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

Thành viên: *Lê Thanh Phát, Nguyễn Hữu Tri, Lư Hồng Phúc, Đỗ Duy Quý*

Phân tích vụ đắm tàu Titanic là điều cần thiết để hiểu rõ dữ liệu lịch sử. Mối tương quan giữa các đặc trưng độc lập và phụ thuộc đã được quan sát nhằm xác định những yếu tố có thể ảnh hưởng đến khả năng sống sót của hành khách. Trong bài báo này, chúng tôi đã khám phá bộ dữ liệu Titanic và triển khai các thuật toán học máy gồm XGBoost, Decision Tree và Random Forest để dự đoán tỷ lệ sống sót của hành khách. Một số yếu tố như "Age", "Sex" và "Pclass" đóng vai trò then chốt trong khả năng sống sót của họ. Phân tích so sánh giữa các thuật toán này đã được thực hiện, và các mô hình được đánh giá bằng nhiều chỉ số đo lường (metrics). Dựa trên kết quả phân tích và đánh giá, XGBoost cho thấy hiệu suất vượt trội hơn so với các thuật toán còn lại trong nghiên cứu này.

I. Giới thiệu

Các thuật toán học máy (machine learning algorithms) đã giúp các nhà phân tích dữ liệu, nhà khoa học dữ liệu và kỹ sư dữ liệu khám phá được những hiểu biết sâu sắc từ các dữ liệu lịch sử. Một trong những vụ đắm tàu nổi tiếng nhất trong lịch sử chính là thảm kịch Titanic, xảy ra vào ngày 15 tháng 4 năm 1912, do tàu va chạm với tảng băng trôi, khiến nhiều phần của con tàu khổng lồ bị vỡ nát.

Nhiều bằng chứng và phân tích đã được công bố nhằm làm rõ nguyên nhân vụ tai nạn cũng như tỷ lệ sống sót của hành khách trên tàu. Dữ liệu liên quan đến thảm kịch này đã được thu thập và công khai .

Một trong những thuật toán học máy mạnh mẽ nhất hiện nay, đạt hiệu suất hàng đầu trong nhiều bài toán thực tế, là Gradient Boosting . Trong nhiều năm qua, thuật toán này vẫn được xem là phương pháp chủ đạo cho các bài toán học có dữ liệu nhiễu, đặc trưng không đồng nhất và mối quan hệ phức tạp giữa các biến . Thuật toán học máy đã được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, từ tài chính, y tế, giao thông cho đến phân tích hành vi người dùng .

Một số tác giả đã nghiên cứu vụ đắm tàu Titanic bằng các hướng tiếp cận khác nhau:

- Eric đã thực hiện phân tích so sánh giữa ba thuật toán học máy gồm Naive Bayes, SVM (Support Vector Machine) và Cây quyết định (Decision Tree), với độ chênh lệch chỉ 2,64% giữa mô hình tốt nhất và kém nhất.
- Cicoria tiến hành phân cụm (cluster analysis) và áp dụng thuật toán cây quyết định để xác định các yếu tố ảnh hưởng đến khả năng sống sót. Trong đó, đặc trưng "Giới tính" (Sex) được xác định là yếu tố quan trọng nhất đối với khả năng sống sót .
- Trevor sử dụng Rừng ngẫu nhiên (Random Forest) và Cây quyết định trên một số đặc trưng chọn lọc trong bộ dữ liệu Titanic, với độ chính xác cao nhất đạt 78%.
- Singh triển khai các thuật toán gồm Hồi quy Logistic, Naive Bayes, Cây quyết định và Rừng ngẫu nhiên trong nghiên cứu của mình.

Các nghiên cứu trước đây đã sử dụng Random Forest để dự đoán khả năng sống sót của hành khách. Thuật toán LogisticRegression được triển khai đạt độ chính xác cao nhất là 94,26%.

Cách tiếp cận của chúng tôi khác biệt so với các nghiên cứu trước. Mục tiêu của bài báo này là thực hiện quy trình khoa học dữ liệu trên bộ dữ liệu Titanic, bao gồm khám phá dữ liệu, làm sạch và biến đổi dữ liệu, xây dựng các mô hình dự đoán như XGBoost, Decision Tree và Random Forest dựa trên các đặc trưng có sẵn. Ngoài ra, chúng tôi xác định các yếu tố ảnh hưởng đến tỷ lệ sống sót của hành khách và tiến hành đánh giá hiệu suất của các mô hình thông qua các chỉ số đo lường.

Phần còn lại của bài báo được cấu trúc như sau:

- Phần II: Phân tích bài toán,
- Phần III và IV: Phương pháp và quá trình thử nghiệm
- Phần V: Kết quả và đánh giá,
- Phần VI: Đề xuất hướng cải tiến mô hình.
- Phần VII: Tham khảo

II. Phân tích bài toán

1. Phân tích bối cảnh và mục tiêu

1.1. Phân tích bối cảnh và mục tiêu

Vụ đắm tàu Titanic năm 1912 là một trong những thảm họa hàng hải nổi tiếng nhất trong lịch sử. Từ dữ liệu hành khách được công bố, bài toán đặt ra là dự đoán khả năng sống sót của một hành khách dựa trên các đặc trưng cá nhân và thông tin chuyến đi.

Mục tiêu: xây dựng mô hình học máy dự đoán biến 'Survived' (1: sống sót, 0: không sống sót) từ những đặc trưng đầu vào: *Sex, Pclass, Sex, Name, AgeSex, SibSp, Parch, Fare, Ticket, Cabin, Embarked.*

1.2. Dạng bài toán

Đây là một bài toán phân loại nhị phân (binary classification) trong học máy có giám sát (supervised learning), với:

- Đầu vào (Input features): thông tin hành khách.
- Đầu ra (Label): trạng thái sống sót (Survived).
- Mục tiêu huấn luyện: tìm hàm ánh xạ f(X) → yf(X)

sao cho mô hình dự đoán đúng xác suất sống sót cao nhất.

1.3. Thách thức chính

- Dữ liệu thiếu: một số cột như Age, Cabin, Embarked có giá trị null.
- Tính không tuyến tính: khả năng sống sót không phụ thuộc tuyến tính vào từng đặc trưng mà phụ thuộc vào sự kết hợp của chúng (ví dụ: phụ nữ + hạng vé 1 → tỷ lệ sống cao).
- Dữ liệu mất cân bằng: số người không sống sót lớn hơn số người sống sót.
- Nhiều đặc trưng phi số: cần mã hóa (Sex, Embarked, Title, CabinPrefix).

2. Data Description (Mô tả dữ liệu)

2.1 Nguồn Dữ liệu

- Bộ dữ liệu được sử dụng trong nghiên cứu là Titanic – Machine Learning from Disaster do Kaggle cung cấp.
- Tập huấn luyện (train.csv) gồm 891 mẫu, tập kiểm thử (test.csv) gồm 418 mẫu.
- Mỗi dòng tương ứng với một hành khách trên tàu Titanic.

