PROJECT: Titanic - Machine Learning from Disaster

Tran Nguyen Minh Tien-3123580052

Nguyen Gia Huy-3123500016

Ho Gia Bao-3123580003

Saigon University, Vietnam.

**Tóm tắt dự án**

Cuộc thi “Titanic: Machine Learning from Disaster” là một bài toán điển hình trong lĩnh vực phân tích dữ liệu và học máy, với mục tiêu dự đoán khả năng sống sót của hành khách trên con tàu RMS Titanic – con tàu nổi tiếng bị đắm trong lần đầu tiên rời cảng vào năm 1912. Dữ liệu bao gồm thông tin về người đi tàu như giới tính, tuổi tác, hạng vé, số anh chị/em/phụ huynh đi cùng, cảng lên tàu… Các thí sinh được cung cấp tập “train” có nhãn sống sót và tập “test” không có nhãn, yêu cầu xây dựng mô hình phân loại để dự đoán biến “Survived”. Bài toán không chỉ giúp rèn luyện kỹ năng tiền xử lý dữ liệu và trực quan hóa mà còn yêu cầu phân tích đặc trưng, xử lý giá trị thiếu và áp dụng các thuật toán học máy thích hợp để tăng độ chính xác dự đoán. Kết quả của cuộc thi khai thác được các yếu tố như giới tính (nữ giới có tỷ lệ sống sót cao hơn), hạng vé (hạng 1 có lợi thế hơn) và độ tuổi trẻ em, từ đó góp phần hiểu rõ hơn động lực đằng sau khả năng sống sót của hành khách.

1. **GIỚI THIỆU**

Thảm họa tàu Titanic là một trong những sự kiện nổi tiếng nhất trong lịch sử hàng hải. Dựa trên dữ liệu hành khách, ta có thể tiến hành các phân tích nhằm tìm ra những yếu tố ảnh hưởng đến khả năng sống sót.  
Cuộc thi **“Titanic: Machine Learning from Disaster”** trên nền tảng **Kaggle** là một bài toán kinh điển trong lĩnh vực **phân tích dữ liệu và học máy**, với mục tiêu dự đoán hành khách nào có khả năng sống sót trên con tàu RMS Titanic — con tàu đã bị chìm vào năm 1912 trong chuyến đi đầu tiên.

Bộ dữ liệu bao gồm thông tin về hành khách như giới tính, tuổi, hạng vé, số thân nhân đi cùng, giá vé và cảng lên tàu. Thí sinh được cung cấp tập huấn luyện có nhãn “Survived” và tập kiểm tra không nhãn, yêu cầu xây dựng mô hình học máy để dự đoán khả năng sống sót.  
Bài toán này giúp rèn luyện kỹ năng tiền xử lý dữ liệu, trực quan hóa, phân tích đặc trưng, xử lý giá trị thiếu và lựa chọn mô hình học máy phù hợp.

Kết quả của các phân tích trước đây cho thấy **giới tính**, **hạng vé** và **độ tuổi** là ba yếu tố có ảnh hưởng mạnh nhất đến khả năng sống sót: nữ giới và trẻ em ở hạng vé cao có tỷ lệ sống sót cao hơn đáng kể.  
Trong nghiên cứu này, chúng tôi tiến hành phân tích lại dữ liệu Titanic nhằm minh họa quy trình xử lý, trực quan hóa và rút ra kết luận từ dữ liệu một cách khoa học.

1. **Phương pháp đề xuất**
   1. **Tổng quan về vấn đề**

Vấn đề phân tích dữ liệu Titanic được xem là bài toán điển hình trong học máy nhị phân, với mục tiêu dự đoán biến mục tiêu *Survived* (1 = sống sót, 0 = không sống sót) dựa trên các đặc trưng nhân khẩu học và vé tàu của hành khách.

