**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**Khoa Viễn thông I**

A logo with a star

Description automatically generated

**BÁO CÁO**

**Học phần: Lưu trữ và Phân tích dữ liệu**

**Đề tài: Phân tích, dự đoán chất lượng không khí tại Delhi**

**thông qua các mô hình học máy**

|  |  |
| --- | --- |
| **Nhóm 2:** | |
| **Hoàng Thành Long** | **B20DCVT233** |
| **Tạ Văn An** | **B20DCVT009** |
| **Nguyễn Trường Xuân** | **B20DCVT425** |
| **Kim Tuấn Kiên** | **B20DCVT201** |
| **Mai Đức Vinh** | **B20DCVT417** |
| **Giảng viên:** | **TS. Nguyễn Minh Tuấn** |

**Mục lục**

[DANH MỤC HÌNH ẢNH 4](#_Toc168255179)

[CHƯƠNG 1: TỔNG QUAN VỀ TẬP DỮ LIỆU VÀ MÔ HÌNH HỌC MÁY 5](#_Toc168255180)

[1.1. Tập dữ liệu sử dụng 5](#_Toc168255181)

[1.1.1. Tổng quan 5](#_Toc168255182)

[1.1.2. Chi tiết 5](#_Toc168255183)

[1.1.3. Mục đích sử dụng và ứng dụng của tập dữ liệu 7](#_Toc168255184)

[1.2. Các mô hình học máy 7](#_Toc168255185)

[1.2.1. Random Forest 7](#_Toc168255186)

[1.2.1.1. Khái niệm 7](#_Toc168255187)

[1.2.1.2. Nguyên lí hoạt động 8](#_Toc168255188)

[1.2.1.3. Đặc điểm 10](#_Toc168255189)

[1.2.2. Decision Tree 10](#_Toc168255190)

[1.2.2.1. Khái niệm 10](#_Toc168255191)

[1.2.2.2. Cấu trúc Decision Tree 11](#_Toc168255192)

[1.2.2.3. Các thuật toán liên quan tới Decition Tree 12](#_Toc168255193)

[1.2.2.4. Ưu, nhược điểm của Decision Tree 13](#_Toc168255194)

[1.2.2.4.1. Ưu điểm 13](#_Toc168255195)

[1.2.2.4.2. Nhược điểm 13](#_Toc168255196)

[1.2.3. Supported Vector Machine 13](#_Toc168255197)

[1.2.4. K-Nearest Neighbors 15](#_Toc168255198)

[CHƯƠNG 2: MÔ HÌNH HỆ THỐNG VÀ ĐÁNH GIÁ KẾT QUẢ 17](#_Toc168255199)

[2.1. Mô hình hệ thống 17](#_Toc168255200)

[2.1.1. Sơ đồ hệ thống 17](#_Toc168255201)

[2.1.2. Thực hiện 17](#_Toc168255202)

[2.1.2.1. Tiền xử lý dữ liệu 17](#_Toc168255203)

[2.1.2.2. Triển khai mô hình 21](#_Toc168255204)

[Random Forest 21](#_Toc168255205)

[Decision Tree 23](#_Toc168255206)

[Support Vector Machine 24](#_Toc168255207)

[K-Nearest Neighbors 25](#_Toc168255208)

[2.2. Đánh giá kết quả phân loại 25](#_Toc168255209)

[2.3. Đánh giá kết quả dự đoán 26](#_Toc168255214)

[CHƯƠNG 3: KẾT LUẬN 29](#_Toc168255215)

[3.1. Mức độ ô nhiễm không khí ở Delhi 29](#_Toc168255216)

[3.2. Khoảng thời gian có chất lượng không khí và ô nhiễm ở Delhi tệ nhất 30](#_Toc168255217)

[3.3. Nguyên nhân chính dẫn đến mức độ ô nhiễm cao ở Delhi 31](#_Toc168255218)

[3.4. Ảnh hưởng sức khỏe bởi ô nhiễm không khí ở Delhi 32](#_Toc168255219)

[3.5. Giải pháp giảm mức độ ô nhiễm và cải thiện chất lượng không khí 33](#_Toc168255220)

# **DANH MỤC HÌNH ẢNH**

[Hình 1‑1: Mô hình học máy RandomForest 8](#_Toc168255157)

[Hình 1‑2: Random Forest 9](#_Toc168255158)

[Hình 1‑3: Cấu trúc Decision Tree 11](#_Toc168255159)

[Hình 1‑4: Siêu phẳng trong không 2 chiều và 3 chiều 14](#_Toc168255160)

[Hình 1‑5: Phân loại tuyến tính và phi tuyến tính 14](#_Toc168255161)

[Hình 1‑6: Ánh xạ dữ liệu từ không gian 2 chiều sang không gian 3 chiều 15](#_Toc168255162)

[Hình 2‑1: Sơ đồ hệ thống cho từng mô hình học máy 17](#_Toc168255163)

[Hình 2‑2: Phần trăm mất mát dữ liệu của từng thành phố trong tập dữ liệu 18](#_Toc168255164)

[Hình 2‑3: Số lượng và tỉ lệ mất mát dữ liệu của từng đặc trưng của tập dữ liệu Delhi 18](#_Toc168255165)

[Hình 2‑4: Chỉ số AQI trung bình theo ngày-tháng-năm của tập dữ liệu Delhi 20](#_Toc168255166)

[Hình 2‑5: Ma trận nhầm lẫn của mô hình Random Forest 21](#_Toc168255167)

[Hình 2‑6: Tham số đánh giá mô hình RandomForest 22](#_Toc168255168)

[Hình 2‑7: Ma trận nhầm lẫn mô hình Decision Tree 23](#_Toc168255169)

[Hình 2‑8: Tham số đánh giá mô hình Decision Tree 23](#_Toc168255170)

[Hình 2‑9: Ma trận nhầm lẫn của mô hình SVM 24](#_Toc168255171)

[Hình 2‑10: Tham số đánh giá mô hình SVM 24](#_Toc168255172)

[Hình 2‑11: Độ chính xác của mô hình KNN theo từng giá trị k 25](#_Toc168255173)

[Hình 2‑12: Ma trận nhầm lẫn của mô hình KNN 25](#_Toc168255174)

[Hình 2‑13: Tập dữ liệu ban đầu và kết quả dự đoán sử dụng mô hình LSTM cho giá trị AQI 26](#_Toc168255175)

[Hình 2‑14: Tỉ lệ giống và khác nhau giữa tập dữ liệu gốc và kết quả dự đoán 27](#_Toc168255176)

[Hình 2‑15: Tỉ lệ giống và khác nhau theo từng nhãn của tập dữ liệu gốc và kết quả dự đoán 28](#_Toc168255177)

[Hình 3‑1: Thành phố Delhi - Ấn Độ 29](#_Toc168255178)

# **TỔNG QUAN VỀ TẬP DỮ LIỆU VÀ MÔ HÌNH HỌC MÁY**

## **Tập dữ liệu sử dụng**

### **Tổng quan**

Thời gian: 01/01/2015 đến 01/07/2020.

Địa điểm: Thành phố Delhi - Ấn Độ.

Phương pháp thu thập tập dữ liệu:

* **Mạng lưới cố định của các trạm giám sát**: Các trạm giám sát cố định được phân bố rải rác trong toàn thành phố để đo lường chất lượng không khí. Các trạm này thường được trang bị các cảm biến đo lường khác nhau để đo các chất gây ô nhiễm như PM2.5, PM10, NO2, SO2, O3, CO…
* **Sử dụng drones**: Đối với các vùng khó tiếp cận hoặc nơi có các vấn đề về an ninh, việc sử dụng drones có thể là một phương pháp hiệu quả để thu thập dữ liệu không khí. Drones có thể được trang bị cảm biến đo chất lượng không khí và có thể bay qua các khu vực khác nhau của thành phố để thu thập dữ liệu.
* **Thiết bị di động giám sát**: Ngoài các trạm cố định, các thiết bị di động như xe hơi, xe buýt hoặc thậm chí là đồng hồ thông minh được trang bị cảm biến đo chất lượng không khí để thu thập dữ liệu khi di chuyển qua các khu vực khác nhau trong thành phố. Điều này giúp thu thập dữ liệu từ nhiều vị trí và điều kiện khác nhau.

