

**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO MÔN HỌC**

**KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:**

**SỬ DỤNG PHƯƠNG PHÁP K NEAREST CLUSTERING**

**VÀ MÔ HÌNH PFM ĐỂ PHÂN TÍCH DỮ LIỆU CỦA THƯƠNG MẠI ĐIỆN TỬ BRAZIL**

|  |  |
| --- | --- |
| **Sinh viên thực hiện:** | **VŨ QUANG LONG**  **NGUYỄN ÁNH DƯƠNG**  **NGÔ MINH HIẾU** |
| **Giảng viên hướng dẫn** | **: VŨ VĂN ĐỊNH** |
| **Khoa** | **: CÔNG NGHỆ THÔNG TIN** |
| **Chuyên ngành** | **: HT THƯƠNG MẠI ĐIỆN TỬ** |
| **Lớp** | **: D13HTTMĐT1** |
| **Khóa** | **: 2018-2023** |

***Hà Nội, tháng 04 năm 2021***

**PHIẾU CHẤM ĐIỂM**

Sinh viên thực hiện :

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Điểm** |
| **Vũ Quang Long** |  |  |
| **Ngô Minh Hiếu** |  |  |
| **Nguyễn Ánh Dương** |  |  |

Giảng viên chấm :

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên 1: |  |  |
| Giảng viên 2: |  |  |

**MỤC LỤC**

[**MỤC LỤC** 3](#_Toc66304555)

[**LỜI CẢM ƠN** 1](#_Toc66304556)

[**TÓM TẮT** 2](#_Toc66304557)

[**CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI** 3](#_Toc66304558)

[1.1 Đặt vấn đề 3](#_Toc66304559)

[1.2 Cơ sở hình thành đề tài 3](#_Toc66304560)

[1.3 Mục tiêu đề tài 5](#_Toc66304561)

[1.4 Đối tượng và phương pháp nghiên cứu 5](#_Toc66304562)

[**CHƯƠNG 2: KHAI PHÁ DỮ LIỆU** 6](#_Toc66304563)

[2.1 Tổng quan về kỹ thuật Khai phá dữ liệu (Data mining) 6](#_Toc66304564)

[2.1.1 Khái niệm về khai phá dữ liệu 6](#_Toc66304565)

[2.1.2 Quy trình khai phá dữ liệu 7](#_Toc66304566)

[2.1.3 Ứng dụng của khai phá dữ liệu 10](#_Toc66304567)

[2.2 Tổng quan về hệ hỗ trợ ra quyết định 10](#_Toc66304568)

[2.3 Phân lớp và ứng dụng 11](#_Toc66304569)

[2.3.1 Mục đích của phân lớp 11](#_Toc66304570)

[2.3.2 Các bước cơ bản để phân lớp 12](#_Toc66304571)

[2.3.4 Các ứng dụng của phân lớp 13](#_Toc66304572)

[2.4 Hồi quy và ứng dụng 15](#_Toc66304573)

[2.4.1 Mục đích của hồi quy 15](#_Toc66304574)

[2.4.2 Các bước cơ bản để lựa chọn mô hình hồi quy 15](#_Toc66304575)

[2.4.3 Các ứng dụng của hồi quy 16](#_Toc66304576)

[**CHƯƠNG 3: KỸ THUẬT PHÂN TÍCH VÀ THUẬT TOÁN** 18](#_Toc66304577)

[3.1 Giới thiệu về kỹ thuật phân tích trong khai phá dữ liệu 18](#_Toc66304578)

[3.1.1 Phân lớp 18](#_Toc66304579)

[3.1.2 Hồi quy 18](#_Toc66304580)

[3.1.3 Phân tích hành vi khách hàng dựa trên RFM 20](#_Toc66304581)

[3.2 Thuật Toán trong phân tích dự liệu 21](#_Toc66304582)

[3.2.1 Thuật toán classifier 21](#_Toc66304583)

[3.2.2 Thuật toán linear regression 22](#_Toc66304584)

[3.2.3 Thuật toán RandomForestClassifier 23](#_Toc66304585)

[3.2.4 Thuật toán ExtraTrees Classifier 24](#_Toc66304586)

# **LỜI CẢM ƠN**

Qua bài tập lớn này, chúng em xin gửi lời cảm ơn tới thầy cô khoa công nghệ thông tin, đặc biệt là thầy Vũ Văn Định đã cho chúng em có cơ hội được tìm hiểu một góc kiến thức mới, hay và bổ ích cùng với đó là sự tận tâm dạy dỗ chúng em, giúp chúng em có thể hoàn thiện đề tài này. Trong quá trình tìm hiểu và hoàn thiện, đề tài sẽ không thể tránh khỏi những sai sót, khuyết điểm. Vì vậy, nhóm thực hiện chúng em hy vọng nhận được sự đánh giá và đóng góp nhiệt tình từ phía thầy và các bạn để bài của nhóm chúng em được hoàn thiện hơn.

Qua bài tập lớn này, chúng em xin cảm ơn các bạn bè lớp D13HTTMDT1 đã giúp đỡ chúng em trong quá trình học tập và làm bài tập lớn, đã chia sẻ kinh nghiệm kiến thức của các bạn đã tạo nên nền tảng kiến thức cho chúng em.

Cuối cùng, chúng em xin gửi lời cảm ơn gia đình đặc biệt là cha mẹ đã tạo điều kiện tốt nhất cho con có đủ khả năng thực hiện bài tập lớn này, trang trải học phí, đông viên tinh thần cho em để học tập trong môi trường đại học tuyệt vời này.

Chúng em xin chân thành cảm ơn!

|  |
| --- |
| Nhóm sinh viên thực hiện |
| Vũ Quang Long |
| Nguyễn Ánh Dương |
| Ngô Minh Hiếu |

# **TÓM TẮT**

Khi nói về khái niệm thương mại điện tử (E-Commerce), nhiều người nhầm lẫn với khái niệm của [Kinh doanh điện tử](https://vi.wikipedia.org/wiki/Kinh_doanh_%C4%91i%E1%BB%87n_t%E1%BB%AD) (E-Business). Tuy nhiên, thương mại điện tử đôi khi được xem là [tập con](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_h%E1%BB%A3p_con) của [kinh doanh điện tử](https://vi.wikipedia.org/wiki/Kinh_doanh_%C4%91i%E1%BB%87n_t%E1%BB%AD). Thương mại điện tử chú trọng đến việc mua bán trực tuyến (tập trung bên ngoài), trong khi đó [kinh doanh điện tử](https://vi.wikipedia.org/wiki/Kinh_doanh_%C4%91i%E1%BB%87n_t%E1%BB%AD) là việc sử dụng [Internet](https://vi.wikipedia.org/wiki/Internet) và các công nghệ trực tuyến tạo ra quá trình hoạt động kinh doanh hiệu quả dù có hay không có lợi nhuận, vì vậy tăng lợi ích với [khách hàng](https://vi.wikipedia.org/wiki/Kh%C3%A1ch_h%C3%A0ng) (tập trung bên trong). Hiện này rất nhiều những sàn thương mại điện tử lớn được tạo ra để phục vụ cho việc mua sắm của con người như: Amazon, taobao, shoppe… Nhưng việc hiểu tâm lý khách hàng và cần tìm ra những khách hàng tiềm năng để chăm sóc đang còn chưa phổ biến và làm sao dựa vào công nghệ để ta có thể nắm bắt được nhu cầu, tâm lý của khác hàng thì phải tìm hiểu và phân tích ngay thương mại điện tử nước Bzail.

# **CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI**

## 1.1 Đặt vấn đề

Ứng dụng công nghệ thông tin vào việc lưu trữ và xử lý thông tin ngày nay được áp dụng hầu hết trong lĩnh vực, điều này đã tạo ra một lượng lớn dữ liệu được lưu trữ với kích thước tăng lên không ngừng. Đây chính là điều kiện tốt cho việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập bẳng biểu và khai phá dữ liệu.

