**BỘ TÀI CHÍNH**

**TRƯỜNG ĐẠI HỌC TÀI CHÍNH – MARKETING**

**KHOA CÔNG NGHỆ THÔNG TIN**

**------------------------🙡🙣------------------------**

Logo, company name

Description automatically generated

**BÁO CÁO ĐỒ ÁN**

**KHAI PHÁ DỮ LIỆU**

**Đề tài:**

**SỬ DỤNG THUẬT TOÁN APRIORI VÀ FP-GROWTH ĐỂ XÁC ĐỊNH XU HƯỚNG TÌM VIỆC VÀ ỨNG TUYỂN**

**Giảng viên hướng dẫn:** Thái Thị Ngọc Lý

**Mã lớp học phần:** 2311112005903

**Nhóm:** 11

**Sinh viên thực hiện 1:** Lê Đình Quý

**Sinh viên thực hiện 2:** Nguyễn Minh Trí

**TpHCM, ngày 15 tháng 4 năm 2023**

**BỘ TÀI CHÍNH**

**TRƯỜNG ĐẠI HỌC TÀI CHÍNH – MARKETING**

**KHOA CÔNG NGHỆ THÔNG TIN**

**------------------------🙡🙣------------------------**

Logo, company name

Description automatically generated

**BÁO CÁO ĐỒ ÁN**

**KHAI PHÁ DỮ LIỆU**

**Đề tài:**

**SỬ DỤNG THUẬT TOÁN APRIORI VÀ FP-GROWTH ĐỂ XÁC ĐỊNH XU HƯỚNG TÌM VIỆC VÀ ỨNG TUYỂN**

**Giảng viên hướng dẫn:** Thái Thị Ngọc Lý

**Mã lớp học phần:** 2311112005903

**Nhóm:** 11

**Sinh viên thực hiện 1:** Lê Đình Quý

**MSSV1:** 2021010167

**Sinh viên thực hiện 2:** Nguyễn Minh Trí

**MSSV2:** 2021010325

**TpHCM, ngày 15 tháng 4 năm 2023**

LỜI CẢM ƠN

Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến giảng viên phụ trách bộ môn Khai Phá Dữ Liệu – Cô Thái Thị Ngọc Lý, người đã gắn bó với chúng em suốt cả học kì vừa qua, cô đã truyền đạt những kiến thức hay, bổ ích và thú vị về bộ môn này một cách rất nhiệt huyết, hỗ trợ tụi chúng em nhiệt tình, giải đáp các câu hỏi một cách dễ hiểu và nhanh chóng. Bằng những kinh nghiệm quý báo và sự nhiệt tình ấy, chúng em và các bạn trong lớp đã tiếp thu được nhiều bài học quý để phục vụ cho việc học tập của bản thân hiện tại cũng như là có thêm hành trang chuẩn bị cho tương lai, vững bước cho công việc sau này.

Chúng em xin gửi đến cô bài đồ án này, đây được xem như là bài báo cáo kết quả những gì chúng em đã học và tiếp thu được sau một khoảng thời gian dài được cô dạy bảo. Nguồn kiến thức thì vô hạn, nhưng với trình độ của chúng em hiện tại có thể nó còn hạn chế, nên trong quá trình làm bài chắc chắn sẽ khó tránh khỏi các sai sót nhất định. Bản thân chúng em rất mong nhận được những lời góp ý đến từ cô sau khi xem bài tiểu luận này để chúng em ngày càng hoàn thiện thêm kĩ năng về tin học của mình.

Lời cuối chúng em xin chúc cô nhiều sức khỏe, đạt được sự thành công trong những điều cô mong muốn, thành công hơn trong con đường sự nghiệp của mình và truyền đạt được nhiều thêm cho các thế hệ kế tiếp.

Chúng em xin chân thành cảm ơn.

NHẬN XÉT ĐÁNH GIÁ CỦA GIẢNG VIÊN 1

Điểm số:

Điểm chữ:

TP. Hồ Chí Minh, ngày…... tháng 4 năm 2023

Giảng viên

(*ký và ghi rõ họ tên)*

NHẬN XÉT ĐÁNH GIÁ CỦA GIẢNG VIÊN 2

Điểm số:

Điểm chữ:

TP. Hồ Chí Minh, ngày…... tháng 4 năm 2023

Giảng viên

(*ký và ghi rõ họ tên)*

DANH MỤC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **STT** | **Từ viết tắt** | **Ý nghĩa** |
| 1 | DW | Data warehouse |
| 2 | CSDL | Cơ sở dữ liệu |
| 3 | CRM | Quản trị mối quan hệ khách hàng |
| 4 | ETL | Extract, Transform, Load |
| 5 | CRISP-DM | Cross-Industry Standard Process for Data Mining |
| 6 | IDE | An Integrated Development Environment |

DANH MỤC THUẬT NGỮ ANH VIỆT

|  |  |  |
| --- | --- | --- |
| **STT** | **Từ tiếng Anh** | **Nghĩa tiếng Việt** |
| 1 | Data Warehouse Schema | Lược đồ kho dữ liệu |
| 2 | Star Schema | Lược đồ sao |
| 3 | Snowflake Schema | lược đồ bông tuyết |
| 4 | Galaxy Schema | lược đồ thiên hà |
| 5 | Financial Data Analysis | Phân tích dữ liệu tài chính |
| 6 | Retail Industry | Ngành Bán lẻ Ngành |
| 7 | Telecommunication Industry | Viễn thông |

DANH MỤC BẢNG BIỂU

[Bảng 2.1. Bảng dữ liệu để ví dụ về Apriori 12](#_Toc132480895)

[Bảng 2.2. Bảng thống kê các phần tử của dữ liệu (C1) 13](#_Toc132480896)

[Bảng 2.3. Bảng dữ liệu sau khi xét min\_support (L1) 13](#_Toc132480897)

[Bảng 2.4. Bảng sau khi kết hợp các Item ở L1 14](#_Toc132480898)

[Bảng 2.5. . Bảng dữ liệu sau khi loại bỏ các phần tử theo điều kiện 14](#_Toc132480899)

[Bảng 2.6. Bảng thống kê các phần tử của dữ liệu (C2) 14](#_Toc132480900)

[Bảng 2.7. Bảng dữ liệu sau khi xét min\_support (L2) 15](#_Toc132480901)

[Bảng 2.8. Bảng dữ liệu sau khi hoàn thành kết hợp (C3) 15](#_Toc132480902)

[Bảng 2.9. Bảng thể hiện luật và độ tin cậy sinh ra từ L2 15](#_Toc132480903)

[Bảng 2.10. Bảng dữ liệu mẫu để chạy FB - Growth 17](#_Toc132480904)

[Bảng 2.11. Bảng thống kê Items (C1) 18](#_Toc132480905)

[Bảng 2.12. Bảng dữ liệu sau khi xét min\_support (L1) 18](#_Toc132480906)

[Bảng 2.13. Bảng xét cái Items phỏ biến từ bảng L1 18](#_Toc132480907)

[Bảng 2.14. Bảng thể hiện các luật được sinh ra 20](#_Toc132480908)

[Bảng 2.15. Bảng dữ liệu chính 20](#_Toc132480909)

[Bảng 4.1. Bảng mô tả các thuộc tính và miền dữ liệu của thuộc tính 23](#_Toc132480910)

DANH MỤC HÌNH ẢNH

[Hình 2.1. Sơ đồ cây để duyệt các Items 19](#_Toc132480920)

[Hình 4.1. Giao diện khi mở WEKA 27](#_Toc132480921)

[Hình 4.2. Gio diện khi mở Explorer và thực hiện thao tác Open file 27](#_Toc132480922)

[Hình 4.3. Giao diện khi thực hiên thành công Open file và xem dữ liệu trên Edit 28](#_Toc132480923)

[Hình 4.4. Link để tải các JDBC 29](#_Toc132480924)

[Hình 4.5. Thư mục khi tải thành công JDBC và chọn phiên bản ‘jre11’ 29](#_Toc132480925)

[Hình 4.6. Giao diện *About* trong phần Setting trên máy tính 30](#_Toc132480926)

[Hình 4.7. Giao diện khi thêm CLASSPATH cho máy 30](#_Toc132480927)

[Hình 4.8. Giao diện thực hiện mỡ file ‘RunWeka” bằng Notepad 31](#_Toc132480928)

[Hình 4.9. Thêm đường dẫn CLASSPATH trong file ‘RunWeka’ 31](#_Toc132480929)

[Hình 4.10. Giao diện giải nén file weka.jar 32](#_Toc132480930)

[Hình 4.11. Đường dẫn để truy cập vào file ‘DatabaseUtils.props’ 32](#_Toc132480931)

[Hình 4.12. Thực hiện chỉnh sửa ‘jdbcURL’ trên file ‘DatabaseUtils.props’’ 33](#_Toc132480932)

[Hình 4.13. Thêm các thuộc tính trong file ‘DatabaseUtils.props.mssqlserver2015’ 33](#_Toc132480933)

[Hình 4.14. Thực hiện sửa file thành ‘DatabaseUtils.props’ 34](#_Toc132480934)

[Hình 4.15. Giao diện để Open các file DataBase 34](#_Toc132480935)

[Hình 4.16. Giao diện để kết nối vào Server 34](#_Toc132480936)

[Hình 4.17. Phần đăng nhập vào SQL Server 35](#_Toc132480937)

[Hình 4.18. Cấp quyền cho người dùng ‘sa’ trong SQL Server 35](#_Toc132480938)

[Hình 4.19. Thực hiện thay đổi cho người dùng ‘sa’ trong SQL Server 36](#_Toc132480939)

[Hình 4.20. Câu lệnh Query để thực hiện thao tác tạo bảng và thêm dữ liệu 36](#_Toc132480940)

[Hình 4.21. Giao diện khi đã kết nối thành công 37](#_Toc132480941)

[Hình 4.22. Sử dụng các lệnh để truy xuất các dữ liệu 37](#_Toc132480942)

[Hình 4.23. Kết quả khi mở dữ liệu từ DataBase 38](#_Toc132480943)

[Hình 4.24. Chọn các thuộc tính cần xóa 39](#_Toc132480944)

[Hình 4.25. Kết quả sau khi xóa các thuộc tính 39](#_Toc132480945)

[Hình 4.26. Thư mục chọn cách thức cho việc lọc dữ liệu 40](#_Toc132480946)

[Hình 4.27. Điều chỉnh các thông số cho việc lọc ‘Discretize’ 41](#_Toc132480947)

[Hình 4.28. Kết quả việc lọc ‘Discretize’ 42](#_Toc132480948)

[Hình 4.29. Thư mục chọn cách thức cho việc lọc dữ liệu 43](#_Toc132480949)

[Hình 4.30. Điều chỉnh các thông số cho việc chuyển đổi ‘NumericToNominal’ 43](#_Toc132480950)

[Hình 4.31. Kết quả sau khi chuyển đổi kiểu dữ liệu numeric thành nominal 44](#_Toc132480951)

[Hình 4.32. Thư mục chọn cách thức cho việc lọc dữ liệu 45](#_Toc132480952)

[Hình 4.33. Điều chỉnh các thông số cho việc phân tập dữ liệu 46](#_Toc132480953)

[Hình 4.34. Kết quả sau khi phân tập dữ liệu 46](#_Toc132480954)

[Hình 4.35. Xóa các thuộc tính không cần và giữ lại các thuộc tính để chạy Apriori 47](#_Toc132480955)

[Hình 4.36. Thực hiện chọn các thư mục cần xóa và ấn ‘Remove’ để xóa 48](#_Toc132480956)

[Hình 4.37. Thực hiện thêm thuộc tính phân lớp cho tập dữ liệu 48](#_Toc132480957)

[Hình 4.38. Kết quả của việc thêm thuộc tính phân lớp 49](#_Toc132480958)

[Hình 4.39. Chọn việc lọc ‘NominalToBinary’ để chuyển kiểu dữ liệu và điều chỉnh thông số cho việc chuyển đổi 49](#_Toc132480959)

[Hình 4.40. Sau khi chuyển đổi thành công thì xóa thuộc tính phân lớp 50](#_Toc132480960)

[Hình 4.41. Xem tập dữ liệu sau khi đã thực hiện phân lớp 51](#_Toc132480961)

[Hình 4.42. Lưu lại tệp sau khi đã xử lý để thực hiện việc chạy FP-Growth 52](#_Toc132480962)

[Hình 4.43. Chọn công cụ ‘Associate’ và chọn ‘Apriori’ để thực hiện chạy 52](#_Toc132480963)

[Hình 4.44. Điều chỉnh các thông số cho việc chạy Apriori 53](#_Toc132480964)

[Hình 4.45. Kết quả chạy Apriori lần 1 54](#_Toc132480965)

[Hình 4.46. Kết quả chạy Apriori lần 2 54](#_Toc132480966)

[Hình 4.47. Giao diện thực hiện ‘Save’ kết quả chạy 55](#_Toc132480967)

[Hình 4.48. Đặt tên cho kết quả chạy và lưu 55](#_Toc132480968)

[Hình 4.49. Check chọn tất cả các dòng dữ liệu 56](#_Toc132480969)

[Hình 4.50. Chọn cách kết hợp FP-Growth 56](#_Toc132480970)

[Hình 4.51. Điều chỉnh các thông số cho việc chạy FP-Growth 57](#_Toc132480971)

[Hình 4.52. Kết quả chạy FP-Growth 57](#_Toc132480972)

[Hình 4.53. Thực hiện lưu kết quả chạy 58](#_Toc132480973)

[Hình 4.54. Chọn các thuộc tính khác cho lần chạy thứ 2 trên FP-Growth 58](#_Toc132480974)

[Hình 4.55. Thực hiện chuyển kiểu dữ liệu 59](#_Toc132480975)

[Hình 4.56. Kết quả lần chạy FP-Growth thứ 2 59](#_Toc132480976)

[Hình 4.57. Kết quả chạy apriori với dữ liệu lần 1 trên R 61](#_Toc132480977)

[Hình 4.58. Kết quả chạy FP-Growth trên R 63](#_Toc132480978)

[Hình 5.1. Điều chỉnh thông số và thực hiện phân lớp cho dữ liệu 74](#_Toc132480979)

[Hình 5.2. Thực hiện lưu file vừa lọc 75](#_Toc132480980)

[Hình 5.3. Giao diện chính của weka và chọn ‘Experrimenter’ 75](#_Toc132480981)

[Hình 5.4. Thực hiện thêm mới một thử nghiệm 76](#_Toc132480982)

[Hình 5.5. Điều chỉnh thông số và thêm vào một tệp dữ liệu 76](#_Toc132480983)

[Hình 5.6. Kết quả khi thêm tập dữ liệu 77](#_Toc132480984)

[Hình 5.7. Thêm mới các thuật toán để chạy bộ dữ liệu 77](#_Toc132480985)

[Hình 5.8. Thuật toán ‘ZeroR’ và các thông số thuộc tính 78](#_Toc132480986)

[Hình 5.9. Thuật toán ‘OneR’ và các thông số thuộc tính 79](#_Toc132480987)

[Hình 5.10. Thuật toán ‘J48’ và các thông sốt thuộc tính 80](#_Toc132480988)

[Hình 5.11. Kết quả sau khi thêm các thuật toán 81](#_Toc132480989)

[Hình 5.12. Chọn vào mục ‘Run’ và thực hiện ‘Start’ để thực hiện 82](#_Toc132480990)

[Hình 5.13. Sau khi thực hiện xong và không báo lỗi ấn qua mục ‘Analyse’ để xem kết quả 83](#_Toc132480991)

[Hình 5.14. Chọn ‘Ranking’ để xem thứ hạng thuật toán 83](#_Toc132480992)

[Hình 5.15. Xem các sai lệch của các thuật toán 84](#_Toc132480993)

[Hình 5.16. Thực hiện lưu lại các kết quả đã chạy 84](#_Toc132480994)

[Hình 5.17. Lưu 3 kết quả đã thực hiện được 85](#_Toc132480995)

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc132481002)

[NHẬN XÉT ĐÁNH GIÁ CỦA GIẢNG VIÊN 1 ii](#_Toc132481003)

[NHẬN XÉT ĐÁNH GIÁ CỦA GIẢNG VIÊN 2 iii](#_Toc132481004)

[DANH MỤC TỪ VIẾT TẮT iv](#_Toc132481005)

[DANH MỤC THUẬT NGỮ ANH VIỆT v](#_Toc132481006)

[DANH MỤC BẢNG BIỂU vi](#_Toc132481007)

[DANH MỤC HÌNH ẢNH vii](#_Toc132481008)

[MỤC LỤC xi](#_Toc132481009)

[CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI NGHIÊN CỨU 1](#_Toc132481010)

[1.1. Tổng quan về đề tài 1](#_Toc132481011)

[1.2. Phạm vi của đề tài 2](#_Toc132481012)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 3](#_Toc132481013)

[2.1. Kho dữ liệu 3](#_Toc132481014)

[2.1.1. Khái niệm kho dữ liệu 3](#_Toc132481015)

[2.1.2. Data Warehouse Schema 4](#_Toc132481016)

[2.1.3. Mô hình dữ liệu đa chiều 5](#_Toc132481017)

[2.2. Khái quát khai phá dữ liệu 7](#_Toc132481018)

[2.2.1. Khái niệm 7](#_Toc132481019)

[2.2.2. Qui trình khai phá dữ liệu 7](#_Toc132481020)

[2.2.3. Các phương pháp khai phá dữ liệu 9](#_Toc132481021)

[2.2.4. Các ứng dụng của khai phá dữ liệu 10](#_Toc132481022)

[2.3. Phương pháp khai phá dữ liệu được sử dụng trong đề tài 12](#_Toc132481023)

[2.3.1. Thuật toán Apriori 12](#_Toc132481024)

[2.3.2. Thuật toán FP-Growth 16](#_Toc132481025)

[CHƯƠNG 3. PHẦN MỀM KHAI PHÁ DỮ LIỆU MÃ NGUỒN MỞ 21](#_Toc132481026)

[3.1. WEKA 21](#_Toc132481027)

[3.1.1. Giới thiệu 21](#_Toc132481028)

[3.1.2. Chức năng 21](#_Toc132481029)

[3.2. Rstudio 22](#_Toc132481030)

[3.2.1. Giới thiệu 22](#_Toc132481031)

[3.2.2. Các tính năng chính của Rstudio 22](#_Toc132481032)

[CHƯƠNG 4. KHAI PHÁ DỮ LIỆU 23](#_Toc132481033)

[4.1. Xác định vấn đề 23](#_Toc132481034)

[4.2. Hiểu dữ liệu 23](#_Toc132481035)

[4.3. Chuẩn bị dữ liệu 26](#_Toc132481036)

[4.3.1. Nạp dữ liệu 26](#_Toc132481037)

[4.3.2. Xử lý dữ liệu tổng 38](#_Toc132481038)

[4.3.3. Xử lý dữ liệu cho Apriori 46](#_Toc132481039)

[4.3.4. Xử lý dữ liệu cho FP-Growth 47](#_Toc132481040)

[4.4. Lập mô hình 52](#_Toc132481041)

[4.4.1. Apriori 52](#_Toc132481042)

[4.4.2. FP-Growth 55](#_Toc132481043)

[4.4.3. Apriori trên R 60](#_Toc132481044)

[4.4.4. FP-Growth trên R 61](#_Toc132481045)

[4.5. Đánh giá mô hình 63](#_Toc132481046)

[4.5.1. Đánh giá mô hình với kết quả xử lý DL1 (KetQua\_Apriori\_DL1 or KetQua\_FBGrowth\_DL1) 63](#_Toc132481047)

[4.5.2. Đánh giá mô hình với kết quả xử lý DL2 (KetQua\_Apriori\_DL2 or KetQua\_FBGrowth\_DL2) 67](#_Toc132481048)

[4.6. Triển khai mô hình 71](#_Toc132481049)

[CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 73](#_Toc132481050)

[5.1. Kết luận 73](#_Toc132481051)

[5.1.1. Những kết quả đạt được 73](#_Toc132481052)

[5.1.2. Hạn chế 73](#_Toc132481053)

[5.2. Hướng phát triển 73](#_Toc132481054)

[5.2.1. Hướng khắc phục những hạn chế 73](#_Toc132481055)

[5.2.2. Hướng mở rộng của đề tài - Experimenter trong weka 74](#_Toc132481056)

[TÀI LIỆU THAM KHẢO 86](#_Toc132481057)

# TỔNG QUAN VỀ ĐỀ TÀI NGHIÊN CỨU

## Tổng quan về đề tài

Trong những năm gần đây – đặc biệt là trong những năm 2022 - 2023, công việc tuyển dụng nguồn nhân lực của các công ty ngày càng khó khăn. Trước làn sóng cắt giảm nhân sự mạnh mẽ cuối năm 2022 và hàng loạt thông tin về tình hình biến động kinh tế giai đoạn đầu năm 2023, rất nhiều người lao động đang tỏ ra e ngại về vấn đề tìm việc làm.

