**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**----------------------------------------**  
  


**KHAI PHÁ DỮ LIỆU**

**BÁO CÁO BÀI TẬP LỚN**

**Đề tài: Phân cụm dữ liệu bệnh nhân sử dụng**

**phương pháp phân cụm K-Means**

Nhóm sinh viên thực hiện: **Nhóm 1**

*Hoàng Anh Tuấn*

*Nguyễn Phú Trường*

*Vũ Văn Thủy*

*Thân Thị Thùy*

**Giảng viên hướng dẫn: TS.Trần Mạnh Tuấn**

**Hà Nội, Tháng 8 Năm 2023.**

# Lời nói đầu

*Trong thời đại ngày nay, sự phát triển không ngừng của khoa học và công nghệ đã mở ra một thế giới tri thức vô cùng phong phú và đa dạng. Kho dữ liệu to lớn và nguồn tri thức ngày càng mở rộ đã tạo ra một thách thức đầy thú vị và quan trọng: làm thế nào chúng ta có thể tận dụng tri thức đó để tạo ra giá trị thực sự cho xã hội và cuộc sống?*

*Trong bối cảnh hiện nay, việc tìm kiếm, xử lý và khai thác thông tin đã trở thành một nhiệm vụ cấp thiết, đặc biệt là trong lĩnh vực quản lý và hoạt động kinh doanh. Để đáp ứng yêu cầu này, nhiều công cụ tìm kiếm và xử lý thông tin đã được phát triển, nhưng vẫn còn nhiều thách thức phải đối mặt. Các phương pháp truyền thống khai thác cơ sở dữ liệu thường không đủ mạnh mẽ để hiểu biết sâu rộng về tri thức ẩn sau dữ liệu.*

*Nhận thức về những thách thức này, nhóm em đã đặt mục tiêu nghiên cứu vào một hướng mới mẻ: "****Khai phá dữ liệu bệnh nhân sử dụng phương pháp phân cụm K-Means****". Mục đích của báo cáo này là tìm hiểu và áp dụng kỹ thuật khai phá dữ liệu để phân loại tình trạng sức khỏe dựa vào dữ liệu của bệnh nhân. Nhóm em hy vọng rằng việc áp dụng phương pháp phân cụm K-Means sẽ giúp chúng ta nhận biết các nhóm bệnh nhân có đặc điểm tương tự, từ đó đưa ra thông tin về nguy cơ mắc bệnh và cách phòng tránh phù hợp cho bệnh nhân.*

*Trong phạm vi báo cáo này, nhóm em sẽ trình bày quy trình thực hiện nghiên cứu, từ việc thu thập dữ liệu, tiền xử lý, áp dụng phương pháp K-Means, đến việc phân tích và trình bày kết quả. Hy vọng rằng thông qua nỗ lực nghiên cứu này, nhóm em sẽ đóng góp một phần nhỏ vào việc tận dụng tri thức từ dữ liệu để nâng cao chất lượng quản lý chăm sóc sức khỏe và cải thiện cuộc sống của bệnh nhân.*

**MỤC LỤC**

[***Lời nói đầu 2***](#_Toc143985833)

[***Bảng phân chia công việc 4***](#_Toc143985834)

[***Chương 1: Tổng quan về khai phá dữ liệu 5***](#_Toc143985835)

[*1.1. Phát hiện tri thức và khai phá dữ liệu 5*](#_Toc143985836)

[*1.2. Quy trình khám phá tri thức trong CSDL 5*](#_Toc143985837)

[*1.3. Một số kỹ thuật khai phá dữ liệu 6*](#_Toc143985838)

[***Chương 2: Mô tả bài toán và phân tích dữ liệu thô 6***](#_Toc143985839)

[*2.1. Mô tả bài toán 6*](#_Toc143985840)

[*2.2. Phân tích dữ liệu thô 6*](#_Toc143985841)

[***Chương 3: Tiền xử lý dữ liệu 8***](#_Toc143985842)

[*3.1. Làm sạch dữ liệu 8*](#_Toc143985843)

[*3.2. Tích hợp dữ liệu 9*](#_Toc143985844)

[*3.3. Biến đổi dữ liệu 9*](#_Toc143985845)

[***Chương 4: Khai phá dữ liệu bằng thuật toán K-means 11***](#_Toc143985846)

[*4.1 Giới thiệu về kỹ thuật phân cụm 11*](#_Toc143985847)

[*4.2. Thuật toán K-means 12*](#_Toc143985848)

[*4.3. Thực hiện Phân cụm dữ liệu bệnh nhân bằng thuật toán K-means 13*](#_Toc143985849)

[*4.3.1. Xác định số tâm cụm tối ưu: 13*](#_Toc143985850)

[*4.3.2. Sử dụng MongoDB để lưu trữ liệu 14*](#_Toc143985851)

[*4.3.3. Cài đặt thuật toán Kmeans bằng Python 15*](#_Toc143985852)

[*4.3.4. Kết quả 18*](#_Toc143985853)

[*4.3.5. Demo ứng dụng dự doán nhóm sức khỏe của bệnh nhân 20*](#_Toc143985854)

[***Chương 5: Kết luận 22***](#_Toc143985855)

[***Tài liệu tham khảo 23***](#_Toc143985856)

# Bảng phân chia công việc

|  |  |
| --- | --- |
| **Họ và tên** | **Công việc thực hiện** |
| Hoàng Anh Tuấn 20510634842 | * Cài đặt thuật toán K-means * Kết nối đến MongoDB lấy dữ liệu thực hiện phân cụm * Làm báo cáo |
| Nguyễn Phú Trường 2051063909 | * Xây dựng ứng dụng bằng Node.js * Xác định số cụm tối ưu * Tiền xử lý dữ liệu |
| Vũ Văn Thủy 2051063873 | * Cài đặt thuật toán K-means * Kết nối đến MongoDB lấy dữ liệu thực hiện phân cụm * Xây dựng ứng dụng bằng Node.js |
| Thân Thị Thùy 2051063434 | * Tìm nguồn dữ liệu * Tiền xử lý dữ liệu * Làm báo cáo |

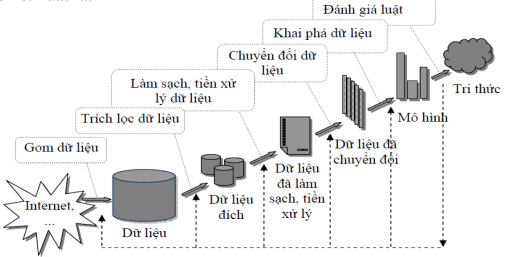
# Chương 1: Tổng quan về khai phá dữ liệu

## Phát hiện tri thức và khai phá dữ liệu

**Phát hiện tri thức (Knowledge Discovery)** trong các cơ sở dữ liệu là một quá trình nhận ra các mẫu hoặc mô hình ẩn chứa trong dữ liệu, mang những đặc điểm: hợp thức, mới mẻ, hữu ích và có thể hiểu được.

**Khai phá dữ liệu (Data Mining)** được định nghĩa: “Data Mining là một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn”. Việc khai thác dữ liệu có thể áp dụng vào nhiều lĩnh vực như y tế, phân tích thị trường, xây dựng và có thể được xem là thành quả của sự phát triển tự nhiên trong lĩnh vực công nghệ thông tin.