Khai phá dữ liệu là một kỹ thuật dựa trên nền tảng của nhiều lý thuyết như xác xuất, thống kê, máy học nhằm tìm kiếm các tri thức tiềm ẩn trong các kho dữ liệu có kích thước lớn mà người dùng khó có thể nhận biết bằng những kỹ thuật thông thường. Nguồn dữ liệu Thương mại điện tử rất lớn, nếu áp dụng khai phá dữ liệu trong lĩnh vực này sẽ mang lại nhiều ý nghĩa. Nó sẽ cung cấp nững thông tin quý giá nhằm hỗ trợ trong việc có thể nắm bắt được nhu cầu, tâm lý của khác hàng.

**Tất nhiên**, mục tiêu cuối cùng của việc hiểu rõ tâm lý khách hàng mua online chính là để bán hàng, tăng doanh thu cho doanh nghiệp. Hơn thế, việc hiểu rõ nhu cầu của người mua còn giúp bạn định hướng, tiếp thị đúng sản phẩm mà người tiêu dùng cần. Và về lâu dài, hiểu rõ khách hàng còn giúp bạn xây dựng lòng tin nơi người mua, xây dựng sự chuyên nghiệp cho chính thương hiệu của mình.

Thậm chí, khi bạn càng hiểu tâm lý khách hàng, bạn càng thuận tiện trong việc phát triển sản phẩm mới đáp ứng đúng nhu cầu của thị trường thực tế và tiếp cận đến đúng phân khúc khách hàng tiềm năng dễ dàng, hiệu quả hơn.

## 1.2 Cơ sở hình thành đề tài

Thời đại công nghệ số 4.0 cùng với sự phát triển như vũ bão của Internet, xu hướng kinh doanh trực tuyến, bán hàng online đã đem lại hiệu quả kinh tế cho không ít lĩnh vực kinh doanh tại nước ta. Chính vì vậy, cụm từ thương mại điện tử hiện nay đã không còn quá xa lạ và lĩnh vực này còn được đánh giá là một lĩnh vực có tiềm năng phát triển mạnh mẽ song hành cùng với sự phát triển của Internet và nền tảng công nghệ thông tin ngày càng được hiện đại hoá.

Theo một kết quả thống kê cho thấy, có khoảng 64 triệu người dân Việt Nam đã sử dụng Internet và trung bình mỗi người sẽ truy cập Internet 2 giờ/ngày; số người sử dụng smart phone hiện nay ở Việt Nam chiếm khoảng 82% dân số và nhiều người còn sử dụng hơn 02 chiếc điện thoại. Điều này cho thấy khả năng truy cập vào các website thương mại điện tử lớn hơn bao giờ hết. Đây chính là nền tảng để lĩnh vực [**Ecommerce**](https://vieclamvui.com/viec-lam-internet-online-media/nganh-thuong-mai-dien-tu-338.html) phát triển mạnh ở nước ta.

Bên cạnh đó, người tiêu dùng Việt Nam cũng đang quen dần với lĩnh vực mua sắm trực tuyến khi các thiết bị điện tử đã không còn quá xa lạ. Ngoài ra, với sự thay đổi của xã hội và nhịp sống hiện đại đầy bận rộn, thời gian dành cho việc ra ngoài mua sắm sẽ hạn chế rất nhiều và thương mại điện tử sẽ trở thành lựa chọn phù hợp.

Doanh số của thương mại điện tử hiện nay chỉ chiếm khoảng 3% trên tổng quy mô của thị trường bán lẻ Việt Nam, vẫn còn khá thấp so với các nước trong khu vực, nhưng với sự cải thiện của các dịch vụ vận chuyển, uy tín người bán và nhận biết của người dân về TMĐT ngày càng được nâng cao, tiềm năng phát triển của thị trường TMĐT Việt Nam sẽ có những bước tăng đột phá trong thời gian tới. Hơn nữa, các doanh nghiệp Việt Nam cũng có thể bồi dưỡng nâng cao kiến thức về thương mại điện tử cho nhân lực công ty qua các khoá học trực tuyến để nắm bắt thị trường và phát triển tốt hơn.

Thêm vào đó, sự gia nhập sân chơi của những những thương hiệu TMĐT lớn từ nước ngoài chắc chắn sẽ góp phần tạo thêm động lực cho thị trường cũng như thúc đẩy các doanh nghiệp thương mại điện tử trong nước quan tâm và có sự đầu tư đúng mức để có thể giành được lợi thế trong cuộc đua giành thị phần. Và lúc này, người hưởng lợi nhất trong thương mại điện tử chính là khách hàng vì có thể mua được sản phẩm rẻ hơn, nhanh hơn, hiệu quả hơn và thuận tiện hơn.

## 1.3 Mục tiêu đề tài

Đề tài tập chung vào nghiên cứu kỹ thuật phân cụm trong khai phá dữu liệu, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và xây dựng ứng dụng cụ thể. Sau khi phân tích đặc điểm của dữu liệu thu nhập được và lựa chọn giải thuật phù hợp với dữ liệu, việc xây dựng và đánh giá chất lượng, độ hiệu quả của hệ thống cũng là mục tiêu chính của đề tài.

## 1.4 Đối tượng và phương pháp nghiên cứu

Đề tài tập chung vào nghiên cứu kỹ thuật phân cụm trong khai phá dữ liệu cụ thể là nghiên cứu thuật toán k-means và mô hình PFM để áp dụng vào việc phân tích cơ sở dữ liệu của tình hình mua hàng hóa. Mục đích đặc biệt là ứng dụng phân khúc khách hàng hỗ trợ các hoạt động marketing, sales hiệu quả. Đặc biệt mô hình PFM là công cụ hỗ trợ phân tích các chiến lược marketing để tìm ra khách hàng giá trị nhất với mỗi công ty.

# **CHƯƠNG 2: KHAI PHÁ DỮ LIỆU**

## 2.1 Tổng quan về kỹ thuật Khai phá dữ liệu (Data mining)

### 2.1.1 Khái niệm về khai phá dữ liệu

Khai phá dữ liệu (data mining) là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp.Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.

Khai phá dữ liệu là một bước của quá trình khai thác tri thức (Knowledge Discovery Process), bao gồm:

• Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (Problem understanding and data understanding).

• Chuẩn bị dữ liệu (Data preparation), bao gồm các quá trình làm sạch dữ liệu (data cleaning), tích hợp dữ liệu (data integration), chọn dữ liệu (data selection), biến đổi dữ liệu (data transformation).

• Khai thác dữ liệu (Data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữ liệu. Kết quả cho ta một nguồn tri thức thô.

• Đánh giá (Evaluation): dựa trên một số tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.

• Triển khai (Deployment).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đến bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

### 2.1.2 Quy trình khai phá dữ liệu

**2.1.2.1 Nghiên cứu lĩnh vực**

Ta cần nghiên cứu lĩnh vực cần sử dụng Data mining để xác định được những tri thức ta cần chất lọc, từ đó định hướng để tránh tốn thời gian cho những tri thức không cần thiết.

**2.1.2.2 Tạo tập tin dữ liễu đầu vào**

Ta xây dựng tập tin để lưu trữ các dữ liệu đầu vào để máy tính có thể lưu trữ và xử lý.

**2.1.2.3 Tiền xử lý, làm sạch, mã hóa**

Ở bước này ta tiến hành bỏ bớt những dữ liệu rườm rà, không cần thiết, tinh chỉnh lại cấu trúc của dữ liệu và mã hóa chúng để tiện cho quá trình xử lý .

**2.1.2.4 Rút gọn chiều**

Thông thường một tập dữ liệu có chiều khá lớn sẽ sinh ra một lượng dự liệu khổng lồ,ví dụ với n chiều ta sẽ có 2^n nguyên tổ hợp .Do đó , đây là một bước quan trọng giúp giảm đáng kể hao tổn hề tài nguyên trong quá trình xử lý tri thức.Thông thường ta sẽ dùng Rough set (http://en.wikipedia.org/wiki/Rough\_set) để giảm số chiều.

**2.1.2.5 Chọn tác vụ khai thác dữ liệu**

Để đạt được mục đích ta cần, ta chọn được tác vụ khai thác dữ liệu sao cho phù hợp.Thông thường có các tác vụ sau:

• Đặc trưng(feature)

• Phân biệt(discrimination)

• Kết hợp(association)

• Phân lớp(classification)

• Gom cụm(clusterity)

• Xu thế(trend analysis)

• Phân tích độ lệch

• Phân tích độ hiếm

**2.1.2.6 Khai thác dữ liệu: Tìm kiếm tri thức**

Sau khi tiến hành các bước trên thì đây là bước chính của cả quá trình , ta sẽ tiến hành khai thác và tìm kiếm tri thức.

