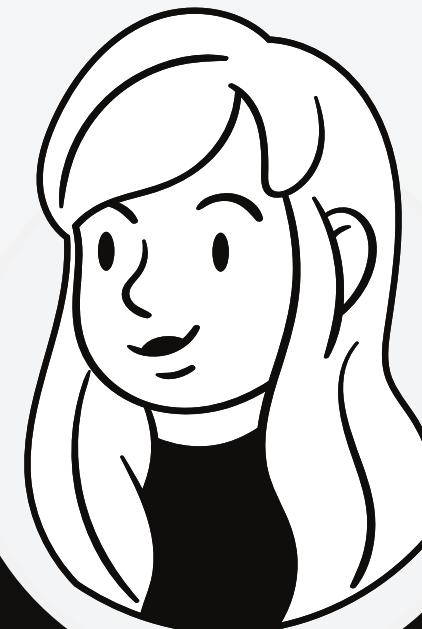


HEART ATTACK ANALYSIS & PREDICTION

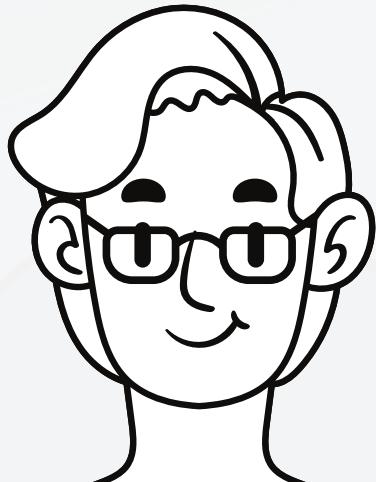
การวิเคราะห์และการทำนายอาการหัวใจวาย

GROUP 3



B6321451

นางสาวขวัญจิรา
พันธุเกตุ



B6326234

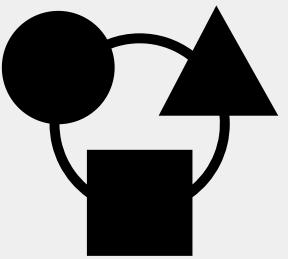
นายกิวากร สีมาเกิด

ที่มา



ต้องการหาแพทเทิร์นของข้อมูลผู้ที่มีอาการการเกิดภาวะหัวใจล้มเหลว หรือภาวะหัวใจวาย เพื่อทำนายโอกาสที่จะเกิดภาวะหัวใจวายในอนาคต

วัตถุประสงค์



ต้องการทดลองนำอัลกอริทึมที่ได้ศึกษาในรายวิชา Knowledge Discovery and Data Mining มาใช้ในการสร้างโมเดล



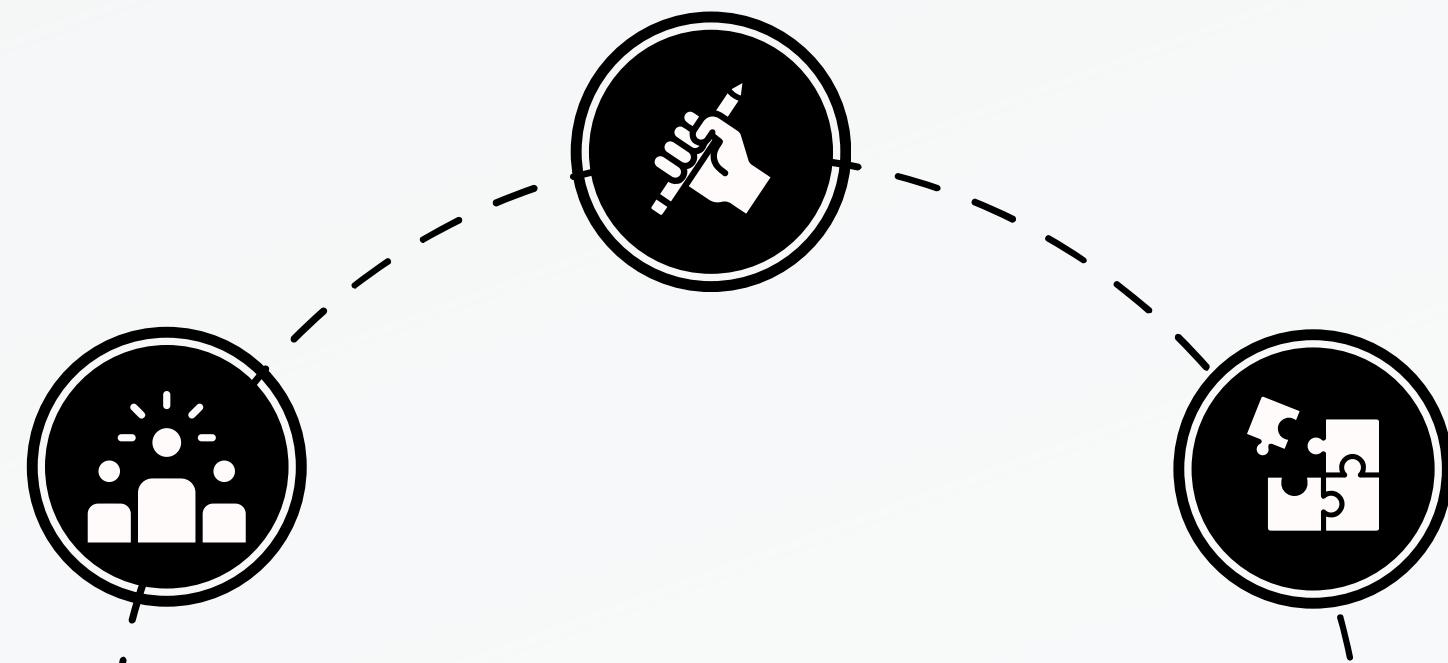
ต้องการเปรียบเทียบโมเดลที่ได้จากการทดลอง เพื่อหาโมเดลที่ดีที่สุด



ຖາມដីកំណើន

1. Data Preparation

ក្របែនការចុះព័ត៌មាន គឺជាក្របែនការកំណើនដែលមានគោលការណ៍ និងគោលការណ៍សម្រាប់ប្រើប្រាស់នៅក្នុងការបង្កើតគម្រោង។ វាបានបង្កើតឡើងដើម្បីរាយការណ៍ និងបង្កើតគម្រោង។



ຖາມເຈົ້າທີ່ເກີຍວິຂອງ

2. OneR

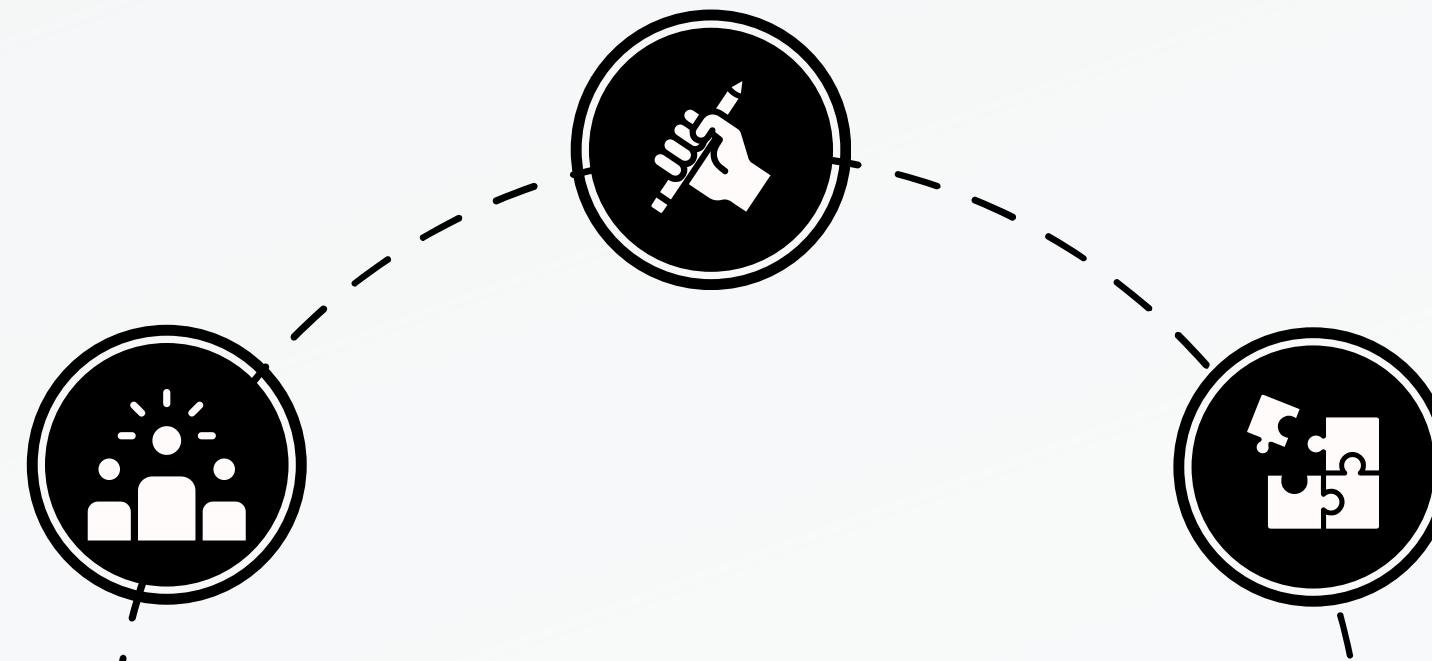
OneR ດັ່ງກ່າວກັບການສ້າງກຸບເພີ່ມກຸບເດືອນ (one rule) ສໍາຮັບແຕ່ລະຕົວທຳນາຍ ໂດຍ
ການພິຈາລະນາຄວາມຄືຂອງແຕ່ລະຄ່າຂອງຕົວທຳນາຍ (predictor) ວ່າຄລາສໄຫມ໌
ຄວາມຄື່ນາກກີ່ສຸດ ແລ້ວກຳນົດໃຫ້ກຸບນີ້ກຳນົດຄລາສນີ້ນໃຫ້ກັບຕົວທຳນາຍນີ້ນັ້ນກັ່ງໝາດ



ក្រុមហ៊ុនកំណើនគ្រប់គ្រង

3. Logistic

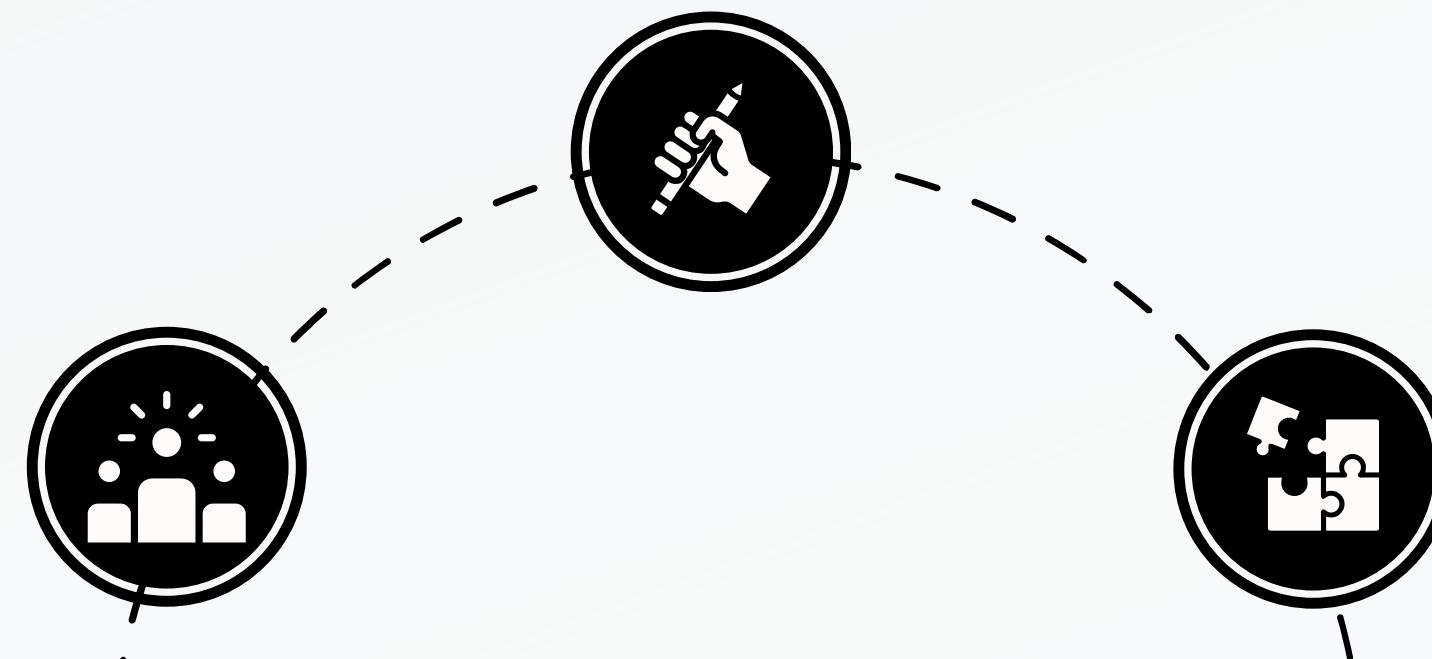
หนึ่งในอัลกอริทึมทางสถิติและเครื่องมือการเรียนรู้เชิงสถิติก็ใช้ในงานทางการวิเคราะห์ข้อมูลและ
การจัดหมวดหมู่
ความน่าจะเป็นที่ตัวแปรตามจะอยู่ในหนึ่งในสองกลุ่มที่กำหนด โดยใช้การหาพิงก์ชันโลจิสติก
(logistic function) เพื่อแปลงผลลัพธ์เป็นความน่าจะเป็นที่อยู่ในช่วง $[0, 1]$



