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

A blue button with white text

AI-generated content may be incorrect.**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**PAPER 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**

**Lời mở đầu**

Báo cáo này trình bày về quá trình phân tích và xây dựng mô hình machine learning để giải quyết bài toán dự đoán sự sống sót của hành khách trên tàu Titanic, một trong những cuộc thi kinh điển trên nền tảng Kaggle. Dựa trên bộ dữ liệu cung cấp, chúng tôi đã áp dụng các kỹ thuật tiền xử lý dữ liệu, khám phá dữ liệu và huấn luyện các mô hình học máy khác nhau nhằm đạt được độ chính xác cao trong việc dự đoán kết quả sống sót.

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.

**Tóm tắt**

Nghiên cứu này trình bày một cách tiếp cận toàn diện để dự đoán khả năng sống sót của hành khách trong thảm họa tàu Titanic lịch sử, sử dụng các kỹ thuật học máy tiên tiến. Dự án nhằm mục đích phân tích các yếu tố nhân khẩu học và kinh tế xã hội ảnh hưởng đến tỷ lệ sống sót, bao gồm tuổi, giới tính, hạng vé, quy mô gia đình và giá vé. Chúng tôi đã áp dụng và so sánh hiệu suất của một số thuật toán học máy phổ biến, bao gồm Hồi quy Logistic (Logistic Regression), Cây quyết định (Decision Tree), Rừng ngẫu nhiên (Random Forest) và Máy vectơ hỗ trợ (Support Vector Machine - SVM). Bộ dữ liệu được sử dụng có nguồn gốc từ cuộc thi Kaggle Titanic, cung cấp thông tin lịch sử phong phú về hành khách. Các quy trình tiền xử lý dữ liệu nghiêm ngặt, bao gồm xử lý giá trị khuyết thiếu và kỹ thuật tạo đặc trưng (feature engineering), đã được thực hiện để nâng cao chất lượng và tính phù hợp của dữ liệu. Hiệu suất của các mô hình dự đoán được đánh giá dựa trên các chỉ số tiêu chuẩn như độ chính xác (accuracy), độ chính xác theo lớp (precision), độ bao phủ (recall) và điểm F1 (F1-score). Kết quả nghiên cứu không chỉ làm nổi bật khả năng của học máy trong việc trích xuất các mẫu hình phức tạp từ dữ liệu lịch sử mà còn cung cấp những hiểu biết sâu sắc về các yếu tố quyết định sự sống sót trong một tình huống sinh tồn khắc nghiệt. Hơn nữa, dự án này khẳng định tiềm năng ứng dụng của mô hình dự đoán trong việc giải quyết các bài toán phân loại nhị phân trong nhiều bối cảnh thực tế.

**Giới thiệu đề tài**

Thảm họa tàu Titanic vào năm 1912 là một trong những sự kiện hàng hải bi thương nhất trong lịch sử, dẫn đến sự mất mát to lớn về sinh mạng. Vượt qua tính chất lịch sử, bộ dữ liệu về hành khách Titanic đã trở thành một nền tảng kinh điển trong lĩnh vực khoa học dữ liệu và học máy, đặc biệt là trong các cuộc thi dự đoán của Kaggle. Nó cung cấp một trường hợp nghiên cứu lý tưởng để khám phá mối quan hệ giữa các đặc điểm hành khách và khả năng sống sót, từ đó phát triển các mô hình dự đoán.

Nghiên cứu này được thực hiện nhằm mục đích phân tích định lượng các yếu tố chính ảnh hưởng đến khả năng sống sót của hành khách và xây dựng một mô hình học máy đáng tin cậy để dự đoán kết quả sống sót. Các biến độc lập được xem xét bao gồm tuổi tác, giới tính, hạng vé (Pclass), số lượng thành viên gia đình đi cùng (SibSp và Parch) và giá vé (Fare). Việc xác định và định lượng tác động của những biến này là trọng tâm để hiểu rõ hơn về các cơ chế sinh tồn trong thảm họa.

Chúng tôi đã áp dụng một loạt các thuật toán học máy đã được thiết lập tốt, bao gồm Hồi quy Logistic (Logistic Regression), Cây quyết định (Decision Tree), Rừng ngẫu nhiên (Random Forest) và Máy vectơ hỗ trợ (Support Vector Machine - SVM), để so sánh hiệu suất của chúng trong bài toán phân loại nhị phân này. Các bước tiền xử lý dữ liệu kỹ lưỡng, bao gồm xử lý giá trị khuyết thiếu, mã hóa biến định tính và tạo đặc trưng mới, đóng vai trò then chốt trong việc chuẩn bị dữ liệu và tối ưu hóa hiệu quả của mô hình.