**2.1.2.7 Đánh giá mẫu tìm được**

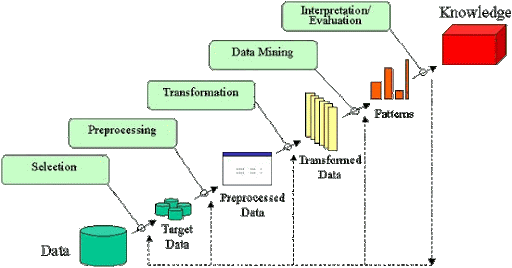
Ta cần đánh giá lại trong các tri thức tìm được , ta sẽ sử dụng được những tri thức nào , những tri thức nào dư thừa,không cần biết.

**2.1.2.8 Biểu diễn tri thức**

Ta biểu diễn tri thức vừa thu nhập được dưới dạng ngôn ngữ tự nhiên và hình thức sao cho người dùng có thể hiểu được những tri thức đó.

**2.1.2.9 Sử dụng các tri thức vừa khám phá**

Ta có thể tham khảo tiến trình KDD( Knowledge Discovery in Databases) để hiểu rõ hơn về khai phá dữ liệu:



Hình 2. 1 Knowledge Discovery in Databases

Chuẩn bị dữ liệu (data preparation), bao gồm các quá trình làm sạch dữ liệu (data cleaning), tích hợp dữu liệu ( data integration), chọn dữ liệu (data selection), biến đổi dữ liệu (data transformation).

Khai thác dữu liệu (data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữu liệu. Kết quả cho ta một nguồn tri thức thô.

Đánh giá (evaluation): dựa trên một tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.

Triển khai (deployment).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đên bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

### 2.1.3 Ứng dụng của khai phá dữ liệu

Kinh tế - ứng dụng trong kinh doanh, tài chính, tiếp thị bán hàng, bảo hiểm, thương mại, ngân hàng,.. Đưa ra các bản báo cáo giàu thông tin, phân tích rửi ro trước khi đưa ra các chiến lược kinh doanh, sản xuất, phân loại khách hàng từ đó phân định ra thị trường, thị phần:…

Khoa học: Thiên văn học - dự đoán đường đi các thiên thể, hành tinh,...; Công nghệ sinh học – tìm ra các gen mới, cây con giống mới,…

Web: các công cụ tìm kiếm.

## 2.2 Tổng quan về hệ hỗ trợ ra quyết định

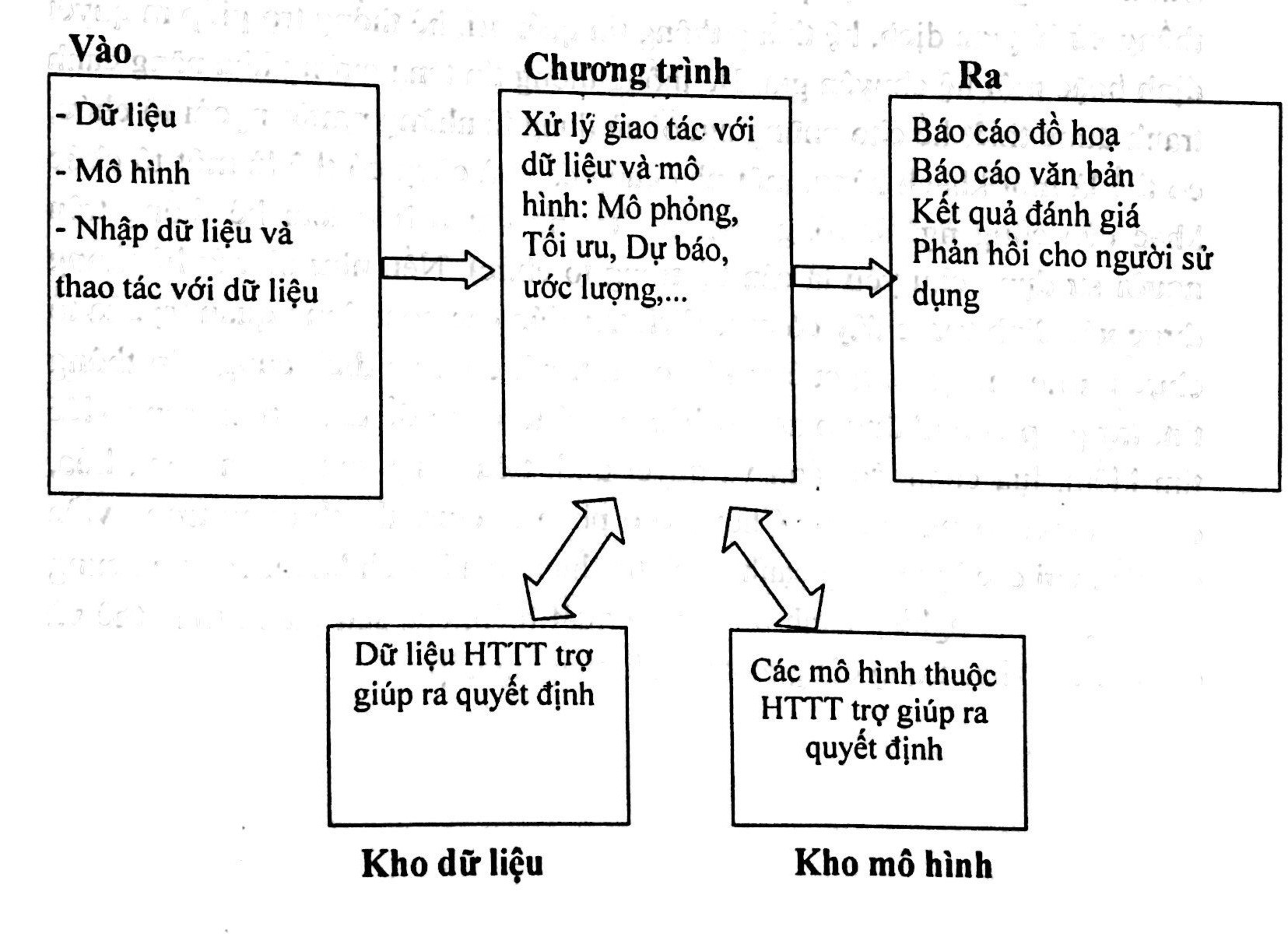
Hệ hỗ trợ ra quyết định là một hệ thống thuộc hệ thống thông tin, có nhiệm vụ cung cấp các thông tin hỗ trợ cho việc ra quyết định để tham khảo và giải quyết vấn đề. Hệ hỗ trợ ra quyết định có thể dùng cho cá nhân hay tổ chức và có thể hỗ trợ gián tiếp hoặc trực tiếp.

Trong lĩnh vực cộng nghệ, hệ hỗ trợ ra quyết định dựa vào tri thức đã học sẽ cung cấp thông tin đơn hàng, tần suất mua, khách hàng uy tín…Thông tin này được trích lọc để cung cấp một cách thông minh có giá trị cho quá trình chuẩn đoán, theo dõi hiệu quả hơn, từ đó ta thấy một số lợi ích của hệ hỗ trợ ra quyết định trong phân tích khách hàng trong TMĐT như sau:

• Tăng cường chất lượng chuẩn đoán, chăm sóc khách hàng.

• Tăng khả năng chốt đơn thành công, khách hành tiềm năng.

• Tăng cường hiệu quả ứng dụng công nghệ thông tin trong việc kinh doanh.



