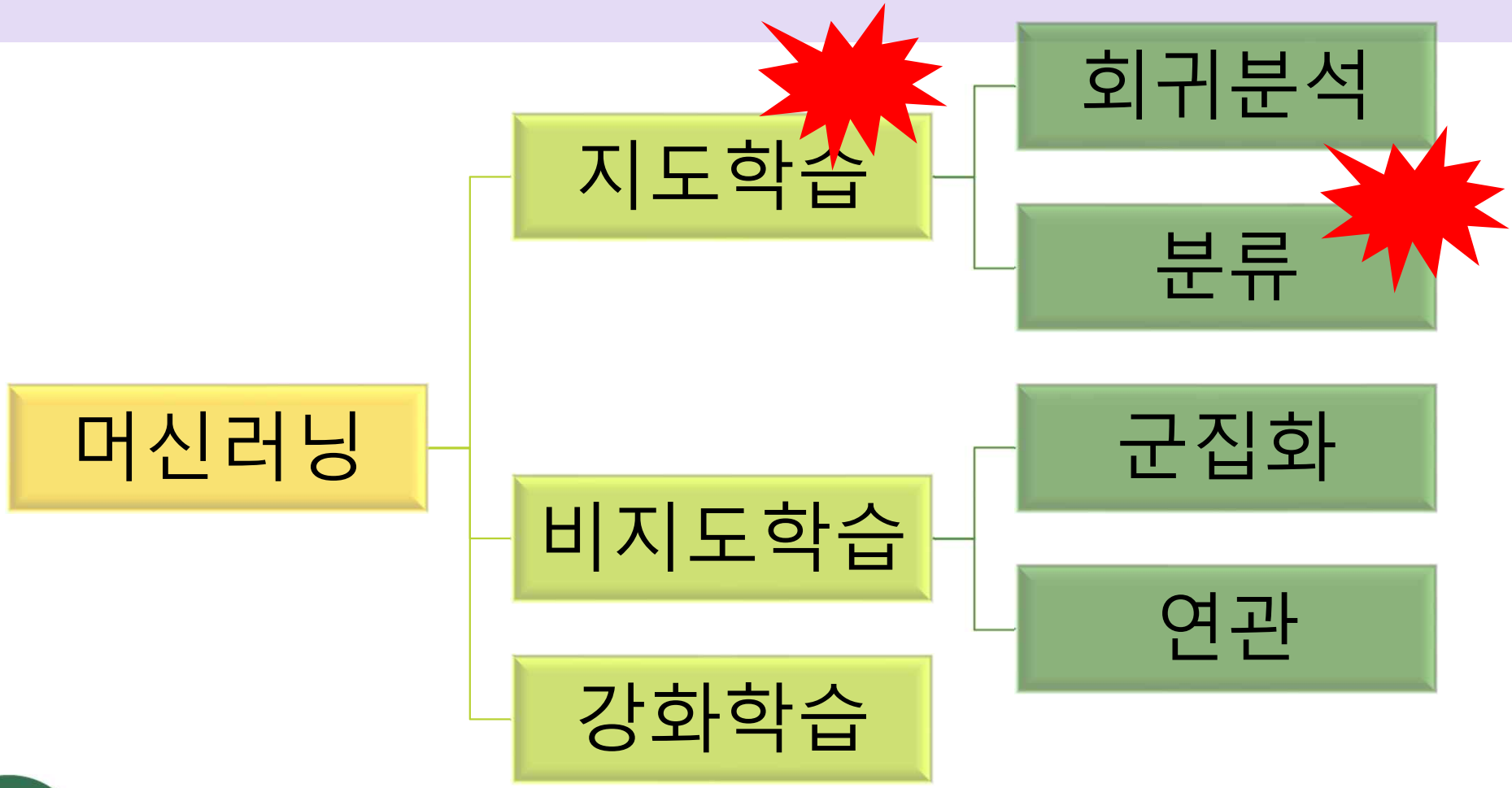
orange 활용 데이터 분석 및 머신 러닝

4차시

지도학습 - 분류 1

- 정형데이터를 활용한 분류 1
  1) 분류모델의 개념 이해
  2) 분류모델의 성능평가지표


머신러닝
지도학습
회귀분석
분류
비지도학습
군집화
연관
강화학습

분류에 활용하는 데이터의 성질

Feature
Attributes
독립변수
문제
: 정형/비정형데이터

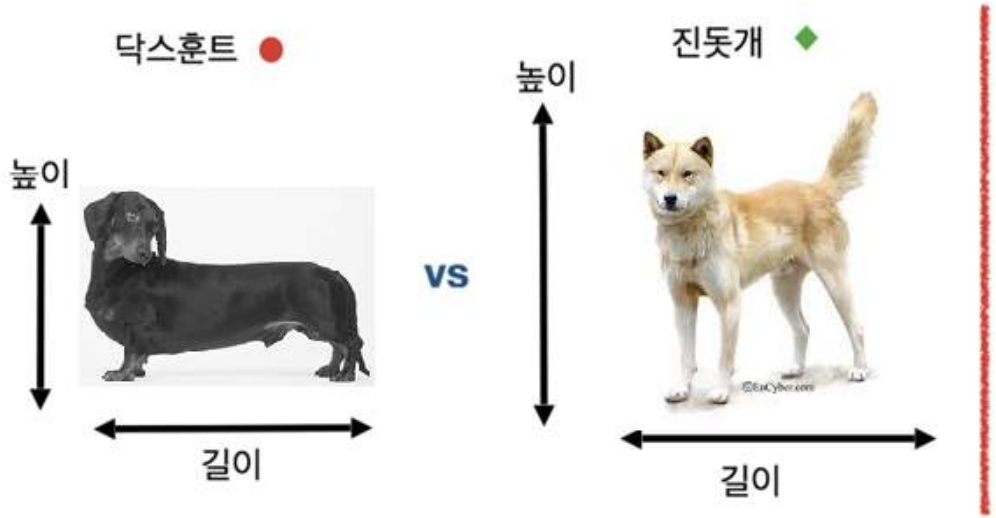
Target : 연속된 숫자가 아닌

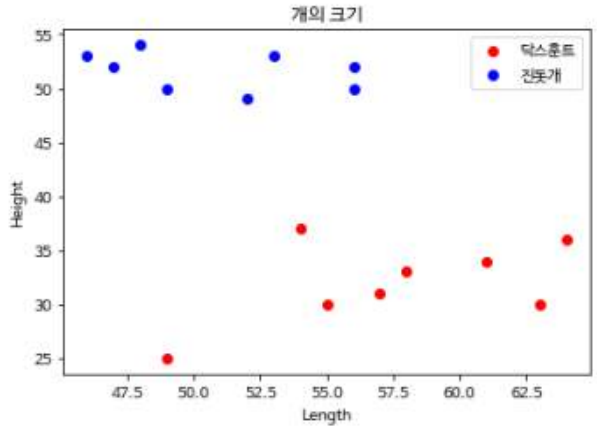
Category 데이터로 이루어짐 :
Class Label

그림 http://ai4school.org

분류

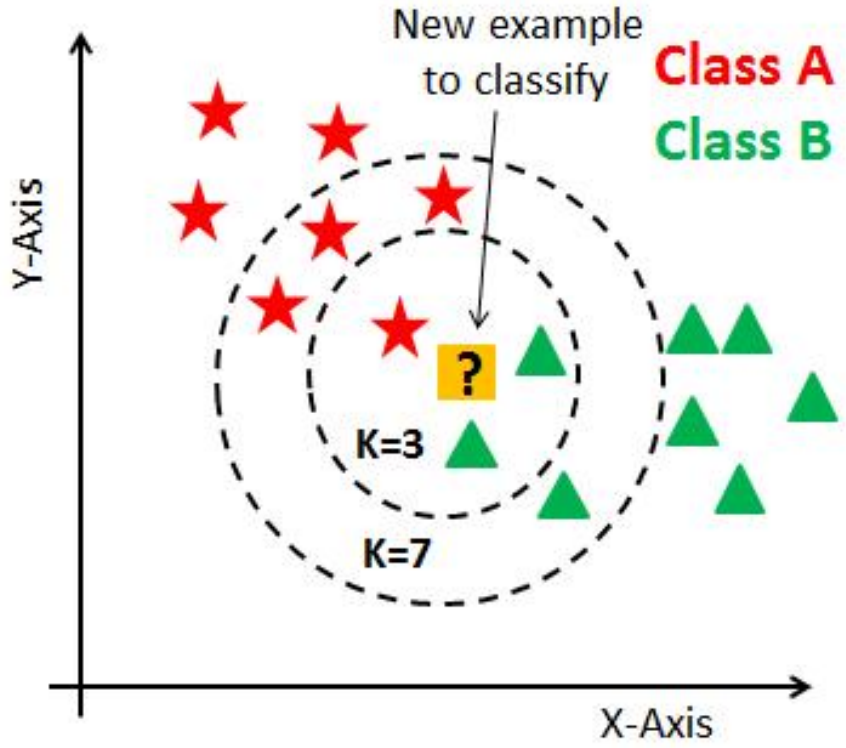
- 산포도 그래프의 위쪽에 분포한 녹색 점은 진돗개이고, 아래쪽의 붉은 점들은 닥스훈트 종이다. 이러한 데이터를 이용하여 효과적으로 **새로운 종에 대해 분류**를 하려면 어떤 기법을 적용하는 것이 바람직할까?



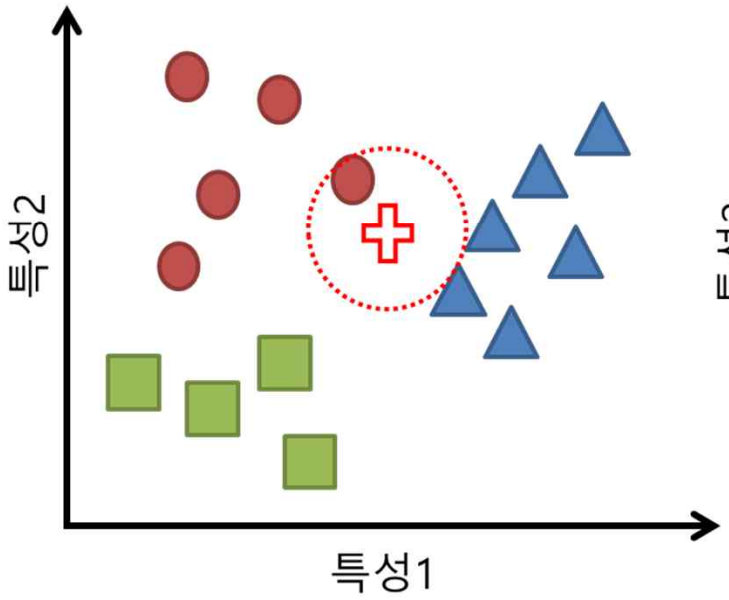


k-NN 알고리즘

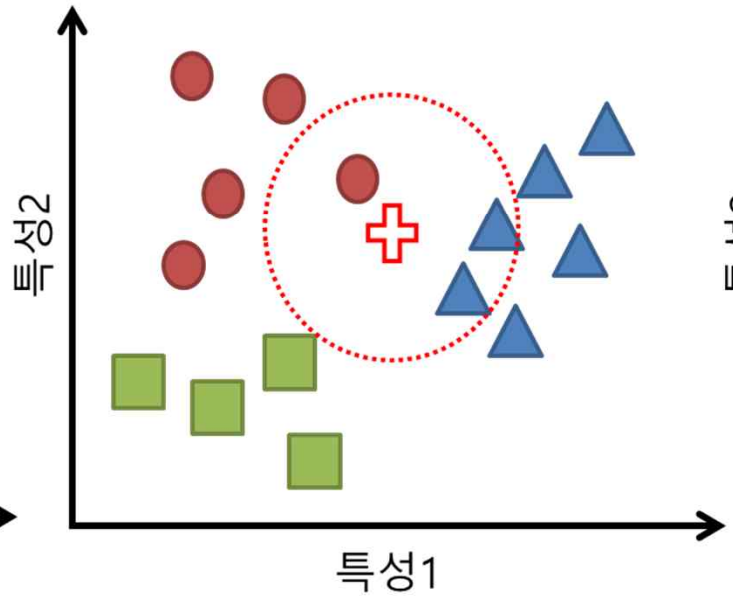
- **k-NN 알고리즘 : k-최근접 이웃**k-Nearest Neighbor의 약자로 특징 공간에 분포하는 데이터에 대하여 k개의 가장 가까운 이웃을 살펴보고 **다수결 방식으로 데이터의 레이블을 할당하는 분류 방식**이다.



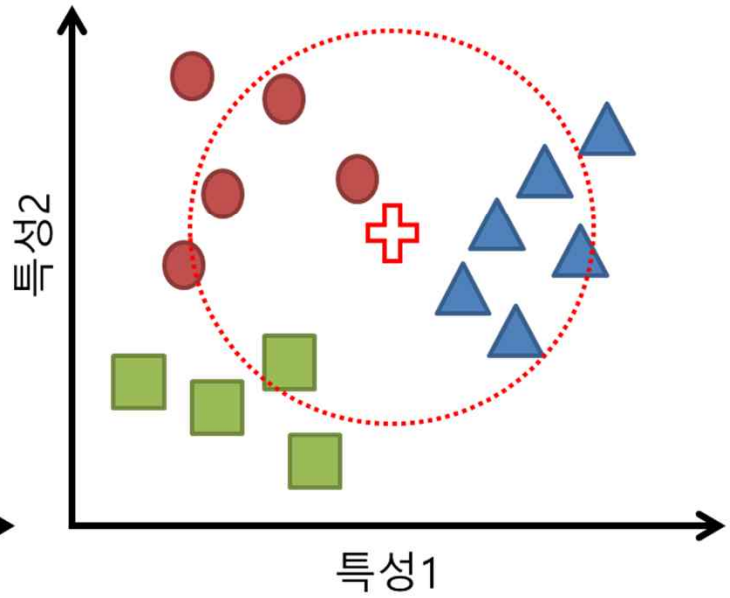
k-NN 분류



k=1일 때



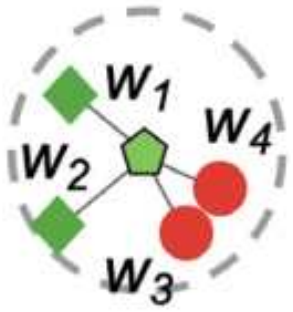
k=3일 때



k=9일 때

https://bskyvision.com/entry/직관적이고도-매우-간단한-분류-모델-kNNk-Nearest-Neighbor

k-NN 알고리즘에 사용할 견종의 표본 집합 시각화



k=4일때

클래스 A :

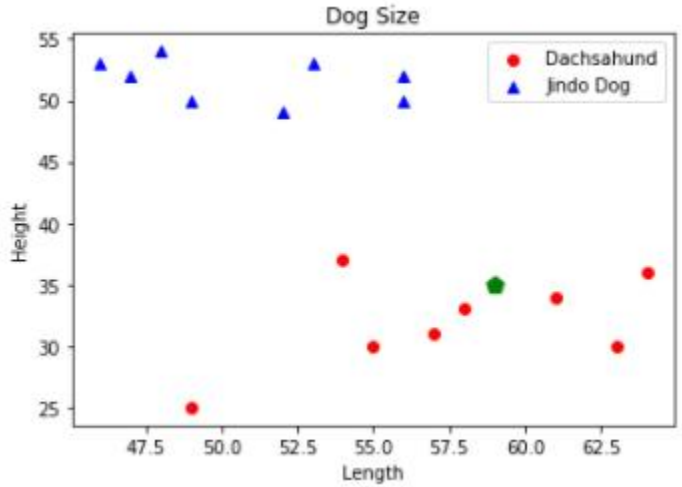
클래스 B :

- k가 짝수이고 가장 가까운 클래스 A와 클래스 B의 개수가 같은 경우 오각형 새 데이터의 클래스를 판정하는 방법 → 이웃과의 거리에 가중치(w_i)를 부여하는 모델

k-NN 알고리즘에 사용할 견종의 표본 집합 시각화

닥스훈트 8마리의 길이와 높이								
길이	55	57	64	63	58	49	54	61
높이	30	31	36	30	33	25	37	34
진돗개 8마리의 길이와 높이								
길이	56	47	46	46	49	53	52	48
높이	52	52	50	53	50	53	49	54

분류하고자 하는 데이터 : [59,35]



orange 를 이용해 견종을 분류해보자.

