**B****Ộ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: MÔ HÌNH DỰ ĐOÁN DOANH SỐ BÁN HÀNG ĐIỆN TỬ**

**Giáo viên hướng dẫn: TS. Lê Thị Thùy Trang**

**TS. Trần Quý Nam**

**Sinh viên thực hiện:**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Mã sinh viên** | **Họ và tên** | **Lớp** |
| 1 | 1671020061 | Phạm Văn Duy | CNTT 16-03 |
| 2 | 1671020316 | Đỗ Trọng Minh Tiến | CNTT 16-03 |
| 3 | 1671020163 | Nguyễn Ngọc Khánh | CNTT 16-03 |
| 4 | 1671020198 | Đào Đức Mạnh | CNTT 16-03 |

**Hà Nội, năm 2025**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC ĐẠI NAM**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**TÊN HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: MÔ HÌNH DỰ ĐOÁN DOANH SỐ BÁN HÀNG ĐIỆN TỬ**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Mã Sinh Viên | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 | 1671020061 | Phạm Văn Duy | 15/01/2004 |  |  |
| 2 | 1671020316 | Đỗ Trọng Minh Tiến | 25/08/2004 |  |  |
| 3 | 1671020163 | Nguyễn Ngọc Khánh | 27/05/2004 |  |  |
| 4 | 1671020198 | Đào Đức Mạnh | 29/02/2004 |  |  |

### CÁN BỘ CHẤM THI

**Hà Nội, năm 2025**

MỤC LỤC

[**MỤC LỤC HÌNH ẢNH 6**](#_Toc192855407)

[**LỜI NÓI ĐẦU 7**](#_Toc192855408)

[**GIỚI THIỆU 8**](#_Toc192855409)

[1. Bối cảnh và Lý do chọn đề tài 8](#_Toc192855410)

[2. Mục tiêu nghiên cứu 9](#_Toc192855411)

[3. Phạm vi nghiên cứu 9](#_Toc192855412)

[4. Ý nghĩa thực tiễn của đề tài 10](#_Toc192855413)

[**CHƯƠNG I: TỔNG QUAN VỀ THƯƠNG MẠI ĐIỆN TỬ VÀ BÀI TOÁN DỰ ĐOÁN DOANH SỐ BÁN HÀNG 12**](#_Toc192855414)

[1.1. Tổng quan về thương mại điện tử 12](#_Toc192855415)

[*1.1.1. Khái niệm thương mại điện tử*  12](#_Toc192855416)

[*1.1.2. Lịch sử phát triển của thương mại điện tử* 13](#_Toc192855417)

[*1.1.3. Các mô hình thương mại điện tử phổ biến* 13](#_Toc192855418)

[*1.1.4. Lợi ích của thương mại điện tử* 14](#_Toc192855419)

[*1.1.5. Thách thức của thương mại điện tử* 15](#_Toc192855420)

[1.2. Vai trò của dữ liệu lớn trong thương mại điện tử 15](#_Toc192855421)

[*1.2.1. Khái niệm dữ liệu lớn (Big Data)* 15](#_Toc192855422)

[*1.2.2. Ứng dụng của dữ liệu lớn trong thương mại điện tử* 16](#_Toc192855423)

[*1.2.3. Thách thức khi ứng dụng dữ liệu lớn trong thương mại điện tử* 17](#_Toc192855424)

[1.3. Bài toán dự đoán doanh số bán hàng và ứng dụng thực tế 17](#_Toc192855425)

[*1.3.1. Giới thiệu bài toán dự đoán doanh số bán hàng* 17](#_Toc192855426)

[*1.3.2. Các phương pháp dự đoán doanh số bán hàng* 17](#_Toc192855427)

[*1.3.3. Ứng dụng thực tế của dự đoán doanh số bán hàng* 18](#_Toc192855428)

[1.4. Kết luận 19](#_Toc192855429)

[**CHƯƠNG II. CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP NGHIÊN CỨU 20**](#_Toc192855430)

[2.1. Các thuật toán machine learning và deep learning được sử dụng 20](#_Toc192855431)

[*2.1.1. Giới thiệu về Machine Learning và Deep Learning* 20](#_Toc192855432)

[*2.1.2. Các thuật toán Machine Learning được sử dụng* 21](#_Toc192855433)

[*2.1.3. Các thuật toán Deep Learning được sử dụng* 22](#_Toc192855434)

[2.2. Giới thiệu về dữ liệu và phương pháp thu thập dữ liệu 22](#_Toc192855435)

[*2.2.1. Đặc điểm của dữ liệu trong dự đoán doanh số bán hàng* 23](#_Toc192855436)

[*2.2.2. Phương pháp thu thập dữ liệu* 23](#_Toc192855437)

[*2.2.3. Tiền xử lý dữ liệu trước khi đưa vào mô hình* 24](#_Toc192855438)

[2.3. Phương pháp đánh giá mô hình dự đoán 25](#_Toc192855439)

[*2.3.1. Các chỉ số đánh giá hiệu suất mô hình* 26](#_Toc192855440)

[*2.3.2. Phương pháp tối ưu hóa mô hình* 27](#_Toc192855441)

[**CHƯƠNG III: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN DOANH SỐ BÁN HÀNG** 28](#_Toc192855442)

[3.1 Tiền xử lý dữ liệu 28](#_Toc192855443)

[*3.1.1 Chuẩn bị dữ liệu* 28](#_Toc192855444)

[*3.1.2 Xử lý dữ liệu* 29](#_Toc192855445)

[3.2. Xây dựng mô hình và kết nối Spark trên RStudio. 30](#_Toc192855447)

[*3.2.1 Cài Đặt và Tải Thư Viện Cần Thiết* 30](#_Toc192855449)

[*3.2.2 Kết nối đến spark và đọc dữ liệu* 31](#_Toc192855450)

[*3.2.3 Chuyển đổi dữ liệu sang spark dataframe và xử lý dữ liệu thiếu* 31](#_Toc192855451)

[*3.2.4 Lưu trữ dataframe trong bộ nhớ và kiểm tra dữ liệu* 32](#_Toc192855452)

[*3.2.5 Chuyển đổi kiểu dữ liệu và tính toán doanh số* 32](#_Toc192855453)

[*3.2.6 Trích xuất thông tin thời gian và thành phố* 32](#_Toc192855454)

[*3.2.7 Tổng hợp doanh số* 33](#_Toc192855455)

[*3.2.8 Trực quan hóa dữ liệu* 33](#_Toc192855456)

[*3.2.9 Xây dựng mô hình dự đoán doanh số* 33](#_Toc192855457)

[*3.2.10 Mã hóa dữ liệu phân loại* 33](#_Toc192855458)

[*3.2.11 Huấn luyện mô hình cây quyết định* 34](#_Toc192855459)

[*3.2.12 Dự đoán và đánh giá mô hình* 34](#_Toc192855460)

[*3.2.13 Xuất dữ liệu đã xử lý* 34](#_Toc192855461)

[**CHƯƠNG IV: ỨNG DỤNG VÀ ĐỀ XUẤT TRIỂN KHAI 36**](#_Toc192855463)

[4.1 Cách Áp Dụng Mô Hình Vào Hệ Thống Thương Mại Điện Tử 36](#_Toc192855464)

[*4.1.1 Các Bước Triển Khai Mô Hình Vào Thực Tế* 36](#_Toc192855465)

[*4.1.2 Lợi ích của việc áp dụng mô hình dự đoán* 37](#_Toc192855466)

[4.2 Những thách thức và hướng phát triển trong tương lai 38](#_Toc192855467)

[*4.2.1 Các thách thức trong việc triển khai mô hình* 38](#_Toc192855468)

[*4.2.2 Xu hướng công nghệ và phát triển trong lĩnh vực thương mại điện tử* 39](#_Toc192855469)

[**CHƯƠNG V. KẾT QUẢ ĐẠT ĐƯỢC** 41](#_Toc192855470)

[5.1 Biểu đồ thống kế số lượng sản phẩm. 41](#_Toc192855471)

[*5.1.1. Biểu đồ doanh số theo từng tháng trong năm* 41](#_Toc192855472)

[5.2. Mô hình dự đoán doanh số 44](#_Toc192855476)

[5.3. Bảng thống kê dữ liệu 45](#_Toc192855477)

[5.4 Lưu trữ thông tin trên Spark 46](#_Toc192855478)

[**KẾT LUẬN 49**](#_Toc192855480)

[**LỜI CẢM ƠN 52**](#_Toc192855481)

[**TÀI LIỆU THAM KHẢO** 53](#_Toc192855482)

# MỤC LỤC HÌNH ẢNH

[Hình 1.1 Ảnh minh họa về thương mại điện tử 12](#_Toc4293)

[Hình 1.2 Lịch sử hơn 50 năm của thương mại điện tử 13](#_Toc17666)

[Hình 1.3 Các mô hình kinh doanh thương mại điện tử phổ biến hiện nay 14](#_Toc10043)

[Hình 1.4 Minh họa về Bigdata 15](#_Toc14643)

[Hình 1.5 Ứng dụng của dữ liệu lớn trong thương mại điện tử 16](#_Toc30992)

[Hình 1.6 Dự đoán doanh số bán hàng 18](#_Toc14621)

[Hình 2.1 Minh họa về Machine Learning và Deep Learning 20](#_Toc5943)

[Hình 2.2 Ví dụ minh họa về thuật toán decision tree 21](#_Toc2715)

[Hình 2.3 Minh họa về thu thập dữ liệu 24](#_Toc3594)

[Hình 2.4 Tối ưu hóa mô hình 27](#_Toc5854)

Hình 3.1. Minh họa về các dữ liệu được lưu vào csv...................................................29

[Hình 3.2 Hình ảnh các dữ liệu trong file.csv 30](#_Toc32556)

[Hình 5.1 Biểu đồ doanh số theo từng tháng trong năm 41](#_Toc18766)

[Hình 5.2 Biểu đồ doanh số bán hàng theo từng thành phố. 42](#_Toc32630)

[Hình 5.3 Biểu đồ theo giờ trong ngày. 43](#_Toc1605)

[Hình 5.4 Mô hình dự đoán doanh số 45](#_Toc302)

[Hình 5.6 Bảng thông tin lưu trữ dữ liệu trên Spark 47](#_Toc229)

# LỜI NÓI ĐẦU

Trong thời đại công nghệ thông tin phát triển mạnh mẽ như hiện nay, dữ liệu đã trở thành một tài nguyên vô giá, có ảnh hưởng sâu rộng đến mọi mặt của đời sống xã hội, kinh tế, khoa học, y tế và nhiều lĩnh vực khác. Với sự bùng nổ của internet, thiết bị di động, mạng xã hội và các hệ thống thông minh, lượng dữ liệu được tạo ra, thu thập và xử lý mỗi ngày đang tăng trưởng theo cấp số nhân. Mỗi cú nhấp chuột, mỗi giao dịch tài chính, mỗi tương tác trên nền tảng số đều góp phần tạo ra một kho dữ liệu khổng lồ, mở ra vô số cơ hội nhưng cũng đặt ra nhiều thách thức trong việc quản lý, phân tích và khai thác dữ liệu một cách hiệu quả.

Thuật ngữ "Dữ liệu lớn" (Big Data) ra đời để mô tả những tập hợp dữ liệu có quy mô khổng lồ, với tốc độ sinh ra nhanh chóng và có cấu trúc đa dạng đến mức các phương pháp xử lý dữ liệu truyền thống không còn đủ khả năng đáp ứng. Sự phát triển của dữ liệu lớn không chỉ đến từ sự gia tăng về số lượng mà còn ở sự phức tạp của dữ liệu, bao gồm cả dữ liệu có cấu trúc (structured data) và dữ liệu phi cấu trúc (unstructured data) đến từ nhiều nguồn khác nhau như văn bản, hình ảnh, video, cảm biến IoT, dữ liệu từ mạng xã hội, thương mại điện tử hay thậm chí là dữ liệu gen trong sinh học.

Để khai thác hiệu quả tiềm năng của dữ liệu lớn, các doanh nghiệp, tổ chức và cá nhân cần trang bị những kiến thức, kỹ thuật và công cụ hiện đại nhằm thu thập, lưu trữ, quản lý và phân tích dữ liệu một cách tối ưu. Môn học **Dữ liệu Lớn** không chỉ cung cấp những khái niệm nền tảng mà còn giúp người học tiếp cận với những công nghệ tiên tiến như hệ thống lưu trữ phân tán (Distributed Storage), xử lý dữ liệu song song (Parallel Processing), các mô hình học máy (Machine Learning) và trí tuệ nhân tạo (Artificial Intelligence) áp dụng trên dữ liệu lớn. Ngoài ra, môn học cũng sẽ giúp bạn hiểu rõ cách sử dụng các nền tảng phổ biến như Hadoop, Spark, NoSQL databases và các hệ thống điện toán đám mây để xử lý dữ liệu hiệu quả hơn.

Dữ liệu lớn không chỉ đơn thuần là một xu hướng công nghệ, mà nó còn trở thành một lĩnh vực mang tính chiến lược, đóng vai trò quan trọng trong việc hỗ trợ ra quyết định, tối ưu hóa quy trình và tạo ra những giá trị đột phá cho các tổ chức và doanh nghiệp trong kỷ nguyên số.

