Titanic - Machine Learning from Disaster

Biện Hữu Khang, Nguyễn Thị Thùy Linh, Mai Hoàng Hải Yến

Trường Đại học Sài Gòn, Việt Nam

Tóm tắt

Vụ đắm tàu Titanic năm 1912 là một trong những thảm họa hàng hải nghiêm trọng nhất trong lịch sử, dẫn đến hơn 1.500 người thiệt mạng. Việc phân tích dữ liệu hành khách của con tàu này đã trở thành một ví dụ kinh điển trong lĩnh vực học máy, giúp các nhà nghiên cứu kiểm chứng các thuật toán phân loại và dự đoán.

Nghiên cứu này tập trung vào việc dự đoán khả năng sống sót của hành khách dựa trên dữ liệu thực tế của Titanic (891 mẫu huấn luyện) bằng hai mô hình học máy phổ biến: Hồi quy Logistic (Logistic Regression) và Rừng ngẫu nhiên (Random Forest). Quy trình thực hiện bao gồm: phân tích dữ liệu khám phá (EDA) để xác định các yếu tố ảnh hưởng, xử lý giá trị thiếu (điền trung vị cho 'Age', loại bỏ 'Cabin' và 'Embarked'), kỹ thuật đặc trưng (tạo 'FamilySize'), mã hóa đặc trưng (One-Hot cho 'Pclass', Map cho 'Sex'), và đánh giá mô hình bằng các chỉ số chuẩn như Accuracy, Precision, Recall, và F1-score trên tập kiểm thử (20% dữ liệu).

Kết quả thực nghiệm cho thấy mô hình Random Forest đạt độ chính xác 81.01%, cao hơn so với Logistic Regression (79.89%). Mô hình Random Forest cũng thể hiện hiệu năng cân bằng hơn, đặc biệt là chỉ số Recall (0.76) cho lớp 'sống sót'. Phân tích độ quan trọng của đặc trưng xác nhận 'Fare' (Giá vé), 'Sex' (Giới tính) và 'Age' (Tuổi) là các yếu tố dự đoán mạnh nhất. Kết quả này khẳng định khả năng ứng dụng của các thuật toán phân loại truyền thống trong bài toán dữ liệu nhỏ và không quá phức tạp.

Từ khóa: Titanic, học máy, Logistic Regression, Random Forest, dự đoán nhị phân, EDA, kỹ thuật đặc trưng.

1. Giới thiệu

Vụ chìm tàu RMS Titanic vào năm 1912 là một trong những thảm họa hàng hải đáng nhớ nhất, lấy đi sinh mạng của hàng ngàn người. Dữ liệu chi tiết về hành khách và thủy thủ đoàn trên con tàu này đã được sử dụng rộng rãi như một bài toán kinh điển trong lĩnh vực khoa học dữ liệu. Nền tảng Kaggle đã cung cấp bộ dữ liệu này để khuyến khích cộng đồng học máy thử nghiệm và phát triển các thuật toán phân loại. Với kích thước vừa phải và chứa nhiều loại thông tin khác nhau (số, chữ, danh mục), bài toán Titanic trở thành một môi trường lý tưởng để minh họa toàn bộ quy trình của học máy có giám sát (supervised learning).

Mục tiêu chính của bài toán là dự đoán xem một hành khách có sống sót (Survived = 1) hay không (Survived = 0) dựa trên các thông tin như giới tính (Sex), tuổi (Age), hạng vé (Pclass), số người thân (SibSp, Parch), và cảng lên tàu (Embarked).

Các phân tích và nghiên cứu trước đây đã chỉ ra rằng giới tính, hạng vé, và tuổi là những yếu tố có ảnh hưởng đáng kể đến khả năng sống sót. Điều này thường phản ánh quy tắc ưu tiên "phụ nữ và trẻ em trước" được áp dụng trong các tình huống khẩn cấp. Tuy nhiên, để hiểu rõ mức độ tác động của từng yếu tố và xây dựng một mô hình dự đoán đáng tin cậy, chúng ta cần một quy trình xử lý dữ liệu tỉ mỉ và một phương pháp đánh giá mô hình định lượng.