File

File - Orange

Source

File: 닥스훈트와 진돗개_train.csv ... Reload

URL: https://www.1ka.si/podatki/141025/72F5B3CC/

File Type

Automatically detect type

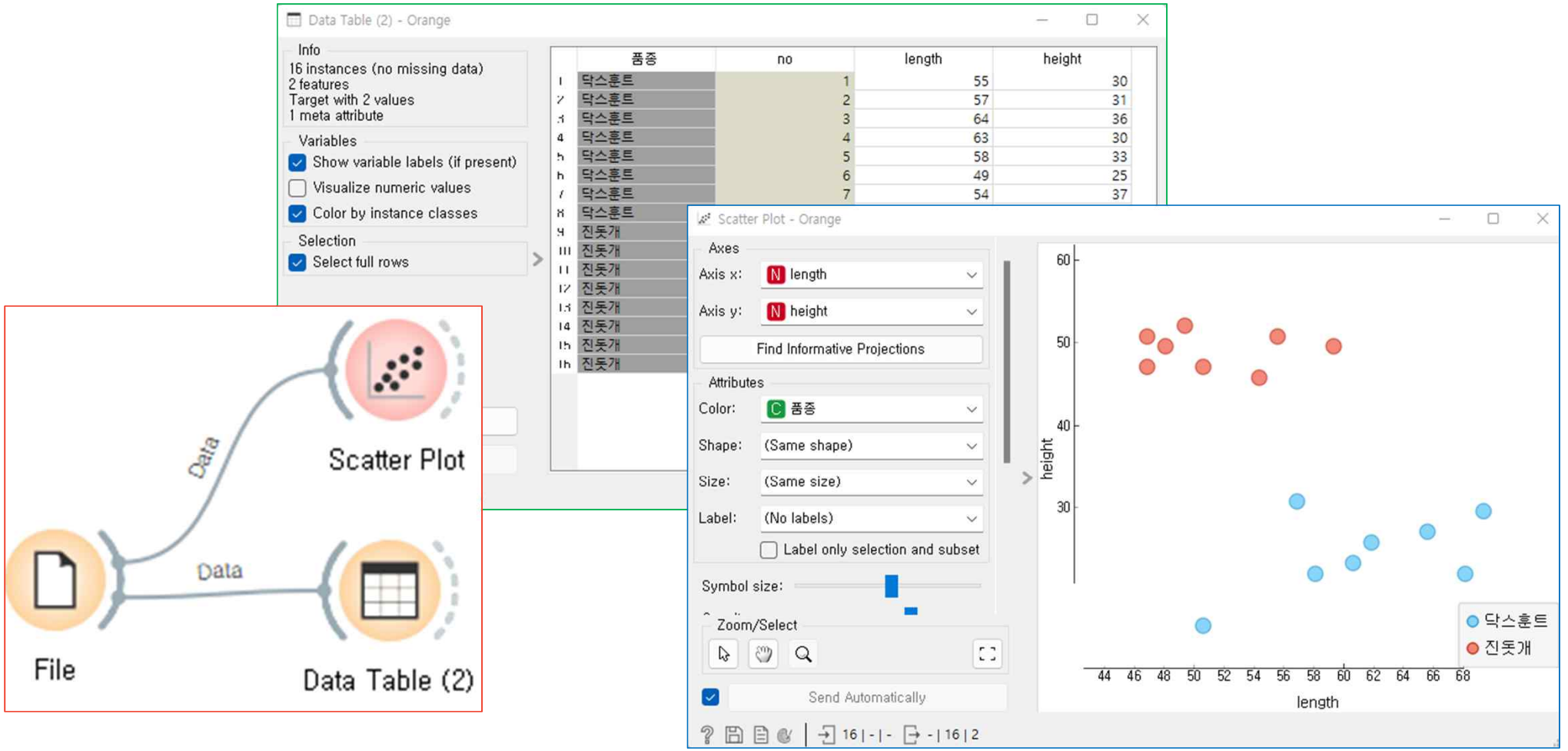
Info

16 instance(s)
4 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

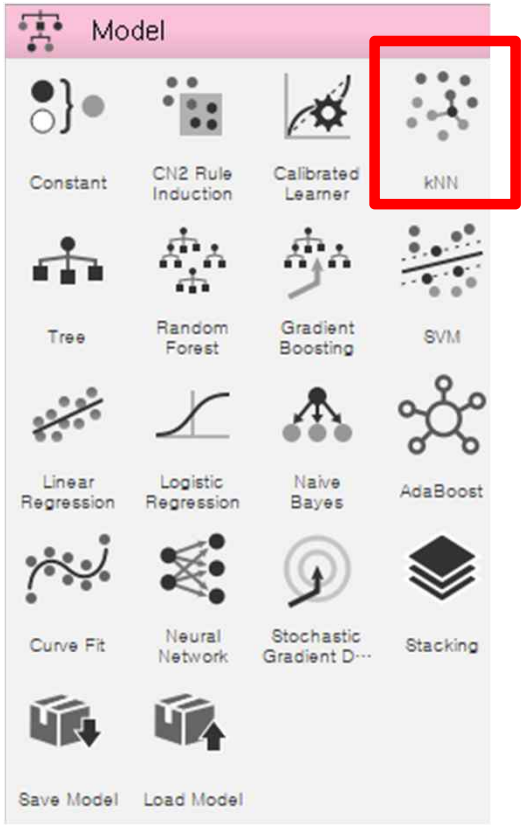
Columns (Double click to edit)

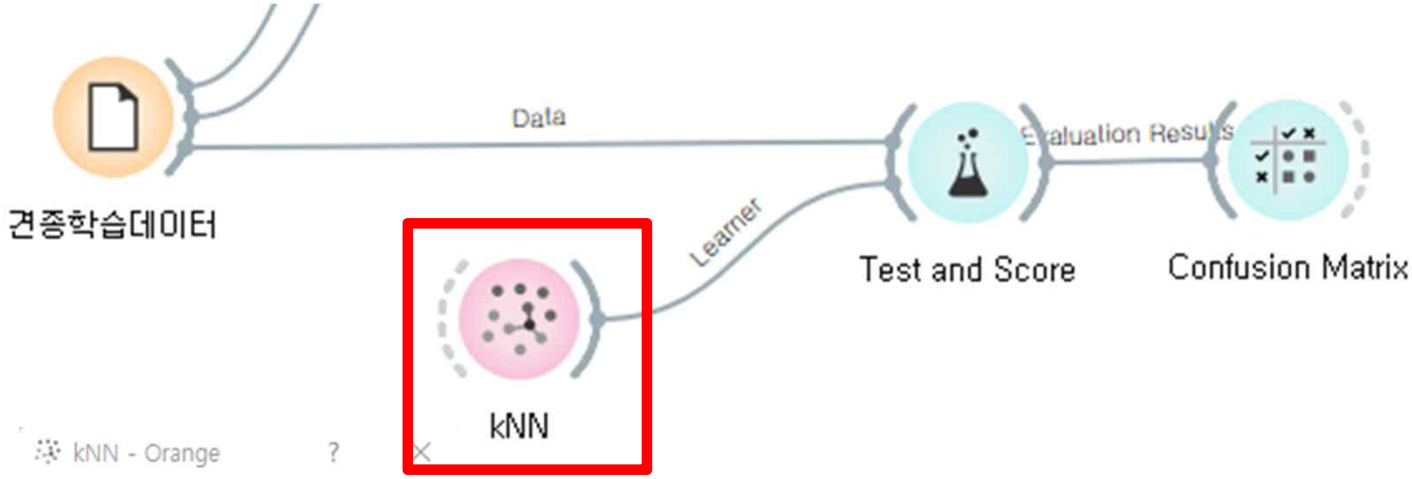
	Name	Type	Role	Values
1	no	numeric	meta	
2	length	numeric	feature	
3	height	numeric	feature	
4	품종	categorical	target	닥스훈트, 진돗개

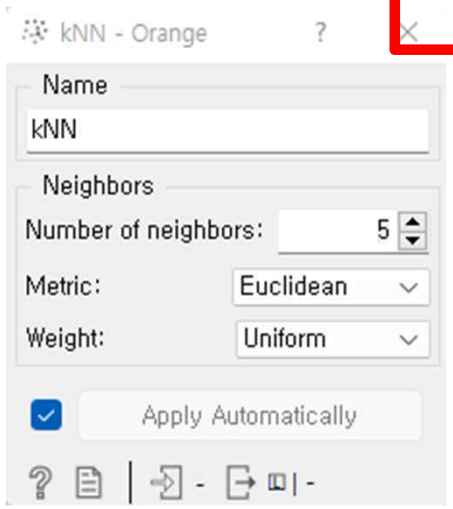
데이터 살펴보기

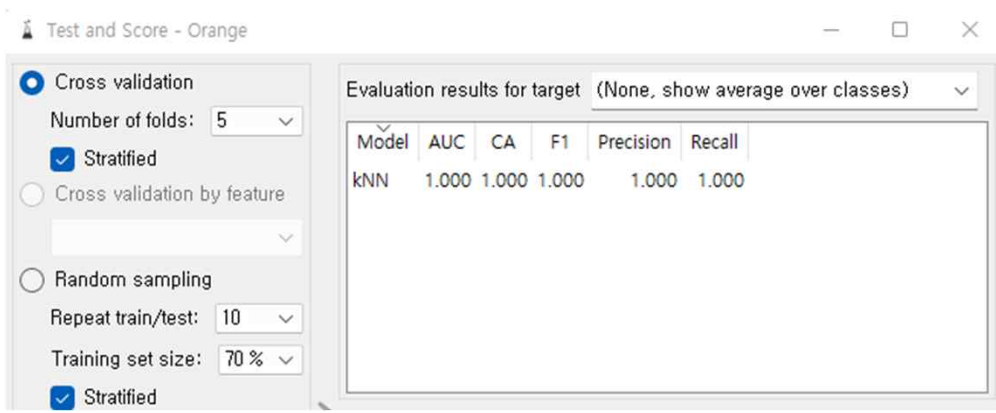

Data Table (2) - Orange
Info
16 instances (no missing data)
2 features
Target with 2 values
1 meta attribute
Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes
Selection
Select full rows
품종 no length height
1 닥스훈트 1 55 30
2 닥스훈트 2 57 31
3 닥스훈트 3 64 36
4 닥스훈트 4 63 30
5 닥스훈트 5 58 33
6 닥스훈트 6 49 25
7 닥스훈트 7 54 37
8 닥스훈트
9 진돗개
10 진돗개
11 진돗개
12 진돗개
13 진돗개
14 진돗개
15 진돗개
16 진돗개
Scatter Plot - Orange
Axes
Axis x: length
Axis y: height
Find Informative Projections
Attributes
Color: 품종
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Label only selection and subset
Symbol size:
Zoom/Select
Send Automatically
16 | - | - - | 16 | 2
60
50
40
30
height
44 46 48 50 52 54 56 58 60 62 64 66 68
length
닥스훈트
진돗개
Data
Data
Scatter Plot
File
Data Table (2)

