Predecting The survival Of Titanic Passengers

Ayhem Belkhamsa
Dept. of EE
ISET Bizerte — Tunisia
ayhem-b

***Abstract* — This project explores machine learning techniques to predict Titanic passenger survival using demographic and socioeconomic data. Logistic regression, K-Nearest Neighbors (KNN), and K-means clustering were applied to analyze the dataset. After preprocessing and feature engineering, the models were evaluated using accuracy and F1-score.**

Results showed that [mention the best-performing model] provided the most accurate predictions. This study highlights the effectiveness of machine learning in predictive analytics and decision-making.

I. Introduction

The Titanic disaster of 1912 remains one of history's most infamous maritime tragedies. This project aims to predict passenger survival using machine learning models based on demographic and socioeconomic data from the Titanic dataset. By employing logistic regression, K-Nearest Neighbors (KNN), and K-means clustering, we analyze key factors influencing survival outcomes. The study highlights the potential of predictive analytics in understanding historical events and making data-driven decision

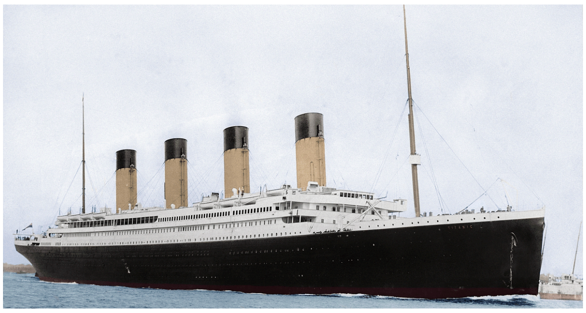


Figure 1: Titanic ship

II. Coding Part

1. *Importing librarys*

```
import numpy as np
np.set_printoptions(precision=3)
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

2. *Importing the data*

```
train_df = pd.read_csv('../titanic_Data/train.csv')
```

3. *Cleaning & Data preparation*

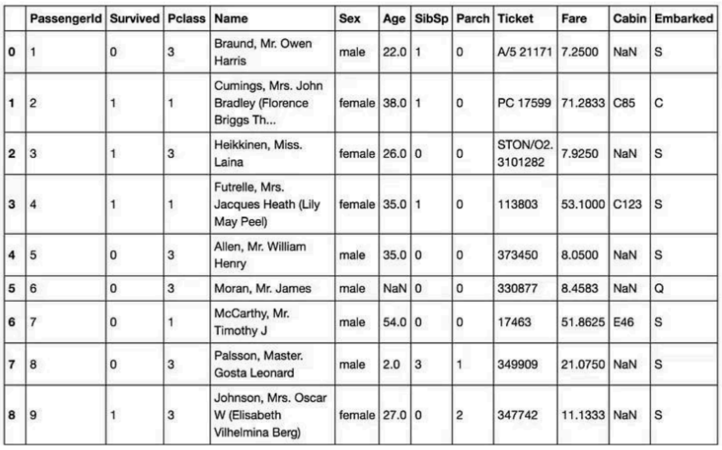
1) *Data info:*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

by looking to the output above we can know that the **Training Data** has

- 891 examples
- 10 features
- 1 target (survived)
- some missing datain **Age** , **Cabin** & **Embarked**

```
train_df.head(9)
```



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S

Figure 2: First 9 examples

2) *removing unnecessary columns & filling missing data:*

so in order to know what coluns has effect in the predection model
we need to know what each one refers to

we are not going to use :

- PassengerId
- Ticket
- Cabin (it has a lot of missing data)