*“Theo khảo sát của JobsGO với hơn 3000 người lao động trên cả nước, phần lớn người tin rằng tìm việc làm năm 2023 sẽ gặp nhiều khó khăn (61.8%), trong đó, tỷ lệ nữ giới cao hơn nam giới (nữ 65.9% và nam 59.2%). 27.9% người nhận thấy không có quá nhiều khác biệt so với năm 2022. Chỉ có khoảng 10.3% người lao động nghĩ tìm việc năm 2023 sẽ dễ dàng hơn. Tuy dự kiến không mấy khả quan, nhưng vẫn có tới 38.9% người lao động đang sẵn sàng tìm việc mới trong năm 2023.*

*Đặc biệt, ứng viên càng trẻ và ít kinh nghiệm lại càng có suy nghĩ tích cực về vấn đề việc làm. Dựa vào kết quả khảo sát, có 54.6% ứng viên dưới 22 tuổi và 55.2% ứng viên là thực tập sinh cho rằng tìm kiếm việc làm năm 2023 sẽ vẫn diễn ra bình thường, thậm chí là dễ hơn năm 2022. Hơn 60% ứng viên trên 27 tuổi với vị trí, chức vụ cao (từ chuyên viên trở lên) lại lo lắng, tỏ ra quan ngại khi tìm việc”*

Trước những khó khăn đó, việc có thể nắm bắt được công nghệ thông tin được coi là một cơ sở hay là một nhân tố hỗ trợ giải quyết hiện trạng này một cách trơn tru hơn. Từ việc thu thập, hiểu được, trích xuất, tính toán, vận dụng thông tin để hành động một các chuẩn xác đều giúp các công ty, doanh nghiệp đưa ra những quyết định đúng đắn trong viêc quyết định tính tuyển dụng của mình. Để phục vụ công tác tuyển dụng đó, chúng ta bắt buộc phải có danh sách dữ liệu về nhu cầu, mong muốn về việc làm của nguồn nhân lực ngoài thị trường thì chúng ra mới có thể đưa ra được những quyết định đúng đắn. Vì vậy các công ty, doanh nghiệp đã dựa vào một vài điểm như sau:

* Bảng dữ liệu có đầy đủ các thông tin liên quan, thống nhất với nhau.
* Quản lý dữ liệu tập trung, tránh thiếu sót và dễ dàng hơn trong các kỹ thuật xử lý dữ liệu để ra báo cáo chi tiết. Nếu các dữ liệu có tính chất khác nhau có thể đưa vào các bảng dữ liệu khác nhau.
* Mỗi thông tin thay đổi cần có thời gian cụ thể, ghi tách riêng thành các dòng dữ liệu mới, không ghi đè lên dữ liệu đã có. Việc này giúp thông tin có tính hệ thống, thành lịch sử và liền mạch.
* Mỗi thông tin quản lý theo 1 cột, không trộn ô trong bảng dữ liệu kể cả tiêu đề. Việc này giúp tránh các lỗi có thể xảy ra khi trích lọc dữ liệu ra báo cáo.

## Phạm vi của đề tài

Xử dụng thuật toán Apriori và FBGrowth xác định xu hướng việc làm và tìm việc:

* Cơ sở lý thuyết về khai phá dữ liệu
* Tìm hiểu về thuật toán Apriori và FBGrowth, ví dụ mẫu
* Tìm hiểu về công cụ hỗ trợ khai phá dữ liệu Weka và Apriori
* Thực hiện kết nối cơ sở dữ liệu giữa Weka và SqlServer 2014
* Thực hiện tiền xử lý dữ liệu
* Thực hiện thuật toán trên Weka và trên R
* Tìm hiểu về Experimenter Enviroment trong Weka.

# CƠ SỞ LÝ THUYẾT

## Kho dữ liệu

### Khái niệm kho dữ liệu

Data warehouse (DW) hay kho dữ liệu là một hệ thống lưu trữ dữ liệu từ nhiều nguồn, nhiều môi trường khác nhau như: phần mềm bán hàng, kế toán, nhân sự hay hệ thống lõi ngân hàng,… giúp tăng cường hiệu suất của các truy vấn cho báo cáo và phân tích.

Data Warehouse hoạt động như một kho lưu trữ trung tâm. Dữ liệu đi vào kho dữ liệu từ hệ thống giao dịch và các cơ sở dữ liệu liên quan khác. Sau đó, dữ liệu được xử lý, chuyển đổi để người dùng có thể truy cập những dữ liệu này thông qua công cụ Business Intelligence, SQL client hay bảng tính.

Một Data Warehouse thường bao gồm các yếu tố như:

* Một cơ sở dữ liệu quan hệ để lưu trữ và quản lý dữ liệu.
* Giải pháp trích xuất, tải và biến đổi ELT để chuẩn bị dữ liệu cho phân tích.
* Khả năng phân tích thống kê, báo cáo và khai thác dữ liệu.
* Các công cụ phân tích khách hàng để trực quan hóa và trình bày dữ liệu cho người dùng doanh nghiệp.
* Các ứng dụng phân tích khác, phức tạp hơn tạo ra thông tin có thể hành động bằng cách áp dụng khoa học dữ liệu và thuật toán trí tuệ nhân tạo AI hoặc các tính năng đồ thị và không gian cho phép nhiều loại phân tích dữ liệu hơn trên quy mô lớn.

Các đặc tính của Data Warehouse:

* **Hướng chủ đề (subject-oriented):** Hướng chủ đề tức thông tin trong Data Warehouse sẽ được tổ chức và sắp xếp theo một chủ đề nhất định. Mục đích của Kho dữ liệu là phục vụ các yêu cầu phân tích, hoặc khai phá cụ thể được gọi là chủ đề.
* **Được tích hợp (integrated):** Với doanh nghiệp, dữ liệu cần phân tích nằm rải rác tại những phòng ban khác nhau và cần tích hợp lại. Từ đó, tổng hợp dữ liệu từ nhiều nguồn vào một kho dữ liệu cho phép chúng ta có thể xem đồng thời nhiều nhóm chỉ tiêu khác nhau. Quá trình tích hợp này sẽ được thực hiện trong quá trình ETL.
* **Có gán nhãn thời gian (time variant):** Vì dữ liệu thay đổi liên tục nên chúng sẽ được gán 1 nhãn thời gian tương ứng tại thời điểm nhập liệu. Việc gắn thời gian này giúp ta dễ dàng so sánh dữ liệu với nhau để biết được các thay đổi đang đi theo chiều hướng tích cực hay tiêu cực.
* **Bất biến (non-volatile):** Dữ liệu trong Kho dữ liệu có chức năng báo cáo lại các chỉ số về hoạt động kinh doanh thực tế đã xảy ra do đó không thể cập nhật, thay đổi vì nó sẽ không phản ánh đúng thực tế. Vì vậy, với kho dữ liệu chỉ có 2 thao tác chính là tải dữ liệu vào kho và truy cập (đọc) dữ liệu từ kho.

### Data Warehouse Schema

**Data Warehouse Schema** là một cấu trúc xác định hợp lý nội dung của Kho dữ liệu, bằng cách tạo điều kiện thuận lợi cho các hoạt động được thực hiện trên Kho dữ liệu và các hoạt động bảo trì của hệ thống Kho dữ liệu, thường bao gồm mô tả chi tiết về cơ sở dữ liệu, bảng, dạng xem, các chỉ mục và Dữ liệu thường được cấu trúc bằng cách sử dụng các loại thiết kế được xác định trước như Star Schema, Snowflake Schema, Galaxy Schema

* **Star Schema:** Star Schema được biết là tạo ra các bảng kích thước không chuẩn hóa – một chiến lược cấu trúc cơ sở dữ liệu tổ chức các bảng để đưa ra dự phòng nhằm cải thiện hiệu suất.
* **Snowflake Schema:** là một lược đồ kho dữ liệu bao gồm sự sắp xếp hợp lý của các bảng thứ nguyên. Lược đồ kho dữ liệu này xây dựng trên lược đồ hình sao bằng cách thêm các bảng thứ nguyên phụ bổ sung liên quan đến các bảng thứ nguyên bậc nhất được nối với bảng thực tế.
* **Galaxy Schema:** còn được gọi là Lược đồ chòm sao thực tế, đóng vai trò là bước lặp tiếp theo của lược đồ kho dữ liệu. Nó có thể được coi là lược đồ sao được liên kết với nhau và được chuẩn hóa hoàn toàn, tránh bất kỳ loại dư thừa hoặc không nhất quán nào của dữ liệu.

### Mô hình dữ liệu đa chiều

**Mô hình dữ liệu đa chiều** có thể được định nghĩa là một phương pháp sắp xếp dữ liệu trong cơ sở dữ liệu, với cấu trúc và tổ chức nội dung trong cơ sở dữ liệu tốt hơn. Không giống như hệ thống có một chiều như danh sách, Mô hình Dữ liệu Đa chiều có thể có hai hoặc ba chiều của các mục từ hệ thống cơ sở dữ liệu.

Các giai đoạn tạo ra mô hình dữ liệu đa chiều:

* **Tập hợp các yêu cầu từ máy khách:** Tương tự như các ứng dụng phần mềm khác, Mô hình dữ liệu cũng yêu cầu chính xác yêu cầu từ máy khách. Hầu hết thời gian, khách hàng có thể không biết những gì có thể đạt được với công nghệ đã chọn.
* **Phân loại các module khác nhau của hệ thống:** Sau quá trình thu thập toàn bộ yêu cầu, bước tiếp theo là xác định và phân loại từng yêu cầu theo module mà chúng thuộc về.
* **Phát hiện các kích thước khác nhau dựa trên đó hệ thống cần được thiết kế:** Sau khi hoàn thành việc tách các yêu cầu khác nhau và chuyển chúng sang các mô-đun phù hợp, bước tiếp theo là xác định các yếu tố chính, theo quan điểm của người dùng.
* **Soạn thảo các thứ nguyên thời gian thực và các thuộc tính tương ứng**: Là một phần của bước tiếp theo, trong quy trình của Mô hình dữ liệu đa chiều, các thứ nguyên được xác định trong bước trước có thể được sử dụng thêm để nhận dạng các thuộc tính liên quan
* **Khám phá sự thật từ các tham số đã được liệt kê và thuộc tính của chúng:** Từ việc thu thập yêu cầu ban đầu, các tham số có thể là sự kết hợp giữa các tham số và dữ kiện. Đó là một bước quan trọng để phân biệt và tách biệt các sự kiện khỏi các chiều.
* **Xây dựng Lược đồ để đặt dữ liệu, đối với thông tin thu thập được từ các bước trên:** Dựa trên thông tin được thu thập cho đến nay, các yêu cầu phức tạp, kích thước, sự kiện và thuộc tính tương ứng của chúng, một Lược đồ có thể được xây dựng.

**Ưu điểm**

* Mô hình dữ liệu đa chiều có thể thực hiện được trên các hệ thống và ứng dụng phức tạp, không giống như các hệ thống cơ sở dữ liệu một chiều đơn giản.
* Tính mô-đun trong loại Cơ sở dữ liệu này là một sự khuyến khích cho các dự án có băng thông thấp hơn cho nhân viên bảo trì.
* Nhìn chung, năng lực tổ chức và định nghĩa cấu trúc của Mô hình dữ liệu đa chiều hỗ trợ giữ dữ liệu sạch hơn và đáng tin cậy trong cơ sở dữ liệu.
* Việc xây dựng các vị trí dữ liệu được xác định rõ ràng làm cho nó không phức tạp, trong các tình huống như một nhóm xây dựng cơ sở dữ liệu, một nhóm khác làm việc trên cơ sở dữ liệu đó và một số nhóm khác làm việc bảo trì
* Vì hệ thống mới và không có rác, hiệu quả của dữ liệu và hiệu suất của hệ thống cơ sở dữ liệu được nâng cao & nâng cao.

**Nhược điểm**

* Vì Mô hình Dữ liệu Đa chiều xử lý các hệ thống phức tạp nên các loại cơ sở dữ liệu này thường phức tạp về bản chất.
* Là một hệ thống phức tạp có nghĩa là nội dung của cơ sở dữ liệu cũng rất lớn. Điều này làm cho hệ thống có rủi ro cao khi có vi phạm an ninh.
* Khi hệ thống lưu vào bộ nhớ cache do các hoạt động trên Mô hình dữ liệu đa chiều, hiệu suất của hệ thống bị ảnh hưởng rất nhiều.
* Mặc dù sản phẩm cuối cùng trong Mô hình dữ liệu đa chiều là thuận lợi, nhưng con đường để đạt được nó hầu hết thời gian đều phức tạp.

## Khái quát khai phá dữ liệu

### Khái niệm

**Khai phá dữ liệu** là kỹ thuật có sự hỗ trợ của máy tính được sử dụng trong hoạt động phân tích để xử lý và khám phá các tập dữ liệu lớn. Nhờ có công cụ và phương pháp khai thác dữ liệu, các tổ chức có thể khám phá những mẫu hình và mối quan hệ ẩn trong dữ liệu của họ. Quá trình khai thác dữ liệu chuyển đổi dữ liệu thô thành kiến thức thực tế.

Quy trình tiêu chuẩn liên ngành đối với khai thác dữ liệu (Cross-Industry Standard Process for Data Mining, CRISP-DM) là một hướng dẫn tuyệt vời để bắt đầu quy trình khai thác dữ liệu. CRISP-DM vừa là phương pháp luận, vừa là mô hình quy trình trung lập với ngành, công cụ và ứng dụng.

* Với vai trò phương pháp luận, CRISP-DM mô tả các giai đoạn điển hình trong một dự án khai thác dữ liệu, phác thảo những nhiệm vụ liên quan trong mỗi giai đoạn và giải thích mối quan hệ giữa những nhiệm vụ này.
* Với vai trò mô hình quy trình, CRISP-DM cung cấp một cái nhìn tổng quan về vòng đời khai thác dữ liệu.

### Qui trình khai phá dữ liệu

Thông qua các giai đoạn CRISP-DM linh hoạt, đội ngũ dữ liệu có thể di chuyển qua lại giữa các giai đoạn nếu cần. Ngoài ra, công nghệ phần mềm có thể thực hiện hoặc hỗ trợ một số nhiệm vụ này.

**Bước 1: Hiểu biết về doanh nghiệp**

Nhà khoa học dữ liệu hoặc người khai thác dữ liệu bắt đầu bằng cách xác định các mục tiêu và phạm vi dự án. Họ hợp tác với các bên liên quan của doanh nghiệp để xác định một số thông tin nhất định.

Vấn đề cần giải quyết

Ràng buộc hoặc giới hạn của dự án

Tác động kinh doanh của các giải pháp tiềm năng

Sau đó, họ sử dụng thông tin này để xác định mục tiêu khai thác dữ liệu cũng như nhận định tài nguyên cần có để khai phá kiến thức.

**Bước 2: Hiểu biết về dữ liệu**

Khi đã nắm được vấn đề kinh doanh, các nhà khoa học dữ liệu bắt đầu phân tích sơ bộ dữ liệu. Họ thu thập các tập dữ liệu từ nhiều nguồn khác nhau, lấy được quyền truy cập và chuẩn bị báo cáo mô tả dữ liệu. Báo cáo này bao gồm các loại dữ liệu, số lượng cũng như yêu cầu về phần cứng và phần mềm để xử lý dữ liệu.

**Bước 3: Chuẩn bị dữ liệu**

Người khai thác dữ liệu dành nhiều thời gian nhất cho giai đoạn này do phần mềm khai thác dữ liệu yêu cầu dữ liệu chất lượng cao. Các quy trình kinh doanh thu thập và lưu trữ dữ liệu vì nhiều lý do khác ngoài việc khai thác và người khai thác dữ liệu phải tinh chỉnh dữ liệu trước khi sử dụng để lập mô hình. Chuẩn bị dữ liệu bao gồm các quy trình sau đây.

* **Làm sạch dữ liệu**: xử lý dữ liệu bị thiếu, lỗi dữ liệu, giá trị mặc định và hiệu chỉnh dữ liệu.
* **Tích hợp dữ liệu**: kết hợp hai tập dữ liệu riêng biệt để có được tập dữ liệu đích cuối cùng.
* **Định dạng dữ liệu**: chuyển đổi loại dữ liệu hoặc cấu hình dữ liệu cho công nghệ khai thác cụ thể đang được sử dụng.

**Bước 4: Lập mô hình dữ liệu**

Người khai thác dữ liệu nhập dữ liệu đã chuẩn bị vào phần mềm khai thác dữ liệu và nghiên cứu kết quả. Để làm điều này, họ có thể chọn trong số nhiều kỹ thuật và công cụ khai thác dữ liệu. Để lập mô hình dữ liệu, các nhà khoa học dữ liệu có thể:

* Đào tạo mô hình máy học (ML) trên các tập dữ liệu nhỏ hơn bằng kết quả đã biết
* Sử dụng mô hình để phân tích thêm các tập dữ liệu chưa biết
* Điều chỉnh và cấu hình lại phần mềm khai thác dữ liệu cho đến khi kết quả thỏa mãn yêu cầu

**Bước 5: Đánh giá**

Sau khi tạo mô hình, người khai thác dữ liệu bắt đầu đo lường mô hình so với mục tiêu kinh doanh ban đầu. Người khai thác dữ liệu có thể thay đổi mô hình, điều chỉnh mục tiêu kinh doanh hoặc xem xét sửa đổi dữ liệu tùy thuộc vào phản hồi của doanh nghiệp. Đánh giá, phản hồi và sửa đổi liên tục là một phần của quy trình khai phá kiến thức.

**Bước 6: Triển khai**

Trong quá trình triển khai, những bên liên quan khác sử dụng mô hình làm việc để tạo ra nghiệp vụ thông minh. Nhà khoa học dữ liệu lên kế hoạch cho quy trình triển khai. Các chuyên viên phân tích nghiệp vụ sử dụng ứng dụng này để tạo báo cáo quản lý, chia sẻ kết quả với khách hàng và cải tiến quy trình kinh doanh.

### Các phương pháp khai phá dữ liệu

* Phương pháp luật kết hợp

Một trong những chủ đề phổ biến của KPDL là khai phá luật kết hợp. Mục đích của khai phá luật kết hợp là xác định mối quan hệ, sự kết hợp giữa các mục dữ liệu (item) trong một CSDL lớn.

* Phương pháp cây quyết định

Mô tả tri thức dạng đơn giản nhằm phân các đối tượng dữ liệu thành một số lớp nhất định. Các nút của cây được gán nhãn là tên các mục dữ liệu, các cạnh được gán các giá trị có thể của các mục dữ liệu, các lá mô tả các lớp khác nhau. Các đối tượng được phân lớp theo các đường đi trên cây, qua các cạnh tương ứng với các giá trị của mục dữ liệu tới lá.

* Phương pháp K-Mean

Có nhiều phương pháp được sử dụng trong phân cụm, phương pháp k-Mean được coi là các kỹ thuật cơ bản của phân cụm. Với phương pháp này sẽ chia tập có n đối tượng thành k cụm sao cho các đối tượng trong cùng một cụm thì giống nhau, các đối tượng khác cụm thì khác nhau.

* Các phương pháp dựa trên mẫu

Phương pháp này sử dụng khai phá chuỗi theo thời gian (Sequential temporal patterns). Xét về mặt kỹ thuật thì tương tự như KPDL bằng luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán bởi chúng có tính dự báo cao.

### Các ứng dụng của khai phá dữ liệu

* Phân tích dữ liệu tài chính (Financial Data Analysis)

Dữ liệu tài chính trong ngân hàng và trong ngành tài chính nói chung thường đáng tin cậy và có chất lượng cao, tạo điều kiện cho khai phá dữ liệu. Dưới đây là một số ứng dụng điển hình trong khai phá dữ liệu tài chính:

* Dự đoán khả năng vay và thanh toán của khách hàng, phân tích chính sách tín dụng đối với khách hàng.
* Phân tích hành vi khách hàng (vay, gửi tiền)
* Phân loại và phân nhóm khách hàng mục tiêu cho tiếp thị tài chính
* Phát hiện các hoạt động rửa tiền và tội phạm tài chính khác
* Công nghiệp bán lẻ (Retail Industry)

Khai phá dữ liệu có vai trò rất quan trọng trong ngành công nghiệp bán lẻ, do dữ liệu  thu thập từ lĩnh vực này rất lớn từ doanh số bán hàng, lịch sử mua hàng của khách hàng, vận chuyển hàng hóa, tiêu thụ và dịch vụ. Dưới đây là một số ứng dụng của khai phá dữ liệu trong ngành công nghiệp bán lẻ:

* Khai phá dữ liệu trên kho dữ liệu khách hàng
* Phân tích đa chiều trên kho dữ liệu khách hàng về doanh số bán hàng, khách hàng, sản phẩm, thời gian và khu vực.
* Phân tích hiệu quả của các chiến dịch bán hàng, Marketing.
* Quản trị mối quan hệ khách hàng (CRM).
* Giới thiệu và tư vấn sản phẩm phù hợp cho khách hàng
* Công nghiệp viễn thông (Telecommunication Industry)

Công nghiệp viễn thông là một trong những ngành công nghiệp mới nổi, cung cấp nhiều dịch vụ như  trên điện thoại di động, Internet, truyền hình ảnh.... Do sự phát triển mạnh của công nghệ máy tính và mạng máy tính, viễn thông đang phát triển với tốc độ rất lớn. Dưới đây là một số ứng dụng của khai phá dữ liệu trong ngành công nghiệp này:

* Phân tích dữ liệu đa chiều viễn thông.
* Xây dựng các mô hình phát hiện gian lận.
* Phát hiện bất thường trong giao dịch viễn thông.
* Phân tích hành vi sử dụng dịch vụ viễn thông của khách hàng
* Sử dụng các công cụ trực quan trong phân tích dữ liệu viễn thông
* Phân tích dữ liệu sinh học (Biological Data Analysis)

Khai phá dữ liệu sinh học là một phần rất quan trọng của lĩnh vực Tin -Sinh học (Bioinformatics). Sau đây là một số ứng dụng của khai phá dữ liệu ứng dụng trong sinh học:

* Lập chỉ mục, tìm kiếm tương tự, bất thường trong cơ sở dữ liệu Gen.
* Xây dựng mô hình khai phá các mạng di truyền và cấu trúc của Gen, protein
* Xây dựng các công cụ trực quan trong phân tích dữ liệu di truyền.
* Phát hiện xâm nhập bất hợp pháp (Intrusion Detection)

Xâm nhập bất hợp pháp là những hành động đe dọa tính toàn vẹn, bảo mật và tính sẵn sàng của tài nguyên mạng. Với sự phát triển của internet và sự sẵn có của các công cụ, thủ thuật trợ giúp cho xâm nhập và tấn công mạng, yêu cầu kiểm soát truy cập bất hợp pháp là yếu tố rất quan trọng đảm bảo cho sự ổn định của hệ thống.

