An Analytical Look At The Deaths Within The Titanic

By: Ali El-Hagagi

# GitHub URL

https://github.com/Ali-Mohamed1999/UCDPA\_Ali-

**Abstract:**

This project aims to analyze the Titanic dataset and predict the factors that influenced survival on the ill-fated ship. The dataset provides information about passengers, including age, gender, socio-economic class, and more. The project's main objectives involve understanding the dataset, performing data preprocessing, conducting exploratory data analysis, and building a predictive model. Through visualizations, the dataset is explored, revealing insights such as the majority of passengers not surviving, a higher percentage of male passengers, and the age distribution. The analysis focuses on the impact of gender, age, passenger class, and port of embarkation on survival.

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, I attempt to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

This dataset has been referred from Kaggle: <https://www.kaggle.com/c/titanic/data>

**Introduction:**

The sinking of the Titanic in 1912 has remained a captivating event, inspiring numerous studies and adaptations in various forms of media. This project delves into the Titanic dataset, offering a unique opportunity to explore and analyze the factors that influenced passengers' chances of survival.

Motivated by a fascination with the Titanic story and a desire to uncover valuable insights, this project aims to gain a better understanding of the passengers' demographics, characteristics, and the factors that contributed to their survival. Additionally, it provides an avenue to apply data preprocessing techniques, conduct exploratory data analysis, and build a predictive model using machine learning algorithms.

Furthermore, this project offers an opportunity to develop important skills in data analysis, visualization, and machine learning. Working with real-world data and employing various techniques will enhance knowledge and proficiency in the field of data science. The predictive modeling aspect of the project allows for testing and improvement of skills in implementing machine learning algorithms and evaluating their performance.

**Dataset:**

The project utilizes the Titanic dataset sourced from Kaggle. This dataset provides detailed information about the passengers aboard the RMS Titanic during its ill-fated maiden voyage. It includes features such as age, gender, socio-economic class (Pclass), ticket fare, cabin, port of embarkation, and survival status.

The dataset consists of two separate files: a training dataset and a test dataset. The training dataset contains information about a subset of passengers, including their survival status. The test dataset includes information about the remaining passengers without the survival status. The project's goal is to build a predictive model using the training dataset to determine the likelihood of survival for passengers in the test dataset.

By selecting the Titanic dataset, this project leverages its historical significance, real-world relevance, and rich feature set to gain insights, develop data analysis skills, and build a predictive model to understand the factors that influenced passenger survival on the ill-fated Titanic.

**Implementation Process:**

Dataset Introduction:

The analysis utilizes a dataset sourced from Kaggle, comprising two main files: "train.csv" and "test.csv." The "train.csv" file serves as the training data, including the target variable "Survived." On the other hand, the "test.csv" file contains similar data without the "Survived" column. The objective is to build a model capable of predicting the survival status of passengers in the test dataset based on the provided features.

Project Goals:

The main goals of this project are as follows:

* Understanding the dataset and its features.
* Performing data preprocessing to handle missing values and prepare the data for analysis.
* Conducting exploratory data analysis using various visualization techniques.
* Building a predictive model to determine the factors that influenced survival on the Titanic.
* Evaluating the model's performance and discussing the results.

**Implementation**

Importing Libraries:

The project utilizes several libraries, including:

* NumPy for numerical operations.
* Pandas for data manipulation and analysis.
* Matplotlib.pyplot for data visualization.
* Seaborn for enhanced data visualization.
* Warnings to handle warnings.
* OS to interact with the operating system.

Reading the Dataset:

The training and test datasets are read into separate dataframes using the Pandas library. The "train.csv" file is loaded into the "train\_df" dataframe, while the "test.csv" file is loaded into the "test\_df\_actual" dataframe. To gain an overview of the data, the first few rows of the "train\_df" dataframe are displayed.

Basic Analysis:

The basic analysis section involves exploring the dataset's shape, column names, data types, and detailed information about the columns. The number of rows and columns in the "train\_df" dataframe is displayed. The column names and data types are shown using the "columns" and "dtypes" attributes of the dataframe, respectively. Additionally, the "info()" function provides detailed information about the columns, including the count of non-null values and data types.

The analysis reveals missing values in some columns such as "Age", "Cabin", and "Embarked." The exact count of null values in each column is displayed using the "isnull().sum()" function.

Data Cleaning and Preprocessing:

To handle missing values, the following steps are performed:

The "Age" column is filled with the mean of the "Age" column from both the training and test datasets.

The "Fare" column in the test dataset is filled with the mean of the "Fare" column.

The "Embarked" column in the training dataset is filled with the mode (most frequent value) of the "Embarked" column.

The "Cabin" column is dropped from both the training and test datasets due to a large number of missing values. Additionally, the "PassengerId", "Name", and "Ticket" columns are dropped as they are not expected to be useful for analysis and prediction.

