# Titanic Survivors Prediction Model

Name: Ned Palom

Date of Completion: 25/11/2024

**Predicting Titanic Survivors Using Machine Learning.**

# Table of Contents

1. Introduction

2. Dataset Description

3. Exploratory Data Analysis (EDA)

4. Data Preprocessing

5. Modeling

6. Results

7. Conclusion

8. Future Work

9. Appendix

# 1. Introduction

**Problem Statement:**  
The project addresses the problem of predicting the survival of passengers aboard the Titanic based on various features such as age, sex, class, and ticket information. This problem is framed as a binary classification task, where the goal is to determine whether a passenger survived or not based on historical data (https://www.kaggle.com/competitions/titanic)

**Objective:**  
The objective of the project is to build a machine learning model that accurately predicts the survival of passengers using the Titanic dataset. By analyzing the dataset, the project also aims to understand the key factors that influenced survival rates during the disaster

**Dataset Overview:**  
The Titanic dataset contains information on 891 passengers, including features such as:

* **Survival** (0 = No, 1 = Yes)
* **Pclass** (Passenger class: 1st, 2nd, 3rd)
* **Sex**
* **Age**
* **SibSp** (Number of siblings/spouses aboard)
* **Parch** (Number of parents/children aboard)
* **Ticket** (Ticket number)
* **Fare** (Ticket fare)
* **Cabin** (Cabin number, with many missing values)
* **Embarked** (Port of embarkation: C = Cherbourg, Q = Queenstown, S = Southampton)

2. Dataset Description

- **Number of rows and columns.**In the Titanic dataset, there are a total of **11** columns and **891** rows

- **Missing Values**

There are **868** missing values in the Train dataset, and **413** missing values in the Test dataset

**Technologies Used**

* **Programming Language:** Python
* **Libraries and Tools:**
  + **Pandas** and **NumPy** for data manipulation and analysis
  + **Scikit-learn** for building machine learning models
  + **Jupyter Notebook** for coding and experimentation​

# 3. Exploratory Data Analysis (EDA)

**Summary statistics (mean, median, mode, etc.)**

**Central Tendency matrix:**

**Train Dataset –** (Cabin and Embarked column I used the mode to impute the missing values) and Use the mean method to fill in the missing ages and change the Age column into ‘Column\_Age’

**Test Dataset** -

Key insights gained from EDA should be highlighted.

# 4. Data Preprocessing

Describe the steps taken to prepare the dataset for modeling:

- Handling missing values.

- Feature engineering.

- Encoding categorical variables.

- Data normalization or scaling.

Include code snippets or explanations for each major preprocessing step.

# 5. Modeling

\*\*Model Selection:\*\* List the models you tried (e.g., Logistic Regression, Decision Trees, etc.).

\*\*Model Training:\*\* Explain how the data was split into training and testing sets.

\*\*Hyperparameter Tuning:\*\* Describe the tuning process if applicable.

\*\*Evaluation Metrics:\*\* Mention metrics used to evaluate model performance.

# 6. Results

Present the performance of the models you tested:

- Include a table comparing metrics for each model.

- Explain which model performed best and why.

# 7. Conclusion

Summarize your findings and the overall performance of the model.

Discuss the strengths and limitations of your approach.

Mention potential improvements.

# 8. Future Work

Highlight opportunities for extending the project:

- Implementing neural networks.

- Using different datasets.

- Deploying the model as a web application or API.

# 9. Appendix

\*\*Code:\*\* Include or link to your project code.

\*\*References:\*\* Cite any external resources, datasets, or tutorials used.

\*\*Additional Visuals:\*\* Include additional charts or tables if necessary.