Trong nghiên cứu này, chúng em tập trung triển khai hai mô hình học máy được sử dụng phổ biến, dễ hiểu và có khả năng hoạt động tốt trên các dữ liệu mới:

1. Hồi quy Logistic: Đây là một mô hình xác suất tuyến tính, thường được dùng để ước tính xác suất xảy ra của một sự kiện (trong trường hợp này là khả năng sống sót).
2. Rừng ngẫu nhiên (Random Forest): Đây là một mô hình mạnh mẽ hơn, thuộc nhóm "ensemble" (kết hợp nhiều mô hình nhỏ). Nó xây dựng nhiều cây quyết định và tổng hợp kết quả của chúng để đưa ra dự đoán cuối cùng, giúp tăng tính ổn định và giảm lỗi.

Mục tiêu: So sánh hiệu suất của hai mô hình này trên cùng bộ dữ liệu đã được xử lý và từ đó đưa ra nhận định về ưu nhược điểm cũng như khả năng ứng dụng thực tế của từng phương pháp.

2. Cơ sở lý thuyết và phương pháp nghiên cứu

2.1. Hồi quy Logistic (Logistic Regression)

Hồi quy Logistic là mô hình xác suất tuyến tính được sử dụng phổ biến cho bài toán phân loại nhị phân. Mô hình này không dự đoán trực tiếp kết quả 0 hay 1, mà dự đoán xác suất $P(y = 1/x)$ (xác suất mẫu x thuộc về lớp 1). Mô hình dựa trên hàm

sigmoid (hoặc logistic) để ánh xạ tổ hợp tuyến tính của các đặc trưng đầu vào vào một giá trị trong khoảng [0, 1]:

$$P(y = 1/x) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Trong đó, z là tổ hợp tuyến tính của các đặc trưng:

$$z = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Ưu điểm của mô hình là đơn giản, huấn luyện nhanh và kết quả dễ giải thích (thông qua các trọng số β_i). Tuy nhiên, nó giả định mối quan hệ tuyến tính giữa các đặc trưng và log-odds của kết quả, do đó có thể hoạt động kém hiệu quả khi dữ liệu có các mối quan hệ phi tuyến phức tạp.

2.2. Rừng ngẫu nhiên (Random Forest)

Random Forest là một mô hình học máy ensemble (tổng hợp) dựa trên kỹ thuật "bagging" (Bootstrap Aggregating) và các cây quyết định (Decision Trees). Thay vì chỉ xây dựng một cây quyết định duy nhất (dễ bị overfitting), Random Forest xây dựng m cây quyết định khác nhau.

Mỗi cây trong rừng được huấn luyện trên một tập dữ liệu con được lấy mẫu ngẫu nhiên có lặp lại (bootstrap) từ tập huấn luyện gốc. Ngoài ra, tại mỗi nút của cây, thay vì xem xét tất cả các đặc trưng, mô hình chỉ chọn ngẫu nhiên một tập con các đặc trưng để tìm ra phép chia tốt nhất.

Kết quả dự đoán cuối cùng được xác định bằng cách lấy đa số phiếu (majority voting) từ tất cả các cây:

$$\hat{y} = mode\{T_1(x), T_2(x), \ldots, T_m(x)\}$$

Trong đó $T_m(x)$ là dự đoán của cây thứ m. Mô hình này rất mạnh mẽ trong việc nắm bắt các mối quan hệ phi tuyến, có khả năng chống overfitting tốt và cung cấp một thước đo quan trọng về độ ảnh hưởng của từng đặc trưng (feature importance).