KNN 알고리즘 적용하여 검증평가하기

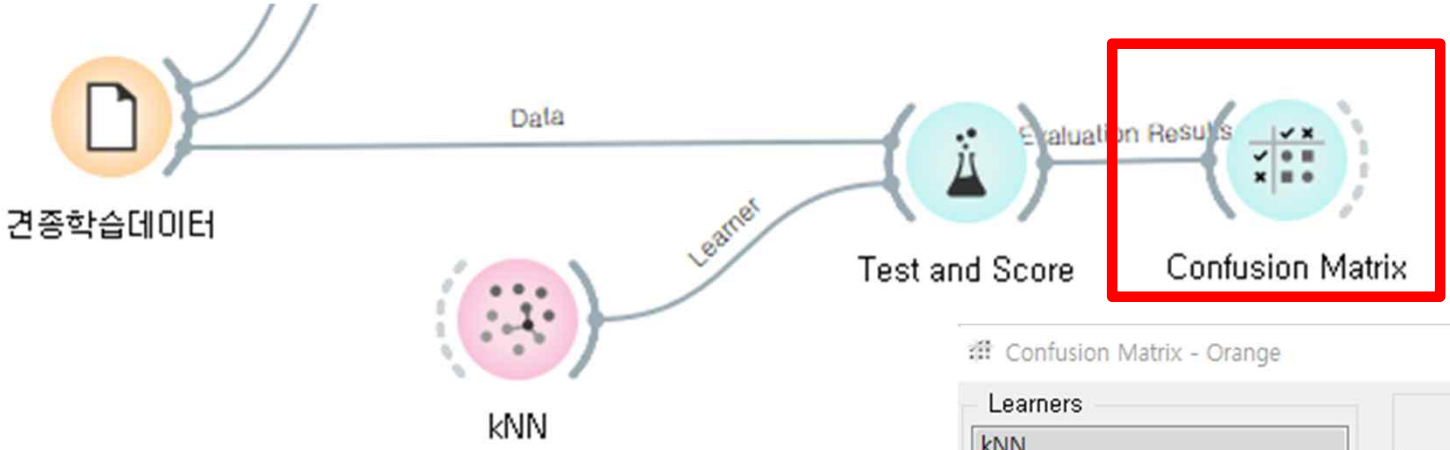


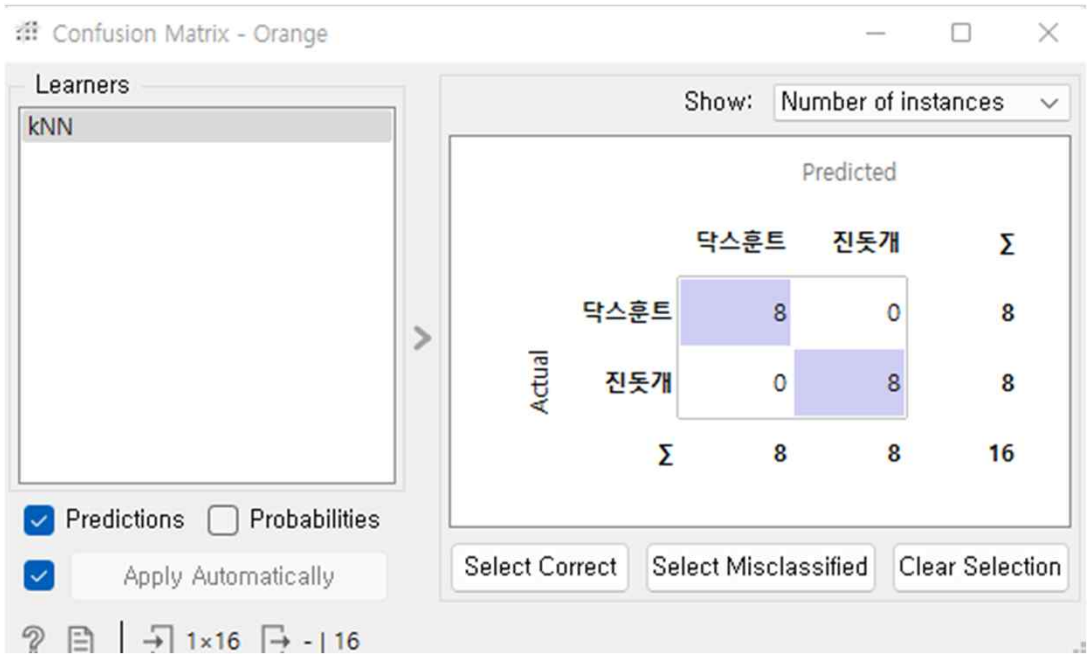




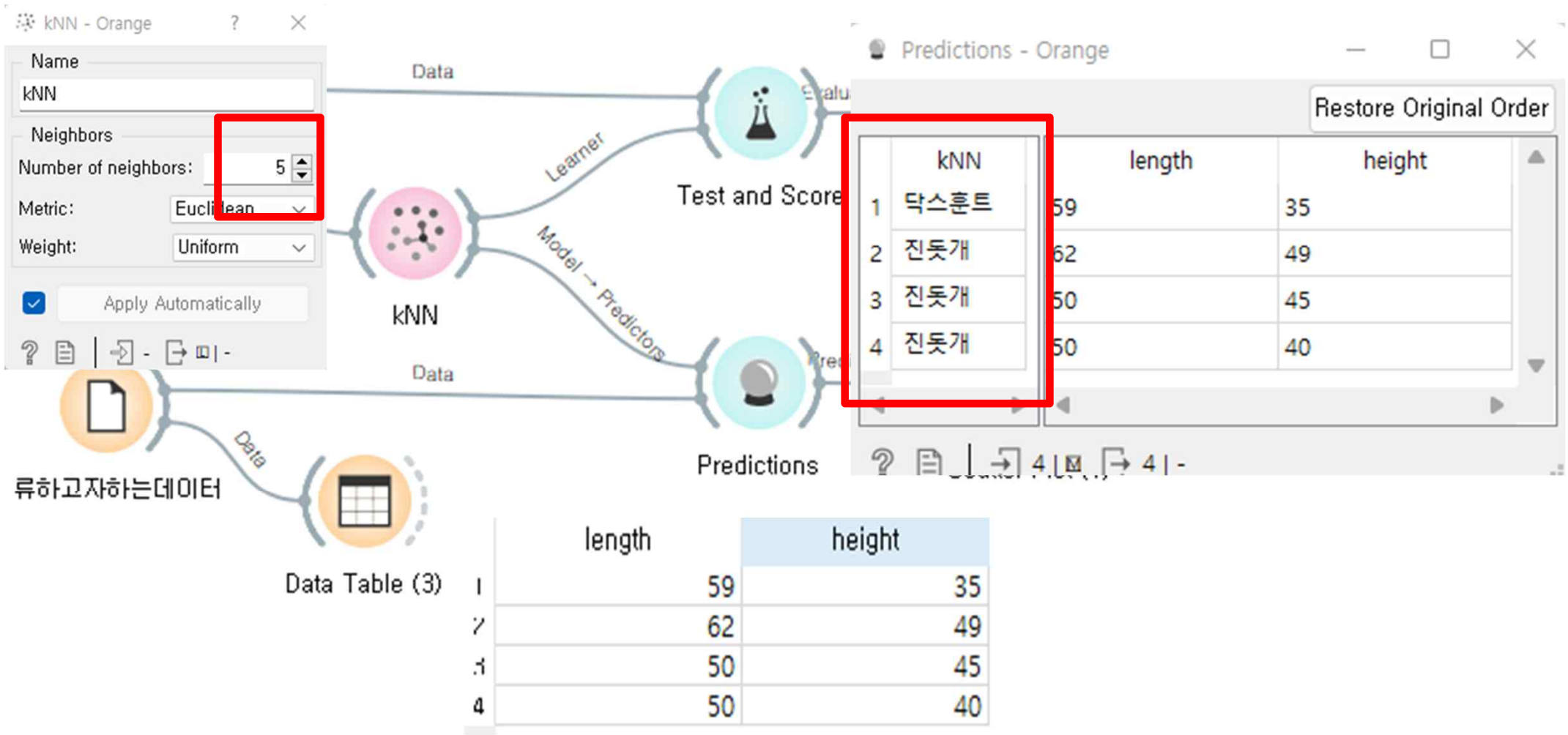


혼동행렬 (Confusion Matrix)



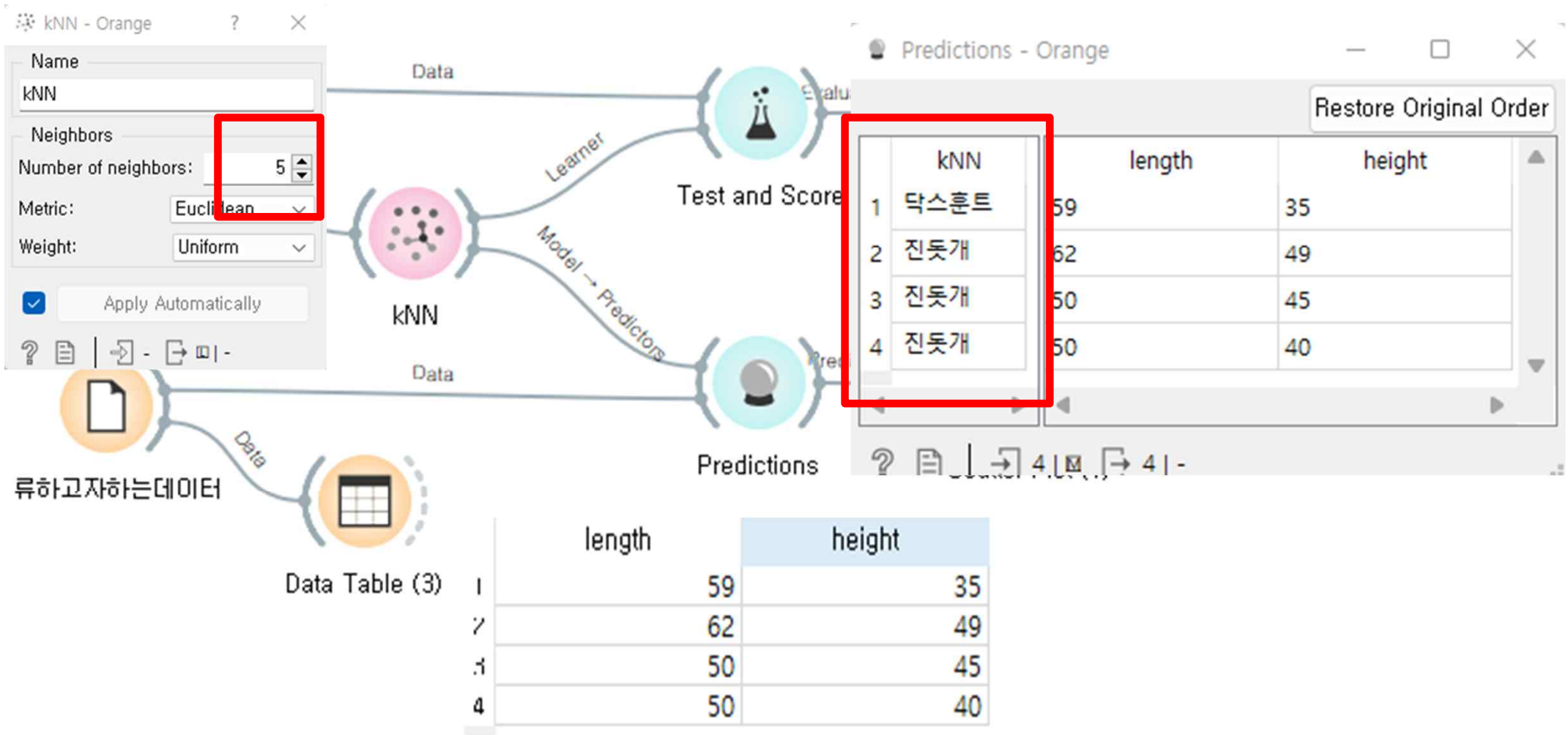


KNN을 활용하여 분류 예측하기



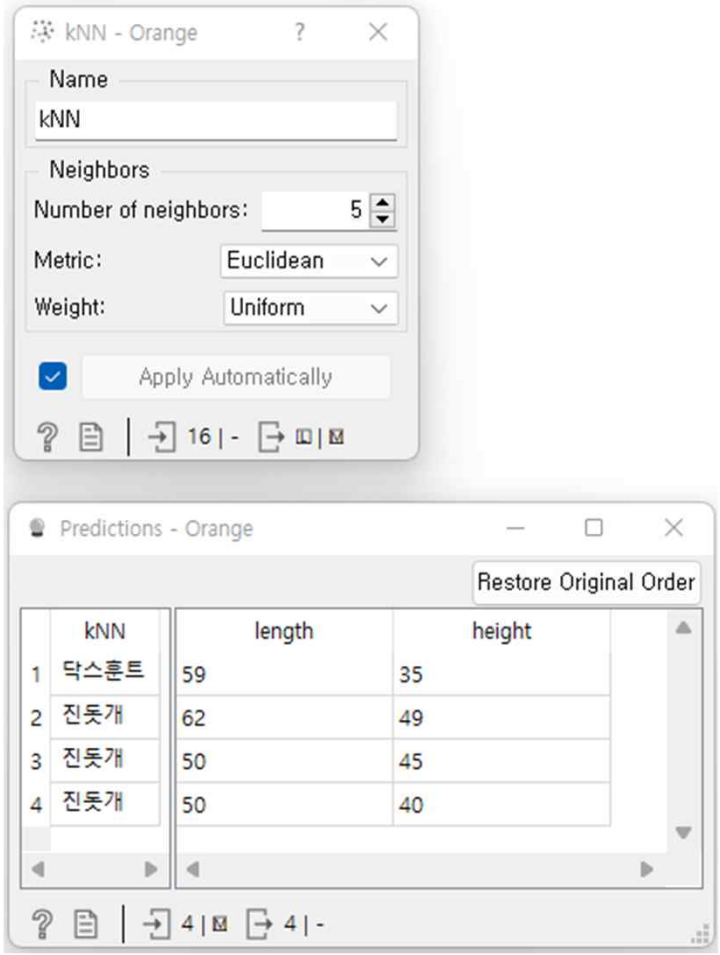
	length	height
1	59	35
2	62	49
3	50	45
4	50	40

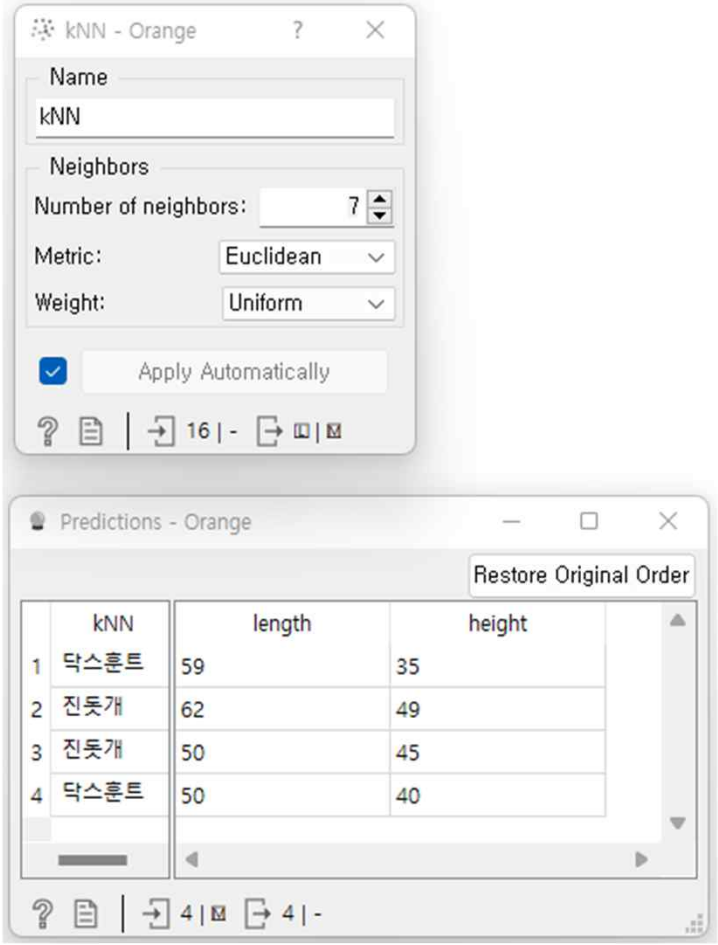
KNN을 활용하여 분류 예측하기



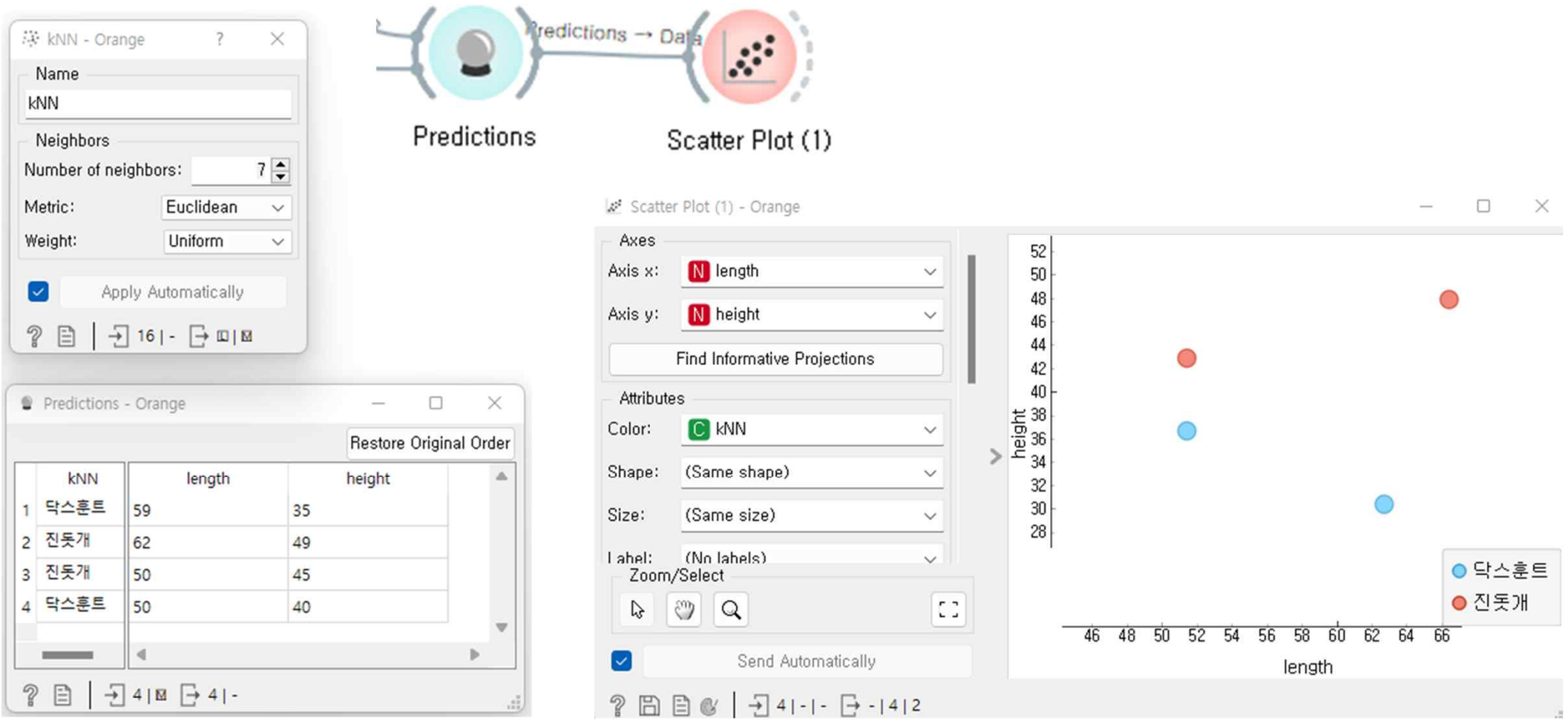
	length	height
1	59	35
2	62	49
3	50	45
4	50	40

