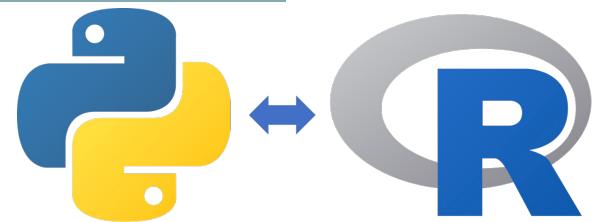


Phân tích dữ liệu với R/Python: [P3] – PHÂN TÍCH DỮ LIỆU VỚI PYTHON

ThS. Nguyễn Quang Phúc
phucnq@uel.edu.vn



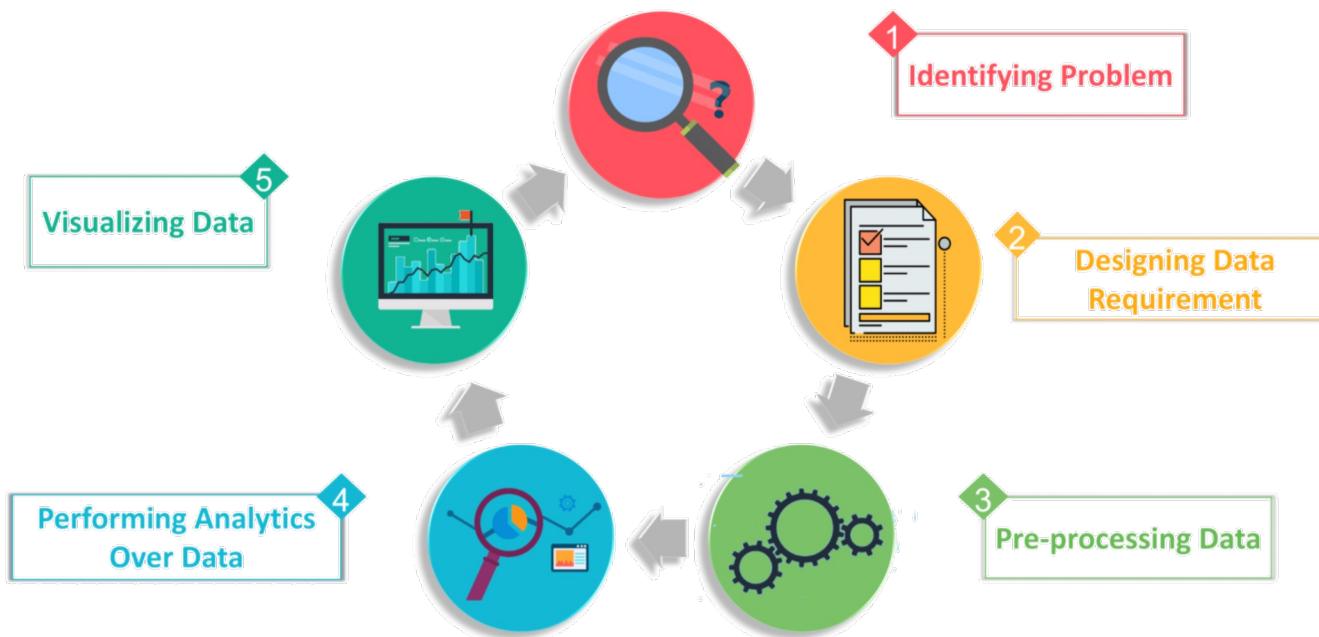
NỘI DUNG

1. Tổng quan về phân tích dữ liệu (data analysis).
2. Các dạng dữ liệu: chuỗi thời gian, dữ liệu chéo, dữ liệu bảng.
3. Một số mô hình cơ bản trong phân tích dự báo.
4. Phân tích dữ liệu chuỗi thời gian.
5. Ứng dụng máy học (machine learning) trong phân tích dữ liệu.

1. Tổng quan về phân tích dữ liệu (data analysis)

»»» Phân tích dữ liệu là gì?

Phân tích dữ liệu là một quá trình *kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu* với mục tiêu khám phá thông tin hữu ích, thông báo kết luận và hỗ trợ ra quyết định.



1. Tổng quan về phân tích dữ liệu (data analysis)

»»» Phân tích dữ liệu là gì?



Phân tích dữ liệu có nhiều khía cạnh và cách tiếp cận, bao gồm các kỹ thuật đa dạng dưới nhiều tên gọi khác nhau và được sử dụng trong các lĩnh vực kinh doanh, khoa học khác nhau.

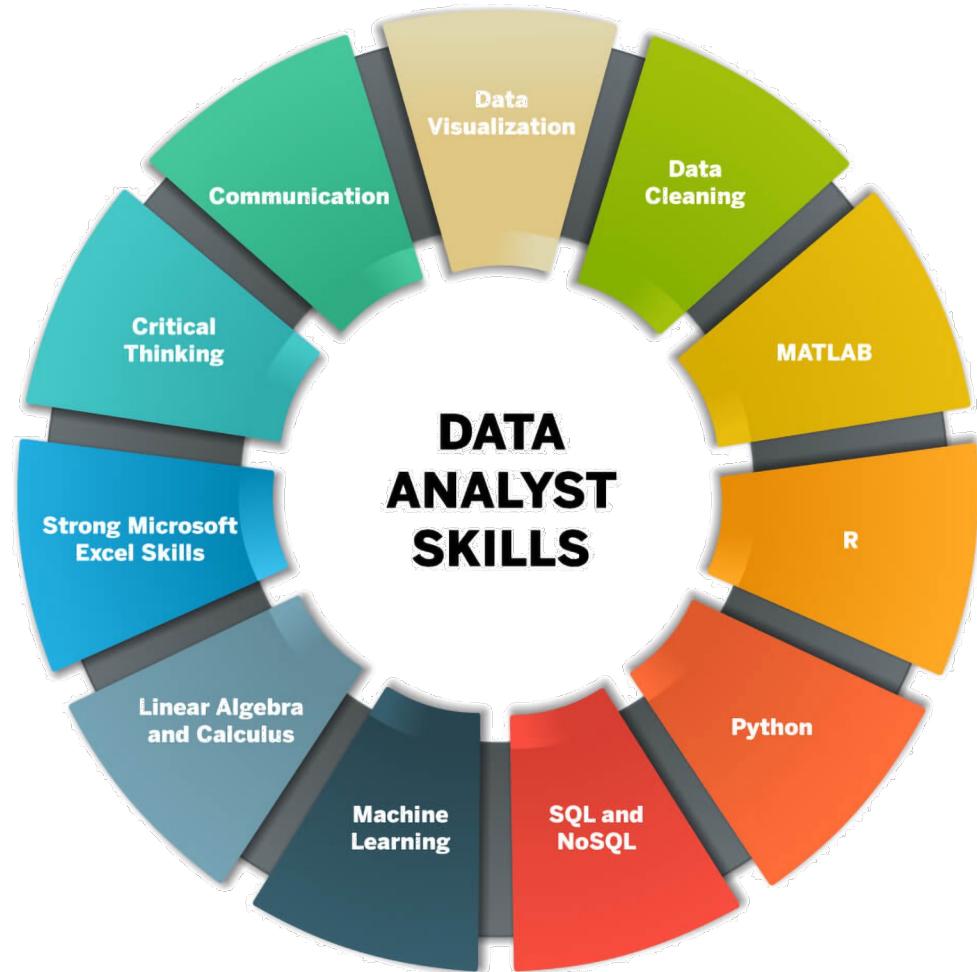


Trong lĩnh vực kinh doanh ngày nay, phân tích dữ liệu đóng vai trò *giúp đưa ra quyết định khoa học* và giúp doanh nghiệp hoạt động hiệu quả hơn.

1. Tổng quan về phân tích dữ liệu (data analysis)

»» Kỹ năng cần thiết?

1. Domain Expertise
2. Programming Skill
3. Visualization Skill
4. Statistical Knowledge
5. Story Telling
6. ...

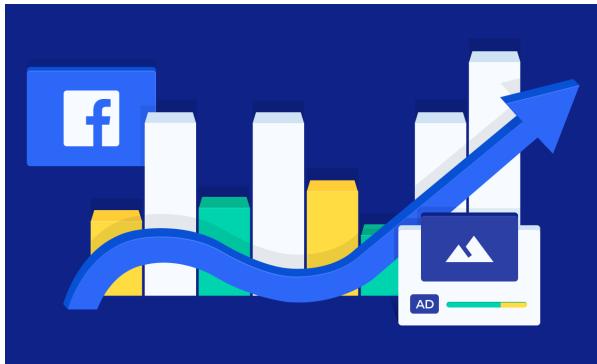


1. Tổng quan về phân tích dữ liệu (data analysis)

»»» Xu hướng nghề nghiệp

Dữ liệu là nguồn tài nguyên quan trọng cho mọi hoạt động sản xuất, kinh doanh.

Với hầu hết các lĩnh vực như: thương mại, marketing, tài chính, ngân hàng, bảo hiểm,... đều cần có nhân sự có chuyên môn về phân tích dữ liệu để làm cơ sở, định hướng cho các hoạt động của doanh nghiệp.



Facebook, Google sở hữu số lượng chuyên viên phân tích dữ liệu hàng đầu thế giới để thực hiện phân tích dữ liệu người dùng, làm cơ sở để hoạch định các chiến lược kinh doanh cũng như định hướng phát triển trong tương lai.

1. Tổng quan về phân tích dữ liệu (data analysis)

»»» Xu hướng nghề nghiệp

- ✓ Chuyên viên phân tích dữ liệu (Data Analyst)
- ✓ Chuyên viên phân tích dữ liệu kinh doanh (Business Analyst)
- ✓ Chuyên viên phân tích định lượng (Quantitative Analyst)
- ✓ Kỹ sư khoa học dữ liệu (Data Scientist)
- ✓ ...



2. Các dạng dữ liệu

»» Chuỗi thời gian

Chuỗi thời gian là một bảng dữ liệu với nhiều cột khác nhau, nhưng *bắt buộc phải có cột dữ liệu thể hiện thời gian và các biến cố thay đổi theo cột thời gian đó*.

	Date	High	Low	Open	Close	Avg	Volume
0	2016-01-04	47.5	45.2	45.6	47.5	46.81	4809120.0
1	2016-01-05	47.7	46.8	47.0	47.5	47.29	2480100.0
2	2016-01-06	47.9	46.7	47.4	47.5	47.16	2001950.0
3	2016-01-07	48.2	46.5	46.8	48.0	47.44	2852010.0
4	2016-01-08	48.0	47.0	47.5	48.0	47.86	1641950.0

1246	2020-12-25	106.4	105.0	105.7	105.9	105.83	370300.0
1247	2020-12-28	106.6	105.4	105.9	105.9	105.96	711710.0
1248	2020-12-29	106.6	105.8	106.1	106.5	106.25	612360.0
1249	2020-12-30	109.6	106.7	106.7	108.5	108.74	1528950.0
1250	2020-12-31	109.3	105.5	108.5	108.2	108.60	656040.0

2. Các dạng dữ liệu

»» Chuỗi thời gian

Chuỗi thời gian với cấu trúc là một bảng dữ liệu có nhiều cột khác nhau, nhưng *bắt buộc phải có cột dữ liệu thể hiện thời gian và các biến cố thay đổi theo cột thời gian đó*.

Year	CPI	Lai_suat	GTSX_CN
2007M1	111.0	6.5	49,212.0
2007M2	113.4	6.5	35,392.0
2007M3	113.1	6.5	45,154.0
2007M4	113.7	6.5	47,344.6
2007M5	114.5	6.5	47,953.4

2. Các dạng dữ liệu

»» Dữ liệu chéo

Dữ liệu chéo là một bảng dữ liệu với nhiều cột khác nhau, nhưng *bắt buộc phải có cột dữ liệu thể hiện chủ thể nghiên cứu khác nhau và các biến khác cần xem xét*. Các chủ thể có thể được hiểu là “không gian” cần phân tích, đánh giá.

Tỉnh	GTSX_CN	GTSX_TM
Hà Nội	2345	1244
HCM	2436	1242
Đà Nẵng	3454	1222
Hải Phòng	2333	1111

2. Các dạng dữ liệu

»»» Dữ liệu bảng

Cấu trúc dữ liệu bảng là sự kết hợp giữa cấu trúc dữ liệu chuỗi thời gian và dữ liệu chéo.

Năm	Tỉnh	GTSX_TM	GTSX_CN
2008	Ha Noi	1244	4577
2009	Ha Noi	1242	4575
2010	Ha Noi	1222	4555
2011	Ha Noi	1111	4444
2008	HCM	2244	5577
2009	HCM	2242	5575
2010	HCM	1422	4755
2011	HCM	1151	4484

3. Một số mô hình cơ bản trong phân tích dự báo

»»» Regression

Regression (hồi quy): là một trong những kỹ thuật thống kê và học máy cơ bản. Hồi quy giúp tìm ra *mối quan hệ giữa các biến* → mối quan hệ này được sử dụng để *dự đoán* (tiên lượng) các giá trị trong tương lai.

Ví dụ: bạn có thể quan sát một số nhân viên của một số công ty và cố gắng tìm hiểu mức lương của họ phụ thuộc vào các yếu tố nào, chẳng hạn như kinh nghiệm, trình độ học vấn, thành phố họ làm việc, ...

Tương tự, bạn có thể cố gắng thiết lập một sự phụ thuộc toán học của giá nhà vào diện tích, số phòng ngủ, thời gian xây dựng, khoảng cách đến trung tâm thành phố, ...

3. Một số mô hình cơ bản trong phân tích dự báo

»»» Linear Regression

Linear Regression (hồi quy tuyến tính): là một trong những kỹ thuật hồi quy được sử dụng rộng rãi. Đây là một trong những phương pháp hồi quy đơn giản nhất. Một trong những ưu điểm chính của nó là dễ dàng giải thích kết quả.

$$Y = \alpha + \beta X + \varepsilon$$

α : intercept

β : gradient / slope

ε : sai số ngẫu nhiên (những giao động về Y trong mỗi giá trị X)

3. Một số mô hình cơ bản trong phân tích dự báo

»»» Linear Regression

Giả định:

- Mỗi liên quan giữa X và Y là tuyến tính và tham số.
- X không có sai số ngẫu nhiên.
- Giá trị của Y là độc lập với nhau.
- Sai số ngẫu nhiên (ε): có phân bố chuẩn, trung bình 0, phương sai bất biến.

$$\varepsilon \sim N(0, \sigma^2)$$

3. Một số mô hình cơ bản trong phân tích dự báo

»» Linear Regression

Mô hình hồi quy tổng thể (PRF):

$$Y = \alpha + \beta X + \varepsilon$$

$$E(Y|X_i) = \alpha + \beta X_i$$

Chúng ta không biết α và β nhưng có thể dùng dữ liệu thực nghiệm để ước tính 2 tham số đó.