*Dưới đây là một số ứng dụng của khai phá dữ liệu có thể được áp dụng để phát hiện xâm nhập:*

* Phát triển các thuật toán khai phá dữ liệu để phát hiện xâm nhập.
* Phân tích kết hợp, tương quan và khác biệt để phát hiện xâm nhập
* Phân tích dòng dữ liệu dữ liệu (Analysis of Stream data) để phát hiện bất thường.

## Phương pháp khai phá dữ liệu được sử dụng trong đề tài

### Thuật toán Apriori

Thuật toán Apriori được công bố bởi R. Agrawal và R. Srikant vào năm 1994 vì để tìm các tập phổ biến trong một bộ dữ liệu lớn. Tên của thuật toán là Apriori vì nó sử dụng kiến thức đã có từ trước (prior) về các thuộc tính, vật phẩm thường xuyên xuất hiện trong cơ sở dữ liệu. Để cải thiện hiệu quả của việc lọc các mục thường xuyên theo cấp độ, một thuộc tính quan trọng được sử dụng gọi là thuộc tính Apriori giúp giảm phạm vi tìm kiếm của thuật toán.

*Thuộc tính Apriori*

Tất cả các tập hợp con không rỗng của tập thường xuyên cũng phải thường xuyên. Khái niệm chủ chốt này của thuật toán Apriori nhằm chống lại đơn điệu của phương pháp tính theo độ hỗ trợ (surport). Apriori cho rằng:

“*Tất cả các tập con của một tập hợp thường xuyên phải là thường xuyên (thuộc tính Apriori).Trong một vật phẩm không thường xuyên, tất cả các tập cha của nó sẽ không thường xuyên. Hãy xem xét các tập dữ liệu sau đây và chúng ta sẽ tìm thấy các tập thường xuyên và tạo quy tắc kết hợp cho chúng.”*

**Ví dụ minh họa:**

Bảng 2.1. Bảng dữ liệu để ví dụ về Apriori

|  |  |
| --- | --- |
| TID | Items |
| T1 | bread, cornflakes, eggs, jam |
| T2 | bread, cornflakes, jam |
| T3 | bread, milk, tea |
| T4 | bread, jam, milk |
| T5 | cornflakes, jam, milk |

Với độ hỗ trợ tối thiểu (min\_support) là 50% và độ tin cậy tối thiểu (min\_confidence) là 70%

***Bước 1:*** Tạo bảng chứa số support của từng mục có trong tập dữ liệu (hay nói cách khác là đếm từng Item trong thuộc tính ‘Items’ của bảng trên) - Được gọi là C1 (tập ứng cử viên)

Bảng 2.2. Bảng thống kê các phần tử của dữ liệu (C1)

|  |  |
| --- | --- |
| **Item\_set** | **Sup\_Count** |
| Bread | 4 |
| Cornflakees | 3 |
| Eggs | 1 |
| Milk | 3 |
| Jam | 4 |
| Tea | 1 |

(II) so sánh số support (cột Sup\_Count) của tập C1 với số lượng hỗ trợ tối thiểu là 2.5 (ở đây min\_support = 50% và tập dữ liệu bao gồm 5 dòng nên số lượng hỗ trợ sẽ bằng: 5 \* 50% = 2.5). Xét các Item\_set nếu support\_count nhỏ hơn số lượng hỗ trợ sẽ xóa các tập đó. Điều này cung cấp cho chúng ta mục L1.

Bảng 2.3. Bảng dữ liệu sau khi xét min\_support (L1)

|  |  |
| --- | --- |
| **Item\_set** | **Sup\_Count** |
| Bread | 4 |
| Cornflakes | 3 |
| Milk | 3 |
| Jam | 4 |

***Bước 2:*** Tạo tập ứng viên C2 bằng L1 (đây được gọi là bước kết hợp). Điều kiện: Giữ lại các cặp phân biệt, bỏ tập các món hàng trùng lặp; Các món hàng giống nhau khác trật tự thì giữ lại 1 tập. Duyệt qua các tập cha của C2, nếu tập cha nào không đạt chuẩn thường xuyên thì tập con đó sẽ bị xóa.

Bảng 2.4. Bảng sau khi kết hợp các Item ở L1

|  |  |  |  |
| --- | --- | --- | --- |
| **C2** | | | |
| Bread, Bread | Cornflakes, Bread | Milk, Bread | Jam, Bread |
| Bread, Cornflakes | Cornflakes, Cornflakes | Milk, Cornflakes | Jam, Cornflakes |
| Bread, Milk | Cornflakes, Milk | Milk, Milk | Jam, Milk |
| Bread, Jam | Cornflakes, Jam | Milk, Jam | Jam, Jam |

Bảng 2.5. . Bảng dữ liệu sau khi loại bỏ các phần tử theo điều kiện

|  |  |
| --- | --- |
| **C2** | |
| Bread, Cornflakes | Cornflakes, Milk |
| Bread, Milk | Cornflakes, Jam |
| Bread, Jam | Milk, Jam |

Thống kê số support của từng mục có trong tập dữ liệu đầu – Tạo ra tập gọi là C2

Bảng 2.6. Bảng thống kê các phần tử của dữ liệu (C2)

|  |  |
| --- | --- |
| **Item\_set** | **Sup\_Count** |
| Bread, Cornflakes | 2 |
| Bread, Milk | 2 |
| Bread, Jam | 3 |
| Cornflakes, Milk | 1 |
| Cornflakes, Jam | 3 |
| Milk, Jam | 2 |

Tiếp tục kiểm tra độ hỗ trợ của các tập trong C2, nếu tập nào không thỏa mãn min\_support thì xóa đi. Ta sẽ nhận được kết quả là tập L2.

Bảng 2.7. Bảng dữ liệu sau khi xét min\_support (L2)

|  |  |
| --- | --- |
| **Item\_set** | **Sup\_Count** |
| Bread, Jam | 3 |
| Cprrnflakes, Jam | 3 |

***Bước 3:*** Lặp lại quy trình như bước 2 ta được hai tập.

C3:

Bảng 2.8. Bảng dữ liệu sau khi hoàn thành kết hợp (C3)

|  |  |
| --- | --- |
| **Item\_set** | **Sup\_Count** |
| Bread, Jam, Cornflakes | 1 |

Tới đây khi ta xét độ hỗ trợ của các tập trong C3 thì ta thấy không thỏa mãn để tiếp tục tạo ra tập L3 nên việc chạy thuật toán tới đây là kết thúc

Tiếp theo chúng ta cần xác định các luật kết hợp thỏa độ tin cậy tối thiểu

Ví dụ về sự tin cậy - Độ tin cậy 70% có nghĩa là 70% khách hàng mà mua sữa và bánh mì cũng sẽ mua bơ.

Confidence(A->B) = Support\_count(A∪B)/Support\_count(A). Ở đây chúng ta sẽ lấy ví dụ với các tập tìm được ở trên bao gồm tập: L2 (Thõa mãn), L3 (Không thõa mãn -> Loại).

Các luật được sinh ra từ L2 và độ tin cậy của mỗi luật

Bảng 2.9. Bảng thể hiện luật và độ tin cậy sinh ra từ L2

|  |  |  |  |
| --- | --- | --- | --- |
| **Rules (Luật)** | **Support\_count (A∪B)** | **Support\_count (A)** | **Độ tin cậy (Confidence )** |
| Bread -> Jam | 3 | 4 | 3/4 = 75% |
| Jam -> Bread | 3 | 4 | 3/4 = 75% |
| CornFlakes -> Jam | 3 | 3 | 3/3 = 100% |
| Jam -> Cornflakes | 3 | 4 | 3/4 = 75% |

Với kết quả trên, nếu độ tin cậy tối thiểu là 70%, thì sinh ra được 4 luật như trên.

### Thuật toán FP-Growth

Thuật toán này là một cải tiến của phương pháp Apriori. Một mẫu phổ biến được tạo ra mà không cần tạo ứng viên. Thuật toán tăng trưởng FP biểu diễn cơ sở dữ liệu dưới dạng cây được gọi là cây mẫu phổ biến hoặc cây FP.

Cấu trúc cây này sẽ duy trì sự liên kết giữa các tập mục. Cơ sở dữ liệu được phân mảnh bằng cách sử dụng một mục phổ biến. Phần bị phân mảnh này được gọi là "mảnh mẫu". Các tập mục của các mẫu phân mảnh này được phân tích. Vì vậy, với phương pháp này, việc tìm kiếm các tập phổ biến được giảm đi một cách tương đối.

***Các bước thực hiện***

**Bước 1 : Nén cơ sở dữ liệu giao dịch gốc vào cây FP-tree**

* Quét cơ sở dữ liệu một lần, tìm các tập phổ biến **1-itemsets** (chỉ có một hạng mục hay phần tử).
* Sắp xếp các tập phổ biến tìm được theo thứ tự giảm dần của độ phổ biến (tần số).
* Quét lại cơ sở dữ liệu lần 2, xây dựng một cây FP-tree bắt đầu với hạng mục phổ biến nhất trong mỗi giao dịch

**Bước 2: Các bước chính để khai thác các tập phổ biến trên cây FP- tree - cây , FP -tree có điều kiện**

* Duyệt từng hạng mục phổ biến **(1-itemsets)** theo thứ tự tăng dần của tần số. Với mỗi hạng mục, xây dựng cơ sở mẫu điều kiện và các cây **FP-tree** có điều kiện tương ứng của nó
* Bắt đầu với hạng mục có tần suất xuất hiện thấp nhất. Cơ sở mẫu điều kiện của nó là tất cả các đường đi tiền tố của cây **FP-Tree** khi duyệt từ nút gốc
* Xây dựng cây **FP-Tree có điều kiện** từ mẫu trên bằng cách trộn tất cả các đường đi và giữ lại các nút có tần số thõa mãn yêu cầu.

**Ưu điểm:**

* Thuật toán này chỉ cần quét cơ sở dữ liệu hai lần so với Apriori quét các giao dịch cho mỗi lần lặp.
* Việc ghép nối các mục không được thực hiện trong thuật toán này và điều này làm cho nó nhanh hơn.
* Cơ sở dữ liệu được lưu trữ trong một phiên bản nhỏ gọn trong bộ nhớ.
* Nó hiệu quả và có khả năng mở rộng để khai thác cả các mẫu phổ biến dài và ngắn.

**Nhược điểm:**

* FP Tree cồng kềnh và khó xây dựng hơn Apriori.
* Nó có thể đắt tiền.
* Khi cơ sở dữ liệu lớn, thuật toán có thể không phù hợp với bộ nhớ dùng chung.

**Ví dụ minh họa:**

Ví dụ: xét tập dữ liệu có 5 giao dịch và 6 món hàng. Tìm các luậtkết hợp với ***độ hỗ trợ 50%*** và ***độ tin cậy 75%*** bằng thuật toán FP-Growth

**Giải:**

Bảng 2.10. Bảng dữ liệu mẫu để chạy FB - Growth

|  |  |
| --- | --- |
| **Tip** | **Items** |
| T1 | Bread, cornflakes,eggs,jam |
| T2 | Bread,conflakes,jam |
| T3 | Bread, Milk, Tea |
| T4 | Bread, Jam, Milk |
| T5 | Cornflakes, Jam, Milk |

**Bước 1: Nén cơ sở dữ liệu giao dịch gốc vào cây FP-tree**

Bảng 2.11. Bảng thống kê Items (C1)

|  |  |
| --- | --- |
| **Item** | **Tần suất** |
| Bread | 4 |
| Jam | 4 |
| Conflakes | 3 |
| Milk | 3 |
| Tea | 1 |

***Độ hỗ trợ Minsupp=0.5 =>***

Bảng 2.12. Bảng dữ liệu sau khi xét min\_support (L1)

|  |  |
| --- | --- |
| **Item** | **Tần suất** |
| Bread | 4 |
| Jam | 4 |
| Conflakes | 3 |
| Milk | 3 |

Tiếp theo sắp xếp các mục phổ biến mức 1 vừa tìm được theo thứ tự giảm dần trong mỗi giao dịch:

Bảng 2.13. Bảng xét cái Items phỏ biến từ bảng L1

|  |  |  |
| --- | --- | --- |
| **Tip** | **Items** | **Items phổ biến** |
| T1 | Bread, cornflakes,eggs,jam | Bread, Jam, Conflakes |
| T2 | Bread,conflakes,jam | Bread, jam, conflakes |
| T3 | Bread, Milk, Tea | Bread, Milk |
| T4 | Bread, Jam, Milk | Bread, Jam, Milk |
| T5 | Cornflakes, Jam, Milk | Jam, Cornflakes, Milk |

Duyệt các Items phổ biến của mỗi giao dịch để xây dựng FP-Tree

Diagram

Description automatically generated

Hình 2.1. Sơ đồ cây để duyệt các Items

***Bước 2:*** Các bước chính để khai thác các tập phổ biến trên cây FP- tree - cây , FP -tree có điều kiện

Thứ tự tăng dần theo tần số xuất hiện: Milk, Conflakes, Jam, Break

Xét items **Milk** ta có:

Xây dựng cây ***FP-Tree có điều kiện*** từ mẫu trên bằng cách trộn tất cả các đường đi và giữ lại các nút có tần số ⩾2,5 **do minsupp**=0.5:

Các đường đi tới Milk: **Break-Jam**: tần suất 1 lần, **Break**: 1 lần, **Jam-Conflakes**: 1 lần

* **Bread**: 2 lần, **Jam**: 2 lần, **conflakes**: 1 lần
* Các mẫu phổ biến của Milk là: Milk

Tương tự với các item trên ta có bảng sau:

Bảng 2.14. Bảng thể hiện các luật được sinh ra

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | **Cơ sở điều kiện** | **FP-Tree điều kiện** | **Các mẫu phổ biến** |
| Milk | { Bread-Jam: 1, Bread:1, Jam- conflakes: 1} | ∅ | Milk |
| Conflakes | { Bread-jam: 2, Jam: 1} | {Jam: 3}-Conflakes | Conflakes, Jam-conflakes |
| Jam | {Break: 3} | {Bread:3}-Jam | Jam, Bread-Jam |
| Break | ∅ | ∅ | Bread |

Ta có các luật sau:

Bảng 2.15. Bảng dữ liệu chính

|  |  |
| --- | --- |
| Tip | Items |
| T1 | Bread, cornflakes,eggs,jam |
| T2 | Bread,conflakes,jam |
| T3 | Bread, Milk, Tea |
| T4 | Bread, Jam, Milk |
| T5 | Cornflakes, Jam, Milk |

Jam-> Conflakes: 0.75

Conflakes -> Jam: 1

Bread -> Jam: 0.75

Jam -> Bread: 0.75

# PHẦN MỀM KHAI PHÁ DỮ LIỆU MÃ NGUỒN MỞ

## WEKA

### Giới thiệu

Weka là một phần mềm nguồn mở bằng Java được thiết kế để cung cấp các công cụ để xử lý trước dữ liệu, triển khai một số thuật toán Học máy và các công cụ trực quan để bạn có thể phát triển các kỹ thuật học máy và áp dụng chúng vào các vấn đề khai thác dữ liệu trong thế giới thực.

Ngoài ra thiết kế để giúp các doanh nghiệp và cá nhân quản lý phân loại dữ liệu, chuẩn bị, phân cụm, hồi quy, liên kết, trực quan hóa và khai thác quy tắc. Nó cho phép người dùng thử các phương pháp hiện có trên bộ dữ liệu mới theo nhiều cách.

### Chức năng

Các tính năng chính của WEKA:

**Tiền xử lý dữ liệu**: là một nhiệm vụ quan trọng trong khai phá dữ liệu. WEKA đưa ra một bộ tùy chọn toàn diện trong danh mục bộ lọc . Ở đây, công cụ này cung cấp cả loại hoạt động được giám sát và không được giám sát.

**Phân loại:** là một trong những chức năng thiết yếu trong học máy, nơi chúng tôi gán các lớp hoặc danh mục cho các mục.

**Phân cụm:** một tập dữ liệu được sắp xếp theo các nhóm/cụm khác nhau dựa trên một số điểm tương đồng. Trong trường hợp này, các mục trong cùng một cụm giống hệt nhau nhưng khác với các cụm khác.

**Kết hợp:** Các quy tắc kết hợp làm nổi bật tất cả các kết hợp và mối tương quan giữa các mục của tập dữ liệu. Công cụ cung cấp các thuật toán Apriori , FilteredAssociator, FPGrowth để khai phá luật kết hợp .

**Chọn thuộc tính:** loại bỏ những thứ không cần thiết và giữ lại những chi tiết liên quan là rất quan trọng để xây dựng một mô hình tốt.

**Hình dung (Mô hình):** Trong tab trực quan hóa, các biểu đồ và ma trận cốt truyện khác nhau có sẵn để hiển thị các xu hướng và lỗi được mô hình xác định.

## Rstudio

### Giới thiệu

RStudio là một IDE nguồn mở linh hoạt và đa chức năng (môi trường phát triển tích hợp) được sử dụng rộng rãi làm giao diện người dùng đồ họa để hoạt động với R của phiên bản 3.0.1 trở lên. Ngoài ra, nó cũng thích ứng với nhiều ngôn ngữ lập trình khác, chẳng hạn như Python hoặc SQL.

### Các tính năng chính của Rstudio

**Source Editor:** Phát triển chương trình trong một cửa sổ console duy nhất bằng cách tổng hợp và tích hợp tất cả các công cụ đang sử dụng. Làm nổi bật cú pháp, xác định chức năng và hoàn thành mã trong bảng điều khiển.

**Ứng dụng web:** Xuất bản ứng dụng, bảng điều khiển và tài liệu lên web bằng gói Shiny Server R được phát triển nội bộ.

**Flexdashboard:** Phát triển bảng điều khiển tương tác với trực quan hóa JavaScript, với sự hỗ trợ cho các tiện ích HTML, với gói R này.

**Trình khởi chạy:** Khởi chạy các quy trình R và Python từ xa hoặc gửi tập lệnh R để tính toán các cụm như SLURM hoặc Kubernetes.

**Tạo API REST:** Phát triển các kết nối API web với gói R của thợ sửa ống nước chỉ với một dòng mã.

**RStudio Connect:** Xuất bản các phân tích dữ liệu thống kê dưới dạng trực quan hóa có tác động trực quan, tổng hợp tất cả các tính năng xuất bản nói trên vào một giao diện.

**Dự đoán và mô hình hóa dữ liệu:** Cung cấp khả năng phân tích dự đoán và theo quy định thông qua khả năng kết nối với TensorFlow. Cải thiện khả năng lập mô hình dữ liệu riêng của sản phẩm thông qua gói Tidymodels R.

# KHAI PHÁ DỮ LIỆU

## Xác định vấn đề

Việc tìm việc làm của những người lao động trong những năm gần đây trở nên khó khăn cho nhiều yếu tố chủ quan và khách quan khác nhau. Bên cạch đó các doanh nghiệp cũng gặp khó khăn trong việc tìm kiếm người lao đông cho mình.

Trước vấn đề đó thì chúng em đã tìm được những công việc đang có nhu cầu và thống kê thành một bảng dữ liệu lớn.

