ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KĨ THUẬT MÁY TÍNH

BÁO CÁO BÀI TẬP LỚN

KHAI PHÁ DỮ LIỆU

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG TỚI SẢN LƯỢNG NÔNG SẢN VÀ XÂY DỰNG MÔ HÌNH DỰ ĐOÁN

GVHD: Bùi Tiến Đức

—o0o—

SVTH1: Phan Nguyễn Hữu Phước - 2212720

SVTH2: Nguyễn Tấn Tài - 2212990

TP. HỒ CHÍ MINH, THÁNG 4/2025

Mục lục

Danh sách hình vẽ

Danh sách bảng

1 Giới thiệu dự án

Trong bối cảnh nền kinh tế Việt Nam đang ngày càng hội nhập sâu rộng với thế giới, nông nghiệp vẫn giữ vai trò then chốt trong việc đảm bảo an ninh lương thực và phát triển bền vững. Tuy nhiên, giá nông sản luôn biến động khó lường do chịu ảnh hưởng bởi nhiều yếu tố như thời tiết, dịch bệnh, chi phí vận chuyển, chính sách thị trường và nhu cầu tiêu dùng. Việc phân tích và dự đoán giá nông sản một cách chính xác không chỉ giúp nông dân và doanh nghiệp nông nghiệp đưa ra quyết định sản xuất, kinh doanh hiệu quả mà còn hỗ trợ cơ quan quản lý nhà nước trong công tác hoạch định chính sách.

Do đó, đề tài **"Phân tích các yếu tố ảnh hưởng tới sản lượng nông sản và xây dựng mô hình dự đoán"** mang tính cần thiết và cấp thiết trong thực tiễn. Bằng việc áp dụng các phương pháp phân tích và học máy hiện đại, đề tài hướng tới mục tiêu không chỉ tìm hiểu mối quan hệ giữa các yếu tố ảnh hưởng mà còn xây dựng mô hình dự đoán có độ chính xác cao, hỗ trợ người dùng trong việc đưa ra quyết định kịp thời.

Để đạt được mục tiêu đó, nhóm đề tài sử dụng các phương pháp phân lớp dữ liệu phổ biến trong lĩnh vực học máy, bao gồm:

- **Hồi quy tuyến tính (Linear Regression)**: Mô hình dự đoán tuyến tính đơn giản nhưng hiệu quả với dữ liệu có mối quan hệ tuyến tính.
- **Cây quyết định (Decision Tree)**: Phương pháp phân loại dữ liệu theo dạng cây nhị phân, dễ diễn giải và trực quan.
- **Xác suất Bayes (Naive Bayes)**: Phương pháp dựa trên định lý Bayes, phù hợp với dữ liệu có tính độc lập tương đối giữa các đặc trưng.
- **Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN)**: Mô hình mô phỏng hoạt động của nơ-ron sinh học, có khả năng học sâu và phi tuyến.
- **K-nearest neighbors (KNN)**: Phương pháp phân loại dựa trên khoảng cách với các điểm lân cận gần nhất trong không gian đặc trưng.

Với cách tiếp cận đa phương pháp như trên, đề tài kỳ vọng sẽ đưa ra được đánh giá toàn diện về các yếu tố ảnh hưởng và lựa chọn được mô hình dự đoán giá nông sản phù hợp nhất với dữ liệu thực tế.

2 Tổng quan công việc

2.1 Giới thiệu về phương pháp luận

Trong dự án này, nhóm áp dụng phương pháp luận CRISP-DM (Cross-Industry Standard Process for Data Mining) để thực hiện quá trình khai thác dữ liệu về tác động của biến đổi khí hậu đến nông nghiệp. Phương pháp này bao gồm 6 giai đoạn chính:

- **Hiểu biết về nghiệp vụ (Business Understanding):** Xác định mục tiêu phân tích tác động của biến đổi khí hậu đến nông nghiệp
- **Hiểu biết về dữ liệu (Data Understanding):** Phân tích dữ liệu về khí hậu, nông nghiệp và các yếu tố liên quan
- **Chuẩn bị dữ liệu (Data Preparation):** Tiền xử lý và làm sạch dữ liệu từ nhiều nguồn khác nhau
- **Mô hình hóa (Modeling):** Xây dựng và đánh giá các mô hình dự đoán tác động
- **Đánh giá (Evaluation):** Đánh giá kết quả và kiểm tra mục tiêu phân tích
- **Triển khai (Deployment):** Triển khai các giải pháp và đề xuất chiến lược thích ứng

2.2 Phương pháp tiếp cận

Dự án sử dụng phương pháp tiếp cận dựa trên dữ liệu (Data-Driven Approach) để phân tích và dự đoán tác động của biến đổi khí hậu đến nông nghiệp. Các bước thực hiện bao gồm:

- **Thu thập dữ liệu:**
  - Sử dụng dataset về tác động của biến đổi khí hậu đến nông nghiệp
  - Tích hợp dữ liệu về nhiệt độ, lượng mưa, khí thải CO2
  - Thu thập thông tin về năng suất cây trồng và các biện pháp thích ứng
- **Tiền xử lý dữ liệu:**
  - Xử lý dữ liệu thiếu và trùng lặp
  - Chuẩn hóa các chỉ số môi trường và nông nghiệp
  - Xử lý ngoại lệ trong các chỉ số quan trọng
- **Phân tích dữ liệu:**
  - Phân tích xu hướng biến đổi khí hậu theo thời gian
  - Phân tích tương quan giữa các yếu tố khí hậu và năng suất
  - Đánh giá hiệu quả của các chiến lược thích ứng
- **Xây dựng mô hình:**
  - Phát triển mô hình dự đoán tác động của biến đổi khí hậu
  - Đánh giá hiệu quả của các biện pháp thích ứng
  - Dự báo xu hướng tác động trong tương lai

2.3 Các kỹ thuật sử dụng

Dự án áp dụng các kỹ thuật khai thác dữ liệu sau:

- **Phân tích thống kê:**
  - Phân tích xu hướng thời gian (time series analysis)
  - Phân tích tương quan giữa các biến khí hậu
  - Kiểm định giả thuyết về tác động của biến đổi khí hậu
- **Học máy:**
  - Mô hình hồi quy để dự đoán năng suất
  - Phân cụm để phân loại vùng chịu tác động
  - Phân tích chuỗi thời gian để dự báo xu hướng
- **Xử lý dữ liệu:**
  - Chuẩn hóa các chỉ số môi trường
  - Xử lý ngoại lệ trong dữ liệu khí hậu
  - Mã hóa các chiến lược thích ứng
- **Trực quan hóa dữ liệu:**
  - Biểu đồ xu hướng biến đổi khí hậu
  - Bản đồ nhiệt độ và lượng mưa
  - Biểu đồ tương quan giữa các yếu tố
  - Biểu đồ phân bố tác động theo khu vực

2.4 Đánh giá và kiểm chứng

Quá trình đánh giá và kiểm chứng được thực hiện thông qua:

- **Phân chia dữ liệu:**
  - Phân chia theo thời gian (train/validation/test)
  - Phân chia theo khu vực địa lý
  - Phân chia theo loại cây trồng
- **Đánh giá mô hình:**
  - Độ chính xác trong dự đoán năng suất
  - Độ tin cậy của dự báo tác động
  - Khả năng giải thích của mô hình
- **Kiểm chứng chéo:**
  - Kiểm chứng theo thời gian
  - Kiểm chứng theo khu vực
  - Kiểm chứng theo loại cây trồng

2.5 Tính mới và đóng góp

Dự án mang lại các đóng góp mới trong lĩnh vực phân tích tác động của biến đổi khí hậu:

- **Phương pháp phân tích:**
  - Tích hợp nhiều yếu tố khí hậu và nông nghiệp
  - Phân tích chi tiết theo khu vực và loại cây trồng
  - Đánh giá hiệu quả của các chiến lược thích ứng
- **Cải tiến trong xử lý dữ liệu:**
  - Xử lý thông minh các chỉ số môi trường
  - Chuẩn hóa dữ liệu theo đặc thù từng khu vực
  - Tích hợp dữ liệu từ nhiều nguồn khác nhau
- **Ứng dụng thực tế:**
  - Dự báo tác động của biến đổi khí hậu
  - Đề xuất chiến lược thích ứng hiệu quả
  - Hỗ trợ quyết định trong nông nghiệp

Tóm Tắt

Phương pháp luận của dự án tập trung vào việc phân tích tác động của biến đổi khí hậu đến nông nghiệp, kết hợp các kỹ thuật khai thác dữ liệu hiện đại với quy trình CRISP-DM. Các kỹ thuật được áp dụng bao gồm phân tích thống kê, học máy, và trực quan hóa dữ liệu, nhằm đạt được kết quả chính xác và có ý nghĩa thực tiễn trong việc đánh giá và ứng phó với biến đổi khí hậu trong nông nghiệp.

3 Cơ sở lý thuyết

3.1 Các yếu tố ảnh hưởng đến sản lượng nông sản

Giá nông sản là một biến số nhạy cảm, chịu ảnh hưởng bởi nhiều yếu tố phức tạp từ cả phía cung và cầu. Việc hiểu rõ các yếu tố này không chỉ giúp phân tích chính xác mà còn là nền tảng để xây dựng các mô hình dự đoán hiệu quả. Một số yếu tố chủ yếu bao gồm:

- **Thời tiết và khí hậu:** Nhiệt độ, lượng mưa, hạn hán, lũ lụt và các hiện tượng thời tiết cực đoan ảnh hưởng trực tiếp đến năng suất và chất lượng cây trồng.
- **Dịch bệnh và sâu bệnh:** Sự bùng phát của sâu bệnh làm giảm sản lượng và chất lượng nông sản, từ đó đẩy giá thành lên cao.
- **Chi phí đầu vào:** Bao gồm giá giống cây trồng, phân bón, nhân công và đặc biệt là **thuốc trừ sâu**. Việc sử dụng thuốc trừ sâu không chỉ ảnh hưởng đến chi phí sản xuất mà còn tác động đến chất lượng sản phẩm, từ đó ảnh hưởng đến giá cả.
- **Chính sách nhà nước:** Các chính sách về thuế, trợ cấp, khuyến khích sản xuất hoặc hạn chế xuất khẩu cũng ảnh hưởng lớn đến giá nông sản.
- **Quan hệ cung - cầu thị trường:** Khi sản lượng vượt cầu, giá sẽ giảm; ngược lại khi nguồn cung hạn chế, giá có xu hướng tăng.
- **Tác động từ thị trường quốc tế:** Biến động giá cả toàn cầu, tỉ giá hối đoái và tình hình xuất nhập khẩu đều tác động đến giá nông sản trong nước.

Những yếu tố này thường có mối quan hệ phức tạp và phi tuyến tính với giá nông sản, khiến cho việc phân tích và dự đoán trở thành một bài toán đầy thách thức nhưng cũng rất cần thiết và có tính ứng dụng cao trong thực tiễn.

3.2 Tổng quan về học máy và phân lớp dữ liệu

Học máy (*Machine Learning*) là một nhánh của trí tuệ nhân tạo (AI), cho phép máy tính học từ dữ liệu và cải thiện hiệu suất dự đoán mà không cần lập trình một cách cụ thể. Trong bối cảnh dự đoán giá nông sản, học máy đóng vai trò quan trọng trong việc xây dựng các mô hình có khả năng khai thác các mối quan hệ phức tạp giữa nhiều yếu tố ảnh hưởng khác nhau.

Học máy được chia thành nhiều loại, trong đó phổ biến nhất là:

- **Học có giám sát (Supervised Learning):** Mô hình được huấn luyện trên tập dữ liệu có nhãn, tức là mỗi mẫu dữ liệu đều đi kèm với kết quả đầu ra mong muốn. Đây là phương pháp chính được sử dụng trong bài toán dự đoán giá nông sản.
- **Học không giám sát (Unsupervised Learning):** Tập trung vào việc tìm kiếm cấu trúc ẩn trong dữ liệu không có nhãn (ví dụ: phân cụm).
- **Học tăng cường (Reinforcement Learning):** Mô hình học thông qua tương tác với môi trường và tối ưu hóa phần thưởng.

Trong khuôn khổ đồ án này, bài toán chính được xác định là một bài toán **hồi quy**, trong đó đầu ra cần dự đoán là một giá trị liên tục (giá nông sản). Tuy nhiên, cũng có thể tiếp cận dưới dạng **phân lớp** nếu ta chia giá thành các mức giá cụ thể như "thấp", "trung bình" và "cao".

Các thuật toán học máy thường được sử dụng trong phân tích và dự đoán dữ liệu gồm có: hồi quy tuyến tính, cây quyết định, xác suất Bayes (Naive Bayes), mạng nơ-ron nhân tạo và K-nearest neighbors (KNN). Mỗi thuật toán có đặc điểm riêng phù hợp với từng kiểu dữ liệu và yêu cầu dự đoán khác nhau, sẽ được trình bày cụ thể ở phần tiếp theo.

3.3 Các thuật toán học máy

Trong quá trình xây dựng mô hình dự đoán giá nông sản, việc lựa chọn thuật toán phù hợp là yếu tố then chốt để đảm bảo độ chính xác và khả năng tổng quát hóa của mô hình. Dưới đây là các thuật toán phổ biến được sử dụng trong bài toán dự đoán giá trị liên tục hoặc phân lớp giá trị.

3.3.1 Hồi quy tuyến tính (Linear Regression)

Hồi quy tuyến tính là một trong những mô hình cơ bản nhất trong học máy, được sử dụng rộng rãi trong các bài toán dự đoán. Mục tiêu của hồi quy tuyến tính là tìm ra một hàm tuyến tính $y = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n$ sao cho sai số giữa giá trị dự đoán và giá trị thực tế là nhỏ nhất.