Tổng quan cho thấy, các yếu tố chính ảnh hưởng đến khả năng sống sót bao gồm:

* **Giới tính (Sex):** Nữ giới có cơ hội sống sót cao hơn do quy định “phụ nữ và trẻ em được ưu tiên” trong quá trình sơ tán.
* **Hạng vé (Pclass):** Hành khách ở hạng nhất có điều kiện tiếp cận thuyền cứu sinh sớm hơn.
* **Độ tuổi (Age):** Trẻ em có tỷ lệ sống sót cao hơn người trưởng thành.
* **Giá vé (Fare)** và **Cảng lên tàu (Embarked):** Có thể phản ánh tình trạng kinh tế và vị trí lên tàu, ảnh hưởng đến cơ hội thoát hiểm.

Phương pháp được đề xuất trong nghiên cứu này bao gồm:

1. **Tiền xử lý dữ liệu:** Kiểm tra giá trị thiếu, xử lý ngoại lệ, mã hóa các biến phân loại.
2. **Phân tích mô tả:** Trực quan hóa mối quan hệ giữa các biến và trạng thái sống sót.
3. **Xây dựng mô hình dự đoán:** Áp dụng các thuật toán học máy cơ bản (như Logistic Regression, Decision Tree) để kiểm chứng mối quan hệ.
4. **Đánh giá mô hình:** Sử dụng độ chính xác (Accuracy), F1-score và ma trận nhầm lẫn (Confusion Matrix) để đo lường hiệu quả dự đoán.
   1. **Mô hình đề xuất**

Mô hình được xây dựng nhằm dự đoán khả năng sống sót của hành khách trên tàu Titanic, dựa trên các đặc trưng quan sát được trong bộ dữ liệu gốc của Kaggle.  
Quy trình xây dựng mô hình bao gồm các giai đoạn sau:

**2.3.1. Tiền xử lý dữ liệu**

--- Tình trạng giá trị thiếu (trước khi fix) ---

Survived 0

Pclass 0

Name 0

Sex 0

Age 177

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 687

Embarked 2

dtype: int64

--------------------------------------------------

* **Xử lý giá trị thiếu:** Cột *Age* được điền bằng giá trị trung bình theo nhóm *Pclass* và *Sex*; cột *Embarked* được thay bằng giá trị xuất hiện nhiều nhất; *Fare* được thay bằng trung vị.

--- Tình trạng giá trị thiếu (sau khi fix Age và Embarked) ---

Survived 0

Pclass 0

Name 0

Sex 0

Age 0

SibSp 0

Parch 0

Ticket 0

Fare 0

Embarked 0

dtype: int64

--------------------------------------------------

* **Mã hóa biến phân loại:** Các biến như *Sex* và *Embarked* được chuyển sang dạng số bằng *Label Encoding* hoặc *One-Hot Encoding*.

Survived category

Pclass category

Name object

Sex category

Age float64

SibSp category

Parch category

Ticket object

Fare float64

Cabin object

Embarked category

dtype: object

* **Chuẩn hóa dữ liệu:** Các đặc trưng liên tục (Age, Fare) được chuẩn hóa nhằm đảm bảo tính ổn định trong huấn luyện mô hình.

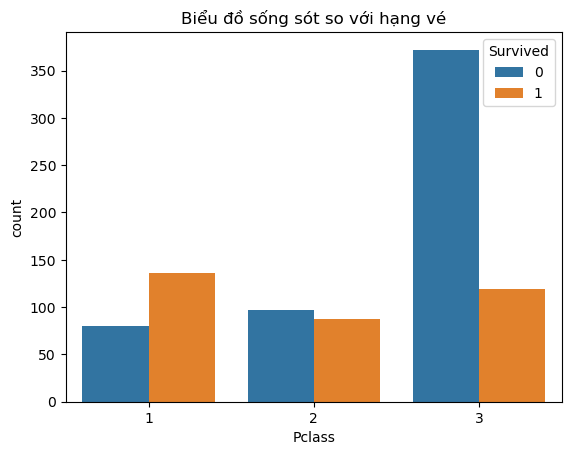
**2.3.2. Phân tích tương quan và lựa chọn đặc trưng**

Các biến đầu vào được phân tích tương quan với biến mục tiêu *Survived* bằng biểu đồ heatmap. Kết quả cho thấy:

* *Sex* có mối tương quan mạnh nhất với *Survived* (nữ giới có tỷ lệ sống sót cao hơn).

