|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN MÔN HỌC**  **KHAI PHÁ DỮ LIỆU LỚN**  **DỰ ĐOÁN ĐỘ AN TOÀN CỦA NƯỚC**  Giảng viên giảng dạy: NGUYỄN QUAN DUY TÙNG  Sinh viên thực hiện: LÊ XUÂN TRỌNG  MSSV : 2100009483  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : KHAI PHÁ DỮ LIỆU LỚN  Khóa : 2021    **Tp.HCM, tháng 10 Năm 2025** |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**  **KHOA CÔNG NGHỆ THÔNG TIN**  Description: C:\Documents and Settings\Administrator\Desktop\logo dai hoc_khong nen.png  **TIỂU LUẬN MÔN HỌC**  **KHAI PHÁ DỮ LIỆU LỚN**  **DỰ ĐOÁN ĐỘ AN TOÀN CỦA NƯỚC**  Giảng viên giảng dạy: NGUYỄN QUAN DUY TÙNG  Sinh viên thực hiện: LÊ XUÂN TRỌNG  MSSV : 2100009483  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : KHAI PHÁ DỮ LIỆU LỚN  Khóa : 2021    **Tp.HCM, tháng 10 Năm 2025** |

|  |  |
| --- | --- |
| Trường Đại học Nguyễn Tất Thành  **Khoa Công Nghệ Thông Tin**  🙜  🙜  🙝  🙝 | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập – Tự do – Hạnh phúc**  🙜 🙜 🙝 🙝 |

**NHIỆM VỤ ĐỒ ÁN**

Họ và tên: **LÊ XUÂN TRỌNG** MSSV: **2100009483**

Chuyên ngành: **Khoa học dữ liệu** Lớp: **21DTH3A** Email: **xuantrongle987@gmail.com** SĐT: **0366394982** Tên đề tài: Dự đoán độ an toàn của nước

Giảng viên hướng dẫn: **Nguyễn Quan Duy Tùng**

**NỘI DUNG ĐỀ TÀI:**

Dự đoán độ an toàn của nước là vấn đề quan trọng trong việc bảo vệ sức khỏe. Đề tài sử dụng ba mô hình học máy: Logistic Regression, Random Forest và Decision Tree để phân tích các đặc trưng dữ liệu ảnh hưởng đến độ an toàn của nước. Kết quả thử nghiệm cho thấy mô hình Random Forest đạt độ chính xác cao nhất và được chọn làm mô hình tối ưu.

**PHƯƠNG PHÁP:**

* Tiền xử lý dữ liệu
* Trực quan hoá các đặc trưng ảnh hưởng đến độ an toàn
* Xây dựng, huấn luyện mô hình và chọn mô hình độ chính xác cao nhất

Thời gian thực hiện: **15 /09/2025 đến 01/ 10 /2025**

Nội dung và yêu cầu đã được thông qua Bộ môn.

*TP.HCM, ngày 01 tháng 10 năm 2025*

|  |  |
| --- | --- |
| **Q. TRƯỞNG BỘ MÔN**  *(Ký và ghi rõ họ tên)* | **GIÁO VIÊN HƯỚNG DẪN**  *(Ký và ghi rõ họ tên)* |

**LỜI MỞ ĐẦU**

Trong bối cảnh phát triển mạnh mẽ của công nghệ thông tin và dữ liệu lớn, học máy (Machine Learning) đã trở thành một công cụ không thể thiếu trong việc giải quyết các bài toán phức tạp thuộc nhiều lĩnh vực khác nhau. Đặc biệt, trong ngành công nghiệp thực phẩm và đồ uống, học máy đang mở ra nhiều cơ hội mới cho việc tối ưu hóa sản phẩm, dự đoán xu hướng tiêu dùng và nâng cao trải nghiệm khách hàng.

Dự án này tập trung vào việc áp dụng Apache Spark – một công cụ mạnh mẽ cho xử lý dữ liệu lớn – để dự đoán loại nước uống dựa trên các đặc trưng đã cho. Spark không chỉ giúp tối ưu hóa quá trình xử lý dữ liệu mà còn cung cấp một môi trường thuận lợi cho việc triển khai các mô hình học máy trên quy mô lớn.

Lời mở đầu của đồ án sẽ giới thiệu tổng quan về tầm quan trọng của học máy trong ngành công nghiệp thực phẩm, vai trò của Apache Spark trong xử lý và phân tích dữ liệu lớn, cùng với mục tiêu của đồ án trong việc áp dụng các thuật toán học máy để dự đoán chính xác loại nước uống từ dữ liệu thô. Qua đó, dự án không chỉ cung cấp một cái nhìn sâu sắc về tiềm năng của công nghệ này mà còn góp phần nâng cao khả năng ứng dụng thực tiễn trong ngành công nghiệp đồ uống.

**LỜI CẢM ƠN**

Nếu như phải gửi một lời cảm ơn dành cho các thầy cô đã trợ giúp cho mình thì cho em xin phép mượn câu nói của ngài William a Warrrd, ông từng nói rằng:

“Một người thầy trung bình chỉ biết nói

Một người thầy giỏi biết giải thích

Một người thầy chúng biết minh họa

Một người thầy vĩ đại biết truyền cảm hứng”

Nếu phải dành lời cảm ơn em muốn mượn câu nói này để dành cho các thầy cô phụ trách, cũng như là thầy Nguyễn Quan Duy Tùng giảng viên phụ trách môn Khai phá dữ liệu lớn đã giúp cũng như hỗ trợ chúng em hoàn thành được đồ án về chủ đề Dự đoán độ an toàn của nước. Cũng từ đồ án này em nhận ra mình cần phải có được tội lỗi phàm ăn để khơi lên cơn đói khát về kiến thức của mình từ đó ăn bớt những hạn chế của mình trong con sông kiến thức cũng như tránh xa các tội lỗi kiêu ngạo, lười biếng và đố kị để làm tròn nên những kinh nghiệm mà mình cần phải học hỏi từ mọi người để bước gần hơn tới thành công. Em cũng mong nhận được những lời gớp ý giúp em cải thiện cũng hoàn thiện đồ án hơn.

Cuối cùng em xin kính chúc quí thầy cô có một ngày tốt lạnh.

Em chân thành cảm ơn

Sinh viên thực hiện

LÊ XUÂN TRỌNG

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ II NĂM HỌC 2023 - 2024** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: Khai phá dữ liệu lớn Lớp học phần:22DHT.HL10

Nhóm sinh viên thực hiện:

1. Lê Xuân Trọng Tham gia đóng góp:100%

2. Tham gia đóng góp:

3. Tham gia đóng góp:

4. Tham gia đóng góp:

5. Tham gia đóng góp:

6. Tham gia đóng góp:

7. Tham gia đóng góp:

8. Tham gia đóng góp:

Ngày thi: 01/10/2025 Phòng thi: VP.K.CNTT-QUẬN 12

Đề tài tiểu luận/báo cáo của sinh viên : Dự đoán độ an toàn của nước

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo | Gồm 3 chương | 1 |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần | Chương 2 và 3 | 5 |  |
| * Lập luận | Chương 1 | 2 |  |
| * Kết luận |  | 1 |  |
| Trình bày | Theo chuẩn format luận văn font chữ 13, canh trái, phải,… | 1 |  |
| **TỔNG ĐIỂM** |  | **10** |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIẢNG VIÊN GIẢNG DẠY**

*Tp.HCM, Ngày 01 tháng 10 năm 2025*

**Giảng viên giảng dạy**

(Ký tên và ghi rõ họ tên)

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc209961052)

[1.1. Giới thiệu đề tài. 1](#_Toc209961053)

[1.2.Lý do chọn đề tài 2](#_Toc209961054)

[1.3.Mục tiêu đề tài. 4](#_Toc209961055)

[1.4. Đối tượng và phạm vi nghiên cứu. 5](#_Toc209961056)

[1.4.1. Đối tượng nghiên cứu 5](#_Toc209961057)

[1.4.2. Phạm Vi Nghiên Cứu 5](#_Toc209961058)

[CHƯƠNG 2 ỨNG DỤNG VÀ THUẬT TOÁN 7](#_Toc209961059)

[2.1 Phương pháp đề tài. 7](#_Toc209961060)

[2.2 Mô tả thuật toán. 8](#_Toc209961061)

[2.2.1. Thuật toán Logistic Regression 8](#_Toc209961062)

[2.2.2. Thuật toán Random Forest 9](#_Toc209961063)

[2.2.3. Thuật toán Decision Tree: 9](#_Toc209961064)

[2.3. Xây dựng bộ dữ liệu. 10](#_Toc209961065)

[2.3.1. Bộ dữ liệu gồm các cột 11](#_Toc209961066)

[2.4. Thực nghiệm với thư viện. 12](#_Toc209961067)

[CHƯƠNG 3 TRỰC QUAN HOÁ DỮ LIỆU 13](#_Toc209961068)

[3.2. Một số ảnh trực quan hóa được phân tích trong dự án 13](#_Toc209961069)

[3.2.1 Heatmap hệ số tương quan giữa các đặc trưng chất lượng nước 13](#_Toc209961070)

[3.2.2. Trực quan theo độ pH 14](#_Toc209961071)

[3.2.3. Trực quan độ pH ảnh hưởng đến độ an toàn của nước 15](#_Toc209961072)

[3.2.4. Trực quan theo độ Hardness 15](#_Toc209961073)

[3.2.5. Trực quan độ Hardness ảnh hưởng đến độ an toàn của nước 16](#_Toc209961074)

[3.2.6. Trực quan Solids 17](#_Toc209961075)

[3.2.7. Trực quan theo Solids ảnh hưởng đến độ an toàn của nước 18](#_Toc209961076)

[3.2.8. Trực quan theo Chloramines 19](#_Toc209961077)

[3.2.9. Trực quan theo Chloramines ảnh hưởng đến độ an toàn của nước 20](#_Toc209961078)

[3.2.10. Trực quan theo Sulfate 21](#_Toc209961079)

[3.2.11. Trực quan theo Sulfate ảnh hưởng đến độ an toàn của nước 22](#_Toc209961080)

[3.2.12. Trực quan theo Conductivity 23](#_Toc209961081)

[3.2.13. Trực quan theo Conductivity ảnh hưởng đến độ an toàn của nước 24](#_Toc209961082)

[3.2.14. Trực quan theo Organic\_carbon 25](#_Toc209961083)

