TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN CUỐI KỲ

Đề tài: Khai phá dữ liệu và dự đoán cấp độ mắc ung thư phổi của bệnh nhân

SVTH :

Phan Văn Thạch Quang	**MSSV: 20133083**
Cao Trọng Nghĩa	**MSSV: 20133071**
Trần Văn Trọng	**MSSV: 20133103**
Nguyễn Minh Đức	**MSSV: 19110350**
Phan Gia Huy	**MSSV: 19110369**

GVHD :

GV. Trần Trọng Bình

Tp. Hồ Chí Minh, tháng 05 năm 2023

Mục Lục

1. Giới thiệu đề tài

Ung thư phổi là nguyên nhân hàng đầu gây tử vong do ung thư trên toàn thế giới, chiếm 1,59 triệu ca tử vong vào năm 2018. Phần lớn các trường hợp ung thư phổi là do hút thuốc, nhưng tiếp xúc với ô nhiễm không khí cũng là một yếu tố rủi ro. Một nghiên cứu mới đã phát hiện ra rằng ô nhiễm không khí có thể liên quan đến việc tăng nguy cơ ung thư phổi, ngay cả ở những người không hút thuốc.

Nghiên cứu được công bố trên tạp chí Y học Tự nhiên đã xem xét dữ liệu từ hơn 462.000 người ở Trung Quốc, những người được theo dõi trong trung bình sáu năm. Những người tham gia được chia thành hai nhóm: những người sống ở khu vực có mức độ ô nhiễm không khí cao và những người sống ở khu vực có mức độ ô nhiễm không khí thấp.

Các nhà nghiên cứu phát hiện ra rằng những người trong nhóm ô nhiễm cao có nhiều khả năng phát triển ung thư phổi hơn những người trong nhóm ô nhiễm thấp. Họ cũng phát hiện ra rằng nguy cơ ở những người không hút thuốc cao hơn so với những người hút thuốc và nguy cơ này tăng lên theo độ tuổi.

Mặc dù nghiên cứu này không chứng minh rằng ô nhiễm không khí gây ra ung thư phổi, nhưng nó gợi ý rằng có thể có mối liên hệ giữa hai vấn đề này. Cần nhiều nghiên cứu hơn để xác nhận những phát hiện này và để xác định ảnh hưởng của các loại và mức độ ô nhiễm không khí khác nhau đối với nguy cơ ung thư phổi.

2. Dữ liệu (dataset)

Nguồn dữ liệu: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

Bộ dữ liệu này chứa thông tin về bệnh nhân ung thư phổi, bao gồm các thông tin tin liên quan về mặt triệu chứng y tế và thông tin sinh học của người bệnh. Bằng cách phân tích dữ liệu này, chúng tôi có thể hiểu rõ hơn về nguyên nhân gây ung thư phổi và cách điều trị tốt nhất

Số cột: 25 cột thuộc tính, trong đó 1 numeric, 24 categories

Số dòng: 2000 dòng (với dữ liệu gốc trên kaggle là 1000 dòng, bổ sung thêm sau khi cập nhật tại trang web gốc của dữ liệu)

+ Tuổi: Tuổi của bệnh nhân. (numeric)

+ Giới tính: Giới tính của bệnh nhân

+ Ô Nhiễm Không Khí: Mức độ tiếp xúc với ô nhiễm không khí của bệnh nhân.

+ Sử dụng rượu: Mức độ sử dụng rượu của bệnh nhân.

+ Dị ứng bụi: Mức độ dị ứng bụi của bệnh nhân.

+ Occupational Hazards: Mức độ nguy hiểm nghề nghiệp của bệnh nhân.

+ Rủi ro di truyền: Mức độ rủi ro di truyền của bệnh nhân.

+ Bệnh Phổi mãn tính: Mức độ bệnh phổi mãn tính của bệnh nhân.

+ Chế độ ăn uống cân bằng: Mức độ ăn uống cân bằng của bệnh nhân.

+ Béo phì: Mức độ béo phì của bệnh nhân.

+ Hút thuốc: Mức độ hút thuốc của bệnh nhân.

+ Passive Smoker: Mức độ hút thuốc lá thụ động của bệnh nhân.

+ Đau Ngực: Mức độ đau ngực của bệnh nhân.

