|  |
| --- |
| **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**  **BIGDATA**  **SỬ DỤNG SPARK DỰ ĐOÁN LOẠI NƯỚC UỐNG**  Giảng viên giảng dạy: ThS PHẠM ĐÌNH TÀI  Sinh viên thực hiện: LÊ XUÂN TRỌNG  MSSV : 2100009483  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : BIGDATA  Khóa : 2021    **Tp.HCM, tháng 09 Năm 2024** |
| **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**  **BIGDATA**  **SỬ DỤNG SPARK DỰ ĐOÁN LOẠI NƯỚC UỐNG**  Giảng viên giảng dạy: ThS PHẠM ĐÌNH TÀI  Sinh viên thực hiện: LÊ XUÂN TRỌNG  MSSV : 2100009483  Chuyên ngành : KHOA HỌC DỮ LIỆU  Môn học : BIGDATA  Khóa : 2021    **Tp.HCM, tháng 09 Năm 2024** |

|  |  |
| --- | --- |
| 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: SỬ DỤNG SPARK DỰ ĐOÁN LOẠI NƯỚC UỐNG

Giảng viên hướng dẫn: **ThS.Phạm Đình Tài**

Thời gian thực hiện:  **24 /06 /2024 đến 05/ 09 /2024**

Nhiệm vụ/nội dung (mô tả chi tiết nội dung, yêu cầu, phương pháp):

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

*TP.HCM, ngày 04 tháng 9 năm 2024*

|  |  |
| --- | --- |
| **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 Phạm Đình Tài giảng viên phụ trách môn bigdata đã giúp cũng như hỗ trợ chúng em hoàn thành được đồ án về chủ đề sử dụng Spark để phân loại nước uống. 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: Bigdata Lớp học phần:21DHT2D

Nhóm sinh viên thực hiện:

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

2. Lê Hoài Ân Tham gia đóng góp:100%

3. Tăng Quốc Trung Tham gia đóng góp:100%

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: 04/09/2024 Phòng thi: L.511

Đề tài tiểu luận/báo cáo của sinh viên : Dự án về sử dụng Spark để phân loại nước uống

|  |  |  |  |
| --- | --- | --- | --- |
| **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 28 tháng 12 năm 2022*

**Giảng viên giảng dạy**

(Ký tên và ghi rõ họ tên)

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 1](#_Toc22404)

[1.1 Giới thiệu đề tài. 1](#_Toc19814)

[1.2.Lý do chọn đề tài 2](#_Toc17418)

[1.3.Mục tiêu đề tài. 3](#_Toc23860)

[1.4.Phương pháp đề tài. 4](#_Toc18596)

[1.5. Đối tượng và phạm vi nghiên cứu. 5](#_Toc5881)

[1.5.1. Đối Tượng Nghiên Cứu 6](#_Toc6589)

[1.5.2. Phạm Vi Nghiên Cứu 6](#_Toc10519)

[CHƯƠNG 2 ỨNG DỤNG VÀ THUẬT TOÁN 7](#_Toc20807)

[2.1 Giới thiệu bài toán. 7](#_Toc15977)

[2.2 Mô tả thuật toán. 8](#_Toc19614)

[2.2.1. Thuật toán Logistic Regression 8](#_Toc19745)

[2.2.2. Thuật toán Random Forest 9](#_Toc1637)

[2.2.3. Thuật toán Decision Tree: 9](#_Toc13732)

[2.2.4. Quy trình tổng quan sử dụng các thuật toán trong Spark 10](#_Toc13427)

[2.3. Xây dựng bộ dữ liệu. 10](#_Toc31056)

[2.3.1. Bộ dữ liệu gồm các cột  11](#_Toc29198)

[2.4. Áp dụng thuật toán vào bài toán. 12](#_Toc20548)

[2.5. Thực nghiệm với thư viện. 12](#_Toc16640)

[CHƯƠNG 3 XÂY DỰNG ỨNG DỤNG NGÔN NGỮ PYTHON 14](#_Toc24233)

[3.1 Xây dựng ứng dụng và giải thích 14](#_Toc16455)

[3.1.1. Logistic Regression trong Spark 14](#_Toc3344)

[3.1.2. Decision Tree Classifier 15](#_Toc11642)

[3.1.3. Random Forest 16](#_Toc25818)

[3.1.4. So sanh các phương pháp dự đoán 17](#_Toc11796)

[3.2. Một số ảnh trực phân hóa được phân tích trong dự án 18](#_Toc6342)

[3.2.1. Trực quan theo độ pH 18](#_Toc2110)

[3.2.2. Trực quan độ pH ảnh hưởng đến độ an toàn của nước 19](#_Toc25374)

[3.2.3. Trực quan theo độ Hardness 20](#_Toc28374)

[3.2.4. Trực quan độ Hardness ảnh hưởng đến độ an toàn của nước 21](#_Toc28184)

[3.2.5. Trực quan Solids 22](#_Toc25283)

[3.2.6. Trực quan theo Solids ảnh hưởng đến độ an toàn của nước 23](#_Toc24673)

[3.2.7. Trực quan theo Chloramines 24](#_Toc31703)

[3.2.8. Trực quan theo Chloramines ảnh hưởng đến độ an toàn của nước 25](#_Toc14068)

[3.2.9. Trực quan theo Sulfate 26](#_Toc15821)

[3.2.10. Trực quan theo Sulfate ảnh hưởng đến độ an toàn của nước 27](#_Toc7845)

[3.2.11. Trực quan theo Conductivity 28](#_Toc548)

[3.2.12. Trực quan theo Conductivity ảnh hưởng đến độ an toàn của nước 29](#_Toc21256)

[3.2.13. Trực quan theo Organic\_carbon 30](#_Toc1508)

[3.2.14. Trực quan theo Organic\_carbon ảnh hưởng đến độ an toàn của nước 31](#_Toc30116)

[3.2.15. Trực quan theo Trihalomethanes 32](#_Toc25243)

[3.2.16. Trực quan theo Trihalomethanes ảnh hưởng đến độ an toàn của nước 33](#_Toc10118)