### **Chi tiết**

Kích thước của tập dữ liệu gồm 48193 mẫu.

Mô tả các đặc trưng (feature hay các cột của tập dữ liệu)

Tập dữ liệu cung cấp các đặc trưng sau:

* Datetime: Đây là thời gian và ngày ghi nhận các thông số khí quyển.
* PM2.5: Nồng độ hạt nhỏ có đường kính nhỏ hơn hoặc bằng 2.5 micromet trong không khí (micrograms/m^3).
* PM10: Nồng độ hạt lớn hơn 10 micromet trong không khí (micrograms/m^3).
* NO: Nồng độ khí oxit nitơ (micrograms/m^3).
* NO2: Nồng độ khí dioxide nitơ (micrograms/m^3).
* NOx: Tổng nồng độ của khí oxit nitơ (NO) và dioxide nitơ (NO2) (micrograms/m^3).
* NH3: Nồng độ khí amoniac (micrograms/m^3).
* CO: Nồng độ khí carbon monoxit (micrograms/m^3).
* SO2: Nồng độ khí dioxide lưu huỳnh (micrograms/m^3).
* O3: Nồng độ khí ozone (micrograms/m^3).
* Benzene: Nồng độ của benzen trong không khí (micrograms/m^3).
* Toluene: Nồng độ của toluen trong không khí (micrograms/m^3).
* Xylene: Nồng độ của xylene trong không khí (micrograms/m^3).
* AQI: Chỉ số chất lượng không khí (Air Quality Index) được tính toán dựa trên nồng độ của các chất gây ô nhiễm khí quyển.
* AQI\_Bucket: Phân loại mức độ ô nhiễm không khí dựa trên AQI, có thể là “Severe”.

Các hàng của tập dữ liệu có thể xem xét một số thông tin quan trọng như giá trị trung bình, giá trị tối đa, giá trị tối thiểu, độ lệch chuẩn, và cũng có thể vẽ biểu đồ để hiểu rõ hơn về xu hướng và phân phối của dữ liệu:

* Thống kê mô tả cơ bản: Tính mean (trung bình), median (trung vị), mode (mođô), min (giá trị nhỏ nhất), max (giá trị lớn nhất), và standard deviation (độ lệch chuẩn) cho từng cột dữ liệu.
* Phân phối dữ liệu: Vẽ histogram hoặc box plot cho từng cột dữ liệu để xem phân phối của chúng.
* Tìm kiếm các giá trị bất thường (outliers): Sử dụng box plot hoặc đơn giản là xem các giá trị rời rạc có giá trị lớn hoặc nhỏ hơn nhiều so với các giá trị khác.
* Phân tích mối tương quan: Tìm mối tương quan giữa các cột dữ liệu bằng cách tính ma trận tương quan hoặc vẽ biểu đồ scatter plot.
* Phân tích xu hướng thời gian (nếu thích hợp): Nếu cột "Datetime" có dạng thời gian, bạn có thể phân tích xu hướng thay đổi của các chỉ số qua thời gian.

### **Mục đích sử dụng và ứng dụng của tập dữ liệu**

Đánh giá chất lượng không khí cung cấp thông tin về chất lượng không khí tại một số địa điểm cụ thể, giúp các nhà quản lý môi trường, chính phủ, và cộng đồng địa phương hiểu rõ hơn về mức độ ô nhiễm không khí.

Dự đoán và phòng ngừa ô nhiễm không khí phân tích xu hướng và biến động của các chỉ số khí quyển có thể giúp dự đoán và phòng ngừa ô nhiễm không khí trong tương lai. Các biện pháp kiểm soát và giảm thiểu ô nhiễm có thể được áp dụng dựa trên thông tin này.

Dữ liệu về chất lượng không khí có thể được sử dụng trong nghiên cứu y tế và môi trường để hiểu tác động của ô nhiễm không khí đối với sức khỏe con người và môi trường.

Dữ liệu có thể được sử dụng để so sánh với các tiêu chuẩn và quy định về chất lượng không khí được đề xuất hoặc áp dụng bởi các tổ chức y tế và môi trường.

## **Các mô hình học máy**

### **Random Forest**

#### Khái niệm

Thuật toán Random Forest là một kỹ thuật học cây mạnh mẽ trong học máy. Nó hoạt động bằng cách tạo ra một số Cây quyết định trong giai đoạn đào tạo. Mỗi cây được xây dựng bằng cách sử dụng một tập hợp con ngẫu nhiên của tập dữ liệu để đo lường một tập hợp con ngẫu nhiên các tính năng trong mỗi phân vùng. Tính ngẫu nhiên này tạo ra sự biến đổi giữa các cây riêng lẻ, giảm nguy cơ trang bị quá mức và cải thiện hiệu suất dự đoán tổng thể.

Trong dự đoán, thuật toán tổng hợp kết quả của tất cả các cây, bằng cách bỏ phiếu (đối với nhiệm vụ phân loại) hoặc bằng cách lấy trung bình (đối với nhiệm vụ hồi quy). Quá trình ra quyết định hợp tác này, được hỗ trợ bởi nhiều cây với thông tin chi tiết của chúng, cung cấp một ví dụ về kết quả ổn định và chính xác .

Random Forest được sử dụng rộng rãi cho các chức năng phân loại và hồi quy, được biết đến với khả năng xử lý dữ liệu phức tạp, giảm tình trạng trang bị quá mức và cung cấp dự báo đáng tin cậy trong các môi trường khác nhau.

