# Phân tích Thảm họa tàu Titanic sử dụng mô hình Machine Learning


## Nhóm Thực Hiện

- 3123560073 - Đỗ Duy Quý
- 3123410274 - Lư Hồng Phúc
- 3123410387 - Nguyễn Hữu Trí
- 3123410258 - Lê Thanh Phát


## Môn Học

**NHẬP MÔN MÁY HỌC**

Nhóm 1 - Chiều thứ 7

# Các Giai Đoạn Chính Của Dự Án

## 1. Giới thiệu Bài toán & Mục tiêu

Xác định rõ bối cảnh, bài toán cần giải quyết và các tiêu chí đánh giá hiệu quả.

## 2. Khám phá & Phân tích Dữ liệu (EDA)

Nghiên cứu cấu trúc, chất lượng và các mối quan hệ quan trọng ẩn chứa trong dữ liệu.

## 3. Phương pháp & Xây dựng Mô hình

Áp dụng các kỹ thuật tạo đặc trưng và lựa chọn thuật toán học máy phù hợp.

## 4. Kết quả & Đánh giá

Đánh giá hiệu năng của mô hình và so sánh các chiến lược đã được triển khai.

## 5. Kết luận & Hướng phát triển

Tổng kết những bài học rút ra và đề xuất các cải tiến tiềm năng trong tương lai.

# 1. Giới thiệu bài toán và Mục tiêu

## Phân Loại Bài Toán

Đây là **bài toán phân loại nhị phân**, nhằm dự đoán khả năng sống sót của mỗi hành khách (0 = tử vong, 1 = sống sót).

**Đặc trưng chính:**

- **Pclass:** Hạng vé (1, 2, 3).
- **Sex, Age:** Thông tin cá nhân.
- **Embarked:** Cảng xuất phát (C, Q, S).
- **PassengerId:** ID duy nhất, không dùng để dự đoán.

**Thách thức dữ liệu:** Xử lý dữ liệu thiếu (ví dụ: `Age`, `Cabin`) và mất cân bằng lớp (`Survived`).

## Tiêu Chí Đánh Giá Mô Hình

Mô hình được đánh giá dựa trên các tiêu chí sau:

- **Accuracy:** Tỷ lệ dự đoán đúng tổng thể.
- **F1-score:** Cân bằng giữa Precision và Recall, phù hợp với dữ liệu mất cân bằng.

Sự kết hợp này đảm bảo đánh giá toàn diện hiệu năng mô hình.

# Tổng Quan Về Các Nghiên Cứu Titanic Trước Đây

### Độ Chính Xác & Phương Pháp

- Độ chính xác thường đạt 78-94%.
- Các phương pháp phổ biến: Hồi quy Logistic, Cây Quyết định, Rừng ngẫu nhiên, SVM.

### Yếu Tố Ảnh Hưởng Chính

- Giới tính (Sex) và Hạng vé (Pclass) có ảnh hưởng lớn nhất.
- Phụ nữ và hành khách hạng cao có tỷ lệ sống sót cao hơn.
- Tuổi (Age) và Cảng lên tàu (Embarked) ít nổi bật hơn.

### Bài Học & Xu Hướng Nghiên Cứu

- Tiền xử lý dữ liệu (dữ liệu thiếu, mã hóa) và tạo đặc trưng rất quan trọng.
- Tập trung vào hiểu đặc trưng và áp dụng đa dạng mô hình học máy để tìm giải pháp tối ưu.

# Hướng Tiếp Cận Của Chúng Tôi

## Khám Phá Dữ Liệu (EDA)

Phân tích sâu để hiểu cấu trúc, phân phối và mối quan hệ giữa các đặc trưng.

