

LOB2203

**ระบบเทคโนโลยีสารสนเทศด้านโลจิสติกส์
สำหรับธุรกิจออนไลน์**

CH 7



rapidminer



สาขาการจัดการโลจิสติกส์สำหรับธุรกิจออนไลน์
วิทยาลัยโลจิสติกส์และซัพพลายเชน มหาวิทยาลัยราชภัฏสุรินทร์



Chapter 7

Classification





การพยากรณ์อากาศ

Current Weather Updated on Wed Jan 22 4:45 PM Next Update in 08:54 mins

Partly cloudy

-14°C
Feels like -15

Snow, blowing snow could impact travel plans this weekend.

Precipitation Outlook
2-4 cm of snow from Wed. Evening to Thurs. Overnight.

Wind: NW 2 km/h **Humidity:** 41% **Pressure:** 101.9kPa **Visibility:** 12.9 km **Ceiling:** 18100ft **Sunrise:** 07:43 **Sunset:** 17:14

Air Quality: Moderate Risk **UV:** Low [View Webcams](#)

Time Period	Weather	Temperature	Feels like	P.O.P.	Snow	Wind	Wind gust	Humidity
Wed. Evening	Few flurries	-17°C	-22	40%	Less than 1 cm	Wind E 10 km/h	-	64%
Wed. Overnight	Scattered flurries	-18°C	-26	40%	1-3 cm	Wind NE 15 km/h	-	77%
Thurs. Morning	Scattered flurries	-18°C	-27	40%	Less than 1 cm	Wind N 15 km/h	-	84%
Thurs. Afternoon	Variable cloudiness	-15°C	-24	30%	-	Wind NW 20 km/h	35 km/h	71%

สภาพอากาศวันปัจจุบัน **สภาพอากาศวันถัดไป**



Speecg recognition





face recognition





Classification in daily life

- spam e-mail

The screenshot shows a Gmail interface with the spam folder selected. The search bar contains 'in:spam'. The main content area displays a list of spam messages with a yellow highlight over the text 'spam e-mail'. The list includes:

Sender	Subject	Date
grmse2014@grmse2014.org	GRMSE-2014, Call For Paper	Apr 18
Melesta	The GREAT GIVEAWAY of ou	Apr 17
mrs fello	How is your family? - From M	Apr 16
william philip	I AM INTERESTED IN MARRI	Apr 16
Hudson Martins	PARTNERSHIP - Dear Future f	Apr 15
Yoast	Yoast: Post Connector updat	Apr 15
Sandra	PLEASE REPLY ASAP. - Dear	Apr 15
Melesta	Kingdom's Heyday - a great ti	Apr 14
support	Yokee - Action needed - Hi Sir	Apr 14
Info	All the bling-bling you'll ever	Apr 12
nofrills	Your weekly flyer is here! - To v	Apr 11



Classification example

- ตัวอย่าง spam e-mail classification
 - ระบุว่า e-mail ใหนบ้างที่เป็น spam e-mail

ID	Text	Type
1	Please call our customer service representative on FREE PHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed £1000 cash	spam
2	You have won \$1,000 cash or a \$2,000 prize! To claim, call 09050000327	spam
3	I'm gonna be home soon and I don't want to talk about this stuff anymore tonight	normal
4	Is that seriously how you spell his name?	normal
5	Double mins and txts 4 6months FREE Bluetooth on Orange. Available on Sony, Nokia Motorola phones.	spam
6	FREE RINGTONE text FIRST to 87131 for a poly or text GET to 87131 for a true tone!	spam
7	Sorry, I'll you call later in meeting.	normal
8	Congratulations - in this week's competition draw u have won the £1450 prize to claim just call 09050002311	spam
9	Thanks a lot for your wishes on my birthday. Thanks you for making my birthday truly memorable.	normal
10	Hello, What are you doing? Did you attend the training course today?	normal



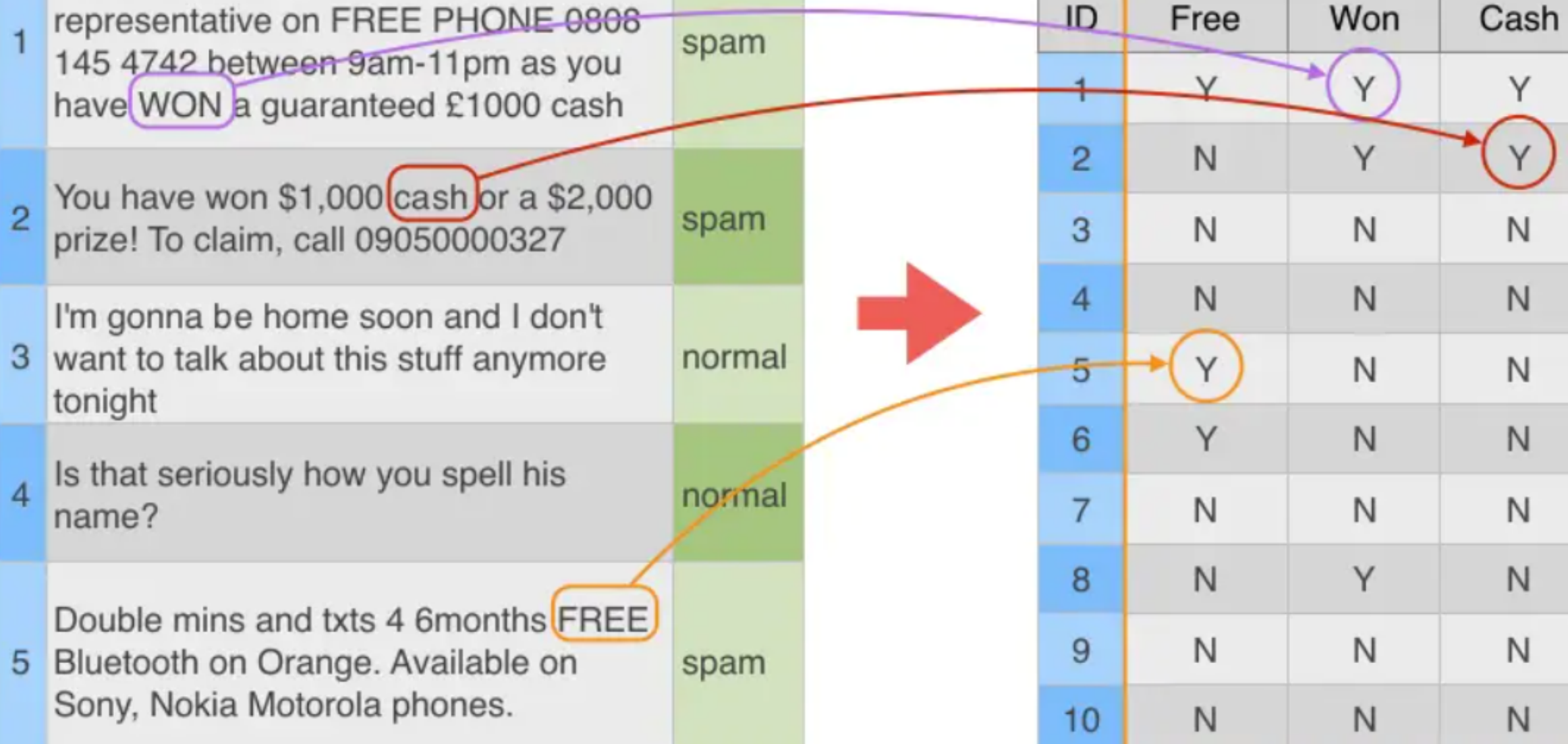
Classification example

- ตัวอย่าง spam e-mail classification
 - หา keyword ที่ใช้บ่งบอกว่า เป็น spam e-mail

ID	Text	Type
1	Please call our customer service representative on FREE PHONE 0800 145 4742 between 9am-11pm as you have WON a guaranteed £1000 cash	spam
2	You have won \$1,000 cash or a \$2,000 prize! To claim, call 09050000327	spam
3	I'm gonna be home soon and I don't want to talk about this stuff anymore tonight	normal
4	Is that seriously how you spell his name?	normal
5	Double mins and txts 4 6months FREE Bluetooth on Orange. Available on Sony, Nokia Motorola phones.	spam



keywords				
ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal



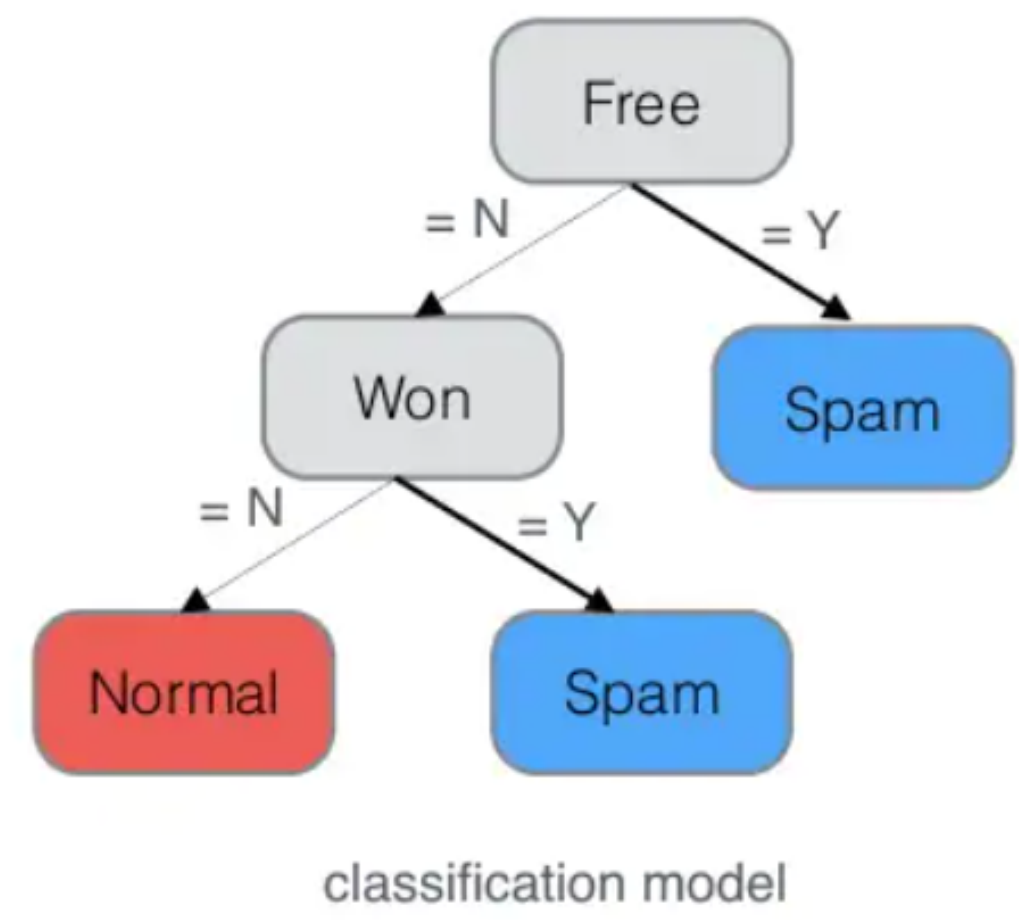


Classification example

- ตัวอย่าง spam e-mail classification
 - สร้างโมเดล (classification model) จากข้อมูล training data ซึ่งมีลาเบล (label)

← attribute → ← label →

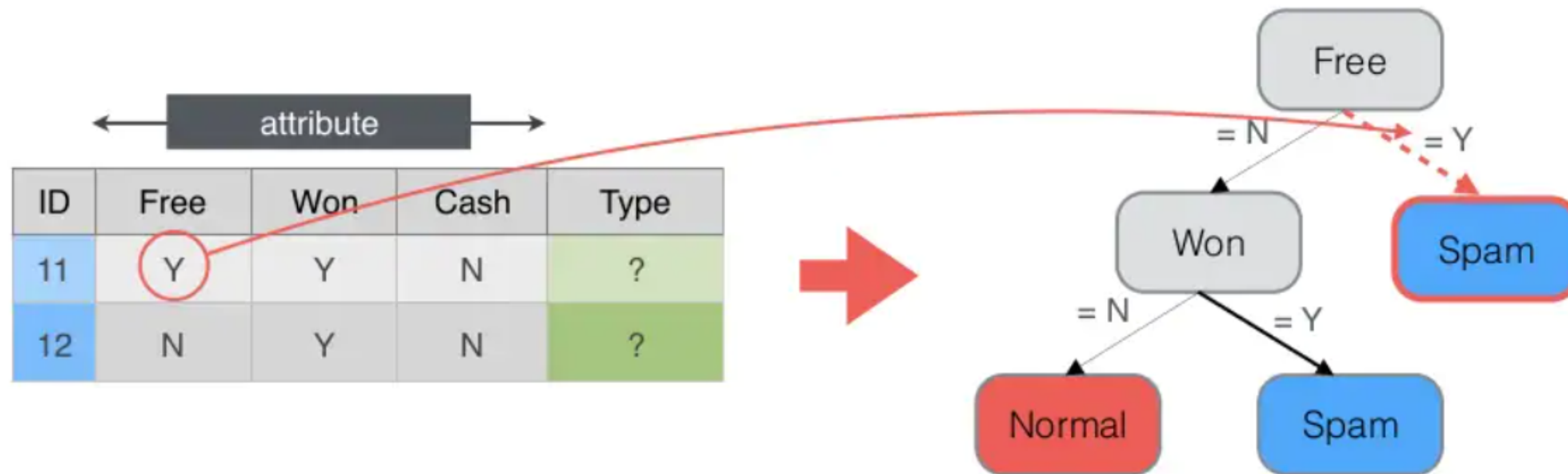
ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal





Classification example

- ตัวอย่าง spam e-mail classification
 - นำข้อมูลใหม่ (unseen data) ทำนายโดยใช้โมเดล

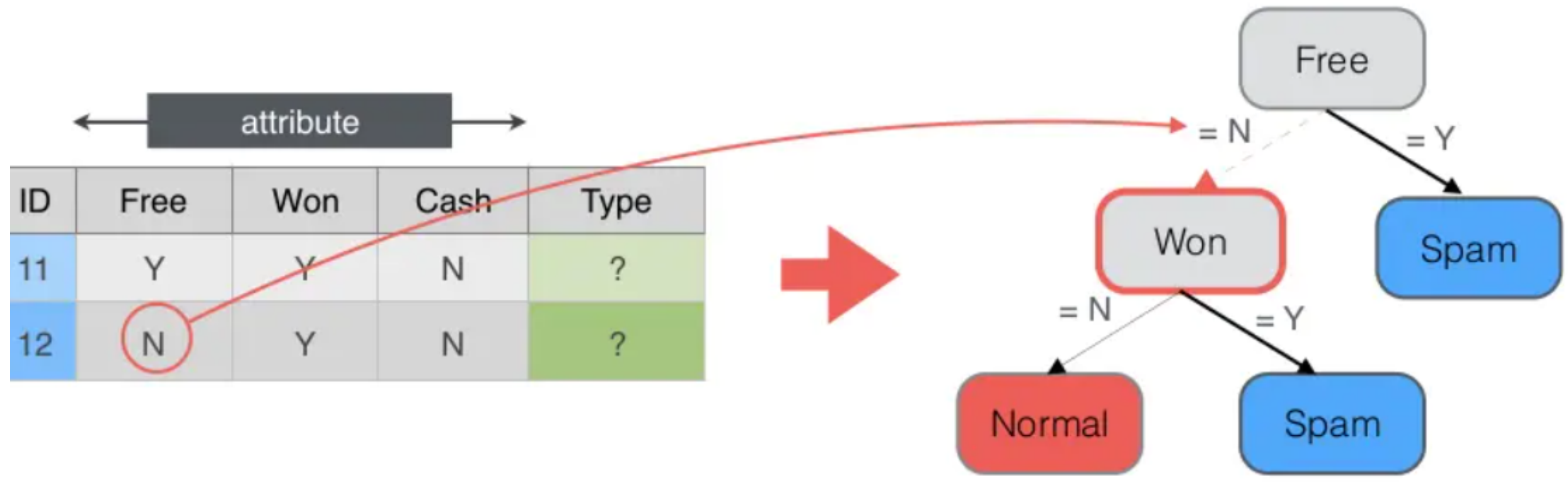




Classification example

ตัวอย่าง spam e-mail classification

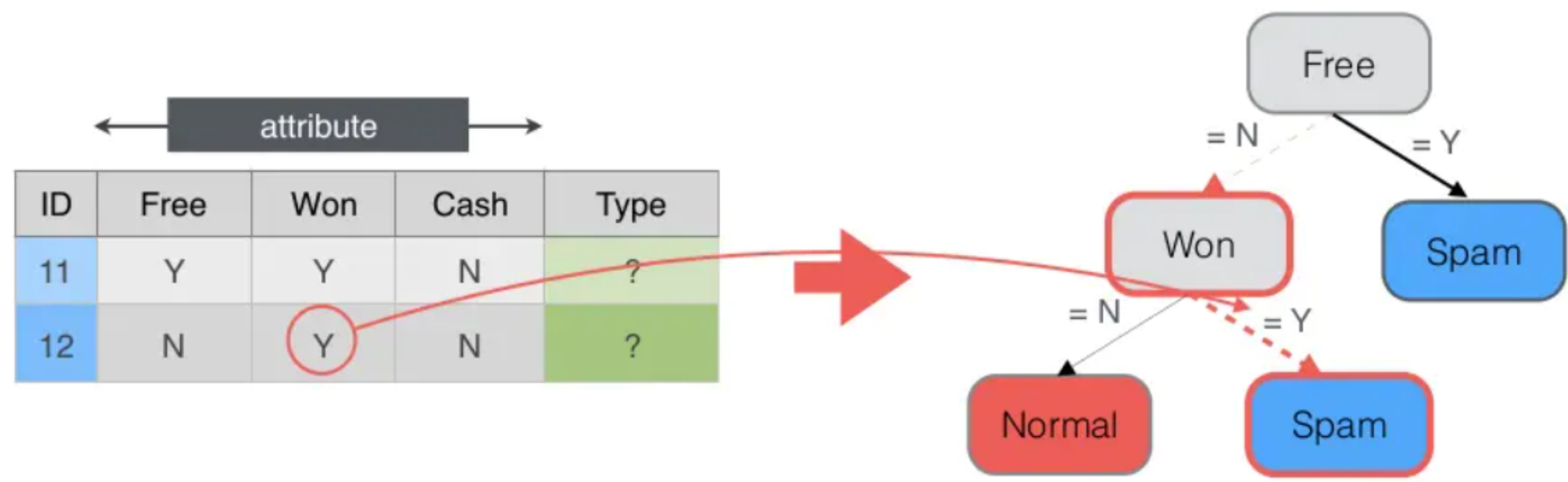
- นำข้อมูลใหม่ (unseen data) ทำนายโดยใช้โมเดล