[3.2.15. Trực quan theo Organic\_carbon ảnh hưởng đến độ an toàn của nước 26](#_Toc209961084)

[3.2.16. Trực quan theo Trihalomethanes 27](#_Toc209961085)

[3.2.17. Trực quan theo Trihalomethanes ảnh hưởng đến độ an toàn của nước 28](#_Toc209961086)

[3.2.18. Trực quan theo Turbidity 29](#_Toc209961087)

[3.2.19. Trực quan theo Turbidity ảnh hưởng đến độ an toàn của nước 30](#_Toc209961088)

[CHƯƠNG 4 HUẤN LUYỆN MÔ HÌNH 32](#_Toc209961089)

[4.1 Xây dựng ứng dụng và giải thích 32](#_Toc209961090)

[4.1.1. Logistic Regression trong Spark 32](#_Toc209961091)

[4.1.2. Decision Tree Classifier 32](#_Toc209961092)

[4.1.3. Random Forest 33](#_Toc209961093)

[4.1.4. So sánh các phương pháp dự đoán 33](#_Toc209961094)

[KẾT LUẬN 35](#_Toc209961095)

[Kết quả đạt được. 35](#_Toc209961096)

[Hạn chế và hướng phát triển . 35](#_Toc209961097)

[TÀI LIỆU THAM KHẢO 37](#_Toc209961098)

**DANH MỤC HÌNH**

[Hình 1: Bảng dữ liệu 10](#_Toc209961099)

[Hình 2: Heatmap hệ số tương quan giữa các đặc trưng chất lượng nước 12](#_Toc209961100)

[Hình 3: Biểu đồ biểu thị sự phân bố của mức độ pH 13](#_Toc209961101)

[Hình 4: Trực quan độ pH ảnh hưởng đến độ an toàn của nước 14](#_Toc209961102)

[Hình 5: Trực quan hóa theo độ Hardness 15](#_Toc209961103)

[Hình 6: Trực quan hóa về độ Hardness ảnh hưởng đến độ an toàn của nước 16](#_Toc209961104)

[Hình 7: Trực quan hoá Solids 17](#_Toc209961105)

[Hình 8: Trực quan theo Solids ảnh hưởng đến độ an toàn của nước 18](#_Toc209961106)

[Hình 9: Trực quan hóa theo Chloramines 19](#_Toc209961107)

[Hình 10: Trực quan theo Chloramines ảnh hưởng đến độ an toàn của nước 20](#_Toc209961108)

[Hình 11: Trực quan theo Sulfate 21](#_Toc209961109)

[Hình 12: Trực quan theo Sulfate ảnh hưởng đến độ an toàn của nước 22](#_Toc209961110)

[Hình 13: Trực quan theo Conductivity 23](#_Toc209961111)

[Hình 14: Trực quan theo Conductivity ảnh hưởng đến độ an toàn của nước 24](#_Toc209961112)

[Hình 15: Trực quan theo Organic\_carbon 25](#_Toc209961113)

[Hình 16: Trực quan theo Organic\_carbon ảnh hưởng đến độ an toàn của nước 26](#_Toc209961114)

[Hình 17: Trực quan theo Trihalomethanes 27](#_Toc209961115)

[Hình 18: Trực quan theo Trihalomethanes ảnh hưởng đến độ an toàn của nước 28](#_Toc209961116)

[Hình 19: Trực quan theo Turbidity 29](#_Toc209961117)

[Hình 20: Trực quan theo Turbidity ảnh hưởng đến độ an toàn của nước 30](#_Toc209961118)

[Hình 21: Dữ liệu chuẩn bị cho phần học máy 31](#_Toc209961119)

[Hình 22: Kết quả mô hình Logistic Regreession trong Spark 31](#_Toc209961120)

[Hình 23: Kết quả mô hình Decision Tree Classifier trong Spark 32](#_Toc209961121)

[Hình 24: Kết quả mô hình Random Forest 32](#_Toc209961122)

[Hình 25: Trực quan hóa của phần trăm của các dự báo 33](#_Toc209961123)

# CHƯƠNG 1: GIỚI THIỆU

## Giới thiệu đề tài.

Trong thời đại số hóa, sự bùng nổ về dữ liệu đã tạo ra cơ hội lớn cho các doanh nghiệp và tổ chức trong nhiều lĩnh vực khác nhau. Ngành công nghiệp thực phẩm và đồ uống cũng không nằm ngoài xu hướng này. Với khối lượng dữ liệu khổng lồ từ các nguồn khác nhau, từ nguyên liệu đầu vào, quy trình sản xuất cho đến hành vi tiêu dùng, việc khai thác và phân tích dữ liệu hiệu quả đang trở thành chìa khóa để đạt được lợi thế cạnh tranh.

Trong bối cảnh đó, việc áp dụng các phương pháp học máy (Machine Learning) không chỉ giúp tối ưu hóa các quy trình nội bộ mà còn mở ra những cơ hội mới trong việc phát triển sản phẩm và nâng cao trải nghiệm khách hàng.

Đề tài “Dự đoán khả năng uống được của nước” được thực hiện nhằm mục tiêu áp dụng các thuật toán học máy để phân loại và dự đoán loại nước uống dựa trên các đặc trưng cụ thể của sản phẩm. Đây là một vấn đề quan trọng trong ngành công nghiệp đồ uống, nơi mà việc nhận diện chính xác các sản phẩm có thể giúp doanh nghiệp cải thiện chất lượng sản phẩm, đáp ứng nhu cầu thị trường một cách hiệu quả hơn, và tăng cường sự hài lòng của khách hàng.

Apache Spark, với khả năng xử lý dữ liệu lớn nhanh chóng và hiệu quả, đã được lựa chọn làm nền tảng để thực hiện dự án này. Được thiết kế để xử lý các tập dữ liệu lớn và phức tạp, Spark cung cấp một môi trường lý tưởng cho việc triển khai các mô hình học máy ở quy mô lớn, từ việc xử lý dữ liệu đầu vào, huấn luyện mô hình, đến việc dự đoán và đánh giá kết quả. Trong dự án này, các thuật toán học máy phổ biến như Logistic Regression, Decision Tree, và Random Forest sẽ được áp dụng để xây dựng các mô hình dự đoán hiệu quả.

Đề tài này không chỉ dừng lại ở việc triển khai các thuật toán học máy mà còn đi sâu vào các bước tiền xử lý dữ liệu, từ việc làm sạch, chuẩn hóa đến việc lựa chọn đặc trưng phù hợp. Việc đánh giá mô hình cũng sẽ được thực hiện một cách kỹ lưỡng, sử dụng các chỉ số như Accuracy, Precision, Recall, và F1-Score để đảm bảo rằng các mô hình được xây dựng đạt hiệu quả cao nhất.

Ngoài ra, dự án còn xem xét các yếu tố thực tế trong quá trình ứng dụng công nghệ này vào ngành công nghiệp đồ uống, như khả năng mở rộng, chi phí tính toán, và tính khả thi của việc triển khai trong môi trường thực tế. Qua đó, đề tài không chỉ mang lại giá trị học thuật mà còn đóng góp những gợi ý thiết thực cho việc ứng dụng công nghệ học máy trong ngành công nghiệp thực phẩm và đồ uống.

Thông qua việc triển khai và đánh giá các mô hình học máy trên nền tảng Apache Spark, đề tài kỳ vọng sẽ mang lại những đóng góp mới mẻ trong việc ứng dụng công nghệ vào quá trình sản xuất và kinh doanh, đồng thời mở ra những hướng nghiên cứu tiềm năng trong tương lai.

## 1.2.Lý do chọn đề tài

Nước an toàn là nước không chứa các tạp chất, vi khuẩn, hóa chất, hoặc các yếu tố độc hại có thể gây hại cho sức khỏe con người. Để nước được coi là an toàn, nó cần đáp ứng các tiêu chuẩn về chất lượng, bao gồm các chỉ số như pH, độ cứng, hàm lượng tạp chất rắn, các hợp chất hóa học (chloramines, sulfate), và độ trong (turbidity). Nước an toàn phải không có vi sinh vật gây bệnh và phải nằm trong giới hạn cho phép của các chất ô nhiễm, đảm bảo không gây nguy hại khi sử dụng.

Trong bối cảnh công nghệ thông tin và dữ liệu lớn phát triển với tốc độ nhanh chóng, ngành công nghiệp thực phẩm và đồ uống đang trải qua những thay đổi đáng kể trong quy trình sản xuất và quản lý. Dữ liệu không chỉ đóng vai trò then chốt trong việc giúp doanh nghiệp ra quyết định chính xác, mà còn là công cụ quan trọng để nâng cao chất lượng sản phẩm và đáp ứng nhu cầu ngày càng đa dạng của thị trường. Tuy nhiên, việc khai thác và phân tích dữ liệu hiệu quả đòi hỏi sự kết hợp của các công nghệ tiên tiến, trong đó học máy (Machine Learning) và xử lý dữ liệu lớn là hai yếu tố quan trọng.

* **Lý do đầu tiên:** Để lựa chọn đề tài này là sự gia tăng trong việc sử dụng kết hợp giữa học máy và Apache Spark, một xu hướng hàng đầu trong việc xử lý và phân tích dữ liệu lớn. Apache Spark, với khả năng xử lý dữ liệu phân tán, không chỉ cung cấp nền tảng mạnh mẽ để triển khai các thuật toán học máy mà còn giúp giảm thời gian xử lý và tối ưu hóa hiệu suất. Điều này rất cần thiết trong ngành công nghiệp thực phẩm và đồ uống, nơi sự cạnh tranh ngày càng gia tăng và yêu cầu về tối ưu hóa quy trình ngày càng khắt khe.
* **Thứ hai:** Việc dự đoán và phân loại sản phẩm trong ngành công nghiệp đồ uống là vô cùng quan trọng. Khả năng nhận diện và phân loại chính xác các loại nước uống giúp doanh nghiệp không chỉ đáp ứng nhu cầu khách hàng mà còn tối ưu hóa chuỗi cung ứng, giảm chi phí sản xuất và nâng cao khả năng cạnh tranh trên thị trường. Ứng dụng các mô hình học máy để thực hiện các nhiệm vụ này mang lại nhiều lợi ích thực tiễn và có thể được áp dụng rộng rãi trong các hoạt động sản xuất và kinh doanh.
* **Thứ ba:** trong thời đại hiện nay, người tiêu dùng không chỉ yêu cầu sản phẩm chất lượng cao mà còn mong muốn có những trải nghiệm cá nhân hóa và phù hợp với nhu cầu riêng của họ. Học máy, với khả năng phân tích và dự đoán từ dữ liệu, giúp doanh nghiệp hiểu rõ hơn về thị hiếu của khách hàng, từ đó cải thiện sản phẩm và dịch vụ để đáp ứng tốt hơn nhu cầu thị trường.
* **Cuối cùng:** việc nghiên cứu đề tài này cũng mang lại giá trị học thuật và nghiên cứu, giúp làm rõ khả năng ứng dụng công nghệ học máy và xử lý dữ liệu lớn trong ngành công nghiệp thực phẩm và đồ uống. Điều này không chỉ góp phần nâng cao kiến thức và kỹ năng chuyên môn mà còn mở ra những hướng nghiên cứu và ứng dụng tiềm năng trong tương lai.

