**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO MÔN HỌC**

**KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI: Khai phá dữ liệu từ bộ dữ liệu ô nhiễm**

**không khí toàn cầu**

**GVHD: Nguyễn Tiến Đạt**

**Nguyễn Văn Bảy**

**LỚP: CS2101**

**SINH VIÊN THỰC HIỆN:**

**Trịnh Huỳnh Thịnh Khang - 2151013036 - Nhóm trưởng**

**Nguyễn Xuân Lộc - 2151013052**

# Lời mở đầu

Trong thời buổi công nghệ số ngày nay, công nghệ thông tin và những ứng dụng của chúng ngày càng phát triển không ngừng. Lượng thông tin, dữ liệu thu được và cần xử lý cũng ngày một nhiều lên, mà nếu chỉ dùng những phương pháp cũ thông thường để xử lý, khai thác dữ liệu thì không thể nào đáp ứng được tốc độ và hiệu quả. Chúng ta cần những phương pháp khác, những thứ có thể xử lý nhanh chóng và đưa ra quyết định một cách chính xác dựa trên lượng dữ liệu khổng lồ mà ta có được. Chính vì điều đó, một khuynh hướng kỹ thuật mới có tên gọi là “Kỹ thuật phát hiện tri thức và khai thác dữ liệu” được ra đời và nhanh chóng phát triển, đáp ứng nhu cầu của người dùng về xử lý thông tin.

Khai thác dữ liệu đã và đang được nghiên cứu, ứng dụng rộng rãi khắp nơi trên thế giới. Ở Việt Nam, kỹ thuật này tuy xuất hiện chưa lâu nhưng đã được nghiên cứu và phát triển không ngừng, áp dụng nhiều trong các lĩnh vực khác nhau. Khai phá dữ liệu là một bước đóng vai trò quan trọng trong quy trình phát hiện tri thức, đưa ra quyết định. Cùng với sự phát triển không ngừng của mã nguồn mở, toán học và các thuật toán, con người không ngừng tìm hiểu, học hỏi tạo ra các kỹ thuật để thực hiện khai phá dữ liệu với mục tiêu nhanh nhất và đem lại hiệu quả tốt nhất.

Trong bài báo cáo này, chúng em đã vận dụng những kiến thức đã học cùng với việc học hỏi, tìm hỏi thêm nhiều kỹ thuật, công nghệ mới để áp dụng khai phá dữ liệu một cách nhanh chóng và hiệu quả nhất. Đề tài của tụi em là “ Khai phá dữ liệu từ bộ dữ liệu ô nhiễm không khí toàn cầu kết hợp ứng dụng các thuật toán máy học”. Thông qua bài báo cáo, đề tài trên chúng em có thể thu được cái nhìn tổng quan, những yếu tố ảnh hưởng đến môi trường, những tri thức có được từ việc khai thác bộ dữ liệu, từ đó có thể đề xuất, đưa ra các phương hướng giải quyết. Trong quá trình làm bài báo cáo trên, chúng em cũng xin gửi lời cảm ơn đến hai thầy Nguyễn Tiến Đạt và Nguyễn Văn Bảy đã rất tận tình chỉ dạy chúng em các phương pháp cũng như hướng dẫn chi tiết cách làm, những kiến thức mà tụi em có được vô cùng hữu ích. Chúng em rất mong nhận được những góp ý từ các thầy ạ. Chúng em xin chân thành cảm ơn!

# Mục lục

[**Lời mở đầu 2**](#_ey71otnreegr)

[**Mục lục 3**](#_3ox4swl2nayj)

[**Chương 1. Tổng quan về khai phá dữ liệu. 5**](#_it5yj2jza9la)

[1.1. Định nghĩa về dữ liệu. 5](#_srbeyl5rzloa)

[1.2. Khai phá dữ liệu là gì? 5](#_58ztr26nybv9)

[1.3. Quy trình khai phá dữ liệu. 5](#_528g317ilmq3)

[1.4. Tính ứng dụng của khai phá dữ liệu. 6](#_781rooqzlffr)

[1.5. Tiền xử lý dữ liệu. 6](#_bx0ewn2a6p85)

[1.5.1. Làm sạch dữ liệu. 7](#_yx2v2zlcrxlv)

[1.5.2. Tích hợp dữ liệu. 7](#_qqnjrn9qu1wi)

[1.5.3. Biến đổi dữ liệu. 7](#_d22sh9arqoq)

[1.5.4. Thu giảm dữ liệu. 7](#_7kok3s1c045s)

[1.6. Các kỹ thuật khai thác dữ liệu thông dụng. 8](#_ie3rakwbw51f)

[1.6.1. Kỹ thuật Gom Cụm. 8](#_eqrqe1ngsnmh)

[1.6.1.1. Kỹ thuật K-means. 9](#_d6adytd6zc1e)

[Hình 1.3. Cách hoạt động của kỹ thuật K-means 9](#_4qf0xp74ny8s)

[1.6.1.2. Kỹ thuật Phân Cấp (Hierarchical Clustering): 10](#_f59gxx11898m)

[1.6.2. Kỹ thuật Phân Lớp: 11](#_1mkuhx7iqlk8)

[1.6.2.1. Kỹ thuật K-nearest-neighbors (KNN): 11](#_qtsy8kgpn89z)

[1.6.2.2. Kỹ thuật Naive-Bayes: 13](#_vtlf1qm9akmh)

[1.6.2.3. Kỹ thuật Cây Quyết Định. 14](#_gxc0mjtn1ec0)

[1.6.3. Kỹ thuật Khai Phá Luật Kết Hợp. 14](#_qxrro52jpwvk)

[1.6.3.1. Kỹ thuật Apriori. 14](#_5opdlt9ca2f)

[**Chương 2. Tiến hành khai phá dữ liệu 15**](#_xkwsdnn2l6tt)

[2.1. Đặt vấn đề, xác định bài toán, mục tiêu: 15](#_i351e0cox90n)

[2.2. Xác định các công cụ và kỹ thuật áp dụng: 15](#_i5mi464no6hl)

[2.3. Tổng quan bộ dữ liệu: 21](#_dir2ctws2zaj)

[2.4. Tiền xử lý dữ liệu 37](#_5ynxg0zficpx)

[2.4.1. Xử lý dữ liệu bị thiếu 37](#_3wlt7khyn3d5)

[2.4.2. Biến đổi dữ liệu: 38](#_12n8p7e0nzu3)

[2.4.3. Chia tách dữ liệu: 40](#_xawbjlw7zpu2)

[2.5. Khai phá dữ liệu: 41](#_ekru1vc6tc5v)

[2.5.1. Bộ dữ liệu df1: 41](#_ayph9dflwac9)

[2.5.1.1. Chuẩn hóa bộ dữ liệu: 41](#_4wpvgpcihgw5)

[2.5.1.2. Phân cụm với kỹ thuật K-means: 42](#_9mjkhrcc05pu)

[2.5.1.3. Đánh giá tổng quan khả năng dự đoán của các thuật toán, mô hình hồi quy, chọn ra mô hình áp dụng cho bài toán. 47](#_ifgscgoioy2q)

[2.5.1.4. Mô hình LinearRegression: 49](#_af3gg1fznpj9)

[a) Train mô hình: 49](#_qzskmjhv9rw2)

[b) Trực quan hóa kết quả thu được: 49](#_2iymtk5pw463)

[c) Đánh giá dựa trên Shap Value: 50](#_hgkcooeiz1i6)

[d) Đánh giá dựa trên kết quả dự đoán với thực tế: 51](#_a9c9pelg2fob)

[e) Đánh giá dựa trên Confusion Matrix: 52](#_fvg5fwg4muhs)

[f) Tổng kết: 53](#_lkdb7mggt0ju)

[2.5.2. Bộ dữ liệu df2: 54](#_odb4hb4sajq0)

[2.5.2.1. Đánh giá tổng quan khả năng dự đoán của các thuật toán, mô hình phân loại, chọn ra mô hình áp dụng cho bài toán. 54](#_nc3h9ratssw8)

[2.5.2.2. Mô hình DecisionTreeClassifier: 55](#_mj1fxvp4qfm0)

[a) Train mô hình: 55](#_qrg1f01j843z)

[b) Trực quan hóa mô hình (thuật toán) DecisionTreeClassifier Classifier thu được: 55](#_st8181u0ecbt)

[Hình 2.46.a. Trực quan hóa mô hình 55](#_7rj8ojuyd3zf)

[c) Đánh giá dựa trên kết quả dự báo so với thực tế trên bộ dữ liệu test : 57](#_xxuf3oj3syye)

[d) Đánh giá dựa trên Confusion Matrix: 58](#_y4wbptc1itc9)

[e) Đánh giá dựa trên mức độ ảnh hưởng của các thuộc tính với mô hình: 58](#_ib9ey094i0t0)

[f) Tổng kết: 59](#_c1kknrevs7g0)

[2.5.3. Bộ dữ liệu df3 60](#_85u637sg337r)

[2.5.3.1. Đánh giá tổng quan khả năng dự đoán của các thuật toán, mô hình phân loại, chọn ra mô hình áp dụng cho bài toán. 60](#_abqmc47vtmne)

[2.5.3.2. Mô hình KNeighborsClassifier: 61](#_ry63c7au955p)

[a) Train mô hình: 61](#_7y2lygft815f)

[b) Đánh giá dựa trên kết quả dự báo so với thực tế trên bộ dữ liệu test: 61](#_u5us01c8q7g6)

[c) Tổng kết: 63](#_wqck0uswtcr6)

[2.5.3.3. Sử dụng thuật toán Apriori để khám phá các luật kết hợp trong bộ dữ liệu 63](#_71gudt5m9c9a)

[a) Chuẩn hóa bộ dữ liệu: 63](#_6kafcu9e6m1x)

[b) Áp dụng thuật toán Apriori để tìm các luật kết hợp: 63](#_j0g65yctc1m)

[c) Trực quan hóa các luật kết hợp: 66](#_72z10igsokvk)

[d) Tổng kết 66](#_wo0eg3ivq78y)

[**KẾT LUẬN 68**](#_qvmtgghbm5l7)

[**TÀI LIỆU THAM KHẢO 69**](#_saxlmq3338vy)

# 

# Chương 1. Tổng quan về khai phá dữ liệu.

## 1.1. Định nghĩa về dữ liệu.

* Dữ liệu là một tập hợp các thông tin được thu thập lưu trữ và xử lý để thực hiện cho các mục đích khác nhau như nghiên cứu, trích xuất, quản lý hỗ trợ ra quyết định [1].
* Một tập dữ liệu (dataset) là một tập hợp các đối tượng (object) và các thuộc tính của chúng [2].
* Mỗi thuộc tính mô tả một đặc điểm, đặc trưng của đối tượng [2].

## 1.2. Khai phá dữ liệu là gì?

Khai phá dữ liệu ( Data mining ) là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó [3].

## 1.3. Quy trình khai phá dữ liệu.

**B1. Xác định vấn đề, bài toán:**

Xác định vấn đề được đặt ra, phân loại thuộc dạng bài toán nào từ đó đề ra phương hướng giải quyết, mục tiêu, kết quả cần thu được, các thông tin cần khai phá.

**B2. Thu thập dữ liệu:**

* Dữ liệu liên quan cho một ứng dụng phân tích được xác định và tập hợp. Dữ liệu có thể được đặt trong các hệ thống nguồn khác nhau, một kho dữ liệu hoặc một bể chứa dữ liệu, đóng vai trò vô cùng quan trong việc khai phá tri thức, giải quyết vấn đề.
* Từ việc xác định được vấn đề, phương hướng giải quyết, mục tiêu phải đạt được, ta cần thu thập, tìm ra những dữ liệu phù hợp và phải đảm bảo tính tin cậy, có khả năng tổng quát hóa cao.

**B3. Chuẩn bị dữ liệu - Tiền xử lý dữ liệu:**

* Chuẩn bị dữ liệu hay tiền xử lý dữ liệu đóng vai trò vô cùng quan trọng, quyết định sự thành công hay thất bại của một mô hình, lượng tri thức, khả năng khai phá dữ liệu thu thập được.
* Thăm dò, lập hồ sơ và xử lý trước dữ liệu, sau đó là làm sạch dữ liệu, sửa lỗi và các vấn đề kiểm tra chất lượng dữ liệu khác. Việc chuyển đổi dữ liệu cũng được thực hiện để làm cho các tập dữ liệu nhất quán, trừ khi một nhà khoa học dữ liệu đang tìm cách phân tích dữ liệu thô chưa được lọc cho một ứng dụng cụ thể [4].

**B4. Khai phá dữ liệu:**

* Sau khi dữ liệu được chuẩn bị, các nhà khoa học dữ liệu chọn kỹ thuật data mining thích hợp và sau đó triển khai một hoặc nhiều thuật toán để thực hiện khai thác [4].
* Trong các ứng dụng học máy, các thuật toán thường phải được đào tạo trên các tập dữ liệu mẫu để tìm kiếm thông tin đang được tìm kiếm trước khi chúng chạy trên toàn bộ tập dữ liệu.

**B5. Đánh giá kết quả:**

* Thông qua việc trực quan hóa dữ liệu, ta đưa ra các nhận định về các đặc điểm, thuộc tính như mối quan hệ tương quan, đồng biến, nghịch biến, thuộc tính nào có ảnh hưởng cao đến kết quả, thuộc tính nào có ít tác dụng, có thể loại bỏ đi giúp giảm chiều dữ liệu.
* Ngoài ra, ta còn có thể dựa vào các chỉ số đánh giá mô hình như MSE (Mean squared error), MAE (Mean absolute error), R2, …. Confusion matrix (prediction, recall, TP, TN, FP, FN….) [5].

**B6. Ứng dụng thông tin thu thập được:**

Sử dụng các kết quả thu được ứng dụng vào thực tế. Từ đó giúp các tổ chức có thể tận dụng được những thông tin quý giá từ dữ liệu của mình để đưa ra các quyết định kinh doanh chính xác, hiệu quả, định hướng kế hoạch trong tương lai, các dự báo cho xã hội [6].

## 1.4. Tính ứng dụng của khai phá dữ liệu.

Công việc khai phá dữ liệu được ứng dụng nhiều trong các lĩnh vực như [6]:

* Kinh doanh, thương mại.
* Môi trường.
* Bảo hiểm.
* Khoa học và y sinh học.
* Viễn thông

Sau khi ứng dụng, tiến hành khai phá các dữ liệu thu thập được ta có thể thu lại các dữ liệu, thông tin, tri thức ẩn bên trong, giúp phân tích, vạch ra lộ trình, hướng đi phát triển, kế hoạch trong tương lai cho các cơ sở, trung tâm, doanh nghiệp.

## 1.5. Tiền xử lý dữ liệu.

Quá trình tiền xử lý dữ liệu đóng vai trò vô cùng quan trọng, góp phần không nhỏ vào việc trích xuất, khai phá dữ liệu thu được kết quả một cách thuận lợi, chính xác, hiệu quả, gặt hái được những tri thức, mối liên hệ trong dữ liệu. Quá trình tiền xử lý dữ liệu bao gồm các bước sau [7]:

### 1.5.1. Làm sạch dữ liệu.

Sau khi thu thập dữ liệu cần kiểm tra tổng quát, xác định các vấn đề ảnh hưởng khiến cho dữ liệu không sạch (rỗng, nhiễu, thiếu giá trị, mất cân bằng các giá trị, dữ liệu không đồng nhất, ….). Nếu ta sử dụng một bộ dữ liệu không sạch để phân tích, khai thác thì kết quả thu được không hiệu quả, không đáng tin cậy, khiến các quyết định đưa ra không còn chính xác. Vì vậy, chúng ta cần phải sửa chữa lại các lỗi lầm này bằng cách gán các giá trị còn thiếu, sửa lại các dữ liệu lỗi, sai định dạng, loại bỏ các giá trị ngoại lai, nhiễu, giải quyết các mâu thuẫn dữ liệu [7].

### 

### 1.5.2. Tích hợp dữ liệu.

Tích hợp dữ liệu là quá trình trộn, kết hợp dữ liệu từ nhiều nguồn khác nhau vào một kho dữ liệu chung trong quá trình khai thác dữ liệu. Chú ý giảm thiểu (tránh được thì càng tốt) các mâu thuẫn dư thừa phát sinh trong quá trình tích hợp dữ liệu. Từ đó ta có thể cải thiện được các mâu thuẫn và nâng cao chất lượng của các tri thức thu được [7].

### 1.5.3. Biến đổi dữ liệu.

Biến đổi dữ liệu là việc thay đổi toàn bộ tập dữ liệu của một thuộc tính từ dạng ban đầu sang dạng khác sao cho phần ý nghĩa, độ lớn cấp bậc dữ liệu không bị thay đổi quá nhiều, phù hợp với mục đích khai thác dữ liệu [7].

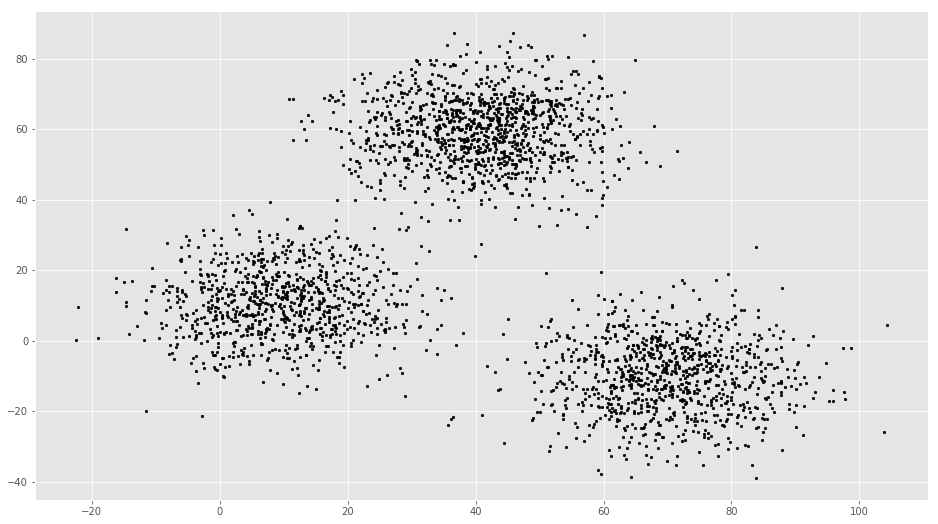
### 

### 1.5.4. Thu giảm dữ liệu.

* Một kho dữ liệu sẽ có kích thước vô cùng lớn khiến cho việc chạy phân tích, khai phá dữ liệu trở nên mất nhiều thời gian. Chưa kể có thể có các đoạn dữ liệu trùng lặp, các thuộc tính không cần thiết, không có ý nghĩa nhiều đối với kết quả trong việc khai phá dữ liệu, do đó ta có thể xem xét thu giảm dữ liệu nhằm tăng tốc độ xử lý và hiệu quả của việc khai thác dữ liệu [7].
* Thu giảm dữ liệu có thể bao gồm một vài phương pháp như: loại bỏ thuộc tính tính không cần thiết, loại bỏ bản ghi trùng lặp, phân tích thành phần chính,…
* Sau khi thu giảm dữ liệu, ta sẽ thu được một tập dữ liệu thu gọn có kích thước nhỏ hơn mà vẫn có thể sinh ra cùng (hoặc xấp xỉ, thậm chí là hơn) độ hiệu quả trong việc khai phá dữ liệu như tập dữ liệu ban đầu [7].

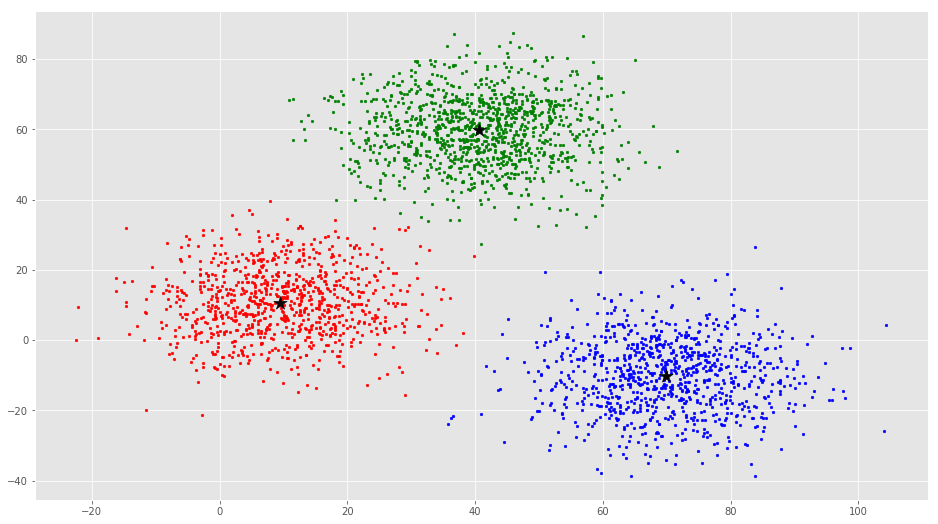
## 1.6. Các kỹ thuật khai thác dữ liệu thông dụng.

### 1.6.1. Kỹ thuật Gom Cụm.



**Hình 1.1. Các điểm dữ liệu trước khi gom cụm**

*(Nguồn:* [*https://mubaris.com/posts/kmeans-clustering/*](https://mubaris.com/posts/kmeans-clustering/)*)*

**

**Hình 1.2. Các điểm dữ liệu sau khi gom cụm**

*(Nguồn:* [*https://mubaris.com/posts/kmeans-clustering/*](https://mubaris.com/posts/kmeans-clustering/)*)*

#### 

#### 1.6.1.1. Kỹ thuật K-means.

* **Mục tiêu:** Gom các điểm dữ liệu thành các cụm sao cho các điểm trong cùng một cụm có tính tương đồng cao và giữa các cụm có sự khác biệt lớn.
* **Cách thức hoạt động [8]:**

**Bước 1:** Chọn k số cụm mong muốn.

**Bước 2:** Gán mỗi điểm dữ liệu vào cụm gần nhất.

**Bước 3:** Cập nhật trung tâm cụm dựa trên điểm trong cụm.

**Bước 4:** Lặp lại quá trình cho đến khi hội tụ.

#### Hình 1.3. Cách hoạt động của kỹ thuật K-means

*(Nguồn:* [*https://medium.com/@vaibhavnohria36/*](https://medium.com/@vaibhavnohria36/k-means-clustering-a3566359c12f)

[*k-means-clustering-a3566359c12f*](https://medium.com/@vaibhavnohria36/k-means-clustering-a3566359c12f)*)*

#### 

#### 1.6.1.2. Kỹ thuật Phân Cấp (Hierarchical Clustering):

* **Mục tiêu:** Xây dựng một cây phân cấp để mô tả sự tương đồng giữa các điểm dữ liệu.
* **Cách thức hoạt động [9]:**

**Bước 1:** Bắt đầu với mỗi điểm là một cụm.

**Bước 2:** Kết hợp các cụm gần nhất cho đến khi có một cụm duy nhất hoặc số lượng cụm mong muốn.

**Bước 3:** Cây phân cấp có thể biểu diễn qua đồ thị dendrogram.

### 

### 

### 

**Hình 1.5. Cách hoạt động của kỹ thuật phân cấp**

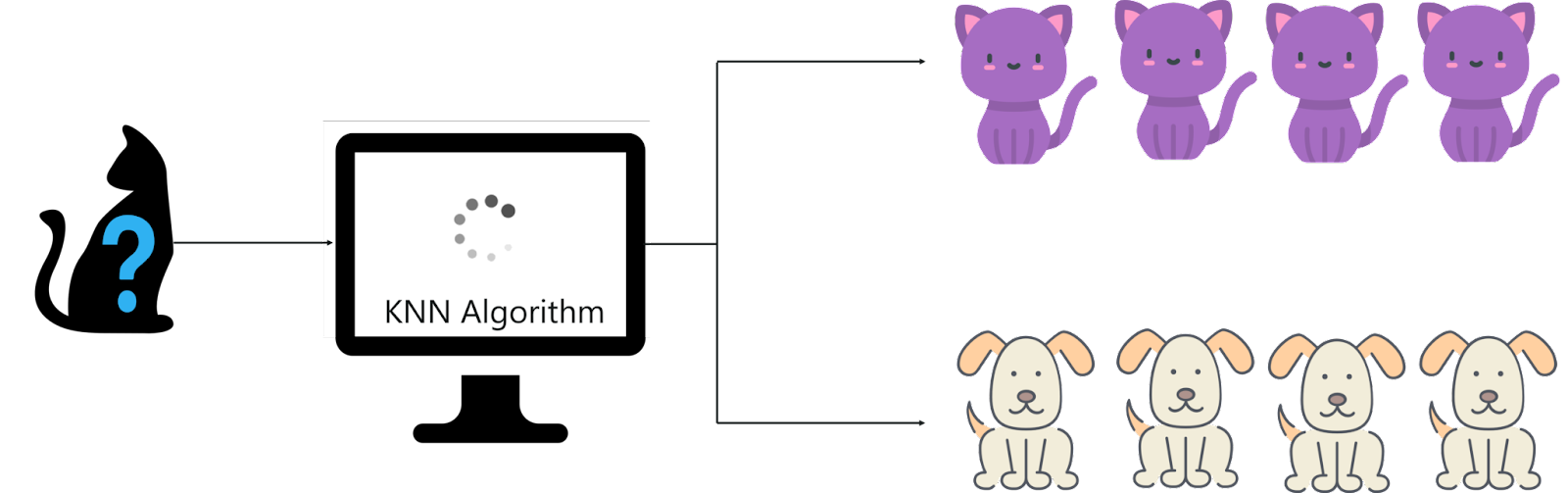
*(Nguồn:* [*https://python3.info/machine-learning*](https://python3.info/machine-learning/clustering/k-means.html)

[*/clustering/k-means.html*](https://python3.info/machine-learning/clustering/k-means.html)*)*

### 

### 1.6.2. Kỹ thuật Phân Lớp:

#### 1.6.2.1. Kỹ thuật K-nearest-neighbors (KNN):

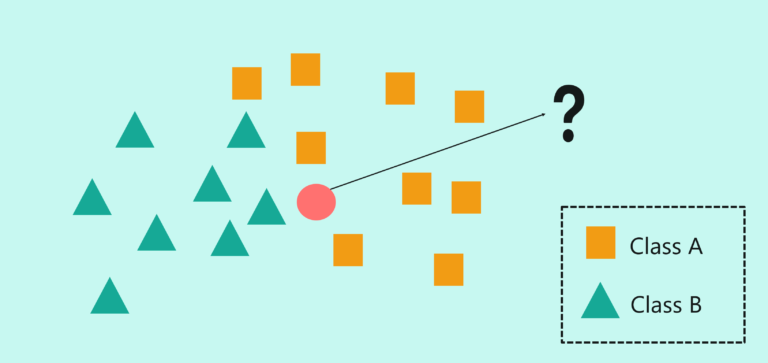


**Hình 1.6. Kỹ thuật K-nearest-neighbors (KNN)**

*(Nguồn:* [*https://medium.com/mlearning-ai/*](https://medium.com/mlearning-ai/k-nearest-neighbor-6d15a2b06772)

[*k-nearest-neighbor-6d15a2b06772*](https://medium.com/mlearning-ai/k-nearest-neighbor-6d15a2b06772)*)*

* **Mục tiêu:** Phân loại một điểm dữ liệu mới dựa trên đa số nhãn của k điểm dữ liệu gần nhất.



