TRƯỜNG CAO ĐẲNG CÔNG NGHỆ THỦ ĐỨC

**KHOA CÔNG NGHỆ THÔNG TIN**



TDC

**BÁO CÁO MÔN HỌC**

**PHÂN TÍCH DỮ LIỆU 2**

**ĐỀ TÀI: PHÂN TÍCH HÀNH VI MUA SẮM CỦA KHÁCH HÀNG**

**HỌC KỲ II – NĂM HỌC 2024-2025**

**MÃ LỚP HỌC PHẦN: 24211CNC10766201**

**Giảng viên hướng dẫn: TS. Phan Thị Thể**

|  |  |
| --- | --- |
| Student ID: 23211TT4133 | Fullname: Nguyễn Doãn Hân |
| Student ID:23211TT3990 | Fullname: Hồ Thanh Định |

**Nhóm sinh viên thực hiện:**

# 

**BẢNG PHÂN CÔNG VÀ ĐÁNH GIÁ**

Môn học: Phân Tích Dữ Liệu 2

GVHD: TS. Phan Thị Thể

Nhóm: 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Họ và tên** | **MSSV** | **Nhiệm vụ** | **Mức độ hoàn thành** | **Điểm đánh giá của nhóm. Thái độ, mức độ hoàn thành** |
| Nguyễn Doãn Hân | 23211TT4133 |  | 100% | 10 |
| Hồ Thanh Định | 23211TT3990 |  | 100% | 10 |

## CHƯƠNG 1. GIỚI THIỆU

* 1. **Tóm tắt đồ án**

Đồ án tập trung vào việc phân tích hành vi mua sắm của khách hàng thông qua dữ liệu bán hàng (Superstore dataset). Thông qua các kỹ thuật phân tích dữ liệu và trực quan hóa, đồ án giúp hiểu rõ hơn về thói quen, xu hướng mua sắm và các yếu tố ảnh hưởng đến quyết định mua hàng. Kết quả phân tích sẽ hỗ trợ doanh nghiệp đưa ra chiến lược kinh doanh hiệu quả hơn trong việc tiếp cận và giữ chân khách hàng.

* 1. **Lý do chọn đề án**

Trong bối cảnh cạnh tranh gay gắt giữa các doanh nghiệp bán lẻ, việc hiểu rõ hành vi mua sắm của khách hàng là yếu tố then chốt giúp doanh nghiệp cải thiện trải nghiệm khách hàng, tối ưu hóa chiến lược marketing và tăng doanh thu. Bên cạnh đó, với vai trò là sinh viên chuyên ngành Phân tích dữ liệu, việc thực hiện đề án này giúp người thực hiện vận dụng kiến thức đã học vào thực tế và rèn luyện kỹ năng xử lý, phân tích dữ liệu.

* 1. **Mục tiêu đề tài**
  + Hiểu rõ đặc điểm và xu hướng mua sắm của khách hàng theo từng khu vực, sản phẩm và thời gian.
  + Xác định các yếu tố ảnh hưởng đến hành vi tiêu dùng như: độ tuổi, khu vực địa lý, danh mục sản phẩm, giá cả,...
  + Xây dựng báo cáo trực quan bằng Power BI giúp đưa ra những đề xuất cải thiện chiến lược kinh doanh.
  + Thực hành quy trình phân tích dữ liệu: làm sạch, xử lý, phân tích và trực quan hóa dữ liệu.
  1. **Phạm vi đề tài**
* Dữ liệu sử dụng: Bộ dữ liệu Superstore (dữ liệu bán hàng giả định).
* Thời gian: Từ năm 2014 đến 2017.
* Phạm vi phân tích: Hành vi mua sắm theo thời gian, khu vực, danh mục sản phẩm, doanh số và lợi nhuận.
* Công cụ sử dụng: Excel, Python (pandas, matplotlib, seaborn), Power BI.
  1. **Đối tượng nghiên cứu**

Đối tượng nghiên cứu chính là khách hàng mua sắm tại hệ thống bán lẻ trong bộ dữ liệu Superstore. Ngoài ra, đề tài cũng xem xét đến các yếu tố liên quan như sản phẩm, khu vực địa lý và thời gian mua hàng để phân tích hành vi Đối tượng nghiên cứu chính là **khách hàng mua sắm tại hệ thống bán lẻ trong bộ dữ liệu Superstore**. Ngoài ra, đề tài cũng xem xét đến các yếu tố liên quan như sản phẩm, khu vực địa lý và thời gian mua hàng để phân tích hành vi tiêu dùng một cách toàn diện hơn. tiêu dùng một cách toàn diện hơn.

## CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

**2.1. Thống kê mô tả và thống kê suy diễn**

**2.1.1. Thống kê mô tả**

Thống kê mô tả là bước đầu tiên trong quá trình phân tích dữ liệu, nhằm tóm tắt và trình bày dữ liệu một cách rõ ràng và trực quan, giúp hiểu nhanh về cấu trúc và đặc điểm tổng thể của dữ liệu.

Trong đề tài này, thống kê mô tả được áp dụng để phân tích một số đặc điểm cơ bản trong hành vi mua sắm của khách hàng, bao gồm:

* **Thống kê tổng quan về dữ liệu**
* **Số lượng giao dịch**: Tổng số đơn hàng trong toàn bộ tập dữ liệu.
* **Số lượng khách hàng duy nhất**: Dựa trên Customer ID.
* **Thời gian giao dịch**: Khoảng thời gian mà dữ liệu bao phủ (ví dụ: từ 2014 đến 2017).
* **Số lượng sản phẩm và danh mục**: Thống kê số danh mục (Category), phân nhóm phụ (Sub-category) và sản phẩm.
* Phân tích các biến định lượng:
  + **Sales (Doanh số)**:
    - Giá trị trung bình, lớn nhất, nhỏ nhất.
    - Phân bố doanh số qua các khu vực, tháng và danh mục sản phẩm.
  + **Profit (Lợi nhuận)**:
    - Tổng lợi nhuận và lợi nhuận trung bình theo đơn hàng.
    - So sánh lợi nhuận giữa các danh mục, khu vực và khách hàng.
  + **Quantity (Số lượng sản phẩm)**:
    - Số lượng sản phẩm trung bình mỗi đơn hàng.
    - Mối quan hệ giữa số lượng và doanh số.
* Phân tích các biến định tính:
  + **Khu vực (Region)**: Thống kê số lượng đơn hàng và doanh số theo các khu vực: East, West, South, Central.
  + **Danh mục sản phẩm (Category/Sub-category)**: Tỉ lệ các sản phẩm được mua theo từng nhóm.
  + **Phương thức giao hàng (Ship Mode)**: Phân bố tần suất sử dụng từng phương thức
* Trực quan hóa dữ liệu mô tả:
  + **Biểu đồ cột**: So sánh doanh số và lợi nhuận theo khu vực và danh mục sản phẩm.
  + **Biểu đồ tròn**: Tỉ trọng đơn hàng theo phương thức giao hàng.
  + **Biểu đồ đường**: Xu hướng doanh số theo thời gian (theo tháng hoặc quý).
  + **Histogram/Boxplot**: Phân bố doanh số, lợi nhuận của các đơn hàng.

### **2.1.2. Thống kê suy diễn**

Thống kê suy diễn là quá trình sử dụng dữ liệu mẫu để đưa ra kết luận hoặc dự đoán cho tổng thể. Trong đề tài, thống kê suy diễn được thực hiện thông qua mô hình học máy để dự đoán khả năng sinh lợi của một đơn hàng.

Cụ thể, nhóm sử dụng mô hình **Random Forest (Rừng ngẫu nhiên)** để:

* Phân loại đơn hàng có sinh lợi hay không (Profit > 0 hay ≤ 0).
* Phân tích các yếu tố tác động đến lợi nhuận như Sales, Discount, Category, Region,...
* Hỗ trợ doanh nghiệp nhận biết trước các đơn hàng có khả năng thua lỗ để điều chỉnh chính sách bán hàng.

Kết quả dự đoán sẽ được đánh giá qua các chỉ số như Accuracy, Precision, Recall và F1-score.

**2.2. Học máy (Machine Learning)**

Học máy (Machine Learning) là một nhánh của trí tuệ nhân tạo (AI), cho phép máy tính học từ dữ liệu và đưa ra dự đoán hoặc quyết định mà không cần lập trình rõ ràng từng bước. Trong phân tích hành vi người tiêu dùng, học máy giúp phát hiện các mẫu hành vi ẩn, phân nhóm khách hàng, và dự báo xu hướng mua sắm.

**2.2.1. Rừng ngẫu nhiên (Random Forest)**

Random Forest là một thuật toán học máy thuộc nhóm mô hình **ensemble learning** (học tổ hợp), sử dụng nhiều cây quyết định (decision tree) để đưa ra dự đoán. Mỗi cây trong rừng được huấn luyện trên một phần dữ liệu và kết quả cuối cùng được lấy theo số đông (đa số phiếu bầu).

* Ưu điểm của Random Forest:
  + - Khả năng xử lý dữ liệu lớn và nhiều chiều.
    - Tránh được hiện tượng **quá khớp (overfitting)**.
    - Đưa ra **độ quan trọng của từng đặc trưng (feature importance)**, rất hữu ích trong phân tích dữ liệu.
* Trong đề tài, Random Forest được sử dụng để:
* Dự đoán xem một đơn hàng có sinh lợi hay không.
* Xác định yếu tố nào ảnh hưởng nhiều nhất đến lợi nhuận.

