**ĐẠI HỌC ĐÀ NẴNG**

***TRƯỜNG ĐẠI HỌC KINH TẾ***

🙞🙜🕮🙞🙜

****

**ĐỀ CƯƠNG CHUYÊN ĐỀ**

**ỨNG DỤNG BUSINESS INTELLIGENCE (BI)   
PHÂN TÍCH MỨC ĐỘ HÀI LÒNG CỦA KHÁCH HÀNG VỀ CHẤT LƯỢNG VÀ DỊCH VỤ MUA XE Ở CỬA HÀNG Ô TÔ HONDA ĐÀ NẴNG**

***SVTH: Trần Ngọc Thái Sơn***

***Lớp: 42K21***

***GVHD: ThS. Cao Thị Nhâm***

***Đà Nẵng, tháng 09 năm 2020.***

NHẬN XÉT CỦA DOANH NGHIỆP THỰC TẬP

1. ***Về ý thức và tinh thần trách nhiệm của sinh viên khi thực tập tại doanh nghiệp:***

1. ***Về năng lực và khả năng thích nghi của sinh viên với môi trường doanh nghiệp:***

1. ***Kiến nghị của Doanh nghiệp đối với Nhà trường để cải thiện chất lượng đào tạo:***

LỜI CẢM ƠN

Lời đầu tiên, em xin chân thành gửi lời cảm ơn tới cô Cao Thị Nhâm – Giảng viên khoa Thống kê – Tin học trường Đại học Kinh tế Đà Nẵng đã nhiệt tình hướng dẫn để em có thể hoàn thành tốt bài báo cáo này.

Em xin chân thành cảm ơn tất cả thầy cô khoa Thống Kê – Tin Học cùng các thầy cô trường Đại Học Kinh Tế - Đại Học Đà Nẵng đã tận tình giảng dạy, truyền đạt nhiệt huyết cho em trong suốt bốn năm học qua.

Em xin gửi lời cảm ơn chân thành tới anh Dương Văn Toàn – Giám đốc điều hành công ty HONDA chi nhánh Đà Nẵng đã tạo điều kiện cung cấp số liệu để em hoàn thành tốt bài báo cáo tốt nghiệp của mình.

Bài báo cáo thực tập được trong khoảng 3 tháng và em mong nhận được những góp ý chân thật từ thầy cô và có thể bỏ qua những sai sót trong bài báo cáo để em có thể hoàn thiện kiến thức và có thể hoàn thành kỳ báo cáo thực tập được suôn sẻ nhất.

Cuối cùng, em xin chúc quý thầy cô Trường Đại Học Kinh Tế Đà Nẵng nhiều sức khỏe và thành công.

Em xin chân thành cảm ơn!

***Sinh Viên thực hiện***

***Trần Ngọc Thái Sơn***

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi thực hiện dưới sự hướng dẫn của cô Cao Thị Nhâm. Các số liệu, kết quả nêu trong Báo cáo là hoàn toàn trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác. Tất cả tài liệu được tôi tham khảo đều sẽ được trích dẫn nguồn gốc rõ ràng.

***Sinh viên thực hiện***

***Trần Ngọc Thái Sơn***

# **MỤC LỤC**

[**NHẬN XÉT CỦA DOANH NGHIỆP THỰC TẬP 2**](#_Toc52117785)

[**LỜI CẢM ƠN 3**](#_Toc52117786)

[**LỜI CAM ĐOAN 4**](#_Toc52117787)

[**MỤC LỤC 5**](#_Toc52117788)

[**DANH MỤC TỪ VIẾT TẮT 5**](#_Toc52117789)

[**DANH MỤC BẢNG BIỂU 6**](#_Toc52117790)

[**DANH MỤC HÌNH ẢNH 7**](#_Toc52117791)

[**MỞ ĐẦU 8**](#_Toc52117792)

[**1. Lý do chọn đề tài 8**](#_Toc52117793)

[**2. Đối tượng và phạm vi nghiên cứu 9**](#_Toc52117794)

[**3. Mục tiêu nghiên cứu 9**](#_Toc52117795)

[**4. Phương pháp nghiên cứu 9**](#_Toc52117796)

[**5. Cấu trúc báo cáo 9**](#_Toc52117797)

[**CHƯƠNG 1: GIỚI THIỆU VỀ CÔNG TY Ô TÔ HONDA 10**](#_Toc52117798)

[**1.1 Giới thiệu về công ty Ô tô Honda Đà Nẵng 10**](#_Toc52117799)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 14**](#_Toc52117800)

[**2.1 Tổng quan về business intelligence (BI) 14**](#_Toc52117801)

[**2.1.1 Business Intelligence (BI) 14**](#_Toc52117802)

[**2.1.2 Lợi ích của BI 14**](#_Toc52117803)

[**2.1.3 Những thành phần chính của Business Intelligence (BI) 15**](#_Toc52117804)

[**2.2 Tổng quan về khai phá dữ liệu (Data Mining) 17**](#_Toc52117805)

[**2.2.1 Khai phá dữ liệu (Data Mining) 17**](#_Toc52117806)

[**2.2.2 Các phương pháp chính trong khai phá dữ liệu 17**](#_Toc52117807)

[**2.2.3 Quy trình khám phá tri thức 18**](#_Toc52117808)

[**2.2.4 Quy trình khai phá dữ liệu 18**](#_Toc52117809)

[**2.3 Tổng quan về luật kết hợp 19**](#_Toc52117810)

[**2.3.1 Định nghĩa về luật kết hợp 19**](#_Toc52117811)

[**2.3.2 Luật kết hợp 21**](#_Toc52117812)

[**2.3.3 Thuật toán Apriori 22**](#_Toc52117813)

[**CHƯƠNG 3: GIỚI THIỆU VỀ CÁC CÔNG CỤ SỬ DỤNG 26**](#_Toc52117814)

[**3.1 Tổng quan về công cụ trực quan hóa Power BI 26**](#_Toc52117815)

[**3.2 Giới Thiệu SQL Server 2014 28**](#_Toc52117816)

[**3.3 Giới thiệu công cụ Visual Studio 2013 31**](#_Toc52117817)

[**3.4 Microsoft SQL Server Analysis Services (SSAS) 32**](#_Toc52117818)

[**3.5 Môi trường triển khai ứng dụng (BIDS) 33**](#_Toc52117819)

