Data 572: Project working title

Avalvir Kaur Sekhon, Zihuan Liu & Jordan Kaseram

February 3, 2026

Abstract

Introduction

hello

Methodology

Data Preprocessing

The Titantic dataset was preprocessed to enhance its suitability for statistical modeling and to reduce noise and redundancy. Identifer features lacking predictive relevance, such as passenger identifers or ticket numbers, were removed prior to analysis. Missing values in numeric features were imputed using the median, whereas missing categorical variables were imputed using the mode. Categorical features were encoded as dummy variables to use in linear modeling frameworks.

To further reduce redundancy and limit multicollinearity, several features were evaluated and pruned. For example, family-related variables were replaced with a single family size

variable. This step is necessary since there are many features that are highly correlated and limiting

Experiment

Expermential Design

Results and Analysis

Discussion

Conclusion

References

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