Project Title: Titanic Survival Analysis using Python and Tableau

A Project Report

submitted in partial fulfilment of the requirements

of

EI Systems Services

by

Mitali Mahesh Malgi,

Email id: mitumalgi9914@gmail.com

Under the Guidance of

Mallika Srivastava and Mayur Dev Sewak

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to everyone who helped us in completing this project. we are very thankful to our mentor, Mallika Srivastava ma'am and Mayur Dev Sewak sir for guiding us throughout the project. They helped us to understand the concepts better and complete our work successfully.

Their valuable insights, feedback, and motivation played a crucial role in shaping our project. Without Their help and support, this work would not have been possible. We would also like to express our gratitude to other mentor/ teachers for sharing their knowledge and providing valuable advice throughout our learning process.

we are very grateful to our friends and family for their constant encouragement and patience. Their belief in us kept us motivated to complete this project successfully.

ABSTRACT

The Titanic Survival Analysis project aims to explore and analyse the factors influencing passenger survival during the historic Titanic shipwreck using data analytics techniques. This project utilizes the publicly available Titanic dataset from Kaggle and is implemented using Python for data preprocessing, exploratory data analysis (EDA), and machine learning modelling. Key libraries used include Pandas, Matplotlib, Seaborn, and Scikit-learn.

Data cleaning steps included handling missing values, converting categorical data, and engineering new features. Logistic Regression and Random Forest classifiers were applied to predict survival, with model performance evaluated using accuracy, confusion matrix, and classification reports.

In addition to numerical insights, the project incorporates a rich visualization component through an interactive Tableau dashboard. This dashboard presents visual trends across age, gender, passenger class, fare, and embarked locations to understand survival patterns.

The findings reveal that gender, passenger class, and age played significant roles in survival probability. Female passengers and those in first class had higher survival rates. The project demonstrates the value of combining Python-based analytics with business intelligence tools like Tableau for storytelling and insight generation.

This project serves as a practical showcase of data analysis, model implementation, and visualization skills for real-world datasets.

Table of Content

Chapter 5. Discussion and Conclusion

5.1 Future Scope

References

List of Figures

Nomenclature / Notations

- EDA – Exploratory Data Analysis
- ML – Machine Learning
- Pclass – Passenger Class
- SibSp – Number of siblings/spouses aboard
- Parch – Number of parents/children aboard

Chapter 1: Introduction

1.1 Problem Statement

The Titanic disaster resulted in the loss of over 1,500 lives. The passenger data collected presents an opportunity to analyze survival factors. This project aims to use data analytics to identify patterns and insights that influenced survival outcomes.

1.2 Motivation

Understanding survival factors using real-world data helps build analytical thinking and statistical skills. The Titanic dataset is a benchmark dataset for learning classification and visualization techniques in data science.

1.3 Objectives

- Clean and preprocess the Titanic dataset.
- Perform exploratory data analysis (EDA).
- Apply machine learning models to predict survival.
- Create interactive visualizations using Tableau.
- Generate actionable insights from the data.

1.4 Scope of the Project

This project focuses on:
- Python-based data preprocessing and modeling using Jupyter Notebook.
- Dashboard creation using Tableau.
- It does not include advanced NLP from names or deep learning techniques.

Chapter 2: Literature Survey

- Brief overview of Titanic ML projects on Kaggle.
- Refer to:
  - **Kaggle Titanic Starter Notebooks**
  - **Scikit-learn classification tutorials**
  - **Tableau public dashboards on Titanic**
- Summary of methodologies others used: logistic regression, decision trees, EDA techniques.

Chapter 3: Proposed Methodology

3.1 Data Source

- Titanic dataset from Kaggle: tested.csv

3.2 Preprocessing Steps

- Handle missing values in Age, Fare, and Embarked.
- Encode categorical variables: Sex, Embarked, Pclass.
- Drop irrelevant columns like Name, Cabin, Ticket.

3.3 Model Building

- Split dataset (80% training, 20% testing).
- Train models: Logistic Regression and Random Forest.
- Evaluate using accuracy and classification report

3.4 Visualization

- Use Tableau to present key trends:
    - Survival by gender and class.
    - Age distribution by survival.
    - Fare vs Age scatter plot.