Classification example

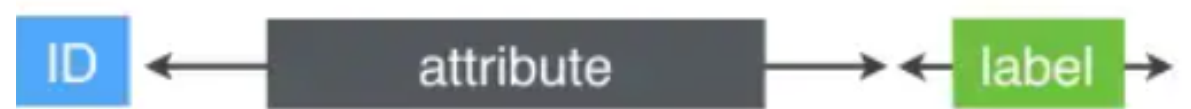
- ตัวอย่าง spam e-mail classification
 - นำข้อมูลใหม่ (unseen data) ทำนายโดยใช้โมเดล





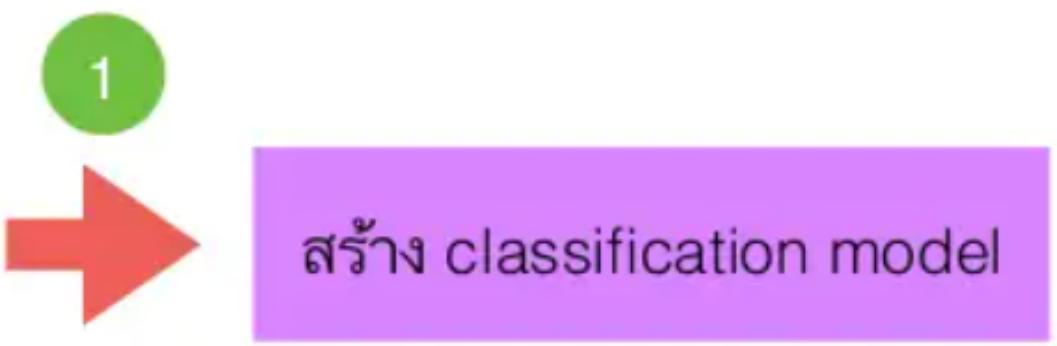
Classification example

- ตัวอย่าง spam e-mail classification



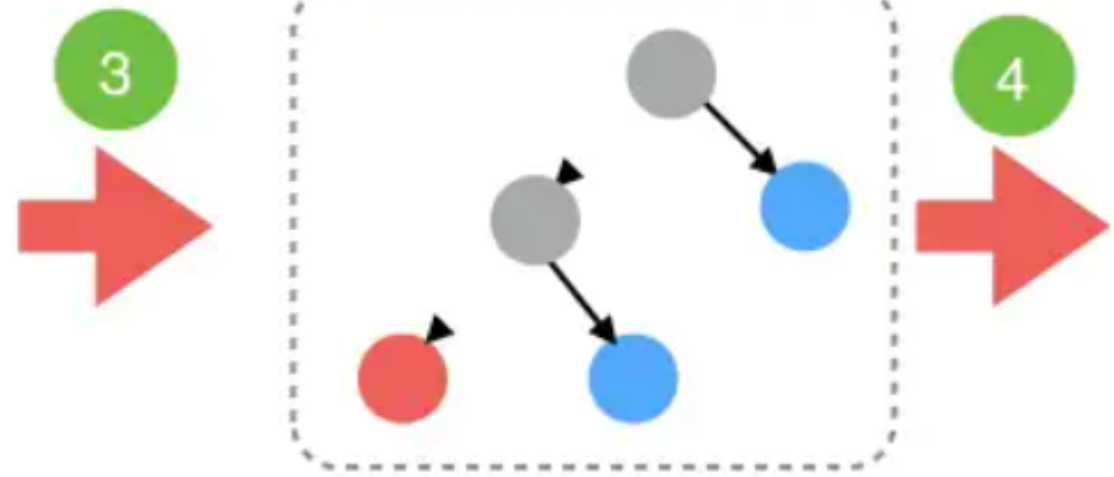
ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam

training data



ID	Free	Won	Cash	Type
11	Y	Y	N	?
12	N	Y	N	?

unseen data



classification model



Classification & Regression

- การจำแนกประเภทข้อมูล (classification)
 - นำข้อมูลเดิมที่มีคำตอบที่สนใจ หรือ คลาส (class) มาสร้างเป็นโมเดล (model) เพื่อหาคำตอบให้กับข้อมูลใหม่ (unseen data)
 - คลาสคำตอบเป็น **ประเภท** (nominal)
 - ผนตก หรือ ไม่ตก
 - spam email หรือ normal email
- การประมาณค่าข้อมูล (regression)
 - มีลักษณะเหมือนกับ classification เพียงแต่คลาสคำตอบที่สนใจเป็น **ตัวเลข** (numeric)
 - อุณหภูมิในวันถัดไป
 - ยอดขายในไตรมาสถัดไป



Classification & Regression task

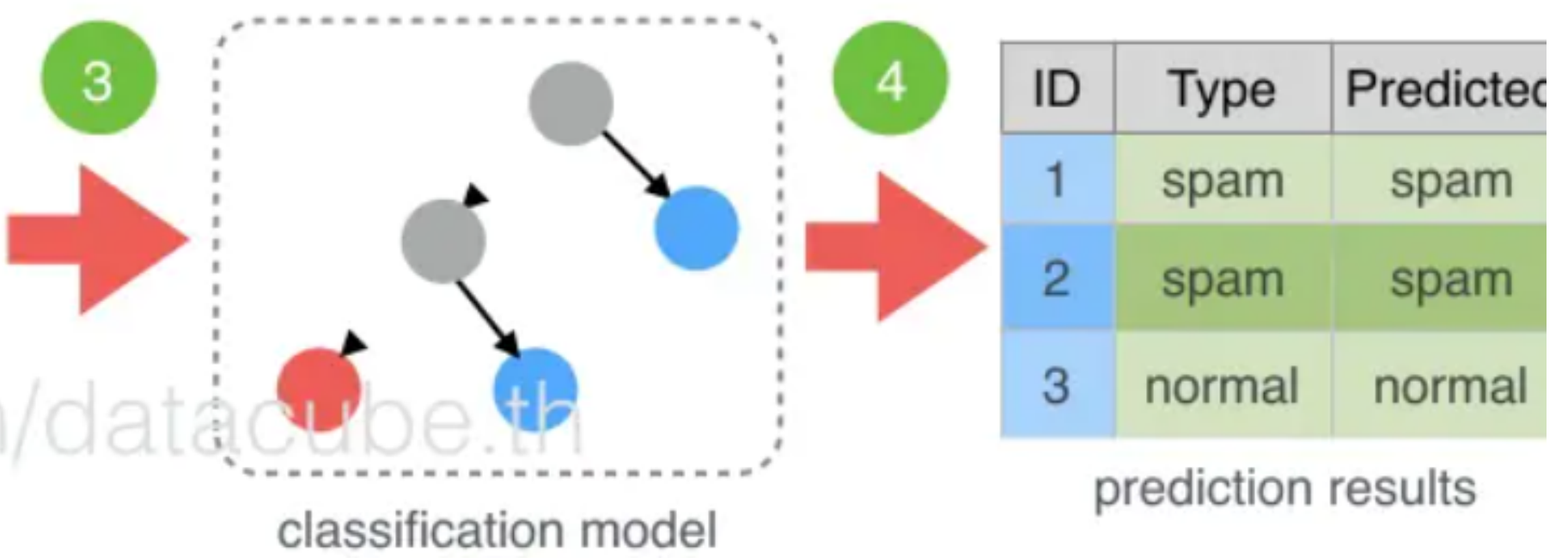
- การสร้างโมเดลและการทดสอบประสิทธิภาพ



ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal



ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal





Performance (classification)

- ตัววัดประสิทธิภาพของโมเดล classification

- **Confusion Matrix**

- True Positive (TP), True Negative (TN)
- False Positive (FP), False Negative (FN)

- Precision and Recall
- Accuracy
- F-Measure
- ROC Graph & Area Under Curve (AUC)





Performance (classification)

ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal



ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal

- พิจารณาคลาส normal

pred.\true.	normal	spam
normal	TP	FP
spam	FN	TN

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)



Performance (classification)

datacube

ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal



ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal

- พิจารณาคลาส normal

pred.\true.	normal	spam
normal	4	FP
spam	FN	TN

- True Positive (TP)

- จำนวนที่ทำนายตรงกับข้อมูลจริงในคลาสที่กำลังพิจารณา

- True Negative (TN)

- False Positive (FP)

- False Negative (FN)



Performance (classification)

ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal



ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal

พิจารณาคลาส normal

pred.\true.	normal	spam
normal	4	FP
spam	FN	6

- True Positive (TP)
- True Negative (TN)
 - จำนวนที่ทำนายตรงกับข้อมูลจริงในคลาสที่ไม่ได้กำลังพิจารณา
- False Positive (FP)
- False Negative (FN)



Performance (classification)

ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal



ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal

- พิจารณาคลาส normal

pred.\true.	normal	spam
normal	4	3
spam	FN	6

- True Positive (TP)
- True Negative (TN)
- **False Positive (FP)**
 - จำนวนที่ทำนายผิดเป็นคลาสที่กำลังพิจารณา
- False Negative (FN)



Performance (classification)

ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal



ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
8	spam	normal
9	normal	normal
10	normal	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	normal
15	normal	normal

- พิจารณาคลาส normal

pred.\true.	normal	spam
normal	4	3
spam	2	6

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)
 - จำนวนที่ทำนายผิดเป็นคลาสที่ไม่ได้กำลังพิจารณา



Performance (classification)

pred.\true.	normal	spam	
normal	TP	FP	Precision
spam	FN	TN	

confusion matrix ของคลาส normal

ID	Type	Predicted
3	normal	normal
8	spam	normal
9	normal	normal
10	normal	normal
13	spam	normal
14	spam	normal
15	normal	normal

predict เป็นคลาส normal

ID	Type	Predicted
1	spam	spam
2	spam	spam
4	normal	spam
5	spam	spam
6	spam	spam
7	normal	spam
11	spam	spam
12	spam	spam

predict เป็นคลาส spam

- Precision
 - จำนวนที่ทำนายถูกจากข้อมูลที่ทำนายว่าเป็นคลาสที่พิจารณาอยู่
- Precision สำหรับ normal
 - $$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$
 - $4/7 \times 100 = 57.12\%$
- Precision สำหรับ spam
 - $6/8 \times 100 = 75\%$



Performance (classification)

pred.\true.	normal	spam
normal	TP	FP
spam	FN	TN

Recall

ID	Type	Predicted
3	normal	normal
4	normal	spam
7	normal	spam
9	normal	normal
10	normal	normal
15	normal	normal

คลาส normal

ID	Type	Predicted
1	spam	spam
2	spam	spam
5	spam	spam
6	spam	spam
8	spam	normal
11	spam	spam
12	spam	spam
13	spam	normal
14	spam	spam

คลาส spam

- Recall
 - จำนวนข้อมูลที่ทำนายถูก
- Recall สำหรับ normal
 - $\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$
 - $4/6 \times 100 = 66.67\%$
- Recall สำหรับ spam
 - $7/9 \times 100 = 77.78\%$



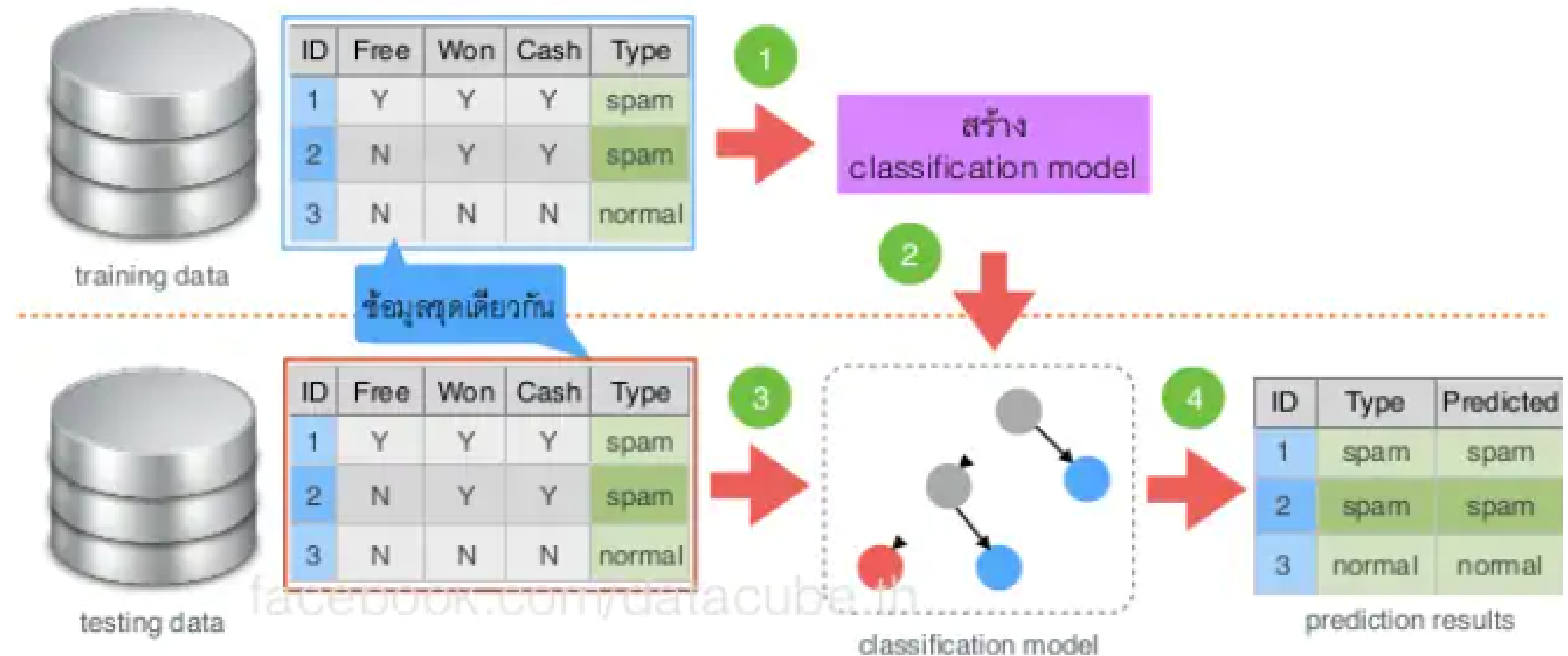
Validation

- การแบ่งข้อมูลเพื่อทดสอบประสิทธิภาพของโมเดล
 - Self consistency test (use training set)
 - Split test
 - Cross-validation test



Self Consistency test

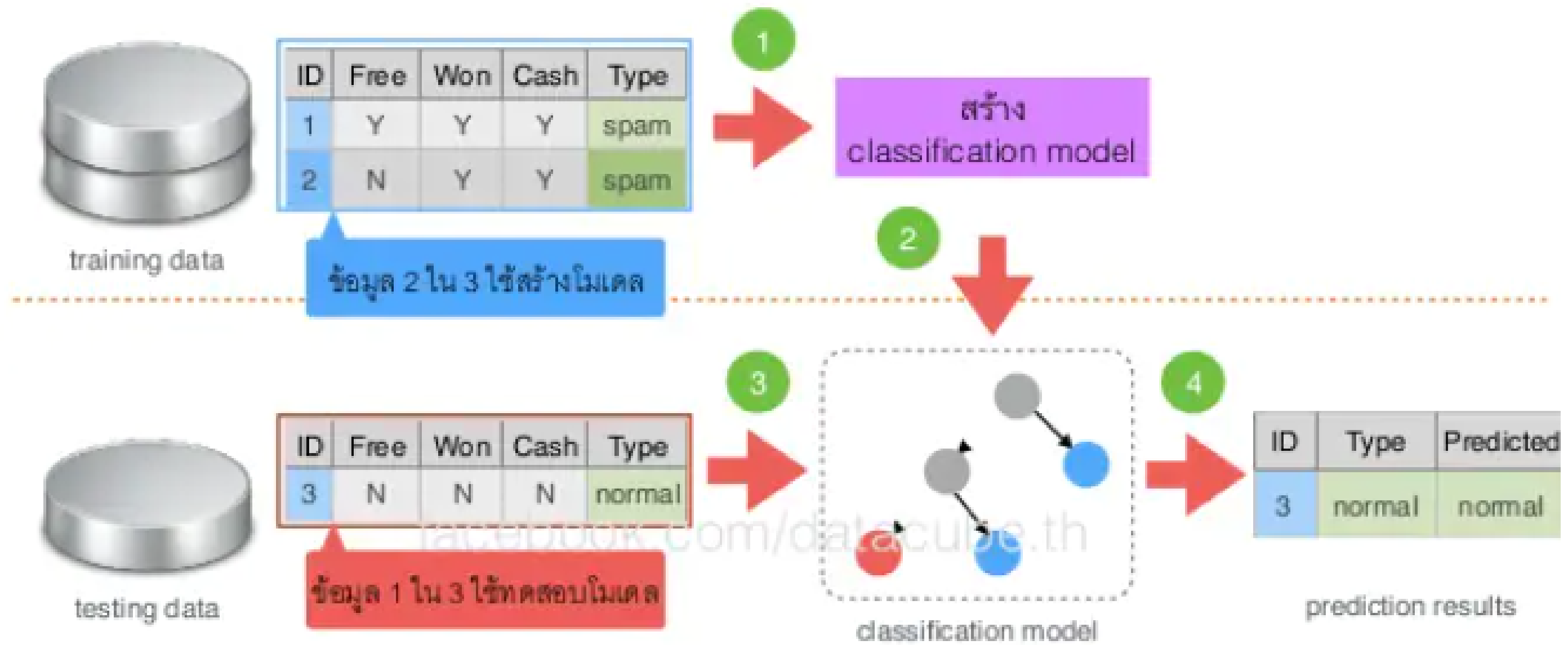
- ใช้ข้อมูล training ในการทดสอบประสิทธิภาพของโมเดล





