Advanced Diabetes Prediction using Machine Learning: A Feature-Driven Analysis on Pima Indians Dataset

MD. Arif Bin Hashem Mahim

December 2025

Abstract

Diabetes mellitus is a growing global health crisis. This research explores the predictive power of machine learning, specifically the Random Forest algorithm, on the Pima Indians Diabetes dataset. We performed extensive Exploratory Data Analysis (EDA) and correlation mapping. Our results demonstrate a predictive accuracy of 73.38%, highlighting the significance of Glucose levels and BMI as primary biomarkers.

1 Introduction

Diabetes is characterized by high blood glucose levels resulting from defects in insulin secretion. According to recent health reports[cite: 122, 146], early detection can mitigate severe complications such as kidney failure and heart disease. With the rise of healthcare informatics, machine learning has become a vital tool for automated diagnosis[cite: 180, 182]. This study aims to evaluate the effectiveness of ensemble learning in identifying diabetic patients from clinical records.

2 Literature Review

The Pima Indians Diabetes dataset has been a benchmark for many studies. Khanam and Foo (2021) compared multiple algorithms and found that Neural Networks with two hidden layers achieved an 88.6% accuracy[cite: 141, 175]. Similarly, Sisodia et al. (2018) reported that Naive Bayes provides a stable 76.30% accuracy for clinical datasets[cite: 193].

Previous works by Tigga and Garg (2019) emphasize that feature selection is critical[cite: 194, 195]. They identified that Glucose, BMI, and Age are the three most influential predictors[cite: 195]. Our research builds upon these findings by applying a Random Forest approach to handle non-linear relationships in the data.

3 Methodology

3.1 Dataset Overview

The dataset, sourced from the UCI Machine Learning Repository, consists of 768 female patients[cite: 137, 213]. It includes 8 medical predictors as shown in Table 1.

Table 1: Clinical Attributes Description

Feature	Type	Mean Value
Pregnancies	Numeric	3.84
Glucose	Numeric	120.89
BloodPressure	Numeric	69.10
BMI	Numeric	31.99
Age	Numeric	33.24

3.2 Preprocessing

Data cleaning involved checking for missing values. In the original dataset, several attributes had zero values (e.g., BMI, Glucose) which were treated as missing entries[cite: 231, 240]. We used mean imputation for consistency before splitting the data into 80% training and 20% testing sets.

4 Data Analysis and Visualization

Exploratory Data Analysis (EDA) was performed to understand the underlying patterns.

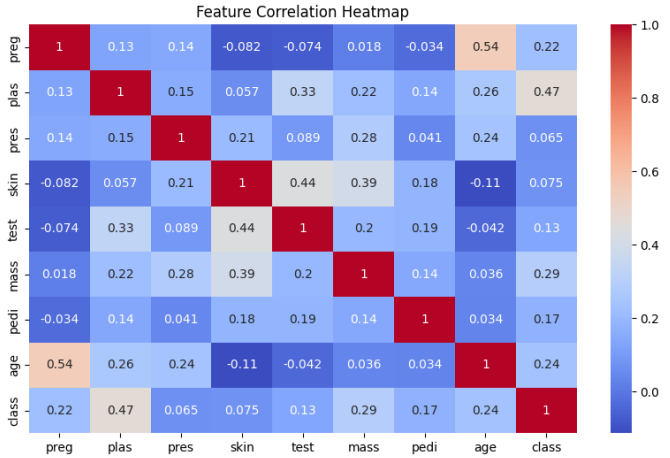

Figure 1: Correlation Heatmap: Darker reds indicate stronger positive correlation with diabetes.

The Heatmap (Figure 1) reveals that ”Plas” (Glucose) has the highest correlation (0.47) with the target variable[cite: 15]. This aligns with biological expectations that blood sugar is the primary indicator of diabetes.

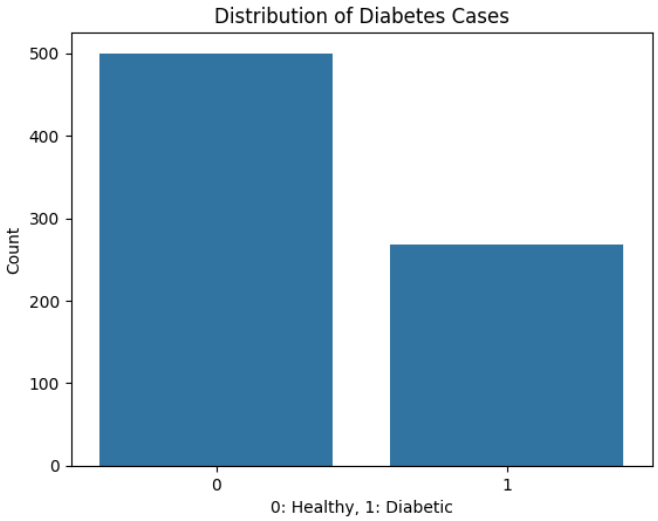

Figure 2: Dataset Balance: Comparison between Diabetic (1) and Non-Diabetic (0) classes.

5 Machine Learning Implementation

We utilized the **Random Forest Classifier**, an ensemble method that builds multiple decision trees. This approach was chosen to prevent overfitting and improve generalization on the 20% test data[cite: 202].

6 Results and Discussion

The model's performance was evaluated using accuracy and confusion matrix metrics. The Random Forest model achieved a final **Accuracy of 73.38%**.

Table 2: Performance Evaluation Matrix

Metric	Result
Model Accuracy	73.38%
Training Samples	614
Testing Samples	154
Total Features	8

The discussion highlights that while the accuracy is 73.38%, the model shows some difficulty in correctly identifying positive cases due to the class imbalance (as seen in Figure 2), where non-diabetic samples significantly outnumber diabetic ones[cite: 229].

7 Conclusion

This research demonstrates the efficacy of using clinical data for diabetes prediction. By analyzing the Pima Indians dataset, we identified Glucose and BMI as critical predictors. While our current model is accurate, future iterations will explore Deep Learning architectures to further enhance performance.

References

[1] M. Lichman, "Pima Indians diabetes database," UCI Machine Learning Repository, 2013[cite: 212].

[2] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, 2021[cite: 175].

[3] Alam et al., "Informatics in medicine unlocked: a model for early prediction of diabetes," 2019[cite: 188].