Predicting Passenger Survival on the Titanic Using Machine Learning
Dự đoán khả năng sống sót của hành khách trên tàu Titanic bằng học máy

Phan Trần Hữu Tấn, Trần Quốc Hoàng, Trần Chí Vỹ

INTRODUCTION

Trong cuộc thi này, bạn sẽ được cung cấp **hai bộ dữ liệu tương tự nhau** chứa thông tin của hành khách như **tên, tuổi, giới tính, tầng lớp xã hội – kinh tế**, v.v. Một bộ dữ liệu có tên là **train.csv**, còn bộ kia là **test.csv**.
train.csv chứa thông tin chi tiết của **một phần hành khách trên tàu (chính xác là 891 người)** và quan trọng hơn, nó cho biết **liệu họ có sống sót hay không** — đây được gọi là **"ground truth" (sự thật gốc)**.
test.csv chứa thông tin tương tự nhưng **không tiết lộ** liệu từng hành khách có sống sót hay không.
Nhiệm vụ của bạn là **dựa trên các mẫu (patterns)** rút ra từ dữ liệu trong **train.csv**, để **dự đoán khả năng sống sót của 418 hành khách còn lại** (có trong **test.csv**).

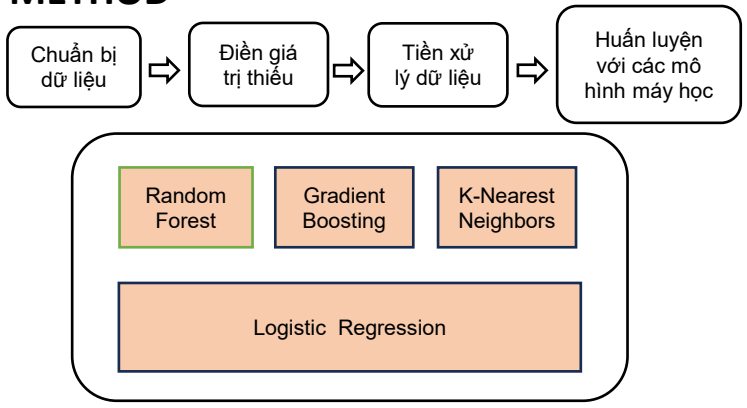
Mục tiêu:
Nhiệm vụ của bạn là **dự đoán xem một hành khách có sống sót sau vụ đắm tàu Titanic hay không**. Đối với mỗi hành khách trong tập kiểm tra, bạn cần **dự đoán giá trị 0 hoặc 1** cho biến "Survived" (0 = không sống sót, 1 = sống sót).

Thước đo đánh giá:
Điểm số của bạn là **tỷ lệ phần trăm số hành khách được dự đoán đúng**, còn được gọi là **độ chính xác (accuracy)**.

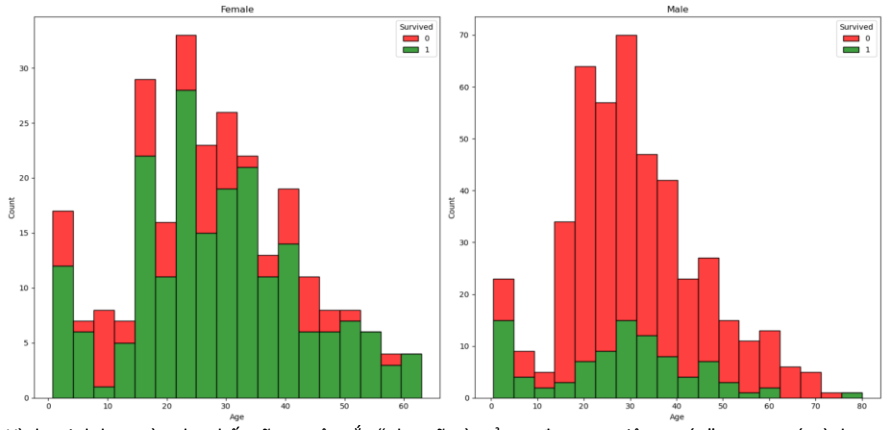
DATASET

Số lượng dữ liệu:
Gồm **891 mẫu trong tập huấn luyện (training set)** và **418 mẫu trong tập kiểm tra (test set)**. Data source: Kaggle Titanic - Machine Learning from Disaster | Kaggle

METHOD



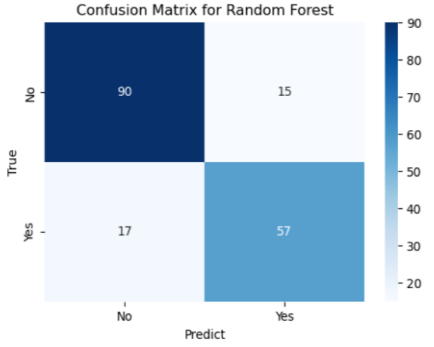
VISUALIZING



Hình minh họa này cho thấy rõ nguyên tắc "phụ nữ và trẻ em được ưu tiên trước" trong quá trình sơ tán, đồng thời thể hiện rằng địa vị kinh tế – xã hội (hạng vé) và giới tính có ảnh hưởng mạnh đến khả năng sống sót.

