Publisher
Vandana Publications
www.vandanapublications.com

International Journal of Engineering and Management Research

2025 Volume 15 Number 3 June

E-ISSN:2250-0758
P-ISSN:2394-6962

Research Article

Predictive Analytics

Comparative Analysis of Machine Learning Models for Diabetes Prediction

Patil P[1*], Lawand D[2], Gambas M[3], Gaikwad D[4]

DOI:10.5281/zenodo.15867611

[1*] Pratiksha Patil, Department of Mathematics, Ramsheth Thakur College of Commerce and Science, Kharghar, Maharashtra, India.

[2] Deepali Lawand, Department of Mathematics, Ramsheth Thakur College of Commerce and Science, Kharghar, Maharashtra, India.

[3] Mohit Gambas, Ramsheth Thakur College of Commerce and Science, Kharghar, Maharashtra, India.

[4] Deepak Gaikwad, Department of Physics, KTSP Mandal's KMC College, Khopoli, Maharashtra, India.

Diabetes is a chronic health condition affecting millions worldwide, and early detection plays a vital role in effective disease management and prevention. In this study, we conduct a comparative analysis of four machine learning models—Logistic Regression, Random Forest, Gradient Boosting, and Linear Regression—applied to the Pima Indian Diabetes dataset obtained from Kaggle. The dataset comprises diagnostic measurements of female patients aged 21 and above of Pima Indian heritage. Each model is evaluated using key classification metrics, including accuracy, precision, recall, and F1-score. Among the models, Logistic Regression and Gradient Boosting achieved the highest accuracy of 75%, while Random Forest and Linear Regression showed slightly lower performance at 72% and 73.16%, respectively. The study highlights the effectiveness of ensemble methods and traditional classifiers in predicting diabetes outcomes and provides insight into their relative strengths for clinical decision support systems. These results suggest that machine learning can be a valuable tool in aiding early diagnosis and improving patient care strategies.

Keywords: Diabetes Prediction, Machine Learning, Logistic Regression, Random Forest, Gradient Boosting, Linear Regression, Predictive Analytics.

Corresponding Author

Pratiksha Patil, Department of Mathematics, Ramsheth Thakur College of Commerce and Science, Kharghar, Maharashtra, India.
Email: pratikshapatil@rtccs.edu.in

How to Cite this Article

Patil P, Lawand D, Gambas M, Gaikwad D, Comparative Analysis of Machine Learning Models for Diabetes Prediction. Int J Engg Mgmt Res. 2025;15(3):89-93.
Available From
https://ijemr.vandanapublications.com/index.php/j/article/view/1771

To Browse

Manuscript Received	Review Round 1	Review Round 2	Review Round 3	Accepted
2025-05-13	2025-06-07			2025-06-21

Conflict of Interest	Funding	Ethical Approval	Plagiarism X-checker	Note
None	Nil	Yes	4.32	

1. Introduction

Diabetes mellitus, particularly Type 2 diabetes, is one of the most pervasive and growing public health challenges worldwide. It is characterized by chronic hyperglycemia due to insulin resistance or insufficient insulin production. According to the International Diabetes Federation (2021), approximately 537 million adults were living with diabetes in 2021, a number expected to rise to 783 million by 2045 if left unchecked. The disease not only impacts individual quality of life but also imposes significant economic and healthcare burdens globally. Early diagnosis and intervention are crucial to managing diabetes and mitigating associated complications such as cardiovascular disease, neuropathy, nephropathy, and retinopathy [1].

Traditional diagnostic methods like fasting plasma glucose and oral glucose tolerance tests, although effective, may be limited by accessibility, time requirements, and cost. In recent years, machine learning (ML) has emerged as a powerful tool to enhance medical diagnostics by identifying complex, nonlinear patterns in patient data that might be overlooked by conventional techniques [2]. ML models can be trained to predict the likelihood of diseases such as diabetes based on various physiological and demographic parameters, thereby enabling more timely and personalized interventions.

