Báo cáo kết quả:dự đoán sống sót trong thảm hoạ Titanic

Họ &tên: Lê Văn Thông

Mssv:3123410362

1. Giới thiệu

Bài toán Titanic là một trong những competition nổi tiếng trên Kaggle.  
Mục tiêu: dự đoán hành khách có sống sót hay không dựa trên thông tin cá nhân (tuổi, giới tính, vé, hạng ghế...).

2. Dữ liệu

train.csv: chứa dữ liệu huấn luyện (891 dòng, 12 cột), có nhãn Survived.

test.csv: chứa dữ liệu test (418 dòng), không có nhãn Survived.

Các đặc trưng quan trọng:

+Pclass: hạng ghế (1, 2, 3).

+Sex: giới tính.

+Age: tuổi.

+SibSp: số anh/chị/em/vợ/chồng đi cùng.

+Parch: số cha/mẹ/con đi cùng.

+Fare: giá vé.

+Embarked: cảng lên tàu.

3. Tiền xử lý dữ liệu

1. Xử lý missing values

Embarked: thay bằng giá trị phổ biến nhất.

Age, Fare: thay bằng giá trị trung bình.

Cabin: bị thiếu quá nhiều => loại bỏ.

1. Loại bỏ cột không cần thiết

Ticket, Name, PassengerId: không ảnh hưởng đến dự đoán.

1. Mã hóa dữ liệu (Encoding)

Sex: chuyển thành số (male=0, female=1).

Embarked: one-hot encoding.

1. Chuẩn hóa dữ liệu (Scaling)

Áp dụng StandardScaler cho Age, Fare.

Áp dụng log-transform cho Fare để giảm outlier.

4. Huấn luyện và Kết quả

* Chia tập train thành train (80%) và validation (20%) để đánh giá.
* So sánh 3 mô hình:

1.Logistic Regression

2.Random Forest

3.XGBoost

Kết quả trên tập validation:

Logistic Regression: Accuracy ~ 0.80

Random Forest: Accuracy ~ 0.83

XGBoost: Accuracy ~ 0.82

5. Kết luận

* Các bước xử lý dữ liệu (điền missing, encode, scale) là bắt buộc để mô hình hoạt động.
* Random Forest cho kết quả tốt nhất trong thử nghiệm.
* Có thể cải thiện thêm bằng: feature engineering (FamilySize, Title từ Name), tuning hyper-parameters, ensemble nhiều model.