Group 05-Final project proposal

Course：Introduction to Machine Learning
Topics：Titanic: Machine Learning from Disaster (binary classification)
Group members：112550112陳景寬、112550189鍾承翰、113511248林昌岳

---

1. Dataset and Task Description

Source and overview ：

The dataset used in this project is **"Titanic: Machine Learning from Disaster"**, available on the Kaggle platform. It provides demographic and travel information for a subset of passengers aboard the RMS Titanic.

Learning task ：

The goal is to solve a **binary classification** problem. We want to predict whether a passenger **survived (Survived = 1)** or **did not survive (Survived = 0)** based on their personal and travel characteristics such as age, gender, and passenger class.

Data characteristics :

- **Type :** Tabular data.
- **Size :** The training set contains **891 samples**, and the test set contains **418 samples**.
- **Features:**
  The dataset includes variables : **Pclass** (Passenger class)、**Sex** (Gender)、**Age** (Age in years)、**SibSp** (Number of siblings/spouses aboard)、**Parch** (Number of parents/children aboard)、**Ticket** (Ticket number)、**Fare** (Ticket fare)、**Cabin** (Cabin number) and **Embarked** (Port of embarkation)

2. Preprocessing & Feature Engineering

1. **Handle missing values :**

   Several features contain missing values, including **'Age'**, **'Cabin'**, and **'Embarked'**. For **'Age'** and **'Embarked'**, we will evaluate filling missing values using measures such as the **median**, **mean**, or **mode**.

   Since **'Cabin'** has a high proportion of missing data, we may consider **dropping** this feature or extracting partial information (e.g., cabin prefix).

2. **Feature transformation :**
   - Convert categorical features such as **'Sex'** and **'Embarked'** into numerical form using one-hot encoding.

    - Extract titles (e.g., Mr., Mrs., Miss) from the **‘Name’** feature to create a new feature **‘Title’**, which may reflect social status and survival likelihood.
    - Combine **‘SibSp’** and **‘Parch’** into a new feature **‘FamilySize’** to represent family presence aboard.

3. **Normalize the data:**

   Apply **standardization** or **normalization** to numerical variables such as **‘Age’** and **‘Fare’** to prevent models from being biased by differing feature scales.

3. Model Implementation & Comparison

To meet the project requirements, we plan to implement and compare at least **three machine learning algorithms**:

1. **Logistic Regression** – a baseline linear classification model.
2. **Decision Tree / Random Forest** – to evaluate the performance of tree-based models.
3. **Support Vector Machine (SVM)** or **K-Nearest Neighbors (KNN)** – to explore alternative classification approaches.

Evaluation Metrics

Since this is a classification task, we will use the following metrics:

- **Accuracy** – as the primary performance indicator.
- **F1-score** – to balance precision and recall for imbalanced outcomes.
- **Confusion Matrix** – to visualize model predictions and analyze classification errors in more detail.