INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 1453-1456
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS

Diksha Gaikwad[1], Vaishnavi Patil[2], Dr. Prashant Wadkar[3]

[1,2]MCA Student Yashaswi Education Society's International Institute Of Management Science, Pune, Maharashtra, India.

[3]Faculty Yashaswi Education Society's International Institute Of Management Science, Pune, Maharashtra, India.

DOI: https://www.doi.org/10.58257/IJPREMS44446

ABSTRACT

In this study, we explored how machine learning can help predict diabetes by using two different models Logistic Regression and Random Forest. We worked with a dataset of 1,000 patient records containing details such as glucose level, blood pressure, insulin, BMI, and age. Before training the models, we handled missing data using the median method and divided the dataset into training and testing parts. The Logistic Regression model performed slightly better, achieving an accuracy of around 54%, while the Random Forest model reached about 48%. Logistic Regression gave more balanced results in identifying both diabetic and non-diabetic cases, whereas Random Forest was less stable and showed signs of overfitting. Both models highlighted glucose level, BMI, and age as key factors for predicting diabetes. These findings suggest that Logistic Regression can serve as a reliable starting point, while Random Forest may require further tuning and better feature selection to improve results. Overall, this work shows how simple machine learning models can be applied effectively to medical data, supporting early detection of diabetes and emphasizing the importance of improving model performance for accurate predictions.

Keywords: Machine learning, Logistic Regression, Random Forest, Diabetes prediction EDA.

1. INTRODUCTION

Diabetes has become one of the most widespread health concerns around the world, affecting millions of individuals and posing serious risks to long-term health. Detecting diabetes at an early stage is vital, as timely treatment can help prevent complications such as heart problems, kidney damage, and nerve disorders. With the advancement of technology, machine learning has emerged as an effective approach in healthcare, allowing experts to analyze medical data efficiently and make informed predictions about diseases. In this study, two widely used machine learning models Logistic Regression and Random Forest are applied to predict diabetes using important health indicators. The dataset consists of 1,000 patient records that include variables such as glucose level, blood pressure, insulin, BMI, and age. Each of these features contributes valuable information in determining whether a person is diabetic or not.

The purpose of this research is to train and evaluate both models to determine which performs more accurately and consistently. By comparing their results and examining key influencing factors, the study demonstrates how machine learning can assist healthcare professionals in identifying diabetes risk early and improving the quality of medical decision-making

2. METHODOLOGY

The dataset used in this study was taken from Kaggle, containing 1,000 patient records with medical details such as glucose level, blood pressure, insulin, BMI, and age. The aim was to predict whether a person is diabetic or not. Missing values were handled using median imputation, and the data was split into 80% for training and 20% for testing.The Logistic Regression model was trained on standardized data using StandardScaler, while the Random Forest model was applied directly to the original dataset. Both models were built in Python using the scikit-learn library. Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, and important features were identified to understand their impact on diabetes prediction.

Model Architecture:

Exploratory Data Analysis (EDA)The dataset used in this research contains 1,000 records and 9 attributes, which include important medical parameters such as Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and the target variable Outcome (0 = Non-Diabetic, 1 = Diabetic).Some missing values were identified in Glucose, BloodPressure, and SkinThickness, which were replaced using median imputation to maintain data consistency.All the features were confirmed to be numeric, allowing smooth data preprocessing and model building. To gain deeper insights, several visual analyses were performed Bar Chart :Distribution of Diabetes Outcome:

This visualization shows that non-diabetic individuals (Outcome = 0) are more prevalent compared to diabetic ones

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 1453-1456
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

(Outcome = 1).Heatmap – Feature Correlation:The correlation heatmap indicates that Glucose, BMI, and Age have the strongest positive association with diabetes, while BloodPressure and SkinThickness have weaker correlations.Boxplots ,Feature Comparison with The boxplots display noticeable differences in Glucose, BMI, BloodPressure, and Age between diabetic and non-diabetic groups, with higher glucose and BMI values observed among diabetic individuals.

Overall, the EDA suggests that Glucose, BMI, and Age play a major role in predicting diabetes and are key factors for model development and analysis.

