**Team Project 1 (Team11)**

CS 53744 Machine Learning Project - Instructor Professor: Jongmin Lee

Submission Date: September 30, 2025

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 (e.g., 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 canonical benchmark for binary classification, aiming to predict whether a passenger survived the disaster based on demographic and travel features. The task illustrates the full machine learning pipeline: establishing baselines, performing EDA, creating new features, and applying predictive models. Beyond accuracy, the assignment emphasizes interpretability, reproducibility, and understanding of model limitations.

**Baseline Models**

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

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)**

EDA highlighted clear patterns in survival rates:  
- Gender: Females had a survival rate exceeding 70%, while males had less than 20%.  
- Passenger Class (Pclass): First-class passengers had the highest survival probability; third-class the lowest.  
- Age: Children (≤12) showed higher survival rates compared to adults.  
- Family Size (SibSp + ParCh + 1): Extremely large families or passengers traveling alone had lower survival chances.  
- Fare: Higher ticket fares correlated with better survival, reflecting socioeconomic advantages.

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 (under 12) 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. Fare distribution by survival status (Figure 7) indicated that passengers who paid higher fares (often first-class) had a better chance of survival.

Missing value analysis (Figure 8 and 9) highlighted incomplete records for Age and Cabin, guiding our feature engineering choices.

* **Figure 1.** Overall survival rate of Titanic passengers.
* **Figure 2.** Survival count and rate by gender.
* **Figure 3.** Survival count and rate by passenger class.
* **Figure 4.** Survival rate 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.
* **Figure 8.** Missing values in training dataset.
* **Figure 9.** Missing values in test dataset.

**Feature Engineering**

To enhance model expressiveness, we created additional features:  
- FamilySize = SibSp + ParCh + 1.  
- Title Extraction from the Name field (e.g., Mr, Mrs, Miss, Master).  
- Age Groups: Child (<13), Adult (13–59), Elderly (≥60).  
- IsAlone: Indicator for passengers with no family aboard.  
  
These engineered features provided more direct signals for survival compared to the raw variables.

**Model Development**

We trained a Logistic Regression classifier using scikit-learn:  
- Data Split: 80% training / 20% validation.  
- Scaling: StandardScaler applied to continuous features (Age, Fare).  
- Encoding: One-hot encoding for categorical variables (Sex, Title, Pclass).  
  
Validation Accuracy: ~80–82%, depending on feature combinations.  
Final Kaggle Submission: Logistic Regression with feature engineering achieved a public leaderboard score higher than the rule-based baseline (~0.78), confirming the benefit of systematic modeling.

**Extended Models**

Beyond Logistic Regression, we briefly tested Random Forest and Support Vector Machines:  
- Random Forest yielded similar or slightly better validation accuracy (~83%), though at the cost of reduced interpretability.  
- Logistic Regression was retained as the primary submission due to simplicity, interpretability, and competitive performance.

**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. This project highlights the balance between model complexity and interpretability in practical prediction problems.