

TRƯỜNG ĐẠI HỌC KINH TẾ THÀNH PHỐ HỒ CHÍ MINH
BÀI LUẬN TRIẾT HỌC CUỐI KỲ
KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH



ĐỀ TÀI: XÂY DỰNG MÔ HÌNH HỌC MÁY PHÂN LOẠI CHẤT LƯỢNG NƯỚC

GVHD: HOÀNG ANH

NHÓM: 9

NGÀNH: KHOA HỌC DỮ LIỆU

KHÓA 48 LỚP DS001

LỚP CHIỀU THỨ 6 – GIẢNG ĐƯỜNG B2-507

LỚP HỌC PHẦN : 24D1INF50906401

DANH SÁCH THÀNH VIÊN NHÓM

Họ tên	Nhiệm vụ	Đánh giá
Trần Mỹ Thiên Tân	Tìm hiểu về kiến thức phân loại nước, các quy định chung. Tiền xử lý dữ liệu. Thiết kế slide.	100%
Nguyễn Trần Thế Anh	Tìm hiểu và áp dụng mô hình học máy để xử lý bài toán với Python, tiền xử lý dữ liệu.	100%
Phạm Bằng	Tìm hiểu và phân tích mô hình học máy trên lý thuyết, xử lý các dữ liệu ngoại lai.	100%
Lê Quyết	Xử lý các dữ liệu bị trống, nhiễu, thống kê mô tả dữ liệu. Thiết kế slide.	100%
Nguyễn Trương Hoàng	Tìm hiểu kiến thức, vấn đề liên quan về đề tài nghiên cứu (các thuộc tính phân loại nước, quy trình phân loại...).	100%
Nguyễn Thành Vinh	Phân chia công việc, tìm hiểu về kiến thức phân loại nước. Định hướng tiền xử lý dữ liệu và mô hình áp dụng.	100%

MỤC LỤC

1. Tổng quan đề tài.....	4
1.1. Lý do chọn đề tài.....	4
1.2. Bố cục đề tài và quy trình phân tích.....	5
2. Cơ sở lý thuyết.....	5
2.1. Mô tả các thuộc tính trong dữ liệu.....	5
2.2. Mô hình học máy.....	7
2.2.1. Lý thuyết về mô hình học máy.....	7
2.2.2. Tiêu chí giá tính hiệu quả của mô hình học máy.....	9
3. Tiến hành nghiên cứu.....	9
3.1. Mô tả dữ liệu.....	9
3.2. Tiền xử lý dữ liệu.....	11
3.2.1. Xử lý missing value.....	11
3.2.2. Xử lý dữ liệu ngoại lai (Outlier).....	12
3.3. Lựa chọn mô hình.....	13
3.4. Dự đoán.....	15
4. Kết luận.....	16

NỘI DUNG

1. Tổng quan đề tài

1.1. Lý do chọn đề tài

Xuyên suốt lịch sử nhân loại, nước là một trong những tài nguyên thiết yếu cho sự tồn tại của con người. Ngày nay, việc sẵn có và dễ dàng tiếp cận nguồn nước sạch và chất lượng là quyền cơ bản của con người và việc đảm bảo nước sạch và vệ sinh cho tất cả mọi người đã được đưa vào mục tiêu phát triển bền vững của Liên Hợp Quốc.

Mặc dù việc tiếp cận nguồn nước sạch được coi là quyền cơ bản của con người, nhưng ở nhiều nơi, nguồn nước có sẵn thường không an toàn cho người dùng và không đủ để đáp ứng các nhu cầu cơ bản về sức khỏe. Theo Tổ chức Y tế Thế giới (WHO) ước tính rằng khoảng 1,1 tỷ người trên toàn cầu đang uống nước không an toàn và hầu hết các bệnh tiêu chảy trên thế giới (88%) là do nước không an toàn, điều kiện vệ sinh kém và các thói quen không hợp vệ sinh.

Chất lượng nước do đô thị cung cấp phải được đo lường theo các tiêu chuẩn quốc gia về nước uống do các cơ quan có thẩm quyền và các tổ chức có liên quan khác xây dựng. Các tiêu chuẩn này coi một số thuộc tính là quan trọng hàng đầu đối với chất lượng nước uống, trong khi một số khác được coi là quan trọng thứ yếu. Thông thường để đánh giá các tiêu chuẩn này bao gồm việc lấy mẫu nước từ các địa điểm khác nhau vào các khoảng thời gian khác nhau và phân tích chúng trong phòng thí nghiệm. Theo tiêu chuẩn của Cục tiêu chuẩn Ấn Độ (2012), để đánh giá được nước là an toàn để uống hay không thì phải đáp ứng được hơn 40 tiêu chí. Việc phân tích tới hơn 40 tiêu chí trong phòng thí nghiệm và lấy mẫu thủ công đối với bất kỳ nguồn nước hoặc quy trình nào cũng có thể tốn kém, tốn thời gian và không hiệu quả trong một số tình huống.

Để khắc phục vấn đề này, một số nhà nghiên cứu đã đề xuất đánh giá chất lượng bằng cách sử dụng các cảm biến tự động, gắn thẻ địa lý và thu thập dữ liệu theo thời gian thực. Mô hình này được kì vọng sẽ đưa ra kết quả một cách kịp thời, tăng tính hiệu quả. Do đó, việc theo dõi chất lượng nước theo thời gian thực là cần thiết và các mô hình học máy được kì vọng là công cụ đắc lực trong việc này. Và để xây dựng được bộ công cụ này, các nhà nghiên cứu đã tạo nên một bộ dữ liệu phân loại chất lượng các mẫu nước trước đó với đúng quy trình chuẩn với 40 tiêu chí và sau đó tiến hành đúc kết gọn về 9 tiêu chí nhằm phục vụ cho quá trình xây dựng mô hình học máy.