2.3. Quy trình nghiên cứu

Nghiên cứu được tiến hành qua 5 bước chính, bám sát quy trình được thực hiện trong notebook `Titanic.ipynb`:

1. Thu thập và Mô tả dữ liệu: Tải và kiểm tra thông tin cơ bản của `train.csv` (891 mẫu, 12 biến) và `test.csv`.
2. Phân tích dữ liệu Khám phá (EDA): Xác định dữ liệu thiếu, dữ liệu trùng lặp, phân tích thống kê mô tả và trực quan hóa các mối liên hệ.
3. Tiền xử lý và Kỹ thuật đặc trưng: Xử lý giá trị thiếu, mã hóa biến phân loại và tạo đặc trưng mới.

4. Huấn luyện mô hình: Chia dữ liệu (80/20) và huấn luyện Logistic Regression và Random Forest.
5. Đánh giá và Dự đoán: Hiệu suất của các mô hình đã huấn luyện được đánh giá trên tập kiểm tra bằng các chỉ số chuẩn như Accuracy, Precision, Recall và F1-score. Mô hình tốt nhất sẽ được sử dụng để đưa ra dự đoán cuối cùng trên tập `test.csv`

3. Dữ liệu và xử lý

3.1. Mô tả dữ liệu

Bộ dữ liệu `train.csv` này chứa 891 bản ghi (mỗi bản ghi đại diện cho một hành khách) và 12 cột thông tin khác nhau. Biến mục tiêu mà chúng ta cần dự đoán là 'Survived', với giá trị 1 nếu hành khách sống sót và 0 nếu không.

Bảng 1: Data Dictionary

Biến	Mô tả	Ghi chú
Survived	Trạng thái sống sót của hành khách	1 = sống sót, 0 = không sống sót (biến mục tiêu)
Pclass	Hạng vé	1 = Hạng nhất, 2 = Hạng hai, 3 = Hạng ba
Sex	Giới tính	male, female
Age	Tuổi	Tính theo năm
SibSp	Số anh chị em/vợ chồng đi cùng	
Parch	Số cha mẹ/con đi cùng	
Ticket	Số vé	
Fare	Giá vé	
Cabin	Số phòng/cabin	
Embarked	Cảng lên tàu	C = Cherbourg, Q = Queenstown, S = Southampton

3.2. Phân tích dữ liệu khám phá (EDA)

3.2.1. Chất lượng dữ liệu

- Giá trị trùng lặp: Kiểm tra bằng `duplicated().sum()` cho thấy không có hàng nào bị trùng lặp hoàn toàn, đảm bảo tính duy nhất của mỗi quan sát.
- Giá trị thiếu: Kiểm tra bằng `isnull().sum()` cho thấy 3 cột có dữ liệu bị thiếu: 'Age' (177 giá trị), 'Cabin' (687 giá trị), và 'Embarked' (2 giá trị). Cột 'Cabin' bị thiếu quá 77% dữ liệu, đây là một tỷ lệ rất lớn.