2.2 Cấu Trúc Dữ Liệu

Các cột (features) chính trong dữ liệu gồm:

Tên cột	Kiểu dữ liệu	Mô tả
PassengerId	int	Mã định danh hành khách
Survived	int (0 hoặc 1)	Nhãn mục tiêu:
Pclass	int (1, 2, 3)	Hạng vé
Name	string	Họ tên hành khách
Sex	string	Giới tính (male/female)
Age	float	Tuổi của hành khách
SibSp	int	Số anh chị em/vợ chồng
Parch	int	Số cha mẹ/con cái đi cùng
Ticket	string	Số vé
Fare	float	Giá vé
Cabin	string	Số buồng
Embarked	string	Cảng lên tàu

2.3. Đặc điểm nổi bật của dữ liệu

- Dữ liệu có giá trị thiếu ở các cột Age, Cabin và Embarked, đòi hỏi bước làm sạch dữ liệu (data cleaning).
- Cột Pclass thể hiện địa vị xã hội và điều kiện kinh tế, ảnh hưởng đáng kể đến cơ hội sống sót.
- Giới tính (Sex) là yếu tố mạnh mẽ nhất, do nguyên tắc "phụ nữ và trẻ em được cứu trước".

3. Phân Tích Dữ Liệu

3.1 Phân tích dữ liệu thiếu

Kết quả thống kê cho thấy một số cột trong cả tập huấn luyện và tập kiểm tra chứa giá trị bị thiếu.
Cụ thể:

- Tập huấn luyện (train) có giá trị thiếu ở các cột Age, Cabin và Embarked.
- Tập kiểm tra (test) có giá trị thiếu ở các cột Age, Cabin và Fare.

Trong một số nghiên cứu, hai tập dữ liệu này được gộp lại trước khi xử lý để đảm bảo tính đồng nhất về phân phối thống kê. Tuy nhiên, trong nghiên cứu này hai tập dữ liệu được giữ tách biệt để tránh rò rỉ thông tin (data leakage) từ tập huấn luyện sang tập kiểm tra.
Việc xử lý giá trị thiếu độc lập cho từng tập giúp mô hình phản ánh điều kiện thực tế của dữ liệu mới (unseen data), thay vì học theo đặc trưng thống kê của toàn bộ tập dữ liệu đã biết. Điều này đặc biệt quan trọng khi mục tiêu là đánh giá năng lực tổng quát hóa của mô hình.
Tỷ lệ dữ liệu thiếu được thống kê như sau:

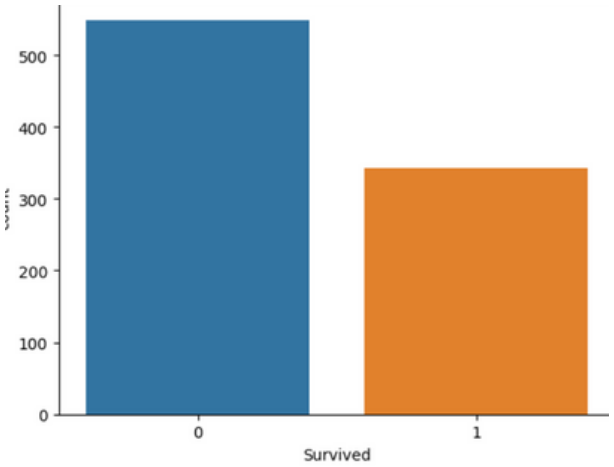
- Các cột Age, Embarked, và Fare chỉ bị thiếu ở mức nhỏ, có thể được điền bằng các thước đo thống kê mô tả (trung vị, trung bình, hoặc mode).
- Riêng cột Cabin bị thiếu gần 80% giá trị, cho thấy dữ liệu này không đủ thông tin để sử dụng trong phân tích định lượng.

Phân tích sơ bộ dữ liệu huấn luyện
Trong tập huấn luyện:

- 38.3% hành khách sống sót
- Phần lớn hành khách thuộc hạng 3 (Pclass = 3)
- 50% hành khách có độ tuổi từ 20 đến 38

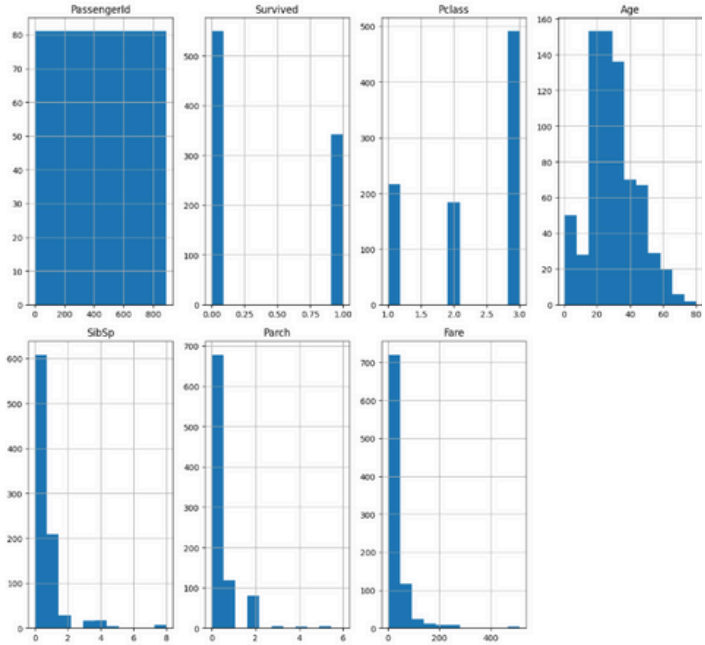
Với tỷ lệ sống sót chỉ 0.383, nếu ta giả định toàn bộ hành khách đều không sống sót, mô hình vẫn có thể đạt độ chính xác 62%.

3.2 Phân tích đơn biến



Biểu đồ trên minh họa rõ ràng sự mất cân bằng về tỷ lệ sống sót trong tập dữ liệu huấn luyện, với số lượng người không sống sót cao hơn đáng kể. Đây là một yếu tố quan trọng cần xem xét khi xây dựng mô hình dự đoán, vì sự mất cân bằng lớp có thể ảnh hưởng tiêu cực đến hiệu suất của mô hình.

Biểu đồ histogram đơn biến (univariate analysis) cho các cột trong bộ dữ liệu Titanic.



