5241 Final Project Proposal

Yongyuan Qu - yq2361
Xiaocong Xuan - xx2438
Yuhao Wang - yw3924
Wenjun Yang - wy2431
Yitong Liu - yl5549

1. Problem Formulation

The Titanic disaster remains one of history's most poignant maritime tragedies, marked by the substantial loss of life it caused. This project seeks to address the problem of predicting the survival outcomes of passengers based on available demographic and socio-economic data. By analyzing the factors that influenced survival rates, this study aims to shed light on historical socio-economic disparities and inform the development of enhanced safety protocols for contemporary maritime transportation.

2. Dataset Description

The dataset used for this project is derived from the Titanic passenger list, sourced from Kaggle (https://www.kaggle.com/competitions/titanic/data). The data is split into two distinct sets:

- Training Set (train.csv): This dataset includes demographic and socioeconomic information for a subset of the Titanic's passengers. It features variables such as passenger class (Pclass), sex, age, number of siblings/spouses aboard (SibSp), number of parents/children aboard (Parch), ticket number, fare, cabin number, and the port of embarkation. Crucially, it also includes the survival status (Survived), where 1 indicates survival and 0 indicates non-survival. This set is used to build and train the machine learning models, providing the 'ground truth' needed for supervised learning.
- Test Set (test.csv): Similar to the training set in structure but without the survival outcome. The purpose of this dataset is to evaluate the performance of the trained models on unseen data. The test set allows for the assessment of how well the predictive model generalizes to new data.

These datasets provide a comprehensive base for developing and testing machine learning models aimed at predicting outcomes from historical events, in this case, the survival of passengers from the Titanic disaster.

3. Methodology

The methodology for predicting survival rates involves several machine-learning models and analytic techniques:

a. Data Preprocessing:

- Handling Missing Data: Missing values in 'Age', 'Cabin', and 'Embarked' will be imputed using statistical methods like median imputation for numerical data and mode imputation for categorical data. In addition, rows containing missing values in any of the critical features will be dropped from the dataset.
- Feature Engineering: Features such as 'Title' extracted from passenger names and a binary 'Alone' indicator based on 'SibSp' and 'Parch' will be created to enhance the model's predictive accuracy.

b. Machine Learning Models:

- Logistic Regression: This model will be used to establish a baseline for comparison. It's particularly useful for binary classification problems and will provide insights into the importance of different features in predicting survival.
- Genetic Algorithm: To optimize feature selection, a genetic algorithm will be employed, improving the efficiency and performance of the predictive models by selecting the most relevant features.
- Neural Network: A neural network model will be constructed using TensorFlow's Keras API. The network will feature layers suitable for binary classification tasks, with dropout layers to prevent overfitting and an Adam optimizer for efficient training.

c. Evaluation:

- Model Evaluation Metrics: Models will be evaluated using accuracy, precision, recall, F1-score, and ROC curves. These metrics will help in understanding the effectiveness of the models in classifying the survival outcomes.
- Cross-Validation: To ensure the models do not overfit, k-fold cross-validation will be employed during the training process.

This methodology outlines a comprehensive approach to applying machine learning techniques to historical data, aiming to provide insights into the factors that influenced survival during the Titanic disaster. By leveraging advanced analytics, this project not only aids in historical analysis but also enhances our understanding of data-driven decision-making in safety and crisis management.