**Diabetes Prediction Early**

**Overview**

This report outlines the process of building and evaluating a machine learning model to predict whether a person has diabetes. The dataset used for this project is available on Kaggle: .

**Data Exploration**

1. \*\*Dataset Size and Structure:\*\* - The dataset contains 768 rows and 9 columns, each describing various health metrics of individuals, including their diabetes status .

2. \*\*Columns:\*\* - Pregnancies - Glucose - BloodPressure - SkinThickness - Insulin - BMI - DiabetesPedigreeFunction - Age - Outcome

3. \*\*Key Observations:\*\* - Basic statistics were computed for numeric columns. - Null values were detected in the columns 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', and 'BMI' after replacing zeros with NaN.

**Data Cleaning**

* Missing values were handled as follows:
  + **Glucose** and **BloodPressure**: Replaced NaN with mean values.
  + **SkinThickness**, **Insulin**, and **BMI**: Replaced NaN with median values based on their data distribution.
* Data distributions were visualized before and after imputing missing values to verify the consistency of the dataset.

**Exploratory Data Analysis (EDA)**

* Outcome distribution was visualized using a count plot.
* Histograms for all numeric features provided insights into their distributions before and after missing value imputation.

**Model Building**

**Train-Test Split**

* The dataset was split into training and testing sets with an 80:20 ratio.

**Feature Scaling**

* StandardScaler was used to scale all features, ensuring consistent ranges for model training.

**Model Selection**

GridSearchCV was used to identify the best model and hyperparameters. Models evaluated include:

* Logistic Regression
* Decision Tree
* Random Forest
* Support Vector Machine (SVM)

**Best Model**

* **Random Forest Classifier** was identified as the best model with the highest accuracy.
* Key hyperparameters tuned:
  + Number of estimators: [10, 15, 20, 50, 100, 200]

**Model Evaluation**

1. **Confusion Matrix:**
   * Visualized for both training and testing datasets to understand model performance.
2. **Accuracy Scores:**
   * **Training Set Accuracy:** 100%
   * **Test Set Accuracy:** 84%
3. **Classification Report:**
   * Precision, recall, and F1-score were calculated for each class (0: No Diabetes, 1: Has Diabetes).
4. **Cross-Validation:**
   * Average accuracy across 5-fold cross-validation was computed for Random Forest Classifier, confirming its reliability.

**Predictions**

A function was created to predict diabetes based on user inputs:

* **Inputs:** Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DPF, Age.
* Example Predictions:
  + **Input:** [2, 81, 72, 15, 76, 30.1, 0.547, 25] → "Oops! You have diabetes."
  + **Input:** [1, 117, 88, 24, 145, 34.5, 0.403, 40] → "Great! You don't have diabetes."
  + **Input:** [5, 120, 92, 10, 81, 26.1, 0.551, 67] → "Great! You don't have diabetes."

**Conclusion**

The Random Forest Classifier was determined to be the best-performing model for predicting diabetes, achieving high accuracy on the test set. The model is reliable, and its predictions can be trusted for this dataset.

**Future Enhancements**

* Perform feature engineering to improve model performance further.
* Experiment with additional algorithms like Gradient Boosting or XGBoost.
* Collect more data to reduce potential biases and improve generalizability.