Finally, the index of the "train\_df" dataframe is reset to ensure proper indexing after removing duplicate rows using the "drop\_duplicates()" function. The duplicated rows are dropped, and the resulting dataframe is displayed.

Analysis, Visualization, and Insights:

This section focuses on analyzing the data and gaining insights through visualization. The following visualizations and insights are presented:

Distribution of Survival: The count and percentage distribution of the "Survived" variable are shown using a countplot and a pie chart. The majority of passengers did not survive the Titanic disaster.

A graph of a number of people

Description automatically generated

Gender Distribution: The distribution of passengers based on gender is displayed using a countplot. There is a higher percentage of male passengers compared to female passengers.

A graph with a red and blue rectangle

Description automatically generated

Age Distribution: A histogram is plotted to visualize the distribution of passenger ages. The average age is around 30 years, with a wide range of ages from infants to elderly passengers.

A graph of a number of passengers

Description automatically generated

Survival Distribution by Gender: A bar plot visualizes the count of survivors and non-survivors based on gender. The number of female survivors is higher than the number of male survivors.

A graph of a person and person

Description automatically generated

Survival Distribution by Age: A bar plot visualizes the count of survivors and non-survivors based on age groups. The majority of non-survivors are between 16-30 years old, while the majority of survivors are below 16 years old.

A graph of a person and person

Description automatically generated

Passenger Class by Gender: A stacked bar plot displays the distribution of passenger class based on gender. Most passengers from both genders prefer the 3rd class.

A graph with red and blue bars

Description automatically generated

Survival Distribution by Port of Embarkation: A bar plot visualizes the count of survivors and non-survivors based on the port of embarkation. Most passengers boarded from Southampton.

A graph with a bar chart

Description automatically generated

Distribution of Numerical Values: Histograms are plotted to show the distribution of age and fare values.

A graph of age and fare

Description automatically generated

**Results:**

Based on the analysis and insights gained from the Titanic dataset, the predictive models can provide valuable information about the factors that influenced survival on the Titanic. By evaluating the performance of the trained models, we can assess their effectiveness in accurately predicting survival status. The following insights can be drawn:

* The majority of passengers did not survive the Titanicdisaster.
* The percentage of male passengers is higher than that of female passengers.
* The average age of passengers is around 30 years, with a wide age range.
* The number of female passengers who survived is higher than the number of male passengers who survived.
* The majority of non-survivors are between 16-30 years old, while the majority of survivors are below 16 years old.
* Regardless of gender, most passengers prefer the 3rd class.
* The majority of passengers boarded from the port of Southampton.

By comparing the performance of different models, we can identify the most effective one in predicting survival on the Titanic. Additionally, feature importance techniques, such as analyzing the coefficients in logistic regression or the feature importances in tree-based models like random forest or gradient boosting, can provide insights into which factors have the most significant influence on survival.

Overall, the results of the predictive models can enhance our understanding of the factors that contributed to the survival of passengers on the ill-fated Titanic and provide a framework for future predictions or analysis of similar scenarios.

**Machine Learning:**

Within the context of the Titanic dataset, machine learning techniques can be employed to build a predictive model that determines the factors influencing survival on the Titanic. Given that the dataset is imbalanced, with more non-survived passengers, it is important to address this issue to avoid bias and overfitting. One approach to tackle class imbalance is to use oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), to create synthetic samples of the minority class (survived) and balance the dataset.

Several machine learning models can be applied to the classification problem of predicting survival status. Here are some commonly used models:

1. Logistic Regression: This model is a simple and interpretable linear classifier that estimates the probability of survival based on the input features. It is a suitable starting point for binary classification problems like predicting survival.
2. Support Vector Classifier (SVC): SVC is a powerful model that can capture complex relationships in the data by mapping it to a higher-dimensional space. It seeks to find an optimal hyperplane that separates the survived and non-survived passengers.
3. K Neighbors Classifier: This model classifies each passenger based on the majority class of its nearest neighbors. It calculates the distance between samples and assigns labels based on the k nearest neighbors. It can capture local patterns in the data.
4. Decision Tree: A decision tree builds a hierarchical structure of if-else rules based on the features' values. It partitions the data into subsets by asking questions at each internal node, leading to leaf nodes that represent the predicted survival status.
5. Random Forest: A random forest is an ensemble model that combines multiple decision trees. It creates a diverse set of decision trees by training each tree on a random subset of features and samples from the dataset. The final prediction is determined by majority voting or averaging across the individual trees.
6. Gradient Boosting: Gradient boosting is another ensemble model that combines weak learners (usually decision trees) in a sequential manner. It fits subsequent models to the residuals of the previous model, gradually improving the prediction accuracy.

These models can be trained using the training dataset, and their performance can be evaluated using appropriate metrics such as accuracy, precision, recall, and F1 score. Cross-validation techniques, such as k-fold cross-validation, can be used to assess the models' generalization ability and ensure robustness.