K 값의 변화에 따른 분류 결과의 변동 확인하기

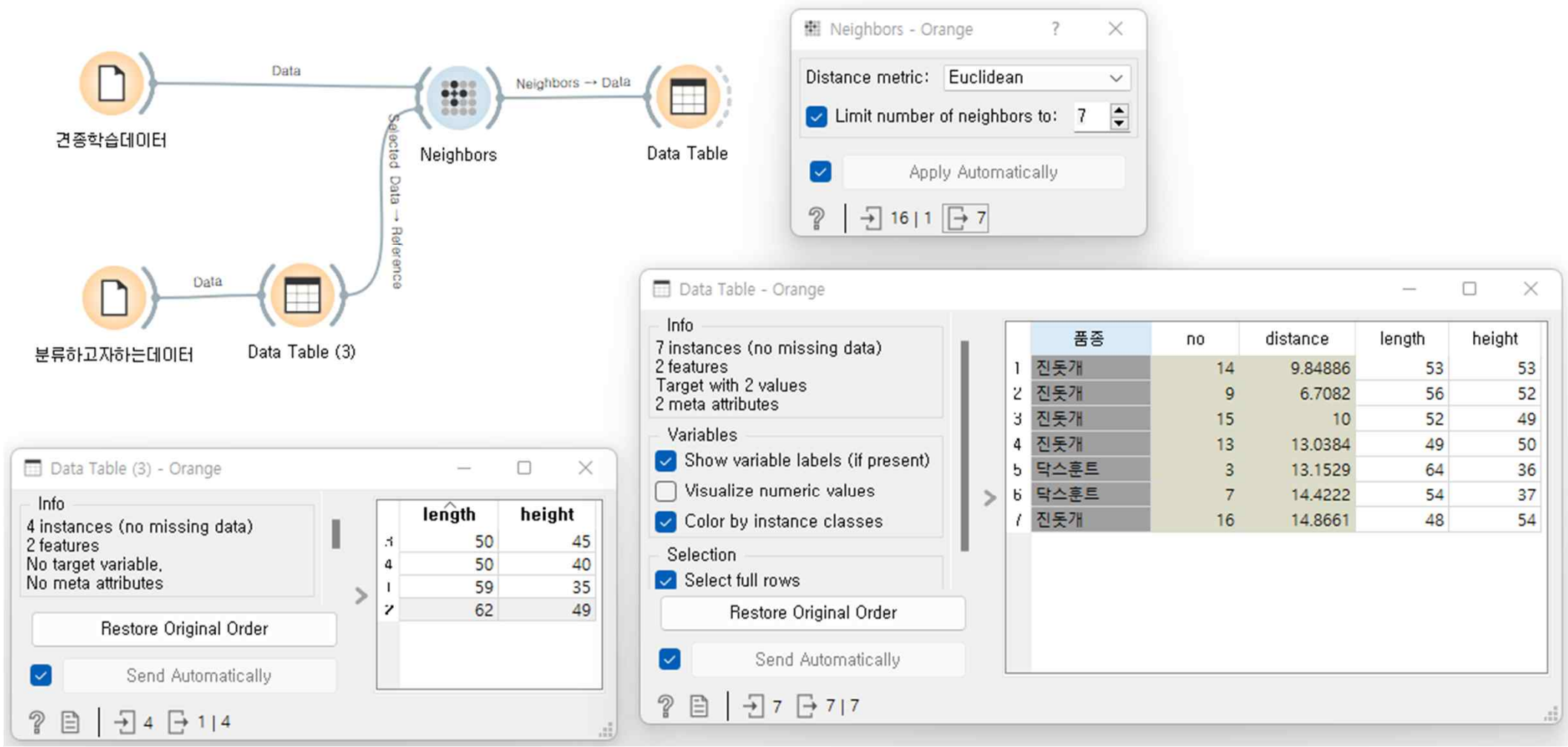




분류결과 그래프로 확인하기



K 개의 최 근접 이웃은 무엇이었을까요?



건강지표 데이터를 활용
심장병 유무를 예측해봅시다.

데이터 읽어오기 - 오렌지에서 제공하는데이터

Datasets

Datasets - Orange

Search for data set ...

	Title	Size	Instances	Variables	Target	Tags
●	Heart Disease	23.5 KB	303	14	C categorical	biology, medicine
●	Housing	33.9 KB	506	14	N numeric	economy
	Illegal waste dumpsites in Slovenia	2.8 MB	13165	25		geo, timeseries, e...
	Imports 1985	25.7 KB	205	25	N numeric	insurance, econo...
	Ionosphere	74.9 KB	351	33	C categorical	physics
●	Iris	4.5 KB	150	5	C categorical	biology
	Kickstarter projects	214.1 KB	1163	15	C categorical	economy
	Lenses	968 bytes	24	5	C categorical	medical
	Liver Disorders	7.2 KB	345	11	C categorical	biology
	Liver cirrhosis - spectral image	3.4 MB	1078	546		spectral, hyperspe...
	Liver spectroscopy (Collagen)	994.8 KB	731	234	C categorical	spectral
	Lymphography	14.6 KB	148	19	C categorical	medical
	METABRIC: Molecular Taxonomy of Breast Ca...	136.3 MB	1904	24403		gene expression, ...
	MONK's 1	12.4 KB	556	7	C categorical	synthetic

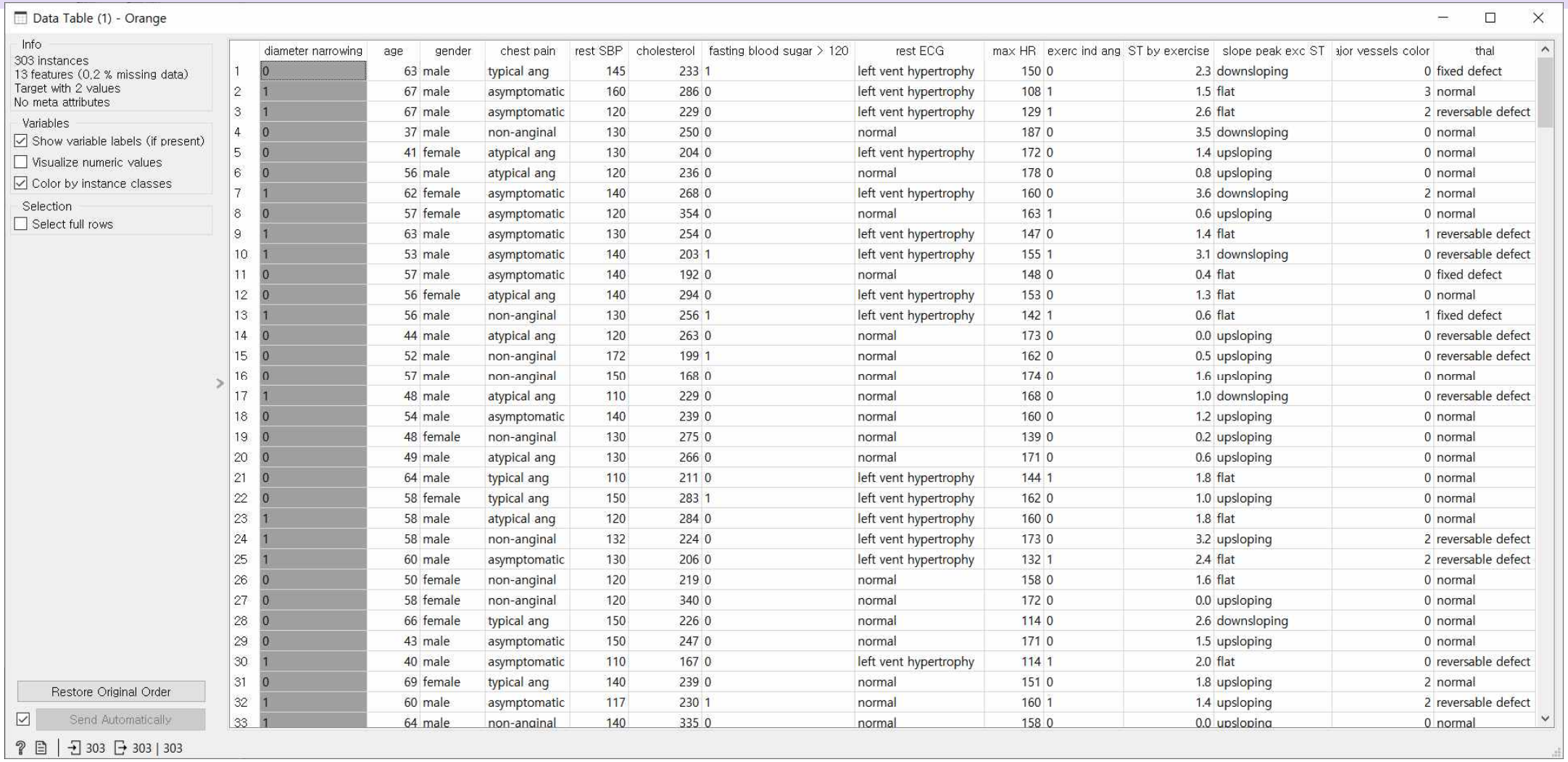
Description

Heart Disease (1988), from UCI ML Repository

This data uses a subset of 14 attributes from the Cleveland database. The 'goal' field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

303

심장병 데이터


Data Table (1) - Orange

Info
303 instances
13 features (0,2 % missing data)
Target with 2 values
No meta attributes

Variables
- [x] Show variable labels (if present)
- [ ] Visualize numeric values
- [x] Color by instance classes

Selection
- [ ] Select full rows

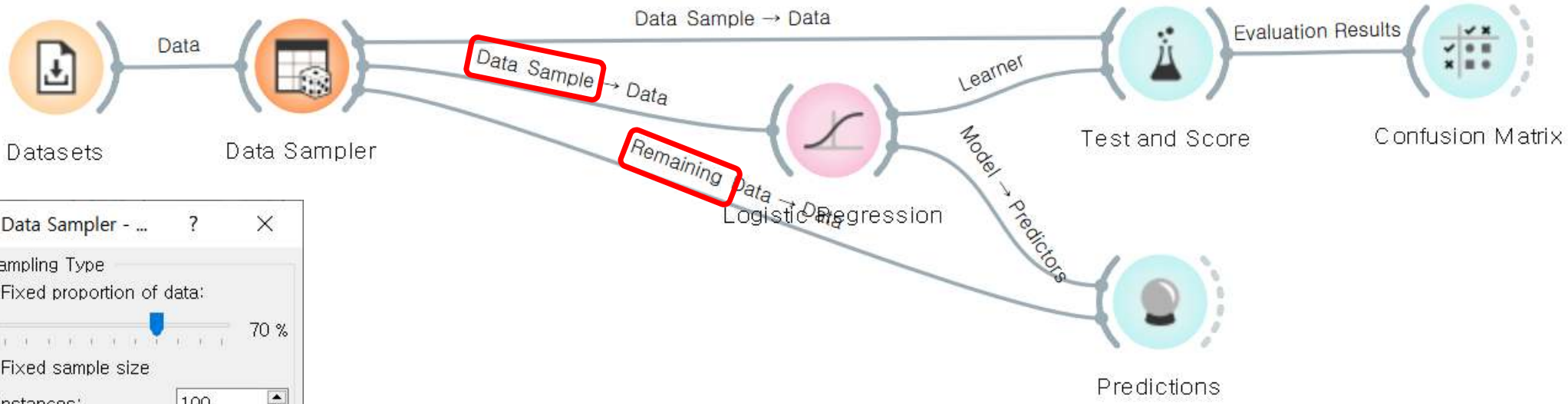
Restore Original Order
Send Automatically
303 | 303 | 303

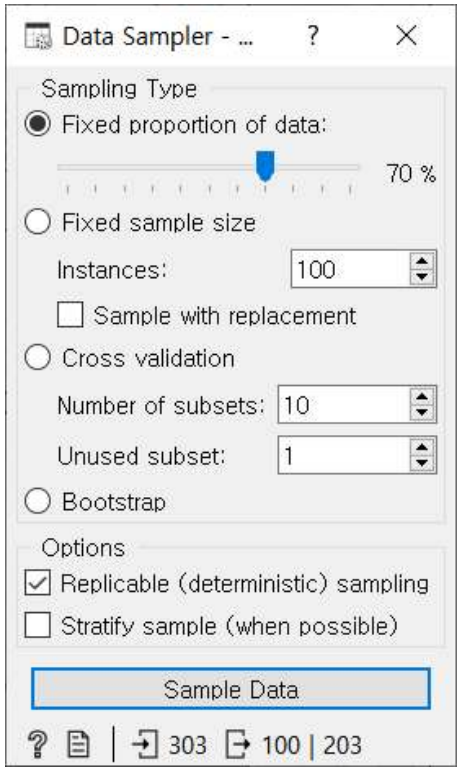
	diameter narrowing	age	gender	chest pain	rest SBP	cholesterol	fasting blood sugar > 120	rest ECG	max HR	exerc ind ang	ST by exercise	slope peak exc ST	ajor vessels color	thal
1	0	63	male	typical ang	145	233	1	left vent hypertrophy	150	0	2.3	downsloping	0	fixed defect
2	1	67	male	asymptomatic	160	286	0	left vent hypertrophy	108	1	1.5	flat	3	normal
3	1	67	male	asymptomatic	120	229	0	left vent hypertrophy	129	1	2.6	flat	2	reversable defect
4	0	37	male	non-anginal	130	250	0	normal	187	0	3.5	downsloping	0	normal
5	0	41	female	atypical ang	130	204	0	left vent hypertrophy	172	0	1.4	upsloping	0	normal
6	0	56	male	atypical ang	120	236	0	normal	178	0	0.8	upsloping	0	normal
7	1	62	female	asymptomatic	140	268	0	left vent hypertrophy	160	0	3.6	downsloping	2	normal
8	0	57	female	asymptomatic	120	354	0	normal	163	1	0.6	upsloping	0	normal
9	1	63	male	asymptomatic	130	254	0	left vent hypertrophy	147	0	1.4	flat	1	reversable defect
10	1	53	male	asymptomatic	140	203	1	left vent hypertrophy	155	1	3.1	downsloping	0	reversable defect
11	0	57	male	asymptomatic	140	192	0	normal	148	0	0.4	flat	0	fixed defect
12	0	56	female	atypical ang	140	294	0	left vent hypertrophy	153	0	1.3	flat	0	normal
13	1	56	male	non-anginal	130	256	1	left vent hypertrophy	142	1	0.6	flat	1	fixed defect
14	0	44	male	atypical ang	120	263	0	normal	173	0	0.0	upsloping	0	reversable defect
15	0	52	male	non-anginal	172	199	1	normal	162	0	0.5	upsloping	0	reversable defect
16	0	57	male	non-anginal	150	168	0	normal	174	0	1.6	upsloping	0	normal
17	1	48	male	atypical ang	110	229	0	normal	168	0	1.0	downsloping	0	reversable defect
18	0	54	male	asymptomatic	140	239	0	normal	160	0	1.2	upsloping	0	normal
19	0	48	female	non-anginal	130	275	0	normal	139	0	0.2	upsloping	0	normal
20	0	49	male	atypical ang	130	266	0	normal	171	0	0.6	upsloping	0	normal
21	0	64	male	typical ang	110	211	0	left vent hypertrophy	144	1	1.8	flat	0	normal
22	0	58	female	typical ang	150	283	1	left vent hypertrophy	162	0	1.0	upsloping	0	normal
23	1	58	male	atypical ang	120	284	0	left vent hypertrophy	160	0	1.8	flat	0	normal
24	1	58	male	non-anginal	132	224	0	left vent hypertrophy	173	0	3.2	upsloping	2	reversable defect
25	1	60	male	asymptomatic	130	206	0	left vent hypertrophy	132	1	2.4	flat	2	reversable defect
26	0	50	female	non-anginal	120	219	0	normal	158	0	1.6	flat	0	normal
27	0	58	female	non-anginal	120	340	0	normal	172	0	0.0	upsloping	0	normal
28	0	66	female	typical ang	150	226	0	normal	114	0	2.6	downsloping	0	normal
29	0	43	male	asymptomatic	150	247	0	normal	171	0	1.5	upsloping	0	normal
30	1	40	male	asymptomatic	110	167	0	left vent hypertrophy	114	1	2.0	flat	0	reversable defect
31	0	69	female	typical ang	140	239	0	normal	151	0	1.8	upsloping	2	normal
32	1	60	male	asymptomatic	117	230	1	normal	160	1	1.4	upsloping	2	reversable defect
33	1	64	male	non-anginal	140	335	0	normal	158	0	0.0	upsloping	0	normal

