**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**A blue circle with text

Description automatically generated**

**BÁO CÁO ĐỒ ÁN**

**MÔN : PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI : PHÂN TÍCH DỮ LIỆU BÁN HÀNG CỦA DOANH NGHIỆP**

**Giáo viên hướng dẫn : Thầy Phan Thành Huấn**

**Sinh viên**

Nguyễn Hoài Lâm – 3120410274

**TP. HỒ CHÍ MINH, Ngày 12 tháng 11 Năm 2023**

MỤC LỤC

[CHƯƠNG I. TỔNG QUAN 5](#_Toc152725541)

[1 . Khái quát về phân tích dữ liệu bán hàng của doanh nghiệp 5](#_Toc152725542)

[1.1. Khái niệm phân tích dữ liệu 5](#_Toc152725543)

[1.2. Phân tích dữ liệu bán hàng. 6](#_Toc152725544)

[2 . Giới thiệu về đề tài 7](#_Toc152725545)

[3 . Lý do chọn đề tài 7](#_Toc152725546)

[4 . Mục tiêu của đề tài 7](#_Toc152725547)

[5 . Đối tượng nghiên cứu 8](#_Toc152725548)

[6 . Công cụ phân tích 8](#_Toc152725549)

[7 . Mô tả dữ liệu và cấu trúc dữ liệu 8](#_Toc152725550)

[CHƯƠNG II . QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ 9](#_Toc152725551)

[1 . Các phương pháp phân tích dữ liệu 9](#_Toc152725552)

[1.1 . Cây quyết định (Descion Tree) 9](#_Toc152725553)

[1.2 . Rừng ngẫu nhiên (Random Forest) 10](#_Toc152725554)

[1.3. Hồi quy tuyến tính 11](#_Toc152725555)

[1.4 . Phân tích biểu đồ và đồ thị 11](#_Toc152725556)

[2 . Tiền xử lý dữ liệu 12](#_Toc152725557)

[3 . Machine learning và trực quan hóa dữ liệu 16](#_Toc152725558)

[3.1. Machine learning 16](#_Toc152725559)

[3.2. Trực quan hóa dữ liệu 19](#_Toc152725560)

[CHƯƠNG III . KẾT LUẬN 23](#_Toc152725561)

LỜI CẢM ƠN

Nhóm em xin gửi lời cảm ơn chân thành đến thầy Phan Thành Huấn - giáo viên của môn học này, vì đã cống hiến thời gian và nỗ lực để giảng dạy và hỗ trợ chúng em trong quá trình học tập. Thầy là người đã truyền đạt cho chúng em những kiến thức và kỹ năng quan trọng trong môn học này, giúp chúng em hiểu sâu hơn về hệ thống thông tin và ứng dụng của nó trong thực tế. Bên cạnh đó, thầy còn luôn sẵn sàng giúp đỡ và hỗ trợ chúng em trong quá trình làm bài tập và nghiên cứu đề tài. Chúng em rất biết ơn vì sự quan tâm và giúp đỡ của thầy đã giúp chúng em hoàn thành môn học này một cách tốt nhất. Chúng em hy vọng sẽ có cơ hội được học tập và làm việc với thầy trong những dự án sắp tới. Một lần nữa, chúng em xin chân thành cảm ơn thầy Phan Thành Huấn.

**ĐÁNH GIÁ MỨC ĐỘ THAM GIA**

|  |  |  |
| --- | --- | --- |
| Danh sách thành viên | MSSV | Mức độ tham gia |
| Nguyễn Hoài Lâm | 3120410274 | 100% |
| Nguyễn Phan Huy Lượng | 3120410312 | 100% |

# CHƯƠNG I. TỔNG QUAN

## . Khái quát về phân tích dữ liệu bán hàng của doanh nghiệp

### 1.1. Khái niệm phân tích dữ liệu

Phân tích dữ liệu (Data Analytics) là quá trình thu thập, xử lý và tìm hiểu dữ liệu để tìm ra các thông tin hữu ích. Các thông tin này có thể được sử dụng để đưa ra quyết định, cải thiện hiệu suất hoặc khám phá các xu hướng mới.

Phân tích dữ liệu có thể được áp dụng cho nhiều lĩnh vực khác nhau, bao gồm kinh doanh, khoa học, y tế, giáo dục, v.v. Trong kinh doanh, phân tích dữ liệu được sử dụng để phân tích dữ liệu khách hàng, dữ liệu bán hàng, dữ liệu sản phẩm, v.v. để hiểu rõ hơn về khách hàng, cải thiện hiệu suất bán hàng và phát triển các sản phẩm mới.