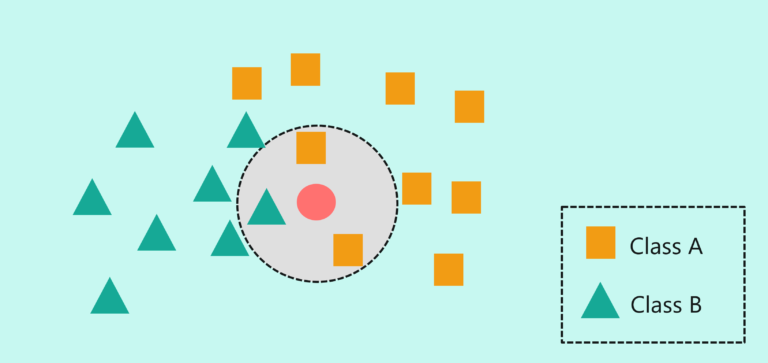
**Hình 1.7. Mục tiêu phân loại điểm dữ liệu mới hình tròn**

*(Nguồn:*[*https://www.edureka.co/blog/knn-algorithm-in-r/*](https://www.edureka.co/blog/knn-algorithm-in-r/)*)*

* **Cách thức hoạt động [10]:**

**Bước 1:** Xác định k điểm gần nhất dựa trên khoảng cách.

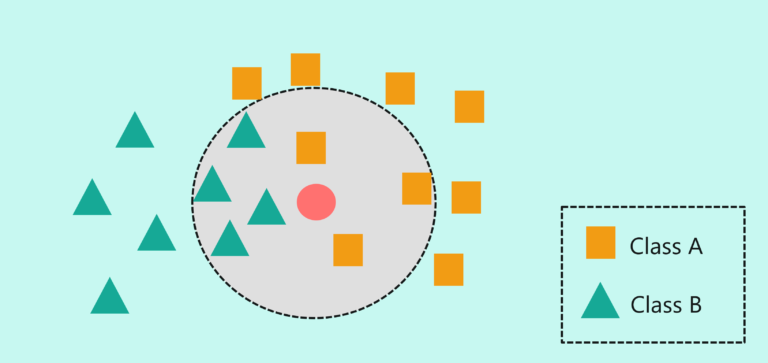
**Bước 2:** Đánh giá đa số nhãn của k điểm này để dự đoán nhãn cho điểm mới.



**Hình 1.8. Kết quả kỹ thuật KNN khi chọn k là 3**

*(Nguồn:*[*https://www.edureka.co/blog/knn-algorithm-in-r/*](https://www.edureka.co/blog/knn-algorithm-in-r/)*)*

\* Với k = 3, ta dự đoán được nhãn cho điểm mới là loại hình vuông (Class A), vì trong ba hàng xóm gần nhất có hai hình vuông (Class A) và một hình tam giác (Class B).



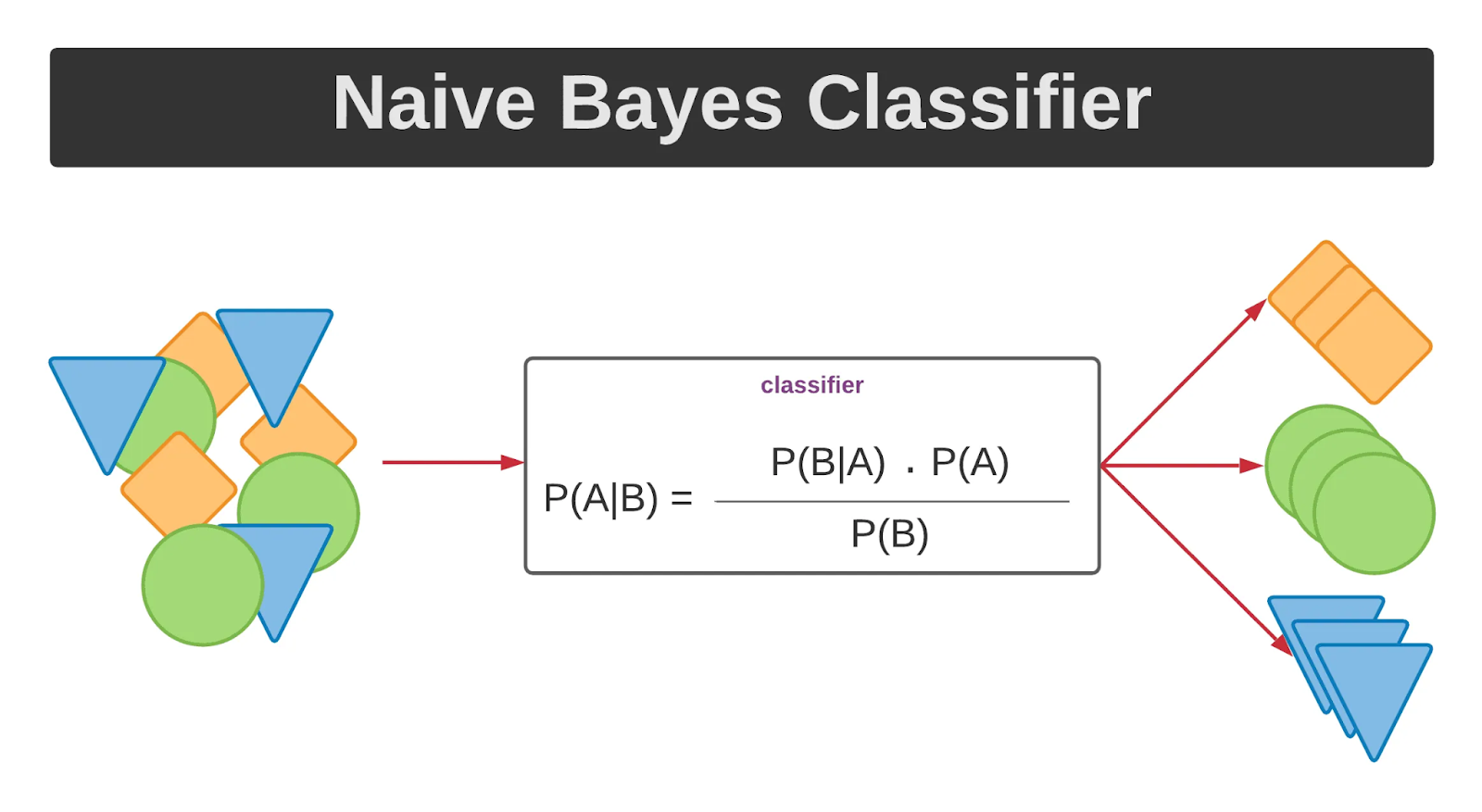
**Hình 1.9. Kết quả kỹ thuật KNN khi chọn k là 7**

*(Nguồn:*[*https://www.edureka.co/blog/knn-algorithm-in-r/*](https://www.edureka.co/blog/knn-algorithm-in-r/)*)*

\* Với k = 7, ta dự đoán được nhãn cho điểm mới là loại hình tam giác (Class B), vì trong bảy hàng xóm gần nhất có ba hình vuông (Class A) và bốn hình tam giác (Class B).

#### 

#### 1.6.2.2. Kỹ thuật Naive-Bayes:



**Hình 1.10. Cách thức hoạt động của kỹ thuật Naive Bayes**

*(Nguồn:* [*https://medium.com/@dancerworld60/demystifying-na%C3%AFve-*](https://medium.com/@dancerworld60/demystifying-na%C3%AFve-bayes-simple-yet-powerful-for-text-classification-ad92b14a5c7)

[*bayes-simple-yet-powerful-for-text-classification-ad92b14a5c7*](https://medium.com/@dancerworld60/demystifying-na%C3%AFve-bayes-simple-yet-powerful-for-text-classification-ad92b14a5c7)*)*

* **Mục tiêu:** Dự đoán nhãn của một điểm dữ liệu mới dựa trên xác suất Bayes.
* **Cách thức hoạt động [11]:**

**Bước 1:** Thu thập dữ liệu đào tạo: Dữ liệu đào tạo là một tập hợp các điểm dữ liệu đã được gắn nhãn, tức là mỗi điểm dữ liệu có một lớp xác định.

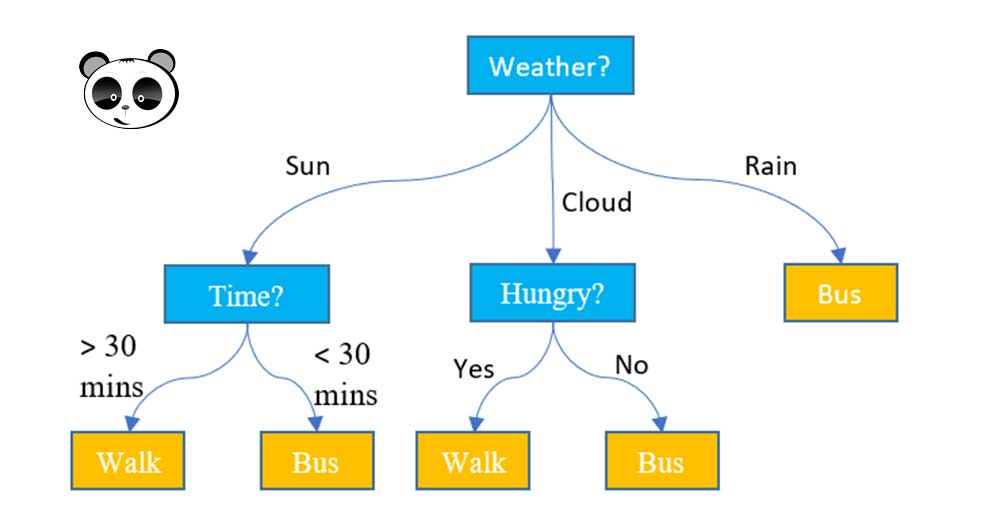
**Bước 2:** Tính toán xác suất xuất hiện của mỗi lớp: Kỹ thuật Naive-Bayes tính toán xác suất xuất hiện của mỗi lớp trong tập dữ liệu đào tạo.

**Bước 3:** Tính toán xác suất xuất hiện của mỗi đặc trưng trong mỗi lớp: Kỹ thuật Naive-Bayes cũng tính toán xác suất xuất hiện của mỗi đặc trưng trong mỗi lớp.

**Bước 4:** Dự đoán lớp của một điểm dữ liệu mới: Để dự đoán lớp của một điểm dữ liệu mới, kỹ thuật Naive-Bayes tính toán xác suất của mỗi lớp đối với điểm dữ liệu đó. Lớp có xác suất cao nhất sẽ được chọn là lớp của điểm dữ liệu mới.

#### 

#### 1.6.2.3. Kỹ thuật Cây Quyết Định.



**Hình 1.11. Đưa ra quyết định đi bộ hay đi xe bus**

*(Nguồn:*[*https://medium.com/analytics-vidhya/ml-algorithms-pros-cons-and-suitable-usages-b377c3c09f1b*](https://medium.com/analytics-vidhya/ml-algorithms-pros-cons-and-suitable-usages-b377c3c09f1b)*)*

* **Mục tiêu:** Xây dựng một cây quyết định để đưa ra quyết định dựa trên các thuộc tính.
* **Cách thức hoạt động [12]:**

**Bước 1:** Chia tập dữ liệu thành các phần con dựa trên các thuộc tính.

**Bước 2:** Lặp lại quá trình cho đến khi đạt được điều kiện dừng.

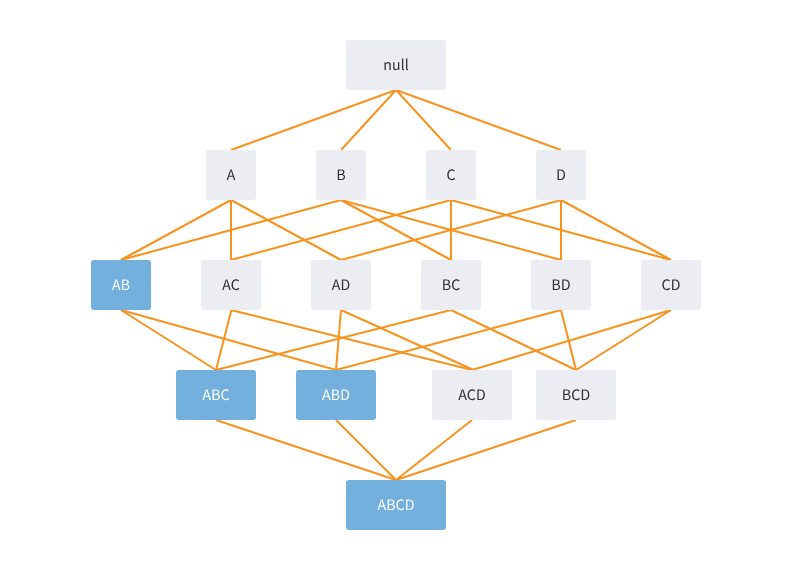
**Bước 3:** Cây có thể được sử dụng để phân loại hoặc dự đoán.

### 1.6.3. Kỹ thuật Khai Phá Luật Kết Hợp.

#### 1.6.3.1. Kỹ thuật Apriori.

* **Mục tiêu:** Khai thác các luật kết hợp từ tập dữ liệu.
* **Cách thức hoạt động [13]:**

**Bước 1:** Xác định tất cả các tập con phổ biến (có sự xuất hiện nhiều hơn ngưỡng).

**Bước 2:** Xây dựng các luật từ các tập con này và lọc bằng ngưỡng độ hỗ trợ và độ tin cậy.

**Hình 1.12. Kỹ thuật Apriori**

*(Nguồn:*[*https://loginom.com/blog/apriorialgorithm*](https://loginom.com/blog/apriorialgorithm)*)*

# Chương 2. Tiến hành khai phá dữ liệu

## 

## 2.1. Đặt vấn đề, xác định bài toán, mục tiêu:

Song song với sự đổi mới tích cực, không ngừng vận động phát triển của xã hội ngày nay thì cũng có những mặt tối không ngừng gia tăng, mà tiêu biểu trong số đó là vấn nạn ô nhiễm không khí, ô nhiễm môi trường, đặc biệt là ở các thành phố, đô thị, các khu công nghiệp đang phát triển mạnh. Vậy ô nhiễm không khí là gì?

Ô nhiễm không khí là sự ô nhiễm môi trường (cả trong nhà và ngoài trời) bởi bất kỳ tác nhân hóa học, vật lý, sinh học làm biến đổi tính tự nhiên của khí quyển. Là sự thay đổi lớn trong thành phần của không khí, chủ yếu do khói bụi, hơi hoặc các khí lạ được đưa vào không khí, có sự tỏa mùi, làm giảm tầm nhìn xa gây biến đổi khí hậu [14]. Các tác nhân cụ thể có thể kể đến như khói, bụi từ các thiết bị đốt trong gia đình, xe cơ giới, các cơ sở công nghiệp, cháy rừng. Tình trạng ô nhiễm môi trường ngày càng gia tăng không giảm, đặc biệt là ở các nước đang phát triển, ảnh hưởng rất lớn đến sức khỏe con người, gia tăng tỷ lệ mắc các loại bệnh về đường hô hấp và các loại bệnh khác. Trong không khí bấy giờ chứa các hạt vật chất, carbon monoxide, ozone, nitrogen, sulfur dioxide, …. không ngừng đi vào phổi của chúng ta [15].

Vậy vấn đề của chúng ta là thông qua bộ dữ liệu ô nhiễm không khí, tiến hành khai thác tri thức, khai phá dữ liệu, với mục tiêu tìm được các mối liên hệ, các yếu tố ảnh hưởng nhiều đến chất lượng không khí, ô nhiễm môi trường từ đó tìm hướng đi, đề ra phương hướng giải quyết. Áp dụng bộ dữ liệu vào các mô hình học máy với mục tiêu dự đoán chất lượng không khí có ảnh hưởng xấu tới sức khỏe không.

## 2.2. Xác định các công cụ và kỹ thuật áp dụng:

* **Công cụ:**
* **Python:** Là một ngôn ngữ lập trình mạnh mẽ và linh hoạt, thường được sử dụng trong lĩnh vực khoa học dữ liệu cũng như học máy. Ngôn ngữ này có một thư viện phong phú các gói phân tích dữ liệu như K-means, Linear Regression, Decision Tree, Naive-Bayes,... Ưu điểm của Python [16]:
  + Cú pháp đơn giản và dễ đọc, phù hợp cho người mới học lập trình và cũng dễ dàng cho những người có kinh nghiệm.
  + Python được sử dụng trong nhiều lĩnh vực như phát triển web, phân tích dữ liệu, trí tuệ nhân tạo, khoa học dữ liệu và nhiều ứng dụng khác. Điều này làm cho Python trở thành một ngôn ngữ lập trình linh hoạt và tiện lợi.
  + Cộng đồng Python khá lớn và luôn hoạt động tích cực nhằm cung cấp nhiều thư viện, framework hữu ích. Người dùng sẽ được nhận sự hỗ trợ từ cộng đồng thông qua tài liệu, diễn đàn và các nguồn thông tin trực tuyến.
  + Python cung cấp nhiều thư viện mạnh mẽ cho phân tích dữ liệu như Pandas, NumPy, và Matplotlib. Bộ ngôn ngữ giúp người dùng xử lý và thể hiện dữ liệu một cách hiệu quả.
  + Python có thể chạy trên nhiều hệ điều hành khác nhau và trở thành một trong những ngôn ngữ lập trình chéo phổ biến.
  + Ngôn ngữ Python là mã nguồn mở và miễn phí, cho phép người dùng tự do sử dụng, phân phối và thay đổi phiên bản của nó.
* **Google Colab:** Là một nền tảng tính toán đám mây cho phép viết và chạy mã Python trong trình duyệt. Nó có những ưu điểm như [17]:
  + **Không cần cấu hình:** Người dùng không cần phải cài Python và các thư viện liên quan trên máy tính của mình mà vẫn chạy được chương trình
  + **Quyền truy cập miễn phí vào GPU:** Google Colab cung cấp quyền truy cập miễn phí vào GPU mạnh mẽ của Google, giúp người dùng chạy mã phân tích dữ liệu nhanh hơn.
  + **Chia sẻ dễ dàng:** Chia sẻ các notebook Colab với người khác để tiến hành làm việc nhóm hiệu quả.
  + **Kết nối dữ liệu:** Có thể tải dữ liệu từ nguồn đám mây, chẳng hạn như Google Drive.
  + **Trực quan hóa dữ liệu:** Có thể trực quan hóa dữ liệu, chẳng hạn như tạo biểu đồ và đồ thị.

Vì những ưu điểm trên, cũng như hạn chế về phần cứng, Google Colab là sự lựa chọn tối ưu để thực hiện dự án này.

* **ydata-profiling:** Là một công cụ phân tích dữ liệu và khai phá dữ liệu chất lượng. Nó cung cấp một số tính năng hữu ích, chẳng hạn như phân tích thống kê, hình ảnh hóa, và phát hiện các vấn đề chất lượng dữ liệu. Ưu điểm [18]:
  + **Cách sử dụng đơn giản.**
  + **Thông tin chi tiết toàn diện trong báo cáo:** Báo cáo bao gồm nhiều số liệu thống kê và trực quan hóa, cung cấp cái nhìn toàn diện về dữ liệu.
  + **Đánh giá được chất lượng dữ liệu:** vượt trội trong việc xác định dữ liệu bị thiếu, các mục trùng lặp và các dữ liệu ngoại lệ. Những thông tin này rất cần thiết cho việc làm sạch và chuẩn bị dữ liệu, đảm bảo độ tin cậy cho kết quả phân tích.
  + **Dễ tích hợp với các luồng khác.**
  + **Khai phá dữ liệu cho các tập dữ liệu lớn:** ngay cả với tập dữ liệu có số lượng hàng lớn, vì ydata-profiling hỗ trợ cả Pandas Dataframes và Spark Dataframes.
* **Lazy Predict:** Là một thư viện Python cung cấp một cách dễ dàng để triển khai các mô hình học máy. Hữu ích để hiểu mối quan hệ giữa các biến và đưa ra dự đoán cho các biến mục tiêu. Ưu điểm [19]:
  + Tiết kiệm thời gian và so sánh nhanh chóng.
  + Dễ sử dụng.
  + Khai phá mô hình tiềm năng.
  + Tích hợp dễ dàng.
* **Kỹ thuật áp dụng:**
* **K-means:** là một thuật toán phân cụm (clustering) được sử dụng trong máy học và khai phá dữ liệu. Nó giúp phân loại một tập dữ liệu thành các nhóm (cụm) sao cho các điểm trong cùng một nhóm có tính chất tương tự và khác biệt so với các nhóm khác. Thuật toán K-means là một trong những phương pháp phân cụm phổ biến do đơn giản và hiệu quả [8].
  + **Thư viện hỗ trợ:** Để sử dụng k-means để phân cụm dữ liệu trong Python, bạn có thể sử dụng thư viện sklearn.cluster. Thư viện này cung cấp một lớp KMeans có thể được sử dụng để thực hiện thuật toán k-means [8].
  + **Chỉ định số lượng cụm:** Để sử dụng lớp KMeans, bạn cần chỉ định số lượng cụm bạn muốn tạo. Bạn cũng có thể chỉ định các tham số khác, chẳng hạn như phương pháp khởi tạo cụm và phương pháp tối ưu hóa [8].
* **Linear Regression (Hồi quy tuyến tính):** Hồi quy tuyến tính là một phương pháp thống kê được sử dụng để mô hình hóa mối quan hệ giữa một biến phụ thuộc (còn gọi là biến mục tiêu) và một hoặc nhiều biến độc lập (còn gọi là biến dự đoán). Mục tiêu của hồi quy tuyến tính là tìm ra một đường thẳng (hoặc mặt phẳng) phù hợp nhất với dữ liệu, cho phép chúng ta dự đoán giá trị của biến phụ thuộc dựa trên các giá trị của các biến độc lập [20].
  + **Ưu điểm của hồi quy tuyến tính [20]:**
    - Dễ hiểu và dễ thực hiện
    - Có thể được sử dụng cho nhiều loại dữ liệu khác nhau
    - Có nhiều công cụ thống kê hỗ trợ
  + **Nhược điểm của hồi quy tuyến tính [20]:**
    - Giả định rằng mối quan hệ giữa các biến là tuyến tính
    - Nhạy cảm với các ngoại lệ (outliers)
    - Có thể bị ảnh hưởng bởi hiện tượng đa cộng tuyến (multicollinearity)
  + **Ví dụ về sử dụng hồi quy tuyến tính:**
    - Giả sử có dữ liệu về giá và các biến độc lập sau:**Hình 2.1. Các biến độc lập**

Chúng ta có thể sử dụng hồi quy tuyến tính để tìm ra phương trình sau:



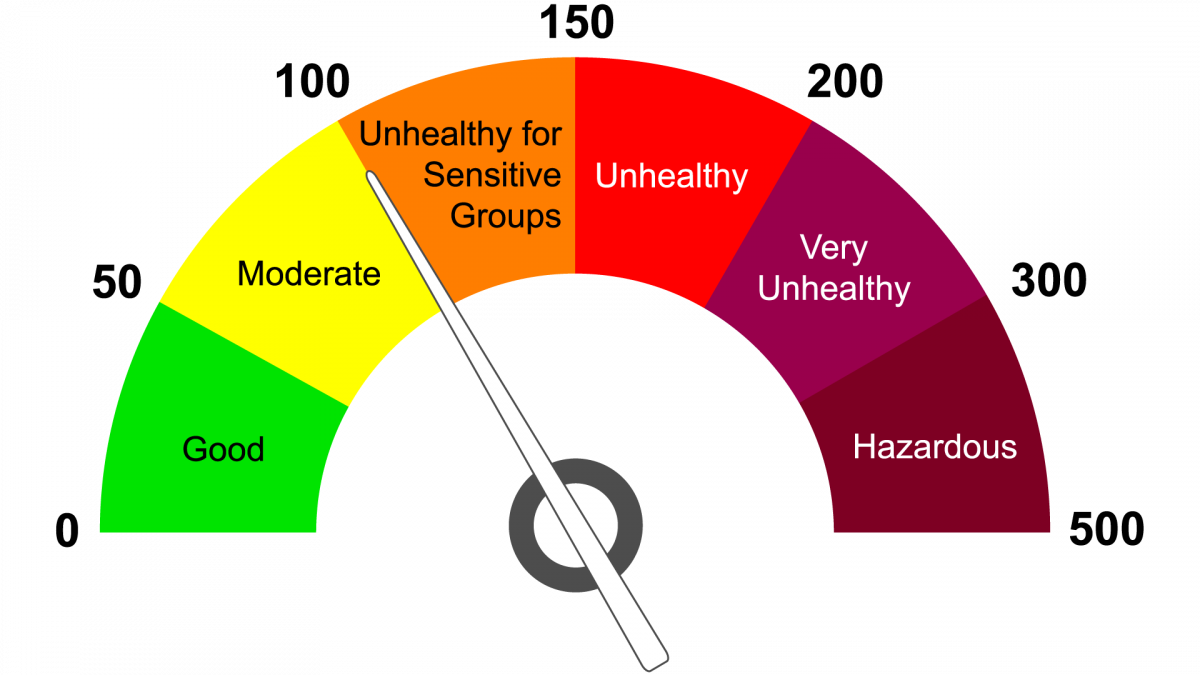
**Hình 2.2.****Phương trình hồi quy tuyến tính thể hiện mối quan hệ về giá và các biến độc lập**

Phương trình này cho biết rằng giá nhà sẽ tăng 50000 đô la cho mỗi đơn vị diện tích thêm, 20000 đô la cho mỗi phòng thêm, và 10000 đô la cho mỗi vị trí tốt hơn.

