**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**----- 🙢🕮🙠 -----**

****

**ĐỀ TÀI: Sử dụng thuật toán Kmean để phân cụm dữ liệu bệnh tiểu đường**

**Nhóm sinh viên thực hiện:** **Nhóm 5**

**Nguyễn Thị Thanh Mai: 2051063842**

**Đỗ Ngọc Lan : 2051063752**

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

**Hà nội ngày 30 tháng 12 năm 2023**

MỤC LỤC

[**CHƯƠNG 1 Tổng quan về khai phá dữ liệu 5**](#_heading=h.2et92p0)

[1.1 Khai phá tri thức và khai phá dữ liệu 5](#_heading=h.tyjcwt)

[1.2 Quá trình khai phá tri thức 5](#_heading=h.3dy6vkm)

[1.3 Mô tả bài toán 5](#_heading=h.4d34og8)

[1.3.1 Tổng quan về bài toán 5](#_heading=h.2s8eyo1)

[**CHƯƠNG 2 Quy trình khai phá dữ liệu 7**](#_heading=h.3rdcrjn)

[2.1 Tiền xử lý dữ liệu 7](#_heading=h.3ziwwcbotrfy)

[2.2 Làm sạch dữ liệu 7](#_heading=h.26in1rg)

[2.1.1 Loại bỏ thuộc tính dư thừa 7](#_heading=h.akhoskwaovql)

[2.3 Tính hợp dữ liệu 11](#_heading=h.lnxbz9)

[2.4 Biến đổi dữ liệu 11](#_heading=h.35nkun2)

[**CHƯƠNG 3 Khai phá dữ liệu bằng mô hình phân cụm 11**](#_heading=h.1ksv4uv)

[3.1 Giới thiệu về phân cụm 11](#_heading=h.44sinio)

[3.1.1 Khái niệm 11](#_heading=h.2jxsxqh)

[3.1.2 Các phương pháp phân cụm 12](#_heading=h.1y810tw)

[3.2 Phân cụm bằng thuật toán K-Means 12](#_heading=h.2xcytpi)

[3.2.1 Các bước thực hiện: 12](#_heading=h.1ci93xb)

[3.2.2 Điều kiện dừng: 13](#_heading=h.3whwml4)

[3.2.3 Đánh giá mô hình 13](#_heading=h.2bn6wsx)

[3.3 Chạy phân cụm với Weka 13](#_heading=h.3as4poj)

[3.4 Chạy phân cụm với python 13](#_heading=h.2p2csry)

[3.5 Ý nghĩa của thuật toán K-Means 13](#_heading=h.32hioqz)

[**CHƯƠNG 4 Kết luận 14**](#_heading=h.3ygebqi)

[**TÀI LIỆU THAM KHẢO 14**](#_heading=h.wywaqj3z9up7)

**MỞ ĐẦU**

Trong bối cảnh của cuộc cách mạng số hóa và phát triển không ngừng của công nghệ thông tin, dữ liệu đã trở thành một nguồn tài nguyên vô cùng quý giá, mở ra những cánh cửa mới trong việc nghiên cứu và ứng dụng trong nhiều lĩnh vực, đặc biệt là lĩnh vực y tế. Để đối mặt với sự phức tạp này, nghiên cứu hiện đại ngày nay đang tập trung vào việc áp dụng các phương pháp phân loại và phân cụm dữ liệu để tìm ra những đặc điểm chung và khác biệt giữa các nhóm bệnh nhân. Tuy các phương pháp quản trị và khai thác dữ liệu truyền thống ngày càng không còn đáp ứng được nhu cầu thực tế, nhưng may mắn là xu hướng mới về việc phát hiện tri thức và khai phá dữ liệu nhanh chóng đã phát triển mạnh mẽ. Điều này tạo ra sự đa dạng lớn trong cách mà bệnh nhân tiếp cận và phản ứng với điều trị.

Chính trong bối cảnh này, việc áp dụng các phương pháp khai phá dữ liệu để xây dựng các mô hình dự đoán bệnh trở nên vô cùng quan trọng. Qua bài tập lớn này, chúng em đã tìm hiểu và trình bày về một kỹ thuật trong lĩnh vực khai phá dữ liệu, với đề tài “Khai phá dữ liệu bằng thuật toán Kmean để phân cụm dữ liệu bệnh tiểu đường”.

Trong nghiên cứu này, chúng em chọn sử dụng thuật toán KMeans, một trong những phương pháp phân cụm hàng đầu trong lĩnh vực học máy, để phân loại dữ liệu về bệnh tiểu đường. Mục tiêu của chúng em không chỉ là phát hiện ra những nhóm có đặc điểm chung mà còn là tìm ra sự đa dạng và sự khác biệt giữa các nhóm này. Qua đó, hy vọng rằng kết quả của nghiên cứu sẽ đóng góp vào việc cá nhân hóa quá trình điều trị và quản lý bệnh tiểu đường, tạo điều kiện thuận lợi cho việc cải thiện chất lượng cuộc sống của bệnh nhân.

**BẢNG PHÂN CHIA CÔNG VIỆC**

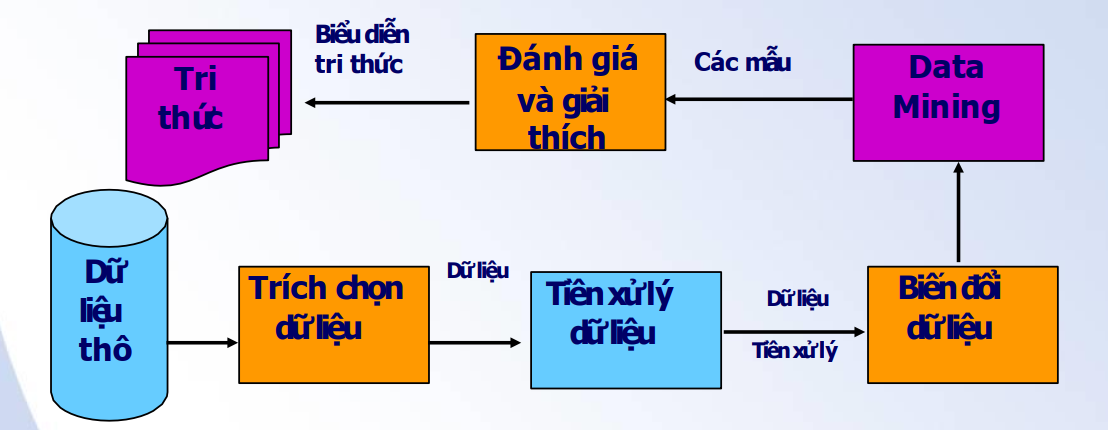
| **Tên thành viên** | **Công việc** |
| --- | --- |
| Nguyễn Thị Thanh Mai | Tìm hiểu thuật toán, code thuật toán, viết báo cáo. |
| Đỗ Ngọc Lan | Tìm hiểu thuật toán, code thuật toán, viết báo cáo. |

