Titanic Survival Prediction Using Logistic Regression

1. Introduction

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. In this project, we use data from the Titanic dataset to predict whether a passenger survived or not based on their demographic and ticket information. The machine learning model chosen for this task is Logistic Regression, a popular method for binary classification.

2. Objective

The primary goal of this project is to develop a model that predicts a passenger's survival based on factors such as age, sex, class, and fare. Using the Titanic dataset, we aim to train a Logistic Regression model and evaluate its accuracy.

3. Dataset

We used the Titanic dataset, which contains information about 891 passengers. Key features include:
Survived: 0 = No, 1 = Yes (target variable)
Pclass: Ticket class (1st, 2nd, 3rd)
Sex: Male or Female
Age: Passenger age in years
SibSp: Number of siblings/spouses aboard
Parch: Number of parents/children aboard
Fare: Fare paid by the passenger
Embarked: Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Missing Values
Age: 177 missing values, filled with the mean age.
Cabin: Mostly missing, hence dropped.
Embarked: 2 missing values, filled with the mode .

4. Exploratory Data Analysis

4.1 Statistical Summary

The dataset contains 891 rows and 12 columns, including categorical and numerical variables. The target variable is "Survived." Descriptive statistics were computed to understand the distribution of key features.

4.2 Data Visualization

We utilized Seaborn to visualize survival rates based on gender, class, and embarkation. Key observations:
- >More males did not survive compared to females.
-> Passengers in 1st class had a higher survival rate compared to 2nd and 3rd class.
- >The majority of passengers boarded from Southampton (S).

---

5. Data Preprocessing

5.1 Handling Missing Values
-The **Cabin** column was dropped due to a large number of missing values.
-The **Age** column's missing values were filled using the mean.
-The **Embarked** column's missing values were filled with the mode (most common value).

5.2 Encoding Categorical Data
Categorical columns such as **Sex** and **Embarked** were encoded into numerical values:
- Sex: Male = 0, Female = 1
- Embarked: S = 0, C = 1, Q = 2

5.3 Feature Selection
The following features were selected for the model:
- Pclass
- Sex
- Age
- SibSp
- Parch
- Fare
- Embarked

The target variable is **Survived**

6. Model Training

6.1 Logistic Regression
Logistic Regression was chosen as it is suitable for binary classification problems like survival prediction. The dataset was split into training (80%) and testing (20%) sets.

6.2 Training Process
The Logistic Regression model was trained on the training set. Here are the key steps:
Feature Selection: Dropped irrelevant columns like PassengerId, Name, and Ticket.
Training-Testing Split: The dataset was split into training (712 samples) and testing (179 samples).

The model was trained using `LogisticRegression()` from the Scikit-learn library.

7. Model Evaluation

7.1 Training Accuracy

After training, the model made predictions on the training dataset. The accuracy score of the training data was calculated using `accuracy_score` from Scikit-learn.

Training Data Accuracy: The model achieved a high accuracy in predicting survival on the training set, indicating that it has learned from the data effectively.

8. Conclusion

This project demonstrates the application of Logistic Regression to predict Titanic passenger survival using demographic and ticket data. The model's accuracy can be further improved with more advanced feature engineering and hyperparameter tuning.