Diabetes Classification Analysis

Ammar Y. Mohamed

January 21, 2025

Introduction

The dataset used in this analysis originates from the National Institute of Diabetes and Digestive and Kidney Diseases. It aims to predict the likelihood of diabetes in patients based on specific diagnostic parameters. The dataset focuses on Pima Indian women aged 21 years or older, with each patient's medical history represented by several predictor variables, including age, BMI, insulin levels, and number of pregnancies. The target variable, `Outcome`, indicates whether a patient has diabetes (1) or not (0).

The primary objective of this analysis is to develop a predictive model that can accurately classify patients as diabetic or non-diabetic based on the provided features.

Methodology

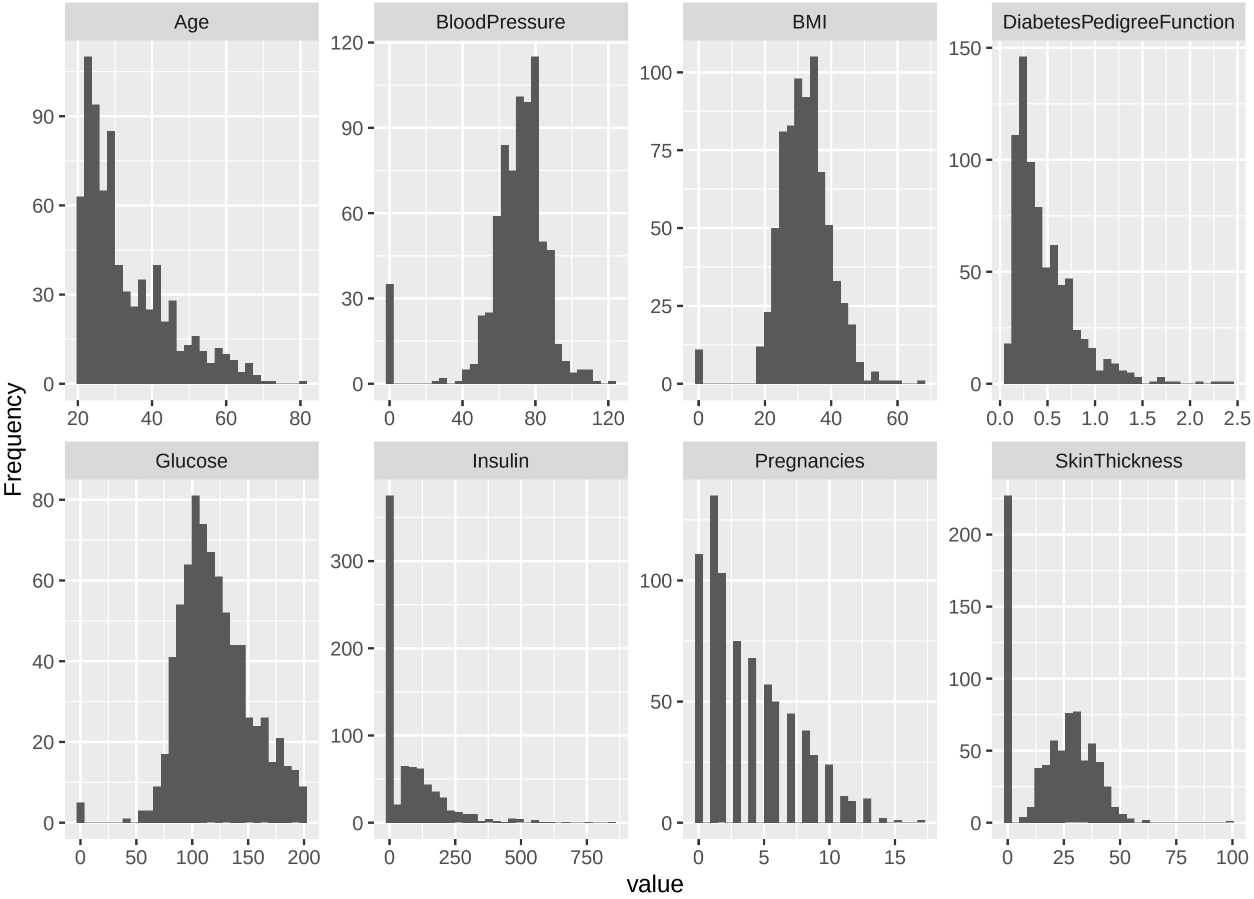
Data Source

The dataset was downloaded using the **Kaggle CLI** from this source.