[3.2.17. Trực quan theo Turbidity 34](#_Toc13317)

[3.2.18. Trực quan theo Turbidity ảnh hưởng đến độ an toàn của nước 35](#_Toc16566)

[KẾT LUẬN 37](#_Toc30788)

[Kết quả đạt được. 37](#_Toc7204)

[Hạn chế và hướng phát triển . 37](#_Toc26263)

[TÀI LIỆU THAM KHẢO 39](#_Toc25590)

**DANH MỤC HÌNH**

[Hình 1 : Bảng dữ liệu 11](#_Toc2580)

[Hình 2 : Các thư viện đã được sử dụng 13](#_Toc4839)

[Hình 3 : Dữ liệu chuẩn bị cho phần học máy 14](#_Toc12390)

[Hình 4 : Khởi tạo mô hình Logistic Regression trong Spark 14](#_Toc10496)

[Hình 5 : Kết quả mô hình Logistic Regreession trong Spark 15](#_Toc826)

[Hình 6 : Khởi tạo mô hình Decision Tree Classifier trong Spark 15](#_Toc2795)

[Hình 7 : Kết quả mô hình Decision Tree Classifier trong Spark 16](#_Toc9588)

[Hình 8 : Khởi tạo Random Forest 16](#_Toc5307)

[Hình 9 : Kết quả mô hình Random Forest 17](#_Toc1775)

[Hình 10 : So sanh các phương pháp dự đoán 17](#_Toc340)

[Hình 11 : Trực quan hóa của phần trăm sai lệch của các dự báo 18](#_Toc26294)

[Hình 12 : Biểu đồ biểu thị sự phân bố của mức độ pH 19](#_Toc30209)

[Hình 13 : Trực quan độ pH ảnh hưởng đến độ an toàn của nước 20](#_Toc4980)

[Hình 14 : Trực quan hóa theo độ Hardness 21](#_Toc18269)

[Hình 15 : Trực quan hóa về độ Hardness ảnh hưởng đến độ an toàn của nước 22](#_Toc26376)

[Hình 16 : Trực quan hoá Solids 23](#_Toc21181)

[Hình 17 : Trực quan theo Solids ảnh hưởng đến độ an toàn của nước 24](#_Toc29375)

[Hình 18 : Trực quan hóa theo Chloramines 25](#_Toc17761)

[Hình 19 : Trực quan theo Chloramines ảnh hưởng đến độ an toàn của nước 26](#_Toc20100)

[Hình 20 : Trực quan theo Sulfate 27](#_Toc31803)

[Hình 21 : Trực quan theo Sulfate ảnh hưởng đến độ an toàn của nước 28](#_Toc29038)

[Hình 22 : Trực quan theo Conductivity 29](#_Toc28334)

[Hình 23 : Trực quan theo Conductivity ảnh hưởng đến độ an toàn của nước 30](#_Toc4767)

[Hình 24 : Trực quan theo Organic\_carbon 31](#_Toc13975)

[Hình 25 : Trực quan theo Organic\_carbon ảnh hưởng đến độ an toàn của nước 32](#_Toc4460)

[Hình 26 : Trực quan theo Trihalomethanes 33](#_Toc22922)

[Hình 27 : Trực quan theo Trihalomethanes ảnh hưởng đến độ an toàn của nước 34](#_Toc10430)

[Hình 28 : Trực quan theo Turbidity 35](#_Toc14174)

[Hình 29 : Trực quan theo Turbidity ảnh hưởng đến độ an toàn của nước 36](#_Toc15042)

# 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 loại nước uống sử dụng Apache Spark” 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

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 "Sử dụng spark dự đoán loại nước uống" 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.Phương pháp đề tài.

Có một số phương pháp tiếp cận đề tài "Sử dụng spark dự đoán loại nước uống" mà em sử dụng để đạt được mục tiêu của nó

* **Thu thập và tiền xử lý dữ liệu:** Dữ liệu sẽ được thu thập từ nhiều nguồn khác nhau như các nhà máy sản xuất, các nghiên cứu thị trường, hoặc từ các thiết bị cảm biến được sử dụng trong quy trình sản xuất. Dữ liệu này bao gồm thông tin về thành phần nguyên liệu, quy trình sản xuất, các đặc tính hóa lý như độ pH, độ ngọt, màu sắc, và các thông số khác liên quan đến sản phẩm. Giai đoạn chuẩn bị dữ liệu là rất quan trọng, bao gồm việc làm sạch dữ liệu, xử lý các giá trị thiếu, và chuẩn hóa dữ liệu để chuẩn bị cho việc phân tích.
* **Phân tích dữ liệu và mô hình hóa:** Các phương pháp học máy như hồi quy logistic, support vector machines, decision trees, random forests, neural networks, và các mô hình dự đoán khác sẽ được sử dụng 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:** Các phương pháp như cross-validation, confusion matrix, và ROC curve sẽ được sử dụng để đánh giá hiệu suất của mô hình. Việc tối ưu hóa mô hình sẽ bao gồm điều chỉnh các siêu tham số và áp dụng kỹ thuật feature selection để nâng cao độ chính xác và hiệu quả của mô hình.
* **Xác định các đặc tính quan trọng:** Sử dụng các phương pháp như feature importance, SHAP values, và permutation importance để xác định những đặc trưng có tác động lớn nhất đến việc phân loại và dự đoán loại nước uống.
* **Triển khai và ứng dụng:** Sau khi phát triển được mô hình phân loại tối ưu, mô hình này sẽ được sử dụng để phân loại các sản phẩm mới, cung cấp thông tin quan trọng hỗ trợ cho quá trình sản xuất và quản lý chất lượng sản phẩm.

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.

## 1.5. 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:

### **1.5.1. 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.

### **1.5.2. 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.
* **Thời gian nghiên cứu:** Quá trình nghiên cứu sẽ diễn ra trong một khoảng thời gian đủ dài để thu thập dữ liệu, phát triển mô hình và đánh giá kết quả. Thời gian này có thể kéo dài từ 6 tháng đến 1 năm, tùy thuộc vào tốc độ thu thập và xử lý dữ liệu, cũng như tiến độ phát triển mô hình.