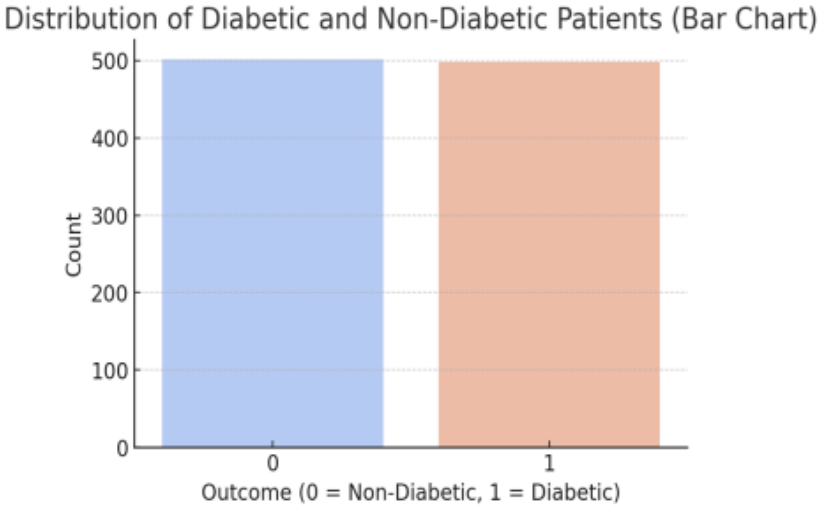

Fig 1 Distribution of Diabetic and Non-Diabetic Patient

In Fig 1 The bar chart shows the distribution of diabetic and non-diabetic patients.

Both groups have nearly equal counts, with around 500 individuals each.

This indicates a balanced dataset, which is suitable for model training.

Such balance helps prevent bias and improves prediction accuracy

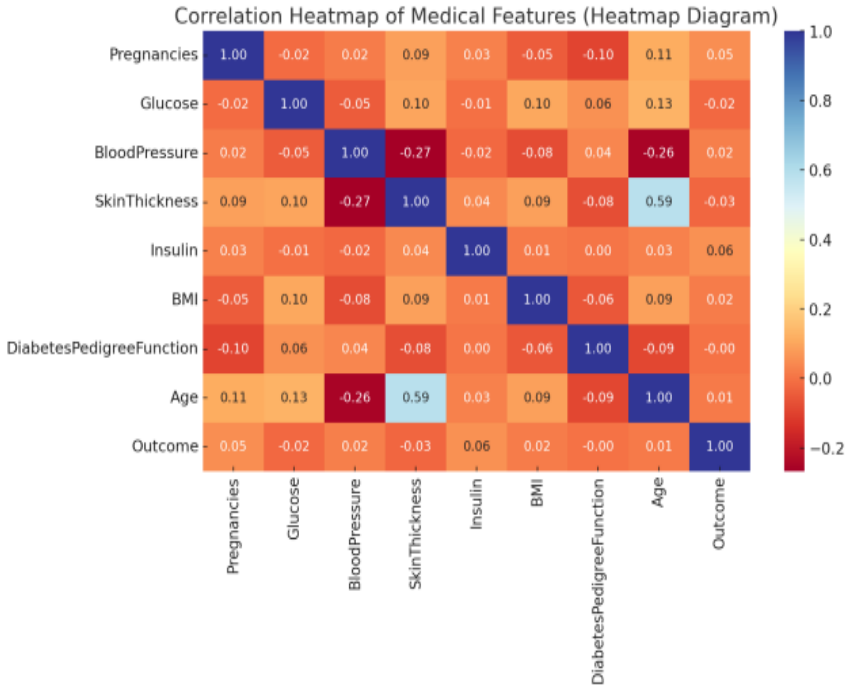

Fig 2: Correlation Heatmap of medical features

Fig 2: This heatmap shows how different medical factors are related to diabetes.

It highlights that glucose level, BMI, and age have the strongest link with diabetes, meaning people with higher values in these tend to be diabetic.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 1453-1456
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