* **Decision Tree:** Cây quyết định là một thuật toán học máy có giám sát được sử dụng để phân loại và hồi quy. Nó hoạt động bằng cách xây dựng một cây, trong đó mỗi nút trong cây đại diện cho một quyết định và mỗi nhánh đại diện cho một kết quả có thể có của quyết định đó [12].
  + **Ưu điểm [12]:**
    - Dễ hiểu và giải thích.
    - Có thể xử lý cả dữ liệu số và dữ liệu phân loại.
    - Không yêu cầu chuẩn hóa dữ liệu.
    - Chịu được nhiễu trong dữ liệu.
    - Hiệu quả với các tập dữ liệu lớn.
  + **Nhược điểm [12]:**
    - Dễ bị quá khớp (overfitting), dẫn đến giảm hiệu suất trên dữ liệu mới.
    - Không phù hợp với các vấn đề có nhiều tương tác phức tạp giữa các thuộc tính.
    - Nhạy cảm với sự thay đổi nhỏ trong dữ liệu huấn luyện.
* **Grid Search:** Grid search là một kỹ thuật tối ưu hóa tham số được sử dụng trong học máy để tìm ra tập hợp tham số tốt nhất cho một mô hình học máy. Kỹ thuật này hoạt động bằng cách tạo ra một lưới các giá trị tham số khác nhau, sau đó đào tạo mô hình với mỗi tập hợp tham số. Cuối cùng, mô hình được đào tạo với tập hợp tham số có hiệu suất tốt nhất [21].
  + **Ưu điểm của grid search [21]**
    - Grid search là một cách hiệu quả để tìm ra tập hợp tham số tốt nhất cho một mô hình học máy.
    - Kỹ thuật này có thể được sử dụng với bất kỳ mô hình học máy nào.
    - Grid search có thể được sử dụng để tối ưu hóa nhiều tham số cùng một lúc.
  + **Nhược điểm của grid search [21]**
    - Grid search có thể tốn thời gian và tài nguyên, đặc biệt nếu có nhiều tham số cần tối ưu hóa.
    - Grid search có thể không tìm ra tập hợp tham số tối ưu, vì nó chỉ kiểm tra một tập hợp hữu hạn các giá trị tham số.
* **Confusion matrix:** Confusion Matrix là Đồ thị trực quan của 2 yếu tố Thực tế và Dự đoán. Nó đo lường hiệu suất của mô hình Classification techniques Machine Learning và trông giống như một cấu trúc giống bảng [22].
  + **Ưu điểm của Confusion Matrix [22]:**
    - Cung cấp cái nhìn tổng thể về hiệu suất của mô hình: Confusion matrix cung cấp thông tin về số lượng dự đoán đúng và sai cho từng lớp trong tập dữ liệu. Điều này giúp cho người dùng có thể đánh giá được hiệu suất của mô hình dự đoán trên từng lớp.
    - Cho phép điều chỉnh ngưỡng quyết định: Confusion matrix cho phép người dùng tùy chỉnh ngưỡng quyết định để điều chỉnh độ nhạy và độ chính xác của mô hình.
    - Giúp xác định vấn đề trong mô hình: Confusion matrix giúp phát hiện ra các vấn đề trong mô hình dự đoán, chẳng hạn như phân loại sai các điểm dữ liệu vào một lớp nhất định.
  + **Nhược điểm của Confusion Matrix [22]:**
    - Không đưa ra được thông tin về chất lượng của dữ liệu: Confusion matrix chỉ cung cấp thông tin về hiệu suất của mô hình trên tập dữ liệu đã có sẵn, không đưa ra được thông tin về chất lượng của dữ liệu.
    - Có thể bị ảnh hưởng bởi sự mất cân bằng trong số lượng mẫu: Nếu có sự mất cân bằng trong số lượng mẫu giữa các lớp, confusion matrix có thể bị ảnh hưởng và không phản ánh đúng hiệu suất của mô hình.
    - Không phù hợp với các bài toán với nhiều lớp: Confusion matrix chỉ phù hợp với các bài toán phân loại hai lớp hoặc phân loại đa lớp đơn giản. Khi có nhiều hơn hai lớp, confusion matrix trở nên khó tính toán và khó hiểu.
* **K-nearest-neighbors (KNN):** KNN là một thuật toán học máy có giám sát được sử dụng cho các tác vụ phân loại. Đây là một thuật toán đơn giản nhưng mạnh mẽ dựa trên ý tưởng về các hàng xóm gần nhất [10].
  + **Ưu điểm [10]:**
    - Đơn giản và dễ hiểu: KNN là một trong những thuật toán phân loại đơn giản nhất, giúp dễ dàng giải thích và triển khai.
    - Không tham số: Nó không đưa ra bất kỳ giả định nào về phân phối dữ liệu cơ bản, khiến nó phù hợp với nhiều loại vấn đề.
    - Hiệu quả cho các tập dữ liệu nhỏ đến vừa: Nó thường hoạt động tốt trên các tập dữ liệu không quá lớn hoặc phức tạp.
  + **Nhược điểm [10]:**
    - Tốn kém về mặt tính toán: Tìm hàng xóm gần nhất có thể tốn nhiều thời gian tính toán, đặc biệt là đối với các tập dữ liệu lớn hoặc không gian tính năng nhiều chiều.
    - Nhạy cảm với nhiễu và ngoại lệ: Ngoại lệ hoặc điểm dữ liệu nhiễu có thể ảnh hưởng đáng kể đến ranh giới quyết định của thuật toán.
* **Apriori:** Thuật toán Apriori là một thuật toán tìm luật kết hợp phổ biến.
  + **Ưu điểm [13]:**
    - Đơn giản và dễ hiểu: Thuật toán Apriori có cách thức hoạt động đơn giản và dễ hiểu, vì vậy nó dễ dàng được triển khai và sử dụng.
    - Hiệu quả: Thuật toán Apriori có thể tìm thấy các luật kết hợp hiệu quả, ngay cả đối với các bộ dữ liệu lớn.
    - Tính linh hoạt: Thuật toán Apriori có thể được sử dụng để tìm các luật kết hợp với các độ hỗ trợ và độ tin cậy khác nhau.
  + **Nhược điểm [13]:**
    - Có thể tạo ra nhiều luật kết hợp không có ý nghĩa: Thuật toán Apriori tạo ra tất cả các luật kết hợp có độ hỗ trợ tối thiểu, bao gồm cả các luật kết hợp không có ý nghĩa hoặc không đáng quan tâm.
    - Thời gian chạy có thể lâu: Thuật toán Apriori có thể mất nhiều thời gian để chạy cho các bộ dữ liệu lớn.

## 2.3. Tổng quan bộ dữ liệu:

* Nguồn bộ dữ liệu: [Global Air Pollution Dataset (kaggle.com)](https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset/data)
* Type: file csv
* Mô tả: bộ dữ liệu cung cấp thông tin về các chất có ảnh hưởng đến tình trạng ô nhiễm không khí tại các nơi trên thế giới:
* **Nitrogen Dioxide [NO2] :** là một trong số các oxit nito. Thông thường chúng được tạo ra từ các hiện tượng tự nhiên như sự xâm nhập từ tầng bình lưu hoặc ánh sáng. Nhưng trong xã hội công nghiệp phát triển ngày nay, ở cấp độ bề mặt chúng được hình thành khí thải ô tô, xe tải hoặc xe buýt, nhà máy và các thiết bị điện. Tiếp xúc với NO2 trong thời gian dài có thể dẫn đến các bệnh hô hấp như hen suyễn, lâu dần sẽ bị nhiễm trùng đường hô hấp. Thường thì những người mắc bệnh hen suyễn và trẻ em sẽ có nguy cơ bị ảnh hưởng bởi NO2 cao hơn [23].
* **Ozone [O3] :** Ozone được cấu tạo từ 3 nguyên tử oxi, không màu, tập trung số lượng lớn có mùi hôi; có tính oxi hóa mạnh các hợp chất hữu cơ. Tầng ozone ở trong khí quyển là tầng bình lưu, có lợi đối với con người, có tác dụng lọc các tia cực tím từ mặt trời và làm giảm bức xạ, bảo vệ sự sống trên trái đất. Còn ô nhiễm Ozone ở mặt đất lại có ảnh hưởng rất lớn đến sức khỏe của con người, gây tổn thương các tế bào mô lót của đường hô hấp. Gây ra viêm phổi, ho, ngứa họng, làm giảm chức năng hô hấp, giảm sức đề kháng của cơ thể, gia tăng các loại bệnh về đường hô hấp. Ngoài ra ô nhiễm ozone ở mặt đất còn làm ảnh hưởng đến thảm thực vật và hệ sinh thái, làm hỏng thảm thực vật nhạy cảm trong mùa sinh trưởng [23].
* **Carbon Monoxide [CO] :** là loại khí không màu, không mùi. Thường được tạo ra bởi các phương tiện, nhà máy sử dụng nhiên liệu hóa thạch, các vật dụng như dầu hỏa, máy sưởi gas, bếp gas,…CO là một trong những nguyên nhân chính dẫn đến tử vong do ngạt khí trong các vụ cháy. Thông thường thì ít xảy ra ngoài trời nhưng trong không gian kín thì lại vô cùng nguy hiểm, gây chóng mặt, mất ý thức tạm thời, bất tỉnh và nghiêm trọng nhất là tử vong [23].
* **Particulate Matter [PM2.5] :** Là các hạt bụi mịn trong không khí có kích thước từ 2.5 micromet, là hỗn hợp phức tạp của nhiều vật chất rắn khác nhau và có cả các hạt chất lỏng. Nếu hít phải có thể gây ra các vấn đề về đường hô hấp và tim mạch như ho, viêm họng, viêm phổi, viêm phế quản, hen suyễn, bệnh phổi tắc nghẽn mãn tính,…Ngoài ra chúng còn có thể xâm nhập vào tủy xương và gậy tổn thương cho các mạch máu. Chúng được cơ quan nghiên cứu Ung thư Quốc tế (IARC) phân loại, liệt vào nhóm chất gây ung thư nhóm 1 [23].
* **AQI [Air Quality Index]:** Chỉ số báo cáo, thang đo về chất lượng không khí hàng ngày [24].



**Hình 2.3. Chỉ số chất lượng không khí (AQI)**

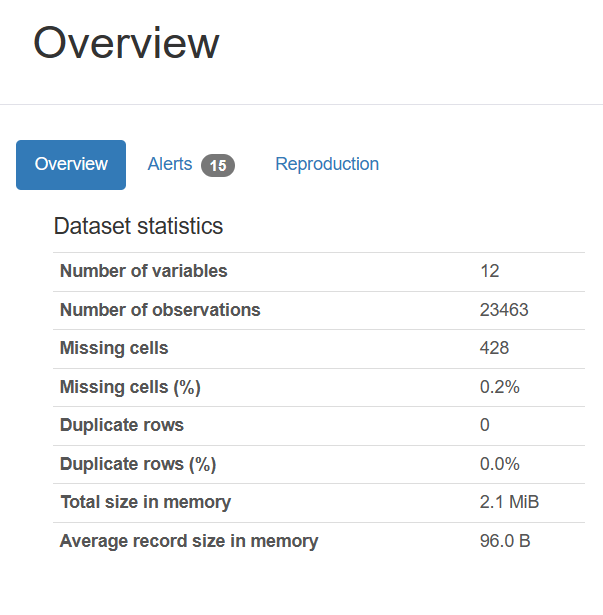
*(Nguồn:* [*http://danapha.com/vi/tu-van-chi-tiet/viem-xoang-ngua-*](http://danapha.com/vi/tu-van-chi-tiet/viem-xoang-ngua-mui-hat-hoi-co-vixolis-voi-di-muon-phien)

[*mui-hat-hoi-co-vixolis-voi-di-muon-phien*](http://danapha.com/vi/tu-van-chi-tiet/viem-xoang-ngua-mui-hat-hoi-co-vixolis-voi-di-muon-phien)*)*

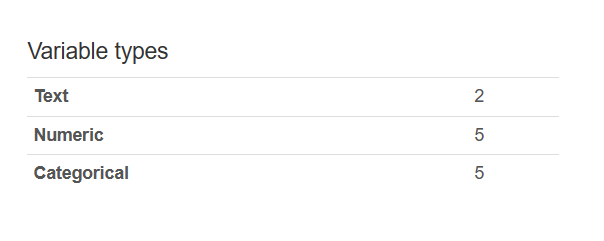
Các trường thuộc tính của bộ dữ liệu [24]:

* Country: Tên quốc gia
* City: Tên thành phố của quốc gia đó
* AQI Value: Giá trị AQI chung của thành phố đó. Được tính bằng cách lấy giá trị AQI thành phần cao nhất.
* AQI Category: Xếp loại AQI của thành phố đó.
* CO AQI Value: Chỉ số AQI của Carbon monoxide.
* CO AQI Category: Xếp loại AQI của Carbon monoxide.
* Ozone AQI Value: Chỉ số AQI của Ozone.
* Ozone AQI Category: Xếp loại AQI của Ozone.
* NO2 AQI Value: Chỉ số AQI của NO2.
* NO2 AQI Category: Xếp loại AQI của NO2.
* PM2.5 AQI Value: Chỉ số AQI các hạt bụi mịn có kích thước 2.5 micromet.
* PM2.5 AQI Category: Xếp loại AQI các hạt bụi mịn có kích thước 2.5 micromet.

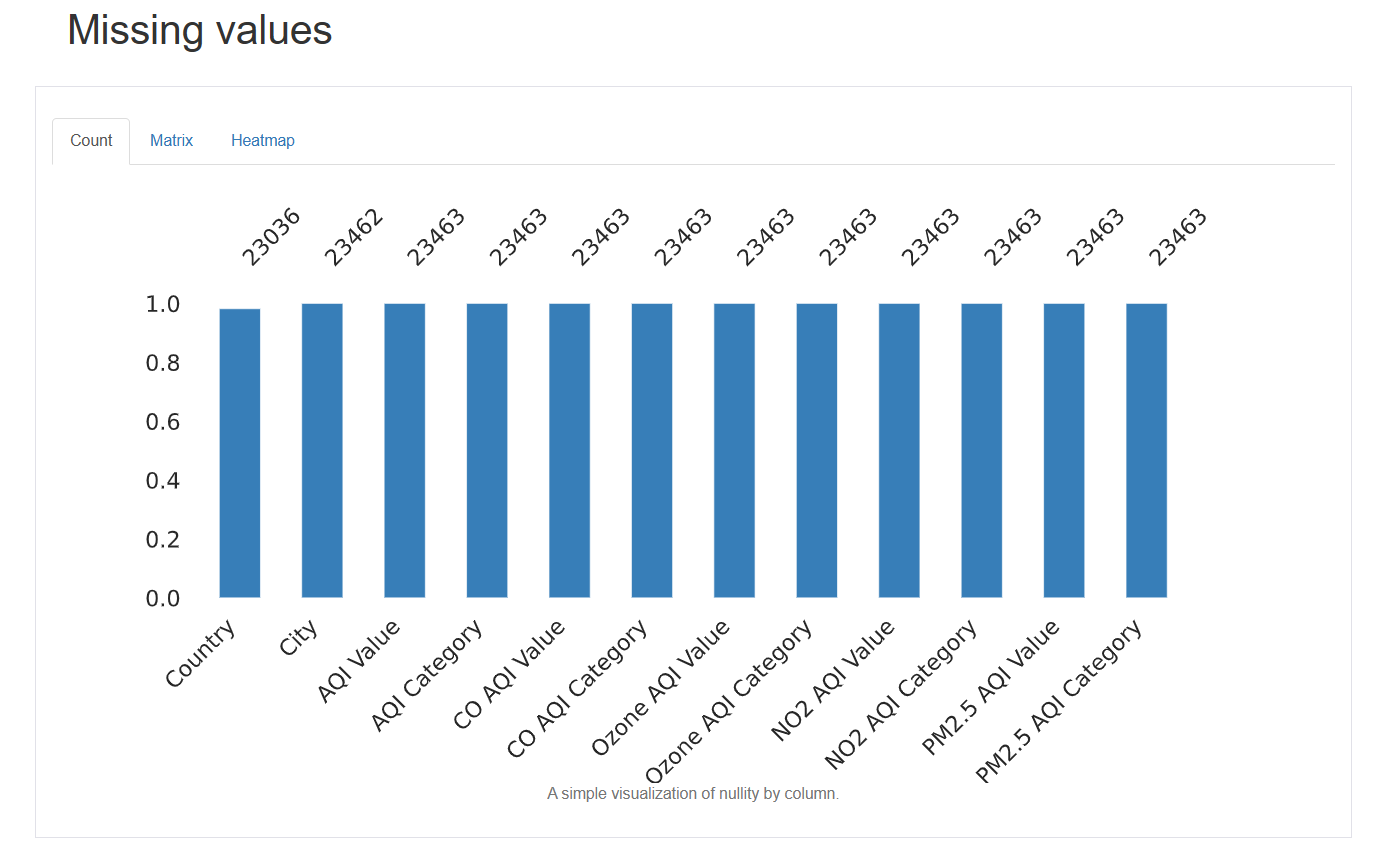
=> Vậy từ đây ta có thể xác định mục tiêu của chúng ta khi sử dụng mô hình dự đoán, phân cụm là AQI Category và AQI Value. Với mô hình dự đoán tuyến tính AQI Value ta sẽ sử dụng các thuộc tính dạng số. Với mô hình dạng đánh nhãn, phân cụm, dự đoán cụm AQI Category ta có thể sử dụng các thuộc tính dạng số và nhãn (chữ), tuy nhiên không thể sử dụng cả hai, vì các thuộc tính có dạng nhãn phân loại đều được suy ra từ các thuộc tính dạng số.

**Tổng quan các thông số của bộ dữ liệu:**

**Hình 2.4. Tổng quan bộ dữ liệu**



**Hình 2.5. Các loại biến**



**Hình 2.6. Dữ liệu thiếu trong bộ dữ liệu**

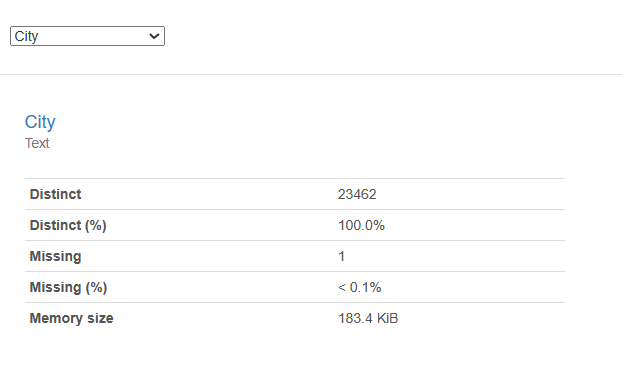
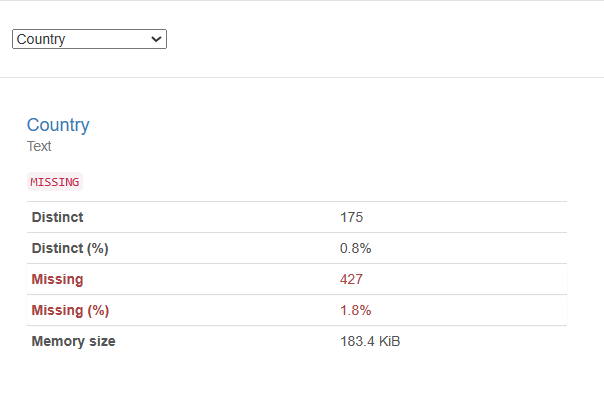
Số lượng thực thể thu được là 23463.

Trong đó:

* Có 428 thực thể không đầy đủ các thuộc tính, chiếm 0.2%.
* Không có thực thể nào lặp lại giá trị.
* Tổng kích thước bộ dữ liệu trong bộ nhớ là: 2.1 MiB.

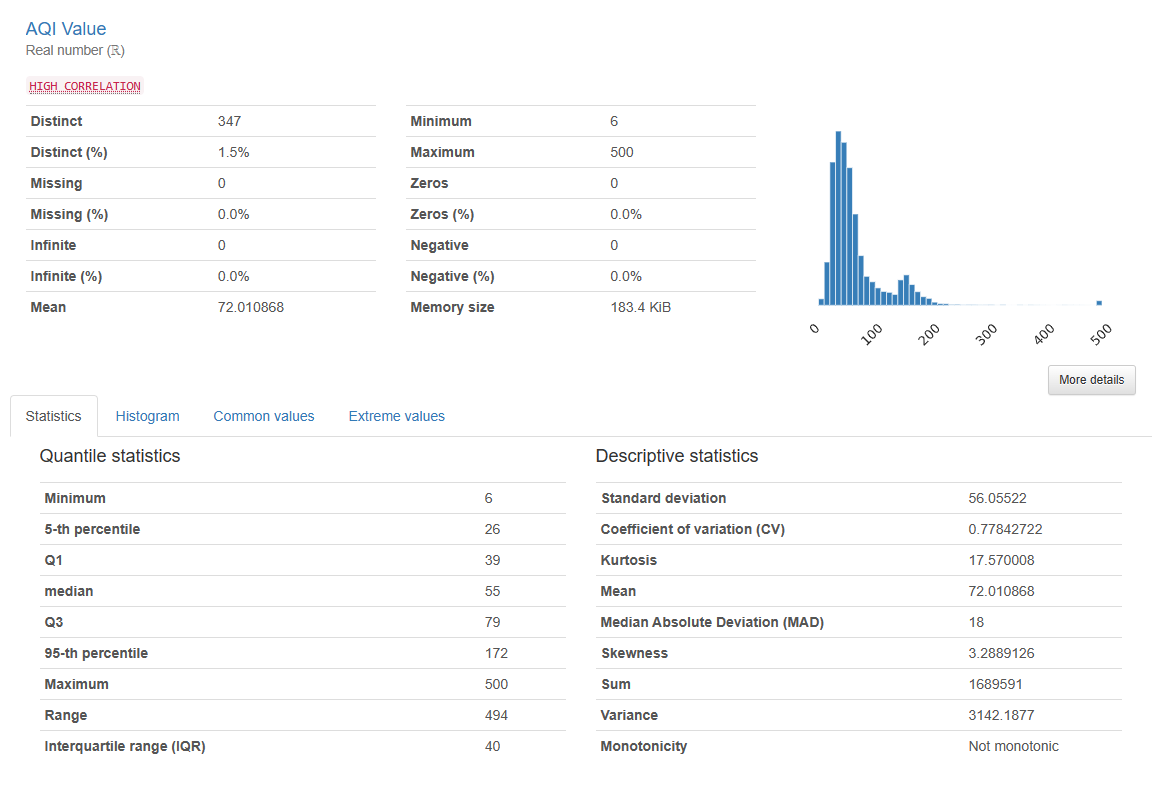
**Phân tích các thuộc tính bộ dữ liệu:**

**Thuộc tính ‘Country’ và ‘City’:**

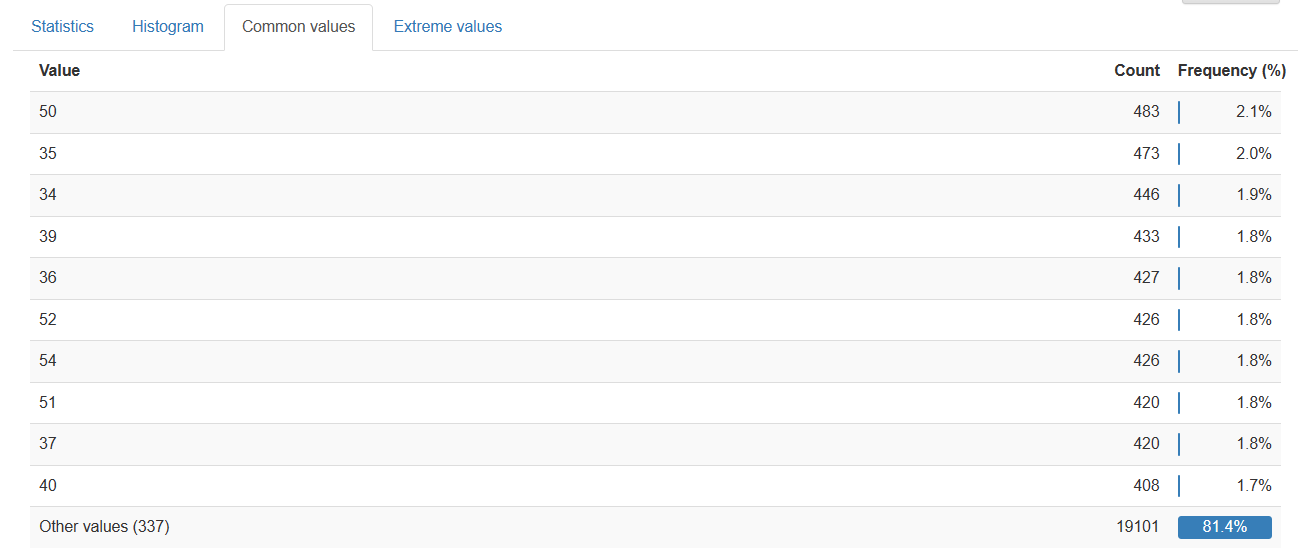
**Hình 2.7. Thuộc tính Country**

**Hình 2.8. Thuộc tính City**

\*Nhận xét: 428 thực thể có giá trị bị thiếu bao gồm 427 trường trường hợp thiếu giá trị Country và 1 trường hợp thiếu giá trị City