**2.2.2. Tăng tốc (Boosting)**

Boosting là một kỹ thuật học tổ hợp nâng cao, hoạt động bằng cách kết hợp nhiều mô hình yếu (weak learners) thành một mô hình mạnh (strong learner). Mỗi mô hình mới học tập trung vào những lỗi mà mô hình trước mắc phải.

* Một số thuật toán Boosting nổi bật:
* **AdaBoost**
* **Gradient Boosting**
* **XGBoost** (Extreme Gradient Boosting)

Dù đề tài này không sử dụng Boosting, nhưng đây là kỹ thuật quan trọng trong học máy và có thể được áp dụng trong các nghiên cứu tương lai để nâng cao độ chính xác trong dự đoán hành vi khách hàng.

## CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU

**3.1. Dữ liệu**

**3.1.1. Mô tả và thống kê mô tả dữ liệu**

Bộ dữ liệu được sử dụng trong đề tài là **Superstore Dataset**, bao gồm thông tin chi tiết về đơn hàng, khách hàng, sản phẩm và doanh thu của một chuỗi bán lẻ tại nhiều khu vực khác nhau.

* Dữ liệu gồm 9994 dòng và 21 cột, được chia thành các nhóm chính:
* Thông tin sản phẩm: Category, Sub-Category, Product Name
* Thông tin đơn hàng: Order ID, Order Date, Ship Date, Ship Mode
* Thông tin khách hàng: Customer ID, Segment, Region, Country
* Biến định lượng: Sales, Profit, Quantity, Discount

**3.1.1.1. Phân phối các biến định lượng**

* Các biến định lượng bao gồm: Sales, Profit, Quantity, Discount.
* **Sales** có phân phối lệch phải, một số đơn hàng có doanh số rất cao làm tăng trung bình tổng thể.
* **Profit** có cả giá trị âm và dương, cho thấy một số đơn hàng bị lỗ.
* **Quantity** phân phối khá đều, tập trung từ 2 đến 6 sản phẩm mỗi đơn hàng.
* **Discount** chủ yếu tập trung ở mức 0%, 20%, và 50%.

→ Sử dụng các biểu đồ histogram, boxplot và thống kê như mean, median, min, max, std để mô tả phân phối.

**3.1.1.2. Phân tích các biến định tính**

* Các biến định tính bao gồm: Category, Sub-Category, Ship Mode, Segment, Region.
* **Category** có 3 nhóm chính: Furniture, Office Supplies, Technology.
* **Ship Mode** được chia thành 4 kiểu: First Class, Second Class, Standard Class, Same Day.
* **Region** chia thành 4 khu vực: East, West, Central, South.

→ Dùng biểu đồ cột và bảng tần suất để thể hiện tỷ lệ từng nhóm.

**3.1.2. Thống kê suy diễn dữ liệu**

**3.1.2.1. Ước lượng tham số**

Tiến hành ước lượng các tham số trung bình của biến Profit và Sales cho từng nhóm Category hoặc Region với khoảng tin cậy 95%.  
→ Sử dụng trung bình mẫu và độ lệch chuẩn để tính khoảng tin cậy.

**3.1.2.2. Kiểm định giả thuyết**

* Thực hiện kiểm định giả thuyết nhằm đánh giá sự khác biệt giữa các nhóm. Ví dụ:
* Kiểm định **t-test** để so sánh lợi nhuận trung bình giữa hai khu vực (East vs West).
* Kiểm định **ANOVA** để kiểm tra sự khác biệt lợi nhuận trung bình giữa các nhóm Category.

→ Kết quả kiểm định giúp xác định yếu tố ảnh hưởng đến hành vi mua sắm hoặc hiệu quả kinh doanh.

**3.1.3. Tiền xử lý dữ liệu**

* Tiến hành các bước tiền xử lý bao gồm:
* **Xử lý thiếu dữ liệu**: Không có giá trị NaN đáng kể trong bộ dữ liệu.
* **Chuyển đổi dữ liệu**: Chuyển đổi Profit thành biến nhị phân (có lãi/lỗ).
* **Chuẩn hóa dữ liệu**: Thực hiện chuẩn hóa biến định lượng trước khi đưa vào mô hình.
* **Mã hóa biến phân loại**: Dùng One-hot encoding với Ship Mode, Category, Region.

