# Practical No.5

**Title:** Write a program to implement the Logistic Regression for a sample training data set stored as a .CSV file. Compute the accuracy of the classifier

## Objective

The objective of this practical is to implement Logistic Regression using a dataset stored in a .CSV file. The model will be trained to classify data into two categories, and its performance will be evaluated using accuracy and other classification metrics.

## Introduction to Logistic Regression

Logistic Regression is a supervised learning algorithm used for binary classification problems. It predicts the probability that an instance belongs to a particular class using the sigmoid function, which maps any real-valued number into a range between 0 and 1.

The model makes predictions based on a decision boundary:

- If the probability is $\geq 0.5$, the instance is classified as 1 (positive class).
- If the probability is $< 0.5$, the instance is classified as 0 (negative class).

This method is widely used in medical diagnosis, fraud detection, and spam filtering due to its efficiency and interpretability.

## Dataset Description

The dataset used for this implementation is stored in a .CSV file. It consists of multiple independent variables (features) and a binary dependent variable (target variable). The features represent input data, while the target variable indicates the classification category.

The dataset should be structured appropriately, with each row representing an observation and each column corresponding to a feature or the target variable.

## Implementation Steps

Step 1: Load and Explore the Dataset

- Read the dataset from a .CSV file.
- Display basic statistics and check for missing values.

- Handle missing data by removing or imputing values.

Step 2: Data Preprocessing

- Convert categorical features into numerical format if necessary.
- Normalize or standardize numerical features for better model performance.
- Split the dataset into training data and testing data to evaluate performance.

Step 3: Train the Logistic Regression Model

- Initialize the Logistic Regression model.
- Train the model using the training dataset by fitting the feature variables to the target variable.

Step 4: Make Predictions

- Use the trained model to predict class labels for the test dataset.
- Obtain predicted probabilities to assess confidence levels.

Step 5: Model Evaluation

To measure the model's effectiveness, several evaluation metrics are used:

**Accuracy Score**

- Measures the proportion of correctly classified instances out of the total observations.

**Confusion Matrix**

- A table summarizing correct and incorrect predictions, providing insights into:
    - True Positives (TP): Correctly classified positive cases.
    - True Negatives (TN): Correctly classified negative cases.
    - False Positives (FP): Incorrectly classified positive cases.
    - False Negatives (FN): Incorrectly classified negative cases.

**Precision, Recall, and F1-Score**

- Precision: The proportion of predicted positive cases that are actually positive.
- Recall: The proportion of actual positive cases that were correctly predicted.
- F1-Score: A balance between precision and recall, providing an overall assessment of model performance.

**ROC Curve and AUC Score**

- The ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR).
- The AUC (Area Under the Curve) score quantifies the model's ability to distinguish between the two classes.

**Conclusion**

This practical demonstrates the implementation of Logistic Regression for binary classification using a dataset stored in a .CSV file. The model was trained and evaluated using various performance metrics, including accuracy, confusion matrix, precision, recall, F1-score, and AUC score.