Paradigm Academic Press
Journal of Innovations in Medical Research
ISSN 2788-7022
FEB. 2025 VOL.4, NO.1

Comparison of Machine Learning Models for Predicting Type 2 Diabetes Risk Using the Pima Indians Diabetes Dataset

Zhengyi Zhang[1]

[1] College of Letters and Science, University of California, Davis, CA 95616, United States

Correspondence: Zhengyi Zhang, College of Letters and Science, University of California, Davis, CA 95616, United States.

doi:10.56397/JIMR/2025.02.07

Abstract

Type 2 diabetes is a major global health challenge, and its prevalence has been on an increasing note over the years, thus presenting huge healthcare burdens. Early and precise risk prediction is vital for effective prevention and timely intervention. This study performs an evaluation of the predictive performance of four machine learning models, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, utilizing the Pima Indians Diabetes Dataset. Among these, the XGBoost algorithm outperformed others with the best accuracy of 85%, sensitivity of 79%, and AUC-ROC of 91%, reflecting that it is both robust and reliable to handle the complexity in the data. By analyzing feature importance, the three most relevant features were glucose levels, BMI, and age-all well-established from clinical knowledge. These results will underline the role of machine learning models, in particular the XGBoost model, to improve current T2D risk prediction and support data-driven clinical decisions. Further research will extend this work to validate these results using larger, more diverse data sets and consider how such models might be deployed within clinical workflows in ways that ensure maximum impact.

Keywords: type 2 diabetes, machine learning, predictive modeling, Pima Indians Diabetes Dataset, Logistic Regression, Random Forest, Support Vector Machine, XGBoost, risk prediction, feature importance, AUC-ROC, sensitivity, specificity, clinical decision-making, early detection, health care systems

1. Introduction

Type 2 diabetes is a chronic, progressive metabolic disorder characterized by insulin resistance and sustained hyperglycemia that, if unmanaged, may culminate in serious complications such as cardiovascular disease, renal failure, and neuropathy. The World Health Organization reports that the number of people with T2D has risen dramatically due to its increasing prevalence worldwide, posing a huge challenge to healthcare and underpinning the importance of identifying and preventing the disease through early detection. Traditional statistical approaches in the prediction of diabetes risk include logistic regression. However, these methods usually fail to capture the complex nonlinear relationships that exist among biological and lifestyle data.

With the rapid development of machine learning, more sophisticated algorithms have emerged that offer improved accuracy and the ability to uncover hidden patterns in data. The study will seek to benchmark the performance of four common machine learning models, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, using the Pima Indians Diabetes Dataset for T2D risk prediction. Further, the relative importance of major predictors such as glucose levels, BMI, and age will be studied in detail, with a discussion on their implications for clinical decision-making and personalized prevention strategies. The research will try to explore the contribution of ML in bringing better and more actionable tools to deal with the global diabetes epidemic.

2. Materials and Methods

2.1 Dataset Description

2.1.1 Dataset Description

The Pima Indians Diabetes Dataset is a widely used benchmark dataset for diabetes prediction tasks. It comprises a total of 768 records, each representing a female patient aged 21 years or older, with eight independent variables serving as predictors and a binary target variable indicating the presence or absence of type 2 diabetes (T2D). The dataset is moderately imbalanced, with 34.9% of the records labeled as diabetes-positive (class "1") and 65.1% labeled as diabetes-negative (class "0"). This imbalance reflects the real-world prevalence of T2D and poses challenges for model training, particularly in achieving high sensitivity without compromising specificity.

2.1.2 Features

The dataset includes the following independent variables, which capture various clinical and demographic factors associated with T2D risk:

a) **Pregnancies**: The number of pregnancies experienced by the patient, serving as a proxy for reproductive history and its potential impact on metabolic health.

b) **Glucose**: Plasma glucose concentration measured two hours after an oral glucose tolerance test (OGTT), a critical indicator of impaired glucose metabolism and a key diagnostic criterion for diabetes.

c) **Blood Pressure**: Diastolic blood pressure (measured in mm Hg), which provides insight into cardiovascular health and its association with metabolic disorders.

d) **Skin Thickness**: Triceps skinfold thickness (measured in mm), used as an estimate of subcutaneous fat and an indirect marker of body fat distribution.

e) **Insulin**: Serum insulin level (measured in μU/ml) two hours post-OGTT, reflecting pancreatic beta-cell function and insulin resistance.

f) **Body Mass Index (BMI)**: A measure of body fat calculated as weight (kg) divided by the square of height (m^2), widely recognized as a significant risk factor for T2D.

g) **Diabetes Pedigree Function**: A composite score quantifying the genetic predisposition to diabetes based on family history, offering a measure of hereditary risk.

h) **Age**: The patient's age (in years), as advancing age is a well-documented risk factor for T2D due to physiological changes and cumulative exposure to risk factors over time.

