TRƯỜNG ĐẠI HỌC THỦY LỢI

KHOA CÔNG NGHỆ THÔNG TIN

----- -----

ĐỀ TÀI:

KHAI PHÁ DỮ LIỆU NHẬN BIẾT TỈ LỆ MẮC UNG THƯ PHỔI

Nhóm sinh viên thực hiện:

1. Phan Thanh Toàn – 185P1063393 – 60TH
2. Nguyễn Hồng Sơn – 185P1063376 – 60TH
3. Nguyễn Ngọc Anh – 185P1063479 – 60TH

Giảng viên hướng dẫn: Giảng viên Trần Mạnh Tuấn

Hà Nội, 03/2022

LỜI MỞ ĐẦU

Trong những năm gần đây cùng với phát triển nhanh chóng của khoa học kỹ thuật là sự bùng nổ về tri thức. Kho dữ liệu, nguồn tri thức của nhân loại cũng trở nên đồ sộ, vô tận làm cho vấn đề khai thác các nguồn tri thức đó ngày càng trở nên nóng bỏng và đặt ra thách thức lớn cho nền công nghệ thông tin thế giới.

Nhu cầu về tìm kiếm và xử lý thông tin, cùng với yêu cầu về khả năng kịp thời khai thác chúng để mạng lại những năng suất và chất lượng cho công tác quản lý, hoạt động kinh doanh... đã trở nên cấp thiết trong xã hội hiện đại. Để đáp ứng phần nào yêu cầu này, người ta đã xây dựng các công cụ tìm kiếm và xử lý thông tin nhằm giúp cho người dùng tìm kiếm được các thông tin cần thiết cho mình.

Vì vậy nhóm em chọn đề tài: "***Khai phá dữ liệu nhận biết tỉ lệ mắc ung thư phổi"***, để làm báo cáo môn học của mình

Báo cáo gồm 5 chương:

Chương 1: Tổng quan về khai phá dữ liệu.

Chương 2: Tiền xử lý dữ liệu.

Chương 3: Khai phá dữ liệu bằng thuật toán K-means clustering

Chương 4: Khai phá dữ liệu bằng thuật toán phân lớp Random Forest

Chương 5: Kết luận và hướng phát triển

BẢNG PHÂN CHIA CÔNG VIỆC

Stt	Họ và tên	Công việc
1	Phan Thanh Toàn	Thuật toán Random Forest
2	Nguyễn Hồng Sơn	Thuật toán K-Means Clustering
3	Nguyễn Ngọc Anh	Tiền xử lý dữ liệu

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1 Phát hiện tri thức và khai phá dữ liệu.

Phát hiện tri thức (*Knowledge Discovery*) trong các cơ sở dữ liệu là một quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được.

Khai phá dữ liệu (Data mining) được định nghĩa như sau: "*Data mining là một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn*".

Khai phá dữ liệu có thể được sử dụng cho các lĩnh vực y tế, phân tích thị trường, xây dựng ... có thể được xem như là kết quả của sự tiến triển tự nhiên của công nghệ thông tin.

1.2 Quy trình khám phá tri thức trong CSDL

Quá trình phát hiện tri thức bao gồm các bước:

- **Làm sạch dữ liệu:** Các nhiễu và dữ liệu không nhất quán sẽ được loại bỏ.
- **Tích hợp dữ liệu:** Dữ liệu từ nhiều nguồn khác có thể được tổ hợp lại.
- **Lựa chọn dữ liệu:** Những dữ liệu thích hợp với nhiệm vụ phân tích sẽ được trích rút ra từ CSDL.
- **Chuyển đổi dữ liệu:** 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:** Quá trình cốt lõi, tất yếu trong đó các phương pháp thông minh sẽ được áp dụng nhằm trích rút ra các mẫu dữ liệu.
- **Đánh giá mẫu:** Các nhà phân tích dữ liệu sẽ dựa trên một số độ đo nào đó để xác định lợi ích thực sự, độ quan trọng của các mẫu biểu diễn tri thức.
- **Biểu diễn tri thức:** 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 Mô tả bài toán chuẩn đoán tỉ lệ.