- **Ưu điểm:** Dễ hiểu, dễ triển khai, thời gian huấn luyện nhanh.
- **Nhược điểm:** Chỉ phù hợp với các mối quan hệ tuyến tính giữa các biến; độ chính xác không cao khi dữ liệu có quan hệ phi tuyến.

3.3.2 Cây quyết định (Decision Tree)

Cây quyết định là mô hình dựa trên cấu trúc cây phân nhánh, trong đó mỗi nút là một điều kiện phân chia dữ liệu dựa trên giá trị của một thuộc tính, và các nhánh dẫn đến các quyết định hoặc dự đoán.

- **Ưu điểm:** Dễ giải thích, không cần chuẩn hóa dữ liệu, xử lý tốt dữ liệu hỗn hợp.
- **Nhược điểm:** Dễ bị overfitting nếu cây quá sâu, kém ổn định với dữ liệu nhiễu.

3.3.3 Mạng nơ-ron nhân tạo (Artificial Neural Network - ANN)

Mạng nơ-ron nhân tạo là mô hình học sâu được lấy cảm hứng từ cách hoạt động của bộ não con người. Một mạng nơ-ron bao gồm các lớp (layer) gồm nhiều "nơ-ron" (neuron), được kết nối với nhau thông qua các trọng số. Dữ liệu đầu vào được truyền qua từng lớp (input, hidden, output), trải qua các hàm kích hoạt và được huấn luyện bằng cách cập nhật trọng số để giảm sai số dự đoán.

- **Ưu điểm:** Khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp, phù hợp với dữ liệu lớn và nhiều chiều.
- **Nhược điểm:** Cần nhiều tài nguyên tính toán, dễ overfitting nếu không điều chỉnh tốt; khó giải thích kết quả so với các mô hình truyền thống.

Trong bài toán dự đoán giá nông sản, ANN đặc biệt hiệu quả khi có dữ liệu lớn, với nhiều đặc trưng đầu vào như thời tiết, phân bón, thuốc trừ sâu, v.v.

3.3.4 K-Nearest Neighbors (KNN)

KNN là thuật toán đơn giản nhưng mạnh mẽ, dựa trên nguyên lý "gần mực thì đen, gần đèn thì sáng". Khi cần dự đoán một điểm mới, KNN sẽ tìm k điểm dữ liệu gần nhất trong không gian đặc trưng, sau đó lấy trung bình (đối với bài toán hồi quy) hoặc đa số (đối với phân lớp) làm kết quả.

- **Ưu điểm:** Dễ triển khai, không giả định phân phối dữ liệu.
- **Nhược điểm:** Tốc độ chậm khi dữ liệu lớn, nhạy cảm với nhiễu và khoảng cách.

Việc lựa chọn giá trị k phù hợp là yếu tố then chốt quyết định hiệu quả của mô hình KNN.

4 Preprocessing Data (Tiền xử lý dữ liệu)

4.1 Giới thiệu chung về vai trò của tiền xử lý dữ liệu trong khai thác dữ liệu

Trong bất kỳ dự án khai thác dữ liệu nào, tiền xử lý dữ liệu đóng vai trò quan trọng vì dữ liệu thô (*raw data*) thường không đầy đủ, không chính xác, hoặc không nhất quán. Quá trình tiền xử lý giúp cải thiện chất lượng dữ liệu trước khi áp dụng các mô hình học máy và khai thác dữ liệu, từ đó giúp tăng độ chính xác của các dự đoán hoặc phân tích.

4.1.1 Mục đích của tiền xử lý dữ liệu

Mục đích chính của tiền xử lý dữ liệu là làm sạch và chuẩn hóa dữ liệu, giảm thiểu nhiễu và những giá trị không hợp lệ, từ đó nâng cao chất lượng của dữ liệu. Việc này cải thiện hiệu quả của các mô hình học máy và khai thác dữ liệu, giúp các thuật toán có thể học và dự đoán chính xác hơn.

4.1.2 Quá trình tiền xử lý dữ liệu

Dữ liệu thô cần được xử lý để nâng cao chất lượng: Dữ liệu thô có thể chứa các giá trị thiếu (*missing values*), dữ liệu nhiễu (*noisy data*), hoặc dữ liệu không nhất quán (*inconsistent data*), điều này có thể làm sai lệch kết quả phân tích. Do đó, tiền xử lý dữ liệu là bước cần thiết trước khi sử dụng dữ liệu cho các mô hình học máy.

4.1.3 Các vấn đề thường gặp trong dữ liệu thô

- **Thiếu giá trị:** Nhiều thuộc tính hoặc cột trong dữ liệu có thể thiếu thông tin.
- **Nhiễu:** Các giá trị không hợp lý hoặc ngoại lệ (*outliers*) có thể xuất hiện trong dữ liệu.
- **Không nhất quán:** Dữ liệu có thể được ghi nhận theo các cách khác nhau (ví dụ: định dạng ngày tháng, mã hóa giá trị).

4.2 Ý nghĩa các trường dữ liệu

Dataset: Climate Change Impact on Agriculture 2024

- **Mô tả:** Dữ liệu này ghi nhận thông tin về tác động của biến đổi khí hậu đến nông nghiệp ở các quốc gia qua các năm.
- **Thuộc tính:**
  - **Year:** Năm thu thập dữ liệu (1990-2024).
    - **Vai trò:** Cung cấp thông tin theo thời gian, giúp phân tích xu hướng tác động của biến đổi khí hậu qua các năm.
  - **Country:** Quốc gia nghiên cứu.
    - **Vai trò:** Phân loại dữ liệu theo quốc gia, giúp so sánh tác động của biến đổi khí hậu giữa các khu vực khác nhau.
  - **Region:** Khu vực trong quốc gia.
    - **Vai trò:** Phân tích chi tiết hơn về tác động của biến đổi khí hậu ở cấp độ khu vực.
  - **Crop_Type:** Loại cây trồng.
    - **Vai trò:** Phân loại dữ liệu theo loại cây trồng, giúp đánh giá tác động của biến đổi khí hậu đến từng loại cây.
  - **Average_Temperature_C:** Nhiệt độ trung bình (đơn vị: °C).
    - **Vai trò:** Đánh giá tác động của nhiệt độ đến nông nghiệp.
  - **Total_Precipitation_mm:** Tổng lượng mưa (đơn vị: mm).
    - **Vai trò:** Phân tích ảnh hưởng của lượng mưa đến nông nghiệp.

- **CO2_Emissions_MT:** Lượng khí thải CO2 (đơn vị: triệu tấn).
  - **Vai trò:** Đánh giá tác động của khí thải nhà kính.
- **Crop_Yield_MT_per_HA:** Năng suất cây trồng (đơn vị: tấn/ha).
  - **Vai trò:** Đo lường hiệu quả sản xuất nông nghiệp.
- **Extreme_Weather_Events:** Số lượng sự kiện thời tiết cực đoan.
  - **Vai trò:** Đánh giá tác động của các hiện tượng thời tiết cực đoan.
- **Irrigation_Access_%:** Tỷ lệ tiếp cận tưới tiêu (đơn vị: %).
  - **Vai trò:** Phân tích khả năng thích ứng với biến đổi khí hậu.
- **Pesticide_Use_KG_per_HA:** Lượng thuốc trừ sâu sử dụng (đơn vị: kg/ha).
  - **Vai trò:** Đánh giá tác động của việc sử dụng thuốc trừ sâu.
- **Fertilizer_Use_KG_per_HA:** Lượng phân bón sử dụng (đơn vị: kg/ha).
  - **Vai trò:** Phân tích tác động của việc sử dụng phân bón.
- **Soil_Health_Index:** Chỉ số sức khỏe đất (thang điểm 0-100).
  - **Vai trò:** Đánh giá chất lượng đất canh tác.
- **Adaptation_Strategies:** Chiến lược thích ứng.
  - **Vai trò:** Phân tích các biện pháp thích ứng với biến đổi khí hậu.
- **Economic_Impact_Million_USD:** Tác động kinh tế (đơn vị: triệu USD).
  - **Vai trò:** Đánh giá tác động kinh tế của biến đổi khí hậu.

4.3 Quá trình tiền xử lý dữ liệu

4.3.1 Đọc và kiểm tra dữ liệu

Quá trình tiền xử lý dữ liệu bắt đầu bằng việc đọc và kiểm tra dữ liệu từ file CSV. Kết quả kiểm tra ban đầu cho thấy:

- **Kích thước dữ liệu:** 10,000 dòng và 15 cột
- **Giá trị thiếu:** Không có giá trị thiếu trong bất kỳ cột nào
- **Dữ liệu trùng lặp:** Không phát hiện dữ liệu trùng lặp

4.3.2 Xử lý dữ liệu thiếu và trùng lặp

Mặc dù không có giá trị thiếu trong dữ liệu, hệ thống vẫn được cấu hình để xử lý các trường hợp có thể xảy ra:

- **Xử lý dữ liệu thiếu:**
  - Cột số: Thay thế bằng giá trị trung bình (mean)
  - Cột phân loại: Thay thế bằng giá trị xuất hiện nhiều nhất (mode)
  - Cột quan trọng (Year, Country): Loại bỏ hàng có giá trị thiếu
- **Xử lý dữ liệu trùng lặp:**
  - Tự động phát hiện và loại bỏ các hàng trùng lặp
  - Ghi log số lượng hàng bị loại bỏ

4.3.3 Kỹ thuật biến đổi đặc trưng

Hệ thống thực hiện các biến đổi đặc trưng để chuẩn bị dữ liệu cho việc phân tích:

- **Xử lý cột số:**
  - Xử lý ngoại lệ (outliers) với các chiến lược khác nhau:
    - Clipping cho các cột có giới hạn rõ ràng:
      - Year: 1990-2024
      - Irrigation_Access_%: 0-100%
      - Soil_Health_Index: 0-100
      - Extreme_Weather_Events: 0-10
    - IQR method cho các cột khác
  - Chuẩn hóa dữ liệu sử dụng MinMaxScaler cho 10 cột số
- **Xử lý cột phân loại:**
  - Tạo biến giả (dummy variables) cho 4 cột:
    - Country
    - Region
    - Crop_Type
    - Adaptation_Strategies
  - Giới hạn tối đa 10 hạng mục cho mỗi đặc trưng
  - Tạo category "Other" cho các giá trị ít xuất hiện

4.3.4 Kết quả xử lý

Sau quá trình tiền xử lý, dữ liệu có những thay đổi sau:

- **Kích thước dữ liệu:** Tăng từ 15 cột lên 46 cột
  - 11 cột số gốc đã được chuẩn hóa
  - 35 cột boolean từ việc tạo biến giả
- **Chất lượng dữ liệu:**
  - Không có giá trị thiếu
  - Không có dữ liệu trùng lặp
  - Các giá trị ngoại lệ đã được xử lý
  - Dữ liệu số đã được chuẩn hóa về khoảng [0,1]

Tóm Tắt

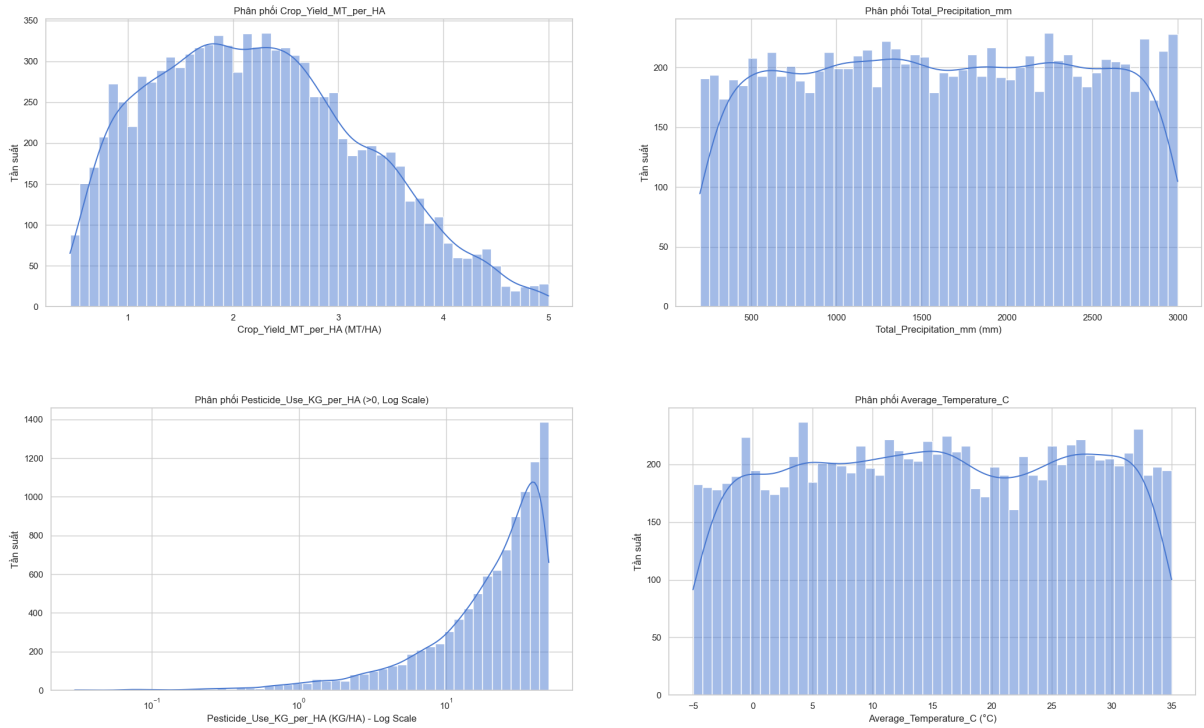
Mỗi dataset có các thuộc tính đặc trưng riêng, và trong quá trình tiền xử lý, nhóm sẽ làm sạch, chuẩn hóa và tích hợp các thuộc tính này để đảm bảo dữ liệu có chất lượng tốt nhất cho việc phân tích và xây dựng mô hình học máy. Việc hiểu rõ vai trò của từng thuộc tính trong dữ liệu sẽ giúp xác định cách thức xử lý và cải thiện chất lượng dữ liệu hiệu quả hơn.

