**The Challenge**

**Titanic** is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the **Titanic** sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (i.e. name, age, gender, socio-economic class, etc).



**The predictive analytics process**

1. **Problem understanding and definition:**understand the problem and how the potential solution would look. Also, define the requirements for solving the problem
2. **Data collection and preparation:** get a dataset that is ready for analysis
3. **Data understanding using Exploratory Data Analysis (EDA):** understand your dataset
4. **Feature Engineering and Data Processing:** process of using raw data to create features that will be used for predictive modeling.
5. **Model building:** produce some predictive models that solve the problem
6. **Model evaluation:** choose the best model among a subset of the most promising ones and determine how good the model is in providing the solution
7. **Communication and/or deployment:** use the predictive model and its results
8. **Deployement:** Submitting entire Project on Kaggle

**1. Problem understanding and definition**

In this challenge, we need to complete the **analysis** of what sorts of people were most likely to **survive**. In particular, we apply the tools of **machine learning** to predict which passengers survived the tragedy.

Predict whether passenger will **survive or not**(I know, it has a morbid connotation).

### Loading the data files

Here we import the data. For this analysis, we will be exclusively working with the Training set. We will be validating based on data from the training set as well. For our final submissions, we will make predictions based on the test set.

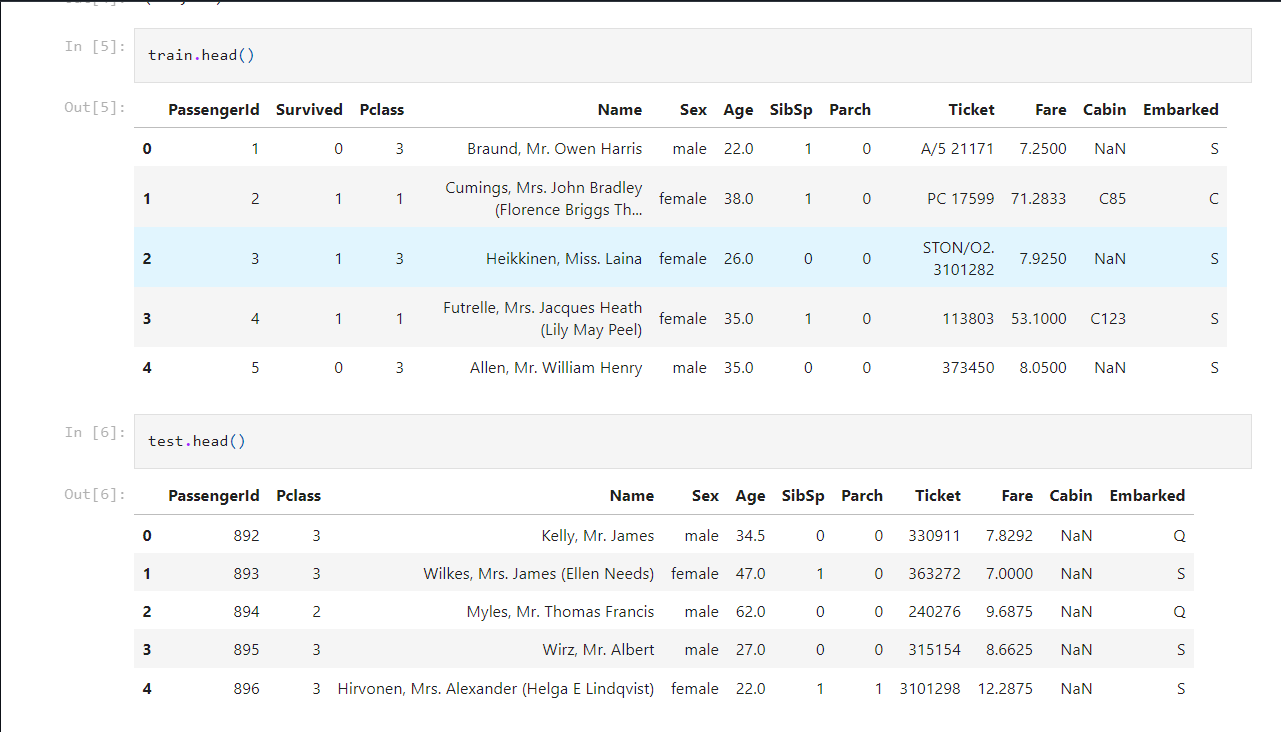
### Data Description

The data has been split into two groups:

* training set (train1.csv)
* test set (test1.csv)

The training set includes passengers survival status (also know as the ground truth from the titanic tragedy) which along with other features like gender, class, fare and pclass (passenger class) is used to create the machine learning model.

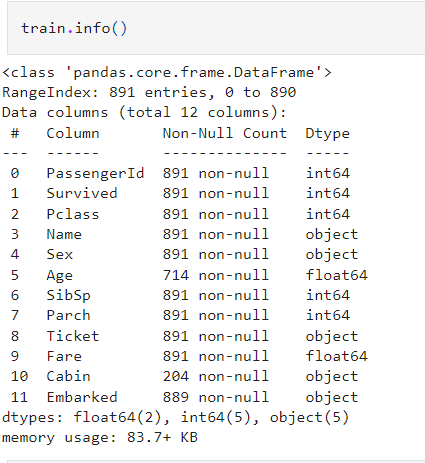
The test set should be used to see how well the model performs on unseen data. The test set does not provide passengers survival status. We are going to use our model to predict passenger survival status.



## 3. Data understanding using Exploratory Data Analysis (EDA)

**Exploratory Data Analysis** refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations [1].

In summary, it's an approach to analyzing data sets to summarize their main characteristics, often with visual methods.



The training-set has 891 rows and 11 features + the **target variable (survived).** 2 of the features are floats, 5 are integers and 5 are objects.

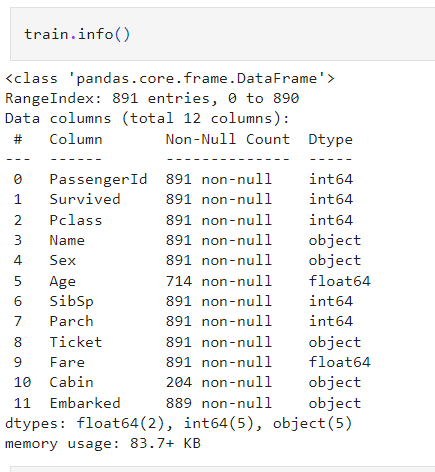
### Conclusions from .describe() method

**.describe()** gives an understanding of the central tendencies of the numeric data.

* Above we can see that **38% out of the training-set survived the Titanic.**
* We can also see that the passenger age range from **0.4 to 80 years old.**
* We can already detect some features that contain **missing values**, like the ‘Age’ feature (714 out of 891 total).
* There's an **outlier** for the 'Fare' price because of the differences between the 75th percentile, standard deviation, and the max value (512). We might want to drop that value.



### Exploring missing data



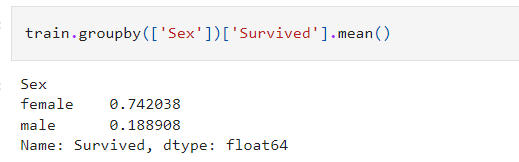
The **'Embarked'** feature has only 2 missing values, which can easily be filled or dropped. It will be much more tricky to deal with the **‘Age’** feature, which has 177 missing values. The **‘Cabin’** feature needs further investigation, but it looks like that we might want to drop it from the dataset since 70% is missing.

**"The captain goes down with the ship"** is a maritime tradition that a sea captain holds ultimate responsibility for both his/her ship and everyone embarked on it, and that in an emergency, he/she will either save them or die trying.

In this case, **Captain Edward Gifford Crosby** went down with Titanic in a heroic gesture trying to save the passengers.