ក្រុមហ៊ុនកំណើនខ្មែរ

4. Prism

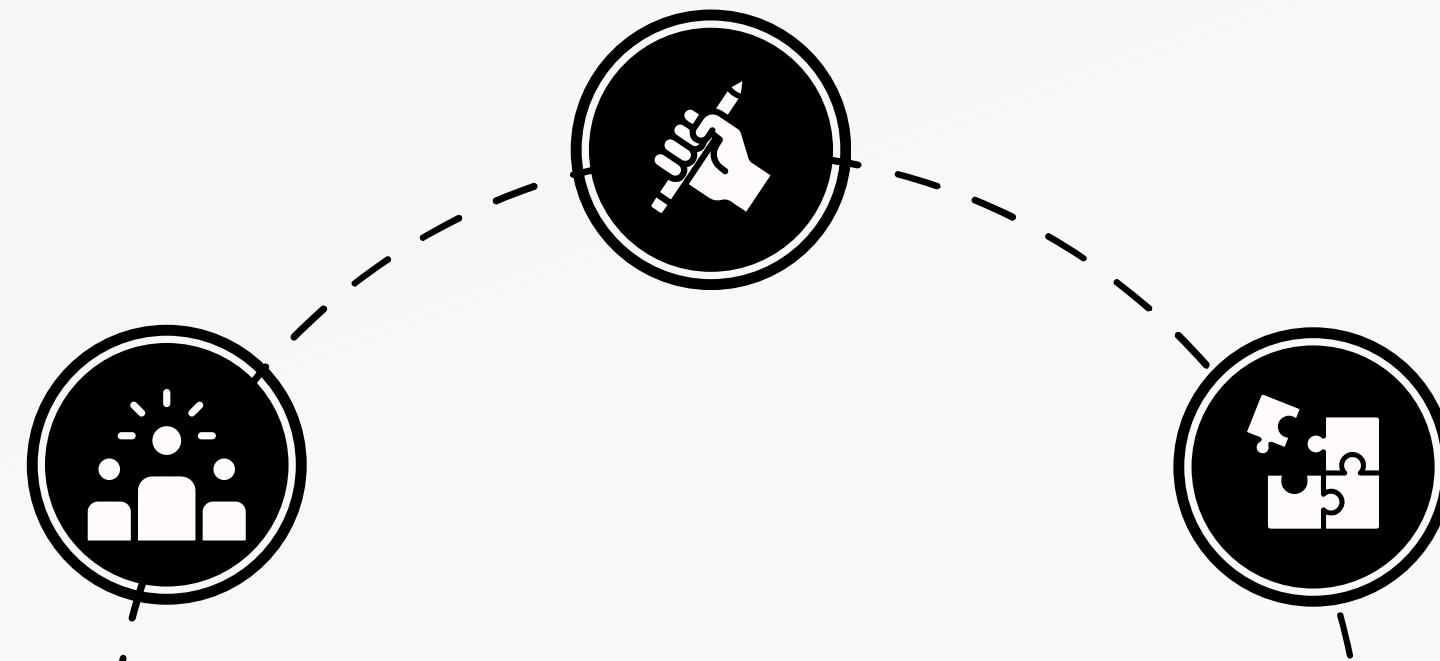
Prism Algorithm เป็นโมเดลที่พยายามหา Positive instances ซึ่ง Positive instances คือ Class ที่เราสนใจ ขณะนั้น เป็นการสร้าง Rule ในลักษณะ IF ? THEN ... (Class ที่เราสนใจ)... PRISM นั้นเป็นอัลกอริทึมที่ใช้ในการประมวลผลข้อมูลลำดับ (sequential data) โดยมุ่งเน้นที่กระบวนการ (process-oriented) และตัวอย่าง (instance-based) เพื่อสร้างโมเดลที่มีความสามารถในการกำหนดเหตุการณ์ตามลำดับของข้อมูล.



ក្រុមភីក់កៀវិយគមខោង

5. Decision Tree

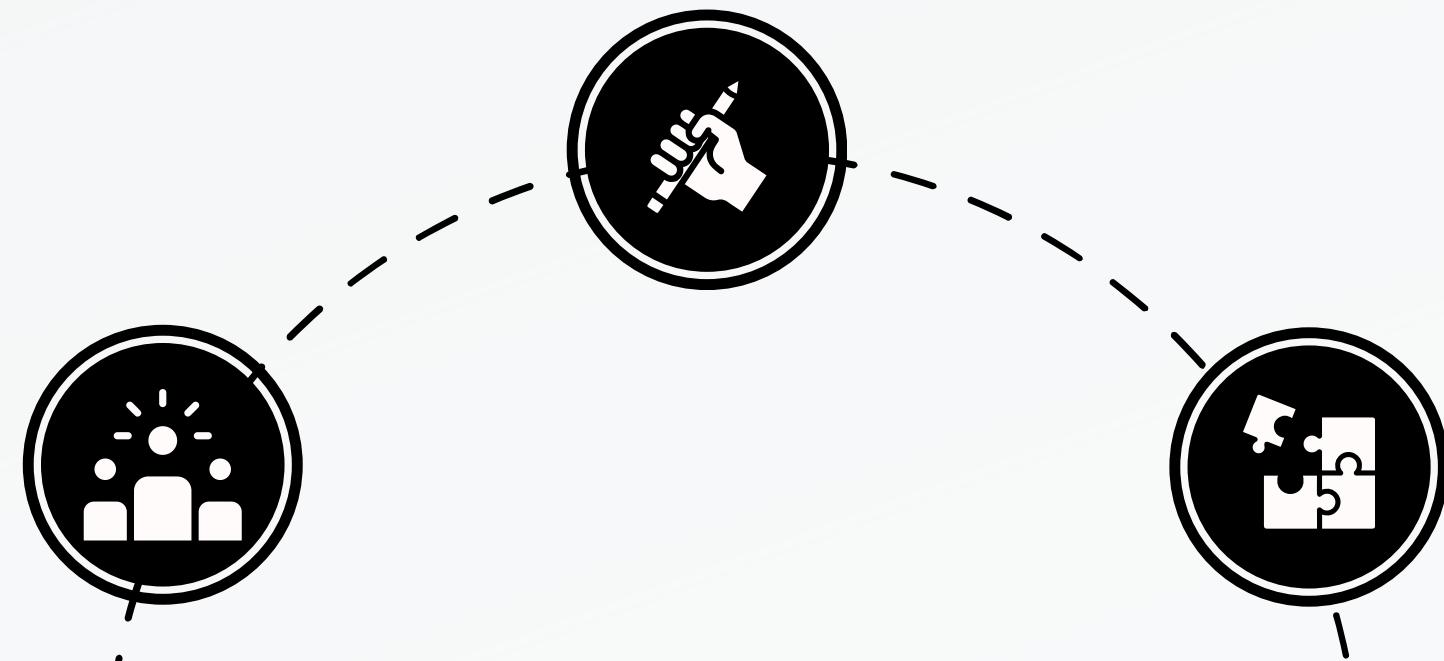
ต้นไม้ตัดสินใจ (Decision Tree) เป็นแบบจำลองการแบ่งข้อมูลออกเป็นกลุ่มย่อยๆ จนเกิดโครงสร้างต้นไม้ที่มีการตัดสินใจเบื้องต้น ซึ่งช่วยในการทำนายผลลัพธ์หรือการตัดสินใจเชิงต่าง ๆ โดยเฉพาะในการจำแนกประเภทของข้อมูล (classification)



ຖរម្មត់កែវប៉ុង

6. Bayes

ត្រូវបានគេប្រើបានជាការស្ថិតិយវត្ថុ ដើម្បីពន្លាសំណងជាមួយនឹងការស្វែងរក។



ក្រុមហ៊ុនកំណើនខ្មែរ

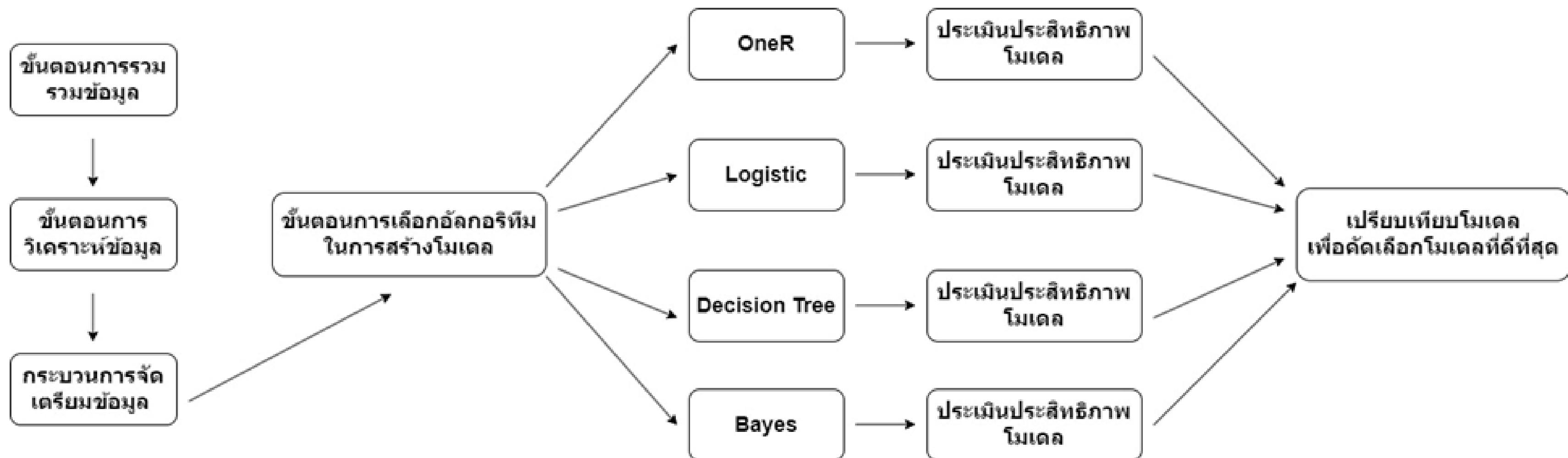
7. Evaluation

เป็นกระบวนการที่ใช้ในการวัดและประเมินประสิทธิภาพของแบบจำลอง เพื่อให้เข้าใจคุณภาพและประสิทธิภาพของการวิเคราะห์หรือการนำนัยที่ทำได้

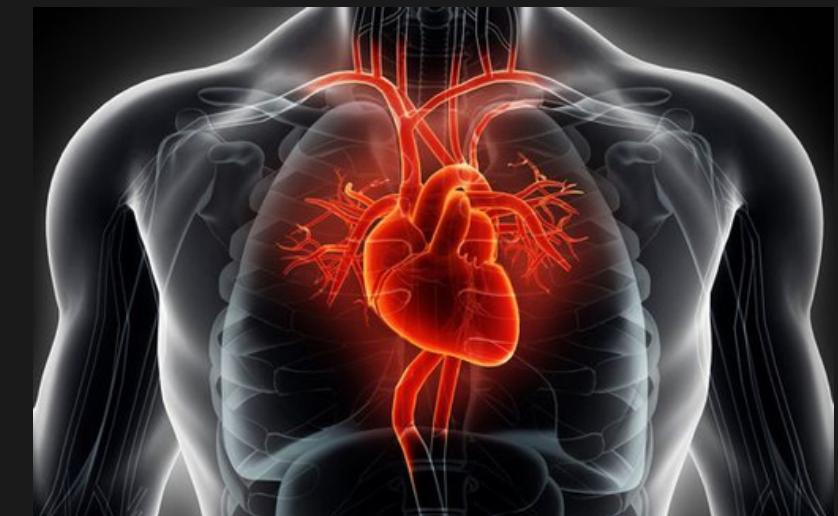


การอุกແບບ

แผนภาพกระบวนการทำงาน



กระบวนการทำนาย

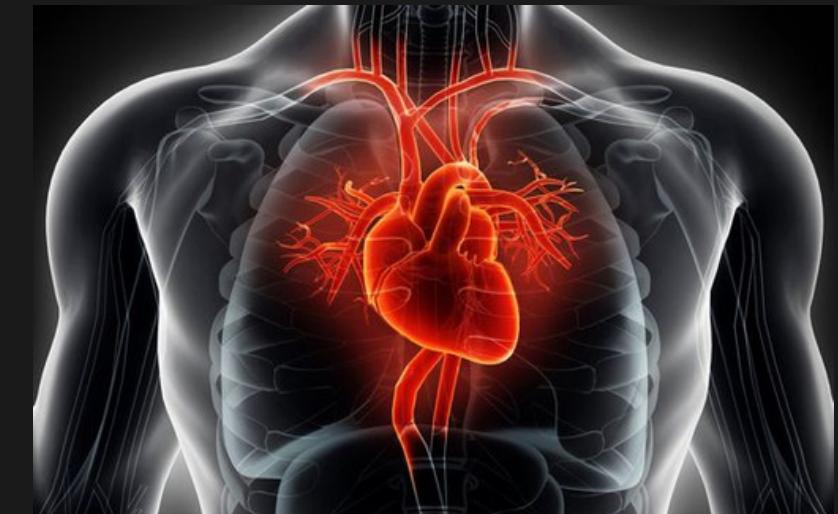


ขั้นตอนการรวมข้อมูล

- ข้อมูลการวิเคราะห์และการทำนายอาการหัวใจวาย (Heart Attack Analysis & Prediction Dataset) กลุ่มของพวคเราได้นำมาจาก เว็บไซต์ของ kaggle

The screenshot shows a web browser displaying a Kaggle dataset page. The URL in the address bar is kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data. The left sidebar has a navigation menu with options: Create, Home, Competitions, Datasets (selected), Models, Code, Discussions, Learn, and More. The main content area features a title "Heart Attack Analysis & Prediction Dataset" and a subtitle "A dataset for heart attack classification". It includes a "Data Card", "Code (1097)", and "Discussion (35)". On the right side, there is a detailed anatomical diagram of the human heart and major blood vessels. At the top right, there are download and notebook creation buttons, and a user profile icon.