Có nhiều loại phân tích dữ liệu khác nhau, mỗi loại có mục đích riêng. Một số loại phân tích dữ liệu phổ biến bao gồm:

* Phân tích mô tả: Trả lời các câu hỏi "Điều gì đã xảy ra?". Loại phân tích này sử dụng các phép thống kê để mô tả dữ liệu, chẳng hạn như tìm xu hướng, phân tích phân phối và xác định giá trị ngoại lệ.
* Phân tích chẩn đoán: Trả lời các câu hỏi "Tại sao điều đó lại xảy ra?". Loại phân tích này sử dụng các kỹ thuật thống kê và máy học để tìm các mối tương quan giữa các biến, chẳng hạn như xác định các yếu tố ảnh hưởng đến doanh số bán hàng.
* Phân tích dự đoán: Trả lời các câu hỏi "Điều gì sẽ xảy ra?". Loại phân tích này sử dụng các mô hình thống kê và máy học để dự đoán các giá trị trong tương lai, chẳng hạn như dự đoán doanh số bán hàng trong tương lai.
* Phân tích đề xuất: Trả lời các câu hỏi "Nên làm gì?". Loại phân tích này sử dụng các mô hình thống kê và máy học để đề xuất các hành động, chẳng hạn như đề xuất các sản phẩm hoặc dịch vụ cho khách hàng.

Phân tích dữ liệu là một công cụ mạnh mẽ có thể được sử dụng để cải thiện hiệu suất và đưa ra quyết định sáng suốt. Tuy nhiên, để phân tích dữ liệu hiệu quả, cần có kiến thức về thống kê, máy học và các kỹ năng phân tích dữ liệu.

Dưới đây là một số lợi ích của phân tích dữ liệu:

* Cải thiện hiệu suất: Phân tích dữ liệu có thể giúp doanh nghiệp hiểu rõ hơn về khách hàng, sản phẩm và quy trình của mình. Điều này có thể giúp doanh nghiệp cải thiện hiệu suất trong các lĩnh vực như bán hàng, marketing, sản xuất và dịch vụ khách hàng.
* Đưa ra quyết định sáng suốt: Phân tích dữ liệu có thể cung cấp cho doanh nghiệp thông tin chi tiết về các xu hướng và mối quan hệ. Điều này có thể giúp doanh nghiệp đưa ra quyết định sáng suốt hơn về chiến lược, sản phẩm và dịch vụ.
* Phát hiện các cơ hội mới: Phân tích dữ liệu có thể giúp doanh nghiệp phát hiện các cơ hội mới trong thị trường. Điều này có thể giúp doanh nghiệp mở rộng kinh doanh hoặc phát triển các sản phẩm hoặc dịch vụ mới.

Phân tích dữ liệu đang trở thành một lĩnh vực quan trọng trong nhiều ngành nghề. Các chuyên gia phân tích dữ liệu có kỹ năng và kiến thức cần thiết để phân tích dữ liệu và đưa ra thông tin chi tiết có giá trị.

### 1.2. Phân tích dữ liệu bán hàng.

Phân tích dữ liệu bán hàng là quá trình thu thập, xử lý và tìm hiểu dữ liệu bán hàng để tìm ra các thông tin hữu ích. Các thông tin này có thể được sử dụng để cải thiện hiệu suất bán hàng, đưa ra quyết định sáng suốt và phát triển các chiến lược bán hàng mới.

Dữ liệu bán hàng là một tập dữ liệu lớn và phức tạp. Nó bao gồm nhiều loại dữ liệu khác nhau, chẳng hạn như:

* Dữ liệu khách hàng: bao gồm thông tin về khách hàng, chẳng hạn như tên, địa chỉ, thông tin liên hệ, lịch sử mua hàng, v.v.
* Dữ liệu sản phẩm: bao gồm thông tin về sản phẩm, chẳng hạn như tên sản phẩm, giá cả, mô tả, v.v.
* Dữ liệu bán hàng: bao gồm thông tin về các giao dịch bán hàng, chẳng hạn như ngày, thời gian, số tiền, sản phẩm được bán, v.v.

Phân tích dữ liệu bán hàng có thể được sử dụng để trả lời nhiều loại câu hỏi, chẳng hạn như:

* Doanh số bán hàng hiện tại và tương lai là gì?
* Các sản phẩm hoặc dịch vụ nào bán chạy nhất?
* Khách hàng nào là khách hàng tiềm năng nhất?
* Các chiến dịch bán hàng nào hiệu quả nhất?

Phân tích dữ liệu bán hàng có thể mang lại nhiều lợi ích cho doanh nghiệp, bao gồm:

* Cải thiện hiệu suất bán hàng: Phân tích dữ liệu bán hàng có thể giúp doanh nghiệp xác định các khu vực cần cải thiện, chẳng hạn như quy trình bán hàng, chiến lược sản phẩm hoặc dịch vụ khách hàng.
* Đưa ra quyết định sáng suốt: Phân tích dữ liệu bán hàng có thể cung cấp cho doanh nghiệp thông tin chi tiết về các xu hướng và mối quan hệ. Điều này có thể giúp doanh nghiệp đưa ra quyết định sáng suốt hơn về chiến lược bán hàng, sản phẩm và dịch vụ.
* Phát hiện các cơ hội mới: Phân tích dữ liệu bán hàng có thể giúp doanh nghiệp phát hiện các cơ hội mới trong thị trường. Điều này có thể giúp doanh nghiệp mở rộng kinh doanh hoặc phát triển các sản phẩm hoặc dịch vụ mới.

