ĐỀ TÀI

DỰ ĐOÁN ĐIỂM THI THÔNG QUA THÓI QUEN HỌC TẬP

Sinh viên thực hiện : NGUYỄN XUÂN ĐỨC

Mã sinh viên : 12423057

Giảng viên hướng dẫn : PGS. TS NGUYỄN VĂN HẬU

Nội dung Thực Hiện

- 1.Lý do chọn đề tài
- 2.Giới thiệu dataset
- 3.Tiền xử lí dữ liệu
- 4.Biểu diễn dữ liệu
- 5.Các mô hình học máy áp dụng
- 6.Kết quả
- 7.Kết luận

Nội dung Thực Hiện

- **1.Lý do chọn đề tài**
- 2.Giới thiệu dataset
- 3.Tiền xử lí dữ liệu
- 4.Biểu diễn dữ liệu
- 5.Các mô hình học máy áp dụng
- 6.Kết quả
- 7.Kết luận

1. Lý do chọn ĐỀ TÀI

Lý do chọn đề tài:

- Thói quen học tập không đi kèm điểm số cụ thể.
- Cần ước lượng điểm tiềm năng để đánh giá kết quả học tập.
- Hỗ trợ học sinh cải thiện phương pháp học.

Mục tiêu nghiên cứu:

- Dự đoán điểm thi từ thói quen học tập.
- Hỗ trợ phân tích hành vi học tập tự động.
- Gợi ý cải thiện thói quen học để nâng cao điểm.

Nội dung Thực Hiện

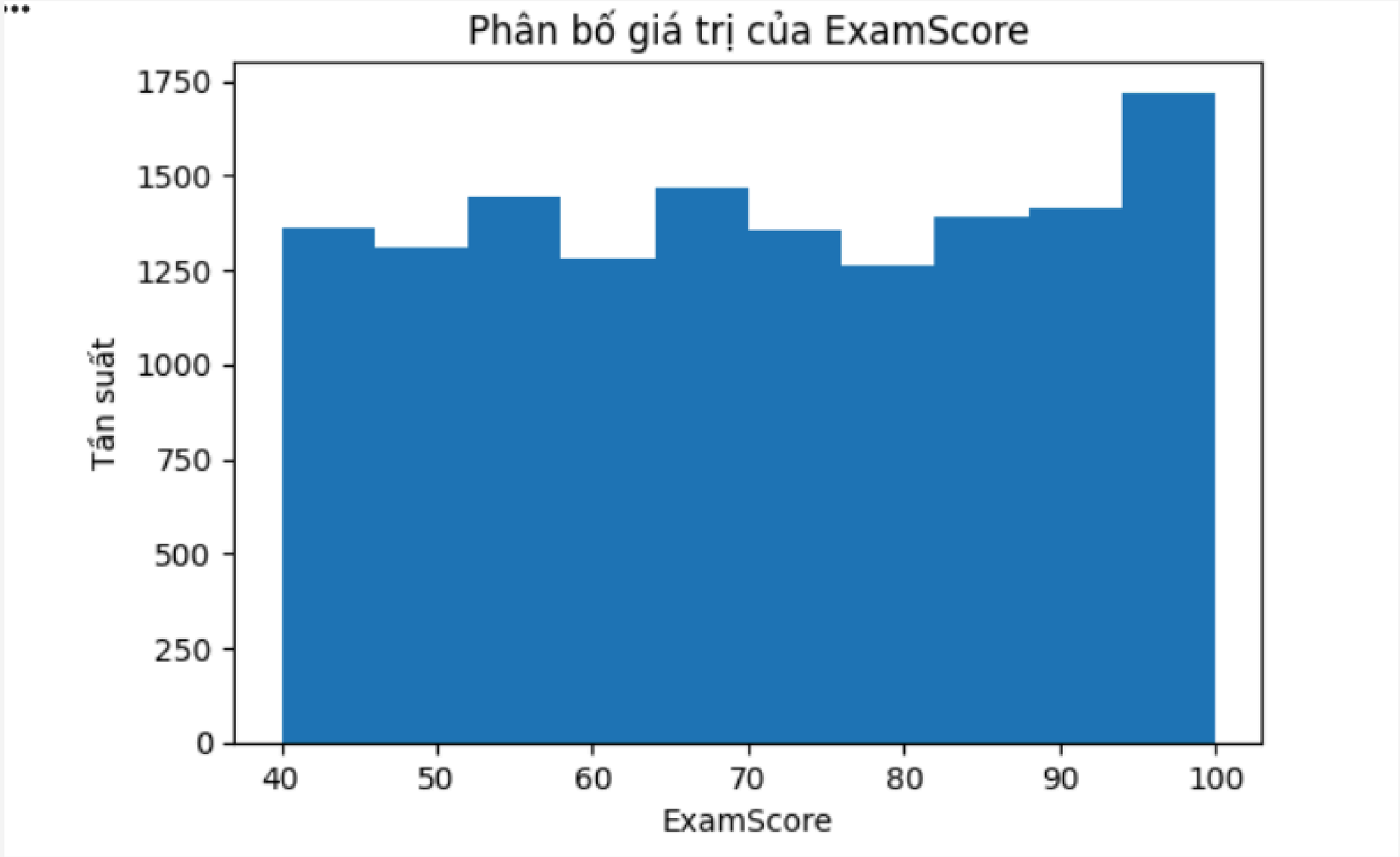
- 1.Lý do chọn đề tài
- **2.Giới thiệu dataset**
- 3.Tiền xử lí dữ liệu
- 4.Biểu diễn dữ liệu
- 5.Các mô hình học máp áp dụng
- 6.Kết quả
- 7.Kết luận

