Diabetes Prediction using Machine Learning

Maitreya Sameer Ganu
IISER Thiruvananthapuram

29 November 2024

Abstract

Diabetes is a chronic condition that poses significant challenges worldwide, emphasizing the need for accurate and early detection. This project employs machine learning techniques to predict diabetes using the Pima Indians Diabetes Database. Through exploratory data analysis, feature scaling, and the application of the K-Nearest Neighbors (KNN) algorithm, the model achieved an accuracy of 80.5% on the test dataset. The study demonstrates the potential of data-driven approaches in healthcare for effective disease prediction and highlights areas for further research.

Introduction

Diabetes is one of the most prevalent metabolic disorders, affecting millions globally. Early detection and intervention are crucial for managing the disease and preventing complications. Machine learning offers a promising solution by enabling data-driven insights from medical datasets. This project focuses on predicting diabetes using a widely studied dataset: the Pima Indians Diabetes Database. The dataset includes features such as glucose levels, BMI, age, and insulin concentration, which are critical for assessing diabetes risk.

The objective of this study is to preprocess and analyze the dataset, identify meaningful patterns, and build a predictive model using the K-Nearest Neighbors (KNN) algorithm. The approach involves feature scaling, hyperparameter tuning, and rigorous evaluation to ensure reliability and accuracy.

Details of the Project

- **Dataset**: Pima Indians Diabetes Database.
- **Preprocessing**: Visualization of feature distributions, detection of outliers, and standard scaling to normalize the data.
- **Algorithm**: K-Nearest Neighbors (KNN) with hyperparameter tuning for selecting the optimal value of k.
- **Performance**: Achieved a test accuracy of 80.5% with $k = 5$.
- **Evaluation Metrics**: Confusion matrix and classification report for precision, recall, and F1-score.