This study utilizes the Pima Indian Diabetes dataset, a widely cited benchmark in medical data science. Provided by the National Institute of Diabetes and Digestive and Kidney Diseases, this dataset contains data from female patients of Pima Indian heritage aged 21 years and older. It includes eight clinical variables: glucose concentration, diastolic blood pressure, skinfold thickness, serum insulin, BMI, diabetes pedigree function, and age, along with a binary outcome indicating diabetes status.

We apply and compare four machine learning models: Logistic Regression, Random Forest, Gradient Boosting, and Linear Regression (adapted for classification via thresholding). Each algorithm is evaluated using standard classification metrics—accuracy, precision, recall, and F1-score—focusing particularly on the correct identification of diabetic individuals. Our results show that Logistic Regression and Gradient Boosting achieved the highest accuracy at 75%, with Logistic Regression also delivering strong performance in recall and F1-score. Random Forest and threshold Linear Regression performed slightly lower but still demonstrated their potential in clinical classification tasks.

The motivation for this research stems from the increasing integration of ML in healthcare and its potential to augment clinical decision-making. As digital health records and wearable technologies continue to expand, ML-based decision support systems can serve as critical tools in early disease detection and health risk assessments [3]. However, issues of model interpretability, fairness, and clinical applicability remain central to successful implementation.

This paper presents a comparative evaluation of these models in predicting diabetes outcomes and offers insights into their practical strengths and limitations. The remainder of the paper details our methodology, model training and validation processes, results, and implications for real-world clinical use.

2. Methodology

Dataset Description

The dataset employed in this study is the Pima Indian Diabetes Dataset, sourced from Kaggle [4], a popular platform for data science competitions and datasets. This dataset provides medical and demographic information for female patients of Pima Indian heritage aged 21 years and older, who are at high risk for developing diabetes. It contains 768 observations with 8 clinical features and 1 binary target indicating diabetes status.

The input features are:

- Glucose – Plasma glucose concentration
- Blood Pressure – Diastolic blood pressure (mm Hg)
- Skin Thickness – Triceps skinfold thickness (mm)
- Insulin – 2-hour serum insulin (mu U/ml)
- BMI – Body Mass Index (weight in kg/(height in m)^2)
- Diabetes Pedigree Function – A function which scores the likelihood of diabetes based on family history
- Age – Age in years

The target variable, Outcome, is binary: 1 represents a positive diabetes diagnosis, and 0 represents a negative result.

This dataset is extensively used for machine learning tasks in the healthcare domain due to its balanced nature and the relevance of features. However, some entries contain biologically implausible zeros (e.g., zero blood pressure or glucose), which necessitates careful preprocessing such as replacement or imputation to ensure data quality.

In this research, we use this dataset to train and compare multiple classification models to predict the presence of diabetes. The preprocessing involved handling missing values, normalization, and data splitting into training and testing sets.

3. Machine Learning Models

Model Descriptions

In this study, we evaluated the performance of four widely-used supervised machine learning models for the prediction of diabetes: Logistic Regression, Random Forest, Linear Regression and Gradient Boosting Classifier. Each model was selected for its proven effectiveness in medical diagnostics and classification problems.

1. Logistic Regression
Logistic Regression is a linear classification model that estimates the probability of a binary outcome using the logistic function. It is widely used in healthcare prediction tasks due to its simplicity and interpretability. In the context of diabetes prediction, it models the log-odds of the outcome as a linear combination of input features [5].

2. Random Forest
Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce over fitting. It is particularly effective for classification tasks with complex relationships between features. In medical datasets with possible noise and non-linearity, Random Forests offer robustness and high performance [6].

3. Linear Regression
Linear regression is a foundational statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It is widely used in prediction and forecasting, as well as in various scientific fields due to its simplicity and interpretability. In the context of battery data, linear regression can model trends such as the relationship between cycle number and capacity or efficiency metrics. [7].

