TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG
KHOA CÔNG NGHỆ - KỸ THUẬT
BỘ MÔN CÔNG NGHỆ THÔNG TIN
oOo

KHÓA LUẬN
TỐT NGHIỆP ĐẠI HỌC

Ngành: Công nghệ thông tin

Đề tài:

TÌM HIỂU
PHÂN TÍCH DỮ LIỆU - DATA MINING
VÀ CÀI ĐẶT BẰNG NGÔN NGỮ PYTHON

Sinh viên thực hiện : NGUYỄN GIA HUY
MSSV : 191106003
Giảng viên hướng dẫn : ThS. LÊ VĂN HẠNH

Tháng 5 -2023

TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG
KHOA KỸ THUẬT VÀ CÔNG NGHỆ
BỘ MÔN CÔNG NGHỆ THÔNG TIN
oOo

ĐỀ CƯƠNG VIẾT KHÓA LUẬN TỐT NGHIỆP

Họ tên và tên sinh viên: Nguyễn Gia Huy MSSV: 191106003

Lớp: TH19DH-TH1 Khóa: 2019

Hệ đào tạo: Chính qui

Ngành / Chuyên ngành: Công nghệ thông tin

Thời gian làm luận:từ 15/02/2022 đến 15/05/2022

Tên đồ án: Tìm hiểu về phân tích dữ liệu và cài đặt bằng ngôn ngữ Python

Đề cương viết đồ án:

1. Giới thiệu về phân tích dữ liệu
2. Khai phá dữ liệu
3. Python trong phân tích dữ liệu
4. Áp dụng thuật toán APRIORI để chẩn đoán dấu hiệu covid-19

.....

.....

.....

.....

.....

.....

Giảng viên hướng dẫn **Sinh viên làm đồ án**

(ký và ghi rõ họ tên) (ký và ghi rõ họ tên)

TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG
KHOA KỸ THUẬT VÀ CÔNG NGHỆ
BỘ MÔN CÔNG NGHỆ THÔNG TIN
oOo

PHIẾU GIAO NHIỆM VỤ
ĐỀ ÁN TỐT NGHIỆP

Lớp: TH19DH-TH1 Khóa: 2019 Hệ đào tạo: Chính qui

Làm đồ án tại: Đại học quốc tế Hồng Bàng

Thời gian làm luận văn: Từ 15/02/2022 đến 15/05/2022

Sinh viên: Nguyễn Gia Huy MSSV: 191106003

Số điện thoại liên lạc khi cần thiết: 0818115119

Giảng viên hướng dẫn: ThS. Lê Văn Hạnh

Số điện thoại: 0913158020

Gặp thầy tại: Đại học quốc tề Hồng Bàng

Email của thầy: hanhlv@hiu.vn

Nội dung và yêu cầu khi làm đồ án:

1. Hàng tuần báo cáo công việc đó làm và kế hoạch làm trong tuần tới .....
2. Thời gian và địa điểm làm việc .....
3. Đề tài tốt nghiệp và các yêu cầu chuyên môn: .....

.....

.....

.....

.....

.....

.....

Thành phố Hồ Chí Minh,ngày tháng năm 2022

Giảng viên hướng dẫn

(ký tên và ghi rõ họ tên)

BẢN NHẬN XÉT ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

Tên đề tài: Tìm hiểu phân tích dữ liệu và cài đặt bằng ngôn ngữ Python

Sinh viên thực hiện: Nguyễn Gia Huy lớp TH19DH-TH1

Khoá DHCQK2019

Ngành/chuyên ngành: Công nghệ thông tin

Nội dung nhận xét: .....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

LỜI CAM ĐOAN

Em – Nguyễn Gia Huy cam đoan khoá luận tốt nghiệp này là công trình nghiên cứu của bản thân em dưới sự hướng dẫn của Thạc sỹ Lê Văn Hạnh. Các kết quả công bố trong khoá luận tốt nghiệp là trung thực và không sao chép từ bất kỳ công trình nào khác.

Sinh viên thực hiện đề tài

Nguyễn Gia Huy

LỜI CẢM ƠN

Trong thời gian thực hiện khoá luận,em nhận được những sự giúp đỡ của quý thầy cô và bạn bè nên khoá luận đã hoàn thành. Em xin chân thành gửi lời cảm ơn đến với:

Thầy Lê Văn Hạnh, giảng viên trường Đại Học Quốc Tế Hồng Bàng đã trực tiếp hướng dẫn và nhiệt tình giúp đỡ tạo điều kiện để em có thể hoàn thành tốt luận án tốt nghiệp và đúng thời hạn.

Em cũng xin chân thành cảm ơn đến các quý thầy trong khoa Kỹ Thuật – Công Nghệ của trường Đại Học Quốc Tế Hồng Bàng đã tận tình dạy dỗ,chỉ bảo, cung cấp cho em những kiến thức nền, chuyên môm làm cơ sở để hoàn thành đề tài này.

Xin gửi lời cảm ơn đến với những người bạn trong khoa Kỹ Thuật – Công Nghệ đã giúp đỡ em để có thể hoàn thành tốt luận án này.

Xin cảm ơn !

Sinh viên thực hiện

Nguyễn Gia Huy

MỤC LỤC

Trang

DANH MỤC BẢNG BIỂU

Trang

DANH SÁCH HÌNH

Trang

DANH MỤC TỪ VIẾT TẮT

Ký hiệu chữ viết tắt	Chữ viết đầy đủ
MCAR	Missing completely at random
MAR	Missing at random
MNAR	Missing not at random
ELT	Extract-Load-Transform
ETL	Extract-Transform-Load

DANH MỤC TỪ TIẾNG ANH CHUYÊN NGÀNH

Tiếng Anh	Tiếng Việt
Missing completely at random	Dữ liệu bị mất hoàn toàn ngẫu nhiên
Missing at random	Dữ liệu bị mất ngẫu nhiên
Missing not at random	Dữ liệu mất không ngẫu nhiên
Data lake	Hồ dữ liệu
Data warehouse	Kho dữ liệu
Data cleaning	Làm sạch dữ liệu
Data transformation	Chuyển hoá dữ liệu
Data integration	Tích hợp dữ liệu
Data normalization	Chuẩn hoá dữ liệu
List-wise deletion	Xoá theo danh sách
Regression imputation	Quy nạp hồi quy
Simple imputation	Quy nạp đơn giản
Mean	Trung bình
Median	Trung vị
Mode	Yếu vị

MỞ ĐẦU

Ngày nay với sự phát triển vượt bậc của công nghệ, một lượng lớn dữ liệu đã thu tập và sử dụng bởi các công ty lớn nhằm tạo ra nhiều chiến lược kinh doanh hoặc đưa ra những dự báo tương lai. Dữ liệu được tạo ra một lượng lớn thông tin từ ghi chép thông tin bệnh nhân từ các bệnh viện, ghi chú khách hàng từ các công ty truyền thông, lịch sử mua hàng và thông tin hàng hoá từ các công ty mua sắm…Các dữ liệu này sau đó sẽ được phân tích và những thông tin hữu ích sẽ được trích lọc. Phương pháp khai phá dữ liệu là một trong những phương pháp được ứng dụng nhiều nhất. Kết hợp với việc áp dụng các thuật toán máy học, các dữ liệu hữu ích sẽ được phân tích và sau đó được sử dụng để hỗ trợ việc đưa ra quyết định.

Cũng vì lý do trên, con người bắt đầu phát triển các loại công cụ và ngôn ngữ để hỗ trợ cho việc phân tích, chọn lọc dữ liệu trở nên hiệu quả hơn. Đặc biệt nhất phải kể đến ngôn ngữ Python, ngôn ngữ này cung cấp rất nhiều các thư viện mạnh cho việc phân tích dữ liệu như: Pandas, Numpy, Matplotlib, … giúp cho việc phân tích dữ liệu trở nên dễ dàng hơn.

Trong luận án này, em sẽ tập trung vào một trong những kỹ thuật của data mining là tìm tập phổ biến và luật kết hợp, với thuật toán chủ đạo là Apriori. Bao gồm giải thích về tập phổ biến, luật kết hợp và cách thức vận hành của thuật toán Apriori cũng như so sánh với các thuật toán khác trong việc tìm ra tập phổ biến và luật kết hợp.

Coronavirus là một tập thể các chủng virus gây ra nhiều các bệnh truyền nhiễm khác nhau. Một trong số những loại phổ biến nhất có thể kể đến: Hội chứng hô hấp Trung Đông (MERS) và hội chứng hô hấp cấp tính (SARS) [1]. Mặc dù đã hơn 2 năm kể từ khi dịch covid 19 hoành hành, tuy nhiên các biến chủng mới của chúng vẫn đang có nguy cơ gây ra các cuộc bùng phát mới trên toàn thế giới.

Trong năm 2019-2022, dịch Covid-19 đã gây ra những tác động lớn đến thế giới, gây ra cái chết cho hàng triệu người trên thế giới và tạo gánh nặng khủng khiếp cho các cơ sở y tế vì số lượng bệnh nhân mắc bệnh là quá lớn. Việc dự đoán và ngăn chặn dịch bệnh được đặt lên hàng đầu và để làm được thế việc chẩn đoán dịch bệnh là hết sức quan trọng. Chính vì vậy mà em đã lựa chọn kỹ thuật tìm tập phổ biến và luật kết hợp và thuật toán Apriori để chẩn đoán các triệu chứng của bệnh và đưa ra các chẩn đoán sơ bộ nhằm có thể giúp làm giảm gánh nặng lên các cơ sở y tế.

Mục tiêu của khoá luận :

- Tìm hiểu về khái niệm của phân tích dữ liệu, quy trình phân tích và các kỹ thuật trong phân tích dữ liệu.
- Tìm hiểu về khai phá dữ liệu, quy trình trong khai phá dữ liệu, khái niệm về tập thường xuyên và luật kết hợp và thuật toán Apriori.
- Tìm hiểu về Python và lý do sử dụng Python trong phân tích dữ liệu.
- Áp dụng thuật toán Apriori để chẩn đoán bệnh covid 19 ở các bệnh nhân đang được theo dõi.

TÓM TẮT KHÓA LUẬN

Chương 1.Tìm hiểu chung về phân tích dữ liệu, quy trình phân tích dữ liệu và giới thiệu về các phương pháp phân tích dữ liệu.

Chương 2. Tìm hiểu khái phá dữ liệu và các kỹ thuật khai phá dữ liệu.

Chương 3. Tìm hiểu về vai trò của Python trong phân tích dữ liệu, trong đó trình bày về lý do sử dụng Python trong phân tích dữ liệu, các thư viện phổ biến.

Chương 4. Áp dụng thuật toán APRIORI trong khai phá dữ liệu để chẩn đoán covid-19.

Chương 5. Kết luận trong đó trình bày các nội và kết quả thực hiện của KHÓA LUẬN, những hạn chế và hướng nghiên cứu tiếp theo.

CHƯƠNG 1. GIỚI THIỆU VỀ PHÂN TÍCH DỮ LIỆU

1.1 Giới thiệu

Theo công ty Amazon, phân tích dữ liệu là quá trình thu thập, lưu trữ, xử lý, làm sạch, phân tích dữ liệu thô để trở thành các thông tin hữu ích, kết luận và hỗ trợ việc quyết định cho các tổ chức, doanh nghiệp. Phân tích dữ liệu có nhiều khía cạnh và hướng tiếp cận khác nhau bao gồm nhiều phương thức phân tích khác nhau trong các lĩnh vực như kinh tế, khoa học và xã hội học [2].

1.2 Vai trò

Phân tích dữ liệu đóng vai trò quan trọng vì nó có thể định hình các quy trình kinh doanh và cải thiện khả năng ra quyết định của doanh nghiệp. Lý do mà việc phân tích dữ liệu có thể làm được như vậy là vì nó có thể giúp cho doanh nghiệp nhìn rõ hơn và hiểu sâu hơn về quá trình và dịch vụ của họ bằng cách thu thập thông tin chuyên sâu chi tiết về trải nghiệm và các vấn đề của khách hàng để rồi phân tích, xây dựng và tạo ra những trải nghiệm khách hàng đã được cá nhân hóa, xây dựng các sản phẩm kỷ thuật số liên quan, tối ưu hoạt động và năng xuất của nhân viên [2].

1.3 Dữ liệu lớn

Việc lượng dữ liệu được sinh ra ngày nay ngày càng lớn lên đến 2.5 tỷ gigabytes mỗi ngày là quá lớn để để có thể xử lý 1 bằng các phương pháp thông thường như xử lý 1 dạng dữ liệu thường mà thay vào đó chúng sẽ được xem là dữ liệu lớn (big data). Dữ liệu lớn (big data) là cách gọi các tệp dữ liệu lớn đa dạng cấu trúc (structured),bán cấu trúc (semi structured) và không cấu trúc (unstructured) được tạo ra liên tục với tốc độ cao và khối lượng lớn. Big data thường được đo bằng terabyte hoặc petabyte. 1 terabyte tương đương 1.000.000 gigabyte [2].

Việc phân tích dữ liệu lớn là quá trình tìm các mẫu, xu hướng và các mỗi quan hệ trong những tập dữ liệu khổng lồ.Việc này đòi hỏi các công cụ, công cụ cụ thể, năng lực điện toán và kho lưu trữ dữ liệu hỗ trợ theo quy mô của dữ liệu.

1.4 Quy trình phân tích dữ liệu

Quy trình phân tích dữ liệu thô sang các thông tin có lợi cho doanh nghiệp, tổ chức,… bao gồm các bước sau [2]:

- Xác định mục tiêu
- Thu thập dữ liệu
- Tiền xử lý dữ liệu
- Xử lý dữ liệu
- Phân tích dữ liệu

- Diễn giải dữ liệu và đưa ra kết luận

1.4.1 Xác định mục tiêu

Đây là bước mà các nhà phân tích dữ liệu phải phát thảo một số mục tiêu rõ ràng. Tìm ra các câu hỏi mà doanh nghiệp, tổ chức, cá nhân,… cần trả lời bằng dữ liệu. VD: Liệu đặt kệ khăn giấy gần quầy thuốc tây có tăng doanh thu của khăn giấy lên không ?

Đây cũng là bước mà các nhà phân tích dữ liệu xác định rõ loại dữ liệu, định dạng của dữ liệu mà họ muốn phân tích.

1.4.2 Thu thập dữ liệu

Bước này bao gồm việc xác định nguồn dữ liệu và thu thập dữ liệu từ nguồn này. Việc thu thập cần tuân theo một trong hai quá trình là ELT hoặc ETL [2].

- ELT – (Extract, Load, Transform) Trích xuất, tải, chuyển đổi:

Trong ELT,trước tiên dữ liệu sau khi được trích xuất từ nguồn sẽ được tải vào kho lưu trữ và sau đó được chuyển đổi thành định dạng yêu cầu

- ETL – (Extract, Transform, Load) Trích xuất, chuyển đổi, tải:

Trong ETL, trước tiên dữ liệu sau khi được trích xuất từ nguồn sẽ được chuyển đổi thành định dạng yêu cầu sau đó được tải và kho lưu trữ.

Sau khi thu thập, dữ liệu cần phải được lưu trữ ở dịch vụ lưu trữ đặc biệt do dữ liệu thu về thường có dung lượng rất lớn.Trong số các dịch vụ lưu trữ dữ liệu: Data warehouse và datalake là 2 dịch vụ phổ biến và hiệu quả nhất.

1.4.2.1 Các phương pháp thu thập dữ liệu

Các phương pháp để thu thập dữ liệu bao gồm [3]:

- Phương pháp quan sát (observation)
- Phương pháp phỏng vấn bằng thư (mail interview)
- Phương pháp phỏng vấn bằng điện thoại (telephone interview)
- Phương pháp phỏng vấn cá nhân trực tiếp (personal interview)
- Phương pháp điều tra nhóm cố định (panels)
- Phương pháp điều tra nhóm chuyên đề (focus groups)

1.4.2.2 Lưu trữ dữ liệu

Dựa vào các loại và độ phức tạp khác nhau mà dữ liệu được lưu trữ ở kho dữ liệu (data warehouse) hoặc hồ dữ liệu (data lake). Đây là các công nghệ giúp cho doanh nghiệp quản lý và tận dụng tiềm năng rộng lớn của Big data.

1.4.2.2.1 Data warehouse

Là cơ sở dữ liệu được tối ưu hóa cho phân tích dữ liệu quan hệ đến từ hệ thống giao dịch và ứng dụng kinh doanh. Cấu trúc dữ liệu và lược đồ được xác định trước để tối ưu hóa việc tìm kiếm và báo cáo nhanh. Dữ liệu được đưa vào kho dữ liệu đã được dọn dẹp và biến đổi để đóng vai trò là "nguồn thông tin " đáng tin cậy mà người dùng có thể tin tưởng. Đa số dữ liệu được lưu trữ trong kho dữ liệu đều là các dữ liệu có cấu trúc hoặc bán cấu trúc.

1.4.2.2.2 Data lake

Data lake thì khác vì có thể lữu trữ cả dữ liệu có cấu trúc và phi cấu trúc mà không cần xử lý thêm. Cấu trúc dữ liệu hoặc lược đồ không cần được xác định khi thu thập, tức là có thể lưu mọi dữ liệu mà không cần cẩn trọng thiết kế, điều này hữu ích khi chưa xác định mục đích sử dụng dữ liệu trong tương lai. Các ví dụ về dữ liệu bao gồm: nội dung truyền thông xã hội, dữ liệu Iot,…

1.4.2.2.3 Sự khác biệt của data lake và data warehouse

So sánh giữa data lake và data warehouse [4]:

	Data lake	Data warehouse
Loại dữ liệu	Tất cả các loại dữ liệu từ hệ thống đều được lưu trữ kể cả dữ liệu bị từ chối lưu trữ trong data ware house. Các loại dữ liệu này chưa được qua xử lý. Có thể lưu cả dữ liệu hiện tại không sử dụng.	Chỉ lưu trữ các dữ liệu trích xuất từ hệ thống giao dịch và các số liệu định lượng để hỗ trợ quá trình phân tích hiệu suất và tình trạng kinh doanh. Dữ liệu trong data warehouse là các dữ liệu có cấu trúc rõ ràng.
Hình thức schema	Data lake sử dụng phương pháp "Schema on Read".Nghĩa là khi có nhu cầu xử dụng dữ liệu để giải quyết các vấn đề kinh doanh thì chỉ có các dữ liệu liên quan đến được chọn và phân tích. Dữ liệu được giữ ở trạng thái nguyên bản. Tiết kiệm chi phí và thời gian	Data warehouse áp dụng phương pháp "Schema on Write". Nghĩa là mô hình này được thiết kế với mục đích là cung cấp báo cáo thông qua quá trình dài phân tích dữ liệu, thấu hiểu quy trình nghiệp vụ, và hình thành một hệ thống xác định để phân tích dữ liệu. Dữ liệu được biến đổi theo cấu trúc.

		Hình thành hệ thống phân tích chi tiết.
Tính linh hoạt	Linh hoạt trong việc tái cấu trúc do dữ liệu được lưu trữ trong data lake đều là những dữ liệu phi cấu trúc.	Khó khăn và tốn kém trong việc tái cấu trúc do dữ liệu được lưu trữ trong data warehouse thường là những dữ liệu có cấu trúc nên dẫn đến việc đòi hỏi thêm các quy trình xử lý phức tạp và tốn kém
Người dùng	Phù hợp với những người có nhu cầu phân tích chuyên sâu như những data scientists. Đa dạng các loại dữ liệu để các nhà phân tích có thể kết hợp và đưa ra các câu hỏi mới để giải đáp.	Phù hợp với doanh nghiệp và người dùng. Dễ dàng sử dụng và chủ yếu được dùng để trả lời các câu truy vấn do dữ liệu được quản lý với cấu trúc chặt chẽ.

Bảng 1-1 Bảng so sánh giữa data lake và data warehouse

1.4.3 Tiền xử lý dữ liệu

Theo tác giả của IBM developer, tiền xử lý dữ liệu (pre data processing) bao gồm các bước làm sạch dữ liệu (data cleaning), chuyển đổi dữ liệu (data transformation), rút gọn dữ liệu (data reduction) [5].

1.4.3.1 Data cleaning

Làm sạch dữ liệu là quá trình loại bỏ hoặc chỉnh sửa các tệp dữ liệu không phù hợp, sai định dạng hoặc bị thiếu thông tin trong tập dữ liệu. Khi kết hợp các tệp dữ liệu lại với nhau có khả năng dữ liệu bị trùng lặp hoặc sai [6].

Việc đảm bảo dữ liệu được chính xác và toàn vẹn trước khi được đưa vào xử lý là vô cùng quan trọng nhằm đảm bảo kết quả và thuật toán được chính xác.

Làm sạch dữ liệu bao gồm các bước:

- Loại bỏ các dữ liệu bị trùng lặp hoặc không phù hợp

  Các dữ liệu bị trùng lặp thường do trong quá trình thu thập dữ liệu và kết hợp các dữ liệu. Khi kết hợp dữ liệu từ nhiều nguồn, đối tượng khác nhau sẽ dễ dẫn đến trùng lặp dữ liệu. Để xử lý dữ liệu bị trùng lặp trước hết cần quan sát

và liệt kê các dữ liệu bị trùng lặp sau đó tiến hành xoá đi các hàng dữ liệu trùng lặp [6].

Các dữ liệu không phù hợp là các dữ liệu không liên quan hoặc không phù hợp với vấn đề đang cần phân tích [6]. VD: Nếu muốn phân tích các hoạt động, sở thích của genz nhưng các tập dữ liệu lại có cả thông tin của những thế hệ lớn hơn thì có thể xoá đi những tập tin của các thế hệ đấy. Điều này góp phần giảm thiểu sự phân tâm và tăng cường hiệu suất trong việc phân tích .

- Sữa lỗi cấu trúc

Lỗi cấu trúc thường xảy ra khi nhà phân tích di chuyển tập dữ liệu hoặc khi đo lường thì dữ liệu thường bị thay đổi như quy ước tên lạ, lỗi chính tả, sai kiểu ký tự, viết hoa sai,… Những điều không nhất quán có thể gây ra sai sót trong phân loại danh mục.VD: Tập dữ liệu đếu xuất hiện đồng thời N/A và NAN,nhưng đáng lý thì chúng phải đều ở cùng 1 phân loại.

- Lọc dữ liệu ngoại lai

Đôi khi trong lúc quan sát tập dữ liệu sẽ xuất hiện các tập dữ liệu không phù hợp với dữ liệu mà nhà phân tích đang phân tích. Nếu dữ liệu khác biệt với các dữ liệu còn lại do sai sót trong lúc nhập liệu thì nhà phân tích hoàn toàn có thể xoá giá trị ngoại lai này đi để tăng chất lượng dữ liệu mà nhà phân tích đang phân tích. Ngược lại nếu dữ liệu đó tồn tại để chứng minh cho lý thuyết mà nhà phân tích đang phân tích thì nhà phân tích nên giữ lại [6].

VD: Trong việc phân tích lương của công nhân viên trong công ty mà lại xuất hiện lương của tổng giảm đốc cao hơn nhiều lần so với các công nhân khác sẽ có thể làm kết quả báo cáo mức lương bình quân của công nhân bị ảnh hưởng nặng. Vì thế các nhà phân tích có thể xoá đi lương của tổng giám đốc đi để đảm bảo kết quả thu về là chính xác. Tuy nhiên nếu trường hợp là mức lương được trải đều từ công nhân viên, trưởng phòng, thư ký, tổng giám đốc,… thì lúc này việc giữ lại lương của tổng giám đốc là quan trọng để tính được mức lương trung bình của cả công ty.