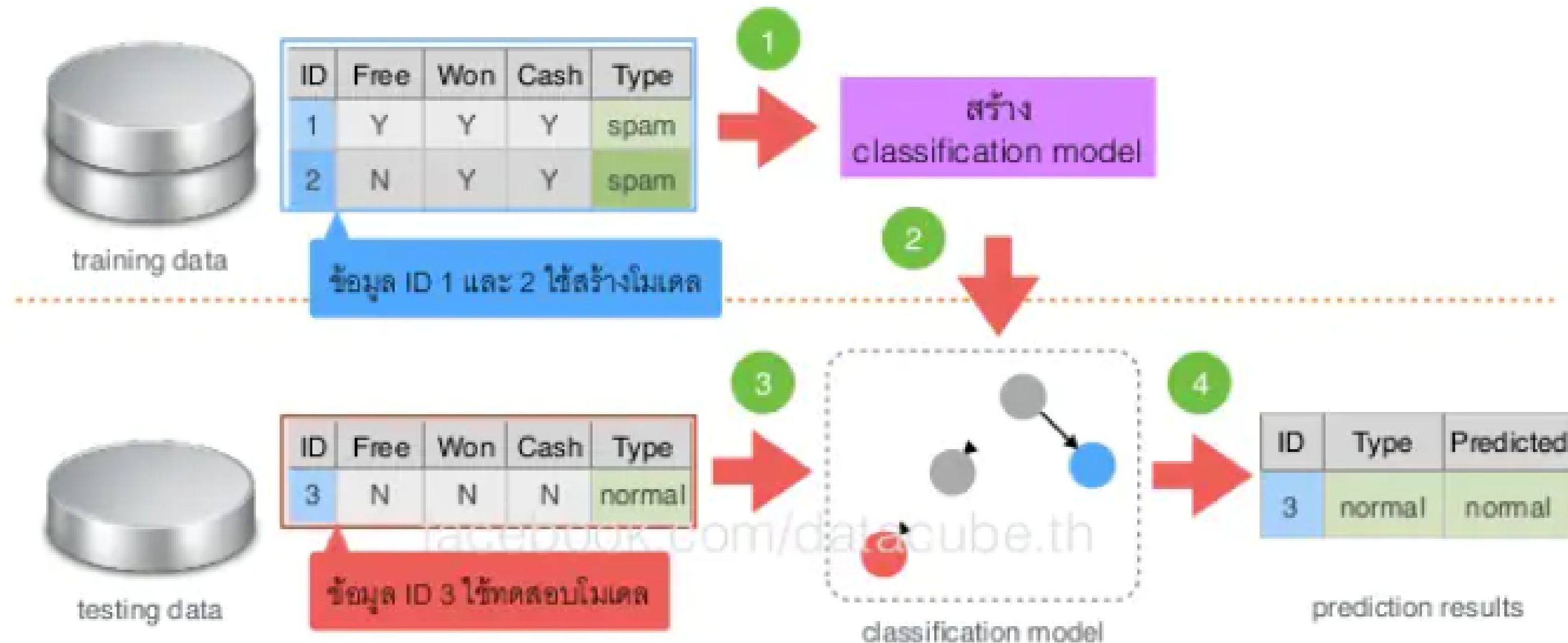
Split test

- แบ่งข้อมูลออกเป็น 2 ชุด
- training data สำหรับสร้างโมเดล และ testing data สำหรับทดสอบ



Cross-validation

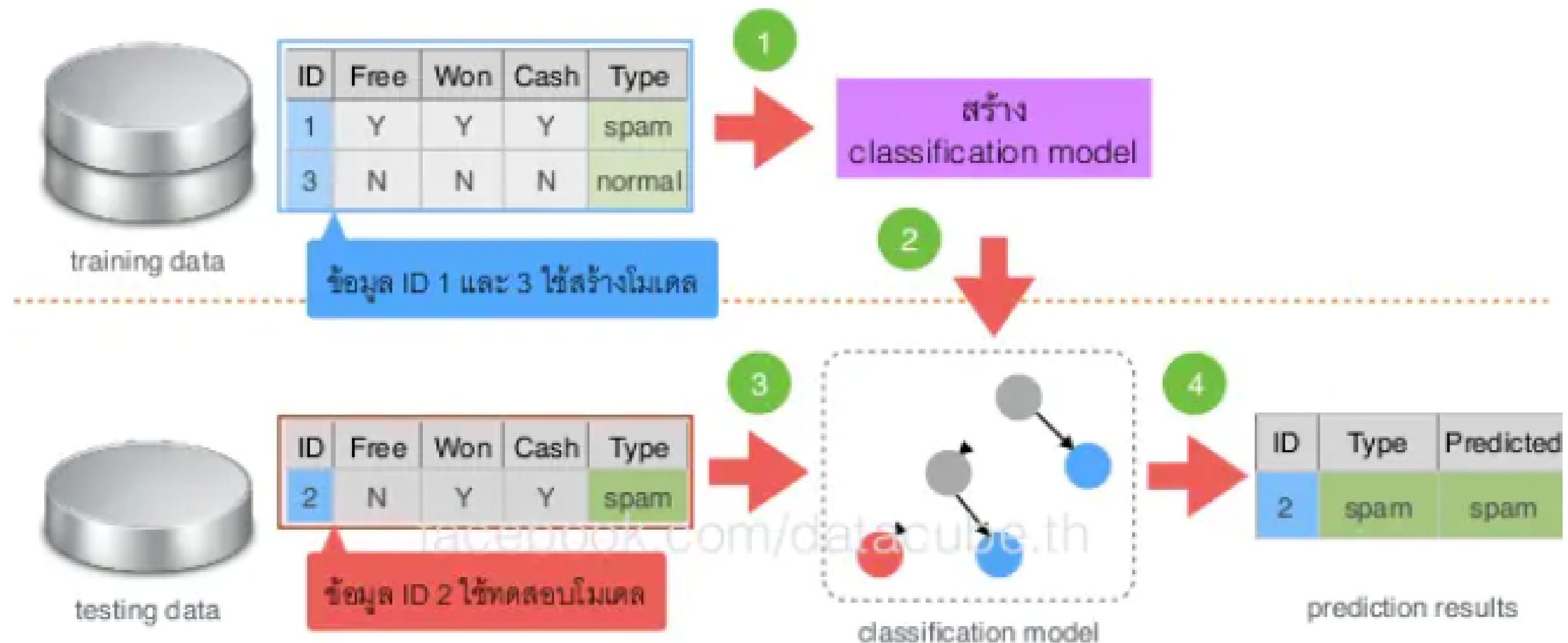
- แบ่งข้อมูลออกเป็น N ชุด เช่น $N = 5$ หรือ 10
- ข้อมูล $N-1$ ชุดสำหรับสร้างโมเดล และ ข้อมูลส่วนที่เหลือสำหรับทดสอบ วนทำจนครบ N





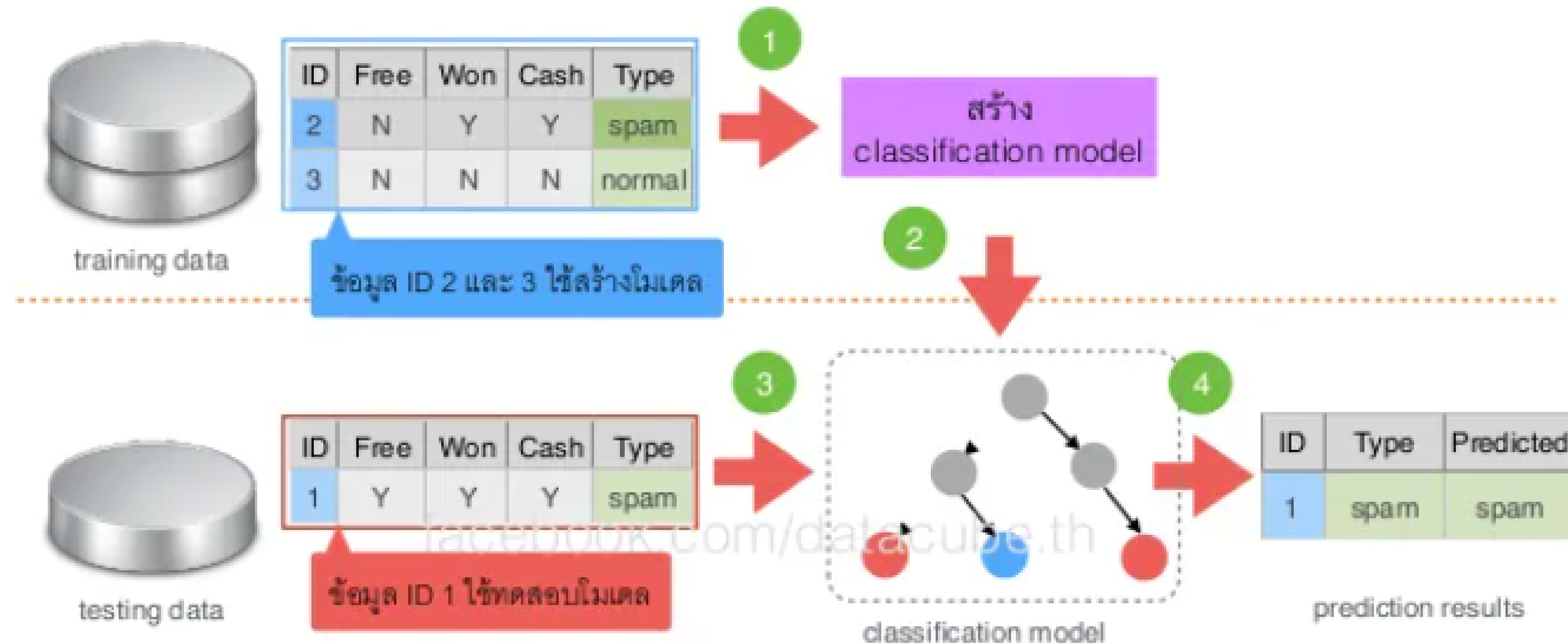
Cross-validation

- แบ่งข้อมูลออกเป็น N ชุด เช่น $N = 5$ หรือ 10
- ข้อมูล $N-1$ ชุดสำหรับสร้างโมเดล และ ข้อมูลส่วนที่เหลือสำหรับทดสอบ วนทำจนครบ N



Cross-validation

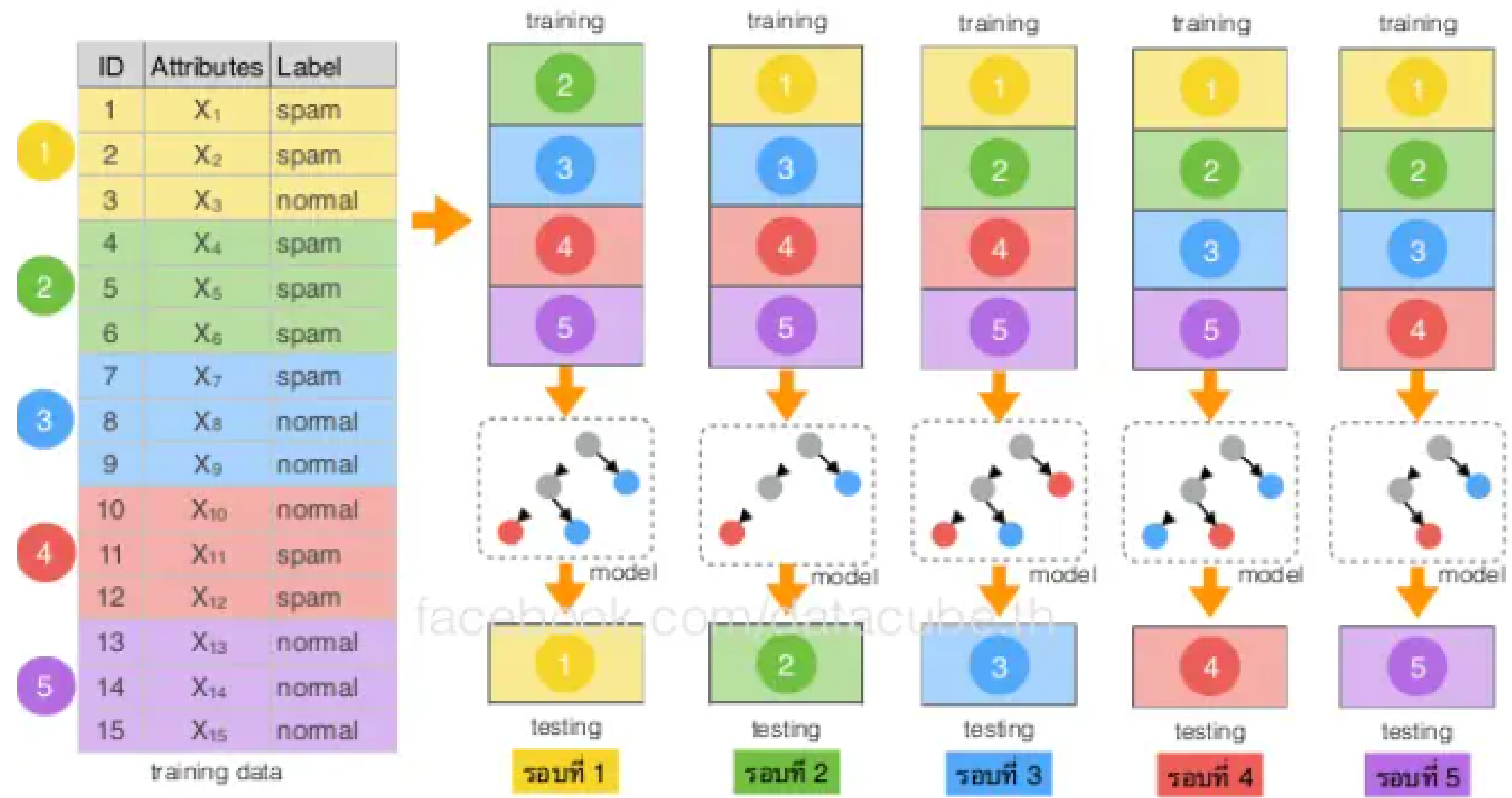
- แบ่งข้อมูลออกเป็น N ชุด เช่น $N = 5$ หรือ 10
- ข้อมูล $N-1$ ชุดสำหรับสร้างโมเดล และ ข้อมูลส่วนที่เหลือสำหรับทดสอบ วนทำงานครบ N





Cross-validation

- ตัวอย่างของ 5-fold cross-validation





Classification Techniques

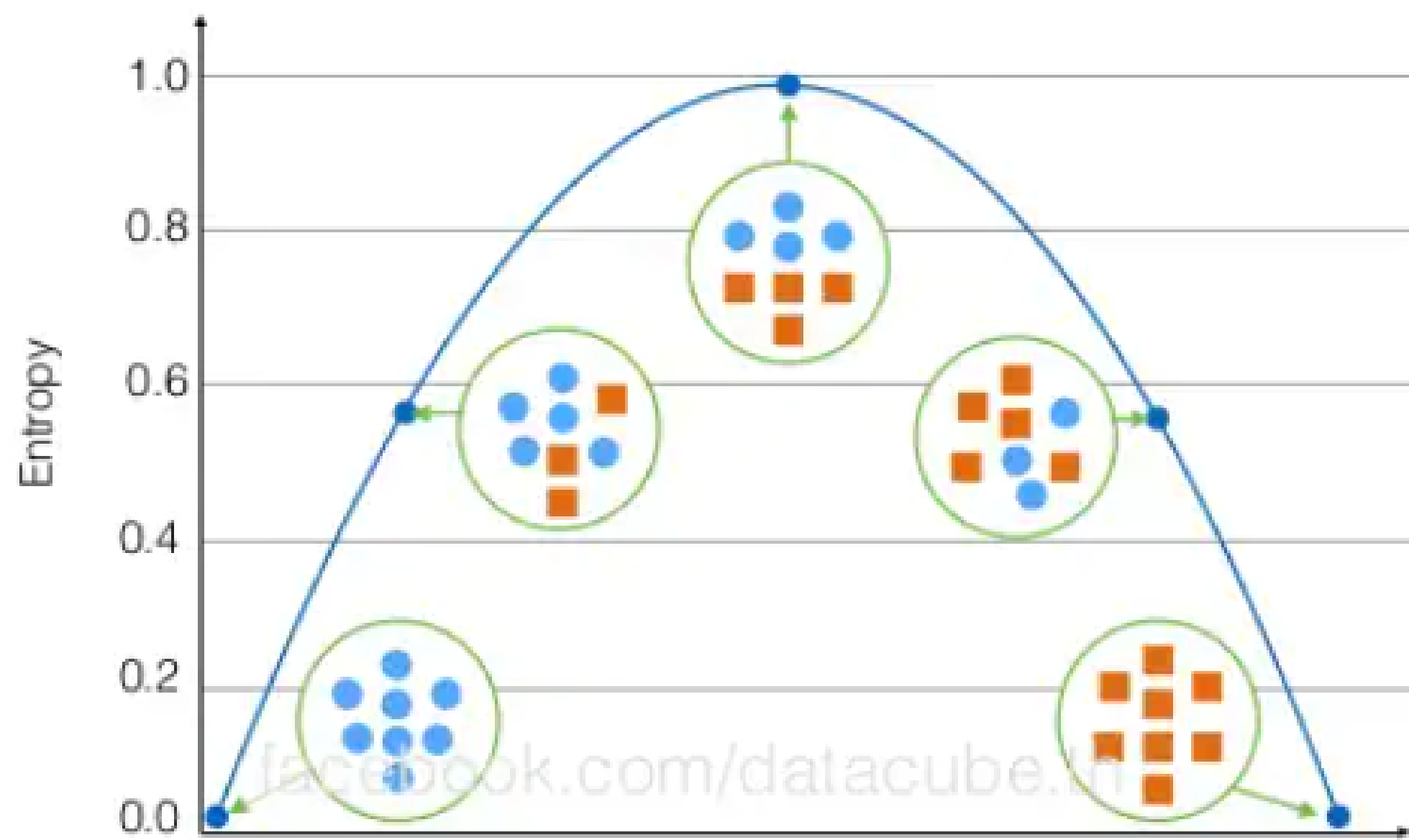
- Decision Tree
 - Naive Bayes
 - K-Nearest Neighbors (kNN)
 - Linear Regression
 - Neural Network
 - Support Vector Machines
-
- Ensemble Classifiers (Vote)
 - Attribute Selection
 - Compare classification performance





