**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**

**TRUNG TÂM ĐÀO TẠO SAU ĐẠI HỌC**



**TIỂU LUẬN PHÂN TÍCH DỮ LIỆU**

***KẾT THÚC HỌC PHẦN***

**Đề tài: Ứng dụng giải thuật Girvan-Newman trong phân tích xu hướng mua hàng đồng thời**

**HỌC VIÊN : VŨ ĐẶNG BẢO KHÁNH**

**VŨ NGỌC DUYỆT**

**ĐÀO MINH TÂM**

**LỚP : CAO HỌC KHÓA 14**

**HỌC PHẦN : PHÂN TÍCH DỮ LIỆU**

**GIẢNG VIÊN : NGUYỄN MẠNH CƯỜNG**

**Hà Nội, tháng 8-2024**

# LỜI CẢM ƠN:

Lời đầu tiên chúng em xin được nói lời cảm ơn đến thầy cô, nhà trường và bạn bè đã hỗ trợ chúng em trong suốt thời gian qua. Sự hỗ trợ của thầy cô và bạn bè chính là động lực để chúng em có thể xác định được hướng đi và đề tài để hoàn thành chương trình học.

Chúng em cũng muốn được cảm ơn thầy TS.Nguyễn Mạnh Cường vì sự tận tâm và chỉ bảo chúng em trong quá trình học và làm bài. Những kiến thức thầy giảng trong chương trình học đã giúp chúng em có được những nguồn thông tin và những ý tưởng để chúng em áp dụng vào đề tài.

Chúng em cũng xin được gửi lời cảm ơn đến tất cả những người đã hỗ trợ chúng em hoàn thành được đề tài này. Những ý kiến, góp ý của mọi người đã hỗ trợ chúng em trong việc nâng cao chất lượng đề tài với thời gian sớm nhất.

Cuối cùng, chúng em xin kết thúc lời cảm ơn này bằng việc cam kết sẽ tiếp tục nỗ lực, trưởng thành hơn trong việc nghiên cứu và học tập. Chúng em sẽ cố gắng vận dụng những gì đã học được trong trường và phát triển những kiến thức đã học để phát triển trong tương lai.

**Học viên**

*Vũ Đặng Bảo Khánh*

*Vũ Ngọc Duyệt*

*Đào Minh Tâm*

# LỜI MỞ ĐẦU

Trong lĩnh vực bán hàng, việc dự đoán được nhu cầu và đáp ứng đúng những gì mà khách hàng đang cần là vấn đề quan trọng. Việc sử dụng các phương pháp áp dụng trí tuệ nhân tạo được rất nhiều nơi sử dụng để phân tích những dự đoán từ dữ liệu đã có sẵn để dự đoán cho tương lai. Trong đó có phương pháp “Giải thuật Girvan-Newman” được chúng em sử dụng trong đề tài này để Phân tích Xu hướng mua hàng đồng thời.

Để đảm bảo kết quả của đề tài nghiên cứu, bài báo cáo này được chia thành các phần như sau:

**Chương 1: Tổng quan về đề tài**

Trình bày tổng quan về đề tài, trình bày khái quát về phân tích dữ liệu và bài toán dự báo, đồng thời nêu ra tình hình nghiên cứu bài toán dự báo cả ở trong nước và nước ngoài. Thêm nữa còn mô tả bài toán phân tích dữ liệu và phân tích xu hướng mua hàng đồng thời bằng Giải thuật Girvan-Newman.

**Chương 2: Các phương pháp kỹ thuật**

Trình bày các phương pháp kỹ thuật, cụ thể là phương pháp phân tích mô tả, giải thuật Girvan-Newman và các công cụ thực hiện bài toán. Đồng thời cũng lựa chọn được ngôn ngữ Python để thực hiện thực nghiệm.

**Chương 3: Thực nghiệm và đánh giá**

Trình bày thực nghiệm và những đánh giá qua các bước đầy đủ từ tiền xử lý dữ liệu cho đến phân tích mô tả và bài toán dự báo. Từ đó đưa ra các đánh giá và những đề xuất phù hợp để cải thiện kết quả của đề tài trong tương lai.

**Chương 4: Xây dựng sản phẩm Demo**

Trình bày các quy trình để xây dựng được sản phẩm Demo, các công cụ, hiển thị chức năng của sản phẩm Demo.

**Kết luận:**

Tổng hợp lại đã làm được những gì trong quá trình làm bài tiểu luận, nêu ra kiến thức đã học được, định hướng được hướng phát triển đề tài để áp dụng vào thực tế.

# CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

* 1. **Tổng quan về phân tích dữ liệu**
     1. **Phân tích dữ liệu là gì?**

Phân tích dữ liệu tập trung vào việc rút ra những hiểu biết sâu sắc từ dữ liệu cần phân tích. Gồm các quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, đưa ra kết luận và hỗ trợ việc ra quyết định.

Các ngành nghề như Khoa học dữ liệu, kinh doanh, tài chính, y tế, marketing, hành chính công và nhiều lĩnh vực khác đều được sử dụng phân tích dữ liệu. Phân tích dữ liệu giúp lập kế hoạch và đưa ra quyết định liên quan đến kinh doanh và nghiên cứu.

Phân tích dữ liệu có thể bao gồm khai phá dữ liệu, thuật toán thống kê, học máy (machine learning), trực quan hóa dữ liệu, mô hình hóa và các phương pháp khác.

