TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:**

**DỰ ĐOÁN TỈ LỆ MẮC VÀ TỬ VONG DO VIRUS COVID-19 BẰNG THUẬT TOÁN PHÂN CỤM K-MEANS**

|  |  |
| --- | --- |
|  |  |
| |  |  |  | | --- | --- | --- | | **Sinh viên thực hiện** | **: LÊ TỰ HỮU - 19810310662**  **CAO THANH SANG - 19810310019**  **LƯU VĂN TIẾN - 19810310683** | | | **Giảng viên hướng dẫn** | **: VŨ VĂN ĐỊNH** | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | **Lớp** | **: D14CNPM1** | | **Khóa** | **: 2019-2024** | |  |  | |  |
|  |  |

***Hà Nội, tháng 06 năm 2022***

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên sinh viên** | **Điểm** | **Chữ ký** |
| 1 | Lê Tự Hữu |  |  |
| 2 | Cao Thanh Sang |  |  |
| 3 | Lưu Văn Tiến |  |  |

|  |  |  |
| --- | --- | --- |
| **Họ và tên giảng viên** | **Chữ ký** | **Ghi chú** |
| Giáo viên chấm 1: |  |  |
| Giáo viên chấm 2: |  |  |

Mục Lục

[LỜI CẢM ƠN 1](#_Toc104883972)

[TÓM TẮT ĐỀ TÀI 2](#_Toc104883973)

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 3](#_Toc104883974)

[1.1 Đặt vấn đề 3](#_Toc104883975)

[1.2 Cơ sở hình thành đề tài 3](#_Toc104883976)

[1.3 Mục tiêu đề tài 4](#_Toc104883977)

[1.4 Đối tượng và phương pháp nghiên cứu 4](#_Toc104883978)

[1.5 Ý nghĩa đề tài 5](#_Toc104883979)

[1.5.1 Ý nghĩa khoa học 5](#_Toc104883980)

[1.5.2 Ý nghĩa thực tiễn 5](#_Toc104883981)

[CHƯƠNG 2: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 6](#_Toc104883982)

[2.1 Tổng quan về kỹ thuật Khai phá dữ liệu (Data mining) 6](#_Toc104883983)

[2.1.1 Khái niệm về khai phá dữ liệu 6](#_Toc104883984)

[2.1.2 Quy trình khai phá dữ liệu 7](#_Toc104883985)

[2.1.3 Ứng dụng của khai phá dữ liệu 9](#_Toc104883986)

[2.2 Phân cụm dữ liệu và ứng dụng 10](#_Toc104883987)

[2.2.1 Mục đích của phân cụm dữ liệu 10](#_Toc104883988)

[2.2.2 Các bước cơ bản để phân cụm 11](#_Toc104883989)

[2.2.3 Các ứng dụng của phân cụm 13](#_Toc104883990)

[2.3 Cơ sở dữ liệu về Covid 19 14](#_Toc104883991)

[2.3.1 Thông tin về dịch Covid 19 14](#_Toc104883992)

[2.3.2 Dấu hiệu và triệu chứng 15](#_Toc104883993)

[CHƯƠNG 3: KỸ THUẬT PHÂN CỤM VÀ THUẬT TOÁN K-MEANS 17](#_Toc104883994)

[3.1 Giới thiệu về kỹ thuật phân cụm trong Khai phá dữ liệu 17](#_Toc104883995)

[3.2 Thuật Toán K-Means 18](#_Toc104883996)

[3.3 Áp dụng và sử dụng thuật toán K-means vào bộ dataset Covid-19 19](#_Toc104883997)

[CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ 19](#_Toc104883998)

[4.1 Xây dựng mô hình bằng Weka 19](#_Toc104883999)

[4.2 Phân tích dữ liệu từ kết quả của weka 20](#_Toc104884000)

[1.Tỉ lệ quốc gia trên số ca nhiễm 20](#_Toc104884001)

[2.Tỉ lệ quốc gia trên tổng số người tử vong / 1 triệu người 22](#_Toc104884002)

[3.Tỉ lệ tiêm chủng vacxin ở các nước 23](#_Toc104884003)

[4.Tỉ lệ xét nghiệm covid 19 24](#_Toc104884004)

[KẾT LUẬN 26](#_Toc104884005)

[TÀI LIỆU THAM KHẢO 27](#_Toc104884006)

# LỜI CẢM ƠN

Trong thời gian làm báo cáo môn học, chúng em đã nhận được nhiều sự giúp đỡ, đóng góp ý kiến và chỉ bảo nhiệt tình của thầy cô, gia đình và bạn bè. Em xin gửi lời cảm ơn chân thành đến giảng viên Vũ Văn Định, giảng viên Bộ môn Khai phá dữ liệu trường Đại học Điện Lực người đã tận tình hướng dẫn, chỉ bảo em trong suốt quá trình học tập, nghiên cứu.

Em cũng xin chân thành cảm ơn các thầy cô trong trường Đại học Điện Lực nói chung, các thầy cô trong khoa công nghệ thông tin nói riêng đã chỉ dẫn cho em kiến thức về các môn đại cương cũng như các môn chuyên ngành, giúp em có được cơ sở lý thuyết vững vàng và tạo điều kiện giúp đỡ em trong suốt quá trình học tập.

# TÓM TẮT ĐỀ TÀI

Ngành y tế và giáo dục luôn là vấn đề sống còn của bất kỳ quốc gia nào trên thế giới. Trong những năm gần đây, chính phủ Việt nam đặc biệt đầu tư cho hai ngành mũi nhọn này thông qua các chính sách , nguồn vốn dành cho trang thiết bị hạ tầng và nghiên cứu khoa học. Trong lĩnh vực kho học, càng ngày càng có nhiều công trình khoa học trong y tế. Tuy nhiên các nghiên cứu khoa học về ứng dụng công nghệ thông tin để giải quyết bài toán về y tế là không nhiều. Do sự nguy hiểm và tình hình lây lan diễn biến phức tạm của đại dịch Covid-19 xảy ra trên toàn thế giới, vậy nên chúng e làm đề tài sử dụng môn học khai phá dữ liệu để xác định đánh gía tỷ lệ mắc bệnh và tử vong của người dân trên 200 quốc gia và vũng lãnh thổ để cho thấy sự nguy hiểm và nhóm các nước bị ảnh hưởng nhiều nhất.

Nghiên cứu tiến hành theo 4 bước chính:

1. Tìm hiểu nghiệp vụ y tế liên quan đến covit-19.
2. Thu nhập và tiền xử lý dữ liệu.
3. Tìm hiểu bài toán phân cụm trong khai phá dữ liệu, lựa chọn thuật toán phù hợp với yêu cầu bài toán đặt ra và dữ liệu thu nhập được.
4. Hiện thực chương trình máy tính và đánh giá ý nghĩa thực tiễn.

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## 1.1 Đặt vấn đề

Công nghệ thông tin ngày càng được phổ biến đến tất cả mọi người, ứng dụng công nghệ thông tin vào việc lưu trữ và xử lý thông tin ngày nay được áp dụng hầu hết vào mọi loại lĩnh vực. điều này đã tạo ra một lượng lớn dữ liệu được lưu trữ với kích thước lớn. Đây chính là điều kiện tốt cho việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập bảng biểu và khai phá dữ liệu.

Trong lĩnh vực y khoa Việt Nam, hiện nay các tuyến y tế phường, xã, vùng sâu, vùng xa còn thiếu nhân lực y tế có trình độ chuyên môn và thiếu các trang thiết bị cần thiết trong chuẩn đoán bệnh. Vì vậy xây dựng hệ thống chuẩn đoán rất cần thiết cho ngành y tế hiện nay ở Việt Nam. Hệ hỗ trợ sẽ kết hợp với cán bộ y tế giúp chuẩn đoán sớm một số bệnh phát hiện sớm được những bệnh nguy hiểm và giảm gánh nặng kinh tế cho gia đình bệnh nhân và xã hội. Để minh chứng cho những lợi ích mà việc chuẩn đoán mang lại, đề tài chọn bộ dữ liệu về virus corona để thử nghiệm và đánh giá.

