TRƯỜNG ĐẠI HỌC SÀI GÒN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN MÔN HỌC

PHƯƠNG PHÁP NGHIÊN CỨU KHOA HỌC

ĐỀ TÀI:

DỰ ĐOÁN SỰ HÀI LÒNG CỦA KHÁCH HÀNG BẰNG CÁC PHƯƠNG PHÁP HỌC MÁY/HỌC SÂU

Nhóm sinh viên thực hiện:

Lê Hồng Sơn - 3121410423

Đỗ Hữu Lộc - 3123410201

Nguyễn Hoàng Thanh Phương - 3122410329

Văn Hoàng Như Ý - 3122410493

Giáo viên hướng dẫn: **TS. Đỗ Như Tài**

Thành phố Hồ Chí Minh, 05/2025

LỜI CẢM ƠN

Để hoàn thành dự án nghiên cứu khoa học "Dự đoán sự hài lòng của khách hàng bằng các phương pháp học máy/học sâu," chúng em đã nhận được rất nhiều sự giúp đỡ và hỗ trợ tận tình. Chúng em xin trân trọng gửi lời cảm ơn sâu sắc đến:

● Khoa Công Nghệ Thông Tin – Trường Đại Học Sài Gòn đã tạo mọi điều kiện thuận lợi để chúng em có thể thực hiện nghiên cứu này.

● Chúng em xin gửi lời tri ân đến thầy Đỗ Như Tài đã tận tình hướng dẫn, chỉ bảo trong suốt quá trình thực hiện đề tài. Sự định hướng và hỗ trợ của thầy đã giúp chúng em hoàn thành bài nghiên cứu một cách thuận lợi và hiệu quả.

● Các thành viên trong nhóm đã luôn đoàn kết, hỗ trợ lẫn nhau và nỗ lực hết mình để hoàn thành dự án với kết quả tốt nhất.

Cuối cùng, chúng em xin kính chúc các thầy cô luôn mạnh khỏe, thành công để tiếp tục dìu dắt các thế hệ học sinh, sinh viên trên con đường học tập và nghiên cứu.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

LỜI MỞ ĐẦU

Trong thời đại bùng nổ dữ liệu như hiện nay, việc hiểu rõ nhu cầu và cảm nhận của khách hàng đóng vai trò then chốt trong chiến lược phát triển của các doanh nghiệp, đặc biệt là trong lĩnh vực thương mại điện tử. Việc nắm bắt được mức độ hài lòng của khách hàng không chỉ giúp cải thiện chất lượng dịch vụ mà còn nâng cao khả năng cạnh tranh và giữ chân người dùng.

Song song với sự phát triển của công nghệ, các phương pháp học máy (Machine Learning) và học sâu (Deep Learning) đã và đang được ứng dụng rộng rãi trong việc phân tích hành vi người dùng và dự đoán xu hướng tiêu dùng. Việc áp dụng các kỹ thuật này vào bài toán dự đoán sự hài lòng của khách hàng mang lại nhiều triển vọng về độ chính xác cũng như hiệu quả xử lý dữ liệu lớn và phức tạp.

Xuất phát từ nhu cầu thực tiễn và tiềm năng nghiên cứu, đề tài *"Dự đoán sự hài lòng của khách hàng bằng phương pháp học máy và học sâu"* được thực hiện nhằm xây dựng mô hình dự đoán hiệu quả, đồng thời so sánh, đánh giá khả năng của các thuật toán học máy và học sâu trong bài toán cụ thể này.

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN ĐỀ TÀI

1.1. Đặt vấn đề

Trong những năm gần đây, thương mại điện tử đã nổi lên như một trong những phương thức mua sắm phổ biến nhất đối với người tiêu dùng do sự tiến bộ của công nghệ (Taher, 2021). Sự phát triển bền vững như vậy của thương mại điện tử đã dẫn đến sự phát triển của các doanh nghiệp trực tuyến [13].

Trong một thị trường cạnh tranh cao như vậy, ý kiến của khách hàng ảnh hưởng trực tiếp đến sự thành công của cả sản phẩm và công ty. Do đó, việc phân tích phản hồi của khách hàng về các khía cạnh khác nhau của sản phẩm là rất quan trọng để hiểu rõ khách hàng, cải thiện chất lượng sản phẩm và đảm bảo sự tồn tại trên thị trường [13].

Các nền tảng thương mại điện tử hiện nay cho phép khách hàng để lại đánh giá và nhận xét – không chỉ là thước đo hiệu quả kinh doanh mà còn là nguồn dữ liệu quý báu phục vụ cho việc dự đoán sự hài lòng của khách hàng [13]. Nhờ sự hỗ trợ của các kỹ thuật phân tích cảm xúc (Sentiment Analysis), học máy (Machine Learning) và học sâu (Deep Learning), doanh nghiệp có thể tự động trích xuất cảm xúc khách hàng từ dữ liệu lớn nhằm ra quyết định kinh doanh hiệu quả hơn.

1.2. Lý do chọn đề tài

Trong kỷ nguyên chuyển đổi số, các doanh nghiệp ngày càng chú trọng đến việc nâng cao trải nghiệm người dùng nhằm giữ chân khách hàng và gia tăng lợi thế cạnh tranh. Trong đó, sự hài lòng của khách hàng được xem là một trong những chỉ số quan trọng phản ánh chất lượng sản phẩm, dịch vụ cũng như hiệu quả vận hành của doanh nghiệp. Đặc biệt trong lĩnh vực thương mại điện tử, nơi quá trình tương tác giữa người dùng và hệ thống chủ yếu diễn ra thông qua nền tảng số, việc đánh giá và cải thiện mức độ hài lòng của khách hàng đóng vai trò then chốt trong việc xây dựng lòng trung thành và tối ưu hóa quy trình kinh doanh [1].

Tuy nhiên, các phương pháp truyền thống như khảo sát bảng hỏi hoặc đánh giá định tính thường tốn nhiều thời gian, chi phí, đồng thời không đảm bảo tính khách quan và kịp thời [4]. Với sự phát triển mạnh mẽ của trí tuệ nhân tạo, đặc biệt là các kỹ thuật học máy (Machine Learning) và học sâu (Deep Learning), việc tự động hóa phân tích dữ liệu phản hồi khách hàng để dự đoán mức độ hài lòng đã trở thành hướng nghiên cứu đầy tiềm năng. Các mô hình như Logistic Regression, Random Forest, Support Vector Machines, Artificial Neural Network (ANN) và Convolutional Neural Network (CNN) đã được chứng minh là có hiệu quả cao trong việc khai thác các đặc trưng tiềm ẩn từ dữ liệu lớn và phi cấu trúc như văn bản hoặc đánh giá người dùng [10].

Tại Việt Nam, một số nghiên cứu bước đầu đã ứng dụng các thuật toán học máy để phân tích bình luận trực tuyến trong lĩnh vực khách sạn và thương mại điện tử, qua đó rút ra các yếu tố chính ảnh hưởng đến sự hài lòng và đưa ra các mô hình phân loại tự động có độ chính xác cao [12]. Tuy vậy, nhiều thách thức vẫn còn tồn tại như dữ liệu không đồng nhất, thiên lệch cảm xúc trong phản hồi khách hàng, hoặc hiện tượng quá khớp (overfitting) khi áp dụng các mô hình học sâu trên tập dữ liệu không đủ đa dạng.

Trước thực tiễn đó, nghiên cứu này được thực hiện nhằm xây dựng và đánh giá hiệu quả của các mô hình học máy và học sâu trong việc dự đoán sự hài lòng của khách hàng trên nền tảng thương mại điện tử, từ đó đề xuất giải pháp ứng dụng thực tế hỗ trợ doanh nghiệp cải thiện chất lượng dịch vụ.

1.3. Mục tiêu nghiên cứu

Nghiên cứu hướng đến các mục tiêu cụ thể như sau:

- Phân tích các yếu tố ảnh hưởng đến sự hài lòng của khách hàng trong lĩnh vực thương mại điện tử.
- Áp dụng các kỹ thuật học máy (Machine Learning) và học sâu (Deep Learning) trong phân tích cảm xúc để dự đoán hành vi khách

hàng và hỗ trợ doanh nghiệp đưa ra quyết định kinh doanh hiệu quả.

- Đề xuất mô hình tối ưu có khả năng ứng dụng thực tế nhằm hỗ trợ doanh nghiệp ra quyết định trong cải tiến dịch vụ.

1.4. Đối tượng và phạm vi nghiên cứu

1.4.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là các phản hồi, đánh giá và bình luận của khách hàng dưới dạng văn bản trên các nền tảng thương mại điện tử. Đây là nguồn dữ liệu phi cấu trúc mang tính chủ quan, thể hiện ý kiến, cảm nhận và mức độ hài lòng của người tiêu dùng sau khi trải nghiệm sản phẩm hoặc dịch vụ.

