**TRƯỜNG ĐẠI HỌC THUỶ LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated

**BÀI TẬP LỚN**

**MÔN HỌC: HỌC MÁY**

**Tên đề tài:**

**DỰ ĐOÁN NĂNG SUẤT CÂY TRỒNG TRONG NÔNG NGHIỆP**

Giáo viên hướng dẫn: TẠ QUANG CHIỂU

Các thành viên trong Nhóm 12\_64TTNT1:

|  |  |
| --- | --- |
| Họ tên sinh viên | Mã sinh viên |
| Nguyễn Thị Quỳnh Anh | 2251262573 |
| Đồng Anh Quân | 2251262626 |

Hà Nội - 2025

**MỤC LỤC**

**[LỜI CẢM ƠN](#_Toc22007)** [1](#_Toc22007)

**[CHƯƠNG I: MÔ TẢ BÀI TOÁN](#_Toc28594)** [2](#_Toc28594)

[1. Lý do chọn đề tài 2](#_Toc8510)

[2. Tổng quan bài toán 2](#_Toc8791)

[3. Quy trình thực hiện 2](#_Toc26628)

[4. Phân tích dữ liệu thô 3](#_Toc23935)

[Bảng các thuộc tính 3](#_Toc1334)

**CHƯƠNG II: QUY TRÌNH XÂY DỰNG MÔ HÌNH HỌC MÁY**..................4

[1. Tiền xử lý và trực quan hóa dữ liệu 4](#_Toc8845)

[1.1 Khám phá dữ liệu ban đầu 4](#_Toc17985)

[1.2 Làm sạch dữ liệu 9](#_Toc31847)

[1.2.1 Xử lý dữ liệu bị thiếu 9](#_Toc30946)

[1.2.2 Xử lý dữ liệu trùng lặp 10](#_Toc24217)

[1.3 Biến đổi dữ liệu 10](#_Toc13356)

[1.3.1 Chuyển đổi dữ liệu 10](#_Toc5821)

[1.3.2 Mã hóa dữ liệu 11](#_Toc7897)

[1.3.3 Chuẩn hóa dữ liệu 12](#_Toc10758)

[2. Xây dựng và huấn luyện mô hình 13](#_Toc17713)

2.1 Lựa chọn mô hình..............................................................................13

2.2 Phân tách dữ liệu................................................................................14

2.3 Xây dựng và huấn luyện với mô hình Linear Regression...................15

[2.3.1 Với thư viện 15](#_Toc2064)

[2.3.2 Tự code 15](#_Toc25344)

2.4 Xây dựng và huấn luyện với mô hình Decision Tree Regressor........16

[2.4.1 Với thư viện 16](#_Toc1397)

[2.4.2 Tự Code 17](#_Toc27258)

[3. Đánh giá mô hình 18](#_Toc25517)

[3.1 Đánh giá mô hình Linear Regression 19](#_Toc5885)

[3.2 Đánh giá mô hình Decision Tree for Regressor 21](#_Toc6466)

**[KẾT LUẬN](#_Toc4872)** [24](#_Toc4872)

**[TÀI LIỆU THAM KHẢO](#_Toc25701)** [24](#_Toc25701)

**LỜI CẢM ƠN**

Ngày nay, việc ứng dụng công nghệ AI đã trở nên phổ biến trong hầu hết mọi mặt của cuộc sống đặc biệt là việc áp dụng các phương pháp học máy để đưa ra dự báo. Trong ít năm trở lại đây, với tốc độ phát triển như vũ bão, AI đang dần làm cho cuộc sống của con người trở nên thú vị và đơn giản hơn. Để bắt kịp với nhịp độ phát triển của xã hội, những kiến thức học được trên giảng đường là vô cùng quan trọng đối với mỗi sinh viên chúng em.

Vì vậy nhóm em chọn đề tài **“Dự đoán năng suất cây trồng trong nông nghiệp”** để làm báo cáo kết thúc môn Học Máy của mình. Nhóm thực hiện chúng em xin gửi lời cảm ơn đặc biệt đến thầy giáo Tạ Quang Chiểu đã tận tình hướng dẫn chúng em trong bộ môn Học Máy và giúp đỡ chúng em hoàn thành bài tập lớn cuối kỳ này.

Bên cạnh những kết quả mà chúng em đạt được thì sẽ không khó tránh khỏi những thiếu sót trong quá trình làm đề tài vì thời gian không cho phép và chưa có kinh nghiệm thực tế. Chính vì vậy chúng em rất mong được sự cảm thông, chỉ bảo góp ý của thầy cô. Những lời nhận xét, góp ý của thầy chính là một bài học, kiến thức cho chúng em trên con đường sau này.

Chúng em xin chân thành cảm ơn!

**CHƯƠNG I: MÔ TẢ BÀI TOÁN**

1. ***Lý do chọn đề tài***

Theo truyền thống, nông dân dựa vào lịch sử, thời tiết và kinh nghiệm của chính mình để ước tính sản lượng. Tuy nhiên, những phương pháp này thường không chính xác và dễ xảy ra lỗi chủ quan do con người. Với AI, nông dân giờ đây có thể tận dụng các thuật toán tiên tiến và kỹ thuật học máy để phân tích lượng dữ liệu khổng lồ và đưa ra dự đoán năng suất chính xác.

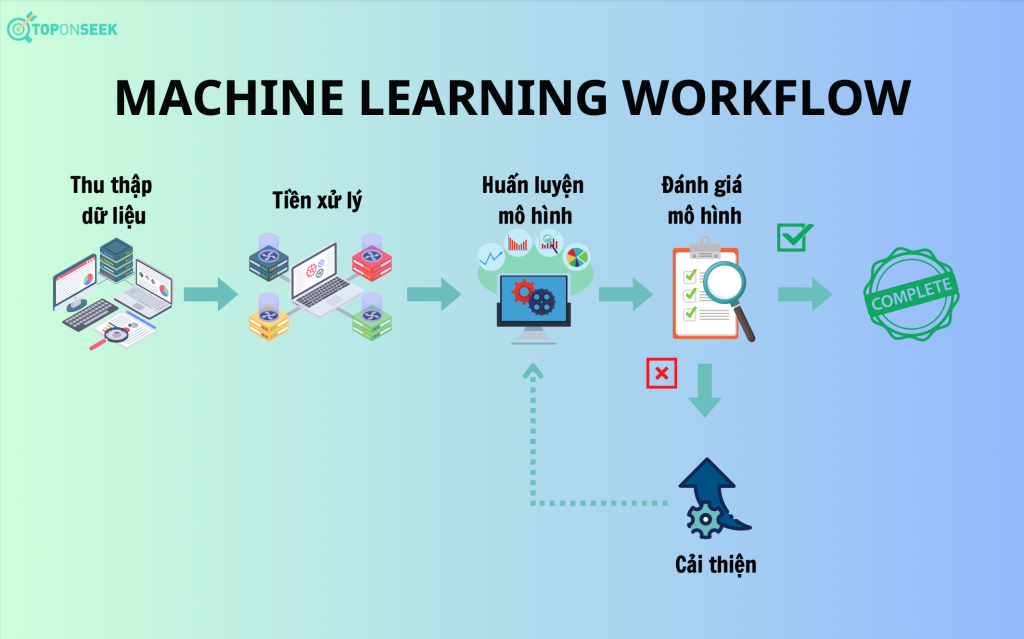
Một trong những lợi thế chính của AI trong dự đoán lợi nhuận là khả năng xử lý các tập dữ liệu lớn. Bằng cách phân tích dữ liệu năng suất lịch sử, kiểu thời tiết, điều kiện đất đai và các yếu tố liên quan khác, thuật toán AI có thể xác định các mô hình và các yếu tố mà con người có thể vô tình bỏ qua.

Điều này cho phép nông dân có được những thông tin sâu về các yếu tố ảnh hưởng đến năng suất cây trồng và đưa ra quyết định dựa trên dữ liệu để tối ưu hóa các biện pháp canh tác của mình.

1. ***Tổng quan bài toán***

Từ Dataset bao gồm các dữ liệu về các thuộc tính tương ứng, ta áp dụng các phương pháp học máy phù hợp để xác định xem với các thuộc tính đã có cho ra năng suất cây trồng đạt sản lượng bao nhiêu tấn trên 1 hecta

***3. Quy trình thực hiện***



Quy trình xây dựng mô hình học máy gồm 6 bước:

Bước 1: Thu thập các dữ liệu để tạo ra 1 bộ dữ liệu

Bước 2: Tiền xử lý, làm sạch tập dữ liệu

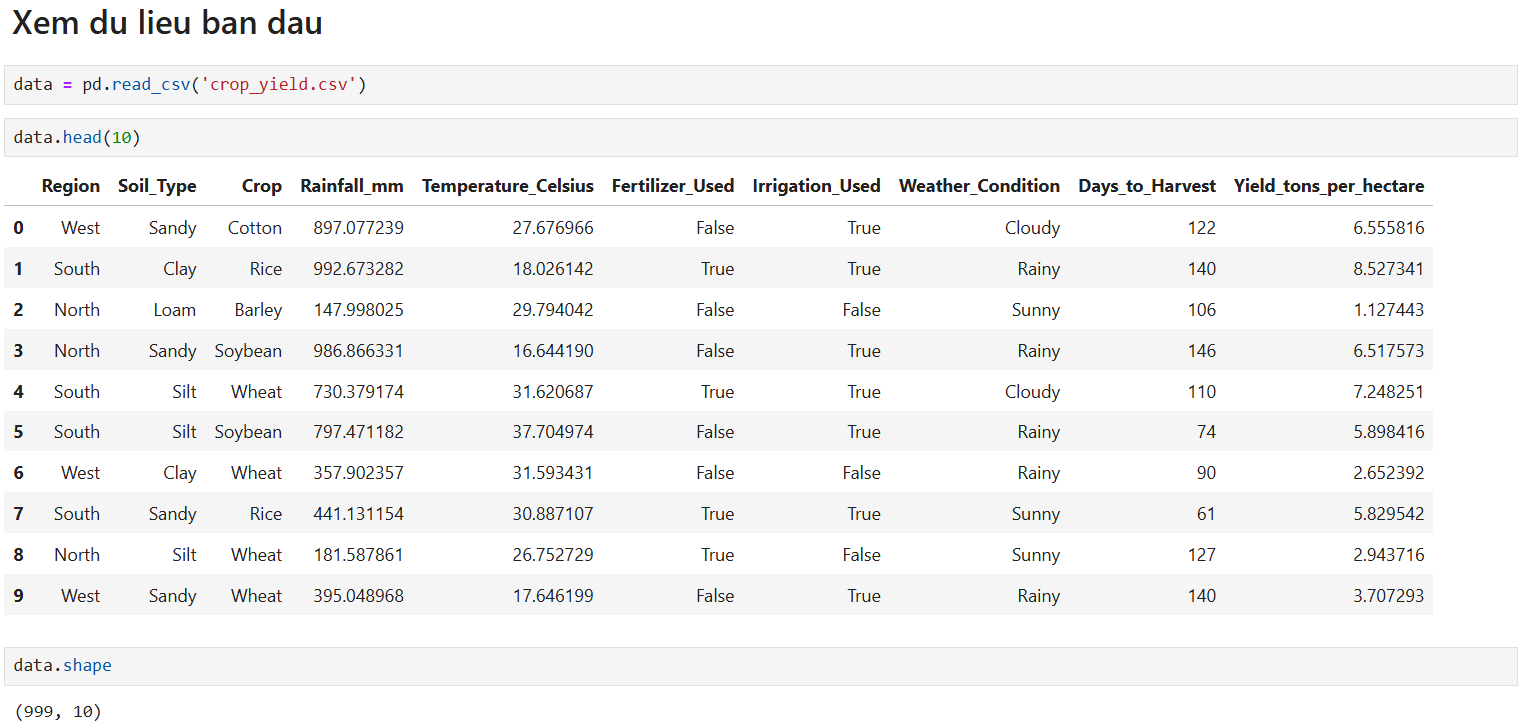
Bước 3: Xây dựng và huấn luyện mô hình

Bước 4: Đánh giá mô hình

Bước 5: Hiệu chỉnh các tham số của mô hình

Bước 6: Triển khai mô hình

***4. Phân tích dữ liệu thô***



*Hình 1.1: Sau khi đọc dữ liệu thô*

- Sau khi hiển thị dữ liệu thô thì ta thấy: bộ dữ liệu gồm có 999 dòng và 10 cột

**Bảng các thuộc tính**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa thuộc tính** |
| 1 | Region | - Vùng trồng cây: ***North (phía bắc), South (phía nam), East (phía đông), West (phía tây)*** |
| 2 | Soil\_Type | - Loại đất trồng: ***Sandy (Cát), Clay (Đất sét), Loam (Đất thịt), Silt (Bùn), Peaty (Than bùn),***  ***Chalky (Phấn)*** |
| 3 | Crop | - Loại cây trồng: ***Cotton (Bông), Rice (Gạo), Barley (Lúa mạch), Soybean (Đậu nành), Wheat (Lúa mì), Maize (Ngô)*** |
| 4 | Rainfall\_mm | Lượng mưa (mm) |
| 5 | Temperature\_Celsius | Nhiệt độ (độ C) |
| 6 | Fertilizer\_Used | - Có sử dụng phân bón cho cây: ***True (Có), False (Không)*** |
| 7 | Irrigation Used | - Có hay tưới tiêu cho cây: ***True (Có), False (Không)*** |
| 8 | Weather\_Condition | - Điều kiện thời tiết: ***Cloudy(Có mây), Rainy (Mưa),***  ***Sunny (Nắng)*** |
| 9 | Days\_to\_Harvest | Ngày được thu hoạch |
| 10 | Yield\_tons\_per\_hectare | Sản lượng lượng thu hoạch (tấn/hecta) |

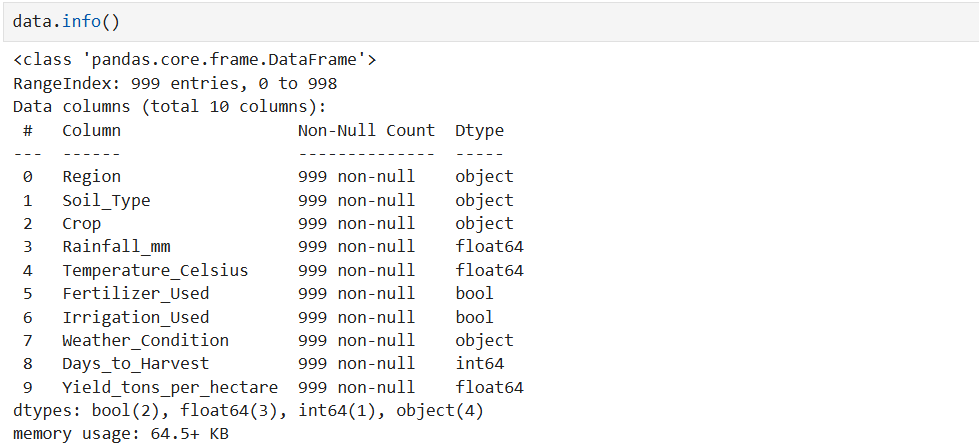
**CHƯƠNG II: QUY TRÌNH XÂY DỰNG MÔ HÌNH HỌC MÁY**

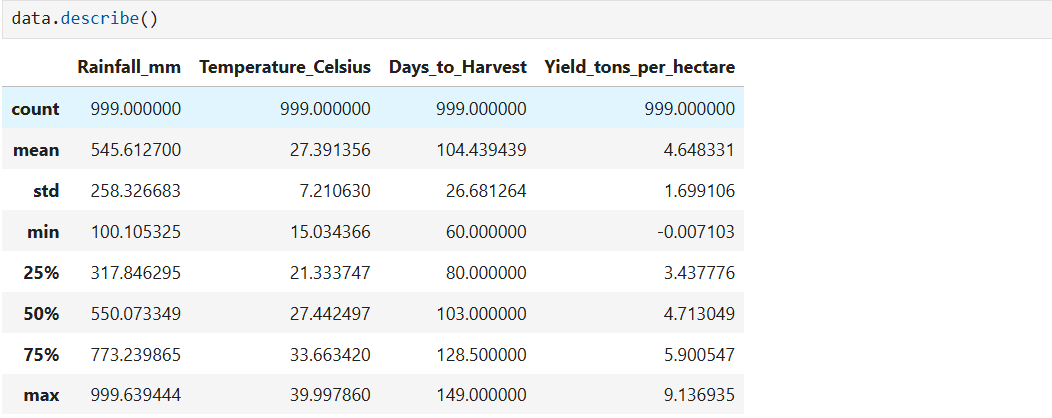
1. ***Tiền xử lý và trực quan hóa dữ liệu***

*- Tiền xử lý:* là quá trình xử lý dữ liệu thô/gốc nhằm cải thiện chất lượng dữ liệu và chất lượng kết quả dự đoán của mô hình

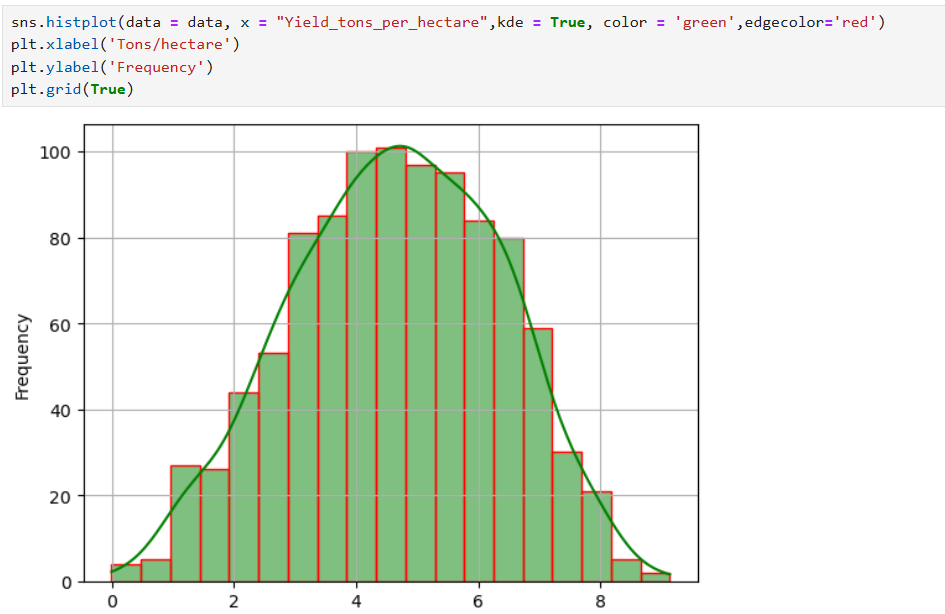
*- Trực quan hóa dữ liệu:* là quá trình biểu diễn dữ liệu dưới dạng các đồ thị, biểu đồ hoặc hình ảnh trực quan. Mục đích chính là làm cho dữ liệu trở nên dễ hiểu hơn, từ đó hỗ trợ việc ra quyết định, phát hiện xu hướng và mẫu mà có thể không dễ dàng nhận ra khi chỉ nhìn vào dữ liệu thô.