* + 1. **Quy trình phân tích dữ liệu:**

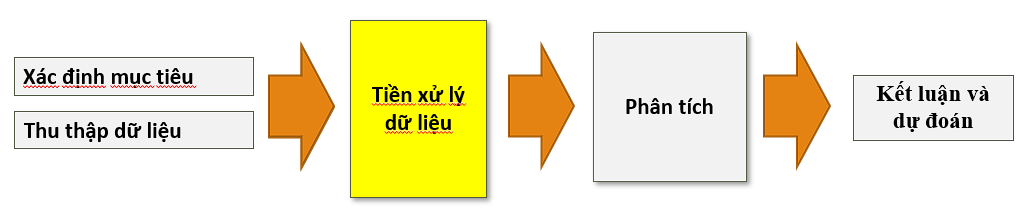


Figure 1: Quy trình phân tích dữ liệu

Quy trình phân tích dữ liệu thường bao gồm các bước chính:

**Xác định mục tiêu và thu thập dữ liệu:**

* Xác định mục tiêu: là những kết quả cụ thể mà chúng ta mong muốn đạt được thông qua việc xử lý và phân tích dữ liệu mà chúng ta đã thu thập. Mục tiêu này xác định hướng đi và phạm vi của quá trình phân tích, giúp chúng ta tập trung vào việc thu thập thông tin quan trọng và thực hiện các phần để đáp ứng các yêu cầu cụ thể.
* Thu thập dữ liệu: là thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, trang web và nhiều nguồn khác. Dữ liệu có thể là hình ảnh, âm thanh, văn bản hoặc số liệu.

**Tiền xử lý dữ liệu:** Dữ liệu thường không chính xác và có thể chứa nhiễu, thiếu hoặc không chính xác. Tiền xử lý dữ liệu là việc chuẩn bị dữ liệu cho phân tích bằng cách tóm lược, làm sạch, tích hợp, chuyển đổi, rút gọn và rời rạc hóa dữ liệu.

**Phân tích dữ liệu:** Điều quan trọng này đòi hỏi kiến thức và phương pháp phân tích để xác định các mối liên hệ và thông tin ẩn sau dữ liệu. Phân tích dữ liệu có thể được thực hiện bằng nhiều phương pháp khác nhau, bao gồm phân tích mô tả, phân tích hồi quy, phân tích sự khác biệt, thống kê, học máy học và lấy dữ liệu.

**Kết luận và dự đoán:** Dựa trên phân tích và thông tin từ dữ liệu, chúng ta có thể rút ra kết luận, hiểu rõ hơn về tình hình, và thậm chí đưa ra dự đoán cho tương lai.

* 1. **Tổng quan về bài toán**

**1.2.1 Phân tích xu hướng mua hàng đồng thời là gì?**

Phân tích xu hướng mua hàng đồng thời, là quá trình nghiên cứu và xác định các mẫu hình hoặc khuynh hướng trong hành vi mua sắm của khách hàng khi họ mua nhiều sản phẩm hoặc dịch vụ cùng một lúc hoặc trong cùng một khoảng thời gian. Mục tiêu của phân tích này là tăng cường hiểu biết của khách hàng về sự liên kết giữa các sản phẩm và dịch vụ họ mua. Điều này sẽ giúp các công ty tìm ra các chiến lược bán hàng tối ưu, bao gồm bán chéo và bán tăng.

Dưới đây là các bước và yếu tố quan trọng trong phân tích xu hướng mua hàng đồng thời:

* **Thu thập dữ liệu giao dịch**: Dữ liệu này có thể bao gồm thông tin về thời gian và số lượng cũng như danh sách các sản phẩm và dịch vụ khác nhau đã được mua, cũng như giá trị giao dịch và thông tin khách hàng.
* **Xử lý và làm sạch dữ liệu**: Loại bỏ các dữ liệu bị lỗi hoặc không hợp lệ, chuẩn hóa các thông tin để dễ dàng phân tích.
* **Phân tích dữ liệu**: Sử dụng các phương pháp phân tích như:
* **Phân tích giỏ hàng (Market Basket Analysis):** Phương pháp này hỗ trợ trong việc xác định mối quan hệ giữa các sản phẩm trong giỏ hàng của người tiêu dùng. Thuật toán Apriori, giúp tìm ra các tập hợp sản phẩm phổ biến, là một ví dụ.
* **Phân tích tần suất mua hàng**: Xác định các sản phẩm hoặc dịch vụ nào được mua cùng nhau thường xuyên nhất.
* **Phân tích phân khúc khách hàng (Customer Segmentation Analysis)**: Để hiểu rõ hơn về từng phân khúc và phát triển các chiến lược bán hàng phù hợp, phân loại khách hàng theo mẫu hình mua sắm của họ.
* **Xác định các yếu tố ảnh hưởng**: Các yếu tố này có thể bao gồm:
* **Mối liên hệ giữa các sản phẩm**: Các sản phẩm bổ sung hoặc tương hỗ.
* **Thời điểm mua hàng**: Các dịp lễ, mùa vụ hoặc thời gian trong ngày.
* **Đặc điểm khách hàng**: Tuổi, giới tính, sở thích, thu nhập, …
* **Chiến lược giá cả và khuyến mãi**: Các chương trình giảm giá, khuyến mãi, gói sản phẩm, …
* **Đưa ra các chiến lược bán hàng**:
* **Cross-selling**: Đề xuất các sản phẩm bổ sung mà khách hàng có thể quan tâm dựa trên sản phẩm họ đã mua.
* **Up-selling**: Đề xuất các phiên bản cao cấp hơn của sản phẩm mà khách hàng đang quan tâm.
* **Thiết kế gói sản phẩm**: Kết hợp các sản phẩm thường được mua cùng nhau thành các gói sản phẩm để tăng giá trị giao dịch.
* **Đo lường và đánh giá hiệu quả**: Theo dõi các chỉ số hiệu suất như tỷ lệ chuyển đổi, tăng trưởng doanh số và giá trị giỏ hàng trung bình và thực hiện các điều chỉnh để tối ưu hóa chiến lược.