# CHƯƠNG 2 ỨNG DỤNG VÀ THUẬT TOÁN

## 2.1 Giới thiệu bài toán.

Trong bối cảnh hiện đại, ngành công nghiệp thực phẩm và đồ uống đang phát triển với tốc độ chóng mặt, đòi hỏi các doanh nghiệp phải thích ứng với nhu cầu đa dạng và không ngừng thay đổi của người tiêu dùng. Điều này đặt ra thách thức lớn trong việc tối ưu hóa quy trình sản xuất, phân loại sản phẩm một cách chính xác, và đáp ứng kịp thời những xu hướng mới của thị trường. Để vượt qua những thách thức này, việc áp dụng các công nghệ tiên tiến như học máy (Machine Learning) và xử lý dữ liệu lớn (Big Data) đã trở thành một xu hướng không thể bỏ qua.

Bài toán "Sử dụng Spark để dự đoán loại nước uống" ra đời từ nhu cầu thực tế của doanh nghiệp trong việc phân loại và dự đoán các loại nước uống dựa trên đặc điểm cụ thể của từng sản phẩm. Mục tiêu là xây dựng một hệ thống có khả năng tự động phân loại và dự đoán loại nước uống dựa trên các yếu tố như thành phần, hương vị, giá trị dinh dưỡng và các đặc trưng liên quan khác. Hệ thống này sẽ hỗ trợ doanh nghiệp trong việc tối ưu hóa quy trình sản xuất, nâng cao chất lượng sản phẩm, và cá nhân hóa trải nghiệm khách hàng.

Apache Spark, một nền tảng xử lý dữ liệu lớn có hiệu suất cao, được chọn để triển khai bài toán này. Spark không chỉ cung cấp khả năng xử lý dữ liệu nhanh chóng mà còn tích hợp các công cụ học máy thông qua thư viện MLlib, giúp việc xây dựng và triển khai các mô hình dự đoán trở nên hiệu quả hơn. Trong bài toán này, Spark sẽ được sử dụng để xử lý các tập dữ liệu lớn liên quan đến nước uống, từ việc làm sạch và chuẩn hóa dữ liệu đến việc xây dựng, huấn luyện và đánh giá các mô hình học máy.

Cụ thể, các thuật toán học máy như Logistic Regression, Decision Tree, Random Forest và Naive Bayes sẽ được áp dụng để phân loại và dự đoán loại nước uống. Những thuật toán này sẽ được triển khai trên nền tảng Spark, tận dụng khả năng xử lý dữ liệu phân tán để cải thiện thời gian và hiệu suất.

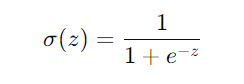
## 2.2 Mô tả thuật toán.

Đề tài "Sử dụng Spark để dự đoán loại nước uống" 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.

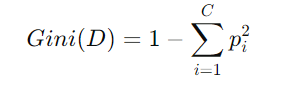
**Huấn luyện mô hình trong Spark:** Spark MLlib triển khai Random Forest bằng cách xây dựng các cây quyết định một cách song song, tận dụng khả năng xử lý phân tán của Spark. Người dùng có thể tùy chỉnh số lượng cây trong rừng (numTrees), độ sâu tối đa của mỗi cây (maxDepth), và các tham số khác để tối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

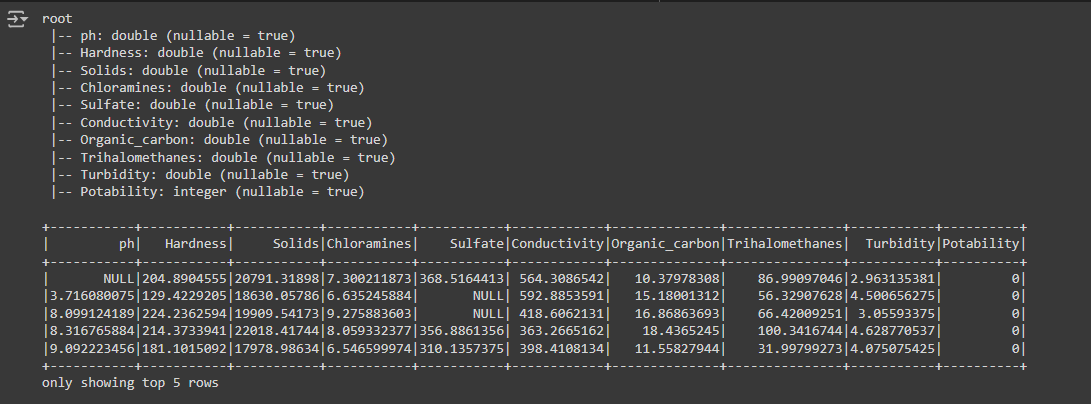
**2.2.4. Quy trình tổng quan sử dụng các thuật toán trong Spark**

* **Chuẩn bị dữ liệu:** Dữ liệu được thu thập, làm sạch, và chuẩn hóa để loại bỏ các giá trị ngoại lai, xử lý dữ liệu thiếu, và 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, và Naive Bayes 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.
* **Tối ưu hóa mô hình:** Quá trình tối ưu hóa như Grid Search và Cross-Validation sẽ được áp dụng để tìm các tham số tốt nhất cho mô hình, đảm bảo độ chính xác cao mà không bị overfitting.

**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**

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. Áp dụng thuật toán vào bài toán.

Trong đề tài “Sử dụng Spark để phân loại nước uống”, có nhiều thuật toán machine learning và deep learning có thể được áp dụng để xây dựng mô hình dự đoán. Dưới đây là một số thuật toán mà em đã áp dụng trong dự án này:

**Logistic Regression:** Đây là một thuật toán phân loại phổ biến, cho phép dự đoán các kết quả nhị phân dựa trên các biến đầu vào. Với đề tài này, hồi quy logistic có thể giúp ước tính xác suất phát sinh bệnh tiểu đường dựa trên các yếu tố nguy cơ cụ thể.