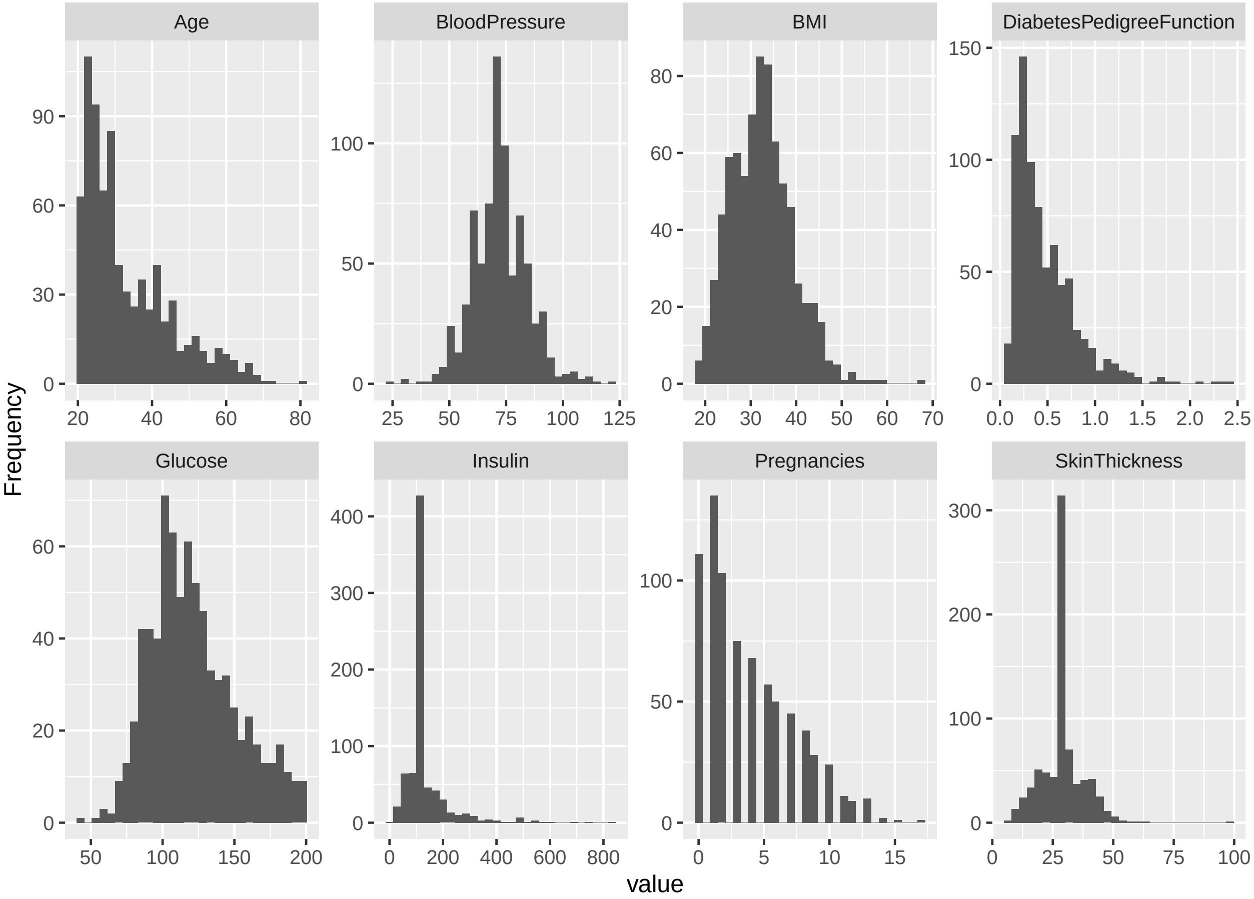
Data Cleaning and Preparation

Several variables in the dataset, such as `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI`, contained meaningless zero values, which were treated as missing data. These values were imputed with the median of their respective columns to ensure data integrity.