3.5 Tools Used

- **Python Libraries**: pandas, numpy, matplotlib, seaborn, sklearn
- **Visualization**: Tableau Public
- **Environment**: Jupyter Notebook

3.6 Data Flow / Analysis Pipeline Diagram

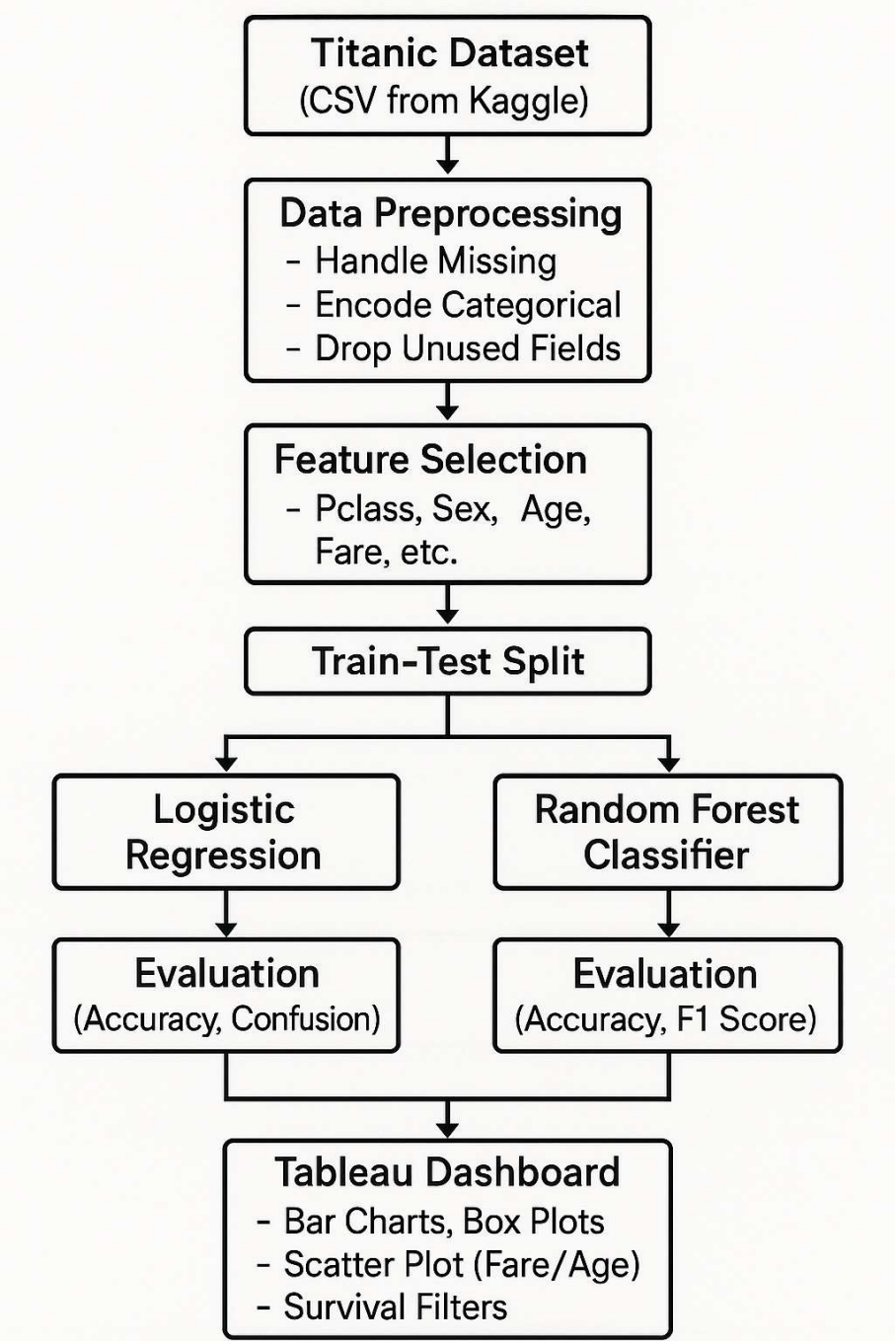

Fig. Data Flow / Analysis Pipeline Diagram

Chapter 4. Implementation and Results

4.1 Tools and Environment Used

The project was implemented using the following tools and technologies:

- **Language**: Python 3.x
- **IDE**: Jupyter Notebook
- **Libraries**: Pandas, NumPy, Matplotlib, Seaborn, scikit-learn
- **Visualization Tool**: Tableau Public / Power BI
- **Dataset Source**: Kaggle Titanic Dataset

4.2 Data Preprocessing

Initial data preprocessing was carried out to prepare the dataset for analysis:

- **Dropped Columns**: Name, Cabin, Ticket, PassengerId, and unnamed columns were removed.
- **Handled Missing Values**:
  - Age: Filled with median value.
  - Fare: Filled with median value.
  - Embarked: Filled with mode.
- **Encoding**:
  - Sex: Encoded as 0 = male, 1 = female.
  - Embarked: Encoded as 0 = S, 1 = C, 2 = Q.