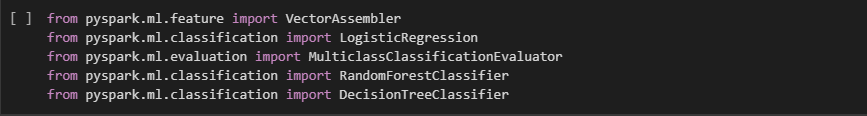
**Decision Trees:** Decision Trees là một kỹ thuật phân loại rất linh hoạt, đặc biệt hiệu quả trong việc xử lý dữ liệu không tuyến tính. Phương pháp này phân chia dữ liệu thành các phân đoạn nhỏ hơn dựa trên những quy tắc được rút ra từ dữ liệu đầu vào. Decision Trees dễ dàng xử lý các biến đầu vào đa chiều và có tính minh bạch cao, giúp dễ dàng diễn giải các kết quả thu được.

**Random Forests:** Random Forests là một thuật toán học máy mạnh mẽ, được tạo thành từ việc kết hợp nhiều Decision Trees lại với nhau. Phương pháp này giúp hạn chế hiện tượng overfitting bằng cách tổng hợp kết quả từ nhiều cây khác nhau, tạo ra một mô hình có độ ổn định và độ tin cậy cao hơn. Random Forests cũng xử lý tốt các dữ liệu không tuyến tính và có thể làm việc hiệu quả với các biến đầu vào phức tạp.

## 2.5. 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.





**Hình 2****: Các thư viện đã được sử dụng**

**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.

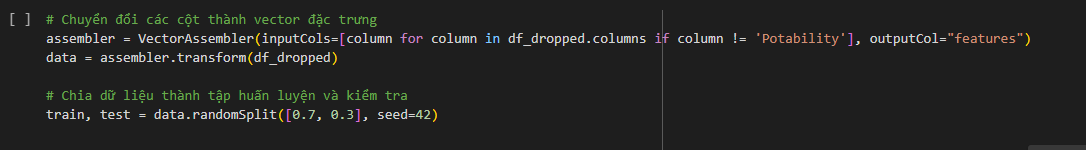
**4. plotly**

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 XÂY DỰNG ỨNG DỤNG NGÔN NGỮ PYTHON

## 3.1 Xây dựng ứng dụng và giải thích

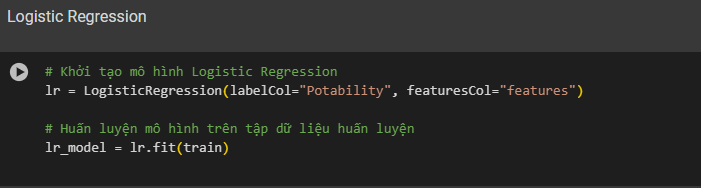
Trước khi đi vào các thuật toán dự đoán thì em xin phép giải thích một ít về phần dữ liệu mà em sẽ dự đoán như sau:



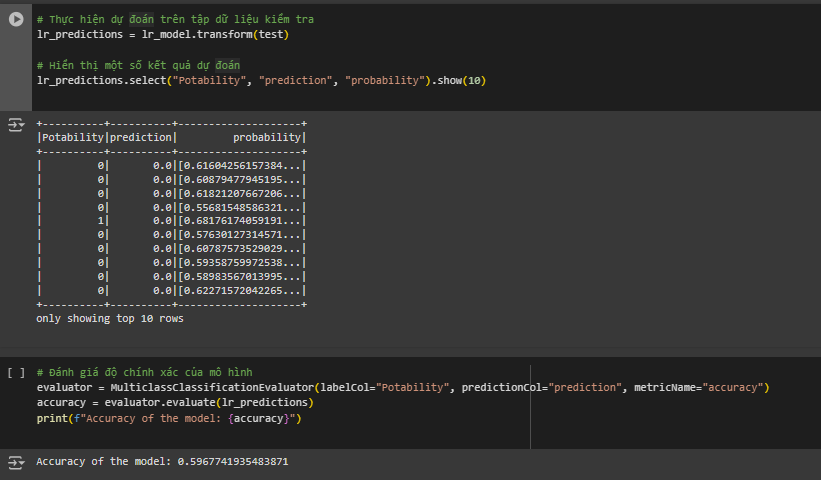
**Hình 3****: Dữ liệu chuẩn bị cho phần học máy**

### **3.1.1. 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 4****: Khởi tạo mô hình Logistic Regression trong Spark**

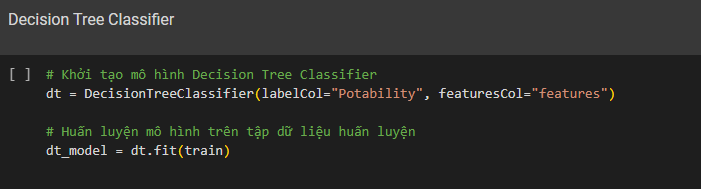


**Hình 5****: Kết quả mô hình Logistic Regreession trong Spark**

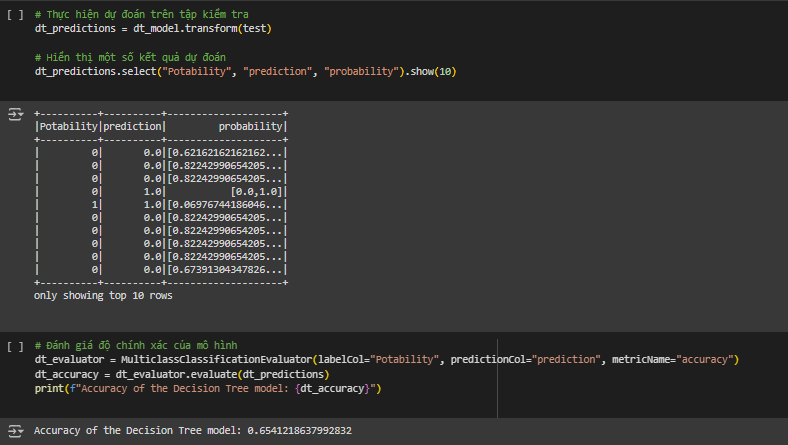
Như mọi người có thể thấy được ở phần kết quả của bài dự đoán này thì dự đoán của mô hình Logistic Regreesion chiếm 0.597