Dựa vào những lý do trên, nhóm tác giả tiến hành nghiên cứu đề tài để giúp các doanh nghiệp có thêm các cơ sở khoa học trong việc ứng dụng các mô hình học máy vào để đánh giá chất lượng nước. Qua đó, giúp người dân có thể tiếp cận đúng với nguồn nước sạch, hạn chế sử dụng các nguồn nước bị ô nhiễm làm ảnh hưởng đến sức khỏe.

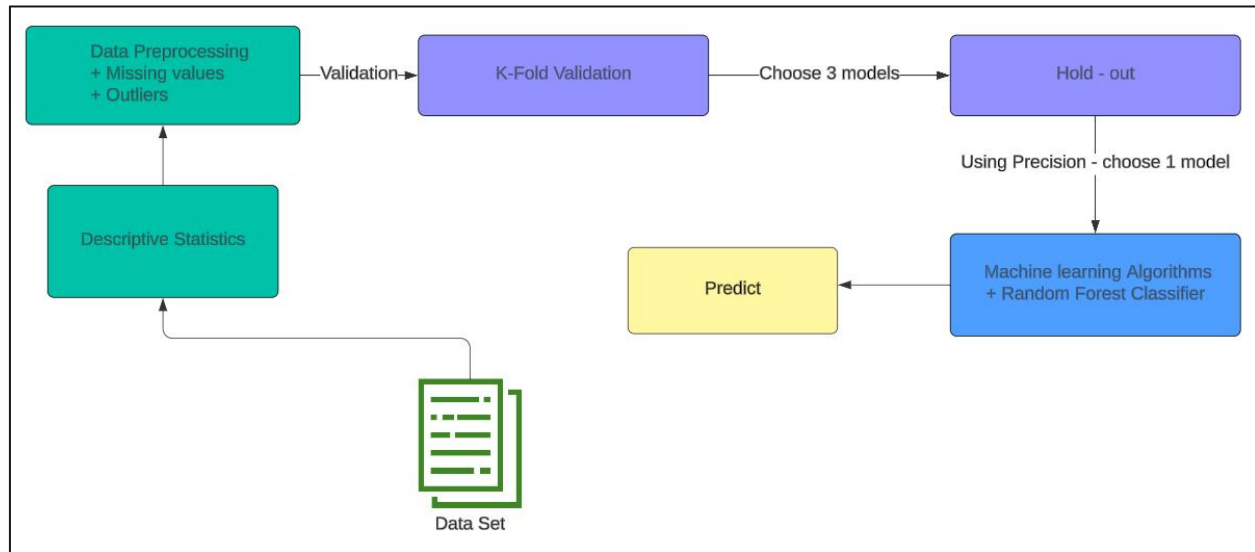
1.2. Bố cục đề tài và quy trình phân tích

Chương 1: Tổng quan đề tài.

Chương 2: Cơ sở lý thuyết.

Chương 3: Tiến hành nghiên cứu.

Chương 4: Kết quả



Hình 1.1: Quy trình phân tích

2. Cơ sở lý thuyết

2.1. Mô tả các thuộc tính trong dữ liệu

pH: pH là một thông số quan trọng trong việc đánh giá sự cân bằng của hàm lượng axit-bazơ có trong nguồn nước. Đồng thời, nồng độ pH cũng là thước đo cho tình trạng axit hay kiềm của nước. Tổ chức Y tế Thế giới (WHO) đã khuyến nghị rằng, giới hạn tối đa cho nồng độ pH trong nguồn nước phải từ 6,5 đến 8,5. Mặc dù pH thường không có tác động trực tiếp đến người tiêu dùng nước, nhưng nó là một trong những thông số chất lượng nước quan trọng nhất trong vận hành. pH của nước đi vào hệ thống phân phối phải được kiểm soát để giảm thiểu sự ăn mòn của các đường ống dẫn nước chính và đường ống trong hệ thống nước gia đình. Nếu không làm như vậy có thể dẫn đến ô nhiễm nước uống và các tác động xấu đến mùi vị, mùi của nước.

Hardness (Độ cứng): Độ cứng của nước chủ yếu được tạo ra bởi hợp chất canxi và magiê (và bởi nhiều kim loại khác) có trong nguồn nước, được hình thành từ các lớp đá mà nước đi qua. Hiện tại theo Tổ chức Y tế Thế giới (WHO) thì độ cứng của nước không có ảnh hưởng đáng kể đến sức khỏe của người sử dụng nhưng đây vẫn là một tiêu chí quan trọng để xác định hình thái của nước.

Solids (Total dissolved solids (TDS)/Tổng chất rắn): Là chỉ số chất rắn hòa tan trong nước; bao gồm các chất vô cơ và chất hữu cơ tồn tại phân tử, ion hóa hoặc vi hạt (Ví dụ: Các thành phần hoá học phổ biến nhất là Canxi, Photphat, Nitrat, Natri, Kali Clorua và các chất kim loại nặng bao gồm: Sắt, đồng, chì, asen, crom, lưu huỳnh....)

Nguồn nước có giá trị TDS cao chứng tỏ nước có độ khoáng hóa (mineralised) cao. Để được xem là “có thể uống được”, giới hạn mong muốn của chỉ số này là 500mg/L và giới hạn tối đa là 1000mg/L.