3. Một số mô hình cơ bản trong phân tích dự báo

»»» Linear Regression

Mô hình hồi quy mẫu (SRF):

$$\hat{Y}_i = a + bX_i$$

- \hat{Y}_i là ước lượng của $E(Y_i|X_i)$.
- a, b là ước lượng của α và β .

$$Y_i = a + bX_i + e_i = \hat{Y}_i + e_i$$

3. Một số mô hình cơ bản trong phân tích dự báo

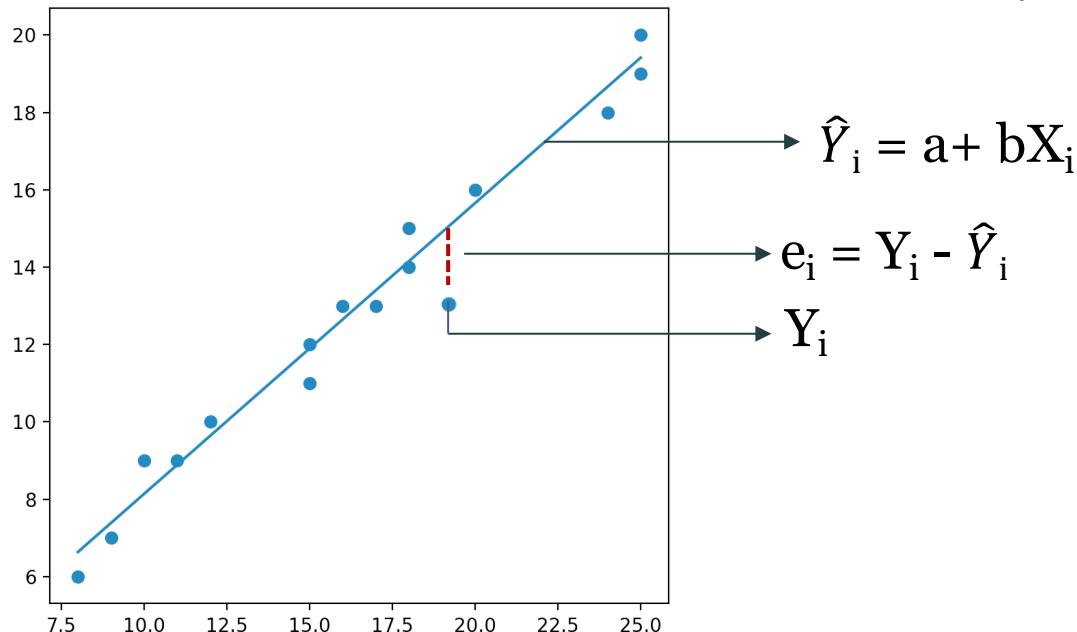
»»» Linear Regression

Mô hình hồi quy mẫu (SRF):

$$Y_i = a + bX_i + e_i = \hat{Y}_i + e_i$$

Tìm a, b sao cho $\sum e_i^2 \rightarrow \min.$

Ordinary Least Square (OLS)



3. Một số mô hình cơ bản trong phân tích dự báo

»»» Linear Regression

Vd: với dữ liệu quan sát thu nhập (triệu đồng) và chi tiêu (triệu đồng) của các hộ gia đình hãy ước tính mối liên quan giữa thu nhập và chi tiêu.

$$\hat{Y}_i = a + bX_i$$

$$\text{Expenditure} = a + b^* \text{Income}$$

	Income	Expenditure
0	8	6
1	9	7
2	10	9
3	11	9
4	12	10
5	15	12
6	15	11
7	16	13
8	17	13
9	18	15
10	18	14
11	20	16
12	24	18
13	25	20
14	25	19

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.983			
Model:	OLS	Adj. R-squared:	0.982			
Method:	Least Squares	F-statistic:	747.5			
Date:	Mon, 10 May 2021	Prob (F-statistic):	7.13e-13			
Time:	10:30:00	Log-Likelihood:	-11.998			
No. Observations:	15	AIC:	28.00			
Df Residuals:	13	BIC:	29.41			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.6207	0.470	1.321	0.209	-0.394	1.636
x1	0.7518	0.027	27.341	0.000	0.692	0.811
Omnibus:	1.567	Durbin-Watson:	2.716			
Prob(Omnibus):	0.457	Jarque-Bera (JB):	0.855			
Skew:	0.045	Prob(JB):	0.652			
Kurtosis:	1.834	Cond. No.	53.9			

3. Một số mô hình cơ bản trong phân tích dự báo

»»» Linear Regression

Vd: dự báo giá căn hộ theo diện tích và số lượng phòng ngủ.

$$\widehat{\text{Price}} = a + b * \text{Area}$$



$$\widehat{\text{Price}} = a + b * \text{Bedrooms}$$

$$\widehat{\text{Price}} = a + b * \text{Area} + c * \text{Bedrooms}$$

	Price	Area	Bedrooms
0	4.70000	74	2.50000
1	5.30000	80	3.00000
2	6.60000	100	3.00000
3	2.60000	50	1.00000
4	5.50000	80	2.50000
5	2.30000	48	1.50000
6	4.50000	74	2.00000
7	7.50000	113	3.00000
8	5.60000	80	2.50000
9	2.40000	48	1.50000
10	4.10000	69	2.50000
11	6.20000	95	3.00000
12	8.30000	135	3.00000
13	3.20000	56	2.00000
14	2.80000	50	1.50000
15	6.30000	95	2.00000
16	5.40000	80	2.00000
17	2.20000	48	1.00000
18	7.70000	113	3.00000

3. Một số mô hình cơ bản trong phân tích dự báo

»»» Linear Regression

Vd: dự đoán lượng khí thải CO₂ của một chiếc xe hơi dựa trên trọng lượng, dung tích xi lanh của động cơ.

$$\widehat{CO_2} = a + b * \text{Weight}$$



$$\widehat{CO_2} = a + b * \text{Volume}$$

$$\widehat{CO_2} = a + b * \text{Weight} + c * \text{Volume}$$

	Car	Model	Volume	Weight	CO2
0	Toyoyt	Aygo	1000	790	99
1	Mitsubishi	Space Star	1200	1160	95
2	Skoda	Citigo	1000	929	95
3	Fiat	500	900	865	90
4	Mini	Cooper	1500	1140	105
5	VW	Up!	1000	929	105
6	Skoda	Fabia	1400	1109	90
7	Mercedes	A-Class	1500	1365	92
8	Ford	Fiesta	1500	1112	98
9	Audi	A1	1600	1150	99
10	Hyundai	I20	1100	980	99
11	Suzuki	Swift	1300	990	101
12	Ford	Fiesta	1000	1112	99
13	Honda	Civic	1600	1252	94
14	Hyundai	I30	1600	1326	97
15	Opel	Astra	1600	1330	97
16	BMW	1	1600	1365	99
17	Mazda	3	2200	1280	104
18	Skoda	Rapid	1600	1119	104
19	Ford	Focus	2000	1328	105
20	Ford	Mondeo	1600	1584	94
21	Opel	Insignia	2000	1428	99
22	Mercedes	C-Class	2100	1365	99
23	Skoda	Octavia	1600	1415	99
24	Volvo	S60	2000	1415	99
25	Mercedes	CLA	1500	1465	102
26	Audi	A4	2000	1490	104
27	Audi	A6	2000	1725	114
28	Volvo	V70	1600	1523	109
29	BMW	5	2000	1705	114
30	Mercedes	E-Class	2100	1605	115

3. Một số mô hình cơ bản trong phân tích dự báo

»» Logistic Regression

Logistic Regression (hàm quy logistic): là một kỹ thuật thống kê xem xét mối liên hệ giữa biến độc lập (*biến liên tục hoặc nhị phân*) và biến thuộc (*biến nhị phân*).

Linear Regression

- Giá nhà
- Giá cổ phiếu
- Doanh số
- Hàng tồn kho
- Sức mua của KH
- ...

Logistic Regression

- Phân loại KH
- Phân loại SP
- Thư spam?
- ...

3. Một số mô hình cơ bản trong phân tích dự báo

»» Logistic Regression

Logistic Regression (hàm quy logistic): là một kỹ thuật thống kê xem xét mối liên hệ giữa biến độc lập (*biến liên tục hoặc nhị phân*) và biến phụ thuộc (*biến nhị phân*).

$$y = \alpha + \beta x + \varepsilon$$

y: biến phụ thuộc với 2 trạng thái (0/1; true/false; yes/no)

→ Mô hình hàm quy logistic được phát biểu như sau:

$$\text{logit}(p) \rightarrow \log\left(\frac{p}{1-p}\right) = \alpha + \beta x \Rightarrow \text{Odds ratio} = \exp(\beta)$$

p là xác suất biến cõi xảy ra và 1-p là xác suất biến cõi không xảy ra

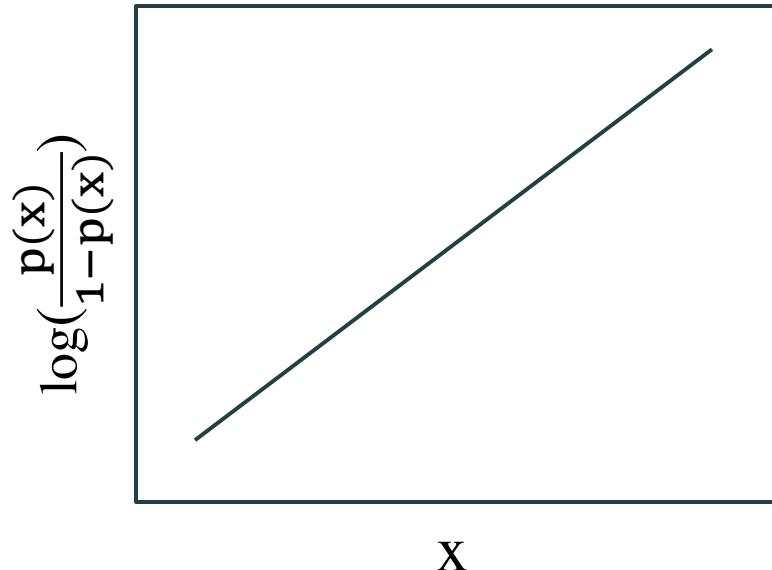
→ Xác suất tiên lượng theo trị số của x:

$$\text{Odds} \rightarrow \frac{p}{1-p} = e^{\alpha + \beta x} \quad p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

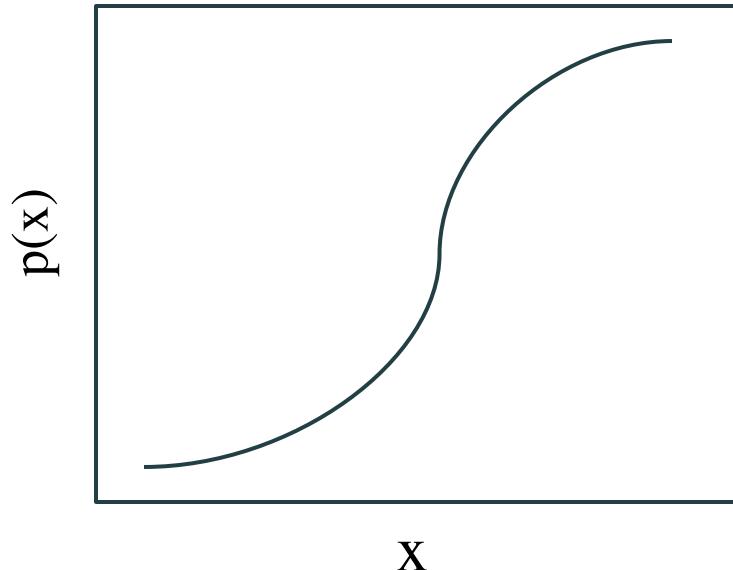
3. Một số mô hình cơ bản trong phân tích dự báo

»» Logistic Regression

$$\log\left(\frac{p}{1-p}\right)$$



$$p = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$



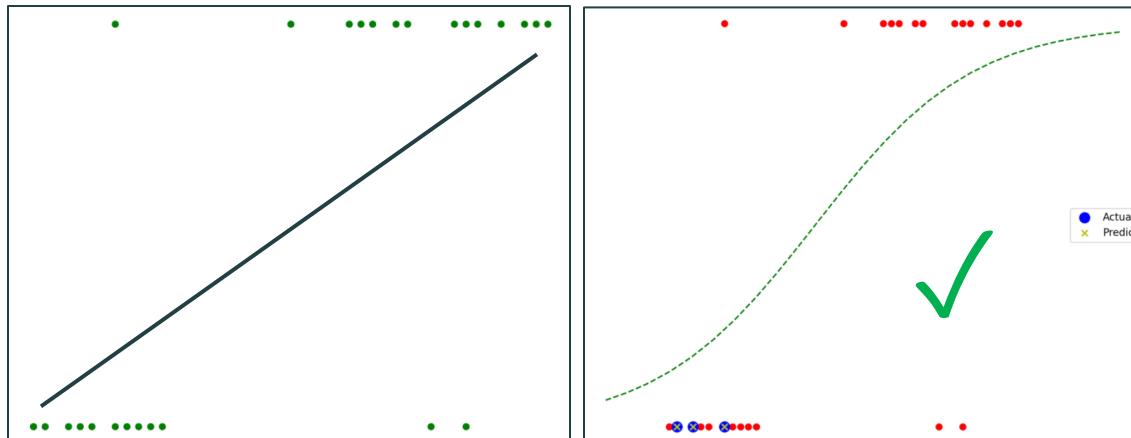
3. Một số mô hình cơ bản trong phân tích dự báo

»»» Logistic Regression

Vd: dự đoán khả năng mua bảo hiểm của khách hàng dựa theo độ tuổi:

$$\text{logit}(p) = a + b * \text{age}$$

Với p là xác suất mua bảo hiểm



$$\text{logit}(p) = -4.0389 + 0.1042 * \text{age}$$

$$\Rightarrow \text{Odds ratio} = \exp(0.1042) = 1.11$$

?

	age	bought_insurance
0	22	0
1	25	0
2	47	1
3	52	0
4	46	1
5	56	1
6	55	0
7	60	1
8	62	1
9	61	1
10	18	0
11	28	0
12	27	0
13	29	0
14	49	1
15	55	1
16	25	1
17	58	1
18	19	0
19	18	0
20	21	0
21	26	0
22	40	1
23	45	1
24	50	1
25	54	1
26	23	0

🏡: SV tìm hiểu thêm về mô hình hồi quy Logistic với biến độc lập là *biến nhị phân*, biến độc lập là *biến thứ bậc*.

3. Một số mô hình cơ bản trong phân tích dự báo

»»» Logistic Regression

Vd: dự đoán khả năng mua bảo hiểm của khách hàng dựa theo độ tuổi:

$$\text{logit}(p) = a + b * \text{age}$$

Với p là xác suất mua bảo hiểm

Logit Regression Results			
Dep. Variable:	bought_insurance	No. Observations:	27
Model:	Logit	Df Residuals:	25
Method:	MLE	Df Model:	1
Date:	Fri, 30 Jul 2021	Pseudo R-squ.:	0.4543
Time:	19:37:50	Log-Likelihood:	-10.203
converged:	True	LL-Null:	-18.696
Covariance Type:	nonrobust	LLR p-value:	3.764e-05

	coef	std err	z	P> z 	[0.025	0.975]
const	-5.2729	1.814	-2.907	0.004	-8.828	-1.718
age	0.1357	0.044	3.118	0.002	0.050	0.221

	♦ age	♦ bought_insurance
0	22	0
1	25	0
2	47	1
3	52	0
4	46	1
5	56	1
6	55	0
7	60	1
8	62	1
9	61	1
10	18	0
11	28	0
12	27	0
13	29	0
14	49	1
15	55	1
16	25	1
17	58	1
18	19	0
19	18	0
20	21	0
21	26	0
22	40	1
23	45	1
24	50	1
25	54	1
26	23	0

4. Phân tích dữ liệu chuỗi thời gian

»» Dữ liệu chuỗi thời gian (Time Series)

Chuỗi thời gian với cấu trúc là một bảng dữ liệu có nhiều cột khác nhau, nhưng *bắt buộc phải có cột dữ liệu thể hiện thời gian và các biến cố thay đổi theo cột thời gian đó*.

Ví dụ:

- Dữ liệu về sức mua của KH theo năm, quý, tháng, ...
- Dữ liệu giá cổ phiếu, lượng giao dịch theo năm, tháng, quý, tuần, ngày, ...
-

Ứng dụng của việc phân tích
dữ liệu chuỗi thời gian?