Visualizations

1. Pairwise plot of all features

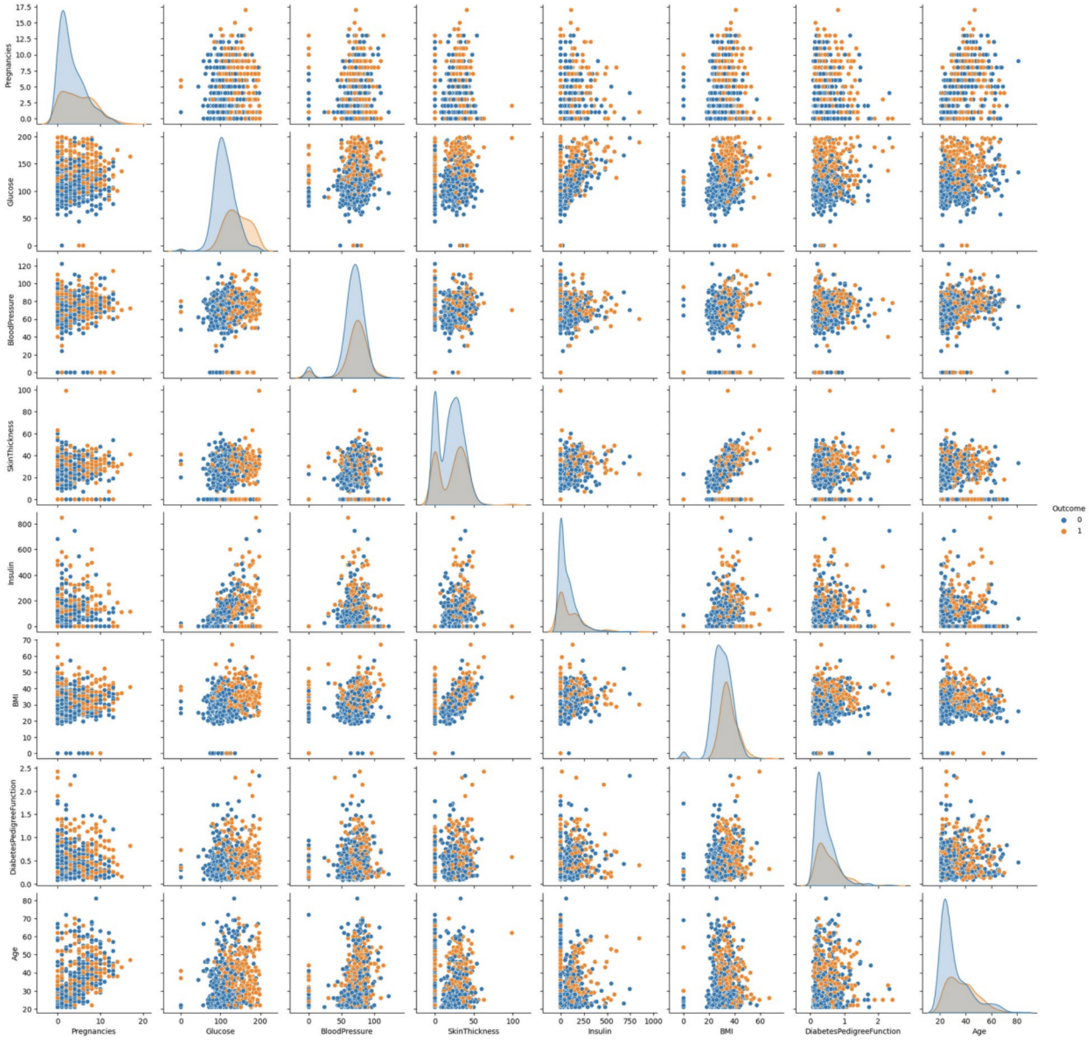

Figure 1: Pairwise distribution of features in the dataset.

