# Report on Logistic Regression for Titanic Dataset

## 1. Objective

The aim of this project is to predict passenger survival on the Titanic using a Logistic Regression model.

## 2. Dataset

• Source: Kaggle Titanic dataset  
• Target Variable: Survived (0 = No, 1 = Yes)  
• Features Used:  
 - Pclass (ticket class)  
 - Sex (male/female)  
 - Age (with missing values filled by median)  
 - SibSp (siblings/spouses aboard)  
 - Parch (parents/children aboard)  
 - Fare (ticket fare)  
 - Embarked (one-hot encoded for C, Q, S)  
  
Dropped irrelevant columns: PassengerId, Name, Ticket, Cabin.

## 3. Methodology

1. Preprocessed the dataset: handled missing values, encoded categorical features.  
2. Split data into 80% training and 20% testing sets.  
3. Trained a Logistic Regression model (max\_iter=1000).  
4. Evaluated performance using accuracy and classification metrics.  
5. Checked feature importance using model coefficients.

## 4. Results

• Accuracy: ~80% (may vary slightly depending on split).

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Precision | Recall | F1-Score |
| Not Survived (0) | 0.82 | 0.89 | 0.85 |
| Survived (1) | 0.77 | 0.66 | 0.71 |
| Overall Accuracy |  |  | 0.80 |

• Key Features Influencing Survival:  
 - Sex (female → higher survival probability)  
 - Pclass (1st class passengers had better survival rates)  
 - Fare (higher fare linked to survival)  
 - Age (younger passengers survived more often)

## 5. Conclusion

• Logistic Regression achieved an accuracy of ~80%, providing a reliable baseline model.  
• Survival was strongly influenced by sex, class, and fare.  
• Future improvements could include:  
 - Feature engineering (e.g., extracting titles from names, creating family size).  
 - Hyperparameter tuning and cross-validation.  
 - Trying more complex models (Random Forest, Gradient Boosting).