Hình 2. 2 Sơ đồ hệ hỗ trợ quyết định

## 2.3 Phân lớp và ứng dụng

### 2.3.1 Mục đích của phân lớp

Phân loại là một trong những vấn đề nguyên cứu mở rộng hiện nay ; tập trung chủ yếu vào thống kê, máy học và mạng nơtrôn . Kĩ thuật phân lớp được đánh giá là một kĩ thuật khai phá dữ liệu được sử dụng rộng rãi nhất với nhiều mở rộng . Sự kết hợp của kỹ thuật phân lớp và cơ sở dữ liệu là một lĩnh vực hứa hẹn bởi vì đáp ứng được một vấn đề hết sức quan trọng của ứng dụng cơ sở dữ liệu đó là tính uyển chuyển cao .. Phân cụm dữ liệu nhằm mục đích chính là khai phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó, cho phép người ta đi sâu vào phân tích và nghiên cứu cho từng cụm dữ liệu này nhằm khai phá và tìm kiếm các thông tin tiềm ẩn, hữu ích phục vụ cho ra quyết định.

Một vài ví dụ về ý nghĩa thực tiễn của phân cụm dữ liệu như sau :

- Khám phá ra các vị trí địa lý thuận lợi cho việc xây dựng các kho hàng

phục vụ mua bàn hàng của một công ty thương mại

- Xác định các cụm ảnh như ảnh của các loài động vật như loài thú, chim,...

trong tập CSDL ảnh về động vật nhằm phục vụ cho việc tìm kiếm ảnh

- Xác định các nhóm người bệnh nhằm cung cấp thông tin cho việc phân phối

các thuốc điều trị trong y tế

- Xác định nhóm các khách hàng trong CSDL ngân hàng có vốn các đầu tư

vào bất động sản cao...

Như vậy, phân cụm dữ liệu là một phương pháp xử lý thông tin quan trọng và phổ biến, nó nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm tương tự.

Tiếp theo, giả sử rằng tất cả các dạng dữ liệu được biểu diễn bởi khái niệm đặc trưng, các đặc trưng hình thành nên vector đặc trưng l- chiều. Thuật ngữ phân cụm được hiểu là phân cụm dữ liệu.

### 2.3.2 Các bước cơ bản để phân lớp

Để xây dựng được mô hình phân lớp và  đánh giá được mô hình chúng ta phải trải quá các quá trình như dưới đây:

Bước 1: Chuẩn bị tập dữ liệu huấn luyện (dataset) và rút trích đặc trưng (feature extraction)

Công đoạn này được xem là công đoạn quan trọng trong các bài toán về Machine Learning. Vì đây là input cho việc học dể tìm ra mô hình của bài toán. Chúng ta phải biết cần chọn ra những đặt trưng tốt (good feature) của dữ liệu, lược bỏ những đặc trưng không tốt của dữ liệu, gây nhiễu (noise). Uớc lượng số chiều của dữ liệu bao nhiêu là tốt hay nói cách khác là chọn bao nhiêu feature. Nếu số chiều quá lớn gây khó khăn cho việc tính toán thì phải giảm số chiều của dữ liệu nhưng vẫn giữ được độ chính xác của dữ liệu (reduce demension).

Ở bước này chúng ta cũng chuẩn bị bộ dữ liệu để test trên mô hình. Thông thường sẽ sử dụng cross-validation (kiểm tra chéo) để chia tập datasets thành hai phần, một phần phục vụ cho training (training datasets) và phần còn lại phục vụ cho mục đích testing trên mô hình (testing dataset). Có hai cách thường sử dụng trong cross-validation là splitting và k-fold.

Bước 2: Xây dựng mô hình phân lớp (classifier model)

Mục đích của mô hình huấn luyện là tìm ra hàm f(x) và thông qua hàm f tìm được để chúng ta gán nhãn cho dữ liệu. Bước này thường được gọi là học hay training.

f(x) = y

                        Trong đó: x là các feature hay input đầu vào của dữ liệu

y là nhãn lớp hay output đầu ra.

Thông thường để xây dựng mô hình phân lớp cho bài toán này chúng ta sử dụng các thuật toán học giám sát (supervised learning) như KNN, Neural Network, SVM, Decision Tree, Navie Bayers.

Bước 3: Kiểm tra dữ liệu với mô hình (make prediction)

Sau khi đã tìm được mô hình phân lớp ở bước 2, thì ở bước này chúng ta sẽ đưa vào các dữ liệu mới để kiểm tra trên mô hình phân lớp.

Bước 4: Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất

Bước cuối cùng chúng ta sẽ đánh giá mô hình bằng cách đánh giá mức độ lỗi của dữ liệu testing và dữ liệu traning thông qua mô hình tìm được. Nếu không đạt được kết quả mong muốn của chúng ta thì phải thay đổi các tham số (turning parameter) của các thuật toán học để tìm ra các mô hình tốt hơn và kiểm tra, đánh giá lại mô hình phân lớp. Và cuối cùng chọn ra mô hình phân lớp tốt nhất cho bài toán của chúng ta.

### 2.3.4 Các ứng dụng của phân lớp

Phân lớp là một công cụ quan trọng trong một số ứng dụng. Sau đây là một số ứng dụng của nó:

Giảm dữ liệu: Giả sử ta có một lượng lớn dữ liệu (N). Phân lớp sẽ nhóm các dữ liệu này thành m phân lớp dữ liệu dễ nhận thấy và m << N. Sau đó xử lý mỗi

Lớp như một đối tượng đơn.

Rút ra các giả thuyết : Các giả thuyết này có liên quan đến tính tự nhiên của dữ liệu và phải được kiểm tra bởi việc dùng một số tập dữ liệu khác.

Kiểm định giả thuyết : Ta sẽ phân lớp để xét xem có tồn tại một tập dữ liệu nào đó trong tập dữ liệu thoả mãn các giả thuyết đã cho hay không. Chẳng

hạn xem xét giả thuyết sau đây: “Các công ty lớn đầu tư ra nước ngoài “. Để

kiểm tra, ta áp dụng kỹ thuật phân lớp với một tập đại diện lớn các công ty.

Giả sử rằng mỗi công ty được đặc trưng bởi tầm vóc, các hoạt động ở nước

ngoài và khả năng hoàn thành các dự án. Nếu sau khi phân lớp, một lớp các

công ty được hình thành gồm các công ty lớn và có vốn đầu tư ra nước ngoài

(không quan tâm đến khả năng hoàn thành các dự án) thì giả thuyết đó được

củng cố bởi kỹ thuật phân lớp đã thực hiện.

Dự đoán dựa trên các lớp : Đầu tiên ta sẽ phân lớp một tập dữ liệu thành các lớp mang đặc điểm của các dạng mà nó chứa. Sau đó, khi có một dạng mới

chưa biết ta sẽ xác định xem nó sẽ có khả năng thuộc về lớp nào nhất và dự

đoán được một số đặc điểm của dạng này nhờ các đặc trưng chung của cả

lớp.

Cụ thể hơn, phân lớp dữ liệu đã được áp dụng cho một số ứng dụng điển hình trong các lĩnh vực sau [13] :

Thương mại : Trong thương mại, phân lớp có thể giúp các thương nhân

khám phá ra các nhóm khách hàng quan trọng có các đặc trưng tương đồng

nhau và đặc tả họ từ các mẫu mua bán trong cơ sở dữ liệu khách hàng.

Sinh học : Trong sinh học, phân lớp được sử dụng để xác định các loại sinh

vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc

trong các mẫu.

Phân tích dữ liệu không gian : Do sự đồ sộ của dữ liệu không gian như dữ

liệu thu được từ các hình ảnh chụp từ vệ tinh các thiết bị y học hoặc hệ thống

thông tin địa lý (GIS), ...làm cho người dùng rất khó để kiểm tra các dữ liệu

không gian một cách chi tiết. Phân lớp có thể trợ giúp người dùng tự động

phân tích và xử lý các dữ liệu không gian như nhận dạng và chiết xuất các

đặc tính hoặc các mẫu dữ liệu quan tâm có thể tồn tại trong cơ sở dữ liệu

không gian.

Lập quy hoạch đô thị : Nhận dạng các nhóm nhà theo kiểu và vị trí địa

lý,...nhằm cung cấp thông tin cho quy hoạch đô thị.

Nghiên cứu trái đất : Phân lớp để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm.

Địa lý : Phân lớp các động vật và thực vật và đưa ra đặc trưng của chúng.

Web Mining : Phân lớp có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc khám phá tri thức từ dữ liệu,...

