Report

Introduction

The aim of this analysis is to build a model that can predict whether a passenger survived or not based on the available features, based off the titanic disaster. The provided Titanic datasets are used to train and test the models then the results are analyzed. Finally the best model is chosen and hyperparameter tuning is used to improve the models performance.

Data Exploration

The two datasets *train.csv* and *test.csv* are loaded and reviewed in a table form to understand the structure and content of the data. While analyzing the dataset its understood which columns are useful for our model building and which columns have missing values and basically the whole plan is formulated.

Data Cleaning

missing values are carefully filled by the mode or median of the column. The columns that are irrelevant are dropped.

Feature Engineering

One hot encoding is used to convert categorical variable values into numerical values. .

Model Building

A logistic regression and Random forest models are built to predict survival.

Model Evaluation

The models are tested and and the results are analyzed then the model with the best results is selected as the most suitable for the prediction of survival.

Model Tuning

Hyperparameter tuning is performed to improve the performance of the model.

Findings

Upon model evaluation a classification report was used to discover the performance of the model. Accuracy, precision, f1-score and recall where the main metrics that were targeted and used to decide which model was better.
The logistic regression model was chosen as the most suitable model for predicting survival.

Conclusion

The approaches and steps taken maximized the suitability of the model to predict survival in the titanic disaster.