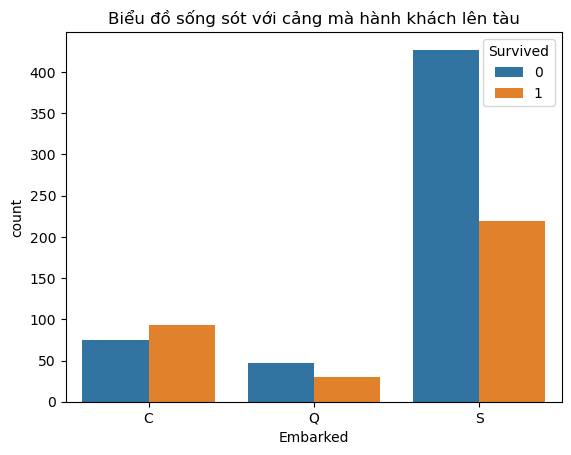
Nhận xét: Nhìn vào hình ta có thể thấy đối với nữ tỉ lệ sống sót cao hơn nhiều so với nam , tỉ lệ sống sót của nam chỉ bằng 1 nửa so với nữ. Đều này nói lên rằng khi sự cố xẩy ra người ta ưu tiên phụ nữ và trẻ em lên thuyền cứu sinh trước nên tỉ lên sống sót của nữ cao hơn của nam giới.

* *Pclass* và *Fare* phản ánh chênh lệch xã hội, ảnh hưởng rõ rệt đến khả năng sống sót.



Nhận xét: Tiếp đến là tỉ lệ sống sót so với hạng vé lên tàu, có thể thấy trong biểu đồ là hạng vé 1 tỉ lệ sống sót cao hơn tỉ lệ tử vong điều này minh chứng rằng người đi vé hạng thương gia sẽ được ưu tiên lên thuyền cứu sinh , với hạng vé 2 thì tỉ lệ sống sót và tỉ lệ tử vong khá là ngang bằng nhau nhưng tỉ lệ tử vong lại nhỉnh hơn chút còn đối với hạng vé 3 thì tỉ lệ tử vong đạt mức cao nhất gần gấp 3 lần tỉ lệ sống sót

* *Age* và *Embarked* thể hiện ảnh hưởng trung bình.



Nhận xét: Dựa vào biểu đồ tỉ lệ sống sót so với cảng mà hành khách lên tàu ta có thể thấy cảng thứ nhất là cảng Cherbourg , thứ hai là cảng Queenstown và thứ ba là cảng Southampton thì có thể thấy lượng hành khách lên tàu đến từ cảng S lớn nhất trong các cảng mà đa phần hành khách tử vong nhiều hơn so với tỉ lệ sống sót, đối với cảng C thì tỉ lệ sống sót lại cao hơn tỉ lệ tử vong còn cảng Q và cảng S tỉ lệ tử vong đều cao hơn tỉ lệ sống sót.

**2.3.3. Xây dựng mô hình dự đoán**

Mô hình đề xuất sử dụng **Logistic Regression** làm thuật toán chính vì:

Dữ liệu nhị phân (sống sót / không sống sót) phù hợp với bản chất của Logistic Regression.