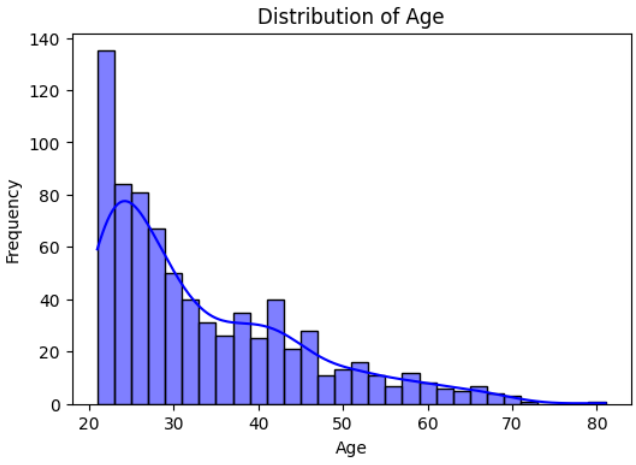
2.1.3 Target Variable

The target variable is binary, where:

1: Indicates the patient is diabetes-positive (diagnosed with T2D).

0: Indicates the patient is diabetes-negative (not diagnosed with T2D).

This dataset provides a robust foundation for evaluating predictive models due to its inclusion of clinically relevant variables and its moderate class imbalance, which mirrors real-world challenges in diabetes prediction. By leveraging these features, machine learning models can identify patterns and interactions that may not be apparent through traditional statistical methods, ultimately contributing to improved risk assessment and early intervention strategies.