1.3.1 Tổng quan bài toán.

Dataset gồm các mô tả về các thuộc tính tương ứng với chẩn đoán ung thư phổi. Áp dụng các thuật toán để xác định xem đối tượng mắc ung thư phổi: YES, NO

1.3.2 Phân tích dữ liệu thô.

Nguồn dữ liệu thô:

https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer?fbclid=IwAR0Ap29_cgYPW-aaoEuL4-QXDpYSHBbhP7fzpzPlJlCqwsBffg1wdHDiTRk

+ *Hiểu dữ liệu:* Dữ liệu dự đoán mắc ung thư phổi dựa trên giá trị các dữ liệu thuộc tính.

+ *Dữ liệu gồm:* Dữ liệu bao gồm 309 bản ghi cùng 16 thuộc tính chẩn đoán ung thư phổi.

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG CANCER
0	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
0		2	1	1	1	2	2	2	1	1	1	2	2	2	YES
1	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
0	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
1	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
1	75	1	2	1	1	2	2	2	2	1	2	2	1	1	YES
0	52	2	1	1	1	1	2	1	2	2	2	2	1	2	YES
1	51	2	2	2	2	1	2	2	1	1	1	2	2	1	YES
1		2	1	2	1	1	2	1	1	1	1	1	1	1	NO
0	53	2	2	2	2	2	1	2	1	2	1	1	2	2	YES
1	61	2	2	2	2	2	2	1	2	1	2	2	2	1	YES
0	72	1	1	1	1	2	2	2	2	2	2	2	1	2	YES
1	60	2	1	1	1	1	2	1	1	1	1	2	1	1	NO

Hình 1.2 Dữ liệu ban đầu.

Hiểu các thuộc tính:

STT	Thuộc tính	Ý nghĩa thuộc tính
1	GENDER	Giới tính của đối tượng.
2	AGE	Tuổi của đối tượng.
3	SMOKING	Thói quen hút thuốc của đối tượng.
4	YELLOW_FINGERS	Màu sắc ngón tay của đối tượng.
5	ANXIETY	Trạng thái lo âu của đối tượng.
6	PEER PRESSURE	Áp lực cuộc sống của đối tượng.
7	CHRONIC DISEASE	Tình trạng mãn tính của đối tượng.
8	FATIGUE	Sự mệt mỏi của đối tượng.
9	ALLERGY	Dị ứng của đối tượng.
10	WHEEZING	Hơi thở khò khè của của đối tượng.
11	ALCOHOL CONSUMING	Mức độ tiêu thụ rượu của đối tượng.
12	COUGHING	Tần suất ho của đối tượng.
13	SHORTNESS OF BREATH	Tần suất thở gấp của đối tượng.
14	SWALLOWING DIFFICULTY	Sự khó khăn khi nuốt của đối tượng.
15	CHEST PAIN	Sự tức ngực của đối tượng.
16	LUNG _ CANCER	Phân loại xem đối tượng mắc ung thư hoặc không. (YES hoặc NO)

CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU

2.1 Làm sạch dữ liệu.

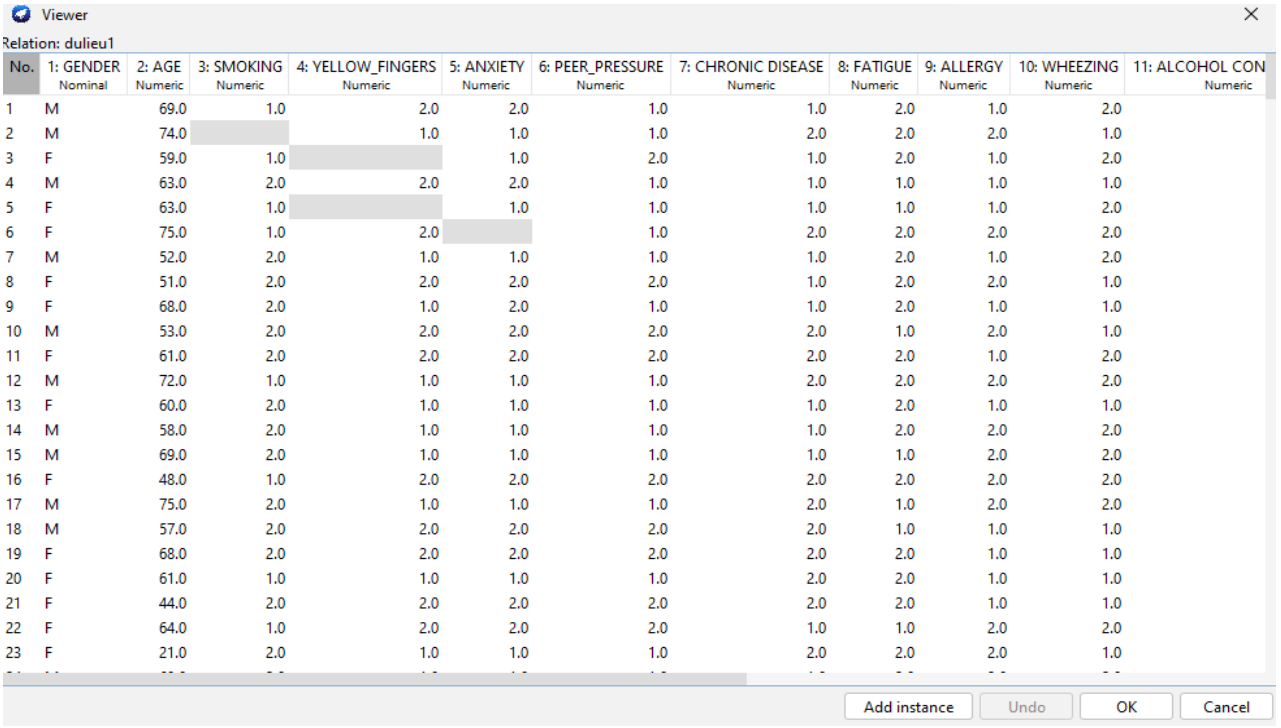
Là quá trình nhận dạng dữ liệu đã có để tiến hành xử lý các dữ liệu bị thiếu (missing data) xử lý dữ liệu bị nhiễu (noisy data) và không nhất quán.

(1) Xử lý dữ liệu bị thiếu (missing data)

(2) Xử lý dữ liệu liệu, không nhất quán (inconsistent data).

Thực hiện:

- *Thực hiện:* Xử lý dữ liệu bị thiếu:
- Kiểm tra những thuộc tính bị thiếu: + Smoking: Bị thiếu 1 dữ liệu
  + Yellow_Fingers: Bị thiếu 2 dữ liệu
  + Anxiety: Bị thiếu 1 dữ liệu



Viewer

Relation: dulieu1

No.	1: GENDER Nominal	2: AGE Numeric	3: SMOKING Numeric	4: YELLOW_FINGERS Numeric	5: ANXIETY Numeric	6: PEER_PRESSURE Numeric	7: CHRONIC DISEASE Numeric	8: FATIGUE Numeric	9: ALLERGY Numeric	10: WHEEZING Numeric	11: ALCOHOL CON Numeric
1	M	69.0	1.0	2.0	2.0	1.0	1.0	2.0	1.0	2.0	
2	M	74.0		1.0	1.0	1.0	2.0	2.0	2.0	1.0	
3	F	59.0	1.0		1.0	2.0	1.0	2.0	1.0	2.0	
4	M	63.0	2.0	2.0	2.0	1.0	1.0	1.0	1.0	1.0	
5	F	63.0	1.0		1.0	1.0	1.0	1.0	1.0	2.0	
6	F	75.0	1.0	2.0		1.0	2.0	2.0	2.0	2.0	
7	M	52.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	2.0	
8	F	51.0	2.0	2.0	2.0	2.0	1.0	2.0	2.0	1.0	
9	F	68.0	2.0	1.0	2.0	1.0	1.0	2.0	1.0	1.0	
10	M	53.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0	1.0	
11	F	61.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0	
12	M	72.0	1.0	1.0	1.0	1.0	2.0	2.0	2.0	2.0	
13	F	60.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	
14	M	58.0	2.0	1.0	1.0	1.0	1.0	2.0	2.0	2.0	
15	M	69.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0	
16	F	48.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	
17	M	75.0	2.0	1.0	1.0	1.0	2.0	1.0	2.0	2.0	
18	M	57.0	2.0	2.0	2.0	2.0	2.0	1.0	1.0	1.0	
19	F	68.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	1.0	
20	F	61.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	1.0	
21	F	44.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	1.0	
22	F	64.0	1.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0	
23	F	21.0	2.0	1.0	1.0	1.0	2.0	2.0	2.0	1.0	

