Project Proposal: Predicting Passenger Survival on the Titanic

Problem Identification

The sinking of the Titanic remains one of the most infamous maritime disasters in history. Out of 2,224 passengers and crew, many people died when the ship sank. The key question of this project is: Can we predict which passengers were more likely to survive the Titanic disaster based on available demographic and socioeconomic data?

Problem Statement

The objective of this project is to create a predictive model that accurately classifies passengers as survivors or non-survivors using various features such as age, gender, ticket class, and fare. By understanding the factors that influenced survival rates, we can gain insights into the socioeconomic and demographic disparities that may have played a role in the outcome of this tragedy.

Context

The Titanic dataset includes information on passenger demographics, ticket details, cabin numbers, and whether they survived or not. This project will use machine learning techniques to predict survival, providing a practical application of classification algorithms and feature engineering.

Criteria for Success

Success will be defined by the following:

- Model Accuracy:The predictive model should achieve an accuracy of at least 70% on the test set.
- Insights: The model should provide interpretable insights into which factors were most significant in determining survival.
- Reproducibility:The process and results should be reproducible by others using the same data and code.

Scope of Solution Space

The project will explore various classification algorithms, including logistic regression, decision trees, random forests, and support vector machines. Feature engineering will play a crucial role, with the potential creation of new features such as family size, title from name, and cabin deck. The solution will include:

- Data Cleaning and Preprocessing: Handling missing data and encoding categorical variables.
- Feature Engineering: Creating and selecting the most relevant features.
- Modeling: Training, tuning, and validating multiple machine learning models.
- Evaluation: Using metrics such as accuracy, precision, recall, and F1-score to evaluate model performance.

Constraints

- Data Quality: The dataset contains missing values (e.g., in age and cabin) that need to be addressed.
- Imbalanced Classes: The dataset has more non-survivors than survivors, which could bias model predictions.

— Computational Resources: The project will be conducted using standard machine learning libraries and tools.

Stakeholders

— Data Science Community: This project serves as a practical example for data scientists and students looking to learn about classification problems.
— Historians and Researchers:Insights gained from the analysis may provide historical context on survival disparities during the Titanic disaster.
— Educational Institutions: The project can be used as a teaching tool for courses in data science, machine learning, and statistics.

Data Sources

— The primary data source is the Titanic dataset available on Kaggle, which includes:
— Passenger information Demographics, ticket class, fare, and cabin details.
— Survival outcome: indicator of whether the passenger survived.