### **3.1.2. 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 6****: Khởi tạo mô hình Decision Tree Classifier trong Spark**

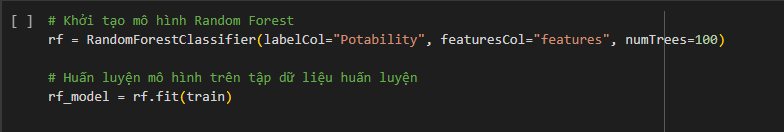


**Hình 7****: Kết quả mô hình Decision Tree Classifier trong Spark**

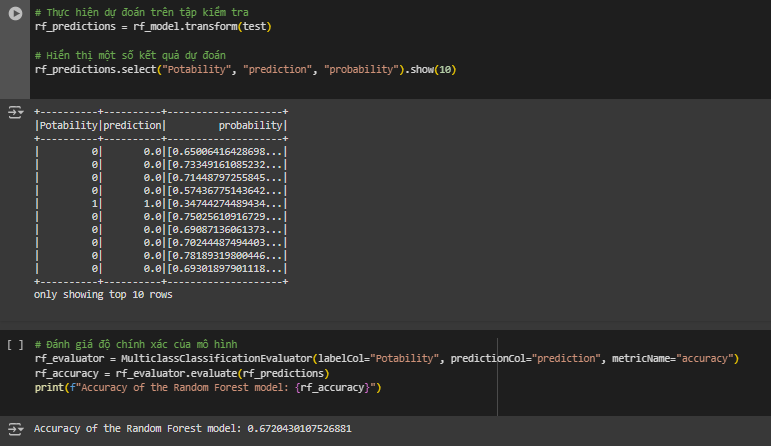
Như mọi người có thể thấy được ở phần kết quả của bài dự đoán này thì dự đoán của mô hình Decision Tree Classifier trong Spark là mấy cũng chiếm 0.654

### **3.1.3. 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



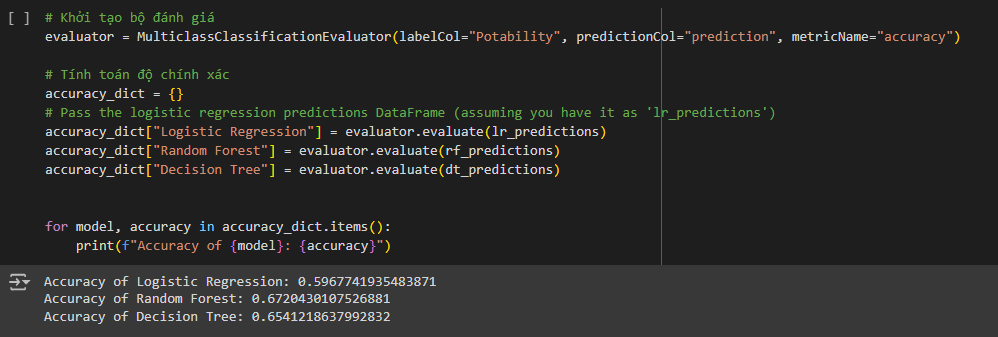
**Hình 8****: Khởi tạo Random Forest**



**Hình 9****: Kết quả mô hình Random Forest**

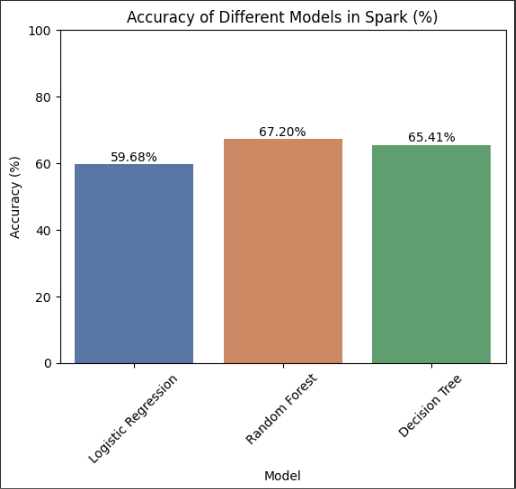
Như mọi người có thể thấy được ở phần kết quả của bài dự đoán này thì dự đoán của mô hình Random Forest trong Spark là mấy cũng chiếm 0.672

**3.1.4. So sanh các phương pháp dự đoán**



**Hình 10****: So sanh các phương pháp dự đoán**

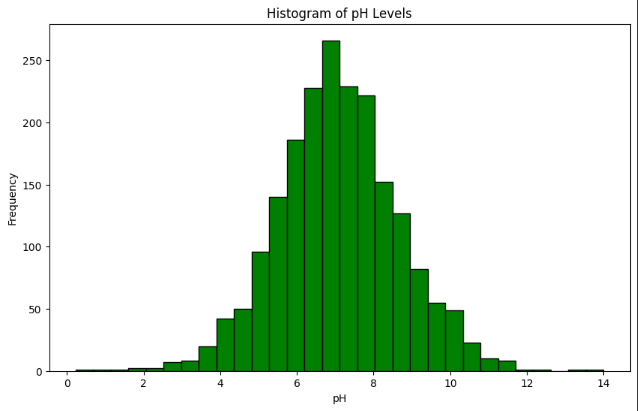
Từ 3 phương pháp dự đoán bao gồm: Logistic Regression, Decision Tree Classifier, Random Forest. Ta có thể thấy được rằng mô hình Logistic Regression là mô hình có độ dự báo sai lệch thấp nhất



**Hình 11****: Trực quan hóa của phần trăm sai lệch của các dự báo**

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

**3.2.1. Trực quan theo độ pH**

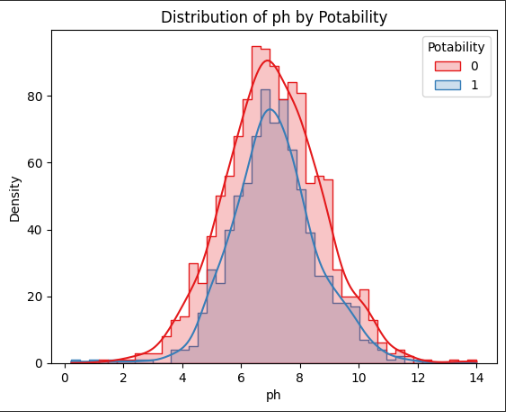


**Hình 12****: Biểu đồ biểu thị sự phân bố của mức độ pH**

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.

