# Titanic Machine Learning from Disaster - InternCareer Report

Your Name

January 10, 2024

## 1 Introduction

The Titanic Machine Learning from Disaster dataset is a classic dataset widely used for practicing and improving machine learning skills. The dataset consists of information about passengers on the Titanic, including features like age, sex, class, and whether they survived or not.

## 2 Data Cleaning

### 2.1 Missing Values

The initial step in the analysis involved checking for missing values in both the training and test datasets. The missing values were handled as follows:

- **Age:** Missing values were replaced with the median value.
- **Embarked:** The missing values were replaced with the most frequent value.
- **Cabin:** Missing values were replaced with 'Unknown'.
- **Fare:** Missing values were replaced with the median value.

## 3 Feature Engineering

### 3.1 Title Extraction

Titles were extracted from the 'Name' column to create a new feature called 'Title'. To simplify the model, titles with fewer occurrences were grouped into an 'Other' category.

### 3.2 Drop Unnecessary Columns

Several columns, including 'Name', 'Ticket', 'Cabin', and 'Age', were dropped as they were deemed unnecessary for the model.

### 3.3 Convert Categorical Variables

Categorical variables like 'Sex', 'Embarked', and 'Title' were converted to numerical format for model compatibility.

## 4 Exploratory Data Analysis (EDA)

EDA was performed to understand the distribution of the target variable ('Survived') and visualize the distributions of numerical and categorical features. Countplots and histograms were used for visualization.

## 5 Bivariate Analysis

Survival rates were analyzed based on different features such as 'Pclass' and 'Sex'. Bar plots were used to illustrate the relationships.

## 6 Model Selection and Training

A Random Forest Classifier was chosen for its ability to handle non-linear relationships and feature importance. The model was trained on the pre-processed training data.

## 7 Model Evaluation

The model was evaluated on a validation set using metrics such as accuracy, classification report, and confusion matrix. The results were satisfactory, indicating the model's ability to generalize.

## 8 Submission

Predictions were made on the test set, and a submission file was created with 'PassengerId' and 'Survived' columns.

# 9 Conclusion

The Random Forest Classifier demonstrated good performance on the validation set. Further improvements could involve hyperparameter tuning or exploring other algorithms. The submission file is ready for assessment, and the documentation provides transparency into the decision-making process.