# Diabetes Prediction Report

## 1. Introduction

Diabetes is a chronic disease that affects millions of people worldwide. Early detection is crucial for prevention and treatment. In this project, we used the **Pima Indians Diabetes Dataset**, which contains medical and health-related attributes of female patients, to build machine learning models for predicting diabetes.

The dataset includes **768 samples and 8 features** such as glucose level, blood pressure, insulin, BMI, age, and pregnancy count. The target variable indicates whether the patient is diabetic (1) or non-diabetic (0).

---

## 2. Dataset Insights

- **Size**: 768 records, 8 features, 1 target variable.
- **Features**: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age.
- **Target**: Diabetes (1 = positive, 0 = negative).
- **Imbalance**: About 65% non-diabetic, 35% diabetic patients.
- **Data Cleaning**: Replaced zero values in certain columns (e.g., BloodPressure, Insulin, BMI) with median values to handle missing information.

---

## 3. Data Visualization Findings

- **Histogram plots** showed that glucose and BMI distributions were skewed, with higher glucose linked to diabetes.
- **Correlation heatmap** indicated that **glucose** and **BMI** had the strongest correlation with diabetes outcome.
- **Boxplots** showed higher median glucose and BMI values among diabetic patients compared to non-diabetic ones.

---

## 4. Model Training and Comparison

We trained three baseline models and later tuned them using **RandomizedSearchCV**.

### Baseline Results

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| KNN | 68.8% | 0.63 | 0.55 | 0.59 |

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Decision Tree | 74.7% | 0.70 | 0.67 | 0.68 |
| Random Forest | 72.7% | 0.69 | 0.63 | 0.65 |

### After Hyperparameter Tuning

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| KNN (tuned) | ~72% | Improved | Improved | Better overall balance |
| Decision Tree (tuned) | ~77% | Improved | Improved | More stable results |
| Random Forest (tuned) | ~79% | Best | Best | Best performance |

---

# 5. Feature Importance

Using the **Random Forest model**, the most important features were:

1. **Glucose** – strongest predictor of diabetes.
2. **BMI** – highly correlated with diabetes.
3. **Age** – older individuals had higher risk.
4. **Pregnancies** – women with more pregnancies had higher likelihood of diabetes.

---

# 6. Evaluation and ROC Curve

The **Random Forest (tuned)** model performed best overall, achieving the highest accuracy and F1-score.

- **Confusion Matrix** showed fewer false negatives compared to other models.
- **ROC Curve (AUC ~0.85)** confirmed good model performance.

---

# 7. Conclusion

- **Best Model**: Random Forest (after hyperparameter tuning).
- **Key Features**: Glucose, BMI, Age, and Pregnancies contributed most to predictions.
- **Impact of Tuning**: Hyperparameter tuning significantly improved model accuracy and stability compared to default settings.
- **Practical Use**: This system could be used in healthcare to assist early diabetes risk prediction and encourage lifestyle or medical interventions.