Như vậy, việc hiểu rõ bản chất, nguyên tắc hoạt động và các ứng dụng thực tiễn của dữ liệu lớn là một trong những kỹ năng quan trọng giúp bạn có thể thích nghi và phát triển trong thế giới hiện đại. Hy vọng rằng, qua môn học này, bạn sẽ có được cái nhìn toàn diện hơn về dữ liệu lớn, từ đó áp dụng hiệu quả vào công việc và nghiên cứu trong tương lai.

**GIỚI THIỆU**

## 1. B**ối cảnh và Lý do chọ**n đề tài

Trong thời đại công nghệ số phát triển mạnh mẽ, thương mại điện tử (E-commerce) đã trở thành một trong những lĩnh vực phát triển nhanh nhất trên toàn cầu. Với sự phát triển của Internet và sự gia tăng của các nền tảng mua sắm trực tuyến như Shopee, Lazada, Tiki, Amazon, Alibaba…, hành vi mua sắm của người tiêu dùng đã có những thay đổi đáng kể. Việc khách hàng có thể dễ dàng tìm kiếm, so sánh và đặt mua sản phẩm chỉ bằng vài cú nhấp chuột đã tạo ra một lượng dữ liệu khổng lồ, cung cấp những thông tin vô cùng giá trị về xu hướng tiêu dùng, nhu cầu thị trường và các yếu tố ảnh hưởng đến doanh số bán hàng.

Trong môi trường cạnh tranh khốc liệt này, các doanh nghiệp thương mại điện tử không chỉ cần tập trung vào việc cải thiện chất lượng sản phẩm, dịch vụ mà còn phải tối ưu hóa chiến lược kinh doanh thông qua việc phân tích dữ liệu và dự đoán xu hướng thị trường. Một trong những bài toán quan trọng mà các doanh nghiệp đang tìm cách giải quyết là **dự đoán doanh số bán hàng**. Việc dự đoán chính xác doanh số có thể giúp doanh nghiệp:

* Quản lý hàng tồn kho hiệu quả, tránh tình trạng thiếu hàng hoặc dư thừa hàng hóa.
* Xác định các chiến lược giá cả phù hợp, tối ưu hóa doanh thu.
* Đưa ra các chương trình khuyến mãi hợp lý nhằm kích thích nhu cầu mua sắm.
* Nâng cao trải nghiệm khách hàng thông qua việc cá nhân hóa sản phẩm, dịch vụ.
* Hỗ trợ ra quyết định chiến lược cho doanh nghiệp dựa trên các xu hướng dữ liệu thực tế.

Dữ liệu lớn (Big Data) đã và đang đóng vai trò quan trọng trong việc phân tích và dự đoán doanh số bán hàng. Nhờ vào sự phát triển của các công nghệ xử lý dữ liệu như **Machine Learning (Học máy), Deep Learning (Học sâu), Data Mining (Khai thác dữ liệu)**, việc xây dựng các mô hình dự đoán doanh số bán hàng trở nên khả thi và mang lại độ chính xác cao. Dựa trên những yếu tố trên, nhóm quyết định chọn đề tài **"Mô hình dự đoán doanh số bán hàng điện tử"** để nghiên cứu và ứng dụng các thuật toán khoa học dữ liệu vào thực tế.

## 2. Mục tiêu nghiên cứu

Mục tiêu chính của đề tài là xây dựng mô hình dự đoán doanh số bán hàng dựa trên dữ liệu lịch sử, từ đó giúp doanh nghiệp thương mại điện tử có thể đưa ra các quyết định kinh doanh chính xác hơn. Cụ thể, nhóm tập trung vào các mục tiêu sau:

* **Thu thập và xử lý dữ liệu bán hàng** từ các nền tảng thương mại điện tử hoặc các bộ dữ liệu công khai.
* **Tiền xử lý dữ liệu** để loại bỏ dữ liệu bị thiếu, nhiễu và chuẩn hóa dữ liệu phục vụ huấn luyện mô hình.
* **Phân tích các yếu tố ảnh hưởng đến doanh số bán hàng**, bao gồm giá cả, xu hướng thị trường, thời gian trong năm, ngày lễ, chương trình khuyến mãi, v.v.
* **Xây dựng và huấn luyện mô hình dự đoán doanh số bán hàng** bằng các thuật toán như hồi quy tuyến tính (Linear Regression), mạng nơ-ron nhân tạo (Neural Network), rừng ngẫu nhiên (Random Forest), XGBoost, LSTM, v.v.
* **Đánh giá và so sánh hiệu suất của các mô hình** bằng các chỉ số như MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) và R² Score.
* **Triển khai mô hình vào thực tế** và đề xuất ứng dụng vào hoạt động kinh doanh của doanh nghiệp thương mại điện tử.

## 3. Phạm vi nghiên cứu

Để đảm bảo tính khả thi và hiệu quả của đề tài, nhóm giới hạn phạm vi nghiên cứu như sau:

**Nguồn dữ liệu**: Nhóm sẽ sử dụng dữ liệu bán hàng của một hoặc một số nền tảng thương mại điện tử (có thể từ các bộ dữ liệu công khai như Kaggle, Google Dataset hoặc từ doanh nghiệp thực tế). Dữ liệu sẽ bao gồm thông tin về sản phẩm, giá bán, số lượng bán, thời gian bán, chương trình khuyến mãi, đánh giá của khách hàng, v.v.

**Phương pháp tiếp cận**: Nhóm sẽ áp dụng các thuật toán Machine Learning và Deep Learning để dự đoán doanh số bán hàng. Một số thuật toán dự kiến sẽ được thử nghiệm bao gồm:

Hồi quy tuyến tính (Linear Regression)

Mô hình rừng ngẫu nhiên (Random Forest)

Gradient Boosting (XGBoost, LightGBM)

Mạng nơ-ron nhân tạo (Neural Network)

LSTM (Long Short-Term Memory) dành cho dữ liệu chuỗi thời gian

**Công cụ và nền tảng**: Nhóm sử dụng các công cụ như **Python, Pandas, NumPy, Scikit-learn, TensorFlow, Keras, Matplotlib, Seaborn** để phân tích và trực quan hóa dữ liệu.

## 

## 4. Ý nghĩa thực tiễn của đề tài

Việc xây dựng mô hình dự đoán doanh số bán hàng không chỉ mang lại lợi ích cho doanh nghiệp thương mại điện tử mà còn đóng góp vào sự phát triển chung của lĩnh vực phân tích dữ liệu lớn. Một số ý nghĩa thực tiễn của đề tài bao gồm:

* **Hỗ trợ doanh nghiệp tối ưu hóa quy trình bán hàng**: Bằng cách dự đoán chính xác doanh số, doanh nghiệp có thể điều chỉnh kế hoạch nhập hàng, kiểm soát tồn kho và đưa ra quyết định kinh doanh phù hợp.
* **Tối ưu hóa chiến lược tiếp thị**: Dữ liệu phân tích giúp doanh nghiệp hiểu rõ hơn về hành vi tiêu dùng, từ đó xây dựng chiến lược marketing hiệu quả hơn.
* **Tăng cường khả năng cạnh tranh**: Các doanh nghiệp có thể tận dụng mô hình dự đoán để nhanh chóng thích ứng với xu hướng thị trường, điều chỉnh giá cả và đưa ra các chương trình khuyến mãi hợp lý.
* **Ứng dụng công nghệ hiện đại vào thực tiễn**: Việc sử dụng Machine Learning trong thương mại điện tử mở ra nhiều cơ hội mới cho các doanh nghiệp muốn áp dụng công nghệ để tối ưu hóa hoạt động kinh doanh.

# 

# CHƯƠNG I: TỔNG QUAN VỀ THƯƠNG MẠI ĐIỆN TỬ VÀ BÀI TOÁN DỰ ĐOÁN DOANH SỐ BÁN HÀNG

## ****1.1. Tổng quan về thương mại điện tử****

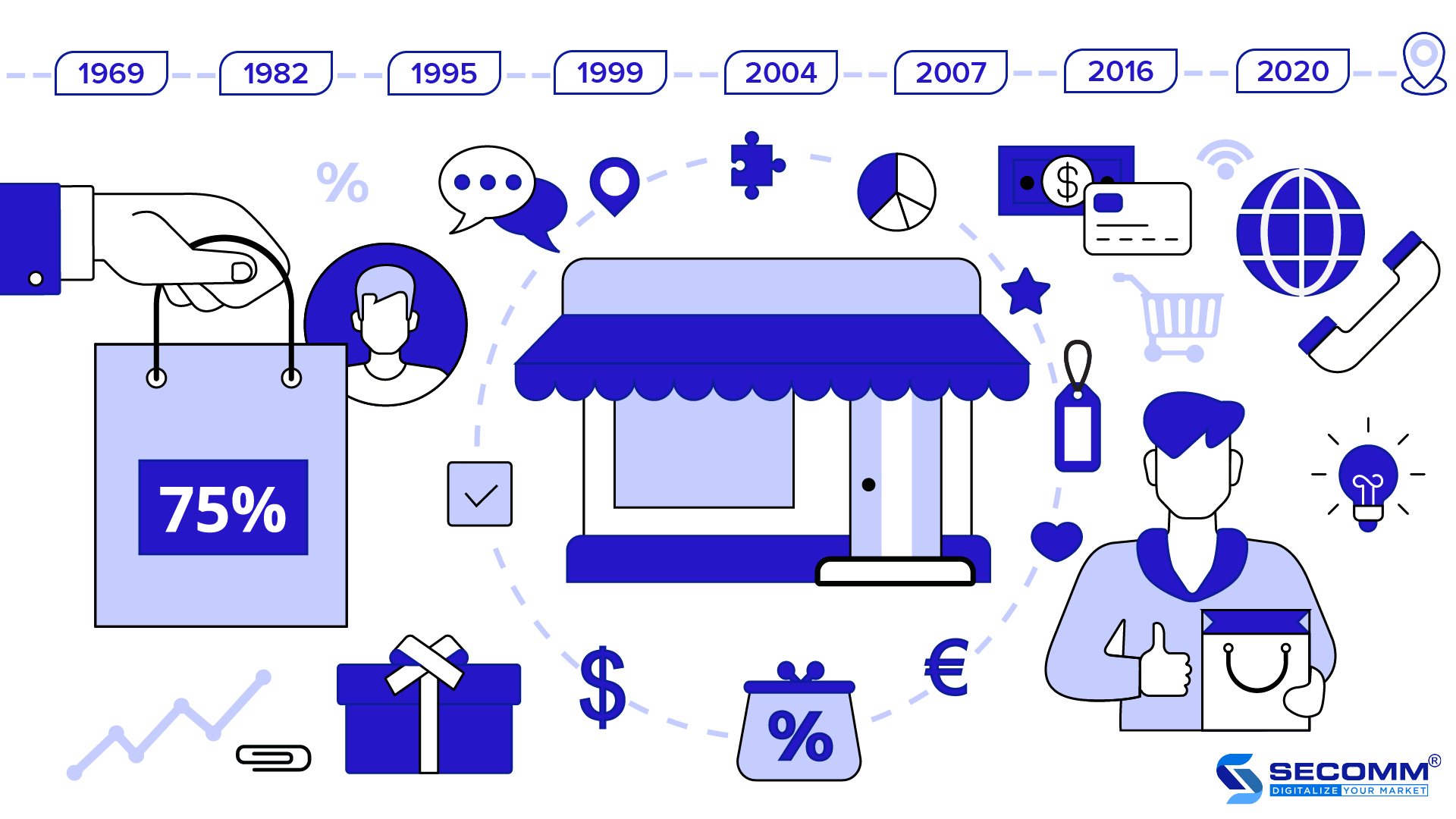
### *****1.1.1. Khái niệm thương mại điện tử***** IMG_256

Hình 1.1 Ảnh minh họa về thương mại điện tử

Thương mại điện tử (E-commerce) là hình thức mua bán hàng hóa và dịch vụ thông qua các nền tảng trực tuyến. Đây là một trong những xu hướng phát triển quan trọng của nền kinh tế hiện đại, giúp kết nối doanh nghiệp với khách hàng một cách nhanh chóng, tiện lợi. Với sự bùng nổ của Internet, điện thoại thông minh và các nền tảng thanh toán số, thương mại điện tử đã trở thành một phần không thể thiếu trong đời sống con người.

### 

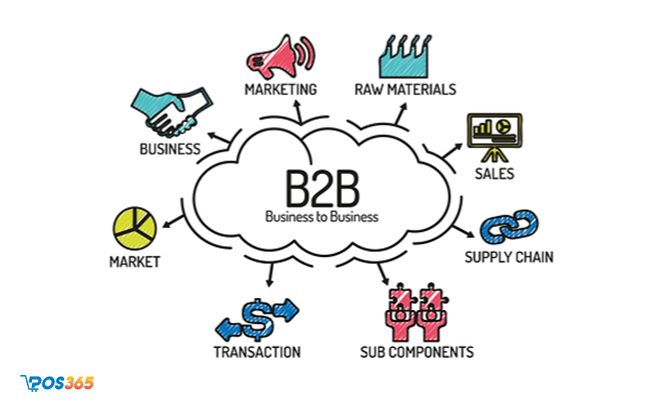
### *****1.1.2. Lịch sử phát triển của thương mại điện tử*****



Hình 1.2 Lịch sử hơn 50 năm của thương mại điện tử

Thương mại điện tử ra đời từ những năm 1960 với sự xuất hiện của hệ thống trao đổi dữ liệu điện tử (EDI), giúp các doanh nghiệp giao dịch với nhau thông qua mạng máy tính. Đến những năm 1990, nhờ sự phát triển mạnh mẽ của Internet, các nền tảng mua sắm trực tuyến như Amazon và eBay ra đời, mở ra một kỷ nguyên mới cho hoạt động thương mại. Từ năm 2010 đến nay, lĩnh vực này tiếp tục phát triển mạnh mẽ với sự ứng dụng của trí tuệ nhân tạo (AI), dữ liệu lớn (Big Data) và mạng xã hội, giúp cá nhân hóa trải nghiệm mua sắm và tối ưu hóa hiệu suất kinh doanh.