## Tiền Xử Lý Dữ Liệu

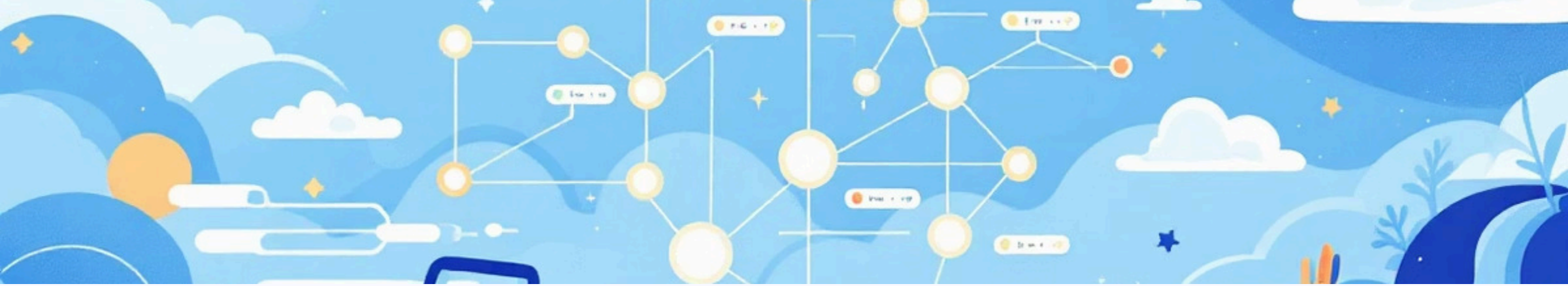
Xử lý giá trị thiếu, mã hóa đặc trưng, và tạo các đặc trưng mới để cải thiện chất lượng dữ liệu.

## Xây Dựng Mô Hình

Phát triển và huấn luyện các mô hình dự đoán, bao gồm XGBoost, Logistic Regression và Random Forest,...

## Đánh Giá, So Sánh Hiệu Suất

Đánh giá hiệu quả của các mô hình thông qua K-Fold Cross-Validation và lựa chọn mô hình tốt nhất.



# Dữ liệu Titanic: Cấu Trúc và Kiểu Dữ Liệu

### Nguồn Dữ Liệu

Bộ dữ liệu công khai từ Kaggle, tập trung vào thảm họa Titanic.

### Cấu Trúc Dữ Liệu

Tập huấn luyện (**train.csv**): 891 dòng, 12 cột (bao gồm biến mục tiêu).

Tập kiểm tra (**test.csv**): 418 dòng, 11 cột.

### Kiểu Dữ Liệu Chính

- Số liên tục (Age, Fare…)
- Categorical có thứ tự (Pclass)
- Categorical không thứ tự (Sex, Embarked)

# Phân Tích Dữ Liệu Thiếu và Trùng Lặp

## Dữ Liệu Thiếu (Missing Values)

Các cột có giá trị thiếu cần xử lý là: `Age`, `Cabin`, và `Embarked`.

`Cabin`: thiếu 687/891 giá trị (khoảng 77%)

`Age`: thiếu 177 giá trị (khoảng 20%)

Cột `Cabin` có tỷ lệ thiếu dữ liệu cao nhất, đòi hỏi phương pháp xử lý đặc biệt thay vì loại bỏ.

## Dữ Liệu Trùng Lặp

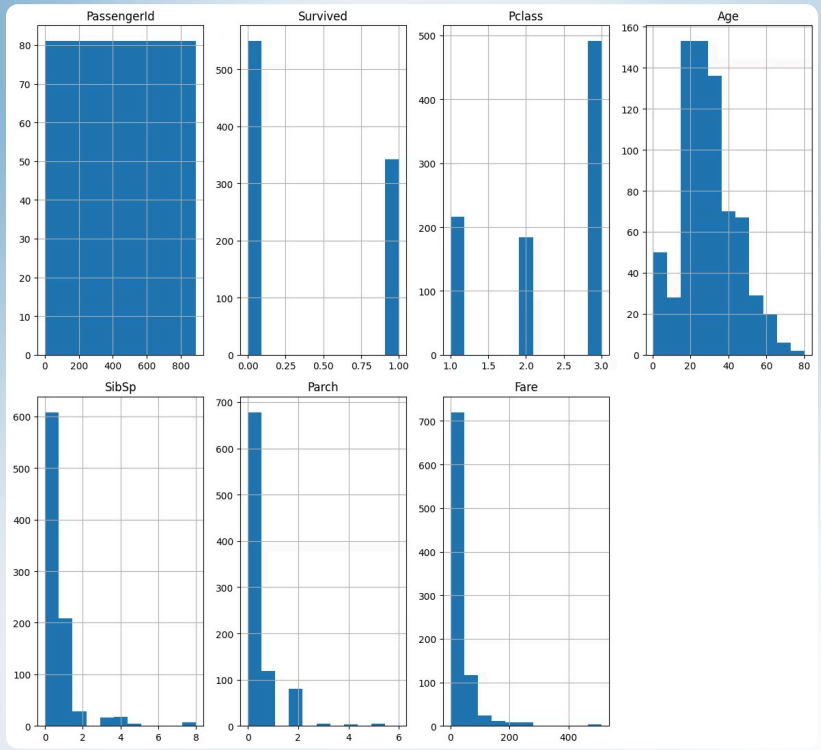
Đã kiểm tra và xác nhận **không có dòng dữ liệu trùng lặp** nào trong tập huấn luyện, đảm bảo tính duy nhất của các bản ghi hành khách.

