Mini-Project Report

On

“Titanic Dataset Analysis and Passenger Survival Prediction using Logistic Regression”

Submitted By

Om Jagzap

Sachin Jaiswal

Vishal kambale

Under the guidance of

Ms. Pratiksha Gujar

DEPARTMENT OF COMPUTER ENGINEERING

Matoshri College of Engineering and Research Centre, Nashik

SAVITRIBAI PHULE UNIVERSITY, NASHIK

Academic Year [2025 – 2026]

INDEX

Introduction

The Titanic dataset is one of the most popular and well-known datasets in data science. It provides detailed information about passengers who were aboard the RMS Titanic, which tragically sank on its maiden voyage in 1912. The dataset includes information such as passenger age, gender, ticket class, fare, and survival status. This project focuses on performing Exploratory Data Analysis (EDA) and building a predictive model using Logistic Regression to determine the likelihood of a passenger's survival based on these attributes.

Exploratory Data Analysis (EDA) helps in understanding data distribution, detecting missing values, identifying patterns, and visualizing relationships between variables. Logistic Regression, a supervised learning algorithm, is then applied to predict survival outcomes. This project provides practical insights into the process of data cleaning, visualization, feature encoding, model training, and accuracy evaluation using Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn.

Problem Statement

The objective of this project is to analyze the Titanic dataset and predict the survival of passengers using machine learning techniques. The main challenges addressed include handling missing data, converting categorical variables into numerical format, visualizing survival patterns, and applying Logistic Regression to achieve accurate predictions. The project also aims to understand how factors like gender, age, passenger class, and family size affected the survival rate.

Requirement Analysis

Hardware Requirements:
- A personal computer or laptop with a minimum of 4 GB RAM.
- Processor: Dual-core or higher.
- Storage: Minimum 500 MB free space.
- Internet connectivity for dataset download.

Software Requirements:
- Operating System: Windows / Linux / macOS
- Programming Language: Python 3.x
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Warnings
- Development Environment: Jupyter Notebook / VS Code / PyCharm

1. Objective of Requirement Analysis

The main objective of this phase is to:

- Identify and specify the tools, software libraries, and hardware configuration needed.
- Ensure compatibility between the dataset, development environment, and programming libraries.

2. Functional Requirements

The system must be able to perform the following functions:

1. Import and explore the Titanic dataset.
2. Perform **Exploratory Data Analysis (EDA)** to understand trends and distributions.

System Analysis

The system is designed to perform data analysis and prediction using a structured workflow consisting of multiple modules. Each module plays a crucial role in transforming raw data into meaningful results.

Module 1: Data Loading and Exploration

- The dataset is imported using Pandas. Initial exploration is done to view data structure, count null values, and check data types using describe() and info().

Module 2: Data Visualization (EDA)

- Various graphs are plotted using Matplotlib and Seaborn to visualize relationships between features such as survival rate, gender, passenger class, age, siblings, and parents on board.

Module 3: Data Cleaning

- Missing data in Age, Cabin, and Embarked columns are handled.
- Unnecessary columns such as Cabin and Ticket are dropped.
- Missing values are filled using mean values or removed if required.

Module 4: Feature Encoding

- Categorical features such as ‘Sex’, ‘Embarked’, and ‘Pclass’ are converted into numerical form using one-hot encoding (get_dummies).

Module 5: Model Building (Logistic Regression)

- The Logistic Regression model is trained using Scikit-learn.
- The dataset is divided into training and testing sets using train_test_split.
- The model learns patterns to predict survival probability.

Motive

The motive of this project is to understand the end-to-end process of data analysis, from raw data cleaning to model prediction, using a real-world dataset. By analyzing the Titanic dataset, the project helps in developing a clear understanding of data preprocessing, visualization, and predictive modeling techniques.

This project provides hands-on experience in applying statistical and machine learning concepts to a real dataset. It also demonstrates how data-driven decision-making can be achieved using Python and modern data science libraries. Furthermore, the study highlights how factors like gender, age, and passenger class influenced survival chances during the Titanic disaster, reflecting the social and historical context of that time.

Result

After completing data cleaning, visualization, and model training, the Logistic Regression model achieved an accuracy of approximately 80.42%. This indicates that the model correctly predicted survival for more than 80% of passengers in the test dataset.

Observed results show that women and first-class passengers had a significantly higher survival rate compared to men and lower-class passengers. Age and fare also had noticeable effects, with younger passengers and those who paid higher fares having better chances of survival.

Metric	**Result**
Model Used	Logistic Regression
Accuracy Achieved	80.42%
Dependent Variable	Survived
Independent Variables	Age, SibSp, Parch, Fare, Sex, Passenger Class, Embarked
Algorithm Type	Supervised Learning (Classification)

Conclusion

The Titanic survival prediction project successfully demonstrates the process of data analysis and model building using Python. Through detailed EDA, it was observed that survival rates were highly dependent on gender, passenger class, and age. The Logistic Regression model effectively classified survivors with over 80% accuracy.

This project enhanced understanding of data preprocessing, encoding techniques, feature selection, and model evaluation. It also emphasizes the importance of data quality and visualization in developing reliable machine learning models. In conclusion, this study serves as a comprehensive example of applying machine learning to solve real-world classification problems.