|  |
| --- |
| **ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**  **KHOA HỆ THỐNG THÔNG TIN**  **LÝ HOA NAM – 15520511**  **PHAN VIỆT ANH – 14520033**  **KHÓA LUẬN TỐT NGHIỆP**  **SỬ DỤNG THUẬT TOÁN KHUYẾN NGHỊ PHÙ HỢP**  **ĐỂ KHAI PHÁ DỮ LIỆU CHO DOANH NGHIỆP**  **USING APPROPRIATE RECOMMENDED ALGORITHMS**  **TO MINING DATA FOR BUSINESSES**  **KỸ SƯ NGÀNH HỆ THỐNG THÔNG TIN**  **TP. HỒ CHÍ MINH, 6/2019** |
| **ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**  **KHOA HỆ THỐNG THÔNG TIN**  **LÝ HOA NAM – 15520511**  **PHAN VIỆT ANH – 14520033**  **KHÓA LUẬN TỐT NGHIỆP**  **SỬ DỤNG THUẬT TOÁN KHUYẾN NGHỊ PHÙ HỢP**  **ĐỂ KHAI PHÁ DỮ LIỆU CHO DOANH NGHIỆP**  **USING APPROPRIATE RECOMMENDED ALGORITHMS**  **TO MINING DATA FOR BUSINESSES**  **KỸ SƯ NGÀNH HỆ THỐNG THÔNG TIN**  **GIẢNG VIÊN HƯỚNG DẪN**  **TH.S THÁI BẢO TRÂN**  **TP. HỒ CHÍ MINH, 6/2019** |

**DANH SÁCH HỘI ĐỒNG BẢO VỆ KHÓA LUẬN**

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số……..ngày…….của hiệu trưởng Trường Đại học Công Nghệ Thông Tin.

1. - Chủ tịch.
2. - Thư ký.
3. - Ủy viên.
4. - Ủy viên.

|  |  |
| --- | --- |
| **ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |
|  | *TP. HCM, ngày…..tháng…..năm……..* |

**NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP CỦA CÁN BỘ HƯỚNG DẪN**

|  |  |  |
| --- | --- | --- |
| **Tên khóa luận:** | | |
| **SỬ DỤNG THUẬT TOÁN KHUYẾN NGHỊ PHÙ HỢP ĐỂ KHAI PHÁ DỮ LIỆU CHO DOANH NGHIỆP** | | |
| **Nhóm SV thực hiện:** | **Cán bộ hướng dẫn:** | |
| Lý Hoa Nam | 15520511 | ThS. Thái Bảo Trân | |
| Phan Việt Anh | 14520033 |  | |
| **Đánh giá Khóa luận**   1. Về cuốn báo cáo:   Số trang Số chương  Số bảng số liệu Số hình vẽ  Số tài liệu tham khảo Sản phẩm  Một số nhận xét về hình thức cuốn báo cáo:  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................   1. Về nội dung nghiên cứu:   ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................   1. Về chương trình ứng dụng:   ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................   1. Về thái độ làm việc của sinh viên:   ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  **Đánh giá chung:**  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  **Điểm từng sinh viên:**  **Lý Hoa Nam:………../10**  **Phan Việt Anh:………../10** | | |

|  |  |
| --- | --- |
|  | **Người nhận xét**  (Ký tên và ghi rõ họ tên) |

|  |  |
| --- | --- |
| **ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |
|  | *TP. HCM, ngày…..tháng…..năm……..* |

**NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP CỦA CÁN BỘ PHẢN BIỆN**

|  |  |  |
| --- | --- | --- |
| **Tên khóa luận:** | | |
| **SỬ DỤNG THUẬT TOÁN KHUYẾN NGHỊ PHÙ HỢP ĐỂ KHAI PHÁ DỮ LIỆU CHO DOANH NGHIỆP** | | |
| **Nhóm SV thực hiện:** | **Cán bộ hướng dẫn:** | |
| Lý Hoa Nam | 15520511 | ThS. Thái Bảo Trân | |
| Phan Việt Anh | 14520033 |  | |
| **Đánh giá Khóa luận**   1. Về cuốn báo cáo:   Số trang Số chương  Số bảng số liệu Số hình vẽ  Số tài liệu tham khảo Sản phẩm  Một số nhận xét về hình thức cuốn báo cáo:  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................   1. Về nội dung nghiên cứu:   ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................   1. Về chương trình ứng dụng:   ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................   1. Về thái độ làm việc của sinh viên:   ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  **Đánh giá chung:**  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  ..............................................................................................................................  **Điểm từng sinh viên:**  **Lý Hoa Nam:………../10**  **Phan Việt Anh:………../10** | | |

|  |  |
| --- | --- |
|  | **Người nhận xét**  (Ký tên và ghi rõ họ tên) |

**LỜI CẢM ƠN**

Khóa luận tốt nghiệp là kết quả của việc nghiên cứu và học tập của nhóm tác giả trong suốt thời gian ngồi trên ghế giảng đường của Trường Đại Học Công Nghệ Thông Tin. Nhóm tác giả không thể hoàn thành tốt khóa luận tốt nghiệp nếu không có sự quan tâm và giúp đỡ tận tình của quý Thầy Cô, bạn bè và gia đình.

Đầu tiên, nhóm tác giả xin gửi lời cảm ơn chân thành đến quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP. Hồ Chí Minh và quý Thầy Cô khoa Hệ thống thông tin đã tạo môi trường và điều kiện thuận lợi, giúp cho nhóm tác giả có những kiến thức cơ bản làm nền tảng để thực hiện đề tài.

Đặc biệt, nhóm tác giả xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới giảng viên hướng dẫn – ThS. Thái Bảo Trân. Cô đã trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp nhiều ý kiến quý báu giúp nhóm tác giả hoàn thành khóa luận tốt nghiệp của mình.

Bên cạnh đó, nhóm tác giả xin được cảm ơn gia đình, người thân và bạn bè đã động viên, khích lệ và hỗ trợ nhóm về mặt tinh thần trong suốt thời gian thực hiện khóa luận tốt nghiệp.

Trong thời gian một học kỳ thực hiện đề tài, nhóm tác giả đã vận dụng những kiến thức nền tảng đã tích lũy được, đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới nhằm hoàn thành một báo cáo đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm tác giả không tránh khỏi những thiếu sót. Chính vì vậy, nhóm tác giả rất mong nhận được những sự góp ý từ phía các Thầy Cô nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài khác trong tương lai.

Xin chân thành cảm ơn quý Thầy Cô !

Nhóm tác giả

|  |  |
| --- | --- |
| ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH  **TRƯỜNG ĐẠI HỌC**  **CÔNG NGHỆ THÔNG TIN** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc Lập - Tự Do - Hạnh Phúc** |

**ĐỀ CƯƠNG CHI TIẾT**

|  |
| --- |
| **TÊN ĐỀ TÀI:**  **Sử dụng thuật toán khuyến nghị phù hợp để khai phá dữ liệu cho doanh nghiệp (Using appropriate recommended algorithms to mining data for businesses)** |
| **Cán bộ hướng dẫn: ThS. Thái Bảo Trân** |
| **Thời gian thực hiện:** Từ tháng 2/2019 đến tháng 6/2019 |
| **Sinh viên thực hiện: Lý Hoa Nam – 15520511**  **Phan Việt Anh – 14520033** |
| **NỘI DUNG ĐỀ TÀI**   1. **Đặt vấn đề**   Hiện nay các webiste vừa và nhỏ đang được phát triển rất nhiều, dù ở loại hình kinh doanh nào, quy mô lớn hay nhỏ nhu cầu website đối với doanh nghiệp đó rất cần thiết. Theo thống kê của bộ thông tin và truyền thông[[1]](#footnote-1), trong năm 2018 đã có tới hơn 1.212.000 website mới được thành lập. Trong đó, các website vừa và nhỏ chiếm tỉ lệ rất cao. Do kinh phí để các doanh nghiệp, cá nhân có thể chi trả cho website có các chức năng phức tạp không cao, vì thế những website này còn thiếu những chức năng cần thiết – điển hình là chức năng “khuyến nghị”.  Hệ thống khuyến nghị (Recommender System) là một loại hình cụ thể của kỹ thuật lọc thông tin (như phim ảnh, âm nhạc, trang web, tin tức) mà người dùng quan tâm và dần dần trở nên phổ cập trong các ứng dụng khác nhau (ví dụ như dự án Netflix, Google tin tức, Amazon). Cụ thể, công cụ khuyến nghị sản phẩm đã góp phần làm tăng doanh số bán hàng của Amazon 29% lên tới 12,83 tỷ USD trong quý tài chính thứ hai từ mức 9,9 tỷ USD trong cùng kỳ năm 2011 và doanh thu quý III năm 2013 tăng tới 24%, lên 17,1 tỷ USD[[2]](#footnote-2).  Chính vì vậy, với mong muốn các website vừa và nhỏ vẫn có thể ứng dụng hệ khuyến nghị dù kinh phí thấp. Nhóm tác giả mong muốn xây dựng một website giúp người dùng có thể đưa dữ liệu đầu vào của họ và website sẽ trả về kết quả hệ khuyến nghị phù hợp với dữ liệu đã cho. Không chỉ vậy, hệ thống còn dashboard và trực quan hóa dữ liệu - giúp người dùng có thể xem bảo cáo, phân tích, thống kê,...tình hình nhập/xuất/mua bán,...của doanh nghiệp một cách dễ dàng. Từ đó, giúp cho chủ doanh nghiệp có thể nắm rõ tình hình kinh doanh, so sánh kết quả, có chiến lược quảng cáo, đầu tư kinh doanh hiệu quả hơn. Mặt khác, khóa luận còn là tài liệu tham khảo giúp các sinh viên khóa sau có thể xác định được thuật toán khuyến nghị phù hợp cho đề tài đồ án, khóa luận của mình.  **2. Mục tiêu đề tài**  Như đã nêu, với mong muốn xây dựng ứng dụng hỗ trợ người dùng ứng dụng hệ  khuyến nghị vào các website của họ để giải quyết vấn đề:   * Đọc và xử lý dữ liệu người dùng. * Hiển thị dữ liệu. * Chạy thuật toán khuyến nghị. * Hiển thị kết quả. * Xuất kết quả quả api để người dùng phát triển website.   **3. Phạm vi và đối tượng nghiên cứu**  **3.1 Phạm vi công nghệ**   * Ngôn ngữ lập trình web: Python, Javascript. * Công nghệ hỗ trợ: HTML5, CSS3, AJAX, jQuery, Bootstrap 3, React Js, surpriselib, flask, pandas.   **3.2 Đối tượng nghiên cứu**   * Các cách tiền xử lý dữ liệu. * Nhu cầu và thực trạng sử dụng hệ khuyến nghị hiện nay. * Các thuật toán hệ khuyến nghị * Các cách biểu diễn dữ liệu (data visualization).   **4. Phương pháp nghiên cứu**  **4.1 Phương pháp khảo sát**   * Tìm hiểu nhu cầu áp dụng hệ khuyến nghị trên các website vừa và nhỏ. * Tìm hiểu đầu vào và đầu ra của thuật toán và chuyển hóa sang dữ liệu đáp ứng nhu cầu của người dùng.   **4.2 Phương pháp phân tích thiết kế**  Phân tích thuật toán trên nền tảng python và hiện thị bằng web application.  **4.3 Phương pháp lập trình**  Sử dụng ngôn ngữ lập trình python để chạy thư viện thuật toán sau đó dùng thư viện flask của python để đưa kết quả bằng api lên và sử dụng javascript cùng với thư viện reactjs để hiện thị kết quả.  **5. Nội dung thực hiện**  **5.1 Khảo sát phân tích hiện trạng và nhu cầu**   * Khảo sát nhu cầu sử dụng hệ khuyến nghị trong các website bán hàng vừa và nhỏ. * Khảo sát cách áp dụng công nghệ hiểu quả khi triển khai trên website người dùng. * Khảo sát thuật toán phù hợp với nhu cầu khuyến nghị và dữ liệu hiện có.   **5.2 Tìm hiểu và lựa chọn công nghệ, giải thuật**  **5.2.1 Công nghệ Web**   * Tìm hiểu công nghệ web ứng dụng để hiện thị và truy xuất dữ liệu nhanh chống và hiệu quả. * Tìm hiểu ngôn ngữ lập trình python xử lý backend và ngôn ngữ javascript hiện thị phần frontend cùng các công nghệ khác.   **5.2.2 Quản lý và tiền xử lý dữ liệu**  Tìm hiểu các kỹ thuật tiền xử lý dữ liệu và lưu trữ truy xuất dữ liệu hiệu quả.  **5.2.3 Tìm hiểu và ứng dụng thuật toán gợi ý**   * Tìm hiểu các thuật toán gợi ý trong thư viện (<http://surpriselib.com/>) như SVD, SVD++, k-NN… và áp dụng thuật toán trong bài toán thực tiễn. * Các thuật toán gợi ý sẽ dùng các kỹ thuật như ma trận, phân lớp … để đưa ra các gợi ý sản phẩm phù hợp với dữ liệu đầu vào có sẵn.   **5.3 Xây dựng môi trường chạy thuật toán**  **5.3.1 Tìm hiểu thuật toán.**   * Tìm hiểu các thuật toán khai phá dữ liệu, nguyên lý hoạt động của nó. * Đề xuất các thuật toán cho dataset, tích hợp, so sánh và đánh giá kết quả.   **5.3.2 Cài đặt môi trường.**   * Các thuật toán được chạy trên môi trường python, với các chức năng yêu cầu để trả về kết quả mong muốn. * Kết quả sẽ được chạy trên môi trường javascript, với các chức năng yêu cầu để trả về kết quả mong muốn.   **5.3 Xây dựng ứng dụng**  **5.4.1 Phân tích thiết kế hệ thống**  Qua khảo sát, tìm hiểu nhu cầu người dùng ở trên. Từ đó, phân tích các tính năng cần có của website: nhập dữ liệu, hiển thị dữ liệu, biểu diễn dữ liệu, …  **5.4.2 Lập trình xây dựng ứng dụng**  Dùng ngôn ngữ lập trình python và javascript, cùng các framework và thư viện liên  quan để xây dựng ứng dụng với các chức năng đáp ứng yêu cầu đã đề ra.  **5.4.3 Kiểm thử và sửa lỗi**  Kiểm thử và sửa lỗi để hoàn chỉnh ứng dụng. |
| **Kế hoạch thực hiện**   |  |  |  |  | | --- | --- | --- | --- | | **STT** | **Nội dung thực hiện** | | **Thời gian dự kiến (tuần)** | | 1 | Khảo sát, phân tích hiện trạngvà nhu cầu | | 1 | | 2 | Tìm hiểu và lựa chọn công nghệ, giải thuật | | 1 | | 3 | Tìm các dữ liệu mẫu | | 1 | | 4 | Cài đặt thuật toán | | 4 | | 4 | Hiện thị (data visualization) | | 5 | | 5 | Kiểm thử và sửa lỗi | | 2 | | **Tổng cộng** | | | 15 tuần | | **Xác nhận của CBHD**  (Ký tên và ghi rõ họ tên) | | | **TP. HCM, ngày 10 tháng 3 năm 2019**  **Sinh viên**  (Ký tên và ghi rõ họ tên) | | | |

**MỤC LỤC**

[DANH MỤC SƠ ĐỒ, HÌNH VẼ i](#_Toc14514809)

[DANH MỤC BẢNG iv](#_Toc14514810)

[DANH MỤC CHỮ VIẾT TẮT v](#_Toc14514811)

[TÓM TẮT 1](#_Toc14514812)

[CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI 2](#_Toc14514813)

[1.1 Đặt vấn đề 2](#_Toc14514814)

[1.2 Khảo sát hiện trạng 3](#_Toc14514815)

[1.3 Mục tiêu khóa luận 5](#_Toc14514816)

[1.4 Đối tượng, phạm vi 6](#_Toc14514817)

[1.4.1 Phạm vi đề tài 6](#_Toc14514818)

[1.4.2 Phạm vi công nghệ 6](#_Toc14514819)

[1.4.3 Đối tượng nghiên cứu 6](#_Toc14514820)

[1.5 Nội dung thực hiện 6](#_Toc14514821)

[1.5.1 Khảo sát phân tích hiện trạng và nhu cầu 6](#_Toc14514822)

[1.5.2 Tìm hiểu và lựa chọn công nghệ, giải thuật 7](#_Toc14514823)

[1.5.3 Xây dựng môi trường chạy thuật toán 7](#_Toc14514824)

[1.5.4 Xây dựng ứng dụng 7](#_Toc14514825)

[1.6 Bố cục báo cáo 8](#_Toc14514826)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 9](#_Toc14514827)

[2.1 K-nearest neighbor 9](#_Toc14514828)

[2.2 K-means Clustering 16](#_Toc14514829)

[2.3 Thuật toán Apriori khai phá luật kết hợp 33](#_Toc14514830)

[2.3.1 Luật kết hợp trong khai phá dữ liệu (Association Rule in Data Mining) 33](#_Toc14514831)

[2.3.2 Thuật toán sinh các luật kết hợp Apriori *(by Agrawal and Srikant 1994)* 35](#_Toc14514832)

[2.4 Thuật toán FP-Growth 37](#_Toc14514833)

[2.5 Thuật toán dựa trên hệ số hóa ma trận (Matrix Factorization-based algorithms) 40](#_Toc14514834)

[2.6 Thuật toán Slope One 42](#_Toc14514835)

[2.7 Môi trường Python, NodeJs 43](#_Toc14514836)

[2.8 Các thư viện hỗ trợ Python 44](#_Toc14514837)

[2.8.1 Tổng quan thư viện trong ứng dụng 44](#_Toc14514838)

[2.8.2 Thư viện Flask-RESTPlus 44](#_Toc14514839)

[2.8.3 Mô hình cấu trúc chức năng (framework functional structure ) 44](#_Toc14514840)

[2.8.4 Thư viện JWT (thư viện tạo token, mã hóa thuật toán HS256) 45](#_Toc14514841)

[2.8.5 Flask\_sqlalchemy (thư viện hỗ trợ kết nối với MySQL) 47](#_Toc14514842)

[2.9 Thư viện thuật toán 47](#_Toc14514843)

[2.9.1 Pandas 47](#_Toc14514844)

[2.9.2 Sklearn (Scikit-learn) 48](#_Toc14514845)

[2.9.3 Surprise 49](#_Toc14514846)

[2.10 Thư viện hỗ trợ chính: Reactjs 50](#_Toc14514847)

[2.11 Các gọi hỗ trợ (package: react-bootstrap, react-easy-chart) 51](#_Toc14514848)

[CHƯƠNG 3. PHÂN TÍCH, THIẾT KẾ HỆ THỐNG 52](#_Toc14514849)

[3.1 Phân tích yêu cầu 52](#_Toc14514850)

[3.1.1 Mô hình phân rã chức năng (BFD – Business Function Diagram) 52](#_Toc14514851)

[3.1.2 Yêu cầu chức năng 52](#_Toc14514852)

[3.1.3 Yêu cầu phi chức năng 52](#_Toc14514853)

[3.2. Phân tích hệ thống về xử lý - Sơ đồ DFD (Data Flow Diagram) 53](#_Toc14514854)

[3.2.1. DFD mức 0 53](#_Toc14514855)

[3.2.2. DFD mức 1 54](#_Toc14514856)

[3.2.3. DFD mức 2 tổng quát 55](#_Toc14514857)

[3.3. Phân tích thiết kế về dữ liệu 59](#_Toc14514858)

[CHƯƠNG 4. CÀI ĐẶT VÀ TRIỂN KHAI HỆ THỐNG 61](#_Toc14514859)

[4.1 Cài đặt và triển khai hệ thống 61](#_Toc14514860)

[4.1.1 Các công cụ xây dựng hệ thống 61](#_Toc14514861)

[4.1.2 Triển khai hệ thống 61](#_Toc14514862)

[4.2 Giao diện của hệ thống 62](#_Toc14514863)

[4.2.1 Sơ đồ tổ chức giao diện 62](#_Toc14514864)

[4.2.2 Một số giao diện chính 63](#_Toc14514865)

[CHƯƠNG 5. THỬ NGHIỆM HỆ THỐNG 75](#_Toc14514866)

[5.1 Thử nghiệm 75](#_Toc14514867)

[5.2 Đánh giá kết quả thử nghiệm 76](#_Toc14514868)

[5.2.1 Kết quả thử nghiệm 76](#_Toc14514869)

[5.2.2 Đánh giá thực nghiệm 88](#_Toc14514870)

[CHƯƠNG 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 90](#_Toc14514914)

[6.1 Kết quả đạt được 90](#_Toc14514915)

[6.2 Đánh giá ưu điểm, khuyết điểm 91](#_Toc14514916)

[6.2.1 Ưu điểm 91](#_Toc14514917)

[6.2.2 Khuyết điểm 91](#_Toc14514918)

[6.3 Hướng phát triển 92](#_Toc14514919)

[TÀI LIỆU THAM KHẢO 93](#_Toc14514920)

# **DANH MỤC SƠ ĐỒ, HÌNH VẼ**

[Hình 2‑1 Kết quả chạy thuật toán KNN 10](#_Toc14465545)

[Hình 2‑2 Iris flower dataset 11](#_Toc14465546)

[Hình 2‑3 Dữ liệu mẫu cho thuật toán KNN 15](#_Toc14465547)

[Hình 2‑4 Sơ đồ thực hiện thuật toán KNN 15](#_Toc14465548)

[Hình 2‑5 Bài toán với 3 clusters 17](#_Toc14465549)

[Hình 2‑6 Biểu đồ thể hiện K-means 18](#_Toc14465550)

[Hình 2‑7 Kết quả thuật toán K-means 25](#_Toc14465551)

[Hình 2‑8 Kết quả thuật toán K-means 27](#_Toc14465552)

[Hình 2‑9 Lần lặp thứ 0 của K-means 28](#_Toc14465553)

[Hình 2‑10 Lần lặp thứ 1 của K-means 28](#_Toc14465554)

[Hình 2‑11 Lần lặp thứ 2 của K-means 29](#_Toc14465555)

[Hình 2‑12 Lần lặp thứ 3 của K-means 29](#_Toc14465556)

[Hình 2‑13 Lần lặp thứ 4 của K-means 30](#_Toc14465557)

[Hình 2‑14 Lần lặp thứ 5 của K-means 30](#_Toc14465558)

[Hình 2‑15 Lần lặp thứ 6 của K-means 31](#_Toc14465559)

[Hình 2‑16 Kết quả sau khi thực hiện K-means của Scikit-learn 32](#_Toc14465560)

[Hình 2‑17 Kết quả thuật toán K-means sau khi đạt tới lần lặp thứ 25 33](#_Toc14465561)

[Hình 2‑18 Cơ sở dữ liệu giao dịch 36](#_Toc14465562)

[Hình 2‑19 Mô tả các bước chạy thuật toán Apriori 36](#_Toc14465563)

[Hình 2‑20 Kết quả chạy thuật toán Apriori 37](#_Toc14465564)

[Hình 2‑21 Bảng dữ liệu 38](#_Toc14465565)

[Hình 2‑22 Các tập mục trong thuật toán FP-Growth 38](#_Toc14465566)

[Hình 2‑23 Mục phổ biến mức 1 FP-Growth 38](#_Toc14465567)

[Hình 2‑24 FP-Tree được xây dựng sau khi duyệt các items phổ biến của các giao dịch 39](#_Toc14465568)

[Hình 2‑25 Mẫu phổ biến 40](#_Toc14465569)

[Hình 2‑26 Giới thiệu JWT 46](#_Toc14465570)

[Hình 2‑27 Logo Scikit-learn 49](#_Toc14465571)

[Hình 2‑28 Logo React 50](#_Toc14465572)

[Hình 3‑1 Mô hình phân rã chức năng (BFD) 52](#_Toc14465573)

[Hình 3‑2 Sơ đồ DFD mức 0 53](#_Toc14465574)

[Hình 3‑3 Sơ đồ DFD mức 1 54](#_Toc14465575)

[Hình 3‑4 Sơ đồ DFD mức 2 55](#_Toc14465576)

[Hình 3‑5 DFD mức 2 xử lý đăng ký 56](#_Toc14465577)

[Hình 3‑6 DFD mức 2 xử lý đăng nhập 56](#_Toc14465578)

[Hình 3‑7 DFD mức 2 xử lý datasets 57](#_Toc14465579)

[Hình 3‑8 DFD mức 2 xử lý thuật toán khuyến nghị và hiển thị kết quả, thống kê 58](#_Toc14465580)

[Hình 3‑9 Cơ sở dữ liệu 59](#_Toc14465581)

[Hình 4‑1 Sơ đồ triển khai của hệ thống 61](#_Toc14465582)

[Hình 4‑2 Sơ đồ tổ chức giao diện 62](#_Toc14465583)

[Hình 4‑3 Giao diện đăng nhập 63](#_Toc14465584)

[Hình 4‑4. Tạo tài khoản người dùng 64](#_Toc14465585)

[Hình 4‑5. Tổng quan trang chủ 65](#_Toc14465586)

[Hình 4‑6. Thông tin file dữ liệu 65](#_Toc14465587)

[Hình 4‑7 Thông tin kiểu dữ liệu 66](#_Toc14465588)

[Hình 4‑8. Biểu đồ số lượng giá trị độc nhất có trong dữ liệu 66](#_Toc14465589)

[Hình 4‑9. Thông tin mô tả giá trị trong file dữ liệu 67](#_Toc14465590)

[Hình 4‑10. Số liệu có trong file 68](#_Toc14465591)

[Hình 4‑11. Thông tin tài khoản người dùng 69](#_Toc14465592)

[Hình 4‑12. Chọn thuật toán 70](#_Toc14465593)

[Hình 4‑13. Chọn cột đầu vào thuật toán K-means 70](#_Toc14465594)

[Hình 4‑14. Chọn đầu vào thuật toán gợi ý 71](#_Toc14465595)

[Hình 4‑15. Tổng quan giao diện khi chạy thuật toán k-mean 72](#_Toc14465596)

[Hình 4‑16. Tổng quan giao diện khi chạy thuật toán Association Rule 73](#_Toc14465597)

[Hình 4‑17. Tổng quan giao diện khi chạy thuật toán predict 74](#_Toc14465598)

[Hình 5‑1 Kết quả khi chạy công thức tính min\_sup và min\_conf 76](#_Toc14465599)

[Hình 5‑2 Kết quả khi chạy thuật toán Apriori 77](#_Toc14465600)

[Hình 5‑3 Kết quả khi chạy thuật toán FP-Growth 78](#_Toc14465601)

[Hình 5‑4 Kết quả khi chạy thuật toán K-means 79](#_Toc14465602)

[Hình 5‑5 Kết quả khi chạy thuật toán K-means 80](#_Toc14465603)

[Hình 5‑6 Kết quả biểu đồ elbow 80](#_Toc14465604)

[Hình 5‑7 Kết quả khi chạy thuật toán KNN 81](#_Toc14465605)

[Hình 5‑8 Kết quả khi chạy thuật toán NMF 82](#_Toc14465606)

[Hình 5‑9 Kết quả khi chạy thuật toán Slope One 83](#_Toc14465607)

# **DANH MỤC BẢNG**

[Bảng 4‑1 Mô tả sơ đồ tổ chức giao diện 63](#_Toc14347972)

[Bảng 5‑1 Mô tả thử nghiệm hệ thống 75](#_Toc14347973)

# **DANH MỤC CHỮ VIẾT TẮT**

|  |  |
| --- | --- |
| KNN | K-nearest neighbors |
| NMF | Non-negative Matrix Factorization |
| Minimum support | min\_sup |
| Minimum confidence | min\_conf |

# **TÓM TẮT**

Khai phá dữ liệu là một công việc rất quan trọng trong hầu hết các loại lĩnh vực, cụ thể là: Trong lĩnh vực tài chính ngân hàng, khai thác dữ liệu được sử dụng để tạo ra các mô hình rủi ro chính xác cho các khoản vay và thế chấp. Họ cũng rất hữu ích khi phát hiện các giao dịch gian lận. Trong tiếp thị, kỹ thuật khai thác dữ liệu được sử dụng để cải thiện chuyển đổi, tăng sự hài lòng của khách hàng và tạo ra các chiến dịch quảng cáo được nhắm mục tiêu. Họ thậm chí có thể được sử dụng khi phân tích nhu cầu trên thị trường và tìm ra ý tưởng cho các dòng sản phẩm hoàn toàn mới. Điều này được thực hiện bằng cách xem dữ liệu khách hàng và bán hàng lịch sử và tạo ra các mô hình dự đoán mạnh mẽ. Các cửa hàng bán lẻ sử dụng các thói quen / chi tiết mua sắm của khách hàng để tối ưu hóa cách bố trí các cửa hàng của họ nhằm nâng cao trải nghiệm của khách hàng và tăng lợi nhuận. Các cơ quan quản lý thuế sử dụng các kỹ thuật khai thác dữ liệu để phát hiện các giao dịch gian lận và khai thuế đáng ngờ hoặc các tài liệu kinh doanh khác.