# Chiến Lược Xử Lý Giá Trị Thiếu

Việc điền giá trị thiếu cần đảm bảo tính logic và chính xác dựa trên đặc điểm của hành khách và tránh rò rỉ dữ liệu.

## Phân Loại & Phương Pháp Xử Lý:

- **Tập huấn luyện (Train):** Các cột có giá trị thiếu bao gồm Age, Cabin, Embarked.
- **Tập kiểm thử (Test):** Các cột có giá trị thiếu bao gồm Age, Cabin, Fare.
- **Cabin:** Thiếu khoảng 77-80% giá trị, do đó rất khó sử dụng hiệu quả cho phân tích định lượng.
- **Age, Embarked, Fare:** Số lượng giá trị thiếu ít hơn, sẽ được điền bằng **giá trị trung vị (median)** hoặc **mốt (mode)** tương ứng.
- **Lưu ý quan trọng:** Cần xử lý tập huấn luyện và tập kiểm thử tách biệt để tránh hiện tượng **rò rỉ dữ liệu (data leakage)**, đảm bảo tính khách quan của mô hình.




# Phân Tích Dữ Liệu Đơn Biến

- **Tỷ lệ sống sót:** Mất cân bằng (38% sống sót, 62% không sống sót) – cần lưu ý để tránh sai lệch mô hình.
- **Hạng vé (Pclass):** Đa số hành khách thuộc hạng 3, phản ánh phân bố thu nhập.
- **Tuổi (Age):** Phân phối lệch phải, tập trung ở độ tuổi 20-40.
- **SibSp & Parch:** Phần lớn hành khách đi một mình hoặc không có người thân đi cùng.
- **Giá vé (Fare):** Phân phối lệch phải, đa số trả giá thấp, một số ít trả giá rất cao.
- **PassengerId:** Định danh duy nhất, không có giá trị dự đoán.