Với những lý do trên, việc lựa chọn đề tài "Sử dụng Spark để phân loại nước uống" là một quyết định đúng đắn và cần thiết, nhằm thúc đẩy việc ứng dụng công nghệ hiện đại trong quản lý và phát triển sản phẩm, đồng thời đáp ứng yêu cầu ngày càng cao của thị trường.

## 1.3.Mục tiêu đề tài.

Mục tiêu mà em đề ra cho đề tài " Dự đoán độ an toàn của nước " bao gôm các mục tiêu sau:

* **Phát triển hệ thống tự động phân loại các loại đồ uống:** Mục tiêu đầu tiên của đề tài là thiết kế một hệ thống sử dụng các thuật toán học máy nhằm tự động phân loại các loại đồ uống dựa trên các đặc trưng cụ thể của sản phẩm. Hệ thống này sẽ hỗ trợ doanh nghiệp trong việc phân loại sản phẩm một cách hiệu quả, tối ưu hóa quy trình sản xuất, đồng thời đảm bảo chất lượng sản phẩm được duy trì đồng nhất.
* **Ứng dụng Apache Spark trong xử lý và phân tích dữ liệu quy mô lớn:** Đề tài nhắm đến việc khai thác sức mạnh của Apache Spark để xử lý và phân tích dữ liệu quy mô lớn, đặc biệt trong lĩnh vực thực phẩm và đồ uống. Bằng cách áp dụng Spark, hệ thống sẽ có khả năng xử lý nhanh chóng và hiệu quả các tập dữ liệu lớn, từ đó nâng cao hiệu suất của các mô hình học máy được triển khai.
* **Đánh giá hiệu quả của các thuật toán học máy:** Một trong những mục tiêu chính của đề tài là thử nghiệm và so sánh hiệu quả của các thuật toán học máy như Logistic Regression, Decision Tree, và Random Forest trong việc phân loại đồ uống. Thông qua việc này, đề tài sẽ đánh giá độ chính xác, độ tin cậy, và khả năng tổng quát hóa của từng thuật toán, từ đó xác định mô hình tối ưu nhất cho bài toán.
* **Xây dựng quy trình chuẩn cho tiền xử lý và phân tích dữ liệu:** Mục tiêu của đề tài là phát triển một quy trình chuẩn cho tiền xử lý dữ liệu, bao gồm các bước như làm sạch, chuẩn hóa, và lựa chọn đặc trưng. Quy trình này nhằm đảm bảo dữ liệu đầu vào có chất lượng tốt, giúp các mô hình học máy được huấn luyện trên dữ liệu đáng tin cậy và chính xác.
* **Đề xuất các ứng dụng thực tiễn cho hệ thống phân loại:** Ngoài việc phát triển hệ thống, đề tài còn tập trung vào việc đề xuất các ứng dụng thực tiễn của hệ thống phân loại đồ uống trong ngành công nghiệp thực phẩm. Các ứng dụng này có thể bao gồm tối ưu hóa chuỗi cung ứng, cải thiện quy trình sản xuất, và cá nhân hóa trải nghiệm của khách hàng.
* **Đóng góp vào nghiên cứu và phát triển công nghệ:** Cuối cùng, đề tài còn nhằm mục tiêu đóng góp vào nghiên cứu học thuật và phát triển công nghệ, đặc biệt trong lĩnh vực học máy và xử lý dữ liệu lớn. Kết quả từ đề tài có thể mở ra những hướng nghiên cứu mới và cung cấp cơ sở cho việc phát triển các ứng dụng tương tự trong tương lai.

## 1.4. Đối tượng và phạm vi nghiên cứu.

Đối tượng và phạm vi nghiên cứu trong dự án sử dụng spark để dự đoán loại nước uống có thể được xác định như sau:

### **1.4.1. Đối tượng nghiên cứu**

Đối tượng nghiên cứu của chúng tôi là các loại nước uống với đặc điểm và thành phần khác nhau, bao gồm những sản phẩm có công thức độc đáo, thành phần nguyên liệu phức tạp, hoặc những loại đồ uống có đặc tính hóa lý đặc biệt. Việc tập trung vào nhóm đối tượng này nhằm giúp mô hình dự đoán phân loại chính xác và hiệu quả, hỗ trợ quá trình cải tiến và phát triển sản phẩm.

### **1.4.2. Phạm Vi Nghiên Cứu**

* **Phạm vi dữ liệu:** Nghiên cứu này sẽ tập trung vào phân loại các loại nước uống, với dữ liệu được lấy từ những nguồn đáng tin cậy. Dữ liệu bao gồm các đặc điểm hóa học, vật lý, và cảm quan của nước uống. Quá trình tiền xử lý sẽ đảm bảo rằng dữ liệu được chuẩn hóa và sẵn sàng để sử dụng trong các mô hình học máy.
* **Phạm vi công nghệ:** Nghiên cứu sẽ sử dụng Apache Spark cùng với thư viện MLlib để xây dựng, huấn luyện, và đánh giá các mô hình dự đoán. Các công cụ và kỹ thuật xử lý dữ liệu lớn, cùng với các thuật toán học máy, sẽ được triển khai trong môi trường Spark để đảm bảo tính hiệu quả và khả năng mở rộng.
* **Phạm vi ứng dụng:** Nghiên cứu sẽ áp dụng vào việc phân loại và dự đoán các loại nước uống. Kết quả sẽ được thử nghiệm trong môi trường thực tế nhằm đánh giá khả năng của mô hình trong việc nâng cao chất lượng sản phẩm, tối ưu hóa quy trình sản xuất, và đáp ứng nhu cầu cá nhân hóa của thị trường.

# CHƯƠNG 2 ỨNG DỤNG VÀ THUẬT TOÁN

## 2.1 Phương pháp đề tài.

Có một số phương pháp tiếp cận đề tài " Dự đoán độ an toàn của nước " mà em sử dụng để đạt được mục tiêu của nó

* Thu thập và tiền xử lý dữ liệu:
  + Bộ dữ liệu được sử dụng lấy từ Kaggle: <https://www.kaggle.com/datasets/adityakadiwal/water-potability?resource=download>
  + Ở giai đoạn này, dữ liệu được làm sạch, xử lý các giá trị bị thiếu và chuẩn hóa để sẵn sàng cho các bước phân tích tiếp theo.
* Khám phá và trực quan hóa dữ liệu:
  + Sử dụng bản đồ nhiệt (heatmap) để mô tả mức độ tương quan giữa các đặc trưng, từ đó đánh giá xem có đặc trưng nào trùng lặp hoặc ảnh hưởng lẫn nhau.
  + Thực hiện trực quan phân phối của từng đặc trưng trong bộ dữ liệu để quan sát sự khác biệt giữa mẫu nước uống được và không uống được.
  + Bước trực quan này giúp hiểu rõ cấu trúc dữ liệu, hỗ trợ cho việc lựa chọn và xây dựng mô hình dự đoán.
* Phân tích dữ liệu và mô hình hóa:
  + Các phương pháp học máy như hồi quy logistic, decision trees, random forests để phân tích dữ liệu. Những mô hình này sẽ giúp xác định và phân loại chính xác các loại nước uống dựa trên các đặc trưng đã được thu thập và chuẩn bị.
* Đánh giá và tối ưu hóa mô hình:
  + Đánh giá bằng các thước đo như accuracy, precision, recall, F1-score.
  + Sử dụng cross-validation và điều chỉnh siêu tham số để tối ưu hiệu quả mô hình.
* Triển khai và ứng dụng:
  + Sau khi chọn được mô hình tối ưu, có thể triển khai để dự đoán trên dữ liệu mới, hỗ trợ cho việc giám sát và quản lý chất lượng nước.

Việc áp dụng một cách hợp lý các phương pháp trên sẽ giúp xây dựng một quy trình phân tích toàn diện và đáng tin cậy trong việc dự đoán và phân loại các loại nước uống.

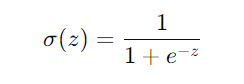
## 2.2 Mô tả thuật toán.

Đề tài " Dự đoán khả năng uống được của nước " tận dụng một loạt các thuật toán học máy khác nhau nhằm phát triển mô hình phân loại chính xác dựa trên các đặc trưng của dữ liệu. Apache Spark được lựa chọn không chỉ vì khả năng xử lý dữ liệu lớn mà còn vì nó cung cấp nền tảng mạnh mẽ để triển khai các thuật toán học máy với hiệu suất cao. Dưới đây là mô tả chi tiết về từng thuật toán:

### **2.2.1. Thuật toán Logistic Regression**

**Mô tả cơ bản:** Logistic Regression là một phương pháp học máy tuyến tính thường được sử dụng cho các bài toán phân loại nhị phân. Tuy nhiên, nó cũng có thể mở rộng cho các bài toán phân loại đa lớp thông qua phương pháp "one-vs-rest" (một so với tất cả).

**Cơ chế hoạt động:** Logistic Regression ước lượng xác suất một mẫu thuộc về một lớp cụ thể dựa trên tổng trọng số của các đặc trưng. Hàm sigmoid được sử dụng để chuyển đổi tổng trọng số này thành một giá trị xác suất nằm trong khoảng từ 0 đến 1. Công thức của hàm sigmoid như sau:



Trong đó, z=wTx+b với w là vector trọng số, x là vector đặc trưng, và b là hằng số bias.

