Predicting Diabetes Risk: A Comparative Analysis of ML Models

Michael Zefanya

2025

About This Project

Diabetes is a global health concern, and early detection is essential for effective management and prevention. This project analyzes a medical dataset to identify patterns and relationships between health metrics and diabetes outcomes using machine learning (ML) techniques.

The objective is to develop predictive models that classify whether a patient is at risk of diabetes. Various ML algorithms are tested and compared to determine the most effective classification method.

The dataset, sourced from Kaggle, includes key medical indicators such as glucose levels, blood pressure, BMI, and insulin levels. By leveraging different ML models, this study provides valuable insights into feature importance and model performance in predicting diabetes risk.

Dataset and Preprocessing

Dataset Overview

- Sourced from Kaggle, containing medical records related to diabetes.
- Features 8 key variables:
  - Glucose, Blood Pressure, BMI, Insulin, Age, Pregnancies, Skin Thickness, Diabetes Pedigree Function
- Target Variable: Diabetes Outcome (1 = Diabetic, 0 = Non-Diabetic).

Preprocessing Steps

Handling Missing Values

- Replaced missing/zero values in Insulin and Skin Thickness with median values.
- Dropped rows with excessive missing data if necessary.

Feature Scaling

- Used Min-Max Scaling to normalize Glucose, Insulin, BMI, and Blood Pressure for better model performance.

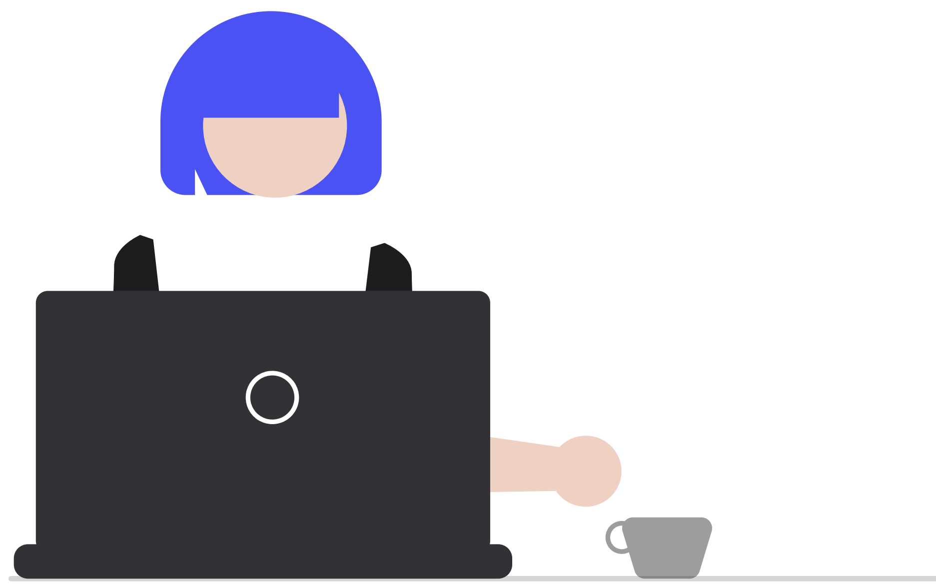
Outlier Detection

- Identified extreme values in BMI & Insulin using box plots.
- Applied Winsorization to limit outliers and improve data consistency.

Final Processed Dataset

- Cleaned, normalized, and ready for machine learning model training.

Ensuring high-quality data is crucial for accurate predictions



Tools I Used?