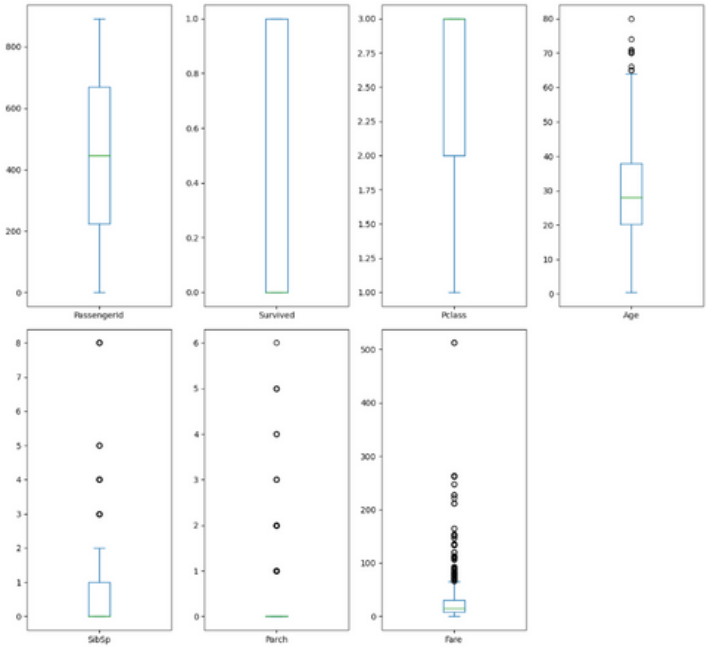
Biểu đồ trên thể hiện phân bố của các biến số trong bộ dữ liệu Titanic, bao gồm: *PassengerId, Survived, Pclass, Age, SibSp, Parch và Fare.*

Phân tích từng biến cho thấy đặc điểm tổng quan của dữ liệu như sau:

- Survived: Dữ liệu mất cân bằng, với khoảng 38% hành khách sống sót và 62% không sống sót. Điều này cho thấy cần chú ý đến vấn đề class imbalance khi huấn luyện mô hình.
- Pclass: Hầu hết hành khách thuộc hạng 3, tiếp theo là hạng 1 và hạng 2. Điều này phản ánh phần lớn hành khách có điều kiện kinh tế trung bình hoặc thấp.
- Age: Phân bố lệch phải (right-skewed), tập trung chủ yếu ở độ tuổi 20–40 tuổi. Rất ít hành khách trên 60 tuổi.
- SibSp (Siblings/Spouses Aboard): Đa số hành khách không có anh chị em hoặc vợ/chồng đi cùng (giá trị = 0), chứng tỏ nhiều người đi một mình.
- Parch (Parents/Children Aboard): Tương tự, phần lớn hành khách không đi cùng cha mẹ hoặc con cái.
- Fare: Phân bố lệch phải rõ rệt, cho thấy chỉ một số ít hành khách trả giá vé rất cao, trong khi phần lớn trả mức giá thấp hơn nhiều.

Phân tích đơn biến giúp hiểu đặc điểm và phân bố của từng biến, đồng thời phát hiện dữ liệu ngoại lai, giá trị thiếu hoặc độ lệch phân phối. Kết quả này là bước tiền đề quan trọng trước khi thực hiện phân tích đa biến và xây dựng mô hình dự đoán

Biểu đồ hộp (Boxplot) cho các biến số trong bộ dữ liệu Titanic, dùng để phát hiện giá trị ngoại lai (outliers) và quan sát độ phân tán của dữ liệu.



Biểu đồ hộp (Boxplot) thể hiện phạm vi giá trị (range), trung vị (median) và các giá trị ngoại lai (outliers) của từng biến trong bộ dữ liệu Titanic.

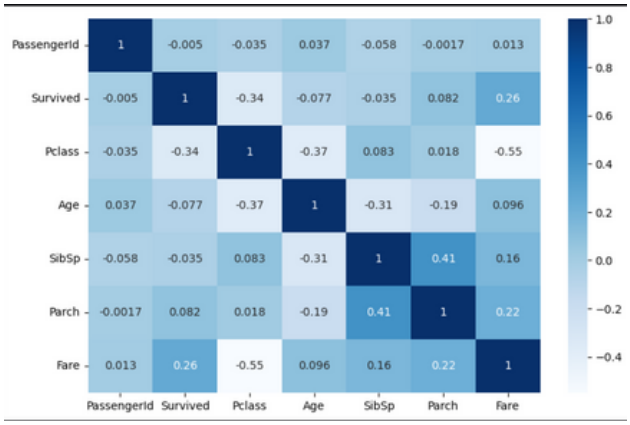
Kết quả quan sát được như sau:

- PassengerId: Phân bố đều, không có ngoại lai đáng kể, chỉ mang tính định danh.
- Survived: Là biến nhị phân (0 và 1), nên không xuất hiện ngoại lai.
- Pclass: Bao gồm 3 giá trị rời rạc (1, 2, 3), không có giá trị bất thường.
- Age: Có một số ngoại lai ở phía trên (tuổi > 65). Tuy nhiên, các giá trị này vẫn hợp lý vì có hành khách lớn tuổi thực sự trên tàu, do đó được giữ nguyên.
- SibSp (Siblings/Spouses Aboard): Xuất hiện nhiều ngoại lai ở giá trị từ 3–8, phản ánh một số hành khách đi cùng nhiều người thân. Những giá trị này được coi là có ý nghĩa thực tế và không bị loại bỏ, chỉ cần chuẩn hóa khi huấn luyện mô hình.
- Parch (Parents/Children Aboard): Có ngoại lai ở các giá trị 4–6, biểu thị những trường hợp hành khách đi cùng gia đình đông người. Các giá trị này được giữ lại vì có thể ảnh hưởng đến khả năng sống sót.
- Fare: Là biến có nhiều ngoại lai rõ rệt. Phần lớn giá vé tập trung ở mức thấp, nhưng một số hành khách (đặc biệt ở hạng nhất) trả giá vé rất cao (trên 200–500), làm phân phối bị lệch phải mạnh. Thay vì loại bỏ, biến này được biến đổi logarit (log-transform) để giảm ảnh hưởng của độ lệch và tăng độ ổn định khi mô hình học.

Tổng thể, các giá trị ngoại lai trong bộ dữ liệu chủ yếu phản ánh đặc trưng thực tế của hành khách, do đó không bị loại bỏ mà được biến đổi hoặc chuẩn hóa hợp lý. Cách tiếp cận này giúp giữ lại thông tin quan trọng, đồng thời đảm bảo dữ liệu đầu vào ổn định cho mô hình dự đoán.

3.3 Phân Tích Đa Biến

Ma trận tương quan (correlation matrix) giữa các biến số trong bộ dữ liệu Titanic

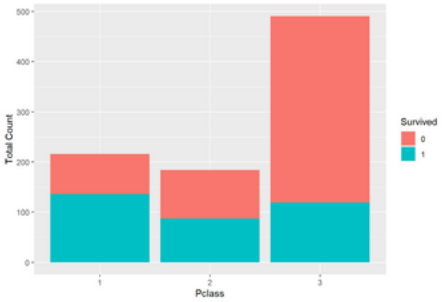


Biểu đồ trên thể hiện mối tương quan giữa các biến số trong tập dữ liệu Titanic, bao gồm: PassengerId, Survived, Pclass, Age, SibSp, Parch và Fare. Mức độ tương quan được biểu diễn bằng các giá trị từ -1 đến 1, trong đó giá trị dương cho thấy mối quan hệ thuận chiều, còn giá trị âm thể hiện mối quan hệ nghịch chiều. Trước hết, biến PassengerId hầu như không có mối tương quan đáng kể với bất kỳ biến nào khác, vì đây chỉ là mã định danh của hành khách, không mang ý nghĩa về hành vi hay đặc điểm sinh tồn. Do đó, biến này có thể được loại bỏ trong quá trình huấn luyện mô hình.