Add instance | Undo | OK | Cancel

Dữ liệu gốc

Thay thế các giá trị bị thiếu bằng giá trị trung bình:

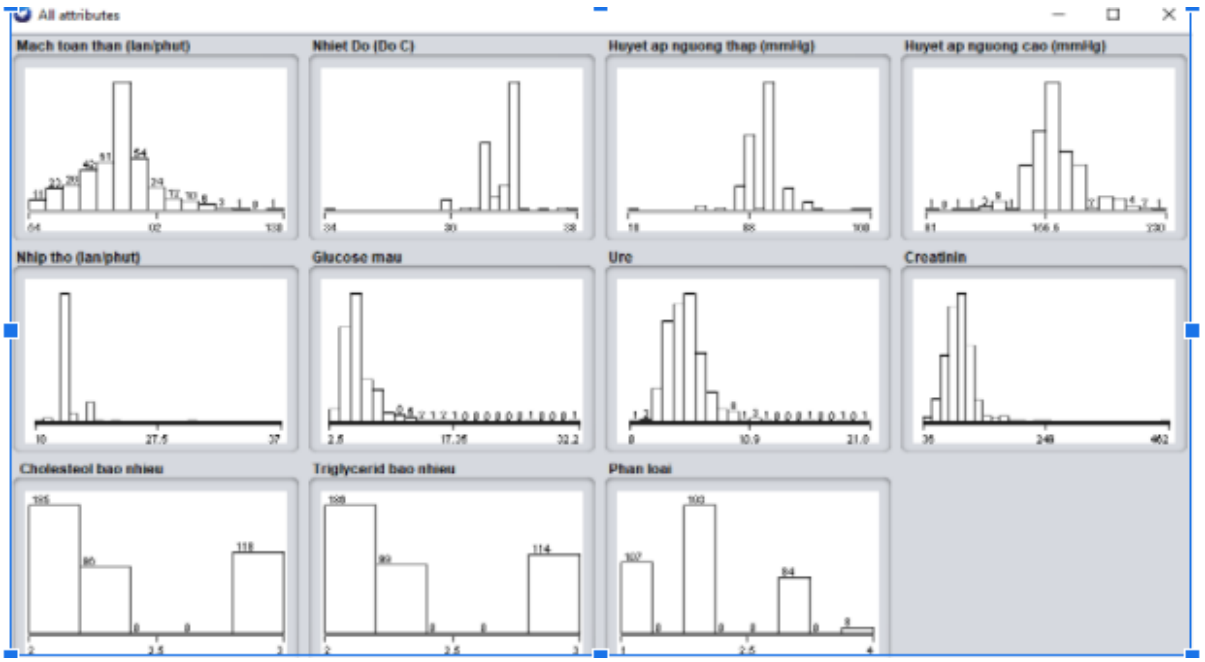
Viewer

Relation: dulieu1-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: GENDER Nominal	2: AGE Numeric	3: SMOKING Numeric	4: YELLOW_FINGERS Numeric	5: ANXIETY Numeric	6: I
1	M	69.0	1.0	2.0	2.0	
2	M	74.0	1.56168831...	1.0	1.0	
3	F	59.0	1.0	1.5700325732899023	1.0	
4	M	63.0	2.0	2.0	2.0	
5	F	63.0	1.0	1.5700325732899023	1.0	
6	F	75.0	1.0	2.0	1.5	
7	M	52.0	2.0	1.0	1.0	
8	F	51.0	2.0	2.0	2.0	
9	F	68.0	2.0	1.0	2.0	

Rig

- Smoking: 1.5616883116883118
- Yellow_Fingers: 1.5700325732899023
- Anxiety: 1.5



Phân bổ dữ liệu

2.2 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.

(1) Tích hợp lược đồ và so trùng đối tượng.

(2) Vấn đề dư thừa.

(3) Phát hiện và xử lý mâu thuẫn giá trị dữ liệu.

=> Dữ liệu lấy từ một nguồn nên không cần thực hiện quá trình này.