ระบบงานการทำงาน



ขั้นตอนการวิเคราะห์ข้อมูล

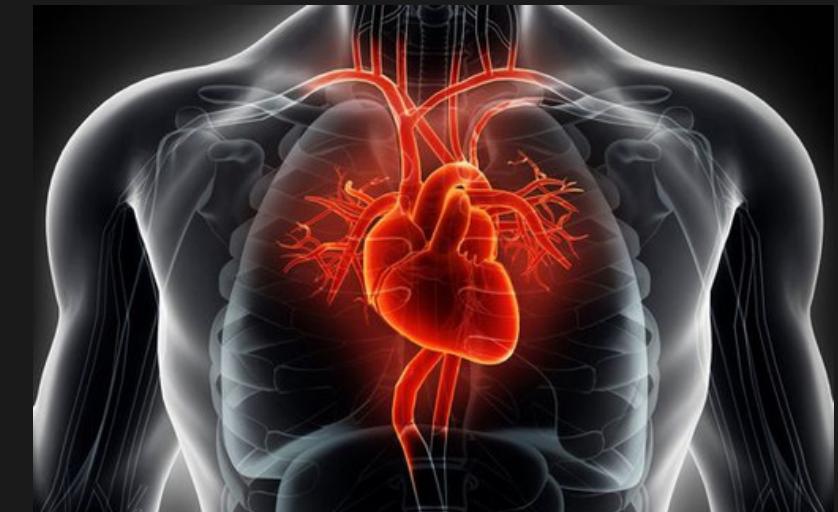


- ลักษณะข้อมูลที่นำมาใช้ในการวิเคราะห์และการทำนายอาการหัวใจวาย เป็นชุดข้อมูลที่ประกอบด้วย 14 attributes 303 instances แต่ละ attributes คือ age, sex, cp, trtbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall, output

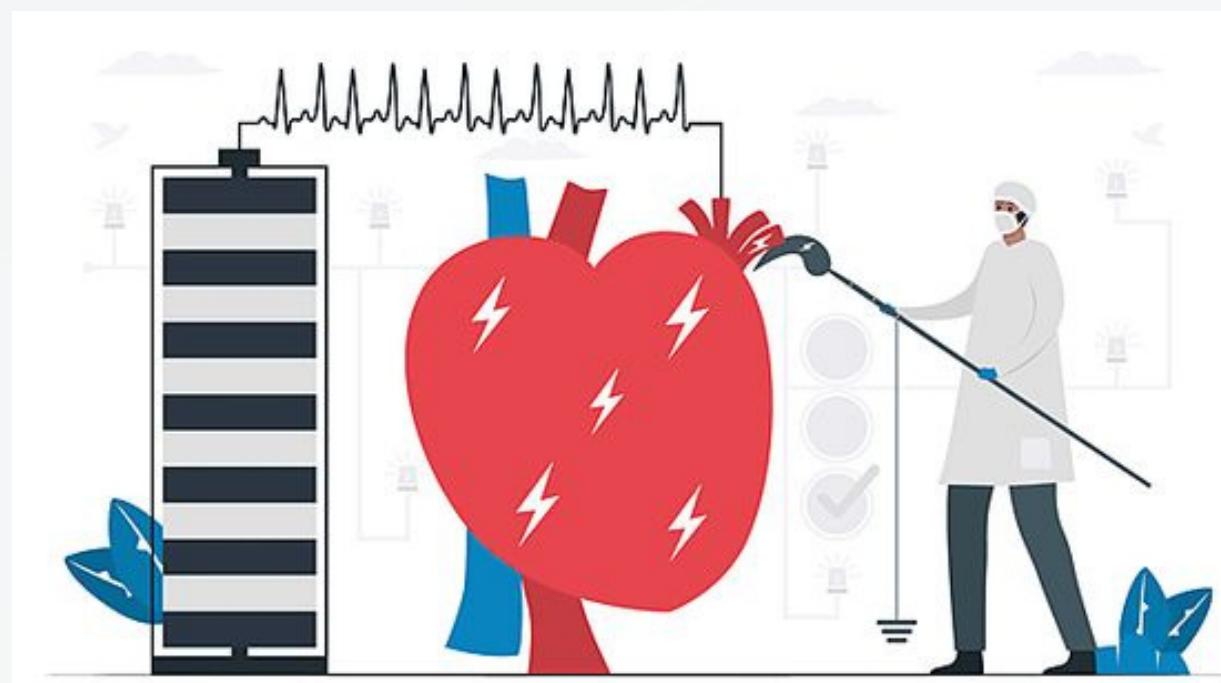
-age อายุของผู้ป่วย
-sex เพศของผู้ป่วย
-cp อาการเจ็บหน้าอกร ประเภทอาการเจ็บหน้าอกร
-trtbps ความดันโลหิตขณะพัก (เป็น mm Hg)
-chol คอเลสเทอโรลในหน่วย mg/dl ที่ดึงมาจากเซ็นเซอร์ BMI
-fbs (น้ำตาลในเลือดขณะอดอาหาร > 120 mg/dl) (1 = true; 0 = false)
-restecg ผลการตรวจคลื่นไฟฟ้าหัวใจขณะพัก

-thalachh อัตราการเต้นของหัวใจสูงสุดที่ได้รับ
-exng เจ็บแปลบจากการอักเสบกล้ามเนื้อ (1 = ใช่; 0 = ไม่ใช่)
oldpeak ค่า ST depression ทำให้เกิดการตีบตันของหัวใจเมื่อมีการอักเสบกล้ามเนื้อ^{พ่อน}
-slp ค่าความชันเมื่ออักเสบกล้ามเนื้อระดับเข้มข้น
-caa จำนวนหลอดเลือดหลักที่ผ่านการตรวจด้วยเครื่องเอ็กซเรย์แบบที่ใช้ร่วมกับสารกึบแสง
-thall การตรวจหัวใจโดยสารกึบแสง
-output เป้าหมาย : 0= มีโอกาสหัวใจวายน้อยลง 1= มีโอกาสหัวใจวายมากขึ้น

กระบวนการทํางาน

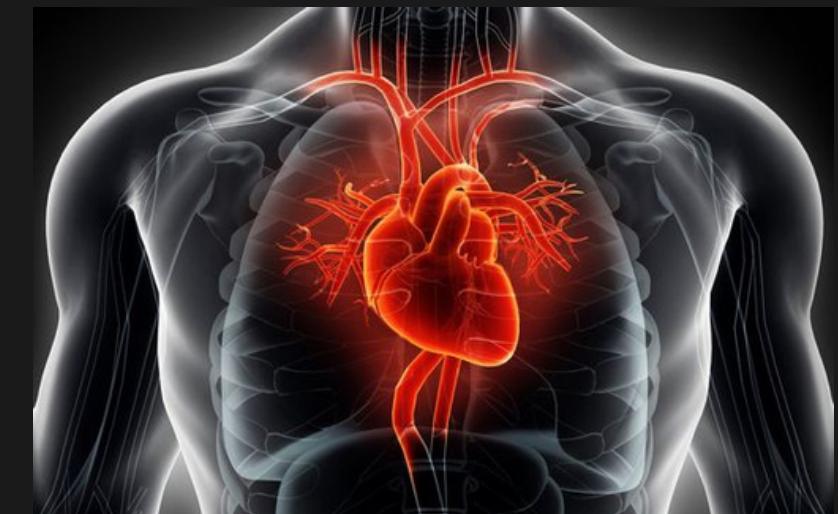


กระบวนการจัดเตรียมข้อมูล



- ทุก attributes มีความสำคัญ และมีผลต่อการสร้างโมเดล จากนั้นจึงได้ทำการเปลี่ยนประเภทของเข้ามูล แต่ละ attributes ให้เหมาะสมต่อการนำไปสร้างโมเดล

กระบวนการทำนาย



กระบวนการจัดเตรียมข้อมูล

- age, trtbps, chol, thalachh, caa = integer
- sex, cp, fbs, restecg, exng, slp, thall, output = polynomial
- oldpeak = real

Import Data - Format your columns.

Format your columns.

Date format Replace errors with missing values

	age integer	sex polynomial	cp polynomial	trtbps integer	chol integer	fbs polynomial	restecg polynomial	thalachh integer
1	63	1	3	145	233	1	0	150
2	37	1	2	130	250	0	1	187
3	41	0	1	130	204	0	0	172
4	56	1	1	120	236	0	1	178
5	57	0	0	120	354	0	1	163
6	57	1	0	140	192	0	1	148
7	56	0	1	140	294	0	0	153
8	44	1	1	120	263	0	1	173
9	52	1	2	172	199	1	1	162
10	57	1	2	150	168	0	1	174
11	54	1	0	140	239	0	1	160
12	48	0	2	130	275	0	1	139
13	49	1	1	130	266	0	1	171
14	64	1	3	110	211	0	0	144
15	58	0	3	150	283	1	0	162
16	50	0	2	120	219	0	1	158
17	58	0	2	120	340	0	1	172

no problems. Previous Next Cancel

Import Data - Format your columns.

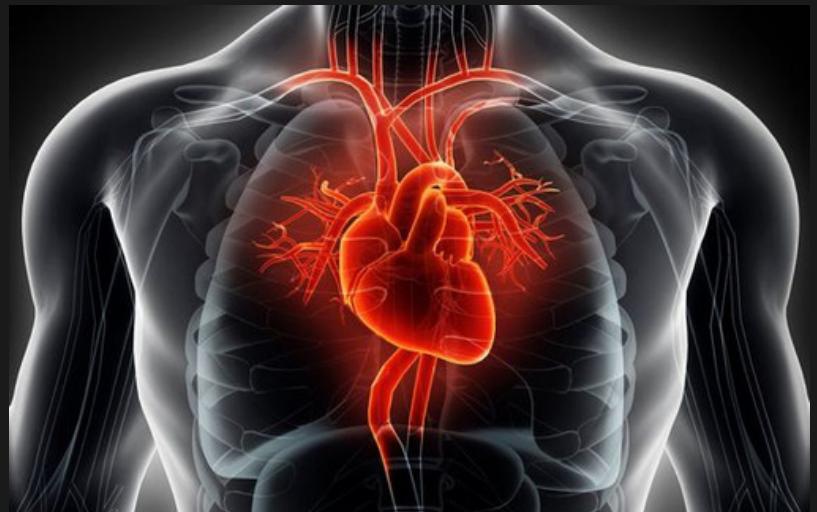
Format your columns.

Date format Replace errors with missing values

	cg ominal	thalachh integer	exng polynomial	oldpeak real	slp polynomial	caa integer	thall polynomial	output polynomial
1		150	0	2.300	0	0	1	1
2		187	0	3.500	0	0	2	1
3		172	0	1.400	2	0	2	1
4		178	0	0.800	2	0	2	1
5		163	1	0.600	2	0	2	1
6		148	0	0.400	1	0	1	1
7		153	0	1.300	1	0	2	1
8		173	0	0.000	2	0	3	1
9		162	0	0.500	2	0	3	1
10		174	0	1.600	2	0	2	1
11		160	0	1.200	2	0	2	1
12		139	0	0.200	2	0	2	1
13		171	0	0.600	2	0	2	1
14		144	1	1.800	1	0	2	1
15		162	0	1.000	2	0	2	1
16		158	0	1.600	1	0	2	1
17		172	0	0.000	2	0	2	1

