Hệ thống đề xuất điều chỉnh các yếu tố thời tiết trong nhà kính để phù hợp với điều kiện phát triển của cây trồng

Võ Anh Quân - 22521192; Nguyễn Cao Thắng - 22521329; Võ Minh Quyền - 22521227

Trường Đại học Công nghệ Thông tin - Đại học Quốc gia TP.HCM

Tóm tắt nội dung

Đề tài tập trung vào việc xây dựng một hệ thống tự động đề xuất các điều chỉnh về nhiệt độ, độ ẩm, ánh sáng và lượng nước tưới trong nhà kính theo thời gian thực nhằm tối ưu hóa điều kiện phát triển của cây trồng. Hệ thống sử dụng dữ liệu thời tiết thời gian thực kết hợp với thông tin về loại cây, loại đất, và các yếu tố môi trường để đưa ra khuyến nghị chính xác. Trong đề tài này, nhóm sẽ sử dụng Kafka để stream dữ liệu thời gian thực từ môi trường, thực hiện xử lý dữ liệu và đưa ra khuyến nghị điều chỉnh các yếu tố thời tiết, đồng thời sử dụng sự hỗ trợ của MLlib để xây dựng mô hình khuyến nghị nên canh tác loại cây nào vào thời điểm nào.

Keywords: Kafka, Spark, Data Streaming, MLlib, Big Data

1 Giới thiệu đề tài

Trong bối cảnh nhu cầu về thực phẩm chất lượng cao ngày càng gia tăng, việc ứng dụng công nghệ vào nông nghiệp đã trở thành xu hướng thiết yếu. Một trong những ứng dụng quan trọng là việc tối ưu hóa điều kiện nhà kính để đảm bảo cây trồng phát triển hiệu quả, đáp ứng năng suất và chất lượng mong muốn.

Hệ thống nhà kính hiện đại không chỉ đơn thuần cung cấp môi trường bảo vệ cây trồng khỏi tác động bên ngoài mà còn có khả năng điều chỉnh các yếu tố thời tiết như nhiệt độ, độ ẩm, ánh sáng, và mức độ tưới tiêu để phù hợp với từng loại cây trồng. Tuy nhiên, việc quản lý và điều chỉnh các yếu tố này đòi hỏi sự chính xác cao và phụ thuộc vào khối lượng dữ liệu lớn được thu thập từ các cảm biến và nguồn dữ liệu thời gian thực.

Đề tài "Hệ thống đề xuất điều chỉnh các yếu tố thời tiết trong nhà kính để phù hợp với điều kiện phát triển của cây trồng"tập trung vào việc xây dựng một hệ thống đề xuất dựa trên các công nghệ như Kafka [3], Spark [4] và MLlib [5].

Đề tài không chỉ giúp nâng cao năng suất và chất lượng cây trồng mà còn góp phần tiết kiệm năng lượng và tài nguyên, hướng đến mô hình nông nghiệp thông minh và bền vững trong tương lai.

2 Các đề tài nghiên cứu liên quan

Trong lĩnh vực nông nghiệp thông minh và hệ thống nhà kính tự động, đã có nhiều nghiên cứu tập trung vào việc tối ưu hóa điều kiện môi trường để nâng cao năng suất và chất lượng cây trồng. Một số nghiên cứu liên quan tiêu biểu bao gồm:

- **Hệ thống IoT giám sát và điều chỉnh điều kiện nhà kính**: Nghiên cứu này triển khai hệ thống IoT để giám sát các thông số như nhiệt độ, độ ẩm, ánh sáng và nồng độ CO_2 trong nhà kính. Hệ thống tích hợp cảm biến IoT và các thuật toán trí tuệ nhân tạo nhằm tự động điều chỉnh các yếu tố môi trường theo thời gian thực, giúp tối ưu hóa điều kiện sinh trưởng của cây trồng.