Decision Tree

- ลักษณะของค่า Entropy





Decision Tree

- ข้อมูล Weather
- เก็บสภาพภูมิอากาศจำนวน 14 วันเพื่อพิจารณาว่าจะมีการแข่งขันกีฬาได้หรือไม่

ID	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	mild	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	mild	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

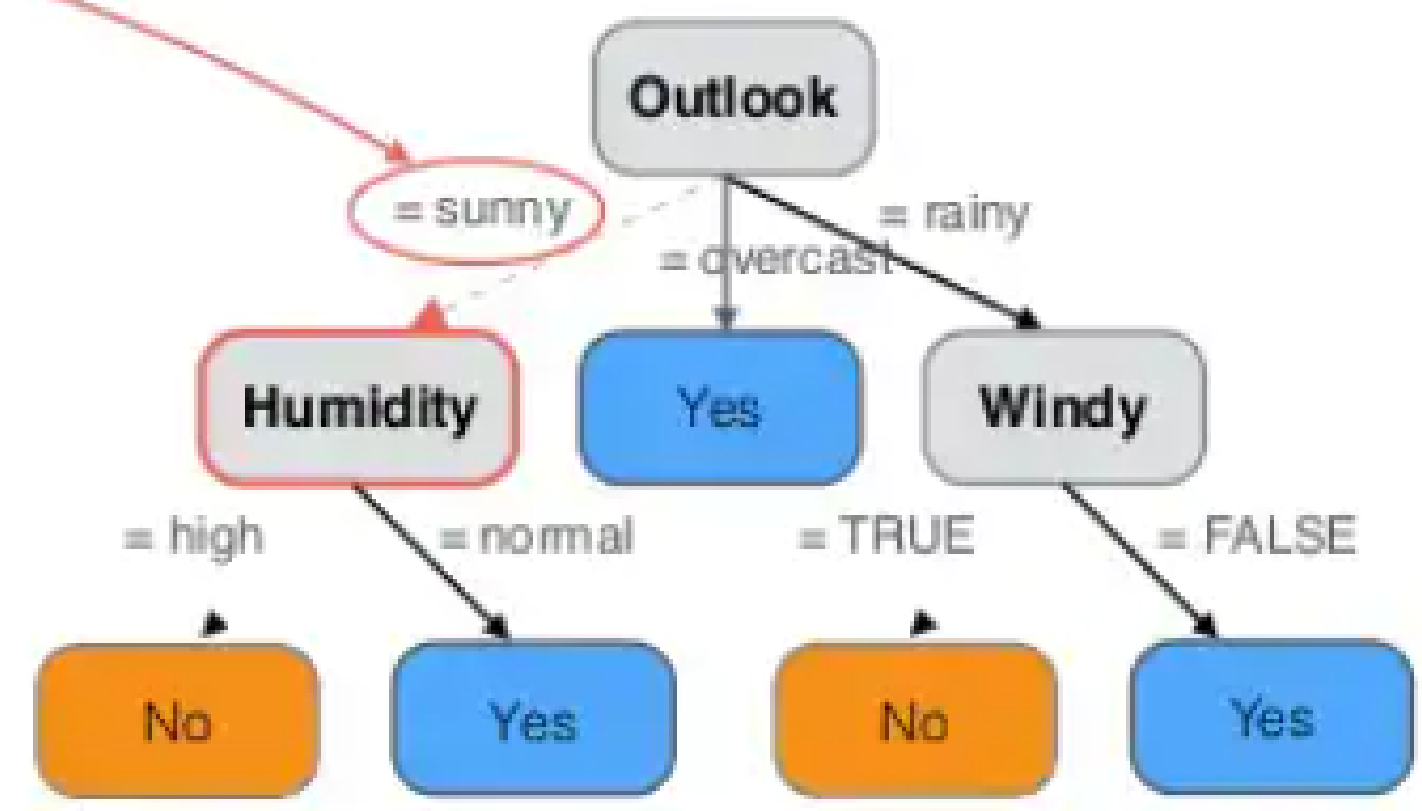


Decision Tree

- การใช้โมเดล predict ข้อมูลใหม่

ID	Outlook	Temperature	Humidity	Windy
1	sunny	not	high	FALSE

ข้อมูลที่ใช้ทดสอบ



โมเดล decision tree

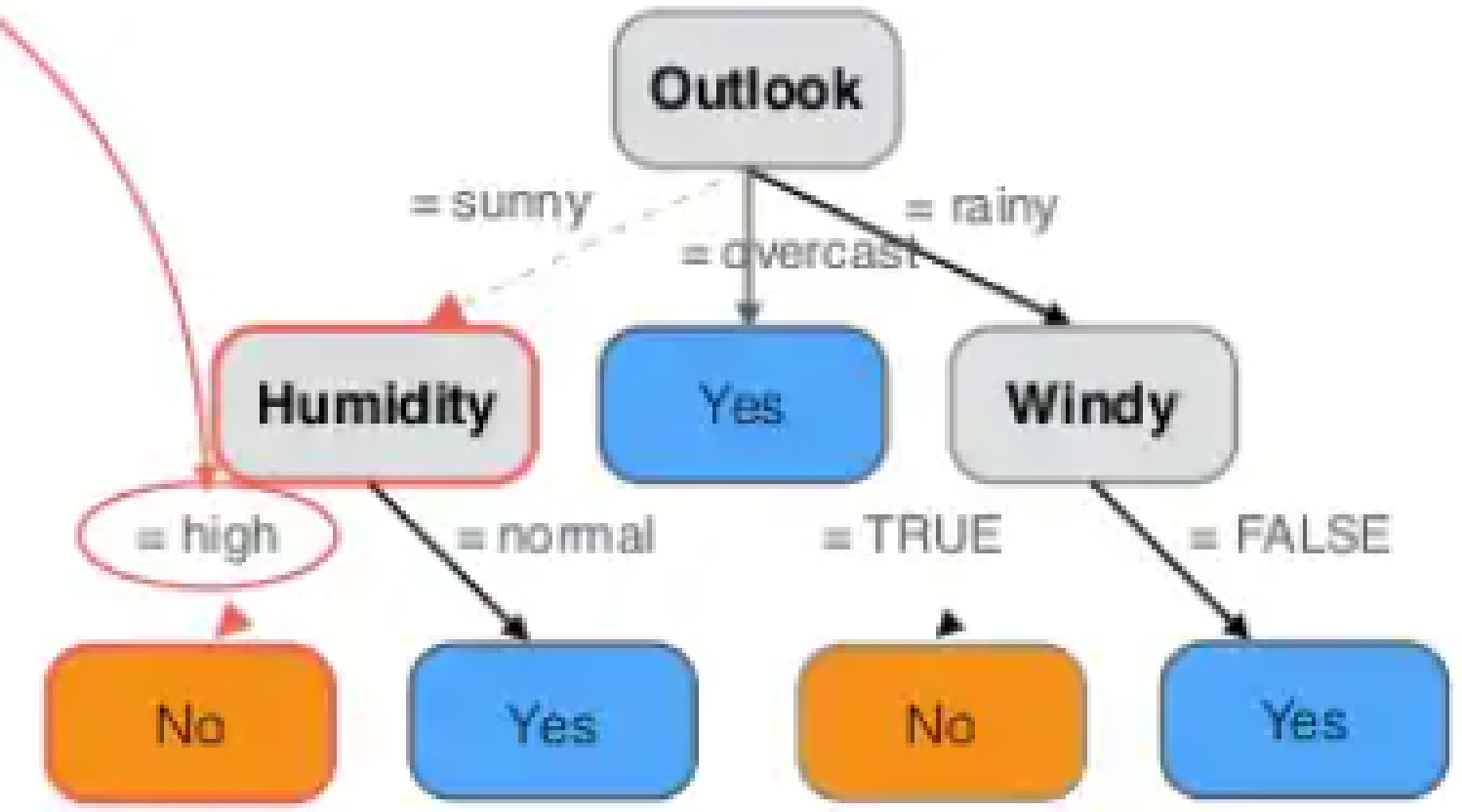


Decision Tree

- การใช้โมเดล predict ข้อมูลใหม่

ID	Outlook	Temperature	Humidity	Windy
1	sunny	hot	high	FALSE

ข้อมูลที่ใช้ทดสอบ



โมเดล decision tree



Decision Tree

- ข้อมูลเป็นตัวเลข
 - เรียงลำดับข้อมูลที่เป็นตัวเลขจากน้อยไปมาก
 - แบ่งข้อมูลออกเป็น 2 ส่วนโดยการหาจุดกึ่งกลางระหว่างค่าตัวเลข 2 ค่า
 - คำนวณค่า Information Gain จากข้อมูล 2 ส่วนที่แบ่งได้
 - เลือกจุดกึ่งกลางที่ให้ค่า Information Gain สูงที่สุดมาใช้งานต่อ



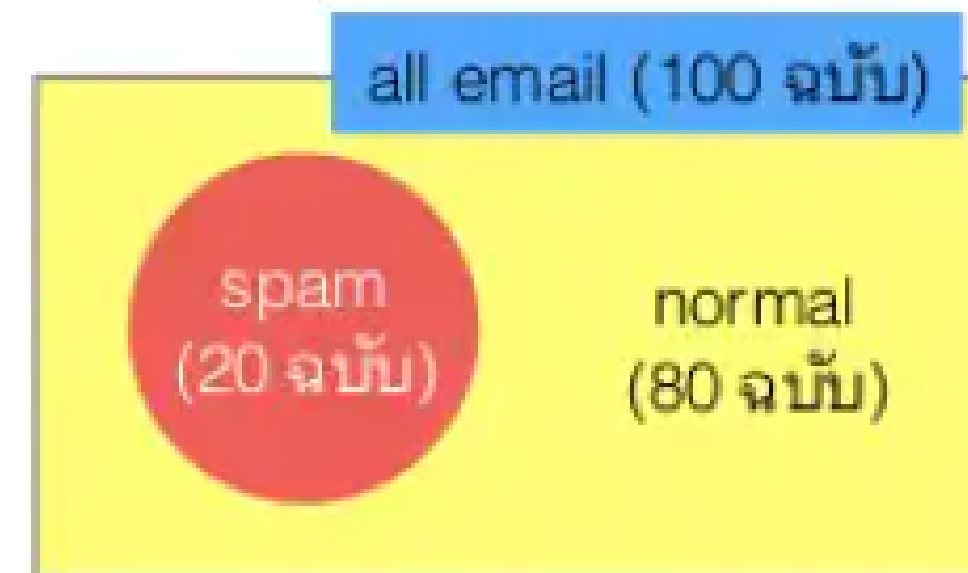
Decision Tree

- Decision Tree
- Naive Bayes
- K-Nearest Neighbors (kNN)
- Linear Regression
- Neural Network
- Support Vector Machines
- Ensemble Classifiers (Vote)
- Attribute Selection
- Compare classification performance



Classification Techniques

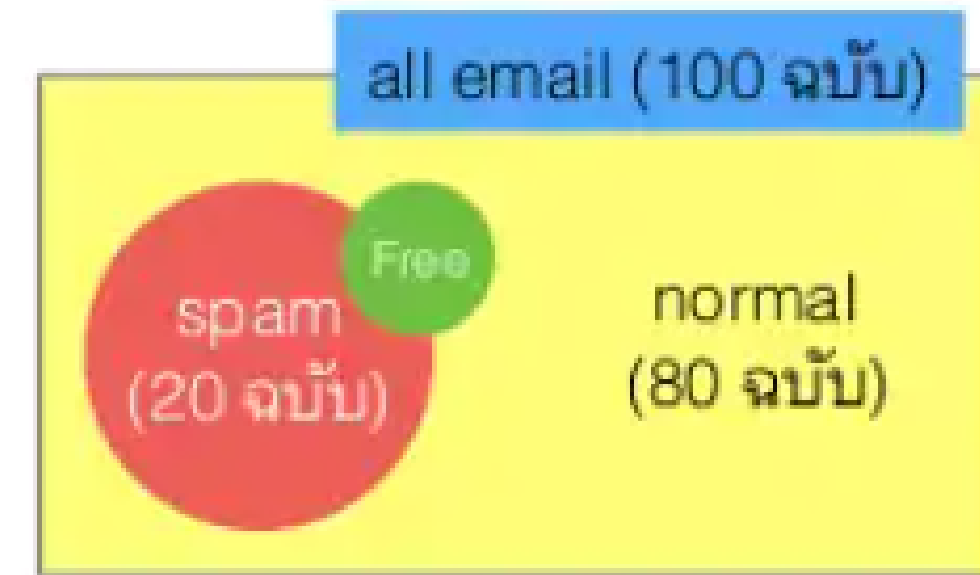
- ความน่าจะเป็น (probability)
 - โอกาสที่เกิดเหตุการณ์จากเหตุการณ์ทั้งหมด ใช้สัญลักษณ์ $P()$ หรือ $Pr()$
 - โยนเหรียญบาท (มีหัวและก้อย)
 - โอกาสได้หัว มีค่าความน่าจะเป็น $1/2 = 0.5$
 - โอกาสได้ก้อย มีค่าความน่าจะเป็น $1/2 = 0.5$
 - ความน่าจะเป็นของการพบ spam email
 - มี email ทั้งหมด 100 ฉบับ
 - มี spam email ทั้งหมด 20 ฉบับ
 - มี normal email ทั้งหมด 80 ฉบับ
 - โอกาสที่ email จะเป็น spam มีความน่าจะเป็น $20/100 = 0.2$ หรือ $P(\text{spam}) = 0.2$
 - โอกาสที่ email จะเป็น normal มีความน่าจะเป็น $80/100 = 0.8$ หรือ $P(\text{normal}) = 0.8$





Probability

- Joint Probability
 - ความน่าจะเป็นที่ 2 เหตุการณ์เกิดร่วมกัน
 - ความน่าจะเป็นที่มีคำว่า Free อยู่ใน spam email
 - สัญลักษณ์ $P(\text{Free}=Y \cap \text{spam})$





Naive Bayes

- ใช้หลักการของความน่าจะเป็น (probability)

ความน่าจะเป็นที่ B เกิด
ก่อนและ A เกิดตามมา

ความน่าจะเป็นที่ A
และ B เกิดร่วมกัน

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$$

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

Bayes Theorem



Probability

Posterior probability

Likelihood

Prior probability

$$P(C|A) = \frac{P(A|C) \times P(C)}{P(A)}$$

- $P(C|A)$ คือ ความน่าจะเป็นของข้อมูลที่มีแอตทริบิวต์ A จะมีคลาส C
- $P(A|C)$ คือ ความน่าจะเป็นของข้อมูลใน training data ที่มีแอตทริบิวต์ A และมีคลาส C
 - $P(A|C) = P(a_1 \cap a_2 \cap a_3 \dots \cap a_M|C)$
 - $P(A|C) = P(a_1|C) \times P(a_2|C) \times \dots \times P(a_M|C)$
- $P(C)$ หรือ $P(A)$ คือ ความน่าจะเป็นของคลาส C หรือ แอตทริบิวต์ A



Naive Bayes

- การใช้โมเดลเพื่อ predict ข้อมูลใหม่

$$P(\text{Type} = \text{normal}) = 5/10 = 0.50$$

$$P(\text{Type} = \text{spam}) = 5/10 = 0.50$$

attribute	Type = normal	Type = spam
Free = Y	0/5 = 0.00	3/5 = 0.60
Free = N	5/5 = 1.00	2/5 = 0.40
Won = Y	0/5 = 0.00	3/5 = 0.60
Won = N	5/5 = 1.00	2/5 = 0.40
Cash = Y	0/5 = 0.00	2/5 = 0.40
Cash = N	5/5 = 1.00	3/5 = 0.60

โมเดล Naive Bayes

ID	Free	Won	Cash
1	Y	Y	Y

ข้อมูลที่ใช้ทดสอบ

$$P(C|A) = P(A|C) \times P(C)$$

$$\begin{aligned}
 P(\text{Type} = \text{normal}|A) &= P(\text{Free} = Y|\text{Type} = \text{normal}) \times \\
 &\quad P(\text{Won} = Y|\text{Type} = \text{normal}) \times \\
 &\quad P(\text{Cash} = Y|\text{Type} = \text{normal}) \times \\
 &\quad P(\text{Type} = \text{normal}) \\
 &= 0.00 \times 0.00 \times 0.00 \times 0.50 \\
 &= 0.00
 \end{aligned}$$

$$\begin{aligned}
 P(\text{Type} = \text{spam}|A) &= P(\text{Free} = Y|\text{Type} = \text{spam}) \times \\
 &\quad P(\text{Won} = Y|\text{Type} = \text{spam}) \times \\
 &\quad P(\text{Cash} = Y|\text{Type} = \text{spam}) \times \\
 &\quad P(\text{Type} = \text{spam}) \\
 &= 0.60 \times 0.60 \times 0.40 \times 0.50 \\
 &= 0.07
 \end{aligned}$$

ค่า prob มากสุด



Classification Techniques

- Decision Tree
- Naive Bayes
- K-Nearest Neighbors (kNN)
- Linear Regression
- Neural Network
- Support Vector Machines
- Ensemble Classifiers (Vote)
- Attribute Selection
- Compare classification performance



K-Nearest Neighbors (KNN)