### *****1.1.3. Các mô hình thương mại điện tử phổ biến*****



Hình 1.3. Các mô hình kinh doanh thương mại điện tử phổ biến hiện nay

Trong thương mại điện tử, các giao dịch chủ yếu được thực hiện thông qua bốn mô hình chính. Mô hình B2C (Business to Consumer) là mô hình phổ biến nhất, trong đó doanh nghiệp bán hàng trực tiếp cho người tiêu dùng qua các nền tảng như Amazon, Shopee hay Lazada. Mô hình B2B (Business to Business) lại tập trung vào việc trao đổi hàng hóa, dịch vụ giữa các doanh nghiệp, điển hình là Alibaba và các nhà cung cấp phần mềm. Mô hình C2C (Consumer to Consumer) cho phép cá nhân giao dịch với nhau thông qua các sàn thương mại như eBay hay Chợ Tốt. Cuối cùng, mô hình C2B (Consumer to Business) là nơi người tiêu dùng cung cấp dịch vụ hoặc sản phẩm cho doanh nghiệp, thường thấy trong lĩnh vực freelancer hay influencer marketing.

### *****1.1.4. Lợi ích của thương mại điện tử*****

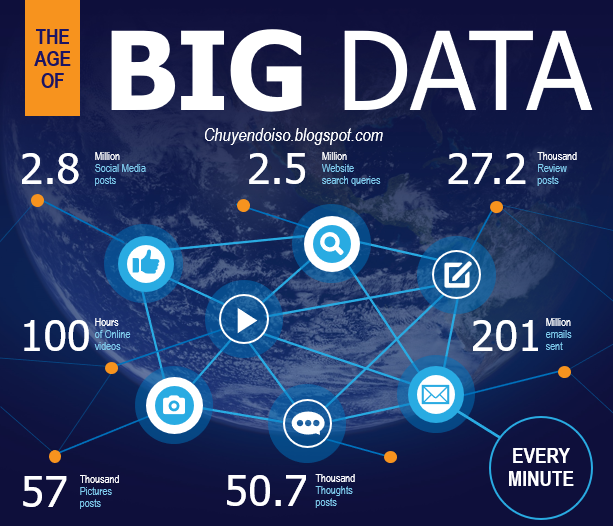
Thương mại điện tử mang lại nhiều lợi ích đáng kể cho cả doanh nghiệp và người tiêu dùng. Người mua có thể dễ dàng tìm kiếm, so sánh sản phẩm và thực hiện giao dịch mọi lúc mọi nơi, không bị giới hạn bởi không gian hay thời gian. Đối với doanh nghiệp, thương mại điện tử giúp tiết kiệm chi phí vận hành, mở rộng thị trường mà không cần đầu tư vào cơ sở hạ tầng vật lý. Đồng thời, nhờ sự hỗ trợ của AI và dữ liệu lớn, các doanh nghiệp có thể phân tích hành vi người tiêu dùng, cá nhân hóa trải nghiệm mua sắm và tối ưu hóa chiến lược tiếp thị.

### *****1.1.5. Thách thức của thương mại điện tử*****

Bên cạnh những lợi ích to lớn, thương mại điện tử cũng đối mặt với không ít thách thức. Sự cạnh tranh khốc liệt buộc các doanh nghiệp phải không ngừng đổi mới để duy trì vị thế trên thị trường. Các vấn đề liên quan đến bảo mật dữ liệu và gian lận trực tuyến cũng là mối quan tâm lớn, đòi hỏi các nền tảng thương mại điện tử phải đầu tư mạnh vào công nghệ bảo mật. Ngoài ra, việc vận chuyển hàng hóa và quản lý đơn hàng cũng đặt ra nhiều khó khăn, đặc biệt là khi tỷ lệ hoàn trả hàng ngày càng tăng.

## ****1.2. Vai trò của dữ liệu lớn trong thương mại điện tử****

### *****1.2.1. Khái niệm dữ liệu lớn (Big Data)*****

Hình 1.4.Minh họa về Bigdata

Dữ liệu lớn (Big Data) là tập hợp dữ liệu có khối lượng khổng lồ, tốc độ tăng trưởng nhanh và có tính đa dạng cao, giúp các doanh nghiệp phân tích, đưa ra quyết định kinh doanh chính xác hơn. Trong thương mại điện tử, dữ liệu lớn đóng vai trò quan trọng trong việc hiểu rõ nhu cầu khách hàng, tối ưu hóa quy trình vận hành và nâng cao trải nghiệm người dùng.

### *****1.2.2. Ứng dụng của dữ liệu lớn trong thương mại điện tử*****



Hình 1.5 Ứng dụng của dữ liệu lớn trong thương mại điện tử

Dữ liệu lớn giúp doanh nghiệp phân tích hành vi khách hàng để cá nhân hóa trải nghiệm mua sắm, từ đó đề xuất sản phẩm phù hợp với từng đối tượng. Nhờ vào khả năng xử lý và phân tích dữ liệu nhanh chóng, doanh nghiệp có thể dự báo xu hướng mua sắm, tối ưu hóa quản lý kho hàng, giảm tình trạng tồn kho hoặc thiếu hụt hàng hóa. Bên cạnh đó, dữ liệu lớn còn hỗ trợ trong việc tối ưu chiến lược tiếp thị, giúp doanh nghiệp quảng cáo đúng đối tượng khách hàng tiềm năng, tiết kiệm chi phí và tăng hiệu quả kinh doanh.

### *****1.2.3. Thách thức khi ứng dụng dữ liệu lớn trong thương mại điện tử*****

Mặc dù mang lại nhiều lợi ích, việc ứng dụng dữ liệu lớn cũng đặt ra nhiều thách thức. Một trong những vấn đề lớn nhất là bảo mật và quyền riêng tư, khi lượng dữ liệu thu thập ngày càng lớn, nguy cơ rò rỉ thông tin cá nhân cũng tăng cao. Bên cạnh đó, việc xử lý và phân tích dữ liệu yêu cầu hệ thống công nghệ tiên tiến, đòi hỏi doanh nghiệp phải đầu tư vào cơ sở hạ tầng và nhân lực có chuyên môn cao.

## ****1.3. Bài toán dự đoán doanh số bán hàng và ứng dụng thực tế****

### *****1.3.1. Giới thiệu bài toán dự đoán doanh số bán hàng*****

Dự đoán doanh số bán hàng là một trong những bài toán quan trọng trong thương mại điện tử, giúp doanh nghiệp ước lượng lượng hàng tiêu thụ trong tương lai dựa trên dữ liệu lịch sử. Việc dự đoán chính xác doanh số giúp doanh nghiệp đưa ra quyết định chiến lược về sản xuất, tồn kho và tiếp thị, giảm thiểu rủi ro kinh doanh và tối ưu hóa lợi nhuận.

### *****1.3.2. Các phương pháp dự đoán doanh số bán hàng*****



Hình 1.6 Dự đoán doanh số bán hàng

Có nhiều phương pháp được sử dụng để dự đoán doanh số bán hàng, từ các phương pháp thống kê truyền thống đến các mô hình học máy hiện đại. Các phương pháp như hồi quy tuyến tính, mô hình chuỗi thời gian ARIMA được sử dụng để phân tích xu hướng dựa trên dữ liệu quá khứ. Trong khi đó, các mô hình học sâu như mạng nơ-ron nhân tạo (ANN) hay mô hình LSTM (Long Short-Term Memory) có khả năng xử lý dữ liệu lớn, nhận diện các mẫu phức tạp và đưa ra dự báo chính xác hơn.

### *****1.3.3. Ứng dụng thực tế của dự đoán doanh số bán hàng*****

Trong thực tế, dự đoán doanh số bán hàng được áp dụng rộng rãi trong nhiều lĩnh vực như bán lẻ, thời trang, thực phẩm và dịch vụ. Các nền tảng thương mại điện tử như Amazon, Shopee hay Tiki sử dụng thuật toán dự đoán doanh số để tối ưu hóa kho hàng, điều chỉnh giá cả và cá nhân hóa đề xuất sản phẩm cho người dùng. Ngoài ra, việc dự báo chính xác doanh số còn giúp doanh nghiệp hoạch định ngân sách, tối ưu hóa chiến dịch quảng cáo và tăng cường hiệu suất kinh doanh.

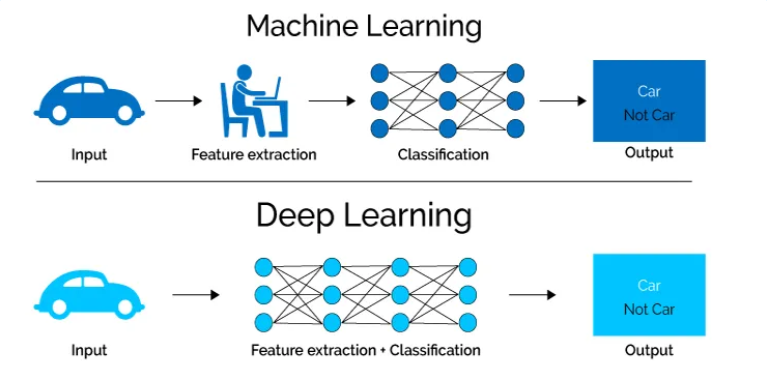
## ****1.4. Kết luận****

Thương mại điện tử đang ngày càng phát triển và dữ liệu lớn đóng vai trò quan trọng trong việc tối ưu hóa hoạt động kinh doanh. Bài toán dự đoán doanh số bán hàng là một trong những ứng dụng quan trọng của dữ liệu lớn, giúp doanh nghiệp đưa ra quyết định kinh doanh hiệu quả hơn. Việc ứng dụng các phương pháp hiện đại trong dự đoán doanh số sẽ giúp doanh nghiệp tăng trưởng bền vững trong bối cảnh thị trường cạnh tranh khốc liệt.

# ****CHƯƠNG II.**** CƠ SỞ LÝ THUYẾT VÀ PHƯƠNG PHÁP NGHIÊN CỨU

## ****2.1. Các thuật toán machine learning và deep learning được sử dụng****

### *****2.1.1. Giới thiệu về Machine Learning và Deep Learning*****

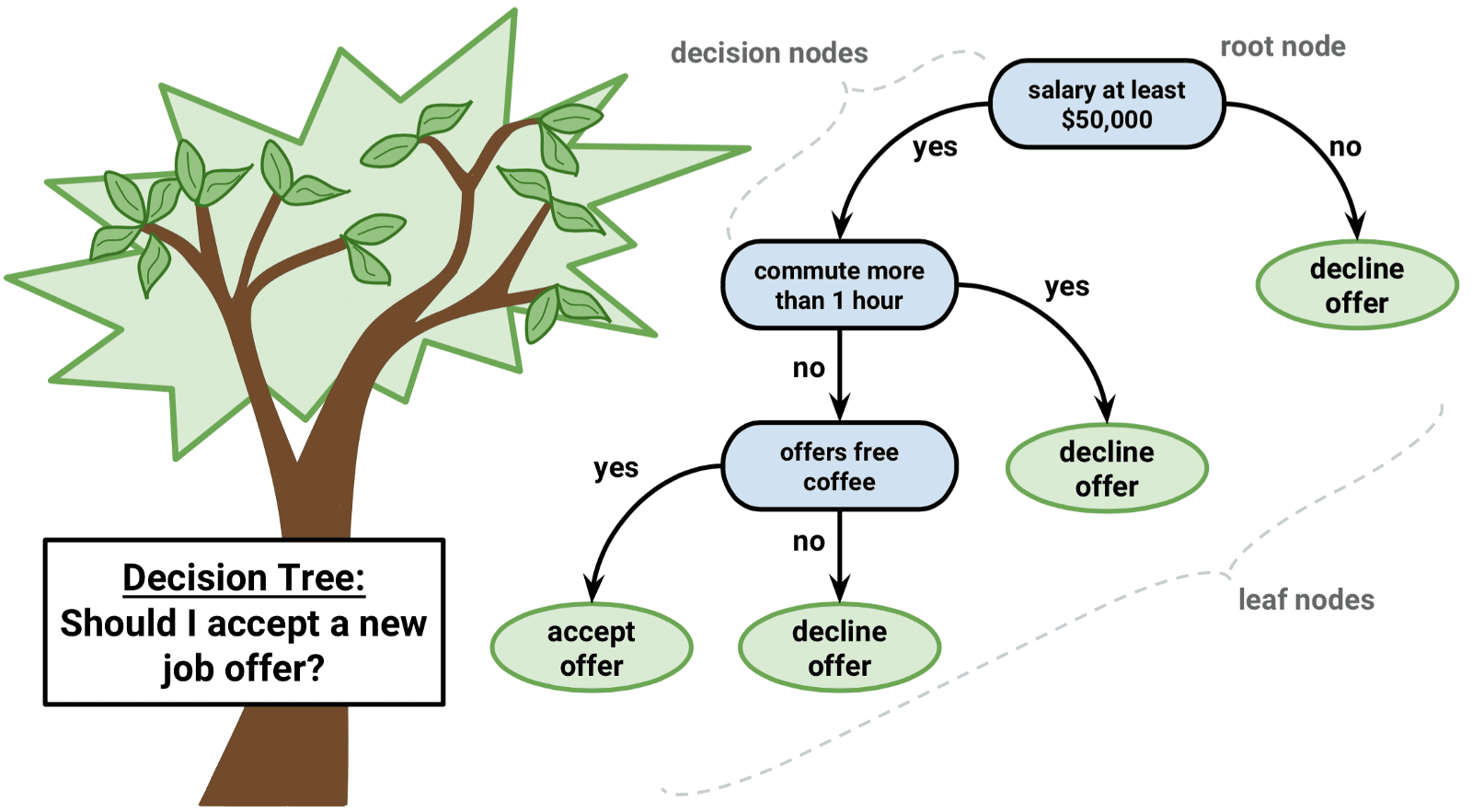


Hình 2.1 Minh họa về Machine Learning và Deep Learning

Trong thời đại công nghệ phát triển mạnh mẽ, trí tuệ nhân tạo (AI) ngày càng khẳng định vai trò quan trọng của mình trong nhiều lĩnh vực, đặc biệt là thương mại điện tử. Trong đó, Machine Learning (học máy) là một nhánh quan trọng của AI, giúp các hệ thống có thể học từ dữ liệu, tự động nhận biết quy luật và đưa ra dự đoán mà không cần lập trình một cách cứng nhắc. Đây chính là nền tảng cốt lõi giúp các doanh nghiệp có thể tối ưu hóa chiến lược kinh doanh, cải thiện hiệu suất và nâng cao trải nghiệm người dùng.