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



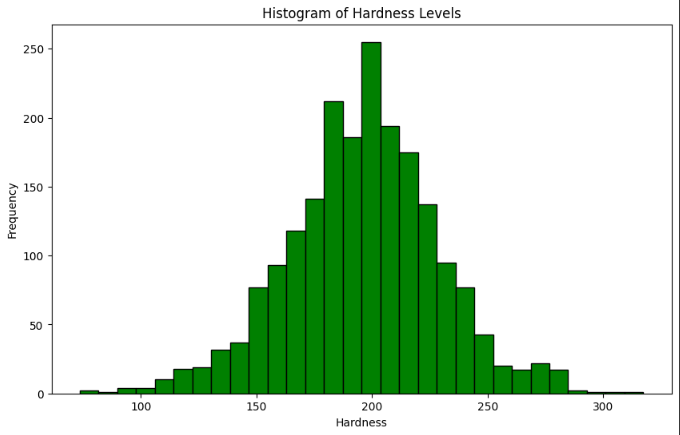
**Hình 13****: Trực quan độ pH ảnh hưởng đến độ an toàn của nước**

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):** Phân bố trải rộng từ pH 5 đến 9, với đỉnh ở khoảng pH 7, cho thấy nước không an toàn có thể có pH khá biến động.
* **Nước an toàn (màu xanh):** Phân bố tập trung hơn trong khoảng pH 6 đến 8, cũng với đỉnh ở pH 7, nhưng mật độ thấp hơn so với nước không an toàn ở hầu hết các giá trị pH.
* **So sánh:** Mức pH từ 6 đến 8 có sự chồng lấn giữa hai nhóm, nhưng nước không an toàn thường có pH nằm ngoài khoảng này

Nhìn chung, pH là một yếu tố quan trọng trong việc xác định độ an toàn của nước uống, nhưng không phải là yếu tố duy nhất.

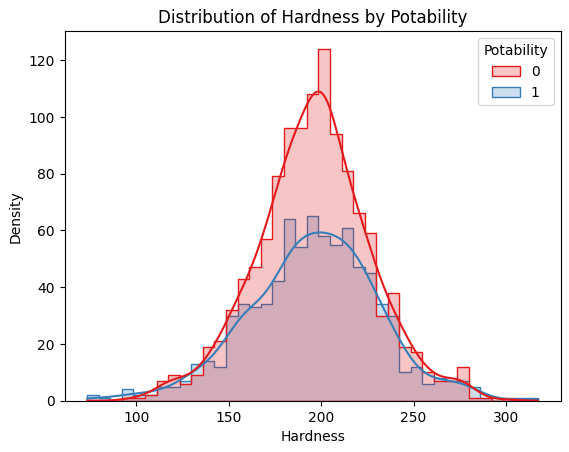
**3.2.3. Trực quan theo độ Hardness**



**Hình 14****: 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.

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



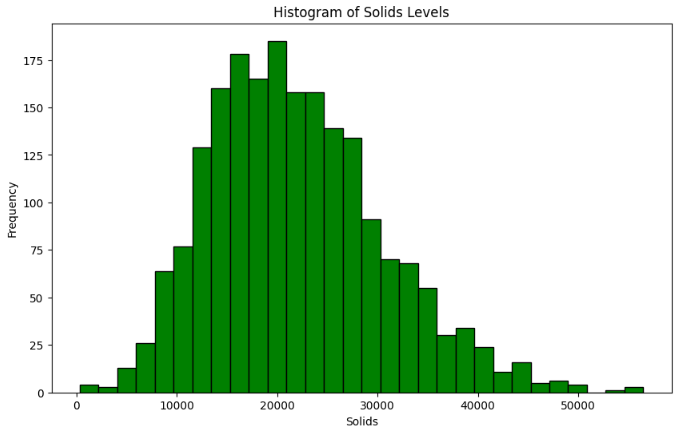
**Hình 15****: Trực quan hóa về độ Hardness ảnh hưởng đến độ an toàn của nước**

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):** Đỉnh phân bố cao nhất nằm ở khoảng 200, với phần lớn các mẫu nằm trong khoảng 150-250. Phân bố này rộng hơn, cho thấy sự đa dạng về độ cứng trong nước không uống được.
* **Nước uống được (màu xanh):** Đỉnh phân bố thấp hơn, cũng tập trung quanh mức 200. Phân bố này hẹp hơn, cho thấy độ cứng nước uống được thường ổn định hơn trong khoảng 150-250.

**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.

**3.2.5. Trực quan Solids**



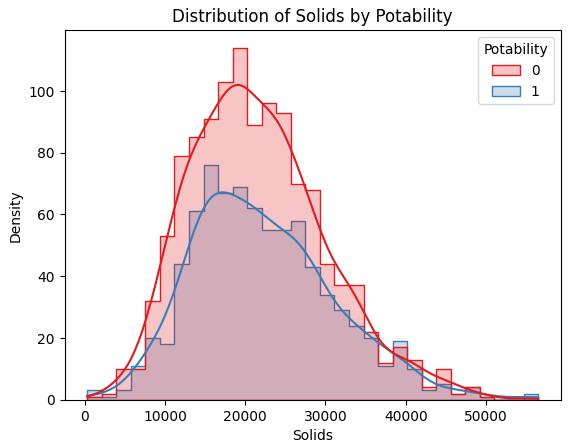
**Hình 16****: Trực quan hoá Solids**

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.