- Xử lý dữ liệu thiếu

Việc dữ liệu bị thiếu là việc xảy ra thường xuyên trong các tập dữ liệu nên việc xử lý các tập dữ liệu có dữ liệu bị thiếu là việc quan trọng trong quá trình phân tích sau này [6].

Việc đầu tiên cần phải xác định là loại dữ liệu bị thiếu là gì ? Có 3 loại dữ liệu bị thiếu. Bao gồm [6]:

- MCAR – Missing completely at random. Trường hợp này sự mất mát là ngẫu nhiên và các dữ liệu bị mất không có mối liên hệ nào tới các dữ liệu nào trong tập dữ liệu.
- MAR – Missing at random. Đối với trường hợp này dữ liệu bị mất một cách ngẫu nhiên. Tuy nhiên, có sự liên hệ giữa các dữ liệu bị mất và các dữ liệu quan sát.
- MNAR – Missing not at random. Dữ liệu bị mất không phải do ngẫu nhiên mà do mối liên hệ giữa giá trị bị mất và giá trị không bị mất trong cùng 1 biến.

Sau khi xác định được các loại dữ liệu bị mất các nhà phân tích tiến hành xử lý các dữ liệu bị mất này nhằm đảm bảo chất lượng dữ liệu trước khi được xử lý.Một số phương pháp xử lý dữ liệu:

Phương pháp	Mô tả	Ưu điểm	Nhược điểm
List-Wise deletion	- Là phương pháp xoá dữ liệu nếu dữ liệu bị thiếu nhiều và những dữ liệu ấy không quan trọng.	- Phương pháp trực quan và dễ dàng sử dụng. - Hiệu quả nếu dữ liệu bị xoá không quan trọng.	- Có thể dẫn đến thất thoát dữ liệu và sai lệch kết quả phân tích. - Không phù hợp nếu dữ liệu bị mất là dạng MNAR
Regression imputation	- Là phương pháp tạo một model để dự đoán giá trị quan sát của 1 biến dựa trên giá trị của một biến khác. Sau đó nhà phân tích dữ liệu có thể dùng dữ liệu model này để fill vào các giá trị bị thiếu.	- Phương pháp trực quan và dễ dàng sử dụng. - Bằng cách sử dụng dữ liệu có sẵn nên không cần phải có thêm dữ liệu bên ngoài.	- Các dữ liệu được bổ sung từ phương pháp này thường không có khái niệm về lỗi và dữ liệu thì phù hợp tới mức không có phương sai thừa. - Làm cho mối quan hệ của các biến bị xác định một cách quá mức và độ chính xác cho kết quả dự đoán lớn hơn cả kết quả đảm bảo.

			- Phương pháp dự đoán được dữ liệu bị thiếu tuy nhiên lại không cung cấp dược sự chắc chắn về dữ liệu mà nó dự đoán.
Simple imputation	- Là phương pháp thay thế các giá trị bị mất bằng giá trị của một số nguyên tắc định danh số học như (mean, median, mode).	- Phương pháp đơn dãng và dễ dàng sử dụng. - Bằng cách sử dụng dữ liệu có sẵn nên không cần phải có thêm dữ liệu bên ngoài. - Mean/median/mode cung cấp cho nhà phân tích ước lượng về giá trị của những dữ liệu bị mất.	- Mean và median chỉ được sử dụng cho các dữ liệu số, đối với dữ liệu kiểu categorical hoặc chữ thì phải dùng phương pháp mode. - Kết quả của phương pháp mean chịu ảnh hưởng nặng từ giá trị ngoại lai - Phương pháp median sẽ ám chỉ dữ liệu thuộc vào dạng MCAR. Điều này là hoàn toàn không chính xác.

Bảng 1-2 Các phương pháp xử lý dữ liệu bị mất

1.4.3.2 Data transformation

Sau khi dữ liệu được làm sạch thì sẽ được đưa qua quá trình chuyển đổi. Đây là bước mà nhà phân tích chuyển những dữ liệu thô thành dữ liệu theo chuẩn để chuẩn bị cho bước phân tích và lập mô hình/sơ đồ. Đây cũng là bước quan trọng trong quá trình phân tích đặc biệt đối với phương pháp data mining nhằm để trích xuất các kiến thức và thông tin hữu ích [7].

Data transformation bao gồm các bước. Bao gồm:

1.4.3.2.1 Data integration

Đây là quá trình kết hợp dữ liệu từ các nguồn khác nhau như từ database, bảng tính, hoá đơn, báo cáo,… về cùng 1 chuẩn.

Các kỹ thuật được dùng trong Data integration bao gồm data warehousing, ETL prosses và data federation.

Data integration được định nghĩa làm 3 phần <G, S, M>. Trong đó [7]:

- G là lược đồ toàn cầu
- S là nguồn khác biệt của lược đồ
- M là cầu nối của các truy vấn từ nguồn và lược đồ toàn cầu

1.4.3.2.2 Data normalization

Đây là quá trình tái cấu trúc các dữ liệu theo tỉ lệ quy định. Đây cũng là quá trình đảm bảo dữ liệu sau khi được chuẩn hoá sẽ là dữ liệu sạch. Ý nghĩa thật sự của data normalization là [7]:

- Chuẩn hoá dữ liệu là quá trình tái cấu trúc dữ liệu sao cho dữ liệu tương tự nhau khắp các vùng và bản ghi dữ liệu.
- Quá trình này cũng tăng cường sự liên kết cho các kiểu dữ liệu đầu vào, tăng cường hiệu quả làm sạch dữ liệu, tạo nên các dữ liệu tiềm năng, phân chia các đoạn dữ liệu và nâng cao chất lượng dữ liệu.

Các phương pháp chuẩn hoá dữ liệu [7]:

- Min-max normalization: Trong kỹ thuật này, phương pháp biến đổi tuyến tính được thực hiện trên dữ liệu gốc. Các giá trị tối đa và tối thiểu được trích lọc sau đó các dữ liệu được thay thế bằng công thức
- Normalization by decimal scaling: Là phương pháp chuẩn hoá dữ liệu bằng cách di chuyển điểm thập phân của giá trị. Để chuẩn hoá dữ liệu bằng kỹ thuật này nhà phân tích chia mỗi giá trị cho giá trị tuyệt đối tối đa của dữ liệu đó bằng công thức sau:
- Z-score normalization: Trong kỹ thuật này, giá trị được chuẩn hoá dựa trên giá trị trung bình và độ lệch chuẩn của dữ liệu A. Công thức của kỹ thuật:

Lợi ích của chuẩn hoá dữ liệu trong quá trình phân tích dữ liệu [7]:

- Tăng cường hiệu quả của các thuật toán máy học: Chuẩn hoá dữ liệu có thể tăng cường hiệu quả của các thuật toán máy học bằng cách tỉ lệ hoá các dữ liệu đầu vào theo tỉ lệ thông thường.
- Xử lý các dữ liệu ngoại lai: Làm giảm sức ảnh hưởng của dữ liệu ngoại lai bằng cách tỷ lệ các dữ liệu đầu vào theo chuẩn thông thường làm cho dữ liệu ngoại lai có sức ảnh hưởng kém hơn.
- Tăng cường khả năng diễn giải của kết quả: Chuẩn hoá làm cho quá trình diễn giải dữ liệu trên mô hình máy học dễ dàng hơn do các dữ liệu đầu vào giờ đã theo tỷ lệ thông thường.

- Tăng cường khả năng khái quát: Chuẩn hoá dữ liệu giúp tăng khả năng khái quát của mô hình bằng cách giảm đi tác động của giá trị ngoại lai và làm giảm sự nhạy cảm với tỉ lệ của dữ liệu đầu vào.

Điểm yếu của chuẩn hoá dữ liệu [7]:

- Mất dữ liệu: Trường hợp này xảy ra khi nhà phân tích thay đổi tỉ lệ của dữ liệu đầu vào theo tỉ lệ thông thường nhưng tỉ lệ ban đầu có các đặc điểm quan trọng
- Ảnh hưởng của dữ liệu ngoại lai: Do dữ liệu ngoại lai cũng sẽ thay đổi theo tỉ lệ mà nhà phân tích chọn như các dữ liệu còn lại vì thế khó phát hiện được dữ liệu ngoại lai.
- Ảnh hưởng đến việc diễn giải mô hình: Việc diễn giải các mô hình máy học sẽ bị ảnh hưởng do các dữ liệu đầu vào sẽ được chuẩn hoá theo tỉ lệ thông thường. Điều này dẫn đến việc chúng không được căn chỉnh đúng với tỉ lệ gốc của dữ liệu.
- Chi phí phát sinh: Với việc thêm bước chuẩn hoá tỉ lệ của dữ liệu đầu vào sẽ dẫn đến việc tăng thêm thời gian xử lý để chuyển từ tỉ lệ gốc sang tỉ lệ thông thường.

1.4.3.3 Data reduction

Data reduction là kỷ thuật làm giảm kích thước của dataset trong khi vẫn giữ lại các thông tin quan trọng. Kỹ thuật này có lợi trong trường hợp dataset quá lớn để có thể được xử lý 1 cách hiệu quả hoặc chứa quá nhiều thông tin thừa hoặc không liên quan đến vấn đề mà những nhà phân tích dữ liệu đang phân tích [8].

Một số phương pháp để giảm bớt dữ liệu:

1.4.3.3.1 Data Sampling

Kỹ thuật này bao gồm việc chọn ra 1 tập con từ tập dữ liệu lớn để thực thit hay vì cả 1 tập dữ liệu. Điều này có ích trong việc giảm thiểu kích thước của tập dữ liệu mà vẫn giữ nguyên được xu hướng và mẫu của dữ liệu [8].

Trong data sampling có 2 loại [8]:

- Random sampling: Trong phương pháp lấy mẫu ngẫu nhiên, mỗi điểm dữ liệu đều có khả năng được chọn như nhau. Kỹ thuật này được sử dụng trong trường hợp dữ liệu được phân bố không thống nhất và các dữ liệu được chọn thường là các dữ liệu đại diện cho toàn bộ tập.
- Statified Sampling: Phương pháp này được sử dụng khi dữ liệu được phân bố không đồng đều giữa các loại hoặc lớp. Dữ liệu được chia thành các nhóm nhỏ dựa trên những loại này và các mẫu dữ liệu sẽ được lấy từ các nhóm này.

Điểm mạnh của phương pháp lấy mẫu dữ liệu [8]:

- Giảm thời gian đào tạo và tính toán: Lấy mẫu dữ liệu làm giảm kích thước của tập dữ liệu cần được đào tạo, giúp giảm thời gian và tài nguyên trong việc tính toán cần thiết để đào tạo 1 mô hình phân tích.
- Giúp cân bằng sự phân bố dữ liệu: Lấy mẫu dữ liệu giúp chỉ ra các lớp bị mất cân bằng, vì lớp này thịnh hành hơn các lớp còn lại trong 1 tập dữ liệu. Bằng cách tăng cường việc lấy mẫu các lớp nhỏ và giảm thiểu việc lấy mẫu các lớp lớn, phương pháp lấy mẫu dữ liệu có thể giúp cân bằng sự phân bố của các lớp và giúp làm tăng hiệu suất của mô hình.
- Tăng cường hiệu suất của mô hình: Phương pháp lấy mẫu dữ liệu có thể làm giảm việc thuần thục quá dữ liệu được huấn luyện và hoạt động kém so với dữ liệu mới. Bằng cách giảm kích thước của dữ liệu của các tập dữ liệu huấn luyện hoặc cân bằng việc phân bổ của các lớp, phương pháp lấy mẫu dữ liệu có thể tăng cường hiệu xuất của mô hình và giảm tình trạng thuần thục quá của mô hình.

Điểm yếu của phương pháp lấy mẫu dữ liệu [8]:

- Tạo ra thiên lệch: Lấy mẫu dữ liệu có thể dẫn đến thiên lệch nếu dữ liệu được chọn không đại diện cho các tập dữ liệu chưa xác dịnh. Ví dụ cho việc này là việc nếu xuất hiện thiên lệch của dữ liệu mẫu thì kết quả cho các phân tích đối với các dữ liệu mới sẽ không khái quát.
- Có thể dẫn đến mất dữ liệu: Trong trường hợp dữ liệu mẫu là quá nhỏ hoặc không đại diện cho các tập dữ liệu chưa xác minh thì nó có thể dẫn đến mất mác dữ liệu. Việc này dẫn đến mô hình bị thu hẹp và xử lý kém khi phải xử lý các dạng dữ liệu mới.
- Có thể gây ảnh hưởng đến kết quả phân tích: Do tính hiệu quả của phương pháp lẫy mẫu dựa vào kỹ thuật lấy mẫu mà nhà phân tích dữ liệu chọn, tập dữ liệu mà họ phân tích và vấn đề mà doanh nghiệp và tổ chức cần giải quyết. Việc chọn đúng kỹ thuật để thực thi sẽ đảm bảo được tính hiệu quả và chính xác của kết quả phân tích.

1.4.3.3.2 Dimensionality reduction

Kỹ thuật này bao gồm việc giảm các đặc tính của dữ liệu trong tập dữ liệu, bằng cách xoá các dữ liệu không phù hợp hoặc kết hợp các nhiều đặc tính vào thành 1 đặc tính [8].

Phương pháp dùng trong kỹ thuật giảm chiều dữ liệu [8]:

- Feature Selection: Lựa chọn các đặc tính tốt nhất để giữ lại và loại bỏ đi các đặc tính không cần thiết. Trong phương pháp này nhà phân tích

dữ liệu còn có thể áp dụng các phương pháp sau để lựa chọn ra các dữ liệu có đặc tính tốt nhất: Missing values ratio, Low-variance filter, High-correlation filter, Random forest, backwards-feature elimination và forward feature selection.

- Feature Extraction (dimensionality reduction): Giảm thiểu các đặc tính của dữ liệu bằng cách kết hợp các đặc tính đang có. Trong phương pháp này nhà phân tích dữ liệu có thể áp dụng các phương pháp sau để kết hợp các đặc tính của dữ liệu: Linear method và Manifold learning hay non-linear method

Điểm mạnh của kỹ thuật giảm chiều dữ liệu:

- Phương pháp làm nén dữ liệu giúp cho giảm dung lượng lưu trữ.
- Giảm thời gian xử lý dữ liệu
- Giảm các đặc tính thừa thải, không quan trọng.

Điểm yếu của kỹ thuật giảm chiều dữ liệu:

- Dẫn đến tình trạng mất dữ liệu
- Trong phương pháp PCA (Principal component analysis) trong phương pháp Linear method thường sẽ tìm các phương sai tuyến tính của biến, điều này đôi khi dẫn đến dữ liệu mà nhà phân tích dữ liệu không mong muốn.
- PCA cũng sẽ không hoạt động tốt nếu như giá trị trung bình và phương sai là không đủ để định nghĩa 1 tập dữ liệu.
- Trong phương pháp Backward Feature Elimination và Forward Feature Selection đều rất tốn thời gian xử lý và tài nguyên để tính toán nên chỉ phù hợp với các tập dữ liệu ít đặc tính.

1.4.3.3.3 Data compression

Nén dữ liệu là kỹ thuật chỉnh sửa dữ liệu, mã hoá dữ liệu hoặc thay đổi cấu trúc dữ liệu sao cho dữ liệu tốn ít dung lượng lưu trữ hơn. Nén dữ liệu bao gồm việc xây dựng cách diện đạt ngắn gọn của dữ liệu bằng cách loại bỏ đi các dữ liệu thừa và biểu diễn chúng dưới dạng nhị phân. Phương pháp nén dữ liệu mà có thể khôi phục dữ liệu về dạng ban đầu được gọi là phương pháp nén không mất mát (Loseless compression). Ngược lại phương pháp nén mà không thể khôi phục lại dạng ban đầu được gọi là nén mất dữ liệu (Lossy compression) [8].

Kỹ thuật này làm giảm kích thước của dữ liệu bằng việc sử dụng các phương pháp mã hoá khác nhau như Huffman Encoding và run-length Encoding. Các nhà phân tích dữ liệu có thể chia chúng làm 2 loại dựa trên cách thức chúng được nén:

- Nén không mất mát: Kỹ thuật mã hoá được dùng là Run-length Encoding. Kỹ thuật này cho phép sự giảm kích thước dữ liệu đơn giản. Kỹ thuật này có thể dùng thuật toán để khôi phục lại chính xác dữ liệu gốc.
- Nén mất dữ liệu: Trong kỹ thuật này, dữ liệu được mã hoá sẽ khác với dữ liệu gốc tuy nhiên vẫn có thể trích xuất ra đầy đủ các thông tin quan trọng của dữ liệu đấy.

1.4.3.3.4 Discretization & Concept Hierarchy Operation

Kỹ thuật rời rạc hoá dữ liệu được dùng để phân chia các thuộc tính liên tục thành các dữ liệu theo từng quãng. Các nhà phân tích dữ liệu có thể thay thế các hằng số bằng các nhãn của các khoảng nhỏ. Điều này có nghĩa là kết quả khai phá được thể hiện 1 cách ngắn gọn và dễ hiểu [8].

Một số phương pháp được dùng trong kỹ thuật rời rạc hoá dữ liệu [8]:

- Top-down discretization: Nếu nhà phân tích dữ liệu cân nhắc 1 hoặc 1 vài điểm (các điểm này được gọi là các điểm ngắt quản hoặc các điểm phân chia) để chia phân chia toàn bộ bộ thuộc tính và lặp lại cho đến khi kết thúc. Quá trình này được gọi là phân chia từ trên xuống hoặc được gọi là phân tách.
- Bottom-up discretization: Nếu nhà phân tích dữ liệu cân nhắc toàn bộ các giá trị hằng số đều là điểm phân chia thì một vài điểm sẽ bị loại bỏ thông qua sự kết hợp của các giá trị lân cận trong khoảng đó. Quá trình này được gọi là phân chia từ dưới lên.

1.4.4 Xử lý dữ liệu

Sau khi đã được lưu trữ và qua các bước tiền xử lý dữ liệu (data preprocessing) dữ liệu sẽ tiến hành chuyển đổi và tổ chức để thu được kết quả chính xác từ các truy vấn (query) phân tích. Tùy thuộc vào tài nguyên điện toán và phân tích sẵn có mà tổ chức, người dùng có thể chọn các cách khác nhau để xử lý dữ liệu [2].

1.4.4.1 Các phương pháp xử lý dữ liệu

- Manual Data Processing: Là phương pháp xử lý dữ liệu mà toàn bộ toàn bộ công đoạn thu thập, trích lọc và tính toán được thực hiện hoàn toàn bằng con người mà không dùng đến các thiết bị điện tử hoặc các phần mềm tự động. Đây là phương pháp ít tốn kém và không cần nhiều công cụ hỗ trợ tuy nhiên lại dễ xảy ra lỗi, chi phí nhân công cao và tiêu tốn nhiều thời gian [2].
- Mechanical Data Processing: Là phương pháp xử lý dữ liệu bằng việc xử dụng các máy móc và thiết bị. Những thiết bị này bao gồm: máy tính

bỏ túi, máy đánh chữ, máy in, … Phương pháp có thể xử lý các loại dữ liệu đơn giản và có ít lỗi hơn trong quá trình xử lý so với phương pháp xử lý thủ công. Tuy nhiên dữ liệu được tạo ra từ phương pháp này sẽ có độ phức tạp và độ khó cao hơn [2].

- Electronic Data Processing: Là phương pháp xử lý dữ liệu bằng các phần mềm và chương trình. Các phần mềm cung cấp các bộ hỗ trợ xử lý và tạo ra thành quả. Phương pháp này là phương pháp tốn kém nhất nhưng lại cung cấp tốc độ xử lý, độ chính xác và độ tin cậy cao nhất [2].

1.4.4.2 Các kỹ thuật xử lý dữ liệu

Có các kỹ thuật dữ liệu khác nhau dựa vào nguồn dữ liệu và các bước của đơn vị xử lý để thu được kết quả. Tuy nhiên không phải phương pháp nào cũng phù hợp để xử lý các dữ liệu thô [2].

- Batch Processing: Trong xử lý theo lô, dữ liệu được thu thập và xử lý theo từng lô và được sử dụng để xử lý 1 lượng dữ liệu lớn. VD: Hệ thống lương bổng của nhân viên.
- Single User Programming Processing: Xử lý cá nhân là loại xử lý được 1 cá nhân sử dụng để xử lý dữ liệu theo nhu cầu của cá nhân đấy. Kỹ thuật này phù hợp với các văn phòng hoặc doanh nghiệp nhỏ.
- Multiple Programming Processing: Xử lý đa chương trình là loại xử lý cho phép đồng thời lưu trữ và tính toán nhiều chương trình trong CPU. Dữ liệu được chia nhỏ thành các khung và tiến trình bằng cách sử dụng từ 2 hoặc nhiều nhân CPU trở lên trong 1 hệ thống máy tính. Chính vì thế kỹ thuật xử lý đa chương trình làm tăng hiệu xuất làm việc của máy tính. Ví dụ phổ biến của kỹ thuật xử lý này là dự báo thời tiết.
- Real-time Processing: Kỹ thuật này tạo điều kiện cho người dùng tương tác trực tiếp với hệ thống máy tính. Kỹ thuật này đơn giản hoá quá trình xử lý dữ liệu vì nó được thiết kế để được xử lý trực tiếp và chỉ được lập trình để xử lý 1 nhiệm vụ. Kỹ thuật là dạng xử lý trực tuyến và luôn được thực hiện. VD: Hệ thống rút tiền trên máy ATM.
- Online Processing: Kỹ thuật này tạo điều kiện cho việc xử lý dữ liệu đầu vào trực tiếp để cho hệ thống không phải lưu trữ sau đó mới phải xử lý. Kỹ thuật này được phát triển để giảm thiểu các lỗi đầu vào vì nó kiểm tra các dữ liệu ở nhiều điểm khác nhau và đảm bảo chỉ có dữ liệu chính xác mới được lưu trữ vào hệ thống. Kỹ thuật này được sử dụng rộng rãi trên các ứng dụng trực tuyến. Ví dụ như quét mã vạch.
- Time-sharing Processing: Đây là một dạng khác của xử lý dữ liệu trực tuyến mà tạo điều kiện cho một vài người dùng chia sẽ dữ liệu của hệ

thống trực tuyến. Kỹ thuật này được sử dụng khi cần có kết quả trong khoản thời gian nhanh nhất. Hơn nữa, hệ thống này dựa theo thời gian. Có nghĩa là nhiều người dùng sẽ được phục vụ cùng lúc và có cùng thời gian xử lý, cũng như có khả năng được tương tác với chương trình.

- Distributed Processing: Đây là dạng xử lý đặc biệt vì các máy tính khác nhau được kết nối với một máy chủ tạo thành một mạng lưới. Các máy tính được kết nối với tốc độ cao. Tuy nhiên máy chủ trung tâm duy trì cơ sở dữ liệu chủ và giám sát hoạt động của các máy khác.

1.4.5 Phân tích dữ liệu

Sau khi đã hoàn thành các bước xử lý dữ liệu, lúc này các nhà phân tích có thể tiến hành phân tích các dữ liệu đấy để thu được các thông tin quan trọng. Ở bước này họ cần tìm xu hướng, mối tương quan, các biến thể và hướng đi để giúp họ có câu trả lời cho vấn đề mà họ đang phân tích. Một vài công nghệ được sử dụng để hỗ trợ các nhà phân tích trong việc quản lý dữ liệu [9].