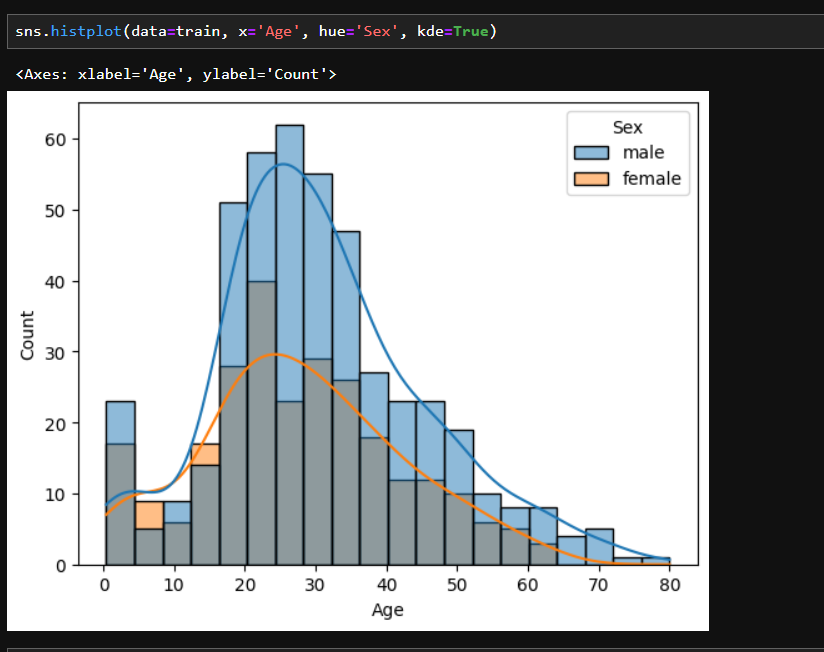


**74% of the women survived, while only 18% of men survived.**



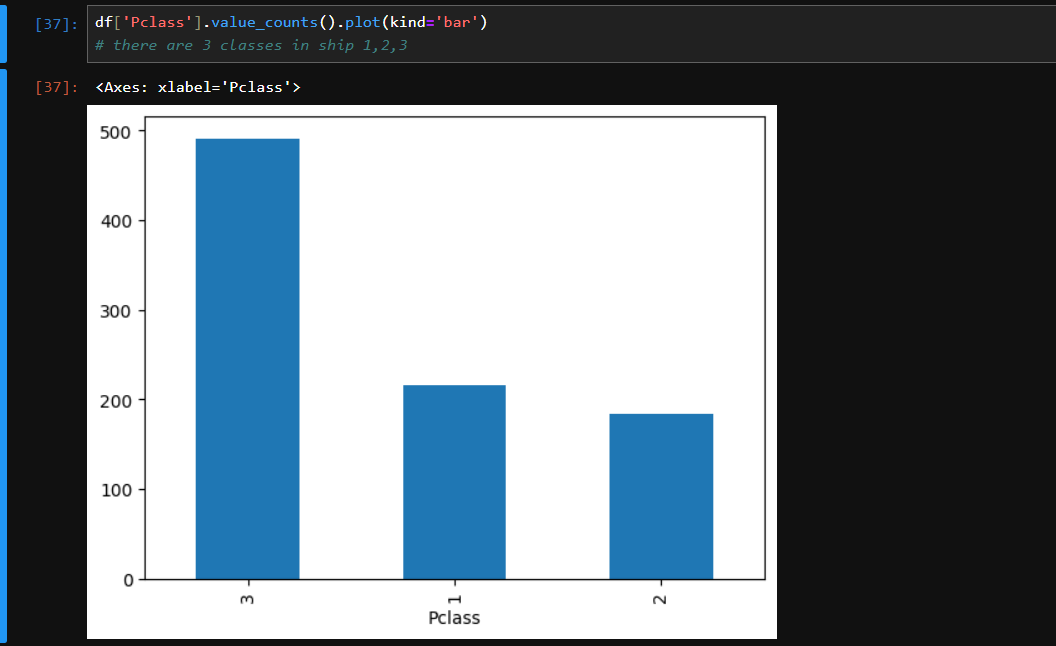
We can see that **men** have a higher probability of survival when they are between **18 and 35 years old.** For **women,** the survival chances are higher between **15 and 40 years old.**

