**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN MÔN DỮ LIỆU LỚN**

**SỬ DỤNG SPARK ĐỂ PHÂN TÍCH**

**LOẠI NƯỚC UỐNG**

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

**Sinh viên thực hiện: CAO TRỌNG NGHĨA**

**MSSV: 2200007309**

**Khoá: 22**

**Ngành/ chuyên ngành: KHOA HỌC DỮ LIỆU**

Tp HCM, ngày 05 tháng 09 năm 2025

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**TIỂU LUẬN MÔN DỮ LIỆU LỚN**

**SỬ DỤNG SPARK ĐỂ PHÂN TÍCH**

**LOẠI NƯỚC UỐNG**

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

**Sinh viên thực hiện: CAO TRỌNG NGHĨA**

**MSSV: 2200007309**

**Khoá: 22**

**Ngành/ chuyên ngành: KHOA HỌC DỮ LIỆU**

Tp HCM, ngày 05 tháng 09 năm 2025

# LỜI MỞ ĐẦU

Nước là nguồn tài nguyên thiết yếu, đóng vai trò quan trọng trong duy trì sự sống và phát triển kinh tế, xã hội. Cùng với quá trình công nghiệp hóa và đô thị hóa nhanh chóng, tình trạng ô nhiễm nguồn nước ngày càng trở nên nghiêm trọng gây ảnh hưởng trực tiếp đến sức khỏe cộng đồng và hệ sinh thái. Theo tổ chức Y tế Thế giới (WHO) mỗi năm có hàng triệu người trên thế giới phải đối mặt với các bệnh tật liên quan đến việc sử dụng nguồn nước không đảm bảo chất lượng. Thực trạng này đặt ra yêu cầu cấp thiết trong việc kiểm tra, giám sát và dự đoán chất lượng nước một cách nhanh chóng, chính xác và tiết kiệm chi phí. Dưới sự phát triển của công nghệ dữ liệu lớn (Big Data) và học máy (Machine Learning) đã mở ra hướng đi mới cho lĩnh vực phân tích và dự đoán chất lượng nước. Đúng với tiêu chí đề bài ta có Apache Spark là một nền tảng tính toán phân tán mạnh mẽ cho phép xử lý khối lượng dữ liệu khổng lồ với tốc độ cao và đồng thời tích hợp các thuật toán học máy để xây dựng mô hình dự đoán hiệu quả. Việc ứng dụng Spark trong phân tích chất lượng nước không chỉ mang ý nghĩa học thuật mà còn có tiềm năng ứng dụng thực tiễn cao trong quản lý tài nguyên nước và bảo vệ sức khỏe cộng đồng.

Xuất phát từ những lý do trên, nhóm chúng em thực hiện đề tài *“Sử dụng Spark dự đoán nước uống”* dưới sự hướng dẫn tận tình của **thầy Phạm Đình Tài**. Đề tài tập trung vào việc thu thập và xử lý dữ liệu, áp dụng các thuật toán học máy trên nền tảng Spark để xây dựng mô hình phân loại, từ đó dự đoán khả năng một nguồn nước có thể sử dụng để uống hay không. Kết quả nghiên cứu kỳ vọng sẽ góp phần minh chứng cho tính khả thi và hiệu quả của việc ứng dụng công nghệ dữ liệu lớn trong giải quyết các vấn đề môi trường. Quá trình thực hiện của nhóm mặc dù nhóm đã cố gắng nỗ lực nhưng do hạn chế về thời gian và kiến thức nên không thể tránh khỏi những thiếu sót. Nhóm rất mong nhận được sự chỉ bảo và đóng góp ý kiến từ thầy cũng như các bạn để đề tài được hoàn thiện hơn.

# LỜI CẢM ƠN

Trong suốt quá trình học tập và thực hiện đề tài *“Sử dụng Spark dự đoán loại nước uống”* em đã nhận được rất nhiều sự quan tâm, giúp đỡ và chỉ dẫn quý báu. Trước hết, em xin gửi lời cảm ơn chân thành và sâu sắc đến **thầy Phạm Đình Tài,** thầy là người đã tận tình hướng dẫn, định hướng và luôn động viên em trong suốt thời gian nghiên cứu. Sự chỉ bảo tận tâm cùng những góp ý quý báu của thầy đã giúp em có thêm kiến thức, hoàn thiện kỹ năng và định hình rõ ràng hơn phương pháp tiếp cận vấn đề. Em cũng xin chân thành cảm ơn quý thầy cô trong **KHOA CÔNG NGHỆ THÔNG TIN – TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH** đã trang bị cho em những kiến thức nền tảng, tạo điều kiện thuận lợi để em có thể vận dụng vào quá trình nghiên cứu và hoàn thành báo cáo này. Đồng thời em muốn bày tỏ lòng biết ơn đến gia đình và bạn bè đã luôn ủng hộ, động viên và hỗ trợ em cả về tinh thần lẫn vật chất trong suốt quá trình học tập. Mặc dù đã nỗ lực hết sức, song do hạn chế về thời gian và kinh nghiệm, báo cáo khó tránh khỏi những thiếu sót, em rất mong nhận được những ý kiến đóng góp quý báu từ thầy cô và các bạn để đề tài được hoàn thiện hơn. Xin trân trọng cảm ơn!

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ 3 NĂM HỌC 2025 – 2026** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

Môn thi Dữ Liệu Lớn Lớp học phần: 22DTH34C

Nhóm sinh viên thực hiện: Nhóm 17

1. Bùi Văn Phúc Tham gia đóng góp: Thực hiện tất cả

2. Su Mỹ Cẩm Tham gia đóng góp: Thực hiện tất cả

3. Cao Trọng Nghĩa Tham gia đóng góp: Thực hiện tất cả

Ngày thi: 05/09/2025 Phòng thi: L.511

Đề tài tiểu luận/báo cáo của sinh viên: “Sử dụng Spark dự đoán loại nước uống”

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  |  |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  |  |  |
| * Lập luận |  |  |  |
| * Kết luận |  |  |  |
| Trình bày |  |  |  |
| **TỔNG ĐIỂM** |  |  |  |

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

**MỤC LỤC**

[LỜI MỞ ĐẦU i](#_Toc207640026)

[LỜI CẢM ƠN ii](#_Toc207640027)

[DANH MỤC HÌNH ẢNH vi](#_Toc207640028)

[KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT vii](#_Toc207640029)

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 1](#_Toc207640030)

[1.1.Giới thiệu chung về chất lượng nước uống 1](#_Toc207640031)

[1.2. Các nghiên cứu và phương pháp truyền thống 2](#_Toc207640032)

[1.3. Xu hướng ứng dụng dữ liệu lớn và học máy 3](#_Toc207640033)

[1.4. Apache Spark và khả năng ứng dụng trong phân tích nước 3](#_Toc207640034)

[1.5. Khoảng trống nghiên cứu và định hướng của đề tài 5](#_Toc207640035)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 6](#_Toc207640036)

[2.1. Các khái niệm cơ bản 6](#_Toc207640037)

[2.1.1. Chất lượng nước uống 6](#_Toc207640038)

[2.1.2. Chỉ số Chất lượng Nước (Water Quality Index – WQI) 7](#_Toc207640039)

[2.1.3. Dữ liệu chất lượng nước 7](#_Toc207640040)

[2.2. Nền tảng công nghệ và hệ thống sử dụng 8](#_Toc207640041)

[2.2.1. Apache Spark 8](#_Toc207640042)

[2.2.2. Pandas và NumPy 8](#_Toc207640043)

[2.2.3. Keras/TensorFlow 8](#_Toc207640044)

[2.2.4. Geopandas và Matplotlib/Seaborn 8](#_Toc207640045)

[2.3. Các bước giải quyết vấn đề 9](#_Toc207640046)

[2.5. Các ràng buộc và thách thức 11](#_Toc207640047)

[2.6. Giải pháp mới và hướng phát triển 12](#_Toc207640048)

[CHƯƠNG 3: MÔ HÌNH THỰC NGHIỆM 14](#_Toc207640049)

[3.1. Giới thiệu chương 14](#_Toc207640050)

[3.2.1. Thiết lập môi trường để cài đặt pyspark 14](#_Toc207640051)

[3.2.2. Đưa file dataset, file map vào Google Colab và tiến hành import các thư viện 15](#_Toc207640052)

[3.2.3. Tải dữ liệu lên khung Spark 17](#_Toc207640053)

[3.2.4. Làm sạch dữ liệu 18](#_Toc207640054)

[3.2.5. Trực quan hóa dữ liệu 19](#_Toc207640055)

[3.2.6. Kỹ thuật tính năng 20](#_Toc207640056)

[3.2.7. Mô hình Hồi quy Tuyến tính Không dựa trên Học sâu 24](#_Toc207640057)

[3.2.8. Mô hình Hồi quy Tuyến tính dựa trên Học sâu 27](#_Toc207640058)

[CHƯƠNG 4: KẾT LUẬN 30](#_Toc207640059)

[4.1. Kết luận 30](#_Toc207640060)

[4.2. Hạn chế và hướng phát triển 30](#_Toc207640061)

[TÀI LIỆU THAM KHẢO 32](#_Toc207640062)