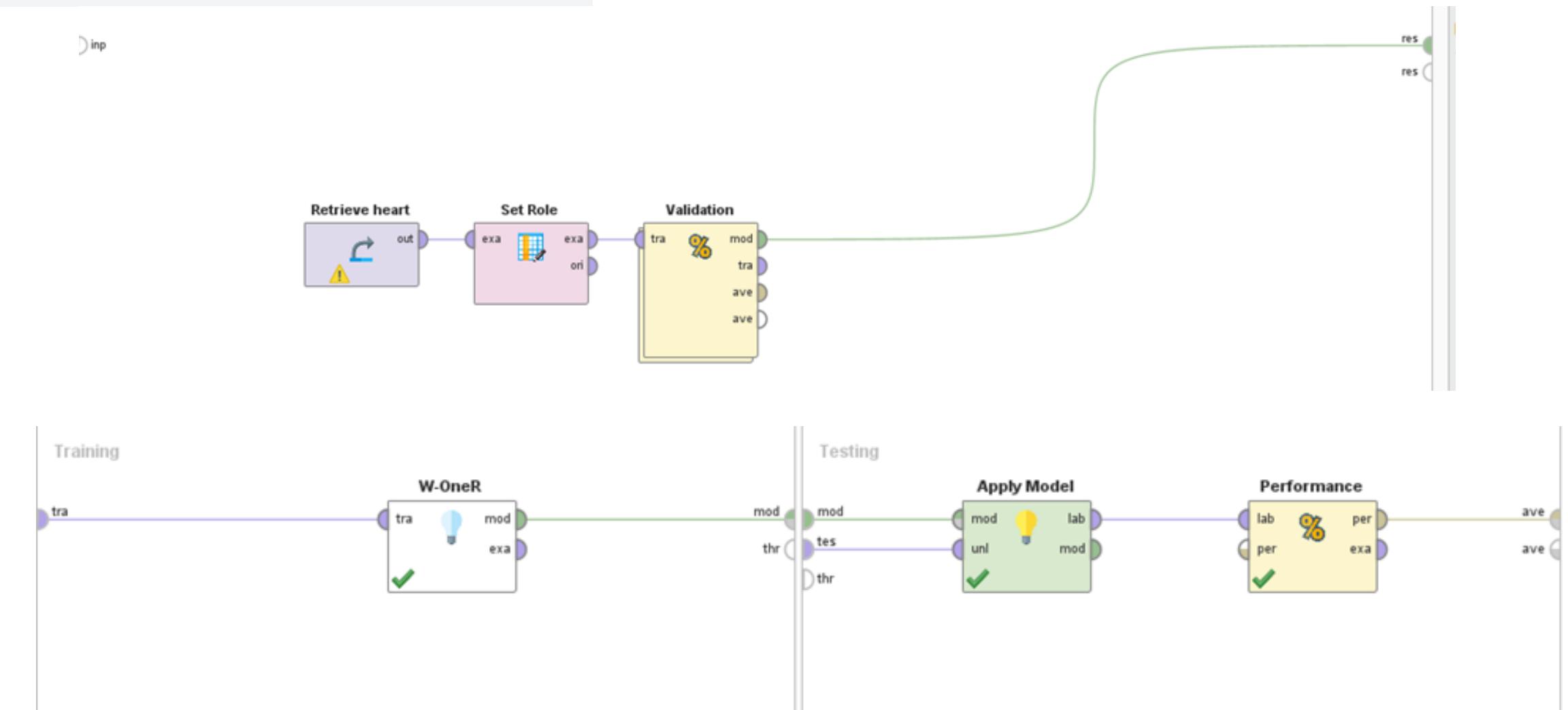
no problems. Previous Next Cancel

กระบวนการทํางาน



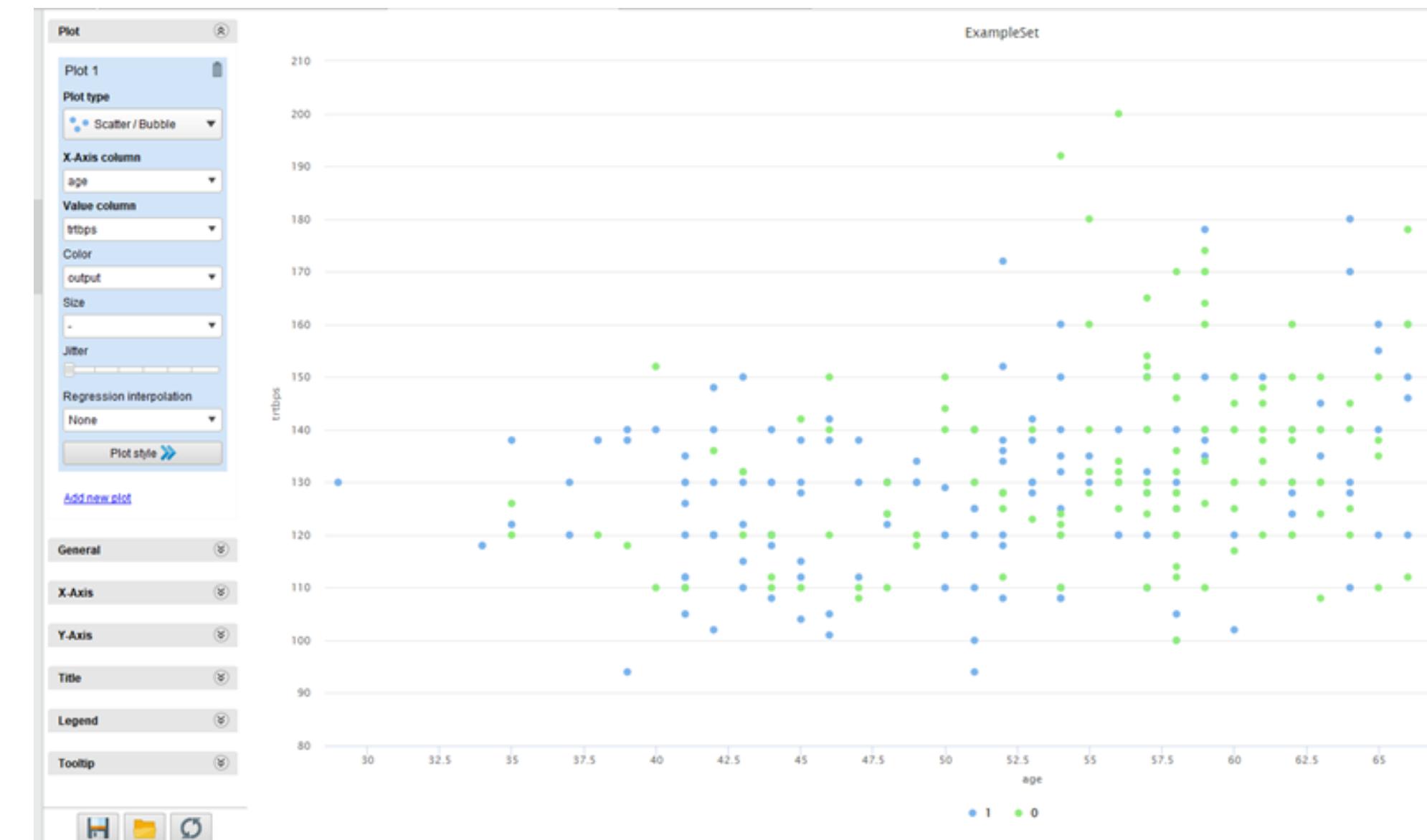
ขั้นตอนการเลือกอัลกอริทึมในการสร้างโมเดล

OneR



ONER

ผลการดำเนินการ



ONER

Model

Result History % PerformanceVector (Performance)

Description

Weka result

W-OneR

thall:

1	-> 0
2	-> 1
3	-> 0
0	-> 1

(232/303 instances correct)

ONER

Confusion Matrix Table

accuracy: 70.33%

	true 1	true 0	class precision
pred. 1	36	13	73.47%
pred. 0	14	28	66.67%
class recall	72.00%	68.29%	

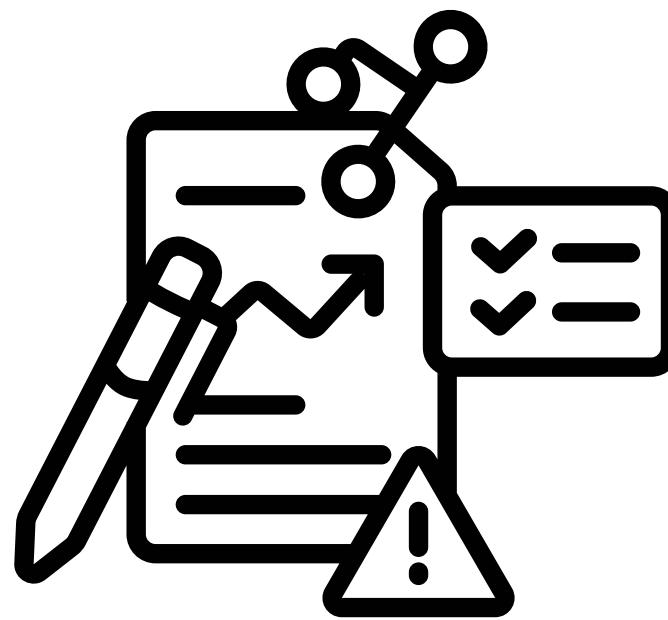
Performance Vector

Performance

PerformanceVector:
accuracy: 70.33%
ConfusionMatrix:
True: 1 0
1: 36 13
0: 14 28
precision: 66.67% (positive class: 0)
ConfusionMatrix:
True: 1 0
1: 36 13
0: 14 28
recall: 68.29% (positive class: 0)
ConfusionMatrix:
True: 1 0
1: 36 13
0: 14 28
AUC (optimistic): 0.911 (positive class: 0)
AUC: 0.500 (positive class: 0)
AUC (pessimistic): 0.492 (positive class: 0)

Description

Annotations

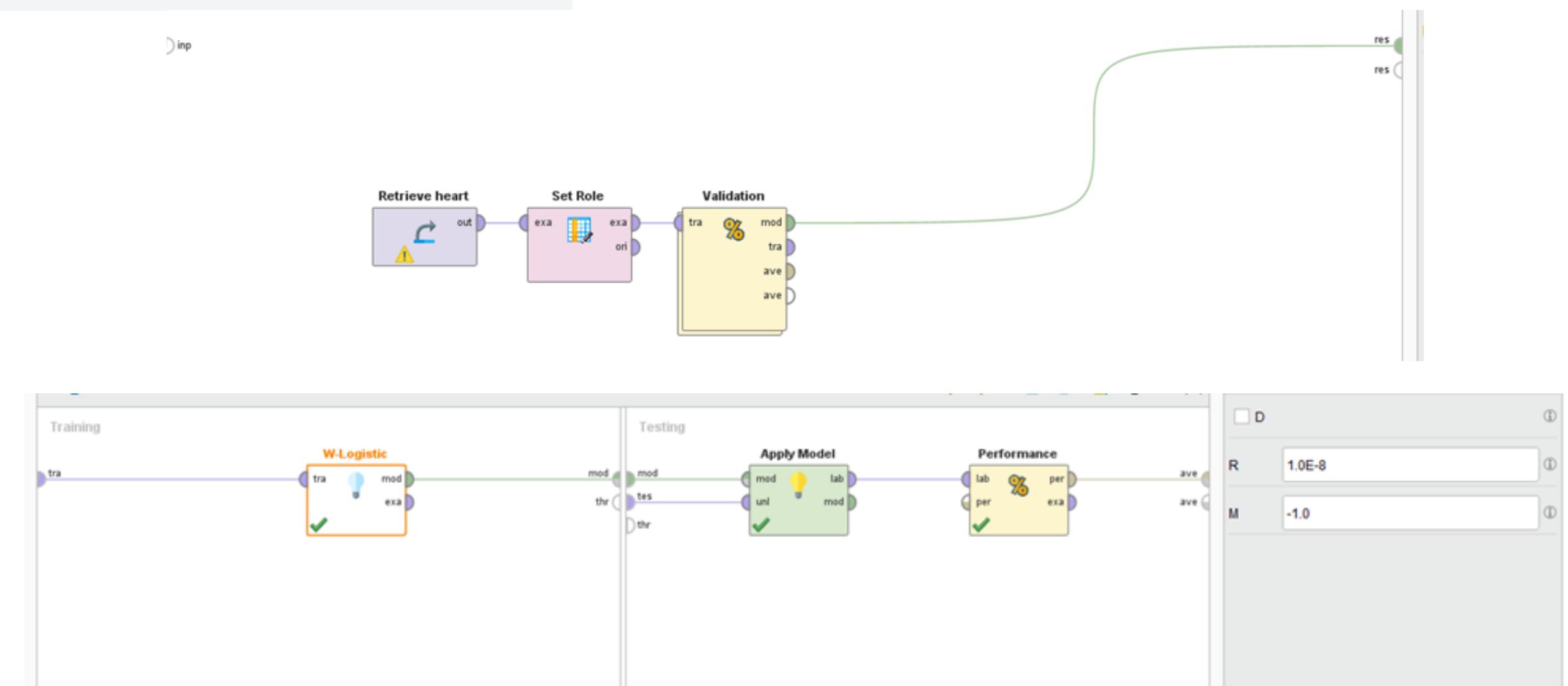


กระบวนการทำนาย



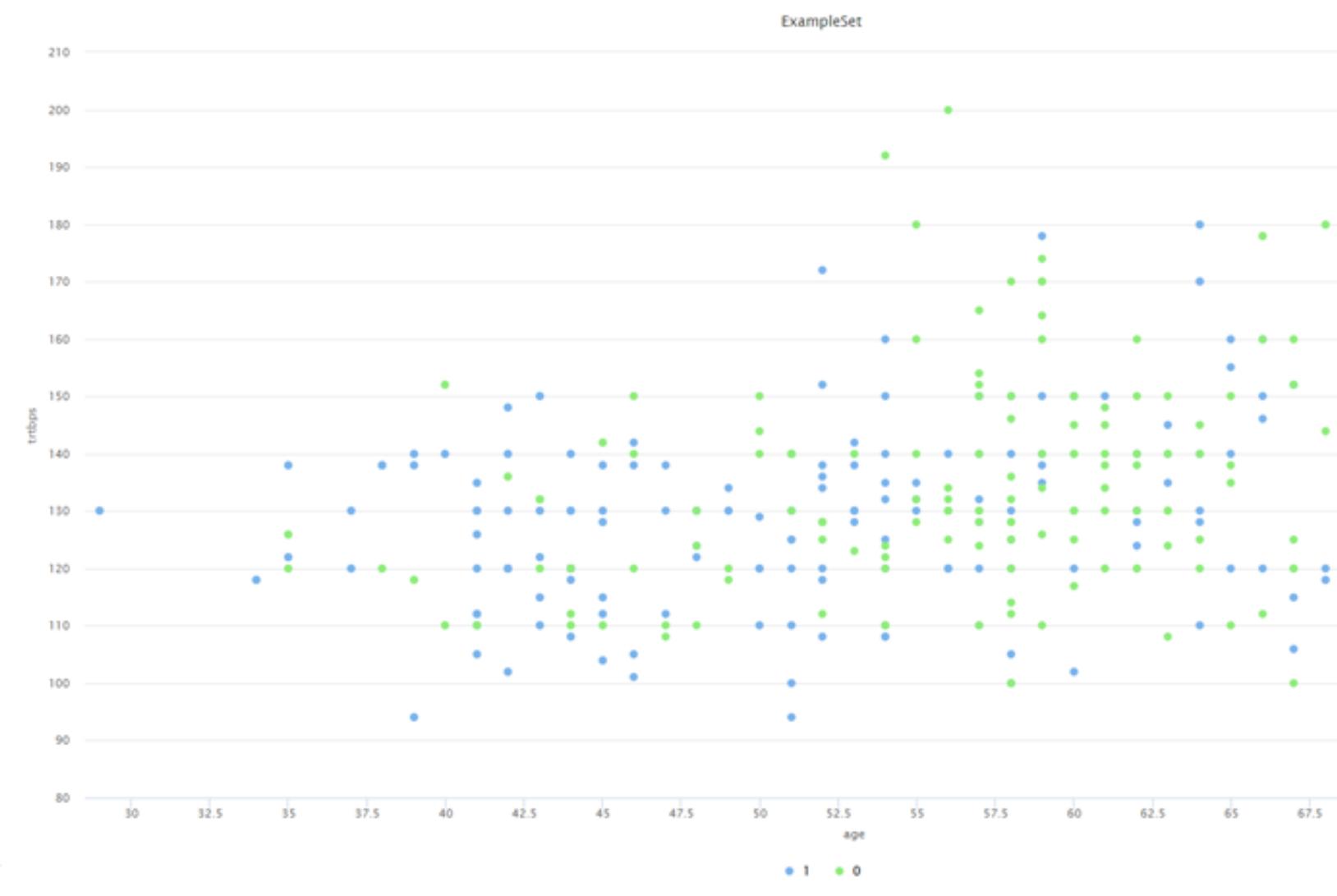
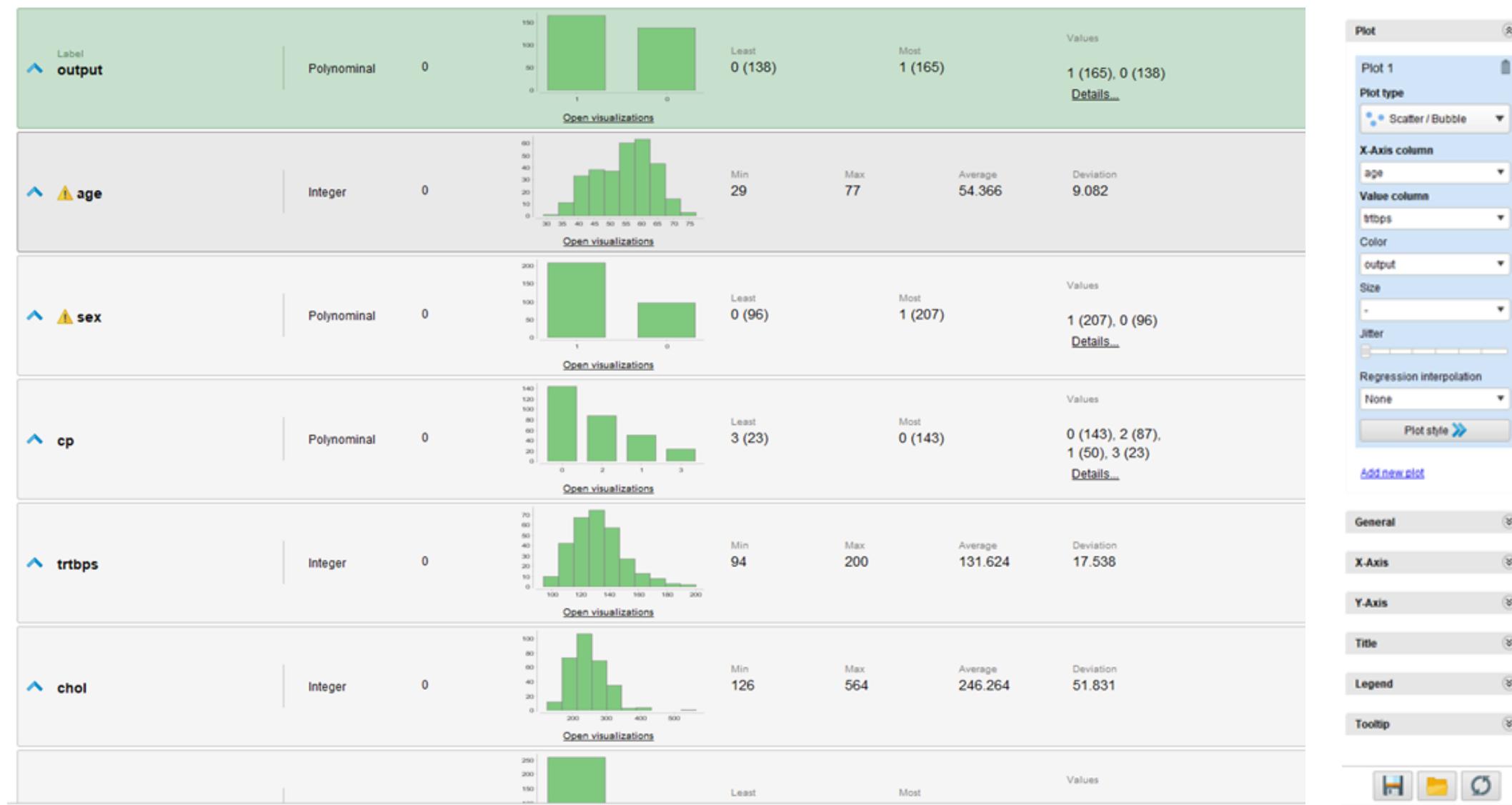
ขั้นตอนการเลือกอัลกอริทึมในการสร้างโมเดล