4. Gradient Boosting Classifier
Gradient Boosting Classifier (GBC) is an ensemble learning algorithm that combines multiple weak learners (usually decision trees) to create a strong classifier. It builds models sequentially, where each new model corrects the errors made by the previous models. The algorithm is effective for binary classification tasks and can handle both linear and non-linear relationships. In battery failure prediction, GBC can classify cells as good or bad based on features like CEF metrics, enabling early detection of cell degradation. [8].

4. Results & Discussion

Model Evaluation and Performance Analysis

Four machine learning models—Logistic Regression, Gradient Boosting, Random Forest, and Linear Regression—were evaluated for diabetes classification using the PIMA Indian Diabetes dataset. Performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices.

Logistic Regression achieved an accuracy of 75% (Figure 1), with strong classification of non-diabetic individuals (Precision: 0.81, Recall: 0.80) and moderate performance for diabetic cases (Precision: 0.65, Recall: 0.67). Gradient Boosting matched this accuracy (75%, Figure 2), showing slightly more balanced class-wise performance with F1-scores of 0.80 and 0.65 for Classes 0 and 1, respectively.

Random Forest (Figure 3) showed lower accuracy at 72%, with reduced recall (0.62) and precision (0.61) for diabetic cases, indicating diminished reliability for detecting positive cases. Linear Regression, adapted for classification via thresholding, achieved a slightly lower accuracy of 73.16% (Figure 4) and is less suitable for binary classification tasks.

These results suggest that Logistic Regression and Gradient Boosting are the most effective models for diabetes prediction in this dataset. The corresponding confusion matrices (Figures 1–4) visually support this conclusion, highlighting class-specific prediction strengths and weaknesses.