Nguồn dữ liệu:
https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer?fbclid=IwAR0Ap29_cgYPW-aaoEuL4-QXDpYSHBbhP7fzpzPlJlCqwsBffg1wdHDiTRk

2.3 Rời rạc hóa dữ liệu

Dữ liệu trước khi rời rạc

```
C:\Users\toan\AppData\Local\Programs\Python\Python39\python.exe "D:/hoc/khai_pha_du_lieu/Khai phá DL - Nhóm 4 - 60TH/btl.py"
dulieugoc
    GENDER  AGE  SMOKING  ...  SWALLOWING DIFFICULTY  CHEST PAIN  LUNG_CANCER
0        M   69      1.0  ...                      2           2          YES
1        M   74      NaN  ...                      2           2          YES
2        F   59      1.0  ...                      1           2           NO
3        M   63      2.0  ...                      2           2           NO
4        F   63      1.0  ...                      1           1           NO
..     ...  ...      ...  ...                    ...         ...          ...
304      F   56      1.0  ...                      2           1          YES
305      M   70      2.0  ...                      1           2          YES
306      M   58      2.0  ...                      1           2          YES
307      M   67      2.0  ...                      1           2          YES
308      M   62      1.0  ...                      2           1          YES
```

Kết quả sau khi khi rời rạc

```
C:\Users\toan\AppData\Local\Programs\Python\Python39\python.exe "D:/hoc/khai_pha_du_lieu/Khai phá DL - Nhóm 4 - 60TH/btl.py"
    GENDER AGE  SMOKING  ...  SWALLOWING DIFFICULTY  CHEST PAIN  LUNG_CANCER
0        M   2      1.0  ...                      2           2          YES
1        M   2      NaN  ...                      2           2          YES
2        F   1      1.0  ...                      1           2           NO
3        M   1      2.0  ...                      2           2           NO
4        F   1      1.0  ...                      1           1           NO
..     ...  ..      ...  ...                    ...         ...          ...
304      F   1      1.0  ...                      2           1          YES
305      M   2      2.0  ...                      1           2          YES
306      M   1      2.0  ...                      1           2          YES
307      M   2      2.0  ...                      1           2          YES
308      M   1      1.0  ...                      2           1          YES
```

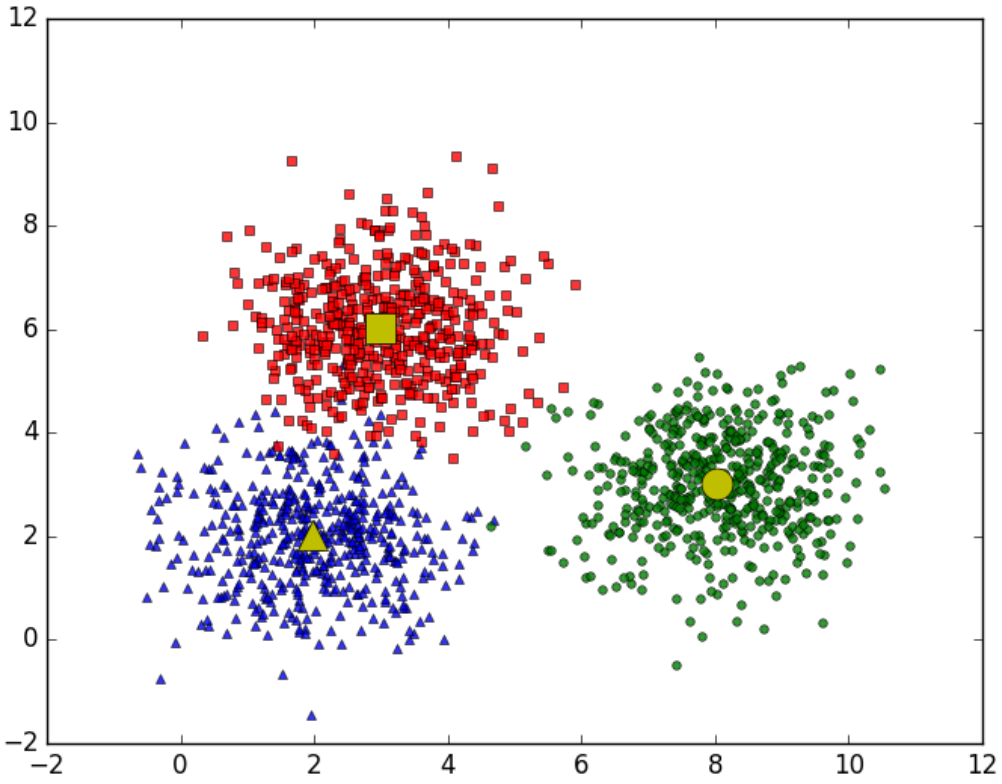
CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN PHÂN CỤM K-MEANS CLUSTERING

