Data 572: Project working title

Avalvir Kaur Sekhon, Zihuan Liu & Jordan Kaseram

February 5, 2026

Abstract

Introduction

hello

Methodology

Titantic Data Set

The Titanic dataset consists of passenger-level information from the sinking of the RMS *Titantic*. Each observation corresponds to an individual passenger, with the response variable indicating survival status (1 = survived, 0 = did not survive). The dataset contains a mixture of demographic, socioeconomic, and voyage-related features, including age, sex, passenger class, fare paid, family composition, and cabin information.

In total, the dataset contains 891 observations with a diverse set of categorical and numerical variables. As is typical of real-world data, the raw dataset presents several challenges, including missing values, highly sparse columns, and correlated features derived from similar underlying information.

These characteristics make the Titanic dataset well suited for evaluating classification methods while also requiring careful preprocessing and feature selection prior to model fitting.

Data Preprocessing

Data preprocessing was performed using Python and the pandas library, which provides flexible and efficient tools for data cleaning, transformation, and preparation prior to machine learning analysis. Identifier variables such as passenger identifiers, ticket numbers, and booking references were removed, as they do not contain predictive information and may introduce unnecessary noise into the models.

Missing data were handled using simple, distribution-preserving imputation techniques. Missing values in passenger age, a numerical variable, were imputed using the median age within each passenger class. This approach allows age estimates to reflect socioeconomic differences across classes while remaining robust to outliers. Missing values in the port of embarkation, a categorical variable, were imputed using the mode, preserving the empirical distribution of this feature.

Categorical predictors were converted into numerical form using one-hot encoding creating dummy variables. To reduce redundancy and limit multicollinearity, several highly correlated or derived features were evaluated and pruned. For example, the variables `SibSp`, `Parch`, and `is_alone` were replaced by a single variable `family_size`.

The resulting feature set provides a stable and interpretable foundation for training supervised machine learning models for binary classification.

Train-Test Splitting Strategy

The dataset was partitioned into training and test sets using a stratified split, with 75% of the data allocated to training and 25% reserved for testing. Stratification ensured that the class distribution of survivors and non-survivors was preserved across both sets. All preprocessing steps involving data-driven quantities, such as imputation values, were computed using the training set only to avoid data leakage.

Model Selection

Experiment

Expermential Design

Results and Analysis

Discussion

Conclusion

References

[1] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: Python edition.

[2] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179-188.