## Quy trình khám phá tri thức trong CSDL



*Hình 1.1 Quá trình khai phá tri thức từ CSDL*

Phát hiện tri thức trong CSDL gồm những bước quan trọng sau:

* **Làm sạch dữ liệu**: Loại bỏ các dữ liệu nhiễu và dữ liệu không nhất quán.
* **Tích hợp dữ liệu**: Tổng hợp dữ liệu từ nhiều nguồn khác nhau để tạo ra một tập dữ liệu toàn diện hơn.
* **Lựa chọn dữ liệu**: Trích xuất các dữ liệu phù hợp với mục tiêu phân tích từ CSDL.
* **Chuyển đổi dữ liệu**: Chuyển đổi và tạo một cấu trúc dữ liệu thích hợp để dễ dàng tiến hành khai phá.
* **Khai phá dữ liệu**: Bước quan trọng, sử dụng các kỹ thuật thông minh để tìm ra những mẫu và thông tin quan trọng từ dữ liệu.
* **Đánh giá mẫu**: Các chuyên gia dữ liệu sẽ áp dụng các chỉ số và phương pháp để định giá giá trị thực sự và tầm quan trọng của các mẫu tri thức.
* **Biểu diễn tri thức**: Ở giai đoạn này, các kỹ thuật biểu diễn và hiển thị dữ liệu được sử dụng để trình bày tri thức đã khám phá cho người dùng.

## Một số kỹ thuật khai phá dữ liệu

* **Kỹ thuật khai phá luật kết hợp (Association Rule Mining)**: Kỹ thuật này tập trung vào việc tìm hiểu mối quan hệ giữa các mục trong dữ liệu. Nó phát hiện ra các luật kết hợp giữa các mục xuất hiện cùng nhau trong tập dữ liệu. Ví dụ, từ dữ liệu bán hàng, kỹ thuật này có thể phát hiện ra rằng nếu khách hàng mua sản phẩm A thì họ cũng có xu hướng mua sản phẩm B.
* **Kỹ thuật phân lớp (Classification)**: Kỹ thuật này dự đoán và gán các mẫu dữ liệu vào một số lớp hay nhãn đã xác định trước dựa trên các đặc điểm của chúng. Ví dụ, trong phân tích email, kỹ thuật này có thể phân loại email là "spam" hoặc "không phải spam" dựa trên nội dung và thuộc tính khác.
* **Kỹ thuật phân cụm (Clustering)**: Kỹ thuật này nhóm các mẫu dữ liệu vào các cụm dựa trên sự tương đồng hoặc thuộc tính chung. Nó giúp phát hiện cấu trúc ẩn trong dữ liệu mà không cần biết trước các nhãn. Ví dụ, trong phân tích dữ liệu khách hàng, kỹ thuật này có thể nhóm khách hàng có sở thích tương tự vào các nhóm.

Những kỹ thuật này cùng với nhiều kỹ thuật khác trong khai phá dữ liệu giúp chúng ta hiểu rõ hơn về dữ liệu, phát hiện ra mẫu ẩn và thông tin hữu ích, và từ đó hỗ trợ trong việc ra quyết định và đưa ra dự đoán.

# Chương 2: Mô tả bài toán và phân tích dữ liệu thô

## 2.1. Mô tả bài toán

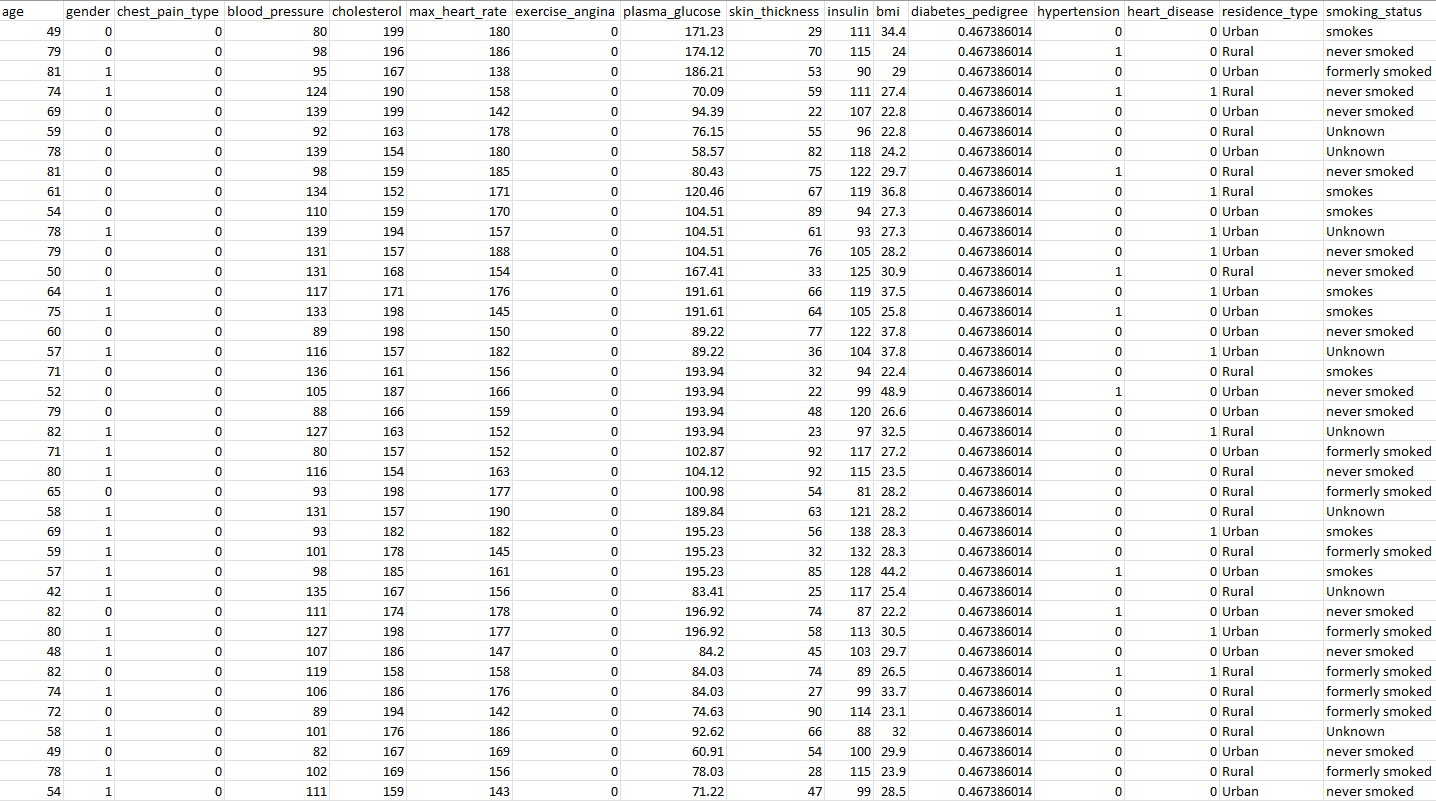
Bài toán đề xuất tập trung vào khai phá dữ liệu bệnh nhân thông qua việc áp dụng phương pháp phân cụm K-means. Bằng cách sử dụng bộ dữ liệu chứa thông tin về các chỉ số sức khỏe của bệnh nhân, mục tiêu của bài toán là thực hiện việc phân cụm để xác định bệnh nhân thuộc nhóm tình trạng sức khỏe nào dựa trên đặc trưng liên quan đến sức khỏe của họ. Từ đó xây dựng ứng dụng dự đoán nhóm sức khỏe của bệnh nhân.