- **Ứng dụng học máy trong dự báo năng suất cây trồng**: Đề tài này sử dụng các mô hình học máy để phân tích và dự đoán năng suất cây trồng dựa trên dữ liệu môi trường trong nhà kính, bao gồm nhiệt độ, độ ẩm, loại đất, và ánh sáng. Kết quả nghiên cứu hỗ trợ người trồng trong việc đưa ra các quyết định điều chỉnh môi trường phù hợp.

- **Hệ thống khuyến nghị cây trồng theo điều kiện môi trường**: Nghiên cứu xây dựng hệ thống gợi ý loại cây trồng phù hợp với điều kiện hiện tại trong nhà kính. Hệ thống này kết hợp dữ liệu thời tiết thời gian thực, loại đất, và thông tin sinh học của cây trồng để đưa ra các khuyến nghị tối ưu.

Các nghiên cứu trên cung cấp cơ sở khoa học và ứng dụng thực tiễn cho việc phát triển hệ thống đề xuất điều chỉnh các yếu tố thời tiết trong nhà kính. Tuy nhiên, đề tài này tập trung sâu hơn vào việc kết hợp dữ liệu thời gian thực từ cảm biến, dữ liệu thời tiết qua Kafka, và hệ thống gợi ý dựa trên các điều kiện lý tưởng để phát triển cây trồng trong từng mùa vụ, nhằm mang lại giải pháp tối ưu hơn cho người trồng.

3 Bộ dữ liệu

3.1 Dữ liệu về khuyến nghị cây trồng dựa trên tính chất đất và các điều kiện thời tiết

Bộ dữ liệu chúng tôi sử dụng được lấy từ FAO [2], về một số loại cây nông nghiệp phổ biến ở châu Âu và Bắc Mỹ như cây Dagussa, cây Teff, cây sorghum, hạt cây niger, khoai tây, lúa mì, lúa mạch, đậu, đậu Hà Lan, ớt đỏ. Trong bộ dữ liệu này đề cập đến những điều kiện về đất và các điều kiện thời tiết ảnh hưởng đến sự phát triển của cây trồng để có thể đạt được năng suất cao nhất. Bộ dữ liệu đã được chỉnh sửa để phù hợp với đề tài, bao gồm các thuộc tính và mô tả như sau:

Bảng 1: Bảng mô tả bộ dữ liệu

Thuộc tính	Mô tả
label (string)	Tên loại cây trồng
soilcolor (string)	Màu sắc của đất dùng để trồng
pH (double)	Độ pH của đất
QV2M-W (double)	Giá trị độ ẩm không khí ở độ cao 2m vào mùa đông
QV2M-Sp (double)	Giá trị độ ẩm không khí ở độ cao 2m vào mùa xuân
QV2M-Su (double)	Giá trị độ ẩm không khí ở độ cao 2m vào mùa hạ
QV2M-Au (double)	Giá trị độ ẩm không khí ở độ cao 2m vào mùa thu
PRECTOTCORR-W (double)	Lượng nước tưới trung bình cho cây vào mỗi ngày mùa đông ($m^3/ha/ngy$)
PRECTOTCORR-Sp (double)	Lượng nước tưới trung bình cho cây vào mỗi ngày mùa xuân ($m^3/ha/ngy$)
PRECTOTCORR-Su (double)	Lượng nước tưới trung bình cho cây vào mỗi ngày mùa hạ ($m^3/ha/ngy$)
PRECTOTCORR-Au (double)	Lượng nước tưới trung bình cho cây vào mỗi ngày mùa thu ($m^3/ha/ngy$)
GWETTOP (double)	Độ ẩm của lớp đất mặt.
CLOUD-AMT (double)	Lượng mây trong không khí, có thể ảnh hưởng đến ánh sáng mặt trời và nhiệt độ.
PS (double)	Áp suất khí quyển.
T2M-AVG-W (double)	Nhiệt độ trung bình của không khí ở độ cao 2 mét trong mùa đông
T2M-AVG-Sp (double)	Nhiệt độ trung bình của không khí ở độ cao 2 mét trong mùa xuân
T2M-AVG-Su (double)	Nhiệt độ trung bình của không khí ở độ cao 2 mét trong mùa hạ
T2M-AVG-Au (double)	Nhiệt độ trung bình của không khí ở độ cao 2 mét trong mùa thu
pH-diff (double)	Độ chênh lệch pH giữa chỉ số hiện tại của đất và mức trung tính