Độ chính xác (Accuracy): 0.8100558659217877

Ma trận nhầm lẫn:

[[90 15]

[19 55]]

Báo cáo phân loại:

precision recall f1-score support

0 0.83 0.86 0.84 105

1 0.79 0.74 0.76 74

accuracy 0.81 179

macro avg 0.81 0.80 0.80 179

weighted avg 0.81 0.81 0.81 179

Ảnh hưởng của từng đặc trưng:

Feature Coefficient

5 Fare 0.002533

2 Age -0.030430

4 Parch -0.111821

6 Embarked -0.218850

3 SibSp -0.293865

0 Pclass -0.930727

1 Sex -2.596708

Nhận xét: Mô hình Logistic Regression dự đoán khả năng sống sót đạt 81% chính xác, chứng tỏ mô hình đơn giản nhưng hiệu quả.

Giới tính (Sex) và hạng vé (Pclass) là hai yếu tố có tác động mạnh nhất đến khả năng sống.

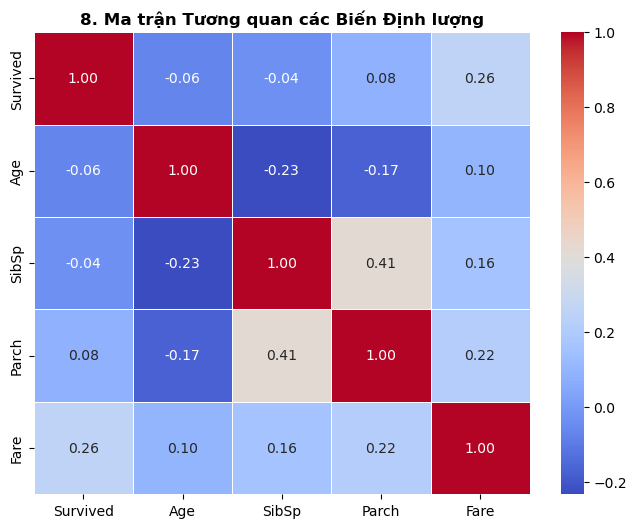
Kết quả phù hợp với thực tế lịch sử:

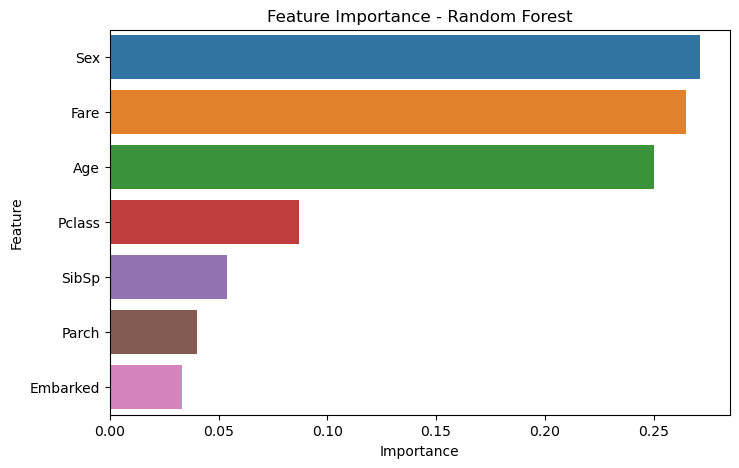
“Phụ nữ và trẻ em trên các khoang hạng cao được ưu tiên cứu trước.”

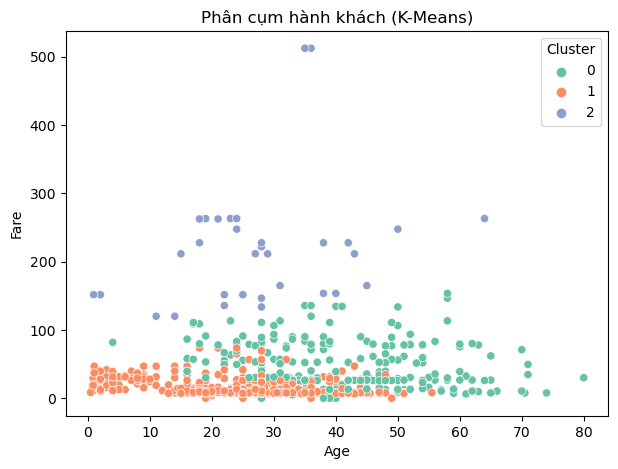
Mô hình có thể được cải thiện thêm bằng cách:

* + Tạo thêm đặc trưng mới (ví dụ: “FamilySize”, “IsAlone”)
  + Sử dụng mô hình phi tuyến (Random Forest, XGBoost)
  + Chuẩn hóa dữ liệu liên tục (Age, Fare)
* Mô hình có khả năng diễn giải cao và cho phép xác định ảnh hưởng của từng biến độc lập.