Logistic



LOGISTIC

ผลการดำเนินการ



LOGISTIC

Model

W-Logistic	
Logistic Regression with ridge parameter of 1.0E-8	
Description	Coefficients...
	Class
Weka result	Variable 1
	=====
	age -0.0006
	sex 1.515
	cp=3 0.996
	cp=2 0.9253
	cp=1 -0.0368
	cp=0 -1.02
	trtbps -0.0171
	chol -0.0043
	fbs -0.1764
	restecg=0 -0.271
	restecg=1 0.2993
	restecg=2 -0.5476
	thalachh 0.0171
	exng -0.7631
	oldpeak -0.4893
	s1p=0 0.2289
	s1p=2 0.4305
	s1p=1 -0.4907
	caa -0.8332
	thall=1 0.5854
	thall=2 0.624
	thall=3 -0.7561
	thall=0 -1.2293
	Intercept 2.0533
Odds Ratios...	
	Class
	Variable 1
	=====
	age 0.9994
	sex 4.5492
	cp=3 2.7074
	cp=2 2.5226
	cp=1 0.9639
	cp=0 0.3606
	trtbps 0.9831
	chol 0.9957
	fbs 0.8383
	restecg=0 0.7627
	restecg=1 1.3489
	restecg=2 0.5783
	thalachh 1.0173
	exng 0.4662
	oldpeak 0.6131
	s1p=0 1.2573
	s1p=2 1.538
	s1p=1 0.6122
	caa 0.4347
	thall=1 1.7956
	thall=2 1.8664
	thall=3 0.4695
	thall=0 0.2925

LOGISTIC

Confusion
Matrix Table

Criterion

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

accuracy: 80.22%

	true 1	true 0	class precision
pred. 1	43	11	79.63%
pred. 0	7	30	81.08%
class recall	86.00%	73.17%	

Performance
Vector

%

Performance

PerformanceVector:

```
accuracy: 80.22%
ConfusionMatrix:
True: 1 0
1: 43 11
0: 7 30
precision: 81.08% (positive class: 0)
```

Description

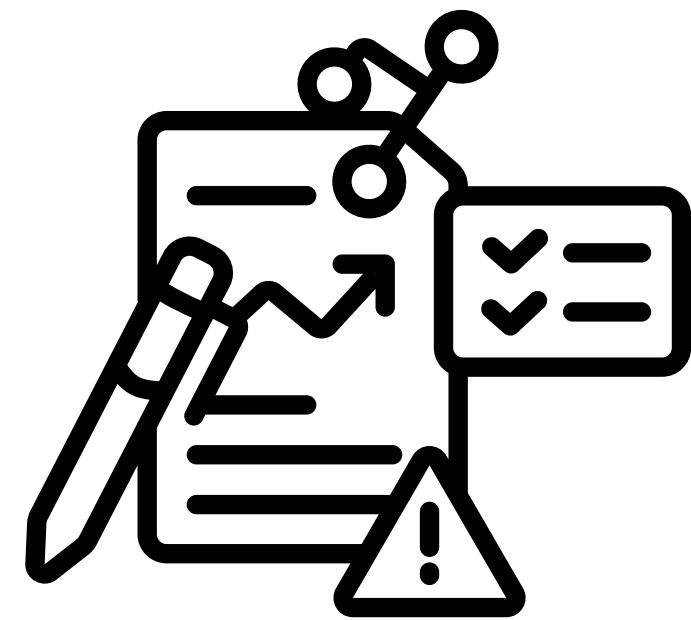
Annotations

ConfusionMatrix:

```
True: 1 0
1: 43 11
0: 7 30
recall: 73.17% (positive class: 0)
```

ConfusionMatrix:

```
True: 1 0
1: 43 11
0: 7 30
AUC (optimistic): 0.861 (positive class: 0)
AUC: 0.861 (positive class: 0)
AUC (pessimistic): 0.861 (positive class: 0)
```



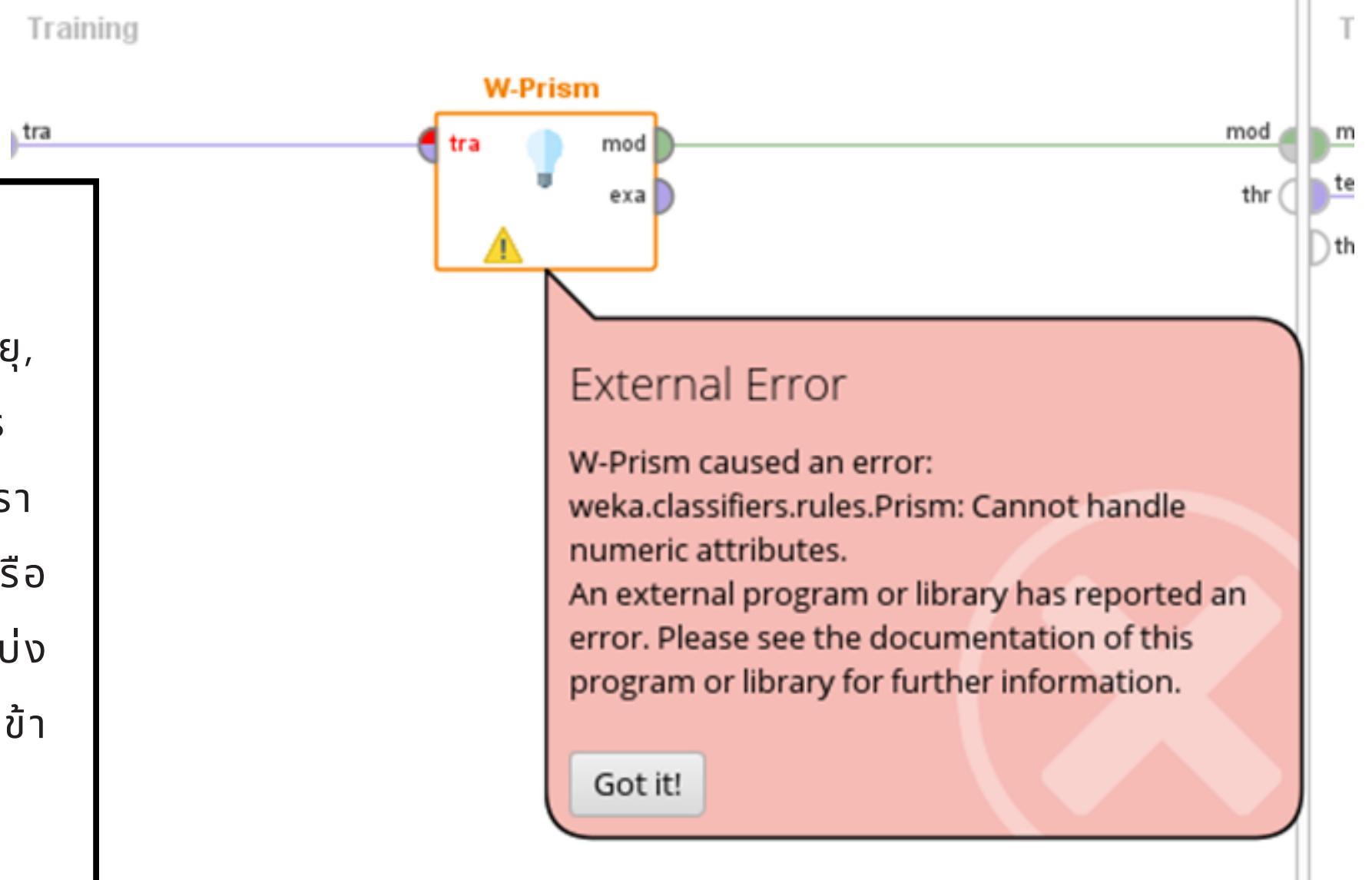
กระบวนการทำนาย



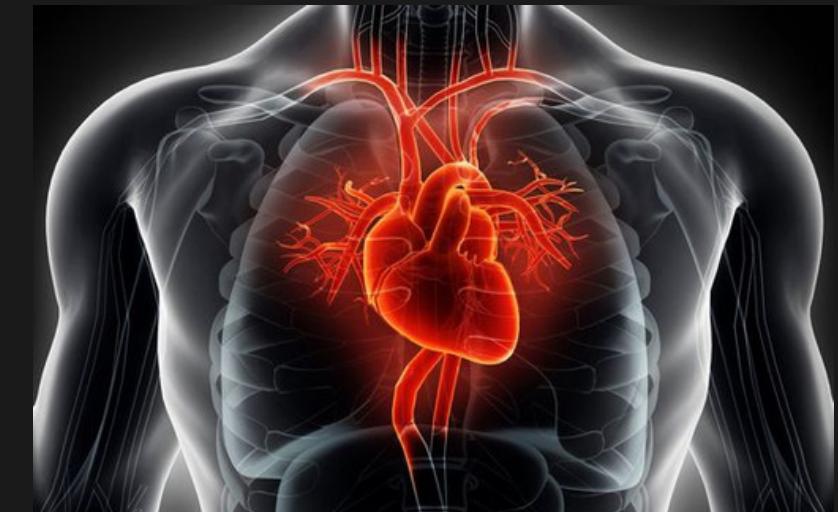
ขั้นตอนการเลือกอัลกอริทึมในการสร้างโมเดล

Prism

ทดลองใช้ Prism กับข้อมูล ผลที่ได้ไม่สามารถหาข้อพิสูจน์ได้ เนื่องจาก Prism ไม่สามารถทำงานกับ ข้อมูลตัวเลขที่มีค่าต่อเนื่อง เช่น อายุ, อัตราการเต้นของหัวใจ, ปริมาณคอเลสเตอรอลในเลือด , old peak การ จำแนกข้อมูลชุดนี้มาใช้กับ Prism จำเป็นต้องทำการแบ่งช่วงของข้อมูลซึ่งเรา จำเป็นต้องทราบถึงความเกี่ยวข้องกันของข้อมูลเพื่อจะทำการแบ่งกลุ่ม หรือ Clustering ให้ข้อมูลมีความเข้ากันได้ต้องทราบถึงจำนวนของกลุ่มที่จะแบ่ง มาใช้กับ Prism ทางผู้จัดทำจึง ไม่ได้นำ Prism มาใช้งานเนื่องด้วยความเข้า กันของตัวชุดข้อมูลและ Algorithm

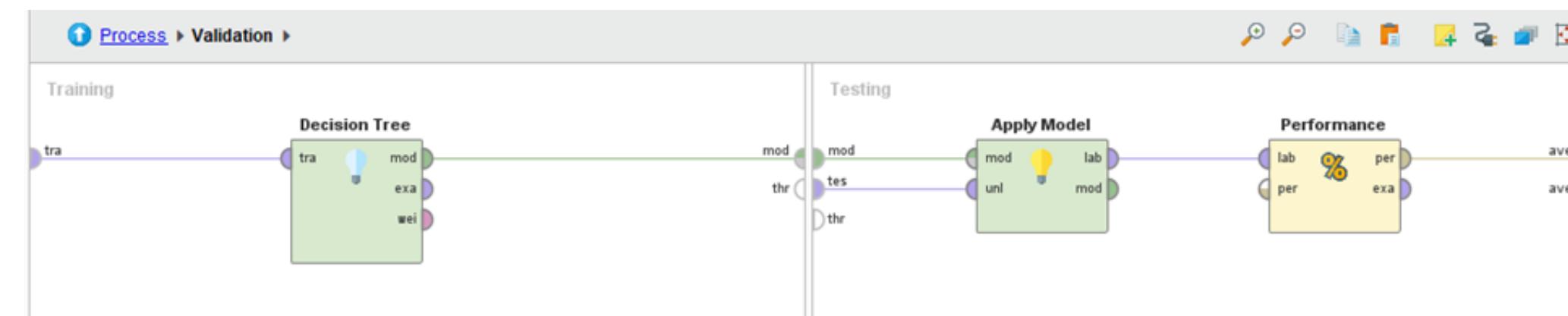
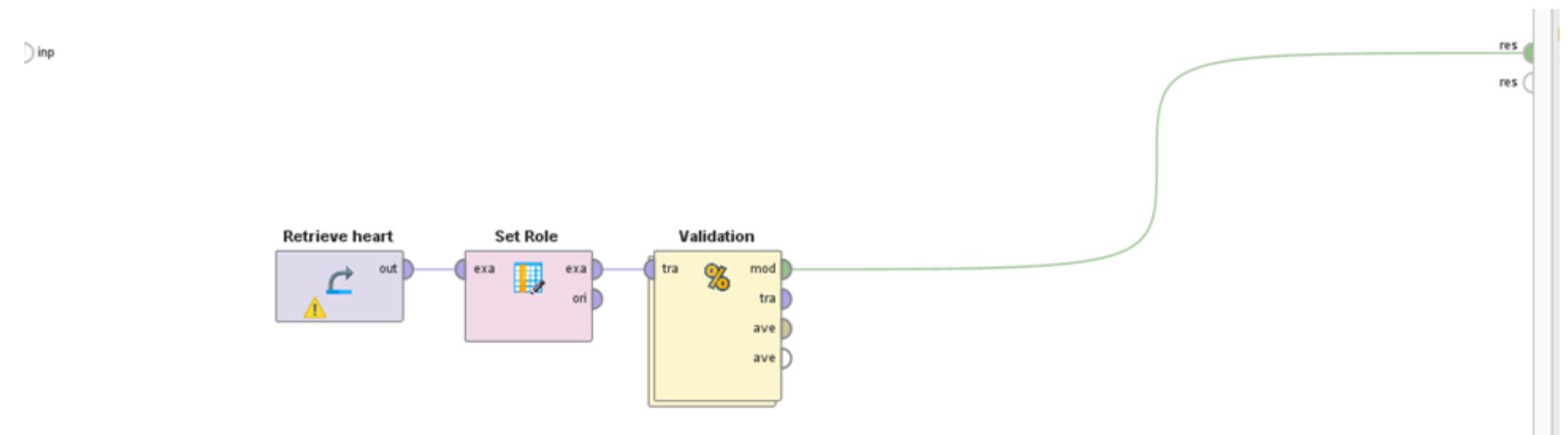