Bên cạnh Machine Learning, Deep Learning (học sâu) được coi là một bước tiến lớn trong lĩnh vực trí tuệ nhân tạo. Deep Learning sử dụng các mạng nơ-ron nhân tạo có nhiều tầng để xử lý thông tin một cách chuyên sâu, cho phép phát hiện những mẫu dữ liệu phức tạp mà các thuật toán truyền thống khó có thể nhận diện được. Trong bài toán dự đoán doanh số bán hàng, cả Machine Learning và Deep Learning đều đóng vai trò quan trọng trong việc phân tích xu hướng tiêu dùng, mô hình hóa hành vi khách hàng cũng như tối ưu hóa chiến lược kinh doanh dựa trên dữ liệu thực tế.

### *****2.1.2. Các thuật toán Machine Learning được sử dụng*****



Hình 2.2 Ví dụ minh họa về thuật toán decision tree

Dự đoán doanh số bán hàng là một bài toán mang tính chất dự báo, trong đó việc lựa chọn thuật toán phù hợp sẽ quyết định đến độ chính xác và hiệu quả của mô hình. Một trong những thuật toán được sử dụng phổ biến trong dự báo doanh số là hồi quy tuyến tính (Linear Regression). Đây là phương pháp cơ bản nhất, trong đó mô hình cố gắng tìm ra mối quan hệ tuyến tính giữa các biến đầu vào và doanh số bán hàng. Bằng cách sử dụng phương trình đường thẳng, thuật toán có thể ước lượng xu hướng tiêu thụ sản phẩm dựa trên các yếu tố như giá cả, mùa vụ hay mức độ quan tâm của khách hàng.

Bên cạnh hồi quy tuyến tính, hồi quy logistic cũng có thể được áp dụng trong một số trường hợp, đặc biệt khi bài toán được đặt dưới dạng phân loại, chẳng hạn như dự đoán doanh số có tăng trưởng hay không trong một giai đoạn cụ thể. Ngoài ra, các mô hình cây quyết định như Random Forest hay Gradient Boosting cũng được sử dụng rộng rãi. Những thuật toán này có khả năng tổng hợp nhiều cây quyết định con để tạo ra một mô hình dự báo mạnh mẽ, giúp giảm thiểu sai số và cải thiện độ chính xác của dự báo.

XGBoost (Extreme Gradient Boosting) cũng là một trong những thuật toán phổ biến trong các bài toán dự báo doanh số nhờ vào hiệu suất cao và khả năng xử lý dữ liệu lớn. Thuật toán này hoạt động dựa trên nguyên tắc học dần dần, liên tục cải thiện độ chính xác của mô hình qua từng vòng lặp, từ đó mang lại kết quả tối ưu hơn so với các phương pháp truyền thống.

### *****2.1.3. Các thuật toán Deep Learning được sử dụng*****

Khi làm việc với các tập dữ liệu lớn và phức tạp, các thuật toán Deep Learning thường mang lại hiệu quả vượt trội. Một trong những mô hình phổ biến nhất trong Deep Learning là mạng nơ-ron nhân tạo (Artificial Neural Network - ANN). ANN mô phỏng cách hoạt động của não bộ con người bằng cách sử dụng các lớp nơ-ron để xử lý dữ liệu đầu vào và tạo ra các dự báo chính xác hơn.

Ngoài ANN, mạng nơ-ron hồi tiếp (Recurrent Neural Network - RNN) cũng thường được sử dụng trong các bài toán chuỗi thời gian, đặc biệt là dự báo doanh số theo từng ngày, từng tháng hoặc từng quý. RNN có khả năng ghi nhớ thông tin từ các bước trước đó, giúp phát hiện các mẫu dữ liệu mang tính chu kỳ, từ đó cải thiện khả năng dự đoán trong tương lai.

Một phiên bản nâng cao hơn của RNN là LSTM (Long Short-Term Memory), giúp khắc phục nhược điểm của RNN truyền thống bằng cách xử lý tốt hơn các chuỗi dữ liệu dài. LSTM đặc biệt hiệu quả trong việc nhận diện xu hướng và mô hình hóa sự biến động của doanh số bán hàng theo thời gian, nhờ đó cung cấp những dự báo chính xác hơn cho doanh nghiệp.

## ****2.2. Giới thiệu về dữ liệu và phương pháp thu thập dữ liệu****

### *****2.2.1. Đặc điểm của dữ liệu trong dự đoán doanh số bán hàng*****

Dữ liệu đóng vai trò quan trọng trong quá trình xây dựng mô hình dự đoán doanh số bán hàng. Để có thể đưa ra những dự báo chính xác, tập dữ liệu cần đảm bảo tính đầy đủ, đa dạng và phản ánh chính xác các yếu tố ảnh hưởng đến doanh số. Trong thương mại điện tử, dữ liệu bán hàng thường bao gồm các thông tin như lịch sử giao dịch, số lượng sản phẩm bán ra, giá bán, chương trình khuyến mãi, phản hồi của khách hàng, thời gian mua sắm, khu vực bán hàng và các yếu tố bên ngoài như mùa vụ hoặc xu hướng thị trường.

Dữ liệu bán hàng thường thuộc nhóm dữ liệu chuỗi thời gian (time series), nghĩa là doanh số bán hàng có thể bị ảnh hưởng bởi các yếu tố thời gian như ngày, tháng, năm hoặc các sự kiện đặc biệt. Do đó, khi thu thập dữ liệu, cần chú ý đến tính liên tục và chu kỳ của dữ liệu để đảm bảo mô hình có thể học được các xu hướng chính xác.

### *****2.2.2. Phương pháp thu thập dữ liệu***** IMG_256

Hình 2.3 Minh họa về thu thập dữ liệu

Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau, tùy thuộc vào phạm vi và mục tiêu của bài toán dự đoán. Một số phương pháp thu thập dữ liệu phổ biến bao gồm:

**Dữ liệu từ hệ thống quản lý bán hàng (POS - Point of Sale):** Đây là nguồn dữ liệu quan trọng, ghi nhận trực tiếp các giao dịch mua bán của khách hàng. Hệ thống POS giúp thu thập thông tin về số lượng sản phẩm bán ra, giá bán, phương thức thanh toán và thời gian giao dịch.

**Dữ liệu từ nền tảng thương mại điện tử:** Các trang web bán hàng trực tuyến như Shopee, Lazada, Tiki, Amazon hoặc các nền tảng thương mại điện tử riêng của doanh nghiệp cung cấp dữ liệu về lượt xem sản phẩm, số lượng đơn hàng, đánh giá của khách hàng và các chương trình khuyến mãi.

**Dữ liệu từ Google Trends và mạng xã hội:** Các công cụ như Google Trends có thể cung cấp thông tin về mức độ quan tâm của khách hàng đối với từng sản phẩm hoặc ngành hàng theo thời gian. Bên cạnh đó, dữ liệu từ Facebook, Instagram, TikTok có thể giúp đánh giá mức độ phổ biến của sản phẩm dựa trên lượt tương tác và phản hồi từ người dùng.

**Dữ liệu từ cảm biến và thiết bị IoT:** Trong một số ngành hàng, doanh số có thể bị ảnh hưởng bởi các yếu tố môi trường như thời tiết, nhiệt độ, độ ẩm. Các cảm biến và thiết bị IoT có thể giúp thu thập dữ liệu về những yếu tố này, từ đó cải thiện mô hình dự báo.

Sau khi thu thập, dữ liệu cần được tổng hợp, làm sạch và tiền xử lý để đảm bảo tính chính xác trước khi đưa vào mô hình huấn luyện.

### *****2.2.3. Tiền xử lý dữ liệu trước khi đưa vào mô hình*****

Dữ liệu thô thường chứa nhiều lỗi như giá trị thiếu, dữ liệu trùng lặp, ngoại lệ hoặc định dạng không đồng nhất. Do đó, quá trình tiền xử lý là một bước quan trọng giúp cải thiện hiệu suất của mô hình dự báo. Một số bước tiền xử lý quan trọng bao gồm:

**Xử lý giá trị thiếu:** Các phương pháp phổ biến để xử lý dữ liệu bị thiếu bao gồm loại bỏ dòng dữ liệu không đầy đủ hoặc sử dụng nội suy (imputation) để thay thế giá trị bị thiếu bằng trung bình, trung vị hoặc mô hình dự đoán.

**Loại bỏ dữ liệu trùng lặp:** Các giao dịch bị ghi nhận nhiều lần hoặc các dòng dữ liệu giống nhau có thể ảnh hưởng đến độ chính xác của mô hình, do đó cần được loại bỏ để đảm bảo dữ liệu không bị sai lệch.

**Phát hiện và xử lý ngoại lệ:** Các điểm dữ liệu bất thường có thể làm sai lệch kết quả dự báo. Các phương pháp như sử dụng Z-score, IQR (Interquartile Range) hoặc thuật toán phân cụm có thể giúp xác định và loại bỏ các giá trị ngoại lệ.

**Chuyển đổi và mã hóa dữ liệu:** Các biến phân loại như tên sản phẩm, danh mục hàng hóa hoặc khu vực bán hàng cần được mã hóa thành dạng số để có thể đưa vào mô hình học máy. Các phương pháp phổ biến bao gồm One-Hot Encoding và Label Encoding.

Sau khi hoàn thành bước tiền xử lý, dữ liệu sẽ được chia thành tập huấn luyện, tập kiểm tra và tập đánh giá để đảm bảo mô hình được đào tạo và kiểm định một cách chính xác.

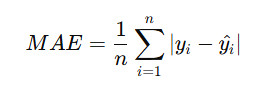
## ****2.3. Phương pháp đánh giá mô hình dự đoán****

Sau khi xây dựng mô hình dự đoán doanh số bán hàng, việc đánh giá hiệu suất của mô hình là một bước quan trọng để kiểm tra độ chính xác và khả năng tổng quát hóa. Các phương pháp đánh giá mô hình giúp xác định mức độ sai số của dự báo so với thực tế, từ đó tối ưu hóa mô hình để đạt kết quả tốt hơn.

