HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG CƠ

SỞ TẠI TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN 2

BÁO CÁO GIỮA KỲ
MÔN HỌC : KHO DỮ LIỆU VÀ KHAI PHÁ

ĐỀ TÀI : Xây dựng kho dữ liệu kinh doanh mặt hàng cá cảnh. Khai phá dữ liệu này cho mục đích xây dựng giá bán và khuyến mãi các mặt hàng bán chậm

GVHD : Ths. Nguyễn Ngọc Duy
Lớp : D21CQCNHT01-N
Sinh viên : Phạm Tiến Anh -N21DCCN100

HỒ CHÍ MINH, NGÀY 20 THÁNG 03 NĂM 2025

MỤC LỤC

Nhận xét của giảng viên

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

I. Giới thiệu đề tài

1.1. Bối cảnh và vấn đề

Khi và đời sống vật chất của người dân đang được phát triển và nâng cao trong những năm gần đây thì đời sống tinh thần đang được mọi người chú trọng nhiều hơn. Ngành kinh doanh cá cảnh đang là một ngành nổi tương đối "hot" trong những năm gần đây vì cá không chỉ đơn giản là vật nuôi mà nhiều người còn quan niệm rằng nó là một con vật mang tính phong thủy.

Sự gia tăng nhu cầu thị trường cùng với sự đa dạng của các loài cá cảnh khiến cho việc quản lý và kinh doanh dần trở nên phức tạp hơn. Các chủ cửa hàng không chỉ quan tâm đến việc duy trì nguồn hàng mà còn phải theo dõi các yếu tố như tình trạng sức khỏe, môi trường sống hay là sự thay đổi của thị trường về những loài cá, định giá sao cho phù hợp để duy trì lợi nhuận và sức cạnh tranh.

Nhiều cửa hàng cá cảnh hiện nay gặp khó khăn trong việc năm bắt xu hướng thị trường và đưa ra các chiến lược hiệu quả. Việc xác định các mặt hàng bán chậm, thực hiện các chương trình khuyến mãi và điều chỉnh giá phù hợp là những yếu tố quan trọng để nâng cao doanh thu và giảm thiểu hàng tồn kho.

1.2. Giải pháp và đề xuất

Để giải quyết các vấn đề trên, cần xây dựng một kho dữ liệu (Data Warehouse) tập chung và đồng nhất, cho phép:

- Theo dõi chi tiết các giao dịch bán hàng, tình trạng hàng tồn kho, và các yếu tố ảnh hưởng đến sức mua.
- Áp dụng các thuật toán phân tích dữ liệu để xác định xu hướng tiêu thụ và sản phẩm bán chậm.
- Đưa ra các chiến lược giá bán và chương trình khuyến mãi phù hợp dựa trên các phân tích dữ liệu.

1.3. Lợi ích của đề tài

Việc xây dựng kho dữ liệu sẽ mang lại nhiều lợi ích cho doanh nghiệp, giúp cho các doanh nghiệp kinh doanh:

- Quản lý dữ liệu tập trung hơn, tăng cường khả năng theo dõi và quản lý thôn tin sản phẩm, giao dịch và tồn kho.
- Phân tích xu hướng tiêu thụ: Giúp xác định các sản phẩm bán chạy và bán chậm, tù đố đưa ra chiến lược bán hàng phù hợp.
- Đưa ra giá bán phù hợp với dữ liệu thực tế về nhu cầu thị trường và lợi nhuận.

- Giảm thiểu hàng tồn khi, tăng hiệu quả trong việc quản lý hàng hóa và trình trạng hàng tồn đọng kéo dài.
- Xây dựng các chiến lược bán hàng và khuyến mãi hợp lý cho các sản phẩm bán chậm để tăng sức mua từ khách hàng.

II. Tổng quan lý thuyết và khảo sát tài liệu

1. Khái niệm về kho dữ liệu, khai phá dữ liệu

1.1. Kho dữ liệu (Data Warehouse)

1.1.1. Khái niệm

Kho dữ liệu (Data Warehouse) là hệ thống lưu trữ dữ liệu tập trung, được thiết kế để hỗ trợ cho các hoạt động phân tích dữ liệu, tạo báo cáo và ra quyết định trong doanh nghiệp. Dữ liệu trong kho dữ liệu được trích xuất từ nhiều nguồn khác nhau như hệ thống giao dịch (OLTP), hệ thống quản lý khách hàng (CRM), các tệp dữ liệu từ nhà cung cấp, dữ liệu từ mạng xã hội hoặc các kênh bán hàng trực tuyến.

Kho dữ liệu được xây dựng dựa trên các đặc điểm chính sau:

- Dữ liệu tích hợp (Atomicity): Dữ liệu tập hợp từ nhiều nguồn khác nhau. Điều này sẽ dẫn đến việc quá trình tập hợp phải thực hiện việc làm sạch, sắp xếp, rút gọn dữ liệu.
- Theo chủ đề (Isolation): Các dữ liệu truy suất không bị ảnh hưởng bởi các dữ liệu khac hoặc tác động lên nhau.
- Dũ liệu cố định (Durable): Khi một Transaction hoàn chỉnh, dữ liệu không thể tạp thêm hay sửa đổi.

Quá trình xây dựng kho dữ liệu bao gồm ba bước quan trọng là trích xuất (Extract), biến đổi (Transform) và tải (Load), gọi chung là quy trình ETL. Dữ liệu sau khi được đưa vào kho sẽ được lưu trữ dưới dạng các bảng dữ liệu quan hệ, trong đó có hai loại bảng chính là bảng Fact và bảng Dimension. Bảng Fact là bảng chứa dữ liệu về các sự kiện hoặc giao dịch (ví dụ: số lượng bán, doanh thu, lợi nhuận…), trong khi bảng Dimension là bảng chứa thông tin mô tả về sản phẩm, khách hàng, thời gian, cửa hàng…

Kho dữ liệu thường được tổ chức theo hai loại lược đồ chính là lược đồ sao (Star Schema) và lược đồ bông tuyết (Snowflake Schema). Hai lược đồ này được sử dụng phổ biến trong các hệ thống phân tích dữ

liệu và có sự khác biệt rõ rệt về cấu trúc, độ phức tạp và hiệu suất xử lý.

1.1.2. Lược đồ sao (Star Schema)

Lược đồ sao là mô hình trong đó bảng Fact được liên kết trực tiếp với các bảng Dimension. Các bảng Dimension trong lược đồ sao thường không được chuẩn hóa, do đó có thể có sự trùng lặp dữ liệu trong các bảng Dimension. Lược đồ sao có cấu trúc đơn giản, dễ hiểu và cho phép thực hiện các truy vấn nhanh hơn do số lượng phép nối (JOIN) giữa các bảng ít hơn.

Trong lược đồ sao, các bảng Dimension thường được thiết kế theo dạng bảng phẳng, trong đó dữ liệu được tổ chức đầy đủ trong một bảng duy nhất mà không cần tham chiếu đến các bảng khác. Điều này giúp cho việc truy vấn dữ liệu diễn ra nhanh hơn và dễ dàng hơn, nhưng đồng thời cũng khiến cho kích thước của kho dữ liệu lớn hơn do dữ liệu trùng lặp.

Lược đồ sao thường được sử dụng trong các hệ thống phân tích trực tuyến (OLAP) và các báo cáo nhanh, nơi mà tốc độ truy vấn là yếu tố quan trọng hơn kích thước của cơ sở dữ liệu. Tuy nhiên, vì các bảng Dimension không được chuẩn hóa nên việc bảo trì và cập nhật dữ liệu có thể trở nên khó khăn hơn khi dữ liệu thay đổi theo thời gian.

1.1.3. Lược đồ bông tuyết (Snowflake Schema)

Lược đồ bông tuyết là mô hình trong đó các bảng Fact được liên kết với các bảng Dimension, nhưng các bảng Dimension lại được chuẩn hóa và phân tách thành nhiều bảng nhỏ hơn. Điều này làm cho cấu trúc dữ liệu trở nên phức tạp hơn nhưng giúp giảm thiểu sự trùng lặp và tiết kiệm không gian lưu trữ.

Trong lược đồ bông tuyết, dữ liệu trong các bảng Dimension được chia thành các bảng con theo nguyên tắc chuẩn hóa (1NF, 2NF, 3NF). Nhờ đó, dữ liệu trong kho sẽ nhất quán và dễ dàng cập nhật hơn khi có sự thay đổi từ các nguồn dữ liệu đầu vào. Tuy nhiên, việc truy vấn trong lược đồ bông tuyết thường chậm hơn so với lược đồ sao vì hệ thống cần thực hiện nhiều phép nối (JOIN) hơn để tổng hợp dữ liệu từ nhiều bảng khác nhau.

