Machine Learning Lab Problem Statement 2: Demonstrating Logistic Regression for Binary Classification

Problem Statement

You are tasked with predicting the survival of passengers on the Titanic based on various features such as age, sex, class, etc. Using the Titanic dataset, you will implement logistic regression to create a predictive model and evaluate its performance.

Objective:

The primary objective is to understand and apply logistic regression for binary classification. You will preprocess the data, build a logistic regression model, and evaluate its performance using appropriate metrics.

Dataset:

The dataset for this lab is the "Titanic: Machine Learning from Disaster" dataset available on Kaggle. This dataset contains information on the passengers aboard the Titanic, including whether they survived or not.

Dataset Link: https://www.kaggle.com/competitions/titanic

Steps to Follow

1. **Load and Explore the Data**

- Load the dataset using pandas.
- Perform exploratory data analysis (EDA) to understand the data structure and identify missing values.

2. **Data Preprocessing**

- Handle missing values appropriately.
- Convert categorical variables into numerical format using one-hot encoding.
- Drop unnecessary columns that do not contribute to the prediction.

3. **Define Features and Target Variable**

- Separate the dataset into features (X) and target variable (y).

4. **Split the Data**

- Split the data into training and testing sets using a 80-20 ratio.

5. **Build and Train the Model**

- Implement a logistic regression model using Scikit-Learn.
- Train the model on the training data.

6. **Make Predictions**

- Use the trained model to make predictions on the test set.

7. **Evaluate the Model**

- Calculate the accuracy of the model.

- Generate a confusion matrix and a classification report to assess precision, recall, and F1-score.

8. I**nterpret the Results**

- Analyze the model coefficients to identify which features are most influential.
- Discuss the model’s performance and suggest potential improvements.

9. **Expected Outcomes**

- Accuracy: Measure how well the model predicts survival.
- Confusion Matrix: Visualize true positives, false positives, true negatives, and false negatives.
- Classification Report: Detailed metrics including precision, recall, and F1-score for both classes.
- Feature Importance: Identify which features significantly impact the survival prediction.

10. **Deliverables**

- Jupyter Notebook: Containing the code and explanations for each step.
- Model Evaluation: Summary of the model’s performance with relevant metrics.
- Feature Analysis: Insights into the most important features influencing survival predictions.

By the end of this lab, you will have a solid understanding of how logistic regression works and how to implement it for binary classification problems. You will also gain experience in data preprocessing, model evaluation, and interpretation of results.