- Giữa biến Survived (tình trạng sống sót) và Pclass (hạng vé) có hệ số tương quan khoảng -0.34, thể hiện mối quan hệ nghịch vừa phải. Nghĩa là hành khách ở hạng vé thấp hơn (ví dụ hạng 3) có xu hướng sống sót ít hơn so với hành khách ở hạng vé cao (hạng 1). Đây là một kết quả hợp lý, phản ánh sự khác biệt về điều kiện và vị trí của các khoang tàu.
- Tương tự, Survived có tương quan thuận nhẹ với Fare (giá vé), với hệ số khoảng 0.26. Điều này cho thấy hành khách trả giá vé cao hơn — thường thuộc tầng lớp giàu có — có cơ hội sống sót cao hơn, có thể vì họ ở gần khu vực có thuyền cứu sinh hoặc được ưu tiên trong quá trình sơ tán.
- Giữa Pclass và Fare tồn tại mối tương quan nghịch mạnh nhất trong toàn bộ ma trận, với hệ số khoảng -0.55. Điều này khẳng định rằng hạng vé càng thấp thì giá vé càng rẻ, và ngược lại — một quan hệ hiển nhiên nhưng rất quan trọng vì nó cho thấy hai biến này mang thông tin tương tự nhau ở một mức độ nhất định.
- Biến SibSp (số anh chị em/vợ chồng đi cùng) và Parch (số cha mẹ/con đi cùng) có mối tương quan thuận vừa, khoảng 0.41, phản ánh xu hướng những hành khách đi cùng gia đình thường có cả người thân và con cái trên tàu. Đây là mối quan hệ hợp lý trong bối cảnh dữ liệu thật.
- Đối với Age, các tương quan với những biến còn lại đều thấp. Cụ thể, mối tương quan giữa Age và Survived chỉ khoảng -0.077, cho thấy tuổi không ảnh hưởng mạnh đến khả năng sống sót một cách tuyến tính. Ngoài ra, Age có mối tương quan âm nhẹ với Pclass (-0.37), nghĩa là hành khách ở hạng thấp hơn thường trẻ hơn một chút so với những người ở hạng cao hơn.

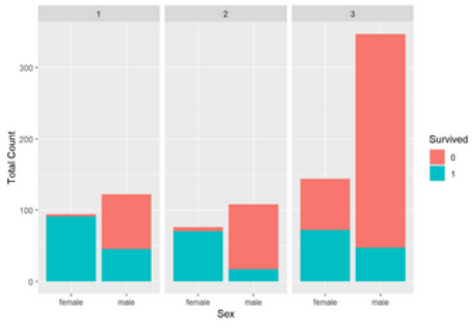
Cuối cùng, mối tương quan giữa các đặc trưng nhìn chung không quá cao (trừ cặp Pclass–Fare), cho thấy tập dữ liệu này ít gặp vấn đề đa cộng tuyến nghiêm trọng. Điều này thuận lợi cho việc huấn luyện các mô hình tuyến tính và cây quyết định.

Pclass vs Survived

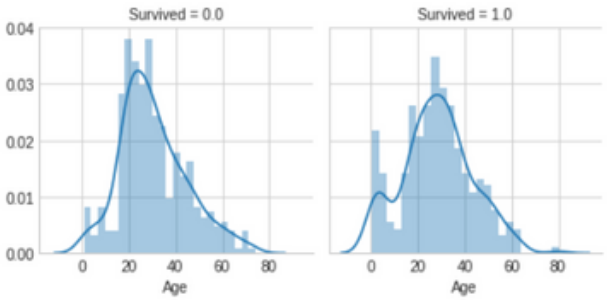


Từ biểu đồ, ta có thể thấy rằng tỷ lệ sống sót của hành khách Pclass 1 cao hơn so với Pclass 2 và Pclass 3, trong khi tỷ lệ sống sót của hành khách Pclass 3 thấp nhất.
→ Không nghi ngờ gì Người giàu có tỷ lệ sống tốt hơn người nghèo

Pclass vs Survived vs Sex



Trong tất cả các khoang, tỷ lệ sống sót của nữ tốt hơn nam



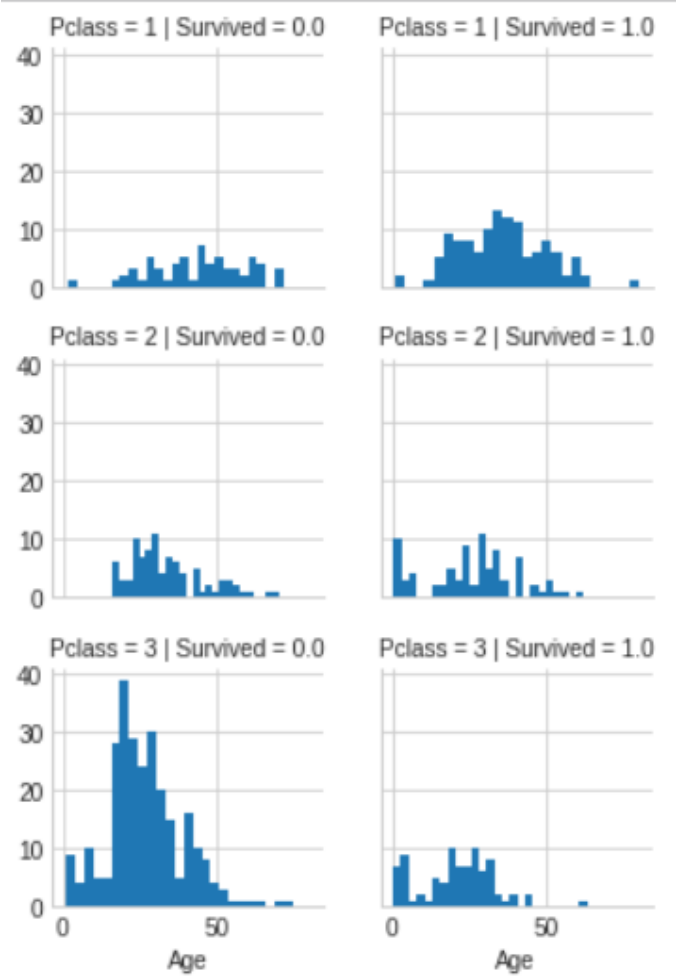
Age vs Survived

Biến tuổi (Age) thể hiện mối quan hệ rõ ràng với khả năng sống sót:

- Hành khách ≤ 10 tuổi có tỷ lệ sống sót cao, cho thấy trẻ em được ưu tiên cứu trước.
- Hành khách cao tuổi nhất (khoảng 80 tuổi) vẫn sống sót, chứng tỏ tuổi tác không hoàn toàn quyết định cơ hội sống.
- Một lượng lớn hành khách khoảng 20 tuổi không sống sót, cho thấy người trẻ tuổi chiếm phần lớn nạn nhân.
- Phần lớn hành khách nằm trong độ tuổi 15–35, là nhóm chính trên tàu.

Vì vậy, biến Age được giữ lại làm đặc trưng quan trọng khi huấn luyện mô hình.
Các giá trị thiếu của Age sẽ được bổ sung (impute) dựa trên phân phối tuổi trong tập dữ liệu, nhằm đảm bảo cấu trúc dữ liệu không bị méo lệch.

