**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙡🕮🙣**

Ảnh có chứa văn bản, mẫu họa

Mô tả được tạo tự động**MÔN HỌC: KHAI PHÁ DỮ LIỆU**

**BÁO CÁO ĐỒ ÁN**

**ĐỀ TÀI**

**THỰC HIỆN CÁC KỸ THUẬT KHAI PHÁ DỮ LIỆU**

**GVHD:** Th.s Trần Trọng Bình

**Mã lớp:** **DAMI330484\_23\_2\_01**

**Nhóm sinh viên thực hiện: Nhóm 5**

1. Huỳnh Nhật Nam – 21110903
2. Nguyễn Thị Lan Anh – 21110368
3. Trịnh Thị Thanh Huyền – 21110477
4. Phan Minh Thuận – 21110667

**TP. Hồ Chí Minh, tháng 05 năm 2024**

**PHẦN 1: MÔ TẢ VỀ TẬP DỮ LIỆU**

1. **Mô tả về tập dữ liệu**

Link dataset: <https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting/code>

Tập dữ liệu bán lẻ của một siêu thị toàn cầu trong 4 năm

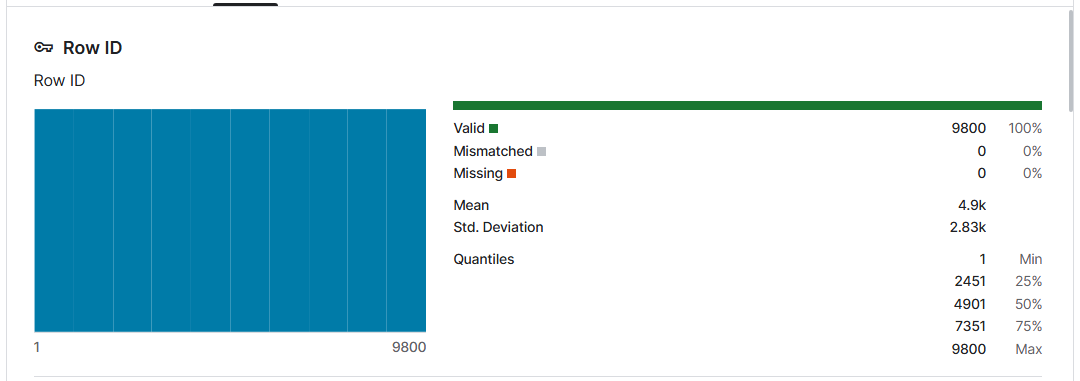
Số lượng dòng : 9800

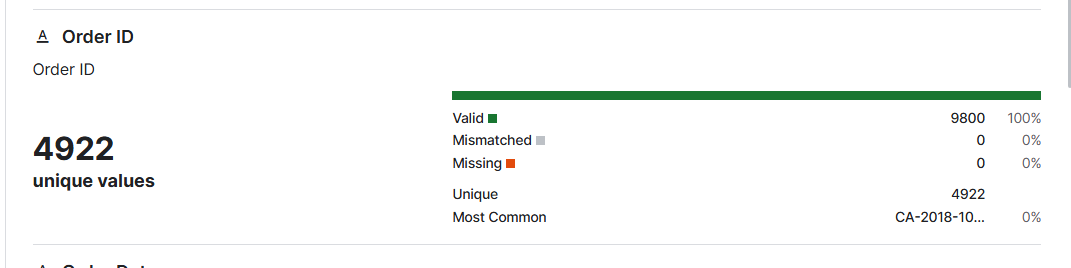
Số lượng cột: 18

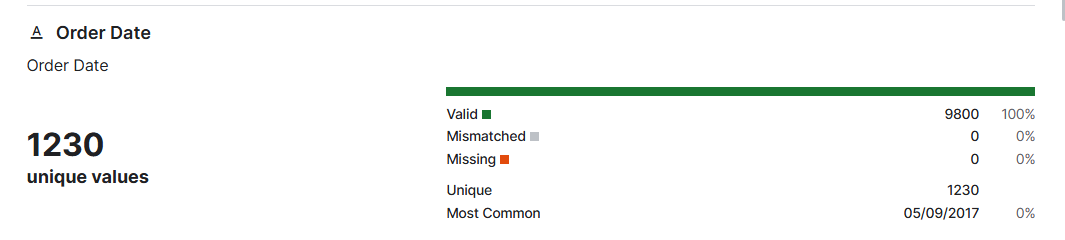
Tập dữ liệu bao gồm 18 cột:

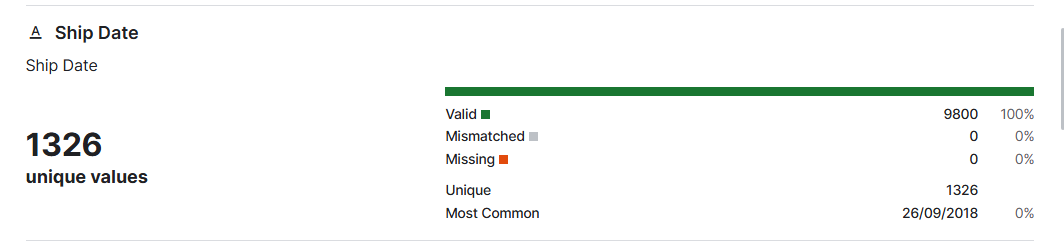
* Row ID : Mã Duy Nhất cho mỗi dòng
* Order ID: Mã Đơn Hàng Duy Nhất cho mỗi Khách Hàng
* Order Date: Ngày Đặt Hàng của sản phẩm
* Ship Date: Ngày Giao Hàng của sản phẩm
* Ship Mode: Phương Thức Giao Hàng do Khách Hàng chỉ định
* Customer ID: Mã Duy Nhất để xác định mỗi Khách Hàng
* Customer Name: Tên của Khách Hàng
* Segment: Phân khúc nơi Khách Hàng thuộc về
* Country: Quốc gia cư trú của Khách Hàng
* City: Thành phố cư trú của Khách Hàng
* State: Bang cư trú của Khách Hàng
* Postal Code: Mã Bưu Chính của mỗi Khách Hàng
* Region: Khu vực nơi Khách Hàng thuộc về
* Product ID: Mã Duy Nhất của sản phẩm
* Category: Danh mục của sản phẩm được đặt hàng
* Sub-Category: Danh mục phụ của sản phẩm được đặt hàng
* Product Name: Tên của sản phẩm
* Sales: Doanh số của sản phẩm

