**Tổng quan về đề tài**

### Giới thiệu bài toán

Tiếp cận nước sạch và an toàn là nhu cầu cấp thiết toàn cầu, ảnh hưởng trực tiếp đến sức khỏe, cuộc sống của con người. Đây là vấn đề được quan tâm hàng đầu ở các cấp quốc gia, khu vực và địa phương.

Ngày nay, việc bảo vệ nguồn nước, đa dạng sinh học và sử dụng hợp lý nguồn tài nguyên nước đã và đang trở nên đáng báo động, đặc biệt sự ô nhiễm các nguồn nước (nhất là nguồn nước ngọt) đang trở nên ngày càng trầm trọng, đe dọa cuộc sống của loài người và gây ra nhiều khó khăn cho sản xuất, cuộc sống.

Dựa vào thông số để đánh giá chất lượng nước là điều cần thiết để đưa ra quyết định sáng suốt trong quản lý tài nguyên nước, sức khỏe cộng đồng và bảo vệ môi trường. Các phương pháp đánh giá truyền thống thường tốn nhiều công sức và thiếu khả năng theo thời gian thực, điều này làm nổi bật sự cần thiết của các mô hình dự đoán mạnh mẽ. Chính vì vậy cần có hệ thống kiểm soát, giám định nguồn nước để giảm các tác động tiêu cực đến sức khỏe và chi phí chăm sóc sức khỏe lớn hơn chi phí thực hiện các biện pháp can thiệp.

Nghiên cứu này giải quyết nhu cầu bằng cách phát triển và đánh giá mô hình dự đoán chất lượng nước.

### Khái niệm và ý nghĩa Khai phá dữ liệu và Học máy

2.1 Khai phá dữ liệu (Data Mining)

* Khai phá dữ liệu (Data Mining) là một lĩnh vực nhắm tự động khai thác những thông tin tri thức đang tiềm ẩn trong dữ liệu
* Khai phá dữ liệu là lĩnh vực phát triển bền vững mang lại nhiều lợi ích, triển vọng ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống
* Các kỹ thuật được áp dụng dựa trên CSDL, học máy, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê và tính toán nâng cao

Quy trình khai phá dữ liệu bao gồm: xác định bài toán, thu thập và tiền xử lý dữ liệu, khai phá dữ liệu, thể hiện tri thức đã được phát hiện, sử dụng tri thức phát hiện được.

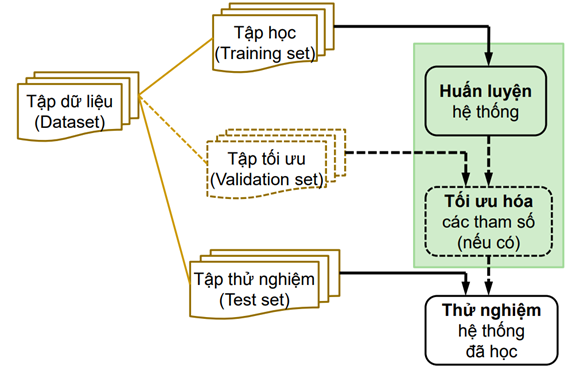
2.2 Học máy

Máy học là chương trình máy tính được cho là học hỏi từ trải nghiệm E đối với một số loại nhiệm vụ T và phép đo hiệu suất P, nếu hiệu suất của nó ở các nhiệm vụ trong T, được đo trong P, được cải thiện với trải nghiệm E.

Máy học từ một tập các ví dụ huấn luyện(training set, tập học){{x1,x2,...,xn} ; {y1,y2,...,ym}}

* x là quan sát(mẫu điểm dữ liệu) của x trong quá khứ
* y là sự quan sát của y trong qua khứ, thường được gọi là nhãn hoặc phản hồi hoặc đầu ra.

Chu trình học máy:

****

Sau khi máy học:

* Xây dựng được mô hình
* Sử dụng mô hình để đưa ra dự đoán hoặc suy luận cho các quan sát trong tương lai.

3.Học máy có giám sát - Supervised Learning:

Học có giám sát là trường hợp dữ liệu huấn luyện được cho một cách tường minh dưới dạng đầu vào và đầu ra của hàm đích, ví dụ, cho trước tập các mẫu cùng nhãn phân loại tương ứng.

Hàm y = f(x) từ tập huấn luyện cho trước {x1, x2,…, xn; y1, y2,…, yn} sao cho yi ≅ f(xi) với mọi i

Học máy khi đó được gọi là có giám sát để thể hiện việc thuật toán nhận được chỉ dẫn trực tiếp về lời giải cho từng trường hợp.