**1.2.2 Tình hình phát triển việc Phân tích xu hướng mua hàng đồng thời ở Việt Nam:**

Phân tích xu hướng mua hàng đồng thời ở Việt Nam đã và đang phát triển mạnh mẽ trong bối cảnh sự bùng nổ của thương mại điện tử và công nghệ số. Dưới đây là một số điểm nổi bật về tình hình phát triển của lĩnh vực này:

* **Tăng trưởng thương mại điện tử**: Với nhiều nền tảng lớn như Shopee, Lazada, Tiki và Sendo, thương mại điện tử tại Việt Nam đang phát triển nhanh chóng. Những nền tảng này hiểu rõ hơn về hành vi mua sắm của người tiêu dùng thông qua phân tích dữ liệu. Điều này giúp tối ưu hóa các chiến lược bán hàng và quảng cáo.
* **Ứng dụng công nghệ và dữ liệu lớn (Big Data)**: Các doanh nghiệp tại Việt Nam đang phát triển nhận thức về tầm quan trọng của việc sử dụng dữ liệu và công nghệ để phân tích hành vi mua sắm của người tiêu dùng. Các doanh nghiệp có thể đồng thời xác định các xu hướng mua hàng bằng cách sử dụng công nghệ dữ liệu lớn và trí tuệ nhân tạo (AI).
* **Phân tích giỏ hàng (Market Basket Analysis)**: Các siêu thị, cửa hàng bán lẻ và các nền tảng thương mại điện tử đều sử dụng phương pháp này. Doanh nghiệp có thể đưa ra các đề xuất sản phẩm phù hợp, cải thiện các chiến lược bán chéo và bán tăng bằng cách phân tích các sản phẩm thường được mua cùng nhau.
* **Tối ưu hóa chiến lược marketing và khuyến mãi**: Các doanh nghiệp sử dụng phân tích xu hướng mua hàng đồng thời để tạo ra các chương trình khuyến mãi hấp dẫn, kết hợp các hàng hóa thường được mua thành các gói hàng hóa hoặc tạo ra các chương trình giảm giá theo thời gian nhằm thu hút khách hàng.
* **Sự phát triển của các công ty công nghệ và dịch vụ phân tích dữ liệu**: Nhiều công ty công nghệ tại Việt Nam cung cấp các giải pháp phân tích dữ liệu và hành vi người tiêu dùng, giúp các công ty dễ dàng áp dụng các phương pháp phân tích hiện đại vào hoạt động kinh doanh của họ. Các công ty công nghệ dữ liệu như FPT, VNG và các startup đang đóng vai trò quan trọng trong lĩnh vực này.
* **Nhận thức và đào tạo**: Tại Việt Nam, các trường đại học và các tổ chức đào tạo đã bắt đầu cung cấp các khóa học và chương trình đào tạo về phân tích dữ liệu, giúp tạo ra nguồn nhân lực chất lượng cao trong lĩnh vực này.
* **Chuyển đổi số trong ngành bán lẻ**: Các doanh nghiệp bán lẻ truyền thống cũng đang chuyển sang số bằng cách sử dụng công nghệ và phân tích dữ liệu để cạnh tranh và đáp ứng nhu cầu của người tiêu dùng ngày càng tăng.

Tóm lại, tăng trưởng thương mại điện tử ở Việt Nam, ứng dụng công nghệ mới và sự nhận thức ngày càng cao của doanh nghiệp về tầm quan trọng của dữ liệu đã khiến phân tích xu hướng mua hàng đồng thời trở nên phổ biến. Trải nghiệm mua sắm tốt hơn cho người tiêu dùng cũng được cải thiện do điều này.

**1.2.3 Tình hình phát triển việc Phân tích xu hướng mua hàng đồng thời trên thế giới:**

Sự phát triển nhanh chóng của công nghệ, bao gồm dữ liệu lớn (Big Data) và trí tuệ nhân tạo (AI), đã thúc đẩy phân tích xu hướng mua hàng trên toàn cầu. Dưới đây là một số điểm nổi bật về tình hình hiện tại liên quan đến phân tích xu hướng mua hàng toàn cầu:

* **Sự bùng nổ của thương mại điện tử:** Các nền tảng thương mại điện tử lớn như Amazon, Alibaba, eBay và các dịch vụ giao hàng như Instacart sử dụng công nghệ phân tích dữ liệu mới để hiểu rõ hơn về hành vi mua sắm của khách hàng và tối ưu hóa trải nghiệm người dùng.
* **Ứng dụng AI và Machine Learning:** Các doanh nghiệp đang sử dụng máy học và trí tuệ nhân tạo để phân tích dữ liệu mua hàng. Các dự đoán và đề xuất sản phẩm được thiết kế riêng cho từng khách hàng được thực hiện bằng cách sử dụng các thuật toán học máy để xác định các mẫu hình phức tạp trong dữ liệu mua sắm.
* **Phân tích giỏ hàng (Market Basket Analysis):** Rộng rãi trên toàn cầu, phương pháp này được sử dụng để xác định các sản phẩm thường được mua cùng nhau. Các công ty bán lẻ và thương mại điện tử sử dụng dữ liệu này để phát triển các chiến lược bán tăng và bán chéo hiệu quả.
* **Phân tích dữ liệu thời gian thực:** Sự gia tăng của công nghệ đám mây và Internet of Things (IoT) đã cho phép các doanh nghiệp thu thập và phân tích thông tin mua sắm trong thời gian thực. Họ có thể nhanh chóng thay đổi hành vi mua sắm để tối ưu hóa chiến lược bán hàng.
* **Phân tích tâm lý người tiêu dùng:** Các doanh nghiệp có thể hiểu rõ hơn về tâm lý và động cơ mua sắm của người tiêu dùng bằng cách sử dụng dữ liệu từ hành vi trực tuyến, đánh giá sản phẩm và dữ liệu từ mạng xã hội. Điều này hỗ trợ họ trong việc phát triển các chiến dịch tiếp thị phù hợp và tăng cường sự tương tác với khách hàng.
* **Sự phát triển của công nghệ dữ liệu lớn (Big Data):** Các công nghệ xử lý dữ liệu lớn như Hadoop và Spark cho phép các công ty xử lý và phân tích một lượng lớn dữ liệu từ nhiều nguồn khác nhau, giúp họ tìm ra các xu hướng và mẫu hình mua sắm quan trọng.
* **Các công cụ phân tích nâng cao:** Các công cụ và nền tảng phân tích như Tableau, Google Analytics, Adobe Analytics và SAS được sử dụng rộng rãi để thu thập, phân tích và trình bày dữ liệu mua sắm. Điều này cho phép các công ty hiểu rõ hơn về hành vi người tiêu dùng và tìm ra cách tối ưu hóa chiến lược kinh doanh của họ.
* **Chuyển đổi số trong ngành bán lẻ:** Các công ty bán lẻ truyền thống đang chuyển sang chuyển đổi số mạnh mẽ bằng cách sử dụng các công nghệ phân tích dữ liệu để đáp ứng nhu cầu ngày càng tăng của khách hàng và cạnh tranh với các đối thủ thương mại điện tử.
* **Đào tạo và phát triển nguồn nhân lực:** Các chương trình đào tạo và khóa học về phân tích dữ liệu và khoa học dữ liệu đang phát triển mạnh mẽ trên toàn cầu tại các trường đại học và tổ chức giáo dục. Điều này giúp tạo ra nguồn nhân lực trong lĩnh vực này

Nhìn chung, sự phát triển toàn cầu của phân tích xu hướng mua hàng đang được thúc đẩy bởi sự phát triển của thương mại điện tử và tiến bộ công nghệ. Điều này không chỉ giúp các công ty hoạt động tốt hơn mà còn mang lại cho người tiêu dùng trải nghiệm mua sắm tốt hơn.

**1.2.4 Bài toán Phân tích xu hướng mua hàng đồng thời bằng giải thuật Girvan-Newman:**

# CHƯƠNG 2: CÁC PHƯƠNG PHÁP KỸ THUẬT

**2.1 Phương pháp phân tích mô tả**

**2.1.1 Phân tích mô tả**

Trong lĩnh vực thống kê và phân tích dữ liệu, phân tích mô tả là một phương pháp nhằm mục đích cung cấp một mô tả ngắn gọn và dễ hiểu về các đặc điểm chính của một tập dữ liệu. Mục tiêu của phân tích mô tả là giúp chúng ta hiểu rõ hơn về dữ liệu mà chúng ta đang làm việc, nhận ra các đặc điểm quan trọng của nó và cung cấp một cái nhìn toàn diện về cách dữ liệu phân phối và biến đổi.

Phân tích mô tả thường bao gồm các khía cạnh sau:

* **Thống kê tóm tắt**: Các số liệu thống kê cơ bản như trung bình, trung vị, độ lệch chuẩn và phân vị được cung cấp ở đây. Hiểu biết về dữ liệu phân tán và trung tâm được cải thiện nhờ các số liệu này.
* **Biểu đồ**: Dữ liệu trực quan thường được trình bày bằng biểu đồ. Các biểu đồ, chẳng hạn như biểu đồ cột, biểu đồ đường, biểu đồ hình tròn và biểu đồ hộp, cung cấp cho chúng ta cơ hội xem xét sự phân bố và xu hướng của dữ liệu.
* **Phân phối dữ liệu**: Phân tích phân phối dữ liệu giúp chúng ta hiểu được tỷ lệ các giá trị khác nhau xuất hiện trong một tập dữ liệu. Xây dựng biểu đồ kernel density hoặc biểu đồ phân phối tần số có thể được sử dụng để đạt được điều này.
* **Kiểm tra sự tương quan**: Việc đánh giá sự tương quan giữa các biến cũng có thể là một phần của phân tích mô tả. Điều này có thể được thực hiện bằng cách sử dụng tính toán hệ số tương quan Pearson hoặc biểu đồ tương quan.
* **Xác định điểm ngoại lệ**: Những giá trị rất khác biệt trong dữ liệu có thể được tìm thấy thông qua phân tích mô tả.
* **Tổng kết và nhận xét**: Cuối cùng, phân tích mô tả thường bao gồm tổng kết và nhận xét về các đặc điểm quan trọng, mẫu thú vị, ưu và nhược điểm của tập dữ liệu.

Phân tích mô tả tạo nền tảng cho các phân tích tiếp theo như dự báo, phân tích hồi quy và học máy. Nó cũng cung cấp một cái nhìn sâu hơn về tập dữ liệu ban đầu.

**2.1.2 Phương pháp phân tích trên từng biến**

Mục tiêu chính khi thực hiện phân tích trên một biến hoặc một thuộc tính là hiểu rõ các đặc điểm cơ bản của biến đó. Xác định và xử lý các giá trị ngoại lai hoặc bất thường thường là một phần của điều này. Các giá trị này rất khác với hầu hết các giá trị khác trong tập dữ liệu. Các giá trị ngoại lai có thể là kết quả của lỗi nhập liệu, lỗi đo lường hoặc đơn giản là sự kiện hiếm gặp.