Chloramines: Clo và Cloramin là chất khử trùng chính được sử dụng trong hệ thống nước công cộng. Chloramin được hình thành khi amoniac được thêm vào Clo trong quá trình xử lý nước uống. Nguồn nước với nồng độ Clo khi đạt tới 4mg/L (hoặc 4ppm) thì sẽ được xem là an toàn để uống. Theo Sở y tế Virginia - Mỹ thì nếu nước chứa chloramine vượt quá mức an toàn (4mg/l) có thể gặp phải kích ứng ở mắt và mũi, làm trầm trọng thêm các vấn đề về hô hấp, gây hắt hơi, nghẹt xoang, ho, nghẹt thở, thở khò khè, khó thở và hen suyễn.

Sulfate: Hợp chất Sulfat thường xuất hiện tự nhiên trong khoáng chất, đất và đá. Chúng có mặt trong không khí, các mạch nước ngầm, trong thực vật hoặc thậm chí trong thực phẩm. Nồng độ Sulfat có trong nước biển là khoảng 2.700mg/L và sẽ giao động từ 3-30mg/L trong hầu hết các nguồn nước ngọt. Theo tổ chức Y tế thế giới (WHO) thì sulfate trong nước uống vượt ngưỡng cũng có thể tạo ra mùi vị đáng chú ý. Ngoài ra Sulfate cao cũng có thể góp phần vào sự ăn mòn của hệ thống đường ống nước.

Conductivity/Electrical conductivity (EC) (Khả năng dẫn điện): EC là chỉ số dùng để đo khả năng dẫn điện của nguồn nước, và chỉ số này có mối quan hệ liên quan trực tiếp tới chỉ số TDS - Total dissolved solids (Aydin 2007). Theo khuyến cáo của Tổ chức Y tế Thế giới (WHO), chỉ số EC không nên vượt quá 400 $\mu\text{S}/\text{cm}$.

Turbidity (Độ đục): Là thước đo độ trong suốt của nước mà mắt thường không thể nhìn thấy được, sử dụng phương pháp chiếu ánh sáng qua một mẫu nước nhất định để định lượng nồng độ hạt lơ lửng. Càng có nhiều hạt trong nguồn nước, độ đục càng cao. Theo BIS (2012), giới hạn mong muốn của độ đục trong nước uống là 5 NTU và giới hạn tối đa cho phép có thể lên tới 10 NTU.

Trihalomethanes (THMs): là loại sản phẩm phụ được hình thành khi clo được thêm vào nước có chất hữu cơ (ví dụ: thực vật và tảo đang phân hủy). Theo tiêu chuẩn của Liên minh Châu Âu (EU), tổng lượng THMs (Trihalomethanes) trong nước uống không được vượt quá 100 $\mu\text{g}/\text{L}$. Tổ chức Y tế Thế giới (WHO) đã thiết lập các giá trị hướng dẫn dựa trên sức khỏe cho chất lượng nước uống đối với THMs. Có bằng chứng khoa học cho thấy một số hóa chất THM riêng lẻ có thể gây ra cả các tác động ngắn hạn và dài hạn đến

sức khỏe, chẳng hạn như: chóng mặt; mệt mỏi; đau đầu; buồn ngủ; mất phối hợp; đau ngực nhẹ; tổn thương gan; tổn thương thận; tổn thương tinh hoàn; và loét da.

Organic Carbon: Theo Cơ quan Bảo vệ Môi trường Hoa Kỳ (EPA), có các loại vi khuẩn vi sinh đặc biệt như Cryptosporidium, có thể gây bệnh và rất kháng cự với các phương pháp khử trùng truyền thống.

2.2. Mô hình học máy

Machine learning (ML) là một phần quan trọng của trí tuệ nhân tạo (AI) cho phép hệ thống tự động học hỏi và cải thiện từ kinh nghiệm thông qua bộ dữ liệu trước đó. Các kỹ thuật được sử dụng trong ML dựa trên việc kiểm tra dữ liệu kỹ lưỡng để phát hiện xu hướng và tự cập nhật theo đó. Trong đề tài này, nhóm nghiên cứu quyết định sử dụng 3 phương pháp sau: Decision tree – Random Forest, Support vector machines (SVM), XGBClassifier để dự đoán chất lượng nước.

2.2.1. Lý thuyết về mô hình học máy

2.2.1.1. Decision tree – Random Forest

Cây quyết định (Decision tree) là một loại thuật toán học có giám sát, thường được sử dụng trong học máy để mô hình hóa và dự đoán kết quả dựa trên dữ liệu đầu vào. Đây là một cấu trúc dạng cây, trong đó mỗi nút tương ứng với một thuộc tính, mỗi nhánh tương ứng với giá trị thuộc tính và mỗi lá đại diện cho quyết định hoặc dự đoán cuối cùng. Được sử dụng để giải quyết bài toán phân loại.

Phương pháp information gain được dùng để xác định nút gốc:

- Giả sử p_i là xác suất mà một tuple lớp D thuộc về lớp C_i , được ước lượng bởi:

$$\left(\frac{|C_i, D|}{|D|} \right)$$

- Thông tin kỳ vọng (entropy) cần thiết để phân loại một bộ tuple trong D :

$$\text{Info}(D) = -\sum p_i \log_2(p_i)$$

- Thông tin cần thiết (sau khi sử dụng A để chia D thành v phần) để phân loại D :

$$\text{Info}_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \text{Info}(D_j)$$

- Thông tin cần thiết (sau khi sử dụng A để chia D thành v phần) để phân loại D :

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

- Chọn thuộc tính có lợi ích thông tin cao nhất

- Lặp lại để xác định các nút tiếp theo

Rừng ngẫu nhiên là một bộ phân loại chứa một số cây quyết định trên các tập con khác nhau của tập dữ liệu đã cho và lấy giá trị trung bình để cải thiện độ chính xác dự đoán của tập dữ liệu đó. Thay vì dựa vào một cây quyết định, rừng ngẫu nhiên lấy dự đoán từ mỗi cây và dựa trên đa số phiếu dự đoán, và nó dự đoán kết quả cuối cùng.