Lược đồ bông tuyết thường được sử dụng khi hệ thống cần lưu trữ một lượng lớn dữ liệu có tính phức tạp cao, và khi tính toàn vẹn của dữ liệu là yếu tố quan trọng hơn tốc độ truy vấn. Do dữ liệu

được chuẩn hóa, các hệ thống sử dụng lược đồ bông tuyết sẽ dễ dàng mở rộng và bảo trì hơn so với lược đồ sao.

1.1.4. So sánh giữa lược đồ sao và lược đồ bông tuyết

Hai loại lược đồ này có sự khác biệt rõ rệt về cấu trúc, hiệu suất và khả năng bảo trì:

Tiêu chí	Lược đồ sao (Star Schema)	Lược đồ bông tuyết (Snowflake Schema)
Cấu trúc	Bảng Fact liên kết trực tiếp với các bảng Dimension	Bảng Fact liên kết với các bảng Dimension được chuẩn hóa và phân tách thành nhiều bảng nhỏ hơn
Mức độ chuẩn hóa	Không chuẩn hóa (Denormalized)	Được chuẩn hóa (Normalized)
Tốc độ truy vấn	Nhanh hơn do số phép nối ít hơn	Chậm hơn do cần nhiều phép nối hơn
Dung lượng lưu trữ	Lớn hơn do dữ liệu trùng lặp trong các bảng Dimension	Nhỏ hơn do dữ liệu đã được chuẩn hóa, không trùng lặp
Khả năng bảo trì	Khó bảo trì hơn khi dữ liệu thay đổi do dữ liệu trùng lặp	Dễ bảo trì hơn do dữ liệu nhất quán và đã được chuẩn hóa
Tính dễ dàng trong phân tích	Dễ dàng cho người dùng vì các bảng Dimension đầy đủ thông tin và dễ dàng truy xuất	Phức tạp hơn vì cần nhiều phép nối khi phân tích dữ liệu
Phù hợp cho	Hệ thống OLAP và các báo cáo nhanh	Hệ thống có dữ liệu lớn, cần đảm bảo tính nhất quán và dễ bảo trì

1.1.5. Kho dữ liệu trong mặt hàng kinh doanh cá cảnh

Trong mặt hàng kinh doanh cá cảnh, kho dữ liệu sẽ giúp:

- Tập trung toàn bộ dữ liệu về các loại cá, thức ăn, môi trường sống, giao dịch bán hàng và tồn kho.
- Theo dõi chi tiết lịch sử bán hàng, xu hướng mua sắm của khách hàng và tình trạng tồn kho.
- Phân tích các yếu tố ảnh hưởng đến giá bán, từ đó đưa ra các chiến lược định giá và khuyến mãi phù hợp.

- Hỗ trợ các chủ kinh doanh ra quyết định dựa trên số liệu thực tế thay vì phán đoán.

1.2. Khai phá dữ liệu (Data Mining)

1.2.1. Khái niệm về khai phá dữ liệu

Khai phá dữ liệu (Data Mining) là quá trình trích xuất thông tin hữu ích, tiềm ẩn và có giá trị từ một tập dữ liệu lớn. Quá trình này nhằm phát hiện ra các mẫu, xu hướng hoặc mối quan hệ trong dữ liệu mà con người khó có thể nhận ra trực tiếp thông qua các công cụ truyền thống. Khai phá dữ liệu là một phần quan trọng của quá trình Phát hiện tri thức từ dữ liệu (KDD - Knowledge Discovery in Database).

Khai phá dữ liệu còn được gọi là:

- Khám phá tri thức (Knowledge Discovery): Phát hiện các mẫu và quy luật tiềm ẩn trong dữ liệu.
- Trích lọc tri thức: Lấy ra các thông tin hữu ích từ tập dữ liệu lớn.
- Phân tích mẫu: Xác định xu hướng và sự khác biệt trong dữ liệu.
- Khảo cổ dữ liệu: Đào sâu vào dữ liệu để tìm ra thông tin giá trị.
- Tri thức kinh doanh (Business Intelligence): Sử dụng thông tin thu được để hỗ trợ cho các quyết định kinh doanh.

1.2.2. Vai trò của khai phá dữ liệu trong quá trình phát hiện tri thức

Khai phá dữ liệu là một bước quan trọng trong quá trình phát hiện tri thức từ dữ liệu (KDD - Knowledge Discovery in Database). Quá trình phát hiện tri thức không chỉ dừng lại ở việc thu thập và lưu trữ dữ liệu, mà còn cần phân tích và khai thác dữ liệu để phát hiện ra các quy luật, mối quan hệ tiềm ẩn nhằm đưa ra các quyết định kinh doanh hoặc cải tiến quy trình hoạt động.

Quá trình phát hiện tri thức từ dữ liệu bao gồm 8 bước cơ bản:

- **Học từ lĩnh vực ứng dụng:**
  - Xác định tri thức cần khai phá và mục tiêu phân tích.
  - Ví dụ: Dự đoán xu hướng tiêu dùng của khách hàng trong tương lai.
- **Tạo tập dữ liệu đích**

  Sau khi xác định được mục tiêu, bước tiếp theo là thu thập và tổng hợp dữ liệu từ các nguồn khác nhau để tạo thành một tập dữ liệu hoàn chỉnh phục vụ cho việc phân tích.

  Nguồn dữ liệu có thể bao gồm:

- Hệ thống giao dịch (OLTP): Thông tin từ các đơn hàng, giao dịch, thanh toán…
- Hệ thống CRM: Dữ liệu về khách hàng (độ tuổi, giới tính, địa chỉ, hành vi mua sắm…).
- Hệ thống quản lý kho: Dữ liệu về số lượng hàng tồn kho, tần suất nhập hàng, chu kỳ mua hàng…
- Dữ liệu từ các kênh trực tuyến: Dữ liệu từ mạng xã hội, đánh giá của khách hàng, hành vi tìm kiếm trên website…

- **Tiền xử lý và làm sạch dữ liệu**

  Dữ liệu thu thập từ nhiều nguồn khác nhau có thể có nhiều vấn đề như:

- Thiếu dữ liệu: Một số giá trị có thể bị bỏ trống.
- Dữ liệu không nhất quán: Định dạng dữ liệu có thể không đồng nhất (ví dụ: ngày tháng có thể ở định dạng khác nhau).
- Dữ liệu trùng lặp: Cùng một thông tin có thể xuất hiện nhiều lần trong các nguồn khác nhau.
- Dữ liệu ngoại lệ: Các giá trị bất thường có thể gây nhiễu cho mô hình phân tích.

- **Chuyển đổi và thu hẹp dữ liệu**

  Dữ liệu sau khi được làm sạch cần được chuyển đổi về định dạng phù hợp để dễ dàng khai thác. Quá trình này bao gồm:

- Giảm chiều dữ liệu: Chỉ giữ lại các thuộc tính quan trọng nhất.
- Chuẩn hóa dữ liệu: Đưa dữ liệu về cùng một khoảng giá trị.
- Mã hóa dữ liệu: Chuyển các giá trị chuỗi thành giá trị số.
- Tạo các thuộc tính mới: Kết hợp các thuộc tính hiện có để tạo thuộc tính mới có ý nghĩa hơn.

- **Lựa chọn phương pháp khai phá dữ liệu**

  Tùy vào mục tiêu và loại dữ liệu, có thể áp dụng các phương pháp khai phá khác nhau như:

- Hồi quy tuyến tính (Linear Regression): Dự đoán các giá trị liên tục như doanh thu, lợi nhuận…
- Phân cụm (Clustering): Phân nhóm các khách hàng có đặc điểm tương tự nhau.
- Cây quyết định (Decision Tree): Xác định yếu tố ảnh hưởng đến hành vi của khách hàng.
- Luật kết hợp (Association Rule): Tìm ra mối quan hệ giữa các sản phẩm trong giỏ hàng của khách hàng.

- **Khai phá dữ liệu**

  Áp dụng các thuật toán học máy hoặc thống kê để phát hiện ra các mẫu, xu hướng hoặc quy luật trong dữ liệu.

  Ví dụ:

- Sử dụng thuật toán K-Means để phân nhóm khách hàng dựa trên hành vi mua sắm.
- Sử dụng thuật toán Apriori để phát hiện mối quan hệ giữa các sản phẩm (nếu khách hàng mua sữa thì thường mua bánh mì).

- **Đánh giá và diễn giải kết quả**

  Đánh giá độ chính xác của mô hình và diễn giải kết quả thông qua các công cụ trực quan hóa (biểu đồ, bảng số liệu…).

- **Sử dụng tri thức**

  Tri thức thu được có thể được sử dụng để:

- Tối ưu hóa giá bán và chương trình khuyến mãi.
- Cải thiện dịch vụ khách hàng.
- Đưa ra chiến lược kinh doanh phù hợp hơn.