Đề tài tập trung xây dựng một website giúp người dùng có thể đưa dữ liệu đầu vào của họ và website sẽ trả về kết quả hệ khuyến nghị phù hợp với dữ liệu đã cho. Không chỉ vậy, hệ thống còn dashboard và trực quan hóa dữ liệu – giúp người dùng có thể xem báo cáo, phân tích, thống kê,…tình hình nhập/xuất/mua/bán,… của doanh nghiệp một cách dễ dàng. Từ đó, giúp cho chủ doanh nghiệp có thể nắm rõ tình hình kinh doanh, so sánh kết quả, có chiến lược quảng cáo, đầu tư kinh doanh hiệu quả hơn. Mặt khác, khóa luận còn là tài liệu tham khảo giúp các sinh viên khóa sau có thể xác định được thuật toán khuyến nghị phù hợp cho đề tài đồ án, khóa luận của mình.

# **TỔNG QUAN ĐỀ TÀI**

* 1. **Đặt vấn đề**

Hiện nay các website vừa và nhỏ đang được phát triển rất nhiều, dù ở loại hình kinh doanh nào, quy mô lớn hay nhỏ nhu cầu website đối với doanh nghiệp đó rất cần thiết. Theo thống kê của bộ thông tin và truyền thông, trong năm 2018 đã có tới hơn 1.212.000 website mới được thành lập. Trong đó, các website vừa và nhỏ chiếm tỉ lệ rất cao [[3]](#footnote-3). Do kinh phí để các doanh nghiệp, cá nhân có thể chi trả cho website có các chức năng phức tạp không cao, vì thế những website này còn thiếu những chức năng cần thiết – điển hình là chức năng “khuyến nghị”

Hệ thống khuyến nghị (Recommender Sytstem) là một loại hình cụ thể của kỹ thuật lọc thông tin (như phim ảnh, âm nhạc, trang web, tin tức) mà người dùng quan tâm và dần dần trở nên phổ cập trong các ứng dụng khác nhau (ví dụ như dự án Netflix, Google tin tức, Amazon). Cụ thể, công cụ khuyến nghị sản phẩm đã góp phần làm tăng doanh số bán hàng của Amazon 29% lên tới 12,83 tỷ USD trong quý tài chính thứ hai từ mức 9,9 tỷ USD trong cùng kỳ năm 2011 và doanh thu quý III năm 2013 tăng tới 24% lên 17,1 tỷ USD 3.

Chính vì vậy, với mong muốn các website vừa và nhỏ vẫn có thể ứng dụng hệ khuyến nghị dù kinh phí thấp. Nhóm tác giả mong muốn xây dựng một website giúp người dùng có thể đưa dữ liệu đầu vào của họ và website sẽ trả về kết quả hệ khuyến nghị phù hợp với dữ liệu đã cho. Không chỉ vậy, hệ thống còn dashboard và trực quan hóa dữ liệu – giúp người dùng có thể xem báo cáo, phân tích, thống kê,…tình hình nhập/xuất/mua/bán,… của doanh nghiệp một cách dễ dàng. Từ đó, giúp cho chủ doanh nghiệp có thể nắm rõ tình hình kinh doanh, so sánh kết quả, có chiến lược quảng cáo, đầu tư kinh doanh hiệu quả hơn. Mặt khác, khóa luận còn là tài liệu tham khảo giúp các sinh viên khóa sau có thể xác định được thuật toán khuyến nghị phù hợp cho đề tài đồ án, khóa luận của mình.

* 1. **Khảo sát hiện trạng**

**Bùng nổ dữ liệu ở thế kỷ 21**

Tập đoàn Dữ liệu quốc tế (IDC), một đơn vị nghiên cứu thông tin thị trường, đã dự đoán rằng sau mỗi năm, “Vũ trụ số” (Digital universe – Dung lượng dữ liệu được tạo ra và sao chép) sẽ đạt 180 zettabyte (180 ngàn tỷ tỷ byte) vào năm 2025 và dung lượng dữ liệu sẽ tăng gấp đôi sau mỗi 2 năm trong thập kỷ tới. Theo tờ báo Wall Street Journal, các tập đoàn lớn như Amazon, Alphabet và Microsoft đã đầu tư 32 tỷ USD cho các “nhà máy lọc dữ liệu” với dự đoán rằng, dữ liệu sẽ ngày một tăng mạnh và nhanh hơn trong tương lai [[4]](#footnote-4).

Trung bình mỗi ngày các doanh nghiệp có thể tạo ra hàng triệu mảnh dữ liệu không cấu trúc (unstructured). Các dữ liệu này có thể ở dưới dạng thư điện tử, bài viết trên mạng xã hội hay là những cuộc trò chuyện về khách hàng, hành vi của công chúng và những xu hướng phát triển mới.

Thế hệ dữ liệu hiện nay cũng có chất lượng cao hơn do chúng nằm trong một thế giới kết nối. Theo nghiên cứu của PwC, hiện có khoảng 8,4 tỷ thiết bị di động thông minh, cảm biến, thiết bị truyền động, xe cộ, máy ghi hình và máy bay không người lái được kết nối với nhau trên toàn cầu.

Thực tế hầu hết các doanh nghiệp đang sỡ hữu nguồn dữ liệu riêng, bao gồm dữ liệu tài chính, thông tin khách hàng, các khảo sát và đánh giá về sản phẩm, quy trình hoạt động,v.v. Tuy nhiên, chỉ một số ít doanh nghiệp có thể tối ưu hóa giá trị của những dữ liệu này.

**Vai trò của dữ liệu**

Sử dụng dữ liệu để thấu hiểu thị trường, khách hàng và bản thân doanh nghiệp đóng một vai trò rất quan trọng, đặc biệt là trong khả năng quản trị doanh nghiệp, dự báo và lên kế hoạch kinh doanh. Trong những năm gần đây, ứng dụng của phân tích dữ liệu (data analytics) đang trở thành ưu tiên hàng đầu, bởi nó giúp quản lý chi phí, giảm thiểu rủi ro và là chìa khóa cho sự tăng trưởng của doanh nghiệp.

Với chiến lược phân tích và sử dụng dữ liệu phù hợp, cùng một đội ngũ chuyên gia phân tích lành nghề, lãnh đạo doanh nghiệp có thể biến chiến lược thành hành động và đặc biệt là tạo ra văn hóa làm việc dựa trên dữ liệu. Khảo sát các lãnh đạo doanh nghiệp toàn cầu 2018 của PwC cho thấy, có tới 64% CEO tin rằng, phương thức quản lý dữ liệu sẽ tạo ra những điểm khác biệt và lợi thế cạnh tranh nổi bật cho doanh nghiệp trong tương lai.

Từ hàng trăm năm trước, đã có những trường hợp sử dụng dữ liệu để giải quyết những vấn đề thiết yếu của xã hội. Trở về với nước Anh năm 1854, khi bệnh dịch tả hoành hành tại Luân Đôn đã lấy đi hàng trăm sinh mạng và trở thành nỗi khiếp đảm của người dân thành phố. Không ai tìm được nguyên nhân của dịch bệnh cho đến khi John Snow, một bác sĩ và là người đi đầu trong lĩnh vực gây tê và y tế dịch tễ, sử dụng dữ liệu để tìm ra gốc rễ vấn đề. Ông đã ghim lên bản đồ thành phố vị trí của những người nhiễm bệnh và nhận ra điểm chung là số lượng nạn nhân ngày càng tăng ở những khu vực gần trạm máy bơm nước. Dữ liệu này đã giúp ông thuyết phục công chúng và chính quyền thành phố rằng cần phải xây dựng một hệ thống thoát nước thải.

Bằng viêc vẽ lên bản đồ các trường hợp tử vong, ông đã sử dụng dữ liệu để xác định nguyên nhân thực tế và cứu hàng triệu người vào thời điểm các biện pháp chữa trị còn hạn chế. Đây là ví dụ minh chứng sức mạnh của việc thu nhập và phân tích dữ liệu, ngay cả ở những dạng nguyên thủy nhất, cũng có thể thay đổi lịch sử của nhân loại.

**Khai phá sức mạnh dữ liệu trong doanh nghiệp**

Lượng dữ liệu đang thực sự bùng nổ hơn bao giờ hết, nhưng chỉ có 0,5% toàn bộ dữ liệu trên thế giới được xử lý và sử dụng mỗi năm, theo Technology Review.

Dữ liệu đang ngày càng dễ tiếp cận và ứng dụng của dữ liệu cũng đa dạng hơn rất nhiều so với những thế kỷ trước. Sử dụng dữ liệu để thấu hiểu khách hàng, từ đó nâng cao giá trị kinh doanh, có thể được áp dụng vào tất cả bộ phận trong chuỗi giá trị, trong mọi lĩnh vực hoạt động kinh doanh của doanh nghiệp.

Để khai phá sức mạnh của dữ liệu, doanh nghiệp cần xây dựng những nguyên lý cơ bản, tập trung vào ba nội dung chính.

* Đầu tiên là doanh nghiệp cần phải tim hiểu nhu cầu đối với dữ lieu để tạo ra nền tảng dữ liệu có giá trị.
* Thứ hai, tạo ra một kho dữ liệu minh bạch và đáng tin cậy thông qua tổ chức, tích hợp, tinh giản và quản trị dữ liệu.
* Cuối cùng doanh nghiệp cần áp dụng những phương pháp, thuật toán, công cụ hỗ trợ trực quan hóa và phân tích dữ liệu nhằm khai phá được tối đa giá trị dữ liệu mang lại.

Với sự bùng nổ của dữ liệu ở thời đại vũ trụ số, thu thập dữ liệu chỉ là bước đi chập chững ban đầu. Để có thể tạo ra sự đột phá, doanh nghiệp cần hiểu rõ luồng thông tin, cách kiểm soát và quản trị dữ liệu và đặc biệt, doanh nghiệp cần phải khai phá và thấu hiểu được những giá trị tiềm tàng mà dữ liệu có thể mang lại.

* 1. **Mục tiêu khóa luận**

Như đã nêu, với mong muốn xây dựng ứng dụng hỗ trợ người dùng ứng dụng hệ thống khuyến nghị vào các website của họ để giải quyết vấn đề:

* Đọc và xử lý dữ liệu người dùng
* Hiển thị dữ liệu
* Chạy thuật toán khuyến nghị
* Hiển thị kết quả
* Xuất kết quả api để người dùng phát triển website
  1. **Đối tượng, phạm vi**

**1.4.1 Phạm vi đề tài**

Vì thời gian thực hiện đề tài có giới hạn, hệ thống được nhóm tác giả tập trung phát triển bao gồm: Xây dựng giao diện cho người dùng tương tác, áp dụng những thuật toán quyết định cài đặt vào hệ thống, chức năng bảo mật cao cho người dùng, lưu trữ dữ liệu của người dùng

**1.4.2 Phạm vi công nghệ**

**Front-end:**

* Reactjs, HTML, CSS3, AJAX, JQuery
* Các thư viện npm kèm theo: easy react chart, boostrap, jquery,…

**Back-end**

* Công nghệ hỗ trợ: surpriselib, flask, pandas, sklearn, python, Flaskrestplus framework

**1.4.3 Đối tượng nghiên cứu**

* Cách cách tiền xử lý dữ liệu
* Nhu cầu và thực trạng sử dụng hệ khuyến nghị hiện nay.
* Các thuật toán hệ khuyến nghị
* Các cách biểu diễn dữ liệu (data visualization).
  1. **Nội dung thực hiện**
     1. **Khảo sát phân tích hiện trạng và nhu cầu**
* Khảo sát nhu cầu sử dụng hệ khuyến nghị trong các website bán hàng vừa và nhỏ.
* Khảo sát cách áp dụng công nghệ hiệu quả khi triển khai trên website người dùng.
* Khảo sát thuật toán phù hợp với nhu cầu khuyến nghị và dữ liệu hiện có.
  + 1. **Tìm hiểu và lựa chọn công nghệ, giải thuật**
       1. **Công nghệ Web**
* Tìm hiểu công nghệ web ứng dụng để hiển thị và truy xuất dữ liệu nhanh chóng và hiệu quả
* Tìm hiểu ngôn ngữ lập trình python xử lý backend và ngôn ngữ javascript hiển thị phần frontend cùng các công nghệ khác.
  + - 1. **Quản lý và tiền xử lý dữ liệu**

Tìm hiểu các kỹ thuật tiền xử lý dữ liệu và lưu trữ truy xuất dữ liệu hiệu quả

* + - 1. **Tìm hiểu và ứng dụng thuật toán gợi ý**
* Tìm hiểu các thuật toán gợi ý trong thư viện (http://surpriselib.com) như SVD, SVD++, k-NN… và áp dụng thuật toán trong bài toán thực tiễn.
* Các thuật toán gợi ý sẽ dùng các kỹ thuật như ma trận, phân lớp … để đưa các gợi ý sản phẩm phù hợp với dữ liệu đầu vào có sẵn.
  + 1. **Xây dựng môi trường chạy thuật toán**
       1. **Tìm hiểu thuật toán**
* Tìm hiểu các thuật toán khai phá dữ liệu, nguyên lý hoạt động của nó.
* Đề xuất các thuật toán cho dataset, tích hợp, so sánh và đánh giá kết quả.
  + - 1. **Cài đặt môi trường**
* Các thuật toán được chạy trên môi trường python, với các chức năng yêu cầu để trả về kết quả mong muốn.
* Kết quả sẽ được chạy trên môi trường javascript, với các chức năng yêu cầu để trả về kết quả mong muốn.
  + 1. **Xây dựng ứng dụng**
       1. **Phân tích thiết kế hệ thống**

Qua khảo sát, tìm hiểu nhu cầu người dùng ở trên. Từ đó, phân tích các tính năng cần có của website: nhập dữ liệu, hiển thị dữ liệu, biểu diễn dữ liệu,…

* + - 1. **Lập trình xây dựng ứng dụng**

Dùng ngôn ngữ lập trình python và javascript cùng các framework và thư viện liên quan để xây dựng ứng dụng với các chức năng đáp ứng yêu cầu đã đề ra.

* + - 1. **Kiểm thử và sửa lỗi**

Kiểm thử và sửa lỗi để hoàn chỉnh ứng dụng.

* 1. **Bố cục báo cáo**

***Chương 1:* Tổng quan đề tài**

Giới thiệu tổng quan nhất về nội dung đề tài. Nội dung chương 1 bao gồm: Đặt vấn đề và lý do chọn đề tài, khảo sát hiện trạng, mục tiêu và phạm vi nghiên cứu, công cụ xây dựng hệ thống và bố cục của báo cáo.

***Chương 2:* Cơ sở lý thuyết**

Giới thiệu lý thuyết sử dụng trong đề tài, bao gồm các thuật toán sử dụng, giải thích các công thức toán học.

***Chương 3:* Phân tích và thiết kế hệ thống**

Trình bày chi tiết quá trình phân tích và thiết kế hệ thống dựa trên mô hình luồng dữ liệu (DFD).

***Chương 4:* Cài đặt và triển khai hệ thống**

Trình bày các kỹ thuật để triển khai hệ thống và một số giao diện của hệ thống.

***Chương 5: Thử nghiệm hệ thống***

Trình bày quá trình thử nghiệm hệ thống, kết quả và tự đánh giá của nhóm tác giả.

***Chương 6:* Kết luận và hướng phát triển**

Tóm tắt lại một lần nữa đề tài đồng thời nêu ra những ưu điểm cũng như hạn chế của hệ thống đó. Từ đó, đưa ra hướng phát triển để ứng dụng được hệ thống trong tương lai.

# **CƠ SỞ LÝ THUYẾT**

## **2.1 K-nearest neighbor**

**1. Giới thiệu**

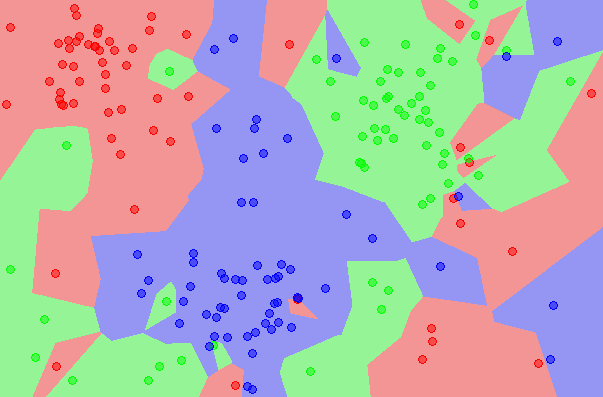
K-nearest neighbor là một trong những thuật toán supervised-learning đơn giản nhất (mà hiệu quả trong một vài trường hợp) trong Machine Learning. Khi training, thuật toán này không học một điều gì từ dữ liệu training (đây cũng là lý do thuật toán này được xếp vào loại lazy learning, mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới. K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classificaiton và Regression. KNN còn được gọi là một thuật toán Instacne-based hay Memory-based learning [1].

Có một vài khái niệm tương ứng người-máy như sau:

| **Ngôn ngữ người** | **Ngôn ngữ Máy Học** | **in Machine Learning** |
| --- | --- | --- |
| Câu hỏi | Điểm dữ liệu | Data point |
| Đáp án | Đầu ra, nhãn | Output, Label |
| Ôn thi | Huấn luyện | Training |
| Tập tài liệu mang vào phòng thi | Tập dữ liệu tập huấn | Training set |
| Đề thi | Tập dữ liểu kiểm thử | Test set |
| Câu hỏi trong dề thi | Dữ liệu kiểm thử | Test data point |
| Câu hỏi có đáp án sai | Nhiễu | Noise, Outlier |
| Câu hỏi gần giống | Điểm dữ liệu gần nhất | Nearest Neighbor |

Với KNN, trong bài toán Classification, label của một điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng major voting (bầu chọn theo số phiếu) giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label. Chi tiết sẽ được nêu trong phần tiếp theo.

Một cách ngắn gọn, KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận), không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu. Hình dưới đây là một ví dụ về KNN trong classification với K = 1.



Hình 2‑1 Kết quả chạy thuật toán KNN

Ví dụ trên đây là bài toán Classification với 3 classes: Đỏ, cam, lục. Mỗi điểm dữ liệu mới (test data point) sẽ được gán label theo màu của điểm mà nó thuộc về. Trong hình này, có một vài vùng nhỏ xem lẫn vào các vùng lớn hơn khác màu. Ví dụ có một điểm màu Lục ở gần góc 11 giờ nằm giữa hai vùng lớn với nhiều dữ liệu màu Đỏ và Lam. Điểm này rất có thể là nhiễu. Dẫn đến nếu dữ liệu test rơi vào vùng này sẽ có nhiều khả năng cho kết quả không chính xác.

**Khoảng cách trong không gian vector**

Trong không gian một chiều, khoảng cách giữa hai điểm là trị tuyệt đối giữa hiệu giá trị của hai điểm đó. Trong không gian nhiều chiều, khoảng cách giữa hai điểm có thể được định nghĩa bằng nhiều hàm số khác nhau, trong đó độ dài đường thẳng nối hai điểm chỉ là một trường hợp đặc biệt trong đó. Nhiều thông tin bổ ích (cho Machine Learning) có thể được tìm thấy tại Norms (chuẩn) của vector trong tab Math.

**2. Ví dụ trên Python**

**Bộ cơ sở dữ liệu Iris (Iris flower dataset).**

Iris flower dataset là một bộ dữ liệu nhỏ (nhỏ hơn rất nhiều so với MNIST. Bộ dữ liệu này bao gồm thông tin của ba loại hoa Iris (một loài hoa lan) khác nhau: Iris setosa, Iris virginica và Iris versicolor. Mỗi loại có 50 bông hoa được đo với dữ liệu là 4 thông tin: chiều dài, chiều rộng đài hoa (sepal), và chiều dài, chiều rộng cánh hoa (petal). Dưới đây là ví dụ về hình ảnh của ba loại hoa. (Chú ý, đây không phải là bộ cơ sở dữ liệu ảnh như MNIST, mỗi điểm dữ liệu trong tập này chỉ là một vector 4 chiều).



Hình 2‑2 Iris flower dataset

**Thí nghiệm**

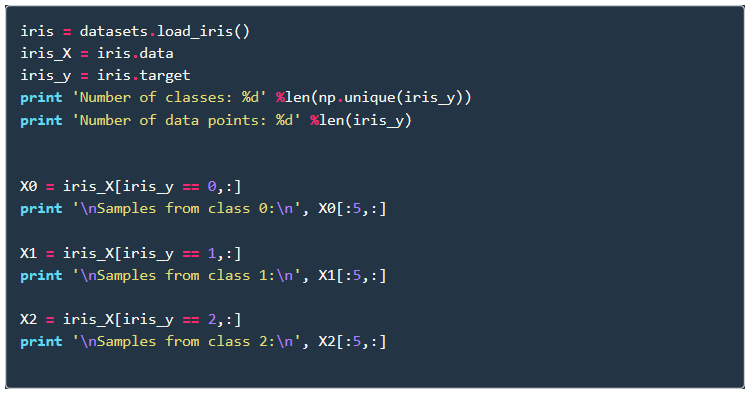
Trong phần này, chúng ta sẽ tách 150 dữ liệu trong Iris flower dataset ra thành 2 phần, gọi là training set và test set. Thuật toán KNN sẽ dựa vào trông tin ở training set để dự đoán xem mỗi dữ liệu trong test set tương ứng với loại hoa nào. Dữ liệu được dự đoán này sẽ được đối chiếu với loại hoa thật của mỗi dữ liệu trong *test set* để đánh giá hiệu quả của KNN.

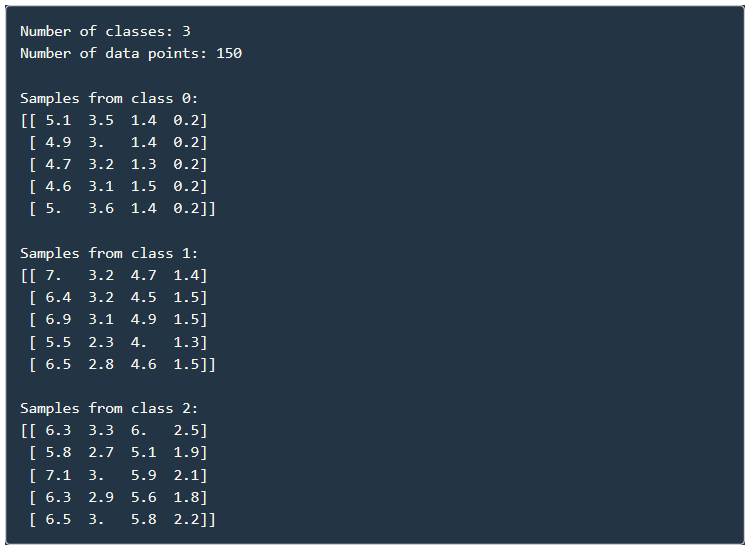
**Trước tiên, chúng ta cần khai báo vài thư viện**.

Iris flower dataset có sẵn trong thư viện scikit-learn.



**Tiếp theo, chúng ta load dữ liệu và hiện thị vài dữ liệu mẫu**. Các class được gán nhãn là 0, 1, và 2.

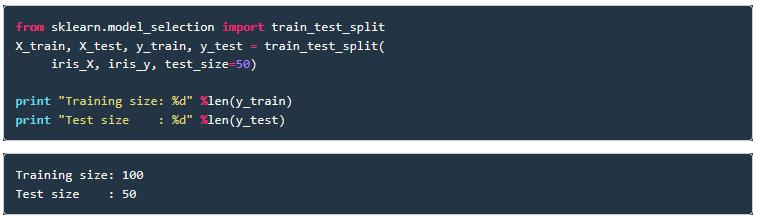




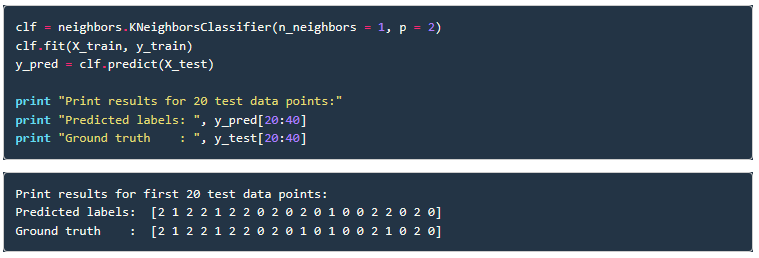
Nếu nhìn vào vài dữ liệu mẫu, chúng ta thấy rằng hai cột cuối mang khá nhiều thông tin giúp chúng ta có thể phân biệt được chúng. Chúng ta dự đoán rằng kết quả classification cho cơ sở dữ liệu này sẽ tương đối cao.