Phân tích dữ liệu được chia ra làm các hạng mục khác nhau tuỳ vào độ phức tạp và nổ lực để đánh giá dữ liệu mà kết quả thu được cũng khác nhau

1.4.5.1 Các hạng mục phân tích chính

Theo Datapine, các hạng mục chính trong phân tích dữ liệu bao gồm [9]:

- Descriptive analysis
- Exploratory analysis
- Diagnostic analysis
- Predictive analysis
- Prescriptive analysis

1.4.5.1.1 Descriptive analysis

Phương pháp phân tích mô tả là điểm bắt đầu cho mọi phản ánh phân tích và mục tiêu của nó là để trả lời cho câu hỏi chuyện gì sẽ xảy ra ? Phương pháp này thực hiện bằng cách sắp xếp, điều khiển và phiên dịch các dữ liệu thô từ các nguồn khác nhau và biến các dữ liệu thô này thành các thông tin chuyên sau có ích cho người phân tích [9].

Thực hiện việc phân tích mô tả là việc quan trọng, phương pháp giúp các nhà phân tích dữ liệu thể hiện góc nhìn của họ theo một cách có nghĩa. Mặc dù phương pháp này sẽ không dự báo trước được các kết quả trong tương lai hoặc trả lời các thắc mắc của họ về việc tại sao nó lại xảy ra ? Tuy nhiên phương pháp này sẽ giúp cho họ tổ chức dữ liệu để sẵn sàng cho các phân tích trong tương lai. [9]

1.4.5.1.2 Exploratory analysis

Mục đích chính của phương pháp này là để khám phá dữ liệu. Trước đó khi thực hiện phân tích khai phá thì vẫn chưa có khái niệm về mối quan hệ giữa dữ liệu và các biến. Khi mà dữ liệu được nghiên cứu, việc phân tích khai phá sẽ giúp các nhà phân tích tìm hiểu được các mối liên kết và tạo ra các giả thuyết và đáp án cho các vấn đề mà họ đang nghiên cứu. Một trong những ứng dụng cho phương pháp này là khai phá dữ liệu (data mining) [9].

1.4.5.1.3 Diagnostic analysis

Phương pháp phân tích chẩn đoán tăng cường việc phân tích và thực thi bằng cách cung cấp cho các nhà phân tích cái nhìn vững chắc về ngữ cảnh của việc tại sao nó lại xảy ra ? Nếu họ biết được vì sao một việc gì đấy xảy ra cũng như cách mà nó xảy ra thì họ sẽ dễ dàng xác định cách thức chính xác để giải quyết vấn đề [9].

Phương pháp này được thiết kế để đưa ra cách thức giải quyết trực tiếp và câu trả lời cho các vấn đề. Đây là một trong những phương pháp quan trọng nhất trên thế giới trong việc nghiên cứu [9].

1.4.5.1.4 Predictive analysis

Phương pháp phân tích dự đoán cho phép các nhà phân tích nhìn vào tương lai đề tìm kiếm câu trả lời cho câu hỏi việc gì sẽ xảy ra ? Để có thể làm được điều này, phương pháp sử dụng kết quả của các phương pháp trước ngoài ra còn kết hợp thêm máy học (machine learning) và trí tuệ nhân tạo (artificial intelligence) để tìm ra xu hướng, các vấn đề hoặc sự thiếu hiệu quả và kết nối trong dữ liệu của họ [9].

Với phương pháp này các nhà phân tích có thể phát triển các khả năng chẩn đoán để tăng cường quá trình xử lý và giúp doanh nghiệp hay tổ chức có lợi thế so với các doanh nghiệp khác.

1.4.5.1.5 Prescriptive analysis

Phương pháp phân tích đề xuất cũng là một trong những phương pháp hiệu quả nhất trong nghiên cứu. Kỹ thuật đề xuất dữ liệu sử dụng các đường đi hoặc xu hướng để phát triển đối sách và các chiến lược kinh doanh hiệu quả [9].

Bằng cách tìm hiểu sâu phương pháp phân tích đề xuất, các nhà phân tích dữ liệu sẽ đóng vai trò chủ động trong quá trình tiêu thụ dữ liệu bằng cách sử dụng những tập dữ liệu trực quan được sắp xếp và dùng chúng như những công cụ hiệu quả cho các vấn đề quan trọng trong một số lĩnh vực như marketing, bán hàng, trải nghiệm khác hàng, nhân lực, tài chính,…

1.4.5.2 Các kỹ thuật phân tích quan trọng

Với kinh nghiệm lâu năm và đã cộng tác với hơn 150 doanh nghiệp khác nhau Datapine đã chia các kỹ thuật phân tích được phân thành 2 hạng mục quan trọng là: Phân tích định lượng và phân tích chất lượng. Mỗi hạng mục đều mang lại kết quả phân tích khác nhau dựa vào tình huống và loại dữ liệu mà nhà phân tích dữ liệu đang phân tích [9].

1.4.5.2.1 Quantitative Methods

Phương pháp phân tích định lượng ám chỉ tất cả các kỹ thuật phân tích mà sử dụng đến các dữ liệu là số hoặc các dữ liệu sẽ chuyển thành số để trích xuất các giá trị chuyên sâu. Phương pháp được sử dụng để trích xuất các kết luận về mối quan hệ, sự khác biệt và các giả thuyết thử nghiệm [9].

Các kỹ thuật thuộc phương pháp phân tích định lượng bao gồm [9]:

- Phân tích theo từng cụm (Cluster analysis)
- Phân tích cohort (Cohort analysis)
- Phân tích hồi quy (Regression analysis)
- Mạng lưới thần kinh (Neural networks)
- Phân tích yếu tố (Factor analysis)
- Khai phá dữ liệu (Data mining)
- Phân tích chuỗi thời gian (Time series analysis)
- Cây quyết định (Decision trees)
- Phân tích liên kết (Conjoint analysis)
- Phân tích tương hợp (Correspondence analysis)
- Mở rộng đa chiều (Multidimensional scaling)

1.4.5.2.2 Qualitive methods

Là phương pháp được định nghĩa khi quan sát về các dữ liệu không phải số dược thu thập và tạo ra từ các phương pháp quan sát như phỏng vấn, nhóm tập trung, bảng câu hỏi, … So với phương pháp phân tích định lượng thì phương pháp phân tích chất lượng mang nhiều tính chủ quan và có giá trị cao hơn khi phân tích phương pháp giữ chân khách hàng và phát triển sản phẩm [9].

Các kỹ thuật thuộc phương pháp phân tích chất lượng bao gồm [9]:

- Phân tích văn bản (Text analysis)
- Phân tích nội dung (Content analysis)
- Phân tích chuyên đề (Thematic analysis)
- Phân tích tường thuật (Narrative analysis)
- Phân tích nghị luận (Discoure analysis)
- Phân tích lý thuyết căn cứ (Grounded theory analysis)

1.4.6 Diễn giải dữ liệu và đưa ra kết luận

Sau khi đã có được kết quả phân tích, các nhà phân tích dữ liệu cần phải diễn giải và giải thích kết quả vừa phân tích để đưa ra được hướng hành động tốt nhất. Với việc kết hợp giữa phân tích và diễn giải dữ liệu có thể giúp tăng cường hiệu suất làm việc và xác định các vấn đề. Việc phát triển sẽ trở nên khó khăn và không đáng tin cậy khi thiếu đi việc diễn giải dữ liệu. Vì những lợi ích mà diễn giải mang lại ảnh hưởng rất lớn đến quá trình lên ý tưởng cho việc phát triển nhất là đối với các doanh nghiệp hay tổ chức lớn.

Các lợi ích mà diễn giải dữ liệu mang lại:

- Đưa ra các quyết định sáng suốt
- Dự đoán nhu cầu bằng cách xác định xu hướng
- Tiết kiệm chi phí
- Làm rõ tầm nhìn trong tương lai

1.4.6.1 Các kỹ thuật để diễn giải dữ liệu

1.4.6.1.1 Trực quan hoá dữ liệu

Các biểu đồ dữ liệu như biểu đồ doanh nghiệp, biểu đồ cột, bảng,… là các công cụ hiệu quả và trực quan để diễn giải dữ liệu. Bởi vì việc sử dụng các biểu đồ sẽ khiến cho các dữ liệu trở nên dễ hiểu khi các thông tin được tóm gọn. Tuy nhiên việc sử dụng không đúng loại biểu đồ sẽ có thể dẫn đến hiểu lầm. Vì thế việc chọn đúng loại biểu đồ là vô cùng quan trọng trong việc diễn giải dữ liệu và kết quả phân tích.

Một số loại biểu đồ phổ biến:

- Biểu đồ cột: Một trong số những loại biểu đồ phổ biến nhất là biểu đồ cột. Loại này được sử dụng để diễn tả mối quan hệ giữa 2 hoặc nhiều biến. Có nhiều loại biểu đồ cột trong đó: biểu đồ cột ngang, biểu đồ cột dọc và biểu đồ xếp chồng.
- Biểu đồ đường: Biểu đồ đường được dùng chủ yếu để diễn giải các dữ liệu mang tính xu hướng như sự tăng, giảm và biến động. Biểu đồ đường thể hiện sự biến đổi của dữ liệu theo thời gian.
- Biểu đồ tròn: Biểu đồ tròn được sử dụng để thể hiện tỷ lệ của một biến. Mặc dù thể hiện dữ liệu theo tỷ lệ phần trăm thì biểu đồ cột sẽ thể hiện được các dữ liệu ở định dạng phức tạp hơn. Tuy nhiên, việc này cũng dựa vào số biến mà nhà phân tích dữ liệu đang so sánh. Để tận dụng tối đa khả năng biểu diễn của biểu đồ tròn thì nên có dưới 10 biến.

- Bảng: Bảng là loại được dùng để miêu tả dữ liệu ở định dạng thô của nó. Nó cũng cung cấp cho nhà phân tích dữ liệu sự tự do trong việc so sánh các giá trị độc lập trong khi vẫn hiển thị tổng số.

1.4.6.1.2 Tuân theo các mục tiêu phân tích

Trong quá trình diễn giải dữ liệu hãy cố gắng tuân theo mục tiêu phân tích ban đầu. Vì kết quả phân tích của nhà phân tích dữ liệu thường có tính chủ quan và vì thế để đảm bảo việc phân tích đi đúng hướng thì có thể cho các đồng nghiệp khác xem qua hoặc những người sẽ sử dụng kết quả phân tích để họ có thể góp ý hoặc chỉ ra các lỗi trong kết quả phân tích trước đó.

1.4.6.2 Kết luận

Sau khi đã có được kết quả từ quá trình phân tích từ dữ liệu thu thập. Các kết quả này sẽ trở thành thông tin quan trọng trong việc nghiên cứu và kết luận. Dựa vào suy nghĩ và lý luận cũng như việc cẩn trọng trước những dữ liệu lỗi, mối tương quan, hậu quả, thông tin không chính xác,… Hãy kiểm tra xem các câu hỏi mà doanh nghiệp hay tổ chức đề ra khi phân tích đã được trả lời hay các gợi ý cho việc phân tích. Khi đã cảm thấy các bước trên đã đạt thì các nhà phân tích dữ liệu có thể tiến đến việc kết luận quá trình phân tích.

CHƯƠNG 2. KHAI PHÁ DỮ LIỆU

2.1 Giới thiệu

2.1.1 Giới thiệu chung

Khai phá dữ liệu, còn được gọi là khám phá tri thức trong dữ liệu, là quá trình khám phá các mẫu và thông tin có giá trị khác từ các tập dữ liệu lớn. Với sự phát triển của công nghệ lưu trữ dữ liệu và sự phát triển của dữ liệu lớn, việc áp dụng các kỹ thuật khai thác dữ liệu đã tăng tốc nhanh chóng trong vài thập kỷ qua, hỗ trợ các doanh nghiệp và tổ chức bằng cách chuyển đổi dữ liệu thô của họ thành kiến thức hữu ích. Tuy nhiên, bất chấp thực tế là công nghệ liên tục phát triển để xử lý dữ liệu ở quy mô lớn, các nhà lãnh đạo vẫn phải đối mặt với những thách thức về khả năng mở rộng và tự động hóa [10].

Khai thác dữ liệu đã cải thiện việc ra quyết định của tổ chức thông qua các phân tích dữ liệu sâu sắc. Các kỹ thuật khai thác dữ liệu làm cơ sở cho các phân tích này có thể được chia thành hai mục đích chính. Các kỹ thuật có thể mô tả tập dữ liệu mục tiêu hoặc các kỹ thuật có thể dự đoán kết quả thông qua việc sử dụng các thuật toán máy học. Các phương pháp này được sử dụng để sắp xếp và lọc dữ liệu, hiển thị thông tin thú vị nhất, từ phát hiện gian lận đến hành vi của người dùng, tắc nghẽn và thậm chí cả vi phạm bảo mật. Các công ty sử dụng phần mềm khai phá dữ liệu để tìm hiểu thêm về khách hàng. Điều này có thể giúp cho các công ty để phát triển các chiến lược marketing, tăng doanh số bán hàng và giảm chi phí. Việc khai phá dữ liệu dựa vào việc thu thập dữ liệu hiệu quả, hệ thống quản lý dữ liệu và xử lý máy tính [10].

Các tri thức được rút ra từ việc khai phá dữ liệu có thể dùng để:

- Giải thích dữ liệu: Cung cấp sự hiểu biết sâu sắc và rất hữu ích về hành vi của các đối tượng, giúp cho các doanh nghiệp hiểu rõ hơn những khách hàng của họ.
- Dự báo: Dự báo giá trị của những đối tượng mới
  - Khuynh hướng mua hàng của khách hàng
  - Xác định rủi ro tính dụng đối với một khách hàng.
  - Định hướng tập trung nguồn lực của doanh nghiệp

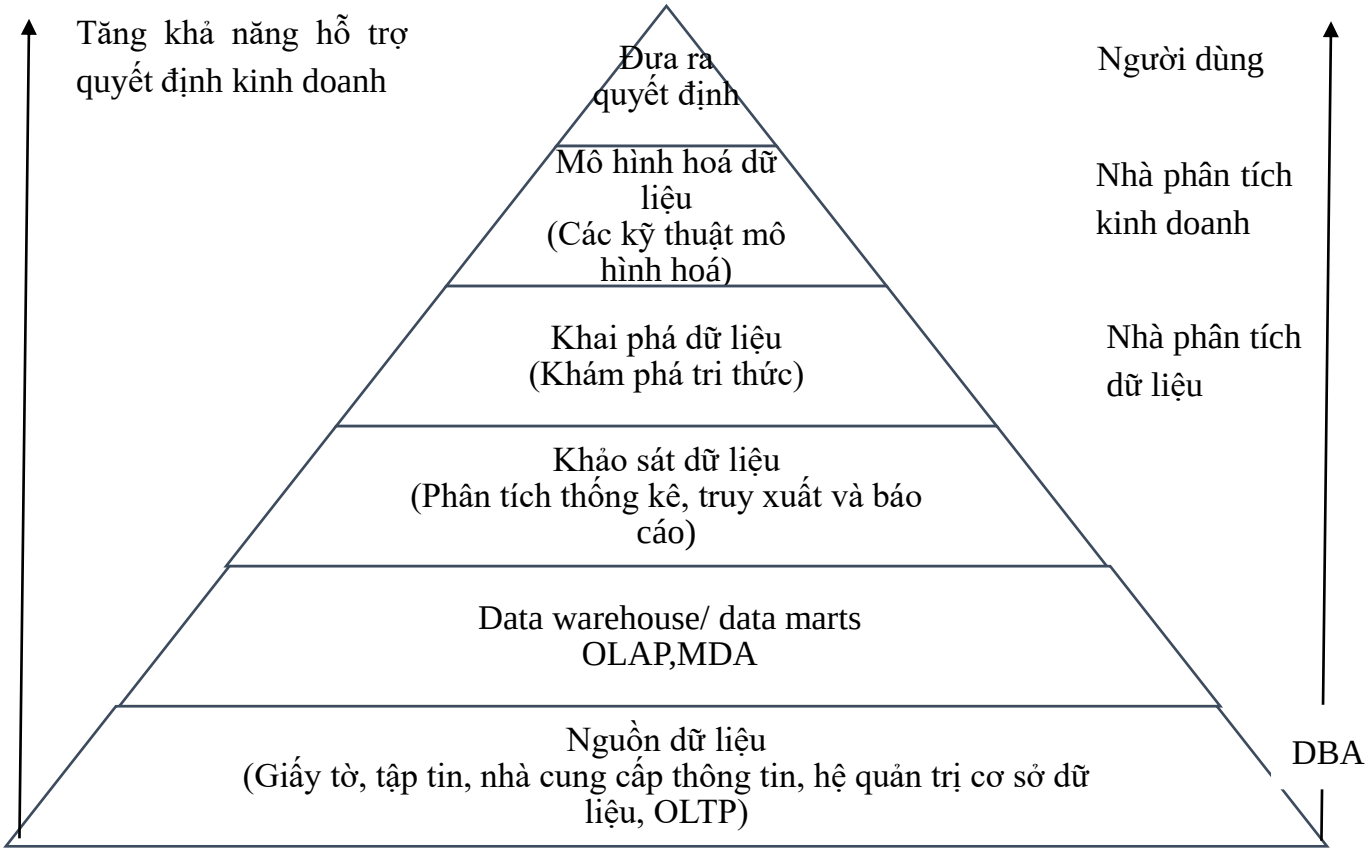
2.1.2 Lịch sử phát triển của khai phá dữ liệu

Một số các cột mốc của sự phát triển của kỹ thuật, công nghệ lưu trữ và khai phá dữ liệu như sau [10]:

- Những năm 1960: xuất hiện cơ sở dữ liệu theo mô hình phân cấp.
- Những năm 1970: thiết lập nền tảng lý thuyết cho cơ sở dữ liệu quan hệ. Các hệ quản trị cơ sở dữ liệu.

- Những năm 1980: hoàn thành lý thuyết về cơ sở dữ liệu quan hệ và các hệ quản trị cơ sở dữ liệu quan hệ, xuất hiện các hệ quản trị cơ sở dữ liệu cao cấp như hướng đối tượng, suy diễn,… và hệ quản trị cơ sở dữ liệu hướng ứng dụng trong lĩnh vực không gian, khoa học, công nghiệp, nông nghiệp,…
- Những năm 1990-2000: Là sự phát triển của khai thác dữ liệu và kho dữ liệu. Cơ sở dữ liệu đa phương tiện và cơ sở dữ liệu web.

Khai thác dữ liệu là công đoạn trong tiến trình lớn hơn là khám phá tri thức từ cơ sở dữ liệu (Knowledge Discovery in Database – KDD). Khai thác dữ liệu mang tính trực giác, cho phép thu được những hiểu biết rõ ràng và sâu sắc hơn. Khai thác dữ liệu còn giúp phát hiện những xu thế phát triển từ những thông tin quá khứ, cũng như cho phép đề xuất các dự báo mang tính thống kê, gom cụm và phân loại dữ liệu. Kho dữ liệu điển hình trong những doanh nghiệp cho phép người dùng hỏi và trả lời những câu hỏi như: "Doanh số bán ra là bao nghiêu tính theo khu vực, theo nhân viên bán hàng từ tháng 7 đến tháng 9 năm 2021". Trong khi đó khai thác dữ liệu sẽ cho phép người ra quyết định kinh doanh hỏi và trả lời các câu hỏi mang tính cá nhân như " Ai là khách hàng chủ yếu của mặt hàng này ? ", " Dòng sản phẩm này ai sẽ là tập khách hàng chủ yếu trong khu vực này dựa vào những sản phẩm tương tự trong khu vực đấy? ". Vị trí của khai thác dữ liệu được thể hiện qua sơ đồ [10]:



Hình 2-1 Vị trí của khai phá dữ liệu

2.1.3 Lý do sử dụng khai phá dữ liệu

Khai thác dữ liệu là cần thiết đối với người dùng vì những lý do sau [10]:

- Ngày càng có nhiều dữ liệu được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu dẫn đến việc hình thành các "mỏ vàng dữ liệu" chứa đầy thông tin chiến lược mà các hệ quản trị cơ sở dữ liệu thông thường không thể phát hiện và quản trị được
- Cơ sở dữ liệu phát triển nhanh về cả kích thước lẫn số lượng. Không xét thông tin mang tính sự kiện được lưu trữ trong cơ sở dữ liệu, những thông tin được suy diễn từ nó cũng hết sức thú vị. Tuy nhiên với các quan hệ có số lượng khổng lồ các bản ghi (record) và có quá nhiều trường (field). Việc này duyệt hàng triểu bản ghi hay hàng trăm trường tin để tìm ra các mẫu và các quy luật là một thách thức và trở ngại thật sự đối với nhà phân tích dữ liệu.
- Không phải ai cũng là nhà thống kê hay nhà phân tích dữ liệu chuyên nghiệp.
- Sử dụng cho các trường hợp tìm kiếm nhưng chưa xác lập rõ hoặc chưa mô tả được điều kiện tìm kiếm. Nếu người dùng biết họ đang tìm kiếm thông tin nào thì họ có thể sử dụng SQL, nhưng trong trường hợp họ chỉ có ý tưởng không rõ ràng thì họ có thể sử dụng khai phá dữ liệu.

Khai phá dữ liệu là công cụ hiệu quả trong các lĩnh vực [10]:

- Sử dụng dữ liệu để xây dựng các mô hình dự báo:
  - Khả năng dự báo tiềm ẩn trong dữ liệu
  - Gợi ý về các chiều và các nhóm dữ liệu có khả năng chứa các tri thức hữu ích.
- Tạo tóm tắt và báo cáo rõ ràng:
  - Tự động tìm những phân đoạn trong dữ liệu
  - Tìm ra những phân đoạn mà nhà phân tích chưa biết hoặc có hiểu biết chưa rõ ràng.
- Cung cấp cơ chế hỗ trợ ra quyết định:
  - Dự báo
  - Mô hình hoá

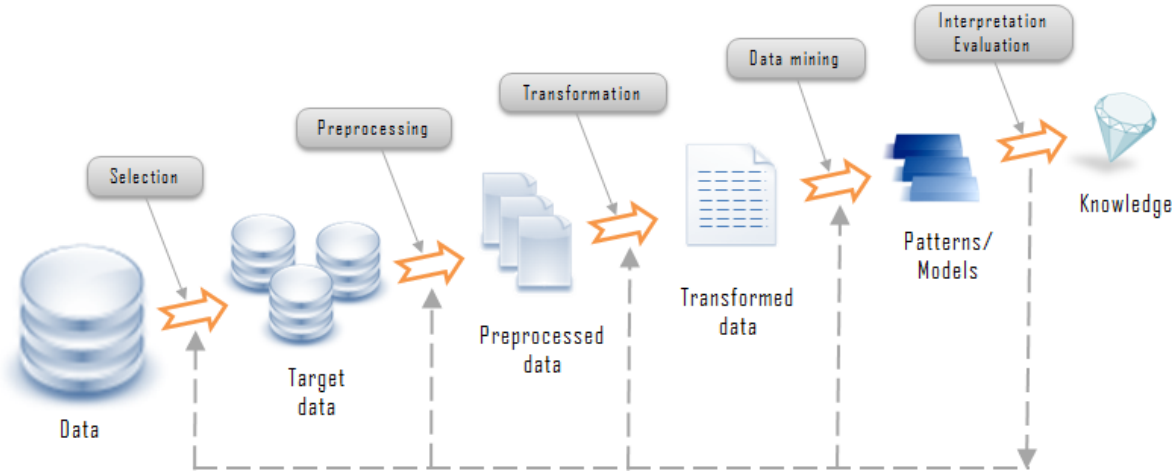
2.2 Các công đoạn khai phá dữ liệu

Tiến trình khai phá dữ liệu từ cơ sở dữ liệu gồm 3 công đoạn [10]:

- Chuẩn bị dữ liệu:
  - Chọn lọc dữ liệu
  - Làm sạch dữ liệu

- Làm giàu dữ liệu
- Mã hoá dữ liệu

- Khai thác dữ liệu
- Tường trình, báo cáo kết quả



Hình 2-2 Các công đoạn khai phá dữ liệu

Tại mỗi công đoạn, tiến trình có thể quay lùi lại một hay nhiều giai đoạn. Ví dụ tại giai đoạn khám phá hay mã hoá dữ liệu, tiến trình vẫn có thể quay trở về giai đoạn xoá bỏ dữ liệu, hay có thể quay về giai đoạn làm giàu dữ liệu nếu có được các dữ liệu mới để sử dụng chúng cho việc làm giàu dữ liệu có sẵn.