+ Ho ra máu: Mức độ ho ra máu của người bệnh.

+ Mệt mỏi: Mức độ mệt mỏi của bệnh nhân.

+ Giảm cân: Mức độ giảm cân của bệnh nhân.

+ Khó thở: Mức độ khó thở của người bệnh.

+ Khò khè: Mức độ thở khò khè của người bệnh

+ Nuốt Khó: Mức độ nuốt khó của người bệnh.

+ Móng tay khoèo: Mức độ móng tay khoèo của bệnh nhân.

Chuẩn hóa dữ liệu : Tập dữ liệu nhóm sử dụng không có missing values . Nhóm có tiền xử lý dữ liệu cho mô hình học máy bằng cách sử dụng các giá trị đã phân loại danh mục sẵn (ngoại trừ độ tuổi) để có thể đưa vào phân tích và dự đoán cấp độ ung thư của bệnh nhân đã thuộc giai đoạn nào.

Thư viện sử dụng : Các thư viện chính được sử dụng trong phân tích như : matplotlib, seaborn, pandas, numpy, plotly, scipy, sklearn

3. Câu hỏi phân tích

3.1. Các yếu tố ảnh hưởng đến cấp độ bệnh ung thư phổi có phân bố như thế nào ?

- **Biến dự đoán (X)** : 22 biến categorical, 1 biến numeric (age, gender, air po llution exposure, alcohol use, dust allergy, occupational hazards, genetic risk , chronic lung disease, balanced diet, obesity, smoking status, passive smoke r status, chest pain, coughing of blood, fatigue levels , weight loss , shortness of breath , wheezing , swallowing difficulty , clubbing of finger nails , frequ ent colds , dry coughs , and snoring)
- **Biến kết quả (Y)** : Level
- **Ý nghĩa** : từ phân tích này giúp chúng ta hiểu được phân bố của các yếu tố ảnh hưởng đến bệnh ung thư phổi.
- **Biểu đồ có thể sử dụng** : histogram
- **Phương pháp sử dụng** : thống kê mô tả (Mô tả các đặc trưng của một phân bố cho biến số ,Mô tả các đặc trưng của một phân bố cho biến phân loại).

3.2. Phân tích các yếu tố có ảnh hưởng quan trọng đến tiến triển của bệnh ung thư phổi? Đâu là những yếu tố quan trọng ?

- **Biến dự đoán (X)** : 22 biến categorical, 1 biến numeric (age, gender, air po llution exposure, alcohol use, dust allergy, occupational hazards, genetic risk , chronic lung disease, balanced diet, obesity, smoking status, passive smoke r status, chest pain, coughing of blood, fatigue levels , weight loss , shortness of breath , wheezing , swallowing difficulty , clubbing of finger nails , frequ ent colds , dry coughs , and snoring)
- **Biến kết quả (Y)** : Level
- **Ý nghĩa** : từ phân tích này chúng ta thấy được một cách tổng quan về các yếu tố ảnh hưởng đến bệnh ung thư, từ biểu đồ rút ra kết luận về các yếu tố chủ yếu gây nên sự tiến triển của bệnh ung thư.
- **Biểu đồ có thể sử dụng** : Bar .
- **Phương pháp sử dụng** : Thống kê mô tả (Mô tả mối quan hệ của hai biến phân loại , Mô tả phân bố của một biến số theo các nhóm của một biến phân loại).

3.3. Có phải đa số bệnh nhân nam thường mắc ung thư phổi hơn so với nữ hay không ?

- **Biến dự đoán (X)** : Gender.
- **Ý nghĩa** : từ kiểm định này chúng ta có thể biết được liệu giới tính nào có nguy cơ mắc bệnh phổi cao hơn .
- **Phương pháp sử dụng** : kiểm định thống kê (two sample t-test for two means ($\mu 1, \mu 2$)).

3.4. Ước lượng độ tuổi có nguy cơ mắc bệnh phổi cao ?

- **Biến dự đoán (X)** : Age .
- **Biến kết quả (Y)** : Level .
- **Ý nghĩa** : từ phân tích này chúng ta biết được khoảng độ tuổi nào là có nguy cơ cao để từ đó có những lời khuyên phù hợp cho bệnh nhân .
- **Phương pháp sử dụng** : Ước lượng population mean

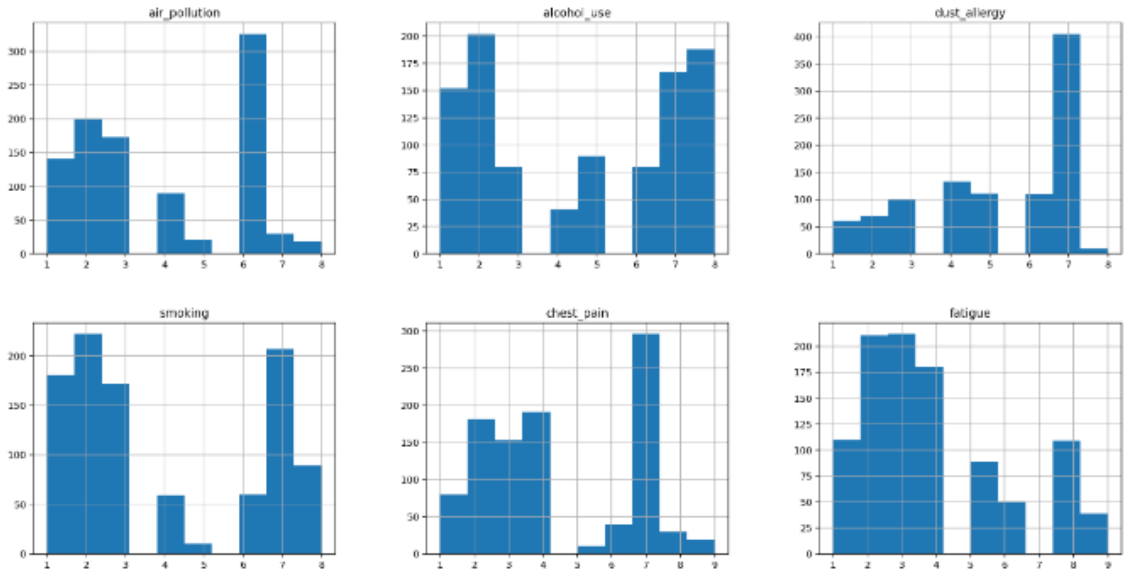
3.5. Xây dựng mô hình chuẩn đoán bệnh ung thư phổi dựa vào bộ dữ liệu

- **Biến dự đoán (X)** : 22 biến categorical, 1 biến numeric (age, gender, air pollution exposure, alcohol use, dust allergy, occupational hazards, genetic risk , chronic lung disease, balanced diet, obesity, smoking status, passive smoker status, chest pain, coughing of blood, fatigue levels , weight loss , shortness of breath , wheezing , swallowing difficulty , clubbing of finger nails , frequent colds , dry coughs , and snoring)
- **Kết quả** là biến : Level (0 là low, 1 là mid, 2 là high)
- **Ý nghĩa** : Từ mô hình học máy giúp dự đoán bệnh phổi nhằm phát hiện sớm để có phương pháp điều trị kịp thời.
- **Phương pháp sử dụng** : Các mô hình dự báo bằng các thuật toán như hồi quy, cây quyết định, Nause Baye,…

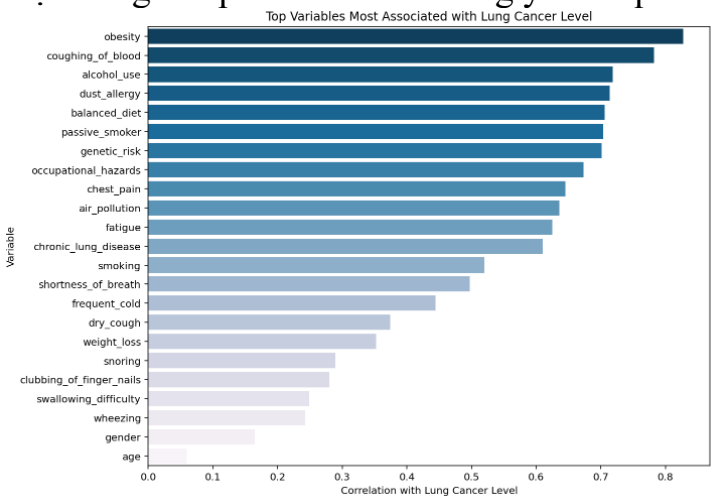
4. Trực quan hóa dữ liệu (data visulization)

4.1. Trả lời task 2.1 : Các yếu tố ảnh hưởng đến cấp độ bệnh ung thư phổi có phân bố như thế nào ?

➔ **Nhận xét** : Nhìn chung các biến phân loại có sự phân tán rõ về cấp độ của mỗi biến, ví dụ với biến air-polution có đa phần là mức độ thấp (dưới mức độ 3)

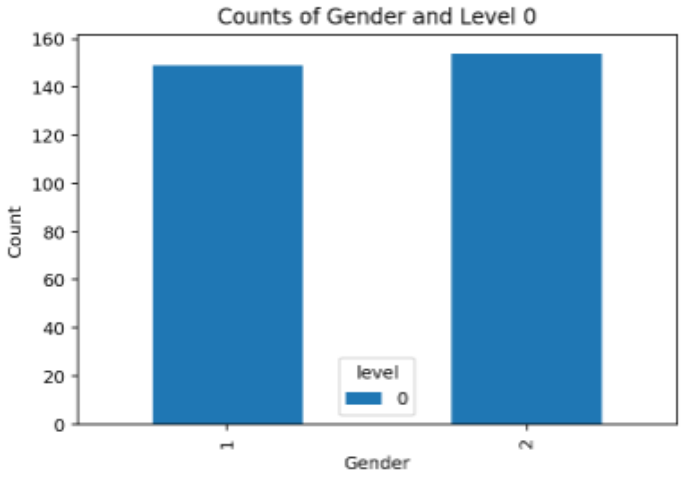


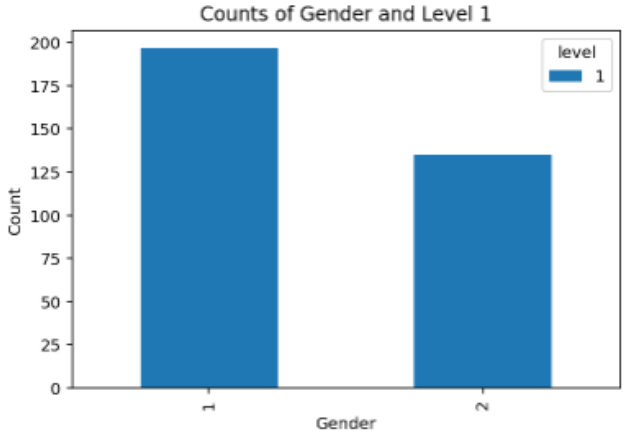
4.2. Trả lời task 2.2 : Phân tích các yếu tố có ảnh hưởng quan trọng đến tiến triển của bệnh ung thư phổi? Đâu là những yếu tố quan trọng ?



➔ **Nhận xét** : Từ biểu đồ có thể thấy các biến có ảnh hưởng đến bệnh phổi nhất chính là: obesity, coughing_of_blood, alcohol_use, dust_allergy, balanced_diet,…

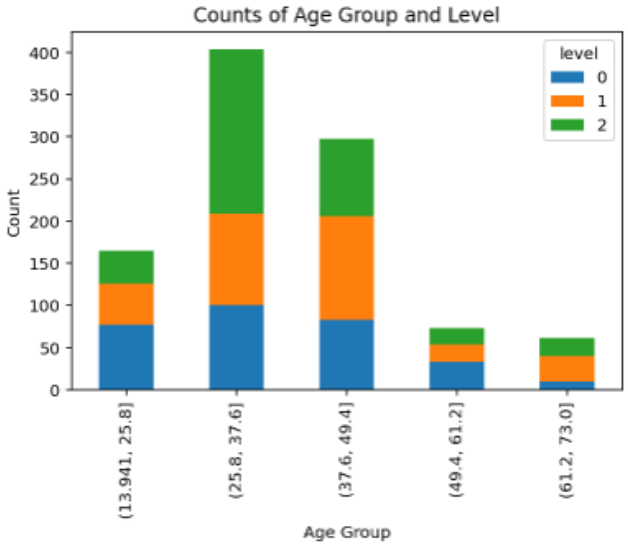
4.3. Trả lời task 2.3 : Có đa số bệnh nhân năm thường mắc ung thư phổi hơn so với nữ hay không ?





➔ **Nhận xét** : Kết quả phân tích trên cho thấy tỉ lệ mắc bệnh đối với nam so với nữ có sự thay đổi tuỳ theo mức độ ung thư (mức độ càng cao thì tỉ lệ nam càng tăng so với tỉ lệ nữ)

4.4. Trả lời task 2.4 : Ước lượng độ tuổi có nguy cơ mắc bệnh cao ?



➔ **Nhận xét** : Kết quả phân tích trên cho thấy nhóm tuổi trung niên từ 25 đến 50 tuổi dễ mắc ung thư phổi hơn so với các nhóm tuổi trẻ hoặc cao tuổi

5. Mô hình hóa dữ liệu (data modeling)

Trả lời cho task 2.5 : Xây dựng mô hình chuẩn đoán cấp độ ung thư dựa vào bộ dữ liệu

5.1. Mô hình Logistic Regression

Mô hình Logistic Regression (hay còn gọi là hồi quy logistic) là một phương pháp thống kê dùng để phân loại các dữ liệu vào một trong hai hoặc nhiều nhóm. Nó được sử dụng phổ biến trong các bài toán phân loại và dự đoán.

5.2. Mô hình Gaussian Naive Bayes

Mô hình Gaussian Naive Bayes là một mô hình phân loại dựa trên giả thuyết Bayes và giả định Naive Bayes. Mô hình này được sử dụng để dự đoán xác suất của một điểm dữ liệu thuộc về một lớp nhất định dựa trên các đặc trưng của nó.

5.3. Mô hình Decision Tree

Cây quyết định (**Decision Tree**) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật.

5.4. Mô hình KNN

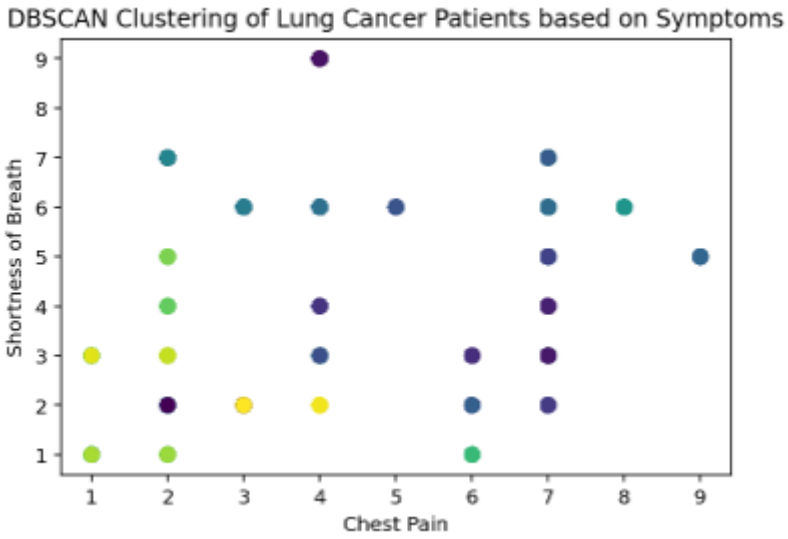
K-nearest neighbors là thuật toán học máy có giám sát, đơn giản và dễ triển khai. Thường được dùng trong các bài toán phân loại và hồi quy.

5.5. Kết quả của các mô hình

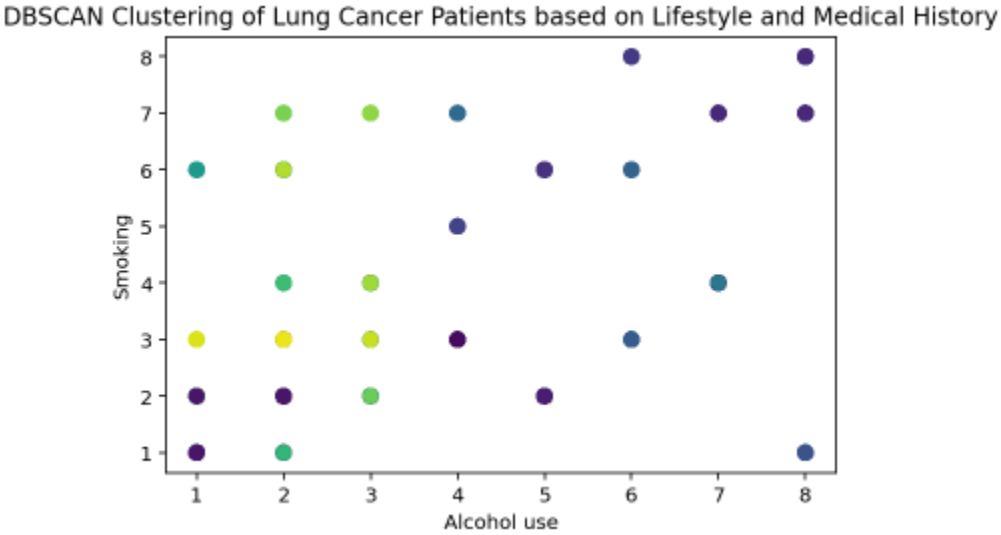
Đối với 3 mô hình Logistic Regression, KNN và Decision Tree cho ra kết quả dự đoán đúng 100%, suy ra các mô hình này có thể bị overfitting so với tập dữ liệu, hoặc dữ liệu trong tập chưa đạt yêu cầu về độ lớn và phân hoá dữ liệu, đối với Gaussian Naive Bayes là trên 90%.

6. Bài toán phân cụm và khai phá luật kết hợp

Sử dụng bài toán phân cụm để phân nhóm các bệnh nhân dựa trên các triệu chứng bệnh hoặc thói quen sống của bệnh nhân bằng thuật toán DBSCAN. DBSCAN là một thuật toán cơ sở để phân nhóm dựa trên mật độ. Nó có thể phát hiện ra các cụm có hình dạng và kích thước khác nhau từ một lượng lớn dữ liệu chứa nhiễu.

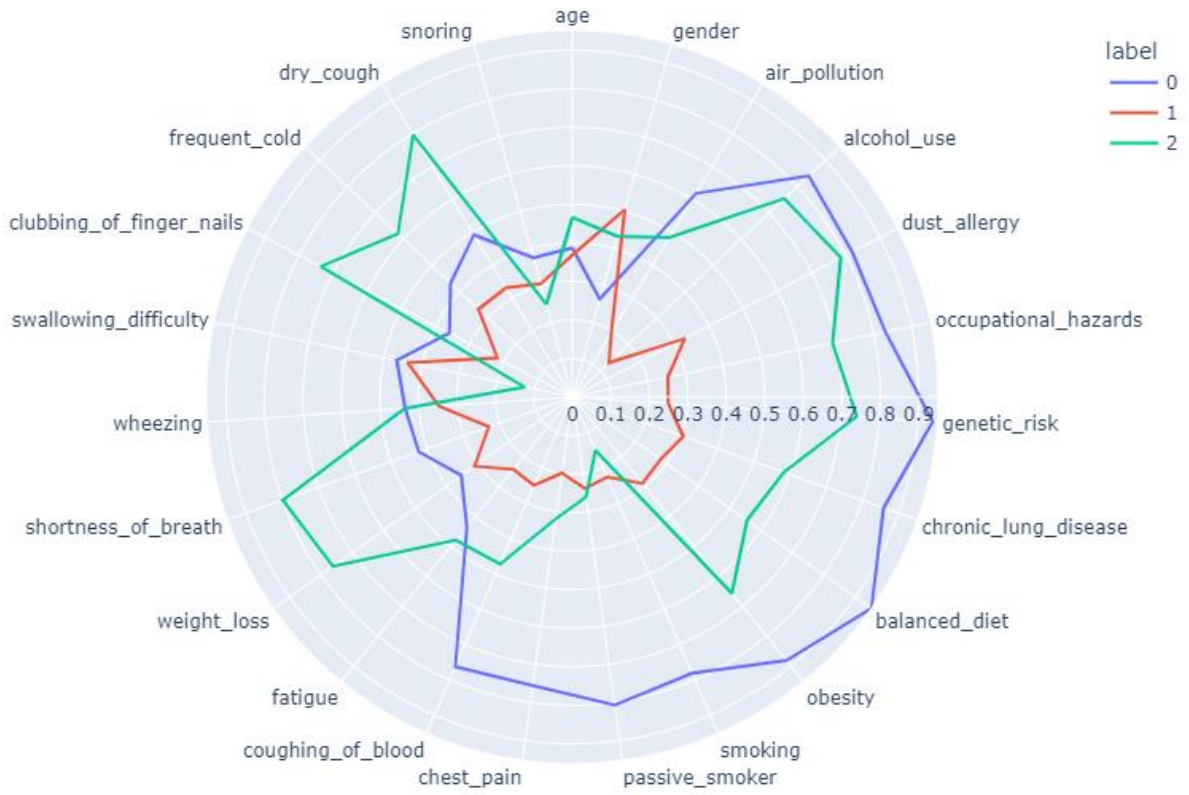


DBSCAN cho phân cụm nhóm bệnh nhân dựa trên triệu chứng



DBSCAN cho phân cụm nhóm bệnh nhân dựa trên lối sống

Sử dụng K-means để gom nhóm các bệnh nhân với tất cả các giá trị triệu chứng bệnh và lối sống, được trực quan hoá qua biểu đồ sau đây:



Với bài toán khai phá luật kết hợp, nhóm sử dụng thuật toán Apriori. Nó có thể được sử dụng để xác định mối liên hệ giữa các yếu tố lối sống khác nhau và sự phát triển của bệnh ung thư phổi. Và cũng có thể được sử dụng để khám phá những triệu chứng hoặc yếu tố rủi ro nào thường liên quan đến một loại hoặc giai đoạn ung thư phổi cụ thể.

```
                                             antecedents  \
311993  (Genetic Risk_7, chronic Lung Disease_6, Dust ...
308692  (Wheezing_7, chronic Lung Disease_6, Dust Alle...
213263  (Genetic Risk_7, chronic Lung Disease_6, Balan...
93732   (chronic Lung Disease_6, Smoking_7, Obesity_7,...
308746  (chronic Lung Disease_6, Smoking_7, Dust Aller...
213262  (Genetic Risk_7, chronic Lung Disease_6, Chest...
308741       (Balanced Diet_7, Smoking_7, Genetic Risk_7)
308732  (chronic Lung Disease_6, Smoking_7, Genetic Ri...
213260  (Genetic Risk_7, chronic Lung Disease_6, Chest...
308717         (chronic Lung Disease_6, Smoking_7, Level)

                                             consequents  antecedent support  \
311993                 (Wheezing_7, Chest Pain_7, Level)            0.108946
308692  (Level, Genetic Risk_7, Chest Pain_7, Balanced...            0.108946
213263                       (Wheezing_7, Chest Pain_7)            0.108946
93732                  (Wheezing_7, Coughing of Blood_7)            0.108946
308746  (Wheezing_7, Level, Genetic Risk_7, Chest Pain...            0.108946
213262                     (Wheezing_7, Balanced Diet_7)            0.108946
308741  (Wheezing_7, Level, chronic Lung Disease_6, Du...            0.108946
308732  (Wheezing_7, Level, Dust Allergy_7, Chest Pain...            0.108946
213260                 (Wheezing_7, Coughing of Blood_7)            0.108946
308717  (Wheezing_7, Genetic Risk_7, Dust Allergy_7, C...            0.108946

        consequent support   support  confidence      lift  leverage  \
...
308741                 inf       1.0
308732                 inf       1.0
213260                 inf       1.0
308717                 inf       1.0
```

Đầu ra của thuật toán khai thác quy tắc kết hợp

Trong kết quả này, thuật toán đã xác định một số nhóm mục thường xuất hiện cùng nhau trong tập dữ liệu của bạn:

+ Cột tiền đề (antecedents column) liệt kê các mục được tìm thấy cùng nhau và cột hệ quả liệt kê các mục có xu hướng theo sau chúng.

+ Cột hỗ trợ (support column) đưa ra tần suất mà các mục tiền đề và hậu quả xảy ra cùng nhau trong tập dữ liệu. Cột độ tin cậy cho biết xác suất có điều kiện của việc tìm thấy các mục kết quả trong các giao dịch có chứa các mục trước đó.

+ Cột nâng (lift column) cho biết mức độ mà sự hiện diện của các mục trước ảnh hưởng đến sự hiện diện của các mục tiếp theo, liên quan đến tần suất riêng lẻ của chúng. Giá trị mức tăng lớn hơn 1 cho biết rằng sự hiện diện của các mục tiền đề làm tăng khả năng tìm thấy các mục tiếp theo, trong khi giá trị nhỏ hơn 1 cho biết điều ngược lại.

+ Cột đòn bẩy (leverage column) cho biết sự khác biệt giữa tần suất quan sát được của các mục tiền đề và hậu quả xảy ra cùng nhau và tần suất dự kiến nếu chúng độc lập với nhau.

+ Cột xác tín (conviction column) là một thước đo khác về mức độ phụ thuộc giữa các mục tiền đề và hậu quả. Nó dựa trên khái niệm về tỷ lệ chênh lệch và đo lường mức độ mà các mục hệ quả phụ thuộc vào sự vắng mặt của các mục trước đó.

+ Cuối cùng, cột số liệu của Zhang là một thước đo khác về mối liên hệ giữa các mục tiền đề và hệ quả, dựa trên khái niệm entropy thông tin. Nó đo lường mức độ thu được thông tin trong việc dự đoán các mục hệ quả, với sự hiện diện của các mục trước đó

7. Kết luận

Qua quá trình thực hiện đề tài nhóm em đã học hỏi được thêm kiến thức về học máy ứng dụng ngôn ngữ Python cho phân tích dữ liệu .

Về tổng quan đề tài đã hoàn thành mức cơ bản , trả lời được các câu hỏi nhóm đã đặt ra với bộ dữ liệu . Tuy nhiên kết quả dự đoán của các mô hình trong đề tài chưa được cao . Sau đây là nhận xét về các thuật toán nhóm đã sử dụng :

Thuật toán	Ưu điểm	Nhược điểm
Logistic Regression	Đây là thuật toán đơn giản và dễ hiểu. Nó hoạt động tốt với các bài toán phân loại nhị phân. Nó tính toán hiệu quả và có thể xử lý được các tập dữ liệu lớn.	Nó có thể không hoạt động tốt với các ranh giới phân chia phi tuyến tính. Nó nhạy cảm với các điểm ngoại lai và có thể bị ảnh hưởng bởi sự tương quan giữa các đặc trưng đầu vào. Nó không thể xử lý các đặc trưng đầu vào không liên quan hoặc trùng lặp một cách tốt.
Gaussian Naive Bayes	Đây là thuật toán đơn giản và nhanh chóng, làm cho nó hiệu quả với các tập dữ liệu lớn. Nó hoạt động tốt với các dữ liệu có số chiều cao. Nó có thể xử lý cả các bài toán phân loại nhị phân và đa lớp. Nó có thể xử lý dữ liệu bị thiếu rất tốt.	Nó giả định rằng các đặc trưng đầu vào là độc lập, điều này có thể không đúng trong một số trường hợp. Nó có thể hoạt động kém nếu giả định về phân phối Gaussian không đúng. Nó có thể gặp vấn đề "tần số bằng không" nếu giá trị của một đặc trưng hạng mục không có mặt trong tập dữ liệu huấn luyện
Decision Tree	Đây là thuật toán đơn giản và dễ hiểu.	Nó có thể dễ dàng bị quá khớp dữ liệu nếu cây quá sâu hoặc số

	Nó có thể xử lý cả dữ liệu hạng mục và số. Nó có thể nắm bắt được các mối quan hệ phi tuyến giữa các đặc trưng. Nó có thể được sử dụng cho cả các bài toán phân loại và hồi quy	mẫu tối thiểu trên mỗi lá quá nhỏ. Nó có thể nhạy cảm với các thay đổi nhỏ trong dữ liệu, dẫn đến các cấu trúc cây khác nhau. Nó có thể tạo ra các cây bị thiên vị nếu một số lớp chiếm nhiều
KNN	Độ phức tạp tính toán của quá trình training là bằng 0. Việc dự đoán kết quả của dữ liệu mới rất đơn giản. Không cần giả sử gì về phân phối của các class.	KNN rất nhạy cảm với nhiễu khi K nhỏ.