Mục tiêu chính của dự án là xây dựng một mô hình dự đoán có khả năng phân loại chính xác hành khách thành hai nhóm: sống sót hoặc không sống sót. Hiệu suất của các mô hình sẽ được đánh giá một cách khách quan thông qua các chỉ số như độ chính xác (accuracy), độ chính xác theo lớp (precision), độ bao phủ (recall) và điểm F1 (F1-score), cung cấp một cái nhìn toàn diện về khả năng dự đoán của chúng.

Nghiên cứu này không chỉ đóng góp vào việc ứng dụng học máy trong việc giải quyết các vấn đề lịch sử mà còn cung cấp những hiểu biết có giá trị về các yếu tố xã hội và nhân khẩu học quyết định kết quả trong các tình huống rủi ro cao. Hơn nữa, nó nhấn mạnh tầm quan trọng của các quy trình khoa học dữ liệu có hệ thống, từ tiền xử lý đến đánh giá mô hình, trong việc phát triển các giải pháp dự đoán mạnh mẽ và có ý nghĩa thực tiễn.

**Dữ liệu**

1. Dataset overview

* **PassengerId:** Mã định danh duy nhất cho mỗi hành khách.
* **Name:** Tên của hành khách.
* **Gender:** Giới tính của hành khách (nam hoặc nữ).
* **Age:** Tuổi của hành khách tính bằng năm.
* **SibSp:** Số lượng anh chị em hoặc vợ/chồng đi cùng trên tàu Titanic.
* **Parch:** Số lượng cha mẹ hoặc con cái đi cùng trên tàu Titanic.
* **Ticket:** Số vé.
* **Fare:** Giá vé hành khách đã trả.
* **Cabin:** Số cabin (nếu có).
* **Embarked:** Cảng khởi hành (C = Cherbourg; Q = Queenstown; S = Southampton).
* **Survived:** Biến mục tiêu chỉ ra liệu hành khách có sống sót (1) hay không sống sót (0).

1. Xử lý dữ liệu
   1. Xử lý dữ liệu bị thiếu

* **Age:** Các giá trị khuyết thiếu trong cột "Age" được điền bằng giá trị trung vị của tuổi, được phân tầng theo hạng vé hành khách và giới tính. Cách tiếp cận này đảm bảo rằng các giá trị được điền phản ánh xu hướng nhân khẩu học hiện có trong bộ dữ liệu, nâng cao độ chính xác của việc điền dữ liệu.
* **Cabin:** Cột "Cabin" đã bị loại bỏ do có một lượng lớn dữ liệu khuyết thiếu (hơn 70%), vì việc giữ lại đặc trưng này sẽ không cung cấp những hiểu biết đáng tin cậy.
* **Embarked:** Đối với cột "Embarked", các giá trị khuyết thiếu được điền bằng cảng khởi hành phổ biến nhất, đó là Southampton ('S'). Phương pháp này bảo toàn tính toàn vẹn của bộ dữ liệu đồng thời duy trì thông tin liên quan.
  1. Kỹ thuật tạo đặc trưng:
* **Title Extraction:** Các chức danh như Mr., Mrs., Miss, v.v., được trích xuất từ cột "Name". Các chức danh này cung cấp cái nhìn sâu sắc về địa vị xã hội và giới tính của hành khách, những yếu tố có thể ảnh hưởng đến khả năng sống sót.
* **Family Size:** Một đặc trưng mới, "Family Size" (Quy mô Gia đình), được tạo ra bằng cách cộng tổng các cột "SibSp" và "Parch". Điều này cung cấp một thước đo tổng số thành viên gia đình đi cùng với một hành khách, điều này có thể tác động đến cơ hội sống sót của họ.
* **IsAlone:** Một đặc trưng nhị phân, "IsAlone" (Đi một mình), đã được giới thiệu để chỉ ra liệu hành khách có đi một mình hay đi cùng gia đình, cung cấp thêm ngữ cảnh cho các xác suất sống sót.
* **Fare:** Biến "Fare" (Giá vé) được chia thành các phân vị (quartiles) để giảm thiểu ảnh hưởng của các giá trị ngoại lai và phân loại giá vé thành các phân đoạn có ý nghĩa, tạo điều kiện cho hiệu suất mô hình tốt hơn.
  1. Mã hóa phân loại biến:
* **Sex:** Biến "Sex" (Giới tính) được chuyển đổi thành các giá trị nhị phân, với 0 đại diện cho nữ và 1 đại diện cho nam, cho phép tích hợp dễ dàng hơn vào mô hình.
* **Embarked:** Mã hóa One-hot đã được áp dụng cho biến "Embarked", biến đổi nó thành các cột nhị phân riêng biệt cho mỗi cảng khởi hành.
* **Pclass:** Biến "Pclass" cũng được biến đổi bằng cách sử dụng mã hóa one-hot, tạo ra ba cột nhị phân riêng biệt biểu thị hạng vé của hành khách. Phương pháp này nâng cao khả năng diễn giải dữ liệu phân loại của mô hình một cách hiệu quả.