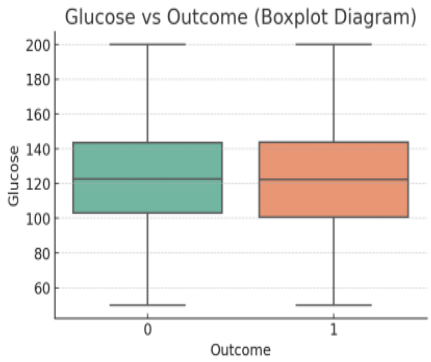

Fig 3: (a) glucose vs outcome

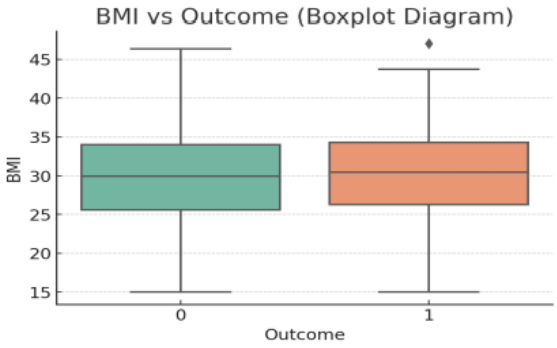

Fig 3: (b) BMI vs Outcome

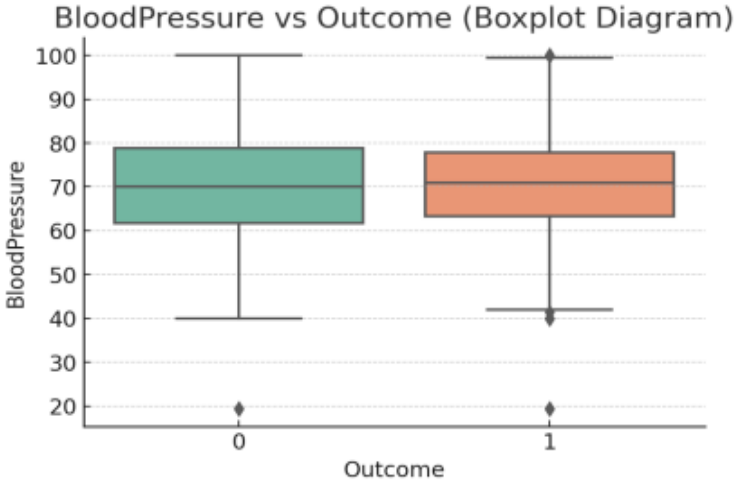

Fig 3: (c) bloodpressure vs outcome

The boxplots illustrate the variation in key medical features between diabetic and non-diabetic groups.in fig 3(a-c)

In Fig 3(a)**,** Glucose values are noticeably higher among diabetic individuals, indicating that blood sugar level is a major factor in diabetes prediction. For above Fig 3(b) reveals that people with diabetes generally have a higher BMI, suggesting that excess body weight increases the risk of diabetes. In Fig 3(c) displays Blood Pressure comparisons, where only slight differences are observed, meaning it has less influence than glucose or BMI.

Model Evaluation:

Two machine learning models Logistic Regression and Random Forest were used to predict diabetes using medical data. The dataset was split into training and testing parts, and performance was measured using accuracy, precision, recall, and F1-score**.** Logistic Regression performed well in identifying diabetic cases but had more false positives. Random Forest showed higher accuracy and more balanced results.

3. RESULTS AND DISCUSSION

This study compared Logistic Regression and Random Forest models for predicting diabetes using Kaggle data. Logistic Regression achieved slightly higher accuracy (52%) than Random Forest (50%) and performed better at identifying diabetic patients. Random Forest showed balanced results but was less consistent. Glucose, BMI, and Age were found to be the most important factors influencing diabetes. Overall, Logistic Regression proved to be the best model for this dataset due to its higher accuracy and reliable performance.

Comparison of Model Performance:

Metric	Logistic Regression	Random Forest
Accuracy	0.5200	0.5050
Precision (Class 0)	0.5426	0.5287
Precision (Class 1)	0.5000	0.4867
Recall (Class 1)	0.5521	0.5729
F1-Score (Class 1)	0.5248	0.5263

4. CONCLUSION

In this study applied two model that is Logistic Regression and Random Forest models to predict diabetes using a dataset of 1,000 patient records obtained from Kaggle. After evaluating model performance, Logistic Regression emerged as the better model, achieving higher accuracy and recall compared to Random Forest.The analysis revealed that Glucose, BMI, and Age are the most significant factors influencing diabetes prediction. Overall, this research demonstrates how simple machine learning techniques can effectively support early diabetes detection and help healthcare professionals make data-driven decisions for better patient care.

5. REFERENCES

[1] Comparative study of different Machine Learning Algorithms used for Credit Card fraud detection, DY Pimpri-2024, National Conference on "Pinnacle Perspectives: Unifying Strategies and Technologies for Tomorrow" 2024, by Prashant Wadkar, Dr. Shivaji Mundhe (Best Paper Award) Abstract ISBN: 978-81-932707-2-4.

[2] Analysis Of Breast Cancer Dataset & Its Prediction Using Machine learning Yashomanthan, International Research Journal ISSN: 2347 - 8039 Vol. XI, Issue I, Feb 2021.

[3] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC, 978-1-5090-3243-3, 2017.

[4] Comparative Analysis of Different Machine Learning Algorithms Used in Breast Cancer Prediction, Education and Society (शिक्षण आशण समाज) ISSN: 2278-6864, (UGC Care Journal) Vol-47, Issue-1, No.10, January-March: 2023