**Huấn luyện mô hình trong Spark:** Spark MLlib triển khai Logistic Regression bằng cách tối ưu hóa hàm mất mát (log-loss) sử dụng thuật toán Gradient Descent. Mô hình sẽ được huấn luyện trên tập dữ liệu đã chuẩn hóa để tìm ra các trọng số tối ưu cho từng đặc trưng. Quá trình huấn luyện này được thực hiện trên toàn bộ cụm máy tính, tận dụng khả năng tính toán phân tán của Spark để giảm thời gian xử lý.

### **2.2.2. Thuật toán Random Forest**

**Mô tả cơ bản:** Random Forest là một phương pháp học máy mạnh mẽ, dựa trên việc kết hợp nhiều cây quyết định để đưa ra kết quả phân loại cuối cùng. Thuật toán này giúp giảm thiểu nguy cơ overfitting và cải thiện độ chính xác của mô hình.

**Cơ chế hoạt động:** Random Forest xây dựng nhiều cây quyết định trên các tập con ngẫu nhiên của dữ liệu và các đặc trưng. Mỗi cây trong rừng đóng góp một "phiếu bầu" cho kết quả phân loại cuối cùng. Kết quả phân loại của Random Forest là kết quả được chọn nhiều nhất từ các cây. Ngoài ra, việc sử dụng các tập con khác nhau và các đặc trưng ngẫu nhiên giúp tăng tính đa dạng giữa các cây, làm tăng độ chính xác tổng thể.

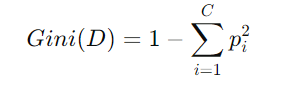
**Huấn luyện mô hình trong Spark:** Spark MLlib triển khai Random Forest bằng cách xây dựng các cây quyết định một cách song song, tận dụng khả năng xử lý phân tán của Spark. Người dùng có thể tùy chỉnh số lượng cây trong rừng (numTrees), độ sâu tối đa của mỗi cây (maxDepth), và các tham số khác để tối ưu hóa hiệu suất của mô hình.

### **2.2.3. Thuật toán Decision Tree:**

**Mô tả cơ bản:** Decision Tree là một thuật toán phân loại phi tuyến tính, chia dữ liệu thành các nhánh dựa trên các đặc trưng. Mỗi nhánh tương ứng với một quyết định về giá trị của một đặc trưng, và các nhánh này tiếp tục phân chia cho đến khi đạt đến kết quả phân loại cuối cùng.

**Cơ chế hoạt động:** Decision Tree hoạt động bằng cách chọn đặc trưng tốt nhất để phân chia dữ liệu tại mỗi nút. Tiêu chí lựa chọn thường là Gini impurity hoặc entropy, nhằm tối đa hóa "độ tinh khiết" của mỗi nhánh. Một cây quyết định có thể được mô tả như một loạt các quyết định "có/không" dẫn đến một kết quả phân loại cụ thể.

**Gini impurity:** Là một thước đo mức độ hỗn loạn của một tập hợp. Công thức tính Gini impurity cho một tập hợp các mẫu như sau:



Trong đó, pi là tỷ lệ các mẫu thuộc lớp i trong tập dữ liệu DDD.

**Huấn luyện mô hình trong Spark:** Trong Spark MLlib, Decision Tree được xây dựng bằng cách sử dụng kỹ thuật đệ quy để tạo ra các phân chia dựa trên các đặc trưng dữ liệu. Cây quyết định sẽ tiếp tục phân chia cho đến khi đạt đến một ngưỡng cụ thể như độ sâu tối đa hoặc số lượng mẫu tối thiểu tại mỗi nút. Spark phân tán quá trình xây dựng cây này trên nhiều nút để tăng tốc độ xử lý.

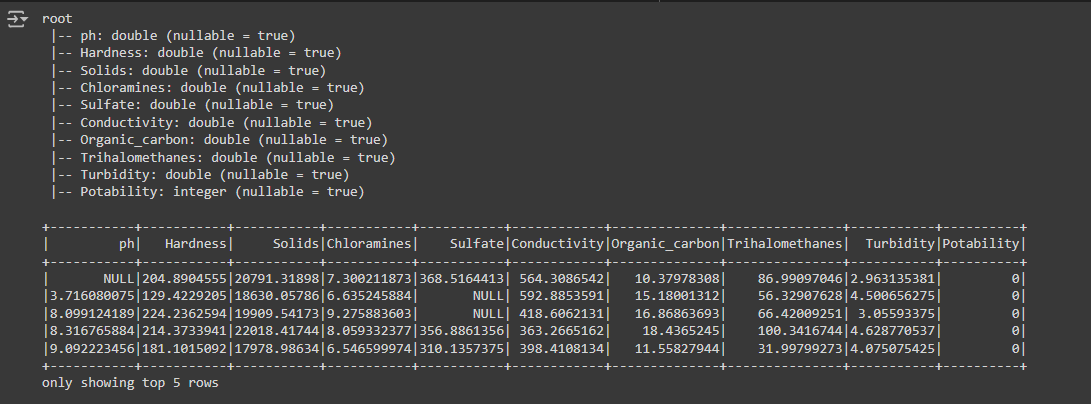
**2.2.4. Quy trình tổng quan sử dụng các thuật toán trong Spark**

* **Chuẩn bị dữ liệu:** Dữ liệu được thu thập, làm sạch, xử lý dữ liệu thiếu để chuẩn bị cho việc huấn luyện mô hình. Dữ liệu sau đó được chia thành tập huấn luyện và tập kiểm thử.
* **Huấn luyện mô hình:** Các thuật toán Logistic Regression, Decision Tree, Random Forest được áp dụng để huấn luyện trên tập dữ liệu đã chuẩn bị. Quá trình huấn luyện tận dụng khả năng xử lý phân tán của Spark để tăng tốc độ và hiệu quả.
* **Đánh giá mô hình:** Mỗi mô hình sau khi huấn luyện sẽ được đánh giá bằng các chỉ số như Accuracy, Precision, Recall, F1-Score, và AUC-ROC. Kết quả này sẽ giúp xác định mô hình nào hoạt động tốt nhất.

**Triển khai mô hình:** Mô hình đã tối ưu sẽ được triển khai trong môi trường thực tế, sẵn sàng cho việc phân loại và dự đoán các loại nước uống dựa trên các đặc trưng mới.

## 2.3. Xây dựng bộ dữ liệu.

### **2.3.1. Bộ dữ liệu gồm các cột**



Hình 1: Bảng dữ liệu

Bộ dữ liệu bao gồm 3277 dòng và 10 thuộc tính gồm:

1. **ph**: Độ pH của nước, biểu thị tính axit hoặc kiềm của mẫu nước.
2. **Hardness**: Độ cứng của nước, thường liên quan đến lượng khoáng chất hòa tan như canxi và magiê trong nước.
3. **Solids**: Hàm lượng chất rắn hòa tan trong nước, được đo bằng mg/L.
4. **Chloramines**: Nồng độ Chloramines, một hợp chất được sử dụng để khử trùng nước, đo bằng mg/L.
5. **Sulfate**: Hàm lượng Sulfate (SO4) trong nước, đo bằng mg/L.
6. **Conductivity**: Độ dẫn điện của nước, biểu thị khả năng nước dẫn điện và thường liên quan đến lượng ion hòa tan, đo bằng µS/cm.
7. **Organic\_carbon**: Hàm lượng carbon hữu cơ trong nước, đo bằng mg/L, chỉ ra mức độ ô nhiễm hữu cơ trong nước.
8. **Trihalomethanes**: Nồng độ Trihalomethanes, một nhóm hợp chất hóa học có thể hình thành khi khử trùng nước, đo bằng µg/L.
9. **Turbidity**: Độ đục của nước, biểu thị mức độ trong suốt hoặc độ mờ đục của nước, đo bằng NTU (Nephelometric Turbidity Units).
10. **Potability**: Khả năng uống được của nước (biến mục tiêu)

* 0 biểu thị nước không an toàn để uống
* 1 biểu thị nước an toàn để uống.

## 2.4. Thực nghiệm với thư viện.

Chắc hẳn bất kỳ ai học lập trình đều quen thuộc với thuật ngữ "Modules" – đề cập đến các thư viện lập trình chứa những đoạn mã thường xuyên được sử dụng, được các lập trình viên phát triển. Để chuẩn bị cho đề tài "Sử dụng Spark dự đoán nước uống," việc sử dụng các thư viện lập trình là điều không thể thiếu. Sau đây, em xin trình bày một số thư viện đã được sử dụng trong quá trình thực hiện dự án này. Những thư viện này không chỉ hỗ trợ tối ưu hóa thời gian phát triển mà còn cung cấp các công cụ hữu ích cho việc xử lý dữ liệu và xây dựng mô hình dự đoán.

**1.pandas (pd):**

Chúng ta sẽ bắt đầu với thư viện gần gũi với những người mới bắt đầu học về python nói tới đây chắc mọi người cũng biết bọn mình đang nhắc tới thư viện gì rồi nhỉ. Bingo đoán đúng rồi đó thư viện mà mình muốn nói tới ở đầy đó là Pandas. Một thư viện mạnh mẽ cho xử lý và phân tích dữ liệu trong Python bên cạnh đó còn cung cấp các cấu trúc dữ liệu mạnh mẽ như DataFrame, Series thuận tiện cho việc dễ dàng trực quan hóa.

**2.matplotlib.pyplot (plt):**

Tiếp theo em xin được phép giới thiệu về thư viện thứ hai mà em sử dụng đó là thư viện matplotlib là một thư viện vẽ đồ thị trong Python.

Pyplot là một module của matplotlib cung cấp các chức năng để tạo và hiển thị đồ thị bằng các dụng cụ như Tkinter, wxPython.

**3.seaborn (sns):**

Seaborn là một thư viện trực quan hóa dữ liệu xây trên matplotlib.

Cung cấp các giao diện cao cấp và mặc định giúp tạo ra các biểu đồ thống kê có trực quan hóa tốt.