- การใช้โมเดลเพื่อ predict ข้อมูลใหม่

ID	Na/K	Age
11	50	30

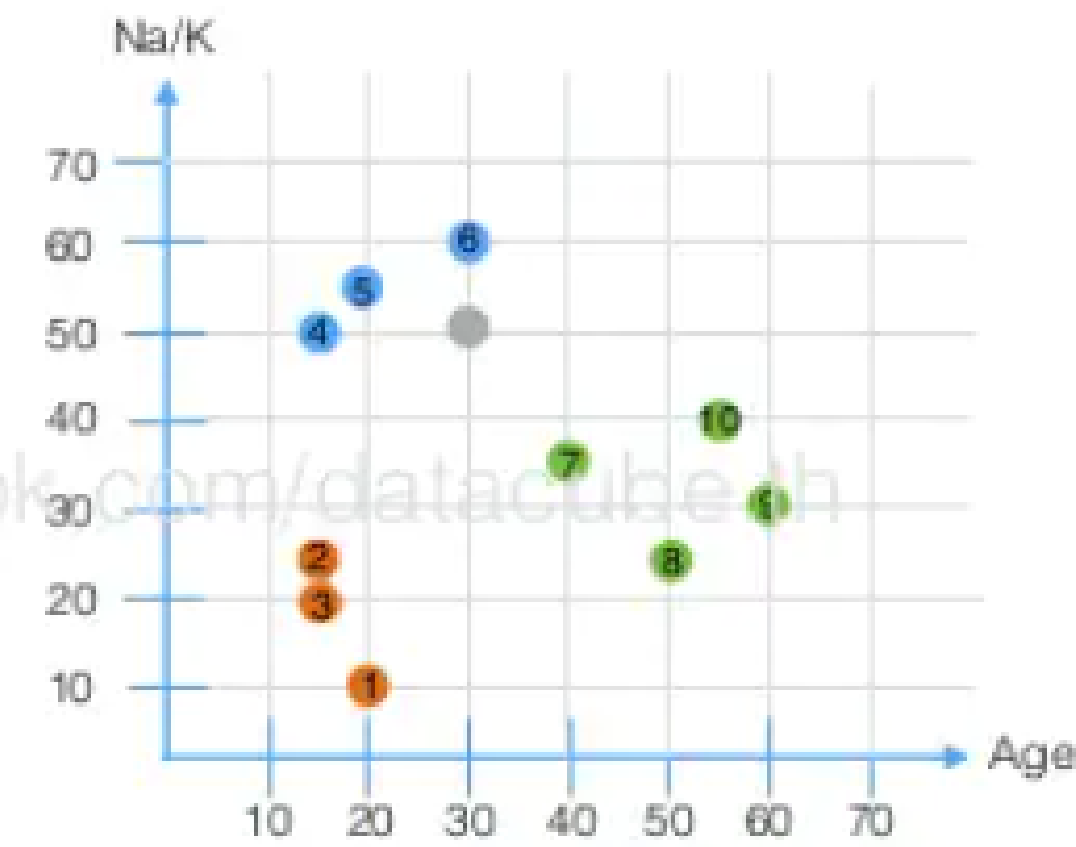
unseen data

ID	Na/K	Age	Type
1	10	20	A
2	25	15	A
3	20	15	A
4	50	15	B
5	55	20	B
6	60	30	B
7	35	40	C
8	25	50	C
9	30	60	C
10	40	55	C



ID	ระยะห่างจากข้อมูล ID = 11
1	$\sqrt{(10-50)^2 + (20-30)^2} = 41.23$
2	$\sqrt{(25-50)^2 + (15-30)^2} = 29.15$
3	$\sqrt{(20-50)^2 + (15-30)^2} = 33.54$
4	$\sqrt{(50-50)^2 + (15-30)^2} = 15.00$
5	$\sqrt{(55-50)^2 + (20-30)^2} = 11.18$
6	$\sqrt{(60-50)^2 + (30-30)^2} = 10.00$
7	$\sqrt{(35-50)^2 + (40-30)^2} = 18.03$
8	$\sqrt{(25-50)^2 + (50-30)^2} = 32.02$
9	$\sqrt{(30-50)^2 + (60-30)^2} = 36.06$
10	$\sqrt{(40-50)^2 + (55-30)^2} = 26.93$

- Type = A (orange dot)
- Type = B (blue dot)
- Type = C (green dot)



โมเดล K-NN



K-Nearest Neighbors (KNN)

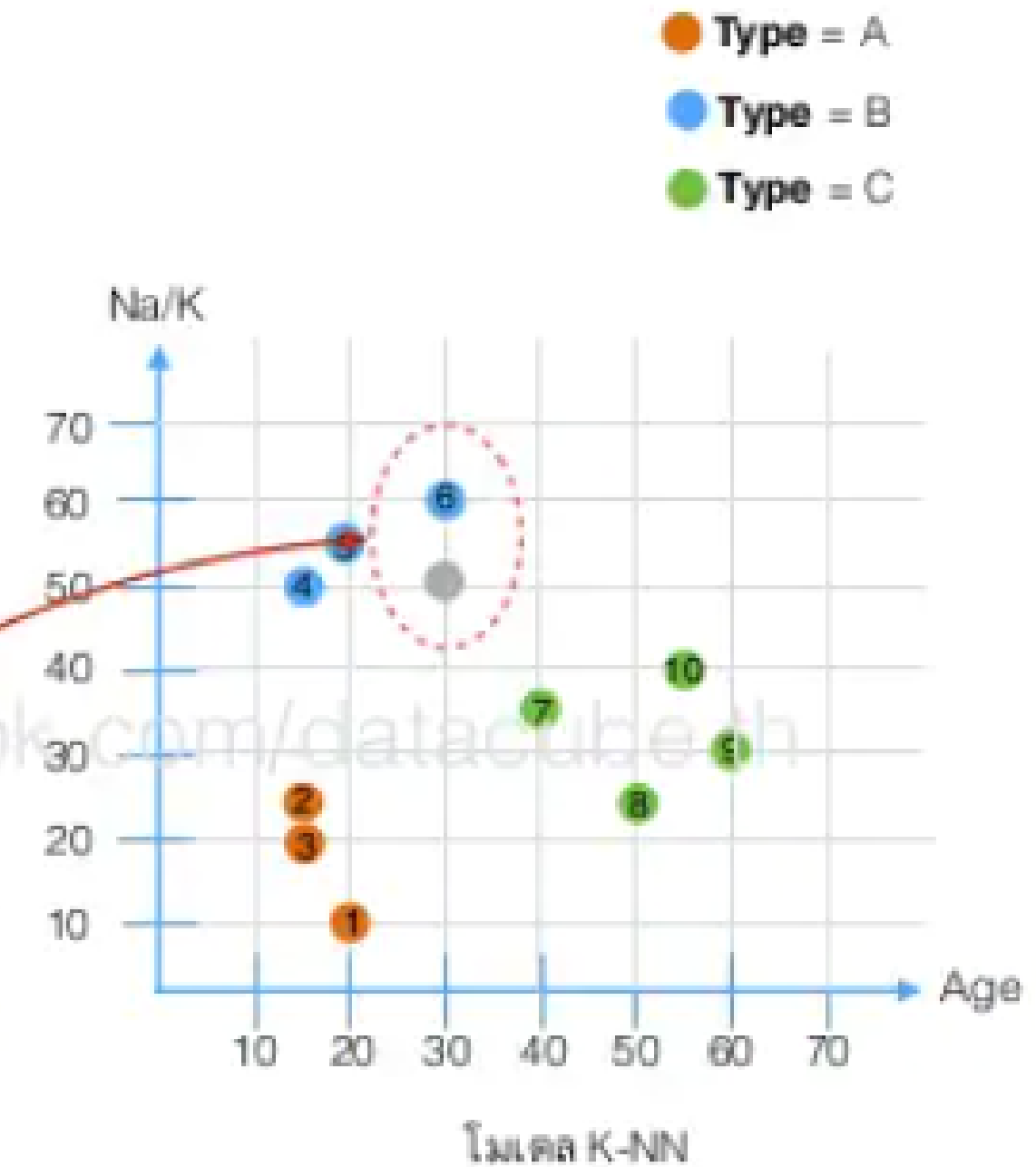
- การใช้โมเดลเพื่อ predict ข้อมูลใหม่ โดยกำหนด $K = 1$

ID	Na/K	Age
11	50	30

unseen data

ID	Na/K	Age	Type
1	10	20	A
2	25	15	A
3	20	15	A
4	50	15	B
5	55	20	B
6	60	30	B
7	35	40	C
8	25	50	C
9	30	60	C
10	40	55	C

ID	ระยะห่างจากข้อมูล ID = 11
1	$\sqrt{(10-50)^2 + (20-30)^2} = 41.23$
2	$\sqrt{(25-50)^2 + (15-30)^2} = 29.15$
3	$\sqrt{(20-50)^2 + (15-30)^2} = 33.54$
4	$\sqrt{(50-50)^2 + (15-30)^2} = 15.00$
5	$\sqrt{(55-50)^2 + (20-30)^2} = 11.18$
6	$\sqrt{(60-50)^2 + (30-30)^2} = 10.00$
7	$\sqrt{(35-50)^2 + (40-30)^2} = 18.03$
8	$\sqrt{(25-50)^2 + (50-30)^2} = 32.02$
9	$\sqrt{(30-50)^2 + (60-30)^2} = 36.06$
10	$\sqrt{(40-50)^2 + (55-30)^2} = 26.93$





K-Nearest Neighbors (KNN)

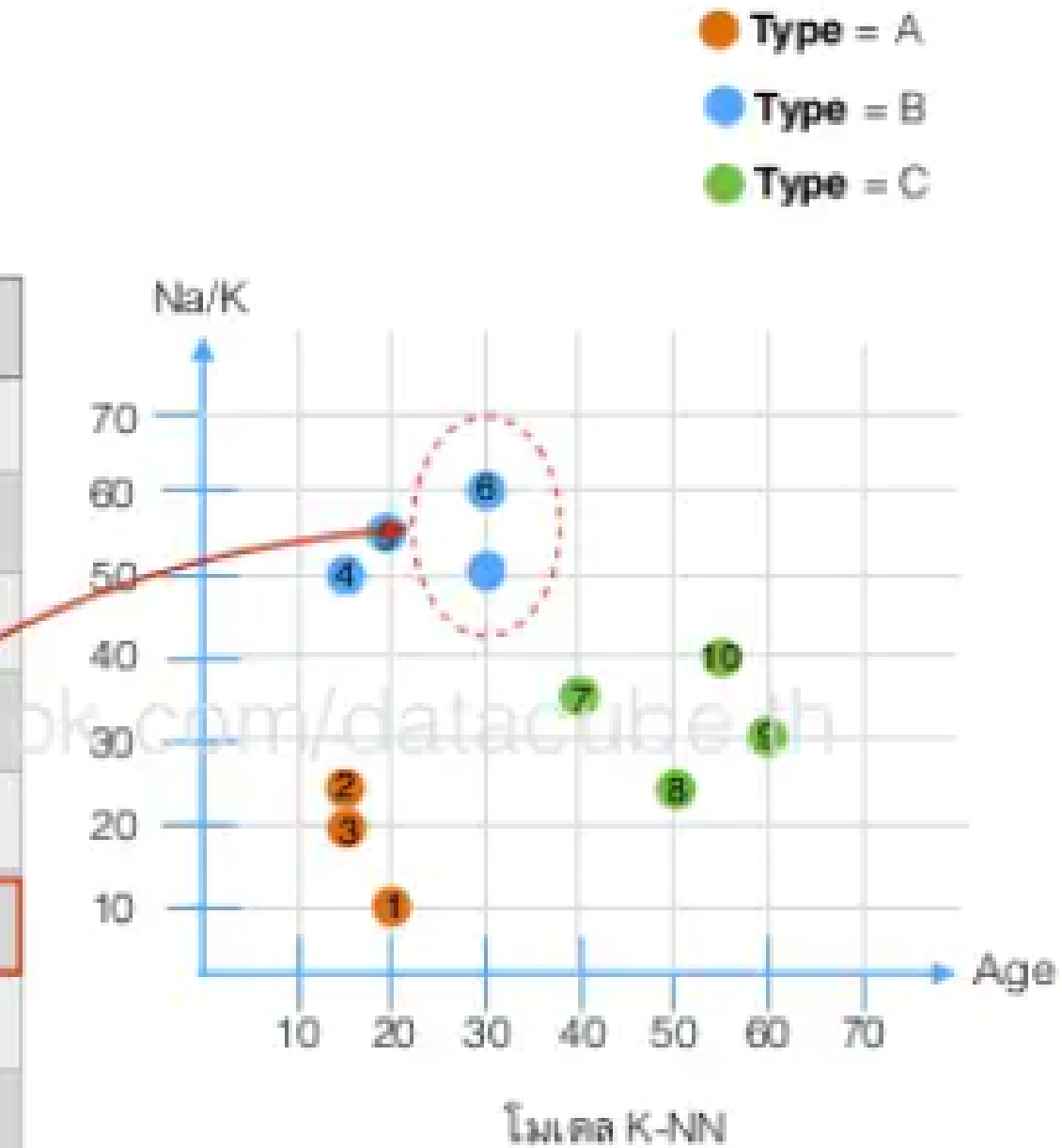
- การใช้โมเดลเพื่อ predict ข้อมูลใหม่ โดยกำหนด $K = 1$

ID	Na/K	Age
11	50	30

unseen data

ID	Na/K	Age	Type
1	10	20	A
2	25	15	A
3	20	15	A
4	50	15	B
5	55	20	B
6	60	30	B
7	35	40	C
8	25	50	C
9	30	60	C
10	40	55	C

ID	ระยะห่างจากข้อมูล ID = 11
1	$\sqrt{(10-50)^2 + (20-30)^2} = 41.23$
2	$\sqrt{(25-50)^2 + (15-30)^2} = 29.15$
3	$\sqrt{(20-50)^2 + (15-30)^2} = 33.54$
4	$\sqrt{(50-50)^2 + (15-30)^2} = 15.00$
5	$\sqrt{(55-50)^2 + (20-30)^2} = 11.18$
6	$\sqrt{(60-50)^2 + (30-30)^2} = 10.00$
7	$\sqrt{(35-50)^2 + (40-30)^2} = 18.03$
8	$\sqrt{(25-50)^2 + (50-30)^2} = 32.02$
9	$\sqrt{(30-50)^2 + (60-30)^2} = 36.06$
10	$\sqrt{(40-50)^2 + (55-30)^2} = 26.93$





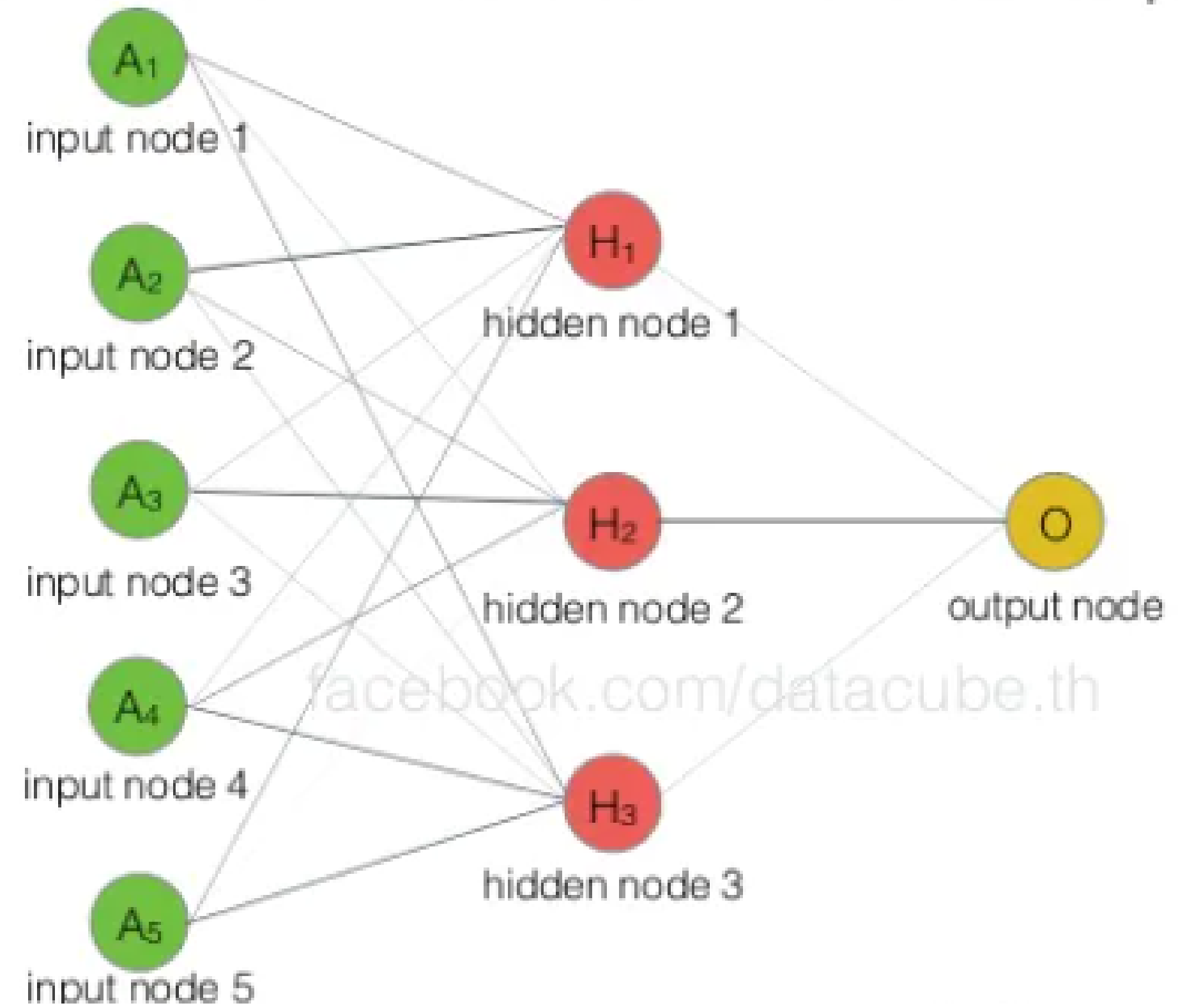
Classification Techniques

- Decision Tree
- Naive Bayes
- K-Nearest Neighbors (kNN)
- Linear Regression
- **Neural Network**
- Support Vector Machines
- Ensemble Classifiers (Vote)
- Attribute Selection
- Compare classification performance



Neural Network

- โมเดลทางคณิตศาสตร์ที่เลียนแบบการทำงานของสมองมนุษย์





Neural Network

- โมเดลที่ได้แปลความหมายได้ยาก จะอยู่ในรูปแบบของสมการคณิตศาสตร์
 - โมเดลจะแสดงค่าน้ำหนัก (weight) ระหว่างโหนดต่างๆ ใน Neural Network
 - โมเดลจะทำการปรับเปลี่ยนค่าน้ำหนักให้เป็นค่าที่เหมาะสม
- สามารถใช้กับ classification (ทำนายค่า nominal) และ regression (ทำนายค่าตัวเลข) ได้