Học máy có giám sát bao gồm phân loại và hồi quy. Phân loại là dạng học có giám sát với hàm đích nhận giá trị **rời rạc** và hồi quy nhận giá trị **liên tục.**

* Phân loại(phân loại, phân lớp): nếu y thuộc một tập rời rạc
* Hồi quy: nếu y là số thực

1. **Đánh giá đề tài liên quan**

* **Nghiên cứu dự báo chất lượng nước dưới đất tại khu vực thành phố Hội An, Quảng Nam, Việt Nam [2]**

Trong nghiên cứu này, các mô hình học máy như hồi quy tuyến tính (LR), rừng ngẫu nhiên (RF), máy hỗ trợ vector (SVM), K- điểm dữ liệu gần nhất (KNN), và mạng lập thể (Cubist) đã được sử dụng để phân tích dữ liệu chất lượng nước dưới đất trong mùa mưa và mùa khô. Các mô hình này đã được tối ưu hóa bằng ngôn ngữ R và kết quả cho thấy mô hình Cubist ở tỷ lệ huấn luyện:kiểm tra 70:30 là tối ưu nhất cho bộ dữ liệu với độ tin cậy R2 lần lượt là 98,8% và 96%.

Nghiên cứu này chỉ ra tiềm năng của học máy trong việc cải thiện khả năng dự báo và phân loại chất lượng nước, đóng góp vào việc quản lý nguồn nước một cách hiệu quả hơn.

* **Phân loại chất lượng nước dựa trên các thuật toán học máy cổ điển [3]**

Các phương pháp học máy cổ điển như cây quyết định (Decision trees), máy vector hỗ trợ (SVM), và Naive Bayes đã được sử dụng để phân loại chất lượng nước. Các phương pháp này thường dựa trên việc xây dựng các mô hình phân loại từ tập dữ liệu đã được gán nhãn. Tuy nhiên, cần lưu ý rằng hiệu suất của các thuật toán này phụ thuộc nhiều vào việc chọn và trích xuất đặc trưng phù hợp từ dữ liệu chất lượng nước. Ứng dụng các mô hình khai phá và phân tích dữ liệu trong việc phân loại mức độ ô nhiễm nước:

Nghiên cứu này khám phá việc sử dụng các mô hình khai phá dữ liệu để phân loại và đánh giá mức độ ô nhiễm của nguồn nước, góp phần vào việc quản lý chất lượng nước

* Tổng quan, các đề tài liên quan đến phân loại chất lượng nước trong khai phá dữ liệu và học máy đang nhận được sự quan tâm và nghiên cứu rộng rãi trong cộng đồng khoa học. Các phương pháp học máy và khai phá dữ liệu có tiềm năng cao để cung cấp thông tin hữu ích và hỗ trợ quyết định trong việc quản lý và bảo vệ chất lượng nước. Tuy nhiên, việc thành công của các đề tài này đòi hỏi sự kết hợp giữa kiến thức về lĩnh vực chất lượng nước và các phương pháp phân loại và khai phá dữ liệu. Nghiên cứu trong lĩnh vực này có tiềm năng mở ra những khía cạnh mới và tiến bộ trong việc giải quyết các vấn đề liên quan đến chất lượng nước và môi trường.

### Mục đích đề tài

Mục đích chính của đề tài "Khai phá dữ liệu về nguồn nước để đánh giá mức độ an toàn" là áp dụng các phương pháp và kỹ thuật học máy, khai phá dữ liệu để xây dựng mô hình phân loại chất lượng nước. Đây là một lĩnh vực nghiên cứu quan trọng vì chất lượng nước ảnh hưởng trực tiếp đến sức khỏe con người và môi trường.

* Dự đoán chất lượng nước dựa trên các chỉ số và phân loại nguồn nước dựa trên các tiêu chuẩn quốc gia quốc tế.
* Hỗ trợ trong việc quản lý và bảo vệ nguồn nước thông qua việc theo dõi liên tục và đánh giá chất lượng nước.
* Nâng cao hiệu quả, tăng cường hiệu quả giám sát.
* Cung cấp các thông tin chính xác và kịp thời cho các nhà quản lý và chính sách để đưa ra quyết định dựa trên dữ liệu phán đoán.

Tóm lại, mục đích của đề tài là tận dụng tiềm năng của học máy và khai phá dữ liệu để xây dựng mô hình phân loại chất lượng nước, cung cấp thông tin hữu ích và hỗ trợ quyết định trong việc quản lý và bảo vệ chất lượng nước trong các ứng dụng thực tế.

### Mục tiêu đề tài

Mục tiêu cụ thể của đề tài bao gồm:

* Sử dụng các phương pháp học máy và khai phá dữ liệu để đánh giá chất lượng an toàn cho các mẫu nước, ví dụ: nước sạch hay bẩn.
* Sử dụng các kỹ thuật học máy và khai phá dữ liệu để tối ưu hóa quy trình phân loại chất lượng nước. Điều này bao gồm việc tìm ra các đặc trưng quan trọng, lựa chọn thuật toán phù hợp và tinh chỉnh các tham số để đạt được hiệu suất tốt nhất trong việc phân loại chất lượng nước.
* Áp dụng các mô hình phân loại chất lượng nước đã xây dựng vào các ứng dụng thực tế như đánh giá chất lượng nước trong các hệ thống cấp nước.