2.2.1 Giai đoạn chuẩn bị dữ liệu

2.2.1.1 Chọn lọc dữ liệu

Đây là giai đoạn chọn lọc và trích xuất dữ liệu cần thiết từ cơ sở dữ liệu tác nghiệp vào một cơ sở dữ liệu riêng. Nhà phân tích chỉ cần chọn ra các dữ liệu phù hợp cho các giai đoạn sau này. Tuy nhiên việc chọn lọc dữ liệu thường rất khó khăn do dữ liệu thường nằm phân tán, rải rác khắp nơi ngoài ra chúng đôi khi còn tồn tại và được tổ chức theo các dạng khác nhau. Ví dụ như một nơi lại dùng kiểu dữ liệu cho mã sản phẩm là kiểu số trong khi nơi khác lại dùng kiểu ký tự cho mã sản phẩm [10].

2.2.1.2 Làm sạch dữ liệu

Phần lớn các cơ sở dữ liệu đều ít nhiều mang tính không nhất quán. Do vậy khi khai phá dữ liệu trên các cơ sở dữ liệu đó thường không đảm bảo tính đúng đắn. Ví dụ: Trong các công ty bảo hiểm thì việc lưu trữ thông tin về ngày sinh của khách hàng cần phải có sự chính xác cao. Trong khi đó, 30-40% dữ liệu về độ tuổi của khách hàng được lưu trong cơ sở dữ liệu đều là để trống hoặc dữ liệu bị sai lệch. Tuy nhiên đối với khai phá dữ liệu việc dữ liệu bị thiếu ảnh hưởng đến độ chính xác của dữ liệu được khai phá. Do đó trước khi bắt đầu việc khai phá dữ liệu thì nhà phân tích phải

tiến hành xoá bỏ dữ liệu không cần thiết. Và việc xoá bỏ dữ liệu này xảy ra nhiều lần do trong quá trình khai phá dữ liệu mới phát hiện ra các bất thường trong dữ liệu [10].

2.2.1.3 Làm giàu dữ liệu

Mục đích của giai đoạn này là bổ sung thêm các thông tin có liên quan vào cơ sở dữ liệu gốc. Bằng cách sử dụng các thông tin trích xuất từ các cơ sở dữ liệu khác để bổ sung vào cơ sở dữ liệu ban đầu để làm giàu thêm thông tin dữ liệu [10].

2.2.2 Giai đoạn mã hoá dữ liệu

Mục đích của giai đoạn mã hoá dữ liệu là chuyển đổi kiểu dữ liệu về những dạng thuận tiện để tiến hành các thuật toán khám phá dữ liệu. Một số cách mã hoá dữ liệu khác nhau [10]:

- Phân vùng: Với dữ liệu dạng chuỗi nằm trong các tập chuỗi cố định
- Biến đổi giá trị năm thành số nguyên là số năm đã trôi qua so với năm hiện tại
- Chia giá trị số theo một hệ số để tập các giá trị nằm trong vùng nhỏ hơn
- Chuyển đổi giá trị yes-no, true-false thành 0-1.

2.2.3 Khai phá dữ liệu

Khai phá dữ liệu là tiến trình điều chỉnh đúng các mô hình dữ liệu.Chức năng biến đổi dữ liệu được đưa vào bước này với mục đích trình diễn dữ liệu.

Bất kỳ vấn đề kinh doanh nào cũng sẽ kiểm tra mô hình thô để xây dựng một mô hình mô tả thông tin và đưa ra các báo cáo để doanh nghiệp sử dụng.Xây dựng mô hình từ các nguồn dữ liệu và các định dạng dữ liệu là quá trình lặp đi lặp lại vì dữ liệu thô có sẵn ở nhiều nguồn và nhiều định dạng khác nhau. Dữ liệu tăng lên từng ngày, do đó khi một nguồn dữ liệu mới được tìm thấy, thì nguồn dữ liệu đấy có thể gây ra thay đổi kết quả [10].

2.2.4 Trình diễn dữ liệu

Trình diễn dữ liệu là quá trình giải thích và hiển thị trực quan các kết quả khai phá dữ liệu để hỗ trợ việc định giá chất lượng dữ liệu. Đánh giá mô hình dữ liệu lựa chọn có phù hợp hay không, và thể hiện mô hình. Mỗi bước (trừ bước lưu trữ dữ liệu) cho phép tương tác người dùng, và một số bước có thể thực hiện hoàn toàn thủ công [9].

2.3 Khái quát các kỹ thuật khai phá dữ liệu

2.3.1 Khai phá tập phổ biến và luật kết hợp

Là phương pháp tìm ra các quy tắc kết hợp thể hiện các điều kiện thuộc tính-xảy ra thường xuyên cùng nhau trong một tập hợp dữ liệu nhất định. Phân tích kết hợp thường xử dụng rộng rãi cho phương pháp giỏ hàng hoặc phân tích dữ liệu giao

dịch. Khai phá kết hợp là một trong những lĩnh vực quan trọng và đặc biệt linh hoạt trong việc nghiên cứu khai phá dữ liệu. Một trong những phương pháp phân loại dựa trên nguyên lý liên kết là phân loại liên kết, bao gồm hai bước. Trong bước chính, các hướng dẫn liên kết được tạo ra bằng thuật toán kết hợp đã được chỉnh sửa gọi là Apriori. Bước tiếp theo là xây dựng một bộ phận phân loại dựa trên các luật kết hợp được khám phá [10]. Trong cơ sở dữ liệu bán hàng, một luật kết hợp tiêu biểu sau :

Có 66% khách hàng mua bia Tiger, thịt bò mỹ thì thường mua kèm măng tây.

Luật kết hợp giúp các nhà phân tích hoạch định hiểu rõ xu thế bán hàng, tâm lý khách hàng,… Từ đó đưa ra các chiến lược bố trí mặt hàng, kinh doanh, tiếp thị,…

2.3.2 Khai thác mẫu tuần tự

Là tiến trình khám phá các mẫu tuần tự phổ biến phản ánh các mối quan hệ giữa các biến cố có trong các cơ sở dữ liệu hướng thời gian. Một luật mô tả mẫu tuần tự có dạng tiêu biểu X->Y phản ánh sự xuất hiện kế tiếp biến cố Y [10]. Một luật thể hiện mẫu tuần tự tiêu biểu :

Có 80% khách hàng mua giày Nike thường có xu hướng sau 2 ngày sẽ mua tất Nike.

2.3.3 Phân lớp dữ liệu

Là tiến trình khám phá các luật phân loại đặc trưng cho các tập dữ liệu đã được xếp lớp. Tập dữ liệu học bao gồm tập đối tượng đã xác định lớp sẽ được dùng để tạo mô hình phân lớp dựa trên đặc trưng của đối tượng trong tập dữ liệu. Các luật phân lớp được sử dụng để xây dựng các bộ phân lớp dữ liệu. Phân lớp dữ liệu có vai trò quan trọng trong tiến trình dự bao khuynh hướng, quy luật phát triển [10].

Trong phân lớp dữ liệu, các kỹ thuật thường được sử dụng cho việc khai phá bao gồm:

- Cây quyết định
  - Cây quyết định gồm các nút trong biểu diễn giá trị thuộc tính, các nhánh biểu diễn đầu ra của kiểm tra, nút lá biểu diễn nhãn lớp. Cây quyết định được tạo theo hai giai đoạn là tạo cây và tỉa nhánh.
  - Các thuật toán nổi tiếng trong cây quyết định: ID3, Gini,…
- Bayes
  - Bộ phân lớp Bayes là một giải thuật thuộc lớp giải thuật thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu. Phân lớp Bayes được dựa trên định lý Bayes (định lý được đặt theo tên tác giả của nó là Thomas Bayes)

2.3.4 Khai thác cụm

Không giống như phân loại và dự đoán bằng cách phân tích các đối tượng hoặc thuộc tính dữ liệu được gắn nhãn lớp, phân cụm phân tích các đối tượng dữ liệu mà không tham khảo nhãn lớp đã xác định. Nói chung, các nhãn lớp không tồn tại trong dữ liệu huấn luyện đơn giản vì chúng không được biết đến từ đầu. Phân cụm có thể được sử dụng để tạo các nhãn này. Các đối tượng được phân cụm dựa trên nguyên tắc tối đa hóa độ tương tự trong lớp và giảm thiểu độ tương tự giữa các lớp. Nghĩa là, các cụm đối tượng được tạo ra sao cho các đối tượng bên trong một cụm có độ tương phản cao với nhau nhưng lại là các đối tượng khác nhau trong các cụm khác. Mỗi cụm được tạo có thể được coi là một lớp đối tượng, từ đó có thể suy ra các quy tắc. Phân cụm cũng có thể tạo thuận lợi cho việc hình thành phân loại, nghĩa là tổ chức các quan sát thành một hệ thống phân cấp của các lớp nhóm các sự kiện tương tự lại với nhau [10].

Khai thác phân cụm sử dụng các kỹ thuật như :

- Phương pháp phân hoạch
  - Đây là phương pháp tạo phân hoạch cơ sở dữ liệu sao cho cơ sỡ dữ liệu
  - Các phương pháp heuristic:
    - K-means
    - K-medoids
- Phương pháp phân cấp
  - Đây là phương pháp tạo phân cấp cụm (hierarchial clustering) chứ không tạo phân hoạch các đối tượng. Phương pháp này không cần phải xác định số cụm từ đầu. Số cụm sẽ đó khoảng cách giữa các cụm hoặc điều kiện dừng quyết định. Tiêu chuẩn gom cụm thường được xác định bởi ma trận khoảng cách. Phân cấp cụm thường được biểu diễn dưới dạng đồ thị dạng cây các cụm (dendogram). Lá của cây biểu diễn đối tượng riêng lẻ, nút trong biểu diễn các cụm.
- Phương pháp dựa trên mật độ
  - Là phương pháp đề cập đến các phương pháp học không giám sát nhằm xác định các cụm phân biệt trong phân phối của dữ liệu, dựa trên ý tưởng rằng một cụm trong không gian dữ liệu là một vùng có mật độ điểm cao được ngăn cách với các cụm khác bằng các vùng liền kề có mật độ điểm thấp .
- Phương pháp dựa trên mô hình
  - Đây là phương pháp dựa trên sự phù hợp giữa dữ liệu và các mô hình toán học. Ý tưởng của phương pháp này là: Dữ liệu phát sinh từ một sự kết hợp nào đó của các phân phối xác xuất ẩn

- Phương pháp dựa trên lưới:
  - Ý tưởng: dùng các cấu trúc dữ liệu dạng lưới với nhiều cấp độ phân giải. Những ô lưới có có mật độ cao sẽ tạo thành những cụm. Phương pháp này rất phù hợp với các phân tích gom cụm ứng dụng trong không gian. Ngoài ra còn có các thuật toán khác như STING, WaveCluster, CLIQUE

2.4 Tập thường xuyên và luật kết hợp

2.4.1 Mở đầu

Giả định chúng ta có rất nhiều mặt hàng, ví dụ như "bánh mì", "sữa",…(coi là tính chất hoặc trường). Khách hàng khi đi siêu thị sẽ bỏ vào giỏ mua hàng của họ một số mặt hàng nào đó, và chúng ta muốn tìm hiểu các khách hàng thường mua các mặt hàng nào đồng thời, chúng ta không cần biết khách hàng cụ thể là ai. Nhà quản lý dùng những thông tin này để điều chỉnh việc nhập hàng về siêu thị, hay đơn giản là để bố trí sắp xếp các mặt hàng gần nhau, hoặc bán các mặt hàng đó theo một gói hàng, giúp cho khắc đỡ mất công tìm kiếm [11].

Khai phá luật kết hợp được mô tả như sự tương quan của các sự kiện những sự kiện xuất hiện thường xuyên một cách đồng thời. Nhiệm vụ chính của khai phá luật kết hợp là phát hiện ra các tập con cùng xuất hiện trong một khối lượng giao dịch lớn của một cơ sở dữ liệu cho trước [11].

2.4.2 Định nghĩa về luật kết hợp

Cho tập dữ liệu riêng biệt có dạng I={I1,I2,…,Im}. Giả sử D là cơ sở dữ liệu với các bản ghi chưa tập con T của các tính chất (có thể coi như T $\subseteq$ I), các bản ghi đều có chỉ số riêng biệt. Một luận kết hợp là một mệnh đề kéo theo có dạng $X \rightarrow Y$, trong đó X,Y $\subseteq$ I, thoả mãn điều kiện $X \cap Y = \emptyset$. Các tập hợp X và Y được gọi là các tập hợp tính chất (itemset). Tập X gọi là nguyên nhân, tập Y gọi là hệ quả [11].

Có 2 độ đo quan trọng đối với độ kết hợp: Độ hỗ trợ (support) và độ tin cậy (confidence), được định nghĩa như phần dưới đây.

2.4.2.1 Độ hỗ trợ

Độ hỗ trợ là một tập hợp X trong cơ sở dữ liệu D là tỷ số giữa các bản ghi T$\subseteq$D có chứa tập X và tổng số bản ghi trong D(hay là phần trăm của các bản ghi trong D có chưa tập X), ký hiệu là support(X) hay supp(X) (support sẽ tự sinh ra khi cài thuật toán) [12].

$$S_0 = \frac{|\{T \subset D : T \supset X\}|}{|D|}$$

Ta có: 0 ≤ supp(X) ≤ 1 với mọi tập hợp X

Độ hỗ trợ của một tập kết hợp $X \rightarrow Y$ là tỷ lệ giữa số lượng các bản ghi chứa tập hợp $X \cup Y$, so với tổng số các bản ghi trong D – Ký hiệu supp($X \rightarrow Y$) [12].

$$Supp(X \rightarrow Y) = \frac{|\{T \subset D: T \supset X \cup Y\}|}{|D|}$$

Trong đó:

D là cơ sở dữ liệu

X, Y là tập hợp có trong D

T là tỷ số giữa các bản ghi có chứa tập X

Khi chúng ta nói rằng độ hỗ trợ của một luật là 50% thì có nghĩa là có 50% tổng số bản ghi chứa $X \cup Y$.Như vậy độ hỗ trợ mang ý nghĩa thống kê của luật.

2.4.2.2 Độ tin cậy

Độ tin cậy của một luật kết hợp $X \rightarrow Y$ là tỷ lệ giữa số lượng các bản ghi trong D chứa $X \cup Y$ với số bản ghi trong D có chứa tập hợp X. Ký hiệu độ tin cậy của một luật là conf(r). Ta có $0 \leq conf(r) \leq 1$ [11].

Độ hỗ trợ và độ tin cậy có xác xuất sau [12]:

$$Supp(X \rightarrow Y) = P(X \cup Y)$$

Trong đó:

X, Y là tập hợp có trong cơ sở dữ liệu

P(X∪Y) là xác xuất kết hợp của X hoặc Y

Supp(X→Y) là độ hỗ trợ hay độ phổ biến của tập chứa đồng thời X và Y [12].

$$Conf(X \rightarrow Y) = P\left(\frac{Y}{X}\right) = \frac{supp(X \cup Y)}{supp(X)}$$

Trong đó:

X, Y là tập hợp có trong cơ sở dữ liệu

P(Y/X) là xác xuất của Y so với X

Supp(X∪Y) là độ hỗ trợ hay độ phổ biến của tập chứa đồng thời X và Y.

Supp(X) là độ hỗ trợ hay độ phổ biến của tập chứa X trong cơ sở dữ liệu

Độ tin cậy của một luật k ết hợp $X \rightarrow Y$ là tỷ lệ giữa số lượng các bản ghi của tập hợp chứa X ∪ Y so với tổng các bản ghi chứa X.

Nhận thấy rằng tri thức đem lại bởi luật kết hợp dạng trên có sự khác biệt rất nhiều so với những thông tin thu được từ các câu lệnh truy vấn dữ liệu thông thường như SQL. Đó là những tri thức, những mối liên hệ chưa biết trước và mang tính dự báo đang tiềm ẩn trong dữ liệu. Những tri thức này không đơn giản là kết quả của phép nhóm, tính tổng hay sắp xếp mà là của một quá trình tính toán khá phức tạp.

2.4.2.3 Tập hợp thường xuyên

Tập hợp X được gọi là tập hợp thường xuyên (Frequent itemset) nếu có supp(X) ≥ minsupp, với minsup là ngưỡng độ hỗ trợ cho trước. Kí hiệu các tập này là FI [11].

2.4.2.3.1 Các tính chất của tập hợp thường xuyên

- Nếu A, B ⊆ I là hai tập hợp với A ⊆ B thì supp(A) ≥ supp(B)
  - Như vậy những bản ghi nào chưa tập hợp B thì cũng chứa tập hợp A
- Nếu A, B là hai tập hợp A,B ⊆ I nếu B là tập hợp thường xuyên và A⊆B thì A cũng là tập hợp thường xuyên.
  - Nếu B là tập hợp thường xuyên thì supp(B) □ minsup, mọi tập hợp A là con của tập hợp B đều là tập hợp thường xuyên trong cơ sở dữ liệu D vì supp(A) □ supp(B).
- Nếu A, B là hai tập hợp, A ⊆ B và A là tập hợp không thường xuyên thì B cũng là tập hợp không thường xuyên

2.4.2.4 Quy trình khai phá luật kết hợp

Khai phá luật kết hợp là công việc phát hiện ra (tìm ra, khám phá, phát hiện) các luật kết hợp thỏa mãn các ngưỡng độ hỗ trợ (δ) và ngưỡng độ tin cậy (α) cho trước. Bài toán khai phá luật kết hợp được chia thành hai bài toán nhỏ, hay như người ta thường nói, việc giải bài toán trải qua hai pha [11]:

- Pha 1: Tìm tất cả các tập phổ biến (tìm FI) trong CSDL T.
- Pha 2: Sử dụng tập FI tìm được ở pha 1 để sinh ra các luật tin cậy (interesting rules). Ý tưởng chung là nếu gọi ABCD và AB là các tập mục phổ biến, thì chúng ta có thể xác định luật AB →CD với tỷ lệ độ tin cậy:

$$conf = \frac{\text{supp(AB)}}{\text{supp(ABCD)}}$$

Nếu conf ≥ minconf thì luật được giữ lại (và thỏa mãn độ hỗ trợ tối thiểu vì ABCD là phổ biến).

Khi các mẫu phổ biến (frequent patterm) dài có từ 15 đến 20 items) thì tập FI, thậm chí cả tập FCI trở nên rất lớn và hầu hết các phương pháp truyền thống phải

đếm quá nhiều tập mục mới có thể thực hiện được. Các thuật toán dựa trên thuật toán Apriori – đếm tất cả 2 k tập con của mỗi k- itemsets mà chúng quét qua, và do đó không thích hợp với các itemsets dài được. Các phương pháp khác sử dụng "lookaheads" để giảm số lượng tập mục được đếm. Tuy nhiên, hầu hết các thuật toán này đều sử dụng tìm kiếm theo chiều rộng. Cách làm này hạn chế hiệu quả của lookaheads, vì các mẫu phổ biến dài hơn mà hữu ích vẫn chưa được tìm ra.

2.4.3 Thuật toán

2.4.3.1 Thuật toán cơ bản

Input: I, D, σ, a

Output: Các luật kết hợp thoả mãn ngưỡng độ hỗ trợ σ, ngưỡng độ tin cậy a.

Thuật toán [12]:

1. Tìm ra tất cả các tập hợp các tính chất có độ hỗ trợ không nhỏ hơn ngưỡng a.
2. Từ các tập mới tìm ra, tạo ra các luật kết hợp có độ tin cậy không nhỏ hơn a.

2.4.3.2 Thuật toán tìm luật kết hợp khi đã biết các tập hợp thường xuyên

Input: I, D, σ, a, S

Output: Các luật kết hợp thoả mãn ngưỡng độ hổ trợ σ ngưỡng độ tin cậy a.

Thuật toán [12]:

1. Lấy ra một tập xuất hiện σ – thường xuyên S $\in S$ và tập con $X \subseteq$ S
2. Xét luật kết hợp có dạng $X \rightarrow (S \cup X)$, đánh giá độ tin cậy của nó xem có nhỏ hơn a hay không.
   Tập hợp S đang xét đóng vai trò của tập hợp giao S = $X \cup V$ và do $X \cap (S - X) = \emptyset$ nên coi như Y = S – X.

2.4.4 Một số thuật toán phát hiện luật kết hợp

2.4.4.1 Thuật toán Apriori

Thuật toán dựa trên nhận xét là bất kỳ tập hợp con nào cũng có xuất hiện tập hợp σ thường xuyên cũng là tập xuất hiện σ-thường xuyên. Do đó, trong quá trình đi tìm các tập ứng cử viên, thuật toán chỉ cần dùng đến các tập ứng cử viên vừa xuất hiện ở bước ngay trước đó, chứ không cần dùng đến tất cả các tập ứng cử viên. Nhờ vậy mà bộ nhớ được giải phóng đáng kể [11].

- Bước 1: Cho trước ngưỡng độ hỗ trợ $0 \leq \sigma \leq 1$. Tìm tất cả các mặt hàng xuất hiện ở σ thường xuyên.

- Bước 2: Tiến hành ghép đôi các phần tử của L1, thu được tập C2, tập gọi là tập các ứng cứ viên có 2 phần tử. Sở dĩ gọi các tập này là "ứng cử viên" vì chưa chắc các tập này đã là σ thường xuyên. Sau khi kiểm tra, tiến hành lọc ra được các tập σ thường xuyên có 2 phần tử. Ký hiệu tập hợp này là L2.
- Bước 3: Với chủ ý đã nêu, tiến hành tìm các ứng cử viên có 3 phần từ (được lấy từ L1). Gọi đó là tập C3. Lưu ý nếu {A,B,C} muốn là "ứng cử viên" thì các tập 2 phần tử {A,B},{B,C} và {C,A} đều phải là σ thường xuyên, tức là chúng phải đều là phần tử của tập L2. Tiến hành kiểm tra các tập trong C3 và lọc ra được các tập hợp σ thường xuyên có 3 phần tử. Tập hợp này được ký hiệu là L3.
- Bước 4: Tiến hành tìm các ứng cử viên có n phần tử. Gọi tập của các ứng cử viên này là tập Cn và từ đây lọc ra Ln là tập hợp các tập σ thường xuyên có n phần tử.

