Report: Titanic Survival Prediction using Logistic Regression

1. Objective

The primary goal of this task is to develop a machine learning model that predicts whether a passenger survived the Titanic disaster based on specific features. We utilize **Logistic Regression** to understand the end-to-end process from data preparation to model interpretation.

---

2. Data Preparation

2.1 Feature Selection

We selected the following features for the model based on their potential impact on survival:

- **Pclass:** Passenger class (1st, 2nd, or 3rd) acts as a proxy for socio-economic status.
- **Sex:** Gender of the passenger.
- **Age:** Age of the passenger.
- **SibSp & Parch:** Indicators of family size (siblings, spouses, parents, or children).
- **Fare:** The price paid for the ticket.
- **Embarked:** The port where the passenger boarded the ship.

Reasoning: Historical records indicate that "Women and Children First" policies and ticket class significantly influenced survival rates.

2.2 Encoding Categorical Variables

Since machine learning algorithms require numerical input, categorical text data was converted as follows:

- **Sex:** Encoded as **0 for male** and **1 for female**.
- **Embarked:** Converted using **One-Hot Encoding** (creating separate columns for each port).

Importance of Encoding: Mathematical models cannot perform calculations on strings like "male" or "Southampton." Encoding transforms these categories into a mathematical format the model can process.

2.3 Train-Test Split

The dataset was split into two parts:

- **80% Training Data:** Used to "teach" the model the patterns in the data.
- **20% Testing Data:** Used to evaluate how well the model predicts outcomes on unseen data.
- **Random State:** We used random_state=42 to ensure that the data split remains identical every time the code is run, allowing for consistent and reproducible results.

---

3. Model Building

3.1 Logistic Regression

Definition: Logistic Regression is a classification algorithm used to predict the probability of a binary outcome (1 or 0).

Suitability: This model is ideal for the Titanic problem because the target variable is binary: a passenger either **Survived (1)** or **Did Not Survive (0)**.

---

4. Model Evaluation

4.1 Accuracy

Accuracy represents the percentage of total predictions that the model got right.

- **Context:** If the model has an accuracy of 80%, it means it correctly predicted the survival status for 80 out of every 100 passengers in the test set.

4.2 Confusion Matrix

The confusion matrix provides a detailed breakdown of the model's performance:

- **True Positives (TP):** Passengers who actually survived and were correctly predicted as survivors.
- **True Negatives (TN):** Passengers who did not survive and were correctly predicted as non-survivors.
- **False Positives (FP):** Passengers who died but were incorrectly predicted by the model to have survived.
- **False Negatives (FN):** Passengers who survived but were incorrectly predicted by the model to have died.

---

5. Interpretation

4.1 Feature Impact

By analyzing the model coefficients, we determined the following:

- **Positive Impact (Increase Survival):** Being **Female** and being in **1st Class** had the highest positive correlation with survival.
- **Negative Impact (Decrease Survival):** Being **Male** and having a **higher Age** generally decreased the probability of survival.

Logical Reasoning: These findings align with the historical "Women and Children First" protocol. Additionally, 1st-class passengers had better access to lifeboats compared to those in 3rd class.

---

6. Conclusion & Limitations

While the Logistic Regression model provides a strong baseline, it has limitations:

1. **Linearity:** It assumes a linear relationship between features and the log-odds of survival.
2. **Missing Data:** Survival predictions are sensitive to how missing values (like Age) were handled during the cleaning phase.