# Tổng quan về khai phá dữ liệu

## Khai phá tri thức và khai phá dữ liệu

* Khai phá dữ liệu là một lĩnh vực nhằm tự động khai thác những thông tin tri thức đang tiềm ẩn trong dữ liệu và là bước chính của khai phá dữ liệu.
* Khai phá tri thức trong cơ sở dữ liệu là một quy trình nhận biết các mẫu hoặc mô hình trong dữ liệu và là mục tiêu chính của khai phá dữ liệu.

## Quá trình khai phá tri thức

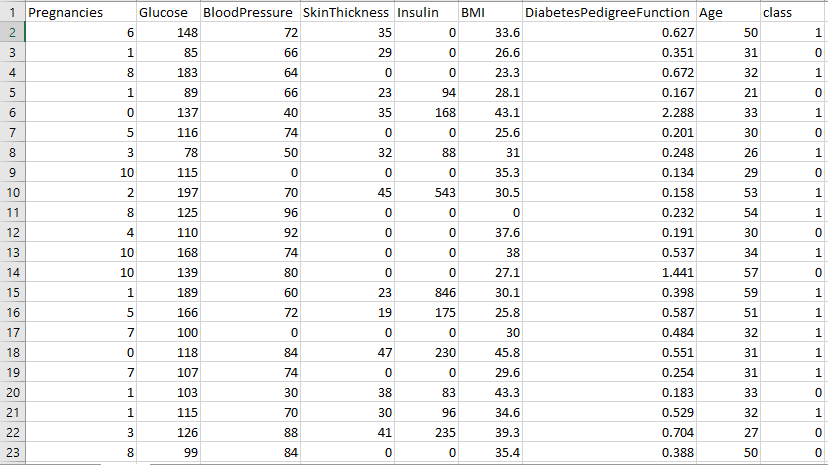


*Hình 1 Quá trình khai phá tri thức*

## Mô tả bài toán

### Tổng quan về bài toán

* Dữ liệu gồm các thuộc tính để dự đoán bệnh tiểu đường. Dữ liệu bao gồm 768 mẫu và 9 thuộc tính để dự đoán bệnh tiểu đường.
* Mục đích của bài toán: Phân cụm bệnh tiểu đường dựa vào các thuộc tính: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Class.



*Hình 2 Dữ liệu ban đầu*

Các thuộc tính là:

| **STT** | **Thuộc tính** | **Ý nghĩa thuộc tính** |
| --- | --- | --- |
| 1 | Pregnancies | số lần mang thai với miền giá trị[0, 17] |
| 2 | Glucose | đường huyết với miền giá trị [0, 199] |
| 3 | BloodPressure | Huyết áp với miền giá trị [0, 122] |
| 4 | SkinThickness | Độ dày da với miền giá trị [0, 99] |
| 5 | Insulin | Insulin với miền giá trị [0, 846] |
| 6 | BMI | Chỉ số khối cơ thể với miền giá trị [0, 67.1] |
| 7 | DiabetesPedigreeFunction | Mức độ di truyền bệnh tiểu đường với miền giá trị [0.078, 2.42] |
| 8 | Age | Tuổi với miền giá giá trị [21, 81] |
| 9 | Class | Bệnh tiểu đường:  - 0: Không bị tiểu đường  - 1: Có bị tiểu đường |

# Quy trình khai phá dữ liệu

## Tiền xử lý dữ liệu

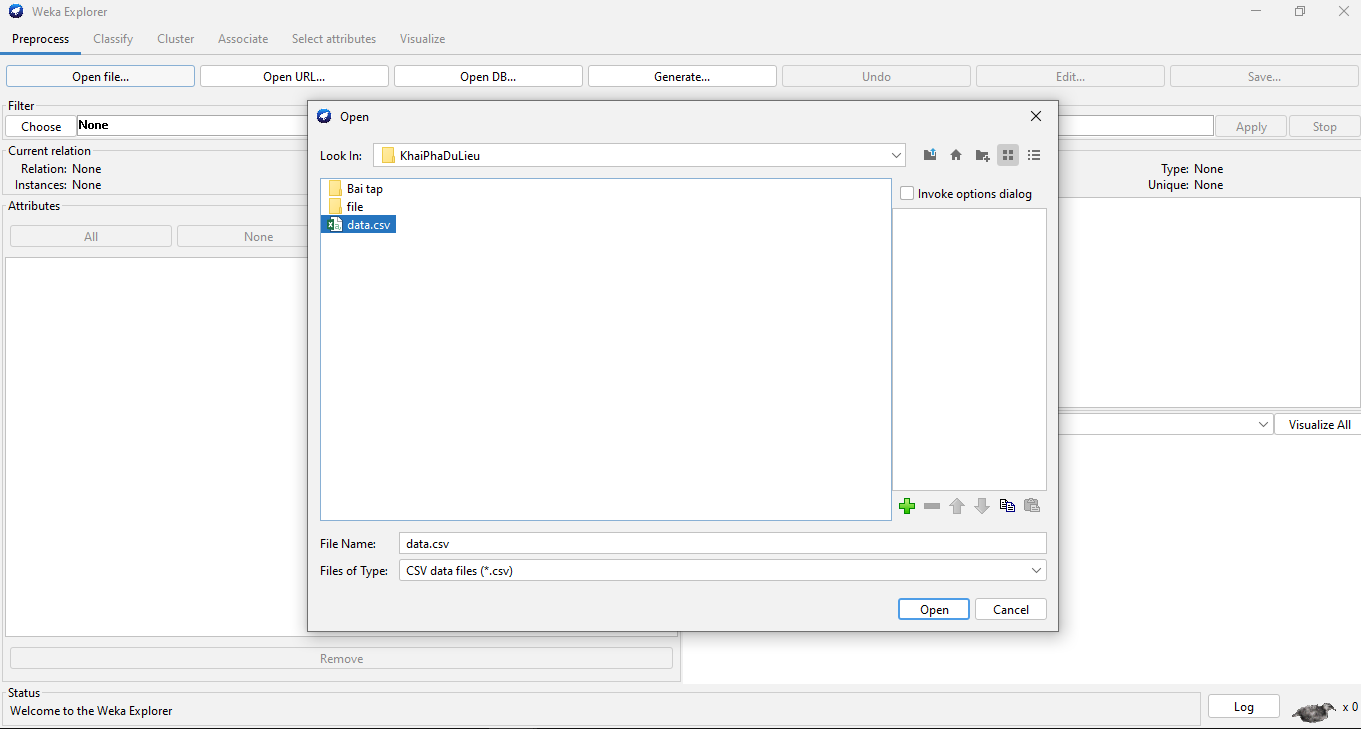
* Là quá trình xử lý dữ liệu thô/gốc nhằm cải thiện chất lượng dữ liệu và chất lượng của kết quả KPDL