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



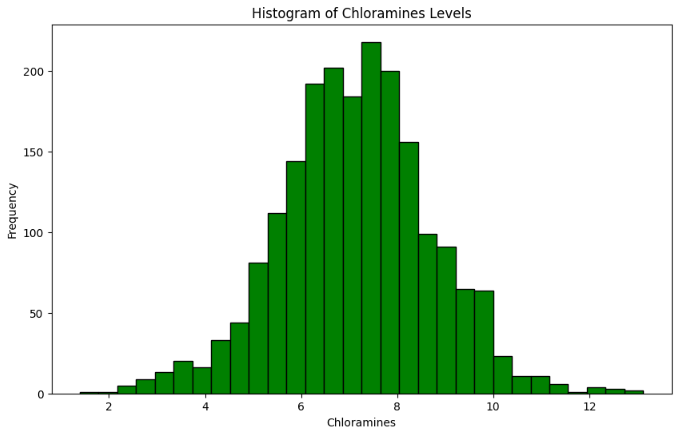
**Hình 17****: Trực quan theo Solids ảnh hưởng đến độ an toàn của nước**

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.

**3.2.7. Trực quan theo Chloramines**



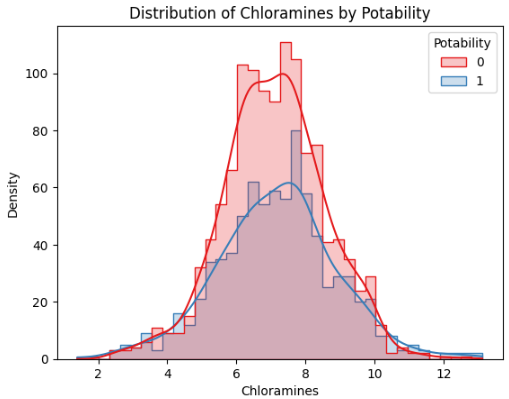
**Hình 18****: Trực quan hóa theo Chloramines**

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.

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



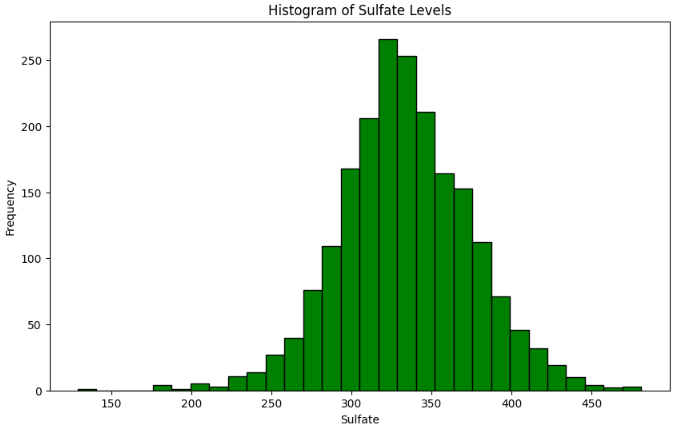
**Hình 19****: Trực quan theo Chloramines ảnh hưởng đến độ an toàn của nước**

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.

**3.2.9. Trực quan theo Sulfate**



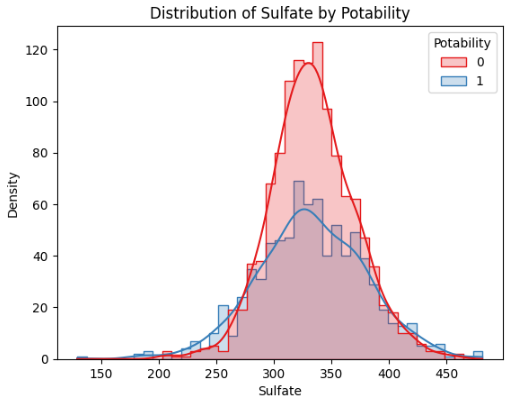
**Hình 20****: Trực quan theo Sulfate**

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.

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



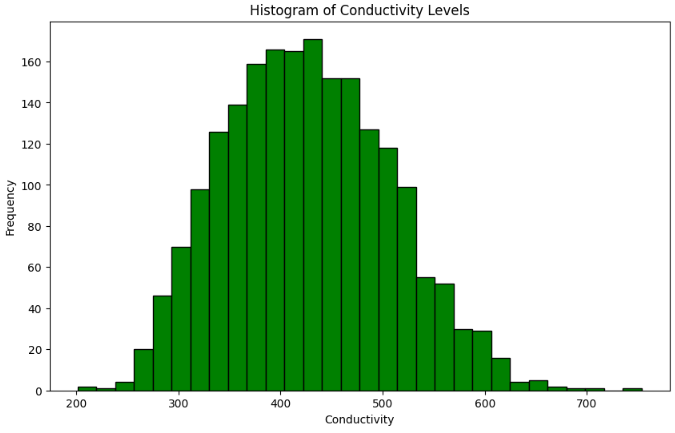
**Hình 21****: Trực quan theo Sulfate ảnh hưởng đến độ an toàn của nước**

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 uống được:** Mật độ cao nhất tập trung quanh mức 325 mg/L, với phân bố rộng hơn.
* **Nước uống được:** Tập trung chủ yếu ở khoảng 300-350 mg/L, nhưng mật độ thấp hơn và phân bố hẹp hơn.

**Nhận xét:** Nước không uống được có mật độ cao hơn ở các giá trị Sulfate từ 300-350 mg/L, cho thấy mức Sulfate cao hơn có thể liên quan đến nước không an toàn.

**3.2.11. Trực quan theo Conductivity**



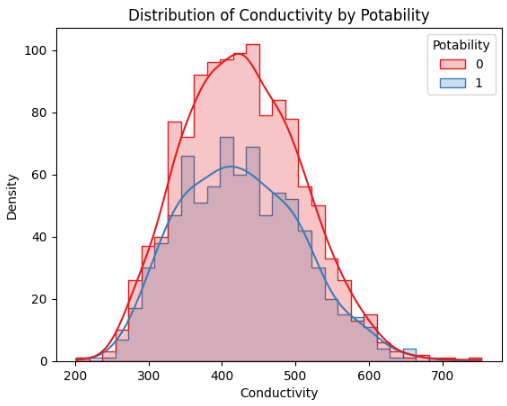
**Hình 22****: Trực quan theo Conductivity**

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.