# DANH MỤC HÌNH ẢNH

[Hình 1. Hình ảnh minh họa về chất lượng nước uống 1](#_Toc207639963)

[Hình 2. Spark 4](#_Toc207639964)

[Hình 3. Bảng tiêu chuẩn của nước uống 6](#_Toc207639965)

[Hình 4. Dataset đồ án 7](#_Toc207639966)

[Hình 5. Hồi quy tuyến tính (Linear Regression) 10](#_Toc207639967)

[Hình 6. Logistic Regression 11](#_Toc207639968)

[Hình 7. Xóa java 11 cài java 8 14](#_Toc207639969)

[Hình 8. Cài Pyspark 15](#_Toc207639970)

[Hình 9. Import các thư viện 15](#_Toc207639971)

[Hình 10. Tải dữ liệu lên khung Spark và xem trước 5 dòng đầu 17](#_Toc207639972)

[Hình 11. Chuyển đôi dữ liệu 18](#_Toc207639973)

[Hình 12. Trực quan hóa dữ liệu 19](#_Toc207639974)

[Hình 13. Biểu đồ trực quan hóa 19](#_Toc207639975)

[Hình 14. Khởi tạo dữ liệu 20](#_Toc207639976)

[Hình 15. Tính xếp hạng chất lượng nước 21](#_Toc207639977)

[Hình 16. Áp dụng công thức wpi 21](#_Toc207639978)

[Hình 17. Phân loại chất lượng nước 21](#_Toc207639979)

[Hình 18. Biểu đồ chất lượng nước tại Ấn Độ 22](#_Toc207639980)

[Hình 19. Biểu đồ thanh ngang chất lượng nước tại Ấn Độ 23](#_Toc207639981)

[Hình 20. Chuẩn hóa dữ liệu sử dụng Normalizer 24](#_Toc207639982)

[Hình 21. Nhập và đưa vào Pipeline 24](#_Toc207639983)

[Hình 22. Phân chia ngẫu nhiên 2 cột 25](#_Toc207639984)

[Hình 23.Kết quả đánh giá mô hình 25](#_Toc207639985)

[Hình 24. Mô hình mạng Nơ-ron và Keras 25](#_Toc207639986)

[Hình 25. Kết quả huấn luyện mô hình 26](#_Toc207639987)

[Hình 26. Kiểm tra hiệu suất của mô hình 27](#_Toc207639988)

[Hình 27. Đánh giá mô hình 28](#_Toc207639989)

[Hình 28. Dự đoán với giá trị thực tế 28](#_Toc207639990)

# KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT

|  |  |
| --- | --- |
| Chữ viết tắt | Ý nghĩa |
| ML | Machine Learning |
| KNN | K-Nearest Neighbors |
| SVM | Support Vector Machine |
| LSTM | Long Short Term-Memory |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## Giới thiệu chung về chất lượng nước uống

Nước uống là một trong những yếu tố quan trọng nhất quyết định trực tiếp đến sức khỏe và sự phát triển bền vững của con người. Theo Tổ chức Y tế Thế giới (WHO), mỗi năm có hàng triệu ca bệnh liên quan đến việc sử dụng nguồn nước ô nhiễm, đặc biệt tại các quốc gia đang phát triển, nơi mà hệ thống xử lý và giám sát chất lượng nước chưa hoàn thiện. Các vấn đề sức khỏe phổ biến có thể kể đến như tiêu chảy, bệnh đường ruột, ngộ độc kim loại nặng và thậm chí là ung thư do phơi nhiễm lâu dài với các chất độc hại trong nước. Các nghiên cứu và giám sát chất lượng nước uống không chỉ mang ý nghĩa khoa học mà còn mang tính xã hội, y tế và môi trường to lớn.

A person drinking water from a glass

AI-generated content may be incorrect.

Hình 1. Hình ảnh minh họa về chất lượng nước uống

Chất lượng nước uống thường được đánh giá dựa trên nhiều chỉ số hóa lý và sinh học. Các thông số cơ bản bao gồm độ pH, tổng chất rắn hòa tan, độ đục, oxy hòa tan, nồng độ nitrat, sunfat, clorua, độ cứng hay sự hiện diện của vi khuẩn và vi sinh vật gây bệnh. Việc phân tích các thông số này cung cấp cái nhìn tổng thể về mức độ an toàn của nguồn nước. Các tổ chức quốc tế như WHO hay Environmental Protection Agency gọi tắt là EPA của Mỹ đều đã công bố các tiêu chuẩn giới hạn tối đa cho từng loại chất trong nước uống nhằm đảm bảo sức khỏe cộng đồng. Vào những năm gần đây có nhiều bộ dữ liệu thực nghiệm về chất lượng nước đã được công bố rộng rãi như bộ Water Potability Dataset trên Kaggle năm 2020 giúp cộng đồng nghiên cứu dễ dàng tiếp cận và ứng dụng các phương pháp phân tích hiện đại.

## 1.2. Các nghiên cứu và phương pháp truyền thống

Trong nhiều thập kỷ qua việc đánh giá chất lượng nước uống chủ yếu dựa trên các phương pháp truyền thống dưa vào các phòng thí nghiệm tiến hành phân tích hóa học và sinh học trực tiếp trên mẫu nước để đo lường hàm lượng các chất. Đây là cách tiếp cận chuẩn mực giúp cung cấp dữ liệu chính xác và chi tiết. Nhưng nhược điểm của nó là chi phí cao, mất nhiều thời gian, đòi hỏi nhân lực chuyên môn và khó mở rộng khi cần theo dõi trên phạm vi rộng hoặc liên tục trong thời gian dài.

Một phương pháp phổ biến khác trong nghiên cứu và quản lý nước là chỉ số chất lượng nước. Chỉ số này được xây dựng bằng cách tổng hợp nhiều thông số đo lường thành một giá trị duy nhất giúp phản ánh mức độ phù hợp của nguồn nước cho mục đích sử dụng. Ưu điểm của chỉ số chất lượng nước là đơn giản, dễ diễn giải và thuận tiện cho công tác quản lý. Sargaonkar và Deshpande (2003) đã đề xuất công thức tính chỉ số chất lượng nước để đánh giá chất lượng nước có mặt tại Ấn Độ, sau đó phương pháp này được áp dụng rộng rãi tại nhiều quốc gia.

Theo thống kê truyền thống cũng đóng vai trò quan trọng trong phân tích dữ liệu nước. Các kỹ thuật như hồi quy tuyến tính, phân tích thành phần chính (PCA) và phân cụm (clustering) thường được dùng để giảm chiều dữ liệu, tìm ra các mẫu tiềm ẩn hoặc phân loại nguồn nước theo mức độ ô nhiễm. Pesce và Wunderlin năm 2000 là một trong những nghiên cứu tiêu biểu áp dụng phân tích đa biến để đánh giá và phân loại chất lượng các con sông ở Argentina. Tuy mang lại nhiều đóng góp quan trọng, các phương pháp này vẫn bị giới hạn khi dữ liệu ngày càng phức tạp và khối lượng ngày càng lớn.

## 1.3. Xu hướng ứng dụng dữ liệu lớn và học máy

Nhờ sự phát triển mạnh mẽ của ngành công nghệ thông tin và khoa học dữ liệu trong thập kỷ gần đây các phương pháp phân tích hiện đại dựa trên học máy (Machine Learning) và dữ liệu lớn (Big Data) đã trở thành xu hướng chủ đạo. Các nghiên cứu không chỉ dừng lại ở phân tích mô tả mà còn hướng đến dự đoán và phân loại tự động giúp nâng cao độ chính xác và khả năng ứng dụng thực tế. Trong học máy có nhiều thuật toán đã được áp dụng thành công cho bài toán phân loại chất lượng nước. Các mô hình như Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM) và Neural Network cho phép dự đoán “tính an toàn” của nguồn nước dựa trên nhiều đặc trưng đầu vào. Vào năm 2020 Al-Barakani và cộng sự đã sử dụng Random Forest và Gradient Boosting để phân loại nước uống giúp đạt độ chính xác trên 80%. Điều này cho thấy tiềm năng to lớn của các thuật toán hiện đại trong việc thay thế hoặc bổ trợ cho các phương pháp truyền thống.

Cùng sự bùng nổ dữ liệu từ Internet of Things (IoT), cảm biến môi trường và vệ tinh đặt ra nhu cầu xử lý dữ liệu quy mô lớn theo thời gian thực. Các trạm quan trắc có thể tạo ra hàng triệu bản ghi mỗi ngày, vượt quá khả năng xử lý của các công cụ truyền thống. Do đó việc tích hợp học máy với nền tảng dữ liệu lớn là xu thế thiết yếu. Rajesh Kumar và cộng sự vào năm 2021 đã triển khai một hệ thống giám sát chất lượng nước theo thời gian thực dựa trên cảm biến IoT, kết hợp Spark Streaming để xử lý dữ liệu tức thì. Nghiên cứu này cho thấy khả năng áp dụng công nghệ dữ liệu lớn không chỉ trong nghiên cứu mà còn trong quản lý và giám sát môi trường thực tế.

## 1.4. Apache Spark và khả năng ứng dụng trong phân tích nước

Apache Spark là một trong những nền tảng xử lý dữ Big Data phổ biến nhất hiện nay. Được phát triển nhằm khắc phục hạn chế về tốc độ của Hadoop MapReduce, Spark cho phép xử lý dữ liệu in-memory giúp mang lại hiệu năng vượt trội trong nhiều bài toán. Spark cũng cung cấp hệ sinh thái phong phú, bao gồm Spark SQL để truy vấn dữ liệu, MLlib cho học máy, GraphX cho phân tích đồ thị và Spark Streaming cho xử lý dữ liệu dòng theo thời gian thực.

A diagram of company logos

AI-generated content may be incorrect.

Hình 2. Spark

Ưu điểm nổi bật của Spark nằm ở khả năng mở rộng quy mô. Spark có thể tích hợp với nhiều nguồn dữ liệu khác nhau như HDFS, Cassandra, MongoDB, Kafka giúp nó trở thành công cụ lý tưởng cho việc xử lý dữ liệu đến từ nhiều hệ thống quan trắc nước. Trong lĩnh vực phân tích chất lượng nước Spark đã được một số nhóm nghiên cứu ứng dụng thành công. Hussain và cộng sự triển khai Spark năm 2019 để phân tích dữ liệu từ hàng triệu bản ghi cảm biến nhằm phát hiện bất thường nhanh chóng, nhờ đó hỗ trợ các cơ quan quản lý môi trường phản ứng kịp thời.Tiếp tục là Patil và Wagh năm 2020 sử dụng Spark MLlib để dự đoán chỉ số WQI dựa trên dữ liệu lịch sử cung cấp công cụ hỗ trợ ra quyết định cho quản lý nước đô thị. Một điểm mạnh khác của Spark là khả năng tích hợp học máy trên quy mô lớn. Với Spark MLlib, người dùng có thể triển khai các thuật toán như hồi quy, phân loại, phân cụm trên tập dữ liệu hàng triệu bản ghi mà vẫn đảm bảo tốc độ xử lý. Điều này mở ra khả năng xây dựng các hệ thống dự đoán chất lượng nước theo thời gian thực, ứng dụng trong giám sát nguồn nước sinh hoạt, nước công nghiệp hoặc nước tưới tiêu nông nghiệp.

## 1.5. Khoảng trống nghiên cứu và định hướng của đề tài

Từ tổng quan trên có thể nhận thấy rằng các phương pháp truyền thống tuy chính xác nhưng khó mở rộng khi dữ liệu lớn và liên tục. Các nghiên cứu ứng dụng học máy đã đạt được nhiều kết quả khả quan nhưng phần lớn vẫn chỉ dừng lại ở quy mô dữ liệu nhỏ hoặc trung bình chưa khai thác tối đa tiềm năng của công nghệ big data. Việc ứng dụng Apache Spark trong phân tích nước đã xuất hiện trong một số công trình tuy chưa phổ biến rộng rãi và chủ yếu tập trung vào các nghiên cứu thử nghiệm. Khoảng trống nghiên cứu hiện nay là thiếu những giải pháp toàn diện kết hợp Spark với học máy, có khả năng xử lý dữ liệu quy mô lớn và đáp ứng yêu cầu thời gian thực của các hệ thống giám sát hiện đại.

Chính vì lí do này mà đề tài “Sử dụng Spark dự đoán loại nước uống” được thực hiện với mục tiêu khảo sát và áp dụng Spark trong xử lý dữ liệu chất lượng nước. Đề tài hướng đến việc tích hợp Spark MLlib để xây dựng các mô hình học máy dự đoán độ an toàn của nguồn nước, đồng thời đánh giá khả năng mở rộng mô hình trong các hệ thống giám sát tài nguyên nước thông minh. Kết quả nghiên cứu kỳ vọng không chỉ mang lại đóng góp về mặt lý thuyết trong lĩnh vực khoa học dữ liệu môi trường, mà còn có giá trị thực tiễn trong việc hỗ trợ quản lý, bảo vệ và nâng cao chất lượng nguồn nước uống cho cộng đồng.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Các khái niệm cơ bản

### 2.1.1. Chất lượng nước uống

Chất lượng nước uống là một chỉ tiêu tổng hợp phản ánh mức độ an toàn và khả năng sử dụng của nguồn nước cho mục đích sinh hoạt, ăn uống. Các yếu tố ảnh hưởng đến chất lượng nước bao gồm thành phần hóa học như pH, kim loại nặng, nitrat, thành phần sinh học ví dụ: vi sinh vật, coliform, và các chỉ số vật lý: độ đục, nhiệt độ, độ dẫn điện. Tổ chức Y tế Thế giới (WHO) và nhiều cơ quan y tế quốc gia đã ban hành tiêu chuẩn chất lượng nước, đặt ra ngưỡng tối đa cho từng thông số.

A chart with different colors of the same color

AI-generated content may be incorrect.

Hình 3. Bảng tiêu chuẩn của nước uống

### 2.1.2. Chỉ số Chất lượng Nước (Water Quality Index – WQI)

WQI là một công cụ phổ biến được sử dụng để lượng hóa chất lượng nước. Công thức chung của WQI như sau:

Trong đó:

* ​: xếp hạng chất lượng của thông số thứ nnn, được chuẩn hóa về thang điểm 0–100.
* trọng số của thông số thứ nnn, phản ánh tầm quan trọng tương đối.

Dựa trên giá trị WQI, nước được phân loại thành nhiều mức: Excellent, Good, Poor, Very Poor, Unsuitable. Đây là nền tảng để so sánh giám sát và dự báo chất lượng nguồn nước.

### 2.1.3. Dữ liệu chất lượng nước

Dữ liệu nghiên cứu trong đề tài được lấy từ Kaggle (Water Quality Dataset), phản ánh chất lượng nước tại nhiều bang của Ấn Độ. Dữ liệu bao gồm các chỉ số: nhiệt độ (TEMP), độ pH, oxy hòa tan (DO), độ dẫn điện (CONDUCTIVITY), nhu cầu oxy sinh hóa (BOD), nồng độ nitrat, coliform phân (FECAL\_COLIFORM). Đủ điều kiện là tập dữ liệu có quy mô đủ lớn để thử nghiệm các mô hình học máy với Spark.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 4. Dataset đồ án

## 2.2. Nền tảng công nghệ và hệ thống sử dụng

### 2.2.1. Apache Spark

Apache Spark là một nền tảng xử lý dữ liệu phân tán, được phát triển nhằm thay thế Hadoop MapReduce, nổi bật với khả năng xử lý in-memory computing, giúp tăng tốc độ xử lý gấp nhiều lần. Spark hỗ trợ nhiều thư viện:

* **Spark SQL**: xử lý dữ liệu có cấu trúc.
* **Spark MLlib**: triển khai các thuật toán học máy.
* **Spark Streaming**: xử lý dữ liệu thời gian thực.
* **GraphX**: phân tích dữ liệu dạng đồ thị.

Trong đề tài của nhóm Spark được dùng để:

* Xử lý dữ liệu chất lượng nước (ETL, tiền xử lý).
* Trực quan hóa dữ liệu kết hợp với Python.
* Xây dựng và huấn luyện mô hình học máy (MLlib).

### 2.2.2. Pandas và NumPy

Pandas và NumPy được sử dụng để chuyển đổi dữ liệu từ Spark DataFrame sang Pandas DataFrame giúp hỗ trợ việc tính toán WQI và thao tác trên dữ liệu dạng bảng. NumPy cung cấp các hàm toán học nhanh chóng cho xử lý vector và ma trận.

### 2.2.3. Keras/TensorFlow

Để thử nghiệm mô hình học sâu (Deep Learning Regression), đề tài sử dụng thư viện Keras với backend TensorFlow. Keras cho phép xây dựng các mạng nơ-ron nhân tạo nhiều tầng (multi-layer perceptron), huấn luyện trên dữ liệu đầu vào nhằm dự đoán WQI liên tục.

### 2.2.4. Geopandas và Matplotlib/Seaborn

Geopandas được sử dụng để trực quan hóa WQI theo bản đồ các bang của Ấn Độ. Matplotlib và Seaborn hỗ trợ vẽ biểu đồ, phân tích dữ liệu khám phá (EDA), giúp hiểu rõ đặc điểm phân bố dữ liệu trước khi xây dựng mô hình.

## 2.3. Các bước giải quyết vấn đề

Bắt đầu từ tiền xử lý dữ liệu là một bước quan trọng nhằm đảm bảo chất lượng đầu vào cho các mô hình học máy. Dữ liệu ban đầu được đọc vào Spark DataFrame với tất cả các cột ở dạng chuỗi. Để có thể tính toán và phân tích, nhóm đã chuyển đổi chúng sang kiểu số thực float. Các hàng chứa giá trị null cũng được loại bỏ để tránh gây sai lệch trong tính toán WQI. Một số cột không cần thiết như TOTAL\_COLIFORM cũng được loại bỏ để tinh gọn dữ liệu. Kết quả của bước này là một tập dữ liệu sạch, sẵn sàng cho các bước phân tích tiếp theo.