## Làm sạch dữ liệu

* Làm sạch dữ liệu là tiến hành xử lý các dữ liệu bị thiếu và xử lý các dữ liệu bị nhiễu

### 2.1.1 Loại bỏ thuộc tính dư thừa

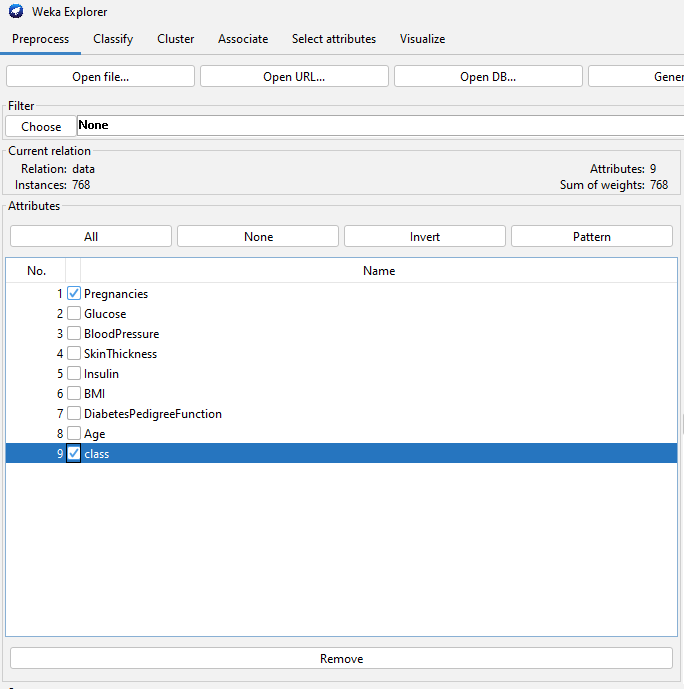
- Đọc dữ liệu vào Weka:



*Hình 3: Chọn file dữ liệu*

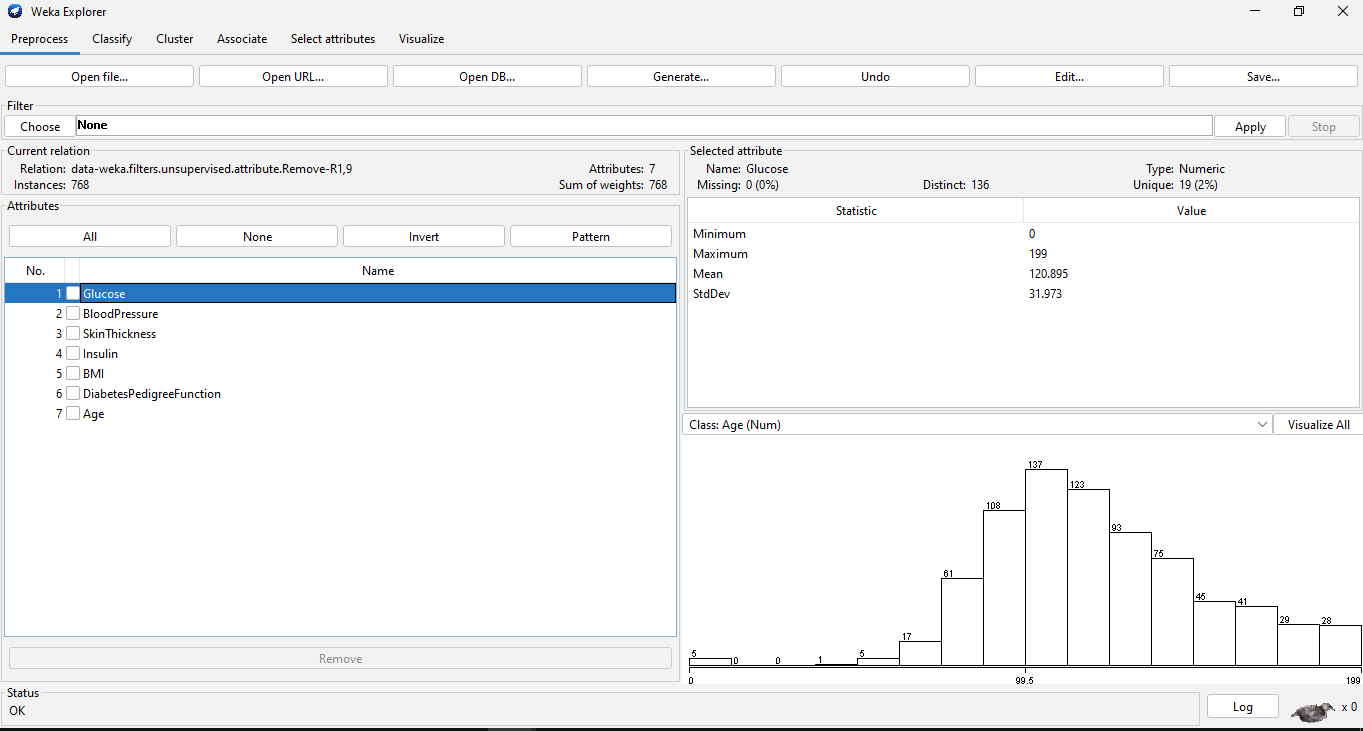
- Loại bỏ thuộc tính dư thừa, không cần thiết: **Pregnancies, Class**