Việc xử lý dữ liệu cũng là một vấn đề. Giải pháp của chúng em là sẽ sử dụng thuật toán kết hợp thông qua 2 thuật toán: Apriori và FP-Growth thông qua 2 ứng dụng là Weka và R để đưa ra cách nhìn khách quan về vấn đề tìm việc và ứng tuyển cho người xin việc

## Hiểu dữ liệu

Bảng 4.1. Bảng mô tả các thuộc tính và miền dữ liệu của thuộc tính

|  |  |  |
| --- | --- | --- |
| ***STT*** | ***Tên dữ liệu*** | ***Tập giá trị*** |
| 1 | Tinh thanh tuyen dung (Tỉnh thành tuyển dụng) | 'Ho Chi Minh','Ha Noi','Binh Duong','Dong Nai','Long An','Thua Thien Hue','Gia Lai', 'Quang Ninh', 'Da Nang', 'Hai Duong', 'Thanh Hoa', 'Binh Dinh','Bac Giang', 'Ho Chi Minh, Ho Chi Minh', 'Hai Phong','Quang Ngai, Da Nang', 'Bac Ninh','Quang Ngai', 'Lang Son', 'Phu Yen','Can Tho', 'Thai Binh','Ho Chi Minh, Ha Noi','Ho Chi Minh, Ha Noi, Binh Duong, Da Nang, Hai Phong','Lam Dong','Quang Nam','Hung Yen','Tay Ninh','Ho Chi Minh, Long An, Vinh Long, Dong Thap, Tien Giang','Tien Giang','Dong Thap','Dak Lak', 'Toan quoc', 'Binh Thuan', 'Nam Dinh','Nghe An', 'Ca Mau','Vinh Phuc','Khanh Hoa','Ho Chi Minh, Binh Duong, Dong Nai','Phu Tho','Ba Ria Vung Tau','Binh Phuoc','Ho Chi Minh, Binh Duong','Ben Tre', 'Thai Binh, Nam Dinh, Hung Yen, Bac Giang, Hai Phong', 'Ha Nam', 'An Giang', 'Ha Tinh', 'Quang Tri','Ho Chi Minh, Hai Phong','Thai Nguyen','Quang Binh','Ha Noi, Hai Phong', 'Kien Giang','Bac Lieu','Son La','Hung Yen, Ha Nam','Hoa Binh','Ha Noi, Bac Ninh','Da Nang, Ho Chi Minh, Ha Noi, Binh Duong, Hai Phong','Vinh Long','Ninh Thuan','Ninh Binh','Tien Giang, Tra Vinh','Kon Tum','Ha Giang' |
| 2 | Chuc vu (Chức vụ) | 'Nhan vien', 'Truong Phong', 'Truong nhom', 'Thuc tap sinh', 'Moi Tot Nghiep','Giam Doc','Pho phong','Quan ly cap trung','Quan ly cap cao' |
| 3 | Muc luong (Mức lương) | '7 - 10 trieu','10 - 15 trieu','15 - 20 trieu', 'Thoa thuan','20 - 30 trieu','5 - 7 trieu','1 - 3 trieu','3 - 5 trieu','Tren 30 trieu','Tren 50 trieu', 'Tren 100 trieu' |
| 4 | Hinh thuc lam viec (Hình thức làm việc) | 'Toan thoi gian co dinh','Ban thoi gian','Toan thoi gian tam thoi',Khac,'Hop dong','Ban thoi gian tam thoi' |
| 5 | Nhu cau (Nhu cầu) | '\'(-inf-50.666667]\'','\'(50.666667-100.333333]\'','\'(100.333333-inf)\'' |
| 6 | Nganh nghe (Ngành nghề) | 'KD bat dong san, Nhan vien kinh doanh, Tu van','Ke toan - Kiem toan','Xay dung','Cham soc khach hang, Tu van, Lao dong pho thong','KD bat dong san','Sinh vien lam them, Nhap lieu, Nganh nghe khac','Khach san - Nha hang, Xay dung, Dau bep - phu bep','Thiet ke web, Truyen thong, Thiet ke - My thuat','Sinh vien lam them, Dau bep - phu bep, Y te - Duoc','Kien truc - Tk noi that, Xay dung','Hanh chinh - Van phong, Y te - Duoc, Cong nghe thuc pham',… |
| 7 | Linh vuc (Lĩnh vực) | ‘nhan vien phat trien mat bang','ke toan noi bo','Ky Su Du Toan','cong tac vien moi gioi bds','nhan vien tu van khach hang','Nhan Vien Kinh Doanh Nha Dat','nhan vien nhap lieu excel','dau bep luong cao','Ui Ux Designer','cong tac vien duoc pham','tro ly giam doc nhan su','nhan vien van phong trinh do 12/12','ke toan tong hop','chuyen vien kinh doanh','nhan vien cham soc khach hang online','nhan vien ky thuat','nhan vien tu van','chuyen vien hoach dinh tai chinh','sinh vien xay dung lam them','part time khach san','y si dong y','Truong Phong Kinh Doanh','nhan vien ke toan kho','Nhan Vien Kinh Doanh Duoc Pham','truong phong ke toan','telesale online','Nhan Vien Kinh Doanh Logistics','nhan vien ban hang luong cao','nhan vien tuyen dung','Nhan Vien Kinh Doanh My Pham','chuyen vien tu van bat dong san','tu van bat dong san','bien phien dich tieng anh','fresher javascript',… |
| 8 | Kinh nghiem (Kinh nghiệm) | 'Khong yeu cau','1 - 2 nam kinh nghiem','2 - 5 nam kinh nghiem','0 - 1 nam kinh nghiem','5 - 10 nam kinh nghiem','Hon 10 nam kinh nghiem' |
| 9 | Id cong ty (Id công ty) | 295,593,607,672,1286,1435,1952,2120,2517,2633,2731,2810,2886,3227,4367,4814,4844,5006,5609,5630,5753,5795,5874,5898,6180,6633,6854,6932,7337,7449,9142,9186,10030,10181,10189,10201,10485,… |

## Chuẩn bị dữ liệu

### Nạp dữ liệu

#### Nạp dữ liệu bằng file.csv

Để tải tải file.csv lên weka, chúng ta thực hiện các bước sau:

***Bước 1:*** Mở Weka -> Explorer

Graphical user interface, text

Description automatically generated

Hình 4.1. Giao diện khi mở WEKA

***Bước 2:*** Chọn chọn tập dữ liệu cần lấy

Ở phần **Preprocess** -> **Open file..** Và sau đó chọn tập dữ liệu cần lấy có kiểu file.csv và chọn **Open**

Graphical user interface, text, application

Description automatically generated

Hình 4.2. Gio diện khi mở Explorer và thực hiện thao tác Open file

***Bước 3:*** Xem dữ liệu

Tại đây người dùng xem thuộc tính của dữ liệu và kiểu của chúng.

Để xem, người dùng chọn edit để xem tập dữ liệu

Graphical user interface

Description automatically generated

Hình 4.3. Giao diện khi thực hiên thành công Open file và xem dữ liệu trên Edit

#### Thực hiện kết nối với cơ sở dữ liệu SQL SERVER để nạp dữ liệu

Để thực hiện kết nối với cơ sở dữ liệu người dùng cần chuẩn bị:

* SQL Server 2014
* Weka 3.8.5
* Microsoft JDBC 8.4:

(<https://learn.microsoft.com/vi-vn/sql/connect/jdbc/release-notes-for-the-jdbc-driver?view=sql-server-ver15#previous-releases>)

Graphical user interface, text, application

Description automatically generated

Hình 4.4. Link để tải các JDBC

Các bước để thực hiện kết nối :

* Sau khi tải về Microsoft JDBC 8.4 thực hiện chuyển file vào thư mục “Microsoft JDBC Driver 6.0 for SQL Server”.
* Chúng ta cần thêm đường dẫn đầy đủ và tên tệp của trình điều khiển Microsoft JDBC 8.4 vào biến môi trường hệ thống CLASSPATH.

Graphical user interface

Description automatically generated with medium confidence

Hình 4.5. Thư mục khi tải thành công JDBC và chọn phiên bản ‘jre11’

Copy đường dẫn ở hình trên. Sau đó chọn biểu tượng của sổ *Windows trên máy -> settings -> About -> Advanced system settings*

Graphical user interface, text, application, email

Description automatically generated

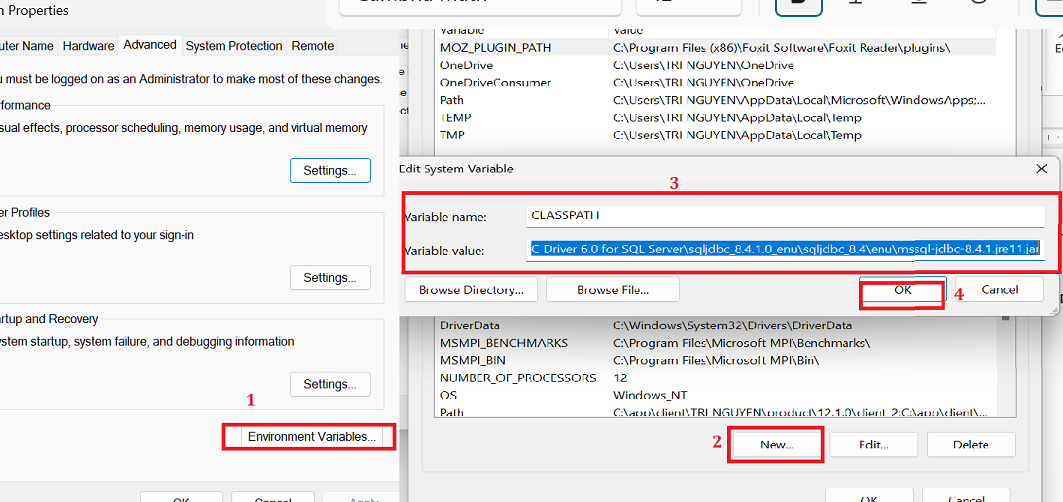
Hình 4.6. Giao diện *About* trong phần Setting trên máy tính

Sau đó chọn chọn Environment Variables.. -> chọn new và chọn thêm đường dẫn vào với:

***Variable name = CLASSPATH***

***Variable value = C:\Program Files\Microsoft JDBC Driver 6.0 for SQL Server\sqljdbc\_8.4.1.0\_enu\sqljdbc\_8.4\enu\mssql-jdbc-8.4.1.jre11.jar***

1. Chọn ok



Hình 4.7. Giao diện khi thêm CLASSPATH cho máy

Tiếp theo thực hiện truy cập đường dẫn C:\Program Files\Weka-3-8-5\RunWeka, cấp quyền full control cho file để thực hiện chỉnh sữa

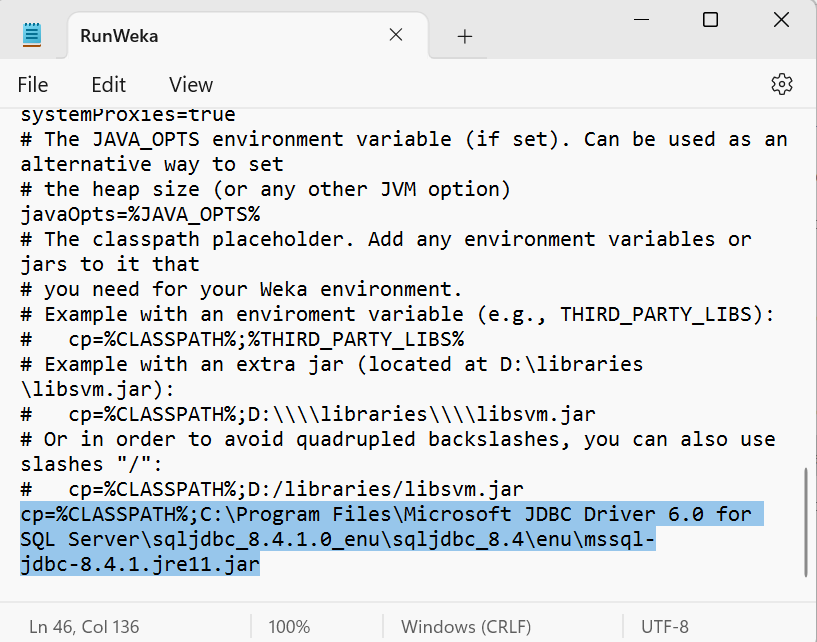
Graphical user interface, text, application

Description automatically generated

Hình 4.8. Giao diện thực hiện mỡ file ‘RunWeka” bằng Notepad

Sau đó thực thêm đường dẫn CLASSPATH như hình dưới

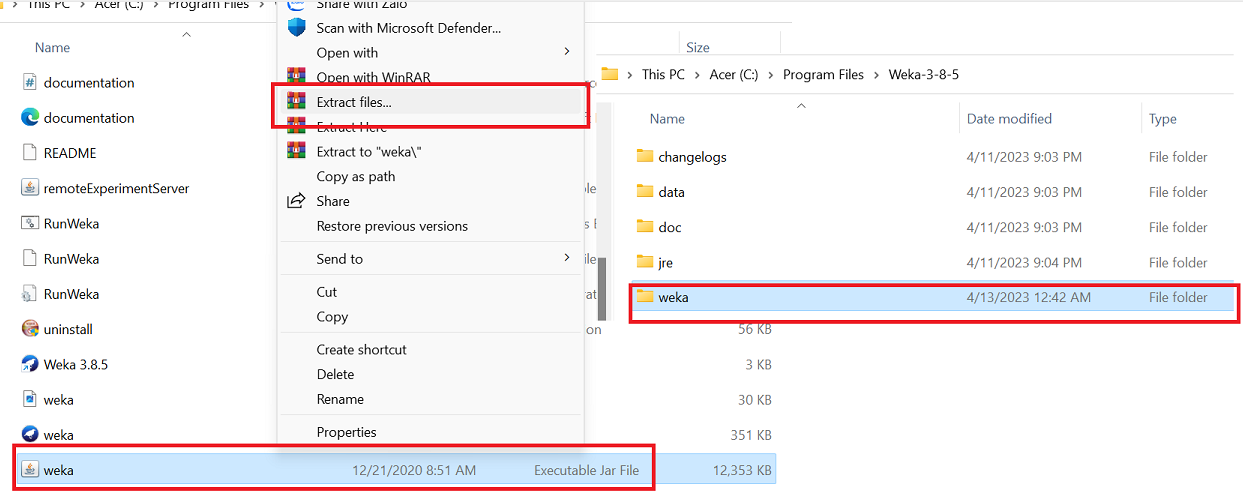
***C:\Program Files\Microsoft JDBC Driver 6.0 for SQL  
Server\sqljdbc\_8.4.1.0\_enu\sqljdbc\_8.4\enu\mssql-jdbc-8.4.1.jre11.jar***



Hình 4.9. Thêm đường dẫn CLASSPATH trong file ‘RunWeka’

Tiếp theo người dùng giải nén file ***Weka.jar***

Ở đường dẫn ‘***C:\Program Files\Weka-3-8-5\Weka.jar’*** ta thu được thư mục Weka



Hình 4.10. Giao diện giải nén file weka.jar

Sau đó truy cập đường dẫn :

***‘C:\Program Files\Weka-3-8-5\weka\weka\experiment’***

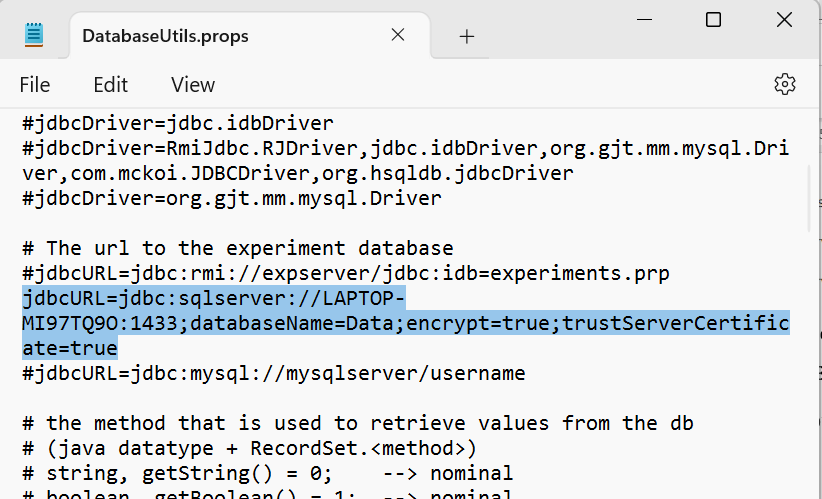
Graphical user interface, application

Description automatically generated

Hình 4.11. Đường dẫn để truy cập vào file ‘DatabaseUtils.props’

Thực hiện một số thay đổi “jdbcURL” đối với tệp DatabaseUtils.props mở bằng trình soạn thỏa văn bản và thay đổi với máy chủ là tên máy chủ của bạn, cổng kết nối là 1433 và tên cơ sở dữ liệu là cơ sở dữ liệu bạn cần lấy

***jdbcURL=jdbc:sqlserver://LAPTOP-MI97TQ9O:1433; databaseName=Data;encrypt=true;trustServerCertificate=true***



Hình 4.12. Thực hiện chỉnh sửa ‘jdbcURL’ trên file ‘DatabaseUtils.props’’

Sau đó thực hiện chỉnh sửa với ***file DatabaseUtils.props.mssqlserver2015*** với thay đổi về ***jdbcURL*** và các phần ánh xạ kiểu dữ liệu giữa weka và sql như hình dưới.

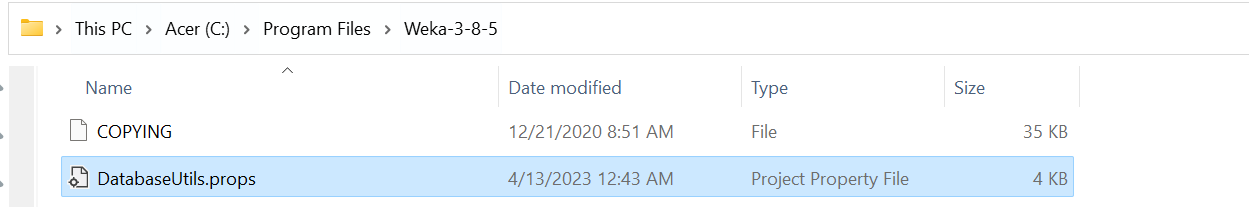
Graphical user interface, text, application, email

Description automatically generated

Hình 4.13. Thêm các thuộc tính trong file ‘DatabaseUtils.props.mssqlserver2015’

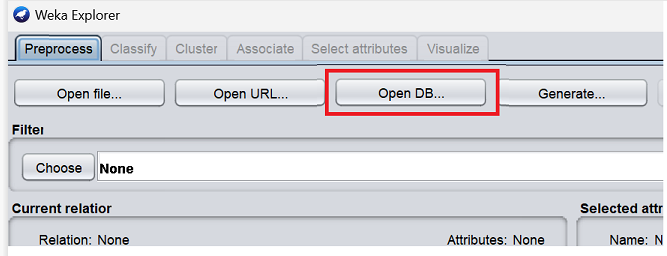
Sau khi sửa xong thực hiện lưu file sau đó thực hiện sao chép file ***DatabaseUtils.props.mssqlserver2015*** và paste vào đường dẫn:

***‘C:\Program Files\Weka-3-8-5’*** đồng thời đổi tên file thành: ***DatabaseUtils.props***



Hình 4.14. Thực hiện sửa file thành ‘DatabaseUtils.props’

Sau đó thực hiện mở ‘***weka -> explorer’***, ở phần ***‘Preprocess -> Open DB…’***



Hình 4.15. Giao diện để Open các file DataBase

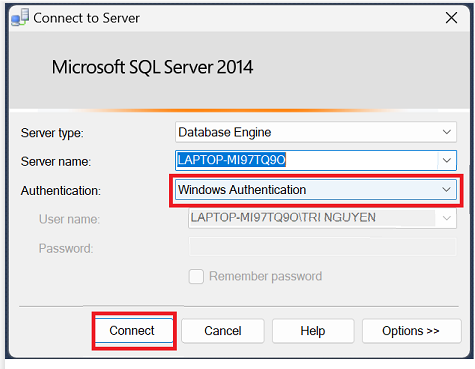
Sau đó chọn biểu tượng người dùng và nhập tài khoản và mật khẩu

Graphical user interface, text, application, Word

Description automatically generated

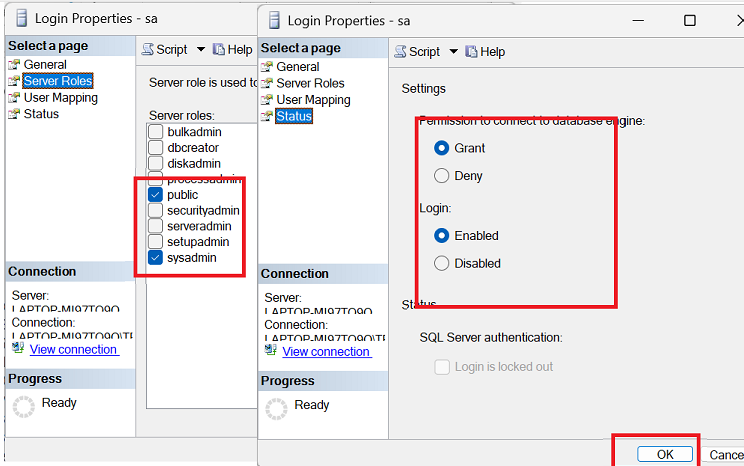
Hình 4.16. Giao diện để kết nối vào Server

Nếu người dùng ch có tài khoản sa, người dùng thực hiện truy cập SQL Server 2014 như hình:

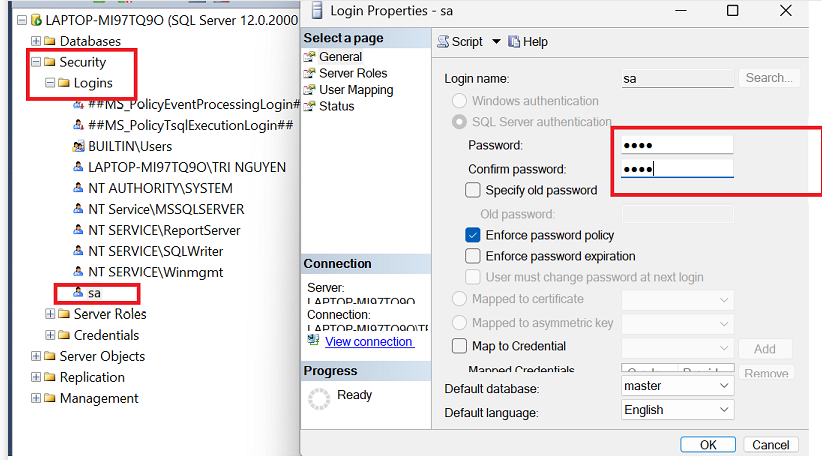


Hình 4.17. Phần đăng nhập vào SQL Server

Sau đó thực hiện truy cập ***‘Security -> Login -> sa’*** chuột trái và chọn property để thực hiện điều chỉnh như các hình dưới.