3.1 Giới thiệu về kỹ thuật K-means clustering.

- **Khái niệm:**

+ Trong thuật toán K-means clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thế nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

+ Giả sử mỗi cluster có một điểm đại diện (center) màu vàng. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó.



Bài toán với 3 clusters.

- **Mô hình và phương trình:**

Thuật toán tối ưu hàm mất mát

Đây là một bài toán khó tìm điểm tối ưu vì nó có thêm các điều kiện ràng buộc. Bài toán này thuộc loại mix - integer programming (điều kiện biến là số nguyên) - là loại rất khó tìm nghiệm tối ưu toàn cục (global optimal point, tức nghiệm làm cho hàm mất mát đạt giá trị nhỏ nhất có thể). Tuy nhiên, trong một số trường hợp chúng ta vẫn có thể tìm được phương pháp để tìm được nghiệm gần đúng hoặc điểm cực tiểu. (Nếu chúng ta vẫn nhớ chương trình toán ôn thi đại học thì điểm cực tiểu chưa chắc đã phải là điểm làm cho hàm số đạt giá trị nhỏ nhất).

Giả sử đã tìm được các centers, hãy tìm các label vector để hàm mất mát đạt giá trị nhỏ nhất. Điều này tương đương với việc tìm cluster cho mỗi điểm dữ liệu.

Khi các centers là cố định, bài toán tìm label vector cho toàn bộ dữ liệu có thể được chia nhỏ thành bài toán tìm label vector cho từng điểm dữ liệu xi như sau:

$$\mathbf{y}_i = \arg\min_{\mathbf{y}_i} \sum_{j=1}^{K} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2 \quad (3)$$

$$\text{subject to: } y_{ij} \in \{0, 1\} \ \forall j; \quad \sum_{j=1}^{K} y_{ij} = 1$$

$$j = \arg\min_{j} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2$$

Vì $\|\mathbf{x}_i - \mathbf{m}_j\|_2^2$ chính là bình phương khoảng cách tính từ điểm xi tới center mj ta có thể kết luận rằng mỗi điểm xi thuộc vào cluster có center gần nó nhất! Từ đó ta có thể dễ dàng suy ra label vector của từng điểm dữ liệu.

Giả sử đã tìm được cluster cho từng điểm, hãy tìm center mới cho mỗi cluster để hàm mất mát đạt giá trị nhỏ nhất.

Một khi chúng ta đã xác định được label vector cho từng điểm dữ liệu, bài toán tìm center cho mỗi cluster được rút gọn thành:

$$\mathbf{m}_j = \arg\min_{\mathbf{m}_j} \sum_{i=1}^{N} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2.$$

Tới đây, ta có thể tìm nghiệm bằng phương pháp giải đạo hàm bằng 0, vì hàm cần tối ưu là một hàm liên tục và có đạo hàm xác định tại mọi điểm. Và quan trọng hơn, hàm này là hàm convex (lồi) theo **mj** nên chúng ta sẽ tìm được giá trị nhỏ nhất và điểm tối ưu tương ứng.

Đặt l(**mj**) là hàm bên trong dấu argmin ta có đạo hàm:

$$\frac{\partial l(\mathbf{m}_j)}{\partial \mathbf{m}_j} = 2 \sum_{i=1}^{N} y_{ij}(\mathbf{m}_j - \mathbf{x}_i)$$

Giải phương trình đạo hàm bằng 0 ta có:

$$\mathbf{m}_j \sum_{i=1}^{N} y_{ij} = \sum_{i=1}^{N} y_{ij}\mathbf{x}_i$$

$$\Rightarrow \mathbf{m}_j = \frac{\sum_{i=1}^{N} y_{ij}\mathbf{x}_i}{\sum_{i=1}^{N} y_{ij}}$$

$\mathbf{m}_j$ là trung bình cộng của các điểm trong cluster j.

