Project Title: Diabetes Patient Data Analysis and Prediction

Aim:

To analyze the dataset of diabetes patients and build a machine learning model to predict whether a patient is diabetic or not based on various health-related features.

Objective:

1. Perform **Exploratory Data Analysis (EDA)** to uncover patterns, correlations, and insights from the dataset.
2. Prepare the data for machine learning by handling missing values, scaling, and splitting into training and test sets.
3. Build and evaluate a classification model to predict the diabetes outcome based on health-related features.
4. Use **metrics** like accuracy, precision, recall, and F1-score to evaluate the model's performance.
5. Generate visualizations to help better understand feature distributions and model predictions.

Dataset Overview:

This dataset consists of 9 columns that provide health and demographic information about patients. Here's a breakdown of each column:

1. **Pregnancies**:
    - Represents the number of times the patient has been pregnant.
    - Type: Integer
2. **Glucose**:
    - Plasma glucose concentration (measured after 2 hours in an oral glucose tolerance test).
    - Type: Float
    - Higher glucose levels may indicate poor insulin control.
3. **BloodPressure**:
    - Diastolic blood pressure (mm Hg).
    - Type: Float
    - Tracks heart health and blood circulation.
4. **SkinThickness**:
    - Triceps skinfold thickness (mm).
    - Type: Float
    - Acts as an indirect measure of body fat.
5. **Insulin**:
    - 2-hour serum insulin (mu U/ml).
    - Type: Float

- Measures insulin function and glucose metabolism.

6. **BMI**:
   - Body mass index (weight in kg/(height in m)^2).
   - Type: Float
   - Used to measure body fat and overall health.
7. **DiabetesPedigreeFunction**:
   - A function that assesses the likelihood of diabetes based on family history.
   - Type: Float
   - This is a probabilistic value derived from genetic factors.
8. **Age**:
   - Age of the patient (years).
   - Type: Integer
   - Age can be a significant factor in diabetes onset.
9. **Outcome**:
   - Target variable indicating whether the patient has diabetes (1) or not (0).
   - Type: Integer (Binary Classification)

EDA Steps:

1. **Data Cleaning**:
   - Check for **missing values** and handle them appropriately (either by imputation or removal).
   - Detect and deal with **outliers** using boxplots or statistical methods.
2. **Univariate Analysis**:
   - Visualize the distribution of individual features (e.g., histograms for continuous variables like glucose, BMI, etc.).
   - Use bar plots for categorical variables like **Pregnancies** and **Outcome**.
3. **Bivariate Analysis**:
   - Investigate relationships between variables, e.g., glucose vs. Outcome using scatter plots, and correlation heatmaps.
   - Use box plots to compare distributions of features across the **Outcome** categories (diabetic vs non-diabetic).
4. **Multivariate Analysis**:
   - Use pair plots and correlation matrices to explore the interactions between multiple features and identify potential multicollinearity.
   - Investigate how age, pregnancies, and BMI jointly affect the likelihood of diabetes.

Machine Learning Approach:

1. **Data Preprocessing**:
    - **Feature scaling** using techniques like standardization (Z-score scaling) for features like glucose, insulin, and BMI.
    - Split the data into **training** and **testing** sets (e.g., 80% train, 20% test).
2. **Model Selection**:
    - Implement classification algorithms like:
        - Logistic Regression
        - Decision Trees
        - Random Forests
        - Support Vector Machines (SVM)
        - Gradient Boosting
3. **Model Evaluation**:
    - Use cross-validation to ensure the model generalizes well.
    - Calculate evaluation metrics:
        - **Accuracy**: Overall correctness of the model.
        - **Precision**: Proportion of positive identifications that are actually correct.
        - **Recall**: The ability to find all the positive samples.
        - **F1-Score**: The harmonic mean of precision and recall.
4. **Model Interpretation**:
    - Feature importance analysis using methods like:
        - Coefficients from logistic regression.
        - Feature importance from tree-based models.
5. **Hyperparameter Tuning**:
    - Use techniques like GridSearchCV or RandomizedSearchCV to optimize model performance.

Key Deliverables:

1. **Visualizations**:
    - Feature distributions, correlation heatmaps, and prediction results.
2. **Classification Model**:
    - A well-tuned machine learning model to predict diabetes.
3. **Insights & Recommendations**:
    - Detailed report on which factors are most important for diabetes prediction.
4. **Performance Metrics**:
    - A table comparing the performance of different models (accuracy, precision, recall, etc.).