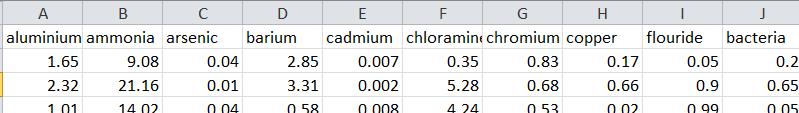
## 

## **II. Tiền xử lý dữ liệu**

### **Phân tích về dữ liệu**

Mục tiêu là dự đoán chính xác chất lượng nước dựa trên bộ dữ liệu toàn diện gồm 21 đặc điểm , bao gồm các hợp chất hóa học, chất gây ô nhiễm, chỉ số vi sinh vật và tính chất vật lý, rất quan trọng để đánh giá chất lượng nước trong các bối cảnh.

* Nguồn dữ liệu: Kaggle (<https://www.kaggle.com/datasets/mssmartypants/water-quality>)
* Thu thập dữ liệu: Dữ liệu bao gồm thông tin về các chỉ số, thông số đo đạc từ nguồn nước.
* Đề tài bao gồm các 20 đặc trưng như sau: aluminum, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, uranium.
* Phân tích đặc trưng:
  + Aluminium: Thường được đo để kiểm tra sự ô nhiễm trong nước từ các nguồn như xỉ nhôm và các hoạt động công nghiệp khác.
  + Ammonia: Mức độ ammonia thường được kiểm tra để đánh giá sự ô nhiễm trong nước từ chất thải động vật hoặc công nghiệp.
  + Arsenic: Một chất độc hại, arsenic thường được đo để đảm bảo nước uống an toàn.
  + Barium: Thường được kiểm tra để đánh giá sự ô nhiễm từ các nguồn như cát thạch anh.
  + Cadmium: Chất này thường xuất hiện trong nước do quá trình sản xuất kim loại và các hoạt động công nghiệp khác.
  + Chloramine: Chloramine thường được thêm vào nước để khử trùng. Việc đo lượng chloramine còn lại giúp đảm bảo an toàn cho người tiêu dùng.
  + Chromium: Được đo để kiểm tra sự ô nhiễm trong nước từ các nguồn như xỉ từ các hoạt động công nghiệp.
  + Copper: Thường được đo để đảm bảo rằng nước không bị ô nhiễm từ các đường ống dẫn nước bằng đồng.
  + Fluoride: Mức độ fluoride thường được kiểm tra để đảm bảo rằng nước uống cung cấp đủ lượng fluoride để bảo vệ răng.
  + Bacteria (vi sinh vật): Vi sinh vật như vi khuẩn và vi khuẩn có thể gây ra nhiều vấn đề sức khỏe nếu xuất hiện trong nước uống.
  + Viruses (vi sinh vật): Như vi khuẩn, virus cũng là một yếu tố quan trọng cần được kiểm tra trong nước để đảm bảo sức khỏe công cộng.
  + Lead: Chất độc hại, lead thường được kiểm tra để đảm bảo rằng nước uống không bị ô nhiễm từ các đường ống chứa chì.
  + Nitrates: Nitrates thường xuất hiện trong nước do nhiều nguồn khác nhau như phân bón và chất thải động vật.
  + Nitrites: Cũng tương tự như nitrates, nitrites cũng là một loại độc tố thường xuất hiện trong nước.
  + Mercury: Một chất độc hại khác, mercury thường được kiểm tra để đảm bảo an toàn cho người tiêu dùng.
  + Perchlorate: Thường xuất hiện trong nước do các loại pháo hoa và hóa chất công nghiệp.
  + Radium: Radium là một chất phóng xạ tự nhiên, thường xuất hiện trong nước dưới dạng cặn.
  + Selenium: Selenium thường được kiểm tra để đảm bảo rằng nước không bị ô nhiễm từ các hoạt động khai thác mỏ và công nghiệp.
  + Silver: Được đo để kiểm tra sự ô nhiễm từ các nguồn như xỉ từ các quá trình sản xuất và công nghiệp.
  + Uranium: Uranium thường xuất hiện trong nước do sự giải phóng từ đất đá và các hoạt động khai thác khoáng sản.

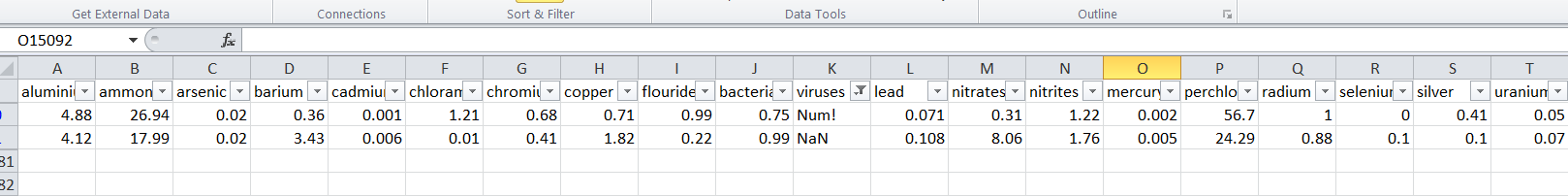


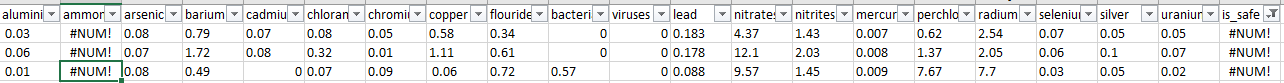


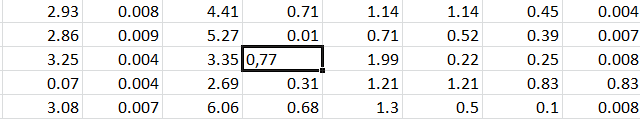
=> Sau khi thu thập được dữ liệu tiến hành xác định các biến đầu vào(input), biến đầu ra (output), các tập kiểm thử, tập huấn luyện.