* Đầu vào: Tập dữ liệu chứa các thuộc tính về chỉ số sức khỏe của bệnh nhân
* Đầu ra: Dữ liệu về các tâm cụm, thể hiện cho đặc trưng về sức khỏe của mỗi cụm

## 2.2. Phân tích dữ liệu thô

Nguồn dữ liệu thô: [Patient Dataset for Clustering | Kaggle](https://www.kaggle.com/datasets/hossamahmedaly/patient-priority-classification), bộ dữ liệu thô về bệnh nhân dùng cho việc phân cụm

* Số lượng thuộc tính: 16
* Số lượng bản ghi: 6963

*Hình 2.1 Dữ liệu ban đầu*

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩ thuộc tính** |
| 1 | Age | Tuổi |
| 2 | Gender | Giới tính |
| 3 | chest\_pain\_type | Mức độ đau ngực |
| 4 | blood\_pressure | Huyết áp |
| 5 | Cholesterol | Cholesterol |
| 6 | max\_heart\_rate | Nhịp tim tối đa |
| 7 | exercise\_angina | Đau ngực do tập thể dục |
| 8 | plasma\_glucose | Glucose huyết tương |
| 9 | skin\_thickness | Độ dày da |
| 10 | Insulin | Insulin |
| 11 | Bmi | Chỉ số BMI |
| 12 | diabetes\_pedigree | Khả năng mắc bệnh tiểu đường tùy thuộc vào độ tuổi hoặc tiền sử gia đình mắc bệnh tiểu đường |
| 13 | Hypertension | Tăng huyết áp |
| 14 | heart\_disease | Bệnh tim |
| 15 | residence\_type | Nơi cư trú |
| 16 | smoking\_status | Tình trạng hút thuốc |

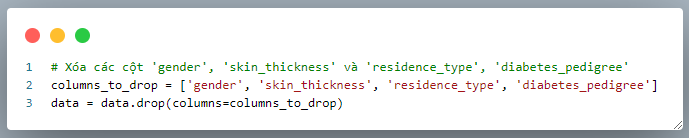
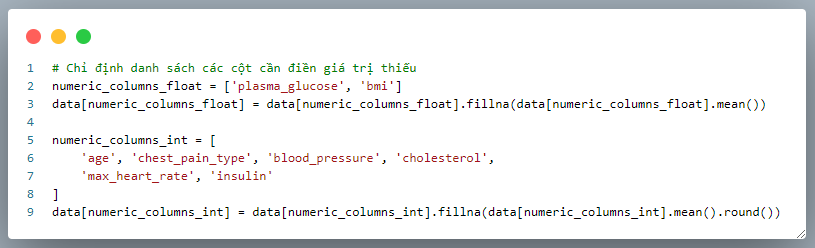
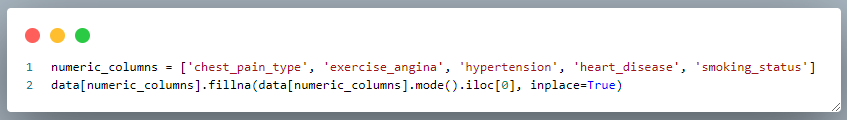
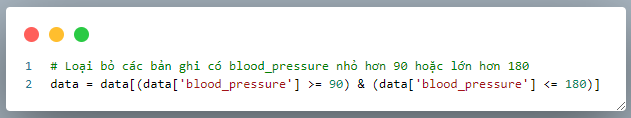
*Bảng 2.1 Liệt kê các thuộc tính của bộ dữ liệu*

# Chương 3: Tiền xử lý dữ liệu

## 3.1. Làm sạch dữ liệu

Làm sạch dữ liệu là quá trình nhận dạng dữ liệu đã có để tiến hành xử lý các dữ liệu bị thiết (missing data), xử lý dữ liệu bị nhiễu (noisy data) và dữ liệu không nhất quán (inconsistent data).

Đọc file dữ liệu và thực hiện làm sạch dữ liệu bằng Python:

* Xóa những thuộc tính không cần thiết: những thuộc tính không cần thiết như ‘gender’, ‘skin\_thickness’, ‘residence\_type’, ‘diabetes\_pedigree’ sẽ được xóa bỏ 
* Xử lý dữ liệu bị thiếu: các dữ liệu bị thiếu thay thế bằng giá trị trung bình của thuộc tính đó sử dụng **mean()**
* Xử lý dữ liệu bị thiếu: thay thế bằng giá trị xuất hiện thường xuyên nhất của thuộc tính đó, sử dụng **mode()**
* Xử lý dữ liệu bị nhiễu: thuộc tính ‘blood\_pressure’ – huyết áp có các giá trị bất thường, không hợp lệ 

## 3.2. Tích hợp dữ liệu

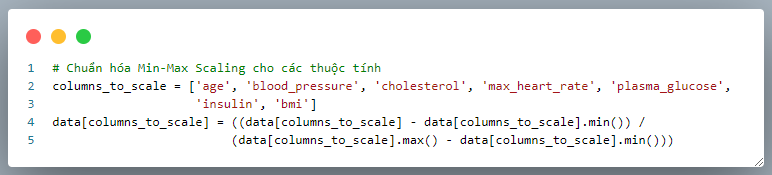
Tích hợp dữ liệu là quá trình kết hợp dữ liệu từ các nguồn khác nhau để tạo thành một tập dữ liệu hoàn chỉnh hơn hoặc để tạo ra các kết hợp mới của dữ liệu.

* Dữ liệu lấy từ một nguồn nên không thực hiện quá trình này

## 3.3. Biến đổi dữ liệu

Biến đổi dữ liệu là quá trình thay đổi cấu trúc hoặc nội dung của dữ liệu để tạo ra các biến mới hoặc thay đổi dạng của dữ liệu hiện có. Mục tiêu của bước này là cải thiện tính phù hợp của dữ liệu với mục tiêu phân tích hoặc mô hình hóa.

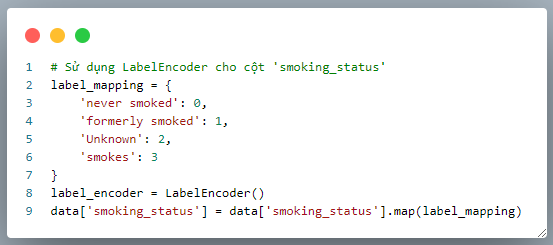
Đọc file dữ liệu và thực hiện biến đổi dữ liệu bằng Python:

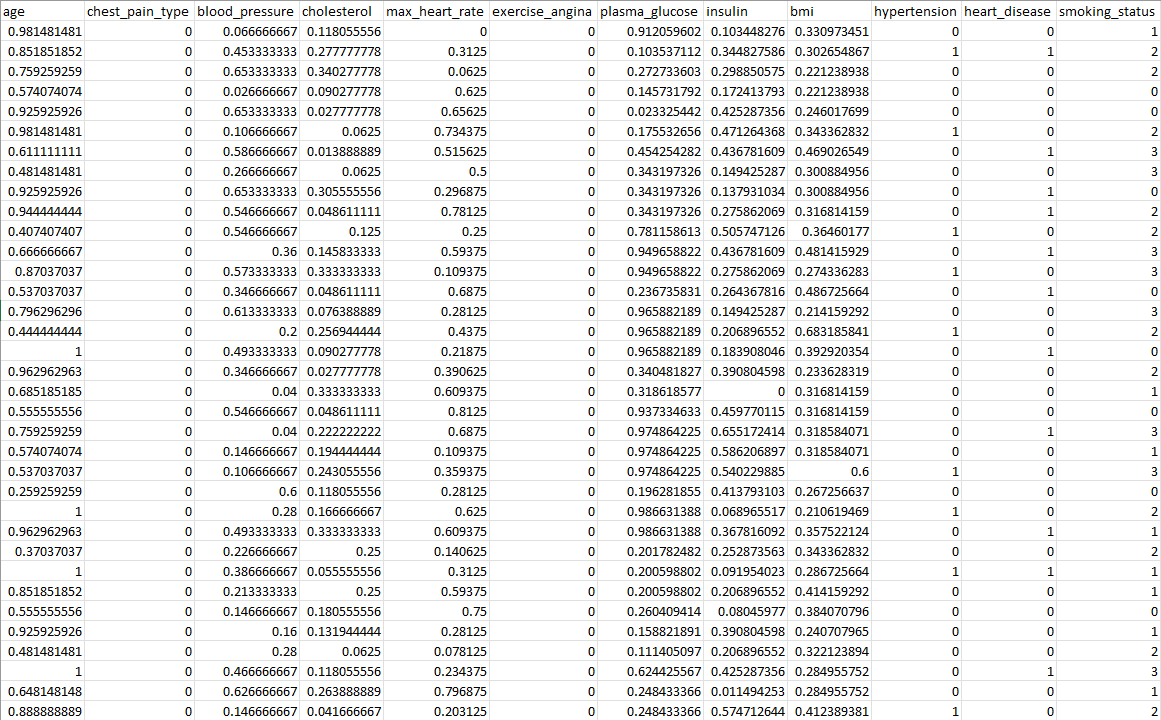
* Biến đổi dữ liệu bằng phương pháp Min Max Scaling: đưa các giá trị dữ liệu về khoảng giá trị cụ thể, thường là trong khoảng từ 0 đến 1. 
* Công thức:

Trong đó:

* v = [minA, maxA] là giá trị cũ
* v’ = [0,1] là giá trị mới

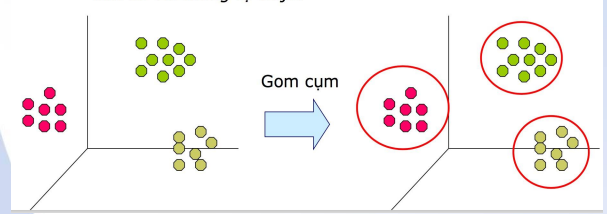
Ví dụ:

* Thuộc tính ‘blood\_pressure’: có v = [90, 180]
* Chuẩn hóa về giá trị: v’ = [0, 1]
* Với v = 105
  + - v’ = (105 - 90) / (180 - 90) = 0.1667
* Biến đổi dữ liệu bằng Label Encoder: biến các giá trị cố định ở dạng ký tự về dạng giá trị số 

*Hình 3.1 Dữ liệu sau khi đã được chuẩn hóa*

# Chương 4: Khai phá dữ liệu bằng thuật toán K-means

## Giới thiệu về kỹ thuật phân cụm

* **Khái niệm:**
* Phân cụm thuộc loại học không giám sát (Unsupervised learning) là một dữ liệu là bài toán gom nhóm các đối tượng dữ liệu vào thánh từng cụm (cluster) sao cho các đối tượng trong cùng một cụm có sự tương đồng theo một tiêu chí nào đó. Ví dụ: phân nhóm khách hàng dựa trên hành vi mua hàng.
* Là một kỹ thuật trong data mining nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho ra quyết định
* **Các hướng tiếp cận trong phân cụm:**
  + Tìm phần tử ngoại lai (Outlier)
  + Xác định một số nhóm nhỏ các đối tượng dữ liệu “Khác thường” so với các dữ liệu trong để tránh sự ảnh hưởng của chúng tới quá trình kết quả của Phân cụm dữ liệu
  + Làm sạch dữ liệu
* **Các bước xây dựng:**
  + Xây dựng hàm tính độ tương tự
  + Xây dựng các tiêu chuẩn phân cụm
  + Xây dựng mô hình cho cấu trúc cụm dữ liệu
  + Xây dựng thuật toán phân cụm và xác lập các điều kiện khởi tạo
  + Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm
* **Kỹ thuật gom cụm:**
  + Là quá trình gom / nhóm cụm dữ liệu / đối tượng vào các lớp / cụm
  + Các đối tượng cùng 1 cụm tương tự với nhau hơn là đối tượng ở các cụm

Hình 4.1 Mô tả về kỹ thuật gom cụm

## 4.2. Thuật toán K-means

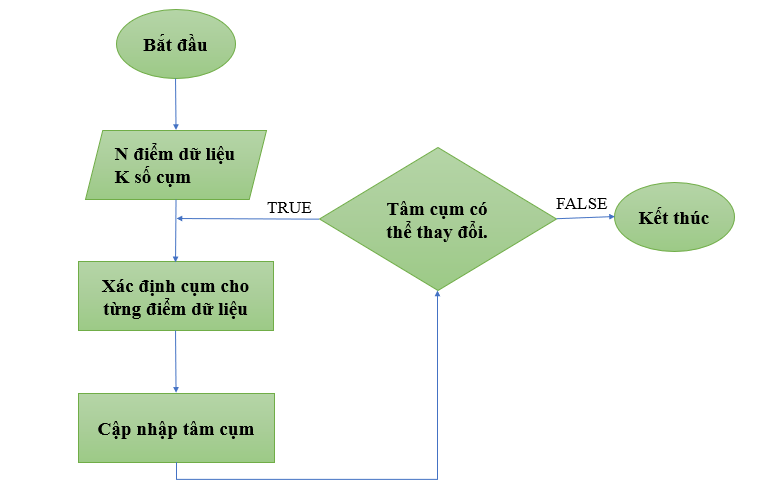
Thuật toán K-means là một trong các thuật toán phân cụm đơn giản và điển hình nhất, do MacQueen đề xuất trong lĩnh vực thống kê năm 1967. Mục đích của thuật toán là sinh ra k cụm dữ liệu từ một tập dữ liệu ban đầu gồm n đối tượng trong không gian p chiều.

**Phát biểu bài toán phân cụm:**

* Input: *n* đối tượng và số các cụm *k*
* Output: các cụm *Ci (i = 1…k)* sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu

**Các bước của thuật toán K-mean: [3]**

* Bước 1: Khởi tạo tâm cụm
  + Chọn *k* đối tượng *mj (j=1…k)* là trọng tâm ban đầu của *k* cụm từ tập dữ liệu
* Bước 2: Tính toán khoảng cách và gán cụm
  + Với mỗi điểm dữ liệu, tính toán khoảng cách từ nó tới mỗi tâm cụm đã được khởi tạo, sau đó tìm tâm cụm gần nhất đối với mỗi đối tượng
  + Công thức tính khoảng cách Euclidean giữa hai điểm A (a1, a2, ..., an) và B (b1, b2, ..., bn) trong không gian n chiều là:
* Bước 3: Cập nhật lại trọng tâm
  + Với mỗi *j=1,…,k* , cập nhật trọng tâm cụm mj bằng cách xác định trung bình cộng của các vector đối tượng dữ liệu
* Bước 4: Kiểm tra điều kiện dừng
  + Lặp lại các bước 2 và 3 cho đế khi các tâm cụm không thay đổi



Hình 4.2 Lưu đồ thuật toán K-means

## 4.3. Thực hiện Phân cụm dữ liệu bệnh nhân bằng thuật toán K-means

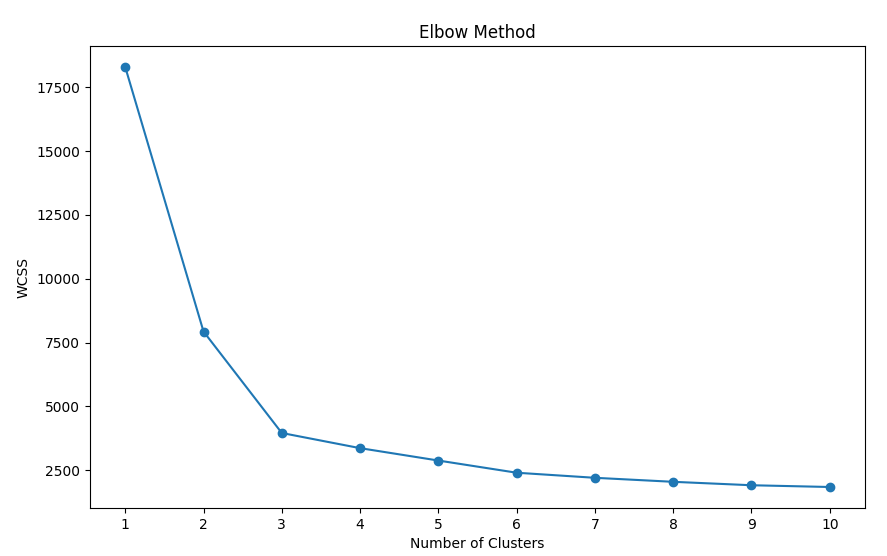
### 4.3.1. Xác định số tâm cụm tối ưu:

Để xác định số lượng cụm K, cần sử dụng một biểu đồ được tạo bằng Elbow Method (phương pháp khuỷu tay) – là một phương pháp thường được sử dụng để chọn số lượng tốt nhất của các cụm trong thuật toán phân cụm. Phương pháp này giúp xác định số lượng cụm tối ưu dựa trên sự thay đổi của tổng bình phương khoảng cách (Within-Cluster Sum of Squares - WCSS) giữa các điểm dữ liệu và trung tâm của cụm.

