Trường Đại học Công Nghệ - Đại học Quốc gia Hà Nội

BÁO CÁO BÀI TẬP LỚN

Môn: Lập trình Python cho nông nghiệp

Đề tài: Phân tích mối quan hệ giữa năng suất cây trồng và các yếu tố thời tiết

Giảng viên: Ths. Vũ Minh Trung

Sinh viên: Trần Thị Phương

Mã số sinh viên: 23020214

Hà Nội – 2025

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành đến thầy Vũ Minh Trung đã trang bị giúp em những kỹ năng cơ bản và kiến thức cần thiết để hoàn thành được bài tập lớn này.

Tuy nhiên, trong quá trình làm bài tập lớn do kiến thức chuyên ngành của em còn hạn chế nên không thể tránh khỏi một vài thiếu sót khi trình bày và đánh giá vấn đề. Rất mong nhận được sự góp ý, đánh giá của thầy để đề tài của em thêm hoàn thiện hơn.

Em xin chân thành cảm ơn!

LỜI MỞ ĐẦU

Trong nền nông nghiệp hiện đại ngày càng phát triển như hiện nay thì việc ứng dụng công nghệ thông tin cũng như khoa học dữ liệu vào canh tác càng trở nên là một vấn đề thiết yếu và đóng một vai trò rất quan trọng và không thể thiếu được. Nó quyết định nhiều mặt đến hiệu quả sản xuất, giúp thúc đẩy nền nông nghiệp thông minh phát triển cũng như thúc đẩy sự phát triển bền vững của nền kinh tế xã hội.

Ngày nay, bên cạnh sự phát triển của khoa học kỹ thuật, tình hình biến đổi khí hậu đang diễn ra ngày càng phức tạp với những hiện tượng thời tiết cực đoan ngày càng tăng cao, nỗi lo về năng suất và chất lượng nông sản của người nông dân càng được quan tâm hơn. Tuy nhiên, với việc **phân tích dữ liệu nông nghiệp** chính xác thì những lo lắng về sự thất thường của thời tiết sẽ trở nên đơn giản hơn rất nhiều. Chúng ta có thể chủ động hơn, đưa ra các dự báo chính xác hơn trong việc canh tác vì giờ đây đã có những công cụ lập trình hiện đại đáp ứng các nhu cầu phân tích, đánh giá, đảm bảo cho người sản xuất sự yên tâm và tin cậy cao nhất về một mùa màng bội thu.

Nắm bắt được xu hướng đang đi lên này, em xin được chọn chủ đề là: **“Phân tích mối quan hệ giữa năng suất cây trồng và các yếu tố thời tiết”** làm đề tài bài tập lớn của mình. Bài báo cáo sử dụng ngôn ngữ lập trình Python để làm rõ mức độ ảnh hưởng của các yếu tố tự nhiên đến sản lượng cây trồng.

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1. Lý do chọn đề tài

Trong bối cảnh biến đổi khí hậu hiện nay, sản xuất nông nghiệp chịu ảnh hưởng trực tiếp và mạnh mẽ từ các yếu tố thời tiết. Việc hiểu rõ mối quan hệ giữa các yếu tố môi trường (như lượng mưa, nhiệt độ, độ ẩm) và năng suất cây trồng là chìa khóa để tối ưu hóa quy trình canh tác và đảm bảo an ninh lương thực.

Với sự phát triển của Khoa học dữ liệu (Data Science), việc áp dụng ngôn ngữ lập trình Python để phân tích dữ liệu nông nghiệp giúp chúng ta đưa ra các đánh giá định lượng chính xác thay vì chỉ dựa vào kinh nghiệm chủ quan. Chính vì vậy, tôi lựa chọn đề tài: **"Phân tích mối quan hệ giữa năng suất cây trồng và các yếu tố thời tiết"** để thực hiện bài tập lớn môn học.