Age vs pclass vs Survived



+ Pclass = 1 (Hạng nhất)
- Biểu đồ bên trái (không sống sót) cho thấy số lượng tử vong ít hơn hẳn so với hạng 2 và 3.
- Ở cột bên phải (sống sót), ta thấy khá nhiều hành khách trong độ tuổi 20–50 được cứu.
- → Kết luận: phần lớn hành khách hạng nhất có cơ hội sống sót cao hơn, đặc biệt là người trưởng thành, thể hiện rõ sự ưu tiên trong cứu hộ.

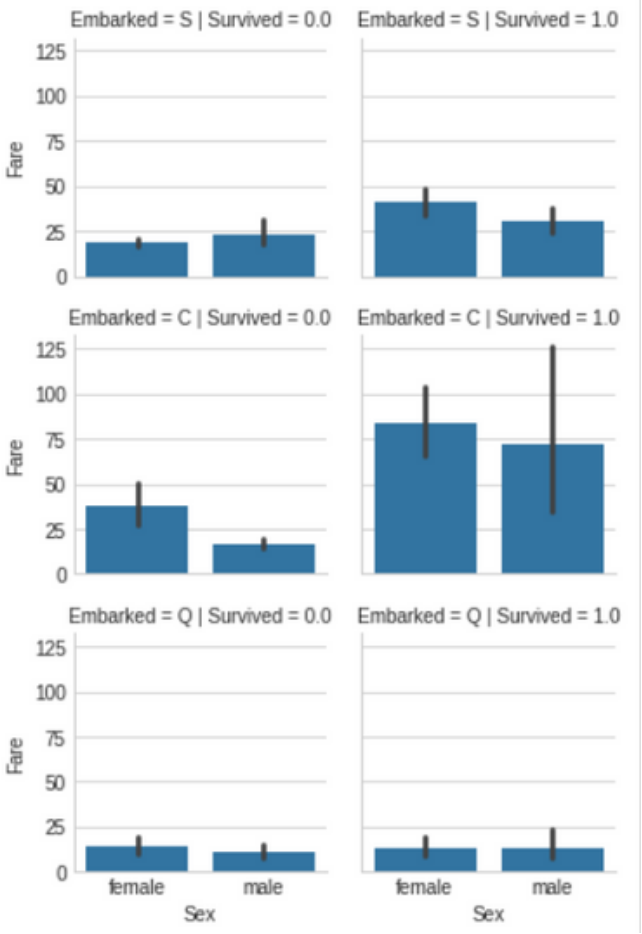
+ Pclass = 2 (Hạng nhì)
- Cả hai biểu đồ cho thấy số lượng tử vong và sống sót khá cân bằng hơn so với hạng 1.
- Độ tuổi sống sót tập trung quanh 20–40 tuổi.
- → Kết luận: tỉ lệ sống sót giảm so với hạng 1, nhưng vẫn còn khá cao, đặc biệt ở nhóm người trưởng thành trẻ tuổi.

+ Pclass = 3 (Hạng ba)
- Số người không sống sót (cột trái) chiếm áp đảo, tập trung nhiều nhất trong nhóm tuổi 15–30.
- Cột bên phải (sống sót) có ít người hơn hẳn, và phân bố rải rác.
- → Kết luận: hành khách hạng 3 có khả năng sống sót thấp nhất, dù có nhiều người trẻ.
- Điều này có thể do vị trí khoang ở tầng dưới của tàu, khó tiếp cận thuyền cứu sinh.

Tổng quan nhận định

1. Tuổi (Age) không phải yếu tố quyết định chính, vì người ở mọi độ tuổi đều có khả năng sống sót nếu ở hạng vé cao.
2. Hạng vé (Pclass) là yếu tố có tác động mạnh: hạng 1 có nhiều người sống sót, hạng 3 hầu như tử vong nhiều.
3. Dù có một số trẻ em ở hạng 3 sống sót, nhưng tổng thể xu hướng vẫn cho thấy sự phân tầng xã hội rõ rệt trong tỷ lệ sống sót.

Embarked vs Sex vs Fare vs Survived



+ Embarked = S (Southampton)
- Người sống sót (Survived = 1) trả giá vé cao hơn rõ rệt so với người tử vong.
- Cả nam và nữ đều có giá vé cao hơn trong nhóm sống sót, nhưng không chênh lệch lớn giữa hai giới.
- Tuy nhiên, ta có thể thấy rằng nhiều người lên từ cảng S có giá vé trung bình thấp hơn so với các cảng khác — thể hiện rằng phần lớn hành khách từ đây là tầng lớp trung lưu hoặc hạng vé thấp.

→ Kết luận: tại cảng Southampton, những người trả giá vé cao hơn (tức là đi hạng cao) có cơ hội sống sót cao hơn, bất kể giới tính.

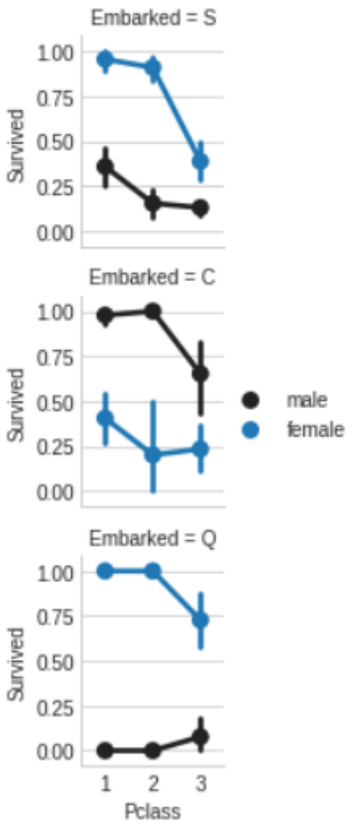
+ Embarked = C (Cherbourg)
- Đây là nhóm có giá vé cao nhất trong toàn bộ biểu đồ, đặc biệt ở nhóm sống sót.
- Người sống sót (Survived = 1), cả nam lẫn nữ, đều có giá vé trung bình rất cao — nhiều người trong số này có thể thuộc hạng vé 1.
- Sự khác biệt về giá vé giữa nhóm sống và tử vong tại Cherbourg là rất rõ:
  - Người tử vong thường mua vé rẻ (khoảng dưới 30).
  - Người sống sót trả vé đắt hơn nhiều (thường trên 70).

→ Kết luận: tại Cherbourg, giá vé cao gắn liền với khả năng sống sót cao, phản ánh tầng lớp thượng lưu có điều kiện sống tốt hơn và được ưu tiên cứu.

+ Embarked = Q (Queenstown)
- Ở cả hai cột (sống và không sống), giá vé trung bình đều rất thấp — khoảng dưới 20.
- Không có sự khác biệt đáng kể giữa nam và nữ.
- Đây là nhóm hành khách nghèo nhất, phần lớn đi hạng 3.

→ Kết luận: tại Queenstown, giá vé thấp đồng nghĩa với hầu như không có lợi thế sống sót; cả nam lẫn nữ đều có tỷ lệ tử vong cao.

Embarked vs Sex vs Pclass vs Survived



Biểu đồ này thể hiện tỷ lệ sống sót (Survived) của hành khách theo hạng vé (Pclass), giới tính (Sex) và bến lên tàu (Embarked).