Cốt lõi của thuật toán Apriori là hàm apriori_gen() do Agrawal đề nghị năm 1994. Hàm hoạt động theo 2 bước, bước 1- tập hợp Lk-1 tự kết nối join với chính nó để tạo ra tập ứng cử viên Ck. Sau đó hàm sẽ loại bỏ các tập hợp con (k-1) phần tử không nằm trong Lk-1.

2.4.4.2 Thuật toán Apriori nhị phân

Thuật toán Apriori nhị phân sử dụng các vector bit cho các thuộc tính, vector nhị phân n chiều ứng với n giao tác trong cơ sở dữ liệu. Có thể biểu diễn cơ sở dữ liệu bằng một ma trận nhị phân trong đó dòng thứ I tương dương với giao tác (bản ghi) i và cột thứ j ương ứng với mục ij [11].

2.4.4.3 Thuật toán Apriori-TID

Thuật toán Apriori là phần mở rộng tiếp theo hướng tiếp cận cơ bản của thuật toán Apiori. Thay vì dựa vào cơ sở dữ liệu thô thuật toán Apriori-TID biểu diễn bên trong một giao dịch bởi các ứng cử viên hiện hành.

Thuật toán Apriori cơ bản đòi hỏi phải quét toàn bộ cơ sở dữ liệu để tính độ hỗ trợ cho các tập hợp ứng cử viên ở mỗi bước dẫn đến lãng phí tài nguyên máy tính. Dựa trên tư tưởng ước đoán và đánh giá độ hỗ trợ Agrawal đề xuất cải tiến Apriori theo hướng chỉ phải quét cơ sở dữ liệu lần đầu tiên, sau đó tính độ hỗ trợ cho các tập hợp 1 phần tử. Từ bước thứ hai trở đi, thuật toán Apriori-TID nhờ lưu trữ song song cả ID giao dịch và các ứng cử viên, có thể đánh giá, ước lượng độ hỗ trợ mà không cần phải quét lại toàn bộ cơ sở dữ liệu [11].

Input: Tập các giao dịch D, minsup

Output: Tập Answer gồm các tập mục thường xuyên trên D

Method:

L1= {large 1 – itemset}

C1 = database D;

For (k=2; Lk-1 ≠ ∅; k++):

Ck;

For all entries t ∈Ck-1:

//Xác định các candidate itemset

//được chứa trong giao dịch với định danh t.TID

C1={c∈Ck|(c-c[k])∈t.set_of_itemset∧(c-c[k-1])∈t.set_of_itemset}

For all candidates c ∈Ct

c.count++

if (C1≠∅) then Ck = Ck <t.TID,Ct >

Luật kết hợp= {c ∈Ck | c.count ≥minsup}

Sự khác nhau giữa Apriori và AprioriTID là: cơ sở dữ liệu không được sử dụng để đếm các hỗ trợ sau lần đâu tiên quét qua cơ sở dữ liệu. Vì sau lần quét đầu tiên các l-itemsets sẽ được sinh ra, các L1 này sẽ được dùng để lọc ra các giao dịch của cơ sở dữ liệu bất kỳ item nào là không phổ biến và những giao dịch trong C_1chỉ chứa những item không phổ biến. Kết quả đó được đưa vào C_2và sử dụng những lần quét đó. Vì vậy kích thước của C_2 là khá nhỏ so với C_1.Sự giống nhau của hai thuật toán này là đều sử dụng bước cắt tỉa trong hàm Apriori_gen() [11].

2.4.4.4 Thuật toán Apriori-Hybrid

Là sự kết hợp giữa thuật toán Apriori và thuật toán Apriori-TID. Trong thuật toán Apriori-Hybrid, được sử dụng khi tổ chức lặp và chuyển sang Apriori-TID khi đã chắc chắn tập C_k đã vào bộ nhớ chính. Thuật toán Apriori-Hybrid còn được coi là tốt hơn so với Apriori và AprioriTID [11].

2.4.4.5 Ví dụ về thuật toán Apriori

Cho dữ liệu về các mặt hàng mà 5 khách hàng đã mua:

TID	Các mặt hàng đã mua
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p

5	a, f, c, e, l, p, m, n

Bảng 2-1 Bảng thông tin mua hàng

Mục tiêu của bài toán là tìm ra các luật kết hợp của các mặt hàng theo thói quen mua hàng của 5 khách hàng. Với độ hỗ trợ (min_support) là 60% và độ tin cậy tối thiểu (min_conf) là 80%.

Với số khách hàng là 5, em tính được tần suất xuất hiện tối thiểu của một mặt hàng phải là: min_sup count = Số đối tượng*min_support = 5*0.6 = 3.

2.4.4.5.1 Lập bảng ứng cử viên C_1

Với k=1 ta tìm các mẫu có chiều dài = 1 để lập nên bảng ứng viên C_1.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*s*
1	1	0	1	1	0	1	1	0	1	0	0	0	1	0	0	1	0
2	1	1	1	0	0	1	0	0	0	0	0	1	1	0	1	0	0
3	0	1	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0
4	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1
5	1	0	1	0	1	1	0	0	0	0	0	1	1	1	0	0	0
sum	3	3	4	1	1	4	1	1	1	1	1	2	3	1	2	3	1

Bảng 2-2 Bảng thống kê số lần xuất hiện của các mặt hàng

Mã hàng	Tần suất xuất hiện
a	3
b	3
c	4
d	1
e	1
f	4
g	1
h	1
i	1
j	1
k	1

l	2
m	3
n	1
o	2
p	3
s	1

Bảng 2-3 Bảng ứng viên C_1

Dựa vào tần suất xuất hiện tối thiểu min_sup count, em chọn ra các mặt hàng có tần suất xuất hiện ≥ min_sup count = 3 để tạo thành bảng thường xuyên L_1.

Bảng tập thường xuyên L_1:

Mã hàng	Tần suất xuất hiện
a	3
b	3
c	4
f	4
m	3
p	3

Bảng 2-4 Bảng tập thường xuyên L_1

2.4.4.5.2 Lập bảng ứng cử viện C_2

Dựa vào bảng tập thường xuyên L_1ta lập nên bảng ứng cử viên C_2 từ các lựa chọn có 2 mặt hàng.

Mã hàng		Tần suất xuất hiện
a	b	1
a	c	3
a	f	3
a	m	3
a	p	1
b	c	2
b	f	2

b	m	1
b	p	1
c	f	3
c	m	3
c	p	3
f	m	3
f	p	1
m	p	1

Bảng 2-5 Bảng ứng viên C_2

Dựa vào tần suất xuất hiện tối thiểu min_sup count, em lọc ra các cặp mặt hàng có tần suất xuất hiện ≥ min_sup count = 3 để lập nên bảng thường xuyên L_2.

Bảng tập thường xuyên L_2:

Mã hàng		Tần suất xuất hiện
a	c	3
a	f	3
a	m	3
c	f	3
c	m	3
c	p	3
f	m	3

Bảng 2-6 Bảng tập thường xuyên L_2

2.4.4.5.3 Lập bảng ứng cử viên C_3

Dựa vào bảng tập thường xuyên L_2 mà em lập nên bảng các ứng viên C_3 từ các lựa chọn có 3 mặt hàng.

Mã hàng			Tần suất xuất hiện
a	c	f	3
a	c	m	3
a	c	p	2

a	f	m	3
a	m	p	2
c	f	m	3
c	f	p	2
c	m	p	2
f	m	p	2

Bảng 2-7 Tập ứng viên C_3

Dựa vào tần suất xuất hiện tối thiểu min_sup count, em lọc ra các tổ hợp mã hàng có tần suất xuất hiện ≥ min_sup count = 3 để lập nên tập thường xuyên L_3.

Mã hàng			Tần suất xuất hiện
a	c	f	3
a	c	m	3
a	f	m	3
c	f	m	3

Bảng 2-8 Tập thường xuyên L_3

2.4.4.5.4 Lập bảng ứng cử viên C_4

Dựa vào bảng tập thường xuyên L_3 mà em lập nên bảng các ứng viên C_4từ các lựa chọn có 4 mặt hàng.

Mã hàng				Tần suất xuất hiện
a	c	f	m	3

Bảng 2-9 Tập ứng viên C_4

Dựa vào tần suất xuất hiện tối thiểu min_sup count, em lọc ra các tổ hợp mã hàng có tần suất xuất hiện ≥ min_sup count = 3. Vì tất cả các ứng viên C_4 đều thoả nên tập ứng viên C_4 là tập thường xuyên L_4.

Vì không thể lập nên kết hợp 5 mặt hàng nên không thể lập bảng các ứng viên C_5.

2.4.4.5.5 Áp dụng luật kết hợp

Tập phổ biến L = $L_1 \cup L_2 \cup L_3 \cup L_4$

L = {a, b, c, f, m, p, ac, af, am, cf, cm, cp,fm, acf, acm, amf, cfm, acfm}

Thu được 52 trường hợp sau:

(1).- Nếu a đã được mua thì c sẽ được mua. a → c

$$S(a,c) = \frac{3}{5} = 0.6 = 60\%$$

$$C(a,c) = \frac{S(a,c)}{S(a)} = \frac{0.6}{0.6} = 1 = 100\%$$

(2).- Nếu c đã được mua thì a sẽ được mua. c→a

$$S(c,a) = \frac{3}{5} = 0.6 = 60\%$$

$$C(c,a) = \frac{S(c,a)}{S(c)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(3).- Nếu a đã được mua thì f sẽ được mua. a→f

$$S(a,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(a,f) = \frac{S(a,f)}{S(a)} = \frac{0.6}{0.6} = 1 = 100\%$$

(4).- Nếu f đã được mua thì a sẽ được mua. f→a

$$S(f,a) = \frac{3}{5} = 0.6 = 60\%$$

$$C(f,a) = \frac{S(f,a)}{S(f)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(5).- Nếu a đã được mua thì m sẽ được mua. a→m

$$S(a,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(a,m) = \frac{S(a,m)}{S(a)} = \frac{0.6}{0.6} = 1 = 100\%$$

(6).- Nếu m đã được mua thì a sẽ được mua. m→a

$$S(m,a) = \frac{3}{5} = 0.6 = 60\%$$

$$C(m,a) = \frac{S(m,a)}{S(m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(7).- Nếu c đã được mua thì f sẽ được mua. c →f

$$S(c,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(c,f) = \frac{S(c,f)}{S(c)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(8).- Nếu f đã được mua thì c sẽ được mua. f→c

$$S(f,c) = \frac{3}{5} = 0.6 = 60\%$$

$$C(f,c) = \frac{S(f,c)}{S(f)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(9).- Nếu c đã được mua thì m sẽ được mua. c→m

$$S(c,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(c,m) = \frac{S(a,c)}{S(c)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(10).- Nếu m đã được mua thì c sẽ được mua. m→c

$$S(m,c) = \frac{3}{5} = 0.6 = 60\%$$

$$C(m,c) = \frac{S(m,c)}{S(m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(11).- Nếu c đã được mua thì p sẽ được mua. c→p

$$S(c,p) = \frac{3}{5} = 0.6 = 60\%$$

$$C(c,p) = \frac{S(c,p)}{S(c)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(12).- Nếu p đã được mua thì c sẽ được mua. p→c

$$S(p,c) = \frac{3}{5} = 0.6 = 60\%$$

$$C(p,c) = \frac{S(p,c)}{S(p)} = \frac{0.6}{0.6} = 1 = 100\%$$

(13).- Nếu f đã được mua thì m sẽ được mua. f→m

$$S(f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(f,m) = \frac{S(f,m)}{S(f)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(14).- Nếu m đã được mua thì f sẽ được mua. m→f

$$S(m,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(m,f) = \frac{S(m,f)}{S(m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(15).- Nếu a đã được mua thì c và f sẽ cùng được mua. a →{c,f}

$$S(a,c,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(a,\{c,f\}) = \frac{S(a,c,f)}{S(a)} = \frac{0.6}{0.6} = 1 = 100\%$$

(16).- Nếu c đã được mua thì a và f sẽ cùng được mua. c→{a,f}

$$S(c,a,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(c,\{a,f\}) = \frac{S(c,a,f)}{S(c)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(17).- Nếu f đã được mua thì c và a sẽ cùng được mua. f→{a,c}

$$S(a,c,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(f,\{a,c\}) = \frac{S(f,a,c)}{S(f)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(18).- Nếu a và c đã được mua cùng nhau thì f sẽ được mua. {a,c}→f

$$S(a,c,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,c\},f) = \frac{S(a,c,f)}{S(a,c)} = \frac{0.6}{0.6} = 1 = 100\%$$

(19).- Nếu a và f đã được mua cùng nhau thì c sẽ được mua. {a,f}→c

$$S(a,c,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,f\},c) = \frac{S(a,c,f)}{S(a,f)} = \frac{0.6}{0.6} = 1 = 100\%$$

(20).- Nếu c và f đã được mua cùng nhau thì a sẽ được mua. {f,c}→a

$$S(a,c,f) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{c,f\},a) = \frac{S(a,c,f)}{S(c,f)} = \frac{0.6}{0.6} = 1 = 100\%$$

(21).- Nếu a được mua thì c và m sẽ được mua cùng nhau. a →{c,m}

$$S(a,c,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(a,\{c,m\}) = \frac{S(a,c,m)}{S(a)} = \frac{0.6}{0.6} = 1 = 100\%$$

(22).- Nếu c được mua thì a và m sẽ được mua cùng nhau. c → {a,m}

$$S(a,c,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(c,\{a,m\}) = \frac{S(a,c,m)}{S(c)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(23).- Nếu m đã được mua thì a và c sẽ được mua cùng nhau. m → {a,c}

$$S(a,c,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(m,\{a,c\}) = \frac{S(a,c,m)}{S(m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(24).- Nếu cả a và c được mua thì m sẽ được mua. {a,c} → m

$$S(a,c,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,c\},m) = \frac{S(a,c,m)}{S(a,c)} = \frac{0.6}{0.6} = 1 = 100\%$$

(25).- Nếu cả a và m cùng được mua thì c sẽ được mua. {a,m}→c

$$S(a,c,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,m\},c) = \frac{S(a,c,m)}{S(a,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(26).- Nếu cả c và m cùng được mua thì a sẽ được mua. {c,m}→a

$$S(a,c,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{c,m\},a) = \frac{S(a,c,m)}{S(c,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(27).- Nếu a được mua thì f và m sẽ được mua cùng nhau. a →{f,m}

$$S(a,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(a,\{f,m\}) = \frac{S(a,f,m)}{S(a)} = \frac{0.6}{0.6} = 1 = 100\%$$

(28).- Nếu f được mua thì a và m sẽ được mua cùng nhau. f →{a,m}

$$S(a,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(f,\{a,m\}) = \frac{S(a,f,m)}{S(f)} = \frac{0.6}{0.8}0.75 = 75\%$$

(29).- Nếu m được mua thì a và f sẽ được mua cùng nhau. m → {a,f}

$$S(a,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(m,\{a,f\}) = \frac{S(a,f,m)}{S(m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(30).- Nếu cả a và m cùng được mua cùng nhau thì f sẽ được mua. {a,m} →f

$$S(a,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,m\},f) = \frac{S(a,f,m)}{S(a,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(31).- Nếu cả a và f cùng được mua cùng nhau thì m sẽ được mua. {a,f} → m

$$S(a,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,f\},m) = \frac{S(a,f,m)}{S(a,f)} = \frac{0.6}{0.6} = 1 = 100\%$$

(32).- Nếu cả m và f cùng được mua cùng nhau thì a sẽ được mua. {m,f} →a

$$S(a,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{m,f\},a) = \frac{S(a,f,m)}{S(m,f)} = \frac{0.6}{0.6} = 1 = 100\%$$

(33).- Nếu c được mua thì f và m sẽ được mua cùng nhau. c →{f,m}

$$S(c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(c,\{f,m\}) = \frac{S(c,f,m)}{S(c)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(34).- Nếu f được mua thì c và m sẽ được mua cùng nhau. f →{c,m}

$$S(c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(f,\{c,m\}) = \frac{S(c,f,m)}{S(f)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(35).- Nếu m được mua thì c và f sẽ được mua cùng nhau. m → {c,f}

$$S(c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(m,\{c,f\}) = \frac{S(c,f,m)}{S(m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(36).- Nếu c và f được mua cùng nhau thì m sẽ được mua. {c,f} → m

$$S(c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{c,f\},m\}) = \frac{S(c,f,m)}{S(c,f)} = \frac{0.6}{0.6} = 1 = 100\%$$

(37).- Nếu c và m được mua cùng nhau thì f sẽ được mua. {c,m} → f

$$S(c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{c,m\},f\}) = \frac{S(c,f,m)}{S(c,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(38).- Nếu m và f được mua cùng nhau thì c sẽ được mua. {f,m} → c

$$S(c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{f,m\},c\}) = \frac{S(c,f,m)}{S(f,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(39).- Nếu a được mua thì cả c, f và m đều được mua cùng nhau. a → {c,f,m}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(a,\{c,f,m\}\}) = \frac{S(a,c,f,m)}{S(a)} = \frac{0.6}{0.6} = 1 = 100\%$$

(40).- Nếu c được mua thì cả a, f và m đều được mua cùng nhau. c →{a,f,m}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(c,\{a,f,m\}\}) = \frac{S(a,c,f,m)}{S(c)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(41).- Nếu f được mua thì cả a, c và m đều được mua cùng nhau. f→ {a,c,m}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(f,\{a,c,m\}\}) = \frac{S(a,c,f,m)}{S(f)} = \frac{0.6}{0.8} = 0.75 = 75\%$$

(42).- Nếu m được mua thì a, c và f đều được mua cùng nhau. m → {a,c,f}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(m,\{a,c,f\}\}) = \frac{S(a,c,f,m)}{S(m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(43).- Nếu a và c đều được mua cùng nhau thì m và f cũng được mua cùng. {a,c}→{f,m}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,c\},\{f,m\}\}) = \frac{S(a,c,f,m)}{S(a,c)} = \frac{0.6}{0.6} = 1 = 100\%$$

(44).- Nếu a và f đều được mua cùng nhau thì c và m cũng được mua cùng. {a,f}→{c,m}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,f\},\{c,m\}\}) = \frac{S(a,c,f,m)}{S(a,f)} = \frac{0.6}{0.6} = 1 = 100\%$$

(45).- Nếu a và m đều được mua cùng nhau thì c và f cũng được mua cùng. {a,m}→{c,f}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,m\},\{c,f\}\}) = \frac{S(a,c,f,m)}{S(a,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(46).- Nếu c và f được mua cùng nhau thì a và m sẽ được mau cùng. {c,f} →{a,m}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{c,f\},\{a,m\}\}) = \frac{S(a,c,f,m)}{S(c,f)} = \frac{0.6}{0.6} = 1 = 100\%$$

(47).- Nếu c và m được mua cùng nhau thì a và f sẽ được mua cùng. {c,m}→{a,f}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{c,m\},\{a,f\}\}) = \frac{S(a,c,f,m)}{S(c,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(48).- Nếu f và m được mua cùng nhau thì a và c sẽ được mua cùng. {f,m} → {a,c}

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{f,m\},\{a,c\}\}) = \frac{S(a,c,f,m)}{S(f,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(49).- Nếu cả a,c, f đều cùng được mua thì m sẽ được mua. {a,c,f}→m

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,c,f\},m) = \frac{S(a,c,f,m)}{S(a,c,f)} = \frac{0.6}{0.6} = 1 = 100\%$$

(50).- Nếu cả a,c,m đều cùng được mua thì f sẽ được mua. {a,c,m}→f

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,c,m\},f) = \frac{S(a,c,f,m)}{S(a,c,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(51).- Nếu cả a, f, m đều cùng được mua thì c sẽ được mua. {a,f,m} → c

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{a,f,m\},c) = \frac{S(a,c,f,m)}{S(a,f,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

(52).- Nếu cả c, f, m đều mua cùng nhau thì a sẽ được mua. {c,f,m}→a

$$S(a,c,f,m) = \frac{3}{5} = 0.6 = 60\%$$

$$C(\{c,f,m\},a) = \frac{S(a,c,f,m)}{S(c,f,m)} = \frac{0.6}{0.6} = 1 = 100\%$$

Từ các trường hợp trên, em lập bảng luật kết hợp

STT	Tập phổ biến	Sup	Các tập phổ biến con	Các luật có thể có	Các luật thoả min_conf(0.8)
1	a	3			
2	b	3			
3	c	4			
4	f	4			
5	m	3			
6	p	3			

7	cf	3	c,f	c→f, f→c	
8	af	3	a,f	a→f, f→a	
9	fm	3	f,m	f→m, m→f	
10	ac	3	a,c	a→c, c→a	
11	cm	3	c,m	c→m, m→c	
12	cp	3	c,p	c→p, p→c	p→c(1)
13	am	3	a,m	a→m, m→a	
14	acf	3	ac,af,cf	a→{c,f}, c→{a,f}, f→{a,c}, {a,c}→ f, {a,f} → c, {c,f} → a	
15	cfm	3	cf,cm,fm	c →{f,m}, f→{c,m}, m→{c,f}, {c,f} → m, {c,m} → f, {f,m} →c	
16	afm	3	af,am,fm	a →{f,m}, f→ {a,m}, m→ {a,f}, {a,f}→ m, {a,m}→ f, {f,m}→ a	
17	acm	3	ac,am,cm	a →{c,m}, c→{a,m}, m→{a,c}, {a,c}→m, {a,m}→c, {c,m} →a	
18	acfm	3	ac, af, am, cf, cm, fm,	a→{c,f,m}, c→{a,f,m}, f→{a,c,m},	a→{c,f,m}(1), m→{a,c,f}(1), {a,c}→{f,m}(1),

			acf, acm, afm, cfm	m→{a,c,f}, {a,c}→{f,m}, {a,f}→{c,m}, {a,m}→{c,f}, {c,f}→{a,m}, {c,m}→{a,f}, {f,m}→{a,c}, {a,c,f}→m, {a,c,m}→f, {a,f,m}→c, {c,f,m}→a	{a,f}→{c,m}(1), {a,m}→{c,f}(1), {c,f}→{a,m}(1), {c,m}→{a,f}(1), {f,m}→{a,c}(1), {a,c,f}→m(1), {a,c,m}→f(1), {a,f,m}→c(1), {c,f,m}→a(1)

Bảng 2-10 Bảng luật kết hợp của 4 mặt hàng

2.5 Ưu điểm và nhược điểm của khai phá dữ liệu

2.5.1 Ưu điểm

- Đưa ra các quyết định chính xác hơn: Bằng cách trích xuất thông tin hữu ích từ các tập dữ liệu lớn để tìm hiểu về xu hướng và mối liên hệ của chúng mà các tổ chức có thể đưa ra các xác định được xu hướng và đưa ra các dự đoán chính xác hơn.
- Tăng cường hiệu quả marketing: Bằng cách phân tích dữ liệu khách hàng mà doanh nghiệp có thể xác định được xu hướng và thói quen của khách hàng để từ đó có thể tạo ra các chiến dịch quảng cáo và các sản phẩm cũng như dịch vụ được cá nhân hoá theo từng tập khách hàng.
- Tăng cường hiệu suất: Bằng cách xác định điểm kém hiệu quả và những vùng cần cả thiện mà khai thác dữ liệu có thể xác định những điểm tắc nghẽn và đưa ra các giải pháp để cải thiện hiệu suất và cắt giảm chi phí.
- Phát hiện lừa đảo/gian lận: Bằng cách xác định các mẫu và mối quan hệ của dữ liệu doanh nghiệp có thể xác định các hoạt động đáng ngờ và ngăn chặn các trường hợp lừa đảo/ gian lận.
- Tăng cường khả năng duy trì khách hàng: Bằng cách phân tích dữ liệu khách hàng, doanh nghiệp có thể xác định các yếu tố làm cho khách hàng của họ rời đi và đưa ra các phương án cải thiện các yếu tố đấy.
- Lợi thế cạnh tranh: Bằng cách phân tích hành vi khách hàng, xu hướng thị trường và hoạt động của đối thủ mà doanh nghiệp có thể xác định được thời cơ để đổi mới hoặc làm khác với đối thủ.