심장병 데이터

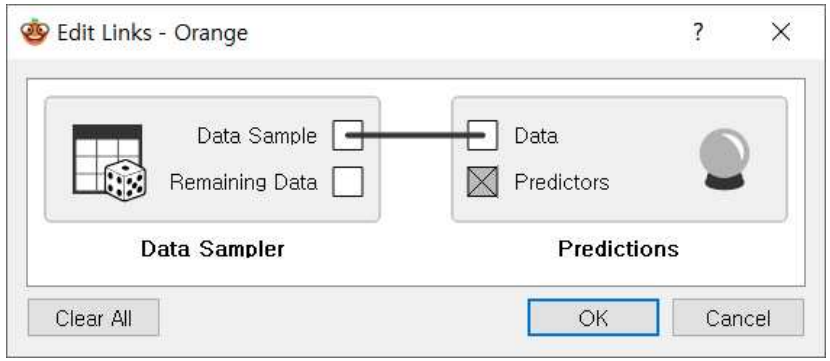
속성	의미	속성	의미
age	나이	**max HR**	최대 심장 박동수
gender	성별(male, female)	**exerc ind ang**	협심증 유발 운동(0, 1))
chest pain	가슴 통증 유형 (4가지)	**ST by exercise**	비교적 안정되기까지 운동에 의해 유발되는 ST 분절 하강 정도
rest SBP	입원 시 안정 혈압(mmHg)	**slope peak exc ST**	최대 운동 ST 분절 기울기(0~3)
cholesterol	혈청 콜레스테롤(mg/dl)	**major vessels colored**	형광 투시된 주요 혈관 수(0~3)
fast blood sugar > 120	공복 혈당(0, 1)	**thal**	탈륨 스트레스 검사 결과(3가지)
rest ECG	안정 심전도 결과(3가지)	**diameter narrowing**	심장병 진단(0, 1)

Data Sampler 위젯 사용하기

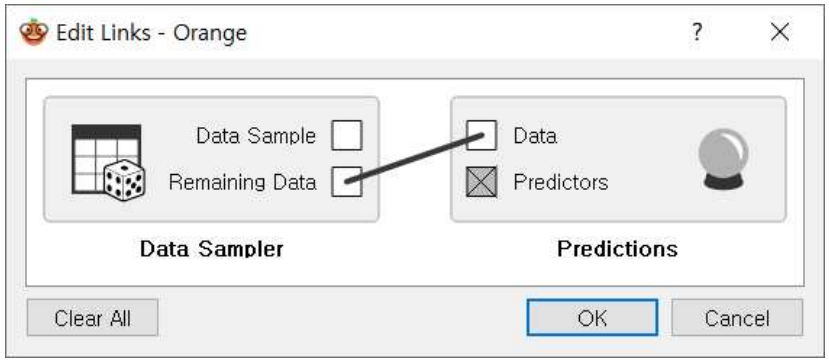




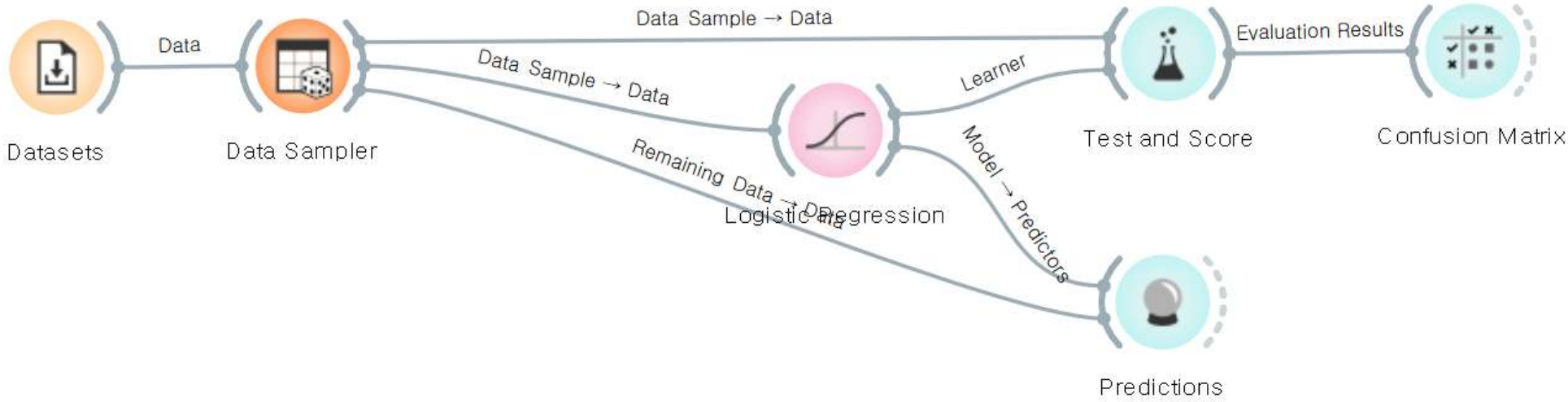
학습에 사용하는 데이터 (Data Sample)



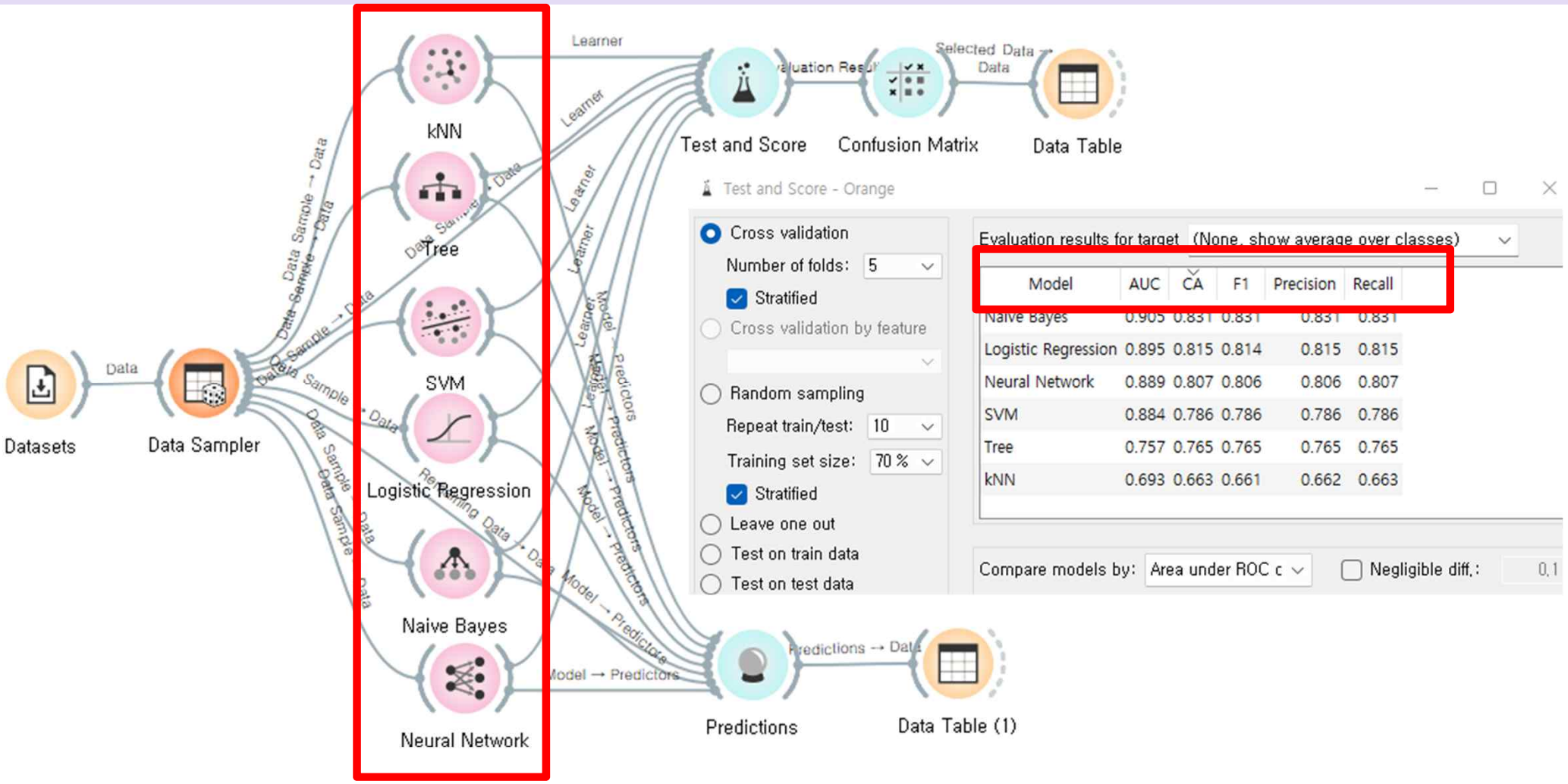
분류에 사용하는 새로운 데이터 (Remaining Data)



학습알고리즘 적용하기

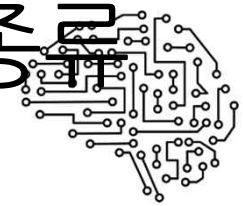

Data
Data Sample → Data
Data Sample → Data
Learner
Evaluation Results
Model → Predictors
Remaining Data → Data
Datasets
Data Sampler
Logistic Regression
Test and Score
Confusion Matrix
Predictions

여러 분류 알고리즘을 활용하여 분류기 성능을 평가해보자.



지도학습 - ‘정형데이터를 활용한 분류2’

1. 여러가지 다양한 분류알고리즘의 종류
2. 분류의 성능평가지표



orange 활용 데이터 분석 및 머신 러닝

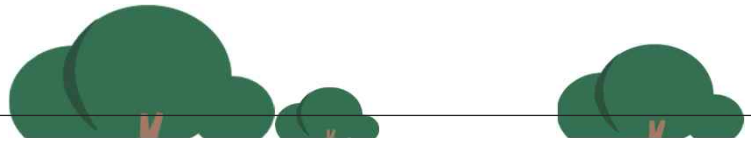
5차시

지도학습 - 분류 2

- 정형데이터를 활용한 분류 2
  1) 여러가지 다양한 분류알고리즘의 종류
  2) 분류의 성능평가지표

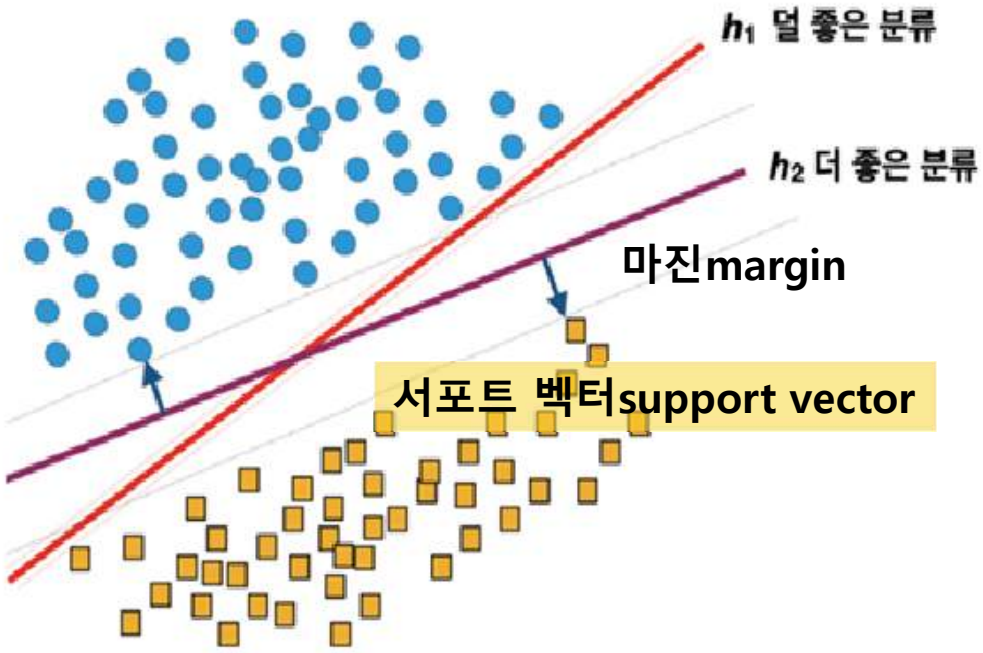
분류에 활용되는 다양한 머신러닝 알고리즘

- 서포트 벡터 머신 (SVM)
- 결정 트리 (Decision Tree)



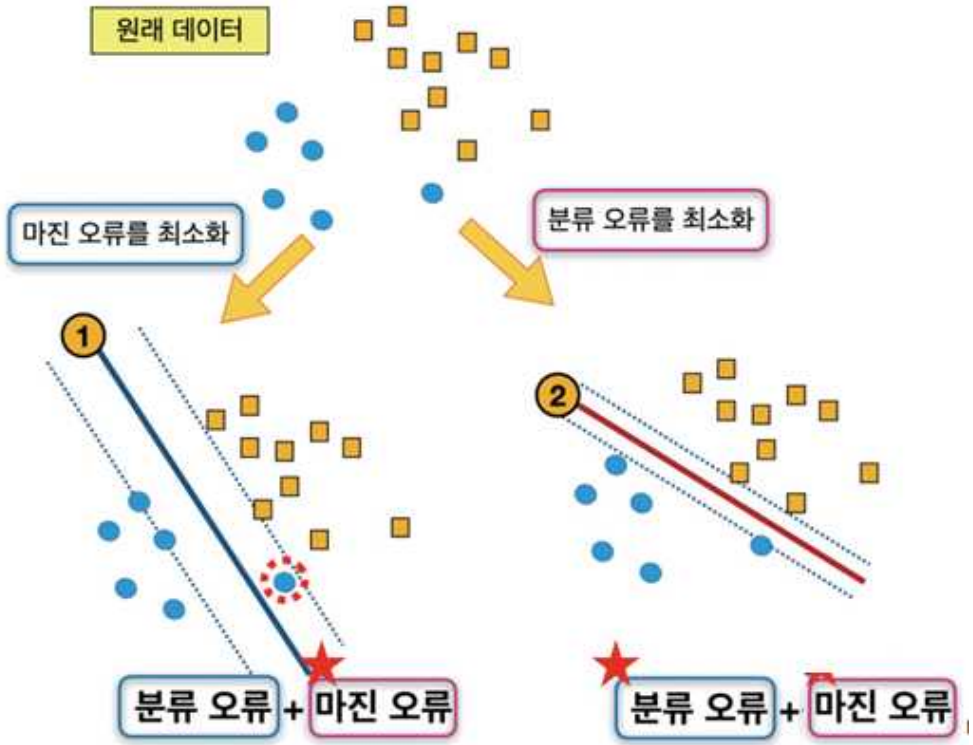
서포트 벡터 머신 (SVM)

- 딥러닝이 떠오르기 전까지 머신러닝 분야에서 가장 인기가 높은 데이터 분류 방법
- 현재의 데이터를 잘 분리하면서도, 새로운 데이터가 들어왔을 때에도 분리를 잘 할 수 있는 **결정경계**를 찾는 것
- 그러기 위해서는 이 평면을 화살표로 표시된 법선normal 벡터 방향으로 움직였을 때 데이터에 닿는 지점이 멀수록 좋을 것이다.



분류 오류와 마진 오류 최소화 사이의 트레이드오프

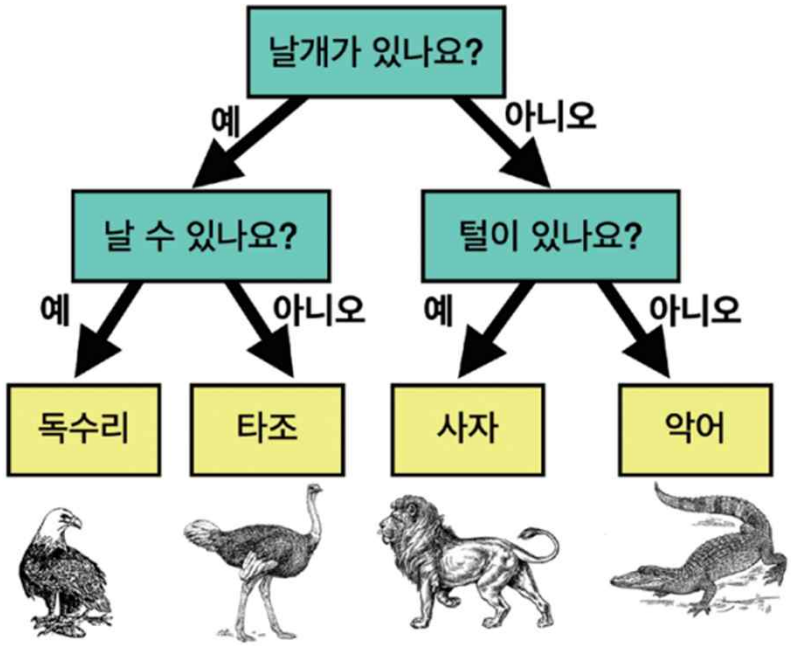
- SVM은 분류 오류와 마진 오류를 모두 줄이는 방향으로 학습을 진행하되 두 오류 중에서 어느 오류를 더 줄이는 방향으로 학습을 할지에 대한 명확한 정책을 정해야만 한다.