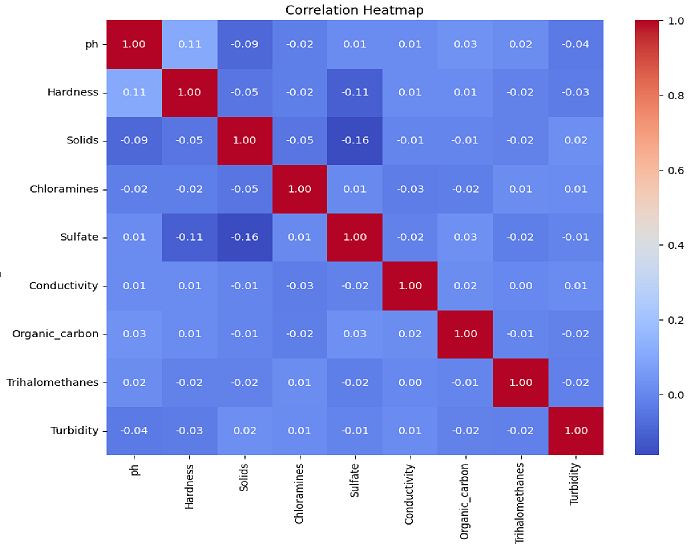
**4. plotly**

Plotly là một thư viện đồ họa tương tác, mã nguồn mở, và dựa trên nền tảng trình duyệt để tạo ra các biểu đồ tương tác, trực quan.

**CHƯƠNG 3 TRỰC QUAN HOÁ DỮ LIỆU**

**3.2. Một số ảnh trực quan hóa được phân tích trong dự án**

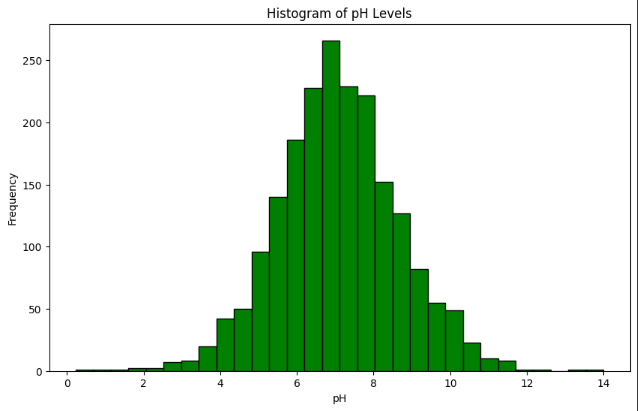
### **3.2.1 Heatmap hệ số tương quan giữa các đặc trưng chất lượng nước**

****

Hình 2: Heatmap hệ số tương quan giữa các đặc trưng chất lượng nước

Kết quả heatmap cho thấy các đặc trưng trong bộ dữ liệu có hệ số tương quan rất thấp (gần 0). Nghĩa là chúng hầu như không phụ thuộc lẫn nhau. Trong trường hợp này, nếu gộp các biến lại sẽ làm mất đi thông tin riêng biệt của từng đặc trưng. Do đó, em quyết định giữ nguyên tất cả các cột thay vì gộp, vì cách này đảm bảo phân tích chính xác và giữ đầy đủ thông tin cho mô hình dự đoán độ an toàn nước.

**3.2.2. Trực quan theo độ pH**

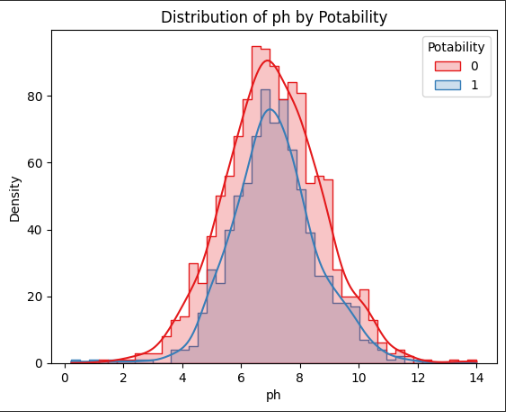


Hình 3: Biểu đồ biểu thị sự phân bố của mức độ pH

Nhìn vào biểu đồ, ta thấy rằng:

* Đa số các giá trị pH tập trung trong khoảng từ 6 đến 8, với đỉnh cao nhất rơi vào khoảng pH 7.
* Các giá trị pH thấp hơn (dưới 4) và cao hơn (trên 10) xuất hiện rất ít, biểu thị rằng những trường hợp có độ pH quá thấp hoặc quá cao là hiếm.
* Phân bố này có dạng gần như đối xứng, cho thấy rằng mức độ pH trong dữ liệu này có xu hướng tập trung xung quanh giá trị trung bình và giảm dần khi di chuyển ra xa khỏi trung tâm.

**3.2.3. Trực quan độ pH ảnh hưởng đến độ an toàn của nước**



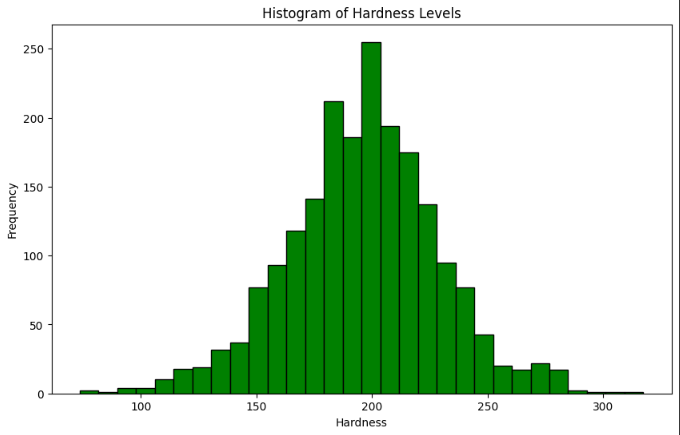
Hình 4: Trực quan độ pH ảnh hưởng đến độ an toàn của nước

Biểu đồ cho thấy sự phân bố độ pH của hai nhóm nước uống: an toàn (Potability = 1) và không an toàn (Potability = 0).

* **Nước không an toàn (màu đỏ):** Có pH chủ yếu dưới 6, cho thấy nước có thể bị axit hóa, không an toàn để uống.
* **Nước an toàn (màu xanh):** Có pH trong khoảng 6-8, là phạm vi pH thường thấy ở nước sạch và an toàn.

Biểu đồ này giúp phân tách nước an toàn và không an toàn theo giá trị pH, nhưng cần lưu ý rằng pH không phải là yếu tố duy nhất quyết định tính an toàn của nước. Các yếu tố khác như độ cứng, nồng độ clo, và các chất hóa học khác cũng cần được xem xét.

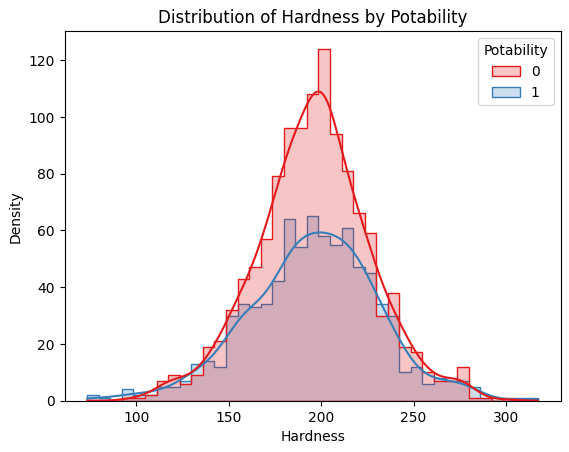
**3.2.4. Trực quan theo độ Hardness**



Hình 5: Trực quan hóa theo độ Hardness

Biểu đồ cho thấy độ cứng của nước chủ yếu nằm trong khoảng 150-250, với đỉnh tại khoảng 200. Mẫu nước có độ cứng dưới 100 và trên 250 rất ít, cho thấy nước thường có độ cứng trung bình, và phân bố này gần giống như phân bố chuẩn.

**3.2.5. Trực quan độ Hardness ảnh hưởng đến độ an toàn của nước**



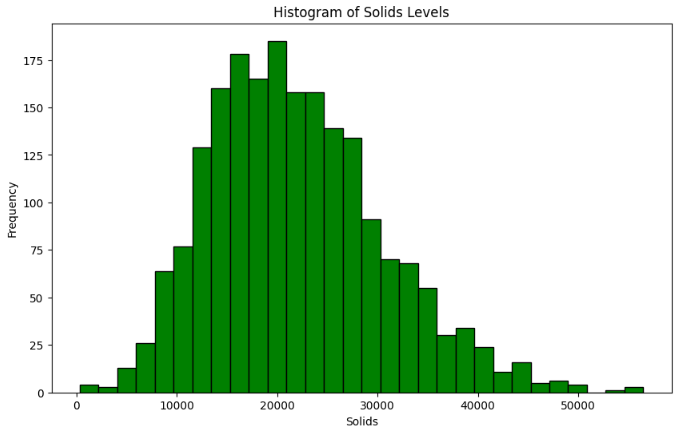
Hình 6: Trực quan hóa về độ Hardness ảnh hưởng đến độ an toàn của nước

Biểu đồ này cho thấy sự phân bố độ cứng nước (Hardness) theo hai nhóm: nước uống được (Potability = 1, màu xanh) và nước không uống được (Potability = 0, màu đỏ).

* **Nước không uống được (màu đỏ):** Có phân phối độ cứng rộng và đỉnh cao hơn ở mức độ cứng từ 180 đến 240, cho thấy rằng nước không an toàn có thể chứa nhiều khoáng chất hòa tan hơn.
* **Nước uống được (màu xanh):** Có độ cứng chủ yếu nằm trong khoảng từ 150 đến 220, và phân phối mượt mà hơn, cho thấy nước an toàn thường có độ cứng thấp hơn và ổn định hơn.

**Nhận xét:** Có sự chồng lấn, nhưng nước không uống được có phân bố rộng hơn, đặc biệt ở các giá trị cực đoan, ngụ ý độ cứng quá cao hoặc quá thấp thường không an toàn.

**3.2.6. Trực quan Solids**



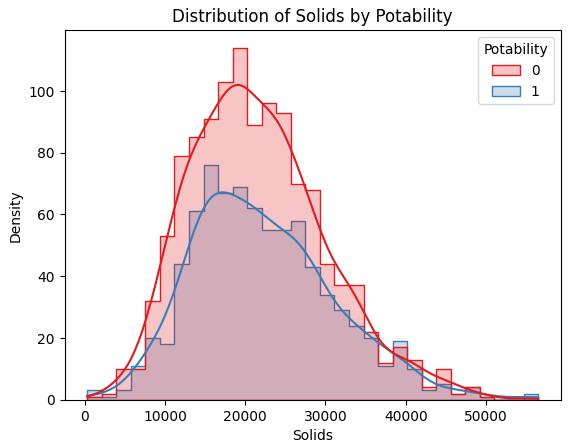
Hình 7: Trực quan hoá Solids

Biểu đồ này hiển thị sự phân bố hàm lượng chất rắn (Solids) trong mẫu nước.

