# Team Project 1 (Team11)

CS 53744 Machine Learning Project - Instructor Professor: Jongmin Lee

Department of Computer Science & Engineering

**20222446 Hyunsoo Kim, 20201876 Sanghyun Na, 20203009 Jaehyun Park**

「A Study on the Development of a Classification Model for Titanic Survival Prediction」

## Abstract

This project addresses the Titanic Survival Prediction task from the Kaggle competition 'Titanic: Machine Learning from Disaster.' We implemented a complete machine learning workflow, starting from simple baseline rules to more advanced predictive modeling. Exploratory Data Analysis (EDA) revealed strong correlations between survival and factors such as gender, passenger class, and family size. Feature engineering introduced new variables such as family size, extracted titles, and age groups that improved model performance.

Logistic Regression served as the primary predictive model, validated through train–validation splits. Our final model achieved competitive accuracy on the Kaggle leaderboard, outperforming the rule-based baselines and demonstrating the importance of feature engineering in tabular prediction problems.

## Introduction

The Titanic dataset is a widely known open dataset for binary classification, aiming to predict whether a passenger survived the disaster based on travel features. The task illustrates the full machine learning pipeline: establishing baselines, performing Exploratory Data Analysis (EDA), creating new features, and applying predictive models.

## Baseline Models

For Step 1. Predict All Deceased, our simplest submission predicted every passenger as not survived (Survived = 0). This produced ~61% accuracy on the Kaggle public leaderboard.

In Step 2. Gender Rule Next, we applied the well-known rule that most females survived, and most males did not: female → Survived = 1, male → Survived = 0.

This rule achieved ~78% accuracy, a substantial improvement over the naive baseline.

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis highlighted several clear patterns in survival rates. Female passengers had a survival rate exceeding 70%, while fewer than 20% of males survived. Passenger class also showed strong effects: first-class passengers exhibited the highest survival probability, whereas third-class passengers had the lowest.

Age was another important factor, as children aged 12 or younger had higher chances of survival compared to adults. Family size influenced outcomes as well; passengers traveling alone or within very large families were less likely to survive.

Finally, fare played a role, with higher ticket prices correlating with improved survival, reflecting socioeconomic advantages

