Titanic Passenger Survival Analysis

Abstract

This project presents an analytical study of passenger survival during the Titanic disaster using Python. The work integrates data preprocessing, statistical analysis, database management, visualization, graphical user interfaces, and automated testing. The objective is to demonstrate a complete data analysis workflow suitable for academic and practical applications.

Keywords

Titanic Dataset, Data Analysis, Python, SQLite, Tkinter, Unit Testing, Visualisation.

1. Introduction

The Titanic dataset is a benchmark dataset in data science education. This project aims to analyze survival patterns among passengers while applying modular programming, database integration, and software testing principles.

2. Tools and Technologies

- Python
- Pandas, numpy
- Matplotlib
- Tkinter
- Unnitest
- Pytest
- Sqlite3, SQLAlchemy

3. Methodology

3.1. Data collection

The dataset used is the Kaggle Titanic training dataset (train.csv).

3.2. Data Preprocessing

- Handling missing values (Age, Embarked)
- Encoding categorical variables (Sex, Embarked)
- Normalizing numerical features (Fare)

3.3. Statistical Analysis

- Survival rate by gender
- Survival rate by passenger class
- Age statistics (mean, min, max)

4. System Architecture

- Data Processing (utils.py)
- Statistical Analysis (module1.py)
- Database Management (module2.py)
- Visualization (visualization.py)
- Graphical User Interface (gui.py)

5. Testing and Validation

5.1. Unit Testing with unittest

Two test modules were implemented:

test_module1.py

- Validates survival analysis logic
- Ensures females have higher survival rate in the test dataset

test_module2.py

- Tests SQLite save and load operations
- Confirms data integrity after database operations

5.2. Test Execution

python -m unittest discover tests
All tests passed successfully, validating correctness of the core functions.

6. Results and Discussion

- Female passengers had significantly higher survival rates
- First-class passengers had better chances of survival
- Majority of passengers were adults between 20 and 40 years old

7. Conclusion

This project demonstrates a complete Python-based data analysis pipeline suitable for academic coursework. It successfully combines data analysis, visualization, database management, GUI development, and testing.