[**CHƯƠNG 4: PHÂN TÍCH PHẢN HỒI KHÁCH HÀNG VỀ VIỆC MUA XE TẠI CỬA HÀNG Ô TÔ HONDA ĐÀ NẴNG 34**](#_Toc52117820)

[**4.1 Thực trạng và yêu cầu 34**](#_Toc52117821)

[**4.2 Đề xuất quy trình 34**](#_Toc52117822)

[**4.2.1 Thu thập dữ liệu 34**](#_Toc52117823)

[**4.2.2 Thiết kế dữ liệu 34**](#_Toc52117824)

[**4.2.3 Cài đặt dữ liệu vào hệ thống 34**](#_Toc52117825)

[**4.2.4 Trực quan hóa dữ liệu 34**](#_Toc52117826)

[**4.2.5 Xây dựng mô hình khai phá dữ liệu trên BIDS 34**](#_Toc52117827)

[**4.2.6 Ứng dụng luật kết hợp phân tích độ hài lòng của khách hàng 34**](#_Toc52117828)

[**CHƯƠNG 5 : KẾT QUẢ PHÂN TÍCH 34**](#_Toc52117829)

[**1.1 Kết quả đánh giá dữ liệu báo cáo BI 34**](#_Toc52117830)

[**1.2 Kết quả đánh giá mô hình luật kết hợp 34**](#_Toc52117831)

[**KẾT LUẬN 34**](#_Toc52117832)

[**ĐỀ XUẤT HƯỚNG PHÁT TRIỂN 34**](#_Toc52117833)

[**TÀI LIỆU THAM KHẢO 34**](#_Toc52117834)

[**PHỤ LỤC 34**](#_Toc52117835)

DANH MỤC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Từ viết tắt** | **Diễn giải** |
| SQL | Structure Query Language |
| BI | Business Intelligence |
| SSAS | SQL Server Analysis Services |
| BIDS | Business Intelligence Development Studio |

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH ẢNH

# **MỞ ĐẦU**

1. **Lý do chọn đề tài**

Trong những năm gần đây nhu cầu mua sắm xe ô tô của người Việt Nam ngày càng trở nên phát triển, lý do đ ầu tiên là do cơ chế mở cửa của Nhà nước, thứ hai vì nhu cầu phục vụ của loại phương tiện này ngày càng lớn với nhiều mục đích khác nhau: phục vụ đi lại, phục vụ nghành vận tải, giao dịch và có thể là để thể hiện đẳng cấp,....

Dự kiến từ đầu năm 2020, ô tô nhập khẩu sẽ được loại bỏ quy định như kiểm tra theo lô, giấy chứng nhận chất lượng kiểu loại, sẽ càng làm cho hoạt động nhập khẩu dễ dàng hơn với chi phí giảm, giá xe có điều kiện giảm tiếp.

Hiện nay cũng có rất nhiều chính sách trả góp lên đến 7 năm và đang có đề xuất giảm 50% phí trước bạ cho các loại ô tô trong nước nên nhu cầu ô tô trong tương lai có thể sẽ rất cao. Do vậy cần khảo sát ý kiến của khách hàng để phân tích và đánh giá, từ đó đưa ra được những kế hoạch phù hợp trong tương lai nhằm nâng cao dịch vụ và chất lượng xe đồng thời cũng cải thiện các chính sách ưu đãi, chăm sóc khách hàng, tăng giá trị và các lợi thế cạnh tranh cho doanh nghiệp trong tương lai.

Từ những kiến thức đã được trao dồi và được công ty Honda Đà Nẵng cung cấp dữ liệu, tôi mong muốn tiến hành những đánh giá về mức độ hài lòng của khách hàng về dịch vụ và chất lượng xe khi mua xe tại các chi nhánh Honda Đà Nẵng, nhờ vào các công cụ khai phá dữ liệu để khai thác những thông tin hữu ích từ kho dữ liệu của công ty, qua đó có thể đưa ra những định hướng đúng đắn trong quá trình kinh doanh cũng như nâng cao lòng tin của khách hàng khi sử dụng dịch vụ.

Vì vậy tôi đã chọn đề tài “Ứng dụng business intelligence (BI) phân tích mức độ hài lòng của **khách hàng về chất lượng và dịch vụ khi mua xe ở cửa hàng Ô tô Honda Đà Nẵng** ” làm đề tài báo cáo tốt nghiệp cho mình.

1. **Đối tượng và phạm vi nghiên cứu**

* **Đối tượng**

Khách hàng mua xe ở cửa hàng ô tô Honda chi nhánh Đà Nẵng.

* **Phạm vi nghiên cứu**

Đánh giá độ hài lòng về dịch vụ và chất lượng xe của khách hàng mua xe ở cửa hàng Ô tô Honda chi nhánh Đà Nẵng.

1. **Mục tiêu nghiên cứu**

Bài báo cáo này được thực hiện nhằm tìm hiểu về Business Intelligence, sử dụng luật kết hợp trong khai phá dữ liệu để ứng dụng vào viêc phân tích mức độ hài lòng của khách hàng về dịch vụ và chất lượng khi mua xe ở cửa hàng Honda Đà Nẵng. Từ đó đưa ra các định hướng, kế hoạch hợp lý cho doanh nghiệp.

1. **Phương pháp nghiên cứu**

Sử dụng các phần mềm chuyên dụng Power BI, Visual Studio và SQL Server phân tích thông tin bằng dữ liệu đã thu thập được tại công ty Honda chi nhánh Đà Nẵng

1. **Cấu trúc báo cáo**

CHƯƠNG 1: Giới thiệu về Honda ô tô Đà Nẵng.

CHƯƠNG 2: Cơ sở lý thuyết.

CHƯƠNG 3: Giới thiệu về các công cụ sử dụng.

CHƯƠNG 4: Phân tích phản hồi khách hàng về việc mua xe tại cửa hàng ô tô Honda Đà Nẵng.

CHƯƠNG 5: Kết quả phân tích.

# **CHƯƠNG 1**: **GIỚI THIỆU VỀ CÔNG TY Ô TÔ HONDA**

## **Giới thiệu về công ty Ô tô Honda Đà Nẵng**

Được thành lập vào năm 1996, công ty **Honda Việt Nam** là công ty liên doanh giữa 3 đơn vị: Công ty Honda Motor (Nhật Bản), Công ty Asian Honda Motor (Thái Lan), Tổng Công ty Máy Động Lực và Máy Nông nghiệp Việt Nam với 2 ngành sản phẩm chính: xe máy và xe ô tô. Sau hơn 20 năm có mặt tại Việt Nam, Honda Việt Nam đã không ngừng phát triển và trở thành một trong những đơn vị hàng đầu trong lĩnh vực sản xuất xe gắn máy và ô tô uy tín tại thị trường Việt Nam.