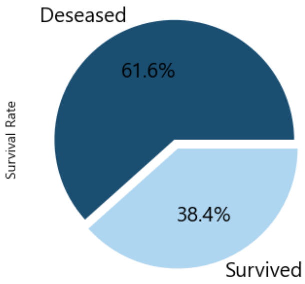


Figure 1. Overall survival rate

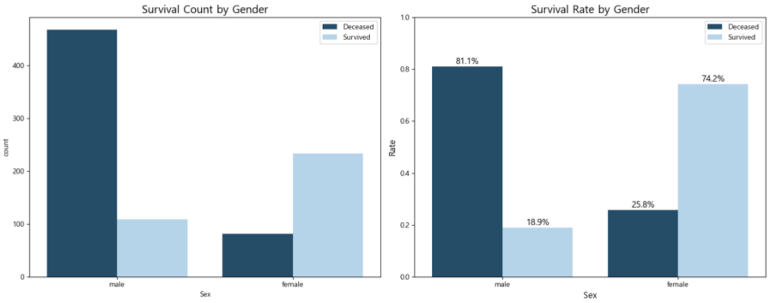


Figure 2. Survival count & rate by gender

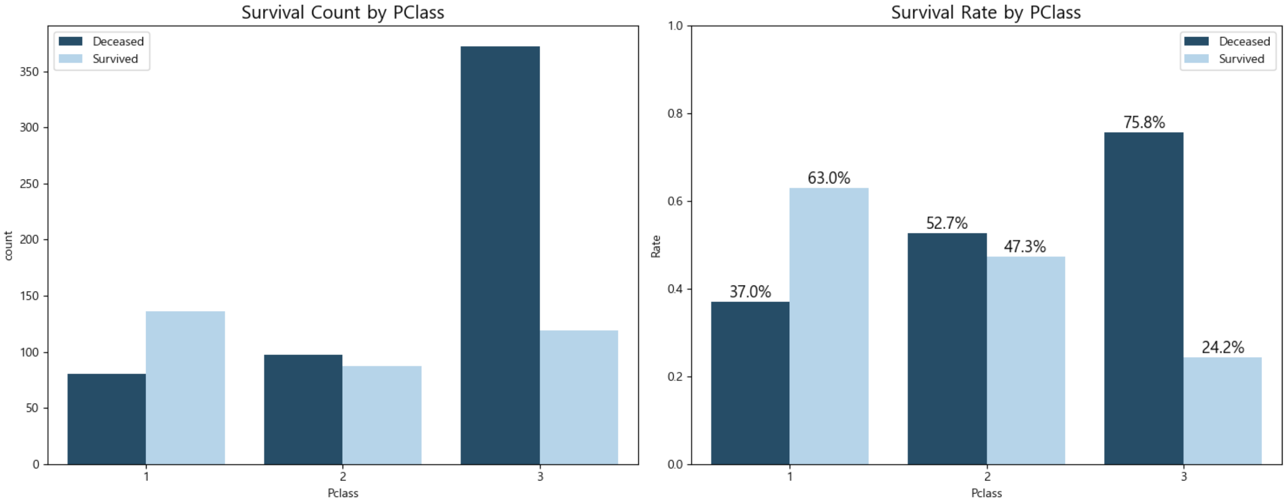


Figure 3. Survival count & rate by PClass

Figure 1 shows the overall survival rate in the dataset, confirming that only about 38% of passengers survived. A clear gender gap is visible in Figure 2, where over 70% of females survived, compared to fewer than 20% of males. Passenger class also strongly influenced survival chances (Figure 3), with first-class passengers having the highest survival probability.

Age groups (Figure 4) showed that children (under12) had a significantly higher chance of survival than adults or the elderly. In addition, survival was affected by family-related features (Figure 5 and 6).

Passengers traveling with small family groups had higher chances of survival, while those traveling alone or with very large families were less likely to survive.