Ứng dụng kỹ thuật phân cụm dữ liệu trong khai phá dữ liệu nhằm xây dựng hệ thống đánh giá là một trong những hướng nghiên cứu chính của đề tài. Sau khi phân tích một số thuật toán cũng như đặc điểm của dữ liệu thu nhập được về visrus covid-19 , đề tài đề xuất ứng dụng mô hình phân cụm và thuật toán K-Means để tìm ra quy luật tiềm ẩn trong dữ liệu.

## 1.2 Cơ sở hình thành đề tài

WHO cuối cùng đã tuyên bố bệnh COVID-19 là đại dịch toàn cầu sau khi dịch bệnh lây lan nhanh tại nhiều quốc gia và vùng lãnh thổ trên thế giới. Bùng phát tại Trung Quốc hồi tháng 12/2019, tính tới tối 11/3 (theo giờ Việt Nam), dịch COVID-19 đã xuất hiện tại 121 quốc gia và vùng lãnh thổ. Tới nay, dịch bệnh đã cướp đi sinh mạng của 4.389 người, khiến 122.289 người mắc bệnh, song cũng có trên 67.000 người được điều trị thành công và khỏi bệnh.

Tâm dịch xuất hiện ở Trung Quốc, Italy, Iran, Hàn Quốc. Trung Quốc trên 80 nghìn ca nhiễm, Italy trên 10 nghìn ca nhiễm, Iran 9 nghìn ca nhiễm, Hàn Quốc trên 7,7 nghìn. Dịch cũng đã lan rộng ra 47 quốc gia/vùng lãnh thổ của châu Âu. Một vài quốc gia châu Âu đang là điểm nóng ngoài Italy như Tây Ban Nha trên 2 nghìn ca nhiễm, Đức trên 1,6 nghìn ca nhiễm. Ở bên kia bờ đại dương, Hoa Kỳ cũng ghi nhận trên 1 nghìn ca nhiễm.

Trong vòng 2 tuần qua, số ca nhiễm COVID-19 bên ngoài Trung Quốc đã tăng gấp 13 lần, số lượng quốc gia/vùng lãnh thổ ghi nhận ca mắc tăng gấp 3 lần, người đứng đầu WHO thừa nhận. WHO lo ngại trong những ngày tới và trong những tuần sắp tới, số ca mắc, số lượng người tử vong cũng như số lượng quốc gia/vùng lãnh thổ có ca mắc COVID-19 sẽ còn tăng lên.

Đề tài áp dụng môn khai phá dữ liệu xây dựng đánh giá các tỷ lệ với bộ dữ liệu thu thập được từ trong nước và ngoài nước

## 1.3 Mục tiêu đề tài

Đề tài tập chung vào nghiên cứu kỹ thuật phân cụm trong khai phá dữu liệu, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và xây dựng ứng dụng cụ thể. Sau khi phân tích đặc điểm của dữu liệu thu nhập được và lựa chọn giải thuật phù hợp với dữ liệu, việc xây dựng và đánh giá chất lượng, độ hiệu quả của hệ thống cũng là mục tiêu chính của đề tài.

## 1.4 Đối tượng và phương pháp nghiên cứu

Đề tài tập chung vào nghiên cứu kỹ thuật phân cụm trong khai phá dữ liệu cụ thể là nghiên cứu thuật toán k-means để áp dụng vào việc phân tích cơ sở dữ liệu tỷ lệ mắc và chết của covid-19 . Thu nhập dữ liệu mắc và tử vong do dịch covid 19 gây ra được thống kê có kiểm chứng bởi các tổ chức nước ngoài.

Dữ liệu được thu nhập từ 237 quốc gia với các trường hợp mắc, tử vong, số ca hồi phục …. Sử dụng phân cụm dữ liệu nhằm định các nhóm người bệnh nhằm cung cấp thông tin cho việc phân phố các thuốc điều trị trong y tế.

## 1.5 Ý nghĩa đề tài

### 1.5.1 Ý nghĩa khoa học

Với sự trợ giúp của máy tính, đề tài đóng góp một biện pháp thực hiện hỗ trợ các cán bộ y tế đánh giá bệnh cho bệnh nhân. Kết quả, Kinh nghiệm thu đưuọc khi thực hiện đề tài này sẽ giúp các cán bộ y tế phát hiện sớm bệnh cho bệnh nhân, đồng thời mong muốn những người đang công tác trong lĩnh vực y khoa và Khoa học máy tính ngồi lại với nhau để tìm ra những giải pháp tốt hơn trong vấn đề điều trị bệnh bằng cách kết hợp giữa 2 lĩnh vực y học và khoa học máy tính.

### 1.5.2 Ý nghĩa thực tiễn

Đánh giá tỷ lệ nhiễm , chết do virus và phát hiện bệnh là cả một quá trình, đòi hỏi các cán bộ y tế không những phải thật vững chuyên môn mà còn có đầy đủ các trang thiết bị y tế mới có thể chuẩn đoán chính xác bệnh cho bệnh nhân. Nếu chuẩn đoán sai bệnh sẽ đưa đến điều trị sai, không phát hiện sớm bệnh cho bệnh nhân,…

# CHƯƠNG 2: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

## 2.1 Tổng quan về kỹ thuật Khai phá dữ liệu (Data mining)

### 2.1.1 Khái niệm về khai phá dữ liệu

**Khai phá dữ liệu** (datamining) được định nghĩa như là một quá trình chắt lọc hay khai phá tri thức từ một lượng lớn dữ liệu. Một ví dụ hay được sử dụng là là việc khai thác vàng từ đá và cát, Dataming được ví như công việc "Đãi cát tìm vàng" trong một tập hợp lớn các dữ liệu cho trước. Thuật ngữ Dataming ám chỉ việc tìm kiếm một tập hợp nhỏ có giá trị từ một số lượng lớn các dữ liệu thô. Có nhiều thuật ngữ hiện được dùng cũng có nghĩa tương tự với từ Datamining như Knowledge Mining (khai phá tri thức), knowledge extraction(chắt lọc tri thức), data/patern analysis(phân tích dữ liệu/mẫu), data archaeoloogy (khảo cổ dữ liệu), datadredging(nạo vét dữ liệu),...

**Định nghĩa:** Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó. Khai phá dữ liệu là một bước trong bảy bước của quá trình KDD (Knowleadge Discovery in Database) và KDD được xem như 7 quá trình khác nhau theo thứ tự sau:

1. Làm sạch dữ liệu (data cleaning & preprocessing)s: Loại bỏ nhiễu và các dữ

liệu không cần thiết.

2. Tích hợp dữ liệu: (data integration): quá trình hợp nhất dữ liệu thành những

kho dữ liệu (data warehouses & data marts) sau khi đã làm sạch và tiền xử lý (data

cleaning & preprocessing).

3. Trích chọn dữ liệu (data selection): trích chọn dữ liệu từ những kho dữ liệu

và sau đó chuyển đổi về dạng thích hợp cho quá trình khai thác tri thức. Quá trình này bao gồm cả việc xử lý với dữ liệu nhiễu (noisy data), dữ liệu không đầy đủ

(incomplete data), .v.v.

4. Chuyển đổi dữ liệu: Các dữ liệu được chuyển đổi sang các dạng phù hợp

cho quá trình xử lý

5. Khai phá dữ liệu(data mining): Là một trong các bước quan trọng nhất, trong đó sử dụng những phương pháp thông minh để chắt lọc ra những mẫu dữ liệu.

6. Ước lượng mẫu (knowledge evaluation): Quá trình đánh giá các kết quả tìm

được thông qua các độ đo nào đó.

7. Biểu diễn tri thức (knowledge presentation): Quá trình này sử dụng các kỹ

thuật để biểu diễn và thể hiện trực quan cho người dùng.