Bắt đầu hoạt động kinh doanh ô tô từ năm 2006, chỉ sau hơn 1 năm, Honda Việt Nam đã xây dựng thành công nhà máy, mạng lưới đại lý, các chương trình đào tạo bán hàng, dịch vụ, lái xe an toàn cho nhân viên các đại lý và ra mắt mẫu xe đầu tiên là Honda Civic vào tháng 8 năm 2006. Không ngừng nỗ lực để đa dạng hóa sản phẩm, mẫu xe Honda CR-V tiếp tục được Honda Việt Nam giới thiệu vào tháng 12 năm 2008 và Honda City vào tháng 6 năm 2013. Ngoài những dòng xe sản xuất trong nước, Honda Việt Nam còn nhập khẩu thêm các mẫu xe Sedan và mẫu xe đa dụng cao cấp lần lượt là Honda Accord và Honda Odyssey nhằm đáp ứng nhu cầu ngày càng cao của khách hàng.

**THÔNG TIN NHÀ MÁY HONDA Ô TÔ VIỆT NAM:**

* Thành lập:  Năm 2005.
* Trụ sở: Xã Phúc Thắng, Huyện Phúc Yên, Tỉnh Vĩnh Phúc, Việt Nam.
* Lao động: 449 người (tháng 09/2017).
* Vốn đầu tư: Khoảng 60 triệu USD.
* Công suất: 10.000 xe/năm.
* Diện tích: 30.000m2.
* **Đại lý Honda chi nhánh Đà Nẵng**

Đại lý Honda ô tô Đà Nẵng thành lập vào ngày 19/10/2007; tọa lạc tại địa chỉ 178, Đường 2/9, Quận Hải Châu, TP Đà Nẵng, Tỉnh Đà Nẵng. Hiện tại, showroom Honda ô tô Đà Nẵng đang cung cấp các sản phẩm ô tô Honda chính hãng, phụ tùng chính hãng cùng với các dịch vụ bảo hành sửa chữa và chăm sóc khách hàng cho toàn bộ khu vực miền trung Việt Nam.



Sau hơn 4 năm thành lập và đi vào hoạt động, so với nhiều hãng xe khác thì showroom Honda Đà Nẵng còn khá non trẻ. Tuy nhiên, công ty đã hoàn thành sứ mệnh và gặt hái được nhiều thành công nhất định, đóng góp vào thành tích của toàn công ty. Honda Ô tô Đà Nẵng có tổng diện tích sử dụng trên 4.500 m2 bao gồm các khu vực như phòng trưng bày xe mới, khu vực tiện ích phục vụ khách hàng, khu vực xưởng dịch vụ, khu vực văn phòng, phòng đào tạo, kho xe mới vv… Do đó, showroom Honda ô tô Đà Nẵng hoàn toàn đáp ứng đầy đủ những yêu cầu khắt khe đối với một showroom đạt tiêu chuẩn 5S của Honda.



Các dịch vụ bán hàng và sau bán hàng của Honda Ô tô Đà Nẵng luôn tuân thủ những quy chuẩn quốc tế chuyên nghiệp nhất, luôn hướng dẫn đến quyền lợi và sự hài lòng của khách hàng nhằm nâng cao uy tín của Honda Ô tô Đà Nẵng nói riêng và hệ thống đại lý ô tô Honda trên toàn quốc nói chung.

# **CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

## **Tổng quan về business intelligence (BI)**

### **Business Intelligence (BI)**

BI là quy trình và công nghệ mà các doanh nghiệp sử dụng để kiểm soát khối lượng dữ liệu khổng lồ, khai phá tri thức giúp cho các doanh nghiệp có thể đưa các các quyết định hiệu quả hơn trong hoạt động kinh doanh của mình. [**Hệ thống báo cáo quản trị thông minh (BI)**](https://ierp.vn/) còn tự động chuyển dữ liệu thành các dạng dashboards (bảng báo cáo), hình ảnh visual, …cho phép người dùng xem và hiểu dữ liệu một cách dễ dàng hơn.

BI cung cấp một cách nhìn toàn cảnh hoạt động của doanh nghiệp từ quá khứ, hiện tại và các dự đoán trong tương lai. [**Lợi ích của BI**](https://ierp.vn/bi-tableau-mang-den-nhung-loi-ich-lon-cho-doanh-nghiep) là hỗ trợ cho doanh nghiệp ra quyết định tốt hơn. BI hỗ trợ tối đa cho nhà quản trị trong quá trình hoạch định chiến lược của doanh nghiệp, giúp thay đổi kỹ năng điều hành từ kinh nghiệm chủ quan bằng cách điều hành dựa trên số liệu và thông tin chính xác thu được từ dữ liệu.

* + 1. **Lợi ích của BI**
* Với việc tổng hợp số liệu nhanh chóng, việc tạo báo cáo một cách trực quan, hệ thống báo cáo quản trị thông minh BI có khả năng đem đến cho nhà quản trị những thông tin, báo cáo gần như ngay lập tức và vô cũng dễ hiểu so với việc phải tốn rất nhiều thời gian như trước đây.
* Quản lý nhiều mảng khác nhau cùng trong 1 bảng báo cáo.
* Giúp các doanh nghiệp sử dụng thông tin một cách hiệu quả, chính xác để thích ứng với môi trường thay đổi liên tục và cạnh tranh khốc liệt trong kinh doanh.
* Hỗ trợ nhà quản trị tối đa trong việc đưa ra các quyết định kinh doanh nhanh chóng, kịp thời, hiệu quả.
* Phân tích hành vi khách hàng.
* Xây dựng chiến lược kinh doanh.
* Tạo lợi thế cạnh tranh, gia tăng cơ hội tìm kiếm và nắm bắt các cơ hội kinh doanh.
  + 1. **Những thành phần chính của Business Intelligence (BI)**

A screenshot of a cell phone

Description automatically generated

* **Data Sources**

- Là cơ sở dữ liệu thô (thường là cơ sở dữ liệu quan hệ) đến từ nhiều nguồn khác nhau như các ứng dụng business như Human Resource Management (HRM), Customer relationship management (CRM), phần mềm bán hàng, website thương mại điện tử…

- Có thể là bất cứ hệ quản trị cơ sở dữ liệu nào như MySQL, Oracle, MSSQL, DB2, …

- Thường được thiết kế theo mô hình cơ sở dữ liệu quan hệ nhưng cũng có thể là dữ liệu lớn, dữ liệu phi quan hệ (như mạng xã hội, NoSQL).