+ Tick chọn thuộc tính **Pregnancies** và **Class** nhấn **Remove** để loại bỏ:



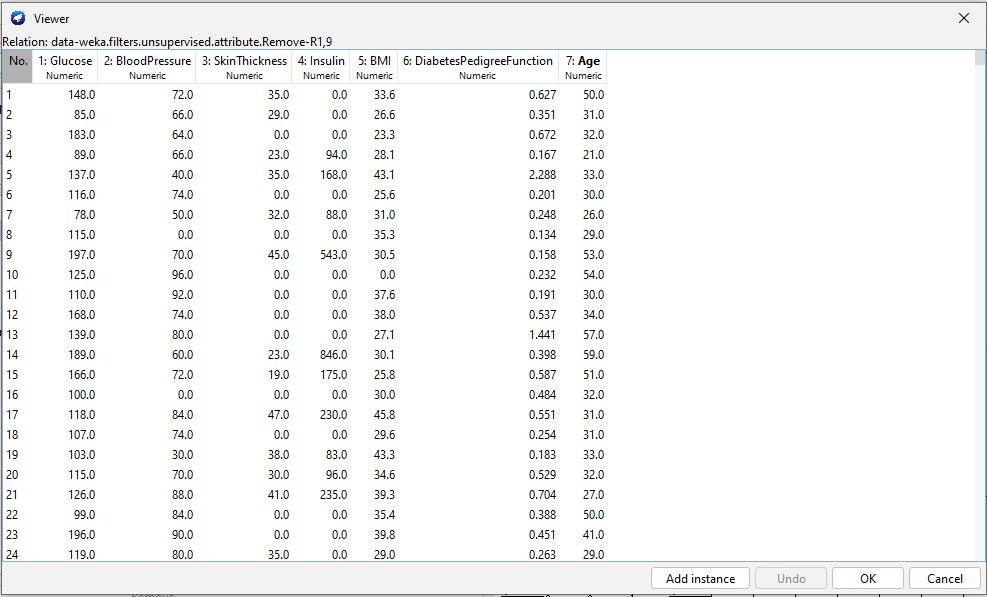
*Hình 4: Xóa bỏ thuộc tính* **Pregnancies** và **Class**

* Kết quả:



*Hình 5: Dữ liệu sau khi loại bỏ thuộc tính dư thừa*

* Dữ liệu sau khi xử lý:



*Hình 6: Dữ liệu sau khi xử lý*

## Tính hợp dữ liệu

* Tính hợp dữ liệu là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu.
* Do ở đây dữ liệu là cùng một nguồn nên bỏ qua quá trình này.

## Biến đổi dữ liệu

* Biến đổi dữ liệu là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu
* Thực hiện chuẩn hóa dữ liệu

+ Sử dụng chuẩn hóa min – max normalization với công thức

Với :

+ Ở đây sẽ đưa giá trị mới về khoảng [0, 1]

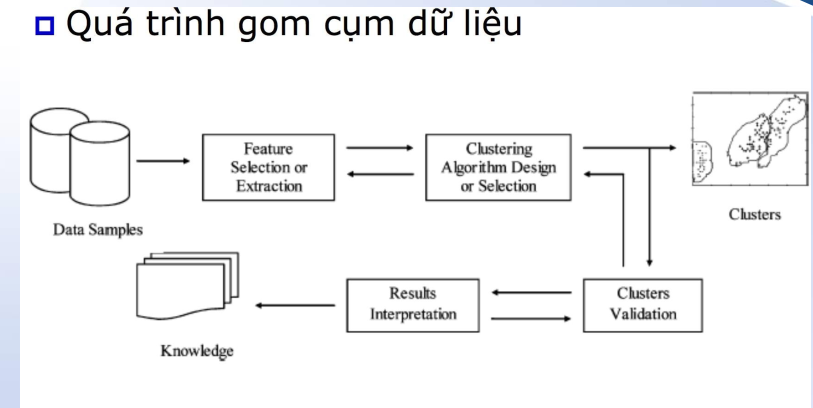
+ Tiến hành chuẩn hóa các thuộc tính về đoạn [0, 1] bằng bộ lọc Normalize, lưu file lại dưới dạng csv.

# Khai phá dữ liệu bằng mô hình phân cụm

## Giới thiệu về phân cụm

### Khái niệm

* PCDL là quá trình phân chia một tập DL ban đầu thành các cụm DL sao cho:
* Các phần tử trong một cụm "tương tự" (Similar) nhau.
* Các phần tử trong các cụm khác nhau sẽ "phi tương tự" (Dissimilar) nhau.
* Số các cụm được xác định trước theo kinh nghiệm hoặc tự động
* Quá trình gom cụm dữ liệu



*Hình 3.1 Quá trình gom cụm dữ liệu*

### Các phương pháp phân cụm

* 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
* Các kỹ thuật đều hướng tới:
* Chất lượng của các cụm
* Tốc độ thực hiện của thuật toán

## Phân cụm bằng thuật toán K-Means

* Thuật toán K-Mean là thuật toán học không giám sát.
* Input: tập dữ liệu không có nhãn và số cụm cần tìm k.
* Output: Các cụm Ci(i=1…k) sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu.

### Các bước thực hiện:

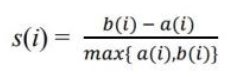
* Bước 1: Khởi tạo tâm cụm:
  + chọn k đối tượng làm trọng tâm ban đầu của k cụm (chọn ngẫu nhiên, theo kinh nghiệm).
* Bước 2: Tính toán khoảng cách và gán cụm:
  + với mỗi điểm tính khoảng cách đến tâm cụm sau đó tìm trọng tâm gần nhất với mỗi điểm (tức là khoảng cách của mỗi điểm tới các tâm cụm, khoảng cách nào ngắn nhất thì ta gán nó vào cụm đó).
* Bước 3: Tính toán và điều chỉnh tâm các cụm:
  + với mỗi cụm cập nhật trọng tâm bằng cách xác định trung bình cộng các vector đối tượng dữ liệu (tức là tính trung bình cộng tọa độ của các điểm thuộc cụm).
* Bước 4: Kiểm tra điều kiện dừng:
  + lặp lại các bước 2 và 3 cho đến khi tâm cụm không thay đổi.

