**Model Evaluation - Comparing Decision Tree and Logistic Regression**

**Objective**:  
In this exercise, you will compare the performance of both the Decision Tree and Logistic Regression models, evaluating which model is more suitable for predicting diabetes based on performance metrics.

**1. Data Dictionary**

**Dataset Link**: [Dataset](https://drive.google.com/file/d/1RRKsGCrRS65vInUn6MQYJJzEck9uYeF3/view?usp=drive_link)  
The dataset contains the following attributes:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Feature Name** |  | **Description** | **Data Type** | | --- | --- | --- | --- | | Number of times pregnant |  | Number of times the individual has been pregnant. | Integer | | Plasma glucose concentration |  | Plasma glucose concentration 2 hours after an oral glucose intake. | Integer | | Diastolic blood pressure |  | Diastolic blood pressure (measured in mmHg). | Integer | | Triceps skin fold thickness |  | Triceps skin fold thickness, a measure of body fat. | Integer | | 2-Hour serum insulin |  | Insulin levels 2 hours after glucose intake. | Integer | | Body mass index (BMI) |  | Weight in kg divided by height squared (m²). | Float | | Diabetes pedigree function |  | A function representing genetic factors related to diabetes. | Float | | Age (years) |  | Age of the individual. | Integer | | Outcome |  | Target variable indicating if the individual has diabetes (YES/NO). | Categorical | |

**2. Problem Statement**

The goal is to evaluate the performance of both Decision Tree and Logistic Regression models using the same dataset. you will compare the metrics to determine which model performs better.

**3. Model Evaluation**

**3.1 Model Comparison**

* Train both Decision Tree and Logistic Regression models on the same training dataset.

**3.2 Evaluation Metrics**

* Evaluate each model using:
  + **Accuracy**: The overall accuracy of the model.
  + **Precision**: The proportion of positive predictions that were correct.
  + **Recall**: The proportion of actual positives correctly identified.
  + **Confusion Matrix**: Analyze the true positives, false positives, true negatives, and false negatives for both models.

**3.3 Visualization**

* Plot ROC curves to compare the models’ performance.
* Visualize the confusion matrices to assess classification performance.

**4. Model Selection**

Based on the evaluation metrics, decide which model (Decision Tree or Logistic Regression) performs better for predicting diabetes. Discuss which model you would recommend for this problem and why.