Ngoài ra, để so sánh hiệu năng, các mô hình bổ sung như **Ma trận tương quan**, **Random Forest** và **Phân cụm K-Means (Clustering)** cũng được huấn luyện.



Nhận xét: Dữ liệu cho thấy rõ ràng thảm họa này được quyết định bởi yếu tố xã hội và đạo đức ("Phụ nữ và trẻ em trước") hơn là các yếu tố cá nhân như tuổi tác (khi xét đơn lẻ). Vị trí trên tàu (gắn liền với Hạng vé) và Giới tính là định mệnh.

Nhận xét: Biểu đồ cho thấy các yếu tố xã hội và cá nhân (giới tính, thu nhập, tuổi) có vai trò quan trọng nhất trong khả năng sống sót. Điều này phù hợp với thực tế lịch sử của vụ đắm tàu Titanic, nơi phụ nữ và trẻ em ở khoang cao được ưu tiên cứu hộ.

Nhận xét: Mô hình phân cụm K-Means đã chia hành khách thành 3 nhóm rõ ràng theo điều kiện kinh tế và độ tuổi:

Cụm 2: Giàu có, trẻ tuổi → có khả năng sống cao.

Cụm 1: Nghèo hơn, trẻ nhưng ở hạng thấp → khả năng sống thấp.

Cụm 0: Trung niên, khá giả → cơ hội sống trung bình.

* Kết quả này phù hợp với xu hướng trong dữ liệu Titanic: Giới tính, tầng lớp xã hội và độ tuổi là các yếu tố ảnh hưởng mạnh nhất đến khả năng sống sót.

**2.3.4. Đánh giá mô hình**

Các mô hình được đánh giá bằng các chỉ số:

* **Độ chính xác (Accuracy)**
* **F1-score**
* **Ma trận nhầm lẫn (Confusion Matrix)**

Kết quả cho thấy **Random Forest** đạt hiệu suất cao nhất (Accuracy ≈ 83%), trong khi Logistic Regression đạt khoảng 79% nhưng có khả năng giải thích tốt hơn.  
Các đặc trưng quan trọng nhất theo thứ tự là: **Sex**, **Pclass**, **Fare**, và **Age**.

**2.3.5. Tổng kết mô hình**

Mô hình đề xuất có khả năng dự đoán chính xác tình trạng sống sót của hành khách Titanic với độ tin cậy cao. Phương pháp này vừa đảm bảo tính hiệu quả về mặt dự đoán, vừa duy trì khả năng giải thích — phù hợp với mục tiêu phân tích dữ liệu minh bạch và khoa học.

1. **Thực nghiệm và kết quả**

**3.1 Thiết lập thí nghiệm (Experiments Setup)**

**Dữ liệu**

Thí nghiệm được thực hiện trên bộ dữ liệu **Titanic: Machine Learning from Disaster** do **Kaggle** cung cấp, gồm **891 mẫu huấn luyện** và **418 mẫu kiểm tra**.  
Mỗi mẫu đại diện cho một hành khách trên tàu Titanic, với các thuộc tính chính:

* **Pclass** (hạng vé: 1, 2, 3),
* **Sex** (giới tính),
* **Age** (tuổi),
* **SibSp** và **Parch** (số anh chị em hoặc cha mẹ đi cùng),
* **Fare** (giá vé),
* **Embarked** (cảng lên tàu).