### 2.1.2 Quy trình khai phá dữ liệu

1) Nghiên cứu lĩnh vực

Ta cần nghiên cứu lĩnh vực cần sử dụng Data mining để xác định được những tri thức ta cần chắt lọc, từ đó định hướng để tránh tốn thời gian cho những tri thức không cần thiết.

2) Tạo tập tin dữ liệu đầu vào

Ta xây dựng tập tin để lưu trữ các dữ liệu đầu vào để máy tính có thể lưu trữ và xử lý.

3) Tiền xử lý, làm sạch, mã hóa

Ở bước này ta tiến hành bỏ bớt những dữ liệu rườm rà, không cần thiết, tinh chỉnh lại cấu trúc của dữ liệu và mã hóa chúng để tiện cho quá trình xử lý.

4) Rút gọn chiều

Thông thường một tập dữ liệu có chiều khá lớn sẽ sinh ra một lượng dữ liệu khổng lồ, ví dụ với n chiều ta sẽ có 2^n tổ hợp. Do đó, đây là một bước quan trọng giúp giảm đáng kể hao tổn về tài nguyên trong quá trình xử lý tri thức. Thông thường ta sẽ dùng Rough set (http://en.wikipedia.org/wiki/Rough\_set) để giảm số chiều.

5) Chọn tác vụ khai thác dữ liệu

Để đạt được mục đích ta cần, ta cần chọn được tác vụ khai thác dữ liệu sao cho phù hợp. Thông thường có các tác vụ sau:

\* Đặc trưng (feature)

\* Phân biệt (discrimination)

\* Kết hợp (association)

\* Phân lớp (classification)

\* Gom cụm (clusterity)

\* Xu thế (trend analysis)

\* Phân Tích độ lệch

\* Phân tích hiếm

6) Chọn các thuật giải Khai thác dữ liệu

7) Khai thác dữ liệu: Tìm kiếm tri thức

Sau khi tiến hành các bước trên thì đây là bước chính của cả quá trình, ta sẽ tiến hành khai thác và tìm kiếm tri thức.

8) Đánh giá mẫu tìm được

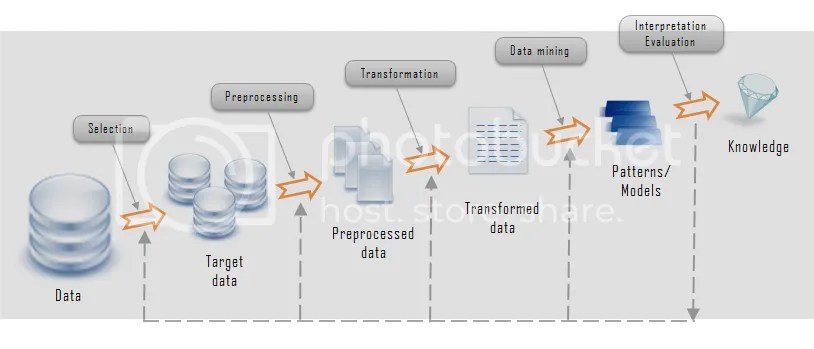
Ta cần đánh giá lại xem trong các tri thức tìm được, ta sẽ sử đụng được những tri thức nào, những tri thức nào dư thừa, không cần thiết

9) Biểu diễn tri thức

Ta biểu diễn tri thức vừa thu thập được dưới dạng ngôn ngữ tự nhiên và hình thức sao cho người dùng có thể hiểu được những tri thức đó.

10) Sử dụng các tri thức vừa khám phá

Ta có thể tham khảo tiến trình KDD (Knowledge Discovery in Databases) để hiểu rõ hơn về khai thác dữ liệu:



*Hình 2. 1 Knowledge Discovery in Databases*

Chuẩn bị dữ liệu (data preparation), bao gồm các quá trình làm sạch dữ liệu (data cleaning), tích hợp dữu liệu ( data integration), chọn dữ liệu (data selection), biến đổi dữ liệu (data transformation).

Khai thác dữu liệu (data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữu liệu. Kết quả cho ta một nguồn tri thức thô.

Đánh giá (evaluation): dựa trên một tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.

Triển khai (deployment).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đên bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

### 2.1.3 Ứng dụng của khai phá dữ liệu

Kinh tế - ứng dụng trong kinh doanh, tài chính, tiếp thị bán hàng, bảo hiểm, thương mại, ngân hàng,.. Đưa ra các bản báo cáo giàu thông tin, phân tích rửi ro trước khi đưa ra các chiến lược kinh doanh, sản xuất, phân loại khách hàng từ đó phân định ra thị trường, thị phần:…

Khoa học: Thiên văn học - dự đoán đường đi các thiên thể, hành tinh,...;

Công nghệ sinh học – tìm ra các gen mới, cây con giống mới,…

Web: các công cụ tìm kiếm.

## 2.2 Phân cụm dữ liệu và ứng dụng

### 2.2.1 Mục đích của phân cụm dữ liệu

Phân loại là một trong những hành vi nguyên thuỷ nhất của con người nhằm nắm giữ lượng thông tin khổng lồ họ nhận được hằng ngày vì sự xử lý mọi thông tin như một thực thể đơn lẻ là không thể. Phân cụm dữ liệu nhằm mục đích chính là khai phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó, cho phép người ta đi sâu vào phân tích và nghiên cứu cho từng cụm dữ liệu này nhằm khai phá và tìm kiếm các thông tin tiềm ẩn, hữu ích phục vụ cho ra quyết định.

Một vài ví dụ về ý nghĩa thực tiễn của phân cụm dữ liệu như sau :

* *Khám phá ra các vị trí địa lý thuận lợi cho việc xây dựng các kho hàng phục vụ mua và bán hàng của một công ty thương mại.*
* *Xác định các cụm ảnh như ảnh của các loài động vật như loài thú, chim,...trong tập CSDL ảnh về động vật nhằm phục vụ cho việc tìm kiếm ảnh - Xác định các nhóm người bệnh nhằm cung cấp thông tin cho việc phân phối*

*các thuốc điều trị trong y tế*

* *Xác định nhóm các khách hàng trong CSDL ngân hàng có vốn các đầu tư*

*vào bất động sản cao...*

Như vậy, phân cụm dữ liệu là một phương pháp xử lý thông tin quan trọng và phổ biến, nó nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm tương tự.

Tiếp theo, giả sử rằng tất cả các dạng dữ liệu được biểu diễn bởi khái niệm đặc trưng, các đặc trưng hình thành nên vector đặc trưng l- chiều. Thuật ngữ phân cụm được hiểu là phân cụm dữ liệu.

### 2.2.2 Các bước cơ bản để phân cụm

**Chọn lựa đặc trưng** : Các đặc trưng phải được chọn lựa một cách hợp lý để có thể “mã hoá” nhiều nhất thông tin liên quan đến công việc quan tâm. Mục tiêu chính là phải giảm thiểu sự dư thừa thông tin giữa các đặc trưng. Các đặc trưng cần được tiền xử lý trước khi dùng trong các bước sau.

**Chọn độ đo gần gũi** : Đây là một độ đo chỉ ra mức độ tương tự hay không tương tự giữa hai vector đặc trưng. Phải đảm bảo rằng tất cả các vector đặc trưng góp phần như nhau trong việc tính toán độ đo gần gũi và không có đặc trưng nào át hẳn đặc trưng nào. Điều này được đảm nhận bởi quá trình tiền xử lý.

**Tiêu chuẩn phân cụm** : Điều này phụ thuộc vào sự giải thích của chuyên gia cho thuật ngữ “dễ nhận thấy” dựa vào loại của các cụm được chuyên gia cho rằng đang ẩn dấu dưới tập dữ liệu. Chẳng hạn, một cụm loại chặt (compact)

của các vector đặc trưng trong không gian l chiều có thể dễ nhận thấy theo một tiêu chuẩn, trong khi một cụm loại “dài và mỏng” lại có thể đươc dễ nhận thấy bởi một tiêu chuẩn khác. Tiêu chuẩn phân loại có thể được diễn đạt bởi

hàm chi phí hay một vài loại quy tắc khác.