## 2 . Giới thiệu về đề tài

Đề tài "Phân tích dữ liệu bán hàng của doanh nghiệp" là một chủ đề quan trọng và hấp dẫn trong lĩnh vực quản lý kinh doanh và phân tích dữ liệu. Nó tập trung vào việc sử dụng công nghệ và phương pháp phân tích dữ liệu để hiểu rõ hơn về quá trình bán hàng của một doanh nghiệp và từ đó đưa ra các quyết định chiến lược. Báo cáo này xoay quanh việc phân tích dữ liệu bán hàng của doanh nghiệp, với mục tiêu rút ra những quan sát sâu sắc, xác định các mô hình và tiết lộ các cơ hội phát triển. Bằng cách sử dụng Python và các thư viện phân tích dữ liệu, nhóm chúng em sẽ tiến hành phân tích và tìm hiểu về dữ liệu bán hàng của doanh nghiệp.

## 3 . Lý do chọn đề tài

Phân tích dữ liệu bán hàng là một hoạt động quan trọng đối với mọi doanh nghiệp, bất kể quy mô hay ngành nghề. Việc phân tích dữ liệu bán hàng giúp đưa ra nhận xét chính xác nhất về xu hướng thị trường, từ đó doanh nghiệp có thể đưa ra hướng phán đoán và xử lý chính xác, tránh gây ra tổn thất cho doanh nghiệp.

Trong thị trường cạnh tranh khốc liệt như hiện nay, việc phân tích dữ liệu bán hàng của đối thủ cũng có thể giúp chúng ta dự đoán phần nào hướng đi tiếp theo của doanh nghiệp đối thủ, từ đó đưa ra chiến lược cạnh tranh hiệu quả hơn, nhờ đó giúp doanh nghiệp đón đầu đối thủ và đưa doanh nghiệp ngày càng phát triển hơn.

Chỉ bấy nhiêu đó thôi cũng có thể thấy được tầm quan trọng của việc Phân tích dữ liệu bán hàng của doanh nghiệp đem lại nhiều lợi ích đến thế nào. Nhận thấy được tiềm năng to lớn đó nên nhóm em quyết định chọn đề tài này để nghiên cứu và cũng như tìm hiểu thêm về lĩnh vực này.

Hy vọng rằng nghiên cứu của chúng em sẽ góp phần giúp các doanh nghiệp hiểu rõ hơn về hoạt động bán hàng của mình và đưa ra những quyết định kinh doanh hiệu quả hơn.

## 4 . Mục tiêu của đề tài

Mục tiêu của đề tài "Phân tích dữ liệu bán hàng của doanh nghiệp" là cung cấp cho các doanh nghiệp những kiến thức và kỹ năng cần thiết để phân tích dữ liệu bán hàng hiệu quả. Cụ thể, mục tiêu của đề tài bao gồm:

* Hiểu rõ về khái niệm, vai trò và ứng dụng của phân tích dữ liệu bán hàng.
* Nắm được các loại dữ liệu bán hàng và phương pháp phân tích dữ liệu bán hàng.
* Thực hành phân tích dữ liệu bán hàng bằng các công cụ và phần mềm chuyên dụng.

## 5 . Đối tượng nghiên cứu

Đối tượng nghiên cứu của đồ án là dựa trên dữ liệu bán hàng - Adidas được công bố bởi các trang web có độ uy tín cao.

## 6 . Công cụ phân tích

Công cụ mà chúng em xử dụng để phân tích dữ liệu bán hàng của doanh nghiệp là Python, ở đây tích hợp nhiều thư viện, chức năng, từ đó thuận tiện hơn cho việc phân tích dữ liệu.

## 7 . Mô tả dữ liệu và cấu trúc dữ liệu

Dữ liệu được tổng hợp từ :

<https://www.kaggle.com/datasets/heemalichaudhari/adidas-sales-dataset/>

Đây là những nguồn được đánh giá là có những bộ dữ liệu đáng tin cậy để phục vụ cho việc thu thập dữ liệu và nghiên cứu.

|  |  |  |
| --- | --- | --- |
| Thuộc tính | Ý nghĩa | Mô tả |
| Retailer | Đơn vị hoặc tổ chức bán sản phẩm Adidas | Object |
| Retailer ID | Mã định danh duy nhất cho mỗi nhà bán lẻ | Object |
| Invoice Date | Ngày diễn ra giao dịch bán hàng | Object |
| Region | Khu vực địa lý nơi nhà bán lẻ hoạt động | Object |
| State | Bang trong khu vực nơi đặt trụ sở của nhà bán lẻ | Object |
| City | Thành phố nơi nhà bán lẻ đặt trụ sở | Object |
| Product | Sản phẩm Adidas đang được bán | Object |
| Price per Unit | Giá của một đơn vị sản phẩm Adidas | Object |
| Units Sold | Số lượng đơn vị sản phẩm Adidas được bán trong một giao dịch cụ thể | Object |
| Total Sales | Tổng doanh thu được tạo ra từ việc bán sản phẩm Adidas trong một giao dịch | Object |
| Operating Profit | Lợi nhuận mà nhà bán lẻ kiếm được từ việc bán hàng sau khi trừ chi phí hoạt động | Object |
| Operating Margin | Tỷ lệ phần trăm lợi nhuận hoạt động so với tổng doanh thu | Object |
| Sales Method | Phương thức hoặc kênh mà qua đó giao dịch bán hàng diễn ra | Object |