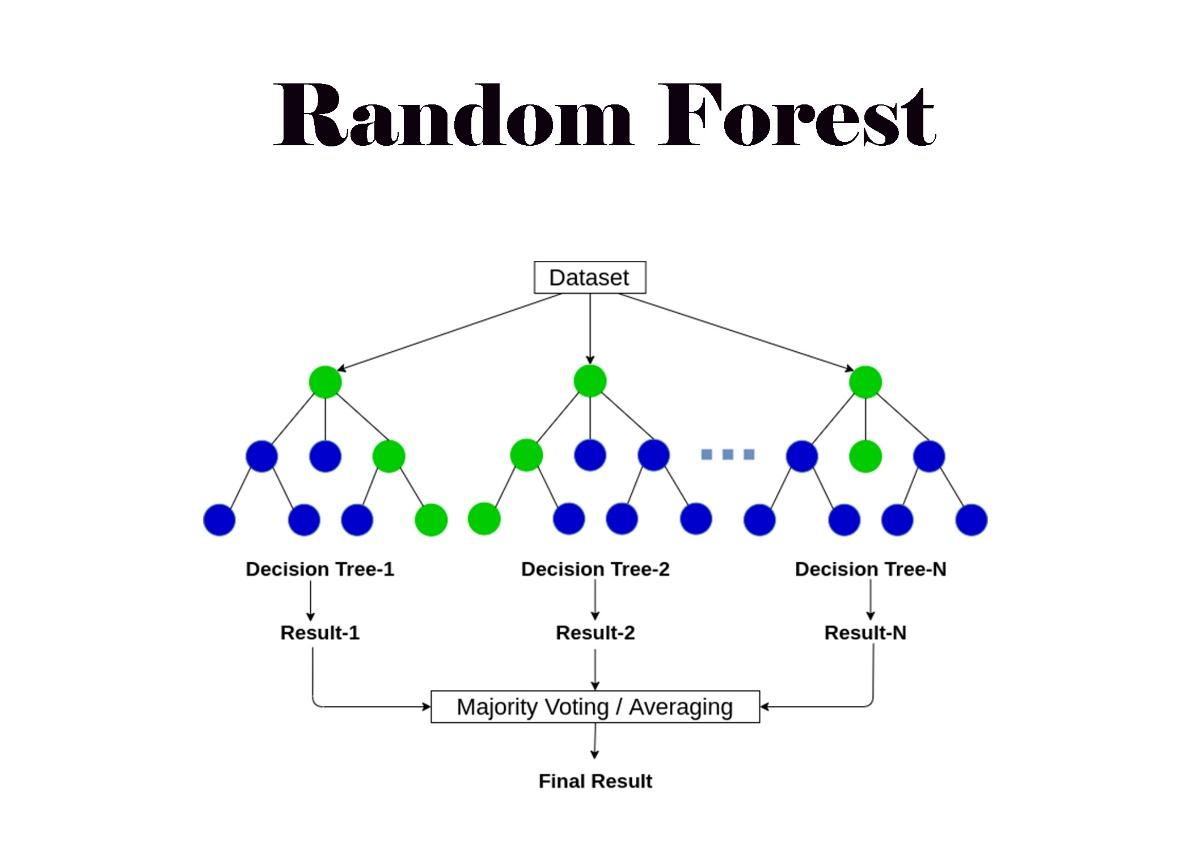
A diagram of a training data

Description automatically generated

Hình 1‑1: Mô hình học máy RandomForest

#### Nguyên lí hoạt động

Tập hợp các Cây quyết định: Random Forest tận dụng sức mạnh của việc học tổng hợp bằng cách xây dựng một tập hợp Cây quyết định. Những cây này giống như các chuyên gia riêng lẻ, mỗi chuyên gia chuyên về một khía cạnh cụ thể của dữ liệu. Điều quan trọng là chúng hoạt động độc lập, giảm thiểu rủi ro mô hình bị ảnh hưởng quá mức bởi các sắc thái của một cây duy nhất.



Hình 1‑2: Random Forest

Lựa chọn tính năng ngẫu nhiên: Để đảm bảo rằng mỗi cây quyết định trong tập hợp mang lại một góc nhìn độc đáo, Random Forest sử dụng lựa chọn tính năng ngẫu nhiên . Trong quá trình huấn luyện mỗi cây, một tập hợp con các đặc tính ngẫu nhiên sẽ được chọn. Tính ngẫu nhiên này đảm bảo rằng mỗi cây tập trung vào các khía cạnh khác nhau của dữ liệu, thúc đẩy một tập hợp các yếu tố dự đoán đa dạng trong quần thể.

Tổng hợp hoặc đóng bao Bootstrap: Kỹ thuật đóng bao là nền tảng trong chiến lược đào tạo của Random Forest, bao gồm việc tạo nhiều mẫu bootstrap từ tập dữ liệu gốc, cho phép lấy mẫu các phiên bản thay thế. Điều này dẫn đến các tập hợp dữ liệu khác nhau cho mỗi cây quyết định, tạo ra sự thay đổi trong quá trình đào tạo và làm cho mô hình trở nên mạnh mẽ hơn.

Ra quyết định và biểu quyết: Khi đưa ra dự đoán, mỗi cây quyết định trong Random Forest sẽ bỏ phiếu. Đối với các nhiệm vụ phân loại , dự đoán cuối cùng được xác định theo chế độ (dự đoán thường xuyên nhất) trên tất cả các cây. Trong các tác vụ hồi quy , giá trị trung bình của các dự đoán cây riêng lẻ được lấy. Cơ chế bỏ phiếu nội bộ này đảm bảo quá trình ra quyết định tập thể và cân bằng.

#### Đặc điểm

* Bản chất Ensemble: Random Forest kết hợp nhiều cây quyết định với nhau. Mỗi cây quyết định được huấn luyện trên một phần của dữ liệu (bootstrap sample), và một số lượng ngẫu nhiên các đặc trưng (feature) được chọn cho mỗi nút trong quá trình phân chia. Việc kết hợp nhiều cây giúp giảm thiểu overfitting và cải thiện tính tổng quát hóa của mô hình.
* Tính linh hoạt: Random Forest có thể được áp dụng cho cả bài toán phân loại và hồi quy. Đối với bài toán phân loại, nó có thể xử lý nhiều loại dữ liệu và có khả năng xử lý cả dữ liệu không cân bằng.
* Hiệu suất cao: Random Forest thường cho hiệu suất tốt trên nhiều loại dữ liệu, bao gồm cả dữ liệu có tính không tuyến tính và dữ liệu không cân bằng. Nó thường không đòi hỏi nhiều thời gian để điều chỉnh các siêu tham số và thường đưa ra các dự đoán tốt mà không cần quá nhiều tinh chỉnh.
* Khả năng xử lý dữ liệu lớn: Random Forest có thể xử lý được các tập dữ liệu lớn với số lượng đặc trưng lớn mà vẫn giữ được hiệu suất.
* Interpretability (Khả năng giải thích): Mặc dù mỗi cây quyết định trong Random Forest có thể không dễ hiểu, nhưng khi kết hợp lại với nhau, mô hình có thể cung cấp cái nhìn tổng quan về cách các đặc trưng ảnh hưởng đến dự đoán cuối cùng.

### **Decision Tree**

#### Khái niệm

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật (series of rules). Khi cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết (unseen data).

Thuật toán cây quyết định thuộc họ thuật toán học có giám sát. Không giống như các thuật toán học có giám sát khác, thuật toán cây quyết định cũng có thể được sử dụng để giải các bài toán hồi quy và phân loại.

Các loại cây quyết định dựa trên mục tiêu mà chúng ta có, nó có thể có 2 loại:

* Cây quyết định biến phân loại: Cây quyết định có biến mục tiêu phân loại nên được gọi là cây quyết định biến phân loại.
* Cây quyết định biến liên tục: Cây quyết định có biến mục tiêu liên tục nên được gọi là cây quyết định biến liên tục.

#### Cấu trúc Decision Tree

Decision Tree gồm 3 phần chính: 1 node gốc (root node), những node lá (leaf nodes) và các nhánh của nó (branches):

* Node gốc là điểm bắt đầu của cây quyết định, nó đại diện cho toàn bộ quần thể hoặc mẫu và phần này được chia thành 2 hay nhiều tập đông nhất.
* Node lá là các node không phân chia, và được gọi là node lá hoặc node đấu cuối.
* Nhánh: một phần con của toàn bộ cây được gọi là nhánh hoặc cây con.

A diagram of a body temperature

Description automatically generated

Hình 1‑3: Cấu trúc Decision Tree

Một số tiêu chí có thể được lựa chọn trong Decision Tree:

* Entropy
* Information gain
* Gini index
* Gain Ratio
* Reduction in Variance
* Chi-Square

#### Các thuật toán liên quan tới Decition Tree

ID3 (Iterative Dichotomiser 3): Xây dụng cây quyết định bằng cách sử dụng phương pháp tìm kiếm tham lam từ trên xuống thông qua không gian của các nhánh có thể mà không cần quay lại. Có thể tóm tắt như sau:

* Lấy tất cả các thuộc tính chưa được sử dụng và đếm entropy liên quan đến mẫu thử của các thuộc tính đó
* Chọn thuộc tính có entropy lớn nhất
* Nối node với thuộc tính đó

C4.5: là thuật toán phân lớp dữ liệu dựa trên cây quyết định hiệu quả và phổ biến trong những ứng dụng khai phá cơ sở dữ liệu có kích thước nhỏ. C4.5 xây dựng cây quyết định từ tập training data tương tự như ID3:

* Tập training data S gồm các mẫu đã được phân  loại sẵn s1, s2,… Mỗi mẫu si = x1,x2,… với x1, x2 là 1 vector biểu diễn cho thuộc tính hoặc đặc điểm của mẫu. Một tập vector C = c1,c2,… với c1, c2 biểu diễn cho mỗi lớp mà mỗi mẫu thuộc về.
* Với mỗi lớp của cây, C4.5 chọn một thuộc tính của dữ liệu mà phân chia tập các mẫu thành nhiều tập con, được nâng cao chất lượng một cách hiệu quả nhất. Tiêu chuẩn của nó là thu được **information gain** (sự khác biệt về entropy) – kết quả từ việc chọn một thuộc tính cho việc chia tách dữ liệu. Quyết định đưa ra dựa trên tập thuộc tính với information gain được chuẩn hóa cao nhất. Thuật toán C4.5 sau đó sẽ lặp lại với các danh sách con nhỏ hơn.