결정 트리(Decision Tree)와 분류

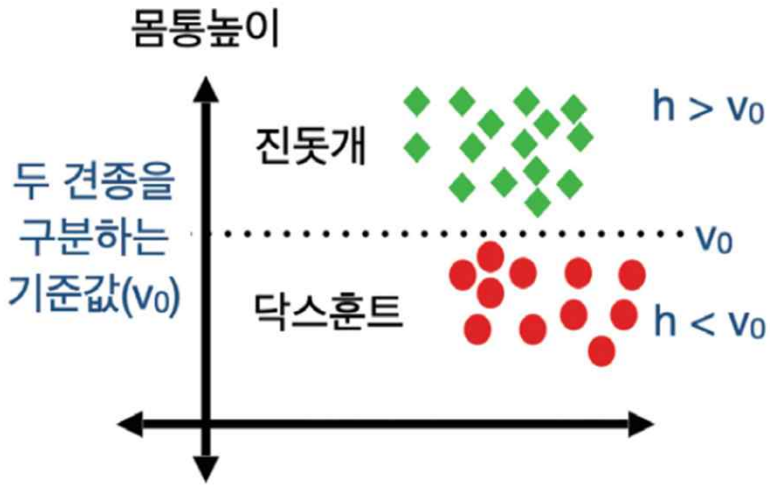
- 결정 트리는 스무고개를 하듯 예 / 아니오 질문을 이어가며 학습을 하는 알고리즘으로 귀납 추론을 위해 자주 사용되는 실용적인 방법
- 분류, 회귀, 다중출력까지 가능
- SVM과 결정트리 차이

- 데이터들을 트리 구조의 루트root에서 시작하여 차례로 중간 노드들을 거쳐 단말 노드에 배정하는 기능을 수행한다.



결정 트리(Decision Tree)와 분류

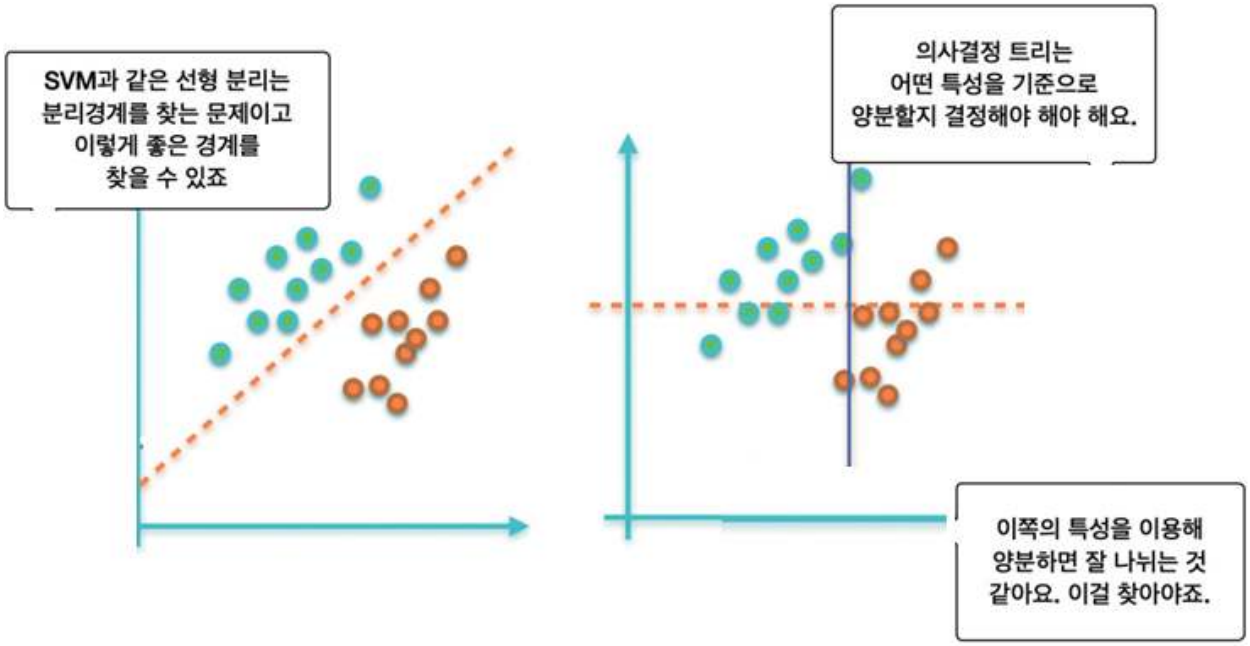
- 진돗개와 닥스훈트의 두 견종의 몸통 길이와 높이가 그림과 같이 분포할 때, 이 두 견종을 구분할 수 있는 몸통 높이(h)의 기준값 v_0가 존재한다고 가정하자.



- 이 경우 몸통 높이를 기준으로 $h > v_0$일 경우 진돗개로, $h < v_0$의 경우를 닥스훈트로 분류할 수 있을 것이다. 이렇게만 본다면 결정 트리와 SVM과 같은 선형 분리 문제는 동일한 문제인 것처럼 보인다.

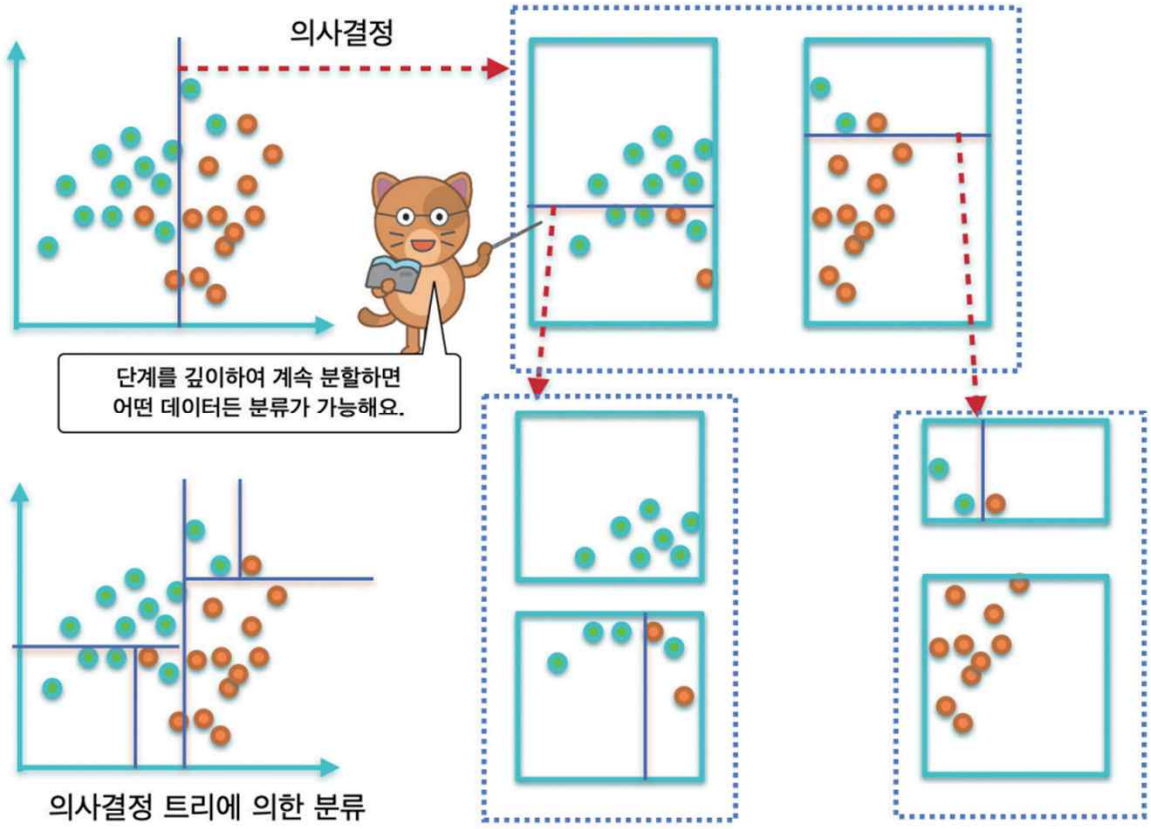
결정 트리(Decision Tree)와 분류

- 선형분리 문제는 아래 그림과 같이 특성 공간 내에서 자유롭게 선택할 수 있지만, 결정 트리는 하나의 특성을 기준으로 데이터를 양분하는 문제가 된다.
- 현실 세계의 데이터는 좀 더 복잡한 경우가 많아서 단순히 하나의 특성으로 깔끔하게 데이터를 양분할 수 없는 경우가 많다.
- 그러면 나누어진 각 영역에 대해 다시 다른 특성을 기준으로 양분하는 일을 해 나간다.

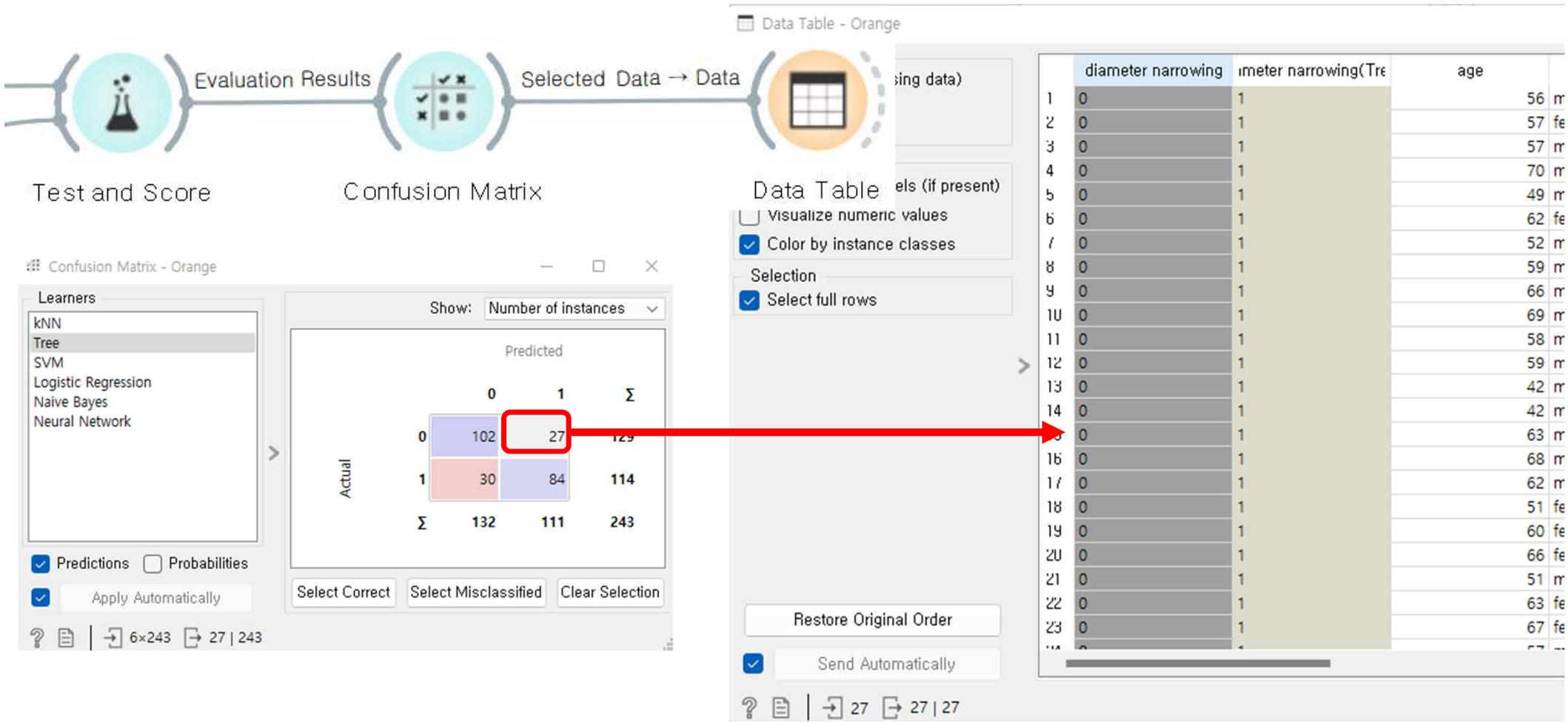


결정 트리(Decision Tree)와 분류

- 가능한 얕은 단계에서 데이터가 분리될 수 있도록 하는 것이 좋다. : 단계를 깊게하여 분리하는 것은 해당 데이터에는 잘 동작하지만, 새로운 데이터에는 잘 적용되지 않는 과적합 발생
- 가능한 얕은 단계에서 데이터를 잘 분리할 수 있도록 하기 위해 가장 데이터 분리를 잘하는 좋은 특성을 찾아, 해당 특성 내에서 어떤 값을 기준으로 분리를 할 것인지를 결정해야 한다.



혼동행렬 (Confusion Matrix)



평가지표 (분류)

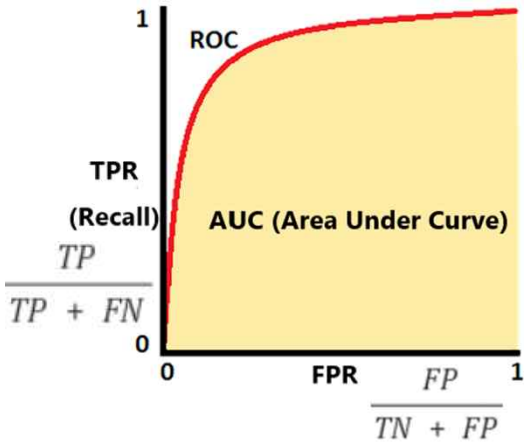
		분류결과	
		Positive	**N**egative
실제값	Positive	**T**rue **P**ositive	**F**alse **N**egative
	Negative	**F**alse **P**ositive	**T**rue **N**egative

혼동행렬(Confusion Matrix)

ROC curve 여러 임계치들을 기준으로

TPR-FPR의 변화를 시각화한 것

AUC ROC 그래프 아래의 면적



Precision(정밀도) 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율 $(Precision) = \frac{TP}{TP + FP}$

Recall(재현율) 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율 $(Recall) = \frac{TP}{TP + FN}$

Accuracy(CA, 정확도) 전체 중 실제 True를 True라고, 실제 False를 False라고 예측한 것의 비율 $(Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}$

F1 score Precision과 Recall의 조화평균

$$(F1\text{-}score) = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

분류 평가지표의 예

정확도: 전체 데이터(FP+FN+TP+TN)중에서 제대로 판정한 데이터(TP + TN)의 비율

재현율: 양성환자 중에서 이 키트가 올바르게 양성이라고 분류한 환자의 비율

정밀도: 검사 키트가 확진자로 분류한 사람들 중 실제 양성인 환자의 비율

$$Acc = \frac{TP+TN}{FP+FN+TP+TN} = \frac{95+89}{11+5+95+89} = 0.92$$

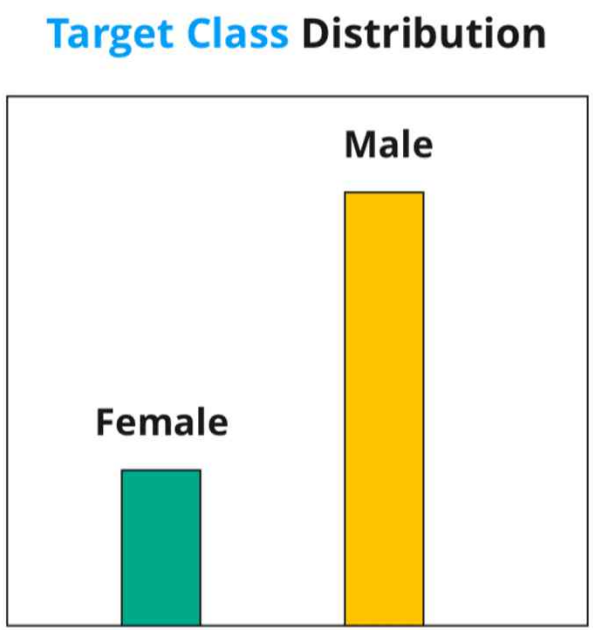
$$TPR = Rec = \frac{TP}{P} = \frac{TP}{FN+TP} = \frac{95}{100} = 0.95$$