2. Giới thiệu DATASET

DATASET: Student Performance and Learning Behavior Dataset .csv

# StudyHours	# Attendance	# Resources	# Extracurri...	# Motivation	# Internet
19	64	1	0	0	1
19	64	1	0	0	1
19	64	1	0	0	1
19	64	1	1	0	1
19	64	1	1	0	1
19	64	1	1	0	1
19	64	0	1	0	1

2. Giới thiệu DATASET



Nội dung Thực Hiện

- 1.Lý do chọn đề tài
- 2.Giới thiệu dataset
- **3.Tiền xử lí dữ liệu**
- 4.Biểu diễn dữ liệu
- 5.Các mô hình học máp áp dụng
- 6.Kết quả
- 7.Kết luận

3. Tiền xử lí dữ liệu

Loại bỏ dữ liệu thiếu, sai lệch, hoặc trùng lặp.

Chuẩn hóa dữ liệu để mô hình học máy xử lý hiệu quả.

Biến đổi dữ liệu thô thành dạng có thể sử dụng cho mô hình.

Tăng độ chính xác và khả năng khái quát của mô hình.

Chia dữ liệu:

Train size: 9802

\

Validation size: 2100

Test size: 2101

3. Tiền xử lí dữ liệu

Kiểm tra dữ liệu và thông tin cơ bản

- **Khái niệm**: Kiểm tra số lượng bản ghi, kiểu dữ liệu, giá trị thiếu, trùng lặp, và phân bố dữ liệu.
- **Tác dụng**: Hiểu đặc điểm dữ liệu, phát hiện vấn đề trước khi xử lý và huấn luyện mô hình.

Xử lý dữ liệu categorical (danh mục)

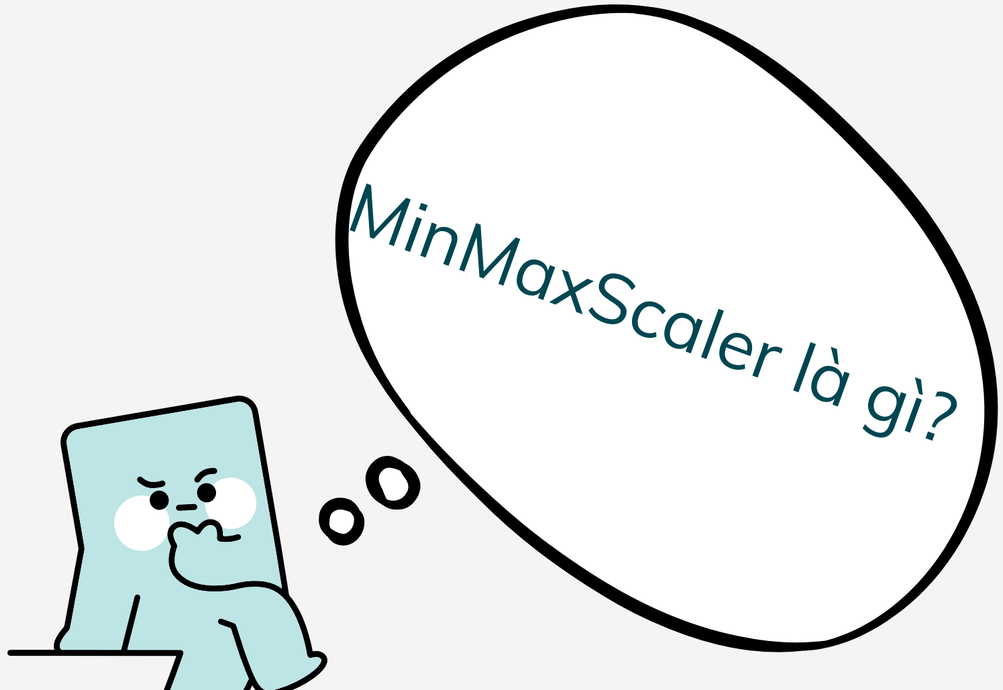
- **Khái niệm**: Biến định tính (ví dụ: phương pháp học, mức độ tập trung) được chuyển sang dạng số.
- **Cách thực hiện**: Sử dụng OneHotEncoder.
- **Tác dụng**: Giúp mô hình học máy hiểu và xử lý dữ liệu categorical.

