**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

---------------------------

**BÁO CÁO ĐỒ ÁN**

**Môn học: Khai Thác Dữ Liệu**

***Đề tài:***

**Tư vấn chọn chuyên ngành cho sinh viên khoa Công nghệ Thông tin bằng K-Means**

**Giảng viên hướng dẫn: Bùi Công Danh**

*Các thành viên:* Tái Thanh Tuấn – 2001181402

Nguyễn Sỹ Thành – 2001180146

Lê Danh Vũ – 2001181427

*TP. HỒ CHÍ MINH, tháng 12 năm 2020*

mục lục

[A. MỞ ĐẦU 3](#_Toc60572366)

[**1.** **Lời mở đầu** 3](#_Toc60572367)

[B. NỘI DUNG ĐỒ ÁN 5](#_Toc60572368)

[I. GIỚI THIỆU ĐỒ ÁN 5](#_Toc60572369)

[**1.** **Tên Đề Tài** 5](#_Toc60572370)

[**2.** **Phạm Vi Đề Tài** 5](#_Toc60572371)

[**3.** **Yêu Cầu Đề Tài** 5](#_Toc60572372)

[**4.** **Lý Do Chọn Đề Tài** 5](#_Toc60572373)

[**5.** **Phần Mềm Để Cài Đặt Yêu Cầu** 5](#_Toc60572374)

[II. PHÂN TÍCH CHI TIẾT NỘI DUNG ĐỀ TÀI 5](#_Toc60572375)

[**1.** **Tổng quan về K-Means** 5](#_Toc60572376)

[**2.** **Các bước thực hiện thuật toán Kmeans** 7](#_Toc60572377)

[**3.Quá trình làm bài tập lớn** 7](#_Toc60572378)

[**4.Cách áp dụng ý tưởng thuật toán vào đồ án môn học:** 12](#_Toc60572379)

[**5. Cài đặt và thiết kế ứng dụng ý tưởng thuật toán:** 13](#_Toc60572380)

[III. KẾT LUẬN VÀ ĐỊNH HƯỚNG PHÁT TRIỂN 17](#_Toc60572381)

[TÀI LIỆU THAM KHẢO 18](#_Toc60572382)

DANH MỤC HÌNH ẢNH

[Hình 1. Tổng quan thuật toán Kmeans 6](#_Toc60572383)

[Hình 2. Trang web thu thập dữ liệu 8](#_Toc60572384)

[Hình 3. Minh họa về thông tin sinh viên 8](#_Toc60572385)

[Hình 4. Minh họa bảng điểm của sinh viên 9](#_Toc60572386)

[Hình 5. Dữ liệu Excel 9](#_Toc60572387)

[Hình 6. Dữ liệu SQL DataBase 10](#_Toc60572388)

[Hình 7. Dữ liệu sau khi được khử nhiễu 11](#_Toc60572389)

[Hình 8. Giao diện demo 12](#_Toc60572390)

[Hình 9. Giao diện thực thi và các chức năng của phần mềm thực nghiệm 13](#_Toc60572391)

[Hình 10. Giao diện thực thi và các chức năng của phần mềm thực nghiệm 14](#_Toc60572392)

[Hình 11. Giao diện thực thi và các chức năng của phần mềm thực nghiệm 14](#_Toc60572393)

[Hình 12. Giao diện thực thi và các chức năng của phần mềm thực nghiệm 15](#_Toc60572394)

[Hình 13. Giao diện thực thi và các chức năng của phần mềm thực nghiệm 15](#_Toc60572395)

[Hình 14. Giao diện thực thi và các chức năng của phần mềm thực nghiệm 16](#_Toc60572396)

[Hình 15. Giao diện thực thi và các chức năng của phần mềm thực nghiệm 16](#_Toc60572397)

# MỞ ĐẦU

### **Lời mở đầu**

**T**

hời gian học tập môn học **Khai thác dữ liệu** cùng với nội dung bài tập lớn chúng em gặp không ít khó khăn về cách làm và kiến thức. Tuy nhiên chúng em đã nhận được rất nhiều sự giúp đỡ quý báu của thầy và các bạn, nhóm chúng em đã hoàn thành báo cáo bài tập lớn môn **Khai thác dữ liệu** với đề tài **Tư vấn chọn chuyên ngành cho sinh viên khoa CNTT bằng K-Means.** Nếu không có sự giúp đỡ của thầy và các bạn, nhóm em sẽ rất khó để có thể hoàn thiện được.

Nhóm em xin gửi lời cám ơn đến Ban giám hiệu trường Đại học Công nghiệp thực phẩm Thành phố Hồ Chí Minh đã tạo cho chúng em điều kiện học tập tốt nhất để học tập và nghiên cứu.

Để có được thành quả này, nhóm em xin gửi lời cám ơn về sự hướng dẫn và chỉ bảo tận tình của giáo viên hướng dẫn thầy **Bùi Công Danh** đã tận tình hướng dẫn và giúp đỡ nhóm em trong suốt quá trình hoàn thành chương trình cũng như bài báo cáo này. Em không chỉ nhận được những kiến thức bổ ích mà còn là hành trang theo em trong suốt thời gian học tập và làm việc sau này.

Vì thời gian có hạn đồng thời kinh nghiệm còn thiếu nên trong chương trình báo cáo này sẽ không tránh khỏi sai sót. Em rất mong nhận được sự góp ý của thầy cô và các bạn để báo cáo đạt được kết quả tốt hơn.

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

..................................................................................................................................

Ngày ....... tháng ........ năm .......

**GIÁO VIÊN HƯỚNG DẪN**

(Ký và ghi rõ họ tên)

1. **NỘI DUNG ĐỒ ÁN**
2. **GIỚI THIỆU ĐỒ ÁN**

### **Tên Đề Tài**

* Tư vấn chọn chuyên ngành cho sinh viên khoa Công Nghệ Thông Tin bằn thuật toán K-Means.

### **Phạm Vi Đề Tài**

* Thuật toán K-Means nằm trong giao trình môn học.
* K-Means có thể tìm hiểu trên các nguồn mạng khác về hàm phân cụm và các cuốn sách khác.

### **Yêu Cầu Đề Tài**

* Tìm hiểu khái quát về K-Means với các vấn đề:
  + K-Means là gì?
  + Kiến trúc của K-Means gồm những thành phần gì và cài đặt thuật toán.
* Tìm hiểu về thuật toán K-Means và cách cài đặt thuật toán.
* Đánh giá sự hoạt động của thuật toán được áp dụng trong đề tài.

### **Lý Do Chọn Đề Tài**

Vì vẫn luôn băn khoăn lượng dữ liệu của các sinh viên,… thu thập mỗi học kì và được xử lý như thế nào mà vẫn đảm bảo tốc độ truy cập tư vấn chọn chuyên ngành một cách nhanh chóng và hiệu quả nhất.

Thuật toán K-Means là một thuật toán phân cụm. Là một thuật toán hiệu quả dựa trên việc tư vấn chọn chuyên ngành cho sinh viên khoa CNTT.

### **Phần Mềm Để Cài Đặt Yêu Cầu**

* Visual Studio 2012.
* Ngôn ngữ C# giao diện .NET Framework.
* Microsoft SQL Server 2012.
* Microsoft Excel.

1. **PHÂN TÍCH CHI TIẾT NỘI DUNG ĐỀ TÀI**

### **Tổng quan về K-Means**

* 1. **Giới thiệu về K-Means?**

K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát(tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.

Thuật toán phân cụm k-means thường được sử dụng trong các ứng dụng cỗ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu,…



Hình 1. Tổng quan thuật toán Kmeans

* 1. **Thuật toán K-Means là gì?**

Thuật toán phân cụm k-means là một phương pháp được sử dụng trong phân tích tính chất cụm của dữ liệu. Nó đặc biệt được sử dụng nhiều trong khai phá dữ liệu và thống kê. Nó phân vùng dữ liệu thành k cụm khác nhau. Giải thuật này giúp chúng ta xác định được dữ liệu của chúng ta nó thực sử thuộc về nhóm nào.

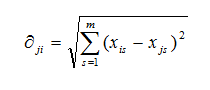
Để các bạn dễ hình dung ứng dụng của thuật toán. Chúng ta hãy quan sát một ví dụ thực tế như sau:

Trong các mô hình kinh doanh, doanh nghiệp sẽ chia nhỏ tệp khách hàng ra thành những nhóm đối tượng khác nhau để có thể áp dụng những chiến lược kinh doanh cụ thể cho từng nhóm đối tượng. Điều này giúp cho khách hàng được tiếp cận với các sản phẩm thật sự phù hợp với bản thân họ. Sự phù hợp đó sẽ kéo doanh số của chúng ta tăng lên. Vấn đề đặt ra là làm sao có thể chia nhỏ tệp khách hàng đó ra khi mà số lượng hóa đơn là rất lớn và chúng ta không thể ngồi để phân tích từng vị khách.

Và mục tiêu của các thuật toán phân cụm là từ tập dữ liệu khổng lồ đó. Làm sao chúng ta biết có những nhóm dữ liệu đặc trưng nào trong đó? Từng dữ liệu trong đó thuộc vào nhóm nào? Đó là cái mà thuật toán phân cụm của chúng ta cần đi tìm câu trả lời.

### **Các bước thực hiện thuật toán Kmeans**

**Bước 1**: Khởi tạo  
Chọn k trọng tâm {mj}kj=1 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 toán khoảng cách  
Đối với mỗi điểm Xi  (1 ≤ i ≤ n), tính toán 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.  
 **Điều kiện dừng:**  
Lặp lại các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

### **3.Quá trình làm bài tập lớn**

* + 1. **Thu thập dữ liệu**

**Đề tài chọn chuyên ngành cho sinh viên khoa công nghệ thông tin nên chúng em đã lấy dữ liệu từ trang** <http://sinhvien.hufi.edu.vn> **ấn vào “Tra cứu thông tin sinh viên”.**



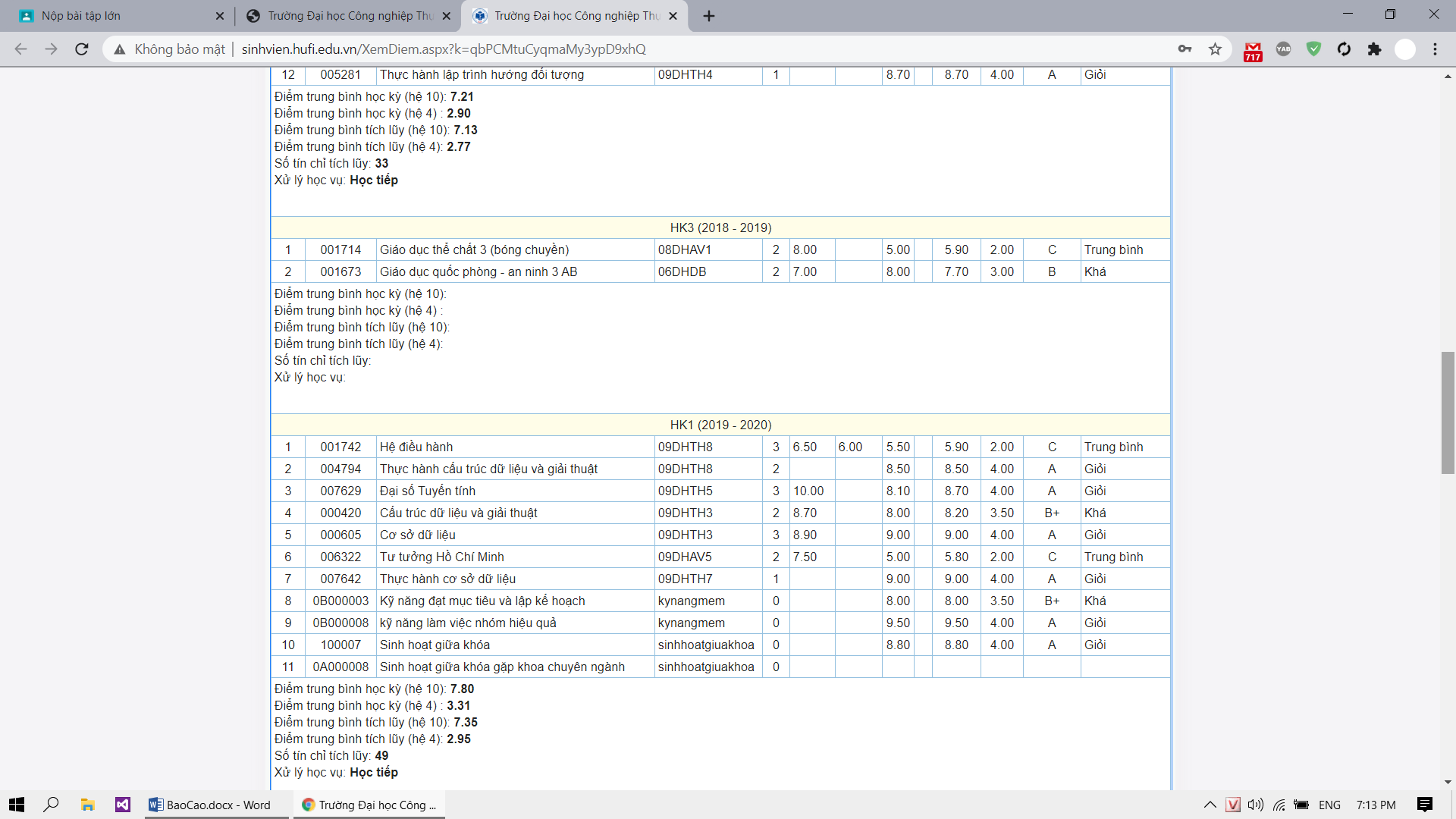
Hình 2. Trang web thu thập dữ liệu

**Sinh viên khoa công nghệ Thông tin khóa 08 thường có mã số sinh viên là “200117xxxx”, “xxxx” ở đây là nhóm em đã nhập từ 0001 -> 0400 và chỉ thu nhập được 117 dữ liệu sinh viên là thỏa mãn. Vậy nên nhóm em sẽ lấy thêm 100 dữ liệu của khóa 09 và cách lấy cũng tương tự như lấy của khóa 08 nhưng mã số sinh viên sẽ là “200118xxxx”.**

**Sau khi nhập mã số sinh viên và mã bảo vệ ta bấm nút “Tra cứu” và bấm vào “Xem điểm” thì ta sẽ dẫn đến trang có bảng kết quả học tập.**

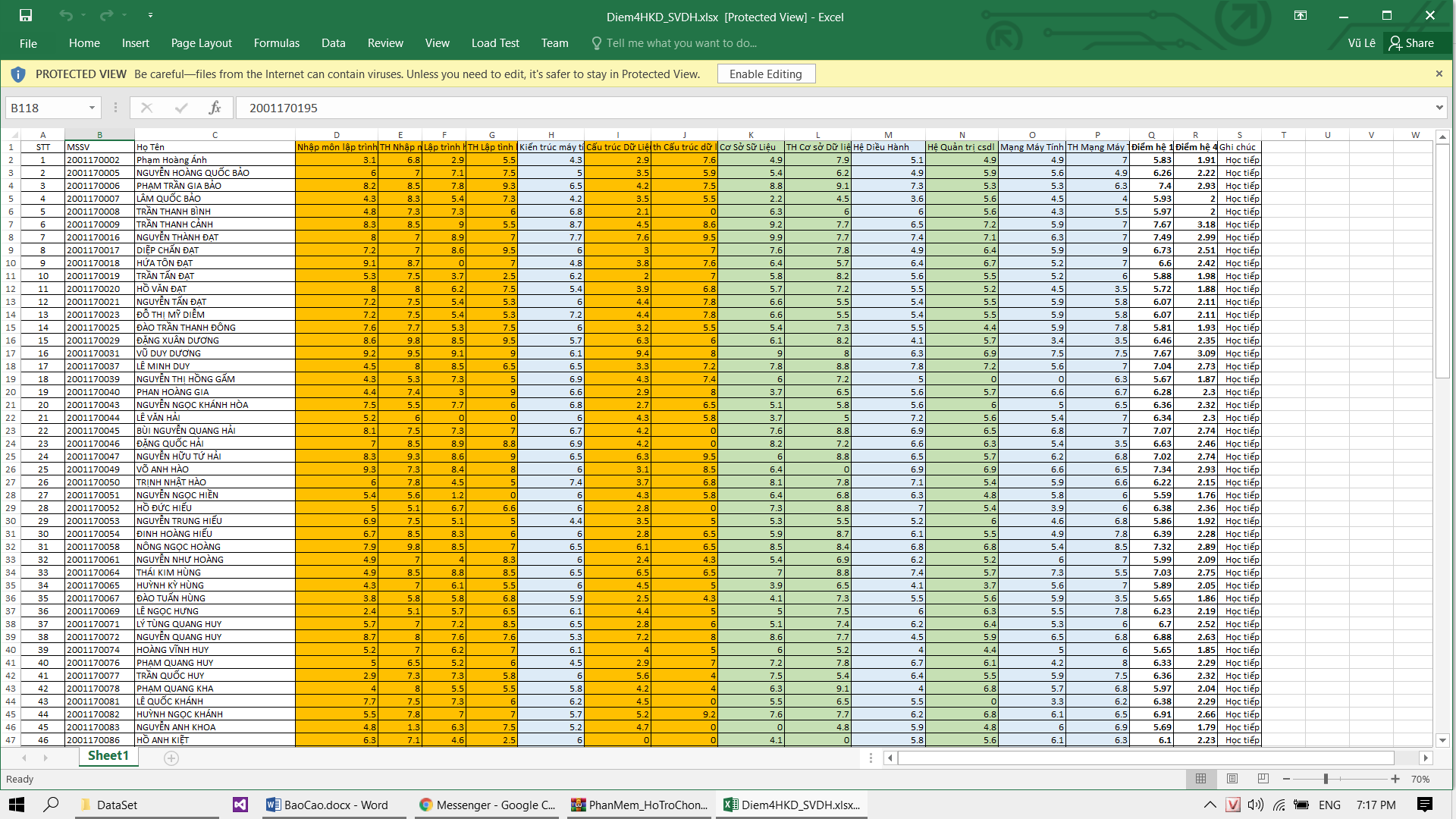


Hình 3. Minh họa về thông tin sinh viên

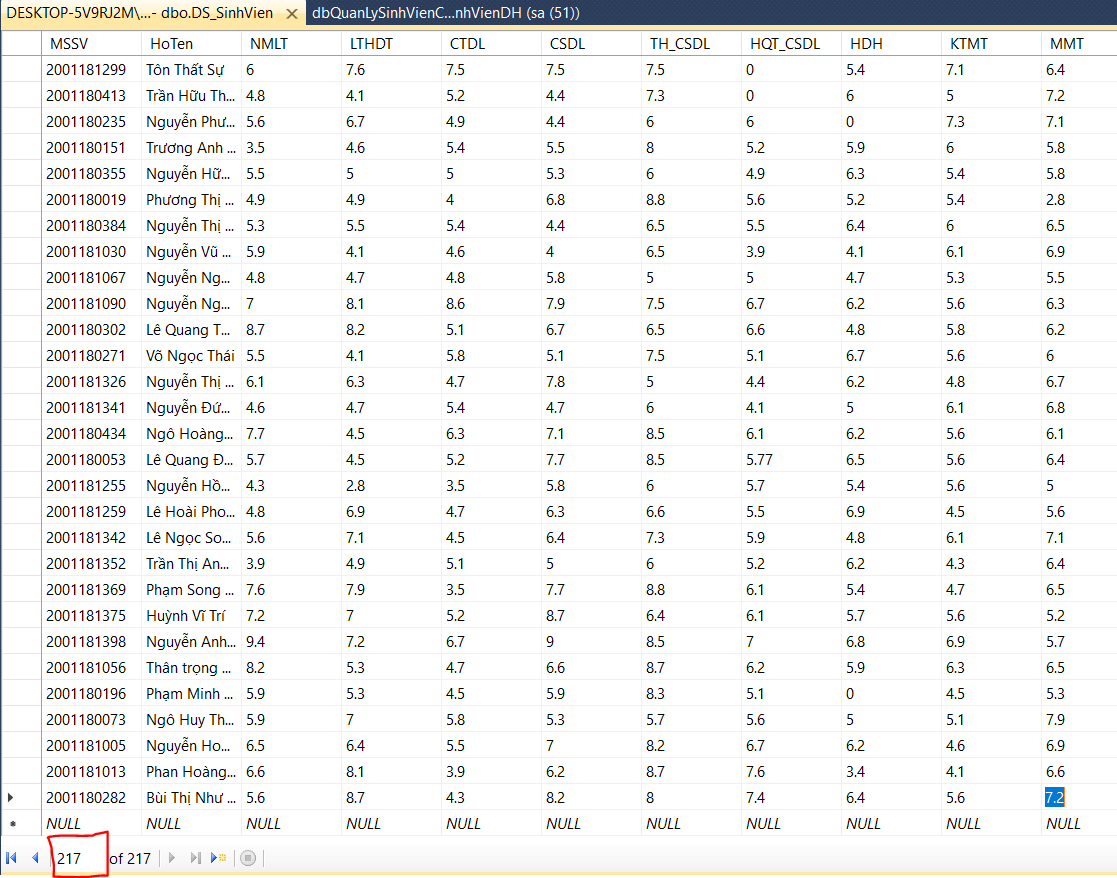


Hình 4. Minh họa bảng điểm của sinh viên

**Từ đó nhóm em sẽ chia ra mỗi người sẽ lấy 80 dữ liệu ổn nhất để cho vào bảng excel. Từ bảng excel ta sẽ chuyển dữ liệu qua SQL Database.**



Hình 5. Dữ liệu Excel



Hình 6. Dữ liệu SQL DataBase

* 1. **Làm sạch dữ liệu**

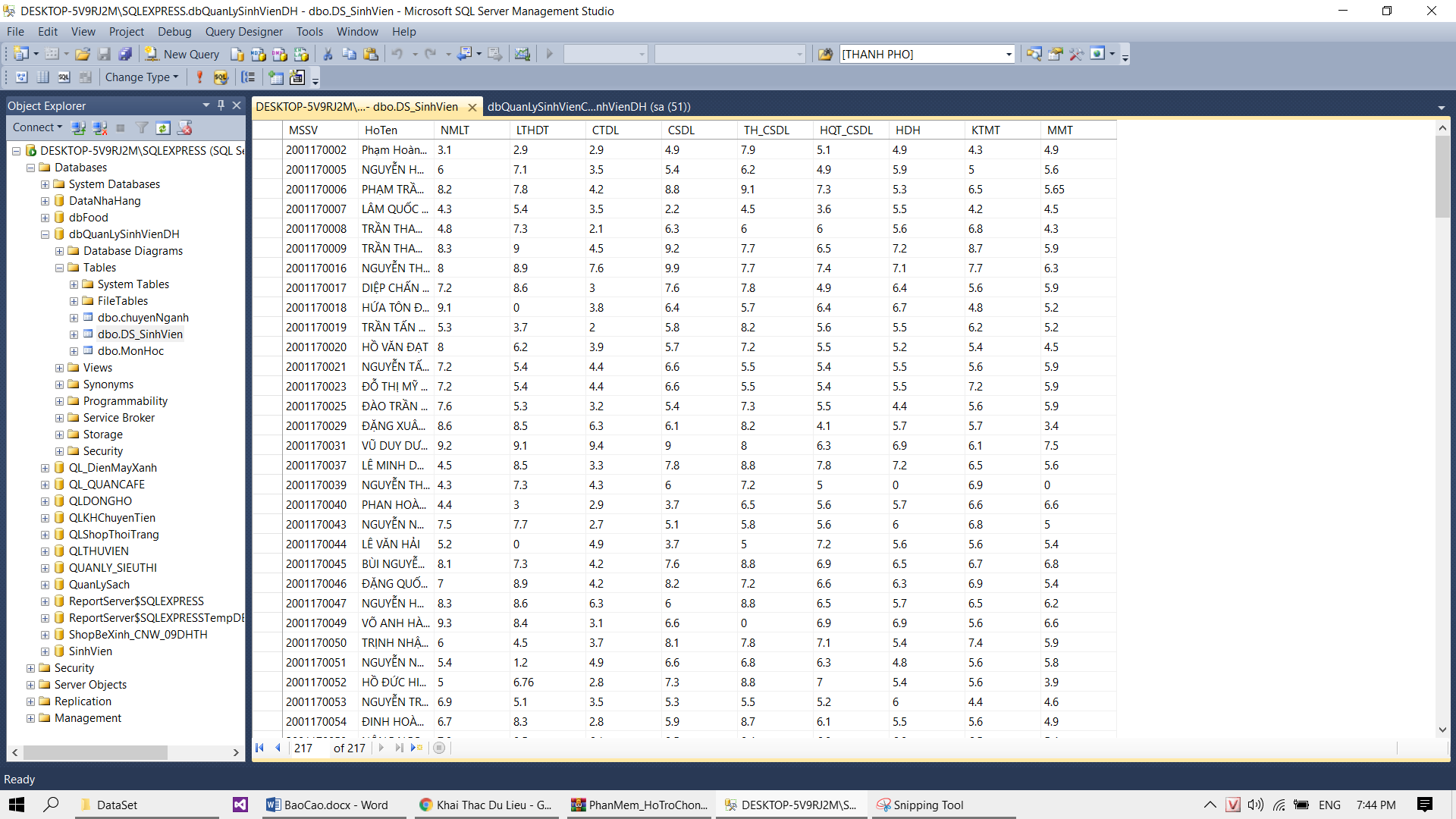
Dùng phương pháp “Chia theo độ sâu” của chương 2 trong cuốn giáo trình “Khai thác dữ liệu” do thầy đã đưa nên chúng em đã áp dụng làm sạch dữ liệu theo phương pháp đó.

Có 3 cách tiến hành khử nhiễu (làm trơn dữ liệu):

* Khử nhiễu bằng giá trị trung vị.
* Khử nhiễu bằng giá trị trung bình.
* Khử nhiễu bằng giá trị biên.

Nhóm em đã áp dụng “khử nhiễu bằng giá trị trung bình”: biến đổi giá trị trong giỏ theo giá trị trung bình của các phần tử trong giỏ.

Chúng em còn tiến hành loại các sinh viên bỏ học và nhiều điểm điểm F để dữ liệu được đẹp hơn.



Hình 7. Dữ liệu sau khi được khử nhiễu

* 1. **Input thuật toán**

-Input vào thuật toán 217 sinh viên với 9 thuộc tính điểm và 2 thuộc tính thông tin sinh viên:mã số sinh viên và họ tên

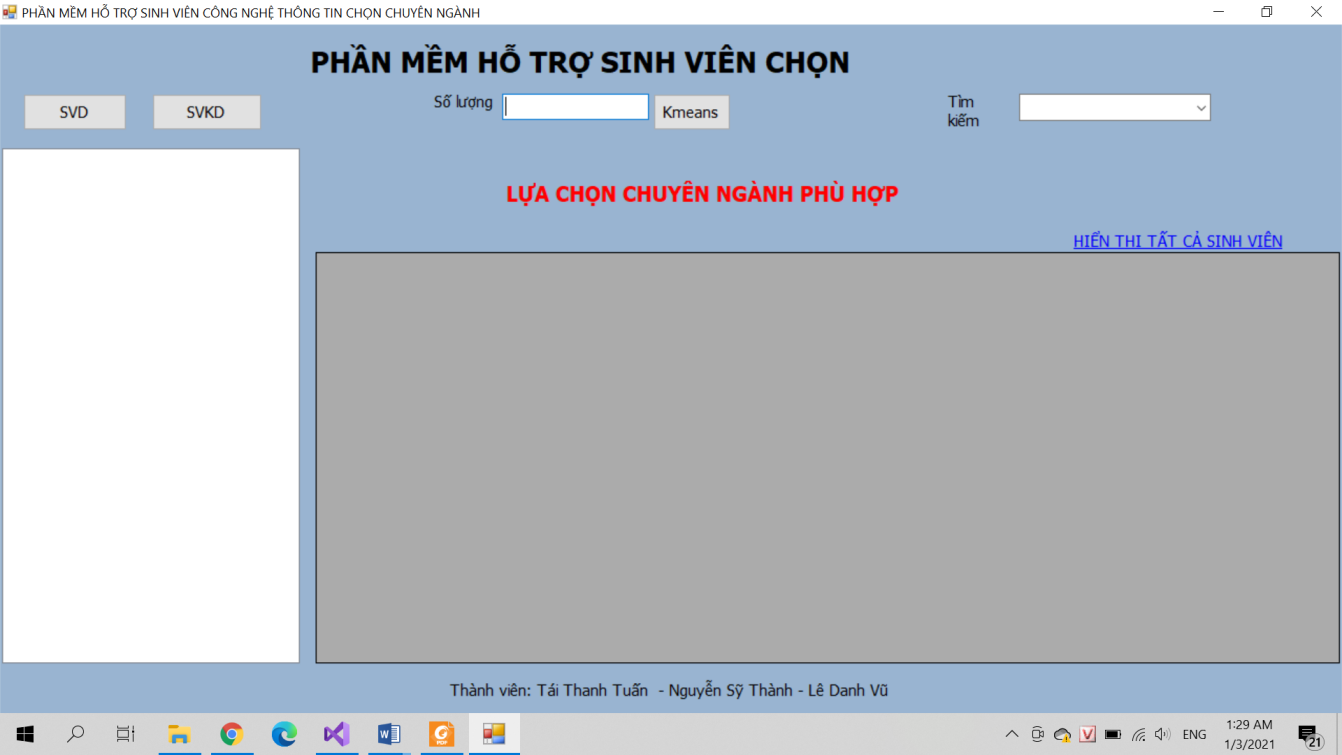
-Tạo ra 3 cụm tương ứng với 3 chuyên ngành:Công nghệ phần mềm,Hệ thống thông tin và Mạng máy tính

+Công nghệ phần mềm:Nhập môn lập trình,Cấu trúc dữ liệu và giải thuật,Lập trình hướng đối tượng

+Hệ thống thông tin:Cơ sở dữ liệu,Hệ quản trị cơ sở dữ liệu,Thực hành cơ sở dữ liệu.

+Mạng máy tính:Kiến trúc máy tính,Hệ điều hành,Mạng máy tính.

* 1. **Giao diện Demo**



Hình 8. Giao diện demo

### **4.Cách áp dụng ý tưởng thuật toán vào đồ án môn học:**

**Bước 1:** Với dữ liệu input đầu vào như đã nói ở trên, bước đầu tiên chúng ta sẽ gom cụm trong từng chuyên ngành thành 2 cụm. Bao gồm:

* Cụm 1: Cụm phù hợp hoặc có khả năng học chuyên ngành đó. Tâm cụm điểm điểm môn học chuyên ngành cụm 1(7,7,7).
* Cụm 2: Cụm không phù hợp với chuyên ngành đó. Tâm cụm điểm môn học chuyên ngành cụm 2(2,2,2).

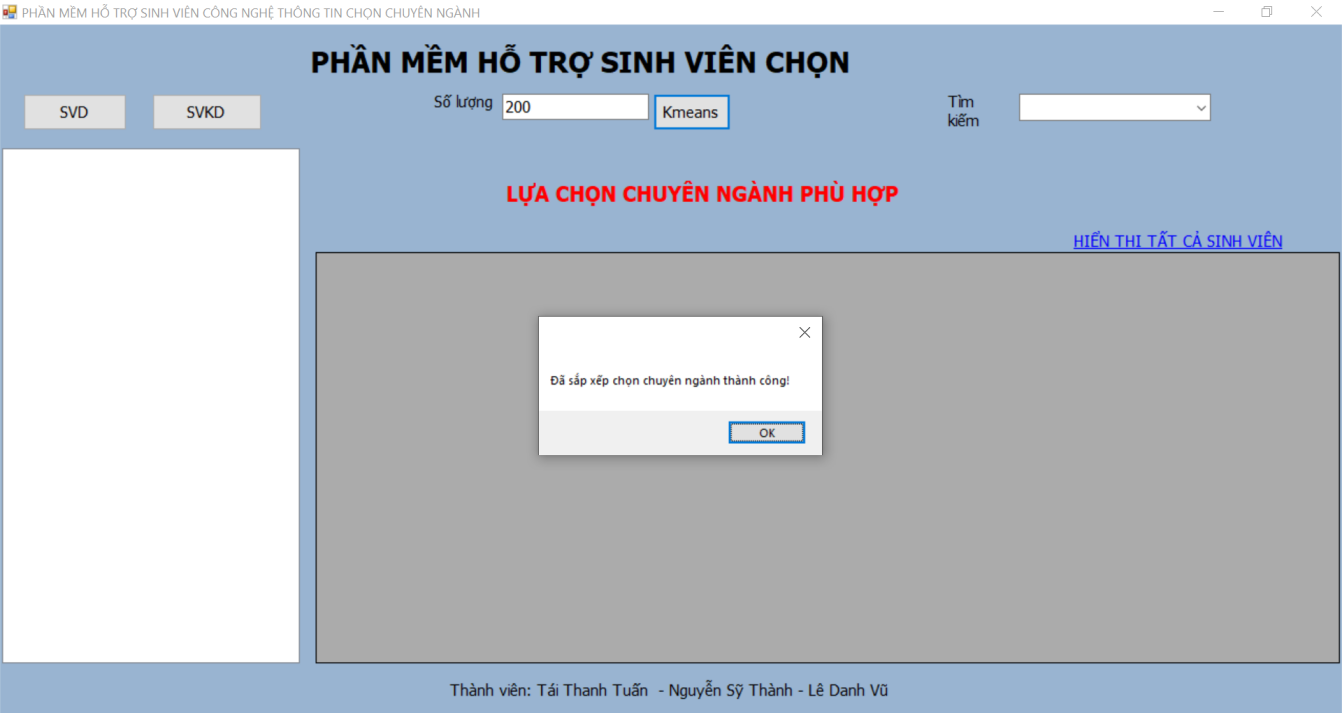
(\*Giải thích: để phù hợp gần sát với yêu cầu thực tế để có thể theo học chuyên ngành đó thì ít nhất sinh viên đó phải đậu được các môn học trong chuyên ngành đó. Vì thế bước 1 sẽ phân ra 2 cụm chủ yếu để sàn lọc ra những sinh viên có khả năng chọn theo học chuyên ngành và danh sách những sinh viên không có khả năng theo học chuyên ngành đó. Còn vì sao chọn 2 tâm cụm bắt đầu sinh viên đạt yêu cầu (7,7,7) và sinh viên không đạt yêu cầu (2,2,2): Dựa theo số môn của từng chuyên ngành mà nhóm đồ án collect được là mỗi chuyên ngành là 3 môn, nên số lượng các môn học trong tâm cụm sẽ là 3, và lấy trung bình giữa 2 khoảng từ 0🡪5 điểm và 5🡪10 điểm là 2 và 7 để chúng ta xét sinh viên điểm số của sinh viên đó thuộc cụm 1 hoặc cụm 2 của 3 chuyên ngành).

**Bước 2:** Tiếp đến đã có danh sách của những sinh viên đạt yêu cầu(Còn danh sách sinh viên không đủ yêu cầu thì không cần phải duyệt xét nữa) thì tiếp tục xét xem trong các cụm đạt yêu cầu của từng chuyên ngành có sinh viên nào trùng không? Nếu có thì xét khoảng cách Euclid điểm số của sinh viên đó gần tâm cụm chuyên ngành nào nhất thì duyệt vào.

**Bước 3:** Tính lại tâm cụm của 3 chuyên ngành(Bằng công thức trung bình cộng) sau khi đã thay đổi các sinh viên vào những chuyên ngành thích hợp.

**Bước 4:** Tiếp tục lấy tâm cụm mới tính ở bước 3 xét lại các danh sách sinh viên đạt chuyên ngành ban đầu xét ở bước 1 và Quay lại bước 2 cho đến khi tâm cụm không còn thay đổi thì dừng.

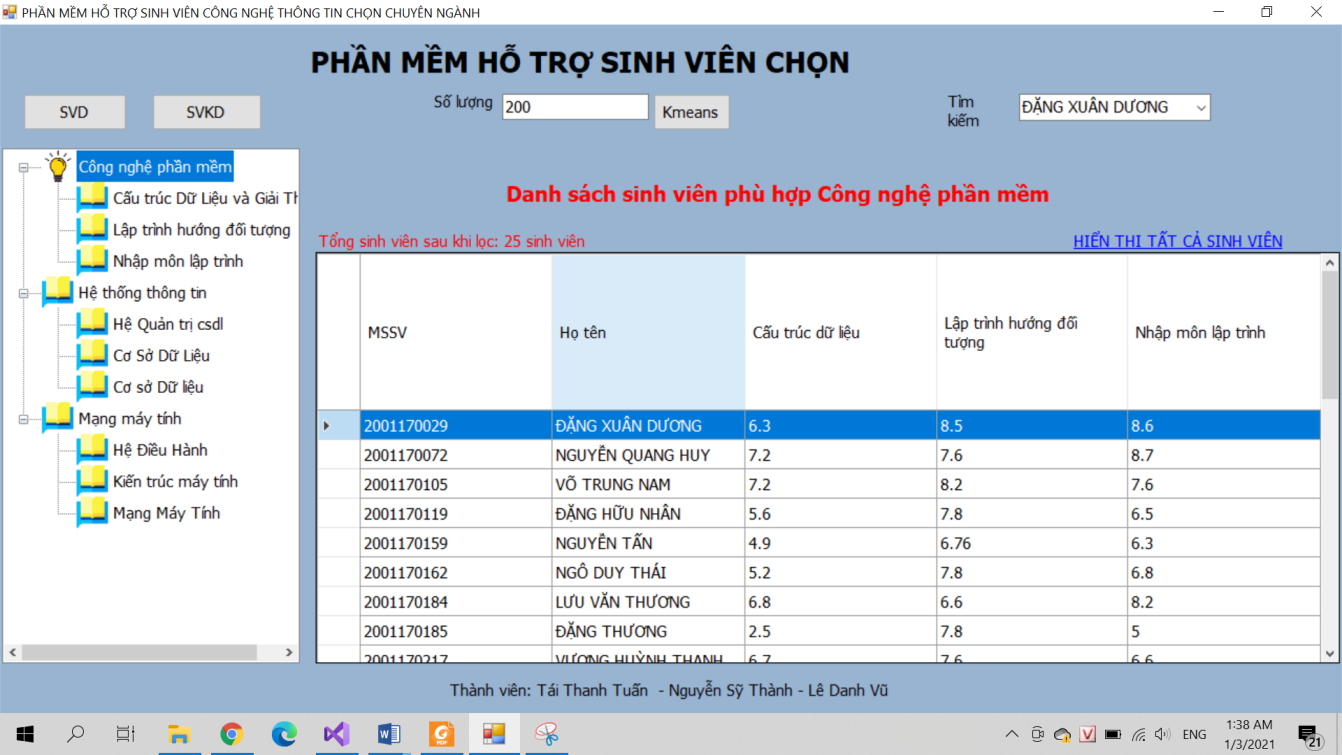
### **5. Cài đặt và thiết kế ứng dụng ý tưởng thuật toán:**



Hình 9. Giao diện thực thi và các chức năng của phần mềm thực nghiệm

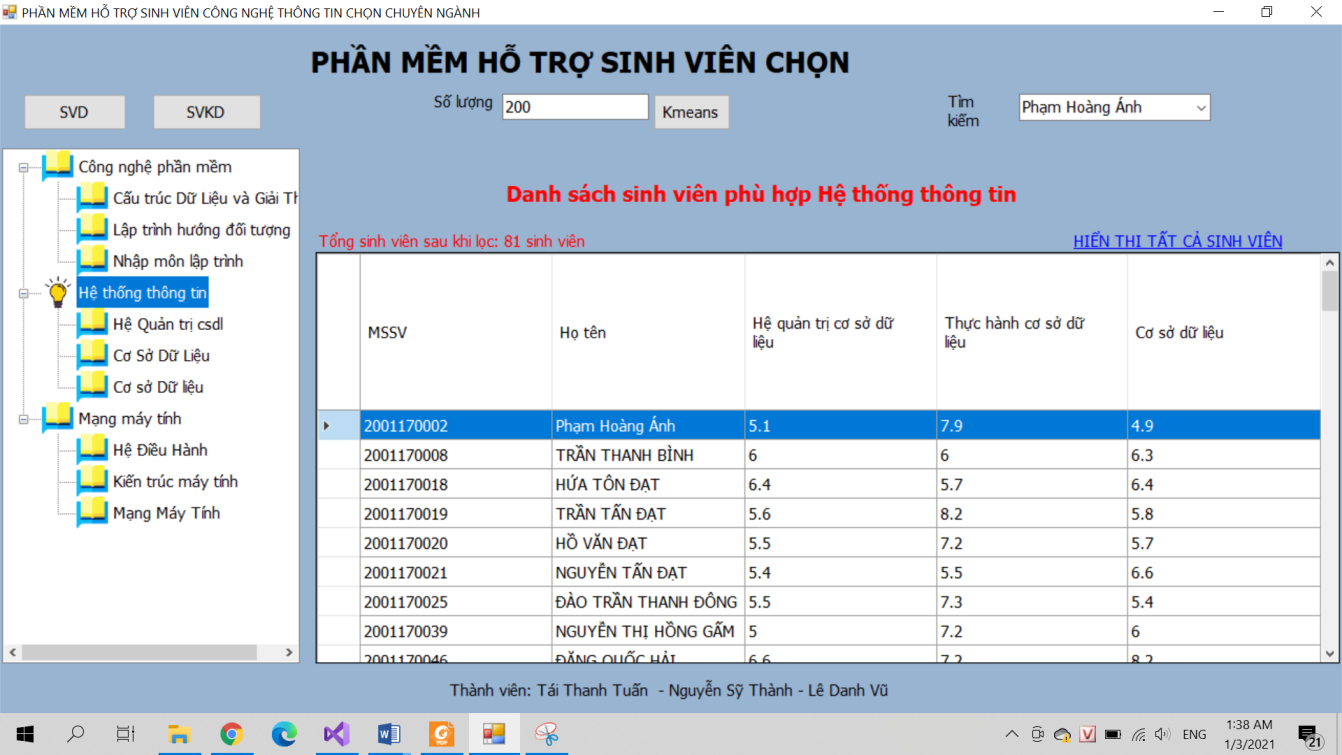
* Trước tiên nhập số lượng danh sách sinh viên để thực thi (Điều kiện bắt buộc- và do số lượng nhóm đồ án thu thập là 217 nên không được nhập quá tối đa).
* Button Kmeans: Thực thi thuật toán kmeans gom cụm vào các sinh viên vừa chọn vào các chuyên ngành phù hợp nhất.
* Button SVD(Sinh viên đạt): Chứa danh sách các sinh viên đạt yêu cầu của từng chuyên ngành. Bao gồm các thông tin sau:

Ngành Công nghệ phần mềm: (Hiển thị tổng số lượng sinh viên và thông tin và điểm số các sinh viên)



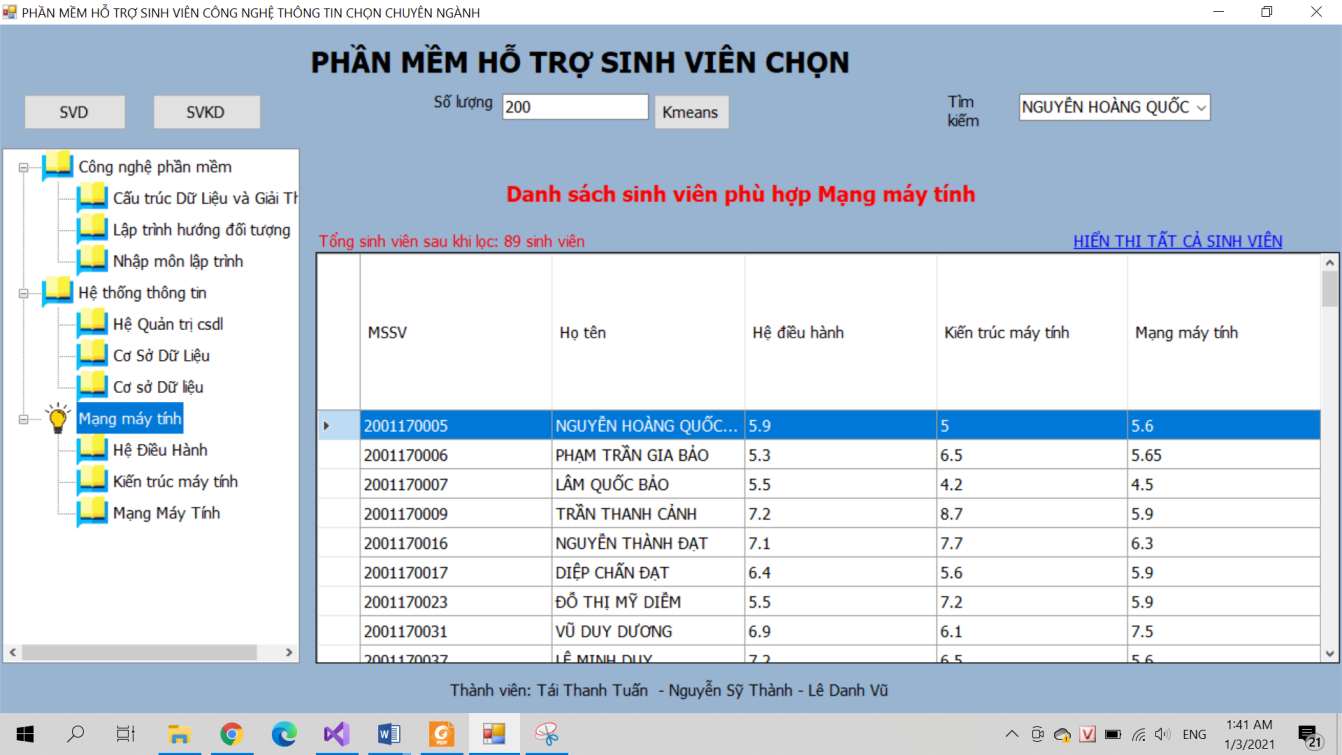
Hình 10. Giao diện thực thi và các chức năng của phần mềm thực nghiệm

Ngành Hệ thống thông tin:



Hình 11. Giao diện thực thi và các chức năng của phần mềm thực nghiệm

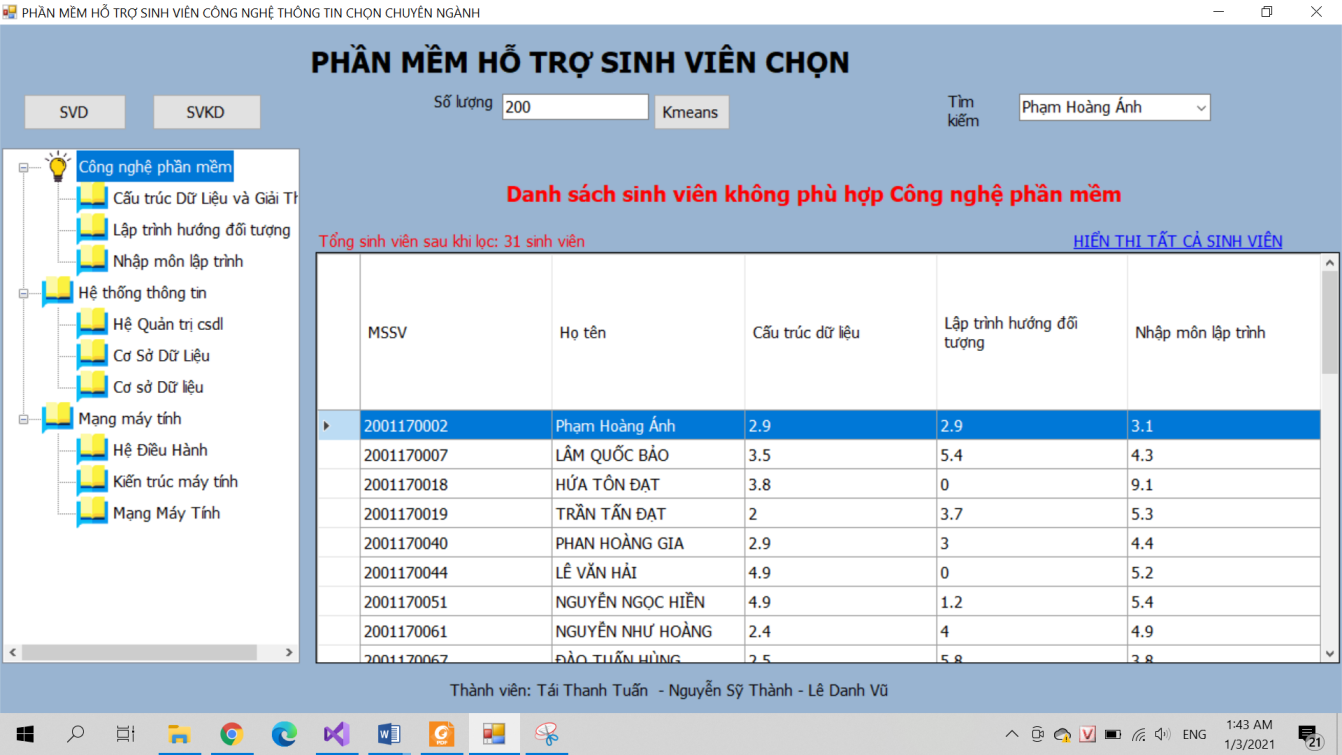
Ngành Mạng máy tính:



Hình 12. Giao diện thực thi và các chức năng của phần mềm thực nghiệm

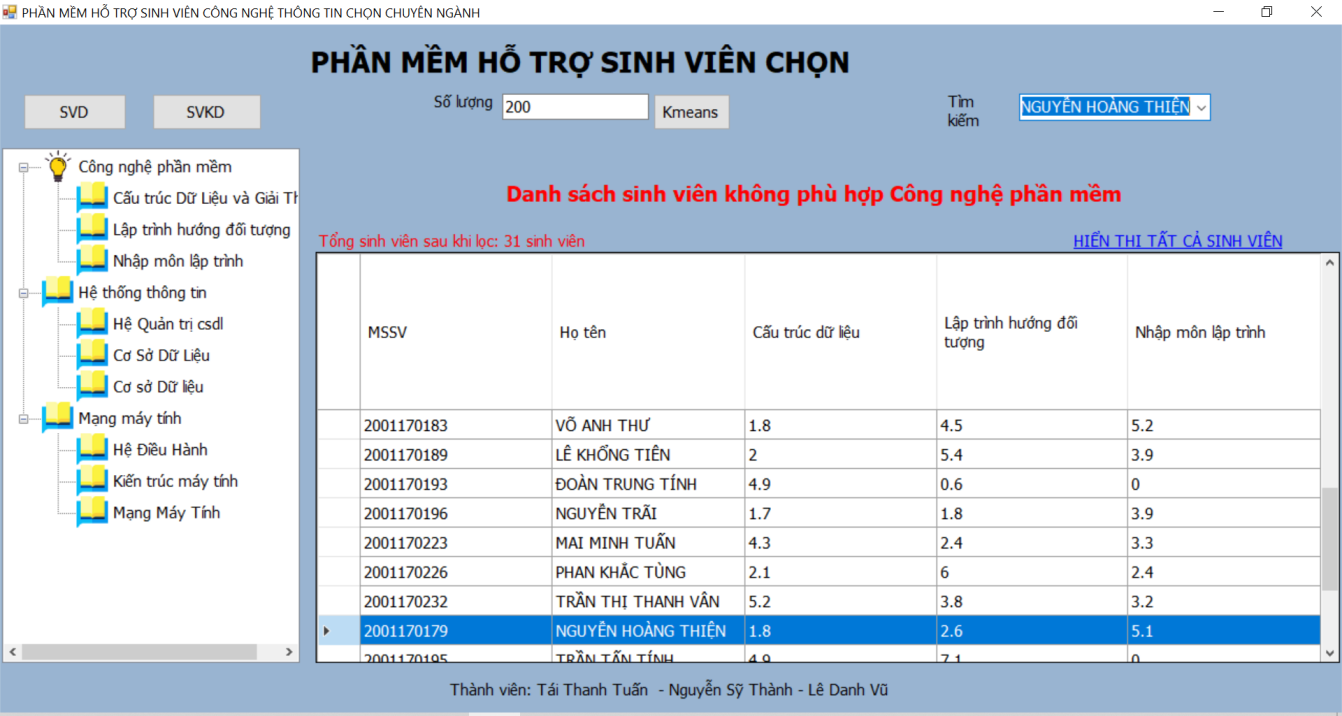
* Button SVKD(Sinh viên không đạt): Chứa danh sách các sinh viên chưa đạt yêu cầu của từng chuyên ngành. Bao gồm các thông tin sau:

Ngành Công nghệ phần mềm( Và các ngành khác cũng tương tự):



Hình 13. Giao diện thực thi và các chức năng của phần mềm thực nghiệm

* ComboBox tìm kiếm: Thực hiện hỗ trợ cho người dùng tìm kiếm nhanh và chính xác tên sinh viên mình muốn tìm. Khi chọn 1 tên bất kỳ trong comboBox thì dataGridView sẽ focus đúng row tên sinh viên trong từng danh sách cho người dùng dễ dàng tìm kiếm nhanh hơn.



Hình 14. Giao diện thực thi và các chức năng của phần mềm thực nghiệm

* LinkLabel “Hiển thị tất cả sinh viên”: show ra danh sách sinh viên dựa vào số lượng mình đã xác định lúc đầu.



Hình 15. Giao diện thực thi và các chức năng của phần mềm thực nghiệm

1. **KẾT LUẬN VÀ ĐỊNH HƯỚNG PHÁT TRIỂN**

Qua đồ án môn học “Khai thác dữ liệu”. Qua nhiều lần thực nghiệm theo từng số lượng thay đổi ta có thể thấy được thuật toán Kmeans cũng còn bị hạn chế khá nhiều trong nghiệp vụ Chọn chuyên ngành phù hợp dựa vào các điểm số sinh viên. Do thuật toán Kmeans phụ thuộc rất nhiều vào dữ liệu dẫn đến một vài trường hợp vẫn còn chưa chính xác.

Ngoài những điều hạn chế đã kể trên. Thực tế ngày nay ngành Công nghệ thông tin cũng đang là 1 trong nhiều ngành hot có rất nhiều sinh viên và ý tưởng đồ án của Nhóm đề tài đưa ra vô cùng sát với nhu cầu thực tế. Nhưng ở đồ án này nhóm chỉ xét và dựa vào phương diện điểm số của các sinh viên mà đưa ra lời khuyên chọn chuyên ngành bất kì nào đó. Để mà có thể đưa và áp dụng dùng được vào trong thực tế thì chúng ta còn cần phải xét đến nhiều trường hợp thực tế khác hơn, ví dụ như sở thích, đam mê, điều kiện gia đình,… Ngoài ra, để có thể khắc phục nhược điểm của thuật toán Kmeans và tăng độ chính xác thì ta cần phải sử dụng nhiều thuật toán nên là từ 2 thuật toán trở lên và có thể thay đổi thuật toán bằng thuật toán gom cụm, phân lớp có độ chính xác tốt hơn kmeans như svm,…

**TÀI LIỆU THAM KHẢO**

https://jstf.hufi.edu.vn/uploads/files/so-tap-chi/nam-2018/tap-15-so-01/15\_152-160.pdf

DoLechChuan.pptx

DataMining\_BuiCongDanh\_Chuong1.pdf

DataMining\_BuiCongDanh\_Chuong2.pdf

DataMining\_BuiCongDanh\_Chuong3.pdf

DataMining\_BuiCongDanh\_Chuong4.pdf

DataMining\_BuiCongDanh\_Chuong5.pdf

<https://machinelearningcoban.com/2017/01/01/kmeans/>

<https://nguyenvanhieu.vn/thuat-toan-phan-cum-k-means/>