2.2.1.2. Support Vector Machine (SVM)

SVM là một thuật toán có giám sát, SVM nhận dữ liệu vào, xem chúng như những các vector trong không gian và phân loại chúng vào các lớp khác nhau bằng cách xây dựng một siêu phẳng trong không gian nhiều chiều làm mặt phân cách các lớp dữ liệu. Để tối ưu kết quả phân lớp thì phải xác định siêu phẳng (hyperplane) có khoảng cách đến các điểm dữ liệu (margin) của tất cả các lớp xa nhất có thể. Thuật toán SVM được dùng để giải quyết bài toán phân loại với cỡ mẫu nhỏ hoặc vừa.

Thuật toán SVM tổng quát:

- Số chiều của không gian bài toán (còn gọi là không gian đặc trưng) tương ứng với số lượng thuộc tính (đặc trưng) của một đối tượng dữ liệu.
- Phương trình biểu diễn siêu phẳng cần tìm (hyperplane) trong không gian đa chiều là: $w^T x + b = 0$ và giá trị:

$$\text{margin} = \frac{2|w^T x + b|}{\|w\|} = \frac{2}{\|w\|}$$

- Mục tiêu của SVM là cần tìm giá trị margin cực đại đồng nghĩa với việc $\|w\|$ đạt cực tiểu với điều kiện:

$$y_n(w^T x_n + b) \geq 1 \quad \forall n = 1, 2, \dots, N$$

- Hàm mục tiêu cần tối ưu là một norm nên là một hàm lồi \Rightarrow bài toán quy hoạch toàn phương (Quadratic Programming).

2.2.1.3. XGB Classifier

XGBoost, hay Extreme Gradient Boosting, là một thuật toán học máy tiên tiến nổi tiếng với hiệu suất dự đoán xuất sắc. Nó được coi là tiêu chuẩn vàng trong học tập tập hợp, đặc biệt khi nói đến các thuật toán tăng cường độ dốc. Thuật toán này phát triển một loạt các mô hình học yếu lần lượt để tạo ra một mô hình dự đoán đáng tin cậy và chính xác.

Cơ bản, XGBoost xây dựng một mô hình dự đoán mạnh bằng cách tổng hợp các dự đoán của nhiều mô hình học yếu, thường là cây quyết định. Nó sử dụng kỹ thuật tăng cường để tạo ra một mô hình tập hợp cực kỳ chính xác bằng cách để mỗi mô hình học yếu sau này sửa chữa những sai lầm của những mô hình trước đó.

Phương pháp tối ưu hóa (gradient) giảm thiểu hàm chi phí bằng cách liên tục thay đổi các tham số của mô hình dựa trên các gradient của lỗi. Thuật toán này cũng giới thiệu ý tưởng "tăng cường độ dốc với cây quyết định", trong đó hàm mục tiêu được giảm thiểu bằng cách tính toán tầm quan trọng của mỗi cây quyết định được thêm vào tập hợp theo thứ tự. Bằng cách thêm một thuật ngữ điều chỉnh và sử dụng một thuật toán tối ưu hóa tiên tiến hơn, XGBoost tiến thêm một bước và cải thiện độ chính xác và hiệu quả.

2.2.2. Tiêu chí giá tính hiệu quả của mô hình học máy

Những mô hình học máy nhóm sử dụng sẽ dựa vào ma trận nhầm lẫn để đưa ra những chỉ số đánh giá từ đó xác định một mô hình có độ chính xác cao nhất và kết quả đáng tin cậy nhất.

- Precision: được định nghĩa là tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm mô hình dự đoán là Positive.
- Recall được định nghĩa là tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm thật sự là Positive (hay tổng số điểm được gán nhãn là Positive ban đầu).
- Accuracy: Là tỷ lệ số mẫu được phân lớp đúng trong toàn bộ tập dữ liệu.
- F1-score: Là tỷ lệ số mẫu được phân lớp đúng trong toàn bộ tập dữ liệu.
- ROC: Là một đồ thị được sử dụng khá phổ biến trong đánh giá các mô hình phân loại nhị phân. Đường cong này được tạo ra bằng cách biểu diễn tỷ lệ dự báo true positive rate (TPR) dựa trên tỷ lệ dự báo false positive rate (FPR). ROC càng tiệm cận với điểm (0;1) trong đồ thị thì mô hình càng hiệu quả.
- AUC: Là diện tích nằm dưới đường cong ROC. Giá trị này càng lớn thì mô hình càng tốt.