### Tiến hành các bước tiền xử lý dữ liệu

* Làm sạch dữ liệu:
* Xử lý giá trị thiếu: Dữ liệu về chất lượng nước thường có thể chứa các giá trị thiếu hoặc không xác định. Cách xử lý giá trị thiếu là loại bỏ các mẫu dữ liệu có giá trị thiếu.
* Xử lý dữ liệu chứa thông số sai, không có kết quả khai thác dữ liệu
* Xóa bỏ các dòng dữ liệu lỗi(hình ảnh):

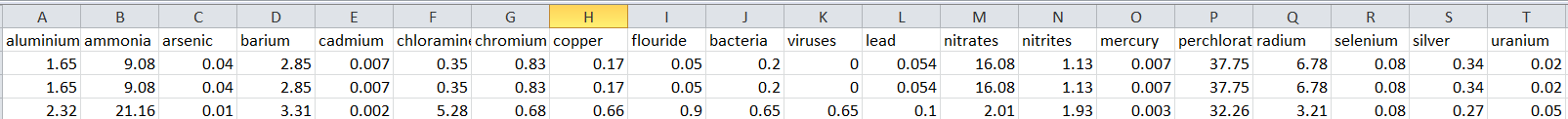




* Xóa bỏ các dữ liệu không hoặc thiếu nhất quán :
* Tích hợp và chuyển đổi dữ liệu:
* Chuẩn hóa dữ liệu: Chuẩn hóa dữ liệu giúp đưa các giá trị về cùng một khoảng giá trị và loại bỏ sự ảnh hưởng của tỷ lệ ban đầu. 

+ Tích hợp nhiều cơ sở dữ liệu:tổng hợp các chỉ số mang tính quyết định để đánh giá chất lượng nước

* Giảm dữ liệu:
* Loại bỏ các dữ liệu trùng lặp



Quá trình tiền xử lý dữ liệu trong đề tài này có thể linh hoạt và phụ thuộc vào đặc thù của dữ liệu cụ thể và mục tiêu phân loại. Việc chọn và áp dụng các phương pháp tiền xử lý phù hợp sẽ giúp cải thiện chất lượng dữ liệu và đảm bảo hiệu suất của mô hình phân loại chất lượng nước.

## **III. Xây dựng mô hình**

Sử dụng Weka làm công cụ khai phá dữ liệu chứa các công cụ phục vụ cho tiền xử lý dữ liệu, phân loại, hồi quy, phân cụm, các luật liên quan và trực quan hóa. Nó cũng phù hợp cho việc phát triển, xây dựng các mô hình học máy và có khả năng chạy được trên nhiều hệ điều hành khác nhau.

Tỉ lệ huấn luyện - kiểm thử : 70 – 30.

Từ kết quả phân lớp xác định được nguồn nước nào đang bị rơi vào tình trạng cảnh báo mức độ ô nhiễm . Nếu có những cảnh bảo nhanh chóng và kịp thời thì hàng năm tỷ lệ nguồn nước ô nhiễm sẽ giảm đi rất nhiều. Bằng phương pháp khai phá dữ liệu, phân lớp sinh viên dựa trên cây quyết định.

Tất cả các thuộc tính đều là các biến số và dữ liệu chi tiết của các chỉ số lần lượt như sau:

* Nhôm - khoảng an toàn (max=5.05 - min=0)
* Amoniac - khoảng an toàn (max=29.84 - min= -0.08)
* Arsenic - khoảng an toàn (max=1.05 - min=0)
* barium - khoảng an toàn (max=4.94 - min=0)
* cadmium - khoảng an toàn (max=0.13 - min=0)
* chloramine - khoảng an toàn (max=8.68 - min=0)
* chromium - khoảng an toàn (max=0.9 - min=0)
* copper - khoảng an toàn (max=2 - min=0)
* fluoride - khoảng an toàn (max=1.5 - min=0)
* bacteria - khoảng an toàn (max=1, - min=0)
* viruses - khoảng an toàn (max=1, - min=0)
* lead - khoảng an toàn (max=0.2 - min=0.001)
* nitrates - khoảng an toàn (max=19.83 - min=0.01)
* nitrites - khoảng an toàn (max=2.92 - min=0)
* mercury- khoảng an toàn (max=60.01 - min=0)
* Perchlorate - khoảng an toàn (max=7.99 - min=0)
* radium - khoảng an toàn (max=0.1 - min=0)
* selenium - khoảng an toàn (max=0.1 - min=0)
* silver - khoảng an toàn (max=0.09 - min=0)
* uranium - khoảng an toàn (max=1 - min=0)
* is\_safe - thuộc tính lớp {0 - không an toàn, 1 - an toàn}