* **Data Warehouse**

- Là cơ sở dữ liệu được thiết kế theo mô hình khác với CSDL OLTP thông thường và là nơi lưu trữ dữ liệu lâu dài của tổ chức.

- Dữ liệu của DWH chỉ có thể đọc, không được sử dụng để ghi hay update bởi ứng dụng thông thường,  nó chỉ được cập nhật/ghi bởi công cụ ETL (Extract Transform Load), công cụ chuyển đổi dữ liệu từ Data Sources vào Data Warehouse.

* **Integrating Server**

- Chịu trách nhiệm trung gian vận hành công cụ ETL để chuyển đổi dữ liệu từ Data Sources vào Data Warehouse.

* **Analysis Server**

- Chịu trách nhiệm thực thi các cube được thiết kế dựa trên các chiều dữ liệu và tri thức nghiệp vụ.

- Cube chịu trách nhiệm nhận dữ liệu đầu vào từ DWH và thực thi theo nghiệp vụ định nghĩa sẵn để trả về kết quả.

* **Reporting Server**

- Thực thi các report với output nhận được từ Analysis Server.

- Nơi quản trị tập trung các report trên nền web, các report này có thể được attach vào ứng dụng web, hay application.

* **Data Mining**

- Là quá trình trích xuất thông tin dữ liệu đã qua xử lý (phù hợp với yêu cầu riêng của doanh nghiệp) từ Data Warehouse rồi kết hợp với các thuật toán để đưa ra ( hoặc dự đoán ) các quyết định có lợi cho việc kinh doanh của doanh nghiệp.

- Đây là một quá trình quan trọng trong BI, thông thường một doanh nghiệp muốn sử dụng giái pháp BI thường kèm theo về Data Mining.

* **Data Presentation**

- Tạo ra các báo cáo, biểu đồ từ quá trình data mining để phục vụ cho nhu cầu của người dùng cuối.

## **Tổng quan về khai phá dữ liệu (Data Mining)**

### **Khai phá dữ liệu (Data Mining)**

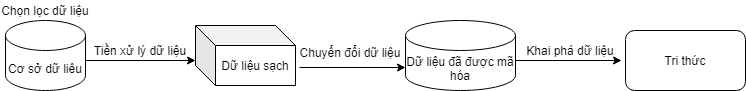
- Data Mining hay còn gọi là khai phá dữ liệu là quy trình mà các doanh nghiệp sử dụng để biến những dữ liệu thô thành thông tin hữu ích. Bằng cách dùng phần mềm để tìm tòi, khai thác và phân tích các khối dữ liệu lớn để chắt lọc ra được những mối liên hệ lẫn nhau giữa các dữ liệu từ đó đưa ra được mẫu hình và xu hướng có ý nghĩa.

### **Các phương pháp chính trong khai phá dữ liệu**

* Luật kết hợp.
* Phương pháp cây quyết định.
* Các phương pháp dựa trên mẫu.
* Phương pháp K-Mean.
* Phân lớp.
* Phân cụm.
* Gom cụm.
* Phân tích độ lệch.
* Hồi quy.

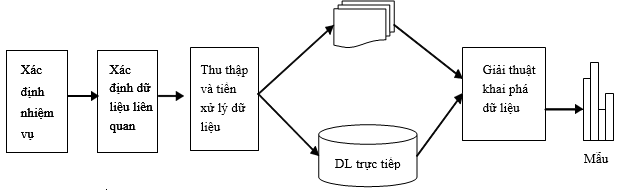
Trong đó có ba phương pháp được các nhà nghiên cứu sử dụng nhiều nhất đó là: Luật kết hợp, Phân lớp dữ liệu và Phân cụm dữ liệu.

### **Quy trình khám phá tri thức**

****

* **Quy trình khám phá tri thức bao gồm các bước như sau:**
* **Chọn lọc dữ liệu:** đây là giai đoạntập hợp các dữ liệu được khai thác từ một CSDL, một kho dữ liệu hay là từ csac nguồn ứng dụng web vào một CSDL riêng, chúng ta chỉ chọn ra những dữ liệu cần thiết cho giai đoạn sau.
* **Tiền xử lý dữ liệu (preprocessing):** phần lớn các CSDL đều ít nhiều mang tính không nhất quán.Vì vậy khi gom dữ liệu có thể mắc một số lỗi như thiếu dữ liệu, không logic, bị trùng lặp,... Do đó cần phải được tiền xử lý trước khi khai phá dữ liệu nếu không sẽ gây nên kết quả sai lệch.
* **Chuyển đổi dữ liệu (Transformation):** trong gia đoạn này dữ liệu sẽ được chuyển đổi về dạng thuận tiện để tiến hành các thuật toán khám phá dữ liệu.
* **Khai phá dữ liệu (Data Mining):** trong giai đoạn này ta sử dụng các kỹ thuật nhằm phát hiện ra các tri thức tiềm ẩn trong dữ liệu, một số kỹ thuật được sử dụng như: phân lớp, gom cụm, luật kết hợp,...
* **Đánh giá kết quả mẫu:** đây là giai đoạn cuối, trong giai đoạn này các mẫu dữ liệu không được chiết xuất bởi các phần mềm khai phá dữ liệu. Không phải mẫu nào cũng đều có ích, thậm chí còn bị sai lệch. Vì vậy cần xác định và lựa chọn những tiêu chuẩn đánh giá sao cho chiết xuất ra được các tri thức cần thiết.

### **Quy trình khai phá dữ liệu**



* **Quy trình khai phá dữ liệu bao gồm các bước sau:**
* **Nghiên cứu lĩnh vực và xác định nhiệm vụ:** ta cần nghiên cứu lĩnh vực cần sử dụng Data mining để xác định được những tri thức ta cần chắt lọc, xác định chính xác những vấn đề cần giải quyết.
* **Xác định các dữ liệu liên quan:** để tìm ra và xây dựng các giải pháp để phân tích.
* **Thu thập và tiền xử lý dữ liệu**: ta tiến hành bỏ bớt những dữ liệu không cần thiết, tinh chỉnh lại cấu trúc của dữ liệu và mã hóa để tiện cho việc xử lý.
* **Rút gọn chiều**: Thông thường một tập dữ liệu có chiều khá lớn sẽ sinh ra một lượng dữ liệu khổng lồ, ví dụ với n chiều ra sẽ có 2^n tổ hợp. Do đó, đây là bước quan trọng giúp đáng kể hao tổn về tài nguyên trong quá trình xử lý tri thức. Thông thường ta sẽ dùng Rough set để giảm số chiều.
* **Chọn tác vụ khai thác dữ liệu**: Để đạt được mục đích ra cần chọn được tác vụ khai phá dữ liệu sao cho phù hợp.