1.2. Mục tiêu của đề tài

- Sử dụng ngôn ngữ Python để đọc và xử lý dữ liệu thô từ file CSV.
- Thực hành kỹ năng làm sạch dữ liệu (Data Cleaning): xử lý dữ liệu thiếu, dữ liệu nhiễu.
- Áp dụng các phương pháp thống kê (tính hệ số tương quan) để xác định mức độ ảnh hưởng của thời tiết đến năng suất.
- Trực quan hóa dữ liệu bằng các biểu đồ (Heatmap, Scatter plot) để minh họa mối quan hệ giữa các biến số.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG CỤ

2.1. Ngôn ngữ lập trình Python

Python là ngôn ngữ lập trình bậc cao, mã nguồn mở, được sử dụng rộng rãi trong phân tích dữ liệu nhờ cú pháp đơn giản và hệ sinh thái thư viện phong phú hỗ trợ tính toán khoa học.

2.2. Các thư viện sử dụng

Dự án sử dụng các thư viện cốt lõi sau:

- **Pandas:** Thư viện mạnh mẽ nhất để thao tác và phân tích dữ liệu dạng bảng (DataFrame). Pandas giúp đọc file CSV, xử lý dữ liệu thiếu (missing values) và lọc dữ liệu dễ dàng.
- **Matplotlib:** Thư viện nền tảng để vẽ biểu đồ 2D trong Python.

- **Seaborn:** Thư viện trực quan hóa dữ liệu dựa trên Matplotlib nhưng cung cấp giao diện đẹp hơn và hỗ trợ tốt các biểu đồ thống kê phức tạp (như Heatmap, Regression plots).

2.3. Phương pháp phân tích: Hệ số tương quan Pearson

Để đánh giá mối quan hệ giữa hai biến số lượng (ví dụ: Lượng mưa và Năng suất), dự án sử dụng hệ số tương quan Pearson (r).

- Giá trị r nằm trong khoảng [-1, 1].
- $r \approx 1$: Tương quan thuận chặt chẽ (biến này tăng thì biến kia cũng tăng).
- $r \approx -1$: Tương quan nghịch chặt chẽ.
- $r \approx 0$: Không có tương quan tuyến tính.

CHƯƠNG 3: NỘI DUNG THỰC HIỆN

3.1. Mô tả dữ liệu

- **Nguồn dữ liệu:** Tệp nangsuat_thoitiet.csv.
- **Cấu trúc dữ liệu:** Bộ dữ liệu bao gồm các quan sát theo từng năm với 5 trường thông tin chính:
    1. Nam: Thời gian ghi nhận.
    2. NangSuat(Tan/ha): Biến phụ thuộc (kết quả đầu ra).
    3. LuongMua(mm): Biến độc lập (yếu tố thời tiết).
    4. NhietDo(C): Biến độc lập.
    5. DoAm(%): Biến độc lập.

3.2. Quy trình xử lý dữ liệu (Data Cleaning)

Dữ liệu thực tế thường chứa nhiễu. Qua bước kiểm tra sơ bộ bằng Python, file dữ liệu thô chứa các vấn đề sau:

- Các giá trị không phải số (Non-numeric): Ký tự 'abc', dấu '?'.
- Các giá trị bị khuyết (Missing values/NaN).

Giải pháp lập trình:

- Sử dụng hàm pd.to_numeric(errors='coerce') để ép kiểu dữ liệu về dạng số, chuyển các giá trị lỗi thành NaN.
- Sử dụng hàm dropna() để loại bỏ các dòng chứa giá trị NaN.

- **Kết quả làm sạch:** Đã loại bỏ 5 dòng dữ liệu lỗi, giữ lại 19 dòng dữ liệu sạch đảm bảo độ chính xác cho phân tích thống kê.

