Early Stage Diabetes Prediction

# Project Overview

This project applies various machine learning algorithms to predict early-stage diabetes using a dataset of diagnostic measurements. The goal is to use predictive analytics to identify at-risk individuals early and ensure timely medical intervention.

# Dataset

The dataset `diabetes\_data\_upload.csv` includes 520 instances with 17 attributes, such as Age, Gender, Polyuria, Polydipsia, and other symptoms related to diabetes, alongside a target variable indicating the presence of diabetes.

# Features

- Data preprocessing and exploration  
- Feature selection and engineering  
- Model training with Logistic Regression, Random Forest, and XGBoost  
- Evaluation using accuracy, precision, recall, F1-score, and ROC AUC

# Technologies Used

- Python 3  
- Pandas for data manipulation  
- NumPy for numerical operations  
- Matplotlib and Seaborn for visualization  
- Scikit-learn for model building and evaluation  
- XGBoost for improved accuracy

# Model Training and Evaluation

The project explores various models, with Random Forest showing the best performance on the test data. Models were evaluated based on their accuracy, precision, recall, F1-score, and ROC AUC score.

# Results

The Random Forest model achieved an accuracy of approximately 99%, demonstrating its effectiveness in predicting early-stage diabetes.