**TRƯỜNG ĐẠI HỌC MỞ HÀ NỘI**

**Khoa Công nghệ thông tin**

****

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**NĂM 2021**

**ỨNG DỤNG KĨ THUẬT PHÂN CỤM DỮ LIỆU HỖ TRỢ SINH VIÊN LỰA CHỌN CHUYÊN NGÀNH**

**Sinh viên thực hiện**

**Nguyễn Mạnh Hùng Lớp: 1910A04 Khoa: Công nghệ thông tin**

Nguyễn Thị Dịu Lớp: 2010A03 Khoa: Công nghệ thông tin

Nguyễn Văn Tiến Lớp: 1910A04 Khoa: Công nghệ thông tin

Trần Quỳnh Trang Lớp: 1910A04 Khoa: Công nghệ thông tin

**Giảng viên hướng dẫn:** ThS. Nguyễn Thị Tâm – Khoa Công nghệ thông tin

**TRƯỜNG ĐẠI HỌC MỞ HÀ NỘI**

**Khoa Công nghệ thông tin**

****

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**NĂM 2021**

**ỨNG DỤNG KĨ THUẬT PHÂN CỤM DỮ LIỆU HỖ TRỢ SINH VIÊN LỰA CHỌN CHUYÊN NGÀNH**

**Sinh viên thực hiện**

**Nguyễn Mạnh Hùng Lớp: 1910A04 Khoa: Công nghệ thông tin**

Nguyễn Thị Diu Lớp: 2010A03 Khoa: Công nghệ thông tin

Nguyễn Văn Tiến Lớp: 1910A04 Khoa: Công nghệ thông tin

Trần Quỳnh Trang Lớp: 1910A04 Khoa: Công nghệ thông tin

**Giảng viên hướng dẫn:** ThS. Nguyễn Thị Tâm – Khoa Công nghệ thông tin

**🙠 MỤC LỤC 🙠**

[**THÔNG TIN KẾT QUẢ NGHIÊN CỨU CỦA ĐỀ TÀI** 4](#_Toc102505372)

[**MỞ ĐẦU** 6](#_Toc102505373)

[**1.** **Tổng quan tình hình nghiên cứu** 6](#_Toc102505374)

[**2.** **Lý do lựa chọn đề tài** 6](#_Toc102505375)

[**3.** **Mục tiêu đề tài** 7](#_Toc102505376)

[**4.** **Phương pháp nghiên cứu** 7](#_Toc102505377)

[**5.** **Đối tượng và phạm vi nghiên cứu** 7](#_Toc102505378)

[**6.** **Bố cục nghiên cứu** 7](#_Toc102505379)

[**KẾT QUẢ NGHIÊN CỨU VÀ PHÂN TÍCH KẾT QUẢ** 8](#_Toc102505380)

[**I.** **Cơ sở lý thuyết** 8](#_Toc102505381)

[1. Phân cụm là gì? 8](#_Toc102505382)

[2. Một số phương pháp phân cụm. 9](#_Toc102505383)

[3. Các bước thực hiện phân cụm. 11](#_Toc102505384)

[4. Ứng dụng của phân cụm dữ liệu 12](#_Toc102505385)

[5. Những vấn đề tồn tại trong phân cụm dữ liệu 12](#_Toc102505386)

[6. Giới thiệu thuật toán K-means 13](#_Toc102505387)

[6.1. Giới thiệu 13](#_Toc102505388)

[6.2. Khoảng cách Euclid 13](#_Toc102505389)

[6.3. Phần tử trọng tâm 14](#_Toc102505390)

[6.4. Thuật toán K-means. [2] 14](#_Toc102505391)

[7. Ví dụ gom cụm với thuật toán K-means 15](#_Toc102505392)

[**II.** **Cách phân cụm chuyên ngành** 18](#_Toc102505393)

[1. Các chuyên ngành và môn cơ sở chuyên ngành 18](#_Toc102505394)

[2. Tiền xử lý dữ liệu 18](#_Toc102505395)

[3. Ứng dụng thuật toán K-means trong phân cụm chuyên ngành 23](#_Toc102505396)

[4. Hiện thực hóa bằng giao diện: 24](#_Toc102505397)

[5. Nhược điểm điểm khi sử dụng thuật toán K-means cho bài toàn phân cụm chuyên ngành. 34](#_Toc102505398)

[6. Giải pháp cho nhược điểm trên là. 34](#_Toc102505399)

[**KẾT LUẬN** 35](#_Toc102505400)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 36](#_Toc102505401)

**DANH MỤC HÌNH ẢNH**

[Hình 1 Ảnh điểm môn Cơ Sở Lập Trình. 19](#_Toc102505349)

[Hình 2 Ảnh điểm môn Cơ Sở Dữ Liệu. 19](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505350)

[Hình 3 Ảnh điểm môn Kỹ Thuật Lập Trình Hướng Đối Tượng 20](#_Toc102505351)

[Hình 4 Ảnh điểm môn Thiết Kết Web 20](#_Toc102505352)

[Hình 5 Ảnh điểm môn Mạng Máy Tính 21](#_Toc102505353)

[Hình 6 Ảnh điểm môn Quản Trị Mạng 21](#_Toc102505354)

[Hình 7 Ảnh điểm môn Lập Trình Web 22](#_Toc102505355)

[Hình 8 Ảnh gộp các môn cơ sở thành một bảng 22](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505356)

[Hình 9 Ảnh đăng nhập với quyền của giảng viên 24](#_Toc102505357)

[Hình 10 Ảnh giảng viên nhập điểm môn cho sinh viên 25](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505358)

[Hình 11 Ảnh admin đăng nhập tài khoản 25](#_Toc102505359)

[Hình 12 Ảnh cửa sổ thực hiện phân cụm dữ liệu điểm sinh viên 26](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505360)

[Hình 13 Ảnh bảng điểm sinh viên qua bước tiền xử lý dữ liệu 27](#_Toc102505361)

[Hình 14 Ảnh cửa sổ lọc điểm sinh viên đủ điều kiện xét chuyên ngành 27](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505362)

[Hình 15 Ảnh Bảng chọn chuyên ngành cần phân cụm và khởi tạo cụm ban đầu 28](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505363)

[Hình 16 Ảnh chuyên ngành Công Nghệ Phần Mềm 29](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505364)

[Hình 17 Ảnh chuyên ngành Đa Phương Tiện 30](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505365)

[Hình 18 Ảnh chuyên ngành Mạng 31](file:///C:\Users\admin\Downloads\CNTT.Nguyễn_Mạnh_Hùng.ỨNG-DỤNG-KĨ-THUẬT-PHÂN-CỤM-DỮ-LIỆU-HỖ-TRỢ-SINH-VIÊN-LỰA-CHỌN-CHUYÊN-NGÀNH.docx#_Toc102505366)

[Hình 19 Ảnh phân tích dữ liệu sau phân cụm Công Nghệ Phần Mềm 32](#_Toc102505367)

[Hình 20 Ảnh đăng nhập với quyền sinh viên 33](#_Toc102505368)

[Hình 21 Ảnh gợi ý chuyên ngành phù hợp cho sinh viên khi đăng nhập tài khoản 33](#_Toc102505369)

[Hình 22 Ảnh những sinh viên 'Đạt' nhưng có đầu điểm dưới 5 34](#_Toc102505370)

[Hình 23 Ảnh những sinh viên điểm 'Đạt' nhưng có điểm dưới 5 đã cập nhập 34](#_Toc102505371)

**THÔNG TIN KẾT QUẢ NGHIÊN CỨU CỦA ĐỀ TÀI**

1. **Thông tin chung:**

**-** Tên đề tài: Ứng dụng kỹ thuật phân cụm dữ liệu hỗ trợ sinh viên lựa chọn chuyên ngành

- Sinh viên thực hiện:

Nguyên Mạnh Hùng - 1910A04 - Sinh viên năm 3

Nguyễn Thị Dịu - 2010A03 - Sinh viên năm 2

Nguyễn Văn Tiến - 1910A04 - Sinh viên năm 3

Trần Quỳnh Trang - 1910A04 - Sinh viên năm 3

- Khoa: Công nghệ thông tin Số năm đào tạo: 4 năm

- Người hướng dẫn: ThS.Nguyễn Thị Tâm

1. **Mục tiêu đề tài:**

Giúp sinh viên, giảng viên trong các trường bậc đại học có thêm cơ sở để định hướng vào các chuyên ngành sau khi trải qua quá trình học tập, tích lũy các môn cơ sở một cách chính xác, rõ ràng và hiệu quả hơn.

1. **Tính cấp thiết, tính mới:**

Việc chọn một ngành học phù hợp để theo đuổi và phát triển theo là một quyết định vô cùng quan trọng của mỗi sinh viên. Bên cạnh những sinh viên đã xác định được con đường mình muốn đi thì còn rất nhiều sinh viên chưa xác định được mình giỏi trong mảng nào và nên theo chuyên ngành gì. Như đã nói ở trên, việc chọn đúng chuyên ngành là vô cùng quan trọng nhưng nhiều sinh viên năm 3, năm 4, năm cuối vẫn chưa biết mình nên học gì, làm gì. Mà hệ lụy của việc chọn sai ngành thì vô cùng lớn, nó tác động lớn tới trường học, xã hội và nhất là chính bản thân sinh viên. Có rất nhiều nguyên nhân dẫn tới việc chọn sai chuyên ngành như: học theo phong trào, học theo bạn bè cho vui, học theo xu hướng hay theo sự quyết định của gia đình. Vậy làm sao để sinh viên có thể đưa ra sự lựa chọn đúng đắn cho mình, giảng viên cũng có cơ sở thể đưa ra lời khuyên cho sinh viên của mình. Chính là nghiên cứu và ứng dụng các kỹ thuật phù hợp để đưa ra những phương án cụ thể, có cơ sở để sinh viên có thể xác định hướng đi đúng của bản thân.

1. **Kết quả nghiên cứu:**

Đưa ra danh sách gợi ý các chuyên ngành phù hợp.

1. **Đóng góp về mặt kinh tế - xã hội,** **giáo dục và đào tạo, an ninh, quốc phòng và khả năng áp dụng của đề tài:**

Việc đưa ra được gợi ý các chuyên ngành phù hợp sẽ giúp chất lượng đào tạo nâng cao, cân bằng việc làm, tránh việc sinh viên đi làm trái ngành do không phù hợp. Bên cạnh đó, sinh viên có thể phát huy được toàn bộ những khả năng của bản thân để góp ích cho nền giáo dục của nước nhà.

1. **Công bố khoa học của sinh viên từ kết quả nghiên cứu của đề tài** *(ghi rõ họ tên tác giả, nhan đề và các yếu tố về xuất bản nếu có)* hoặc nhận xét, đánh giá của cơ sở đã áp dụng các kết quả nghiên cứu *(nếu có)*:

Ngày tháng năm

|  |  |
| --- | --- |
|  | **Sinh viên chịu trách nhiệm chính**  **thực hiện đề tài**  *(ký, họ và tên)* |

**Nhận xét của người hướng dẫn về những đóng góp khoa học của sinh viên thực hiện đề tài** *(phần này do người hướng dẫn ghi):*

Ngày tháng năm

|  |  |
| --- | --- |
|  | **Người hướng dẫn**  *(ký, họ và tên)* |

# **MỞ ĐẦU**

## **Tổng quan tình hình nghiên cứu**

Hiện nay, sinh viên lựa chọn chuyên ngành phù hợp với bản thân trong những năm cuối của đại học vô cùng quan trọng.

Để sinh viên có những cái nhìn khách quan về các chuyên ngành kèm theo các số liệu cụ thể để sinh viên có cơ sở để xác định đúng hướng đi thì việc xây dựng thuật toán giúp phân loại chuyên ngành là giải pháp tối ưu nhất. Sinh viên có thể xem điểm những học phần mình đã học, từ đó có thể đánh giá được khả năng tiếp thu và sự để tâm của sinh viên vào những môn học liên quan tới chuyên ngành đó.

Giảng viên cũng có thể dựa vào kết quả nghiên cứu để áp dụng vào việc chia sinh viên vào các chuyên ngành sao cho phù hợp, đồng thời đưa ra các phương hướng giúp sinh viên chọn đúng chuyên ngành hơn.

## **Lý do lựa chọn đề tài**

Chọn đúng ngành phù hợp với mình là điều vô cùng quan trọng và cấp thiết, có anh hưởng đến tương lai của mỗi người và cả xã hội lao động.

Việc chọn đúng nghề sẽ giúp giảm lãng phí nguồn lực, giảm thời gian công sức, đào tạo đổi nghề trong quá trình học tập, làm việc.

Việc sinh viên chọn đúng nghề còn giúp xã hội phân bổ nguồn lực đồng đều trong các lĩnh vực của xã hội. Từ đó, giúp giảm tình trạng thất nghiệp trong xã hội.

Năng suất lao động cho việc chọn đúng nghề cũng được nâng cao và phát triển không ngừng, góp phần phát triển kinh tế và xã hội cho đất nước.

Tuy nhiên việc chọn ngành của một số sinh viên hiện tạo lại chỉ mang tính chất là cảm tính. Theo thống kê của Phòng thưởng mại và công nghiệp Việt Nam, mỗi năm cả nước có 38% sinh viên mới ra trường không có định hướng nghề nghiệp cụ thể, 60% làm trái ngành. Trong khi, nguồn lao động chất lượng cao lại chưa đáp ứng đủ của các doanh nghiệp.

Vậy nên, việc đưa ra gợi ý định hướng điểm mạnh để sinh viên tham khảo sẽ giúp sinh viên có định hướng học tập, nâng cao chất lượng đào tạo, giảm tình trạng đi làm trái ngành trong thị trường lao động.

## **Mục tiêu đề tài**

Phân tích dữ liệu điểm số và đưa ra dự đoán chuyên ngành sinh viên có thể theo học và chuyên ngành không thể theo. Giúp cho các nhân sinh viên và nhưỡng người quan tâm có thêm căn cứ để đưa ra lựa chọn chuyên ngành phù hợp nhất.

## **Phương pháp nghiên cứu**

* Nghiên cứu, tìm hiểu các tài liệu về truy vấn dữ liệu và tối ưu truy vấn với ngôn ngữ SQL.
* Thực hành áp dụng trên hệ quản trị cơ sở dữ liệu SQL Server.
* Nguyên cứu, tìm hiểu thuật toán K-mean và tìm cách nâng cao độ chính xác của thuật toán.
* Tiến hành kiểm tra thực tế trên cơ sở dữ liệu SQL Server
* Xây dựng ứng dụng Winform để trình bày giới thiệu cho người dùng thực tế.

## **Đối tượng và phạm vi nghiên cứu**

* **Mục tiêu nghiên cứu:**

Giúp sinh viên lựa chọn được chuyên ngành phù hợp với khả năng bản thân.

* **Phạm vi nghiên cứu:**

3 chuyên ngành của Khoa công nghệ thông tin Đại học Mở Hà Nội.

* Công nghệ phầm mềm
* Công nghệ đa phương tiện
* Mạng

## **Bố cục nghiên cứu**

Toàn bộ nội dung báo cáo được chia thành các chương như sau:

**Chương 1: Cơ sở lý thuyết**

1. Phân cụm là gì?.
2. Một số phương pháp phân cụm.
3. Các bước để thực hiện phân cụm.
4. Ứng dụng của phân cụm dữ liệu
5. Những vấn đề tồn tại trong phân cụm dữ liệu
6. Giới thiệu thuật toán K-mean.

**Chương 2: Cách phân cụm chuyên ngành**

1. Các chuyên ngành và môn sơ sở chuyên ngành.
2. Tiền xử lý dữ liệu.
3. Ứng dụng thuật toán K-mean trong phân cụm chuyên ngành.
4. Hiện thực hóa bằng giao diện
5. Nhược điểm khi sử dụng thuật toán K – mean cho bài toán phân cụm chuyên ngành
6. Giải pháp cho nhược điểm trên.

# **KẾT QUẢ NGHIÊN CỨU VÀ PHÂN TÍCH KẾT QUẢ**

## **Cơ sở lý thuyết**

### Phân cụm là gì?

Phân tích cụm (hay phân nhóm, gom cụm, tiếng Anh: cluster analysis) là một tác vụ gom nhóm một tập các đối tượng theo cách các đối tượng cùng nhóm (gọi là cụm, cluster) sẽ có tính giống nhau (theo các đặc tính nào đó) hơn so với các đối tượng ngoài nhóm hoặc thuộc các nhóm khác.

Bản thân phân tích cụm không phải là một thuật toán cụ thể , mà là nhiệm vụ chung cần được giải quyết. Nó có thể đạt được bằng nhiều thuật toán khác nhau, khác nhau đáng kể trong hiểu biết của chúng về những gì tạo thành một cụm và cách tìm chúng một cách hiệu quả. Do đó, phân cụm có thể được xây dựng như một bài toán tối ưu hóa đa mục tiêu.

Thuật toán phân cụm phụ thuộc vào từng cá nhân tập dữ liệu và mục đích sử dụng kết quả. Phân tích cụm như vậy không phải là một nhiệm vụ tự động, mà là một quá trình lặp đi lặp lại để khám phá kiến ​​thức hoặc tối ưu hóa đa mục tiêu tương tác liên quan đến thử nghiệm và thất bại. Thông thường cần phải sửa đổi các tham số tiền xử lý dữ liệu và mô hình cho đến khi kết quả đạt được các thuộc tính mong muốn.

Phân tích cụm là một tác vụ chính của khai phá dữ liệu, và là một kỹ thuật phổ biến trong thống kê phân tích dữ liệu, được dùng trong nhiều lĩnh vực, bao gồm nhận dạng mẫu, phân tích ảnh, truy hồi thông tin, tin sinh học, nén dữ liệu, đồ họa máy tính và học máy.

### Một số phương pháp phân cụm.

Một số phương pháp phân cụm điển hình: Phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ, phân cụm dựa trên lưới, phân cụm dựa trên mô hình, phân cụm có ràng buộc, ...

* Phân cụm phân hoạch

Trong phân cụm dựa trên centroid, mỗi cụm được biểu diễn bằng một vectơ trung tâm, vectơ này không nhất thiết phải là thành viên của tập dữ liệu. Khi số lượng cụm được cố định thành k , phân cụm k -means đưa ra một định nghĩa chính thức như một bài toán tối ưu hóa: tìm k trung tâm cụm và gán các đối tượng cho trung tâm cụm gần nhất, sao cho khoảng cách bình phương từ cụm được giảm thiểu.

Hầu hết các thuật toán kiểu k -means yêu cầu số lượng cụm - k - phải được chỉ định trước, đây được coi là một trong những nhược điểm lớn nhất của các thuật toán này. Hơn nữa, các thuật toán thích các cụm có kích thước gần giống nhau, vì chúng sẽ luôn gán một đối tượng cho tâm gần nhất. Điều này thường dẫn đến việc cắt đường viền của các cụm không chính xác (điều này không có gì đáng ngạc nhiên vì thuật toán tối ưu hóa các trung tâm cụm, không phải đường viền cụm).

* Phân cụm phân cấp

Dựa trên ý tưởng cốt lõi về các đối tượng có liên quan nhiều hơn đến các đối tượng ở gần hơn là với các đối tượng ở xa hơn. Các thuật toán này kết nối các "đối tượng" để tạo thành "cụm" dựa trên khoảng cách của chúng. Một cụm có thể được mô tả phần lớn bằng khoảng cách tối đa cần thiết để kết nối các phần của cụm. Ở các khoảng cách khác nhau, các cụm khác nhau sẽ hình thành, có thể được biểu thị bằng biểu đồ hình hạt , giải thích vị trí tên chung " phân cụm phân cấp"đến từ: các thuật toán này không cung cấp một phân vùng duy nhất cho tập dữ liệu, mà thay vào đó cung cấp một hệ thống phân cấp mở rộng của các cụm hợp nhất với nhau ở những khoảng cách nhất định. Trong một biểu đồ hình ảnh, trục y đánh dấu khoảng cách mà các cụm hợp nhất , trong khi các đối tượng được đặt dọc theo trục x để các cụm không trộn lẫn.

Phân cụm phân cấp là một nhóm các phương pháp khác nhau theo cách tính khoảng cách. Ngoài sự lựa chọn thông thường về chức năng khoảng cách , người dùng cũng cần quyết định tiêu chí liên kết (vì một cụm bao gồm nhiều đối tượng, có nhiều ứng cử viên để tính khoảng cách) để sử dụng. Các lựa chọn phổ biến được gọi là phân cụm liên kết đơn (tối thiểu khoảng cách đối tượng), phân cụm liên kết hoàn chỉnh (tối đa khoảng cách đối tượng) và UPGMA hoặc WPGMA("Phương pháp nhóm cặp không trọng số hoặc có trọng số với trung bình số học", còn được gọi là phân cụm liên kết trung bình). Hơn nữa, phân cụm phân cấp có thể là gộp (bắt đầu với các phần tử đơn lẻ và tổng hợp chúng thành các cụm) hoặc phân chia (bắt đầu với tập dữ liệu hoàn chỉnh và chia nó thành các phân vùng).

* Phân cụm dựa trên mật độ

Trong phân cụm dựa trên mật độ, cụm được định nghĩa là các khu vực có mật độ cao hơn phần còn lại của tập dữ liệu. Các đối tượng ở các khu vực thưa thớt - bắt buộc phải tách các cụm - thường được coi là nhiễu và các điểm biên giới.

Phương pháp phân cụm dựa trên mật độ phổ biến nhất là DBSCAN. Trái ngược với nhiều phương pháp mới hơn, nó có mô hình cụm được xác định rõ ràng được gọi là "khả năng tiếp cận mật độ". Tương tự như phân cụm dựa trên liên kết, nó dựa trên việc kết nối các điểm trong các ngưỡng khoảng cách nhất định. Tuy nhiên, nó chỉ kết nối các điểm thỏa mãn tiêu chí mật độ, trong biến thể ban đầu được xác định là số lượng tối thiểu các đối tượng khác trong bán kính này. Một cụm bao gồm tất cả các đối tượng được kết nối với mật độ (có thể tạo thành một cụm có hình dạng tùy ý, trái ngược với nhiều phương pháp khác) cộng với tất cả các đối tượng nằm trong phạm vi của các đối tượng này. Một thuộc tính thú vị khác của DBSCAN là độ phức tạp của nó khá thấp - nó yêu cầu một số lượng tuyến tính các truy vấn phạm vi trên cơ sở dữ liệu - và nó sẽ phát hiện ra các kết quả về cơ bản giống nhau (nó mang tính xác định đối với điểm lõi và điểm nhiễu, nhưng không phải đối với điểm biên giới) trong mỗi lần chạy, do đó không cần phải chạy nó nhiều lần.

* Phân cụm dựa trên lưới

Kỹ thuật phân cụm dựa trên lưới thích hợp với dữ liệu nhiều chiều, dựa trên cấu trúc dữ liệu lưới để phân cụm, phương pháp này chủ yếu tập trung áp dụng cho lớp dữ liệu không gian. Mục tiêu của phương pháp này là lượng hóa dữ liệu thành các ô tạo thành cấu trúc dữ liệu lưới. Sau đó, các thao tác phân cụm chỉ cần làm việc với các đối tượng trong từng ở trên lưới chứ không phải các đối tượng dữ liệu. Cách tiếp cận dựa trên lưới này không di chuyền các đối tượng trong các ô mà xây dựng nhiều mức phân cấp của nhóm các đối tượng trong một ô. Phương pháp này gần giống với phương pháp phân cụm phân cấp nhưng chúng không trộn các ô, đồng thời giai quyết khác phục yêu cầu đối với dữ liệu nhiều chiều mà phương pháp phân phân cụm dựa trên mật độ không giải quyết được ưu điểm của phương pháp phân cụm dựa trên lưới là thời gian xử lí nhanh và độc lập với số đối tượng dữ liệu trong tập dữ liệu ban đầu, thay vào đó là chúng phụ thuộc vào số ở trong mỗi chiều không gian lưới.

* Phân cụm dựa trên mô hình

Phân cụm dựa trên mô hình là một công cụ phổ biến nổi tiếng với các nền tảng xác suất và tính linh hoạt của nó. Tuy nhiên, dữ liệu chiều cao ngày nay ngày càng thường xuyên hơn và, thật không may, các kỹ thuật phân nhóm dựa trên mô hình cổ điển cho thấy một hành vi đáng thất vọng trong không gian chiều cao. Điều này chủ yếu là do các phương pháp phân cụm dựa trên mô hình được tham số quá mức trong trường hợp này. Tuy nhiên, không gian chiều cao có những đặc điểm cụ thể hữu ích cho việc phân cụm và các kỹ thuật gần đây khai thác những đặc điểm đó. Sau khi nhớ lại các cơ sở của phân cụm dựa trên mô hình, các phương pháp tiếp cận giảm thứ nguyên, kỹ thuật dựa trên chính quy hóa, mô hình phân tích cú pháp, phương pháp phân cụm không gian con và phương pháp phân cụm dựa trên lựa chọn biến được xem xét.

Ở bài toán này cúng ta chọn phương pháp phân cụm phân hoạch nhằm gom cụm các sinh viên theo chuyên ngành phù hợp dựa trên điểm số một số học phần đã tích lũy được trong thời gian học tập.

### Các bước thực hiện phân cụm.

Chọn đặc trưng: Các đặc trưng lựa chọn phải hợp lý để có thể mã hóa nhiều nhất các thông tin liên quan đến đối tượng được quan tâm.

Chọn độ đo gần nhất: Một độ đo chỉ ra mức độ tương tự hay không tương tự giữa hai vector đặc trưng.

Tiêu chuẩn phân cụm: tiêu chuẩn phân cụm có thể được biểu diễn bởi hàm chi phí hoặc một vài quy tắc khác.

Công nhận kết quả: sau khi có kết quả phân cụm, cần kiểm tra tính đúng đắn của nó.

Giải thích kết quả: bằng kết quả thực nghiệm cần phân tích để đưa ra kết quả đúng đắn.

### Ứng dụng của phân cụm dữ liệu

Phân cụm dữ liệu có thể ứng dụng trong nhiều lĩnh vực như:

* Thương mại tìm kiếm nhóm các khách hàng quan trọng dựa vào các thuộc tính đặc trưng tương đồng và những đặc tả của họ trong các bản ghi mua bán của cơ sở dữ liệu.
* Sinh học phân loại động, thực vật qua các chức năng gen trong đồng của chúng.
* Thư viện phân loại các cụm sách có nội dung và ý nghĩa trong đồng nhau để cung cấp cho độc giả cũng như đặt hàng với nhà cung cấp
* Bảo hiểm nhận dạng nhóm tham gia bảo hiểm có chi phi yêu cầu bồi thưởng trung bình cao, xác định gian lận trong bảo hiểm thông qua các mẫu cả biệt.
* Quy hoạch đô thị : nhận dạng các nhóm nhà theo kiểu, vị trí địa lí, giá trị …. nhằm cung cấp thông tin cho quy hoạch đô thị,
* Nghiên cứu địa chấn : phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho việc nhận dạng các vùng nguy hiểm
* Tài liệu phân loại, phân nhóm dữ liệu weblog để khám phá các nhóm về các hình thức tiếp cận tương tự trợ giúp cho việc khai phá thông tin từ dữ liệu.

### Những vấn đề tồn tại trong phân cụm dữ liệu

* Kỹ thuật clustering hiện nay không trình bày được tất cả các yêu cầu đầy đủ (và đồng thời);
* Giao dịch với số lượng lớn các mẫu và số lượng lớn các mẫu tin của dữ liệu có thể gặp vấn đề phức tạp về thời gian
* Hiệu quả của phương pháp phụ thuộc vào định nghĩa của khoảng cách" (đối với phân cụm dữ liệu dựa trên khoảng cách) Nếu không tồn tại một thước đo khoảng cách rõ ràng chúng ta “phải tự xác định” một điều mà không thật sự dễ dàng chút nào, nhất là trong không gian đã chiều:
* Kết quả của thuật toán phân cụm dữ liệu có thể được giải thích theo nhiều cách khác nhau (mà trong nhiều trường hợp chỉ có thể được giải thich theo ý riêng của mỗi người).

### Giới thiệu thuật toán K-means

#### **Giới** thiệu

K-mean là kỹ thuật rất quan trọng và được sử dụng phổ biến trong ký thuật phân cụm dữ liệu [1]. Thuật tán này tìm cách phân cụm các đối tượng đã cho vào k cụm (k là số là số xác định trước, k là số nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm cụm (centroid) là nhỏ nhất. Về nguyên lý, có n đối tượng, mỗi đối tượng có m thuộc tính, các đối tượng được phân chia thành k cụm dựa trên thuộc tính của đối tượng bằng cách sử dụng thuật toán K-mean. Bài toán này xem mỗi thuộc tính của đối tượng (đối tượng có m thuộc tính) như một tọa độ của không gian m chiều và biểu diễn đối tượng và biểu diễn đối tượng như một điểm trong không gian m chiều, đó là:

***ai = (xi1, xi2, xi3, xi4, …)*** (1)

Trong đó:

ai (i=1, 2, …, n) là đối tượng thứ i.

xij (i=1, 2, …, n, j=1, 2, …, m) là thuộc tính thứ j của đối tượng i.

#### Khoảng cách Euclid

Phương pháp phân cụm dữ liệu thực hiện dựa trên khoảng cách Euclid là khoảng cách nhỏ nhất từ đối tượng đến phần tử trọng tâm của các cụm. Phần tử trọng tâm của cụm được xác định bằng giá trị trung bình các phần tử trong cụm.

ai = (xi1, xi2 , xim); i = 1..n là đối tượng thứ i cần phân cụm.

C = (xj1, xj2, xjm); j = 1..k là phần tử trọng tâm cụm j

Khoảng cách Euclid từ đối tượng a, đến phần tử trọng tâm nhóm j; c; được tính toán dựa trên công thức:

(2)

Trong đó:

: Khoảng cách Euclid thừ ai đến cj.

xis: Thuộc tính thứ s của đối tượng ai

xjs: Thuộc tính thứ s của phần tử trọng tâm cj

#### Phần tử trọng tâm

K phần tử trọng tâm ban đầu được chọn ngẫu nhiên, sau mỗi lần gom các đối tượng vào các cụm, phần tử trọng tâm được tính toán lại:

ci = {a1, a2, a3, …, at} – cụm thứ i

i = 1…k, k là số cụm

j = 1…m, m là số thuộc tính.

t là số phần tử hiện có của nhóm thứ i

xsj thuộc tính thứ j của phần tử s; s = 1…t

cij phần tử thứ j của phần tử trung tâm cụm i;

(3)

#### Thuật toán K-means. [2]

**Input:** Số cụm k và các trọng tâm cụm {𝑚𝑖}; 𝑘𝑖 = 1

**Output:** Các cụm [𝑖] (1 ≤ 𝑖 ≤ 𝑘) và hàm tiêu chuẩn E đạt giá trị tối thiểu.

**Bước 1**: Khởi tạo

Chọn k trọng tâm {𝑚𝑗} (1 ≤ 𝑗 ≤ 𝑘), ban đầu trong không gian Rd (d là số chiều của dữ liệu). Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm.

**Bước 2:** Tính khoảng cách

Đối với mỗi điểm x𝑖 (1 ≤ 𝑖 ≤ 𝑛), tính khoảng cách của nó tới mỗi trọng tâm {𝑚𝑗} (1 ≤ 𝑗 ≤ 𝑘). Sau đó tìm trọng tâm gần nhất đối với mỗi điểm.

**Bước 3**: Cập nhật lại trọng tâm

Đối với mỗi 1 ≤ 𝑗 ≤ 𝑘, cập nhật trọng tâm cụm mj bằng cách xác định trung bình cộng các vectơ đối tượng dữ liệu.

**Bước4:** Lặp lại các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

### Ví dụ gom cụm với thuật toán K-means

Đề bài: Có 20 học sinh với 3 đầu điểm Toán, Lý, Hóa. Yêu cầu gom thành 3 cụm

Bảng điểm học sinh

|  |  |  |  |
| --- | --- | --- | --- |
| Mã HS | Toán | Lý | Hóa |
| A01 | 6 | 5 | 10 |
| A02 | 9 | 10 | 8 |
| A03 | 8 | 3 | 7 |
| A04 | 6 | 4 | 3 |
| A05 | 8 | 5 | 4 |
| A06 | 6 | 5 | 9 |
| A07 | 7 | 8 | 8 |
| A08 | 5 | 4 | 6 |
| A09 | 7 | 3 | 4 |
| A10 | 9 | 10 | 7 |
| A11 | 2 | 5 | 3 |
| A12 | 8 | 8 | 6 |
| A13 | 3 | 10 | 3 |
| A14 | 3 | 10 | 10 |
| A15 | 7 | 6 | 5 |
| A16 | 10 | 7 | 3 |
| A17 | 7 | 4 | 4 |
| A18 | 7 | 9 | 5 |
| A19 | 4 | 10 | 7 |
| A20 | 1 | 3 | 4 |

Trọng tâm cụm lần 1: Chọn ngẫu nhiên (ở đây ta chọn vị trí 1, 9, 20)

Cụm 1, khởi tạo trọng tâm C1 (6,5,10).

Cụm 2, khởi tạo trọng tâm C2 (7,3,4).

Cụm 3, khởi tạo trọng tâm C3 (1,3,4).

Tính Khoảng cách Euclid và phân cụm lần 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mã HS | Euclid C1 | Euclid C2 | Euclid C3 | Lần 1 | | |
| Gần C1 | Gần C2 | Gần C3 |
| A01 | 0.0 | 6.4 | 8.0 | x |  |  |
| A02 | 6.1 | 8.3 | 11.3 | x |  |  |
| A03 | 4.1 | 3.1 | 7.6 |  | x |  |
| A04 | 7.0 | 1.7 | 5.1 |  | x |  |
| A05 | 6.3 | 2.2 | 7.2 |  | x |  |
| A06 | 1.0 | 5.4 | 7.3 | x |  |  |
| A07 | 3.7 | 6.4 | 8.7 | x |  |  |
| A08 | 4.2 | 3.0 | 4.5 |  | x |  |
| A09 | 6.4 | 0.0 | 6.0 |  | x |  |
| A10 | 6.5 | 7.8 | 11.0 | x |  |  |
| A11 | 8.0 | 5.4 | 2.4 |  |  | x |
| A12 | 5.3 | 5.4 | 8.8 | x |  |  |
| A13 | 9.1 | 8.1 | 7.3 |  |  | x |
| A14 | 5.8 | 10.0 | 9.4 | x |  |  |
| A15 | 5.1 | 3.1 | 6.7 |  | x |  |
| A16 | 8.3 | 5.0 | 9.8 |  | x |  |
| A17 | 6.1 | 1.0 | 6.0 |  | x |  |
| A18 | 6.4 | 6.0 | 8.5 |  | x |  |
| A19 | 6.1 | 8.1 | 8.1 | x |  |  |
| A20 | 8.0 | 6.0 | 0.0 |  |  | x |

Trọng tâm cụm lần 2: Tính lại trọng tâm cụm với các phần tử cụm sau lần 1

Cụm 1, khởi tạo trọng tâm C1 (6.5, 8.2, 8.1).

Cụm 2, khởi tạo trọng tâm C2 (7.2, 5.0, 4.5).

Cụm 3, khởi tạo trọng tâm C3 (2.0, 6.0, 3.3).

Tính Khoảng cách Euclid và phân cụm lần 2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mã HS | Euclid C1 | Euclid C2 | Euclid C3 | Lần 2 | | |
| Gần C1 | Gần C2 | Gần C3 |
| A01 | 3.7 | 56 | 78 | x |  |  |
| A02 | 3..0 | 6.3 | 9.3 | x |  |  |
| A03 | 5.5 | 3.3 | 7.6 |  | x |  |
| A04 | 6.6 | 2.1 | 4.4 |  | x |  |
| A05 | 54 | 0.9 | 6.1 |  | x |  |
| A06 | 3.3 | 4.6 | 7.0 | x |  |  |
| A07 | 0.5 | 4.6 | 7.1 | x |  |  |
| A08 | 4.9 | 2.8 | 4.5 |  | x |  |
| A09 | 6.6 | 2.0 | 5.8 |  | x |  |
| A10 | 3.2 | 5.8 | 8.8 | x |  |  |
| A11 | 7.5 | 5.4 | 1.0 |  |  | x |
| A12 | 2.5 | 3.4 | 6.8 | x |  |  |
| A13 | 67 | 6.7 | 4.1 |  |  | x |
| A14 | 4.3 | 8.5 | 7.8 | x |  |  |
| A15 | 3.8 | 1.1 | 5.2 |  | x |  |
| A16 | 6.3 | 3.7 | 8.0 |  | x |  |
| A17 | 5.8 | 1.1 | 5.4 |  | x |  |
| A18 | 3.2 | 4.0 | 6.0 | x |  |  |
| A19 | 3.2 | 6.4 | 5.8 | x |  |  |
| A20 | 6.8 | 6.5 | 3.2 |  |  | x |

Trọng tâm cụm lần 3: Tính lại trọng tâm cụm với các phần tử cụm sau lần 2

Cụm 1, khởi tạo trọng tâm C1 (6.5, 8.3, 7.7).

Cụm 2, khởi tạo trọng tâm C2 (7.2, 4.5., 4.5).

Cụm 3, khởi tạo trọng tâm C3 (2.0, 6.0, 3.3).

Tính Khoảng cách Euclid và phân cụm lần 3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mã HS | Euclid C1 | Euclid C2 | Euclid C3 | Lần 2 | | |
| Gần C1 | Gần C2 | Gần C3 |
| A01 | 4.0 | 5.6 | 7.8 | x |  |  |
| A02 | 3.0 | 6.7 | 9.3 | x |  |  |
| A03 | 5.5 | 3.0 | 7.6 |  | x |  |
| A04 | 6.3 | 1.9 | 4.4 |  | x |  |
| A05 | 5.1 | 1.0 | 6.1 |  | x |  |
| A06 | 3.5 | 4.6 | 7.0 | x |  |  |
| A07 | 0.9 | 4.9 | 7.1 | x |  |  |
| A08 | 4.8 | 2.7 | 7.5 |  | x |  |
| A09 | 6.4 | 1.5 | 5.8 |  | x |  |
| A10 | 3.1 | 6.3 | 8.8 | x |  |  |
| A11 | 7.2 | 5.4 | 1.0 |  |  | x |
| A12 | 2.2 | 3.8 | 6.8 | x |  |  |
| A13 | 6.1 | 7.0 | 4.1 |  |  | x |
| A14 | 4.5 | 8.8 | 7.8 | x |  |  |
| A15 | 3.5 | 1.5 | 5.2 |  | x |  |
| A16 | 6.0 | 4.0 | 8.0 |  | x |  |
| A17 | 5.6 | 0.7 | 5.4 |  | x |  |
| A18 | 2.8 | 4.5 | 6.0 | x |  |  |
| A19 | 3.1 | 6.8 | 5.8 | x |  |  |
| A20 | 8.4 | 6.3 | 3.2 |  |  | x |

Trọng tâm cụm lần 4: Tính lại trọng tâm cụm với các phần tử cụm sau lần 3

Cụm 1, khởi tạo trọng tâm C1 (6.5, 8.3, 7.7).

Cụm 2, khởi tạo trọng tâm C2 (7.2, 4.5., 4.5).

Cụm 3, khởi tạo trọng tâm C3 (2.0, 6.0, 3.3).

(\*) Ta thấy tính lại sau lần 3 không có thay đổi so với lần 2, ta dừng lại thuật toán.

🡪 Vậy:

Cụm 1 (6.5, 8.3, 7.7) gồm 9 học sinh (A1, A2, A6, A7, A10, A12, A14, A18, A18).

Cụm 2 (7.2, 4.5, 4.5) gồm 8 học sinh (A3, A4, A5, A8, A9, A15, A16, A17).

Cụm 3 (2.0, 6.0, 3.3) gồm 3 học sinh (A11, A13, A20).

## **Cách phân cụm chuyên ngành**

### Các chuyên ngành và môn cơ sở chuyên ngành

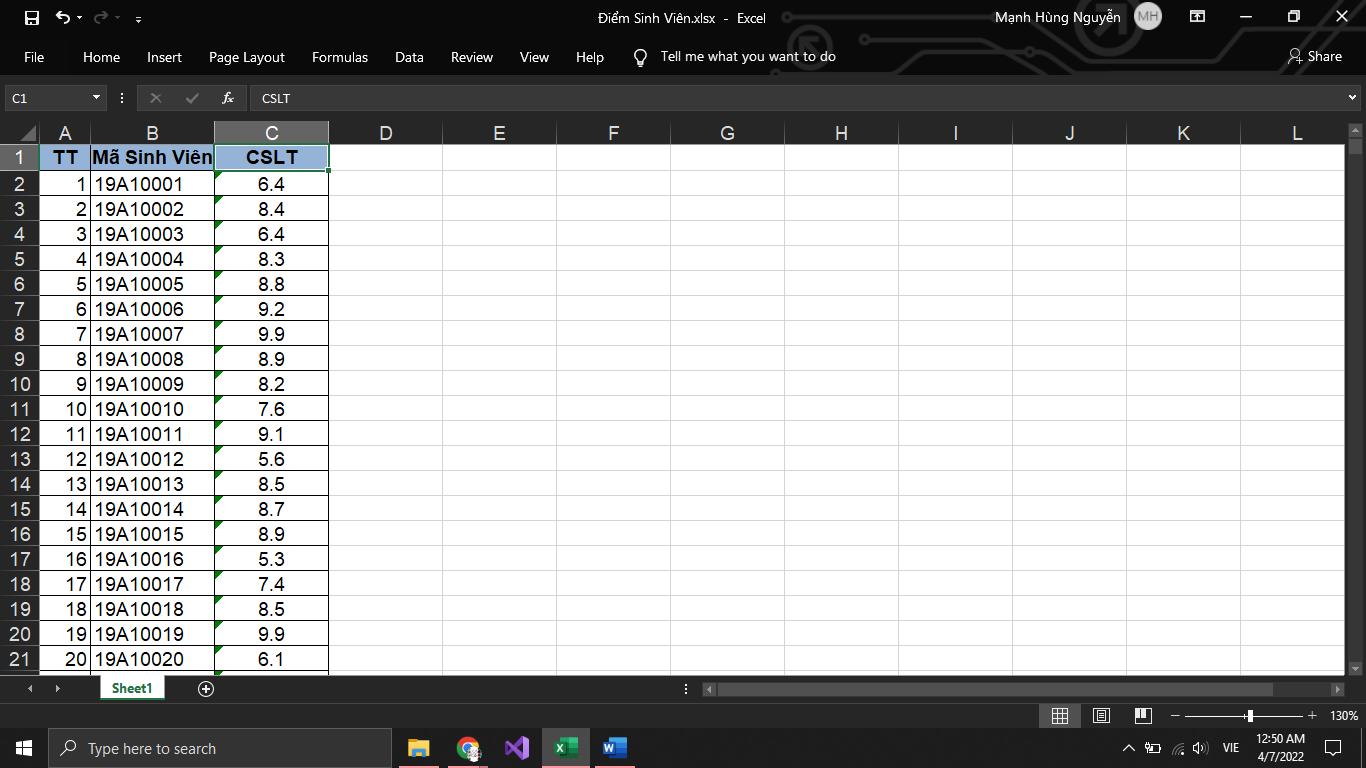
Chương trình đào tạo hệ đại học chính quy ngành Công nghệ thông tin được chia làm 4 chuyên ngành, nhưng từ trước đến giờ chỉ có 3 chuyên ngành được sinh viên đăng ký theo học gồm: Công nghệ phần mềm (CNPM), Đa phương tiện (DPT) và Chuyên ngành Mạng.

Mỗi chuyên ngành gồm những học phần chuyên sâu thể hiện khối kiến thức đặc thù của chuyên ngành đó. Định hướng chuyên ngành cho sinh viên một cách tự động căn cứ vào kết quả học tập những học phần có kiến thức bổ trợ cho từng chuyên ngành. Các học phần này được chọn theo từng chuyên ngành như sau:

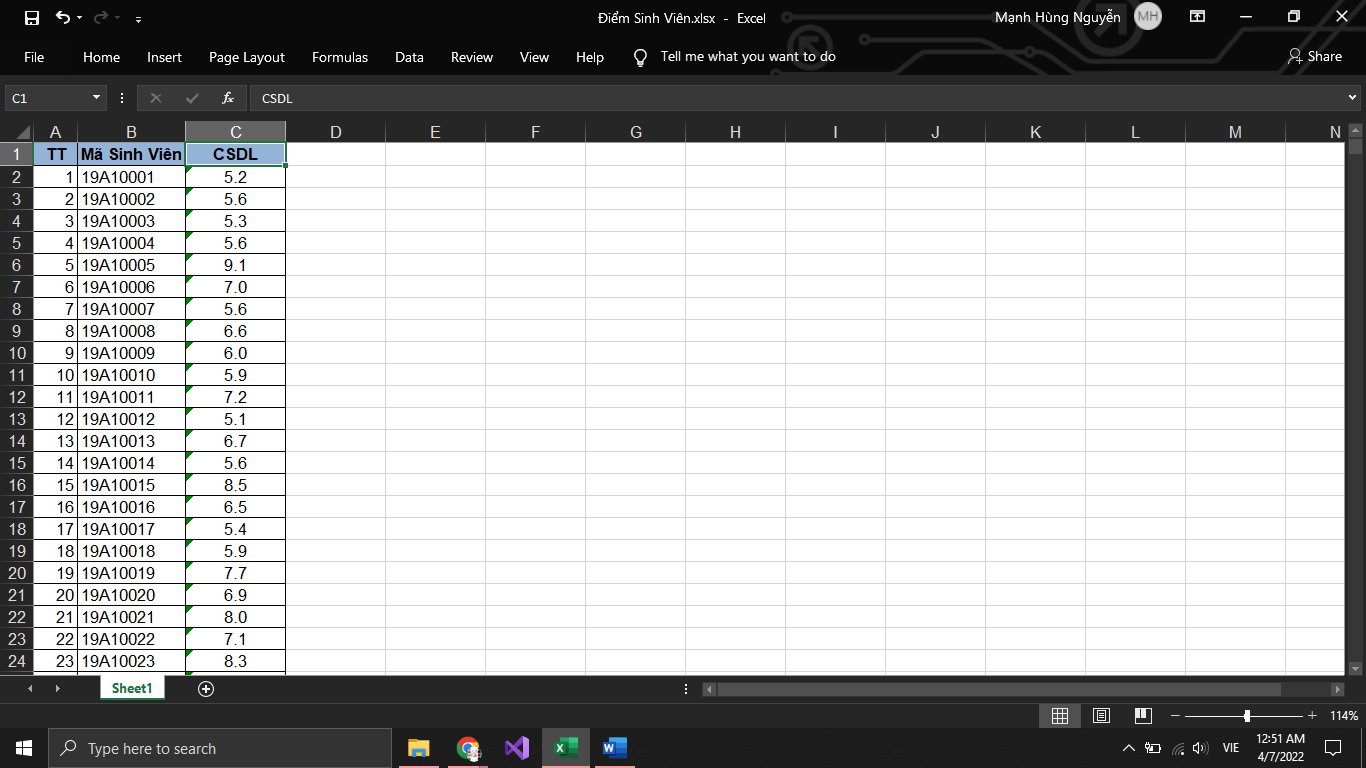
* Chuyên ngành công nghệ phần mềm gồm: Cơ sở dữ liệu, Cơ sở lập trình, Lập trình hướng đối tượng.
* Chuyên ngành đa phương tiện gồm: Cơ sở lập trình Thiết kế web, Lập trình web.
* Chuyên ngành mạng: Cơ sở lập trình, Mạng máy tính, Quản trị mạng.

### Tiền xử lý dữ liệu

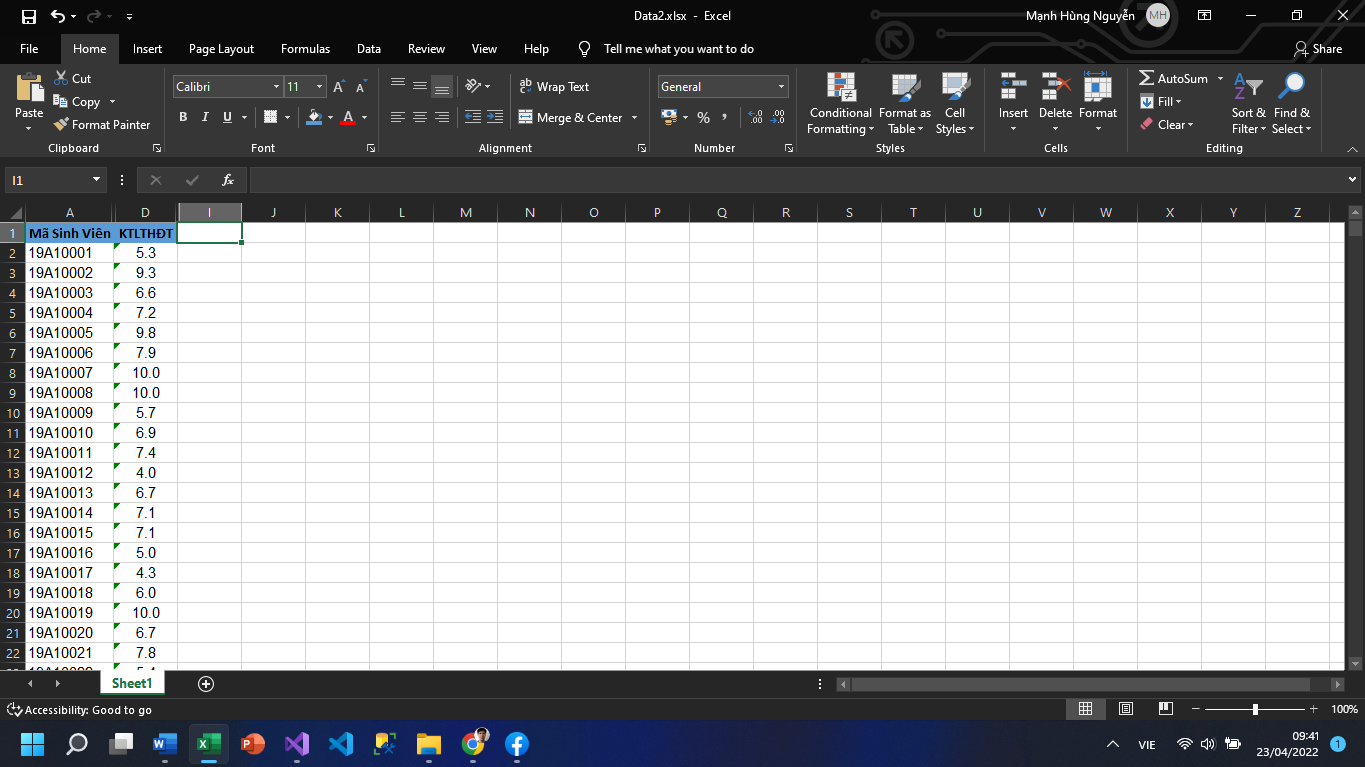
Dữ liệu thu thập ban đầu là các tập tin excel chứa thông tin điểm học tập của sinh viên. Mỗi tập tin chứa thông tin điểm số các môn trong một học kỳ được tổ chức thành bảng có nhiều cột và dòng, trong đó là kết quả học tập của sinh viên trong học kỳ đó.



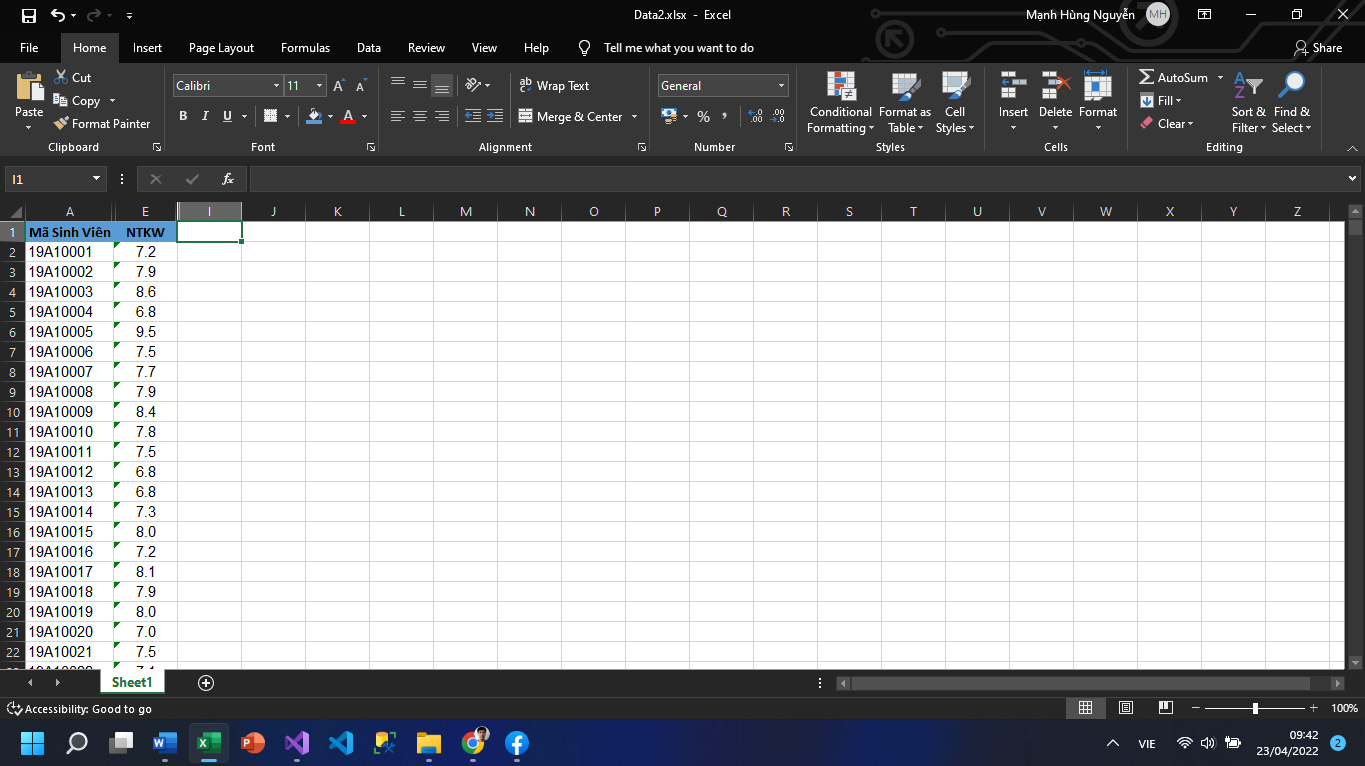
Hình 1 Ảnh điểm môn Cơ Sở Lập Trình.



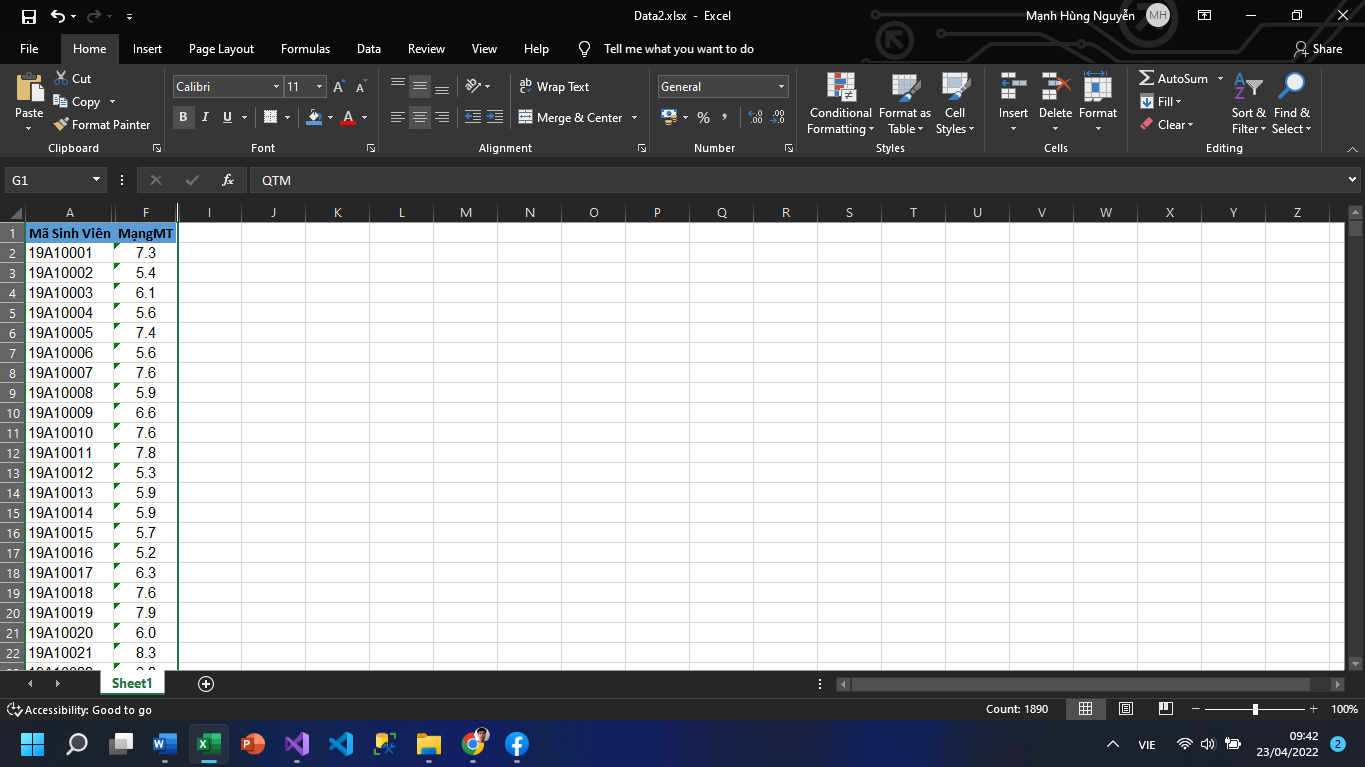
Hình 2 Ảnh điểm môn Cơ Sở Dữ Liệu.



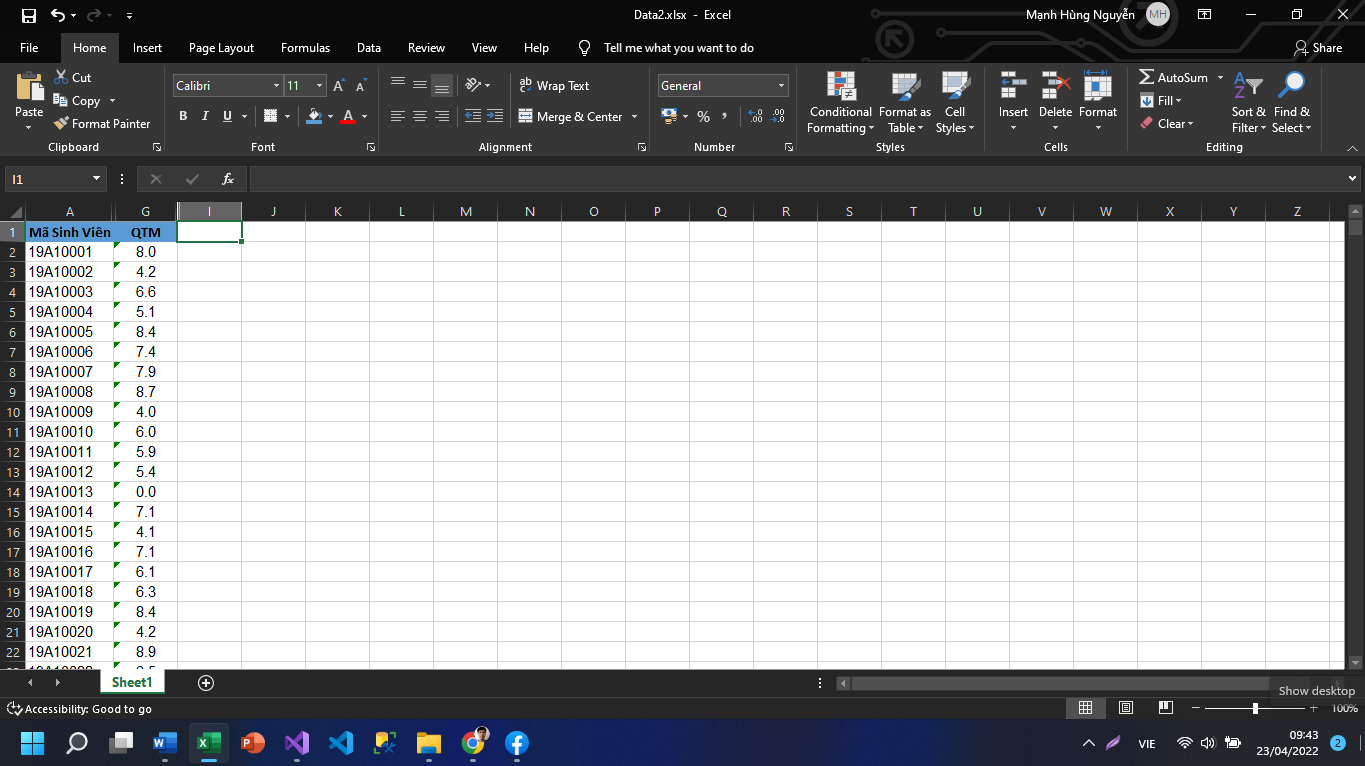
Hình 3 Ảnh điểm môn Kỹ Thuật Lập Trình Hướng Đối Tượng



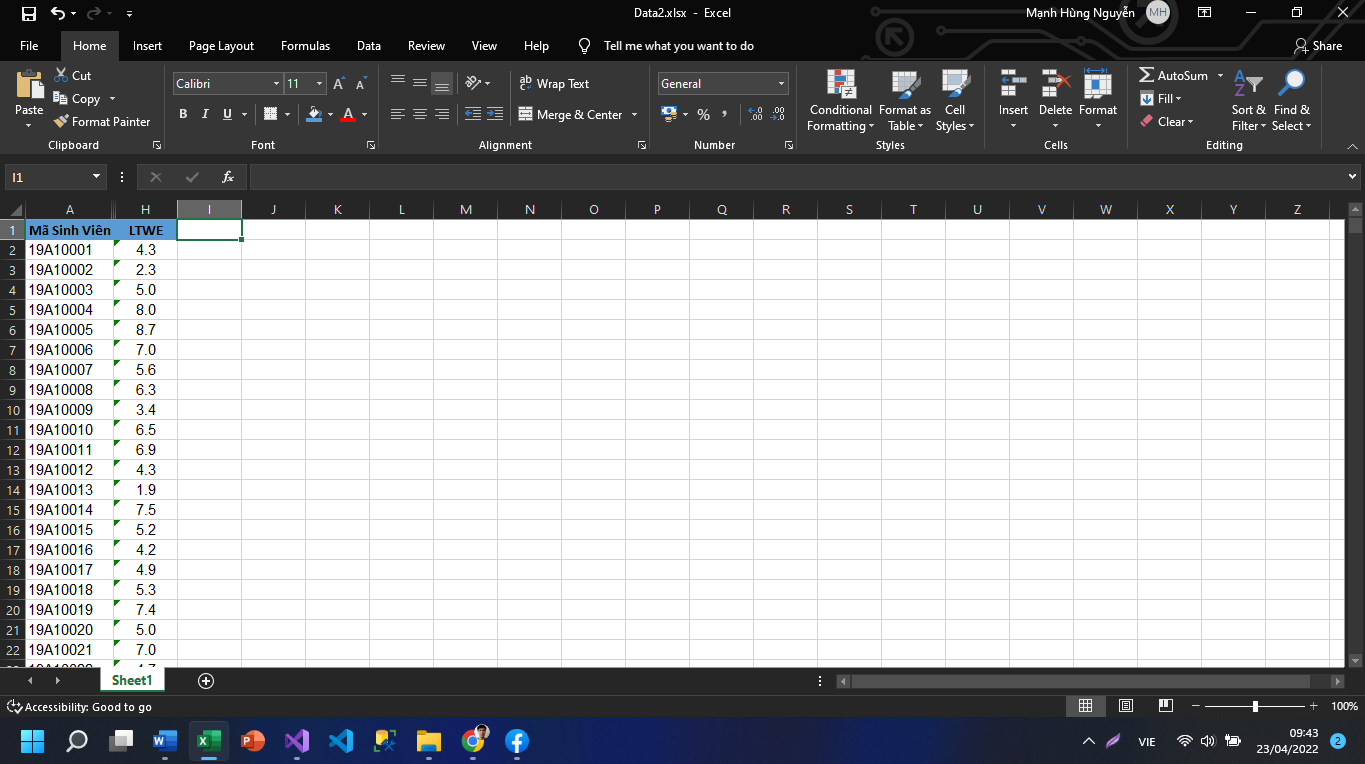
Hình 4 Ảnh điểm môn Thiết Kết Web



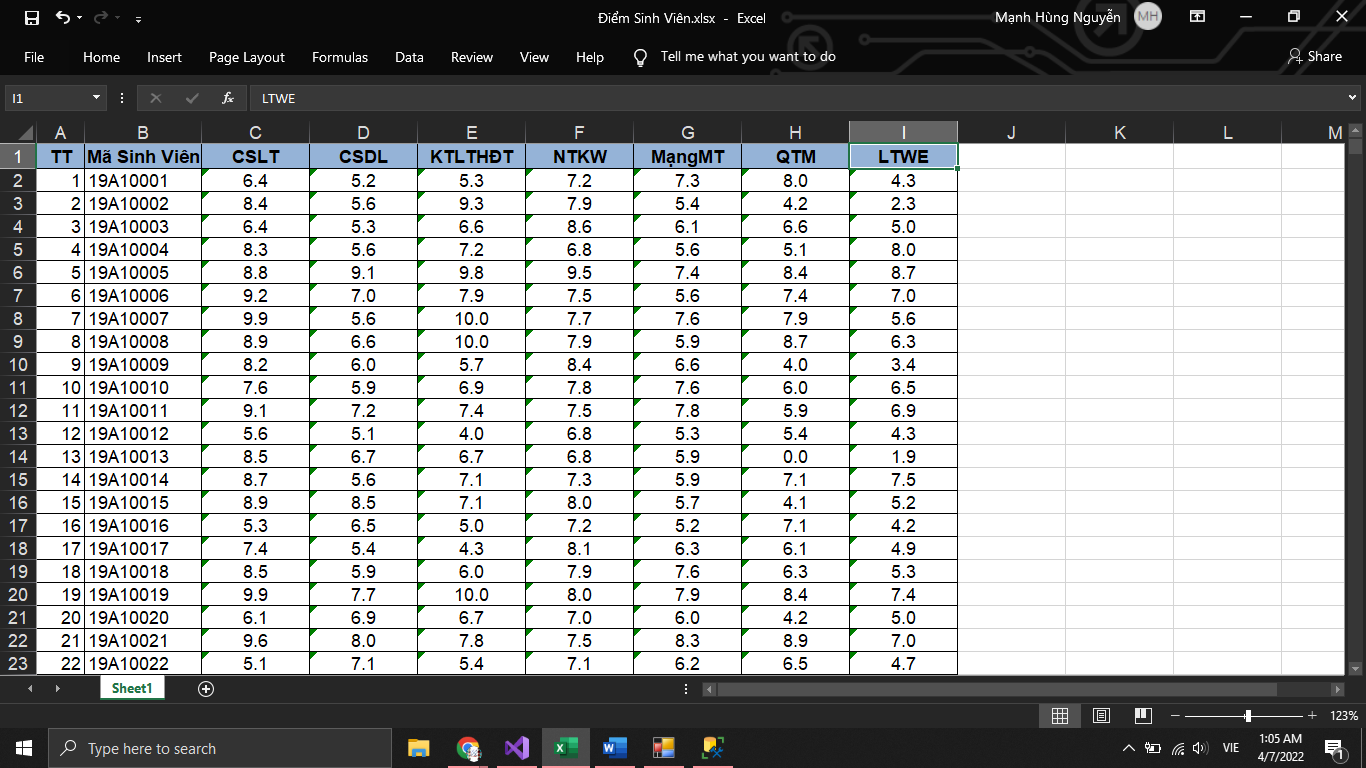
Hình 5 Ảnh điểm môn Mạng Máy Tính



Hình 6 Ảnh điểm môn Quản Trị Mạng



Hình 7 Ảnh điểm môn Lập Trình Web

Do mỗi tập tin excel chỉ chứa thông tin về điểm của một số môn học nên cần thực hiện tổng hợp dữ liệu từ nhiều tập tin, sau đó loại bỏ các môn học chung, chỉ giữ lại các môn học cơ sở ngành có kiến thức bổ trợ cho từng chuyên ngành nhằm phục vụ cho bài toán phân cụm chuyên ngành.

Hình 8 Ảnh gộp các môn cơ sở thành một bảng

### Ứng dụng thuật toán K-means trong phân cụm chuyên ngành

Thuật toán K-Means được áp dụng phân cụm cho từng chuyên ngành, mỗi chuyên  
ngành có kết quả là 2 cụm, một cụm gồm các sinh viên có khả năng theo học chuyên ngành đó và cụm còn lại là sinh viên không có khả năng học, lặp lại thuật toán cho đến hết các chuyên ngành. Trọng tâm ban đầu của mỗi cụm trong từng chuyên ngành được chỉ định với 2 mức gọi là ngưỡng trên và ngưỡng dưới. Ngưỡng trên sẽ gom cụm gồm những sinh viên có khả năng học chuyên ngành, ngưỡng dưới gom cụm gồm những sinh viên không có khả năng học chuyên ngành.

Thuật toán K-means áp dụng cho bài toán:

Input:

* Bảng điểm sinh viên tổng hợp đã qua bước tiền xử lý.
* Lọc sinh viên đủ điều kiện để xét chuyên ngành
* Trọng tâm của 2 cụm ứng với ngưỡng trên (có khả năng) và ngưỡng dưới (không có khả năng).

Output: Danh sách sinh viên được phân cụm theo từng chuyên ngành và danh sách sinh viên không thuộc chuyên ngành nào

Các bước thực hiện:

**Bước 1**: Khởi tạo

Trọng tâm ban đầu theo từng chuyên ngành 𝐶ki(c0, 𝑐1, … , 𝑐𝑚); trong đó ki là trọng tâm của cụm i thuộc chuyên ngành k; m là điểm số thứ m thuộc trọng tâm 𝐶ki

Danh sách điểm số 𝐷ki (𝑑0, 𝑑1, … , 𝑑𝑚); trong đó ki là nhóm điểm thứ i thuộc chuyên ngành k; m là điểm số thứ m thuộc nhóm điểm 𝐷ki

**Bước 2**: Phân cụm cho mỗi chuyên ngành

RQ = K-means(Q); // xử lý phân cụm cho chuyên ngành Q

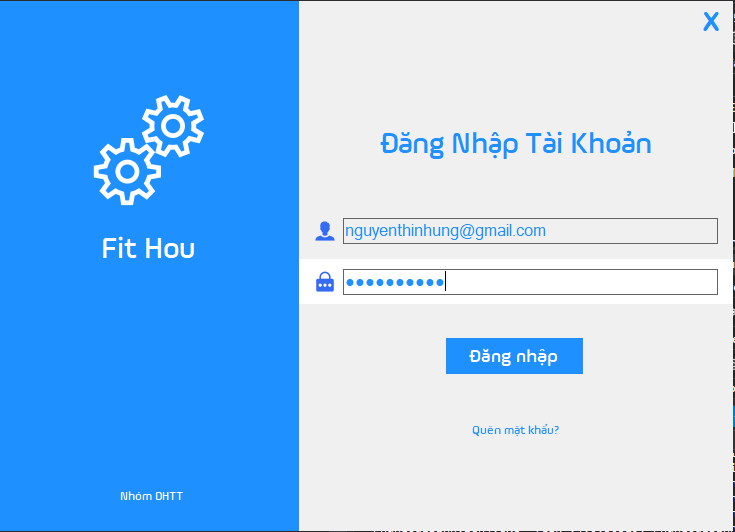
// RQ gồm 2 tập: RQ0 là tập các sinh viên thuộc vào chuyên ngành Q, RQ1 là tập hợp các sinh viên không thuộc chuyên ngành Q.

**Bước 3**: Xử lý kết quả

Hiện kết quả thống kê sinh viên mỗi tập chuyên ngành

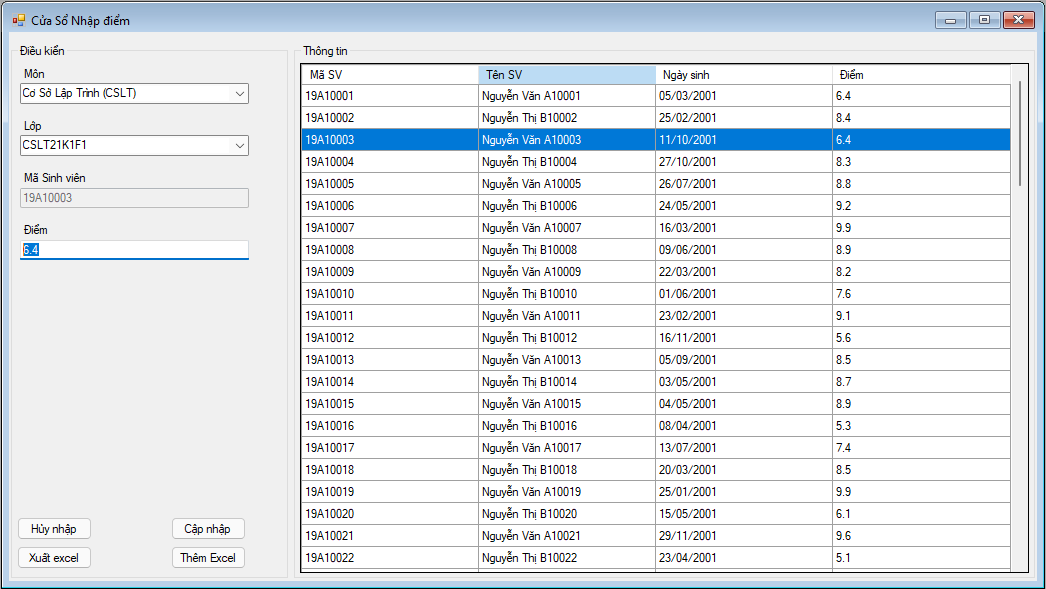
### Hiện thực hóa bằng giao diện:

* Giảng viên đăng nhập vào để quản lý sinh viên lớp hành chính mình quản lý và nhập điểm cho sinh viên theo môn học:



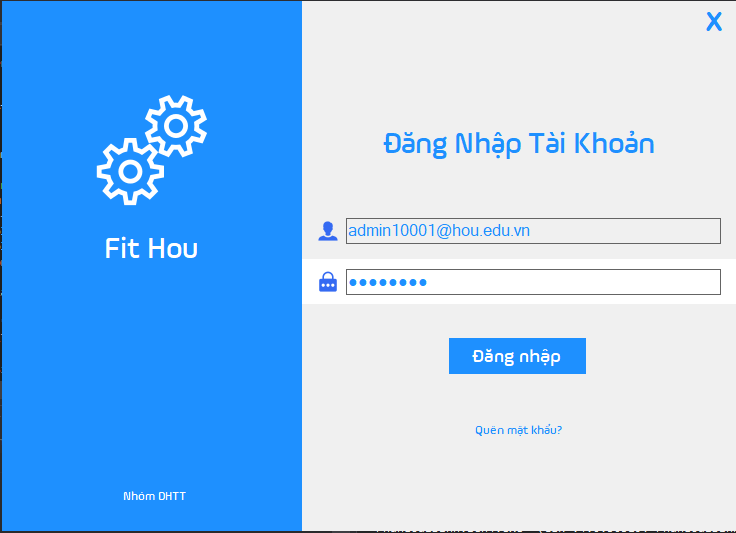
Hình 9 Ảnh đăng nhập với quyền của giảng viên

* Giảng viên nhập điểm cho sinh viên:

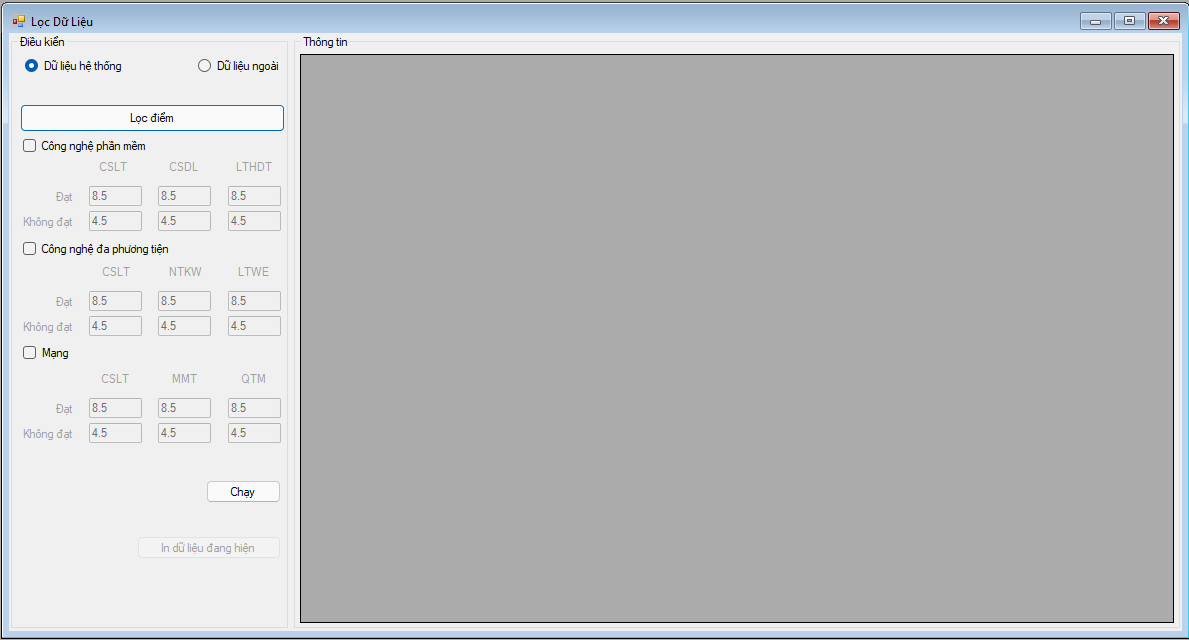


Hình 10 Ảnh giảng viên nhập điểm môn cho sinh viên

* Đăng nhập với tài khoản Admin để có quyền thực hiện phân cụm chuyên ngành:



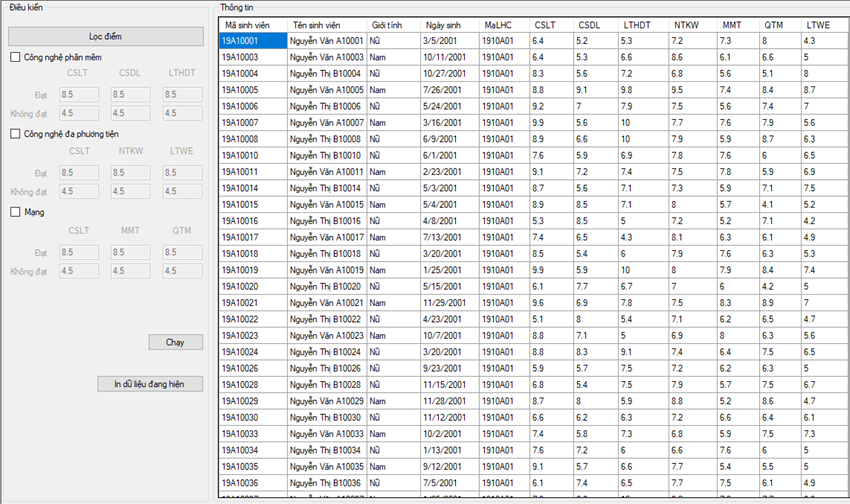
Hình 11 Ảnh admin đăng nhập tài khoản

* Vào cửa sổ chuyên ngành để tiến hành phân cụm

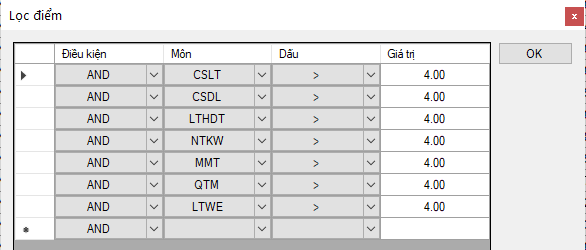
Hình 12 Ảnh cửa sổ thực hiện phân cụm dữ liệu điểm sinh viên

* Tổng hợp điểm của sinh viên sau bước tiền xử lý:

Với những sinh viên có điểm thi các môn Cơ sở lập trình (CSLT), Cơ sở dữ liệu (CSDL), Lập trình hướng đối tượng (LTHDT), Lập trình web (LTWE), Thiết kế web (NTKW), Mạng máy tính (MMT), Quản trị mạng (QTM) chưa có điểm hoặc điểm dưới 4 thì điều kiện xét chuyên ngành là không đạt sẽ được loại bỏ sinh viên đó ra khỏi dữ diệu cần được xử lý.



Hình 13 Ảnh bảng điểm sinh viên qua bước tiền xử lý dữ liệu



Hình 14 Ảnh cửa sổ lọc điểm sinh viên đủ điều kiện xét chuyên ngành

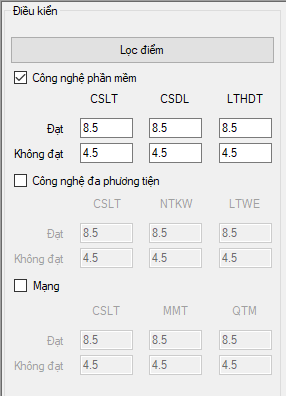
Trọng tâm của 2 cụm ứng với ngưỡng trên (có khả năng) và ngưỡng dưới (không có khả năng).

Chúng ta sẽ khởi tạo 2 cụm ban đầu với hai ngưỡng:

+ Ngưỡng trên được khởi tạo cụm ban đầu là đạt

+ Ngưỡng dưới được khởi tạo cụm ban đầu là không đạt

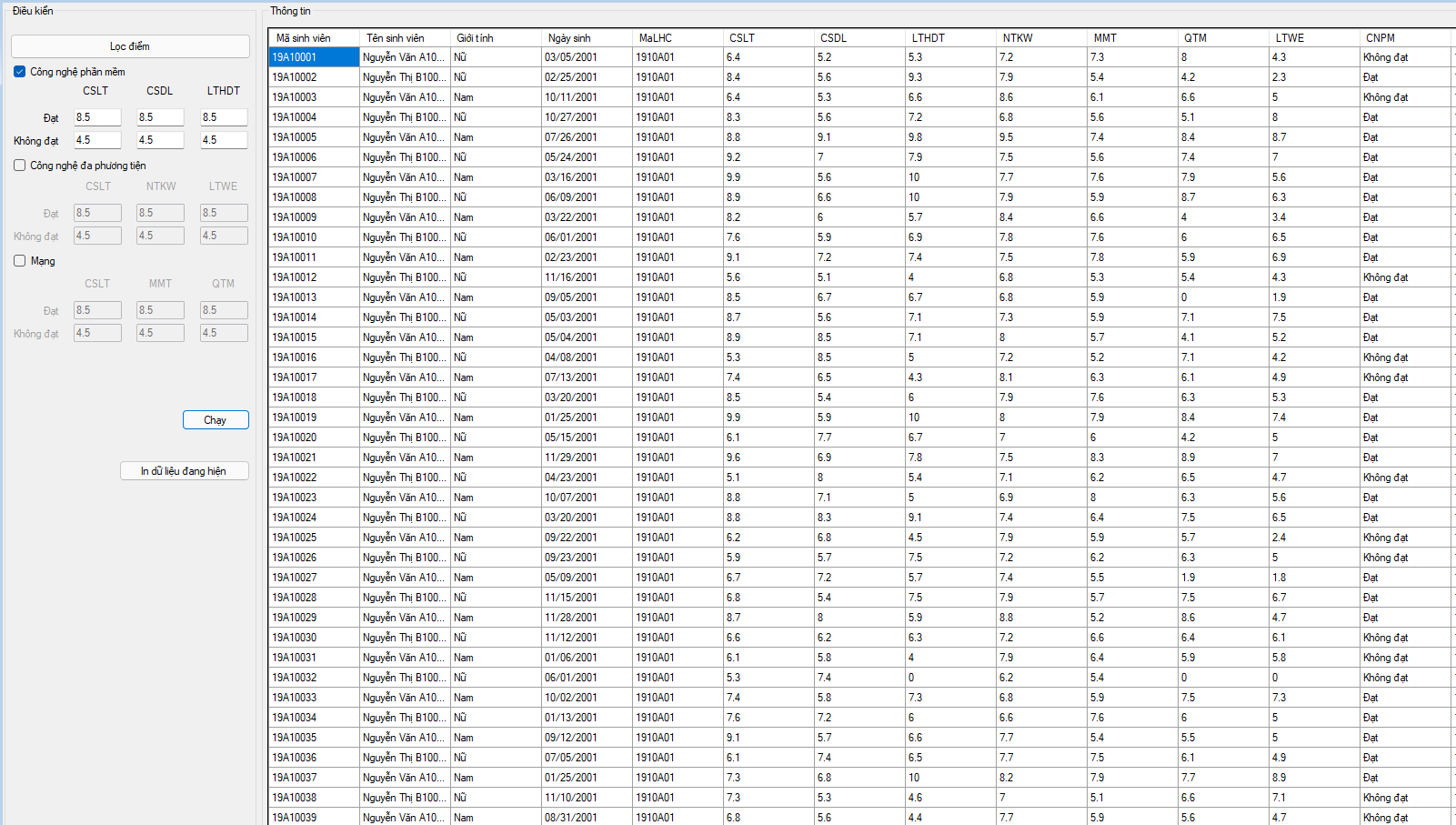
Tương tự đối với 2 chuyên ngành còn lại chúng ta sẽ khởi tạo tương tự.



Hình 15 Ảnh Bảng chọn chuyên ngành cần phân cụm và khởi tạo cụm ban đầu

Kết quả:

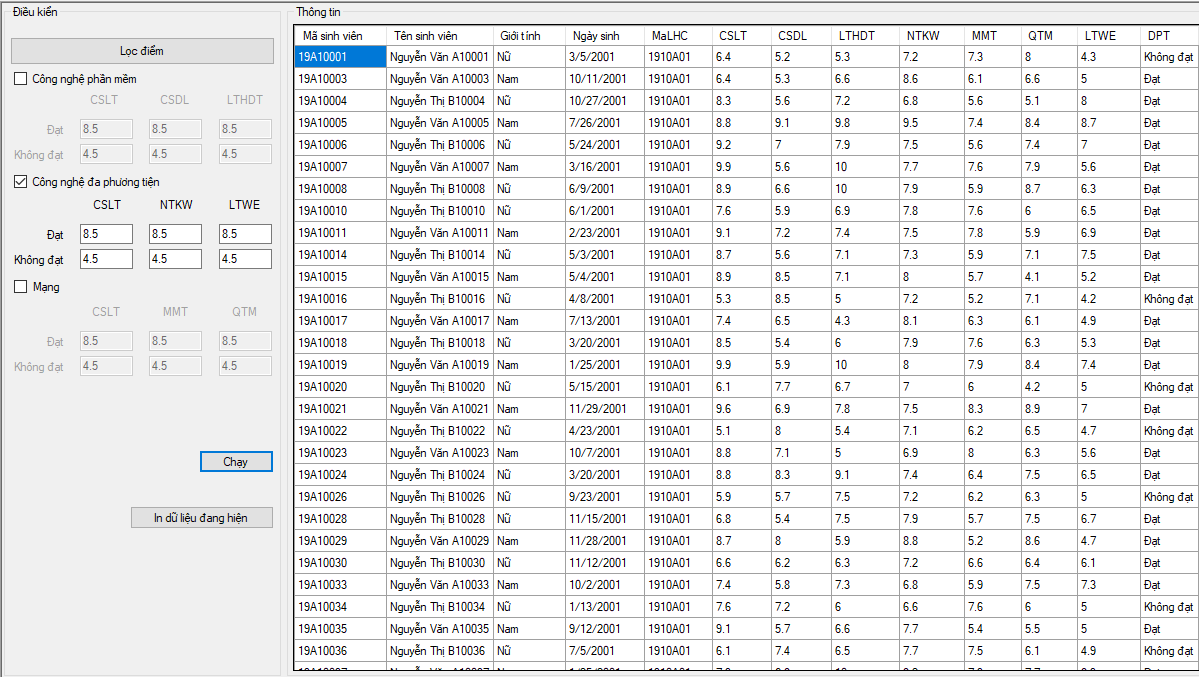
Sau khi phân cụm chuyên ngành Công nghệ phần mềm theo 2 cụm thì các môn được xét là CSLT, CSDL, LTHDT sẽ được phân vào 2 cụm theo thuật toán K-means. Những điểm nào gần ngưỡng trên thì sẽ được phân vào cụm 1 với kết luận “Đạt”. Tương tự với điểm gần ngưỡng dưới thì sẽ được phân vào cụm 2 với kết luận “Không đạt”.

Ví dụ trong bảng điểm dưới đây. Sinh viên có mã 19A10001 có điểm các môn CSLT, CSDL, LTHDT lần lượt là 6.4, 5.2, 5.3. Như ta thấy những điểm này gần ngưỡng 2 nên sẽ được phân vào cụm 2 “Không đạt”.

Hình 16 Ảnh chuyên ngành Công Nghệ Phần Mềm

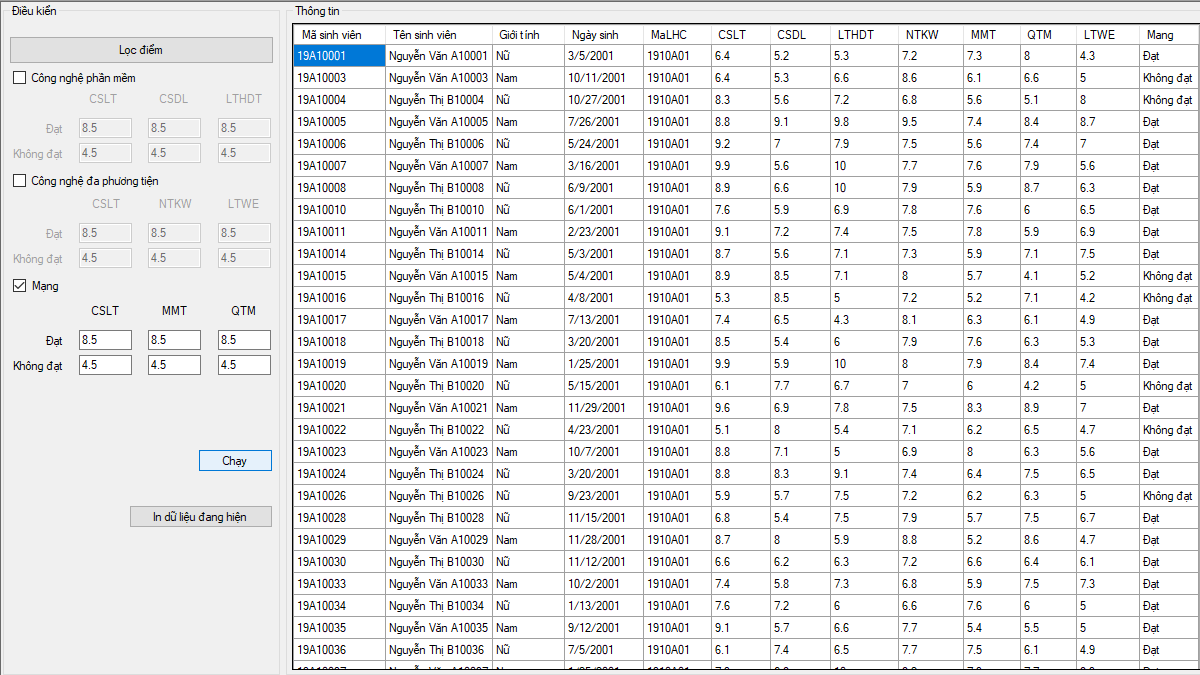
Sau khi phân cụm chuyên ngành Công nghệ đa phương tiện theo 2 cụm thì các môn được xét là CSTL, NTKW, LTWE sẽ được phân vào 2 cụm theo thuật toán K-means. Những điểm nào gần ngưỡng trên thì sẽ được phân vào cụm 1 với kết luận “Đạt”. Tương tự với điểm gần ngưỡng dưới thì sẽ được phân vào cụm 2 với kết luận “Không đạt”.

Ví dụ trong bảng điểm dưới đây. Sinh viên có mã 19A10001 có điểm các môn CSTL, NTKW, LTWE lần lượt là 6.4, 7.2, 4.3. Như ta thấy những điểm này gần ngưỡng 2 nên sẽ tiếp tục được phân vào cụm 2 “Không đạt”.



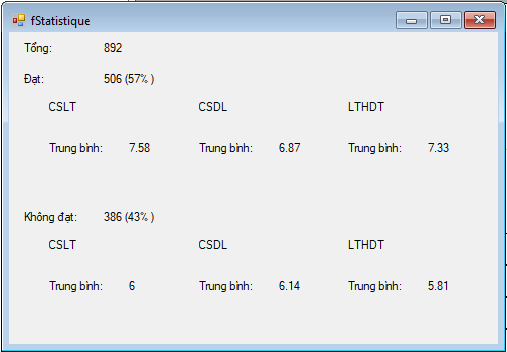
Hình 17 Ảnh chuyên ngành Đa Phương Tiện

Sau khi phân cụm chuyên ngành Mạng theo 2 cụm thì các môn được xét là CSTL,MMT, QTM sẽ được phân vào 2 cụm theo thuật toán K-means. Những điểm nào gần ngưỡng trên thì sẽ được phân vào cụm 1 với kết luận “Đạt”. Tương tự với điểm gần ngưỡng dưới thì sẽ được phân vào cụm 2 với kết luận “Không đạt”.

Ví dụ trong bảng điểm dưới đây. Sinh viên có mã 19A10001 có điểm các môn CSTL, MMT, QTM lần lượt là 6.4, 7.3, 8. Như ta thấy những điểm này gần ngưỡng 1 nên được phân vào cụm 2 “Đạt”.

Hình 18 Ảnh chuyên ngành Mạng

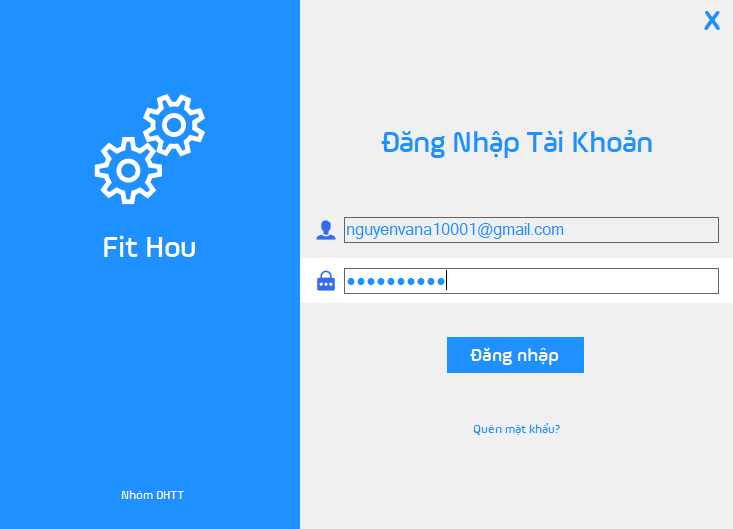
Như vậy sau khi lọc điểm theo 3 chuyên ngành lấy ví dụ là sinh viên mã 19A10001 được xác định đạt chuyên ngành Mạng, không đạt chuyên ngành Công nghệ đa phương tiện và chuyên ngành Công nghệ phần mềm.



Hình 19 Ảnh phân tích dữ liệu sau phân cụm Công Nghệ Phần Mềm

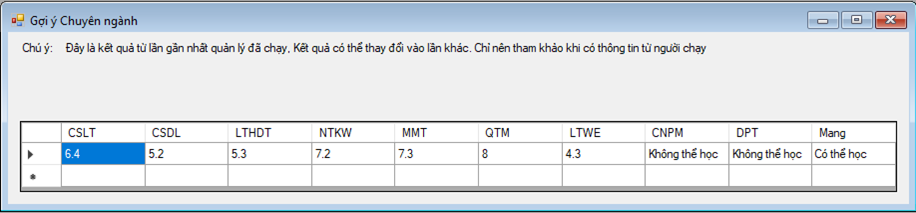
Từ đó đã có thêm cơ sở để xác định sinh viên phù hợp với chuyên ngành nào hơn, chọn chuyên ngành một cách có định hướng cơ sở.

* Sinh viên đăng nhập tài khoản để xem điểm và xem gợi ý chuyên ngành phù hợp:



Hình 20 Ảnh đăng nhập với quyền sinh viên

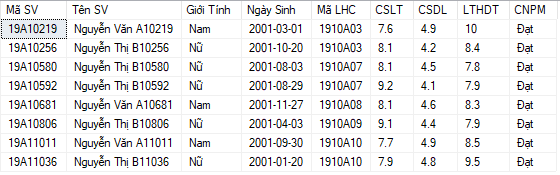
* Xem gợi ý chuyên ngành



Hình 21 Ảnh gợi ý chuyên ngành phù hợp cho sinh viên khi đăng nhập tài khoản

### Nhược điểm điểm khi sử dụng thuật toán K-means cho bài toàn phân cụm chuyên ngành.

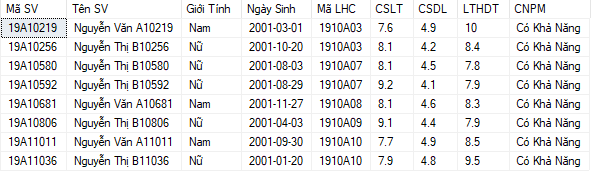
Do thuật toán K-means sử dụng khoảng cách Euclid để so sánh khoảng cách tử phần tử đến các cụm. Và sẽ gán chỉ số cụm cho phần tử khi thỏa mãn khoảng cách từ phần tử đó đến cụm được chọn là ngắn nhất trong tất cả các cụm. vì vậy sẽ xảy ra trường hợp sinh viên có điểm 2 trong 3 môn xét chuyên ngành rất cao (ví dụ 8, 9, 10), nhưng điểm còn lại thì lại rất thấp (ví dụ 4, 4.1 ). Điều này tạo ra trường hợp sinh viên ‘Đạt’ chuyên ngành nhưng thực tế là không.



Hình 22 Ảnh những sinh viên 'Đạt' nhưng có đầu điểm dưới 5

### Giải pháp cho nhược điểm trên là.

Chúng ta sẽ đánh dấu những sinh viên ‘Đạt’ chuyên ngành nhưng lại có điểm môn nào đó quá thấp và không đăng ký chuyên ngành đó. Điều này giúp cho việc cán bộ xét chuyên ngành chính xác trong việc xét đơn, nhưng cũng là gợi ý cho sinh viên nếu thực sự muốn đăng ký chuyên ngành đó có thể đăng ký học cải thiện điểm để đợt xét tuyển sau có thể đăng ký được chuyên ngành mong muốn.



Hình 23 Ảnh những sinh viên điểm 'Đạt' nhưng có điểm dưới 5 đã cập nhập

**KẾT LUẬN**

Những kết quả nghiên cứu của báo cáo cho phép rút ra những kết luận sau:

**Về mặt nghiên cứu khoa học:** Việc xây dựng thuật toán phân loại sinh viên theo chuyên ngành dựa trên thuật toán phân cụm dữ liệu K-means sẽ giúp cho việc nghiên cứu, định hướng sinh viên một cách có cơ sở, tiết kiệm thời gian.

**Về mặt thực tế:** Nhờ việc phân tích, tổng hợp một cách chi tiết, số liệu thực hiện đồng nhất, dựa trên các thuật toán hiệu quả nhất, các bược thực hiện có thứ tự đưa ra phương pháp tối ưu cho việc phân loại sinh viên vào chuyên ngành có hiệu quả nhất. Giúp cho cá nhân sinh viên, giảng viên, nhà trường có một sản phẩm ứng dụng để áp dụng phân loại chuyên ngành sao cho phù hợp với từng sinh viên. Từ đó, có thể nâng cao nền giáo dục, đào tạo thêm được các nhân tài cho đất nước.

**DANH MỤC TÀI LIỆU THAM KHẢO**

|  |  |
| --- | --- |
| [1] | vinhvinhit@gmail.com, "http://ungdung.khoa-hnvd.com/," 23 05 2016. [Online]. Available: http://ungdung.khoa-hnvd.com/Hoc\_thuat/KMeans.html. |
| [2] | J. McCaffrey, "https://visualstudiomagazine.com," 12 10 2013. [Online]. Available: https://visualstudiomagazine.com/Articles/2013/12/01/K-Means-Data-Clustering-Using-C.aspx?Page=1. [Accessed 17 03 2022]. |
| [3] | Đ. M. Tường, Các kỹ thuật cơ bản và hiện đại, Hà Nội: Đại học Quốc gia Hà Nội, 2015, pp. 480-481. |
| [4] | N. V. Chức, "http://bis.net.vn," 12 02 2010. [Online]. Available: http://bis.net.vn/forums/t/374.aspx. [Accessed 21 02 2022]. |