1.2.3. Khai phá dữ liệu trong mặt hàng kinh doanh cá cảnh

- Phân tích xu hướng tiêu thụ: Giúp cửa hàng xác định các loại cá được ưa chuông theo từng mùa.
- Phát hiện sản phẩm ban chậm: Phát hiện các loại cá bán chậm để đưa ra chương trình khuyến mãi.

- Tối ưu giá bán: Đưa ra mức giá phù hợp dựa trên xu hướng mua hàng và nhu cầu thực tế của thị trường.
- Phân loại khách hàng: Xác định các nhóm khách hàng để triển khai các chiến lược tiếp thị phù hợp.

2. Các phương pháp

2.1. Phương pháp ETL

Phương pháp ETL là quy trình chuẩn trong việc xây dựng kho dữ liệu, bao gồm 3 bước:

- Extract (Trích xuất): Trích xuất dữ liệu từ nhiều nguồn khác nhau như hệ thống bán hàng, hệ thống quản lý kho và phản hồi từ khách hàng.
- Transform (Chuyển đổi): Làm sạch, chuẩn hóa và chuyển đổi dữ liệu về cùng định dạng.
- Load (Tải): Nạp dữ liệu đã xử lý vào kho dữ liệu sẵn sàng cho việc phân tích.

2.2. Các phương pháp khai phá dữ liệu

Khai phá dữ liệu là quá trình áp dụng các phương pháp và thuật toán khác nhau để phát hiện tri thức từ dữ liệu. Mỗi phương pháp có những thuật toán đặc trưng, phù hợp với từng loại bài toán cụ thể và đặc điểm của dữ liệu. Dưới đây là các phương pháp khai phá dữ liệu chính, cùng với các thuật toán phổ biến được sử dụng cho từng phương pháp.

2.2.1. Phương pháp phân loại (Classification)

Phân loại là bài toán dự đoán nhãn (class) của một đối tượng mới dựa trên các dữ liệu đã được gán nhãn trước đó. Mục tiêu của phân loại là xây dựng một mô hình từ dữ liệu huấn luyện và sử dụng mô hình đó để phân loại các đối tượng chưa được biết trước.

Nguyên tắc hoạt động:

- Dữ liệu huấn luyện bao gồm các bản ghi (record) có nhãn phân loại.
- Mô hình được huấn luyện trên tập dữ liệu đã gán nhãn.
- Khi có một bản ghi mới, mô hình sẽ dự đoán nhãn của bản ghi đó dựa trên các đặc điểm đã học được từ dữ liệu huấn luyện.

Thuật toán phân loại phổ biến:

a. Thuật toán cây quyết định (Decision Tree)

- Là thuật toán phân loại dựa trên cấu trúc cây.
- Trong cây quyết định:
  + Mỗi nút (node) là một điều kiện kiểm tra thuộc tính.

+ Mỗi nhánh (branch) là một kết quả của điều kiện kiểm tra.
+ Mỗi lá (leaf) là một nhãn của dữ liệu.

- Cách hoạt động:
  + Chọn thuộc tính phân chia tốt nhất bằng cách tính Entropy (ID3) hoặc Gini Index (CART).
  + Tách dữ liệu theo thuộc tính được chọn.
  + Tiếp tục chia cho đến khi đạt điều kiện dừng (các lá chứa các nhãn cụ thể).

b. Thuật toán hồi quy logistic (Logistic Regression)

Là thuật toán phân loại nhị phân (có/không) dựa trên xác suất. Sử dụng hàm sigmoid để đưa kết quả đầu ra về khoảng [0, 1]:

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+...+b_nx_n)}}$$

Kết quả là xác suất của đối tượng thuộc vào một lớp cụ thể.

c. Thuật toán K-Nearest Neighbors (KNN)

- Dựa trên khoảng cách giữa các điểm dữ liệu để phân loại.
- Nguyên tắc hoạt động:
  + Xác định giá trị K (số lượng điểm lân cận gần nhất).
  + Tính khoảng cách giữa điểm mới và các điểm lân cận (thường sử dụng khoảng cách Euclidean):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

  + Dự đoán nhãn của điểm mới dựa trên số lượng lớn nhất của các điểm lân cận.

d. Thuật toán Naive Bayes

- Dựa trên lý thuyết xác suất Bayes.
- Giả định rằng các thuộc tính độc lập với nhau.

Công thức:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

2.2.2. Phương pháp hồi quy (Regression)

Hồi quy là bài toán dự đoán giá trị của một biến liên tục (continuous value) dựa trên mối quan hệ giữa các biến đầu vào và biến đầu ra.

Nguyên tắc hoạt động:

- Xây dựng mô hình hồi quy dựa trên mối quan hệ giữa các biến đầu vào và biến đầu ra.
- Sử dụng các phương pháp tối ưu để tìm ra mối quan hệ tốt nhất.

Thuật toán hồi quy phổ biến:

- Hồi quy tuyến tính (Linear Regression)
  Giả định rằng có mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc.
- Mô hình có dạng:

$$Y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$$

- Phương pháp tối ưu: Tối thiểu hóa sai số bình phương (Least Squares).
- Hồi quy phi tuyến (Non-Linear Regression)
  Khi mối quan hệ giữa biến độc lập và biến phụ thuộc là phi tuyến.
  Sử dụng các phương pháp như:
  + Hồi quy bậc hai
  + Hồi quy hàm mũ
  + Hồi quy logarit

2.2.3. Phương pháp phân cụm (Clustering)

Phân cụm là bài toán nhóm các đối tượng có đặc điểm tương đồng vào cùng một cụm mà không cần nhãn trước đó.
Thuật toán phân cụm phổ biến:

- K-Means
  Xác định số lượng cụm K.
  Tính khoảng cách giữa các điểm và trọng tâm của các cụm.
  Lặp lại cho đến khi các cụm ổn định.
- Hierarchical Clustering
  Tạo phân cấp các cụm từ dưới lên hoặc từ trên xuống.
  Sử dụng khoảng cách Euclidean hoặc Cosine để xác định các cụm gần nhất.

2.2.4. Phương pháp phát hiện luật kết hợp (Association Rule Mining)

Phát hiện luật kết hợp là quá trình tìm ra các mối quan hệ trong dữ liệu.

2.2.5. Phương pháp phát hiện bất thường (Anomaly Detection)

Phát hiện bất thường là bài toán tìm kiếm các điểm dữ liệu khác biệt so với số đông.

3. Công nghệ

3.1. SQL Server

3.1.1. Khái niệm

Microsoft SQL Server là một hệ quản trị cơ sở dữ liệu quan hệ được phát triển bởi Microsoft. Là một máy chủ cơ sở dữ liệu, nó là một sản phẩm phần mềm có chức năng chính là lưu trữ và truy xuất dữ liệu theo yêu cầu của các ứng dụng phần mềm khác. Có thể chạy trên cùng một máy tính hoặc trên một máy tính khác trên mạng (bao gồm cả Internet).

Microsoft tiếp thị ít nhất một chục phiên bản Microsoft SQL Server khác nhau, nhắm vào các đối tượng khác nhau và cho khối lượng công việc khác nhau, từ các ứng dụng máy đơn nhỏ đến các ứng dụng Internet lớn có nhiều người dùng đồng thời.

Microsoft cung cấp tính năng quản lý dữ liệu cùng **SQL Server** với các dịch vụ tích hợp lập trình SQL Server, SQL Server Data Quality và SQL Server Master. Ngoài ra, hai bộ công cụ dành riêng cho quản trị viên cơ sở dữ liệu (DBAs) và lập trình viên:

- SQL Server Data Tools: Được sử dụng trong việc phát triển cơ sở dữ liệu.
- SQL Server Management Studio được ứng dụng để triển khai, giám sát và quản lý cơ sở dữ liệu.

SQL Server còn được trang bị tính năng kinh doanh giúp người dùng có thể thực hiện phân tích dữ liệu thông qua:

- SQL Server Analysis Services (SSAS): sử dụng để phân tích các dữ liệu.
- SQL Server Reporting Services: để tạo ra báo cáo dễ dàng hơn.

3.1.2. SQL server trong kho dữ liệu cá cảnh

Trong kho dữ liệu cá cảnh, SQL Server đóng vai trò là nền tảng lưu trữ và quả lý dữ liệu bao gồm: lưu trữ dữ liệu về các loại cá, số lượng tồn kho, giá bán, môi trường sống, thức ăn, giao dịch,..; tối ưu hóa truy vấn để trích xuất dữ liệu một cách nhanh hơn cho việc phân tích và báo cáo.

3.2. Python

3.2.1. Khái niệm

Python là một ngôn ngữ lập trình mã nguồn mở và đa nền tảng được sử dụng rộng rãi trong phát triển phần mềm, khoa học dữ liệu và học máy. Python có cú pháp đơn giản, dễ học và thư

viện phong phú, phù hợp cho cả người mới bắt đầu và các lập trình viên chuyên nghiệp.