Xong bước làm sạch dữ liệu nhóm tiến hành tính toán chỉ số chất lượng nước (WQI). Do dữ liệu gốc không có sẵn cột WQI nên nhóm đã xây dựng công thức dựa trên các thông số thành phần. Mỗi thông số pH, DO, BOD, coliform, độ dẫn, nitrat được ánh xạ sang một giá trị chuẩn hóa trong khoảng 0 đến 100, dựa trên các ngưỡng an toàn do WHO và các nghiên cứu trước đây đề xuất. Ví dụ giá trị pH trong khoảng 7–8.5 được coi là tối ưu và gán điểm 100 trong khi đó các giá trị lệch nhiều sẽ được gán điểm thấp hơn. Chuẩn hóa hoàn thành thì các thông số này được nhân với trọng số phản ánh tầm quan trọng và cộng lại để tạo thành WQI. Từ đó mỗi mẫu nước được phân loại vào các mức chất lượng khác nhau.

Tiếp theo thì nhóm tiến hành phân tích dữ liệu khám phá. Đây là bước nhằm hiểu rõ đặc điểm phân bố dữ liệu và mối quan hệ giữa các biến. Bằng cách sử dụng Spark SQL và Matplotlib, nhóm vẽ các biểu đồ thể hiện sự thay đổi của DO, pH, BOD, nitrat và coliform. Điều này giúp nhận diện xu hướng, phát hiện dữ liệu bất thường và đánh giá sự đóng góp của từng thông số vào WQI. Đồng thời áp dụng Geopandas để trực quan hóa WQI theo từng bang trên bản đồ Ấn Độ. Cách tiếp cận này không chỉ cho thấy giá trị trung bình mà còn thể hiện sự khác biệt về chất lượng nước giữa các khu vực, giúp hình dung rõ ràng hơn về tình trạng môi trường.

## 2.4. Các mô hình dự đoán

Để dự đoán WQI và phân loại chất lượng nước, nhóm đã triển khai ba mô hình khác nhau nhằm so sánh và đánh giá hiệu quả.

Đầu tiên là mô hình hồi quy tuyến tính (Linear Regression). Đây là phương pháp cổ điển nhưng vẫn hữu ích khi dự đoán biến liên tục như WQI. Trong Spark MLlib, dữ liệu đầu vào được chuyển thành vector bằng VectorAssembler, sau đó được chuẩn hóa bằng Normalizer. Mô hình hồi quy tuyến tính được huấn luyện trên tập dữ liệu này và kết quả được đánh giá bằng chỉ số R2R^2R2. Mặc dù đơn giản, mô hình này cho thấy khả năng dự đoán WQI khá ổn định và là cơ sở để so sánh với các mô hình phức tạp hơn.

A diagram of a linear regression

AI-generated content may be incorrect.

Hình 5. Hồi quy tuyến tính (Linear Regression)

Tiếp đó nhóm thử nghiệm mô hình hồi quy tuyến tính có áp dụng học sâu. Bằng Keras/TensorFlow, một mạng nơ-ron nhân tạo nhiều lớp được xây dựng với hàng trăm nơ-ron trong mỗi lớp ẩn. Hàm kích hoạt ReLU được sử dụng để đảm bảo khả năng học phi tuyến tính, trong khi thuật toán tối ưu Adam và hàm mất mát MSE giúp mô hình hội tụ nhanh. Sau khi huấn luyện mô hình được kiểm tra bằng cách so sánh giá trị dự đoán với dữ liệu thực tế. Kết quả cho thấy mạng nơ-ron có khả năng nắm bắt mối quan hệ phức tạp giữa các biến đầu vào và WQI, nhưng đồng thời cũng đòi hỏi nhiều tài nguyên tính toán hơn.

A diagram of a brain

AI-generated content may be incorrect.

Hình 6. Logistic Regression

Cuối cùng là mô hình hồi quy Logistic (Logistic Regression). Khác với hai mô hình trước, Logistic Regression không dự đoán WQI liên tục mà phân loại trực tiếp chất lượng nước thành các nhãn như Excellent, Good, Poor, Very Poor, Unsuitable. Để thực hiện cột “quality” trước tiên được chuyển thành dạng số bằng StringIndexer. Sau đó dữ liệu đầu vào được tạo vector rồi chuẩn hóa và đưa vào mô hình Logistic Regression. Kết quả dự đoán được so sánh với nhãn thực tế và hiệu quả mô hình được đánh giá bằng thước đo Accuracy. Đây là phương pháp đơn giản nhưng nó cho phép trực tiếp phân loại nguồn nước mà không cần bước trung gian tính WQI và phù hợp cho các ứng dụng giám sát tự động.

## 2.5. Các ràng buộc và thách thức

Trong các bước để triển khai mô hình nhóm nghiên cứu gặp phải một số ràng buộc đáng chú ý. Vấn đề đầu tiên liên quan đến chất lượng dữ liệu. Bộ dữ liệu gốc chứa nhiều giá trị thiếu và chưa đồng nhất, khiến cho quá trình làm sạch và chuẩn hóa trở nên phức tạp. Nếu không xử lý cẩn thận, các mô hình học máy có thể bị sai lệch hoặc cho kết quả thiếu chính xác. Thách thức tiếp theo là việc chuẩn hóa dữ liệu để tính WQI. Các ngưỡng điểm số cho từng thông số như pH, DO, BOD được thiết lập dựa trên quy chuẩn và giả định từ các nghiên cứu trước. Tuy nhiên, trong thực tế, điều kiện môi trường của từng khu vực có thể khác nhau, dẫn đến việc áp dụng công thức chung có thể không phản ánh chính xác tình hình cụ thể.

Về mặt kỹ thuật thì môi trường Spark cũng có những yêu cầu nhất định. Để chạy PySpark, cần cài đặt đúng phiên bản Java (Java 8 thay vì Java 11), cũng như cài đặt các thư viện bổ trợ. Điều này đôi khi gây khó khăn cho người mới tiếp cận. Với mô hình học sâu, thách thức là thời gian huấn luyện dài và nguy cơ overfitting khi mạng nơ-ron quá phức tạp so với lượng dữ liệu. Do đó, việc điều chỉnh tham số và sử dụng các kỹ thuật chống overfitting là cần thiết.

## 2.6. Giải pháp mới và hướng phát triển

Đề tài đóng góp ở chỗ kết hợp nhiều công nghệ và phương pháp khác nhau. Việc áp dụng Spark MLlib cho hồi quy tuyến tính và logistic giúp khai thác sức mạnh xử lý dữ liệu phân tán, trong khi việc tích hợp Keras/TensorFlow cho phép thử nghiệm mạng nơ-ron hiện đại. Nhờ đó giúp nghiên cứu không chỉ giới hạn trong một phương pháp duy nhất mà còn tạo ra sự so sánh khách quan giữa các cách tiếp cận.

Trong tương lai nhóm sẽ có nhiều hướng phát triển khả thi. Trước hết về việc tích hợp Spark Streaming sẽ cho phép xây dựng hệ thống giám sát chất lượng nước theo thời gian thực, xử lý dữ liệu liên tục từ cảm biến IoT. Có thể mở rộng dữ liệu bằng cách kết hợp nguồn từ vệ tinh hoặc các trạm quan trắc tự động giúp nâng cao độ chính xác của mô hình. Các mô hình nâng cao như Random Forest, Gradient Boosting hoặc Deep Neural Networks trên Spark MLlib hoàn toàn có thể được áp dụng để cải thiện kết quả phân loại và dự đoán. Cuối cùng về việc xây dựng một hệ thống dashboard trực quan hóa chất lượng nước sẽ giúp các nhà quản lý dễ dàng tiếp cận thông tin và ra quyết định kịp thời.

# CHƯƠNG 3: MÔ HÌNH THỰC NGHIỆM

## 3.1. Giới thiệu chương

Trong chương mô hình thực nghiệm áp dụng cho dữ liệu chất lượng nước nhằm đánh giá hiệu quả của các phương pháp dự đoán và phân loại. Toàn bộ quy trình được thực hiện trên nền tảng Apache Spark kết hợp với các thư viện hỗ trợ trong Python. Các bước chính bao gồm: tiền xử lý dữ liệu, tính toán chỉ số chất lượng nước (WQI), phân tích và trực quan hóa dữ liệu, xây dựng mô hình dự đoán, và cuối cùng là đánh giá kết quả. Kết quả thực nghiệm được so sánh giữa các phương pháp khác nhau, từ đó đưa ra những nhận xét và phân tích khách quan.

## 3.2. Chuẩn bị dữ liệu và tiền xử lý

### 3.2.1. Thiết lập môi trường để cài đặt pyspark

Thiết lập môi trường trước khi bắt đầu nên nhóm cần thay đổi phiên bản Java vì nếu sử dụng phiên bản 11 thì sẽ gặp một số lỗi và không thể sử dụng PySpark bình thường. Vì vậy nên nhóm sẽ xóa Java phiên bản 11 và cài đặt Java phiên bản 8.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 7. Xóa java 11 cài java 8



Hình 8. Cài Pyspark

Khi cài đặt xong phiên bản java 8 thì tiếp tục với bước cài đặt Pyspark.

### 3.2.2. Đưa file dataset, file map vào Google Colab và tiến hành import các thư viện

A screenshot of a computer

AI-generated content may be incorrect.

