

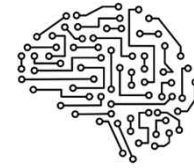


orange 활용 데이터 분석 및 머신 러닝



2차시

지도학습 - 회귀1



간단한 정형데이터를 이용하여
선형회귀모델을 만들어봅시다.

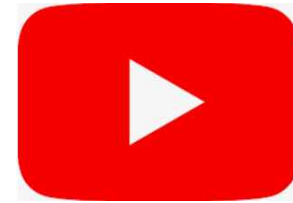
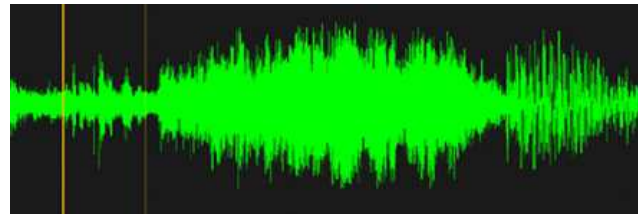
데이터의 종류 - ① 정형 데이터(Structured Data)

ID	Name	AGE	SEX
01	KIM	32	M
02	LEE	26	F
03	PARK	72	F
04	CHOI	15	M

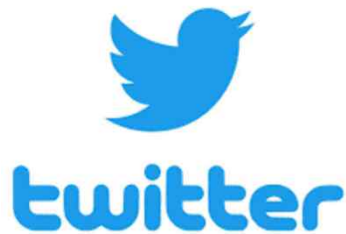
	A	B	C	D	E	F
1	번호	이름	국어	수학	사회	과학
2	1	김석진	85	90	95	80
3	2	민윤기	80	95	100	70
4	3	정호석	85	90	85	95
5	4	김남준	100	80	90	90
6	5	박지민	80	95	90	100
7	6	김태형	90	100	80	100
8	7	전정국	85	75	100	85

[참고 : ITA정보통신용어사전](#)

데이터의 종류 - ② 비정형 데이터(Unstructured Data)