Đặc điểm của python:

- Ngôn ngữ bậc cao: Dễ đọc, dễ viết và dễ duy trì.
- Hỗ trợ nhiều thư viện: Như NumPy, Pandas, Scikit-learn, TensorFlow, ...
- Hỗ trợ lập trình hướng đối tượng và lập trình hàm: Tăng tính linh hoạt trong xử lý dữ liệu.
- Khả năng tích hợp: Kết nối tốt với các hệ quản trị cơ sở dữ liệu như SQL Server, MySQL, PostgreSQL,...

3.2.2. Python với kho dữ liệu cá cảnh

Python được sử dụng để xử lý, phân tích và khai phá dữ liệu từ kho dữ liệu cá cảnh:

- Xử lý dữ liệu: Làm sạch, chuẩn hóa dữ liệu từ kho dữ liệu trước khi phân tích.
- Phân tích xu hướng: Dự đoán xu hướng tiêu thụ dựa trên dữ liệu lịch sử.
- Phát hiện các mẫu (pattern): Phát hiện các loại cá bán chậm để điều chỉnh giá và triển khai khuyến mãi.
- Trực quan hóa dữ liệu: Dùng thư viện Matplotlib hoặc Seaborn để vẽ biểu đồ.
- Học máy (Machine Learning): Dùng các thuật toán hồi quy và phân loại để dự đoán nhu cầu và hành vi mua hàng

III. Xây dựng kho dữ liệu và khai phá kho dữ liệu

I. Tổng quan về dữ liệu

1.1. Nguồn dữ liệu sử dụng

Dữ liệu trong kho cá cảnh được thu thập từ nhiều nguồn khác nhau, bao gồm các website thương mai điện tử và các nhà cung cấp. Dữ liệu được thu tập từ các công cụ thu thập dữ liệu tự động, được tổng hợp và xuất thành file excel để đưa vào quá trình tiền xử lý và xây dựng kho dữ liệu.

1.2. Dữ liệu sau khi trích xuất

Dữ liệu sau khi được thu thập lấy ra 194 dòng và 4 cột phục vụ mục đích xây dựng kho dữ liệu.

Dữ liệu sau khi trích xuất từ dữ liệu gốc:

	Neon Lemon Mắt Đỏ - Lemon Tetra Albino	Hết hàng	6.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bizweb.dkt	Neon Rubi - Rubi Tetra		30.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ,bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Pristella Tetra Golden	Hết hàng	20.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bizweb.dkt	Rosy Tetra	Hết hàng	14.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bizweb.dkt	Trâm Kalochroma - Rasbora Kalochroma		45.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ...Tập tính: sống theo đàn,có cần nhau
https://bizweb.dkt	White Fin Tetra	Hết hàng	32.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bizweb.dkt	White Spot Tetra	Hết hàng	20.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bizweb.dkt	White Tetra	Hết hàng	16.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
	Cá Dĩa Bồ Câu Đỏ - Pigeon Red Discus	Còn hàng	160.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Bồ Câu Panda - Pigeon Red Panda Discus	Còn hàng	160.000đ	#N/A
	Cá Dĩa Bồ Câu Vàng - Pigeon Yellow Discus	Còn hàng	160.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Bông Đỏ - Red Spotted Discus	Còn hàng	160.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Bông Xanh - Mixed Blue Discus	Còn hàng	160.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Đỏ Trắng - Red-White Leopand Discus	Còn hàng	160.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Hoa Hồng - Pink Rose Red	Còn hàng	160.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Lam Colban - Blue Solid Discus	Còn hàng	270.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Lam Đức - Blue Diamond Discus	Còn hàng	160.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Man Đỏ - Manrlboro Red Discus	Còn hàng	160.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Man Vàng - Marlboro Yellow Discus	Còn hàng	270.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Trắng - White Glass Pigeon Discus	Còn hàng	160.000đ	#N/A
	Cá Dĩa Beo - Leopar Red Discus	Hết hàng	450.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
	Cá Dĩa Bồ Câu Vàng Mắt Đỏ - Pegion Yellow Albino Discus	Hết hàng	450.000đ	#N/A
https://bizweb.dkt	Cá Dĩa Đỏ Mắt Đỏ - Manrlboro Red Albino	Hết hàng	450.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
https://bizweb.dkt	Cá Dĩa Đỏ Trắng Mắt Đỏ - Red-White Albino Discus	Hết hàng	450.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Lam Mắt Đỏ - Platinum Albino Discus	Còn hàng	450.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
	Cá Dĩa Vàng Mắt Đỏ - Manrlboro Yellow Albino	Còn hàng	450.000đ	pH: 7 Nhiệt độ: 28°C-30°CThức ăn: trùn chỉ, tim bò...Tập tính: sống bày đàn
https://bizweb.dkt	Black And White Angelfish	Hết hàng	64.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám, trùn chỉ,bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Black And White Angelfish Longfin	Hết hàng	70.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám, trùn chỉ,bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Blue Paraiba Angelfish	Hết hàng	90.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám, trùn chỉ,bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Cá Betta Super Yellow		120.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám, trùn chỉ, bobo...Tập tính:hay cắn nhau, sống ri
https://bizweb.dkt	Cầu Vồng Parkinsoni (Full Trống) - Parkinsoni	Còn hàng	170.000đ	pH: 7 Nhiệt độ: 26-28°CThức ăn: cám, trùn chỉ, bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Cá Cầu Vồng Nắng Vàng - Forktail Rainbowfish	Hết hàng	30.000đ	pH: 7 Nhiệt độ: 26-28°CThức ăn: cám, trùn chỉ, bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Cá Cầu Vồng Táo Đỏ	Hết hàng	100.000đ	pH: 7 Nhiệt độ: 26-28°CThức ăn: cám, trùn chỉ, bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Cá Cầu Vồng Thạch Mỹ Nhân (Full Trống) - Boesemani Rainbowfi	Hết hàng	140.000đ	pH: 7 Nhiệt độ: 26-28°CThức ăn: cám, trùn chỉ, bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Cá Cầu Vồng Thạch Mỹ Nhân Shortbody - Boesemani	Hết hàng	198.000đ	pH: 7 Nhiệt độ: 26-28°CThức ăn: cám, trùn chỉ, bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Cá Cầu Vồng Xanh - Dwarf Rainbow Fish	Còn hàng	8.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ,bobo...Tập tính:sống theo đàn
	Cá Cầu Vồng Xanh Indo - Blue Rainbowfish	Hết hàng	60.000đ	
	Cầu Vồng Ivantsoff'S Blue-Eye	Hết hàng	48.000đ	pH: 7 Nhiệt độ: 26-28°CThức ăn: cám, trùn chỉ, bobo...Tập tính:sống theo đàn
	Cầu Vồng Maculochi - Maccullochi Rainbowfish	Hết hàng	60.000đ	pH: 7 Nhiệt độ: 26-28°CThức ăn: cám, trùn chỉ, bobo...Tập tính:sống theo đàn
https://bizweb.dkt	Cầu Vồng Madagascar - Madagascar Rainbowfish	Hết hàng	40.000đ	
https://bizweb.dkt	Cầu Vồng Red Neon - Paska'S Blue-Eye	Hết hàng	60.000đ	
https://bizweb.dkt	Cầu Vồng Vây Dài (Full Trống) - Threadfin	Hết hàng	28.000đ	
https://bizweb.dkt	Tép Aura		18.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Dreamblue		18.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Mix Màu		5.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Ong Đà Nẵng		10.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Red Chery		8.000đ	
https://bizweb.dkt	Tép Rili Đen		30.000đ	
https://bizweb.dkt	Tép Rili Đỏ - Red Rili Shrimp		8.000đ	
https://bizweb.dkt	Tép Rili Vàng		8.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Socola		30.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Src Đỏ		8.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Thanh Mai		10.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Tiger		10.000đ	pH: 7 Nhiệt độ: 24°CThức ăn: tảo, rêu, cám dùng cho tép,...Tập tính: sống theo đà
https://bizweb.dkt	Tép Vàng Cam		8.000đ	
https://bizweb.dkt	Tép Yamato Vn		8.000đ	

II. Tiền xử lý dữ liệu

2.1. Xử lý các giá trị khuyết thiếu

Trong quá trình thu thập dữ liệu, một số giá trị trong các cột bị thiếu hoặc không hợp lệ. Để đảm bảo tính toàn vẹn và nhất quán, các dữ liệu khuyết thiếu được xử lý:

- Các giá trị khuyết thiếu trong price được thay thế bằng giá trị trung bình của loại cá tương ứng để đảm bảo tính hợp lý khi phân tích.
- Nếu số lượng bị thiếu, các giá trị được điền mặc định là 0.