2 . Outlier Detection

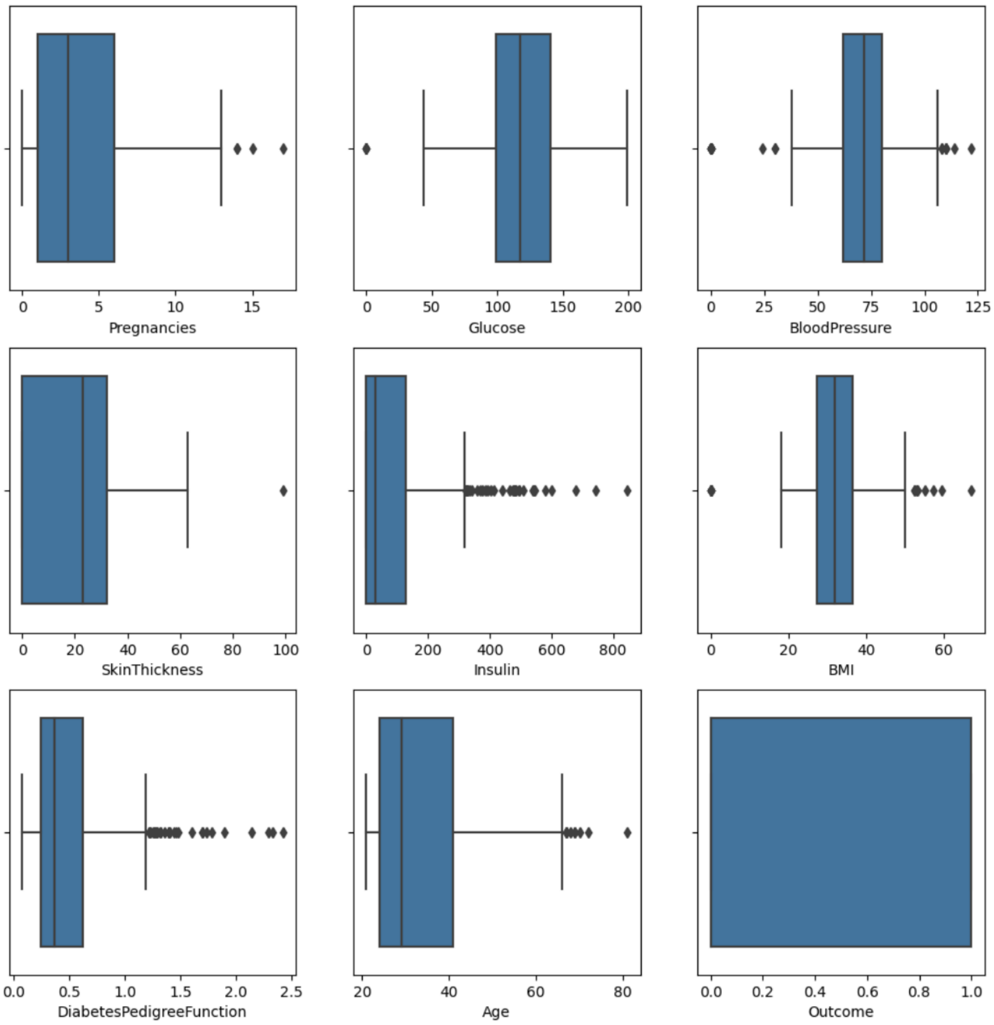

Figure 2: Boxplot for Outlier Analysis.

3 . Train vs Test Scores

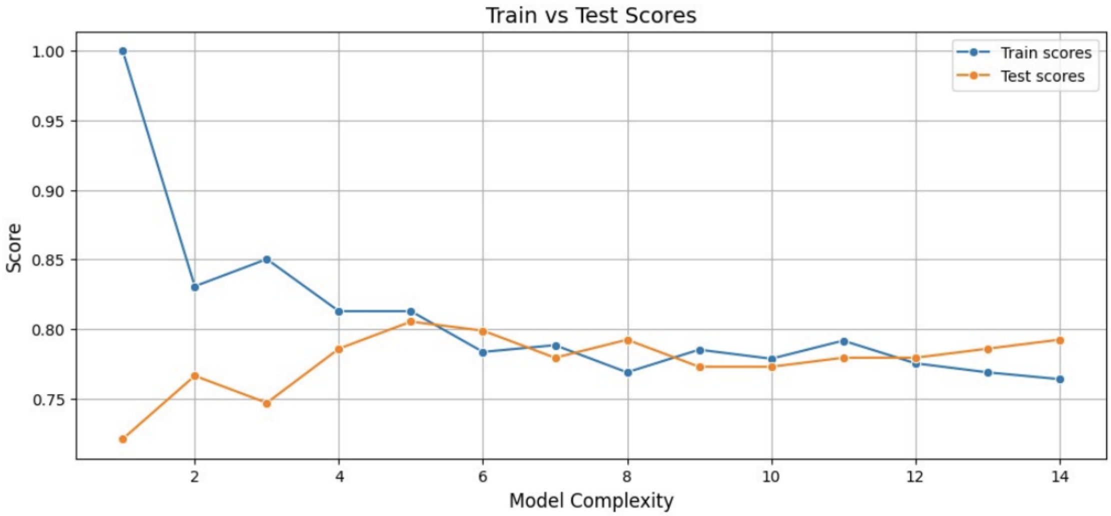

Figure 3: Train vs Test Scores for Different k Values.