Hình 9. Import các thư viện

Bắt đầu giai đoạn đầu nhóm tiến hành tải dữ liệu và thiết lập môi trường phân tích. Trước hết phải có file dữ liệu waterquality.csv được tải trực tiếp từ máy tính cá nhân lên Google Colab bằng thư viện “google.colab.files”. Đây là bước khởi tạo quan trọng, giúp dữ liệu được lưu trong môi trường Colab để có thể gọi lại trong các bước xử lý tiếp theo. Sau đó, nhóm tiến hành nhập các thư viện cần thiết cho phân tích và mô hình hóa, bao gồm pandas và numpy để thao tác dữ liệu dạng bảng và xử lý số học, matplotlib và seaborn để trực quan hóa dữ liệu, geopandas để làm việc với dữ liệu không gian và warnings để loại bỏ các cảnh báo không cần thiết.

Nhóm còn sử dụng các thư viện của Spark như là pyspark.sql và pyspark.ml nhằm xử lý dữ liệu lớn theo mô hình phân tán. Trong đoạn mã, một SparkContext được khởi tạo với chế độ chạy cục bộ cho phép tận dụng toàn bộ lõi CPU của môi trường Colab. Trên cơ sở đó, một SparkSession và SQLContext cũng được thiết lập giúp cung cấp giao diện để xử lý dữ liệu theo phong cách SQL và quản lý DataFrame trong Spark. Việc khởi tạo này đảm bảo rằng dữ liệu có thể được nạp vào khung dữ liệu Spark và sẵn sàng cho các bước tiền xử lý, phân tích và mô hình hóa tiếp theo.

### 3.2.3. Tải dữ liệu lên khung Spark

A screenshot of a computer

AI-generated content may be incorrect.

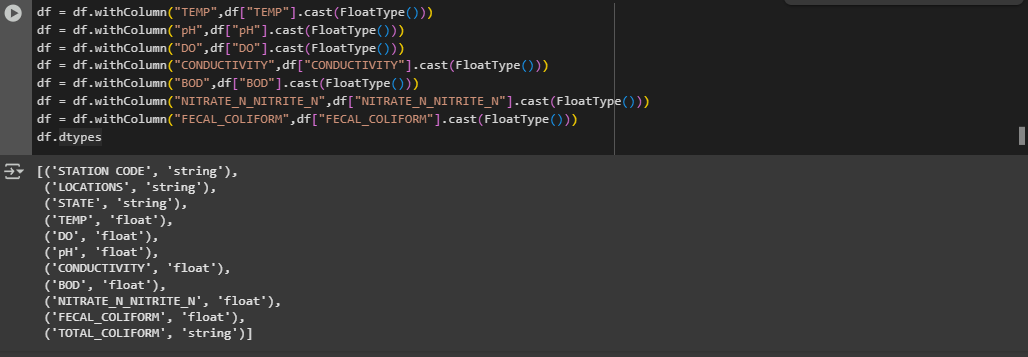
Hình 10. Tải dữ liệu lên khung Spark và xem trước 5 dòng đầu

Sau khi khởi tạo môi trường Spark, nhóm tiến hành tải dữ liệu chất lượng nước từ file waterquality.csv bằng câu lệnh spark.read.format("csv").option("header","true").load(). tùy chọn header = true, Spark tự động nhận hàng đầu tiên trong file CSV là tên cột, giúp định dạng dữ liệu trực quan và thuận tiện hơn cho các thao tác tiếp theo. Song song với đó, nhóm cũng nạp dữ liệu bản đồ hành chính Ấn Độ (Indian\_States.shp) thông qua thư viện geopandas, nhằm phục vụ việc trực quan hóa sau này.

Để có cái nhìn ban đầu nhóm sử dụng “df.show(5)” để hiển thị 5 dòng dữ liệu đầu tiên. Bảng kết quả cho thấy mỗi bản ghi bao gồm các thông tin: mã trạm đo STATION CODE, tên địa điểm LOCATIONS, bang STATE, cùng các chỉ số chất lượng nước như nhiệt độ TEMP, oxy hòa tan DO, độ pH, độ dẫn điện CONDUCTIVITY, nhu cầu oxy sinh hóa BOD, nồng độ nitrat/nitrit NITRATE\_N\_NITRITR\_N, coliform phân (FECAL\_COLIFORM) và tổng coliform TOTAL\_COLIFORM.

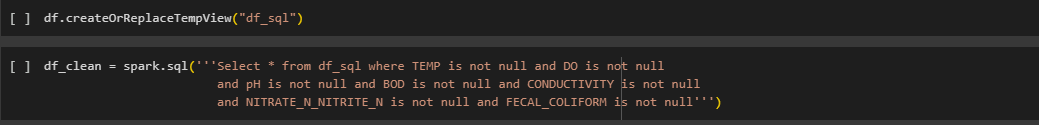
Tiếp đến, nhóm sử dụng “df.dtypes” để kiểm tra kiểu dữ liệu của từng cột. Kết quả cho thấy toàn bộ các trường đều đang ở dạng chuỗi (string). Đây là một đặc điểm thường gặp khi đọc dữ liệu CSV bằng Spark. Tuy nhiên, để có thể tính toán và phân tích, các cột số liệu như TEMP, DO, pH, BOD, hay CONDUCTIVITY cần được chuyển đổi sang kiểu số thực (float). Do đó, bước tiền xử lý tiếp theo sẽ tập trung vào việc chuyển đổi kiểu dữ liệu và loại bỏ những trường không cần thiết, đảm bảo dữ liệu có thể dùng để tính toán chỉ số WQI và huấn luyện mô hình dự đoán.

### 3.2.4. Làm sạch dữ liệu



Hình 11. Chuyển đôi dữ liệu

Tất cả các cột đều có kiểu dữ liệu chuỗi, nhưng để tính toán chỉ số chất lượng nước, chúng ta cần chuyển đổi chúng sang kiểu dữ liệu float. vậy nên chúng ta sẽ chuyển đổi các cột cần thiết sang kiểu dữ liệu float.



Nhóm muốn xóa tất cả các hàng chứa bất kỳ giá trị null nào. Để áp dụng truy vấn SQL, trước tiên chúng ta phải đăng ký bảng tạm ảo (virtual temporary table) và sau đó sẽ thực hiện truy vấn SQL. Nhóm làm điều này vì việc thực hiện các bước làm sạch dữ liệu là rất quan trọng vì nó sẽ giúp mô hình của chúng ta hoạt động tốt hơn.

### 3.2.5. Trực quan hóa dữ liệu



Hình 12. Trực quan hóa dữ liệu

Sau khi đã làm sạch dữ liệu và lưu vào df\_clean, bước tiếp theo trong phân tích dữ liệu là trích xuất các cột cần thiết để thực hiện các phép tính cụ thể (ví dụ như tính chỉ số WQI – Water Quality Index). Đoạn mã dưới đây dùng để lấy dữ liệu từ các cột như DO, pH, BOD, và NITRATE\_N\_NITRITE\_N, và chuyển chúng từ DataFrame PySpark sang danh sách (list) của Python.



Hình 13. Biểu đồ trực quan hóa

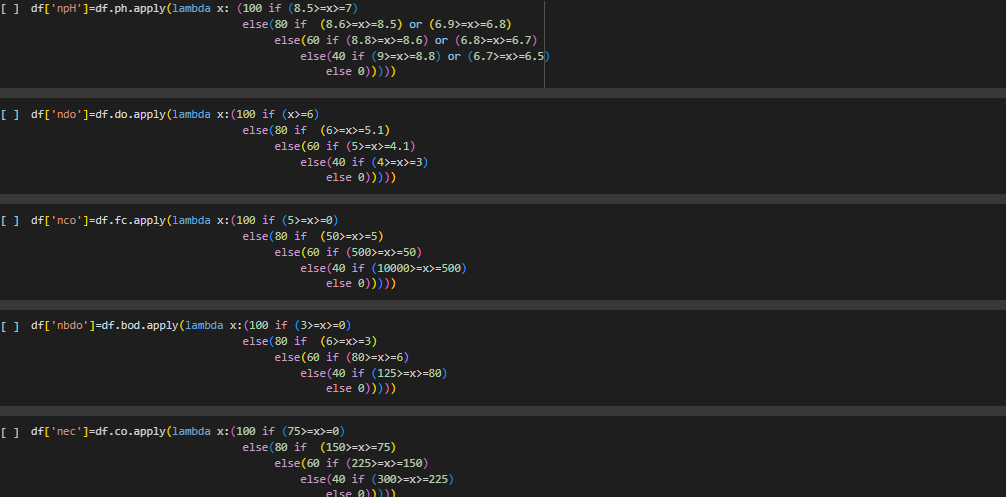
Qua biểu đồ, có thể thấy chỉ số DO biến động mạnh, có những thời điểm giảm xuống thấp. Nồng độ oxy hòa tan thấp thường phản ánh tình trạng ô nhiễm hữu cơ cao, ảnh hưởng xấu đến khả năng tự làm sạch của nguồn nước. Chỉ số pH nhìn chung ổn định trong khoảng trung tính (7–8), phù hợp với tiêu chuẩn nước uống, tuy nhiên vẫn tồn tại một số đỉnh bất thường cần được kiểm chứng để loại trừ sai số đo hoặc tác động từ nguồn thải đột ngột. Đối với BOD, nhiều giá trị vượt ngưỡng cho phép, cho thấy áp lực ô nhiễm hữu cơ từ chất thải sinh hoạt hoặc công nghiệp. Trong khi đó, chỉ số NN xuất hiện một số đỉnh cao, phản ánh nguy cơ nhiễm nitrat hoặc amoni, vốn có thể ảnh hưởng đến sức khỏe nếu vượt quá mức an toàn.