https://bi:	Cá Buồm [	0	6.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Buồm \	475	6.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Cánh B	437	6.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Chỉ Đỏ	0	16.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Hắc Kỳ	380	8.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Hồng N	409	6.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Hồng N	0	30.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Kim Tơ	78	13.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Na Na -	0	6.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon B	0	18.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon B	0	16.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon Đ	0	7.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon H	0	18.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon H	0	20.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon H	0	14.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon V	0	14.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
	Cá Neon V	7005	8.000đ	pH: 7 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon X	752	7.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon X	0	20.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Neon X	0	35.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn
https://bi:	Cá Ngân B	0	8.000đ	pH: 6 Nhiệt độ: 28°CThức ăn: cám,trùn chỉ, bobo...Tập tính: sống bày đàn

2.2. Sử dụng đoạn code python để phân phân chia chi tiết hơn

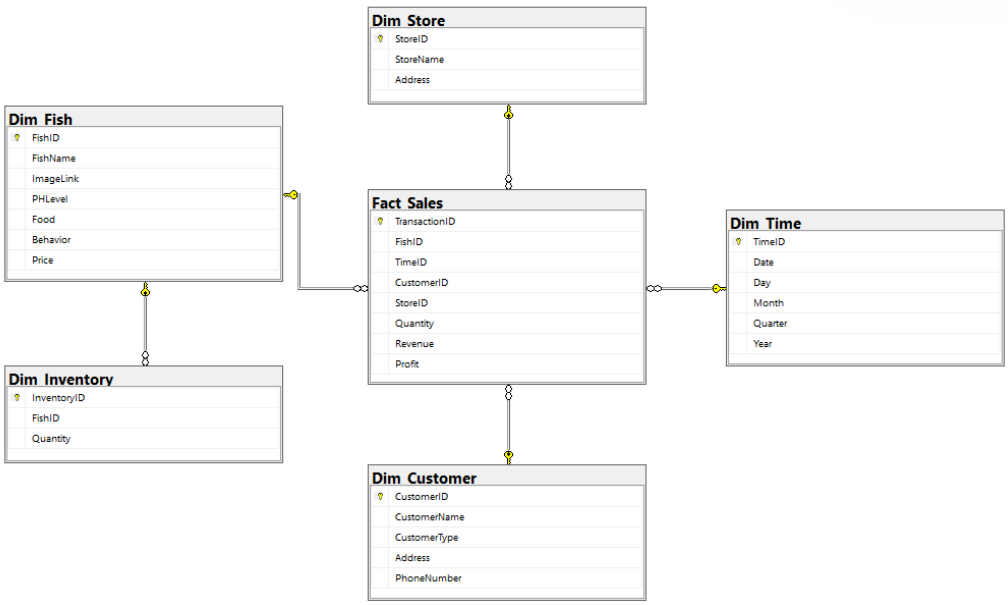
Sử dụn python để phân chia các đặc điểm môi trường và tập tính của cá chi tiết hơn thành độ PH môi trường, thức ăn của cá và tập tính của các loại cá.

```
def extract_condition(text):
    if isinstance(text, str) and text.strip():

        ph_match = re.search(r'pH:\s*([\d.]+)', text)
        ph = float(ph_match.group(1)) if ph_match else None

        temp_match = re.search(r'Nhiệt độ:\s*(\d+°C)', text)
        temperature = temp_match.group(1) if temp_match else None

        food_match = re.search(r'Thức ăn:\s*(.*)', text)
        food = food_match.group(1).strip() if food_match else None

        return pd.Series([ph, temperature, food])
    else:
        return pd.Series([None, None, None])
```

7	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
7	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...
6	28°C	cám,trùn chỉ, bobo...

III. Snowflake Schema (Lược đồ hinh bông tuyết)

Trong kho dữ liệu kinh doanh cá cảnh, Snowflake được áp dụng để tổ chức dữ liệu theo cấu trúc phân cấp, bao gồm:



3.1. Fact_Sales

Lưu thông tin các lịch sử giao dịch để dễ dàng phân tích

Thuộc tính	Kiểu dữ liệu	Mô tả
TransactionID	Int	ID của từng giao dịch
FishID	Int	Mã cá
StoreID	Int	Mã của hàng
CustomerID	Int	Mã khách hàng
TimeID	Int	Mã thời gian
Quantity	Int	Số lượng bán
Revenue	Money	Doanh thu
Profit	Float	Lợi nhuận

3.2. Dim_Time

Lưu thông tin chi tiết về thời gian dễ dàng truy vấn phân tích theo kỳ

Thuộc tính	Kiểu dữ liệu	Mô tả
TimeID	Int	ID thời gian
Date	Date	Ngày cụ thể
Day	Int	Ngày trong tháng
Month	Int	Tháng trong năm
Year	Int	Năm

Quarter	Int	Quý trong năm

- Phân tích doanh thu theo ngày, tháng, quý, năm
- Dự đoán xu hướng khách hàng mua theo mùa
- Điều chỉnh giá bán phù hợp với chu kỳ của khách hàng

3.3. Dim_Store

Thuộc tính	Kiểu dữ liệu	Mô tả
StoreID	Int	ID cửa hàng, chi nhánh
StoreName	Nvarchar	Tên cửa hàng
Address	Nvarchar	Địa chỉ cửa hàng

- Phân tích doanh thu theo từng cửa hàng
- Điều chỉnh giá bán theo khu vực cho phù hợp

3.4. Dim_Fish

Thuộc tính	Kiểu dữ liệu	Mô tả
FishID	Int	ID cá
FishName	Nvarchar	Tên ca
ImageLink	Nvarchar	Lưu trữ linh hình ảnh mẫu của các loại cá
PHLevel	Float	Nồng độ PH
Food	Nvarchar	Đồ ăn của cá
Behavior	Nvarchar	Tập tính của cá
Price	Money	Giá thành của cá

- Xác định loại cá bán chạy nhất
- Dự đoán xu hướng mua hàng
- Điều chỉnh giá bán dựa trên nhu cầu và điều kiện của khách hàng
- Đưa ra các gợi ý về thức ăn và môi trường sống phù hợp cho khách hàng.

3.5. Dim_Customer

Lưu trữ thông tin người mua hàng

Thuộc tính	Kiểu dữ liệu	Mô tả
CustomerID	Int	Id khách hàng
CustomerName	Nvarchar	Tên khách hàng
CustomerType	Enum	Cá nhân hoặc doanh nghiệp
Address	Nvarchar	Địa chỉ người mua hàng

PhoneNumber	Nvachar	Số điện thoại người mua hàng

- Xác định tần suất mua hàng cảu từng nhóm hàng
- Có thể xác định được mặt hàng được ưa chuộng để triển khai các chương trình khuyến mãi phù hợp
- Xác định khu vực mua hàng tập trung nhất để mở có thể phát triển thêm trong tương lai

3.6. Dim_Inventory

Lưu trữ thông tin tồn kho của các loại cá

Thuộc tính	Kiểu dữ liệu	Mô tả
InventoryID	Int	ID
FishID	Int	Mã cá
Quantity	Int	Số lượng cá trong kho

IV. Khai phá dữ liệu

4.1. Chuẩn hóa dữ liệu

4.1.1. Chuyển đổi dữ liệu thời gian

Dữ liệu thời gian từ cở sở dữ liệu được lưu dưới dạng chuỗi. Để sử dụng dữ liệu để khai phá, cần chuyển đổi sang dạng timestamp để biểu diễn dưới dạng số nguyên.

Từ dữ liệu thời gian chuẩn hóa, tiến hành trích xuất các đặc trưng quan trọng sau:

- Year → Phát hiện xu hướng theo năm.
- Month → Phát hiện xu hướng theo tháng.
- Quarter → Phát hiện xu hướng theo quý.
- Weekday → Phát hiện xu hướng theo ngày trong tuần.

```
def process_datetime(df, column):
    # Chuyển dữ liệu thời gian về dạng datetime
    df[column] = pd.to_datetime(df[column])

    # Chuyển về dạng số (timestamp) theo giây
    df[column + '_timestamp'] = df[column].view('int64') // 10**9

    # Trích xuất các thuộc tính từ dữ liệu thời gian
    df['Year'] = df[column].dt.year
    df['Month'] = df[column].dt.month
    df['Quarter'] = df[column].dt.quarter
    df['Weekday'] = df[column].dt.weekday
```