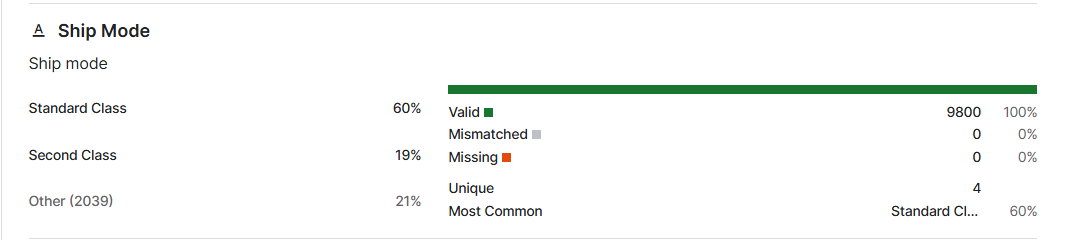
1. **Phân tích tập dữ liệu**

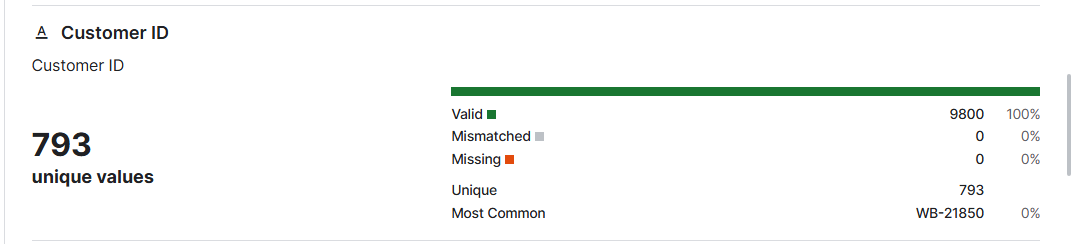


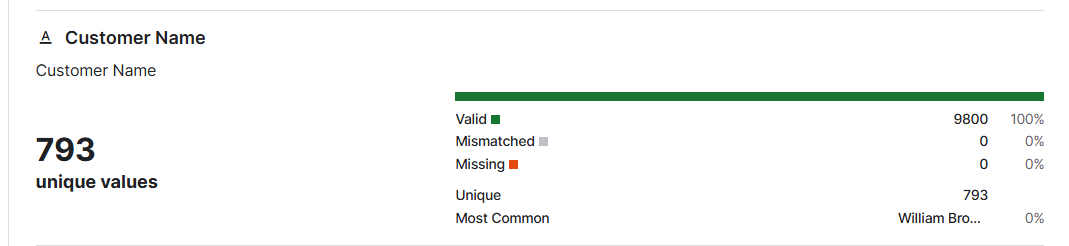


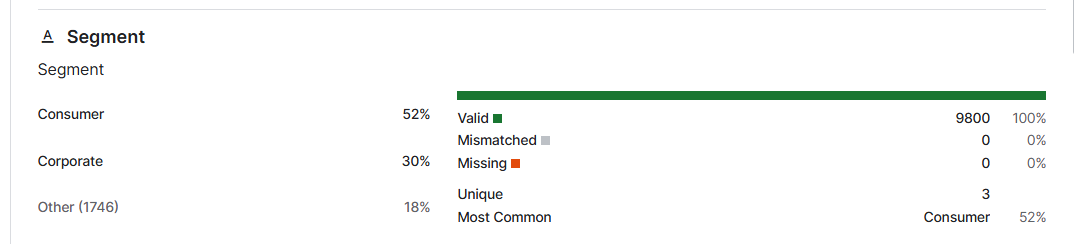


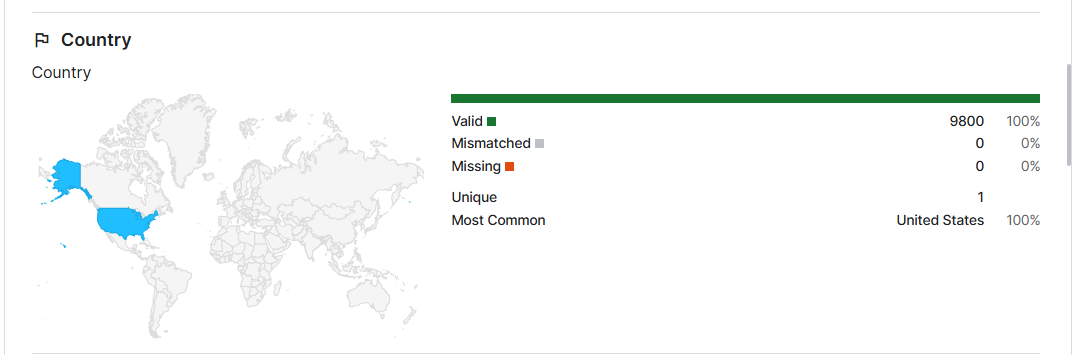


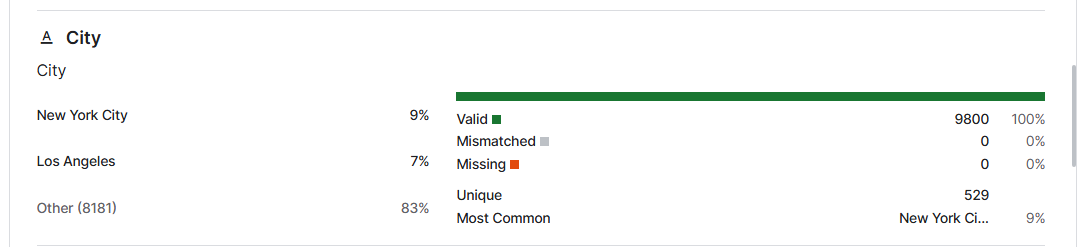


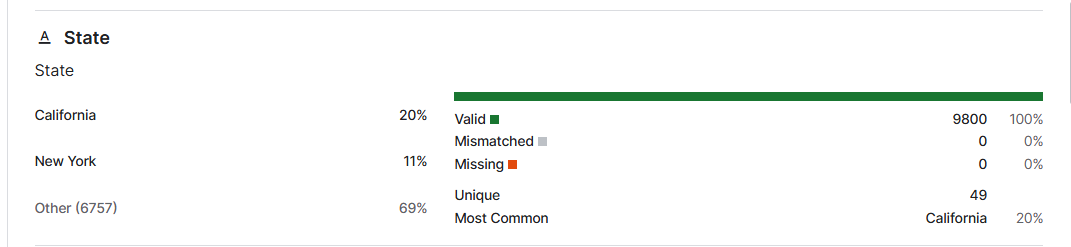


