**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**PHÂN TÍCH TỶ LỆ SỐNG SÓT TITANIC – PHÂN TÍCH THĂM DÒ VÀ XÂY DỰNG MÔ HÌNH**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhóm 14** | | | |
| Sinh viên thực hiện: | | | |
| STT | Họ tên | MSSV | Ngành |
| 1 | Đoàn Chí Hưng | 24550014 | CNTT |
| 2 | Hoàng Minh Nhật | 24550031 | CNTT |

**TP. HỒ CHÍ MINH – 12/2022**

# GIỚI THIỆU

Đề tài nhóm tập trung vào việc phân tích dữ liệu các hành khách tàu Titanic nhằm xác định ra các biến tương quan mạnh đến khả năng sống sót, tiếp theo, nhóm tinh chỉnh dữ liệu cho quá trình huấn luyện model, và cuối cùng, nhóm tổng hợp và đánh giá các thuật toán đánh giá để huấn luyện ra mô hình tổng hợp (Ensemble Model). Nhóm triển khai chu trình chuẩn của môn Phân tích Dữ liệu IE224: làm sạch, khám phá, xây dựng đặc trưng và mô hình hóa. Bộ công cụ chính gồm pandas cho tiền xử lý, seaborn/matplotlib để trực quan hóa và scikit-learn cho mô hình học máy. Bộ dữ liệu được tham khảo từ Kaggle [1]; nhóm không sử dụng lại đồ án trước đây và xây dựng quy trình xử lý riêng. Nhóm có sử dụng quy trình phân tích và huấn luyện từ dự án mẫu tại Kaggle Code [2] và có bổ sung thêm các phương pháp trực quan và tiền xử lý lý, đồng thời mở rộng lập luận và nhận xét cho các kết quả. Các bước chính bao gồm đánh giá chất lượng dữ liệu (đặc biệt là thiếu vắng ở Age và Cabin), trích xuất đặc trưng mở rộng (Title, Family Size, Deck, Ticket Group) và thử nghiệm ba thuật toán giám sát (Logistic Regression, Decision Tree, Random Forest) trước khi tổng hợp. Kết quả hiện tại cho thấy độ chính xác trên tập kiểm tra đạt 83.16% với Logistic Regression baseline và 81.52% với mô hình VotingClassifier sử dụng các mô hình tốt nhất sau tinh chỉnh.

# MÔ TẢ BỘ DỮ LỆU (BẮT BUỘC)

Bộ dữ liệu Titanic gồm 891 dòng và 12 biến mô tả thông tin hành khách. Trong đó có 3 biến số (Age, Fare, PassengerId) và 9 biến phân loại/định danh (Survived, Pclass, Name, Sex, SibSp, Parch, Ticket, Cabin, Embarked). Bộ dữ liệu được tham khảo tại Kaggle [1]; quá trình xử lý trong notebook đọc bản sao trên GitHub với cấu trúc giữ nguyên so với nguồn gốc.

|  |  |
| --- | --- |
| Tên biến | Ý nghĩa |
| PassengerId | Mã định danh cho hành khách |
| Survived | Trạng thái sống sót (1 sống sót, 0 l ngược lại |
| Pclass | Hạng vé (1, 2 hoặc 3) |
| Name | Họ tên đầy đủ (bao gồm chức danh) |
| Sex | Giới tính (Male/Female) |
| Age | Tuổi |
| SibSp | Số anh chị em/vợ chồng đi cùng |
| Parch | Số con cái và cha mẹ đi cùng |
| Ticket | Số vé |
| Fare | Giá vé |
| Cabin | Số cabin |
| Embarked | Tên bến cảng lên tàu (C/Q/S) |

Quá trình thăm dò chung tìm được một số đặc điểm nổi bật sau: tuổi trung bình 29.7 (độ lệch chuẩn 14.5), giá vé trung bình 32.2 bảng với giá trị lớn nhất 512.33. Tỷ lệ sống sót chung là 38.38%.

Ngoài ra, ba biến Age, Cabin và Embarked có tình trạng thiếu dữ liệu:

|  |  |  |  |
| --- | --- | --- | --- |
| Biến | Số giá trị thiếu | Tỷ lệ | Ghi chú |
| Age | 177 | 19.9% | Ảnh hưởng trực tiếp đến phân tầng độ tuổi và cần ước lượng lại theo nhóm Pclass, SibSp, Parch. |
| Cabin | 687 | 77.1% | Dữ liệu cabin rất rời rạc; giải pháp là trích deck từ ký tự đầu và gán Unknown khi khuyết. |
| Embarked | 2 | 0.2% | Thiếu không đáng kể; có thể điền giá trị phổ biến C. |

# PHƯƠNG PHÁP PHÂN TÍCH

Nhóm đã áp dụng những ký thuật phân tích thăm dò được học, đồng thời sử dụng thêm những kỹ thuật nâng cao cho quá trình xử lý nhãn bị thiếu, tiền xử lý và đặc biệt là quá trình lựa chọn thuật toán và huấn luyện mô hình tổng hợp.

A diagram of a purple rectangle with black arrows

AI-generated content may be incorrect.

Hình 1. Tổng quan quy trình phân tích và huấn luyện

## Đặt tên bài toán

Dựa vào bộ dữ liệu Titanic, nhóm quyết định sử dụng những phương pháp phân tích, kiểm tra, tiền xử lý và huấn luyện để tạo một model phân loại (classifier): Hành khách hoặc là sống sót hoặc không.

## Kiểm tra và tiền xử lý

* Nhóm bắt đầu quá trình phân tích thăm dò bằng các phương pháp thăm dò chung như: kiểm tra kích thước dataset; kiểu dữ liệu; phân bổ giá trị của từng biến; kiểm tra các nhãn bị thiếu.
* Tiếp theo, nhóm tìm ra các đặc điểm nổi bật của từng biến hoặc của các cặp biến. Ví dụ: Tương quan giữa Sex – Pclass - Fare đối với survived.
* Sau khi tìm được các đặc trưng nổi bật và những mối tương quan có giá trị, nhóm tiếp tục tìm các phương pháp xử lý giá trị bị thiếu.
* Đối với biến Age (thiếu ~20%), nhóm áp dụng phương pháp tìm giá trị trung vị (median) của ba biến có tương quan cao nhất đến Age là Pclass, SibSp và Parch.
* Đối với biến Embarked (thiếu 2 ô), nhóm sử dụng catplot để tìm ra bến cảng phổ biến nhất đối với hành khách có Fare = 80 (Pclass =1), nhóm kết luận rằng đối với Fare = 80, khả năng Embark cao nhất là C ( Cherbourgh).

## Xây dựng đặc trưng và mã hóa

Quá trình xây dựng đặc trưng và mã hóa bao gồm những điểm chính sau:

* One-hot encoding cho Sex, Embarked, Pclass và Ticket.
* Trích Title từ họ tên và chuẩn hóa về 4 nhóm chính nhằm phản ánh địa vị xã hội.
* Chuẩn hóa Fare  phân vị (quantile).

|  |  |  |
| --- | --- | --- |
| Biến mới | Biến được định dạng lại | Biến được mã hóa one-hot |
| Fsize & Family\_size (Parch + SibSp + 1) | Fare (quantile) | Embarked |
| IsAlone (Fsize = 0?) |  | Sex |
| Deck (trích từ chữ cái đầu của Cabin) |  | Pclass |
| Title (trích từ chức danh trong name) |  | Ticket |

## Lựa chọn thuật toán và huấn luyện

* Quá trình huấn luyện bắt đầu từ việc chia Train/Test để huấn luyện dự đoán từ ba mô hình riêng lẻ (Decision Tree, Random Forest, Logistic Regression).
* Từ kết quả của 3 thuật toán trên, nhóm tiếp tục sử dụng mô hình tổng hợp VotingClassifier với voting="soft". Mô hình cuối cùng sẽ tổng hợp kết quả dự đoán từ ba mô hình trước đó và sau đó cộng các xác suất này lại để đưa ra quyết định cuối cùng.
* Điều này giúp tận dụng điểm mạnh của từng mô hình và thường dẫn đến kết quả dự đoán tốt hơn so với bất kỳ mô hình đơn lẻ nào.

A close-up of several signs

AI-generated content may be incorrect.

Hình 2. Tổng quan quá trình lựa chọn thuật toán và huấn luyện model tổng hợp

# PHÂN TÍCH THĂM DÒ/SƠ BỘ

A group of blue bars

AI-generated content may be incorrect.

Qua quá trình phân tích thăm dò sơ bộ, nhóm tìm được các đặc trưng đáng chú ý sau:

* Sex và Pclass có sự chên lệch sống sót lớn nhất, nữ vé hạng 1 có khả năng sống sót gấp 3 lần nam hạng 3
* Age: Trẻ em dưới 12 tuổi có khả năng sống sót cao, nhóm 20-35 tuổi có khả năng sống sót thấp nhất.
* SibSp/Parch: Gia đình nhỏ có 1-2 người đi cùng có khả năng sống sót cao, trong khi đi một mình hoặc đi quá đông sẽ giảm khả năng.
* Embarked: Khách từ bến Cherbourgh có khả năng sống sót cao hơn 2 bến cảng còn lại.

# KẾT QUẢ PHÂN TÍCH (gợi ý)

# CHỈNH SỬA SAU BÁO CÁO

Style viết báo cáo

# KẾT LUẬN

Tóm tắt lại quá trình và kết quả hiện có.

Tóm tắt kết quả đạt được? Hạn chế?

Viết 10 dòng, hoặc nửa trang.

**Sinh viên trình bày không quá 10 trang cho tổng cộng 3 phần:**

**Giới thiệu.**

**Nội dung.**

**Kết luận.**

**(Không tính số trang tài liệu tham khảo)**

**Chú ý:**

* **Tại footer của báo cáo, nếu Họ tên SV dài có thể ghi tắt.   
  Ví dụ: Nguyễn Thị Thu Trang => Nguyễn T.T. Trang, hoặc N.T.T. Trang**
* **Nếu nhóm nhiều thành viên, không đủ chỗ ghi tên tại footer thì có thể không cần ghi.**

**TÀI LIỆU THAM KHẢO**

Kaggle. Link:<https://www.kaggle.com/datasets/yasserh/titanic-dataset/> (04/10/2025) .

Kaggle. <https://www.kaggle.com/code/berkaydogan1/titanic-ml-pipeline-end-to-end-classification> (04/10/2025)

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

***(Chú ý: Ghi rõ từng nhiệm vụ chi tiết của mỗi thành viên).***

Khi làm bài thu hoạch SV phải thực hiện đúng các quy định sau:

1. Sử đụng đúng Template này. Không đúng thì **– 2 điểm**.
2. Trình bày đề tài tối thiểu là 05 trang và tối đa là 10 trang A4
   * Không tính Trang Bìa, và Phụ lục phân công công việc trong template.
   * Phụ lục code
3. Copy y chan trên Internet thì gọi là đạo văn và **nhận 0 điểm** cho bài thu hoạch.
4. Nộp đúng hạn theo thông báo của GV.
5. SV đầu tiên trong danh sách đăng ký sẽ phụ trách nộp.
6. Không trình bày code trong nội dung bài thu hoạch. Nếu muốn thì trình bày tại phụ lục.

Sản phẩm:

1. File word (thêm pdf) báo cáo, ko in ra, nộp file mềm.
2. Code Demo kết quả.
3. Slide thuyết trình.
4. Bộ dữ liệu