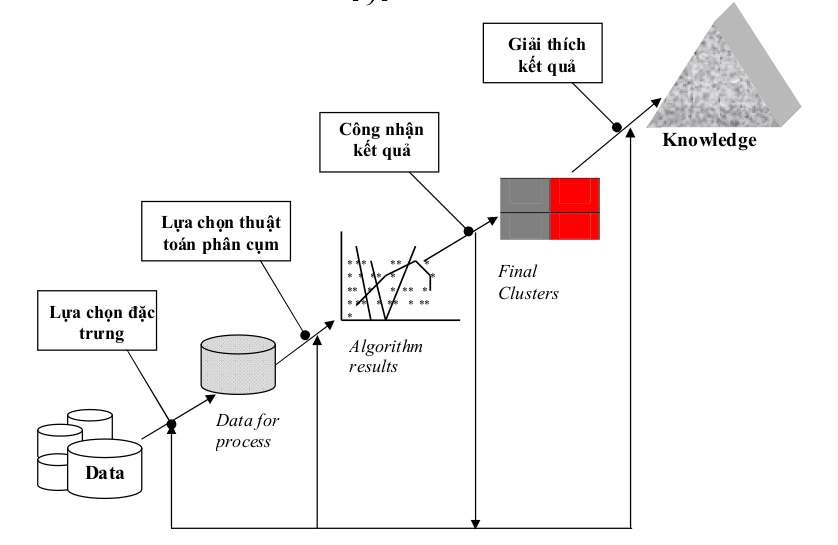
**Thuật toán phân loại** : Cần lựa chọn một sơ đồ thuật toán riêng biệt nhằm làm sáng tỏ cấu trúc cụm của tập dữ liệu.

**Công nhận kết quả** : Khi đã có kết quả phân loại thì ta phải kiểm tra tính đúng đắn của nó. Điều này thường được thực hiện bởi việc dùng các kiểm định phù hợp.

**Giải thích kết quả** : Trong nhiều trường hợp, chuyên gia trong lĩnh vực ứng dụng phải kết hợp kết quả phân loại với bằng chứng thực nghiệm và phân tích để đưa ra các kết luận đúng đắn. Trong một số trường hợp, nên có cả bước khuynh hướng phân cụm; trong bước này có các kiểm định khác nhau để chỉ ra một dữ liệu có hay không một cấu trúc phân cụm. Ví dụ như tập dữ liệu của

ta có thể hoàn toàn ngẫu nhiên vì vậy mọi cố gắng phân cụm đều vô nghĩa.

Các lựa chọn khác nhau của các đặc trưng, độ đo gần gũi, tiêu chuẩn phân cụm có thể dẫn tới các kết quả phân cụm khác nhau. Do đó, việc lựa chọn một cách hợp lý nhất hoàn toàn dựa vào kiến thức và kinh nghiệm của chuyên gia. Tính chủ quan (của chuyên gia) là một thực tế mà ta phải chấp nhận.



*Hình 2. 3 Các bước trong quá trình phân cụm*

### 2.2.3 Các ứng dụng của phân cụm

Phân cụm là một công cụ quan trọng trong một số ứng dụng. Sau đây là một số ứng dụng của nó:

* **Giảm dữ liệu**: Giả sử ta có một lượng lớn dữ liệu (N). Phân cụm sẽ nhóm các dữ liệu này thành m cụm dữ liệu dễ nhận thấy và m << N. Sau đó xử lý mỗi cụm như một đối tượng đơn.
* **Rút ra các giả thuyết** : Các giả thuyết này có liên quan đến tính tự nhiên của dữ liệu và phải được kiểm tra bởi việc dùng một số tập dữ liệu khác.
* **Kiểm định giả thuyết** : Ta sẽ phân cụm để xét xem có tồn tại một tập dữ liệu nào đó trong tập dữ liệu thoả mãn các giả thuyết đã cho hay không. Chẳng hạn xem xét giả thuyết sau đây: “Các công ty lớn đầu tư ra nước ngoài “. Để kiểm tra, ta áp dụng kỹ thuật phân cụm với một tập đại diện lớn các công ty. Giả sử rằng mỗi công ty được đặc trưng bởi tầm vóc, các hoạt động ở nước ngoài và khả năng hoàn thành các dự án. Nếu sau khi phân cụm, một cụm các công ty được hình thành gồm các công ty lớn và có vốn đầu tư ra nước ngoài (không quan tâm đến khả năng hoàn thành các dự án) thì giả thuyết đó được

củng cố bởi kỹ thuật phân cụm đã thực hiện.

* **Dự đoán dựa trên các cụm** : Đầu tiên ta sẽ phân cụm một tập dữ liệu thành các cụm mang đặc điểm của các dạng mà nó chứa. Sau đó, khi có một dạng mới chưa biết ta sẽ xác định xem nó sẽ có khả năng thuộc về cụm nào nhất và dự đoán được một số đặc điểm của dạng này nhờ các đặc trưng chung của cả cụm.

Cụ thể hơn, phân cụm dữ liệu đã được áp dụng cho một số ứng dụng điển hình trong các lĩnh vực sau [13] :

* *Thương mại* : Trong thương mại, phân cụm có thể giúp các thương nhân khám phá ra các nhóm khách hàng quan trọng có các đặc trưng tương đồng nhau và đặc tả họ từ các mẫu mua bán trong cơ sở dữ liệu khách hàng. Sinh học : Trong sinh học, phân cụm được sử dụng để xác định các loại sinh vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu.
* *Phân tích dữ liệu không gian* : Do sự đồ sộ của dữ liệu không gian như dữ liệu thu được từ các hình ảnh chụp từ vệ tinh các thiết bị y học hoặc hệ thống thông tin địa lý (GIS), ...làm cho người dùng rất khó để kiểm tra các dữ liệu không gian một cách chi tiết. Phân cụm có thể trợ giúp người dùng tự động phân tích và xử lý các dữ liệu không gian như nhận dạng và chiết xuất các đặc tính hoặc các mẫu dữ liệu quan tâm có thể tồn tại trong cơ sở dữ liệu không gian.
* *Lập quy hoạch đô thị* : Nhận dạng các nhóm nhà theo kiểu và vị trí địa lý,...nhằm cung cấp thông tin cho quy hoạch đô thị.
* *Nghiên cứu trái đất* : Phân cụm để theo dõi các tâm động đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm.
* *Địa lý* : Phân lớp các động vật và thực vật và đưa ra đặc trưng của chúng.
* *Web Mining* : Phân cụm có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc khám phá tri thức từ dữ liệu,...

## 2.3 Cơ sở dữ liệu về Covid 19

### 2.3.1 Thông tin về dịch Covid 19

COVID-19 (bệnh vi-rút corona 2019) là một bệnh do vi-rút có tên SARS-CoV-2 gây ra và được phát hiện vào ​​​​​ tháng 12 năm 2019 ở Vũ Hán, Trung Quốc. Căn bệnh này rất dễ lây lan và đã nhanh chóng lan ra khắp thế giới.

COVID-19 thường gây ra các triệu chứng hô hấp, có thể cảm thấy giống như cảm lạnh, cúm hoặc viêm phổi. COVID-19 có thể tấn công không chỉ phổi và hệ hô hấp của quý vị. Các bộ phận khác của cơ thể quý vị cũng có thể bị ảnh hưởng bởi căn bệnh này.

* Hầu hết những người bị COVID-19 có các triệu chứng nhẹ, nhưng một số người trở nên nặng.
* Một số người, kể cả những người có triệu chứng nhẹ hoặc không có triệu chứng có thể bị hội chứng hậu COVID - hoặc "di chứng COVID".
* Những người cao tuổi hoặc người có các bệnh lý nền nhất định có nguy cơ cao mắc bệnh nghiêm trọng do COVID-19.
* Hàng nghìn người đã tử vong do COVID-19 tại Hoa Kỳ.
* Vắc-xin ngừa COVID-19 là an toàn và hiệu quả.