1.4.2. Phạm vi nghiên cứu

Phạm vi nghiên cứu tập trung vào môi trường thương mại điện tử nói chung, không giới hạn trong một ngành hàng cụ thể như mỹ phẩm, thời trang hay điện tử. Dữ liệu khảo sát và phân tích được thu thập từ các nền tảng mua sắm trực tuyến phổ biến tại Việt Nam và/hoặc quốc tế (ví dụ: Shopee, Lazada, Tiki, Amazon...), với thời gian thu thập nằm trong giai đoạn nhất định, đảm bảo tính cập nhật và phù hợp với mục tiêu nghiên cứu. Phạm vi kỹ thuật bao gồm việc áp dụng các mô hình học máy và học sâu để xử lý ngôn ngữ tự nhiên (NLP), trích xuất cảm xúc từ văn bản, và đánh giá mức độ hài lòng của khách hàng.

1.5. Phương pháp nghiên cứu

1.5.1. Cách tiếp cận nghiên cứu

Đề tài sử dụng cách tiếp cận định lượng kết hợp với phương pháp phân tích văn bản để khai thác thông tin từ các đánh giá, nhận xét của khách hàng trên nền tảng thương mại điện tử. Nghiên cứu hướng đến việc áp dụng các kỹ

thuật học máy (Machine Learning) và học sâu (Deep Learning) trong phân tích cảm xúc nhằm phát hiện xu hướng và mức độ hài lòng của người tiêu dùng.

1.5.2. Các nguồn dữ liệu và cách thu thập dữ liệu

Dữ liệu phục vụ cho nghiên cứu được lấy từ cuộc thi trên nền tảng Kaggle có tên "Predict the Customer Satisfaction - CSE 22". Dataset bao gồm các thông tin phản hồi, đánh giá của khách hàng liên quan đến trải nghiệm mua sắm trên các nền tảng thương mại điện tử. Đây là nguồn dữ liệu phong phú và đa dạng, giúp mô hình học máy và học sâu có thể phân tích và dự đoán chính xác mức độ hài lòng của khách hàng dựa trên các phản hồi thực tế.

1.5.3. Quy trình thực hiện

- Xác định bài toán nghiên cứu.
- Thu thập dữ liệu đầu vào.
- Tiền xử lý dữ liệu.
- Xây dựng mô hình học máy/học sâu.
- Đánh giá và so sánh mô hình
- Kết luận và đề xuất hướng ứng dụng.

1.5.4. Dự kiến các phương pháp nghiên cứu

Đề tài dự kiến sử dụng phương pháp nghiên cứu định lượng với sự hỗ trợ của các công cụ khai phá dữ liệu và học máy/học sâu. Cụ thể:

- Phân tích thống kê mô tả để hiểu rõ đặc điểm dữ liệu.
- Học máy (Machine Learning): Sử dụng mô hình Random Forest để xử lý các đặc trưng dạng bảng và phát hiện các yếu tố quan trọng ảnh hưởng đến sự hài lòng. Áp dụng LightGBM để tận dụng ưu điểm về tốc độ và độ chính xác trong xử lý dữ liệu lớn và phức tạp.
- Học sâu (Deep Learning): Sử dụng mô hình CNN để khai thác đặc trưng trong dữ liệu văn bản phản ánh cảm xúc/hài lòng của khách hàng.

- Công cụ và thư viện: Sử dụng Python và các thư viện xử lý dữ liệu như**:** numpy (hỗ trợ tính toán số học, xử lý mảng và ma trận), pandas (hỗ trợ xử lý dữ liệu dạng bảng (DataFrame), thao tác dữ liệu dễ dàng).

1.6. Câu hỏi nghiên cứu

Đề tài đặt ra các câu hỏi nghiên cứu chính như sau:

- Những yếu tố nào có ảnh hưởng đáng kể đến mức độ hài lòng của khách hàng trên nền tảng thương mại điện tử?
- Mô hình học máy hoặc học sâu nào đạt độ chính xác cao nhất trong việc dự đoán sự hài lòng?
- Làm thế nào để tối ưu hóa mô hình nhằm cải thiện hiệu suất dự đoán và khả năng ứng dụng thực tiễn?

CHƯƠNG 2. TỔNG QUAN LÝ THUYẾT

2.1. Sự hài lòng của khách hàng

2.1.1. Định nghĩa sự hài lòng của khách hàng

Sự hài lòng của khách hàng là sự đánh giá dựa trên kinh nghiệm sử dụng một dịch vụ hoặc sản phẩm cụ thể trong một thời gian. Sự hài lòng của khách hàng là kim chỉ nam cho các doanh nghiệp khi triển khai marketing cho thấy sự tin tưởng và hài lòng của khách hàng khi trải nghiệm các dịch vụ và sản phẩm marketing có liên quan đến mức độ gắn bó với thương hiệu và lợi thế cạnh tranh so với đối thủ, từ đó giúp doanh nghiệp cải thiện [13].

2.1.2. Vai trò của sự hài lòng trong kinh doanh

- Đảm bảo sự trung thành và duy trì khách hàng: bởi sự hài lòng của khách hàng là yếu tố then chốt để biết khách hàng có trung thành không, và nó ảnh hưởng trực tiếp đến sự thành công lâu dài của doanh nghiệp [13].
- Yếu tố thúc đẩy doanh thu và lan tỏa thương hiệu: khách hàng hài lòng thường sẽ quay lại mua hàng, giới thiệu cho người khác, từ đó giúp doanh nghiệp phát triển hơn [13]. Những công ty có điểm hài lòng khách hàng cao thường sẽ đạt doanh thu lớn hơn và xây dựng được lòng trung thành với thương hiệu [10].
- Tác động đến quyết định chiến lược kinh doanh, nâng cao dịch vụ: việc có thể dự đoán chính xác sự hài lòng giúp doanh nghiệp chủ động hơn trong quản lý và điều chỉnh dịch vụ – điều rất quan trọng trong thị trường cạnh tranh hiện nay [10].

2.1.3. Các cách tiếp cận đo lường sự hài lòng

Cách tiếp cận	Loại dữ liệu	Ưu điểm	Hạn chế
Khảo sát truyền thống	Định lượng (Likert)	Dễ triển khai, trực tiếp	Thiên lệch chủ quan, tỉ lệ phản hồi thấp

NPS (Net Promoter Score)	Định lượng	Đơn giản, phổ biến	Không cho biết nguyên nhân cụ thể
Phân tích phản hồi	Văn bản tự do	Phản ánh cảm xúc chân thực	Cần kỹ thuật NLP nâng cao
Học máy & học sâu	Dữ liệu lớn (review, log)	Dự đoán chính xác, hiệu quả	Cần dữ liệu nhiều và kỹ thuật phức tạp
UX & hành vi người dùng	Log truy cập, hành vi click	Không cần khảo sát, đo tự động	Khó liên kết trực tiếp đến hài lòng

Bảng 2.1. Các cách tiếp cận đo lường sự hài lòng

2.2. Các yếu tố ảnh hưởng đến sự hài lòng

2.2.1. Chất lượng sản phẩm dịch vụ

Sự hài lòng bị ảnh hưởng nhiều nhất bởi chất lượng của sản phẩm, dịch vụ doanh nghiệp cung cấp. Chất lượng của sản phẩm chính là yếu tố quyết định việc nhu cầu của khách hàng có được thỏa mãn hay không.

2.2.2. Chất lượng dịch vụ khách hàng

Dịch vụ chăm sóc khách hàng tốt là điểm cộng cho doanh nghiệp, khiến khách hàng hài lòng hơn khi sử dụng sản phẩm, dịch vụ. Một quy trình chăm sóc khách hàng nhiệt tình, sát sao và chuyên nghiệp chắc chắn gây được ấn tượng tốt đẹp trong lòng khách hàng.

2.2.3. Giá cả sản phẩm

Giá cả của sản phẩm cũng là yếu tố gia tăng sự hài lòng trong tâm trí khách hàng. Người mua hàng mong muốn giá cả và chất lượng sản phẩm phải đi đôi với nhau. Chính vì vậy, doanh nghiệp không chỉ cần chú trọng về chất lượng mà còn phải nghiên cứu kỹ giá thành của thị trường để đưa ra giá bán hợp lý.