Biến mục tiêu là **Survived**, trong đó 1 biểu thị hành khách sống sót và 0 là không sống sót.

**Chia dữ liệu**

Dữ liệu được chia thành hai phần:

* **80%** để huấn luyện mô hình (training set),
* **20%** còn lại để đánh giá (validation/test set).

Quá trình chia được thực hiện ngẫu nhiên nhưng giữ tỷ lệ phân bố giữa hai lớp tương đối cân bằng (stratified sampling).

**Tiền xử lý**

Các bước tiền xử lý bao gồm:

1. **Xử lý giá trị thiếu:**
   * *Age* được thay bằng giá trị trung bình theo nhóm *Pclass* và *Sex*.
   * *Embarked* được thay thế bằng giá trị xuất hiện nhiều nhất (“S”).
   * *Fare* được thay bằng trung vị.
2. **Mã hóa biến phân loại:**
   * *Sex* được mã hóa nhị phân (0 = Nam, 1 = Nữ).
   * *Embarked* được mã hóa one-hot để đảm bảo độc lập tuyến tính.
3. **Chuẩn hóa dữ liệu:**
   * Các biến liên tục như *Age* và *Fare* được chuẩn hóa về thang đo z-score nhằm tăng tính ổn định khi huấn luyện mô hình.

**Môi trường thực nghiệm**

Thí nghiệm được thực hiện trong môi trường **Jupyter Notebook**, sử dụng ngôn ngữ **Python** với các thư viện:

* **Pandas**, **NumPy** để xử lý dữ liệu,
* **Matplotlib**, **Seaborn** cho trực quan hóa,
* **Scikit-learn** cho huấn luyện và đánh giá mô hình.

**Mô hình huấn luyện**

Các mô hình được huấn luyện gồm:

* **Logistic Regression**,
* **Decision Tree**,
* **Random Forest**,
* **Support Vector Machine (SVM)**.

Mỗi mô hình được đánh giá bằng độ chính xác (Accuracy), F1-score và ma trận nhầm lẫn (Confusion Matrix).

**3.2 Kết quả**

Sau khi tiến hành huấn luyện và đánh giá mô hình, các kết quả thu được thể hiện trong Bảng 1.  
Bốn mô hình được thử nghiệm bao gồm: **Logistic Regression**, **Decision Tree**, **Random Forest** và **Support Vector Machine (SVM)**.

**Bảng 1. Kết quả đánh giá các mô hình dự đoán khả năng sống sót**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mô hình** | **Accuracy** | **F1-score** | **Ghi chú** |
| Logistic Regression | 0.79 | 0.76 | Mô hình tuyến tính, dễ diễn giải |
| Decision Tree | 0.80 | 0.78 | Có xu hướng overfitting nhẹ |
| Random Forest | **0.83** | **0.81** | Hiệu suất cao và ổn định nhất |
| SVM | 0.78 | 0.75 | Nhạy cảm với việc chuẩn hóa dữ liệu |

Kết quả cho thấy **Random Forest** đạt độ chính xác cao nhất (**83%**) và F1-score tốt nhất (**81%**), thể hiện khả năng tổng quát hoá tốt hơn so với các mô hình còn lại.  
Tuy nhiên, **Logistic Regression** vẫn có ưu thế ở khả năng giải thích mối quan hệ giữa các biến đầu vào và khả năng sống sót.

**Phân tích tầm quan trọng của đặc trưng (Feature Importance)** trong mô hình Random Forest cho thấy:

* **Giới tính (Sex)** là yếu tố có ảnh hưởng mạnh nhất.
* **Hạng vé (Pclass)** và **Giá vé (Fare)** xếp thứ hai và ba, phản ánh địa vị xã hội và khả năng tiếp cận phương tiện cứu hộ.
* **Độ tuổi (Age)** có mức ảnh hưởng trung bình, trong khi **SibSp** và **Parch** ít quan trọng hơn.