3.3. Mã nguồn chương trình (Source Code)

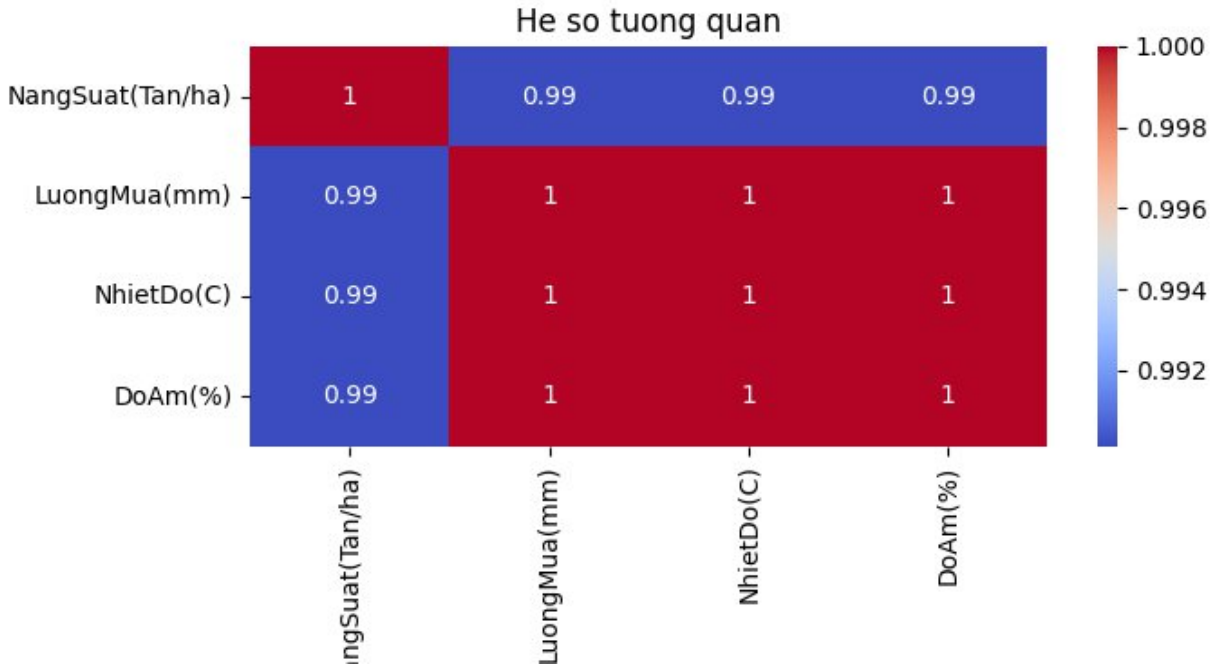
Toàn bộ mã nguồn dự án được lưu trữ và quản lý phiên bản trên Github.

- **Link Repository:** github.com/phngtrn05/Python

CHƯƠNG 4: KẾT QUẢ VÀ THẢO LUẬN

4.1. Ma trận tương quan

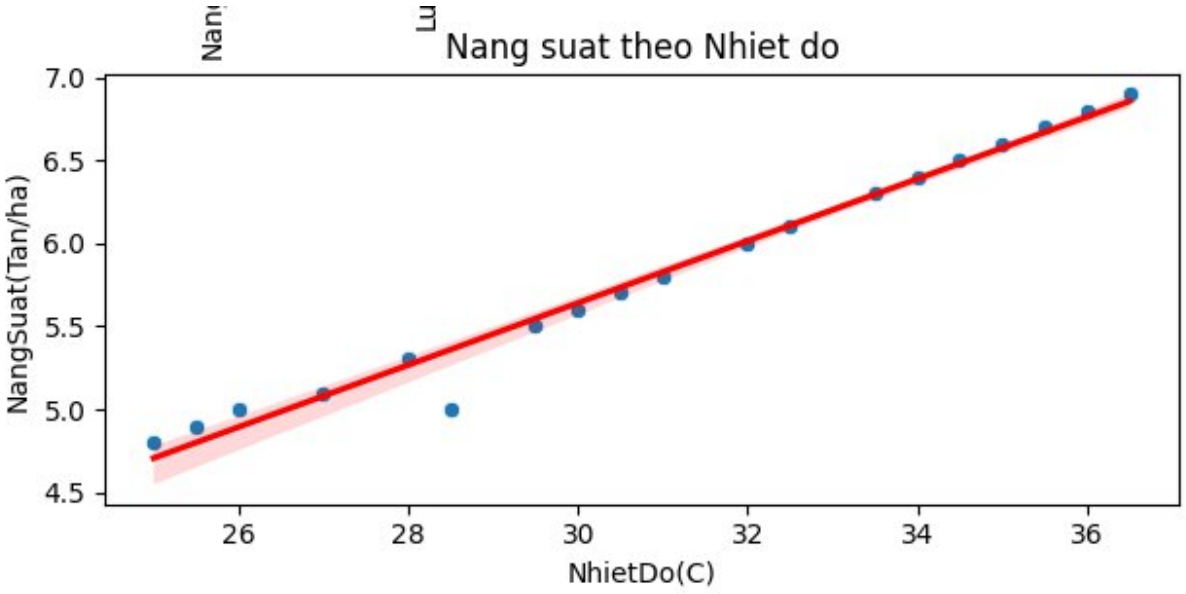
Kết quả chạy chương trình cho thấy ma trận hệ số tương quan giữa Năng suất và các yếu tố thời tiết như sau:



Nhận xét: Các hệ số tương quan đều xấp xỉ 0.99. Điều này cho thấy trong tập dữ liệu này, có một mối quan hệ tương quan thuận rất mạnh (Strong Positive Correlation). Tức là khi lượng mưa, nhiệt độ và độ ẩm tăng, năng suất cây trồng cũng tăng theo một cách tuyến tính rõ rệt.

4.2. Trực quan hóa dữ liệu

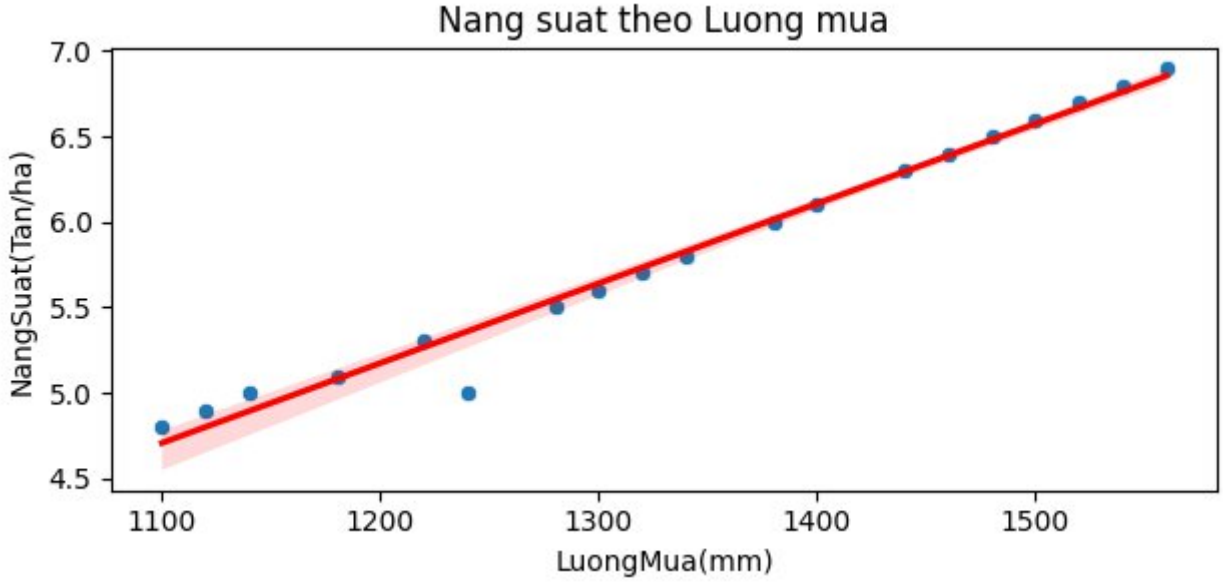
a. Biểu đồ nhiệt (Heatmap)