## **Tổng quan về luật kết hợp**

### **Định nghĩa về luật kết hợp**

Trong lĩnh vực Data Mining, mục đích của luật kết hợp (Association Rule - AR) là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu. Nội dung cơ bản của luật kết hợp được tóm tắt như dưới đây.

Cho cơ sở dữ liệu gồm các giao dịch T là tập các giao dịch t1, t2, …, tn.

**T = {t1, t2, …, tn}. T gọi là cơ sở dữ liệu giao dịch (Transaction Database)**.

Mỗi giao dịch ti bao gồm tập các đối tượng I (gọi là itemset).

**I = {i1, i2, …, im}.**Một itemset gồm k items gọi là k-itemset.

Mục đích của luật kết hợp là tìm ra sự kết hợp (association) hay tương quan (correlation) giữa các items. Những luật kết hợp này có dạng***X =>Y***

Trong Basket Analysis, luật kết hợp X =>Y có thể hiểu rằng những người mua các mặt hàng trong tập X cũng thường mua các mặt hàng trong tập Y. (X và Y gọi là itemset).

Ví dụ, nếu X = {Apple, Banana} và Y = {Cherry, Durian} và ta có luật kết hợp X =>Y thì chúng ta có thể nói rằng những người mua Apple và Banana thì cũng thường mua Cherry và Durian.

Theo quan điểm thống kê, X được xem là biến độc lập (Independent variable) còn Y được xem là biến phụ thuộc (Dependent variable).

**Độ hỗ trợ (Support) và độ tin cây (Confidence) là 2 tham số dùng để đo lường luật kết hợp.**

Độ hỗ trợ (Support) của luật kết hợp X =>Y là tần suất của giao dịch chứa tất cả các items trong cả hai tập X và Y. Ví dụ, support của luật X =>Y là 5% có nghĩa là  5% các giao dịch X và Y được mua cùng nhau.

Công thức để tính support của luật X =>Y như sau:

http://bis.net.vn/photos/storage/image0011.png

Trong đó: N là tổng số giao dịch.

Độ tin cậy (Confidence) của luật kết hợp X =>Y là xác suất xảy ra Y khi đã biết X. Ví dụ độ tin cậy của luật kết hợp {Apple} =>Banana} là 80% có nghĩa là 80% khách hàng mua Apple cũng mua Banana.

Công thức để tính độ tin cậy của luật kết hợp X =>là xác suất có điều kiện Y khi đã biết X như sau:

http://bis.net.vn/photos/storage/image005.gif

Trong đó: n(X) là số giao dịch chứa X

* **Một số tính chất liên quan đến các hạng mục phổ biến**

***Tính chất 1****:* (Độ hỗ trợ của tập con)

* Với A và B là tập các mục, nếu A ⊆ B thì sup(A) ≥ sup(B).
* Điều này là rõ ràng vì tất cả các giao tác của D hỗ trợ B thì cũng hỗ trợ A.

***Tính chất 2:***

* Một tập chứa một tập không phổ biến thì cũng là tập không phổ biến.
* Nếu một mục trong B không có độ hỗ trợ tối thiểu trên D nghĩa là sup(B)< minsup thì một tập con A của B sẽ không phải là một tập phổ biến vì support(B) ≤ support(A) < minsup (theo tính chất 1).

***Tính chất 3:*** Các tập con của tập phổ biến cũng là tập phổ biến.

Nếu mục B là mục phổ biến trên D, nghĩa là support(B) ≥ minsup thì mọi tập con A của B là tập phổ biến trên D vì support(A) ≥ support(B) > minsup.

### **Luật kết hợp**

***Tính chất 1:***( Không hợp các luật kết hợp)

Nếu có X→Z và Y→Z trong D thì không nhất thiết X∪Y→Z là đúng.

Xét trường hợp X ∩Z =∅ và các tác vụ trong D hỗ trợ Z nếu và chỉ nếu chúng hỗ trợ mỗi X hoặc Y, khi đó luật X∪Y→Z có độ hỗ trợ 0%.

Tương tự : X→Y ∧ X→Z ⇒ X→Y∪Z

***Tính chất 2:*** (Không tách luật)

Nếu X∪Y→Z thì X→Z và Y→Z chưa chắc xảy ra

Ví dụ trường hợp Z có mặt trong một giao tác chỉ khi cả hai X và Y cũng có mặt, tức là sup(X∪Y)= sup(Z), nếu độ hỗ trợ của X và Y đủ lớn hơn sup(X∪Y), tức là sup(X) > sup(X∪Y) và sup(Y) > sup(X∪Y) thì hai luật riêng biệt sẽ không đủ độ tin cậy.

Tuy nhiên đảo lại: X→Y∪Z ⇒ X→Y ∧ X→Z

***Tính chất 3:*** (Các luật kết hợp không có tính bắc cầu)

Nếu X→Y và Y→Z, chúng ta không thể suy ra X→Z.

Ví dụ: giả sử T(X) ⊂ T(Y) ⊂ T(Z), ở đó T(X), T(Y), T(Z) tương ứng là các giao dịch chứa X,Y,Z, và độ tin cậy cực tiểu minconf.

conf(X→Y) = conf(Y→Z) = minconf thế thì: conf(X→Y) = minconf2 < minconf vì minconf < 1, do đó luật X→Z không đủ độ tin cậy.

***Tính chất 4:***

Nếu A→ (L - A) không thoả mãn độ tin cậy cực tiểu thì luật.

B → (L -B) cũng không thoả mãn, với các tập mục L, A, B và B ⊆ A ⊂ L

Vì supp(B) ≥ sup(A) (theo tính chất 1) và định nghĩa độ tin cậy, chúng ta nhận được: 

Cũng như vậy: Nếu có (L-C)→ C thì ta cũng có luật (L – D) →D, với D⊆C và D≠∅.

