1. Problem Statement

The Titanic disaster is one of the most infamous shipwrecks in history. In 1912, the Titanic sank after colliding with an iceberg, leading to the deaths of most of its passengers and crew. In this project, we analyze a dataset containing demographics and passenger information from 891 of the 2224 passengers and crew on board. The challenge is to use machine learning to create a model that predicts which passengers would have survived the Titanic shipwreck.

2. Approach

Our approach include the several steps listed below:

- Data Acquisition: The dataset was obtained, which includes various features such as age, sex, passenger class, and whether the passenger survived.
- Data Cleaning: The data was cleaned to handle missing values, erroneous entries, and outliers. Features like ‘Cabin’ were engineered to extract useful information.
- Exploratory Data Analysis (EDA): Conducted thorough analysis to understand the relationships between different features and survival rates.
- Feature Engineering: Created new features such as ‘Family Size’ from ‘SibSp’ (siblings aboard) and ‘Parch’ (parents/children aboard). Converted categorical variables into numerical variables for modeling.
- Modeling: Started with basic models like Logistic Regression and Decision Trees. Progressed to more complex models like Gradient Boosting and XGBoost to improve prediction accuracy.
- Evaluation: Models were evaluated using accuracy, precision, recall, F1-score, and ROC curves. The best-performing model was identified through these metrics.

3. Findings

In our analysis, we identified several key predictors of survival on the Titanic, notably including passenger sex, class, and age. Women, higher-class passengers, and younger individuals were significantly more likely to survive, reflecting social hierarchies and physical abilities impacting survival chances during the disaster. Our comparative evaluation of various predictive models showed that the Gradient Boosting and XGBoost models outperformed others, achieving the highest ROC AUC scores, which indicate their strong capability in distinguishing between survivor and non-survivor classes. The ROC curves and confusion matrices further substantiated the robustness of these models, demonstrating their high sensitivity and specificity in predicting outcomes based on the dataset. These findings provide a solid foundation for understanding the dynamics of survival in maritime disasters and for developing more effective safety protocols.

4. Recommendations for the Client

Based on our analysis, we recommend that safety protocols should prioritize women, children, and elderly passengers, as these groups demonstrated higher survival rates and could benefit significantly from targeted support in emergencies. Additionally, we suggest using our predictive model to identify passengers at higher risk during emergencies to optimize the allocation of critical resources such as lifeboats and safety gear. This targeted approach allows for more efficient use of resources, ensuring that those most in need are adequately prepared and supported during crises.

5. Ideas for Further Research

For further research, we propose several avenues to enhance the predictive capabilities and applicability of our models. Firstly, enriching the dataset with additional variables such as ticket price and detailed cabin location could provide deeper insights and refine model accuracy. Exploring alternative modeling techniques, including deep learning, could also offer improved performance and handle larger datasets more effectively. Additionally, analyzing the duration of survival post-impact would provide valuable information on the effectiveness of emergency responses and could guide improvements in safety protocols and training programs. These expansions would not only broaden the understanding of survival factors but also contribute to more robust safety strategies in similar scenarios.