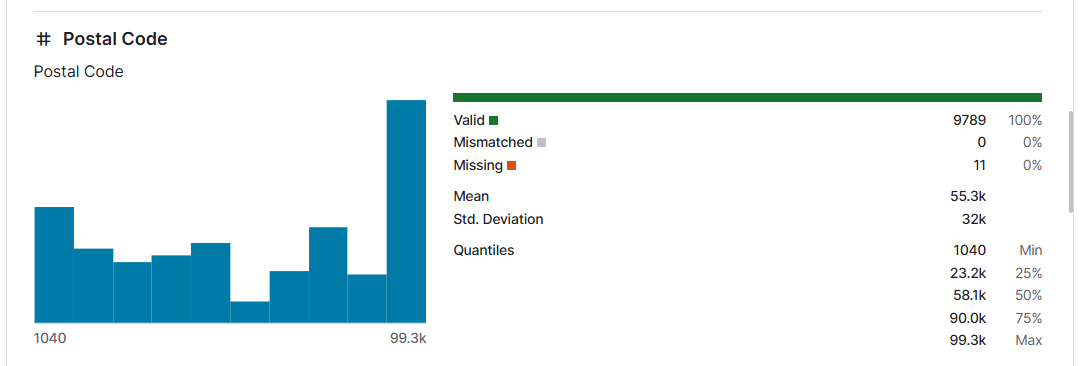


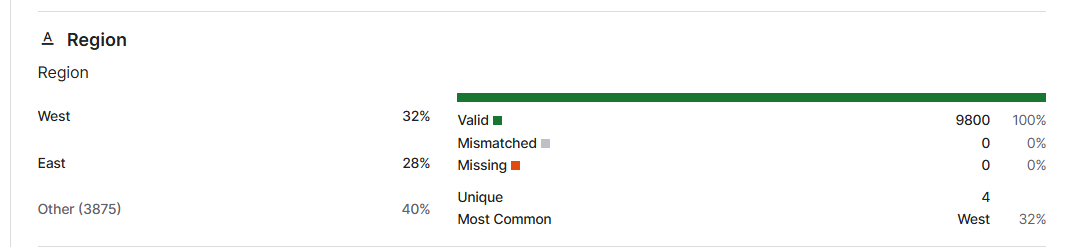


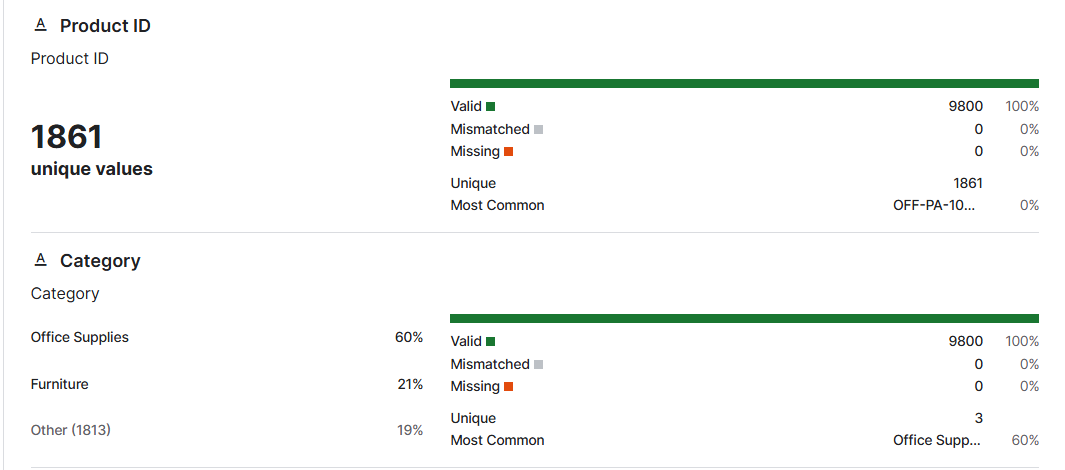


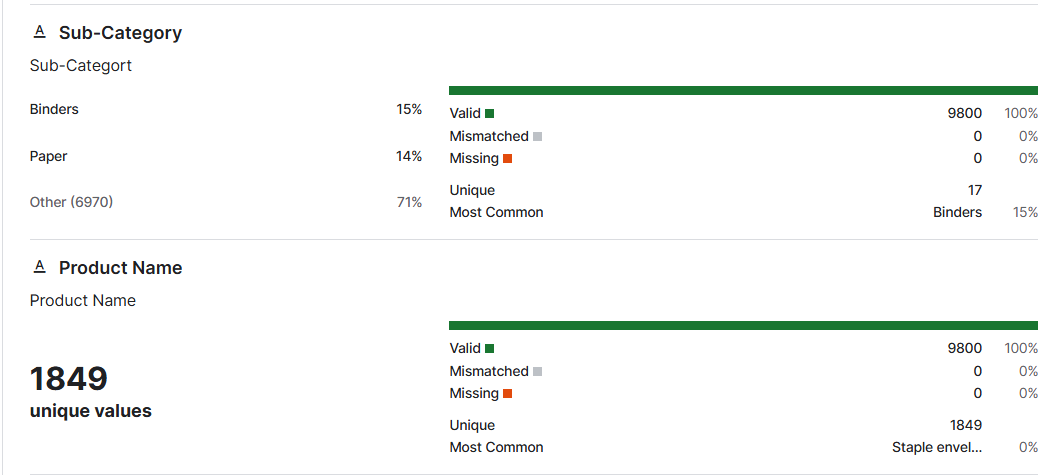


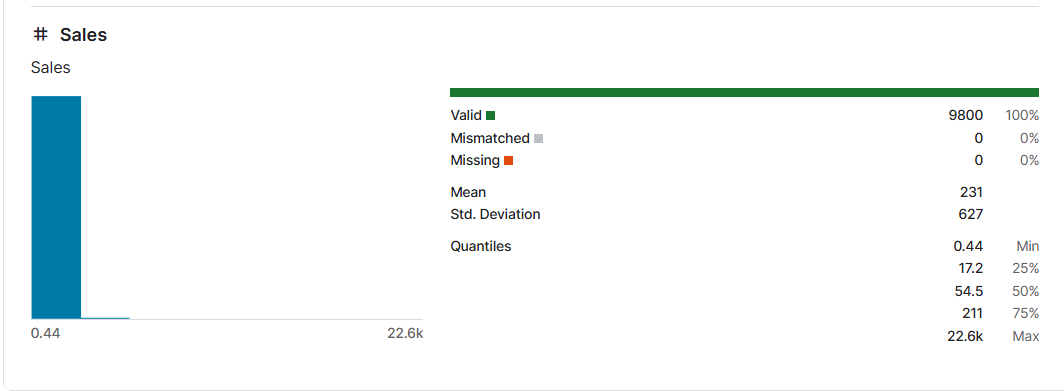












=> Dữ liệu đa phần đầy đủ, không bị trường hợp missing data, chỉ có cột Postal Code