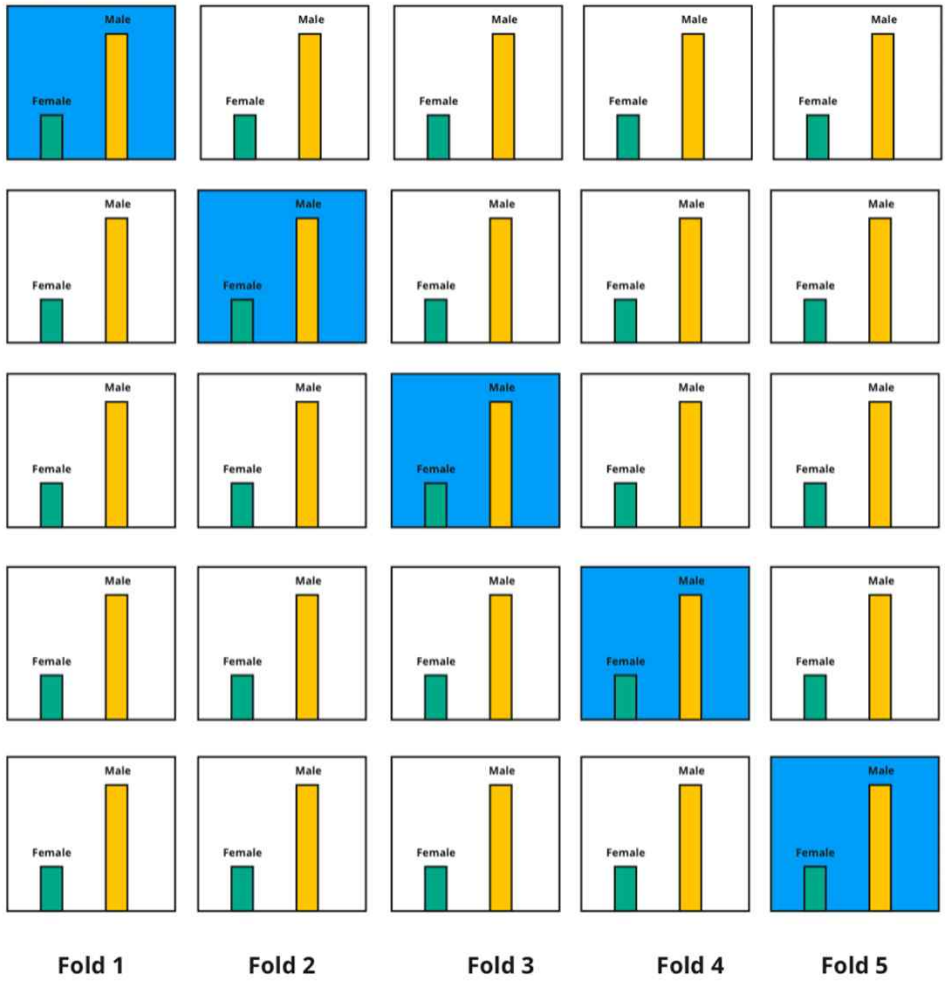
$$Pre = \frac{TP}{TP+FP} = \frac{95}{106} = 0.896$$

		KoKIT22의 예측값 (검사결과)					
		음성			양성		
환자의 실제 상태값		N			P		
음성 (COVID 안걸림)	N	89 TN	T 일치	N 예측	11 FP	F 불일치	P 예측
양성 (COVID 걸림)	P	5 FN	F 불일치	N 예측	95 TP	T 일치	P 예측

교차 검증 (학습용 세트가 적을 때 주로 사용)

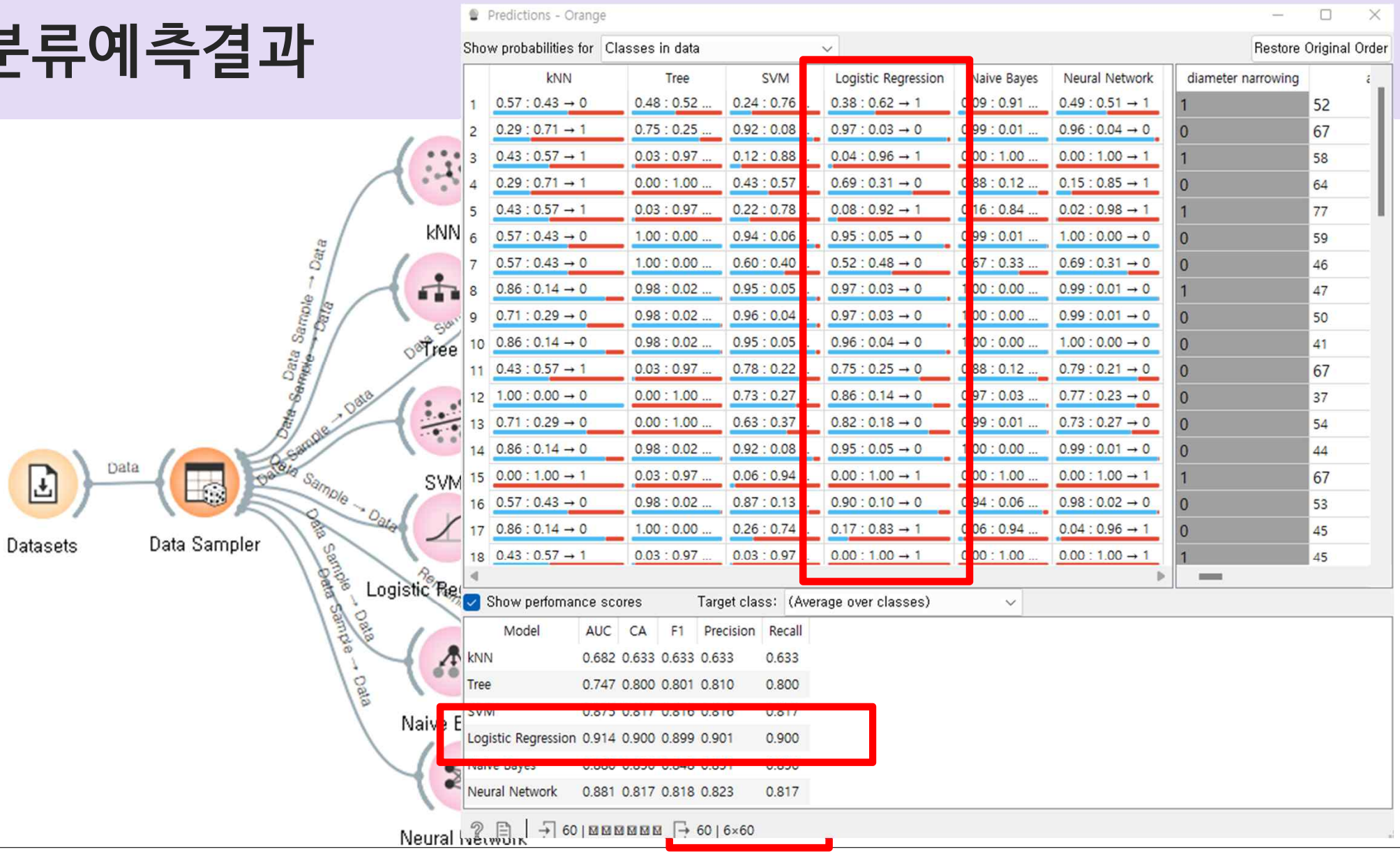
Stratified k-fold (계층별 k겹 교차 검증)





- 데이터가 편향되어 있는 경우, 데이터 클래스 별 분포를 고려해 나눔
- 분류 모델에 사용

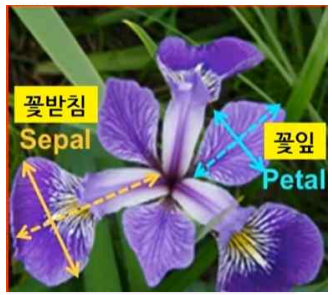
분류예측결과


Predictions - Orange
Show probabilities for
Classes in data
Restore Original Order
kNN
Tree
SVM
Logistic Regression
Naive Bayes
Neural Network
diameter narrowing
1 0.57 : 0.43 → 0 0.48 : 0.52 ... 0.24 : 0.76 ... 0.38 : 0.62 → 1 0.09 : 0.91 ... 0.49 : 0.51 → 1 1 52
2 0.29 : 0.71 → 1 0.75 : 0.25 ... 0.92 : 0.08 ... 0.97 : 0.03 → 0 0.99 : 0.01 ... 0.96 : 0.04 → 0 0 67
3 0.43 : 0.57 → 1 0.03 : 0.97 ... 0.12 : 0.88 ... 0.04 : 0.96 → 1 0.00 : 1.00 ... 0.00 : 1.00 → 1 1 58
4 0.29 : 0.71 → 1 0.00 : 1.00 ... 0.43 : 0.57 ... 0.69 : 0.31 → 0 0.88 : 0.12 ... 0.15 : 0.85 → 1 0 64
5 0.43 : 0.57 → 1 0.03 : 0.97 ... 0.22 : 0.78 ... 0.08 : 0.92 → 1 0.16 : 0.84 ... 0.02 : 0.98 → 1 1 77
6 0.57 : 0.43 → 0 1.00 : 0.00 ... 0.94 : 0.06 ... 0.95 : 0.05 → 0 0.99 : 0.01 ... 1.00 : 0.00 → 0 0 59
7 0.57 : 0.43 → 0 1.00 : 0.00 ... 0.60 : 0.40 ... 0.52 : 0.48 → 0 0.67 : 0.33 ... 0.69 : 0.31 → 0 0 46
8 0.86 : 0.14 → 0 0.98 : 0.02 ... 0.95 : 0.05 ... 0.97 : 0.03 → 0 1.00 : 0.00 ... 0.99 : 0.01 → 0 1 47
9 0.71 : 0.29 → 0 0.98 : 0.02 ... 0.96 : 0.04 ... 0.97 : 0.03 → 0 1.00 : 0.00 ... 0.99 : 0.01 → 0 0 50
10 0.86 : 0.14 → 0 0.98 : 0.02 ... 0.95 : 0.05 ... 0.96 : 0.04 → 0 1.00 : 0.00 ... 1.00 : 0.00 → 0 0 41
11 0.43 : 0.57 → 1 0.03 : 0.97 ... 0.78 : 0.22 ... 0.75 : 0.25 → 0 0.88 : 0.12 ... 0.79 : 0.21 → 0 0 67
12 1.00 : 0.00 → 0 0.00 : 1.00 ... 0.73 : 0.27 ... 0.86 : 0.14 → 0 0.97 : 0.03 ... 0.77 : 0.23 → 0 0 37
13 0.71 : 0.29 → 0 0.00 : 1.00 ... 0.63 : 0.37 ... 0.82 : 0.18 → 0 0.99 : 0.01 ... 0.73 : 0.27 → 0 0 54
14 0.86 : 0.14 → 0 0.98 : 0.02 ... 0.92 : 0.08 ... 0.95 : 0.05 → 0 1.00 : 0.00 ... 0.99 : 0.01 → 0 0 44
15 0.00 : 1.00 → 1 0.03 : 0.97 ... 0.06 : 0.94 ... 0.00 : 1.00 → 1 0.00 : 1.00 ... 0.00 : 1.00 → 1 1 67
16 0.57 : 0.43 → 0 0.98 : 0.02 ... 0.87 : 0.13 ... 0.90 : 0.10 → 0 0.94 : 0.06 ... 0.98 : 0.02 → 0 0 53
17 0.86 : 0.14 → 0 1.00 : 0.00 ... 0.26 : 0.74 ... 0.17 : 0.83 → 1 0.06 : 0.94 ... 0.04 : 0.96 → 1 0 45
18 0.43 : 0.57 → 1 0.03 : 0.97 ... 0.03 : 0.97 ... 0.00 : 1.00 → 1 0.00 : 1.00 ... 0.00 : 1.00 → 1 1 45
Show perfomance scores
Target class: (Average over classes)
Model AUC CA F1 Precision Recall
kNN 0.682 0.633 0.633 0.633 0.633
Tree 0.747 0.800 0.801 0.810 0.800
Logistic Regression 0.914 0.900 0.899 0.901 0.900
Neural Network 0.881 0.817 0.818 0.823 0.817
60 | 60 | 6×60
Datasets
Data
Data Sampler
Data Sample → Data

[활용] 아이리스 데이터셋을 활용한 품종 분류

◦ 붓꽃(아이리스) 데이터셋 https://www.kaggle.com/uciml/iris

Iris setosa



Iris versicolor

Iris virginica

◦ 문제정의

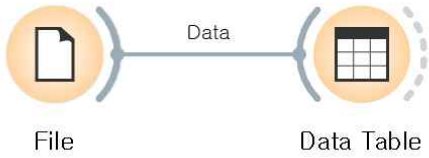
◦ 붓꽃의 꽃받침(Sepal)과 꽃잎(Petal)의 길이와 너비에 따라 3가지 종류(setosa, versicolor, virginica)로 구별

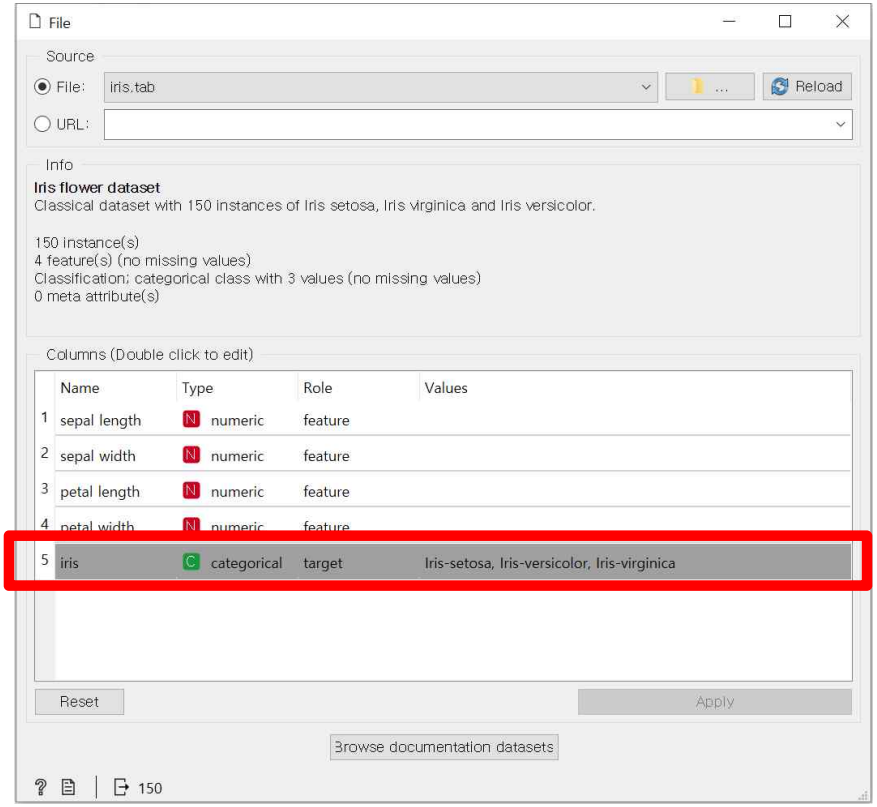
◦ Iris(붓꽃) 데이터셋은 1936년에 영국 통계 학자 및 생물학자인 로널드 피셔(Ronald Fisher)가 다변량 통계 분석 데이터셋으로 소개

◦ 해당 데이터는 총 150개로 이루어져 있으며, 꽃받침(Sepal)과 꽃잎(Petal)의 길이와 너비의 특성으로 3가지 품종(setosa, versicolor, virginica)으로 구별

◦ 붓꽃 데이터는 측정이 잘 되어 있어 기계학습 알고리즘을 학습하는데 데이터로 활용, 데이터분석의 “Hello World”