MACHINE LEARNING MODELS AND RESULTS

1.Random Forest



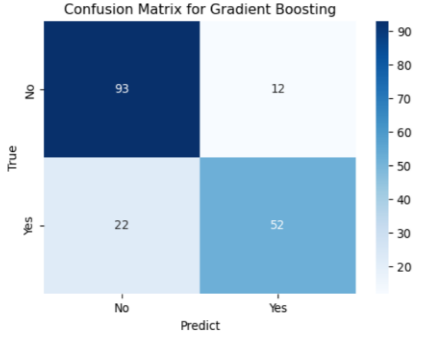
Random Forest thường đạt độ chính xác cao nhất nhờ việc kết hợp nhiều cây quyết định (decision trees) và khả năng nắm bắt các mối quan hệ phi tuyến (non-linear relationships).

Random Forest là mô hình tổng thể tốt nhất trên bộ dữ liệu Titanic nhờ khả năng xử lý dữ liệu phức tạp và các mối quan hệ phi tuyến.

2.Gradient Boosting

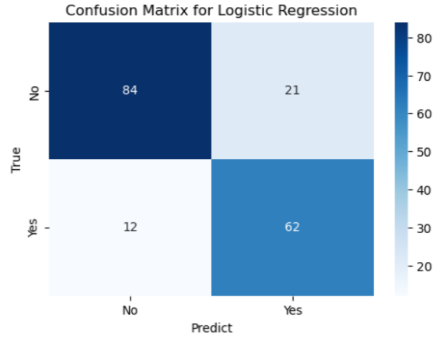
Gradient Boosting cũng hoạt động hiệu quả, đặc biệt khi dữ liệu có nhiều **tương tác phức tạp giữa các đặc trưng (features)**.

Gradient Boosting là một lựa chọn thay thế mạnh mẽ, đặc biệt khi các **siêu tham số (hyperparameters)** được tinh chỉnh phù hợp.



Random Forest và **Gradient Boosting** thường có **ít giả trị âm giả (False Negative)** hơn, nghĩa là chúng **ít bỏ sót việc dự đoán những hành khách sống sót**.

3. Logistic Regression



Logistic Regression là một mô hình tuyến tính đơn giản, hoạt động hiệu quả khi dữ liệu có quan hệ gần tuyến tính, nhưng **có thể cho độ chính xác thấp hơn so với các mô hình dựa trên cây (tree-based models)**.

Logistic Regression phù hợp với các mô hình đơn giản và **dễ giải thích, dễ hiểu kết quả**.

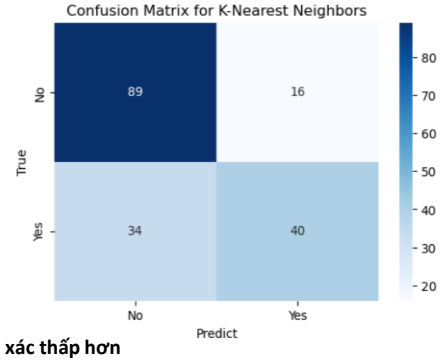
Logistic Regression hoạt động tốt với các mối quan hệ tuyến tính, nhưng **có thể đạt điểm F1 thấp hơn ở các lớp thiểu số (minority classes)**.

4. K-Nearest Neighbors

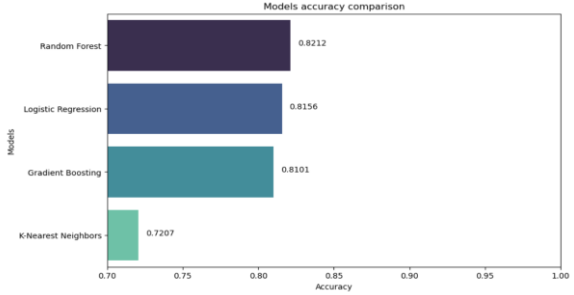
K-Nearest Neighbors (KNN) phụ thuộc vào **giá trị k** được chọn và **cách chuẩn hóa đặc trưng (feature scaling)**; mô hình này **có thể hoạt động kém hơn** nếu tập dữ liệu có **nhiều đặc trưng** hoặc **phân bố không đồng đều**.

KNN phù hợp với các mô hình đơn giản và **dễ diễn giải, dễ hiểu kết quả**.

KNN và **Logistic Regression** có thể **mắc nhiều sai sót hơn ở một số lớp dữ liệu nhất định**.



KNN và **Logistic Regression** có **độ chính xác thấp hơn một chút**, nhưng vẫn là **những lựa chọn hợp lý cho các mô hình đơn giản hoặc có tốc độ xử lý nhanh**.



Tổng kết:
Biểu đồ cho thấy **Random Forest** là mô hình tốt nhất về độ chính xác, trong khi **KNN** hoạt động kém hiệu quả hơn. **Logistic Regression** và **Gradient Boosting** vẫn là các lựa chọn ổn định, cân bằng giữa độ chính xác và độ phức tạp mô hình.