2.5.2 Nhược điểm

- Chất lượng dữ liệu: Phương pháp có yêu cầu cao về chất lượng dữ liệu được sử dụng cho phân tích. Nếu dữ liệu bị thiếu, không chính xác hoặc không đồng nhất sẽ có thể dẫn đến kết quả phân tích không đáng tin cậy.
- Dữ liệu riêng tư và tính bảo mật: Vì sử dụng một lượng lớn dữ liệu mà trong đó có thể bao gồm các thông tin nhạy cảm về cá nhân hay tổ chức. Và nếu các tập dữ liệu này bị rò rĩ ra ngoài thì chúng có thể được dùng cho các mục đích xấu như trộm danh tính hay gián điệp công ty.
- Độ phức tạp cao: Vì phương pháp yêu cầu chuyên môn ở nhiều lĩnh vực, bao gồm số liệu thống kê, khoa học máy tính và kiến thức lĩnh vực. Các vấn đề phức tạp này có thể là rào cản đối với một số doanh nghiệp hoặc tổ chức.
- Chi phí: Khai phá dữ liệu rất tốn kém nhất là khi một tập dữ liệu lớn cần được phân tích.
- Khó khăn trong việc diễn giải kết quả: Vì các thuật toán của phương pháp sẽ tạo ra một lượng lớn dữ liệu mà chúng có thể khó để diễn giải. Đây có thể là thử thách đối với các doanh nghiệp khi họ muốn xác định ý nghĩa của các kết quả phân tích.
- Phụ thuộc vào thiết bị: Khác với các phương pháp khác có thể dùng nhân lực để phân tích thì khai phá dữ liệu phụ thuộc nặng nề vào thiết bị. Nếu như các thiết bị gặp trục trặc như lỗi phần cứng hoặc phần mềm thì rất có thể dẫn đến mất dữ liệu hoặc dữ liệu bị hư hại.

CHƯƠNG 3.
PYTHON TRONG PHÂN TÍCH DỮ LIỆU

3.1 Giới thiệu

Python là một ngôn ngữ lập trình được sử dụng rộng rãi trong các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (Machine Learning). Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Python là ngôn ngữ mã nguồn mở, được tải xuống miễn phí, tích hợp tốt với tất cả các loại hệ thống và tăng tốc độ phát triển.

3.2 Lý do để sử dụng Python cho việc phân tích dữ liệu

- Python có thể được sử dụng trên nhiều nền tảng khác nhau và trên hầu hết các thiết bị.
- Python hỗ trợ cực mạnh việc phân tích dữ liệu với vô số các thư viện sẵn sàng hỗ trợ nhu cầu từ phân tích dữ liệu, tự động hoá,…
- Python là một mã nguồn mở vì thế nên nó miễn phí và các nhà phân tích dữ liệu không cần phải lo đến vấn đề bản quyền.
- Khả năng tự động hoá cao. Không cần phải thực hiện thao tác thủ công mà nhà phân tích dữ liệu hoàn toàn chỉ cần viết code và để nó thực hiện các tác vụ thay mình.
- Làm việc được với hầu hết các định dạng dữ liệu, file với tốc độ nhanh. Nếu đọc các file CSV lên đến cả GB thì Power PI hay Excel sẽ gặp tình trạng đứng máy hoặc tải rất lâu để hiển thị toàn bộ dữ liệu. Với Python, các nhà phân tích dữ liệu hoàn toàn có thể tải dữ liệu vào dataframe và xem trước với thư viện Pandas cực kỳ nhanh chóng.

3.3 Các thư viện hỗ trợ trong python

3.3.1 Pandas

3.3.1.1 Giới thiệu

Pandas là một thư viện Python cung cấp các cấu trúc dữ liệu nhanh, mạnh mẽ và linh hoạt. Pandas được thiết kế để làm việc dễ dàng và trực quan với dữ liệu có cấu trúc (dạng bảng, đa chiều, không đồng nhất) và dữ liệu chuỗi thời gian [13].

Pandas phù hợp với nhiều dạng dữ liệu khác nhau:

- Dữ liệu dạng bảng với các cột không đồng nhất, như các bảng SQL hoặc Excel.
- Dữ liệu chuỗi thời gian theo thứ tự và không có thứ tự

- Dữ liệu ma trận với các hàng và cột. Bất kỳ các dạng hình thức khác nhau của các bộ dữ liệu quan sát / thống kê.
- Dữ liệu thực sự không cần phải dán nhãn vào cấu trúc dữ liệu pandas. Pandas được xây dựng dựa trên Numpy.
- Hai cấu trúc dữ liệu chính của pandas là Series đối với dữ liệu một chiều và Dataframe đối với dữ liệu 2 chiều. Hai dạng cấu trúc dữ liệu có thể xử lý được phần lớn các trường hợp điển hình trong tài chính, thống kê, khoa học,…

3.3.1.2 Các hàm quan trọng trong thư viện Pandas

Các hàm quan trọng [13]:

Tên hàm	*Công dụng*
Read_csv()	Là một trong những hàm quan trọng nhất của Pandas. Hàm này cho phép đọc các tập dữ liệu csv thành các Dataframe tách biệt nhau bởi dấu phẩy.
Head()	Hàm trả về số dòng đầu tiên của tập dữ liệu. Hàm này cho phép xem trước các dòng của tập dữ liệu (thấy được các cột và hàng của tập dữ liệu).
Info()	Hàm cho biết thông tin của tập dữ liệu như số dòng, số cột, loại dữ liệu của từng cột và số dòng dữ liệu có giá trị khác null.
Describe()	Giúp nắm được các thông tin cơ bản của DataFrame. Max, min, … của từng cột.
Loc[]	Là lệnh giúp lấy ra được các hàng tương ứng khi thỏa mãn được điều kiện và người dùng có thể chỉ định cột trả về.
Iloc[]	Là hàm sẽ giúp lấy ra các hàng trùng với số thứ tự được chỉ định. Cũng có thể chỉ định các cột trả về tương tự với hàm loc[].
Groupby()	Là hàm sẽ giúp nhóm các hàng có theo giá trị chỉ định
Sort_value()	Là hàm giúp sắp xếp lại các cột theo thứ tự tăng hoặc giảm
Fillna()	Đây là hàm để thay thế các giá trị null với giá trị do người dùng quy định.

Bảng 3-1 Bảng các hàm quan trọng trong Pandas

3.3.2 Numpy

3.3.2.1 Giới thiệu

Đó là một thư viện Python cung cấp một đối tượng mảng nhiều chiều, nhiều đối tượng dẫn xuất khác nhau (chẳng hạn như mảng và ma trận bị che khuất) và một

loạt các thói quen cho các thao tác nhanh trên mảng, bao gồm thao tác toán học, logic, hình dạng, sắp xếp, chọn, I/O, biến đổi Fourier rời rạc, đại số tuyến tính cơ bản, phép toán thống kê cơ bản, mô phỏng ngẫu nhiên, v.v. [14].

Cốt lõi của Numpy là các mảng đa chiều. Điều này đóng gói các mảng n chiều của các kiểu dữ liệu đồng nhất, với nhiều thao tác được thực hiện trong mã được biên dịch để thực hiện. Có một số khác biệt quan trọng giữa mảng NumPy và chuỗi Python tiêu chuẩn:

- Mảng của Numpy có kích thước được quy định trước, khác với mảng của Python có thể tăng một cách linh hoạt. Thay đổi kích thước của mảng có thể dẫn đến việc tạo ra mảng mới và xoá đi mảng ban đầu.
- Các yếu tố trong Numpy được yêu cầu phải cùng kiểu dữ liệu và do đó sẽ cùng kích thước.
- Các mảng NumPy tạo điều kiện thuận lợi cho toán học nâng cao và các loại hoạt động khác trên số lượng lớn dữ liệu. Thông thường, các thao tác như vậy được thực thi hiệu quả hơn và sử dụng ít code hơn so với khả năng sử dụng trình tự dựng sẵn của Python.
    - Ngày càng có nhiều package khoa học và toán học dựa trên Python đang sử dụng mảng NumPy; mặc dù chúng thường hỗ trợ đầu vào chuỗi Python, nhưng chúng chuyển đổi đầu vào đó thành mảng NumPy trước khi xử lý và chúng thường xuất ra mảng NumPy. Nói cách khác, để sử dụng hiệu quả nhiều (thậm chí là hầu hết) phần mềm khoa học/toán học dựa trên Python ngày nay, chỉ biết cách sử dụng các kiểu trình tự dựng sẵn của Python là chưa đủ - người ta cũng cần biết cách sử dụng mảng NumPy.

3.3.2.2 Các loại dữ liệu trong Numpy

Các loại dữ liệu trong Numpy [14] :

STT	Loại dữ liệu	Mô tả
1	Bool	Boolean(True hoặc False) được lưu trữ dưới dạng type
2	Int	Kiểu số nguyên mặc định, tương tự với kiểu long của C (thường có dạng là int32 hoặc int 64)
3	Intc	Giống hệt với int C (thường là int32 hoặc int64)
4	Intp	Số nguyên được sử dụng để lập chỉ mục (giống như C ssize_t; thông thường là int32 hoặc int64)
5	Int8	Byte (-128 đến 127)
6	Int16	Số nguyên (-32768 đến 32767)

7	Int32	Số nguyên (-2147483648 đến 2147483647)
8	Int64	Số nguyên (-9223372036854775808 đến 9223372036854775807)
9	Uint8	Số nguyên không dấu (0 đến 255)
10	Uint16	Số nguyên không dấu (0 đến 65535)
11	Uint32	Số nguyên không dấu (0 đến 4294967295)
12	Uint64	Số nguyên không dấu (0 đến 18446744073709551615)
13	Float	Viết tắt cho float64
14	Float16	float: bit dấu, số mũ 5 bit, phần định trị 10 bit
15	Float32	float: bit dấu, số mũ 8 bit, phần định trị 23 bit
16	Float64	float: bit dấu, số mũ 11 bit, phần định trị 52 bit
17	Complex	Viết tắt cho complex128
18	Complex64	Số phức, được biểu diễn bằng hai số thực 32 bit (thành phần thực và ảo)
19	Complex128	Số phức, được biểu thị bằng hai số thực 64 bit (thành phần thực và ảo)

Bảng 3-2 Bảng các loại dữ liệu của numpy

3.3.2.3 Các hàm quan trọng trong Numpy

Các hàm quan trọng trong Numpy [14]:

Tên hàm	*Công dụng*
min	Tìm giá trị nhỏ nhất trong mảng Numpy
max	Tìm giá trị lớn nhất trong mảng Numpy
mean	Tìm giá trị trung bình của mảng Numpy
std	Tìm độ lệch chuẩn của giá trị trong mảng Numpy
meadian	Tìm giá trị trung vị của mảng Numpy
percentile	Tìm phân vị của giá trị trong mảng Numpy
linspace	Tìm các số cách đều nhau trong một khoảng xác định
shape	Xác định hình dạng của mảng Numpy
reshape	Thay đổi hình dạng của mảng Numpy

copyto	Sao chép các giá trị trong một mảng tới 1 mảng khác.
transpose	Đảo ngược các trục của một mảng
stack	Hợp mảng theo 1 trục
vstack	Hợp mảng theo trục dọc
hstack	Hợp mảng theo trục ngang
sort	Sắp xếp các giá trị trong mảng

Bảng 3-3 Bảng các hàm quan trọng trong Numpy

3.3.3 Matplotlib

3.3.3.1 Giới thiệu

Matplotlib là thư viện vẽ đồ thị rất mạnh mẽ hữu ích cho những người làm phân tích dữ liệu. Module mà được sử dụng nhiều nhất trong Matplotlib là module Pyplot [15].

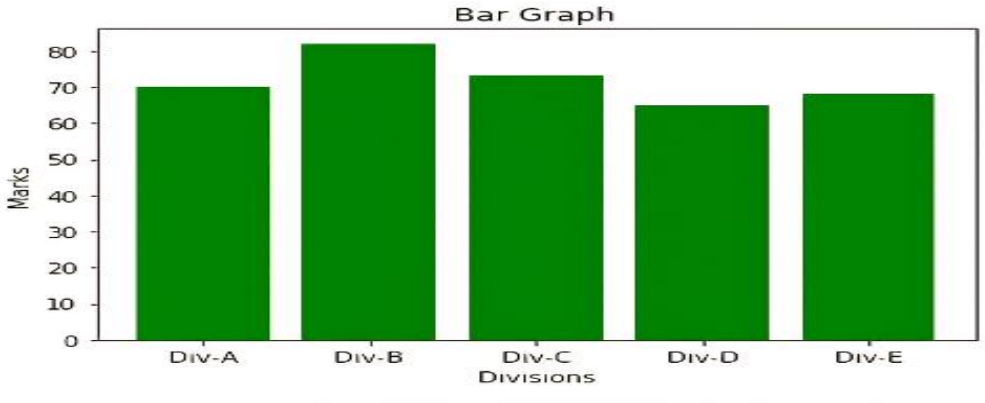
Module pyplot cung cấp cho người dùng nhiều công cụ để mô hình hoá dữ liệu do đa số các dữ liệu mà module xử đều là ở dạng mảng. Ngoài ra, pyplot còn cung cấp cho người dùng khả năng tạo ra các subplot trong cùng một hình.

Một Matplotlib figure có thể phân ra làm nhiều phần như sau:

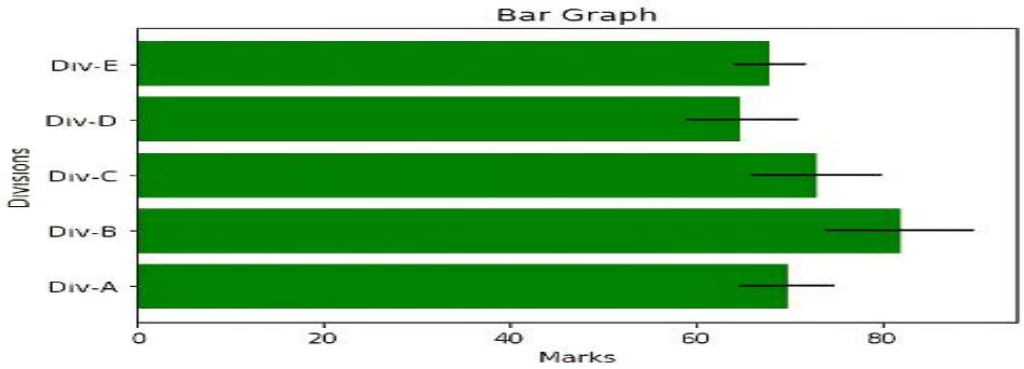
- Figure: Như một cái cửa sổ chứa tất cả những gì được vẽ bên trong đó.
- Axes: Thành phần chính của một figure là các axes.Một figure có thể có một hoặc nhiều axes. Nói cách khác, figure chỉ là khung chứa, chính các axes mới thật sự là nơi các hình vẽ lên.
- Axis: Chúng là dòng số giống như các đối tượng và đảm nhiệm việc tạo giới hạn biểu đồ.
- Artist: Mọi thứ có thể thấy trên figure là một artist như Text object, Line2D object, collection object.Hầu hế các Artist được gắn kết với Axes.

3.3.3.2 Các loại biểu đồ có trong matplotlib

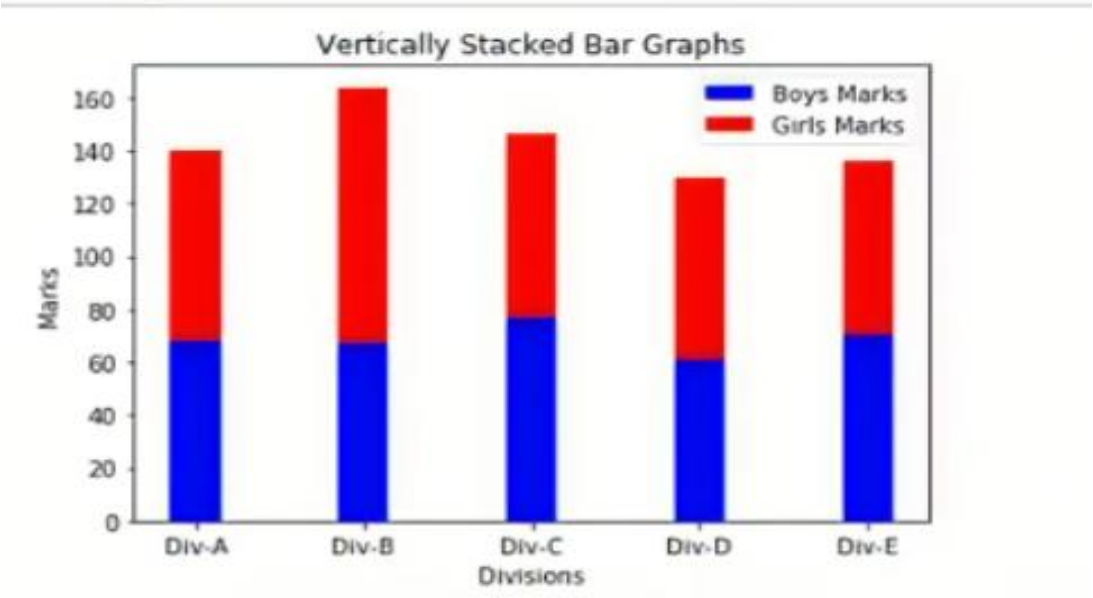
- Biểu đồ cột: Đây là dạng biểu đồ phổ biến nhất để hiển thị dữ liệu và phân loại các biến.



Hình 3-1 Biểu độ cột dọc trong pyplot

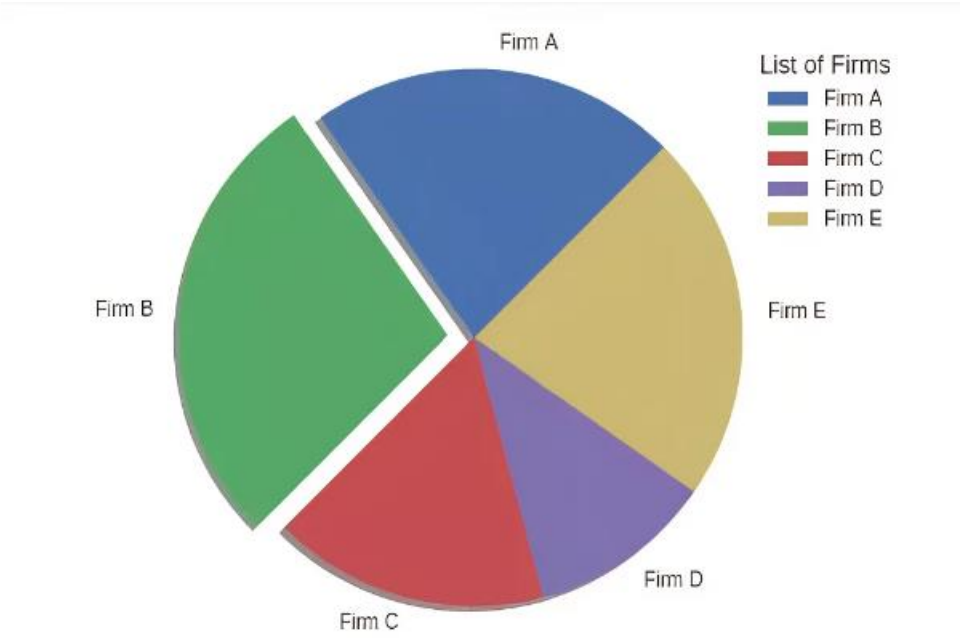


Hình 3-2 Biểu đồ cột ngang trong pyplot



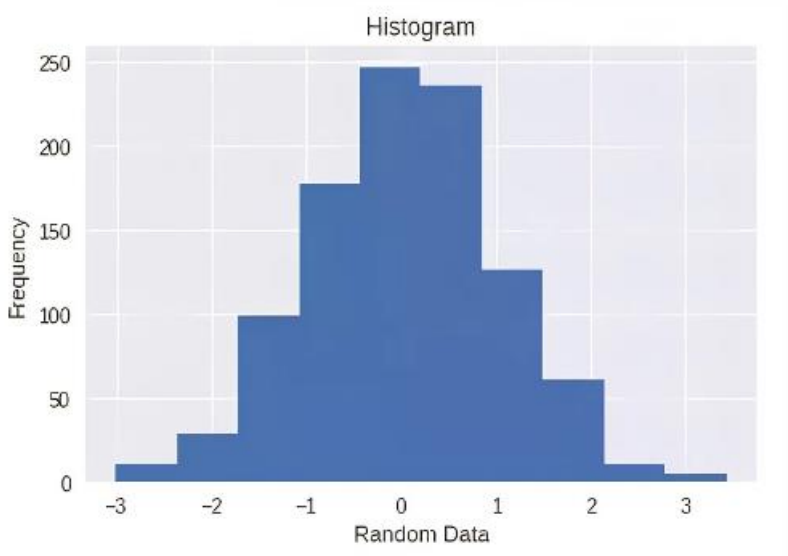
Hình 3-3 Biểu đồ cột xếp chồng trong pyplot

- Biểu đồ tròn: Biểu đồ tròn là một dạng biểu đồ cơ bản trong pyplot. Các nhà phân tích dữ liệu có thể chuyển đổi các thông số để tuỳ chỉnh biểu đồ theo ý em muốn.



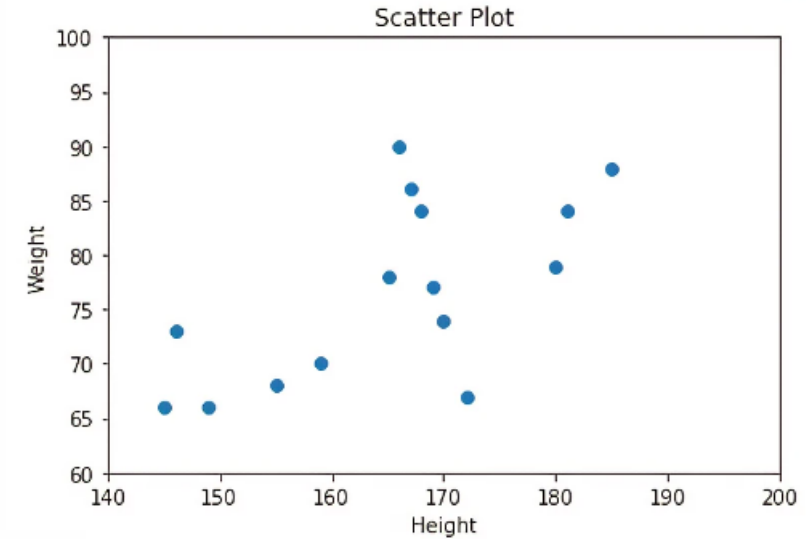
Hình 3-4 Biểu đồ tròn trong pyplot

- Histogram: Là biểu đồ phổ biến khi biểu đồ cho phép các nhà phân tích dữ liệu xem xét dữ liệu như chiều cao, cân nặng, giá cả, … Histogram's data được vẽ trong phạm vi so với tần số của nó. Là một trong những dạng biểu đồ phổ biến nhất trong xác xuất và thống kê.