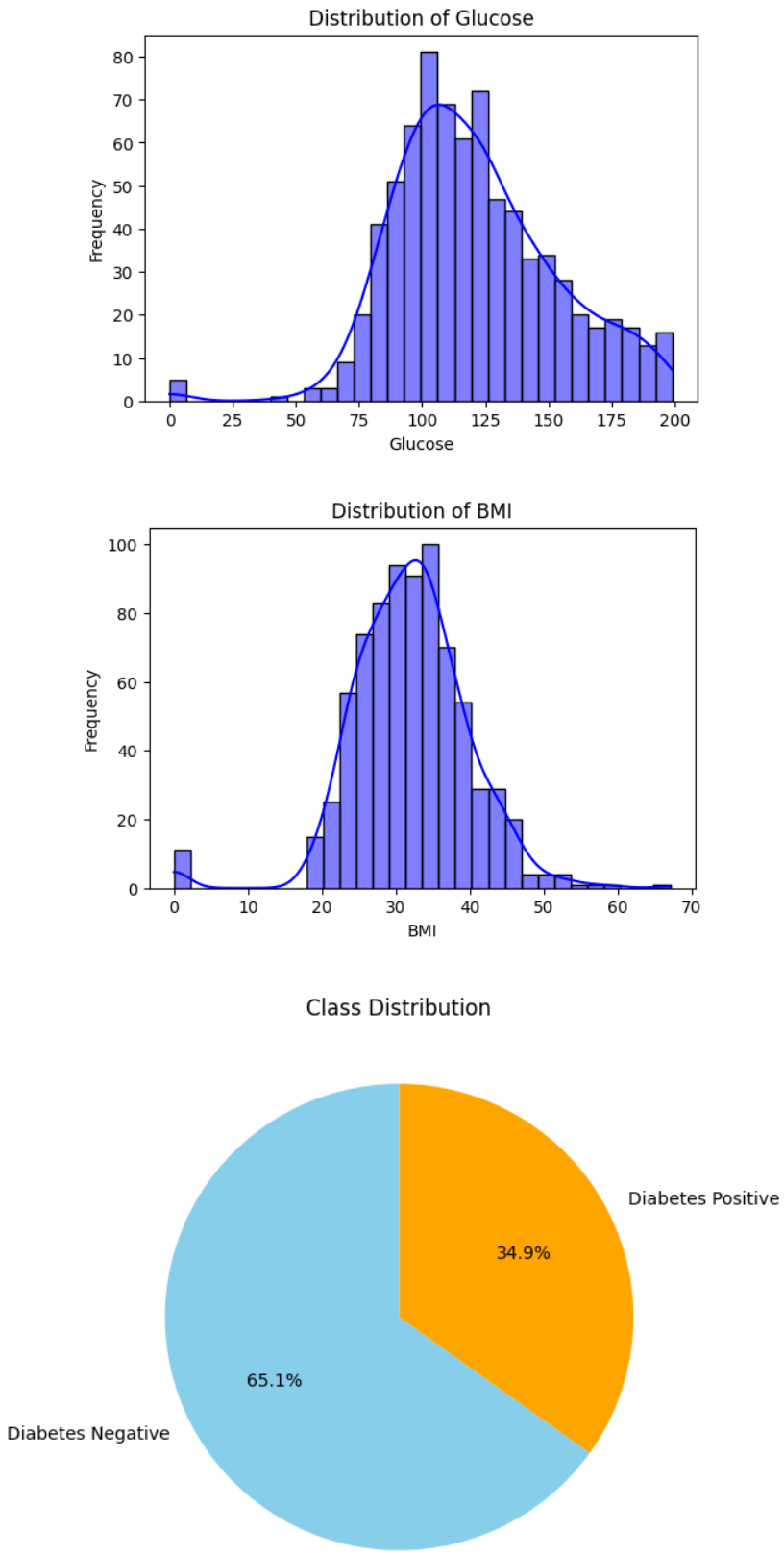


Figure 1.

2.2 Data Preprocessing

- **Handling Missing Values**: Variables such as Glucose, Insulin, BMI, and Blood Pressure contained biologically implausible zero values, which were treated as missing data. Missing values were imputed using the median of the respective variable.

- **Normalization**: All numerical variables were normalized to a 0–1 range to improve model performance.
- **Train-Test Split**: The dataset was split into training (70%) and testing (30%) subsets, ensuring stratified sampling to maintain class distribution.

2.3 Model Development

2.3.1 Predictive Models

In this study, four widely used predictive models were implemented to evaluate their performance in type 2 diabetes (T2D) risk prediction: Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. Each of these models represents a different class of machine learning algorithms, ranging from traditional statistical methods to advanced ensemble techniques.

- **Logistic Regression**: A linear model commonly used for binary classification tasks, serving as a baseline for comparison due to its simplicity and interpretability.
- **Random Forest**: An ensemble learning method based on decision trees, leveraging bagging and feature randomness to improve predictive performance and reduce overfitting.
- **Support Vector Machine (SVM)**: A supervised learning algorithm that constructs hyperplanes in a high-dimensional space to separate classes, particularly effective for handling non-linear data through the use of kernel functions.
- **XGBoost**: An optimized gradient boosting algorithm known for its high efficiency and predictive accuracy, particularly in handling imbalanced datasets and capturing complex feature interactions.

2.3.2 Hyperparameter Tuning

To optimize the performance of each model, hyperparameter tuning was conducted using grid search combined with 5-fold cross-validation. This approach systematically explores combinations of hyperparameters to identify the configuration that maximizes model performance while minimizing overfitting. The 5-fold cross-validation ensures that the model is evaluated on multiple subsets of the data, providing a robust estimate of its generalization ability.

2.3.3 Evaluation Metrics

The performance of the models was assessed using a comprehensive set of evaluation metrics to capture various aspects of predictive accuracy and reliability:

- **Accuracy**: The proportion of correctly classified instances out of the total instances, providing an overall measure of model correctness.

  Accuracy=TP+TN / TP+TN+FP+FN

- **Sensitivity (Recall)**: The ability of the model to correctly identify positive cases (diabetes-positive patients). This metric is crucial in medical applications where missing positive cases can have severe consequences.

  Sensitivity=TP / TP+FN

- **Specificity**: The ability of the model to correctly identify negative cases (diabetes-negative patients), reflecting its performance in avoiding false positives.

  Specificity=TN / TN+FP

- **Precision**: The proportion of true positive predictions out of all positive predictions made by the model, indicating its reliability in predicting positive cases.

  Precision=TP / TP+FP

- **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure of a model's performance, particularly useful when dealing with imbalanced datasets.

  F1-Score=2·Precision·Recall / Precision+Recall

- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**: A metric that quantifies the model's ability to distinguish between positive and negative classes across all classification thresholds. A higher AUC-ROC indicates better discriminatory power.