3.2 Bộ dữ liệu thời gian thực về các điều kiện đất đai và thời tiết

Trong đề tài này, chúng tôi sử dụng dữ liệu điều kiện đất đai và thời tiết theo thời gian thực ở khu vực thử nghiệm mẫu là San Francisco (Hoa Kỳ). Bộ dữ liệu được thu thập từ API của Agromonitoring [1], cung cấp các dữ liệu theo thời gian thực về nông nghiệp (trong đề tài này chúng tôi thực hiện cập nhật dữ liệu mỗi 30 phút một lần). Sau khi request dữ liệu từ Agromonitoring, dữ liệu này được đưa lên Kafka và lưu trữ bằng hệ quản trị cơ sở dữ liệu MongoDB dưới dạng các tài liệu JSON

Bảng 2: Bảng mô tả bộ dữ liệu

Thuộc tính	Mô tả
id (string)	ID bản ghi
temperature (double)	Nhiệt độ không khí tại thời điểm đo
humidity-air (integer)	Độ ẩm không khí
pressure (integer)	Áp suất không khí (hPa)
soil-moisture (double)	Độ ẩm đất
timestamp	Thời gian đo đạc (YYYY-MM-DD HH:mm:ss)

4 Triển khai

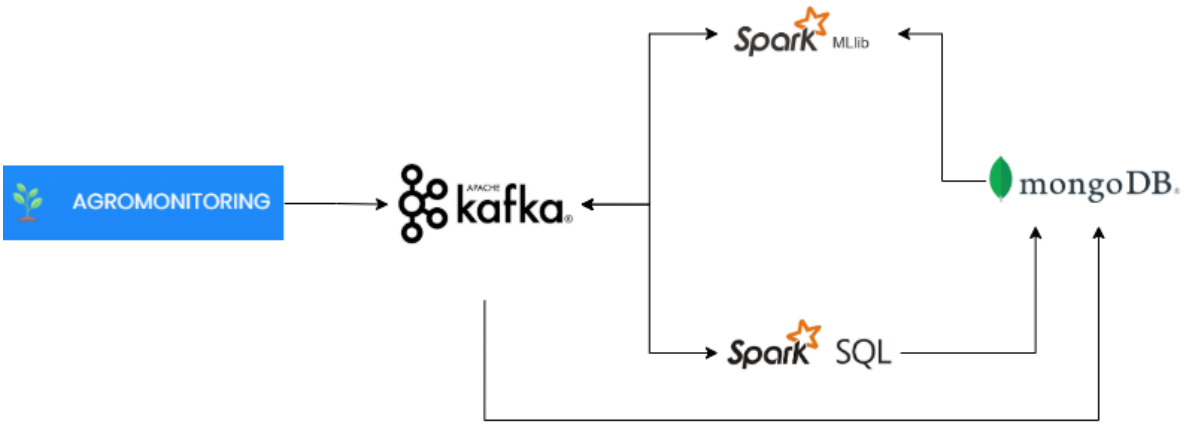
4.1 Tổng quan phương pháp

Đầu tiên, dữ liệu thời tiết thời gian thực ở San Francisco được stream thông qua request API của Agromonitoring bằng Kafka. Trong đề tài này, chúng tôi thực hiện stream dữ liệu mỗi 30 phút một lần để có thể dễ nhận biết các biến động về các yếu tố thời tiết hơn.

Request sẽ được điều chỉnh để dữ liệu stream về có đơn vị của các thuộc tính về nhiệt độ, áp suất phù hợp. Sau đó dữ liệu này sẽ được lưu trữ trên một cơ sở dữ liệu phân tán quản lý bởi hệ quản trị cơ sở dữ liệu MongoDB [6].