3. Tiến hành nghiên cứu

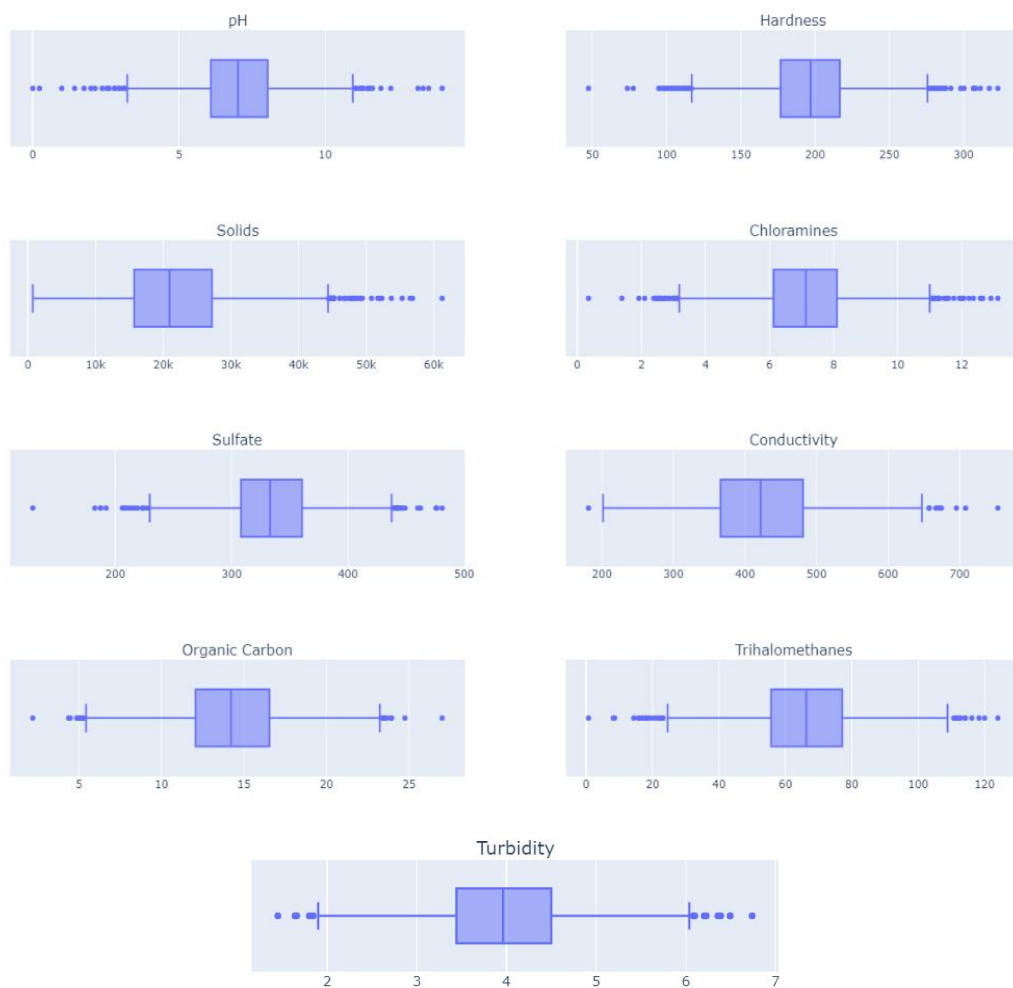
3.1. Mô tả dữ liệu

Dữ liệu về chất lượng nước uống này đã được thu thập từ các bang khác nhau ở Ấn Độ. Tổng cộng 2620 mẫu đã được thu thập và phân tích trong phòng thí nghiệm theo bộ tiêu chí của Cục tiêu chuẩn Ấn Độ 2012 để đánh giá khả năng uống được của nước. Sau khi phân tích đánh giá thì chuyên gia đã gán nhãn về các mẫu nước uống được và không được. Tiếp theo, nhóm chuyên gia đã tổng kết từ bộ tiêu chí của Cục tiêu chuẩn Ấn Độ 2012 về thành 9 tiêu chí để tiến hành xây dựng mô hình học máy. Những tiêu chí được sử dụng là pH, Hardness, Solids, Choloramines, Sulfate, Conductivity, Organic Carbon,

Trihalomethanes, Turbidity. Tuy nhiên, trong quá trình nhập liệu và lưu trữ thì đã xảy ra một số vấn đề khiến bộ dữ liệu xuất hiện những giá trị bị khuyết (missing value).