có 11 hàng không có giá trị

**PHẦN 2: CÁC KỸ THUẬT KHAI PHÁ DỮ LIỆU**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Thuật toán | Ưu điểm | Nhược điểm | Ý nghĩa |
| Gom cụm | K-means | - Tốc độ tính toán nhanh  - Dễ hiểu dễ cài đặt  - Cluster có số lượng điểm tương đối bằng nhau | - Phải chỉ ra số lượng cluster  - Có một số mẫu nằm gần biên cụm | - Bằng cách phân cụm khách hàng dựa trên hành vi mua hàng, bao gồm ba tiêu chí: Recency (Thời gian gần nhất mua hàng), Frequency (Tần suất mua hàng), Monetary (Giá trị tiền mỗi lần mua hàng), đã phân khúc khách hàng thành ba thành cụm chính 'Loyal' (Khách hàng trung thành) là Những khách hàng này không chỉ mua sắm thường xuyên mà còn chi tiêu nhiều và thường quay lại mua hàng trong khoảng thời gian gần đây. 'Casual' (Khách hàng không thường xuyên) là Những khách hàng này mua hàng không quá thường xuyên và không chi tiêu nhiều như khách hàng trung thành. 'At-Risk' (Khách hàng có rủi ro) là Những khách hàng này đã không mua hàng trong một thời gian dài, ít khi mua hàng và chi tiêu ít. Thông qua phương pháp phân cụm bao gồm K-means, Hierarchical Clustering, DBSCAN giúp phát hiện các nhóm khách hàng có đặc điểm tương tự nhau một cách chi tiết. Từ đó đưa ra chiến lược kinh doanh cụ thể: **a. Khách hàng trung thành (Loyal Customers)** **- Đặc điểm**: Có giá trị Recency thấp, Frequency và Monetary cao.  - **Mục tiêu**: Duy trì và tăng cường sự trung thành của khách hàng, thúc đẩy họ tiếp tục mua sắm và trở thành đại sứ thương hiệu.  **- Chiến lược**: duy trì recency ở mức thấp, bằng cách:  **Ưu đãi độc quyền**: Cung cấp các ưu đãi chỉ dành riêng cho nhóm này như giảm giá đặc biệt, chương trình tích điểm đổi quà.  **Gửi email cảm ơn**: Tạo cảm giác được quan tâm và đánh giá cao bằng cách gửi email cảm ơn sau mỗi giao dịch.  **Voucher chất lượng và quà tặng tri ân**: Gửi tặng các voucher giá trị hoặc quà tặng để tri ân sự trung thành của họ. **b. Khách hàng bình thường (Casual Customers)** **- Đặc điểm**: Có giá trị Frequency và Monetary trung bình, Recency trung bình hoặc cao.  **Mục tiêu:** Chuyển đổi khách hàng thường xuyên thành khách hàng trung thành bằng cách tăng cường tần suất mua hàng và sự gắn kết với thương hiệu.  **- Chiến lược**: làm giảm recency, tăng cường frequency và Monetary bằng cách sau:  **Chương trình khuyến mãi đặc biệt**: Tổ chức các chương trình khuyến mãi hấp dẫn để kích thích họ mua sắm nhiều hơn.  **Tăng cường tương tác**: Gửi thông tin về sản phẩm mới, ưu đãi đặc biệt qua email hoặc SMS để duy trì sự quan tâm của họ.  **Chương trình khách hàng thân thiết**: Tạo chương trình tích điểm hoặc nhận thưởng khi mua hàng nhiều hơn để tăng tần suất mua hàng. **c. Khách hàng rủi ro (At-Risk Customers)** **- Đặc điểm**: Có giá trị Recency cao, Frequency và Monetary cao hoặc trung bình.  - **Mục tiêu:** Thu hút lại sự quan tâm của khách hàng, và khuyến khích họ quay lại mua sắm.  **- Chiến lược**: Làm cho recency giảm và làm tăng frequency và Monetary bằng cách sau:  **Gửi email nhắc nhở**: Gửi email nhắc nhở về các sản phẩm yêu thích hoặc khuyến mãi để thu hút họ quay lại mua hàng.  **Khuyến mãi đặc biệt để quay lại**: Cung cấp các khuyến mãi đặc biệt, giảm giá sâu để thu hút họ quay lại.  **Khảo sát ý kiến**: Gửi khảo sát để hiểu lý do họ không quay lại mua hàng, từ đó cải thiện dịch vụ và sản phẩm.  Qua đó không chỉ giúp hiểu rõ hơn về khách hàng mà còn hỗ trợ trong việc xây dựng các chiến lược tiếp thị và chăm sóc khách hàng phù hợp, từ đó tối ưu hóa doanh thu và sự hài lòng của khách hàng. Việc áp dụng các biện pháp chiến lược phù hợp cho từng nhóm khách hàng sẽ giúp doanh nghiệp duy trì và phát triển mối quan hệ bền vững với khách hàng, tăng cường lòng trung thành và tối đa hóa giá trị lâu dài. |