.jpg / .txt / .mp3 / .doc



facebook



지도학습, 비지도학습, 강화학습

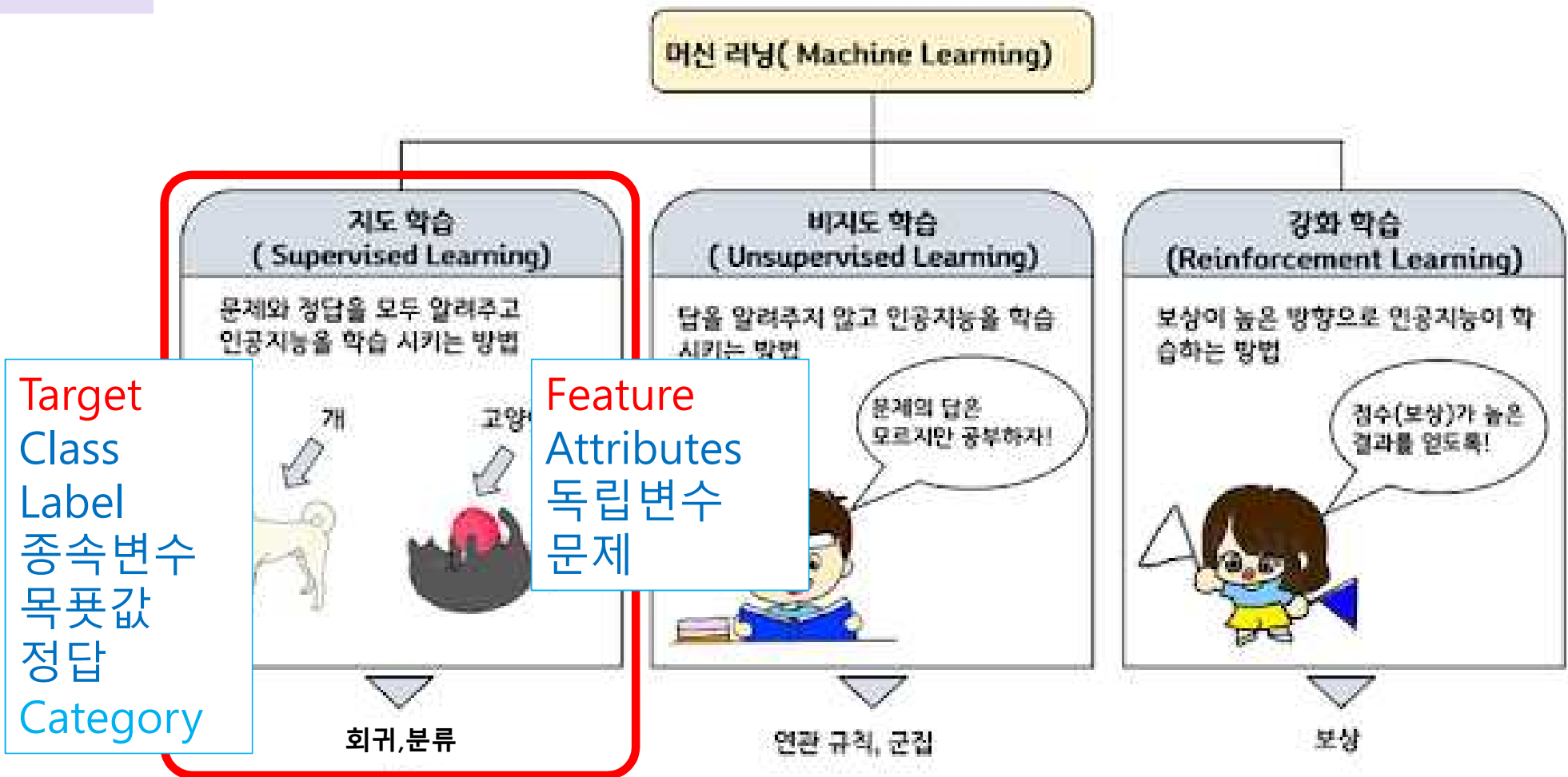