Bởi vì chúng ta có thể làm sạch các giá trị này trong giai đoạn tiền xử lý dữ liệu của phân tích hồi quy, việc xác định các outliers là một yếu tố quan trọng giúp kết nối phân tích mô tả với phân tích hồi quy. Các loại dữ liệu sau đây sẽ được phân tích cụ thể:

**Dữ liệu số:**

* **Biểu đồ Histogram:** Biểu đồ hiển thị tần suất xuất hiện của các khoảng giá trị dữ liệu.
* **Các đại lượng thống kê:** Bao gồm mean (trung bình), stdev (độ lệch chuẩn), median (trung vị), quartile (phân vị) ... Các giá trị này giúp mô tả trung bình, phương sai và phân phối của dữ liệu.
* **Biểu đồ Box & Whisker (Boxplot):** Biểu đồ hiển thị tổng quan giá trị đó bao gồm các giá trị đại lượng thống kê đã tính được.

**Dữ liệu phi số:**

* **Bảng tần suất (Frequency table):** Biểu đồ liệt kê các giá trị khác nhau của biến và số lần xuất hiện của mỗi giá trị.
* **Biểu đồ cột (Bar chart):** Biểu đồ thể hiện tần suất của từng giá trị dữ liệu dưới dạng các cột đứng.
* **Biểu đồ hình tròn hoặc Donut (Pie chart, Donut chart):** Biểu đồ thể hiện phần trăm tần suất của từng giá trị trong tổng số.

**2.1.3 Phương pháp phân tích trên nhiều biến**

Phân tích nhiều biến giúp bạn hiểu được các mối quan hệ và tương tác trong tập dữ liệu. Điều này có thể giúp bạn tìm ra các tương quan, xu hướng hoặc mẫu có thể tồn tại giữa chúng.

Mối liên hệ tuyến tính, không tuyến tính hoặc ngược đã được gọi là các mối liên hệ giữa các biến. Có thể xem xét và xác định cách các biến tác động và ảnh hưởng lẫn nhau trong mỗi mối liên hệ.

Phân tích hồi quy giúp xác định các giá trị ngoại lai của dữ liệu, vì vậy phân tích trên nhiều biến cũng rất quan trọng. Có ba loại dữ liệu phân tích khác nhau do phân tích nhiều biến: số, phi số và hỗn hợp (cả số và phi số):

**Dữ liệu số:**

* **Scatter Plot (Biểu đồ Scatter):** Biểu đồ thể hiện mối quan hệ giữa hai biến số. Mỗi điểm trên biểu đồ thể hiện một cặp giá trị của hai biến trên trục ngang và dọc. Biểu đồ này dùng để tìm kiếm sự tương quan giữa 2 biến số như tương quan tuyến tính hoặc không tuyến tính.
* **Bảng dữ liệu thống kê (Statistical Summary Table):** Tạo bảng để liệt kê các đại lượng thống kê (mean, median, stdev…) giữa các biến số của dữ liệu.

**Dữ liệu phi số:**

* **Bảng dữ liệu thống kê (Statistical Summary Table):** Cũng là bảng dữ liệu thống kê nhưng với giá trị phi số, đó sẽ chỉ có giá trị tần suất xuất hiện (mode) của dữ liệu.

**Dữ liệu hỗn hợp:**

* **Bảng thống kê tổng hợp:** Đây là sự kết hợp giữa bảng dữ liệu thống kê của dữ liệu số và phi số. Sự kết hợp tổng quan này sẽ cho ta bao quát được phân bổ của dữ liệu.
* **Biểu đồ Box & Whisker (Boxplot):** Được sử dụng để so sánh phân phối của một dữ liệu số với tần suất của một dữ liệu phi số. Biểu đồ này sẽ cho ta mối quan hệ mật thiết về sự ảnh hưởng của các giá trị phi số lên giá trị số được phân tích.

**2.2 Phương pháp phân tích hồi quy**

**2.2.1 Tổng quan về phân tích hồi quy**

Phân tích hồi quy là một tập hợp các chiến lược thống kê được sử dụng để xác định cách một hoặc nhiều biến độc lập tương tác với một biến phụ thuộc. Nó có thể được sử dụng để đánh giá sức mạnh của cách các biến tương tác với nhau và để mô hình hóa cách chúng sẽ tương tác trong tương lai.

Phân tích hồi quy là một phương pháp phân loại toán học để xác định biến nào có tác động thực sự. Nó cung cấp câu trả lời cho các câu hỏi như: yếu tố nào là quan trọng nhất? Cái nào không nên bỏ qua? Những thành phần đó tương tác với nhau như thế nào? Và quan trọng nhất, chúng ta có chắc chắn về tất cả những yếu tố này như thế nào?

Phân tích hồi quy đòi hỏi phải xác định một biến phụ thuộc, tức là yếu tố quan trọng mà chúng ta đang cố gắng hiểu hoặc dự đoán. Một số biến thể, bao gồm tuyến tính, nhiều tuyến tính và phi tuyến tính, được xem xét trong phân tích hồi quy. Mô hình tuyến tính và nhiều tuyến tính phổ biến nhất. Chúng thường được sử dụng trong phân tích hồi quy phi tuyến cho các tập dữ liệu phức tạp hơn trong đó các biến phụ thuộc và độc lập thể hiện các mối quan hệ phi tuyến.