**Các chỉ số AQI:**

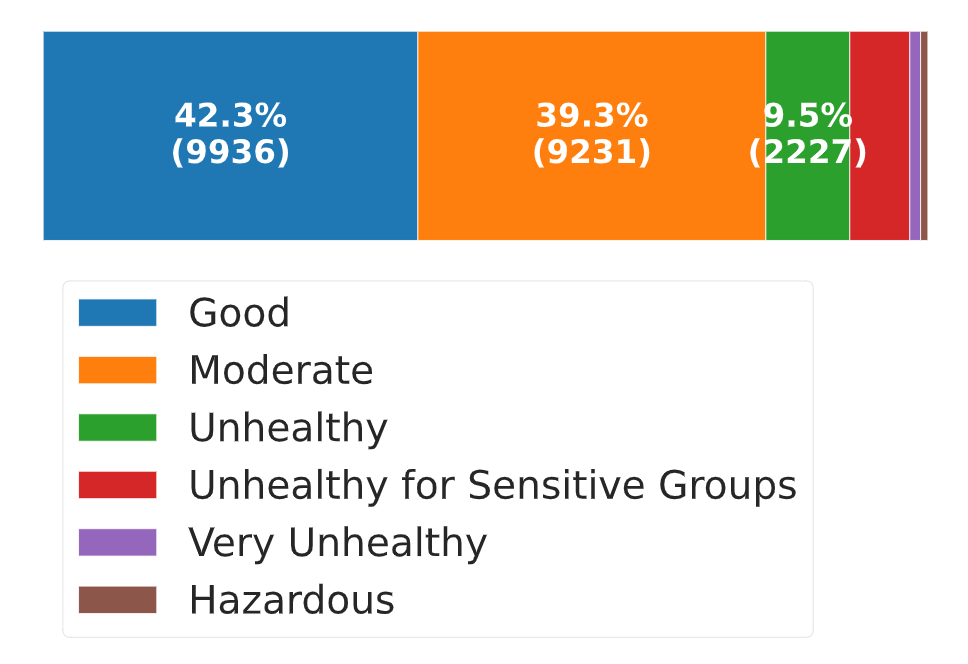
**Hình 2.9.a. AQI Value**



**Hình 2.9.b. AQI Value**

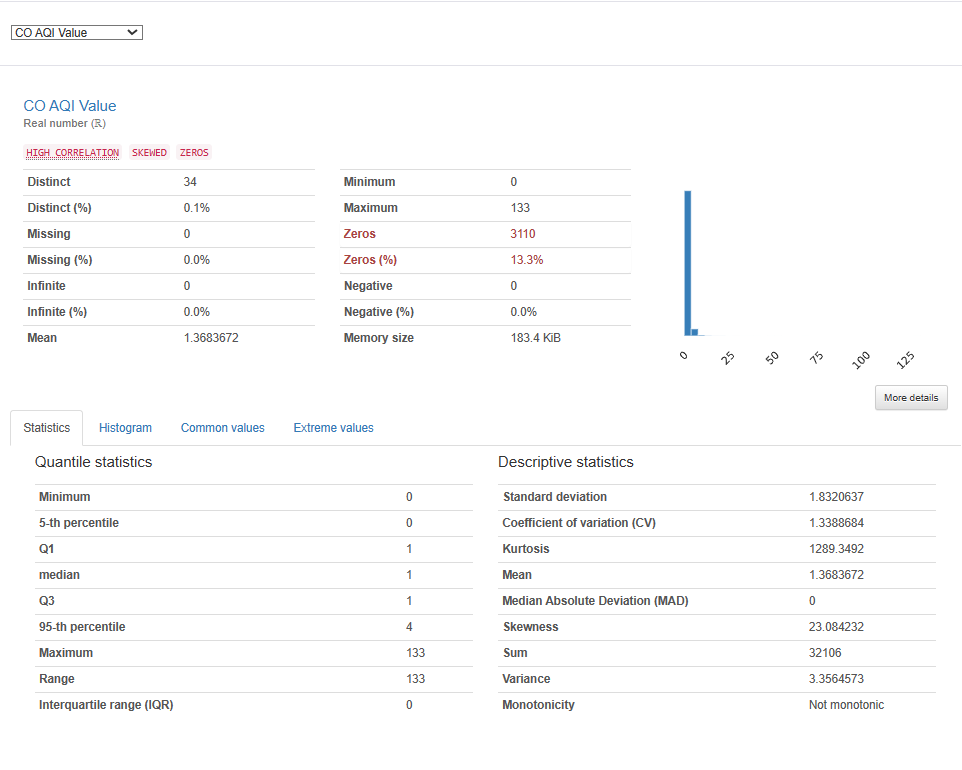
\*Nhận xét:Với thuộc tính AQI Value, dữ liệu tập trung nhiều trong khoảng từ 0 đến 100. Giá trị xuất hiện nhiều nhất là 50 (2.1%). Chất lượng không khí ở mặt bằng chung vẫn duy trì ở mức tạm ổn đến tốt.



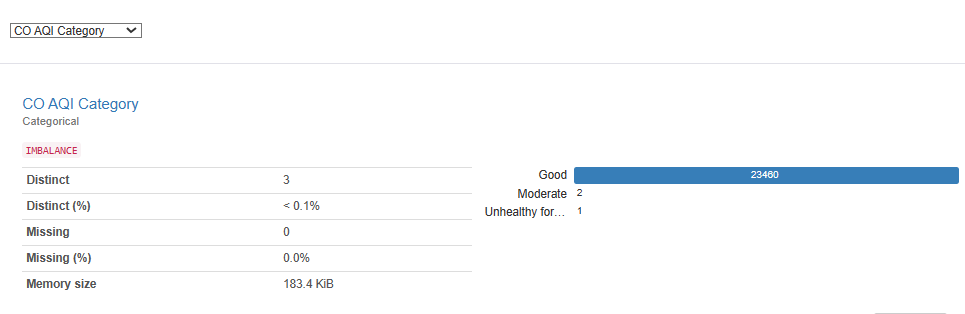
**Hình 2.10.a. AQI Category**

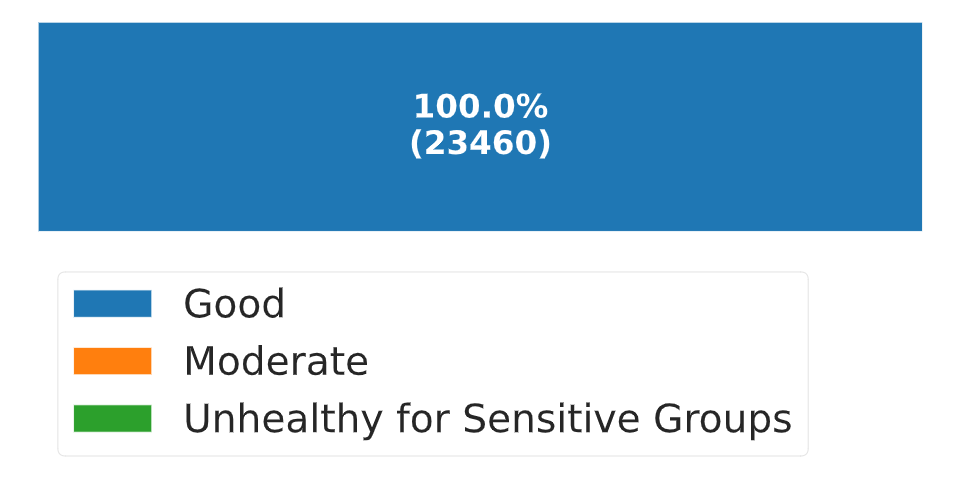
**Hình 2.10.b. AQI Category**

\*Nhận xét: với thuộc tính AQI Category, chất lượng không khí nói chung ở mức tạm ổn đến tốt chiếm đa số.

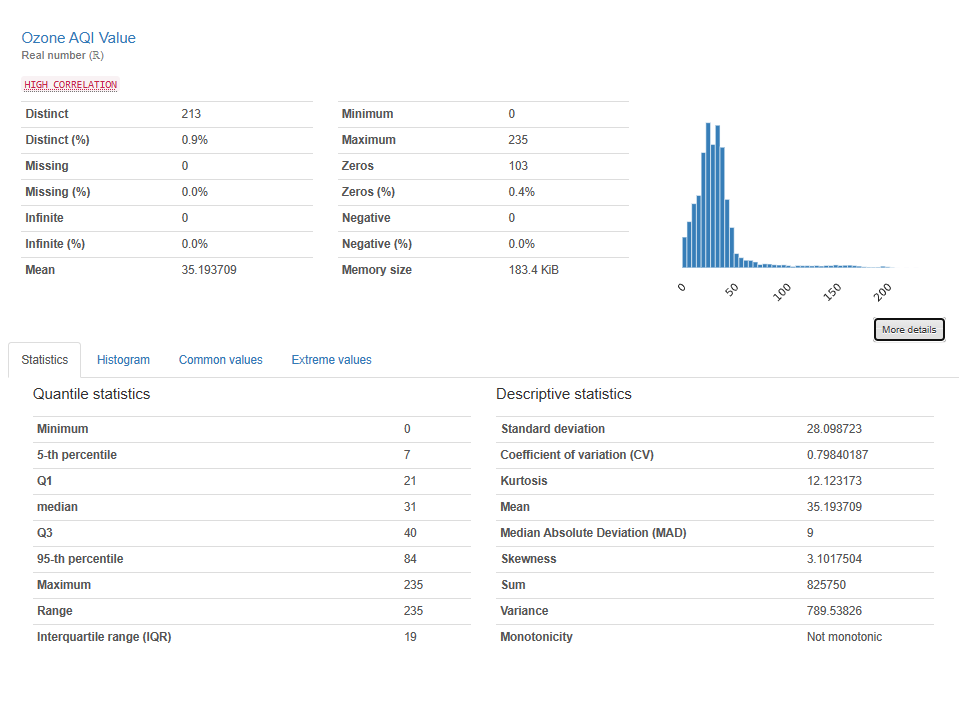
**Hình 2.11.a. Chỉ số CO AQI Value**

\*Nhận xét: với thuộc tính CO AQI Value, chỉ số tương đối thấp, điều này cũng đúng với thực tế vì CO thông thường không dễ hình thành với số lượng lớn, chỉ khi các vụ cháy nghiêm trọng xảy ra như cháy rừng, cháy nhà, các cơ sở thì lúc này lượng CO sản sinh ra mới tương đối đáng kể.

**Hình 2.11.a. Chỉ số CO AQI Category**

**Hình 2.11.b. Chỉ số CO AQI Category**

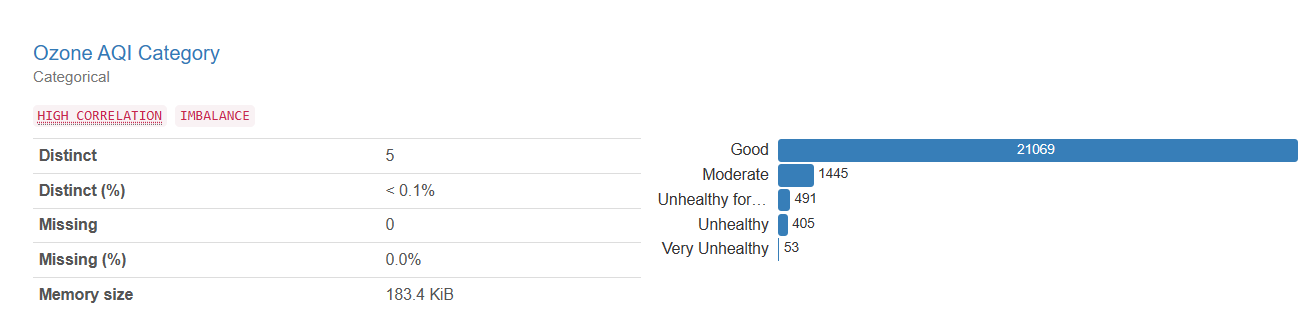
\*Nhận xét: với thuộc tính CO AQI Category, như một cách phân phân loại, phản ánh lại thuộc tính CO AQI Value, dữ liệu đạt gần như 100% mức độ an toàn với nồng độ CO trong không khí. Lượng CO sản sinh ra quá nhỏ, hầu như không tác động nhiều đến việc ô nhiễm không khí, ảnh hưởng sức khỏe. Tuy nhiên ta cũng có thể thấy với bộ dữ liệu này mô hình có thể sẽ không tổng quát được trường hợp nồng độ CO ảnh hưởng đến sức khỏe.

**Hình 2.12.a. Chỉ số Ozone AQI Value**

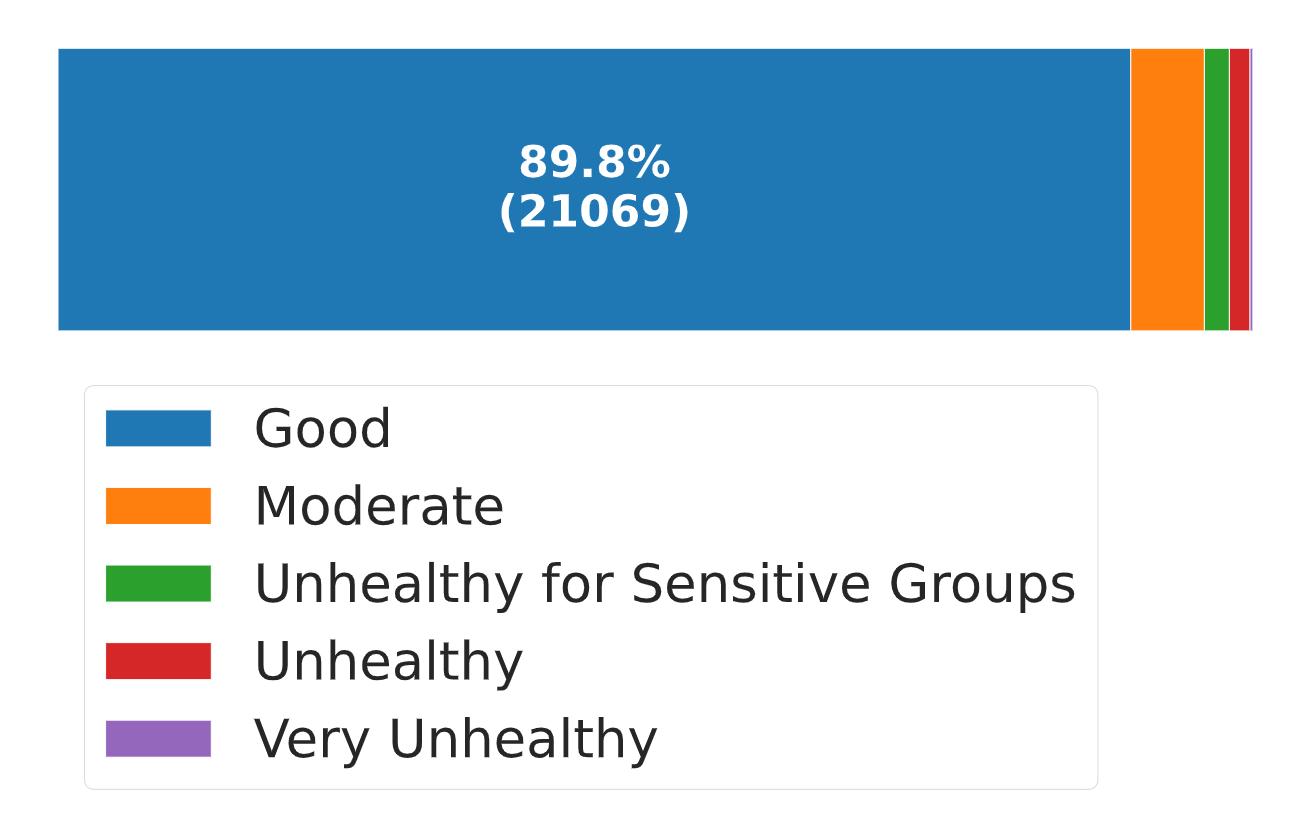


**Hình 2.12.b. Chỉ số Ozone AQI Value**

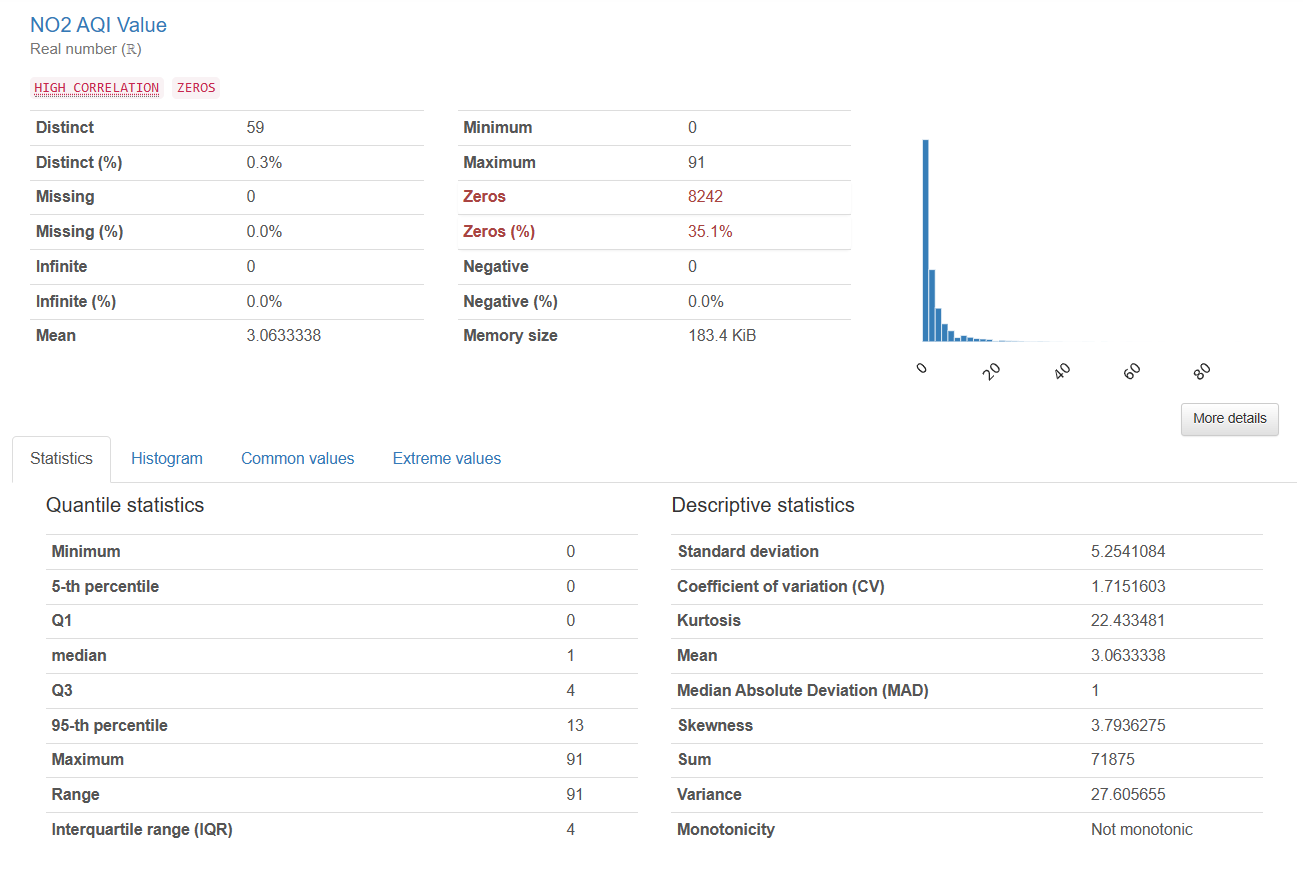
\*Nhận xét: với thuộc tính Ozone, lượng khí Ozone trong không khí đa số rơi vào khoảng từ 0 đến 50, vẫn nằm trong ngưỡng an toàn đối với con người, chỉ có một vài trường hợp nhỏ số lượng vượt quá 50 và khi vượt quá 100 ảnh hưởng đến sức khỏe con người.



**Hình 2.13.a. Chỉ số Ozone AQI Category**

**Hình 2.13.b. Chỉ số Ozone AQI Category**

\*Nhận xét: với thuộc tính Ozone AQI Category, đa số vẫn nằm ở mức ổn cho đến tốt đối với sức khỏe con người.

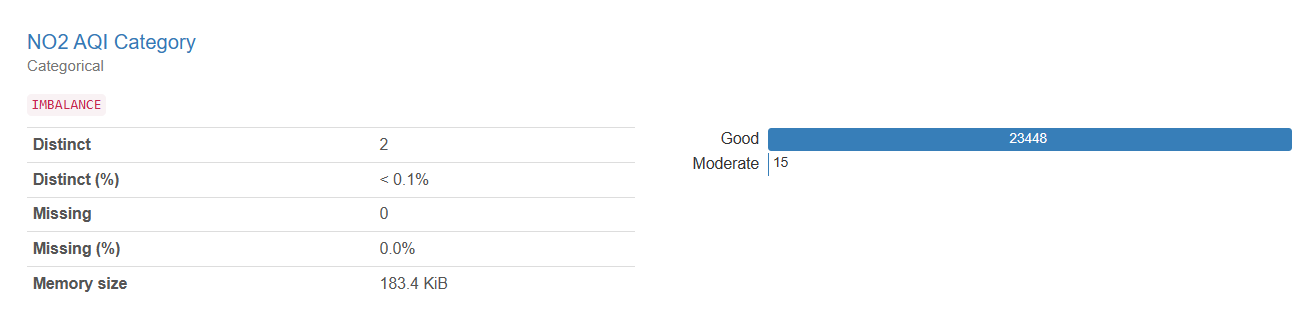


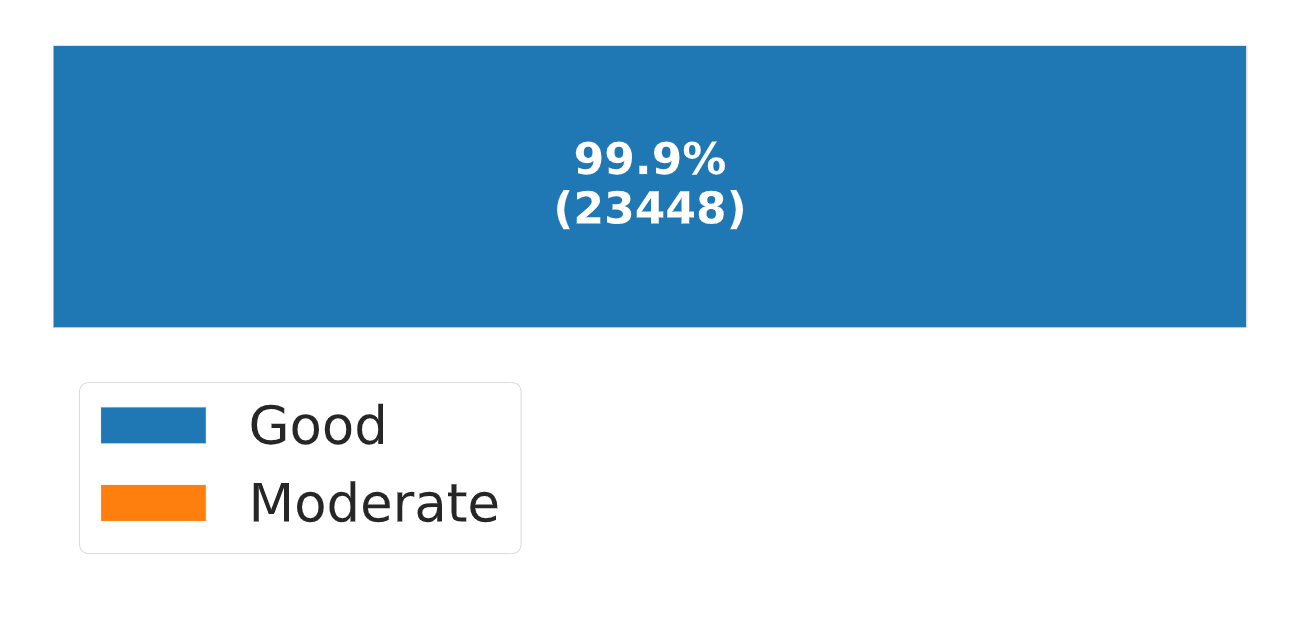
**Hình 2.14.a. Chỉ số NO2 AQI Value**



**Hình 2.14.b. Chỉ số NO2 AQI Value**

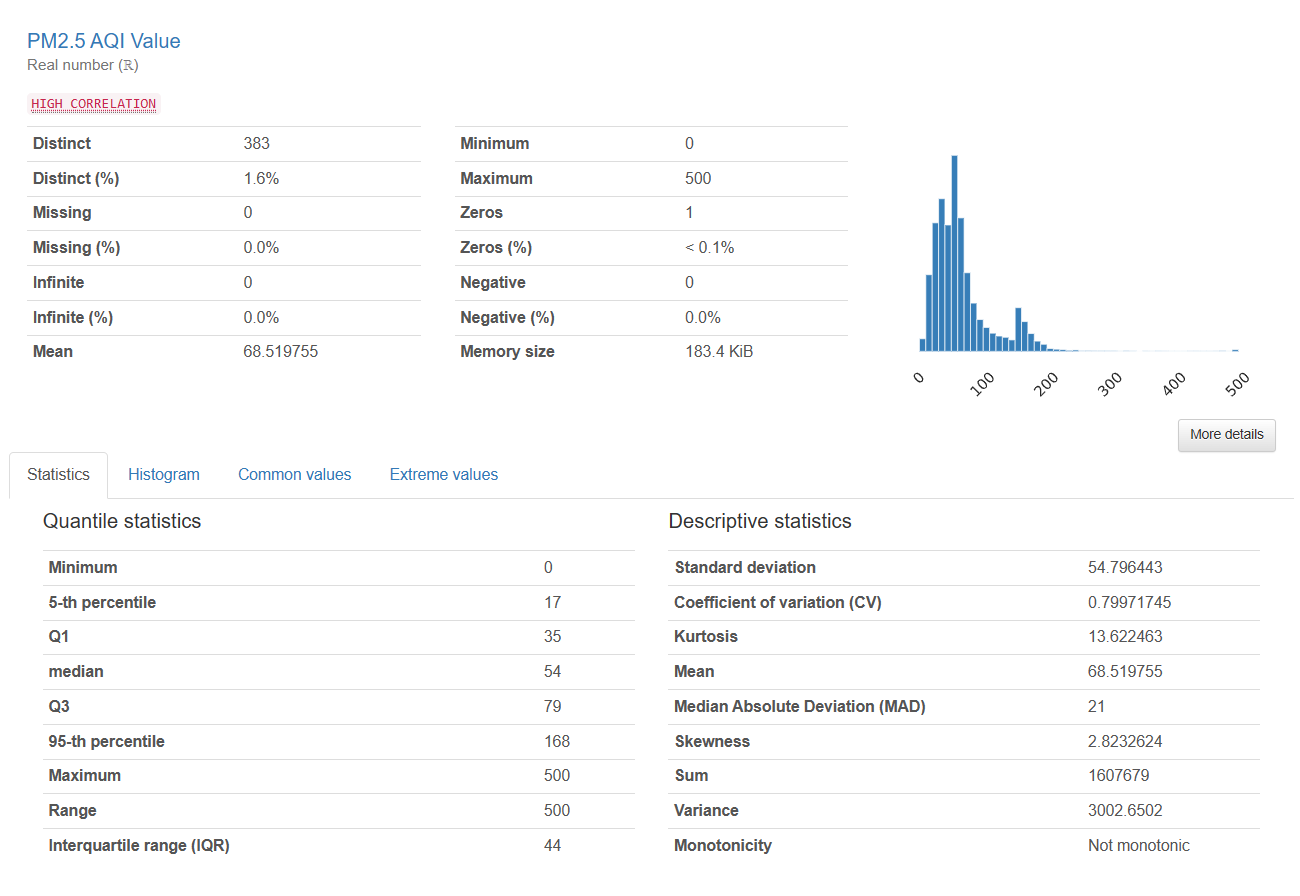
\*Nhận xét:với NO2 AQI Value, qua biểu đồ ta có thể thấy rõ NO2 gần như đạt mức an toàn tuyệt đối, giá trị AQI rất thấp.