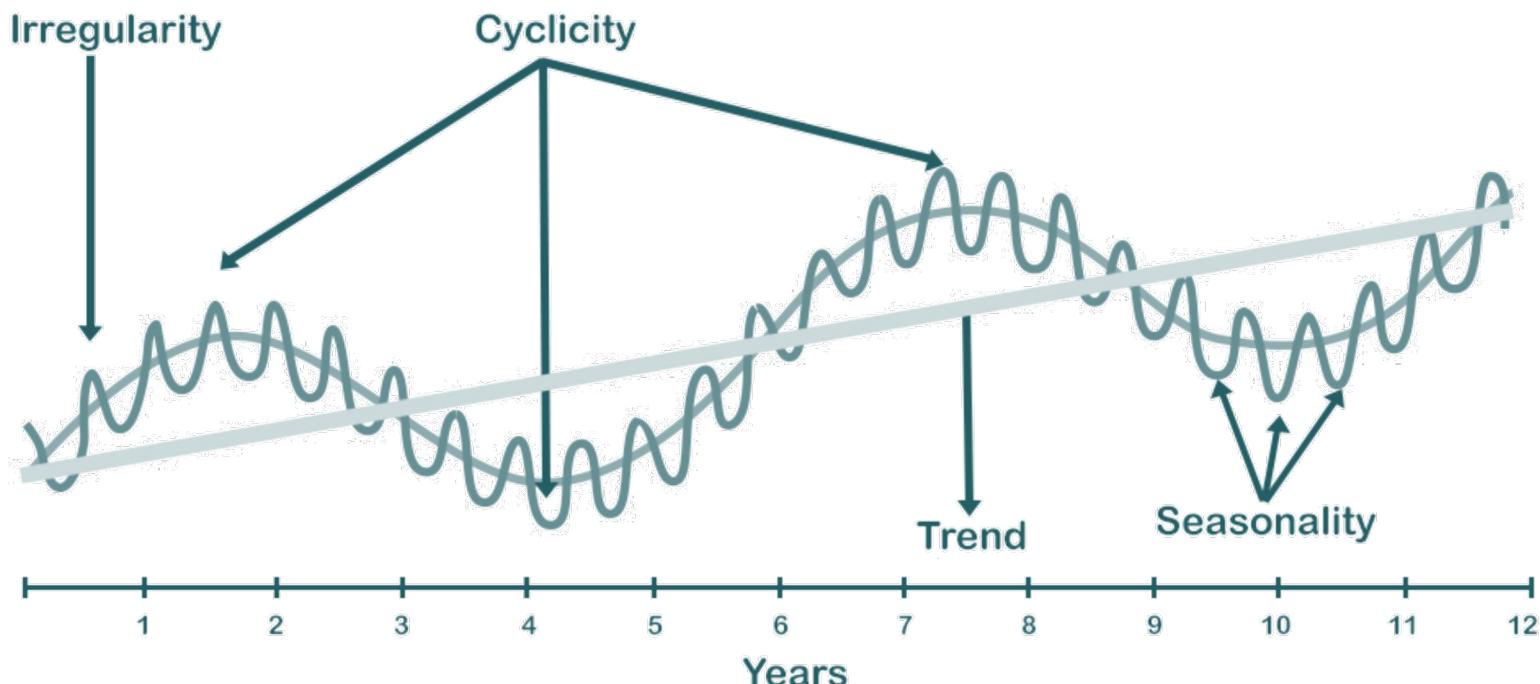
	Date	High	Low	Open	Close	Avg	Volume
0	2016-01-04	47.5	45.2	45.6	47.5	46.81	4809120.0
1	2016-01-05	47.7	46.8	47.0	47.5	47.29	2480100.0
2	2016-01-06	47.9	46.7	47.4	47.5	47.16	2001950.0
3	2016-01-07	48.2	46.5	46.8	48.0	47.44	2852010.0
4	2016-01-08	48.0	47.0	47.5	48.0	47.86	1641950.0

1246	2020-12-25	106.4	105.0	105.7	105.9	105.83	370300.0
1247	2020-12-28	106.6	105.4	105.9	105.9	105.96	711710.0
1248	2020-12-29	106.6	105.8	106.1	106.5	106.25	612360.0
1249	2020-12-30	109.6	106.7	106.7	108.5	108.74	1528950.0
1250	2020-12-31	109.3	105.5	108.5	108.2	108.60	656040.0

4. Phân tích dữ liệu chuỗi thời gian

»» Dữ liệu chuỗi thời gian (Time Series)

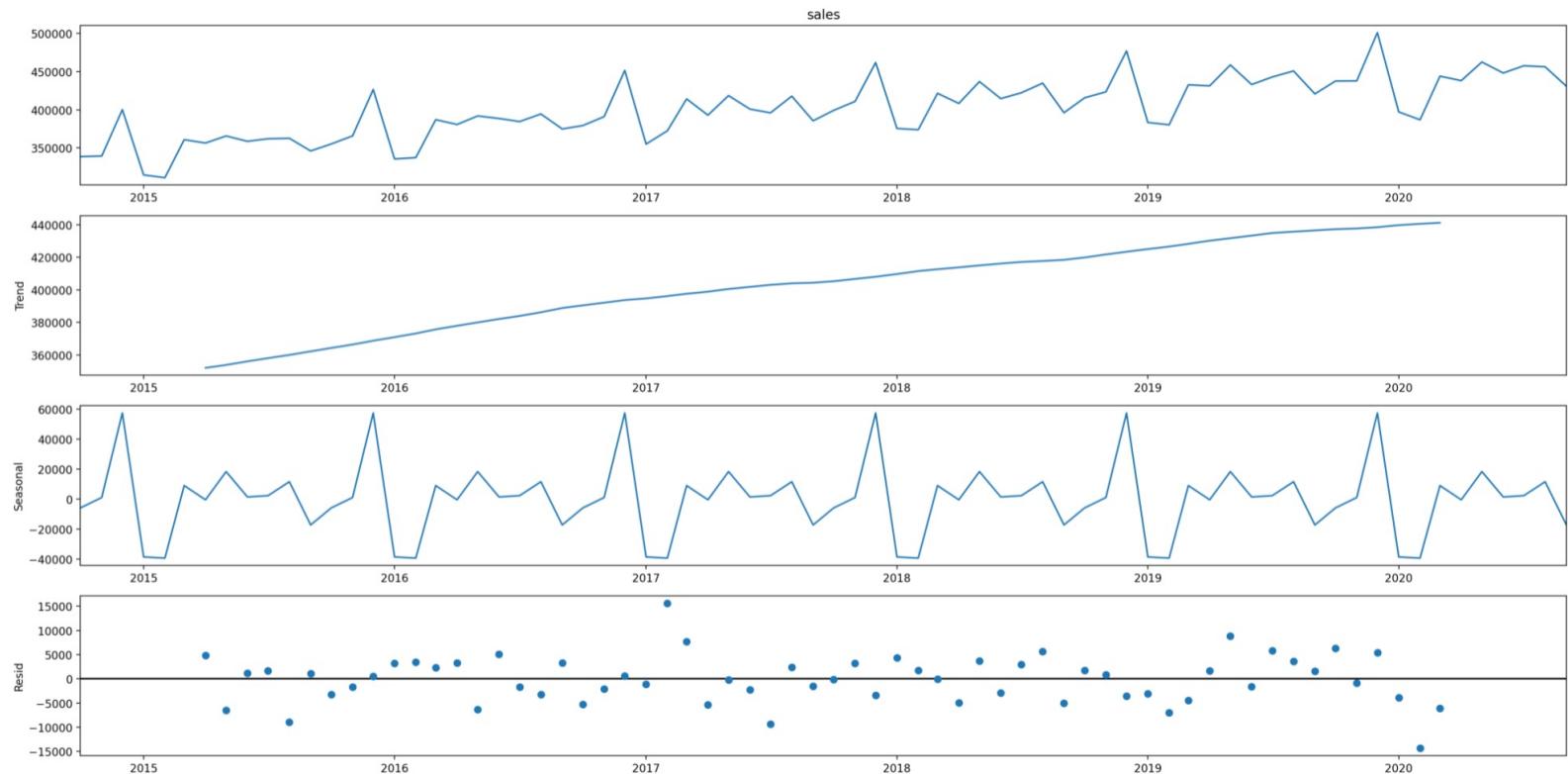
▷ Các thành phần



4. Phân tích dữ liệu chuỗi thời gian

»» Dữ liệu chuỗi thời gian (Time Series)

▷ *Các thành phần*



4. Phân tích dữ liệu chuỗi thời gian

»» Dữ liệu chuỗi thời gian (Time Series)

$$Y_t = \text{Signal}_t + \text{Noise}_t$$

▷ Mô hình tự hồi quy bậc k

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_k Y_{t-k} + \varepsilon_t$$

Ví dụ: $\text{Exp}_t = \alpha + \beta_1 \text{Exp}_{t-1} + \varepsilon_t$

▷ Mô hình trễ phân phối hữu hạn

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \dots + \beta_k X_{t-k} + \varepsilon_t$$

Ví dụ: $\text{Exp}_t = \alpha + \beta_0 \text{Inc}_t + \beta_1 \text{Inc}_{t-1} + \varepsilon_t$

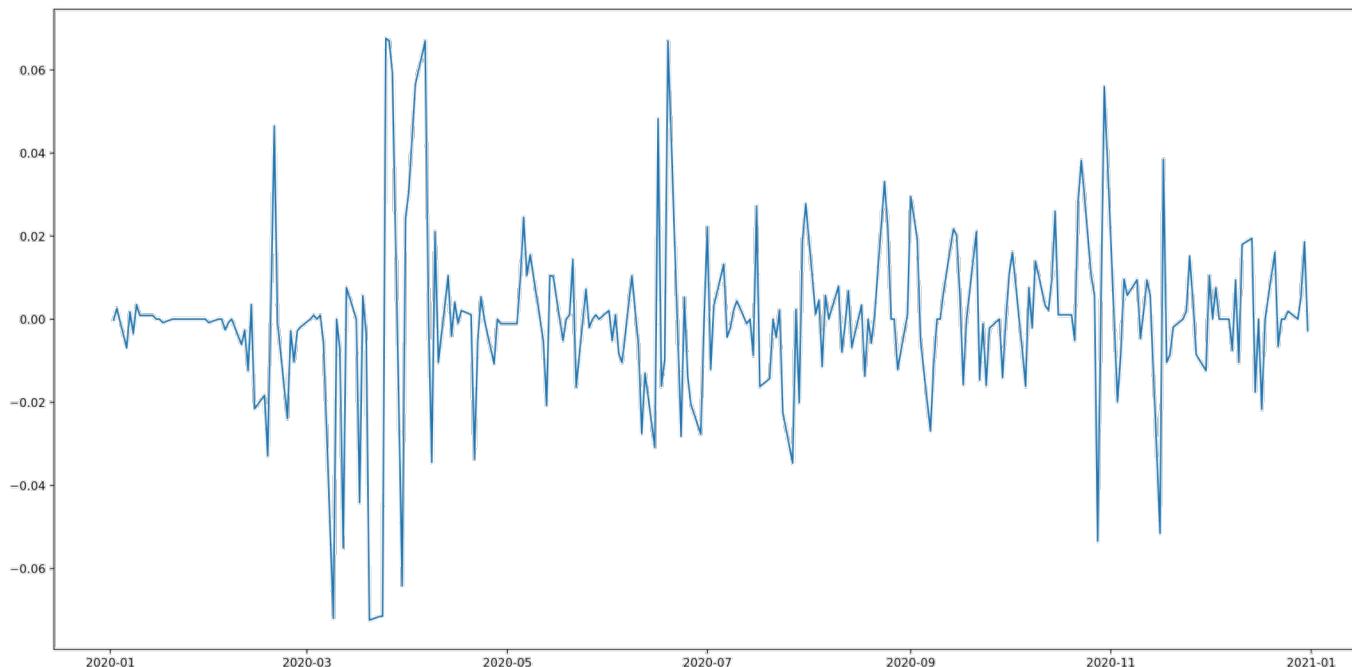
4. Phân tích dữ liệu chuỗi thời gian

»» Dữ liệu chuỗi thời gian (Time Series)

▷ *Tính dừng - Stationary*

$$Y_t \text{ dừng} \Leftrightarrow \begin{cases} E(Y_t) = \mu \\ \text{Var}(Y_t) = \sigma^2 \\ \text{Cov}(Y_t, Y_{t-p}) = \gamma_p \notin t \end{cases}$$

Tại sao cần kiểm tra tính dừng trong phân tích dữ liệu chuỗi thời gian?

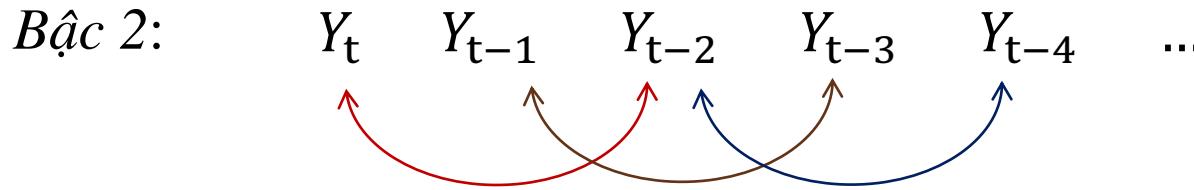


4. Phân tích dữ liệu chuỗi thời gian

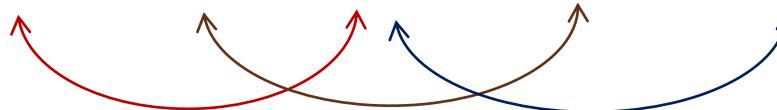
»» Dữ liệu chuỗi thời gian (Time Series)

▷ *Tương quan - Correlation*

Bậc 1: $Y_t \leftrightarrow Y_{t-1} \leftrightarrow Y_{t-2} \leftrightarrow Y_{t-3} \leftrightarrow Y_{t-4} \leftrightarrow \dots$



Bậc 1&2: $Y_t \leftrightarrow Y_{t-1} \leftrightarrow Y_{t-2} \leftrightarrow Y_{t-3} \leftrightarrow Y_{t-4} \leftrightarrow \dots$



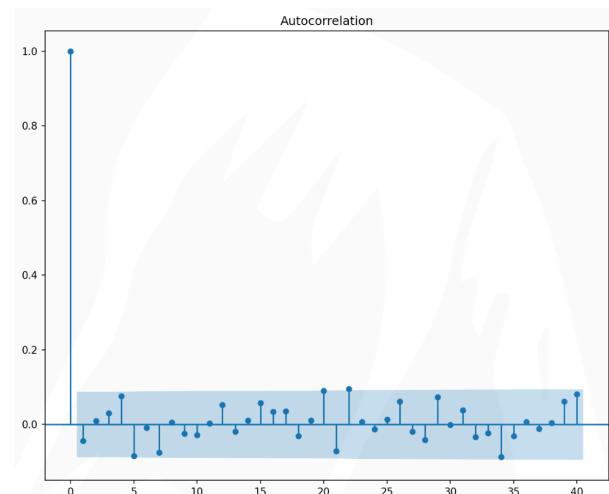
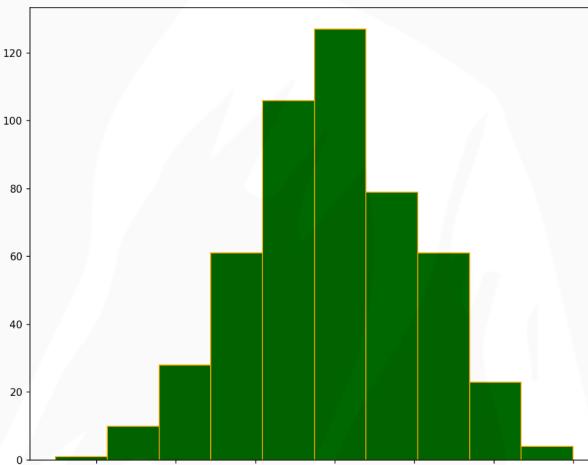
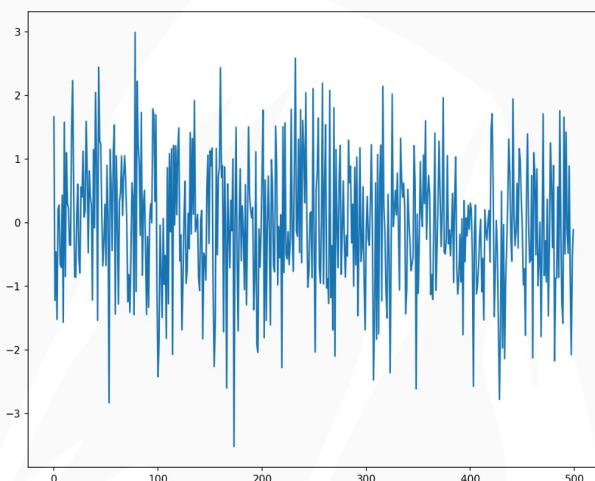
4. Phân tích dữ liệu chuỗi thời gian

»»» Dữ liệu chuỗi thời gian (Time Series)

▷ *Nhiễu trắng – White noise*

ε_t nhiễu trắng \Leftrightarrow

$$\left\{ \begin{array}{l} E(\varepsilon_t) = 0 \\ \text{Var}(\varepsilon_t) = \sigma^2 \\ \text{Cov}(\varepsilon_t, \varepsilon_{t-p}) = 0 \end{array} \right.$$



Tương tự nhiễu trắng

$$Y_t = \alpha_1 + \alpha_2 \varepsilon_t$$

Tại sao cần nhận diện “nhiễu trắng” trong phân tích dữ liệu chuỗi thời gian?