Dữ liệu thời gia sau khi được chuẩn hóa

FishID	FishName	SaleDate	Year	Month	Quarter	QuantitySc	Revenue	Profit	InventoryC	Price	SaleDate_	Weekday
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	475	2850000	285000	475	6000	1.74E+09	2
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	475	2850000	342000	475	6000	1.74E+09	2
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	475	2850000	627000	475	6000	1.74E+09	2
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	437	2622000	419520	437	6000	1.74E+09	2
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	437	2622000	786600	437	6000	1.74E+09	2
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	437	2622000	707940	437	6000	1.74E+09	2
781	CÃ¡ HÃ°c	1/1/2025	2025	1	1	380	3040000	668800	380	8000	1.74E+09	2
781	CÃ¡ HÃ°c	1/1/2025	2025	1	1	380	3040000	304000	380	8000	1.74E+09	2
781	CÃ¡ HÃ°c	1/1/2025	2025	1	1	380	3040000	516800	380	8000	1.74E+09	2
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	409	2454000	588960	409	6000	1.74E+09	2
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	409	2454000	294480	409	6000	1.74E+09	2
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	409	2454000	269940	409	6000	1.74E+09	2
784	CÃ¡ Kim T	1/1/2025	2025	1	1	78	1014000	243360	78	13000	1.74E+09	2
784	CÃ¡ Kim T	1/1/2025	2025	1	1	78	1014000	131820	78	13000	1.74E+09	2
784	CÃ¡ Kim T	1/1/2025	2025	1	1	78	1014000	111540	78	13000	1.74E+09	2
793	CÃ¡ Neon	1/1/2025	2025	1	1	7005	56040000	14010000	7005	8000	1.74E+09	2
793	CÃ¡ Neon	1/1/2025	2025	1	1	7005	56040000	10647600	7005	8000	1.74E+09	2
793	CÃ¡ Neon	1/1/2025	2025	1	1	7005	56040000	14570400	7005	8000	1.74E+09	2
794	CÃ¡ Neon	1/1/2025	2025	1	1	752	5264000	894880	752	7000	1.74E+09	2
794	CÃ¡ Neon	1/1/2025	2025	1	1	752	5264000	1316000	752	7000	1.74E+09	2
794	CÃ¡ Neon	1/1/2025	2025	1	1	752	5264000	1421280	752	7000	1.74E+09	2

4.1.2. Chuẩn hóa dữ liệu

Việc chuẩn hóa dữ liệu là một bước quan trọng trong quá trình xử lý dữ liệu, đặc biệt khi sử dụng các mô hình học máy. Mục tiêu của việc chuẩn hóa dữ liệu bao gồm:

- Đưa dữ liệu về cùng một thang đo để đảm bảo các thuộc tính có trọng số tương đương nhau.
- Tránh tình trạng mô hình bị ảnh hưởng bởi các giá trị có đơn vị hoặc giá trị quá lớn.
- Tăng độ hội tụ và cải thiện hiệu suất của mô hình.

a. Sử dụng MinMaxScaler

MinMaxScaler giúp giữ nguyên cấu trúc dữ liệu ban đầu, thích hợp cho các thuật toán dựa trên khoảng cách như hồi quy tuyến tính, SVM, k-means, đưa tất cả các thuộc tính về cùng thang đo, giúp mô hình dễ dàng học các trọng số phù hợp.

Phương pháp Min-Max Scaling sẽ đưa các giá trị của từng thuộc tính về khoảng [0, 1] theo công thức:

$$X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Trong đó:

- X = Giá trị thuộc tính
- min(X) = Giá trị nhỏ nhất của thuộc tính
- max(X) = Giá trị lớn nhất của thuộc tính
- X_{scaled} = Giá trị sau khi chuẩn hóa

b. Các thuộc tính cần chuẩn hóa

Các thuộc tính số được chọn để chuẩn hóa:

- QuantitySold → Số lượng bán ra
- Revenue → Doanh thu
- Profit → Lợi nhuận
- Quantity → Lượng hàng tồn kho
- Price → Giá bán

c. Hàm chuẩn hóa dữ liệu
Sử dụng câu lệnh query để lấy dữ liệu từ trong SQL Server

```
query = """
SELECT
    S.FishID,
    F.FishName,
    T.Date AS SaleDate,
    T.Year,
    T.Month,
    T.Quarter,
    S.Quantity AS QuantitySold,
    S.Revenue,
    S.Profit,
    I.Quantity AS InventoryQuantity,
    F.Price
FROM
    Fact_Sales S
JOIN
    Dim_Fish F ON S.FishID = F.FishID
JOIN
    S.Revenue,
    S.Profit,
    I.Quantity AS InventoryQuantity,
    F.Price
FROM
    Fact_Sales S
JOIN
    Dim_Fish F ON S.FishID = F.FishID
JOIN
    Dim_Inventory I ON F.FishID = I.FishID
JOIN
    Dim_Time T ON S.TimeID = T.TimeID;
"""
scaler = MinMaxScaler()
cols_to_scale = ['QuantitySold', 'Revenue', 'Profit', 'InventoryQuantity', 'Price']

missing_cols = [col for col in cols_to_scale if col not in df.columns]
if missing_cols:
    raise ValueError(f"Các cột sau không tồn tại trong dữ liệu: {missing_cols}")
```

Dữ liệu sau khi chuẩn hóa

FishID	FishName	SaleDate	Year	Month	Quarter	QuantityS	Revenue	Profit	InventoryC	Price
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	0.0533	0.0311	0.01062	0.0533	0.00225
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	0.0533	0.0311	0.01276	0.0533	0.00225
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	0.0533	0.0311	0.02349	0.0533	0.00225
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	0.04901	0.02859	0.01568	0.04901	0.00225
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	0.04901	0.02859	0.02949	0.04901	0.00225
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	0.04901	0.02859	0.02653	0.04901	0.00225
781	CÃ¡ Há°c	1/1/2025	2025	1	1	0.04257	0.03318	0.02506	0.04257	0.00674
781	CÃ¡ Há°c	1/1/2025	2025	1	1	0.04257	0.03318	0.01133	0.04257	0.00674
781	CÃ¡ Há°c	1/1/2025	2025	1	1	0.04257	0.03318	0.01934	0.04257	0.00674
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	0.04585	0.02674	0.02205	0.04585	0.00225
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	0.04585	0.02674	0.01097	0.04585	0.00225
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	0.04585	0.02674	0.01005	0.04585	0.00225
784	CÃ¡ Kim T	1/1/2025	2025	1	1	0.00847	0.01091	0.00905	0.00847	0.01798
784	CÃ¡ Kim T	1/1/2025	2025	1	1	0.00847	0.01091	0.00485	0.00847	0.01798
784	CÃ¡ Kim T	1/1/2025	2025	1	1	0.00847	0.01091	0.00409	0.00847	0.01798
793	CÃ¡ Neon	1/1/2025	2025	1	1	0.79074	0.61574	0.52715	0.79074	0.00674
793	CÃ¡ Neon	1/1/2025	2025	1	1	0.79074	0.61574	0.40061	0.79074	0.00674
793	CÃ¡ Neon	1/1/2025	2025	1	1	0.79074	0.61574	0.54825	0.79074	0.00674
794	CÃ¡ Neon	1/1/2025	2025	1	1	0.08458	0.05763	0.03357	0.08458	0.00449
794	CÃ¡ Neon	1/1/2025	2025	1	1	0.08458	0.05763	0.04942	0.08458	0.00449
794	CÃ¡ Neon	1/1/2025	2025	1	1	0.08458	0.05763	0.05338	0.08458	0.00449
798	CÃ¡ SÃ³c Ä	1/1/2025	2025	1	1	1	0.9734	1	1	0.01124
798	CÃ¡ SÃ³c Ä	1/1/2025	2025	1	1	1	0.9734	0.46661	1	0.01124
798	CÃ¡ SÃ³c Ä	1/1/2025	2025	1	1	1	0.9734	0.89999	1	0.01124
802	CÃ¡ TrÃ¢m	1/1/2025	2025	1	1	0.05342	0.20905	0.20053	0.05342	0.07865
819	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00791	0.12815	0.10978	0.00791	0.34831
819	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00791	0.12815	0.08341	0.00791	0.34831
819	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00791	0.12815	0.06583	0.00791	0.34831
820	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00056	0.01384	0.01049	0.00056	0.34831
820	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00056	0.01384	0.01434	0.00056	0.34831
820	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00056	0.01384	0.01338	0.00056	0.34831
823	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00881	0.14222	0.09256	0.00881	0.34831
823	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00881	0.14222	0.12183	0.00881	0.34831
823	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00881	0.14222	0.09256	0.00881	0.34831
825	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00937	0.15101	0.13453	0.00937	0.34831
825	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00937	0.15101	0.10864	0.00937	0.34831
825	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00937	0.15101	0.15007	0.00937	0.34831
826	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00791	0.12815	0.06143	0.00791	0.34831
826	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00791	0.12815	0.13176	0.00791	0.34831
826	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00791	0.12815	0.04824	0.00791	0.34831
827	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00553	0.15409	0.1267	0.00553	0.59551
827	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00553	0.15409	0.12142	0.00553	0.59551
827	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.00553	0.15409	0.11085	0.00553	0.59551
833	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.01095	0.49439	0.42328	0.01095	1
833	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.01095	0.49439	0.38941	0.01095	1
833	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.01095	0.49439	0.50796	0.01095	1
834	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.01073	0.4845	0.46461	0.01073	1
834	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.01073	0.4845	0.36502	0.01073	1
834	CÃ¡ DÄ©a	1/1/2025	2025	1	1	0.01073	0.4845	0.46461	0.01073	1
839	CÃ¡ Chim	1/1/2025	2025	1	1	0.01615	0.19234	0.12517	0.01615	0.25843
839	CÃ¡ Chim	1/1/2025	2025	1	1	0.01615	0.19234	0.11857	0.01615	0.25843