### Điều kiện dừng:

* Không có(hoặc có không đáng kể) việc gán lại các ví dụ vào các cụm khác.
* Không có(hoặc có không đáng kể) việc gán lại các điểm trung tâm(centroiids) của các cụm.
* Không giảm hoặc giảm không đáng kể về tổng lỗi phân cụm E.

### Đánh giá mô hình

* Độ bóng (silhouette) có giá trị [-1,1]: giá trị càng tiến đến 1 thì mô hình càng tốt.
* Giả có k cụm, t có công thức:
* a(i): khoảng cách trung bình từ điểm i tới tất cả các điểm dữ liệu có chung cụm với i.
* b(i): khoảng cách trung bình ngắn nhất từ i tới bất kì cụm nào không chứa i. Cụm tương ứng với b(i) này được gọi là cụm hàng xóm của i.



*Hình 3.2.3 Công thức silhouette\_score*

## Chạy phân cụm với Weka

Bước 1: Cho Weka đọc dữ liệu phân cụm đã chuẩn hoá ở trên.

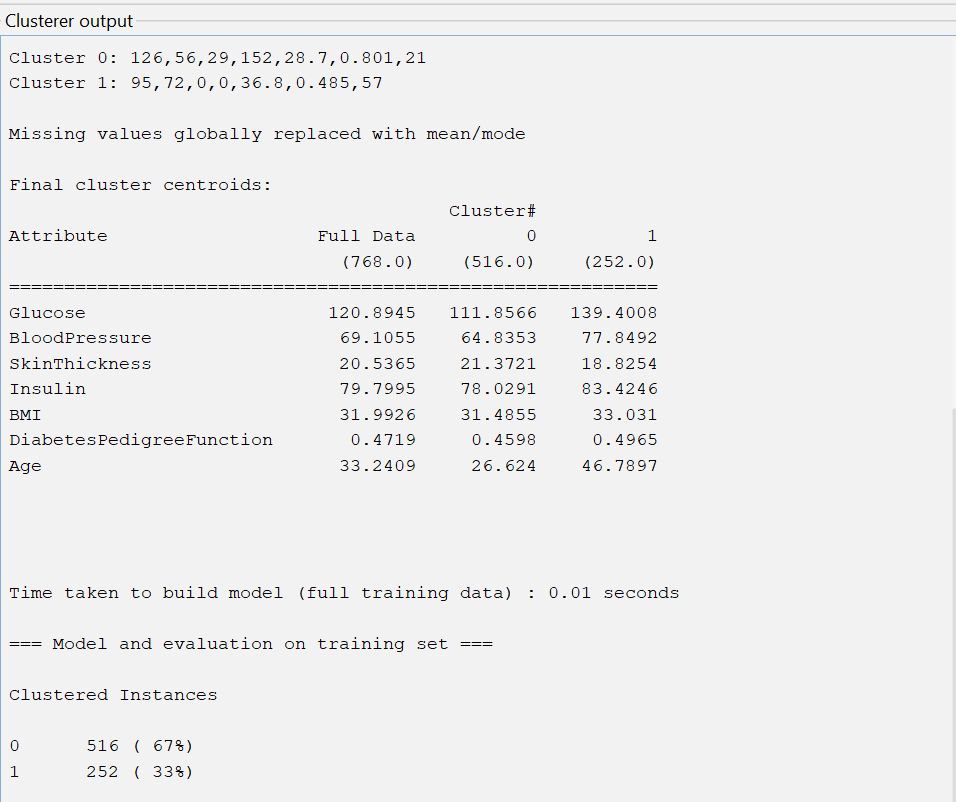
Bước 2: Chon Cluster và chọn SimpleKMeans.

Bước 3: Chọn numClustes = 2

Bước 4: Tiến hành chạy thuật toán phân cụm.

- Kết quả chạy thu được

## *Hình 3.2 Kết quả chạy phân cụm bằng weka*



*Hình 3.3 Kết quả chạy phân cụm bằng weka*

- Giải thích kết quả:

* *Clustering Model Information:*
  + Mô hình được sử dụng là K-means.
  + Số lần lặp được thực hiện trong quá trình phân cụm là 16.
  + Tổng các lỗi bình phương trong cụm (WCSS) là thước đo mức độ gần của các điểm dữ liệu trong mỗi cụm với tâm cụm đó. Giá trị WCSS được báo cáo là 104.113. WCSS thấp hơn thường biểu thị phân cụm tốt hơn.
* *Initial Starting Points:*
  + Phần này hiển thị trọng tâm ban đầu của mỗi cụm. Mỗi dòng tương ứng với một cụm và các giá trị đại diện cho các thuộc tính/tính năng của tâm của cụm đó.
* *Missing Values Globally:*
  + Các giá trị bị thiếu trong tập dữ liệu đã được thay thế trên toàn cầu bằng giá trị trung bình hoặc chế độ của thuộc tính tương ứng.
* *Final Cluster Centroids:*
  + Bảng này hiển thị các giá trị thuộc tính cho trọng tâm của mỗi cụm ( Cột “#Cụm” cho biết số cụm)
  + Hàng "Full Data" đại diện cho các giá trị thuộc tính trên toàn bộ tập dữ liệu.
  + Các số trong ngoặc đơn biểu thị số lượng phiên bản trong mỗi cụm.
  + Các giá trị trong bảng cho biết các giá trị thuộc tính trung bình cho trọng tâm của mỗi cụm. Ví dụ, trong Cụm 0, thuộc tính "glucose" có giá trị trung bình 111.8566.
* *Time Taken:*
  + Mất khoảng 0.01 giây để xây dựng mô hình phân cụm bằng dữ liệu huấn luyện đầy đủ.
* *Model Evaluation:* 
  + Cụm 0 chứa 516 (67% tập dữ liệu)
  + Cụm 1 chứa 252 (33% tập dữ liệu)