**Tách training và test sets**

Giả sử chúng ta muốn dùng 50 điểm dữ liệu cho test set, 100 điểm còn lại cho training set. Scikit-learn có một hàm số cho phép chúng ta ngẫu nhiên lựa chọn các điểm này, như sau:



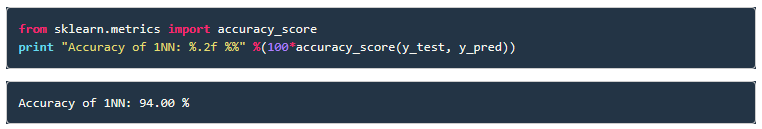
Sau đây, tôi trước hết xét trường hợp đơn giản K = 1, tức là với mỗi điểm test data, ta chỉ xét 1 điểm training data gần nhất và lấy label của điểm đó để dự đoán cho điểm test này.



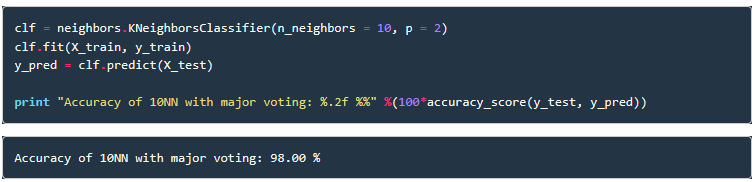
Kết quả cho thấy label dự đoán gần giống với label thật của test data, chỉ có 2 điểm trong số 20 điểm được hiển thị có kết quả sai lệch. Ở đây chúng ta làm quen với khái niệm mới: *ground truth*. Một cách đơn giản, *ground truth* chính là nhãn/label/đầu ra thực sự của các điểm trong test data.

**Phương pháp đánh giá (evaluation method)**

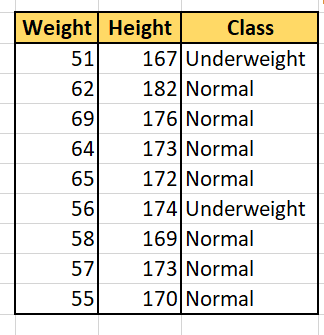
Để đánh giá độ chính xác của thuật toán KNN classifier này, chúng ta xem xem có bao nhiêu điểm trong test data được dự đoán đúng. Lấy số lượng này chia cho tổng số lượng trong tập test data sẽ ra độ chính xác. Scikit-learn cung cấp hàm số accuracy\_score để thực hiện công việc này.



Nhận thấy rằng chỉ xét 1 điểm gần nhất có thể dẫn đến kết quả sai nếu điểm đó là nhiễu. Một cách có thể làm tăng độ chính xác là tăng số lượng điểm lân cận lên, ví dụ 10 điểm, và xem xem trong 10 điểm gần nhất, class nào chiếm đa số thì dự đoán kết quả là class đó. Kỹ thuật dựa vào đa số này được gọi là major voting.

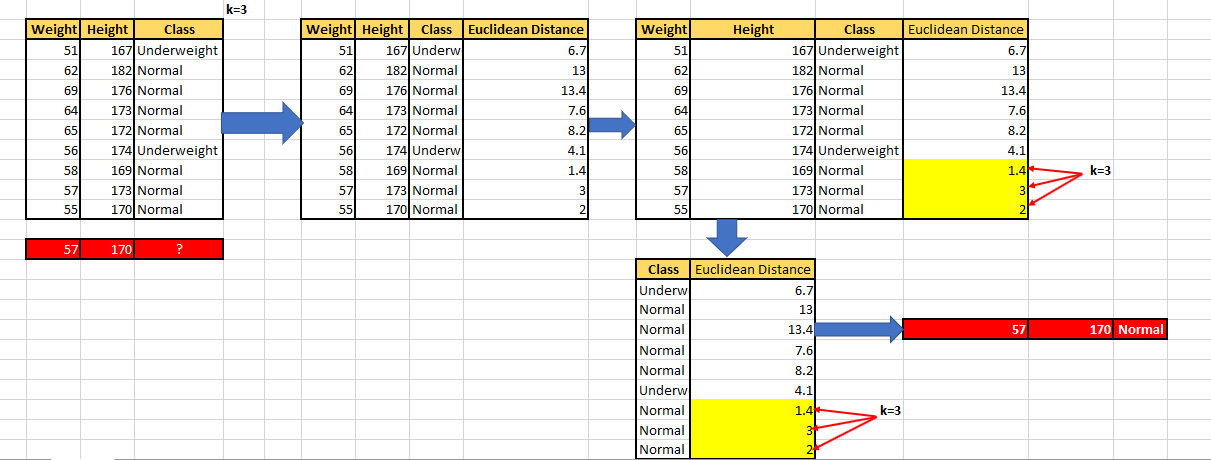


Ví dụ: Giả sử ta có dữ liệu cho trước gồm thuộc tính height (cm) & weight (kg) và mỗi dòng dữ liệu được phân lớp là Normal hoặc Underweight



Hình 2‑3 Dữ liệu mẫu cho thuật toán KNN

Thuật toán KNN được mô tả qua các bước sau



Hình 2‑4 Sơ đồ thực hiện thuật toán KNN

**Ưu điểm:**

1. Thuật toán đơn giản và dễ cài đặt
2. Không cần xây dựng model, điều chỉnh tham số hoặc thêm những giả thiết
3. Thuật toán đáp ứng được nhiều chức năng như có thể sử dụng cho phân lớp, hồi quy và tìm kiếm

**Khuyết điểm:**

Thuật toán sẽ chậm đáng kể khi số lượng dòng dữ liệu hoặc biến dự đoán/phụ thuộc tăng.

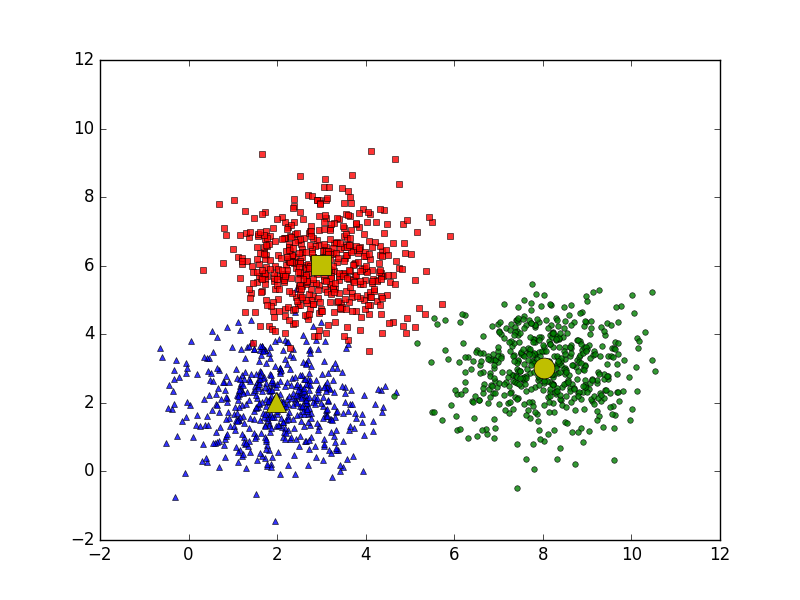
## **2.2 K-means Clustering**

**Giới thiệu**

Trong thuật toán K-means clustering [2], chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thể nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Ví dụ: Một công ty muốn tạo ra những chính sách ưu đãi cho những nhóm khách hàng khác nhau dựa trên sự tương tác giữa mỗi khách hàng với công ty đó (số năm là khách hàng; số tiền khách hàng đã chi trả cho công ty; độ tuổi; giới tính; thành phố; nghề nghiệp; …). Giả sử công ty đó có rất nhiều dữ liệu của rất nhiều khách hàng nhưng chưa có cách nào chia toàn bộ khách hàng đó thành một số nhóm/cụm khác nhau. Nếu một người biết Machine Learning được đặt câu hỏi này, phương pháp đầu tiên anh (chị) ta nghĩ đến sẽ là K-means Clustering. Vì nó là một trong những thuật toán đầu tiên mà anh ấy tìm được trong các cuốn sách, khóa học về Machine Learning. Và tôi cũng chắc rằng anh ấy đã đọc blog Machine Learning cơ bản. Sau khi đã phân ra được từng nhóm, nhân viên công ty đó có thể lựa chọn ra một vài khách hàng trong mỗi nhóm để quyết định xem mỗi nhóm tương ứng với nhóm khách hàng nào. Phần việc cuối cùng này cần sự can thiệp của con người, nhưng lượng công việc đã được rút gọn đi rất nhiều.

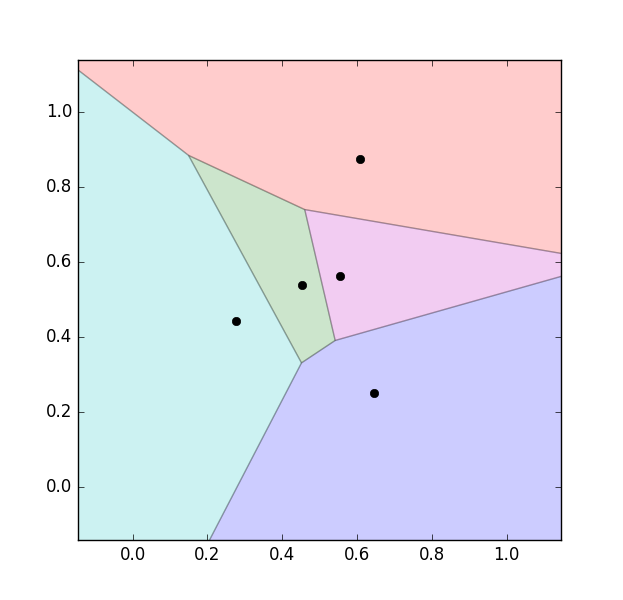
Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 3 cụm dữ liệu (từ giờ tôi sẽ viết gọn là cluster).



Hình 2‑5 Bài toán với 3 clusters

Giả sử mỗi cluster có một điểm đại diện (center) màu vàng. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó. Tới đây, chúng ta có một bài toán thú vị: Trên một vùng biển hình vuông lớn có ba đảo hình vuông, tam giác, và tròn màu vàng như hình trên. Một điểm trên biển được gọi là thuộc lãnh hải của một đảo nếu nó nằm gần đảo này hơn so với hai đảo kia . Hãy xác định ranh giới lãnh hải của các đảo.

Hình dưới đây là một hình minh họa cho việc phân chia lãnh hải nếu có 5 đảo khác nhau được biểu diễn bằng các hình tròn màu đen:



Hình 2‑6 Biểu đồ thể hiện K-means

.

Chúng ta thấy rằng đường phân định giữa các lãnh hải là các đường thẳng (chính xác hơn thì chúng là các đường trung trực của các cặp điểm gần nhau). Vì vậy, lãnh hải của một đảo sẽ là một hình đa giác.

Cách phân chia này trong toán học được gọi là Voronoi Diagram.

Trong không gian ba chiều, lấy ví dụ là các hành tinh, thì (tạm gọi là) lãnh không của mỗi hành tinh sẽ là một đa diện. Trong không gian nhiều chiều hơn, chúng ta sẽ có những thứ (mà tôi gọi là) siêu đa diện (hyperpolygon).

Quay lại với bài toán phân nhóm và cụ thể là thuật toán K-means clustering, chúng ta cần một chút phân tích toán học trước khi đi tới phần tóm tắt thuật toán ở phần dưới.

**Phân tích toán học**

Mục đích cuối cùng của thuật toán phân nhóm này là: từ dữ liệu đầu vào và số lượng nhóm chúng ta muốn tìm, hãy chỉ ra center của mỗi nhóm và phân các điểm dữ liệu vào các nhóm tương ứng. Giả sử thêm rằng mỗi điểm dữ liệu chỉ thuộc vào đúng một nhóm.

**Một số ký hiệu toán học**

Giả sử có N điểm dữ liệu là X=[x1,x2,…,xN] ∈Rd x N và *K* < N là số cluster chúng ta muốn phân chia. Chúng ta cần tìm các center m1, m2,…, mk ∈ Rd x N và label của mỗi điểm dữ liệu.

Với mỗi điểm dữ liệu xi đặt yi = [yi1,yi2,…,yik] là label vector của nó, trong đó nếu xi được phân vào cluster k thì yik = 1 và yịj = 0, ∀j ≠ kyij=0. Điều này có nghĩa là có đúng một phần tử của vector yi là bằng 1 (tương ứng với cluster của xi), các phần tử còn lại bằng 0. Ví dụ: nếu một điểm dữ liệu có label vector là [1,0,0,…,0] thì nó thuộc vào cluster 1, là [0,1,0,…,0] nó thuộc vào cluster 2,… Cách mã hóa label của dữ liệu như thế này được gọi là biểu diễn one-hot. Chúng ta sẽ thấy cách biểu diễn one-hot này rất phổ biến trong Machine Learning ở các bài tiếp theo.

Ràng buộc của yi có thể viết được dưới dạng toán học như sau:

Yik∈{0,1},   (1)

Hàm mất mát và bài toán tối ưu

Nếu ta coi center mk là center (hoặc representative) của mỗi cluster và ước lượng tất cả các điểm được phân vào cluster này bởi mk, thì một điểm dữ liệu xi được phân vào cluster k sẽ bị sai số là . Chúng ta mong muốn sai số này có trị tuyệt đối nhỏ nhất nên (giống như trong phần Linear Regression) ta sẽ tìm cách để đại lượng sau đây đạt giá trị nhỏ nhất:

Hơn nữa, vì xi được phân vào cluster k nên yik = 1, yij = 0, . Khi đó, biểu thức bên trên sẽ được viết lại là:

=

Sai số cho toàn bộ dữ liệu là:

Trong đó Y= [, M = [] lần lượt là các ma trận được tạo bởi label vector của mỗi điểm dữ liệu và center của mỗi cluster. Hàm số mất mát trong bài toán K-means clustering của chúng ta là hàm với ràng buộc như được nêu trong phương trình (1).

Tóm lại, chúng ta cần tối ưu bài toán sau:

Y, M = (2)

Subject to: ;

(subject có nghĩa là thỏa mãn điều kiện)

Nhắc lại khái niệm argmin: Chúng ta biết ký hiệu min là giá trị nhỏ nhất của hàm số, argmin chính là giá trị của biến số để hàm số đó đạt giá trị nhỏ nhất đó. Nếu f(x) = x2−2x+1 = (x−1)2thì giá trị nhỏ nhất của hàm số này bằng 0, đạt được khi x = 1. Trong ví dụ này minx f(x) = 0 và argminx f(x) = 1.

Thêm ví dụ khác, nếu x1 = 0, x2 = 10, x3 = 5 thì ta nói argminixi = 1 vì 1 là chỉ số để xi đạt giá trị nhỏ nhất (bằng 0). Biến số viết bên dưới min là biến số chúng ta cần tối ưu. Trong các bài toán tối ưu, ta thường quan tâm tới argmin hơn là min.

**Thuật toán tối ưu hàm mất mát**

Bài toán (2) là một bài toán khó tìm điểm tối ưu vì nó có thêm các điều kiện ràng buộc. Bài toán này thuộc loại mix-integer programming (điều kiện biến là số nguyên) – là loại rất khó tìm nghiệm tối ưu toàn cục (global optimal point, tức nghiệm làm cho hàm mất mát đạt giá trị nhỏ nhất có thể). Tuy nhiên, trong một số trường hợp chúng ta vẫn có thể tìm được phương pháp để tìm được nghiệm gần đúng hoặc điểm cực tiểu. (Nếu chúng ta vẫn nhớ chương trình toán ôn thi đại học thì điểm cực tiểu chưa chắc đã phải là điểm làm cho hàm số đạt giá trị nhỏ nhất).

Một cách đơn giản để giải bài toán (2) là xen kẽ giải Y và M khi biến còn lại được cố định. Đây là một thuật toán lặp, cũng là kỹ thuật phổ biến khi giải bài toán tối ưu. Chúng ta sẽ lần lượt giải quyết hai bài toán sau đây:

Cố định M, tìm Y

**Giả sử đã tìm được các centers, hãy tìm các label vector để hàm mất mát đạt giá trị nhỏ nhất.** Điều này tương đương với việc tìm cluster cho mỗi điểm dữ liệu.

Khi các centers là cố định, bài toán tìm label vector cho toàn bộ dữ liệu có thể được chia nhỏ thành bài toán tìm label vector cho từng điểm dữ liệu xi như sau:

(3)

Subject to: ;

Vì chỉ có một phần tử của label vector yi bằng 1 nên bài toán (3) có thể tiếp tục được viết dưới dạng đơn giản hơn:

j =

Vì chính là bình phương khoảng cách tính từ điểm xi tới center mj, ta có thể kết luận rằng mỗi điểm xi thuộc vào cluster có center gần nó nhất! Từ đó ta có thể dễ dàng suy ra label vector của từng điểm cố định.

**Cố định Y, tìm M**

**Giả sử đã tìm được cluster cho từng điểm, hãy tìm center mới cho mỗi cluster để hàm mất mát đạt giá trị nhỏ nhất.**

Một khi chúng ta đã xác định được label vector cho từng điểm dữ liệu, bài toán tìm center cho mỗi cluster được rút gọn thành:

Tới đây, ta có thể tìm nghiệm bằng phương pháp giải đạo hàm bằng 0, vì hàm cần tối ưu là một hàm liên tục và có đạo hàm xác định tại mọi điểm. Và quan trọng hơn, hàm này là hàm convex (lồi) theo mj nên chúng ta sẽ tìm được giá trị nhỏ nhất và điểm tối ưu tương ứng.

Đặt l(mj) là hàm bên trong dấu arg min, ta có đạo hàm:

= 2

Giải phương trình đạo hàm bằng 0 ta có:

mj =

Nếu để ý một chút, chúng ta sẽ thấy rằng mẫu số chính là phép đếm số lượng các điểm dữ liệu trong cluster j. Còn tử số chính là tổng các điểm dữ liệu trong cluster j.

Hay nói một cách đơn giản: mj  là trung bình cộng của các điểm trong cluster j.

Tên gọi K-means clustering cũng xuất phát từ đây.

Tóm tắt thuật toán

Đầu vào: Dữ liệu X và số lượng cluster cần tìm K.

Đầu ra: Các center M và label vector cho từng điểm dữ liệu Y

1. Chọn K điểm bất kỳ làm các center ban đầu.
2. Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
3. Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì ta dừng thuật toán.
4. Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
5. Quay lại bước 2.

Chúng ta có thể đảm bảo rằng thuật toán sẽ dừng lại sau một số hữu hạn vòng lặp. Thật vậy, vì hàm mất mát là một số dương và sau mỗi bước 2 hoặc 3, giá trị của hàm mất mát bị giảm đi. Theo kiến thức về dãy số trong chương trình cấp 3: nếu một dãy số giảm và bị chặn dưới thì nó hội tụ! Hơn nữa, số lượng cách phân nhóm cho toàn bộ dữ liệu là hữu hạn nên đến một lúc nào đó, hàm mất mát sẽ không thể thay đổi, và chúng ta có thể dừng thuật toán tại đây.

**Ví dụ trên Python**

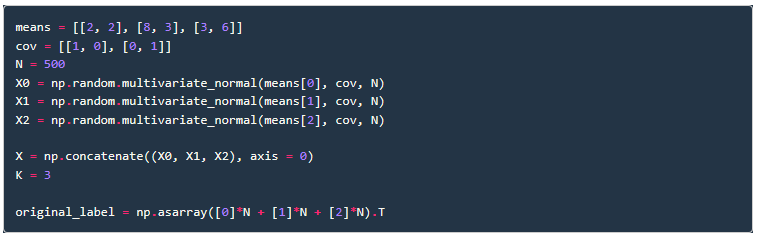
**Giới thiệu bài toán**

Để kiểm tra mức độ hiểu quả của một thuật toán, chúng ta sẽ làm một ví dụ đơn giản. Trước hết, chúng ta chọn center cho từng cluster và tạo dữ liệu cho từng cluster bằng cách lấy mẫu theo phân phối chuẩn có kỳ vọng là center của cluster đó và ma trận hiệp phương sai (covariance matrix) là ma trận đơn vị.

Trước tiên, chúng ta cần khai báo các thư viện cần dùng. Chúng ta cần numpy và matplotlib cho việc tính toán ma trận và hiển thị dữ liệu. Chúng ta cũng cần thêm thư viện scipy.spatial.distance để tính khoảng cách giữa các cặp điểm trong hai tập hợp một cách hiệu quả.

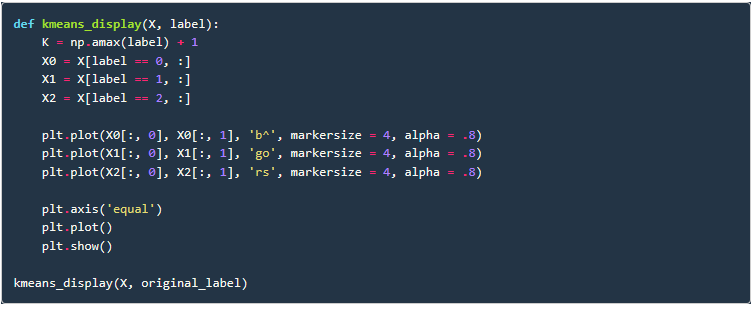


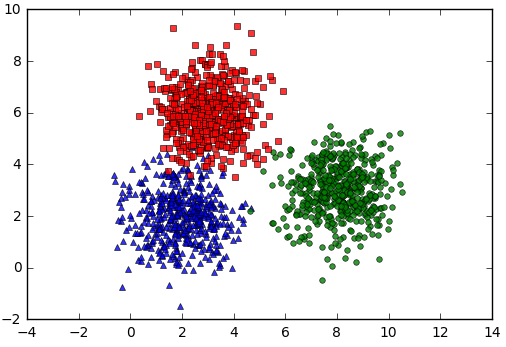
Tiếp theo, ta tạo dữ liệu bằng cách lấy các điểm theo phân phối chuẩn có kỳ vọng tại các điểm có tọa độ (2, 2), (8, 3) và (3, 6), ma trận hiệp phương sai giống nhau và là ma trận đơn vị. Mỗi cluster có 500 điểm. (Chú ý rằng mỗi điểm dữ liệu là một hàng của ma trận dữ liệu.



**Hiển thị dữ liệu trên đồ thị**

Chúng ta cần một hàm kmeans\_display để hiển thị dữ liệu. Sau đó hiển thị dữ liệu theo nhãn ban đầu.





Hình 2‑7 Kết quả thuật toán K-means

Trong đồ thị trên, mỗi cluster tương ứng với một màu. Có thể nhận thấy rằng có một vài điểm màu đỏ bị lẫn sang phần cluster màu xanh.

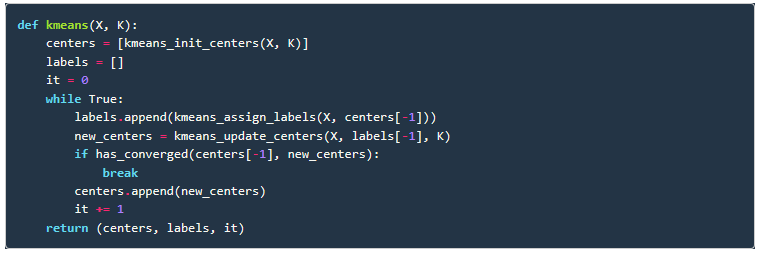
**Các hàm số cần thiết cho K-means clustering**

Viết các hàm:

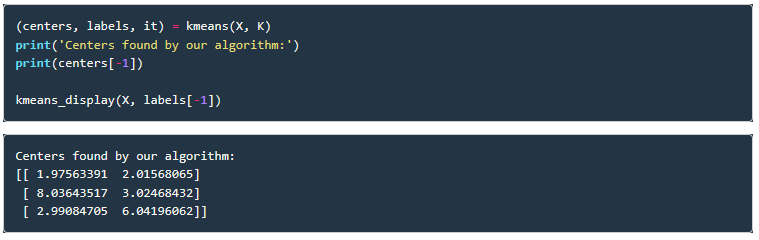
1. kmeans\_init\_centers để khởi tạo các centers ban đầu.
2. kmeans\_asign\_labels để gán nhán mới cho các điểm khi biết các centers.
3. kmeans\_update\_centers để cập nhật các centers mới dữa trên dữ liệu vừa được gán nhãn.
4. has\_converged để kiểm tra điều kiện dừng của thuật toán.

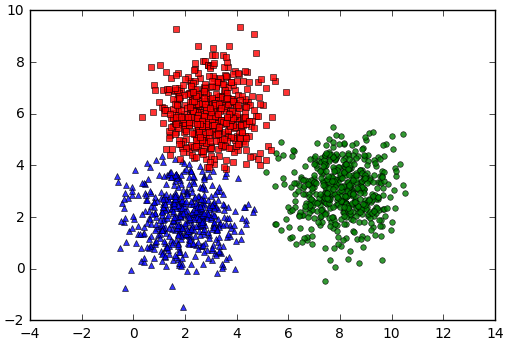


Phần chính của K-means clustering:



Áp dụng thuật toán vừa viết vào dữ liệu ban đầu, hiển thị kết quả cuối cùng.