4 . Histogram plot for number of entries in each feature

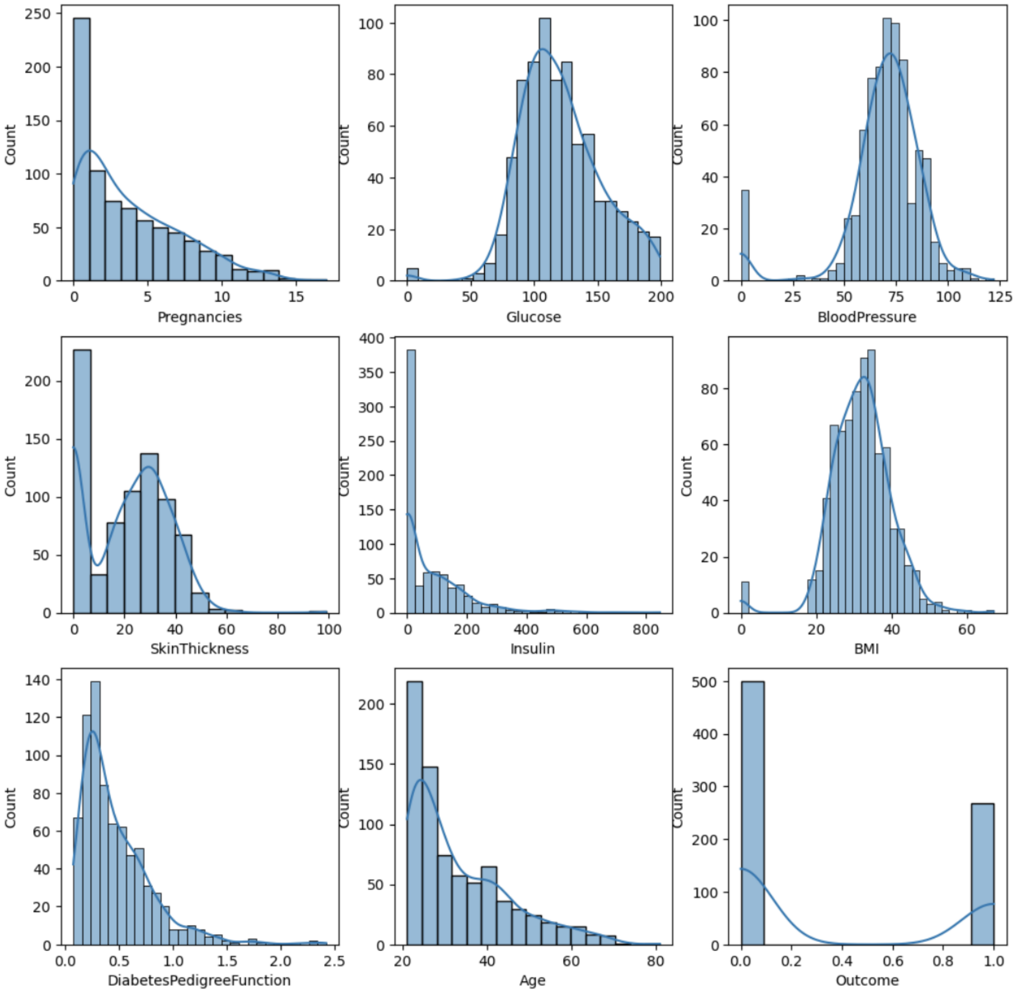

Figure 4: This diagram shows the Kernel Density Estimation (KDE) plots of all features