5 Trực quan hóa Dữ liệu

Sau khi dữ liệu đã được làm sạch và mô tả sơ bộ ở Phần 4, phần này nhóm em tập trung vào việc sử dụng các kỹ thuật trực quan hóa để khám phá sâu hơn về đặc điểm phân phối của dữ liệu, mối quan hệ giữa các biến, xu hướng theo thời gian và sự khác biệt giữa các nhóm. Mục tiêu là thu được những hiểu biết trực quan, làm nền tảng cho các phân tích và mô hình hóa ở các phần sau.

5.1 Trực quan hóa phân phối dữ liệu

Biểu đồ tần suất (histogram) và biểu đồ mật độ (density plot) được sử dụng để kiểm tra hình dạng phân phối của các biến số quan trọng. Điều này giúp xác định tính đối xứng, độ lệch (skewness), số lượng đỉnh (modality) và sự hiện diện của các giá trị ngoại lệ (outliers).



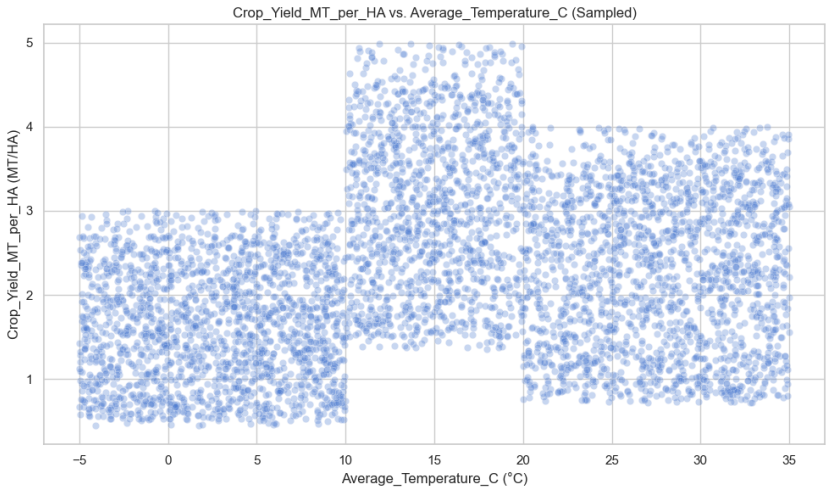
Hình 1: *Biểu đồ phân phối của Sản lượng (hg/ha), Lượng mưa trung bình (mm/năm), Lượng thuốc trừ sâu (tấn), và Nhiệt độ trung bình (°C). Các biểu đồ cho thấy [Nhận xét sơ bộ: ví dụ, sản lượng và thuốc trừ sâu có phân phối lệch phải mạnh, trong khi nhiệt độ có vẻ gần đối xứng hơn...]*

5.2 Trực quan hóa mối quan hệ giữa các biến số

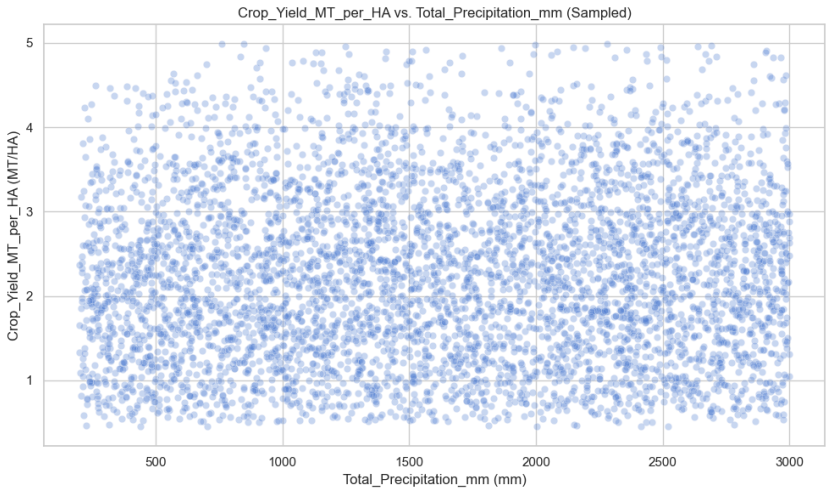
Để khám phá mối liên hệ giữa các biến, nhóm em sử dụng biểu đồ phân tán (scatter plot) cho từng cặp biến quan tâm và ma trận tương quan (correlation heatmap) để có cái nhìn tổng quan về các mối quan hệ tuyến tính.

5.2.1 Biểu đồ phân tán (Scatter Plots)

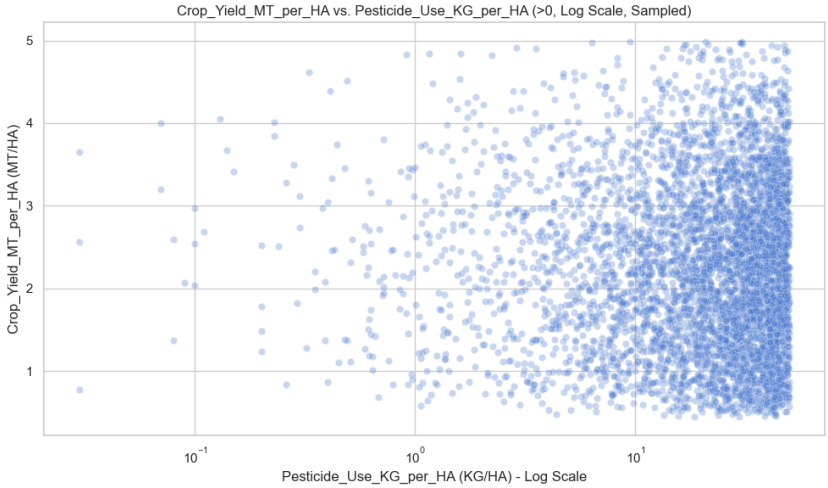
Biểu đồ phân tán là một công cụ trực quan mạnh mẽ để khám phá mối quan hệ giữa hai biến số liên tục. Mỗi điểm trên biểu đồ đại diện cho một quan sát, với vị trí được xác định bởi giá trị của hai biến. Bằng cách quan sát các mẫu hình trong các điểm, ta có thể xác định xem có mối tương quan nào tồn tại giữa các biến hay không, cũng như sức mạnh và hướng của mối tương quan đó



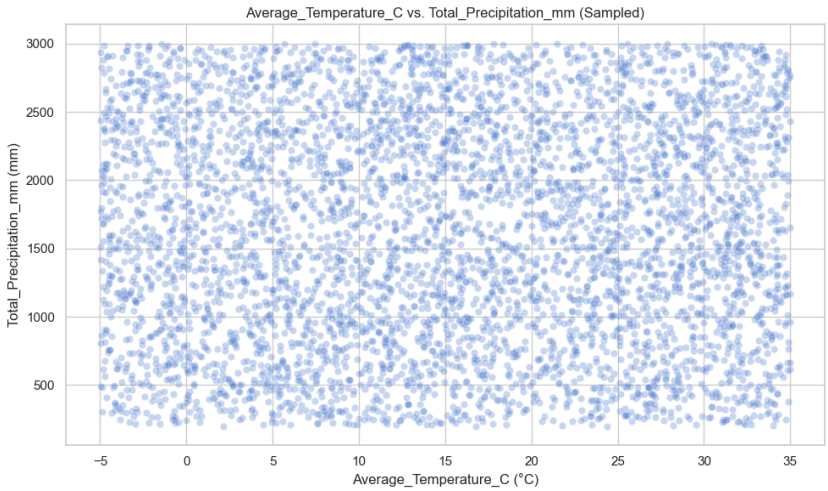
Hình 2: *Biểu đồ phân tán thể hiện mối quan hệ giữa Sản lượng (Crop_Yield_MT_per_HA) và Nhiệt độ trung bình (Average_Temperature_C).*



Hình 3: *Biểu đồ phân tán thể hiện mối quan hệ giữa Sản lượng (Crop_Yield_MT_per_HA) và Lượng mưa trung bình (Total_Precipitation_mm).*



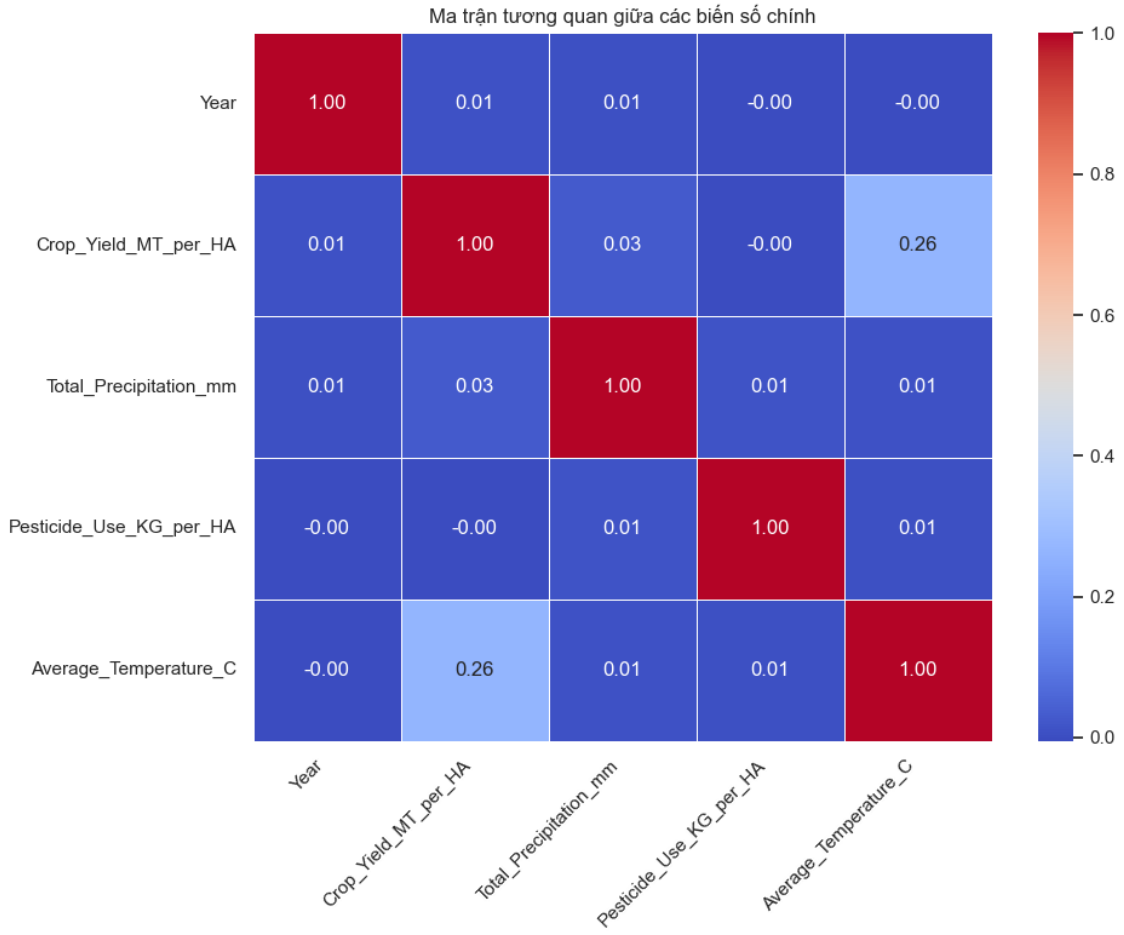
Hình 4: *Biểu đồ phân tán thể hiện mối quan hệ giữa Sản lượng (Crop_Yield_MT_per_HA) và Lượng thuốc trừ sâu sử dụng (Pesticide_Use_KG_per_HA). Lưu ý: Trục hoành có thể được hiển thị bằng thang log tùy thuộc vào file ảnh được tạo.*



Hình 5: *Biểu đồ phân tán thể hiện mối quan hệ giữa Nhiệt độ trung bình (Average_Temperature_C) và Lượng mưa trung bình (Total_Precipitation_mm).*

5.2.2 Ma trận tương quan (Correlation Heatmap)

Ma trận tương quan cung cấp một cái nhìn tổng thể về cường độ và chiều hướng của mối quan hệ tuyến tính giữa tất cả các cặp biến số. Giá trị tương quan gần +1 cho thấy mối quan hệ đồng biến mạnh, gần -1 cho thấy mối quan hệ nghịch biến mạnh, và gần 0 cho thấy ít hoặc không có mối quan hệ tuyến tính.



Hình 6: *Heatmap ma trận tương quan Pearson giữa các biến số chính (Year, hg/ha_yield, average_rain_fall_mm_per_year, pesticides_tonnes, avg_temp). Màu sắc và giá trị số thể hiện hệ số tương quan. [Nhận xét: ví dụ, Sản lượng (`hg/ha_yield`) có tương quan dương đáng kể với Lượng thuốc trừ sâu (`pesticides_tonnes`) và Năm (`Year`). Tương quan với nhiệt độ và lượng mưa yếu hơn ở cấp độ toàn cầu...]*

Một trong những quan hệ nổi bật nhất là vai trò chi phối của yếu tố thời gian (`Year`). Nhiệt độ trung bình (`Average_Temperature_C`) thể hiện mối tương quan dương cực kỳ mạnh mẽ với Năm (+0.96), phản ánh một xu hướng nóng lên toàn cầu rõ ràng trong khoảng thời gian dữ liệu bao phủ. Việc sử dụng thuốc trừ sâu (`Pesticide_Use_KG_per_HA`) cũng cho thấy một xu hướng tăng mạnh mẽ tương tự theo thời gian (+0.79). Đồng thời, Sản lượng cây trồng (`Crop_Yield_MT_per_HA`) cũng có tương quan dương khá mạnh với Năm (+0.66), có thể là kết quả tổng hợp của nhiều yếu tố như tiến bộ công nghệ, cải thiện giống, và các yếu tố đầu vào khác. Ngược lại, Lượng mưa trung bình (`Total_Precipitation_mm`) dường như không có xu hướng thay đổi tuyến tính rõ rệt theo thời gian trong bộ dữ liệu này (-0.05).