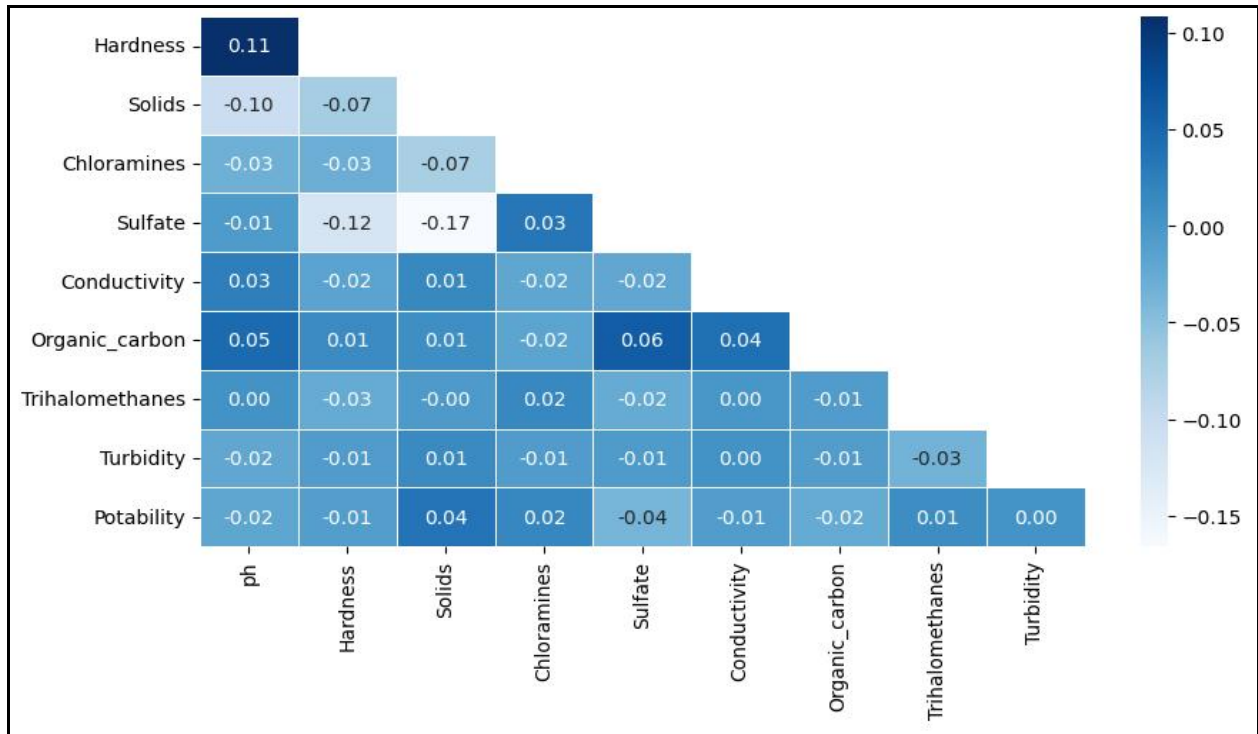
	count	mean	std	min	25%	50%	75%	max
ph	2234.000000	7.071024	1.588089	0.000000	6.089577	7.022285	8.035837	14.000000
Hardness	2620.000000	196.200911	33.137494	47.432000	176.734160	196.926487	216.657847	323.124000
Solids	2620.000000	21973.377381	8702.501618	728.750830	15715.740730	20926.882155	27216.632550	61227.196008
Chloramines	2620.000000	7.130793	1.584109	0.352000	6.125502	7.131972	8.105680	13.127000
Sulfate	1999.000000	334.040923	41.573027	129.000000	307.992545	333.073546	360.601377	481.030642
Conductivity	2620.000000	425.749573	81.038498	181.483754	365.842780	421.884968	481.089353	753.342620
Organic_carbon	2620.000000	14.292469	3.272972	2.200000	12.067417	14.220645	16.541731	27.006707
Trihalomethanes	2495.000000	66.194111	16.186110	0.738000	55.711220	66.299162	77.127254	124.000000
Turbidity	2620.000000	3.965649	0.783198	1.450000	3.441806	3.962234	4.497554	6.739000
Potability	2620.000000	0.394275	0.488788	0.000000	0.000000	0.000000	1.000000	1.000000

Hình 3. 1: Thống kê mô tả



Hình 3. 2: Box plot của các biến giải thích

Thông qua các biểu đồ hộp và thống kê mô tả của các thuộc tính, ta thấy có một vài điểm dữ liệu có giá trị nằm ngoài đường kẻ biên ngoài (tức nằm ngoài khoảng 25% và 75% dữ liệu trung tâm).



Hình 3. 3: Ma trận tương quan

3.2. Tiền xử lý dữ liệu

3.2.1. Xử lý missing value

Missing values hay còn gọi là những giá trị bị thiếu, không được điền hoặc không được cập nhật vào bộ dữ liệu, đây có thể là kết quả của một quá trình sai sót trong quá trình nhập liệu. Việc làm sạch dữ liệu đầu vào bao giờ cũng đi kèm với công việc xử lý missing values.

Thông thường, có 2 cách để xử lý missing values:

Cách 1: Loại bỏ missing values (trong trường hợp missing values đó không quan trọng đối với dữ liệu của chúng ta hoặc số lượng missing values quá ít - chỉ chiếm khoảng dưới 3% tổng số quan sát trong 1 biến nhất định).

Cách 2: Thay thế missing values bằng một giá trị khác. Việc thay thế bằng giá trị nào sẽ phụ thuộc vào việc bản chất của missing values trong những trường hợp đó là gì.



Hình 3. 4: Số giá trị bị khuyết

Missing value chỉ xuất hiện ở 3 biến quan sát đó là pH (14.7%), Sulfate (23.7%) và Trihalomethanes (4.8%). Trong trường hợp này, việc xóa đi các dữ liệu bị thiếu là không hợp lý. Do đó, nhóm tiến hành điền dữ liệu bị thiếu bằng một giá trị thay thế. Giá trị thay thế được nhóm chọn đó là

trung vị (median). Lý do được nhóm đề ra là vì trong dữ liệu có những điểm dữ liệu quá lớn hoặc quá bé so với các điểm dữ liệu còn lại, việc thay missing value bằng giá trị trung bình sẽ làm dữ liệu mất tính chính xác.

3.2.2. Xử lý dữ liệu ngoại lai (Outlier)

Như đã nhắc đến ở trên, trong bộ dữ liệu tồn tại các điểm dữ liệu quá lớn hoặc quá bé so với phần lớn các điểm dữ liệu còn lại, nhóm xác định chúng là điểm dữ liệu ngoại lai. Vậy, chính xác, dữ liệu ngoại lai là gì hay như thế nào là dữ liệu bình thường, biên kiểm định là bao nhiêu và tác hại của chúng như thế nào?

Dữ liệu ngoại lai là những điểm dữ liệu khác biệt đáng kể so với các quan sát khác trong tập dữ liệu. Những điểm này có thể gây ảnh hưởng đến kết quả phân tích như trung bình, phương sai, hồi quy và làm giảm độ chính xác của mô hình phân tích, dự đoán.

Trong bài nghiên cứu, nhóm xác định dữ liệu ngoại lệ thông qua phương pháp Interquartile range (IQR). Phương pháp IQR dùng các giá trị phần tư (Q1, Q3) để xác

định giới hạn thấp nhất để một giá trị được coi là dữ liệu ngoại lai. Trong đó, các giá trị nằm trong phạm vi Q1 và Q3 chắc chắn không được coi là giá trị ngoại lai.