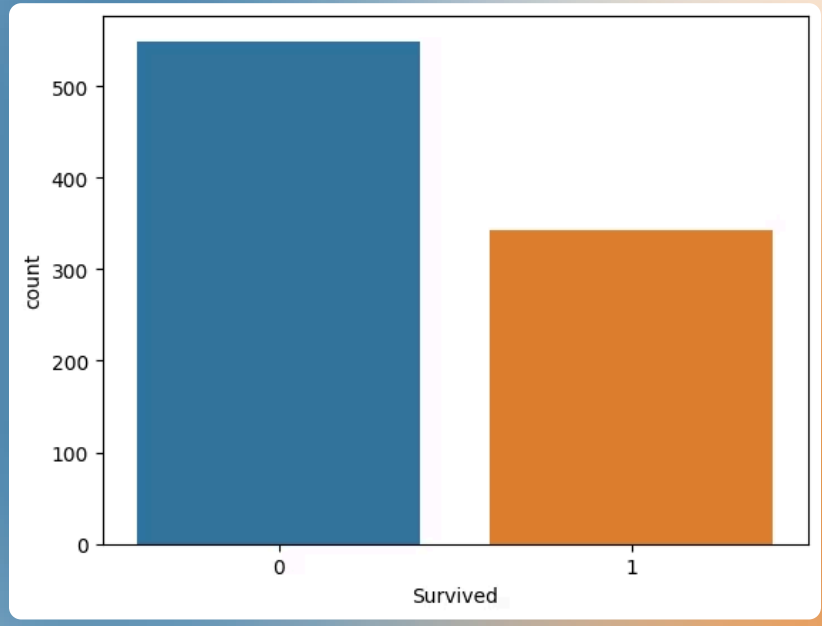
# Phân Tích Dữ Liệu Đơn Biến

## Tỷ lệ sống sót

Dữ liệu **mất cân bằng**: 38% sống sót, 62% không sống sót. Cần xử lý để tránh **sai lệch mô hình**.

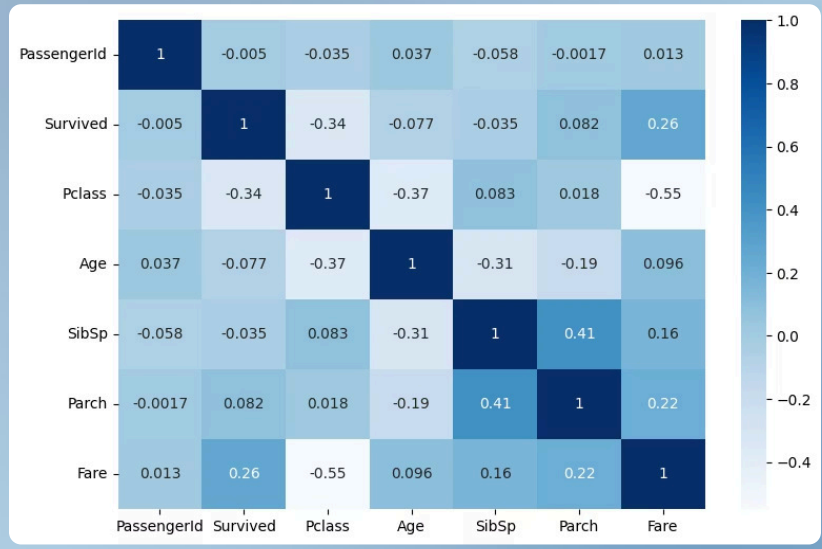
## Hiểu rõ phân phối

Phân tích từng biến giúp có cái nhìn tổng quan, chuẩn bị cho **phân tích đa biến** và **xây dựng mô hình** hiệu quả.



# Phân tích đa biến

- **PassengerId**: Tương quan không đáng kể, loại bỏ khi huấn luyện mô hình.
- **Survived – Pclass** (-0.34): Hành khách hạng thấp, tỷ lệ sống sót thấp hơn.
- **Survived – Fare** (0.26): Giá vé cao, cơ hội sống sót cao hơn.
- **Pclass – Fare** (-0.55): Hạng thấp, vé rẻ hơn (có thể dư thừa thông tin).
- **SibSp – Parch** (0.41): Hành khách đi cùng gia đình.
- **Age – Survived** (-0.077): Tuổi tác ít ảnh hưởng đến sống sót.
- **Age – Pclass** (-0.37): Hành khách hạng thấp có xu hướng trẻ hơn.



# Tương Quan Giữa Các Đặc Trưng

Việc khám phá mối quan hệ giữa các đặc trưng giúp chúng ta hiểu sâu hơn về những yếu tố ảnh hưởng đến khả năng sống sót và định hình chiến lược xây dựng mô hình hiệu quả.

### Giới Tính & Pclass

Đây là hai đặc trưng có tương quan mạnh mẽ nhất với **Survived**. Phụ nữ và trẻ em, đặc biệt ở hạng vé cao, có tỷ lệ sống sót cao hơn rõ rệt.

### Deck & Pclass

Đặc trưng **Deck** (trích xuất từ Cabin) thể hiện mối liên hệ chặt chẽ với **Pclass**. Hành khách ở các boong trên thường thuộc hạng sang và có vị trí thoát hiểm thuận lợi hơn.