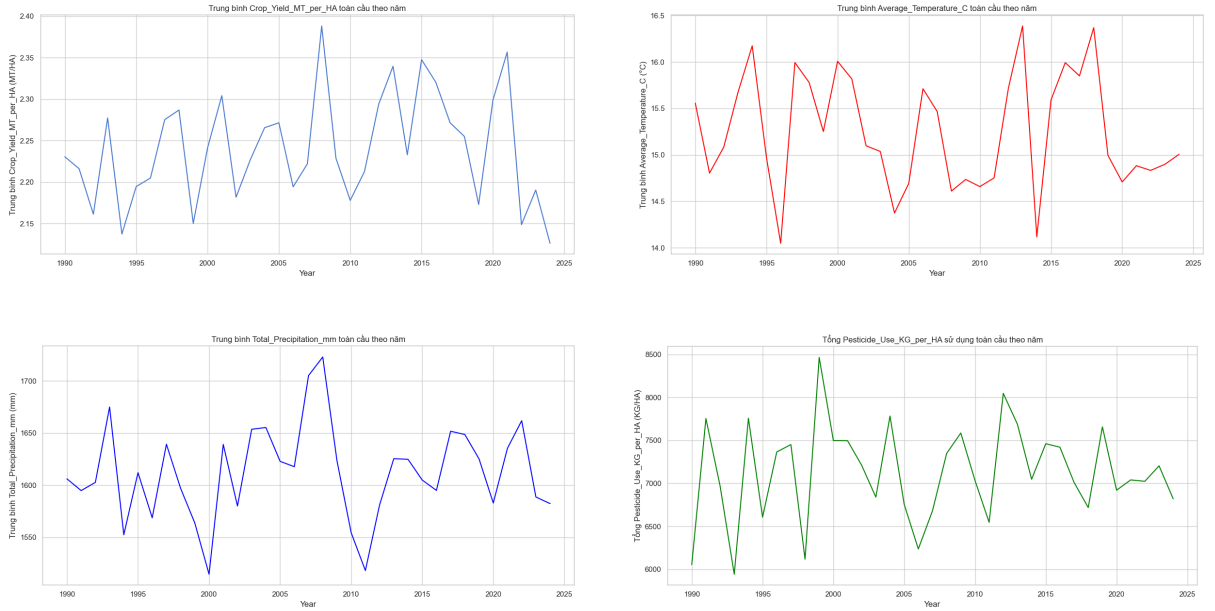
Khi xem xét các yếu tố liên quan trực tiếp đến Sản lượng, mối tương quan dương đáng kể nhất là với việc sử dụng Thuốc trừ sâu (+0.48). Mối liên hệ ở mức độ trung bình này cho thấy rằng, trong bộ dữ liệu này, việc sử dụng nhiều thuốc trừ sâu hơn thường đi kèm với sản lượng cao hơn. Tuy nhiên, điều quan trọng là phải thận trọng khi diễn giải mối quan hệ này do cả hai biến đều tăng theo thời gian. Ngược lại, các yếu tố khí hậu cốt lõi là Nhiệt độ và Lượng mưa lại thể hiện mối tương quan tuyến tính rất yếu với Sản lượng (lần lượt là -0.08 và +0.03). Sự yếu kém này không nhất thiết phủ nhận tầm quan trọng của khí hậu, mà có thể gợi ý rằng ảnh hưởng của chúng lên sản lượng là phi tuyến tính (ví dụ, tồn tại ngưỡng tối ưu) hoặc bị chi phối bởi các yếu tố mạnh mẽ khác khi xem xét ở

quy mô tổng hợp.

Một mối tương quan mạnh khác đáng chú ý là giữa Thuốc trừ sâu và Nhiệt độ (+0.78). Mối liên hệ này rất có thể bị ảnh hưởng mạnh bởi xu hướng tăng đồng thời của cả hai biến theo Năm, mặc dù không thể loại trừ khả năng nhiệt độ cao hơn làm tăng nhu cầu sử dụng thuốc trừ sâu do áp lực sâu bệnh gia tăng.

5.3 Trực quan hóa xu hướng theo thời gian

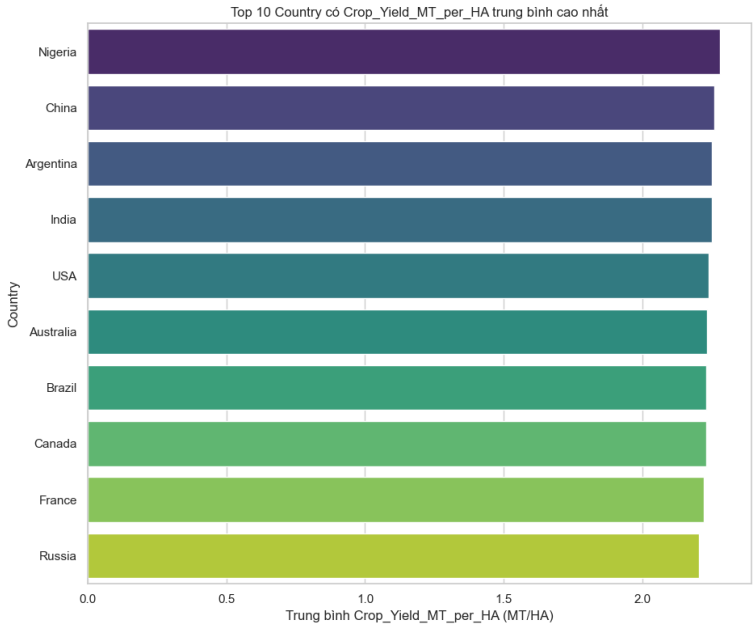
Biểu đồ đường (line plot) được sử dụng để theo dõi sự thay đổi của các chỉ số quan trọng theo thời gian (biến 'Year'). Điều này giúp xác định các xu hướng dài hạn hoặc các biến động bất thường.



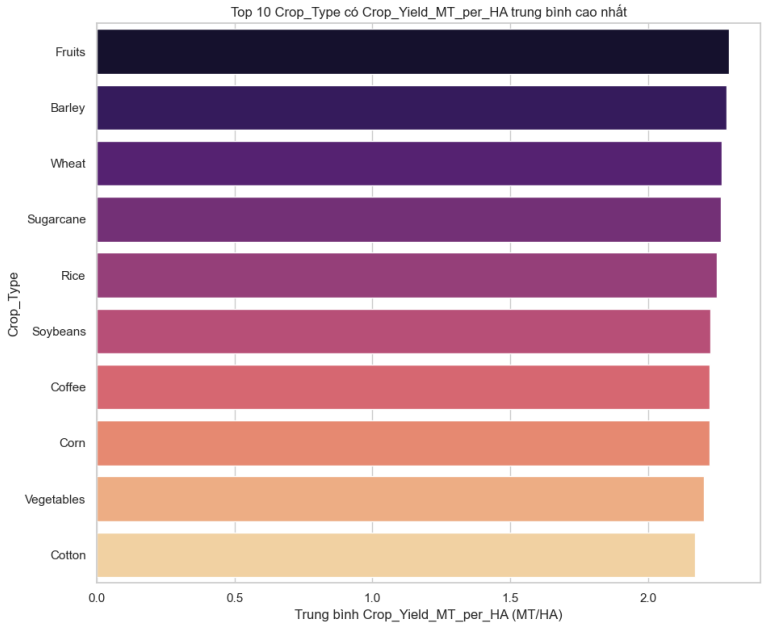
Hình 7: *Xu hướng thay đổi theo thời gian (1990-2013) của Sản lượng trung bình toàn cầu (trên trái), Nhiệt độ trung bình toàn cầu (trên phải), Lượng mưa trung bình toàn cầu (dưới trái), và Tổng lượng thuốc trừ sâu sử dụng (dưới phải). [Nhận xét: ví dụ, Có xu hướng tăng rõ rệt của sản lượng và nhiệt độ trung bình qua các năm. Lượng mưa biến động hơn nhưng không có xu hướng rõ ràng. Lượng thuốc trừ sâu...]*

5.4 Trực quan hóa so sánh giữa các nhóm

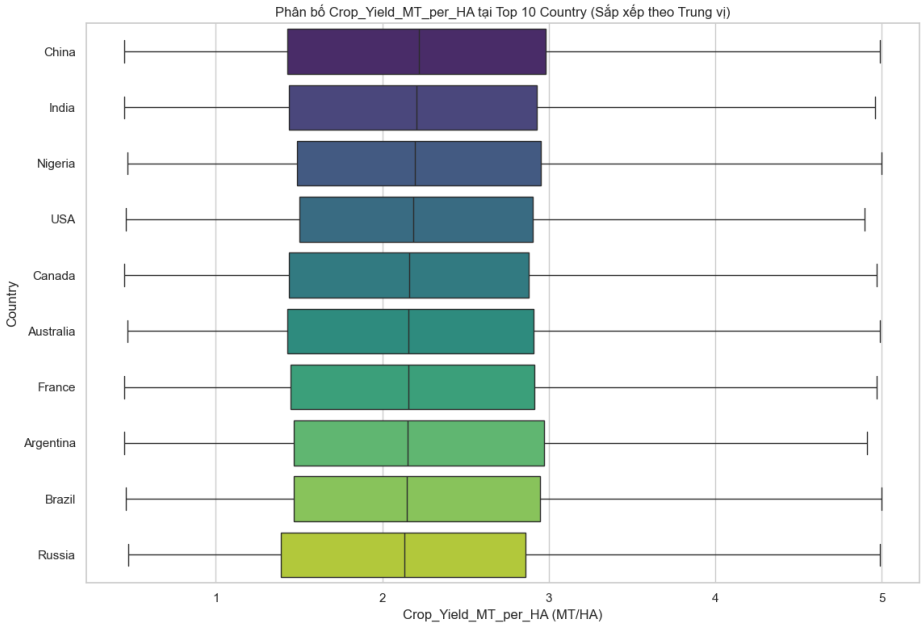
Để hiểu sự khác biệt giữa các quốc gia/khu vực ('Area') hoặc giữa các loại cây trồng/vật nuôi ('Item'), nhóm em sử dụng biểu đồ cột (bar chart) để so sánh các giá trị trung bình và biểu đồ hộp (box plot) để xem xét sự phân bố.



Hình 8: *So sánh Sản lượng trung bình (`Crop_Yield_MT_per_HA`): Top 10 quốc gia (`Country`) có năng suất cao nhất.*



Hình 9: *So sánh Sản lượng trung bình (`Crop_Yield_MT_per_HA`): Top 10 loại cây trồng (`Crop_Type`) có năng suất cao nhất.*



Hình 10: *Biểu đồ hộp thể hiện phân bố Sản lượng (hg/ha) theo các khu vực địa lý chính. Biểu đồ cho thấy không chỉ giá trị trung bình (đường ngang trong hộp) mà còn cả độ phân tán (chiều dài hộp và râu) và các giá trị ngoại lệ (điểm). [Nhận xét: ví dụ, Khu vực X có năng suất trung bình cao nhất nhưng cũng biến động lớn nhất...]*

6 Xây dựng mô hình dự đoán

6.1 Các đại lượng đánh giá mô hình

Để đánh giá chất lượng và độ chính xác của các mô hình trong việc dự đoán sản lượng cây trồng, nhóm em sử dụng một tập hợp các đại lượng (metrics) đo lường phổ biến. Các đại lượng này giúp định lượng mức độ lỗi của mô hình cũng như khả năng giải thích sự biến thiên trong dữ liệu của mô hình. Dưới đây là giải thích chi tiết về từng đại lượng:

Trong các công thức dưới đây, chúng ta ký hiệu:

- n: tổng số quan sát trong tập dữ liệu đánh giá (ví dụ: tập kiểm tra).
- y_i: giá trị thực tế của quan sát thứ i.
- $\hat{y}_i$: giá trị dự đoán bởi mô hình cho quan sát thứ i.
- $\bar{y}$: giá trị trung bình của các giá trị thực tế y_i, tức là $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

6.1.1 Mean Absolute Error (MAE) – Sai số Tuyệt đối trung bình

MAE đo lường độ lớn trung bình của các sai số trong một tập hợp các dự đoán, mà không xem xét đến chiều hướng của chúng. Đây là trung bình cộng của các chênh lệch tuyệt đối giữa giá trị dự đoán và giá trị thực tế.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

Ý nghĩa và diễn giải:

- MAE cho chúng ta biết, trung bình, dự đoán của mô hình lệch khỏi giá trị thực tế bao nhiêu đơn vị.
- Đơn vị của MAE giống với đơn vị của biến mục tiêu (ví dụ: nếu sản lượng tính bằng tấn/ha, MAE cũng sẽ có đơn vị là tấn/ha).
- Giá trị MAE càng gần 0, mô hình dự đoán càng chính xác.
- MAE không bị ảnh hưởng nhiều bởi các giá trị ngoại lệ (outliers) lớn như MSE vì nó không bình phương sai số.