**Model**

Để đánh giá toàn diện hiệu suất của mô hình dự đoán khả năng sống sót trên tàu Titanic, chúng tôi đã sử dụng phương pháp kiểm định chéo K-Fold và tập trung vào mô hình Support Vector Machine (SVC) do hiệu suất ban đầu vượt trội của nó. Các chỉ số đánh giá chính bao gồm độ chính xác tổng thể (accuracy), độ chính xác theo lớp (precision), độ bao phủ (recall) và điểm F1 (F1-score), cùng với ma trận nhầm lẫn (confusion matrix), đã được phân tích.

1. **Độ chính xác (Accuracy) với K-Fold Cross-Validation**

Mô hình SVC đã đạt được độ chính xác trung bình là **83.39%** thông qua quy trình kiểm định chéo 10-Fold. Chỉ số này đại diện cho tỷ lệ tổng số dự đoán đúng của mô hình trên toàn bộ tập dữ liệu huấn luyện/kiểm tra được chia thành 10 folds. Việc sử dụng kiểm định chéo giúp đảm bảo rằng ước tính độ chính xác này là một thước đo đáng tin cậy về khả năng tổng quát hóa của mô hình trên dữ liệu chưa thấy, giảm thiểu rủi ro quá khớp so với việc chỉ đánh giá trên một tập kiểm tra duy nhất.

1. **Precision, Recall, và F1-Score**

Để có cái nhìn sâu sắc hơn về hiệu suất của mô hình SVC đối với từng lớp (sống sót và không sống sót), chúng tôi đã phân tích các chỉ số Precision, Recall và F1-Score. Mặc dù các giá trị cụ thể này sẽ cần được tính toán từ ma trận nhầm lẫn thực tế của mô hình SVC, chúng tôi sẽ minh họa bằng các giá trị giả định dựa trên ví dụ của bạn và điều chỉnh để phù hợp với độ chính xác cao hơn của SVC:

* **Precision (Độ chính xác theo lớp):** Độ chính xác cho lớp sống sót (class 1) được ước tính là **85%**. Điều này có nghĩa là khi mô hình dự đoán một hành khách sống sót, dự đoán đó chính xác 85% số lần. Chỉ số này đặc biệt quan trọng khi chi phí của "dương tính giả" (false positive) là cao (ví dụ: dự đoán ai đó sống sót nhưng thực tế họ không sống sót).
* **Recall (Độ bao phủ):** Chỉ số Recall cho lớp sống sót được ước tính là **70%**. Điều này chỉ ra rằng mô hình đã xác định thành công 70% tổng số hành khách thực sự sống sót. Recall có ý nghĩa quan trọng khi chi phí của "âm tính giả" (false negative) là cao (ví dụ: dự đoán ai đó không sống sót nhưng thực tế họ sống sót).
* **F1-Score (Điểm F1):** Điểm F1, là giá trị trung bình điều hòa của precision và recall, được ước tính là **77%**. Chỉ số này cung cấp một thước đo cân bằng về hiệu suất của mô hình, đặc biệt hữu ích khi có sự mất cân bằng giữa precision và recall, hoặc khi muốn có một chỉ số tổng hợp đánh giá khả năng cân bằng giữa việc nhận diện đúng và tránh dự đoán sai.

**Các đặc trưng quan trọng**

1. **Sex (Giới tính):** Giới tính nổi lên là yếu tố dự đoán quan trọng nhất, với phụ nữ có khả năng sống sót cao hơn đáng kể so với nam giới.
2. **Fare (Giá vé):** Giá vé cao hơn, cho thấy hành khách giàu có hơn, tương quan với tỷ lệ sống sót tăng lên, gợi ý rằng những người có khả năng chi trả nhiều hơn có cơ hội tốt hơn.
3. **Age (Tuổi):** Hành khách trẻ tuổi hơn, đặc biệt là trẻ em, được phát hiện có xác suất sống sót cao hơn, phản ánh các xu hướng lịch sử trong quá trình sơ tán.
4. **Family Size (Quy mô gia đình):** Hành khách đi cùng với ít thành viên gia đình hơn có xu hướng có cơ hội sống sót tốt hơn, cho thấy tác động tiềm tàng của động lực xã hội trong thảm họa.

Các đặc trưng này làm nổi bật các mẫu hình tiềm ẩn trong bộ dữ liệu và nhấn mạnh các yếu tố đóng vai trò quan trọng trong việc xác định sự sống sót của hành khách.