957	TÃ©p Aura	1/1/2025	2025	1	1	0.04269	0.07515	0.04119	0.04269	0.02921
957	TÃ©p Aura	1/1/2025	2025	1	1	0.04269	0.07515	0.02828	0.04269	0.02921
958	TÃ©p Drea	1/1/2025	2025	1	1	0.01231	0.02193	0.0181	0.01231	0.02921
958	TÃ©p Drea	1/1/2025	2025	1	1	0.01231	0.02193	0.02037	0.01231	0.02921
958	TÃ©p Drea	1/1/2025	2025	1	1	0.01231	0.02193	0.02265	0.01231	0.02921
959	TÃ©p Mix l	1/1/2025	2025	1	1	0.02146	0.01038	0.00752	0.02146	0
959	TÃ©p Mix l	1/1/2025	2025	1	1	0.02146	0.01038	0.00497	0.02146	0
959	TÃ©p Mix l	1/1/2025	2025	1	1	0.02146	0.01038	0.00425	0.02146	0
960	TÃ©p Ong	1/1/2025	2025	1	1	0.00689	0.0068	0.00374	0.00689	0.01124
960	TÃ©p Ong	1/1/2025	2025	1	1	0.00689	0.0068	0.00543	0.00689	0.01124
960	TÃ©p Ong	1/1/2025	2025	1	1	0.00689	0.0068	0.00422	0.00689	0.01124
961	TÃ©p Red	1/1/2025	2025	1	1	0.02101	0.01639	0.01241	0.02101	0.00674
961	TÃ©p Red	1/1/2025	2025	1	1	0.02101	0.01639	0.01298	0.02101	0.00674
961	TÃ©p Red	1/1/2025	2025	1	1	0.02101	0.01639	0.01241	0.02101	0.00674
962	TÃ©p Rilli	1/1/2025	2025	1	1	0.02202	0.06506	0.02448	0.02202	0.05618
962	TÃ©p Rilli	1/1/2025	2025	1	1	0.02202	0.06506	0.06695	0.02202	0.05618
962	TÃ©p Rilli	1/1/2025	2025	1	1	0.02202	0.06506	0.04236	0.02202	0.05618
963	TÃ©p Rilli	1/1/2025	2025	1	1	0.02021	0.01577	0.0092	0.02021	0.00674
963	TÃ©p Rilli	1/1/2025	2025	1	1	0.02021	0.01577	0.01194	0.02021	0.00674
963	TÃ©p Rilli	1/1/2025	2025	1	1	0.02021	0.01577	0.01414	0.02021	0.00674
964	TÃ©p Rilli	1/1/2025	2025	1	1	0.03794	0.02958	0.0203	0.03794	0.00674
964	TÃ©p Rilli	1/1/2025	2025	1	1	0.03794	0.02958	0.02847	0.03794	0.00674
964	TÃ©p Rilli	1/1/2025	2025	1	1	0.03794	0.02958	0.02745	0.03794	0.00674
965	TÃ©p Soc	1/1/2025	2025	1	1	0.03365	0.09902	0.09844	0.03365	0.05618
965	TÃ©p Soc	1/1/2025	2025	1	1	0.03365	0.09902	0.04747	0.03365	0.05618
965	TÃ©p Soc	1/1/2025	2025	1	1	0.03365	0.09902	0.09505	0.03365	0.05618

4.2. Khai phá dữ liệu

4.2.1. Tổng quan về khai phá dữ liệu trong kho cá cảnh

Khai phá dữ liệu (Data Mining) là quá trình phát hiện và trích xuất thông tin hữu ích từ dữ liệu thô trong các hệ thống cơ sở dữ liệu hoặc kho dữ liệu. Đối với kho dữ liệu về kinh doanh mặt hàng cá cảnh, khai phá dữ liệu được thực hiện nhằm phát hiện các xu hướng, mẫu ẩn trong dữ liệu, từ đó hỗ trợ cho các quyết định kinh doanh như tối ưu hóa giá bán và đưa ra các chương trình khuyến mãi phù hợp.

Kho dữ liệu kinh doanh mặt hàng cá cảnh được xây dựng theo mô hình Snowflake Schema. Khai phá dữ liệu giúp doanh nghiệp hiểu rõ hơn về hành vi mua sắm của khách hàng, xác định các sản phẩm tiêu thụ chậm và đưa ra chiến lược kinh doanh phù hợp.

4.2.2. Mục tiêu khai phá dữ liệu

Quá trình khai phá dữ liệu trong kho dữ liệu cá cảnh có các mục tiêu chính sau:

- Dự đoán lượng tiêu thụ – Dự đoán số lượng bán ra của từng loại cá dựa trên các yếu tố như giá bán, thời gian, doanh thu và lợi nhuận.
- Xác định sản phẩm bán chậm – Xác định các sản phẩm có số lượng tiêu thụ thấp nhưng tồn kho cao để đưa ra các chiến lược khuyến mãi.
- Tối ưu hóa giá bán – Dự đoán giá bán tối ưu để vừa tối đa hóa lợi nhuận, vừa thúc đẩy tiêu thụ sản phẩm.
- Phân tích xu hướng tiêu dùng – Phân tích các xu hướng mua sắm

theo từng thời kỳ (tháng, quý, năm).

- Hỗ trợ ra quyết định kinh doanh – Cung cấp thông tin để hỗ trợ nhà quản lý ra quyết định về giá cả và tồn kho.

4.2.3. Tổng quan về thuật toán sử dụng

a. Hồi quy tuyến tính (Linear Regression)

Hồi quy tuyến tính là một phương pháp thống kê được sử dụng để mô hình hóa mối quan hệ giữa một biến phụ thuộc yyy và một hoặc nhiều biến độc lập xxx.

Phương trình hồi quy tuyến tính được biểu diễn như sau:

$$y = w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n + \epsilon$$

Trong đó:

- Y – Giá trị cần dự đoán (Số lượng tiêu thụ)
- w0, w1, ..., wn – Các hệ số của mô hình (được huấn luyện từ dữ liệu)
- x1, x2, ..., xn – Các đặc trưng đầu vào (giá bán, tồn kho, doanh thu, lợi nhuận)
- ϵ – Sai số ngẫu nhiên

b. Hồi quy Ridge (Ridge Regression)

Hồi quy Ridge là một biến thể của hồi quy tuyến tính có thêm tham số phạt (regularization term) để giảm thiểu hiện tượng overfitting.

Phương trình Ridge Regression là:

$$J(w) = \sum_{i=1}^{n}(y_i - w_0 - \sum_{j=1}^{m} w_j x_{ij})^2 + \lambda \sum_{j=1}^{m} w_j^2$$

Trong đó:

- λ – Hệ số phạt (điều chỉnh mức độ regularization)

- Khi λ lớn → mô hình bị phạt nhiều hơn → đơn giản hóa mô hình và tránh overfitting.

c. Hồi quy Lasso (Lasso Regression)

Hồi quy Lasso cũng là một biến thể của hồi quy tuyến tính, nhưng thay vì sử dụng bình phương các trọng số, nó sử dụng trị tuyệt đối của các trọng số làm hệ số phạt:

$$J(w) = \sum_{i=1}^{n}(y_i - w_0 - \sum_{j=1}^{m} w_j x_{ij})^2 + \lambda \sum_{j=1}^{m} |w_j|$$

Hồi quy Lasso có tác dụng loại bỏ các trọng số không quan trọng, dẫn đến mô hình đơn giản hơn và dễ giải thích hơn.