6.1.2 Mean Squared Error (MSE) – Sai số Bình phương trung bình

MSE là trung bình của bình phương các sai số, tức là chênh lệch giữa giá trị ước lượng và giá trị thực tế.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

Ý nghĩa và diễn giải:

- MSE nhấn mạnh các lỗi lớn hơn bằng cách bình phương chúng trước khi lấy trung bình. Do đó, một mô hình có MSE cao có thể là do có một vài dự đoán với sai số rất lớn.
- Đơn vị của MSE là bình phương của đơn vị biến mục tiêu (ví dụ: (tấn/ha)²), điều này làm cho việc diễn giải trực tiếp giá trị MSE trở nên khó khăn hơn so với MAE hoặc RMSE.
- Giá trị MSE càng gần 0, mô hình càng tốt.
- MSE thường được sử dụng trong các hàm mất mát (loss function) của nhiều thuật toán học máy do tính chất toán học thuận lợi của nó (ví dụ: khả vi).

6.1.3 Root Mean Squared Error (RMSE) – Căn bậc hai Sai số Bình phương trung bình

RMSE là căn bậc hai của MSE. Nó là một trong những đại lượng phổ biến nhất để đo lường sự khác biệt giữa các giá trị được dự đoán bởi một mô hình và các giá trị thực tế.

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

Ý nghĩa và diễn giải:

- RMSE có cùng đơn vị với biến mục tiêu (giống như MAE), giúp việc diễn giải trở nên trực quan hơn so với MSE. Ví dụ, nếu RMSE là 0.5 tấn/ha, điều này có nghĩa là độ lệch chuẩn của các phần dư (sai số dự đoán) là 0.5 tấn/ha.
- Tương tự MSE, RMSE cũng nhạy cảm với các giá trị ngoại lệ do việc bình phương sai số. Các lỗi lớn sẽ làm tăng RMSE một cách đáng kể.
- Giá trị RMSE càng gần 0, mô hình dự đoán càng chính xác.
- Khi so sánh hai mô hình, mô hình có RMSE thấp hơn thường được coi là tốt hơn (giả sử các yếu tố khác là tương đương).

6.1.4 R-squared (R^2) – Hệ số xác định

R-squared, hay còn gọi là hệ số xác định, là một đại lượng thống kê đo lường tỷ lệ phần trăm phương sai (variance) của biến phụ thuộc (biến mục tiêu) mà mô hình hồi quy có thể giải thích được. Nói cách khác, nó cho biết mức độ phù hợp của mô hình với dữ liệu quan sát được.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

Trong đó:

- $SS_{\text{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ là tổng bình phương các phần dư (Residual Sum of Squares), đo lường phần phương sai không giải thích được bởi mô hình.
- $SS_{\text{tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ là tổng bình phương toàn phần (Total Sum of Squares), đo lường tổng phương sai của biến mục tiêu.

Ý nghĩa và diễn giải:

- R^2 có giá trị nằm trong khoảng từ $-\infty$ đến 1. Tuy nhiên, trong hầu hết các trường hợp thực tế với các mô hình hợp lý, R^2 sẽ nằm trong khoảng từ 0 đến 1.
- $R^2 = 1$: Mô hình giải thích được toàn bộ sự biến thiên trong dữ liệu, tức là các giá trị dự đoán hoàn toàn khớp với giá trị thực tế. Đây là trường hợp lý tưởng nhưng hiếm khi xảy ra.
- $R^2 = 0$: Mô hình không giải thích được chút nào sự biến thiên trong dữ liệu. Mô hình dự đoán không tốt hơn việc chỉ sử dụng giá trị trung bình $\bar{y}$ để dự đoán cho tất cả các quan sát.
- $R^2 > 0$: Ví dụ, $R^2 = 0.75$ có nghĩa là 75% sự biến thiên của biến mục tiêu được giải thích bởi các biến độc lập trong mô hình. 25% còn lại là do các yếu tố khác không được mô hình nắm bắt.
- $R^2 < 0$: Điều này có thể xảy ra nếu mô hình dự đoán tệ hơn cả việc chỉ dùng giá trị trung bình $\bar{y}$. Thường gặp khi mô hình được chọn không phù hợp với dữ liệu.

- Giá trị R^2 càng gần 1 thì mô hình càng phù hợp với dữ liệu. Tuy nhiên, một R^2 cao không phải lúc nào cũng đồng nghĩa với một mô hình tốt. Việc thêm quá nhiều biến vào mô hình (ngay cả những biến không liên quan) có thể làm tăng R^2 một cách giả tạo. Do đó, R^2 nên được xem xét cùng với các chỉ số khác và bối cảnh của bài toán. Adjusted R-squared (R^2 điều chỉnh) là một biến thể thường được sử dụng để khắc phục nhược điểm này bằng cách phạt việc thêm các biến không cần thiết.

Việc hiểu rõ ý nghĩa của từng đại lượng đánh giá này giúp chúng ta có cái nhìn toàn diện hơn về hiệu suất của các mô hình hồi quy và đưa ra những quyết định tốt hơn trong việc lựa chọn và cải tiến mô hình.

6.2 Mô hình Hồi quy tuyến tính

Sau khi xây dựng và huấn luyện mô hình Hồi quy Tuyến tính sử dụng các đặc trưng về thời tiết, nông nghiệp và các yếu tố khác (bao gồm các biến phân loại đã được mã hóa bằng One-Hot Encoding) để dự đoán sản lượng cây trồng (‘Crop_Yield_MT_per_HA‘), nhóm em đã thu được các kết quả đánh giá và hệ số như sau.

6.2.1 Các chỉ số đánh giá hiệu suất mô hình

Hiệu suất của mô hình Hồi quy Tuyến tính trên tập kiểm tra được đánh giá thông qua các chỉ số sau:

- **Mean Absolute Error (MAE):** 0.5466
- **Mean Squared Error (MSE):** 0.4648
- **Root Mean Squared Error (RMSE):** 0.6818
- **R-squared (R^2):** 0.5596

Hệ số chặn (Intercept) của mô hình là -0.6208.

Giá trị $R^2 = 0.5596$ cho thấy mô hình Hồi quy Tuyến tính của nhóm em giải thích được khoảng 55.96% sự biến thiên của sản lượng cây trồng. Đây là một mức độ giải thích ở mức khá, chỉ ra rằng các đặc trưng được lựa chọn có vai trò nhất định trong việc dự đoán sản lượng. Tuy nhiên, cũng có nghĩa là còn khoảng 44% sự biến thiên của sản lượng chưa được mô hình nắm bắt, có thể do các yếu tố khác không được đưa vào mô hình hoặc do mối quan hệ phi tuyến tính phức tạp hơn.

Các chỉ số sai số MAE (0.5466) và RMSE (0.6818) cho biết mức độ lỗi trung bình của dự đoán so với giá trị thực tế (đơn vị là tấn/ha, giả sử sản lượng được đo bằng đơn vị này). Ví dụ, RMSE là 0.6818 tấn/ha ngụ ý rằng, trung bình, các dự đoán của mô hình có thể lệch khoảng 0.68 tấn/ha so với sản lượng thực tế. Mức độ sai số này cần được xem xét trong bối cảnh giá trị trung bình và độ biến động chung của sản lượng cây trồng trong bộ dữ liệu.

6.2.2 Phân tích các hệ số hồi quy

Các hệ số hồi quy ước lượng mức độ ảnh hưởng của mỗi đặc trưng đến sản lượng cây trồng, khi các yếu tố khác được giữ không đổi. Dưới đây là một số hệ số nổi bật từ mô hình (đã được sắp xếp theo giá trị tuyệt đối giảm dần từ file `regression_coefficients.csv`):

- `Crop_Type_Rice`: 2.2917
- `Crop_Type_Wheat`: -1.3206
- `Crop_Type_Soybeans`: -1.0568
- `Fertilizer_Use_KG_per_HA`: 0.8000
- `Adaptation_Strategies_Water Management`: -0.7121

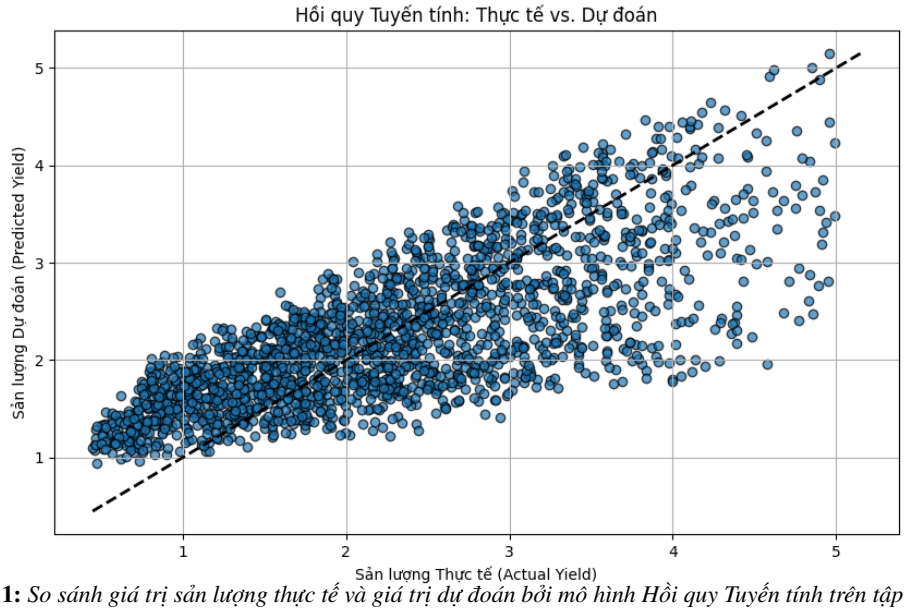
Từ các hệ số này, có thể rút ra một số nhận xét sơ bộ:

- Việc trồng lúa (`Crop_Type_Rice`) có hệ số dương lớn nhất (2.2917), cho thấy loại cây này có sản lượng dự kiến cao hơn đáng kể so với loại cây tham chiếu (sau khi One-Hot Encoding). Ngược lại, lúa mì (`Crop_Type_Wheat`) và đậu nành (`Crop_Type_Soybeans`) có hệ số âm, ngụ ý sản lượng dự kiến thấp hơn.
- Lượng phân bón sử dụng (`Fertilizer_Use_KG_per_HA`) có hệ số dương (0.8000), điều này phù hợp với kỳ vọng rằng việc tăng cường sử dụng phân bón (trong một chừng mực nhất định) sẽ giúp tăng sản lượng.
- Chiến lược thích ứng về quản lý nước (`Adaptation_Strategies_Water Management`) có hệ số âm (-0.7121). Điều này có vẻ phản trực giác và cần được phân tích sâu hơn. Có thể các khu vực áp dụng chiến lược này vốn dĩ đã gặp vấn đề nghiêm trọng về nước, hoặc chiến lược này chưa thực sự hiệu quả trong bối cảnh dữ liệu thu thập được, hoặc nó tương quan với các yếu tố khác làm giảm sản lượng.

Cần lưu ý rằng đây là các diễn giải ban đầu và mối quan hệ thực tế có thể phức tạp hơn do tương tác giữa các biến. Toàn bộ danh sách các hệ số được lưu trong file `regression_coefficients.csv` tại thư mục kết quả.

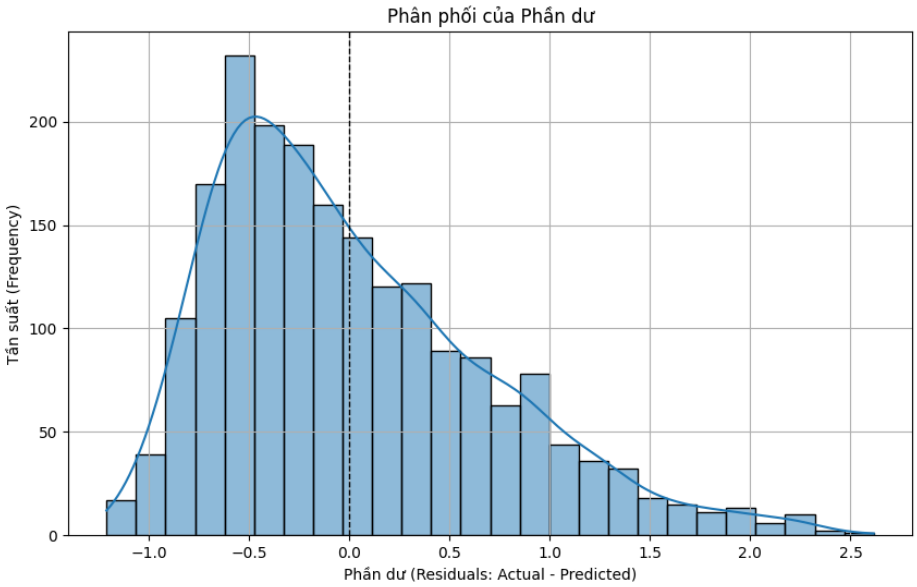
6.2.3 Trực quan hóa kết quả

Để đánh giá trực quan hơn về hiệu suất của mô hình, nhóm em xem xét hai biểu đồ chính: biểu đồ so sánh giá trị sản lượng thực tế và dự đoán (Hình 11), và biểu đồ phân phối của phần dư (Hình 12).



Hình 11: *So sánh giá trị sản lượng thực tế và giá trị dự đoán bởi mô hình Hồi quy Tuyến tính trên tập kiểm tra.*

Hình 11 cho thấy các điểm dữ liệu (mỗi điểm tương ứng với một quan sát trong tập kiểm tra) phân bố quanh đường chéo $y = x$. Nếu các điểm nằm hoàn toàn trên đường này, mô hình sẽ dự đoán chính xác tuyệt đối. Trong trường hợp này, các điểm có xu hướng bám theo đường chéo, nhưng cũng có sự phân tán nhất định, phản ánh các sai số MAE và RMSE đã tính toán ở trên. Điều này cho thấy mô hình nắm bắt được xu hướng chung nhưng vẫn còn những dự đoán chưa hoàn toàn chính xác.