Before Imputation

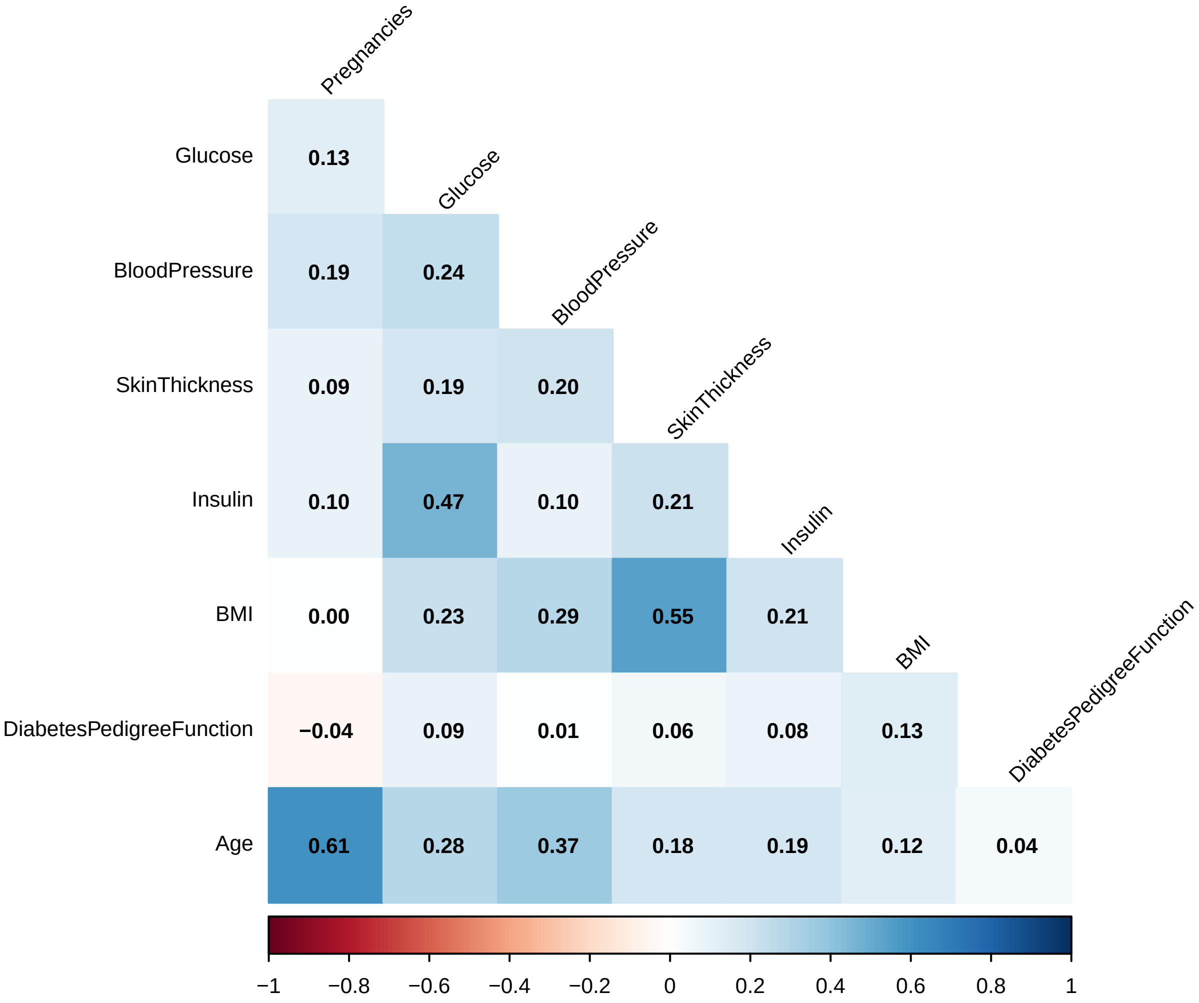


After Imputation



Correlation Analysis

A correlation analysis was conducted using Spearman's method to identify relationships between the predictor variables. The correlation plot reveals no strong correlations among the variables, indicating that multicollinearity is not a significant concern in this dataset.



Training Machine Learning Algorithms

Five models were trained and evaluated: Logistic Regression (with and without PCA preprocessing), Random Forest, XGBoost, and an Ensemble model combining predictions from all four models. The dataset was split into an 80% training set and a 20% test set to evaluate model performance.

Results

The performance of the models was evaluated using **Accuracy** and **F1 Score**, which measure overall correctness and the balance between precision and recall, respectively. The results are summarized in the table below:

Key Findings: The **Ensemble model** achieved the highest performance, with an **Accuracy of 82.4%** and an **F1 Score of 0.738**, demonstrating the effectiveness of combining predictions from multiple models. **XGBoost** also performed well, achieving an **Accuracy of 81.7%** and an **F1 Score of 0.720**, highlighting the strength of gradient boosting algorithms. The base **Logistic Regression** model achieved an **Accuracy of 80.4%** and an **F1 Score of 0.681**, indicating solid performance despite its simplicity. The **Random Forest** model performed comparably, with an **Accuracy of 79.7%** and an **F1 Score of 0.687**. The **Logistic Regression model with PCA preprocessing** achieved similar results to the base Logistic Regression model, with an **Accuracy of 79.7%** and an **F1 Score of 0.680**, suggesting that PCA did not significantly improve performance in this case.

Model Evaluation Results

	Performance Metrics	
Model	Accuracy	F1 Score
Logistic Regression	0.804	0.681
Random Forest	0.797	0.687
XGBoost	0.817	0.720
Updated Logistic Regression	0.797	0.680
Ensemble	0.824	0.738

Conclusion

This analysis demonstrates the effectiveness of ensemble methods in improving predictive performance for diabetes classification. The Ensemble model outperformed all individual models, achieving the highest accuracy and F1 score. XGBoost also showed strong performance, making it a viable alternative for this task. While Logistic Regression and Random Forest provided reasonable results, their performance was slightly lower than that of the Ensemble and XGBoost models. The use of PCA preprocessing did not yield significant improvements, indicating that feature engineering may require further exploration. Overall, this study highlights the importance of model selection and the potential benefits of combining diverse algorithms for enhanced predictive accuracy.

GitHub: Ammarymo