Đồng thời, dữ liệu được stream về bằng Kafka sẽ được thực hiện xử lý để đưa ra các khuyến nghị phù hợp như tăng/giảm nhiệt độ, độ ẩm,... dưới sự hỗ trợ của Spark SQL. Các khuyến nghị được đưa ra cũng sẽ được đồng bộ cùng với mỗi lần lấy dữ liệu của Kafka (mỗi 30 phút một lần) và được lưu trữ vào một cơ sở dữ liệu quản lý bởi MongoDB.

Những dữ liệu thời tiết thời gian thực nhận về cũng sẽ được sử dụng để thực hiện khuyến nghị về loại cây trồng phù hợp với tình hình thời tiết hiện tại. Trong đề tài này, chúng tôi sử dụng mô hình Random Forrest dưới sự hỗ trợ của Spark MLlib để thực hiện huấn luyện trên tập dữ liệu về khuyến nghị cây trồng dựa trên tính chất đất và các điều kiện thời tiết.



Hình 1: Sơ đồ tổng quan phương pháp thực hiện

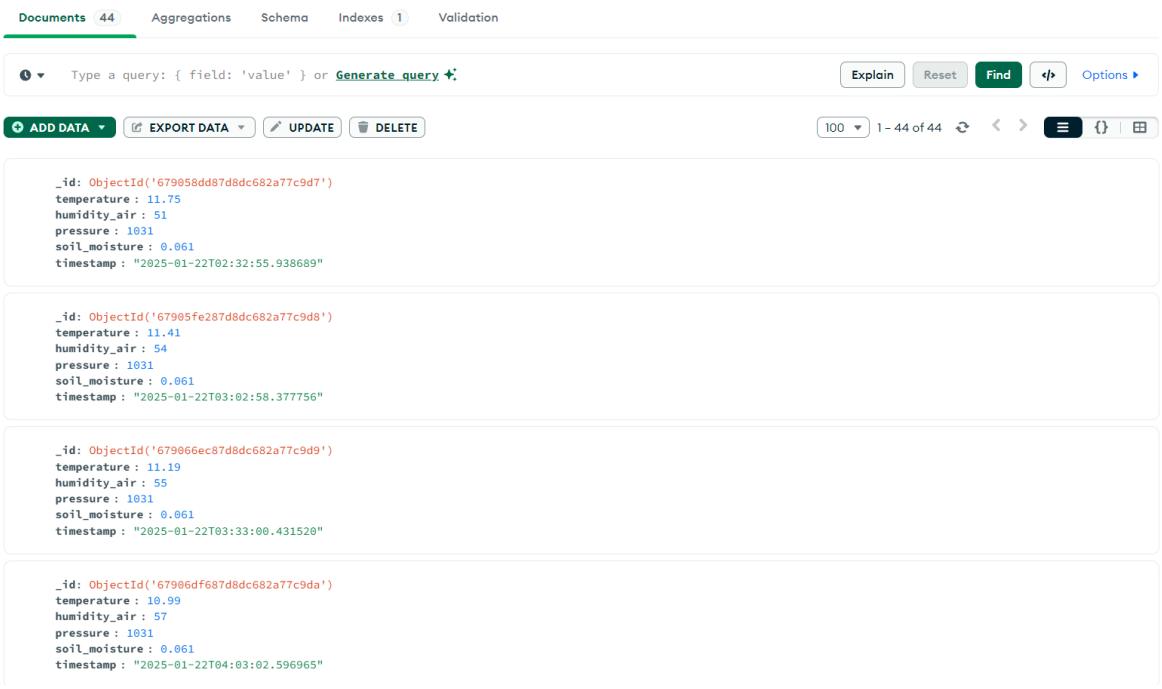
4.2 Đầu vào và tiền xử lý dữ liệu

Producer thực hiện việc thu thập dữ liệu thời tiết và độ ẩm đất thông qua Agromonitoring API. Dữ liệu thời tiết, bao gồm nhiệt độ, độ ẩm không khí, và áp suất, được lấy dựa trên tọa độ địa lý, trong khi dữ liệu đất (độ ẩm đất) được lấy dựa trên polygon ID đại diện cho một khu vực cụ thể. Sau khi thu thập, dữ liệu từ hai API được kết hợp thành một cấu trúc JSON duy nhất, bao gồm các thông tin thời tiết, đất, và dấu thời gian hiện tại. Dữ liệu này sau đó được gửi đến một topic Kafka (`SF_weather_data`) thông qua Kafka Producer. Toàn bộ quy trình được thực hiện định kỳ mỗi 30 phút bằng cách sử dụng vòng lặp `while`.