5 .Correlation heat map

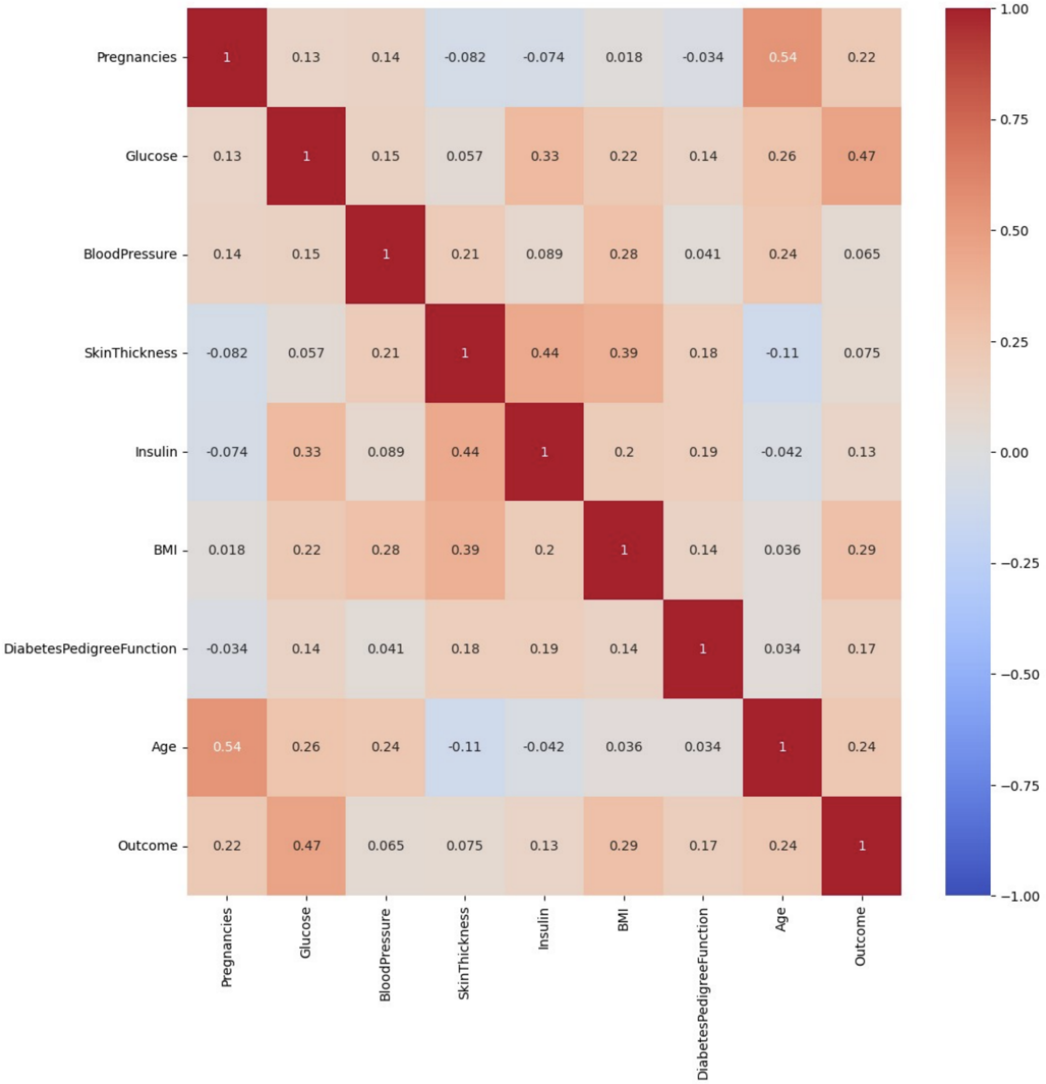

Figure 5: Heatmap depicting the correlation coefficients among features in the dataset. Strong positive correlations (closer to 1) are shown in red, while strong negative correlations (closer to -1) are shown in blue. This visualization highlights relationships between features, aiding in the identification of potential multicollinearity and feature importance.

Conclusion

This study demonstrates the effective application of machine learning for diabetes prediction. The K-Nearest Neighbors algorithm, combined with pre-processing techniques, achieved a test accuracy of 80.5%, showcasing its potential for healthcare applications. While the results are promising, there is scope for improvement through the use of advanced models such as ensemble methods or deep learning. Future work can also explore integrating additional datasets to enhance prediction robustness and enable personalized healthcare solutions.In this diabetes prediction project, the Pareto Principle was applied to split the dataset for training and testing purposes. Following the rule, approximately 20% of the dataset was reserved for testing the model, while 80% was used for training. This approach aligns with the principle by focusing the majority of resources (training data) on creating a robust model, while a smaller portion (test data) evaluates its performance.