Titanic Survival Prediction: Project Report

1. Introduction

This report details the process of building a machine learning model to predict whether a passenger on the Titanic survived or not. The project uses a logistic regression model, a common and effective algorithm for binary classification tasks. The primary dataset is the well-known "Titanic: Machine Learning from Disaster" dataset from Kaggle.

2. Data Exploration and Visualization

The initial step involved loading the dataset and performing exploratory data analysis (EDA) to understand its structure and features.

- **Initial Data Inspection:** The dataset contains 891 passenger records with features like Pclass (Passenger Class), Sex, Age, SibSp (Siblings/Spouses Aboard), Parch (Parents/Children Aboard), and Fare.
- **Missing Values:** A check for missing data revealed significant gaps:
  - Age: 177 missing values.
  - Cabin: 687 missing values.
  - Embarked: 2 missing values.
- **Visualizations:**
  - **Survival Count:** The overall survival count showed that more passengers perished than survived.
  - **Survival by Gender:** A key insight was that females had a much higher survival rate than males. - **Survival by Passenger Class:** Passengers in First Class (Pclass=1) had a significantly higher chance of survival compared to those in Third Class (Pclass=3). This suggests a strong correlation between socio-economic status and survival.

3. Data Preprocessing and Cleaning

To prepare the data for the model, several cleaning and transformation steps were necessary:

- **Handling Missing Age:** The missing Age values were imputed using the median age of the passenger's corresponding class (Pclass). This is a more robust approach than using the overall median age, as age distribution varied across classes.
- **Dropping Columns:**
  - The Cabin column was dropped entirely due to the high number of missing values.
  - PassengerId, Name, and Ticket were dropped as they are unique identifiers and not useful for prediction.
- **Handling Missing Embarked:** The two missing Embarked values were filled with the mode (the most common port of embarkation).
- **Converting Categorical Features:** The model requires numerical input. Therefore,

categorical columns like Sex and Embarked were converted into numerical format using one-hot encoding (pd.get_dummies). To avoid multicollinearity, one category from each feature (male, C for Cherbourg) was dropped.

4. Model Training

- **Feature and Target Selection:** The Survived column was designated as the target variable (y), while all other processed columns were used as features (X).
- **Data Splitting:** The dataset was split into a training set (80%) and a testing set (20%) using train_test_split. A random_state was set to ensure reproducibility.
- **Model Instantiation:** A LogisticRegression model was instantiated. The max_iter parameter was increased to 1000 to ensure the model's optimization algorithm had enough iterations to converge.
- **Training:** The model was trained using the .fit() method on the training data (X_train, y_train).

5. Model Evaluation

The model's performance was evaluated on the unseen test data.

- **Classification Report:**
  - **Accuracy:** The model achieved an overall accuracy of approximately 82%, meaning it correctly predicted the outcome for about 82% of the passengers in the test set.
  - **Precision:** The precision for predicting non-survival (0) was 83%, and for survival (1) was 80%. This indicates that when the model predicts a passenger did not survive, it is correct 83% of the time.
  - **Recall:** The recall for non-survival was 87%, while for survival it was 74%. This means the model successfully identified 74% of all the passengers who actually survived.
- **Confusion Matrix:**
  - **True Negatives (TN):** 91 passengers were correctly predicted as not having survived.
  - **True Positives (TP):** 55 passengers were correctly predicted as having survived.
  - **False Positives (FP):** 14 passengers were incorrectly predicted as having survived (they did not).
  - **False Negatives (FN):** 19 passengers were incorrectly predicted as not having survived (they did).

6. Conclusion

The logistic regression model provides a solid baseline for the Titanic survival prediction task, achieving an accuracy of 82%. The analysis confirms that key factors influencing survival were gender and passenger class.