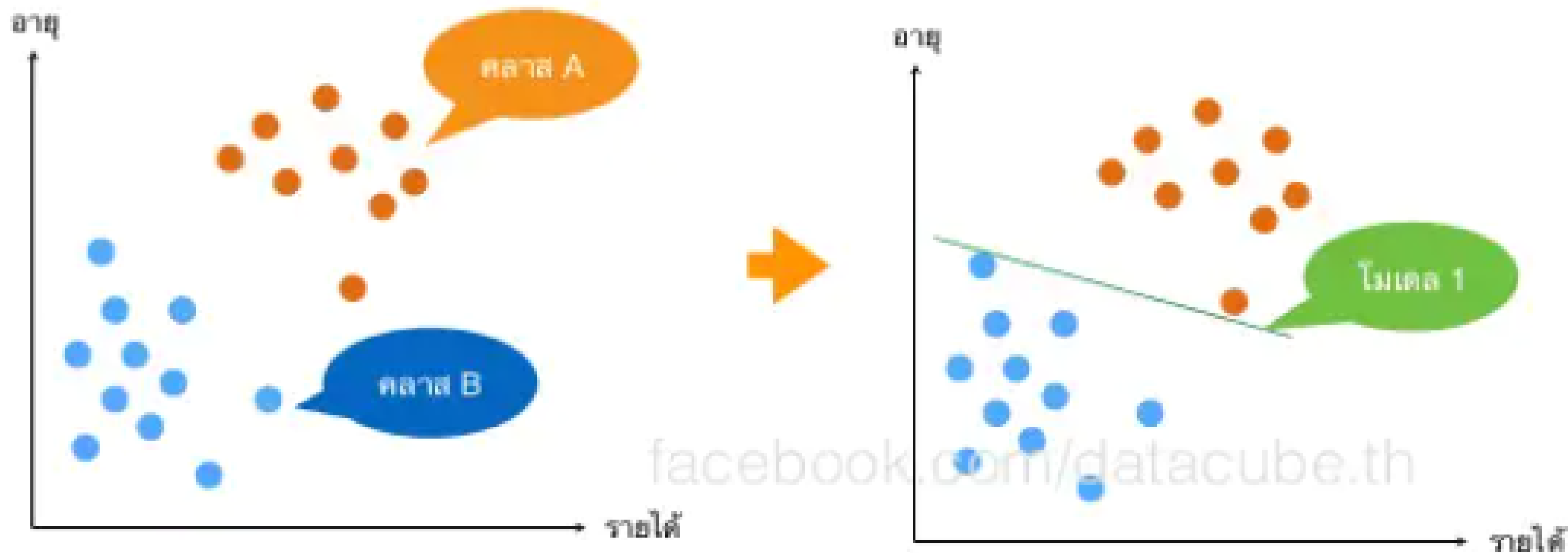


Classification Techniques

- Decision Tree
- Naive Bayes
- K-Nearest Neighbors (kNN)
- Linear Regression
- Neural Network
- Support Vector Machines
- Ensemble Classifiers (Vote)
- Attribute Selection
- Compare classification performance

Support Vector Machines (SVM)

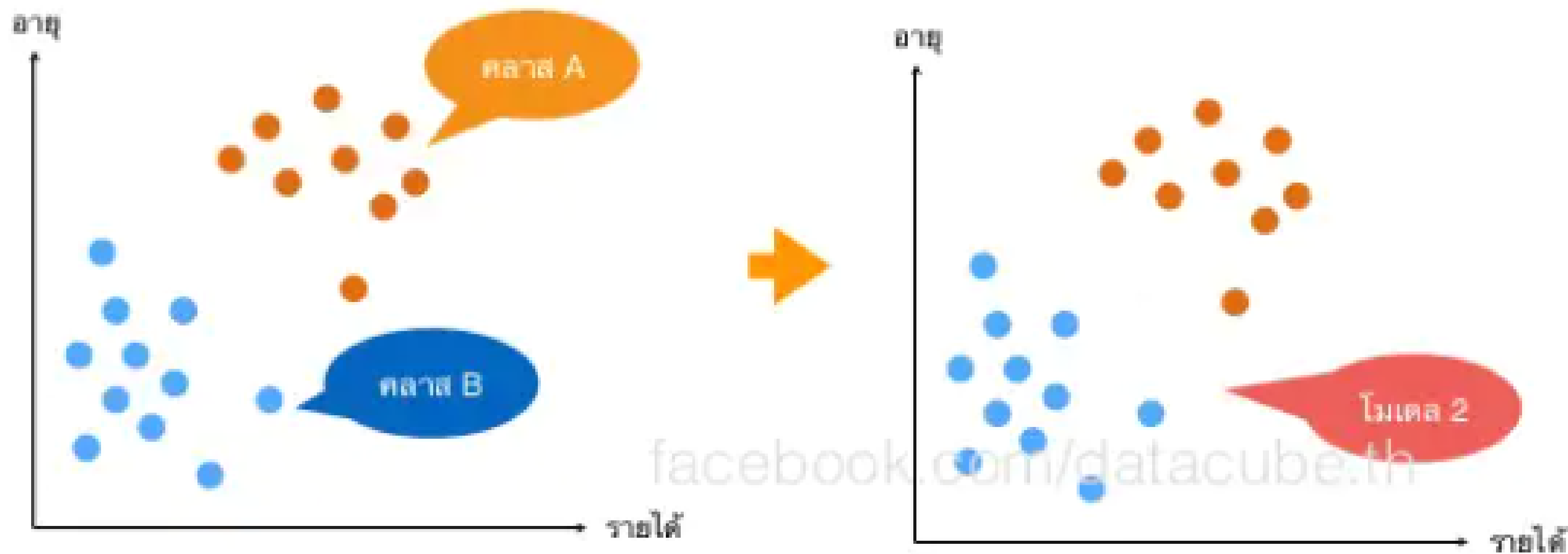
- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส





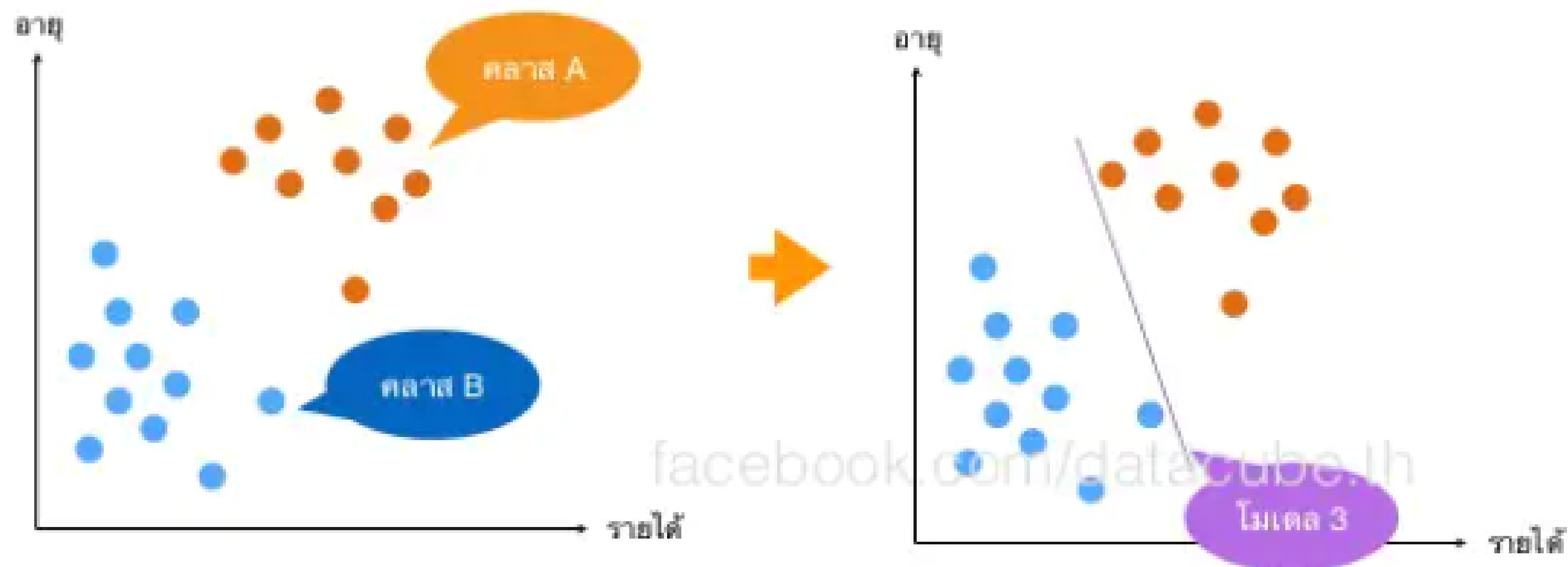
Support Vector Machines (SVM)

- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส



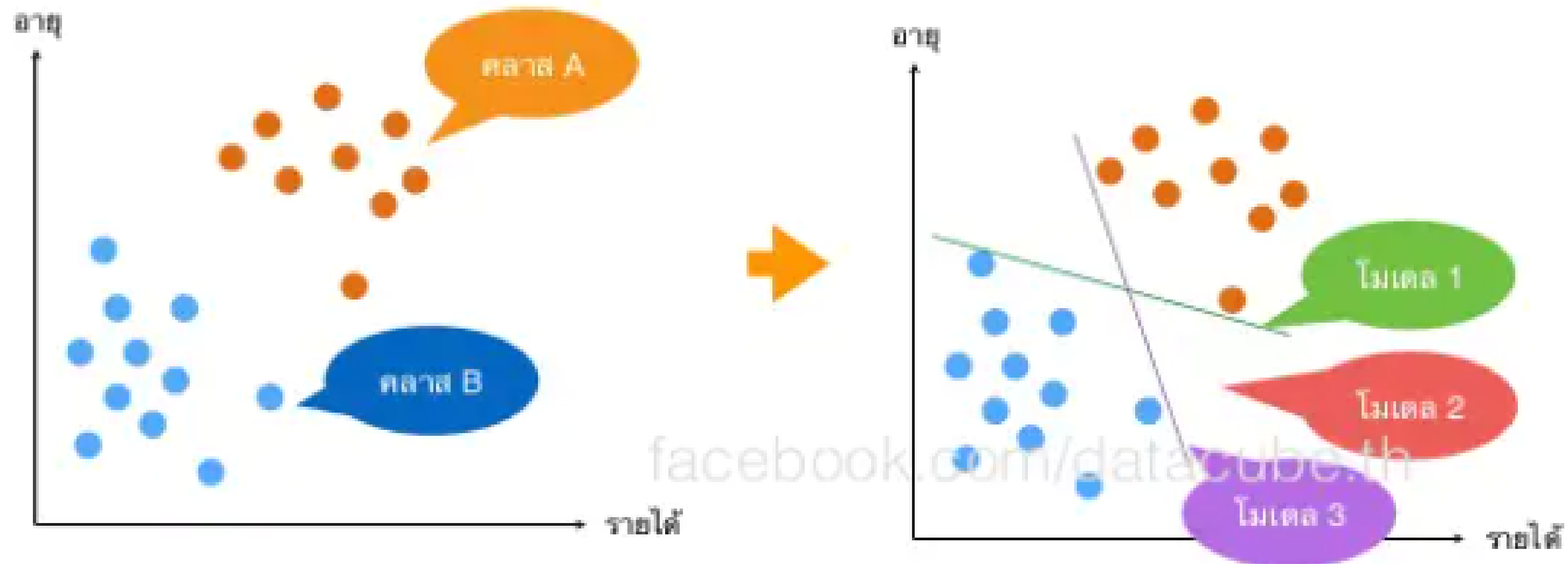
Support Vector Machines (SVM)

- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส



Support Vector Machines (SVM)

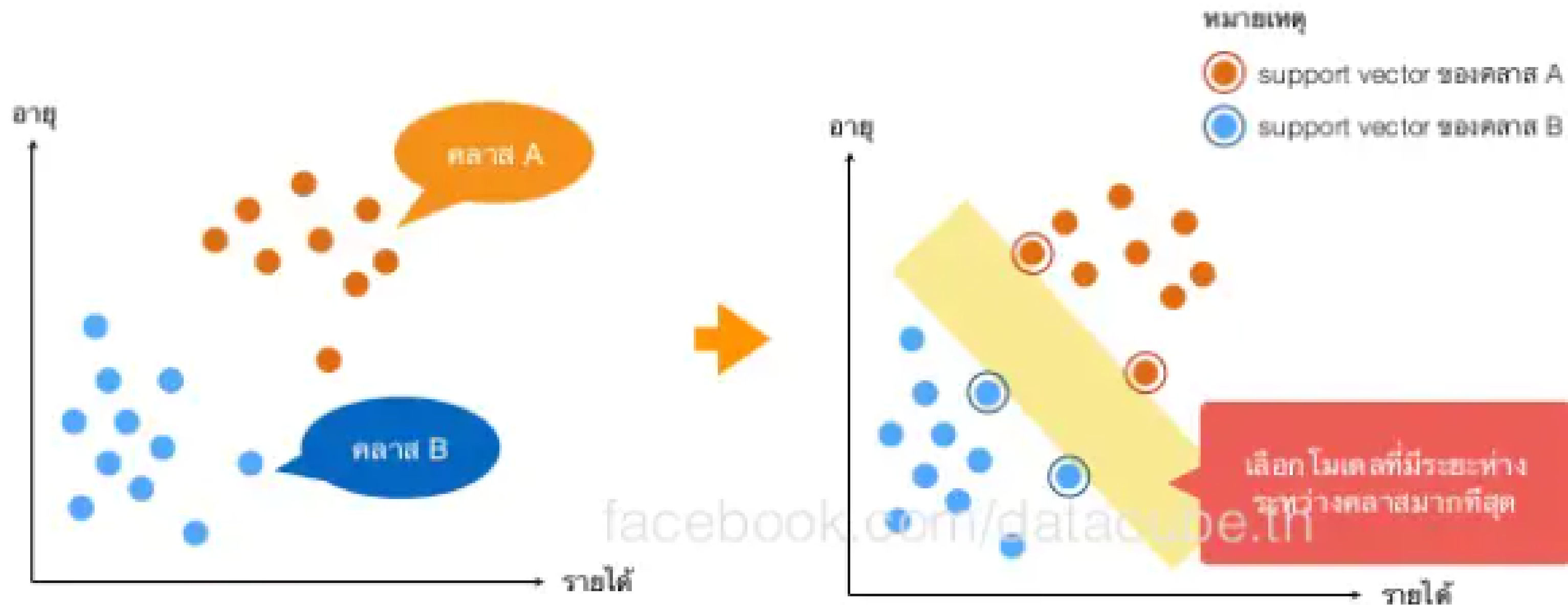
- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส





Support Vector Machines (SVM)

- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส



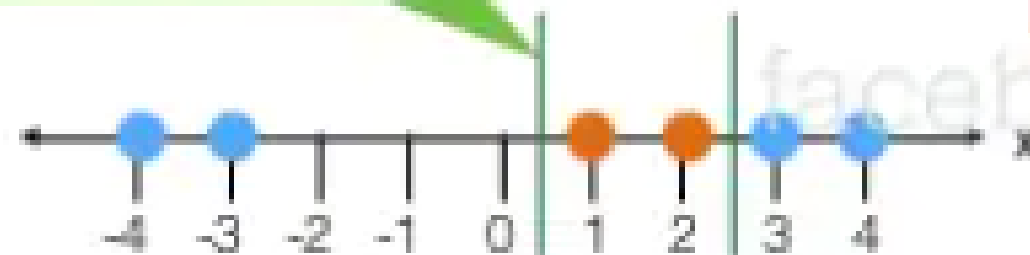
Support Vector Machines (SVM)

- SVM มีการใช้ kernel function เพื่อทำการแปลงข้อมูลที่ไม่สามารถใช้โมเดล linear ไปอยู่มิติ (dimension) ที่สูงขึ้น
- ใช้โมเดล linear แบ่งข้อมูลได้ง่ายขึ้น