Producer sử dụng Apache Kafka làm nền tảng xử lý luồng dữ liệu lớn, cho phép truyền dữ liệu thời gian thực với độ tin cậy và khả năng mở rộng cao. Kafka đóng vai trò là trung tâm xử lý, đảm bảo dữ liệu từ ứng dụng được truyền đến các hệ thống downstream. Ngoài ra, việc tích hợp Agromonitoring API cho phép thu thập dữ liệu thời gian thực tự động trong việc lập lịch thu thập và gửi dữ liệu định kỳ.

Consumer thực hiện việc thu thập dữ liệu từ Kafka topic (`SF_weather_data`) và lưu trữ vào cơ sở dữ liệu MongoDB. Dữ liệu được lấy từ Kafka Consumer, nơi các thông tin thời tiết và đất đã được gửi trước đó bởi Kafka Producer. Consumer được cấu hình để tự động lấy dữ liệu từ đầu topic và giải mã chúng từ định dạng JSON. Mỗi bản ghi dữ liệu nhận được sẽ được lưu vào một cơ sở dữ liệu MongoDB với tên `Weather-condition` và được lưu cụ thể trong collection `WeatherData`. Toàn bộ quy trình chạy liên tục, đảm bảo tất cả dữ liệu đến từ Kafka đều được lưu trữ vào MongoDB.

Consumer sử dụng Apache Kafka để quản lý và truyền luồng dữ liệu thời gian thực từ hệ thống upstream (Producer) đến downstream (Consumer). Kafka Consumer đảm bảo dữ liệu được lấy đúng thứ tự và không bị mất dữ liệu. Bên cạnh đó, MongoDB được sử dụng làm cơ sở dữ liệu NoSQL để lưu trữ dữ liệu không có cấu trúc hoặc bán cấu trúc (JSON). Việc sử dụng MongoDB giúp xử lý và truy vấn dữ liệu linh hoạt, phù hợp với các hệ thống lưu trữ dữ liệu lớn.



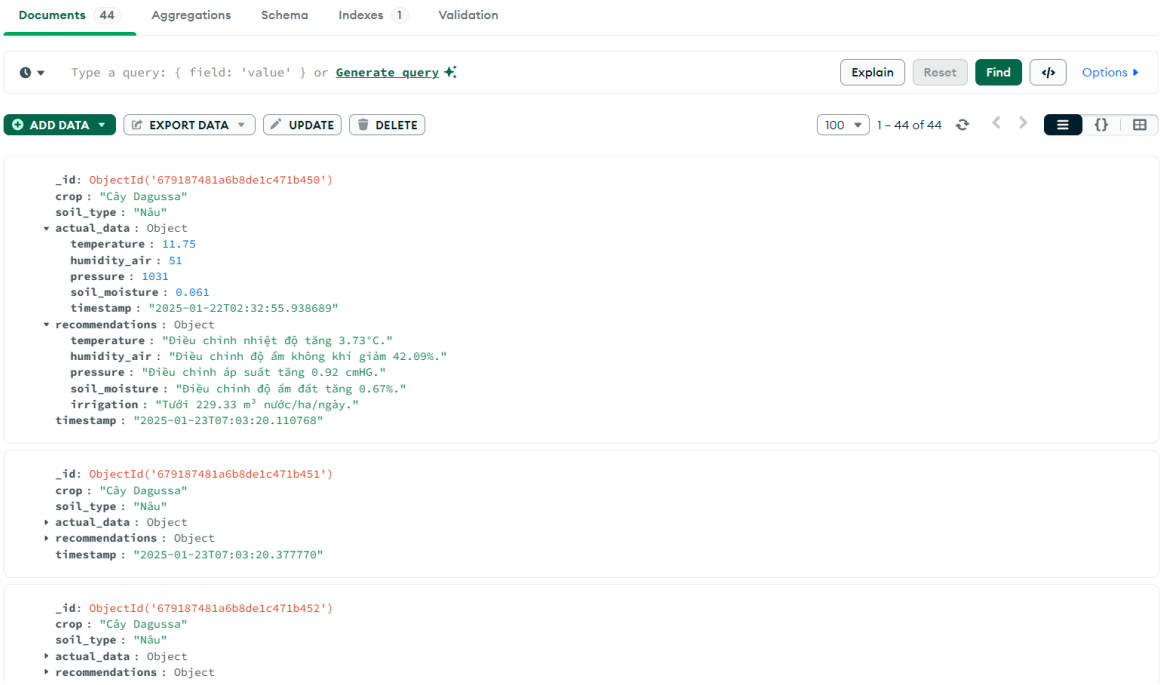
Hình 2: Dữ liệu đất đai và thời tiết thu thập được