3. Tiền xử lí dữ liệu

Chuẩn hóa dữ liệu số

- **Khái niệm:** Biến đổi các giá trị số về cùng thang đo (0–1) để mô hình học máy xử lý hiệu quả.
- **Cách thực hiện:** Sử dụng MinMaxScaler để chuyển tất cả các cột số về khoảng 0–1.
- **Tác dụng:** Tránh biến số lớn chi phối mô hình, giúp mô hình học nhanh hơn và dự đoán chính xác hơn

\



MinMaxScaler

Khái niệm : Là phương pháp chuẩn hóa dữ liệu số, biến đổi các giá trị về cùng một khoảng cố định, thường là 0–1.

Công thức :

$$\text{MinMaxScaler}: \ x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Tác dụng : MinMaxScaler hoạt động bằng cách trừ đi giá trị nhỏ nhất (để tạo giá trị bắt đầu từ 0) rồi chia cho (x_max - x_min) để giá trị nhỏ hơn hoặc bằng 1.

Chuẩn hóa số giờ học mỗi tuần

Dữ liệu

Học sinh	Giờ học/tuần
	21
	06
	16

Nội dung Thực Hiện

- 1.Lý do chọn đề tài
- 2.Giới thiệu dataset
- 3.Tiền xử lí dữ liệu
- **4.Biểu diễn dữ liệu**
- 5.Các mô hình học máp áp dụng
- 6.Kết quả
- 7.Kết luận

4.Biểu diễn dữ liệu

Histogram- Phân bố dữ liệu

Mô tả : Cho thấy phân bố điểm thi và số giờ học mỗi ngày. Dữ liệu gần chuẩn phân phối , giúp mô hình học máy dự đoán hiệu quả

Boxplot-Mức độ phân tán

Mô tả : Hiển thị mức độ phân tán của các số, giúp nhận diện các giá trị ngoại lai để xử lí trước khi huấn luyện mô hình

Bar chart- Biến nhị phân(0/1)

Mô tả : Thể hiện số lượng học sinh tham gia học nhóm. Giúp hình dung tần suất thói quen học tập trong dữ liệu

4.Biểu diễn dữ liệu

Scatter Plot - Mối quan hệ với điểm thi

Mô tả : Minh họa mối quan hệ giữa số giờ học và điểm thi. Xu hướng rõ ràng cho thấy giờ học có ảnh hưởng tích cực đến kết quả học tập