Ngoài ID3, C4.5 thì còn một số thuật toán khác như:

* Thuật toán CHAID: tạo cây quyết định bằng cách sử dụng thống kê chi-square để xác định các phân tách tối ưu. Các biến mục tiêu đầu vào có thể là số (liên tục) hoặc phân loại.
* Thuật toán C&R: sử dụng phân vùng đệ quy để chia cây. Tham biến mục tiêu có thể dạng số hoặc phân loại.
* MARS.
* Conditional Inference Trees.

#### Ưu, nhược điểm của Decision Tree

##### Ưu điểm

Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bới những lợi ích của nó:

* Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
* Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả
* Có thể làm việc với cả dữ liệu số và dữ liệu phân loại
* Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
* Có khả năng là việc với dữ liệu lớn

##### Nhược điểm

Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể:

* Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.
* Cây quyết định hay gặp vấn đề [over-fitting](https://trituenhantao.io/tu-dien-thuat-ngu/overfitting/).

### **Supported Vector Machine**

Là thuật toán học máy có giám sát sử dụng phân loại dữ liệu bằng cách tìm 1 đường thẳng hoặc 1 siêu phẳng tối ưu nhằm tối đa hóa khoảng cách giữa mỗi lớp trong không gian N chiều

A graph of a line and a hyperplane line

Description automatically generated

Hình 1‑4: Siêu phẳng trong không 2 chiều và 3 chiều

SVM được sử dụng rộng rãi vì nó có thể xử lý các bài toán phân loại tuyến tính và phi tuyến. Với các bài toán không thể phân chia tuyến tính, các hàm kernel được sử dụng để chuyển đổi dữ liệu không gian có chiều cao hơn nhằm cho phép phân chia tuyến tính

A diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of a diagram of

Description automatically generated

Hình 1‑5: Phân loại tuyến tính và phi tuyến tính

Các hàm kernel hay thủ thuật kernel (kernel trick) được sử dụng với bài toán phi tuyến. Các hàm này sẽ ánh xạ dữ liệu từ không gian ít chiều sang không gian nhiều chiều hơn để tìm ra siêu mặt phẳng phân tách dữ liệu.

A diagram of a graph

Description automatically generated

Hình 1‑6: Ánh xạ dữ liệu từ không gian 2 chiều sang không gian 3 chiều

Có một số hàm kernel: Linear kernel, Polynimial kernel, Radial basis function (RBF) kernel, Sigmoid kernel.

### **K-Nearest Neighbors**

KNN là một trong những thuật toán phân loại cơ bản nhưng cần thiết nhất trong học máy. Nó thuộc miền học có giám sát và có ứng dụng mạnh mẽ trong nhận dạng mẫu, khai thác dữ liệu và phát hiện xâm nhập. Nó được sử dụng rộng rãi trong các tình huống thực tế vì nó không có tham số, nghĩa là nó không đưa ra bất kỳ giả định cơ bản nào về việc phân phối dữ liệu (trái ngược với các thuật toán khác như GMM, giả định phân phối Gaussian của dữ liệu đã cho). Chúng tôi được cung cấp một số dữ liệu trước đó (còn gọi là dữ liệu huấn luyện), phân loại tọa độ thành các nhóm được xác định bởi một thuộc tính.

K-NN không yêu cầu bất kỳ giả định nào về phân phối dữ liệu cơ bản. Nó cũng có thể xử lý cả dữ liệu số và dữ liệu phân loại, khiến nó trở thành lựa chọn linh hoạt cho nhiều loại tập dữ liệu khác nhau trong các nhiệm vụ phân loại và hồi quy. Đây là một phương pháp phi tham số đưa ra dự đoán dựa trên sự giống nhau của các điểm dữ liệu trong một tập dữ liệu nhất định. K-NN ít nhạy cảm hơn với các ngoại lệ so với các thuật toán khác. Thuật toán K-NN hoạt động bằng cách tìm K lân cận gần nhất với một điểm dữ liệu nhất định dựa trên thước đo khoảng cách, chẳng hạn như khoảng cách Euclide. Lớp hoặc giá trị của điểm dữ liệu sau đó được xác định bằng đa số phiếu bầu hoặc mức trung bình của K hàng xóm. Cách tiếp cận này cho phép thuật toán thích ứng với các mẫu khác nhau và đưa ra dự đoán dựa trên cấu trúc cục bộ của dữ liệu. Số liệu khoảng cách được sử dụng trong thuật toán KNN Như chúng ta biết rằng thuật toán KNN giúp chúng ta xác định các điểm hoặc nhóm gần nhất cho một điểm truy vấn. Nhưng để xác định các nhóm gần nhất hoặc các điểm gần nhất cho một điểm truy vấn, chúng ta cần một số số liệu. Với mục đích này, chúng tôi sử dụng các số liệu khoảng cách dưới đây:

Khoảng cách Euclid:

Khoảng cách Manhattan:

Khoảng cách Minkowski:

# **MÔ HÌNH HỆ THỐNG VÀ ĐÁNH GIÁ KẾT QUẢ**

## **Mô hình hệ thống**

### **Sơ đồ hệ thống**

A diagram of a diagram

Description automatically generated with medium confidence

Hình 2‑1: Sơ đồ hệ thống cho từng mô hình học máy

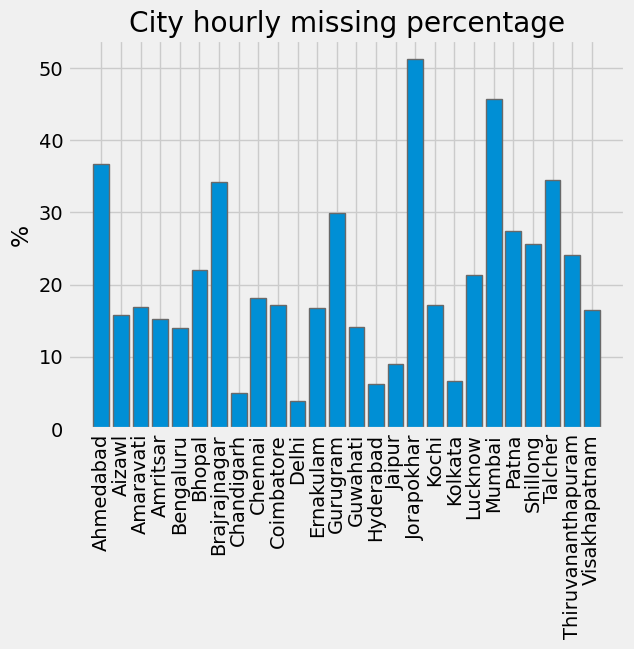
Các bước thực hiện:

1. Trích xuất dữ liệu
2. Xử lý mất mát dữ liệu
3. Xử lý giá trị ngoại lệ
4. Tính chỉ số AQI
5. Đưa tập dữ liệu vào các mô hình học máy khác nhau
6. Biểu diễn kết quả tương ứng với từng mô hình học máy
7. So sánh kết quả để chọn mô hình phù hợp nhất

### **Thực hiện**

#### Tiền xử lý dữ liệu

**Trích xuất dữ liệu**



Hình 2‑2: Phần trăm mất mát dữ liệu của từng thành phố trong tập dữ liệu

Trích xuất dữ liệu của thành phố có tỉ lệ mất mát dữ liệu thấp nhất – Delhi.