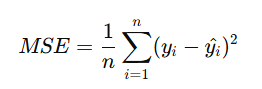
### *****2.3.1. Các chỉ số đánh giá hiệu suất mô hình*****

Có nhiều tiêu chí để đánh giá mô hình dự đoán doanh số, trong đó các chỉ số phổ biến nhất bao gồm:

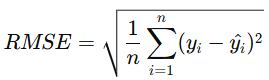
**Mean Absolute Error (MAE - Sai số trung bình tuyệt đối):** Đây là chỉ số đo lường mức độ sai lệch trung bình giữa giá trị thực tế và giá trị dự báo. MAE càng nhỏ, mô hình càng chính xác. Công thức tính MAE như sau:



**Mean Squared Error (MSE - Sai số trung bình bình phương):** Chỉ số này tính toán độ sai số trung bình bằng cách bình phương sai lệch giữa giá trị thực tế và giá trị dự báo. MSE có tác dụng phạt các sai số lớn nhiều hơn so với MAE, giúp phát hiện những dự báo kém chính xác. Công thức tính MSE:

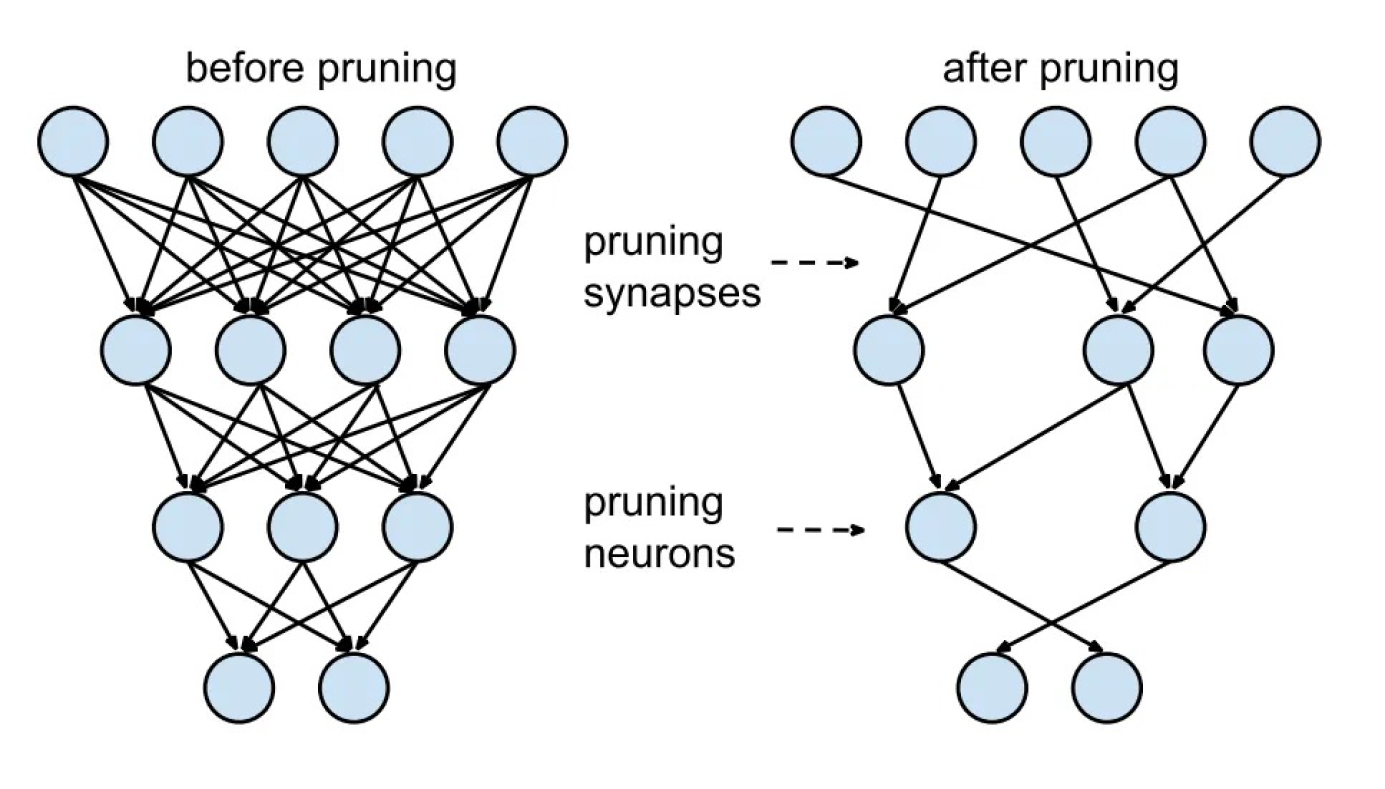


**Root Mean Squared Error (RMSE - Căn bậc hai của sai số trung bình bình phương):** RMSE là một biến thể của MSE, giúp đánh giá sai số theo đơn vị gốc của dữ liệu. Đây là chỉ số hữu ích khi làm việc với dữ liệu có giá trị lớn. Công thức tính RMSE: ​



**R-squared (R² - Hệ số xác định):** Đây là chỉ số đo lường mức độ phù hợp của mô hình với dữ liệu thực tế. R² dao động từ 0 đến 1, trong đó giá trị càng cao cho thấy mô hình càng phù hợp với dữ liệu.

### *****2.3.2. Phương pháp tối ưu hóa mô hình*****



Hình 2.4. Tối ưu hóa mô hình

Sau khi đánh giá, nếu mô hình chưa đạt hiệu suất mong muốn, có thể áp dụng các phương pháp tối ưu hóa như:

**Điều chỉnh siêu tham số (Hyperparameter Tuning):** Tinh chỉnh các tham số như learning rate, số lượng tầng ẩn (trong mạng nơ-ron) hoặc số cây trong Random Forest để tối ưu hiệu suất mô hình.

**Kỹ thuật giảm quá khớp (Regularization):** Sử dụng L1, L2 Regularization hoặc Dropout (trong Deep Learning) để tránh mô hình học quá chi tiết vào dữ liệu huấn luyện mà không tổng quát hóa tốt.

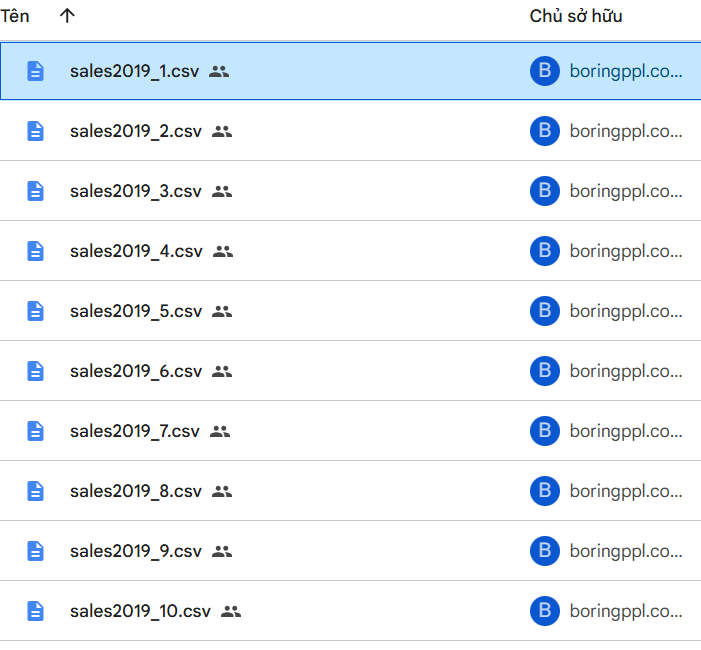
**Sử dụng tập dữ liệu lớn hơn:** Mô hình dự báo doanh số có thể hoạt động tốt hơn nếu có nhiều dữ liệu đầu vào đa dạng hơn.

Nhờ vào việc áp dụng các phương pháp đánh giá và tối ưu hóa, mô hình dự đoán doanh số bán hàng có thể được cải thiện để cung cấp dự báo chính xác, hỗ trợ doanh nghiệp đưa ra quyết định kinh doanh hiệu quả hơn.

**CHƯƠNG III: XÂY DỰNG MÔ HÌNH DỰ ĐOÁN DOANH SỐ BÁN HÀNG**

## 3.1 Tiền xử lý dữ liệu

### *3.1.1 Chuẩn bị dữ liệu*

Lấy dữ liệu : <https://drive.google.com/drive/folders/1Y2rY-n_iO1dJYcnOgaFhzjPYHr_fD_pD>  
Hình 3.1. Minh họa về các dữ liệu được lưu vào csv

Ví dụ về các dữ liệu ở bên trong file.csv  


Hình 3.2 Hình ảnh các dữ liệu trong file.csv

### *3.1.2 Xử lý dữ liệu*

### Từ các dữ liệu của file.csv trên tổng hợp thành 1 file và làm sạch dữ liệu. S**ử dụng Python & Pandas**

**Cách làm**:

python

import pandas as pd

# Đọc file CSV

df = pd.read\_csv("sales2019clean.csv")

# 1. Kiểm tra dữ liệu trống

df = df.dropna() # Xóa dòng trống hoặc df.fillna(value) để thay thế giá trị trống

# 2. Xử lý dữ liệu không hợp lệ

df = df[df["Quantity Ordered"] >= 0] # Loại bỏ giá trị âm

# 3. Chuẩn hóa định dạng ngày tháng

df["Order Date"] = pd.to\_datetime(df["Order Date"], errors='coerce')

# 4. Chuyển đổi kiểu dữ liệu

df["Quantity Ordered"] = pd.to\_numeric(df["Quantity Ordered"], errors='coerce')

# 5. Xóa dữ liệu trùng lặp

df = df.drop\_duplicates()

# 6. Tạo cột mới (Tổng giá trị đơn hàng)

df["Total Price"] = df["Quantity Ordered"] \* df["Price Each"]

# 7. Lưu file sạch

df.to\_csv("sales2019clean.csv", index=False)

## 3.2 Xây dựng mô hình và kết nối Spark trên RStudio.

## Các Bước Xây Dựng Mô Hình Dự Đoán Doanh Số Bán Hàng

### *3.2.1* *****Cài Đặt và Tải Thư Viện Cần Thiết*****

- library(tidyr)

- library(dplyr)

- library(ggplot2)

- library(caret)

- library(rpart)

- Library(rpart.plot)

- library(sparklyr)

- library(Metrics)

* **Mục đích**: Tải các thư viện cần thiết để xử lý và phân tích dữ liệu, cũng như xây dựng mô hình. Mỗi thư viện có chức năng riêng:
  + tidyr và dplyr giúp thao tác với dữ liệu.
  + ggplot2 sử dụng để trực quan hóa dữ liệu.
  + caret hỗ trợ xây dựng và đánh giá mô hình.
  + rpart và rpart.plot được sử dụng cho cây quyết định.
  + sparklyr kết nối với Spark để xử lý dữ liệu lớn.
  + Metrics cung cấp các hàm để đánh giá mô hình.

### *3.2.2 K******ết nối đến spark và đọc dữ liệu*****

sc <- spark\_connect(master = "local")

file\_path <- "sales2019clean.csv"

df\_combined <- read.csv(file\_path, stringsAsFactors = FALSE)

* **Mục đích**: Thiết lập kết nối với Spark và đọc dữ liệu từ file CSV vào DataFrame trong R. Kết nối với Spark cho phép xử lý dữ liệu lớn một cách hiệu quả, trong khi việc đọc dữ liệu từ CSV là bước đầu tiên để phân tích.

### *3.2.3 C******huyển đổi dữ liệu sang spark dataframe và xử lý dữ liệu thiếu*****

df\_spark <- copy\_to(sc, df\_combined, "df\_combined", overwrite = TRUE)

df\_spark <- df\_spark %>% filter(!is.na('Quantity.Ordered') & !is.na('Price.Each'))

* **Mục đích**: Sao chép dữ liệu sang Spark để tận dụng khả năng xử lý song song. Sau đó, lọc bỏ các hàng có giá trị thiếu trong các cột quan trọng (Quantity.Ordered và Price.Each), vì dữ liệu thiếu có thể dẫn đến sai lệch trong mô hình.

### *3.2.4 L******ưu trữ dataframe trong bộ nhớ và kiểm tra dữ liệu*****

df\_spark <- df\_spark %>% sparklyr::sdf\_persist()

df\_spark %>% summarize(count = n()) %>% collect()

* **Mục đích**: Lưu trữ DataFrame trong bộ nhớ để tăng hiệu suất. Sau đó, kiểm tra tổng số hàng để xác minh rằng dữ liệu đã được nạp thành công và không bị mất.

### *3.2.5.* *****Chuyển đổi kiểu dữ liệu và tính toán doanh số*****

df\_combined <- df\_combined %>%

filter(!is.na(Quantity.Ordered) & !is.na(Price.Each)) %>%

mutate(

Quantity.Ordered = as.integer(Quantity.Ordered),

Price.Each = as.numeric(Price.Each),

Sales = Quantity.Ordered \* Price.Each

)

* **Mục đích**: Chuyển đổi kiểu dữ liệu của các cột để đảm bảo chúng phù hợp với mô hình. Tính toán doanh số (Sales) bằng cách nhân số lượng đã đặt (Quantity.Ordered) với giá mỗi sản phẩm (Price.Each). Doanh số là biến mục tiêu cho mô hình dự đoán.

### *3.2.6.* *****Trích xuất thông tin thời gian và thành phố*****

df\_combined$Order.Date <- as.POSIXct(df\_combined$Order.Date, format="%m/%d/%Y %H:%M")

df\_combined$Month <- format(df\_combined$Order.Date, "%m")

df\_combined$Hours <- format(df\_combined$Order.Date, "%H")

df\_combined$City <- sapply(strsplit(df\_combined$Purchase.Address, ","),

function(x) ifelse(length(x) >= 2, trimws(x[2]), NA))

* **Mục đích**: Trích xuất tháng và giờ từ ngày đặt hàng, tạo ra các biến mới để sử dụng trong mô hình. Thông tin về thành phố cũng được trích xuất từ địa chỉ mua hàng, điều này cho phép phân tích theo địa lý.

### *3.2.7.* *****Tổng hợp doanh số*****

sales\_value\_month <- aggregate(Sales ~ Month, data = df\_combined, sum)

sales\_value\_city <- aggregate(Sales ~ City, data = df\_combined, sum)

sales\_value\_hours <- aggregate(Sales ~ Hours, data = df\_combined, sum)

* **Mục đích**: Tổng hợp doanh số theo tháng, thành phố và giờ. Việc tổng hợp này giúp hiểu rõ hơn về xu hướng và mẫu trong doanh số bán hàng.

### *3.2.8.* *****Trực quan hóa dữ liệu*****

par(mfrow=c(2, 2)) # Thiết lập bố cục

# Vẽ các biểu đồ doanh số theo tháng, thành phố, giờ và số lượng sản phẩm

* **Mục đích**: Trực quan hóa dữ liệu để dễ dàng nhận diện xu hướng và mẫu. Biểu đồ giúp người dùng dễ dàng hiểu hơn về dữ liệu.

### *3.2.9.* *****Xây dựng mô hình dự đoán doanh số*****

df\_combined <- df\_combined %>% filter(!is.na(Quantity.Ordered))

set.seed(42) # Đặt hạt giống

trainIndex <- createDataPartition(df\_combined$Quantity.Ordered, p = 0.7, list = FALSE)

train\_data <- df\_combined[trainIndex, ] # Tập huấn luyện

test\_data <- df\_combined[-trainIndex, ] # Tập kiểm tra

* **Mục đích**: Lọc dữ liệu để loại bỏ các hàng thiếu và chia dữ liệu thành tập huấn luyện (70%) và kiểm tra (30%). Việc đặt hạt giống giúp tái tạo kết quả trong các lần chạy khác nhau.

### *3.2.10.* *****Mã hóa dữ liệu phân loại*****

train\_data <- train\_data %>% mutate(across(where(is.character), as.factor))

test\_data <- test\_data %>% mutate(across(where(is.character), as.factor))

* **Mục đích**: Chuyển đổi các biến phân loại thành kiểu factor, giúp môhình dễ dàng xử lý.

### *3.2.11.* *****Huấn luyện mô hình cây quyết định*****

model <- rpart(Quantity.Ordered ~ Month + Hours + City + Price.Each,

data = train\_data, method = "anova")

* **Mục đích**: Huấn luyện mô hình cây quyết định để dự đoán số lượng đã đặt dựa trên các biến đầu vào (tháng, giờ, thành phố và giá). Cây quyết định là một trong những phương pháp phổ biến cho hồi quy.

### *3.2.12.* *****Dự đoán và đánh giá mô hình*****

y\_pred <- predict(model, test\_data) # Dự đoán trên tập test

accuracy <- cor(y\_pred, test\_data$Quantity.Ordered) # Tính toán hệ số tương quan

rmse\_value <- rmse(test\_data$Quantity.Ordered, y\_pred) # Tính RMSE

* **Mục đích**: Dự đoán số lượng đã đặt trên tập kiểm tra và tính toán độ chính xác bằng hệ số tương quan. RMSE được tính để đánh giá độ chính xác của mô hình, với giá trị thấp hơn cho thấy mô hình dự đoán tốt hơn.

### *3.2.13.* *****Xuất dữ liệu đã xử lý*****

write.csv(df\_combined, “sales2019final\_clean.csv”, row.names = FALSE)

* **Mục đích**: Xuất dữ liệu đã xử lý và các biến mới ra file CSV để sử dụng cho các phân tích trong tương lai hoặc để lưu trữ.

## \*Kết Luận

Quá trình xây dựng mô hình dự đoán doanh số bán hàng bao gồm nhiều bước quan trọng, từ việc chuẩn bị dữ liệu, xử lý dữ liệu thiếu, trích xuất thông tin, cho đến việc huấn luyện và đánh giá mô hình. Mỗi bước đều có vai trò quan trọng trong việc đảm bảo rằng mô hình cuối cùng có thể dự đoán chính xác doanh số bán hàng dựa trên các yếu tố như thời gian và vị trí địa lý. Việc trực quan hóa dữ liệu cũng giúp cho người phân tích có cái nhìn sâu sắc hơn về xu hướng và mẫu trong dữ liệu.

# CHƯƠNG IV: ỨNG DỤNG VÀ ĐỀ XUẤT TRIỂN KHAI

## 4.1. Cách Áp Dụng Mô Hình Vào Hệ Thống Thương Mại Điện Tử

### *4.1.1. Các Bước Triển Khai Mô Hình Vào Thực Tế*

**Tích Hợp Mô Hình Vào Hệ Thống:**

Để đưa mô hình dự đoán vào sử dụng thực tế trong hệ thống thương mại điện tử, cần thực hiện một quy trình tích hợp rõ ràng và hiệu quả. Một trong những phương pháp phổ biến là phát triển API (Application Programming Interface) hoặc giao diện người dùng.

**Phát triển API:**

* 1. **Xác định các điểm cuối (endpoints):** Các điểm cuối là các địa chỉ mà ứng dụng khác có thể gửi dữ liệu đến. Ví dụ, một điểm cuối có thể nhận thông tin về đơn hàng và trả về dự đoán về doanh số.
  2. **Chọn công nghệ:** Quyết định sử dụng ngôn ngữ lập trình và framework nào cho việc phát triển API, chẳng hạn như Flask cho Python hoặc Express cho Node.js.
  3. **Bảo mật API:** Thiết lập các biện pháp bảo mật như xác thực và phân quyền để đảm bảo chỉ những người dùng hợp lệ mới có thể truy cập vào mô hình.
  4. **Kiểm tra API:** Thực hiện kiểm tra để đảm bảo rằng API hoạt động đúng và có thể xử lý các lỗi có thể xảy ra.

**Xây dựng giao diện người dùng:**

* 1. **Thiết kế giao diện:** Tạo giao diện thân thiện với người dùng để các bộ phận trong doanh nghiệp có thể dễ dàng tương tác với mô hình. Ví dụ, nhân viên bán hàng có thể nhập dữ liệu và xem kết quả dự đoán một cách trực quan.
  2. **Tích hợp với hệ thống hiện tại:** Đảm bảo giao diện tương thích và tích hợp tốt với các phần mềm mà doanh nghiệp đang sử dụng như CRM hay ERP.

**Theo dõi và cập nhật mô hình:**

Sau khi mô hình đã được tích hợp, việc theo dõi và cập nhật là rất quan trọng để duy trì hiệu suất.

**Thiết lập hệ thống giám sát:**

* 1. **Theo dõi hiệu suất mô hình:** Sử dụng các chỉ số như RMSE, MAE và R² để liên tục đánh giá hiệu suất của mô hình. Hệ thống nên cảnh báo khi các chỉ số này vượt quá ngưỡng cho phép.
  2. **Ghi log dữ liệu:** Lưu trữ dữ liệu đầu vào và kết quả đầu ra để phân tích sau này.

**Cập nhật mô hình:**

* 1. **Định kỳ cập nhật dữ liệu:** Mô hình cần được cập nhật với dữ liệu mới để cải thiện độ chính xác. Điều này có thể thực hiện hàng tháng hoặc hàng quý.
  2. **Điều chỉnh mô hình:** Sử dụng các kỹ thuật như retraining để điều chỉnh mô hình sao cho phù hợp với dữ liệu mới.

### *4.1.2. Lợi ích của việc áp dụng mô hình dự đoán*

**Tăng cường khả năng ra quyết định dựa trên dữ liệu:**

Việc áp dụng mô hình dự đoán không chỉ giúp doanh nghiệp đưa ra quyết định thông minh hơn mà còn giúp giảm thiểu rủi ro. Khi doanh nghiệp có thể dự đoán chính xác doanh số bán hàng, họ có thể điều chỉnh chiến lược tiếp thị, giá cả và khuyến mãi một cách hợp lý.

* **Phân tích dữ liệu lớn:** Mô hình giúp doanh nghiệp khai thác dữ liệu lớn để tìm ra xu hướng và mẫu hành vi khách hàng, từ đó tối ưu hóa chiến lược kinh doanh.
* **Quyết định kịp thời:** Dựa trên các dự đoán, doanh nghiệp có thể đưa ra quyết định nhanh chóng, phù hợp với tình hình thực tế.

**Tối ưu hóa quản lý tồn kho:**

Một trong những lợi ích lớn nhất khi áp dụng mô hình dự đoán là khả năng tối ưu hóa quản lý tồn kho.

* **Giảm thiểu tình trạng hết hàng:** Bằng cách dự đoán chính xác nhu cầu, doanh nghiệp có thể đảm bảo luôn có đủ hàng hóa để phục vụ khách hàng.
* **Giảm thiểu tồn kho dư thừa:** Mô hình cũng giúp doanh nghiệp tránh việc tích trữ hàng hóa không cần thiết, từ đó giảm chi phí lưu kho và tối ưu hóa nguồn lực.
* **Cải thiện quy trình đặt hàng:** Doanh nghiệp có thể điều chỉnh quy trình đặt hàng dựa trên dự đoán, giúp tiết kiệm thời gian và chi phí.

## 4.2. Những thách thức và hướng phát triển trong tương lai

### *4.2.1. Các thách thức trong việc triển khai mô hình*

**Độ Chính Xác:**

Một trong những thách thức lớn nhất khi triển khai mô hình là đảm bảo rằng nó luôn chính xác. Mô hình cần được cập nhật thường xuyên với dữ liệu mới và điều chỉnh để phản ánh những thay đổi trong thị trường và hành vi khách hàng.

* **Phân tích độ chính xác:** Doanh nghiệp cần thường xuyên xem xét các chỉ số hiệu suất để xác định hiệu quả của mô hình.
* **Kiểm định mô hình:** Sử dụng các phương pháp kiểm định như cross-validation để đảm bảo mô hình có thể tổng quát và không chỉ phù hợp với dữ liệu huấn luyện.

**Thay Đổi Hành Vi Khách Hàng:**

Hành vi của khách hàng có thể thay đổi theo thời gian do nhiều yếu tố như xu hướng, mùa vụ hoặc sự kiện đặc biệt. Do đó, việc theo dõi và điều chỉnh mô hình theo những thay đổi này là rất cần thiết để duy trì hiệu suất cao.

* **Phân tích thị trường:** Doanh nghiệp cần thường xuyên theo dõi xu hướng thị trường để kịp thời điều chỉnh mô hình dự đoán.
* **Thu thập phản hồi:** Nhận phản hồi từ khách hàng có thể giúp điều chỉnh mô hình sao cho phù hợp hơn với nhu cầu thực tế.

### *4.2.2. Xu hướng công nghệ và phát triển trong lĩnh vực thương mại điện tử*

**Sự phát triển của AI và Machine Learning:**

Trong tương lai, AI và Machine Learning sẽ tiếp tục phát triển mạnh mẽ và đóng vai trò quan trọng trong thương mại điện tử.

* **Cá nhân hóa trải nghiệm khách hàng:** Sử dụng dữ liệu để tạo ra trải nghiệm mua sắm cá nhân hóa cho từng khách hàng. Các thuật toán học máy có thể phân tích hành vi trước đó của khách hàng để đề xuất sản phẩm phù hợp.
* **Dự đoán xu hướng thị trường:** Phân tích dữ liệu lớn để dự đoán xu hướng thị trường và nhu cầu của khách hàng. Điều này không chỉ giúp doanh nghiệp chuẩn bị tốt hơn mà còn tạo ra cơ hội cạnh tranh lớn hơn.
* **Tự động hóa quy trình kinh doanh:** Sử dụng AI để tự động hóa từ quản lý tồn kho đến chăm sóc khách hàng. Ví dụ, chatbot có thể giúp trả lời câu hỏi của khách hàng 24/7, giảm thiểu gánh nặng cho nhân viên.

**Công Nghệ Mới Nhất:**

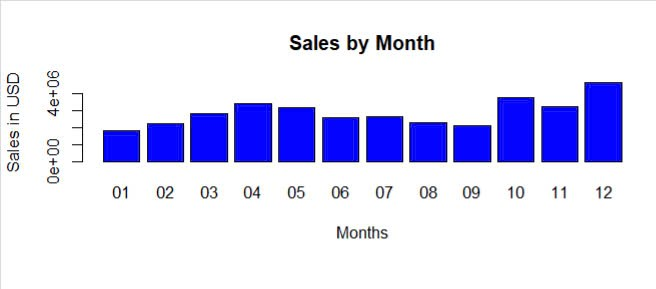
* **Blockchain:** Công nghệ blockchain có thể tăng cường tính minh bạch và an toàn trong giao dịch thương mại điện tử, giúp xây dựng niềm tin với khách hàng.
* **IoT (Internet of Things):** Các thiết bị kết nối Internet có thể cung cấp dữ liệu thời gian thực về hành vi của khách hàng, từ đó cải thiện khả năng dự đoán.
* **Phân tích dữ liệu lớn:** Các công cụ phân tích dữ liệu lớn sẽ giúp doanh nghiệp hiểu rõ hơn về khách hàng và thị trường, từ đó tối ưu hóa chiến lược kinh doanh.

Những xu hướng này không chỉ giúp doanh nghiệp tối ưu hóa hoạt động mà còn nâng cao trải nghiệm của khách hàng, tạo ra lợi thế cạnh tranh bền vững và mở ra nhiều cơ hội mới trong lĩnh vực thương mại điện tử.

**CHƯƠNG V. KẾT QUẢ ĐẠT ĐƯỢC**

**5.1 Biểu đồ thống kế số lượng sản phẩm.**

***5.1.1. Biểu đồ doanh số theo từng tháng trong năm***



Hình 5.1. Biểu đồ doanh số theo từng tháng trong năm

Thành phần chính:

Trục X: Các tháng từ tháng 1 đến tháng 12

Trục Y: Doanh số bán hàng tính bằng USD.

Diễn giải:

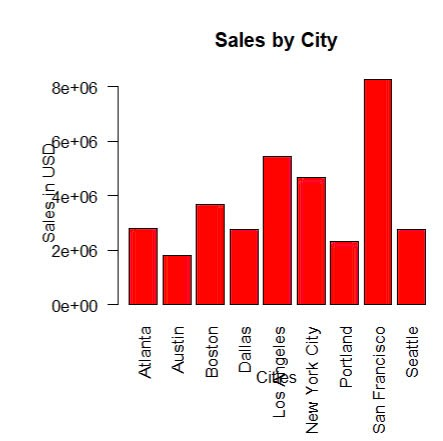
Cột màu xanh cho thấy mức doanh thu của từng tháng.

Tháng 11 và 12 thường có doanh số cao, phản ánh các dịp lễ hội.

Các tháng như tháng 01 có doanh số thấp hơn, có thể do hiệu ứng mùa vụ.

Ý nghĩa:

### Biểu đồ giúp xác định tháng nào có doanh số tốt hoặc kém, từ đó điều chỉnh chiến lược kinh doanh cho phù hợp. *5.1.2. Biểu đồ doanh số bán hàng theo từng thành phố.*

  
*Hình 5.2. Biểu đồ doanh số bán hàng theo từng thành phố.*

Thành phần chính:

Trục X: Các thành phố bao gồm Atlanta, Austin, Boston, Dallas, Los Angeles, New York City, Portland, San Francisco và Seattle.

Trục Y: Doanh số bán hàng tính bằng USD.

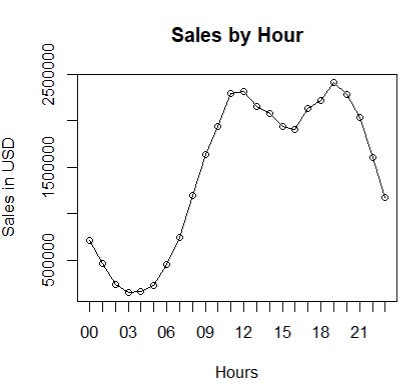
Diễn giải:

Cột màu đỏ thể hiện doanh thu của từng thành phố.

San Francisco và Portland có doanh số cao nhất, trong khi Atlanta và Austin có doanh số thấp hơn.

Ý nghĩa:

### Biểu đồ giúp xác định thành phố nào đóng góp nhiều nhất vào doanh thu, từ đó hỗ trợ quyết định marketing và phân phối. *5.1.3. Biểu đồ theo giờ trong ngày.*



Hình 5.3. Biểu đồ theo giờ trong ngày.

Thành phần chính:

Trục X: Các giờ trong ngày, từ 00 đến 23.

Trục Y: Doanh số bán hàng tính bằng USD.

Diễn giải:

Đường biểu diễn cho thấy sự thay đổi doanh số theo từng giờ.

Doanh số cao nhất diễn ra trong khoảng thời gian từ 12h đến 15h, với mức trên 2,500,000 USD, và thấp nhất ở các khung giờ 03h.

Ý nghĩa:

### Biểu đồ giúp xác định giờ cao điểm bán hàng, hỗ trợ trong việc lên kế hoạch cho marketing và quản lý nguồn lực. *5.1.4. Biểu đồ hiển thị mức độ nhu cầu đối với sản phẩm.*

Thành phần chính:

Trục X: Các sản phẩm khác nhau, bao gồm màn hình, tai nghe, TV và máy tính xách tay.

Trục Y: Số lượng đặt hàng (Quantity Ordered) tính bằng số lượng.

Diễn giải:

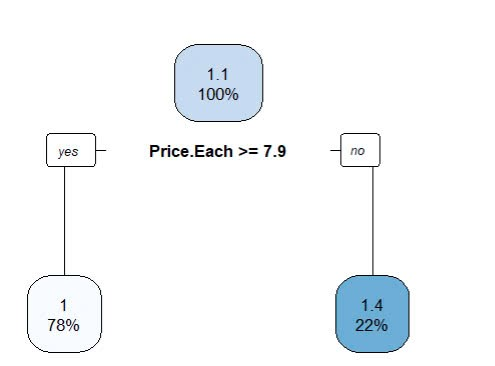
Các cột màu xanh cho thấy số lượng sản phẩm được đặt hàng.

Tai nghe (headphones) có số lượng đặt hàng cao nhất, với các cột khác có nhu cầu thấp hơn và có sự biến động.

Ý nghĩa:

Biểu đồ giúp xác định sản phẩm nào có nhu cầu cao nhất, hỗ trợ trong việc quản lý hàng tồn kho và lập kế hoạch sản xuất.

## 5.2. Mô hình dự đoán doanh số



Hình 5.4. Mô hình dự đoán doanh số

Biểu đồ quyết định này mô tả cách phân loại giá mỗi sản phẩm:

Sử dụng mô hình cây quyết định (decion trees)

Thành phần chính:

Điều kiện phân loại: Nếu giá mỗi sản phẩm (Price Each) lớn hơn hoặc bằng 7.9.

Hai nhánh:

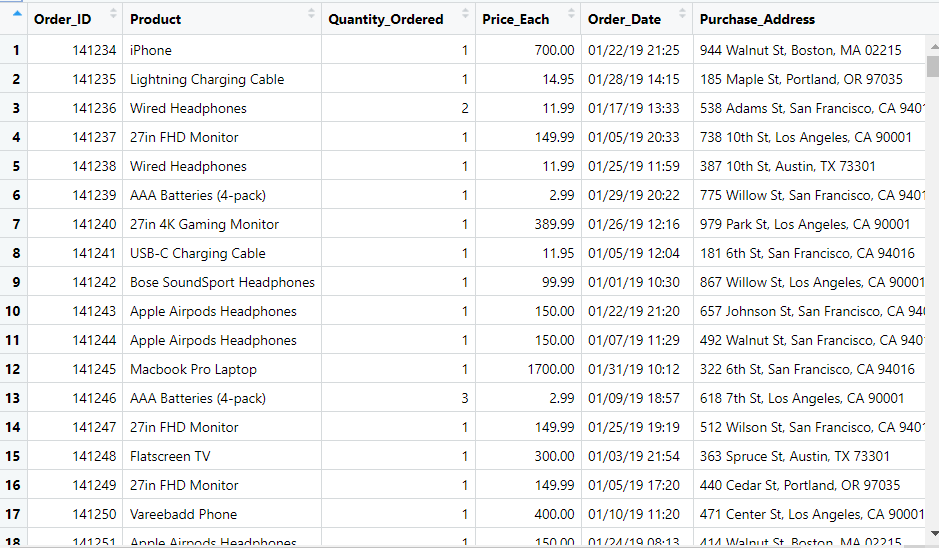
Nhánh "Yes": Nếu điều kiện đúng (78% tổng số), sản phẩm có giá 1.1.

Nhánh "No": Nếu điều kiện sai (22% tổng số), sản phẩm có giá 1.4.

Ý nghĩa:

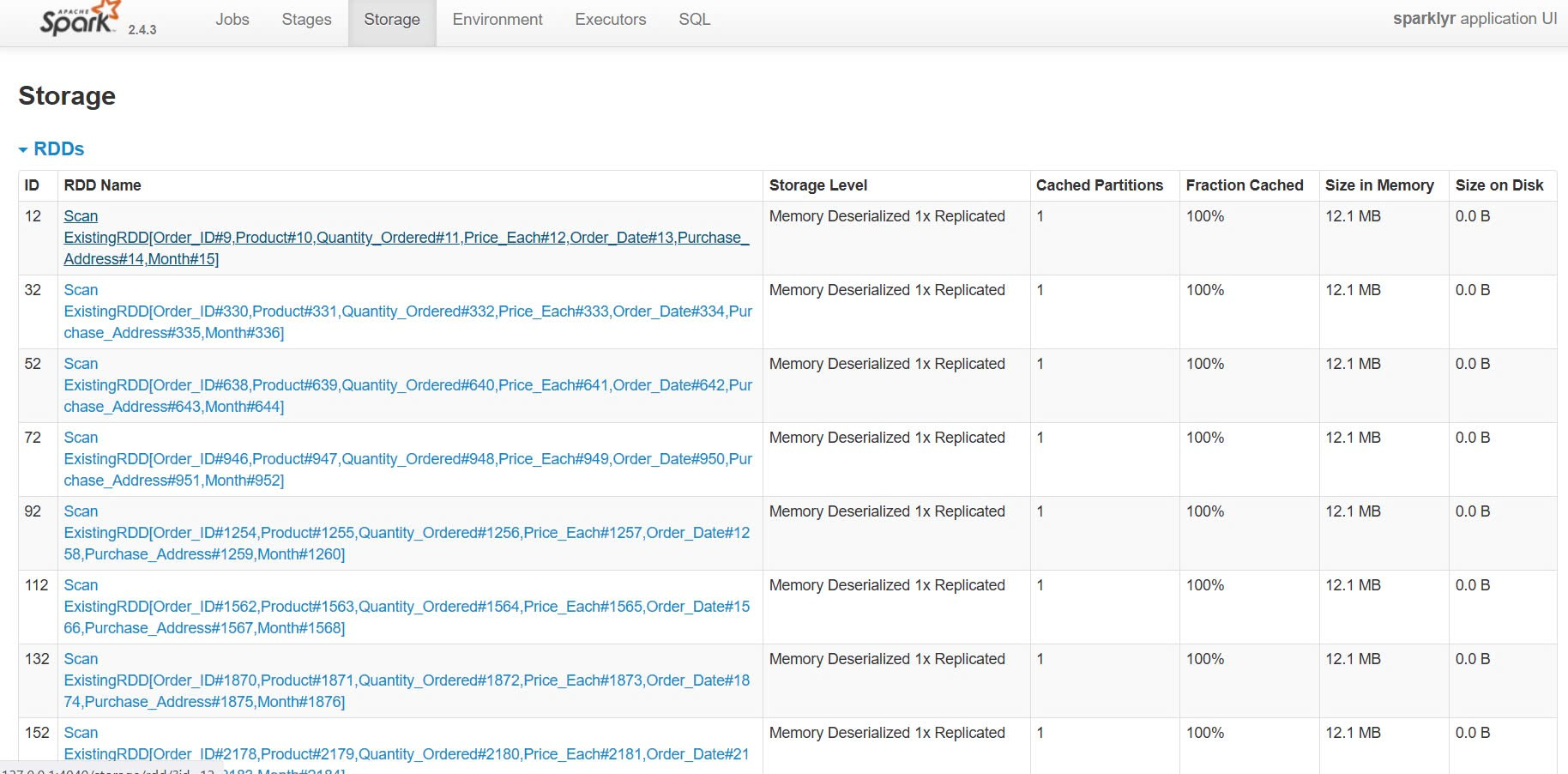
Biểu đồ giúp xác định tỷ lệ và giá của sản phẩm dựa trên mức giá đã đặt ra, hỗ trợ trong việc ra quyết định kinh doanh liên quan đến giá cả.  
Hình 5.5. Bảng thông tin lưu trữ dữ liệu trên RStudio

## 5.3. Bảng thống kê dữ liệu



* Sau khi chạy ra kết quả xong và đoạn mã code này:
* “file\_path <- "sales2019clean.csv"
* df\_combined <- read.csv(file\_path, stringsAsFactors = FALSE)”
* Có chức năng đọc và xuất dữ liệu có sẵn trong đó và hiển thị thông tin các sản phẩm ra một file có tên là “df\_combined” như hình ảnh trên.

## 5.4 Lưu trữ thông tin trên Spark



Hình 5.6. Bảng thông tin lưu trữ dữ liệu trên Spark

Giao diện **Storage** của Spark UI, cụ thể là danh sách **RDDs (Resilient Distributed Datasets)** đang được lưu trữ trong bộ nhớ.

Dưới đây là giải thích chi tiết về các cột trong bảng:

1. **ID**: Mã định danh (ID) của mỗi RDD trong Spark.
2. **RDD Name**: Tên của RDD. Dựa vào tên, ta thấy đây là dữ liệu liên quan đến đơn hàng (**Order\_ID, Product\_ID, Quantity, Price, Order\_Date, Purchase\_Address**).
3. **Storage Level**: Mức lưu trữ của RDD. Ở đây, tất cả các RDD đều được lưu dưới dạng **Memory Deserialized 1x Replicated**, có nghĩa là dữ liệu được giải nén vào bộ nhớ (RAM) và chỉ lưu một bản sao.
4. **Cached Partitions**: Số phân vùng (partitions) của RDD được lưu vào bộ nhớ. Mỗi RDD ở đây có 1 partition.
5. **Fraction Cached**: Phần trăm dữ liệu của RDD được lưu trong bộ nhớ. Ở đây tất cả đều là **100%**, có nghĩa là toàn bộ dữ liệu của RDD được lưu trữ mà không bị đổ xuống đĩa.
6. **Size in Memory**: Dung lượng của mỗi RDD trong bộ nhớ RAM. Mỗi RDD chiếm khoảng **12.1 MB**.
7. **Size on Disk**: Dung lượng của mỗi RDD trên đĩa. Ở đây, tất cả đều là **0.0 B**, có nghĩa là dữ liệu không bị ghi xuống đĩa, toàn bộ được giữ trong RAM.

### **Ý nghĩa**

* Spark đang lưu trữ dữ liệu của bạn trong bộ nhớ RAM để tăng tốc độ xử lý.
* Mức lưu trữ **Memory Deserialized 1x Replicated** cho thấy dữ liệu đã được giải nén, giúp truy xuất nhanh hơn nhưng tiêu tốn nhiều bộ nhớ hơn so với lưu trữ ở dạng nén.
* Không có dữ liệu nào bị đổ xuống ổ đĩa (**Size on Disk = 0.0 B**), nghĩa là bộ nhớ vẫn đủ để chứa toàn bộ dữ liệu.
* Dung lượng bộ nhớ bị chiếm bởi mỗi RDD là **12.1 MB**, nếu dữ liệu tiếp tục tăng, bạn có thể gặp vấn đề về bộ nhớ tràn (Out of Memory).