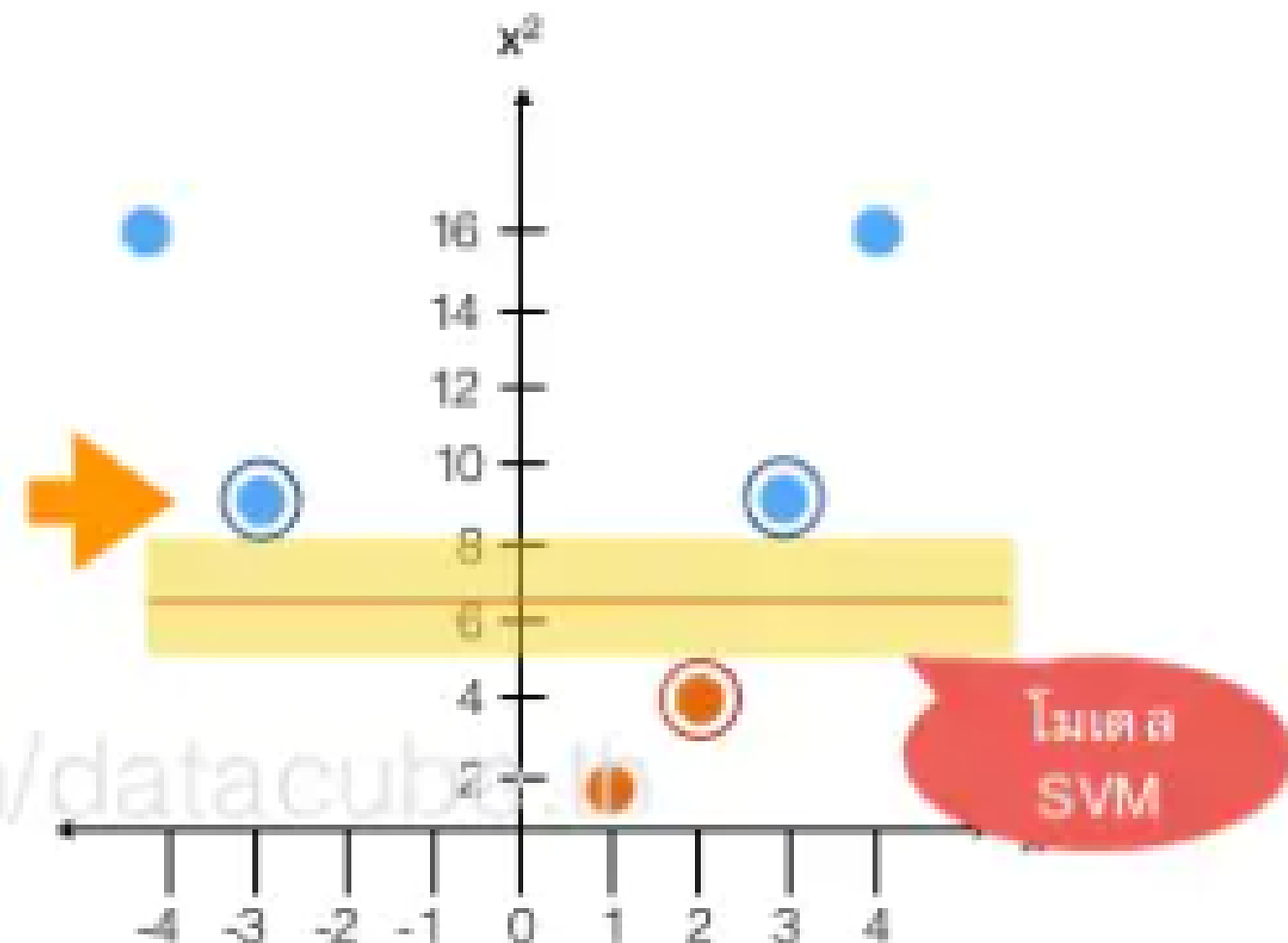
หมายเหตุ

- คลาส A
- คลาส B

IF $(x < 0.5 \parallel x > 2.5)$
THEN class = B



kernel
function



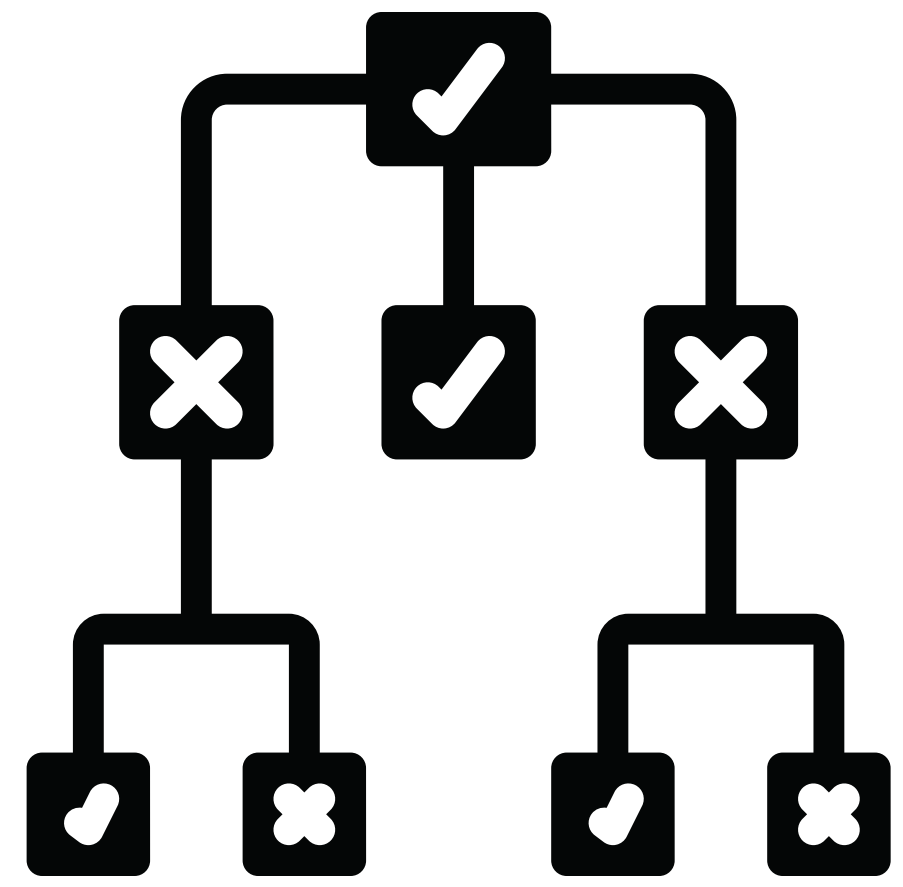


Classification Techniques

- Decision Tree
- Naive Bayes
- K-Nearest Neighbors (kNN)
- Linear Regression
- Neural Network
- Support Vector Machines
- Ensemble Classifiers (Vote)
- Attribute Selection
- Compare classification performance



Validation



rapidminer

Decision Tree



Decision Tree

ID	outlook	Temp	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	mild	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	mild	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	FALSE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no



Decision Tree

<new process*> – RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments More

Repository

- Import Data
- Book1 (8/24/21 11:25 PM – 4 k)
- CSV 1 (8/24/21 11:02 PM – 4 k)
- Sunny Play (8/24/21 11:03 PM)
- Sunny Play -V2 (8/31/21 11:23)
- Sunny Play -V3 (8/31/21 11:40)
- process
- TONG (Local)

Process

Process

inp

Retrieve Sunny Play ...

Decision Tree

tra mod
exa
wei

res
res



Decision Tree

<new process*> – RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Depl

Result History Tree (Decision Tree) X

Zoom

Graph

Tree

Node Labels

Edge Labels

Description

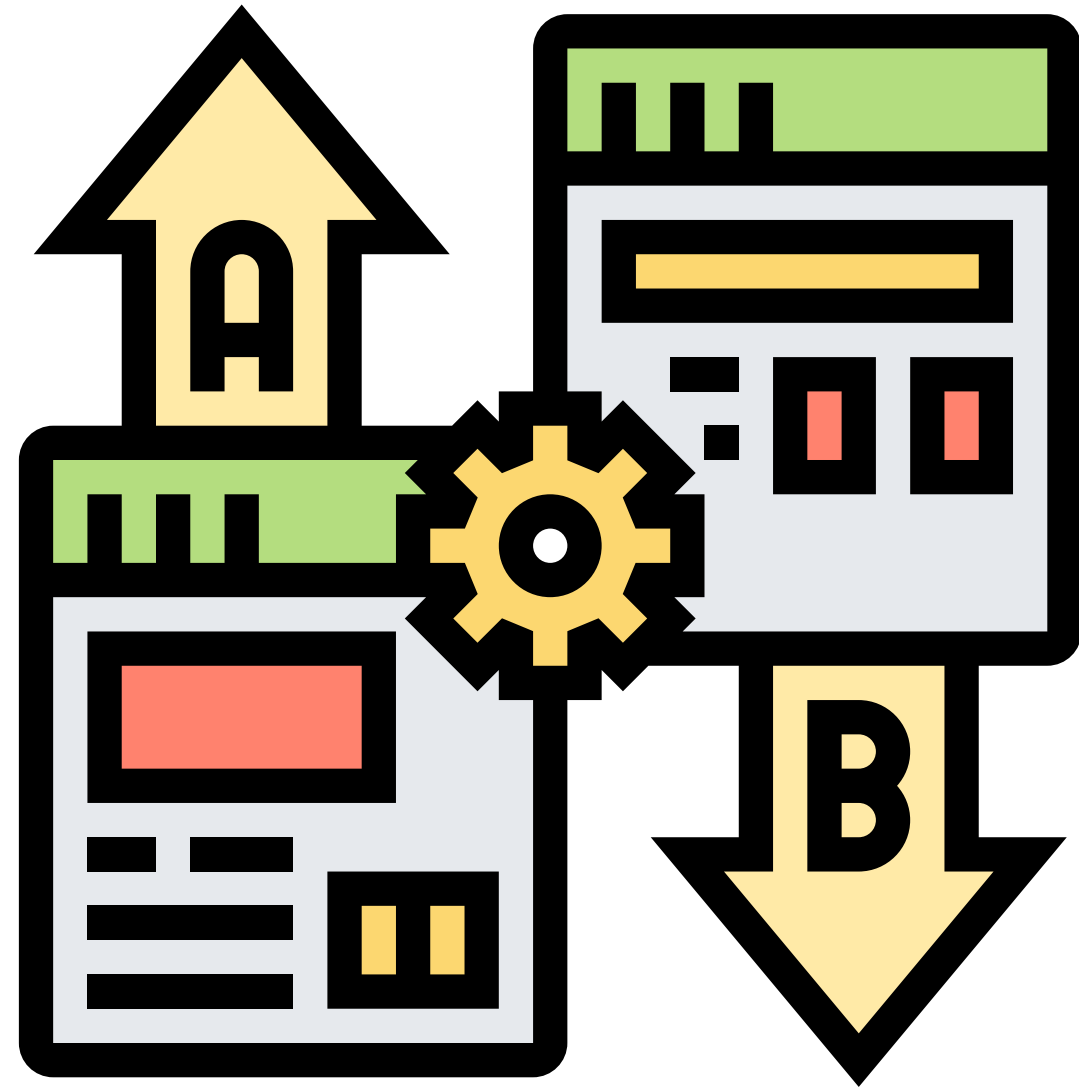
Annotations

```
graph TD; outlook[outlook] -- overcast --> yes1[yes]; outlook -- rainy --> windy[windy]; outlook -- sunny --> humidity[humidity]; windy -- FALSE --> yes2[yes]; windy -- TRUE --> no1[no]; humidity -- high --> no2[no]; humidity -- normal --> yes3[yes];
```

The diagram shows a decision tree with the root node 'outlook'. The 'outlook' node has three branches: 'overcast', 'rainy', and 'sunny'. The 'overcast' branch leads to a leaf node 'yes' with a red bar. The 'rainy' branch leads to a node 'windy', which has two branches: 'FALSE' leading to a leaf node 'yes' with a red bar, and 'TRUE' leading to a leaf node 'no' with a blue bar. The 'sunny' branch leads to a node 'humidity', which has two branches: 'high' leading to a leaf node 'no' with a blue bar, and 'normal' leading to a leaf node 'yes' with a red bar.



validation

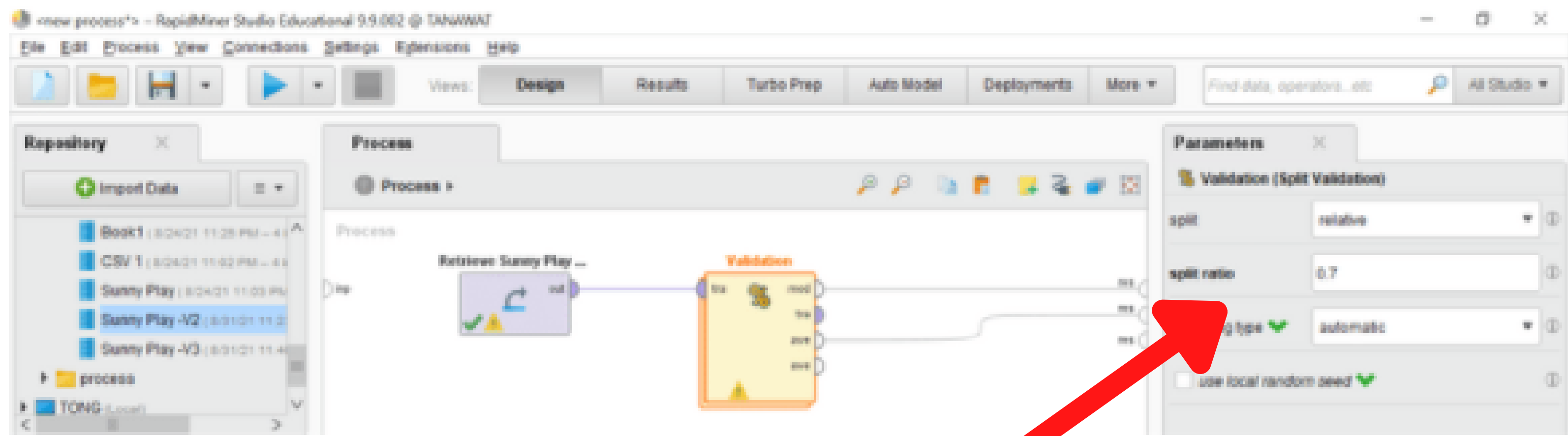


rapidminer

Self Consistency test | Split test | Cross-validation



Split test



70 % สำหรับสร้างโมเดล 30% ทดสอบโมเดล



Split test

The screenshot shows the RapidMiner Studio interface. In the 'Process' view, a 'Retrieve Sunny Play ...' operator is connected to a 'Validation' operator. A hand icon with the text 'CLICK' points to the 'Validation' operator. The 'Parameters' panel on the right shows the 'Validation (Split Validation)' settings. The 'split' parameter is set to 'relative', and the 'split ratio' is set to 0.7. A red arrow points to the 'split ratio' field.

70 % สำหรับสร้างโมเดล 30% ทดสอบโมเดล



Split test

The screenshot displays the RapidMiner Studio interface. The main workspace shows a process flow with a 'Validation' operator. The 'Parameters' panel on the right is open, showing the configuration for 'Validation (Split Validation)':

- split: relative
- split ratio: 0.7
- sampling type: automatic
- use local random seed: checked

The process flow is divided into 'Training' and 'Testing' sections. The 'Training' section contains a 'tra' operator. The 'Testing' section contains 'mod', 'tes', and 'thr' operators. The 'Parameters' panel also shows 'ave' and 'ave' operators.



Split test

The screenshot displays the RapidMiner Studio interface with a workflow designed for split testing. The workflow is divided into two phases: Training and Testing.

- Training Phase:** A **Decision Tree** operator is connected to the process. Its output port is labeled 'tra'.
- Testing Phase:** The output from the Decision Tree is fed into an **Apply Model** operator. The output of the Apply Model operator is then fed into a **Performance** operator.

The **Performance** operator is configured for **Validation (Split Validation)**. The parameters for this operator are:

- split:** relative
- split ratio:** 0.7
- sampling type:** automatic
- use local random seed:** checked

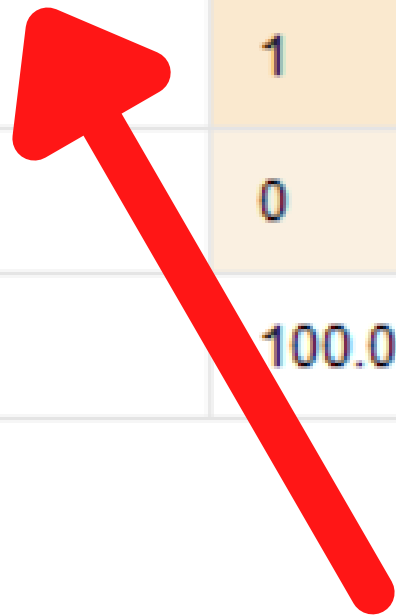
The interface also shows a **Repository** on the left with several data sources, including 'Sunny Play -V2'. The **Parameters** panel on the right provides a detailed view of the configuration for the selected operator.



Split test

accuracy: 25.00%

	true no	true yes	class precision
pred. no	1	3	25.00%
pred. yes	0	0	0.00%
class recall	100.00%	0.00%	



ถูกต้อง 25%



Cross-validation

The screenshot displays the RapidMiner Studio Educational 9.9.002 interface. The main workspace shows a process flow starting with 'Retrieve Sunny Play ...' (with a warning icon), followed by a 'Validation' operator, and then a 'Cross Validation' operator (highlighted with an orange box and a green checkmark). The 'Cross Validation' operator is connected to the 'Validation' operator. The 'Parameters' panel on the right is open for the 'Cross Validation' operator, showing the following settings:

- split on batch attribute
- leave one out
- number of folds 10
- sampling type automatic
- use local random seed
- enable parallel execution

The 'Repository' panel on the left shows a list of data sources, including 'Sunny Play -V2' and 'Sunny Play -V3'. The 'Operators' panel at the bottom left shows a search for 'perf' and a list of operator categories: Validation (20), Performance (18), and Predictive (7).