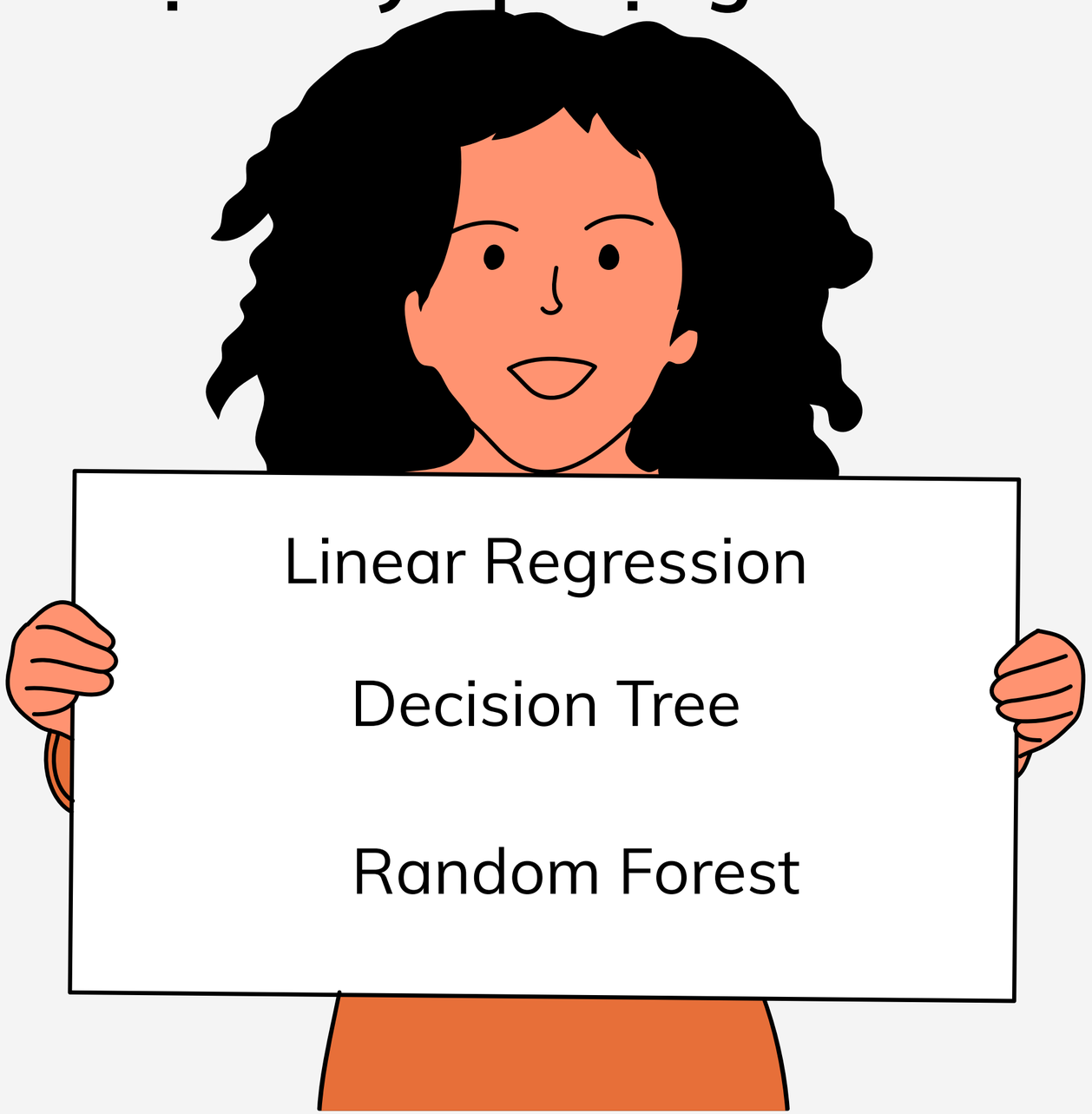
Heatmap- Mức độ tương quan

Mô tả : Hiển thị mức độ tương quan giữa các yếu tố thói quen học tập và điểm thi. Giúp lựa chọn các biến quan trọng cho mô hình dự đoán

Nội dung Thực Hiện

- 1.Lý do chọn đề tài
- 2.Giới thiệu dataset
- 3.Tiền xử lí dữ liệu
- 4.Biểu diễn dữ liệu
- **5.Các mô hình học máp áp dụng**
- 6.Kết quả
- 7.Kết luận

5.Các mô hình học máy áp dụng



5.Các mô hình học máy áp dụng

Linear Regression

Khái niệm : Linear Regression là mô hình dự đoán giá trị liên tục của biến đầu ra y y dựa trên mối quan hệ tuyến tính giữa các biến đầu vào x 1 , x 2 , . . . , x n x 1 ,x 2 ,...,x n

Công thức :

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

Trong đó :

- $\hat{y}$: giá trị dự đoán (điểm thi)
- x_i : biến đầu vào (thói quen học tập thứ i)
- w_i : trọng số phản ánh ảnh hưởng của x_i lên điểm thi
- b : bias (hằng số) – giá trị dự đoán khi tất cả $x_i = 0$

5.Các mô hình học máy áp dụng

Hàm mất mát (Loss Function – MSE)