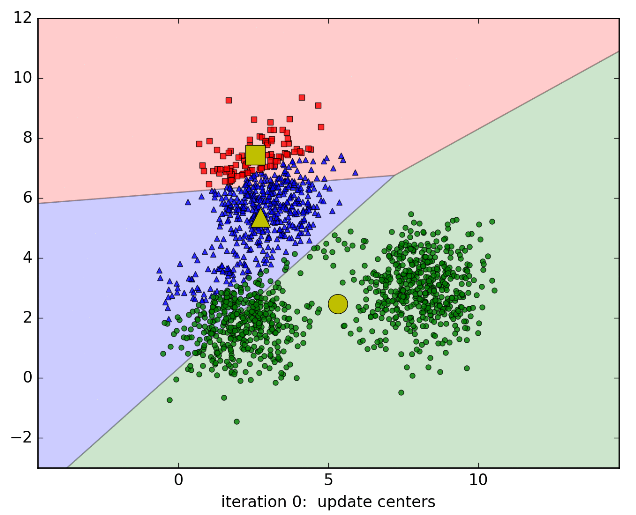
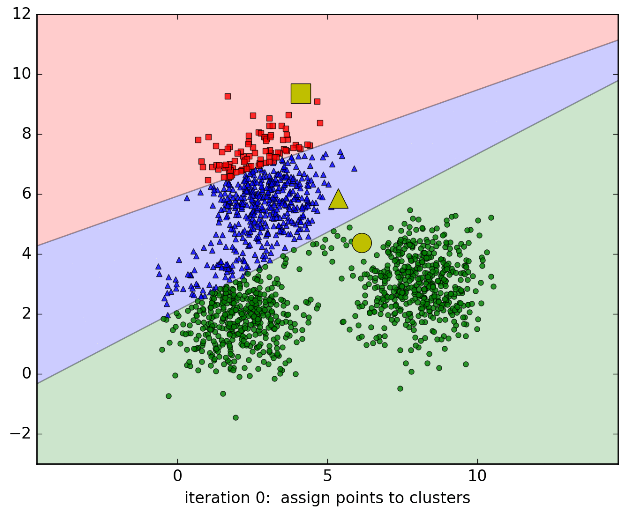


Hình 2‑8 Kết quả thuật toán K-means

Từ kết quả này chúng ta thấy rằng thuật toán K-means clustering làm việc khá thành công, các centers tìm được khá gần với kỳ vọng ban đầu. Các điểm thuộc cùng một cluster hầu như được phân vào cùng một cluster (trừ một số diểm màu đỏ ban đầu đã bị phân nhầm vào cluster màu xanh da trời, nhưng tỉ lệ là nhỏ và có thể chấp nhận được).

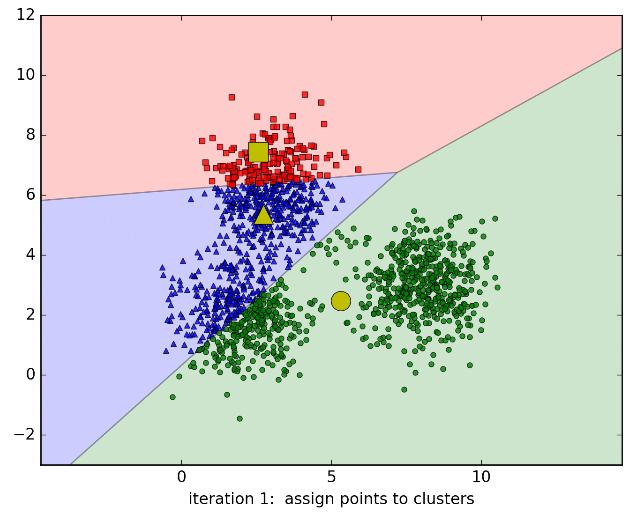
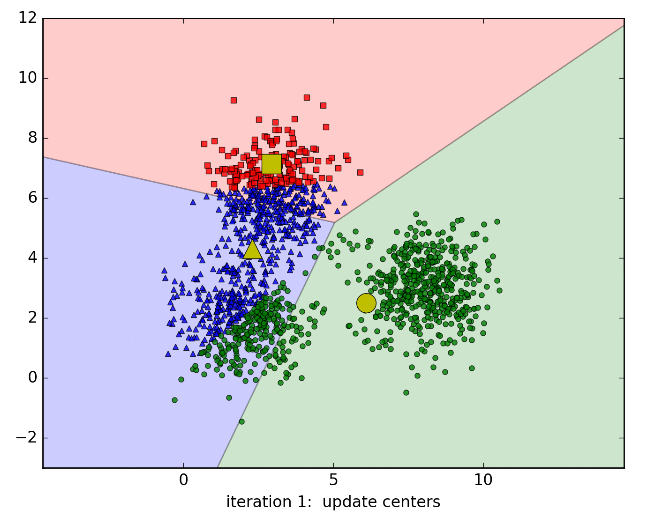
Dưới đây là hình ảnh động minh họa thuật toán qua từng vòng lặp, chúng ta thấy rằng thuật toán trên hội tụ rất nhanh, chỉ cần 6 vòng lặp để có được kết quả cuối cùng:

Lần lặp thứ 0:



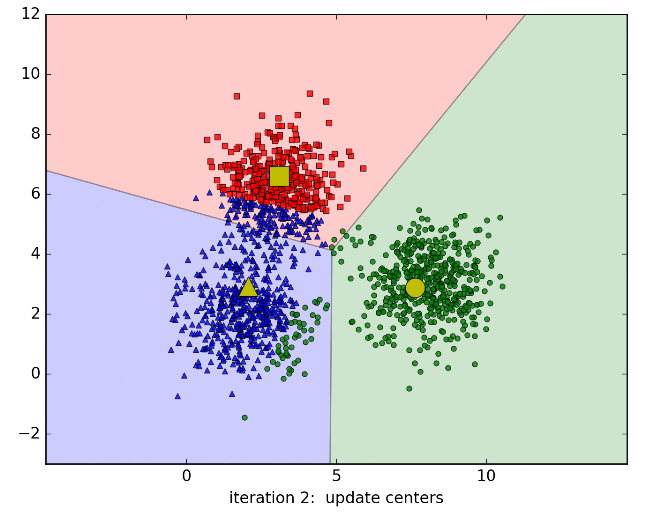
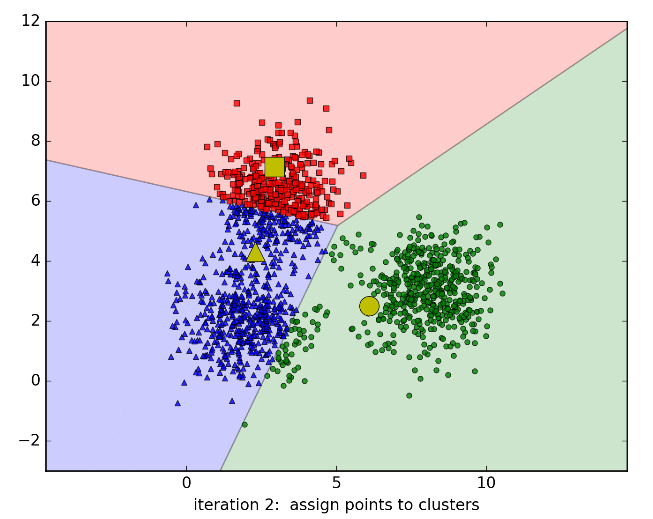
Hình 2‑9 Lần lặp thứ 0 của K-means

Lần lặp thứ 1:



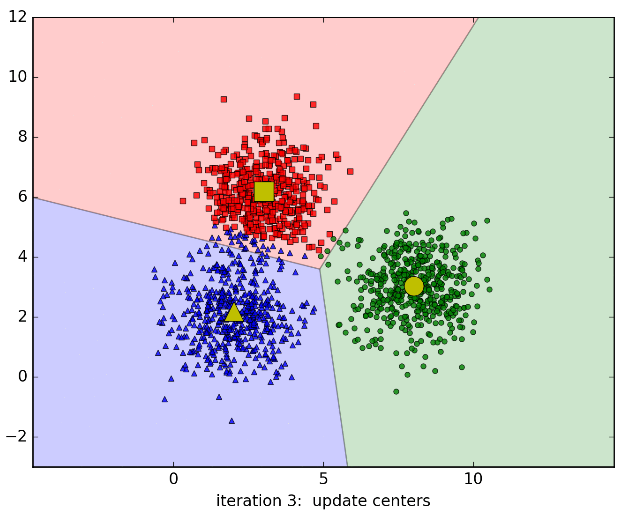
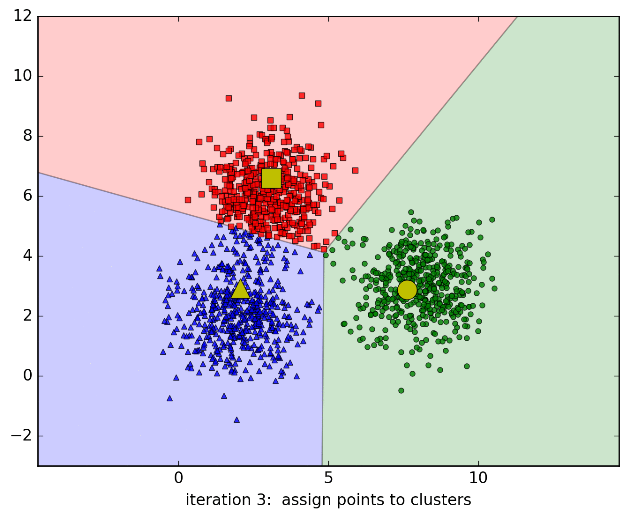
Hình 2‑10 Lần lặp thứ 1 của K-means

Lần lặp thứ 2:



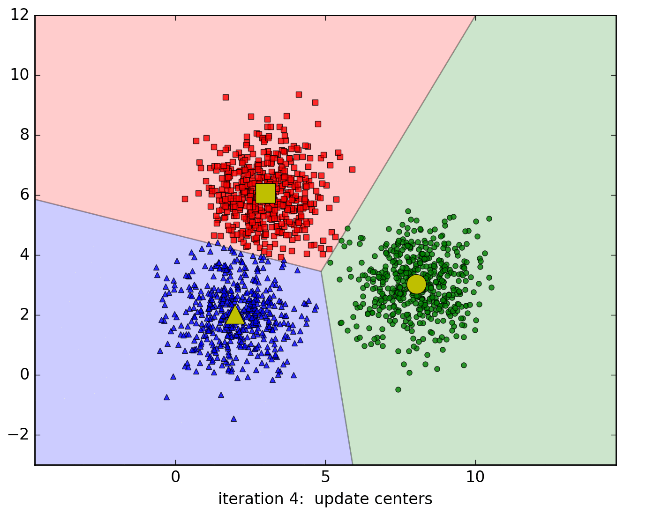
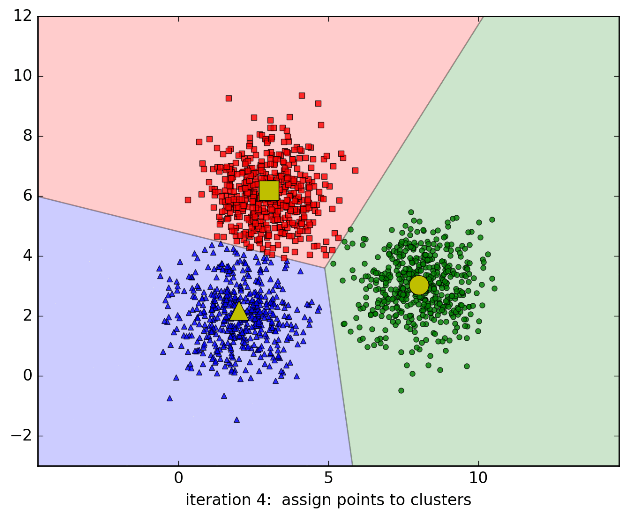
Hình 2‑11 Lần lặp thứ 2 của K-means

Lần lặp thứ 3:



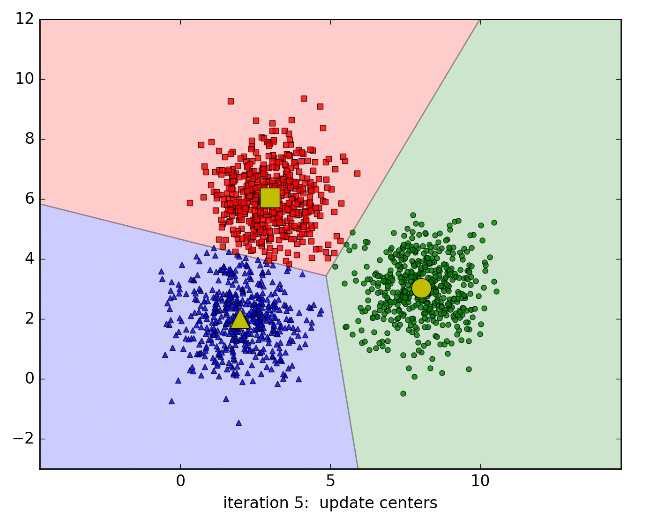
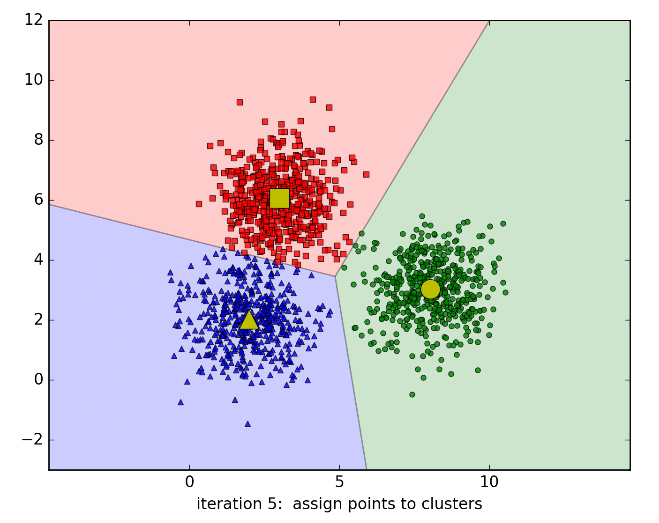
Hình 2‑12 Lần lặp thứ 3 của K-means

Lần lặp thứ 4:



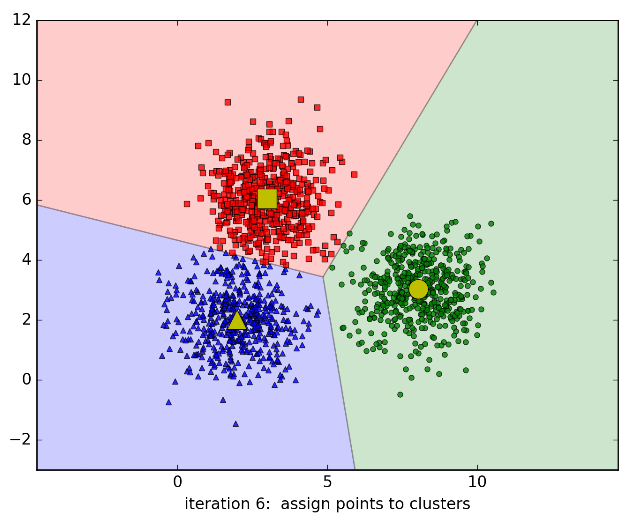
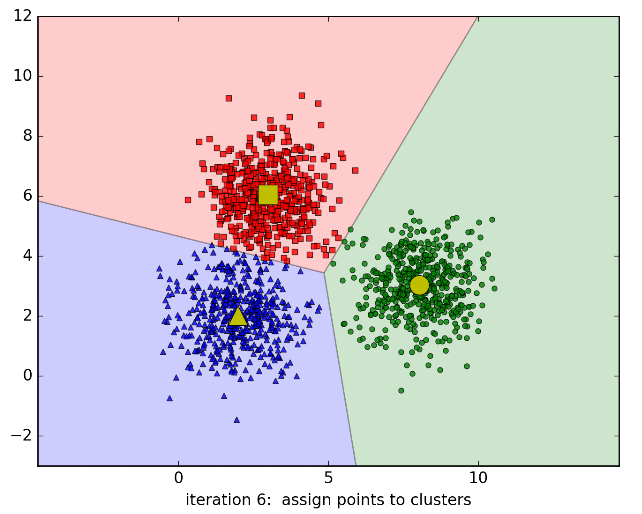
Hình 2‑13 Lần lặp thứ 4 của K-means

Lần lặp thứ 5:



Hình 2‑14 Lần lặp thứ 5 của K-means

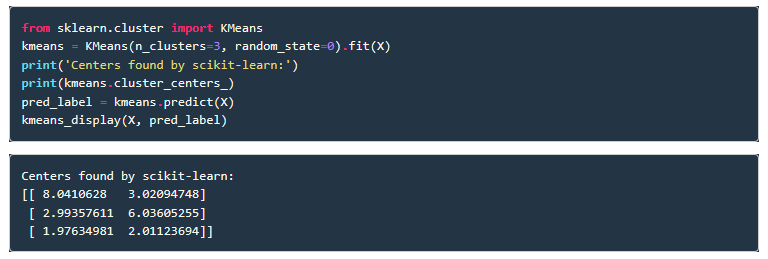
Lần lặp thứ 6:

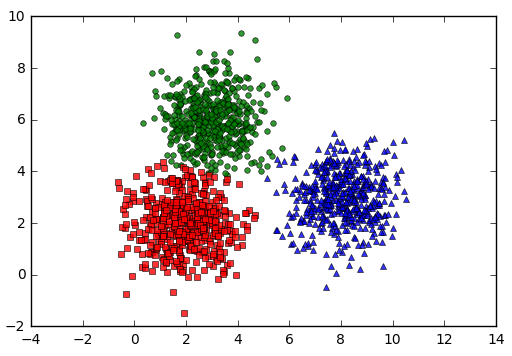


Hình 2‑15 Lần lặp thứ 6 của K-means

Kết quả tìm được bằng thư viện scikit-learn

Để kiểm tra thêm, chúng ta hãy so sánh kết quả trên với kết quả thu được bằng cách sử dụng thư viện scikit-learn.





Hình 2‑16 Kết quả sau khi thực hiện K-means của Scikit-learn

**Hạn chế**

Có một vài hạn chế của thuật toán K-means clustering:

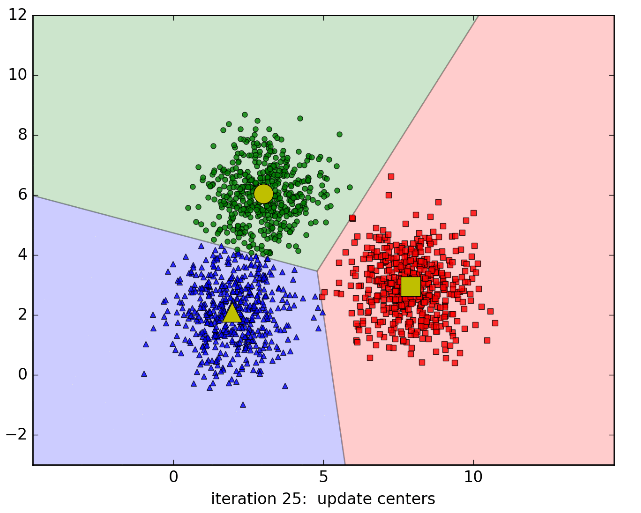
* **Chúng ta cần biết số lượng cluster cần clustering**

Để ý thấy rằng trong thuật toán nêu trên, chúng ta cần biết đại lượng KK là số lượng clusters. Trong thực tế, nhiều trường hợp chúng ta không xác định được giá trị này.

* **Nghiệm cuối cùng phụ thuộc vào các centers được khởi tạo ban đầu**

Tùy vào các center ban đầu mà thuật toán có thể có tốc độ hội tụ rất chậm, ví dụ:

* Phải đạt tới lần lặp thứ 25



Hình 2‑17 Kết quả thuật toán K-means sau khi đạt tới lần lặp thứ 25

## **2.3 Thuật toán Apriori khai phá luật kết hợp**

### **2.3.1 Luật kết hợp trong khai phá dữ liệu (****Association Rule in Data Mining)**

Trong lĩnh vực Data Mining, mục đích của luật kết hợp (Association Rule - AR) là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu. Nội dung cơ bản của luật kết hợp được tóm tắt như dưới đây [3].

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

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

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

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

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

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

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

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

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

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

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

Trong đó: N là tổng số giao dịch.

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

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

Trong đó: n(X) là số giao dịch chứa X

**Để thu được các luật kết hợp, ta thường áp dụng 2 tiêu chí:** min\_sup và  min\_conf

Các luật thỏa mãn có support và confidence thỏa mãn (lớn hơn hoặc bằng)  cả min\_sup và min\_conf gọi là các luật mạnh (Strong Rule)

min\_sup và min\_conf gọi là các giá trị ngưỡng (threshold) và phải xác định trước khi sinh các luật kết hợp.

**Một itemsets mà tần suất xuất hiện của nó >= min\_sup goi là frequent itemsets**

**Một số loại luật kết hợp**

Binary association rules (luật kết hợp nhị phân): Apple => Banana

Quantitative association rules (luật kết hợp định lượng):

weight in [70kg – 90kg] => height in [170cm – 190cm]

Fuzzy association rules (Luật kết hợp mờ):  weight in HEAVY => height in TALL

Thuật toán phổ biến nhất tìm các luật kết hợp là Apriori sử dụng Binary association rules.

### **2.3.2 Thuật toán sinh các luật kết hợp Apriori** *(by Agrawal and Srikant 1994)*

Tư tưởng chính của thuật toán Apriori là:

- Tìm tất cả frequent itemsets:

k-itemset (itemsets gồm k items) được dùng để tìm (k+1)- itemset.

Đầu tiên tìm 1-itemset (ký hiệu L1). L1 được dùng để tìm L2 (2-itemsets). L2 được dùng để tìm L3 (3-itemset) và tiếp tục cho đến khi không có k-itemset được tìm thấy.

- Từ frequent itemsets sinh ra các luật kết hợp mạnh (các luật kết hợp thỏa mãn 2 tham số min\_sup và min\_conf)

**Apriori Algorithm**

1. Duyệt (Scan)  toàn bộ transaction database để có được support S của 1-itemset, so sánh S với min\_sup, để có được 1-itemset (L1)

2. Sử dụng Lk-1 nối (join) Lk-1  để sinh ra candidate k-itemset. Loại bỏ các itemsets không phải là frequent itemsets thu được k-itemset

3. Scan transaction database để có được support của mỗi candidate k-itemset, so sánh S với min\_sup để thu được frequent k –itemset (Lk)

4. Lặp lại từ bước 2 cho đến khi Candidate set (C) trống (không tìm thấy frequent itemsets)

5. Với mỗi frequent itemset I, sinh tất cả các tập con s không rỗng của I

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

Chẳng hạn với I= {A1,A2,A5}, các tập con của I:

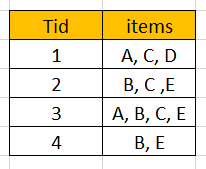
{A1}, {A2}, {A5}, {A1,A2},{A1,A5},{A2,A5}

sẽ có các luật sau

{A1} => {A2,A5}, {A2} =>{A1,A5}, {A5} =>{A1,A2}

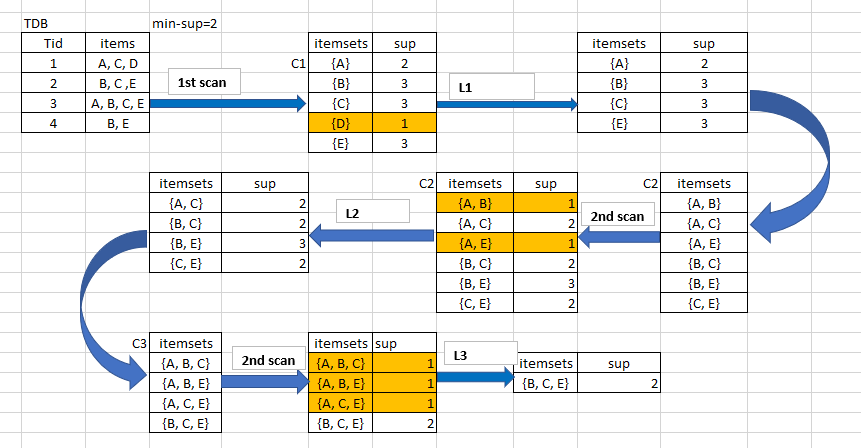
{A1,A2} =>{A5}, {A1,A5} =>{A2}, {A2,A5} => {A1}

Ví dụ: Giả sử ta có có sở dữ liệu giao dịch (Transaction Database -TDB) như sau :



Hình 2‑18 Cơ sở dữ liệu giao dịch

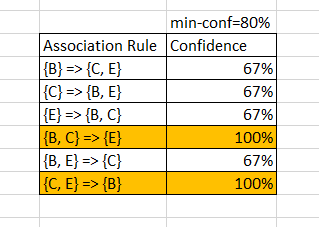
Thuật toán Apriori khai phá luật kết hợp được mô tả qua các bước sau



Hình 2‑19 Mô tả các bước chạy thuật toán Apriori

Ta có frequent itemsets I ={B,C,E}, với min\_conf =80% ta có 2 luật kết hợp là

{B,C} => {E} và {C,E} => {B}



Hình 2‑20 Kết quả chạy thuật toán Apriori

## **2.4 Thuật toán FP-Growth**

FP-Growth biểu diễn dữ liệu các giao dịch bằng một cấu trúc dữ liệu gọi là FP-Tree.

FP-Growth sử dụng FP-Tree để xác định trực tiếp các tập hạng mục phổ biến (không sinh các tập hạng mục ứng viên từ các tập hạng mục ứng viên trước) [4].

Khi một FP-Tree đã được xây dựng, FP-Growth sử dụng cách tiếp cận chia để trị đệ quy để khai thác các tập phổ biến.

Với mỗi giao dịch, FP-Tree xây dựng một đường đi (path) trong cây.

Hai giao dịch có chứa cùng một số các mục, thì đường đi của chúng sẽ có phần (đoạn) chung.

Càng nhiều các đường đi có phần tử chung, thì việc biểu diễn bằng FP-Tree sẽ càng gọn (compressed/compacted).

**Xây dựng FP-Tree**

Ban đầu, FP-Tree chỉ chứa duy nhất nút gốc (được biểu diễn bởi ký hiệu null).

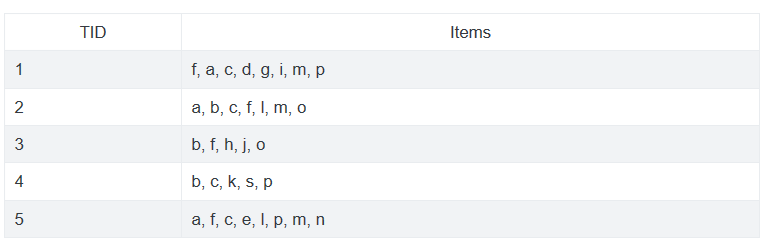
Cơ sở dữ liệu các giao dịch được duyệt lần 1, để xác định độ hỗ trợ của mỗi mục.

* Các mục không thường xuyên (infrequent items) bị loại bỏ.
* Các mục thường xuyên (frequent items) được sắp xếp theo thứ tự giảm dần về độ hỗ trợ.

Cơ sở dữ liệu các giao dịch được duyệt lần thứ 2, để xây dựng FP-Tree.

Để hiểu rõ hơn thuật toán, xem lời giải dưới đây.