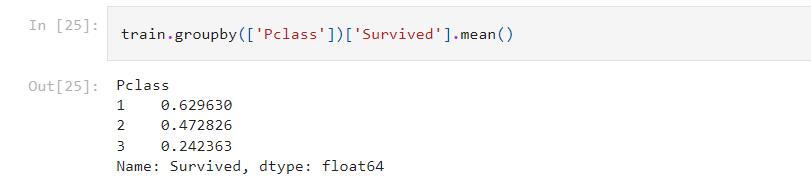
For men the probability of survival is very low between the **ages of 5 and 18**, and **after 35**, but that isn’t true for women. Another thing to note is that **infants have a higher probability of survival.**



**Children below 18 years of age** have higher chances of surviving.

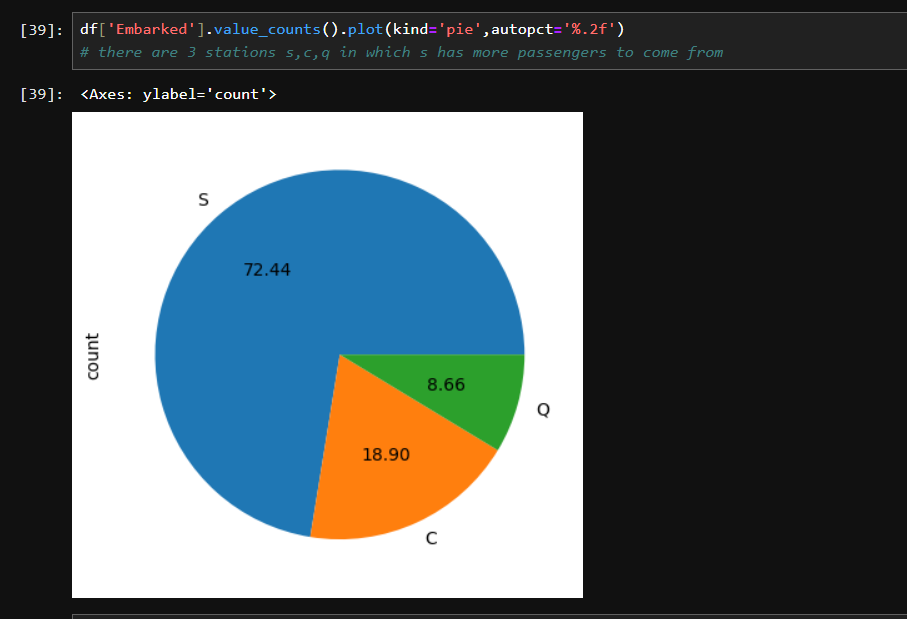
### Passenger class distribution; Survived vs Non-Survived





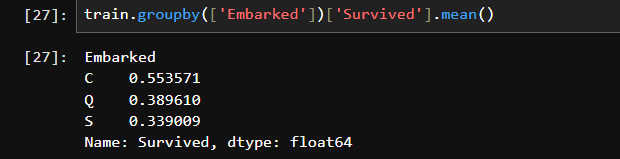
The graphs above clearly shows that **economic status (Pclass)** played an important role regarding the potential survival of the Titanic passengers. First class passengers had a much higher chance of survival than passengers in the 3rd class. We note that:

* 63% of the 1st class passengers survived the Titanic wreck
* 47% of the 2nd class passenger survived
* Only 24% of the 3rd class passengers survived



**1. Southampton (S) is the most popular embarkation location, with more than 70% of passengers boarding there.**

**2. Cherbourg (C) has a higher chance of survival than Southampton or Queenstown, with 55% of passengers who embarked in Cherbourg surviving compared to 34% and 39% at Southampton and Queenstown respectively.**