scikit learn

Pandas

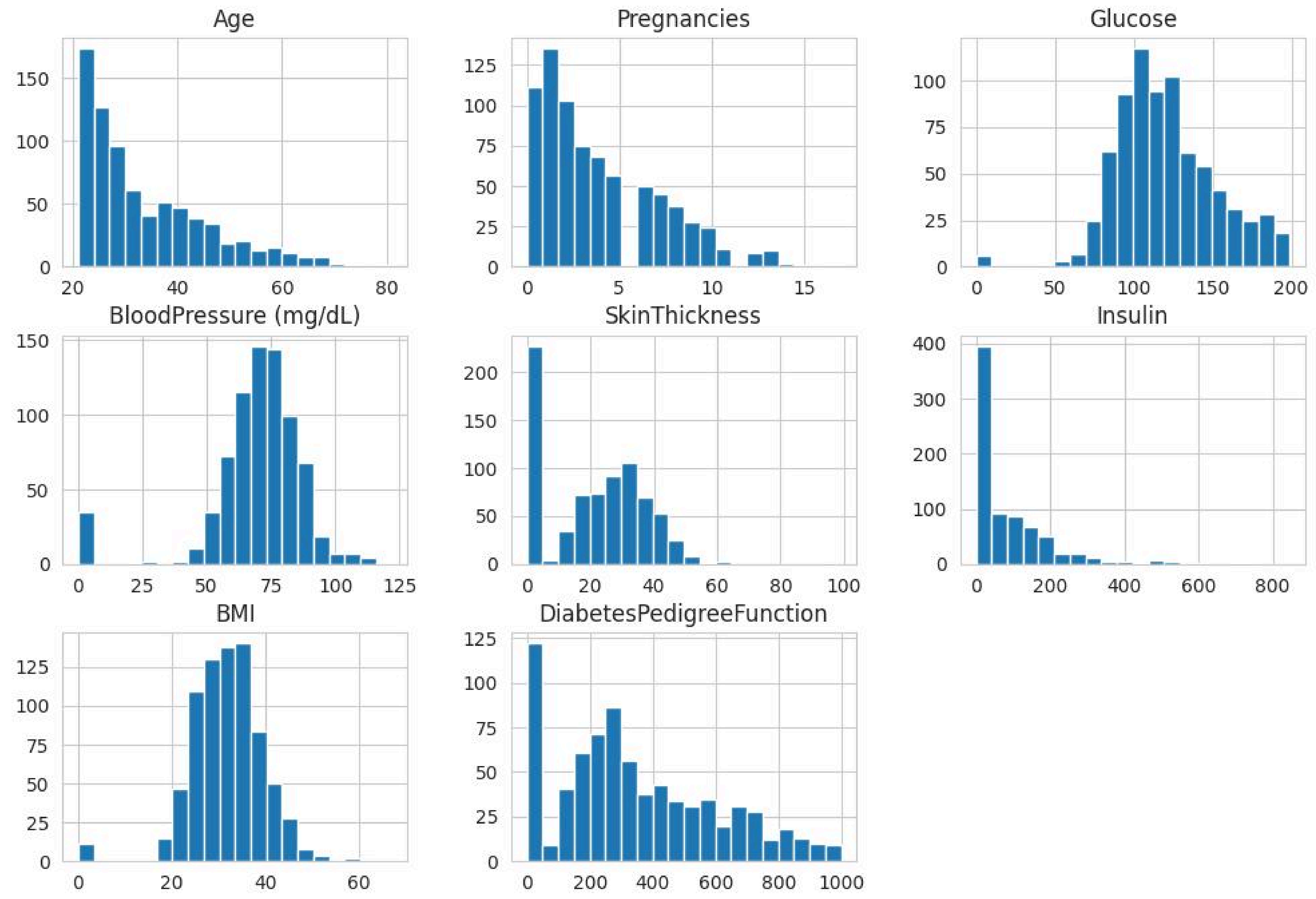
matplotlib

k

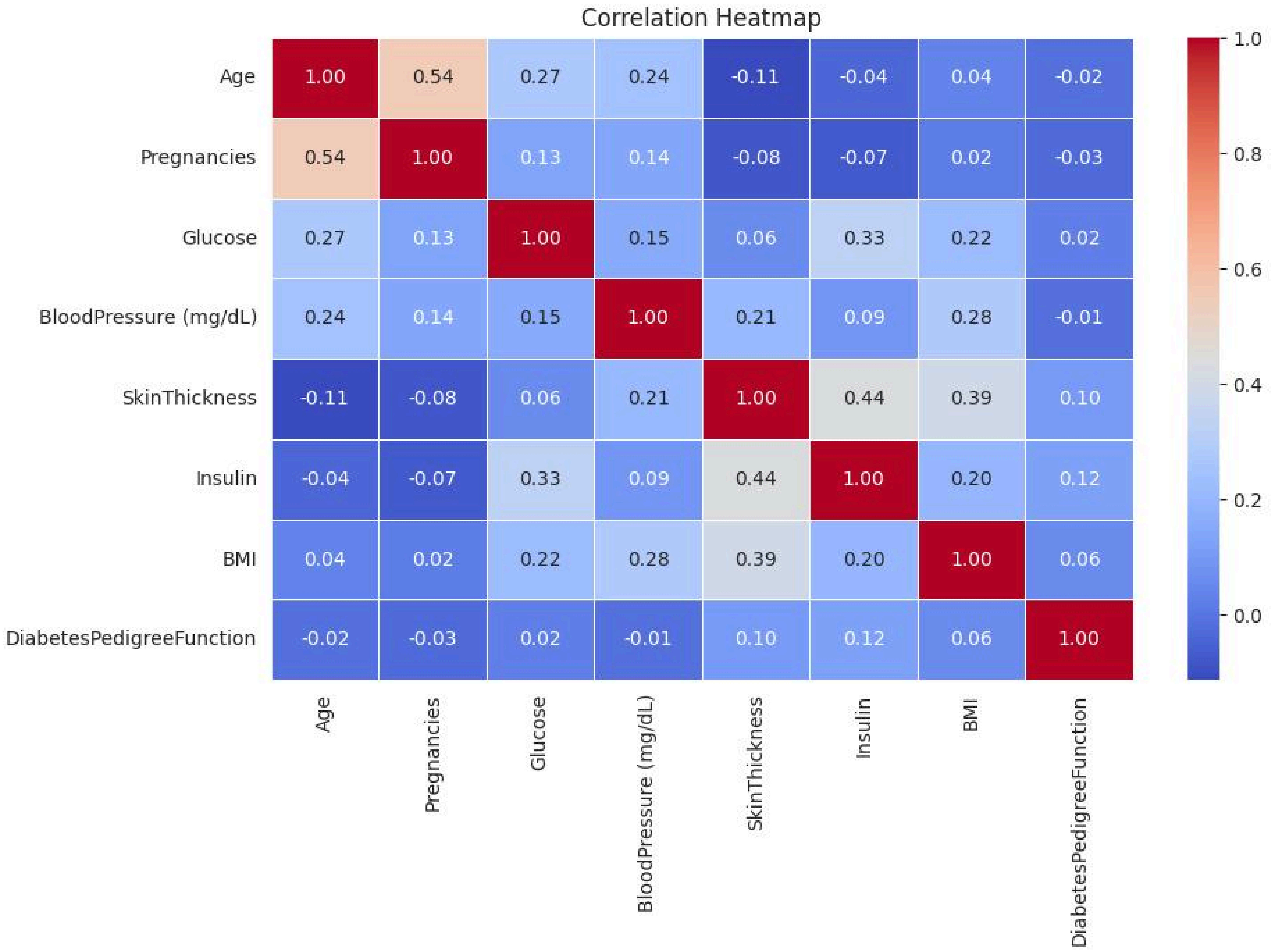
CO

seaborn

Exploratory Data Analysis (EDA)



Correlation Analysis



The correlation heatmap highlights that glucose has the strongest positive correlation with diabetes, followed by BMI and age, while insulin and skin thickness show a moderate relationship, and blood pressure and diabetes pedigree function have weaker correlations.