2.2.4. Tính tiện lợi

Tính tiện lợi ở đây đề cập đến sự thuận tiện trong quá trình mua hàng, sử dụng và gặp các vấn đề liên quan. Một số khâu quan trọng được khách hàng đánh giá mức độ tiện lợi khi mua hàng là:

- Tìm kiếm thông tin: Thông tin về sản phẩm được doanh nghiệp cung cấp đầy đủ, rõ ràng, dễ dàng tìm kiếm.
- Tư vấn mua hàng: Khách hàng được tư vấn kỹ lưỡng, được trả lời những thắc mắc nhanh chóng để có thể đưa ra quyết định mua hàng sớm.
- Thanh toán: Quy trình mua hàng, thanh toán tiện lợi với nhiều phương thức trả tiền khác nhau như tiền mặt, chuyển khoản, thẻ tín dụng,…
- Giao nhận hàng: Quá trình giao hàng được thông tin rõ ràng, cập nhật thường xuyên, sản phẩm có thể được nhận linh hoạt ở nhiều nơi giúp gia tăng sự hài lòng của khách hàng.
- Sau khi mua hàng: Khách hàng được hỗ trợ 24/7 khi gặp vấn đề về sản phẩm,…

2.2.5. Chính sách - chương trình ưu đãi

Chính sách, chương trình ưu đãi sẽ có thể làm tăng mạnh mức độ hài lòng của khách hàng về sản phẩm. Những chương trình ưu đãi đang được nhiều người tiêu dùng săn đón, nhất là trong bối cảnh mua hàng trực tuyến phát triển mạnh mẽ như hiện nay.

2.3. Tổng quan về học máy (Machine Learning)

2.3.1. Khái niệm và phân loại học máy

Machine Learning (Học máy) là tập hợp con của AI tập trung vào phát triển các thuật toán có khả năng học hỏi mà không cần lập trình rõ ràng

2.3.2. Các thuật toán học máy phổ biến

Trong học máy, có nhiều thuật toán được sử dụng để phân loại, dự đoán hoặc phân cụm dữ liệu. Trong phạm vi đề tài này, hai thuật toán phổ biến được sử dụng gồm:

- Random Forest: Là mô hình học máy theo phương pháp Bagging, kết hợp nhiều cây quyết định (Decision Tree) để tăng độ chính xác và giảm overfitting. Mỗi cây trong rừng được huấn luyện trên một tập con khác nhau của dữ liệu, sau đó kết quả được tổng hợp bằng cách bỏ phiếu (voting).
- LightGBM (Light Gradient Boosting Machine): Là thuật toán boosting được phát triển bởi Microsoft, nổi bật với khả năng xử lý dữ liệu lớn, tốc độ huấn luyện nhanh và hiệu quả cao. LightGBM hoạt động bằng cách xây dựng các cây quyết định theo hướng giảm dần độ lỗi (gradient-based) và sử dụng phương pháp Leaf-wise để chia nhánh.

2.3.3. Ứng dụng của học máy trong dự đoán sự hài lòng

Trong lĩnh vực thương mại điện tử, việc dự đoán sự hài lòng của khách hàng dựa trên các đặc trưng như hành vi mua sắm, phản hồi, lịch sử giao dịch,... là một ứng dụng quan trọng của học máy.

- Các thuật toán học máy như Random Forest và LightGBM có thể được huấn luyện để phân loại mức độ hài lòng dựa trên dữ liệu sẵn có.
- Việc ứng dụng các mô hình này không chỉ giúp doanh nghiệp dự đoán chính xác mức độ hài lòng mà còn đưa ra các chiến lược cải thiện dịch vụ phù hợp.
- Trong đề tài này, các mô hình học máy được xây dựng và đánh giá nhằm tìm ra thuật toán phù hợp nhất để giải quyết bài toán phân loại sự hài lòng của khách hàng từ dữ liệu thương mại điện tử.

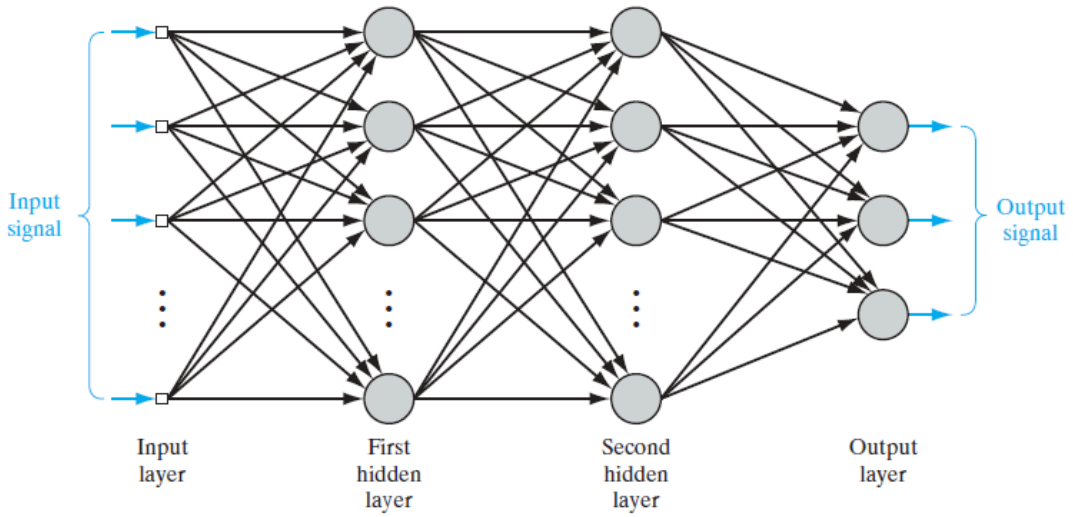
2.4. Tổng quan về học sâu (Deep Learning)

2.4.1. Khái niệm và nguyên lý hoạt động

Deep Learning (Học sâu) là tập hợp con của Machine Learning. Các thuật toán của Deep Learning được lấy cảm hứng từ cấu trúc não bộ con người và hoạt động cực kỳ hiệu quả với dữ liệu phi cấu trúc như hình ảnh, video hoặc văn bản.

2.4.2. Mạng nơ-ron nhân tạo (ANN), CNN, RNN

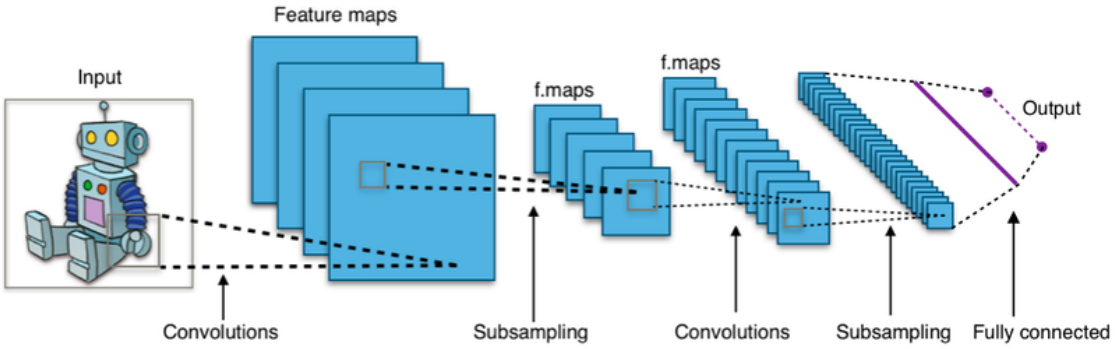
- Mạng nơ-ron nhân tạo (ANN - Artificial Neural Network): ANN là cấu trúc cơ bản nhất trong Deep Learning, mô phỏng hoạt động của nơ-ron sinh học. Mạng bao gồm nhiều lớp (layer) với các nơ-ron liên kết chặt chẽ, có thể học các hàm ánh xạ phức tạp từ dữ liệu đầu vào đến đầu ra. ANN thường được dùng trong các bài toán phân loại hoặc hồi quy tổng quát.



Hình 2.1. Mô hình ANN

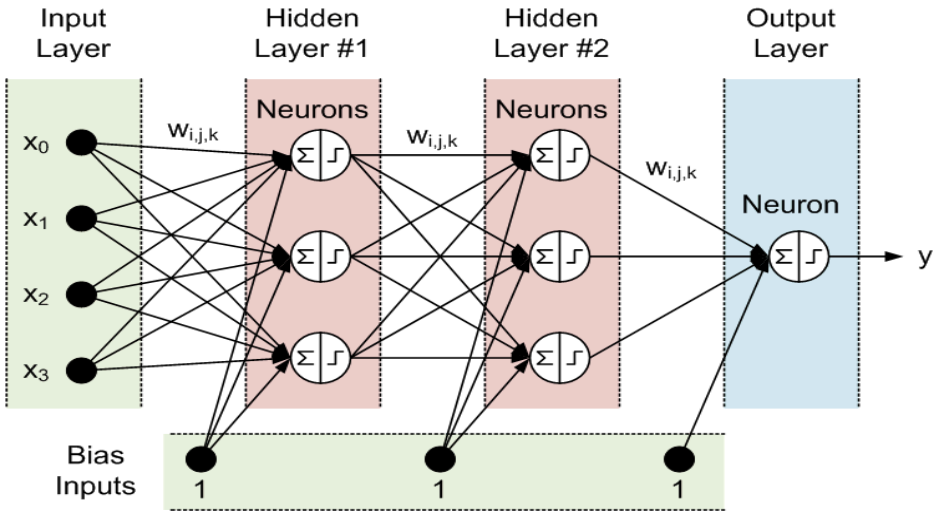
- Mạng nơ-ron tích chập (CNN - Convolutional Neural Network): CNN được thiết kế chủ yếu cho xử lý dữ liệu có cấu trúc dạng lưới như hình ảnh, nhưng cũng có thể áp dụng cho chuỗi dữ liệu sau khi được xử lý phù hợp. CNN sử dụng các lớp tích chập và gộp (pooling) để tự động trích

xuất đặc trưng, rất hiệu quả trong bài toán dự đoán sự hài lòng khách hàng khi có nhiều đặc trưng tương quan.