**Hình 2.15.a. Chỉ số NO2 AQI Category**

**Hình 2.15.b. Chỉ số NO2 AQI Category**

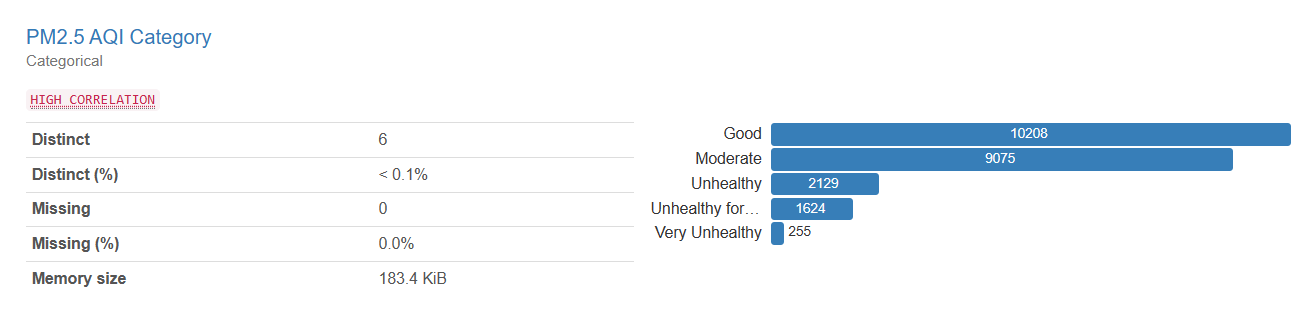
\*Nhận xét: NO2 AQI Category, dữ liệu cho thấy NO2 không hề có tác động xấu gì đối với sức khỏe con người tại các địa phương trong bộ dữ liệu.

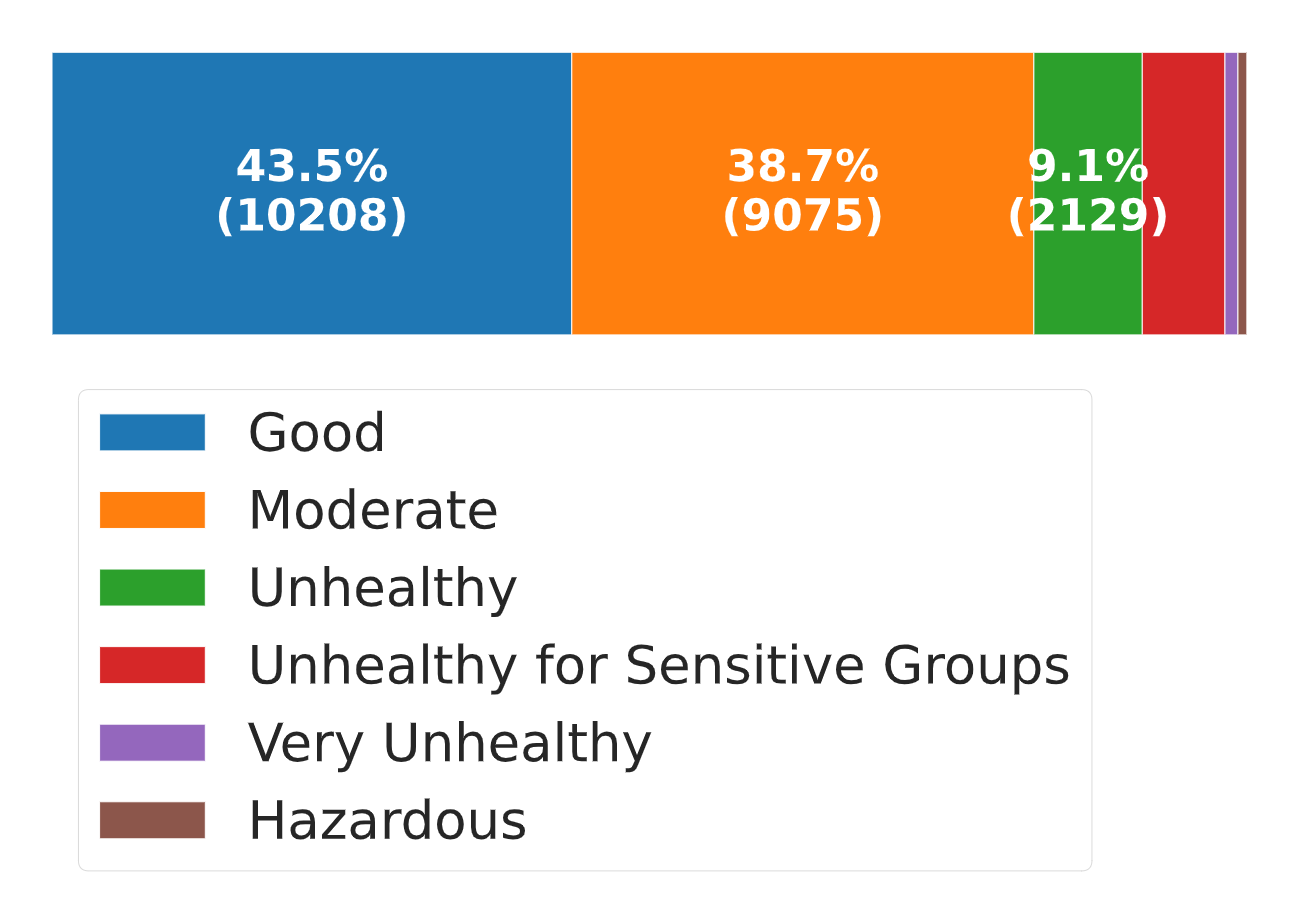
**Hình 2.16.a. Chỉ số PM2.5 AQI Value**



**Hình 2.16.b. Chỉ số PM2.5 AQI Value**

\*Nhận xét: với thuộc tính PM2.5 AQI Value, chỉ số bụi mịn kích thước 2.5 micromet xuất hiện nhiều trong khoảng từ 20 đến 80. Tuy nhiên số lượng giá trị trên 100 cũng đáng kể, có thể nói là nhiều hơn các thuộc tính khác, mức độ ảnh hưởng sức khỏe cao hơn.



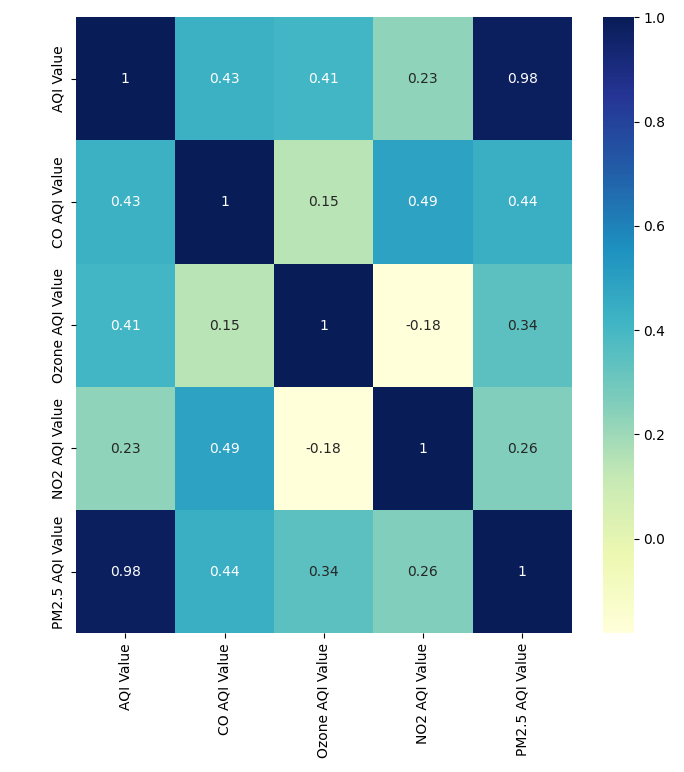
**Hình 2.17.a. Chỉ số PM2.5 AQI Category**

**Hình 2.17.b. Chỉ số PM2.5 AQI Category**

\*Nhận xét:với thuộc tính PM2.5 AQI Category: đúng như những gì ta đã thấy rõ và nhận xét trên biểu đồ cột về chỉ số PM2.5 AQI Value, số lượng ảnh hưởng xấu tới sức khỏe chiếm tới 17.8 %.

**Nhận xét tổng quan:** bộ dữ liệu có tình trạng mất cân bằng nghiêm trọng về số lượng các nhãn mức độ AQI ảnh hưởng tới sức khỏe con người, số lượng nhãn ở mức tạm ổn đến tốt chiếm đa số, đúng với tình trạng thực tế giữa các nước.

**Xét mối quan hệ tương quan giữa các trường thuộc tính:**



**Hình 2.18. Mối tương quan giữa các thuộc tính**

Ta có thể thấy rõ chỉ số bụi mịn (PM2.5 AQI Value) có ảnh hưởng nhiều nhất đến chỉ số AQI chung (AQI Value), tiếp đến là Ozone AQI Value và NO2 AQI Value.

## 2.4. Tiền xử lý dữ liệu

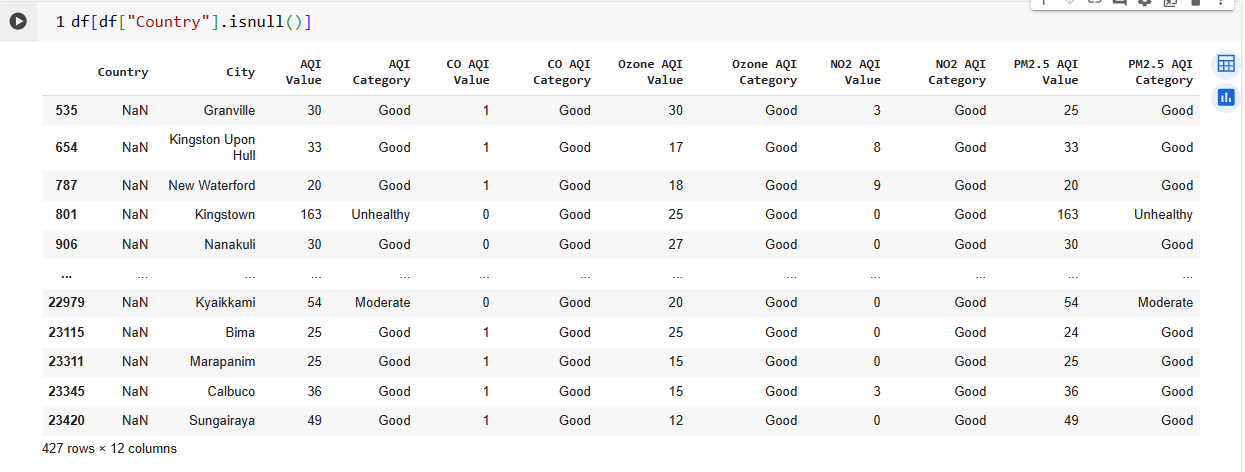
### 2.4.1. Xử lý dữ liệu bị thiếu

Như đã đề cập ở trên, 2 thuộc tính có dữ liệu bị thiếu là Country và City, số lượng thực thể bị thiếu dữ liệu là 428, chiếm 0.2%.

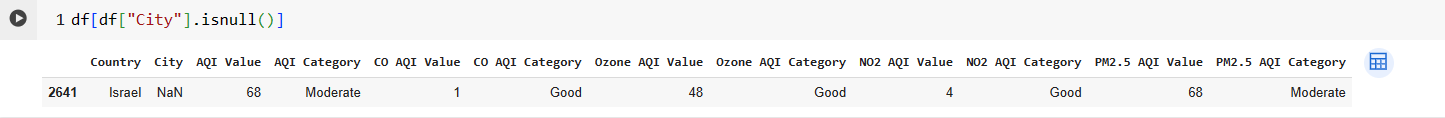
Có 2 phương pháp xử lý dữ liệu bị thiếu là:

* Loại bỏ missing values: thường được sử dụng trong trường hợp các missing values không quá quan trọng hoặc số lượng quá ít, không đáng kể (<5%).
* Thay thế missing values: thay thế các missing values đó bằng một giá trị khác. Với dữ liệu dạng số ta có thể thay bằng 0, median, mode. average, … Với dữ liệu dạng chữ, phân loại ta có thể nhóm chúng thành một nhóm có tên là “missing”.

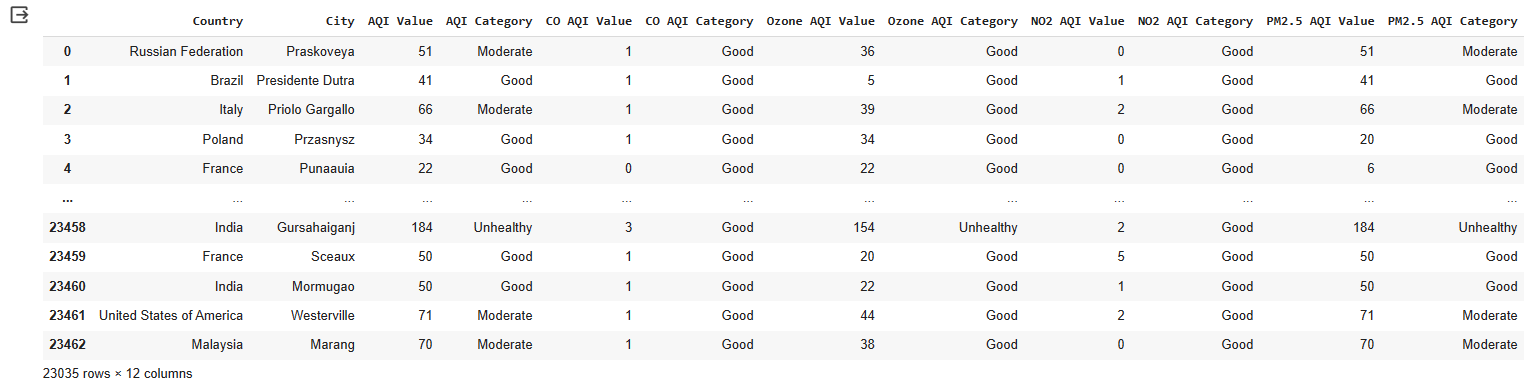
\* Đối với bộ dữ liệu như trên, chúng em đề xuất sử dụng phương pháp loại bỏ missing values.

**Các thực thể không có thuộc tính Country:**

**Hình 2.19.a. Thiếu dữ liệu Country**

**Các thực thể không có giá trị City**

**Hình 2.19.b. Thiếu dữ liệu City**

Tiến hành loại bỏ các thực thể thiếu giá trị, số lượng thực thể còn lại là 23035 thực thể:

**Hình 2.19.c. Số lượng thực thể còn lại sau khi loại bỏ thực thể thiếu giá trị**

### 2.4.2. Biến đổi dữ liệu:

Mã hóa các thuộc tính phân loại thành dạng số bằng phương pháp Label Encoding, tuy nhiên chúng ta sẽ không sử dụng LabelEncoding() của sklearn. Lý do, LabelEncoding() của sklearn mã hóa theo thứ tự bảng chữ cái, do đó sẽ không biết được độ lớn, thứ tự xếp hạng giá trị của các biến. Thay vào đó ta sẽ tự mã hóa nhãn thủ công.

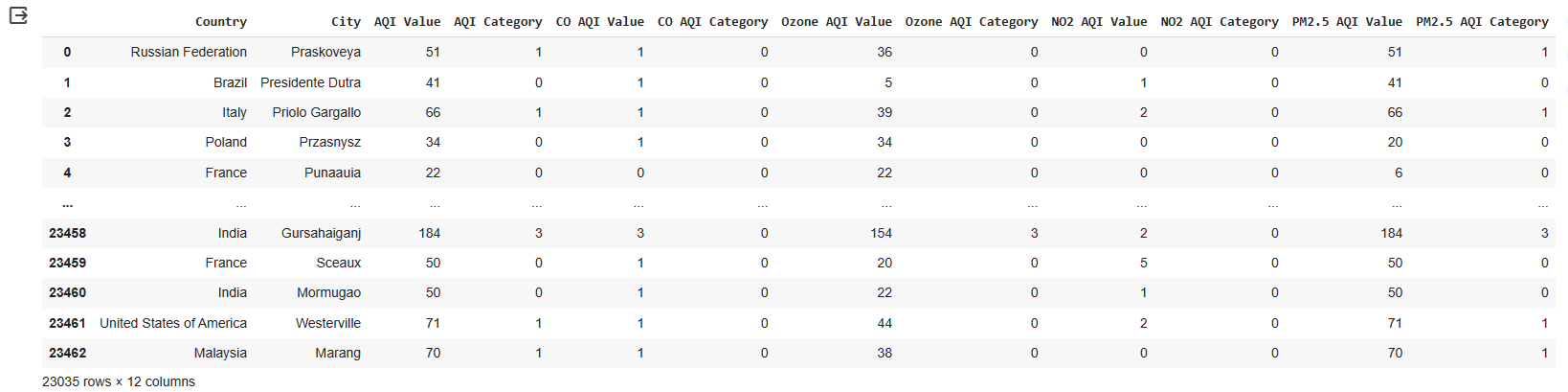
Giá trị tương đương của các biến:

* Good: 0
* Moderate: 1
* Unhealthy for Sensitive Groups: 2
* Unhealthy: 3
* Very Unhealthy: 4
* Hazardous: 5

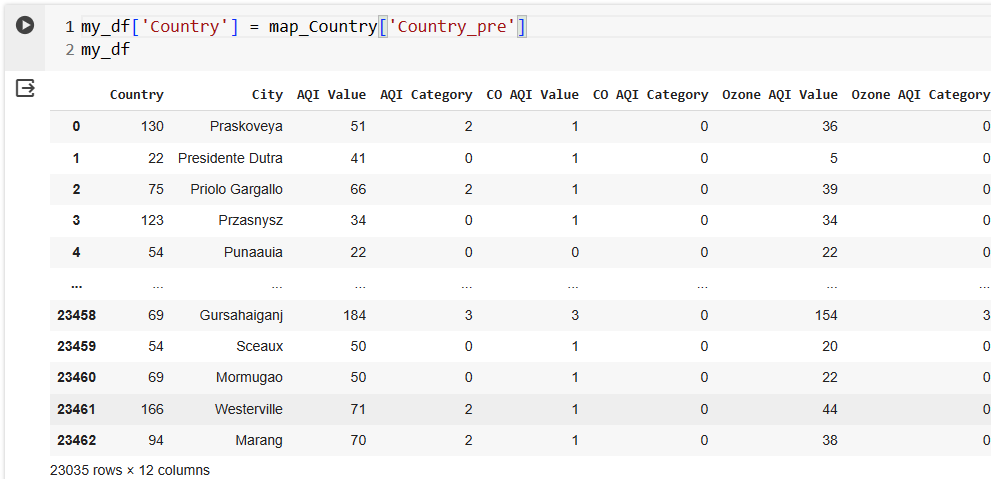
Ta sẽ xếp theo thứ tự tăng dần mức độ nguy hiểm tới sức khỏe vì mục đích của chúng ta là xác định dự đoán chất lượng không khí có ảnh hưởng tới sức khỏe hay không, các giá trị, trường hợp ảnh hưởng tới sức khỏe không nhiều nên ta cần tăng giá trị lên để mô hình có thể học được nhiều hơn, tổng quát hơn.

Các cột thuộc tính bị thay đổi giá trị là:

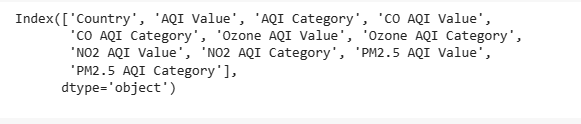
* AQI Category
* NO2 AQI Category
* CO AQI Category
* Ozone AQI Category
* PM2.5 AQI Category

**Hình 2.20. Sau khi mã hóa các cột AQI**

Tiến hành mã hóa tên các nước: sử dụng LabelEncoder của Sklearn

**Hình 2.21. Sau khi mã hóa cột Country**

Xét các cột thuộc tính: mục tiêu bài toán là triển khai trên quy mô phạm vi các nước nên tụi em đề xuất bỏ thuộc tính “City” đi. Các cột thuộc tính còn lại của bộ dữ liệu là:



**Hình 2.22. Các cột thuộc tính còn lại của bộ dữ liệu**

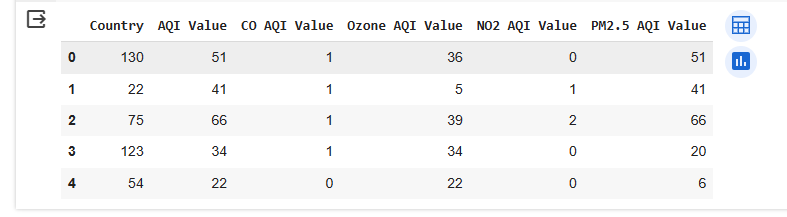
### 2.4.3. Chia tách dữ liệu:

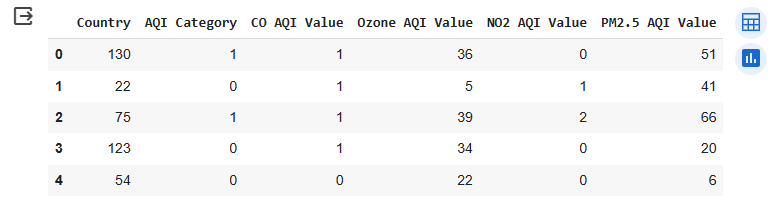
Trong bộ dữ liệu này, ta có các cột thuộc tính với 2 kiểu dữ liệu có dạng numeric và category.

* Numeric: AQI Value, CO AQI Value, Ozone AQI Value, NO2 AQI Value, PM2.5 AQI Value.
* Category: Country, City, AQI Category, CO AQI Category, Ozone AQI Category, NO2 AQI Category, PM2.5 AQI Category.

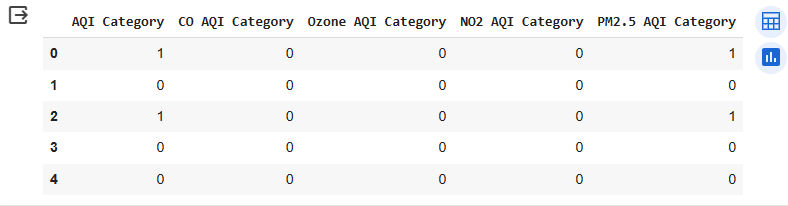
Nhằm phục vụ cho việc trích xuất thông tin, các luật cũng như đưa vào các mô hình, thuật toán máy học tụi em chia dữ liệu ra làm 3 bộ:

* Bộ dữ liệu 1 (df1): bao gồm các thuộc tính dạng số (Country, AQI Value, CO AQI Value, Ozone AQI Value, NO2 AQI Value, PM2.5 AQI Value) sử dụng cho các thuật toán hồi quy, dự đoán số AQI Value.