* 1. ***Khám phá dữ liệu ban đầu***





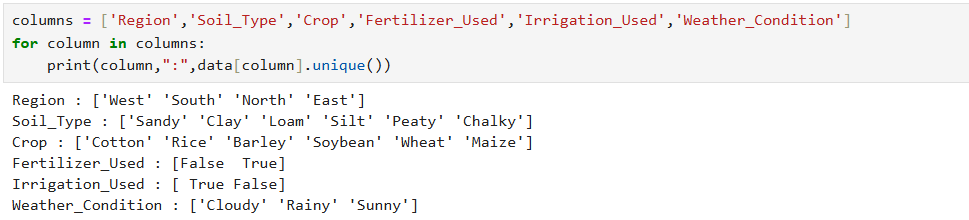
*Hình 2.1: Mô tả chung của dữ liệu ban đầu*



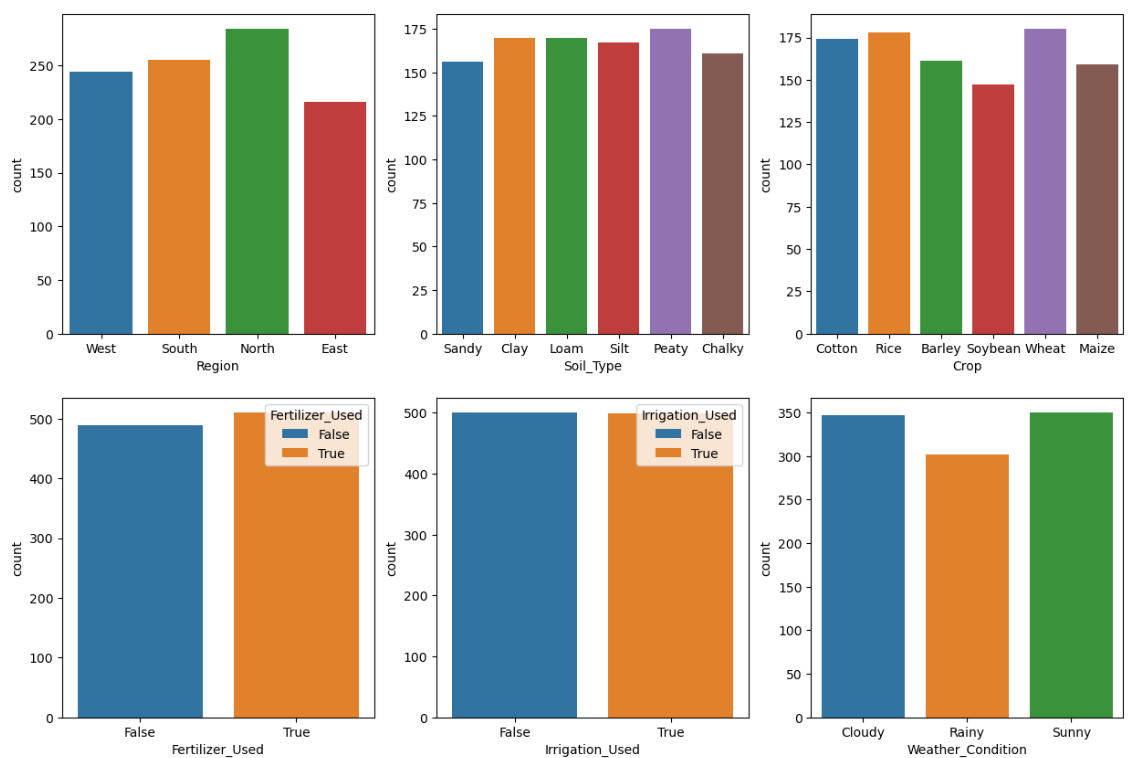
*Hình 2.2: Sau khi trực quan hóa biến mục tiêu*

*- Nhận xét:* Từ biểu đồ ta thấy sản lượng thu hoạch phân bố theo dạng chuẩn Gauss và phân bố tập trung nhiều vào khoảng từ 3 đến 6 tấn/hecta

- Từ các thuộc tính ở trong bảng thuộc tính ta xem các thuộc tính có chứa dữ liệu rời rạc và trực quan hóa chúng:



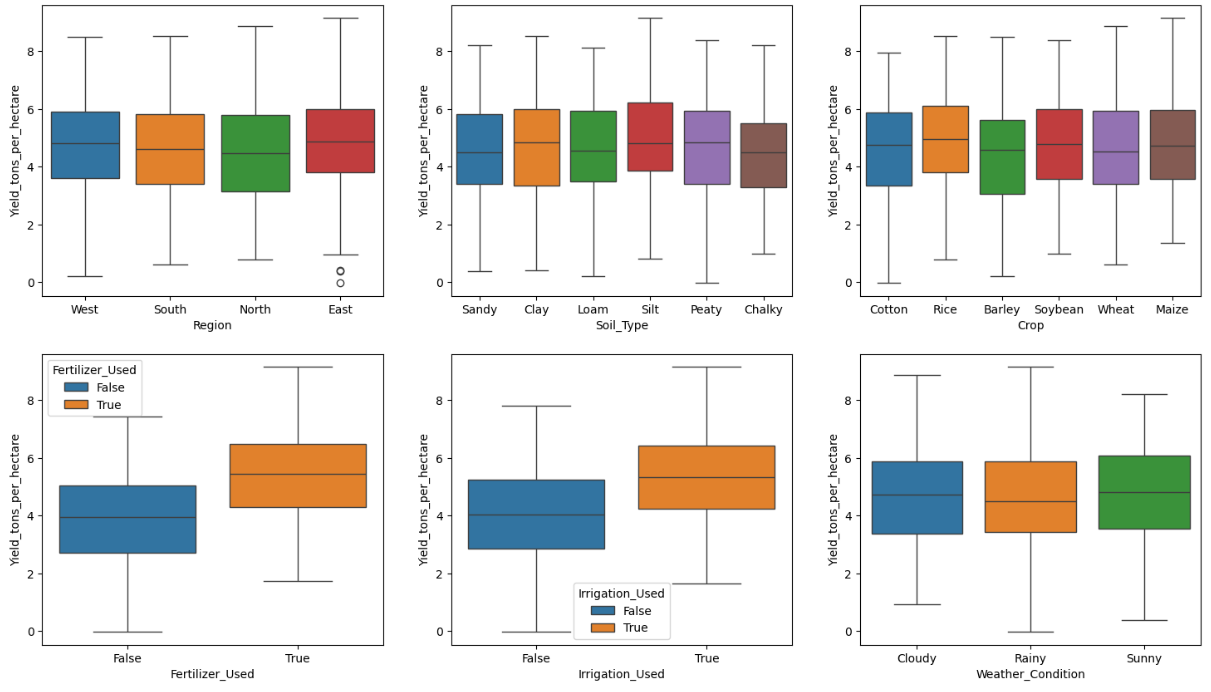
*Hình 2.3: Ta xem các thuộc tính có chứa dữ liệu rời rạc*



*Hình 2.4: Sau khi trực quan các thuộc tính chứa dữ liệu rời rạc*

*- Nhận xét:* Sau khi trực quan hóa ta thấy dữ liệu rời rạc chứa trong các thuốc tính phân bố khá cân bằng với nhau

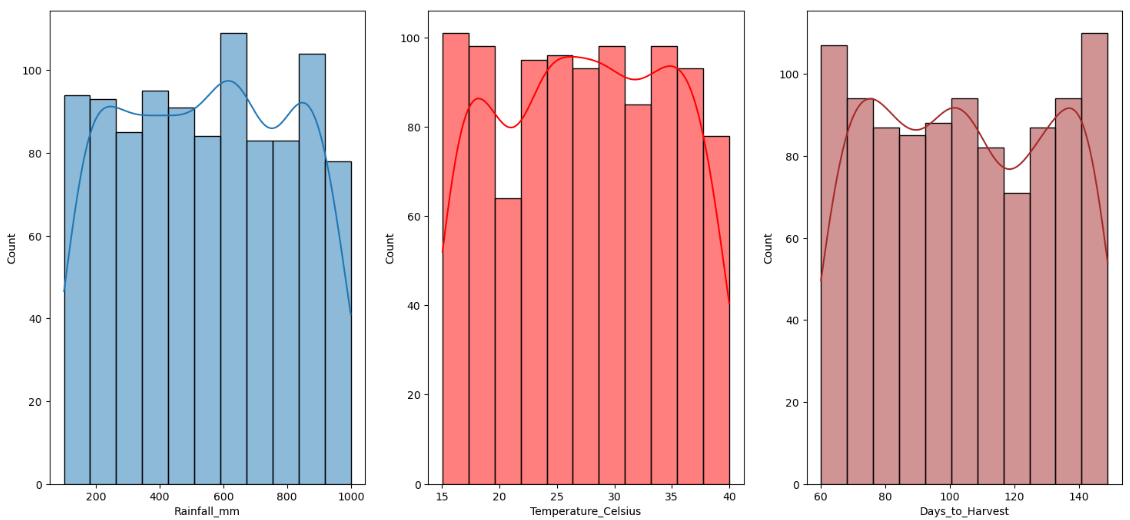
- Tương tự ta trực quan hóa các thuộc tính chứa dữ liệu rời rạc với biến mục tiêu để xem sự tương quan giữa chúng:

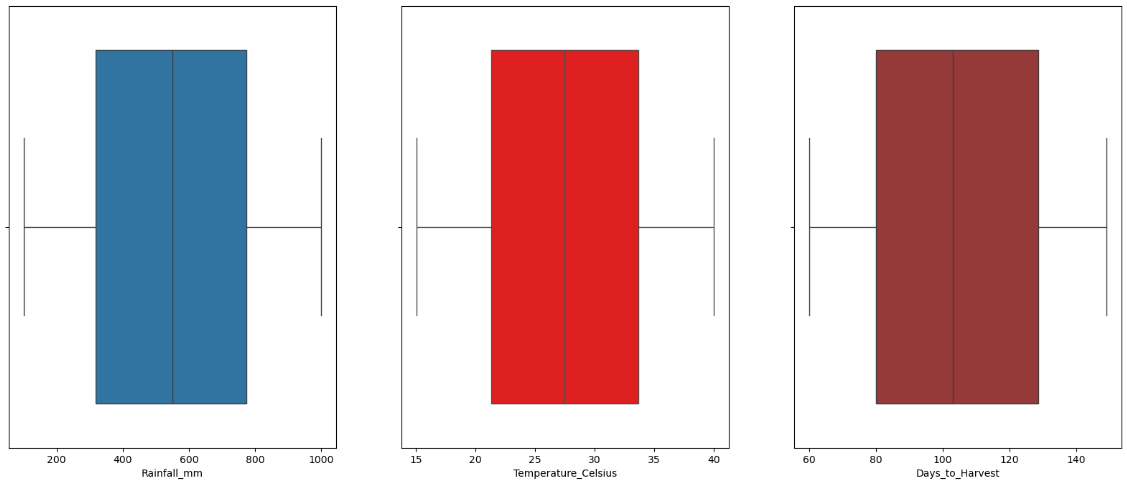


*Hình 2.5: Sau khi trực quan hóa thuộc tính chứa dữ liệu rời rạc với biến mục tiêu*

*- Nhận xét:* Các yếu tố như tưới tiêu, phân bón có ảnh hưởng tích cực rõ rệt đến năng suất. Loại đất, loại cây trồng và khu vực cũng có vai trò đáng kể, nhưng có sự ảnh hướng lớn đến năng suất.

- Ta tiếp tục trực quan hóa các thuộc tính chứa dữ liệu liên tục:





*Hình 2.6: Sau khi trực quan hóa thuộc tính chứa dữ liệu liên tục*

*- Nhận xét:*

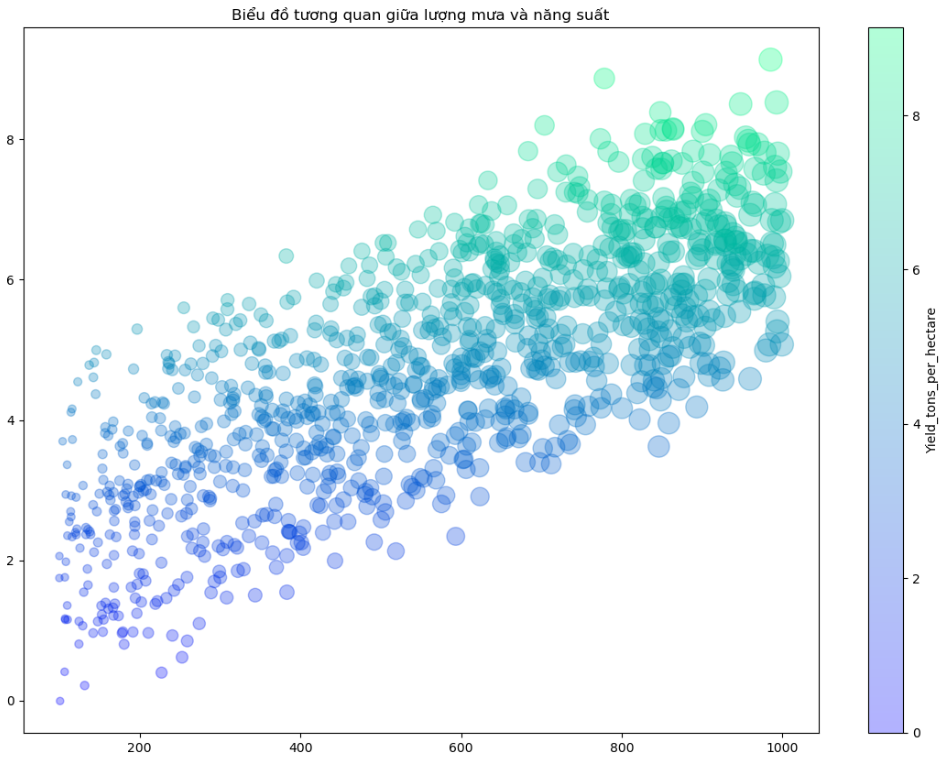
+> Với thuộc tính Lượng mưa (Rainfall\_mm): Phân bố lượng mưa tương đối cân đối, khoảng phân bố tập trung từ khoảng 300 mm đến 800 mm.

+> Với thuộc tính Temperature\_Celsius (Nhiệt độ): Phân bố nhiệt độ tương đối cân đối, với khoảng phân bố tâp trung từ 22°C đến 35°C.

+> Với thuộc tính Days\_to\_Harvest (Sô ngày thu hoạch): Phân bố nhiệt độ tương đối cân đối, với khoảng phân bố tâp trung từ 80 ngày đến 130 ngày.

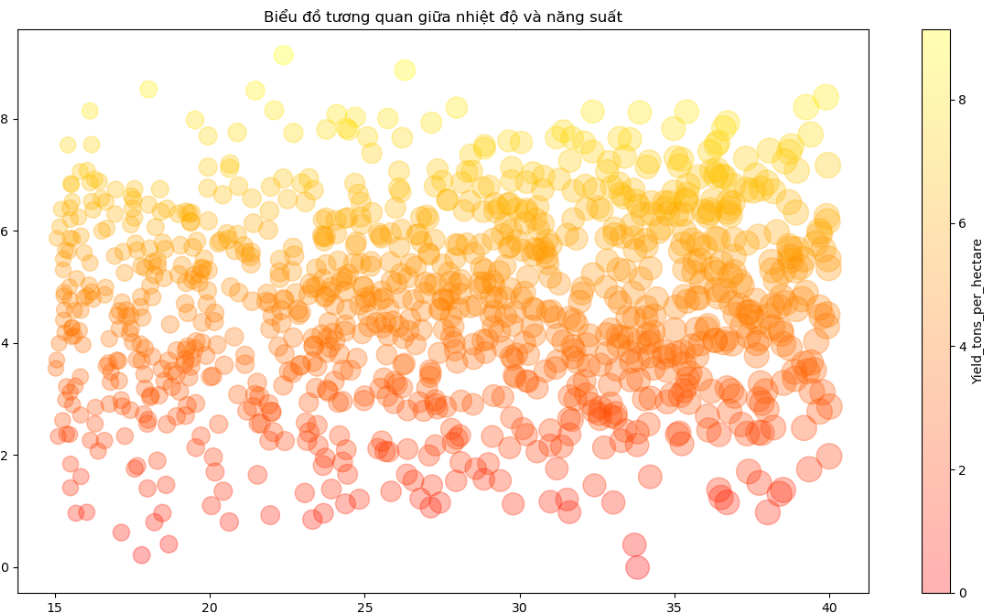
- Các yếu tố lượng mưa, nhiệt độ và số ngày thu hoạch đều không có điểm ngoại lệ, cho thấy tính đồng nhất và ổn định trong dữ liệu.

- Tương tự ta trực quan hóa các thuộc tính chứa dữ liệu liên tục với biến mục tiêu để xem sự tương quan giữa chúng:

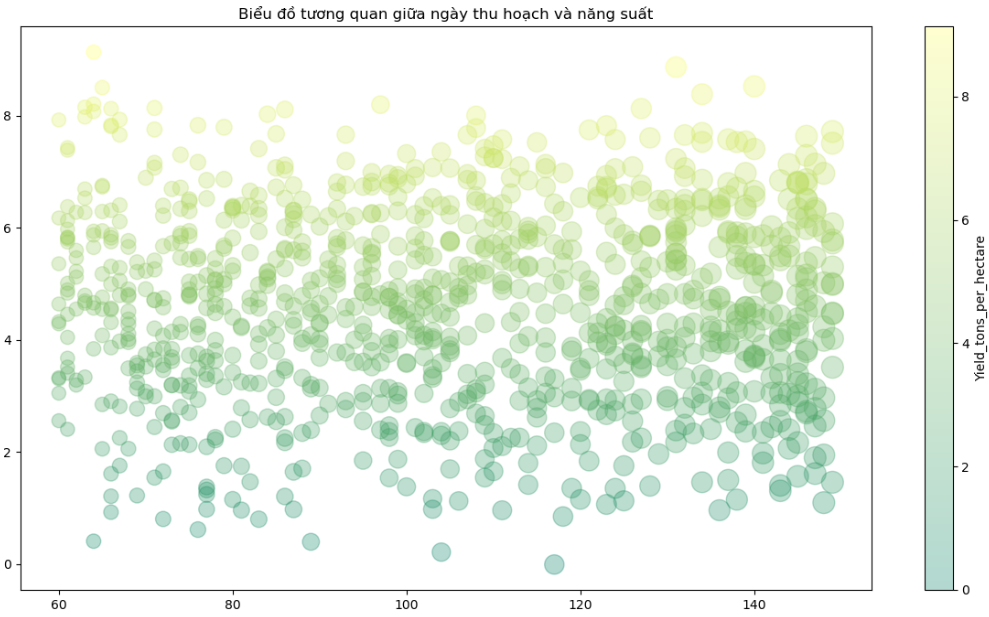


*Hình 2.7: Tương quan giữa thuộc tính lượng mưa với biến mục tiêu*

*- Nhận xét:* Ta thấy lượng mưa ảnh hưởng tuyến tính đến sản lượng (Lượng mưa nhiều sản lượng lớn)



