TP 2 : Titanic Survival Prediction with Logistic Regression

Problem:

Predict whether a passenger survived the Titanic disaster.

Data:

Use the "Titanic: Machine Learning from Disaster" dataset available on Kaggle: Titanic Dataset.

Tasks:

1. Data Exploration and Pre-processing:
   - Load the Titanic dataset.
   - Explore the dataset to understand its structure and features.
   - Handle missing values and perform necessary data cleaning.
2. Feature Engineering:
   - Select relevant features for prediction (e.g., age, gender, ticket class).
   - Encode categorical variables using one-hot encoding or label encoding.
3. Train a Linear Regression Model:
   - Split the dataset into training and testing sets.
   - Implement logistic regression using a library like scikit-learn and using the one we wrote on class.
4. Model Evaluation:
   - Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1 score.
   - Visualize the confusion matrix and ROC curve.
5. Feature Importance Analysis:
   - Analyze the coefficients of the logistic regression model to understand the impact of each selected feature on survival.
6. Bonus: Make Predictions for New Unseen Cars:
   - Acquire or generate information for new passengers.
   - Use the trained model to predict whether these passengers would survive or not.
   - Discuss potential use cases and limitations of the model.

Deliverables:

- Collab Notebook containing the code for data exploration, pre-processing, model training, and evaluation.
- Visualizations illustrating the dataset features, confusion matrix, ROC curve, and any relevant insights.
- A report summarizing key findings, insights from the coefficient analysis, and considerations for using the model in a real-world scenario.

This TP leverages a well-known dataset while introducing students to logistic regression in the context of predicting survival on the Titanic. The dataset is relatively simple, making it suitable for learners who are new to logistic regression.