4.3 Xử lý dữ liệu và đưa ra khuyến nghị

Hệ thống khuyến nghị điều chỉnh các yếu tố thời tiết và đất trong nhà kính dựa trên dữ liệu thực tế từ Kafka và dữ liệu lý tưởng từ file CSV. Dữ liệu thời gian thực về nhiệt độ, độ ẩm không khí, áp suất, và độ ẩm đất được lắng nghe từ topic Kafka (`SF_weather_data`). Sau đó, dữ liệu này được kết hợp với thông tin lý tưởng (nhiệt độ, độ ẩm, lượng nước tưới, v.v.) cho từng loại cây trồng và loại đất, được lưu trữ trong file CSV. Hệ thống sử dụng SparkSQL để xử lý dữ liệu CSV, xác định thời điểm hiện tại và tính toán sự khác biệt giữa các điều kiện thực tế và lý tưởng. Từ đó, các khuyến nghị cụ thể được xây dựng để điều chỉnh môi trường. Mỗi khuyến nghị được lưu trữ vào MongoDB để sử dụng sau này.

Kafka được sử dụng như một hệ thống quản lý luồng dữ liệu, cho phép thu thập và truyền tải dữ liệu thời gian thực từ các cảm biến hoặc hệ thống upstream, đảm bảo dữ liệu được tiếp nhận một cách liên tục và chính xác. SparkSQL đóng vai trò quan trọng trong việc xử lý dữ liệu lớn, giúp lọc, truy vấn và phân tích các thông tin lý tưởng về cây trồng và đất đai từ file CSV, đồng thời hỗ trợ khả năng

mở rộng và tính toán hiệu quả trên tập dữ liệu lớn.



Hình 3: Kết quả khuyến nghị điều chỉnh yếu tố môi trường dựa theo tình hình thời tiết và đất đaiđai

Đồng thời, dữ liệu thời tiết thu thập được cũng được export ra dạng CSV và thực hiện xây dựng khuyến nghị về loại cây trồng phù hợp với tình hình thời tiết hiện tại. Quy trình huấn luyện và dự đoán sử dụng mô hình Random Forest trên dữ liệu thời tiết và cây trồng theo quy trình cụ thể sau:

1. Tạo SparkSession: Đầu tiên, mã tạo một phiên làm việc với SparkSession, cho phép sử dụng các API của Spark để xử lý dữ liệu phân tán.

2. Đọc và xử lý dữ liệu: Các tệp CSV chứa dữ liệu về khuyến nghị cây trồng và dữ liệu thời tiết được đọc vào Spark DataFrame. Dữ liệu thời tiết được biến đổi để tính toán các giá trị nhiệt độ, độ ẩm, và áp suất cho từng mùa (winter, spring, summer, autumn) từ dữ liệu gốc.