Mỗi cụm biểu đồ (S, C, Q) tương ứng với một bến cảng nơi hành khách lên tàu:

- S = Southampton
- C = Cherbourg
- Q = Queenstown

Cùng đi qua từng phần chi tiết nhé:

Embarked = S (Southampton)

- Đây là bến có số lượng hành khách nhiều nhất.
- Nữ giới (đường xanh) có tỷ lệ sống sót cao hơn rõ rệt so với nam giới ở tất cả các hạng vé.
- Tuy nhiên, tỷ lệ sống sót của nữ giảm mạnh khi hạng vé giảm — từ gần 100% ở hạng 1 xuống còn khoảng 30% ở hạng 3.
- Nam giới (đường đen) có tỷ lệ sống sót thấp toàn bộ, dưới 30% ở mọi hạng vé.

→ Kết luận: tại cảng S, giới tính và hạng vé đều ảnh hưởng rõ rệt đến khả năng sống sót — nữ giới và hạng vé cao có lợi thế hơn nhiều.

Embarked = C (Cherbourg)

- Biểu đồ cho thấy tỷ lệ sống sót của nữ giới ở hạng 1 và 2 đều rất cao (xấp xỉ 100%).
- Đáng chú ý, nam giới ở hạng 1 tại cảng này cũng có tỷ lệ sống sót cao (trên 70%) — cao hơn nhiều so với hai bến còn lại.

→ Có thể giả định rằng hành khách ở Cherbourg phần lớn là giàu có, hoặc có điều kiện tiếp cận cứu sinh tốt hơn. Tuy nhiên, ở hạng 3, cả nam lẫn nữ đều có tỷ lệ sống sót giảm mạnh.

Embarked = Q (Queenstown)

- Cảng này có rất ít hành khách, nhưng vẫn thể hiện rõ xu hướng chung:
  - Nữ giới ở hạng 1 và 2 có cơ hội sống cao.
  - Nam giới ở mọi hạng hầu như không sống sót.

→ Đây là nhóm hành khách nghèo hơn, nhiều người thuộc hạng vé 3, nên khả năng sống sót thấp.

Nhận Xét

- Giới tính là yếu tố ảnh hưởng mạnh nhất: nữ giới có khả năng sống sót cao hơn rõ rệt ở mọi bến và hạng vé.
- Hạng vé (Pclass) cũng có ảnh hưởng mạnh: hạng càng cao thì tỷ lệ sống càng lớn.
- Bến lên tàu (Embarked) thể hiện sự khác biệt xã hội — hành khách từ Cherbourg (C) có tỷ lệ sống cao nhất, đặc biệt ở hạng nhất; trong khi Southampton (S) và Queenstown (Q) có nhiều hành khách nghèo và tỷ lệ sống thấp.

4. Nhận xét tổng thể

1. Survived (Biến mục tiêu)

- Nhận xét: Tỷ lệ sống sót là 38.3%, nghĩa là chỉ khoảng 1/3 hành khách sống sót.
- Ý nghĩa: Dữ liệu mất cân bằng nhẹ (imbalance) → cần chú ý khi huấn luyện mô hình (không nên chỉ dùng accuracy làm thước đo).

2. Pclass (Hạng vé)

- Nhận xét:
  - Hành khách hạng 3 chiếm nhiều nhất, sau đó là hạng 1 và hạng 2.
  - Tỷ lệ sống sót giảm dần theo hạng vé: Hạng 1 sống cao nhất, hạng 3 thấp nhất.
- Ý nghĩa: Hạng vé liên quan chặt chẽ đến khả năng sống sót — có thể phản ánh điều kiện phòng, vị trí trên tàu, hoặc khả năng tiếp cận xuồng cứu sinh.

3. Sex (Giới tính)

- Nhận xét:
  - Số lượng nam nhiều hơn nữ, nhưng nữ có tỷ lệ sống sót cao hơn nhiều.
- Ý nghĩa: Đây là biến có mối quan hệ mạnh nhất với sống sót → xác nhận chính sách "Women and children first".

4. Age (Tuổi)

- Nhận xét:
  - Độ tuổi phổ biến nhất: 20–38 tuổi, chiếm khoảng 50%.
  - Trẻ nhỏ có tỷ lệ sống cao hơn, còn nhóm trung niên và cao tuổi sống sót ít hơn.
- Ý nghĩa: Tuổi có ảnh hưởng đến khả năng sống sót, nhưng không mạnh bằng giới tính hay hạng vé.

5. Fare (Giá vé)

- Nhận xét:
  - Phân phối lệch phải (right-skewed): phần lớn hành khách trả giá vé thấp, chỉ vài người trả rất cao.
  - Người trả vé cao thường thuộc hạng 1 và có tỷ lệ sống sót lớn hơn.
- Ý nghĩa: Fare có thể dùng như biến đại diện cho điều kiện kinh tế hoặc tầng lớp xã hội.

6. Embarked (Cảng lên tàu)

- Nhận xét:
  - Đa số hành khách lên tàu tại Southampton (S).
  - Tỷ lệ sống sót cao hơn ở Queenstown (Q) và Cherbourg (C) so với S.
- Ý nghĩa: Có thể liên quan đến vị trí khoang tàu hoặc đặc điểm nhóm khách.

7. SibSp & Parch (Thành viên gia đình)

- Nhận xét:
  - Hầu hết hành khách đi một mình (SibSp = 0, Parch = 0).
  - Có ít người thân đi cùng → tỷ lệ sống thấp hơn.
  - Nhóm có 1–2 người thân thường sống cao hơn.
- Ý nghĩa: Cho thấy yếu tố "đi cùng gia đình" giúp tăng khả năng sống, có thể do hỗ trợ lẫn nhau.

III. Phương pháp và quá trình thử nghiệm

A. Phương pháp

- Sau khi phân tích và đánh giá dữ liệu, nhận thấy dataset khá nhỏ, chỉ 891 mẫu và có một số cột bị thiếu giá trị rất nhiều. Thêm vào đó, để xử lí thuận tiện và tránh tình trạng 'data leakage', dùng kỹ thuật *K-fold* kết hợp với *Pipeline*.
- Vì là bài toán phân loại, ta có nhiều thuật toán để lựa chọn cho mô hình, có thể kể đến như: RandomForestClassifier, DecisionTreeClassifier, SVM, LogisticRegression,...Do đó, để có thể đánh giá và chọn ra thuật toán phù hợp và tốt nhất cho bài toán, ta dùng kỹ thuật *GridSearchCV* để chọn ra thuật toán tốt nhất.