กระบวนการทำนาย



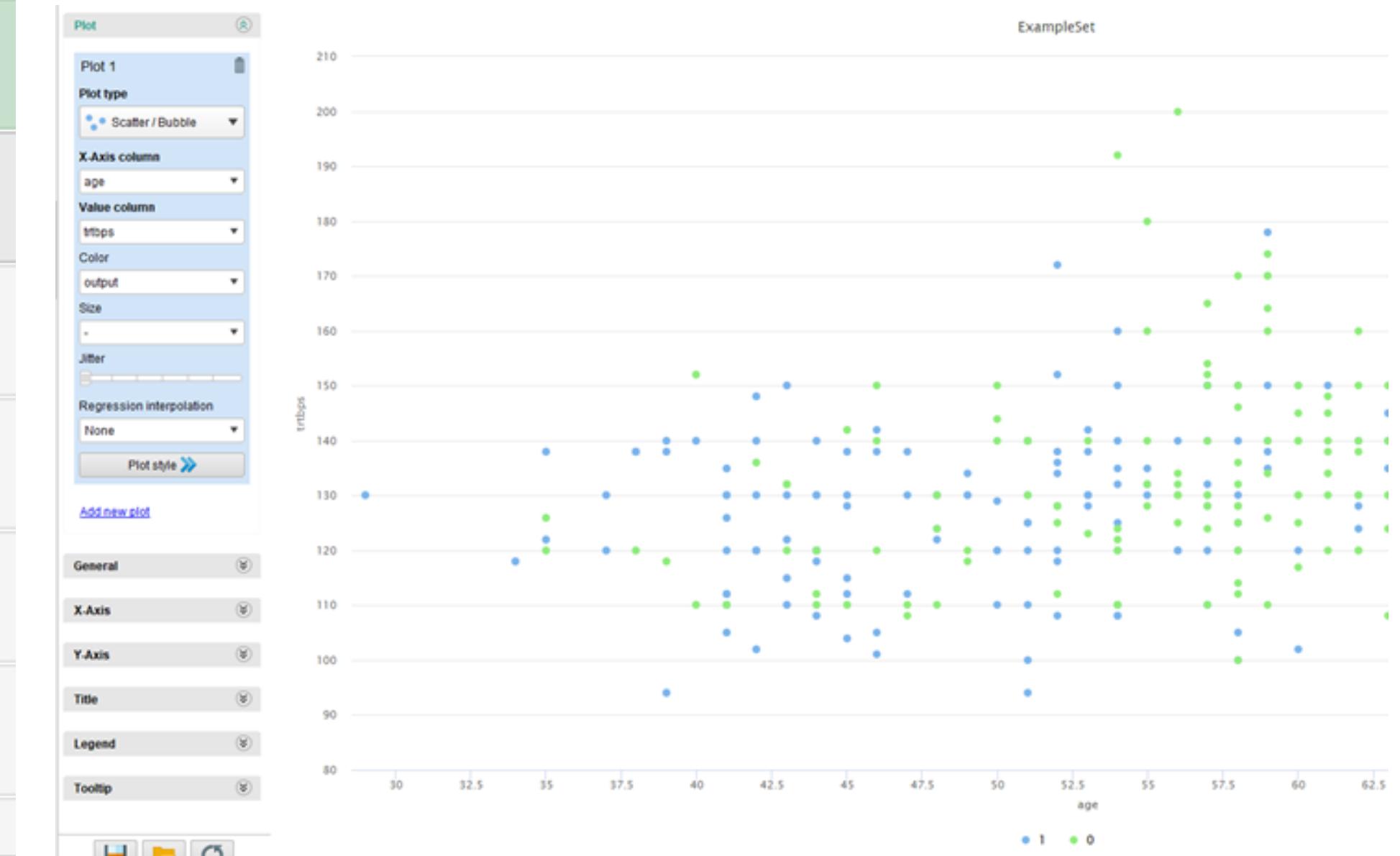
ขั้นตอนการเลือกอัลกอริทึมในการสร้างโมเดล

Decision Tree



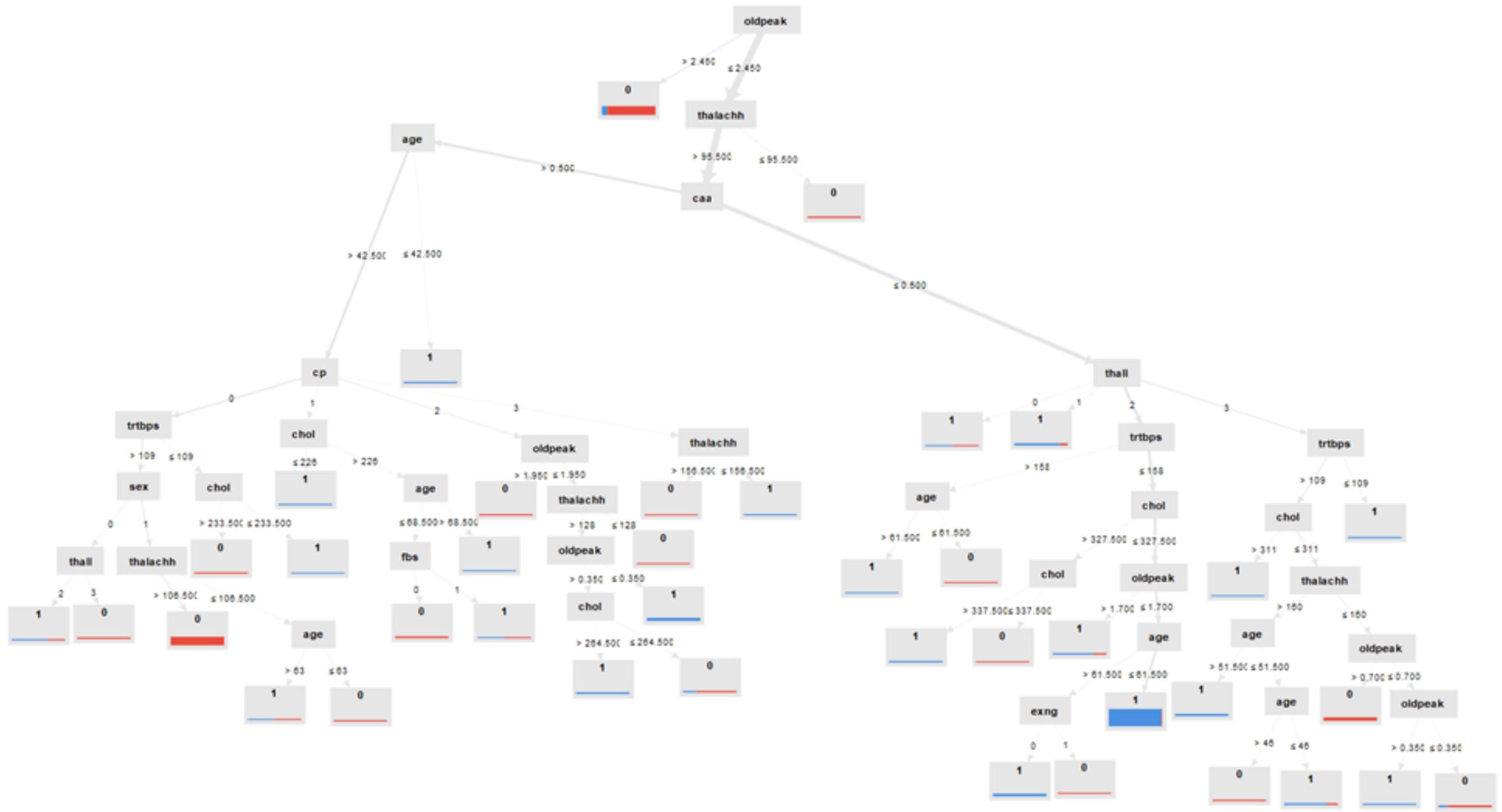
DECISION TREE

ผลการดำเนินการ



DECISION TREE

Model



DECISION TREE

Confusion Matrix Table

Criterion

- Table View Plot View
- accuracy
- precision
- recall
- AUC (optimistic)
- AUC
- AUC (pessimistic)

accuracy: 78.02%

	true 1	true 0	class precision
pred. 1	42	12	77.78%
pred. 0	8	29	78.38%
class recall	84.00%	70.73%	

Performance Vector

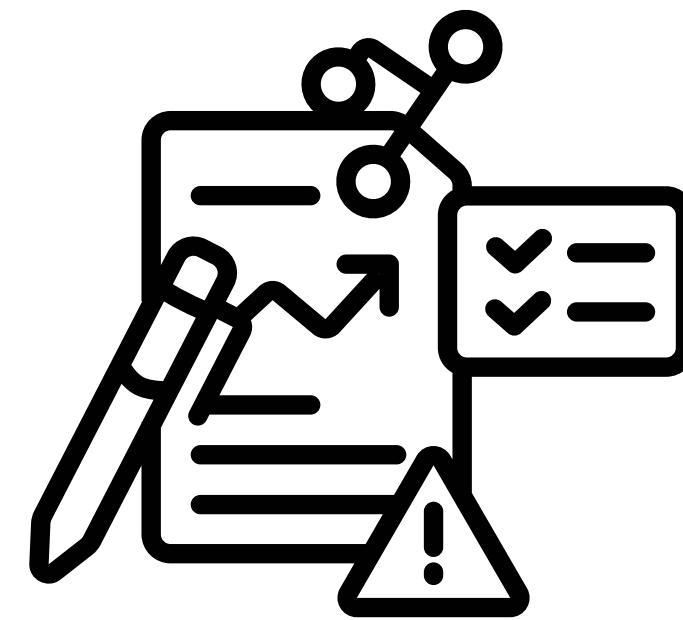
%

Performance

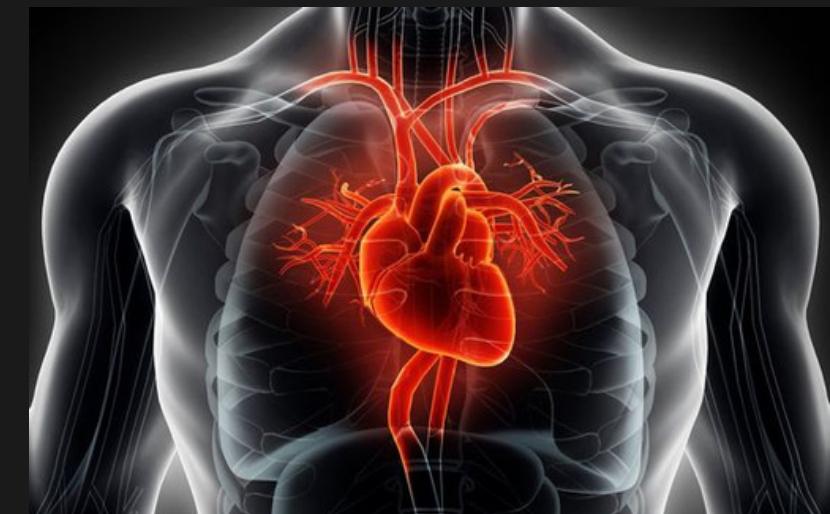
PerformanceVector:
accuracy: 78.02%
ConfusionMatrix:
True: 1 0
1: 42 12
0: 8 29
precision: 78.38% (positive class: 0)
ConfusionMatrix:
True: 1 0
1: 42 12
0: 8 29
recall: 70.73% (positive class: 0)
ConfusionMatrix:
True: 1 0
1: 42 12
0: 8 29
AUC (optimistic): 0.810 (positive class: 0)
AUC: 0.740 (positive class: 0)
AUC (pessimistic): 0.699 (positive class: 0)

Description

Annotations

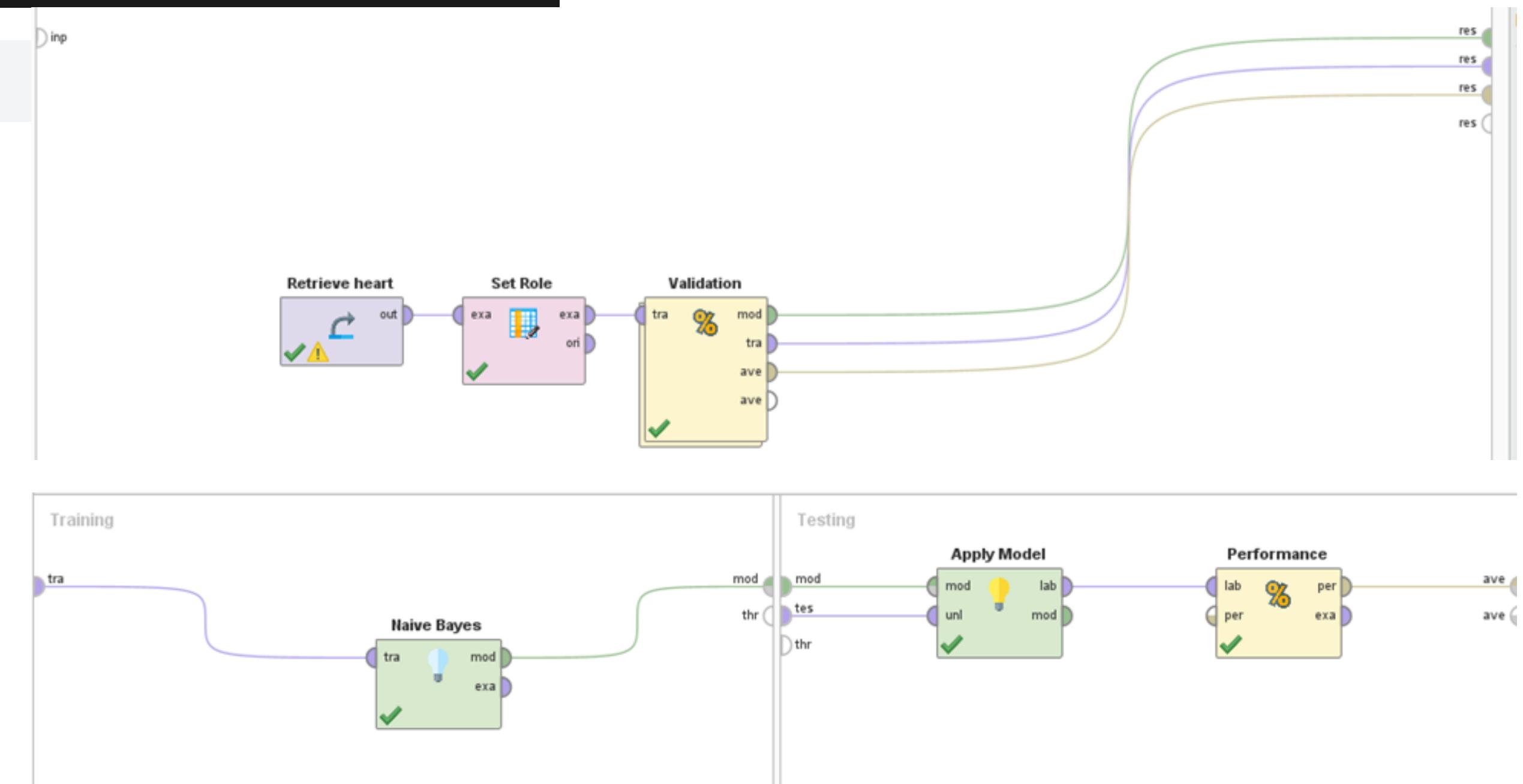


กระบวนการทํางาน



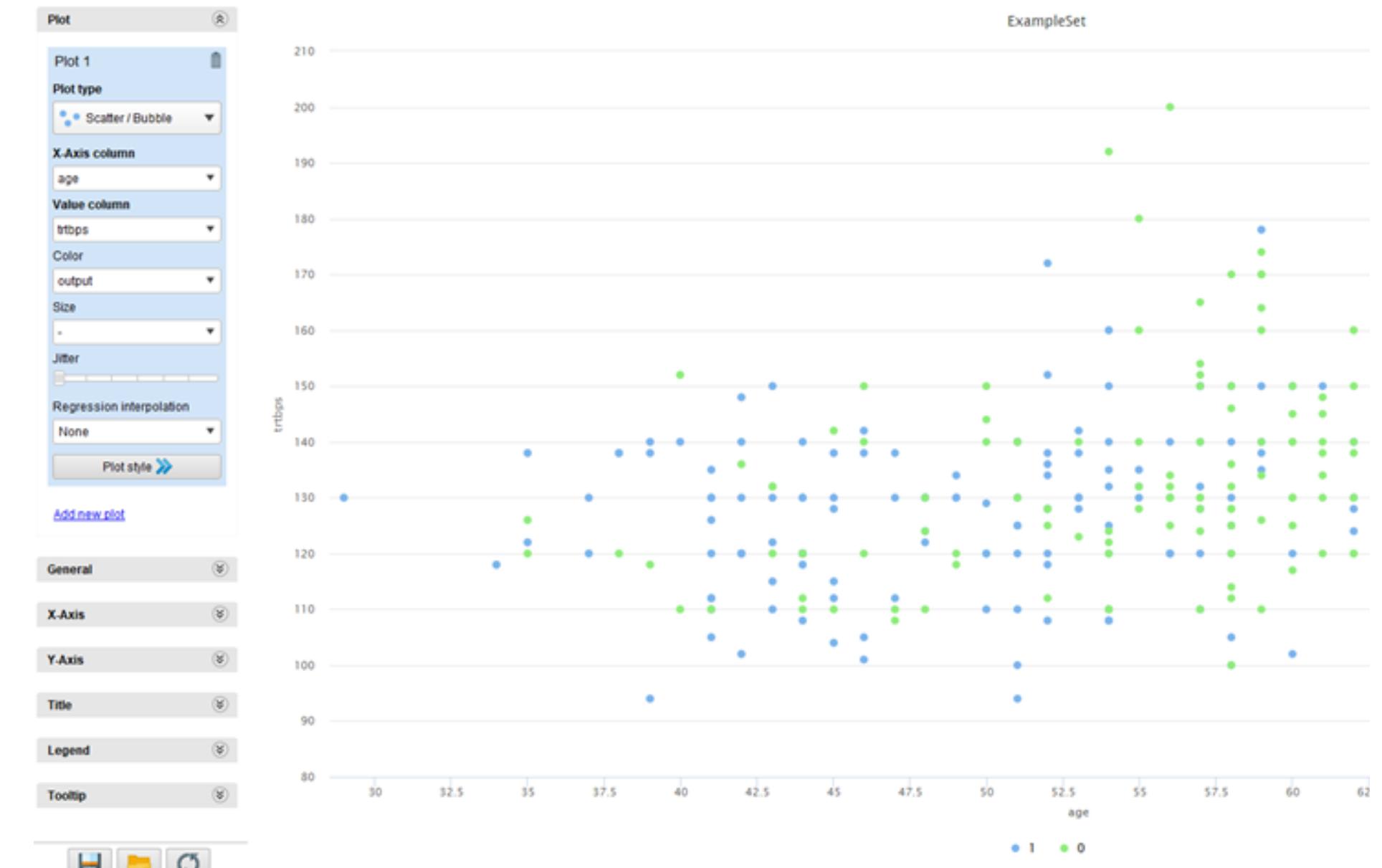
ขั้นตอนการเลือกอัลกอริทึมในการสร้างโมเดล

Bayes



BAYES

ผลการดำเนินการ



BAYES

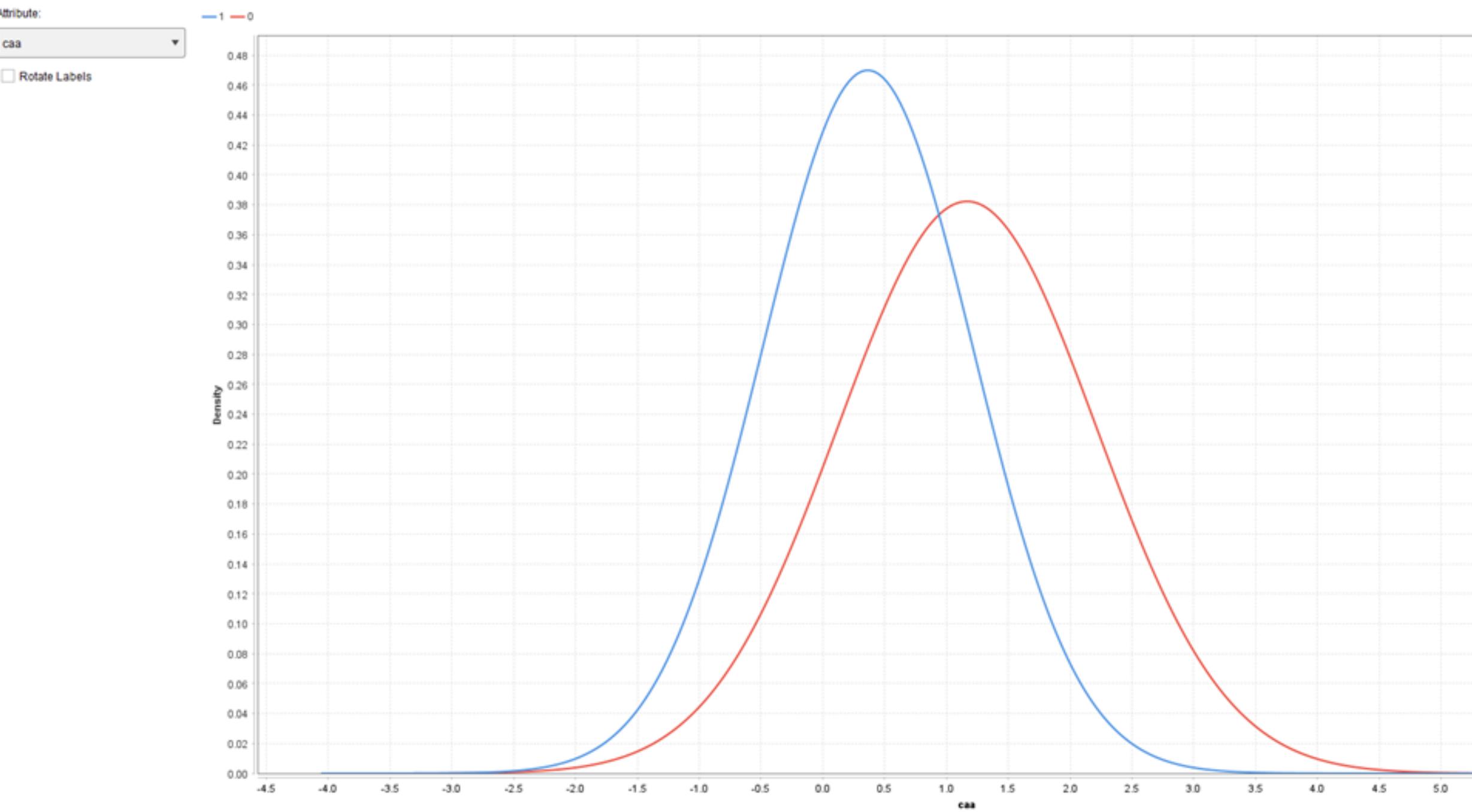
Model

SimpleDistribution

Distribution model for label attribute output

Class 1 (0.545)
13 distributions

Class 0 (0.455)
13 distributions



BAYES

Confusion Matrix Table

Criterion

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

accuracy: 78.02%

	true 1	true 0	class precision
pred. 1	40	10	80.00%
pred. 0	10	31	75.61%
class recall	80.00%	75.61%	

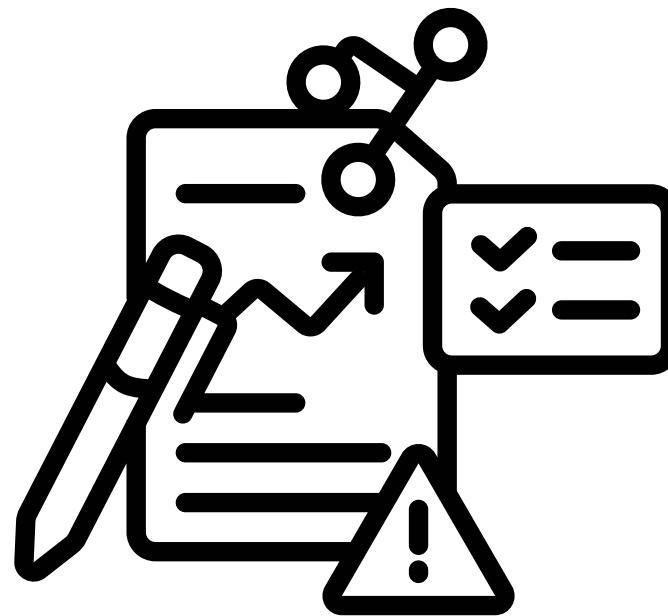
Performance Vector

Performance

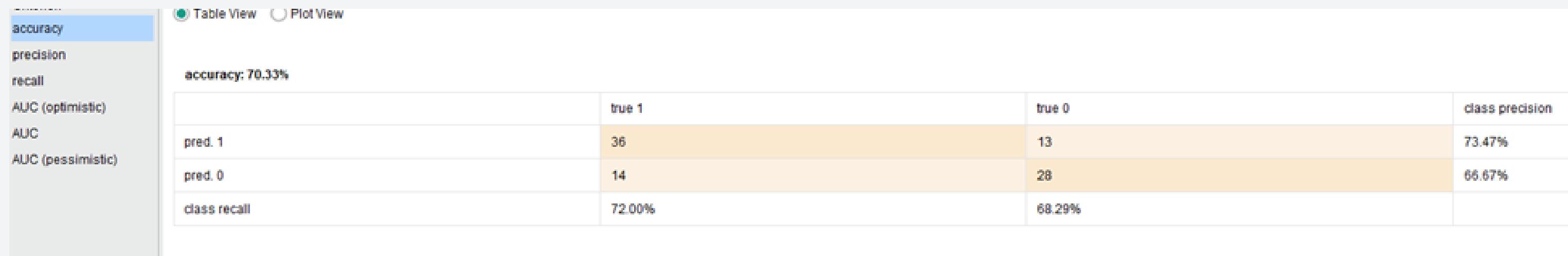
PerformanceVector:
accuracy: 78.02%
ConfusionMatrix:
True: 1 0
1: 40 10
0: 10 31
precision: 75.61% (positive class: 0)
ConfusionMatrix:
True: 1 0
1: 40 10
0: 10 31
recall: 75.61% (positive class: 0)
ConfusionMatrix:
True: 1 0
1: 40 10
0: 10 31
AUC (optimistic): 0.880 (positive class: 0)
AUC: 0.880 (positive class: 0)
AUC (pessimistic): 0.880 (positive class: 0)

Description

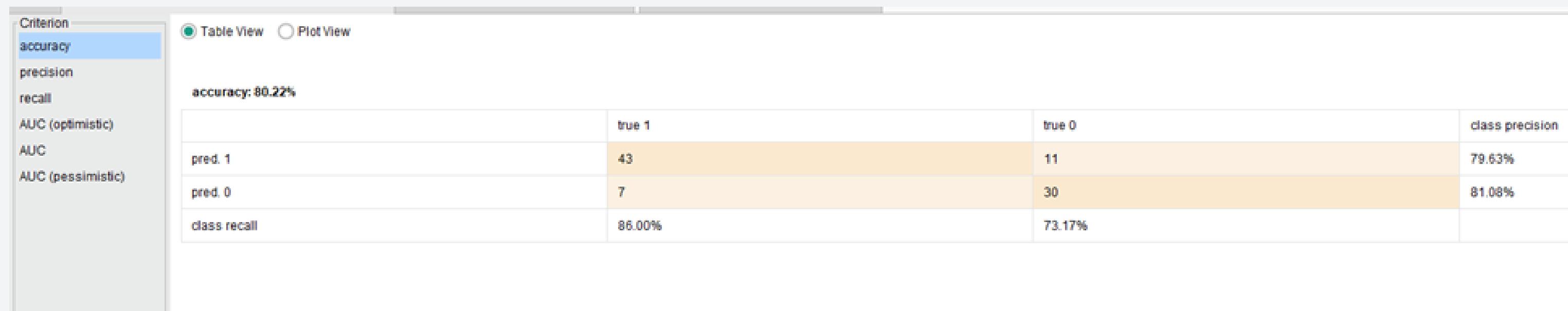
Annotations



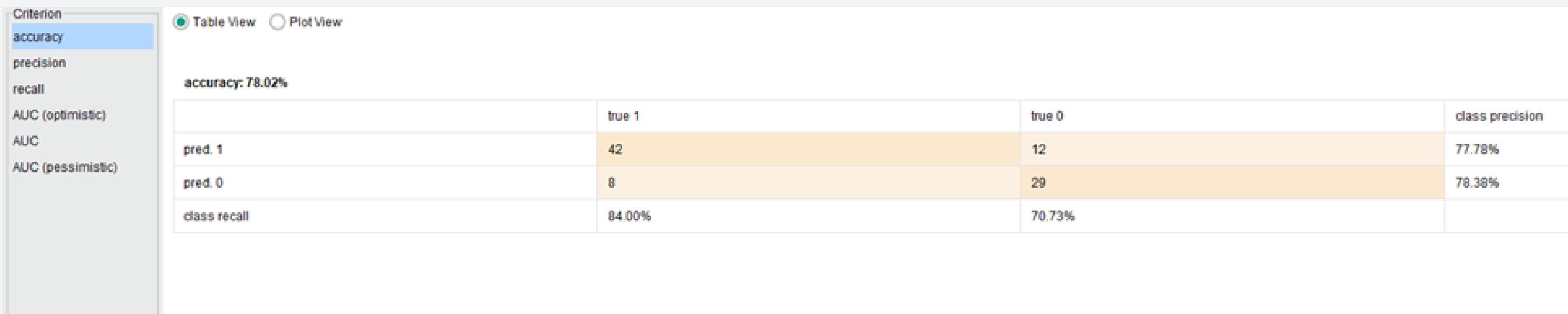
ONER



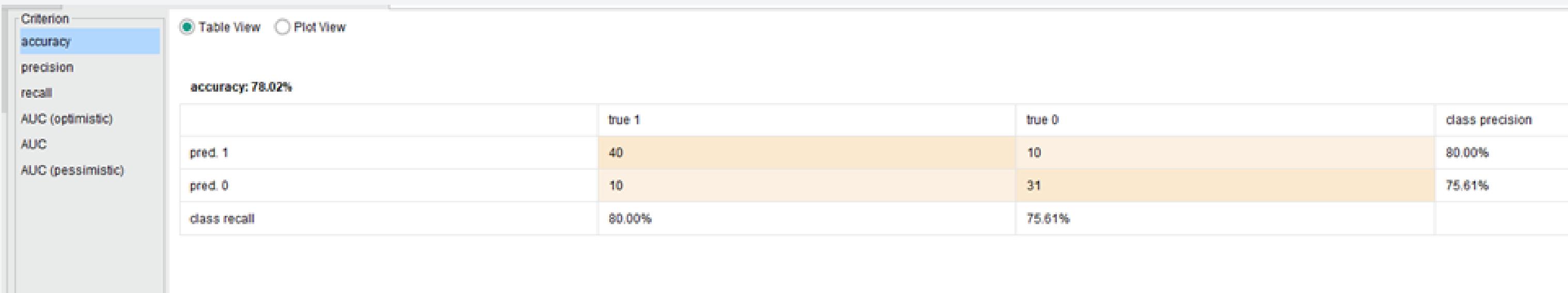
LOGISTIC



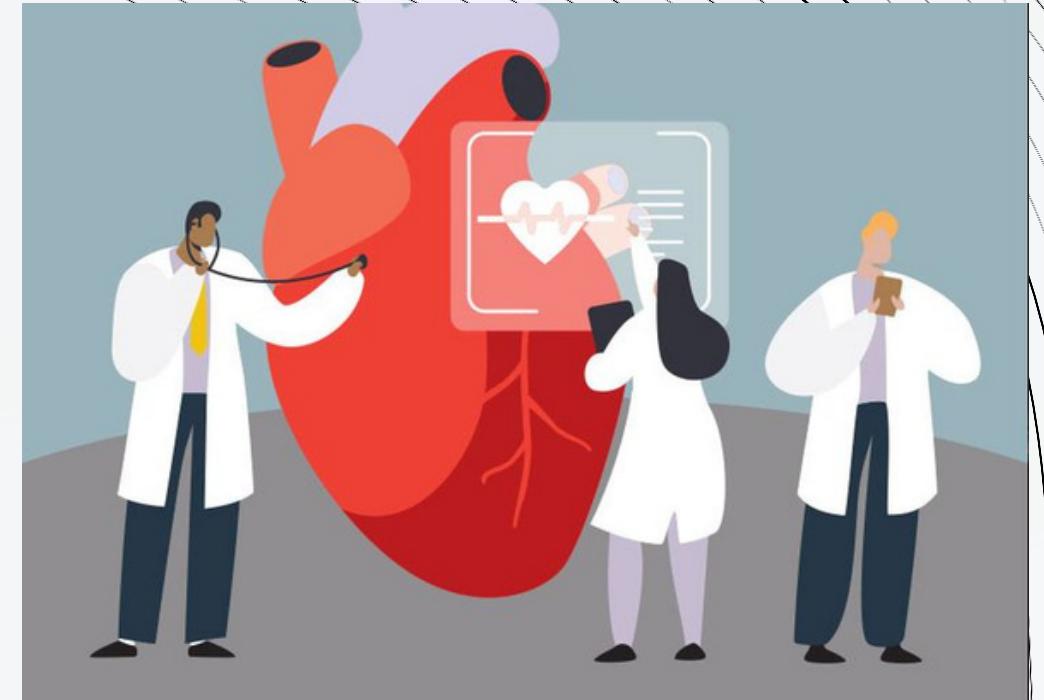
DECISION TREE



BAYES

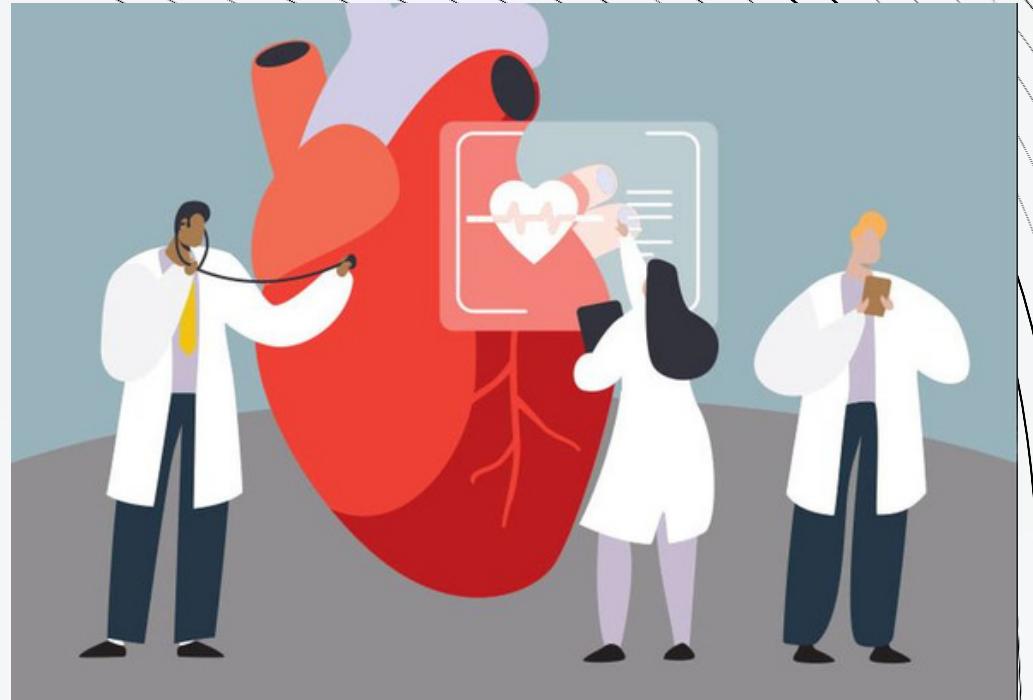


ผลการดำเนินงาน



จากผลการทดลอง การนำเอา Algorithm Logistic Regression มาใช้กับข้อมูลทำให้ได้ค่า Accuracy ที่สูงที่สุดเนื่องจากตัวของ Algorithm Logistic Regression ไม่ใช่การพยายามคำนวณค่าต่อเนื่อง แต่จะคำนวณความน่าจะเป็นที่ตัวแปรตามจะอยู่ในหนึ่งในสองกลุ่มที่กำหนด ในที่นี้คือ 0 – 1 การเป็นโรค หรือไม่เป็นโรคโดยใช้ฟังก์ชันโลจิสติก (logistic function) เพื่อแปลงผลลัพธ์เป็นความน่าจะเป็นที่อยู่ในช่วง [0, 1]. การใช้ฟังก์ชันนี้ช่วยให้ Logistic Regression สามารถจัดการกับปัญหาการจัดหมวดหมู่ได้ และทำให้เราได้ผลลัพธ์ที่มีค่า Accuracy สูงที่สุดคือ 80.22% เนื่องด้วย ข้อมูลมีการผสมของข้อมูลแบบ Continuous Value และ ข้อมูล แบบ Independent Value หรือก็คือทั้ง Numerical และ polynominal

ผลการดำเนินงาน



ในส่วนของ Algorithm Decision Tree และ Algorithm Navie Bayes เราจะเห็นได้ว่าค่าของ Accuracy มีค่าเท่ากับคือ 78.02 % แต่ความต่างของ recall นั้นต่างกันตรงที่ Algorithm Decision Tree จะมีค่า recall ของการทำนายว่าใครเป็นโรคได้แม่นยำกว่าผู้ของ Algorithm Navie Bayes ทางผู้กดลงมองเห็น หากต้องเลือกความพิดพลาดในการทำนาย ควรจะเลือกว่าคนที่ไม่เป็นโรคบัน เป็นจะมีค่าความเสียหายที่มากกว่า การทำนายว่าคนนี้ไม่ได้เป็นโรคก็ที่เป็นแล้วไม่ได้รับการรักษา

ก้ายกีสุดแล้วหากต้องเลือกเพียงอัลกอริทึมเดียวมาใช้ในการสร้างโมเดล ระหว่าง Algorithm Logistic Regression กับ Algorithm Decision Tree ผู้กดลงจะเลือก Logistic เหตุผล เพราะให้ค่า Accuracy สูงที่สุด

สรุปผลการดำเนินงานและข้อเสนอแนะ

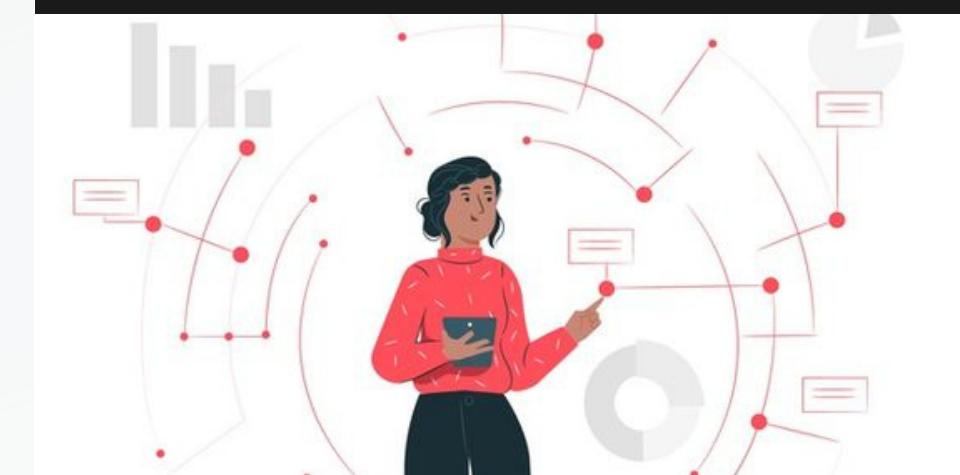
สรุปผลการดำเนินงาน



จากการทดลองเพื่อหารูปแบบความเกี่ยวข้องของการเกิดภาวะหัวใจล้มเหลว หรือภาวะหัวใจวาย
พบว่าแนวโน้มในการเกิดโรคจะมีความเกี่ยวข้องกับ ค่า ST depression ทำให้เกิดการตีบตันของหัวใจเมื่อมีการออกกำลังกาย , อัตราการเต้นของหัวใจสูงสุดที่ได้รับ , จำนวนหลอดเลือดหลักที่ผ่านการตรวจด้วยเครื่องเอ็กซเรย์แบบที่ใช้ร่วมกับสารกีบแสง ข้อมูลข้างต้นส่งผลถึงแนวโน้มในการเกิดภาวะหัวใจล้มเหลว โดยเลือกการวิเคราะห์ข้อมูลจากผลการดำเนินการของ Algorithm ที่ได้ทำการทดลอง

ปัญหาและข้อเสนอแนะที่พบคือ ในการประยุกต์ใช้งานควรเพิ่มจำนวนของข้อมูลที่นำมาใช้ Training & Testing เพราะในการทดลองนี้ได้มีการทำข้อมูลมาใช้เพียง 303 Instance ทำให้ยังมีช่วงของข้อผิดพลาด เช่น จำนวนข้อมูลที่บออย และ ขาดข้อมูลที่นำมาทำการทดลองเป็นข้อมูลที่ไม่เป็นปัจจุบัน ทำให้ปัจจัยในการเกิดอาจมีแนวโน้มที่ต่างออกไปจากรูปแบบการเกิดที่ได้จากโมเดลของชุดข้อมูลนี้

ปัญหาและข้อเสนอแนะ





คำ ค่า น