Linear Regression

# Abstract

This report is on the Titanic dataset, determining the prediction of survival using linear regression. The dataset is first loaded and explored, with “sex”, “class”, and “age” the main things looked at. Then data cleaning techniques are used and one hot encoding as well. Finally, a logistic regression model is applied on the dataset.

# Introduction

The Titanic dataset is a widely used dataset for predictive modeling and machine learning. The goal of this analysis is to predict passenger survival based on various features such as sex, age, and class. The process involves exploratory data analysis, data cleaning, and the implementation of a logistic regression model for classification.

# Methodology

## Data Exploration

The dataset is loaded, and initial exploratory data analysis is done to understand it better. Descriptive statistics are examined, and visualizations are created to understand the distribution of key variables, such as survival, gender, and class. In addition, data cleaning is done. Missing values are visualized using a heatmap, and then appropriate actions are taken, including dropping rows with missing values. Categorical variables like sex, embarked, and class are one-hot encoded for model compatibility.

## Model Training and Testing

The logistic regression model is then used for survival prediction. The dataset is split into training and testing sets, and the model is trained on the training set as per usual. The trained model is then used to make predictions on the test set. Performance metrics such as confusion matrix and accuracy score are calculated to evaluate the model's predictive capabilities.

# Results

The graphs and the model itself illustrate the distribution of survival across different variables, highlighting potential factors influencing survival rates. The logistic regression model, after training and testing, achieves an accuracy score of 76.4%, indicating a reasonably good predictive performance. The confusion matrix provides additional insights into the model's ability to correctly classify survival outcomes.

# Discussion

The exploratory analysis emphasizes the importance of variables such as sex, age, and class in understanding survival patterns. The logistic regression model, while providing a reasonable accuracy, may benefit from more training and data to improve its accuracy.

# Conclusion

In conclusion, this report presents an analysis of the Titanic dataset, focusing on survival prediction through logistic regression. The process includes data exploration, cleaning, and model training, with a discussion of the results and potential areas for improvement.