### 2.3.2 Dấu hiệu và triệu chứng

Những người mắc bệnh COVID-19 đã có một loạt các triệu chứng được báo cáo - từ các triệu chứng nhẹ đến bệnh nghiêm trọng. Các triệu chứng có thể xuất hiện 2-14 ngày sau khi phơi nhiễm với vi-rút. Bất cứ ai cũng có thể có các triệu chứng từ nhẹ đến nghiêm trọng. Những người có các triệu chứng này có thể nhiễm COVID-19:

* Sốt hoặc ớn lạnh
* Ho
* Hụt hơi hoặc khó thở
* Mệt mỏi
* Đau cơ hoặc đau người
* Đau đầu, họng
* Ngạt mũi hoặc chảy nước mũi
* Buồn nôn hoặc nôn mửa
* Tiêu chảy

Danh sách này không bao gồm tất cả các triệu chứng có thể xuất hiện. CDC sẽ tiếp tục cập nhật danh sách này khi chúng tôi tìm hiểu thêm về COVID-19. Người cao tuổi và những người có các tình trạng bệnh nền nghiêm trọng như bệnh tim hoặc phổi hoặc bệnh tiểu đường dường như có nguy cơ cao hơn phát triển các biến chứng nghiêm trọng hơn từ bệnh

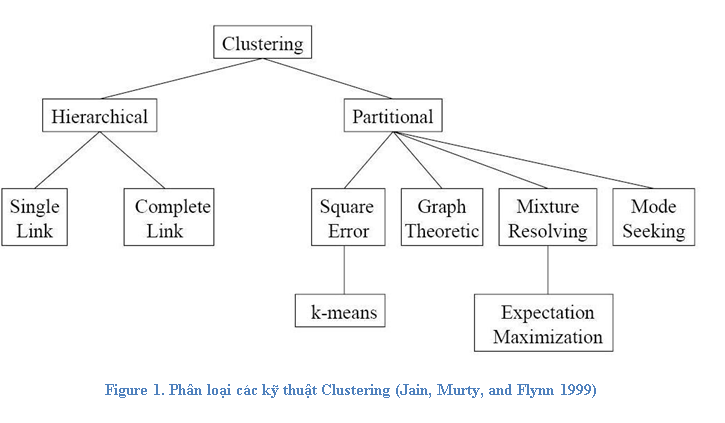
# CHƯƠNG 3: KỸ THUẬT PHÂN CỤM VÀ THUẬT TOÁN K-MEANS

## 3.1 Giới thiệu về kỹ thuật phân cụm trong Khai phá dữ liệu

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp **Unsupervised Learning** trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu.

Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, “natural clusters”, “useful” clusters, outlier detection.

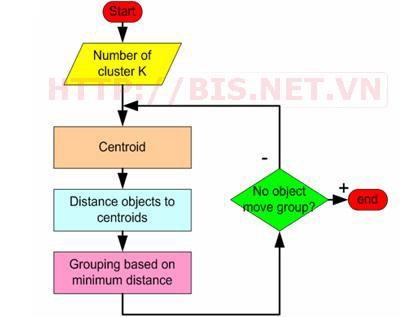


*Hình 3. 1 Các kỹ thuật phân cụm*

## 3.2 Thuật Toán K-Means

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid ) là nhỏ nhất.

***Thuật toán K-Means được mô tả như sau***



*Hình 3. 2 Mô tả thuật toán K-Means*

***Thuật toán K-Means thực hiện qua các bước chính sau:***

1. Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.
2. Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean)
3. Nhóm các đối tượng vào nhóm gần nhất
4. Xác định lại tâm mới cho các nhóm
5. Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng

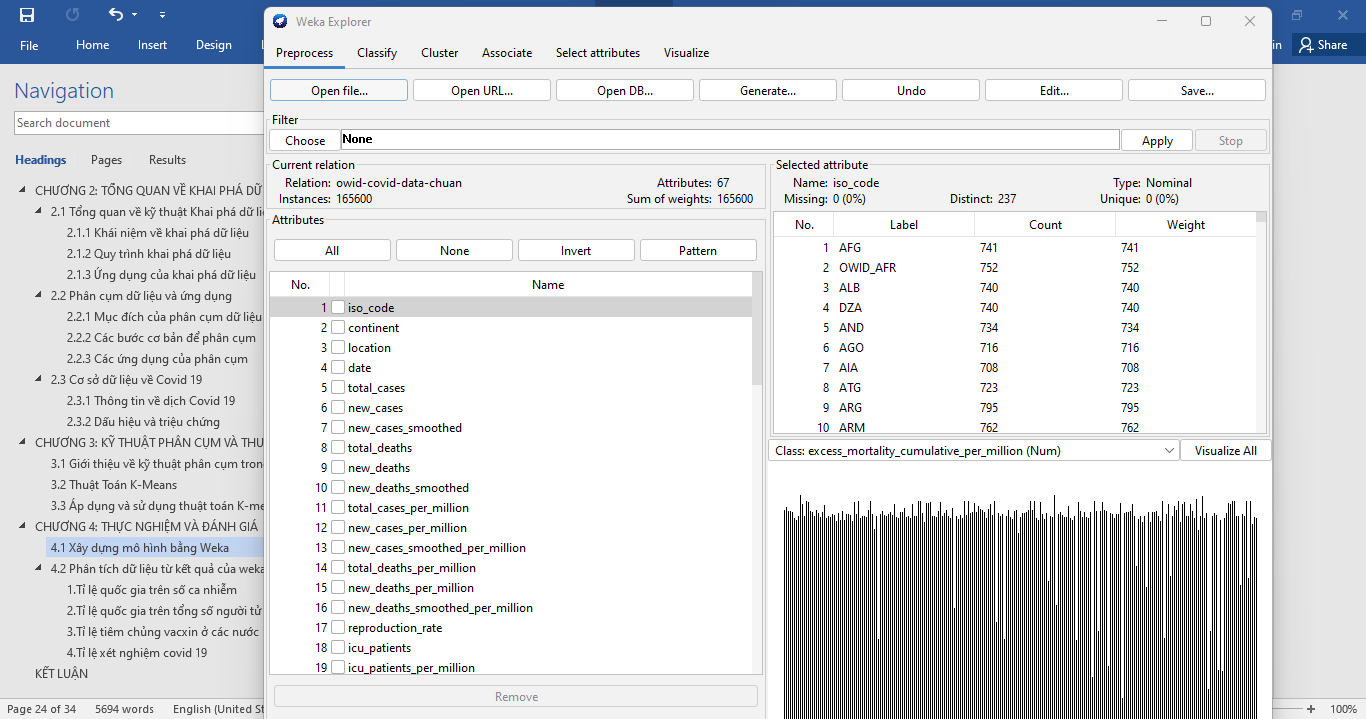
## 3.3 Áp dụng và sử dụng thuật toán K-means vào bộ dataset Covid-19

Tập dữ liệu covid-19 bao gồm dữ liệu của 200 quốc gia gồm những nước đã có người chết vì virus covid-19 và các nước chưa có nguời chết vì covid-19. Tập dữ liệu bao gồm các thuộc tính như sau:

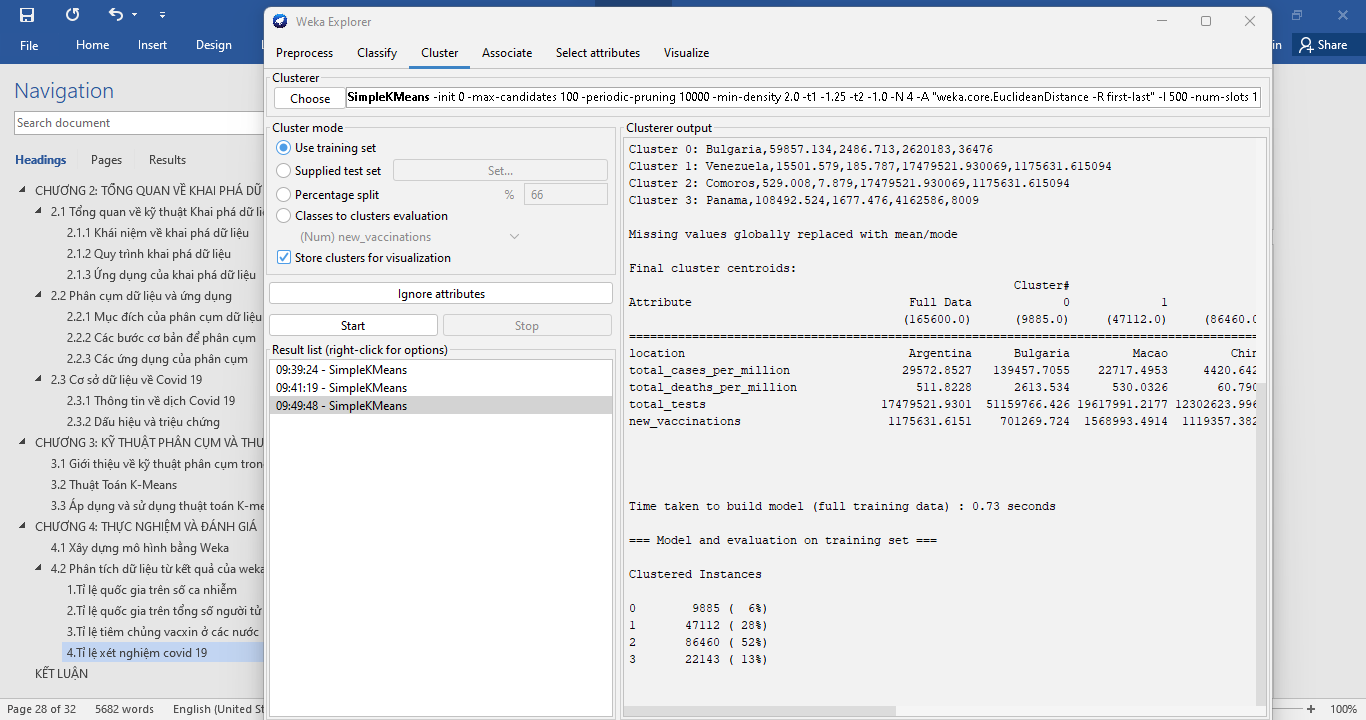
1. Quốc Gia
2. Tỉ lệ người nhiễm trên 1 triệu người.
3. Tỷ lệ người chết trên 1 triệu người.
4. Tỷ lệ tiêm chủng vacxin trên 1 trăm người
5. Tỷ lệ xét nghiệm trên 1 triệu người

# CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ

## 4.1 Xây dựng mô hình bằng Weka



*Hình 4. 1 Nhập dữ liệu từ file csv vào Weka*



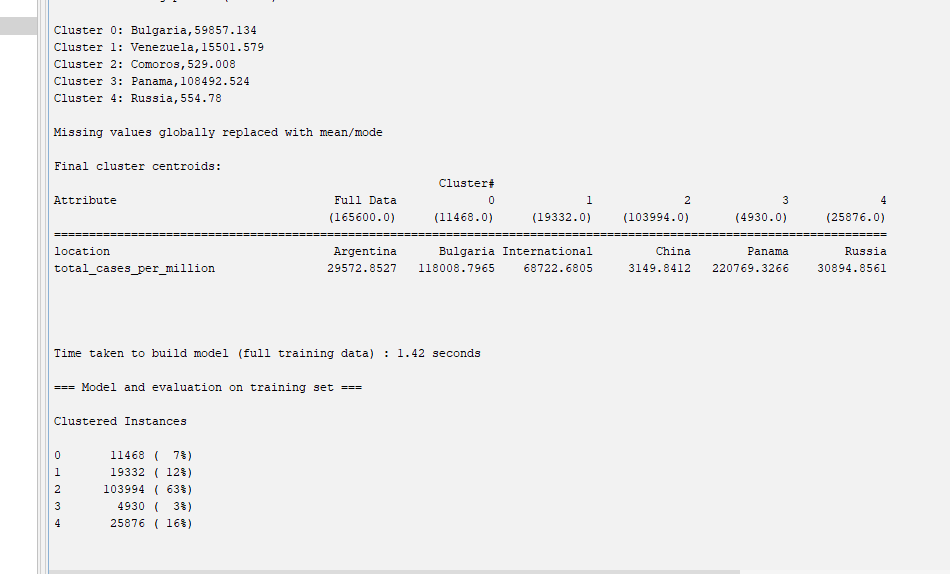
*Hình 4. 4 Đầu ra phân lớp cho tất cả các thuộc tính ( 5 Cluster)*

**• Đầu ra phân lớp:** Các thuộc tính:

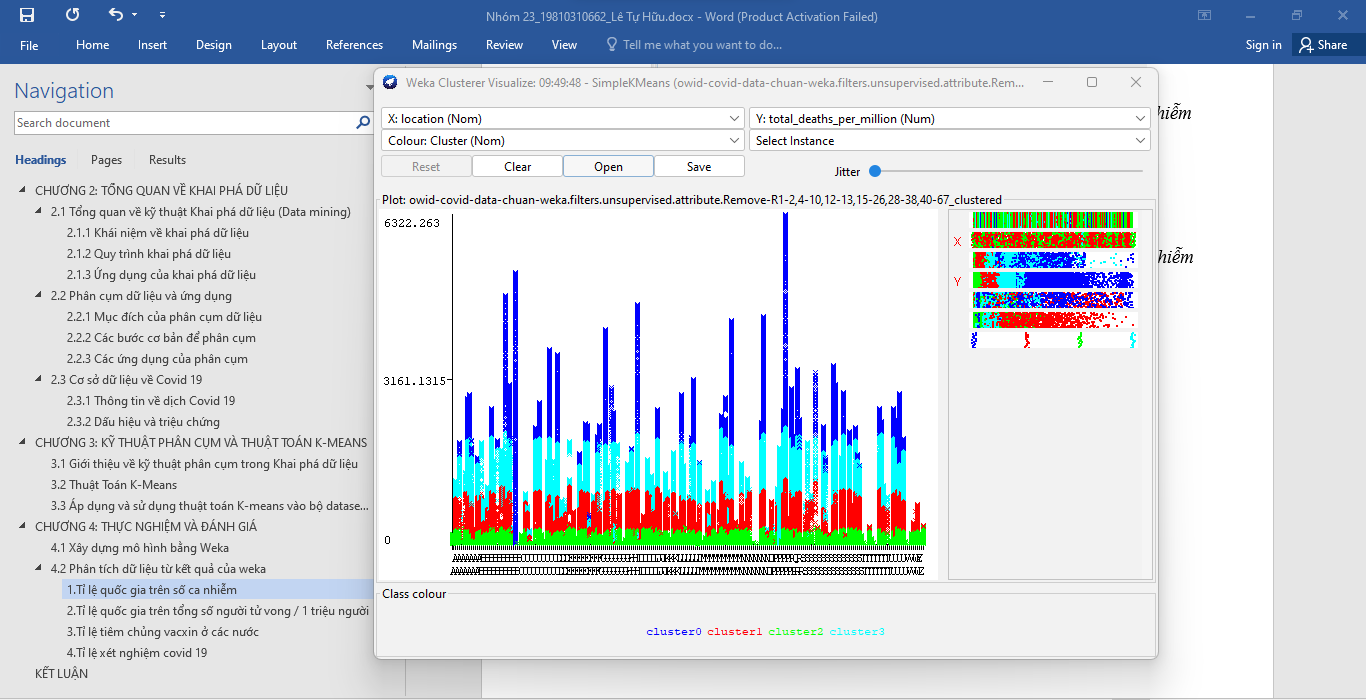
1. Quốc Gia
2. Tỉ lệ người nhiễm trên 1 triệu người.
3. Tỷ lệ người chết trên 1 triệu người.
4. Tỷ lệ tiêm chủng vacxin trên 1 triệu người
5. Tỷ lệ xét nghiệm trên 1 triệu người