| Hierarchical Clustering | - Không cần xác định số lượng cụm trước.  - Tạo ra biểu đồ cây thể hiện cấu trúc phân cụm dữ liệu  -Phát hiện được cụm có hình dạng không đều, không phụ thuộc vào kích thước hoặc hình dạng cụm  - Thể hiện mức độ tương đồng giữa các các điểm dữ liệu trong cụm | - Tính toán phức tạp  - Không thể điều chỉnh sau khi tạo cụm, muốn chỉnh thì phải chạy lại từ đầu.  - Khó khăn trong việc xác định số cụm tối ưu |
| DBSCAN | - Không yêu cầu số cụm ban đầu  - Xác định cụm có hình dạng bất kỳ  - Xử lý tốt nhiễu  - Khả năng mở rộng tốt với tập dữ liệu lớn | - Lựa chọn tham số nhạy cảm  - Khả năng mở rộng hạn chế với dữ liệu cao chiều |
| Khai phá luật kết hợp | Apriori | - Dễ hiểu và triển khai  - Hiệu quả  - Tìm kiếm không bắt buộc toàn bộ không gian tìm kiếm | - Duyệt CSDL nhiều lần.  - Tập ứng viên sinh ra rất lớn  - Việc tính độ phổ biến nhiều. | - Với những phân tích trên, thì thấy binders là danh mục sản phẩm phổ biến nhất, bán được nhiều nhất  - Khai phá luật kết hợp sẽ giúp nhà bán hàng xác định mối quan hệ giữa các sản phẩm, hiểu rõ hơn hành vi mua sắm của khách hàng, để thiết lập các chiến dịch khuyến mãi, thiết kế các combo bán hàng cùng nhau cho hiệu quả, giúp tăng lợi nhuận, đáp ứng nhu cầu của khách hàng  - Ví dụ: (Storage, Appliances) -> (Binders) , sau khi biết được mối quan hệ giữa (Storage, Appliances) và (Binders), khi khách hàng mua (Storage, Appliances) thì khả năng cao họ cũng sẽ mua (Binders) thì nhà bán hàng sẽ có chiến lược như là sẽ giảm giá binders khi mua kèm với Storage, Appliances , hoặc là bố trí vị trí các danh mục sản phẩm này ở gần nhau. |
| FP - Growth | - Sử dụng bộ nhớ ít hơn  - Không cần quét lại cơ sở dữ liệu nhiều lần  - Mặc dù Apriori nhanh hơn trong trường hợp này, sự chênh lệch về thời gian không quá lớn. Tuy nhiên, sự khác biệt về bộ nhớ sử dụng là rất đáng kể. FP-Growth sử dụng ít bộ nhớ hơn gấp nhiều lần so với Apriori. Đơn hàng ngày một tăng, xử lý dữ liệu ngày càng lớn thì FP-Growth sẽ nhanh hơn cả về thời gian, vì nó có ưu điểm giảm số lần quét cơ sở dữ liệu, khắc phục được nhược điểm của Apriori khi phải cần nhiều lần quét cơ sở dữ liệu để tạo ra các tập con của các mục có kích thước tăng dần  - Đề xuất: Dựa trên kết quả và phân tích, cửa hàng nên sử dụng FP-Growth vì lợi ích vượt trội về bộ nhớ, thời gian thực thi |  |
|  | RNN | -RNN dễ hiểu và triển khai hơn so với LSTM và GRU.  -Đòi hỏi ít tài nguyên tính toán và bộ nhớ hơn, nên thời gian huấn luyện và suy luận nhanh hơn. | RNN gặp khó khăn trong việc học các phụ thuộc dài hạn do vấn đề gradient biến mất, làm cho chúng kém hiệu quả với chuỗi này, vì để dự đoán giá trị hiện tại phải cần tới 25 giá trị trước đó. Có hiệu suất kém hơn so với LSTM và GRU | Bởi vì dữ liệu mang tính tuần tự theo thời gian nên việc áp dụng 3 thuật toán này giúp việc dự đoán doanh thu trong tương lai được chính xác hơn.  Khả năng ghi nhớ thông tin từ các bước thời gian trước giúp các mô hình này dự đoán doanh thu dựa trên các yếu tố lịch sử, như các sự kiện đặc biệt(Black Friday), mùa vụ(thấp vào những tháng đầu năm và cao những tháng cuối năm), các xu hướng kinh doanh dài hạn.  