Tiến hành triển khai xây dựng mô hình bằng cách sử dụng một số các kĩ thuật Data mining và Machine learning dựa trên cơ sở 5 giải thuật Data Mining sau:

### Thuật toán Naive Bayes:

3.1.1 Cơ sở lý thuyết

* Phân loại Naive Bayes là một trong các phương pháp học máy được áp dụng phổ biến nhất trong thực tế, dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê
* Định lý Bayes:

Trong đó:

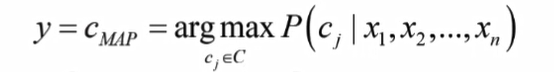
P(A): Xác suất xảy ra của A

P(B): Xác suất xảy ra của B

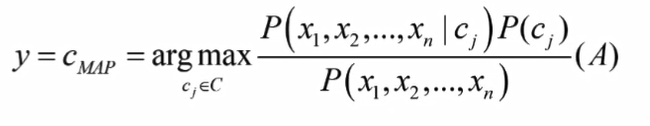
P(B|A): Xác suất của B khi biết A (xác suất có điều kiện)

P(A|B): Xác suất của A khi biết B (xác suất có điều kiện)

* Việc phân loại sẽ dựa trên các giá trị xác suất của các khả năng xảy ra các giả thuyết (phân loại)
* Mặc dù được đặt giả sử về sự độc lập có điều kiện của các thuộc tính đối với các phân lớp, nhưng phương pháp phân loại Naive Bayes vẫn thu được các kết quả phân loại tốt trong nhiều lĩnh vực ứng dụng thực tế
* Trong giai đoạn huấn luyện, dữ liệu huấn luyện được cung cấp dưới dạng các mẫu <xi, yi>. Sau khi huấn luyện xong, bộ phân loại cần dự đoán nhãn cho mẫu mới x.
* Theo lý thuyết học Bayes, nhãn phân loại được xác định bằng cách tính xác suất điều kiện của nhãn khi quan sát thấy tổ hợp giá trị thuộc tính {x1, x2, x3, …, xn}. Thuộc tính được chọn, ký hiệu CMAP là thuộc tính có xác suất điều kiện cao nhất (MAP là viết tắt của maximum a posterior), tức là



* Sử dụng quy tắc Bayes, biểu thức trên được viết lại như sau



* Tại công thức A nhận thấy giá trị của P{x1, x2, x3, …, xn} không phụ thuộc vào giá trị của c, do đó không làm ảnh hưởng đến CMAP nên có thể viết lại như sau:



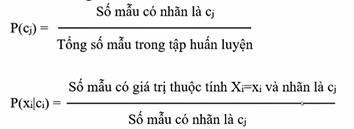
* Giá trị P(cj) được tính bằng tần suất quan sát thấy nhãn cj trên tập huấn luyện. Việc tính P{x1, x2, x3, …, xn|cj) khó khăn hơn nhiều. Vấn đề là số tổ hợp giá trị của n thuộc tính cùng với nhãn phân loại là rất lớn khi n lớn. Để tính xác suất này được chính xác, mỗi tổ hợp giá trị thuộc tính phải xuất hiện cùng nhãn phân loại đủ nhiều, trong khi số mẫu huấn luyện thường không đủ lớn.
* Để giải quyết vấn đề trên, giả sử các thuộc tính là độc lập về xác suất với nhau khi biết nhãn phân loại cj
* Với giả thiết về tính độc lập xác suất có điều kiện, có thể viết như sau:



tức là xác suất đồng thời quan sát thấy các thuộc tính bằng tích xác suất điều kiện của từng thuộc tính riêng lẻ. Thay vào biểu thức ở trên, ta được bộ phân loại Bayes đơn giản (có đầu ra ký hiệu là cNB) như sau:



* Quá trình huấn luyện hay học Bayes đơn giản là quá trình tính các xác suất P(cj) và các xác suất điều kiện P(xi|cj) bằng cách đếm trên tập dữ liệu huấn luyện. Các xác suất P(cj) và các xác suất điều kiện P(xi|cj) được tính trên tập dữ liệu huấn luyện theo công thức sau:



3.1.2. Ưu điểm của Naive Bayes

* Đơn giản dễ hiểu dễ triển khai để dự đoán lớp của tập dữ liệu thử nghiệm.
* Giả định độc lập: khi giả định về sự độc lập giữa các đặc trưng là đúng, Naive Bayes có khả năng hoạt động tốt hơn các mô hình khác.
* Cần ít dữ liệu hơn và quá trình dự đoán diễn ra nhanh chóng
* Naive Bayes có khả năng xử lý dữ liệu bị thiếu hay nhiễu một cách hiệu quả.
* Naive Bayes có khả năng hoạt động tốt trong các bài toán nhiều lớp phân loại.