# CHƯƠNG II . QUY TRÌNH THỰC HIỆN VÀ KẾT QUẢ

## 1 . Các phương pháp phân tích dữ liệu

### 1.1 . Cây quyết định (Descion Tree)

Cây quyết định là một mô hình trong machine learning thuộc họ supervised learning, được sử dụng cho cả vấn đề phân loại (classification) và hồi quy (regression). Mô hình này dựa trên việc xây dựng một cây quyết định dựa trên dữ liệu huấn luyện để đưa ra dự đoán.

**Cách Cây Quyết Định Hoạt Động:**

1. Chọn Đặc Trưng: Chọn đặc trưng (feature) dựa trên tiêu chí nào đó sao cho việc phân loại hoặc dự đoán có tính phân chia tốt nhất.
2. Chia Dữ Liệu: Phân tách dữ liệu thành các nhóm con dựa trên giá trị của đặc trưng đã chọn. Mục tiêu là tạo ra các nhóm có tính đồng nhất cao về phân loại hoặc dự đoán.
3. Lặp Lại Quá Trình: Lặp lại quá trình trên với từng nhóm con. Tiếp tục chia đến khi một điều kiện dừng được đạt được, chẳng hạn như đạt đến một độ sâu tối đa hoặc không còn có thể phân chia dữ liệu hiệu quả.
4. Tạo Nút Lá (Leaf Node): Khi một điều kiện dừng được đạt được, tạo nút lá và gán nhãn (phân loại) hoặc giá trị dự đoán (hồi quy) cho nút lá.
5. Dự Đoán: Khi có dữ liệu mới cần dự đoán, mô hình đi qua cây quyết định từ nút gốc đến một nút lá và trả về kết quả của nút lá đó.

**Điểm Mạnh và Yếu Điểm của Cây Quyết Định:**

**Điểm Mạnh:**

* Dễ Hiểu và Giải Thích:Cây quyết định mang lại quyết định rõ ràng và dễ giải thích, giúp hiểu rõ quá trình ra quyết định của mô hình.
* Khả Năng Xử Lý Đặc Trưng Phi Tuyến:Có khả năng xử lý đặc trưng phi tuyến, tức là không cần phải chuẩn bị đặc trưng một cách tuyến tính.
* Khả Năng Xử Lý Dữ Liệu Nhiễu:Cây quyết định có khả năng xử lý dữ liệu nhiễu mà không cần nhiều tiền xử lý.

**Yếu Điểm:**

* Overfitting:Cây quyết định có thể dễ bị overfitting trên dữ liệu huấn luyện, đặc biệt là khi cây quá sâu và quá phức tạp.
* Khả Năng Dễ Bị Nhiễu:Cây quyết định có thể dễ bị nhiễu bởi dữ liệu nhiễu, dẫn đến việc tạo ra quyết định không ổn định.

### 1.2 . Rừng ngẫu nhiên (Random Forest)

Random Forest (Rừng Ngẫu Nhiên) là một phương pháp trong machine learning thuộc họ ensemble learning. Nó được thiết kế để cải thiện hiệu suất dự đoán và giảm nguy cơ overfitting so với các mô hình quyết định đơn lẻ, chẳng hạn như cây quyết định.

**Cách Random Forest hoạt động:**

1. Tạo Rừng Ngẫu Nhiên:

* Random Forest xây dựng nhiều cây quyết định (có thể là hàng trăm hoặc thậm chí hàng nghìn) trên các tập dữ liệu con được rút ra từ tập dữ liệu huấn luyện gốc thông qua quy trình rút ngẫu nhiên.
* Mỗi cây quyết định được xây dựng độc lập từ nhau.

1. Rút Ngẫu Nhiên Dữ Liệu:

* Mỗi lần xây dựng một cây, một tập con ngẫu nhiên được rút ra từ tập dữ liệu huấn luyện gốc. Điều này giúp đa dạng hóa các cây và giảm nguy cơ overfitting.

1. Xây Dựng Cây Quyết Định:

* Mỗi cây quyết định được xây dựng dựa trên tập dữ liệu con được rút ra và một số đặc trưng được chọn ngẫu nhiên ở mỗi bước.

1. Voting hoặc Trung Bình Hóa:

* Khi có dữ liệu mới cần dự đoán, mỗi cây quyết định trong rừng đưa ra dự đoán của mình.
* Đối với vấn đề phân loại, kết quả cuối cùng là dự đoán được đưa ra thông qua quá trình bỏ phiếu (voting) từ tất cả các cây.
* Đối với vấn đề hồi quy, kết quả có thể được tính toán dựa trên trung bình hoặc trung vị của dự đoán từ các cây.