3.2 Thuật toán phân cụm K-means clustering

- **Các bước thực hiện:**

- Bước 1: Khởi tạo ngẫu nhiên tâm cụm: chọn ngẫu nhiên k ví dụ trong tập data làm k cụm khởi đầu
- Bước 2: Gán từng ví dụ vào cụm có tâm gần nó nhất. Việc tính khoảng cách từ một điểm tới một tâm cụm có thể tính dựa theo khoảng cách hình học Euclid
- Bước 3: Điều chỉnh tâm cụm: tọa độ của tâm cụm mới bằng tọa độ trung bình của tất cả các ví dụ trong cụm đó
- Bước 4: Kiểm tra điều kiện dừng: nếu thuật toán chưa hội tụ, quay lại bước 2

3.3 Ưu điểm

- K-means là có độ phức tạp tính toán O(tkn).
- K-means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn.

3.4 Nhược điểm

- K-means không khắc phục được nhiễu và giá trị k phải được cho bởi người dùng.
- Chỉ thích hợp áp dụng với dữ liệu có thuộc tính số và khám ra các cụm có dạng hình cầu

3.5 Kết quả phương pháp K- means.

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=2).fit(x_train)
y_pred_kmeans = kmeans.predict(x_test)

print("confusion_matrix Kmeans: ")
print(metrics.confusion_matrix(y_test, y_pred_kmeans))

print(len(x_test[1]))
```

```
confusion_matrix Kmeans:
[[ 17  22]
 [145 125]]
```

- Tỷ lệ dự đoán đúng: 55.05%
- Tỷ lệ dự đoán sai: 44.95%
- Phân tích kết quả:
    + Có 22 trường hợp dự đoán đúng, 17 trường hợp dự đoán sai về tỉ lệ người mắc ung thư phổi.
    + Có 145 trường hợp dự đoán đúng, 125 trường hợp dự đoán sai về tỉ lệ người không mắc ung thư phổi.

CHƯƠNG 4: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN PHÂN LỚP RANDOM FOREST

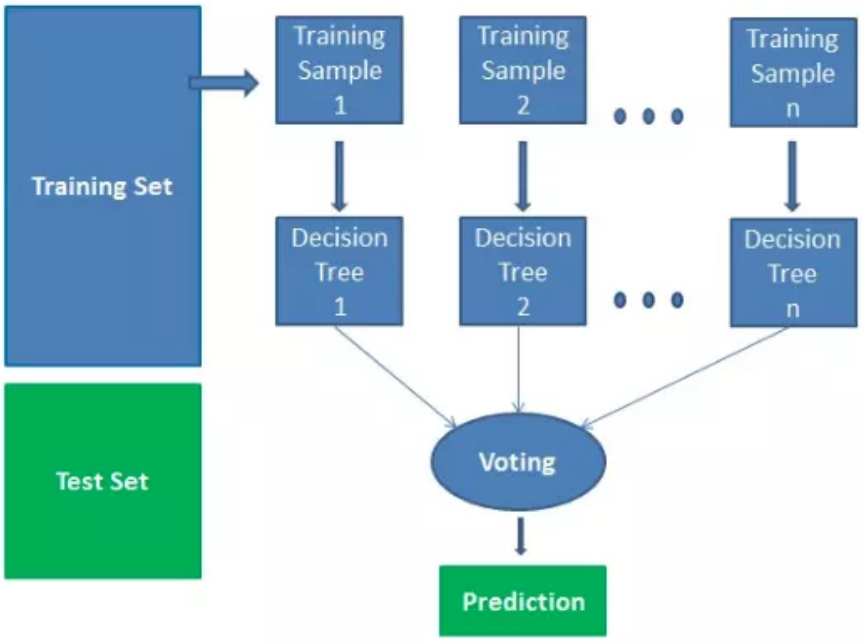
4.1 Khái niệm

Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest mình sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

4.2 Cách thức hoạt động

Nó hoạt động theo bốn bước:

- ❖ Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
- ❖ Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
- ❖ Hãy bỏ phiếu cho mỗi kết quả dự đoán.
- ❖ Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.



4.3 Ưu điểm

Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này. Nó không bị vấn đề overfitting. Có thể nhận được tầm quan trọng của tính năng tương đối, giúp chọn các tính năng đóng góp nhiều nhất cho trình phân loại.

4.4 Nhược điểm

Random forests chậm tạo dự đoán bởi vì nó có nhiều cây quyết định. Bất cứ khi nào nó đưa ra dự đoán, tất cả các cây trong rừng phải đưa ra dự đoán cho cùng một đầu vào cho trước và sau đó thực hiện bỏ phiếu trên đó. Toàn bộ quá trình này tốn thời gian. Mô hình khó hiểu hơn so với cây quyết định, nơi bạn có thể dễ dàng đưa ra quyết định bằng cách đi theo đường dẫn trong cây.

4.5 Kết quả phương pháp học máy Random Forest.