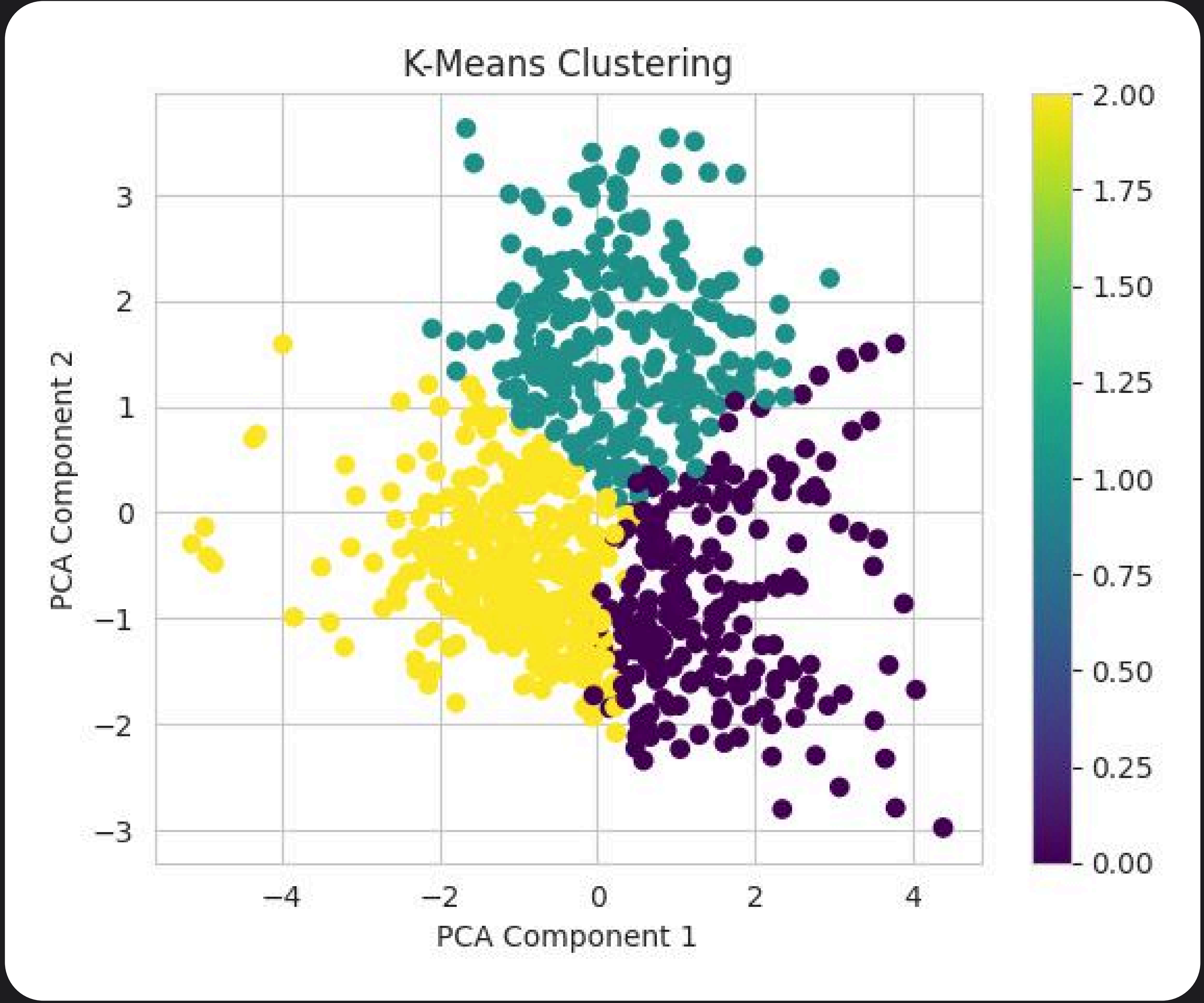
Clustering Results Using K-Means

K-Means clustering segmented individuals based on health characteristics, with the optimal clusters determined by the elbow method, highlighting varying diabetes risk levels and potential subgroups for targeted medical interventions.

01. Individuals with high glucose levels and BMI, potentially at higher risk.

02. Individuals with moderate glucose and BMI levels, possibly borderline cases

03. Individuals with low glucose and BMI, likely lower diabetes risk.



Conlusion

This study demonstrates the effectiveness of machine learning in predicting diabetes risk, with glucose levels (0.47), BMI (0.29), and age (0.24) identified as key predictors. SVM achieved the highest accuracy (85.4%), followed by Logistic Regression (83.7%), while Naïve Bayes had the fastest training time (78.5% accuracy). KNN showed overfitting, with 92.1% training accuracy but only 79.6% test accuracy. Future improvements include hyperparameter tuning, feature selection, and integrating real-time data for better performance. Deploying a web-based diagnostic tool using the best model could enhance accessibility and practical application in healthcare.

Thanks for Your Attention!

Thank you for your time and attention. If you have any questions or feedback, I would be delighted to hear from you. As I continue to learn and grow, I truly value the insights and engagement you've provided, and I look forward to any further discussions

linkedin.com/in/michaelzefanya

michaelzefanya04@gmail.com

github.com/michaelzefanya