**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**CAO TIẾN HIẾU**

**PHÂN TÍCH DỮ LIỆU TỶ PHÚ VÀ CÔNG TY ĐA QUỐC GIA**

**ĐỒ ÁN NGÀNH**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**TP. HỒ CHÍ MINH, 2023BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**CAO TIẾN HIẾU**

**PHÂN TÍCH DỮ LIỆU TỶ PHÚ VÀ CÔNG TY ĐA QUỐC GIA**

**Mã số sinh viên: 2051052038**

**ĐỒ ÁN NGÀNH**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Giảng viên hướng dẫn: VÕ THỊ KIM ANH**

**TP. HỒ CHÍ MINH, 2023**

LỜI CẢM ƠN

Báo cáo đồ án ngành với đề tài “Phân tích dữ liệu tỷ phú và công ty đa quốc gia”, là một quá trình cố gắng và phát triển không ngừng của bản thân em và được sự chỉ bảo giúp đỡ của các quý thầy, cô cho em những lời khuyên, những kiến thức bổ ích. Để hoàn thành báo cáo lần này, em xin cảm ơn cô Võ Thị Kim Anh, đã tận tình hướng dẫn trong suốt quá trình làm bài và viết báo cáo Đồ án ngành.

Em xin chân thành cảm ơn quý thầy, cô khoa Công Nghệ Thông Tin, Trường Đại học Mở Thành phố Hồ Chí Minh đã tận tình chỉ dạy, truyền đạt kiến thức trong những năm em học tập tại trường. Với những kiến thức được tiếp thu trong quá trình học tập vừa là nền tảng để em có thể làm được bài báo cáo Đồ án lần này còn là kiến thức để hỗ trợ trong việc thực tập sắp tới.

Cuối cùng em xin chúc tất cả quý thầy, cô tại trường luôn luôn đạt nhiều thành công tốt đẹp trong sự nghiệp và trong công việc.

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

TÓM TẮT ĐỒ ÁN NGÀNH

Ở đồ án phân tích dữ liệu lần này, em đã chọn phân tích dữ liệu tỷ phú và các công ty đa quốc gia. Với mục đích là tìm ra được những nhóm tỷ phú với các ngành nghề liên quan và các ngành nghề tiềm năng ở các năm sắp tới. Trong đồ án em chọn việc phân tích top 100 tỷ phú và các công ty để phân tích nhằm đem lại giá trị tốt nhất về sự ảnh hưởng của những công ty đi đầu xu hướng.

Trong đồ án, em đã áp dụng 3 phương pháp đã học từ môn Khai phá dữ liệu là: Gom cụm, Luật kết hợp và Phân lớp. Đầu tiên sẽ là việc thu thập dữ liệu từ trang Forbes và Kaggle, sau đó xử lý các dữ liệu lỗi hoặc thiếu và giảm chiều dữ liệu. Thứ hai là phân tích tài sản và ngành nghề để xác định xu hướng và mô hình. Tiếp theo là so sánh và đánh giá các biểu đồ để đánh giá sự khác biệt và tầm quan trọng của ngành nghề hoặc quốc gia. Và cuối cùng là đưa ra dự đoán về ngành nghề tiềm năng trong tương lai.

Các bước đầu tiên là Tiền xử lý dữ liệu thì em sẽ làm thủ công bằng Excel để xử lý các dữ liệu thiếu, bị lỗi. Ở kỹ thuật Gom cụm em sử dụng công cụ phân tích dữ liệu là Weka để hỗ trợ. Kỹ thuật Phân lớp em sử dụng cây quyết định, cũng sử dụng công cụ Weka những sẽ có phần trực quan hóa bằng hình ảnh để đưa ra nhận xét về các lớp. Riêng kỹ thuật Luật kết hợp em sử dụng ngôn ngữ Python ở trang Google Colab để code và tính toán ra được những luật có tác động tới dữ liệu và lựa chọn luật nào phù hợp. Và từ các luật có tác động sẽ đưa ra được những dự đoán liên quan về ngành nghề. Cuối cùng tổng hợp lại bằng các biểu đồ từ Excel và đưa ra nhận xét ở trong Word.

ABSTRACT

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc148877388)

[NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN 2](#_Toc148877389)

[TÓM TẮT ĐỒ ÁN NGÀNH 3](#_Toc148877390)

[ABSTRACT 4](#_Toc148877391)

[DANH MỤC TỪ VIẾT TẮT 7](#_Toc148877392)

[DANH MỤC HÌNH VẼ 8](#_Toc148877393)

[DANH MỤC BẢNG 9](#_Toc148877394)

[MỞ ĐẦU 10](#_Toc148877395)

[Chương 1. TIỀN XỬ LÝ DỮ LIỆU 11](#_Toc148877396)

[1.1. Thu thập dữ liệu 11](#_Toc148877397)

[1.2. Giảm nhỏ dữ liệu 11](#_Toc148877398)

[1.2.1. Giảm chiều dọc 11](#_Toc148877399)

[1.2.2. Giảm chiều ngang 11](#_Toc148877400)

[1.3. Mô tả dữ liệu 12](#_Toc148877401)

[1.4. Làm sạch dữ liệu 14](#_Toc148877402)

[Chương 2. GOM CỤM DỮ LIỆU 16](#_Toc148877403)

[2.1. Lý thuyết gom cụm 16](#_Toc148877404)

[2.1.1. Các độ đo khoảng cách cụm 16](#_Toc148877405)

[2.1.2. Chuẩn hóa dữ liệu 17](#_Toc148877406)

[2.2. Gom cụm áp dụng thuật toán K-means 17](#_Toc148877407)

[Chương 3. MÔ HÌNH DỰ ĐOÁN SỬ DỤNG LUẬT KẾT HỢP 23](#_Toc148877408)

[3.1. Lý thuyết luật kết hợp 23](#_Toc148877409)

[3.1.1. Thuật toán Apriori khai phá luật kết hợp 23](#_Toc148877410)

[3.2. Áp dụng thuật toán Apriori qua ngôn ngữ lập trình 23](#_Toc148877411)

[Chương 4. PHÂN LỚP BẰNG CÂY QUYẾT ĐỊNH 39](#_Toc148877412)

[4.1. Lý thuyết phân lớp 39](#_Toc148877413)

[4.1.1. Phân lớp với cây quyết định 39](#_Toc148877414)

[4.2. Áp dụng phân lớp trong KPDL 40](#_Toc148877415)

[Chương 5. TỔNG KẾT 48](#_Toc148877416)

[TÀI LIỆU THAM KHẢO 49](#_Toc148877417)

[PHỤ LỤC 50](#_Toc148877418)

DANH MỤC TỪ VIẾT TẮT

CSDL: Cơ Sở Dữ Liệu

KPDL: Khai Phá Dữ Liệu

MTSS: Ma Trận Sai Sót

DANH MỤC HÌNH VẼ

[Hình 1. 1: Dữ liệu đưa vào Weka khi chưa xử lý bị lỗi 14](#_Toc148574031)

[Hình 2. 1: Biểu đồ gom cụm dữ liệu tỷ phú 18](#_Toc148574040)

[Hình 2. 2: Biểu đồ gom cụm dữ liệu công ty năm 2020 19](#_Toc148574041)

[Hình 2. 3: Biểu đồ gom cụm dữ liệu công ty năm 2021 20](#_Toc148574042)

[Hình 2. 4: Biểu đồ gom cụm dữ liệu công ty năm 2022 21](#_Toc148574043)

[Hình 3. 1: Biểu đồ tổng tài sản theo ngành nghề năm 2022 ở Mỹ 26](#_Toc148574045)

[Hình 3. 2: Biểu đồ tổng tài sản theo quốc gia năm 2022 26](#_Toc148574046)

[Hình 3. 3: Biểu đồ tổng tài sản cổ phiếu theo quốc gia năm 2020 28](#_Toc148574047)

[Hình 3. 4: Biểu đồ tổng lợi nhuận theo quốc gia năm 2020 29](#_Toc148574048)

[Hình 3. 5: Biểu đồ tổng kinh doanh theo quốc gia năm 2020 29](#_Toc148574049)

[Hình 3. 6: Biểu đồ trung bình giá trị theo quốc gia năm 2020 30](#_Toc148574050)

[Hình 3. 7: Biểu đồ tổng tài sản cổ phiếu theo quốc gia năm 2021 32](#_Toc148574051)

[Hình 3. 8: Biểu đồ tổng lợi nhuận theo quốc gia năm 2021 32](#_Toc148574052)

[Hình 3. 9: Biểu đồ tổng kinh doanh theo quốc gia năm 2021 33](#_Toc148574053)

[Hình 3. 10: Biểu đồ trung bình giá trị theo quốc gia năm 2021 33](#_Toc148574054)

[Hình 3. 11: Biểu đồ tổng tài sản cổ phiếu theo quốc gia năm 2022 35](#_Toc148574055)

[Hình 3. 12: Biểu đồ tổng lợi nhuận theo quốc gia năm 2022 36](#_Toc148574056)

[Hình 3. 13: Biểu đồ tổng kinh doanh theo quốc gia năm 2022 36](#_Toc148574057)

[Hình 3. 14: Biểu đồ trung bình giá trị theo quốc gia năm 2022 37](#_Toc148574058)

[Hình 4. 1: Khái quát đầu vào dữ liệu tỷ phú trong phân lớp 40](#_Toc148574059)

[Hình 4. 2: Mô tả kết quả sau khi chạy mô hình phân lớp dữ liệu tỷ phú 41](#_Toc148574060)

[Hình 4. 3: Cây quyết định dữ liệu tỷ phú 42](#_Toc148574061)

[Hình 4. 4: Khái quát đầu vào dữ liệu công ty trong phân lớp 45](#_Toc148574062)

[Hình 4. 5: Mô tả kết quả sau khi chạy mô hình phân lớp dữ liệu công ty 46](#_Toc148574063)

[Hình 4. 6: Cây quyết định dữ liệu công ty 47](#_Toc148574064)

DANH MỤC BẢNG

[Bảng 1. 1: Bảng luật kết hợp dữ liệu tỷ phú 24](#_Toc148572368)

[Bảng 1. 2: Bảng luật kết hợp ý nghĩa dữ liệu tỷ phú 24](#_Toc148572369)

[Bảng 2. 1: Bảng luật kết hợp dữ liệu công ty năm 2020 26](#_Toc148572376)

[Bảng 2. 2: Bảng luật kết hợp ý nghĩa dữ liệu công ty năm 2020 26](#_Toc148572377)

[Bảng 3. 1: Bảng luật kết hợp dữ liệu công ty năm 2021 29](#_Toc148572419)

[Bảng 3. 2: Bảng luật kết hợp ý nghĩa dữ liệu công ty năm 2021 30](#_Toc148572420)

[Bảng 4. 1: Bảng luật kết hợp dữ liệu công ty năm 2022 32](#_Toc148572423)

[Bảng 4. 2: Bảng luật kết hợp ý nghĩa dữ liệu công ty năm 2022 33](#_Toc148572424)

MỞ ĐẦU

Hiện nay, phân tích dữ liệu đang là xu hướng phát triển tiềm năng của các công ty. Nhằm mục đích để nâng cao những giá trị và khả năng phát triển của khách hàng cũng như sản phẩm thông qua dữ liệu để đưa ra được những mối liên kết. Vì sự phát triển về CSDL trong cuộc sống nên cũng thúc đẩy việc thu thập dữ liệu, xử lý hay đưa ra các mô hình cần phải tỉ mỉ hơn. Các chuyên gia phải đưa ra được những phương án xử lý tốt hơn, những mô hình thông minh và hiệu quả hơn. Qua đó nâng cao hiệu suất cho công ty.

Một trong những kỹ thuật được áp dụng đó chính là Khai phá dữ liệu (Data Mining). Mặc dù kỹ thuật này được áp dụng ở rất nhiều quốc gia trên thế giới những vẫn còn khá mới mẻ so với ở nước ta. **Khai phá dữ liệu – Data Mining** là quá trình phân loại, sắp xếp các tập hợp dữ liệu nhất định để xác định xu hướng, các mẫu và thiết lập các mối liên hệ hữu ích nhằm giải quyết các vấn đề nhờ phân tích dữ liệu. Mục tiêu của việc này là cho phép các doanh nghiệp có thể dự đoán được xu hướng tương lai, nhằm đưa ra các quyết định được hỗ trợ dữ liệu từ các tập dữ liệu khổng lồ [1]. Trong đồ án lần này em sẽ áp dụng những kỹ thuật cơ bản của Khai phá dữ liệu (Data Mining) như Gom cụm, Phân lớp và một kỹ thuật quan trọng chính là Luật kết hợp để đưa ra dự đoán.

Trong đồ án khi làm các kỹ thuật trong KPDL sẽ cần có sự trợ giúp của các ứng dụng phân tích dữ liệu như Weka và sử dụng Power BI, Excel để trực quan hóa dữ liệu. Qua những ứng dụng này giúp chúng ta có cái nhìn tổng quát hơn và chi tiết hơn về dữ liệu cũng như cách vận hành của từng kỹ thuật.

Vậy nên việc chọn đề tài phân tích tỷ phú và các công ty đa quốc gia để em có thể đưa ra được những ngành nghề tiềm năng trong tương lai. Và khi phát triển sâu thì sẽ giúp ích trong việc tìm hiểu các ngành nghề nổi trội đem lại nguồn lợi và tiềm năng hơn.

# TIỀN XỬ LÝ DỮ LIỆU

## Thu thập dữ liệu

- Về dữ liệu tỷ phú (Billionaires) là dữ liệu được lấy ở năm gần đây nhất là 2022 vì sẽ có thống kê chính xác cho từng vị trí vào đúng thời điểm kết thúc năm 2022 và bảng thống kê có tên là “The Richest in 2022” ở trang Forbes. Dữ liệu gốc bao gồm 2668 tỷ phú.

- Về dữ liệu các công ty đa quốc gia (Companies) sẽ lấy 3 năm gần đây nhất là 2020, 2021 và 2022 ở trang Kaggle. Dữ liệu gốc bao gồm 2000 công ty đa quốc gia.

## Giảm nhỏ dữ liệu

- Khi một CSDL đưa vào mô hình có nhiều chiều dữ liệu thì sẽ gây ra việc tiêu tốn dữ liệu và mất rất nhiều thời gian để huấn luyện. Vì vậy khi đối diện với những kiểu dữ liệu như vậy ta thường phải chọn giảm nhỏ chiều dữ liệu lại từ không gian đa chiều (high dimensionality) về lại không gian thấp chiều (low dimensionality) mà vẫn phải giữ nguyên đặc trưng của dữ liệu [2].

### Giảm chiều dọc

- Ở dữ liệu Billionaires và Companies em đều giảm số lượng dữ liệu thành 100 instance. Mục đích để lấy được những bộ dữ liệu top đầu và có sức ảnh hưởng cao nhất chắc chắn có tác động tới những thuộc tính khác. Ngoài ra vẫn sẽ đảm bảo được dữ liệu không bị thay đổi về thuộc tính chính.

### Giảm chiều ngang

- Dữ liệu gốc sẽ có những thuộc tính sau khi đưa vào mô hình vừa không có sự ảnh hưởng tới dữ liệu khác, vừa tốn bộ nhớ dữ liệu, vừa tốn rất nhiều thời gian vào việc huấn luyện mô hình. Những dữ liệu đó em sẽ giảm bớt khi đưa vào mô hình để tạo được tính đặc trưng và có độ ảnh hưởng nhất định trong mô hình.

- Ở data Billionaires giảm chiều của các cột không cần thiết như State, Organization, City, BirthDate, Title, PhilanthropyScore, Bio và About.

- Ở data Companies của năm 2020 và 2021 sẽ giảm chiều của cột Year.

## Mô tả dữ liệu

**\*Data Billionaires:**

- Các thông tin khai phá:

+ Có 9 instance và bộ dữ liệu không missing.

+ Attribute gồm: Rank, Name, Age, NetWorth, Industry, Source, Country, SelfMade, Gender.

+ Các attribute thuộc dạng nominal: Name, Industry, Source, Country, SelfMade, Gender.

+ Các attribute thuộc dạng numeric: Rank, Age, NetWorth.

+ Các attribute Rank, Age, NetWorth thuộc dạng numeric có số liệu chi tiết và dùng để phân tích.

+ Thuộc tính Industry gồm: 16 (thuộc tính).

+ Thuộc tính Source gồm: 77 (thuộc tính).

+ Thuộc tính Country gồm: 22 (thuộc tính).

**\*Data Companies:**

**Năm 2020:**

- Các thông tin khai phá:  
+ Có 7 instance và bộ dữ liệu không missing.  
+ Attribute gồm: Rank, Name, Country, Sales, Profit, Assets, Market Value.

+ Các attribute thuộc dạng nominal: Name, Country.

+ Các attribute thuộc dạng numeric: Rank, Sales, Profit, Assets, Market Value.

+ Các attribute Rank, Sales, Profit, Assets, Market Value thuộc dạng numeric có số liệu chi tiết và dùng để phân tích.

+ Thuộc tính Country gồm: 19 (thuộc tính).

**Năm 2021:**

- Các thông tin khai phá:  
+ Có 7 instance và bộ dữ liệu không missing.  
+ Attribute gồm: Rank, Name, Country, Sales, Profit, Assets, Market Value.

+ Các attribute thuộc dạng nominal: Name, Country.

+ Các attribute thuộc dạng numeric: Rank, Sales, Profit, Assets, Market Value.

+ Các attribute Rank, Sales, Profit, Assets, Market Value thuộc dạng numeric có số liệu chi tiết và dùng để phân tích.

+ Thuộc tính Country gồm: 17 (thuộc tính).

**Năm 2022:**

- Các thông tin khai phá:  
+ Có 7 instance và bộ dữ liệu không missing.  
+ Attribute gồm: Rank, Name, Country, Sales, Profit, Assets, Market Value.

+ Các attribute thuộc dạng nominal: Name, Country.

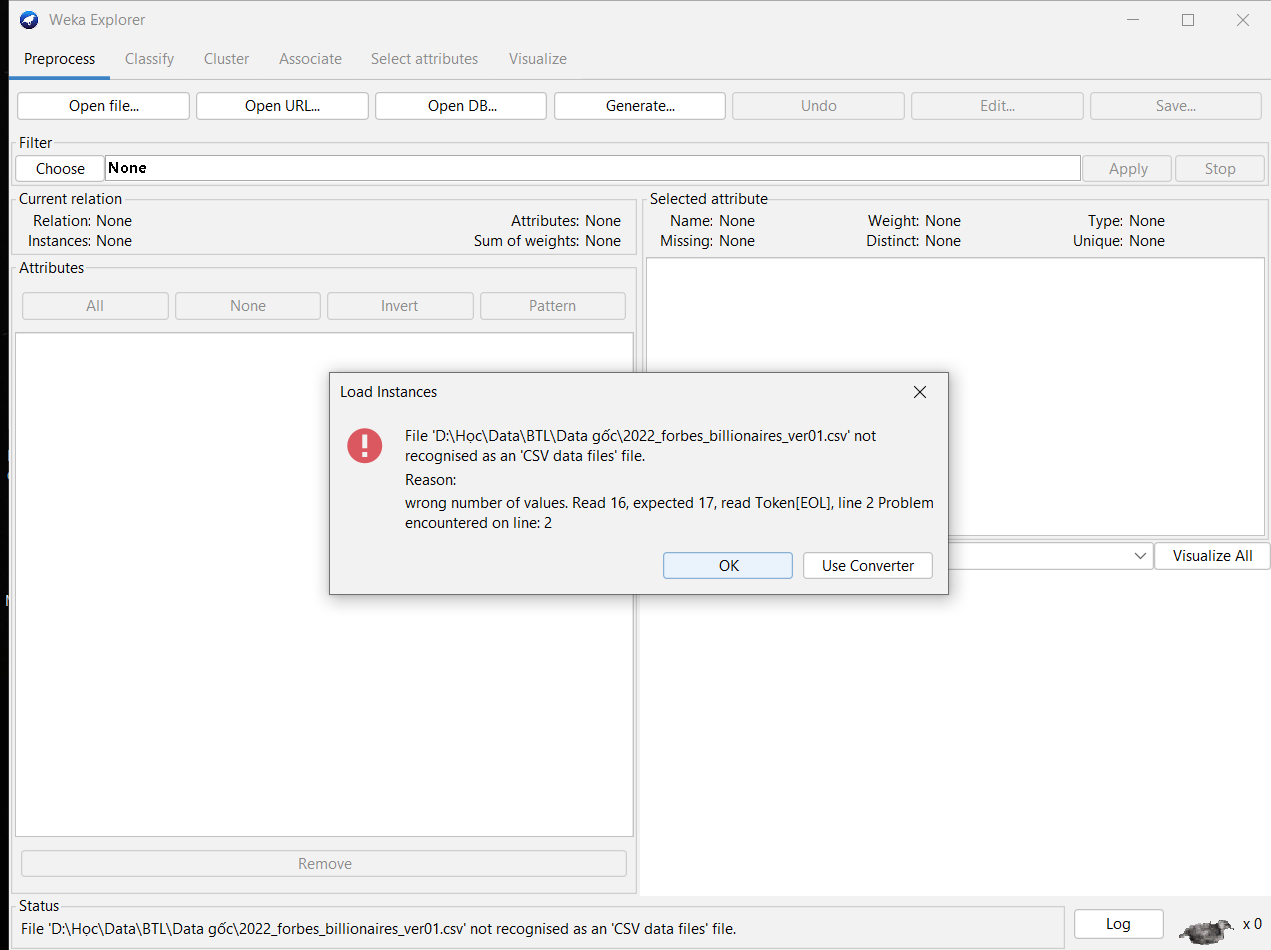
+ Các attribute thuộc dạng numeric: Rank, Sales, Profit, Assets, Market Value.

+ Các attribute Rank, Sales, Profit, Assets, Market Value thuộc dạng numeric có số liệu chi tiết và dùng để phân tích.

+ Thuộc tính Country gồm: 20 (thuộc tính).

## Làm sạch dữ liệu

- Với dữ liệu nguyên bản thì việc sai sót dữ liệu là không thể tránh khỏi. Điều này làm dữ liệu sau khi đưa vào ứng dụng Weka bị lỗi và ứng dụng sẽ không chạy.



Hình 1. 1: Dữ liệu đưa vào Weka khi chưa xử lý bị lỗi

- Vậy nên chúng ta cần phải làm sạch dữ liệu để khi đưa vào các mô hình trong KPDL sẽ không bị lỗi nữa. Công đoạn làm sạch dữ liệu cực kì quan trọng đối với tất cả các mô hình vì chỉ khi có dữ liệu sạch thì mô hình mới hoạt động một cách tốt nhất, đem về hiệu quả cao nhất.

**Cách làm sạch dữ liệu:**

**\*Data Billionaires:**

- Ghi hoa tất cả chữ cái đầu mỗi chữ.

- Chỉnh lại số thứ tự rank.

- Cột NetWorth:

+ Million -> Billion.

- Cột Name:

+ Bỏ dấu ','.

- Thay 'Rodolphe Saadã© & family' -> 'Rodolphe Saade'.

- Thay 'Franã§Ois Pinault & family'-> 'Francois Pinault & family'.

- Thay 'Beate Heister & Karl Albrecht Jr. & family'-> 'Karl Albrecht Jr. & family'.

- Cột Age:

+ Thay dòng 36 có dữ liệu trống -> Mean Age.

- Cột Global Company:

+ Thay 'L'Orã©Al -> LOréal'.

- Cột Country:

+ Thay dòng 67 có dữ liệu trống -> 'Switzerland'.

**\*Data Companies:**

- Bỏ $ và B ở các cột Sales, Profit, Assets, Market Value và thêm cột rank cho năm 2021.

**Năm 2020:**

- Thay 'NestlÃ©' -> 'Nestlé'.

- Thay 'LVMH MoÃ«t Hennessy Louis Vuitton' -> 'LVMH'.

- Thay 'ItaÃº Unianco Holding' -> 'Itaú Unibanco Holding'.

**Năm 2021:**

- Thay 'LVMH MoÃ«t Hennessy Louis Vuitton' -> 'LVMH'.

**Năm 2022:**

- Thay 'NestlÃ©' -> 'Nestlé'.

- Thay 'LVMH MoÃ«t Hennessy Louis Vuitton' -> 'LVMH'.

# GOM CỤM DỮ LIỆU

## Lý thuyết gom cụm

- Gom cụm dữ liệu là bài toán gom nhóm các đối tượng dữ liệu vào thành từng cụm (cluster) sao cho các đối tượng trong cùng một cụm có sự tương đồng theo một tiêu chí [3] nào đó được đặt ra trước.

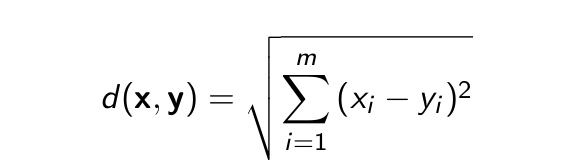
- Chia dữ liệu thành K nhóm (K là bao nhiêu cần test và chọn), sau khi có kết quả gom cụm: tìm hiểu các phần tử trong từng cụm để phát hiện đặc điểm của cụm. Qua đó khi ta có được một hoặc nhiều dữ liệu mới thì dựa vào sự tương đồng mà sắp xếp dữ liệu đó thuộc về cụm nào.

### Các độ đo khoảng cách cụm

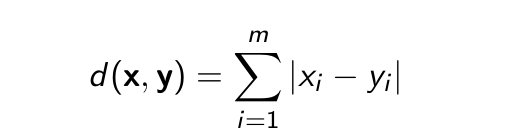
- Tính khoảng cách của cụm: có 2 cách đo khoảng cách phổ biến nhất là độ đo Euclid và độ đo Manhattan.

- Cho hai điểm x = (x1, x2, ..., xm) và y = (y1, y2, ..., ym) và các độ đo được xác định theo công thức:

+ Độ đo Euclid:



+ Độ đo Manhattan:

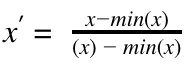


### Chuẩn hóa dữ liệu

- Trước khi vào tính khoảng cách của dữ liệu thì những attribute thuộc dạng numeric sẽ phải cần được chuẩn hóa hoặc rời rạc hóa dữ liệu. Điều này làm cho khoảng cách của dữ liệu không bị quá xa gây nhiễu loạn vị trí và không thể xác định được khoảng cách khi đưa vào mô hình.

- Thường chúng ta sẽ sử dụng phương pháp chuẩn hóa hoặc rời rạc hóa dữ liệu:

+ Chuẩn hóa: đây là phương pháp biến các số từ giá trị cao về giá trị có phạm vi từ [0,1] hoặc từ [-1,1] thông qua công thức



+ Rời rạc hóa: ở phương pháp này có 2 cách phổ biến để rời rạc là rời rạc bằng thuật toán Equal-Width và rời rạc bằng thuật toán Equal- Frequency.

## Gom cụm áp dụng thuật toán K-means

**\*Data Billionaires:**

- Đầu tiên chuẩn hóa dữ liệu của cột Tuổi và Tài sản về phạm vi từ [0,1].

- Sau đó sử dụng phần mềm Weka để gom cụm bằng thuật toán K-means với số cụm là 7, gom cụm dựa theo tiêu chí là Age – Net Worth và cho ra kết quả tương đối như sau:

**Cluster 0:** đa số là các tỷ phú ở các quốc gia Châu Âu là nam, có sự hậu thuẫn từ người thân, làm trong ngành Kinh doanh hàng hóa thời trang, có độ tuổi trung bình là 72, có tài sản trung bình là $47.4B.

**Cluster 1:** đa số là các tỷ phú ở Mỹ và Trung Quốc là nam, tự lập, làm trong ngành Công nghệ thông tin, có độ tuổi trung bình là 62, có tài sản trung bình là $34.2B.

**Cluster 2:** đa số là các tỷ phú ở Trung Quốc là nam, tự lập, làm trong ngành Công nghệ Ô tô, có độ tuổi trung bình là 54, có tài sản trung bình là $53.2B.

**Cluster 3:** đa số là các tỷ phú ở Mỹ và Ấn Độ là nam, tự lập, làm Đa dạng ngành, có độ tuổi trung bình là 73, có tài sản trung bình là $66.4B.

**Cluster 4:** đa số là các tỷ phú ở Mỹ là nam, tự lập, làm trong ngành Công nghệ thông tin, có độ tuổi trung bình là 61, có tài sản trung bình là $119.2B.

**Cluster 5:** đa số là các tỷ phú ở các quốc gia Châu Á là nam, tự lập, làm trong ngành dịch vụ Ăn Uống, Khai thác Mỏ và Kim loại, có độ tuổi trung bình là 69, có tài sản trung bình là $23.2B.

**Cluster 6:** đa số là các tỷ phú ở Mỹ là nam, tự lập, làm trong ngành Đầu tư tài chính, có độ tuổi trung bình là 65, có tài sản trung bình là $34.5B.

Hình 2. 1: Biểu đồ gom cụm dữ liệu tỷ phú

***\*Nhận xét***: từ biểu đồ và các cụm đã phân tích bên trên ta có thể nhìn thấy được ngành Công nghệ thông tin vẫn đang dẫn đầu và phát triển nhất trong tất cả các ngành trong mục tài sản của tỷ phú, thường thì ở cột này trung bình cao gấp đôi hoặc gấp 3 thậm chí là gấp 4 các ngành khác trong cụm mặc dù tỷ phú ở cụm này ít nhất.

**\*Data Companies:**

**-** Chuẩn hóa tất cả các cột Sales, Profit, Assets và Market Value về phạm vi [0,1].

**Năm 2020:**

- Sử dụng phần mềm Weka để gom cụm bằng thuật toán K-means với số cụm là 3, gom cụm dựa theo tiêu chí là Country – Profit và cho ra kết quả tương đối như sau:

**Cluster 0 (42%):** đa phần các công ty ở Mỹ có Doanh thu trung bình là $144B, Lợi nhuận trung bình là $17.87B, tổng Tài sản trung bình là $490.6B, tổng Giá trị cổ phiếu vào trung bình là $308.4B.

**Cluster 1 (37%):** đa phần các công ty ở các quốc gia khác có Doanh thu trung bình là $81.14B, Lợi nhuận trung bình là $7.45B, tổng Tài sản trung bình là $562.3B, tổng Giá trị cổ phiếu trung bình là $84.3B.

**Cluster 2 (21%):** đa phần các công ty ở Trung Quốc có tài sản Doanh thu là $124.6B, Lợi nhuận trung bình là $14.4B, tổng Tài sản trung bình là $1536.7B, tổng Giá trị cổ phiếu trung bình là $130.2B.

Hình 2. 2: Biểu đồ gom cụm dữ liệu công ty năm 2020

***\*Nhận xét***:

Mặc dù các công ty ở Mỹ đều vượt qua Trung quốc và các quốc gia khác về mặt Doanh thu, Lợi nhuận và tổng Giá trị cổ phiếu nhưng về mặt Tài sản thì Trung Quốc lại bỏ xa Mỹ. Vì sao lại có vấn đề này? Lý do chính là năm 2020, giá trị “bất động sản” tăng rất cao. Đây là lần thứ 2 Mỹ bị soán ngôi sau 30 năm.

**Năm 2021:**

- Sử dụng phần mềm Weka để gom cụm bằng thuật toán K-means với số cụm là 2, gom cụm dựa theo tiêu chí là Country – Assets và cho ra kết quả tương đối như sau:

**Cluster 0 (26%):** đa phần các công ty ở Trung Quốc có Doanh thu trung bình là $110B, Lợi nhuận trung bình là $13.34B, tổng Tài sản trung bình là $1839B, tổng Giá trị cổ phiếu vào trung bình là $149B.

**Cluster 1 (74%):** đa phần các công ty ở Mỹ có Doanh thu trung bình là $108B, Lợi nhuận trung bình là $12.63B, tổng Tài sản trung bình là $453B, tổng Giá trị cổ phiếu trung bình là $307.4B.

Hình 2. 3: Biểu đồ gom cụm dữ liệu công ty năm 2021

***\*Nhận xét:***

Ở năm này thì dịch Covid 19 bắt đầu bùng phát mạnh ở Mỹ làm cho Doanh thu và Lợi nhuận bị thụt giảm có phần lép vế hơn so với Trung Quốc nhưng Mỹ vẫn dẫn trước về tổng Giá trị cổ phiếu còn Trung Quốc thì vẫn hơn về mặt Tài sản.

**Năm 2022:**

- Sử dụng phần mềm Weka để gom cụm bằng thuật toán K-means với số cụm là 2, gom cụm dựa theo tiêu chí là Country – Assets và cho ra kết quả tương đối như sau:

**Cluster 0 (22%):** đa phần các công ty ở Trung Quốc có Doanh thu trung bình là $136.28B, Lợi nhuận trung bình là $18.3B, tổng Tài sản trung bình là $2018.54B, tổng Giá trị cổ phiếu trung bình là $133.8B.

**Cluster 1 (78%):** đa phần các công ty ở Mỹ có Doanh thu trung bình là $136.28B, Lợi nhuận trung bình là $19.86B, tổng Tài sản trung bình là $483.76B, tổng Giá trị cổ phiếu trung bình là $304.30B.

Hình 2. 4: Biểu đồ gom cụm dữ liệu công ty năm 2022

***\*Nhận xét:***

Ở năm 2022 dịch bệnh ở Mỹ đã được kiểm soát tạm ổn nên nhìn vào biểu đồ ta thấy về Doanh thu và Lợi nhuận đã ngang bằng hoặc hơn Trung Quốc còn lại thì vẫn như năm 2021.

# MÔ HÌNH DỰ ĐOÁN SỬ DỤNG LUẬT KẾT HỢP

## Lý thuyết luật kết hợp

- Luật kết hợp hay còn gọi là Khai phá luật kết hợp - Association rule in data mining, là một kĩ thuật cực kì quan trọng trong KPDL. Với mục tiêu là nhằm tìm ra được những mối liên hệ từ các mục dữ liệu trong CSDL [4] có ý nghĩa nhất dựa vào ngưỡng độ hỗ trợ tối thiểu (minsup) và độ tin cậy tối thiểu (minconf).

- Ban đầu luật kết hợp là một bài toán tìm ra mối liên hệ từ việc A->B làm sao cho thỏa mãn sup(A->B) > minsup và conf(A->B) > minconf. Từ đó được chia ra làm 2 bài toán con là tìm các tập mục thường xuyên và tìm ra được luật từ các tập mục thường xuyên.

### Thuật toán Apriori khai phá luật kết hợp

- Thuật toán Apriori là một thuật toán nhằm gộp tất cả những phần tử sản phẩm chung trong 1 dãy phần tử dựa theo độ xuất hiện của sản phẩm và sau khi gộp qua n lần chúng ta có được những luật phổ biến để đưa ra tỷ lệ xuất hiện của việc tiếp theo.

- Thuật toán Apriori có mục đích dự đoán được tỷ lệ bao nhiêu phần trăm xuất hiện của một việc gì đó xảy ra tiếp theo dựa vào một việc gì đó trong quá khứ. Việc này giống như chúng ta có thể dựa vào số lượng mua phổ biến của nhiều món hàng của các khách hàng mua trong quá khứ để sắp xếp chúng gần lại với nhau [5] nhằm gia tăng giá trị hóa đơn.

## Áp dụng thuật toán Apriori qua ngôn ngữ lập trình

- Ở đây em chỉ quan tâm tới các luật, ngưỡng độ hỗ trợ và độ tin cậy còn những cột khác em không đào sâu vào trong đó có các cột:

+ antecedents: tiền thân hay còn gọi là sự việc xảy ra trước.

+ consequents: kết quả hay còn gọi là việc được dự đoán xảy ra.

+ support: độ hỗ trợ.

+ confidence: độ tin cậy.

**\*Data Billionaires:**

**-** Dưới đây là đoạn code Python trên Google Colab áp dụng thuật toán Apriori và ý nghĩa:

from numpy.lib.shape\_base import column\_stack

from mlxtend.frequent\_patterns import apriori

from mlxtend.frequent\_patterns import association\_rules

import pandas as pd

#Đọc CSDL từ file excel

df = pd.read\_excel('/content/drive/MyDrive/Data Mining/BTL/Bill.xlsx')

# Đoạn mã này sử dụng phương thức applymap() để duyệt qua từng phần tử trong DataFrame df.

# Mỗi phần tử x được kiểm tra xem có bằng 1 không. Nếu bằng 1, thì được chuyển thành True, ngược lại, chuyển thành False.

# Mục đích có thể là chuyển đổi dữ liệu nhị phân (có giá trị 1 và 0) thành dạng Boolean (True và False).

df = df.applymap(lambda x: True if x == 1 else False)

# Dòng này đơn giản là để hiển thị DataFrame df sau khi dữ liệu đã được chuyển đổi ở bước trước.

# Điều này có thể giúp bạn xác định liệu việc chuyển đổi đã được thực hiện đúng hay không.

df

data\_endcoded = df

# Thuật toán Apriori được sử dụng để tìm các tập hỗ trợ và luật kết hợp trong dữ liệu đã được chuyển đổi.

# min\_support=0.2 là ngưỡng hỗ trợ tối thiểu, tức là chúng ta chỉ quan tâm đến các tập mẫu thường xuyên xuất hiện ít nhất 20% trong dữ liệu.

# use\_colnames=True sử dụng tên cột thay vì chỉ số để biểu diễn các tập hỗ trợ và luật kết hợp.

frequency\_set=apriori(data\_endcoded,min\_support=0.2,use\_colnames=True)

# Dựa trên tập hỗ trợ và ngưỡng độ tin cậy tối thiểu là 80%, tạo ra các luật kết hợp từ dữ liệu đã được chuyển đổi.

rules = association\_rules(frequency\_set, min\_threshold=0.8)

print(rules)

# Đưa bảng có luật kết hợp vào trong file Billres.csv

rules.to\_csv("/content/drive/MyDrive/Data Mining/BTL/Billres.csv")

**-** Sau đó ta được bảng chứa tất cả các luật kết hợp như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| antecedents | consequents | support | confidence |
| frozenset({'NetWorth'}) | frozenset({'Country'}) | 0.26 | 0.866666667 |
| frozenset({'Country'}) | frozenset({'Gender'}) | 0.63 | 0.851351351 |
| frozenset({'Industry'}) | frozenset({'Gender'}) | 0.61 | 0.871428571 |
| frozenset({'Age'}) | frozenset({'Gender'}) | 0.42 | 0.857142857 |
| frozenset({'NetWorth'}) | frozenset({'Gender'}) | 0.26 | 0.866666667 |
| frozenset({'SelfMade'}) | frozenset({'Gender'}) | 0.62 | 0.984126984 |
| frozenset({'NetWorth', 'Industry'}) | frozenset({'Country'}) | 0.2 | 0.909090909 |
| frozenset({'Country', 'Industry'}) | frozenset({'Gender'}) | 0.46 | 0.867924528 |
| frozenset({'Country', 'Age'}) | frozenset({'Gender'}) | 0.28 | 0.848484848 |
| frozenset({'Country', 'NetWorth'}) | frozenset({'Gender'}) | 0.22 | 0.846153846 |
| frozenset({'NetWorth', 'Gender'}) | frozenset({'Country'}) | 0.22 | 0.846153846 |
| frozenset({'Country', 'SelfMade'}) | frozenset({'Gender'}) | 0.46 | 0.978723404 |
| frozenset({'Industry', 'Age'}) | frozenset({'Gender'}) | 0.26 | 0.8125 |
| frozenset({'Industry', 'SelfMade'}) | frozenset({'Gender'}) | 0.44 | 1 |
| frozenset({'Age', 'SelfMade'}) | frozenset({'Gender'}) | 0.24 | 1 |
| frozenset({'Country', 'Industry', 'SelfMade'}) | frozenset({'Gender'}) | 0.34 | 1 |

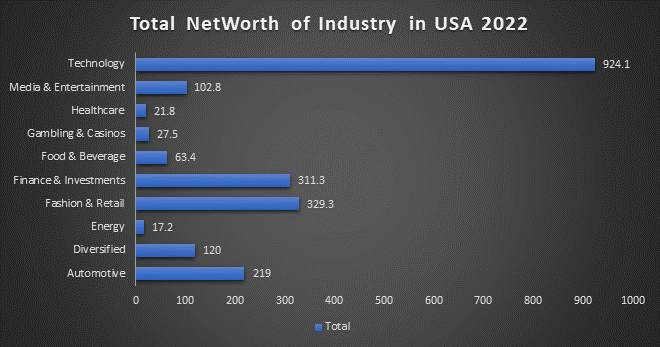
Bảng 1. 1: Bảng luật kết hợp dữ liệu tỷ phú

- Lựa chọn ra luật có ý nghĩa nhất:

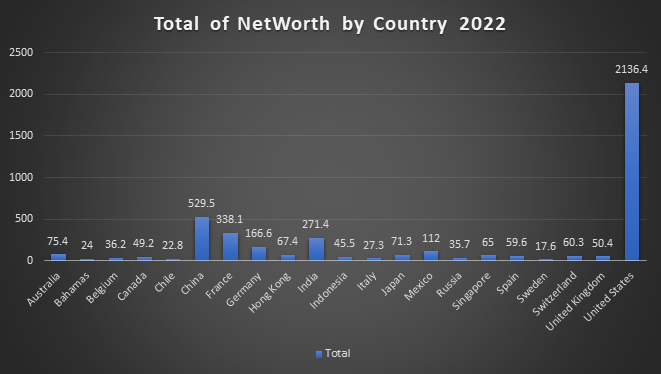
|  |  |  |  |
| --- | --- | --- | --- |
| antecedents | consequents | support | confidence |
| frozenset ({'NetWorth', 'Industry'}) | frozenset({'Country'}) | 0.2 | 0.909090909 |

Bảng 1. 2: Bảng luật kết hợp ý nghĩa dữ liệu tỷ phú

- Từ đây ta thấy được mối liên hệ giữa Ngành nghề (Industry) và Tài sản ròng (NetWorth) đối với Quốc gia (Country).



Hình 3. 1: Biểu đồ tổng tài sản theo ngành nghề năm 2022 ở Mỹ



Hình 3. 2: Biểu đồ tổng tài sản theo quốc gia năm 2022

***\*Nhận xét:***

Mỹ bị ảnh hưởng nhiều bởi Tài sản ròng của các tỷ phú trong ngành Công nghệ thông tin bởi vì các tỷ phú này đều lọt top 30, có lượng tài sản rất cao và qua đó cho ta thấy được ngành Công nghệ thông tin đang vượt trội hơn tất cả những ngành khác khi chiếm tới hơn 43% tổng tài sản tỷ phú ở Mỹ với lượng tài sản trung bình khoảng $84B và chiếm gần 70% tổng tài sản của ngành Công nghệ thông tin trong bộ dữ liệu.

**\*Data Companies:**

**Năm 2020:**

**-** Dưới đây là đoạn code Python trên Google Colab áp dụng thuật toán Apriori với minsup = 0,14 và minconf = 0,88:

from numpy.lib.shape\_base import column\_stack

from mlxtend.frequent\_patterns import apriori

from mlxtend.frequent\_patterns import association\_rules

import pandas as pd

df = pd.read\_excel('/content/drive/MyDrive/Data Mining/BTL/2020.xlsx')

df = df.applymap(lambda x: True if x == 1 else False)

df

data\_endcoded = df

frequency\_set=apriori(data\_endcoded,min\_support=0.14,use\_colnames=True)

rules = association\_rules(frequency\_set, min\_threshold=0.88)

print(rules)

rules.to\_csv("/content/drive/MyDrive/Data Mining/BTL/2020-res.csv")

**-** Sau đó ta được bảng chứa tất cả các luật kết hợp như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| antecedents | consequents | support | confidence |
| Frozenset ({'Market Value', 'Profit'}) | Frozenset ({'Country'}) | 0.18 | 0.9 |
| Frozenset ({'Market Value', 'Sales'}) | Frozenset ({'Profit'}) | 0.14 | 0.9333333333 |

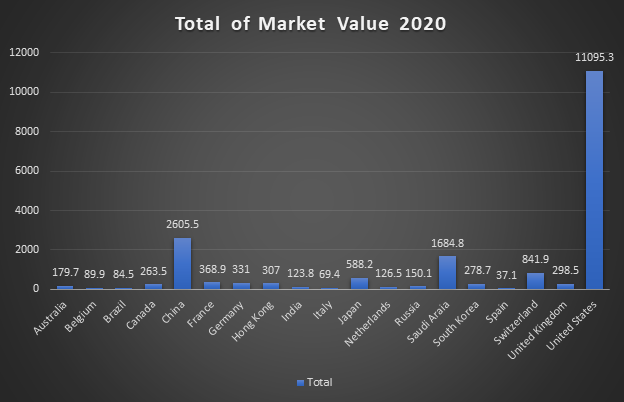
Bảng 2. 1: Bảng luật kết hợp dữ liệu công ty năm 2020

- Lựa chọn ra luật có ý nghĩa nhất:

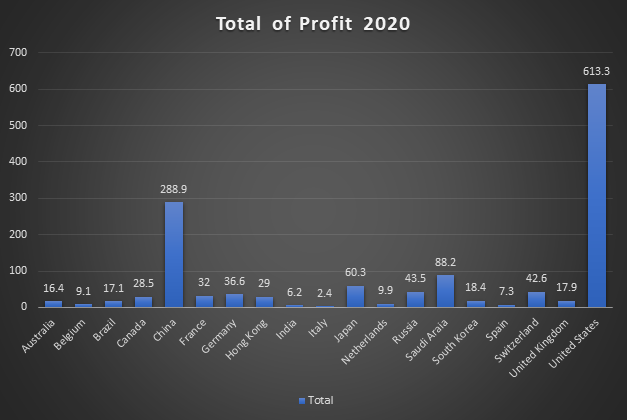
|  |  |  |  |
| --- | --- | --- | --- |
| antecedents | consequents | support | confidence |
| Frozenset ({'Market Value', 'Profit'}) | Frozenset ({'Country'}) | 0.18 | 0.9 |
| Frozenset ({'Market Value', 'Sales'}) | Frozenset ({'Profit'}) | 0.14 | 0.9333333333 |

Bảng 2. 2: Bảng luật kết hợp ý nghĩa dữ liệu công ty năm 2020

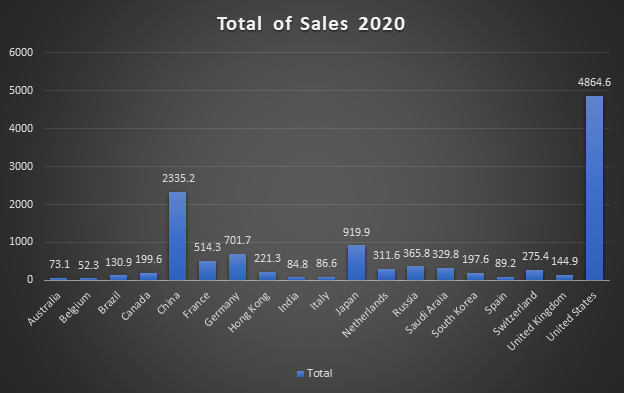
- Từ đây ta thấy được mối liên hệ giữa Tài sản cổ phiếu (Market Value) và Lợi nhuận (Profit) đối với Quốc gia (Country) và mối liên hệ giữa Tài sản cổ phiếu (Market Value) và Tổng kinh doanh (Sales) đối với Lợi nhuận (Profit).



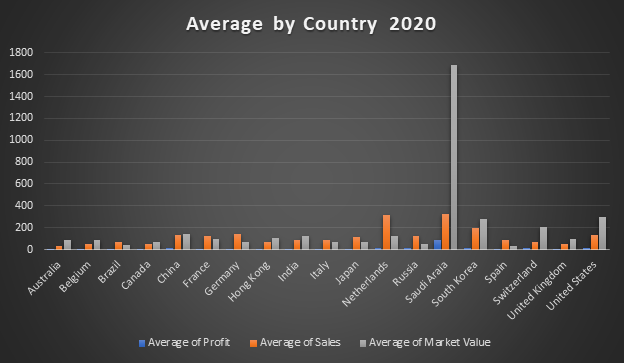
Hình 3. 3: Biểu đồ tổng tài sản cổ phiếu theo quốc gia năm 2020



Hình 3. 4: Biểu đồ tổng lợi nhuận theo quốc gia năm 2020



Hình 3. 5: Biểu đồ tổng kinh doanh theo quốc gia năm 2020



Hình 3. 6: Biểu đồ trung bình giá trị theo quốc gia năm 2020

***\*Nhận xét:***

Tổng Giá trị cổ phiếu, Lợi nhuận và Doanh thu ở Mỹ lần lượt là: $11095.3B chiếm tới hơn 57%; $613.3B chiếm tới gần 45% và $4864.6B chiếm tới hơn 40% trong bộ dữ liệu.

Còn ở Trung Quốc tổng Giá trị cổ phiếu, Lợi nhuận và Doanh thu lần lượt là: $2605.5B chiếm tới hơn 13%; $288.9B chiếm tới hơn 21% và $2335.2B chiếm tới gần 20% trong bộ dữ liệu.

**Năm 2021:**

**-** Dưới đây là đoạn code Python trên Google Colab áp dụng thuật toán Apriori với minsup = 0,15 và minconf = 0,8:

from numpy.lib.shape\_base import column\_stack

from mlxtend.frequent\_patterns import apriori

from mlxtend.frequent\_patterns import association\_rules

import pandas as pd

df = pd.read\_excel('/content/drive/MyDrive/Data Mining/BTL/2021.xlsx')

df = df.applymap(lambda x: True if x == 1 else False)

df

data\_endcoded = df

frequency\_set=apriori(data\_endcoded,min\_support=0.15,use\_colnames=True)

rules = association\_rules(frequency\_set, min\_threshold=0.8)

print(rules)

rules.to\_csv("/content/drive/MyDrive/Data Mining/BTL/2021-res.csv")

**-** Sau đó ta được bảng chứa tất cả các luật kết hợp như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| antecedents | consequents | support | confidence |
| frozenset({'Sales'}) | frozenset({'Country'}) | 0.31 | 0.861111111 |
| frozenset({'Profit'}) | frozenset({'Country'}) | 0.28 | 0.8 |
| frozenset({'Market Value'}) | frozenset({'Profit'}) | 0.2 | 0.952380952 |
| frozenset({'Profit', 'Sales'}) | frozenset({'Country'}) | 0.16 | 0.842105263 |
| frozenset({'Country', 'Market Value'}) | frozenset({'Profit'}) | 0.15 | 1 |

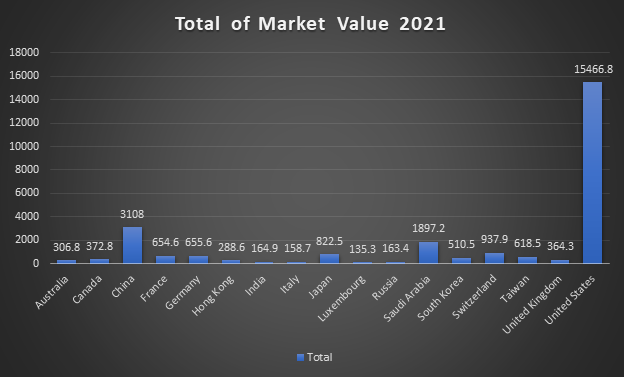
Bảng 3. 1: Bảng luật kết hợp dữ liệu công ty năm 2021

- Lựa chọn ra luật có ý nghĩa nhất:

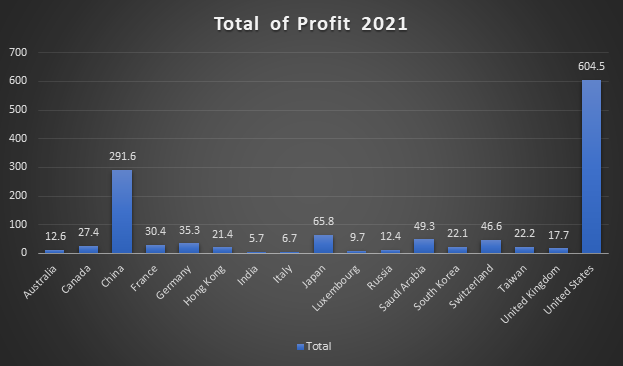
|  |  |  |  |
| --- | --- | --- | --- |
| antecedents | consequents | support | confidence |
| frozenset ({'Profit', 'Sales'}) | frozenset ({'Country'}) | 0.19 | 0.7 |
| frozenset ({'Country', 'Market Value'}) | frozenset ({'Profit'}) | 0.15 | 0.35 |

Bảng 3. 2: Bảng luật kết hợp ý nghĩa dữ liệu công ty năm 2021

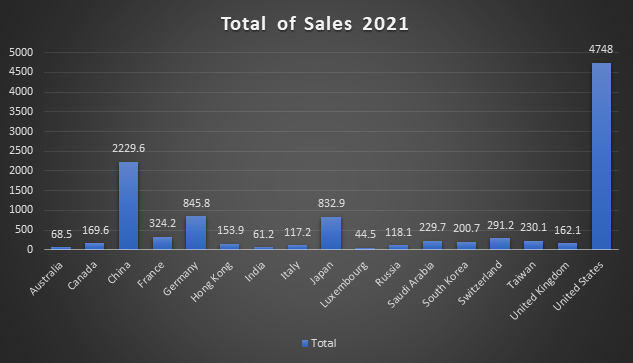
- Từ đây ta thấy được mối liên hệ giữa Lợi nhuận (Profit) và Tổng kinh doanh (Sales) đối với Quốc gia (Country) và mối liên hệ giữa Tài sản cổ phiếu (Market Value) và Quốc gia (Country) đối với Lợi nhuận (Profit).



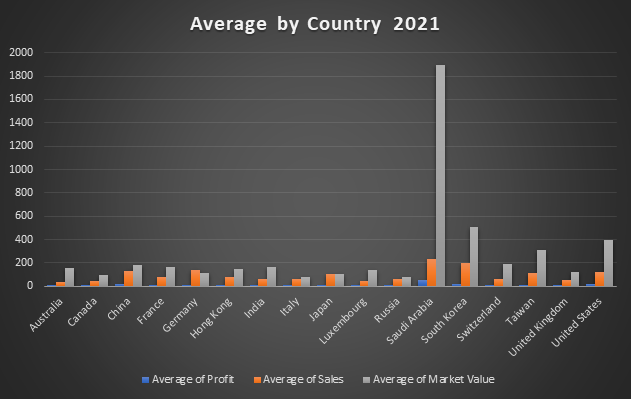
Hình 3. 7: Biểu đồ tổng tài sản cổ phiếu theo quốc gia năm 2021



Hình 3. 8: Biểu đồ tổng lợi nhuận theo quốc gia năm 2021



Hình 3. 9: Biểu đồ tổng kinh doanh theo quốc gia năm 2021



Hình 3. 10: Biểu đồ trung bình giá trị theo quốc gia năm 2021

***\*Nhận xét:***

Tổng Giá trị cổ phiếu, Lợi nhuận và Doanh thu ở Mỹ lần lượt là: $15466.8B chiếm tới hơn 58%; $604.5B chiếm tới hơn 47% và $4748B chiếm tới gần 44% trong bộ dữ liệu.

Còn ở Trung Quốc tổng Giá trị cổ phiếu, Lợi nhuận và Doanh thu lần lượt là: $3108B chiếm tới hơn 11%; $291.6B chiếm tới gần 23% và $2229.6B chiếm tới hơn 20% trong bộ dữ liệu.

**Năm 2022:**

**-** Dưới đây là đoạn code Python trên Google Colab áp dụng thuật toán Apriori với minsup = 0,15 và minconf = 0,88:

from numpy.lib.shape\_base import column\_stack

from mlxtend.frequent\_patterns import apriori

from mlxtend.frequent\_patterns import association\_rules

import pandas as pd

df = pd.read\_excel('/content/drive/MyDrive/Data Mining/BTL/2022.xlsx')

df = df.applymap(lambda x: True if x == 1 else False)

df

data\_endcoded = df

frequency\_set=apriori(data\_endcoded,min\_support=0.15,use\_colnames=True)

rules = association\_rules(frequency\_set, min\_threshold=0.7)

print(rules)

rules.to\_csv("/content/drive/MyDrive/Data Mining/BTL/2022-res.csv")

**-** Sau đó ta được bảng chứa tất cả các luật kết hợp như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| antecedents | consequents | support | confidence |
| frozenset ({'Sales'}) | frozenset ({'Country'}) | 0.26 | 0.7647058824 |
| frozenset ({'Profit'}) | frozenset ({'Country'}) | 0.21 | 0.7 |
| frozenset ({'Market Value'}) | frozenset ({'Country'}) | 0.2 | 0.8333333333 |

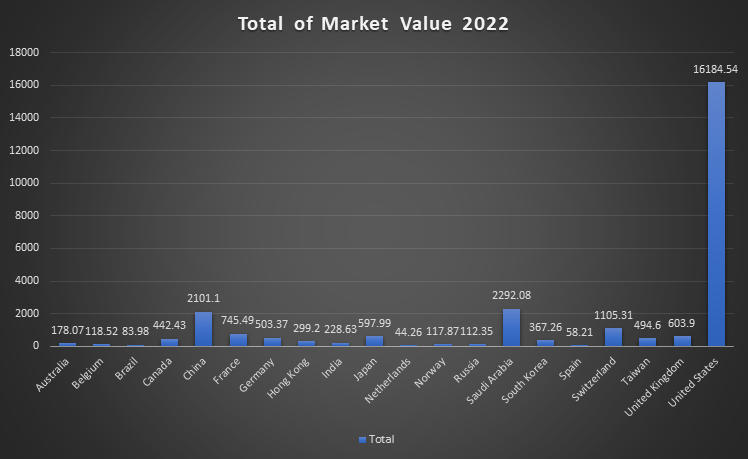
Bảng 4. 1: Bảng luật kết hợp dữ liệu công ty năm 2022

- Lựa chọn ra luật có ý nghĩa nhất:

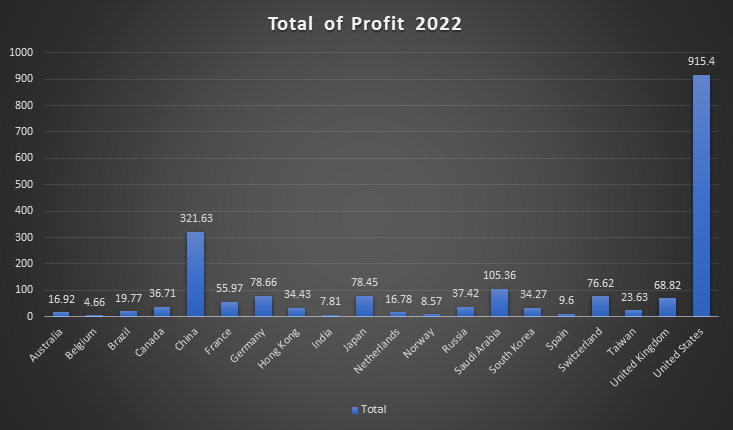
|  |  |  |  |
| --- | --- | --- | --- |
| antecedents | consequents | support | confidence |
| frozenset ({'Sales'}) | frozenset ({'Country'}) | 0.26 | 0.7647058824 |
| frozenset ({'Profit'}) | frozenset ({'Country'}) | 0.21 | 0.7 |
| frozenset ({'Market Value'}) | frozenset ({'Country'}) | 0.2 | 0.8333333333 |

Bảng 4. 2: Bảng luật kết hợp ý nghĩa dữ liệu công ty năm 2022

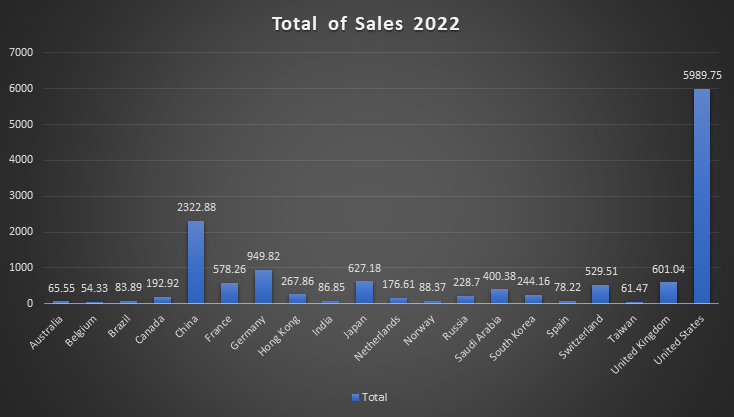
- Từ đây ta thấy được mối liên hệ giữa Tổng kinh doanh (Sales), Lợi nhuận (Profit) và Tài sản cổ phiếu (Market Value) đối với Quốc gia (Country).



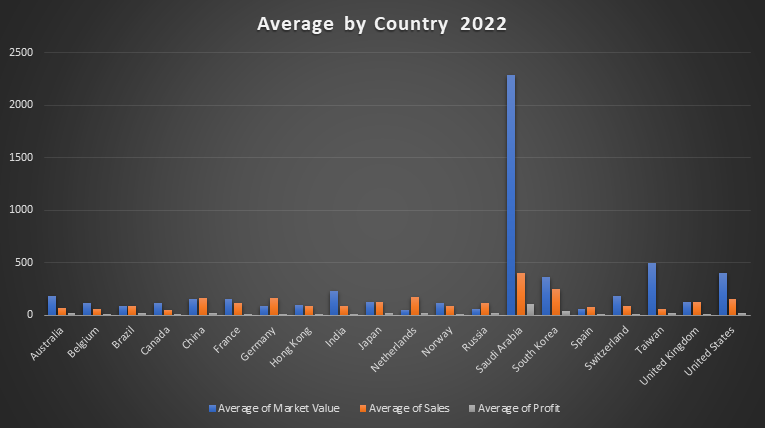
Hình 3. 11: Biểu đồ tổng tài sản cổ phiếu theo quốc gia năm 2022



Hình 3. 12: Biểu đồ tổng lợi nhuận theo quốc gia năm 2022



Hình 3. 13: Biểu đồ tổng kinh doanh theo quốc gia năm 2022



Hình 3. 14: Biểu đồ trung bình giá trị theo quốc gia năm 2022

***\*Nhận xét:***

Ở năm 2022 khi tìm ra được luật Doanh thu cao sẽ tác động tới quốc gia thì ngoài 2 nước Mỹ và Trung Quốc ra ta còn có thêm nước Đức điều này cho thấy Đức cũng đang thúc đẩy nền kinh tế của mình vượt bậc khi những công ty có Doanh thu cao đều lọt top 45.

Mặc dù các công ty ở Trung Quốc đều lọt top 45 nhưng nhìn chung đã bị thụt giảm đáng kể so với 2 năm trước về mặt tổng Giá trị cổ phiếu khoảng 32%.

Về Lợi nhuận và Giá cổ phiếu thì vẫn như 2 năm trước thì đa phần tác động tới 2 nước là Mỹ và Trung Quốc.

Tổng Giá trị cổ phiếu, Lợi nhuận và Doanh thu ở Mỹ lần lượt là: $ 16184.54B chiếm tới hơn 60%; $ 915.4B chiếm tới gần 47% và $5989.75B chiếm tới gần 44% trong bộ dữ liệu.

Còn ở Trung Quốc tổng Giá trị cổ phiếu, Lợi nhuận và Doanh thu lần lượt là: $2101.1B chiếm tới gần 8%; $321.63B chiếm tới hơn 16% và $2322.88B chiếm tới hơn 17% trong bộ dữ liệu.

Và ở Đức tổng Doanh thu là $949.82B chiếm khoảng 7%.

***\*Nhận xét chung:***

Quốc gia có Giá trị cổ phiếu, Lợi nhuận và Doanh thu tác động phần lớn là Mỹ và Trung Quốc bởi vì nền kinh tế 2 nước này vẫn đang ở top đầu mặc dù trong năm 2020 chịu sự ảnh hưởng rất lớn từ đại dịch Covid-19.

Đa số những Quốc gia có Doanh thu và Giá cổ phiếu cao thì sẽ có Lợi nhuận cao như Mỹ và Trung Quốc.

Ở Trung Quốc (chỉ một vài công ty) mặc dù không vượt trội hơn Mỹ về Doanh thu, Lợi nhuận và Giá trị cổ phiếu nhưng về lượng Tài sản lại vượt bậc nên vẫn có sự ảnh hưởng tới nền kinh tế cả nước và nằm trong top của bộ dữ liệu.

Nhìn vào biểu đồ Trung bình ta thấy được quốc gia Saudi Arabia cực kì vượt trội về mặt tổng Giá trị cổ phiếu thị trường. Lý do bởi vì Saudi Arabia chỉ có 1 công ty duy nhất lọt vào top đó là Saudi Arabian Oil Company (Saudi Aramco), đây là Công ty Dầu Ả Rập Xê Út và cũng là công ty có giá trị thị trường đứng thứ 2 trên thế giới sau Apple với mức độ khai thác dầu mỏ cả trên bờ lẫn trên biển lớn nhất thế giới.

# PHÂN LỚP BẰNG CÂY QUYẾT ĐỊNH

## Lý thuyết phân lớp

- Phân loại dữ liệu là một kỹ thuật dựa trên việc sử dụng tập huấn luyện và các giá trị hoặc nhãn của lớp trong một thuộc tính phân loại, với mục tiêu sử dụng nó để phân loại dữ liệu mới. Nó cũng đồng nghĩa với việc dự đoán lớp của các mẫu dữ liệu. Bên cạnh phân loại, chúng ta cũng có một phương pháp tương tự là dự đoán, với sự khác biệt là phân loại liên quan đến việc dự đoán lớp của nhãn, trong khi dự đoán sẽ tạo ra các hàm đánh giá liên tục.

- Kỹ thuật phân loại thực hiện qua hai giai đoạn: Xây dựng mô hình và Sử dụng mô hình. Giai đoạn Xây dựng mô hình mô tả một tập các lớp đã được xác định trước, trong đó mỗi mẫu hoặc bộ dữ liệu được gán cho một lớp cụ thể dựa trên thuộc tính nhãn lớp. Tập hợp các bộ dữ liệu được sử dụng trong giai đoạn này được gọi là tập huấn luyện. Mô hình có thể được biểu diễn bằng các luật phân loại, cây quyết định hoặc các phương trình toán học.

- Trong giai đoạn Sử dụng mô hình, mô hình được áp dụng để phân loại dữ liệu trong tương lai hoặc để phân loại cho các đối tượng mới mà chưa được biết đến trước đó. Trước khi sử dụng mô hình, chúng ta thường phải đánh giá tính chính xác của nó. Điều này bao gồm việc so sánh nhãn biết trước của các mẫu kiểm tra với kết quả phân loại của mô hình. Độ chính xác được đo bằng phần trăm của các mẫu kiểm tra được phân loại đúng bởi mô hình, với tập kiểm tra là độc lập với tập huấn luyện. Phân loại là một hình thức của học có giám sát, trong đó tập dữ liệu huấn luyện bao gồm quan sát và nhãn lớp đã được xác định, và việc phân loại dữ liệu mới dựa trên thông tin trong tập huấn luyện. Ngược lại, học không có giám sát không biết nhãn lớp của tập dữ liệu huấn luyện [6].

### Phân lớp với cây quyết định

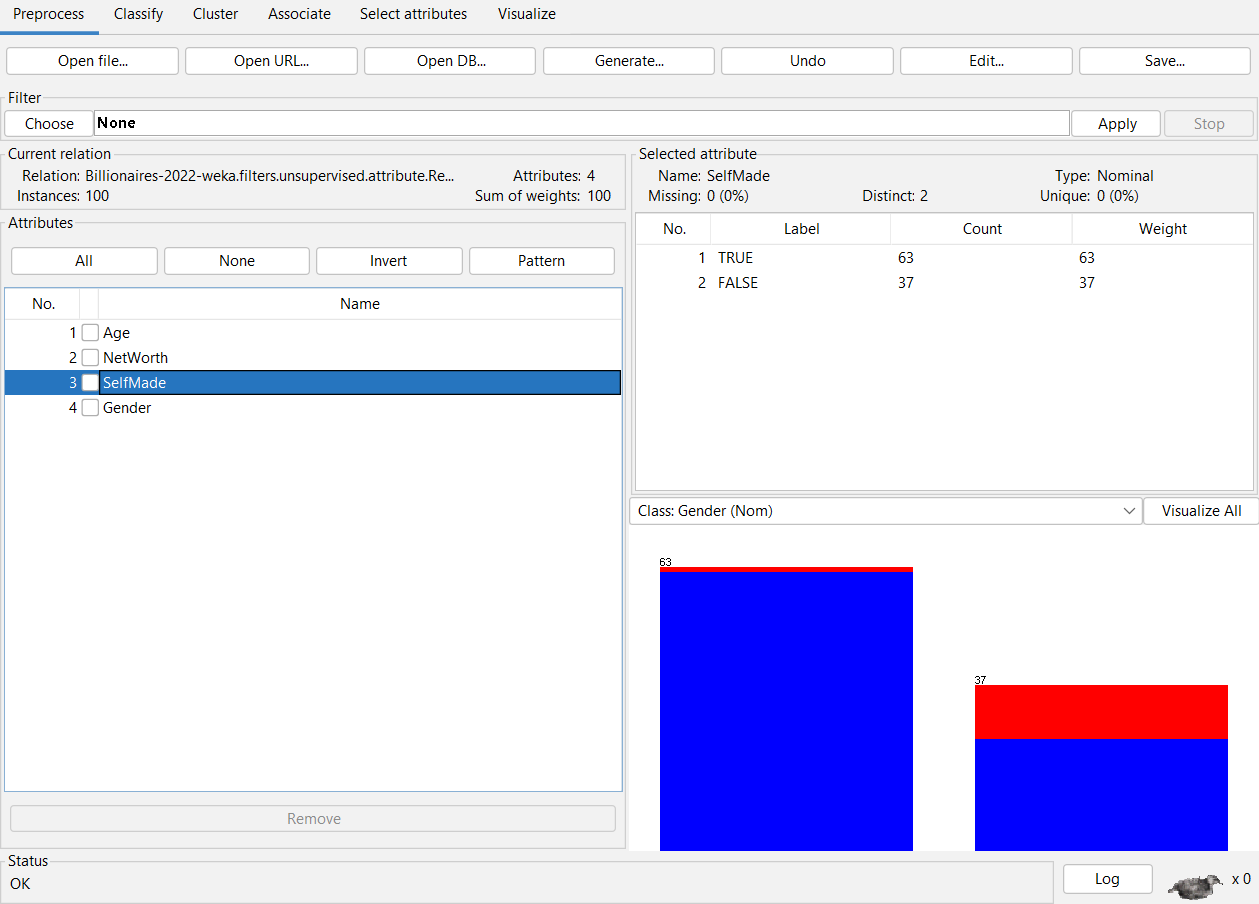
- Cây quyết định, một loại cấu trúc phân cấp, và nó được hoạt động bằng cách xác định lớp hoặc đầu ra của các đối tượng dựa trên một loạt các luật. Khi nó phân tích dữ liệu, nó không quan tâm đến loại thuộc tính của đối tượng, có thể là binary, nominal, ordinal hoặc quantitative. Nhưng đối với thuộc tính phân lớp, cây quyết định yêu cầu chúng phải thuộc loại binary hoặc ordinal.

- Đơn giản mà nói, khi chúng ta đưa vào mô hình một tập dữ liệu chứa các đối tượng với các thuộc tính và các lớp tương ứng của chúng, cây quyết định sẽ làm nhiệm vụ tạo ra những quy tắc dự đoán cho các đối tượng chưa biết lớp của chúng.

## Áp dụng phân lớp trong KPDL

**\*Data Billionaires:**

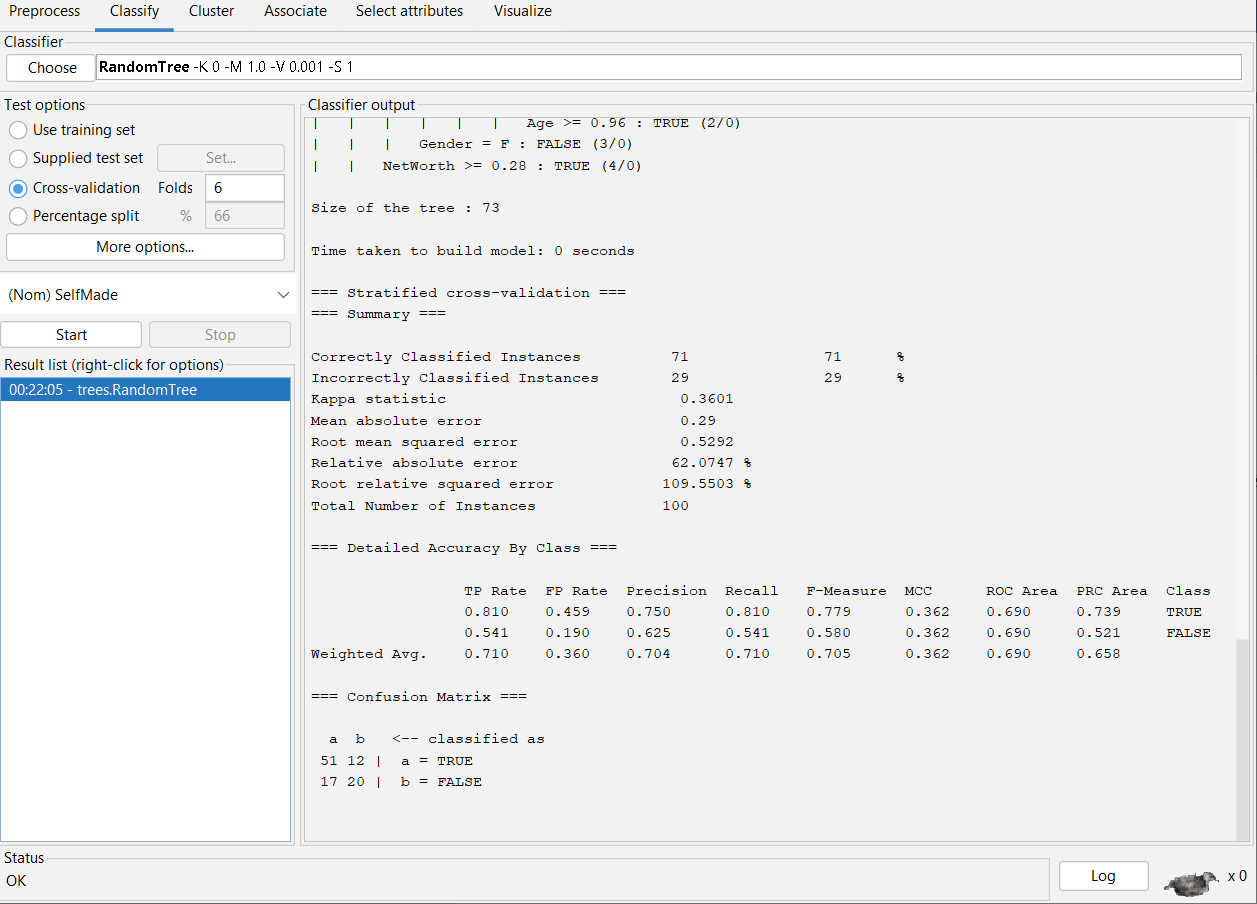
**-** Ở dữ liệu này chúng ta chỉ quan tâm tới việc tỷ phú có thuộc lớp dạng thừa kế hay tự lập (SelfMade) dựa vào việc phân lớp về Độ tuổi (Age), Tài sản ròng (NetWorth) và Giới tính (Gender).



Hình 4. 1: Khái quát đầu vào dữ liệu tỷ phú trong phân lớp

- Nhìn vào dữ liệu ta thấy đa phần các tỷ phú sẽ thuộc về Tự lập (True) với sai số rất ít còn về Thừa kế (False) tuy có sai số khá đáng kể nhưng không làm ảnh hưởng tới phần lớp Tự lập.

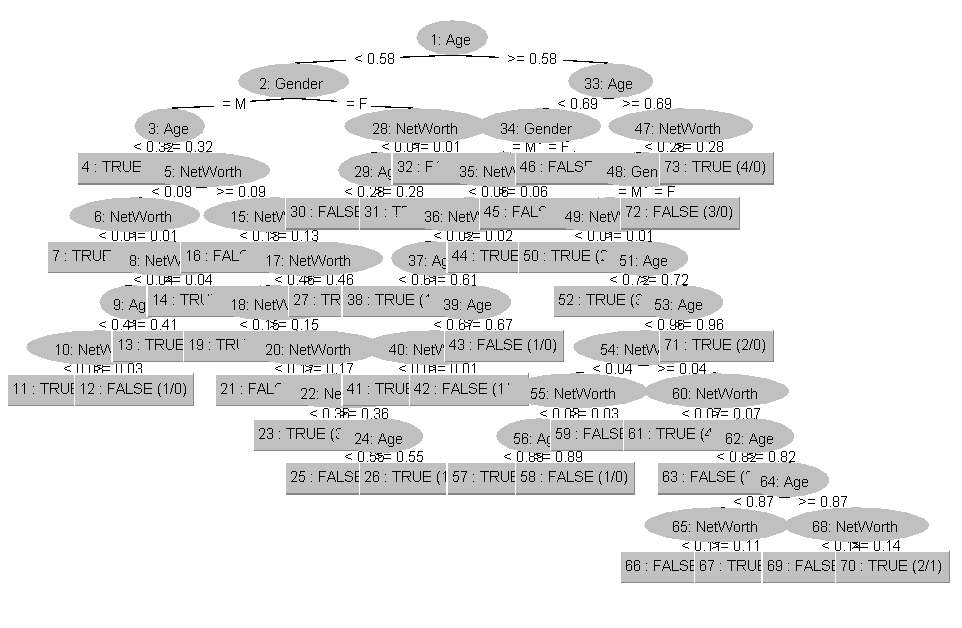
**-** Sử dụng thuật toán cây quyết định RandomTree với K Cross-validation = 6 để xét tỷ phú thuộc dạng tỷ phú tự lập hay tỷ phú thừa kế với Tự lập (True), Thừa kế (False):



Hình 4. 2: Mô tả kết quả sau khi chạy mô hình phân lớp dữ liệu tỷ phú

**-** Ở trong dữ liệu sau khi huấn luyện ta thấy với K = 6 thì mô hình có tỷ lệ đúng tạm ổn là 71%.

- Nhìn vào MTSS – Confusion Matrix ta thấy việc phân lớp vào lớp Tự lập có độ chính xác cao hơn rất nhiều bởi vì trong top 100 tỷ phú thì hết 63 người là tỷ phú tự lập. Vậy nên em huấn luyện mô hình lấy độ chính xác về lớp Tự lập sẽ cao hơn so với lớp Thừa kế.



Hình 4. 3: Cây quyết định dữ liệu tỷ phú

- Chi tiết cây quyết định:

Age < 0.58

| Gender = M

| | Age < 0.32 : TRUE (10/0)

| | Age >= 0.32

| | | NetWorth < 0.09

| | | | NetWorth < 0.01 : TRUE (9/0)

| | | | NetWorth >= 0.01

| | | | | NetWorth < 0.04

| | | | | | Age < 0.41

| | | | | | | NetWorth < 0.03 : TRUE (2/1)

| | | | | | | NetWorth >= 0.03 : FALSE (1/0)

| | | | | | Age >= 0.41 : TRUE (7/0)

| | | | | NetWorth >= 0.04 : TRUE (4/0)

| | | NetWorth >= 0.09

| | | | NetWorth < 0.13 : FALSE (3/0)

| | | | NetWorth >= 0.13

| | | | | NetWorth < 0.46

| | | | | | NetWorth < 0.15 : TRUE (1/0)

| | | | | | NetWorth >= 0.15

| | | | | | | NetWorth < 0.17 : FALSE (1/0)

| | | | | | | NetWorth >= 0.17

| | | | | | | | NetWorth < 0.36 : TRUE (3/0)

| | | | | | | | NetWorth >= 0.36

| | | | | | | | | Age < 0.55 : FALSE (1/0)

| | | | | | | | | Age >= 0.55 : TRUE (1/0)

| | | | | NetWorth >= 0.46 : TRUE (2/0)

| Gender = F

| | NetWorth < 0.01

| | | Age < 0.28 : FALSE (1/0)

| | | Age >= 0.28 : TRUE (1/0)

| | NetWorth >= 0.01 : FALSE (4/0)

Age >= 0.58

| Age < 0.69

| | Gender = M

| | | NetWorth < 0.06

| | | | NetWorth < 0.02

| | | | | Age < 0.61 : TRUE (1/0)

| | | | | Age >= 0.61

| | | | | | Age < 0.67

| | | | | | | NetWorth < 0.01 : TRUE (2/1)

| | | | | | | NetWorth >= 0.01 : FALSE (1/0)

| | | | | | Age >= 0.67 : FALSE (1/0)

| | | | NetWorth >= 0.02 : TRUE (2/0)

| | | NetWorth >= 0.06 : FALSE (5/0)

| | Gender = F : FALSE (4/0)

| Age >= 0.69

| | NetWorth < 0.28

| | | Gender = M

| | | | NetWorth < 0.01 : TRUE (2/0)

| | | | NetWorth >= 0.01

| | | | | Age < 0.72 : TRUE (3/0)

| | | | | Age >= 0.72

| | | | | | Age < 0.96

| | | | | | | NetWorth < 0.04

| | | | | | | | NetWorth < 0.03

| | | | | | | | | Age < 0.89 : TRUE (2/0)

| | | | | | | | | Age >= 0.89 : FALSE (1/0)

| | | | | | | | NetWorth >= 0.03 : FALSE (4/0)

| | | | | | | NetWorth >= 0.04

| | | | | | | | NetWorth < 0.07 : TRUE (4/0)

| | | | | | | | NetWorth >= 0.07

| | | | | | | | | Age < 0.82 : FALSE (2/0)

| | | | | | | | | Age >= 0.82

| | | | | | | | | | Age < 0.87

| | | | | | | | | | | NetWorth < 0.11 : FALSE (1/0)

| | | | | | | | | | | NetWorth >= 0.11 : TRUE (2/0)

| | | | | | | | | | Age >= 0.87

| | | | | | | | | | | NetWorth < 0.14 : FALSE (1/0)

| | | | | | | | | | | NetWorth >= 0.14 : TRUE (2/1)

| | | | | | Age >= 0.96 : TRUE (2/0)

| | | Gender = F : FALSE (3/0)

| | NetWorth >= 0.28 : TRUE (4/0)

***\*Nhận xét:***

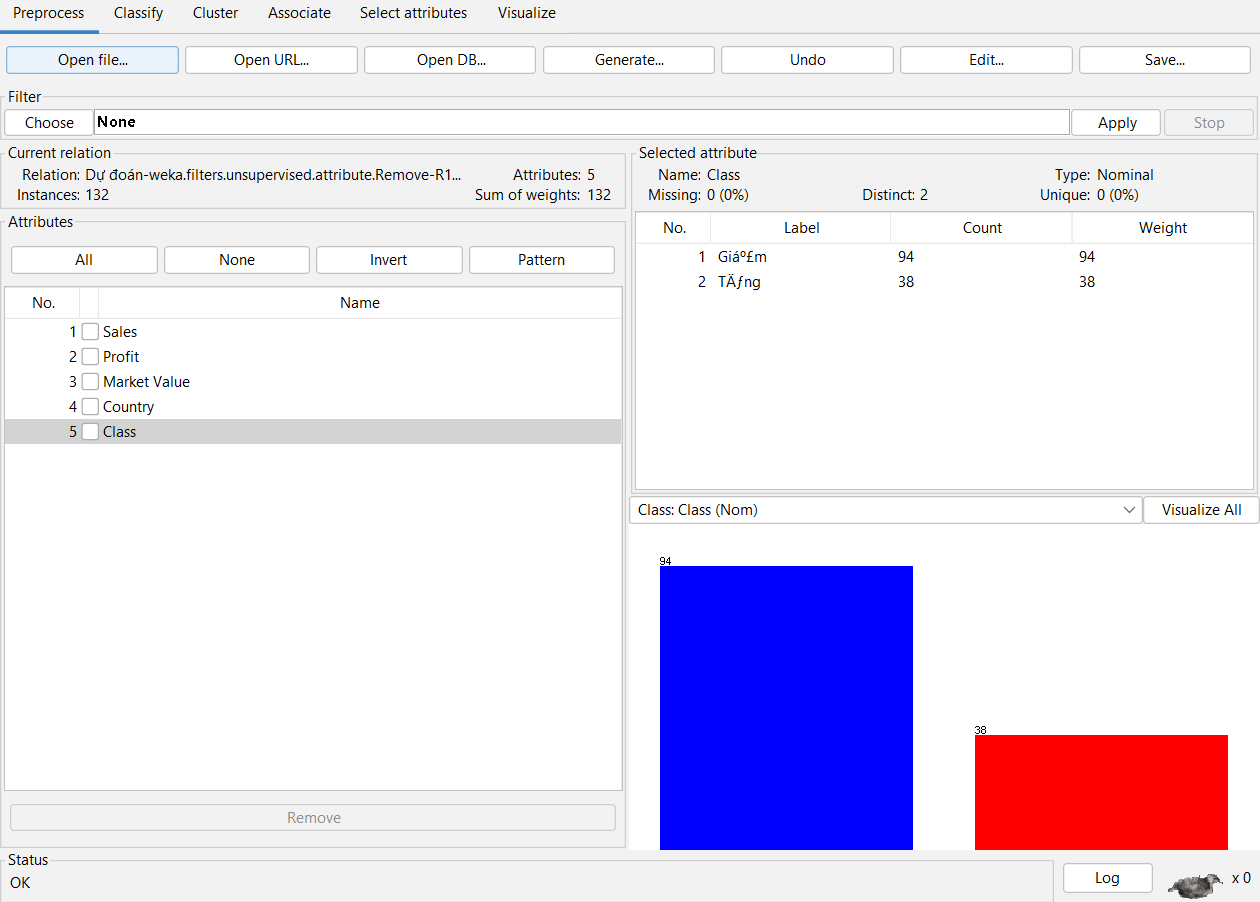
Độ tuổi của các tỷ phú có thể ảnh hưởng đến cách họ tự lập. Một số người có thể đạt được thành công tài chính ở tuổi trẻ thông qua khởi nghiệp, trong khi người khác có thể mất nhiều thập kỷ để tích lũy tài sản đủ lớn.

Giới tính cũng có vai trò quan trọng trong việc tự lập của các tỷ phú. Trong quá khứ, các tỷ phú có giới tính là nữ thường gặp nhiều rào cản và khó khăn hơn trong việc thăng tiến trong thế giới kinh doanh. Tuy nhiên, động thái và sự thay đổi trong xã hội đã mở ra nhiều cơ hội hơn cho phụ nữ để trở thành tỷ phú và tự lập.

Tài sản ban đầu của các tỷ phú có thể ảnh hưởng đáng kể đến khả năng tự lập. Người có tài sản khởi đầu lớn hơn có thể dễ dàng hơn để đầu tư vào các dự án mới, trong khi người khác có thể phải bắt đầu từ con số 0 và phụ thuộc vào việc khởi nghiệp hoặc đầu tư để tích lũy tài sản.

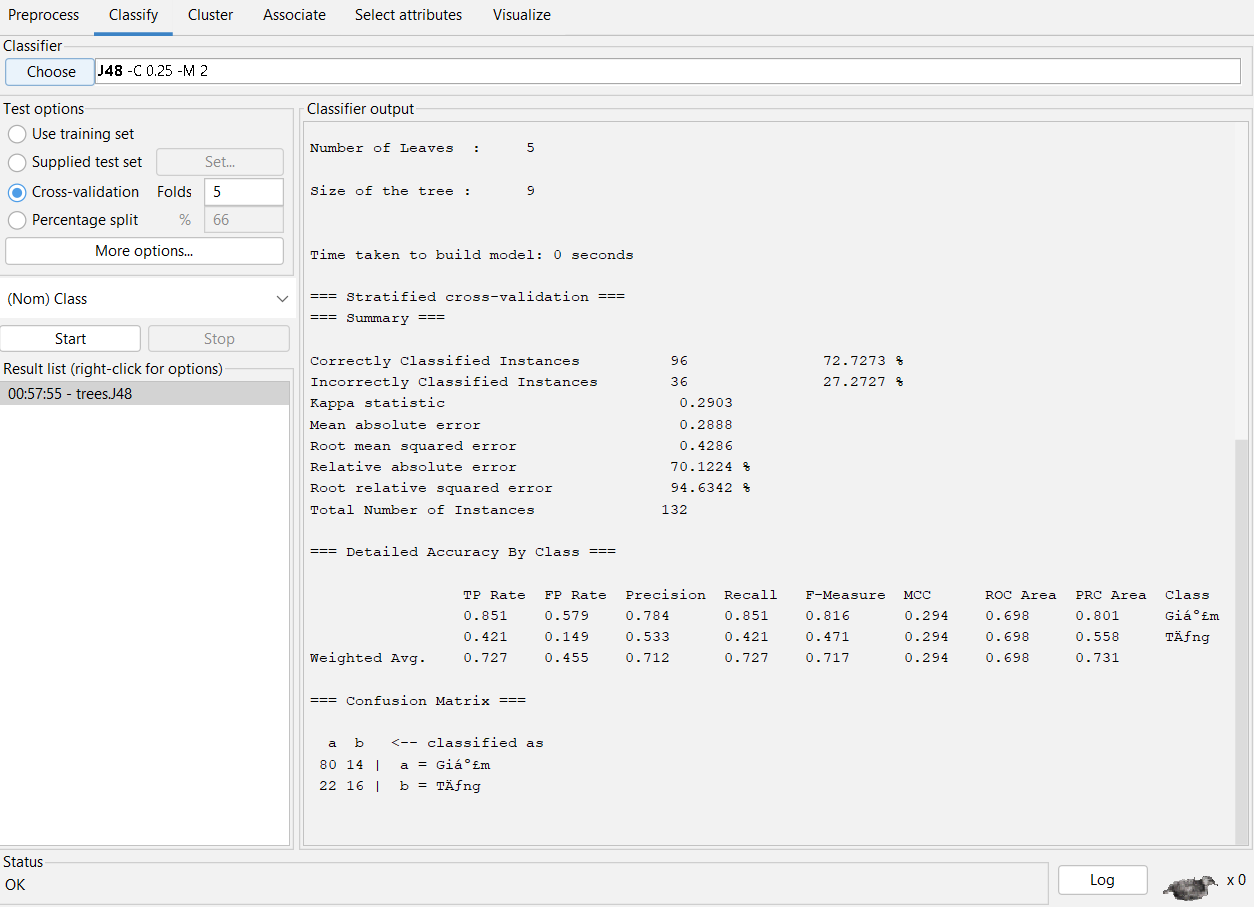
**\*Data Companies:**

**-** Ở dữ liệu này chúng ta sẽ quan tâm tới việc các công ty đa quốc gia có thuộc lớp dạng tăng trưởng hay là không (Class) dựa vào việc phân lớp về Tổng kinh doanh (Sales), Lợi nhuận (Profit), Tài sản cổ phiếu (Market Value) và Quốc gia (Country).



Hình 4. 4: Khái quát đầu vào dữ liệu công ty trong phân lớp

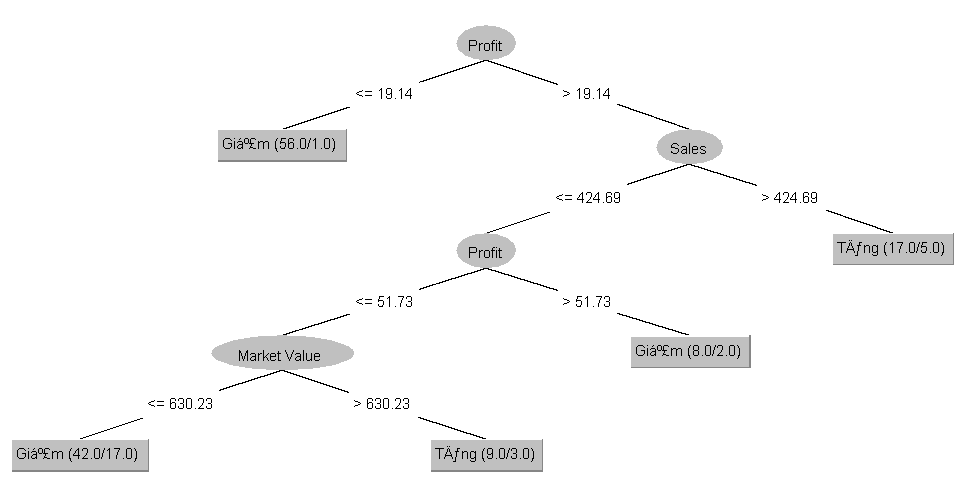
- Sử dụng thuật toán cây quyết định J48 với K Cross-validation = 5 để xét mức độ tăng trưởng hoặc giảm sút của từng công ty qua 3 năm 2020-2021-2022:



Hình 4. 5: Mô tả kết quả sau khi chạy mô hình phân lớp dữ liệu công ty

-Ở trong dữ liệu sau khi huấn luyện ta thấy với K = 5 thì mô hình có tỷ lệ đúng tạm ổn là 72,72%.

- Nhìn vào MTSS – Confusion Matrix ta thấy việc phân lớp vào lớp Giảm có độ chính xác cao hơn rất nhiều bởi vì trong các công ty đa quốc gia hàng đầu thì hết 94 công ty thuộc về nhóm lớp Giảm so với 38 công ty thuộc lớp Tăng. Vậy nên em huấn luyện mô hình lấy độ chính xác về lớp Giảm sẽ cao hơn so với lớp Tăng. Nhằm mục đích chọn lọc những ngành nghề tiềm năng chính xác nhất.



Hình 4. 6: Cây quyết định dữ liệu công ty

- Từ cây quyết định em đem vào Excel để xem xét những ngành nghề dự đoán thông qua các lớp được phân như bên trên.

***\*Nhận xét:***

Từ cây quyết định em có dự đoán với những công ty thuộc các nhóm lĩnh vực như Công nghệ, Ô tô, Dịch vụ, Bảo hiểm, Ngân hàng và Truyền thông là những công ty có tiềm năng phát triển mạnh mẽ trong năm tới bởi vì nhìn vào cây và xét từ dữ liệu gốc cho ta thấy những công ty thuộc lớp là Tăng đều xuất hiện trong 3 năm với mức độ tăng trưởng lên theo từng năm.

# TỔNG KẾT

Từ đề tài lần này đã cho thấy được tổng quan sự đặc trưng trong thuật toán về việc áp dụng KPDL trong việc xử lý dữ liệu và cách áp dụng các thuật toán vào bài làm.

Về lý thuyết thì KPDL gồm các bước như: xác định bài toán cần xử lý là gì sau đó thu thập dữ liệu rồi tiền xử lý dữ liệu, áp dụng các thuật toán để khai phá dữ liệu cần thiết vào bài làm và đưa ra các tri thức mới có thể áp dụng được. KPDL có thể là những phương pháp như: gom cụm, phân lớp, luật kết hợp, tổ hợp, mẫu tuần tự… nhưng em áp dụng các phương pháp đã học vào đồ án chứ chưa thể đào sâu vào các vấn đề và áp dụng được nhiều phương pháp khai phá hơn.

Về phương pháp thì đồ án của em đã áp dụng các thuật toán như K-means trong gom cụm dữ liệu để đưa ra các cụm dữ liệu có thuộc tính tương đồng nhau, Apriori trong luật kết hợp để đưa ra dự đoán các ngành nghề tiềm năng trong tương lai và Cây quyết định trong phân lớp để đưa ra các bộ dữ liệu với các lớp xác định.

Trong quá trình làm báo cáo thì em cũng đã cố gắng tham khảo ở rất nhiều nguồn và chắc chắn không thể tránh khỏi những sai sót. Rất mong được sự góp ý của các thầy cô để hoàn thiện hơn cho các kiến thức của em.

TÀI LIỆU THAM KHẢO

[1] H. Đinh, “Data Mining – Khai phá dữ liệu – [Data Science Series]”, 2022. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/data-mining-khai-pha-du-lieu-data-science-series-1Je5EAX45nL>. [Truy cập 15/09/2023].

[2] P. Đ. Khánh, “Machine Learning lý thuyết tới thực hành”, 2021. [Trực tuyến]. Địa chỉ: <https://phamdinhkhanh.github.io/deepai-book/ch_ml/index_PCA.html>. [Truy cập 15/09/2023].

[3] N. T. Hop, “Hierarchical Clustering – Phân cụm dữ liệu”, 2020. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/hierarchical-clustering-phan-cum-du-lieu-maGK7q2elj2>. [Truy cập 16/09/2023].

[4] T. Nhi, “Khai phá luật kết hợp trong cơ sở dữ liệu (Association rule in data mining) là gì?”, 2020. [Trực tuyến]. Địa chỉ: <https://vietnambiz.vn/khai-pha-luat-ket-hop-trong-co-so-du-lieu-association-rule-in-data-mining-la-gi-2020052614221747.htm>. [Truy cập 18/09/2023].

[5] N. M. Duc, “Thuật toán Apriori khai phá luật kết hợp trong Data Mining”, 2019. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/thuat-toan-apriori-khai-pha-luat-ket-hop-trong-data-mining-3P0lPEv85ox>. [Truy cập 18/09/2023].

[6] N. T. T. Linh, “*Nghiên cứu các thuật toán phân lớp dữ liệu dựa trên cây quyết định,*” Luận văn tốt nghiệp, Đại học Quốc gia Hà Nội Trường Đại học Công Nghệ, 2005.

PHỤ LỤC