## 2.4 Hồi quy và ứng dụng

### 2.4.1 Mục đích của hồi quy

**Phân tích hồi quy (regression analysis)** là kỹ thuật thống kê dùng để ước lượng phương trình phù hợp nhất với các tập hợp kết quả quan sát của biến phụ thuộc và biến độc lập. Nó cho phép đạt được kết quả ước lượng tốt nhất về mối quan hệ chân thực giữa các biến số. Từ phương trình ước lượng được này, người ta có thể dự báo về biến phụ thuộc (chưa biết) dựa vào giá trị cho trước của biến độc lập (đã biết).

### 2.4.2 Các bước cơ bản để lựa chọn mô hình hồi quy

Bước 1: Xác định danh sách các biến độc lập có thể có trong mô hình. Dựa vào ý nghĩa thực tế của bài toán đã được đặt ta, ta cần liệt kê tất cả những biến độc lập có thể có của mô hình, là những biến có khả năng ảnh hưởng đến giá trị của biến phụ thuộc. Những biến này có thể đã có sẵn trong danh sách các biến của bộ số liệu, song cũng có thể là 10 biến được tạo ra từ các biến trong danh sách đó thông qua các phép biến đổi.

Bước 2: Kiếm tra sự vi phạm các giả thiết của mô hình hồi quy. Bước này bao gồm việc tiến hành kiểm định các vấn đề như đa cộng tuyến, phương sai thay đổi, tự tương quan, phân không chuẩn của sai số, và khắc phục các vi phạm phát hiện được.

Bước 3: Chọn dạng hàm hồi quy. Dạng hàm hồi quy có thể được xác định dựa trên kiến thức chuyên ngành liên quan đến số liệu hoặc dựa vào dạng hàm đã được sử dụng trong các nghiên cứu trước đó. Bên cạnh đó, có thể xác định dạng hàm hồi quy thông qua việc khảo sát các đồ thị biểu diễn sơ bộ mối hệ giữa biến độc quan lập và biến phụ thuộc. Chẳng hạn nếu trên đồ thị, các chấm tương ứng với các quan sát của tập số liệu nằm tập trung hai bên một đường thăng nào đó, thi có thể chọn dang hàm hồi quy tuyến tính. Còn nếu các chấm đó lại nằm hai bên một đường cong thì có thể dựa vào dạng của đường cong đó mà đưa ra dạng hàm hồi quy phi tuyến thích hợp.

Bước 4: Áp dụng các tiêu chuẩn để đánh giá và lựa chọn mô hình. Hệ số xác định là một thước đo thường được dùng đầu tiên để đánh giá chất lượng của mô hình hồi quy. Nếu hệ số xác định có giá trị lớn hơn 50% thì có thể coi mô hình khá phù hợp với tập số liệu. Còn nếu hệ số xác định nhỏ hơn 30% thì có thể khẳng định mô hình không phù hợp và nên tìm mô hình khác.

### 2.4.3 Các ứng dụng của hồi quy

Một số ví dụ về hồi quy có thể là:

* Dự báo mưa bằng cách sử dụng nhiệt độ và các yếu tố khác
* Xác định xu hướng thị trường
* Dự đoán tai nạn đường bộ do lái xe hấp tấp.

Tại sao chúng ta sử dụng phân tích hồi quy?

Như đã đề cập ở trên, phân tích hồi quy giúp dự đoán một biến liên tục. Có nhiều tình huống khác nhau trong thế giới thực mà chúng ta cần một số dự đoán trong tương lai như điều kiện thời tiết, dự đoán bán hàng, xu hướng tiếp thị, v.v., trong trường hợp đó, chúng tôi cần một số công nghệ có thể đưa ra dự đoán chính xác hơn. Vì vậy, đối với trường hợp như vậy, chúng ta cần phân tích hồi quy, một phương pháp thống kê và được sử dụng trong học máy và khoa học dữ liệu. Dưới đây là một số lý do khác để sử dụng phân tích hồi quy:

* Hồi quy ước tính mối quan hệ giữa mục tiêu và biến độc lập.
* Nó được sử dụng để tìm các xu hướng trong dữ liệu.
* Nó giúp dự đoán các giá trị thực / liên tục.
* Bằng cách thực hiện hồi quy, chúng ta có thể tự tin xác định **yếu tố quan trọng nhất, yếu tố ít quan trọng nhất và mỗi yếu tố ảnh hưởng đến các yếu tố khác như thế nào.**

# **CHƯƠNG 3: KỸ THUẬT PHÂN TÍCH VÀ THUẬT TOÁN**

## 3.1 Giới thiệu về kỹ thuật phân tích trong khai phá dữ liệu

### 3.1.1 Phân lớp

Phân loại là một nhiệm vụ yêu cầu sử dụng các thuật toán học máy học cách gán nhãn lớp cho các ví dụ từ miền vấn đề. Một ví dụ dễ hiểu là phân loại email là “ spam ” hoặc “ không phải spam ”.

Có nhiều loại nhiệm vụ phân loại khác nhau mà bạn có thể gặp phải trong học máy và các phương pháp tiếp cận chuyên biệt để lập mô hình có thể được sử dụng cho từng loại.

Trong hướng dẫn này, bạn sẽ khám phá các loại mô hình dự đoán phân loại khác nhau trong học máy.

Sau khi hoàn thành hướng dẫn này, bạn sẽ biết:

* Mô hình dự đoán phân loại liên quan đến việc gán nhãn lớp cho các ví dụ đầu vào.
* Phân loại nhị phân đề cập đến việc dự đoán một trong hai lớp và phân loại nhiều lớp liên quan đến việc dự đoán một trong nhiều hơn hai lớp.
* Phân loại nhiều nhãn liên quan đến việc dự đoán một hoặc nhiều lớp cho mỗi ví dụ và phân loại không cân bằng đề cập đến các nhiệm vụ phân loại trong đó sự phân bố của các ví dụ trên các lớp không bằng nhau.

Các loại phân lớp:

* Mô hình dự đoán phân loại
* Phân loại nhị phân
* Phân loại nhiều lớp
* Phân loại nhiều nhãn
* Phân loại không cân bằng

### 3.1.2 Hồi quy

Hồi quy là một kỹ thuật học có giám sát giúp tìm ra mối tương quan giữa các biến và cho phép chúng tôi dự đoán biến đầu ra liên tục dựa trên một hoặc nhiều biến dự báo. Nó chủ yếu được sử dụng để **dự đoán, dự báo, mô hình chuỗi thời gian và xác định mối quan hệ nhân quả giữa các biến**.

Trong Hồi quy, chúng tôi vẽ biểu đồ giữa các biến phù hợp nhất với điểm dữ liệu đã cho, bằng cách sử dụng biểu đồ này, mô hình học máy có thể đưa ra dự đoán về dữ liệu. Nói cách đơn giản, "Hồi quy hiển thị một đường hoặc đường cong đi qua tất cả các điểm dữ liệu trên biểu đồ dự báo mục tiêu theo cách sao cho khoảng cách dọc giữa điểm dữ liệu và đường hồi quy là nhỏ nhất.**"**Khoảng cách giữa các điểm dữ liệu và dòng cho biết liệu một mô hình có nắm bắt được mối quan hệ chặt chẽ hay không.