Hình 4.18. Cấp quyền cho người dùng ‘sa’ trong SQL Server



Hình 4.19. Thực hiện thay đổi cho người dùng ‘sa’ trong SQL Server

Tạo cơ sở dữ liệu

Graphical user interface, text, application

Description automatically generated

Hình 4.20. Câu lệnh Query để thực hiện thao tác tạo bảng và thêm dữ liệu

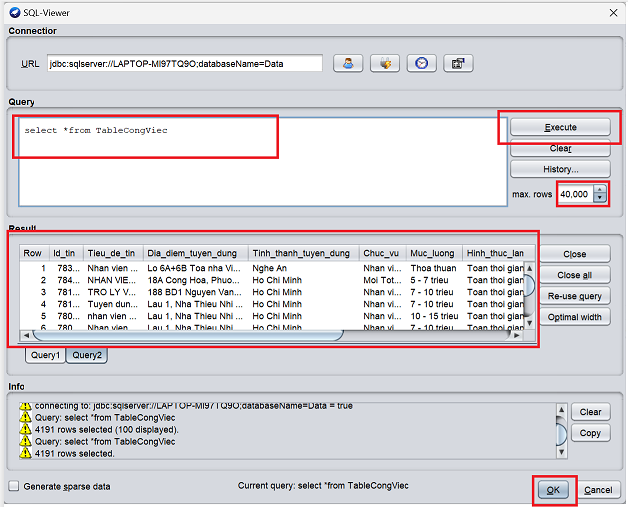
Sau khi có tài khoản sa người dùng đăng thực hiện nhập user name và pass sau đó chọn biểu tượng plugin để thực hiện kết nối giữa Weka với Sql server. Nếu kết nối thành công, thì ở phần Info hệ thống sẽ thông báo lên kết quả kết nối.

Graphical user interface, text, application, email

Description automatically generated

Hình 4.21. Giao diện khi đã kết nối thành công

Sau khi kết nối thành công người dùng thực hiện câu lệnh select để lấy dữ liệu từ cơ sỡ dữ liệu đã kết nối và chọn ok để nạp dữ liệu lên weka.



Hình 4.22. Sử dụng các lệnh để truy xuất các dữ liệu

***Kết quả:***

Graphical user interface, application

Description automatically generated

Hình 4.23. Kết quả khi mở dữ liệu từ DataBase

### Xử lý dữ liệu tổng

#### Xoá các thuộc tính không cần thiết.

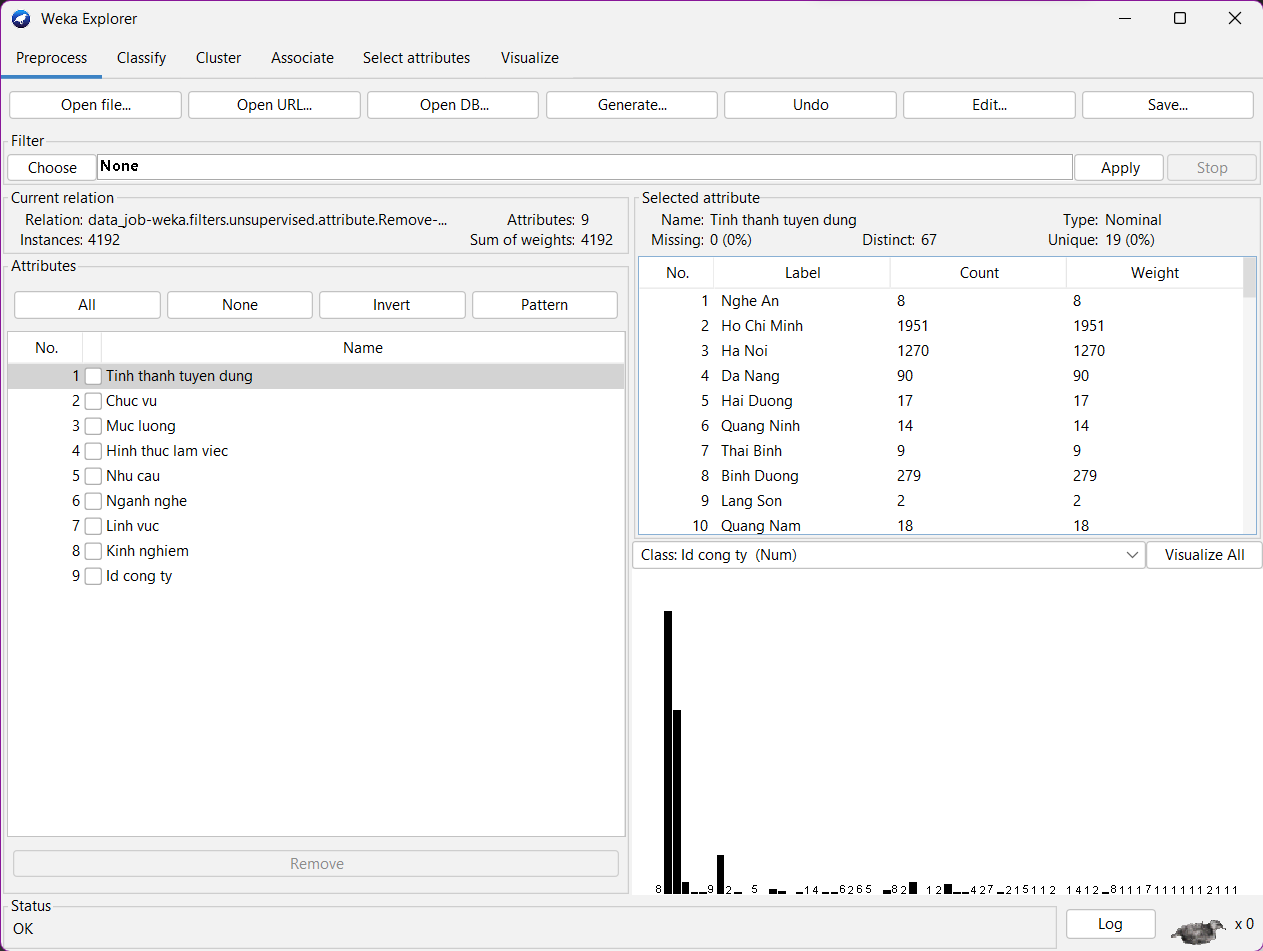
***Bước 1:*** Chọn những cột không liên quan như hình sau:



Hình 4.24. Chọn các thuộc tính cần xóa

***Bước 2:*** Nhấn ‘Remove’

Kết quả:

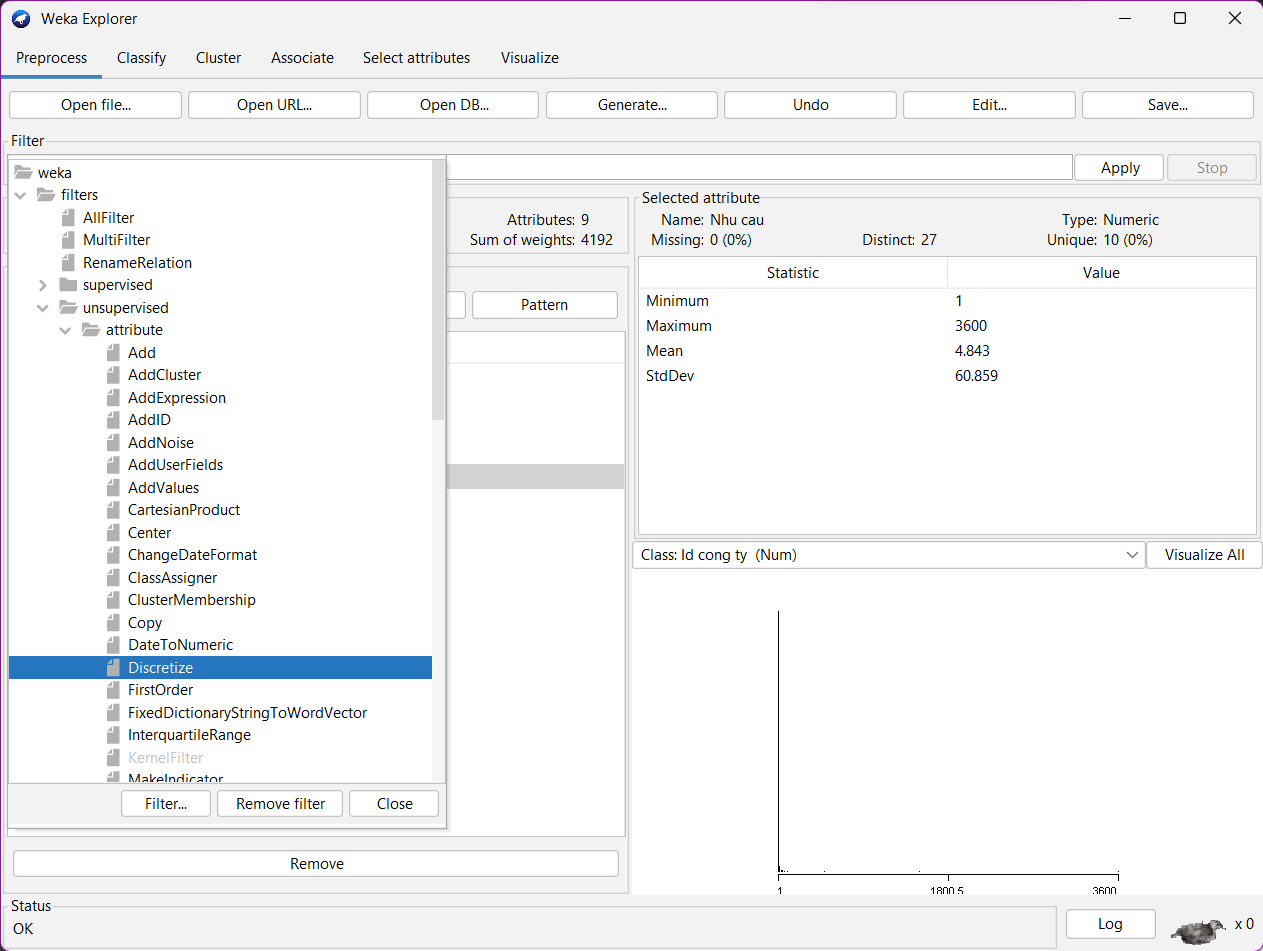


Hình 4.25. Kết quả sau khi xóa các thuộc tính

#### Chia phạm vi cho thuôc tính

***Bước 1:*** Chọn ‘Choose’ và tiếp tục nhấn chọn:

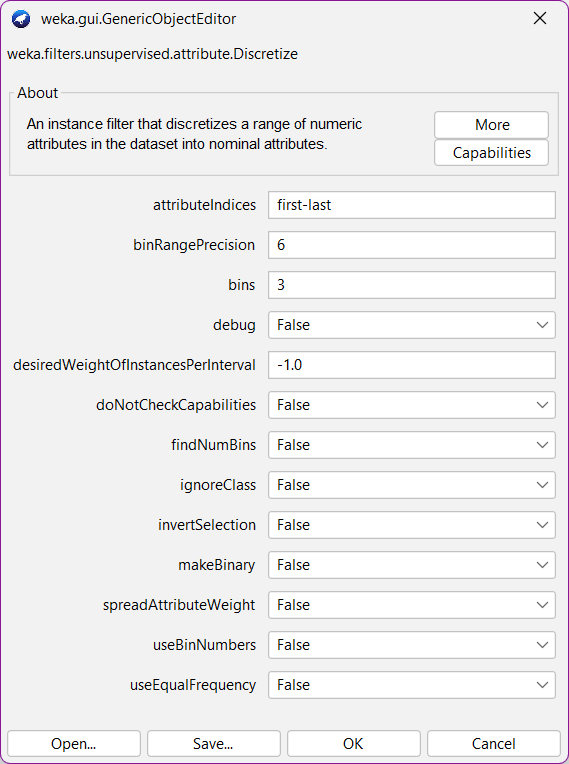
***‘weka/filters/unsupervided/attribute/discretize’***



Hình 4.26. Thư mục chọn cách thức cho việc lọc dữ liệu

***Bước 2:*** Click vào dòng Discretize trên cùng dòng với nút chọn ‘Choose’ để sửa các thông số

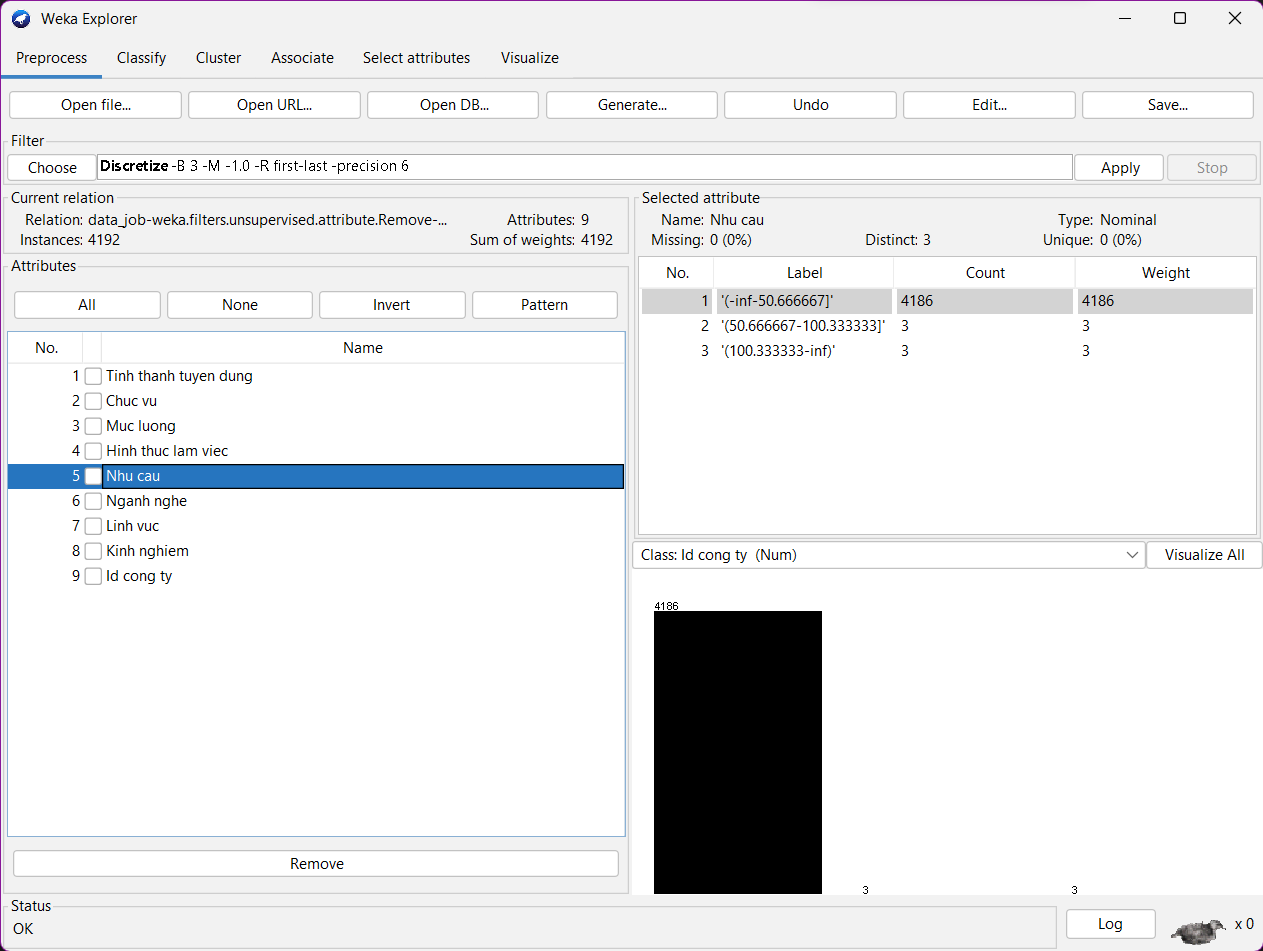
Ở đâu tập dữ liệu nhu cầu sẽ phân làm 3 vùng nên thuộc tính ‘bins’ sẽ để là 3



Hình 4.27. Điều chỉnh các thông số cho việc lọc ‘Discretize’

***Bước 3:*** Sau khi chỉnh sửa xong nhấn ok để hoàn thành và ấn ‘Apply’ trên ô ‘Filters’ để chạy lọc dữ liệu

***Kết quả:***

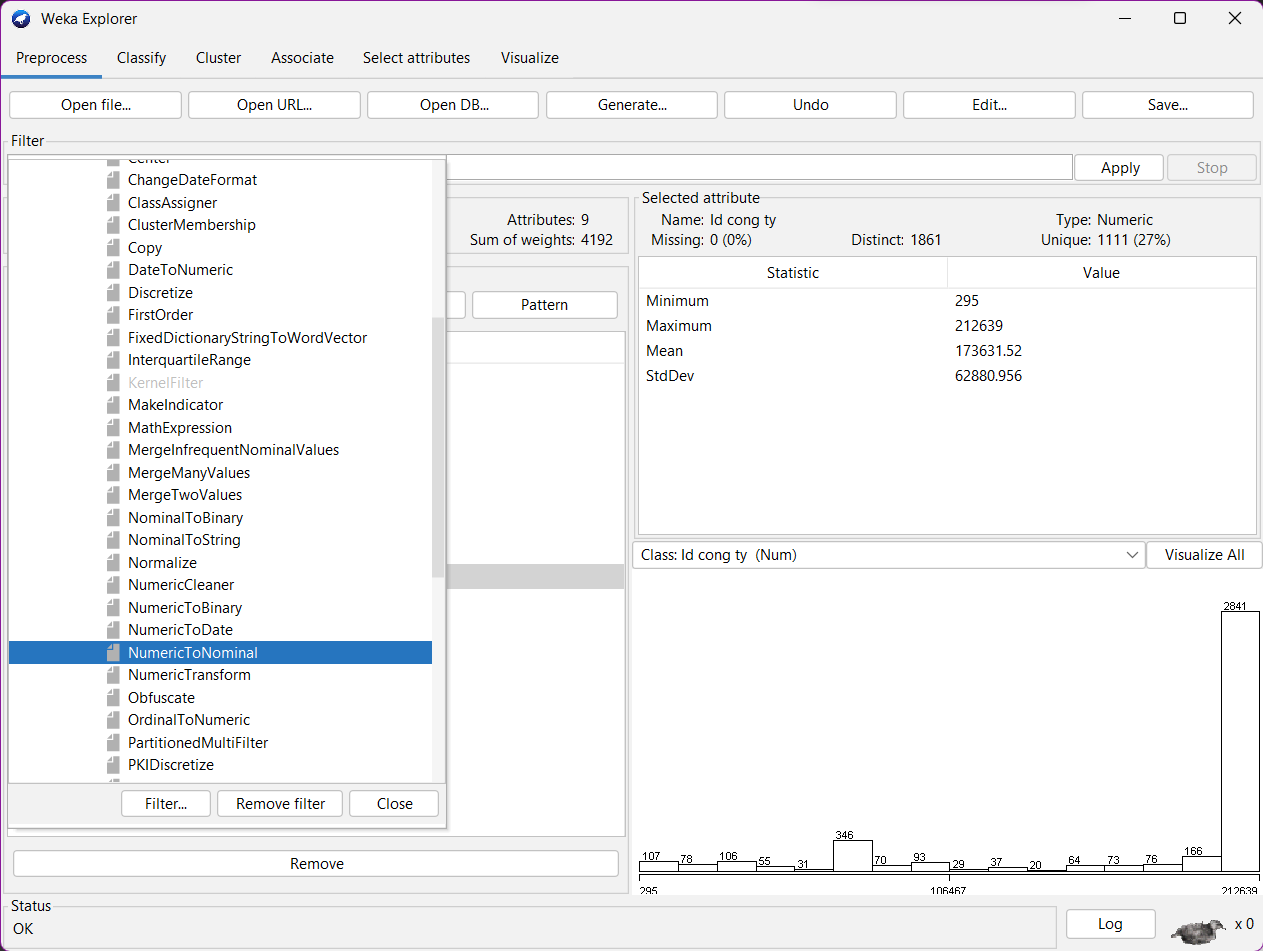


Hình 4.28. Kết quả việc lọc ‘Discretize’

#### Chuyển đổi kiểu dữ liệu numeric thành nominal

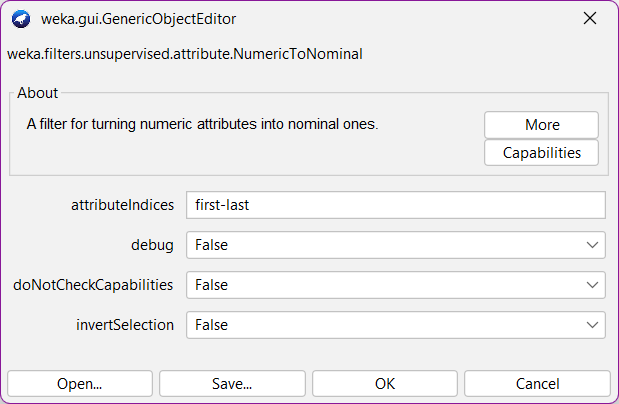
***Bước 1:*** Chọn ‘Choose’ và tiếp tục nhấn chọn:

‘***weka/filters/unsupervided/attribute/numerictonominal’***



Hình 4.29. Thư mục chọn cách thức cho việc lọc dữ liệu

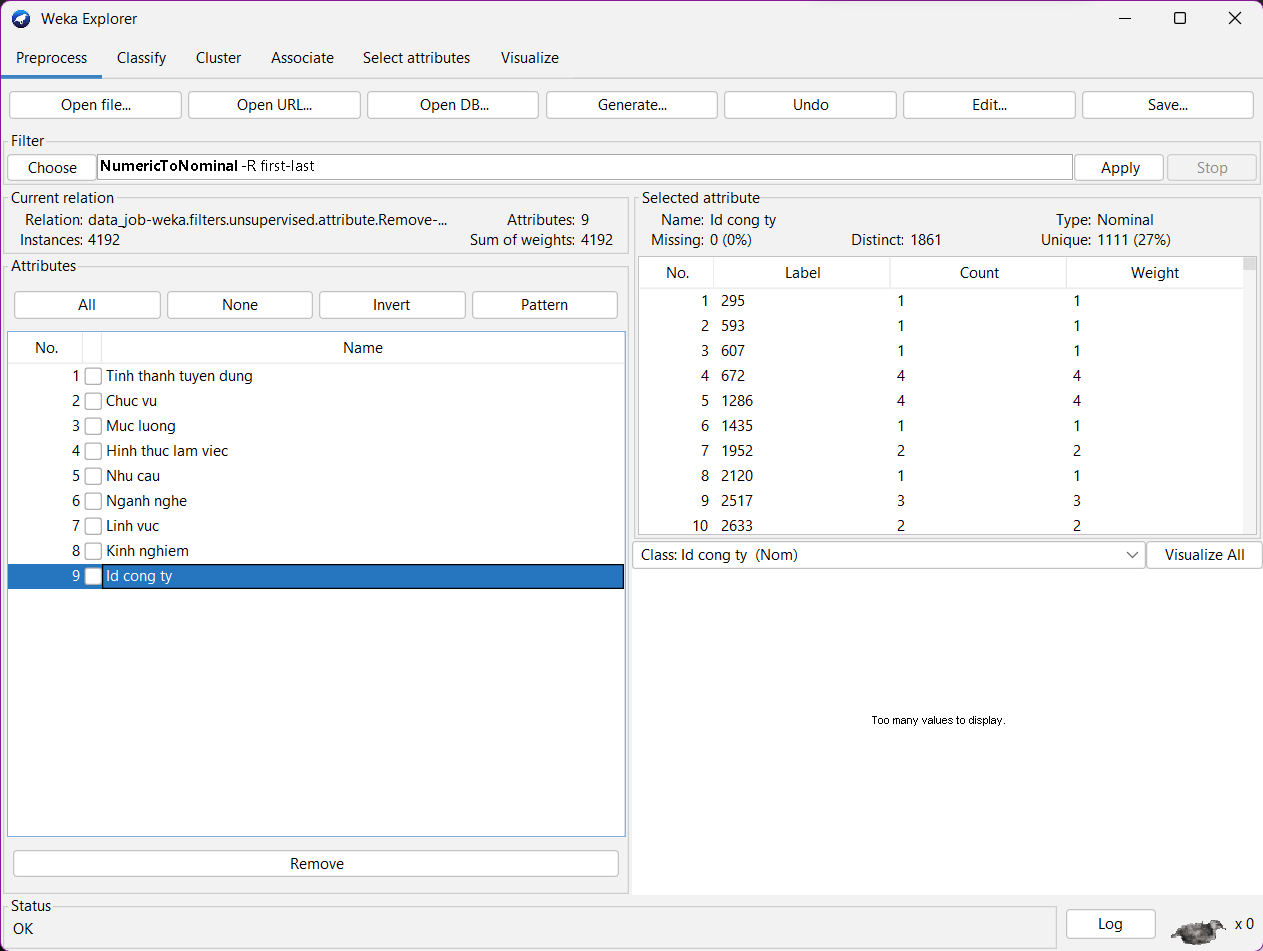
***Bước 2:*** Chọn vào dòng ***‘NumericToNominal’*** trên hộp ***‘Filters’*** để chỉnh sửa các thuộc tính



Hình 4.30. Điều chỉnh các thông số cho việc chuyển đổi ‘NumericToNominal’

***Bước 3:*** Nếu không có chỉnh sửa gì khác nhấn ***‘OK’*** sau đó ấn ***‘Apply’*** trên hộp ***‘Filters’*** để thực hiện chuyển đổi.

***Kết quả:***



Hình 4.31. Kết quả sau khi chuyển đổi kiểu dữ liệu numeric thành nominal

#### Lấy 30% của tập dữ liệu chính

***Bước 1:*** Chọn ***‘Choose -> filters-> Unsupervised -> Resample’***

Graphical user interface

Description automatically generated with low confidence

Hình 4.32. Thư mục chọn cách thức cho việc lọc dữ liệu

***Bước 2:*** Click vào ‘***Resample’*** và sau sửa ***‘sampleSizePercent = 30’***

Graphical user interface, application

Description automatically generated

Hình 4.33. Điều chỉnh các thông số cho việc phân tập dữ liệu

***Bước 3:*** Sau đó chọn apply

Graphical user interface

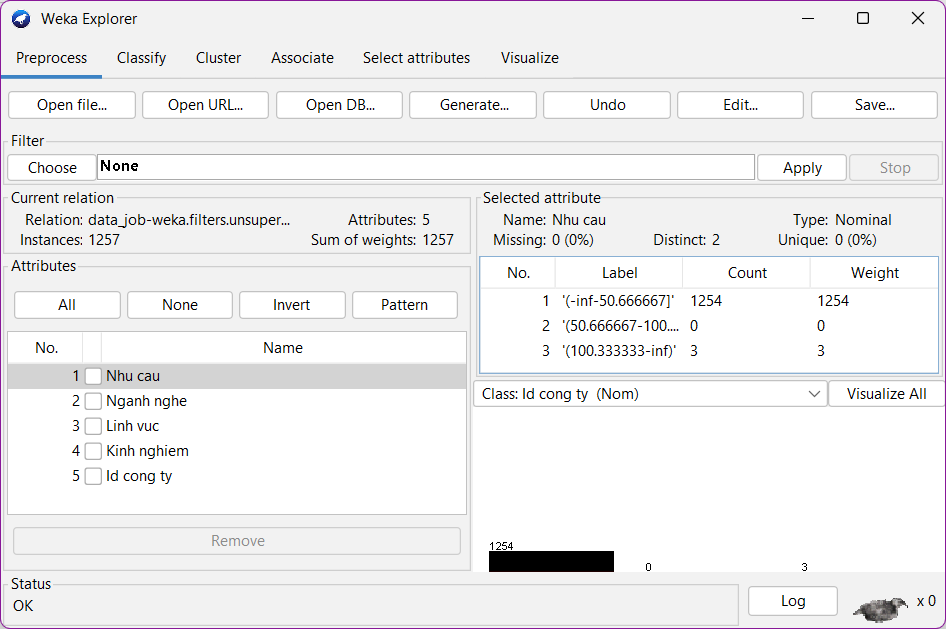
Description automatically generated

Hình 4.34. Kết quả sau khi phân tập dữ liệu

### Xử lý dữ liệu cho Apriori

Ví dụ: Xét luật kết hợp Apriori theo các yếu tố (NhuCau, NganhNghe, LinhVuc, KinhNghiem, IdCongTY)

Xử lý tương tự như phần ***‘Xóa các thuộc tính không cần thiết’*** ở mục ***‘4.3.2/a.’***



Hình 4.35. Xóa các thuộc tính không cần và giữ lại các thuộc tính để chạy Apriori

Sau đó bạn có thể ‘Associate’ để sinh ra các luật từ các thuộc tính.

### Xử lý dữ liệu cho FP-Growth

***Bước 1:*** Xóa các thuộc tính ko liên quan

Chọn các thuộc tính ko liên quan và chọn remove để xóa

Graphical user interface, application

Description automatically generated

Hình 4.36. Thực hiện chọn các thư mục cần xóa và ấn ‘Remove’ để xóa

***Bước 2:*** Ở ***‘filters -> choose -> filters -> unsupervised -> attributes -> AddCluster’***

Graphical user interface, text, application

Description automatically generated

Hình 4.37. Thực hiện thêm thuộc tính phân lớp cho tập dữ liệu

Sau đó chọn ***‘Apply’*** để tạo ra một lớp

Graphical user interface, application

Description automatically generated

Hình 4.38. Kết quả của việc thêm thuộc tính phân lớp

***Bước 3:*** Chuyển từ Kiểu Nominal sang Binary

Ở ***‘Filters -> choose -> unsupervised -> attributes -> nominaltobinary’*** và điều chỉnh thuộc tính

Ở phần ***binaryattributesnomial = true*** để sau khi thay đổi thì các thuộc tính vẫn ở dạng ***Nominal***

Graphical user interface, text, application

Description automatically generated

Hình 4.39. Chọn việc lọc ‘NominalToBinary’ để chuyển kiểu dữ liệu và điều chỉnh thông số cho việc chuyển đổi

Sau đó chọn apply và thực hiện xóa thuộc tính phân lớp cluster

Graphical user interface, application

Description automatically generated

Hình 4.40. Sau khi chuyển đổi thành công thì xóa thuộc tính phân lớp

Sau khi xóa chọn edit để xem tập dữ liệu

Graphical user interface

Description automatically generated

Hình 4.41. Xem tập dữ liệu sau khi đã thực hiện phân lớp

Để lưu tập dữ liệu, người dùng chọn save để lưu và sau đó đặt tên thư mục và chọn nơi lưu.

Graphical user interface, text

Description automatically generated

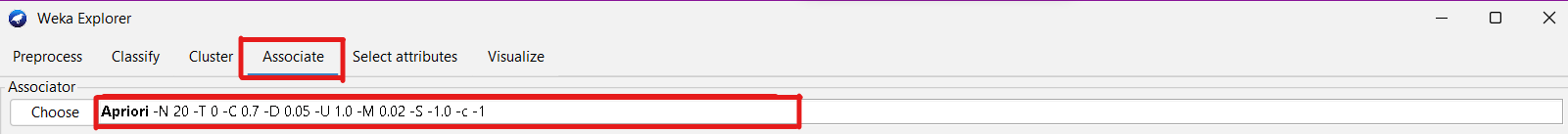
Hình 4.42. Lưu lại tệp sau khi đã xử lý để thực hiện việc chạy FP-Growth

## Lập mô hình

### Apriori

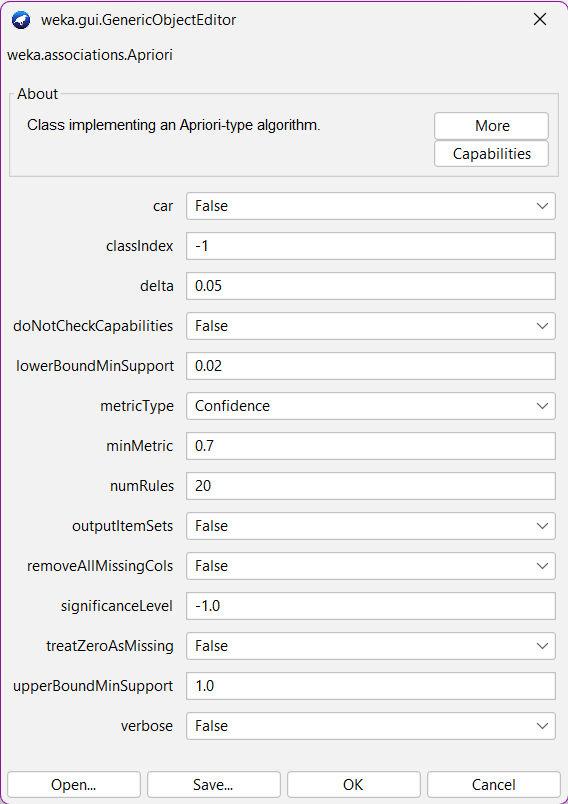
Với độ hỗ trợ tối thiểu (min\_support): = 2% và độ tin cậy tối thiểu là 70%

Để thay đổi các thông số đó, thực hiện các bước sau:



Hình 4.43. Chọn công cụ ‘Associate’ và chọn ‘Apriori’ để thực hiện chạy

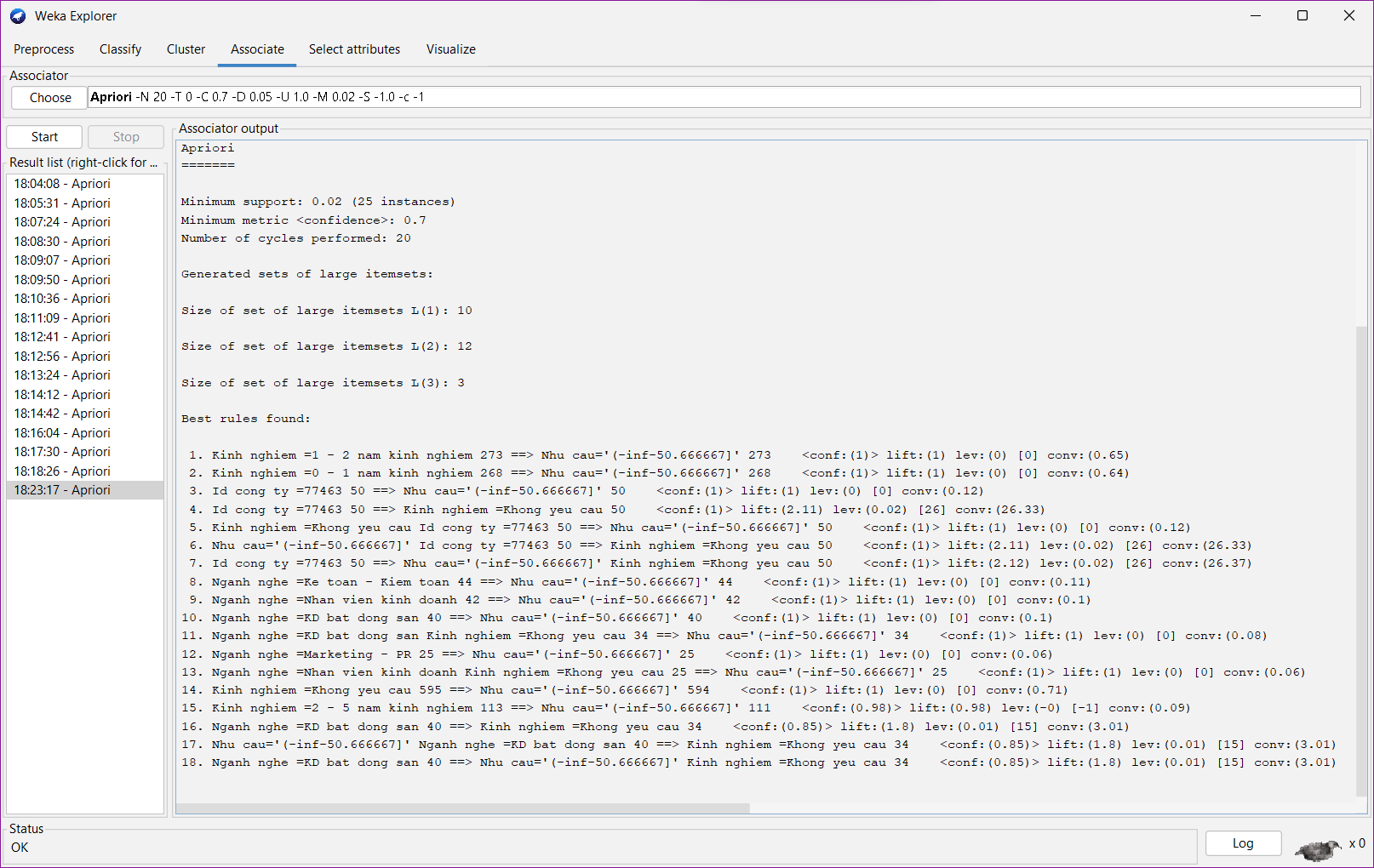
Chọn ‘Associate’ trong hộp “Associator’ click vào ‘Apriori’



Hình 4.44. Điều chỉnh các thông số cho việc chạy Apriori

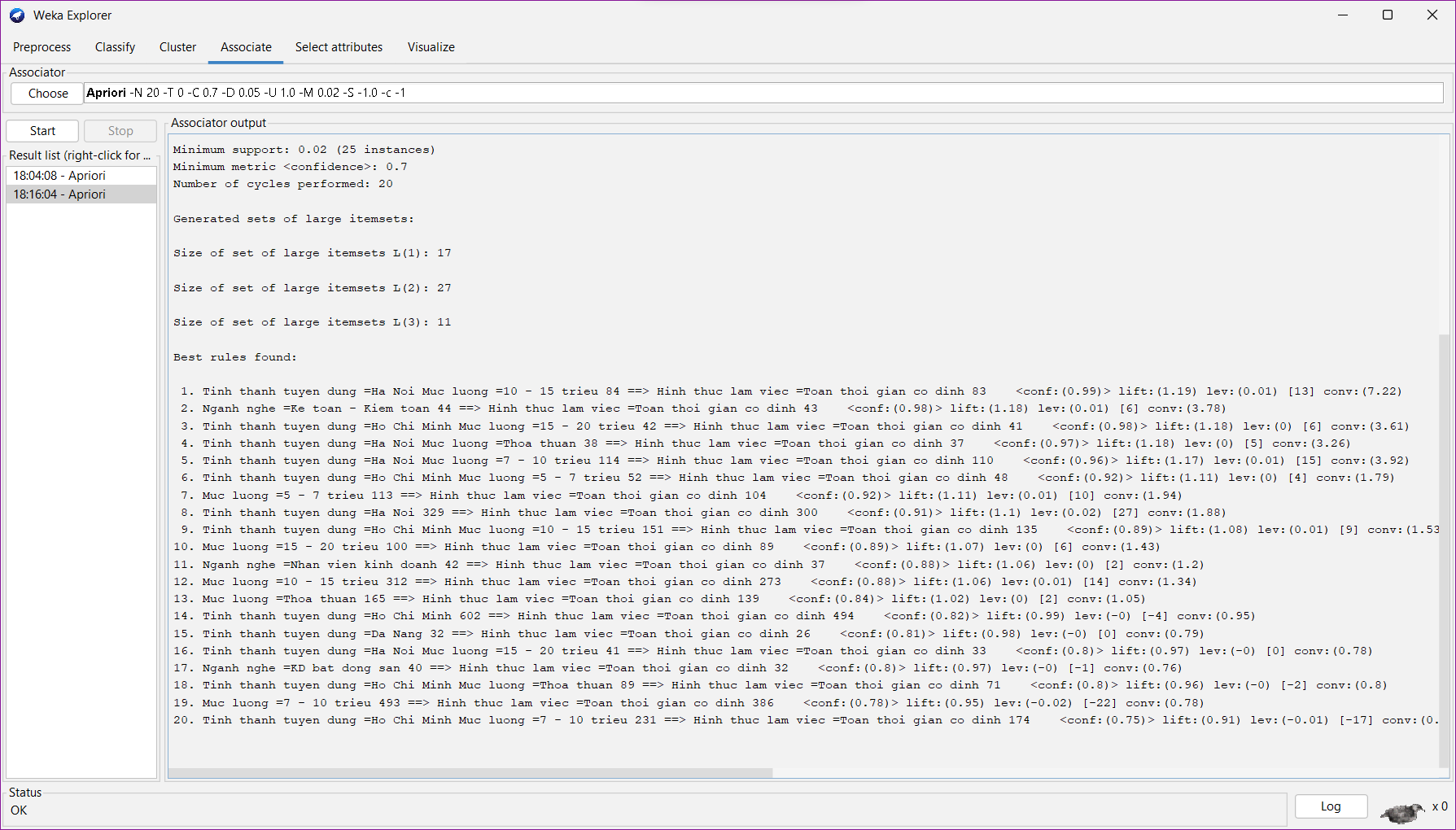
Chỉnh sửa các thông số ***‘lowerBoundMinSupport’*** là độ hỗ trợ tối thiểu và ***‘minMetric’*** là độ tin cậy tối thiểu

Xét luật kết hợp Apriori theo các yếu tố (NhuCau, NganhNghe, LinhVuc, KinhNghiem, IdCongTY)



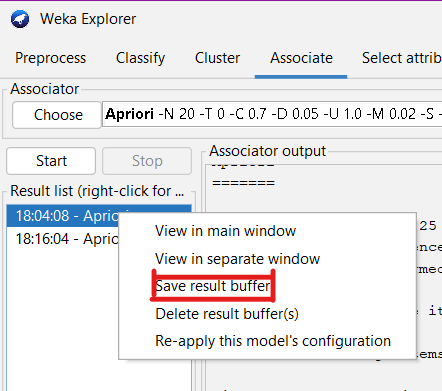
Hình 4.45. Kết quả chạy Apriori lần 1

Xét luật kết hợp Apriori theo các yếu tố (Tinh thanh tuyen dung, Muc luong, Hinh thuc lam viec, Nganh nghe, Linh vuc)

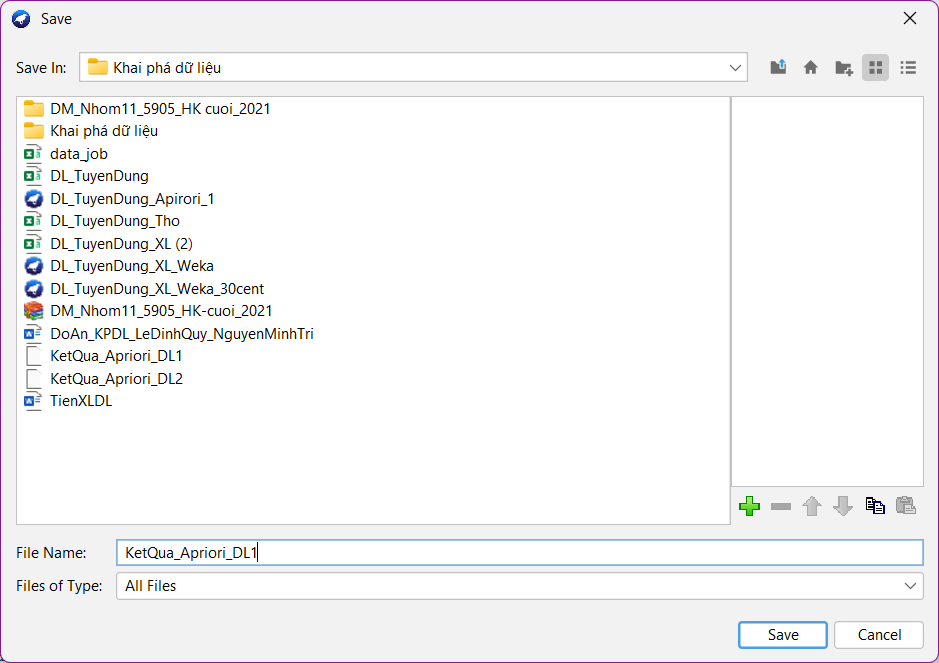


Hình 4.46. Kết quả chạy Apriori lần 2

Lưu 2 file kết quả chạy lại bằng cách click chuột phải và chọn như hình. Lưu với tên: ***‘KetQua\_Apriori\_DL1’*** và ***‘KetQua\_Apriori\_DL2’***



Hình 4.47. Giao diện thực hiện ‘Save’ kết quả chạy



Hình 4.48. Đặt tên cho kết quả chạy và lưu

### FP-Growth

Tạo luật theo Nhu cầu, ngành nghề, lĩnh vực, kinh nghiệm và ID công ty.