→ Kết hợp hai điều trên ta có phương pháp:

K-fold + Pipeline + GridSearchCV

IV. Quá trình thử nghiệm

Lần 1:

- *Tiền xử lí dữ liệu (*xử lí các giá trị thiếu)
  - Cột số ('Age', 'Fare', 'SibSp', 'Parch'): điền bằng giá trị *trung vị* (median)
  - Cột phân loại ('Sex', 'Embarked', 'Pclass') : điền bằng giá trị có *tần suất cao nhất*
- *Lựa chọn đặc trưng*
  - *Feature selection:*
    - Cột '*Cabin*': xóa, vì thiếu hơn 77% giá trị (687/891 mẫu)
    - Cột '*Name*': xóa, vì chưa khai thác được thông tin gì ẩn trong nó
    - Cột '*PassengerId*': xóa, vì không có giá trị
  - *Feature engineering:*
    - *Cột phân loại ('Sex', 'Embarked', 'Pclass'): biến đổi giá trị về kiểu thứ tự 0, 1, 2, ... (OrdinalEncoder)*
  - *Feature extraction:* Không có.
- *Các thuật toán và tham số tương ứng:*
  - *Logistic Regression*
    - max_iter=10000
    - C: [0.1, 1.0, 10],
    - solver: ["lbfgs", "liblinear"]
  - *Random Forest*
    - random_state=21
    - n_estimators: [70, 100, 200]
    - max_depth: [3, 5, 10]
  - *XGBoost*
    - use_label_encoder=False
    - eval_metric='logloss'
    - n_estimators: [70, 100, 200]
    - learning_rate: [0.05, 0.1]
    - max_depth": [3, 5]
- *K-fold:*
  - k = 10
- *Score matric: dùng cả 3 độ đo, là:*
  - 'accuracy': 'accuracy',
  - 'f1': 'f1',
  - 'roc_auc': 'roc_auc'

→ *Mô hình tốt nhất và tham số: RandomForestClassifier*
  - *max_depth:* 10
  - *n_estimators:* 200

Kết quả lần thử nghiệm 1 (V5-Kaggle)	
Train score	*0.7736*
Submit score	*0.7559*

Xem xét lại dữ liệu, tự đặt câu hỏi rằng: "*Tại sao giá trị của cột 'Cabin' lại thiếu nhiều như vậy? Có phải chỉ có những người sống sót mới có thể cho biết khoang khi ở trên tàu của họ không?"* Do đó, ta không xóa cột 'Cabin' mà lấy các khoang của cột 'Cabin' và điền giá trị thiếu bằng giá trị nào đó, ví dụ: 'Z'.

Ngoài ra, ta suy luận: "*Nếu như các hành khách đi chung với người thân thì họ sẽ có tỉ lệ sống thấp hơn vì họ tốn thời gian nhiều hơn để cứu những người thân của mình."*Vì vậy, ta tạo thêm cột '*Relationship*' để tính các tổng người thân của hành khách:

'Relationship' = 'SibSp' + 'Parch'

Lần 2:

- *Tiền xử lí dữ liệu: giống như lần 1, và:*
  - Cột *'Cabin'*: điền các giá trị thiếu bằng 'Z'
- *Lựa chọn đặc trưng*
  - *Feature selection:*
    - Cột *'Name'*: xóa, vì chưa khai thác được thông tin gì ẩn trong nó
    - Cột *'PassengerId'*: xóa, vì nhận thấy không có giá trị
    - Cột *'SibSp' & 'Parch'*: xóa, vì được thay thế bởi cột mới - *'Relationship'*
  - *Feature engineering:*
    - *Cột mới 'Relationship': giá trị bằng tổng giá trị của hai cột: 'SibSp' & 'Parch'*
  - *Feature extraction:* Không có.
- *Các thuật toán và tham số tương ứng:*
  - *Logistic Regression*
    - max_iter=10000
    - C: [0.1, 1.0, 10],
    - solver: ["lbfgs", "liblinear"]
  - *Random Forest*
    - random_state=21
    - n_estimators: [50, 100, 200, 300, 400, 500]
    - max_depth: [1, 3, 5, 7, 9]
  - *XGBoost*
    - use_label_encoder=False,
    - eval_metric='logloss'
    - n_estimators: [50, 100, 200, 300, 400, 500]
    - learning_rate: [0.01, 0.05, 0.1]
    - max_depth": [1, 3, 5, 7, 9]
- *K-fold:*
  - k = 10
- *Score matric: dùng cả 3 độ đo, là:*
  - 'accuracy': 'accuracy',
  - 'f1': 'f1',
  - 'roc_auc': 'roc_auc'

→ *Mô hình tốt nhất và tham số: RandomForestClassifier*
  - *max_depth: 9*
  - *n_estimators:* 300

Kết quả lần thử nghiệm 2 (V9-Kaggle)	
Train score	*0.7788*
Submit score	*0.7775*

Nhận xét: mô hình đã tốt hơn khi tăng tỉ lệ dự đoán đúng tăng thêm gần *3%*

Xem xét kĩ hơn vào dữ liệu ta phát hiện ra các danh xưng trong cột '*Name*', khác nhau giữa mỗi người và cho biết thêm thông tin về họ. Ví dụ với 3 cái tên trong cột '*Name*':

- *Futrelle, Mrs. Jacques Heath (Lily May Peel)*
- *Skoog, Master. Harald*
- *Minahan, Dr. William Edward*

Lần lượt, thông tin về họ là:

- *Mrs: phụ nữ đã có chồng,*
- *Master: bé trai (cách gọi lúc trước),*
- *Dr: tiến sĩ.*

Các thông tin này cho ta thêm suy đoán khả năng họ có thể sống sót cao hơn (vì khi có tai nạn thì người ta luôn ưu tiên phụ nữ, trẻ em và những người có học thức cao, chức vụ cao trong xã hội. Ngoài ra, trong cột 'Cabin', các hành khách ở các khoang khác nhau như '*C85', 'B42*',... cho thấy rằng họ ở các vị trị khác nhau trên tàu, nên tỉ lệ sống sót của họ cũng khác nhau.

Do đó, ta thêm cột '*Title*' hay danh xứng, trích xuất từ trong cột '*Name*' và trích xuất từ cột 'Cabin' các khoang khác nhau của từng hành khách tạo thành cột 'Deck'

Lần 3:

- *Tiền xử lí dữ liệu:* giống như lần 2.
- *Lựa chọn đặc trưng*
  - *Feature selection:* giống như lần 2 nhưng không xóa cột '*Name*'
  - *Feature engineering:* giống như lần 2
  - *Feature extraction:*
    - *Cột 'Title': trích xuất danh xưng từ cột 'Name'*
    - *Cột 'Deck': trích xuất khoang từ cột 'Cabin'*
- *Các thuật toán và tham số tương ứng:* giống như lần 2
- *K-fold:* giống như lần 2
- *Score matric:* giống như lần 2

→ *Mô hình tốt nhất và tham số: XGBClassifier*

- *learning_rate: 0.01*
- *max_depth: 7*
- *n_estimators: 5*00

Kết quả lần thử nghiệm 3 (V10-Kaggle)	
Train score	*0.7902*
Submit score	*0.75837*

Nhận xét:

- Điểm khi train của mô hình tăng, chứng tỏ mô hình đã học tốt hơn trên tập train này.
- Song. điểm khi nôp (Submit) thì giảm. Điều này chứng to mô hình của chúng ta đã có một chút 'overfitting' hay quá khớp.

Do vậy, ta cần kiểm tra lại cách chúng ta tạo đặc trưng và có thể giảm, điều chỉnh các tham số của mô hìnhl để xử lí việc quá khớp với dữ liệu train của model.