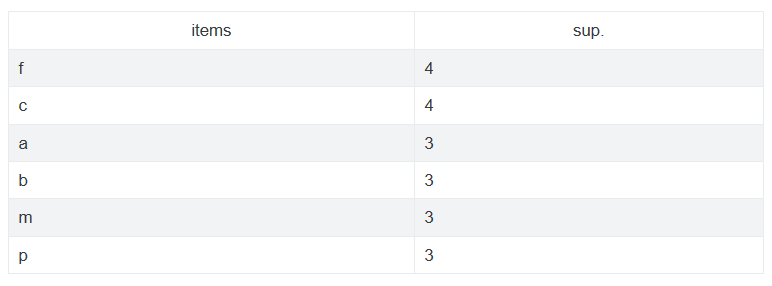
Cho bảng dữ liệu bao gồm các giao dịch (tid) sau:



Hình 2‑21 Bảng dữ liệu

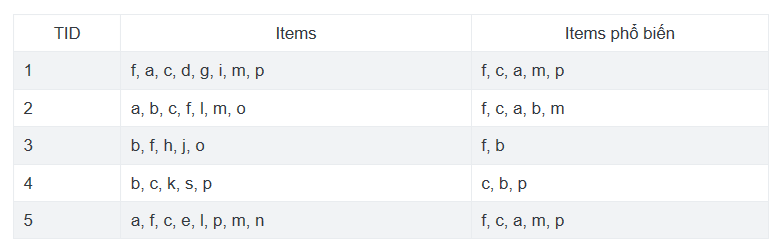
Tìm các tập mục có độ hỗ trợ ≥ 0.6 (tức tần số sup. ≥ 3).

Đầu tiên tìm các item mức 1 có sup. ≥ 3, và sắp xếp theo thứ tự giảm dần:



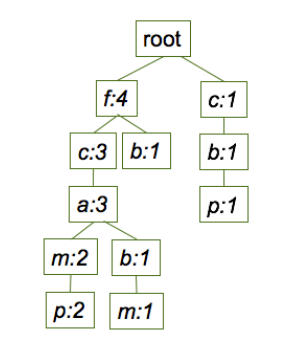
Hình 2‑22 Các tập mục trong thuật toán FP-Growth

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



Hình 2‑23 Mục phổ biến mức 1 FP-Growth

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



Hình 2‑24 FP-Tree được xây dựng sau khi duyệt các items phổ biến của các giao dịch

Tiếp theo, duyệt các item phổ biến mức 1 theo thứ tự tăng dần độ hỗ trợ là p, m, b, a, c, f. Với mỗi item, xây dựng các **cơ sở mẫu điều kiện** *(conditional pattern-base)* và sau đó là các **FP-Tree điều kiện** *(conditional FP-Tree)* của nó.

**Tính chất**: bất kì mẫu phổ biến nào có chứa mục Ii đều được chứa trên các nhánh (đường dẫn) của cây FP-Tree chứa Ii, số lần xuất hiện của mẫu chứa các nút trong đường dẫn tiền tố bằng số lần xuất hiện của nút Ii.

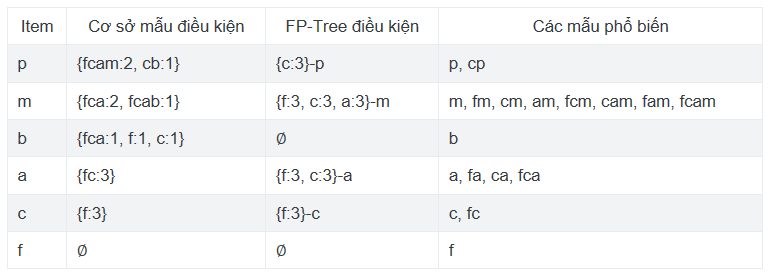
Bắt đầu với item p, cơ sở mẫu điều kiện của nó là tất cả các đường dẫn tiền tố của cây FP-Tree khi duyệt từ gốc root = null đến nút p, chính là fcam:2 và cb:1 (số theo sau là số lần xuất hiện của nút p tương ứng với mỗi tiền tố đó).

Tiếp theo ta xây dựng FP-Tree điều kiện từ mẫu này bằng cách trộn tất cả các đường dẫn và giữ lại các nút có tổng các số đếm ≥ sup. = 3:

fcam:2 và cb:1 trộn lại thành f:2, c:3, a:2, m:2, b:1; chỉ có c:3 là thoả mãn điều kiện.

Do đó, các mẫu phổ biến chứa p là: p, cp.

Làm tương tự cho các item còn lại, ta sẽ tìm được các mẫu phổ biến chứa các item đó. Cuối cùng có bảng sau đây:



Hình 2‑25 Mẫu phổ biến

## **2.5 Thuật toán dựa trên hệ số hóa ma trận (Matrix Factorization-based algorithms)**

Trong phạm vi đề tài, nhóm tác giả sẽ tập trung vào thuật toán lọc cộng tác (collaborative filtering algorithm) dựa trên hệ số hóa ma trận không âm (Non-negative Matrix Factorization) [5].

Thuật toán này rất tương tự với SVD (Singular Value Decomposition - Phân tích giá trị đơn).

Dự đoán được tính theo công thức:

,

Trong đó:

Các hệ số user và item được đặt là dương.

ước tính sự xếp hạng của user cho item .

ma trận chuyển vị mảng numpy các hệ số i có kích thước (n\_items, n\_factors).

mảng numpy các phần tử các hệ số u có kích thước (n\_users, n\_factors).

Mỗi một bước của phương thức SGD, hệ số hoặc user và item được cập nhật như sau:

Trong đó:

* λu và λi chính là các tham số chính quy.
* : mảng numpy kích thước (n\_users, n\_factors) tương ứng với hệ số f.
* : mảng numpy kích thước (n\_items, n\_factors) tương ứng với hệ số f.
* : đánh giá true của user cho item .
* : tập hợp tất cả các items được đánh giá bởi user .
* : tập hợp tất cả các user đánh giá item .
* n\_users: số các phần tử user.
* n\_factors: số các phần tử factors. Mặc định là 15.

Thuật toán này phụ thuộc cao vào giá trị khởi tạo. Các hệ số user và item được khởi tạo một cách thống nhất giữa init\_low và init\_high. Nếu thay đổi các hệ số đó theo ý của chúng ta sẽ tạo ra các rủi ro.

* init\_low: vùng dưới cho khởi tạo ngẫu nhiên các hệ số. Phải lớn hơn 0 để đảm bảo các hệ số không âm. Mặc định là 0
* init\_high: Vùng trên cho khởi tạo ngẫu nhiên các hệ số. Mặc định là 1.

Trong trường hợp này, độ dự đoán được tính bởi công thức:

,

Trong đó:

* : giá trị trung bình của tất cả các đánh giá.
* : mảng numpy các phần tử chịu ảnh hưởng u có kích thước (n\_users).
* : mảng numpy các phần tử chịu ảnh hưởng I có kích thước (n\_items).
* , : tương tự như trên.

## **2.6 Thuật toán Slope One**

Thuật toán lọc cộng tác đơn giản nhưng rất chính xác [6].

Đây là một triển khai đơn giản của thuật toán Slopeone

Dự đoán được tính bởi công thức:

,

Trong đó Ri(u) là tập hợp các phần tử items liên quan với nhau, là giá trị trung bình của tất cả các đánh giá được đưa ra bởi user . Ví dụ, tập hợp các items được đánh giá bởi có ít nhất một phần tử phổ biến user với . dev(i,j) được tính bằng độ lệch giữa các đánh giá và của :

* : tập hợp tất cả các users đã đánh giá item và .
* : đánh giá true của user cho item .
* : đánh giá true của user cho item .

## **2.7 Môi trường Python, NodeJs**

**Giới thiệu về Python** [7]

Python là một ngôn ngữ lập trình phổ biến. Được tạo ra bởi Guido van Rossum vào năm 1991.

Ngày nay, Python được sử dụng trong các mục đích như:

* Phát triển web (trên máy chủ).
* Phát triển phần mềm.
* Tính toán, khoa học.
* Lên kịch bản cho hệ thống (system scripting).

**Python Làm Được Những Gì?**

* Python khi sử dụng trên máy chủ có thể tạo ra các ứng dụng nền web (web application).
* Python có thể chạy song song cùng các phần mềm khác để dễ phân luồng công việc.
* Python có thể kết nối dễ dàng đến cơ sở dữ liệu, hay cả việc đọc và ghi file.
* Với Python, việc xử lý Big Data và các phép toán phức tạp trở nên dễ dàng.
* Dễ dàng sử dụng Python để tạo ra các sản phẩm demo một cách nhanh chóng. Hơn nữa, các công ty, hoặc lập trình viên chuyên nghiệp luôn ưu tiên sử dụng Python cho việc phát triển những sản phẩm chất lượng.

Hiện nay, với khả năng xử lý các phép toán phức tạp của mình, Python đang được sử dụng nhiều trong việc phát triển Trí Tuệ Nhân Tạo và các nghiên cứu trong lĩnh vực Machine Learning.

**Giới thiệu về Node.js** [8]

NodeJS là một nền tảng Server side được xây dựng dựa trên Javascript Engine (V8 Engine). Node.js được phát triển bởi Ryan Dahl năm 2009 và phiên bản cuối cùng là v0.10.36. Định nghĩa NodeJs bởi tài liệu chính thức như sau:

Node.js là một nền tảng dựa vào Chrome Javascript runtime để xây dựng các ứng dụng nhanh, có độ lớn. Node.js sử dụng các phần phát sinh các sự kiện (event-driven), mô hình non-blocking I/O để tạo ra các ứng dụng nhẹ và hiệu quả cho các ứng dụng về dữ liệu thời gian thực chạy trên các thiết bị phân tán.

NodeJs là một mã nguồn mở, đa nền tảng cho phát triển các ứng dụng phía Server và các ứng dụng liên quan đến mạng. Ứng dụng Node.js được viết bằng Javascript và có thể chạy trong môi trường Node.js trên hệ điều hành Window, Linux...

Node.js cũng cung cấp cho chúng ta các module Javascript đa dạng, có thể đơn giản hóa sự phát triển của các ứng dụng web sử dụng Node.js với các phần mở rộng.

Node.js = Môi trường Runtime + các thư viện Javascript

## **2.8 Các thư viện hỗ trợ Python**

### **2.8.1 Tổng quan thư viện trong ứng dụng**

Nhóm sử dụng ngôn ngữ Python (như đã được giới thiệu ở mục môi trường Python) để xây dựng phía backend. Trong đó, Python hỗ trợ rất nhiều bao gồm các thư viện thuật toán (sklearn), Surprise, framework Flask-RESTplus,… là một môi trường lý tưởng để thực hiện đề tài khóa luận này.

### **2.8.2 Thư viện Flask-RESTPlus**

Flask-RESTPlus [9]là phần mở rộng của Flask để hỗ trợ xây dựng REST APIs một cách nhanh chóng. Flask-RESTPlus khuyến khích thực hành tốt với sự thiết lập tối thiểu. Nó cung cấp một tập hợp rõ ràng của công cụ để diễn tả API và thể hiện tài liệu của nó một cách chính xác [10].

### **2.8.3 Mô hình cấu trúc chức năng (framework functional structure )**

**Blueprints**

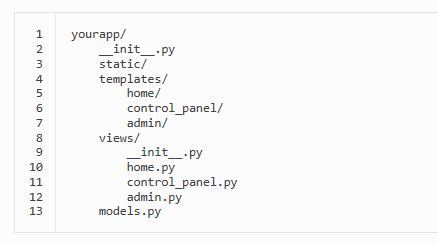
Blueprint định nghĩa tập hợp các views, templates, static files và các thành phần có thể được sử dụng trong ứng dụng. Ví dụ, tưởng tượng chúng ta có một bản thiết kế cho một admin panel. Bản thiết kế (blueprint) sẽ định nghĩa cho các routes chẳng hạn /admin/login và /admin/dashboard. Nó có thể bao gồm cả các templates và static files phục vụ cho các routes đó [11].

**Vì sao sử dụng Blueprints ?**

Sử dụng bluesprints để tổ chức ứng dụng thành các thành phần riêng biệt. Ví dụ cho Twitte-like microblog, blueprint có trong các pages như index.html và about.html. Sau đó chúng ta có logged-in dashboard nơi hiển thị tất cả các bài đăng mới nhất và administator’s panel. Mỗi một khu vực của site được phân chia như trong code. Nó làm cho cấu trúc của ứng dụng trở nên nhỏ hơn để làm.

**Blueprint được sử dụng trong functional structure**

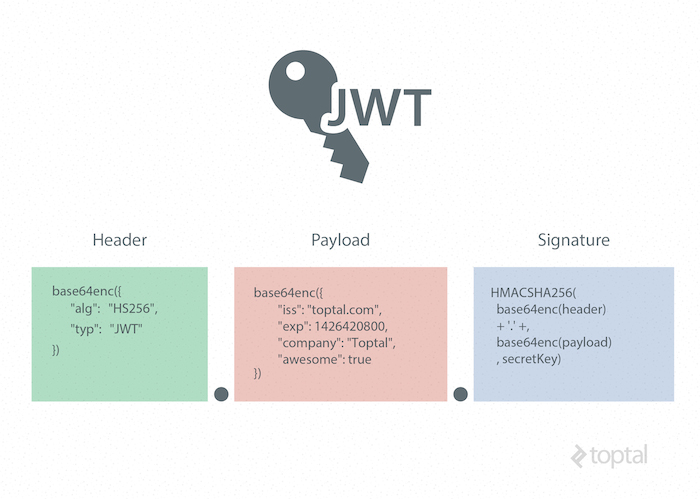
Với việc xây dựng app theo cấu trúc functional structure, chúng ta sẽ tổ chức các thành phần của app theo chức năng của nó. Theo hình dưới đây: templates được gom chung lại với nhau theo chung một thư mục, tương tự với static và views



Ngoại trừ *yourapp/views/\_\_init\_\_.py,* mỗi file .py trong *yourapp/views/* đều là blueprint. Trong *yourapp/\_\_init\_\_.py* chúng ta import các blueprints đó và register chúng qua Flask() object.

### **2.8.4 Thư viện JWT (thư viện tạo token, mã hóa thuật toán HS256)**

JSON Web Token (JWT) là 1 tiêu chuẩn mở (RFC 7519)[[5]](#footnote-5) định nghĩa cách thức truyền tin an toàn giữa các thành viên bằng 1 đối tượng JSON. Thông tin này có thể được xác thực và đánh dấu tin cậy nhờ vào "chữ ký" của nó. Phần chữ ký của JWT sẽ được mã hóa lại bằng HMAC hoặc RSA [12].

Hình 2‑26 Giới thiệu JWT

**Cấu trúc của JSON Web Token:**

JSON Web Token bao gồm 3 phần, được ngăn cách nhau bởi dấu chấm (.):

1. Header
2. Payload
3. Signature (chữ ký)

Tổng quát thì nó có dạng như sau:



**Header:** Phần Header dùng để khai báo kiểu chữ ký và thuật toán mã hóa sẽ dùng cho cái token của chúng ta.

**Ví dụ cho phần Header:**



Đoạn Header này khai báo rằng đối tượng được mã hóa là 1 JWT (để phân biệt với JWS hay JWE), và chữ ký của nó sử dụng thuật toán mã hóa HMAC SHA-256.

Đoạn Header này sẽ được mã hóa **base64url**, và ta thu được phần đầu tiên của JWT:



### **2.8.5 Flask\_sqlalchemy (thư viện hỗ trợ kết nối với MySQL)**

SQLAlchemy là một Python toolkit với thế mạnh về OR Mapper giúp người phát triển ứng dụng có được đầy đủ sức mạnh và linh hoạt trong SQL. Flask\_ sqlalchemy là Flask extension hỗ trợ cho ứng dụng Flask [13].

ORM (Objects Relation Mapping)

Dữ liệu trong RDBMS servers được lưu trữ dưới dạng tables. Object relation mapping là một kỹ thuật dùng để mapping object parameters thành cấu trúc bảng RDBMS. ORM API cung cấp các phương thức để thực hiện các hoạt động CRUD mà không cần viết các câu lệnh SQL cơ bản.

## **2.9 Thư viện thuật toán**

### **2.9.1 Pandas**

**Thư viện pandas**

Thư viện pandas trong python là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu. Đây cũng là bộ công cụ phân tích và xử lý dữ liệu mạnh mẽ của ngôn ngữ lập trình python. Thư viện này được sử dụng rộng rãi trong cả nghiên cứu lẫn phát triển các ứng dụng về khoa học dữ liệu. Thư viện này sử dụng một cấu trúc dữ liệu riêng là Dataframe. Pandas cung cấp rất nhiều chức năng xử lý và làm việc trên cấu trúc dữ liệu này. Chính sự linh hoạt và hiệu quả đã khiến cho pandas được sử dụng rộng rãi [14].

Nhóm sử dụng thư viện pandas vì:

* DataFrame đem lại sự linh hoạt và hiệu quả trong thao tác dữ liệu và lập chỉ mục;
* Là một công cụ cho phép đọc/ ghi dữ liệu giữa bộ nhớ và nhiều định dạng file: CSV, text, excel, sql database, hdf5;
* Liên kết dữ liệu thông minh, xử lý được trường hợp dữ liệu bị thiếu. Tự động đưa dữ liệu lộn xộn về dạng có cấu trúc;
* Dễ dàng thay đổi bố cục của dữ liệu;
* Tích hợp cơ chế trượt, lập chỉ mục, lấy ra tập con từ tập dữ liệu lớn.
* Có thể thêm, xóa các cột dữ liệu;
* Tập hợp hoặc thay đổi dữ liệu với group by cho phép bạn thực hiện các toán tử trên tập dữ liệu;
* Hiệu quả cao trong trộn và kết hợp các tập dữ liệu;
* Lập chỉ mục theo các chiều của dữ liệu giúp thao tác giữa dữ liệu cao chiều và dữ liệu thấp chiều;
* Tối ưu về hiệu năng;
* Pandas được sử dụng rộng rãi trong cả học thuật và thương mại. Bao gồm thống kê, thương mại, phân tích, quảng cáo,…

### **2.9.2 Sklearn (Scikit-learn)**

Sklearn là một thư viện thực sự hữu ích cho máy học trong Python. Cùng với các thư viện tương tự như NumPy, SciPy và matplotlib, thư viện Sklearn chứa các công cụ hết sức hiệu quả cho machine learning và statistical modeling bao gồm các thuật toán phân lớp (classification), hồi quy (regression), phân cụm (clustering) và giảm chiều dữ liệu (Dimensionality Reduction) [15].

Chú ý rằng scikit-learn (sklearn) được sử dụng để xây dựng models. Cho nên nó sẽ không được sử dụng cho việc đọc (reading), thao tác (manipulating) và tổng hợp (summarizing) dữ liệu. Có những thư viện đảm nhiệm việc đó chẳng hạn: Numpy, Pandas,…

[](https://www.analyticsvidhya.com/blog/wp-content/uploads/2015/01/scikit-learn-logo.png)

Hình 2‑27 Logo Scikit-learn

Các thành phần trong scikit-learn:

Đi cùng với thư viện scikit-learn là rất nhiều các tính năng/đặc điểm (features):

**Supervised learning algorithms:** scikit-learn cung cấp rất nhiều các loại thuật toán về supervised learning cụ thể như Linear Regression, Support Vector Machines (SVM), Decision Trees, Bayesian,… Với một công cụ hỗ trợ nhiều thuật toán như vậy tạo nên sự tiện lợi trong việc sử dụng nên nhóm tác giả đã chọn thư viện scikit-learn để vận dụng vào khóa luận.

Cross-validation: Có nhiều các phương pháp để kiểm tra độ chính xác của supervised model trên dữ liệu không xem trước.

**Unsupervised learning algorithms:** Tương tự như Supervised learning algorithms thì scikit-learn cũng hỗ trợ số lượng lớn các thuật toán như: clustering (phân cụm), factor analysis, principle component analysis, unsupervised neural networks.

**Various toy datasets:** scikit-learn cung cấp rất nhiều các thư viện có sẵn như IRIS dataset, Boston House prices dataset,…

**Feature extraction:** Rất hữu ích cho việc rút trích các đặc điểm từ hình ảnh và văn bản (Bag of Words)

### **2.9.3 Surprise**

Surprise là một Python scikit building ( Scikits viết tắt của Scipy Toolkits là phần package add-on cho SciPy) và phân tích các hệ thống khuyến nghị [16]

Surprise được thiết kế nhằm mục đích:

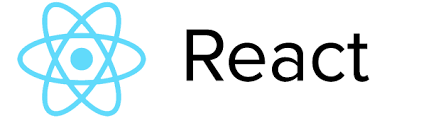
* Giúp user hoàn thiện trong việc kiểm soát toàn bộ các thí nghiệm.
* Giảm bớt khó khăn trong việc xử lý dữ liệu. Users có thể sử dụng dữ liệu được xây dựng sẵn (Movielens, Jester) và dữ liệu của chính họ.
* Cung cấp rất nhiều thuật toán sẵn sàng để sử dụng như prediction algorithms chẳng hạn baseline algorithms, neighborhood methods, matrix factorization-based (SVD, PMF, SVD++, NMF) và các loại khác. Thật vậy, có rất nhiều similarity measures (cosine, MSD, pearson…) được xây dựng sẵn.
* Dễ dàng triển khai thực hiện một ý tưởng thuật toán mới.
* Cung cấp các công cụ để đánh giá, phân tích và so sánh hiệu suất các thuật toán.

Tên Surprise được dựa trên Simple Python Recommendation System Engine.

## **2.10 Thư viện hỗ trợ chính: Reactjs**

React (Hay ReactJS, React.js) là một thư viện Javascript mã nguồn mở để xây dựng các thành phần giao diện có thể tái sử dụng. Nó được tạo ra bởi Jordan Walke, một kỹ sư phần mềm tại Facebook. Người bị ảnh hưởng bởi XHP (Một nền tảng thành phần HTML cho PHP). React lần đầu tiên được triển khai cho ứng dụng Newsfeed của Facebook năm 2011, sau đó được triển khai cho Instagram.com năm 2012. Nó được mở mã nguồn (open-sourced) tại JSConf US tháng 5 năm 2013 [17].

Mã nguồn của React được mở trên GitHub: https://github.com/facebook/react



Hình 2‑28 Logo React

Hiện nay, thư viện này nhận được rất nhiều sự quan tâm đến từ cộng đồng. Nó đang được bảo trì (maintain) bởi Facebook và Instagram, cũng với sự đóng góp của cộng đồng các lập trình viên giỏi trên Thế giới.

Về cơ bản, việc xây dựng một ứng dụng **MVC** phía client với ràng buộc dữ liệu 2 chiều (2 way data-binding) là khá đơn giản. Tuy nhiên nếu dự án ngày càng mở rộng, nhiều tính năng hơn, làm cho việc bảo trì dự án gặp khó khăn, đồng thời hiệu năng cũng bị giảm.

Bạn cũng có thể giải quyết vấn đề đó bằng các thư viện khác như Backbone.js hay Angular.js, tuy nhiên bạn sẽ thấy các hạn chế của chúng khi dự án của bạn ngày càng lớn.

React ra đời sau AngularJS, nó sinh ra để dành cho các ứng dụng lớn dễ dàng quản lý và mở rộng. Mục tiêu chính của React là nhanh, đơn giản, hiệu năng cao và dễ dàng mở rộng.

## **2.11 Các gọi hỗ trợ (package: react-bootstrap, react-easy-chart)**

**React-Bootstrap**

React-Bootstrap cung cấp sẵn một bộ các React component với look-and-feel của Twitter Bootstrap component từ đó giúp việc tạo dựng UI cho React app dễ dàng hơn bao giờ hết. Ví dụ khi ta mô tả một button nhỏ, text là “Something”, click event trigger hàm callback “someCallback”[[6]](#footnote-6).

**React-easy-chart**

React-easy-chart tập hợp các thành phần đồ thị với mục đích tạo điều kiện thuận lợi và nhanh nhất cho việc triển khai thực hiện đồ thị trong một ứng dụng React [18].

# **PHÂN TÍCH, THIẾT KẾ HỆ THỐNG**

## **3.1 Phân tích yêu cầu**

### **3.1.1 Mô hình phân rã chức năng (BFD – Business Function Diagram)**

Hình 3‑1 Mô hình phân rã chức năng (BFD)

### **3.1.2 Yêu cầu chức năng**

- Yêu cầu đọc và xử lý dữ liệu người dùng

- Yêu cầu hiển thị dữ liệu

- Yêu cầu chạy thuật toán khuyến nghị

- Yêu cầu hiển thị kết quả

- Yêu cầu xuất kết quả

### **3.1.3 Yêu cầu phi chức năng**

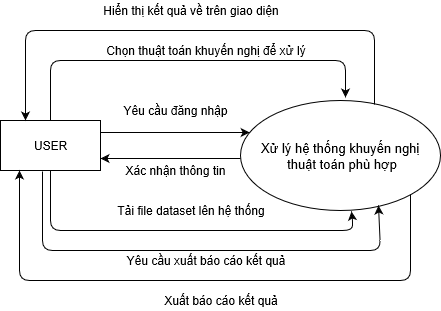
- Yêu cầu về giao diện: Giao diện quản lý gần gũi, đơn giản. Giao diện dùng cho người dùng thân thiện, sinh động, tương tác cao.

- Yêu cầu về chất lượng:

* Tính tiến hóa:
* Dễ dàng mở rộng khi hệ thống phát triển.
* Tính tiện dụng:
* Hệ thống có giao diện trực quan, thân thiện và dễ sử dụng
* Các đặc tả trên giao diện và hướng dẫn sử dụng rõ ràng.
* Tính hiệu quả:
* Trang web thể hiện trực quan các trạng thái thông tin, giúp người dùng dễ thao tác.
* Trang web hoạt động ổn định và đáng tin cậy.
* Tính tương thích:
* Hoạt động tốt với tất cả các trình duyệt web phổ biến hiện nay (Internet Exploler, FireFox, Google Chrome.

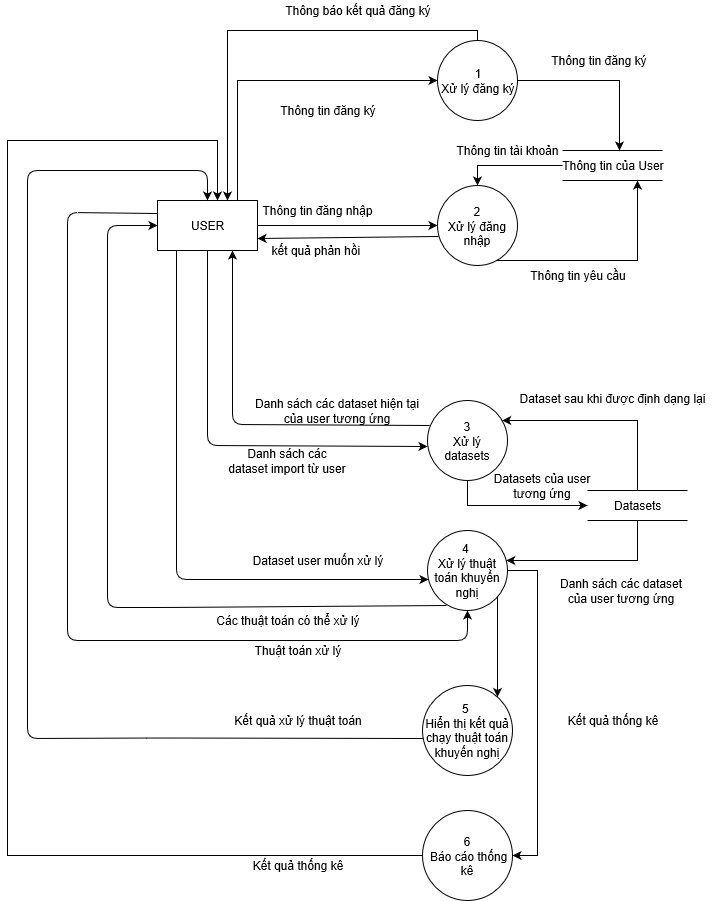
## **3.2. Phân tích hệ thống về xử lý - Sơ đồ DFD (Data Flow Diagram)**

### **3.2.1. DFD mức 0**



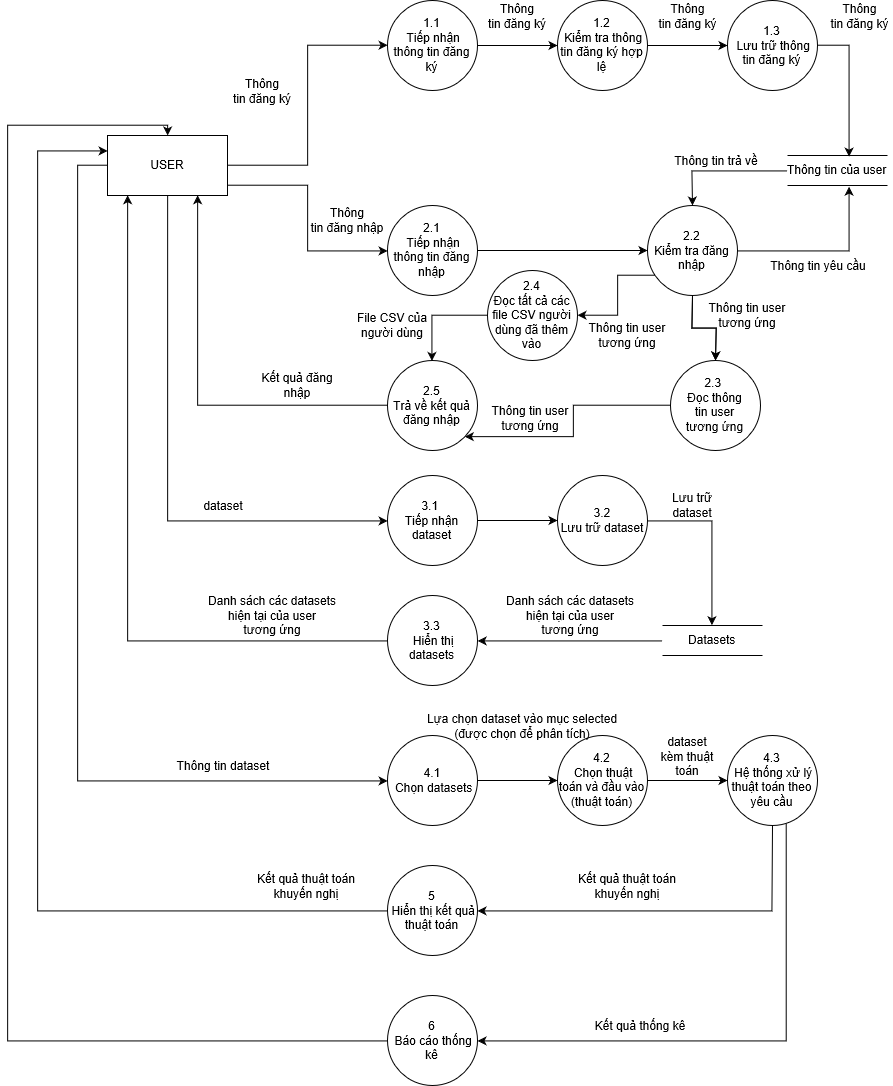
Hình 3‑2 Sơ đồ DFD mức 0

### **3.2.2. DFD mức 1**



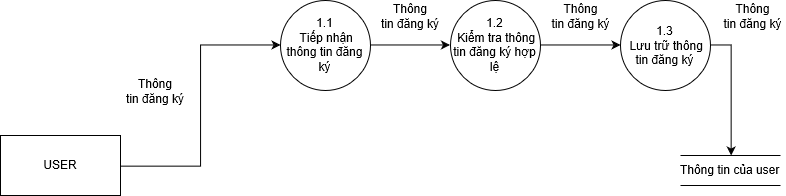
Hình 3‑3 Sơ đồ DFD mức 1

### **3.2.3. DFD mức 2 tổng quát**



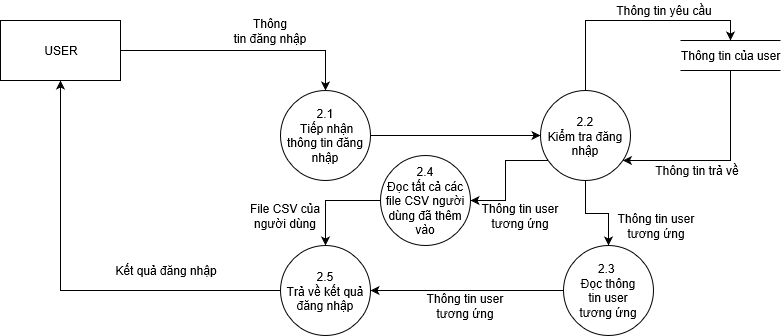
Hình 3‑4 Sơ đồ DFD mức 2

#### **3.2.3.1 DFD mức 2 xử lý đăng ký**

****

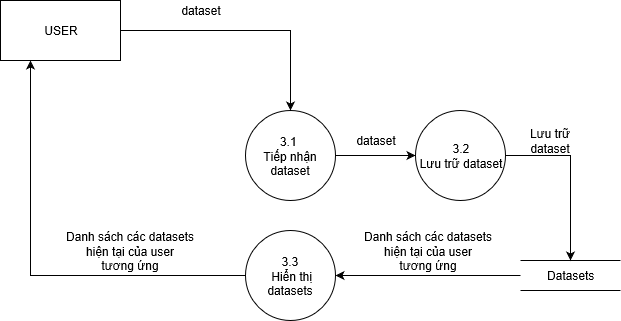
Hình 3‑5 DFD mức 2 xử lý đăng ký

#### **3.2.3.2 DFD mức 2 xử lý đăng nhập**

****

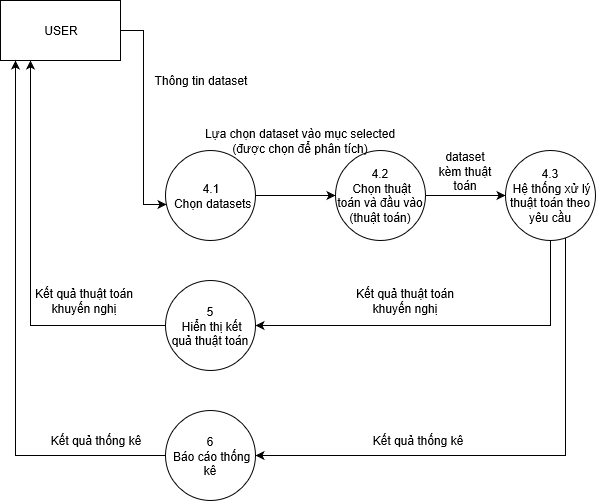
Hình 3‑6 DFD mức 2 xử lý đăng nhập

#### **3.2.3.3 DFD mức 2 xử lý datasets**

****

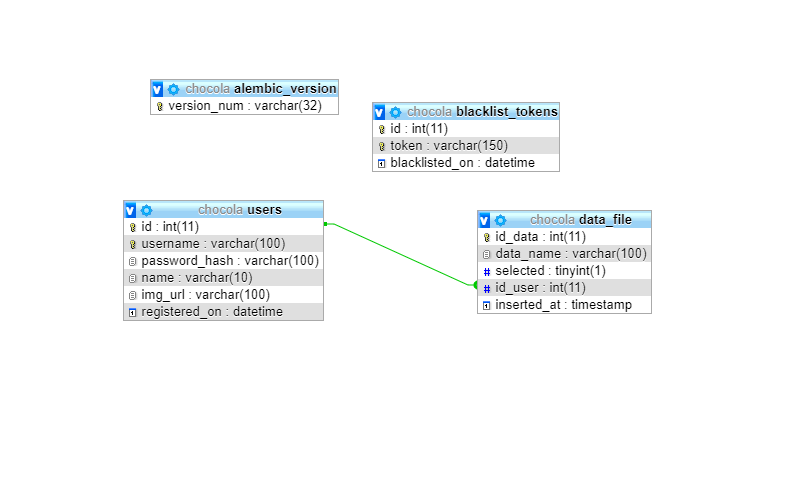
Hình 3‑7 DFD mức 2 xử lý datasets

#### **3.2.3.4 DFD mức 2 xử lý thuật toán khuyến nghị và hiển thị kết quả, thống kê**

****

Hình 3‑8 DFD mức 2 xử lý thuật toán khuyến nghị và hiển thị kết quả, thống kê

## **3.3. Phân tích thiết kế về dữ liệu**



Hình 3‑9 Cơ sở dữ liệu

1. **blacklist\_tokens (id, token, blacklisted\_on):** Bảng chứa dữ liệu token đã dùng, tính số session người dùng đăng nhập.
2. **alembic\_version (version\_num):** Bảng chứa dữ liệu thông tin version.
3. **users (id, username, password\_hash, name, img\_url, registered\_on):** Bảng chứa dữ liệu thông tin người dùng.
4. **data\_file (id\_data, data\_name, selected, id\_user, inserted\_at):** Bảng chứa dữ liệu các data file của user tương ứng.

Mô tả thông tin của các bảng

* Đặc tả bảng “blacklist\_tokens**”**

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên | Kiểu | Ý nghĩa |
| 1 | Id | int | Khóa chính của token |
| 2 | Token | varchar | Chuỗi token |
| 3 | blacklisted\_on | datetime | Ngày giờ user sử dụng hết phiên làm việc của mình |

* Đặc tả bảng “alembic\_version”

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên | Kiểu | Ý nghĩa |
| 1 | Version\_num | varchar | Phiên bản của phần mềm |

* Đặc tả bảng “users”

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên | Kiểu | Ý nghĩa |
| 1 | Id | int | Khóa chính của user |
| 2 | Username | varchar | Username của user |
| 3 | Password\_hash | varchar | Password của user đã được mã hóa |
| 4 | Name | varchar | Tên của user |
| 5 | img\_url | varchar | Hình đại diện của user |
| 6 | registered\_on | datetime | Thời điểm đăng ký tài khoản của user |

* Đặc tả bảng “data-file”

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên | Kiểu | Ý nghĩa |
| 1 | id\_data | int | Khóa chính của data |
| 2 | Data\_name | varchar | Tên của data |
| 3 | Selected | tinyint | Thể hiện data có được chọn hay không |
| 4 | id\_user | int | Khóa ngoại của data tham chiếu tới khóa chính của user |
| 5 | Inserted\_at | timestamp | Thời điểm file data được thêm vào |

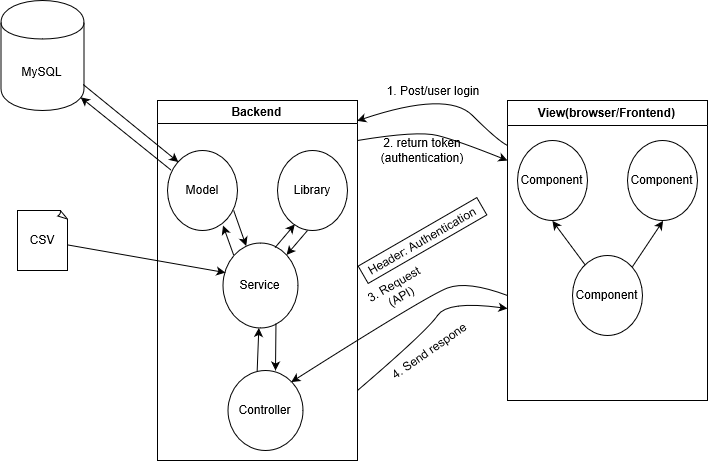
# **CÀI ĐẶT VÀ TRIỂN KHAI HỆ THỐNG**

## **4.1 Cài đặt và triển khai hệ thống**

### **4.1.1 Các công cụ xây dựng hệ thống**

* **Phía Server:**
* Sử dụng framework của Python: Flask.
* Hệ quản trị cơ sở dữ liệu: Mysql.
* Và một số thư viện hỗ trợ khác.
* **Phía client:**
* Reactjs, HTML, CSS3, Ajax, Jquery.
* Các thư viện npm kèm theo: easy react chart, boostrap, jquery,…

### **4.1.2 Triển khai hệ thống**



Hình 4‑1 Sơ đồ triển khai của hệ thống

Hệ thống được phát triển và triển khai trên môi trường:

* Trình duyệt: trình duyệt web.
* Hệ quản trị cơ sở dữ liệu: Mysql.
* Các thư viện, dịch vụ: Python (Flask), Javascript (ReactJS).
* Công cụ phân tích hệ thống: draw.io.
* IDE xây dựng ứng dụng: Visual Studio Code 1.35, Sublime Text 3.
* Công cụ quản lý mã nguồn: Github.

## **4.2 Giao diện của hệ thống**

### **4.2.1 Sơ đồ tổ chức giao diện**

Hình 4‑2 Sơ đồ tổ chức giao diện

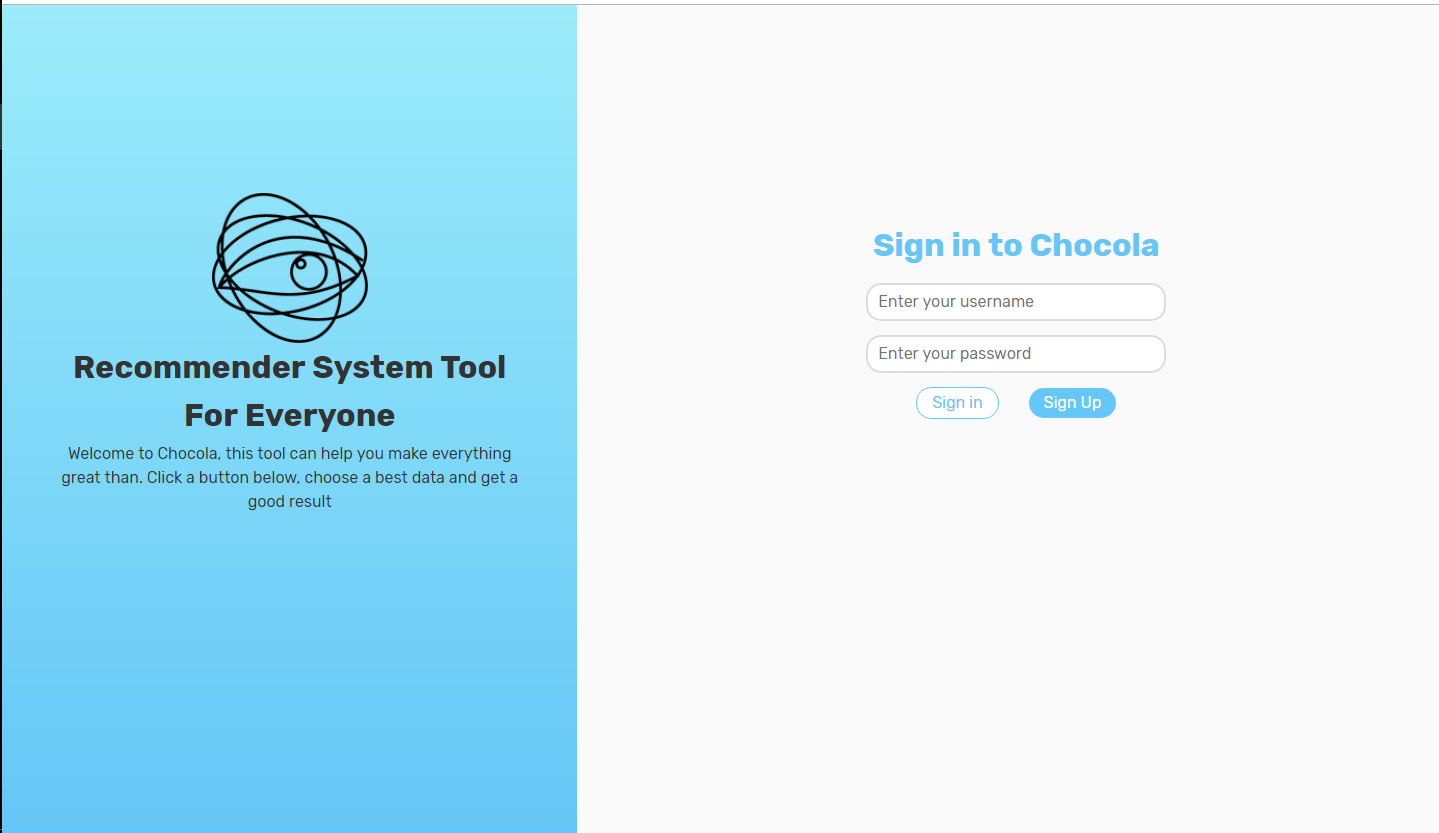
Bảng 4‑1 Mô tả sơ đồ tổ chức giao diện

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên giao diện** | **Chức năng-Ý nghĩa** | **Đối tượng tương tác** |
| 1 | Trang tạo tài khoản | Người dùng đăng ký vào hệ thống | Người dùng |
| 2 | Trang đăng nhập | Người dùng đăng nhập vào hệ thống | Người dùng |
| 3 | Trang thông tin tài khoản | Hiển thị thông tin bao gồm thông tin người dùng, số lượng data, danh sách các data, button import data | Người dùng |
| 4 | Trang thông tin dữ liệu | Hiển thị thông tin dữ liệu đang được chọn bao gồm tên dữ liệu, số lượng cột, dòng, thông tin chi tiết dữ liệu, biểu đồ | Người dùng |
| 5 | Trang kết quả chạy thuật toán | Hiển thị kết quả chạy thuật toán bao gồm biểu đồ, thời gian chạy, kết quả của từng loại thuật toán | Người dùng |

### **4.2.2 Một số giao diện chính**

* Giao diện đăng nhập

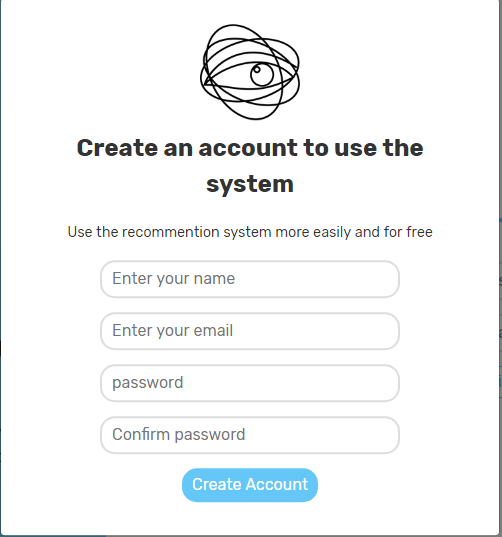
Nội dung: Giao diện đăng nhập



Hình 4‑3 Giao diện đăng nhập

* Tạo tài khoản

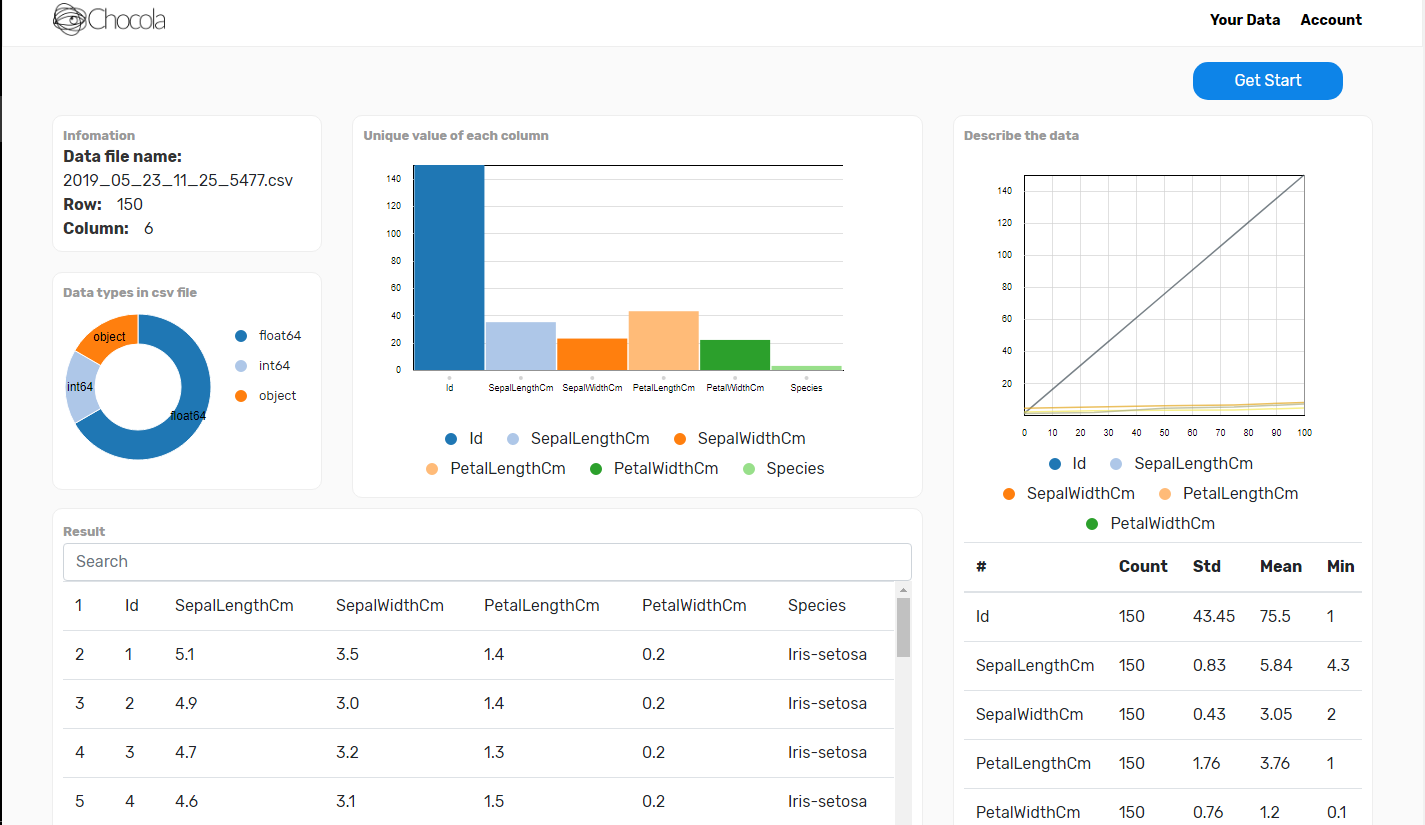
Nội dung: Tạo tài khoản người dùng

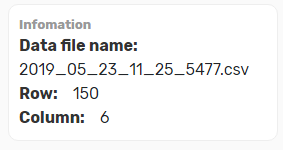


Hình 4‑4. Tạo tài khoản người dùng

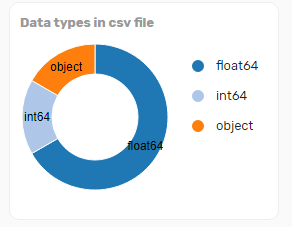
* Trang chủ

Nội dung: Tổng quan trang chủ

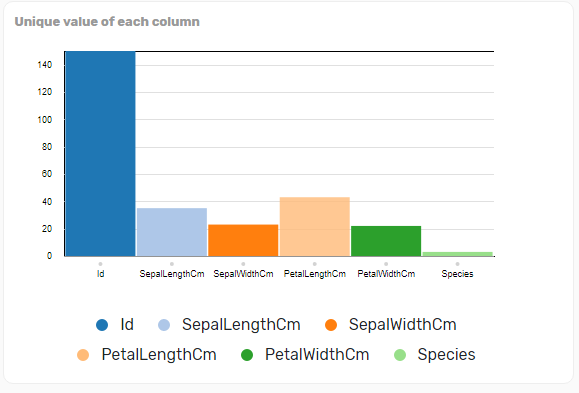


Hình 4‑5. Tổng quan trang chủ

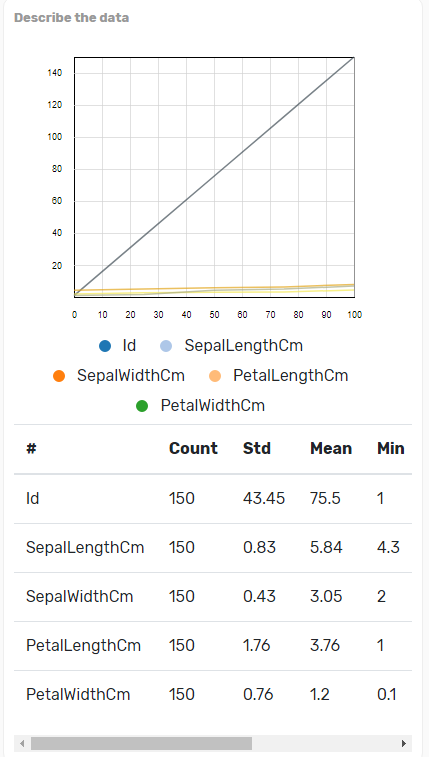
Hình 4‑6. Thông tin file dữ liệu



Hình 4‑7 Thông tin kiểu dữ liệu

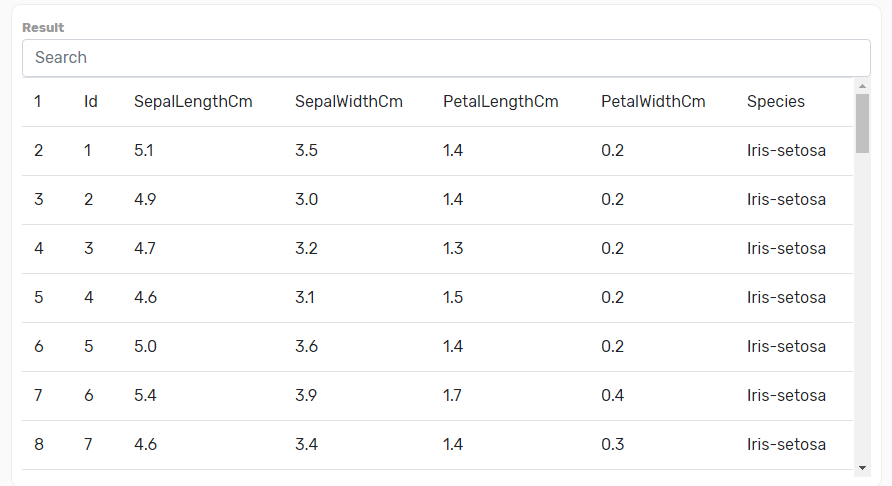


Hình 4‑8. Biểu đồ số lượng giá trị độc nhất có trong dữ liệu



Hình 4‑9. Thông tin mô tả giá trị trong file dữ liệu

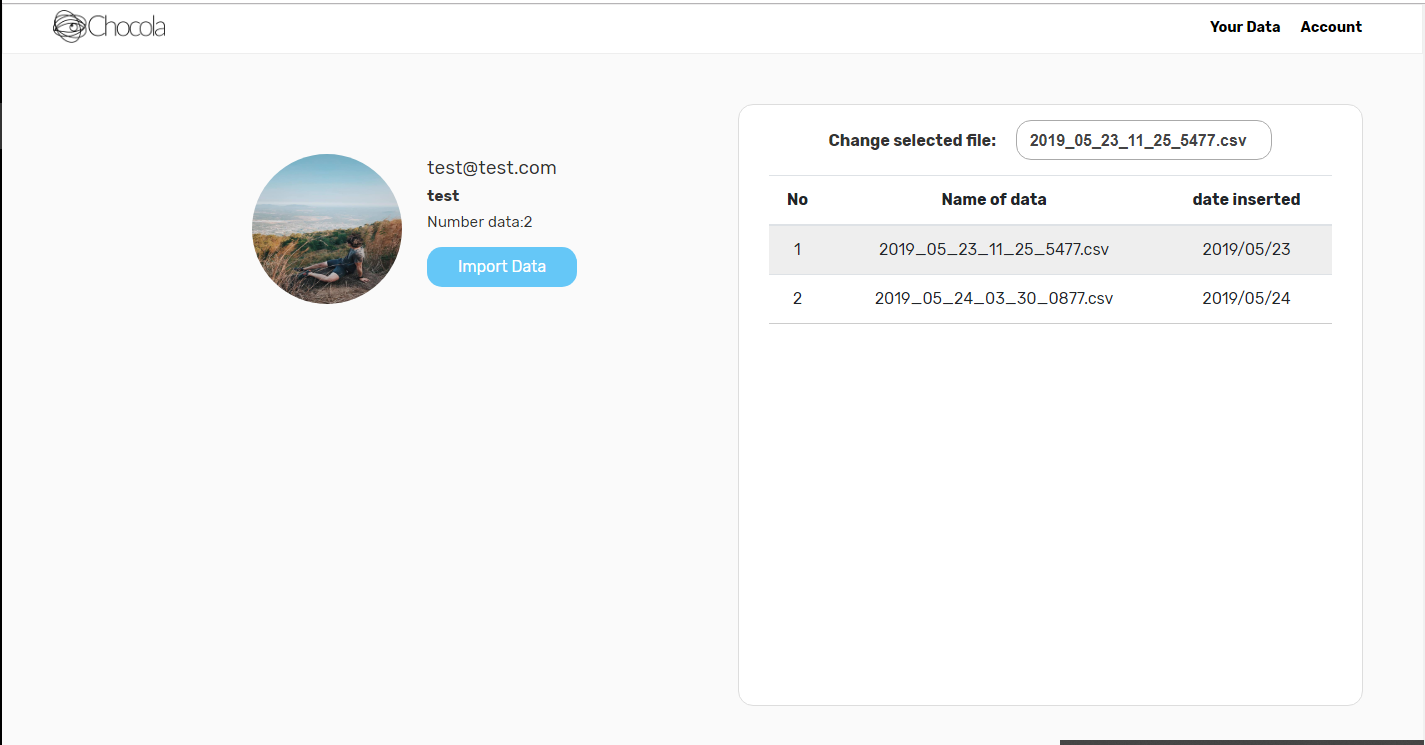
Nội dung: Số liệu có trong file



Hình 4‑10. Số liệu có trong file

* Thông tin tài khoản

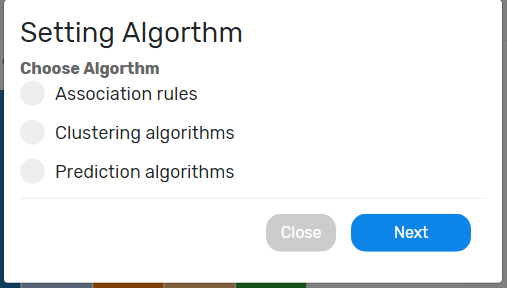
Nội dung: Thông tin tài khoản người dùng



Hình 4‑11. Thông tin tài khoản người dùng

* Chọn thuật toán

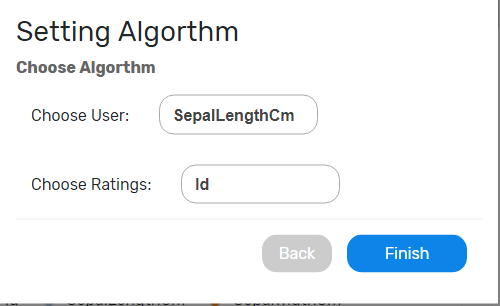
Nội dung: Chọn thuật toán



Hình 4‑12. Chọn thuật toán

* Chọn cột đầu vào thuật toán k-mean

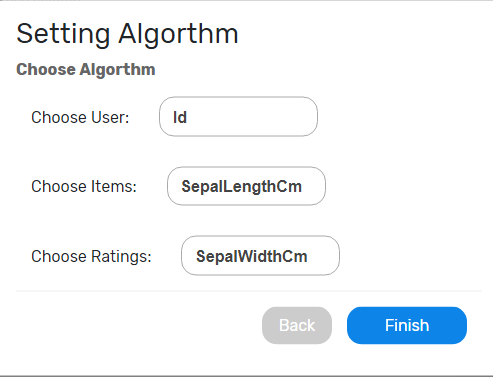
Nội dung: Chọn cột đầu vào thuật toán k mean



Hình 4‑13. Chọn cột đầu vào thuật toán K-means

* Chọn đầu vào thuật toán gợi ý

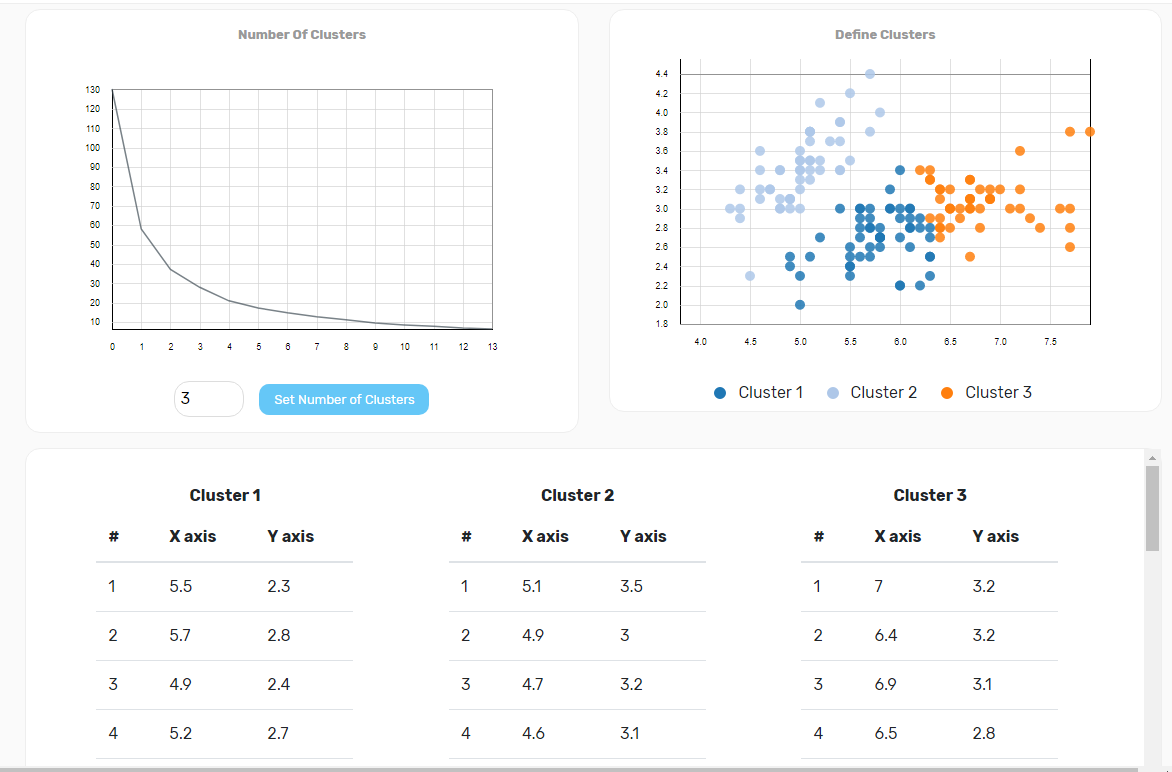
Nội dung: Chọn đầu vào thuật toán gợi ý



Hình 4‑14. Chọn đầu vào thuật toán gợi ý

* Tổng quan giao diện khi chạy thuật toán k mean

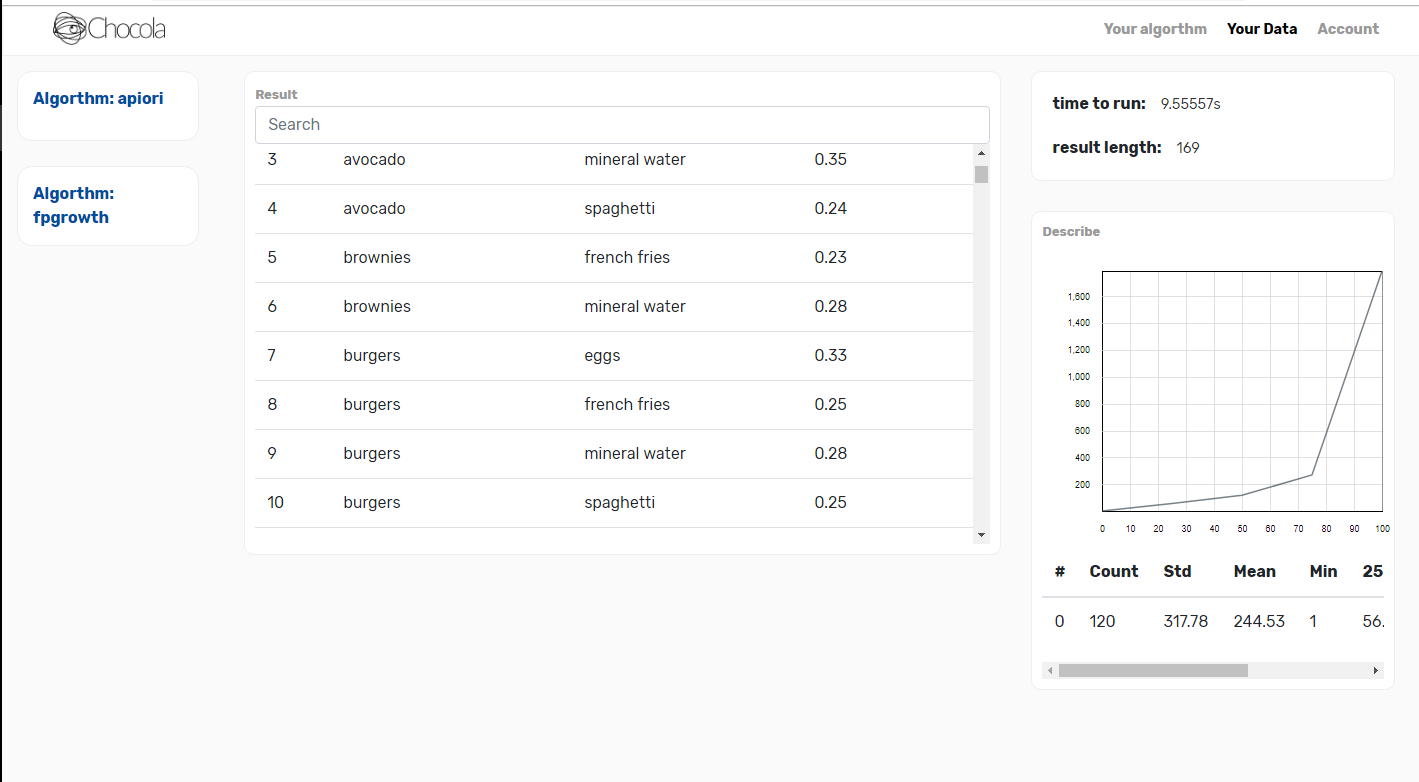
Nội dung: Tổng quan giao diện khi chạy thuật toán k mean



Hình 4‑15. Tổng quan giao diện khi chạy thuật toán k-mean

* Tổng quan giao diện khi chạy thuật toán Association Rule

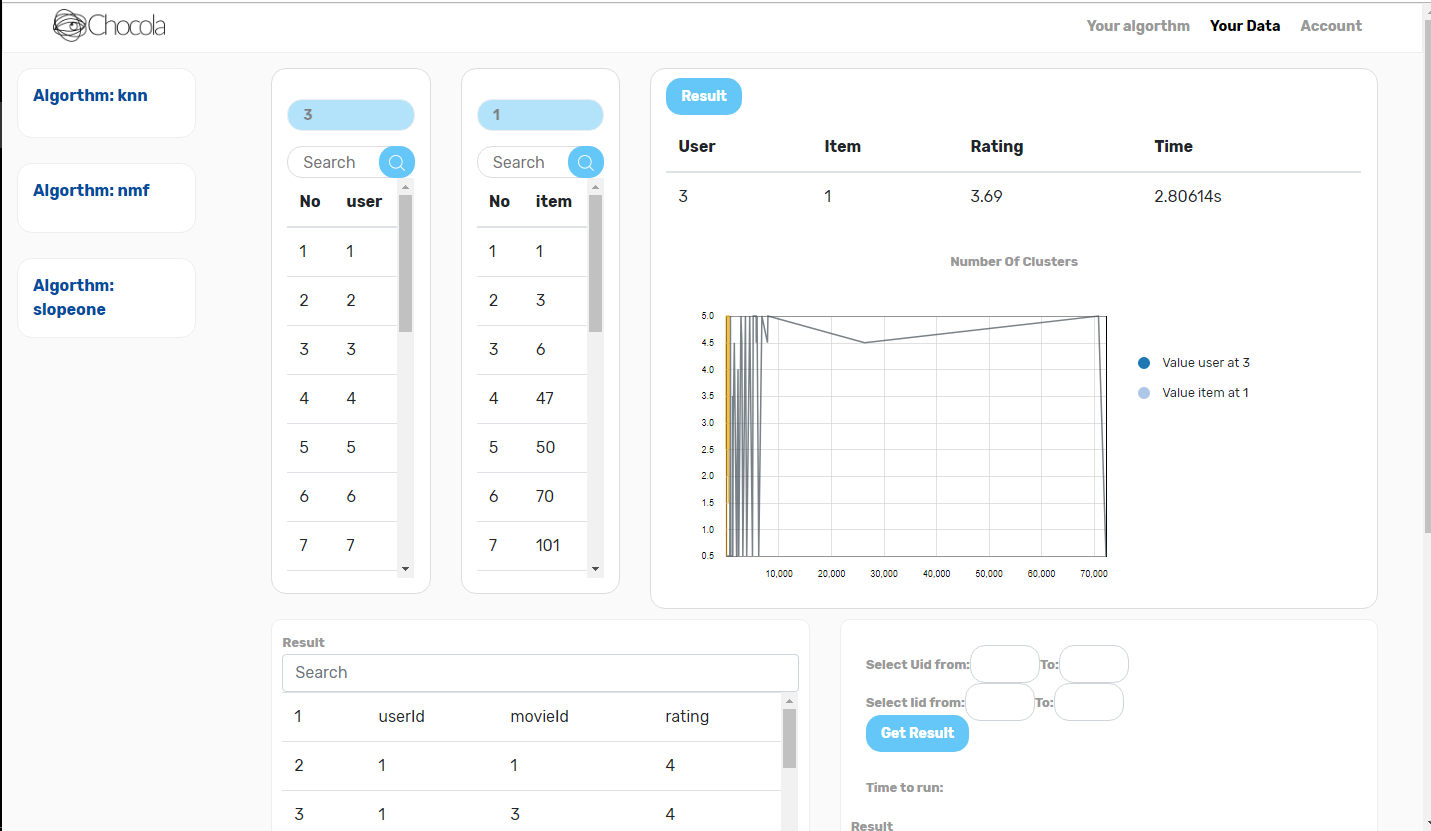
Nội dung: Tổng quan giao diện khi chạy thuật toán Association Rule

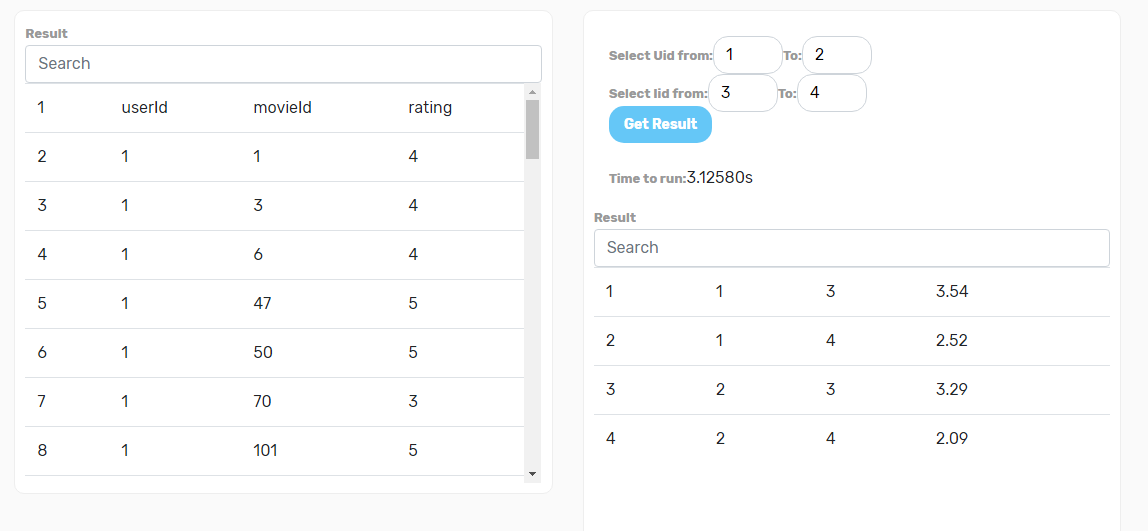


Hình 4‑16. Tổng quan giao diện khi chạy thuật toán Association Rule

* Tổng quan giao diện khi chạy thuật toán predict

Nội dung: Tổng quan giao diện khi chạy thuật toán predict





Hình 4‑17. Tổng quan giao diện khi chạy thuật toán predict

# **THỬ NGHIỆM HỆ THỐNG**

## **Thử nghiệm**

Bảng 5‑1 Mô tả thử nghiệm hệ thống

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Trường hợp thực hiện** | **Người thực hiện** | **Các bước thực hiện** | **Kết quả** | **Kết luận** |
| 1 | Đăng ký tài khoản | Người dùng hệ thống | Nhập thông tin | Hoàn thành | Tốt |
| 2 | Đăng nhập | Người dùng hệ thống | Nhập tài khoản và mật khẩu | Hoàn thành | Tốt |
| 3 | Thêm dữ liệu | Người dùng hệ thống | Thêm file CSV vào website | Hoàn thành | Tốt |
| 4 | Thay đổi dữ liệu được chọn | Người dùng hệ thống | Chọn dữ liệu muốn thay đổi | Hoàn thành | Tốt |
| 5 | Trực quan file dữ liệu | Người dùng hệ thống | Chọn chức năng trực quan | Hoàn thành | Hoàn thành tốt ở các kiểu data bình thường |
| 6 | Tự động tìm ra min\_sup và min\_conf của thuật toán Association Rule | Hệ thống | Hệ thống tự động chạy ngầm | Hoàn thành | Hoàn thành tốt ở các kiểu data bình thường |
| 7 | Chạy thuật toán Association Rule | Hệ thống | Hệ thống chạy khi được kích hoạt | Hoàn thành | Hoàn thành tốt ở các kiểu data bình thường |
| 8 | Chạy thuật toán predict | Hệ thống | Hệ thống chạy khi được kích hoạt | Hoàn thành | Hoàn thành tốt ở các kiểu data bình thường |
| 9 | Chạy thuật toán clustering | Hệ thống | Hệ thống chạy khi được kích hoạt | Hoàn thành | Hoàn thành tốt ở các kiểu data bình thường |

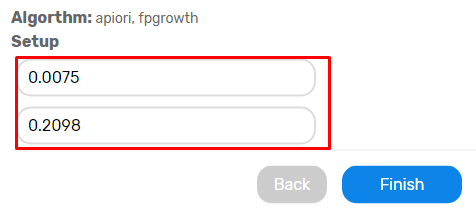
## **5.2 Đánh giá kết quả thử nghiệm**

### **5.2.1 Kết quả thử nghiệm**

#### **5.2.1.1 Kết quả khi tự động tìm ra min\_sup và min\_conf của thuật toán Association Rule**

Thông tin dữ liệu: Tập dữ liệu 7500 dòng, 20 cột, dữ liệu là số lần mua các mặt hàng trong vòng một tuần của một của hàng tại Pháp[[7]](#footnote-7). Mỗi dòng thể hiện một lần mua hàng. Có tất cả 120 sản phẩm.

* Kết quả:



Hình 5‑1 Kết quả khi chạy công thức tính min\_sup và min\_conf

* Kết quả khi sử dụng chạy thuật toán:

Thuật toán Apriori:

* Thời gian chạy: 8s
* Số luật đưa ra được: 169

Thuật toán Fp-Growth:

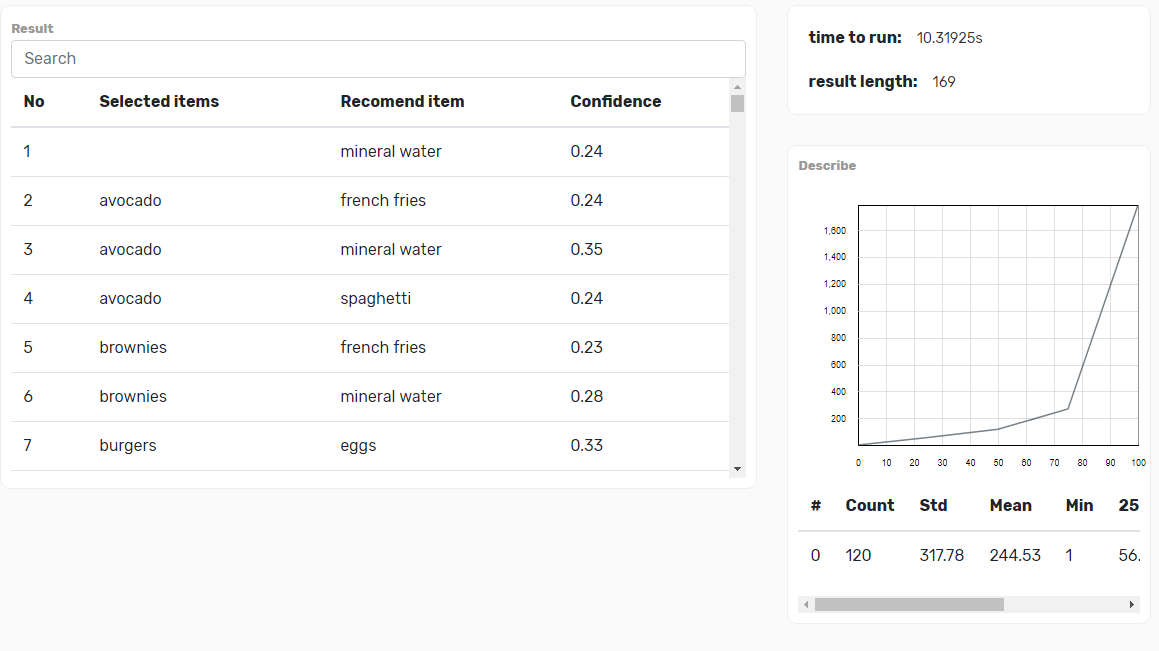
* Thời gian chạy: 10s
* Số luật đưa ra được: 107
* Đánh giá:

Công thức chạy được với phần lớn các trường hợp, đưa ra được số luật nhiều, phù hợp với những dữ liệu không thực sự quá lớn.

#### **5.2.1.2 Chạy thuật toán Association Rule**

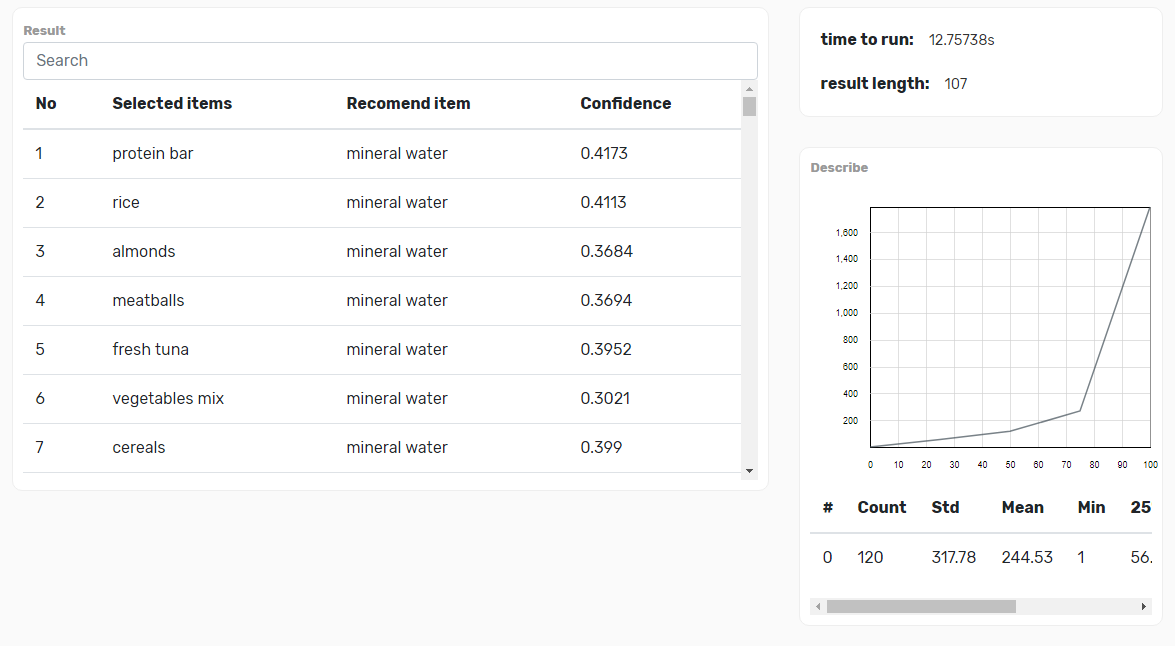
Thông tin dữ liệu: Tập dữ liệu 7500 dòng, 20 cột, dữ liệu là số lần mua các mặt hàng trong vòng một tuần của một của hàng tại Pháp[[8]](#footnote-8). Mỗi dòng thể hiện một lần mua hàng. Có tất cả 120 sản phẩm.

* Kết quả chạy thuật toán Apriori:



Hình 5‑2 Kết quả khi chạy thuật toán Apriori

* Kết quả chạy thuật toán FP-Growth:



Hình 5‑3 Kết quả khi chạy thuật toán FP-Growth

* Đánh giá thuật toán:

Thuật toán Apriori:

* Thời gian chạy: 8s
* Số luật đưa ra được: 169

Thuật toán FP-Growth:

* Thời gian chạy: 10s
* Số luật đưa ra được: 107

Số liệu trên cho thấy thuật toán Apriori chạy tốt hơn về thời gian và số luật đưa ra khi cùng một dữ liệu đầu vào. Nhưng thuật toán FP-Growth sẽ tốt hơn khi với các bộ dữ liệu nhỏ hơn.

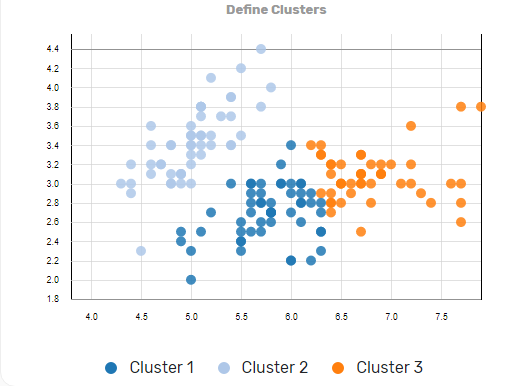
#### **5.2.1.3 Kết quả khi chạy thuật toán clustering**

Thông tin dữ liệu: Tập dữ liệu 150 dòng, 6 cột, dữ liệu là thông tin chi tiết của cây Iris[[9]](#footnote-9).

Dữ liệu đầu vào:5 dòng đầu của dữ liệu đầu vào

|  |  |
| --- | --- |
| SepalLengthCm | SepalWidthCm |
| 5.1 | 3.5 |
| 4.9 | 3.0 |
| 4.7 | 3.2 |
| 4.6 | 3.1 |
| 5.0 | 3.6 |

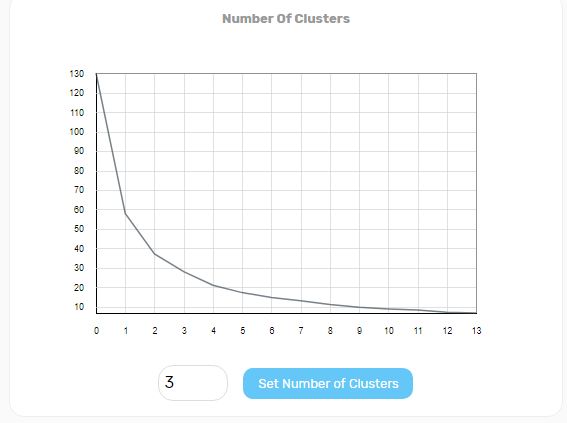
* Kết quả:



Hình 5‑4 Kết quả khi chạy thuật toán K-means



Hình 5‑5 Kết quả khi chạy thuật toán K-means



Hình 5‑6 Kết quả biểu đồ elbow

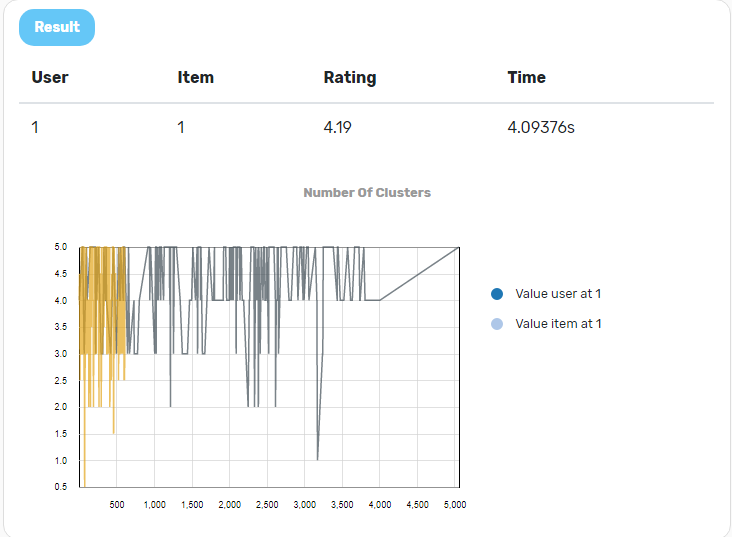
* Đánh giá:

Thuật toán chạy nhanh và đưa kết quả khá chính xác.

#### **5.2.1.4 Chạy thuật toán predict**

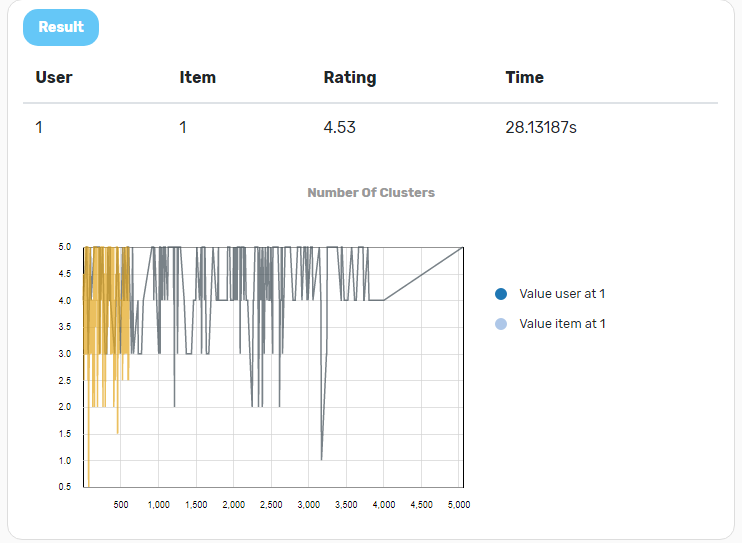
Thông tin dữ liệu: Tập dữ liệu 100836 dòng, 3 cột, dữ liệu là các đánh giá phim (rating) của người dùng (userId) và phim (movieId)[[10]](#footnote-10).

* Kết quả chạy thuật toán KNN khi dự đoán user = 1 và item = 1;



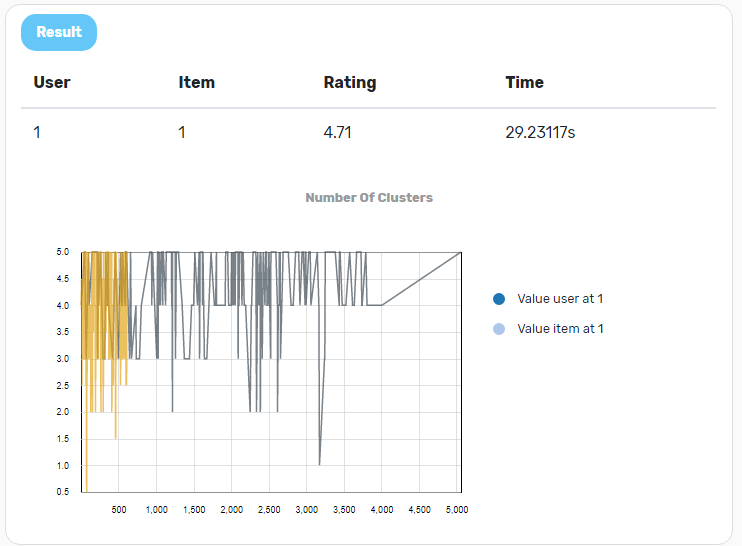
Hình 5‑7 Kết quả khi chạy thuật toán KNN

* Kết quả chạy thuật toán NMF khi dự đoán user = 1 và item = 1;



Hình 5‑8 Kết quả khi chạy thuật toán NMF

* Kết quả chạy thuật toán Slope One khi dự đoán user = 1 và item = 1;



Hình 5‑9 Kết quả khi chạy thuật toán Slope One

* Đánh giá thuật toán khi dự đoán một giá trị:

Thuật toán KNN:

* Giá trị dự đoán: 4.19
* Thời gian chạy: 4.09

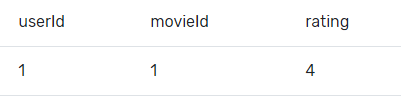
Thuật toán NMF:

* Giá trị dự đoán: 4.53s
* Thời gian chạy: 28s

Thuật toán Slope One:

* Giá trị dự đoán: 4.17s
* Thời gian chạy: 29s

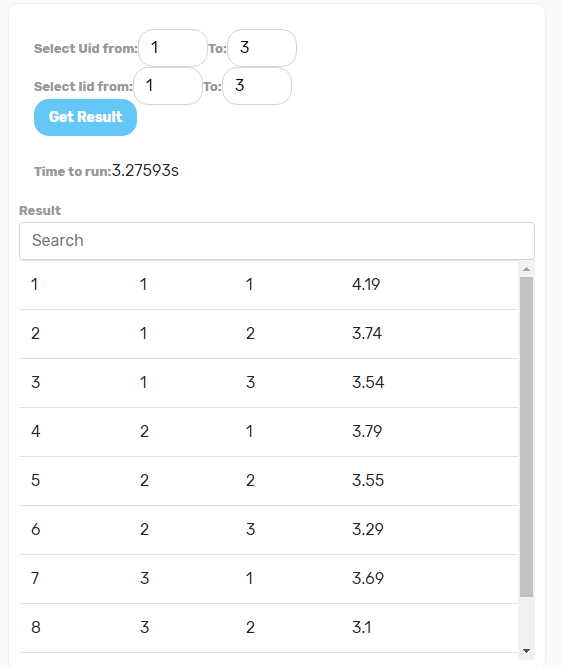
Giá trị thực tế:



Hình 5‑10 Giá trị thực tế

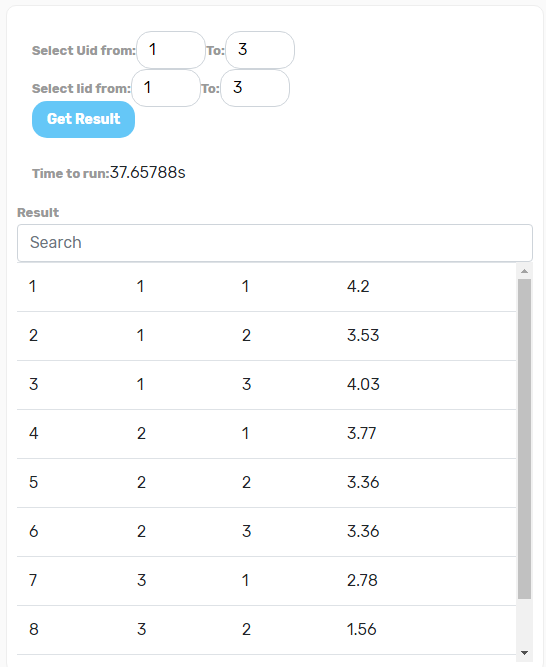
Ta có thể thấy rằng với cùng một giá trị là userId =1 và movieId =1 thì thuật toán KNN chạy tốt hơn rất nhiều với 2 thuật toán còn lại về độ chính xác và thời gian.

* Kết quả chạy thuật toán KNN khi dự đoán user từ 1 đến 3 và item từ 1 đến 3



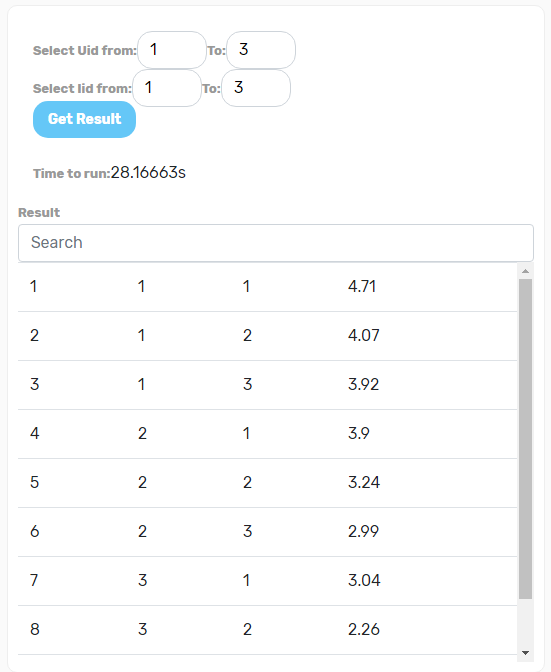
Hình 5‑11 Kết quả khi chạy thuật toán KNN

* Kết quả chạy thuật toán NMF khi dự đoán user từ 1 đến 3 và item từ 1 đến 3



Hình 5‑12 Kết quả khi chạy thuật toán NMF

* Kết quả chạy thuật toán Slope One khi dự đoán user từ 1 đến 3 và item từ 1 đến 3



Hình 5‑13 Kết quả khi chạy thuật toán Slope One

* Đánh giá thuật toán khi dự đoán nhiều giá trị:

Thuật toán KNN:

* Thời gian chạy: 3.2

Thuật toán NMF:

* Thời gian chạy: 37.6s

Thuật toán Slope One:

* Thời gian chạy: 28s

Ta có thể thấy rằng khi chạy giá trị dự đoán nhiều hơn một. Cụ thể là từ 1 tới 3, thì thuật toán KNN vẫn chiếm ưu thế cao hơn. Về tốc độ, dù kết quả vẫn tương đối gần bằng nhau (độ chênh lệch mỗi giá trị không chênh lệch quá 1). Và điều đặc biệt, khi chạy với nhiều giá trị thì thuật toán KNN và Slope One cho ra thời gian nhanh hơn hẳn chạy với 1 giá trị.

### **5.2.2 Đánh giá thực nghiệm**

* **Đánh giá thuật toán luật kết hợp.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * + 1. **Thuật toán** | * + 1. **Thời gian chạy** | * + 1. **Kết quả** | * + 1. **Đánh giá** | * + 1. **Ghi chú** |
| * + 1. Apriori | * + 1. 8s | * + 1. 169 luật | * + 1. Thuật toán tốt hơn, đưa ra số luật nhiều hơn và thời gian chạy nhanh hơn. | * + 1. Kết quả cho cùng một giá trị với 5 lần chạy (cùng một bộ dữ liệu) |
| * + 1. Fp-growth | * + 1. 10s | * + 1. 107 luật | * + 1. Thuật toán chạy không tốt bằng thuật toán Apriori với bộ dữ liệu trên 2000 dòng | * + 1. Thuật toán chạy tốt hơn khi thực nghiệm với bộ dữ liệu ít hơn 2000 dòng. |

* **Đánh giá thuật toán phân cụm.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * + 1. **Thuật toán** | * + 1. **Thời gian chạy** | * + 1. **Kết quả** | * + 1. **Đánh giá** | * + 1. **Ghi chú** |
| * + 1. K-mean | * + 1. 3s | * + 1. 3 cụm,   Khoảng cách điểm xa nhất với tâm là 3.67 (khoảng cách điểm đầu đến điểm cuối là 9) | * + 1. Thuật toán đưa ra kết quả tốt. Không có các điểm “nhiễu” (phần tử tách lẻ với với phần còn lại) | * + 1. Kết quả cho cùng một giá trị với 5 lần chạy (cùng một bộ dữ liệu) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * + 1. **Thuật toán** | * + 1. **Thời gian chạy** | * + 1. **Kết quả** | * + 1. **Đánh giá** | * + 1. **Ghi chú** |
| * + 1. KNN | * + 1. 4.09s | Giá trị dự đoán 4.19 | * + 1. Thuật toán cho ra kết quả gần đúng nhất với thời gian chạy nhanh nhất | * + 1. Giá trị thực tế 4 |
| * + 1. NMF | * + 1. 28s | Giá trị dự đoán 4.83 | * + 1. Thuật toán đưa ra kết quả ít chính xác nhất, với thời gian chạy lâu | * + 1. Giá trị thực tế 4 |
| * + 1. Slope One | * + 1. 29s | Giá trị dự đoán 4.17s | * + 1. Thuật toán cho kết quả tương đối tốt nhưng thời gian chạy khá lâu. | * + 1. Giá trị thực tế 4 |
| * + 1. *Qua nhiều lần chạy thuật toán với bộ test 1000 dòng, cho được kết quả thuật toán KNN tốt nhất tiếp đó là thuật toán Slope One và cuối cùng là NMF.* | | | | |

* **Đánh giá thuật toán gợi ý.**

# **KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

## **6.1 Kết quả đạt được**

Trong quá trình thực hiện đề tài, nhóm tác giả tiếp cận với nhiều nghiên cứu liên quan cả trong nước và ngoài nước. Điều đó giúp chúng tôi hiểu rộng hơn, sâu hơn và kỹ hơn về đề tài.

Sau quá trình tìm hiểu và thực hiện đề tài, nhóm tác giả đã hoàn thành được các mục tiêu đề ra:

* Đọc file dữ liệu CSV của người dùng.
* Nghiên cứu công thức tính min\_sup và min\_conf đầu vào phù hợp.
* Tìm hiểu và đưa ra được cách trực quan dữ liệu, giúp người dùng dễ dàng phân tích dữ liệu đang có.
* Nghiên cứu các thuật toán Association Rule (luật kết hợp) và đưa ra được giải pháp tự động tìm min\_sup và min\_conf cũng như chạy hai thuật toán này.
* Nghiên cứu thuật toán clustering (phân cụm) là K-means để chạy thuật toán cũng như trực quan kết quả có được.
* Nghiên cứu thuật toán predict và để chạy thuật toán cũng như trực quan kết quả có được.
* Tìm hiểu và ứng dụng các thư viện thuật toán vào hệ thống.
* Xây dựng website trên ngôn ngữ lập trình python và javascript.

## **6.2 Đánh giá ưu điểm, khuyết điểm**

### **6.2.1 Ưu điểm**

Sau thời gian nghiên cứu và thực hiện, nhóm tác giả đã nghiên cứu một số thuật toán thường dùng trong khai phá dữ liệu cho các dữ liệu thu thập được, kết quả của việc phân tích dữ liệu cũng được trực quan hóa. Hệ thống cũng đạt được mục tiêu đặt ra và có một số ưu điểm như sau:

* Giao diện thân thiện, có các biểu đồ minh họa rõ ràng giúp người dùng dễ dàng phân tích và đánh giá được dữ liệu của mình như biết các loại dữ liệu có trong file, các giá trị duy nhất ở mỗi cột, giá trị tăng của mỗi cột, cũng như các biểu đồ hỗ trợ trực quan hóa thuật toán.
* Tự động hóa các dữ liệu đầu vào của thuật toán, người dùng không cần hiểu rõ về các thuật toán cũng có thể sử dụng được qua vài click chuột (những thuật toán không thể chạy với dữ liệu đó thì không thể click được).
* Chạy thuật toán tương đối nhanh và ổn định.
* Độ bảo mật tương đối tốt, sử dụng công nghệ JSON Web Token (JWT) để truyền tin giữa user và server.
* Có nhiều thuật toán, giúp người dùng có thể tùy biến sử dụng và so sánh, đánh giá.
* Xuất kết quả chạy thuật toán ra các file CSV, PDF nhanh chóng và dễ dàng.

### **6.2.2 Khuyết điểm**

Trong quá trình thực hiện đề tài, nhóm tác giả không thể tránh khỏi những sai  
sót. Sau đây, một số khuyết điểm mà chúng tôi nhận thấy:

* Đối với người dùng mới (không biết gì về Công nghệ Thông tin): hệ thống sẽ hơi khó để tiếp cận vì mang tính chuyên môn cao.
* Chưa thêm được các loại dữ liệu đầu vào khác ngoài file CSV.
* Người dùng khó xác định rõ nên đưa dữ liệu đầu vào nào khi sử dụng thuật toán để đưa kết quả mong muốn.
* Ứng dụng chưa tự động đánh giá thuật toán.
* Phân chia ngẫu nhiên tập dữ liệu training và testing, chọn các tham số ngẫu nhiên cho từng thuật toán.
* Số bộ dữ liệu của từng tập dữ liệu thử nghiệm ít.
* Chưa đánh giá được độ chính xác của kết quả sau khi chạy thuật toán.

## **6.3 Hướng phát triển**

Như đã trình bày, hệ thống hiện tại vẫn còn một số hạn chế, nhóm tác giả đề xuất các mục tiêu trong tương lai để cải tiến và hoàn thiện hệ thống như sau:

* Nghiên cứu thêm nhiều thuật toán mới về khuyến nghị để đưa vào hệ thống.
* Thực hiện chạy mô hình kết hợp 2 hay nhiều thuật toán trên cùng một dữ liệu.
* Tự động chạy thuật toán mà không cần phải chọn.

# **TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | V. H. Tiệp, "K-nearest neighbors," 8 Jan 2017. [Online]. Available: https://machinelearningcoban.com/2017/01/08/knn/. |
| [2] | V. H. Tiệp, "K-means Clustering," 01 01 2017. [Online]. Available: https://machinelearningcoban.com/2017/01/01/kmeans/. |
| [3] | xuandungpy, "Thuật toán Apriori khai phá luật kết hợp," 09 08 2014. [Online]. Available: http://bis.net.vn/forums/p/389/683.aspx. |
| [4] | N. Nguyen, "Bài tập khai phá tập phổ biến bằng thuật toán FP-Growth," 13 08 2017. [Online]. Available: https://nhannguyen95.github.io/bai-tap-khai-pha-tap-pho-bien-bang-thuat-toan-fp-growth/. |
| [5] | N. Hug, " Matrix Factorization-based algorithms," [Online]. Available: https://surprise.readthedocs.io/en/stable/matrix\_factorization.html . |
| [6] | N. Hug, "Slope One," [Online]. Available: https://surprise.readthedocs.io/en/stable/slope\_one.html. |
| [7] | H. Võ, "Giới Thiệu Ngôn Ngữ Lập Trình Python," 11 07 2018. [Online]. Available: https://www.codehub.vn/Gioi-Thieu-Ngon-Ngu-Lap-Trinh-Python. |
| [8] | vietjack, "Node.js là gì ?," [Online]. Available: https://vietjack.com/nodejs/nodejs\_la\_gi.jsp. |
| [9] | Grinberg, Miguel, An Introduction to Flask First Steps in Web Development with Python, O'Reilly Media, 2014. |
| [10] | G. Obinna, "How to structure a Flask-RESTPlus web service for production builds," 15 4 2018. [Online]. Available: https://medium.com/free-code-camp/structuring-a-flask-restplus-web-service-for-production-builds-c2ec676de563. [Accessed 06 01 2019]. |
| [11] | R. Picard, "Blueprints," [Online]. Available: http://exploreflask.com/en/latest/blueprints.html#functional-structur. |
| [12] | P. Đ. Việt, "Khái niệm về JSON Web Token," 10 7 2016. [Online]. Available: https://techmaster.vn/posts/33959/khai-niem-ve-json-web-token. |
| [13] | T. Point, "Flask – SQLAlchemy," [Online]. Available: https://www.tutorialspoint.com/flask/flask\_sqlalchemy.htm. |
| [14] | N. V. Hiếu, "Pandas Python Tutorial," 02 10 2018. [Online]. Available: https://viblo.asia/p/huong-dan-su-dung-thu-vien-pandas-trong-python-XL6lAxaDZek. |
| [15] | K. Jain, "Scikit-learn in Python – the most important Machine Learning tool I learnt last year!," 05 01 2015. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/. |
| [16] | N. Hug, "A Python scikit for recommender systems.," [Online]. Available: http://surpriselib.com/. |
| [17] | o7planning, "Giới thiệu về ReactJS," [Online]. Available: https://o7planning.org/vi/12115/gioi-thieu-ve-react. |
| [18] | I. npm, "React Easy Chart," 2017. [Online]. Available: https://www.npmjs.com/package/react-easy-chart. |

1. Nguồn: <https://www.thongkeinternet.vn> [↑](#footnote-ref-1)
2. Nguồn: <https://www.thongkeinternet.vn> [↑](#footnote-ref-2)
3. Nguồn: <https://www.thongkeinternet.vn/> [↑](#footnote-ref-3)
4. Nguồn: <https://baodautu.vn/khai-pha-suc-manh-du-lieu-d80470.html> [↑](#footnote-ref-4)
5. Nguồn: <https://tools.ietf.org/html/rfc7519> [↑](#footnote-ref-5)
6. Nguồn: <https://viblo.asia/p/tai-sao-chung-ta-nen-dung-react-bootstrap-roaerwDOkRM> [↑](#footnote-ref-6)
7. Nguồn: [https://www.kaggle.com/shazadudwadia/supermarket](https://www.kaggle.com/shazadudwadia/supermarket?fbclid=IwAR1sH3KWjjr1kfUQ4zV-TZwziYf1l_dORLUkOk8KPJ0JxIMx2N-TRLwzhhA) [↑](#footnote-ref-7)
8. Nguồn: [https://www.kaggle.com/shazadudwadia/supermarket](https://www.kaggle.com/shazadudwadia/supermarket?fbclid=IwAR3JgCo88F6zA6g2VXMGaL9wz4VvALoSei0wPq9mybUnwOQ0S1iIMkm0viw) [↑](#footnote-ref-8)
9. Nguồn: [https://www.kaggle.com/uciml/iris](https://www.kaggle.com/uciml/iris?fbclid=IwAR13ynA9_TBvyBFaKPycurx2Fa4lN7IUZkDABLLdlu8PkZDPv1Jp82xX_oI) [↑](#footnote-ref-9)
10. Nguồn: [https://grouplens.org/datasets/movielens/](https://grouplens.org/datasets/movielens/?fbclid=IwAR1JL8OBacpFlbnoKAGzCseZ2otvbnbAF_5FqYQy-qv6eyR-klbPPdc6c4E) [↑](#footnote-ref-10)