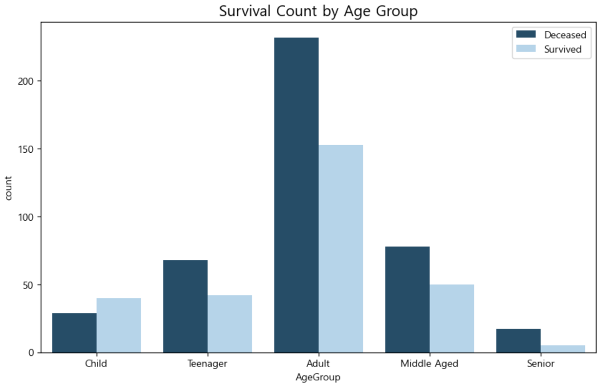


Figure 4. Survival count by age group

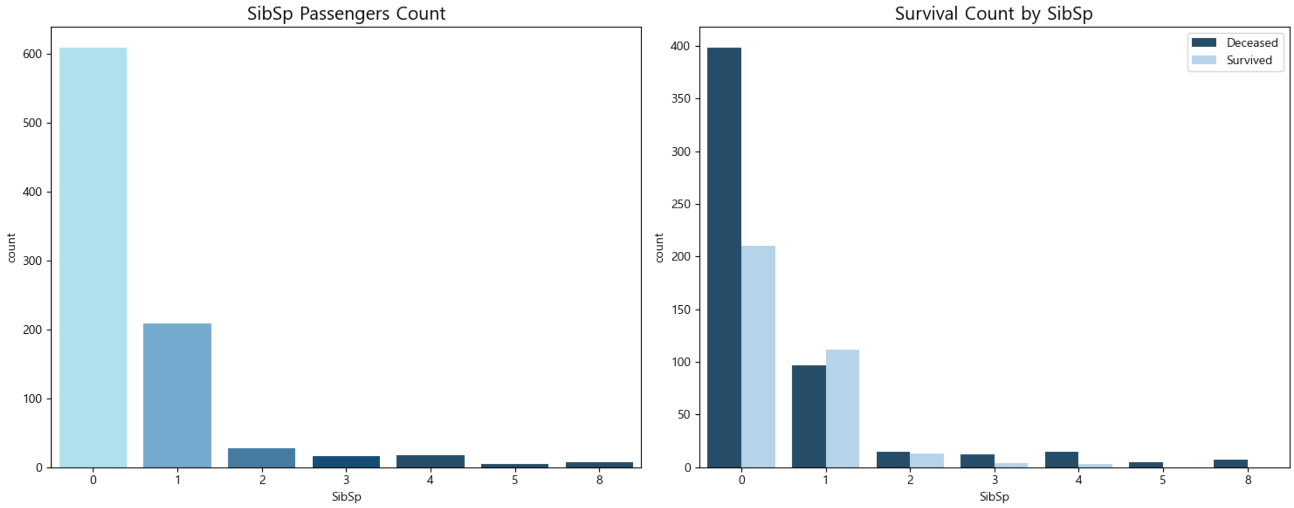


**Figure 5. Survival by number of siblings / spouses aboard (SibSp)**

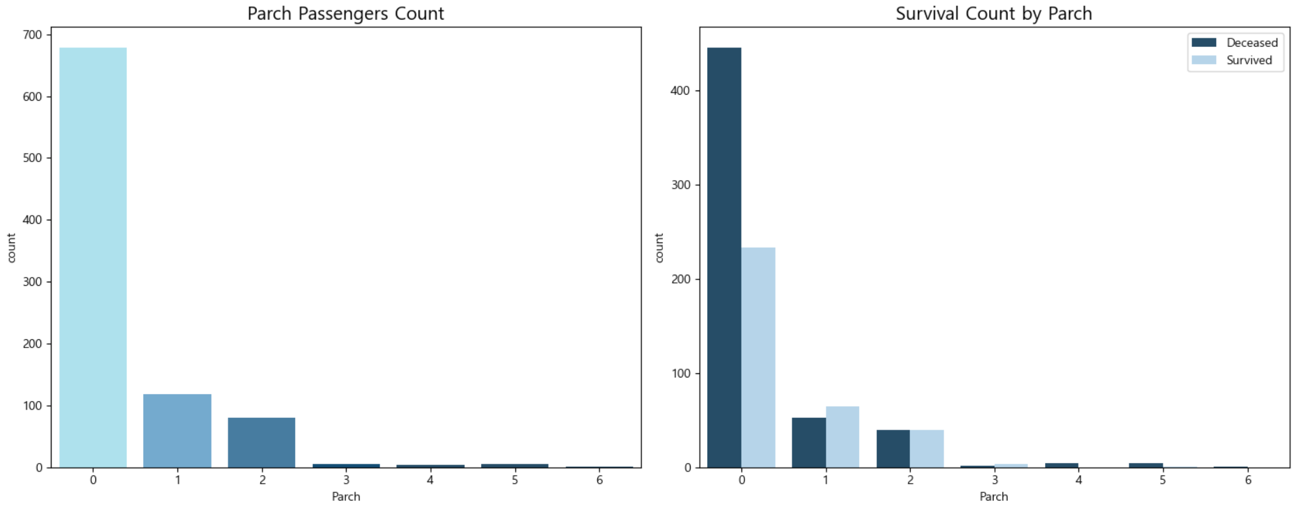


**Figure 6. Survival by number of parents / children aboard (ParCh)**

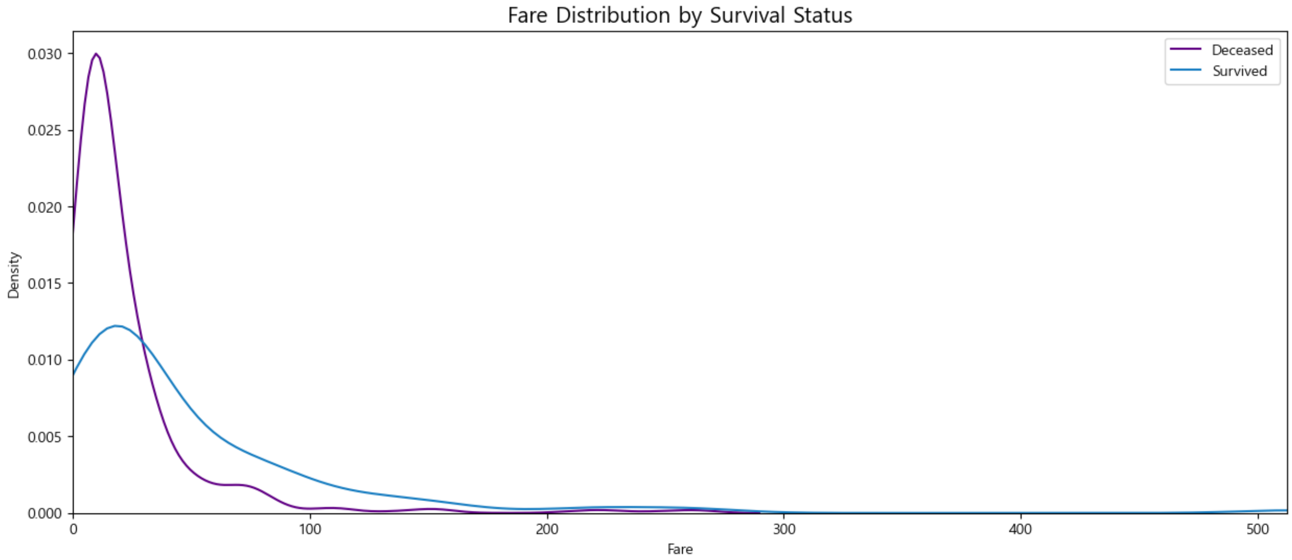


**Figure 7. Fare distribution by survival status**

Fare distribution by survival status (Figure 7) indicated that passengers who paid higher fares (often first-class) had a better chance of survival.

