Machine Learning
Mini Project Report

Title: Predicting Titanic Survival

Group Members: Ashish Avhad (BE-A-37)

Tushar Pawar (BE-A-38)

Snehal Salunke (BE-A-39)

Atharva Shewale (BE-A-40)

Abstract

This project aims to predict the survival of passengers on the Titanic using a machine learning approach. Utilizing the Titanic dataset from Kaggle, we explore various features such as age, gender, and socio-economic class to build a predictive model. Through data preprocessing, feature engineering, and model training with a Random Forest classifier, we achieve an accuracy of approximately 81% on the test set. This report details the methodology, results, and potential areas for future improvement in the predictive modeling of Titanic survival.

Introduction

The objective of this project is to predict the survival of passengers on the Titanic using a machine learning model. This prediction is based on various features such as age, gender, socio-economic class, and more. The dataset used for this project is obtained from Kaggle.

Dataset Overview

The Titanic dataset consists of data on 891 passengers with the following features:

- **PassengerId**: Unique identifier for each passenger
- **Survived**: Survival status (0 = No, 1 = Yes)

- **Pclass**: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **Name**: Passenger name
- **Sex**: Gender
- **Age**: Age of the passenger
- **SibSp**: Number of siblings/spouses aboard the Titanic
- **Parch**: Number of parents/children aboard the Titanic
- **Ticket**: Ticket number
- **Fare**: Passenger fare
- **Cabin**: Cabin number
- **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Methodology

1. Data Preprocessing

Data preprocessing is crucial for handling missing values, encoding categorical variables, and ensuring the dataset is suitable for training a machine learning model.

1.1 Data Cleaning and Feature Engineering

- **Dropping Unnecessary Columns**: Columns such as PassengerId, Name, Ticket, and Cabin were dropped as they do not directly contribute to the prediction model.
- **Handling Missing Values**:
    - The Age column had missing values, which were filled using the mean value of the column.
    - The Fare column in the test data also had missing values, filled similarly.
    - The Embarked column had missing values, filled with the mode (most frequent value).
- **Encoding Categorical Variables**:

- The Sex and Embarked columns were converted to numerical values using label encoding.

1.2 Train-Test Split

The dataset was split into training and testing sets to evaluate the model's performance. 80% of the data was used for training, and 20% was reserved for testing.

1.3 Feature Scaling

Feature scaling was applied to standardize the range of the features, ensuring that each feature contributes equally to the model training.

2. Model Training

A Random Forest classifier was chosen for this task due to its robustness and ability to handle both numerical and categorical data. The model was trained on the training set using the processed features and the survival status as the target variable.

3. Model Evaluation

The model's performance was evaluated on the test set using the following metrics:

- **Accuracy**: The proportion of correctly predicted survival statuses.
- **Confusion Matrix**: A table used to describe the performance of the classification model.
- **Classification Report**: A detailed report showing the precision, recall, and F1-score for each class.

Results

The Random Forest classifier achieved an accuracy of approximately 81% on the test set. The confusion matrix and classification report indicated that the model performed well in distinguishing between passengers who survived and those who did not.

Discussion

The results demonstrate that certain features, such as gender, age, and socio-economic class, significantly influence the survival chances of passengers. The preprocessing steps ensured that the model was trained on clean and well-prepared data, contributing to its good performance.

Conclusion

This project successfully built a machine learning model to predict Titanic survival with reasonable accuracy. Future improvements could include trying other machine learning algorithms, tuning hyperparameters, and incorporating more sophisticated feature engineering techniques.