3. Chuyển đổi dữ liệu: Dữ liệu về cây trồng và loại đất được mã hóa thành các chỉ số số bằng cách sử dụng `StringIndexer`. Các cột dữ liệu thời tiết được chuẩn hóa và điều chỉnh để phù hợp với mô hình.

4. Tiền xử lý dữ liệu: Dữ liệu được chuẩn hóa bằng StandardScaler và kết hợp các đặc trưng thành một vector để đưa vào mô hình học máy. Dữ liệu được chia thành hai phần: 80% cho huấn luyện và 20% cho kiểm tra mô hình.

5. Huấn luyện mô hình: RandomForestClassifier được sử dụng để huấn luyện mô hình dựa trên các đặc trưng dữ liệu, sau đó mô hình được kiểm tra và đánh giá bằng MulticlassClassificationEvaluator, với độ chính xác được tính toán.

6. Dự đoán và kết hợp dữ liệu: Mô hình huấn luyện được sử dụng để dự đoán các giá trị cho dữ liệu thời tiết mới. Kết quả dự đoán được kết hợp với các thông tin về cây trồng và loại đất, tạo thành kết quả cuối cùng.

Pyspark SQL và Spark MLlib đóng vai trò quan trọng trong việc xử lý và phân tích dữ liệu lớn. pyspark.sql được sử dụng để đọc và xử lý dữ liệu dạng bảng (DataFrame) từ các tệp CSV, cung cấp các hàm như col(), withColumn() để thao tác với các cột trong dữ liệu, đồng thời cho phép chuẩn hóa dữ liệu và thực hiện các phép toán phức tạp. Các thao tác như chuyển đổi giá trị chuỗi thành chỉ số số

(sử dụng StringIndexer), kết hợp các cột đặc trưng thành vector (sử dụng VectorAssembler), và chuẩn hóa dữ liệu (sử dụng StandardScaler) giúp dữ liệu sẵn sàng cho các mô hình học máy. Spark MLlib được sử dụng để huấn luyện mô hình phân loại bằng RandomForestClassifier, một thuật toán học máy phổ biến để phân loại dữ liệu, cùng với MulticlassClassificationEvaluator để đánh giá độ chính xác của mô hình. Nhờ sự kết hợp của cả hai thư viện này, đoạn mã có thể xử lý và phân tích dữ liệu quy mô lớn, huấn luyện mô hình học máy hiệu quả, đồng thời cung cấp các kết quả dự đoán chính xác cho các tình huống thực tế.

```
+--------+---------+---+-----------+-----------+-------+-------+-------------+--------------+--------------+--------------+-------+---------+-----+----
|   label|Soilcolor| Ph|     QV2M-W|    QV2M-Sp|QV2M-Su|QV2M-Au|PRECTOTCORR-W|PRECTOTCORR-Sp|PRECTOTCORR-Su|PRECTOTCORR-Au|GWETTOP|CLOUD_AMT|   PS|T2M_
+--------+---------+---+-----------+-----------+-------+-------+-------------+--------------+--------------+--------------+-------+---------+-----+----
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
|Lúa mạch|     Khác|7.1|7.586666667|8.136666667|  11.43|  10.15|          9.1|   22.66666667|   117.6666667|          29.6|   0.63|    59.11|77.15|
+--------+---------+---+-----------+-----------+-------+-------+-------------+--------------+--------------+--------------+-------+---------+-----+----
only showing top 20 rows
```

Hình 4: Kết quả khuyến nghị loại cây trồng phù hợp theo tình hình thời tiết và đất đai