Hình 3-5 Biểu đồ histogram trong pyplot

- Biểu đồ phân tán: Là dạng biểu đồ phổ biến trong việc hình dung một vấn đề về hồi quy.



Hình 3-6 Biểu đồ phân tán trong pyplot

3.3.4 Mlxtend

3.3.4.1 Giới thiệu

Mlxtend là thư viện máy học cung cấp các công cụ hữu ích cho các tác vụ khoa học dữ liệu. Thư viện được thiết kế bởi Sebastian Raschka, một chuyên viên phân tích thống kê tại đại học Wisconsin-Madison [16].

3.3.4.2 Các module hỗ trợ máy học

3.3.4.2.1 Classifier

Trong module Classifier hỗ trợ các thuật toán [16]:

- Adaline
- EnsembleVoteClassifier
- LogisticRegression
- MultilayerPerceptron
- OneRClassifier
- Perceptron
- SoftmaxRegression
- StackingClassifier
- StackingCVClassifier

3.3.4.2.2 Phân cụm (Cluster)

Trong module Cluster hỗ trợ thuật toán K-means [16].

3.3.4.2.3 Data

Trong module Data, một số hàm hỗ trợ xử lý dữ liệu [16]:

- autompg_data

- boston_housing_data
- iris_data
- loadlocal_mnist
- make_multiplexer_dataset
- mnist_data
- three_blobs_data
- wine_data

3.3.4.2.4 Evaluate

Trong module Evaluate, một số hàm được sử dụng để đánh giá [16]:

- accuracy_score
- bias_variance_decomp
- bootstrap
- bootstrap_point632_score
- BootstrapOutOfBag
- cochrans_q
- combined_ftest_5x2cv
- confusion_matrix
- create_counterfactual
- feature_importance_permutation
- ftest
- GroupTimeSeriesSplit
- lift_score
- mcnemar_table
- mcnemar_tables
- mcnemar
- paired_ttest_5x2cv
- paired_ttest_kfold_cv
- paired_ttest_resampled
- permutation_test
- PredefinedHoldoutSplit
- proportion_difference
- RandomHoldoutSplit
- scoring

3.3.4.2.5 Feature_extraction

Trong module feature_extraction, một số hàm hỗ trợ việc phân tích trích lọc yếu tố [16]:

- LinearDiscriminantAnalysis

- PrincipalComponentAnalysis
- RBFKernelPCA

3.3.4.2.6 Feature_selection

Trong module feature_selection, một số hàm hỗ trợ việc phân tích chọn lọc yếu tố [16]:

- ColumnSelector
- ExhaustiveFeatureSelector
- SequentialFeatureSelector

3.3.4.2.7 File_io

Trong module file_io, một số hàm hỗ trợ việc xử lý file [16]:

- find_filegroups
- find_files

3.3.4.2.8 Frequent_patterns

Trong module frequent_patterns, một số hàm hỗ trợ việc phân tích các tập phổ biến:

- Apriori [17]:
  - Hàm Apriori được sử dụng để tạo ra tập phổ biến của các items có trong tập dữ liệu. Hàm sử dụng thông số min_support để đánh giá việc tập dữ liệu có phải là tập phổ biến. Ví dụ như nếu min_support = 0.5(50%) thì để một tập items được xét là tập phổ biến thì tập đấy phải xuất hiện ít nhất trong 50% số giao dịch.
  - Hàm apriori có dạng:

apriori("tập dữ liệu", min_support = …, use_colnames = True/False)

- Association_rules [18]:
  - Luật kết hợp (Association_rules) là công cụ phổ biến cho việc khai phá các tập thường xuyên.Một luật phổ biến có dạng X→Y. Luật kết hợp sử dụng các hệ số để đánh giá và lựa chọn ra các luật phù hợp.
  - Các hệ số bao gồm:
    - Support: Là hệ số đánh giá độ phong phú hoặc độ phổ biến của tập items trong các giao dịch.

$$support(A \rightarrow C) = support(A \cup C)$$

    - Confidence: Là hệ số xác xuất xảy ra hệ quả trong giao dịch khi có tiền đề. Độ tin cậy không có tính đối xứng vì A→C có tỷ lệ xảy ra khác so với C→A

$$confidence(A \rightarrow C) = \frac{support(A \rightarrow C)}{support(A)}$$

- Lift: Là hệ số thường được dùng để đo mức độ xuất hiện thường xuyên của tiền đề và hệ quả so với việc hệ quả xảy ra một cách độc lập.

$$lift(A \rightarrow C) = \frac{confidence(A \rightarrow C)}{support(C)}$$

- Leverage: Là sự khác biệt giữa tần xuất quan sát được của A và C xuất hiện cùng nhau và tần suất có thể xảy ra nếu A và C độc lập.

$$levarage(A \rightarrow C) = support(A \rightarrow C) - support(A) X support(C)$$

- Conviction: Là sự phụ thuộc của hệ quả vào tiền đề. Trong trường hợp độ tin cậy là hoàn hảo (100%) tử số trở thành 0 (do 1-1) và điểm thuyết phục là "inf".

$$conviction(A \rightarrow C) = \frac{1 - support(C)}{1 - confidence(A \rightarrow C)}$$

- Zhang_metrics: Là thang đó kết hợp hoặc phân ly. Giá trị nằm trong khoảng từ -1 đến 1. Giá trị dương thì xác định là Kết hợp, ngược lại nếu âm thì là phân ly.

$$zhangs\ metric(A \rightarrow C) = \frac{confidence(A \rightarrow C) - confidence(A' \rightarrow C)}{Max[confidence(A \rightarrow C), confidence(A' \rightarrow C)]}$$

- Hàm association_rules có dạng:

association_rules("Tập dữ liệu thường xuyên","Hệ số = ","Ngưỡng tối thiểu = ")

- fpgrowth
- fpmax

3.3.4.2.9 Image

Image là module cung cấp hàm trích xuất điểm ảnh của gương mặt thành mảng 2 chiều [16].

3.3.4.2.10 Math

Math là module cung cấp các hàm tính toán kết hợp hoặc hoán vị [16]:

- num_combinations
- num_permutations

3.3.4.2.11 Plotting

Plotting là module cung cấp các hàm để phát hoạ dữ liệu thành các đồ thị minh hoạ [16]:

- category_scatter
- checkerboard_plot
- ecdf
- enrichment_plot
- heatmap
- plot_confusion_matrix
- plot_pca_correlation_graph
- plot_decision_regions
- plot_learning_curves
- plot_linear_regression
- plot_sequential_feature_selection
- scatterplotmatrix
- scatter_hist
- stacked_barplot

3.3.4.2.12 Preprocessing

Preprocessing là module cung cấp các hàm xử lý dữ liệu trước khi tập dữ liệu được phân tích [16]:

- CopyTransformer
- DenseTransformer
- MeanCenterer
- minmax_scaling
- one-hot_encoding
- shuffle_arrays_unison
- standardize
- TransactionEncoder

  TransactionEncoder là hàm thường xuyên được dùng để chuyển tập dữ liệu thành một mảng phù hợp với thuật toán máy học. Bằng cách học các nhãn độc nhất (unique labels) trong tập dữ liệu sau đó là chuyển tập dữ liệu đầu vào thành một chuỗi mã hoá Boolean của Numpy.

3.3.4.2.13 Regressor

Regressor là module cung cấp các hàm phân tích hồi quy [16].

- LinearRegression
- StackingCVRegressor
- StackingRegressor

3.3.4.2.14 Text

Text là module cung cấp các hàm liên quan đến xử lý hoặc chuyển đổi ký tự [16].

- generalize_names
- generalize_names_duplcheck
- tokenizer

CHƯƠNG 4.
ÁP DỤNG THUẬT TOÁN APRIORI ĐỂ CHẨN ĐOÁN COVID-19

4.1 Mục tiêu

Mục tiêu của nghiên cứu là xác định được 3 triệu chứng phổ biến để chẩn đoán các ca mắc covid-19 trong các bệnh nhân đang được quan sát tại đại học Irvine California.

Với hai tiêu chí để đánh giá kết quả chẩn đoán:

- Xác định các luật kết hợp bệnh có tỷ lệ xuất hiện cao hơn 60% hay min_sup ≥ 60%.
- Xác định các luật kết hợp của các bệnh có độ tin cậy cao lớn hơn 75% hay min_conf ≥75%

4.2 Dữ liệu và phương pháp

Trong nghiên cứu này, thuật toán Apriori được sử dụng trong các thuật toán luật kết hợp trong khai phá dữ liệu. Trong phương pháp này, tần số xuất hiện của các đối tượng kéo theo sẽ được xác định. Bảng phân loại quốc tế đầu tiên được thực hiện bởi viện thống kê quốc tế vào năm 1893. Đến năm 1948, tổ chức y tế thế giới (WHO) đã công bố bảng danh sách phân loại được biết đến là ICD-6 và sau 52 năm, bảng danh sách phân loại ICD-10 được công bố. Bảng danh sách ICD-11 được phát hành vào ngày 18 tháng 6 năm 2018 và được dự định sẽ được các quốc gia thành viên dựa theo để báo cáo về tình hình y tế quốc gia trong năm 2023.

Bảng dưới đây là danh sách các căn bệnh truyền nhiễm được WHO liệt kê trong ICD-10 [19]:

Mã bệnh	Tên bệnh
A00	Dịch tả
A01	Sốt thương hàn và phó thương hàn
A02	Nhiễm khuẩn salmonella khác
A03	Bệnh nhiễm khuẩn Shigella
A04	Nhiễm trùng đường ruột do vi khuẩn khác
A05	Nhiễm độc thực phẩm do vi khuẩn khác
A06	Bệnh lỵ amip

A07	Các bệnh đường ruột đơn bào khác
A08	Virus và nhiễm trùng đường ruột được chỉ định khác
A09	Viêm dạ dày ruột và viêm đại tràng nhiễm trùng khác

Bảng 4-1 Bảng danh sách mã bệnh ICD của WHO

Đối với dữ liệu được sử dụng được lấy từ kho lưu trữ máy học của đại học Irvine California về dữ liệu quan sát dấu hiệu covid 19 của các bệnh nhân [20].

	A01	A02	A03	A04	A05	A06	A07	Hạng mục
0	+	+	+	+	+	-	-	PUS
1	+	+	-	+	+	-	-	PUS
2	+	+	+	+	-	+	-	PUS
3	+	+	-	+	-	+	-	PUS
4	+	-	-	-	-	-	+	PUS
5	+	+	+	-	-	-	+	PUS
6	+	+	-	-	-	-	+	PUS
7	+	+	+	+	-	-	-	PUS
8	+	-	-	+	+	-	-	PIM
9	-	+	-	+	+	-	-	PIM
10	+	-	-	+	-	+	-	PIM
11	-	+	-	+	-	+	-	PIM
12	-	+	-	-	-	-	+	PIM
13	-	-	-	-	-	-	+	PWS

Bảng 4-2 Tập dữ liệu quan sát bệnh nhân

Thông tin về tập dữ liệu :

- Tập dữ liệu có 14 dòng (14 records)
- Tập dữ liệu có 8 cột. Mỗi cột đại diện cho 1 trường (fields) với các trường từ A01-A07 đại diện cho các bệnh mà bệnh nhân mắc và Categories (hạng mục) đại diện cho loại đối tượng quan sát.
- Tập dữ liệu có kích thước là >1Kb.

Tập dữ liệu thể hiện triệu chứng của những đối tượng mắc là bệnh nhân mắc covid, đối tượng được quan sát và đối tượng không có triệu chứng. PUS (Patient Under Supervision) là những bệnh nhân được quan sát trực tiếp, PIM(Person in

Monitoring) là những tối tượng đang được quan sát và PWS(Person without Symptoms) là những đối tượng không có dấu hiệu mắc bệnh. Với ký tự "+" là bệnh nhân có triệu chứng và "-" là bệnh nhân không có triệu chứng.

4.3 Tiền xử lý dữ liệu

Biến đổi dữ liệu bao gồm việc thay đổi các ký tự có giá trị là "+" và "-" lần lượt thành "1" và "0". Với 1 là giá trị đại diện cho True và 0 là giá trị đại diện cho False.

	A01	A02	A03	A04	A05	A06	A07	Hạng mục
0	1	1	1	1	1	0	0	PUS
1	1	1	0	1	1	0	0	PUS
2	1	1	1	1	0	1	0	PUS
3	1	1	0	1	0	1	0	PUS
4	1	0	0	0	0	0	1	PUS
5	1	1	1	0	0	0	1	PUS
6	1	1	0	0	0	0	1	PUS
7	1	1	1	1	0	0	0	PUS
8	1	0	0	1	1	0	0	PIM
9	0	1	0	1	1	0	0	PIM
10	1	0	0	1	0	1	0	PIM
11	0	1	0	1	0	1	0	PIM
12	0	1	0	0	0	0	1	PIM
13	0	0	0	0	0	0	1	PWS

Bảng 4-3 Tập dữ liệu qua xử lý

4.4 Áp dụng thuật toán Apriori

Với $S_{min} = 0.6$ và $C_{min} = 0.75$

4.4.1 Lập tập ứng viên candidate C_1

Xét các bệnh nhân với 1 triệu chứng, em lặp nên được bảng danh sách các ứng viên C_1.

Số bệnh nhân đang được quan sát (PUS): 8 người

Số đối tượng đang được quan sát (PIM): 5 người

Số đối tượng không có triệu chứng (PWS): 1 người

$$\min\_\sup count(PUS) = 8 * 0.6 = 4.8 \approx 5$$

$$\min\_\sup count(PIM) = 5 * 0.6 = 3$$

$$\min\_\sup count(PWS) = 1 * 0.6 = 0.6 \approx 1$$

ICD Code	PUS	PIM	PWS
A01	8	2	0
A02	7	3	0
A03	4	0	0
A04	5	4	0
A05	2	2	0
A06	2	2	0
A07	3	1	1

Bảng 4-4 Tập dữ liệu ứng viên C1

Số bệnh nhân (PUS) với triệu chứng A01(8) chiếm tỷ lệ cao nhất với độ hỗ trợ là.Số bệnh nhân (PUS) có triệu chứng A02(7) chiếm tỷ lệ cao thứ hai. Đối với hai triệu chứng là A05(2) và A06(2) chiếm tỷ lệ thấp nhất. Không có đối tượng được quan sát (PIM) là có triệu chứng A03(0). Đối với đối tượng không có triệu chứng (PWS) thì lại chỉ có 1 triệu chứng là A07(1).

Vì tần suất xuất hiện của A03,A05,A06 và A07 < min_sup count = 5 nên các mã ICD này sẽ bị xoá trong bảng các ứng viên L_1 của các bệnh nhân.

Xét bảng C_1 của các đối tượng là bệnh nhân (PUS) với min_sup count =5

Ta có bảng L_1 của các bệnh nhân:

ICD Code	Tần suất
A01	8
A02	7
A04	5

Bảng 4-5 Tập thường xuyên L1 của bệnh nhân

Vì tần suất xuất hiện của các triệu chứng A01,A03,A05,A06 và A07 < min_sup count =3 nên sẽ xoá các mã này ra hỏi tập thường xuyên L_1 của các đối tượng là đối tượng được quan sát

Xét bảng C_1 của các đối tượng là đối tượng được quan sát (PIM) với min_sup count =3

Ta có bảng L_1 của các đối tượng quan sát

ICD Code	Tần suất
A02	3
A04	4

Bảng 4-6 Tập thường xuyên L1 của đối tượng quan sát

4.4.2 Lập tập ứng viên candidate C_2

Xét các bệnh nhân (PUS) với hai triệu chứng, em lập được bảng danh sách ứng viên C_2.

ICD Code		Tần suất
A01	A02	7
A01	A04	5
A02	A04	5

Bảng 4-7 Tập ứng viên C2 của bệnh nhân

Theo quan sát, bệnh nhân có đồng hai triệu chứng A01(7) và A02(7) chiếm tỷ lệ cao nhất và hai dấu hiệu này thường đi kèm với nhau.

Vì tần suất xuất hiện của {A01,A02},{A01,A04} và {A02,A04} là ≥ min_sup count = 5 nên bảng ứng cử viên C_2 sẽ là bảng tập thường xuyên L_2 của các bệnh nhân.

Xét các đối tượng được quan sát (PIM) với hai triệu chứng em lập được bảng danh sách ứng viên C_2.

ICD Code		Tần suất
A02	A04	2

Bảng 4-8 Bảng ứng viên C2 của các đối tượng quan sát

Theo quan sát, các đối tượng được quan sát có đồng thời hai triệu chứng A02 và A04 là cao nhất. Tuy nhiên do tần suất xuất hiên của {A02,A04} < min_sup count = 3 nên tập ứng viên C_2 của các đối tượng quan sát sẽ không được xét là tập thường xuyên. Vì không còn tập thường xuyên của các đối tượng được quan sát và việc các đối tượng được quan sát mắc đồng thời cả 3 triệu chứng là không cao nên việc nghiên cứu các tối tượng này sẽ dừng lại.

4.4.3 Lập tập ứng viên candidate C_3

Dựa vào tập thường xuyên L_2 mà em lập nên bảng ứng cử viên C_3

ICD Code			Tần suất
A01	A02	A04	5

Bảng 4-9 Bảng thường xuyên L3 của các bệnh nhân

Vì tần suất xuất hiện của {A01,A02,A04} = min_sup count = 5 nên bảng ứng cử viên C_3 sẽ là bảng tập tập thường xuyên L_3.

4.4.4 Áp dụng luật kết hợp

Có 6 trường hợp với 3 triệu chứng.

1. Trường hợp 1:

$$\{A01\} \rightarrow \{A02, A04\}$$

Nếu A01(8) hiện diện thì A02 và A04(5) sẽ tồn tại cùng nhau.

$$\mathrm{S(A01, \{A02, A04\})} = \frac{5}{8} = 0.625 = 62.5\%$$

$$\mathrm{C}(A01, \{A02, A04\}) = \frac{S(A01, \{A02, A04\})}{S(A01)} = \frac{0.625}{1} = 0.625 = 62.5\%$$

2. Trường hợp 2:

$$\{A02\} \rightarrow \{A01, A04\}$$

Nếu A02(7) hiện diện thì A01 và A04(5) sẽ tồn tại cùng nhau.

$$S(A02, \{A01, A05\}) = \frac{5}{8} = 0.625 = 62.5\%$$

$$\mathrm{C}(A02, \{A01, A04\}) = \frac{S(A02, \{A01, A04\})}{S(A02)} = \frac{0.625}{0.875} = 0.714 = 71.4\%$$

3. Trường hợp 3:

$$\{A04\} \rightarrow \{A01, A02\}$$

Nếu A04(5) hiện diện thì A01 và A02(5) sẽ tồn tại cùng nhau.

$$S(A04, \{A01, A02\}) = \frac{5}{8} = 0.625 = 62.5\%$$

$$C(A04, \{A01, A02\}) = \frac{S(A04, \{A01, A02\})}{S(A04)} = \frac{0.625}{0.625} = 1 = 100\%$$

4. Trường hợp 4:

$$\{A01, A02\} \rightarrow \{A04\}$$

Nếu cả A01 và A02(7) cùng tồn tại thì A04 sẽ hiện diện.

$$S(\{A01, A02\}, A04) = \frac{5}{8} = 0.625 = 62.5\%$$

$$S(A01, A02) = \frac{7}{8} = 0.875 = 87{,}5\%$$

$$C(\{A01, A02\}, A04) = \frac{S(\{A01, A02\}, A04)}{S(A01, A02)} = \frac{0.625}{0.875} = 0.714$$

$$= 71.4\%$$

5. Trường hợp 5:

$$\{A01, A04\} \rightarrow \{A02\}$$

Nếu cả A01 và A04(5) đều cùng tồn tại thì A02 sẽ hiện diện

$$S(\{A01, A04\}, A02\} = \frac{5}{8} = 0.625 = 62.5\%$$

$$S(A01, A04) = \frac{5}{8} = 0.625 = 62.5\%$$

$$C(\{A01, A04\}, A02) = \frac{S(\{A01, A04\}, A02)}{S(A01, A04)} = \frac{0.625}{0.625} = 1 = 100\%$$

6. Trường hợp 6:

$$\{A02, A04\} \rightarrow \{A01\}$$

Nếu cả A02 và A04(5) đều cùng tồn tại thì A01 sẽ hiện diện

$$S(\{A02, A04\}, A01) = \frac{5}{8} = 0.625 = 62.5\%$$

$$S(A02, A04) = \frac{5}{8} = 0.625 = 62.5\%$$

$$C(\{A02, A04\}, A01) = \frac{S(\{A02, A04\}, A01\})}{S(A02, A04)} = \frac{0.625}{0.625} = 1 = 100\%$$

Các triệu chứng cùng tồn tại trên bệnh nhân(PUS)		Độ hỗ trợ(%)	Độ tin cậy(%)
Trường hợp 1	Nếu A01 hiện diện thì A02 và A04 sẽ cùng tồn tại.	62.5%	62.5%
Trường hợp 2	Nếu A02 hiện diện thì A01 và A04 sẽ cùng tồn tại.	62.5%	71.4%
Trường hợp 3	Nếu A04 hiện diện thì A01 và A02 sẽ cùng tồn tại	62.5%	100%
Trường hợp 4	Nếu cả A01 và A02 đều cùng tồn tại thì A04 sẽ hiện diện	62.5%	71.4%
Trường hợp 5	Nếu A01 và A04 đều cùng tồn tại thì A02 sẽ hiện diện	62.5%	100%
Trường hợp 6	Nếu A02 và A04 đều cùng tồn tại thì A01 sẽ hiện diện	62.5%	100%

Bảng 4-10 Bảng tổ hợp các trường hợp cùng với độ hỗ trợ và độ tin cậy

Mục tiêu đặt ra với tỷ lệ hỗ trợ là 60% và độ tin cậy là 75%. Tất cả trường hợp đều có độ hỗ trợ cao hơn độ hỗ trợ đề ra, tuy nhiên chỉ có 3 trường hợp là:"Trường hợp 3", "Trường hợp 5" và "Trường hợp 6" là có độ tin cậy cao hơn mục tiêu đề ra. Tuy nhiên với độ tin cậy của cả 3 trường hợp đạt 100% thì có thể dùng điều kiện của các trường hợp này để chẩn đoán ca mắc covid-19.

4.4.5 Sử dụng thư viện mlxtend để tìm ra luật kết hợp

Áp dụng hàm "apriori" từ thư viện mlxtend vào tập dữ liệu với độ hỗ trợ tối thiểu là 0.6(60%), ta có thể tạo ra các tập dữ liệu thường xuyên C.

	support	itemsets	length
0	1.000	(A01)	1
1	0.875	(A02)	1
2	0.625	(A04)	1
3	0.875	(A01, A02)	2
4	0.625	(A04, A01)	2
5	0.625	(A04, A02)	2
6	0.625	(A04, A01, A02)	3

Hình 4-1 Tập thường xuyên L

Tập dữ liệu thường xuyên bao gồm các tập items cùng với độ hỗ trợ của của các items đấy cũng như là số item trong mỗi tập thường xuyên.

Áp dụng hàm "association_rule" vào tập dữ liệu thường xuyên vừa tạo với độ tin cậy tối thiểu là 0.8(80%), ta có thể tập luật kết hợp .

