A logo with a ball and a circle

Description automatically generated with medium confidence

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**MÔN HỌC: KHAI PHÁ DỮ LIỆU**

**BÁO CÁO ĐỒ ÁN**

**ĐỀ TÀI:** **PHÂN TÍCH VÀ XÂY DỰNG MÔ HÌNH DỰ ĐOÁN UNG THƯ PHỔI CỦA BỆNH NHÂN**

**GVHD:** *GV. Trần Trọng Bình*

**Lớp HP:** *DAMI330484\_23\_2\_01*

**Nhóm thực hiện:** *Nhóm 12*

**Học kỳ:** *2*

**Năm học:** *2023 - 2024*

*Thành phố Hồ Chí Minh, tháng 5 năm 2024*

**DANH SÁCH SINH VIÊN THỰC HIỆN**

**HỌC KÌ 2 NĂM HỌC 2023 – 2034**

**NHÓM 12**

***Đề tài: Dự đoán cấp độ ung thư phổi của bệnh nhân***

|  |  |
| --- | --- |
| **Họ tên Sinh viên** | **Mã số Sinh viên** |
| Lê Hoàn Nhật | 21110576 |
| Ngô Minh Lợi | 21110898 |
| Mai Trần Bách | 21110379 |
| Thái Bảo An | 21110365 |

*TP. Hồ Chí Minh, ngày 23 tháng 5 năm 2024*

*Giảng viên hướng dẫn*

*(Ký và ghi rõ họ tên)*

***Nhận xét của GV:***

**MỤC LỤC**

[1. Giới thiệu đề tài 1](#_Toc167400564)

[2. Mô tả tập dữ liệu 2](#_Toc167400565)

[3. Câu hỏi phân tích: 4](#_Toc167400566)

[3.1. Các yếu tố nào có ảnh hưởng đến bệnh ung thư phổi và chúng có phân bố như thế nào? 4](#_Toc167400567)

[3.2. Trong các mức độ của yếu tố gây ra bệnh thì ở mức độ nào người bệnh dễ mắc ung thư phổi nhất ? 4](#_Toc167400568)

[3.3. Xây dựng mô hình chuẩn đoán bệnh ung thư phổi dựa vào bộ dữ liệu 4](#_Toc167400569)

[4. Trực quan hóa dữ liệu 6](#_Toc167400570)

[4.1. Các yếu tố nào có ảnh hưởng đến bệnh ung thư phổi và chúng có phân bố như thế nào? 6](#_Toc167400571)

[4.2. Trong các mức độ của yếu tố gây ra bệnh thì ở mức độ nào người bệnh dễ mắc ung thư phổi nhất ? 6](#_Toc167400572)

[5. Mô hình hóa dữ liệu 8](#_Toc167400573)

[5.1. Logistic Regression 8](#_Toc167400574)

[5.2. Gaussian Navie Bayes 8](#_Toc167400575)

[5.3. Random Forest 8](#_Toc167400576)

[5.4. Kmeans 8](#_Toc167400577)

[5.5. SVM 8](#_Toc167400578)

[5.6. KNN 9](#_Toc167400579)

[5.7. ANN 9](#_Toc167400580)

[5.8. Khai phá luật kết hợp 9](#_Toc167400581)

[5.9. Kết quả 9](#_Toc167400582)

[6. Kết luận 10](#_Toc167400583)

[TÀI LIỆU THAM KHẢO 12](#_Toc167400584)

# **Giới thiệu đề tài**

Ung thư phổi là một loại ung thư phổ biến trên thế giới, lên tới con số 1,59 triệu ca tử vong vào năm 2018. Phần lớn nguyên nhân gây ra ung thư phổi theo mặc định của chúng ta là do hút thuốc hoặc tiếp xúc với bầu không khí ô nhiễm, độc hại trong khoảng thời gian lâu dài. Nhưng trong tình trạng bầu không khí đang ngày càng ô nhiễm từ khói bụi của các chất thải nhà máy, phương tiện giao thông, v.v thì liệu có trở thành một nguyên nhân tiềm ẩn gây ra bệnh này không.

Nghiên cứu được công bố trên tạp chí Y học Tự nhiên đã đánh giá dữ liệu từ hơn 462.000 người ở Trung Quốc trong một cuộc thử nghiệm trong vòng 6 tháng. 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.

Sau quá trình quan sát, nghiên cứu kĩ lưỡng thì các nhà nghiên cứu đã đút kết được 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. Lại có thêm một điều tưởng chừng phi lý là 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.

# **Mô tả tập dữ liệu**

Dữ liệu nghiên cứu được lấy từ nguồn: 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 em 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: 1000 dòng (với dữ liệu gốc trên kaggle là 1000 dòng, bổ sung thêm sau khi cập nhật tại trang web gốc của dữ liệu)

+ Age: Tuổi của bệnh nhân. (numeric)

+ Gender: Giới tính của bệnh nhân

+Air Pollution: Mức độ tiếp xúc với ô nhiễm không khí của bệnh nhân.

+Alcohol use: Mức độ sử dụng rượu của bệnh nhân.

+Dust Allergy: Mức độ dị ứng bụi của bệnh nhân.

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

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

+ Chronic Lung Disease: Mức độ bệnh phổi mãn tính của bệnh nhân.

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

+ Obesity: Mức độ béo phì của bệnh nhân.

+ Smoking: Mức độ hút thuốc của bệnh nhân.

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

+ Chest Pain: Mức độ đau ngực của bệnh nhân.

+ Coughing of Blood: Mức độ ho ra máu của người bệnh.

+ Fatigue: Mức độ mệt mỏi của bệnh nhân.

+ Weight Loss: Mức độ giảm cân của bệnh nhân.

+ Shortness of Breath: Mức độ khó thở của người bệnh.

+ Wheezing: Mức độ thở khò khè của người bệnh

+ Swallowing Difficulty: Mức độ nuốt khó của người bệnh.

+ Clubbing of Finger Nails: 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.

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

# **Câu hỏi phân tích:**

1. **Các yếu tố nào có ảnh hưởng đến bệnh ung thư phổi và chúng có phân bố như thế nào?**

- **Biến dự đoán (X)** : 23 biến categorical ('age', 'gender', 'genetic\_risk', 'chronic\_lung\_disease', 'balanced\_diet', 'obesity', 'chest\_pain', 'coughing\_of\_blood', 'fatigue','weight\_loss','shortness\_of\_breath','wheezing','swallowing\_difficulty','clubbing\_of\_finger\_nails', 'frequent\_cold','dry\_cough', 'snoring' ,'air\_pollution', 'alcohol\_use', 'dust\_allergy','occupational\_hazards','smoking','passive\_smoker')

**- Biến kết quả (Y)** : ‘level’

**- Ý nghĩa** : từ phân tích này giúp chúng ta hiểu được các yếu tố có mức độ ảnh hưởng đến bệnh ung thư phổi ra sai

**- Biểu đồ có thể sử dụng** : bar

**- 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).

1. **Trong các mức độ của yếu tố gây ra bệnh thì ở mức độ nào người bệnh dễ mắc ung thư phổi nhất ?**

- **Biến dự đoán (X)** : 23 biến categorical ('age', 'gender', 'genetic\_risk', 'chronic\_lung\_disease','balanced\_diet','obesity','chest\_pain','coughing\_of\_blood', 'fatigue','weight\_loss','shortness\_of\_breath','wheezing','swallowing\_difficulty','clubbing\_of\_finger\_nails', 'frequent\_cold','dry\_cough', 'snoring' ,'air\_pollution', 'alcohol\_use', 'dust\_allergy','occupational\_hazards','smoking','passive\_smoker')

- **Biến kết quả (Y)** : level

- **Ý nghĩa** : từ phân tích này chúng ta thấy được ở các yếu tố ảnh hưởng thì mức độ nào có ảnh hưởng nhât đến bệnh phổi.

**- Biểu đồ có thể sử dụng** : bar

- **Phương pháp sử dụng** : Thống kê mô tả (đếm số lượng mắc bệnh ở từng mức độ và thống kê)

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

- **Biến dự đoán (X)** : 23 biến categorical ('age', 'gender', 'genetic\_risk', 'chronic\_lung\_disease','balanced\_diet','obesity','chest\_pain','coughing\_of\_blood','fatigue','weight\_loss','shortness\_of\_breath','wheezing','swallowing\_difficulty','clubbing\_of\_finger\_nails', 'frequent\_cold','dry\_cough', 'snoring' ,'air\_pollution', 'alcohol\_use', 'dust\_allergy','occupational\_hazards','smoking','passive\_smoker')

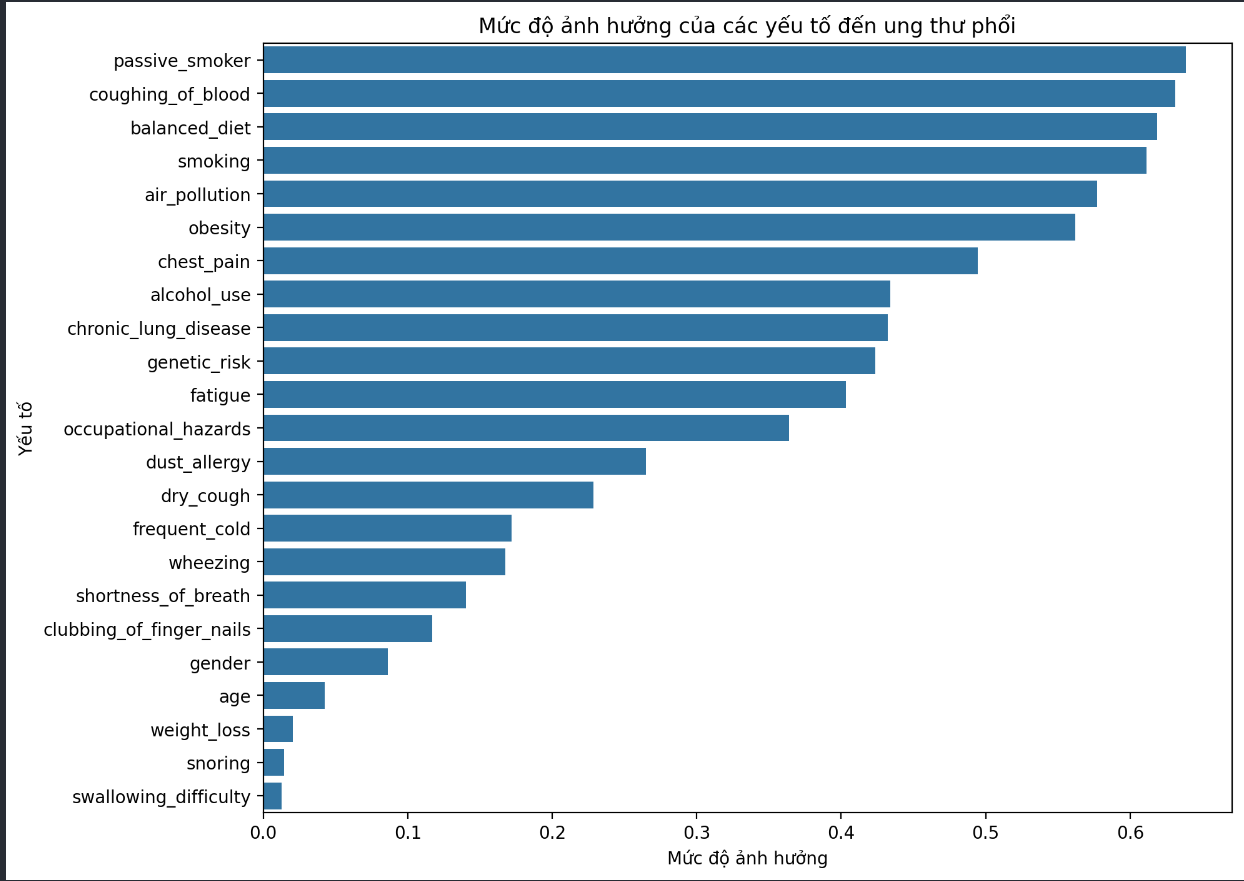
**- Kết quả là biến** : level

- **Ý nghĩa** : từ kết quả dự đoán của mô hình có thể giúp phát hiện bệnh sớm nhất có thể từ đó có những phương pháp điều trị thích hợp.

- **Phương pháp sử dụng** : Các mô hình dự báo bằng các thuật toán như Logistic Regression,Random Forest, Gaussian Bayes,…

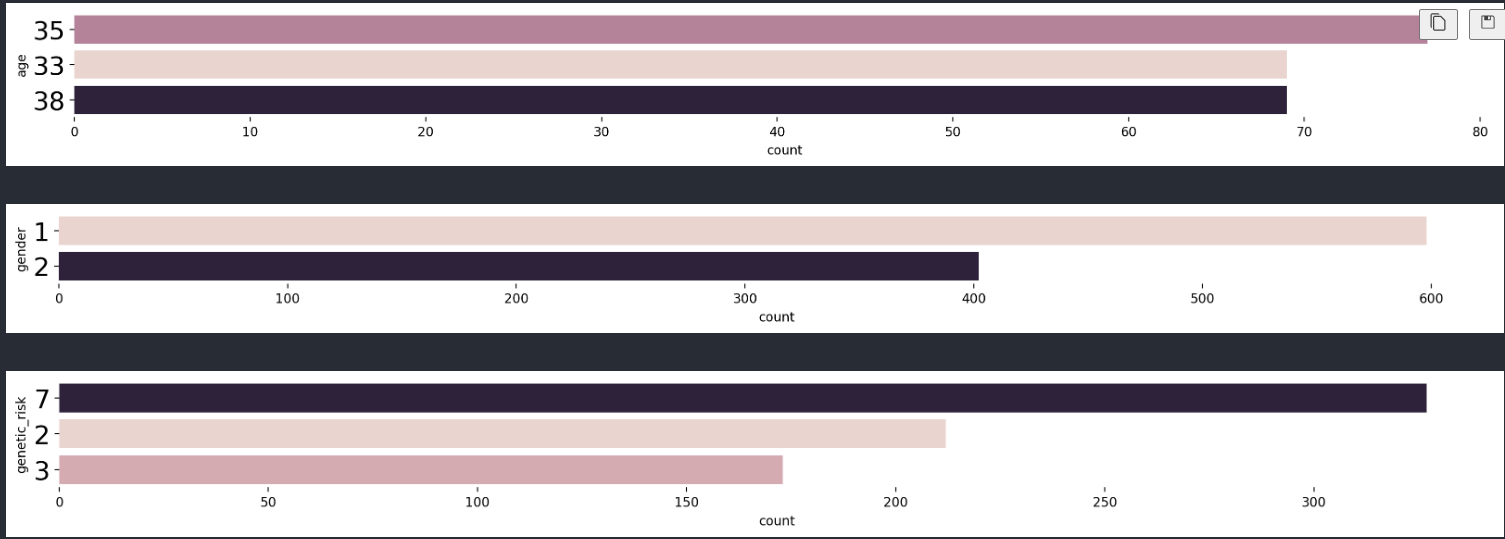
# **Trực quan hóa dữ liệu**

## **Các yếu tố nào có ảnh hưởng đến bệnh ung thư phổi và chúng có phân bố như thế nào?**

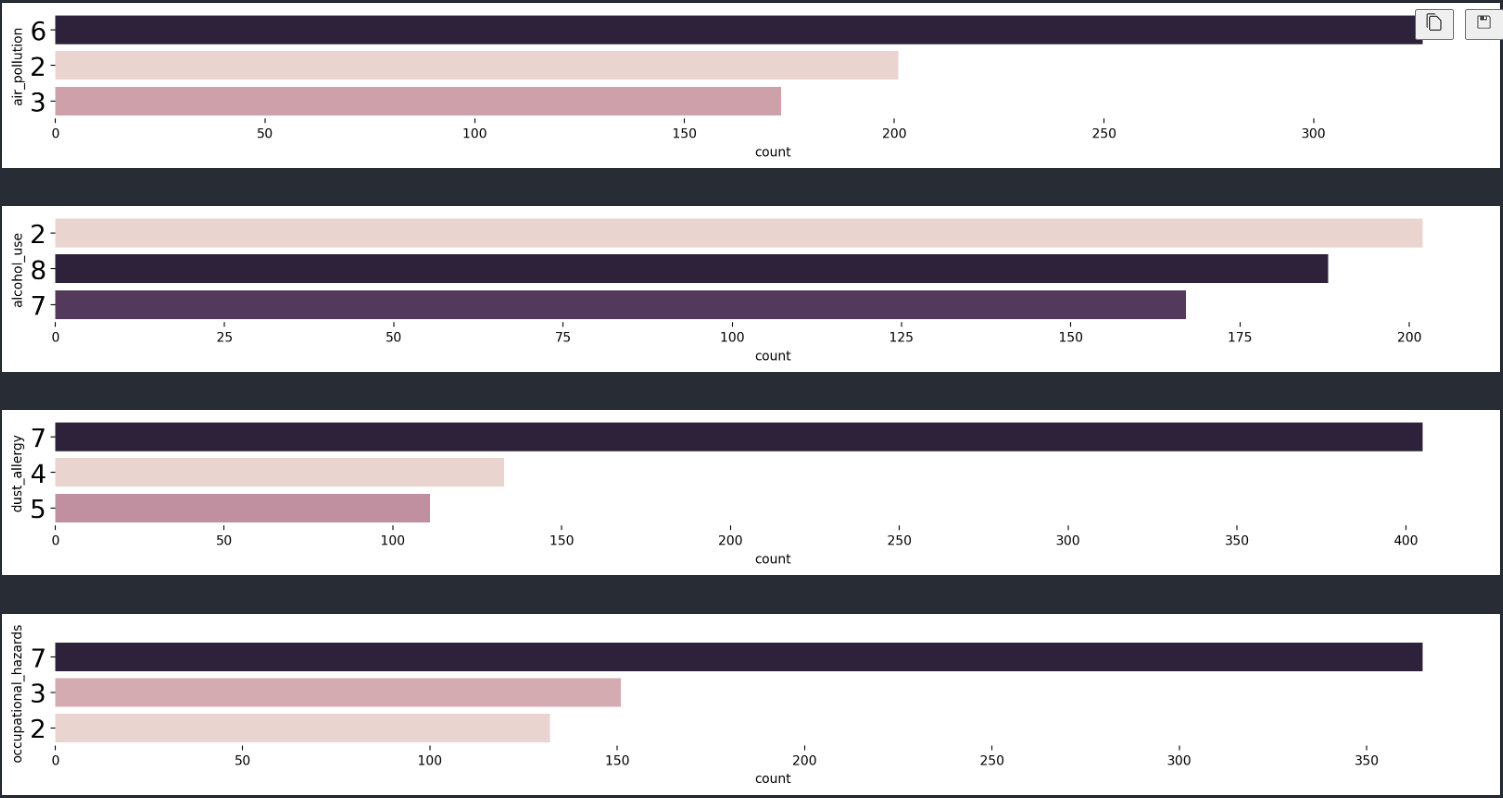
******

***Nhận xét:*** *các yếu tố ảnh hưởng nhiều nhất gây ra ung thư phổi là passive\_smoker, balanced\_diet, smoking,...*

## **Trong các mức độ của yếu tố gây ra bệnh thì ở mức độ nào người bệnh dễ mắc ung thư phổi nhất ?**



**Nhận xét:** có thể thấy về yếu tố tự nhiên thì ở độ tuổi 35, giới tính nam,... người ta dễ mắc bệnh phổi nhất



**Nhận xét:** có thể thấy về yếu tố xã hội thì ô nhiễm không khí ở mức độ cao, thường xuyên tiêu thụ rượu,... khiến người ta dễ mắc bệnh phổi nhất

# **Mô hình hóa dữ liệu**

## **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.

## **Gaussian Navie Bayes**

Gaussian Naive Bayes là một phương pháp phân loại dựa trên định lý Bayes, mà giả định rằng các đặc trưng là độc lập với nhau và có phân phối Gaussian (phân phối chuẩn). Đây là một trong những phương pháp phân loại đơn giản nhất và phổ biến trong lĩnh vực học máy. Phương pháp này thường được sử dụng cho các bài toán phân loại với dữ liệu liên tục, nhưng có thể được mở rộng để xử lý dữ liệu rời rạc thông qua việc ước lượng các tham số phân phối Gaussian từ dữ liệu mẫu.

## **Random Forest**

Random Forest là một phương pháp học máy dựa trên việc xây dựng một tập hợp (ensemble) của các cây quyết định (decision trees). Ý tưởng cơ bản của Random Forest là kết hợp sức mạnh của nhiều cây quyết định để tạo ra một mô hình phân loại hoặc dự đoán mạnh mẽ hơn.

Random Forest thường được ưa chuộng vì nó có khả năng xử lý tốt các dữ liệu lớn, không cần nhiều điều chỉnh siêu tham số và ít bị overfitting hơn so với một cây quyết định đơn lẻ. Đặc biệt, Random Forest có khả năng xử lý các vấn đề phức tạp và dữ liệu không đồng nhất, làm cho nó trở thành một trong những phương pháp phổ biến nhất trong học máy.

## **Kmeans**

K-means là một thuật toán phân cụm trong lĩnh vực học không giám sát. Mục tiêu của thuật toán này là phân chia các điểm dữ liệu thành các nhóm (cụm) sao cho các điểm trong cùng một nhóm có tính chất tương đồng với nhau.

Kết quả của thuật toán K-means là các trọng tâm của các cụm và phân chia của các điểm dữ liệu vào các cụm tương ứng. K-means thường được sử dụng để phân cụm dữ liệu không gắn nhãn và tìm ra cấu trúc ẩn bên trong dữ liệu. Nó có thể được áp dụng trong nhiều lĩnh vực, từ phân tích hành vi người dùng trên mạng đến phân tích khách hàng trong kinh doanh. Tuy nhiên, điều quan trọng là phải chú ý đến việc chọn số lượng cụm phù hợp và kiểm tra sự ổn định của kết quả phân cụm.

## **SVM**

Support Vector Machine (SVM) là một thuật toán học máy phổ biến được sử dụng cho các bài toán phân loại và hồi quy. Thuật toán này hoạt động bằng cách tìm ra ranh giới phân chia tốt nhất giữa các lớp dữ liệu.

SVM được sử dụng rộng rãi trong nhiều lĩnh vực, từ phân loại văn bản, nhận dạng hình ảnh đến dự đoán thị trường tài chính. Nó là một trong những thuật toán mạnh mẽ trong học máy với khả năng xử lý tốt các bài toán phân loại có số lượng đặc trưng lớn và dữ liệu không tuyến tính.

## **KNN**

K-nearest neighbors (KNN) là một thuật toán học máy đơn giản nhưng mạnh mẽ được sử dụng cho các bài toán phân loại và hồi quy. Ý tưởng cơ bản của KNN là dự đoán nhãn của một điểm dữ liệu mới bằng cách xem xét các điểm dữ liệu gần nhất trong tập huấn luyện và dự đoán dựa trên đa số nhãn của các điểm gần nhất đó.

KNN không yêu cầu giả định về phân phối của dữ liệu và có thể được sử dụng cho cả các bài toán phân loại lẫn hồi quy. Tuy nhiên, điều quan trọng cần lưu ý là việc chọn giá trị K phù hợp, vì giá trị K ảnh hưởng đến tính chính xác và hiệu suất của mô hình. KNN thường được sử dụng trong các bài toán mà không có cấu trúc đặc trưng phức tạp hoặc khi dữ liệu có ít nhiễu.

## **ANN**

Artificial Neural Network (ANN), hay còn được gọi là mạng nơ-ron nhân tạo, là một mô hình học máy được lấy cảm hứng từ cấu trúc của não người. Mạng nơ-ron nhân tạo được xây dựng từ các đơn vị tính toán gọi là nơ-ron, được tổ chức thành các lớp và kết nối với nhau qua các trọng số.

ANN được sử dụng rộng rãi trong nhiều lĩnh vực, từ nhận dạng hình ảnh và âm thanh đến dự đoán dữ liệu tài chính và dự báo chuỗi thời gian. Đặc biệt, mạng nơ-ron sâu (Deep Neural Networks - DNN) là một biến thể mạnh mẽ của ANN, có khả năng học biểu diễn phức tạp và xử lý hiệu quả các bài toán có dữ liệu lớn.

## **Khai phá luật kết hợp**

Khai phá luật kết hợp (Association Rule Mining) là một phương pháp trong lĩnh vực khai phá dữ liệu (Data Mining) nhằm tìm ra các mối quan hệ và quy luật xuất hiện đồng thời trong tập dữ liệu. Mục tiêu chính của khai phá luật kết hợp là phát hiện các quy luật tương quan giữa các mục (items) trong dữ liệu và biểu diễn chúng dưới dạng các luật kết hợp.

Khai phá luật kết hợp được áp dụng rộng rãi trong nhiều lĩnh vực, từ bán lẻ và tiếp thị đến y sinh và khoa học xã hội, để phát hiện ra các mẫu mua hàng, hành vi tiêu dùng, hay mối quan hệ trong dữ liệu phức tạp.

## **Kết quả**

Các mô hình dự đoán mà nhóm đã sử dụng đều cho kết quả trên 90% cho thấy rằng dữ liệu tập dữ liệu mà nhóm chọn chưa thật sự đủ lớn và đa dạng về phân loại hóa dữ liệu, dẫn đến các mô hình có thể dễ bị overfitting.

1. **Kết luận**

Qua quá trình thực hiện đề tài này, nhóm đã được ôn lại những kiến thức về các mô hình phân loại đã học, đồng thời cũng có thêm những kiến thức mới trong quá trình trực quan hóa dữ liệu và tối ưu kết quả máy học.

Về tổng quan nhóm đã trả lời được những câu hỏi về tập dữ liệu. Tuy nhiên do tập dữ liệu chưa thật sự phù hợp nên kết quả của các mô hình học máy chưa thật sự phản ánh đúng. Sau đây là nhận xét về các thuật toán mà nhóm đã sử dụng:

|  |  |  |
| --- | --- | --- |
| **Thuật toán** | **Ưu điểm** | **Nhược điểm** |
| **Logistic Regression** | * Dễ hiểu và dễ triển khai * Hiệu quả với dữ liệu tuyến tính * Ít cần điều chỉnh tham số | * Không xử lý tốt với dữ liệu phi tuyến tính * Dễ bị ảnh hưởng bởi nhiễu |
| **Gaussian Naive Bayes** | * Đơn giản và hiệu quả, đặc biệt là khi dữ liệu tuân theo phân phối Gaussian. * Xử lý tốt với dữ liệu lớn | * Giả định về sự độc lập giữa các biến * Không phù hợp với dữ liệu phi tuyến tính |
| **Random Forest** | * Mạnh mẽ, có khả năng xử lý cả các bài toán phân loại và hồi quy, cũng như dữ liệu lớn và không gian đặc trưng lớn. * Giảm thiểu overfitting | * Tốn tài nguyên tính toán |
| **Support Vector Machine (SVM)** | * Hiệu quả với dữ liệu lớn * Mạnh mẽ trong không gian đặc trưng cao | * Đòi hỏi lựa chọn kernel phù hợp * Tốn tài nguyên tính toán |
| **K-nearest neighbors (KNN):** | * Đơn giản và dễ triển khai * Không cần giả định về phân phối dữ liệu | * Yêu cầu lưu toàn bộ tập dữ liệu trong bộ nhớ |
| **Artificial Neural Network (ANN)** | * Linh hoạt và mạnh mẽ * Có thể xử lý dữ liệu phi tuyến tính * Tích hợp được với các kiến trúc mạng phức tạp | * Đòi hỏi nhiều dữ liệu huấn luyện * Cần nhiều thời gian và tài nguyên tính toán * Dễ bị overfitting |
| **Kmeans** | * Đơn giản và dễ triển khai * Hiệu quả với dữ liệu lớn | * Cần chọn số lượng cụm * Nhạy cảm với điểm khởi tạo ban đầu * Không xử lý tốt với các cụm có kích thước và hình dạng không đồng nhất |

# **TÀI LIỆU THAM KHẢO**

The Devastator, “Lung Cancer Prediction”, Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link. Accessed on: April 28, 2024.

Jay Singh, “Training of ANN in Data Mining”, tutorialspoint, 2023. [Online]. Available: https://www.tutorialspoint.com/training-of-ann-in-data-mining. Accessed on: April 29, 2024.