* **Phân bố:** Phần lớn các mẫu nước có hàm lượng chất rắn nằm trong khoảng từ 10,000 đến 30,000 mg/L, với đỉnh cao nhất ở khoảng 20,000 mg/L.
* **Dạng phân bố:** Phân bố này có dạng gần như chuẩn, với số lượng mẫu giảm dần ở cả hai phía của đỉnh phân bố.
* **Mẫu cực trị:** Có rất ít mẫu nước có hàm lượng chất rắn thấp hơn 5,000 mg/L hoặc cao hơn 40,000 mg/L.

Nhìn chung, hàm lượng chất rắn trong nước thường tập trung ở mức trung bình, với các giá trị cực đoan là khá hiếm.

**3.2.7. Trực quan theo Solids ảnh hưởng đến độ an toàn của nước**



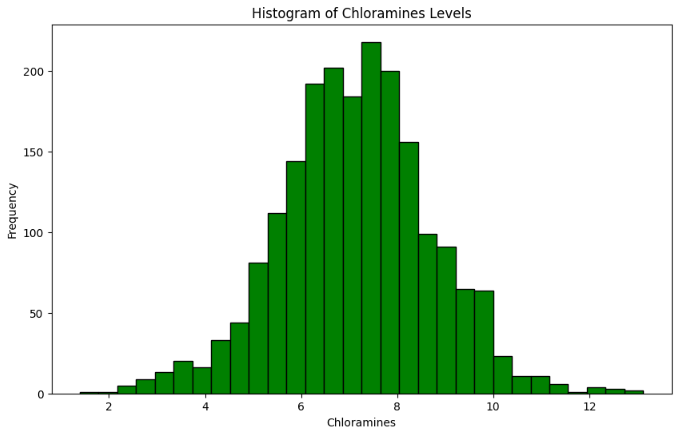
Hình 8: Trực quan theo Solids ảnh hưởng đến độ an toàn của nước

Biểu đồ này cho thấy sự phân bố hàm lượng chất rắn giữa hai nhóm nước: uống được (màu xanh) và không uống được (màu đỏ).

* **Nước không uống được:** Tập trung quanh 20,000 mg/L với mật độ cao hơn trong khoảng 10,000-30,000 mg/L..
* **Nước uống được:** Cũng tập trung quanh 20,000 mg/L, nhưng có mật độ thấp hơn và phân bố hẹp hơn.

**Nhận xét:** Nước có hàm lượng chất rắn cao hơn thường không an toàn để uống.

**3.2.8. Trực quan theo Chloramines**



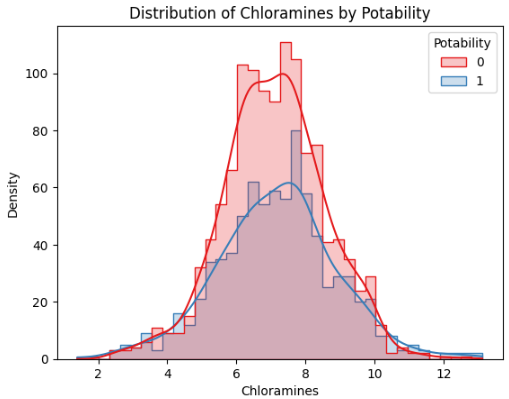
Hình 9: Trực quan hóa theo Chloramines

Biểu đồ này hiển thị sự phân bố mức độ Chloramines trong mẫu nước.

* **Phân bố:** Mức Chloramines chủ yếu nằm trong khoảng từ 4 đến 10 mg/L, với đỉnh cao nhất ở khoảng 7-8 mg/L..
* **Dạng phân bố:** Phân bố có dạng gần chuẩn, với số lượng mẫu giảm dần ở hai đầu của dải giá trị.
* **Mẫu cực trị:** Ít mẫu có mức Chloramines thấp hơn 2 mg/L hoặc cao hơn 12 mg/L.

Tóm lại, mức Chloramines trong nước thường tập trung quanh mức trung bình, với các giá trị cực đoan là khá hiếm.

**3.2.9. Trực quan theo Chloramines ảnh hưởng đến độ an toàn của nước**



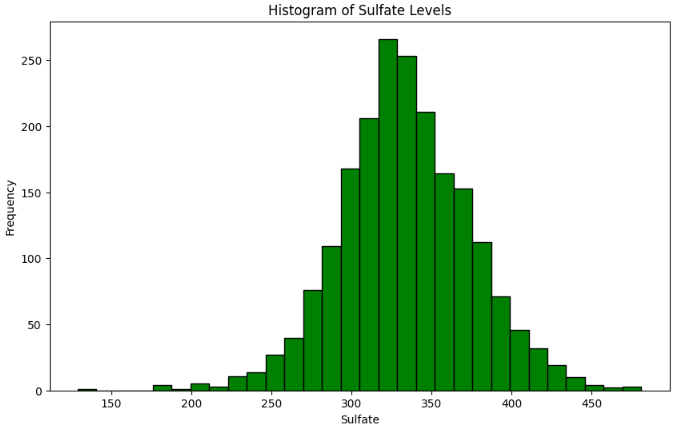
Hình 10: Trực quan theo Chloramines ảnh hưởng đến độ an toàn của nước

Biểu đồ này cho thấy sự phân bố mức Chloramines giữa hai nhóm nước: uống được (màu xanh) và không uống được (màu đỏ).

* **Nước không uống được:** Mật độ cao nhất tập trung quanh mức 7-8 mg/L, với phân bố rộng hơn.
* **Nước uống được:** Tập trung chủ yếu từ 6-8 mg/L, nhưng với mật độ thấp hơn và phân bố hẹp hơn.

**Nhận xét:** Nước không uống được có xu hướng xuất hiện ở cả hai phía của khoảng 6-8 mg/L với mật độ cao hơn, cho thấy mức Chloramines cao hoặc thấp đều có thể liên quan đến nước không an toàn.

**3.2.10. Trực quan theo Sulfate**



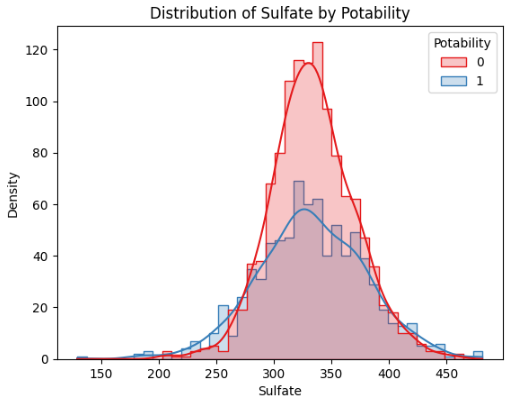
Hình 11: Trực quan theo Sulfate

Biểu đồ này cho thấy sự phân bố mức Sulfate trong nước.

* **Phân bố:** Mức Sulfate chủ yếu nằm trong khoảng 250-400 mg/L, với đỉnh cao nhất quanh 325 mg/L.
* **Dạng phân bố:** Phân bố gần chuẩn, với số lượng mẫu giảm dần ở cả hai phía của đỉnh.
* **Mẫu cực trị:** Ít mẫu có mức Sulfate dưới 200 mg/L hoặc trên 450 mg/L.

Tóm lại, mức Sulfate trong nước tập trung quanh mức trung bình, với các giá trị cực đoan là khá hiếm.

**3.2.11. Trực quan theo Sulfate ảnh hưởng đến độ an toàn của nước**



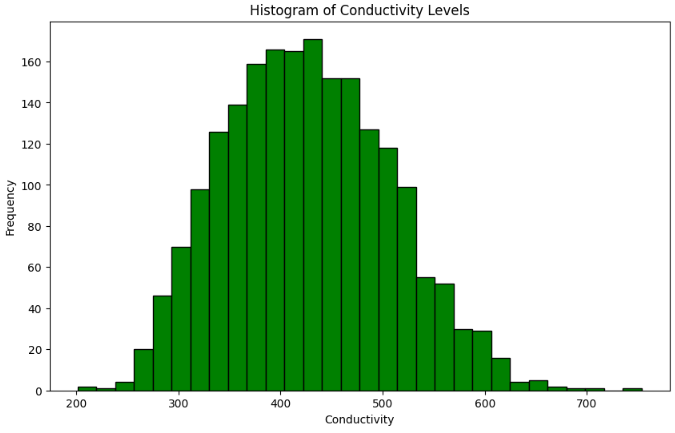
Hình 12: Trực quan theo Sulfate ảnh hưởng đến độ an toàn của nước

Biểu đồ này thể hiện sự phân bố mức Sulfate giữa hai nhóm nước: uống được (màu xanh) và không uống được (màu đỏ).

* **Nước không an toàn (đường màu đỏ) có phân phối Sulfate rộng hơn, chủ yếu nằm trong khoảng từ 280 đến 340, với đỉnh ở khoảng 300.**
* **Nước an toàn (đường màu xanh dương) có phân phối Sulfate hẹp hơn, chủ yếu nằm trong khoảng từ 290 đến 320, cho thấy mức Sulfate vừa phải.**

**Nhận xét:** Mặc dù có sự phân tách giữa các nhóm, vẫn có sự chồng lấp giữa hai nhóm ở khoảng từ 290 đến 320, điều này cho thấy rằng Sulfate không phải là yếu tố duy nhất quyết định tính an toàn của nước. Các yếu tố khác như pH, Chloramines, và độ cứng (Hardness) cũng cần được xem xét khi đánh giá tính an toàn của nước.

**3.2.12. Trực quan theo Conductivity**



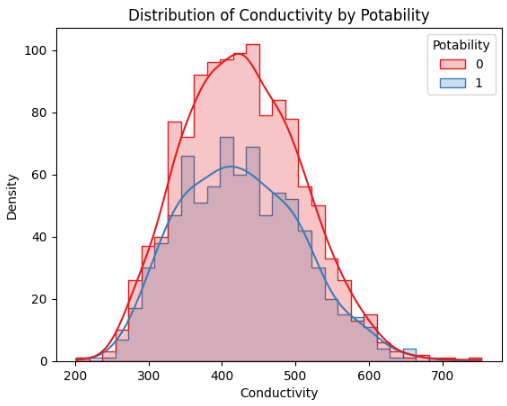
Hình 13: Trực quan theo Conductivity

Biểu đồ này cho thấy sự phân bố của độ dẫn điện trong nước.