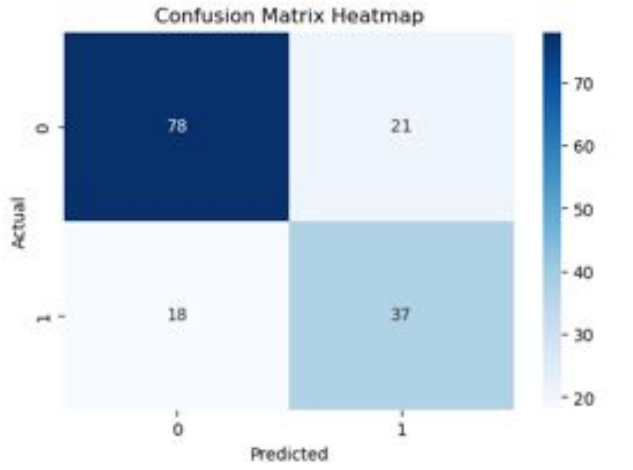

Figure 1: Confusion Metrics of Logistic Regression Model

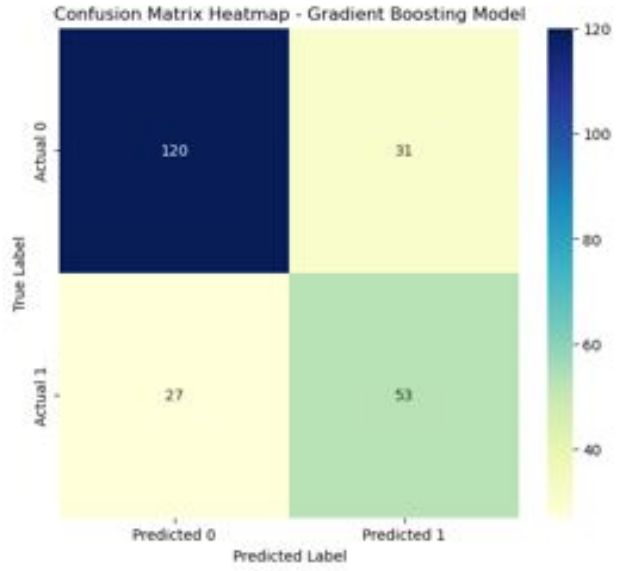

Figure 2: Confusion Metrics of Gradient Boosting Model

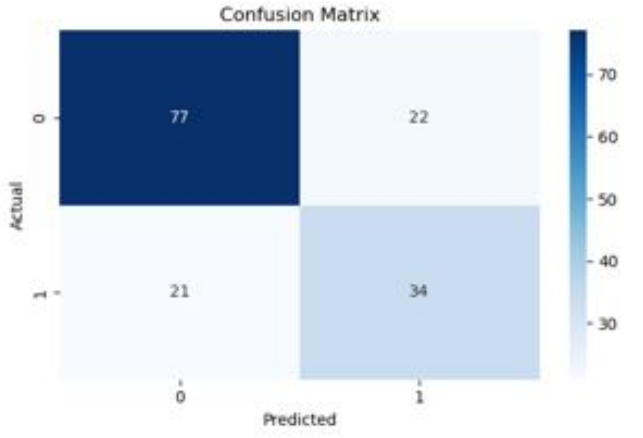

Figure 3: Confusion Metrics of Random Forest Model

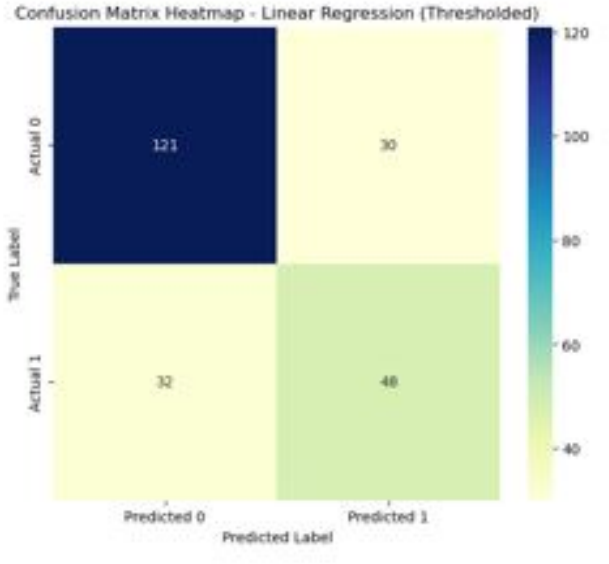

Figure 4: Confusion Metrics of Linear Regression Model

5. Conclusion

This study reinforces the promise of machine learning (ML) in supporting early diagnosis of diabetes mellitus, particularly Type 2 diabetes, by leveraging widely available clinical and demographic parameters. Using the Pima Indian Diabetes dataset, we implemented and compared the predictive performance of four ML algorithms: Logistic Regression, Random Forest, Gradient Boosting, and Linear Regression (adapted for classification). Among these, Logistic Regression and Gradient Boosting emerged as the most effective models, both achieving an accuracy of 75%. Logistic Regression further demonstrated strong recall and F1-score, making it particularly suitable for high-stakes medical contexts where minimizing false negatives is critical.

These results emphasize the feasibility of integrating ML models into healthcare workflows for timely identification of diabetes risk, especially in resource-constrained settings where traditional diagnostic tests may be limited. Simpler models like Logistic Regression offer transparency and ease of deployment, while more complex models such as Gradient Boosting provide improved performance by capturing intricate patterns in the data. Despite their advantages, challenges related to model generalizability, interpretability, and ethical deployment persist.

References

[1] International Diabetes Federation. (2021). *IDF diabetes atlas.* (10th ed.). Brussels, Belgium. Retrieved from: https://diabetesatlas.org/.

[2] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.

[3] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.

[4] Kaggle. (n.d.). *Pima Indians diabetes database*. Retrieved May 6, 2025, from: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[5] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. (3rd ed.). Wiley.

[6] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[7] Faraway, J. J. (2016). *Linear models with R*. (2nd ed.). CRC Press.

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