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



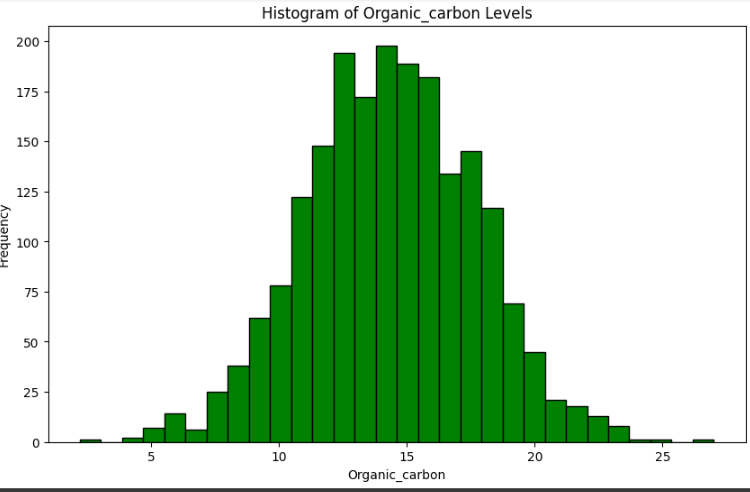
**Hình 23****: Trực quan theo Conductivity ảnh hưởng đến độ an toàn của nước**

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.

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



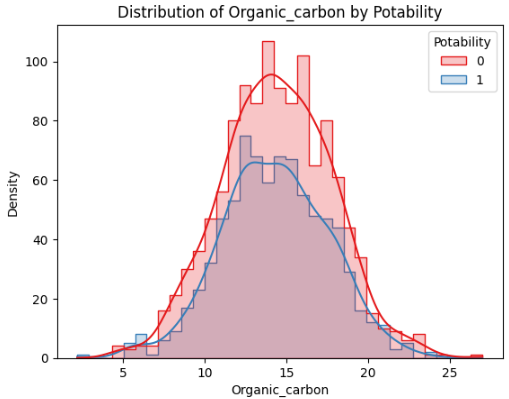
**Hình 24****: Trực quan theo Organic\_carbon**

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.

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



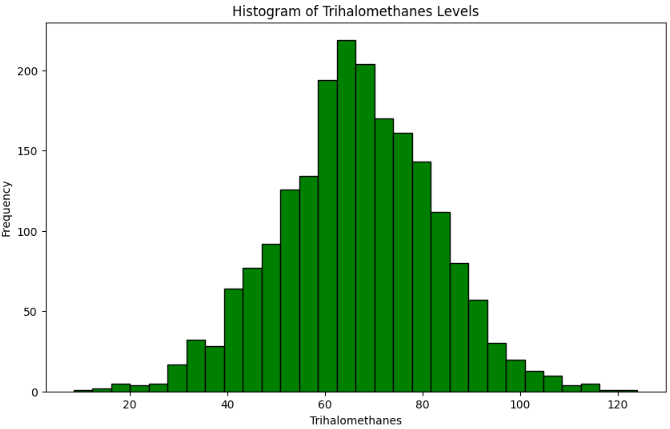
**Hình 25****: Trực quan theo Organic\_carbon ảnh hưởng đến độ an toàn của nước**

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.

**3.2.15. Trực quan theo Trihalomethanes**



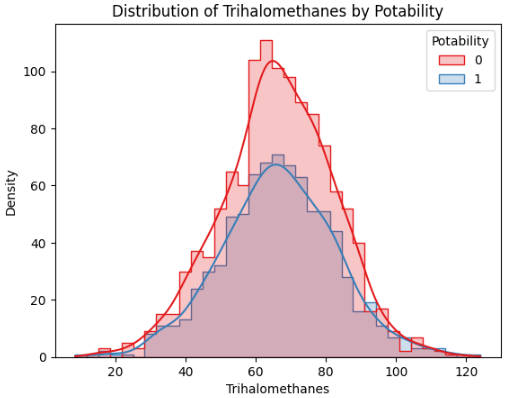
**Hình 26****: Trực quan theo Trihalomethanes**

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.

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



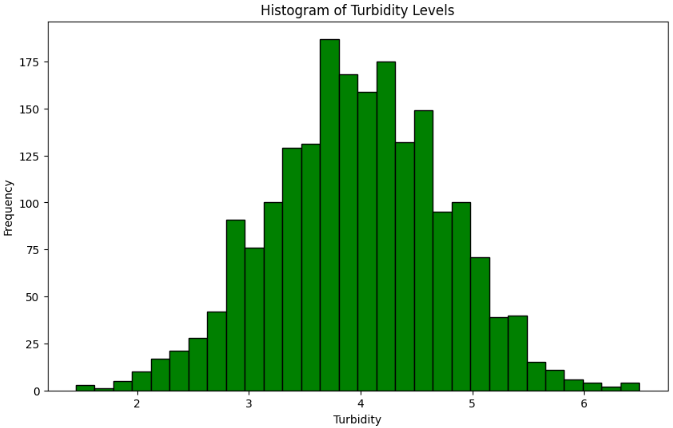
**Hình 27****: Trực quan theo Trihalomethanes ảnh hưởng đến độ an toàn của nước**

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.

**3.2.17. Trực quan theo Turbidity**



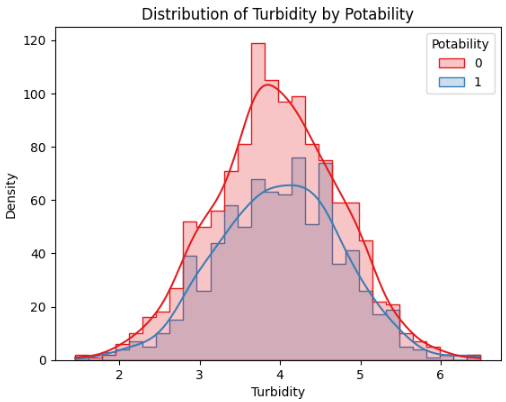
**Hình 28****: Trực quan theo Turbidity**

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.

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



**Hình 29****: 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.

# 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