*Hình 2.8: Tương quan giữa thuộc tính nhiệt độ với biến mục tiêu*



*Hình 2.9: Tương quan giữa thuộc tính ngày thu hoạch với biến mục tiêu*

*- Nhận xét:*

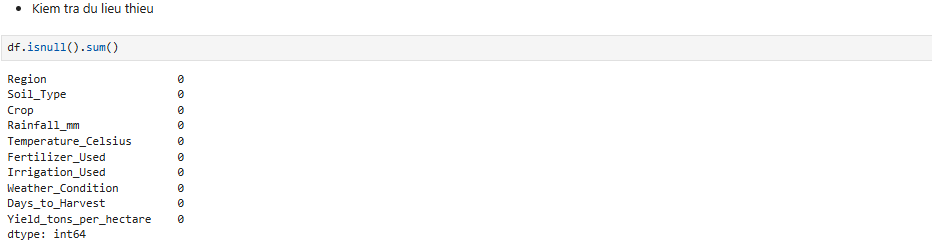
*+> Hình 2.8:* Ở mỗi mức nhiệt độ, năng suất không cố định mà có sự phân tán lớn. Điều này chỉ ra rằng ngoài nhiệt độ, còn nhiều yếu tố khác như loại đất, lượng mưa, hoặc cây trồng ảnh hưởng đến năng suất.

+> *Tương tự* *Hình 2.9:* Ở số ngày thu hoạch, năng suất không cố định mà có sự phân tán lớn.

* 1. ***Làm sạch dữ liệu***

*- Làm sạch dữ liệu:* là quá trình phát hiện và sửa chữa hoặc loại bỏ các lỗi, giá trị không hợp lệ, thiếu dữ liệu, hoặc các điểm bất thường trong bộ dữ liệu nhằm cải thiện chất lượng và độ tin cậy của dữ liệu

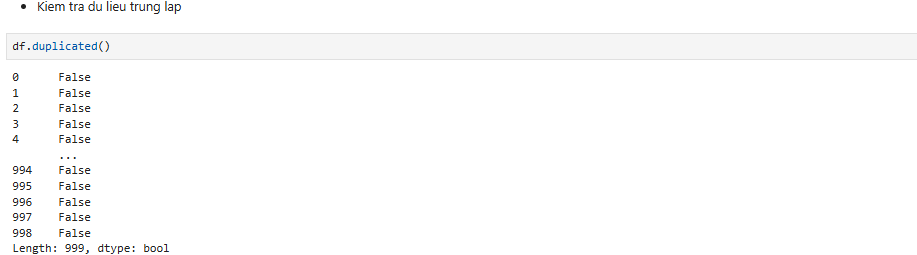
***1.2.1 Xử lý dữ liệu bị thiếu***



*Hình 2.10: Thực hiện kiểm tra dữ liệu thiếu*

*- Nhận xét:* Từ hình trên ta thấy bộ dữ liệu không có dữ liệu bị thiếu

***1.2.2 Xử lý dữ liệu trùng lặp***



*Hình 2.11: Thực hiện kiểm tra dữ liệu trùng lặp*

*- Nhận xét:* Từ hình trên ta thấy bộ dữ liệu không bị trùng lặp

* 1. ***Biến đổi dữ liệu***

*- Mục đích:* đưa các dữ liệu có trong từng thuộc tính về cùng 1 kiểu dữ liệu để tránh lỗi xung đột xảy ra khi huấn luyện mô hình

***1.3.1 Chuyển đổi dữ liệu***

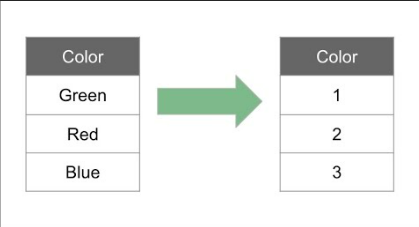


*Hình 2.12: Thực hiện biến đổi các dữ liệu trong thuộc tính rời rạc thành kiểu dữ liệu phân loại (category variable)*

***1.3.2 Mã hóa dữ liệu***

*- Mục đích:* Mã hóa dữ liệu là việc mã hóa các dữ liệu dạng chuỗi về dữ liệu dạng số để máy có thể hiểu được

- Ta sẽ dùng phương pháp LabelEncoder để mã hóa dữ liệu



*Hình 2.13: Ý tưởng của Label Encoder*

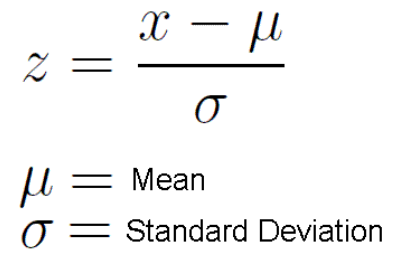


*Hình 2.14: Quá trình thực hiện mã hóa dữ liệu*

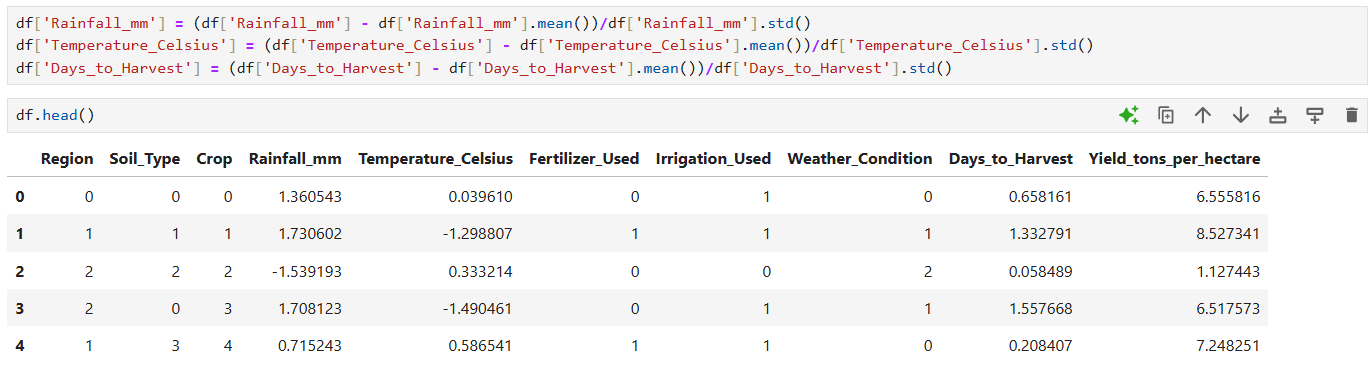
***1.3.3 Chuẩn hóa dữ liệu***

*- Mục đích:* là quá trình biến đổi các đặc trưng (features) của dữ liệu về cùng một thang đo (scale), thường là trong khoảng giá trị nhất định. Mục tiêu chính của chuẩn hóa là đảm bảo rằng tất cả các đặc trưng được đối xử công bằng khi mô hình học máy hoặc thống kê xử lý dữ liệu.

- Trong bài này ta sử dụng phương pháp chuẩn hóa là StandardScaler để chuẩn hóa dữ liệu cho các dữ liệu liên tục

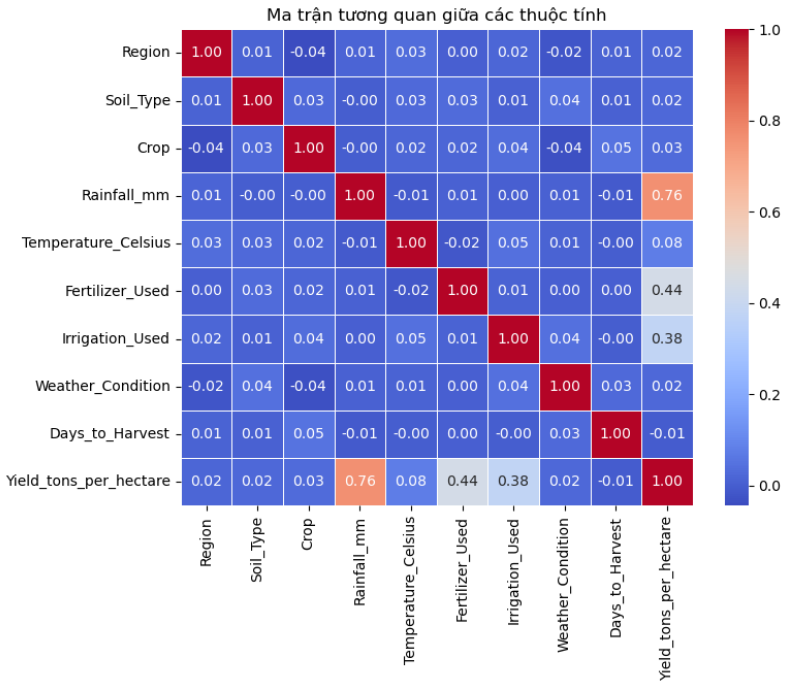


*Hình 2.15: Công thức của StandardScaler*

****

*Hình 2.16: Thực hiện chuẩn hóa dữ liệu*

1. ***Xây dựng và huấn luyện mô hình***
   1. ***Lựa chọn mô hình***



*Hình 2.17: Ma trận tương quan giữa các thuộc tính trong bộ dữ liệu*

- Dựa vào hình trên ta thấy các thuộc tính: Rainfall\_mm (Lượng mưa), Fertilizer\_Used (Có sử dụng phân bón), Irrigation\_Used (Có hay tưới tiêu) có sự ảnh hưởng lớn đến biến mục tiêu dựa vào hệ số tương quan lần lượt là: 0.76, 0.44, 0.38

