# Data Collection And Preprocessing

## Abou the Titanic dataset

The Titanic dataset is a well-known dataset used in machine learning for classification tasks. The goal is to predict whether a passenger survived or not based on various features. It comes from the passenger manifest of the Titanic ship and includes several features (columns) with labels corresponding to survival status.

Features:

1. PassengerId: Unique identifier for each passenger (not particularly useful for modeling).

2. Survived: Target label (0 = No, 1 = Yes) indicating whether the passenger survived.

3. Pclass: Ticket class (1st = 1, 2nd = 2, 3rd = 3), representing the socioeconomic status of the passengers.

4. Name: The full name of the passenger, which can sometimes be parsed for additional information (e.g., titles like Mr., Mrs.).

5. Sex: Gender of the passenger (male, female).

6. Age: Age of the passenger in years. Some values are missing and may require imputation.

7. SibSp: Number of siblings or spouses aboard the Titanic.

8. Parch: Number of parents or children aboard the Titanic.

9. Ticket: Ticket number (can be useful in some advanced feature engineering, but often not directly valuable).

10. Fare: The fare paid by the passenger.

11. Cabin: Cabin number (many missing values, but can potentially indicate passenger class or location on the ship).

12. Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton), representing the departure point.

Label:

Survived: This is the target variable, indicating whether the passenger survived or not (1 = Survived, 0 = Did not survive).

Features that Add Value:

1. Pclass: There is a strong correlation between the ticket class and survival, with higherclass passengers having better survival rates.

2. Sex: Gender plays a significant role; women were more likely to survive due to the "women and children first" policy.

3. Age: Younger passengers were often given preference during evacuation, so age is a crucial factor.

4. Fare: Higher farepaying passengers (typically in higher classes) were more likely to survive.

5. Embarked: The port of embarkation can offer insight, as passengers from different ports might belong to different socioeconomic classes.

6. SibSp and Parch: Having family members aboard can affect survival probabilities; passengers with more family members may have been more likely to survive.

These key features often contribute to building a successful prediction model for the Titanic dataset.

**url = "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv"**