Diabetes Risk Prediction Framework

A Data-Driven Diagnostic Strategy using Machine Learning

Author: Gaurav Mali

Project Link: github.com/Gaurav-mali12/diabetes-risk-prediction

Date: February 2026

1. Executive Summary

This project analyzes physiological health markers to predict diabetic outcomes. The workflow encompasses Exploratory Data Analysis (EDA), Outlier Mitigation, and Algorithm Benchmarking to identify the most robust predictive model for clinical assistance. Our final model achieved a predictive benchmark of ~80%.

2. Problem Statement

Diabetes is a chronic condition requiring early detection. This analysis utilizes clinical measurements—such as Glucose levels, BMI, and Age—to build a classification system capable of identifying high-risk individuals.

3. Methodology & Data Refinement

A significant challenge was the presence of 'Biological Zeros' (e.g., Blood Pressure at 0). These were treated as missing values and mitigated to ensure data integrity. Furthermore, StandardScaler was applied to normalize features, ensuring balanced algorithmic weighting.

4. Model Performance Comparison

Algorithm	Accuracy (%)
Random Forest	79.75%
Logistic Regression	77.21%
K-Nearest Neighbors	74.68%
Decision Tree	70.89%

5. Conclusion

The integration of rigorous data cleaning and ensemble learning (Random Forest) provided the most reliable results. This framework is ready for clinical validation and demonstrates the power of diagnostic machine learning.