- (1) Dựa vào việc trực quan hóa các thuộc tính trên với biến mục tiêu (đã làm ở mục 1.1) và ma trận tương quan ta thấy:

+> Lượng mưa nhiều cho sản lượng năng suất lớn

+> Việc sử dụng phân bón cho năng suất cao hơn

+> Việc hay tưới tiêu cho năng suất cao hơn

- (2) Dựa vào đầu ra của bài toán là năng suất của cây trồng (Biến liên tục)

=> Dựa trên (1) và (2) ta lựa chọn 2 mô hình phù hợp là: ***Linear Regression*** và ***Decision Tree Regressor*** cho bài toán dự đoán năng suất cây trồng trong nông nghiệp

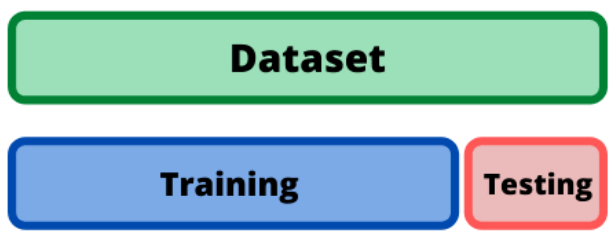
* 1. ***Phân tách dữ liệu***

*- Mục đích:*Việc chia tập dữ liệu giúp đảm bảo mô hình học máy hoạt động tốt trên dữ liệu thực tế, tránh sai lệch và overfitting, đồng thời cung cấp cách đánh giá khách quan để chọn mô hình tối ưu

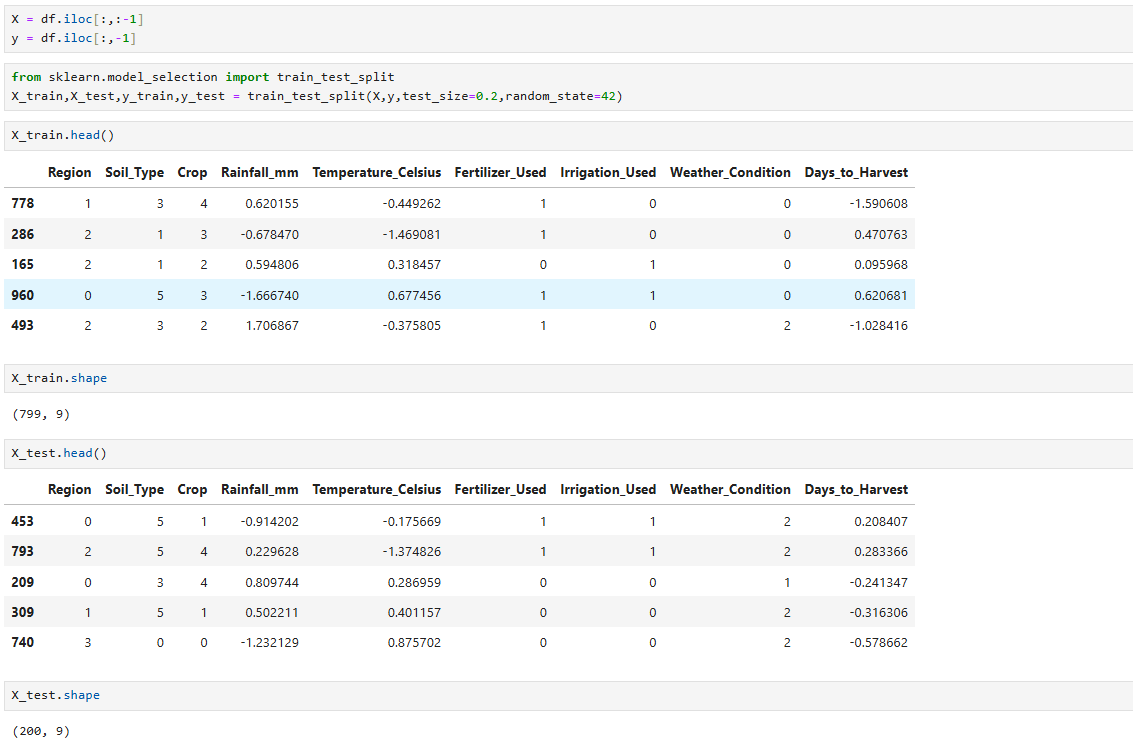
*- Cách chia:* Ta chia bộ dữ liệu ban đầu thành 2 phần

+> Phần Train (80%): Phần này dùng để huấn luyện mô hình

+> Phần Test (20%): Phần này dùng để kiểm tra và đánh giá mô hình

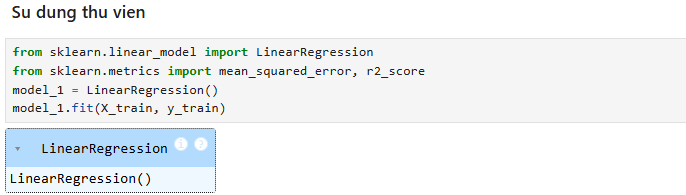


*Hình 2.18: Ảnh minh họa cách chia bộ dữ liệu*

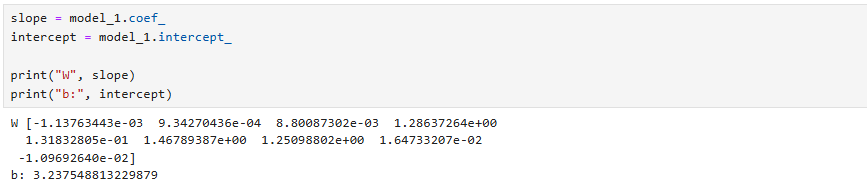


*Hình 2.19: Thực hiện phân tách bộ dữ liệu*

* 1. ***Xây dựng và huấn luyện với mô hình Linear Regression***
     1. ***Với thư viện***

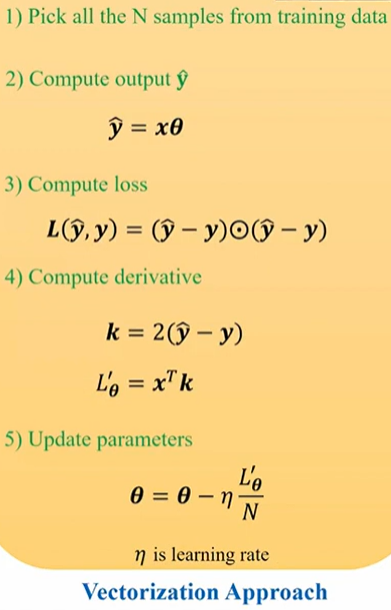
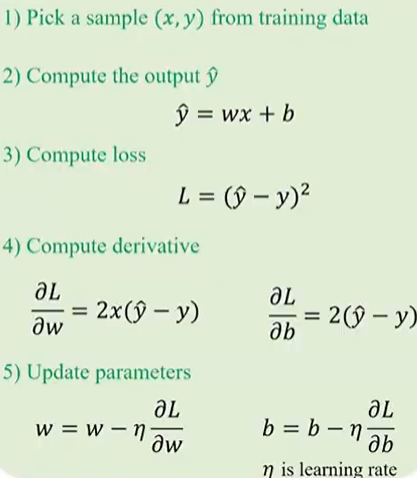


