Họ tên: Phạm Minh Dương  
Mã số SV: 3123410065  
Môn học: Nhập môn máy học

Bài: Titanic – Machine Learning from Disaster (Kaggle)

# 1. Giới thiệu

Bài thực hành này sử dụng dữ liệu Titanic từ Kaggle để làm quen với quá trình xử lý dữ liệu, huấn luyện mô hình và đánh giá kết quả. Mục tiêu là dự đoán khả năng sống sót của hành khách dựa trên các đặc trưng như hạng vé, giới tính, tuổi, số người đi cùng, giá vé, và cảng lên tàu.

# 2. Dữ liệu Titanic

**Tổng quan:** 2 file CSV:

* train.csv (có nhãn Survived) — dùng để học.
* test.csv (không có Survived) — dùng để dự đoán và tạo submission.csv.

**Các cột chính :**

* **Pclass** (1/2/3) — proxy cho tầng lớp xã hội; thường quan trọng (hạng cao có tỉ lệ sống cao hơn).
* **Sex** (male/female) — rất quan trọng; thường chuyển thành binary (0/1).
* **Age** (số thực, có missing) — quan trọng nhưng nhiều giá trị thiếu → cần impute.
* **SibSp, Parch** — kích thước gia đình/quan hệ; có thể dùng riêng hoặc ghép thành FamilySize = SibSp + Parch + 1.
* **Fare** (giá vé, lệch phải) — liên quan tới Pclass; thường log-transform để ổn định phân phối.
* **Embarked** (C/Q/S) — thông tin vị trí lên tàu; có thể ảnh hưởng (một số cảng nhiều hành khách giàu hơn).
* **Name, Ticket, Cabin, PassengerId** — không phải feature trực tiếp:
  + Name có thể **trích Title** (Mr/Mrs/Miss) — feature mạnh.
  + Cabin có nhiều missing (>70%) — thường drop, hoặc tạo flag HasCabin.
  + Ticket thường lộn xộn nhưng có thể group theo prefix.
  + PassengerId chỉ dùng để tạo file submission, không dùng làm feature.

# 3. Tiền xử lý

- Bổ sung giá trị thiếu (Age, Fare, Embarked)  
- Loại bỏ các cột không cần thiết (Cabin, Ticket, Name, PassengerId)  
- Chuyển đổi dữ liệu dạng chữ (Sex, Embarked) sang dạng số  
- Chuẩn hóa các cột số (Age, Fare)

# 4. Huấn luyện mô hình

Sử dụng 3 mô hình:

* Logistic Regression
* Random Forest
* XGBoost

Kết quả trên tập validation cho thấy **Random Forest** hoạt động tốt nhất.

# 5. Kết quả

Sau khi tiền xử lý dữ liệu và huấn luyện mô hình, tiến hành nộp file dự đoán (submission.csv) lên Kaggle.  
Kết quả đạt được:

