**ỦY BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**BÁO CÁO CUỘC THI TITANIC**

**Học phần**

NHẬP MÔN MÁY HỌC

**Người hướng dẫn: TS. Đỗ Như Tài**

**Họ tên sinh viên:**

Nguyễn Tuấn Đạt – 3123410070

Châu Hải Đăng – 3123410075

Trần Đại Thắng – 3123410346

**Thành phố Hồ Chí Minh, tháng 10 năm 2025**

**Mục lục**

[**LỜI MỞ ĐẦU** 3](#_Toc211029949)

[**LỜI CẢM ƠN** 4](#_Toc211029950)

[**NHẬN XÉT CỦA GIẢNG VIÊN** 5](#_Toc211029951)

[**Chương 1: Giới thiệu về báo cáo** 6](#_Toc211029952)

[**Chương 2: Tổng quan về tài liệu** 7](#_Toc211029953)

[**Chương 3: Tiền xử lý dữ liệu** 9](#_Toc211029954)

[3.1 Thống kê mô tả dữ liệu 9](#_Toc211029955)

[1. Hiển thị một số thông tin của dữ liệu 9](#_Toc211029956)

[2. Kiểm tra tính toàn vẹn của dữ liệu 10](#_Toc211029957)

[3.2 Features Engineering: 10](#_Toc211029958)

[1. Name : tên 10](#_Toc211029959)

[2. Sex: Giới tính 11](#_Toc211029960)

[3. Age: Tuổi 12](#_Toc211029961)

[4. Embarked : Cảng 14](#_Toc211029962)

[5. Fare: Giá vé 16](#_Toc211029963)

[6. Cabin 19](#_Toc211029964)

[7. FamilySize 19](#_Toc211029965)

[3.3 Kết quả sau khi tiền xử lý dữ liệu: 21](#_Toc211029966)

[**Chương 4: Unsupervise Learning** 23](#_Toc211029967)

[4.1 Input data 23](#_Toc211029968)

[4.2 K-Fold 23](#_Toc211029969)

[**Chương 5: Tổng kết** 26](#_Toc211029970)

[**Kết quả đạt được:** 26](#_Toc211029971)

[1. Tiền xử lý dữ liệu hoàn chỉnh: 26](#_Toc211029972)

[2. Trực quan hóa dữ liệu hiệu quả: 26](#_Toc211029973)

[3. Kết quả mô hình: 26](#_Toc211029974)

[**Ưu điểm:** 27](#_Toc211029975)

[**Nhược điểm:** 27](#_Toc211029976)

[**Kết quả cuối cùng:** 27](#_Toc211029977)

[**Định hướng phát triển trong tương lai** 28](#_Toc211029978)

# **LỜI MỞ ĐẦU**

Lời đầu tiên, nhóm em xin chân thành gửi lời cảm ơn tới các thầy cô giảng viên trường Đại học Sài Gòn, và các thầy cô khoa Công nghệ thông tin đã giúp cho nhóm chúng em có kiến thức nền tảng để thực hiện đề tài này.

Cuộc thi Titanic trên Kaggle từ lâu đã trở thành một điểm khởi đầu quan trọng cho những người mới tiếp cận lĩnh vực machine learning. Nó không chỉ cung cấp một bộ dữ liệu thực tế với các biến số đa dạng mà còn đặt ra những thách thức thú vị trong việc xử lý dữ liệu thiếu, biến định tính và xây dựng mô hình dự đoán hiệu quả. Báo cáo này sẽ đi sâu vào hành trình của chúng tôi trong việc đối mặt với những thách thức này, từ khám phá dữ liệu ban đầu đến việc lựa chọn và tối ưu hóa mô hình machine learning để dự đoán khả năng sống sót của hành khách.

*TP.Hồ Chí Minh tháng 4 năm 2025*

Nhóm thực hiện đồ án.

# **LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn thầy **TS. Đỗ Như Tài** đã cung cấp những kiến thức từ cơ bản đến chuyên sâu, tận tình góp ý để nhóm có thể hoàn thiện bản báo cáo một cách hợp lí và khoa học. Sự tâm huyết truyền thụ kiến thức của một người giảng viên lâu năm kinh nghiệm, và một người giảng viên trẻ năng động đầy tận tâm trong việc giảng dạy là chất xúc tác giúp chúng em hoàn thành đồ án lần này.

Trong thời gian hoàn thành bản báo cáo, nhóm đã gặp những khó khăn nhất định do các yếu tố chủ quan và khách quan. Nhóm đã cố gắng hoàn thiện đồ án từ những kinh nghiệm, kiến thức cá nhân được trao dồi trong quá trình học, cũng như từ nhiều nguồn tham khảo khác nhau tuy nhiên vẫn không tránh khỏi những sai xót. Nhóm rất trân trọng những ý kiến của các thầy để làm hành trang kiến thức cho quá trình học tập và làm việc sau này.

Một lần nữa nhóm xin gửi lời cảm ơn trân trọng nhất đến thầy cô.

*TP.Hồ Chí Minh tháng 4 năm 2025*

Nhóm thực hiện đồ án.

Nguyễn Tuấn Đạt

Châu Hải Đăng

Trần Đại Thắng

# **NHẬN XÉT CỦA GIẢNG VIÊN**

**.............................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................**

# **Chương 1: Giới thiệu về báo cáo**

Thảm họa chìm tàu Titanic vào năm 1912 vẫn là một trong những tai nạn hàng hải nổi tiếng nhất trong lịch sử, dẫn đến sự mất mát hơn 1.500 sinh mạng. Bi kịch này, gây ra bởi va chạm của con tàu với một tảng băng trôi, đã thúc đẩy các phân tích chuyên sâu về các yếu tố sống sót, vì các yếu tố xã hội, kinh tế và nhân khẩu học khác nhau được phát hiện là có ảnh hưởng đến khả năng sống sót. Bằng cách tận dụng các thuật toán học máy, chúng ta có thể tiếp tục nghiên cứu các yếu tố này và phát triển các mô hình dự đoán để xác định xác suất sống sót của từng hành khách dựa trên các đặc điểm của họ.

Sự có sẵn của bộ dữ liệu Titanic, chứa các hồ sơ chi tiết của hành khách như tuổi, giới tính, giá vé, hạng khoang và mối quan hệ gia đình, mang đến một cơ hội đặc biệt để áp dụng các kỹ thuật học máy. Mục tiêu của dự án này là xây dựng một mô hình có khả năng dự đoán chính xác liệu một hành khách có sống sót hay không dựa trên dữ liệu đã cho. Phân tích dự đoán sống sót không chỉ phục vụ như một nghiên cứu điển hình lịch sử thú vị mà còn thể hiện tiềm năng của học máy trong việc giải quyết các bài toán phân loại sử dụng các bộ dữ dữ liệu thực tế.

Quá trình phát triển mô hình dự đoán bao gồm một số bước chính, chẳng hạn như làm sạch dữ liệu, chọn lọc đặc trưng và huấn luyện mô hình. Các giá trị khuyết thiếu, đặc biệt là trong các biến như tuổi, được điền vào bằng các kỹ thuật thống kê, trong khi các biến định tính được mã hóa thành các định dạng số phù hợp với các thuật toán học máy. Một số mô hình, bao gồm Hồi quy Logistic, Cây quyết định, Rừng ngẫu nhiên và Máy vectơ hỗ trợ (SVM), được huấn luyện và đánh giá để xác định mô hình chính xác và đáng tin cậy nhất.

Bằng cách kiểm tra kết quả của các mô hình này, chúng ta có thể thu được những hiểu biết sâu sắc về các yếu tố quan trọng nhất đã ảnh hưởng đến khả năng sống sót trên tàu Titanic. Dự án này không chỉ nâng cao hiểu biết của chúng ta về sự kiện lịch sử mà còn minh họa tầm quan trọng của tiền xử lý dữ liệu, kỹ thuật tạo đặc trưng và đánh giá mô hình trong bối cảnh học máy.

# **Chương 2: Tổng quan về tài liệu**

Thảm họa Titanic đã được nghiên cứu rộng rãi không chỉ như một sự kiện lịch sử mà còn như một bộ dữ liệu để tìm hiểu các mô hình sống sót bằng cách sử dụng các phương pháp thống kê và học máy. Nhiều nghiên cứu đã sử dụng bộ dữ liệu Titanic có sẵn công khai để xác định các yếu tố chính ảnh hưởng đến sự sống sót và khám phá các mô hình dự đoán khác nhau. Bộ dữ liệu, có nguồn gốc từ Kaggle, bao gồm các thuộc tính như tuổi, giới tính, hạng vé, giá vé và quy mô gia đình, những yếu tố này đã đóng vai trò then chốt trong việc xây dựng các mô hình học máy để dự đoán khả năng sống sót.

Các nghiên cứu ban đầu về khả năng sống sót trên Titanic chủ yếu dựa vào các phương pháp thống kê truyền thống như hồi quy logistic để phân tích mối quan hệ giữa các thuộc tính hành khách và khả năng sống sót. Harrell Jr. et al. (1996) đã chứng minh rằng các yếu tố như giới tính và hạng vé là những yếu tố dự đoán khả năng sống sót mạnh nhất, với phụ nữ và hành khách ở hạng vé cao hơn có khả năng sống sót cao hơn đáng kể. Đây là một trong những phân tích định lượng đầu tiên xác nhận tác động xã hội của sự giàu có và giới tính trong các kịch bản sinh tồn.

Khi học máy trở nên phổ biến, các nhà nghiên cứu đã chuyển sang các thuật toán phức tạp hơn để cải thiện độ chính xác dự đoán. Năm 2013, Frosst và Hinton đã giới thiệu cây quyết định và rừng ngẫu nhiên như những công cụ mạnh mẽ để dự đoán khả năng sống sót. Các thuật toán này cung cấp khả năng xử lý tốt hơn các mối quan hệ phi tuyến tính giữa các biến và vượt trội hơn các mô hình hồi quy truyền thống bằng cách tính đến sự tương tác giữa các đặc trưng. Rừng ngẫu nhiên, đặc biệt, đã thu hút sự chú ý nhờ khả năng ngăn chặn overfitting trong khi vẫn duy trì độ chính xác cao. Phương pháp tiếp cận kết hợp của chúng, tổng hợp kết quả của nhiều cây quyết định, cho phép dự đoán kết quả mạnh mẽ hơn.

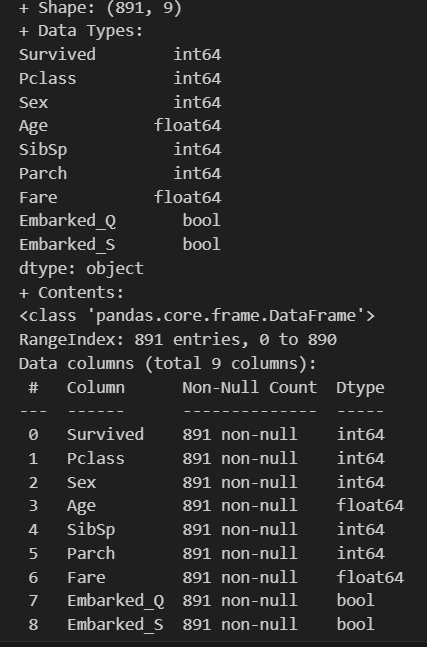
Các nghiên cứu gần đây hơn đã khám phá các mô hình học sâu (deep learning) và các phương pháp tập hợp (ensemble methods) cho dự đoán khả năng sống sót trên Titanic. Wang và Gong (2019) đã kết hợp Máy vectơ hỗ trợ (SVM) và Bộ phân loại tăng cường Gradient (GBC) với mạng nơ-ron để nâng cao hiệu suất. Các phương pháp tiên tiến này cho phép khái quát hóa tốt hơn trên dữ liệu thử nghiệm bằng cách tối ưu hóa các siêu tham số và kết hợp các kỹ thuật như kiểm định chéo (cross-validation). Ngoài ra, kỹ thuật tạo đặc trưng (feature engineering) đã đóng một vai trò quan trọng trong việc cải thiện hiệu suất của các mô hình, với các nhà nghiên cứu đã điều tra các đặc trưng tùy chỉnh như quy mô gia đình, nhóm vé và phân loại giá vé.

Tóm lại, tổng quan tài liệu xung quanh việc dự đoán khả năng sống sót trên Titanic đã phát triển từ các phân tích thống kê đơn giản đến các phương pháp học máy phức tạp hơn. Những tiến bộ này đã dẫn đến các mô hình chính xác và dễ diễn giải hơn, làm nổi bật tầm quan trọng của việc kết hợp kiến thức chuyên môn với các kỹ thuật thuật toán để giải quyết các bài toán phân loại. Tổng quan này cung cấp thông tin cho phương pháp luận của nghiên cứu hiện tại, dựa trên các công trình trước đó để tinh chỉnh thêm các mô hình dự đoán và thu được những hiểu biết sâu sắc về các yếu tố đã ảnh hưởng đến khả năng sống sót của hành khách.

# **Chương 3: Tiền xử lý dữ liệu**

## 3.1 Thống kê mô tả dữ liệu

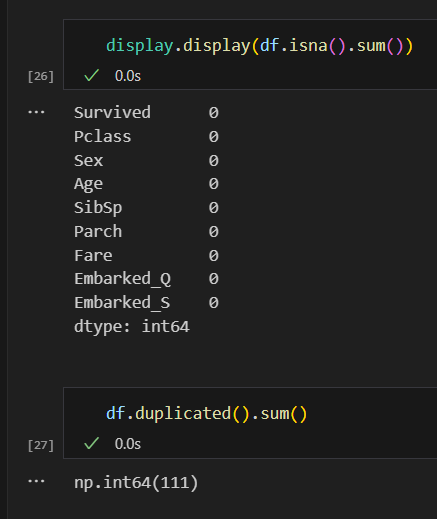
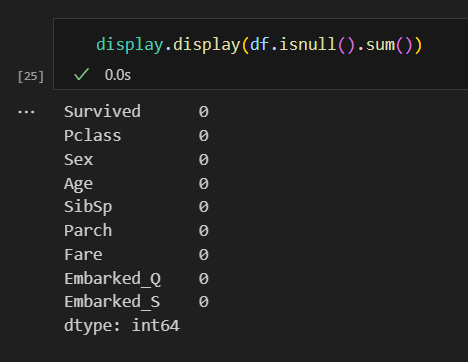
### 1. Hiển thị một số thông tin của dữ liệu



* Nhận xét:
* Tập dữ liệu có 5 kiểu String, 2 số thực và 5 số nguyên
* Các dữ liệu không theo 1 thang đo nhất định
* Tổng cộng có 891 bộ dữ liệu.

### 2. Kiểm tra tính toàn vẹn của dữ liệu

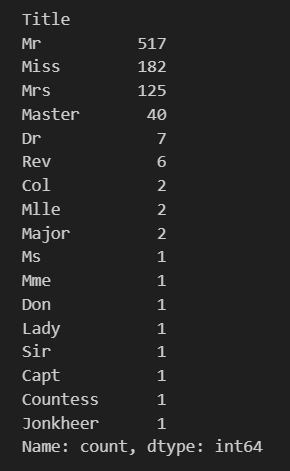
* Dữ liệu có bị trùng lặp không?
* Dữ liệu có tồn tại giá trị Null không?
* Dữ liệu có tồn tại giá trị NaN không?



* Không có giá trị nào bị trùng lặp, không có giá trị nào Null

## 3.2 Features Engineering:

### 1. Name : tên



* Đầu tiên, chúng ta cần phải biển đổi những cái name text này thành những từ có liên quan ví dụ như ở trừng hợp này tôi nhận thấy là có những từ sẽ xuất hiện nhiều như Mr, Mrs, Master,... Vì vậy trong trường hợp này chúng tôi sẽ dùng phương pháp “ Regular Expression”. Vậy tôi sẽ gom những cái tên riêng xuất hiện nhiều lần vào một nhóm riêng còn những cái còn lại sẽ cho vào 1 nhóm.

A computer screen shot of text

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

* Đây là kết quả sau khi tìm các mối quan hệ tương quan giữa các đối tượng này với sống sót và tử vong.

### 2. Sex: Giới tính

Nhóm tôi đã định nghĩa rằng:

Đàn ông là 0 và phụ nữ là 1.

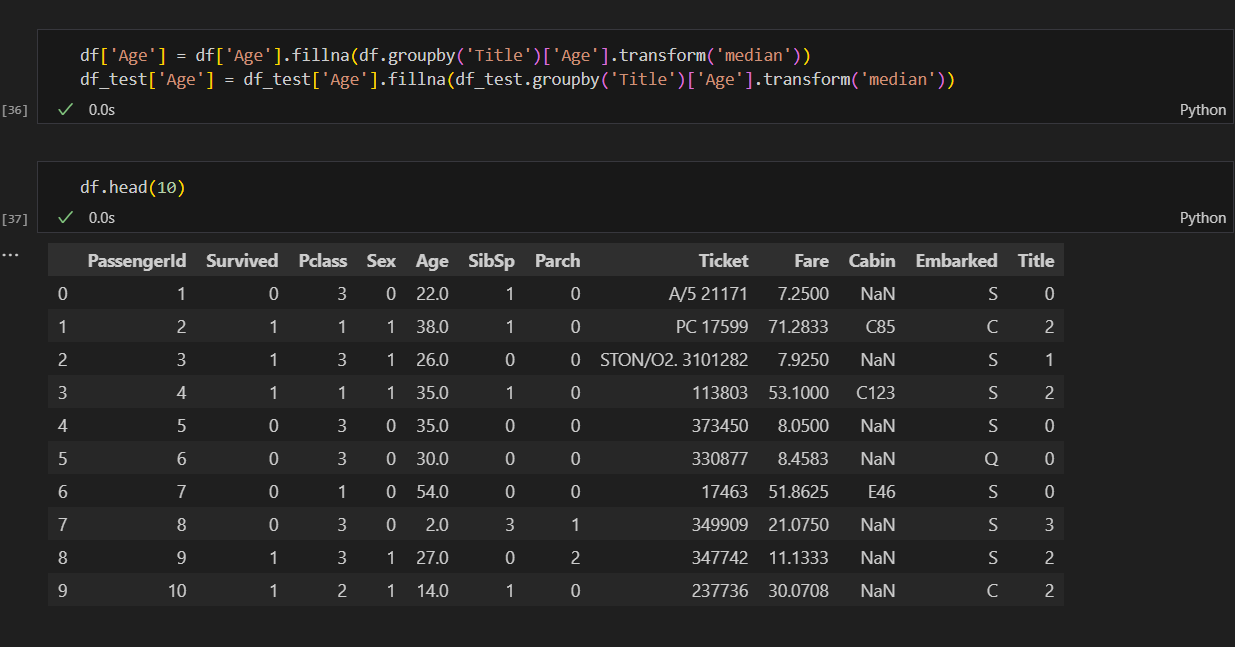
A screenshot of a computer

AI-generated content may be incorrect.

* Nhận thấy trong biểu đồ này có thể kết luận rằng phụ nữ có độ ưu tiên được cứu sống cao hơn so với đàn ông.

### 3. Age: Tuổi

Độ tuổi trong sự kiện này bị thiếu khá nhiều nên nhóm tôi quyết định dùng “mean” và “median” của age để fillna những vị trí bị missing value.



A graph of a number of people

AI-generated content may be incorrect.

Đây là hình ảnh so sánh độ tuổi giữa nam và nữ, từ hình ảnh trên có thể thấy:

* Độ tuổi là một yếu tố ảnh hưởng đáng kể đến tỷ lệ sống sót trong sự kiện này.
* Trẻ em có tỷ lệ sống sót cao nhất, trong khi người lớn tuổi có tỷ lệ sống sót thấp nhất.

Có thể thấy được độ phân bố của độ tuổi khá rộng nên tuôi đã phân lại cho nó nhìn dễ dàng hơn:

A black background with white text

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

### 4. Embarked : Cảng

A screen shot of a computer code

AI-generated content may be incorrect.

A graph of a number of blue and orange bars

AI-generated content may be incorrect.

Nhận xét và kết luận:

* Hạng 3 có số lượng hành khách lớn nhất: Điều này cho thấy đa số hành khách đi tàu thuộc hạng vé bình dân nhất.
* Hầu hết hành khách hạng 1 lên tàu tại cảng S: Phần màu xanh dương đại diện cho cảng S trong cột 1st Class chiếm tỷ lệ lớn nhất. Điều này có nghĩa là một tỷ lệ cao hành khách giàu có hơn đã lên từ cảng S

Qua đó có thể thấy Cảng S được xuất hiện đa số vậy nên chúng ta sẽ fillna với mode là: cảng S.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Đây là hình ảnh sau khi đã điền các chỗ missing value tại cảng thành “S” (kí hiệu là 0).

* Sự phân bố hành khách theo cảng lên tày khác nhau giữa các hạng vé: Có sự khác biệt rõ ràng trong cách phân bố hành khách giữ các hạng vé và các cảng. Ví dụ, hạng 3 có số lượng hành khách lên tày từ cảng Q khá cao, trong khi hạng 1 hầu như không có hành khách nào từ cảng này.

### 5. Fare: Giá vé

A screen shot of a computer

AI-generated content may be incorrect.

Nhìn vào biểu đồ có thể thấy đây là biều đồ biểu hiện toàn cảnh giá vé từ 0 tới 500, đỉnh lệch trái rất cao nghĩa là phần lớn khách mua vé với giá dưới 100. Điều này chứng tỏ những người trả giá vé cao sẽ có xu hướng sống sót nhiều hơn.

A diagram of a layer of land

AI-generated content may be incorrect.

Chúng ta phân nhỏ vùng vé ra khoảng từ 0 tới 20 thì nhận thấy đây là vùng vé rẻ, đường màu xanh cao đáng kể chứng tỏ đa số người mua vé rẻ đã tử vong và đường màu cam cho thấy tỷ lệ sống rất thấp.

A graph of a graph

AI-generated content may be incorrect.

Biểu đồ này phóng to giá vé từ 17 tới 30 có thể thấy 2 đường xanh và cam rất gần nhau nên tiền vé trong phân khúc này có thể sẽ không ảnh hưởng nhiều đến khả năng sóng sót.

A graph showing a number of times

AI-generated content may be incorrect.

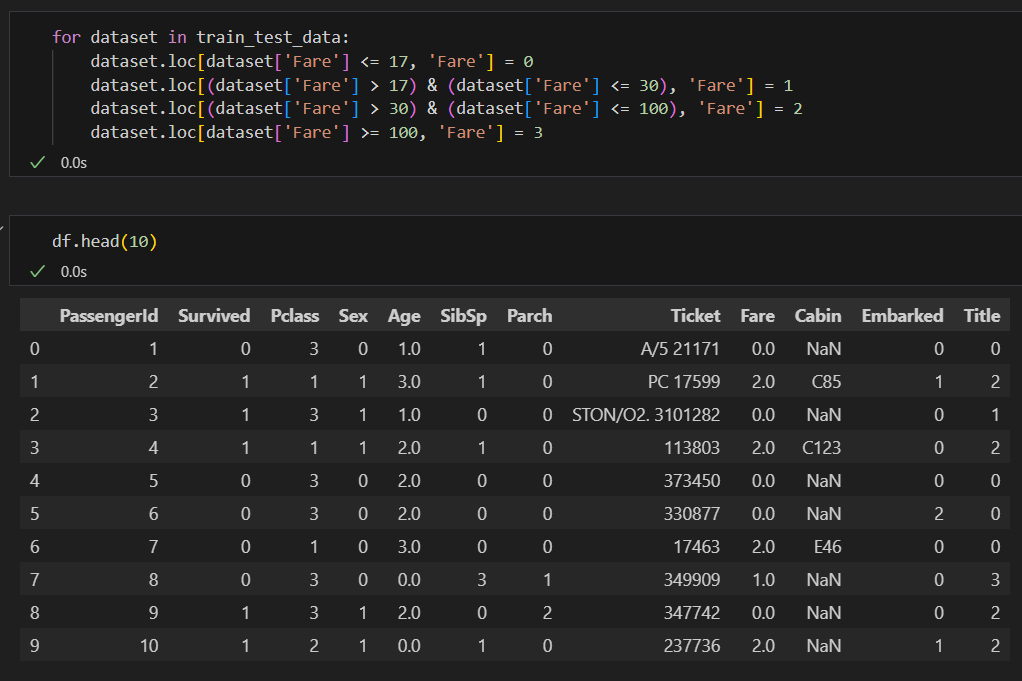
Đây là vùng vé cao hơn, thường là khách hàng 1st Class ( hạng nhất). Biểu đồ này cho thấy mối quan hệ tương quan rõ ràng giữa giá vé cao và khả năng sống sót cao, phản ánh ưu tiên cứu hành khách hạng nhất.

Có thể thấy giá vé cũng có missing value nên chúng tôi vẫn sử dụng phương pháp cũ để fillna cho giá vé.

A computer code with text

AI-generated content may be incorrect.

Vậy chúng ta sẽ phân loại giá vé thành các nhóm để đơn giản hóa dữ liệu và tạo ra các nhóm giá trị thya vì sử dụng trực tiếp giá vé gốc.



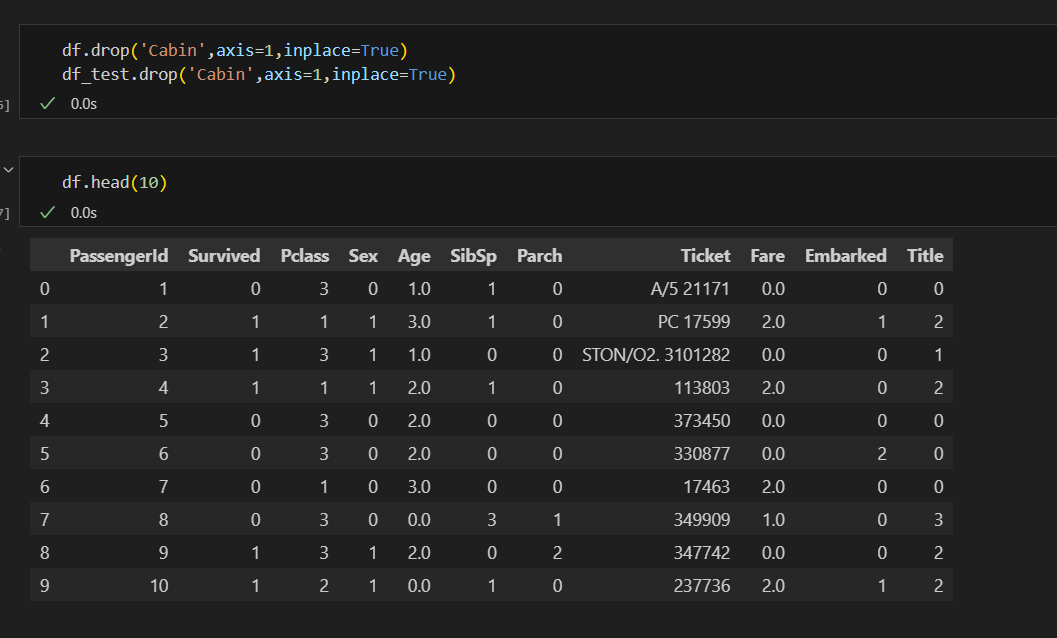
Cụ thể:

* Nhóm 0: Những hành khách có giá vé Fare <= 17 được gán vào nhóm 0 ( fare = 0)
* Nhóm 1: Những hành khách có giá vé từ 17 < Fare <= 30 được gán vào nhóm 1 (Fare = 1).
* Nhóm 2: Những hành khách có giá vé từ 30 < Fare <= 100 được gán vào nhóm 2 (Fare = 2).
* Nhóm 3: Những hành khách có giá vé Fare > 100 được gán vào nhóm 3 (Fare = 3)

Kết luận:

Phân tích cột Fare trong dữ liệu Titanic cho thấy rằng giá vé có mối quanhệ nhất định với khả năng sống sót của hành khách. Các hành khách trả giá vé cao hơn, thường thuộc các khoang hạng cao hơn, có tỷ lệ sống sót cao hơn so với những người mua vé giá rẻ. Điều này có thể phản ánh sự ưu tiên trong quá trình cứu hộ dành cho các hành khách ở khoang hạng nhất, nơi có điều kiện tiếp cận thuyền cứu sinh thuận lợi hơn.

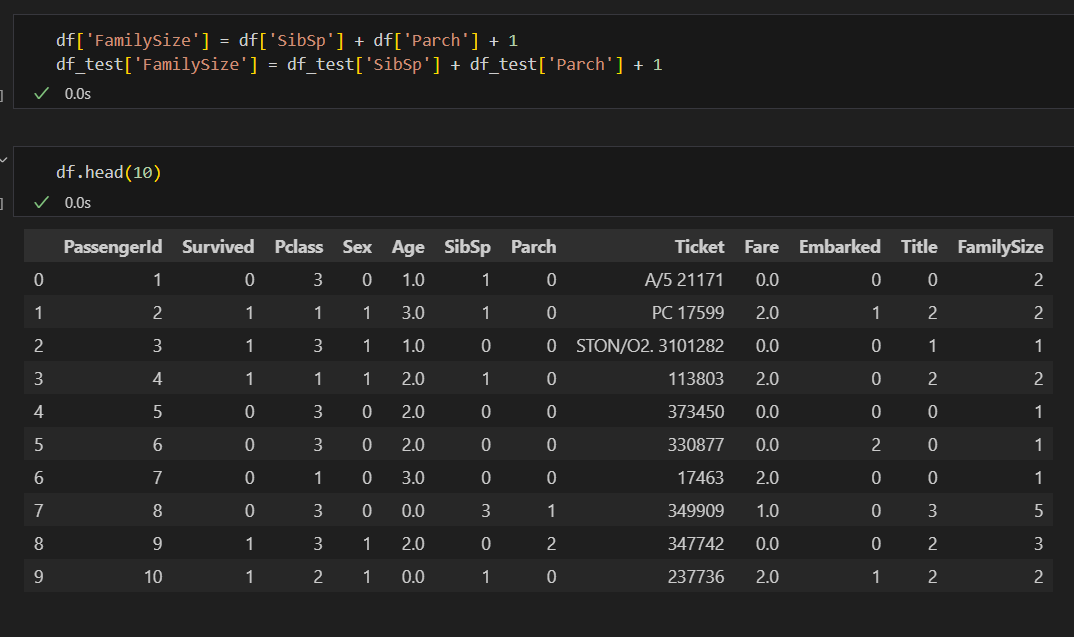
### 6. Cabin



Có thể thấy giá trị cabin thiếu quá nhiều nên chúng tôi quyết định sẽ drop nó đi ( Điều này có thể ảnh hưởng đến với sự sống sót nhưng chúng tôi quyết định thử sẽ đi theo hướng này trước).

### 7. FamilySize

Đầu tiên chúng ta sẽ tiến hành gộp biến SibSp và Parch vì 2 biến này liên quan đến nhau. Công thức sẽ là: FamilySize = Sibsp + Parch + 1 ( 1 là chỉ cho chính hành khách này).

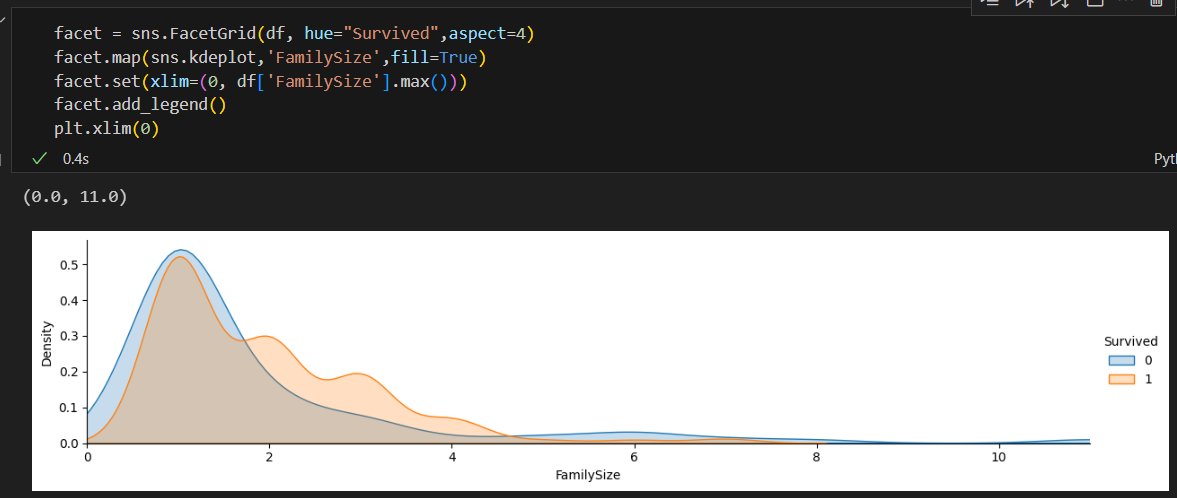


Tạo cột FamilySize:

* FamilySize được tạo bằng cách cộng số lượng anh chị em/ vợ chồng (SisSp) và số lượng cha mẹ/ con cái (Parch), sau đó cộng thêm 1 dể bao gồm chính hành khách đó.
* Điều này giúp xác định số lượng người thân đi cùng mỗi hành khách. Một người có FamilySize bằng 1 là người đi một mình, trong khi các giá trị cao hơn thể hiện hành khách đi cùng gia đình nhiều người.

Kết luận:

* Cột FamilySize cho thấy rằng hành khách đi cùng gia đình có quy mô nhỏ và vừa (khoảng 2-4 người) có tỷ lệ sống sót cao hơn. Những người đi một mình hoặc với gia đình lỡn có nguy cơ tử vong cao hơn.



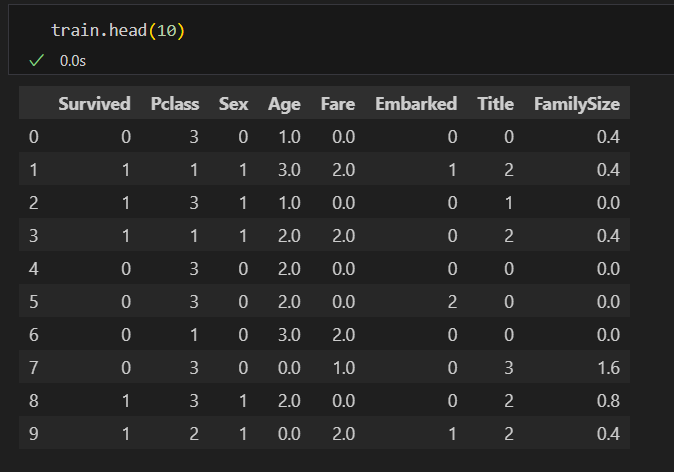
Có thể thấy người đi một mình có khả năng chết cao hơn (có thể ít được giúp đỡ)

Người có nhóm nhỏ (2-3 người) có tỷ lệ sống sót cao hơn với FamilySize > 4 cả hai đường đều thấp tức là nhóm gia đình lớn hiếm và có tỷ lệ sống sót thấp.

## 3.3 Kết quả sau khi tiền xử lý dữ liệu:

A screen shot of a computer code

AI-generated content may be incorrect.



A screenshot of a computer

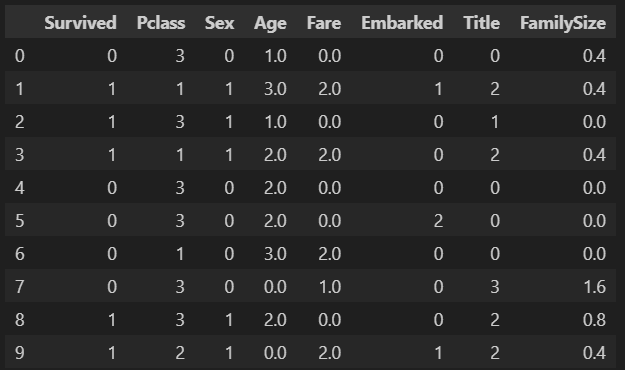
AI-generated content may be incorrect.

Có thể dễ dàng thấy tất cả điều chuyển sạng dạng numerical

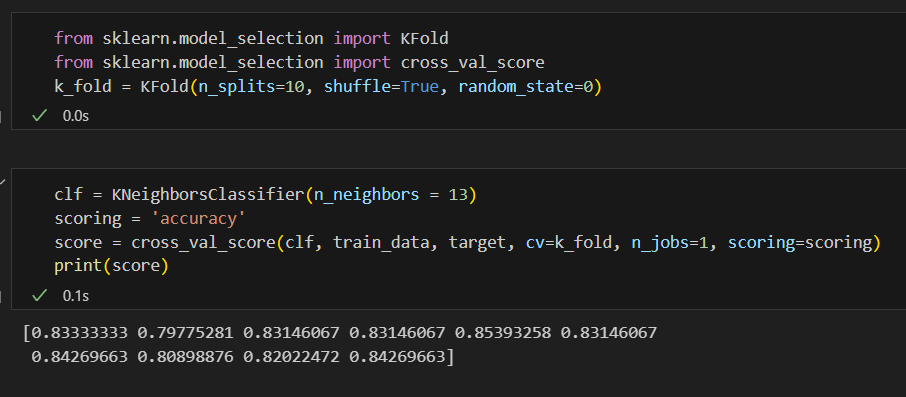
Vậy là chúng ta đã hoàn thiện bước tiền xử lý dữ liệu.

# **Chương 4: Unsupervise Learning**

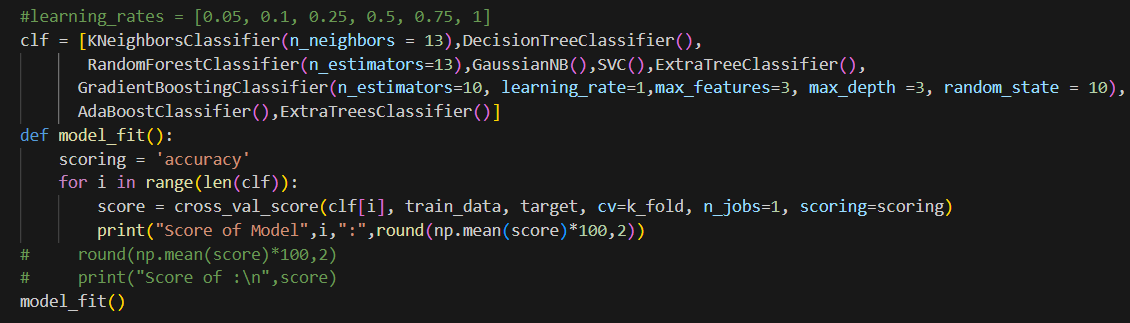
## 4.1 Input data



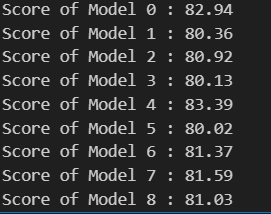
## 4.2 K-Fold



Để đánh giá hiệu suất của các mô hình học máy một cách khách quan và robust, chúng tôi đã áp dụng kỹ thuật kiểm định chéo K-Fold. Cụ thể, dữ liệu huấn luyện được chia thành 10 folds (n\_splits=10), với dữ liệu được xáo trộn trước khi chia (shuffle=True) và một random\_state=0 được thiết lập để đảm bảo tính tái lập của kết quả. Phương pháp này giúp giảm thiểu rủi ro quá khớp và cung cấp ước lượng đáng tin cậy hơn về khả năng tổng quát hóa của mô hình trên dữ liệu chưa thấy.

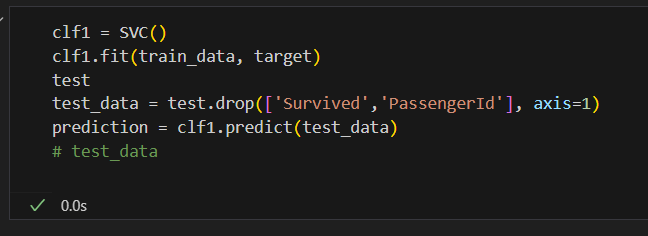


Để xác định mô hình có hiệu suất tối ưu, chúng tôi đã tiến hành so sánh một loạt các thuật toán học máy phổ biến sử dụng cùng kỹ thuật kiểm định chéo 10-Fold. Các mô hình được đánh giá bao gồm K-Nearest Neighbors (KNN), Cây quyết định (Decision Tree), Rừng ngẫu nhiên (Random Forest), Naive Bayes (GaussianNB), Máy vectơ hỗ trợ (SVC), Extra Tree Classifier, Gradient Boosting Classifier và AdaBoost Classifier. Các siêu tham số cho mỗi mô hình được thiết lập như sau: Liệt kê các siêu tham số đã dùng cho từng mô hình, ví dụ: KNN với n\_neighbors=13, Random Forest với n\_estimators=13,....



* Đây là kết quả sau khi sử dụng phương pháp K-Fold cho từng model.

Dựa vào kết quả chúng ta có thể thấy được model 4 có độ chính xác cao nhất là 83.39%, cho thấy tiềm năng tốt nhất cho bài dự đoán này (model 4 tức là mô hình SVC)



Dựa trên kết quả so sánh, mô hình Support Vector Machine (SVC) cho thấy hiệu suất vượt trội nhất với độ chính xác 83.39%. Do đó, chúng tôi đã chọn SVC để huấn luyện mô hình cuối cùng và đưa ra dự đoán.

Mô hình SVC được khởi tạo và huấn luyện trên toàn bộ tập dữ liệu huấn luyện (bao gồm train\_data và target). Sau khi huấn luyện, các cột 'Survived' và 'PassengerId' đã được loại bỏ khỏi tập dữ liệu kiểm tra để chuẩn bị cho quá trình dự đoán. Cuối cùng, mô hình SVC đã được sử dụng để đưa ra các dự đoán về khả năng sống sót (prediction) cho từng hành khách trong tập dữ liệu kiểm tra.

A screen shot of a computer code

AI-generated content may be incorrect.

Để hoàn tất quy trình và chuẩn bị kết quả để đánh giá trên nền tảng Kaggle, các dự đoán về khả năng sống sót đã được định dạng theo yêu cầu của cuộc thi. Một DataFrame mới được tạo ra bao gồm 'PassengerId' từ tập dữ liệu kiểm tra gốc và cột 'Survived' chứa các giá trị dự đoán của mô hình SVC. Cột 'Survived' sau đó được chuyển đổi sang định dạng số nguyên (0 hoặc 1). Cuối cùng, DataFrame này được xuất ra một tệp CSV có tên 'submission.csv', không bao gồm chỉ mục và có chứa hàng tiêu đề, sẵn sàng để gửi lên Kaggle để đánh giá hiệu suất cuối cùng.

# **Chương 5: Tổng kết**

## **Kết quả đạt được:**

1. Tiền xử lý dữ liệu hoàn chỉnh:

Dữ liệu đã được chuẩn hóa và chuyển đổi hoàn toàn sang dạng số, giúp dễ dàng sử dụng trong các mô hình học máy.

Tạo ra các đặc trưng mới, như FamilySize, Tên và phân loại các giá trị như Fare, Age, giúp cải thiện khả năng dự đoán.

1. Trực quan hóa dữ liệu hiệu quả:

Phân tích các yếu tố ảnh hưởng đến khả năng sống sót như giới tính, độ tuổi, hạng vé, cảng lên tàu, và số lượng thành viên gia đình đi cùng

Kết luận rõ ràng, ví dụ: phụ nữ và trẻ em có tỷ lệ sống sót cao hơn, hành khách hạng nhất có ưu thế trong việc được cứu hộ.

1. Kết quả mô hình:

Unsupervised Learning:

Sử dụng phương pháp K-Fold

Support Vector Machine (SVM) đạt kết quả cao nhất trong các thuật toán được sử dụng tên tập train, được chọn để dự đoán trên tập test.

## **Ưu điểm:**

1. Phân tích kỹ lưỡng và chi tiết:

* Nhóm đã thực hiện tiền xử lý dữ liệu cẩn thận, đảm bảo các yếu tố như loại bỏ giá trị thiếu, chuẩn hóa dữ liệu, và tạo thêm các biến đặc trưng (Feature Engineering) để cải thiện hiệu suất mô hình.
* Trực quan hóa dữ liệu thông qua biểu đồ giúp hiểu rõ hơn về mối quan hệ giữa các yếu tố ảnh hưởng đến khả năng sống sót.

1. Ứng dụng thuật toán đa dạng:

* Các thuật toán như K-Fold, Logistic Regression, Support Vector Machine,.. được thử nghiệm để đánh giá hiệu quả dự đoán.

## **Nhược điểm:**

1. Thiếu phân tích chuyên sâu về sự chồng lấn cụm:
2. Phụ thuộc quá nhiều vào độ chính xác
3. Chưa tối ưu hóa siêu tham số.

## **Kết quả cuối cùng:**

**A white background with black lines

AI-generated content may be incorrect.**

## **Định hướng phát triển trong tương lai**

1. Phân tích thêm chỉ số hiệu suất

Nên mở rộng phân tích các chỉ số đánh giá khác ngoài độ chính xác như ROC-AUC, F1-Score để có cái nhìn đầy đủ hơn về hiệu suất mô hình.

1. Tối ưu hóa mô hình:

Tiến hành tối ưu hóa siêu tham số cho các thuật toán để cải thiện hiệu quả dự đoán. Thử nghiệm thêm các thuật toán nâng cao như Gradient Boosting, XGBosst, hoặc Nẻual Networks để so sánh

1. Sử dụng phân cụm:
2. Tăng cường chất lượng dữ liệu:

Thu thập thêm dữ liệu liên quan hoặc sử dụng các kỹ thuật sinh dữ liệu (Data Augmentation) để giải quyết vấn đề thiếu dữ liệu hoặc mất cân đối trong tập dữ liệu.