*Hình 2.20: Thực hiện huấn luyện mô hình với bộ Train*



*Hình 2.21: Sau khi huấn luyện ta được 1 bộ trọng số cho từng thuộc tính*

* + 1. ***Tự code***

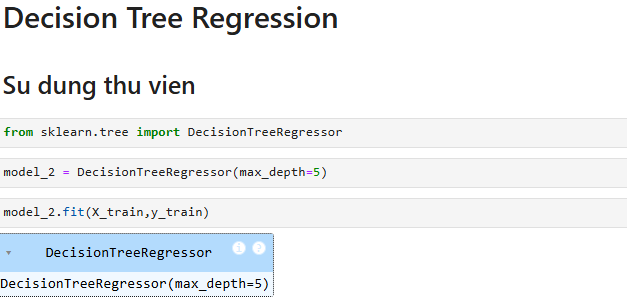


*Hình 2.22: Mô tả cách hoạt động của mô hình Linear Regression*

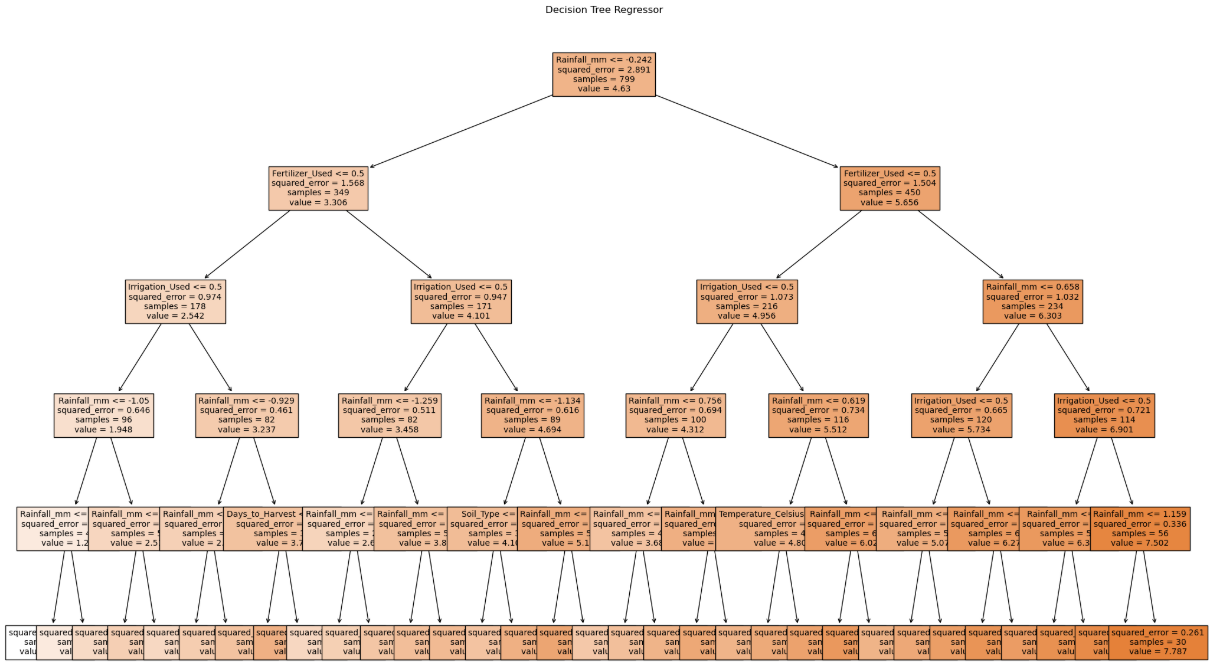


*Hình 2.23: Code theo các bước hoạt động ở trên*

* 1. ***Xây dựng và huấn luyện với mô hình Decision Tree Regressor***
     1. ***Với thư viện***

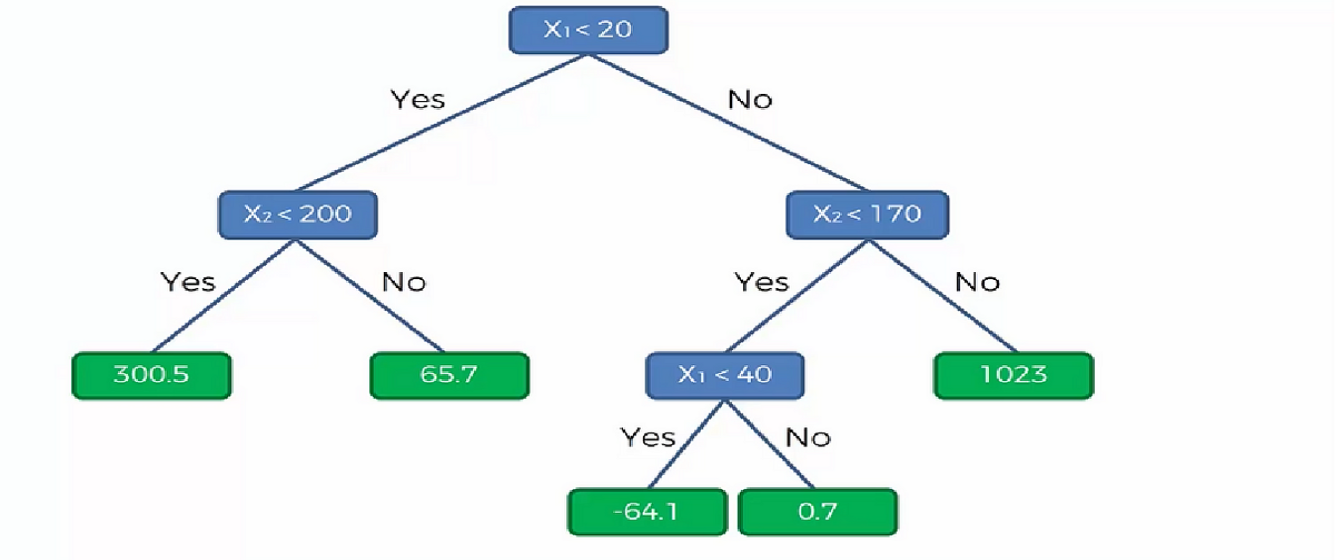


*Hình 2.24: Huấn luyện mô hình với bộ Train với chiều sâu của cây là 5*

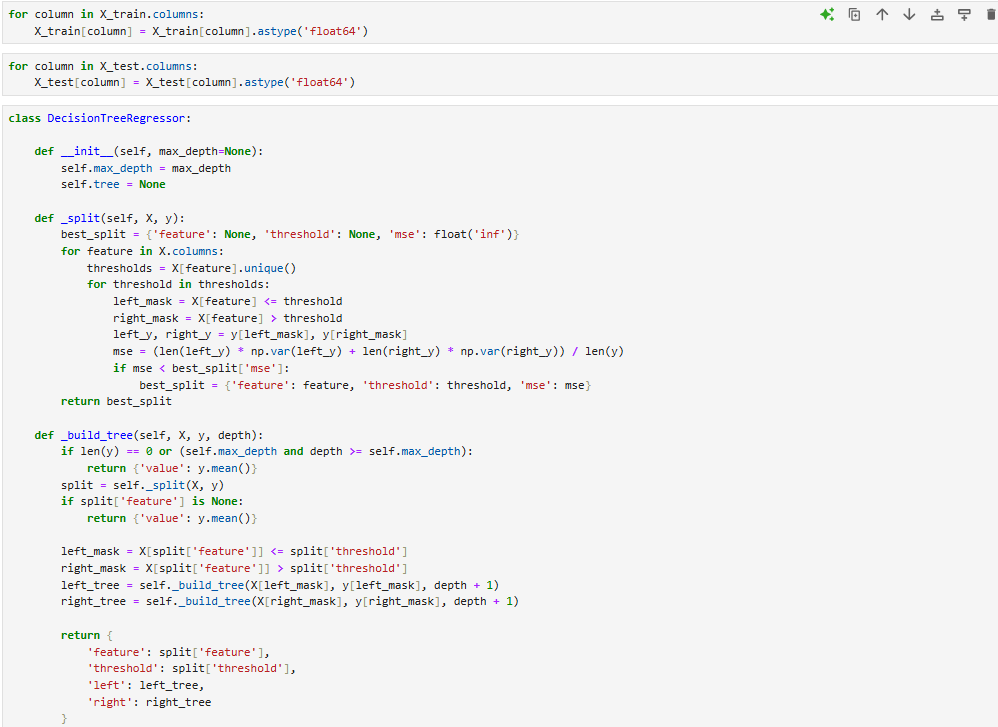


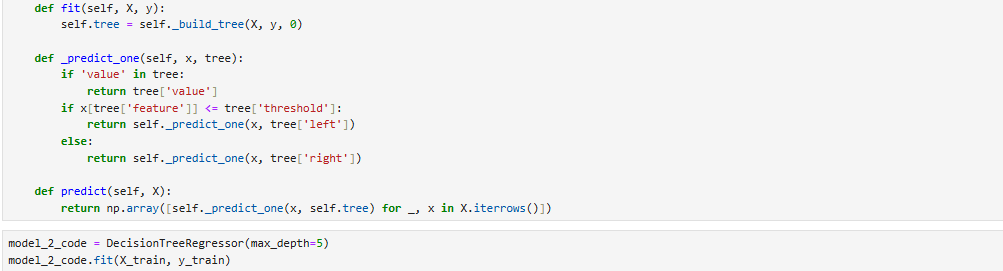
*Hình 2.25: Cây được xây dựng sau khi huấn luyện mô hình*

* + 1. ***Tự Code***

******

*Hình 2.26: Minh họa cách hoạt động của mô hình*



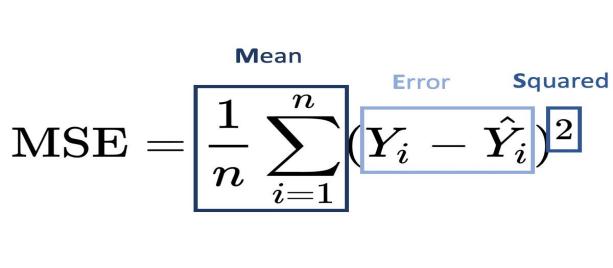


*Hình 2.27: Xây dựng mô hình Decision Tree for Regressor*

1. ***Đánh giá mô hình***

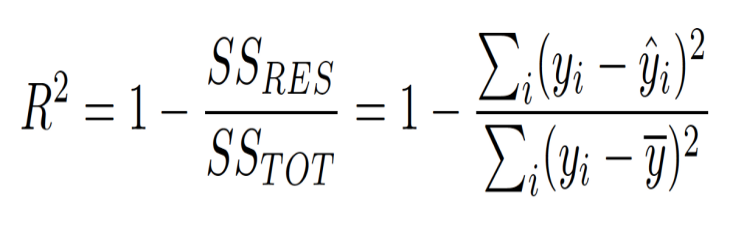
- Các thang đánh giá được sử dụng trong bài:

*+>MSE (Mean Squared Error):* Sai số trung bình bình phương thể hiện sai số giữa giá trị thực tế và giá trị dự đoán