### Fare & Survived

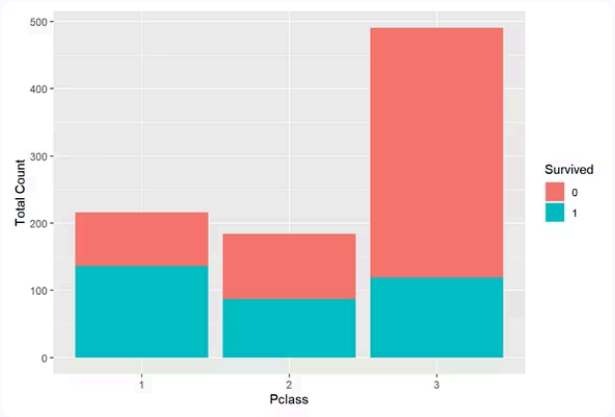
Giá vé (**Fare**) có mối tương quan dương với khả năng sống sót. Hành khách trả giá vé cao hơn có xu hướng sống sót nhiều hơn, thường đi kèm với hạng vé cao và vị trí boong tốt.

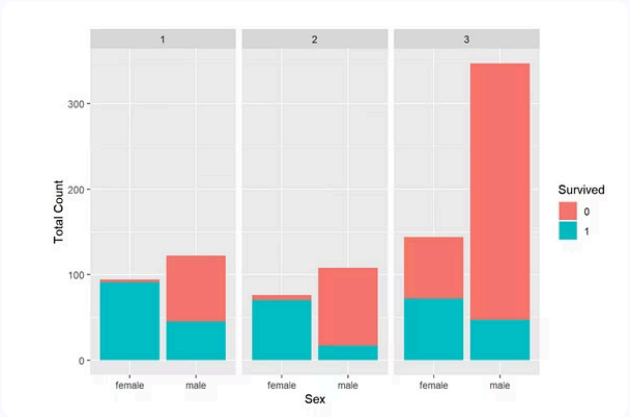
### SibSp & Parch

Hai đặc trưng này tương quan với nhau, phản ánh kích thước gia đình. Việc có người thân đi cùng có thể ảnh hưởng đến quyết định sống còn và khả năng được cứu hộ.

# Phân Tích Tương Quan Trực Quan

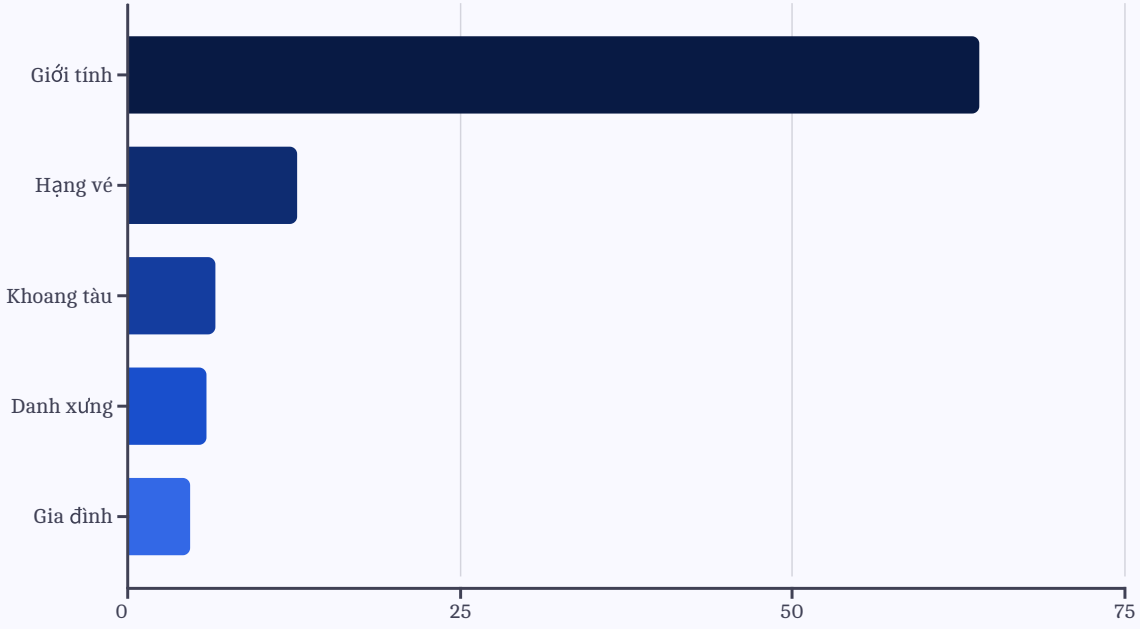
Để xây dựng mô hình dự đoán hiệu quả, việc trực quan hóa và đánh giá mối quan hệ giữa các đặc trưng là vô cùng quan trọng. Các biểu đồ dưới đây cung cấp cái nhìn sâu sắc về sự tương quan giữa các biến trong tập dữ liệu.




