**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

**----- 🙢🕮🙠 -----**

****

**ĐỀ TÀI: Khai phá dữ liệu bằng phương pháp phân cụm và phân lớp để đánh giá chất lượng sữa**

**Nhóm sinh viên thực hiện:** **Nhóm 4**

**Nguyễn Thị Mây: 2051063448**

**Vũ Thị Thanh: 2051063658**

**Nguyễn Hồng Thương: 2051063623**

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

**Hà nội ngày 25 tháng 8 năm 2023**

MỤC LỤC

DANH MỤC CÁC HÌNH ẢNH 4

MỞ ĐẦU 4

Bảng phân chia công việc: 5

CHƯƠNG 1 Tổng quan về khai phá dữ liệu 7

1.1 Khai phá tri thức và khai phá dữ liệu 7

1.2 Quá trình khai phá tri thức 7

1.3 Mô tả bài toán 7

1.3.1 Tổng quan về bài toán 7

CHƯƠNG 2 Tiền xử lý dữ liệu 9

2.1 Làm sạch dữ liệu 9

2.2 Tính hợp dữ liệu 9

2.3 Biến đổi dữ liệu 10

CHƯƠNG 3 Khai phá dữ liệu bằng mô hình phân cụm 10

3.1 Giới thiệu về phân cụm 10

3.1.1 Khái niệm 10

3.1.2 Các phương pháp phân cụm 11

3.1.3 Các kỹ thuật đều hướng tới: 11

3.2 Thuật toán K-Means 11

3.2.1 Các bước thực hiện: 11

3.2.2 Điều kiện dừng: 12

3.2.3 Đánh giá mô hình 12

3.3 Chạy phân cụm với Weka 12

3.4 Chạy phân cụm với python 15

3.5 Ý nghĩa của thuật toán K-Means 17

CHƯƠNG 4 Khai phá dữ liệu bằng mô hình phân lớp 18

4.1 Giới thiệu về phân lớp 18

4.1.1 Khái niệm 18

4.1.2 Các phương pháp phân lớp 19

4.2 Phân lớp bằng cây quyết định 19

4.2.1 Khái niệm 19

4.2.2 Thuật toán tạo cây quyết định 19

4.2.3 Điều kiện dừng phân chia tập học 20

4.2.4 Độ lợi thông tin trong cây quết định 20

4.3 Thuật toán phân lớp ID3 20

4.3.1 Chạy phân lớp với weka 21

4.3.2 Chạy phân lớp với python 24

4.4 Ý nghĩa của thuật toán ID3 25

CHƯƠNG 5 Kết luận 25

TÀI LIỆU THAM KHẢO 27

DANH MỤC CÁC HÌNH ẢNH

[Hình 1.2 Quá trình khai phá tri thức 7](#_Toc144061885)

[Hình 1.3 Dữ liệu ban đầu 8](#_Toc144061886)

[Hình 3.1 Quá trình gom cụm dữ liệu 10](#_Toc144061887)

[Hình 3.2.3 Công thức silhouette\_score 12](#_Toc144061888)

[Hình 3.2 Kết quả chạy phân cụm bằng weka 13](#_Toc144061889)

[Hình 3.3 Kết quả chạy phân cụm bằng weka 14](#_Toc144061890)

[Hình 3.4 Chương trình chạy thuật toán k-mean trên python 16](#_Toc144061891)

[Hình 3.5 Kết quả tọa độ tâm cụm 16](#_Toc144061893)

[Hình 3.6 Kết quả độ đo silhouette 17](#_Toc144061894)

[Hình 4.1 Kết quả chạy phân lớp bằng weka 21](#_Toc144061895)

[Hình 4.2.Kết quả chạy phân lớp bằng weka 22](#_Toc144061896)

[Hình 4.3.2.1 .Chương trình chạy thuật toán id3 trên python 23](#_Toc144061897)

[Hình 4.3.2.2 .Chương trình chạy thuật toán id3 trên python 24](#_Toc144061898)

[Hình 4.3.2.3 Kết quả cây quyết định lưu dưới dạng từ điển 25](#_Toc144061899)

[Hình 4.3.2.4 Kết quả cây quyết định được triển khai lại dưới dạng cây. 25](#_Toc144061901)

[Hình 4.3.2.5 Kết quả nhãn dự đoán trên tập test 26](#_Toc144061902)

[Hình 4.3.2.6 Kết quả đánh giá trên độ đo precision, recall, f1-score 26](#_Toc144061903)

MỞ ĐẦU

* Sữa cung cấp các chất dinh dưỡng cần thiết và là một nguồn cung cấp năng lượng dồi dào, giàu protein và chất béo. Sữa đóng góp một lượng đáng kể các chất dinh dưỡng như: canxi, magie, selen, riboflavin, vitamin B12 và axit pantothenic.
* Chất lượng vệ sinh của sữa có vai trò rất quan trọng trong việc sản xuất sữa và các sản phẩm từ sữa an toàn. Nếu sản phẩm sữa có hàm lượng Bacillus Cerus cao sẽ ảnh hưởng không tốt đến thời hạn sử dụng của sữa. Hay nếu hàm lượng Clostridium trong sữa quá cao cũng sẽ tác động đến quá trình sản xuất các sản phẩm phô mai.
* Vì thế việc đánh giá chất lượng sữa là rất quan trọng. Do đó nhóm em chọn đề tài “Khai phá dữ liệu bằng phương pháp phân lớp và phân cụm để đánh giá chất lượng sữa” để làm báo cáo cho môn học này.

Bảng phân chia công việc:

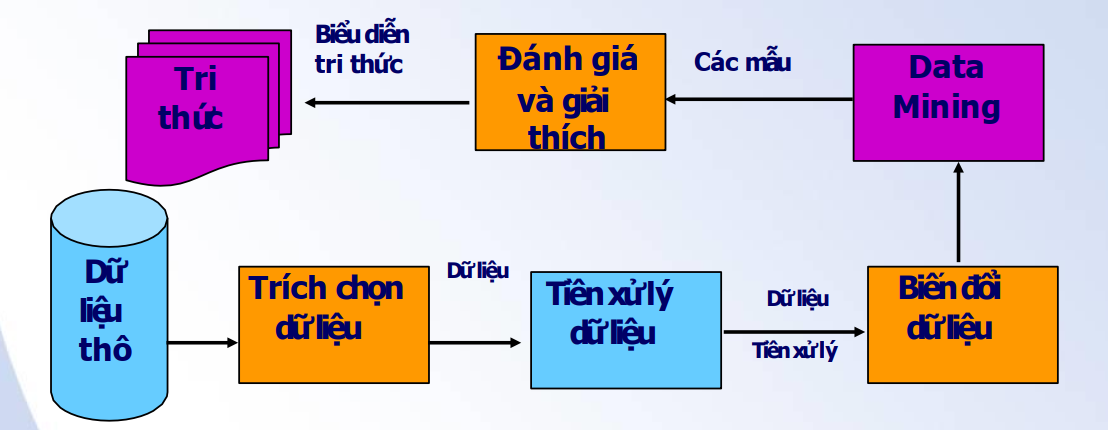
|  |  |
| --- | --- |
| **Tên thành viên** | **Công việc** |
| Nguyễn Thị Mây | Tìm hiểu thuật toán, code thuật toán. |
| Vũ Thị Thanh | Tìm hiểu thuật toán, viết báo cáo. |
| Nguyễn Hồng Thương | Tìm hiểu thuật toán, viết báo cáo. |

# 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

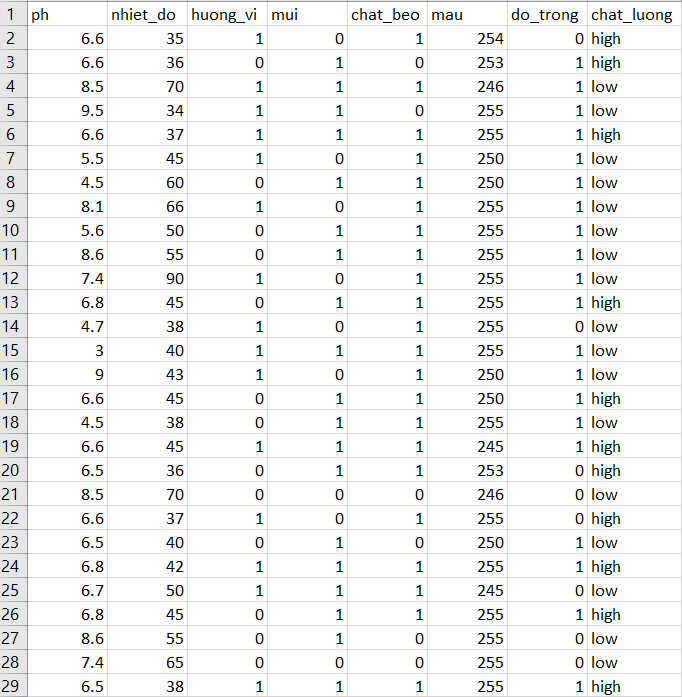


Hình 1.2 Quá trình khai phá tri thức

## Mô tả bài toán

### Tổng quan về bài toán

* Dữ liệu gồm các thuộc tính để dự đoán chất lượng sữa. Dữ liệu bao gồm 685 mẫu và 8 thuộc tính để đánh giá chất lượng sữa.
* Mục đích của bài toán: Phân cụm và phân lớp chất lượng sữa dựa vào các thuộc tính: pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong, Màu sữa, Chất lượng.



Hình 1.3 Dữ liệu ban đầu

Các thuộc tính là:

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa thuộc tính** |
| 1 | pH | Độ pH của sữa với miền giá trị [3, 9.5] |
| 2 | nhiet\_do | Nhiệt độ của sữa với miền giá trị [34, 90] |
| 3 | huong\_vi | Hương vị của sữa với miền giá trị [0, 1] |
| 4 | mui | Mùi của sữa với miền giá trị [0, 1] |
| 5 | chat\_beo | Chất béo trong sữa với miền giá trị [0, 1] |
| 6 | do\_trong | Độ trong của sữa với miền giá trị [0, 1] |
| 7 | mau | Màu của sữa với miền giá trị [240, 255] |
| 8 | chat\_luong | Phân loại sữa   * Low: Thấp * High: Cao |

# Tiền xử lý dữ liệu

## 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
* Thực hiện

+ Xử lý trên excel

* Xử lý nhất quán dữ liệu trên file và lưu file thành file csv để làm sạch dữ liệu trên weka

+ Xử lý dữ liệu còn thiếu trên weka với mô hình phân lớp

* Các dữ liệu bị thiếu sẽ được thay thế bằng giá trị trung bình của thuộc tính.
* Dùng bộ lọc ReplaceMissingValue để thực hiện thao tác này.

+ Xử lý dữ liệu còn thiếu trên weka với mô hình phân cụm

* Sau đó các dữ liệu bị thiếu sẽ được thay thế bằng giá trị trung bình của thuộc tính.
* Dùng bộ lọc ReplaceMissingValue để thực hiện thao tác này.

## 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]

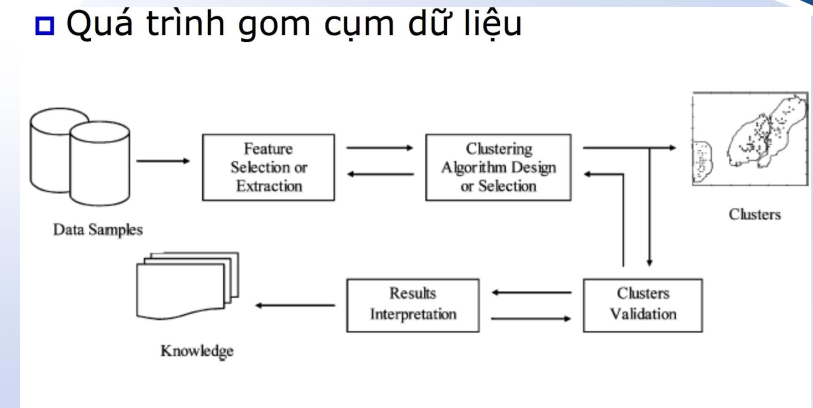
+ Tiến hành chuẩn hóa các thuộc tính về đoạn [0, 1] bằng bộ lọc Normalize, lưu file lại dưới dạng csv.

# 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 giái trị tối thiểu.

### Các bước thực hiện:

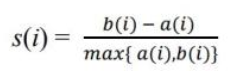
* Bước 1: Khởi tạo tâm cụm:
  + chọn k đối tượng làm trọng tâm ban đầu của k cụm (chọn ngẫu nhiên, theo kinh nghiệm).
* Bước 2: Tính toán khoảng cách và gán cụm:
  + với mỗi điểm tính khoảng cách đến tâm cụm sau đó tìm trọng tâm gần nhất với mỗi điểm (tức là khoảng cách của mỗi điểm tới các tâm cụm, 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

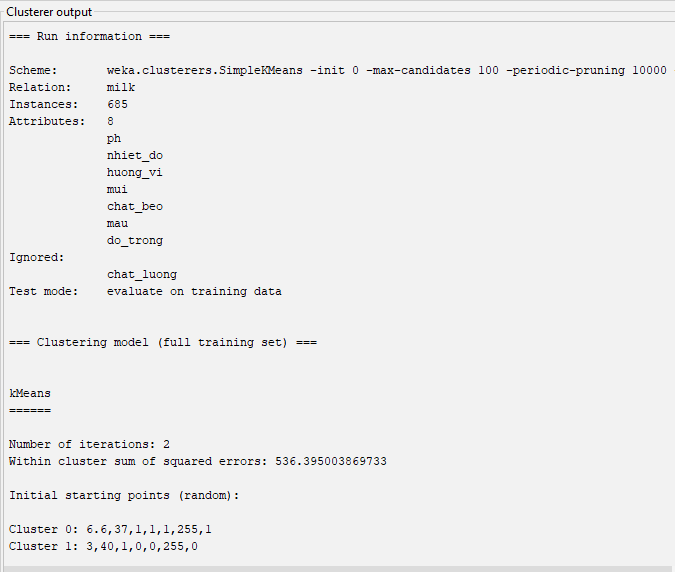
Bước 1: Cho Weka đọc dữ liệu phân cụm đã chuẩn hoá ở trên.

Bước 2: Chon Cluster và chọn SimpleKMeans.

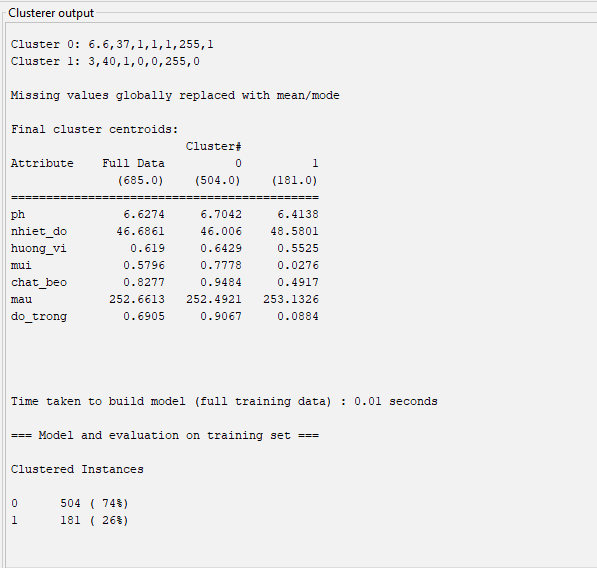
Bước 3: Chọn numClustes = 2

Bước 4: Tiến hành chạy thuật toán phân cụm.

* Kết quả chạy thu được

.  


Hình 3.2 Kết quả chạy phân cụm bằng weka

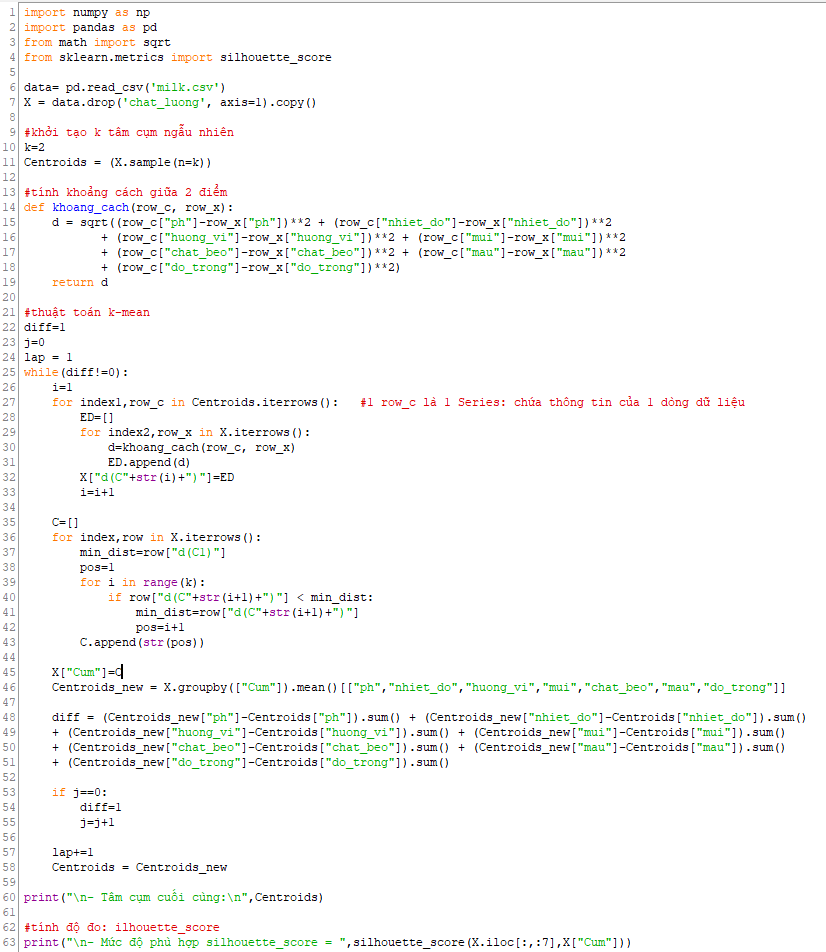


Hình 3. Kết quả chạy phân cụm bằng weka

* 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 3.
  + Tổng các lỗi bình phương trong cụm (WCSS) là thước đo mức độ gần của các điểm dữ liệu trong mỗi cụm với tâm của cụm đó. Giá trị WCSS được báo cáo là 536.40. 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 của 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 "huong\_vi" có giá trị trung bình 0.6429.
* 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 504 mẫu (74% tập dữ liệu), Cụm 1 chứa 181 mẫu (26% tập dữ liệu).

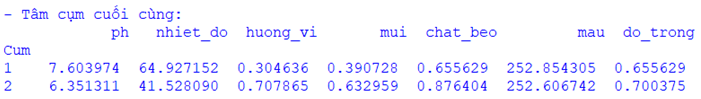
## Chạy phân cụm với python

* Trương trình thực hiện thuật toán k-mean:



Hình . Chương trình chạy thuật toán k-mean trên python

* Kết quả chạy thu được:



Hình .5 Kết quả tọa độ tâm cụm



Hình .6 Kết quả độ đo silhouette

## 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.

# Khai phá dữ liệu bằng mô hình phân lớp

## Giới thiệu về phân lớp

### Khái niệm

* Phân lớp dữ liệu là xếp đối tượng dữ liệu vào trong các lớp đã được xác định trước
* Phân lớp gồm 2 bước:

+ Bước 1: Xây dựng mô hình

* Mô tả tập các lớp xác định trước.
* Tập học/ huấn luyện: các mẫu dành cho xây dựng mô hình.
* Mỗi mẫu thuộc về 1 lớp đã định nghĩa trước.
* Tìm luật phân lớp, cây quyết định hoặc công thức toán mô tả lớp.

+ Bước 2: Vận hành mô hình

* Phân lớp các đối tượng chưa biết.
* Xác định độ chính xác của mô hình, sử dụng tập dữ liệu kiểm tra độc lập.
* Độ chính xác chấp nhận được => áp dụng mô hình để phân lớp các mẫu chưa xác định được nhãn lớp.

### Các phương pháp phân lớp

* Phân lớp bằng cây quyết định
* Phân lớp bằng Bayesian
* Phân lớp bằng KNN

## Phân lớp bằng cây quyết định

### Khái niệm

* Cây quyết định là một flow-chart giống cấu trúc cây, nút bên trong biểu thị một kiểm tra trên một thuộc tính, nhánh viểu diễn đầu ra của kiểm tra, nút lá biểu diễn nhãn lớp hoặc sự phân bố của lớp.
* Cây được tạo theo hai giai đoạn: tạo cây và tỉa nhánh.
* Giai đoạn tạo cây:

Bắt đầu tất cả các mẫu học đều nằm ở nút gốc, Sau đó các mẫu học được phân chia một cách đệ quy dựa trên thuộc tính được chọn.

* Bước tỉa nhánh: tìm và xóa những nhánh có phẩn tử không thể xếp vào lớp nào cả.
* Bước vận hành: kiểm tra những giá trị thuộc tính của mẫu đối với các giá trị trên nhánh của cây.

### Thuật toán tạo cây quyết định

* Bước 1: Cây được xây dựng đệ quy từ trên xuống và theo cách chia để trị.
* Bước 2: ban đầu tất cả mẫu học đều nằm ở gốc.
* Bước 3: Thuộc tính được phân loại (nếu là giá trị liên tục thì được rời rạc hóa)
* Bước 4: Các mẫu học được phân chia đệ quy dựa trên thuộc tính chọn lựa.
* Bước 5: Kiểm tra những thuộc tính được chọn dựa trên kinh nghiệm hoặc của một tiêu chuẩn thống kê.

### Điều kiện dừng phân chia tập học

* Tất cả những mẫu học đối với một nút cho trước đều cùng lớp.
* Không còn thuộc tính nào để phân chia tiếp.
* Không còn mẫu học

### Độ lợi thông tin trong cây quết định

Cho hai lớp P (Positive) và N (Negative), tập học S. Lớp P có p phần tử và lớp N có n phần tử. Khối lượng thông tin cần để quyết định các mẫu trong S thuộc về lớp P hay lớp N được xác định bởi:

G/S thuộc tính A được chọn để phân hoạch S thành các tập hợp {S1,S2,…,Sv}. Nếu Si chứa pi mẫu của lớp P và ni mẫu của lớp N thì entropy cần để phân loại các đối tượng trong cây con Si là:

Độ lợi thông tin của nhánh A là:

Gain(A) = I(p,n) – E(A)

## Thuật toán phân lớp ID3

* Mục đích của phương pháp: Tạo ra cây quyết định bằng cách sử dụng hàm “entropy” làm thước đo.
* Ý tưởng:
* Tạo cây quyết định bằng việc sử dụng cách tìm kiếm từ trên xuống trong tập học.
* Sử dụng độ lợi thông tin để chọn thuộc tính có khả năng phân loại tốt nhất.
* Cách thực hiện:
  + ID3(S,D,A)
  + Input: Tập học S, thuộc tính quyết định D, tập thuộc tính A.
  + Output: Nút gốc của cây quyết định.

Begin

* + - Tạo ‘Nút gốc’ cho cây quyết định.
    - If tất cả mẫu học của S đều có trị của D là P, trả về cây có 1 nút duy nhất là Nút\_gốc với nhãn P.
    - If tất cả mẫu học của S đều có trị của D là N, trả về cây có 1 nút duy nhất là “Nút\_gốc” với nhãn N
    - If A là rỗng, trả về cây có nút duy nhất là Nút\_gốc với nhãn là trị phổ biến nhất của D trong tập mẫu.
    - Else Begin:
      * Gọi X là thuộc tính của A phân lớp S tốt nhất // tính độ lợi
      * Gán X vào thuộc tính quyết định D của Nút\_gốc
      * For each trị v của X:
        + Thêm 1 nhánh cây mới dưới Nút\_gốc ứng với X=v
        + Gọi Sv là tập con của v trị của X là v
        + If Sv rỗng:

Thêm dưới nhánh mới này, 1 nút lá có nhãn là trị phổ biến nhất của thuộc tính quyết định trong S.

* + - * + Else:

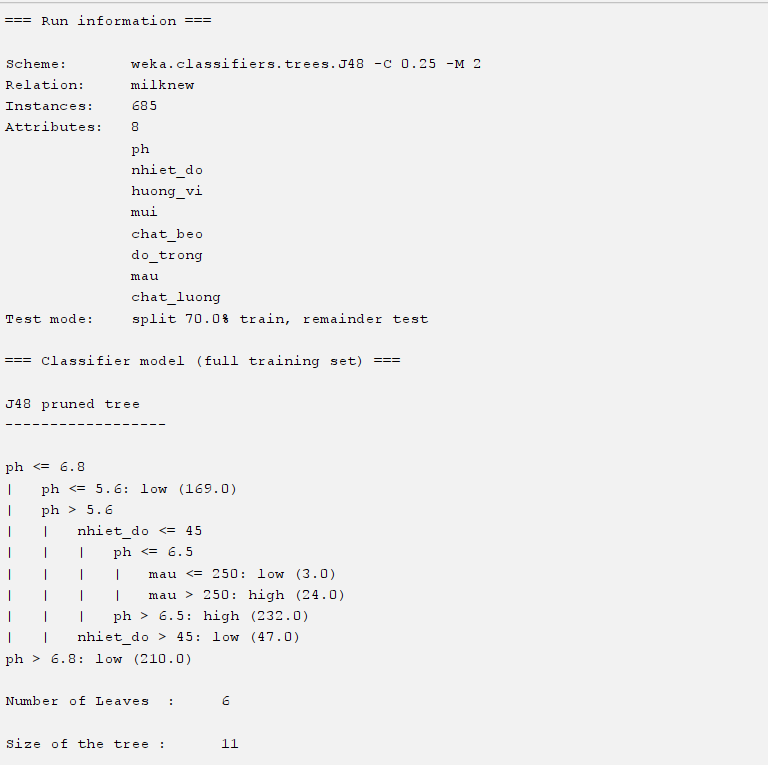
Thêm cây con vào dưới nhánh mới này.

ID3(Sv, D, A-{X})

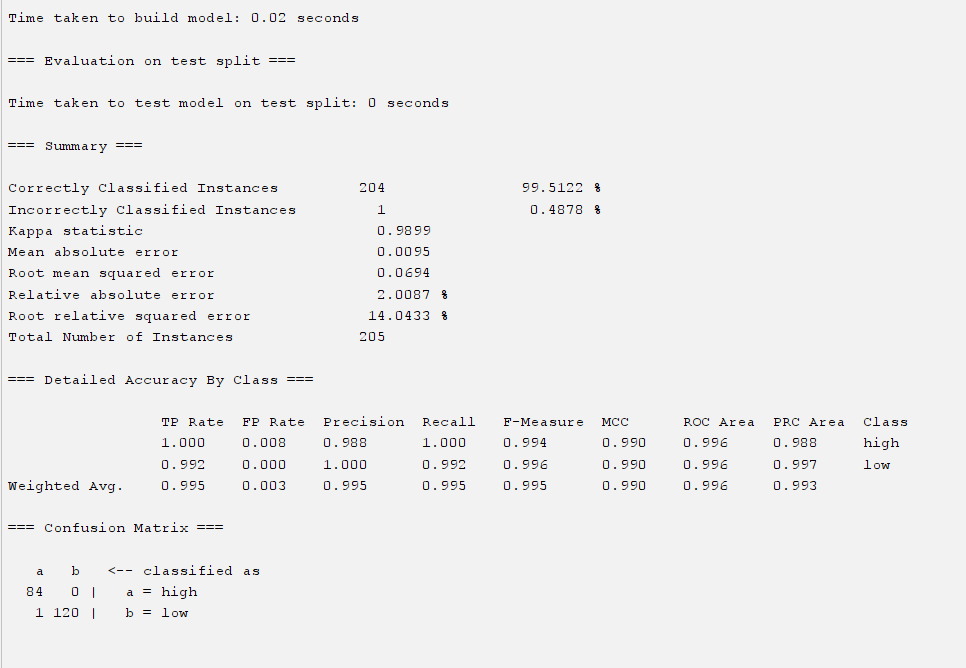
* + - * End
    - Return Nút\_gốc

### Chạy phân lớp với weka

* Bước 1: Cho Weka đọc dữ liệu hồi quy đã chuẩn hoá ở trên.
* Bước 2: Chon Classify và chọn J48.
* Bước 3: Chọn Percentage split
* Chia tệp dữ liệu ra 70% dữ liệu là dùng để train và 30% dữ liệu dùng để test.
* Kết quả chạy thu được

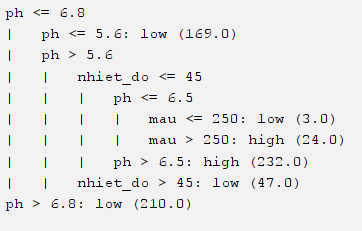


Hình 4. Kết quả chạy phân lớp bằng weka



Hình 4..Kết quả chạy phân lớp bằng weka

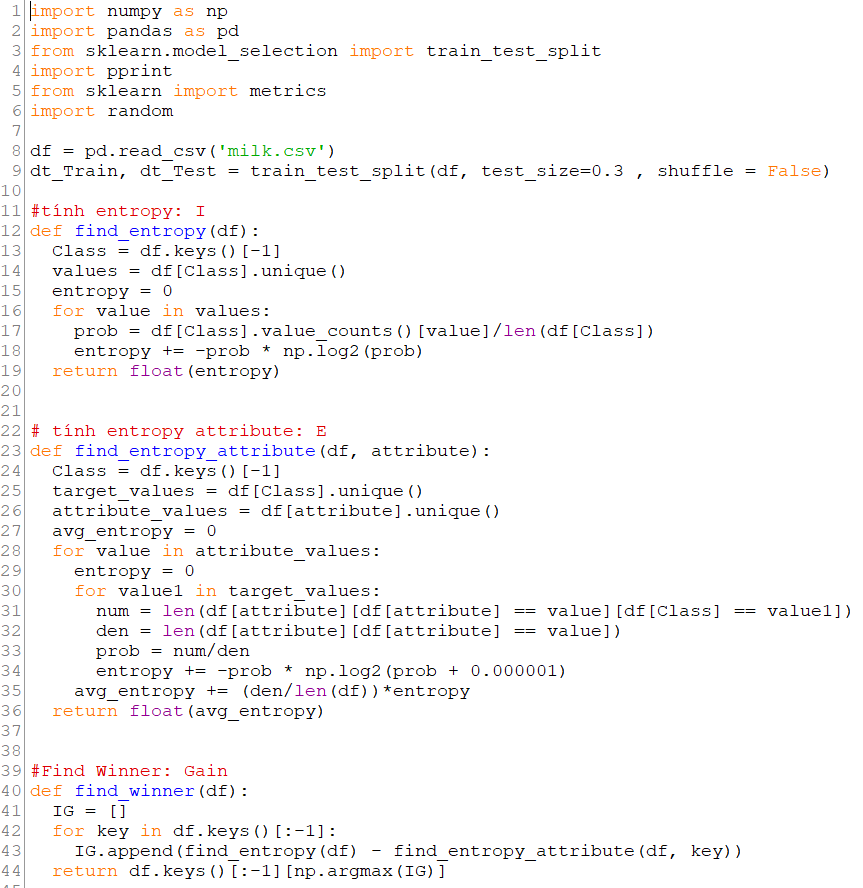
* Giải thích kết quả
* Có tất cả 204 dữ liệu được dự đoán đúng (99.5122%) và 1 dữ liệu bị dự đoán sai (0.4878%)
* Mô hình cây quyết định là:



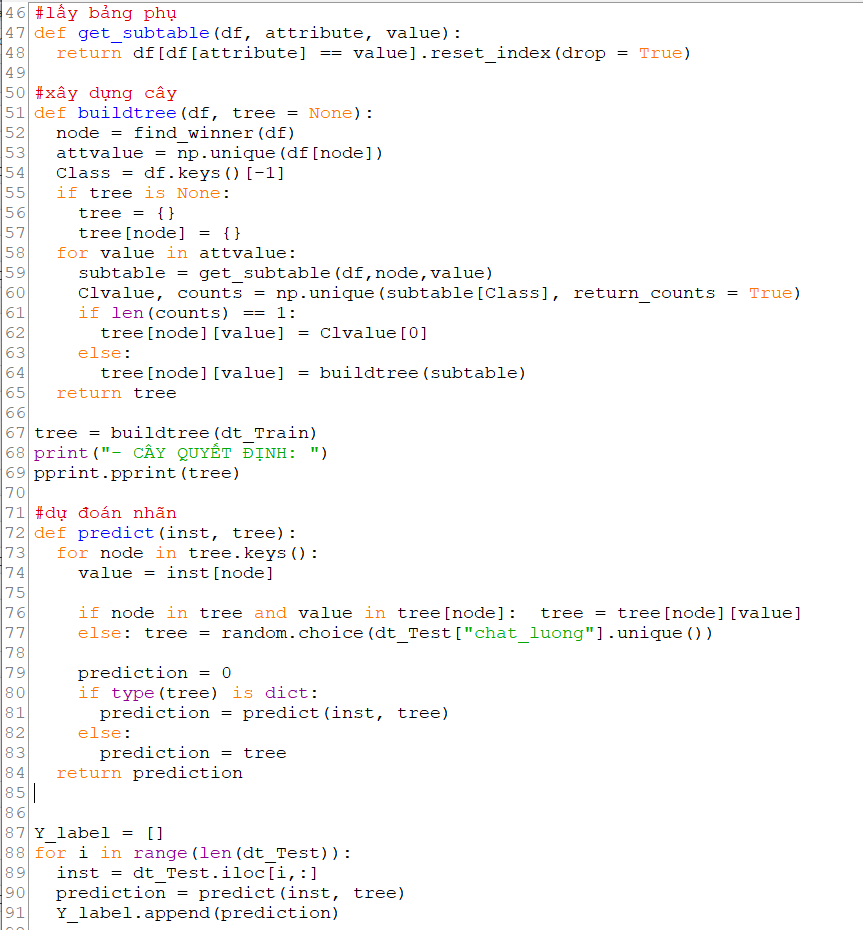
* Number of Leaves: Số lá là 6
* Size of the tree: kích thước cây là 11

### Chạy phân lớp với python

1. Trương trình chạy thuật toán id3.

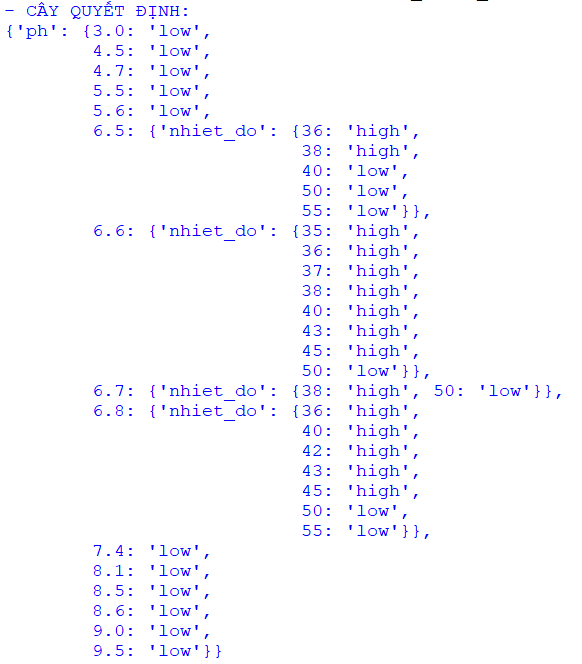


Hình 4..2.1 .Chương trình chạy thuật toán id3 trên python

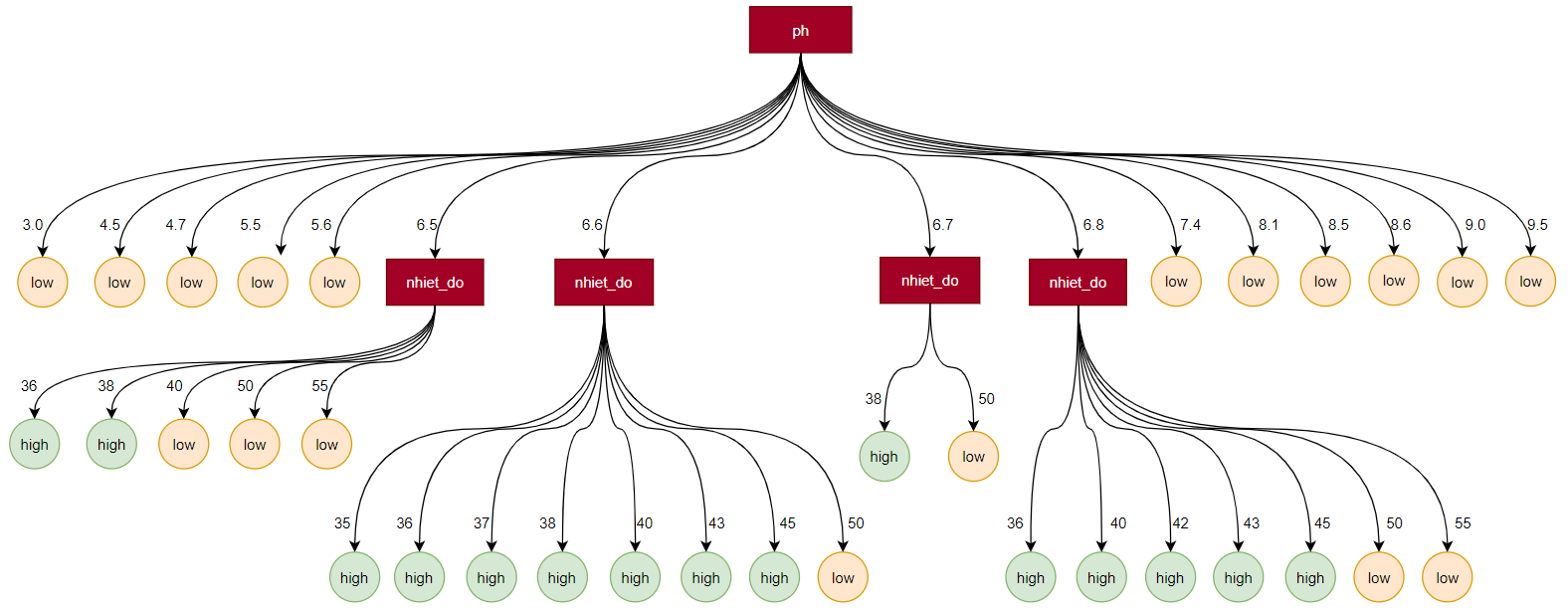


Hình ..2.2 .Chương trình chạy thuật toán id3 trên python

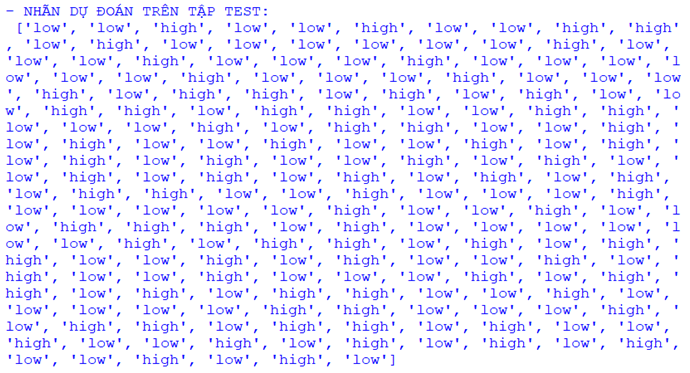
1. Kết quả chạy chương trình.



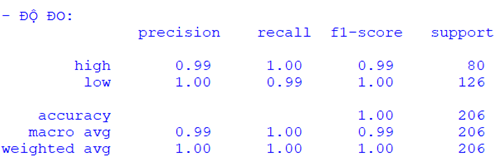
Hình 4.3.2.3 Kết quả cây quyết định lưu dưới dạng từ điển



Hình .3.2.4 Kết quả cây quyết định được triển khai lại dưới dạng cây.



Hình .3.2.5 Kết quả nhãn dự đoán trên tập test



Hình .3.2.6 Kết quả đánh giá trên độ đo precision, recall, f1-score

## Ý nghĩa của thuật toán ID3

* ID3 là một giải thuật hay vì cách biểu diễn tri thức học được của nó, tiếp cận của nó trong việc quản lý tính phức tạp, heuristic của nó dùng cho việc chọn lựa các khái niệm ứng viên, và tiềm năng của nó đối với việc xử lý dữ liệu nhiễu. ID3 biểu diễn các khái niệm (concept) ở dạng các cây quyết định (decision tree).

# Kết luận

* Phân lớp và phân cụm là 2 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 lớp, bài toán phân cụm từ đó xây dựng được mô hình phân lớp và mô hình phân cụm để đánh giá chất lượng sữa

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

* Nhưng vẫn có 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 vào để đánh giá chất lượng sữa chứ chưa hỗ trợ để đánh giá những phần khác

* Trong quá trình thự 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/datasets/yrohit199/milk-quality>