Để tính các giới hạn ngoại lai, chúng ta thực hiện các bước sau:

- Tính giá trị IQR.

$$\text{IQR} = Q3 - Q1$$

- Tính giới hạn sàn và giới hạn trần:

$$\text{lower} = Q1 - 1.5 * \text{IQR}$$

$$\text{upper} = Q3 + 1.5 * \text{IQR}$$

Như vậy, các giá trị nằm trong đoạn [lower, upper] được coi là dữ liệu bình thường, các điểm dữ liệu còn lại được coi là ngoại lai.

3.3. Lựa chọn mô hình

Để lựa chọn mô hình học máy phù hợp, nhóm sử dụng phương pháp K-Fold Cross Validation nhằm đánh giá các mô hình Random Forest Classifier, XGB Classifier, Decision Tree Classifier, Logistic Regression, K-Neighbors Classifier và SVC.

Với K-Fold Cross Validation, tập dữ liệu training sẽ được chia thành K phần, thường K được chọn là 5 hoặc 10, ở đây, nhóm chọn K = 5.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Hình 3. 5: Chia dữ liệu theo phương pháp K-Fold

Sau khi chia bộ dữ liệu, tiến hành train mô hình K lần, mỗi lần chọn một phần làm dữ liệu validation và K-1 phần còn lại làm dữ liệu training. Kết quả đánh giá mô hình cuối

cùng sẽ là trung bình cộng kết quả đánh giá của K lần train. Như vậy, kết quả sẽ trở nên khách quan và chính xác hơn việc chọn phương pháp tách bộ dữ liệu thành 2 phần với tỷ lệ 8:2 hoặc 7:3.

Sau khi có kết quả đánh giá các mô hình, ta chọn 3 mô hình có kết quả (ACC trung bình) cao nhất và thực hiện một trong 2 cách sau để tạo ra mô hình dự đoán:

- Ta lưu lại mô hình tốt nhất trong quá trình K-Fold và sử dụng chúng để dự đoán cho bộ dữ liệu test. Tuy nhiên, cách này có nhược điểm, đó là, mô hình được chọn không được hình thành nên từ toàn bộ dữ liệu.
- Train các mô hình đã chọn một lần nữa và tiến hành dự đoán.

	Mean	Standard Deviation
model_name		
RandomForestClassifier	0.781298	0.019396
XGBClassifier	0.762214	0.009212
DecisionTreeClassifier	0.700382	0.029932
SVC	0.605725	0.001045
KNeighborsClassifier	0.553053	0.032482

Hình 3. 6: Kết quả của các mô hình K-Fold

Như đã trình bày ở trên, nhóm quyết định chọn 3 mô hình có điểm cao nhất để tiến hành dự đoán, đó là: Random Forest Classifier, XGB Classifier và Decision Tree Classifier.

Đối với việc phân loại nước uống được và không uống được, tham số đặc biệt quan trọng để đánh giá mô hình tốt hay không chính là độ chính xác (Precision) và Recall (0).

	Precision	Recall
RandomForestClassifier	0.75	0.88
DecisionTreeClassifier	0.73	0.88
XGBClassifier	0.66	0.81

Hình 3. 7: Kết của 3 mô hình (K-Fold)

Precision cho biết trong số những mẫu được dự đoán uống được (Potability = 1), có tỷ lệ bao nhiêu mẫu dự đoán đúng. Việc lấy tham số này làm quy chuẩn chọn mô hình bởi vì tính nghiêm trọng của việc dự đoán mẫu nước vào nhóm uống được nhưng thực tế thì không sẽ rất nguy hiểm cho người dùng.

- **Precision** = $TP / (TP + FP)$

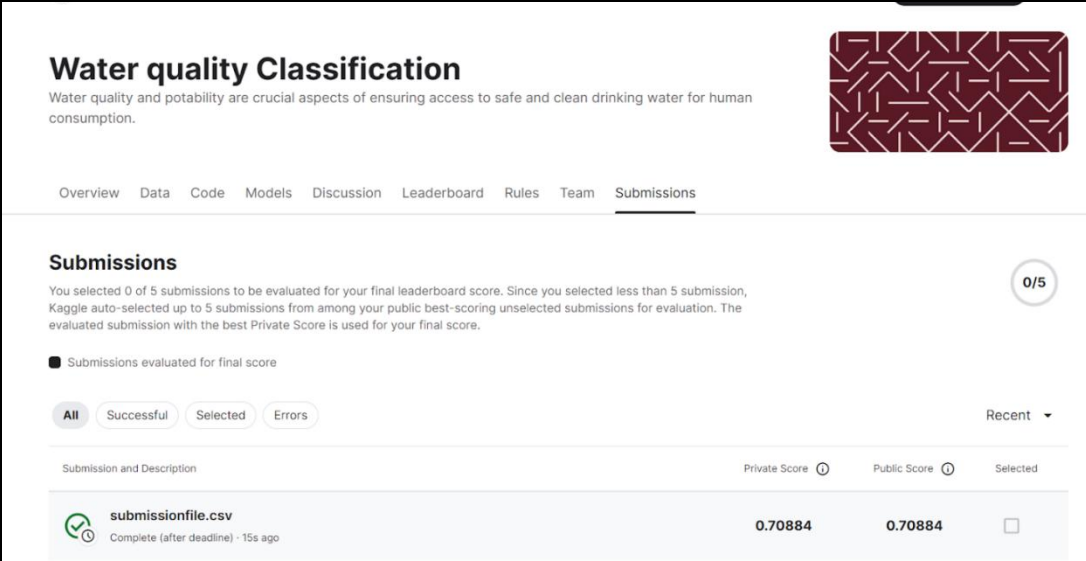
Recall (đôi khi còn được gọi là Sensitivity): trong những mẫu nước thực sự không uống được, bao nhiêu trong số đó được dự đoán đúng bởi mô hình. Nói cách khác, có bao nhiêu dự đoán “negative” đúng là do mô hình đưa ra.