3.2.2. Phân tích thống kê mô tả

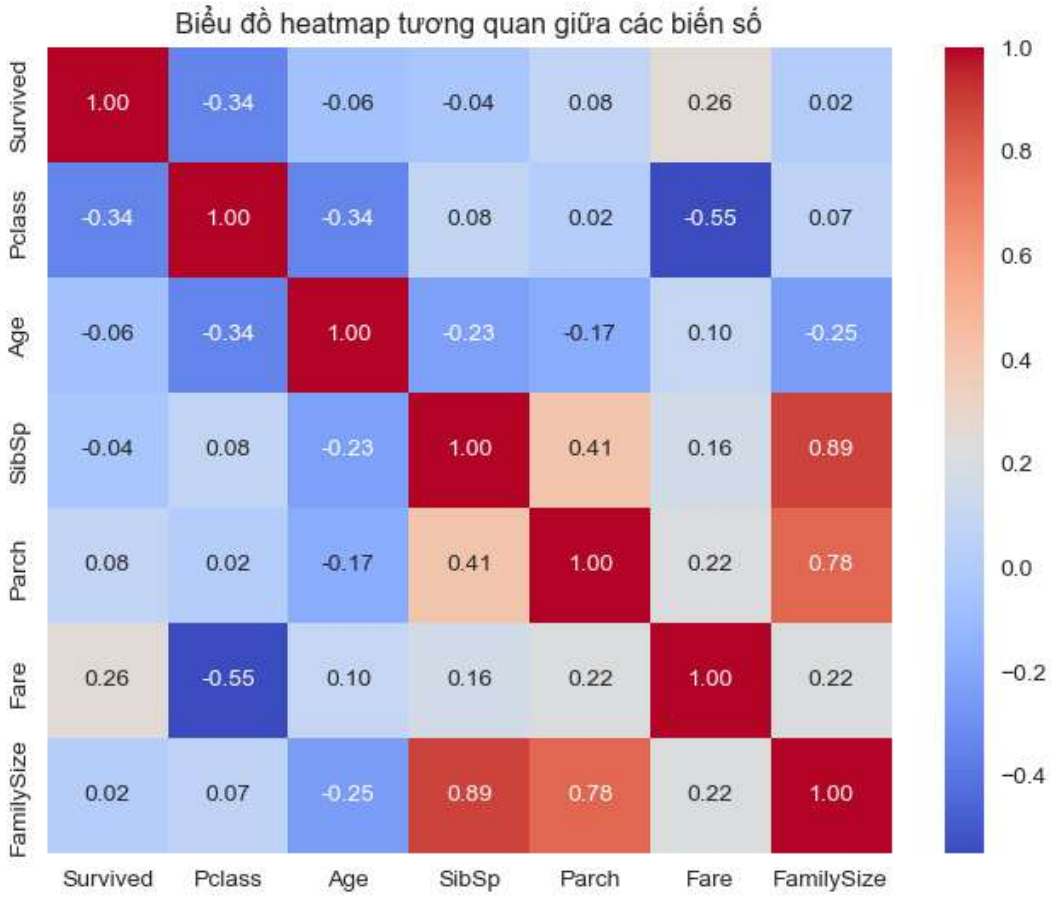
Kết quả từ `describe()` cung cấp cái nhìn tổng quan về các biến số định lượng:

- Survived: Giá trị trung bình là 0.384. Điều này có nghĩa là chỉ có khoảng 38.4% hành khách trong tập huấn luyện sống sót, cho thấy một sự mất cân bằng giữa hai lớp (sống sót/không sống sót).
- Pclass: Giá trị trung bình là 2.3. Điều này ngụ ý rằng đa số hành khách thuộc hạng vé 2 và 3, với số lượng hành khách hạng 1 ít hơn.
- Age: Tuổi của hành khách dao động từ trẻ sơ sinh (0.42 tuổi) đến người già (80 tuổi), với độ lệch chuẩn khá lớn (13.0), cho thấy sự đa dạng về độ tuổi.
- Fare: Giá vé có độ lệch chuẩn cực kỳ cao (49.69) và giá trị tối đa lên tới 512.3. Sự khác biệt lớn giữa giá trị trung bình (32.2) và giá trị tối đa, cùng với phân vị 75 (31.0), cho thấy phân bố giá vé bị lệch phải nặng, với một số ít hành khách trả giá rất cao.

3.2.3. Trực quan hóa và Phân tích tương quan

Phân tích trực quan được thực hiện để tìm hiểu mối quan hệ giữa các đặc trưng và khả năng sống sót:

- Phân bố biến số (Histograms): Các biến ‘Age’, ‘Fare, ‘SibSp’, ‘Parch và ‘FamilySize’ đều cho thấy phân bố lệch phải.
- Biến phân loại (Count Plots):
    - Survived vs. Sex: Biểu đồ cho thấy nữ giới có tỷ lệ sống sót cao hơn hẳn nam giới.
    - Survived vs. Pclass: Biểu đồ cho thấy hành khách hạng 3 (Pclass=3) có tỷ lệ tử vong cao nhất, trong khi hành khách hạng 1 có tỷ lệ sống sót cao nhất.
- Tuổi (Age) và khả năng sống sót (KDE Plot): Biểu đồ mật độ ước tính hạt nhân (KDE Plot) so sánh phân bố tuổi giữa nhóm người sống sót và nhóm người không sống sót. Kết quả cho thấy tỷ lệ tử vong cao nhất tập trung ở nhóm tuổi 20 - 40. Đáng chú ý, trẻ em (đặc biệt dưới 12 tuổi) có tỷ lệ sống sót cao hơn rõ rệt, phù hợp với quy tắc ưu tiên.
- Ma trận tương quan (Heatmap): Biểu đồ heatmap thể hiện mối tương quan tuyến tính giữa tất cả các biến số định lượng. ‘Survived’ có tương quan âm đáng kể với ‘Pclass’ (-0.34) và tương quan dương đáng kể với ‘Fare’ (0.26). Điều này củng cố giả thuyết rằng hạng vé và giá vé là những yếu tố mạnh mẽ ảnh hưởng đến khả năng sống sót.



Hình 1: Biểu đồ heatmap tương quan giữa các biến số

3.3. Tiền xử lý và Kỹ thuật đặc trưng

Dựa trên EDA, các bước tiền xử lý và kỹ thuật đặc trưng sau được áp dụng:

1. Xử lý giá trị thiếu:
    - Age (Tuổi): Các giá trị thiếu được điền bằng trung vị (median) của cột 'Age' (là 28.0). Phương pháp này ít bị ảnh hưởng bởi các giá trị ngoại lai hơn so với giá trị trung bình. Sau đó, kiểu dữ liệu của cột 'Age' được chuyển đổi sang số nguyên (int).
    - Cabin (Số phòng): Loại bỏ cột này do thiếu quá nhiều dữ liệu.
    - Embarked (Cảng lên tàu): Loại bỏ cột này.
2. Kỹ thuật đặc trưng:
    - Tạo biến 'FamilySize'= SibSp + Parch + 1.
3. Mã hóa biến:
    - Sex (Giới tính): Mã hóa thành số (male: 0, female: 1).
    - Pclass (Hạng vé): Mã hóa One-Hot Encoding, tạo ra các cột 'Pclass_2' và 'Pclass_3'.
4. Chuẩn hóa:
    - Các biến số 'Age', 'Fare', 'FamilySize' được chuẩn hóa bằng `StandardScaler`.

5. Lựa chọn đặc trưng:
    - Các cột không cần thiết (‘Name’, ‘Ticket’, ‘SibSp’, ‘Parch’, ‘PassengerId’, ‘AgeGroup’) bị loại bỏ.
    - Các đặc trưng cuối cùng được chọn để huấn luyện là: `Age`, `Fare`, `FamilySize`, `Sex`, `Pclass_2`, `Pclass_3`.

4. Thực nghiệm và kết quả

4.1. Huấn luyện mô hình

Dữ liệu đặc trưng (X) và biến mục tiêu (y) được chia thành tập huấn luyện (80%) và tập kiểm thử (validation) (20%) bằng `train_test_split` với `random_state=42`. Các mô hình được huấn luyện trên tập `X_train` và `y_train`, sau đó đánh giá trên tập `X_val` và `y_val`.

Các chỉ số đánh giá chính bao gồm:

- Accuracy: Tỷ lệ dự đoán đúng.
- Confusion Matrix: Ma trận nhầm lẫn (TN, FP, FN, TP).
- Classification Report: Bao gồm Precision, Recall, và F1-score cho từng lớp.

4.2. Kết quả mô hình Logistic Regression

Mô hình `LogisticRegression` được huấn luyện. Kết quả đánh giá trên tập kiểm thử (179 mẫu) được trình bày trong Bảng 2.

