**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**

# **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