Cách thức thực hoạt động của Elbow Method như sau:

* Thực hiện thuật toán KMeans với một loạt số lượng cụm khác nhau (thường từ 1 đến một giới hạn tối đa đã xác định từ trước).
* Đối với mỗi số lượng cụm, tính giá trị WCSS (tổng bình phương khoảng cách giữa các điểm dữ liệu và trung tâm cụm tương ứng).
* Vẽ biểu đồ Elbow Method với trục x là số lượng cụm và trục y là giá trị WCSS. Đường cong trên biểu đồ thường có dạng "cổ điển" như cùi chỏ, giống như đoạn của một cánh tay và một bàn tay đang giơ lên.
* Xác định điểm "bẻ" của đường cong (nơi mà đường cong thay đổi hướng từ việc giảm dần sang giảm chậm hơn). Điểm này thường gọi là "đỉnh của cùi chỏ" (elbow point).
* Số lượng cụm tốt nhất sẽ là số lượng cụm tương ứng với điểm "bẻ" trên biểu đồ Elbow Method.

Ý nghĩa của Elbow Method là khi số lượng cụm tăng, WCSS thường giảm vì các điểm dữ liệu được gom vào cụm một cách gần hơn với trung tâm cụm. Tuy nhiên, tại một điểm nào đó, việc tăng số lượng cụm không còn gây ra sự giảm đáng kể trong WCSS, và đây chính là điểm mà đường cong bẻ. Điều này xảy ra vì thêm một cụm mới không còn đáng giá trong việc giảm tổng bình phương khoảng cách nữa, và sự thay đổi giữa số lượng cụm đã đạt một mức ổn định.



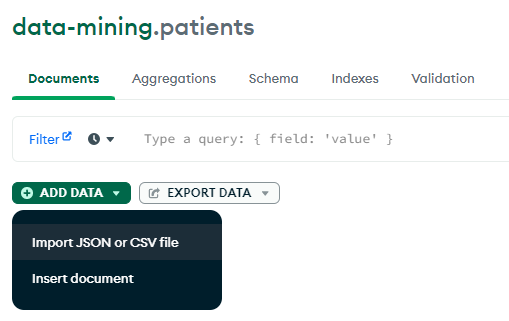
Hình 4.3 Biểu đồ Elbow

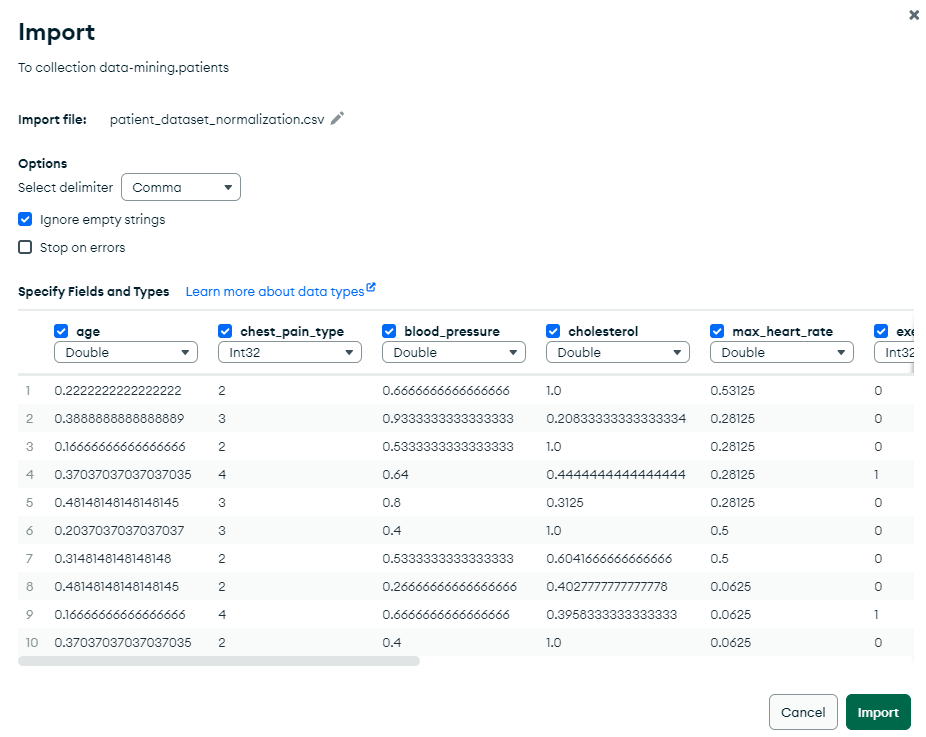
Từ biểu đồ trên, nhận thấy rằng tại số tượng cụm bằng 3 thì biểu đồ thay đổi hướng từ việc giảm dần sang giảm chậm hơn. Kết luận rằng số lượng cụm tối ưu cho dữ liệu là 3 cụm

### 4.3.2. Sử dụng MongoDB để lưu trữ liệu

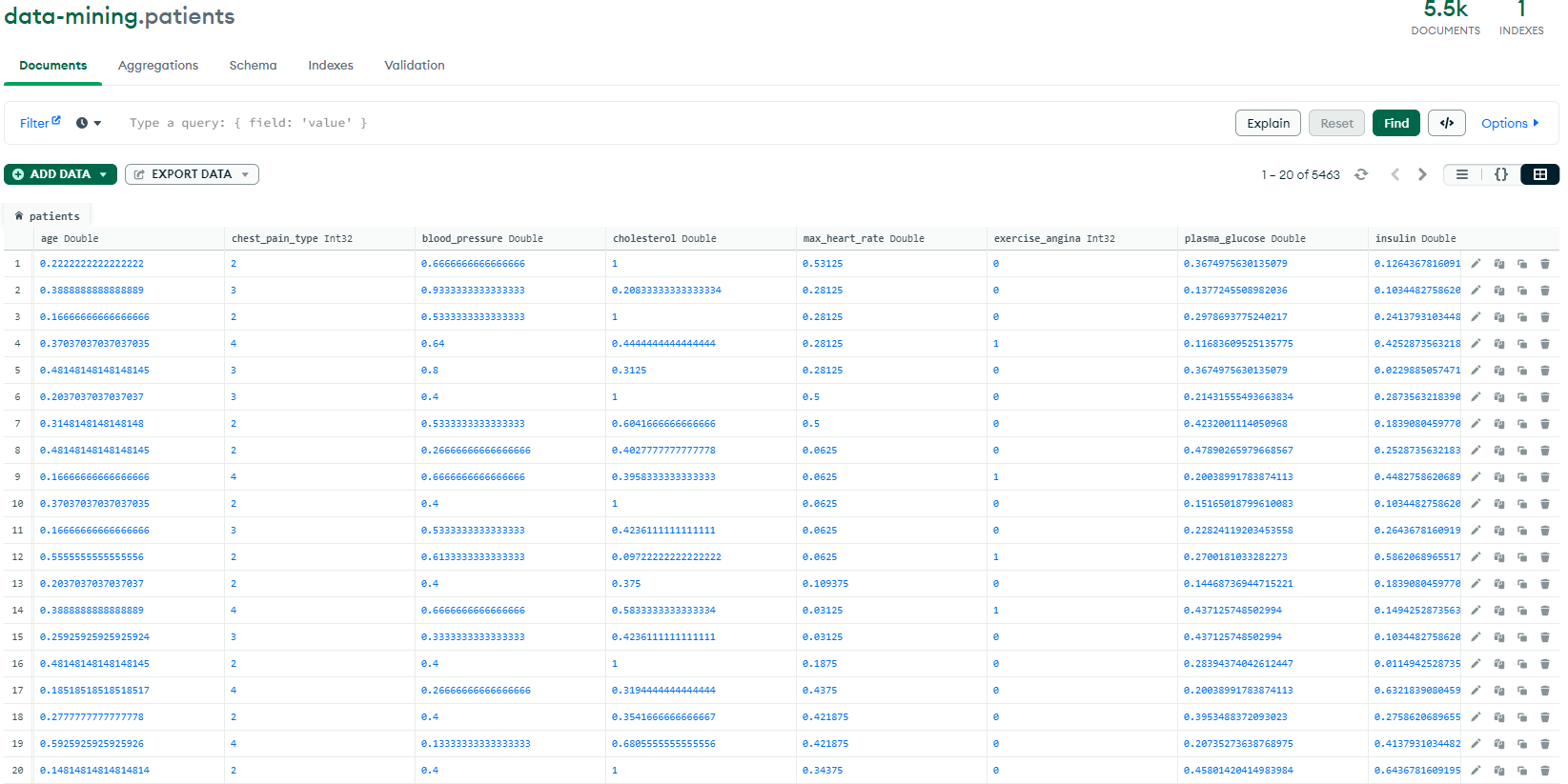
Dữ liệu sau khi đã được chuẩn hóa, thực hiện import file csv vào trong MongoDB:

* Chọn ‘Import JSON or CSV file’ trong MongoDB Compass
* Chọn file CSV cần import
* Kiểm tra lại kiểu dữ liệu của các trường dữ liệu
* Ấn ‘Import’ và kiểm tra dữ liệu đã được import đầy đủ hay chưa

  
Hình 4.4 Import file CSV vào MongoDB (1)



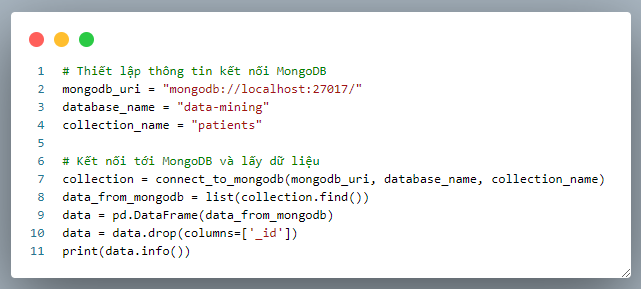
Hình 4.5 Import file CSV vào MongoDB (2)

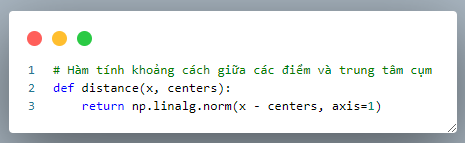


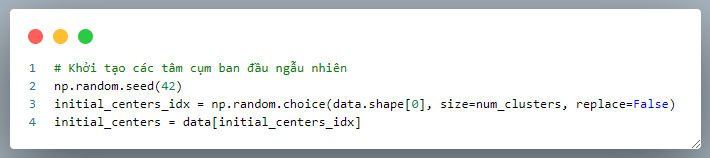
Hình 4.6 Dữ liệu đã được import thành công vào MongoDB

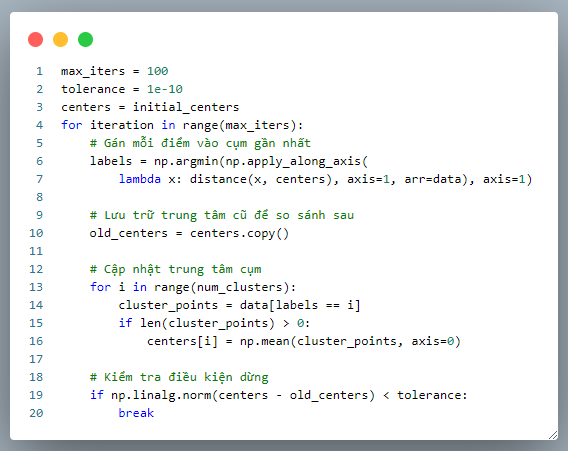
### 4.3.3. Cài đặt thuật toán Kmeans bằng Python

Các bước cài đặt thuật toán Kmeans:

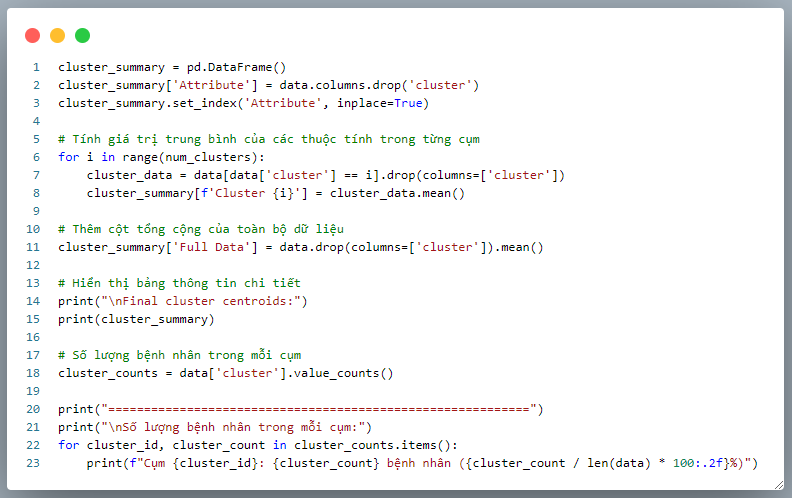
* Bước 1: Thực hiện kết nối đến MongoDB, lấy dữ liệu của các bệnh nhân để thực hiện phân cụm 
* Bước 2: Cài đặt hàm thực hiện tính toán khoảng cách từ một điểm dữ liệu đến các tâm cụm



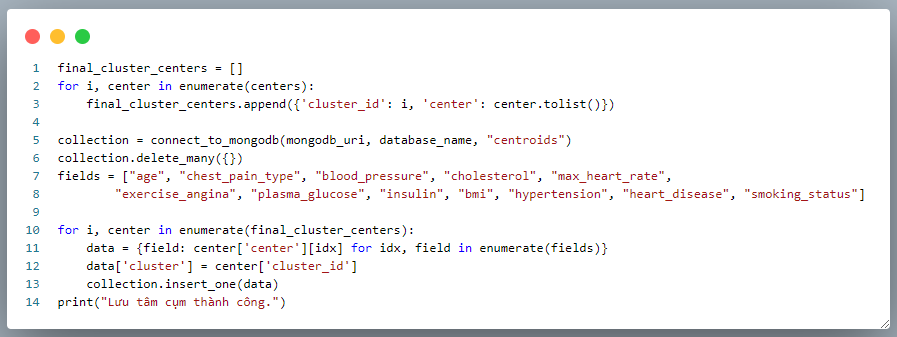
* Bước 3: Khởi tạo ngẫu nhiên 3 tâm cụm ban đầu 
* Bước 4: Vòng lặp chính của thuật toán K-means
  + Cấu hình số lần lặp tối đa, ngưỡng dừng thuật toán
  + Mỗi bước lặp lại thực hiện
    - Gán mỗi điểm dữ liệu vào tâm cụm gần nhất
    - Lưu lại tâm cụm cũ để thực hiện so sánh
    - Tính trụng bình các điểm dữ liệu của từng tâm cụm để tìm tâm cụm mới
    - So sánh tâm cụm cũ và tâm cụm mới để kiểm tra điều kiện dừng