4. Phân tích dữ liệu chuỗi thời gian

»»» Một số mô hình phổ biến



AR (Auto-Regression) – Mô hình tự hồi quy: mô hình hóa mối quan hệ giữa dữ liệu và độ trễ của chính nó → thể hiện qua tham số ‘p’ trong mô hình ARIMA

I (Integrated) – Tích hợp: việc sử dụng sai phân nhằm loại bỏ tính xu hướng trong dữ liệu → thể hiện qua tham số ‘d’ trong mô hình ARIMA

MA (Moving Average) – Trung bình trượt: mô hình hóa mối quan hệ giữa dữ liệu và sai số so với trung bình của các độ trễ khác → thể hiện qua tham số ‘q’ trong mô hình ARIMA

➔ ARIMA(p, d, q)

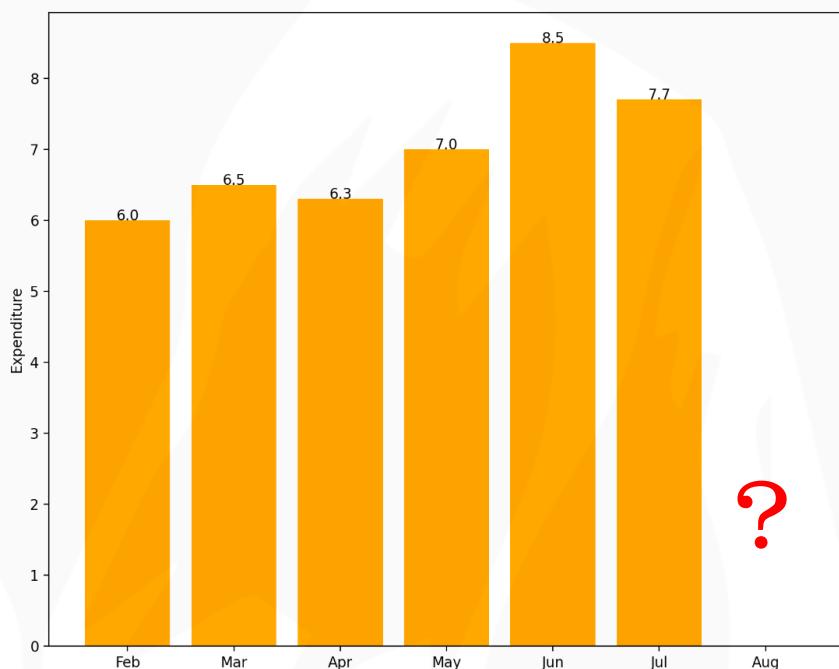
4. Phân tích dữ liệu chuỗi thời gian

»»» Một số mô hình phổ biến

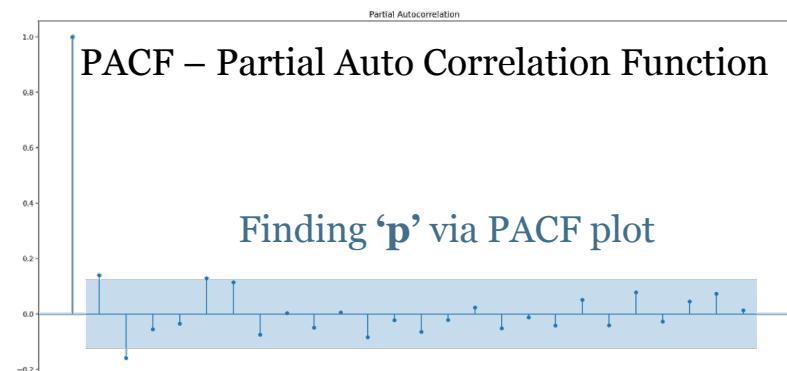
▷ $AR(p)$ – AutoRegression

$$AR(1): Y_t = \alpha + \beta_1 Y_{t-1} + \varepsilon_t$$

$$AR(2): Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t$$



Xác định p ?



$$Exp_{Aug} = \alpha + \beta_1 Exp_{Jul} + \beta_2 Exp_{Jun} + Error$$

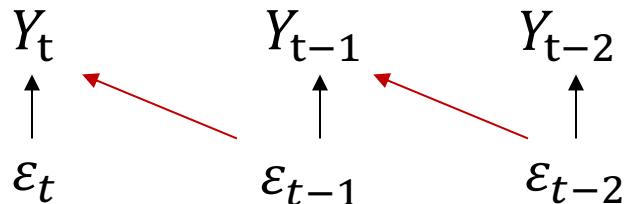
4. Phân tích dữ liệu chuỗi thời gian

»»» Một số mô hình phổ biến

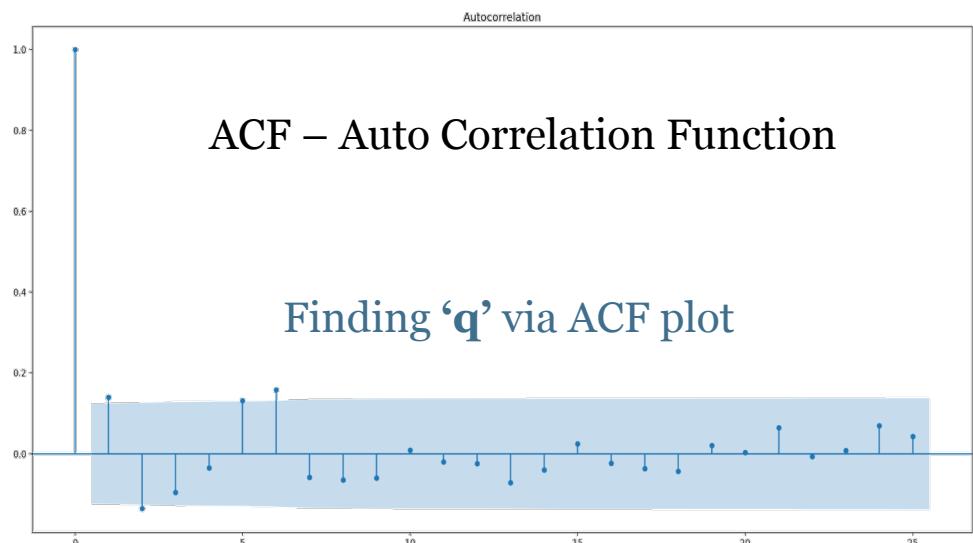
▷ $MA(q)$ – Moving Average

$$AR(1): Y_t = \alpha + \beta_1 Y_{t-1} + \varepsilon_t$$

$$MA(1): Y_t = \mu + \beta_1 \varepsilon_{t-1} + \varepsilon_t$$



Xác định q?



4. Phân tích dữ liệu chuỗi thời gian

»»» Một số mô hình phổ biến

▷ ARMA – Auto Regression Moving Average

AR(p) (*Auto-Regression*): sử dụng giá trị quá khứ (past values)

$$AR(1): Y_t = \alpha + \beta_1 Y_{t-1} + \varepsilon_t$$

MA(q) (*Moving Average*): sử dụng sai số (nhiều) quá khứ (past errors)

$$MA(1): Y_t = \mu + \beta_1 \varepsilon_{t-1} + \varepsilon_t$$

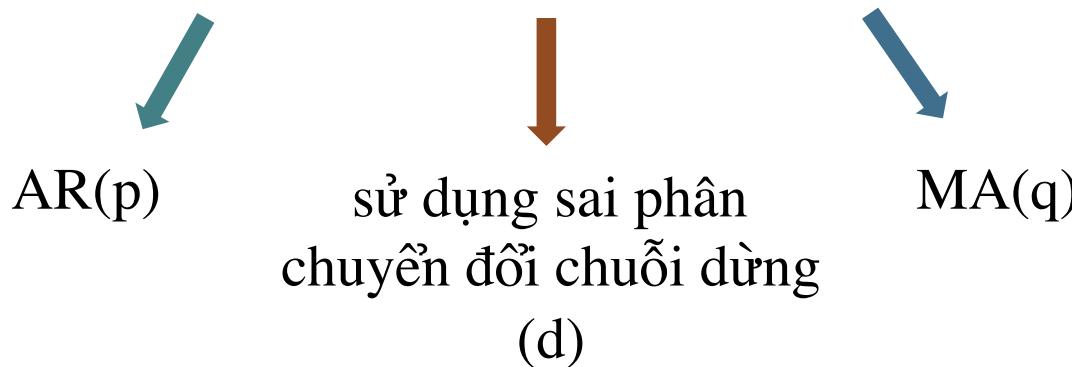
ARMA(p,q) model:

$$ARMA(1,1): Y_t = \alpha + \beta Y_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t$$

4. Phân tích dữ liệu chuỗi thời gian

»»» Một số mô hình phổ biến

▷ ARIMA – *Auto Regression Integrated Moving Average*



→ ARIMA(p, d, q)

4. Phân tích dữ liệu chuỗi thời gian

»» Quy trình phân tích dữ liệu chuỗi thời gian

Xác định
vấn đề?

1

Lựa chọn hướng
tiếp cận (kỹ thuật)?

3

2

Phân tích sơ
bộ dữ liệu?

4

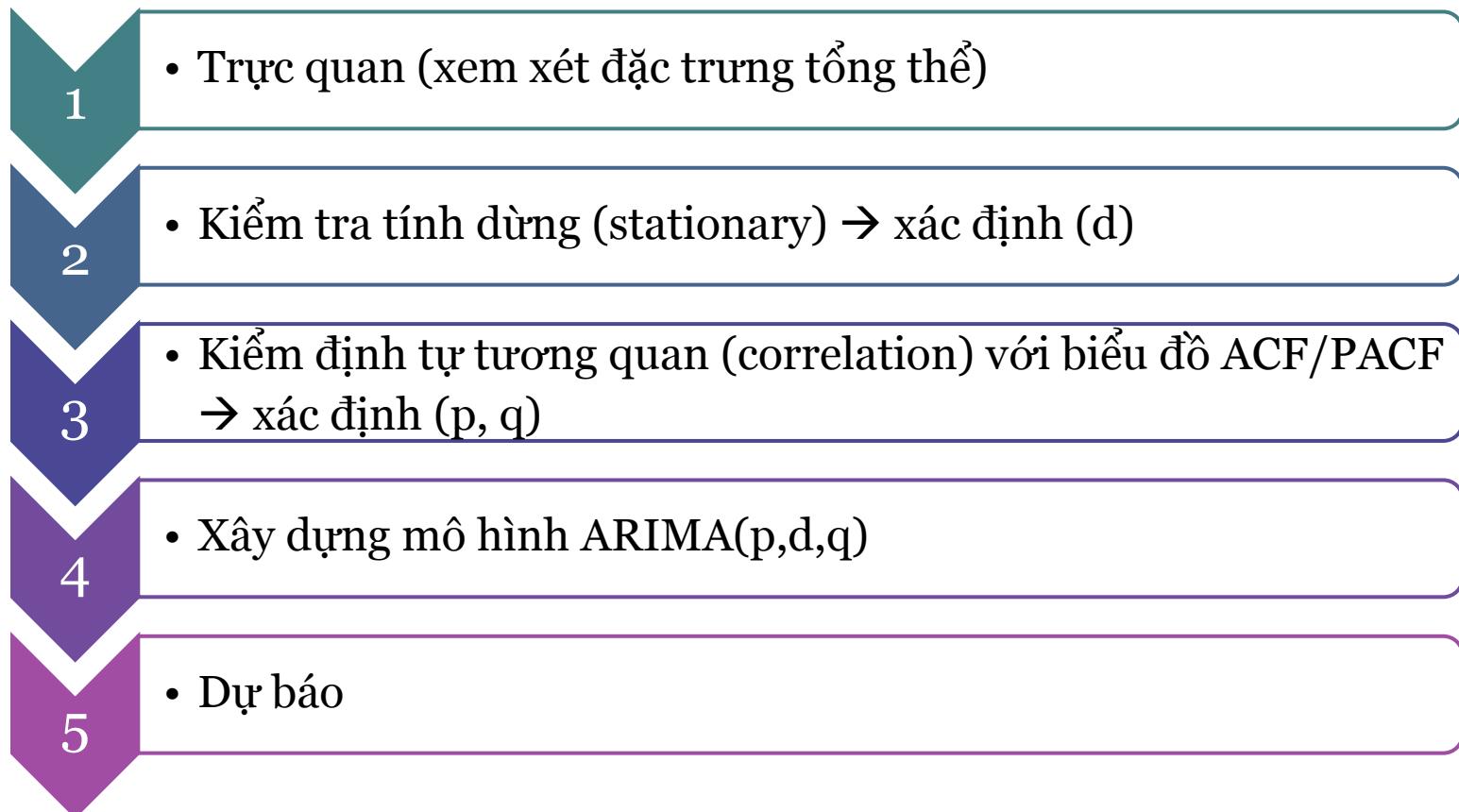
Dự báo



4. Phân tích dữ liệu chuỗi thời gian

»» Quy trình phân tích dữ liệu chuỗi thời gian

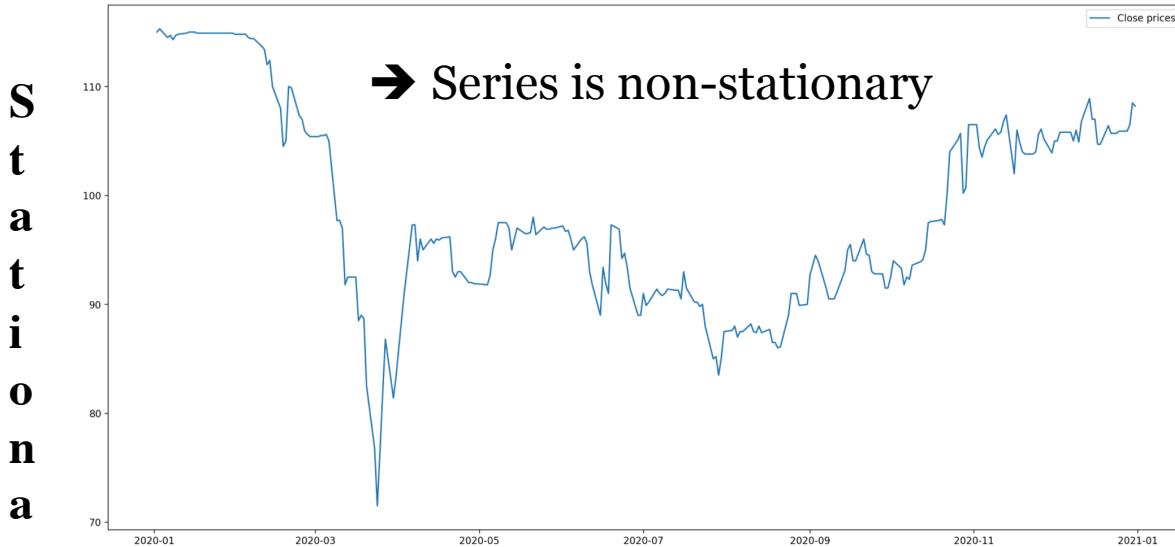
▷ Quy trình phân tích dữ liệu với mô hình ARIMA



4. Phân tích dữ liệu chuỗi thời gian

»» Quy trình phân tích dữ liệu chuỗi thời gian

▷ Quy trình phân tích dữ liệu với mô hình ARIMA



ADF: Test statistic	-2.294636
p value	0.173704
# of Lags	6.000000
# of Observations	245.000000
Critical Value (1%)	-3.457326
Critical Value (5%)	-2.873410
Critical Value (10%)	-2.573096
KPSS: Test statistic	0.312411
p value	0.100000
# of Lags	16.000000
Critical Value (10%)	0.347000
Critical Value (5%)	0.463000
Critical Value (2.5%)	0.574000
Critical Value (1%)	0.739000
dtype:	float64

erry Kiểm định ADF (*Augmented Dickey-Fuller Test*) với giả thuyết không (Ho) (null hypothesis) là chuỗi không dừng.