### 3.2.6. Kỹ thuật tính năng



Hình 14. Khởi tạo dữ liệu

Đoạn mã dưới đây sử dụng thư viện Pandas để xử lý dữ liệu dưới dạng bảng (DataFrame). Nhóm chọn 448 dòng đầu tiên trong tập dữ liệu, sau đó trích xuất các cột cần thiết như trạm đo, địa điểm, DO, pH, BOD,... để chuẩn bị cho việc phân tích chất lượng nước.



Hình 15. Tính xếp hạng chất lượng nước

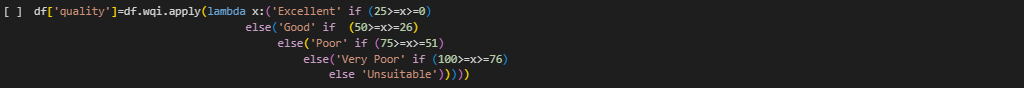
Chỉ số Chất lượng Nước được tính bằng cách cộng tuyến tính xếp hạng chất lượng với trọng số,WQI = ∑ (qn x Wn) trong đó qn = Xếp hạng chất lượng cho thông số chất lượng nước thứ n, Wn = trọng số đơn vị cho thông số thứ n.

Mặc dù chúng ta có công thức chuẩn để tính toán qn nhưng trong trường hợp này không thể thực hiện được, vì vậy chúng tôi đã áp dụng phương pháp chuẩn để tính xếp hạng chất lượng cho từng thông số.

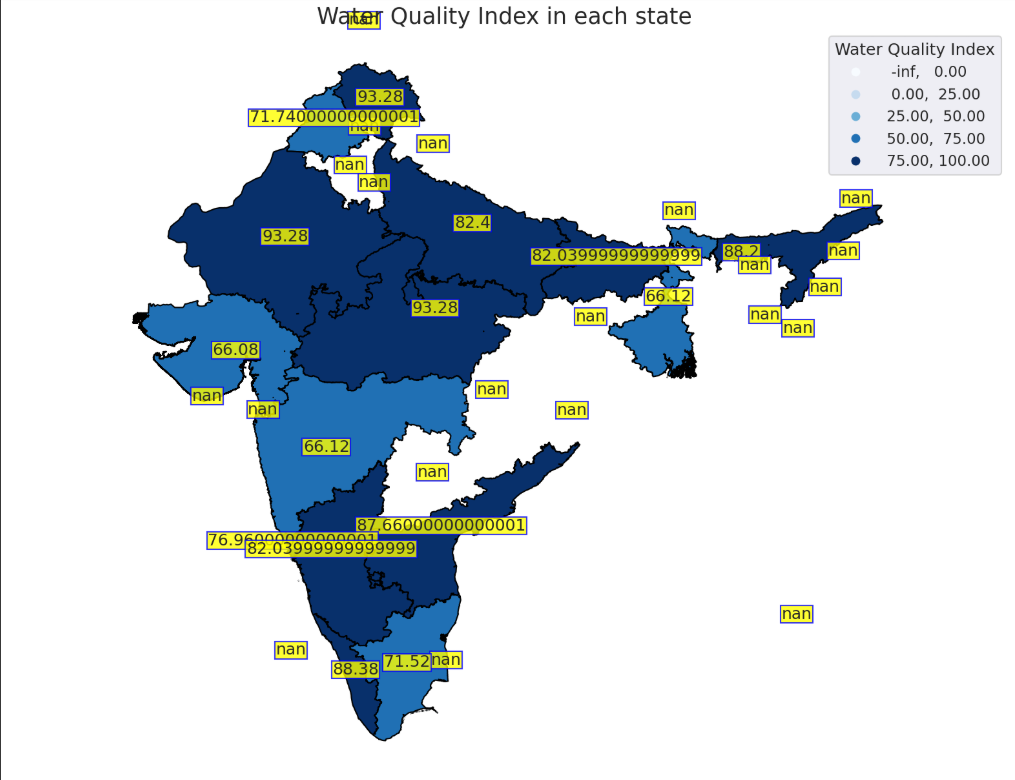


Hình 16. Áp dụng công thức wpi

Nhóm áp dụng công thức wqi bằng cách đầu tiên nhân tất cả xếp hạng chất lượng với trọng số của nó và sau đó cộng tất cả các giá trị.



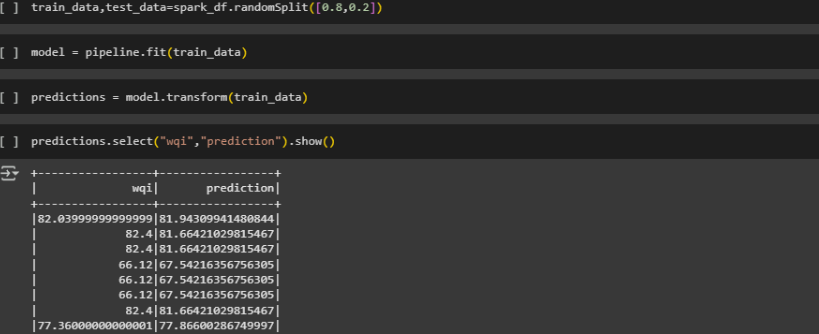
Hình 17. Phân loại chất lượng nước



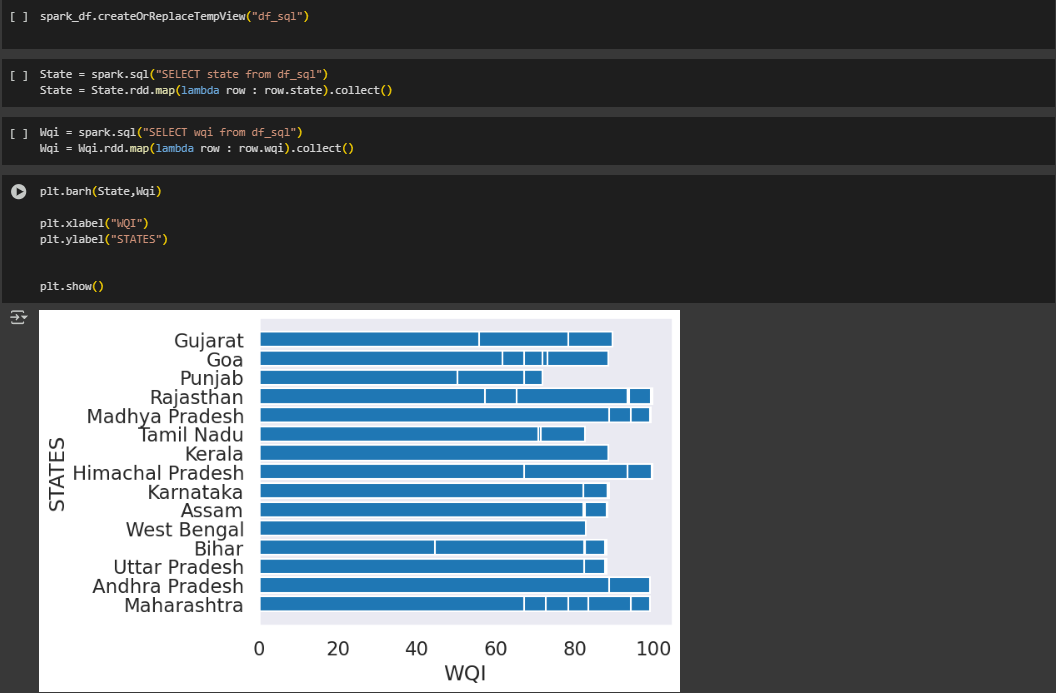
Hình 18. Biểu đồ chất lượng nước tại Ấn Độ

Biểu đồ trên là bản đồ trực quan thể hiện chỉ số chất lượng nước (Water Quality Index - WQI) tại các bang của Ấn Độ. Mỗi bang được tô màu theo thang giá trị WQI, với các cấp độ được chia theo khoảng: 0–25 (kém), 25–50 (trung bình), 50–75 (tốt), và 75–100 (rất tốt). Màu sắc càng đậm thể hiện chất lượng nước càng cao.

Trên bản đồ, giá trị WQI cụ thể của từng bang được hiển thị ngay tại vị trí đại diện, giúp người xem dễ dàng nhận biết mức độ chất lượng nước của từng khu vực. Việc trực quan hóa này không chỉ giúp so sánh sự khác biệt giữa các vùng mà còn hỗ trợ đánh giá tình trạng môi trường nước một cách tổng thể trên toàn quốc. Đây là một công cụ hữu hiệu để xác định các khu vực cần quan tâm và ưu tiên trong công tác bảo vệ nguồn nước.



Chúng ta hãy chuyển đổi lại toàn bộ dữ liệu trong khung tia lửa để thực hiện các quy trình tiếp theo.