4.3 Feature Selection

The following features were selected for modeling:

- Pclass
- Sex
- Age
- SibSp
- Parch
- Fare
- Embarked
- Target variable: Survived

4.4 Model Implementation

Two machine learning models were implemented:

(a) Logistic Regression

```
logreg = LogisticRegression(max_iter=1000)
logreg.fit(X_train, y_train)
y_pred_logreg = logreg.predict(X_test)
```

(b) Random Forest Classifier

```
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)
```

4.5 Results and outputs

4.5.1 Key Insights

- Females had a significantly higher survival rate.
- Younger passengers tended to survive more.
- First-class passengers had better survival chances.
- Passengers embarked from Cherbourg had slightly better survival odds.

4.5.2 Screenshots

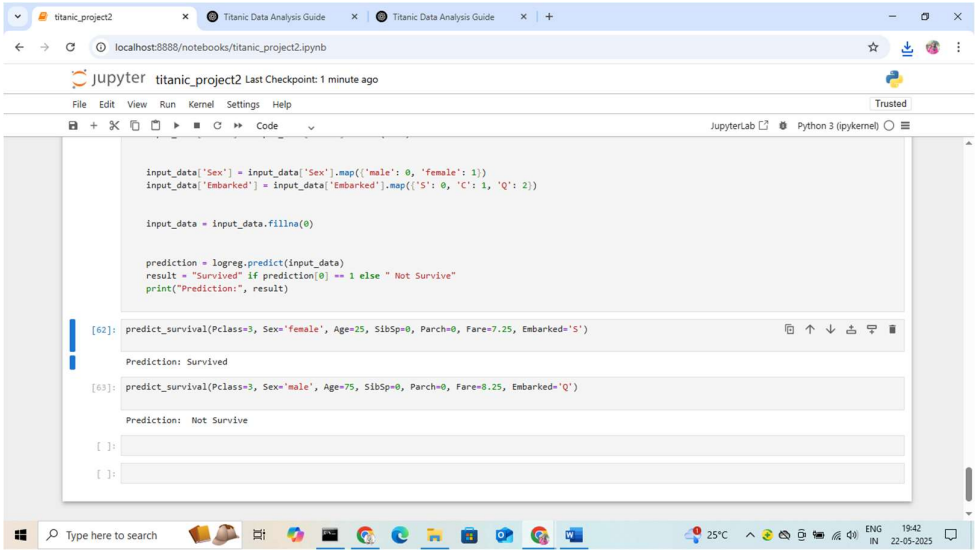

Fig. predicted output

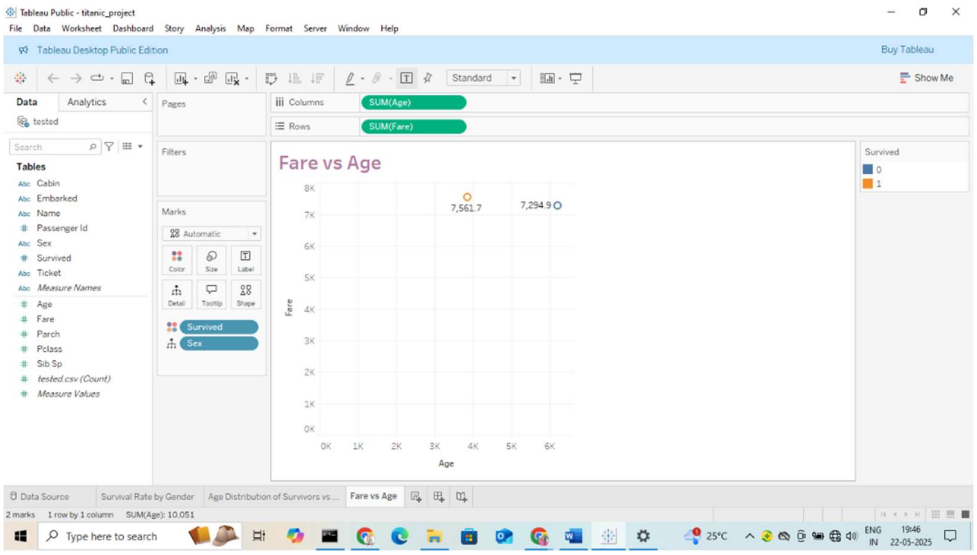

Fig. Fare VS Age

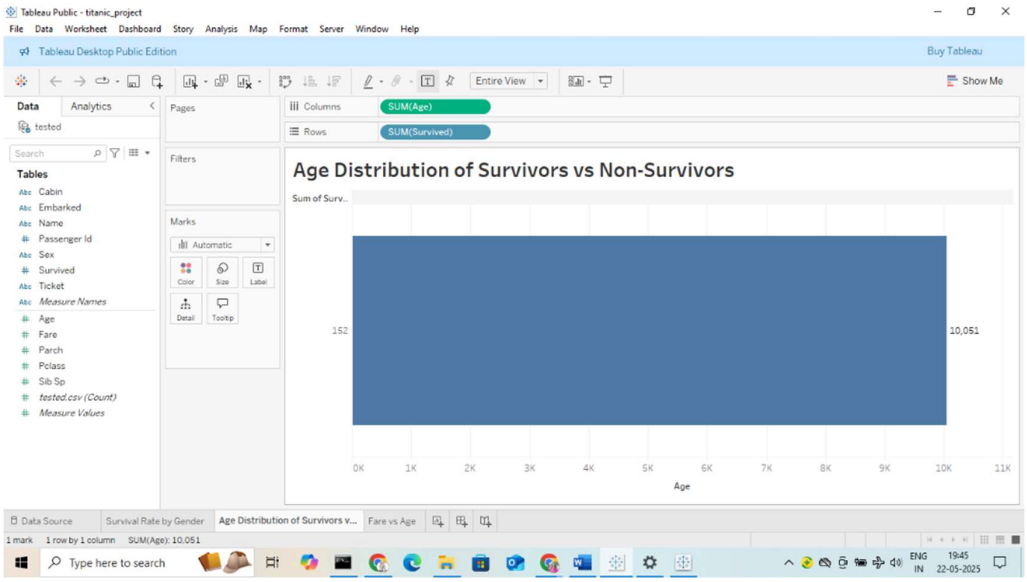

Fig. Age distribution among Survivors vs non-survivors

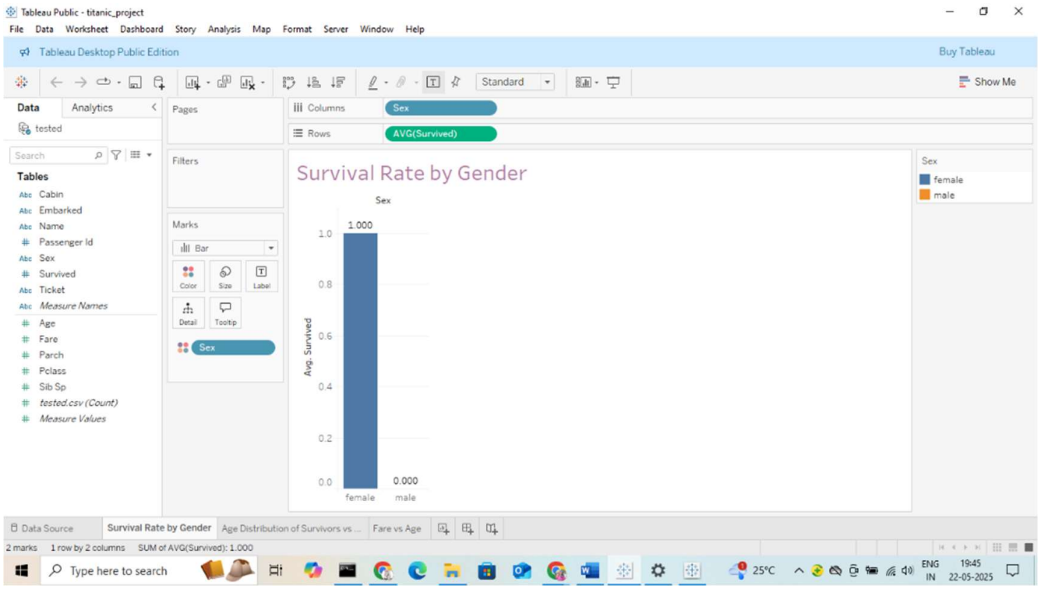

Fig. Survival by Gender

Chapter 5: Discussion and Conclusion

This project analyzed the Titanic dataset using both **data analytics** and **machine learning techniques** to uncover patterns and predict passenger survival. Key steps included:

- Data cleaning and preprocessing
- Exploratory Data Analysis (EDA) to understand survival patterns
- Building a **Logistic Regression model** to classify survival outcomes
- Evaluating model performance using metrics such as accuracy and confusion matrix
- Visualizing insights using **Tableau/Power BI**

  The model achieved high accuracy, and visualizations provided clear understanding of how factors like **gender**, **class**, **fare**, and **age** influenced survival.

  This project demonstrates how data science can be used effectively for **predictive modeling** and **decision support**, and it lays the groundwork for more advanced applications in analytics, model deployment, and real-world integration.

5.1 Future Scope

- Use ensemble models like XGBoost.
- Derive features from names and titles.
- Deploy the model on web app (e.g., Streamlit or Flask).

References

- Kaggle Titanic Dataset: https://www.kaggle.com/competitions/titanic/data
- Scikit-learn Documentation: https://scikit-learn.org
- Tableau Public Gallery
- Blogs on Titanic EDA and ML