Bởi vì D ⊆ C nên (L - D) ⊇ (L - C), do đó sup(L - D) ≤ sup(L-C)

⇒≥ minconf

### **Thuật toán Apriori**

Để hình dung rõ hơn các giai đoạn trong quá trình phát hiện luật kết hợp, chúng ta sẽ xem xét giải thuật Apriori phát hiện luật kết hợp từ cơ sở dữ liệu tác vụ.

* **Tính chất Apriori:**

Mọi tập con khác rỗng của tập mục phổ biến cũng phải phổ biến. Ví dụ, nếu {bia, bỉm, hạnh nhân} là phổ biến thì {bia, bỉm} cũng vậy, mọi giao dịch chứa {bia, bỉm, hạnh nhân} cũng chứa {bia, bỉm}.

* **Giải thuật Apriori:**

Duyệt (Scan) toàn bộ cơ sở dữ liệu giao dịch để có được độ hỗ trợ (support) S của 1- itemset, so sánh S với độ hỗ trợ tối thiểu min\_sup, để có được 1-itemset (F1).

Sử dụng Fk-1 nối (kết nối) Fk-1 để sinh ra candidate k-itemset. Loại bỏ các itemsets không phải là tập phổ biến thu được k-itemset (bước cắt tỉa).

Tiếp tục duyệt cơ sở dữ liệu giao dịch để có được độ hỗ trợ S của mỗi candidate k-itemset, so sánh S với min\_sup để thu được tập phổ biến k –itemset (Fk).

Lặp lại từ bước 2 cho đến khi Candidate set (C) trống (không tìm thấy tập phổ biến).

Với mỗi tập phổ biến I, sinh tất cả các tập con s không rỗng của I.

Với mỗi tập con s không rỗng của I, sinh ra các luật s => (I-s) nếu độ tin cậy (Confidence) của nó > =min\_conf.

* **Thuật toán Apriori:**

**Các kí hiệu**:

Lk: Tập các k-mục phổ biến (large k-itemset) (tức tập các itemset có sup tối thiểu và có lực lượng bằng k).

Mỗi phần tử của tập này có 2 trường: itemset và support-count.

Ck: Tập các candidate k-itemset (tập các tập k-mục ứng cử viên). Mỗi phần tử trong tập này cũng có 2 trường itemset và support-count.

Nội dung thuật toán Apriori được trình bày như sau:

**Input:** Tập các giao dịch D, ngưỡng support tối thiểu minsup.

**Output:** L- tập mục phổ biến trong D.

**Method:**

L1={large 1-itemset (tập 1- mục phổ biến)} //tìm tất cả các tập mục phổ biến: nhận được L1.

**For** (k=2; Lk-1 ≠ ∅; k++) **do**

**Begin**

Ck=apriori-gen(Lk-1); //sinh ra tập ứng cử viên từ Lk-1

**For** (mỗi một giao dịch TD) **do**

**Begin**

CT = subset(Ck, T); //lấy tập con của T là ứng cử viên trong Ck

**For** (mỗi một ứng cử viên c CT) **do**

c.count++; //tăng bộ đếm tần xuất 1 đơn vị

**End;**

Lk = {c ∈ Ck| c.count ≥ minsup}

**End;**

return ∪kLk

Trong thuật toán này, giai đoạn đầu đơn giản chỉ là việc đếm support cho các mục(item). Để xác định tập 1-mục phổ biến (L1), người ta chỉ giữ lại các mục (item) mà sup của nó lớn hơn hoặc bằng *minsup.*

Tiếp theo CSDL D sẽ được quét để tính support cho mỗi ứng viên trong Ck. Để việc đếm được nhanh, cần phải có một giải pháp hiệu quả để xác định các ứng viên trong Ck là có mặt trong một giao dịch T cho trước.

Vấn đề sinh tập candidate (tập ứng viên) của Apriori – Hàm Apriori\_gen: Hàm Apriori\_gen với đối số là Lk-1(tập các large(k-1)-itemset) sẽ cho lại kết quả là một superset, tập của tất cả các large k – itemset. Sơ đồ sau là thuật toán cho hàm này.

**Input:** tập mục phổ biến Lk-1 có kích thước k-1

**Output:** tập ứng cử viên Ck

**Method:**

**function** apriori-gen(Lk-1: tập mục phổ biến có kích thước k-1)

**Begin**

**For** (mỗi L1 ∈ Lk-1) **do**

**For** (mỗi L2 ∈ Lk-1) **do**

**begin**

**If** ((L1[1]=L2[1]) ∩ (L1[2]=L2[2]) ∩ ... ∩ (L1[k-2]=L2[k-2]) ∩ (L1[k-1]=L2[k-1])) **then**

c = L1 ⊕ L2; // kết nối L1 với L2 sinh ra ứng cử viên C

**If** has\_infrequent\_subset(c, Lk-1) **then** //có tập con không phổ biến của c trong Lk-1

remove (C) // bước tỉa (xoá ứng cử viên C)

**else** Ck = Ck ∪ {c}; kết tập c vào Ck

**end**;

Return Ck;

**End**;

Với nội dung trên, ta thấy hàm này có 2 bước:

Bước nối (join step).

Bước cắt tỉa (prune step).

Hàm Subset (trong bước tỉa) Các tập ứng cử viên Ck được lưu trữ trong một cây băm.

Nút lá của cây băm chứa danh sách một tập mục và đếm.

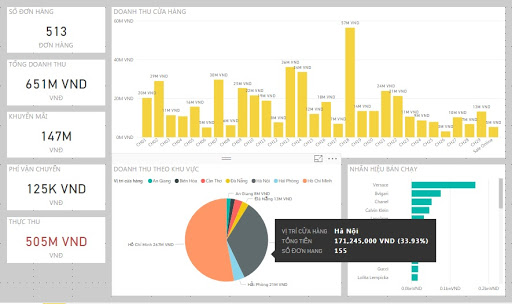
Các nút trong chứa ở trong bảng băm.

Hàm subset: tìm tất cả các ứng cử viên được chứa trong giao tác.

# **CHƯƠNG 3: GIỚI THIỆU VỀ CÁC CÔNG CỤ SỬ DỤNG**

## **Tổng quan về công cụ trực quan hóa Power BI**

**Power BI** là một giải pháp phân tích kinh doanh cho phép bạn trực quan hóa dữ liệu của mình và chia sẻ thông tin chi tiết về tổ chức của bạn hoặc nhúng chúng vào ứng dụng hoặc trang web của bạn, lập báo cáo, biểu diễn dữ liệu trực quan và xuất bản báo cáo dữ liệu lên môi trường web rất dễ dàng và thuận lợi.