5 Kết luận

Trong đề tài này, chúng tôi đã xây dựng một hệ thống kết hợp nhiều công nghệ dữ liệu lớn nhằm khuyến nghị tối ưu hóa các yếu tố môi trường trong nhà kính, giúp cây trồng phát triển tốt nhất trong các điều kiện thời tiết biến động. Cụ thể, hệ thống thu thập dữ liệu thời tiết thực tế và độ ẩm đất từ các nguồn bên ngoài như Agromonitoring API và Kafka, sau đó sử dụng Apache Spark để xử lý và phân tích dữ liệu. Dựa trên dữ liệu này, mô hình học máy được huấn luyện để đưa ra các khuyến nghị về việc điều chỉnh nhiệt độ, độ ẩm không khí, áp suất và độ ẩm đất trong nhà kính, với mục tiêu tối ưu hóa các yếu tố này theo từng mùa và điều kiện đất đai cụ thể.

Một trong những ưu điểm lớn của hệ thống là khả năng xử lý dữ liệu lớn và không ngừng cập nhật từ nhiều nguồn khác nhau nhờ vào sự kết hợp của Kafka, Spark, và MongoDB. Kafka giúp stream dữ liệu thời gian thực, trong khi Spark cho phép xử lý dữ liệu một cách hiệu quả và thực hiện các phép toán phức tạp trong môi trường phân tán. Bên cạnh đó, việc sử dụng mô hình học máy như Random Forest giúp tự động đưa ra các dự đoán và khuyến nghị chính xác, hỗ trợ người sử dụng trong việc điều chỉnh các yếu tố môi trường một cách kịp thời và chính xác.

Tuy nhiên, hệ thống cũng có một số nhược điểm cần được cải thiện. Thứ nhất, mặc dù việc sử dụng dữ liệu thời gian thực từ Kafka giúp cập nhật nhanh chóng, nhưng hệ thống vẫn có thể gặp phải độ trễ trong việc xử lý dữ liệu khi khối lượng dữ liệu quá lớn hoặc không ổn định. Thứ hai, độ chính xác của các dự đoán còn phụ thuộc vào chất lượng và độ chính xác của dữ liệu đầu vào, đặc biệt là trong trường hợp thiếu sót hoặc sai lệch về dữ liệu thời tiết và độ ẩm đất. Cuối cùng, hệ thống hiện tại chỉ mới dừng lại ở các khuyến nghị cơ bản về việc điều chỉnh các yếu tố môi trường, và có thể mở rộng thêm để đưa ra các giải pháp cụ thể hơn như tối ưu hóa quá trình tưới nước, quản lý tài nguyên năng lượng trong nhà kính.

Tóm lại, hệ thống đã chứng tỏ được tính khả thi và tiềm năng trong việc ứng dụng công nghệ dữ liệu lớn để hỗ trợ nông nghiệp thông minh. Các công nghệ như Apache Kafka, Spark và MLlib đã giúp hệ thống hoạt động hiệu quả, xử lý dữ liệu lớn và đưa ra khuyến nghị chính xác. Tuy nhiên, để hệ thống hoàn thiện hơn và đáp ứng tốt hơn yêu cầu thực tế, cần tiếp tục cải tiến các thuật toán học máy, tối ưu hóa hiệu suất xử lý dữ liệu và đảm bảo chất lượng dữ liệu đầu vào.

Tài liệu

[1] Agromonitoring. *Agromonitoring - Agricultural Data API*. Accessed: 2025-01-23. 2025. URL: https://agromonitoring.com.

[2] FAO. *Food and Agriculture Organization of the United Nations*. Accessed: 2025-01-23. 2025. URL: https://www.fao.org.

[3] The Apache Software Foundation. *Apache Kafka Documentation*. Accessed: 2025-01-24. 2025. URL: https://kafka.apache.org/documentation/.

[4] The Apache Software Foundation. *Apache Spark Documentation*. Accessed: 2025-01-23. 2025. URL: https://spark.apache.org/docs/latest/.

[5] The Apache Software Foundation. *Spark MLlib Documentation*. Accessed: 2025-01-24. 2025. URL: https://spark.apache.org/mllib/.

[6] Inc. MongoDB. *PyMongo Documentation*. Accessed: 2025-01-24. 2025. URL: https://pymongo.readthedocs.io/.