Cross-validation

The screenshot displays the RapidMiner Studio interface for a cross-validation process. The main workspace is divided into Training and Testing sections. In the Training section, a 'Decision Tree (2)' operator is connected to a 'tra' input. In the Testing section, the 'Decision Tree (2)' operator is connected to an 'Apply Model (2)' operator, which is then connected to a 'Performance (2)' operator. The 'Performance (2)' operator is configured for cross-validation. The Parameters panel on the right shows the following settings for the 'Cross Validation' operator:

- split on batch attribute
- leave one out
- number of folds: 10
- sampling type: automatic
- use local random seed
- enable parallel execution



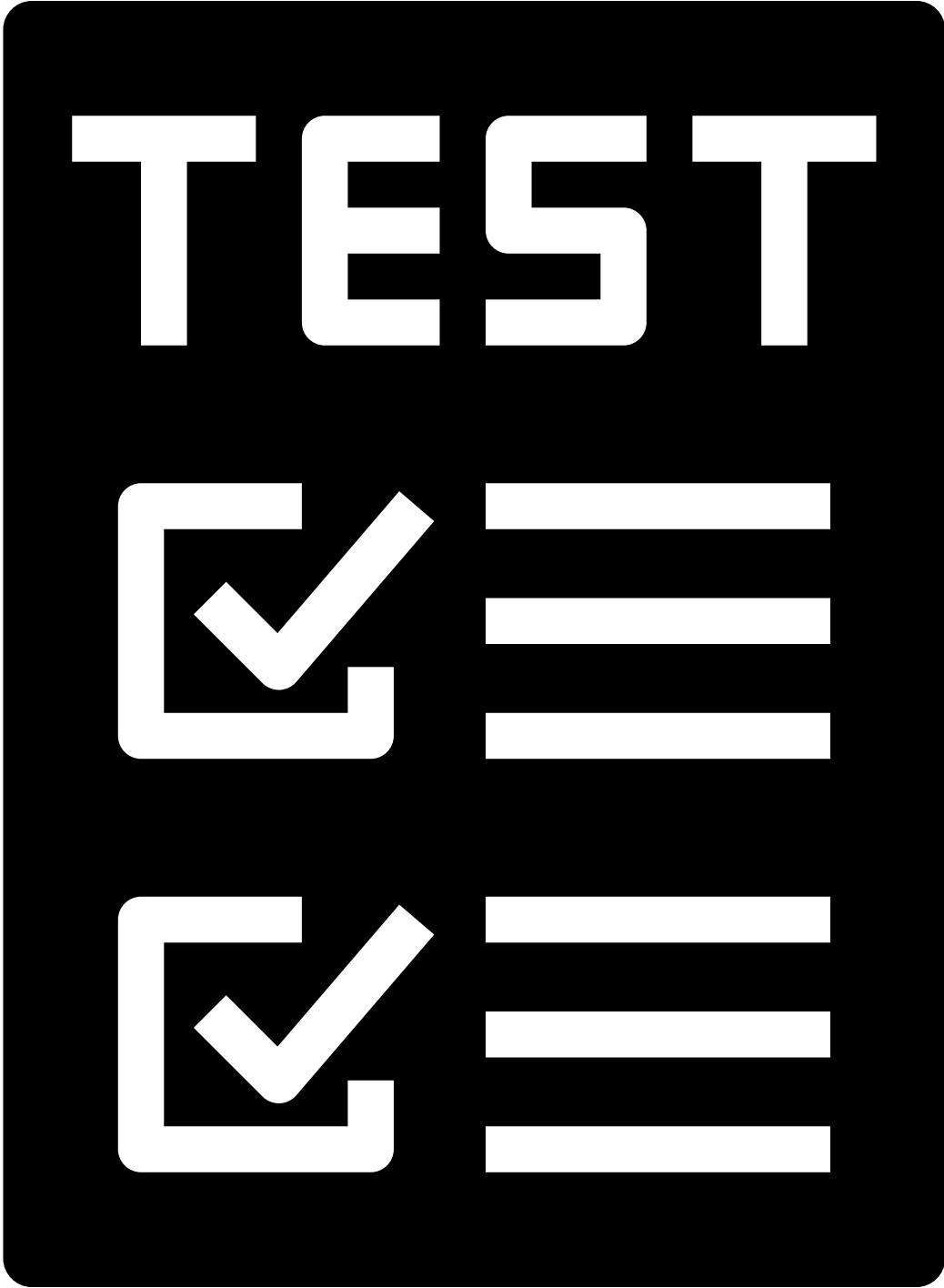
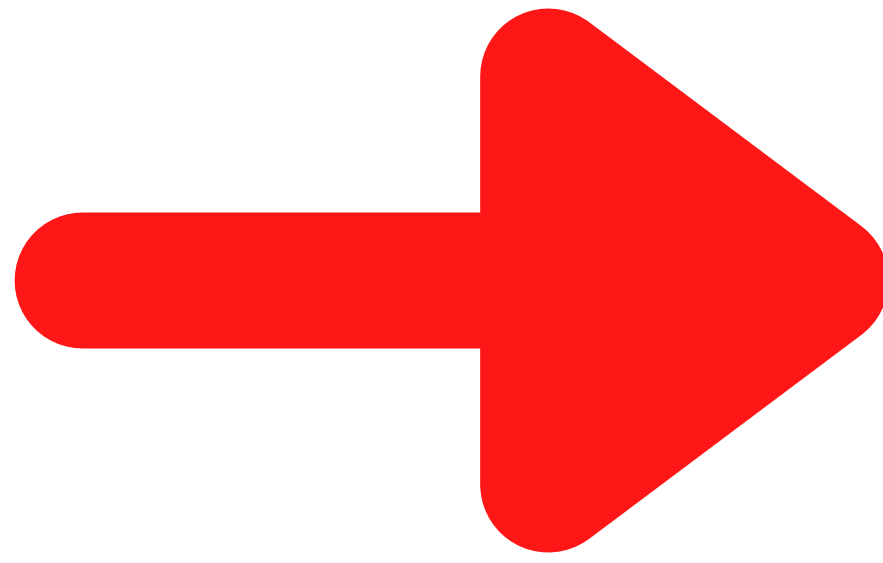
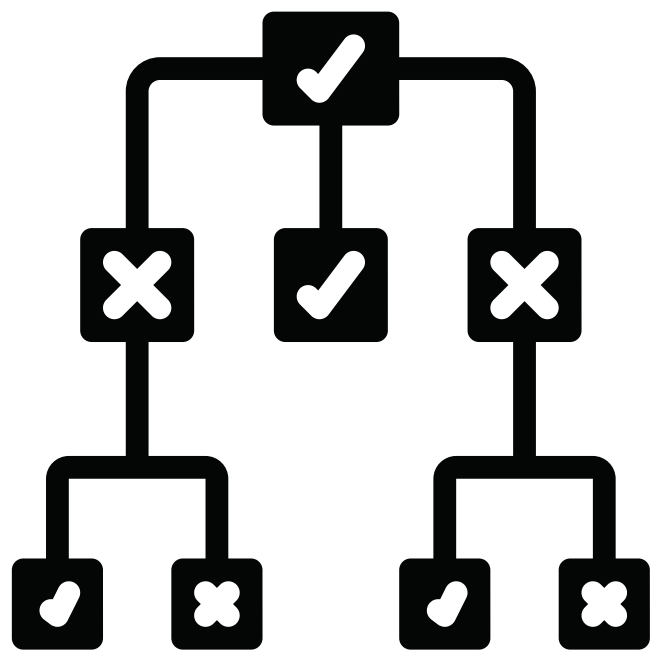
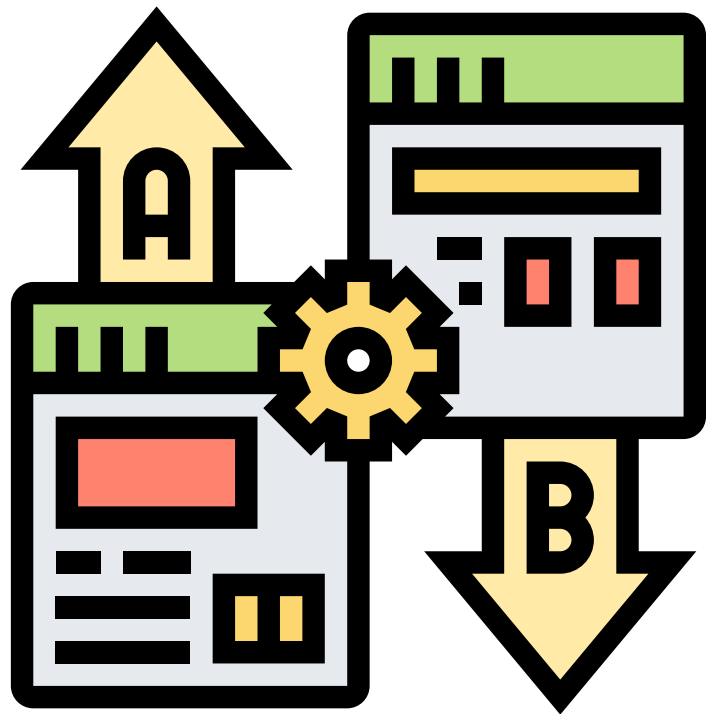
Cross-validation

accuracy: 40.00% +/- 45.95% (micro average: 42.86%)

	true no	true yes	class precision
pred. no	2	5	28.57%
pred. yes	3	4	57.14%
class recall	40.00%	44.44%	



นำข้อมูลมาทดลองจริง





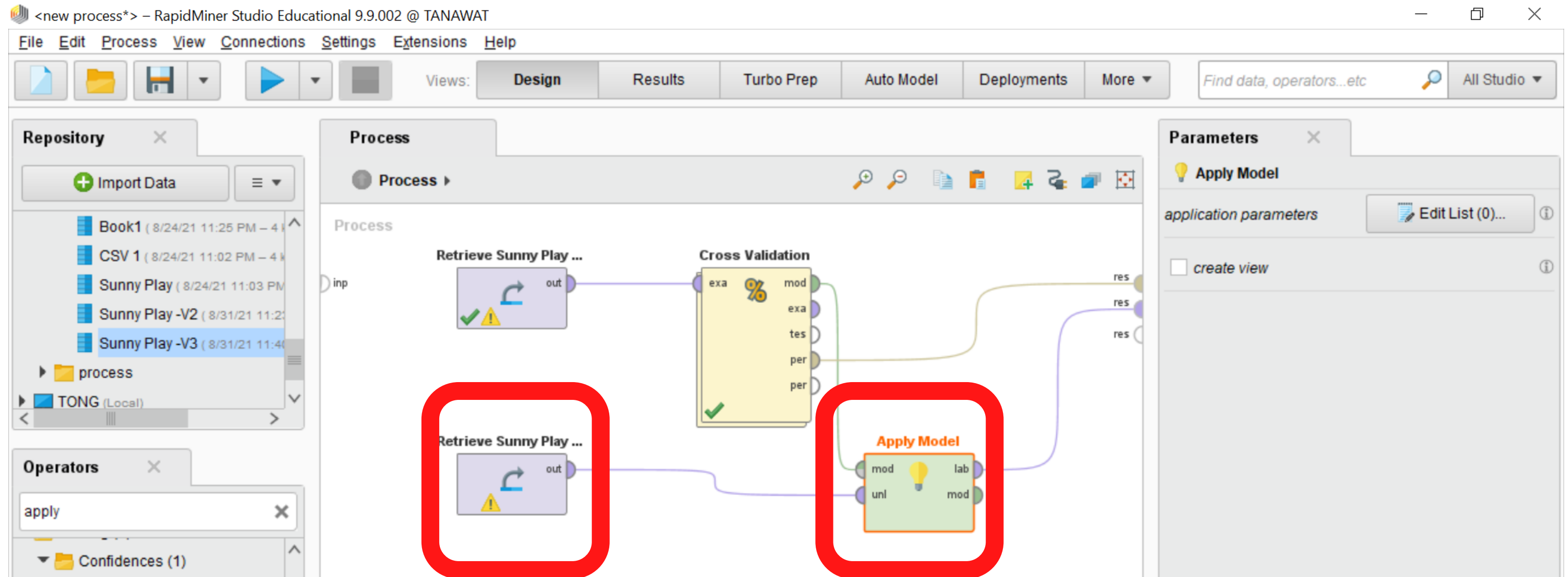
นำข้อมูลมาทดลองจริง

The screenshot shows the RapidMiner Studio interface with the following components:

- Repository:** Lists data sources including 'Book1', 'CSV 1', and three 'Sunny Play' datasets (V1, V2, V3). A 'process' folder and 'TONG (Local)' are also visible.
- Process Design:** A workflow starting with an 'inp' connector, followed by two 'Retrieve Sunny Play ...' operators. The top one connects to a 'Cross Validation' operator, which then connects to an 'Apply Model' operator. The bottom 'Retrieve Sunny Play ...' operator also connects to the 'Apply Model' operator. The 'Apply Model' operator has three output ports labeled 'res'.
- Parameters Panel:** Shows the 'Apply Model' operator's configuration, including 'application parameters' and a 'create view' checkbox.
- Operators Panel:** A search bar containing 'apply' and a list of operators under 'Confidences (1)'.



นำข้อมูลมาทดลองจริง



ข้อมูลที่ต้องการทดสอบ

ประยุกต์โมเดล



นำข้อมูลมาทดลองจริง

ExampleSet (Apply Model) PerformanceVector (Performance (2))

Open in Turbo Prep Auto Model Filter (14 / 14 examples): all

Row No. ↑	ID	play	prediction(pl...	confidence(...	confidence(...	outlook	Temp	humidity
1	1	?	yes	0	1	rainy	mild	high
2	2	?	yes	0	1	rainy	cool	normal
3	3	?	no	1	0	sunny	mild	high
4	4	?	yes	0	1	sunny	mild	normal
5	5	?	yes	0	1	rainy	hot	normal
6	6	?	yes	0	1	sunny	hot	normal
7	7	?	yes	0	1	rainy	hot	normal
8	8	?	no	1	0	sunny	hot	high
9	9	?	yes	0	1	sunny	hot	normal
10	10	?	yes	0	1	rainy	hot	normal
11	11	?	yes	0	1	sunny	hot	normal
12	12	?	yes	0	1	overcast	hot	high
13	13	?	yes	0	1	overcast	hot	normal



เปลี่ยนใช้ Model [Naive]

<new process*> – RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments More

Repository

- Import Data
- 7. Asso Rule (8/25/21 1:21 PM – 2 kB)
- 8. Group (8/25/21 1:49 PM – 1 kB)
- 9. Big Data - SMS PHONE (8/25/21 1:53 PM – 3 kB)
- 9. Sunny Play - Model (9/1/21 10:31 AM – 3 kB)
- 10. Sunny Play - TEST (9/1/21 10:40 AM – 3 kB)
- Process
- DB (Legacy)

Process

Process > Validation

Training

Naive Bayes

Testing

Apply Model

Performance

The screenshot shows a workflow in RapidMiner Studio. The 'Process' view is active, showing a 'Validation' process. The workflow is divided into 'Training' and 'Testing' phases. In the Training phase, a 'Naive Bayes' operator (green box with a lightbulb icon) is connected to a 'tra' port. In the Testing phase, an 'Apply Model' operator (green box with a lightbulb icon and a checkmark) is connected to 'mod' and 'thr' ports. The 'Apply Model' operator is connected to a 'Performance' operator (yellow box with a percentage sign and a checkmark). The 'Performance' operator is connected to 'lab' and 'per' ports. The 'Performance' operator has two output ports labeled 'ave'.



เปลี่ยนใช้ Model [Naive]

<new process*> – RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments More

ExampleSet (Apply Model (2)) PerformanceVector (Performance)

Result History SimpleDistribution (Naive Bayes)

Performance

Criterion

- accuracy
- precision
- recall
- AUC (optimistic)
- AUC
- AUC (pessimistic)

Description

Table View Plot View

accuracy: 71.43%

	true no	true yes	class precision
pred. no	1	1	50.00%
pred. yes	1	4	80.00%
class recall	50.00%	80.00%	



เปลี่ยนใช้ Model [KNN]

<new process*> – RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments More knn

Repository

- Import Data
- 7. Asso Rule (8/25/21 1:21 PM – 2 kB)
- 8. Group (8/25/21 1:49 PM – 1 kB)
- 9. Big Data - SMS PHONE (8/25/21 1:53 PM – 1 kB)
- 9. Sunny Play - Model (9/1/21 10:31 AM – 3 kB)
- 10. Sunny Play - TEST (9/1/21 10:40 AM – 3 kB)
- Process
- DB (Legacy)

Process

Process > Validation

Training

Testing

k-NN

Apply Model

Performance



เปลี่ยนใช้ Model [KNN]

<new process*> – RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments More knn

Result History KNNClassification (k-NN) ExampleSet (Apply Model (2)) PerformanceVector (Performance)

Criterion

- accuracy
- precision
- recall
- AUC (optimistic)
- AUC
- AUC (pessimistic)

Table View Plot View

accuracy: 57.14%

	true no	true yes	class precision
pred. no	0	1	0.00%
pred. yes	2	4	66.67%
class recall	0.00%	80.00%	

Performance

Description



เปลี่ยนใช้ Model [Network]

<new process*> - RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments More network

Repository

- Import Data
- 7. Asso Rule (8/25/21 1:21 PM - 2 kB)
- 8. Group (8/25/21 1:49 PM - 1 kB)
- 9. Big Data - SMS PHONE (8/25/21 1:53 PM - 1 kB)
- 9. Sunny Play - Model (9/1/21 10:31 AM - 3 kB)
- 10. Sunny Play - TEST (9/1/21 10:40 AM - 3 kB)
- Process
- DB (Legacy)

Process

Process > Validation

Training

tra -> **Deep Learning** (tra, mod, exa, wei) [✓]

Testing

mod, thr -> **Apply Model** (mod, lab, uni, mod) [✓] -> **Performance** (lab, per, per, exa) [✓] -> ave



เปลี่ยนใช้ Model [Network]

<new process*> – RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments More network

ExampleSet (Apply Model (2)) PerformanceVector (Performance)

Result History Deep Learning Model (Deep Learning)

Performance

Criterion

- accuracy
- precision
- recall
- AUC (optimistic)
- AUC
- AUC (pessimistic)

Description

Table View Plot View

accuracy: 57.14%

	true no	true yes	class precision
pred. no	2	3	40.00%
pred. yes	0	2	100.00%
class recall	100.00%	40.00%	



เปลี่ยนใช้ Model [SVM]

<new process*> – RapidMiner Studio Educational 9.9.002 @ TANAWAT

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments More svm

Repository

- 7. Asso Rule (8/25/21 1:21 PM – 2 kB)
- 8. Group (8/25/21 1:49 PM – 1 kB)
- 9. Big Data - SMS PHONE (8/25/21 1:53 PM – 1 kB)
- 9. Sunny Play - Model (9/1/21 10:31 AM – 3 kB)
- 10. Sunny Play - TEST (9/1/21 10:40 AM – 3 kB)

Process

Process > Validation

Training

tra → **SVM** (mod, est, wei, exa)

Testing

mod, thr → **Apply Model** (mod, lab, unl, mod) → **Performance** (lab, per, % per, exa) → ave, ave



เปลี่ยน Cross Validation

The screenshot shows the RapidMiner Studio Educational 9.9.002 interface. The main workspace displays a workflow with the following components:

- Retrieve 9. Sunny PI...:** A data retrieval operator.
- Cross Validation:** A central operator highlighted with an orange border, used for model validation. It has ports for 'exa' (examples), 'mod' (model), 'exa' (examples), 'tes' (test), 'per' (percentage), and 'per' (percentage).
- Retrieve 10. Sunny P...:** A second data retrieval operator.
- Apply Model (2):** An operator that applies the model to the data. It has ports for 'mod' (model), 'lab' (label), 'unl' (unlabeled), and 'mod' (model).

The **Parameters** panel on the right is open for the **Cross Validation** operator, showing the following settings:

- split on batch attribute
- leave one out
- number of folds 14
- sampling type all
- use local random seed
- enable parallel execution

The **Repository** panel on the left shows the **models** folder expanded, with **SVM Model Plain** selected. The **Operators** panel at the bottom shows a search for **svm**, with **Weight by SVM** under **Feature Weights (1)** visible.



ทำซ้ำๆ ด้วยกัน