Các thuật ngữ liên quan đến phân tích hồi quy:

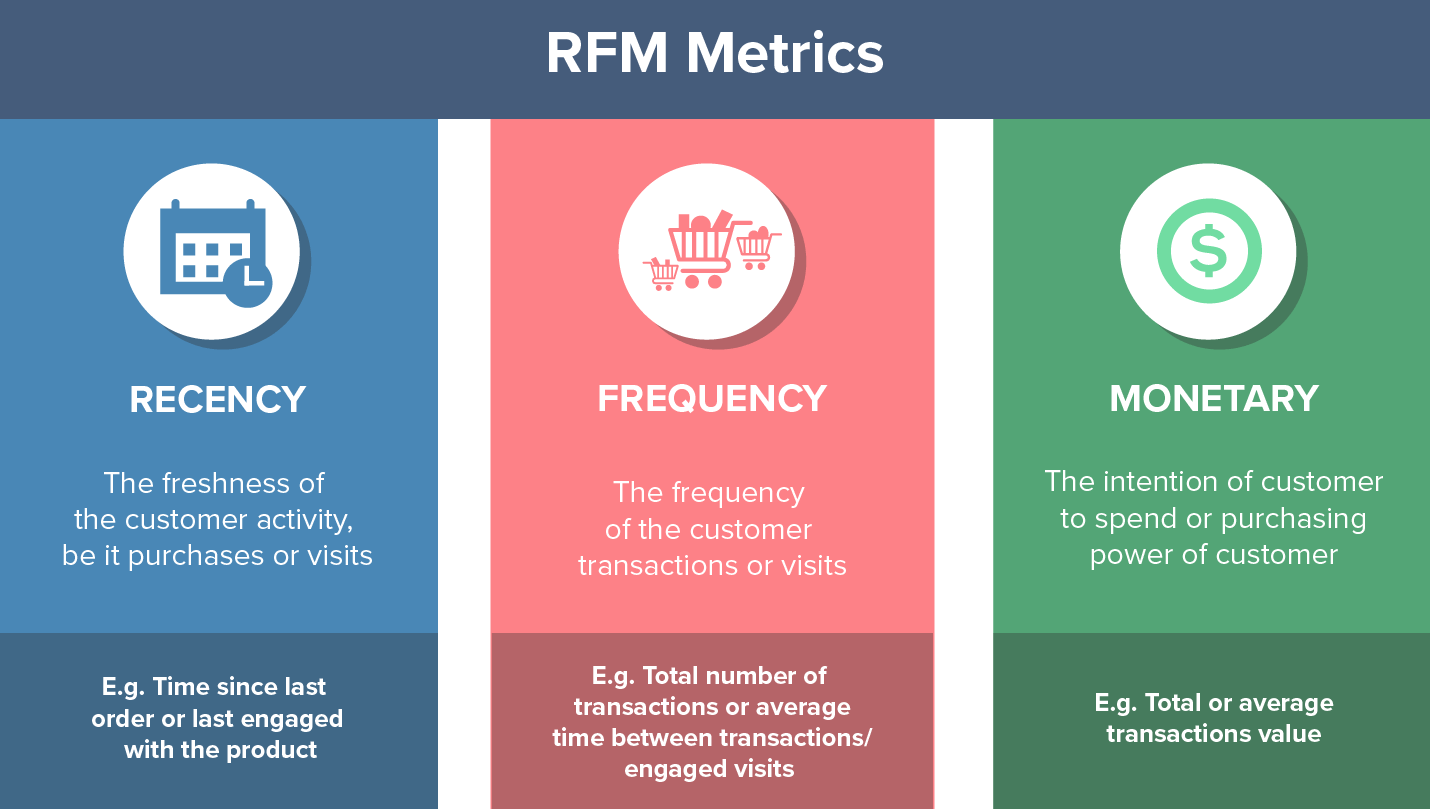
* **Biến phụ thuộc:** Yếu tố chính trong phân tích hồi quy mà chúng ta muốn dự đoán hoặc hiểu được gọi là biến phụ thuộc. Nó còn được gọi là**biến mục tiêu.**
* **Biến độc lập:** Các yếu tố ảnh hưởng đến các biến phụ thuộc hoặc được sử dụng để dự đoán giá trị của các biến phụ thuộc được gọi là biến độc lập, còn được gọi là **yếu tố dự báo.**
* **Giá trị ngoại lai:** Giá trị **ngoại lệ**là quan sát chứa giá trị rất thấp hoặc giá trị rất cao so với các giá trị quan sát khác. Một ngoại lệ có thể cản trở kết quả, vì vậy cần tránh.
* **Đa cộng tuyến:** Nếu các biến độc lập có tương quan cao với nhau hơn các biến khác, thì điều kiện như vậy được gọi là Đa cộng tuyến. Nó không nên xuất hiện trong tập dữ liệu, vì nó tạo ra vấn đề trong khi xếp hạng biến ảnh hưởng nhiều nhất.
* **Underfitting và Overfitting:** Nếu thuật toán của chúng tôi hoạt động tốt với tập dữ liệu đào tạo nhưng không tốt với tập dữ liệu thử nghiệm, thì vấn đề như vậy được gọi là **Overfitting.** Và nếu thuật toán của chúng tôi không hoạt động tốt ngay cả với tập dữ liệu đào tạo, thì vấn đề như vậy được gọi là **thiếu trang bị.**

Các loại hồi quy: Có nhiều loại hồi quy khác nhau được sử dụng trong khoa học dữ liệu và học máy. Mỗi loại có tầm quan trọng riêng đối với các kịch bản khác nhau, nhưng cốt lõi, tất cả các phương pháp hồi quy đều phân tích ảnh hưởng của biến độc lập lên các biến phụ thuộc. Ở đây chúng tôi đang thảo luận về một số loại hồi quy quan trọng được đưa ra dưới đây:

* **Hồi quy tuyến tính**
* **Hồi quy logistic**
* **Hồi quy đa thức**
* **Hỗ trợ hồi quy vectơ**
* **Hồi quy cây quyết định**
* **Hồi quy rừng ngẫu nhiên**
* **Hồi quy Ridge**
* **Hồi quy Lasso:**

### 3.1.3 Phân tích hành vi khách hàng dựa trên RFM

RFM là viết tắt của giá trị Lần truy cập gần đây, Tần suất và Tiền tệ, mỗi giá trị tương ứng với một số đặc điểm khách hàng chính. Các chỉ số RFM này là các chỉ số quan trọng về hành vi của khách hàng vì tần suất và giá trị tiền tệ ảnh hưởng đến giá trị lâu dài của khách hàng và lần truy cập gần đây ảnh hưởng đến tỷ lệ giữ chân, một thước đo mức độ tương tác.



Các doanh nghiệp thiếu khía cạnh tiền tệ, như lượng người xem, lượng người đọc hoặc các sản phẩm hướng đến lướt web, có thể sử dụng các thông số Mức độ tương tác thay vì các yếu tố Tiền tệ. Điều này dẫn đến việc sử dụng RFE - một biến thể của RFM. Hơn nữa, thông số Cam kết này có thể được xác định là giá trị tổng hợp dựa trên các chỉ số như tỷ lệ thoát, thời lượng truy cập, số trang đã truy cập, thời gian dành cho mỗi trang, v.v.

Các yếu tố RFM minh họa những sự kiện này:

* mua hàng càng gần đây, khách hàng càng phản ứng nhanh hơn với các chương trình khuyến mãi
* khách hàng mua càng thường xuyên, họ càng tương tác và hài lòng hơn
* giá trị tiền tệ phân biệt những người chi tiêu nhiều và những người mua hàng có giá trị thấp

Phân tích phân đoạn RFM:

* **Nhà vô địch** là những khách hàng tốt nhất của bạn, những người đã mua gần đây nhất, thường xuyên nhất và là những người chi tiêu nhiều. Thưởng cho những khách hàng này. Họ có thể trở thành những người sớm chấp nhận các sản phẩm mới và sẽ giúp quảng bá thương hiệu của bạn.
* **Người trung thành tiềm năng** là những khách hàng gần đây của bạn với tần suất trung bình và những người đã chi tiêu nhiều. Cung cấp các chương trình thành viên hoặc khách hàng thân thiết hoặc giới thiệu các sản phẩm liên quan để bán thêm và giúp họ trở thành Khách hàng thân thiết hoặc Nhà vô địch của bạn.
* **Khách hàng mới** là khách hàng của bạn có điểm RFM tổng thể cao nhưng không phải là người mua sắm thường xuyên. Bắt đầu xây dựng mối quan hệ với những khách hàng này bằng cách cung cấp hỗ trợ giới thiệu và ưu đãi đặc biệt để tăng lượt truy cập của họ.
* **Khách hàng** có **rủi ro** là những khách hàng của bạn đã mua hàng thường xuyên và chi số tiền lớn, nhưng gần đây không mua. Gửi cho họ các chiến dịch kích hoạt lại được cá nhân hóa để kết nối lại, đồng thời đề nghị gia hạn và các sản phẩm hữu ích để khuyến khích mua hàng khác.
* **Không thể mất họ** là những khách hàng đã từng ghé thăm và mua hàng khá thường xuyên, nhưng gần đây không ghé thăm nữa. Đưa họ trở lại với các chương trình khuyến mãi có liên quan và thực hiện các cuộc khảo sát để tìm ra vấn đề và tránh để mất chúng vào tay đối thủ cạnh tranh.