Thông qua các biểu đồ tương quan, chúng ta có thể dễ dàng xác định những đặc trưng có ảnh hưởng mạnh mẽ nhất đến biến mục tiêu **Survived** (ví dụ: Sex, Pclass, Fare) và nhận diện các đặc trưng ít liên quan hoặc có thể loại bỏ (ví dụ: PassengerId). Việc phân tích này không chỉ hỗ trợ trong việc chọn lọc đặc trưng quan trọng mà còn giúp tránh hiện tượng overfitting, từ đó cải thiện độ chính xác và khả năng tổng quát hóa của mô hình.

# Phân tích các yếu tố quyết định khả năng sống sót trên tàu Titanic




Biểu đồ minh họa mức độ quan trọng của các đặc trưng trong mô hình dự đoán. Rõ ràng, **Giới tính** và **Hạng vé** là hai yếu tố có ảnh hưởng vượt trội, đóng góp gần 77% vào khả năng sống sót của hành khách.

Các đặc trưng khác như **Khoang tàu** (Deck), **Danh xưng** (Title từ trường Name), và **Gia đình** (Family size, từ SibSp và Parch) cũng cho thấy vai trò đáng kể, phản ánh sự phức tạp của các yếu tố sinh tồn trên tàu Titanic.

# Phương pháp và quá trình thử nghiệm

Tổng quan về các bước xử lý dữ liệu, tạo đặc trưng và quy trình xây dựng, đánh giá mô hình dự đoán.

### 1. Xử Lý Dữ Liệu Thiếu

- **Tuổi (Age):**

  Sử dụng giá trị trung vị (median) của nhóm dựa trên tổ hợp `Pclass` và `Sex` để điền vào các giá trị thiếu.

- **Cảng Khởi Hành (Embarked):**

  Điền giá trị thiếu bằng chế độ (mode) của cột, tức là cảng xuất phát phổ biến nhất.

- **Giá Vé (Fare):**

  Điền giá trị thiếu bằng giá trị trung vị (median) của cột

### 2. Tạo Đặc Trưng Mới (Feature Engineering)

- **Kích Thước Gia Đình (FamilySize):**

  Tổng số thành viên trong gia đình (`SibSp` + `Parch` + 1)

- **Đi Một Mình (IsAlone):**

  Một biến nhị phân chỉ ra liệu hành khách có đi một mình hay không (`FamilySize` == 1).

- **Danh Xưng (Title):**

  Trích xuất danh xưng (ví dụ: Mr, Miss, Mrs, Master, Dr) từ cột `Name`.

# Phương pháp và quá trình thử nghiệm

### 3. Quy Trình Thử Nghiệm (Experimentation Pipeline)

Quy trình thử nghiệm bao gồm tiền xử lý dữ liệu, lựa chọn các đặc trưng phù hợp, xây dựng và huấn luyện mô hình, cũng như đánh giá hiệu suất dựa trên các chỉ số như độ chính xác và F1-score. Việc thử nghiệm liên tục giúp cải thiện khả năng dự đoán sống sót của mô hình trên tập dữ liệu kiểm tra.

### 4. Đánh Giá Mô Hình (Model Evaluation)

Để đánh giá hiệu suất của mô hình, chúng tôi sử dụng các chỉ số sau:

- **Độ chính xác (Accuracy)**

  Tỷ lệ các dự đoán đúng trên tổng số dự đoán.