```
from sklearn.ensemble import RandomForestClassifier

clf = RandomForestClassifier(max_depth=2)
clf.fit(x_train, y_train)
kq1 = clf.predict(x_test)
print("Accuracy_Random:", clf.score(x_test, y_test))
print("confusion_matrix: ")
print(metrics.confusion_matrix(y_test, kq1))
```

```
Accuracy_Random: 0.8387096774193549
confusion_matrix:
[[ 0 15]
 [ 0 78]]
```

- Tỷ lệ dự đoán đúng: 83.87%
- Tỷ lệ dự đoán sai: 16.13%
- Phân tích kết quả:
    + Có 0 trường hợp dự đoán đúng, 15 trường hợp dự đoán sai về tỉ lệ người mắc ung thư phổi.
    + Có 78 trường hợp dự đoán đúng, 0 trường hợp dự đoán sai về tỉ lệ người không mắc ung thư phổi.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận.

Phân cụm và phân lớp là 2 lĩnh vực quan trọng trong khai phá dữ liệu nhằm đưa ra những dự đoán, xu hướng trong tương lai, nó được ứng dụng trong nhiều ngành nghề và nhiều lĩnh vực. Trong quá trình tìm hiểu và hoàn thành bài tập lớn với đề tài "***Khai phá dữ liệu nhận biết tỉ lệ mắc ung thư phổi***". Nhóm 4 đã đạt được một số kết quả như sau:

- Tìm hiểu tổng quan về khai phá dữ liệu, bài toán phân lớp, phương pháp phân lớp, phân cụm, phương pháp phân cụm để từ đó xây dựng mô hình phân lớp, phân cụm.
- Thu thập dữ liệu, tiền xử lý dữ liệu. Xây dựng nên mô hình phân lớp, phân cụm trên python sử dụng các thư viện có sẵn.
- Đánh giá mô hình phân lớp qua thuật toán Random Forest
- Đánh giá mô hình phân cụm qua thuật toán K-means clustering

5.2 Hướng phát triển

Xây dựng, cải tiến mô hình chẩn đoán bệnh với phương pháp học máy khác như SVM, KNN,…

Trong quá trình hoàn thành bài tập lớn, chúng em đã cố gắng tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, thời gian có hạn nên chúng em sẽ không tránh khỏi những thiếu sót, rất mong nhận được sự đóng góp ý kiến của quý thầy cô và các bạn để báo cáo và kỹ năng của chúng em ngày được hoàn thiện hơn và có thế áp dụng được trong thực tiễn.

Tuy nhiên bài tập nhóm vẫn còn một số hạn chế:

Việc thu thập dữ liệu chưa đầy đủ, thông số về bệnh còn hạn chế nên kết quả dự đoán tương đối lệch và vẫn chưa được tốt nhất.

TÀI LIỆU THAM KHẢO

1. Tài liệu tiếng việt.

[1] TS.Trần Mạnh Tuấn, Bài giảng "Khai Phá dữ liệu" .

https://piazza.com/class/kzeswerja0j1ae

[2] https://ongxuanhong.wordpress.com/2015/08/20/tien-xu-ly-du-lieu-horse-colic-dataset/

[3] https://www.slideshare.net/tenzou2411/tiu-lun-khai-ph-d-liu-s-dng-wekazphn-lp-trn-dataset-weather-arff

[4] https://cuongndh.blogspot.com/p/khai-pha-du-lieu.html

[5] https://machinelearningcoban.com/2016/12/28/linearregression/

2. Tài liệu tiếng anh.

[7] https://machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/

[8] https://en.wikipedia.org/wiki/Mean_squared_error.