1. Decision Tree

## **Justification for Algorithm Choice:** We have chosen the Decision Tree algorithm based on the following reasons –

## **Interpretability:** Decision Trees are highly interpretable compared to many other machine learning models. The flowchart-like structure allows stakeholders (e.g., healthcare providers) to understand how decisions are made based on specific features. This transparency is crucial in healthcare, where understanding the rationale behind predictions can inform clinical decisions.

## **Handling Non-Linear Relationships:** The nature of the data in diabetes prediction often includes non-linear relationships between features and the target variable. Decision Trees can capture these relationships effectively without requiring extensive preprocessing or transformation of the data.

## **Feature Importance:** Decision Trees inherently provide a mechanism to assess feature importance, allowing us to identify which demographic and medical factors are most influential in predicting diabetes risk. This is particularly useful for informing further research and preventive strategies.

## B. **Model Training and Tuning**

## **1. Data Preprocessing**

## Before training the model, essential preprocessing steps were taken, including:

* *Missing values were checked and handled accordingly.*
* *Categorical features were converted to numerical format using ordinal encoding.*
* *The dataset was split into training and testing subsets to evaluate model performance accurately.*

2. **Model Training**:

* *The initial model was trained using the default parameters of the DecisionTreeClassifier.*
* *The model was fitted to the training data using dt\_classifier.fit()*

## **3. Feature Selection**:

## Features were assessed for their importance, which helped refine the feature set and remove irrelevant or redundant features, improving model accuracy and interpretability.

## C. **Effectiveness of the Algorithm**

## **1. Performance Metrics**

## Several metrics were utilized to evaluate the effectiveness of the logistic regression model:

## **Accuracy**: Measures the overall correctness of the model. Given the class distribution in the dataset, accuracy alone might be misleading.

## **Precision and Recall**: These metrics are crucial in healthcare applications where false negatives (missed diabetes cases) can have serious consequences. High precision indicates a low false positive rate, while high recall indicates a low false negative rate.

## **2. Model Performance Results**

## Upon evaluation, the logistic regression model achieved:

## **Accuracy**: Approximately 95%, indicates a high level of correct predictions overall.

## **Precision**:

## For Class 0 (non-diabetic): 0.97 (This means 97% of the instances predicted as non-diabetic were correct.)

## For Class 1 (diabetic): 0.72 (This means 72% of the instances predicted as diabetic were correct.)

## **Recall**:

## For Class 0 (non-diabetic): 0.97 (This means 97% of the actual non-diabetic cases were correctly identified.)

## For Class 1 (diabetic): 0.74 (This means 74% of the actual diabetic cases were correctly identified.)

## **F1-Score**:

## For Class 0 (non-diabetic): 0.97 (This indicates a very strong performance for predicting non-diabetic patients.)

## For Class 1 (diabetic): 0.73 (This shows moderate performance for predicting diabetic patients.)

## D. Insights Gained from the Algorithm

* Feature importance analysis revealed that certain factors (e.g., BMI, age, and gender) were significantly associated with diabetes risk.
* The decision tree visualization helped identify specific thresholds for features (e.g., BMI > 30) that are critical indicators of diabetes, which can guide public health interventions and individual risk assessments.
* The model's performance on different classes provided insights into the need for targeted interventions for populations at risk, especially older individuals and those with high BMI.