Doanh nghiệp có thể sử dụng các dự đoán doanh thu để lập kế hoạch kinh doanh, bao gồm quản lý tồn kho, lên lịch nhân viên, và lập kế hoạch marketing: doanh thu thấp sẽ cắt bớt nhân lực, giảm số lượng sản phẩm tồn kho, doanh thu cao sẽ tăng số lượng sản phẩm tồn kho, tăng nhân lực  Các dự đoán chính xác giúp doanh nghiệp quản lý dòng tiền và các nguồn lực tài chính hiệu quả hơn, giảm thiểu rủi ro tài chính.  Dựa trên các dự đoán doanh thu, doanh nghiệp có thể tối ưu hóa chiến lược tiếp thị để tăng doanh số vào các thời điểm có nhu cầu cao hoặc giảm chi phí vào các thời điểm dự đoán doanh thu thấp.  Các dự đoán doanh thu giúp doanh nghiệp phân tích hiệu quả kinh doanh qua các thời kỳ khác nhau và xác định các yếu tố ảnh hưởng đến doanh thu. |
| LSTM | LSTM được thiết kế để ghi nhớ các phụ thuộc dài hạn, giúp chúng hiệu quả với chuỗi này vì có nhiều thông tin quá khứ quan trọng.  Bao gồm các cơ chế (cổng quên, cổng vào và cổng ra) để kiểm soát thông tin nào được nhớ hoặc quên. | LSTM phức tạp hơn RNN, làm cho chúng khó hiểu và triển khai hơn.  Đòi hỏi nhiều tài nguyên tính toán và bộ nhớ hơn, dẫn đến thời gian huấn luyện và suy luận lâu hơn. |
| GRU | GRU đơn giản hơn LSTM vì sử dụng ít cổng hơn (cổng đặt lại và cổng cập nhật), làm cho chúng nhanh hơn trong huấn luyện và suy luận.  Có hiệu suất tương đương với LSTM Có chút nhỉnh hơn | Tương tự như LSTM |
|  | Time series analysis | **Dự đoán tương lai:** Phân tích chuỗi thời gian giúp dự đoán xu hướng và biến động trong tương lai, từ đó giúp cửa hàng lập kế hoạch và đưa ra quyết định có tính chiến lược.  **Phát hiện mẫu và xu hướng:** Bằng cách phân tích chuỗi thời gian, chúng ta đã phát hiện các mẫu, chu kỳ, và xu hướng trong dữ liệu này. Giúp hiểu rõ hơn về cơ chế hoạt động của các quy trình kinh doanh  **Quản lý rủi ro:** Phân tích chuỗi thời gian cho phép dự đoán các biến động và rủi ro trong tương lai, giúp doanh nghiệp chuẩn bị và phản ứng kịp thời.  **Tối ưu hóa quy trình sản xuất và cung ứng:** Bằng cách dự đoán nhu cầu và nguồn cung trong tương lai, các doanh nghiệp tối ưu hóa quy trình sản xuất và cung ứng, từ đó giảm thiểu lãng phí và tăng hiệu suất. | Dữ liệu ngày bị missing 228 giá trị, do đó phải đi tiền xử lý để đi phân tích và dự đoán trên chuỗi này  Một số mô hình chuỗi thời gian như SARIMA khá phức tạp về mặt toán học và đòi hỏi nhiều tiền xử lý dữ liệu  Mặc dù phân tích chuỗi thời gian dự đoán được các mẫu và xu hướng, nhưng nó gặp khó khăn trong việc dự đoán các biến động không xác định hoặc các sự kiện bất ngờ. | Phân tích chuỗi thời gian cung cấp thông tin cần thiết để cải thiện dự đoán và quyết định kinh doanh, từ việc dự đoán doanh thu tới quản lý rủi ro và lập kế hoạch sản xuất: tăng số lượng sản phẩm trong kho, tăng số lượng nhân viên nếu doanh thu cao, giảm số lượng sản phẩm tồn kho giảm số lượng nhân viên khi doanh thu thấp  Hiểu rõ hơn về mẫu và xu hướng trong dữ liệu chuỗi thời gian giúp tối ưu hóa quy trình sản xuất, cung ứng và tiếp thị, từ đó tăng hiệu suất và giảm lãng phí: ví dụ như biết được doanh thu sẽ thấp vào những tháng đầu năm và cao vào những tháng cuối năm đặc biệt là tháng 11  Dựa trên những thông tin từ phân tích chuỗi thời gian, doanh nghiệp phát triển chiến lược kinh doanh dài hạn và thích ứng với môi trường thị trường đang biến đổi. |