* **Phân bố:** Độ dẫn điện tập trung chủ yếu từ 300-600 µS/cm, với đỉnh cao nhất ở khoảng 400-450 µS/cm.
* **Dạng phân bố:** Gần với phân bố chuẩn, số mẫu giảm dần khi đi xa khỏi đỉnh.
* **Mẫu cực trị:** Rất ít mẫu có độ dẫn điện dưới 200 µS/cm hoặc trên 700 µS/cm.

Tóm lại, độ dẫn điện thường nằm quanh mức trung bình, với các giá trị cực đoan xuất hiện hiếm hoi.

**3.2.13. Trực quan theo Conductivity ảnh hưởng đến độ an toàn của nước**



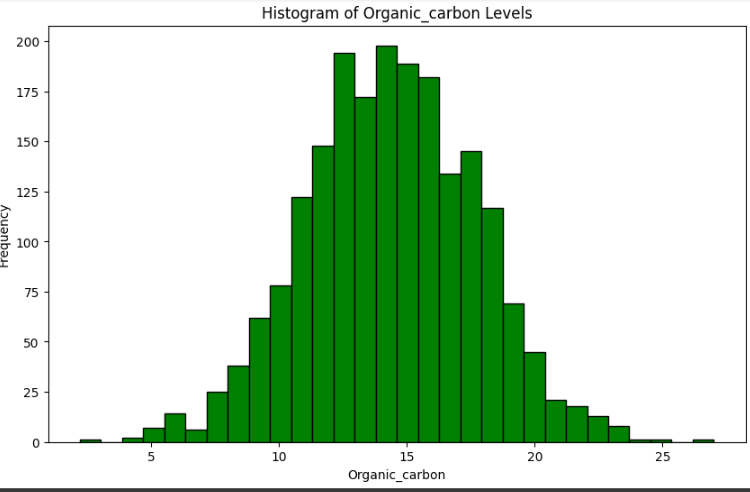
Hình 14: Trực quan theo Conductivity ảnh hưởng đến độ an toàn của nước

Biểu đồ này thể hiện sự phân bố độ dẫn điện (Conductivity) của nước giữa hai nhóm: nước uống được (màu xanh) và nước không uống được (màu đỏ).

* **Nước không uống được:** Tập trung nhiều quanh mức 400-450 µS/cm, với phân bố rộng hơn.
* **Nước uống được:** Tập trung chủ yếu từ 350-500 µS/cm, nhưng có mật độ thấp hơn và phân bố hẹp hơn.

**Nhận xét:** Nước không uống được thường có độ dẫn điện cao hơn, đặc biệt trong khoảng 400-450 µS/cm, cho thấy mức độ dẫn điện cao có thể liên quan đến nước không an toàn.

**3.2.14. Trực quan theo Organic\_carbon**



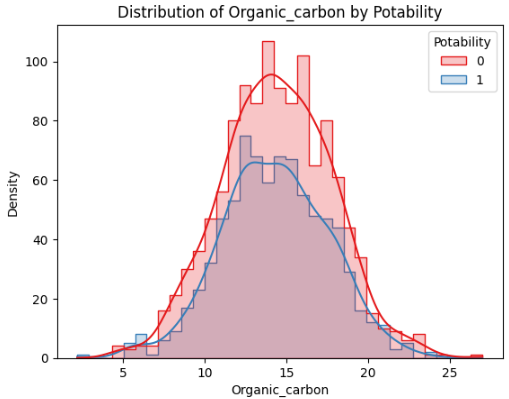
Hình 15: Trực quan theo Organic\_carbon

Biểu đồ này hiển thị sự phân bố mức độ carbon hữu cơ (Organic Carbon) trong nước.

* **Phân bố:** Mức độ carbon hữu cơ chủ yếu nằm trong khoảng 10-20 mg/L, với đỉnh cao nhất ở khoảng 15 mg/L.
* **Dạng phân bố:** Phân bố gần chuẩn, với số lượng mẫu giảm dần ở cả hai phía của đỉnh.
* **Mẫu cực trị:** Ít mẫu có mức carbon hữu cơ dưới 5 mg/L hoặc trên 25 mg/L.

Tóm lại, mức độ carbon hữu cơ trong nước thường tập trung quanh mức trung bình, với các giá trị cực đoan xuất hiện hiếm hoi.

**3.2.15. Trực quan theo Organic\_carbon ảnh hưởng đến độ an toàn của nước**



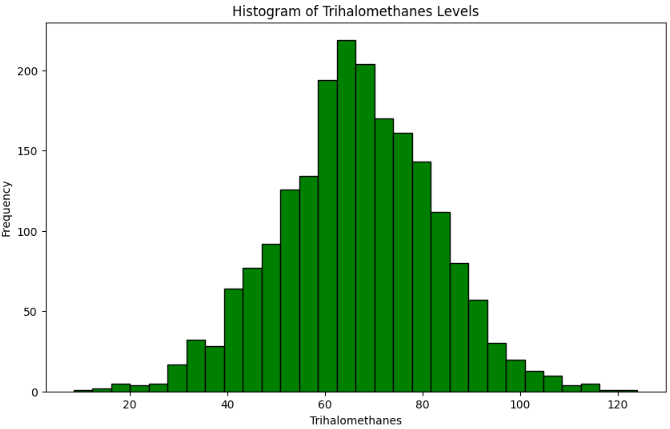
Hình 16: Trực quan theo Organic\_carbon ảnh hưởng đến độ an toàn của nước

Biểu đồ này cho thấy sự phân bố mức độ carbon hữu cơ (Organic Carbon) giữa hai nhóm nước: uống được (màu xanh) và không uống được (màu đỏ).

* **Nước không uống được:** Tập trung chủ yếu quanh mức 15 mg/L, với phân bố rộng hơn.
* **Nước uống được:** Cũng tập trung ở khoảng 15 mg/L, nhưng có mật độ thấp hơn và phân bố hẹp hơn.

**Nhận xét:** Nước không uống được thường có mức độ carbon hữu cơ cao hơn, đặc biệt trong khoảng 10-20 mg/L, cho thấy mức carbon hữu cơ cao có thể liên quan đến nước không an toàn.

**3.2.16. Trực quan theo Trihalomethanes**



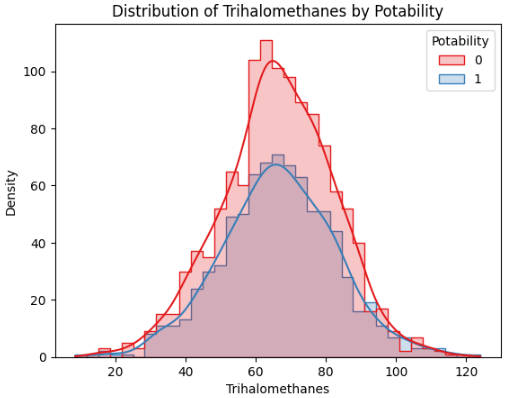
Hình 17: Trực quan theo Trihalomethanes

Biểu đồ này cho thấy sự phân bố mức độ Trihalomethanes trong nước.

* **Phân bố:** Mức Trihalomethanes chủ yếu nằm trong khoảng 40-80 µg/L, với đỉnh cao nhất ở khoảng 60 µg/L.
* **Dạng phân bố:** Phân bố gần chuẩn, với số lượng mẫu giảm dần ở cả hai phía của đỉnh.
* **Mẫu cực trị:** Rất ít mẫu có mức Trihalomethanes dưới 20 µg/L hoặc trên 100 µg/L.

Tóm lại, mức Trihalomethanes trong nước thường tập trung quanh mức trung bình, với các giá trị cực đoan xuất hiện rất ít.

**3.2.17. Trực quan theo Trihalomethanes ảnh hưởng đến độ an toàn của nước**



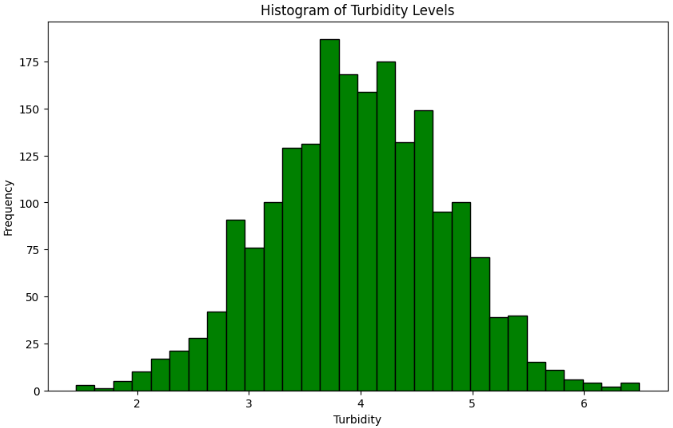
Hình 18: Trực quan theo Trihalomethanes ảnh hưởng đến độ an toàn của nước

Biểu đồ này cho thấy sự phân bố mức độ Trihalomethanes giữa hai nhóm nước: uống được (màu xanh) và không uống được (màu đỏ).

* **Nước không uống được:** Tập trung nhiều nhất quanh mức 60 µg/L, với phân bố rộng hơn.
* **Nước uống được:** Tập trung chủ yếu trong khoảng 50-70 µg/L, nhưng mật độ thấp hơn và phân bố hẹp hơn.

**Nhận xét:** Nước không uống được thường có mức Trihalomethanes cao hơn, đặc biệt quanh mức 60 µg/L, cho thấy mức Trihalomethanes cao có thể liên quan đến nước không an toàn.

**3.2.18. Trực quan theo Turbidity**



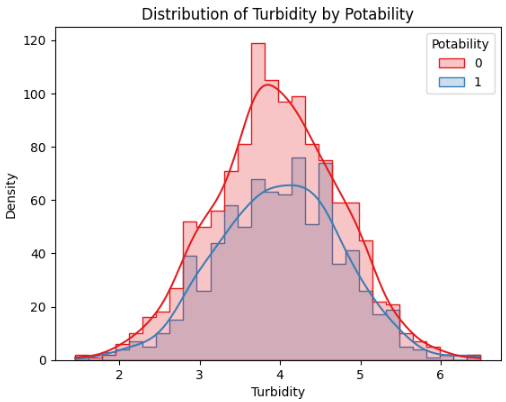
Hình 19: Trực quan theo Turbidity

Biểu đồ này hiển thị sự phân bố mức độ đục (Turbidity) trong nước.

