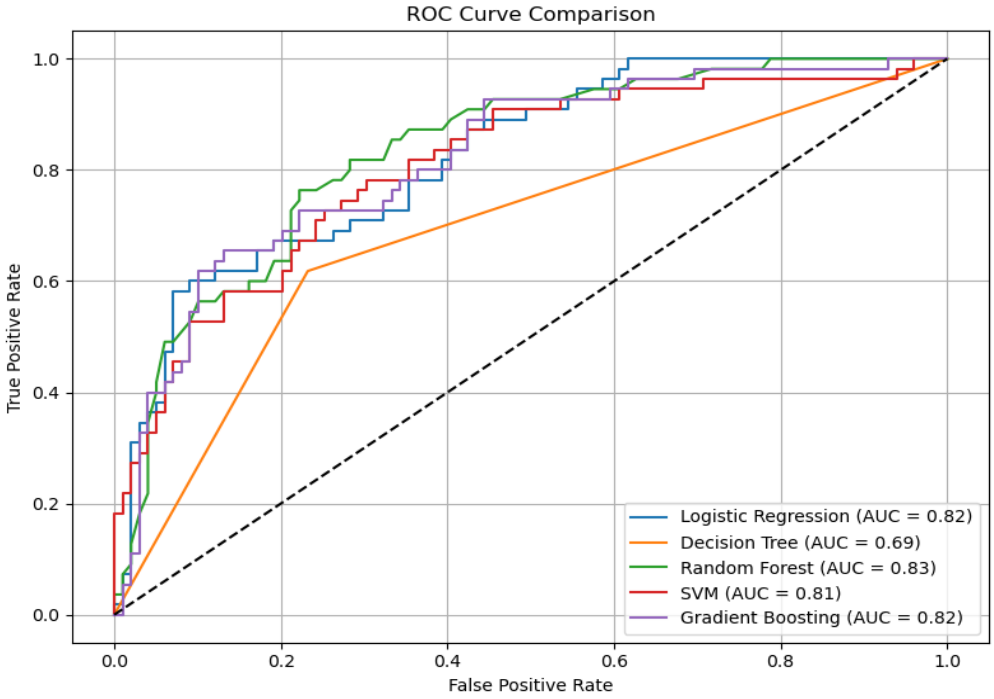

Figure 1. ROC curve (Picture credit: Original)

Figure 1 further visualizes model discrimination through ROC curves. Random Forest consistently outperforms others with its curve closest to the top-left corner, aligning with its superior AUC. Logistic Regression and Gradient Boosting follow closely, with steep curves indicating strong sensitivity and specificity. SVM also performs reasonably well. In contrast, Decision Tree's ROC curve is flatter and falls below the others, reinforcing its weaker overall performance.

In summary, ensemble models—particularly Random Forest and Gradient Boosting—consistently achieve higher accuracy, greater sensitivity for diabetic cases, and stronger overall discriminative ability. These characteristics make them well-suited for clinical applications where early and accurate identification of high-risk individuals is essential.

4. Conclusion

This study conducted a comparative evaluation of five machine learning models for predicting type 2 diabetes using clinical attributes. The results demonstrated that ensemble methods—particularly Random Forest and Gradient Boosting—consistently outperform traditional models in terms of overall accuracy and discriminatory power. Random Forest achieved the highest AUC (0.833), while Gradient Boosting yielded the best classification accuracy (75.97%) and strong balance across both classes.

Despite these encouraging results, a notable limitation across all models was the relatively low score for class 1 (diabetic patients), ranging from 0.61 to 0.67. This highlights an important challenge in diabetes risk prediction: the reduced sensitivity of models when identifying true positive cases. In real-world healthcare settings, such false negatives could result in delayed diagnosis and treatment, undermining the utility of predictive screening tools. Therefore, improving model performance on minority classes remains an essential direction for further refinement.

The findings of this research support the integration of ensemble-based machine learning models into early screening frameworks, where predictive reliability and interpretability are both important. Future work should focus on incorporating more diverse and representative datasets, including

behavioral and genetic to improve generalizability. Additionally, enhancing class-specific recall and reducing algorithmic bias will be critical to making these tools clinically robust and ethically sound.

References

[1] Global increase in diabetes prevalence imposes a substantial health and economic burden: Published by Journal of Health Economics and Outcomes Research. Journal of Health Economics and Outcomes Research, 2021. Available: https://jheor.org/post/1265-global-increase-in-diabetes-prevalence-imposes-a-substantial-health-and-economic-burden.

[2] Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2021. Results. Institute for Health Metrics and Evaluation, 2024.

[3] Sisodia, D., & Sisodia, D. S. Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578 - 1585, 2018.

[4] Laila, U. E., Mahboob, K., Khan, A. W., Khan, F., & Taekeun, W. An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. Sensors, 22 (14), 5247, 2022.

[5] Hasan, M. K., Saeed, R. A., Alsuhibany, S. A., & Abdel-Khalek, S. An empirical model to predict the diabetic positive using stacked ensemble approach. Frontiers in Public Health, 9, 792124, 2022.

[6] Linear regression. scikit-learn. (n.d.). Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[7] Decision trees. scikit-learn. (n.d.). Available: https://scikit-learn.org/stable/modules/tree.html#decision-trees.

[8] Random forests. scikit-learn. (n.d.). Available: https://scikit-learn.org/stable/modules/ensemble.html#random-forests.

[9] Histogram-Based Gradient Boosting. scikit-learn. (n.d.). Available: https://scikit-learn.org/stable/modules/ensemble.html#histogram-based-gradient-boosting.

[10] Support vector machines. scikit-learn. (n.d.). Available: https://scikit-learn.org/stable/modules/svm.html#support-vector-machines.