**2.2.2 Các phương pháp phân tích hồi quy**

Để phân tích hồi quy có rất nhiều phương pháp để phân tích. Dưới đây sẽ là một số phương pháp quan trọng dùng để phân tích hồi quy:

* **Hồi quy tuyến tính (Linear Regression):** Hồi quy tuyến tính tìm đường thẳng "tốt nhất" qua dữ liệu để dự đoán giá trị mục tiêu dựa trên biến độc lập. Phương pháp này đơn giản và phù hợp với dữ liệu tuyến tính. Nhưng nó có thể không xử lý được dữ liệu phi tuyến và bị nhiễu dữ liệu ảnh hưởng.
* **Hồi quy Ridge (Ridge Regression):** Hồi quy Ridge, thêm hệ số điều chuẩn l2 vào hàm mất mát, là phiên bản cải tiến của hồi quy tuyến tính. Điều này có lợi cho việc kiểm soát độ phức tạp của mô hình và ngăn chặn tình trạng quá khớp. Mặc dù xử lý đa cộng tuyến và giảm overfitting là ưu điểm, nhưng cần lựa chọn tham số điều chuẩn chính xác.
* **Hồi quy Lasso (Lasso Regression):** Ngoài ra, hồi quy Lasso là một cải tiến từ hồi quy tuyến tính; tuy nhiên, nó sử dụng hệ số điều chuẩn l1 để thúc đẩy một số hệ số về 0 thay vì l2. Điều này dẫn đến việc giảm biến quan trọng và chọn biến tự động. Mặc dù Lasso có thể "chọn biến", nhưng nó cần có các tham số điều chuẩn chính xác.

**2.2.3 Khái quát về các thuật toán phát hiện cộng đồng**

Santo Fortunato, đã tổng hợp các phương pháp/thuật toán phát hiện cộng đồng trong mạng xã hội . (Fortunato, 2009)

Theo tác giả, các nhóm phương pháp điển hình là:

* Các phương pháp truyền thống
* Các thuật toán chia
* Các phương pháp dựa trên mô đun hóa
* Các thuật toán dựa trên phổ
* Các thuật toán động
* Các phương pháp dựa trên suy luận thống kê
* Các phương pháp phát hiện chồng chéo cộng đồng
* Các phương pháp nhiều lời giải và hệ thống phân cụm phân cấp
* Các phương pháp phát hiện cộng đồng động

Các phương pháp truyền thống bao gồm 4 phương pháp, đó là **phân vùng đồ thị**, **phân cụm phân cấp**, **phân cụm theo vùng** và **phân cụm theo phổ**. Nội dung tóm lược của từng phương pháp như sau:

* Mục tiêu của phương pháp phân vùng đồ thị là tối thiểu hóa đồ thị thành các nhóm có kích thước cụ thể.Phương pháp này cần xác định số cụm được phân chia thành và kích thước của mỗi cụm.Tính toán song song, phân vùng mạch và một số giải thuật nối tiếp sử dụng phương pháp phân vùng đồ thị.
* Khi mạng xã hội được phân cấp, phương pháp phân cụm phân cấp thường được sử dụng. Điều này có nghĩa là mạng xã hội được chia làm nhiều cộng đồng con, trong đó mỗi cộng đồng con được chia ra làm nhiều cộng đồng con khác. Ý tưởng cơ bản của thuật toán phân cụm phân cấp là tìm các đỉnh trong đồ thị 13 mạng có độ tương tự cao và sau đó xếp những đỉnh có độ tương tự cao vào cùng một nhóm.Phân tích mạng xã hội, sinh học, kỹ thuật, tiếp thị, v.v. sử dụng phổ biến phương pháp phân cụm phân cấp.
* Phương pháp phân cụm theo vùng dựa trên khái niệm sau: biểu diễn đồ thị trong một không gian metric sao cho mỗi đỉnh của nó được biểu diễn bằng một điểm trong không gian đó, sau đó định nghĩa trước một số k là số lượng các cụm. Sau đó, sự khác nhau giữa các đỉnh trong đồ thị được đo bằng cách tính khoảng cách giữa các điểm trong không gian.Mục tiêu của phương pháp là phân chia không gian trên thành k cụm các điểm sao cho hàm chi phí dựa trên khoảng cách giữa các điểm trong cụm và tâm của cụm là lớn nhất hoặc nhỏ nhất.Cụm trong các tập điểm dữ liệu được xác định bằng cách sử dụng phương pháp phân cụm theo vùng.
* Tất cả các phương pháp và kỹ thuật để chia tập đối tượng thành các cụm sử dụng vector riêng của ma trận được định nghĩa từ tập đối tượng đó được sử dụng trong phương pháp phân cụm theo phổ.Đối tượng này có thể là các đỉnh của đồ thị hoặc các điểm của một không gian metric.Trong phương pháp phân cụm, các tập đối tượng được chuyển thành các điểm trong một không gian, được coi là các thành phần của vector đặc trưng. Sau đó, các điểm này được phân cụm dựa trên các phương pháp chuNn, chẳng hạn như phương pháp phân cụm k-mean.

Mục tiêu cơ bản của các phương pháp sử dụng thuật toán phân chia là xác định các cạnh nối giữa các đỉnh của các cộng đồng khác nhau. Sau đó, chúng được loại bỏ khỏi đồ thị.Các cụm trong đồ thị sẽ được kết nối với nhau như vậy, điều này cho phép chúng ta chia đồ thị thành các cộng đồng.Điểm mấu chốt của phương pháp này là xác định và loại bỏ các đặc điểm nào đó của các cạnh nối các cộng đồng được hiển thị trong đồ thị. Phương pháp sử dụng thuật toán chia có thể được coi là một kiểu thuật toán phân cụm phân cấp. Nó chỉ khác với việc tìm các cạnh có độ tương đồng cao để ghép các đỉnh của các cạnh đó thành các cộng đồng, và thay vào đó, nó tìm cách loại bỏ các cạnh nối giữa các cộng đồng đểDo đó, các mô hình phân cấp có thể được tạo thành từ kết quả của các thuật toán chia.