**Lợi Ích của Random Forest:**

* Ổn Định và Hiệu Quả: Random Forest thường cho hiệu suất dự đoán tốt và ít nhạy cảm với overfitting hơn so với các mô hình đơn lẻ.
* Xử Lý Dữ Liệu Nhiễu: Khả năng xử lý dữ liệu nhiễu và giữ lại đặc trưng quan trọng là một điểm mạnh của Random Forest.
* Thuật Toán Tích Hợp: Có thể sử dụng cho cả vấn đề phân loại và hồi quy.
* Dễ Điều Chỉnh: Có thể điều chỉnh các tham số như số cây, độ sâu cây để tối ưu hóa mô hình.

### 1.3. Hồi quy tuyến tính

Hồi quy tuyến tính là một trong những phương pháp quan trọng trong machine learning, được sử dụng để mô hình hóa mối quan hệ tuyến tính giữa biến độc lập và biến mục tiêu. Nó được sử dụng chủ yếu cho các vấn đề dự đoán giá trị liên tục, và mục tiêu là tìm ra đường tuyến tính sao cho dự đoán của mô hình gần với giá trị thực tế nhất.

**Cách Hồi Quy Tuyến Tính Hoạt Động**:

1. **Giả định mô hình tuyến tính**: Mô hình giả định rằng có một mối quan hệ tuyến tính giữa biến độc lập và biến mục tiêu. Điều này được biểu diễn bằng một đường thẳng.
2. **Tìm Tham Số Tối Ưu**: Sử dụng phương pháp như phương pháp tối thiểu hóa sai số bình phương (Least Squares) để tìm ra tham số của đường tuyến tính sao cho tổng các bình phương của sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất.
3. **Đánh Giá Mô Hình**: Mô hình được đánh giá thông qua các thước đo như Mean Squared Error (MSE) hoặc R-squared để đo lường khả năng giải thích của mô hình.

### 1.4 . Phân tích biểu đồ và đồ thị

Phân tích biểu đồ và đồ thị là một phần quan trọng trong phân tích dữ liệu. Biểu đồ và đồ thị giúp chúng ta có thể trực quan hóa dữ liệu, từ đó dễ dàng nhận biết các xu hướng, mối tương quan và các vấn đề tiềm ẩn trong dữ liệu.

Có nhiều phương pháp khác nhau để phân tích biểu đồ và đồ thị trong phân tích dữ liệu. Một số phương pháp phổ biến bao gồm:

* Xem xét tổng thể: Bước đầu tiên trong việc phân tích biểu đồ và đồ thị là xem xét tổng thể biểu đồ. Điều này bao gồm việc xác định các yếu tố sau:
  + Dữ liệu được thể hiện theo cách nào?
  + Các biến được thể hiện trên mỗi trục?
  + Biểu đồ sử dụng các yếu tố trực quan nào?
* Xác định các xu hướng: Sau khi xem xét tổng thể biểu đồ, chúng ta cần xác định các xu hướng trong dữ liệu. Các xu hướng có thể được xác định bằng cách xem xét các điểm dữ liệu theo thời gian, theo nhóm hoặc theo các biến khác.
* Xác định mối tương quan: Chúng ta cũng có thể sử dụng biểu đồ và đồ thị để xác định mối tương quan giữa các biến. Mối tương quan là mức độ liên quan giữa hai biến. Có nhiều loại mối tương quan khác nhau, bao gồm mối tương quan tuyến tính, mối tương quan phi tuyến tính và mối tương quan không tương quan.
* Xác định các vấn đề tiềm ẩn: Biểu đồ và đồ thị cũng có thể được sử dụng để xác định các vấn đề tiềm ẩn trong dữ liệu. Các vấn đề tiềm ẩn có thể bao gồm các lỗi dữ liệu, các xu hướng bất thường hoặc các mẫu không mong muốn.

Dưới đây là một số ví dụ cụ thể về cách phân tích biểu đồ và đồ thị trong phân tích dữ liệu:

* Biểu đồ đường: Biểu đồ đường thường được sử dụng để hiển thị sự thay đổi của một biến theo thời gian. Ví dụ, một nhà phân tích kinh doanh có thể sử dụng biểu đồ đường để theo dõi doanh số bán hàng theo tháng.
* Biểu đồ thanh: Biểu đồ thanh thường được sử dụng để so sánh các giá trị của một biến giữa các nhóm. Ví dụ, một nhà nghiên cứu có thể sử dụng biểu đồ thanh để so sánh mức độ hài lòng của khách hàng giữa các phân khúc khách hàng khác nhau.
* Biểu đồ phân tán: Biểu đồ phân tán thường được sử dụng để xác định mối tương quan giữa hai biến. Ví dụ, một nhà khoa học có thể sử dụng biểu đồ phân tán để xác định mối tương quan giữa cân nặng và chiều cao.
* Bảng nhiệt: Bảng nhiệt thường được sử dụng để hiển thị các giá trị của một biến trên một lưới. Ví dụ, một nhà quản lý dự án có thể sử dụng bảng nhiệt để theo dõi tiến độ của một dự án.