- **Độ chính xác (Precision)**

  Khả năng của mô hình không gắn nhãn tích cực cho một mẫu tiêu cực.

- **Độ bao phủ (Recall)**

  Khả năng của mô hình tìm thấy tất cả các mẫu tích cực.

- **Điểm F1 (F1-Score)**

  Trung bình điều hòa của Precision và Recall.

- **ROC-AUC**

  Đánh giá hiệu suất phân loại ở các ngưỡng khác nhau. Giá trị cao hơn cho thấy khả năng phân loại tốt hơn.

# Phương pháp

## Dataset

- Dataset nhỏ (891 mẫu)
- Một số cột thiếu nhiều giá trị

## Tiền xử lý dữ liệu

- Sử dụng K-fold và Pipeline để tránh Data Leakage.
- Tổng hợp phương pháp: K-fold + Pipeline + GridSearchCV.

## Thuật toán phân loại

- RandomForestClassifier
- DecisionTreeClassifier
- SVM
- LogisticRegression

## Chọn mô hình tối ưu

- Dùng GridSearchCV để chọn thuật toán tốt nhất.
- Đánh giá hiệu suất mô hình đã chọn.

# Kết Quả Đánh Giá Mô Hình

## 1 So sánh hiệu suất các thuật toán

Áp dụng K-fold kết hợp Pipeline và GridSearchCV để đánh giá các mô hình:

- RandomForestClassifier
- DecisionTreeClassifier
- SVM
- LogisticRegression

## 2 Mô hình tối ưu

Sau quá trình tinh chỉnh và đánh giá, **RandomForestClassifier** cho thấy hiệu suất vượt trội nhất trên các chỉ số chính (Accuracy, Precision, Recall, F1-Score), phù hợp với yêu cầu bài toán phân loại.

## 3 Ý nghĩa kết quả

Mô hình được chọn có khả năng dự đoán cao, giảm thiểu lỗi trên tập dữ liệu nhỏ và không đầy đủ, chứng minh tính hiệu quả của phương pháp đã áp dụng.

# Hành Trình Tối Ưu Hóa Mô Hình Titanic: Từ Baseline đến Ensembling

### Lần 1: Baseline

Áp dụng xử lý dữ liệu cơ bản và xây dựng mô hình dự đoán đầu tiên.

Mô hình tốt nhất: **Random Forest**.

Điểm số: **0.7559**.

### Lần 2: Mở Rộng Đặc Trưng (Relationship)

Thêm đặc trưng **Relationship**, tính toán từ số lượng người thân (SibSp, Parch) trên tàu.

Điểm số cải thiện rõ rệt: **0.7775**.

### Lần 3: Kỹ Thuật Đặc Trưng Sâu Hơn (Title & Deck)

Trích xuất đặc trưng **Title** từ trường Name và **Deck** từ trường Cabin, giúp nắm bắt thông tin chi tiết hơn.

Mô hình tốt nhất: **XGBoost**.

Điểm số: **0.75837** (lưu ý dấu hiệu overfitting).

### Lần 4: Kết hợp giữa base model và best model dùng VottingClassifier

Tạo 'base_model' chỉ với 2 đặc trưng 'mạnh' nhất là 'Sex' và 'Pclass' kết hợp với model chính  đã thiết kế lúc trước và dùng thuật toán 'VotingClassifier' để thực hiện điều này.

### Lần 5: Ensembling & Đặc Trưng Nâng Cao

Kết hợp sức mạnh của nhiều mô hình bằng **VotingClassifier** và bổ sung các đặc trưng chuyên sâu (IsAlone, male_rich) để đạt hiệu suất tối đa.

Điểm số cao nhất: **0.78229**.

### Những lần tiếp theo

Liên tục cải tiến, đưa ra đặc trưng, phương pháp mới nhưng kết quả không khả quan

# Kết quả và Thảo luận

### Mô hình hiệu quả nhất

Mô hình kết hợp (Ensemble) sử dụng **VotingClassifier** với các thuật toán *Random Forest, XGBoost, SVC, và LGBM Classifier* đã cho kết quả tốt nhất trên Kaggle: **0.78229**.