**Hình 2.23. Bộ dữ liệu 1 (df1)**

* Bộ dữ liệu 2 (df2): tương tự như bộ dữ liệu 1, tuy nhiên mục tiêu của chúng ta không phải là dự đoán AQI Value mà là AQI Category, các trường thuộc tính của bộ dữ liệu là Country, AQI Category, CO AQI Value, Ozone AQI Value, NO2 AQI Value, PM2.5 AQI Value

**Hình 2.24. Bộ dữ liệu 2 (df2)**

* Bộ dữ liệu 3 (df3): mục tiêu của bộ dữ liệu giống với bộ dữ liệu 2, đều dự đoán kết quả của AQI Category, ngoài ra ở bộ dữ liệu này ta sẽ dùng phương pháp luật kết hợp để phát hiện các luật. Bộ dữ liệu bao gồm các thuộc tính sau: AQI Category, CO AQI Category, Ozone AQI Category, NO2 AQI Category, PM2.5 AQI Category

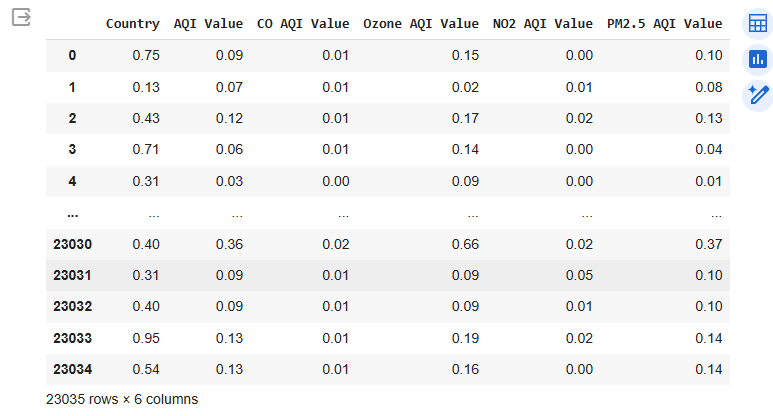
**Hình 2.25. Bộ dữ liệu 3 (df3)**

## 2.5. Khai phá dữ liệu:

### 2.5.1. Bộ dữ liệu df1:

#### 2.5.1.1. Chuẩn hóa bộ dữ liệu:

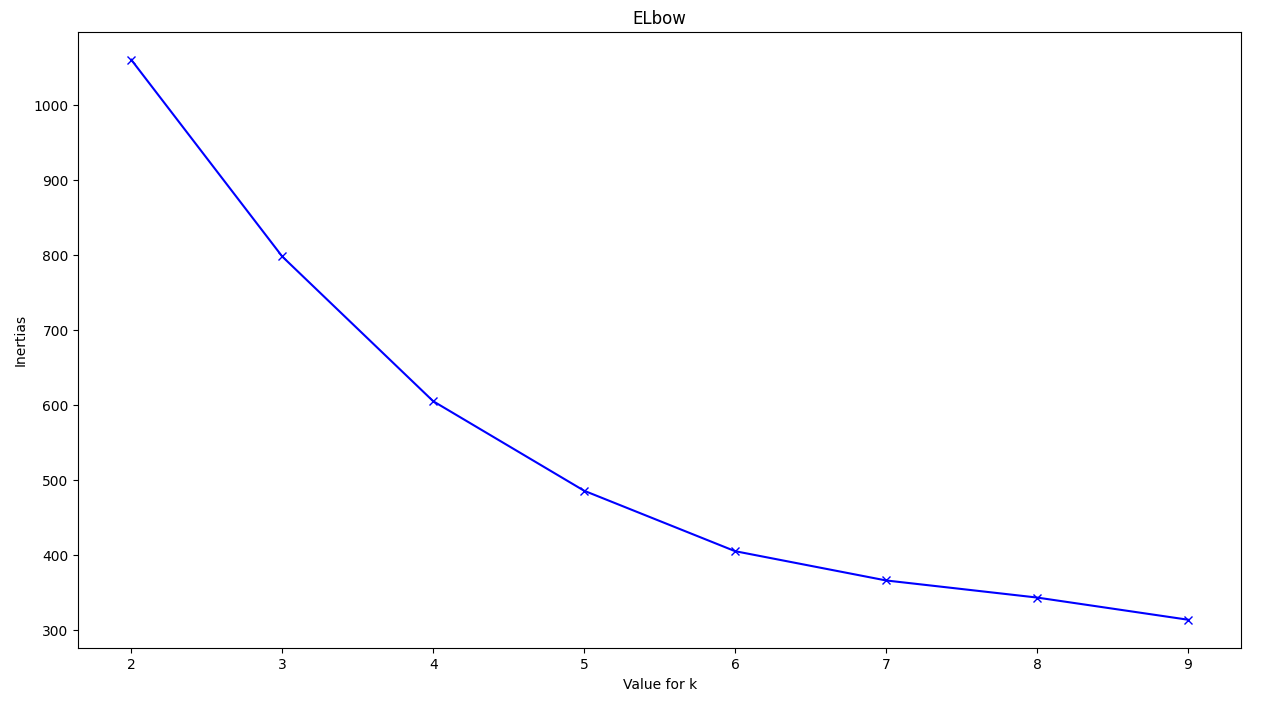
Dùng phương pháp MinMaxScaler để chuẩn hóa các thuộc tính theo tỉ lệ.



**Hình 2.26. Df1 sau khi chuẩn hóa dữ liệu**

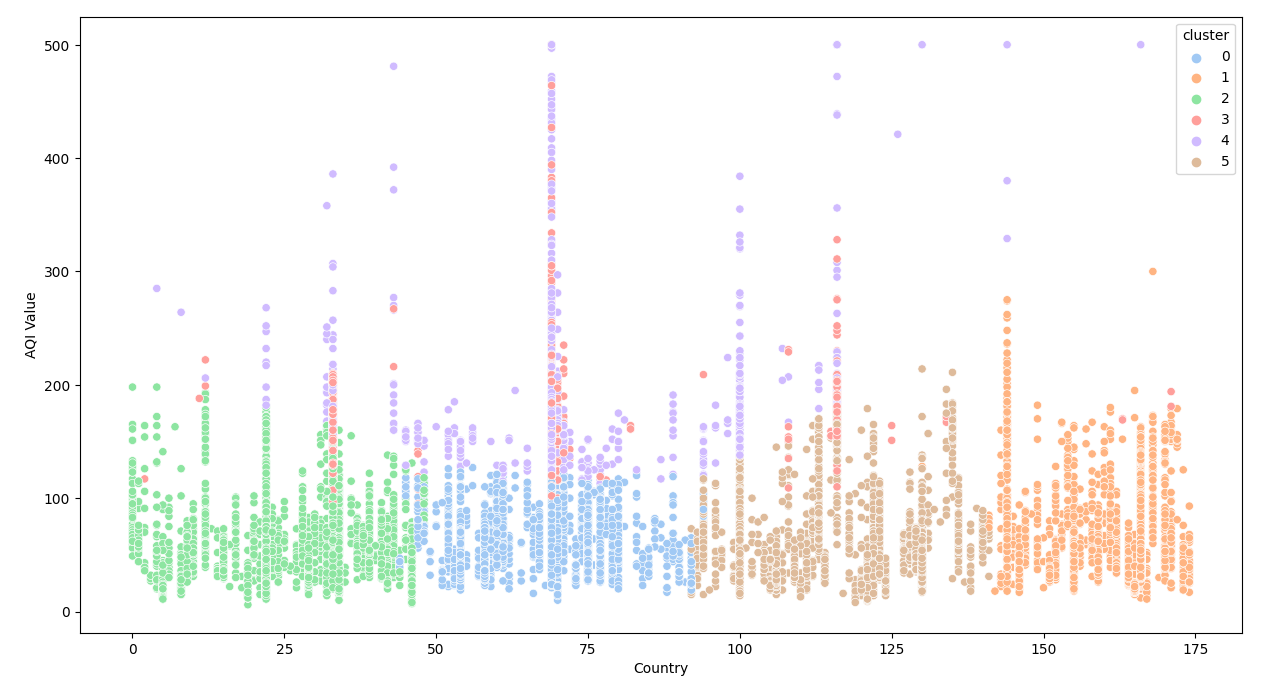
#### 

#### 2.5.1.2. Phân cụm với kỹ thuật K-means:

Sử dụng phương pháp Elbow để chọn ra số cụm thích hợp cho thuật toán. 

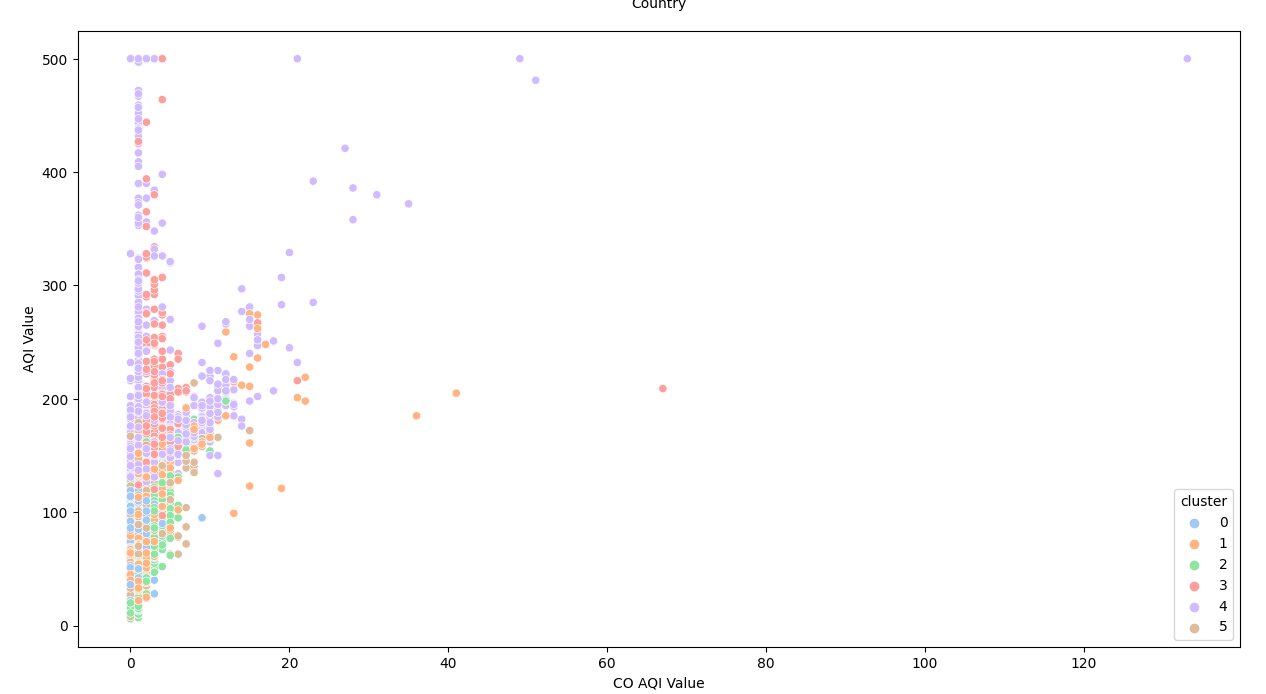
**Hình 2.27. Chọn số cụm bằng phương pháp Elbow**

\*Nhận xét: sau khi quan sát thì tụi em chọn số cụm áp dụng cho thuật toán K-mean là 6, vì ở số cụm này Inertia có dấu hiệu hội tụ.

**Hình 2.28. Áp dụng thuật toán K-means theo Country**

\*Nhận xét: Ta có thể thấy có sự phân cụm rõ rệt theo các quốc gia được đánh nhãn.

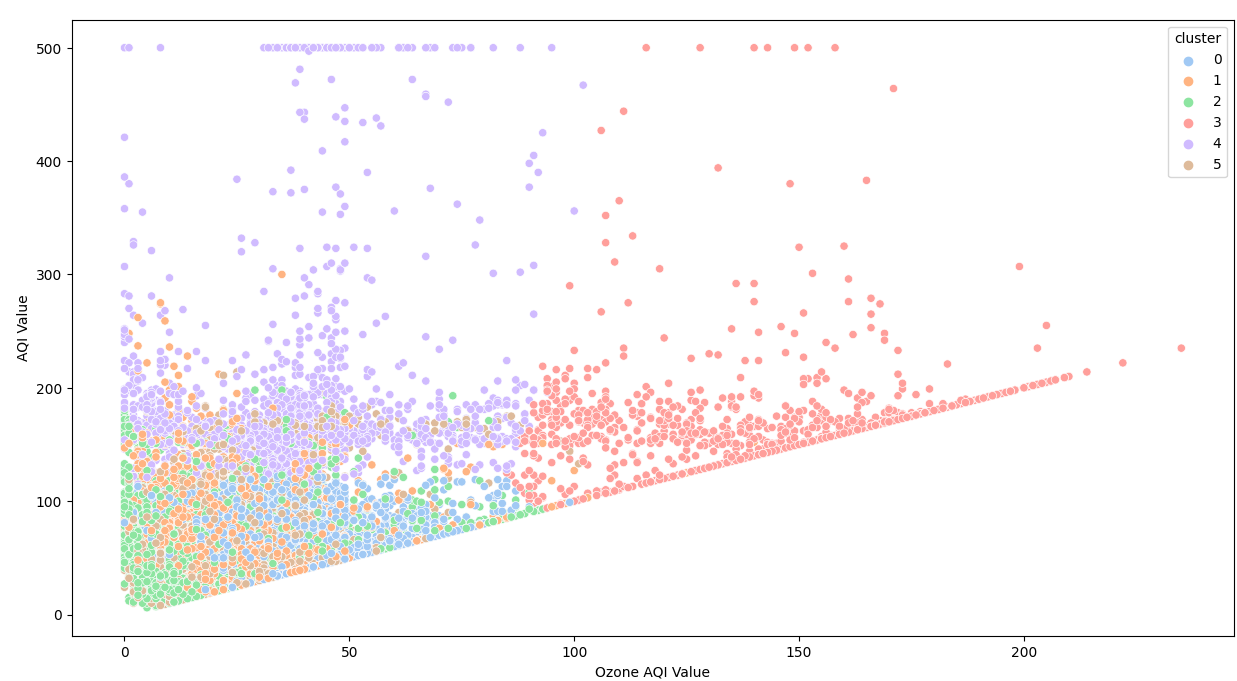
* Cụm 0: Các quốc gia có nhãn trong khoảng từ 40 đến 90 với AQI Value từ 120 trở xuống.
* Cụm 1: Các quốc gia có nhãn trong khoảng từ 140 đến 175 với AQI Value từ 300 trở xuống.
* Cụm 2: Các quốc gia có nhãn trong khoảng từ 0 đến 40 với AQI Value từ 200 trở xuống.
* Cụm 3: Các quốc gia rải rác có giá trị AQI Value ở mức 100 đến 450.
* Cụm 4: Các quốc gia rải rác có nhãn tập trung chủ yếu từ 25 đến 120 với mức AQI Value dao động từ 100 đến 500.
* Cụm 5: Các quốc gia có nhãn trong khoảng từ 90 đến 140 với mức AQI Value từ 200 trở xuống.



**Hình 2.29. Áp dụng thuật toán K-means theo CO AQI Value**

\*Nhận xét: Các cụm không có sự phân cụm rõ rệt với chỉ số CO AQI Value, rõ nhất chỉ có cụm 3 và 4.

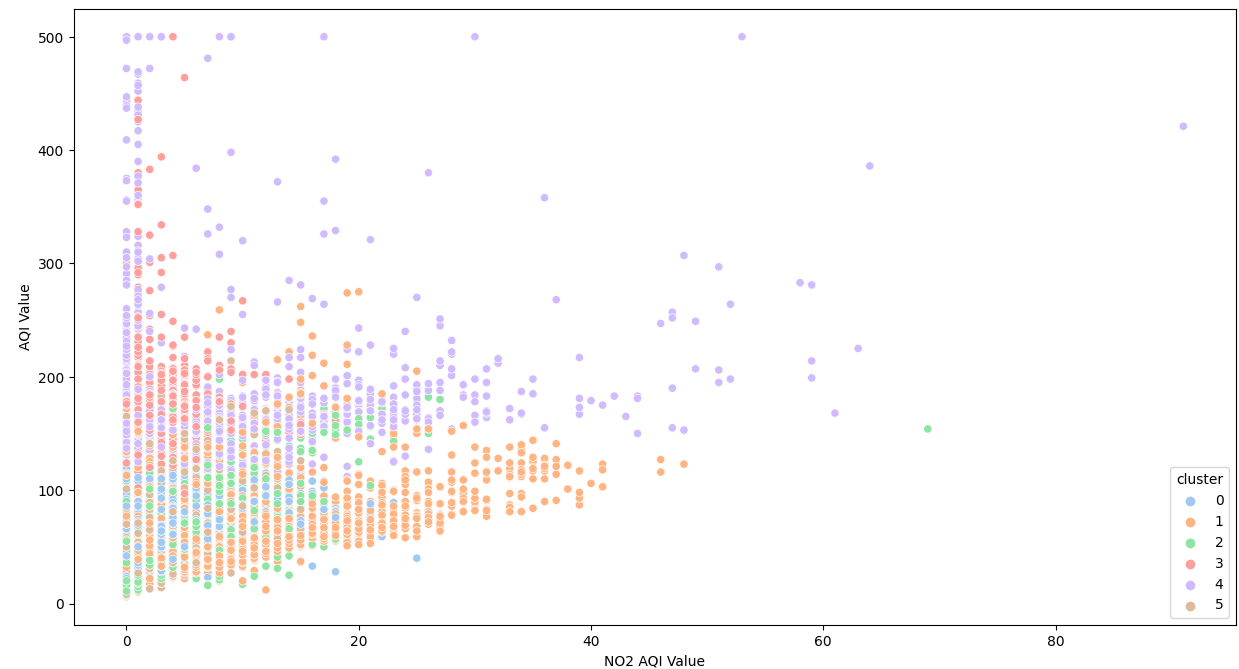
* Cụm 0, 1, 2, 5 có chỉ số CO AQI Value tập trung từ 0 đến 10 là chủ yếu.
* Cụm 3 có mức CO AQI Value tập trung trong khoảng từ 5 đến 10 với AQI Value tập trung nhiều trong khoảng 100 đến 250.
* Cụm 4 có mức CO AQI Value phân bố rộng trong khoảng 0 đến 25, cao nhất đạt gần 50.



**Hình 2.30. Áp dụng thuật toán K-means theo Ozone AQI Value**

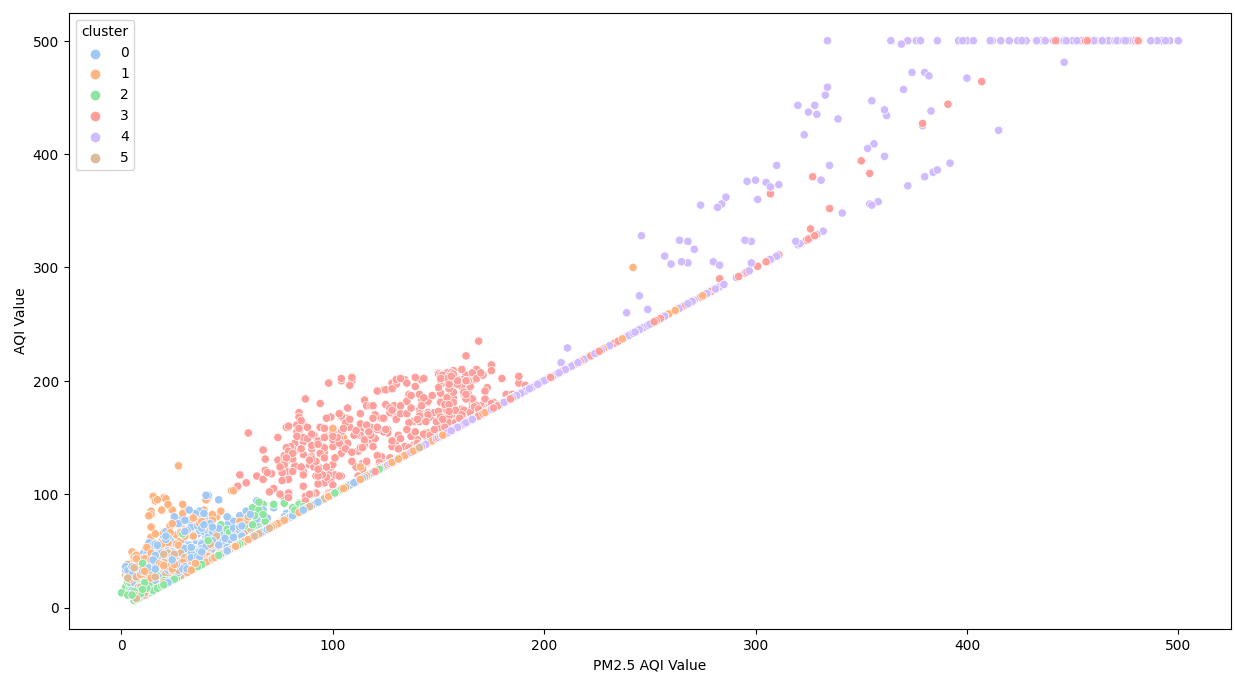
\*Nhận xét: Chỉ có vài cụm có sự phân hóa rõ rệt dựa theo chỉ số Ozone AQI Value:

* Cụm 0, 1, 2, 5: tập trung ở mức Ozone AQI Value trong khoảng từ 0 đến 90 và có mức AQI Value đa số từ 200 đổ xuống, chỉ có một số ít giá trị nhóm 1 có thể đạt giá trị AQI Value từ 300 đổ xuống.
* Cụm 4 có sự phân cụm rõ với mức Ozone AQI Value trong khoảng từ 0 đến 90 và giá trị AQI Value tập trung nhiều trong khoảng từ 150 đến 300, cao nhất có thể đạt 500.
* Cụm 3 cũng tương tự cụm 4, có sự phân cụm rõ rệt với mức Ozone AQI Value trong khoảng từ 90 trở lên, giá trị AQI Value tập trung nhiều trong khoảng từ 100 đến 250, cao nhất có thể đạt tới 500.

**Hình 2.31. Áp dụng thuật toán K-means theo NO2 AQI Value**

\*Nhận xét: Có sự phân cụm nhưng không rõ rệt, không có tính phân hóa mạnh.

* Cụm 0, 2, 5 có giá trị NO2 AQI Value trong khoảng từ 0 đến 20.
* Cụm 1 có giá trị NO2 AQI Value trong khoảng từ 0 đến 50.
* Cụm 3 có giá trị NO2 AQI Value trong khoảng từ 0 đến 15.
* Cụm 4 có có giá trị NO2 AQI Value tập trung trong khoảng từ 0 đến 40, cao nhất có thể đạt gần 65.



**Hình 2.32. Áp dụng thuật toán K-means theo PM2.5 AQI Value**

\*Nhận xét: Chỉ có cụm 3 và 4 có sự phân hóa rõ rệt theo chỉ số PM2.5 AQI Value.

* Cụm 3 có giá trị PM2.5 AQI Value tập trung nhiều trong khoảng từ 50 đến 200, cao nhất có thể gần 500.
* Cụm 4 có giá trị PM 2.5 AQI Value phân bổ rời rạc trong khoảng từ 150 đến 500.

**\*Tổng kết nhận xét sau khi áp dụng kỹ thuật phân cụm K-means với bộ dữ liệu df1:**

* Các nước thuộc cụm 0, 1, 2, 5 thường sẽ có mức độ ô nhiễm không khí tương đối thấp, ở mức tốt cho đến không tốt với nhóm người nhạy cảm. Trong khi đó nhóm nước thuộc cụm 3, 4 sẽ có mức độ ô nhiễm không khí đạt mức xấu đến báo động đối với sức khỏe con người. Qua đó ta có thể thấy chất lượng không khí có sự phân hóa rõ rệt theo vị trí địa lý các nước.
* Chỉ số Ozone AQI Value và PM2.5 AQI Value có tính liên hệ cao với chất lượng không khí (AQI Value) tạo nên sự phân hóa rõ rệt. Tuy là giá trị cao nhất của Ozone AQI Value vẫn chưa đạt đến mức gây nguy hiểm cho sức khỏe con người nhưng lại tạo ra sự phân cụm rõ trong phạm vi từ ảnh hưởng xấu đến tốt đối với sức khỏe. Chỉ số PM2.5 AQI có khả năng tác động nhiều nhất tới AQI Value, dựa theo biểu đồ phân bố có thể thấy rõ mối quan hệ tuyến tính của chúng.

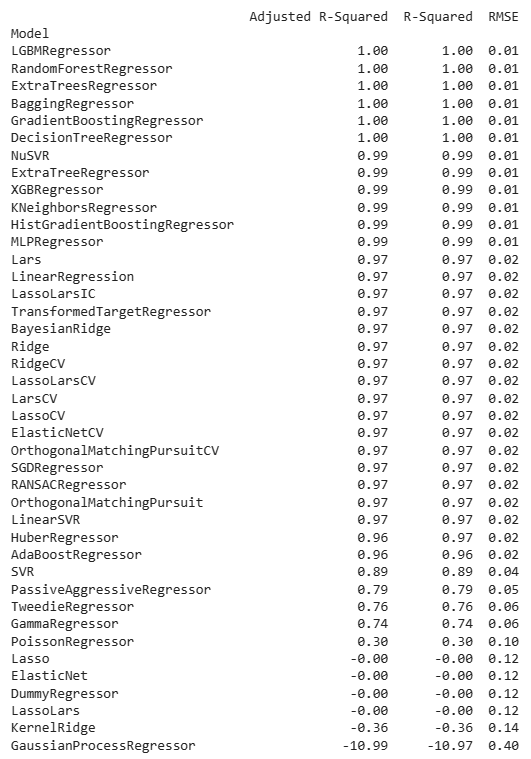
#### 

#### 2.5.1.3. Đánh giá tổng quan khả năng dự đoán của các thuật toán, mô hình hồi quy, chọn ra mô hình áp dụng cho bài toán.

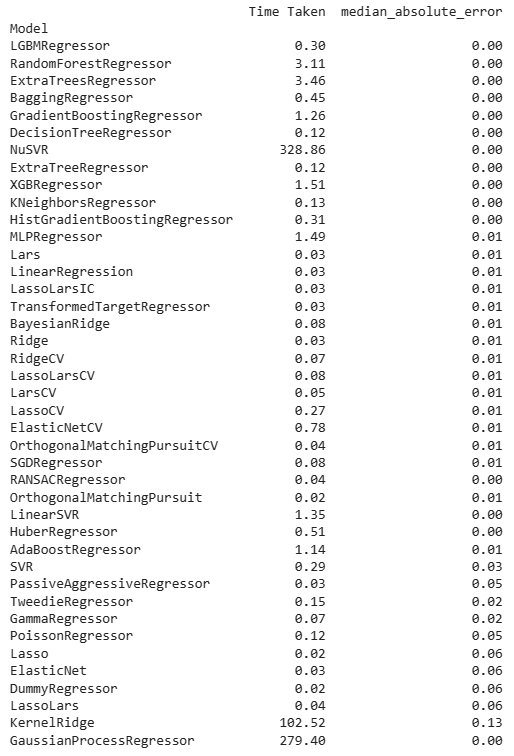
Ta sẽ sử dụng thư viện Lazy Predict[] để trực quan hóa một cách tổng quan, đánh giá khả năng dự đoán của các mô hình.