Kiểm định KPSS (*Kwiatkowski-Phillips-Schmidt-Shin Test*) với giả thuyết không (Ho) là chuỗi dừng.

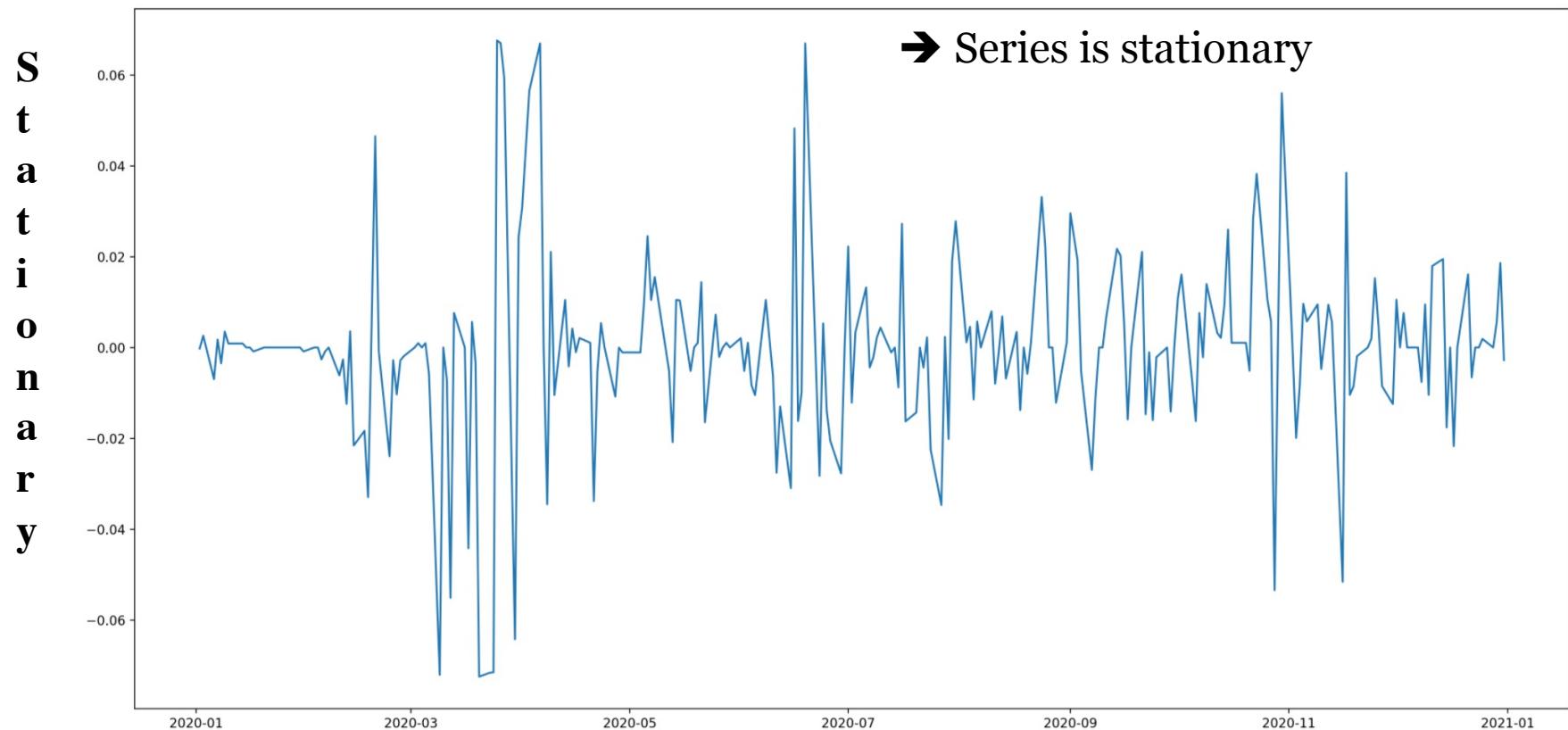
Tại sao cần xem xét tính dừng của dữ liệu?

SV tìm đọc thêm giáo trình Kinh tế lượng về hai phương pháp kiểm định ADF, KPSS.

4. Phân tích dữ liệu chuỗi thời gian

»» Quy trình phân tích dữ liệu chuỗi thời gian

▷ Quy trình phân tích dữ liệu với mô hình ARIMA

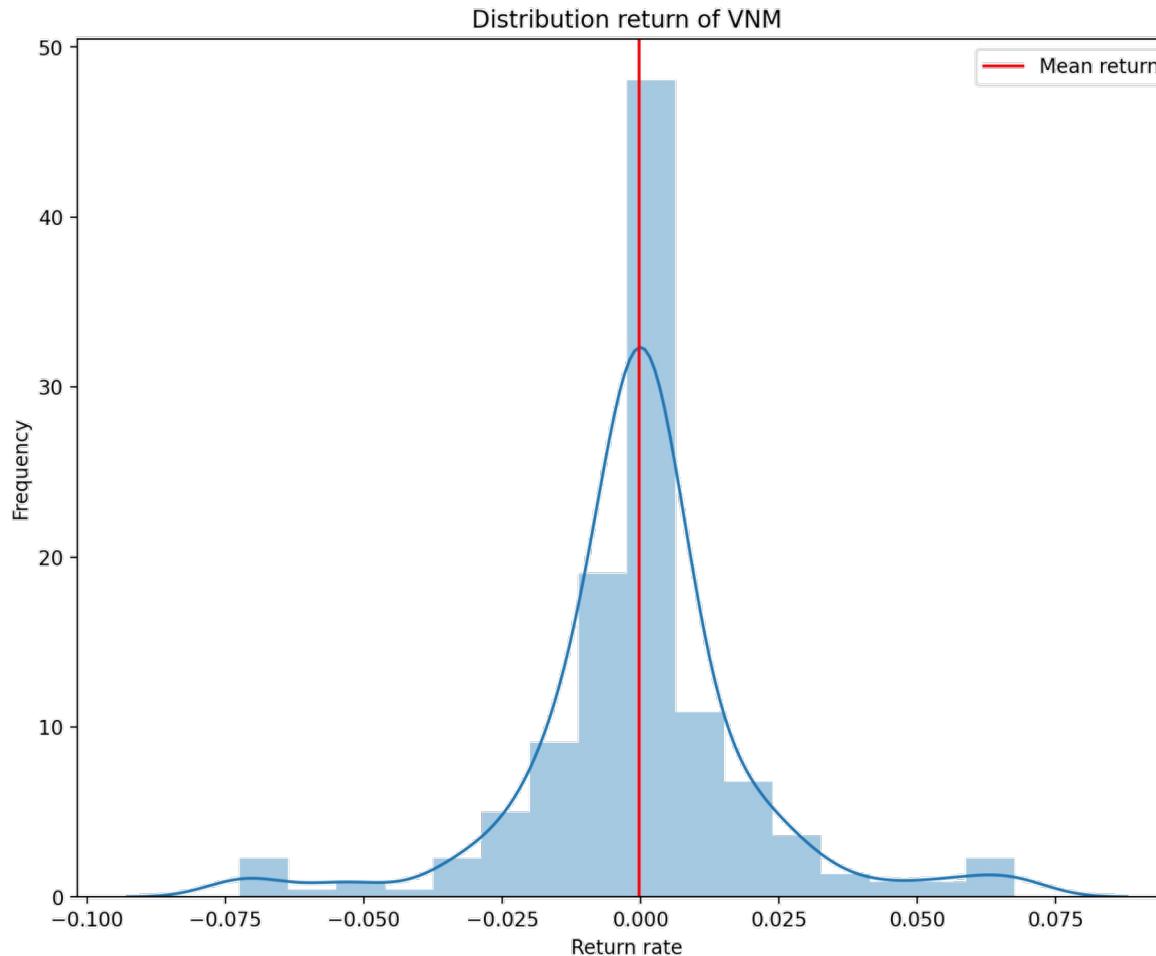


4. Phân tích dữ liệu chuỗi thời gian

»» Quy trình phân tích dữ liệu chuỗi thời gian

▷ Quy trình phân tích dữ liệu với mô hình ARIMA

Stationary



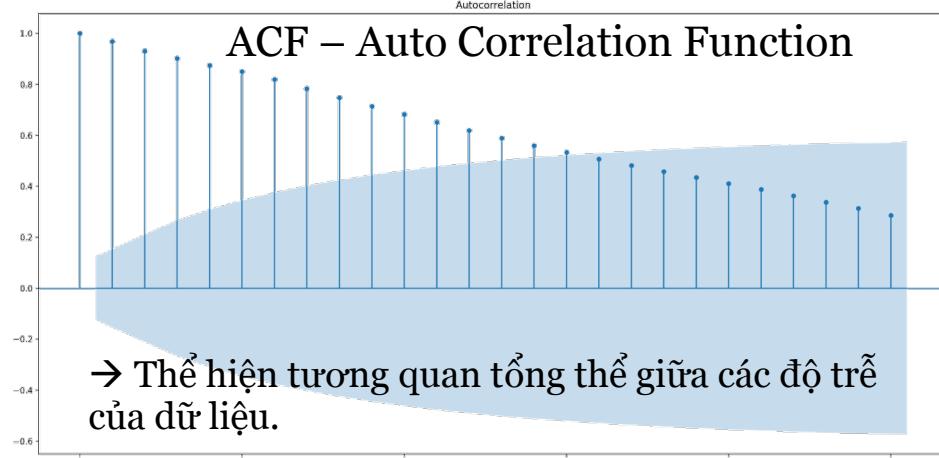
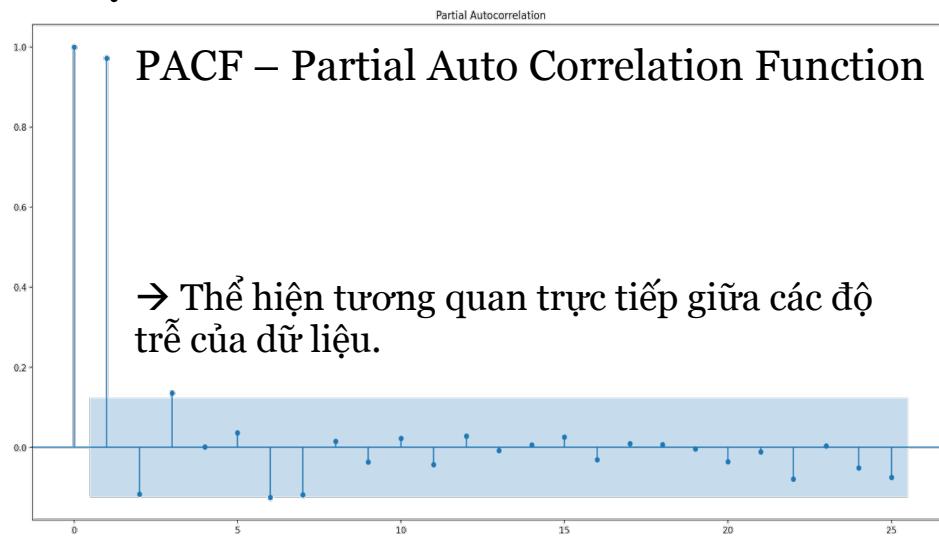
4. Phân tích dữ liệu chuỗi thời gian

»» Quy trình phân tích dữ liệu chuỗi thời gian

▷ Quy trình phân tích dữ liệu với mô hình ARIMA

C
o
r
e
l
a
t
i
o
n

Tự tương quan là hiện tượng các *giá trị trong chuỗi thời gian có quan hệ tương quan lẫn nhau*.