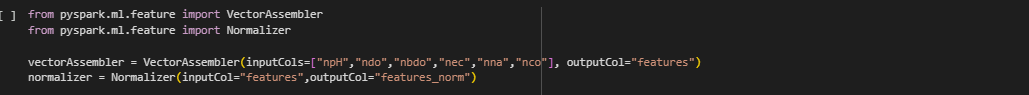
Hình 19. Biểu đồ thanh ngang chất lượng nước tại Ấn Độ

Biểu đồ trên là biểu đồ thanh ngang thể hiện chỉ số chất lượng nước (Water Quality Index - WQI) của các bang tại Ấn Độ. Trục tung liệt kê tên các bang, còn trục hoành biểu diễn giá trị WQI tương ứng. Mỗi thanh ngang đại diện cho một bang, với chiều dài của thanh thể hiện mức độ chất lượng nước: thanh càng dài thì WQI càng cao, tức là chất lượng nước của bang đó càng tốt.

Việc sử dụng biểu đồ thanh ngang giúp dễ dàng so sánh trực tiếp giữa các bang, từ đó xác định được khu vực nào có nguồn nước đạt chuẩn và khu vực nào cần được quan tâm cải thiện. Nhìn vào biểu đồ, có thể thấy sự chênh lệch rõ rệt giữa các vùng, cho thấy tình trạng môi trường nước không đồng đều trên toàn quốc.

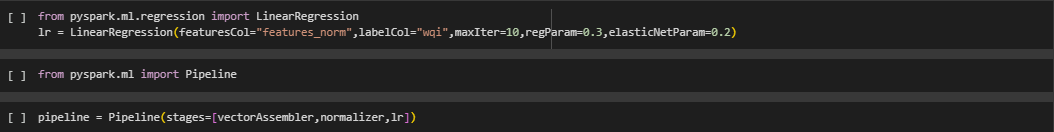
Biểu đồ này đóng vai trò quan trọng trong việc minh họa dữ liệu một cách trực quan, giúp người đọc nhanh chóng nắm bắt thông tin và hỗ trợ việc đưa ra các quyết định liên quan đến quản lý và bảo vệ tài nguyên nước.

### 3.2.7. Mô hình Hồi quy Tuyến tính Không dựa trên Học sâu



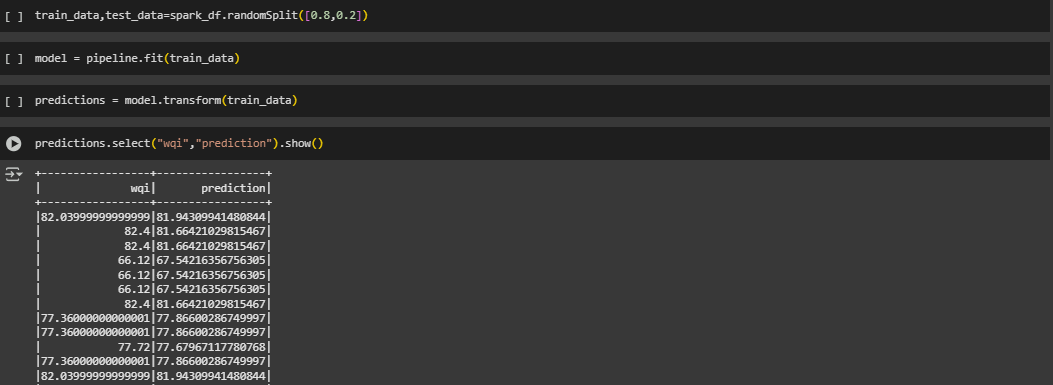
Hình 20. Chuẩn hóa dữ liệu sử dụng Normalizer

Trong mô hình này, dữ liệu đầu tiên cần thiết để dự đoán WQI được chuyển đổi sang dạng vector bằng cách sử dụng VectorAssembler. Sau đó, chúng ta chuẩn hóa dữ liệu bằng cách sử dụng Normalizer.



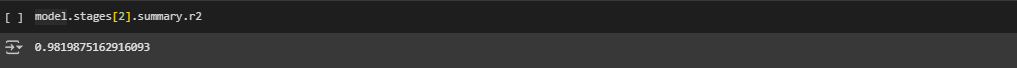
Hình 21. Nhập và đưa vào Pipeline

Sau đó, nhập LinearRegression từ pyspark.ml.regression và áp dụng nó vào dữ liệu đã chuẩn hóa của chúng tôi. Sau đó, nhập Pipeline từ pyspark.ml và đưa tất cả các bước đã thực hiện vào pipeline.



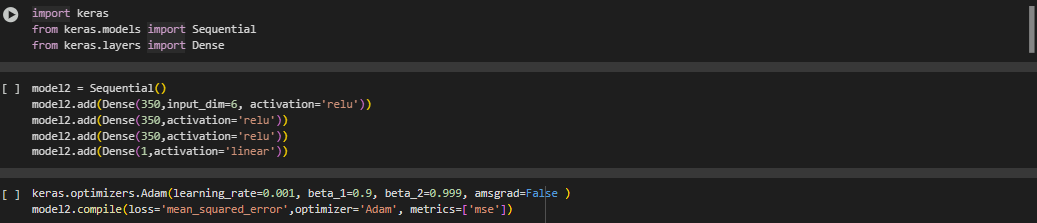
Hình 22. Phân chia ngẫu nhiên 2 cột

Trước khi đào tạo, dữ liệu của chúng tôi được chia ngẫu nhiên thành hai phần để tránh quá khớp và sau đó quá trình đào tạo được thực hiện.



Hình 23.Kết quả đánh giá mô hình

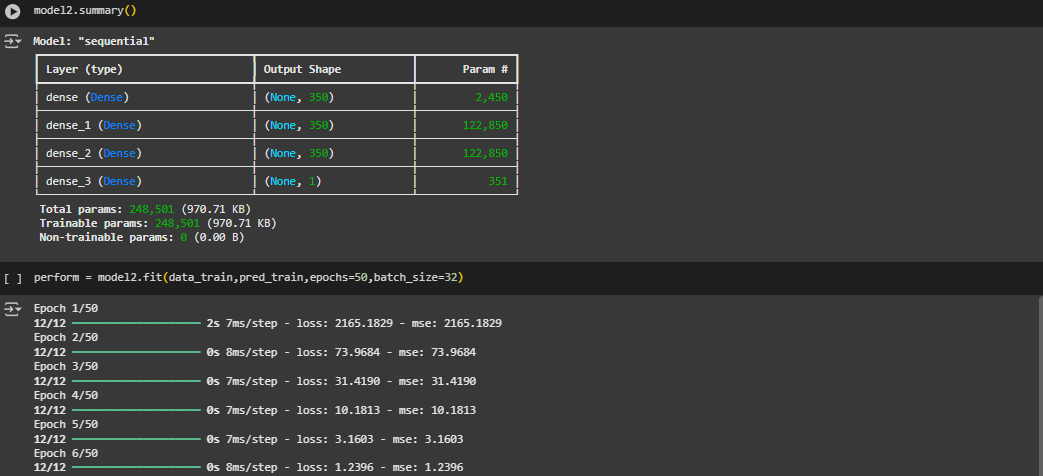
Mô hình hồi quy tuyến tính được sử dụng trong bài toán dự đoán chỉ số chất lượng nước (WQI) đạt được hệ số xác định R2=0.9819R^2 = 0.9819R2=0.9819. Đây là một chỉ số đo lường mức độ phù hợp của mô hình với dữ liệu huấn luyện. Giá trị R2R^2R2 nằm trong khoảng từ 0 đến 1, trong đó giá trị càng gần 1 cho thấy mô hình càng dự đoán tốt. Với R2≈0.98R^2 \approx 0.98R2≈0.98, có thể kết luận rằng mô hình có khả năng giải thích khoảng 98,19% phương sai của dữ liệu đầu ra, chứng tỏ mô hình có độ chính xác rất cao và phù hợp để áp dụng vào bài toán phân tích chất lượng nước.



Hình 24. Mô hình mạng Nơ-ron và Keras

Để dự đoán chỉ số chất lượng nước (WQI), nhóm đã xây dựng một mô hình học sâu (deep learning) sử dụng kiến trúc mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) thông qua thư viện Keras. Mô hình được tạo theo dạng Sequential, bao gồm 4 lớp (layers).

Việc sử dụng mô hình ANN giúp cải thiện độ chính xác so với các mô hình tuyến tính truyền thống, nhờ khả năng học và mô hình hóa mối quan hệ phi tuyến giữa các yếu tố ảnh hưởng đến chất lượng nước.



Hình 25. Kết quả huấn luyện mô hình

Mô hình ANN được xây dựng gồm 4 lớp Dense (fully connected): ba lớp ẩn với 350 nơ-ron mỗi lớp và một lớp đầu ra với 1 nơ-ron. Các lớp ẩn sử dụng hàm kích hoạt ReLU nhằm giúp mô hình học các đặc trưng phi tuyến phức tạp, trong khi lớp đầu ra dùng hàm kích hoạt tuyến tính để phù hợp với bài toán hồi quy.

Tổng số tham số cần học của mô hình là 248,501, trong đó tất cả đều có thể được huấn luyện (trainable). Điều này cho thấy mô hình có độ phức tạp đủ lớn để nắm bắt mối quan hệ giữa các biến đầu vào và đầu ra.