Một góc nhìn khác, ta kiểm tra xem các đặc trưng quan trọng như thế nào và chúng đóng góp bao nhiêu phần trăm trong kết quả dự đoán model hay trong tổng số các đặc trưng, các đặc trưng nào là 'mạnh', ta có được kết quả (theo độ 'mạnh' giảm dần) như sau:

1. *categorical__Sex: 0.6406*
2. *categorical__Pclass: 0.1274*
3. *categorical__Deck: 0.0660*
4. *categorical__Title: 0.0595*
5. *numeric__Relationship: 0.0473*
6. *numeric__Fare: 0.0227*
7. *numeric__Age: 0.0201*
8. *categorical__Embarked: 0.0163*

Ta thấy rằng, '*Sex*' và '*Pclass*' là 2 đặc trưng 'mạnh' nhất, chiếm phần lớn ảnh hưởng đối với kết quả dự đoán của model. Từ đó, một ý tưởng nảy ra: *Tạo một 'base_model' chỉ với 2 đặc trưng 'mạnh' nhất là 'Sex' và 'Pclass' kết hợp với model chính của chúng ta đã thiết kế lúc trước và dùng thuật toán 'VotingClassifier' để thực hiện điều này.*

Lần 4:

- *Tiền xử lí dữ liệu:* giống như lần 3
- *Lựa chọn đặc trưng*
  - *Feature selection:* giống như lần 3
  - *Feature engineering:* giống như lần 3
  - *Feature extraction: giống như lần 3*
- *Các thuật toán và tham số tương ứng:* giống như lần 3
- *K-fold:* giống như lần 3
- *Score matric:* giống như lần 3
- Thuật toán VotingClassifier:
  - estimators=[
    ('baseline', baseline_model),
    ('full', best_model)
    ]
  - voting='soft',
  - *voting*n_jobs=-1

Kết quả lần thử nghiệm 4 (V12-Kaggle)	
Base score	*0.7868*
Train score	*0.7902*
Submit score	*0.77272*

Nhận xét:

- Điểm base khá cao, gần với điểm train. Điều này chứng minh rằng nhận định của chúng ta đúng khi cho rằng hai đặc trưng 'Sex' và 'Pclass' là 'mạnh' nhất.
- Điểm submit đã được cải thiện. Tuy nhiên nó không vượt qua được so với lần thử nghiệm 2 (0.7775). Vậy các phương pháp của chúng ta đã tới ngưỡng cao nhất.

Nhưng nó đưa chúng ta đến một câu hỏi: *Tại sao không sử dụng VotingClassifier cho nhiều thuật toán phân loại khác?*

Thực hiện ý tưởng vừa nảy ra từ lần thử nghiệm thứ 4 ở trên, ta dùng thuật toán VotingClassifier với các thuật toán:

- *RandomForestClassifier*
- *XGBClassifier*
- *SVC*
- *LGBMClassifier*

Và cải tiến các đặc trưng, thêm các đặc trưng mới:

- *Cột* '*IsAlone*': giá trị bằng 1 khi khi giá trị ở cột '*Relationship*' bằng 1, còn lại bằng 0. Vì các hành khách đi một mình thì có khả năng sống cao hơn vì họ không dành thời gian cứu ai cả.
- Cột '*male_rich*': những người nam giới có giá vé, cột 'Fare', cao hơn 50 (EDA cho ta số này), vì những người giàu có này thì sẽ có ở vị trí tốt trên tàu và họ khôn ngoan hơn.
- Cột '*Ticket_freq*': đếm tần suất vé giống nhau (những người thân mua vé chung), thì họ có tỉ lệ sống thấp hơn.

Lần 5:

- *Tiền xử lí dữ liệu:* giống như lần 4.

- *Lựa chọn đặc trưng*
  - *Feature selection:* giống như lần 4
  - *Feature engineering:* giống như lần 4, thêm vào:
    - *Cột* '*IsAlone*': giá trị bằng 1 khi khi giá trị ở cột '*Relationship*' bằng 1, còn lại bằng 0.
    - Cột '*male_rich*': những người nam giới có giá vé, cột 'Fare', cao hơn 50.
    - Cột '*Ticket_freq*': đếm tần suất vé giống nhau
  - *Feature extraction:* giống như lần 4

- *Các thuật toán và tham số tương ứng:* giống như lần 4

- *K-fold:* giống như lần 4

- *Score matric:* giống như lần 4

- Thuật toán VotingClassifier:
  - estimators=[
    ('rf, best_models.get('RandomForestClassifier')),
    ('xgb', best_models.get('XGBClassifier')),
    ('svc', best_models.get('SVC')),
    ('lgb', best_models.get('LGBMClassifier'))
    ]

  - voting='soft',

Kết quả lần thử nghiệm 5 (V17-Kaggle)	
Train score	*0.7773*
Submit score	*0.78229*

Nhận xét:

- Điểm đã được cải thiện lên 1% cho thấy VotingClassifier đã có hiệu quả.
- Các đặc trưng xây dựng như 'IsAlone', 'Ticket_freg', 'male_rich',... là những đặc trưng tốt, đóng góp vào kết quả dự đoán của model

IV. Kết quả và đánh giá

Sau nhiều lần tinh chỉnh và cải tiến mô hình, ta thu được kết quả với số điểm ***0.78229***

Kết quả lần thử nghiệm 5 (V17-Kaggle)	
Train score	*0.7773*
Submit score	*0.78229*

Đánh giá:

- Mô hình đã cho ra dự đoán đúng khoảng 80%
- Mới chỉ khai thác và phát hiện được một phần thông in bên trong dữ liệu.
- Chưa hiểu rõ được lịch sử tại thời gian lúc tàu Titanic gặp tai nạn và chưa hiểu rõ được văn hóa lúc đó để đánh giá và liên kết dữ liệu tốt hơn.
- Các mô hình và phương pháp làm còn đơn sơ và chưa đi sâu vào tinh chỉnh mô hình phù hợp.

VI. Hướng cải tiến mô hình

Hướng 1: Hiểu rõ dữ liệu, lịch sử và văn hóa tại thời gian lúc tàu Titanic gặp tai nạn và cấu tạo của tàu, nhằm:

- Xử lí dữ liệu phù hợp hơn
  - Cách điền giá trị thiếu thông minh hơn
  - Chuẩn hóa dữ liệu
  - Rời rạc hóa dữ liệu
- Thiết kế đặc trưng tốt hơn
  - Dùng PCA, T-sne,... để trích xuất
  - Dùng Chi2, ANOVA,.. để lựa chọn những đặc trưng phù hợp

Hướng 2: Phân tích sâu hơn và đánh giá chính xác hơn từng mô hình, nhằm:

- Lựa chọn được mô hình phù hợp với dữ liệu và phương pháp đang sử dụng:
  - Logistic Regression
  - Genetic Algorithms
- Tinh chỉnh sâu vào mô hình
- Thử nghiệm các mô hình mới.

Hướng 3: **Sử dụng phương pháp mới.**

VII. Tham khảo

[1] Kaggle Titanic Alexis Cook Titanic Tutorial
[2] A data science framwork