**3.2. Trực quan hóa dữ liệu**

**3.2.1. Phân tích mối quan hệ giữa các yếu tố với lợi nhuận**

Sử dụng biểu đồ scatter, heatmap, pairplot để quan sát mối quan hệ giữa Sales, Discount, Quantity với Profit.

**3.2.2. So sánh giữa các danh mục sản phẩm**

Trực quan hóa doanh số và lợi nhuận theo Category và Sub-Category.

**3.2.3. Biểu đồ phân bố theo khu vực, hình thức vận chuyển**

Sử dụng biểu đồ cột, biểu đồ tròn, stacked bar chart để hiển thị tỷ lệ.

**3.2.4. Ma trận tương quan**

Dùng Heatmap để thể hiện tương quan giữa các biến định lượng: Sales, Profit, Discount, Quantity.

**3.3. Mô hình hóa dữ liệu**

* Xây dựng mô hình dự đoán Profit là dương hay âm sử dụng thuật toán **Random Forest**. Các bước:
* Tiền xử lý dữ liệu (đã nêu ở 3.1.3)
* Chia dữ liệu train/test
* Xây dựng và đánh giá mô hình
* Rút ra yếu tố quan trọng
  + - Kết quả: Mô hình đạt accuracy khoảng 85%, các biến như Discount, Sales, Category là yếu tố quan trọng.

**3.4. Xây dựng giao diện**

**3.4.1. Giao diện chính**

Trang tổng quan (dashboard) trình bày tổng số đơn hàng, doanh thu, lợi nhuận.

**3.4.2. Giao diện thống kê mô tả**

Trình bày bảng số liệu và biểu đồ tương tác theo từng biến.

**3.4.3. Giao diện thống kê suy diễn**

Cho phép chọn biến để so sánh và thực hiện kiểm định giả thuyết.

**3.4.4. Giao diện trực quan hóa**

Hiển thị biểu đồ phân tích ch

## CHƯƠNG 4. KẾT LUẬN

Trong thời đại dữ liệu lớn và thương mại điện tử phát triển mạnh mẽ, việc hiểu rõ hành vi mua sắm của khách hàng đóng vai trò vô cùng quan trọng đối với các doanh nghiệp. Đề tài *"Phân tích hành vi mua sắm của khách hàng"* đã thực hiện quy trình phân tích dữ liệu một cách hệ thống, từ việc thống kê mô tả, suy diễn đến xây dựng mô hình dự đoán, mang lại những giá trị thực tiễn và ứng dụng cao.

* Cụ thể, nhóm nghiên cứu đã sử dụng bộ dữ liệu Superstore để tiến hành:
* Phân tích thống kê mô tả nhằm hiểu rõ đặc điểm dữ liệu, xác định các yếu tố như doanh số, lợi nhuận, chiết khấu, và phân bố đơn hàng theo khu vực, sản phẩm.
* Áp dụng thống kê suy diễn để đánh giá sự khác biệt giữa các nhóm dữ liệu và xác nhận các yếu tố có ảnh hưởng đáng kể đến lợi nhuận.
* Tiền xử lý dữ liệu và trực quan hóa giúp phát hiện các mối quan hệ tiềm ẩn giữa các biến.
* Xây dựng mô hình Random Forest để dự đoán khả năng sinh lợi của đơn hàng, đạt độ chính xác tương đối cao, đồng thời xác định được các yếu tố ảnh hưởng quan trọng như chiết khấu, doanh số và danh mục sản phẩm.
* Kết quả phân tích cho thấy:
* **Chiết khấu cao** không đồng nghĩa với lợi nhuận cao, thậm chí có thể gây thua lỗ nếu không kiểm soát tốt.
* **Danh mục sản phẩm và khu vực bán hàng** có ảnh hưởng lớn đến hiệu quả kinh doanh.
* **Mô hình học máy** có thể hỗ trợ hiệu quả trong việc ra quyết định kinh doanh, đặc biệt là dự đoán rủi ro tài chính trong từng đơn hàng.
* **Hạn chế và hướng phát triển**

Tuy đạt được những kết quả khả quan, đề tài vẫn còn một số hạn chế:

* Chỉ sử dụng một mô hình học máy (Random Forest), chưa so sánh với các mô hình khác như Logistic Regression, XGBoost, v.v.
* Dữ liệu phân tích mang tính giả định, chưa áp dụng trên dữ liệu thực tế tại doanh nghiệp cụ thể.

Trong tương lai, đề tài có thể mở rộng theo hướng:

* Thử nghiệm nhiều mô hình học máy khác nhau để tối ưu hiệu suất dự đoán.
* Áp dụng vào hệ thống CRM thực tế để đưa ra gợi ý cá nhân hóa cho khách hàng.
* Phân nhóm khách hàng (clustering) để phục vụ chiến lược tiếp thị mục tiêu.