Diabetes Classification Using Sowutuom Clinic Dataset

1. Introduction

The Sowutuom Clinic Dataset contains 500 synthetic patient records collected from 10 clinics. The goal is to build a machine learning model that classifies patients as diabetic (1) or non-diabetic (0) using clinical features such as age, BMI, glucose level, blood group, and genotype.

2. Dataset Overview

• Rows: 500 • Columns: 9 • Source: Synthetic • Target Column: diabetic (0 = non-diabetic, 1 = diabetic) The dataset includes: clinic, age, height, weight, BMI, glucose level, blood group, genotype, and diabetes status. No missing values were found, and all data types are consistent.

3. Preprocessing Steps

1. Load dataset in Python (pandas) 2. Encode categorical variables using One-Hot Encoding: - clinic - blood_group - genotype 3. Split dataset into training and testing sets (80/20) 4. Prepare features (X) and label (y) for model training

4. Models Used

Two machine learning models were trained and compared: 1. Logistic Regression - Baseline model - Good interpretability using coefficients 2. Random Forest Classifier - Handles nonlinear relationships - Provides feature importance ranking

5. Evaluation Metrics

The following metrics were computed for both models: • Accuracy – overall correctness of predictions • Precision – how many predicted diabetics were actually diabetic • Recall – ability to detect actual diabetic patients • F1 Score – harmonic mean of precision and recall

6. Feature Importance

Features influencing predictions were examined using: • Logistic Regression coefficients • Random Forest feature_importances_ Strong predictors typically include: • Glucose Level • BMI • Age

7. Model Comparison

Model performance is compared using all evaluation metrics. Random Forest generally performs better, while Logistic Regression remains easier to interpret.

8. Conclusion

The Sowutuom Clinic Dataset can effectively support diabetes prediction. By training Logistic Regression and Random Forest models, evaluating them with standard metrics, and analyzing feature importance, we gain both predictive power and insights into clinical factors that contribute most to diabetes diagnosis.