**Xử lý mất mát dữ liệu**

|  |  |
| --- | --- |
| A graph of blue rectangular bars with black text  Description automatically generated | A graph with blue bars  Description automatically generated |

Hình 2‑3: Số lượng và tỉ lệ mất mát dữ liệu của từng đặc trưng của tập dữ liệu Delhi

Để xử lý mất mát dữ liệu hàm *interpolate* của thư viện *pandas* với phương thức mặc định là *linear(tuyến tính).*

**Tính chỉ số AQI**

Trong đó,

* Iₚ: chỉ số chất ô nhiễm
* Cₚ: nồng độ chất ô nhiễm
* : giá trị nồng độ cao nhất trong khoảng
* : giá trị nồng độ nhỏ nhất trong khoảng
* : chỉ số AQI cao nhất tương ứng với
* : chỉ số AQI thấp nhất tương ứng với

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AQI  (range) | PM10  (24hr) | PM2.5  (24hr) | NO2  (24hr) | O3  (8hr) | CO  (8hr) | SO2  (24hr) | NH3  (24hr) | Pb  (24hr) |
| Good  (0–50) | 0–50 | 0–30 | 0–40 | 0–50 | 0–1.0 | 0–40 | 0–200 | 0–0.5 |
| Satisfactory  (51–100) | 51–100 | 31–60 | 41–80 | 51–100 | 1.1–2.0 | 41–80 | 201–400 | 0.5–1.0 |
| Moderate  (101–200) | 101–250 | 61–90 | 81–180 | 101–168 | 2.1–10 | 81–380 | 401–800 | 1.1–2.0 |
| Poor  (201–300) | 251–350 | 91–120 | 181–280 | 169–208 | 10–17 | 381–800 | 801–1200 | 2.1–3.0 |
| Very Poor  (301–400) | 351–430 | 121–250 | 281–400 | 209–748 | 17–34 | 801–1600 | 1200–1800 | 3.1–3.5 |
| Severe  (401–500) | 430+ | 250+ | 400+ | 748+ | 34+ | 1600+ | 1800+ | 3.5+ |

Bảng tính toán các chỉ số để đánh giá mức độ ô nhiễm

|  |  |
| --- | --- |
| Đồ thị AQI trung bình | Biểu đồ hộp – Boxplot |
| A graph with blue lines  Description automatically generated | A graph showing different colored lines  Description automatically generated with medium confidence |
| A graph with a line  Description automatically generated | A graph of different colored squares  Description automatically generated |
| A graph with a line going up  Description automatically generated | A graph of different colored squares  Description automatically generated |

Hình 2‑4: Chỉ số AQI trung bình theo ngày-tháng-năm của tập dữ liệu Delhi

#### Triển khai mô hình

##### Random Forest

**A graph showing the difference between a number of different numbers

Description automatically generated with medium confidence**

Hình 2‑5: Ma trận nhầm lẫn của mô hình Random Forest

**A chart of different colors

Description automatically generated**

**A diagram of a model evaluation metrics

Description automatically generated**

Hình 2‑6: Tham số đánh giá mô hình RandomForest

##### Decision Tree

|  |  |
| --- | --- |
| **A chart with numbers and symbols  Description automatically generated**  Hình 2‑7: Ma trận nhầm lẫn mô hình Decision Tree | |
| **A graph of a number of columns  Description automatically generated with medium confidence** | **A graph of a model evaluation  Description automatically generated with medium confidence** |

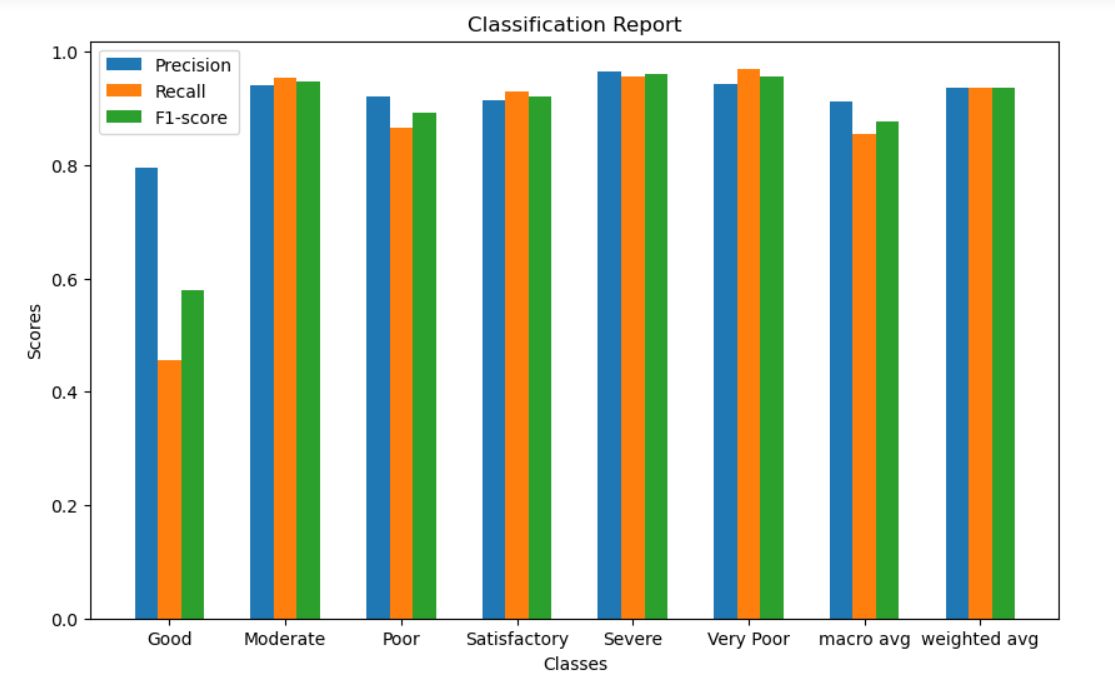
Hình 2‑8: Tham số đánh giá mô hình Decision Tree