Quá trình huấn luyện mô hình được thực hiện trong 50 epoch với batch size là 32. Qua các epoch đầu tiên, giá trị hàm mất mát (loss) và sai số bình phương trung bình (MSE) giảm mạnh từ hơn 2,000 xuống còn khoảng 3, cho thấy mô hình đang dần hội tụ và cải thiện hiệu suất dự đoán. Đây là tín hiệu cho thấy mạng nơ-ron đang học được các đặc trưng quan trọng của dữ liệu và dự đoán chỉ số chất lượng nước ngày càng chính xác hơn.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 26. Kiểm tra hiệu suất của mô hình

Biểu đồ bên trái thể hiện quá trình giảm giá trị hàm mất mát (loss) trong suốt 50 epoch huấn luyện của mô hình mạng nơ-ron. Ban đầu, giá trị loss rất cao, vượt ngưỡng 1000, nhưng nhanh chóng giảm mạnh trong những epoch đầu tiên và sau đó ổn định ở mức thấp hơn rất nhiều, dưới 10. Điều này cho thấy mô hình đã học được các đặc trưng quan trọng từ dữ liệu và khả năng dự đoán được cải thiện rõ rệt trong quá trình huấn luyện.

Biểu đồ bên phải so sánh giá trị thực tế (biểu diễn bằng dấu chấm xanh) với giá trị dự đoán (dấu cộng màu xanh lá) của mô hình trên tập dữ liệu huấn luyện. Ta có thể thấy các điểm dự đoán nằm khá gần với các điểm dữ liệu thực tế, đặc biệt ở các vùng có giá trị chỉ số chất lượng nước cao (gần 80-100). Điều này chứng minh rằng mô hình không chỉ hội tụ tốt mà còn dự đoán chính xác các giá trị đầu ra trong phạm vi dữ liệu được cung cấp.

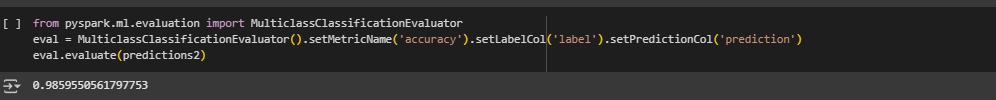
### 3.2.8. Mô hình Hồi quy Tuyến tính dựa trên Học sâu



Vì cột chất lượng của chúng tôi chứa các giá trị ở định dạng chuỗi, nên trước tiên chúng tôi lập chỉ mục chúng bằng StringIndexer. Sau đó, dữ liệu cần thiết để dự đoán chất lượng nước được chuyển đổi sang dạng vector bằng VectorAssembler. Cuối cùng, chúng tôi chuẩn hóa dữ liệu bằng Normalizer.



Sau đó, nhập LogisticRegression từ pyspark.ml.classification và áp dụng vào dữ liệu đã chuẩn hóa của chúng tôi. Tiếp theo, nhập Pipeline từ pyspark.ml và đưa tất cả các bước đã thực hiện vào pipeline.



Hình 27. Đánh giá mô hình

Với kết quả đạt 98.59%, sử dụng thư viện PySpark. Điều này cho thấy mô hình dự đoán đúng gần như toàn bộ các trường hợp trên tập dữ liệu kiểm tra, chứng tỏ hiệu suất rất cao và phù hợp để áp dụng vào bài toán phân loại chất lượng nước.



Hình 28. Dự đoán với giá trị thực tế

Sau khi đánh giá độ chính xác tổng thể của mô hình bằng chỉ số accuracy (98.59%), nhóm tiếp tục thực hiện kiểm tra trực tiếp đầu ra của mô hình bằng cách so sánh 100 nhãn dự đoán đầu tiên với giá trị thực tế từ cột quality. Các chỉ số đầu ra từ mô hình (dạng số nguyên) được ánh xạ về các nhãn phân loại tương ứng với mức chất lượng nước: "Very Poor", "Poor", "Good", "Unsuitable", "Excellent".

Đoạn mã sử dụng Spark SQL để truy xuất hai cột prediction và quality, sau đó chuyển sang RDD và thu thập dưới dạng list để có thể lặp qua. Việc in trực tiếp cặp giá trị "Predicted" và "Actual" giúp kiểm tra định tính mức độ chính xác của mô hình ở từng dòng dữ liệu cụ thể.

Kết quả cho thấy mô hình dự đoán chính xác gần như toàn bộ trường hợp, đặc biệt ở các lớp "Very Poor" và "Good", là hai lớp phổ biến trong tập dữ liệu. Điều này không chỉ xác nhận hiệu quả của mô hình trên tổng thể, mà còn chứng minh rằng mô hình có thể phân loại đúng các mức chất lượng nước cụ thể, phù hợp để triển khai trong thực tế hoặc sử dụng trong các hệ thống giám sát môi trường tự động.

# CHƯƠNG 4: KẾT LUẬN

### 4.1. Kết luận

Đề tài đã triển khai thành công quy trình xử lý, phân tích và xây dựng mô hình dự đoán chỉ số chất lượng nước (Water Quality Index - WQI) dựa trên các thông số hóa lý như DO, pH, BOD, Nitrate v.v với sự hỗ trợ của PySpark trong xử lý dữ liệu lớn và mô hình học máy (Linear Regression, Neural Network) trong dự đoán. Các biểu đồ trực quan và kết quả thực nghiệm cho thấy mô hình có hiệu năng cao, trong đó mô hình Linear Regression đạt R² = 0.98 và mô hình phân loại đạt độ chính xác 98.59%. Kết quả này chứng minh khả năng áp dụng của các mô hình học máy trong việc hỗ trợ giám sát, đánh giá chất lượng nguồn nước một cách nhanh chóng và hiệu quả.

### 4.2. Hạn chế và hướng phát triển

Hạn chế:

* Dữ liệu thiếu đồng đều giữa các bang và khu vực, khiến một số tỉnh không có giá trị WQI hoặc dự đoán sai lệch.
* Sự mất cân bằng dữ liệu phân loại, ví dụ lớp "Very Poor" xuất hiện nhiều hơn "Excellent", có thể ảnh hưởng đến mô hình phân loại.
* Mô hình hiện tại chưa xét đến yếu tố thời gian (time series) nên không đánh giá được xu hướng thay đổi chất lượng nước theo mùa hoặc năm.
* Việc sử dụng dữ liệu tĩnh không phản ánh tính thời gian thực, hạn chế khả năng ứng dụng trong hệ thống cảnh báo sớm.

Hướng phát triển

* Tích hợp dữ liệu thời gian để xây dựng các mô hình dự đoán theo chuỗi thời gian (time series forecasting).
* Ứng dụng học sâu (Deep Learning) như LSTM, GRU cho các bài toán dự đoán WQI nâng cao theo thời gian thực.
* Thu thập thêm dữ liệu từ nhiều nguồn khác nhau, đặc biệt các cảm biến IoT, để làm phong phú tập dữ liệu và tăng tính chính xác mô hình.
* Triển khai hệ thống giám sát chất lượng nước tự động, tích hợp mô hình đã huấn luyện vào một dashboard hoặc API để hỗ trợ ra quyết định nhanh chóng cho các cơ quan quản lý môi trường.

# 

# TÀI LIỆU THAM KHẢO

[1] “Chỉ số TDS là gì? Nước có chỉ số TDS bao nhiêu thì uống được?” *máy lọc nước Geyser*,25 April 2022, <https://www.geyser.com.vn/tin-tuc/chi-so-tds-la-gi/>. Accessed 11 August 2025.

[2] “QCVN 6-1:2010/BYT Quy chuẩn kỹ thuật quốc gia đối với nước khoáng thiên nhiên và nước uống đóng chai?” *Thư Viện Pháp Luật*, 1 January 2025, <https://thuvienphapluat.vn/hoi-dap-phap-luat/839DB2D-hd-qcvn-612010byt-quy-chuan-ky-thuat-quoc-gia-doi-voi-nuoc-khoang-thien-nhien-va-nuoc-uong-dong-chai.html>. Accessed 11 August 2025.

[3] “Machine Learning cơ bản.” *Machine Learning cơ bản*, 26 December 2016, <https://machinelearningcoban.com/2016/12/26/introduce/>. Accessed 11 August 2025.

[4] “Spark là gì? – Giới thiệu về Apache Spark và Phân tích.” *AWS*, [https://aws.amazon.com/vi/what-is/apache-spark/. Accessed 2 September 2025](https://aws.amazon.com/vi/what-is/apache-spark/.%20Accessed%202%20September%202025).

[5] “Apache Spark #1: kiến thức cơ bản về nền tảng của Apache Spark.” *Viblo*, [https://viblo.asia/p/apache-spark-1-kien-thuc-co-ban-ve-nen-tang-cua-apache-spark-MG24BkEWJz3. Accessed 2 September 2025](https://viblo.asia/p/apache-spark-1-kien-thuc-co-ban-ve-nen-tang-cua-apache-spark-MG24BkEWJz3.%20Accessed%202%20September%202025).

[6] “What is Deep Learning? - Deep Learning AI Explained.” *AWS*, <https://aws.amazon.com/what-is/deep-learning/>. Accessed 2 September 2025.