**REPORT LAB 03 – MACHINE LEARNING**

Họ và tên: Nguyễn Dương Bảo Trân

Mã SV: 3123410386

A screenshot of a computer

AI-generated content may be incorrect.

Dữ liệu Titanic được cung cấp dưới dạng file train.csv và test.csv.

- PassengerId: Mã số hành khách (dùng để phân biệt từng người, không phải feature dự đoán).

- Survived: Nhãn mục tiêu (0 = chết, 1 = sống).

- Pclass: Hạng vé (1, 2, 3).

- Name: Tên hành khách (không dùng trực tiếp làm feature, có thể trích Title).

- Sex: Giới tính (male/female).

- Age: Tuổi (có giá trị missing).

- SibSp: Số anh/chị/em hoặc vợ/chồng đi cùng.

- Parch: Số cha/mẹ/con đi cùng.

- Ticket: Số vé (thường bỏ hoặc dùng để trích feature).

- Fare: Giá vé.

- Cabin: Số phòng (nhiều missing).

- Embarked: Cảng lên tàu (C = Cherbourg, Q = Queenstown, S = Southampton).

A screenshot of a computer

AI-generated content may be incorrect.

Dữ liệu test được cung cấp trong file test.csv, dùng để dự đoán nhãn Survived cho các hành khách.

Các cột trong dữ liệu test giống với train.csv, nhưng **không có cột Survived.**

A screenshot of a computer

AI-generated content may be incorrect.

Dataset train.csv (891 bản ghi, 12 cột):

* **Numerical:** PassengerId, Survived (target), Pclass, Age (177 missing), SibSp, Parch, Fare
* **Categorical (object):** Name, Sex, Ticket, Cabin (687 missing), Embarked (2 missing)

**Tiền xử lý:**

* Xử lý missing: Age, Cabin
* Encode categorical: Sex, Embarked (label hoặc one-hot)

A screenshot of a computer

AI-generated content may be incorrect.

Sau khi kiểm tra dữ liệu, thấy cột Age có **177 giá trị thiếu**, cột Cabin **thiếu 687 giá trị**, và cột Embarked **thiếu 2 giá trị**. Các cột còn lại đều đầy đủ dữ liệu. Trong xử lý dữ liệu, có thể điền trung bình/median cho Age, bỏ hoặc điền giá trị phổ biến cho Cabin và Embarked.

A screenshot of a data table

AI-generated content may be incorrect.

Sau khi làm sạch dữ liệu và chọn lọc các đặc trưng quan trọng, tập dữ liệu huấn luyện gồm 891 quan sát với 8 biến. Biến **Survived** là nhãn mục tiêu, thể hiện tình trạng sống sót của hành khách (0 = chết, 1 = sống). Các biến **Pclass**, **Sex**, **Age**, **SibSp**, **Parch**, **Fare** và **Embarked** được sử dụng làm đặc trưng đầu vào để dự đoán khả năng sống sót. Trong đó, **Age** và **Embarked** có các giá trị thiếu đã được xử lý, còn **Sex** và **Embarked** được mã hóa dưới dạng số hoặc nhãn để phù hợp với mô hình học máy. Tập dữ liệu này là cơ sở để huấn luyện và đánh giá hiệu suất các mô hình dự đoán.

A screenshot of a computer

AI-generated content may be incorrect.

Trên tập dữ liệu **test\_data**, các biến **Pclass**, **Sex**, **SibSp**, **Parch** và **Embarked** không có giá trị thiếu, trong khi **Age** còn 86 giá trị thiếu và **Fare** còn 1 giá trị thiếu. Những giá trị này cần được xử lý trước khi sử dụng tập dữ liệu để dự đoán kết quả sống sót của hành khách bằng mô hình.

A screenshot of a computer code

AI-generated content may be incorrect.

Sau khi tiền xử lý dữ liệu Titanic, cả tập huấn luyện (**train\_data**) và tập kiểm tra (**test\_data**) đều không còn giá trị thiếu, với tổng số giá trị missing bằng 0 ở mỗi tập. Trong tập huấn luyện, các biến kiểu đối tượng còn lại là **Sex** và **Embarked**, cần được mã hóa sang dạng số để phục vụ việc huấn luyện mô hình dự đoán khả năng sống sót của hành khách.

A screenshot of a computer code

AI-generated content may be incorrect.

Sau khi mã hóa các biến phân loại, tập huấn luyện **train\_data** không còn cột kiểu đối tượng (**object**). Tiếp theo, phân tích biến **Fare** bằng biểu đồ hộp (boxplot) cho thấy sự phân bố của giá vé trên các hành khách. Biểu đồ giúp nhận diện các giá trị ngoại lai (outliers) có thể ảnh hưởng đến mô hình dự đoán khả năng sống sót.

A graph with numbers and circles

AI-generated content may be incorrect.

*Hình 1 minh họa boxplot của biến Fare.*

A screenshot of a computer

AI-generated content may be incorrect.

Để dự đoán khả năng sống sót của hành khách, ba mô hình học máy đã được huấn luyện và đánh giá trên tập xác thực (**validation set**): **Logistic Regression**, **Random Forest** và **XGBoost**. Các chỉ số hiệu suất được sử dụng gồm **Accuracy**, **F1 Score** và **ROC AUC**.

→ Trong ba mô hình, **Random Forest** cho Accuracy và F1 Score cao nhất, trong khi **Logistic Regression** có ROC AUC tốt nhất, cho thấy mô hình này dự đoán phân loại cân bằng giữa các lớp hơi nhỉnh hơn. Các kết quả này giúp lựa chọn mô hình phù hợp để dự đoán khả năng sống sót của hành khách trên dữ liệu thực tế.

A screenshot of a computer program

AI-generated content may be incorrect.

Sau khi so sánh hiệu suất các mô hình, **Random Forest** được chọn là mô hình tốt nhất để dự đoán khả năng sống sót. Mô hình được huấn luyện trên tập huấn luyện (**X\_train, y\_train**) và dự đoán trên tập xác thực (**X\_val**).

→ Kết quả này cho thấy mô hình dự đoán khá sát với nhãn thực tế, mặc dù vẫn có một số trường hợp dự đoán sai. Phân tích này giúp đánh giá trực quan khả năng dự đoán của mô hình trước khi áp dụng trên dữ liệu thực tế.

A screenshot of a computer program

AI-generated content may be incorrect.

File này được lưu dưới tên **submission.csv**, sẵn sàng để nộp lên Kaggle. Kết quả đầu ra thông báo rằng file submission đã được tạo thành công.