##### Support Vector Machine

**A graph of a number of people

Description automatically generated with medium confidence**

Hình 2‑9: Ma trận nhầm lẫn của mô hình SVM

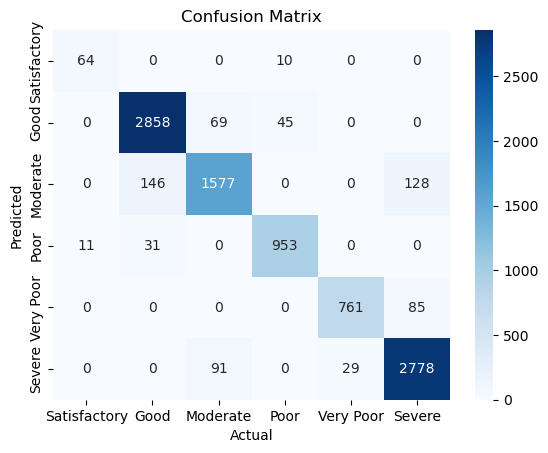
****

Hình 2‑10: Tham số đánh giá mô hình SVM

##### K-Nearest Neighbors

|  |  |
| --- | --- |
|  |  |

Hình 2‑11: Độ chính xác của mô hình KNN theo từng giá trị k

****

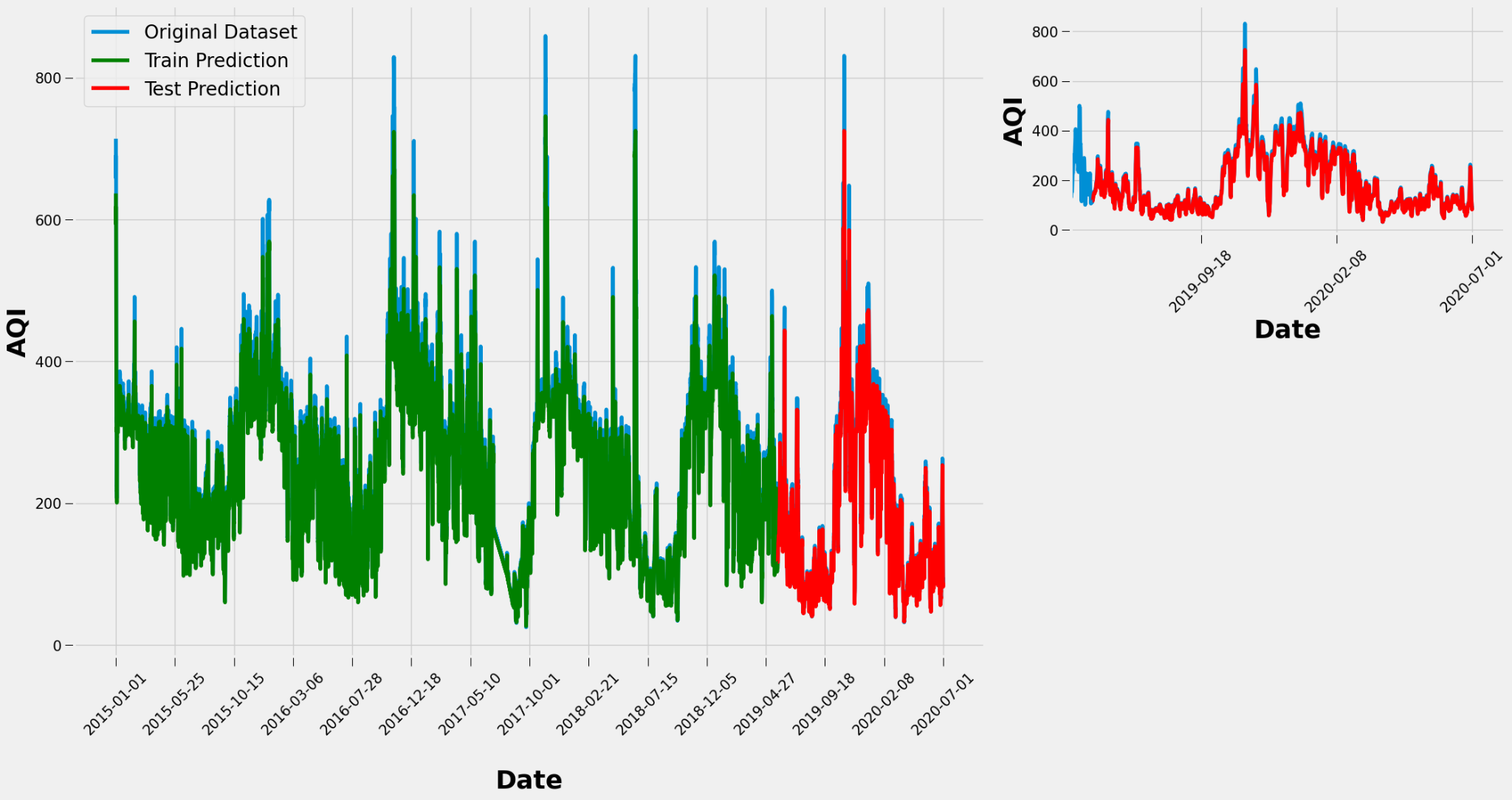
Hình 2‑12: Ma trận nhầm lẫn của mô hình KNN

## **Đánh giá kết quả phân loại**

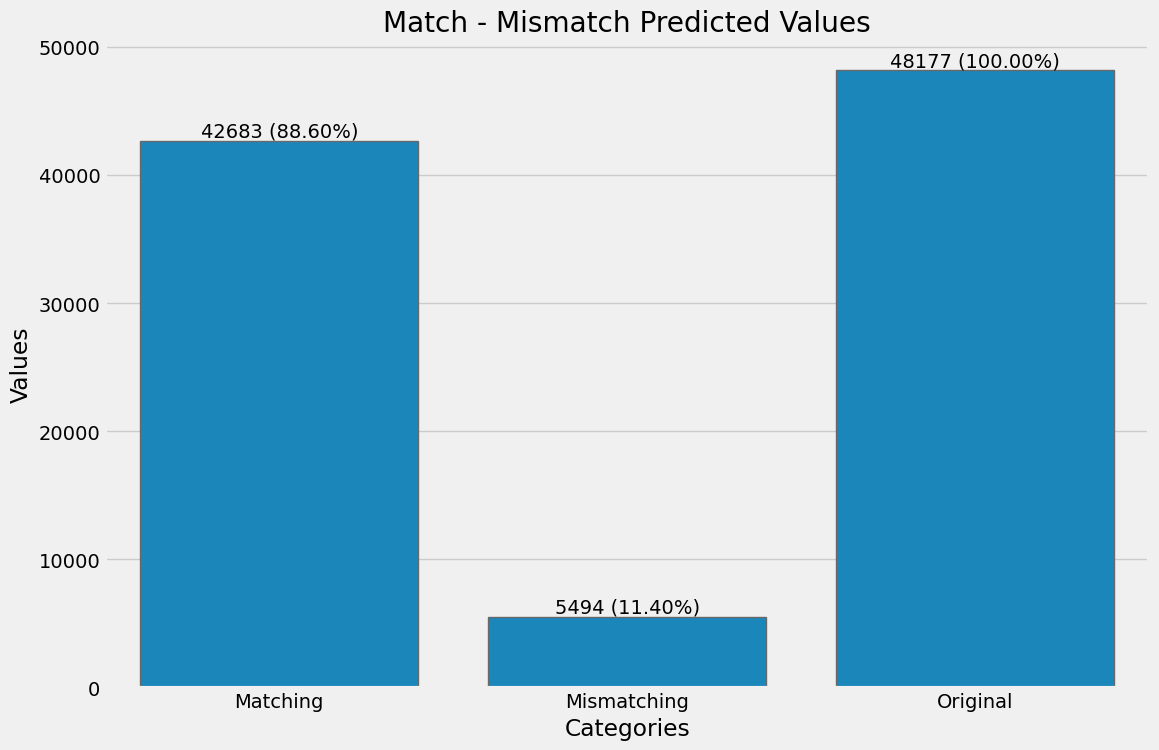
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | AuC |
| RandomForest | 0.9993 | 0.9551 | 0.9551 | 0.9972 |
| DecisionTree | 0.9754 | 0.9699 | 0.9754 | 0.9871 |
| SVM | 0.9322 | 0.9318 | 0.9322 | 0.9965 |
| K-NN | 0.933 | 0.9329 | 0.933 | 0.9968 |



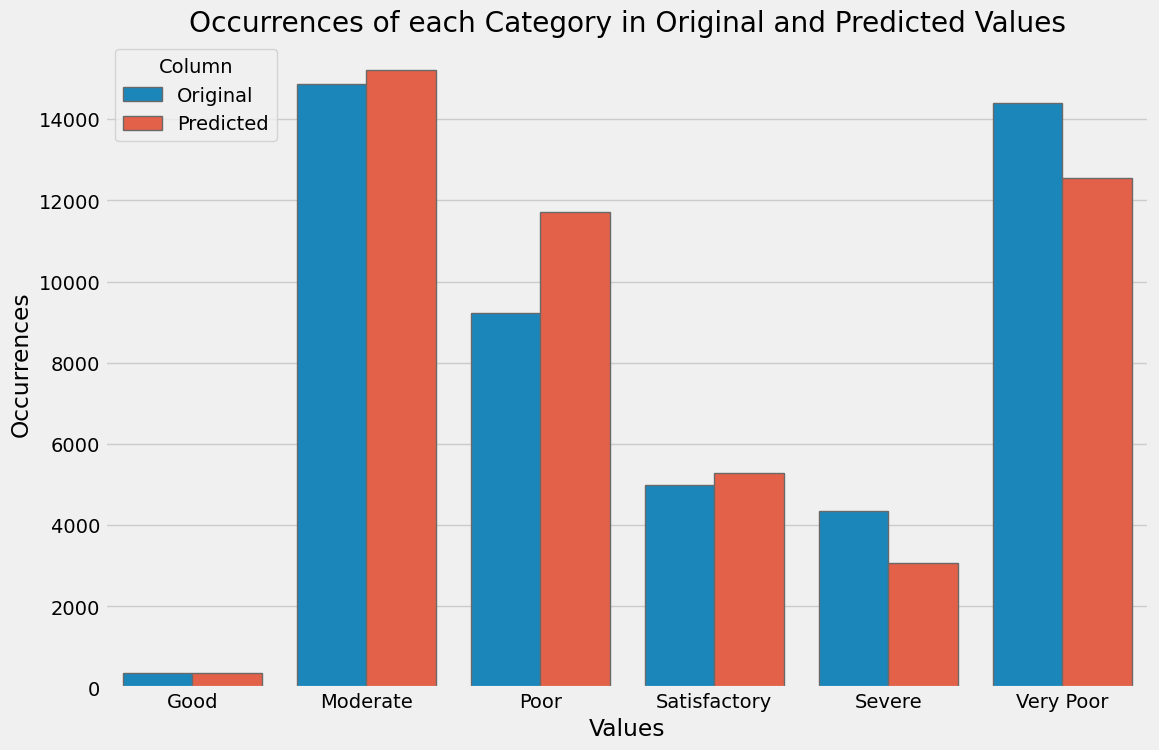
## **Đánh giá kết quả dự đoán**

****

Hình 2‑13: Tập dữ liệu ban đầu và kết quả dự đoán sử dụng mô hình LSTM cho giá trị AQI