Thuật toán điển hình nhất trong các thuật toán chia này là thuật toán **Girvan-Newman** (Hiếu, 2011). Thuật toán này có ý nghĩa lịch sử (đánh dấu bắt đầu một thời kỳ mới trong sự phát triển của phát hiện cộng đồng) và phương pháp, vì vậy nó rất phổ biến. Do có nhiều lợi thế, thuật toán **Girvan-Newman** được sử dụng rộng rãi và sau này đã được phát triển thành nhiều thuật toán khác, tạo thành 14 họ thuật toán **Girvan-Newman**. Do đó, khóa luận này tập trung vào nghiên cứu các thuật toán GN được sử dụng để phát hiện cộng đồng trên Twitter, mạng xã hội trực tuyến.

**2.2.4 Thuật toán Girvan-Newman**

Các yêu cầu cơ bản của một thuật toán phát hiện cộng đồng nói chung là nó phải chỉ ra một phép phân chia tự nhiên giữa các đỉnh. Thuật toán không cần chỉ ra các đường đi rõ ràng như trong thuật toán phân cụm phân cấp thông thường, như đã nêu ở trên.Girvan và Newman đã đưa ra một cách để đáp ứng các mục tiêu này, trong đó có 3 tính chất như sau (Fortunato, 2009):

* Đây là một phương pháp chia nhỏ, trong đó các cạnh được loại bỏ dần dần ra khỏi đồ thị, trái ngược với phương pháp phân cụm phân cấp thông thường.
* Các cạnh cần loại bỏ trong mỗi bước được xác định qua một độ đo trung gian
* Sau khi loại bỏ mỗi cạnh, độ đo trung gian của các cạnh còn lại được tính lại dựa vào cạnh bị loại bỏ đó.

Thuật toán này được thiết kế để tìm và loại bỏ các cạnh kết nối các cộng đồng trong mạng; những cạnh này hoạt động như các nút giao thông trong mạng.Một đường đi giữa hai đỉnh của một mạng thuộc hai cộng đồng khác nhau nhất thiết phải đi qua ít nhất một trong những cạnh này. Do đó, nếu chúng ta thiết lập các đường đi giữa tất cả các đỉnh trong mạng và xác định cạnh nào trong đồ thị được các đường đi này đi qua nhiều nhất, ta có thể loại bỏ nó để phân chia mạng thành các cộng đồng riêng biệt.

Độ đo trung gian giữa các cạnh được Girvan và Newman sử dụng để xác định cạnh nào có lưu lượng giao thông nhiều nhất trong mạng theo ý tưởng được trình bày ở trên. Độ đo này được phát triển từ khái niệm độ đo trung gian giữa các đỉnh rất phổ biến do Freeman đề xuất. Tổng số đường đi trắc địa giữa các cặp đỉnh trong toàn bộ đồ thị mà đi qua cạnh được xét tới được gọi là độ đo trung gian giữa các cạnh."Đường đi ngắn nhất giữa hai đỉnh" là một loại đường đi trắc địa được sử dụng.Trong trường hợp đồ 15 có thị m cạnh và n đỉnh, thời gian tính toán cho giai đoạn này là O(mn).Độ đo trung gian của một cạnh có trọng số đơn giản có thể được xác định bằng cách sử dụng độ đo trung gian của một cạnh không có trọng số chia cho trọng số của cạnh đó trên đồ thị có trọng số.

Thuật toán Girvan và Newman rất đơn giản; mục đích chính của chúng chỉ là tìm cạnh có độ trung gian lớn nhất và loại bỏ nó. Sau đó, độ trung gian được tính lại và các cạnh tiếp tục được loại bỏ cho đến khi không còn cạnh nào trong đồ thị.Nếu có hai cạnh có độ trung gian như nhau, chúng ta có thể loại bỏ một trong số chúng hoặc cả hai cạnh cùng một lúc. Một dendrogram có thể được sử dụng để hiển thị toàn bộ thuật toán. Trong một dendrogram, thuật toán có thể được hiểu từ gốc đến các lá.Các nhánh của cây biểu diễn cho phép chia đồ thị thành các cộng đồng riêng rẽ bằng cách loại bỏ các cạnh.

Mặc dù có hai nhược điểm chính, nhưng thuật toán Girvan và Newman hoạt động tương đối tốt trong nhiều trường hợp.Đầu tiên, thuật toán Girvan-Newman không xác định được số lượng cộng đồng mà mạng sẽ phân chia ra. Điều này là do có quá nhiều phép phân vùng, vì vậy khó có thể xác định được phép phân vùng nào mang lại hiệu quả tốt nhất. Thứ hai, thuật toán có độ phức tạp thời gian cao hơn so với tổng số m cạnh cần loại bỏ. Trong mỗi bước lặp, nó cần thời gian tính toán O(mn) và tổng thời gian tính toán là ( ) 2 O m n hoặc có thể là ( ) 3 O m n [3]. Ngoài ra, thuật toán Girvan và Newman không giải quyết được sự chồng chéo cộng đồng, điều này có nghĩa là một đỉnh có thể thuộc vào một trong nhiều cộng đồng khác nhau trong mạng.