Việc lựa chọn phương pháp phân tích biểu đồ và đồ thị phù hợp sẽ phụ thuộc vào loại dữ liệu và mục tiêu phân tích. Tuy nhiên, việc hiểu các phương pháp phân tích này là cần thiết để có thể khai thác tối đa thông tin từ dữ liệu.

## 2 . Tiền xử lý dữ liệu

Import thư viện cần dùng trong phân tích dữ liệu

A screen shot of a computer program

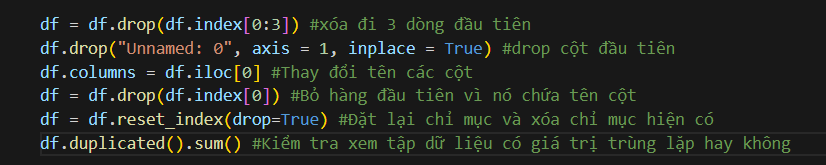
Description automatically generated

Tiến hành đọc file dữ liệu

A screenshot of a computer

Description automatically generated

Nhận thấy dữ liệu còn chứa những khoảng trống không có dữ liệu (dữ liệu chưa được sạch), chúng ta tiến hành làm sạch dữ liệu



A screenshot of a computer

Description automatically generated

Tiến hành kiểm tra kiểu dữ liệu của từng cột

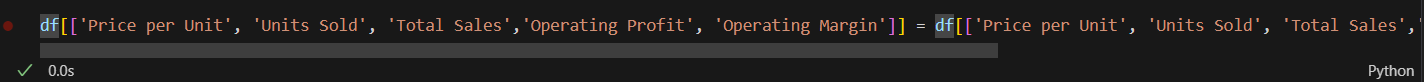
A screenshot of a computer

Description automatically generated

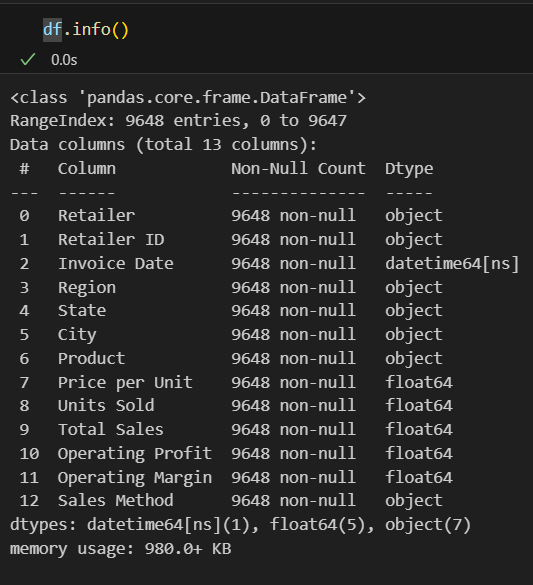
Thay đổi kiểu dữ liệu của cột “Invoice Date” thành kiểu Datetime



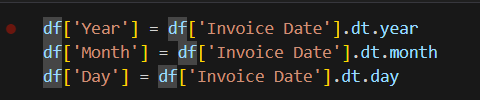
Thay đổi kiểu dữ liệu của những cột “Price per Unit”, “Units Sold”, “Total Sales”, “Operating Profit”, “Operating Margin” thành kiểu Float



Kiểu dữ liệu sau khi thay đổi



Tách cột “Invoice Date” ra thành 3 cột ngày, tháng, năm cho tiện trong việc Phân tích



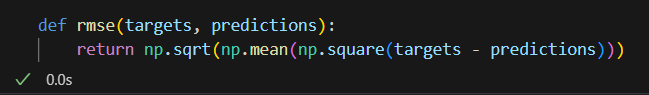
A screen shot of a black screen

Description automatically generated

## 3 . Machine learning và trực quan hóa dữ liệu

### 3.1. Machine learning

Sử dụng **Cây quyết định** để dự đoán lợi nhuận

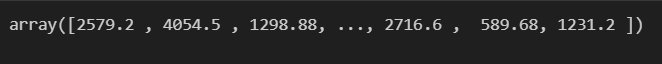


A black background with colorful text

Description automatically generated

A screen shot of a computer program

Description automatically generated



A screenshot of a computer code

Description automatically generated

A screen shot of a computer code

Description automatically generated

Phương thức model.score tính toán điểm R^2 , một đo lường cho việc dự đoán của mô hình so với giá trị thực tế, trên bộ kiểm tra (X\_test, y\_test). Kết quả sau đó được in ra.

Điểm R^2 = khoảng 0.998 cho thấy mô hình cây quyết định giải thích được khoảng 99.8% sự biến động của 'Total Sales' dựa trên các đặc trưng đã chọn.