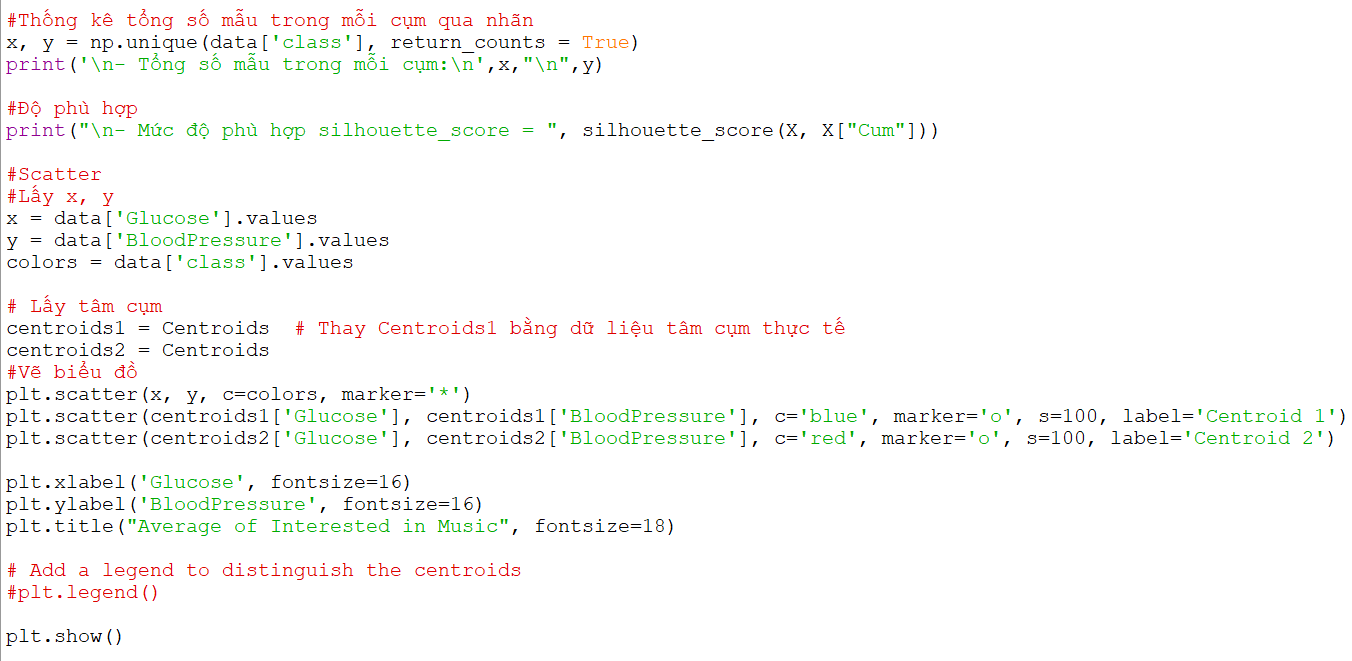
## 

## Chạy phân cụm với python

* Chương trình thực hiện thuật toán K-means:

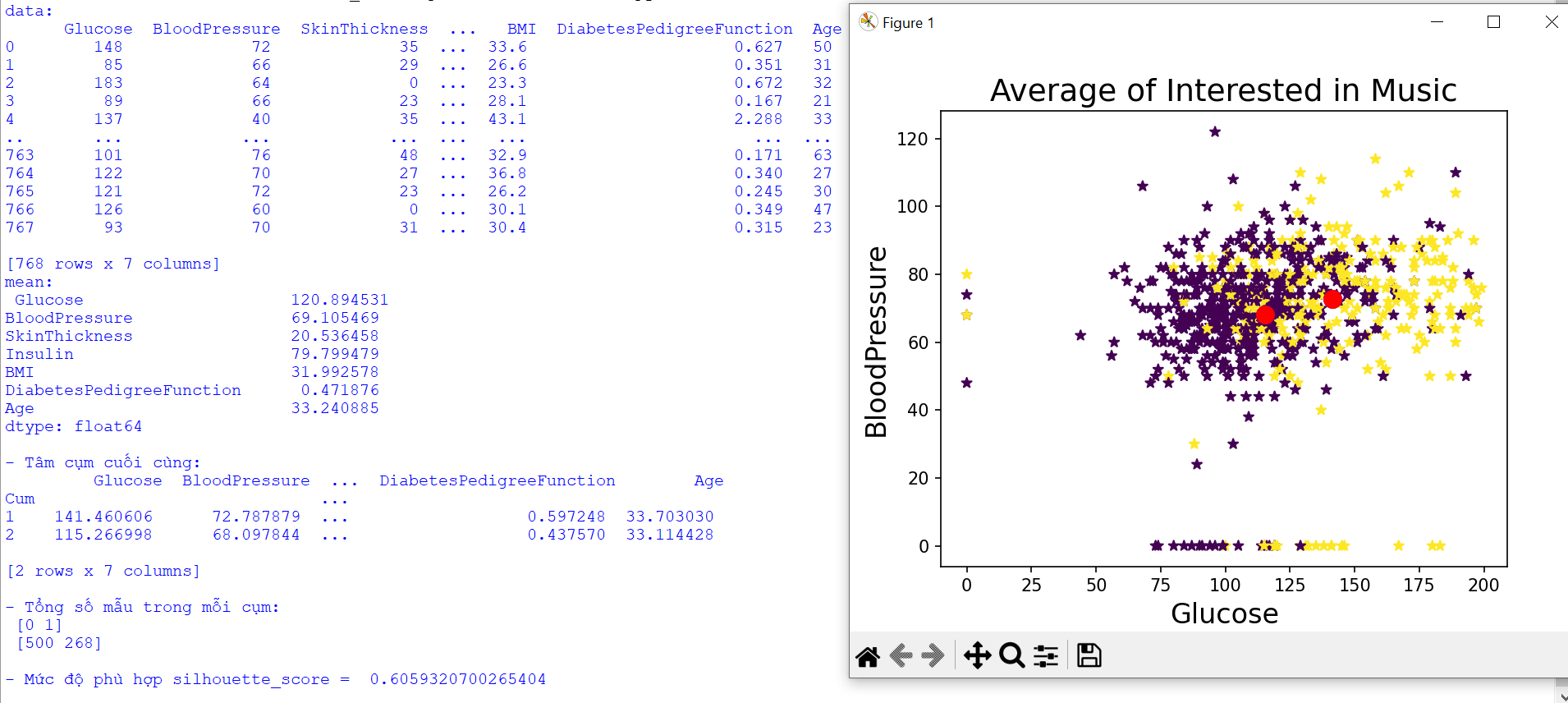


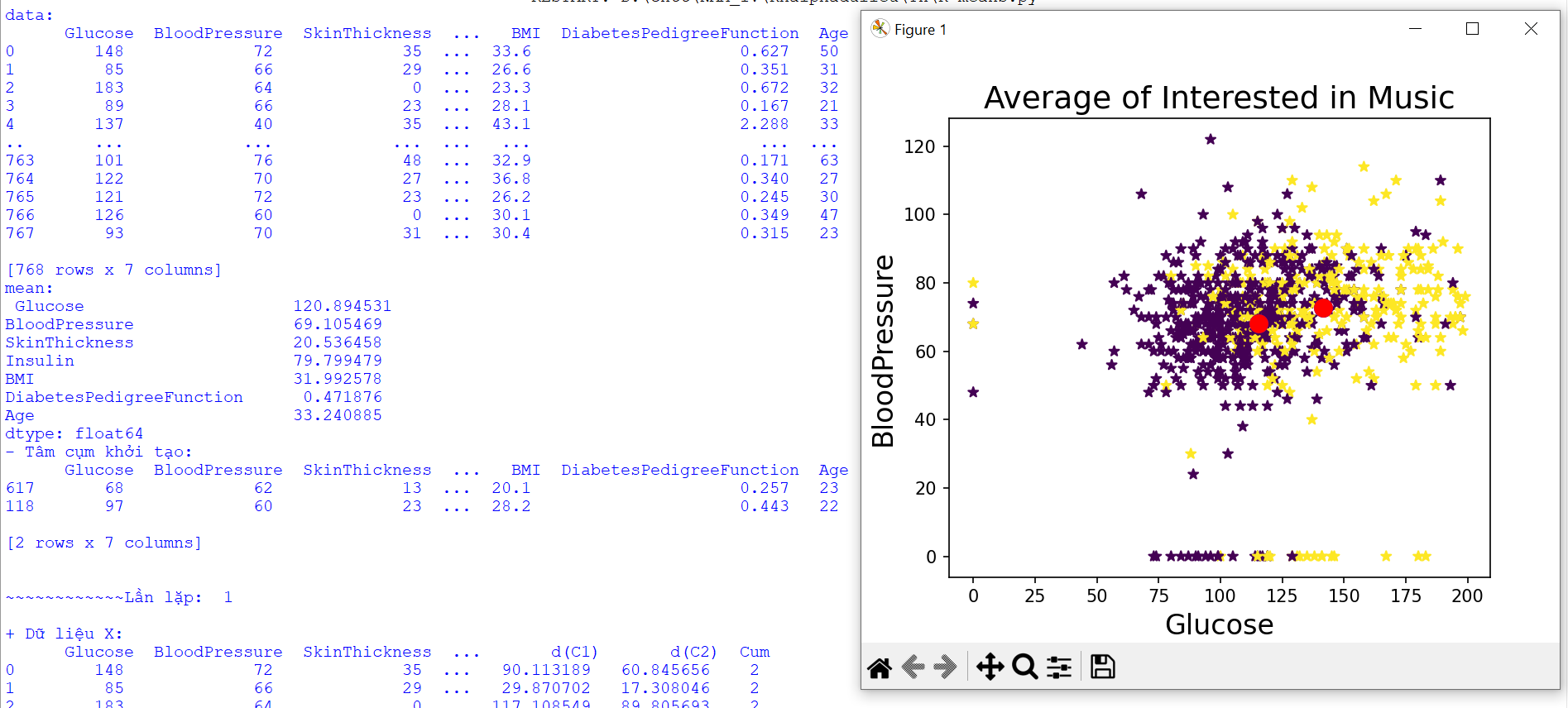


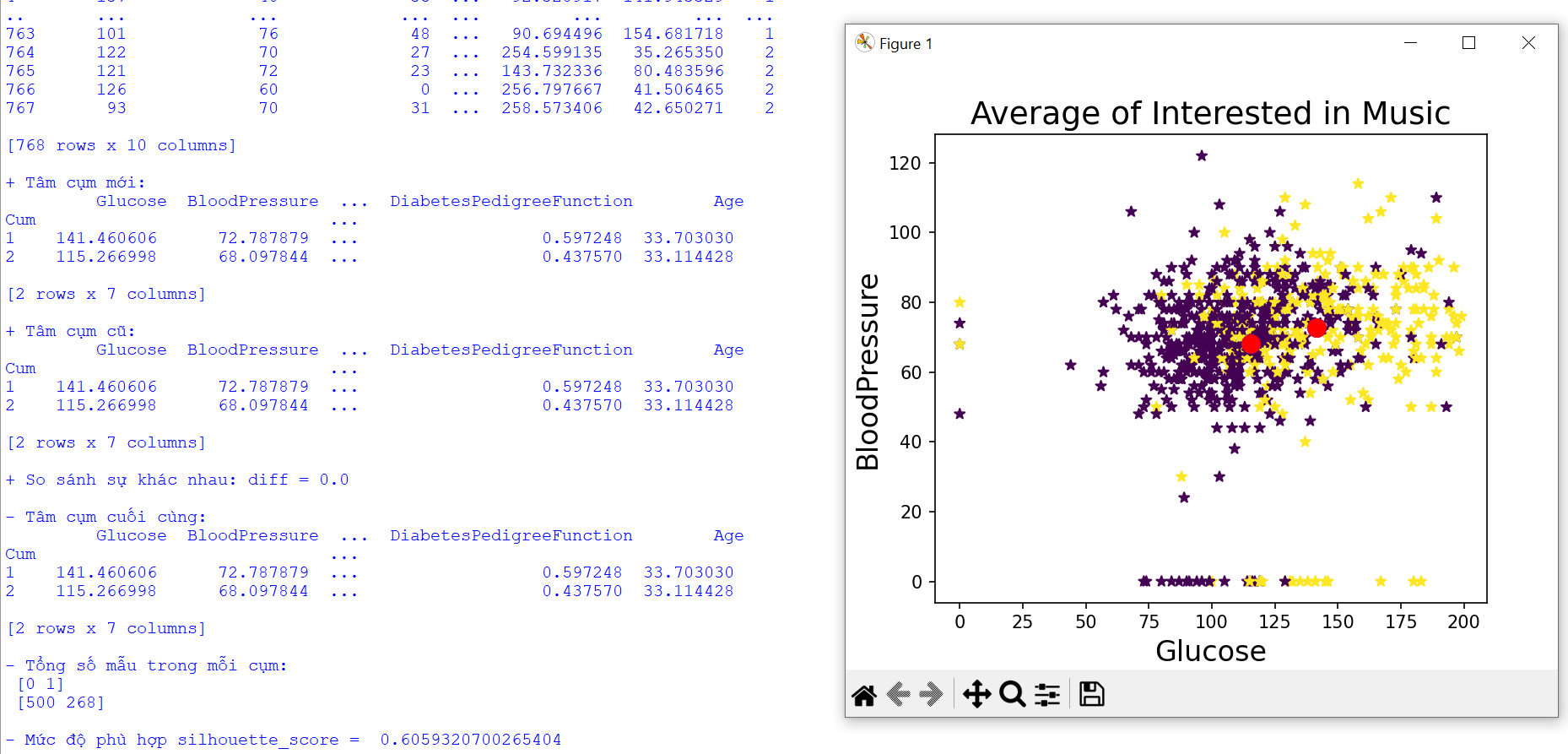


*Chương trình chạy thuật toán K-means trên python*

* Kết quả chạy thu được:







- Data: In ra danh sách các thuộc tính trong dữ liệu ban đầu

- Mean: In ra giá trị trung bình của từng thuộc tính

- Tâm cụm cuối cùng: In ra tọa độ của các tâm cụm cuối cùng sau khi thực hiện thuật toán K-means

- Tổng số mẫu trong mỗi cụm: Thống kê tổng số mẫu trong mỗi cụm thông qua nhãn

- Mức độ phù hợp: Tính toán mức độ phù hợp bằng độ đo Silhouette

- Biểu đồ Scatter: Tạo một biểu đồ dựa trên 2 thuộc tính tùy chọn để hiển thị kết quả phân cụm của thuật toán thông qua nhãn

## Ý nghĩa của thuật toán K-Means

- Phân nhóm dữ liệu: K-means giúp tự động phát hiện cấu trúc và nhóm dữ liệu dựa trên sự tương đồng giữa các điểm. Điều này giúp trong việc tổ chức và tóm tắt thông tin từ tập dữ liệu lớn.

- Khám phá thông tin ẩn: Khi dữ liệu không có nhãn (unlabeled data), K-means có thể giúp tìm ra các cụm tự nhiên trong dữ liệu, từ đó có thể giúp hiểu sâu hơn về sự phân bố và mối quan hệ giữa các điểm dữ liệu.

- Nén dữ liệu: K-means có thể được sử dụng để biểu diễn dữ liệu dưới dạng một số lượng cụm nhỏ hơn, giúp giảm kích thước của dữ liệu mà vẫn giữ lại phần lớn thông tin quan trọng.

- Phát hiện ngoại lệ: Các điểm dữ liệu không rơi vào bất kỳ cụm nào hoặc rơi vào cụm rất xa có thể được xem xét là ngoại lệ. K-means có thể giúp xác định những điểm này.

- Phân loại không giám sát (Unsupervised classification): Khi có các điểm dữ liệu mới, K-means có thể được sử dụng để dự đoán cụm mà điểm mới thuộc về dựa trên các trung tâm cụm đã học từ dữ liệu huấn luyện.