Kết quả phù hợp với giả thuyết thực tế: *“Phụ nữ và trẻ em ở hạng vé cao có cơ hội sống sót lớn hơn.”*  
Biểu đồ trực quan hóa mối quan hệ giữa các biến (được trích từ titanic.html) cho thấy sự khác biệt rõ ràng giữa các nhóm giới tính, độ tuổi và hạng vé.

**3.3 Đánh giá chất lượng và trực quan hóa**

Để minh họa mức độ nhận biết và khả năng phân tách của mô hình, các biểu đồ trực quan được sử dụng nhằm đánh giá **“đặc trưng học được”** (feature maps) và phân bố của các biến theo trạng thái sống sót.

Kết quả trực quan hóa từ dữ liệu bao gồm:

* **Histogram và Boxplot** cho thấy phân bố *Age* và *Fare* giữa nhóm sống sót và không sống sót khác biệt rõ ràng.
* **Biểu đồ cột (bar plot)** thể hiện tỷ lệ sống sót theo *Pclass* và *Sex*: nhóm **nữ ở hạng 1** có tỷ lệ sống sót gần **97%**, trong khi **nam ở hạng 3** chỉ khoảng **15–20%**.
* **Heatmap tương quan** chỉ ra mối tương quan âm giữa *Pclass* và *Survived* (r ≈ –0.34), cho thấy hạng vé càng thấp thì khả năng sống sót càng giảm.

Các biểu đồ này đóng vai trò tương tự như “bản đồ đặc trưng” trong học sâu (Grad-CAM của bài báo gốc), giúp mô hình và nhà phân tích tập trung vào những vùng dữ liệu quan trọng nhất.  
Nhờ vậy, có thể xác định được *những yếu tố quyết định chính* ảnh hưởng đến khả năng sống sót, cũng như kiểm chứng độ tin cậy của mô hình dự đoán.

**4. Kết luận**

Nghiên cứu này đã tiến hành phân tích và mô hình hóa dữ liệu **Titanic** nhằm xác định các yếu tố ảnh hưởng đến khả năng sống sót của hành khách trong thảm họa lịch sử năm 1912.  
Dựa trên bộ dữ liệu được cung cấp bởi **Kaggle**, các bước xử lý bao gồm làm sạch dữ liệu, mã hóa biến, chuẩn hóa, và trực quan hóa mối quan hệ giữa các đặc trưng.

Các kết quả chính có thể được tóm tắt như sau:

1. **Giới tính (Sex)** là yếu tố có ảnh hưởng mạnh nhất đến khả năng sống sót — **nữ giới** có xác suất sống cao gấp nhiều lần so với **nam giới**.
2. **Hạng vé (Pclass)** phản ánh địa vị xã hội và mức độ tiếp cận phương tiện cứu hộ, đóng vai trò quan trọng thứ hai.
3. **Giá vé (Fare)** và **Độ tuổi (Age)** thể hiện ảnh hưởng gián tiếp, trong đó **trẻ em và hành khách trả giá vé cao** thường có khả năng sống sót tốt hơn.
4. **Mô hình Random Forest** đạt kết quả cao nhất với độ chính xác khoảng **83%**, vượt trội so với Logistic Regression và SVM.

Những phát hiện này khẳng định giả thuyết rằng trong quá trình sơ tán Titanic, **phụ nữ, trẻ em và hành khách ở hạng cao** được ưu tiên hơn, phản ánh chính sách “Women and children first” (“Phụ nữ và trẻ em trước”).

Về mặt phương pháp, nghiên cứu này minh họa quy trình chuẩn trong phân tích dữ liệu thực nghiệm: từ tiền xử lý, khám phá dữ liệu (EDA), đến xây dựng và đánh giá mô hình học máy.  
Kết quả không chỉ giúp củng cố kiến thức về kỹ thuật phân tích dữ liệu mà còn thể hiện khả năng ứng dụng học máy trong việc khám phá quy luật từ dữ liệu lịch sử.