4.2.4. Huấn luyện mô hình

```
import pyodbc
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns
```

Chia dữ liệu thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80/20%
Đặt các biến x là biến độc lập, các biến y là biến phụ thuộc.

```
def split_data(df):
    X = df[['InventoryQuantity', 'Price', 'Revenue', 'Profit']]
    y = df['QuantitySold']

    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.2, random_state=42
    )

    return X_train, X_test, y_train, y_test
```

(variable) y_test: Any

Xây dựng và khởi tạo mô hình hồi quy tuyến tính

```
def train_linear_regression(X_train, X_test, y_train, y_test):
    model = LinearRegression()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
```

Khởi tạo và huấn luyện mô hình Ridge Regression thêm tham số phạt để tránh overfitting.

```
def train_ridge_regression(X_train, X_test, y_train, y_test):
    model = Ridge(alpha=1.0)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
```

Khởi tạo và huấn luyện Lasso Regression thêm tham số phạt L1 để làm mất hiệu lực các tham số không quan trọng.

```
def train_lasso_regression(X_train, X_test, y_train, y_test):
    model = Lasso(alpha=0.01)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    mse = mean_squared_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
```

Để khuyến mãi cho các mặt hàng bán chậm, đầu tiên tính trung bình doanh số (Tổng số lượng bán ra trong một khoảng thời gian) và doanh thu (Tổng số tiền thu được) để làm ngưỡng so sánh. Nếu như sản phẩm nào cso số doanh thu và doanh số đều thấp hơn trung binh thì sản phẩm không bán chạy, từ đó tiến hành giảm giá khuyến mãi.

```
def apply_discount_on_unpopular_products(df, discount_rate=0.05):
    # Xác định ngưỡng trung bình
    mean_quantity = df['QuantitySold'].mean()
    mean_revenue = df['Revenue'].mean()
    mean_inventory = df['InventoryQuantity'].mean()

    # Đánh dấu các mặt hàng không bán chạy
    df['Low_Sales'] = (df['QuantitySold'] < mean_quantity) & (df['Revenue'] < mean_revenue)

    # Tính giá sau khi giảm cho các mặt hàng không bán chạy
    df['Discounted_Price'] = df['Price']
    df.loc[df['Low_Sales'], 'Discounted_Price'] = df['Price'] * (1 - discount_rate)
```

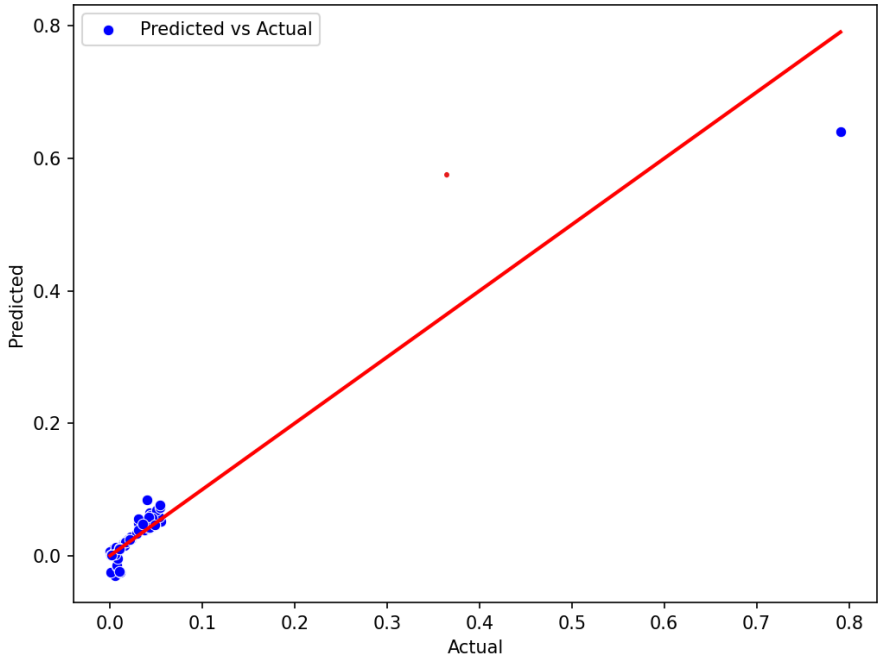
4.2.5. Đánh giá mô hình

Sau khi huấn luyện các mô hình, kết quả phân tích được cho thấy

Mô hình	MSE (Mean Squared Error)	R^2 (Hệ số xác định)	Đánh giá
Linear Regression		1	Dự đoán hoàn hảo (quá tốt để là thật) → Overfitting

Ridge Regression	0.000525	0.9418	Dự đoán rất tốt, nhưng vẫn có sai số nhỏ → Mô hình tối ưu nhất
Lasso Regression	0.00354	0.6074	Sai số lớn hơn, độ phù hợp thấp → Kém hiệu quả nhất

Ta có biểu đồ so sánh giá trị thực tế và giá trị dự đoán từ mô hình hồi quy Ridge



Dựa theo thuật toán, các mặt hàng có thể khuyến mãi được là:

FishID	FishName	SaleDate	Year	Month	Quarter	QuantitySc	Revenue	Profit	InventoryC	Price	Low_Sales	Discounted_Price
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	0.053303	0.031095	0.010615	0.053303	0.002247	FALSE	0.002247
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	0.053303	0.031095	0.01276	0.053303	0.002247	FALSE	0.002247
778	CÃ¡ Buá»“	1/1/2025	2025	1	1	0.053303	0.031095	0.023486	0.053303	0.002247	FALSE	0.002247
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	0.049012	0.028589	0.015678	0.049012	0.002247	FALSE	0.002247
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	0.049012	0.028589	0.029493	0.049012	0.002247	FALSE	0.002247
779	CÃ¡ CÃ¡nh	1/1/2025	2025	1	1	0.049012	0.028589	0.026533	0.049012	0.002247	FALSE	0.002247
781	CÃ¡ Há°¯c	1/1/2025	2025	1	1	0.042575	0.033183	0.02506	0.042575	0.006742	FALSE	0.006742
781	CÃ¡ Há°¯c	1/1/2025	2025	1	1	0.042575	0.033183	0.01133	0.042575	0.006742	FALSE	0.006742
781	CÃ¡ Há°¯c	1/1/2025	2025	1	1	0.042575	0.033183	0.019339	0.042575	0.006742	FALSE	0.006742
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	0.04585	0.026742	0.022055	0.04585	0.002247	FALSE	0.002247
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	0.04585	0.026742	0.010972	0.04585	0.002247	FALSE	0.002247
782	CÃ¡ Há»“n	1/1/2025	2025	1	1	0.04585	0.026742	0.010049	0.04585	0.002247	FALSE	0.002247
784	CÃ¡ Kim T	1/1/2025	2025	1	1	0.00847	0.010915	0.009048	0.00847	0.017978	TRUE	0.017079
784	CÃ¡ Kim T	1/1/2025	2025	1	1	0.00847	0.010915	0.00485	0.00847	0.017978	TRUE	0.017079
784	CÃ¡ Kim T	1/1/2025	2025	1	1	0.00847	0.010915	0.004087	0.00847	0.017978	TRUE	0.017079

V. Kết luận

Đề tài xây dựng kho dữ liệu kinh doanh mặt hàng cá cảnh và khai phá dữ liệu nhằm xác định giá bán cùng các chiến lược khuyến mãi cho các sản phẩm tiêu thụ chậm đã chứng minh được tính khả thi của việc áp dụng công nghệ hiện đại vào quản lý và ra quyết định kinh doanh. Qua quá trình thu thập dữ liệu từ website thương mại điện tử, thông tin từ nhà cung cấp, dữ liệu được làm sạch,

chuẩn hóa và tổ chức theo mô hình Snowflake, giúp tập trung và quản lý thông tin một cách hiệu quả.

Hệ thống kho dữ liệu được xây dựng một cách khoa học, với các bảng dữ liệu chứa thông tin chi tiết về khách hàng, sản phẩm, giao dịch và tồn kho. Qua đó, dữ liệu trở nên nhất quán, giúp cho quá trình phân tích và khai phá có nền tảng vững chắc. Các phương pháp phân tích tiên tiến như hồi quy tuyến tính, phân cụm và luật kết hợp đã được áp dụng nhằm khám phá mối quan hệ giữa giá bán, tồn kho, thời gian giao dịch và lượng tiêu thụ.

Trong tương lai, đề tài hướng đến việc tích hợp thêm các nguồn dữ liệu mới như phản hồi từ khách hàng và dữ liệu thị trường để mở rộng kho dữ liệu và làm giàu thông tin phân tích. Bên cạnh đó, việc áp dụng các thuật toán học máy tiên tiến hơn như Gradient Boosting hoặc Neural Networks sẽ được nghiên cứu để dự báo xu hướng tiêu thụ một cách chính xác hơn, từ đó hỗ trợ doanh nghiệp điều chỉnh chiến lược kinh doanh một cách linh hoạt và hiệu quả.

TÀI LIỆU THAM KHẢO

https://fr.slideshare.net/slideshow/ti-liu-data-warehouse-vietsub/42992427

https://github.com/0xl4p/Giao-Trinh-PTIT/blob/main/B%C3%A0i%20gi%E1%BA%A3ng%20Kho%20d%E1%BB%AF%20li%E1%BB%87u%20v%C3%A0%20k%E1%BB%B9%20thu%E1%BA%ADt%20khai%20ph%C3%A1.pdf

https://vi.wikipedia.org/wiki/Python_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh)

https://vi.wikipedia.org/wiki/Kho_d%E1%BB%AF_li%E1%BB%87u