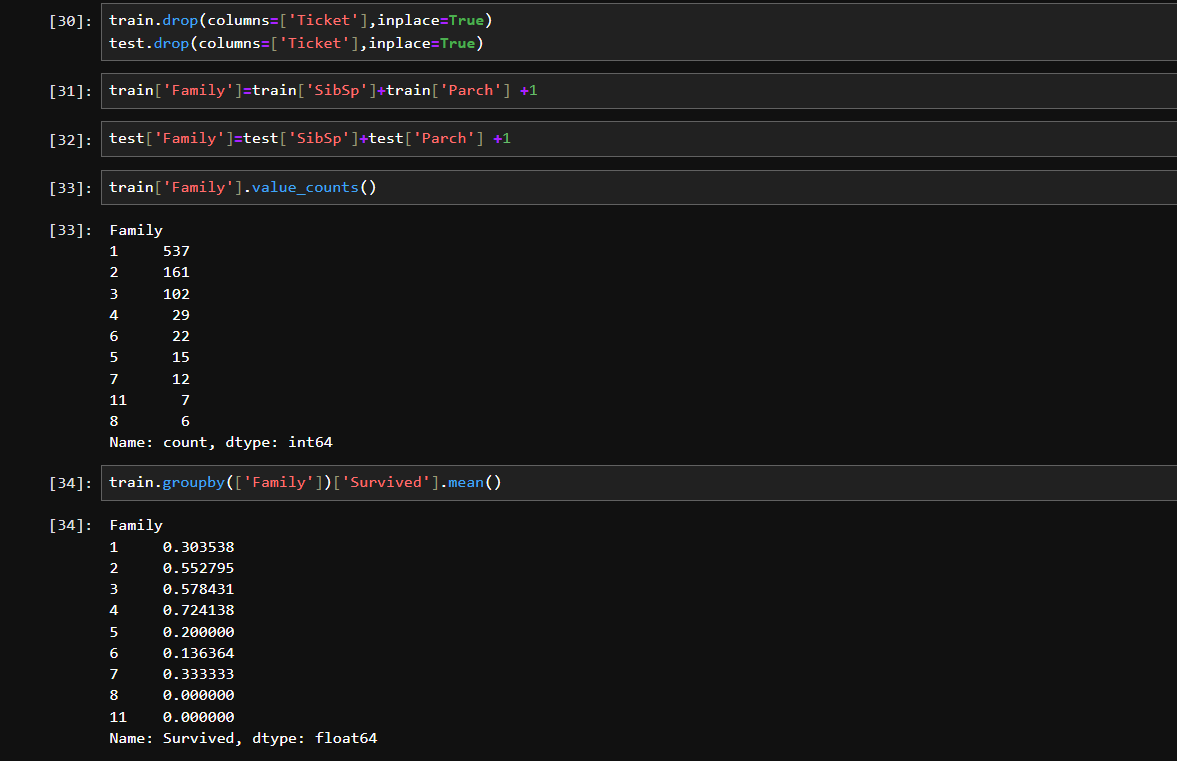
## 4. Feature Engineering and Data Processing

### Drop 'PassengerId'

First, I will drop ‘PassengerId’ from the train set, because it does not contribute to a persons' survival probability.

### Combining SibSp and Parch

SibSp and Parch would make more sense as a combined feature that shows the total number of relatives a person has on the Titanic. I will create the new feature 'relative' below, and also a value that shows if someone is not alone.





### Missing Data

As a reminder, we have to deal with **Cabin (687 missing values), Embarked (2 missing values)** and **Age (177 missing values).**

### Cabin

We will drop it as it contains more than 60 percent of missing values

### Age

As seen previously on **"Exploring missing data"**, there are a lot of missing 'Age' values (177 data points). We can normalize the 'Age' feature by creating an array that contains normalised random numbers, which are computed based on the mean age value in regards to the standard deviation and is\_null.