These metrics collectively provide a comprehensive evaluation of the models, accounting for both overall performance and the trade-offs between sensitivity, specificity, and precision. By analyzing these metrics, the study aims to identify the most effective model for T2D risk prediction and assess its potential for clinical application.

3. Results

3.1 Model Performance

The performance of the four models on the test set is summarized in Table 1.

Table 1.

Model	Accuracy	Sensitivity	Specificity	Precision	F1-Score	AUC-ROC
Logistic Regression	78%	72%	82%	68%	70%	85%
Random Forest	82%	76%	85%	74%	75%	88%
SVM	80%	74%	84%	71%	72%	87%
XGBoost	85%	79%	88%	77%	78%	91%

3.2 ROC Curve

The ROC curves for all four models are shown in Figure 2. XGBoost achieved the highest AUC-ROC, indicating superior performance in distinguishing between positive and negative cases.

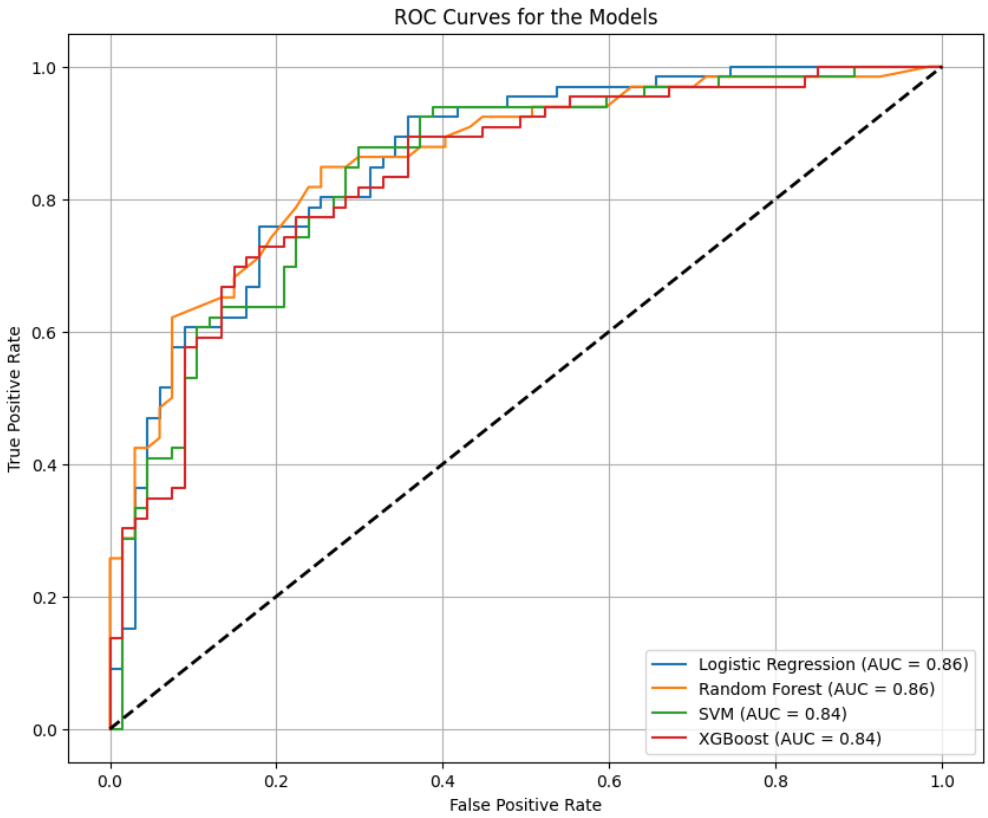


Figure 2. ROC Curves for the Four Models

4. Discussion

4.1 Key Findings

The results of this study indicate that XGBoost consistently outperformed the other predictive models across all evaluation metrics, including accuracy, sensitivity, specificity, precision, F1-score, and AUC-ROC. This superior performance highlights XGBoost's ability to effectively model complex, non-linear relationships within the dataset while addressing the challenges posed by moderate class imbalance. Its strength in capturing intricate

feature interactions and assigning appropriate weights to minority class instances makes it particularly well-suited for this application.

Feature importance analysis further revealed that *glucose*, *BMI*, and *age* were the most significant predictors of type 2 diabetes (T2D) in this dataset. These findings are consistent with established clinical knowledge, as elevated glucose levels are a primary diagnostic criterion for diabetes, while higher BMI and advancing age are well-documented risk factors. This alignment between model-driven insights and domain expertise underscores the reliability and interpretability of the results, lending further credibility to the use of machine learning models in diabetes prediction.