Hình 2‑14: Tỉ lệ giống và khác nhau giữa tập dữ liệu gốc và kết quả dự đoán



Hình 2‑15: Tỉ lệ giống và khác nhau theo từng nhãn của tập dữ liệu gốc và kết quả dự đoán

# **KẾT LUẬN**

## **Mức độ ô nhiễm không khí ở Delhi**

**A large arch in the middle of a road

Description automatically generated**

Hình 3‑1: Thành phố Delhi - Ấn Độ

Là thủ đô của Ấn Độ, Delhi phải chịu mức độ ô nhiễm cao quanh năm. Mật độ hạt mịn và thô, được gọi tương ứng là PM2.5 và PM10, thường phổ biến trong không khí, cũng như các dạng chất ô nhiễm và hóa chất độc hại khác xâm nhập vào khí quyển, mỗi dạng đều có tác động bất lợi riêng đối với sức khỏe con người. Ước tính có khoảng 30,2 triệu người đăng ký sống ở Delhi tính đến năm 2020, tất cả đều bị dồn vào một khu vực tương đối nhỏ với diện tích 1.484 km2, khiến nơi đây có mật độ dân số cực cao.

Vào năm 2019, nó được xếp hạng với chỉ số PM2.5 là 98,6 µg/m³, xếp thành phố vào nhóm ‘không lành mạnh’ theo Chỉ số Chất lượng Không khí Hoa Kỳ, yêu cầu chỉ số PM2.5 nằm trong khoảng từ 55,5 đến 150,4 µg/m³. Việc một thành phố nằm trong nhóm không lành mạnh tính theo mức trung bình quanh năm, đó là một dấu hiệu cho thấy chất lượng không khí có những mối nguy hiểm đáng kể đối với sức khỏe. Mức độ ô nhiễm trên mặt đất phải cực kỳ cao mới có thể duy trì được mức trung bình như vậy thành phố cũng phải vật lộn để theo kịp tốc độ phát triển không ngừng. mở rộng dân số. Điều này sẽ đặt thành phố dưới áp lực to lớn để cải thiện cơ sở hạ tầng dân cư và thương mại. Như vậy, kết quả của việc này sẽ là sự gia tăng ở hầu hết các ngành công nghiệp và sự gia tăng ô nhiễm có liên quan trực tiếp đến tăng trưởng công nghiệp và kinh tế.

Dân số nhiều hơn có nghĩa là có nhiều phương tiện giao thông di chuyển trên đường hơn. Lượng khí thải từ do những phương tiện này thải ra sẽ có khối lượng cực kỳ cao, ước tính có khoảng 11,2 triệu phương tiện cơ giới đã đăng ký trên đường vào năm 2018, tăng khoảng 27% kể từ năm 2015. Delhi là một trong những thành phố có chất lượng không khí tệ nhất thế giới.

## **Khoảng thời gian có chất lượng không khí và ô nhiễm ở Delhi tệ nhất**

Xếp hạng AQI của Hoa Kỳ cũng như chỉ số PM10 và PM2.5 ở mức tồi tệ nhất trong những tháng cuối năm cũng như tháng 1, do thời tiết thay đổi và nạn đốt cháy hoa màu, đất rừng. Những hoạt động này, được gọi là canh tác nương rẫy, đã có truyền thống từ khá lâu, nhưng gần đây đã bị giám sát chặt chẽ hơn và gây áp lực cả trong nước và quốc tế để yêu cầu chúng phải dừng lại, vì tác động của những hoạt động này trên quy mô lớn, cùng với với sự bùng nổ dân số và công nghiệp, tất cả kết hợp lại để tạo ra mức độ ô nhiễm khủng khiếp như hiện nay.

Phương pháp đốt nương làm rẫy đang ngày càng trở thành vấn đề ở một số nơi trên thế giới, trong đó phần lớn châu Á cũng như các nước Đông Nam Á đang phải gánh chịu hậu quả ô nhiễm do nó gây ra. Liên quan đến Delhi, có thể thấy trên xếp hạng PM2.5 năm 2019 rằng những tháng tồi tệ nhất lần lượt là tháng 1, tháng 11 và tháng 12, với số lượng vượt xa các tháng khác một cách khá lớn. Tháng 1 đến với chỉ số 191,7 µg/m³, sau đó giảm một nửa xuống còn 84,8 µg/m³ vào tháng 2 và tiếp tục giảm trong thời gian còn lại của năm cho đến khi quan sát thấy mức tăng nhẹ vào tháng 9, rồi đột ngột tăng lên. Chỉ số PM2.5 tăng đột biến vào tháng 10 là 116,7 µg/m³, tăng 79,5 µg/m³ trong khoảng thời gian từ tháng 9 đến tháng 10, cho thấy sự thay đổi về mức độ ô nhiễm thực sự đáng kể như thế nào vào cuối năm và điều gì là phù hợp. hành động có thể được thực hiện chống lại họ, điều này sẽ được thảo luận chi tiết hơn sau này. Cuối cùng, chỉ số PM2,5 trung bình của tháng 11 và tháng 12 là 220,7 µg/m³ và 194,8 µg/m³, cả hai đều nằm trong mức 'rất có hại cho sức khỏe', một mức chất lượng không khí tệ đến mức sẽ gây ra một số thảm họa kéo dài và kéo dài. những ảnh hưởng ngắn hạn, bất kể người đó khỏe mạnh hay có bệnh lý về hệ hô hấp từ trước. Vì vậy, trong khi ba tháng trong năm có mức AQI và PM2.5 trung bình tệ nhất, thì tháng 11 lại là tháng có mức độ xấu nhất.

## **Nguyên nhân chính dẫn đến mức độ ô nhiễm cao ở Delhi**

Một trong những thủ phạm chính tất nhiên là khí thải từ phương tiện giao thông. Với số lượng ô tô khổng lồ như vậy hiện diện trong thành phố, nhiều chiếc trong số đó có động cơ có chất lượng kém (đặc biệt là đối với các xe tải, xe tải và xe buýt cũ), chúng sẽ tạo ra một lượng khói và khí thải cực lớn, bao phủ thành phố trong muội than và carbon đen, một dạng carbon rất có hại cho mọi sinh vật sống, do đốt nhiên liệu hóa thạch không đúng cách và các dạng chất hữu cơ khác nhau, như chất thải thực phẩm và đất rừng/nông nghiệp bị đốt cháy.

Như đã đề cập trước đó, sự thay đổi trong các mùa cũng góp phần khiến chất lượng không khí trở nên tồi tệ hơn, trong đó tình trạng ô nhiễm không khí vào mùa đông đặc biệt nghiêm trọng do nông dân ở các vùng xung quanh đốt xác thực vật còn sót lại hoặc gốc rạ để dọn đất sau vụ thu hoạch tháng 9.