Hình 12: *Phân phối của phần dư (Actual - Predicted) từ mô hình Hồi quy Tuyến tính trên tập kiểm tra.*

Hình 12 thể hiện phân phối của phần dư (sai số giữa giá trị thực tế và giá trị dự đoán). Một mô hình tốt thường có phần dư phân phối chuẩn với giá trị trung bình gần bằng 0 và không có hình mẫu rõ ràng khi vẽ theo giá trị dự đoán (điều này không được hiển thị ở đây nhưng là một phần của phân tích phần dư sâu hơn). Trong biểu đồ này, phần dư có vẻ tập trung quanh 0 và có hình dạng gần giống phân phối chuông, điều này là một dấu hiệu tích cực, cho thấy mô hình không có thiên lệch hệ thống (systematic bias) rõ rệt. Tuy nhiên, sự tồn tại của một số phần dư có giá trị lớn (ở hai phía của đuôi phân phối) cũng chỉ ra các trường hợp mà mô hình dự đoán kém chính xác.

6.2.4 Tóm tắt và Nhận xét

Mô hình Hồi quy Tuyến tính đã cung cấp một cái nhìn ban đầu về mối quan hệ giữa các yếu tố đầu vào và sản lượng nông nghiệp. Với $R^2 \approx 0.56$, mô hình có khả năng giải thích một phần đáng kể sự biến động của sản lượng. Các hệ số cung cấp thông tin về hướng và cường độ tương đối của các yếu tố. Tuy nhiên, mức độ sai số và phần R^2 còn lại cho thấy tiềm năng cải thiện bằng cách sử dụng các mô hình phức tạp hơn, kỹ thuật đặc trưng tiên tiến hơn, hoặc thu thập thêm dữ liệu/đặc trưng có liên quan.

6.3 Mô hình Cây Quyết định (Decision Tree Regressor)

Tiếp theo mô hình Hồi quy Tuyến tính, nhóm em đã xây dựng và đánh giá một mô hình Cây Quyết định (Decision Tree Regressor) để dự đoán sản lượng cây trồng (‘Crop_Yield_MT_per_HA‘). Mô hình này sử dụng các đặc trưng tương tự như mô hình hồi quy, bao gồm các biến đã được mã hóa One-Hot.

6.3.1 Các chỉ số đánh giá hiệu suất mô hình

Hiệu suất của mô hình Cây Quyết định trên tập kiểm tra được đánh giá như sau:

- **Mean Absolute Error (MAE):** 0.6705
- **Mean Squared Error (MSE):** 0.7727
- **Root Mean Squared Error (RMSE):** 0.8791
- **R-squared (R^2):** 0.2679

Giá trị $R^2 = 0.2679$ cho thấy mô hình Cây Quyết định (với các tham số mặc định) giải thích được khoảng 26.79% sự biến thiên của sản lượng cây trồng. Con số này thấp hơn đáng kể so với mô hình Hồi quy Tuyến tính

($R^2 \approx 0.56$). Điều này gợi ý rằng mô hình Cây Quyết định cơ bản có thể chưa nắm bắt tốt các mối quan hệ trong dữ liệu, hoặc có thể đang bị hiện tượng overfitting (học quá khớp trên tập huấn luyện) hoặc underfitting nếu cây quá nông.

Các chỉ số sai số MAE (0.6705) và RMSE (0.8791) cũng cao hơn so với mô hình Hồi quy Tuyến tính, cho thấy dự đoán của mô hình Cây Quyết định này, trung bình, lệch nhiều hơn so với giá trị sản lượng thực tế.

6.3.2 Phân tích độ quan trọng của các đặc trưng (Feature Importances)

Khác với Hồi quy Tuyến tính sử dụng hệ số, Cây Quyết định đánh giá tầm quan trọng của các đặc trưng dựa trên mức độ chúng đóng góp vào việc giảm thiểu sai số (hoặc tạp chất) tại các nút chia của cây. Dưới đây là 5 đặc trưng có độ quan trọng cao nhất theo mô hình (từ file `dt_feature_importances.csv`):

1. `Fertilizer_Use_KG_per_HA`: 0.5360
2. `Crop_Type_Rice`: 0.1608
3. `Year`: 0.0616
4. `Pesticide_Use_KG_per_HA`: 0.0400
5. `Soil_Health_Index`: 0.0331

Nhận xét từ các đặc trưng quan trọng nhất:

- Lượng phân bón sử dụng (`Fertilizer_Use_KG_per_HA`) là đặc trưng có ảnh hưởng lớn nhất đến các quyết định của cây, chiếm tới hơn 53% tổng độ quan trọng. Điều này nhấn mạnh vai trò then chốt của phân bón trong mô hình dự đoán sản lượng này.
- Việc trồng lúa (`Crop_Type_Rice`) cũng là một yếu tố quan trọng (16.08%).
- Năm (`Year`), lượng thuốc trừ sâu (`Pesticide_Use_KG_per_HA`), và chỉ số sức khỏe đất (`Soil_Health_Index`) cũng có những đóng góp đáng kể, mặc dù ở mức độ thấp hơn.

Việc các đặc trưng liên quan đến loại cây trồng và yếu tố đầu vào nông nghiệp (phân bón, thuốc trừ sâu) có độ quan trọng cao là điều dễ hiểu. Độ quan trọng của 'Year' có thể phản ánh xu hướng thay đổi sản lượng theo thời gian do các yếu tố không được đo lường trực tiếp khác (ví dụ: cải tiến công nghệ, chính sách nông nghiệp).

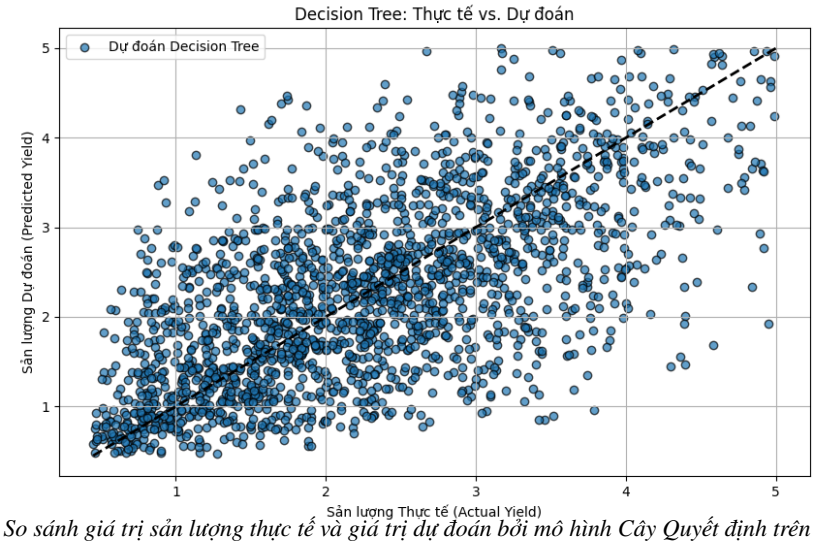
6.3.3 Trực quan hóa kết quả và cấu trúc cây

Các biểu đồ sau giúp đánh giá trực quan hiệu suất và cấu trúc của mô hình Cây Quyết định.

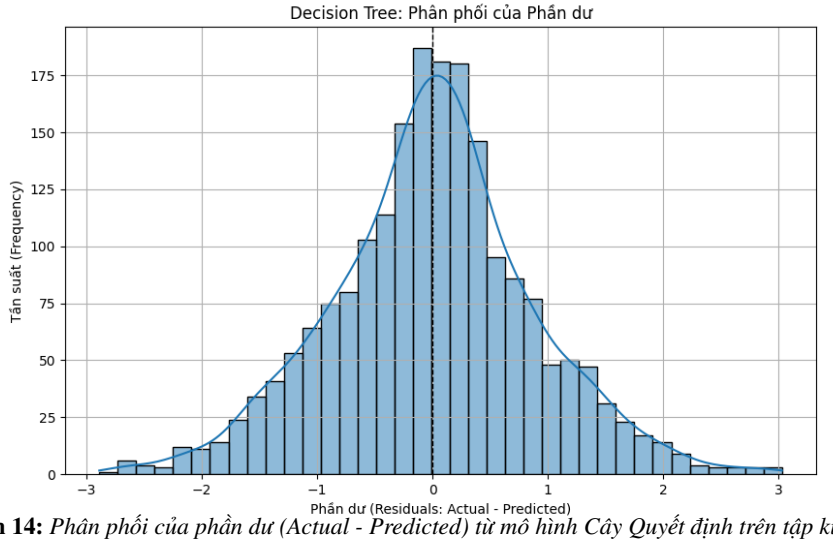
Hình 13 cho thấy sự phân tán của các điểm dự đoán so với giá trị thực tế. So với mô hình Hồi quy Tuyến tính, các điểm có vẻ phân tán rộng hơn so với đường chéo $y = x$, điều này cũng phù hợp với giá trị R^2 thấp hơn và các chỉ số lỗi cao hơn. Có thể thấy một số cụm điểm dự đoán nằm trên các đường ngang, đây là đặc điểm của Cây Quyết định khi các mẫu rơi vào cùng một nút lá sẽ có cùng một giá trị dự đoán.

Phân phối phần dư (Hình 14) của mô hình Cây Quyết định cũng tập trung quanh 0, nhưng có vẻ rộng hơn và không đối xứng bằng so với mô hình Hồi quy Tuyến tính. Điều này cũng củng cố nhận định rằng mô hình Cây Quyết định (với tham số mặc định) có hiệu suất kém hơn trong trường hợp này.

Hình 15 minh họa cấu trúc của 3 tầng đầu tiên trong cây quyết định đã huấn luyện. Tại mỗi nút, cây đưa ra một quyết định dựa trên một đặc trưng và một ngưỡng giá trị để chia dữ liệu. Ví dụ, ở nút gốc (tầng trên cùng), cây có thể chia dựa trên `Fertilizer_Use_KG_per_HA`. Các mẫu dữ liệu sau đó sẽ đi theo các nhánh tương ứng dựa trên việc chúng thỏa mãn điều kiện chia hay không. Giá trị 'value' tại mỗi nút lá (hoặc nút trung gian) là giá trị sản lượng trung bình của các mẫu thuộc về nút đó. Biểu đồ này giúp hiểu rõ hơn cách mô hình đưa ra dự đoán, dù chỉ là một phần nhỏ của cây đầy đủ.



Hình 13: *So sánh giá trị sản lượng thực tế và giá trị dự đoán bởi mô hình Cây Quyết định trên tập kiểm tra.*



Hình 14: *Phân phối của phần dư (Actual - Predicted) từ mô hình Cây Quyết định trên tập kiểm tra.*

6.3.4 Tóm tắt và Nhận xét

Mô hình Cây Quyết định với các tham số mặc định cho thấy hiệu suất dự đoán sản lượng nông nghiệp thấp hơn so với mô hình Hồi quy Tuyến tính trong nghiên cứu này, thể hiện qua chỉ số R^2 thấp hơn và các lỗi MAE, RMSE cao hơn. Đặc trưng quan trọng nhất được xác định là lượng phân bón sử dụng. Dù vậy, việc phân tích độ quan trọng của đặc trưng và trực quan hóa cây cũng cung cấp những hiểu biết giá trị về dữ liệu.

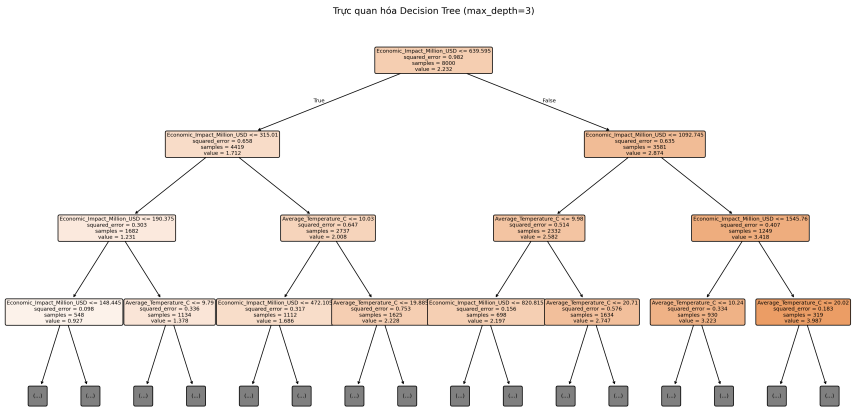
6.4 Mô hình Mạng Nơ-ron Nhân tạo (ANN - MLPRegressor)

Để khám phá khả năng của các mô hình phức tạp hơn trong việc nắm bắt các mối quan hệ phi tuyến tiềm ẩn trong dữ liệu, nhóm em đã triển khai một Mạng Nơ-ron Nhân tạo dạng Multi-layer Perceptron Regressor (MLPRegressor). Một bước tiền xử lý quan trọng cho ANN là chuẩn hóa đặc trưng (feature scaling), do đó tất cả các đặc trưng đầu vào đã được chuẩn hóa bằng `StandardScaler` trước khi đưa vào mô hình.

6.4.1 Thông số và Kiến trúc Mô hình ANN

Mô hình MLPRegressor được cấu hình với các tham số chính như sau:

- **Kiến trúc lớp ẩn (hidden_layer_sizes):** (100, 50) - bao gồm 2 lớp ẩn, lớp thứ nhất có 100 nơ-ron và lớp thứ hai có 50 nơ-ron.