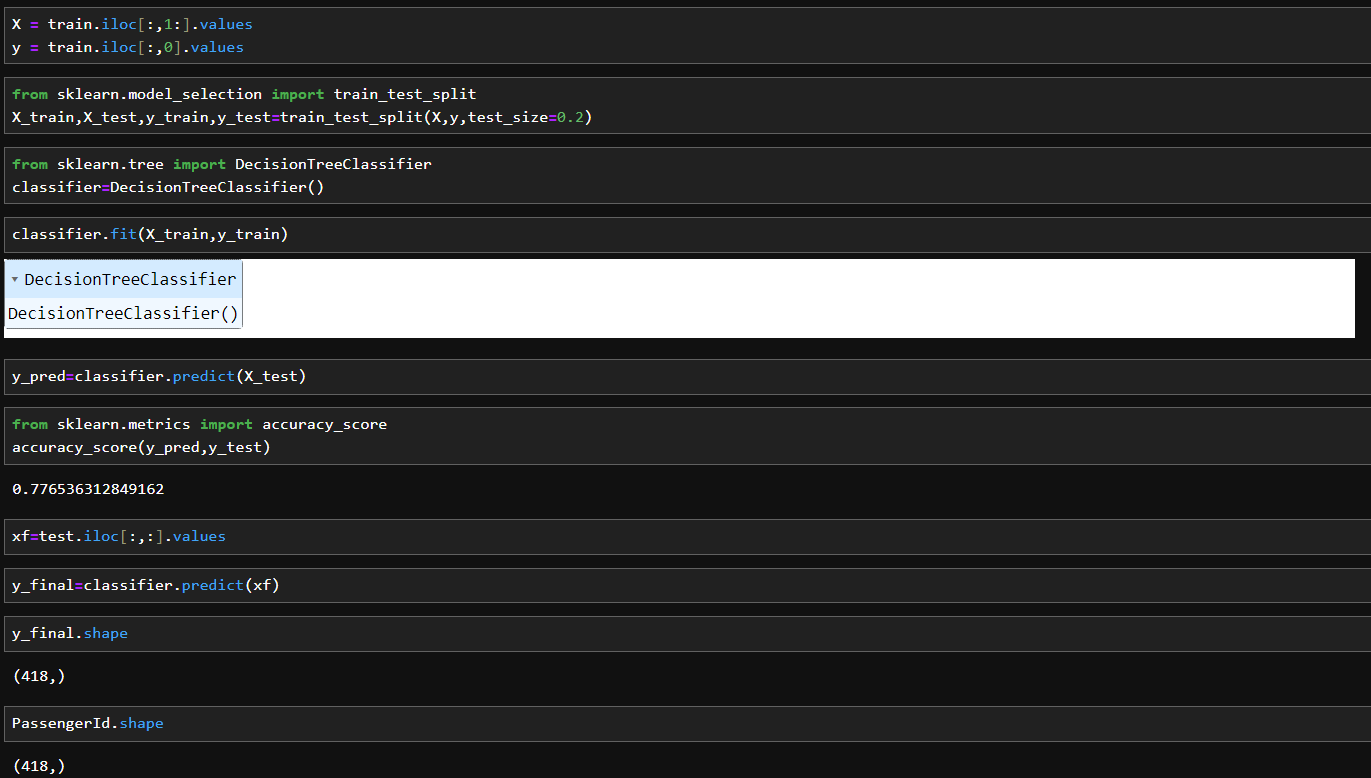
### Embarked

Since the Embarked feature has only 2 missing values, we will fill these with the most common one (S).