### Thuật toán cây quyết định (Decision Trees):

3.2.1 Cơ sở lý thuyết về cây quyết định

* Cây quyết định nhận đầu vào là một bộ giá trị thuộc tính mô tả một đối tượng hay một tình huống nào đó và trả về một giá trị rời rạc.
* Mỗi bộ thuộc tính đầu vào được gọi là một mẫu hay một ví dụ, đầu ra gọi là loại hay nhãn phân loại. Thuộc tính đầu vào còn được gọi là đặc trưng và có thể nhận giá trị rời rạc hoặc liên tục.
* Để cho đơn giản, trước tiên sẽ xem xét thuộc tính rời rạc, sau đó sẽ mở rộng cho trường hợp thuộc tính nhận giá trị liên tục.
* Thuật toán xây dựng

Input:Tập dữ liệu huấn luyện

Output:Cây quyết định

* Khởi đầu:nút hiện thời là nút gốc chứa toàn bộ tập dữ liệu huấn luyện
* Tại nút hiện thời n chọn thuộc tính

Chưa được sử dụng( tức là nút nằm trên đường đi từ gốc tới nút hiện thời)

Cho phép phân chia tập dữ liệu hiện thời thành các tập con tốt nhất

* Với mỗi giá trị thuộc tính được chọn:

Thêm một nút con bên dưới

Chia mẫu ở nút hiện thời về các nút con theo giá trị thuộc tính được chọn

* Lặp (đệ quy) với mỗi nút con cho tới khi

Tất cả các thuộc tính đã được sử dụng ở các nút phía trên, hoặc

Tất cả ví dụ tại nút hiện thời có cùng nhãn phân loại

Nhãn của nút được lấy theo đa số nhãn của ví dụ tại nút hiện thời

3.2.2 Ưu điểm của cây quyết định

* Cây quyết định dễ hiểu có thể hiểu mô hình cây quyết định sau khi được giải thích ngắn.
* Dữ liệu chuẩn bị cho cây quyết định là cơ bản không đòi hỏi các kĩ thuật chuẩn hóa dữ liệu tạo biến phụ…
* Cây quyết định có thể xử lí cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại.
* Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn. Có thể dùng máy tính cá nhân để phân tích các lượng dữ liệu lớn trong 1 thời gian đủ ngắn

## **IV. Đánh giá mô hình học máy**

Đánh giá mô hình thông qua ma trận nhầm lẫn (Confusing Matrix).

* Ma trận nhầm lẫn (confusion matrix) là công cụ trực quan và hiệu quả để đánh giá hiệu suất của mô hình phân loại. Nó thể hiện rõ ràng mức độ chính xác của mô hình trong việc dự đoán các lớp (classes) khác nhau.
* Phương pháp đánh giá:
* Tính toán các thước đo
  + - * Accuracy: tỉ lệ giữa số điểm dữ liệu được dự đoán đúng và tổng số điểm dữ liệu.
      * Precision: tỉ lệ số điểm true positive trong số những điểm được phân loại là positive
      * Recall: tỉ lệ số điểm true
      * positive trong số những điểm thực sự là positive
      * F1 Score: Điểm cân bằng giữa precision và recall.
* Phân tích các thước đo trong ma trận: Một mô hình tốt khi cả Precision và Recall đều cao có nghĩa là ít phân loại nhầm và tỷ lệ bỏ sót thấp
* Công thức tính:
* Accuracy= (TP+TN)/(TP+TN+FP+FN)
* Precision= TP/(TP+ FP)
* Recall= TP/(TP+ FN)
* Accuracy=2\* (Precision\*Recall )/(Precision+Recall)

Trong đó:

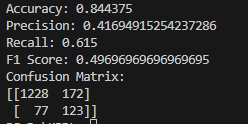
TP (True Positive): Số lượng các trường hợp mà mô hình dự đoán đúng là positive (dương tính) và thực tế cũng là positive.

TN (True Negative): Số lượng các trường hợp mà mô hình dự đoán đúng là negative và thực tế cũng là negative.

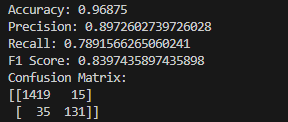
FP (False Positive): Số lượng các trường hợp mà mô hình dự đoán là positive nhưng thực tế lại là negative (âm tính).

FN (False Negative): Số lượng các trường hợp mà mô hình dự đoán là negative nhưng thực tế lại là positive.

* Đánh giá mô hình NaïveBayes:



* + Nhận xét:
* Mô hình có Accuracy cao (84.4%)
* Mô hình có Precision thấp (41.7%)
* Mô hình có Recall trung bình (62%)
* Mô hình có: F1 Score trung bình (50%)
  + - * Mô hình có hiệu suất thấp
* Đánh giá mô hình Decision Trees:



* + Nhận xét:
* Mô hình có Accuracy rất cao (96.8%)
* Mô hình có Precision cao (89.7%)
* Mô hình có Recall cao (79%)
* Mô hình có: F1 Score cao (84%)
  + - * Mô hình có hiệu suất rất tốt
* So sánh 2 mô hình

|  | Naive Bayes | Decision Tree |
| --- | --- | --- |
| Accuracy | 84.4% | 96.8% |
| Precision | 41.7% | 89.7% |
| Recall | 62% | 79% |
| F1 – Score | 50% | 84% |

Nhận xét

Decision Tree có độ chính xác cao hơn so với Naive Bayes (96.8% so với 84.4%). Điều này cho thấy Decision Tree có khả năng phân loại chính xác hơn trên tập dữ liệu kiểm tra.

