**Logistic Regression**

**Objective**:  
This exercise allows students to apply Logistic Regression to the same diabetes prediction dataset.

The goal is to Build the Logistic Regression model.

**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 task is to build a Logistic Regression model to predict whether an individual has diabetes based on the provided features.

**3. Data Pre-Processing**

Follow the same steps as in Decision Tree Exercise, ensuring the data is clean and outliers are handled.

Decision Tree Exercise Data Preprocessing steps

**(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)**

Carry out the same exploratory analysis as in Decision Tree Exercise to understand the data and relationships between features.

**5. Model Building**

**5.1 Model Training**

* Split the dataset into training and testing sets.
* Use Logistic Regression to train the model on the training data.
* Predict the train model on test data