- Tiền xử lý dữ liệu: K-means có thể được sử dụng để tiền xử lý dữ liệu bằng cách gán nhãn cụm cho từng điểm dữ liệu, từ đó tạo ra một biểu diễn mới của dữ liệu.

- Hạn chế: Nó cần phải biết trước số lượng cụm cần phân chia, và nó có thể bị ảnh hưởng bởi điểm khởi tạo ban đầu. Điều này có thể dẫn đến kết quả phân chia khác nhau nếu thuật toán được chạy nhiều lần với các điểm khởi tạo khác nhau.

## 

# Kết luận

# - Phân lớp dữ liệu là lĩnh vực quan trọng trong khai phá dữ liệu, nó được ứng dụng trong nhiều ngành như y tế, kinh tế,… Với bài tập lớn này chúng em đã đạt được một số kết quả:

# + Hiểu về khai phá dữ liệu, bài toán phân cụm từ đó xây dựng được mô hình phân cụm để đánh giá độ yêu thích.

# + Hiểu, biết cách xử lý các dữ liệu bị thiếu và chuẩn hóa dữ liệu

# - Những hạn chế:

# + Do đây là thuật toán học có giám sát nên độ chính xác phụ thuộc vào dữ liệu mà chúng ta thu thập được và số dữ liệu nhiễu trong tập dữ liệu.

# + Tập dữ liệu này chỉ mới có tập trung để đánh giá độ yêu thích chứ chưa hỗ trợ để đánh giá những phần khác.

# - Trong quá trình thực hiện bài tập lớn, chúng em đã cố gắng tìm hiểu và tham khảo những tài liệu liên quan. Nhưng do thời gian và trình độ có hạn nên không tránh khỏi những hạn chế và thiếu sót. Chúng em rất mong nhận được các nhận xét và góp ý của thầy để hoàn thiện hơn kết quả thực nghiệm.

# 

# 

# 

# **TÀI LIỆU THAM KHẢO**

1. TS. Trần Mạnh Tuấn, Bài giảng Khai phá dữ liệu

2. <https://machinelearningcoban.com>

3. <https://www.kaggle.com/code/khnhtrng/m-h-nh-d-o-n-b-nh-ti-u-ng/>