## 4.2 Phân tích dữ liệu từ kết quả của weka

### 1.Tỉ lệ quốc gia trên số ca nhiễm

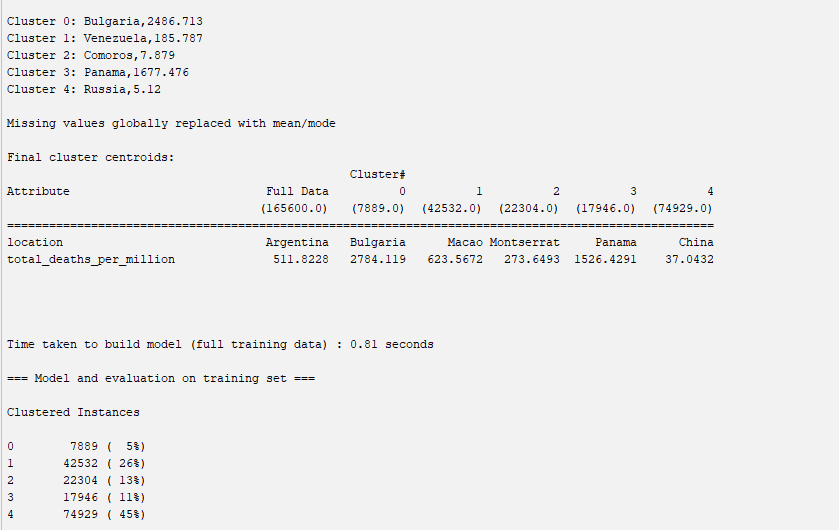


*Hình 4. 5 Đầu ra phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ nhiễm trên 1 triệu người*

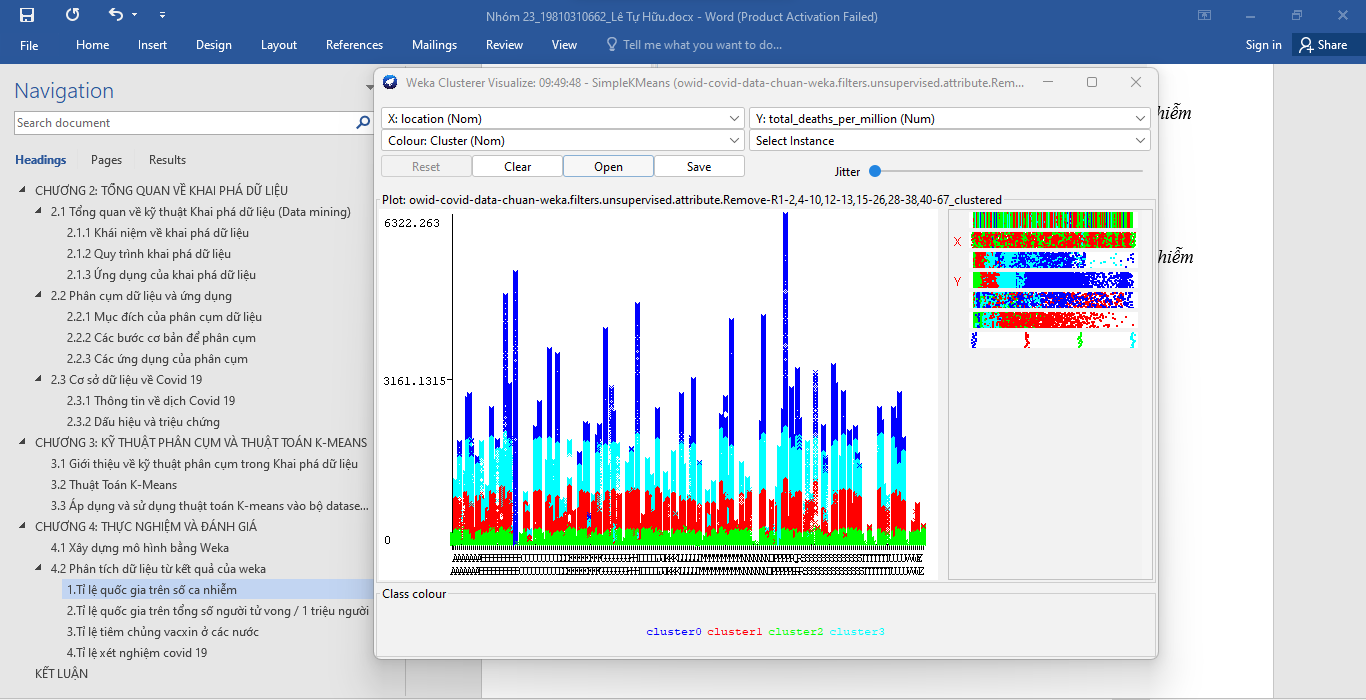
****

*Hình 4. 6 Biểu đồ phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ nhiễm trên 1 triệu người*

### 2.Tỉ lệ quốc gia trên tổng số người tử vong / 1 triệu người

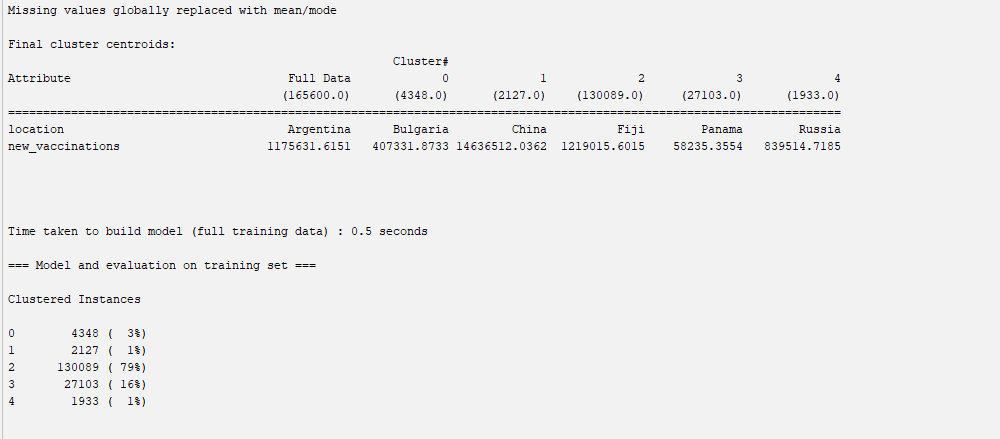


*Hình 4. 5 Đầu ra phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ ca tử cong trên 1 triệu người*

