Title of Project: Predicting whether a patient is diabetic or not using supervised machine learning algorithms.

Objective:

The objective of this project is to apply supervised machine learning techniques to predict whether a patient is diabetic based on specific health measurements and personal information.

Problem Statement:

You are provided with a dataset containing various health-related features of individuals. Your task is to build a predictive model that can accurately classify whether an individual is diabetic or not.

Dataset:

You can download the dataset from the following link: (URL)

The dataset consists of the following features:

1. **Pregnancies:** Number of times the patient has been pregnant.
2. **Glucose:** Plasma glucose concentration after 2 hours in an oral glucose tolerance test.
3. **BloodPressure:** Diastolic blood pressure (mm Hg).
4. **SkinThickness:** Triceps skin fold thickness (mm).
5. **Insulin:** 2-Hour serum insulin (mu U/ml).

6. **BMI:** Body mass index (weight in kg/(height in m)^2).
7. **DiabetesPedigreeFunction:** A function which scores the likelihood of diabetes based on family history.
8. **Age:** Age of the patient (years).
9. **Outcome:** Class variable (0 or 1) indicating whether the patient is diabetic (1) or not (0).

1. **Understanding the Problem Statement:**
    - Discuss what each feature represents and its potential impact on diabetes prediction.

      Example - Glucose: Glucose refers to the amount of sugar in the blood, measured two hours after drinking a sugary liquid. It is measured in milligrams per decilitre (mg/dL). This test is called an Oral Glucose Tolerance Test (OGTT) and shows how well a person's body processes sugar.

      Potential Impact on Diabetes Prediction: The glucose level is a very important indicator for predicting diabetes. Here's why:

        - **Direct Connection with Diabetes:** High blood sugar levels are a clear sign of hyperglycemia, which is often linked to diabetes. When the body doesn't make enough insulin or can't use it properly, sugar stays in the blood instead of being used for energy.
    - Understand the objective of classifying patients as diabetic or non-diabetic.
2. **Observing the Dataset:**
    - Load the dataset and display the first few rows. (pd.read_csv("file_name"))
    - Check the dataset for missing values and outliers.
    - Use statistical measures (mean, median, standard deviation) to understand the distribution of features.
3. **Data Pre-processing:**
    - Handle missing values (if any) using appropriate techniques such as mean imputation, median imputation, or removal. (For example, code for mean imputation for Glucose column: df['Glucose'].fillna(df['Glucose'].mean(), inplace=True))
    - Normalize or standardize the data if necessary to ensure all features contribute equally to the model.
    - Perform feature engineering if necessary to create new features that might improve the model's performance.
4. **Model Building:**
    - Select an appropriate supervised learning algorithm (e.g., Logistic Regression, Decision Tree, Random Forest, Support Vector Machine).

- Train the model using the training data.
- Explain in 6-8 sentences why you chose the above model and how it is performing.

5. **Model Evaluation:**
    - Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.
    - Use a confusion matrix to understand the performance in detail.
6. **Conclusion:**
    - Summarize the findings, including the final model's performance and any insights gained from the data.
7. **Documentation:**
    - Document all the steps, observations, and decisions made during the project.
    - Prepare a report detailing the methodology, results, and any challenges faced during the project.