Mục tiêu: Tìm w1,w2,...,wn và b sao cho giảm thiểu sai số bình phương trung bình

Công thức :

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

Trong đó :

- m : số lượng mẫu dữ liệu
- $y^{(i)}$: giá trị thực tế (điểm thi)
- $\hat{y}^{(i)}$: giá trị dự đoán từ mô hình

5.Các mô hình học máy áp dụng

Gradient Descent để cập nhật w và b

Định Nghĩa : Gradient Descent là thuật toán dùng để tìm cực tiểu của một hàm số, thường là hàm mất mát (loss function) trong quá trình huấn luyện mô hình

Công thức :

$$w_j := w_j - \alpha \frac{\partial J}{\partial w_j}, \quad b := b - \alpha \frac{\partial J}{\partial b}$$

Trong đó :

- α : learning rate
- $\frac{\partial J}{\partial w_j}$: đạo hàm MSE theo trọng số w_j
- $\frac{\partial J}{\partial b}$: đạo hàm MSE theo bias

5.Các mô hình học máy áp dụng

Decision Tree

Khái niệm: Là mô hình dự đoán liên tục bằng cách chia dữ liệu thành các nhánh dựa trên điều kiện của biến đầu vào.

Cách hoạt động và mục tiêu tối ưu :

Tại mỗi node (parent node) có N mẫu:

$$\bar{y}_{parent} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

MSE tại node :

$$MSE_{parent} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}_{parent})^2$$

5.Các mô hình học máy áp dụng

Random Forest

Khái niệm:

- Là tập hợp nhiều cây hồi quy (Decision Tree).
- Dự đoán cuối cùng bằng trung bình kết quả của tất cả các cây.

Cách hoạt động và mục tiêu tối ưu :

- Bootstrap sampling: Lấy nhiều tập con ngẫu nhiên từ dữ liệu huấn luyện.
- Xây dựng nhiều cây hồi quy: Mỗi cây tối ưu MSE tại node (như Decision Tree).
- Kết hợp dự đoán:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)}$$

Trong đó :