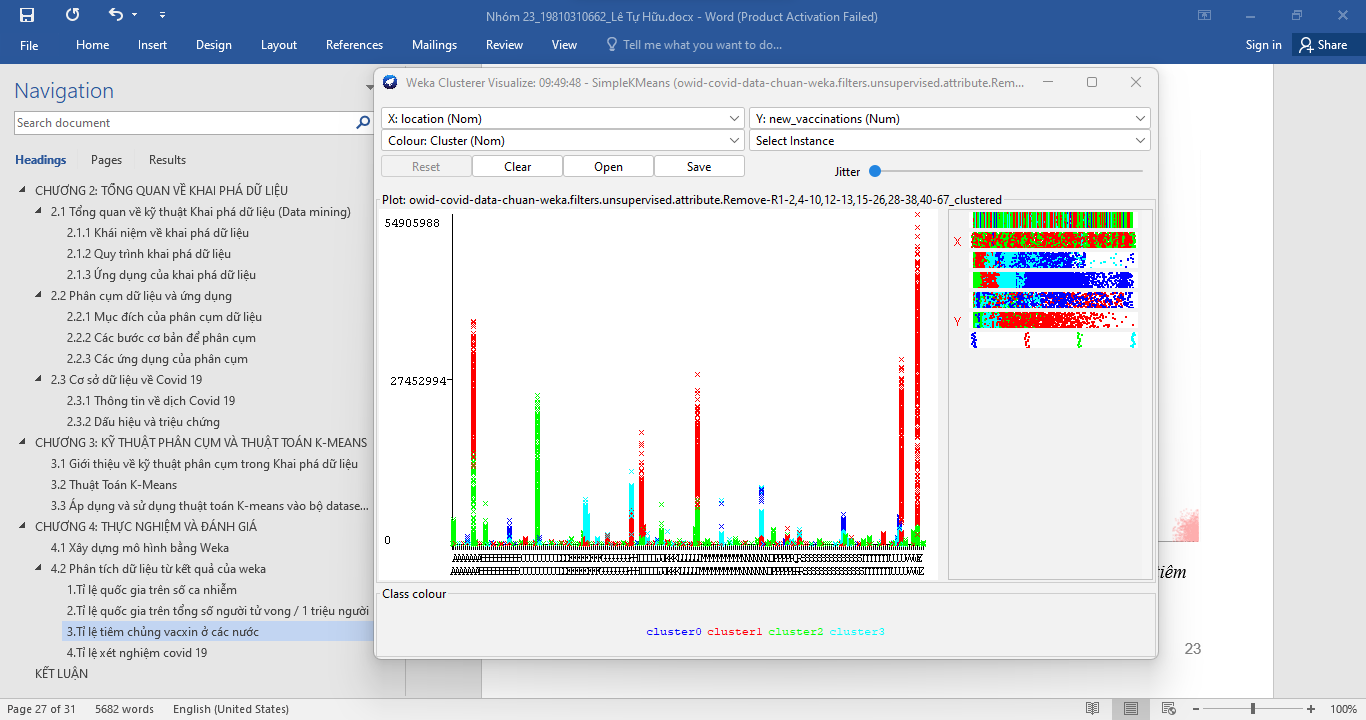
****

*Hình 4. 6 Biểu đồ phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ nhiễm trên 1 triệu người*

### 3.Tỉ lệ tiêm chủng vacxin ở các nước

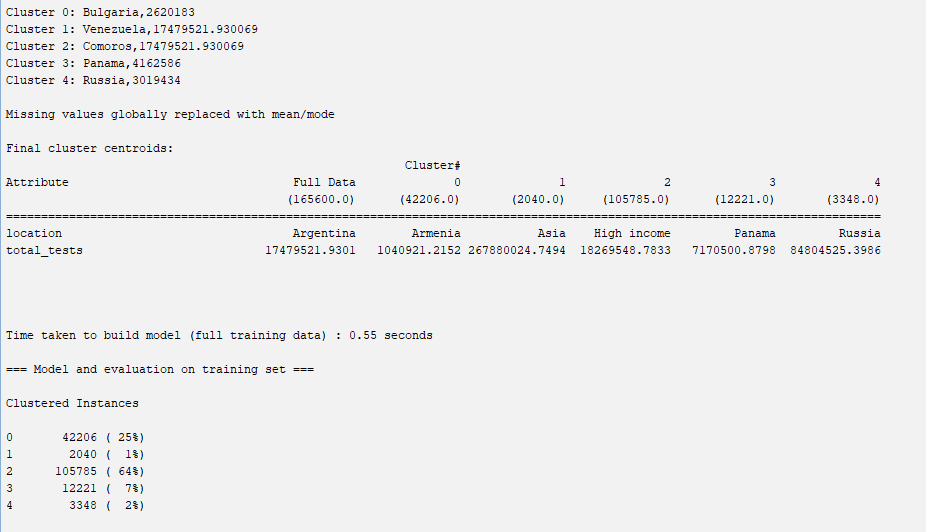


*Hình 4. 5 Đầu ra phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ tiêm chủng vacxin*

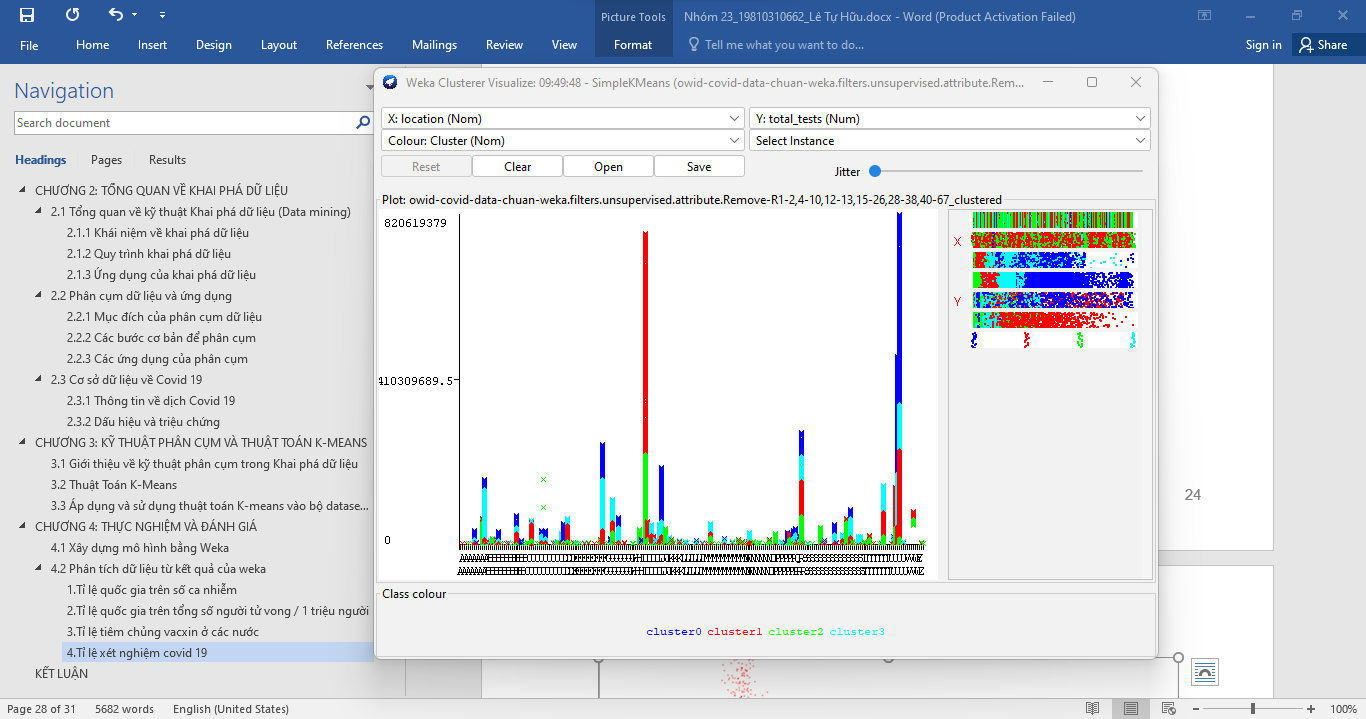
****

*Hình 4. 6 Biểu đồ phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ tiêm chủng vacxin*

### 4.Tỉ lệ xét nghiệm covid 19



*Hình 4. 5 Đầu ra phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ xét nghiệm covid 19*

****

*Hình 4. 6 Biểu đồ phân cụm bằng K-means với thuộc tính quốc gia và tỷ lệ tiêm chủng vacxin*

# KẾT LUẬN

Sau khi hoàn thành nội dung về báo em đã tổng kết và đưa ra khái niệm về kết quả sau như tìm hiểu được về khai phá dữ liệu, nêu ra tổng quan về đề tài, phương pháp nghiên cứu và thực nghiệm đánh giá kết quả về đề tài trên phần mềm Weka. Kết quả đạt được vẫn chưa thật sự tốt, và vẫn cần có nhiều điểm thiếu sót cần được khắc phục nhằm đem lại kết quả chính xác hơn khi thực nghiệm các lần kế tiếp.

# TÀI LIỆU THAM KHẢO

1 Giáo trình Khai phá dữ liệu - đại học Điện Lực

2 Dữ liệu về covid : <https://vietnamese.cdc.gov/coronavirus/2019-nCoV/index.html>