**Decision Tree Prediction**

**Objective:**

In this exercise, we will use the Decision Tree Classifier to predict whether a person has diabetes based on various health-related features from a dataset.

**1. Data Dictionary**

**Dataset Link**: [Dataset](https://drive.google.com/file/d/1RRKsGCrRS65vInUn6MQYJJzEck9uYeF3/view?usp=drive_link)

The dataset includes the following attributes:

| **Feature Name** | **Description** | **Data Type** |
| --- | --- | --- |
| Number of times pregnant | Number of pregnancies the individual has had. | Integer |
| Plasma glucose concentration | Plasma glucose concentration 2 hours after an oral glucose intake. | Integer |
| Diastolic blood pressure | Diastolic blood pressure (mmHg). | Integer |
| Triceps skin fold thickness | Measurement of body fat from the triceps skin fold. | Integer |
| 2-Hour serum insulin | Insulin levels measured 2 hours after glucose intake. | Integer |
| Body mass index (BMI) | Body mass index (weight in kg / height in m²). | Float |
| Diabetes pedigree function | A measure of genetic factors related to diabetes. | Float |
| Age (years) | Age of the individual. | Integer |
| Outcome | Whether the individual has diabetes (YES/NO). | Categorical |

**2. Problem Statement**

The goal is to build a Decision Tree model to predict whether an individual has diabetes based on the provided features. The challenge is to create an accurate classifier using these features.

**3. Data Preprocessing**

**3.1 Data Cleaning and Feature Engineering**

* **Handling Missing Values:** Check for and handle any missing data, either by removing or imputing values where appropriate.
* **Encoding Categorical Variables:** Convert the "Outcome" variable (YES/NO) to binary format (1 for YES, 0 for NO).

**3.2 Outlier Detection and Treatment**

* **Identifying Outliers:** Detect outliers in features like plasma glucose and BMI using methods like box plots or Z-scores.
* **Outlier Handling:** Treat outliers by either removing them or applying capping techniques, depending on the severity.

**4. Exploratory Data Analysis (EDA)**

**4.1 Summary Statistics**

* Calculate basic statistics (mean, median, standard deviation, etc.) for all numerical features to understand the distribution of the data.

**4.2 Univariate Analysis**

* **Numerical Features:** Visualize distributions using histograms and box plots to identify the spread of the data.
* **Categorical Features:** Display the distribution of the outcome variable (diabetes) using bar charts.

**4.3 Bivariate Analysis**

* **Relationships Between Features:** Use scatter plots to visualize relationships between numerical features (e.g., age vs. BMI).
* **Outcome vs. Features:** Use box plots to examine how numerical features vary between individuals with and without diabetes.
* **Correlation Analysis:** Create a heatmap of the correlation matrix to identify any significant relationships between features.

**5. Model Building**

**5.1 Model Training**

* **Data Splitting:** Split the dataset into training and testing sets (commonly 70% for training and 30% for testing).
* **Training the Model:** Train a Decision Tree Classifier on the training data.