- T : số cây trong rừng
- $\hat{y}^{(t)}$: dự đoán của cây thứ t

5.Các mô hình học máy áp dụng

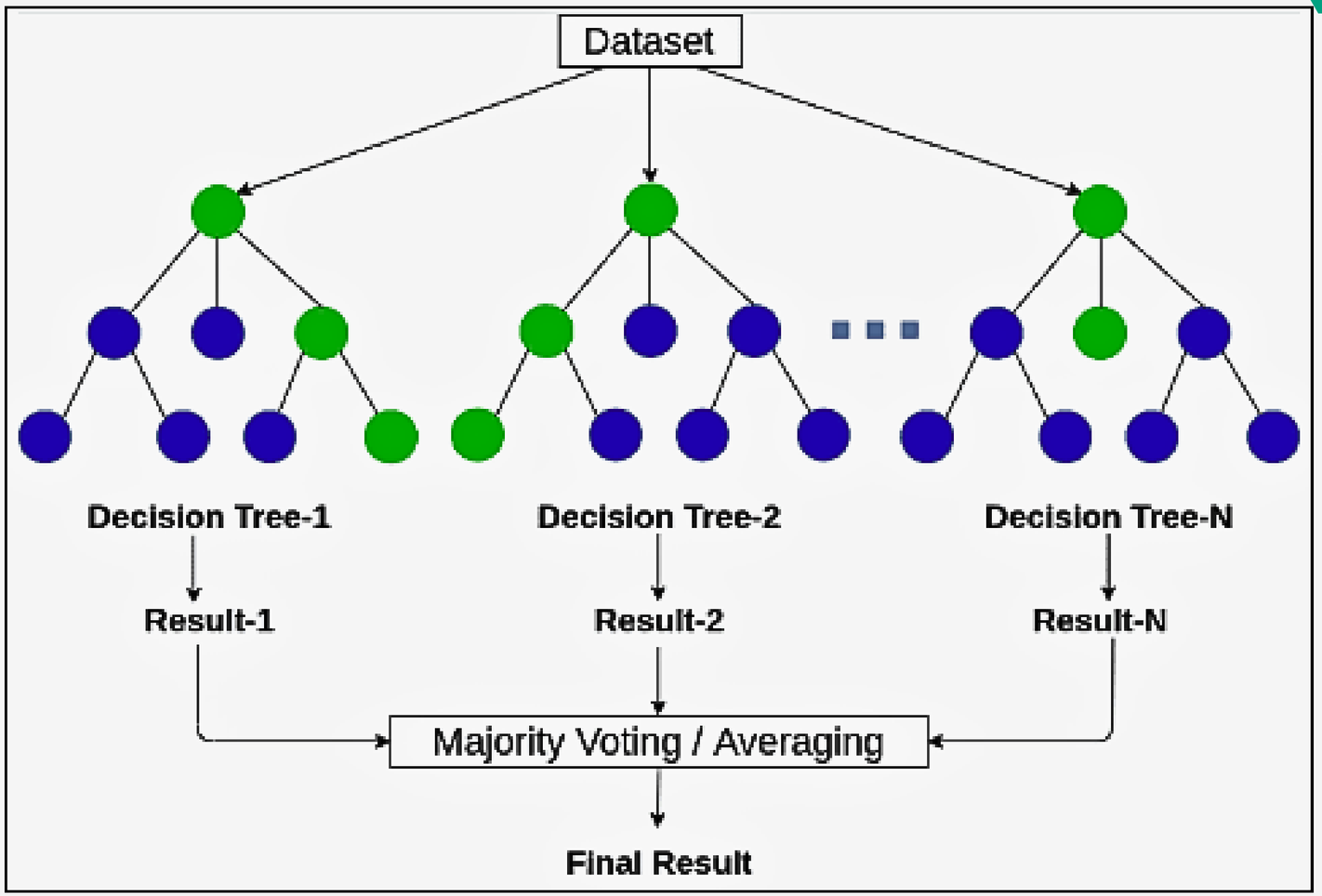
Random Forest

Ưu điểm:

- Dự đoán ổn định, chính xác.
- Giảm overfitting so với Decision Tree đơn lẻ.

Nhược điểm:

- Tốn tài nguyên tính toán.



5.Các mô hình học máy áp dụng

Metric đánh giá:

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$RMSE = \sqrt{MSE}$$

Nội dung Thực Hiện

- 1.Lý do chọn đề tài
- 2.Giới thiệu dataset
- 3.Tiền xử lí dữ liệu
- 4.Biểu diễn dữ liệu
- 5.Các mô hình học máp áp dụng
- **6.KẾT QUẢ**
- 7.Kết luận

6. Kết quả

Train & validation

	Model	Dataset	MAE	RMSE	R2
0	Linear Regression	Train	0.063480	0.073460	0.937670
1	Linear Regression	Validation	0.063310	0.072910	0.939210
3	Decision Tree	Train	0.000000	0.000000	1.000000
4	Decision Tree	Validation	0.011660	0.038460	0.983080
6	Random Forest	Train	0.007030	0.012220	0.998280
7	Random Forest	Validation	0.017800	0.031590	0.988590

Test

	Model	Dataset	MAE	RMSE	R2
2	Linear Regression	Test	0.063830	0.073780	0.938070
5	Decision Tree	Test	0.012340	0.038120	0.983470
8	Random Forest	Test	0.018540	0.031860	0.988450

Nội dung Thực Hiện

- 1.Lý do chọn đề tài
- 2.Giới thiệu dataset
- 3.Tiền xử lí dữ liệu
- 4.Biểu diễn dữ liệu
- 5.Các mô hình học máy áp dụng
- 6.Kết quả
- **7.KẾT LUẬN**

7. Kết luận

Mô hình	Ưu điểm	Nhược điểm	Hiệu quả / MSE	Ứng dụng / Bài toán phù hợp
Linear Regression	Đơn giản, dễ giải thích, dự đoán nhanh	Giả định tuyến tính, khó xử lý phi tuyến	MSE trung bình, tốt với dữ liệu tuyến tính	Dữ liệu tuyến tính, dự đoán điểm trung bình, phân tích ảnh hưởng tuyến tính
Decision Tree	Trực quan, xử lý tốt phi tuyến	Dễ overfitting, dự đoán thô	MSE cao hơn Linear Regression	Dữ liệu phi tuyến, muốn giải thích trực quan, phân nhánh học sinh theo nhóm
Random Forest	Dự đoán ổn định, giảm overfitting, xử lý dữ liệu phi tuyến tốt	Ít trực quan, tính toán phức tạp hơn	MSE thấp nhất, dự đoán gần đúng thực tế	Dữ liệu phức tạp, dự đoán điểm thi chính xác, các bài toán hồi quy tổng quát

DEMO

Q/A