Hình 2.2. Mô hình CNN

- Mạng nơ-ron hồi tiếp (RNN - Recurrent Neural Network): RNN chuyên dùng cho dữ liệu chuỗi thời gian hoặc văn bản vì khả năng ghi nhớ thông tin trước đó trong chuỗi. Nó thích hợp cho việc phân tích hành vi khách hàng qua thời gian như lịch sử giao dịch, phản hồi, hoặc truy vấn.



Hình 2.3. Mô hình RNN

2.4.3. Lợi ích và hạn chế của học sâu trong xử lý dữ liệu khách hàng

- Lợi ích: Deep Learning (học sâu) mang lại nhiều lợi ích đáng kể trong việc phân tích và dự đoán sự hài lòng của khách hàng trong thương mại điện tử:

+ Khả năng phát hiện mẫu ẩn trong dữ liệu: các mô hình học sâu như CNN có thể nhận diện các mẫu tiềm ẩn phức tạp liên quan đến hành vi và cảm nhận của khách hàng mà các mô hình truyền thống khó phát hiện [13].
+ Xử lý mối quan hệ phi tuyến: Deep Learning được ghi nhận là vượt trội trong việc học các mối quan hệ phi tuyến tính giữa các thuộc tính dữ liệu, giúp cải thiện đáng kể độ chính xác của mô hình [10].
+ Hiệu quả với dữ liệu phi cấu trúc: Do đặc trưng kiến trúc mạng nơ-ron sâu, học sâu xử lý tốt dữ liệu như văn bản đánh giá, hình ảnh sản phẩm hoặc phản hồi khách hàng, mở rộng phạm vi ứng dụng so với các thuật toán học máy truyền thống.

- Hạn chế:
  + Yêu cầu tài nguyên tính toán cao: Các mô hình học sâu cần khối lượng dữ liệu lớn và phần cứng mạnh để huấn luyện (theo bài báo thứ nhất), gây khó khăn khi triển khai thực tế trong các doanh nghiệp vừa và nhỏ.
  + Thiếu khả năng giải thích: Dù đạt độ chính xác cao, mô hình học sâu thường bị xem là "hộp đen" (black-box), khó giải thích nguyên nhân dẫn đến một dự đoán cụ thể (theo bài báo thứ hai). Điều này là một trở ngại trong việc ra quyết định dựa trên kết quả mô hình, đặc biệt trong các chiến lược chăm sóc khách hàng.

2.5. Các mô hình dự đoán phổ biến

2.5.1. Mô hình hồi quy tuyến tính (Linear Regression):

Mô hình dự đoán giá trị liên tục dựa trên mối quan hệ tuyến tính giữa các biến.Đơn giản, dễ hiểu, thường dùng cho bài toán dự đoán số liệu (giá, điểm, doanh thu...).

2.5.2. Rừng ngẫu nhiên (Random Forest):

Mô hình phân nhánh theo điều kiện để đưa ra quyết định hoặc dự đoán. Dễ trực quan hóa, hoạt động tốt với dữ liệu có cấu trúc rõ ràng, dễ bị overfitting nếu không kiểm soát độ sâu.

2.5.3. Máy vector hỗ trợ (SVM):

Mô hình phân loại tìm đường biên tối ưu giữa các nhóm dữ liệu. Hiệu quả cao với dữ liệu có ít nhiễu, hoạt động tốt với cả bài toán tuyến tính và phi tuyến.

2.5.4. Mạng nơ-ron tích chập (CNN):

Mô hình học sâu mô phỏng não người, đặc biệt hiệu quả trong xử lý ảnh. Tự động trích xuất đặc trưng, mạnh với dữ liệu thị giác (ảnh, video), yêu cầu dữ liệu lớn và tài nguyên tính toán cao.

2.5.5. So sánh và lựa chọn mô hình phù hợp

Mô hình	Ưu điểm	Hạn chế	Phù hợp với
Linear Regression	Dễ hiểu, nhanh, dễ triển khai	Không phù hợp với quan hệ phi tuyến	Dữ liệu tuyến tính, bài toán đơn giản
Random Forest	Chính xác cao hơn Decision Tree, giảm overfitting	Khó giải thích hơn, chậm hơn một chút	Bài toán dự đoán phân loại phức tạp
SVM	Tốt với dữ liệu nhiều chiều, phân nhóm rõ ràng	Chạy chậm với tập dữ liệu lớn	Dự đoán chính xác, dữ liệu vừa phải
CNN (Deep Learning)	Khả năng tự học đặc trưng, hiệu quả với dữ liệu phi cấu trúc (ảnh, text)	Đòi hỏi dữ liệu lớn và tài nguyên tính toán cao	Phân tích cảm xúc, văn bản, ảnh

Bảng 2.2. So sánh các mô hình dự đoán phổ biến

=> Nếu bài toán yêu cầu độ chính xác cao, có đủ tài nguyên tính toán và dữ liệu đa dạng (bao gồm văn bản hoặc ảnh), các mô hình học sâu như CNN sẽ là lựa chọn tối ưu. Nếu yêu cầu tốc độ và dễ triển khai, Random Forest hoặc SVM là lựa chọn phù hợp hơn.

CHƯƠNG 3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Quy trình nghiên cứu

Để dự đoán sự hài lòng của khách hàng bằng các phương pháp học máy và học sâu, nhóm đã xây dựng quy trình nghiên cứu gồm 6 bước chính, đảm bảo logic và tính thực tiễn:

- *Xác định bài toán nghiên cứu:* Làm rõ mục tiêu chính là dự đoán mức độ hài lòng của khách hàng trong thương mại điện tử dựa trên các đặc trưng hành vi, lịch sử mua hàng và phản hồi.
- *Thu thập dữ liệu đầu vào***:** Tiến hành thu thập dữ liệu, trong đó chứa các thông tin chi tiết về giao dịch, hành vi khách hàng, phương thức thanh toán, ưu đãi nhận được và trải nghiệm người dùng.
- *Tiền xử lý dữ liệu:* Làm sạch và chuẩn hóa dữ liệu: loại bỏ các cột không cần thiết, xử lý giá trị thiếu, xử lý dữ liệu phân loại (giới tính, hình thức thanh toán...), mã hóa các trường văn bản.
- *Xây dựng mô hình học máy/học sâu*
    + Áp dụng nhiều thuật toán như Logistic Regression, Random Forest, hoặc ANN nhằm tìm ra mô hình dự đoán tối ưu.
    + Lựa chọn mô hình dựa trên loại dữ liệu và tính chất bài toán (phân loại hay hồi quy).
- *Đánh giá và so sánh mô hình*
    + Sử dụng các chỉ số đánh giá như Accuracy, F1-Score, RMSE để đo lường hiệu suất mô hình.
    + So sánh giữa các mô hình nhằm chọn ra mô hình tốt nhất phục vụ cho việc triển khai thực tế.
- *Kết luận và đề xuất hướng ứng dụng***:** Tổng kết kết quả thực nghiệm, từ đó đề xuất giải pháp cụ thể giúp doanh nghiệp ứng dụng mô hình vào việc cải thiện chất lượng dịch vụ, tăng mức độ hài lòng của khách hàng.

3.2. Thu thập dữ liệu

Dữ liệu nghiên cứu được lấy từ một tập tin CSV có tên train_dataset.csv. Tập dữ liệu này mô phỏng phản hồi từ người dùng trong môi trường thương mại điện tử, bao gồm các thông tin như ID người dùng, hình thức thanh toán, phương thức mua hàng, mức chiết khấu, điểm trung thành đã sử dụng, v.v.

Sau khi tải lên, dữ liệu được đọc vào bằng thư viện pandas, từ đó phục vụ cho các bước phân tích và huấn luyện mô hình.