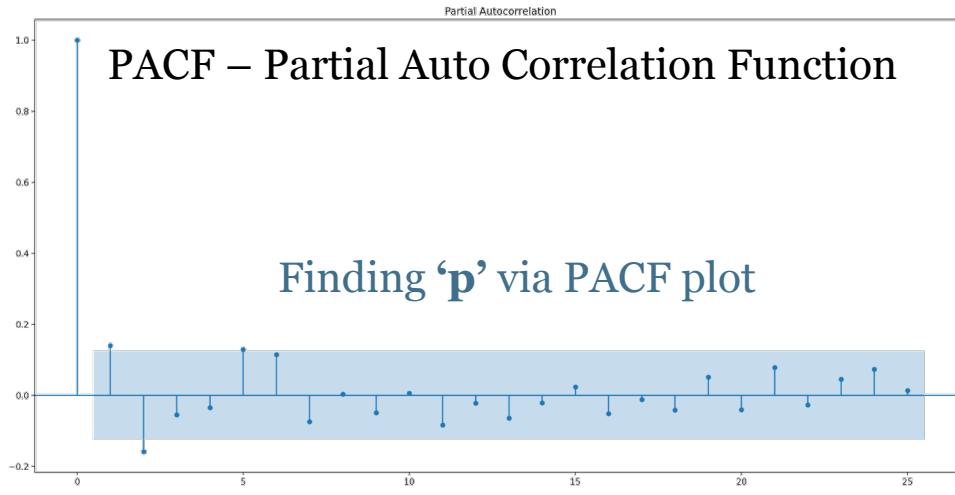


4. Phân tích dữ liệu chuỗi thời gian

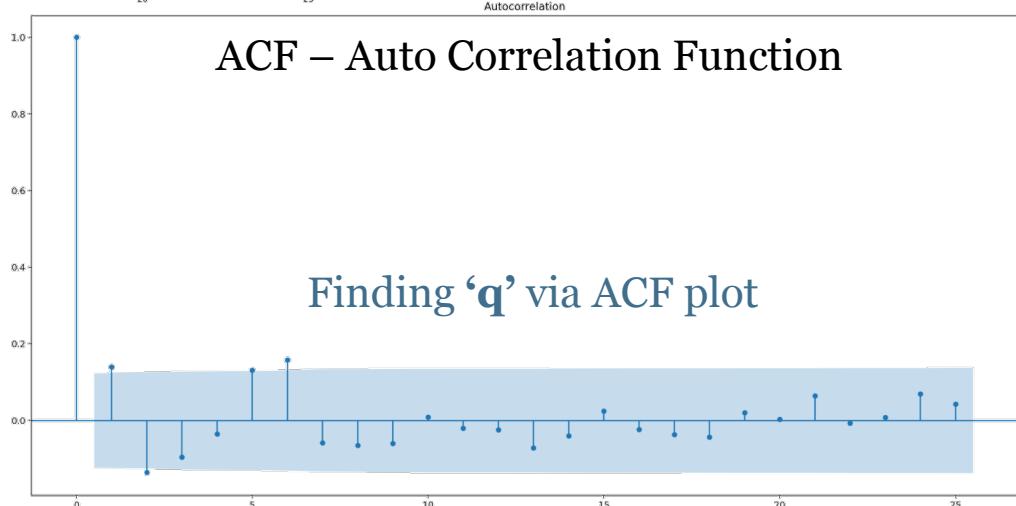
»» Quy trình phân tích dữ liệu chuỗi thời gian

▷ Quy trình phân tích dữ liệu với mô hình ARIMA

C
o
r
r
e
l
a
t
i
o
n



ARIMA(p, d, q)



ARIMA(p, d, q)

4. Phân tích dữ liệu chuỗi thời gian

»» Quy trình phân tích dữ liệu chuỗi thời gian

▷ Quy trình phân tích dữ liệu với mô hình ARIMA

Xác định tham số tự động cho mô hình:

Với Python, thư viện hỗ trợ xác định bộ tham số (p, d, q) tối ưu dựa trên tiêu chí AIC (Akaike Information Criterion) → giảm nguy cơ “overfitting” và “underfitting” của mô hình.

```
Performing stepwise search to minimize aic
ARIMA(2,1,2)(0,0,0)[0] intercept : AIC=-1345.946, Time=0.26 sec
ARIMA(0,1,0)(0,0,0)[0] intercept : AIC=-1182.808, Time=0.03 sec
ARIMA(1,1,0)(0,0,0)[0] intercept : AIC=-1256.807, Time=0.07 sec
ARIMA(0,1,1)(0,0,0)[0] intercept : AIC=-1343.980, Time=0.05 sec
ARIMA(0,1,0)(0,0,0)[0] intercept : AIC=-1184.808, Time=0.02 sec
ARIMA(1,1,2)(0,0,0)[0] intercept : AIC=-1350.810, Time=0.15 sec
ARIMA(0,1,2)(0,0,0)[0] intercept : AIC=-1352.711, Time=0.18 sec
ARIMA(0,1,3)(0,0,0)[0] intercept : AIC=-1350.045, Time=0.19 sec
ARIMA(1,1,1)(0,0,0)[0] intercept : AIC=-1351.023, Time=0.16 sec
ARIMA(1,1,3)(0,0,0)[0] intercept : AIC=-1349.182, Time=0.36 sec
ARIMA(0,1,2)(0,0,0)[0] intercept : AIC=-1358.041, Time=0.16 sec
ARIMA(0,1,1)(0,0,0)[0] intercept : AIC=-1359.395, Time=0.08 sec
ARIMA(1,1,1)(0,0,0)[0] intercept : AIC=-1358.009, Time=0.14 sec
ARIMA(1,1,0)(0,0,0)[0] intercept : AIC=-1258.805, Time=0.02 sec
ARIMA(1,1,2)(0,0,0)[0] intercept : AIC=-1355.827, Time=0.10 sec

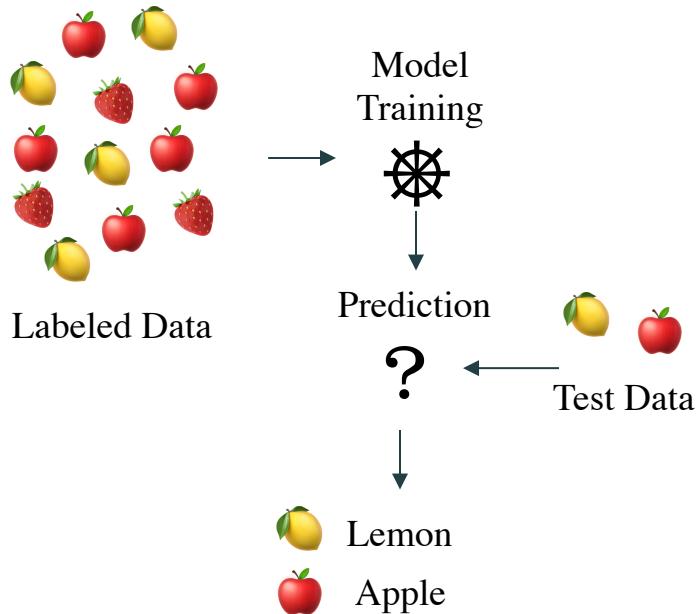
Best model: ARIMA(0,1,1)(0,0,0)[0]
Total fit time: 1.965 seconds
```

Mô hình có điểm AIC thấp hơn được xem là tốt hơn

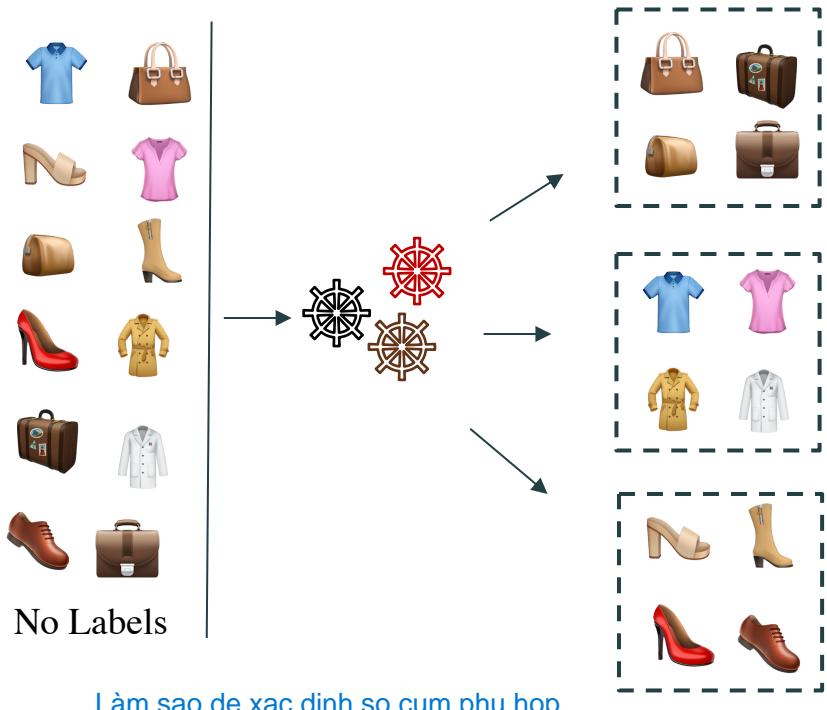
5. Machine Learning trong phân tích dữ liệu

Machine Learning Model

Supervised Learning
(Học có giám sát)

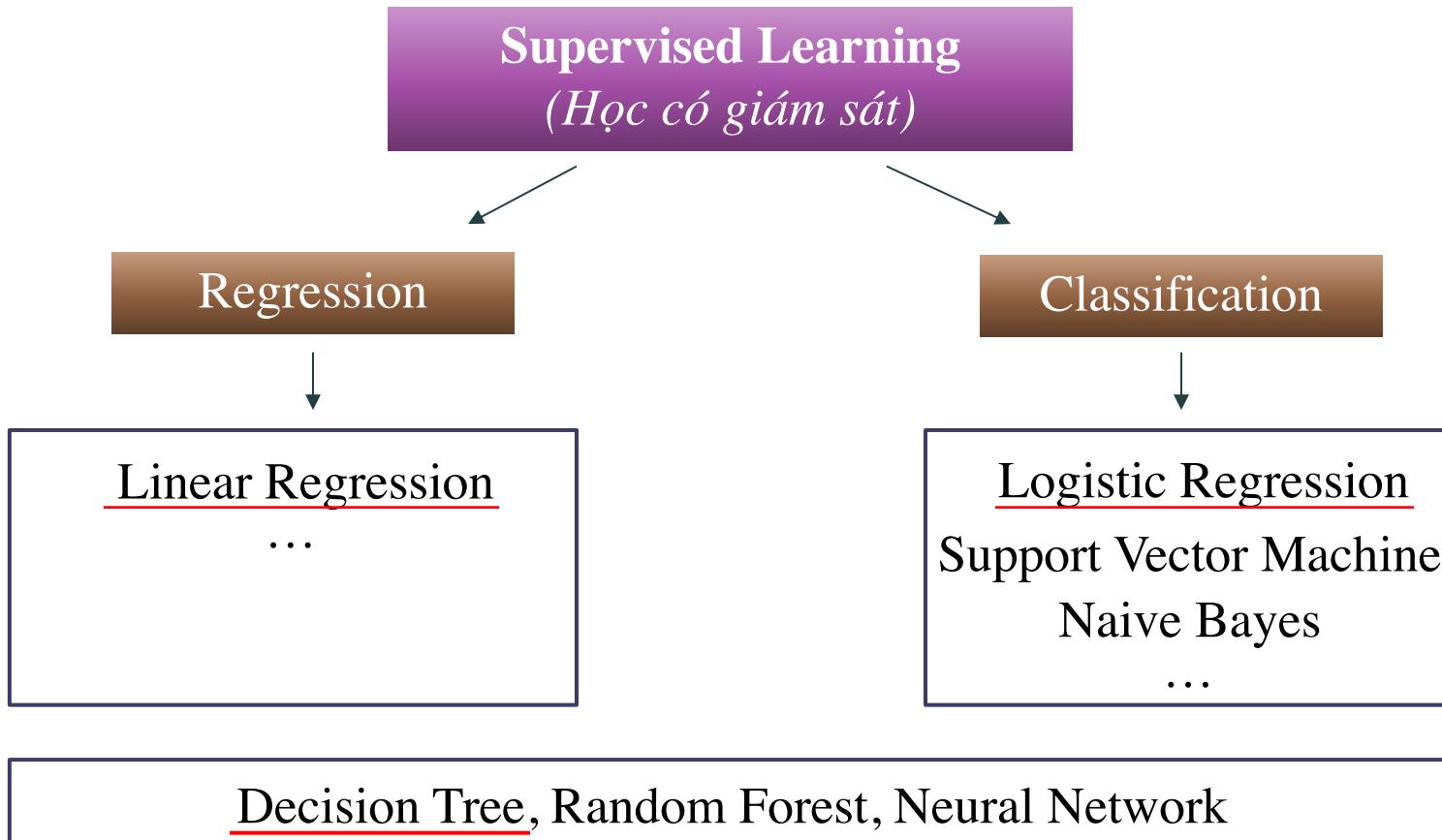


Unsupervised Learning
(Học không giám sát)



5. Machine Learning trong phân tích dữ liệu

»»» Supervised Learning

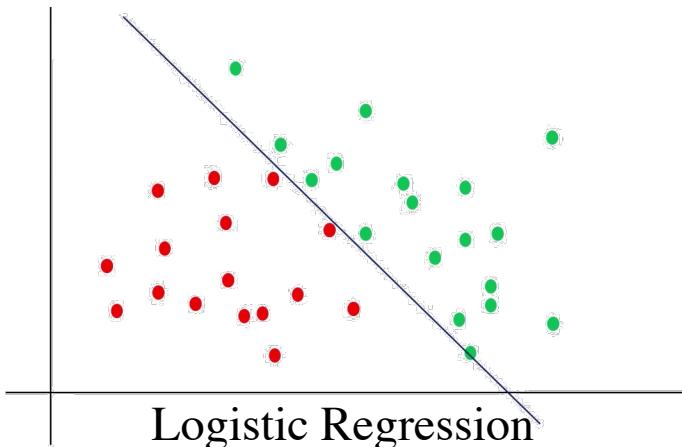


: SV tìm đọc thêm tài liệu về các thuật toán học có giám sát ứng dụng thực tiễn.

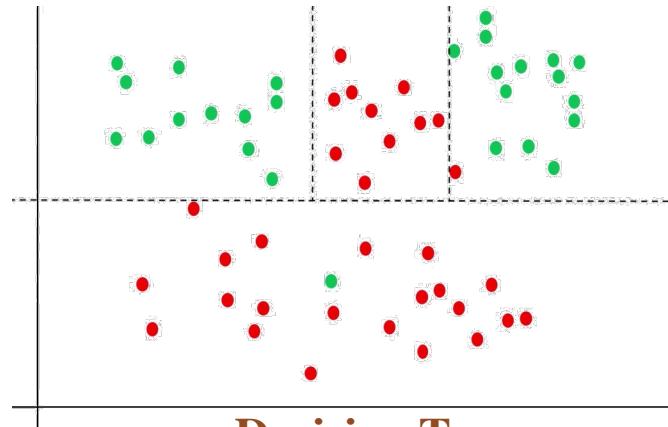
5. Machine Learning trong phân tích dữ liệu

»»» Supervised Learning

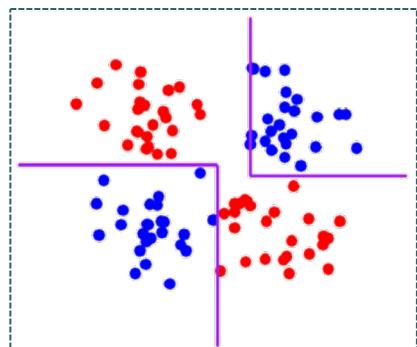
▷ Cây quyết định (Decision Tree)



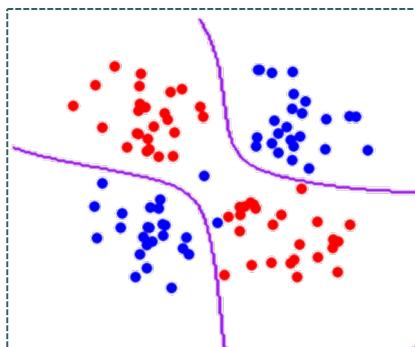
Logistic Regression



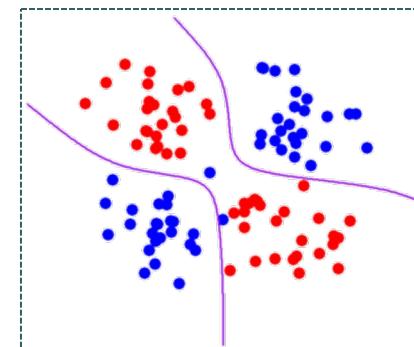
Decision Tree



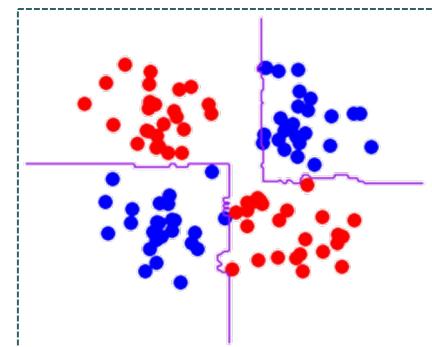
Decision Tree



SVM



Neural Network



Random Forest

La chn information gain cao nh t làm nút gc

5. Machine Learning trong phân tích dữ liệu

»»» Supervised Learning

▷ Cây quyết định (Decision Tree)