Các thông số dùng để đánh giá [4]:

* Adjusted R-Squared: R bình phương hiệu chỉnh
* R-Squared: R bình phương
* RMSE: Độ lệch chuẩn



**Hình 2.33.a. Sử dụng Lazy Predict để đánh giá các mô hình**



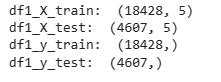
**Hình 2.33.b. Sử dụng Lazy Predict để đánh giá các mô hình**

Ta có kết quả tổng hợp của các mô hình xếp theo thứ tự độ hiệu quả giảm dần từ trên xuống dưới. Qua các chỉ số ta có thể thấy mô hình LinearRegression hoạt động khá tốt dựa trên các chỉ số đánh giá và cũng là một trong những thuật toán đơn giản giúp khai thác nhanh bộ dữ liệu. Chúng ta sẽ chọn mô hình này làm tiêu biểu để khai phá dữ liệu.

#### 2.5.1.4. Mô hình LinearRegression:

##### Train mô hình:

Chia tập dữ liệu train và test với tỉ lệ 8:2

**Hình 2.34. Chia tập dữ liệu train và test**

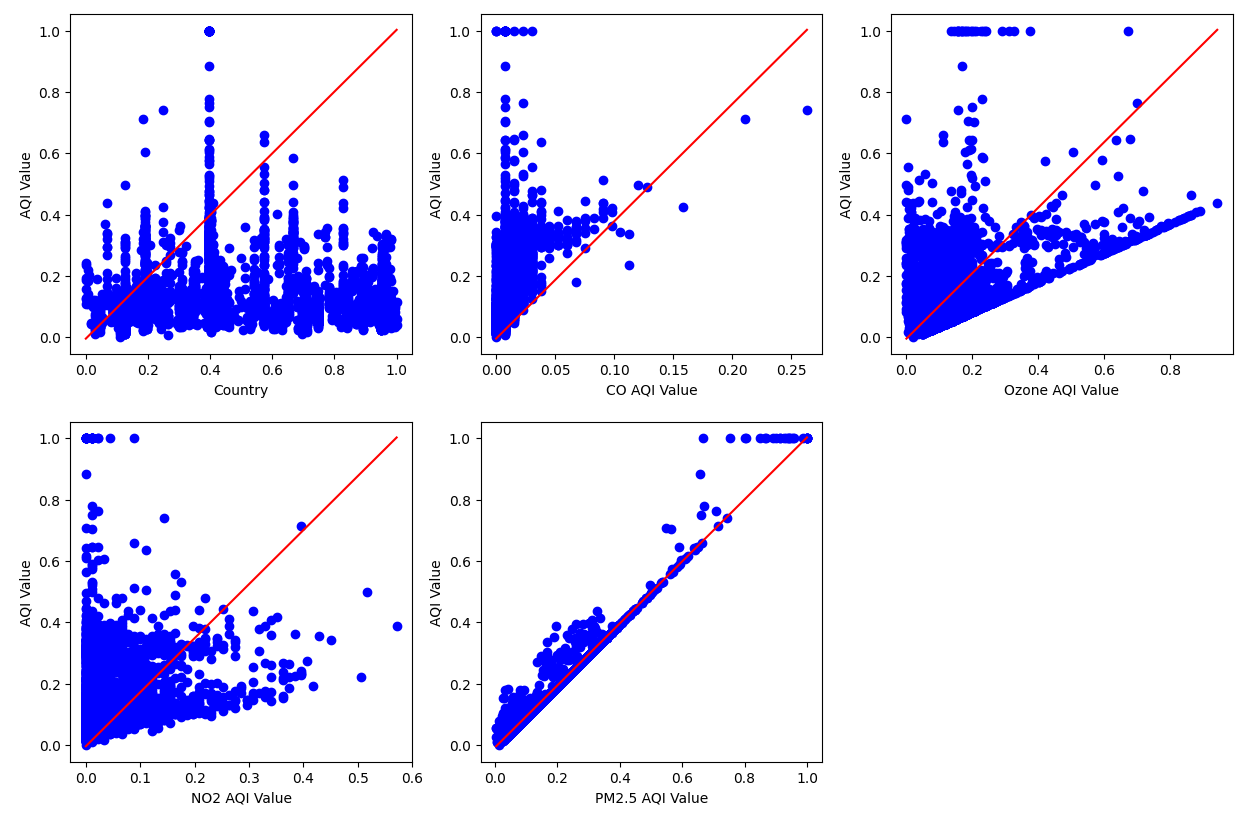
Kết quả thu được sau khi train model:



**Hình 2.35. Kết quả train model**

Đây là một con số rất tốt, gần như đạt tới 1 cho thấy mô hình có độ phù hợp cao với bộ dữ liệu.

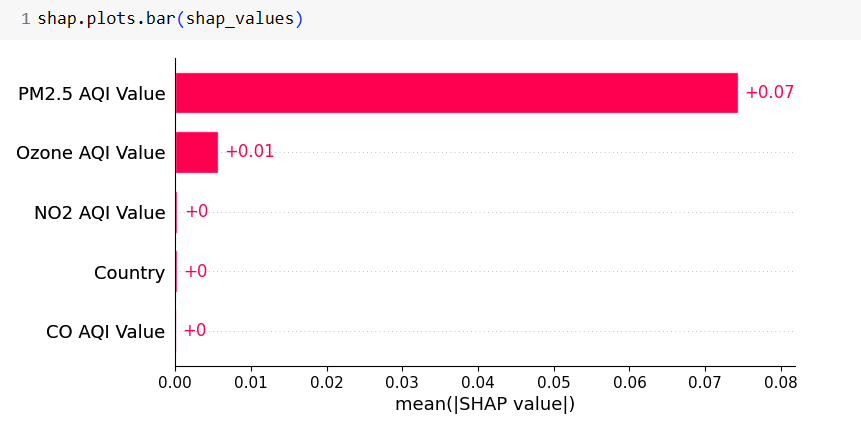
##### Trực quan hóa kết quả thu được:

**Hình 2.36. Trực quan hóa kết quả**

\*Nhận xét: đường thẳng dự báo màu đỏ cho ta thấy mức độ ảnh hưởng của các thuộc tính tới kết quả dự đoán. Ở đây có thuộc tính PM2.5 AQI Value có ảnh hưởng, khớp nhiều nhất với kết quả dự đoán AQI Value thu được. Ta có thể thấy PM2.5 AQI Value là thuộc tính có độ ảnh hưởng cao nhất tới kết quả dự đoán, 2 thuộc tính tiếp theo cũng tác động ảnh hưởng nhiều đến kết quả là Ozone AQI Value và NO2 AQI Value. Các thuộc tính còn lại có mối quan hệ khá thấp, gần như là không tới kết quả.

##### Đánh giá dựa trên Shap Value:

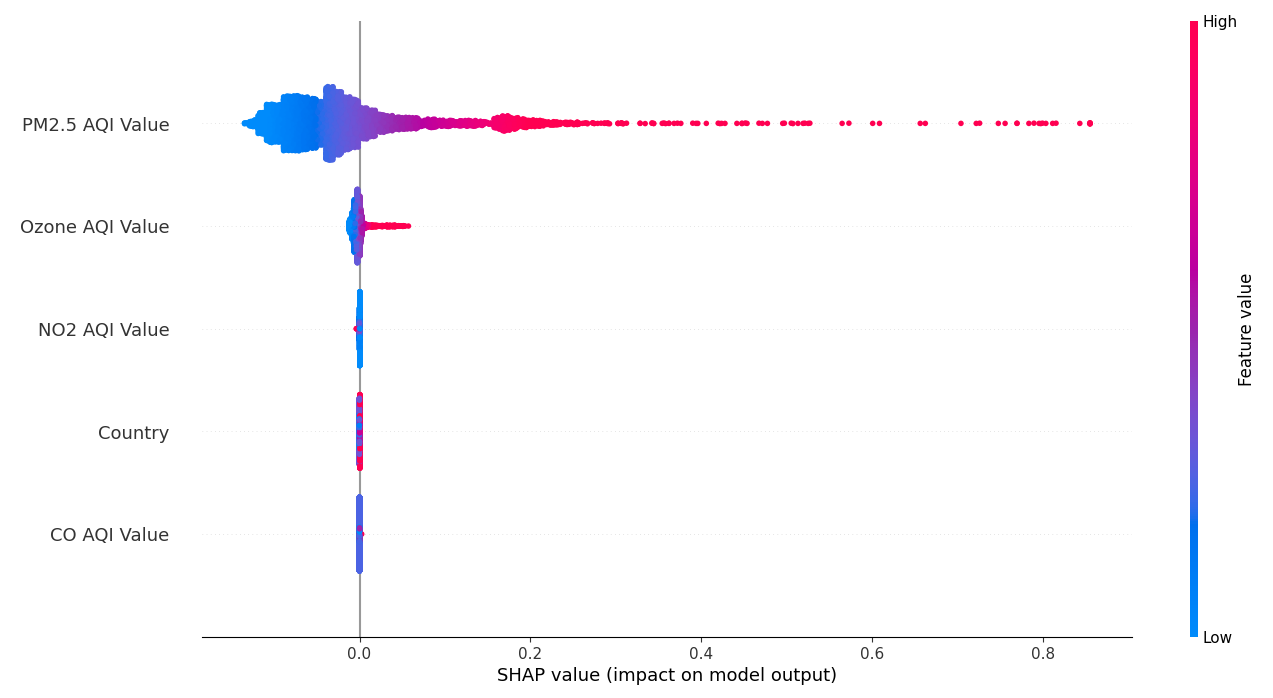
Đo mức độ ảnh hưởng của các thuộc tính tới kết quả:



**Hình 2.37. Mức độ ảnh hưởng của các thuộc tính đến kết quả**

\*Nhận xét: Kết quả dự đoán hầu như chịu ảnh hưởng hoàn toàn bởi thuộc tính PM2.5 AQI Value kể cả theo hướng tích cực hay tiêu cực,

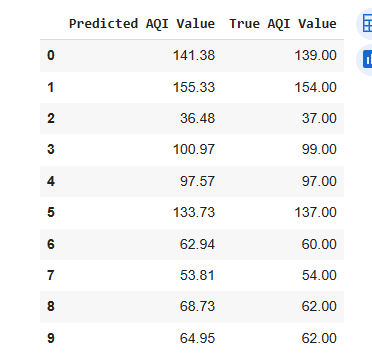
Tính ảnh hưởng trung bình của các thuộc tính:



**Hình 2.38. Tính ảnh hưởng trung bình của các thuộc tính đến kết quả**

\*Nhận xét: Qua đây ta có thể khẳng định chắc chắn PM2.5 AQI Value là thuộc tính có ảnh hưởng lớn nhất, chi phối gần như tuyệt đối đến các kết quả dự đoán. Đúng với tình trạng thực tế của không khí toàn cầu, trong khi các yếu tố khác khó phát sinh, ảnh hưởng thì lượng khí thải bụi mịn từ các phương tiện di chuyển, hoạt động công nghiệp thì lại không ngừng thải ra, không những thế các hoạt động, thiết bị còn tạo ra khí ozone gây hại cho sức khỏe con người.

##### Đánh giá dựa trên kết quả dự đoán với thực tế:



**Hình 2.39. Kết quả dự đoán với thực tế**

Độ chênh lệch giữa giá trị AQI Value dự đoán so với thực tế hầu hết tương đối thấp, không đáng kể.

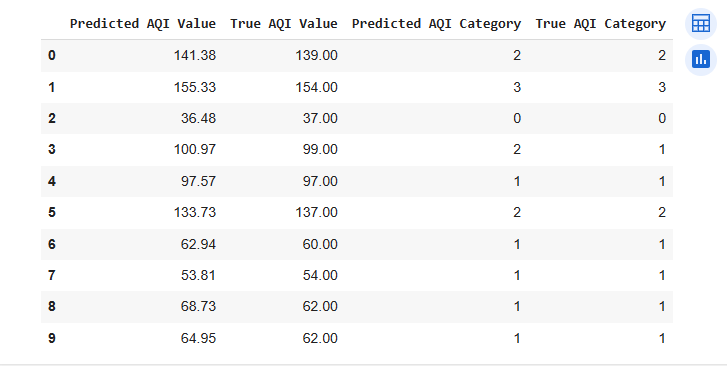


**Hình 2.40.a. Giá trị chênh lệch**



**Hình 2.40.b. Giá trị chênh lệch**

Chuẩn hóa chúng về dạng AQI Category để đo độ chính xác:



**Hình 2.41. Kết quả dự đoán với thực tế sau khi chuẩn hóa**

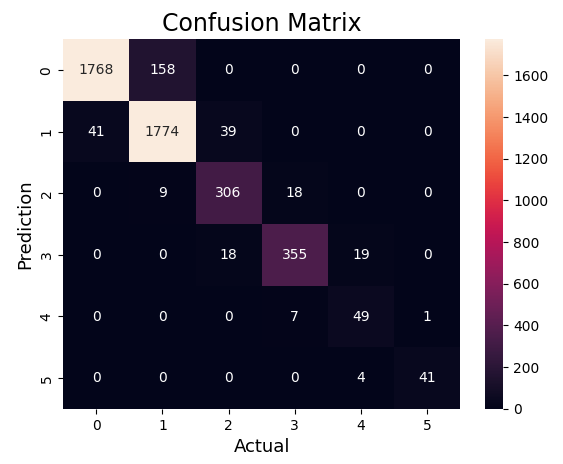
Độ chính xác thu được dựa theo các nhãn là khoảng 93.18%

Tuy khi chuyển sang giá trị dạng nhãn chịu sự ảnh hưởng bởi chênh lệch giá trị số ở các ranh giới các mức nhưng độ chính xác ta thu được vẫn tương đối cao cho thấy độ hiệu quả, phù hợp của thuật toán đối với bộ dữ liệu.

Vì đây là bộ dữ liệu mất cân bằng về số lượng giá trị các mức độ ảnh hưởng đến sức khỏe, mức độ từ xấu trở lên chiếm khá ít, do đó ta cần phải đo thêm tỉ lệ chính xác với các nhãn đại diện mức độ ảnh hưởng xấu trở lên đối với sức khỏe.

Độ chính xác thu được dựa trên các nhãn xấu trở lên là khoảng 90.08%

##### Đánh giá dựa trên Confusion Matrix:

****

**Hình 2.42. Ma trận nhầm lẫn**

Nhìn chung các nhãn chủ yếu bị nhầm lẫn với các nhãn liền kề, không có chênh lệch lớn. Sự nhầm lẫn ở đây đến từ những giá trị AQI Value gần khoảng ranh giới, mô hình Linear Regression có thể dự đoán giá trị số nhưng không thể đúng hoàn toàn sẽ có những khoảng chênh lệch nhỏ khiến AQI Value ở mức này lệch sang mức khác chênh dưới 1 bậc.

* Ở nhãn 0 và 1, số lượng nhầm lẫn giữa 2 nhãn này tương đối nhiều, tuy nhiên đều là mức độ không ảnh hưởng đến sức khỏe nên không gây ra sự nhầm lẫn nguy hiểm nào.
* Bắt đầu từ nhãn 2 (gây ảnh hưởng xấu đến nhóm người nhạy cảm) thì lại bị nhầm lẫn với nhãn 1 (tương đối ổn, không ảnh hưởng xấu), tuy là ở mức độ này thì tính nghiêm trọng không cao nhưng vẫn gây ra rủi ro nguy hiểm.
* Bắt đầu từ các nhãn 3 trở lên thì số lượng nhầm lẫn ít hơn (tập mẫu ít đi) tuy nhiên tỷ lệ nhầm lẫn lại tăng lên (trên 10%), phạm vi nhẫm lần là trên dưới 1 bậc. Tuy là ở phạm vi này đều gây ra ảnh hưởng xấu đến con người, đưa ra cảnh báo để phòng bị, kiểm soát và khắc phục nhưng chúng ta không thể ý thức rõ được mức độ nguy hiểm mà chúng mang lại để đưa ra động thái chuẩn bị phù hợp. Do đó ta cần phải luôn có biện pháp phòng bị trên 1 bậc.

##### Tổng kết:

Mô hình (thuật toán) Linear Regression hoạt động tốt trên bộ dữ liệu, cho ra được kết quả dự đoán với tỉ lệ chính xác tương đối cao kể cả với các nhãn đại diện mức độ ảnh hưởng xấu trở lên đối với sức khỏe, đồng thời cho ta thấy thực trạng ô nhiễm môi trường không khí, chỉ số AQI Value chịu ảnh hưởng nhiều bởi chỉ số bụi mịn (PM 2.5 AQI value) và chỉ số ozone trong không khí.

### 2.5.2. Bộ dữ liệu df2:

#### 2.5.2.1. Đánh giá tổng quan khả năng dự đoán của các thuật toán, mô hình phân loại, chọn ra mô hình áp dụng cho bài toán.

Ta sẽ sử dụng thư viện Lazy Predict[] để trực quan hóa một cách tổng quan, đánh giá khả năng dự đoán của các mô hình.

Các thông số dùng để đánh giá [4]:

* Accuracy: độ chính xác các nhãn
* Balanced Accuracy: độ chính xác cân bằng, thường được sử dụng đối với các tập dữ liệu có số lượng nhãn mất cân bằng.
* ROC AUC: đường cong phân loại.
* F1 Score:



**Hình 2.43. Sử dụng Lazy Predict để đánh giá các mô hình**

Qua thống kê cho thấy mô hình (thuật toán) DecisionTreeClassifier cho ra hiệu quả tốt nhất. Độ chính xác cân bằng đạt gần như tuyệt đối, tuy nhiên ta cũng cần phải lưu ý vì việc đánh giá mô hình này chỉ dựa trên tỷ lệ dự đoán chính xác là chính, đối với mô hình (thuật toán) DecisionTreeClassifier khi độ sâu càng tăng thì độ phức tạp của mô hình sẽ càng cao ngoài ra còn dễ dẫn đến tình trang Overfitting. Chúng ta cần tiến hành phân tích thuật toán DecisionTreeClassifier với bộ dữ liệu này.

#### 2.5.2.2. Mô hình DecisionTreeClassifier:

##### Train mô hình:

Các tham số thử nghiệm cho mô hình:



**Hình 2.44. Tham số thử nghiệm cho mô hình**

Sử dụng GridSearch() để chọn ra tham số phù hợp nhất.

Kết quả thu được là:



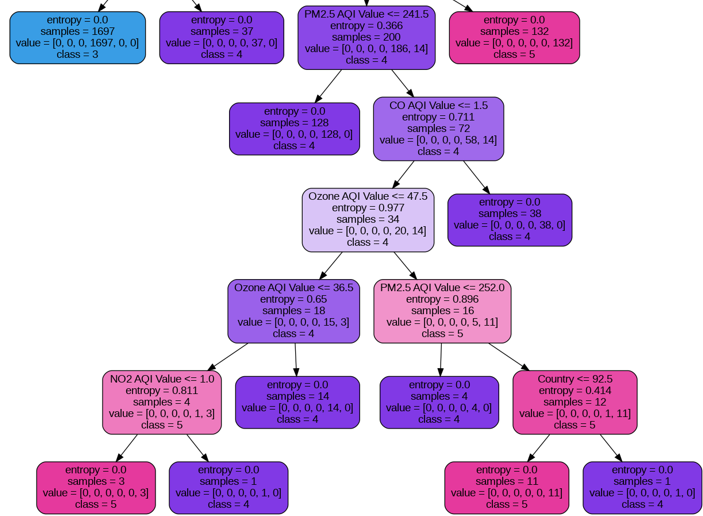
**Hình 2.45. Kết quả**

Thiết lập mô hình(thuật toán) DecisionTreeClassifier với tham số trên.

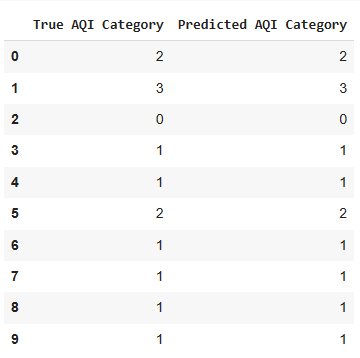
##### Trực quan hóa mô hình (thuật toán) DecisionTreeClassifier Classifier thu được:

##### Hình 2.46.a. Trực quan hóa mô hình

##### 

**Hình 2.46.b. Trực quan hóa mô hình****Hình 2.46.c. Trực quan hóa mô hình**

##### Đánh giá dựa trên kết quả dự báo so với thực tế trên bộ dữ liệu test :



**Hình 2.47. Kết quả dự báo so với thực tế**

Độ chính xác và độ chính xác cân bằng thu được là:

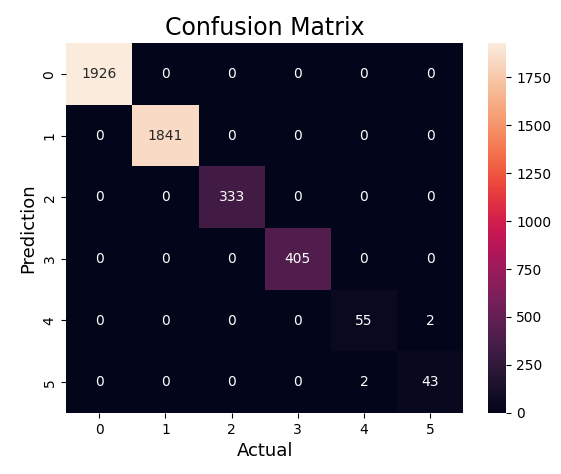


**Hình 2.48. Độ chính xác**

\*Nhận xét: Độ chính xác và độ chính xác cân bằng thu được là rất cao cho thấy mô hình rất hiệu quả, phù hợp với bộ tham số và bộ dữ liệu df2.

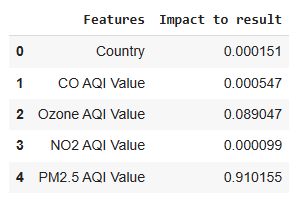
##### 

##### Đánh giá dựa trên Confusion Matrix:

**Hình 2.49. Ma trận nhầm lẫn**

Bảng Confusion Matrix cho thấy kết quả hết sức ngạc, chỉ có 2 trường hợp nhầm lẫn ở nhãn 4 và 5. Nguyên do có thể do dữ liệu học tập ở 2 loại nhãn này tương đối hạn chế, nhưng đây vẫn là một kết quả tuyệt vời.

##### Đánh giá dựa trên mức độ ảnh hưởng của các thuộc tính với mô hình:



**Hình 2.50. Độ ảnh hưởng của các thuộc tính**

Thông qua thuộc tính feature\_importances\_ của DecisionTreeClassification ta có thể thấy thuộc tính ảnh hưởng nhiều nhất đến kết quả chất lượng không khí (AQI Value) vẫn là chỉ số PM2.5 AQI Value và tiếp đến là Ozone AQI Value.

##### Tổng kết:

Mô hình DecisionTreeClassification có độ tương thích phù hợp cao với bộ dữ liệu df2, có thể đưa ra dự đoán với tỷ lệ chính xác gần như tuyệt đối. Thuộc tính PM2.5 AQI Value có độ ảnh hưởng cao nhất tới kết quả, phản ánh được tình trạng thực tế vấn đề toàn cầu khi lượng khí thải bụi mịn từ các phương tiện, nhà máy, cơ sở công nghiệp thải ra là vô cùng nghiêm trọng. Tiếp đến ảnh hưởng đến lượng Ozone trong không khí dưới mặt đất. Đây là 2 nguyên nhân chính tác động đến chất lượng không khí, ảnh hưởng của vấn đề ô nhiễm không khí tới sức khỏe con người.

### 

### 2.5.3. Bộ dữ liệu df3

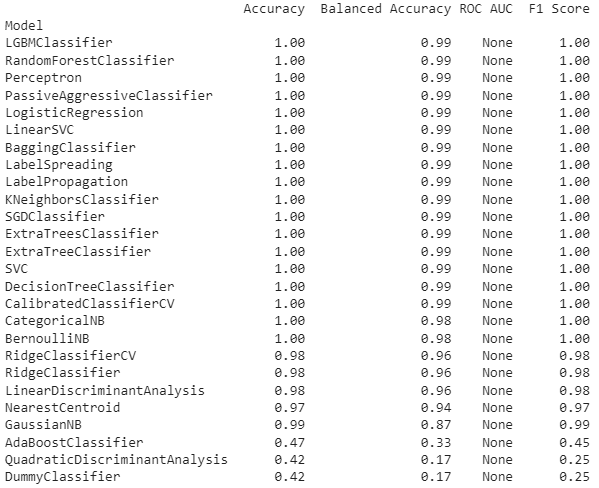
#### 2.5.3.1. Đánh giá tổng quan khả năng dự đoán của các thuật toán, mô hình phân loại, chọn ra mô hình áp dụng cho bài toán.

Cũng như bộ dữ liệu df1 và df2, ta sẽ sử dụng thư viện Lazy Predict[] để trực quan hóa một cách tổng quan, đánh giá khả năng dự đoán của các mô hình trên bộ dữ liệu df3.

Các thông số dùng để đánh giá [4]:

* Accuracy: độ chính xác các nhãn
* Balanced Accuracy: độ chính xác cân bằng, thường được sử dụng đối với các tập dữ liệu có số lượng nhãn mất cân bằng.
* ROC AUC: đường cong phân loại.
* F1 Score: Là sự kết hợp giữa độ chính xác (precision) và độ phủ (recall) của mô hình.