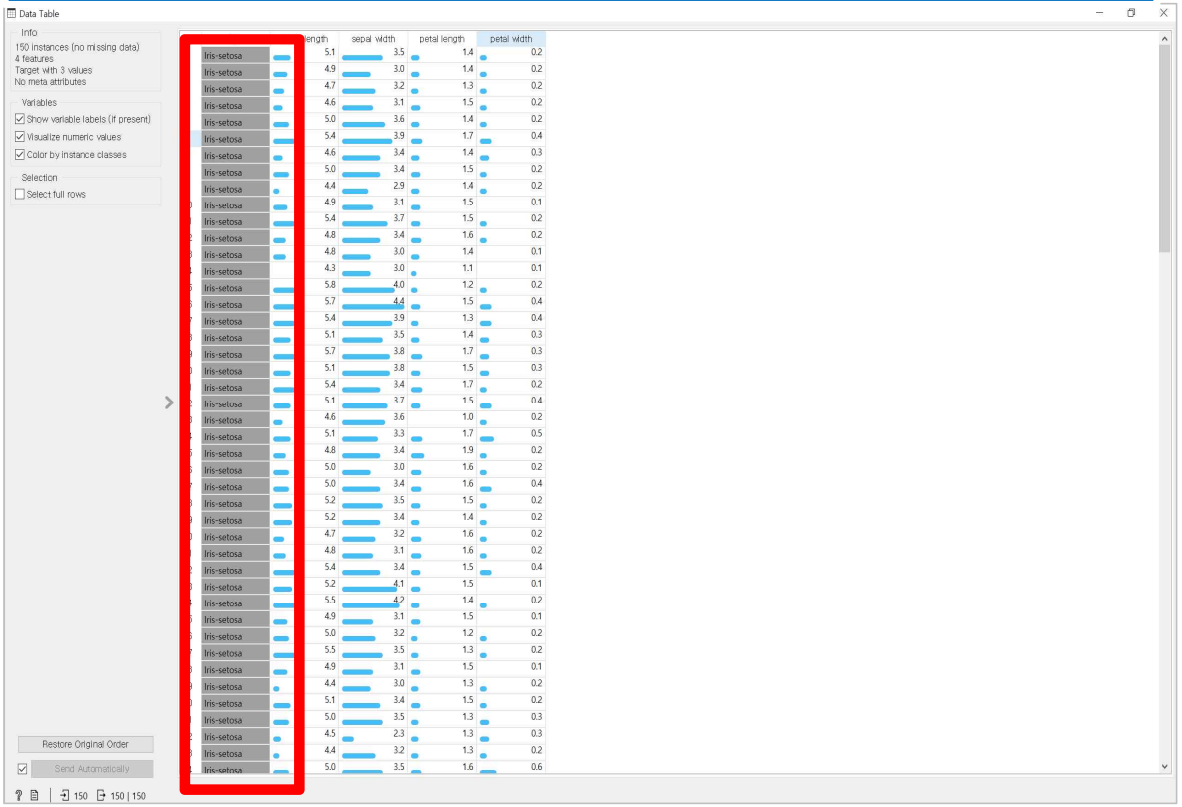
데이터 준비 및 탐색



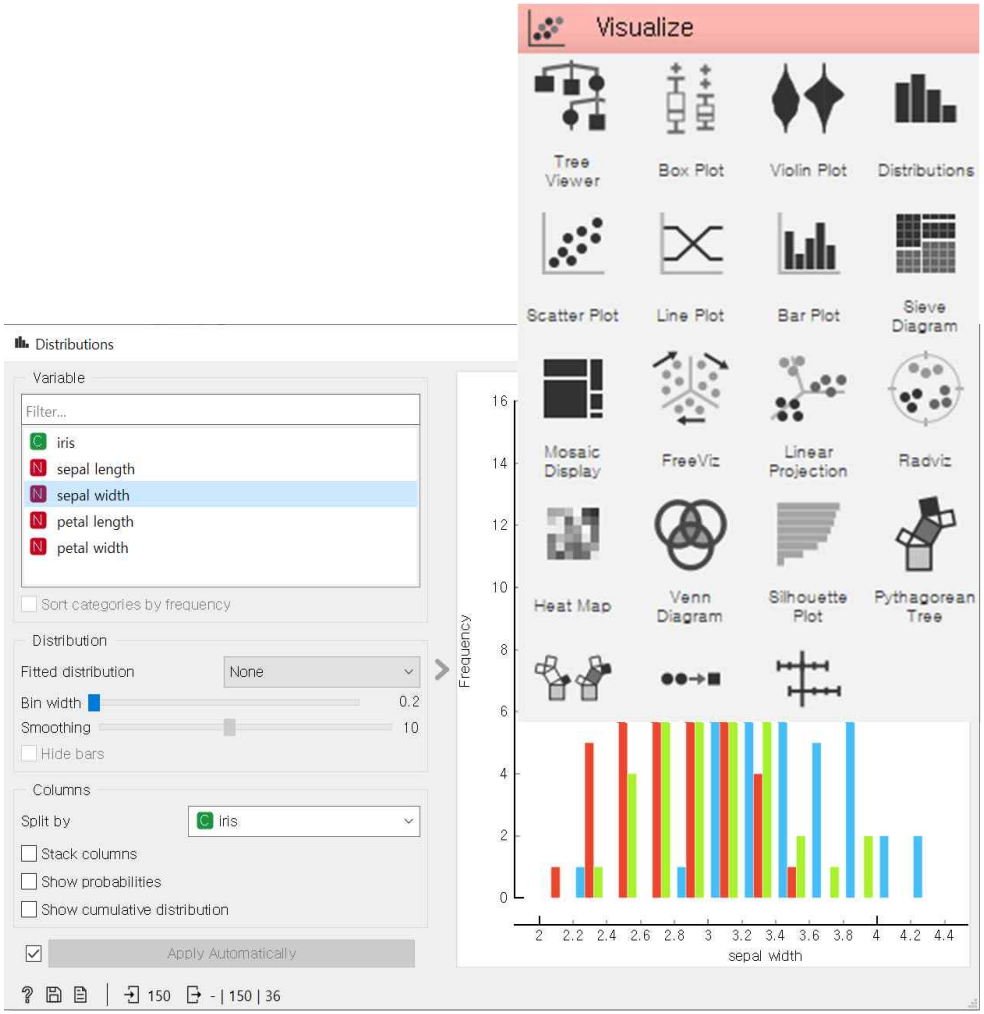


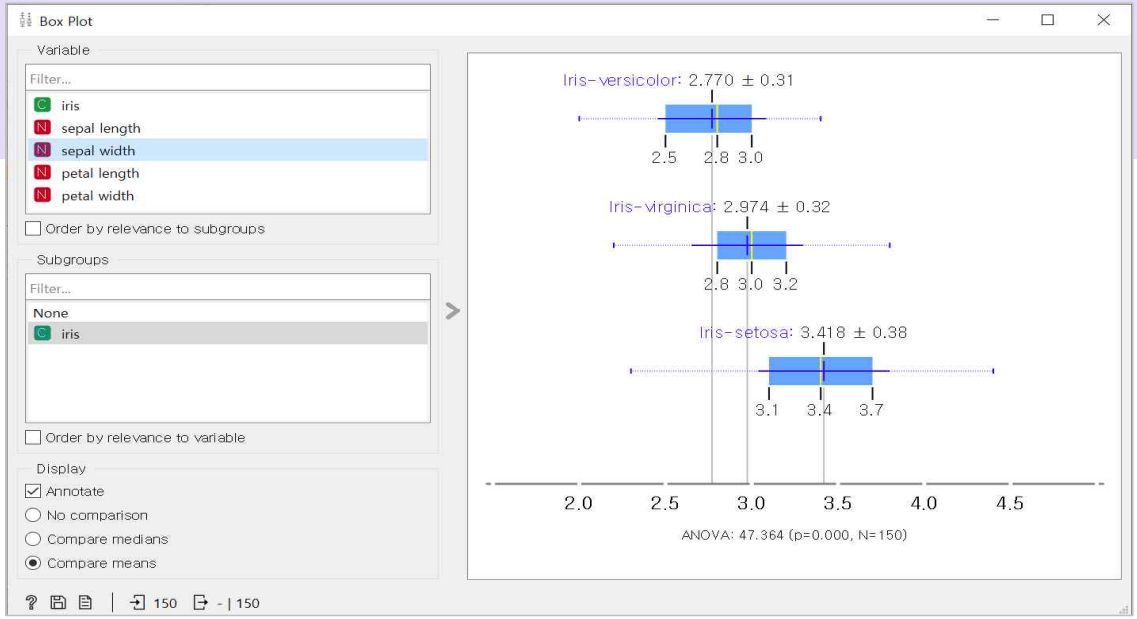
Role

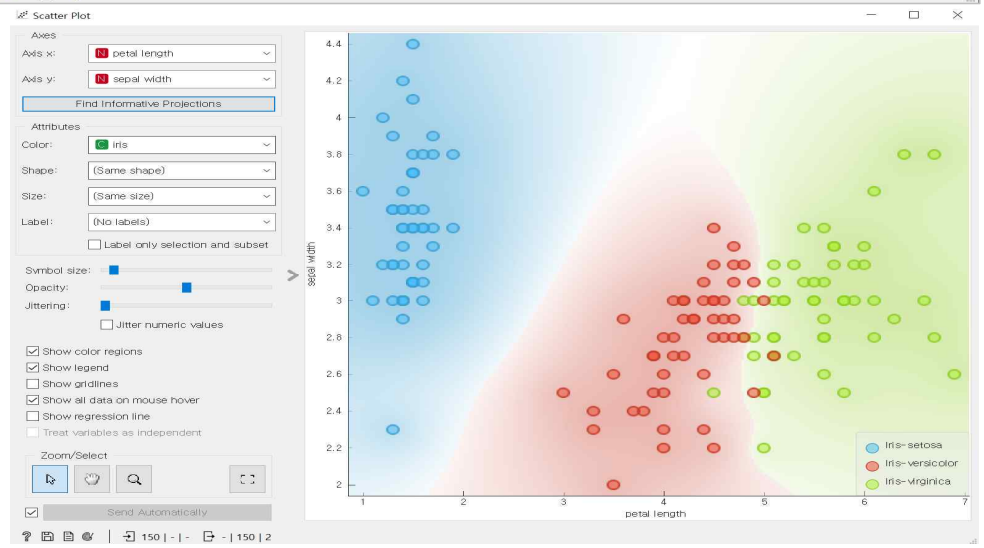
- Skip : 분석 작업에서 아예 무시해도 되는 데이터는 건너뛰기
- Meta : 실제로 작업에 사용하진 않지만, 정보성 데이터로 남겨 두는게 좋은 것은 meta로 표시
- Target : 예측하고자 하는 열(종속변수)을 target으로 지정
- Feature : 예측하는 데 사용될 원인(독립변수)로 지정



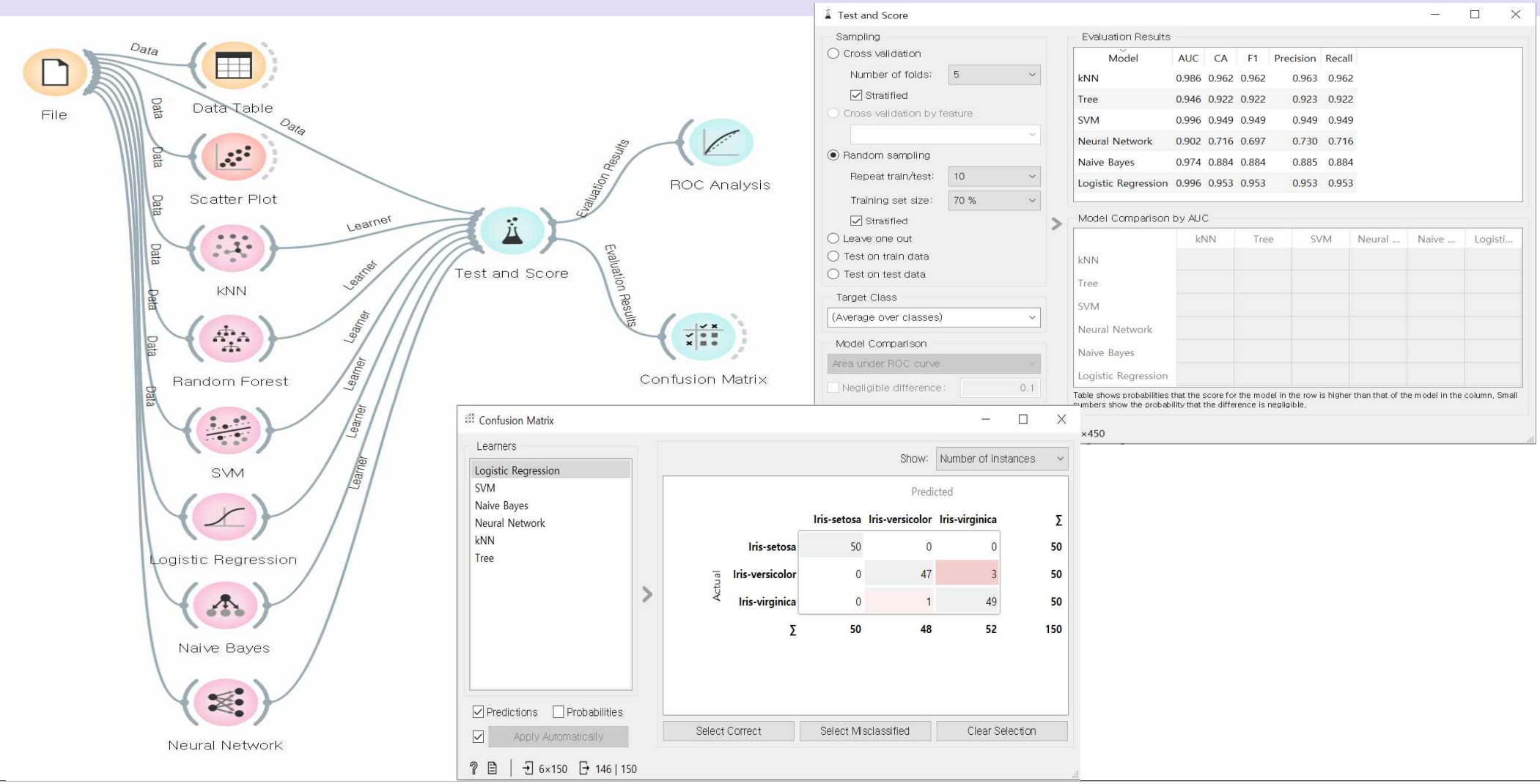
데이터 시각화


Visualize
Tree Viewer
Box Plot
Violin Plot
Distributions
Scatter Plot
Line Plot
Bar Plot
Sieve Diagram
Mosaic Display
FreeViz
Linear Projection
Radviz
Heat Map
Venn Diagram
Silhouette Plot
Pythagorean Tree
Distributions
Variable
Filter...
iris
sepal length
sepal width
petal length
petal width
Sort categories by frequency
Distribution
Fitted distribution
None
Bin width
0.2
Smoothing
10
Hide bars
Columns
Split by
iris
Stack columns
Show probabilities
Show cumulative distribution
Apply Automatically
150
- | 150 | 36
Frequency
sepal width


Box Plot
Variable
Filter...
iris
sepal length
sepal width
petal length
petal width
Order by relevance to subgroups
Subgroups
Filter...
None
iris
Order by relevance to variable
Display
Annotate
No comparison
Compare medians
Compare means
150
- | 150
Iris-versicolor: 2.770 ± 0.31
Iris-virginica: 2.974 ± 0.32
Iris-setosa: 3.418 ± 0.38
ANOVA: 47.364 (p=0.000, N=150)


Scatter Plot
Axes
Axis x:
petal length
Axis y:
sepal width
Find Informative Projections
Attributes
Color:
iris
Shape:
(Same shape)
Size:
(Same size)
Label:
(No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Show gridlines
Show all data on mouse hover
Show regression line
Treat variables as independent
Zoom/Select
Send Automatically
Iris-setosa
Iris-versicolor
Iris-virginica
petal length
sepal width

모델 설정, 학습 , Test and Score



다음 시간에는 **비지도학습 - 군집화**

‘정형데이터를 활용한 군집화와 지도 시각화’에 대해 알아봅시다.