Bảng 2: Kết quả Logistic Regression

Chỉ số	Lớp 0 (Không sống)	Lớp 1 (Sống sót)	Tổng/Trung bình
Precision	0.80	0.80	Accuracy: 0.80
Recall	0.88	0.69	Macro Avg: 0.78
F1-score	0.84	0.74	Weighted Avg: 0.79
Support	105	74	179

- Ma trận nhầm lẫn:
    - True Negative (Dự đoán đúng không sống): 92
    - False Positive (Dự đoán sai thành sống): 13
    - False Negative (Dự đoán sai thành không sống): 23
    - True Positive (Dự đoán đúng sống): 51

Mô hình Hồi quy Logistic đạt độ chính xác 79.89%. Phân tích sâu hơn cho thấy mô hình này hoạt động khá tốt trong việc dự đoán hành khách không sống sót (Recall = 0.88 cho Lớp 0). Tuy nhiên, nó có xu hướng bỏ sót nhiều trường hợp hành khách thực sự sống sót (Recall = 0.69 cho Lớp 1), dẫn đến số False Negative cao hơn.

4.3. Kết quả mô hình Random Forest

Mô hình Random Forest, với 100 cây quyết định (`n_estimators=100`), được huấn luyện trên tập huấn luyện. Kết quả đánh giá chi tiết trên tập kiểm thử được trình bày trong Bảng 3.

Bảng 3: Kết quả Random Forest

Chỉ số	Lớp 0 (Không sống)	Lớp 1 (Sống sót)	Tổng/Trung bình
Precision	0.83	0.78	Accuracy: 0.81
Recall	0.85	0.76	Macro Avg: 0.80
F1-score	0.84	0.77	Weighted Avg: 0.81
Support	105	74	179

- Ma trận nhầm lẫn:
  - True Negative (Dự đoán đúng không sống): 89
  - False Positive (Dự đoán sai thành sống): 16
  - False Negative (Dự đoán sai thành không sống): 18
  - True Positive (Dự đoán đúng sống): 56

Mô hình Random Forest đạt độ chính xác 81.01%. So với Hồi quy Logistic, mô hình này không chỉ có độ chính xác tổng thể cao hơn mà còn cho thấy hiệu suất cân bằng hơn giữa hai lớp. Cụ thể, Recall cho Lớp 1 (sống sót) đã cải thiện đáng kể lên 0.76, cho thấy mô hình này ít bỏ sót các trường hợp sống sót hơn.

4.4. So sánh tổng hợp các mô hình

Bảng 4 tóm tắt và so sánh các chỉ số hiệu suất quan trọng của cả hai mô hình, tập trung vào độ chính xác tổng thể và khả năng dự đoán lớp 1 (sống sót) – vốn là lớp thiểu số trong bài toán này.

Bảng 4: So sánh hiệu suất hai mô hình

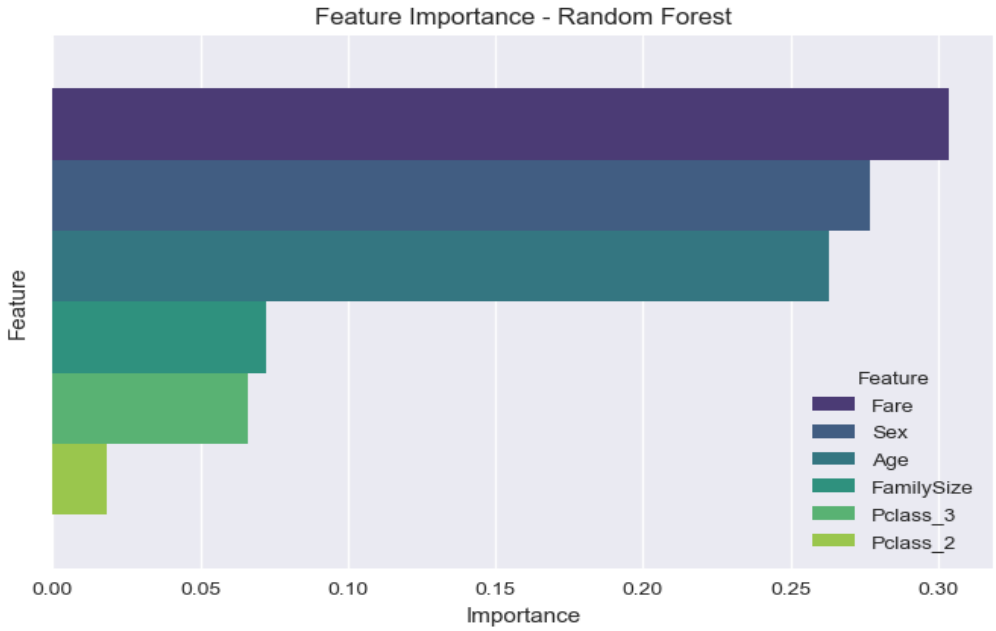
Mô hình	Accuracy	Recall (Lớp 1)	F1-score (Lớp 1)
Logistic Regression	79.9%	0.69	0.74
Random Forest	81.0%	0.76	0.77