- **Recall(0)** = $TN / (TN + FP)$

Với **Precision = 75%** và **Recall (0) = 78%** ở mô hình Random Forest Classifier, đây là kết quả cao nhất. Vậy nhóm chọn mô hình này để dự đoán.

3.4. Dự đoán

Nhóm tiến hành dự đoán trên file test bằng mô hình Random Forest Classifier và submit trên Kaggle là đạt được kết quả 0.70884.



Water quality Classification

Water quality and potability are crucial aspects of ensuring access to safe and clean drinking water for human consumption.


Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

You selected 0 of 5 submissions to be evaluated for your final leaderboard score. Since you selected less than 5 submission, Kaggle auto-selected up to 5 submissions from among your public best-scoring unselected submissions for evaluation. The evaluated submission with the best Private Score is used for your final score.

Submissions evaluated for final score

All Successful Selected Errors Recent

Submission and Description	Private Score	Public Score	Selected
 submissionfile.csv Complete (after deadline) - 15s ago	0.70884	0.70884	<input type="checkbox"/>

So sánh với các bài nộp trước:

Water quality Classification						
<div>Overview Data Code Models Discussion Leaderboard Rules Team Submissions</div> <div>Submitted by Nguyen Tran The Anh · Submitted 2 minutes ago</div> <div>Jump to your leaderboard position</div> <div>Search leaderboard</div> <div>The private leaderboard is calculated over the same rows as the public leaderboard in this competition. This competition has completed. This leaderboard reflects the final standings.</div>						
#	Team	Members	Score	Entries	Last	Solution
1	maneesh_23		0.71493	2	17d	
2	ROSHANN S		0.71189	8	17d	
3	Tharukesh SD		0.70884	4	17d	

4. Kết luận

Nghiên cứu này tập trung vào việc phân loại chất lượng nước bằng cách sử dụng các kỹ thuật máy học và đề xuất một phương pháp giám sát chất lượng nước theo thời gian thực thông minh. Bộ dữ liệu bao gồm hơn 2620 mẫu được thu thập từ các địa điểm khác nhau ở Ấn Độ. Bộ dữ liệu bao gồm các thuộc tính: pH, Hardness, Solids, Choloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity. Các bộ phân loại máy học được sử dụng trong nghiên cứu là Decision tree – Random Forest, Support vector machines (SVM), XGBClassifier. Các chỉ số hiệu suất để đánh giá về tính hiệu quả của mô hình bao gồm độ chính xác (Precision) và Recall (0). Precision được lựa chọn làm tiêu chí chọn mô hình bởi vì tính nghiêm trọng của việc dự đoán mẫu nước vào nhóm uống được nhưng thực tế thì không sẽ rất nguy hiểm cho người dùng. Còn Recall (đôi khi còn được gọi là Sensitivity) để đánh giá trong các mẫu nước thực sự không uống được thì bao nhiêu được dự đoán đúng bởi mô hình.

Về độ chính xác, Với Precision = 75% và Recall (0) = 78% ở mô hình Random Forest Classifier, đây là kết quả cao nhất. Vì vậy mô hình Random Forest Classifier này được sử dụng để dự đoán. Điều này mở ra cơ hội cho sự hợp tác liên ngành giữa các nhà nghiên cứu máy học và các chuyên gia trong các lĩnh vực phân tích chất lượng nước. Phân tích này cũng có thể được mở rộng theo nhiều hướng, bao gồm triển khai hệ thống giám sát nước theo thời gian thực, ứng dụng Internet vạn vật (IoT), nghiên cứu y sinh, và các lĩnh vực khác.

TÀI LIỆU THAM KHẢO

- Abuzir, S. Y., & Abuzir, Y. S. (2022). Machine learning for water quality classification. *Water Quality Research Journal*, 57(3), 152-164.
- Alomani, S. M., & Alhawiti, N. I. (2022). Prediction of Quality of Water According to a Random Forest Classifier. *International Journal of Advanced Computer Science and Applications*, 13(6).
- Chidiac, S., El Najjar, P., Ouaini, N., El Rayess, Y., & El Azzi, D. (2023). A comprehensive review of water quality indices (WQIs): History, models, attempts and perspectives. *Reviews in Environmental Science and Bio/Technology*, 22(2), 349-395.
- Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal*, 53(1), 3-13.
- Hirani, P., & Dimble, V. (2019). Water pollution is killing millions of Indians. Here's how technology and reliable data can change that. In *World Economic Forum*.
- IS, B. (2012). 10500: 2012 Drinking Water—Specification.
- Khademikia, S., Rafiee, Z., Amin, M. M., Poursafa, P., Mansourian, M., & Modaberi, A. (2013). Association of nitrate, nitrite, and total organic carbon (TOC) in drinking water and gastrointestinal disease. *Journal of Environmental and public Health*, 2013.
- Khurana, I., Sen, R., & India, W. (2008). Drinking water quality in rural India: Issues and approaches.
- Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, 48, 102920.
- Shams, M. Y., Elshewey, A. M., El-kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2023). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, 1-28.
- Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., ... & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, 1(2), 107-116.