# KẾT LUẬN

**Ưu điểm**

Việc xây dựng mô hình dự đoán doanh số bán hàng mang lại nhiều lợi ích quan trọng cho doanh nghiệp. Trước hết, mô hình giúp cải thiện chiến lược kinh doanh bằng cách xác định thời điểm doanh số cao và thấp trong năm, từ đó tối ưu hóa các chiến dịch tiếp thị và khuyến mãi. Chẳng hạn, dữ liệu cho thấy doanh thu thường tăng mạnh vào tháng 11 và 12 do các dịp lễ hội, vì vậy doanh nghiệp có thể đẩy mạnh quảng cáo và giảm giá vào thời điểm này. Bên cạnh đó, mô hình hỗ trợ tối ưu hóa hoạt động phân phối bằng cách xác định khu vực có doanh số cao, chẳng hạn như San Francisco và Portland, giúp doanh nghiệp tập trung nguồn lực phù hợp. Ngoài ra, việc phân tích nhu cầu sản phẩm thông qua biểu đồ số lượng đặt hàng giúp quản lý hàng tồn kho hiệu quả hơn. Ví dụ, nếu tai nghe có số lượng đặt hàng cao nhất, doanh nghiệp có thể điều chỉnh sản xuất hoặc nhập hàng để đáp ứng nhu cầu. Hơn nữa, mô hình còn giúp nhận diện khung giờ cao điểm của doanh số, điển hình như khoảng thời gian từ 12h đến 15h, cho phép tối ưu hóa nhân sự và chiến lược marketing vào những thời điểm quan trọng này. Ngoài ra, sử dụng mô hình cây quyết định còn giúp doanh nghiệp dự báo xu hướng giá cả và điều chỉnh giá bán sản phẩm để tối ưu lợi nhuận. Đặc biệt, với sự hỗ trợ của Spark trong RStudio, quá trình xử lý dữ liệu lớn diễn ra nhanh chóng và hiệu quả hơn nhờ cơ chế lưu trữ trong bộ nhớ RAM, giúp truy xuất và phân tích dữ liệu một cách chính xác và kịp thời. Nhìn chung, việc ứng dụng mô hình dự đoán doanh số bán hàng không chỉ giúp doanh nghiệp tối ưu hóa doanh thu mà còn hỗ trợ ra quyết định kinh doanh thông minh, cải thiện quản lý hàng hóa và tăng hiệu quả vận hành.  
**Nhược điểm** Mặc dù mô hình dự đoán doanh số bán hàng mang lại nhiều lợi ích, nhưng nó cũng tồn tại một số nhược điểm cần lưu ý. Trước hết, độ chính xác của mô hình phụ thuộc rất lớn vào chất lượng dữ liệu đầu vào. Nếu dữ liệu bị thiếu, sai lệch hoặc không phản ánh đúng xu hướng thị trường, dự báo có thể không chính xác, dẫn đến quyết định kinh doanh sai lầm. Ngoài ra, mô hình có thể không thích ứng tốt với những biến động bất ngờ của thị trường, chẳng hạn như thay đổi hành vi mua sắm do khủng hoảng kinh tế, xu hướng tiêu dùng mới hoặc các yếu tố bất khả kháng như đại dịch. Hơn nữa, việc triển khai và duy trì mô hình đòi hỏi doanh nghiệp phải có đội ngũ nhân sự am hiểu về phân tích dữ liệu, lập trình và các công cụ như Spark, RStudio, điều này có thể gây khó khăn cho các doanh nghiệp nhỏ chưa có đủ nguồn lực. Bên cạnh đó, mô hình cây quyết định có thể trở nên quá phức tạp và dễ bị overfitting khi áp dụng trên dữ liệu lớn, khiến dự báo kém chính xác khi gặp dữ liệu mới. Hệ thống lưu trữ dữ liệu trên Spark mặc dù giúp tăng tốc xử lý, nhưng lại tiêu tốn nhiều bộ nhớ RAM, và nếu dữ liệu tiếp tục mở rộng, doanh nghiệp có thể gặp vấn đề về hiệu suất và chi phí hạ tầng. Cuối cùng, các quyết định kinh doanh không nên chỉ dựa hoàn toàn vào dự đoán từ mô hình mà cần kết hợp với kinh nghiệm thực tế và đánh giá thị trường để đưa ra những chiến lược phù hợp nhất.

**Hướng phát triển**

Trong tương lai, mô hình dự đoán doanh số bán hàng có thể được cải thiện bằng cách áp dụng các thuật toán tiên tiến hơn như Random Forest, XGBoost hoặc mạng nơ-ron nhân tạo (Neural Networks) để nâng cao độ chính xác trong dự báo. Việc kết hợp nhiều nguồn dữ liệu đa dạng như mạng xã hội, phản hồi khách hàng, xu hướng tìm kiếm trên Google, thông tin thời tiết hay tình hình kinh tế sẽ giúp mô hình có cái nhìn toàn diện hơn về yếu tố ảnh hưởng đến doanh số. Ngoài ra, việc áp dụng xử lý dữ liệu theo thời gian thực bằng Apache Spark kết hợp với hệ thống điện toán đám mây như AWS, Google Cloud hay Microsoft Azure sẽ giúp tăng tốc độ tính toán và mở rộng quy mô mô hình khi dữ liệu ngày càng lớn. Bên cạnh đó, cá nhân hóa dự đoán doanh số theo từng khu vực, nhóm khách hàng hay sản phẩm sẽ giúp doanh nghiệp tối ưu chiến lược tiếp thị, quản lý hàng tồn kho và xây dựng các chương trình khuyến mãi phù hợp. Chẳng hạn, AI có thể phân tích thói quen mua sắm của từng khách hàng để đưa ra gợi ý sản phẩm và khuyến mãi cá nhân hóa, từ đó tăng doanh thu và mức độ hài lòng của khách hàng. Hơn nữa, để nâng cao khả năng phân tích dữ liệu, các doanh nghiệp có thể triển khai các công cụ trực quan hóa như R Shiny, Tableau hoặc Power BI, giúp dễ dàng theo dõi xu hướng kinh doanh, xác định các điểm bất thường và đưa ra quyết định kịp thời.

Một hướng phát triển quan trọng khác là cải thiện khả năng tự động hóa của mô hình, giúp hệ thống có thể tự học từ dữ liệu mới và điều chỉnh dự báo mà không cần can thiệp quá nhiều từ con người. Điều này đòi hỏi các thuật toán học máy có khả năng cập nhật liên tục, cùng với hệ thống giám sát để phát hiện và điều chỉnh khi mô hình có sai số lớn. Ngoài ra, việc đào tạo nhân sự về khoa học dữ liệu, phân tích dữ liệu lớn (Big Data) và ứng dụng AI trong kinh doanh cũng đóng vai trò quan trọng để đảm bảo mô hình được vận hành và tối ưu hóa hiệu quả. Tóm lại, bằng cách kết hợp công nghệ tiên tiến, mở rộng dữ liệu đầu vào và nâng cao khả năng phân tích, mô hình dự đoán doanh số bán hàng sẽ ngày càng chính xác, giúp doanh nghiệp đưa ra quyết định kinh doanh hiệu quả hơn.

# LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành đến **Trường Đại học Đại Nam** cùng hai giảng viên môn **Dữ liệu lớn** – **cô Lê Thị Thùy Trang và thầy Trần Quý Nam** – những người đã tận tình giảng dạy, truyền đạt kiến thức và hỗ trợ em trong suốt quá trình học tập. Nhờ sự hướng dẫn của thầy cô, em đã có cơ hội tiếp cận và hiểu rõ hơn về lĩnh vực **dữ liệu lớn**, một trong những công nghệ quan trọng có tầm ảnh hưởng sâu rộng trong kỷ nguyên số hóa hiện nay.

Trước khi học môn này, em chưa thực sự nhận thức rõ tầm quan trọng của dữ liệu lớn trong khoa học, kinh doanh và đời sống. Tuy nhiên, qua mỗi buổi học, thầy cô đã giúp em nắm vững các khái niệm từ **xử lý dữ liệu phân tán, khai thác dữ liệu**, đến các công nghệ tiên tiến như **Hadoop, Spark, NoSQL**. Thầy cô không chỉ cung cấp kiến thức lý thuyết mà còn hướng dẫn ứng dụng thực tế thông qua các bài tập thực hành, giúp em hiểu rõ hơn về cách khai thác dữ liệu một cách hiệu quả và có hệ thống.

Trong quá trình học tập và làm bài tập nhóm, em gặp không ít khó khăn khi phải xử lý lượng dữ liệu khổng lồ, tối ưu hóa mô hình phân tích và hiểu sâu về thuật toán. Nhưng nhờ sự hướng dẫn tận tình của thầy cô, em đã từng bước làm chủ các công cụ, rèn luyện tư duy logic và tiếp cận vấn đề một cách khoa học hơn. Sự nhiệt tình giảng dạy và những chỉ dẫn cụ thể của thầy cô giúp em có cái nhìn toàn diện hơn về tiềm năng cũng như thách thức của dữ liệu lớn trong thực tế.

Môn **Dữ liệu lớn** không chỉ giúp em nâng cao kiến thức chuyên môn mà còn rèn luyện **tư duy phân tích, kỹ năng lập trình, làm việc nhóm và giải quyết vấn đề**. Thông qua các bài tập thực tế, em nhận ra rằng dữ liệu không chỉ là những con số vô tri mà còn là nguồn tài nguyên quý giá, giúp doanh nghiệp đưa ra các quyết định chiến lược và tối ưu hóa hiệu suất hoạt động.

Bên cạnh đó, em cũng nhận thức rõ hơn về những **thách thức trong việc quản lý dữ liệu**, đặc biệt là các vấn đề liên quan đến **tốc độ xử lý, bảo mật thông tin và khả năng mở rộng hệ thống**. Khi khối lượng dữ liệu ngày càng tăng, việc áp dụng các công nghệ phù hợp để tối ưu hóa xử lý và lưu trữ trở nên vô cùng quan trọng. Nhờ sự giảng dạy tận tâm của thầy cô, em đã biết cách tiếp cận và áp dụng những kỹ thuật phù hợp để giải quyết các bài toán dữ liệu thực tế.

Một lần nữa, em xin bày tỏ lòng biết ơn sâu sắc đến **cô Lê Thị Thùy Trang và thầy Trần Quý Nam** vì sự hướng dẫn tận tình và những bài giảng đầy tâm huyết. Em kính chúc thầy cô **luôn dồi dào sức khỏe, thành công trong sự nghiệp giảng dạy** và tiếp tục truyền cảm hứng cho nhiều thế hệ sinh viên. Những kiến thức và kỹ năng mà em học được từ môn **Dữ liệu lớn** sẽ là hành trang quý giá giúp em tự tin hơn trên con đường học tập và phát triển sự nghiệp trong tương lai.

**TÀI LIỆU THAM KHẢO**

[1] Giáo trình Dữ liệu lớn, Tác giả: Trường Đại học Bách Khoa Hà Nội, Năm xuất bản: 2021.

[2] Big Data: Nguyên lý và ứng dụng, Tác giả: Nguyễn Văn Hùng, Năm xuất bản: 2022.

[3] Phân tích dữ liệu lớn với Python, Tác giả: Lê Minh Tuấn, Năm xuất bản: 2023.

[4] Hệ thống quản lý dữ liệu lớn, Tác giả: Trần Quốc Bảo, Năm xuất bản: 2024.

[5] Khoa học dữ liệu và ứng dụng trong kinh doanh, Tác giả: Hoàng Anh Dũng, Năm xuất bản: 2020.

[6] Xử lý dữ liệu lớn với Apache Spark, Tác giả: Phạm Thanh Sơn, Năm xuất bản: 2023.

[7] Học máy và dữ liệu lớn, Tác giả: Đặng Thị Lan, Năm xuất bản: 2022.

[8] Cơ sở dữ liệu phân tán và dữ liệu lớn, Tác giả: Nguyễn Quốc Trung, Năm xuất bản: 2021.

[9] Quản lý và khai thác dữ liệu lớn, Tác giả: Bùi Hữu Thắng, Năm xuất bản: 2024.

[10] Big Data trong trí tuệ nhân tạo, Tác giả: Trần Ngọc Hải, Năm xuất bản: 2023.

[11] Phân tích dữ liệu lớn trong thương mại điện tử, Tác giả: Lương Bảo Châu, Năm xuất bản: 2025.

[12] Ứng dụng dữ liệu lớn trong tài chính và ngân hàng, Tác giả: Vũ Hoàng Nam, Năm xuất bản: 2024.

[13] Hệ thống lưu trữ và xử lý dữ liệu lớn, Tác giả: Nguyễn Văn Tài, Năm xuất bản: 2022.

[14] Bảo mật và an toàn dữ liệu lớn, Tác giả: Hoàng Minh Tuấn, Năm xuất bản: 2023.

[15] Phân tích dữ liệu lớn với R, Tác giả: Lê Thị Hồng, Năm xuất bản: 2024.

[16] Kỹ thuật khai phá dữ liệu lớn, Tác giả: Phạm Quốc Huy, Năm xuất bản: 2025.

[17] Ứng dụng dữ liệu lớn trong y tế, Tác giả: Nguyễn Thị Hoa, Năm xuất bản: 2023.

[18] Hướng dẫn sử dụng Hadoop cho dữ liệu lớn, Tác giả: Đỗ Mạnh Hùng, Năm xuất bản: 2024.