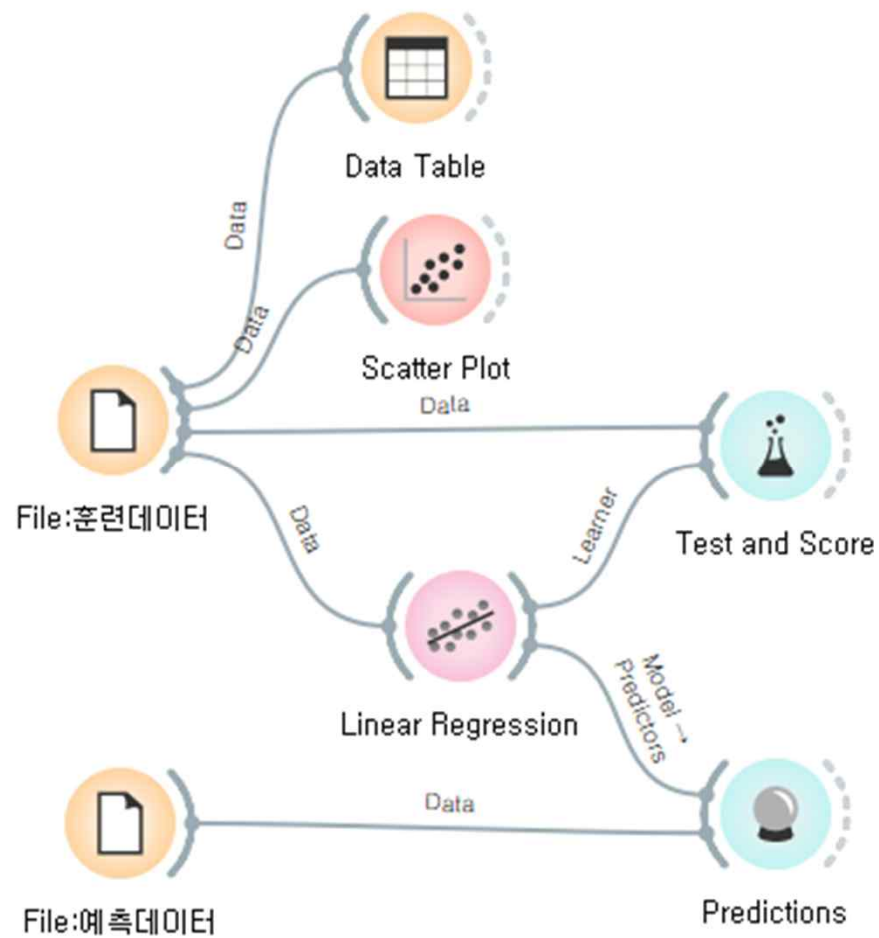
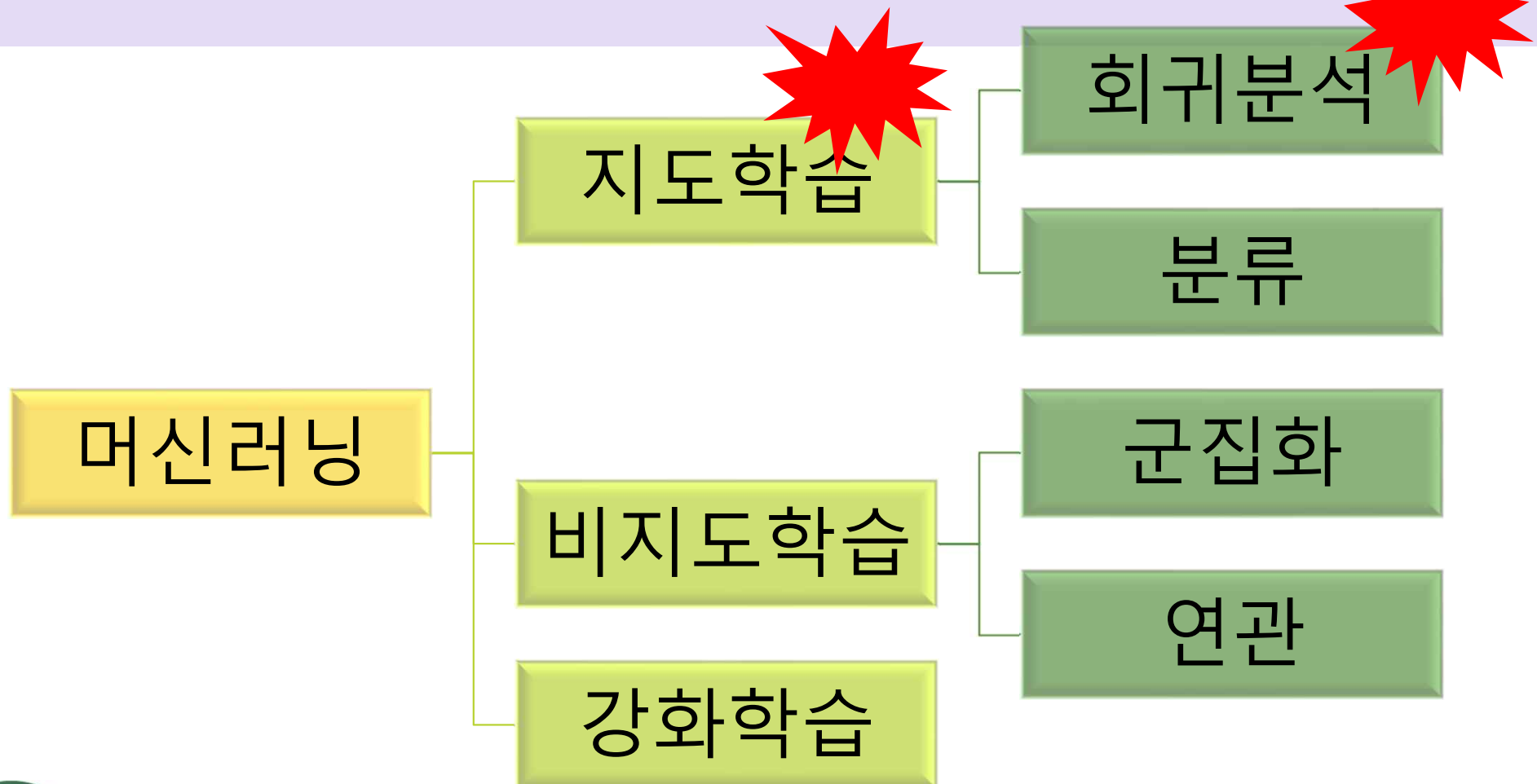