Hình 2.28: Công thức của MSE

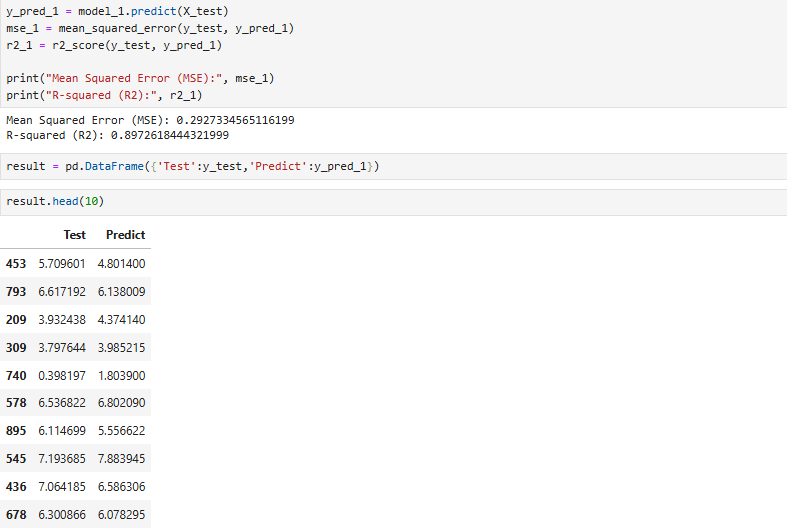
*+> R2 (hệ số tương quan r bình phương):* Cho biết mô hình đó hợp với dữ liệu ở mức bao nhiêu %. Giá trị R2 càng cao thì mối quan hệ giữa nhân tố độc lập (biến độc lập) và nhân tố phụ thuộc càng chặt chẽ.



*Hình 2.29: Công thức của R2*

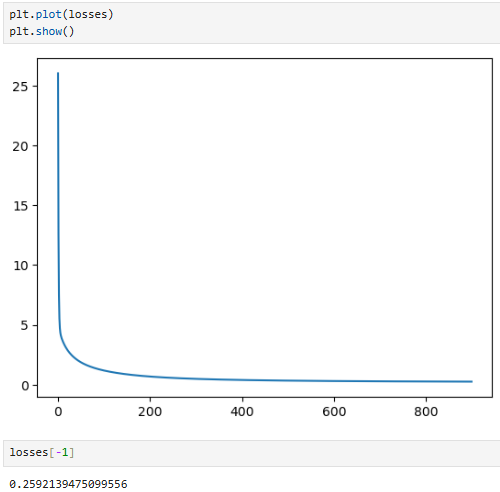
* 1. ***Đánh giá mô hình Linear Regression***

*- Với thư viện:*

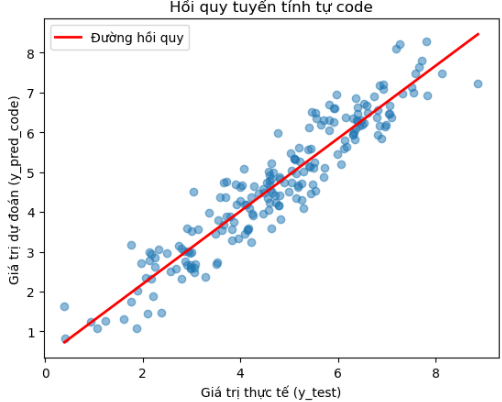
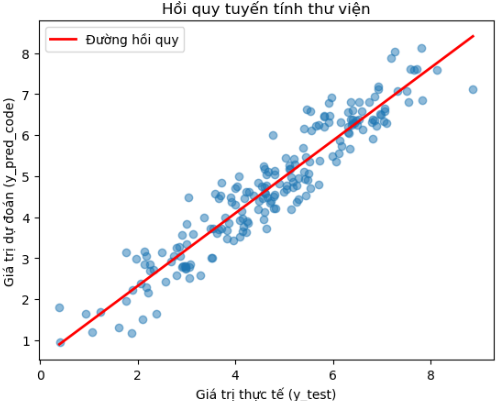


*Hình 2.30: Đánh giá mô hình với việc sử dụng thư viện*

*- Tự code:*



*Hình 2.31: Đánh giá mô hình tự code*



*Hình 2.32: Đường hồi quy của 2 mô hình so với bộ dữ liệu*

*- Nhận xét:*

+> Với hệ số MSE của mô hình tự code tốt hơn so với mô hình thư viện (0.2592<0.2927)

+> Với hệ số R2 của mô hình thư viện tốt hơn so với mô hình tự code khoảng 0.1 (0.8972>0.8880)

+> Đường hồi quy của 2 mô hình đã khớp 1 cách tốt nhất với bộ dữ liệu mặc dù vẫn còn sai số và độ chính xác chưa phải là tốt nhất

* 1. ***Đánh giá mô hình Decision Tree Regressor***

*- Với thư viện:*

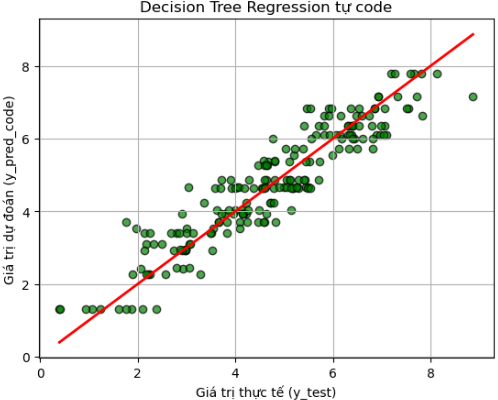


*Hình 2.33: Đánh giá mô hình với việc sử dụng thư viện*

*- Tự code:*



*Hình 2.34: Đánh giá mô hình với việc tự code*



*Hình 2.35: Đường hồi quy của 2 mô hình so với bộ dữ liệu*

*- Nhận xét:*

+> Với hệ số MSE (0.3663 < 0.3790) và R2 (0.8714>0.8669)của mô hình thư viện tốt hơn so với mô hình tự code, sai số vào khoảng 0.01

+> Đường hồi quy của 2 mô hình đã khớp 1 cách tốt nhất với bộ dữ liệu có

*- So sánh giữa 2 mô hình:*

* *Linear Regression:*

- Phù hợp nếu dữ liệu có mối quan hệ tuyến tính hoặc gần tuyến tính.

- Dùng khi cần giải thích rõ ràng về tác động của các biến độc lập.

- Thích hợp với bài toán cần đơn giản, nhanh chóng, và khả năng tổng quát hóa tốt.

* *Decision Tree Regressor:*

- Phù hợp với dữ liệu tuyến tính, phi tuyến tính hoặc có tương tác phức tạp giữa các biến.

- Dùng khi ưu tiên tính linh hoạt hoặc xử lý các giá trị ngoại lệ tốt.

- Cần cẩn thận với overfitting

*=> Đánh giá chung:*

+> Dựa vào nhận xét chung và thang đánh giá của 2 mô hình Linear Regression và Decision Tree for Regression ta thấy cả 2 mô hình đều phù hợp với bộ dữ liệu với sai số MSE thấp và R2 cao trong khoảng 87-89%, đường hồi quy khá khớp với dữ liệu.

+> Mô hình Linear Regression nhìn chung tốt hơn so với mô hình Decision Tree for Regression với mức MSE và R2 tốt hơn

+> Ta có thể hiểu chỉnh tham số, thu thập thêm dữ liệu hoặc sử dụng các phương pháp khác nhau để cải thiện độ chính xác cho mô hình

**KẾT LUẬN**

Trong quá trình thực hiện bài tập lớn môn Học Máy, nhóm chúng em đã hiểu và áp dụng các bước trong quy trình xây dựng và huấn luyện mô hình học máy để làm khung xây dựng lên bài tập lớn này.

Việc lựa chọn và xây dựng 2 mô hình Linear Regression và Decision Tree Regression đã giúp nhóm em hiểu thêm về cách hoạt động của các mô hình, cách đánh giá 1 mô hình là tốt hay tệ. Bên cạnh đó, phần quan trọng nhất là nhóm em biết được rằng việc tiền xử lý dữ liệu đã giúp cải thiện chất lượng của mô hình, tạo điều kiện cho các mô hình học máy hoạt động hiệu quả hơn.

Nhóm chúng em nhận thấy rằng việc áp dụng các phương pháp trong lĩnh vực nông nghiệp, có thể mang lại lợi ích lớn cho việc phát triển và tăng năng suất cho cây trồng. Tuy nhiên, để đạt được kết quả tốt hơn nữa, nhóm chúng em sẽ tìm thêm dữ liệu phong phú và chính xác hơn cũng như áp dụng các công cụ và phương pháp khác để phát triển mô hình dự đoán tối ưu hơn, áp dụng tốt vào trong thực tiễn.

**TÀI LIỆU THAM KHẢO**

1. <https://machinelearningcoban.com/2016/12/28/linearregression/>
2. <https://machinelearningcoban.com/2017/01/16/gradientdescent2/>
3. <https://seaborn.pydata.org/tutorial.html>
4. <https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot.html>
5. <https://machinelearningcoban.com/2018/01/14/id3/>
6. <https://scikit-learn.org/1.5/auto_examples/tree/plot_tree_regression.html>
7. https://openscience.vn/chi-tiet-du-lieu/bo-du-lieu-du-doan-nang-suat-cay-

trong-11369

1. "Machine Learning Yearning" - Andrew Ng
2. "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" - Aurélien Géron
3. https://scikit-learn.org/stable/