### Bài học về Feature Engineering

Việc tạo ra các đặc trưng mới một cách thông minh (như Deck, Title, Family_Size) có tác động lớn đến hiệu suất mô hình, thậm chí còn quan trọng hơn việc chỉ tinh chỉnh tham số.

### Vấn đề Overfitting

Trong quá trình thử nghiệm (Lần 3), chúng tôi nhận thấy việc thêm quá nhiều đặc trưng phức tạp có thể dẫn đến **overfitting**, khi mô hình học quá tốt trên tập huấn luyện nhưng dự đoán kém trên dữ liệu mới.

# Kết luận & Hướng phát triển

Dự án đã xây dựng thành công mô hình dự đoán khả năng sống sót trên tàu Titanic với độ chính xác cao. Phân tích đã tái khẳng định các yếu tố lịch sử quan trọng như giới tính, địa vị xã hội (hạng vé) và tình trạng gia đình có ảnh hưởng trực tiếp đến tỷ lệ sống sót, và chúng đã được tích hợp hiệu quả vào mô hình.

## Hướng phát triển

### Tối ưu hóa siêu tham số

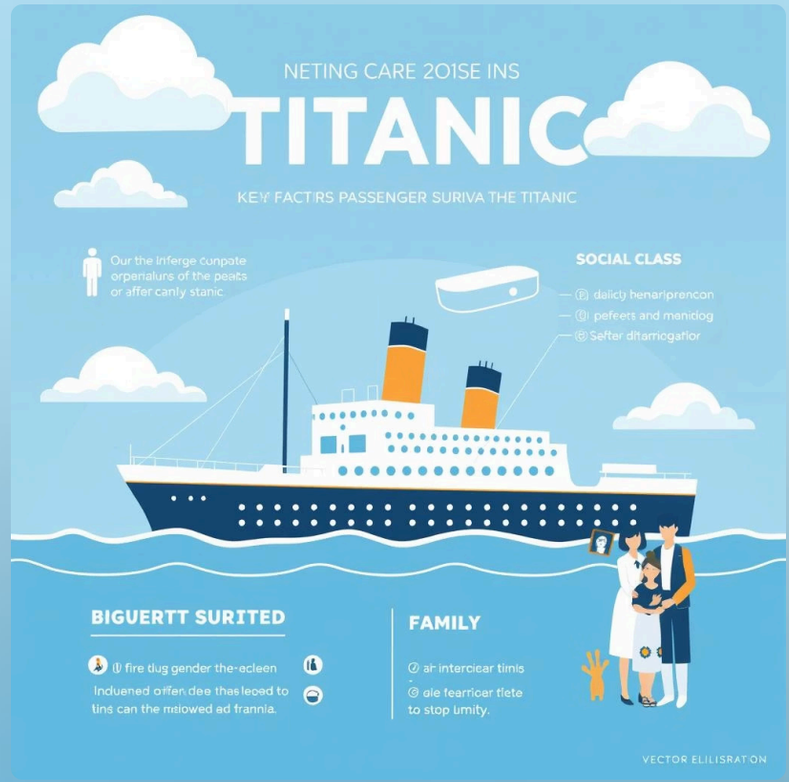
Để tối ưu hóa hiệu suất, chúng tôi sẽ sử dụng các kỹ thuật tìm kiếm siêu tham số nâng cao như Optuna hoặc Bayesian Optimization.

### Thử nghiệm mô hình nâng cao

Tiếp tục khám phá các thuật toán học máy khác như CatBoost (hiệu quả với biến phân loại) hoặc các mô hình Deep Learning đơn giản nhằm nâng cao hơn nữa độ chính xác.

### Phân tích chuyên sâu nhóm hành khách

Đi sâu phân tích các nhóm nhỏ hơn (ví dụ: "nam giới giàu có đi một mình", "phụ nữ có con nhỏ ở hạng 3") để tìm kiếm các insight ẩn và điều chỉnh mô hình phù hợp.



# Cảm ơn quý vị đã lắng nghe