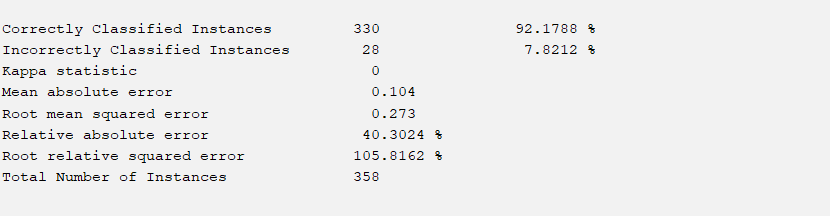
Decision Tree có Precision cao hơn gấp đôi Naive Bayes (89.7% so với 41.7%). Điều này cho thấy Decision Tree có tỉ lệ dự đoán đúng positive cao hơn Naive Bayes.

Naive Bayes có Recall mức trung bình (62%), trong khi Decision Tree là 79%. Điều này chỉ ra rằng Decision Tree có khả năng phát hiện các trường hợp positive tốt hơn Naive Bayes trên tập dữ liệu kiểm tra.

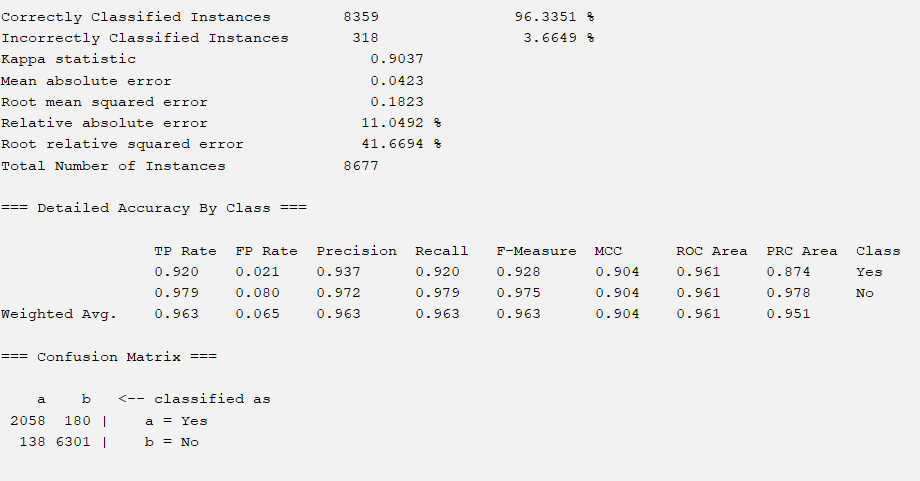
Decision Tree có F1 Score cao hơn (84% so với 50%). F1 Score kết hợp cả Precision và Recall, do đó, mô hình Decision Tree có sự cân bằng giữa việc dự đoán chính xác positive và phát hiện được nhiều positive hơn.

* + - * Mô hình Decision Tree có hiệu suất tốt hơn cho bài toán đặt ra

Test mô hình Decision Tree với file test được trích ra từ file gốc với 2000 bản ghi



Cross validation về Decision Tree



## **V. Tích hợp mô hình vào hệ thống**

1. Thuật toán

from tkinter import messagebox

import tkinter as tk

import joblib

import numpy as np

column\_names = ["aluminium", "ammonia", "arsenic", "barium", "cadmium", "chloramine", "chromium", "copper", "flouride", "bacteria", "viruses", "lead",

"nitrates", "nitrites", "mercury", "perchlorate", "radium", "selenium", "silver", "uranium"]

model = joblib.load("model/Wquality.joblib")

def predict\_water\_quality(attributes):

X = np.array(attributes).reshape(1, -1)

prediction = model.predict(X)

return "Sạch" if prediction[0] == 1 else "Bẩn"

def on\_predict():

try:

attributes = [float(entry.get()) for entry in entry\_boxes]

result = predict\_water\_quality(attributes)

messagebox.showinfo("Kết quả", f"Chất lượng nước được dự đoán là: {result}")

except ValueError:

messagebox.showerror("Lỗi", "Vui lòng nhập các giá trị số cho tất cả các thuộc tính!")

def clear\_entries():

for entry in entry\_boxes:

entry.delete(0, tk.END)

root = tk.Tk()

root.title("Xác định chất lượng nước")

entry\_boxes = []

for i, column\_name in enumerate(column\_names):

label = tk.Label(root, text=f"Thuộc tính {column\_name}:")

label.grid(row=i, column=0, padx=5, pady=5)

entry = tk.Entry(root)

entry.grid(row=i, column=1, padx=5, pady=5)

entry\_boxes.append(entry)

predict\_button = tk.Button(root, text="Dự đoán", command=on\_predict)