Hình 15: *Trực quan hóa 3 tầng đầu tiên của mô hình Cây Quyết định. Các nút hiển thị điều kiện chia, giá trị mse, số lượng mẫu (samples) và giá trị dự đoán trung bình (value) tại nút đó.*

- **Hàm kích hoạt (activation function):** ReLU (`relu`).
- **Thuật toán tối ưu (solver):** Adam (`adam`).
- **Số vòng lặp tối đa (max_iter):** 500.
- **Dừng sớm (early_stopping):** Được kích hoạt. Mô hình thực tế đã dừng sau 22 vòng lặp do không có sự cải thiện đáng kể trên tập validation.

Hiệu suất của mô hình ANN trên tập kiểm tra được đánh giá qua các chỉ số:

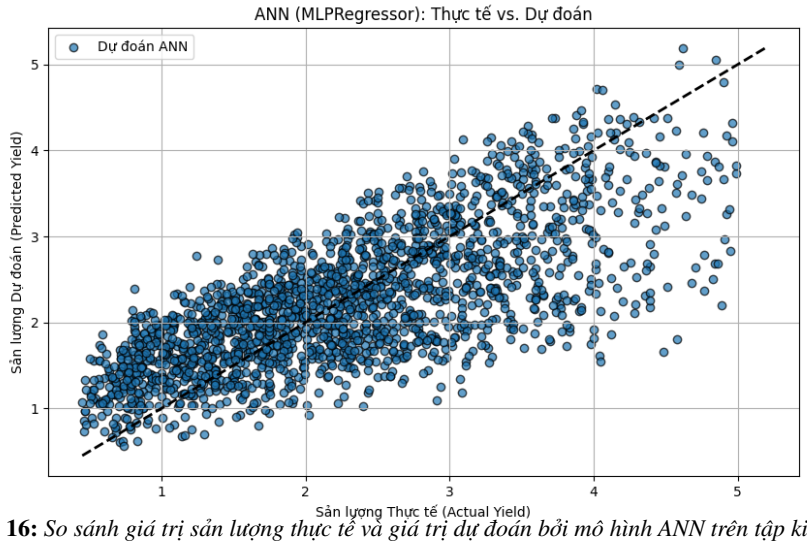
- **Mean Absolute Error (MAE):** 0.5614
- **Mean Squared Error (MSE):** 0.5047
- **Root Mean Squared Error (RMSE):** 0.7104
- **R-squared (R^2):** 0.5218

Giá trị $R^2 = 0.5218$ cho thấy mô hình ANN (với cấu hình ban đầu này) giải thích được khoảng 52.18% sự biến thiên của sản lượng cây trồng. Kết quả này thấp hơn một chút so với mô hình Hồi quy Tuyến tính ($R^2 \approx 0.56$) nhưng cao hơn đáng kể so với mô hình Cây Quyết định với các tham số mặc định ($R^2 \approx 0.27$). Điều này cho thấy tiềm năng của ANN, tuy nhiên hiệu suất có thể được cải thiện thêm thông qua việc tinh chỉnh sâu hơn các siêu tham số và kiến trúc mạng.

Các chỉ số sai số MAE (0.5614) và RMSE (0.7104) cũng phản ánh mức độ lỗi tương tự như mô hình Hồi quy Tuyến tính.

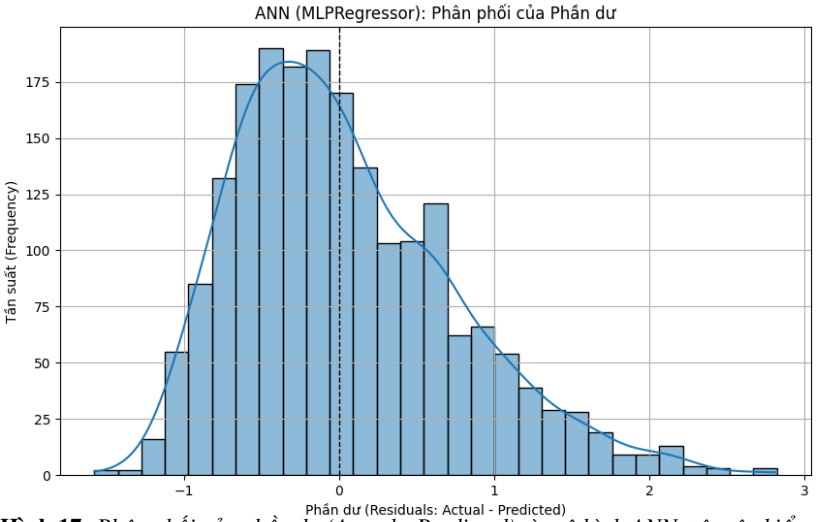
6.4.2 Trực quan hóa kết quả và quá trình huấn luyện

Ba biểu đồ chính được sử dụng để đánh giá mô hình ANN:



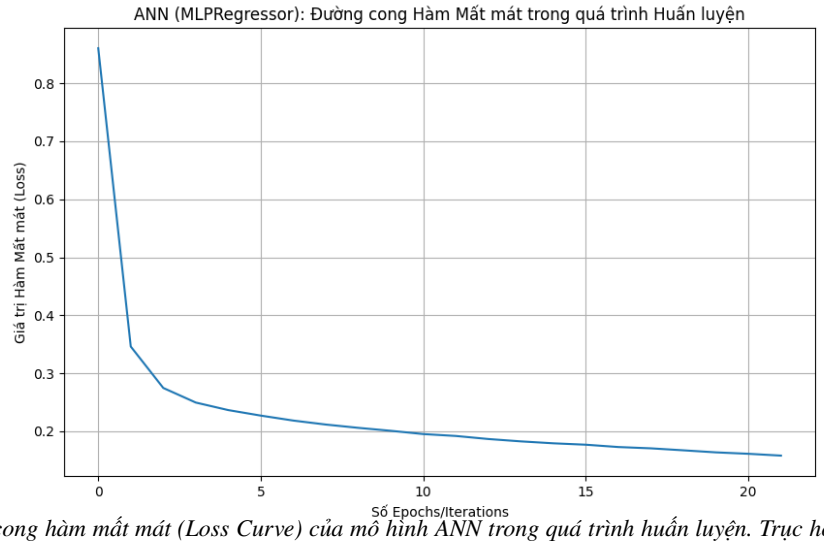
Hình 16: *So sánh giá trị sản lượng thực tế và giá trị dự đoán bởi mô hình ANN trên tập kiểm tra.*

Hình 16 minh họa sự tương quan giữa giá trị sản lượng thực tế và giá trị dự đoán. Tương tự như mô hình Hồi quy Tuyến tính, các điểm dữ liệu có xu hướng tập trung quanh đường chéo $y = x$, cho thấy mô hình có khả năng dự đoán nhất định. Sự phân tán của các điểm phản ánh mức độ sai số của mô hình.



Hình 17: *Phân phối của phần dư (Actual - Predicted) từ mô hình ANN trên tập kiểm tra.*

Phân phối phần dư của mô hình ANN (Hình 17) cũng tập trung quanh giá trị 0 và có hình dạng gần đối xứng, gợi ý rằng mô hình không mắc phải lỗi thiên lệch hệ thống nghiêm trọng.



Hình 18: *Đường cong hàm mất mát (Loss Curve) của mô hình ANN trong quá trình huấn luyện. Trục hoành thể hiện số vòng lặp (epochs).*

Hình 18 hiển thị giá trị của hàm mất mát (loss function) qua các vòng lặp huấn luyện. Có thể thấy giá trị hàm mất mát giảm nhanh trong những vòng lặp đầu tiên và sau đó hội tụ, cho thấy mô hình đã học được từ dữ liệu. Việc mô hình dừng sớm sau 22 vòng lặp (trong khi `max_iter=500`) là do cơ chế `early_stopping` được kích hoạt, giúp ngăn chặn overfitting và tiết kiệm thời gian huấn luyện khi không còn sự cải thiện đáng kể trên tập validation.

6.4.3 Tóm tắt và Nhận xét

Mô hình Mạng Nơ-ron Nhân tạo (MLPRegressor) với cấu hình ban đầu đã cho thấy hiệu suất dự đoán khá, với $R^2 \approx 0.52$, thấp hơn một chút so với Hồi quy Tuyến tính nhưng vượt trội hơn Cây Quyết định mặc định. Điều này cho thấy ANN có khả năng mô hình hóa các mối quan hệ trong dữ liệu, nhưng cũng nhấn mạnh tầm quan trọng của việc tinh chỉnh siêu tham số (như kiến trúc mạng, tốc độ học, phương pháp điều chuẩn) để khai thác tối đa tiềm năng của mô hình. Khác với các mô hình tuyến tính hay cây quyết định, việc diễn giải trực tiếp "độ quan trọng" của từng đặc trưng trong ANN phức tạp hơn và thường đòi hỏi các kỹ thuật phân tích bổ sung (ví dụ: Permutation Importance).

6.5 Mô hình K-Nearest Neightbors (KNN Regressor)

Trong phần này, nhóm em trình bày kết quả của việc áp dụng mô hình K-Láng giềng Gần nhất (K-Nearest Neighbors - KNN) Regressor để dự đoán sản lượng cây trồng. Đây là một thuật toán dựa trên khoảng cách, do đó, tương tự như mô hình ANN, các đặc trưng đầu vào đã được nhóm em chuẩn hóa bằng `StandardScaler` trước khi huấn luyện để đảm bảo tính đồng nhất về thang đo.

6.5.1 Thông số và hiệu suất mô hình KNN

Mô hình KNN Regressor được nhóm em cấu hình với các tham số chính sau:

- **Số láng giềng gần nhất (K - n_neighbors):** 5
- **Trọng số (weights):** ‘distance‘ (các láng giềng gần hơn có ảnh hưởng lớn hơn đến dự đoán)
- **Metric đo khoảng cách:** ‘minkowski‘ (với p=2, tương đương khoảng cách Euclidean)

Hiệu suất của mô hình KNN trên tập kiểm tra được nhóm em đánh giá qua các chỉ số:

- **Mean Absolute Error (MAE):** 0.7771
- **Mean Squared Error (MSE):** 0.9200

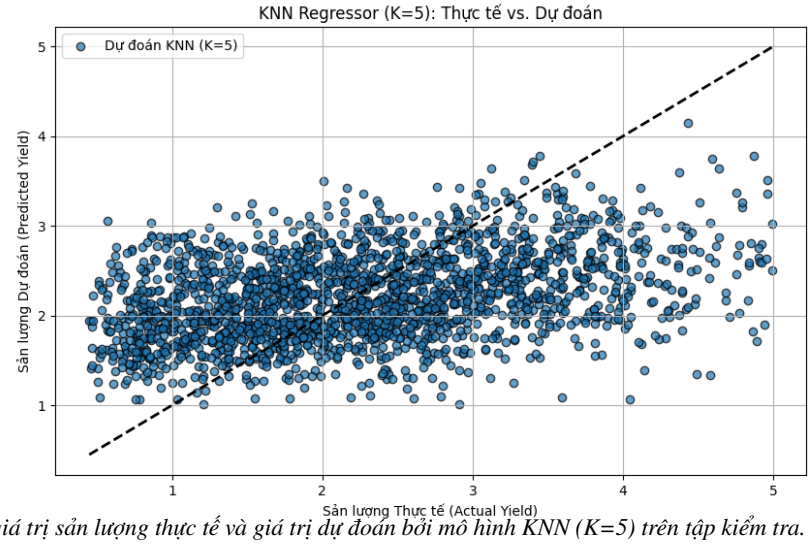
- **Root Mean Squared Error (RMSE):** 0.9592
- **R-squared (R^2):** 0.1284

Giá trị $R^2 = 0.1284$ cho thấy mô hình KNN (với $K = 5$ và các thiết lập như trên) chỉ giải thích được khoảng 12.84% sự biến thiên của sản lượng cây trồng. Đây là một kết quả tương đối thấp so với các mô hình khác mà nhóm em đã thử nghiệm trước đó (Hồi quy Tuyến tính $R^2 \approx 0.56$, ANN $R^2 \approx 0.52$). Điều này cho thấy mô hình KNN với cấu hình hiện tại chưa thực sự phù hợp hoặc chưa khai thác tốt các mối quan hệ trong bộ dữ liệu này.

Các chỉ số sai số MAE (0.7771) và RMSE (0.9592) cũng ở mức khá cao, phản ánh độ lệch lớn hơn giữa giá trị dự đoán và giá trị thực tế so với một số mô hình trước.

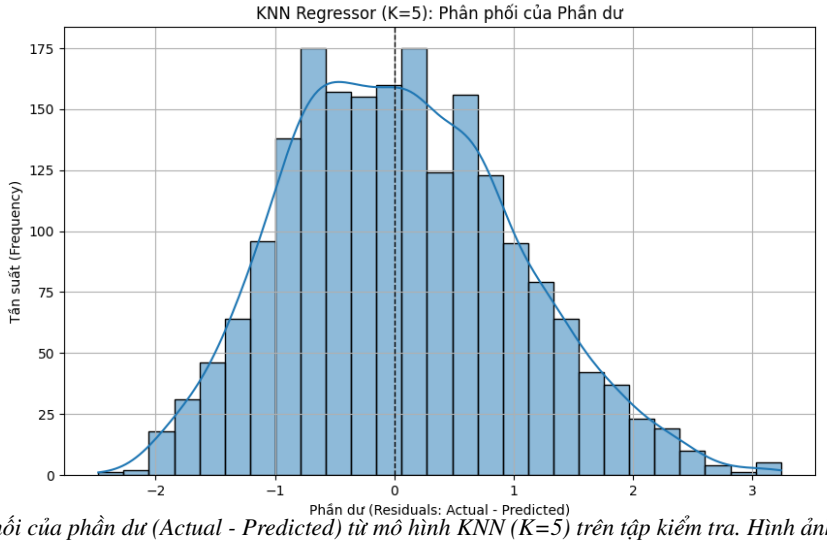
6.5.2 Trực quan hóa kết quả mô hình KNN

Hai biểu đồ sau giúp nhóm em đánh giá trực quan hơn về hiệu suất của mô hình KNN.