Power BI là công cụ báo cáo và trực quan cá nhân của bạn, đồng thời cũng đóng vai trò là công cụ phân tích và quyết định cho các dự án nhóm, bộ phận hoặc toàn bộ tập đoàn.

A picture containing clock

Description automatically generated

* **Các môi trường làm việc chính của Power BI**
* **Power BI Desktop:** Là một ứng dụng dùng trên máy tính để bàn.
* **Power BI Service:** Dịch vụ SaaS hỗ trợ trực tuyến.
* **Power BI mobile:** Là phiên bản dùng trên điện thoại di động thông minh có hệ điều hành iOS, Android.

A screenshot of a computer

Description automatically generated

* **Công dụng của Power BI:**
* Giúp bạn đưa ra được những phân tích dữ liệu và ra quyết định nhanh chóng, thuận lợi và dễ dàng.
* Power BI còn tích hợp với các công cụ khác, bao gồm Microsoft Excel, do đó bạn có thể làm việc nhanh chóng và đơn giản hóa các công việc.
* Kết nối, mô hình hóa và sau đó khám phá dữ liệu của bạn với các báo cáo trực quan mà bạn có thể cộng tác, xuất bản và chia sẻ.
* **Lý do bạn nên sử dụng Power BI :**
* Tương tác trực quan, dễ dàng tích hợp với các phần mềm tin học khác.
* Truy cập không giới hạn vào dữ liệu cục bộ và dữ liệu đám mây.
* Tích hợp không giới hạn với ứng dụng của bên thứ ba.
* Chia sẻ an toàn các báo cáo và phân tích dữ liệu.
* Khả năng tìm kiếm thông minh.
* Bảo mật thông tin tốt.

## **Giới Thiệu SQL Server 2014**

* **SQL Server 2014:**

**Microsoft SQL Server** (MS SQL Server) là một hệ quản trị cơ sở dữ liệu (CSDL) quan hệ do hãng Microsoft phát triển sử dụng ngôn ngữ truy vấn Transact- SQL (T-SQL) để trao đổi dữ liệu giữa máy khách (Client) và máy chủ (Server) – Mô hình Client-Server.

Hiện nay MS SQL Server có thể thao tác với các CSDL có kích thước cực lớn, phục vụ nhiều máy khách cùng lúc và kết hợp ăn ý với các chương trình khác (như Microsoft Internet  Information Server – IIS, Microsoft Team Foundation Server – TFS, Microsoft Visual Studio - VS,…). MS SQL Server chỉ chạy trên hệ điều hành windows của Microsoft. Nhưng việc kết nối đến MS SQL Server có thể thực hiện từ các hệ điều hành khác nhau dựa vào thư viện các nhà phát triển cung cấp.

* **Các thành phần cơ bản trong SQL Server**



* **Database Engine**
* Cái lõi của SQL Server.
* Ðây là một engine có khả năng chứa data ở các quy mô khác nhau dưới dạng table và support tất cả các kiểu kết nối (data connection) thông dụng của Microsoft như ActiveX Data Objects (ADO), OLE DB, and Open Database Connectivity (ODBC).
* Ngoài ra nó còn có khả năng tự điều chỉnh (tune up) ví dụ như sử dụng thêm các tài nguyên (resource) của máy khi cần và trả lại tài nguyên cho hệ điều hành khi một user log off.
* **Replication**
* Cơ chế tạo bản sao (Replica).
* Giả sử bạn có một database dùng để chứa dữ liệu được các ứng dụng thường xuyên cập nhật. Khi bạn muốn có một cái database giống y hệt như thế trên một server khác để chạy báo cáo (report database). Vấn đề là report server của bạn cũng cần phải được cập nhật thường xuyên để đảm bảo tính chính xác của báo cáo. Bạn không thể dùng cơ chế back up and restore trong trường hợp này. Lúc đó cơ chế replication của SQL Server sẽ được sử dụng để đảm bảo dữ liệu ở 2 database được đồng bộ (synchronized).
* **Integration Services (DTS)**
* Integration Services là một tập hợp các công cụ đồ họa và các đối tượng lập trình cho việc di chuyển, sao chép và chuyển đổi dữ liệu.
* Nếu bạn làm việc trong một công ty lớn trong đó data được chứa trong nhiều nơi khác nhau và ở các dạng khác nhau cụ thể như chứa trong Oracle, DB2 (của IBM), SQL Server, Microsoft Access… Bạn chắc chắn sẽ có nhu cầu di chuyển data giữa các server này (migrate hay transfer) và không chỉ di chuyển bạn còn muốn định dạng (format) nó trước khi lưu vào database khác, khi đó bạn sẽ thấy DTS giúp bạn giải quyết công việc trên dễ dàng.
* **Analysis Services**
* Dữ liệu (Data) chứa trong database sẽ chẳng có ý nghĩa gì nhiều nếu như bạn không thể lấy được những thông tin (Information) bổ ích từ đó. Do đó Microsoft cung cấp cho bạn một công cụ rất mạnh giúp cho việc phân tích dữ liệu trở nên dễ dàng và hiệu quả bằng cách dùng khái niệm hình khối nhiều chiều (multi-dimension cubes) và kỹ thuật “khai phá dữ liệu” (data mining).
* **Notification Services**
* Dịch vụ thông báo Notification Services là nền tảng cho sự phát triển và triển khai các ứng dụng tạo và gửi thông báo. Notification Services có thể gửi thông báo theo địch thời đến hàng ngàn người đăng ký sử dụng nhiều loại thiết bị khác nhau.
* **Reporting Services**
* Reporting Services bao gồm các thành phần server và client cho việc tạo, quản lý và triển khai các báo cáo. Reporting Services cũng là nền tảng cho việc phát triển và xây dựng các ứng dụng báo cáo.
* **Full Text Search Service**
* Dịch vụ SQL Server Full Text Search là một dịch vụ đặc biệt cho đánh chỉ mục và truy vấn cho dữ liệu văn bản không cấu trúc được lưu trữ trong các CSDL SQL Server. Đánh chỉ mục với Full Text Search có thể được tạo trên bất kỳ cột dựa trên dữ liệu văn bản. Nó sẽ rất hiệu quả cho việc tìm các sử dụng toán tử LIKE trong SQL với trường hợp tìm văn bản.
* **Service Broker**
* Được sử dụng bên trong mỗi Instance, là môi trường lập trình cho việc các ứng dụng nhảy qua các Instance. Service Broker giao tiếp qua giao thức TCP/IP và cho phép các component khác nhau có thể được đồng bộ cùng nhau theo hướng trao đổi các message. Service Broker chạy như một phần của bộ máy cơ sở dữ liệu, cung cấp một nền tảng truyền message tin cậy và theo hàng đợi cho các ứng dụng SQL Server.

