**TRƯỜNG ĐẠI HỌC THỦY LỢI KHOA CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP LỚN**

**MÔN : KHAI PHÁ DỮ LIỆU**

ĐỀ TÀI:

**KHAI PHÁ DỮ LIỆU DỰ ĐOÁN GIÁ LAPTOP**

**BẰNG PHƯƠNG PHÁP K-MEANS**

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

Nhóm sinh viên thực hiện: Nhóm 7\_64TTNT1

|  |  |
| --- | --- |
| **Họ tên sinh viên** | **Mã sinh viên** |
| Thái Văn Sáng | 2251262634 |
| Vũ Thị Như Quỳnh | 2251262633 |
| Nguyễn Văn Quân | 2251262627 |

**Mục lục**

[Chương 1: Tổng quan khai phá dữ liệu 3](#_Toc180867858)

[1.1 Khái niệm 3](#_Toc180867859)

[1.2 Quy trình khám phá tri thức trong CSDL 3](#_Toc180867860)

[1.3 Các kỹ thuật khai phá dữ liệu 4](#_Toc180867861)

[1.3.1 Kỹ thuật phân cụm(Clustering) 4](#_Toc180867862)

[1.3.2 Kỹ thuật phân lớp(Classification) 5](#_Toc180867863)

[1.3.3 Khai phá bằng luật kết hợp(Association rule) 5](#_Toc180867864)

[1.3.4 Phát hiện bất thường(Anomaly detection) 5](#_Toc180867865)

[Chương 2: Quy trình khai phá dữu liệu 6](#_Toc180867866)

[2.1 Tiền xử lý dữ liệu 7](#_Toc180867867)

[2.1.1 Làm sạch dữ liệu 7](#_Toc180867868)

[2.1.2 Tích hợp dữ liệu 10](#_Toc180867869)

[2.1.3 Biến đổi dữ liệu (chuẩn hóa dữ liệu) 10](#_Toc180867870)

[2.1.4 Dữ liệu sau khi tiền xử lý 14](#_Toc180867871)

[Chương 3: Khai phá dữ liệu bằng thuật toán K-Means 15](#_Toc180867872)

[3.1 Giới thiệu về bài toán phân cụm K-mean. 15](#_Toc180867873)

[3.2 Thuật toán phân cụm K-means 15](#_Toc180867874)

[\* Các bước thực hiện thuật toán K-means trên Weka : 16](#_Toc180867875)

[Chương 4: Demo ứng dụng 22](#_Toc180867876)

[4.1 Giao diện bắt đầu 22](#_Toc180867877)

[4.2 Giao diện dự đoán giá laptop 22](#_Toc180867878)

[Chương 5: Kết luận và hướng phát triển 23](#_Toc180867879)

[5.1 Đánh giá mô hình 23](#_Toc180867880)

[5.2 Kết luận 23](#_Toc180867881)

[5.3 Hướng phát triển 23](#_Toc180867882)

**MỞ ĐẦU**

Trong thời đại công nghệ số phát triển mạnh mẽ như hiện nay, máy tính xách tay (laptop) đã trở thành một công cụ không thể thiếu trong cuộc sống và làm việc của con người. Sự đa dạng về mẫu mã, cấu hình và giá cả của các dòng laptop trên thị trường khiến người tiêu dùng gặp không ít khó khăn trong việc lựa chọn sản phẩm phù hợp. Do đó, việc xây dựng một mô hình dự đoán giá laptop chính xác trở nên cấp thiết hơn bao giờ hết. Vì vậy, chúng em chọn đề tài **“Khai phá dữ liệu dự đoán giá laptop bằng phương pháp K-means”** nhằm thực hiện xây dựng một mô hình dự đoán giá laptop dựa trên thuật toán K-means, một trong những thuật toán phân cụm phổ biến và hiệu quả nhất.

Đề tài này được thực hiện dựa trên cơ sở lý thuyết về khai thác dữ liệu và các thuật toán phân cụm. Phần lớn dữ liệu sử dụng trong đề tài được thu thập từ các trang web bán hàng trực tuyến.

# 

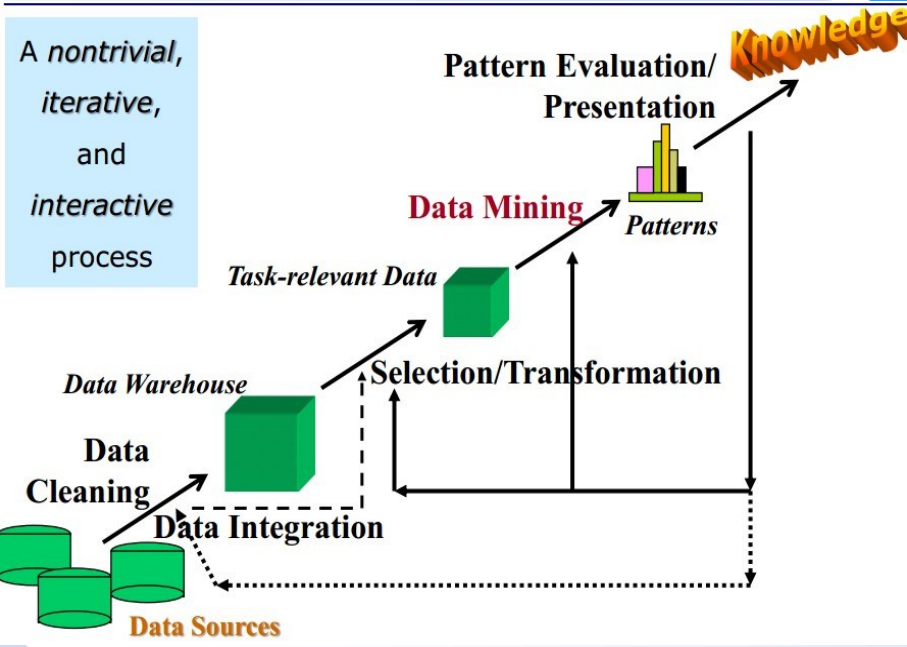
# Chương 1: Tổng quan khai phá dữ liệu

## 1.1 Khái niệm

- Khai phá dữ liệu (hay còn gọi là **data mining**) là quá trình trích xuất thông tin có giá trị và hữu ích từ một tập dữ liệu lớn. Mục tiêu của khai phá dữ liệu là tìm ra các mẫu, xu hướng, mối quan hệ ẩn chứa trong dữ liệu mà ban đầu không rõ ràng. Đây là một phần quan trọng trong lĩnh vực khoa học dữ liệu, thường được sử dụng để hỗ trợ việc ra quyết định, dự đoán và phân tích.

## 1.2 Quy trình khám phá tri thức trong CSDL

- Khai phá dữ liệu là một giai đoạn quan trong trong quá trình KDD ( Knowledge Discovery in Databases - Khai phá tri thức từ cơ sở dữ liệu)



- Quá trình phát hiện tri thức bao gồm các bước:

**+** Làm sạch dữ liệu (Data cleaning & preprocessing): Các nhiễu và dữ liệu không nhất quán sẽ được loại bỏ

**+** Tích hợp dữ liệu (Data integration ): Tích hợp các nguồn dữ liệu khác nhau vào một nguồn chung. Quá trình này đảm bảo tính nhất quán của dữ liệu

**+** Lựa chọn dữ liệu ( Data selection ): Những dữ liệu thích hợp với nhiệm vụ phân tích sẽ được trích rút ra từ CSDL.

**+** Chuyển đổi dữ liệu (Data transformation): Dữ liệu sau khi được chọn lọc sẽ được chuyển đổi hay hợp nhất về dạng thích hợp cho việc khai phá.

**+** Khai phá dữ liệu (Data mining): Đây là bước quan trọng nhất, sử dụng kĩ thuật và thuật toán để kham phá mẫu, mối quan hệ và tri thức ẩn trog dữ liệu. Điều này có thể liên quan đến việc phân loại, gom cụm, dự đoán và tìm kiếm các luật kết hợp

**+** Đánh giá mẫu (Pattrn evaluation): Đánh giá tính hợp hợp lý và giá trị của các mẫu được khám phá.Thường được đánh giá dựa trên các độ đo nhất định

**+** Biểu diễn tri thức (Knowledge presentation): Giai đoạn này các kĩ thuật biểu diễn và hiển thị tri thức sẽ được sử dụng để đưa tri thức đã lấy ra đến người dùng.

## 1.3 Các kỹ thuật khai phá dữ liệu

### 1.3.1 Kỹ thuật phân cụm(Clustering)

- Một nhiệm vụ mô tả phổ biến trong đó người ta tìm cách xác định một tập hợp hữu hạn các cụm để mô tả dữ liệu. Ví dụ, phân nhóm khách hàng theo sở thích, phân nhóm sản phẩm theo đặc tính,…

### 1.3.2 Kỹ thuật phân lớp(Classification)

- Là phương pháp dự báo, cho phép phân loại một đối tượng vào một hoặc một số lớp cho trước. Ví dụ, phân loại email là rác hay không rác, phân loại khách hàng là tiềm năng hay không tiềm năng,…

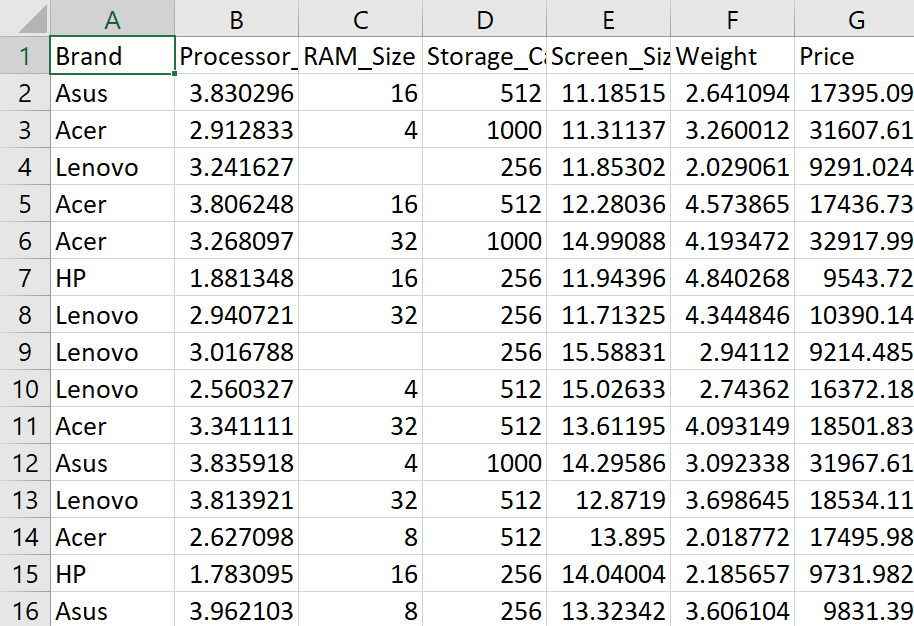
### 1.3.3 Khai phá bằng luật kết hợp(Association rule)

- Là phương pháp để xác định các mối quan hệ giữa các biến trong cơ sở dữ liệu. Ví dụ, luật kết hợp trong giỏ hàng để tìm ra các sản phẩm thường được mua cùng nhau, luật kết hợp trong web mining để tìm ra các trang web thường được truy cập cùng nhau,…

### 1.3.4 Phát hiện bất thường(Anomaly detection)

- Là phương pháp để nhận diện các mục dữ liệu bất thường hoặc khác biệt so với mẫu chung. Ví dụ, phát hiện gian lận thẻ tín dụng, phát hiện xâm nhập mạng, phát hiện bệnh lý y tế,…

* Dữ liệu đầu vào: dữ liệu dự đoán giá cả máy tính



* Số lượng thuộc tính: 7 thuộc tính
* Số lượng mẫu: 1001 mẫu

*Bảng mô tả thuộc tính của dữ liệu*

|  |  |  |
| --- | --- | --- |
| STT | Thuộc tính | Mô tả |
| 1 | brand | Tên hãng |
| 2 | processor\_Speed | Tốc độ xử lý |
| 3 | ram\_size | Kích thước bộ nhớ tạm (đơn vị GB) |
| 4 | Storage\_Capacity | Dung lượng lưu trữ (lượng dữ liệu có thể được lưu trữa trong thiết bị) |
| 5 | Screen\_Size | Kích thước của màn hình |
| 6 | Weight | Khối lượng của thiết bị |
| 7 | Price | Gía thành |

# Chương 2: Quy trình khai phá dữu liệu

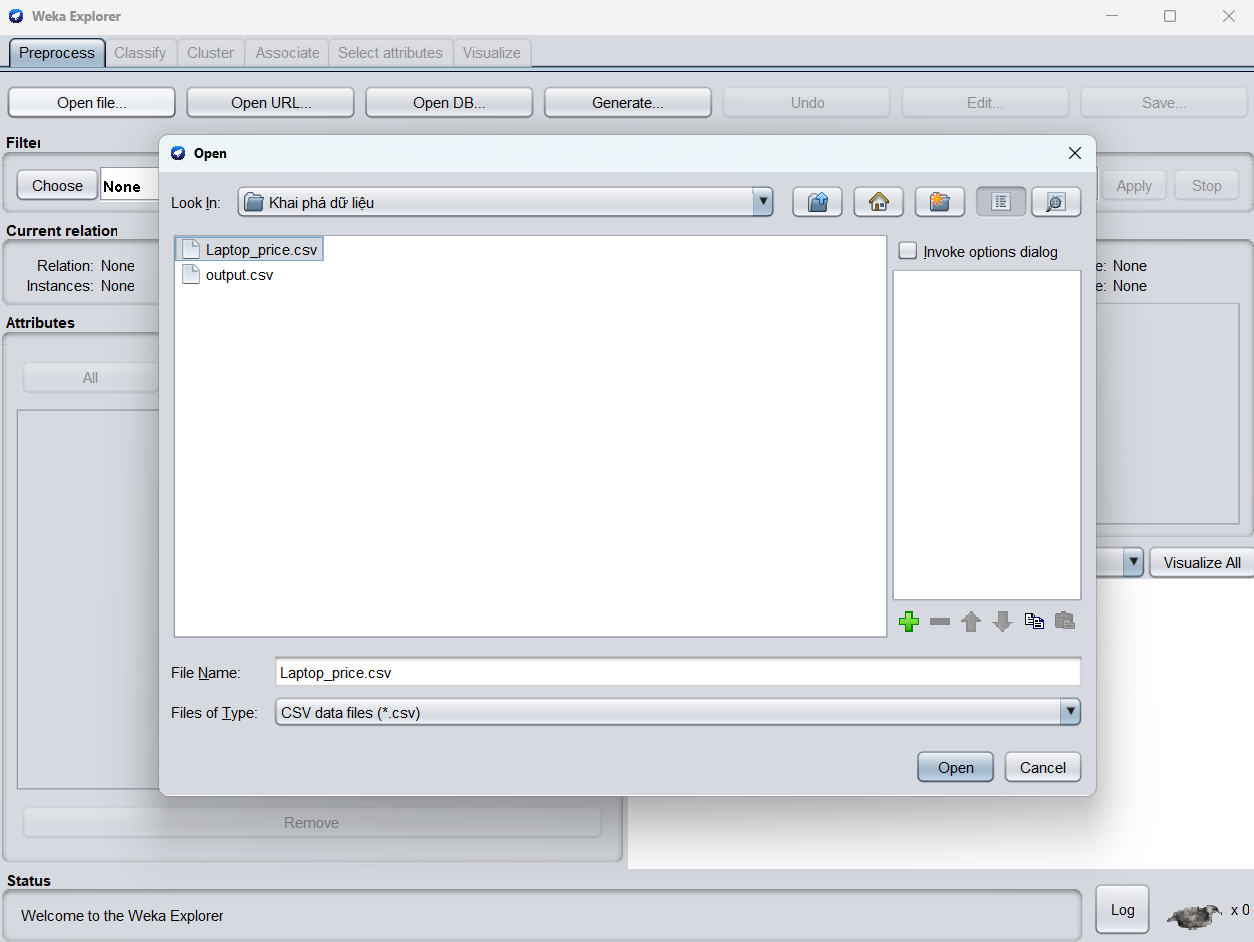
# 2.1 Tiền xử lý dữ liệu

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

**2.1.1 Làm sạch dữ liệu**

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

Đọc dữ liệu vào Weka:

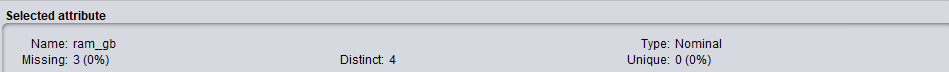


**- Loại bỏ thuộc tính dư thừa, không cần thiết**

=> Không có thuộc tính cần loại bỏ

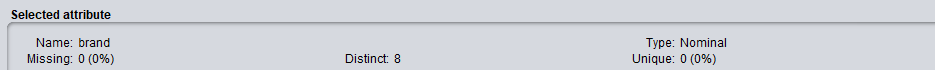
**- Xử lý dữ liệu bị thiếu (missing data)**

* Dữ liệu thiếu ở phần processor\_name :



* Thay thế dữ liệu thiếu:

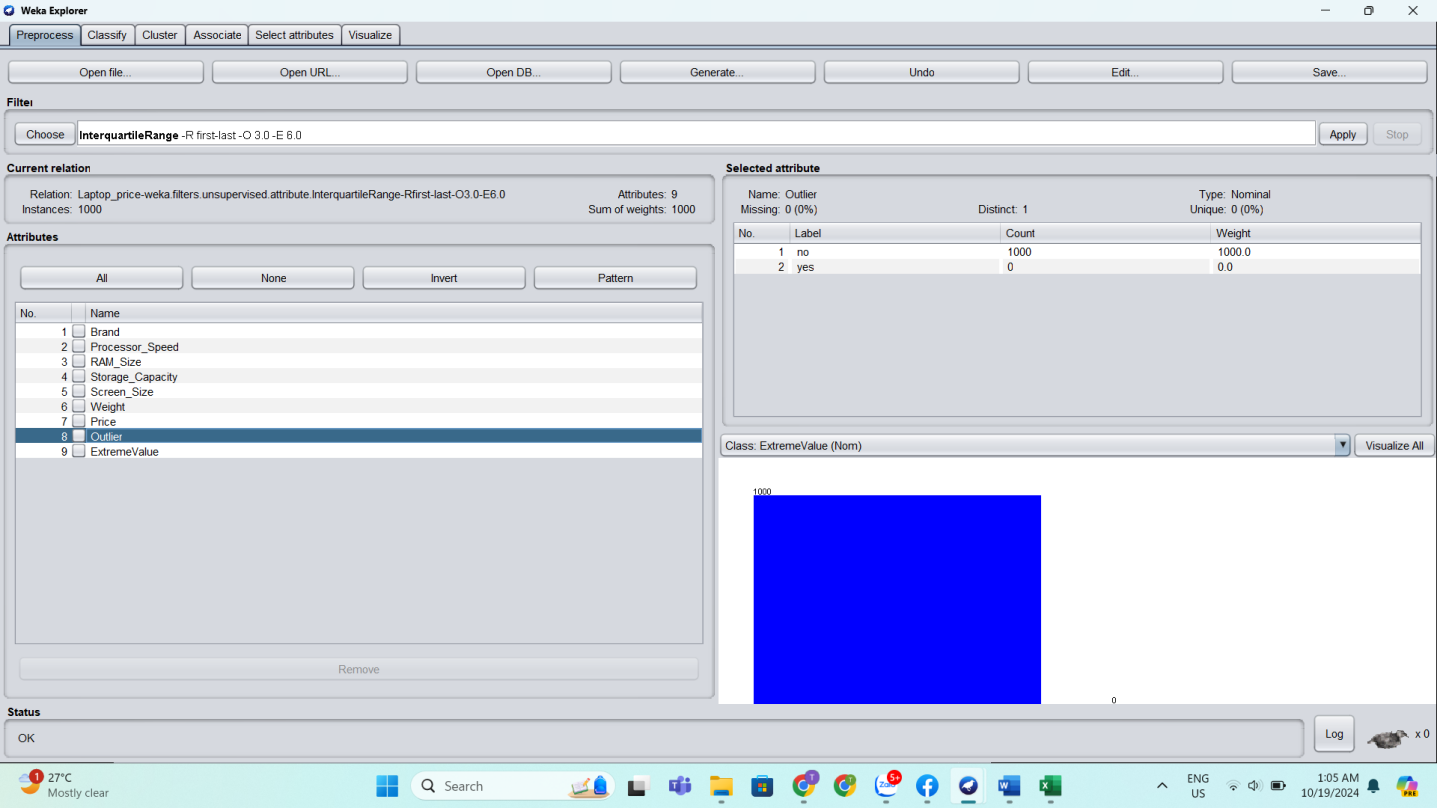
**Filter -> Unsupervised -> Attribute -> ReplaceMissingValues-> Apply**



**- Loại bỏ giá trị ngoại lệ và cực đoan:**

* + Những điểm dữ liệu nằm ngoài phạm vi bình thường của tập dữ liệu, có thể ảnh hưởng đáng kể đến kết quả phân tích.
  + Sử dụng bộ lọc **InterquartileRange** để phát hiện các giá trị ngoại lệ (outlier) và giá trị cực đoan (extreme values).

**Filter -> Unsupervised -> Attribute -> InterquartileRange -> Apply**



* + Nhận thấy không có phần tử ngoại lệ và cực đoan nên không cần loại bỏ.

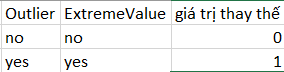
**- Loại bỏ nhiễu :**

Xử lý dữ liệu nhiễu, không hợp lý có thể ảnh hưởng đến kết quả phân tích:

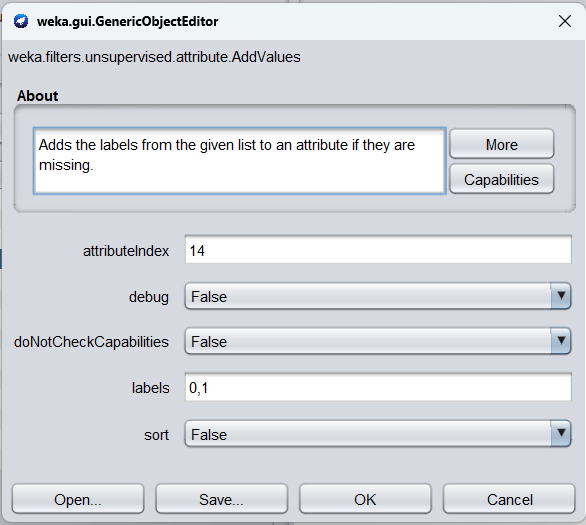
* + Không có dữ liệu gây nhiễu

Xử lý dữ liệu không nhất quán

* + Thực hiện thay thế các giá trị:

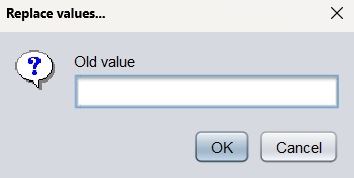
 

* Sử dụng **AddValues** để thêm các giá trị mới để thay thế

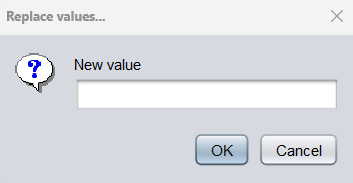


*Sử dụng AddValues*

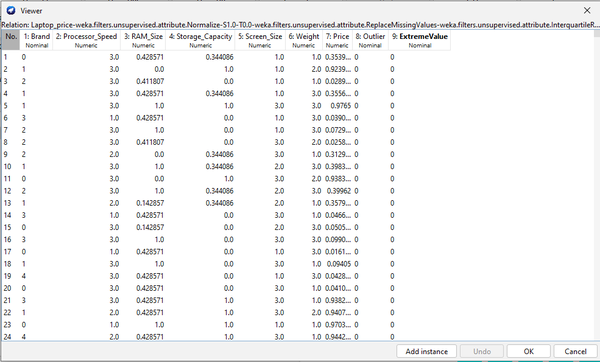
* Sử dụng **Replace values with** để chuyển đổi giá trị cũ sang giá trị mới

**

*Sử dụng Old value*

**

*Sử dụng New value*



*Dữ liệu sau khi thay thế*

### 2.1.2 Tích 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

- Liên quan đến cấu trúc và tính không thuần nhất (heterogeneity) về ngữ nghĩa (semantics) của dữ liệu

- Hỗ trợ việc giảm và tránh dư thừa và không nhất quán về dữ liệu > cải thiện tính chính xác và tốc độ quá trình khai phá dữ liệu

**⇨ Bộ dữ liệu đang dùng được lấy từ một nguồn nên không cần thực hiện quá trình này**

### 2.1.3 Biến đổi dữ liệu *(chuẩn hóa dữ liệu)*

- Là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu.

- Bao gồm :

● Làm trơn dữ liệu

● Kết hợp dữ liệu

● Tổng quát hóa

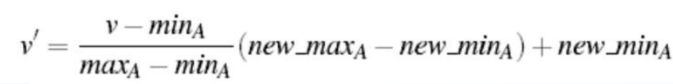
● Chuẩn hóa

● Xây dựng thuộc tính/đặc tính

● Thu giảm dữ liệu

**\* Chuẩn hóa dữ liệu** :

⇨ *Thực hiện chuẩn hóa dữ liệu:*



⇨ Chuyển dữ liệu về dạng min-max của thuộc tính Prosessor\_Speed

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Mô tả được tạo tự động

*Sau khi chuẩn hóa*

Ảnh có chứa văn bản, ảnh chụp màn hình, màn hình, phần mềm

Mô tả được tạo tự động

⇨ Chuyển dữ liệu về dạng min-max của thuộc tính Screen\_Size

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Mô tả được tạo tự động

*Sau khi chuẩn hóa*

Ảnh có chứa văn bản, ảnh chụp màn hình, màn hình, số

Mô tả được tạo tự động

⇨ Chuyển dữ liệu về dạng min-max của thuộc tính weight

Ảnh có chứa văn bản, ảnh chụp màn hình, số, màn hình

Mô tả được tạo tự động

*Sau khi chuẩn hóa*

Ảnh có chứa văn bản, ảnh chụp màn hình, số, màn hình

Mô tả được tạo tự động

### 2.1.4 Dữ liệu sau khi tiền xử lý

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

**Chương 3: Khai phá dữ liệu bằng thuật toán K-Means**

## 3.1 Giới thiệu về bài toán phân cụm K-mean.

- 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 quy 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

* 1. **Thuật toán phân cụm K-means**

- Giới thiệu thuật toán 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 định 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 thực hiện qua các bước chính sau:

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

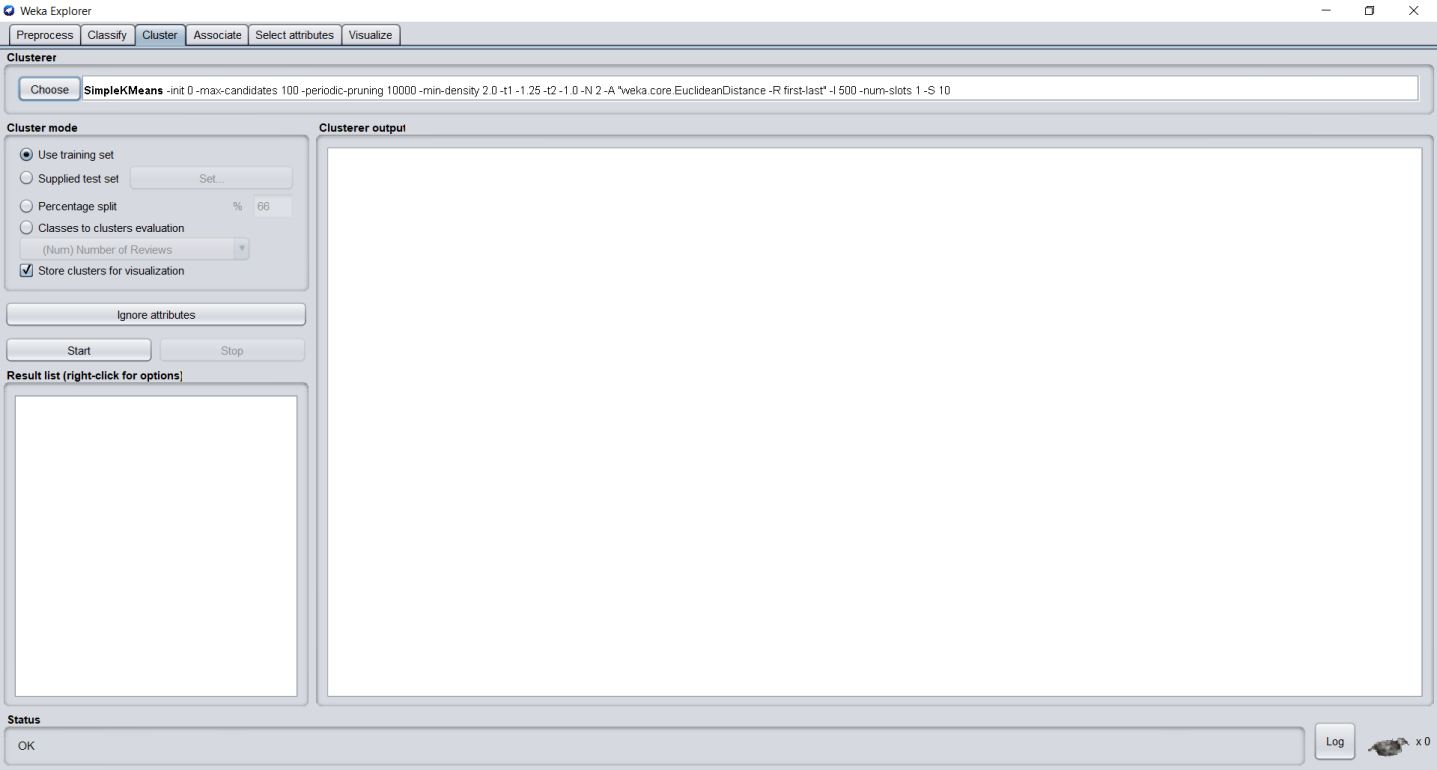
**\* Các bước thực hiện thuật toán K-means trên Weka :**

- Bước 1: Mở file .csv đã tiền xử lý bằng Weka trước đó trên Weka Dataset như sau:

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự động

- Bước 2: Trên giao diện Weka chọn Cluster sau đó chọn SimpleKMeans.



- Bước 3: Tiến hành đánh giá hiệu quả phân cụm của thuật toán đối với tập dữ liệu được dùng theo hai phương pháp:

* Phương pháp 1: Use training set

Ảnh có chứa văn bản, ảnh chụp màn hình, màn hình, phần mềm

Mô tả được tạo tự động

* *Kết quả thu được*
* Cụm khởi tạo ngẫu nhiên

+ Cluser 0: 2,2,0.4,0.3,3,3,0.3,0,0

+ Cluser 1: 2,1,0.1,1,2,2,0.9,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1

(1000.0) (610.0) (390.0)

========================================

Brand 4 4 0

Processor\_Speed 2.2 2.2 2.1

RAM\_Size 0.4 0.4 0.4

Storage\_Capacity 0.4 0.1 0.9

Screen\_Size 2.2 2.3 2.2

Weight 2.2 2.2 2.2

Price 0.4 0.1 0.8

Outlier 0 0 0

ExtremeValue 0 0 0

Time taken to build model (full training data): 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 610 (61%) - tỉ lệ dữ đoán đúng 61%

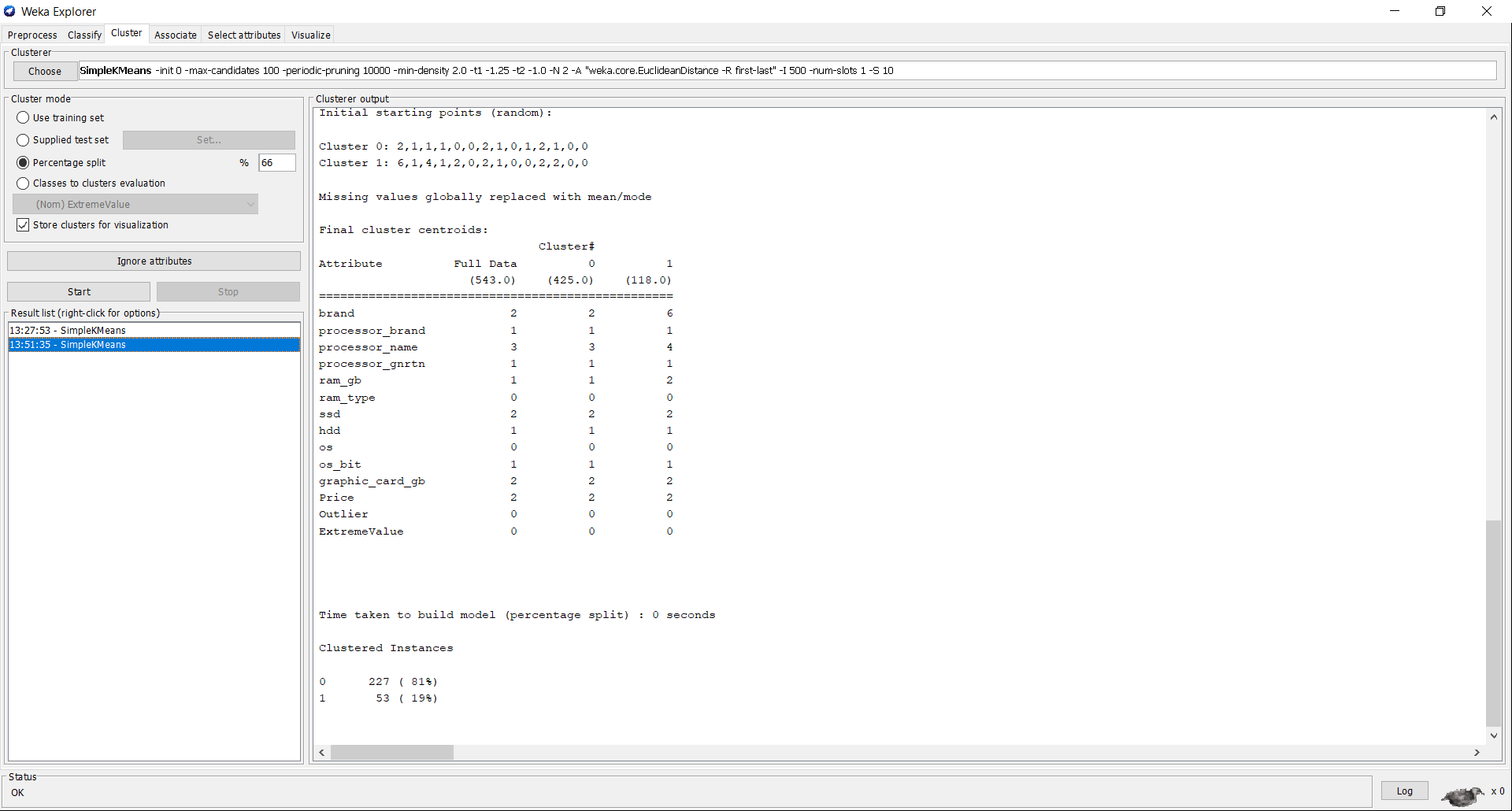
1 390 (39%) - tỉ lệ dự đoán sai 39%

*-* Các cụm sau khi được phân chia bằng Use training set

Ảnh có chứa văn bản, ảnh chụp màn hình, phần mềm

Mô tả được tạo tự động

* Phương pháp 2: Percentage split



* *Kết quả thu được*
* Cụm khởi tạo ngẫu nhiên

Cụm 0: 1,1,1,1,1,2,0.9,0,0

Cụm 1: 1,1,0.1,0.3,3,1,0.3,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1

(660.0) (255.0) (405.0)

===================================================

Brand 0 2 4

Processor\_Speed 2.2 2.2 2.2

RAM\_Size 0.4 0.4 0.4

Storage\_Capacity 0.4 0.9 0.1

Screen\_Size 2.2 2.3 2.2

Weight 2.2 2.3 2.1

Price 0.4 0.8 0.1

Outlier 0 0 0

ExtremeValue 0 0 0

Time taken to build model (percentage split): 0 seconds

Clustered Instances

0 130 (38%) – tỷ lệ dự đoán đúng: 38%

1. 210 (62%) – tỷ lệ dự đoán sai: 62%

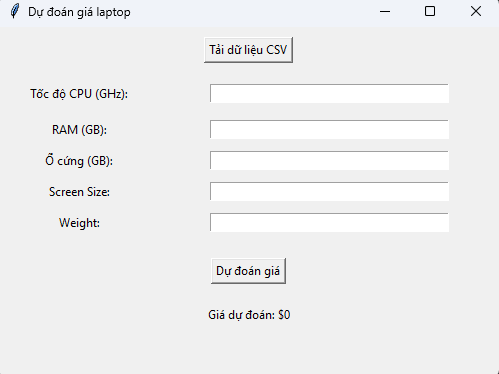
* Các cụm sau khi được phân chia bằng Use training set

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, phần mềm

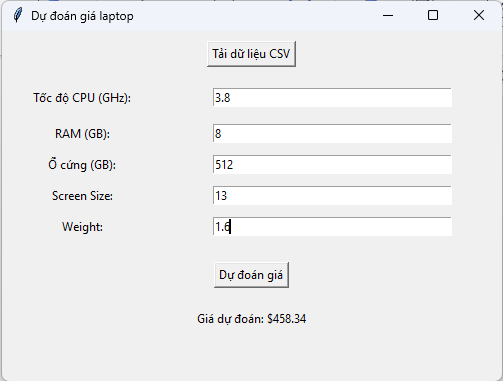
Mô tả được tạo tự động

**Chương 4: Demo ứng dụng**

**4.1 Giao diện bắt đầu**



**4.2 Giao diện dự đoán giá laptop**



# Chương 5: Kết luận và hướng phát triển

## 5.1 Đánh giá mô hình

- Sử dụng phương pháp "Use training set", mô hình đạt tỷ lệ dự đoán đúng là 61% và sai là 39%.

- Khi áp dụng phương pháp "Percentage split", tỷ lệ dự đoán đúng là 38% và sai là 62%.

=> Sử dụng phương pháp "Use training set" cho kết quả dự đoán cao hơn

## 5.2 Kết luận

-Phương pháp phân cụm K-Means có thể hữu ích trong việc nhóm các laptop có những đặc điểm tương đồng lại với nhau, chẳng hạn như dựa trên các thông số kỹ thuật (RAM, CPU, ổ cứng, màn hình, trọng lượng) và giá cả. Điều này giúp chia thị trường thành các phân khúc giá khác nhau. Trong quá trình tìm hiểu và làm đề tài “ Dự đoán giá laptop bằng phương pháp K-Means “ nhóm chúng em đã đạt được những kết quả sau:

+ Tìm hiểu tổng quan về khai phá dữ liệu, các thuật toán từ đó xây dựng mô hình dự đoán giá laptop

+ Thu thập dữ liệu , tiền xử lý dữ liệu trên weka

+ So sánh kết quả tỷ lệ train/test để lựa chọn tỷ lệ đánh giá mô hình tốt nhất

## 5.3 Hướng phát triển

- Ứng dụng các phương pháp khác như KNN, **DBSCAN** để so sánh với K-Means và tìm ra phương pháp phù hợp hơn.

- **Kết hợp các thuật toán khác:** Để tăng độ chính xác của việc dự đoán như **hồi quy tuyến tính, hồi quy Ridge/Lasso, Random Forest, hoặc mô hình học sâu (Deep Learning)**

\* Trong quá trình làm bài tập lớn chúng em đã tham khảo rất nhiều nguồn . Tuy nhiên , chúng em vẫn có những thiếu sót, rất mong sự góp ý của thầy cô và các bạn để bài báo cáo của chúng em hoàn thiện hơn