Sử dụng **Rừng ngẫu nhiên** để dự đoán lợi nhuận

A black background with colorful text

Description automatically generated

A black screen with colorful text

Description automatically generated

A computer screen with text

Description automatically generated



A screenshot of a computer program

Description automatically generated

Sử dụng **hồi quy tuyến tính** để dự đoán lợi nhuận

A screen shot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

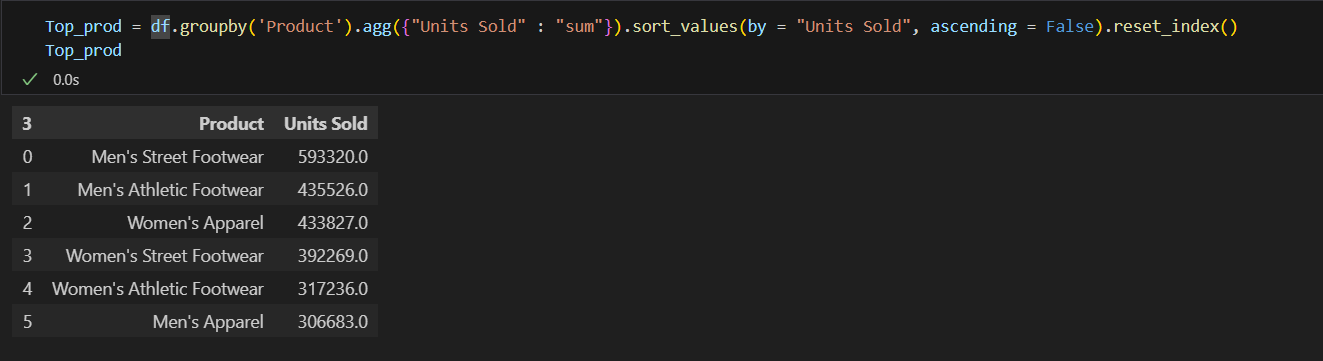
**So sánh** mức độ hiệu quả giữa các phương pháp phân tích

|  |  |  |
| --- | --- | --- |
| Phương pháp | R^2 Square | RMSE |
| Cây quyết định | 0,9986 | 1935,83 |
| Rừng ngẫu nhiên | 0,9988 | 1784,47 |
| Hồi quy tuyến tính | 0,9986 | 1935,83 |

Sau khi so sánh hiệu suất hoạt động và mức độ chênh lệch giữa giá trị thực tế và kết quả dự đoán giữa các phương pháp hồi quy, nhận thấy rằng mô hình **Rừng ngẫu nhiên** cho ra hiệu suất cao và mức độ chêch lệch thấp nhất, vậy nên trong trường hợp này ta nên sử dụng mô hình **Rừng ngẫu nhiên** để tiến hành phân tích dữ liệu bán hàng của doanh nghiệp.