## Feature Engineering

To enhance the model's expressiveness, we engineered several new features. Family size was calculated as the sum of siblings, spouses, parents, and adding one more for children aboard. Titles were extracted from the Name field (Mr, Mrs, Miss, and Master) to capture social status and gender cues. Age was grouped into categories of Child (under 13), Adult (13–59), and Elderly (60 and above). Additionally, we created 'Solo' tag to mark passengers traveling without family members.

Beyond these, we also introduced features derived from ticket information. The Ticket Frequency feature measured how many passengers shared the same ticket number. We discretized both Fare (into nine quantile bins) and Age (into ten quantile bins) to reduce the effect of extreme outliers and to better capture non-linear relationships with survival.

Finally, categorical features (Sex, Embarked, and Titles) were converted into one-hot encoded vectors, while numerical features were normalized using Min-Max Scaling to ensure balanced contributions across the model. These preprocessing steps ensured that the feature set was both expressive and consistent, enabling the downstream machine learning models to capture survival patterns more effectively.

### Model Development

We initially trained a Logistic Regression classifier using scikit-learn. The dataset was divided into training and validation sets with an 80/20 split. Continuous features such as Age and Fare were scaled with StandardScaler, while categorical variables including Sex, Title, and Pclass were transformed using one-hot encoding.

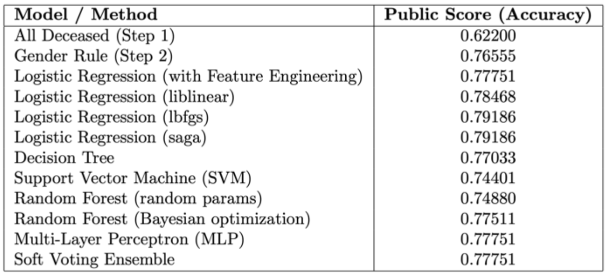
Validation accuracy ranged from approximately 80% to 82%, depending on the chosen feature combinations. Our first Kaggle submission with feature engineering and Logistic Regression achieved a public leaderboard score of 0.77751**,** which was higher than the rule-based baselines (0.62200 for all-deceased and 0.76555 for gender rule). This confirmed that systematic modeling and engineered features improve predictive performance beyond simple heuristics.

### Extended Models

We experimented with multiple solvers for Logistic Regression. Both **lbfgs** and **saga** solvers achieved the highest score of **0.79186**, while **liblinear** produced a slightly lower result of **0.78468**. This suggests that gradient-based solvers provided a small but consistent improvement over liblinear.

Beyond Logistic Regression, we extended our experiments to a range of classical machine learning models. A Decision Tree reached **0.77033**, while Support Vector Machine (SVM) scored **0.74401**. Random Forest showed varied results depending on hyperparameter tuning: the random baseline achieved **0.74880**, whereas Bayesian optimization improved it to **0.77511**. A Multi-Layer Perceptron (MLP) neural network yielded **0.77751**, comparable to Logistic Regression with feature engineering.

Finally, a Soft Voting Ensemble combining multiple models also achieved **0.77751**, showing that ensembling produced stable but not necessarily superior performance compared to the best individual Logistic Regression solvers.



| Model / Method | Public Score (Accuracy) |
|---|---|
| All Deceased (Step 1) | 0.62200 |
| Gender Rule (Step 2) | 0.76555 |
| Logistic Regression (with Feature Engineering) | 0.77751 |
| Logistic Regression (liblinear) | 0.78468 |
| Logistic Regression (lbfgs) | 0.79186 |
| Logistic Regression (saga) | 0.79186 |
| Decision Tree | 0.77033 |
| Support Vector Machine (SVM) | 0.74401 |
| Random Forest (random params) | 0.74880 |
| Random Forest (Bayesian optimization) | 0.77511 |
| Multi-Layer Perceptron (MLP) | 0.77751 |
| Soft Voting Ensemble | 0.77751 |

## Conclusion

The Titanic prediction task demonstrated the progression from naive rules to machine learning models. While simple heuristics already achieved strong performance, feature engineering and logistic regression provided measurable improvements. Key insights included the importance of gender, passenger class, and family status in survival. Overall, this project was a useful exercise in balancing model performance with interpretability, and it helped us better understand the practical steps involved in building predictive models.