## 3.2 Thuật Toán trong phân tích dự liệu

### 3.2.1 Thuật toán classifier

Classifier là quá trình nhận biết, hiểu và nhóm các ý tưởng và đối tượng thành các danh mục đặt trước hoặc “tiểu quần thể”. Sử dụng các tập dữ liệu đào tạo được phân loại trước, các chương trình học máy sử dụng nhiều thuật toán khác nhau để phân loại các tập dữ liệu trong tương lai thành các loại.

Các thuật toán phân loại trong học máy sử dụng dữ liệu đào tạo đầu vào để dự đoán khả năng dữ liệu tiếp theo sẽ thuộc một trong các danh mục được xác định trước. Một trong những cách sử dụng phổ biến nhất của phân loại là lọc email thành “spam” hoặc “không phải spam”.

Nói tóm lại, phân loại là một dạng “nhận dạng mẫu”, với các thuật toán phân loại được áp dụng cho dữ liệu huấn luyện để tìm ra mẫu giống nhau (từ hoặc tình cảm tương tự, chuỗi số, v.v.) trong các tập dữ liệu trong tương lai.

Sử dụng các thuật toán phân loại mà chúng ta sẽ đi vào chi tiết hơn bên dưới, phần mềm phân tích văn bản có thể thực hiện những việc như phân tích cảm tính để phân loại văn bản phi cấu trúc theo các cực của ý kiến ​​(tích cực, tiêu cực, trung lập và hơn thế nữa).

**Các loại thuật toán phân loại khác nhau:**

* Hồi quy logistic
* Naive Bayes Classifier
* K-Những người hàng xóm gần nhất
* Cây quyết định
* Rừng ngẫu nhiên
* Hỗ trợ Máy Vector

### 3.2.2 Thuật toán linear regression

Đây là dạng hồi quy đơn giản nhất. Nó là một kỹ thuật trong đó biến phụ thuộc có bản chất liên tục. Mối quan hệ giữa biến phụ thuộc và các biến độc lập được giả định là tuyến tính, chúng ta có thể quan sát rằng biểu đồ đã cho thể hiện mối quan hệ tuyến tính nào đó giữa quãng đường đi được và chuyển vị của ô tô. Các điểm màu xanh lá cây là các quan sát thực tế trong khi đường màu đen được trang bị là đường hồi quy.

|  |
| --- |
| [Phân tích hồi quy](https://4.bp.blogspot.com/-IOOxgPaXMVc/Wlj3LWvcnjI/AAAAAAAACKE/UeTFYvAxDmUDel5UBjdifeWaApB3-dXVgCLcBGAs/s1600/img1.jpg) |
| Phân tích hồi quy |

Khi bạn có chỉ có 1 biến độc lập và 1 biến phụ thuộc, nó được gọi là hồi quy tuyến tính đơn giản. Khi bạn có nhiều hơn 1 biến độc lập và 1 biến phụ thuộc, nó được gọi là hồi quy tuyến tính nhiều.

Phương trình của hồi quy tuyến tính bội được liệt kê dưới đây:

|  |
| --- |
| <https://2.bp.blogspot.com/-xbqTM5K3bIU/WkzhtHMPEmI/AAAAAAAACFs/RULnlMKw_0U14oRWOUcuETJNt9TBYiJEgCLcBGAs/s1600/b.jpg> |

Ở đây 'y' là biến phụ thuộc được ước lượng, và X là các biến độc lập và ε là thuật ngữ lỗi. βi là hệ số hồi quy.

Các giả định của hồi quy tuyến tính**:**

* Giữa các biến độc lập và phụ thuộc phải có mối quan hệ tuyến tính.
* Không nên có bất kỳ ngoại lệ nào.
* Không có phương sai thay đổi
* Các quan sát mẫu phải độc lập.
* Các thuật ngữ lỗi phải được phân phối bình thường với giá trị trung bình 0 và phương sai không đổi.
* Không có đa cộng tuyến và tự động tương quan.

### 3.2.3 Thuật toán RandomForestClassifier

Random Forests là thuật toán học có giám sát (supervised learning). Nó có thể được sử dụng cho cả phân lớp và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Một khu rừng bao gồm cây cối. Người ta nói rằng càng có nhiều cây thì rừng càng mạnh. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu. Nó cũng cung cấp một chỉ báo khá tốt về tầm quan trọng của tính năng. Random forests có nhiều ứng dụng, chẳng hạn như công cụ đề xuất, phân loại hình ảnh và lựa chọn tính năng. Nó có thể được sử dụng để phân loại các ứng viên cho vay trung thành, xác định hoạt động gian lận và dự đoán các bệnh. Nó nằm ở cơ sở của thuật toán Boruta, chọn các tính năng quan trọng trong tập dữ liệu.

### 3.2.4 Thuật toán ExtraTrees Classifier

Bộ phân loại cây cực kỳ ngẫu nhiên (Extra Trees Classifier) là một loại kỹ thuật học tập tổng hợp, tổng hợp kết quả của nhiều cây quyết định loại bỏ tương quan được thu thập trong một “khu rừng” để đưa ra kết quả phân loại. Về khái niệm, nó rất giống với Máy phân loại rừng ngẫu nhiên và chỉ khác nó ở cách xây dựng các cây quyết định trong rừng.

Mỗi Cây quyết định trong Rừng cây phụ được xây dựng từ mẫu đào tạo ban đầu. Sau đó, tại mỗi nút kiểm tra, Mỗi cây được cung cấp một mẫu ngẫu nhiên gồm k tính năng từ tập hợp đặc trưng mà từ đó mỗi cây quyết định phải chọn tính năng tốt nhất để tách dữ liệu dựa trên một số tiêu chí toán học (điển hình là Chỉ số Gini). Mẫu tính năng ngẫu nhiên này dẫn đến việc tạo ra nhiều cây quyết định không tương quan.

Để thực hiện lựa chọn đối tượng địa lý bằng cách sử dụng cấu trúc rừng ở trên, trong quá trình xây dựng rừng, đối với mỗi đối tượng địa lý, tổng số giảm chuẩn hóa của các tiêu chí toán học được sử dụng để quyết định đối tượng phân tách (Chỉ số Gini nếu Chỉ số Gini được sử dụng trong việc xây dựng rừng) được tính toán. Giá trị này được gọi là Tầm quan trọng Gini của đối tượng địa lý. Để thực hiện lựa chọn tính năng, mỗi tính năng được sắp xếp theo thứ tự giảm dần theo Mức độ quan trọng của từng tính năng và người dùng chọn k tính năng hàng đầu theo lựa chọn của mình.

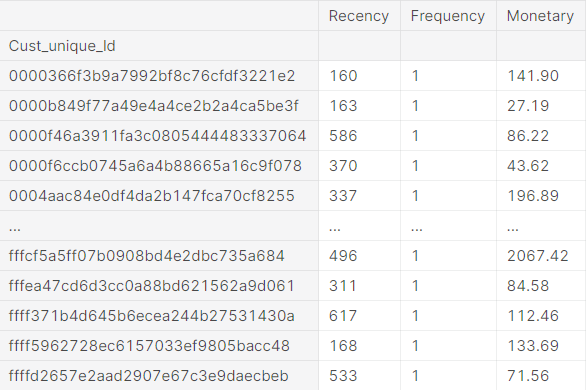
## 3.3 Áp dụng các phương pháp phân tích và thuật toán vào bộ dữ liệu ecommerce brazil

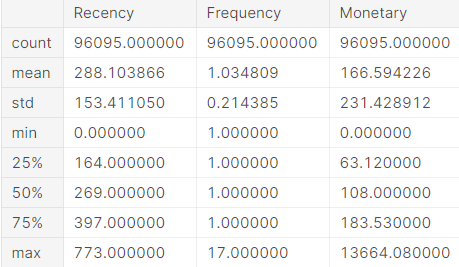
### 3.3.1 Xác định phân khúc khách hàng dựa trên RFM

Để xác định được phân khúc khách hàng ta dựa vào hành vi theo 3 điểm sau :