Chọn thực hiện thuật toán Fpgrowth (lưu ý các thuộc tính phải ở kiểu Nominal, nếu ko ở dạng này thì chúng ta cần chuyển thuộc tính về dạng này)

Chọn tất cả thuộc tính và chọn mục Associate

Graphical user interface, text, application

Description automatically generated

Hình 4.49. Check chọn tất cả các dòng dữ liệu

Ở ***‘Associator -> choose -> Association -> fpgrowth’***

Graphical user interface, text, application, email

Description automatically generated

Hình 4.50. Chọn cách kết hợp FP-Growth

Sau đó thực hiện thay đổi LowerBoundMinSupport = 0.02, Metric type = Confidence, MinMetric = 0.7 và số luật là 20

Graphical user interface, application

Description automatically generated

Hình 4.51. Điều chỉnh các thông số cho việc chạy FP-Growth

Sau đó chọn Start. Ta thu được kết quả

Graphical user interface, text, application, email

Description automatically generated

Hình 4.52. Kết quả chạy FP-Growth

Sau đó chuột trái vào mới thực hiện, chọn save, sau đó đặt tên và chọn nơi lưu file.

Graphical user interface, text, application

Description automatically generated

Hình 4.53. Thực hiện lưu kết quả chạy

Tạo luật với Fpgrowth theo các trường ‘Tinh thanh tuyen dung, Muc luong, Hinh thuc, Nganh nghe, Linh vuc’

Tương tự như các bước ở phần “Tạo luật theo Nhu cầu, ngành nghề, lĩnh vực, kinh nghiệm và ID công ty” ta có:

Graphical user interface, application, Word

Description automatically generated

Hình 4.54. Chọn các thuộc tính khác cho lần chạy thứ 2 trên FP-Growth

Sau khi chuyển từ Nominal sang Binary

Graphical user interface, application

Description automatically generated

Hình 4.55. Thực hiện chuyển kiểu dữ liệu

Thực hiện thuật toán Fpgrowth với tập dữ liệu DL2\_Fpgrowth.ar trên weka với độ hỗ trợ, luật tương tự như ở phần trên ta thu được kết quả:

Graphical user interface, text, application

Description automatically generated

Hình 4.56. Kết quả lần chạy FP-Growth thứ 2

Lưu file với tên KetQua\_Fpgrowth\_DL2

### Apriori trên R

# Cài đặt gói khai thác liên kết - hỗ trợ chạy Apiori

install.packages ("arules")

# Cài đặt gói "RWeka" hỗ trợ đọc file arff

install.packages("RWeka")

#Tải thư viện “arules”

library(arules)

#Tải thư viện "RWeka"

library("RWeka")

#Thuật toán Apriori

#Lấy dữ liệu từ các file trong máy tính và gán vào data

data <- read.arff("D://DL\_Apriori\_Lan1.arff")

(Ở đây nếu đặt file ở những nơi khác thì sẽ thay đổi đường dẫn. Ví dụ:

D:...//Tài liệu/Khai phá dữ liệu/DL\_TuyenDung\_XL\_Weka.arff)

#Xem dữ liệu

# Chúng ta có thể xem tên thuộc tính

names(data)

#xem dữ liệu có trong data

data

# Thống kê của dữ liệu

summary (data)

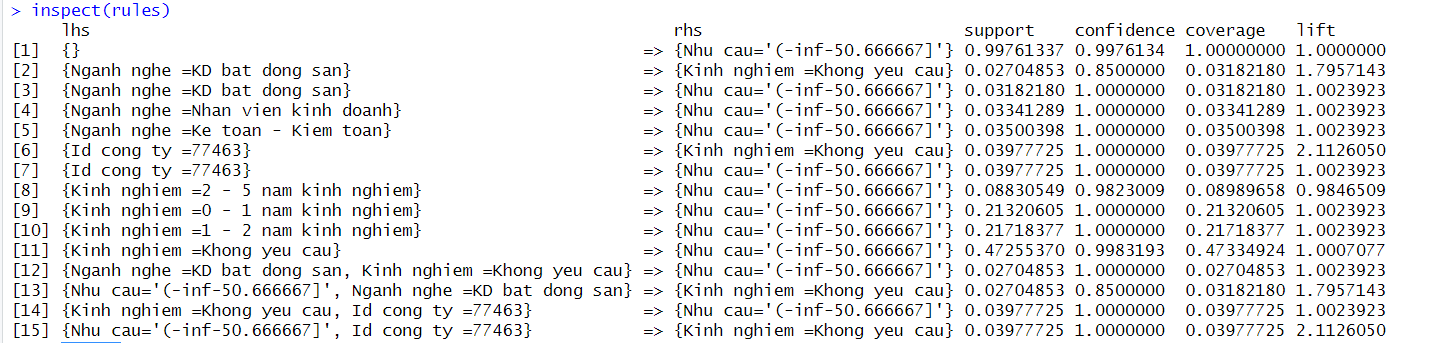
#Lệnh chạy thuật toán Apiori

rules <- apriori(data, parameter=list(target="rules"))

rules <-apriori(data, parameter=list(sup=0.02, conf=0.7, target="rules"))

#Hiện kết quả chạy Apiori

inspect(rules)



Hình 4.57. Kết quả chạy apriori với dữ liệu lần 1 trên R

### FP-Growth trên R

# Cài đặt gói khai thác liên kết - hỗ trợ chạy FP-Growth

install.packages ("rJava")

install.packages ("rCBA")

# Cài đặt gói "RWeka" hỗ trợ đọc file arff

install.packages("RWeka")

# Cài đặt gói hỗ trợ chuyển đổi dạng dữ liệu

install.packages("tidycensus")

install.packages("arules")

install.packages("arulesViz")

# Cài đặt gói phụ hỗ trợ chuyển đổi dạng dữ liệu (Có thể không cài)

install.packages("readxl")

install.packages("knitr")

install.packages("ggplot2")

install.packages("lubridate")

install.packages("plyr")

#Tải thư viện “arules”

library("rJava")

#Tải thư viện "RWeka"

library("RWeka")

#Tải thư viện "rCBA"

library("rCBA")

#Tải thư viện hỗ trợ chuyển đổi dạng dữ liệu

library("tidycensus")

library("arules")

library("arulesViz")

#Tải thư viện phụ

library("readxl")

library("knitr")

library("ggplot2")

library("lubridate")

library("plyr")

#Thuật toán FP-Growth

#Lấy dữ liệu từ các file trong máy tính và gán vào data

data <- read.arff("D://DL\_Apriori\_Lan1.arff")

#Chuyển đổi kiểu dữ liệu

train <- sapply(data,as.factor)

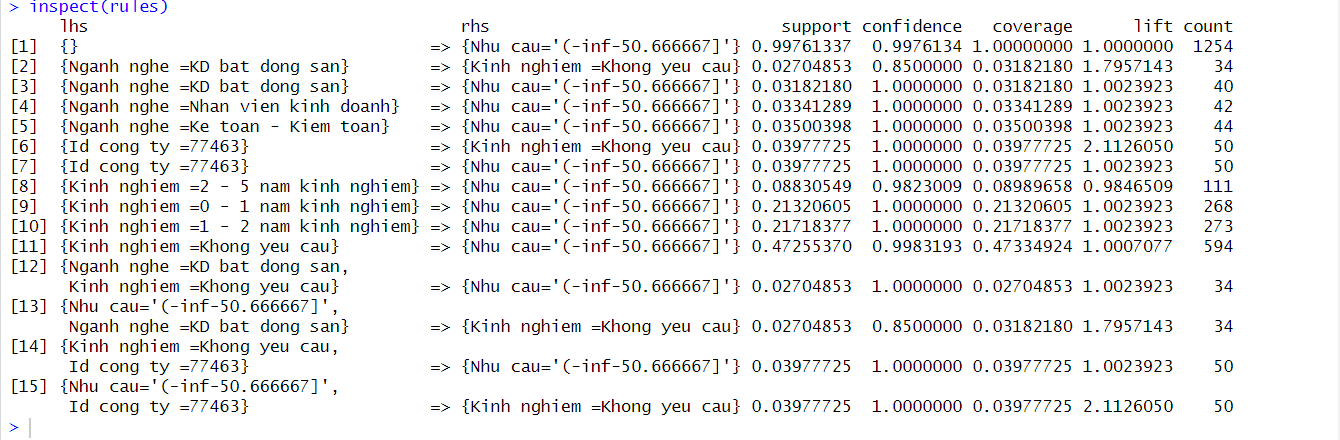
train <- data.frame(train, check.names=FALSE)

txns = as(train,"transactions")

#Chạy thuật toán FP-Growwth

rules <- rCBA::fpgrowth(txns, support=0.02, confidence=0.7, maxLength=1, consequent="Species", parallel=FALSE)

inspect(rules)



Hình 4.58. Kết quả chạy FP-Growth trên R

## Đánh giá mô hình

### Đánh giá mô hình với kết quả xử lý DL1 (KetQua\_Apriori\_DL1 or KetQua\_FBGrowth\_DL1)

Kết quả xử lý với độ hỗ trợ là 2% và Độ tin cậy tối thiểu là 70% của(Nhu cau, Nganh nghe Linh vuc , Kinh nghiem, Id cong ty ) ta thu được các luật sau:

1. [Kinh nghiem =1 - 2 nam kinh nghiem=t]: 273 ==> [Nhu cau='(-inf-50.666667]'=t]: 273 <conf:(1)> lift:(1) lev:(0) conv:(0.65)

***Giải thích:*** *Với kinh nghiệm từ 1 đến 2 năm với mật độ xuất hiện là 273, nhu cầu tuyển dụng là từ 0 tới 51 với mật độ là 273 lần. Độ tin cậy là 100 %*

2. [Kinh nghiem =0 - 1 nam kinh nghiem=t]: 268 ==> [Nhu cau='(-inf-50.666667]'=t]: 268 <conf:(1)> lift:(1) lev:(0) conv:(0.64)

***Giải thích:*** *Với kinh nghiệm từ 0 tới 1 năm với mật độ xuất hiện là 268 lần, nhu cầu tuyển dụng là từ 0 tới 51 với mật độ xuất hiện là 268 lần nên độ tin cậy là 100%*

3. [Id cong ty =77463=t]: 50 ==> [Nhu cau='(-inf-50.666667]'=t]: 50 <conf:(1)> lift:(1) lev:(0) conv:(0.12)

***Giải thích:*** *Với công ty có id là 77463 với mật độ xuất hiện là 50, nhu cầu tuyển dụng là từ 0 tới 51 với mật độ xuất hiện là 50 nên độ tin cậy là 100%*

4. [Nganh nghe =Ke toan - Kiem toan=t]: 44 ==> [Nhu cau='(-inf-50.666667]'=t]: 44 <conf:(1)> lift:(1) lev:(0) conv:(0.11)

***Giải thích:*** *Với ngành nghề là Kế toán – Kiểm toán với mật độ xuất hiện là 44, nhu cầu tuyển dụng là từ 0 tới 51 với mật độ xuất hiện là 44 nên độ tin cậy là 100%*

5. [Nganh nghe =Nhan vien kinh doanh=t]: 42 ==> [Nhu cau='(-inf-50.666667]'=t]: 42 <conf:(1)> lift:(1) lev:(0) conv:(0.1)

***Giải thích:*** *Với ngành nghề là nhân viên kinh doanh với mật độ xuất hiện là 42, nhu cầu tuyển dụng là từ 0 tới 51 với mật độ xuất hiện là 42 nên độ tin cậy là 100%*

6. [Nganh nghe =KD bat dong san=t]: 40 ==> [Nhu cau='(-inf-50.666667]'=t]: 40 <conf:(1)> lift:(1) lev:(0) conv:(0.1) )

***Giải thích:*** *Với ngành nghề là kinh doanh bất động sản với mật độ xuất hiện là 40, nhu cầu tuyển dụng là từ 0 tới 51 với mật độ xuất hiện là 40 thì độ tin cậy là 100%*

7. [Id cong ty =77463=t]: 50 ==> [Kinh nghiem =Khong yeu cau=t]: 50 <conf:(1)> lift:(2.11) lev:(0.02) conv:(26.33)

***Giải thích:*** *Với công ty có ID là 77463 có mật độ xuất hiện là 50, yêu cầu kinh nghiệm để tuyển dụng là không yêu cầu với mật độ xuát hiện là 50 nên độ tin cậy là 100%*

8. [Id cong ty =77463=t]: 50 ==> [Nhu cau='(-inf-50.666667]'=t, Kinh nghiem =Khong yeu cau=t]: 50 <conf:(1)> lift:(2.12) lev:(0.02) conv:(26.37)

***Giải thích:*** *Với công ty có ID là 77463 với mật độ xuất hiện là 50, nhu cầu tuyển dụng là từ 0 tới 51 với yêu cầu kinh nghiệm để tuyển dụng là không yêu cầu có mật độ xuất hiện là 50 nên độ tin cậy là 100%*

9. [Nhu cau='(-inf-50.666667]'=t, Id cong ty =77463=t]: 50 ==> [Kinh nghiem =Khong yeu cau=t]: 50 <conf:(1)> lift:(2.11) lev:(0.02) conv:(26.33)

***Giải thích:*** *Với nhu cầu tuyển dụng là từ 0 tới 51 thì công ty có ID là 77463 có mật độ xuất hiện là 50, yêu cầu về kinh nghiệm là không yêu cầu xuất hiện 50 lần nên độ tin cậy là 100%*

10. [Kinh nghiem =Khong yeu cau=t, Id cong ty =77463=t]: 50 ==> [Nhu cau='(-inf-50.666667]'=t]: 50 <conf:(1)> lift:(1) lev:(0) conv:(0.12)

***Giải thích:*** *Không yêu cầu về kinh nghiệm với công ty có ID là 77463 có mật độ xuất hiện là 50, nhu cầu tuyển dụng với số lượng từ 0 tới 51 có mật độ xuất hiện là 50 nên Độ tin cậy là 100%*

11. [Kinh nghiem =Khong yeu cau=t, Nganh nghe =KD bat dong san=t]: 34 ==> [Nhu cau='(-inf-50.666667]'=t]: 34 <conf:(1)> lift:(1) lev:(0) conv:(0.08)

***Giải thích:*** *Kinh nghiệm làm việc là không yêu cầu và ngành nghề là kinh doanh bất động sản có mật độ xuất hiện là 34, Nhu cầu tuyển dụng là từ 0 tới 51 có mật độ xuất hiện là 34 nên độ tin cậy là 100 %.*

12. [Kinh nghiem =Khong yeu cau=t]: 595 ==> [Nhu cau='(-inf-50.666667]'=t]: 594 <conf:(1)> lift:(1) lev:(0) conv:(0.71)

***Giải thích:*** *Không yêu cầu về kinh nghiệm có mật độ xuất hiện là 595, nhu cầu tuyển dụng có số lượng từ 0 đến 51 có mật độ xuất hiện là 594 nên độ tin cậy xấp xỉ là 100%*

13. [Kinh nghiem =2 - 5 nam kinh nghiem=t]: 113 ==> [Nhu cau='(-inf-50.666667]'=t]: 111 <conf:(0.98)> lift:(0.98) lev:(-0) conv:(0.09)

***Giải thích:*** *Kinh nghiệm từ 2 tới 5 năm kinh nghiệm có mật độ xuất hiện là 113, nhu cầu tuyển dụng từ 0 tới 51 có mật độ xuất hiện là 111 nên độ tin cậy là 98%*

14. [Nganh nghe =KD bat dong san=t]: 40 ==> [Kinh nghiem =Khong yeu cau=t]: 34 <conf:(0.85)> lift:(1.8) lev:(0.01) conv:(3.01)

***Giải thích:*** *Ngành nghề kinh doanh bất động sản có mật độ xuất hiện là 40, nhu cầu tuyển dụng với kinh nghiệm là không yêu cầu có mật độ xuất hiện là 34 nên Độ tin cậy là 85%*

15. [Nganh nghe =KD bat dong san=t]: 40 ==> [Nhu cau='(-inf-50.666667]'=t, Kinh nghiem =Khong yeu cau=t]: 34 <conf:(0.85)> lift:(1.8) lev:(0.01) conv:(3.01)

***Giải thích:*** *Ngành nghề là kinh doanh bất động sản có mật độ xuất hiện là 40, nhu cầu tuyển dụng với số lượng từ 0 đến 51 và không yêu cầu về kinh nghiệm có mật độ xuất hiện là 34 nên Độ tin cậy là 85%*

16. [Nhu cau='(-inf-50.666667]'=t, Nganh nghe =KD bat dong san=t]: 40 ==> [Kinh nghiem =Khong yeu cau=t]: 34 <conf:(0.85)> lift:(1.8) lev:(0.01) conv:(3.01)

***Giải thích:*** *Nhu cầu tuyển dụng từ 0 đến 50 và ngành nghề là kinh doanh bất động sản có mật độ xuất hiện là 40, Kinh nghiệm là không yêu cầu về kinh nghiệm có mật độ xuất hiện là 34 thì Độ tin cậy là 85%.*

***Kết luận 1:*** *qua đó ta có thể thấy giữa ngành nghề, kinh nghiệm, ID công ty có mối quan hệ qua lại lẫn nhau và với nhu cầu và hầu hết nhu cầu nằm trong khoảng 1 đến 50.*

*Ta có thể đánh giá các yếu tố được rằng thứ tự ưu tiên giảm dần và độ tin cậy tốt :*

* Ngành nghề: Ngành kiểm toán kế toán, nhân viên kinh doanh, kinh doanh bất động sản.
* Kinh nghiệm: Không yêu cầu, 1 đến 2 năm, 0 đến 1 năm, 2 đến 5 năm
* ID công ty: 77463
* Ngành nghề, kinh nghiệm: (Kinh doanh bất động sản, không yêu cầu kinh nghiệm),
* ID công ty, kinh nghiệm: (77463, không yêu cầu)

### Đánh giá mô hình với kết quả xử lý DL2 (KetQua\_Apriori\_DL2 or KetQua\_FBGrowth\_DL2)

Kết quả xử lý với độ hỗ trợ là 2% và Độ tin cậy tối thiểu là 70% của( Tinh thanh tuyen dung, Muc luong, Hinh thuc lam viec, Nganh nghe, Linh vuc) ta thu được các luật sau:

1. [Tinh thanh tuyen dung =Ha Noi=t, Muc luong =10 - 15 trieu=t]: 84 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 83 <conf:(0.99)> lift:(1.19) lev:(0.01) conv:(7.22)

***Giải thích****: Tỉnh thành tuyển dụng là Hà Nội và mức lương giao động từ 10 đến 15 triệu có mật độ xuất hiện là 84, hình thức làm việc là toàn thời gian có mật độ xuất hiện là 83 nên Độ tin cậy 99%.*

2. [Nganh nghe =Ke toan - Kiem toan=t]: 44 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 43 <conf:(0.98)> lift:(1.18) lev:(0.01) conv:(3.78)

***Giải thích****: Ngành nghề là Kế toán – Kiểm toán có mật độ xuất hiện là 44, thì hình thức làm việc là toàn thời gian cố định có mật độ xuất hiện là 43 nên Độ tin cậy là 98%.*

3. [Tinh thanh tuyen dung =Ho Chi Minh=t, Muc luong =15 - 20 trieu=t]: 42 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 41 <conf:(0.98)> lift:(1.18) lev:(0) conv:(3.61)

***Giải thích:*** *Tỉnh thành tuyển dụng là Hồ Chí Minh và mức lương là từ 15 tới 20 triệu có mật độ xuất hiện là 42, Hình thức làm việc là toàn thời gian cố định với mật độ xuất hiện là 41 nên độ tị cậy là 98%.*

4. [Tinh thanh tuyen dung =Ha Noi=t, Muc luong =Thoa thuan=t]: 38 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 37 <conf:(0.97)> lift:(1.18) lev:(0) conv:(3.26)

***Giải thích****: Tỉnh thành tuyển dụng là Hà Nội và mức lương là thỏa thuận có mật độ xuất hiện là 38, hình thức làm việc là toàn thời gian cố định với mật độ xuất hiện là 37 nên Độ tin cậy 97%.*

5. [Muc luong =7 - 10 trieu=t, Tinh thanh tuyen dung =Ha Noi=t]: 114 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 110 <conf:(0.96)> lift:(1.17) lev:(0.01) conv:(3.92)