### 3.2. Trực quan hóa dữ liệu

**Danh mục sản phẩm bán chạy nhất**

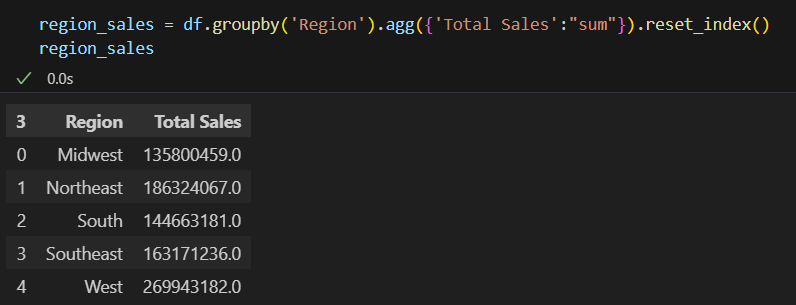


Biểu đồ trực quan

A graph of sales

Description automatically generated with medium confidence

**Tổng doanh thu theo khu vực**

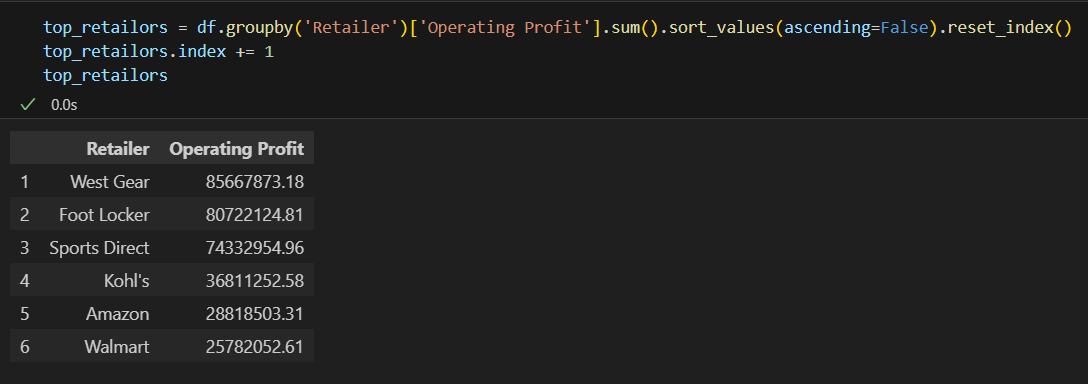


Biểu đồ

A colorful pie chart with different colored sections

Description automatically generated

**Top nhà bán lẻ có lợi nhuận cao nhất**

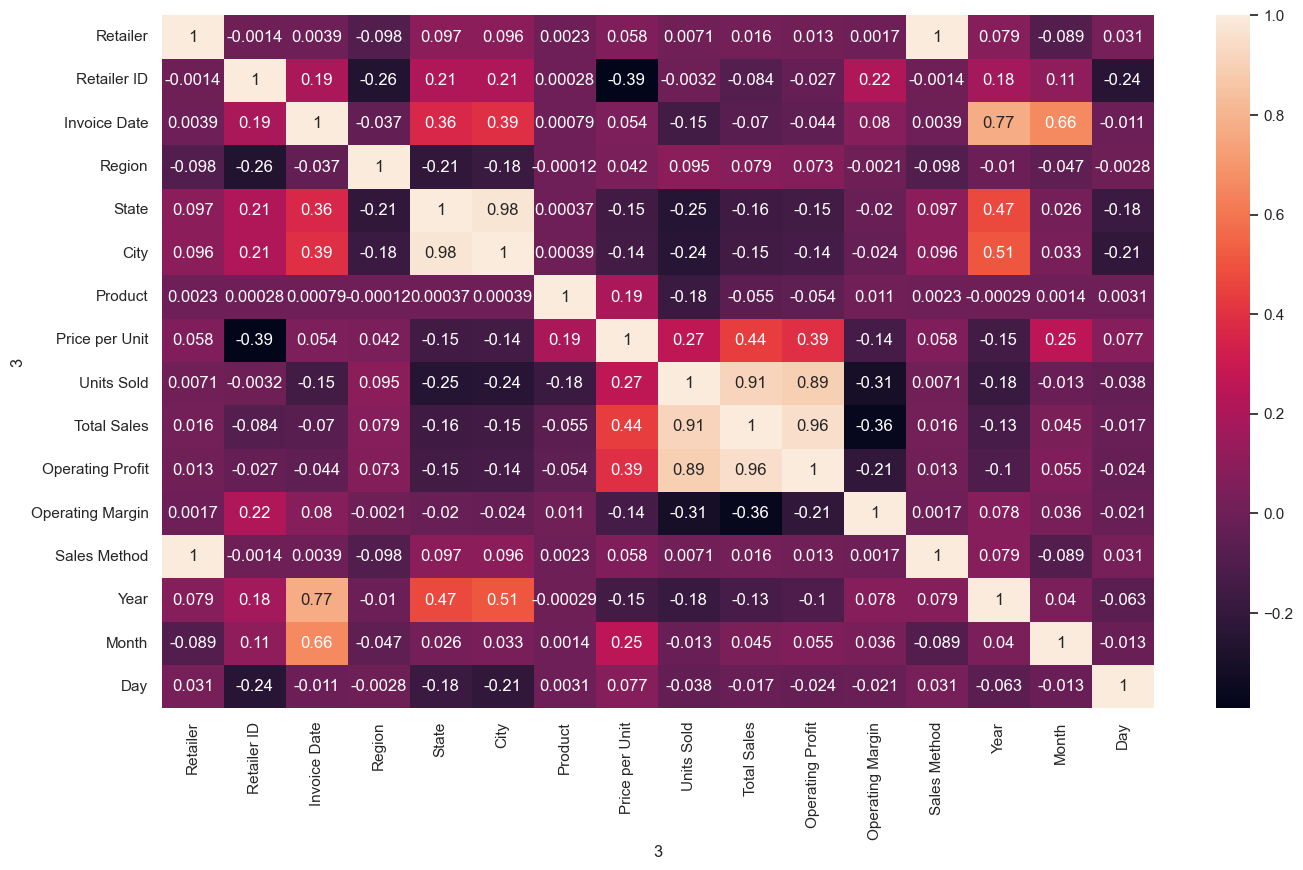


Biểu đồ

A graph of a bar chart

Description automatically generated with medium confidence

**Độ tương quan giữa các yếu tố**



# CHƯƠNG III . KẾT LUẬN

Phân tích dữ liệu bán hàng của doanh nghiệp là một đề tài có tính thực tế và ứng dụng cao. Phân tích dữ liệu giúp doanh nghiệp dự đoán tình hình thị trường trong tương lai. Qua việc nhận biết xu hướng và biến động, doanh nghiệp có thể thích nghi một cách linh hoạt và đưa ra những quyết định chiến lược.

Qua bài báo cáo này, nhóm đã hoàn thành mục tiêu đề ra là giải quyết những bài toán liên quan đến phân tích dữ liệu bán hàng của doanh nghiệp, giúp cho doanh nghiệp có thể thuận lợi hơn trong việc kiểm soát tình hình biến động của các sản phẩm được bày bán trên thị trường, từ đó đưa ra những quyết định chiến lược chính xác, hạn chế xác suất thất bại trong dự án. Có thể coi những phương pháp này là cần thiết và vô cùng quan trọng giúp nâng cao hiểu quả phân tích dữ liệu.

Bài báo cáo của nhóm em chỉ được phân tích dựa trên cơ sở lý thuyết đã được học tại trường lớp, chưa thông qua trải nghiệm thực tế, vì vậy đây chỉ là một bài báo cáo mang tính khách quan, mong nhận được những lời góp ý quý giá của Thầy để nhóm em có thể hoàn thiện hơn trong tương lai.