## **Giới thiệu công cụ Visual Studio 2013**

[**Microsoft Visual Studio**](https://www.visualstudio.com/) là một môi trường phát triển tích hợp từ Microsoft. Nó được sử dụng để phát triển chương trình máy tính cho Microsoft Windows, cũng như các trang web, các ứng dụng web và các dịch vụ web. Visual Studio sử dụng nền tảng phát triển phần mềm của [Microsoft](https://www.microsoft.com/vi-vn/) như Windows API, Windows Forms, Windows Presentation Foundation, Windows Store và Microsoft Silverlight. Nó có thể sản xuất cả hai ngôn ngữ máy và mã số quản lý.

**Visual Studio** bao gồm một trình soạn thảo mã hỗ trợ IntelliSense cũng như cải tiến mã nguồn. Trình gỡ lỗi tích hợp hoạt động cả về trình gỡ lỗi mức độ mã nguồn và gỡ lỗi mức độ máy. Công cụ tích hợp khác bao gồm một mẫu thiết kế các hình thức xây dựng giao diện ứng dụng, thiết kế web, thiết kế lớp và thiết kế giản đồ cơ sở dữ liệu. Nó chấp nhận các plug-in nâng cao các chức năng ở hầu hết các cấp bao gồm thêm hỗ trợ cho các hệ thống quản lý phiên bản (như Subversion) và bổ sung thêm bộ công cụ mới như biên tập và thiết kế trực quan cho các miền ngôn ngữ cụ thể hoặc bộ công cụ dành cho các khía cạnh khác trong quy trình phát triển phần mềm.

## **Microsoft SQL Server Analysis Services (SSAS)**

Analysis Services là công cụ dễ sử dụng, tích hợp và linh động giúp định nghĩa các chiều, xây dựng các khối dữ liệu đa chiều và cung cấp các chương trình ứng dụng truy xuất tới các khối này.

Microsoft SQL Server Analysis Services (SSAS) đáp ứng nhu cầu xử lý phân tích dữ liệu OLAP và các ứng dụng khai phá dữ liệu. Analysis Services hỗ trợ OLAP bằng cách cho phép bạn thiết kế, tạo ra, và quản lý các cấu trúc đa chiều có chứa dữ liệu được tổng hợp từ các nguồn dữ liệu khác, chẳng hạn như cơ sở dữ liệu quan hệ. Đối với các ứng dụng khai thác dữ liệu Analysis Services cho phép bạn thiết kế tạo ra và hình dung mô hình khai thác dữ liệu được xây dựng từ các nguồn dữ liệu khác bằng cách sử dụng một loạt các tiêu chuẩn các thuật toán khai thác dữ liệu.

SSAS cho phép các ứng dụng hoạt động theo chế độ Client/Server. Thành phần server được triển khai như một dịch vụ của Microsoft Windows bao gồm một số thành phần thực hiện các chức năng:

* **Phân tích câu lệnh nhận được từ client.**
* **Quản lý metadata.**
* **Quản lý giao dịch.**
* **Xử lý tính toán.**
* **Lưu trữ dữ liệu.**
* **Tạo các khối tích hợp.**
* **Lập lịch truy vấn.**
* **Quản lý tài nguyên.**

SSAS hỗ trợ nhiều dịch vụ trên cùng một máy chủ, mỗi dịch vụ là một đối tượng riêng biệt. Thành phần client truyền thông với server bằng giao thức XML for Analysis (XML/A) thông qua cơ chế truy vấn và trả lời tương tự như web service. Câu lệnh truy vấn sử dụng ngôn ngữ SQL, MDX (Multidimensional Expressions), DMX (Data Mining Extensions). Vài provider cũng được SSAS cung cấp để hỗ trợ các ngôn ngữ lập trình khác nhau.

## **Môi trường triển khai ứng dụng (BIDS)**

Business Intelligence Development Studio (BIDS) là môi trường phát triển ứng dụng SSAS. Môi trường này được tích hợp với Microsoft Analysis Service. Phần lớn thời gian làm việc với SQL Server Data Mining là làm việc với BIDS. Nó là công cụ cho phép tổ chức và khai thác kho dữ liệu (xử lý phân tích trực tuyến) cũng như xây dựng mô hình khai phá dữ liệu rất dễ sử dụng và hiệu quả của Microsoft.

* ***BIDS có hai chế độ hoạt động:* Offline và Immediate.**

Làm việc với chế độ Immediate, người dùng kết nối trực tiếp và liên tục đến Analysis Service server. Khi tạo mới hoặc thay đổi các đối tượng đã có, các hoạt động này đều được thay đổi trên server. Mặc dù chế độ làm việc này rất trực quan và dễ hiểu. Tuy nhiên, nó cũng có nhiều hạn chế như các thao tác sai không thể phục hồi, các tác động không mong muốn từ các phiên làm việc khác, tính nhất quán giữa các ứng dụng sử dụng chung dữ liệu. Ngoài ra chế độ này chỉ bó hẹp trong một cơ sở dữ liệu nhất định.

# **CHƯƠNG 4: PHÂN TÍCH PHẢN HỒI KHÁCH HÀNG VỀ VIỆC MUA XE TẠI CỬA HÀNG Ô TÔ HONDA ĐÀ NẴNG**

## **Thực trạng và yêu cầu**

## **Đề xuất quy trình**

### **Thu thập dữ liệu**

### **Thiết kế dữ liệu**

### **Cài đặt dữ liệu vào hệ thống**

### **Trực quan hóa dữ liệu**

### **Xây dựng mô hình khai phá dữ liệu trên BIDS**

### **Ứng dụng luật kết hợp phân tích độ hài lòng của khách hàng**

# **CHƯƠNG 5: KẾT QUẢ PHÂN TÍCH**

## **Kết quả đánh giá dữ liệu báo cáo BI**

## **Kết quả đánh giá mô hình luật kết hợp**

# **KẾT LUẬN**

# **ĐỀ XUẤT HƯỚNG PHÁT TRIỂN**

# **TÀI LIỆU THAM KHẢO**

# **PHỤ LỤC**