* **Phân bố:** Mức độ đục chủ yếu nằm trong khoảng 3-5 NTU, với đỉnh cao nhất ở khoảng 4 NTU.
* **Dạng phân bố:** Phân bố gần chuẩn, với số lượng mẫu giảm dần ở cả hai phía của đỉnh.
* **Mẫu cực trị:** Rất ít mẫu có mức độ đục dưới 2 NTU hoặc trên 6 NTU.

Tóm lại, mức độ đục của nước thường tập trung quanh mức trung bình, với các giá trị cực đoan là khá hiếm.

**3.2.19. Trực quan theo Turbidity ảnh hưởng đến độ an toàn của nước**



Hình 20: Trực quan theo Turbidity ảnh hưởng đến độ an toàn của nước

Biểu đồ này cho thấy sự phân bố mức độ đục (Turbidity) giữa hai nhóm nước: uống được (màu xanh) và không uống được (màu đỏ).

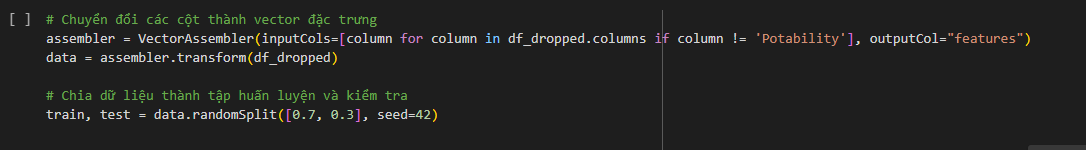
* **Nước không uống được:** Mật độ cao nhất tập trung quanh mức 4 NTU, với phân bố rộng hơn.
* **Nước uống được:** Cũng tập trung quanh mức 4 NTU, nhưng có mật độ thấp hơn và phân bố hẹp hơn.

**Nhận xét:** Nước không uống được thường có mức độ đục cao hơn, đặc biệt quanh mức 4 NTU, cho thấy mức độ đục cao có thể liên quan đến nước không an toàn.

# CHƯƠNG 4 HUẤN LUYỆN MÔ HÌNH

## 4.1 Xây dựng ứng dụng và giải thích

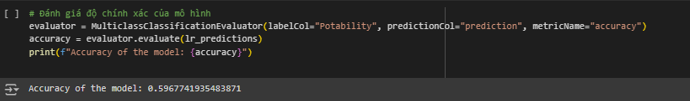
Trước khi đi vào các thuật toán dự đoán thì em xin phép giải thích một ít về phần dữ liệu mà em sẽ dự đoán như sau:



Hình 21: Dữ liệu chuẩn bị cho phần học máy

### **4.1.1. Logistic Regression trong Spark**

Spark MLlib triển khai Logistic Regression bằng cách tối ưu hóa hàm mất mát (log-loss) sử dụng thuật toán Gradient Descent. Mô hình sẽ được huấn luyện trên tập dữ liệu đã chuẩn hóa để tìm ra các trọng số tối ưu cho từng đặc trưng. Quá trình huấn luyện này được thực hiện trên toàn bộ cụm máy tính, tận dụng khả năng tính toán phân tán của Spark để giảm thời gian xử lý.

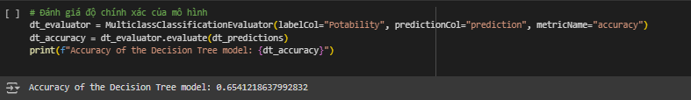


Hình 22: Kết quả mô hình Logistic Regreession trong Spark

Như mọi người có thể thấy được ở phần kết quả của bài dự đoán này thì dự đoán của mô hình Logistic Regreesion chiếm 0.597

### **4.1.2. Decision Tree Classifier**

Trong Spark MLlib, Decision Tree được xây dựng bằng cách sử dụng kỹ thuật đệ quy để tạo ra các phân chia dựa trên các đặc trưng dữ liệu. Cây quyết định sẽ tiếp tục phân chia cho đến khi đạt đến một ngưỡng cụ thể như độ sâu tối đa hoặc số lượng mẫu tối thiểu tại mỗi nút. Spark phân tán quá trình xây dựng cây này trên nhiều nút để tăng tốc độ xử lý.

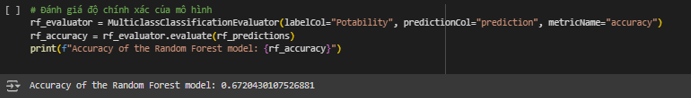


Hình 23: Kết quả mô hình Decision Tree Classifier trong Spark

Như mọi người có thể thấy được ở phần kết quả của bài dự đoán này thì dự đoán của mô hình Decision Tree Classifier trong Spark là mấy cũng chiếm 0.654

### **4.1.3. Random Forest**

Spark MLlib triển khai Random Forest bằng cách xây dựng các cây quyết định một cách song song, tận dụng khả năng xử lý phân tán của Spark. Người dùng có thể tùy chỉnh số lượng cây trong rừng (numTrees), độ sâu tối đa của mỗi cây (maxDepth), và các tham số khác để tối ưu hóa hiệu suất của mô hình

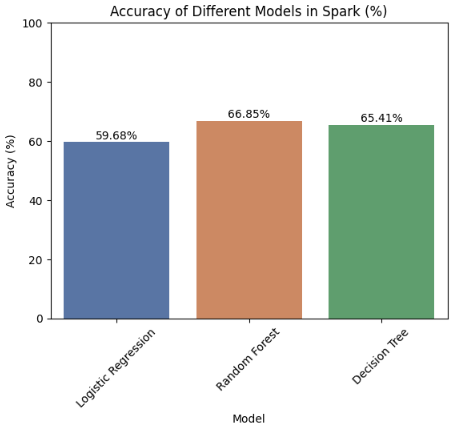


Hình 24: Kết quả mô hình Random Forest

Như mọi người có thể thấy được ở phần kết quả của bài dự đoán này thì dự đoán của mô hình Random Forest trong Spark là mấy cũng chiếm 0.6685

**4.1.4. So sánh các phương pháp dự đoán**

Từ 3 phương pháp dự đoán bao gồm: Logistic Regression, Decision Tree Classifier, Random Forest. Ta có thể thấy được rằng mô hình Random Forest là mô hình dự đoán tốt nhất.



Hình 25: Trực quan hóa của phần trăm của các dự báo

# KẾT LUẬN

## Kết quả đạt được.

Kết quả đạt được từ mô hình phân loại nước uống sử dụng Apache Spark đã mang lại nhiều lợi ích quan trọng trong quản lý chất lượng nước. Thông qua mô hình này, chúng em đã thu được một số kết quả đáng chú ý:

* **Cải thiện độ chính xác trong phân loại nước uống:** Mô hình giúp phân loại nước uống một cách chính xác dựa trên các chỉ số hóa lý, từ đó hỗ trợ các cơ sở sản xuất và xử lý nước nâng cao chất lượng sản phẩm, đảm bảo nước cung cấp an toàn cho người tiêu dùng.
* **Hiệu quả của khoa học dữ liệu trong quản lý nước:** Dự án đã chứng minh rằng các kỹ thuật học máy và xử lý dữ liệu lớn như Apache Spark có thể được ứng dụng thành công trong quản lý chất lượng nước. Qua việc phân tích và trực quan hóa dữ liệu, các thông tin quan trọng có thể được cung cấp cho nhà quản lý và chuyên gia để giám sát và kiểm tra chất lượng nước.
* **Phát triển công cụ dự báo và phòng ngừa:** Mô hình này có khả năng dự báo chất lượng nước dựa trên các thông số như pH, độ cứng, hàm lượng chất rắn, v.v. Điều này cho phép các cơ quan quản lý can thiệp kịp thời để ngăn chặn nguy cơ cung cấp nước không đạt chuẩn đến người dân.
* **Lợi ích đối với cộng đồng:** Việc áp dụng mô hình dự đoán giúp đảm bảo rằng nước uống đạt tiêu chuẩn an toàn, góp phần bảo vệ sức khỏe cộng đồng. Các biện pháp can thiệp kịp thời có thể giảm thiểu nguy cơ mắc bệnh liên quan đến nước uống không an toàn.

Những kết quả này không chỉ giúp tối ưu hóa quy trình sản xuất và quản lý chất lượng nước mà còn mang lại lợi ích lâu dài trong việc bảo vệ sức khỏe của cộng đồng.

## Hạn chế và hướng phát triển .

**Hạn chế**

Bên cạnh những thuận lợi thì cũng có những hạn chế như: Vì thời gian làm dự án này khá ngắn nên bọn em không thể tự tổng hợp cũng như raw được bộ dữ liệu không ưng ý nên không thể crawl dự liệu như mong muốn dẫn đến bộ dữ liệu không đủ lớn và còn khá thiếu sót dẫn đến tập dữ liệu không đầy đủ và đa dạng từ đó dẫn đến việc mô hình không tổng quát hoá tốt trên các trường hợp mới hoặc không đưa ra dự đoán chính xác đó là những hạn chế về bộ dữ liệu này sau khi bọn em phân tích.

**Hướng phát triển**

Nếu có nhiều thời gian hơn bọn em sẽ mở rộng và cải thiện dữ liệu: Nghiên cứu có thể tìm kiếm và tích hợp thêm dữ liệu từ nhiều nguồn khác nhau để nó phong phú hơn và cải thiện khả năng dự đoán của mô hình.

Tinh chỉnh Mô hình: Cần tiếp tục tinh chỉnh các tham số của mô hình, thử nghiệm với nhiều cấu hình khác nhau để tìm ra mô hình tốt nhất và tránh overfitting hoặc underfitting.

Đó là những hạn chế cũng như hướng phát triển thêm mà nhóm em có thể nghĩ ra được rất mong được quý thầy cô xem xét và cho ý kiến thêm.

Em chân thành cảm ơn!

# TÀI LIỆU THAM KHẢO

[1] DATA MINING: CONCEPTS AND TECHNIQUES 3RD EDITION

[2] Wenqiang Feang(2021) , Learning Apache Spark With Python

[3] Nguyễn Thanh Tuấn(2009), Deeplearning cơ bản The Legrand Orange Book Template by Mathias Legrand is used : <https://nttuan8.com/sach-deep-learning-co-ban/>

[4] Lecturer in deep learning at the Faculty of Information Technology