**2.3 Các công cụ hỗ trợ thực hiện bài toán**

**2.3.1 Python**

**2.4 Kết luận chương 2**

# CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

## Dữ liệu thực nghiệm

Trong project này, bộ dữ liệu được phân tích ở đây là file dataset (.csv) có tên là housing.csv chứa 20640 dòng thông tin về giá nhà California.

Link dataset: <https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset/data>

Cụ thể thông tin như sau:

* **Customer ID (Mã Khách Hàng)**: Một định danh duy nhất được gán cho từng khách hàng, giúp theo dõi và phân tích hành vi mua sắm của họ theo thời gian.
* **Age (Tuổi)**: Tuổi của khách hàng, cung cấp thông tin về nhân khẩu học để phân khúc thị trường và triển khai các chiến lược tiếp thị mục tiêu.
* **Gender (Giới Tính)**: Nhận diện giới tính của khách hàng, là một biến nhân khẩu học chính ảnh hưởng đến sở thích sản phẩm và mẫu mua hàng.
* **Item Purchased (Sản Phẩm Mua)**: Sản phẩm hoặc mặt hàng cụ thể mà khách hàng đã chọn trong giao dịch.
* **Category (Danh Mục)**: Phân loại rộng hoặc nhóm mà sản phẩm mua thuộc về (ví dụ: quần áo, điện tử, thực phẩm).
* **Purchase Amount (USD) (Số Tiền Mua Hàng (USD))**: Giá trị tiền tệ của giao dịch, tính bằng Đô la Mỹ (USD), cho biết chi phí của sản phẩm (hoặc các sản phẩm) đã mua.
* **Location (Vị Trí)**: Vị trí địa lý nơi giao dịch được thực hiện, cung cấp thông tin về sở thích khu vực và xu hướng thị trường.
* **Size (Kích Cỡ)**: Thông số kích cỡ (nếu có) của sản phẩm mua, liên quan đến quần áo, giày dép và một số hàng tiêu dùng nhất định.
* **Color (Màu Sắc)**: Biến thể màu sắc hoặc lựa chọn liên quan đến sản phẩm đã mua, ảnh hưởng đến sở thích của khách hàng và tình trạng hàng hóa.
* **Season (Mùa)**: Tính phù hợp theo mùa của sản phẩm đã mua (ví dụ: xuân, hè, thu, đông), ảnh hưởng đến việc quản lý tồn kho và các chiến lược tiếp thị.
* **Review Rating (Đánh Giá) (Đánh Giá)**: Đánh giá định lượng hoặc định tính được cung cấp bởi khách hàng về mức độ hài lòng với sản phẩm đã mua.
* **Subscription Status (Tình Trạng Đăng Ký)**: Cho biết liệu khách hàng đã đăng ký dịch vụ theo gói không, cung cấp thông tin về mức độ trung thành và khả năng tạo doanh thu định kỳ.
* **Shipping Type (Loại Giao Hàng)**: Chỉ định phương thức giao hàng cho sản phẩm đã mua (ví dụ: giao hàng tiêu chuẩn, giao hàng nhanh), ảnh hưởng đến thời gian và chi phí giao hàng.
* **Discount Applied (Giảm Giá Áp Dụng)**: Cho biết liệu có bất kỳ giảm giá khuyến mãi nào được áp dụng cho giao dịch hay không, làm sáng tỏ độ nhạy cảm với giá và hiệu quả của các chương trình khuyến mãi.
* **Promo Code Used (Mã Khuyến Mãi Được Sử Dụng)**: Ghi chú liệu có mã khuyến mãi hoặc phiếu giảm giá nào được sử dụng trong giao dịch, hỗ trợ đánh giá thành công của các chiến dịch tiếp thị.
* **Previous Purchases (Giao Dịch Trước Đây)**: Cung cấp thông tin về số lượng hoặc tần suất các giao dịch trước đây của khách hàng, góp phần vào phân khúc khách hàng và các chiến lược giữ chân khách hàng.
* **Payment Method (Phương Thức Thanh Toán)**: Chỉ định phương thức thanh toán mà khách hàng sử dụng (ví dụ: thẻ tín dụng, tiền mặt), cung cấp thông tin về các tùy chọn thanh toán ưu thích.
* **Frequency of Purchases (Tần Suất Mua Hàng)**: Cho biết mức độ thường xuyên khách hàng thực hiện các hoạt động mua sắm, là chỉ số quan trọng để đánh giá sự trung thành của khách hàng và giá trị vòng đời của khách hàng.

## Môi trường thực nghiệm

**Cấu hình máy**: CPU Intel Core i5 13600KF, RAM 32GB, Window 11 64-bit

**Công cụ hỗ trợ**: IDE Pycharm, Python 3.12, Jupyter Notebook

Tập dữ liệu sử dụng để thực nghiệm: …

## Quy trình thực nghiệm

Biểu đồ / hình ảnh mô tả quy trình

### Đặt mục tiêu

…

### Tiền xử lý dữ liệu

Mô tả …

A screenshot of a computer

Description automatically generated

Hình . Thông tin tóm lược dữ liệu của cột dữ liệu dạng số

- Làm sạch dữ liệu:

Mô tả …

Xóa một số dữ liệu trong dataset cho case dữ liệu khuyết

A screen shot of a computer

Description automatically generated

Hình . Kiểm tra dữ liệu bị khuyết

***-*** Chuyển đổi dữ liệu

Xây Dựng Mạng Lưới Mua Hàng

(100 first records)

A network of dots and lines

Description automatically generated

### Áp Dụng Giải Thuật Girvan-Newman

A screen shot of a computer

Description automatically generated

# CHƯƠNG 4: XÂY DỰNG SẢN PHẨM DEMO

# KẾT LUẬN: