LOB2203
ระบบเทคโนโลยีสารสนเทศด้านโลจิสติกส์สำหรับธุรกิจออนไลน์
CH 7

rapidminer

สาขาการจัดการโลจิสติกสำหรับธุรกิจออนไลน์
วิทยาลัยโลจิสติกส์และซัพพลายเชน มหาวิทยาลัยราชภัฏสวนสุนันทา

Chapter 7

Classification

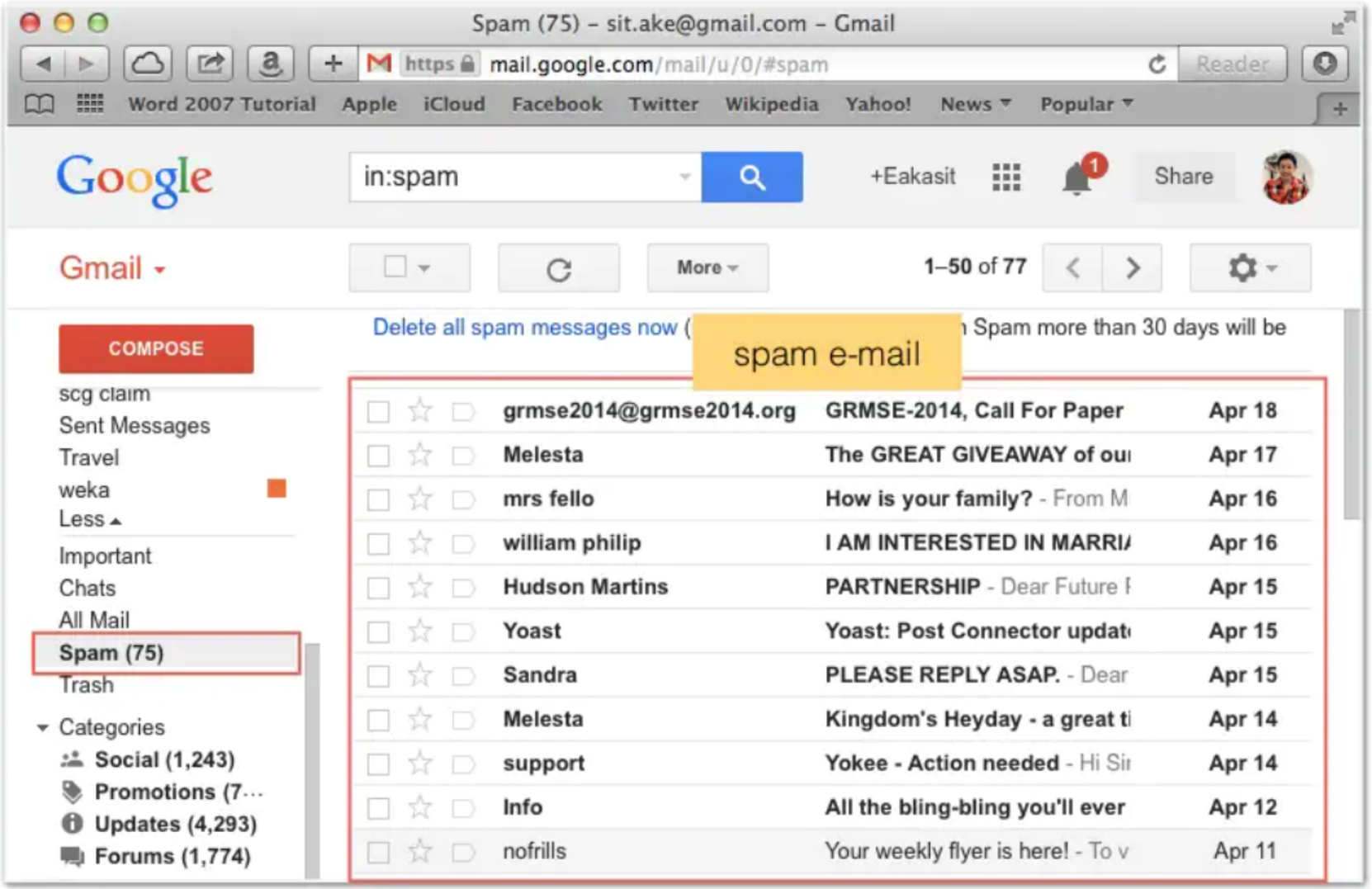
การพยากรณ์อากาศ

Speecg recognition

face recognition

Classification in daily life

- spam e-mail



Classification example

- ตัวอย่าง spam e-mail classification
  - ระบุว่า e-mail ไหนบ้างที่เป็น 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 ที่ใช้บ่งบอกว่าเป็น 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

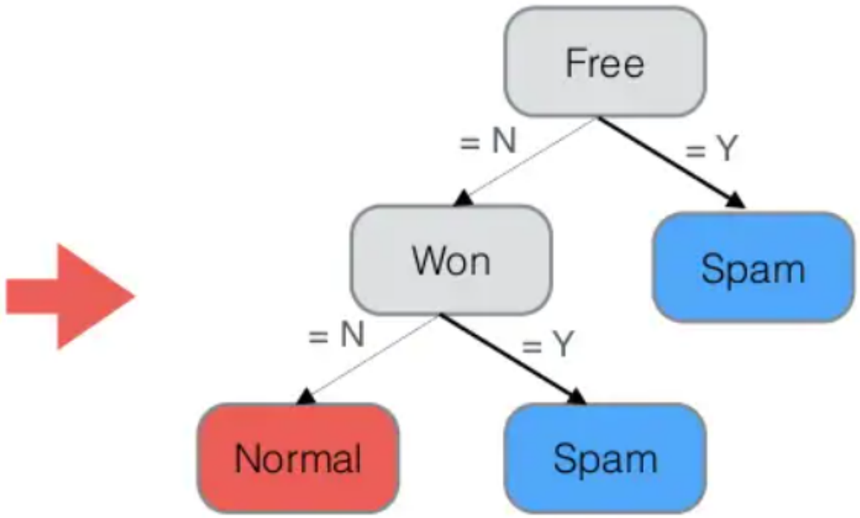
	keywords			
ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal

Classification example

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

attribute → | label →

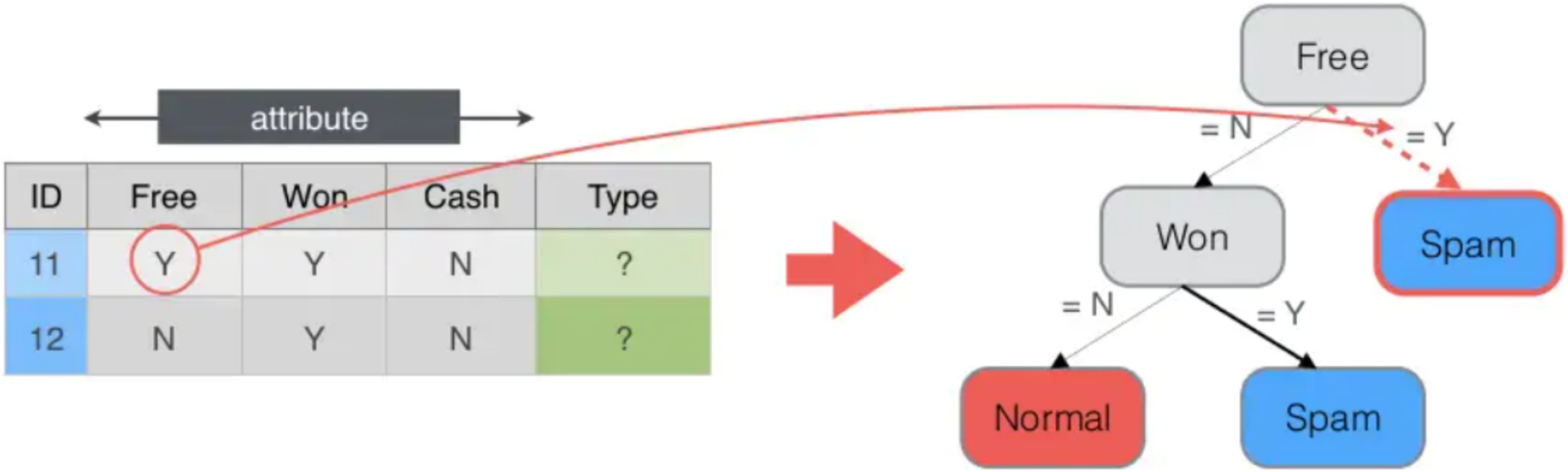
ID	Free	Won	Cash	Type
1	Y	Y	Y	spam
2	N	Y	Y	spam
3	N	N	N	normal
4	N	N	N	normal
5	Y	N	N	spam
6	Y	N	N	spam
7	N	N	N	normal
8	N	Y	N	spam
9	N	N	N	normal
10	N	N	N	normal



classification model

Classification example

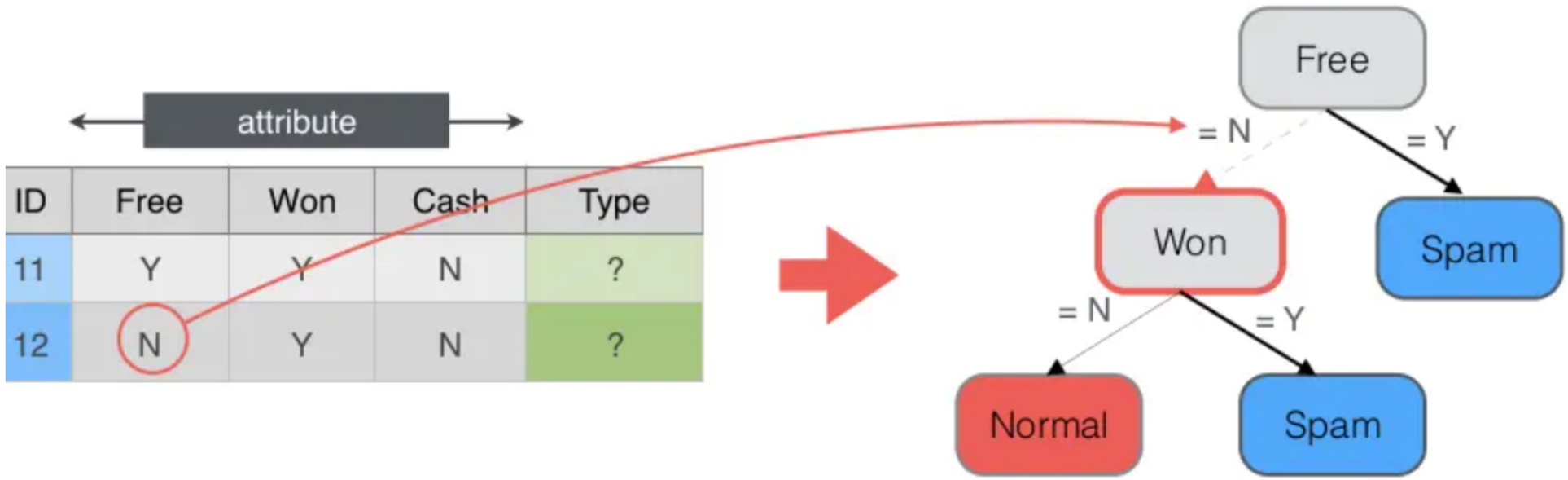
- ตัวอย่าง spam e-mail classification
  - นำข้อมูลใหม่ (unseen data) ทำนายโดยใช้โมเดล



Classification example

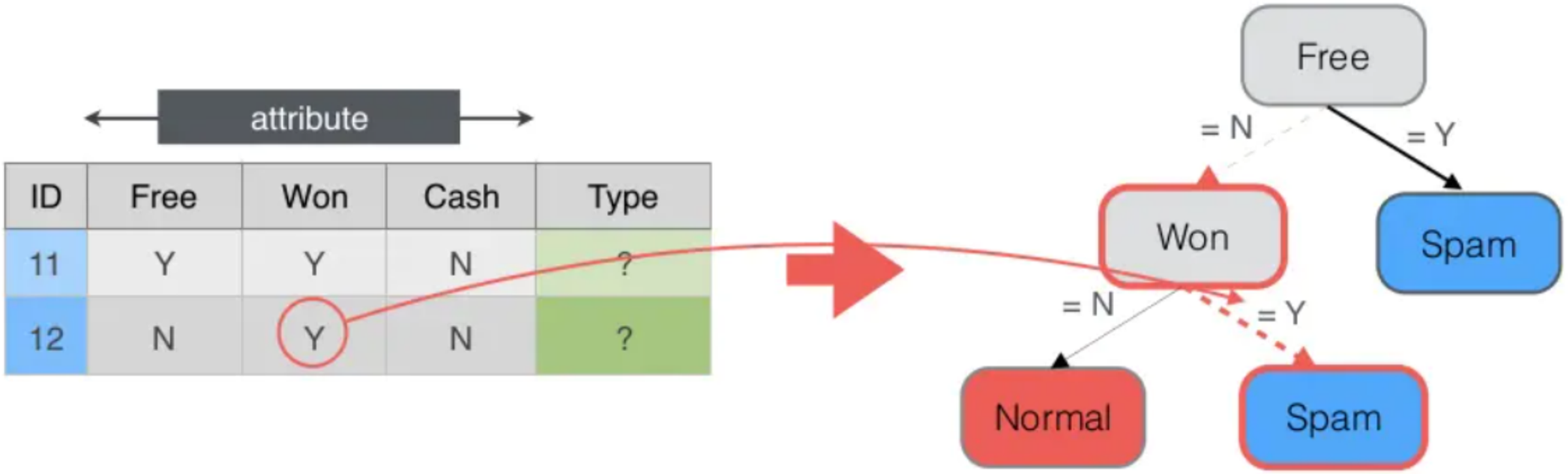
ตัวอย่าง spam e-mail classification

- นำข้อมูลใหม่ (unseen data) ทำนายโดยใช้โมเดล



Classification example

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



Classification example

- ตัวอย่าง spam e-mail classification

ID ← 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

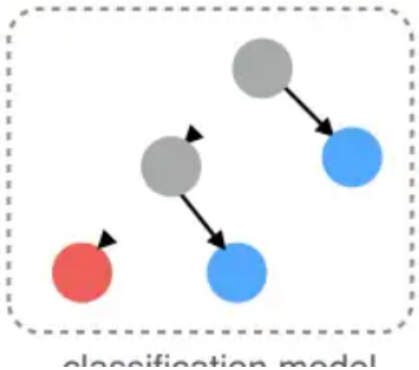
training data

1 → สร้าง classification model

2

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

unseen data

3 → 

4 →

ID	Type
11	spam
12	spam

Classification & Regression

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

Classification & Regression task

- การสร้างโมเดลและการทดสอบประสิทธิภาพ

training data

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

1

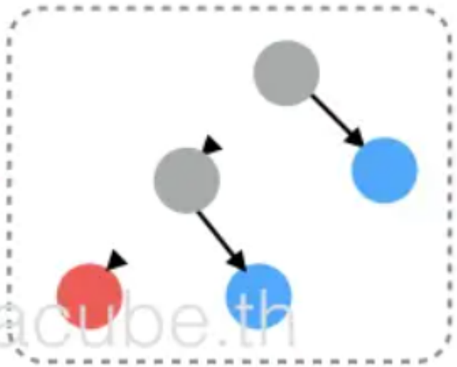
สร้าง classification model

2

testing data

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

3



classification model

4

ID	Type	Predicted
1	spam	spam
2	spam	spam
3	normal	normal

prediction results

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

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

- 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 x 100 = 57.12%
- Precision สำหรับ spam
  - 6/8 x 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

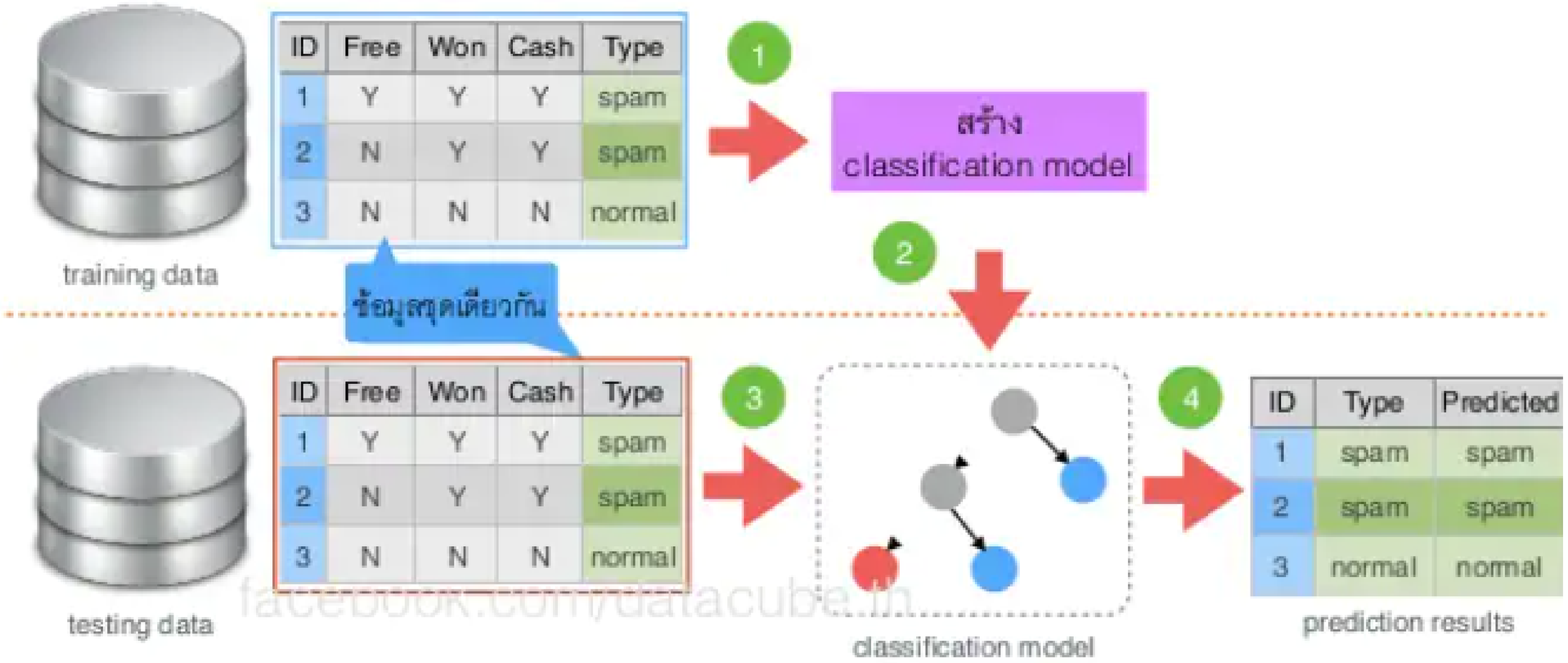
- Recall
  - จำนวนข้อมูลที่ทำนายถูก
- Recall สำหรับ normal
  - $$\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$
  - 4/6 x 100 = 66.67%
- Recall สำหรับ spam
  - 7/9 x 100 = 77.78%

Validation

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

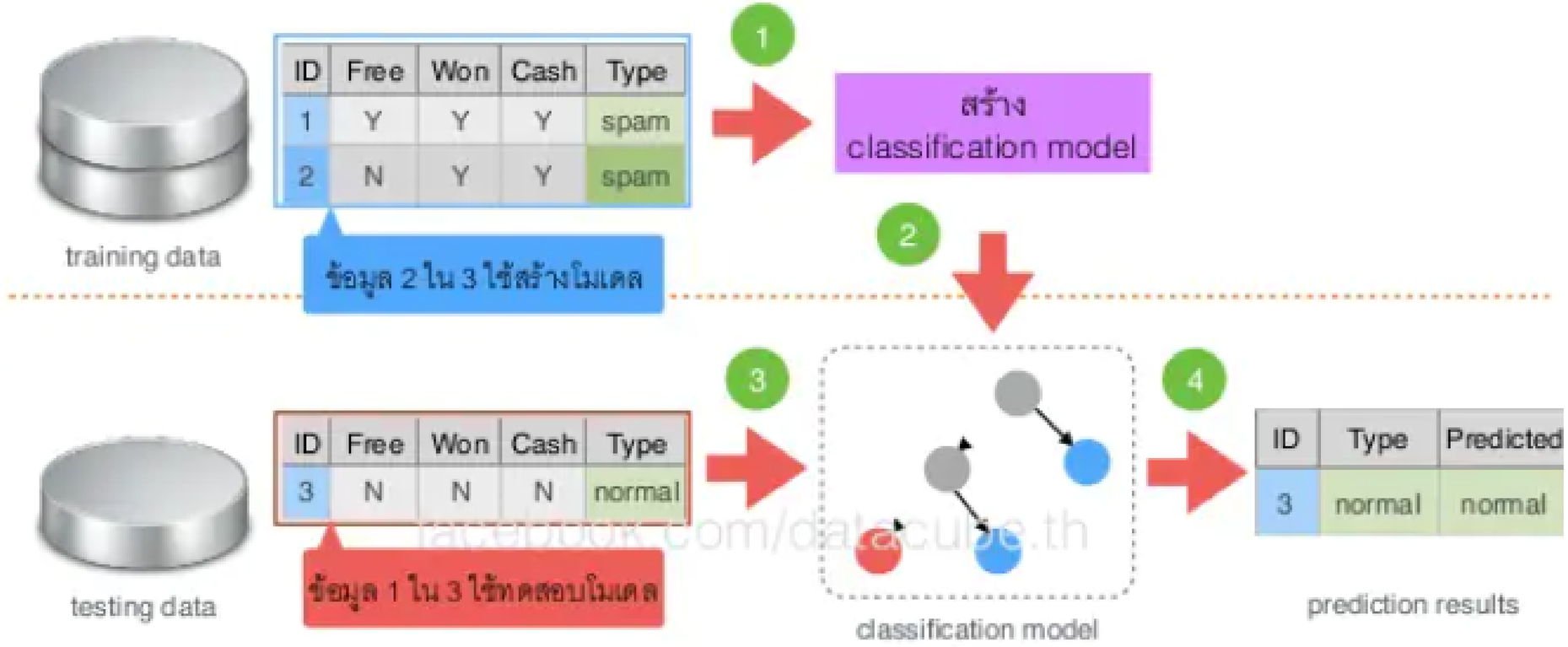
Self Consistency test

- ใช้ข้อมูล training ในการทดสอบประสิทธิภาพของโมเดล



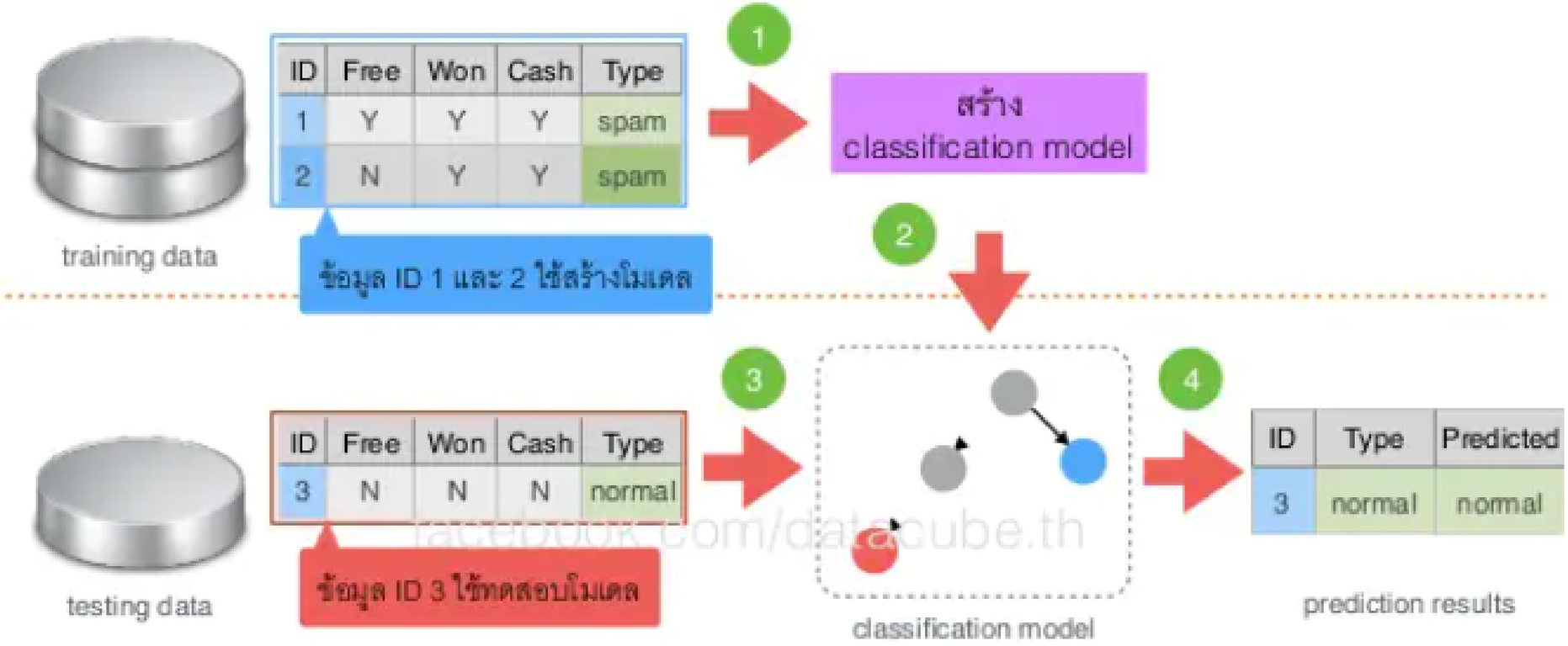
Split test

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



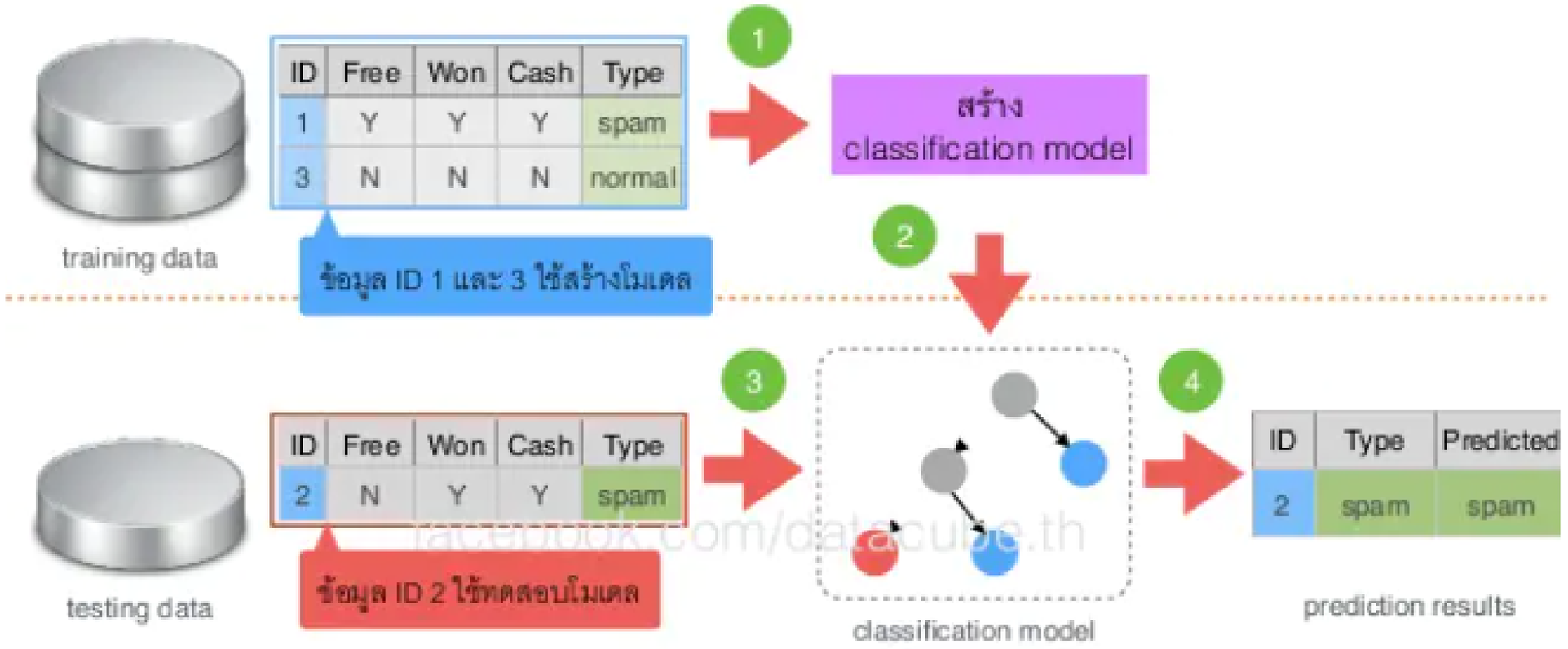
Cross-validation

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



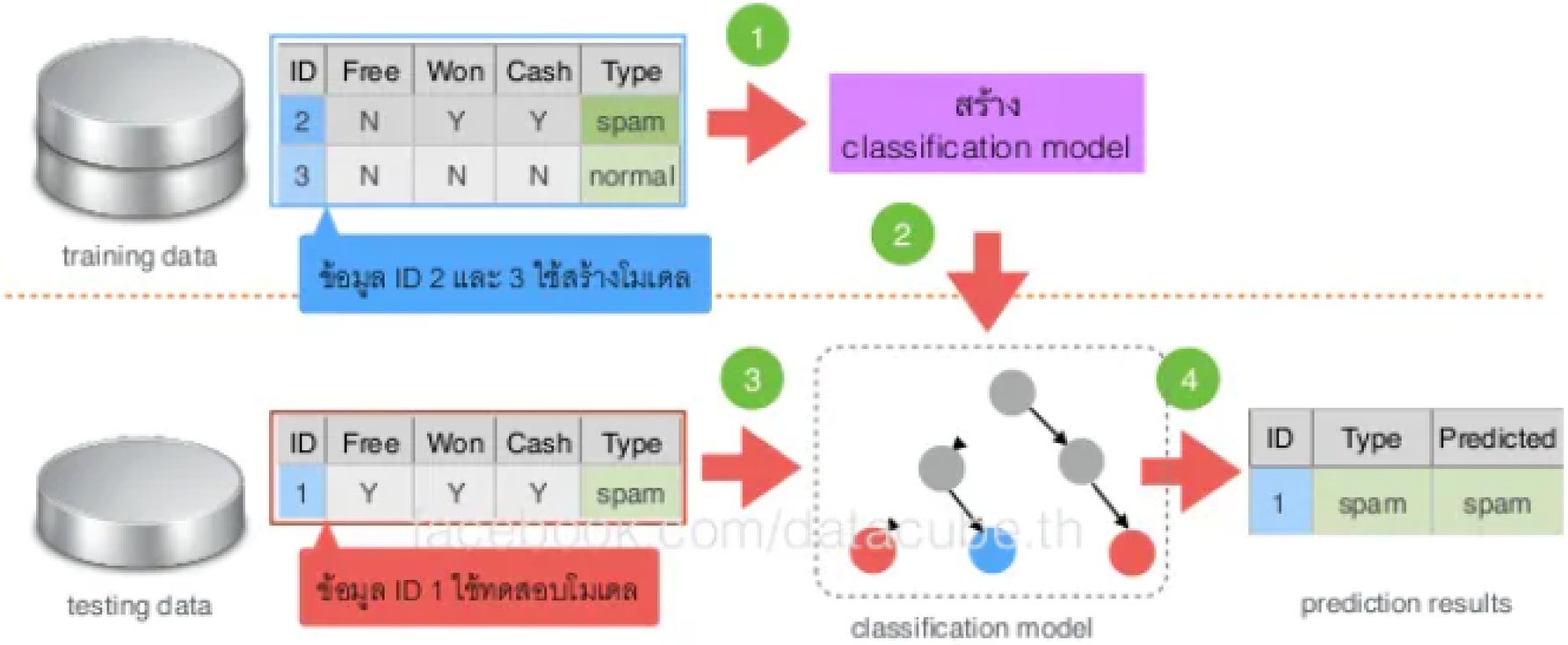
Cross-validation

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



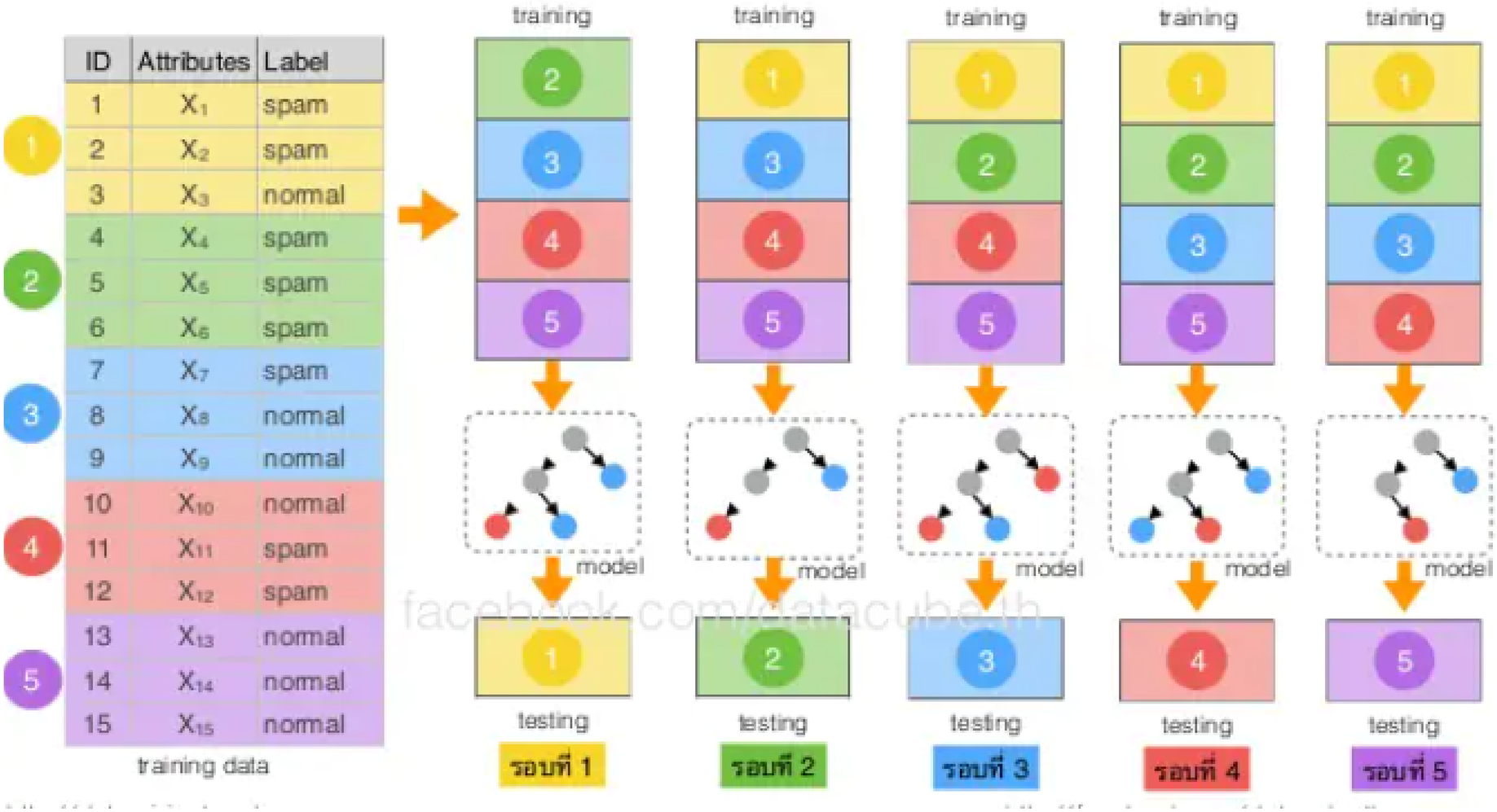
Cross-validation

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



Cross-validation

- ตัวอย่างของ 5-fold cross-validation



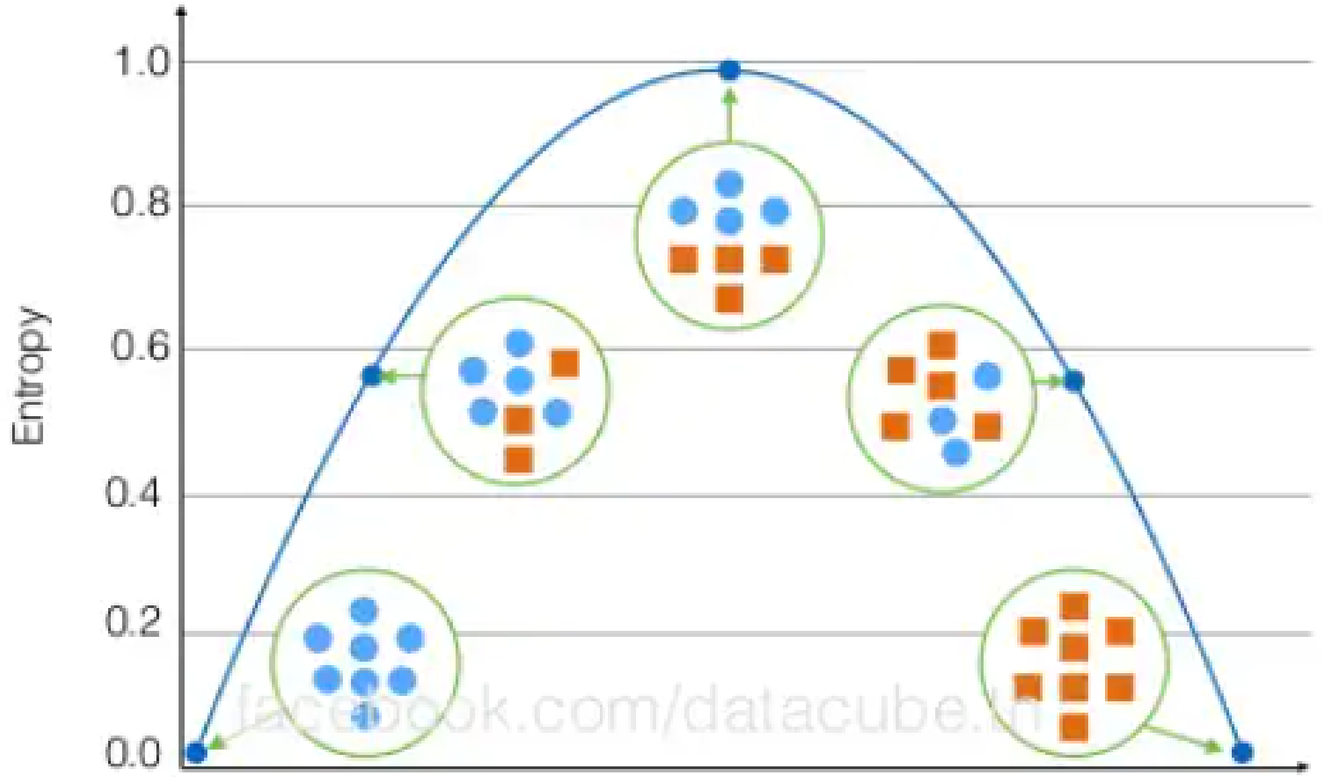
Classification Techniques

- 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

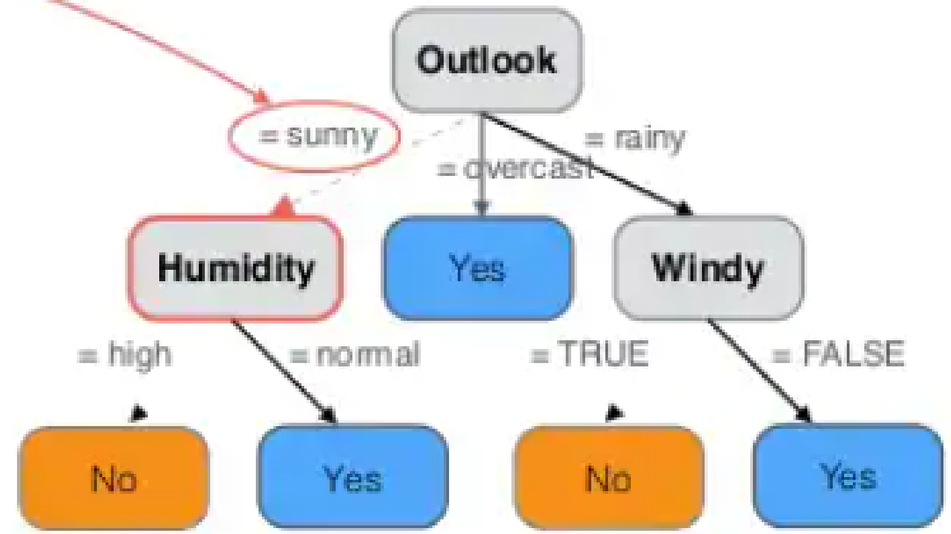
ID	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	mild	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	mild	normal	FALSE	yes
10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

ข้อมูลที่ใช้ทดสอบ



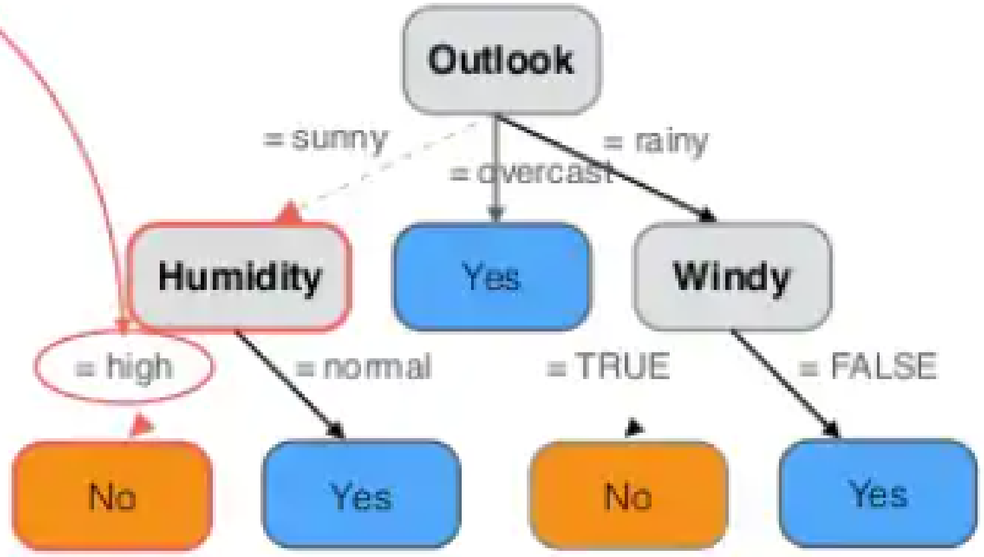
โมเดล decision tree

Decision Tree

- การใช้โมเดล predict ข้อมูลใหม่

ID	Outlook	Temperature	Humidity	Windy
1	sunny	hot	high	FALSE

ข้อมูลที่ใช้ทดสอบ



โมเดล decision tree

Decision Tree

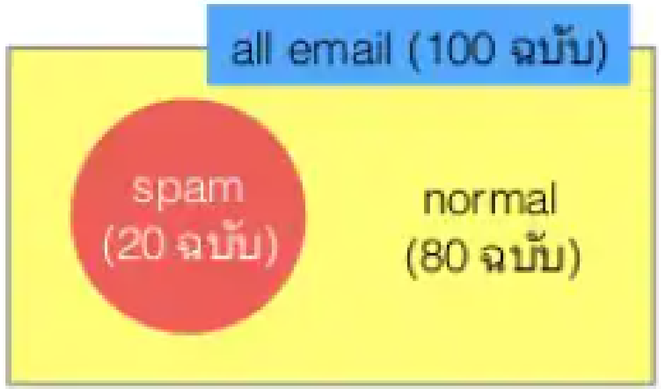
- ข้อมูลเป็นตัวเลข
  - เรียงลำดับข้อมูลที่เป็นตัวเลขจากน้อยไปมาก
  - แบ่งข้อมูลออกเป็น 2 ส่วนโดยการหาจุดกึ่งกลางระหว่างค่าตัวเลข 2 ค่า
  - คำนวณค่า Information Gain จากข้อมูล 2 ส่วนที่แบ่งได้
  - เลือกจุดกึ่งกลางที่ให้ค่า Information Gain สูงที่สุดมาใช้งานต่อ

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

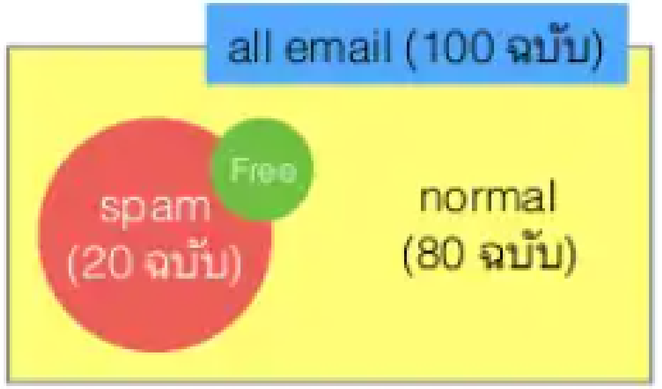
Classification Techniques

- 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 emai จะเป็น spam มีความน่าจะเป็น 20/100 = 0.2 หรือ P(spam) = 0.2
    - โอกาสที่ emai จะเป็น normal มีความน่าจะเป็น 80/100 = 0.8 หรือ P(normal) = 0.8



Probability

- Joint Probability
  - ความน่าจะเป็นที่ 2 เหตุการณ์เกิดร่วมกัน
  - ความน่าจะเป็นที่มีคำว่า Free อยู่ใน spam email
  - สัญลักษณ์ P(Free=Y ∩ spam)



Naive Bayes

- ใช้หลักการของความน่าจะเป็น (probability)

ความน่าจะเป็นที่ B เกิดก่อนและ A เกิดตามมา

ความน่าจะเป็นที่ A และ B เกิดร่วมกัน

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

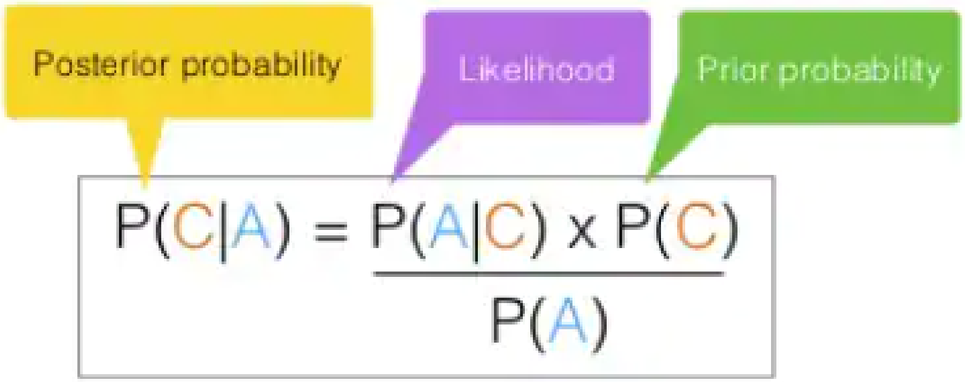
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

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

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

Bayes Theorem

Probability



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

Naive Bayes

- การใช้โมเดลเพื่อ predict ข้อมูลใหม่

P(**Type** = normal) = 5/10 = 0.50

P(**Type** = 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

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

P(**Type** = normal|A) = P(**Free** = Y|**Type** = normal) x
P(**Won** = Y|**Type** = normal) x
P(**Cash** = Y|**Type** = normal) x
P(**Type** = normal)
= 0.00 x 0.00 x 0.00 x 0.50
= 0.00

P(**Type** = spam|A) = P(**Free** = Y|**Type** = spam) x
P(**Won** = Y|**Type** = spam) x
P(**Cash** = Y|**Type** = spam) x
P(**Type** = spam)
= 0.60 x 0.60 x 0.40 x 0.50
= 0.07

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)

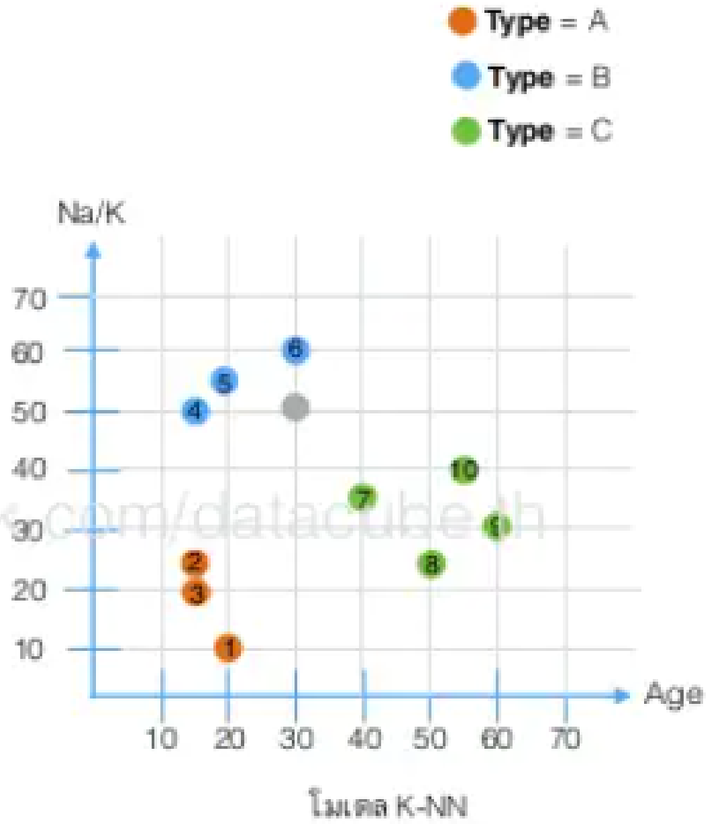
- การใช้โมเดลเพื่อ predict ข้อมูลใหม่

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	$\sqrt{(40-50)^2 + (55-30)^2} = 26.93$



K-Nearest Neighbors (KNN)

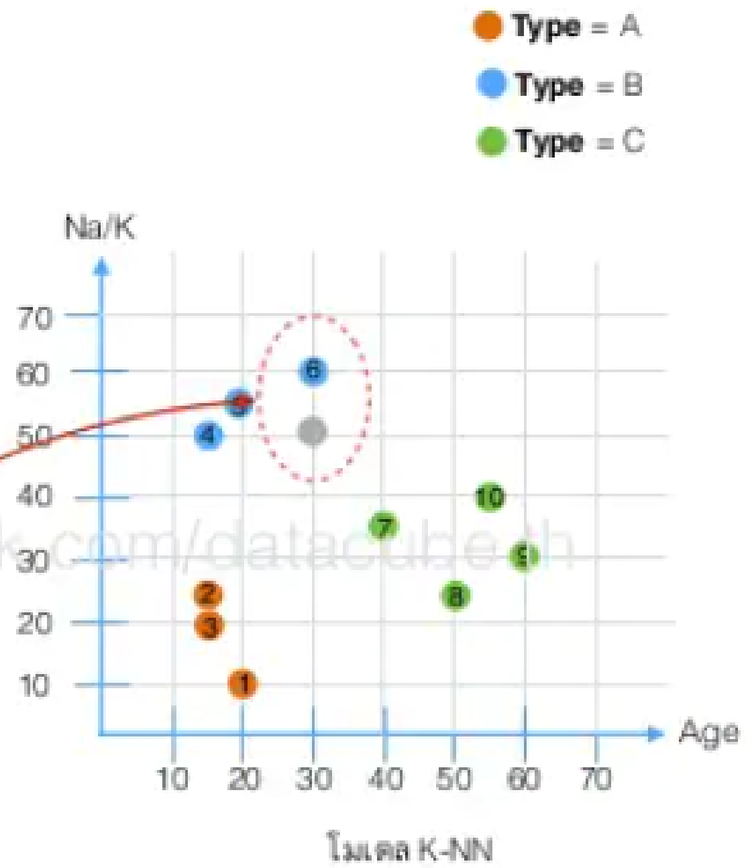
- การใช้โมเดลเพื่อ predict ข้อมูลใหม่ โดยกำหนด K = 1

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)

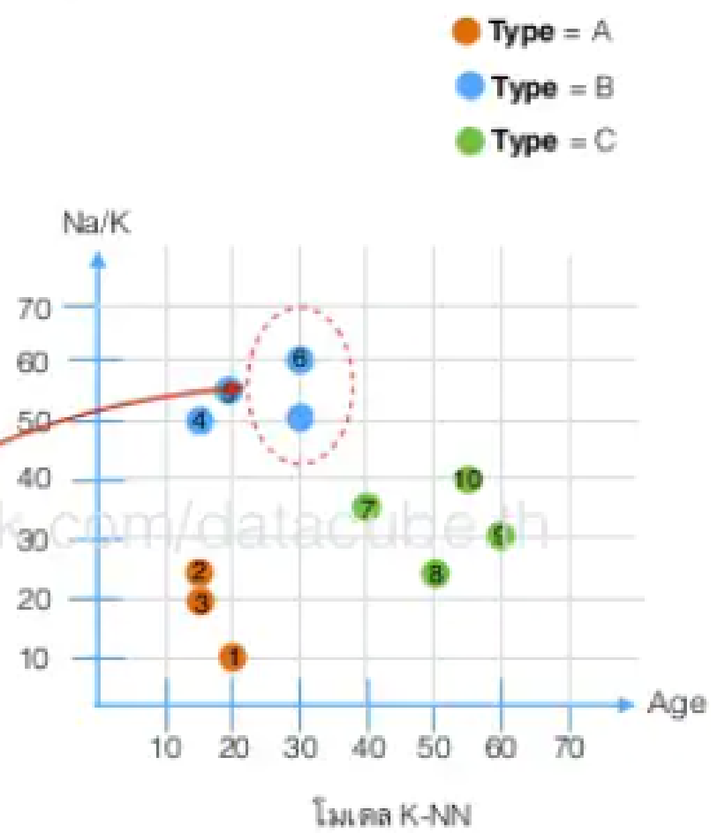
- การใช้โมเดลเพื่อ predict ข้อมูลใหม่ โดยกำหนด K = 1

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	$\sqrt{(40-50)^2 + (55-30)^2} = 26.93$

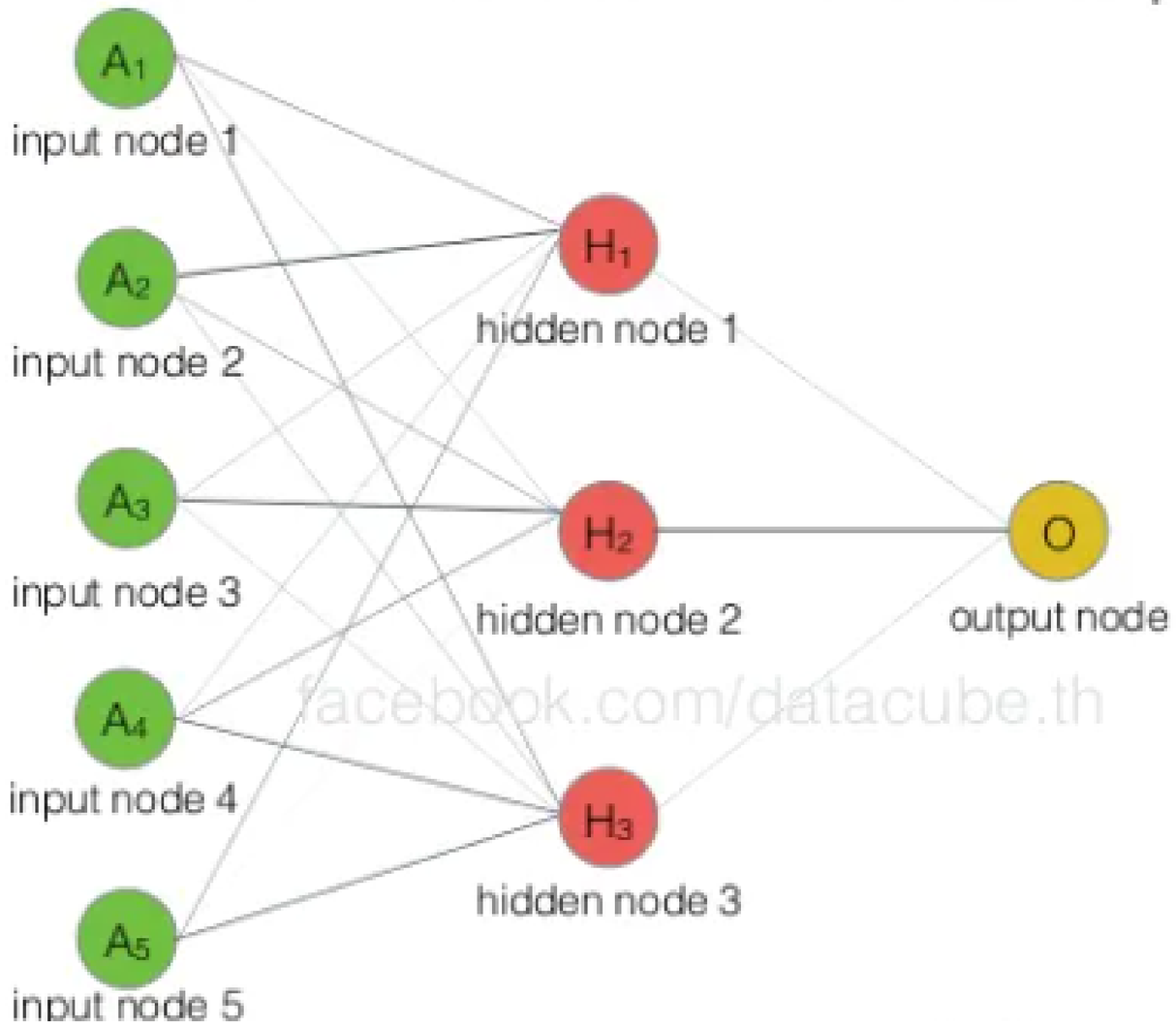


Classification Techniques

- 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

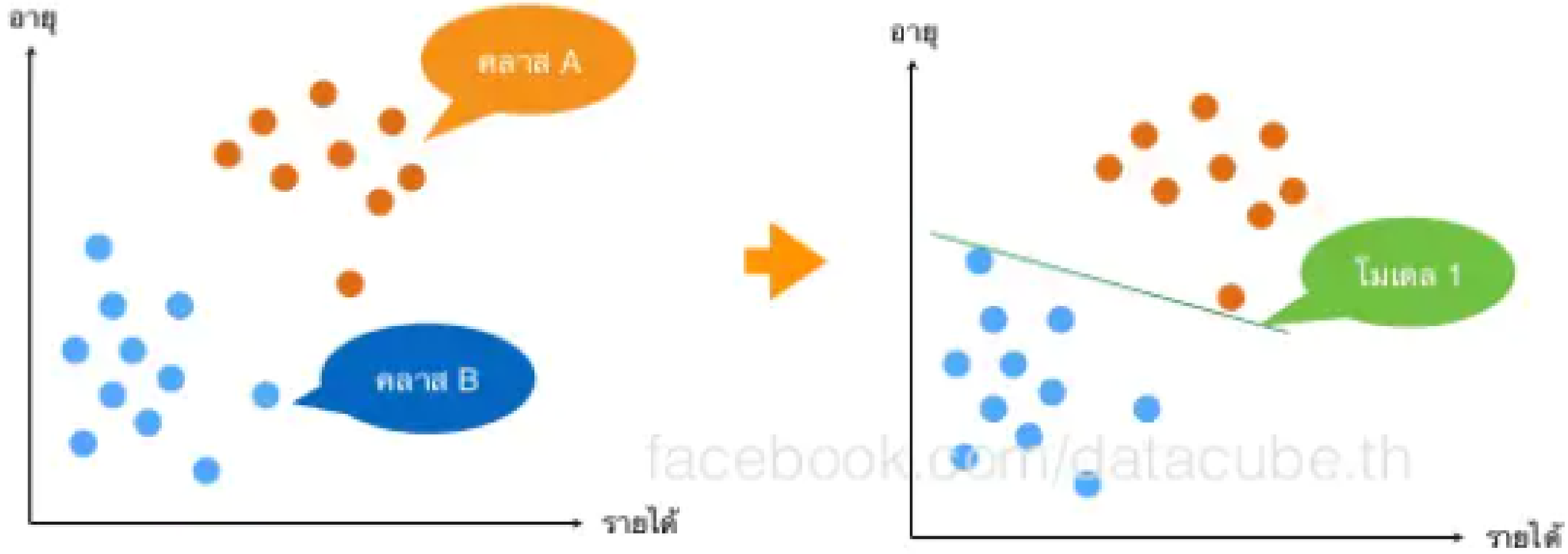
- โมเดลที่ได้แปลความหมายได้ยาก จะอยู่ในรูปแบบของสมการคณิตศาสตร์
  - โมเดลจะแสดงค่าน้ำหนัก (weight) ระหว่างโหนดต่างๆ ใน Neural Network
  - โมเดลจะทำการปรับเปลี่ยนค่าน้ำหนักให้เป็นค่าที่เหมาะสม
- สามารถใช้กับ classification (ทำนายค่า nominal) และ regression (ทำนายค่าตัวเลข) ได้

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

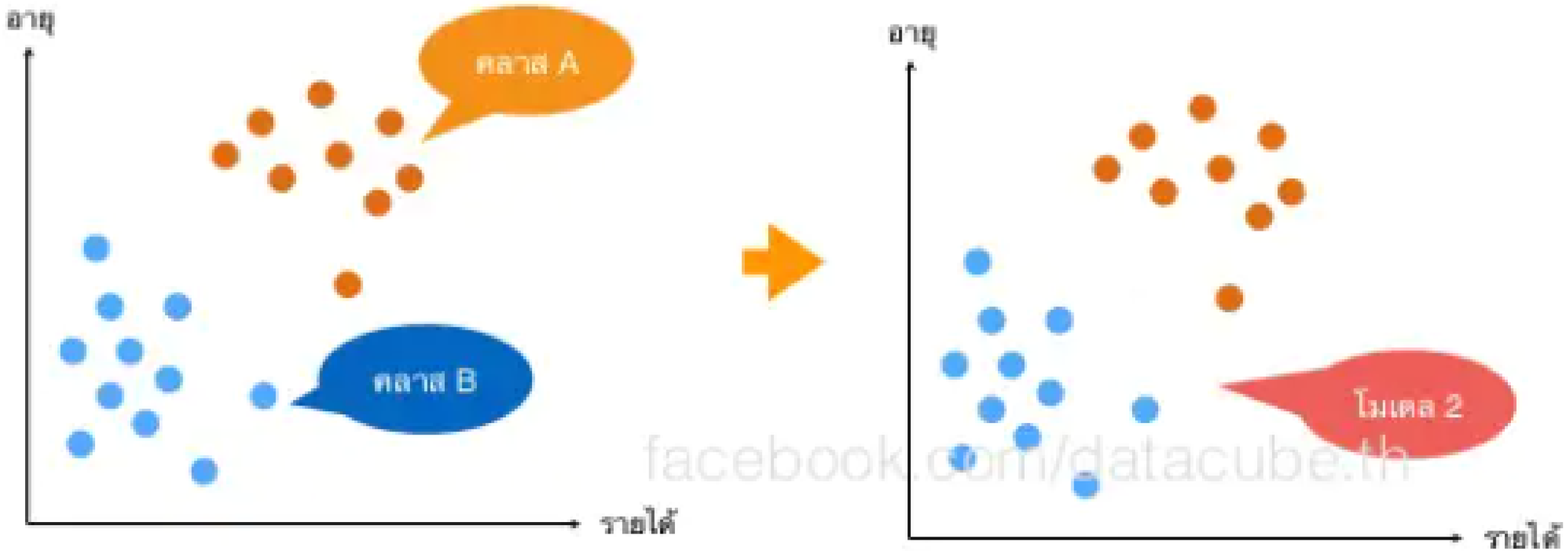
Support Vector Machines (SVM)

- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส



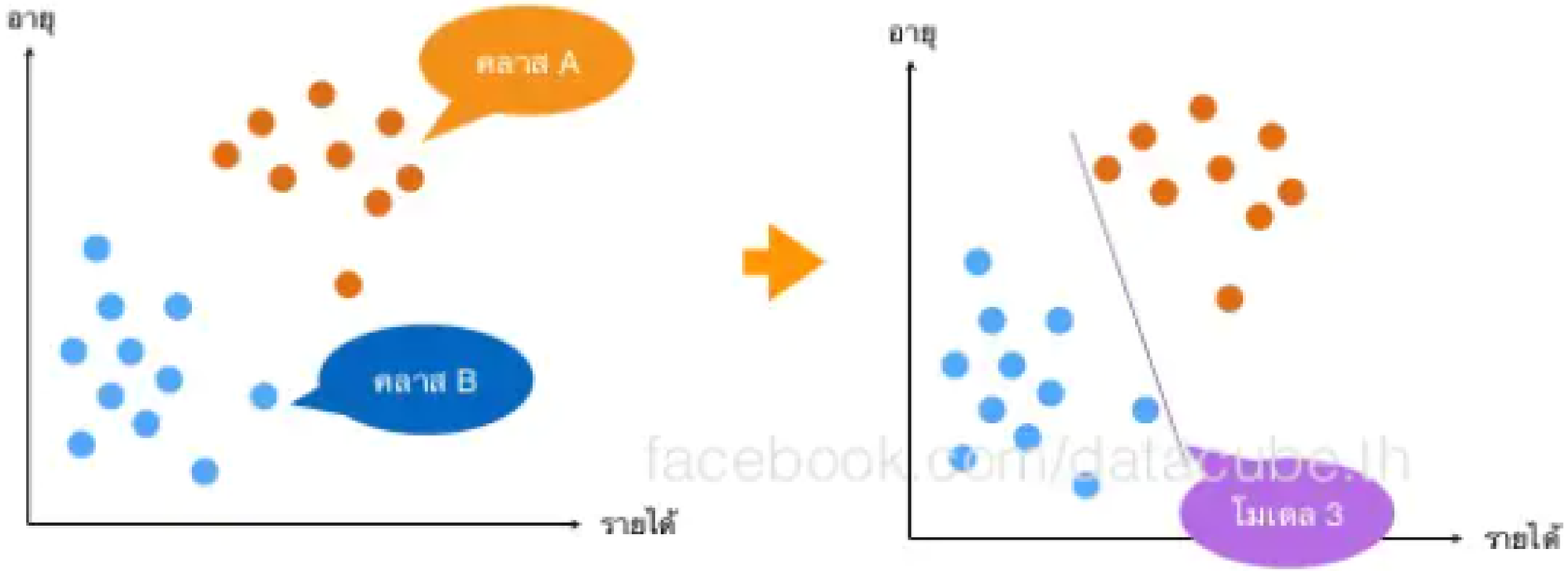
Support Vector Machines (SVM)

- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส



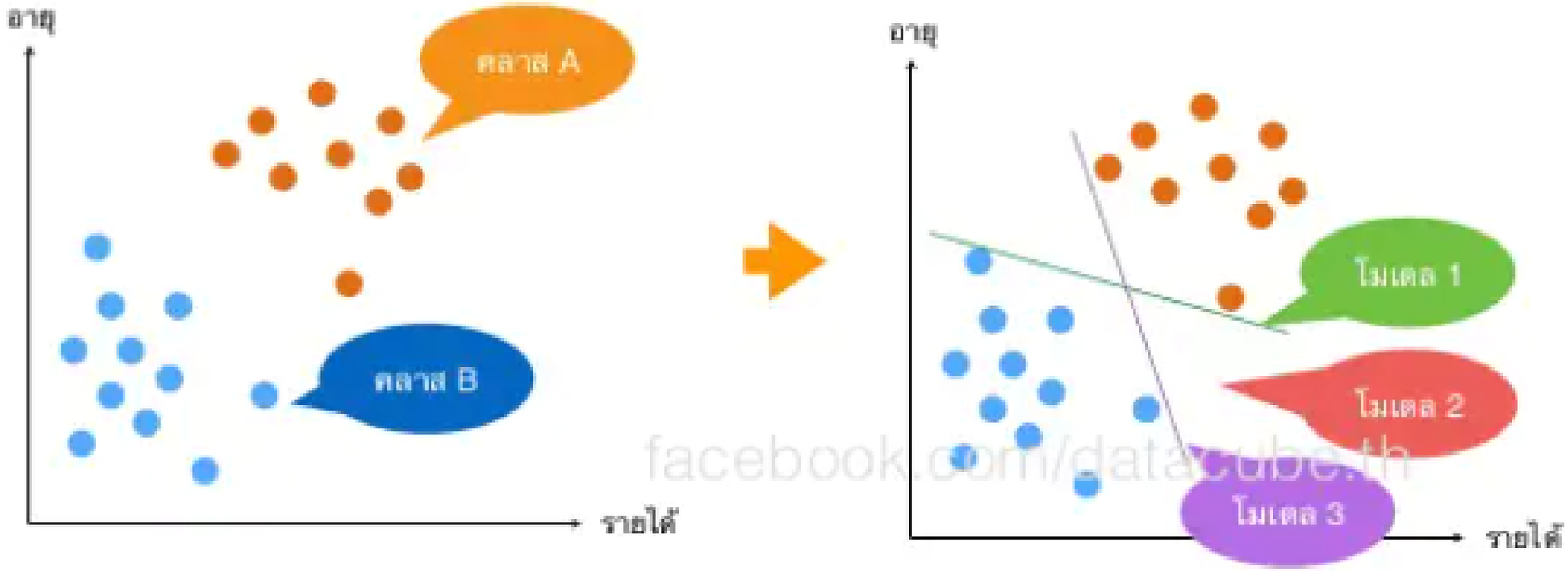
Support Vector Machines (SVM)

- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส



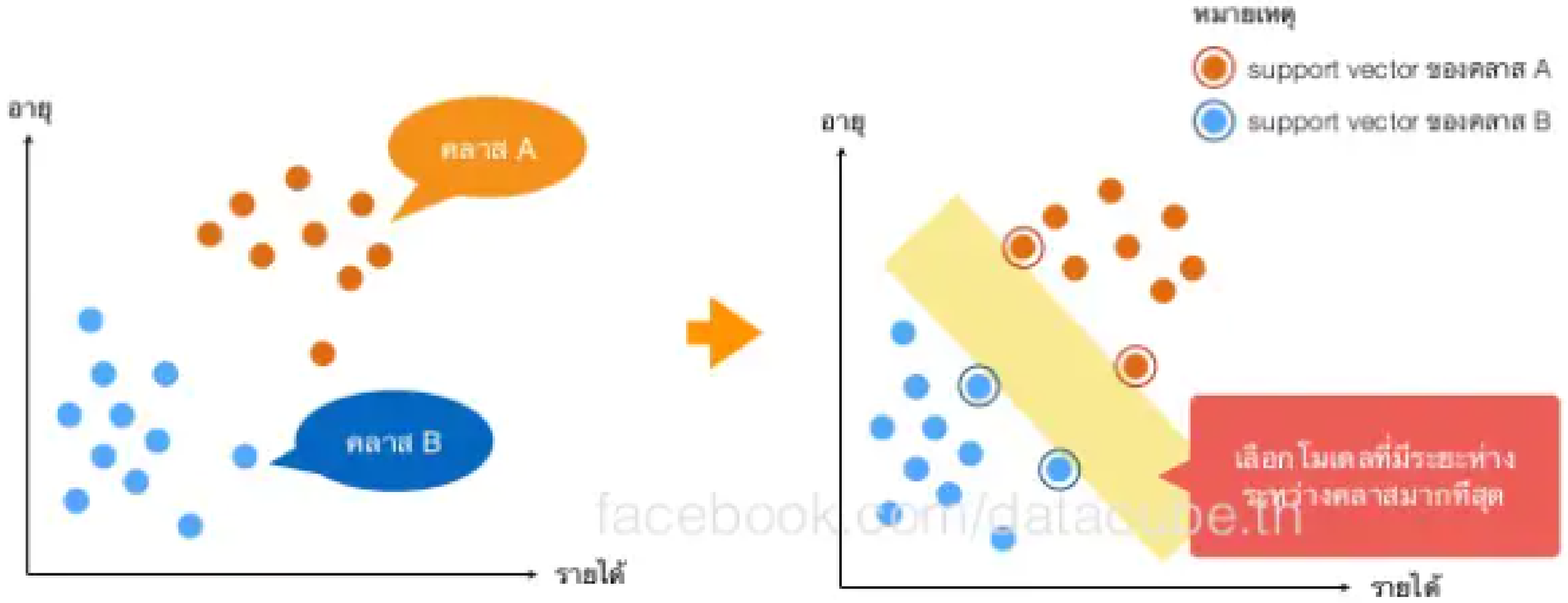
Support Vector Machines (SVM)

- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส



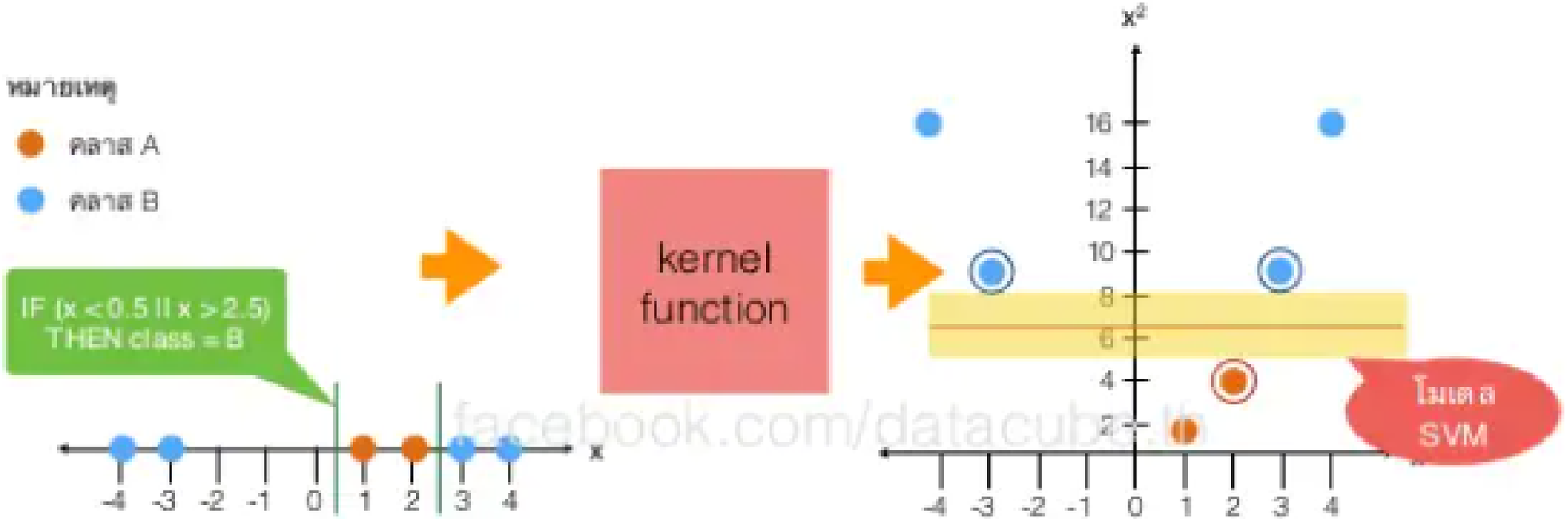
Support Vector Machines (SVM)

- เป็นเทคนิคที่พัฒนาขึ้นมาใหม่ และมีประสิทธิภาพสูง
- ใช้หลักการของโมเดล linear ที่แบ่งข้อมูลออกเป็น 2 คลาส



Support Vector Machines (SVM)

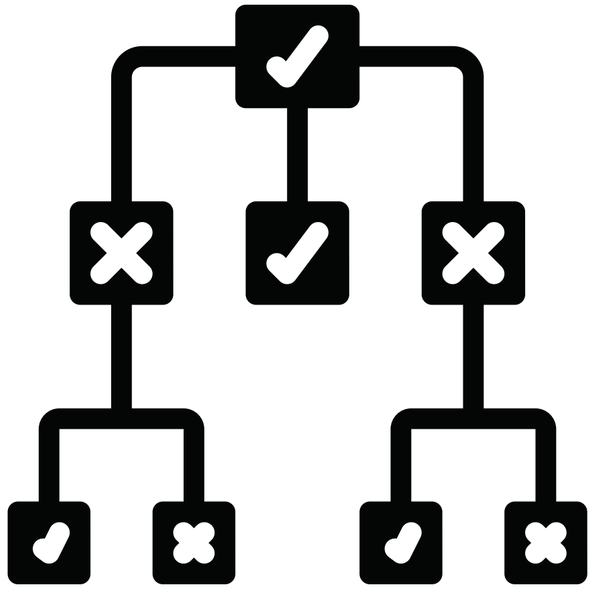
- SVM มีการใช้ kernel function เพื่อทำการแปลงข้อมูลที่ไม่สามารถใช้โมเดล linear ไปอยู่มิติ (dimension) ที่สูงขึ้น
- ใช้โมเดล linear แบ่งข้อมูลได้ง่ายขึ้น



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



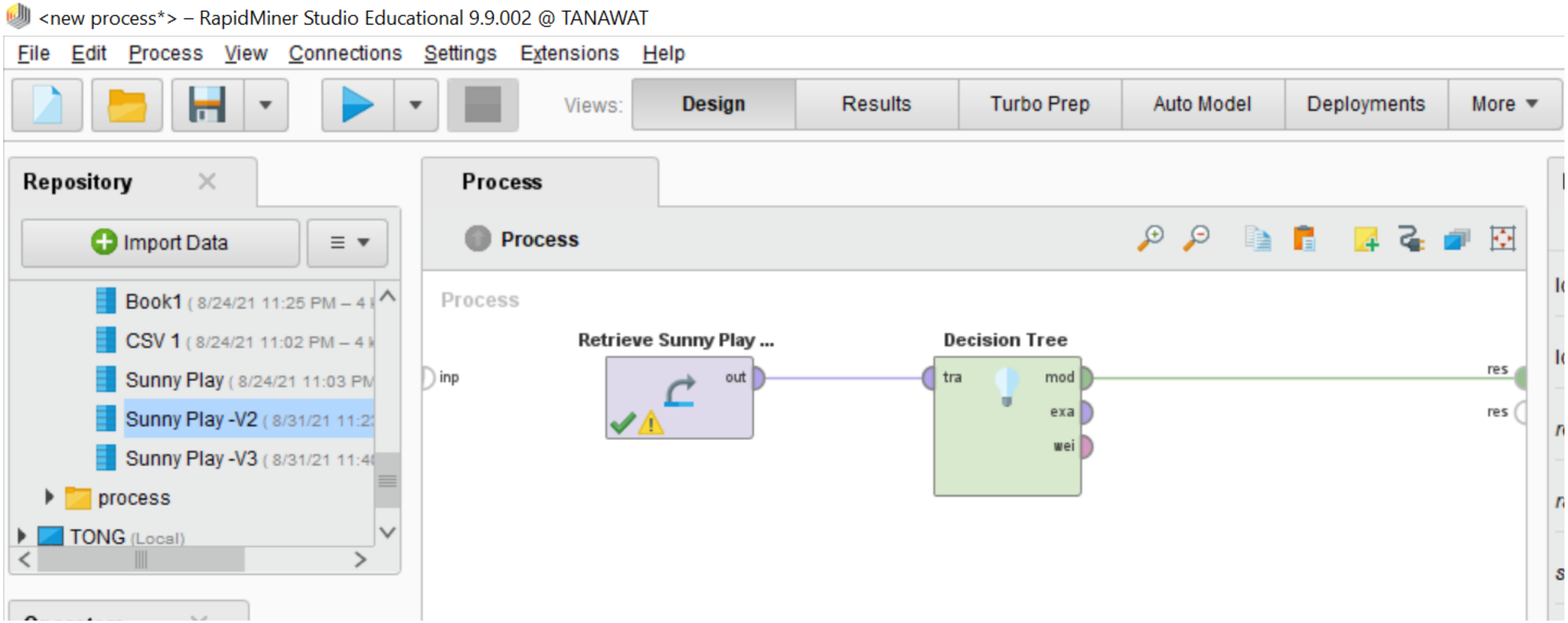
rapidminer

Decision Tree

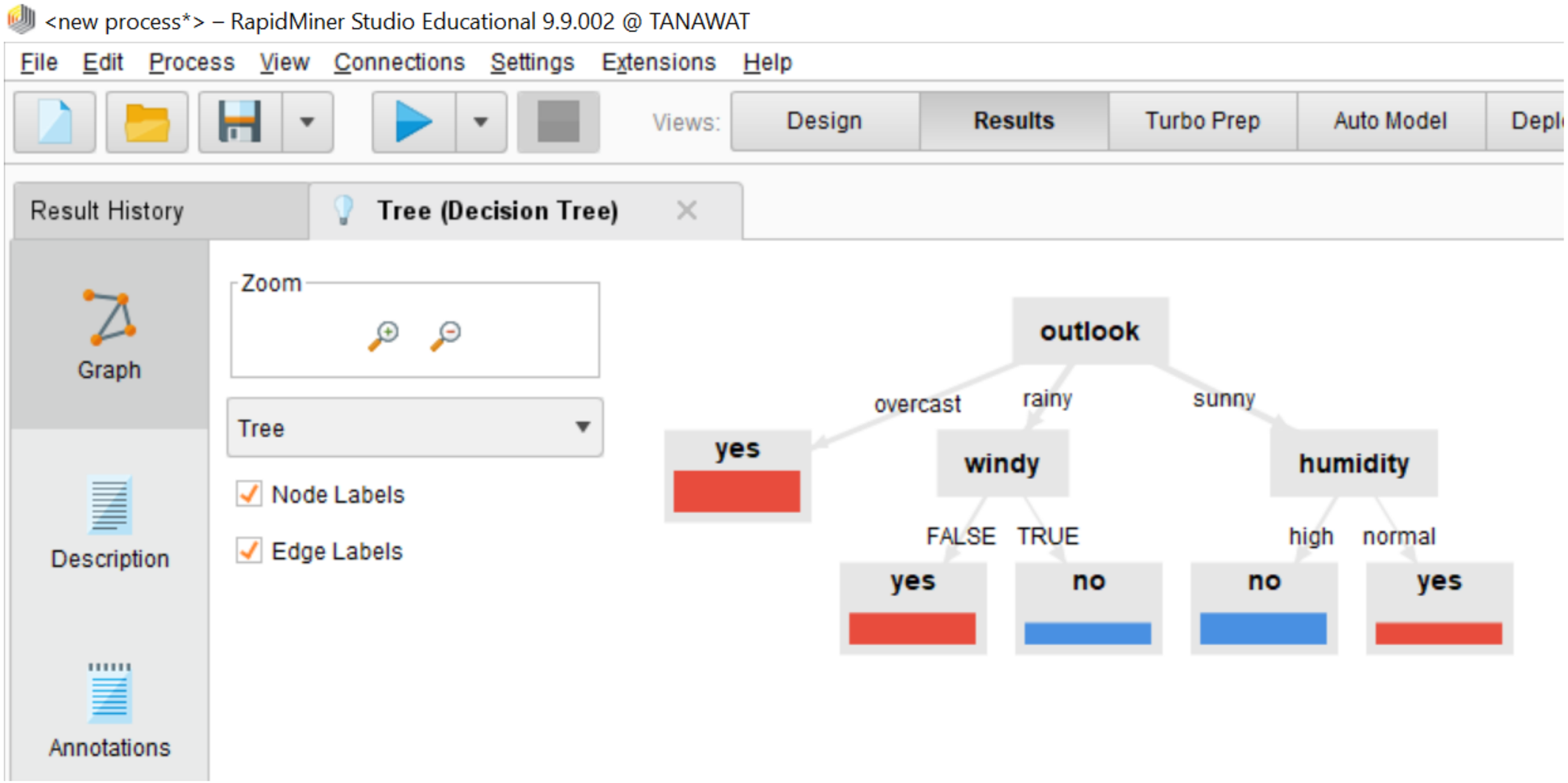
Decision Tree

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

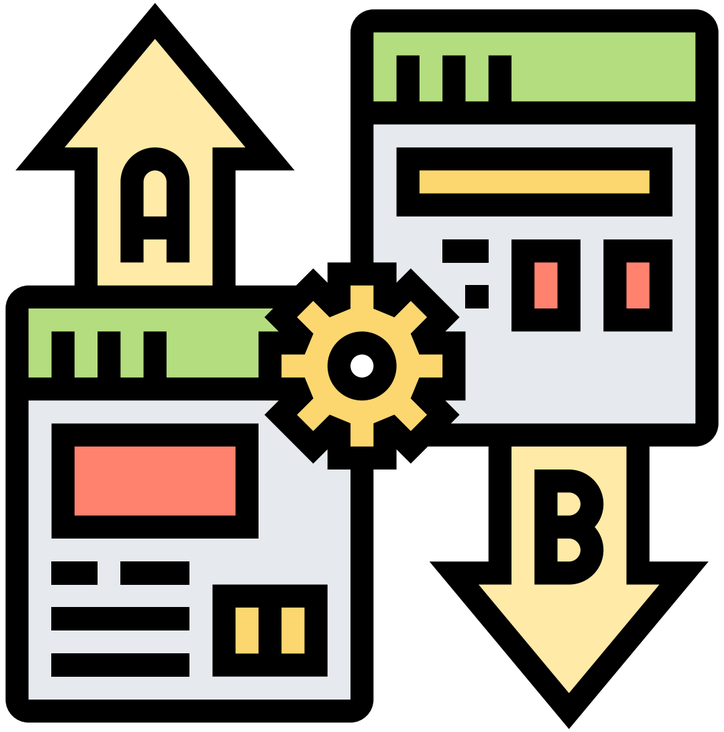
Decision Tree



Decision Tree



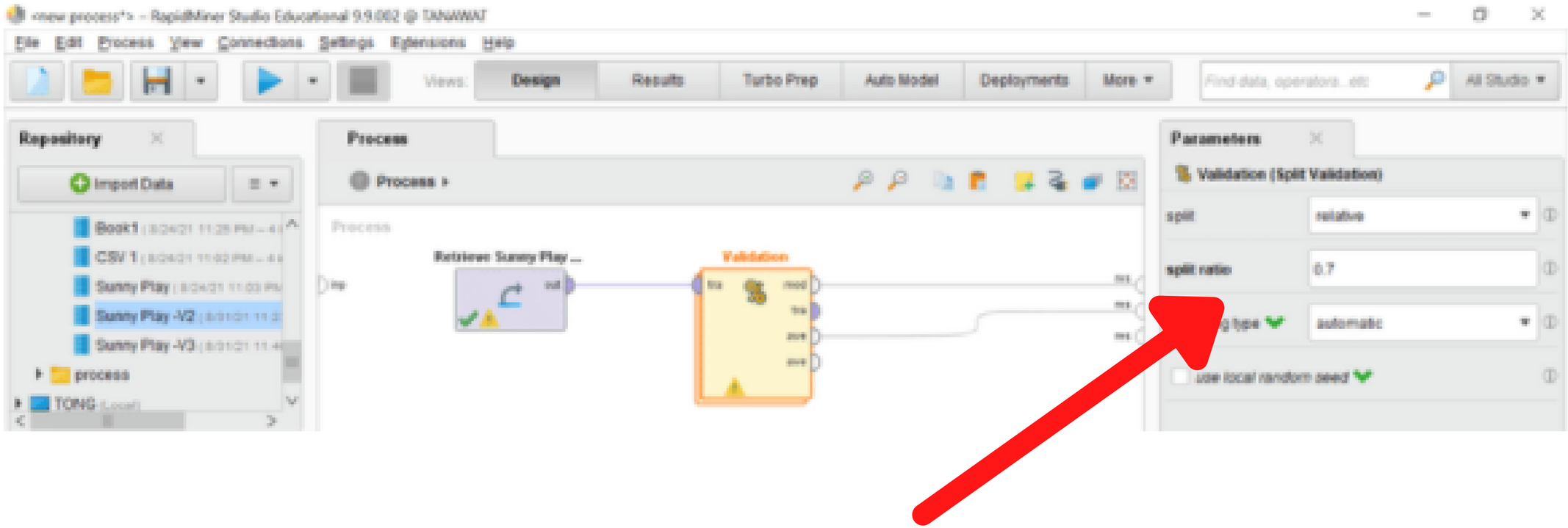
validation



rapidminer

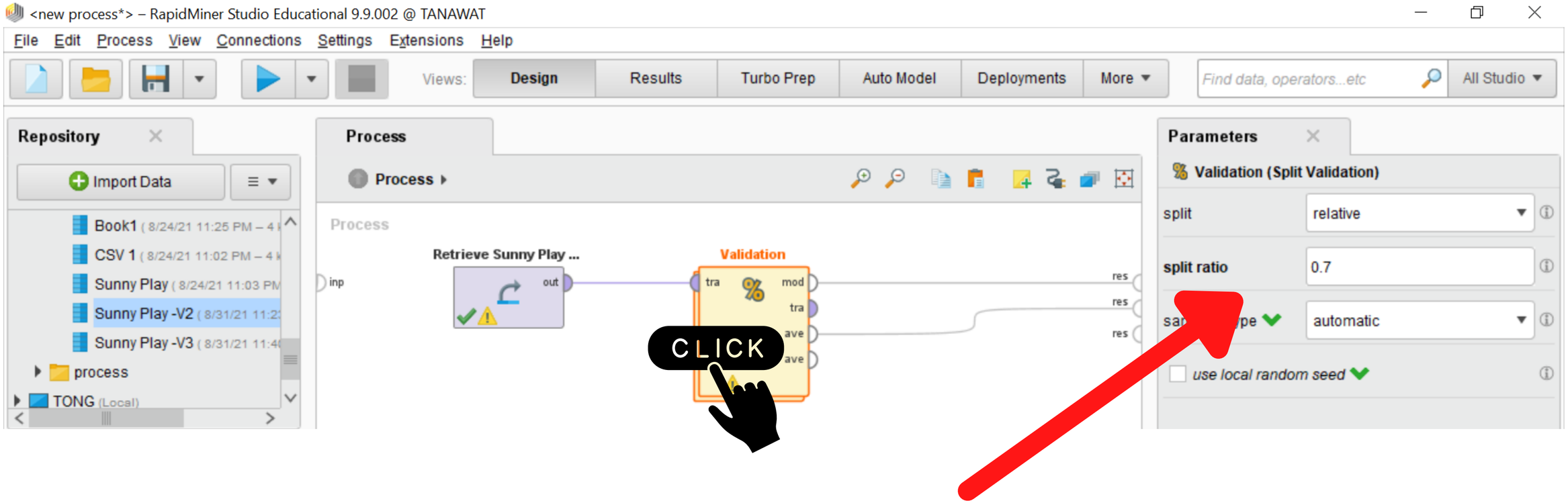
Self Consistency test | Split test | Cross-validation

Split test



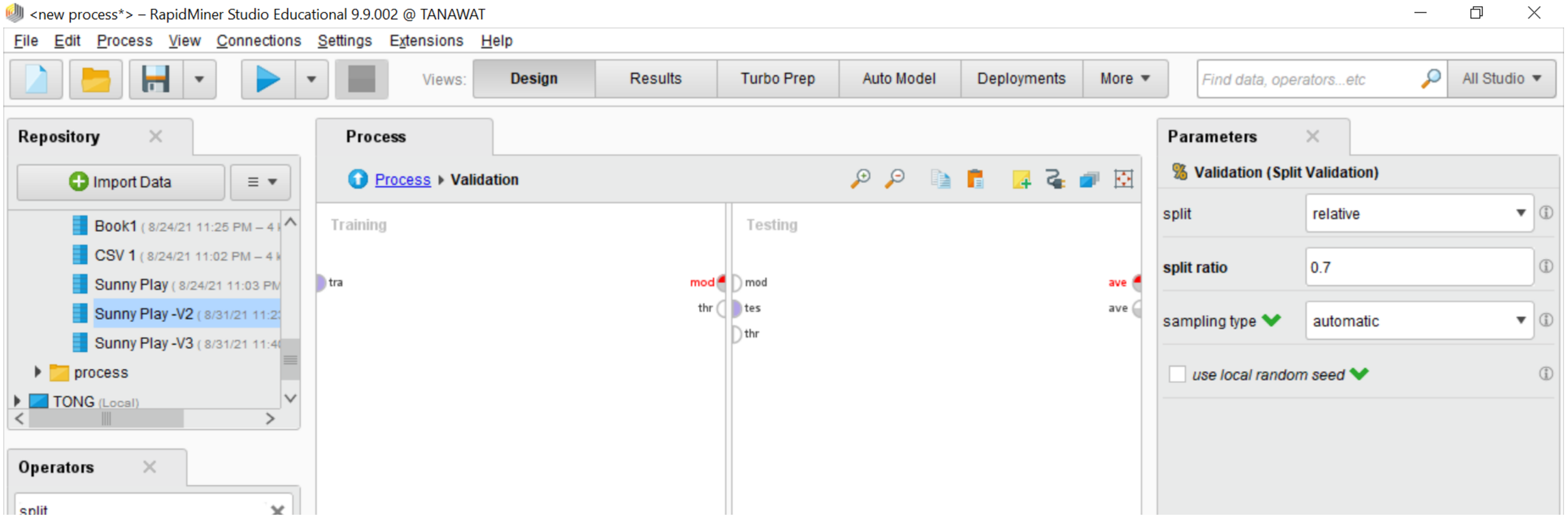
70 % สำหรับสร้างโมเดล 30% ทดสอบโมเดล

Split test

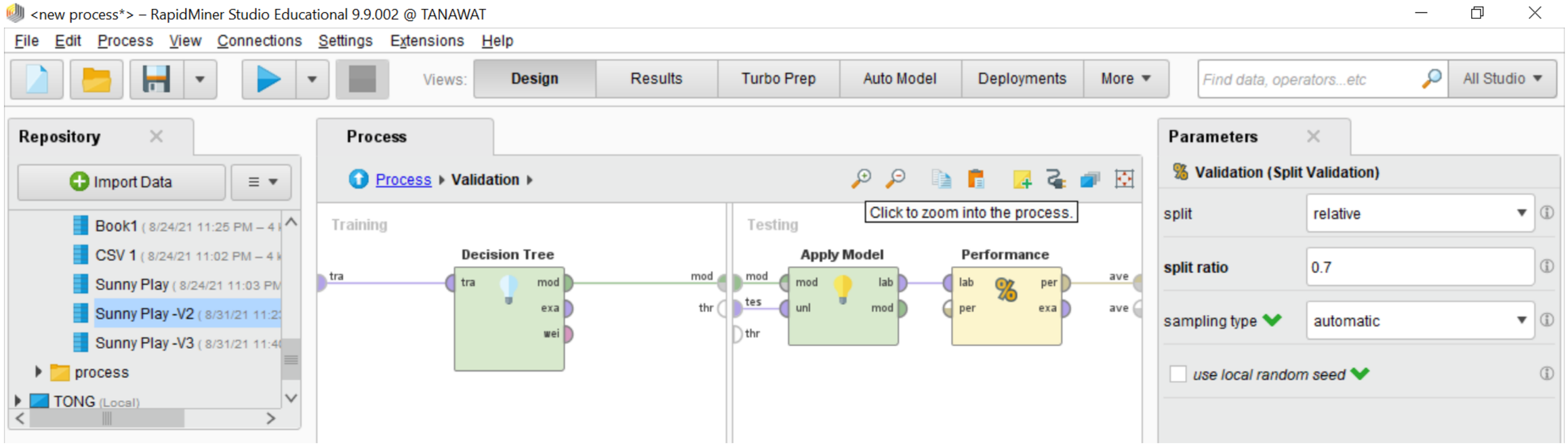


70 % สำหรับสร้างโมเดล 30% ทดสอบโมเดล

Split test


<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
Find data, operators...etc
All Studio
Repository
Import Data
Book1 (8/24/21 11:25 PM – 4
CSV 1 (8/24/21 11:02 PM – 4
Sunny Play (8/24/21 11:03 PM
Sunny Play -V2 (8/31/21 11:2
Sunny Play -V3 (8/31/21 11:4
process
TONG (Local)
Operators
split
Process
Process ▸ Validation
Training
Testing
tra
mod
thr
mod
tes
thr
ave
ave
Parameters
Validation (Split Validation)
split
relative
split ratio
0.7
sampling type
automatic
use local random seed

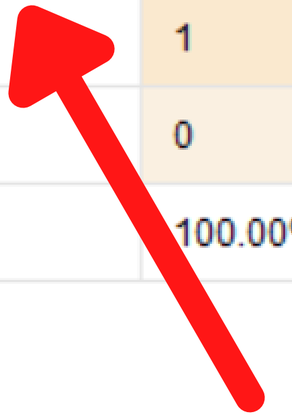
Split test


<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
Find data, operators...etc
All Studio
Repository
Import Data
Book1 (8/24/21 11:25 PM – 4
CSV 1 (8/24/21 11:02 PM – 4
Sunny Play (8/24/21 11:03 PM
Sunny Play -V2 (8/31/21 11:2
Sunny Play -V3 (8/31/21 11:4
process
TONG (Local)
Process
Process ▸ Validation ▸
Click to zoom into the process.
Training
Testing
Decision Tree
Apply Model
Performance
tra
mod
exa
wei
thr
tes
unl
lab
per
ave
Parameters
Validation (Split Validation)
split
relative
split ratio
0.7
sampling type
automatic
use local random seed

Split test

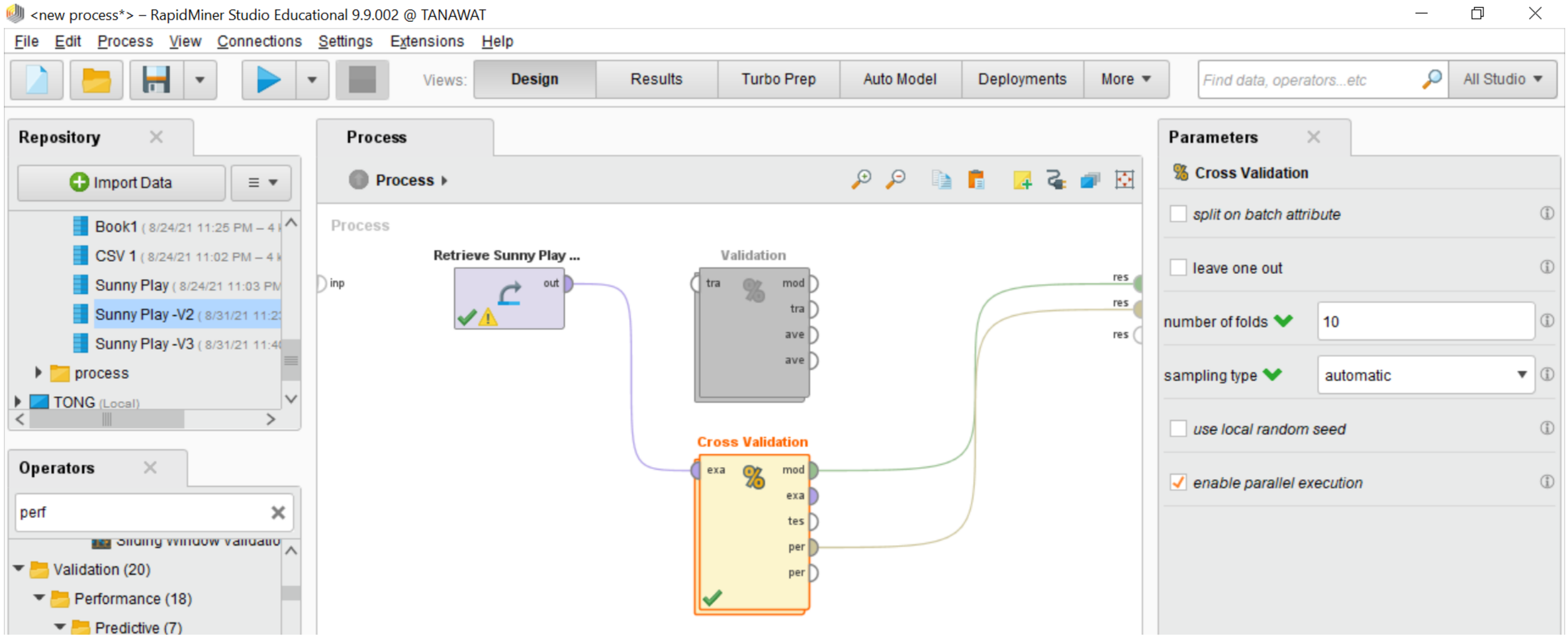
accuracy: 25.00%

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

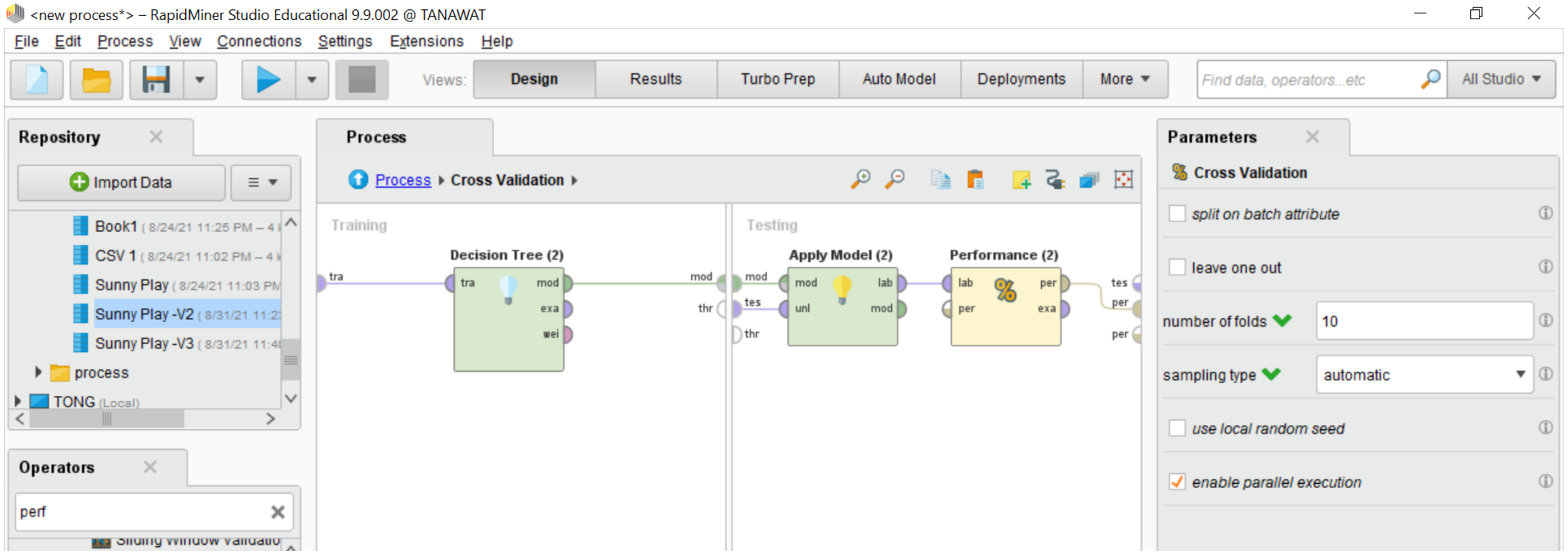


ถูกต้อง 25%

Cross-validation



Cross-validation

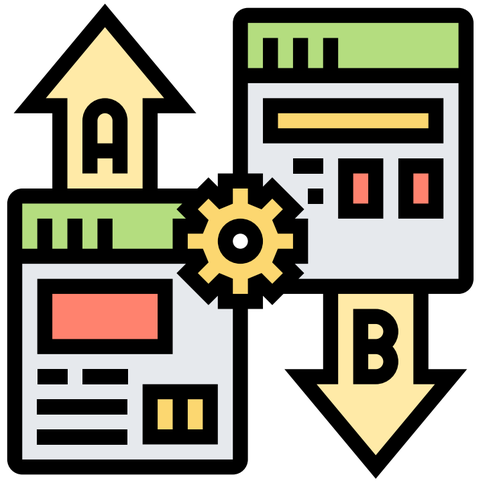


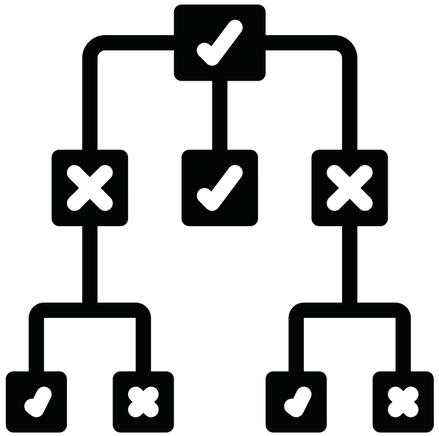
Cross-validation

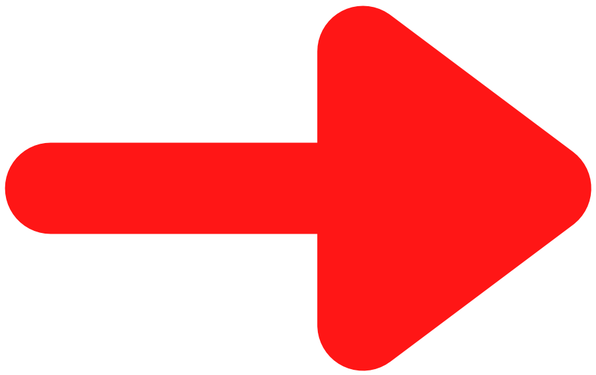
accuracy: 40.00% +/- 45.95% (micro average: 42.86%)

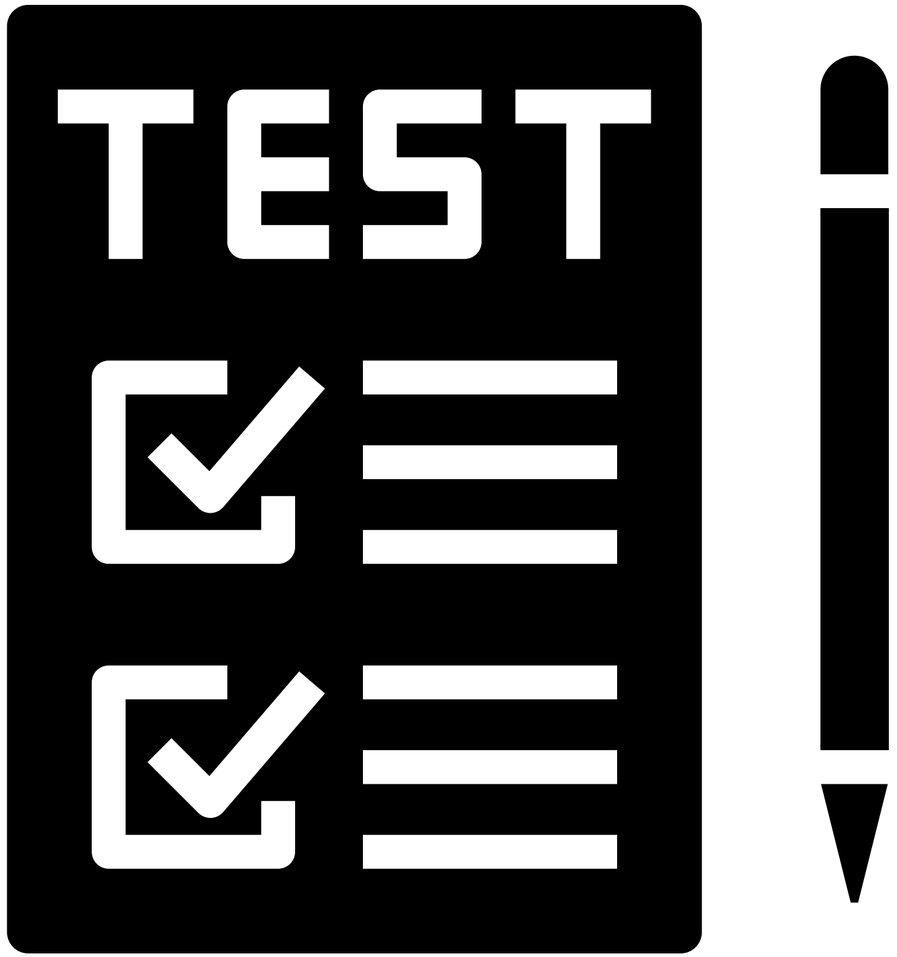
	true no	true yes	class precision
pred. no	2	5	28.57%
pred. yes	3	4	57.14%
class recall	40.00%	44.44%	

นำข้อมูลมาทดลองจริง

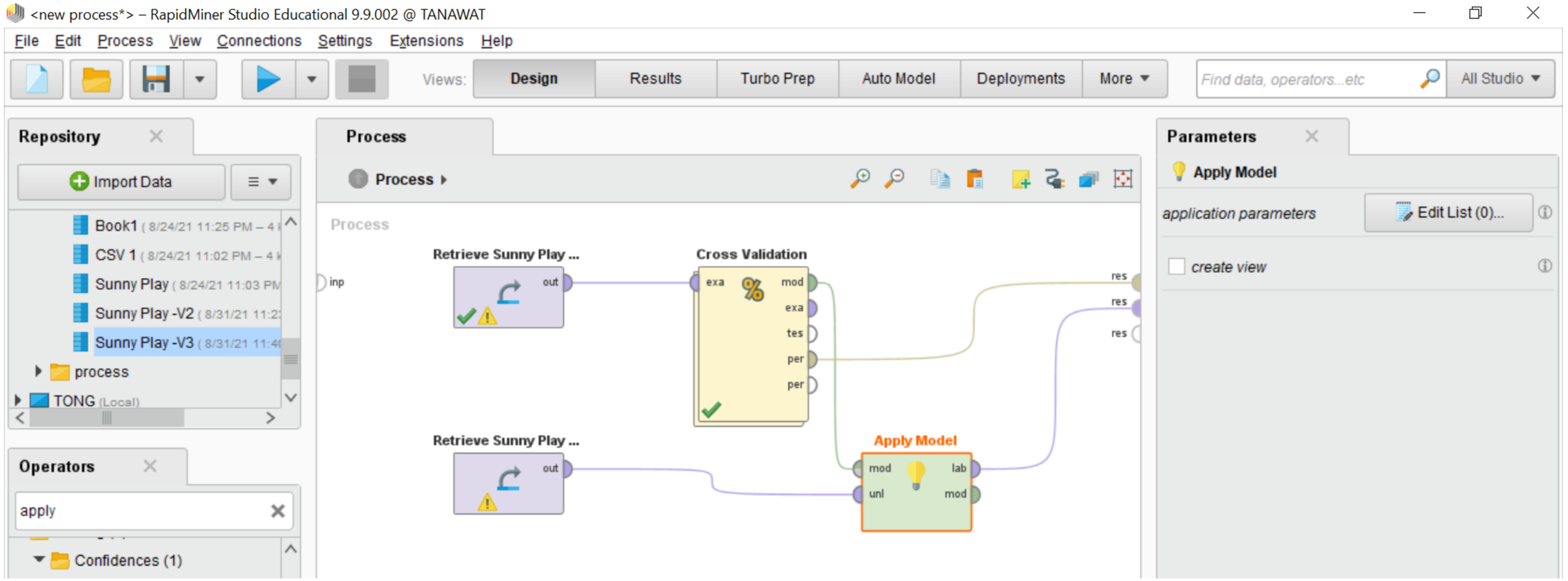




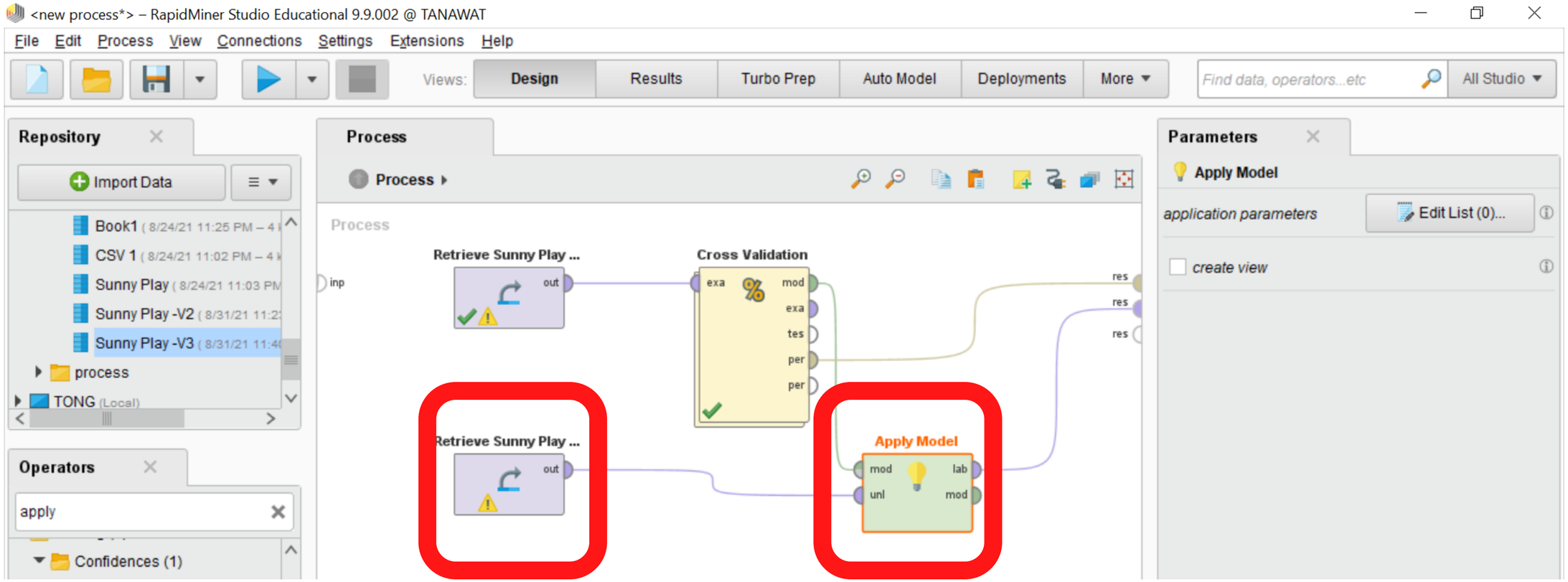




นำข้อมูลมาทดลองจริง



นำข้อมูลมาทดลองจริง



ข้อมูลที่ต้องการทดสอบ

ประยุกต์โมเดล

นำข้อมูลมาทดลองจริง

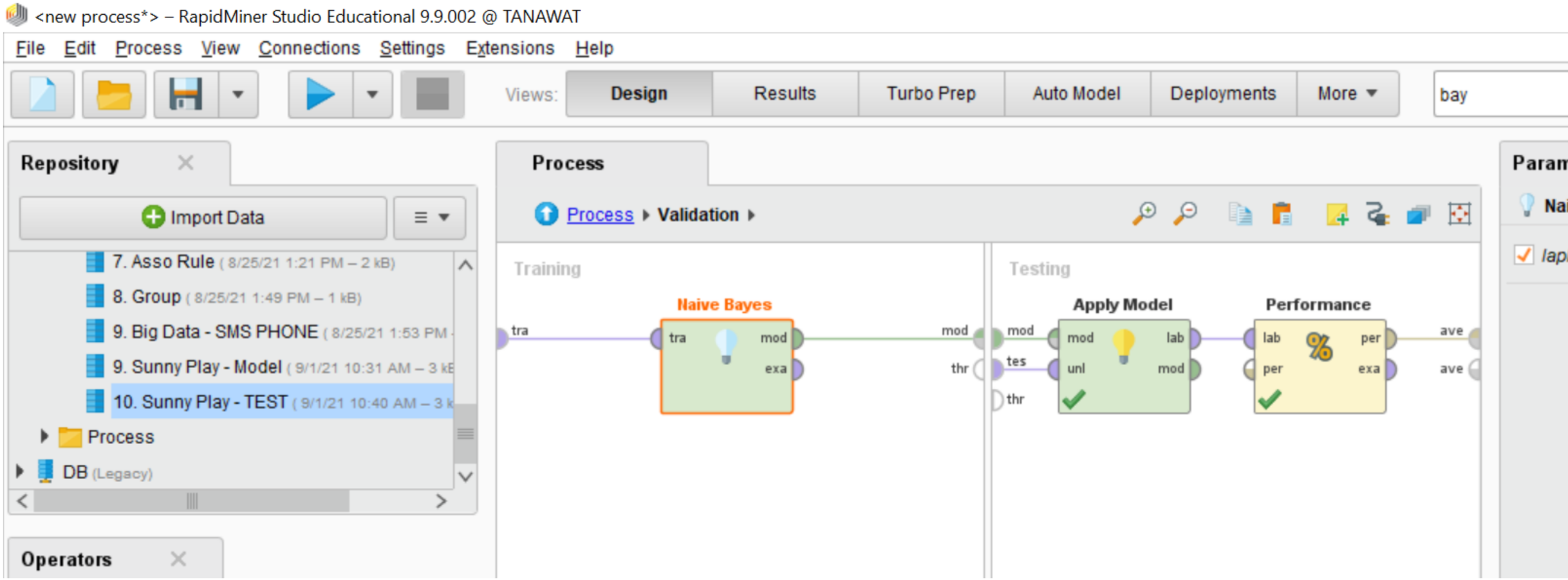
ExampleSet (Apply Model) | PerformanceVector (Performance (2))

Open in Turbo Prep | Auto Model

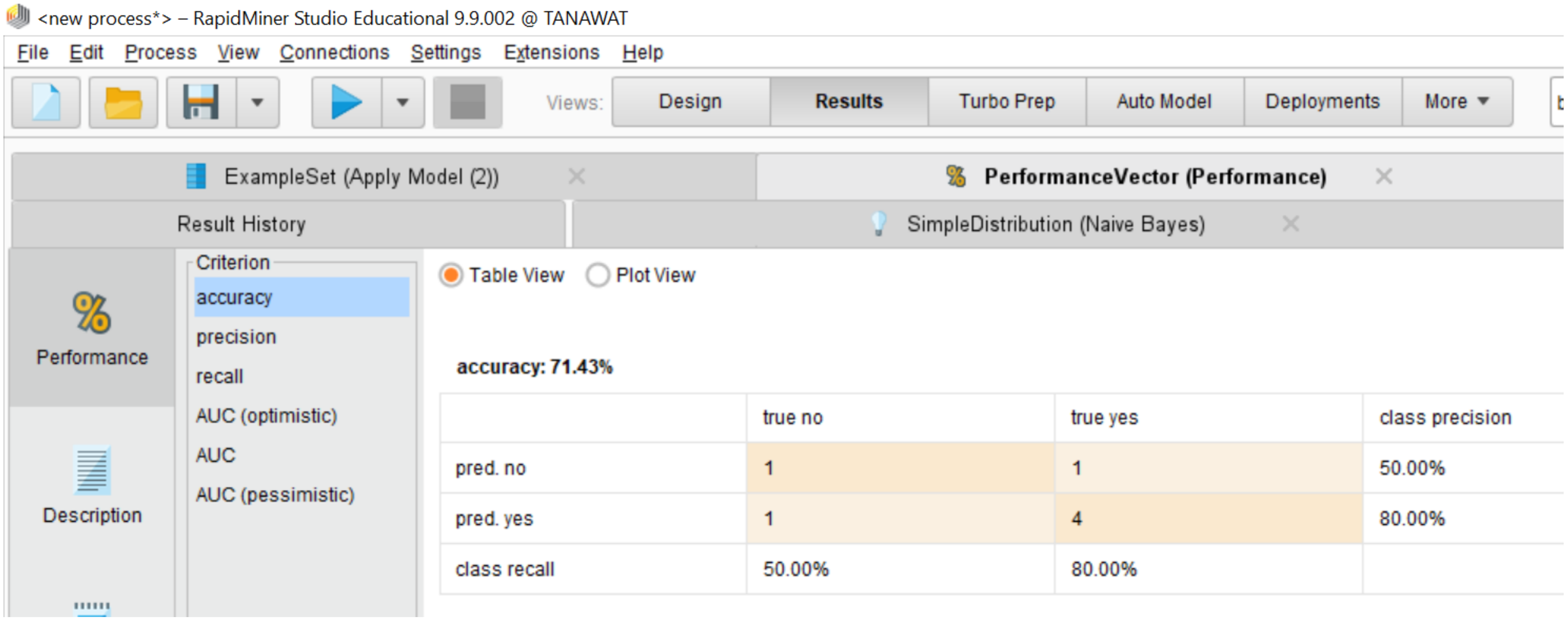
Filter (14 / 14 examples): all

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

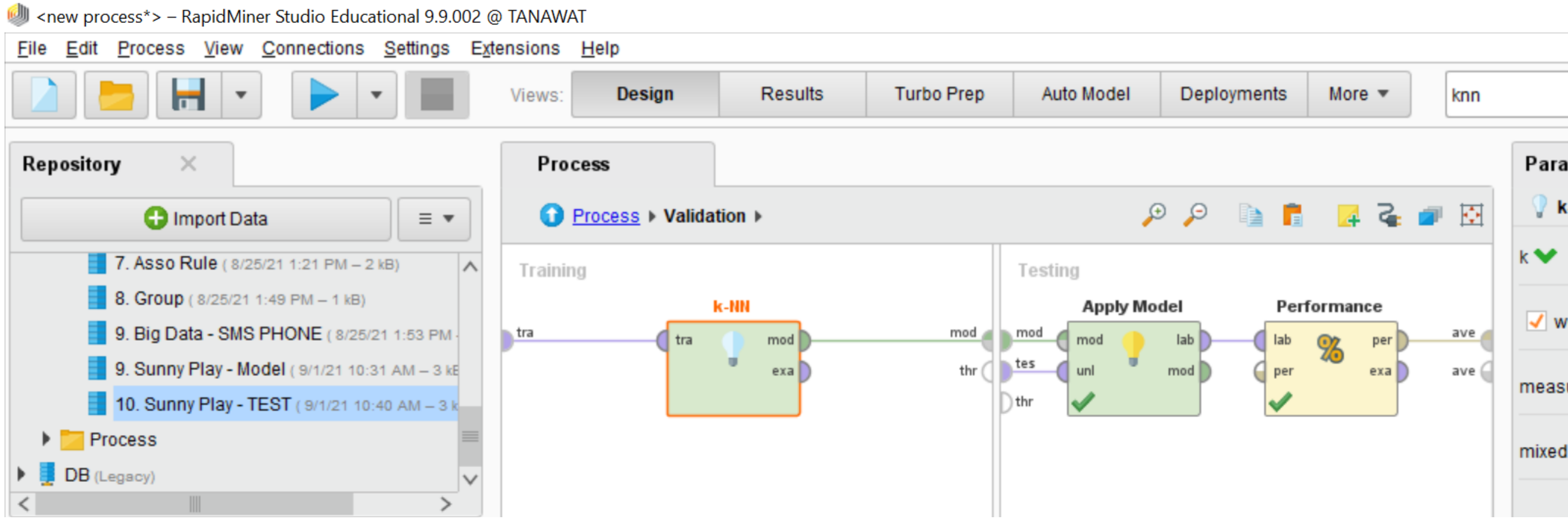
เปลี่ยนใช้ Model [Naive]



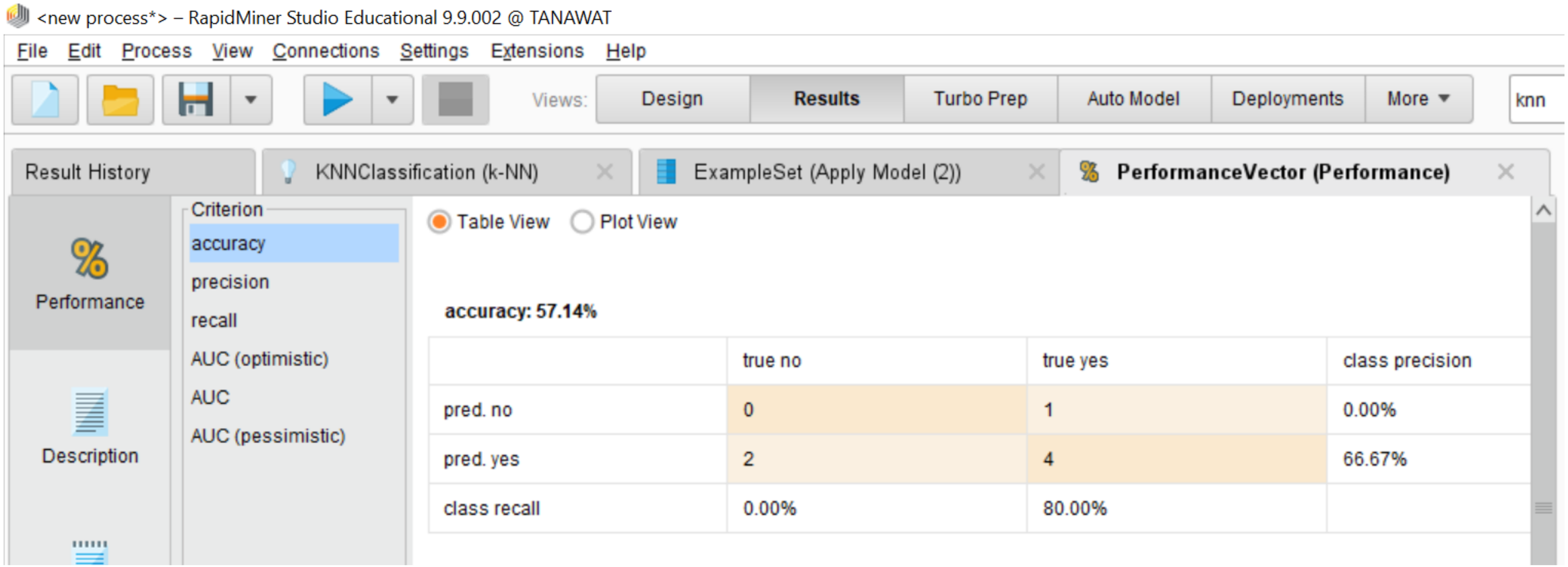
เปลี่ยนใช้ Model [Naive]



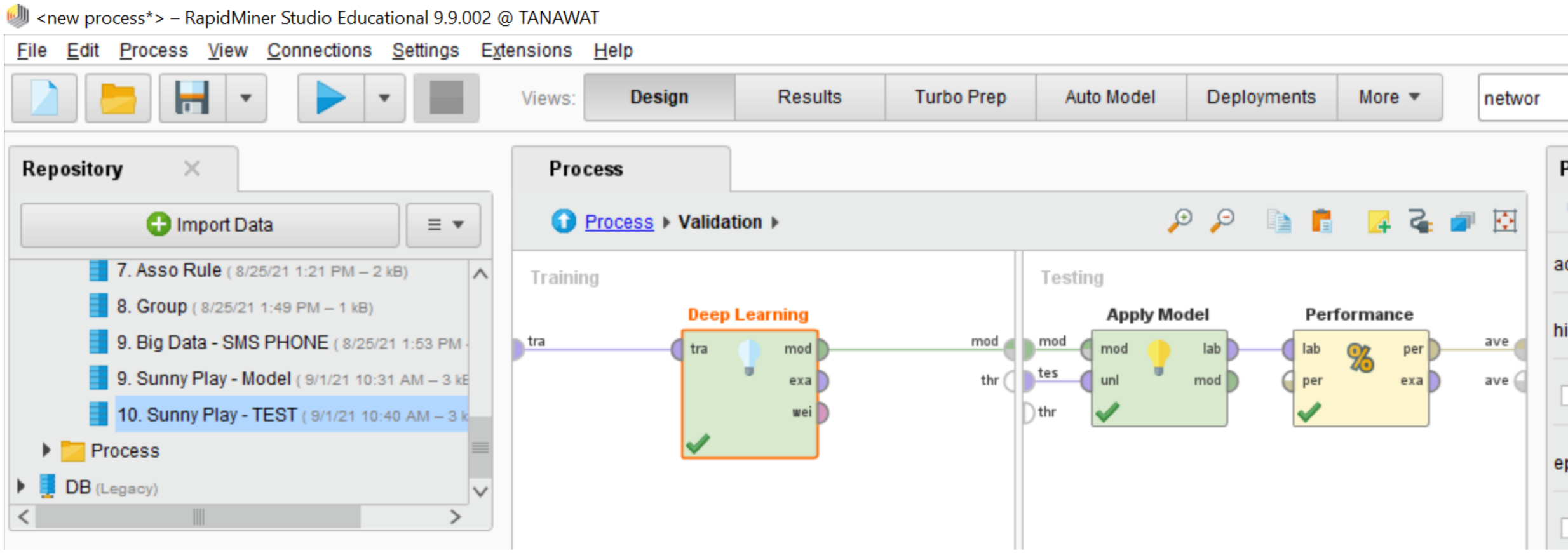
เปลี่ยนใช้ Model [KNN]



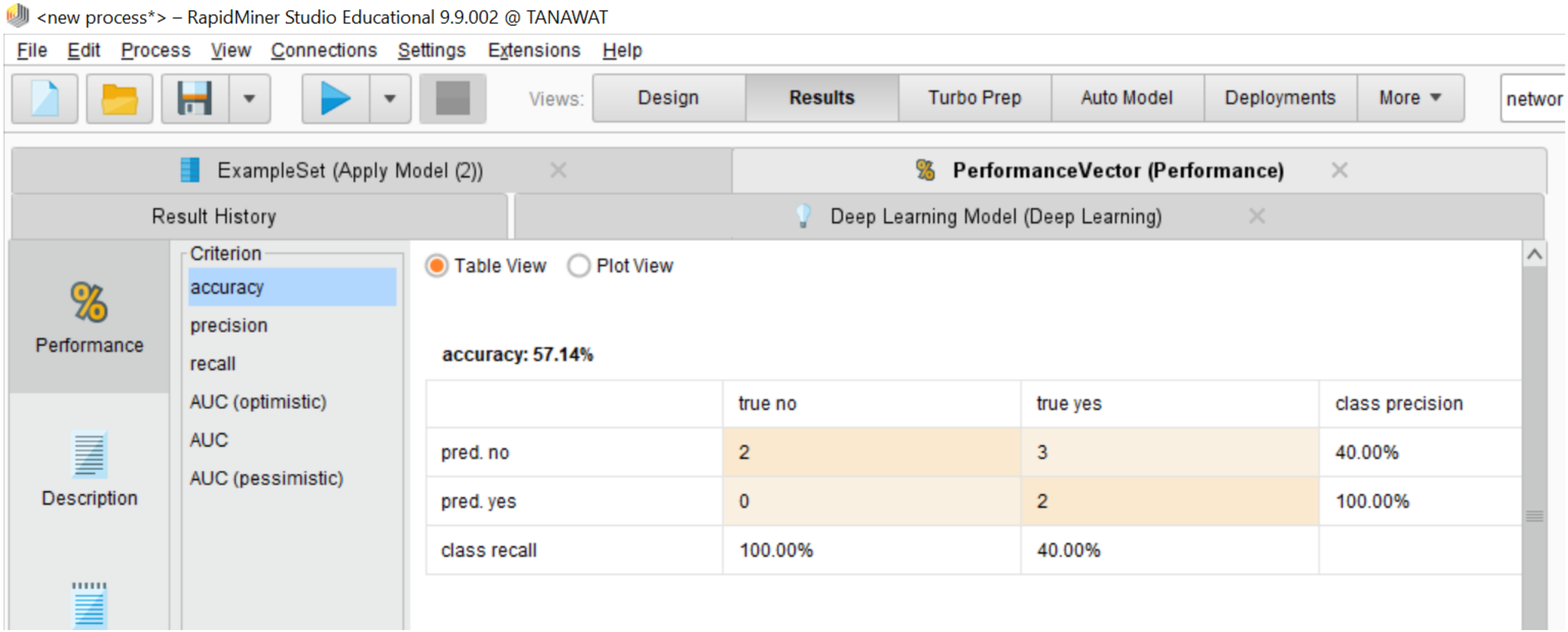
เปลี่ยนใช้ Model [KNN]



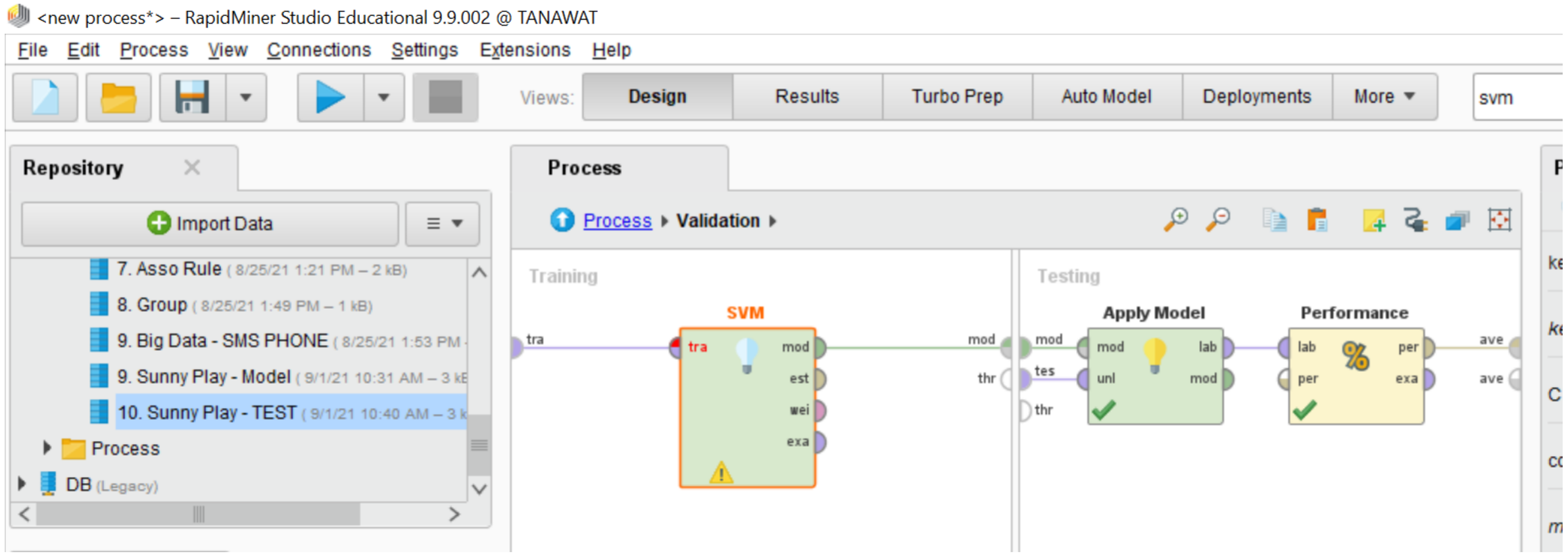
เปลี่ยนใช้ Model [Netword]



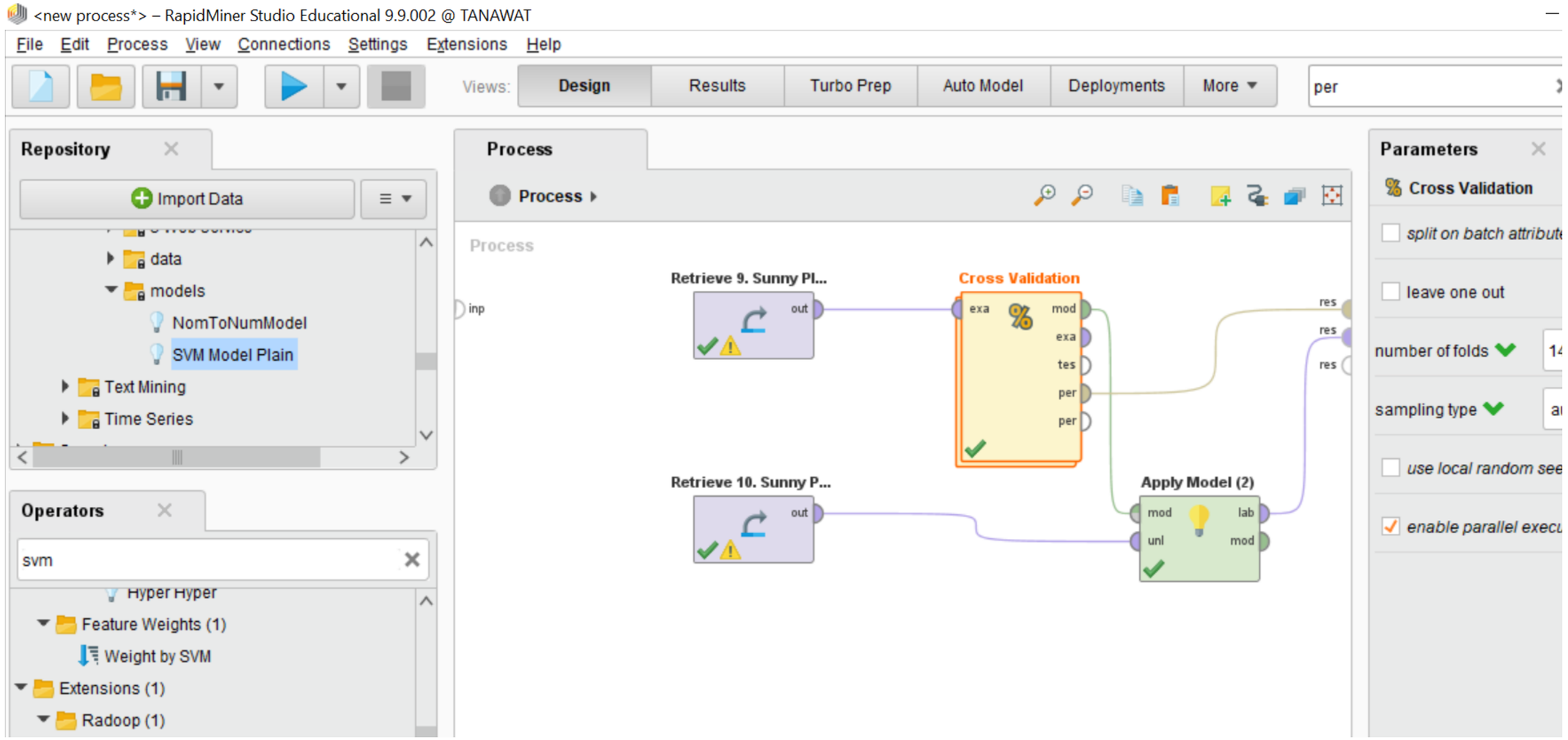
เปลี่ยนใช้ Model [Netword]



เปลี่ยนใช้ Model [SVM]



เปลี่ยน Cross Validation



ทำช้าๆ ด้วยกัน