Kết quả đánh giá các mô hình:



**Hình 2.51. Sử dụng Lazy Predict để đánh giá các mô hình**

Qua thống kê cho thấy mô hình (thuật toán) KNeighborsClassifier cho ra hiệu quả khá tốt với độ chính xác và độ chính xác cân bằng đạt gần như tuyệt đối. KNN thích hợp trong trường hợp này vì dễ hiểu và dễ triển khai, thêm nữa, dữ liệu df3 phân phối không đồng nhất, không theo phân phối nhất định và số lượng mẫu trong df3 khá nhỏ, KNN có thể mang lại hiệu suất rất tốt.

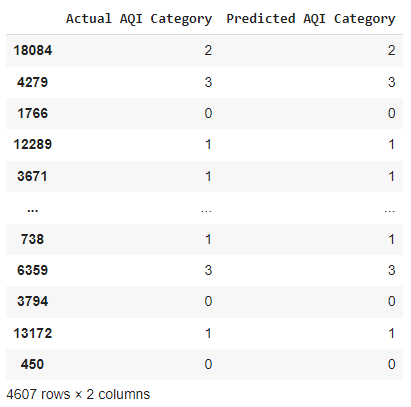
#### 2.5.3.2. Mô hình KNeighborsClassifier:

##### Train mô hình:

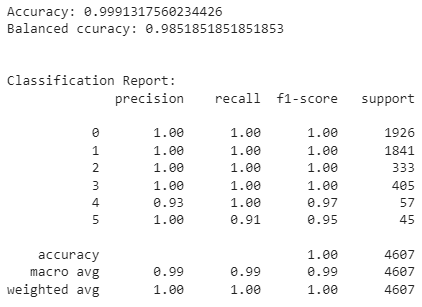
Để huấn luyện mô hình ta cần chọn ra giá trị ‘k’ (n\_neighbors) phù hợp nhất, do đó ta cũng sử dụng kỹ thuật Grid Search thu được kết quả best\_n\_neighbors là 5

Từ giá trị “n\_neighbors” tốt nhất trên, ta tiến hành huấn luyện mô hình.

##### Đánh giá dựa trên kết quả dự báo so với thực tế trên bộ dữ liệu test:



**Hình 2.52. Kết quả dự báo so với thực tế trên bộ dữ liệu**



**Hình 2.53. Báo cáo đánh giá mô hình**

\*Nhận xét:

* Bảng Actual AQI Category và Predicted AQI Category cho thấy kết quả dự đoán đúng gần như 100% với độ chính xác và độ chính xác cân bằng rất cao, xấp xỉ mô hình Decision Tree ở bộ dữ liệu df2. Cho thấy độ hiệu quả của mô hình.
* Nhìn vào báo cáo phân loại (Classification Report) ta thấy Class 0 và 1 (tương ứng Good và Moderate) có số lượng thực sự (Support) trong tập dữ liệu kiểm tra lớn nhất (1926 và 1841 trên tổng số 4607) nhưng các chỉ số khác đều đạt tối đa (Precision và Recall).
* Ngoài ra còn có các chỉ số khác như:
  + **Accuracy (Độ chính xác):** Đây là tỷ lệ giữa số lượng dự đoán đúng trên tổng số dự đoán. Trong trường hợp này, độ chính xác là 1.00, tức là 100%. Mô hình KNN dự đoán đúng toàn bộ tập dữ liệu kiểm tra.
  + **Macro Avg (Trung bình đa lớp):** Trung bình của các metric (Precision, Recall, F1-Score) tính toán cho từng lớp và sau đó lấy trung bình của chúng. Trong trường hợp này, trung bình đa lớp là 0.99, có nghĩa là mô hình có hiệu suất cao trên các lớp.
  + **Weighted Avg (Trung bình có trọng số):** Trung bình có trọng số của các metric tính toán cho từng lớp và sau đó lấy trung bình của chúng, với trọng số dựa trên số lượng mẫu trong mỗi lớp. Trong trường hợp này, trung bình có trọng số cũng là 1.00, cho thấy hiệu suất tốt trên toàn bộ tập dữ liệu.

##### Tổng kết:

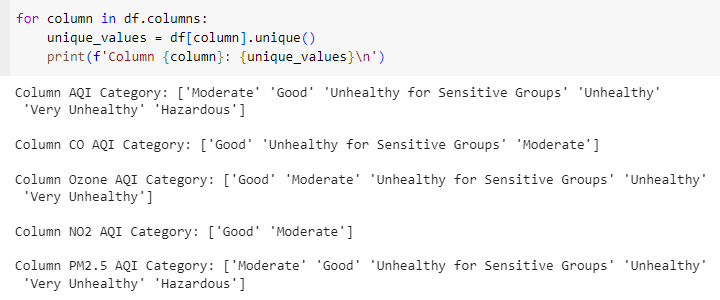
Mô hình KNeighborsClassifier dường như rất phù hợp với bộ dữ liệu test, đạt hiệu suất rất cao trên mỗi lớp. Về độ chính xác mô hình đạt 1.00, tức là dự đoán đúng toàn bộ tập dữ liệu kiểm tra. Điều này có vẻ ấn tượng, nhưng cũng có thể là dấu hiệu của overfitting. Support cho mỗi lớp thường không đồng đều, điều này là bình thường trong thực tế. Các lớp có số lượng mẫu lớn hơn thường có ảnh hưởng lớn hơn đến trung bình có trọng số.

#### 2.5.3.3. Sử dụng thuật toán Apriori để khám phá các luật kết hợp trong bộ dữ liệu

##### Chuẩn hóa bộ dữ liệu:

Phương pháp mã hóa dữ liệu one-hot encoding sẽ được chọn để chuẩn hóa bộ dữ liệu vì những lý do sau:

* Trong thuật toán Apriori, thông thường ta phải sử dụng ma trận tần suất (frequency matrix) để xác định các mục xuất hiện cùng nhau.. Apriori thường yêu cầu dữ liệu đầu vào là một ma trận chỉ chứa giá trị 0 hoặc 1 để xác định sự xuất hiện hoặc không xuất hiện của mỗi mục có trong dữ liệu.
* Khi in những giá trị unique của từng cột, ta thu được:

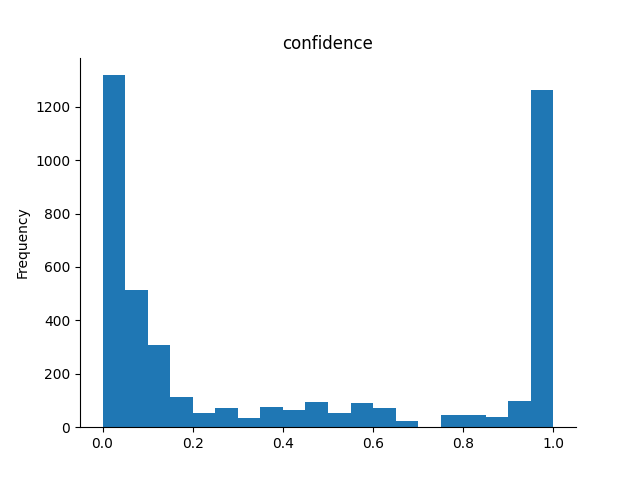


**Hình 2.54. Giá trị Unique của từng cột**

Việc sử dụng one-hot encoding đòi hỏi về mặt kiểm soát dữ liệu khác nhau không được quá nhiều, nếu không sẽ gây ra tình trạng tăng kích thước dữ liệu khi số lượng giá trị khác nhau của biến phân loại rất lớn. Điều này có thể làm tăng độ phức tạp của mô hình và yêu cầu nhiều tài nguyên và thời gian hơn để huấn luyện một mô hình. Trong bộ dữ liệu df3, số lượng giá trị khác nhau tương đối ít, phù hợp để sử dụng phương pháp one-hot encoding.

##### Áp dụng thuật toán Apriori để tìm các luật kết hợp:

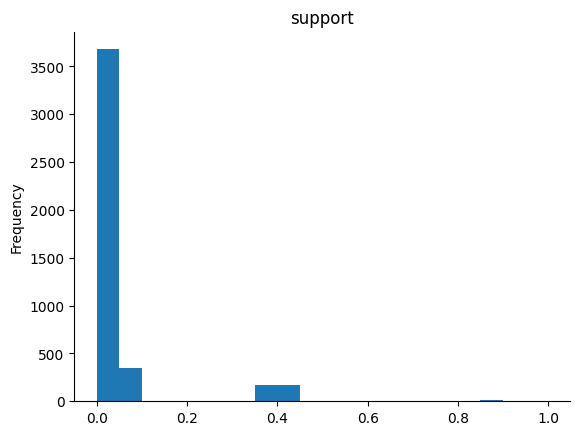
* **Tổng quan:**
  + Áp dụng thuật toán Apriori, tìm được tổng cộng 4374 luật kết hợp.
  + Phân bố các luật dựa theo confidence:



**Hình 2.55. Sự phân bố của các luật theo Confidence**

Các luật phân bố chủ yếu ở khoảng confidence 0.0 đến 0 .1 và 0.95 đến 1.0. Điều này có nghĩa là khoảng hơn 1250 luật không đáng tin, khi gặp trường hợp này hệ thống dự đoán khả năng rất cao là sai. Khoảng 1230 luật có confidence xấp xỉ tuyệt đối, khi gặp trường hợp này khả năng hệ thống dự đoán đúng là rất cao.

* + Phân bố các luật dựa theo support:

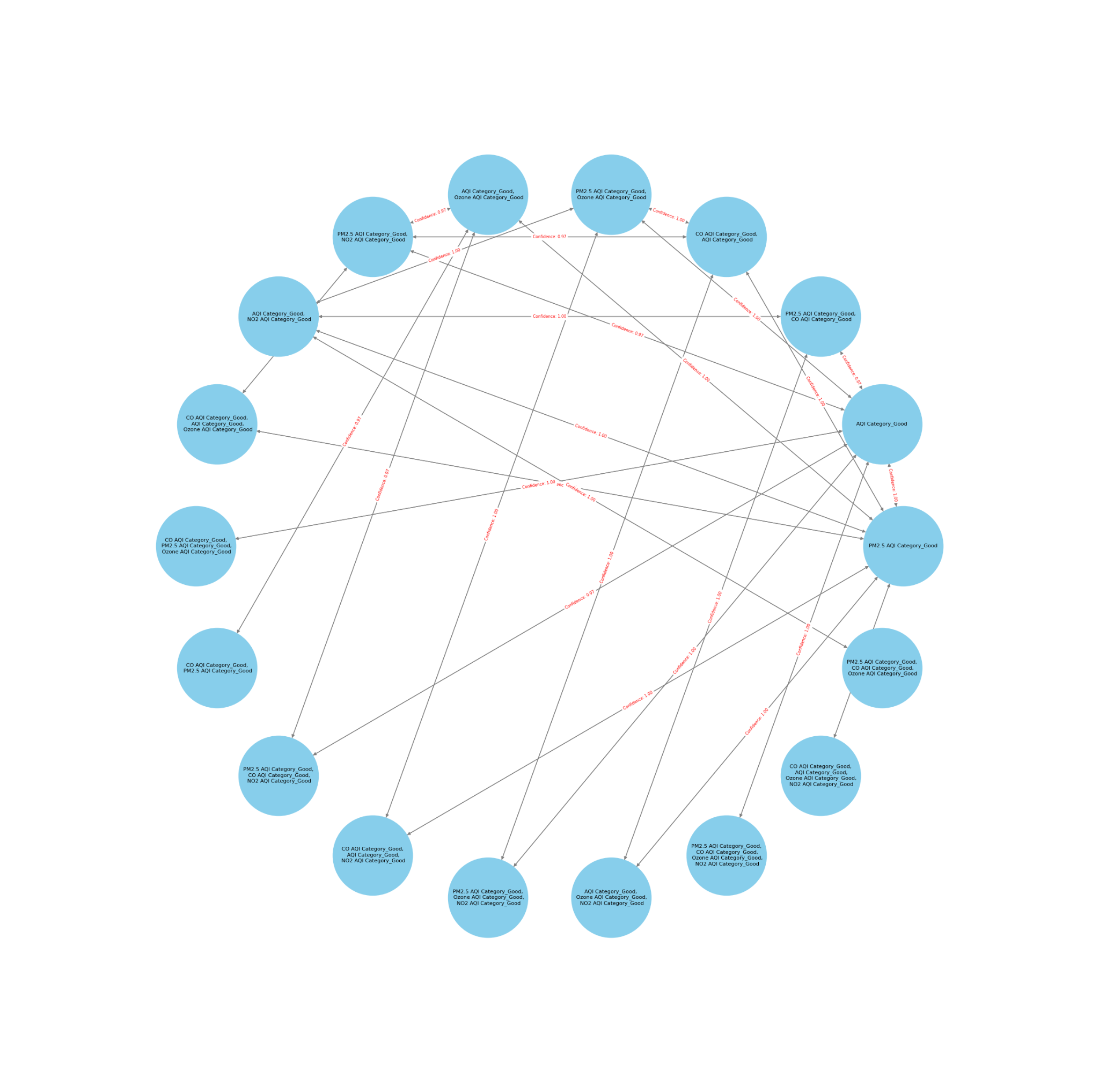


**Hình 2.56. Sự phân bố của các luật theo Support**

Các luật phân bố chủ yếu ở khoảng support 0.0 đến 0.1, cho thấy các trường hợp xuất hiện ít thường xuyên trong bộ dữ liệu.

* + Loại bỏ những luật có confidence thấp (chỉ lấy những luật có confidence > 0.8) ta thấy support vẫn rất thấp. Tình huống confidence cao nhưng support thấp có thể là dấu hiệu của hiện tượng quá khớp. Các luật bị quá khớp thường chỉ áp dụng cho một số ít dữ liệu trong tập dữ liệu đào tạo, và có thể không áp dụng được cho các dữ liệu mới.
  + Loại bỏ những có support thấp (chỉ lấy những luật có support > 0.4) và chỉ lấy những luật có sự liên kết mạnh mẽ và có ý nghĩa (trong trường hợp này luật phân bố chủ yếu ở lift < 1.2 và lift > 2.3, ta sẽ loại bỏ những luật có lift < 1.2) thu được kết quả 54 luật.

##### Trực quan hóa các luật kết hợp:



**Hình 2.57. Trực quan hóa các luật**

##### Tổng kết

Thuật toán Apriori phù hợp với bộ dữ liệu. Cụ thể:

* **Thời gian chạy nhanh:** Đây là một ưu điểm quan trọng của thuật toán Apriori. Nếu thời gian chạy của thuật toán quá lâu, thì thuật toán sẽ không thể áp dụng cho các bộ dữ liệu lớn.
* **Số lượng luật kết hợp được tạo ra không nhiều:** Điều này cho thấy rằng các mối quan hệ trong dữ liệu không quá phức tạp. Nếu số lượng luật kết hợp được tạo ra quá lớn, thì có thể các luật kết hợp không có ý nghĩa hoặc không đáng quan tâm.
* **Các luật kết hợp có ý nghĩa và độ tin cậy cao:** Đây là những yếu tố quan trọng nhất để đánh giá một thuật toán tìm luật kết hợp. Nếu các luật kết hợp có ý nghĩa và độ tin cậy cao, thì thuật toán đó có thể được sử dụng để giải thích hoặc dự đoán các xu hướng trong dữ liệu.

Các luật kết hợp cho thấy, muốn đạt được “AQI Category\_Good”, tối thiểu phải đạt được “PM2.5 AQI category\_Good”, bởi vì 100% luật ảnh hưởng đến “AQI Category\_Good” đều có sự xuất hiện của “PM2.5 AQI Category\_Good”. Một lần nữa cho thấy mức độ quan trọng của chỉ số bụi mịn trong không khí đã ảnh hưởng lớn như thế nào đến chỉ số chất lượng không khí. Kết quả thu được đã củng cố kết luận ở df1 và df2.

# KẾT LUẬN

Thông qua việc khai phá bộ dữ liệu về vấn đề ô nhiễm không khí toàn cầu với các chỉ số AQI, bao gồm PM2.5, Ozone, NO2, SO2 và CO. Bằng cách sử dụng các phương pháp thống kê và học máy, bài báo cáo đã tìm ra được yếu tố tác động nhất đến vấn đề ô nhiễm là AQI PM2.5, là 1 loại bụi mịn có kích thước nhỏ hơn 2.5 micromet, có thể xâm nhập sâu vào phổi và máu, gây ra các bệnh về hô hấp, tim mạch và ung thư. Yếu tố tiếp theo là Ozone, là khí gây ô nhiễm có màu xanh nhạt, có thể gây kích ứng mắt, họng, khó thở và suy giảm chức năng phổi. Từ bộ dữ liệu, bài nghiên cứu cũng đã xây dựng được một mô hình dự đoán mức độ ảnh hưởng đến sức khỏe con người dựa trên các mức AQI, từ đó đưa ra các khuyến nghị phòng ngừa và cải thiện chất lượng không khí.

Bài báo cáo đã đóng góp cho việc nâng cao nhận thức và hiểu biết về ô nhiễm không khí và ảnh hưởng của nó đến sức khỏe con người. Tuy nhiên, bài báo cáo vẫn còn có một số hạn chế, như dữ liệu chỉ được thu thập từ một số địa điểm có trạm đo AQI, chưa phản ánh được chính xác tình trạng ô nhiễm ở các khu vực khác. Do đó, bài báo cáo có thể đề xuất các hướng nghiên cứu tiếp theo, như mở rộng dữ liệu từ nhiều nguồn khác nhau, áp dụng các phương pháp khai phá dữ liệu mới và nâng cao độ chính xác của mô hình dự đoán. Ngoài ra, thông qua việc nghiên cứu, báo cáo này cũng kêu gọi sự hợp tác của các cơ quan chức năng, các tổ chức xã hội và cộng đồng dân cư trong việc giảm thiểu nguồn thải gây ô nhiễm và bảo vệ môi trường sống.

# TÀI LIỆU THAM KHẢO

[1] Kieu Hoa, Pum, “DỮ LIỆU LÀ GÌ? CÁC LOẠI DỮ LIỆU VÀ CÁCH PHÂN TÍCH DỮ LIỆU,” 07/04/2023. [Trực tuyến]. Địa chỉ: <https://200lab.io/blog/du-lieu-la-gi/#:~:text=D%E1%BB%AF%20li%E1%BB%87u%20l%C3%A0%20g%C3%AC%3F%20D%E1%BB%AF%20li%E1%BB%87u%20l%C3%A0%20t%E1%BA%ADp,t%E1%BB%AD%2C%20s%E1%BB%91%20li%E1%BB%87u%2C%20h%C3%ACnh%20%E1%BA%A3nh%2C%20%C3%A2m%20thanh%2C%20video%2C>. [Truy cập 12/2023].

[2] N.Đ. Quân, N.V.T. Nam và P.M. Đức, “DỮ LIỆU LÀ GÌ? CÁC LOẠI DỮ LIỆU VÀ CÁCH PHÂN TÍCH DỮ LIỆU,” 03/2021. [Trực tuyến]. Địa chỉ: <https://www.studocu.com/vn/document/dai-hoc-dien-luc/he-thong-thong-tin/nhom-4-khai-pha-du-lieu-tai-lieu-hoc-tap/31544152>. [Truy cập 12/2023].

[3] T.V. Hoàn, N.Đ.T. Minh và Đ.V. Hùng, “THUẬT TOÁN HỒI QUY TUYẾN TÍNH ỨNG DỤNG VÀO DỰ ĐOÁN BỆNH NHÂN BỊ BỆNH UNG THƯ PHỔI,” 04/2023. [Trực tuyến]. Địa chỉ: <https://www.studocu.com/vn/document/dai-hoc-dien-luc/khai-pha-du-lieu/baocaokhaiphadulieu-bao-cao-khai-pha-du-lieu/74291850>. [Truy cập 12/2023].

[4] VTC Academy Plus, “Data Mining là gì? Quy trình khai phá dữ liệu như thế nào?,” 15/09/2022. [Trực tuyến]. Địa chỉ: <https://plus.vtc.edu.vn/data-mining-la-gi>. [Truy cập 12/2023].

[5] N.T. Thịnh, “Đánh giá model trong Machine Learing,” 29/05/2022. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/danh-gia-model-trong-machine-learing-RnB5pAq7KPG.> [Truy cập 12/2023].

[6] GMO Internet Group, “Data mining là gì? Các công cụ khai phá dữ liệu tốt nhất hiện nay,” 17/07/2023. [Trực tuyến]. Địa chỉ: <https://cloud.z.com/vn/news/data-mining/>. [Truy cập 12/2023].

[7] deepak\_jain, “Data Preprocessing in Data Mining,” 06/05/2023. [Trực tuyến]. Địa chỉ: <https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/>. [Truy cập 12/2023].

[8] Machine Learning cơ bản, “Bài 4: K-means Clustering,” 01/01/2017. [Trực tuyến]. Địa chỉ: <https://machinelearningcoban.com/2017/01/01/kmeans/>. [Truy cập 12/2023].

[9] N.T. Hop, “Hierarchical clustering - Phân cụm dữ liệu,” 18/02/2020. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/hierarchical-clustering-phan-cum-du-lieu-maGK7q2elj2>. [Truy cập 12/2023].

[10] Machine Learning cơ bản, “Bài 6: K-nearest neighbors,” 08/01/2017. [Trực tuyến]. Địa chỉ: <https://machinelearningcoban.com/2017/01/08/knn/>. [Truy cập 12/2023].

[11] Machine Learning cơ bản, “Bài 32: Naive Bayes Classifier,” 08/08/2017. [Trực tuyến]. Địa chỉ: <https://machinelearningcoban.com/2017/08/08/nbc/>. [Truy cập 12/2023].

[12] Machine Learning cơ bản, “Bài 34: Decision Trees (1): Iterative Dichotomiser 3,” 14/01/2018. [Trực tuyến]. Địa chỉ: <https://machinelearningcoban.com/2018/01/14/id3/>. [Truy cập 12/2023].

[13] N.M. Đức, “Thuật toán Apriori khai phá luật kết hợp trong Data Mining,” 20/08/2019. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/thuat-toan-apriori-khai-pha-luat-ket-hop-trong-data-mining-3P0lPEv85ox>. [Truy cập 01/2024].

[14] N.N. Linh, “Ô nhiễm môi trường không khí: nguyên nhân và giải pháp khắc phục,” 20/08/2019. [Trực tuyến]. Địa chỉ: [https://www.linkedin.com/pulse/ô-nhiễm-môi-trường-không-khí-nguyên-nhân-và-giải-pháp-nguyễn](https://www.linkedin.com/pulse/%C3%B4-nhi%E1%BB%85m-m%C3%B4i-tr%C6%B0%E1%BB%9Dng-kh%C3%B4ng-kh%C3%AD-nguy%C3%AAn-nh%C3%A2n-v%C3%A0-gi%E1%BA%A3i-ph%C3%A1p-nguy%E1%BB%85n). [Truy cập 12/2023].

[15] Báo Điện Tử - ĐCSVN, “Nguyên nhân gây tình trạng ô nhiễm môi trường không khí và giải pháp khắc phục,” 19/10/2021. [Trực tuyến]. Địa chỉ: <https://dangcongsan.vn/xay-dung-xa-hoi-an-toan-truoc-thien-tai/nguyen-nhan-gay-tinh-trang-o-nhiem-moi-truong-khong-khi-va-giai-phap-khac-phuc-594455.html>. [Truy cập 01/2024].

[16] N. Liên, “Python là gì? Tổng hợp tất tần tật kiến thức về ngôn ngữ Python có thể bạn chưa biết,” 15/12/2923. [Trực tuyến]. Địa chỉ: <https://fptshop.com.vn/tin-tuc/danh-gia/python-la-gi-168825>. [Truy cập 01/2024].

[17] Pum, “Google Colab là gì? Hướng dẫn sử dụng Google Colab,” 29/07/2022. [Trực tuyến]. Địa chỉ: <https://200lab.io/blog/google-colab-la-gi/>. [Truy cập 01/2024].

[18] S. Tripathi, “Pandas Profiling (ydata-profiling) in Python: A Guide for Beginners,” 11/2023. [Trực tuyến]. Địa chỉ: <https://www.datacamp.com/tutorial/pandas-profiling-ydata-profiling-in-python-guide>. [Truy cập 01/2024].

[19] Vikashar, “Lazy Predict Library in Python for Machine Learning,” 02/01/2023. [Trực tuyến]. Địa chỉ: <https://www.geeksforgeeks.org/lazy-predict-library-in-python-for-machine-learning/>. [Truy cập 01/2024].

[20] N. Dương, “Linear Regression - Hồi quy tuyến tính trong Machine Learning,” 30/05/2017. [Trực tuyến]. Địa chỉ: <https://viblo.asia/p/linear-regression-hoi-quy-tuyen-tinh-trong-machine-learning-4P856akRlY3>. [Truy cập 01/2024].

[21] M. Vivekanandan, “Grid Search Technique for Hyperparameter Tuning,” 17/08/2023. [Trực tuyến]. Địa chỉ: <https://www.linkedin.com/pulse/grid-search-technique-hyperparameter-tuning-madhavan-vivekanandan>. [Truy cập 01/2024].

[22] GeeksforGeeks, “Confusion Matrix in Machine Learning,” 16/03/2023. [Trực tuyến]. Địa chỉ: <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>. [Truy cập 01/2024].

[23] Dr.Air, “Điểm tên các chất nguy hại gây ô nhiễm không khí,” 06/2021. [Trực tuyến]. Địa chỉ: <https://xulykhoibep.com/diem-ten-cac-chat-nguy-hai-gay-o-nhiem-khong-khi/#:~:text=C%C3%A1c%20ch%E1%BA%A5t%20g%C3%A2y%20%C3%B4%20nhi%E1%BB%85m%20kh%C3%B4ng%20kh%C3%AD%20%C4%91%C6%B0%E1%BB%A3c%20quan%20t%C3%A2m,v%C3%A0%20c%C3%A1c%20ngu%E1%BB%93n%20%C4%91%E1%BB%91t%20kh%C3%A1c.>. [Truy cập 01/2024].

[24] IQAir, “Chỉ số chất lượng không khí (AQI) là gì?,” 04/08/2018. [Trực tuyến]. Địa chỉ:<https://www.iqair.com/vi/newsroom/what-is-aqi>. [Truy cập 01/2024].