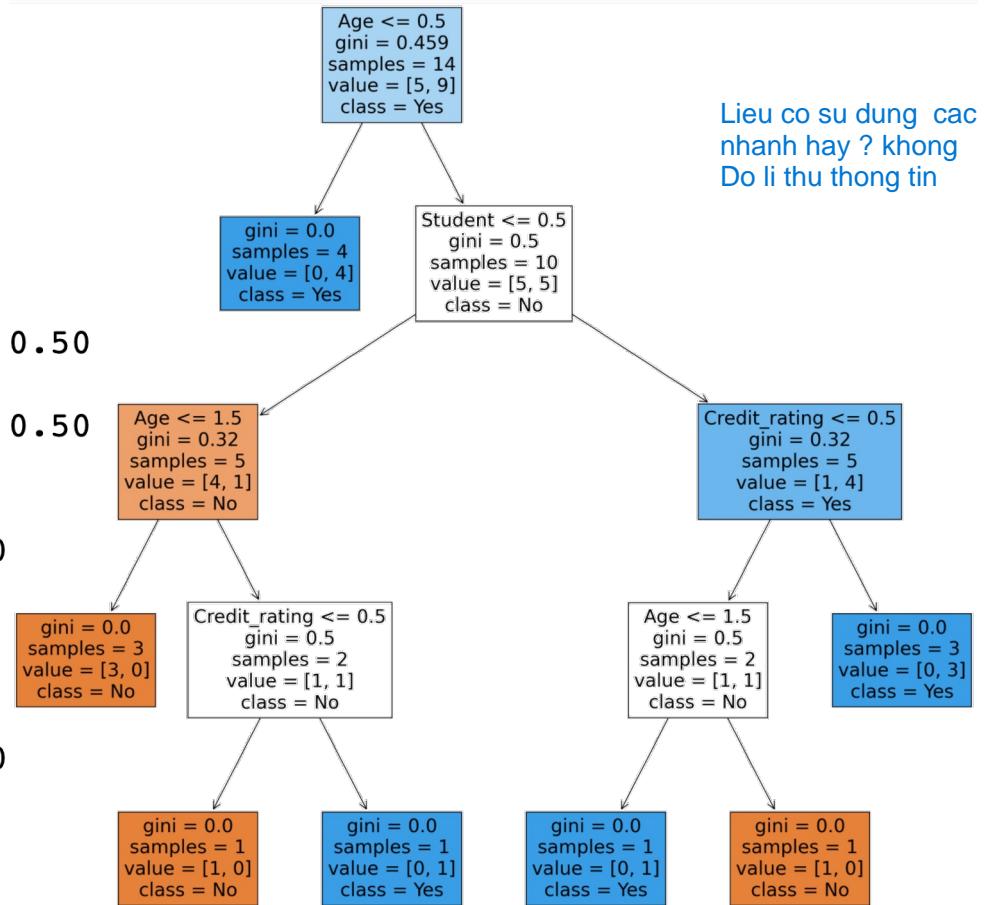
	♦ Age	♦ Income	♦ Student	♦ Credit_rating	♦ Buys_computer
0	<=30	high	no	fair	no
1	<=30	high	no	excellent	no
2	31..40	high	no	fair	yes
3	>40	medium	no	fair	yes
4	>40	low	yes	fair	yes
5	>40	low	yes	excellent	no
6	31..40	low	yes	excellent	yes
7	<=30	medium	no	fair	no
8	<=30	low	yes	fair	yes
9	>40	medium	yes	fair	yes
10	<=30	medium	yes	excellent	yes
11	31..40	medium	no	excellent	yes
12	31..40	high	yes	fair	yes
13	>40	medium	no	excellent	no

5. Machine Learning trong phân tích dữ liệu

»»» Supervised Learning

```

    --- Age <= 0.50
      |--- class: 1
    --- Age > 0.50
      |--- Student <= 0.50
        |--- Age <= 1.50
          |--- class: 0
        |--- Age > 1.50
          |--- Credit_rating <= 0.50
            |--- class: 0
            |--- Credit_rating > 0.50
              |--- class: 1
      --- Student > 0.50
        |--- Credit_rating <= 0.50
          |--- Age <= 1.50
            |--- class: 1
          |--- Age > 1.50
            |--- class: 0
        |--- Credit_rating > 0.50
          |--- class: 1
  
```



5. Machine Learning trong phân tích dữ liệu

»» Supervised Learning

▷ Cây quyết định (Decision Tree)

✓ Entropy (Information Gain)

Với:

- D : tập dữ liệu huấn luyện.
- $C_{i,D}$: tập các mẫu của D thuộc lớp C_i với $i = \{1, 2, \dots, m\}$.
- $|C_{i,D}|, |D|$: số lượng mẫu của tập $C_{i,D}$ và D tương ứng.
- p_i là xác suất để một mẫu bất kỳ của D thuộc về lớp C_i .

Thông tin kỳ vọng để phân lớp một mẫu trong D :

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i) \quad p_i = \frac{|C_{i,D}|}{|D|}$$

5. Machine Learning trong phân tích dữ liệu

»»» Supervised Learning

- ▷ Cây quyết định (Decision Tree)
- ✓ Entropy (Information Gain)

Với 14 mẫu tin trong ví dụ có 9 mẫu là “mua máy tính”:

$$|D| = 14; m = 2; C_1 = \text{“mua”}; C_2 = \text{“không mua”}$$

$$|C_{1,D}| = 9; |C_{2,D}| = 5$$

Thông tin kỳ vọng để phân lớp một mẫu trong D là:

$$Info(D) = I(9,5) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940$$

5. Machine Learning trong phân tích dữ liệu

»»» Supervised Learning

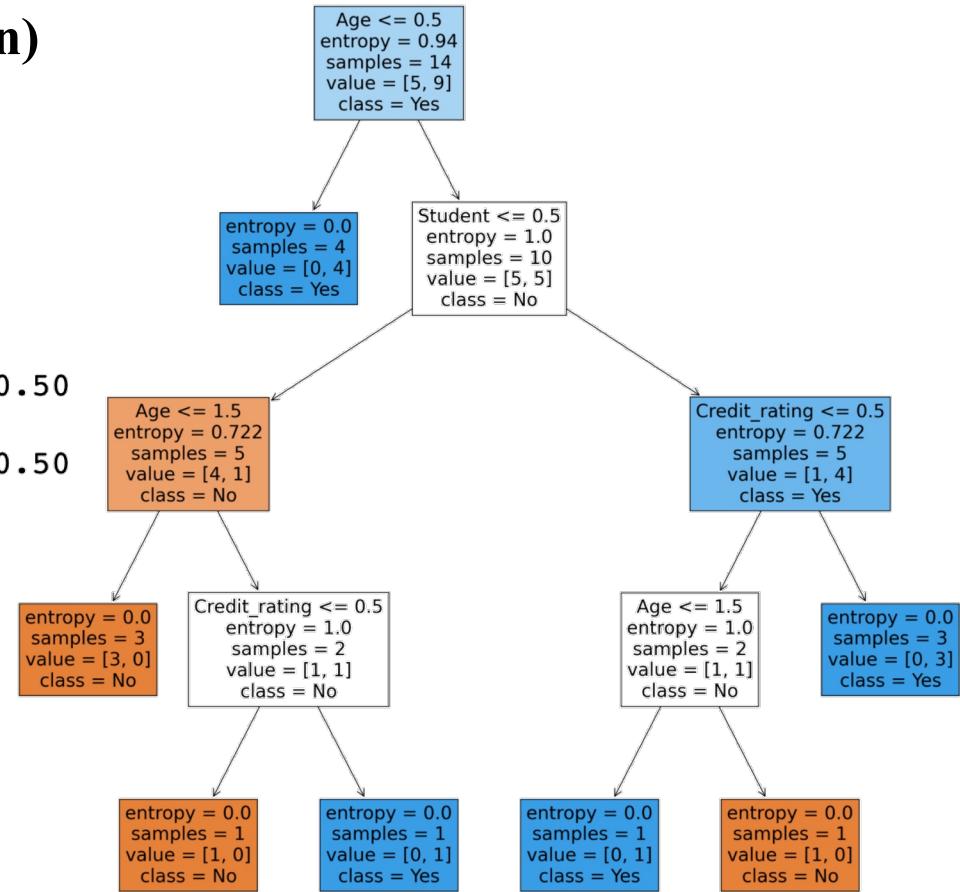
▷ Cây quyết định (Decision Tree)

✓ Entropy (Information Gain)

```

--- Age <= 0.50
|   --- class: 1
--- Age > 0.50
|   --- Student <= 0.50
|       --- Age <= 1.50
|           --- class: 0
|       --- Age > 1.50
|           --- Credit_rating <= 0.50
|               --- class: 0
|               --- Credit_rating > 0.50
|                   --- class: 1
|   --- Student > 0.50
|       --- Credit_rating <= 0.50
|           --- Age <= 1.50
|               --- class: 1
|               --- Age > 1.50
|                   --- class: 0
|       --- Credit_rating > 0.50
|           --- class: 1

```



5. Machine Learning trong phân tích dữ liệu

»» Supervised Learning

▷ Cây quyết định (Decision Tree)

✓ Gini Index

Với tập huấn luyện D chứa các mẫu của m lớp.

Ta có chỉ mục Gini của tập D – $gini(D)$ là:

$$gini(D) = 1 - \sum_{i=1}^m p_i^2$$

p_i là tần suất của lớp C_i trong D

Với tập DL của ví dụ trên, ta có $gini(D)$ là:

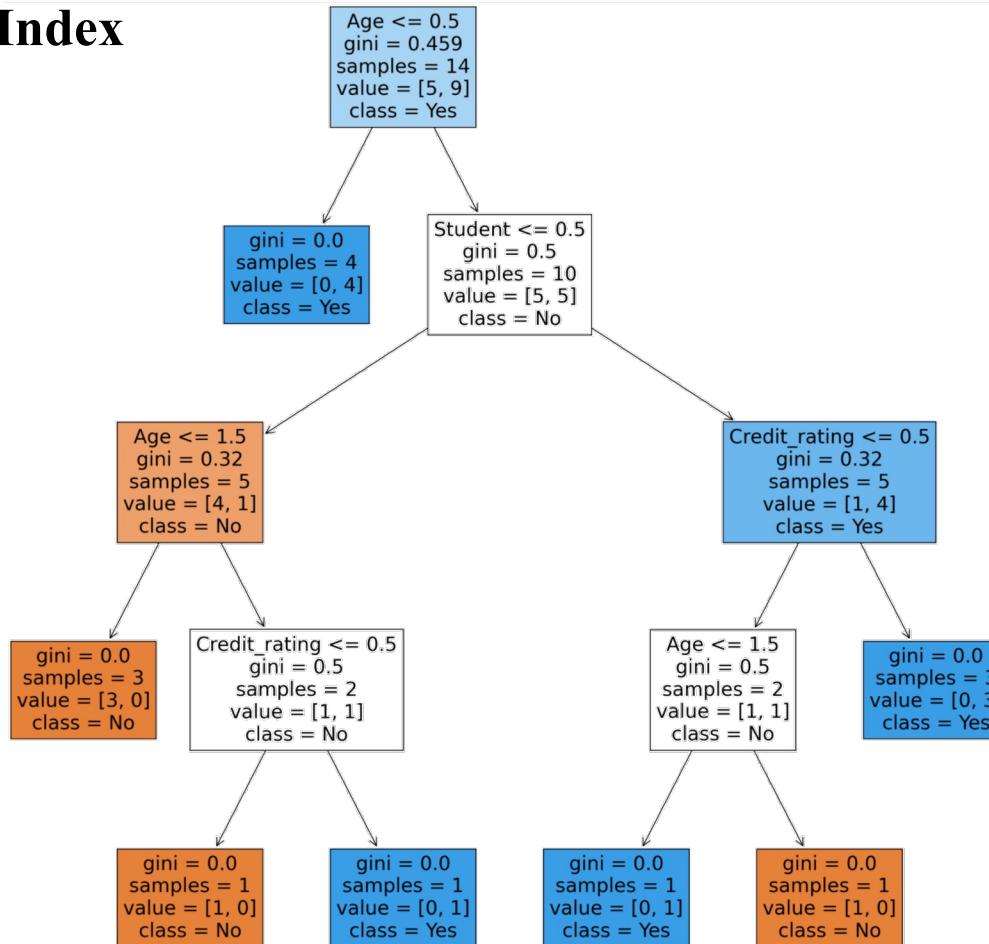
$$gini(D) = 1 - \left(\frac{9}{14} \right)^2 - \left(\frac{5}{14} \right)^2 = 0.459$$

5. Machine Learning trong phân tích dữ liệu

»»» Supervised Learning

▷ Cây quyết định (Decision Tree)