그림 <http://ai4school.org>

아주 단순한 선형회귀



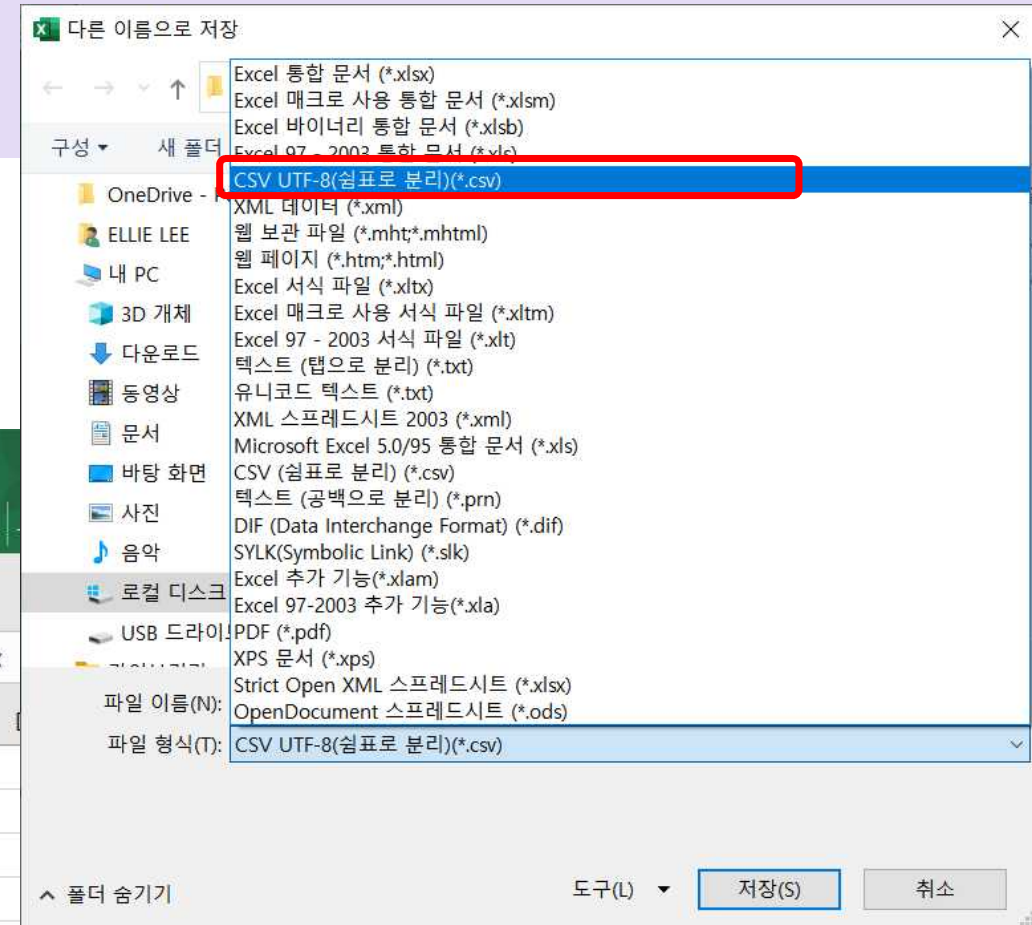


데이터 만들기

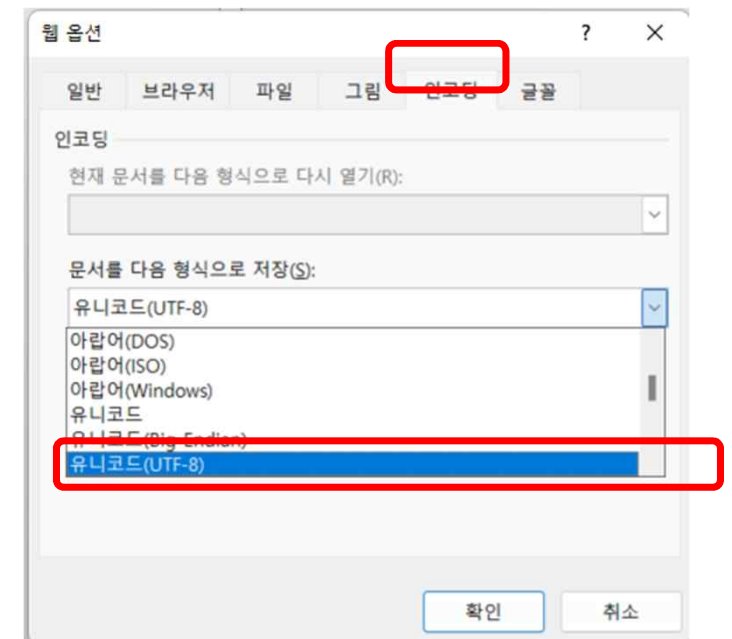
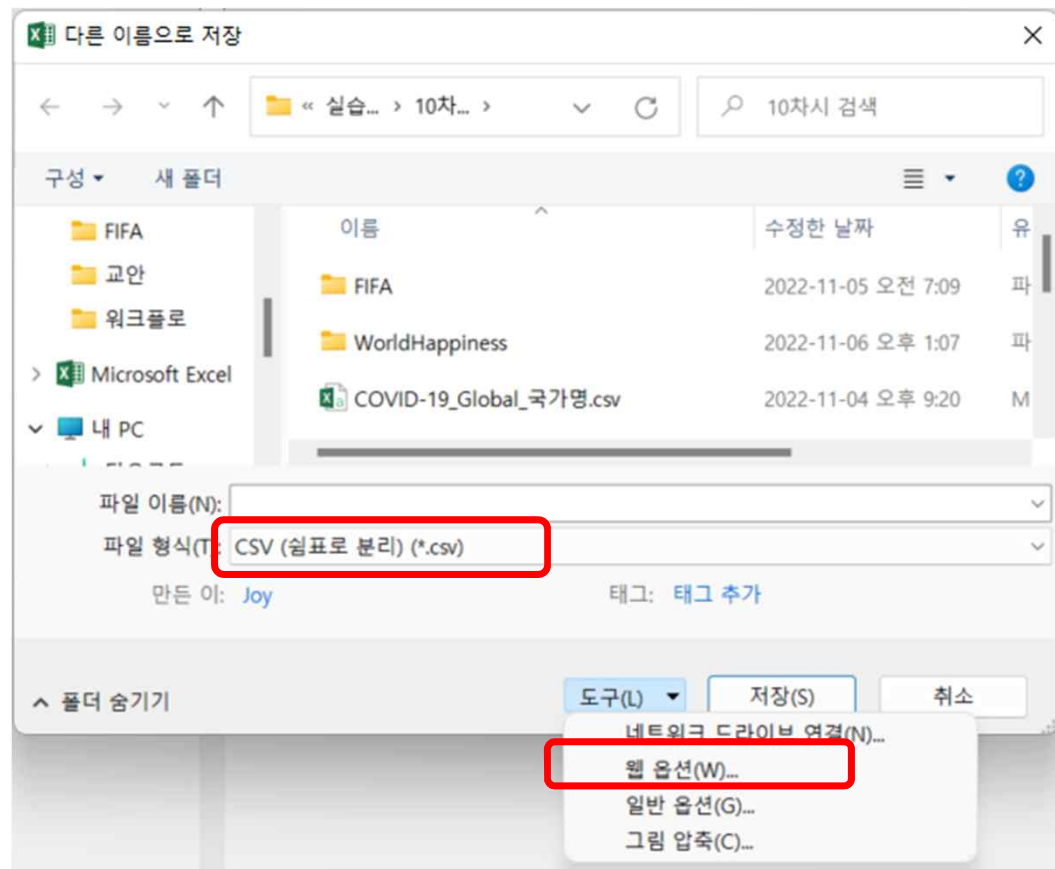
CSV UTF-8(쉼표로 분리)(*.csv) 선택하여
각각 **train.csv**와 **test.csv** 로 저장

	A	B	C	D
1	x	y		
2	1	1		
3	2	2		
4	3	3		
5	4	4		
6	5	5		
7				
8				

	A	B	C	D
1	x			
2		6		
3		7		
4		8		
5		9		
6		10		
7				
8				



- CSV UTF-8(쉼표로 분리)(*.csv) 가 없을 경우 CSV (쉼표로 분리)(*.csv) 선택 -> 도구



데이터 가져오기

train.csv →

File

File (1)

test.csv →

File - Orange

Source

☒ File: train.csv ... Reload

☐ URL: JPAdr_yvitTMjo2VXT_E3ktsdcnPDr5eO3lO0/edit?usp=sharing

File Type

Automatically detect type

Info

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

Columns (Double click to edit)