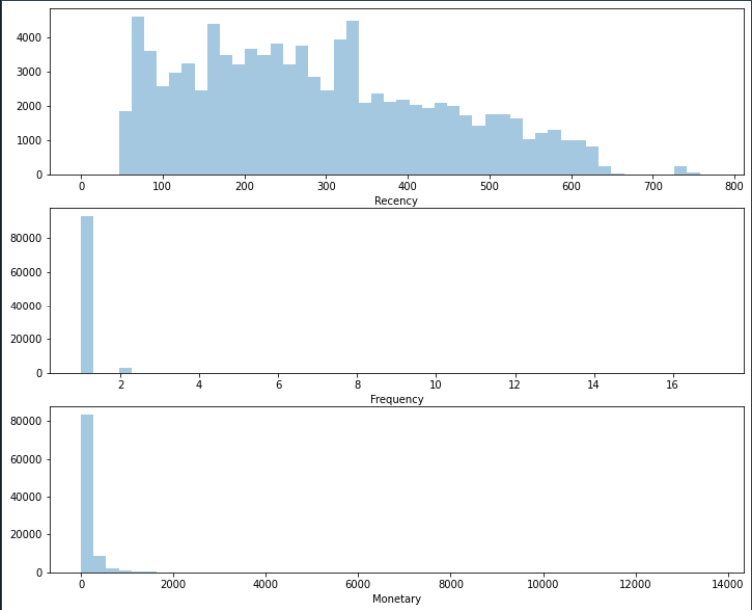
* Recency — Số ngày mua của khách kể từ lần mua cuối cùng
* Frequency — Tần suất mua hàng của khách hàng
* Monetary — Số tiền khách đã chi tiêu trong 1 khoảng thời gian

Kết quả thống kê khách hàng trên 3 yếu tố trên :

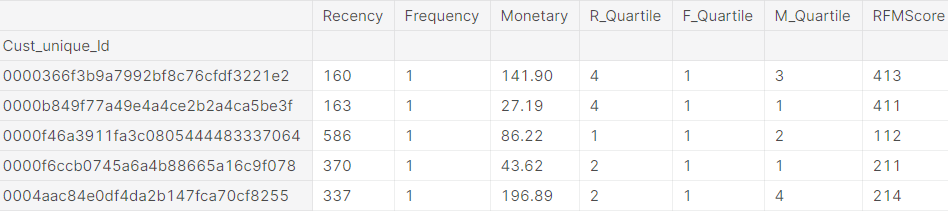


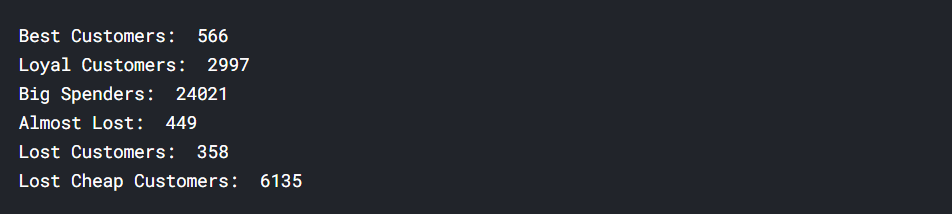


Biểu đồ thống kê :



Đánh giá phân khúc khách hàng sử dụng RFM quantity:



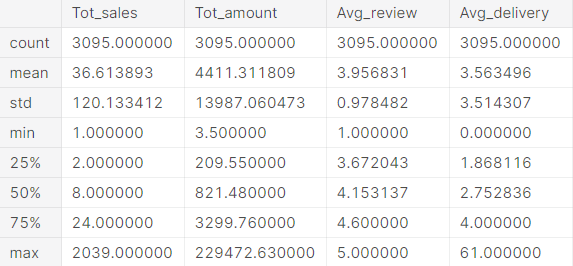


### 3.3.2 Đánh giá nhân viên bán hàng

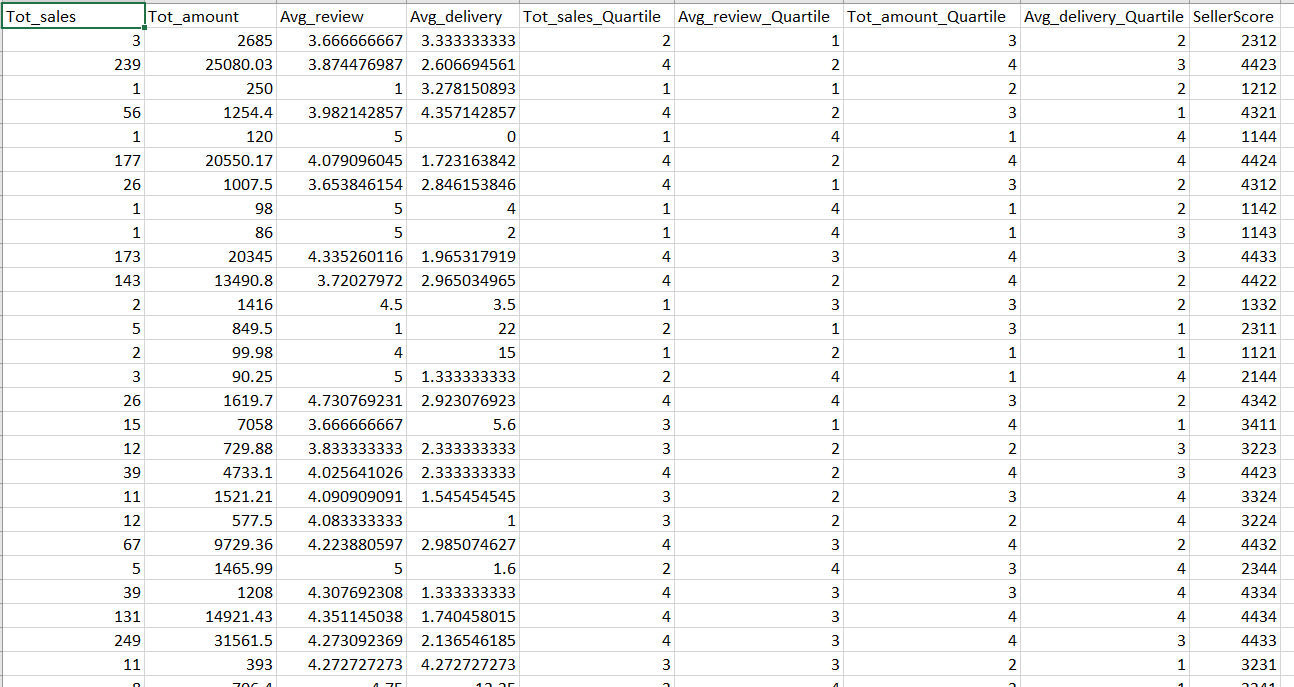
**Đánh giá người bán dựa trên các yếu tố sau :**

* Tổng số lần bán hàng đã được thực hiện
* Thời gian trung bình cần thiết để người bán cung cấp cho người vận chuyển
* Tổng tiền thực hiện trong tổng doanh thu
* Đánh giá nhận được cho đơn hàng

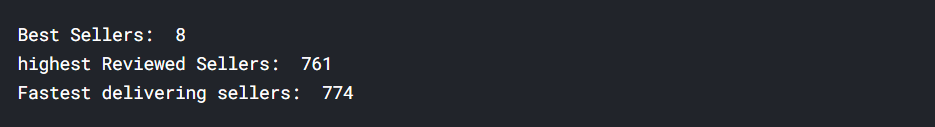
**Kết quả thống kê**



**Sử dụng RFM quartiles**



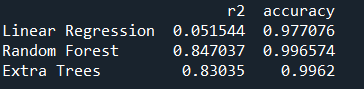
**Đánh giá nhân viên bán hàng :**



### 3.3.3 Dự đoán khả năng mua hàng của khách hàng

So sánh các model machine learning bao gồm : Linear regression, Random Forest, Extra Trees và Classifier để tìm ra model đánh giá tốt nhất.

**Kết quả so sánh :**





Như ta thấy điểm của Random Forest gần với 1 nhất, suy ra mô hình Random Forest có khả năng dự đoán xác suất mua hàng tốt nhất. Vậy Random Forest là công cụ dự đoán tốt nhất