3.3. Tiền xử lý dữ liệu

3.3.1 Giới thiệu tập dữ liệu

Tập dữ liệu được sử dụng trong dự án bao gồm thông tin chi tiết về người dùng, chương trình khách hàng thân thiết, sản phẩm, giao dịch, vận chuyển và trải nghiệm khách hàng. Cụ thể:

- Người dùng:
    + user_id
    + age
    + Gender
    + Date_Registered
- Chương trình khách hàng thân thiết:
    + Is_current_loyalty_program_member
    + loyalty_points_redeemed, loyalty_tier
    + Received_tier_discount_percentage
    + Received_card_discount_percentage
    + Received_coupon_discount_percentage
- Sản phẩm:
    + product_category
    + Product_value
- Giao dịch:
    + transaction_id

+ order_id
+ payment_method
+ payment_datetime
+ purchased_datetime
+ purchase_medium
+ final_payment

- Giao hàng:
  + released_date
  + estimated_delivery_date
  + received_date
  + shipping_method
  + tracking_number
- Biến mục tiêu: customer_experience (có giá trị: good, neutral, bad)

3.3.2 Các bước tiền xử lý dữ liệu

Bước 1. Làm sạch dữ liệu:

- Xoá các cột không cần thiết như id, transaction_id, order_id, tracking_number.

```
train_set = train_set.drop(['id','transaction_id', 'order_id',  'tracking_number'], axis=1)
```

- Xử lý định dạng user_id, loại bỏ ký tự đặc biệt.

```
train_set['user_id'] = train_set['user_id'].str.replace('****', '', regex=False)
```

- Điền giá trị thiếu bằng 0 cho các cột như: loyalty_tier, Received_tier_discount_percentage, received_card_discount_percentage.

```
train_set['loyalty_tier'] = train_set['loyalty_tier'].fillna(0)
train_set['Received_tier_discount_percentage'] = train_set['Received_tier_discount_percentage'].fillna(0)
train_set['Received_card_discount_percentage'] = train_set['Received_card_discount_percentage'].fillna(0)
```

Bước 2. Chuẩn hóa dữ liệu:

- Chuẩn hóa giá trị trong các cột như Gender (M, F, O → Male, Female, Other) và payment_method (visa_c, mastercard_c, gcash,... thành các nhóm phương thức thanh toán thống nhất).

```
train_set['Gender'] = train_set['Gender'].replace({
    'O': 'Other',
    'F': 'Female',
    'M': 'Male'
})
```

```
train_set['payment_method'] = train_set['payment_method'].replace({
    'amex': 'American Expres',
    'visa_c': 'Credit Card',
    'mastercard_c':'Credit Card',
    'visa_d': 'Debit Card',
    'mastercard_d': 'Debit Card',
    'gcash':'G-cash',
    'maya':'Maya',
    'coinsph':'Coin-Sph',
    'grabpay':'Grab-Pay',
    'shopeepay':'Shopee-Pay'
})
```

```
for item in columns:
    print(f"Nhãn dữ liệu từ {item}:", train_set[item].unique())
```

Bước 3. Loại bỏ dữ liệu trùng lặp:

- Sau khi kiểm tra, một số bản ghi bị trùng lặp đã được phát hiện và loại bỏ nhằm đảm bảo tính chính xác và độ tin cậy của mô hình huấn luyện.

- Danh sách dữ liệu trùng lặp

```
user_id = train_set['user_id']
train_set = train_set.drop(['user_id'], axis=1)
```

- Loại bỏ trùng lặp

```
train_set.drop_duplicates(inplace=True)
```

```
train_set = Interquartile_Range(train_set, Quartile='Product_value')
train_set = Interquartile_Range(train_set, Quartile='final_payment')
```

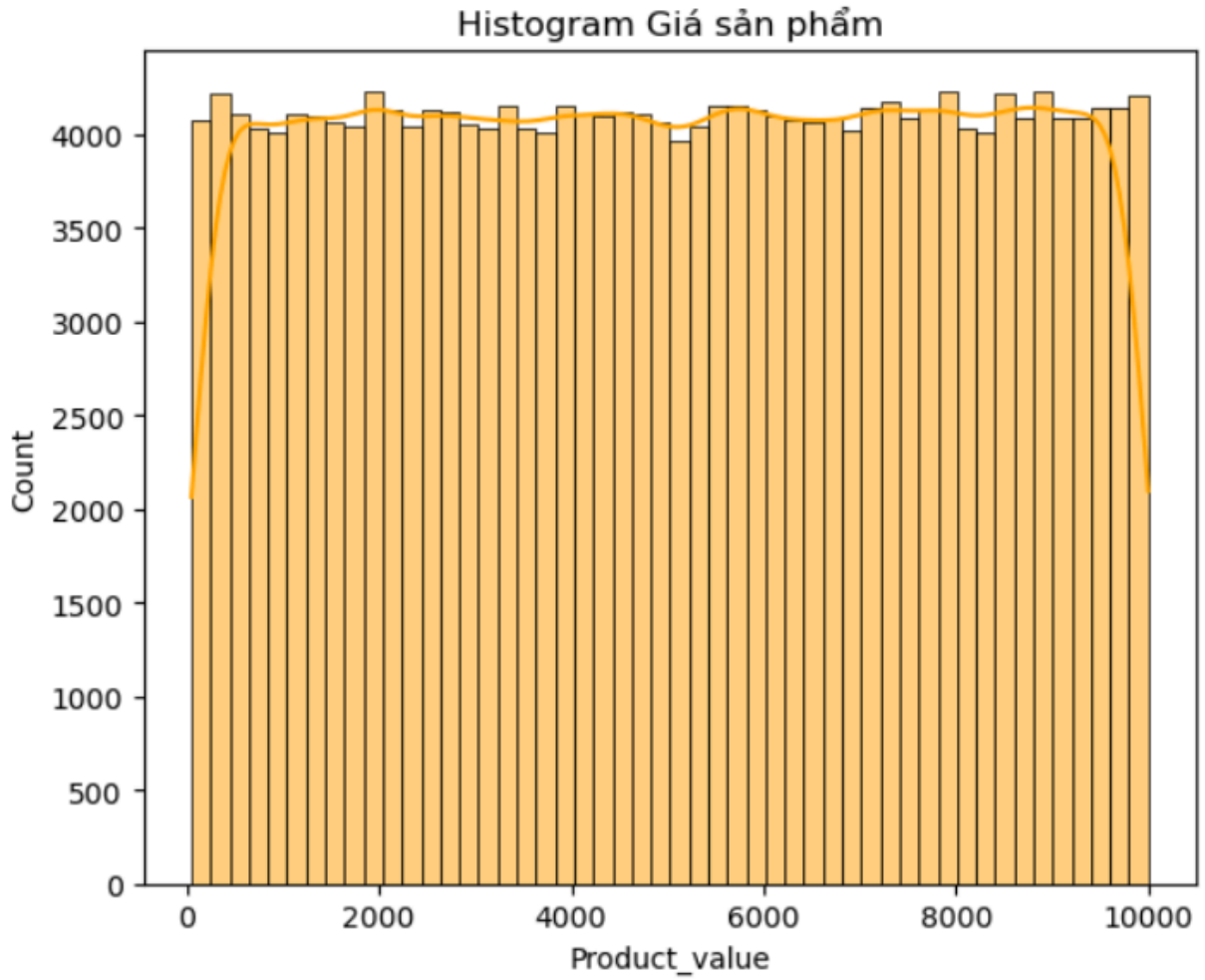
Bước 4. Xử lý dữ liệu ngoại lai:

- Sử dụng phương pháp Interquartile Range (IQR) để loại bỏ các giá trị ngoại lai trong các cột Product_value và final_payment.

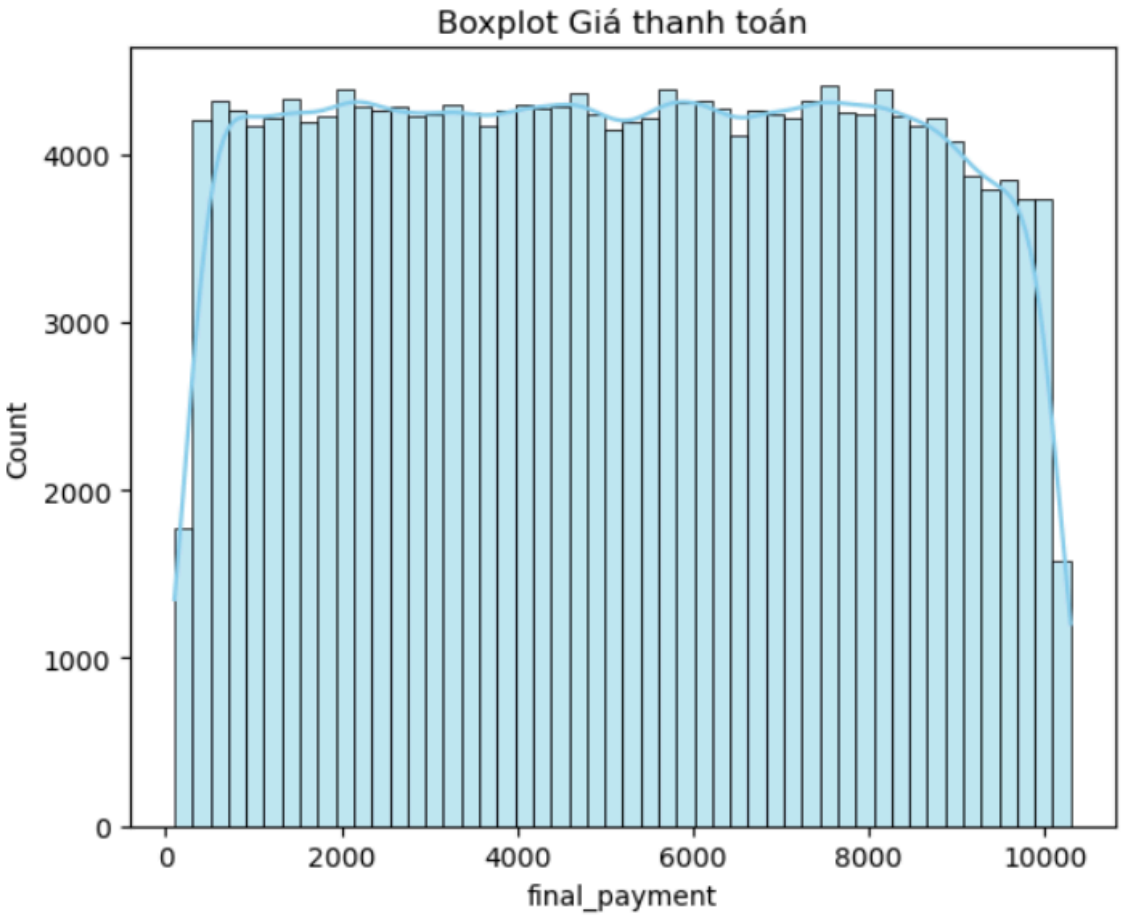
```
def Interquartile_Range(data,Quartile):
    Q1 = data[Quartile].quantile(0.25)
    Q3 = data[Quartile].quantile(0.75)
    Iqr = Q3 - Q1
    data_clean = data[(data[Quartile] >= Q1 - 1.5 * Iqr) & (data[Quartile] <= Q3 + 1.5 * Iqr)]
    return data_clean
```

Bước 5. Trực quan hóa dữ liệu:

- Vẽ biểu đồ để kiểm tra phân phối của các cột liên quan như giá sản phẩm, điểm thưởng, độ tuổi khách hàng.



Biểu đồ histogram



Biểu đồ Boxplot

Bước 6. Feature Engineering

- Phân nhóm độ tuổi thành các nhóm: Children, Adolescent, Teenager, Adult, Middle, Elderly.

```
ranks = [0,10,15,20,30,45,100]
labels_age = ['Children' , 'Adolescent' , 'Teenager', 'Adult', 'Middle', 'Elderly']
```

```
train_set['Age-Group'] = pd.cut(train_set['age'], bins=ranks, labels=labels_age)
```

- Thống kê tần suất xuất hiện của từng user_id.

```
Counter_id = user_id.value_counts()
```

```
Counter_id.describe()
```

```
count    186931.000000
mean          1.107195
std           0.333045
min           1.000000
25%           1.000000
50%           1.000000
75%           1.000000
max           5.000000
Name: count, dtype: float64
```

```
user_id.describe()
```

```
count     206969
unique    186931
top       256449
freq           5
Name: user_id, dtype: object
```

- Chuyển đổi các cột ngày sang định dạng thời gian (datetime).

```
date_time = ['released_date', 'received_date', 'Date_Registered','payment_datetime',
             'purchased_datetime', 'estimated_delivery_date']
```

```
for item in date_time:
    train set[item] = pd.to datetime(train set[item])
```

Bước 7. Lưu dữ liệu đã xử lý

- Dữ liệu sau khi làm sạch được lưu lại dưới dạng file Train_clean.csv.

```
train_set.to_csv('Train_clean.csv', index=False, encoding='utf-8')
```

3.4. Xây dựng mô hình dự đoán

3.4.1. Tiền xử lý dữ liệu

Dữ liệu từ cuộc thi bao gồm nhiều đặc trưng liên quan đến thông tin người dùng, chi tiết giao dịch và chương trình khách hàng thân thiết. Trước khi xây dựng mô hình, dữ liệu cần được xử lý như sau:

- Xử lý giá trị thiếu: Loại bỏ hoặc thay thế các giá trị thiếu bằng các phương pháp như trung bình, trung vị hoặc sử dụng mô hình dự đoán.

```
data.fillna({
    'Received_tier_discount_percentage': 0,
    'Received_card_discount_percentage': 0,
    'Received_coupon_discount_percentage': 0,
    'loyalty_tier': -1
}, inplace=True)
```

- Mã hóa biến phân loại: Sử dụng phương pháp One-Hot Encoding hoặc Label Encoding để chuyển đổi các biến phân loại thành dạng số.

```
from sklearn.preprocessing import OneHotEncoder

def one_hot_encode_and_add(df, column):
    one_hot_encoder = OneHotEncoder(sparse_output=False)
    one_hot_encoded = one_hot_encoder.fit_transform(df[[column]])
    encoded_columns = pd.DataFrame(one_hot_encoded, columns=one_hot_encoder.get_feature_names_out([column]))
    encoded_columns.index = df.index
    df = pd.concat([df, encoded_columns], axis=1)
    df = df.drop(columns=[column])
    return df

columns_to_encode = ['Gender', 'Is_current_loyalty_program_member', 'purchase_medium']

for col in columns_to_encode:
    df = one_hot_encode_and_add(df, col)
```

- Chuẩn hóa dữ liệu: Áp dụng chuẩn hóa Min-Max hoặc Z-score để đưa các đặc trưng về cùng một thang đo, giúp mô hình học hiệu quả hơn.

```
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

3.4.2. Phân chia dữ liệu

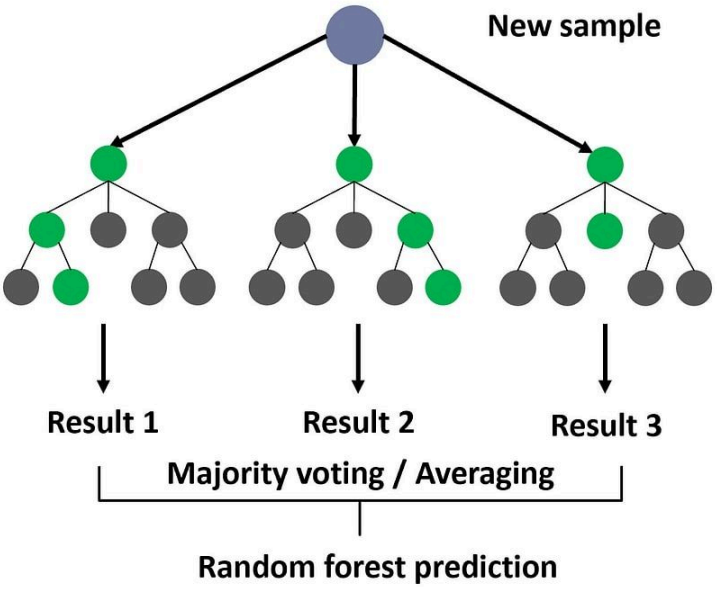
Dữ liệu được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80:20. Ngoài ra, sử dụng phương pháp K-Fold Cross Validation với k=5 để đánh giá hiệu suất mô hình một cách toàn diện.

```
# Create feature set
X = data[selected_features].copy()

# Target variable
y = data['customer_experience']

# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

3.4.3. Lựa chọn mô hình

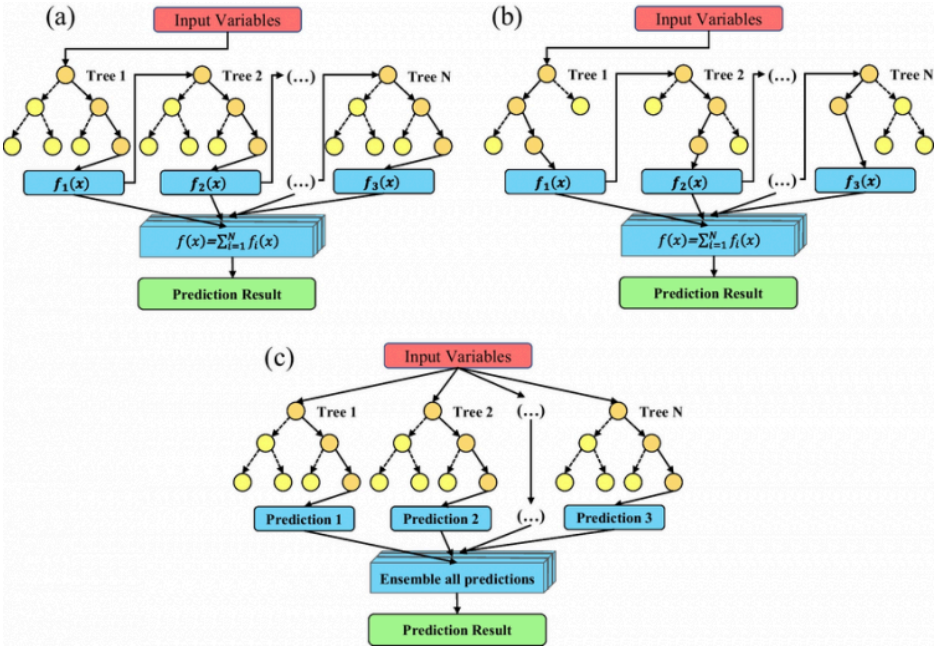
Các mô hình được thử nghiệm bao gồm:

- **Random Forest**



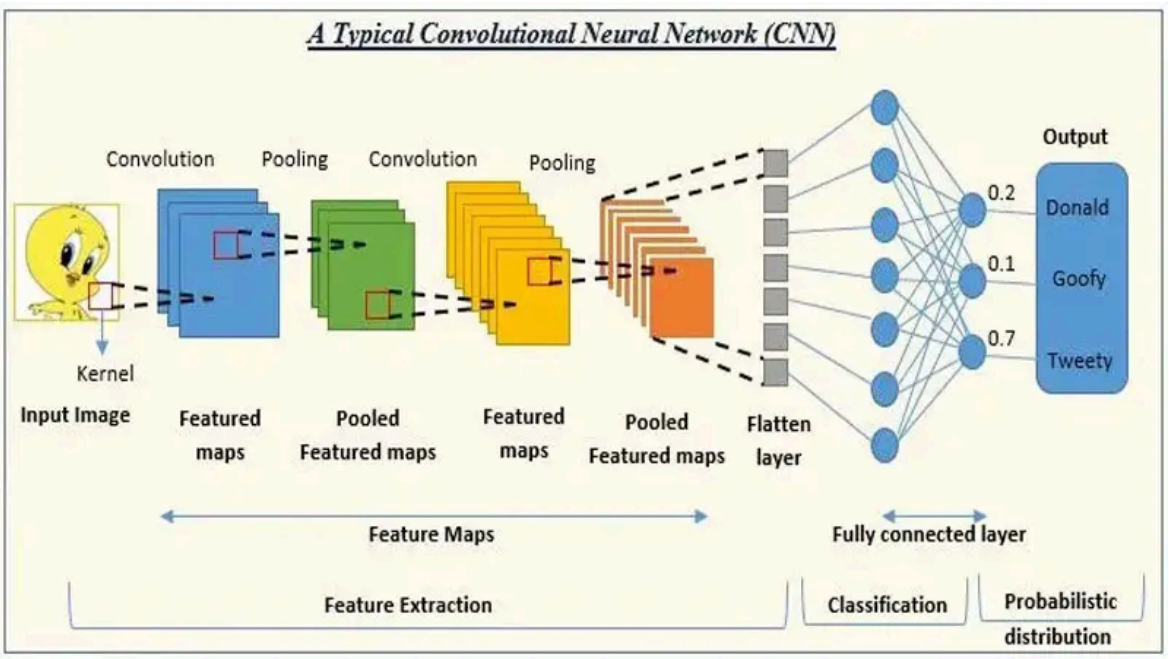
Hình 3.1. Mô hình Random Forest

- **LightGBM**



Hình 3.2. Mô hình LightGBM

- **CNN**



Hình 3.3. Mô hình CNN

- Định nghĩa mô hình LightGBM và Random Forest

```
models = {
    'LightGBM': LGBMClassifier(
        objective='multiclass',
        num_class=3,
        learning_rate=0.1,
        n_estimators=90,
        num_leaves=64,
        feature_fraction=0.9,
        bagging_fraction=0.9,
        lambda_l1=0.1,
        lambda_l2=0.1,
        random_state=42
    ),
    'RandomForest': RandomForestClassifier(
         n_estimators=200,
         max_depth=10,
         min_samples_split=5,
         min_samples_leaf=2,
         random_state=42
    )
}
```

- Định nghĩa mô hình CNN

```
model = keras.Sequential([
    layers.Input(shape=(X_train.shape[1],)),  # Lớp đầu vào
    layers.Reshape((X_train.shape[1], 1)),  # Định hình lại dữ liệu thành dạng 1D
    layers.Conv1D(filters=32, kernel_size=3, activation='relu'),  # Lớp tích chập 1D
    layers.MaxPooling1D(pool_size=2),  # Lớp gộp tối đa
    layers.Dense(64, activation='relu'),  # Lớp ẩn
    layers.Dense(3, activation='softmax')  # Lớp đầu ra với 3 lớp (bad, good, neutral)
])
```

3.4.4. Huấn luyện mô hình

Mỗi mô hình được huấn luyện trên tập huấn luyện và đánh giá trên tập kiểm tra. Sử dụng các kỹ thuật như Grid Search để tìm kiếm siêu tham số tối ưu cho từng mô hình.

- Huấn luyện mô hình LightGBM và Random Forest

```
for model_name, model in models.items():
    evaluate_model(model, X_train, X_test, y_train, y_test, model_name, output_file)
```

- Huấn luyện mô hình CNN

```
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

```
model.fit(X_train, y_train, epochs=10, batch_size=32, validation_data=(X_test, y_test))
```

3.5. Đánh giá mô hình

- Hàm Đánh Giá Hiệu Suất Mô Hình (evaluate_model)

```
def evaluate_model(model, X_train, X_test, y_train, y_test, model_name, output_file):
    with open(output_file, 'a', encoding='utf-8') as f:
        f.write(f"\n========== {model_name} ==========\n")
        model.fit(X_train, y_train)
        y_pred = model.predict(X_test)
        acc = accuracy_score(y_test, y_pred)
        f1 = f1_score(y_test, y_pred, average='weighted')
        f.write(f"Accuracy: {acc:.4f}\n")
        f.write(f"F1 Score: {f1:.4f}\n\n")
        f.write("Classification Report:\n")
        f.write(classification_report(y_test, y_pred))
        f.write("\nConfusion Matrix:\n")
        f.write(str(confusion_matrix(y_test, y_pred)))
        f.write("\n")
```

3.5.1. Mô hình LightGBM

- Tổng quan:
    + Mô hình boosting dựa trên cây quyết định, tối ưu hóa về tốc độ và hiệu quả bộ nhớ.
    + Hiệu quả trên dữ liệu lớn, nhiều đặc trưng, cả liên tục và phân loại.
    + Kết quả tốt nhất trong các mô hình thử nghiệm.

- Kết quả đánh giá:
    + Accuracy: 0.6731
    + F1-Score: 0.6741
    + Macro Avg Precision: 0.68
    + Macro Avg Recall: 0.68
    + Macro Avg F1: 0.68

Lớp	Precision	Recall	F1-Score	Support
0	0.60	0.67	0.63	13,707
1	0.73	0.73	0.73	9,408
2	0.71	0.65	0.68	17,416

Bảng 3.1. Kết quả đánh giá mô hình LightGBM theo từng lớp phân loại

+ Mô hình dự đoán tốt nhất cho lớp 1 (trung tính) với F1-Score là 0.73.
+ Lớp 0 (không hài lòng) thường bị nhầm lẫn với lớp 2 (hài lòng), thể hiện ở 3061 trường hợp.
+ Lớp 2 vẫn được dự đoán tương đối ổn định với precision và recall khá cao.

=> Độ chính xác tổng thể ở mức khá, phản ánh mô hình có thể ứng dụng trong thực tế với các điều chỉnh bổ sung.

3.5.2. Mô hình Random Forest

- Tổng quan:
    + Là một mô hình ensemble đơn giản và phổ biến.

+ Dễ triển khai, tốc độ nhanh nhưng không đạt hiệu suất cao như LightGBM trong bài toán này.

- Kết quả đánh giá:
  + Accuracy: 0.6043
  + F1-Score: 0.6054
  + Macro Avg F1: 0.60

Lớp	Precision	Recall	F1-Score	Support
0	0.57	0.54	0.55	13,707
1	0.52	0.63	0.57	9,408
2	0.69	0.64	0.67	17,416

Bảng 3.2. Kết quả đánh giá mô hình Random Forest theo từng lớp phân loại

- Kết luận:
  + LightGBM là mô hình có hiệu suất tổng thể cao hơn với độ chính xác và F1-Score vượt trội.
  + Mặc dù chưa đạt mức rất cao, nhưng với các kỹ thuật cải tiến như xử lý mất cân bằng lớp, lựa chọn đặc trưng (feature selection), tuning hyperparameter, mô hình này có thể áp dụng tốt trong môi trường thực tế.
  + Random Forest thích hợp cho mục đích baseline hoặc phân tích sơ bộ, nhưng không phải lựa chọn tối ưu cho bài toán này.

Chỉ số	LightGBM	Random Forest
Accuracy	0.6731	0.6043
F1-Score	0.6741	0.6054
F1 lớp 0	0.63	0.55
F1 lớp 1	0.73	0.57
F1 lớp 2	0.68	0.67
Tính ổn định	Tốt	Trung bình
Khả năng tổng quát	Tốt	Cần cải thiện

Bảng 3.3. So sánh kết quả đánh giá giữa mô hình LightGBM và Random Forest

CHƯƠNG 4. Ý NGHĨA THỰC TIỄN CỦA ĐỀ TÀI

Việc dự đoán mức độ hài lòng của khách hàng từ dữ liệu hành vi và giao dịch là một trong những ứng dụng quan trọng nhất của trí tuệ nhân tạo trong lĩnh vực kinh doanh hiện đại. Đề tài "Dự đoán sự hài lòng của khách hàng" không chỉ là một bài toán học thuật mà còn mang ý nghĩa ứng dụng rộng rãi trong thực tế, đặc biệt là trong các lĩnh vực như ngân hàng, bảo hiểm, thương mại điện tử và dịch vụ tiêu dùng.

Trước hết, hệ thống dự đoán mức độ hài lòng giúp các doanh nghiệp hiểu rõ hơn về khách hàng của mình. Trong thời đại cạnh tranh khốc liệt hiện nay, việc giữ chân khách hàng quan trọng không kém gì việc tìm kiếm khách hàng mới. Một khách hàng không hài lòng có thể rời bỏ doanh nghiệp, đồng thời lan truyền những nhận xét tiêu cực, làm giảm uy tín thương hiệu. Thông qua mô hình dự đoán được xây dựng trong đề tài này, doanh nghiệp có thể phát hiện sớm những khách hàng có nguy cơ không hài lòng để từ đó thực hiện các biện pháp can thiệp kịp thời như cung cấp ưu đãi, cải thiện dịch vụ, hoặc cá nhân hóa trải nghiệm nhằm giữ chân khách hàng.

Thứ hai, kết quả từ đề tài góp phần nâng cao hiệu quả trong việc xây dựng các chiến lược tiếp thị và chăm sóc khách hàng. Thay vì sử dụng một chiến dịch quảng bá đồng loạt và tốn kém, doanh nghiệp có thể tập trung nguồn lực vào các nhóm khách hàng đang có xu hướng giảm mức độ hài lòng hoặc có nguy cơ rời bỏ. Việc phân khúc khách hàng dựa trên mức độ hài lòng không chỉ giúp tối ưu chi phí mà còn nâng cao tỷ lệ chuyển đổi và tăng doanh thu dài hạn. Trong thời đại mà dữ liệu lớn (Big Data) và trí tuệ nhân tạo đang đóng vai trò cốt lõi, việc tích hợp mô hình dự đoán như trong đề tài này là bước đi chiến lược giúp các doanh nghiệp bắt kịp xu thế chuyển đổi số.

Thứ ba, mô hình học máy (machine learning) được áp dụng trong đề tài còn có tính khả chuyển cao. Điều này có nghĩa là cùng một mô hình có thể được điều chỉnh để áp dụng cho các lĩnh vực khác như đánh giá sự hài lòng của bệnh nhân trong y tế, học viên trong giáo dục, hay khách du lịch trong ngành lữ hành.

Mỗi khi tổ chức có dữ liệu tương tác từ phía người dùng, mô hình có thể được huấn luyện lại hoặc fine-tune để đưa ra dự đoán chính xác trong bối cảnh mới. Đây là giá trị bền vững, lâu dài của một mô hình học máy có tính tổng quát và mở rộng cao.

Ngoài ra, đề tài còn góp phần thúc đẩy quá trình nghiên cứu và ứng dụng học sâu (deep learning), machine learning vào các bài toán thực tế tại Việt Nam. Trong khi các tập đoàn lớn trên thế giới như Amazon, Google, và Netflix đã triển khai rất mạnh mô hình dự đoán hành vi người dùng để tối ưu trải nghiệm, thì nhiều doanh nghiệp trong nước vẫn đang trong giai đoạn đầu của quá trình chuyển đổi số. Do đó, đề tài này có thể xem như một tài liệu tham khảo, một nền tảng khởi đầu cho các nhóm nghiên cứu, kỹ sư dữ liệu, hoặc đội ngũ IT trong doanh nghiệp Việt ứng dụng hiệu quả các kỹ thuật hiện đại vào bài toán kinh doanh cụ thể.

Cuối cùng, một ý nghĩa xã hội quan trọng khác là mô hình này giúp cải thiện mối quan hệ giữa doanh nghiệp và khách hàng. Việc thấu hiểu khách hàng không chỉ là cơ sở để kinh doanh hiệu quả, mà còn giúp doanh nghiệp xây dựng được sự tin tưởng và gắn bó từ phía người tiêu dùng. Khi khách hàng cảm thấy họ được lắng nghe, được phục vụ tốt hơn nhờ những dự đoán chính xác từ hệ thống, họ sẽ có xu hướng trung thành hơn với thương hiệu, góp phần xây dựng một môi trường tiêu dùng tích cực và bền vững.

CHƯƠNG 5. KHÓ KHĂN DỰ KIẾN VÀ HƯỚNG GIẢI QUYẾT

- Mặc dù LightGBM cho kết quả tốt nhất trong các mô hình đã thử, độ chính xác (accuracy) và F1-Score vẫn còn thấp (khoảng 67%). Điều này đặt ra yêu cầu cần cải thiện thêm về chất lượng dự đoán.
- Một số đặc trưng trong tập dữ liệu có thể chứa nhiễu, lỗi nhập liệu hoặc giá trị ngoại lệ, ảnh hưởng đến khả năng học của mô hình.
- Hướng giải quyết:
    + Tối ưu siêu tham số bằng Grid Search hoặc Bayesian Optimization thay vì sử dụng các giá trị mặc định.
    + Áp dụng kỹ thuật stacking, bagging hoặc boosting đa tầng để kết hợp nhiều mô hình, giúp tăng khả năng tổng quát.
    + Triển khai ensemble học sâu (deep ensemble) hoặc thử nghiệm mạng nơ-ron như MLP (Multi-layer Perceptron) trong các tập Jupyter để so sánh hiệu quả.
    + Áp dụng các kỹ thuật phát hiện và loại bỏ outlier (ví dụ: IQR, Isolation Forest).
    + Sử dụng feature engineering có định hướng: tạo thêm các đặc trưng tổng hợp có ý nghĩa cao, loại bỏ các đặc trưng kém tương quan.
    + Chuẩn hóa và mã hóa dữ liệu đầu vào hợp lý: sử dụng các phương pháp như StandardScaler, MinMaxScaler, hoặc Quantile Transformer.

TÀI LIỆU THAM KHẢO

[1] Azizi, M., & Djouhri, L. Predicting customer satisfaction using machine learning techniques.

[2] Research Team. Predicting customer satisfaction: An approach based on machine learning.

[3] C Research Group. Predicting customer satisfaction for distribution companies using machine learning.

[4] Author(s) Unknown. Predicting E-commerce customer satisfaction: Traditional machine learning versus deep learning.

[5] Lao Research Team. Performance of machine learning models to predict customer satisfaction scores for Lao National Convention Center.

[6] RIT Research Team. Predicting & optimizing airlines customer satisfaction using machine learning techniques.

[7] Can machine learning techniques predict customer dissatisfaction? A feasibility study for the automotive industry.

[8] OPTIMIZING E-COMMERCE PRICING STRATEGIES

[9] Enhancing the Prediction of User Satisfaction with Metaverse Service Through Machine Learning.

[10] Predicting E-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches.

[11] Passenger satisfaction analysis using supervised machine learning model.

[12] Ứng dụng phương pháp máy học trong đo lường sự hài lòng của khách hàng dựa trên các bình luận trực tuyến.

[13] Predictive model for customer satisfaction analytics in E-commerce sector using machine learning and deep learning.