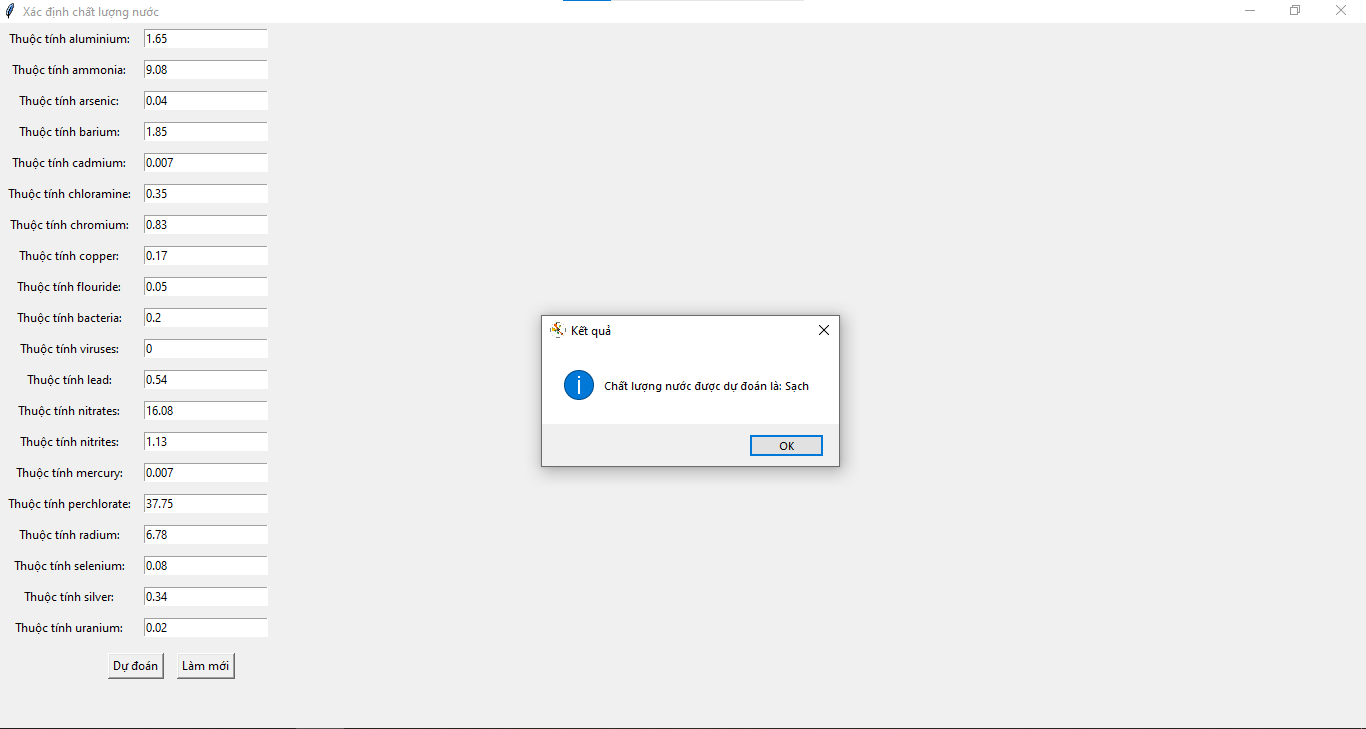
predict\_button.grid(row=20, column=0, columnspan=2, pady=10)

clear\_button = tk.Button(root, text="Làm mới", command=clear\_entries)

clear\_button.grid(row=len(column\_names), column=1, padx=5, pady=10)

root.mainloop()

1. Kết quả chương trình



## **VI. Kết luận**

Kết luận theo từng mục tiêu của đề tài

Trong quá trình thực hiện bài tập lớn, nhóm đã thực hiện các công việc sau:

* Nghiên cứu về các phương pháp học máy có giám sát.
* Áp dụng các giai đoạn của chu trình học máy vào bài toán thực tế
* Thực hiện đánh giá, so sánh 2 mô hình học để lựa chọn mô hình học máy tốt nhất
* Tích hợp được mô hình học máy vào ứng dụng để giải quyết các vấn đề trong lĩnh vực nghiên cứu đánh giá chất lượng nước trong các hệ thống cấp nước.

# Tài liệu tham khảo

[1] Bài giảng môn Nhập môn Khai phá dữ liệu và Máy học, TS Vũ Xuân Hạnh, Khoa Công nghệ thông tin, Trường Đại Học Mở Hà Nội

[2] "Nghiên cứu dự báo chất lượng nước dưới đất tại khu vực thành phố Hội An, Quảng Nam, Việt Nam", Tạp Chí Khoa Học Và Công Nghệ, Lê Phước Cường, Ngô Viết Thắng, [jst-ud.vn/jst-ud/article/view/7694](http://jst-ud.vn/jst-ud/article/view/7694).

[3] "Water Quality." Kaggle: Your Machine Learning and Data Science Community, www.kaggle.com/datasets/mssmartypants/water-quality.

[4] joykimaiyo18. "Quality Prediction Decision Tree Classifier." Kaggle: Your Machine Learning and Data Science Community, 2 July 2023, [www.kaggle.com/code/joykimaiyo18/quality-prediction-decision-tree-classifier/notebook#Data-Pre-processing](http://www.kaggle.com/code/joykimaiyo18/quality-prediction-decision-tree-classifier/notebook#Data-Pre-processing).

[5] PhilArchive: The Philosophy E-Print Archive, philarchive.org/archive/ALMNNW.