**PHẦN 3: QUÁ TRÌNH LÀM VIỆC NHÓM**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Họ và tên | Nhiệm vụ | Ưu điểm | Nhược điểm | Cải thiện | Đánh giá(%) |
| Huỳnh Nhật Nam  21110903 | - Time series analysis  - Mạng thần kinh nhân tạo (RNN, LSTM, GRU) | - Hoàn thành tiến độ  - Giúp đỡ các bạn  - Chủ động tìm hiểu | Cho nhóm chạy deadline quá trễ. Triển khai muộn. Hậu quả là kết quả không như mong muốn | Rút kinh nghiệm triển khai project sớm hơn, phân công rõ ràng. | 99 |
| Nguyễn Thị Lan Anh 21110368 | - Phương pháp phân cụm dựa trên mật độ - DBSCAN | - Hoàn thành tiến độ  - Giúp đỡ các bạn  - Chủ động tìm hiểu | - Còn khó khăn trong việc tìm lựa chọn tham số áp dụng phân cụm | Cần tìm hiểu kỹ lý thuyết trước khi bắt tay vào làm | 99 |
| Trịnh Thị Thanh Huyền  21110477 | - Phương pháp phân cụm K-means  - Khai phá luật kết hợp | - Hoàn thành tiến độ  - Giúp đỡ các bạn  - Chủ động tìm hiểu | - Đã có những phút giây muốn đổi dataset khi phân cụm không được | Cần kiên trì tìm nhiều cách trước khi quyết định (Cuối cùng đã tìm ra cách phân cụm và giữ nguyên dataset) | 99 |
| Phan Minh Thuận  21110667 | - Phương pháp phân cụm phân hoạch - Hierarchical | - Hoàn thành tiến độ  - Giúp đỡ các bạn  - Chủ động tìm hiểu | Dù đã cố gắng nhưng độ hiệu quả của mô hình vẫn không thể tốt hơn được nữa | Cố gắng tìm cách nâng độ hiệu quả của các mô hình | 99 |

|  |  |  |
| --- | --- | --- |
| Số điểm GV đã chấm | Nội dung chỉnh sửa | Đánh giá lại kết quả làm việc nhóm |
| 8.8 | - Bỏ bài toán Dự đoán doanh thu bằng phương pháp Linear Regression, SVM, Decision Tree, Random Forest. - Thêm thuật toán LSTM, RNN, GRU cho bài toán dự báo doanh thu | - Tất cả các thành viên đã kiểm tra, chỉnh sửa hoàn thiện đầy đủ.  - Đánh giá: 99%. |