***Giải thích****: Với mức lương từ 7 tới 10 triệu và tỉnh thành tuyển dụng là Hà Nội có mật độ xuất hiện là 114 , hình thức làm việc là toàn thời gian cố định có mật độ xuất hiện là 110 nên Độ tin cậy là 96%.*

6. [Tinh thanh tuyen dung =Ho Chi Minh=t, Muc luong =5 - 7 trieu=t]: 52 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 48 <conf:(0.92)> lift:(1.11) lev:(0) conv:(1.79)

***Giải thích****: Với tỉnh thành tuyển dụng là Hồ Chí Ming, mức lương từ 5 đến 7 triệu có mật độ xuất hiện là 52, hình thức làm việc toàn thời gian cố định với mật độ xuất hiện là 48 nên độ tin cậy là 92%.*

7. [Muc luong =5 - 7 trieu=t]: 113 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 104 <conf:(0.92)> lift:(1.11) lev:(0.01) conv:(1.94)

***Giải thích****: Với mức lương từ 5 đến 7 triệu có mật độ xuất hiện là 113, Hình thức làm việc là toàn thời gian cố định có mật độ là 104 nên độ tin cậy là 92%.*

8. [Tinh thanh tuyen dung =Ha Noi=t]: 329 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 300 <conf:(0.91)> lift:(1.1) lev:(0.02) conv:(1.88)

***Giải thích****: Vớitỉnh thành tuyển dụng là Hà Nội có mật độ xuất hiện là 329, hình thức làm việc toàn thời gian với mật độ xuất hiện là 300 nên độ tin cậy tối là 91%.*

9. [Tinh thanh tuyen dung =Ho Chi Minh=t, Muc luong =10 - 15 trieu=t]: 151 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 135 <conf:(0.89)> lift:(1.08) lev:(0.01) conv:(1.53)

***Giải thích****: Với tỉnh thành tuyển dụng là Hồ Chí Minh mức lương từ 10 tới 15 triệu có mật độ xuất hiện là 151, hìn thức làm việc toàn thời gian cố định với mật độ xuất hiện là 135 nên độ tin cậy là 89 %.*

10. [Muc luong =15 - 20 trieu=t]: 100 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 89 <conf:(0.89)> lift:(1.07) lev:(0) conv:(1.43)

***Giải thích****: Với mức lương từ 10 tới 20 triệu có mật độ xuất hiện là 100, hình thức làm việc toàn thời gian cố định với mật độ xuất hiện là 89 nên độ tin cậy là 89%.*

11. [Nganh nghe =Nhan vien kinh doanh=t]: 42 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 37 <conf:(0.88)> lift:(1.06) lev:(0) conv:(1.2)

***Giải thích****: Với ngành nghề là nhân viên kinh doanh có mật độ xuất hiện là 42, hình thức làm việc là toàn thời gian cố định với mật độ xuất hiện là 37 nên độ tin cậy là 88%.*

12. [Muc luong =10 - 15 trieu=t]: 312 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 273 <conf:(0.88)> lift:(1.06) lev:(0.01) conv:(1.34)

***Giải thích****: Với mức lương từ 10 đến 15 triệu có mật độ xuất hiện là 312, hình thức làm việc là toàn thời gian cố định có mật độ xuất hiện là 273 nên độ tin cậy là 88%.*

13. [Muc luong =Thoa thuan=t]: 165 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 139 <conf:(0.84)> lift:(1.02) lev:(0) conv:(1.05)

***Giải thích****: Với mức lương thỏa thuận có mật độ xuất hiện là 165, hình thức làm việc toàn thời gian cố định có mật độ xuất hiện là 139 thì độ tin cậy là 84%.*

14. [Tinh thanh tuyen dung =Ho Chi Minh=t]: 602 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 494 <conf:(0.82)> lift:(0.99) lev:(-0) conv:(0.95)

***Giải thích****: Với tỉnh thành tuyển dụng là Hồ Chí Minh có mật độ xuất hiện là 602, hình thức làm việc toàn thời gian cố định có mật độ xuất hiện là 494 nên có độ tin cậy là 82%*

15. [Tinh thanh tuyen dung =Da Nang=t]: 32 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 26 <conf:(0.81)> lift:(0.98) lev:(-0) conv:(0.79)

***Giải thích****: Với tỉnh thành tuyển dụng là Đà Nẵng có mật độ xuất hiện là 32, hình thức làm việc toàn thời cố định có mật độ xuất hiện là 26 nên độ tin cậy là 81%.*

16. [Tinh thanh tuyen dung =Ha Noi=t, Muc luong =15 - 20 trieu=t]: 41 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 33 <conf:(0.8)> lift:(0.97) lev:(-0) conv:(0.78)

***Giải thích****: Với tỉnh thành tuyển dụng là Hà Nội và mức lương từ 15 tới 20 triệu có mật độ xuất hiện là 41, hình thức làm việc là toàn thời gian có mật độ xuất hiện là 33 nên độ tin cậy là 80%.*

17. [Nganh nghe =KD bat dong san=t]: 40 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 32 <conf:(0.8)> lift:(0.97) lev:(-0) conv:(0.76)

***Giải thích****: Với ngành nghề là KD bất độn sản có mật độ xuất hiện là 40, hình thức làm việc là toàn thời gian cố định có mật độ xuất hiện là 32 nên độ tin cậy là 80%.*

18. [Tinh thanh tuyen dung =Ho Chi Minh=t, Muc luong =Thoa thuan=t]: 89 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 71 <conf:(0.8)> lift:(0.96) lev:(-0) conv:(0.8)

***Giải thích****: Với tỉnh thành tuyển dụng là Hồ Chí Minh và mức lương là thỏa thuận có mật độ xuất hiện là 89, Hình thức làm việc toàn thời gian cố định với mật độ xuất hiện là 71 nên độ tin cậy là 80%.*

19. [Muc luong =7 - 10 trieu=t]: 493 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 386 <conf:(0.78)> lift:(0.95) lev:(-0.02) conv:(0.78).

***Giải thích****: Với mức lương từ 7 tới 10 triệu có mật độ xuất hiện là 493, hình thức làm việc là toàn thời gian cố định có mật độ là 386 nên độ tin cậy là 78%*

20. [Tinh thanh tuyen dung =Ho Chi Minh=t, Muc luong =7 - 10 trieu=t]: 231 ==> [Hinh thuc lam viec =Toan thoi gian co dinh=t]: 174 <conf:(0.75)> lift:(0.91) lev:(-0.01) conv:(0.68)

***Giải thích****: Với tỉnh thành tuyển dụng là Hồ Chí Minh, mức lương là 7 tới 10 triệu có mật độ xuất hiện là 231, hình thức làm việc là toàn thời gian cố định có mật độ xuất hiện alf 174 nên độ tin cậy là 75%.*

**Kết luận 2:** Qua đó ta có thể thấy các yếu tố như tình thành, mức lương, hình thức làm việc nó có mối quan hệ qua lại lẫn nhau.

*Ta có thể đánh giá các yếu tố được rằng thứ tự ưu tiên giảm dần và độ tin cậy tốt :*

* Tỉnh thành, mức lương và hình thức làm việc: ( Hà Nội, 10 – 15 triệu, Toàn thời gian cố định), (Hồ Chí Minh, 15 – 20 triệu, Toàn thời gian cố định), (Hà Nội, Thỏa thuận, Toàn thời gian cố định),(Hà Nội, 7 – 10 triệu, toàn thời gian cố định),(Hồ Chí Minh, 5 – 7 triệu, toàn thời gian cố định),(Hà Nội, 15 -20 triệu, toàn thời gian cố định),(Hồ Chí Minh, Thỏa thuận, Toàn thời gian cố định),(Hồ Chí Minh, 7-10 triệu, Toàn thời gian cố định).
* Ngành nghề, hình thức làm việc: (Kiểm toán kế toán, Toàn thời gian cố định), (Kinh doanh bất động sản , toàn thời gian cố định),(Nhân viên kinh doanh, toàn thời gian cố định).

## Triển khai mô hình

Từ kết quả kết luận 1 và kết luận 2, ta có thể thấy được hiện nay ngành nghề kế toán kiểm toán, nhân viên kinh doanh và kinh doanh bất động sản đang là xu hướng. Vậy người làm trong lĩnh vực này có cơ hội ứng tuyển các vị trí công việc cao và để ứng tuyển hiệu quả thì người ứng tuyển nên xét đến các yếu tố chính bao gồm: Kinh nghiệm, tỉnh thành, mức lương, hình thức làm việc.

Cụ thể người ứng tuyển và đang tìm những công việc này nên quan tâm đến các yếu tố bên dưới và giảm dần theo độ ưu tiên:

Ứng tuyển tại hai thành phố lớn: Hà Nội, Thành phố Hồ Chí Minh

Kinh nghiệm làm việc: Không yêu cầu kinh nghiệm (Chủ yếu là các bạn sinh viên mới tốt nghiệp hoặc đang đi học), Kinh nghiệm từ 1 đến 2 năm, 0 – 1 năm, 2 – 5 năm.

Hình thức làm việc: ưu tiên làm việc toàn thời gian cố định

Mức lương phổ biến: 10 -15 triệu, Thỏa Thuận, 5 – 7 Triệu, 7 – 10 triệu, 15 – 20 triệu.

Cụ thể Mức lương theo tỉnh thành: ( Hà Nội, 10 – 15 triệu), (Hồ Chí Minh, 15 – 20 triệu), (Hà Nội, Thỏa thuận),(Hà Nội, 7 – 10 triệu),(Hồ Chí Minh, 5 – 7 triệu ),(Hà Nội, 15 -20 triệu),(Hồ Chí Minh, Thỏa thuận),(Hồ Chí Minh, 7-10 triệu).

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Kết luận

### Những kết quả đạt được

* Nắm được tầm quan trọng của việc áp dụng công nghệ thông tin vào việc giải quyết các vấn đề về tìm việc và ứng tuển.
* Hiểu được công dụng của công cụ Weka và R trong việc khai phá dữ liệu
* Nắm rõ kiến thức về các thuật toán Apriori và FP-Growth và cách thức chạy thuật toán
* Hiểu thêm về một số chức năng khác của Weka.
* Giải đáp được bài tìm việc, giúp đánh giá được mức độ liên quan đến từng phương pháp từ đó có thể lựa chọn được phương hiệu quả nhất.

### Hạn chế

Vì thời gian còn hạn chế, trong quá trình làm đồ án chúng em sẽ không tránh khỏi những sai sót. Một trong những hạn chế mà đồ án của chúng em còn gặp phải như sau:

* + - * Nhiều công cụ trong Weka mà nhóm khai thác còn bị hạn chế trong quá trình làm đồ án.
      * Khi chạy Code trên R mặc dù chạy được nhưng vẫn phát sinh lỗi và có thể đối với một số máy khác sẽ không thể chạy được.
      * Đồ án chỉ có tính chất đúng đắn trong một số trường hợp cụ thể và trong khoảng thời gian cụ thể. Vì thế có nhiều sai sót về mặt dữ liệu dẫn đến việc thống kê khó tránh khỏi các lỗi.

## Hướng phát triển

### Hướng khắc phục những hạn chế

* Chuẩn hóa và xử lý triệt để hơn bộ dữ liệu
* Tìm hiểu kỹ hơn về R: các gói hỗ trợ và các lệnh để thực hiện
* Tìm hiểu rõ hơn về các công cụ trong Weka để thực hiện dễ dang trong việc lọc dữ liệu

### Hướng mở rộng của đề tài - Experimenter trong weka

Không giống như Weka Explorer dùng để lọc dữ liệu và thử các thuật toán khác nhau, Experimenter dùng để thiết kế và chạy thử nghiệm. Các kết quả thử nghiệm mà nó tạo ra rất mạnh mẽ và đủ tốt để được xuất bản.

Weka Experimenter cho phép bạn thiết kế các thử nghiệm chạy thuật toán trên bộ dữ liệu của riêng mình, chạy thử nghiệm và phân tích kết quả. Đó là một công cụ mạnh mẽ.

***Chuẩn bị dữ liệu:***

Lấy 251 dòng dữ liệu từ tập dữ liệu ở phần chính ở phần trên và thêm thuộc tính phân lớp với numClusster là 4

Graphical user interface, application

Description automatically generated

Hình 5.1. Điều chỉnh thông số và thực hiện phân lớp cho dữ liệu

Sau đó lưu lại file với tên là DL\_TuyenDung\_ThuNghiem.arff

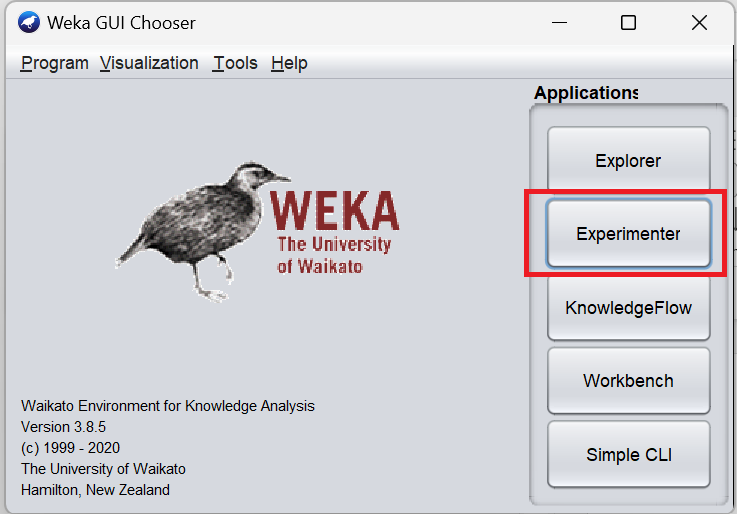
Graphical user interface, text

Description automatically generated

Hình 5.2. Thực hiện lưu file vừa lọc

Thực hiện thử nghiệm

***Bước 1:*** Mở Weka và chọn Experimenter



Hình 5.3. Giao diện chính của weka và chọn ‘Experrimenter’

Sau khi giao diện Weka Experimenter hiện lên, ở mục Setup chọn new để tạo mới thử nghiệm

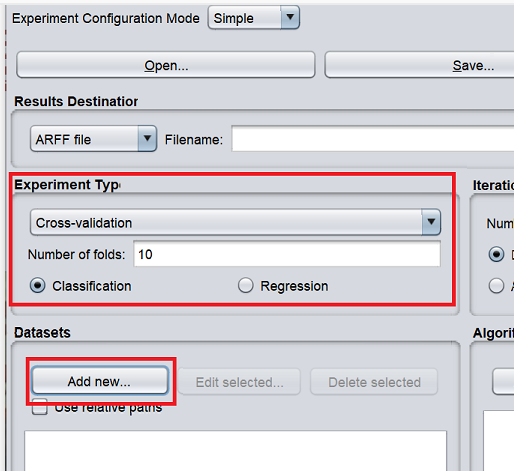
A screenshot of a computer

Description automatically generated

Hình 5.4. Thực hiện thêm mới một thử nghiệm

***Bước 2:*** Chọn tập dữ liệu

Ở phần Experiment Type, sử dụng Cross-Vadilation với 10 folds và sau đó ở **Datasets** chọn **Add New** để chọn tập dữ liệu



Hình 5.5. Điều chỉnh thông số và thêm vào một tệp dữ liệu

Chọn tập dữ liệu cần lấy

Graphical user interface, text, application

Description automatically generated

Hình 5.6. Kết quả khi thêm tập dữ liệu

***Bước 3:*** Chọn thuật toán để chạy bộ dữ liệu

Thực hiện chọn 3 thuật toán để chạy bộ dữ liệu

Bằng cách: Ở ***‘Algorithms -> Add news’***

Graphical user interface

Description automatically generated

Hình 5.7. Thêm mới các thuật toán để chạy bộ dữ liệu

**Thêm thuật toán ZeroR**

ZeroR là thuật toán đơn giản. Nó chọn giá trị lớp chiếm đa số trong tập dữ liệu và đưa ra giá trị đó cho tất cả các dự đoán

Chọn ***‘Choose -> rules-> ZeroR’*** thay đổi các tính chất liên quan và chọn ok để thêm thuật toán

Graphical user interface, application

Description automatically generated

Hình 5.8. Thuật toán ‘ZeroR’ và các thông số thuộc tính

**Thêm thuật toán OneR**

OneR giống như thuật toán đơn giản. Nó chọn một thuộc tính tương quan tốt nhất với giá trị lớp và chia nhỏ nó để có được độ chính xác dự đoán tốt nhất có thể

Chọn choose -> rules-> OneR thay đổi các tính chất liên quan và chọn ok để thêm thuật toán

Graphical user interface, text, application

Description automatically generated

Hình 5.9. Thuật toán ‘OneR’ và các thông số thuộc tính

**Thêm thuật toán J48**

**J48** là thuật toán cây quyết định. Nó là một triển khai của thuật toán C4.8 trong Java (“J” cho Java và 48 cho C4.8). Thuật toán C4.8 là một phần mở rộng nhỏ của thuật toán C4.5 nổi tiếng và là một thuật toán dự đoán rất mạnh. (Phải có thuộc tính phân lớp để chạy)

Chọn choose -> Trees-> J48 thay đổi các tính chất liên quan và chọn ok để thêm thuật toán

Graphical user interface, application

Description automatically generated

Hình 5.10. Thuật toán ‘J48’ và các thông sốt thuộc tính

Chọn OK

Sau đó xem danh sách các thuật toán đã hoặc nếu muốn thay đổi tính chất của 1 thuật toán, có thể chọn edit để thay đổi, tương tự nếu muốn xóa có thể chọn delete để xóa thuật toán

Graphical user interface

Description automatically generated

Hình 5.11. Kết quả sau khi thêm các thuật toán

**Bước 4: Chạy thử nghiệm**

Chọn tab Run -> start để bắt đầu để bắt đầu thử nghiệm và xem phần “Nhật ký” và “Trạng thái” để theo dõi hoạt động của thử nghiệm.

Graphical user interface, text, application, email

Description automatically generated

Hình 5.12. Chọn vào mục ‘Run’ và thực hiện ‘Start’ để thực hiện

Nếu không có lỗi thì thử nghiệm đã thành công và chuyển sang bước tiếp theo

**Bước 5: Đánh giá thử nghiệm**

Chọn tab ***‘Analyse -> Experiment’***. Thao tác này sẽ mở bảng phân tích kết quả thử nghiệm.

Graphical user interface, text, application

Description automatically generated

Hình 5.13. Sau khi thực hiện xong và không báo lỗi ấn qua mục ‘Analyse’ để xem kết quả

Ở phần TestBase chọn select -> chọn Ranking -> Perform test để thực hiện xem thứ hạng của thuật toán

Graphical user interface, application

Description automatically generated

Hình 5.14. Chọn ‘Ranking’ để xem thứ hạng thuật toán

Bảng xếp hạng cho biết độ chính xác của các thuật toán và sự khác biệt có ý nghĩa thống kê

Như trên hình ta có thể thấy J48 có độ chính xác cao hơn của OneR và ZeroR

Để xem được các thuật toán đạt được điểm số nào, ta thực hiện:

***‘Testbase -> select -> ZeroR -> select’***, sau đó tích vào Show std. deviation và chọn Perform test

Để thực hiện kiểm tra sai lệch giữa các thuật toán.

Graphical user interface

Description automatically generated

Hình 5.15. Xem các sai lệch của các thuật toán

Ta có thể thấy ở kết quả của trees.j48 có chữ ‘v’ so với kết quả của ZeroR và OneR, điều này có nghĩa là sự khác biệt về độ chính xác của J48 so với ZeroR và OneR là có ý nghĩa thống kê và thấy đc độ chính xác và sự sai lệch của J48 so với ZeroR và OneR.

**Bước 6: Lưu kết quả thử nghiệm**

Chọn Save Output , đặt tên và chọn nơi lưu các mẫu thử nghiệm

Graphical user interface, text, application, chat or text message

Description automatically generated

Hình 5.16. Thực hiện lưu lại các kết quả đã chạy

Graphical user interface, application

Description automatically generated

Hình 5.17. Lưu 3 kết quả đã thực hiện được

TÀI LIỆU THAM KHẢO

1. Apriori\_DataMining.pdf. (n.d.).
2. https://app.datacamp.com/workspace/preview?\_tag=rdocs&rdocsPath=packages%2FrCBA%2Fversions%2F0.4.3%2Ftopics%2Ffpgrowth&utm\_source=r-docs&utm\_medium=docs&utm\_term=fpgrowth&utm\_content=run\_example\_in\_workspace. (n.d.).
3. https://cafef.vn/618-nguoi-lao-dong-tin-rang-tim-viec-nam-2023-se-kho-khan-hon-20230215065755291.chn. (n.d.).
4. https://machinelearningmastery.com/design-and-run-your-first-experiment-in-weka/. (n.d.).
5. https://viblo.asia/p/thuat-toan-apriori-khai-pha-luat-ket-hop-trong-data-mining-3P0lPEv85ox. (n.d.).
6. https://www.academia.edu/16029167/Fp\_tree. (n.d.).
7. https://www.mssqltips.com/sqlservertip/4630/connecting-weka-to-sql-server-2016/?fbclid=IwAR1xzWYzr8kOzBm8orVHbq\_90C2f4zh5RahZXKrL31Oo5EK6wJgSaoFXRGM. (n.d.).
8. Lý, S. B.-G. (n.d.).
9. Mot\_so\_phan\_mem\_khac\_khai\_pha\_du\_lieu\_[DataMining].pdf. (n.d.).
10. R\_baisic.pdf. (n.d.).