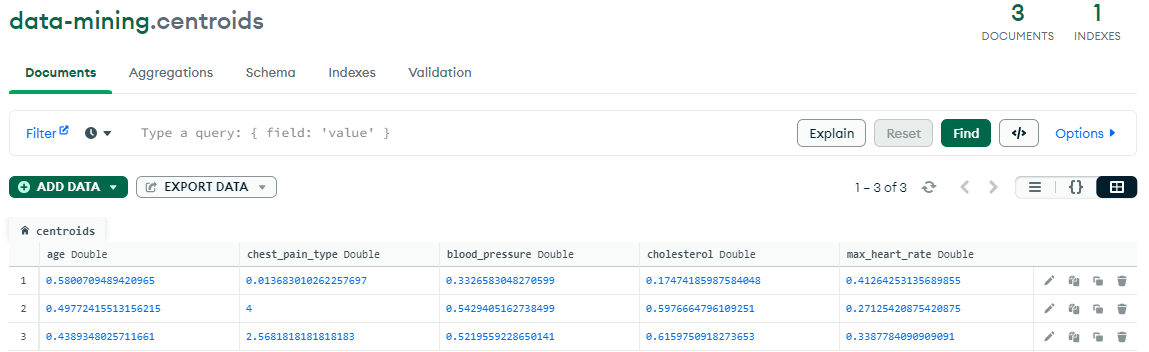


* Bước 5: In ra chi tiết các kết quả:
  + Ba tâm cụm ban đầu
  + Số lần lặp
  + Ba tâm cụm cuối cùng
  + Số lượng điểm dữ liệu trong mỗi cụm



* Bước 6: Thực hiện lưu 3 tâm cụm cuối cùng vào MongoDB, phục vụ cho việc xây dựng ứng dụng dự đoán nhóm sức khỏe của bệnh nhân

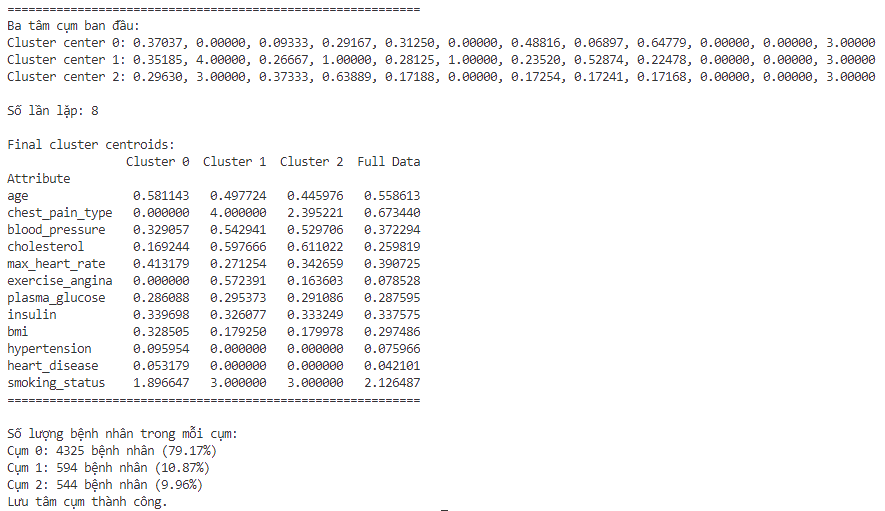




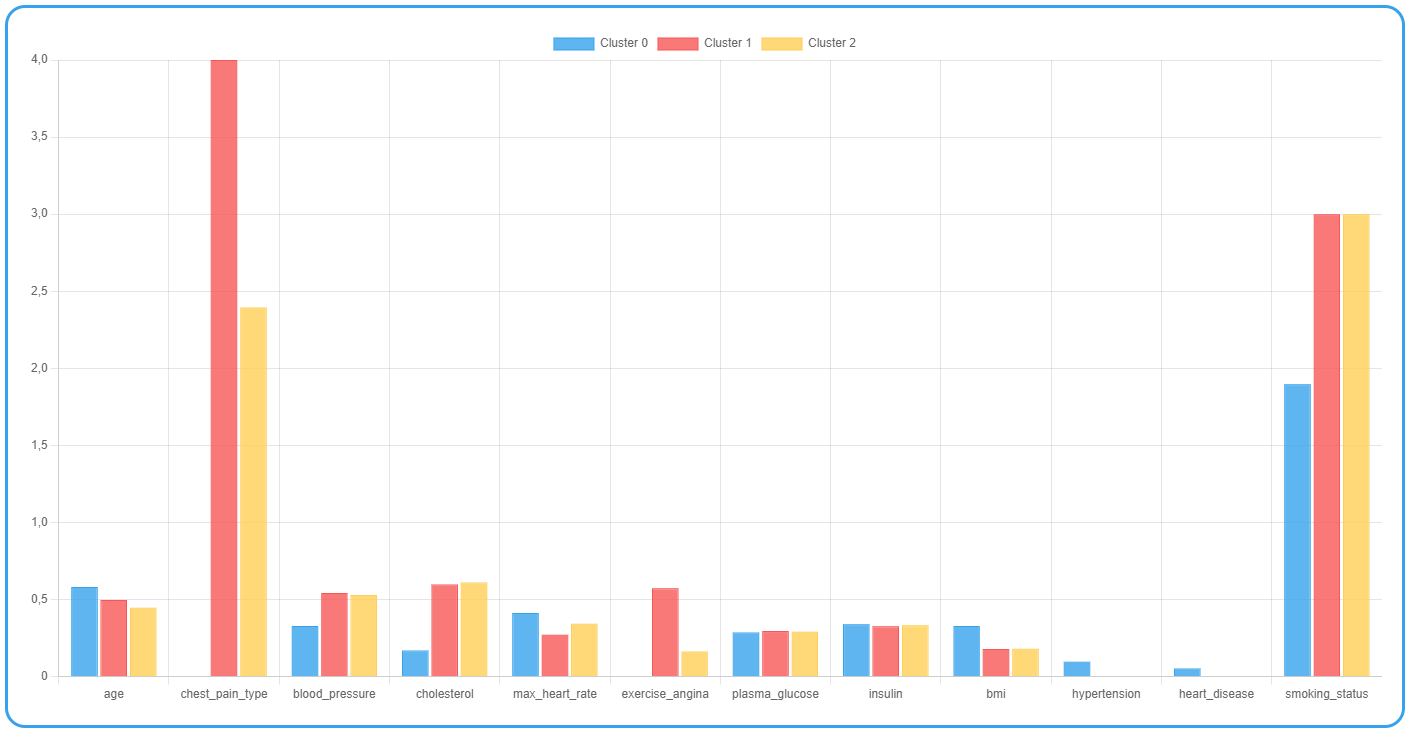
Hình 4.7 Dữ liệu 3 tâm cụm đã lưu vào MongoDB

### 4.3.4. Kết quả

Sau khi thực hiện Phân cụm K-means bằng Python, kết quả thu được như sau:



Hình 4.8 Kết quả thu được



Hình 4.9 Biểu đồ cột thể hiện các trường dữ liệu của 3 tâm cụm

Kết quả sau khi phân cụm:

* 3 tâm cụm:
  + Cluster 0: 4325 bệnh nhân, chiếm 79.17%
  + Cluster 1: 594 bệnh nhân, chiếm 10,87%
  + Cluster 2: 484 bệnh nhân, chiếm 9.96%

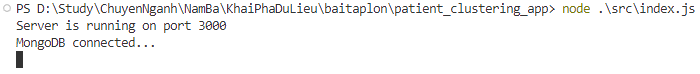
Phân tích chi tiết đặc điểm của từng tâm cụm:

* **Cluster 0:** 
  + Độ tuổi trung bình lớn hơn 2 cụm còn lại
  + Mức độ đau ngực nhẹ
  + Các chỉ số về ‘blood\_pressure’, ‘cholesterol’ và ‘max\_heart\_rate’ ở mức tốt
  + Các chỉ số ‘plasma\_glucose’, ‘insulin’, và ‘bmi’ ở mức trung bình
  + Chỉ số ‘exercise\_angina’ là 0, tức là không đau thắt ngực khi tập thể dục
  + Do cụm này có số lượng bệnh nhân cao hơn, và độ tuổi trung bình cao hơn, nên số chỉ số bệnh nhân mắc tăng huyết áp cao hơn 2 cụm còn lại
  + Đa phần bệnh nhân không hút thuốc, hoặc đã hút thuốc nhưng bỏ
* **Cụm 0 có các bệnh nhân có đặc điểm sức khỏe ổn định.**
* **Cluster 1:**
  + Độ tuổi trung bình
  + Mức độ đau ngực cao nhất
  + Các chỉ số như ‘chest\_pain\_type’, ‘blood\_pressure’, ‘cholesterol’ có giá trị cao
  + Các chỉ số ‘plasma\_glucose’, ‘insulin’, và ‘bmi’ ở mức trung bình
  + Chỉ số ‘exercise\_angina’ cao, tức là số lượng bệnh nhân đau thắt ngực khi tập thể dục là lớn
  + Đa phần bệnh nhân đều hút thuốc
* **Cụm 1 thể hiện các bệnh nhân có nguy cơ rất lớn mắc các bệnh về tim mạch**
* **Cluster 2:**
  + Độ tuổi trung bình thấp hơn 2 cụm còn lại
  + Mức độ đau ngực trung bình
  + Các chỉ số như ‘chest\_pain\_type’, ‘blood\_pressure’, ‘cholesterol’ có giá trị cao
  + Chỉ số ‘exercise\_angina’ ở mức thấp, tức là tồn tại số ít bệnh nhân đau thắt ngực khi tập thể dục
  + Các chỉ số ‘plasma\_glucose’, ‘insulin’, và ‘bmi’ ở mức trung bình
  + Đa phần các bệnh nhân đều hút thuốc
* **Cụm 2 thể hiện các bệnh nhân có nguy cơ mắc các bệnh về tim mạch**

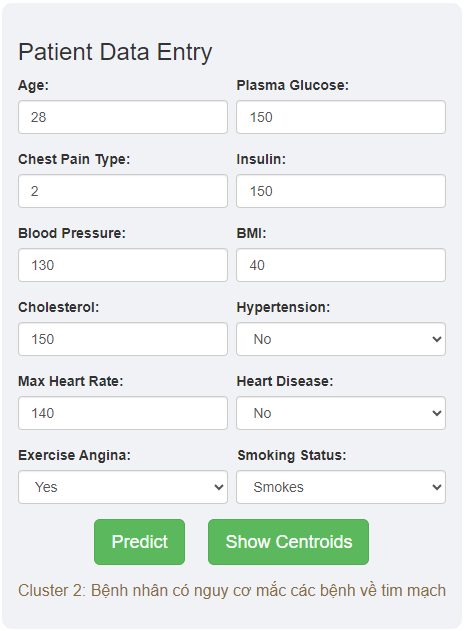
### 4.3.5. Demo ứng dụng dự doán nhóm sức khỏe của bệnh nhân

Từ những phân tích đặc điểm dữ liệu của các cụm trên, xây dựng ứng dụng dự đoán nhóm sức khỏe của bệnh nhân bằng Node.js:

* Khởi chạy ứng dụng Node.js
* Truy cập <http://localhost:3000/>
* Nhập dữ liệu của một bệnh nhân và ấn nút ‘Predict’
* Dữ liệu sau đó sẽ được chuẩn hóa theo đúng dạng đã sử dụng để phân cụm
* Sau đó thực hiện truy vấn đến 3 tâm cụm trong MongoDB rồi thực hiện tính toán khoảng cách từ điểm dữ liệu nhập vào tới 3 tâm cụm
* Tìm ra tâm cụm gần nhất và hiển thị đặc trưng của tâm cụm đó lên màn hình



Hình 4.10 Khởi chạy ứng dụng Node.js



Hình 4.11 Ứng dụng dự đoán nhóm sức khỏe của bệnh nhân

# Chương 5: Kết luận

Trong quá trình tìm hiểu và thực hiện bài tập lớn với mục tiêu khai phá dữ liệu bệnh nhân bằng phương pháp phân cụm K-means, chúng em đã thu được những kết quả quan trọng sau:

* Tìm hiểu về K-means và phân cụm dữ liệu: Chúng em đã nắm vững kiến thức về thuật toán K-means và quy trình phân cụm dữ liệu. Điều này bao gồm việc lựa chọn số lượng cụm, chuẩn bị dữ liệu và cách hoạt động của thuật toán K-means.
* Tiền xử lý dữ liệu: Chúng em đã thực hiện tiền xử lý dữ liệu bệnh nhân để chuẩn bị cho việc sử dụng thuật toán K-means. Việc này bao gồm xử lý các giá trị thiếu, chuẩn hóa dữ liệu và chọn thuộc tính phù hợp.
* Xây dựng mô hình K-means: Chúng em đã triển khai thuật toán K-means trên dữ liệu bệnh nhân và tạo ra các cụm dựa trên đặc điểm của họ. Qua đó, chúng em đã tạo ra một cấu trúc phân cụm dữ liệu giúp hiểu rõ hơn về sự đa dạng trong dữ liệu bệnh nhân.
* Phân tích kết quả: Chúng em đã phân tích kết quả phân cụm để hiểu rõ hơn về những đặc điểm chung và khác biệt giữa các cụm. Việc này giúp chúng em xác định các đặc trưng quan trọng có thể ảnh hưởng đến sức khỏe của các bệnh nhân.

Tuy nhiên, trong quá trình thực hiện, bài tập nhóm vẫn còn một số hạn chế nhất định:

* Lựa chọn số lượng cụm: Việc xác định số lượng cụm là một thách thức và có thể ảnh hưởng đến kết quả phân cụm. Chúng em đã cố gắng sử dụng các phương pháp như Elbow Method để giúp xác định số lượng cụm tốt nhất.
* Dữ liệu không đầy đủ: Dữ liệu bệnh nhân có thể không đầy đủ và thiếu một số thông tin quan trọng. Điều này có thể làm giảm tính chính xác của kết quả phân cụm.
* Tính khách quan: Việc tạo lựa chọn ban đầu về số lượng cụm có phần tính khách quan và phải dựa vào kiến thức chuyên gia hoặc các phương pháp thống kê.

Tổng kết, việc sử dụng thuật toán K-means trong việc phân cụm dữ liệu bệnh nhân đã mang lại những kết quả quan trọng và cung cấp cái nhìn mới về sự phân phối và đa dạng của dữ liệu. Tuy nhiên, việc cải tiến và phát triển thêm các phương pháp phân cụm khác và áp dụng thực tế sẽ giúp nâng cao tính ứng dụng của kết quả trong lĩnh vực chăm sóc sức khỏe.

# Tài liệu tham khảo

*[1] Giáo trình đại học thuỷ lợi môn Khai phá dữ liệu của TS. Trần Mạnh Tuấn*

*[2] Tìm hiểu về Khai phá dữ liệu trên* [*Wikipedia*](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u)

*[3] Tìm hiểu về thuật toán Kmeans* [*Wikipedia*](https://en.wikipedia.org/wiki/K-means_clustering)

*[4] Tìm hiểu thêm về thuật toán Kmeans qua* [*Machine Learning cơ bản*](https://machinelearningcoban.com/2017/01/01/kmeans/)

*[5] Nguồn dữ liệu bệnh nhân trên Kaggle* [*Patient Dataset for Clustering | Kaggle*](https://www.kaggle.com/datasets/hossamahmedaly/patient-priority-classification)