Về mặt địa hình, Delhi đang ở thế bất lợi rõ rệt vì vị trí cũng như hướng gió và hướng gió ảnh hưởng trực tiếp đến sự tích tụ ô nhiễm trong không khí. Người ta biết rằng những cơn gió thổi vào Delhi có thể mang theo một lượng lớn khói và sương mù từ các khu vực khác trong khu vực, và khi mùa đông đến, ít khi có gió hơn, khiến ô nhiễm tích tụ bị mắc kẹt trong bầu khí quyển của Delhi. Hơn nữa, khi nhiệt độ không khí giảm trong những tháng mùa đông, các chất ô nhiễm trong không khí sẽ giảm theo độ cao, dẫn đến một loạt các chất ô nhiễm trộn lẫn với nhau, tạo thành đám mây mù đã đề cập trước đó bao phủ thành phố. Điều này đóng một vai trò lớn hơn đối với mức độ ô nhiễm hơn người ta có thể tưởng tượng, với ước tính vào năm 2019 cho biết bụi do gió thổi vào chiếm khoảng 21,5% tổng lượng ô nhiễm ở Delhi.

Xem xét lại vấn đề ô nhiễm do ô tô gây ra, với hơn 11 triệu phương tiện trên đường, lượng khói và PM2.5 do các ô tô, xe buýt và xe tải khác nhau thải ra sẽ đủ để đảm bảo đây là ô nhiễm lớn nhất gây ra mức độ ô nhiễm cao ở Delhi, với các phương tiện giao thông chiếm khoảng 41% tổng lượng ô nhiễm ở thủ đô.

Mặc dù có nhiều nguyên nhân gây ô nhiễm và khói bụi cần được xác định chính xác, nhưng nguyên nhân chính cuối cùng được thảo luận là lĩnh vực công nghiệp, với một số lượng lớn các nhà máy và nhà máy sản xuất thuộc nhiều loại khác nhau rải rác khắp thành phố, chúng cũng sẽ đóng vai trò đáng kể. vai trò liên quan đến việc thải khói vào khí quyển. Trên thực tế, trong các giai đoạn khẩn cấp do ô nhiễm, người ta biết rằng các nhà máy sản xuất nhiên liệu và than đá đã được lệnh đóng cửa như một biện pháp quyết liệt nhằm ngăn chặn mức độ ô nhiễm gia tăng thảm khốc. Ước tính khoảng 18,6% tổng lượng ô nhiễm không khí ở Delhi đến từ lĩnh vực công nghiệp, với khoảng vài trăm đến hơn một nghìn tấn chất ô nhiễm thải vào khí quyển trong suốt cả năm chỉ từ các nhà máy. Không có gì ngạc nhiên khi chỉ số PM2.5 đã được ghi nhận tăng vọt lên tới 829,2 µg/m³, gấp hơn 80 lần mức khuyến nghị hàng năm của WHO (Tổ chức Y tế Thế giới).

## **Ảnh hưởng sức khỏe bởi ô nhiễm không khí ở Delhi**

Chỉ riêng năm 2017, ô nhiễm không khí ở Ấn Độ có liên quan đến cái chết của 1,24 triệu người, với 54% số ca tử vong do ô nhiễm không khí xung quanh và 46% số ca tử vong do ô nhiễm hộ gia đình như nhiên liệu rắn để nấu ăn. Người ta ước tính rằng số ca tử vong do ô nhiễm không khí chiếm 12,5% tổng số ca tử vong được ghi nhận trong năm đó.

Tập trung vào Delhi trong những năm gần đây và những ảnh hưởng sức khỏe do hít phải không khí ô nhiễm cao như vậy có thể mang lại, người ta có thể xem xét tác động của PM10 và PM2.5 đối với sức khỏe của một người, do cả những ảnh hưởng dài hạn và ngắn hạn. tài liệu tốt. Các hạt vật chất có đường kính từ 10 micromet trở xuống (PM10) có thể xâm nhập vào đường hô hấp của con người, gây kích ứng và nhiễm trùng xảy ra ở mắt, mũi và cổ họng, cũng như phổi nếu hạt vật chất được hít vào đủ sâu trong thời gian dài hơn của thời gian. Về bản chất, PM2.5 có tính chất nguy hiểm hơn nhiều, với kích thước đặc biệt nhỏ giúp nó có khả năng xâm nhập vào các mô sâu của phổi và sau đó đi sâu hơn vào máu. Điều này có thể gây ra nhiều vấn đề cho cả hệ hô hấp và tuần hoàn. Bệnh phổi tắc nghẽn mạn tính (COPD) có khả năng xảy ra rất cao khi hít thở không khí có chất lượng thấp như vậy, với các bệnh như khí thũng và viêm phế quản mãn tính thuộc thuật ngữ COPD, do nó được sử dụng để mô tả nhiều loại bệnh liên quan đến phổi và hô hấp.

Các biến chứng về tim và các bệnh tim khác nhau cũng có thể xảy ra khi tiếp xúc với không khí của Delhi, trong đó người trẻ và người già cũng như những người có hệ thống miễn dịch bị tổn thương sẽ có nguy cơ đặc biệt, mặc dù điều quan trọng cần lưu ý là khi nồng độ PM2.5 trong không khí rơi vào nhóm không lành mạnh (như trường hợp nhiều lần trong năm ở Delhi), rủi ro bắt đầu áp dụng cho tất cả mọi người, và ngay cả những người không có bất kỳ bệnh trạng nào từ trước hoặc có khuynh hướng mắc bệnh cũng có thể thấy mình mắc các bệnh mới nếu tiếp xúc nhiều. tiếp tục trong một khoảng thời gian đủ dài. Ung thư phổi cũng là một trường hợp khác.

## **Giải pháp giảm mức độ ô nhiễm và cải thiện chất lượng không khí**

Bước hợp lý đầu tiên để đi đúng hướng trong việc giải quyết cuộc khủng hoảng không khí ở Delhi là cung cấp nhiều dữ liệu về chất lượng không khí theo thời gian thực hơn cho công chúng và tăng cường tính minh bạch của dữ liệu về chất lượng không khí, cũng như các sáng kiến giáo dục người dân về hậu quả sâu rộng của việc hít thở không khí có chất lượng thấp. Khi biết mức độ nghiêm trọng của tình trạng ô nhiễm mà họ đang hít thở, mọi người có thể thực hiện nhiều biện pháp phòng ngừa để bảo vệ bản thân và những người xung quanh.

Chìa khóa để cải thiện ô nhiễm không khí là giảm lượng khí thải tổng thể từ tất cả các nguồn vi phạm, chẳng hạn như những nguồn phát sinh từ ngành giao thông cũng như ngành công nghiệp, một nhiệm vụ nói dễ hơn làm. Chính quyền Ấn Độ đã triển khai Chương trình Không khí Sạch Quốc gia (NCAP) nhằm mục đích cắt giảm ô nhiễm ở 102 thành phố ô nhiễm nhất từ 20-30% vào năm 2024. Theo NCAP, chính phủ có kế hoạch cắt giảm khí thải công nghiệp và giao thông, giảm ô nhiễm bụi và áp đặt các quy định chặt chẽ hơn về việc đốt sinh khối. Ngoài ra còn có kế hoạch nâng cấp và tăng cường hệ thống giám sát không khí.

Liên quan đến việc đốt sinh khối, các quy định nghiêm ngặt hơn có thể được áp dụng đối với những người chịu trách nhiệm đốt hàng loạt đất nông nghiệp và diện tích rừng để sử dụng cho nông nghiệp. Những quy tắc này không chỉ phải có hiệu lực mà còn cần phải được thực thi.

Các cá nhân có thể thực hiện các bước trong cuộc sống hàng ngày để giảm lượng khí thải cá nhân bằng cách ít phụ thuộc vào ô tô cá nhân hoặc sử dụng phương tiện giao thông công cộng, cũng như chuyển sang sử dụng nhiên liệu thay thế xanh hơn, với ý thức được giáo dục về những gì họ đang đốt cháy, cho dù đó là đống rác thải nhựa, chất hữu cơ chết hoặc thậm chí một số nguyên liệu thực phẩm nhất định, tất cả những bước nhỏ hơn này có thể giúp Delhi giảm mức độ ô nhiễm và dần bắt đầu tiến tới cải thiện chất lượng không khí