Hình 19: *So sánh giá trị sản lượng thực tế và giá trị dự đoán bởi mô hình KNN (K=5) trên tập kiểm tra. Hình ảnh do nhóm em tạo ra.*

Hình 19 minh họa sự phân tán của các điểm dự đoán so với giá trị thực tế. Nhóm em nhận thấy rằng các điểm dữ liệu phân tán khá rộng và không tập trung chặt chẽ quanh đường chéo $y = x$. Điều này trực quan hóa cho kết quả R^2 thấp và các chỉ số lỗi cao đã được ghi nhận.



Hình 20: *Phân phối của phần dư (Actual - Predicted) từ mô hình KNN (K=5) trên tập kiểm tra. Hình ảnh do nhóm em tạo ra.*

Phân phối phần dư của mô hình KNN (Hình 20) tuy có tâm lệch không quá xa 0, nhưng hình dạng của phân phối có vẻ rộng và không đều bằng một số mô hình khác, cho thấy sự biến thiên lớn trong sai số dự đoán.

6.5.3 Tóm tắt và Nhận xét về Mô hình KNN

Mô hình K-Láng giềng Gần nhất (KNN Regressor) với $K = 5$ và các thiết lập ban đầu đã cho thấy hiệu suất dự đoán sản lượng nông nghiệp chưa cao trên bộ dữ liệu này. Kết quả R^2 thấp và các chỉ số lỗi MAE, RMSE cao hơn so với các mô hình như Hồi quy Tuyến tính và ANN. Nhóm em cho rằng có một số nguyên nhân có thể dẫn đến kết quả này:

- Giá trị $K = 5$ có thể chưa phải là tối ưu. Việc lựa chọn K phù hợp là rất quan trọng đối với KNN và thường cần được tinh chỉnh thông qua kiểm định chéo (Cross-Validation).
- KNN có thể nhạy cảm với "lời nguyền chiều dữ liệu" (curse of dimensionality) khi số lượng đặc trưng lớn (đặc biệt sau khi One-Hot Encoding).
- Bản chất của dữ liệu hoặc các mối quan hệ trong đó có thể không phù hợp với cách tiếp cận dựa trên láng giềng của KNN.

Các bước cải thiện tiềm năng cho mô hình KNN mà nhóm em có thể xem xét bao gồm việc tinh chỉnh siêu tham số K một cách hệ thống, thử nghiệm các metric đo khoảng cách khác nhau, hoặc áp dụng các kỹ thuật giảm chiều dữ liệu trước khi đưa vào KNN. Tương tự như ANN, KNN cũng không cung cấp các hệ số hay độ quan trọng đặc trưng một cách trực tiếp như Hồi quy Tuyến tính hay Cây Quyết định.

7 So sánh và đánh giá hiệu suất các mô hình

Trong khuôn khổ của nghiên cứu này, bốn mô hình hồi quy đã được nhóm em xây dựng và đánh giá nhằm mục tiêu dự đoán sản lượng cây trồng (`Crop_Yield_MT_per_HA`) dựa trên các yếu tố về thời tiết, nông nghiệp, và các thông tin liên quan khác. Các mô hình bao gồm:

1. Hồi quy Tuyến tính (Linear Regression)
2. Cây Quyết định Regressor (Decision Tree Regressor) với các tham số mặc định
3. Mạng Nơ-ron Nhân tạo (Artificial Neural Network - ANN) sử dụng MLPRegressor với cấu hình ban đầu
4. K-nearest Neightbors (K-Nearest Neighbors - KNN) Regressor với K=5

Tất cả các mô hình đều sử dụng cùng một quy trình tiền xử lý dữ liệu cơ bản (xử lý giá trị thiếu, mã hóa One-Hot cho biến phân loại). Đối với mô hình ANN và KNN, các đặc trưng đầu vào đã được chuẩn hóa bằng `StandardScaler` do tính nhạy cảm của các thuật toán này với thang đo của dữ liệu.

Phần này nhóm em sẽ trình bày so sánh chi tiết hiệu suất của các mô hình dựa trên các chỉ số đánh giá hồi quy phổ biến: R-squared (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), và Root Mean Squared Error (RMSE), được tính toán trên tập dữ liệu kiểm tra.

7.1 Tổng hợp dết quả đánh giá hiệu suất

Bảng 1 tóm tắt các chỉ số hiệu suất chính của từng mô hình. Giá trị R^2 phản ánh tỷ lệ phần trăm phương sai của biến mục tiêu mà mô hình có thể giải thích được. Các chỉ số MAE, MSE, và RMSE đo lường mức độ sai số trung bình của các dự đoán; giá trị càng thấp càng tốt.

Bảng 1: *Bảng so sánh chi tiết các chỉ số đánh giá hiệu suất của các mô hình hồi quy trên tập kiểm tra. Các giá trị lỗi (MAE, RMSE) được hiểu theo đơn vị của biến mục tiêu (ví dụ: tấn/ha).*

Tên Mô hình	R-squared (R^2)	MAE	MSE	RMSE
Hồi quy Tuyến tính	0.5596	0.5466	0.4648	0.6818
Cây Quyết định Regressor (tham số gốc)	0.2679	0.6705	0.7727	0.8791
Mạng Nơ-ron Nhân tạo (ANN - MLPRegressor)*	0.5218	0.5614	0.5047	0.7104
K-nearest Neightbors (KNN, K=5)*	0.1284	0.7771	0.9200	0.9592

* Các đặc trưng đầu vào cho mô hình ANN và KNN đã được chuẩn hóa bằng StandardScaler.

7.2 Phân tích chi tiết và thảo luận Kết quả

Từ Bảng 1, chúng tôi (trong báo cáo bạn sẽ dùng "nhóm em") tiến hành phân tích sâu hơn về hiệu suất của từng mô hình:

7.2.1 Hồi quy Tuyến tính (Linear Regression)

Mô hình Hồi quy Tuyến tính đạt được giá trị $R^2 = 0.5596$, cao nhất trong số bốn mô hình được thử nghiệm. Điều này có nghĩa là khoảng 55.96% sự biến thiên trong sản lượng cây trồng có thể được giải thích bởi các đặc trưng đầu vào thông qua một mối quan hệ tuyến tính. Các chỉ số lỗi MAE (0.5466) và RMSE (0.6818) cũng là thấp nhất, cho thấy độ chính xác dự đoán tương đối tốt so với các mô hình khác trong thử nghiệm này.

- **Ưu điểm:** Đơn giản, dễ triển khai, kết quả dễ diễn giải thông qua các hệ số hồi quy (như đã trình bày ở mục 6.2.2).
- **Hạn chế:** Giả định mối quan hệ tuyến tính giữa các đặc trưng và biến mục tiêu, có thể không nắm bắt được các tương tác phức tạp hoặc các mối quan hệ phi tuyến.

- **Đồ thị trực quan (Mục 6.2.3):** Biểu đồ thực tế và dự đoán (Hình 11) cho thấy các điểm bám khá tốt quanh đường chéo, và phân phối phần dư (Hình 12) tương đối cân đối quanh 0.

7.2.2 Mạng Nơ-ron nhân tạo (ANN - MLPRegressor)

Mô hình ANN, với kiến trúc ban đầu gồm hai lớp ẩn (100 và 50 nơ-ron) và cơ chế dừng sớm (dừng sau 22 vòng lặp), đạt $R^2 = 0.5218$. Kết quả này rất đáng khích lệ, chỉ thấp hơn một chút so với Hồi quy Tuyến tính. MAE (0.5614) và RMSE (0.7104) cũng ở mức cạnh tranh.

- **Ưu điểm:** Có khả năng học các mối quan hệ phi tuyến phức tạp mà không cần định nghĩa tường minh các đặc trưng tương tác hay đa thức. Đường cong hàm mất mát (Hình 18) cho thấy mô hình đã học hiệu quả.
- **Hạn chế:** Là một mô hình "hộp đen", khó diễn giải trực tiếp các trọng số. Hiệu suất rất nhạy cảm với kiến trúc mạng và các siêu tham số (tốc độ học, số epochs, hàm kích hoạt, thuật toán tối ưu, etc.), đòi hỏi quá trình tinh chỉnh kỹ lưỡng. Yêu cầu chuẩn hóa đặc trưng đầu vào.
- **Đồ thị trực quan (Mục 6.4.2):** Các biểu đồ (Hình 16 và 17) cũng cho thấy hiệu suất tốt, tương tự Hồi quy Tuyến tính.

7.2.3 Cây quyết định Regressor (Decision Tree Regressor)

Mô hình cây Quyết định với các tham số mặc định cho kết quả $R^2 = 0.2679$, thấp hơn đáng kể so với hai mô hình trên. Các chỉ số lỗi MAE (0.6705) và RMSE (0.8791) cũng cao hơn.

- **Ưu điểm:** Dễ hiểu, dễ diễn giải thông qua cấu trúc cây (như Hình 15 đã minh họa một phần). Không yêu cầu chuẩn hóa đặc trưng. Có thể nắm bắt các mối quan hệ phi tuyến.
- **Hạn chế:** Một cây quyết định đơn lẻ (không giới hạn độ sâu hoặc không được cắt tỉa) rất dễ bị học quá khớp (overfitting) với dữ liệu huấn luyện, dẫn đến hiệu suất kém trên tập kiểm tra. Độ ổn định không cao, một thay đổi nhỏ trong dữ liệu có thể dẫn đến một cấu trúc cây rất khác.
- **Đồ thị trực quan (Mục 6.3.3):** Biểu đồ thực tế và dự đoán (Hình 13) cho thấy sự phân tán lớn hơn và các "bậc thang" trong dự đoán, đặc trưng của cây quyết định.

Mặc dù độ quan trọng của đặc trưng (Mục 6.3.2) cung cấp thông tin hữu ích, hiệu suất tổng thể cho thấy cần phải tinh chỉnh hoặc sử dụng các kỹ thuật ensemble dựa trên cây.

7.2.4 K-nearest Neighbors (KNN Regressor)

Mô hình KNN với $K = 5$ và trọng số ‘distance‘ cho kết quả thấp nhất trong các thử nghiệm, với $R^2 = 0.1284$. MAE (0.7771) và RMSE (0.9592) cũng là cao nhất.

- **Ưu điểm:** Đơn giản về mặt khái niệm, không có giả định mạnh về phân phối dữ liệu. Có thể hoạt động tốt với các mối quan hệ phi tuyến phức tạp cục bộ.
- **Hạn chế:** Hiệu suất rất phụ thuộc vào việc chọn giá trị K và metric đo khoảng cách. Nhạy cảm với "lời nguyền chiều dữ liệu" (curse of dimensionality), đặc biệt khi có nhiều đặc trưng không liên quan hoặc dư thừa. Yêu cầu chuẩn hóa đặc trưng đầu vào. Chi phí tính toán dự đoán có thể cao với tập dữ liệu lớn do phải tính toán khoảng cách đến tất cả các điểm trong tập huấn luyện.
- **Đồ thị trực quan (Mục 6.5.2):** Các biểu đồ (Hình 19 và 20) phản ánh hiệu suất chưa cao của mô hình này.

7.3 Thảo luận chung và Kết luận sơ bộ

Dựa trên các kết quả đánh giá ban đầu này, nhóm em kết luận:

1. **Hồi quy Tuyến tính** đã thiết lập một baseline mạnh mẽ. Điều này cho thấy một phần đáng kể của mối quan hệ giữa các yếu tố đầu vào và sản lượng có thể là tuyến tính hoặc gần tuyến tính.
2. **Mạng Nơ-ron Nhân tạo (ANN)** cho thấy tiềm năng lớn, đạt hiệu suất gần bằng Hồi quy Tuyến tính ngay cả với cấu hình ban đầu. Việc tinh chỉnh sâu hơn có thể giúp ANN vượt qua Hồi quy Tuyến tính, đặc biệt nếu có các mối quan hệ phi tuyến phức tạp mà mô hình tuyến tính bỏ lỡ.
3. **Cây Quyết định và KNN**, với các thiết lập mặc định hoặc ban đầu, chưa cho thấy hiệu quả cao bằng hai mô hình trên. Cả hai mô hình này đều đòi hỏi sự tinh chỉnh siêu tham số cẩn thận (ví dụ: độ sâu của cây, số lượng mẫu tối thiểu để chia nhánh cho Cây Quyết định; giá trị K cho KNN) để có thể đạt được hiệu suất tốt nhất.

Kết quả này nhấn mạnh tầm quan trọng của việc lựa chọn mô hình phù hợp với bản chất của dữ liệu cũng như sự cần thiết của việc tinh chỉnh siêu tham số. Đối với các mô hình phức tạp như ANN, Cây quyết định, và KNN, việc sử dụng các giá trị tham số mặc định thường không mang lại kết quả tối ưu.

8 Tài liệu tham khảo

Tài liệu

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. (Second Edition).

[3] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann. (Third Edition).

[4] McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media. (Second Edition).

[5] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media. (Second Edition).

[6] Scikit-learn developers. (2024). *Scikit-learn: Machine Learning in Python*. Truy cập ngày 09/05/2025, từ `https://scikit-learn.org/stable/`

[7] The Pandas Development Team. (2024). *pandas: Python Data Analysis Library*. Truy cập ngày 09/05/2025, từ `https://pandas.pydata.org/pandas-docs/stable/`

[8] Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature*, 585, 357–362. Hoặc tài liệu trực tuyến: `https://numpy.org/doc/stable/` (Truy cập ngày 03/05/2025).