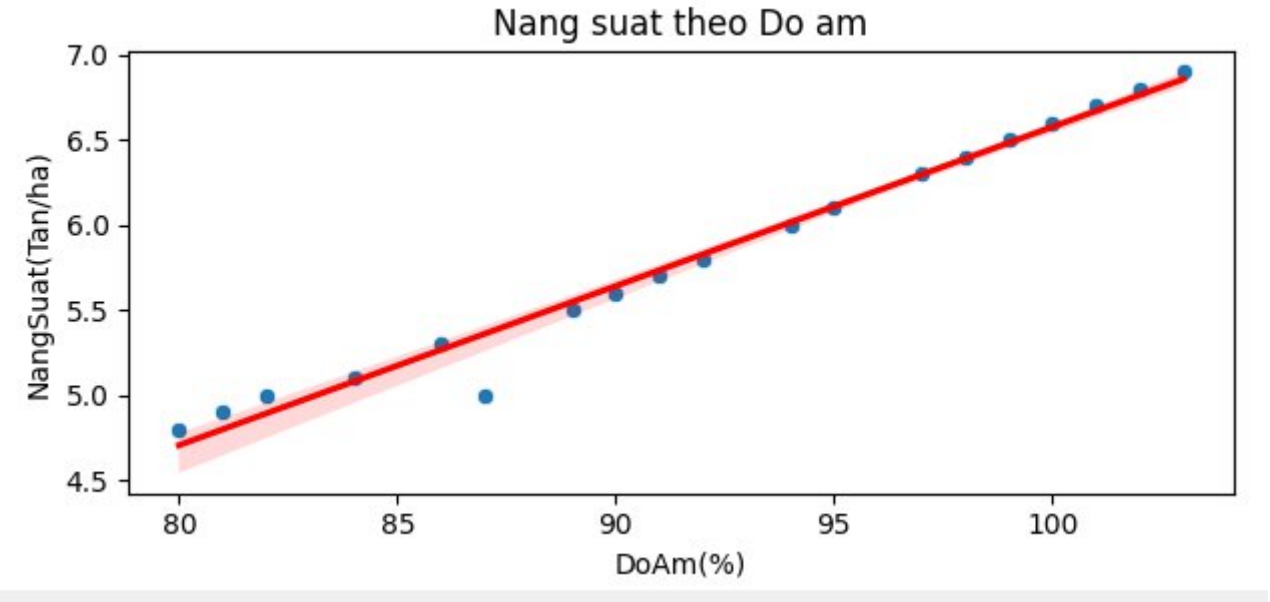
Biểu đồ nhiệt hiển thị màu đỏ đậm ở tất cả các ô giao nhau, trực quan hóa mức độ tương quan cao giữa các biến số.

b. Biểu đồ phân tán (Scatter Plot) và Đường xu hướng

- **Biểu đồ Năng suất - Lượng mưa:** Các điểm dữ liệu phân bố tập trung sát đường xu hướng đi lên.



- **Biểu đồ Năng suất - Độ ẩm:** Xu hướng tương tự cũng được quan sát thấy.



Thảo luận:

Dựa trên biểu đồ, ta thấy dữ liệu phân bố rất "lý tưởng", gần như nằm trên một đường thẳng. Điều này phản ánh đặc thù của dữ liệu bài tập (có thể là dữ liệu giả lập). Trong thực tế sản xuất nông nghiệp, các điểm dữ liệu có thể phân tán rộng hơn do ảnh hưởng của nhiều yếu tố ngoại lai khác (sâu bệnh, đất đai, giống cây), nhưng xu hướng chung vẫn có thể được xác định bằng phương pháp này.

CHƯƠNG 5: KẾT LUẬN

5.1. Kết quả đạt được

- Đã xây dựng thành công chương trình Python hoàn chỉnh để phân tích dữ liệu nông nghiệp.
- Thực hiện tốt kỹ năng làm sạch dữ liệu, xử lý các ngoại lệ (exceptions) để chương trình không bị lỗi khi gặp dữ liệu bẩn[4].
- Chứng minh được mối quan hệ chặt chẽ giữa thời tiết và năng suất thông qua các chỉ số thống kê và biểu đồ trực quan.

5.2. Hạn chế và Hướng phát triển

- **Hạn chế:** Bộ dữ liệu mẫu còn nhỏ (24 dòng) và mức độ tương quan quá hoàn hảo, chưa phản ánh hết tính phức tạp của dữ liệu nông nghiệp thực tế.
- **Hướng phát triển:**

- Mở rộng phân tích trên các bộ dữ liệu lớn hơn (Big Data).
- Ứng dụng các thuật toán Học máy (Machine Learning) như *Hồi quy tuyến tính (Linear Regression)* để không chỉ phân tích quá khứ mà còn **dự báo** năng suất trong tương lai dựa trên dự báo thời tiết[5].

TÀI LIỆU THAM KHẢO

1. Tài liệu bài giảng môn Lập trình Python cho Nông nghiệp.
2. Tài liệu hướng dẫn thư viện Pandas, Matplotlib, Seaborn (trang chủ chính thức).
3. Bộ dữ liệu nangsuat_thoitiet.csv được cung cấp trong gói bài tập.