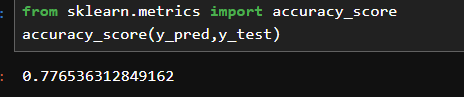
## 5. Model building

I will be using most popular Machine Learning model in Data Science. I won't dive too deep on specific characteristics of each one of the models. The models is:

**Decision Tree**

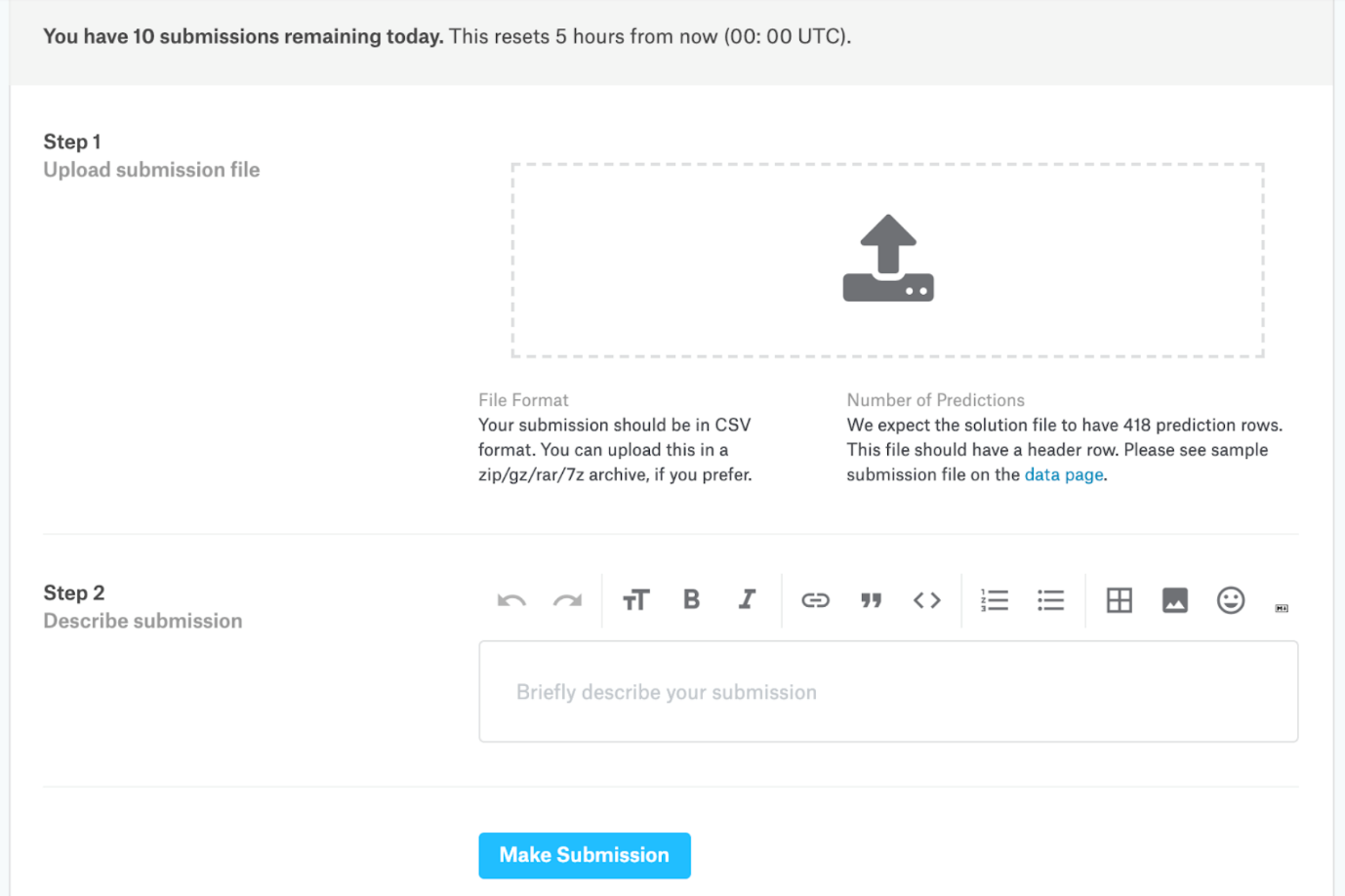


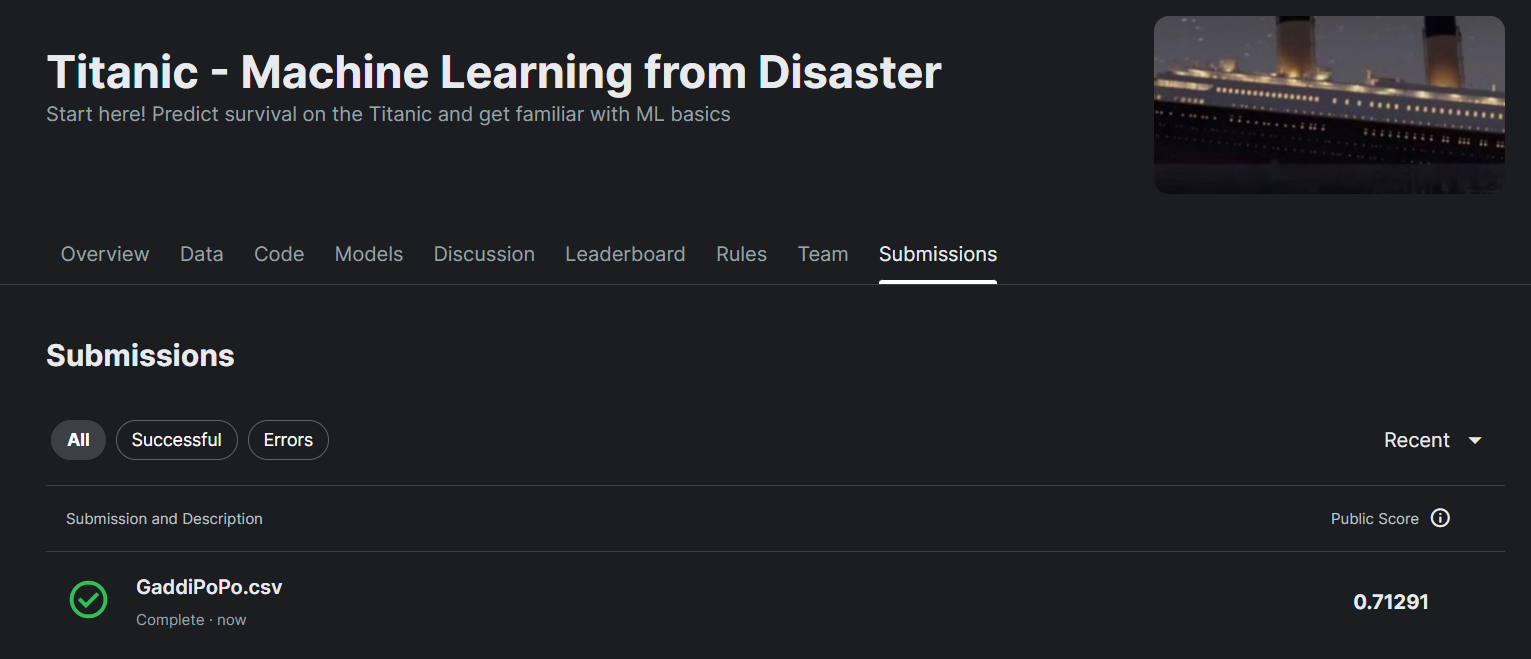
## 6. Model evaluation



Converting to csv file and submitting on Kaggle







### **Conclusion**

I'm really glad I got to do another Data Science project as I learned valuable concepts that can be transferable to other projects. It significantly deepened my machine learning knowledge and strengthened my problem-solving and analytical skills, as well as applying concepts learned from textbooks, articles, classroom, and various other sources, on a challenging problem. Looking forward to keep learning new things and tacking new projects in my Data Science journey!

## Sources

Github-https://github.com/GaddiPoPo/Titanic-Dataset-/tree/main/Titanic%20Full%20Project

Kaggle-https://www.kaggle.com/competitions/titanic