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(A01)	(A02)	1.000	0.875	0.875	0.875000	1.000000	0.000000	1.0000	0.000000
1	(A02)	(A01)	0.875	1.000	0.875	1.000000	1.000000	0.000000	inf	0.000000
2	(A04)	(A01)	0.625	1.000	0.625	1.000000	1.000000	0.000000	inf	0.000000
3	(A01)	(A04)	1.000	0.625	0.625	0.625000	1.000000	0.000000	1.0000	0.000000
4	(A04)	(A02)	0.625	0.875	0.625	1.000000	1.142857	0.078125	inf	0.333333
5	(A02)	(A04)	0.875	0.625	0.625	0.714286	1.142857	0.078125	1.3125	1.000000
6	(A04, A01)	(A02)	0.625	0.875	0.625	1.000000	1.142857	0.078125	inf	0.333333
7	(A04, A02)	(A01)	0.625	1.000	0.625	1.000000	1.000000	0.000000	inf	0.000000
8	(A01, A02)	(A04)	0.875	0.625	0.625	0.714286	1.142857	0.078125	1.3125	1.000000
9	(A04)	(A01, A02)	0.625	0.875	0.625	1.000000	1.142857	0.078125	inf	0.333333
10	(A01)	(A04, A02)	1.000	0.625	0.625	0.625000	1.000000	0.000000	1.0000	0.000000
11	(A02)	(A04, A01)	0.875	0.625	0.625	0.714286	1.142857	0.078125	1.3125	1.000000

Hình 4-2 Tập luật kết hợp

Tập luật kết hợp từ tập tập thường xuyên bao gồm: antecedents, consequents, antecedent support, consequent support, support, confidence, lift, leverage, conviction, zhangs_metric.

Trong đấy, em tập trung vào các luật kết hợp có được từ tập thường xuyên mà có từ 3 items trở lên bao gồm: (A04,A01) → A02, (A04,A02) → (A01,A02)→A04, A01 →(A02,A04), A02 →(A01,A04), A04 →(A01,A02).

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	antecedents_len	consequents_len	Cases
(A01)	(A02, A04)	1.000	0.625	62.5	62.500000	1.000000	0.000000	1.0000	0.000000	1	2	Case 1
(A02)	(A04, A01)	0.875	0.625	62.5	71.428571	1.142857	0.078125	1.3125	1.000000	1	2	Case 2
(A04)	(A02, A01)	0.625	0.875	62.5	100.000000	1.142857	0.078125	inf	0.333333	1	2	Case 3
(A02, A01)	(A04)	0.875	0.625	62.5	71.428571	1.142857	0.078125	1.3125	1.000000	2	1	Case 4
(A02, A04)	(A01)	0.625	1.000	62.5	100.000000	1.000000	0.000000	inf	0.000000	2	1	Case 5
(A04, A01)	(A02)	0.625	0.875	62.5	100.000000	1.142857	0.078125	inf	0.333333	2	1	Case 6

Hình 4-3 Tập luật kết hợp có 3 items trở lên

Trong các luật kết hợp trên, có 3 luật thoả với min_conf ≥ 0.8(80%) là: A04→(A01, A02), (A04, A01)→A02, (A04, A02)→A01. Cả 3 luật trên đều có độ tin cậy tối đa là 100%, điều này cho thấy độ chính xác của 3 luật này là rất cao. Ngoài support và confidence ra thì hàm "association_rules" còn cung cấp thêm các hệ số khác. Bao gồm:

- Lift: Là hệ số ám chỉ khả năng cả xuất hiện của cả luật kết hợp so với chỉ hệ quả . Ví dụ trong Case 3 có lift = 1.14, điều này có nghĩa khả năng xảy ra đồng thời A04→(A02,A01) so với khả năng chỉ xảy ra A02→A01 cao hơn 1.14 lần.
- Leverage: Là hệ số khác biệt giữa độ thường xuyên cùng nhau của tiền đề và hệ quả so với độ thường xuyên của tiền đề và hệ quả độc lập với nhau. Ví dụ trong case 3 có leverage = 0.078125, điều này có nghĩa sự khác biệt giữa độ phổ biến của (A04,{A02,A01}) so với A04 và {A02,A01} là 0.078125.
- Conviction: Là hệ số phụ thuộc của hệ quả đối với tiền đề. Hệ số này càng cao thì độ phụ thuộc của hệ quả vào tiền đề là càng cao. Ví dụ trong case 5, độ phụ thuộc của {A02,A04} → A01 là inf do độ tin cậy (confidence) = 1. Điều này có nghĩa là sự phụ thuộc của A01 với {A02,A04} là chắc chắn.
- Zhang_metrics: Là hệ số xác định sự kết hợp hoặc phân ly. Giá trị giao động khoảng từ -1 đến 1. Nếu giá trị dương là kết hợp còn giá trị âm là phân ly. Ví dụ trong case 4, hệ số zhang_metric là 1 thể hiện luật (A02,A01)→A04 là luật kết hợp.

4.5 Đối chiếu kết quả

Dựa vào kết quả có được từ thuật toán Apriori, em nhận thấy 3 triệu chứng phổ biến nhất của các bệnh nhân mắc covid-19 được theo dõi là :

- A01 : Sốt thương hàn và phó thương hàn

- A02 : Nhiễm khuẩn salmonella
- A04 : Nhiễm trùng đường ruột do vi khuẩn

Dựa vào 3 triệu chứng phổ biến và có mức độ tin cậy cao nhất mà em có được từ thuật toán Apriori, em tiến hành đối chiếu kết quả có được với các nghiên cứu khoa học đã có với các triệu chứng lần lượt theo thứ tự là : "Sốt thương hàn và phó thương hàn", "Nhiễm khuẩn salmonella" và cuối cùng là "Nhiễm trùng đường ruột do vi khuẩn".

4.5.1 Mối quan hệ giữa covid-19 và sốt thương hàn và phó thương hàn

Theo SETA (Chương trình giám sát bệnh thương hàn nghiêm trọng ở châu Phi) chỉ ra số ca mắc bệnh thương hàn đã tăng 6.4 lần trong khoảng thời gian đại dịch hoành hành ở Madagascar. Trong nghiên cứu của họ, trong khoảng từ tháng 8 năm 2016 đến tháng 1 năm 2020 (tiền đại dịch covid-19) thì tỷ lệ mắc của bệnh sốt thương hàn là 2.1/100000 người/năm và số ca tử vong là 0.6/100000 người/năm so với khoảng thời gian đại dịch là 13.2/100000 người/năm và số ca tử vong 5.1/100000 người/năm [21].

4.5.2 Mối quan hệ giữa covid-19 và nhiễm khuẩn salmonella

Trong báo cáo vào tháng 2 năm 2021 từ thư viện y học quốc gia của hoa kỳ về ca mắc covid có triệu chứng của khuẩn salmonella trên bệnh nhân 56 tuổi tại Pakistan với các triệu chứng thông thường của bệnh covid 19 như sốt, khó thở, đau cơ kéo dài. Ngoài ra các bác sĩ đã xác nhận ông bị nhiễm khuẩn salmonella khi họ xét nghiệm và tìm ra trong máu ông có chứa Salmonella enterica spp. enterica, là một trong những loại vi khuẩn thuộc salmonella [22].

Các bác sĩ đã đưa ra các suy luận rằng việc mắc bệnh covid 19 đã dẫn đến cơ chế miễn dịch của cơ thể bị ức chế và làm cho việc bị nhiễm khuẩn salmonella trở nên dễ dàng hơn [22].

4.5.3 Mối quan hệ giữa covid-19 và bệnh nhiễm khuẩn đường ruột

Trong bài nghiên cứu của tiến sĩ Ken Cadwell cùng đồng nghiệp của ông đã chỉ ra sự ảnh hưởng của dịch covid đến hệ vi khuẩn đường ruột khi họ xét nghiệm mẫu dịch của 96 bệnh nhân và nhận thấy một chi vi khuẩn chiếm ưu thế. Những vi khuẩn chiếm ưu thế này bao gồm các mầm bệnh cơ hội và kháng kháng sinh [23].

Họ đã đưa ra kết luận rằng covid-19 đã phá vỡ hệ vi khuẩn đường ruột. Điều này cho phép nhiễm trùng thứ phát do vi khuẩn, cả bằng cách cho phép vi khuẩn gây bệnh xâm chiếm ruột và bằng cách thay đổi niêm mạc ruột để cho phép những vi khuẩn này dễ dàng lây lan từ ruột vào máu hơn [23].

4.5.4 So sánh kết quả giữa hai thuật toán là Apriori và C4.5

Trong bài nghiên cứu của 2 nghiên cứu sinh trên cùng tập dữ liệu là "Covid 19 Surveillance Data set" được đăng vào 03/2020 trên tờ Jurnal PILAR NUSA Mandiri Vol 16. Hai nghiên cứu sinh đã dùng kỹ thuật cây quyết định với thuật toán C4.5 là một trong những thuật toán nổi bật trong kỹ thuật khai phá phân lớp. Kết quả cho ra từ nghiên cứu [24]:

1. Nếu G01 là true và G02 là true thì việc chẩn đoán sẽ bao gồm trong danh mục PDP
2. Nếu G01 là true, G02 là false và G04 là true thì việc chẩn đoán sẽ bao gồm trong danh mục ODP.
3. Nếu G01 là true, G02 là false và G04 là false thì việc chẩn đoán sẽ bao gồm trong danh mục PDP.
4. Nếu G01 là false và G02 là true thì việc chẩn đoán sẽ bao gồm trong danh mục ODP.
5. Nếu G01 là false và G02 là false thì việc chẩn đoán sẽ bao gồm trong danh mục OTG.

Trong đó :

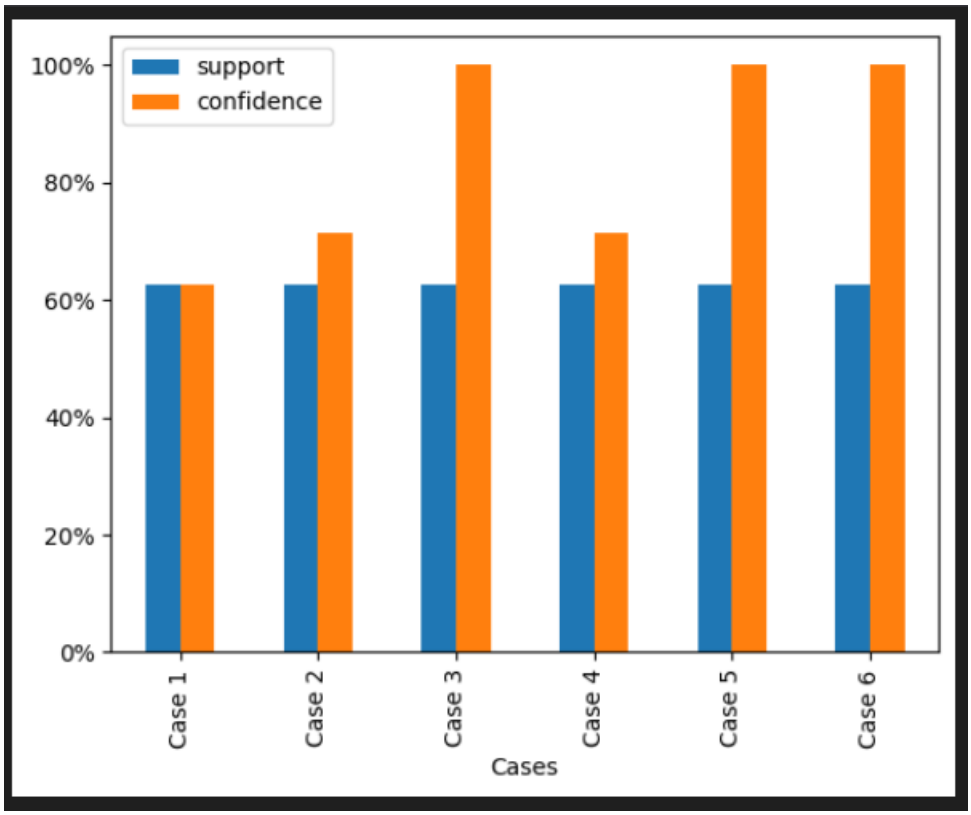
- G01 tương đương với mã bệnh A01
- G02 tương đương với mã bệnh A02
- G04 tương đương với mã bệnh A04
- PDP tương đương với loại bệnh nhân PUS
- ODP tương đương với loại bệnh nhân PIM
- OTG tương đương với loại bệnh nhân PWS

So sánh giữa hai thuật toán đều cho ra được 3 triệu chứng phổ biến nhất là A01(Sốt thương hàn và phó thương hàn), A02(Nhiễm khuẩn salmonella) và A04(Nhiễm trùng đường ruột do vi khuẩn). Trong thuật toán C4.5 của hai nghiên cứu sinh phân lớp dựa vào việc phân loại dựa trên loại bệnh nhân PDP, ODP và OTG với độ chính xác lên đến 92.86% [24]. Đối với thuật toán Apriori trong bài nghiên cứu của em có 3 trường hợp đạt độ tin cây 100% và cả ba trường hợp đều có sự hiện diện của cả 3 triệu chứng A01,A02 và A04.

4.6 Mô hình hoá và kết luận

4.6.1 Mô hình hoá

Dựa vào tập luật kết hợp L, em tiến hành lập biểu đồ cột các trường hợp có từ 3 triệu chứng.



Hình 4-4 Đồ thị biểu diễn độ hỗ trợ và độ tin cậy của các trường hợp có 3 triệu chứng

Dựa vào mô hình trên, có thể xác định rằng độ hỗ trợ của cả 6 case đều bằng nhau (0.625), điều này có nghĩa độ phong phú của cả 6 trường hợp trên đều là ngang nhau. Đối với độ tin cậy của các trường hợp, trường hợp 1, 2, 4 có độ tin cậy thấp hơn ngưỡng tối thiểu là 0.8. Tuy nhiên, các trường hợp 3, 5 và 6 lại có độ tin cậy tối đa là 100%. Cho thấy các trường hợp trên là đáng tin cậy cho việc chẩn đoán.

4.6.2 Kết luận

Trong bài áp dụng, kỹ thuật khai phá dữ liệu đã được sử dụng để chẩn đoán các triệu chứng của covid-19. Trong đó thuật toán Apriori đã được sử dụng làm phương pháp khai phá dữ liệu. Đối với tập dữ liệu sử dụng là "COVID-19 Surveliance" từ đại học California-Irvine.Tập dữ liệu bao gồm các thông tin của 14 đối tượng được quan sát bao gồm thông tin mã bệnh ICD, các loại dấu hiệu mà họ đang mắc và phân loại các đối tượng quan sát.

Trong bài áp dụng, các luật kết hợp từ việc áp dụng thuật toán Apriori trên tập dữ liệu về đối tượng quan sát. Thuật toán cho em thấy rằng, "nếu A04 tồn tại thì A01 và A02 sẽ cùng tồn tại ", "nếu A01 và A04 cùng tồn tại thì A02 sẽ tồn tại" và "nếu A02 và A04 cùng tồn tại thì A01 sẽ tồn tại" tất cả 3 trường hợp trên đều xảy ra trong 100% bệnh nhân. Có nghĩa rằng nếu bệnh nhân bị nhiễm trùng đường ruột do vi khuẩn khác, nhiễm khuẩn salmonella khác và sốt thường hàn,phó thương hàn cùng

nhau thì bệnh nhân dương tính với covid-19. Trong trường hợp 2, nếu bệnh nhân bị sốt thương hàn, phó thương hàn và nhiễm trùng đường ruột khác cùng nhau, họ cũng bị nhiễm khuẩn salmonella khác thì bệnh nhân được xác định dương tính với covid-19. Trong trường hợp 3, nếu bệnh nhân có triệu chứng của nhiễm khuẩn salmonella khác và bệnh nhiễm khuẩn đường ruột, họ cũng bị sốt thương hàn và phó thương hàn thì bệnh nhân được xác định dương tính với covid-19. Đối với các đối tượng chỉ có triệu chứng của bệnh đường ruột đơn bào khác thì họ không dương tính với covid-19. Tuy nhiên vẫn cần phải xét nghiệm PCR để có thể chắc chắn về việc bị nhiễm covid-19. Bài áp dụng cũng cho thấy việc chẩn đoán có thể được thực hiện nhanh hơn nhờ vào việc đánh giá việc các triệu chứng cùng tồn tại.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Trong 3 tháng tìm hiểu, em đã hoàn thành được các mục tiêu đề ra tuy nhiên vẫn còn một số thiếu sót trong quá trình tìm hiểu.

Đề tài đã hoàn thành các mục tiêu đã đề ra :

- Em đã tìm hiểu được các khái niệm về phân tích dữ liệu, quy trình phân tích và các kỹ thuật cơ bản trong phân tích dữ liệu.
- Em đã tìm hiểu được khái niệm về khai phá dữ liệu, đặc biệt là kỹ thuật tìm tập phổ biến và luật kết hợp. Áp dụng được thuật toán Apriori trong việc tìm ra các kết hợp sản phẩm trong ví dụ và áp dụng vào việc chẩn đoán Covid-19 trong bài áp dụng.
- Em đã tìm hiểu về việc áp dụng Python và sử dụng các thư viện của Python trong việc tiền xử lý dữ liệu, kiểm tra file, mô hình hoá kết quả khai phá và áp dụng thuật toán Apriori một cách dễ dàng nhờ vào thư viện mlxtend.
- Em đã áp dụng thành công thuật toán Apriori trong việc chẩn đoán Covid-19 đối với các bệnh nhân được quan sát. Thuật toán cho ra được kết quả với độ hỗ trợ và độ tin cậy cao.

Đề tài còn các điểm hạn chế như :

- Em vẫn chưa tìm hiểu được các thuật toán khác dùng để tìm tập phổ biến và luật kết hợp như các kỹ thuật nâng cao của thuật toán Apriori, thuật toán SETM và thuật toán AIS.
- Tập dữ liệu mà em sử dụng trong bài áp dụng còn hạn chế và các triệu chứng mà em dùng để đánh giá đa phần là các triệu chứng chuyên ngành dẫn đến việc áp dụng kết quả khai phá để chẩn đoán diện rộng là không khả quan.
- Em chỉ áp dụng một thuật toán là Apriori cơ bản nên dẫn đến kết quả khai phá vẫn chưa mang tính thuyết phục cao do chưa có các kết quả khác để đối chiếu với kết quả khai phá.

Đặc biệt, em xin chân thành cảm ơn các thầy/cô trong khoa Kỹ thuật - Công nghệ, đặc biệt là ThS. LÊ VĂN HẠNH vì đã tận tình giúp đỡ em hoàn thành đề tài này.

5.2 Hướng phát triển

Trong hướng phát triển tiếp theo, em sẽ tìm hiểu thêm về các kỹ thuật khai phá tập phổ biến và luật kết hợp khác như Apriori TID, SETM và AIS. Ngoài ra, em cũng sẽ tìm các tập dữ liệu có kích thước lớn hơn với các triệu chứng phổ thông hơn như: sốt, ho, nhức đầu, … Để cho kết quả trở nên phổ thông và dễ dàng chẩn đoán hơn.

TÀI LIỆU THAM KHẢO

[1]. Wiguna, W., Riana, D., "Diagnosis of Coronavirus disease 2019(Covid-19) surveillance using C4.5 algorithm", Journal PILAR Nusa Mandiri, 16 (2020): (71-80).

[2]. Amazon, "What is data analytics", What is Data Analytics? - Data Analytics Explained - AWS (amazon.com), (2023).

[3]. Lưu Hà Chi, "CÁC PHƯƠNG PHÁP THU THẬP DỮ LIỆU SƠ CẤP VÀ THỨ CẤP", https://luanvanviet.com/cac-phuong-phap-thu-thap-du-lieu-cap-va-thu-cap/, (2021)

[4]. Hồ Nguyễn, "Data Lake là gì? Phân biệt Data Warehouse và Data Lake", https://blog.trginternational.com/vi/phan-biet-su-khac-nhau-giua-data-lake-va-data-warehouse , (2018).

[5]. Sana Mushtaq, "Data preprocessing in detail", Data preprocessing in detail - IBM Developer, (2019)

[6]. Ibrahim Abayomi Ogunbiyi, "How to handle missing data in a dataset", How to Handle Missing Data in a Dataset (freecodecamp.org), (24/06/2022).

[7]. Geeksforgeeks, "Data Transformation in Data Mining", Data Transformation in Data Mining - GeeksforGeeks,(2023).

[8]. Geeksforgeeks, "Data Reduction in Data Mining", https://www.geeksforgeeks.org/data-reduction-in-data-mining/,(2023)

[9]. Bernardita Calzon, "Your Modern Business Guide To Data Analysis Methods And Techniques", https://www.datapine.com/blog/data-analysis-methods-and-techniques, (2023)

[10]. TS.Đỗ Phúc, "Tổng quan khai thác dữ liệu , "Khai thác dữ liệu", Đại học công nghệ thông tin (2012) : (5-19).

[11]. Nguyễn Thị Biên, "Luật kết hợp" ,"Khai phá luật kết hợp trong cơ sở dữ liệu đa phương tiện", luận văn thạc sỹ ngành công nghệ thông tin (2012) : (9-16).

[12]. Rakesh Agrawal, "Fast Algorithms for mining Association Rules", "IBM Almaden Research Center", (1994).

[13]. Pandas, "Pandas documentation", pandas documentation — pandas 2.0.1 documentation (pydata.org), version 2.0.2 (2023)

[14]. Numpy ,"Numpy documentation", NumPy Documentation, version 1.24 (2022)

[15]. Matplotlib, "Matplotlib documentation", Matplotlib documentation — Matplotlib 3.7.1 documentation, version 3.7.0 (2023)

[16]. Sebastian Raschka, "mlxtend documentation",mlxtend (rasbt.github.io), (2023)

[17]. Sebastian Rashka, “apriori: Frequent itemsets via the Apriori algorithm”, “mlxtend documentation”, Apriori - mlxtend (rasbt.github.io)

[18]. Sebastian Rashka, “association_rules: Association rules generation from frequent itemsets”, “mlxtend documentation”, Association rules - mlxtend (rasbt.github.io)

[19]. WHO,“Chapter I Certain infectious and parasitic diseases”, ICD-10 Version:2016 (who.int) (2019)

[20]. Dua, D., Graff, C., “Covid 19 Surveillance Data set”, University of California, School of Information and Computer Science. Irvine, USA, (2019)

[21]. Đại học y dược Cambridge, “Surge of Typhoid Intestinal Perforations as Possible Result of COVID-19–Associated Delays in Seeking Care, Madagascar”, Surge of Typhoid Intestinal Perforations as Possible Result of COVID-19–Associated Delays in Seeking Care, Madagascar - PMC (nih.gov) , (2021)

[22]. Fatma Yekta Ürkmez, Tuğba Atalay, “Salmonella Bacteremia Accompanying COVID-19: The First Salmonella Co-Infection in the World Unrelated to Pakistan”, [Salmonella Bacteremia Accompanying COVID-19: The First Salmonella Co-Infection in the World Unrelated to Pakistan] - PubMed (nih.gov), (2022)

[23]. Drs. Ken Cadwell and Jonas Schluter, “COVID-19 disrupts gut microbiome”, COVID-19 disrupts gut microbiome | National Institutes of Health (NIH), (2022)

[24]. Wildan Wiguna & Dwiza Riana, “*DIAGNOSIS OF CORONAVIRUS DISEASE 2019 (COVID-19) SURVEILLANCE USING C4.5 ALGORITHM*”, Jurnal PILAR Nusa Mandiri Vol. 16, (2020).