Kết quả so sánh rõ ràng cho thấy mô hình Random Forest mang lại hiệu suất tổng thể tốt hơn trên bộ dữ liệu này, đặc biệt là trong việc dự đoán chính xác các trường hợp sống sót.

5. Thảo luận

5.1. Phân tích độ quan trọng của đặc trưng

Một trong những ưu điểm nổi bật của mô hình Random Forest là khả năng ước tính và cung cấp độ quan trọng (feature importance) của từng đặc trưng. Điều này giúp chúng ta hiểu rõ hơn về những yếu tố nào đóng góp nhiều nhất vào quyết định dự đoán của mô hình.



Hình 2: Biểu đồ độ quan trọng của đặc trưng từ mô hình Random Forest

Kết quả từ hình trên cho thấy thứ tự quan trọng của các đặc trưng được sử dụng như sau:

1. Fare (Giá vé): Đây là yếu tố quan trọng nhất. Biểu đồ cho thấy ‘Fare’ có mức ảnh hưởng (Importance) cao vượt trội (khoảng 0.305). Điều này xác nhận mạnh mẽ giả thuyết từ EDA rằng điều kiện kinh tế (phản ánh qua giá vé) là yếu tố dự đoán hàng đầu cho sự sống sót.
2. Sex (Giới tính): Là yếu tố quan trọng thứ hai (khoảng 0.27). Mặc dù EDA cho thấy đây là yếu tố phân biệt rất rõ ràng (tỷ lệ nữ sống sót cao hơn nhiều), tầm quan trọng của nó trong mô hình chỉ đứng sau ‘Fare’.
3. Age (Tuổi): Là yếu tố quan trọng thứ ba (khoảng 0.26), có mức ảnh hưởng gần bằng với ‘Sex’. Tuổi tác, như đã phân tích trong EDA, có vai trò rõ rệt, phù hợp với quy tắc ưu tiên "phụ nữ và trẻ em" trong lịch sử.

4. FamilySize (Kích thước gia đình): Yếu tố này có ảnh hưởng ở mức độ thấp hơn đáng kể (khoảng 0.075).
5. Pclass_3 (Hạng vé 3): Có ảnh hưởng thấp (khoảng 0.065).
6. Pclass_2 (Hạng vé 2): Có ảnh hưởng thấp nhất trong các đặc trưng (khoảng 0.02).

Kết luận: Kết quả này xác nhận rằng điều kiện kinh tế (phản ánh qua Giá vé) và các yếu tố nhân khẩu học cơ bản (Giới tính, Tuổi) là ba yếu tố dự đoán mạnh mẽ nhất theo mô hình Random Forest này. Các đặc trưng về Hạng vé (‘Pclass’) có thể có ảnh hưởng thấp hơn vì thông tin của chúng đã bị "hấp thụ" phần lớn bởi đặc trưng ‘Fare’ (Giá vé), vốn có tương quan cao với ‘Pclass’.

5.2. Hạn chế của nghiên cứu

Mặc dù đạt kết quả khả quan, nghiên cứu này có một số hạn chế:

- Việc xử lý dữ liệu thiếu còn đơn giản, chỉ dừng lại ở việc điền trung vị (median) cho ‘Age’ và loại bỏ hoàn toàn cột ‘Embarked’.
- Kỹ thuật đặc trưng còn hạn chế khi chỉ tạo ra biến ‘FamilySize’ , bỏ qua các thông tin tiềm năng khác như ‘Title’ (chức danh) từ cột ‘Name’.
- Các mô hình được huấn luyện với tham số mặc định , chưa qua tinh chỉnh siêu tham số (Hyperparameter Tuning) để tối ưu hiệu suất.
- Vấn đề mất cân bằng dữ liệu (chỉ 38.4% sống sót) chưa được xử lý bằng các kỹ thuật chuyên biệt như SMOTE hay Class Weight.

5.3. Hướng phát triển tương lai

Dựa trên các hạn chế đã nêu, hướng phát triển tương lai sẽ tập trung vào việc cải thiện đặc trưng và mô hình:

- Về dữ liệu: Áp dụng kỹ thuật điền dữ liệu (imputation) nâng cao cho ‘Age’, đồng thời thử nghiệm giữ lại và mã hóa cột ‘Embarked’.
- Về đặc trưng: Trích xuất và sử dụng đặc trưng Title’ từ cột ‘Name’ để tăng cường thông tin cho mô hình.
- Về mô hình: Áp dụng các thuật toán mạnh mẽ hơn như XGBoost, LightGBM hoặc Mạng Nơ-ron đơn giản để so sánh và cải thiện kết quả dự đoán.

6. Kết luận

Nghiên cứu này đã triển khai thành công một quy trình phân tích và học máy cơ bản để dự đoán khả năng sống sót của hành khách trên tàu Titanic. Qua các bước từ

phân tích dữ liệu khám phá đến tiền xử lý và huấn luyện mô hình, chúng em đã đánh giá hiệu suất của hai thuật toán phân loại phổ biến.

Kết quả so sánh cho thấy mô hình Rừng ngẫu nhiên (Random Forest) mang lại hiệu suất dự đoán cao hơn (Độ chính xác 81.01%) và khả năng cân bằng hơn trong việc xác định cả hai nhóm hành khách (sống sót và không sống sót) so với mô hình Hồi quy Logistic (Logistic Regression) (Độ chính xác 79.89%) trên tập dữ liệu này. Phân tích độ quan trọng của đặc trưng đã xác nhận rằng Giới tính, Giá vé và Tuổi là những yếu tố then chốt, có ảnh hưởng mạnh mẽ nhất đến sự sống còn của hành khách.

Nghiên cứu này khẳng định rằng, ngay cả với tập dữ liệu tương đối nhỏ và không quá phức tạp, các mô hình học máy truyền thống vẫn có thể đạt được hiệu năng đáng kể nếu quy trình tiền xử lý và kỹ thuật đặc trưng được thực hiện một cách cẩn thận và phù hợp. Dựa trên hiệu suất vượt trội, mô hình Random Forest đã được lựa chọn để thực hiện dự đoán trên tập `test.csv` và tạo ra tệp `submission.csv` cuối cùng cho cuộc thi.

Tài liệu tham khảo

[1] Kaggle. (2012). *Titanic: Machine Learning from Disaster* [Data set]. Kaggle. https://www.kaggle.com/c/titanic

[2] Scikit-learn Developers. (2023). *Scikit-learn: Machine learning in Python* [Computer software]. https://scikit-learn.org

[3] Géron, A. (2023). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

[6] Bruce, P., Bruce, A., & Gedeck, P. (2022). *Practical statistics for data scientists* (2nd ed.). O'Reilly Media.