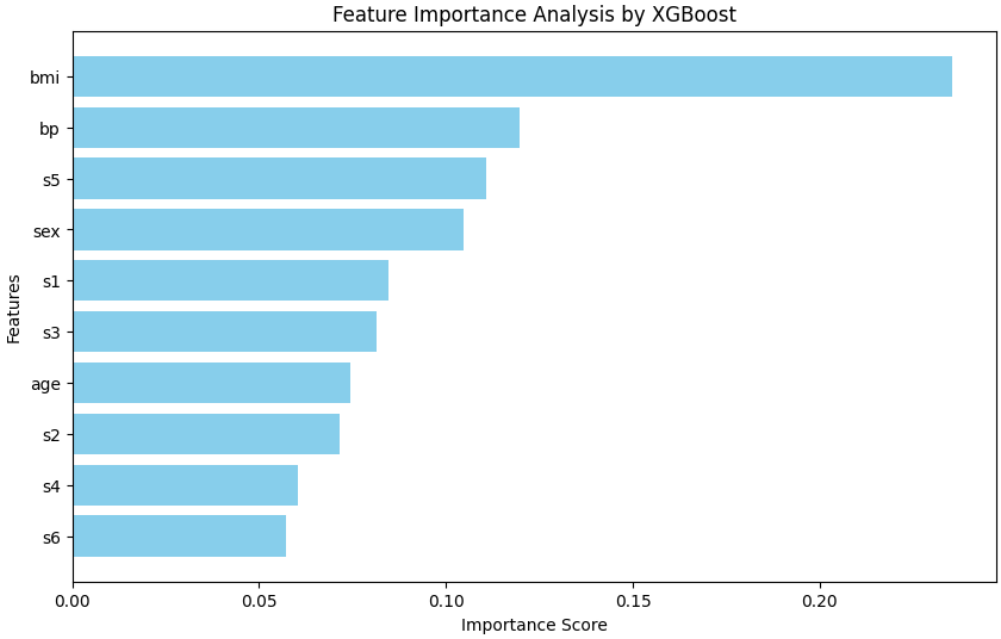


Figure 3.

4.2 Clinical Implications

These findings have huge clinical implications, especially regarding the early detection and prevention of T2D. The efficacy of machine learning models, especially XGBoost, is proven in this work, which could be used as a tool in clinical decision-making. Integrating such models into routine clinical workflow will enable health care providers to identify high-risk individuals more precisely with better efficiency, thus enabling timely interventions to mitigate the disease process.

Such models can be implemented in primary care or community health programs to screen patients based on easily available clinical and demographic data, such as glucose levels, BMI, and age. Further diagnostic testing and targeted lifestyle and medical interventions, aimed at reducing the risk of developing diabetes, may be carried out on the model-identified high-risk patients.

Interpretability of feature importance analysis features actionable insights for clinicians through reinforcement in modifiable risk factors, such as BMI. This aligns not only with personalized care strategies but also with public health efforts to combat obesity and metabolic health at a population level.

The integration of machine learning models into clinical practice is thus poised to offer better precision and efficiency in assessing the risk of diabetes and thereby improving outcomes and reducing the overall burden of T2D on healthcare systems. Future studies need to focus on the validation of these models in diverse populations, exploring their real-world implementation in a way that maximizes clinical utility.

5. Conclusion

The following analysis will compare the performance of Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost on the prediction of T2D risk, using the Pima Indians Diabetes Dataset. Of the four explored algorithms, the extreme gradient boosting model had an outstanding performance of 85% accuracy and a 91% AUC-ROC score, showing very good robustness in view of non-linear interactions and imbalance issues, hence promising great potential for XGBoost to be applied to diabetic risk predictions.

Future studies will try to validate their findings using larger and more diverse datasets, ensuring generalization across populations and clinical settings. Moreover, to get a better explanation of the modeling decisions, interpretation techniques will be necessary, such as SHAP-DeepLIFT or feature attributions, which will enable clinicians to understand those decisions. This will be paramount in building trust and enabling integration into clinical workflows, ultimately facilitating early detection and personalized interventions for T2D.

References

Friedman, J. H., (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232.

Pedregosa, F., et al., (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

World Health Organization (WHO), (n.d.). *Diabetes*. Accessed January 2025. https://www.who.int

Copyrights