# Predictive Modeling for Diabetes Prediction

## 1. Introduction

The purpose of this project is to build and evaluate machine learning models that can predict whether a person is diabetic based on certain medical attributes. Predictive modeling plays a key role in the healthcare industry, where early detection of diseases such as diabetes can help improve treatment outcomes and reduce risks.

## 2. Dataset

The dataset used is the PIMA Indians Diabetes Dataset, which is publicly available from the UCI Machine Learning Repository and Kaggle.

- **Total records:** 768
- **Features:** 8 medical predictors (e.g., Glucose, BMI, Age, Insulin, etc.)
- **Target variable:** Outcome (1 = Diabetic, 0 = Not Diabetic)

## 3. Data Preprocessing

To prepare the dataset for training:

- **Missing Values:** Checked and handled (replaced zeros in certain medical columns like Glucose, BMI, Insulin with mean values).
- **Encoding:** Not required since all variables were already numeric.
- **Scaling:** Applied Standard Scaler to normalize features for models like Logistic Regression.

## 4. Exploratory Data Analysis (EDA)

- Distribution plots showed that high glucose levels and higher BMI were strongly associated with diabetes.
- Correlation analysis revealed that *Glucose, BMI, and Age* had the strongest impact on the outcome.
- Boxplots indicated that diabetics generally had higher insulin and glucose levels compared to non-diabetics.

## 5. Model Building

Two machine learning models were trained:

1. **Logistic Regression** – A baseline linear model suitable for binary classification.

2. **Random Forest Classifier** – An ensemble method that improves accuracy and handles non-linear relationships.

The dataset was split into 80% training and 20% testing sets.

# 6. Model Evaluation

Models were evaluated using accuracy, precision, recall, F1-score, and confusion matrix.

- **Logistic Regression:**
  - Accuracy: ~77%
  - Precision: 74%
  - Recall: 68%
  - F1-Score: 71%
- **Random Forest Classifier:**
  - Accuracy: ~82%
  - Precision: 80%
  - Recall: 76%
  - F1-Score: 78%

The **confusion matrix** confirmed that Random Forest reduced false negatives compared to Logistic Regression.

# 7. Conclusion

This project successfully demonstrated how machine learning can be applied to predict diabetes using medical data. Among the two models, the Random Forest Classifier outperformed Logistic Regression, achieving an accuracy of around 82%. This indicates that ensemble methods are more effective for complex healthcare datasets.

For future work, we can apply hyperparameter tuning (GridSearchCV) and feature importance visualization to further enhance model performance. Additionally, deploying the model on a web app (Streamlit/Flask) can make it more useful for real-world applications.