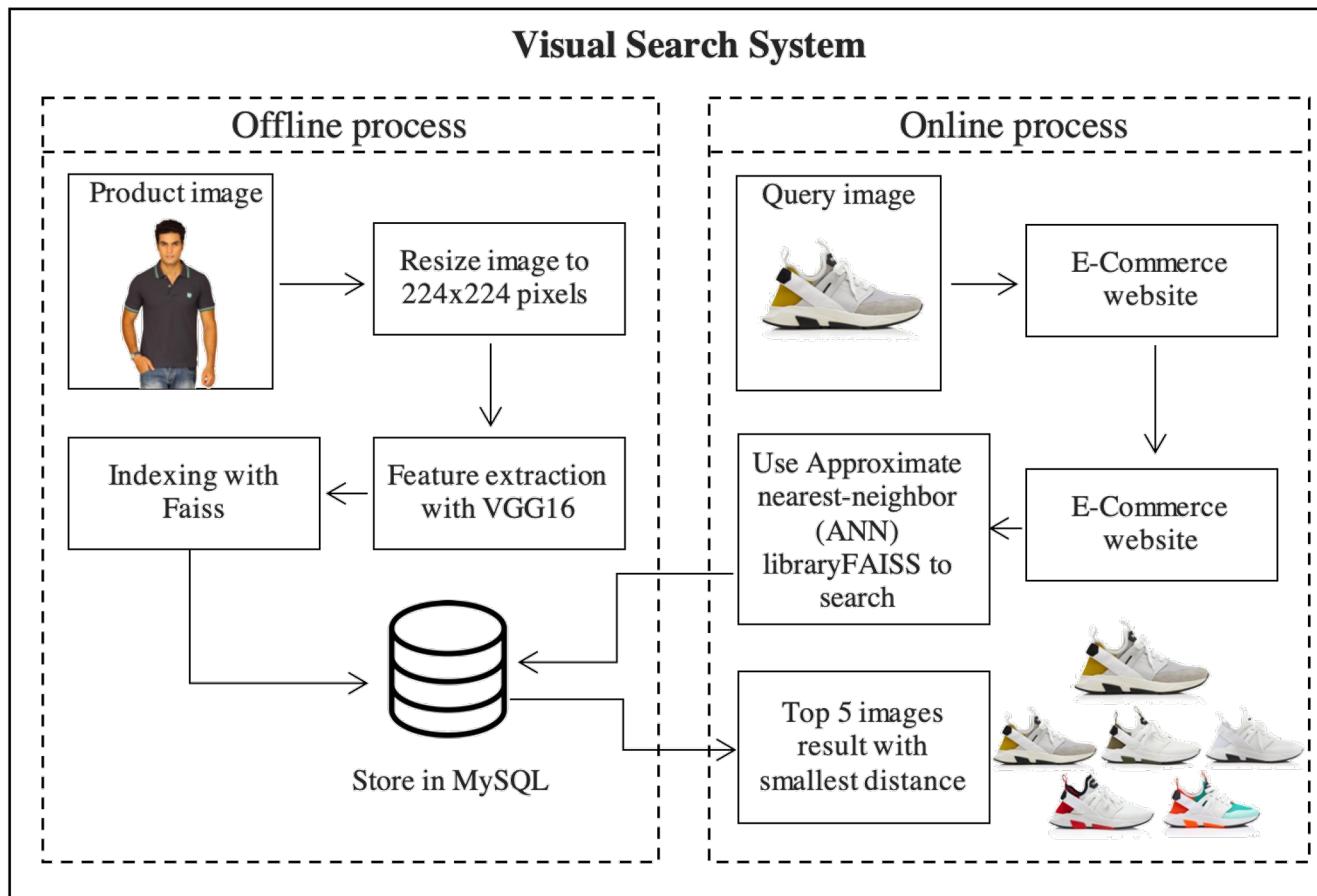
✓ Gini Index



5. Machine Learning trong phân tích dữ liệu

» Supervised Learning

▷ Một số ứng dụng thực tế

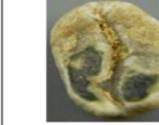


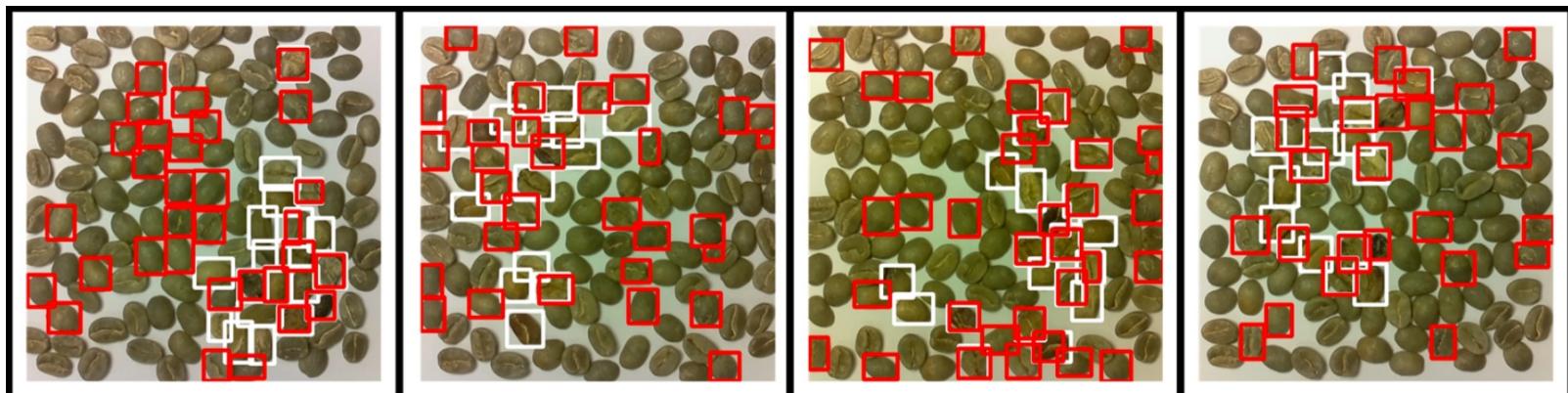
5. Machine Learning trong phân tích dữ liệu

» Supervised Learning

▷ Một số ứng dụng thực tế

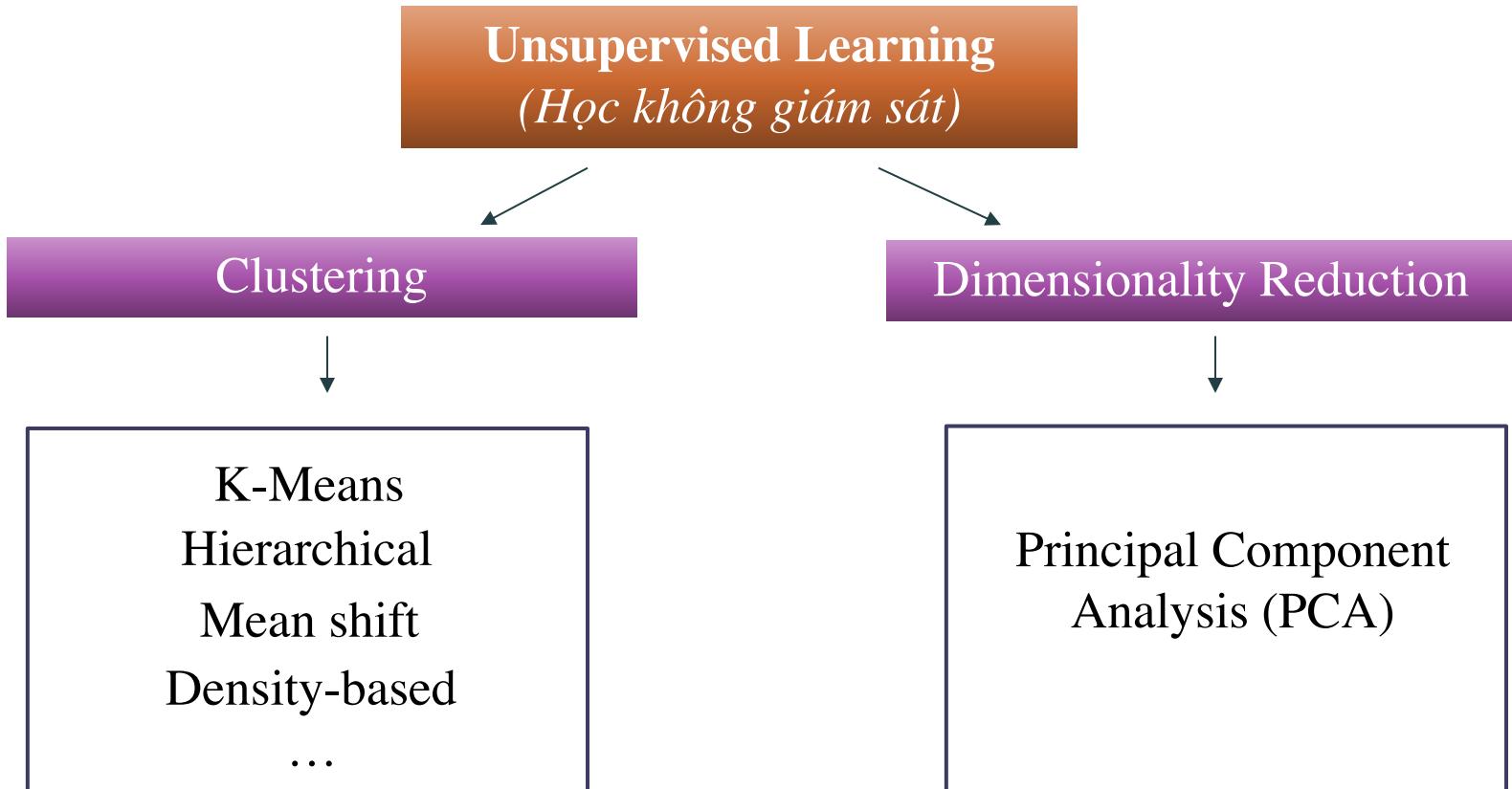
Defective Green Bean Removal

						
Normal	Insect damage	Water damage	Unhulled	Shell	Moldy	Floater
						
Immature	Sour	Black	Faded	Broken	Ferment	Dead



5. Machine Learning trong phân tích dữ liệu

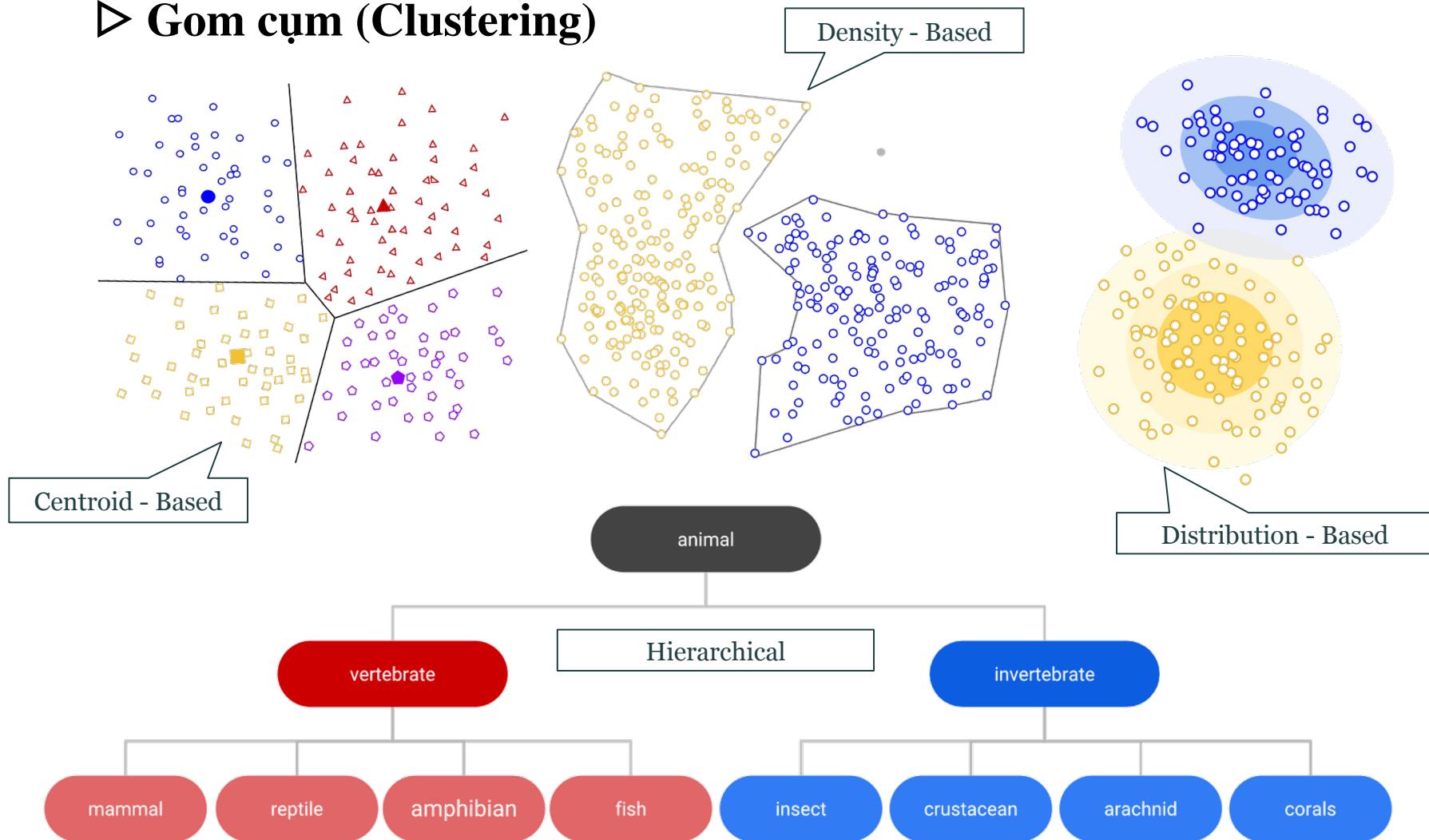
»»» Unsupervised Learning



5. Machine Learning trong phân tích dữ liệu

»»» Unsupervised Learning

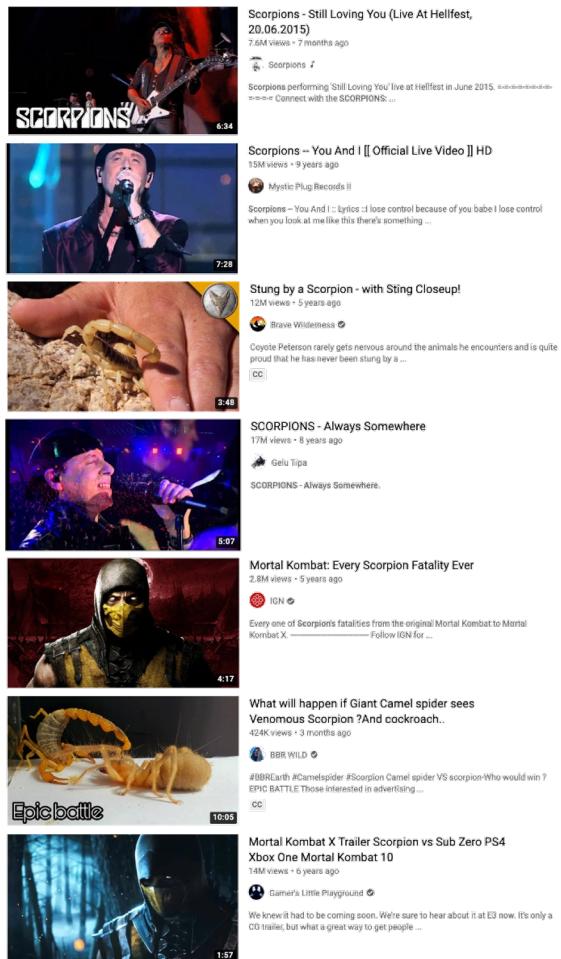
▷ Gom cụm (Clustering)



5. Machine Learning trong phân tích dữ liệu

»»» Unsupervised Learning

▷ Vd: gom cụm video theo chủ đề



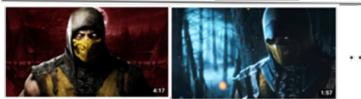
Cluster #1



Cluster #2



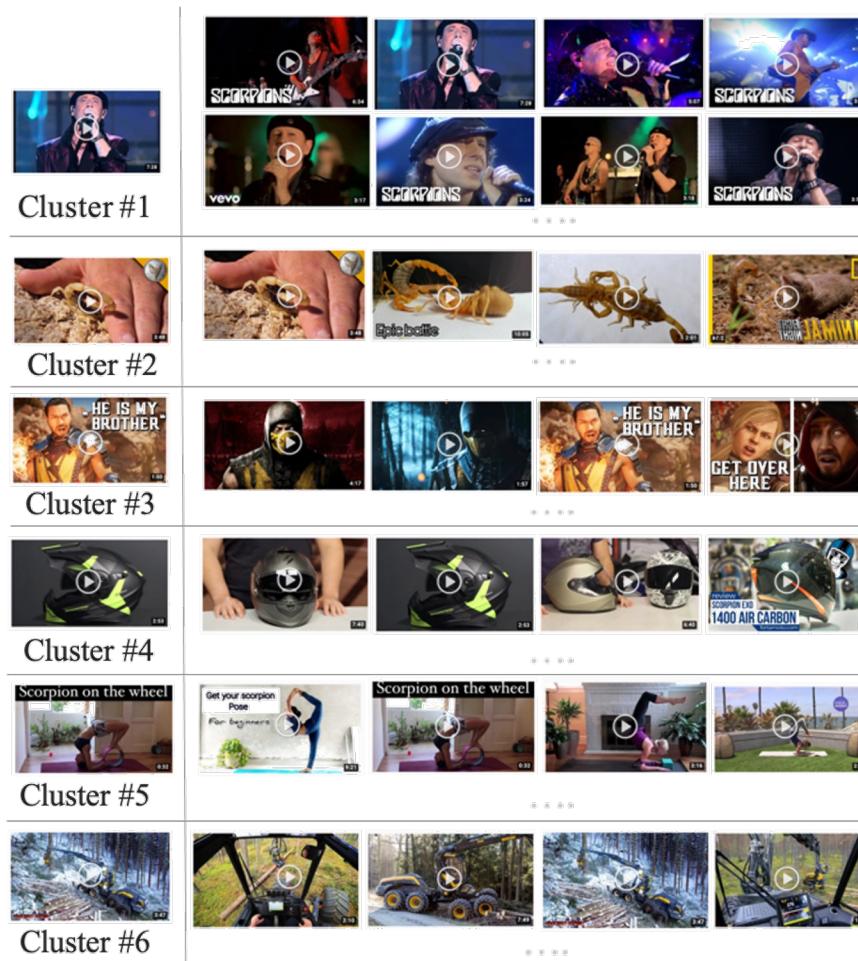
Cluster #3



5. Machine Learning trong phân tích dữ liệu

»»» Unsupervised Learning

▷ Vd: gom cụm video theo chủ đề



Q

A