**Họ tên:** Giang Hào Tường

**MSSV:** 3123410421

**LAB\_03:** LÀM QUEN VỚI KỲ THI KAGGLE

**1. Giới thiệu tổng quan**

Bài toán Titanic là một trong những bài toán cơ bản trong Machine Learning, với mục tiêu dự đoán khả năng sống sót của hành khách trên chuyến tàu Titanic dựa vào các thuộc tính (tuổi, giới tính, hạng vé, số người đi cùng...).

Báo cáo này sẽ trình bày về hiệu suất dự đoán của các mô hình máy học để dự đoán bài toán phân loại Titanic, bộ dữ liệu Titanic từ Kaggle được sử dụng để thực hiện phân tích. Các mô hình máy học được sử dụng bao gồm: Hồi quy Logistic (Logistic Regression), Rừng ngẫu nhiên (Random Forest) và XGBoost (XGBoost Classifier). Các mô hình được huấn luyện và đánh giá dựa trên các chỉ số chính: Độ chính xác (Accuracy), điểm F1 (F1 Score) và ROC AUC (ROC Area Under Curve).

**2. Các chỉ số đánh giá**

- Độ chính xác (Accuracy): Tỉ lệ các dự đoán đúng trên tổng số dự đoán

- Điểm F1 (F1 score): Trung bình điều hòa của độ chính xác (Precision) và độ thu hồi (Recall). F1 Score đặc biệt hữu ích khi các lớp dữ liệu không cân bằng.

- ROC AUC (ROC Area Under Curve): Đo lường khả năng phân biệt giữa các lớp của mô hình. Giá trị càng gần 1, mô hình càng tốt.

**3. Kết quả đánh giá**

- Hồi quy Logistic (Logistic Regression)

A black text with black letters

AI-generated content may be incorrect.

Mô hình cho độ chính xác khá tốt, cân bằng giữa Precision và Recall. Chỉ số ROC AUC đạt 0.8387 cho thấy mô hình có khả năng phân loại khá tốt giữa các lớp.

- Rừng ngẫu nhiên (Random Forest)

A number of numbers on a white background

AI-generated content may be incorrect.

Mô hình Random Forest cho hiệu suất cao nhất về Accuracy và F1 Score trong số 3 mô hình được đánh giá.

- XGBoost (XGBoost Classifier)

A screenshot of a computer screen

AI-generated content may be incorrect.

Mô hình XGBoost cho kết quả thấp hơn 2 mô hình còn lại, đặc biệt ROC AUC thấp nhất. Mặc dù vẫn đạt độ chính xác gần 80% nhưng hiệu suất thấp nhất trong 3 mô hình.

**4. Nhận xét**

Dựa trên kết quả phân tích, mô hình Random Forest là mô hình có hiệu suất tốt nhất với Accuracy và F1 Score cao nhất. Mô hình này là lựa chọn tối ưu để sử dụng trong bài toán này. Trong tương lai, việc tinh chỉnh các tham số cho mô hình có thể giúp cải thiện hiệu suất hơn nữa

Kết quả khi chạy thực tế trên cuộc thi Titanic là 0.7488 và xếp hạng trên leaderboard là 11793