	Name	Type	Role	Values
1	x	N numeric	feature	
2	y	N numeric	target	

Reset Apply

Browse documentation datasets

? | 5

File (1) - Orange

Source

☒ File: test.csv ... Reload

☐ URL: JPAdr_yvitTMjo2VXT_E3ktsdcnPDr5eO3lO0/edit?usp=sharing

File Type

Automatically detect type

Info

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

Columns (Double click to edit)

	Name	Type	Role	Values
1	x	N numeric	feature	

Reset Apply

Browse documentation datasets

? | 5

구글 스프레드시트에서 데이터 읽어오기

The screenshot shows a Google Sheets interface with a spreadsheet titled 'train.csv'. A sharing dialog box is open, titled '"train.csv" 공유'. The dialog shows the user 'ELLIE LEE(나)' as the owner. Under the '일반 액세스' (General Access) section, the option '링크가 있는 모든 사용자' (Anyone with the link) is selected and highlighted with a red box. Below this, there is a '링크 복사' (Copy link) button and an '완료' (Done) button.

train.csv - Google Sheets

docs.google.com/spreadsheets/d/17Y9yHGhSk3-0G1DYlBUayTqa_IO62zZly7tKxWG7VMI/edit#gid=0

train.csv

파일 수정 보기 삽입 서식 데이터 도구 확장 프로그램 도움말 4분 전에 마지막으로 수정했...

100% W % .0 .00 123 기본값 (Arl... 10 B I U A

K26 fx

1 x y

2 1 1

3 2 2

4 3 3

5 4 4

6 5 5

"train.csv" 공유

사용자 및 그룹 추가

액세스 권한이 있는 사용자

ELLIE LEE(나) 소유자
lee.ellie@gmail.com

일반 액세스

링크가 있는 모든 사용자 링크가 있는 인터넷상의 모든 사용자가 볼 수 있음 뷰어

링크 복사 완료

The screenshot shows the Orange data mining software interface. The 'File' widget is selected, and its configuration window is open. The 'Source' tab is active, showing 'train.csv' as the file source. The 'URL' option is selected and highlighted with a red box, with the URL '/9yHGhSk3-0G1DYlBUayTqa_IO62zZly7tKxWG7VMI/edit?usp=sharing' entered. The 'File Type' is set to 'Automatically detect type'. The 'Info' section shows 5 instance(s), 2 feature(s), and 0 meta attribute(s). The 'Columns' section shows two columns: 'x' (numeric, feature) and 'y' (numeric, target). The 'y' column is highlighted with a red box. The 'Reset' and 'Apply' buttons are at the bottom.

File - Orange

Source

File: train.csv Reload

URL: /9yHGhSk3-0G1DYlBUayTqa_IO62zZly7tKxWG7VMI/edit?usp=sharing

File Type

Automatically detect type

Info

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

Columns (Double click to edit)

	Name	Type	Role	Values
1	x	numeric	feature	
2	y	numeric	target	

Reset Apply

Browse documentation datasets

데이터 분석 과정 알아보기

- 훈련 데이터(train.csv) : 학습용
- 테스트 데이터(test.csv) : 모델 학습 결과 평가용

Feature
Attributes
독립변수
문제

	A	B	C
1	x	y	
2		1	2
3		2	4
4		3	6
5		4	8
6		5	10

train.csv

Target
Class
Label
종속변수
목표값
정답
Category

Feature 제시

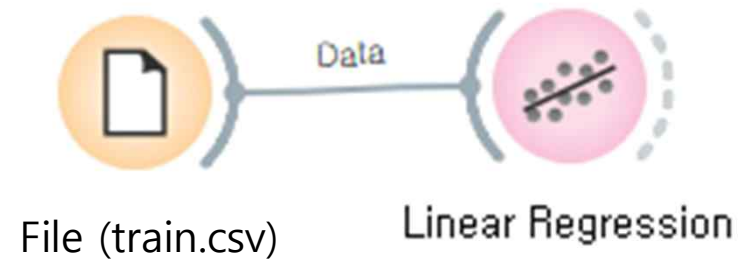
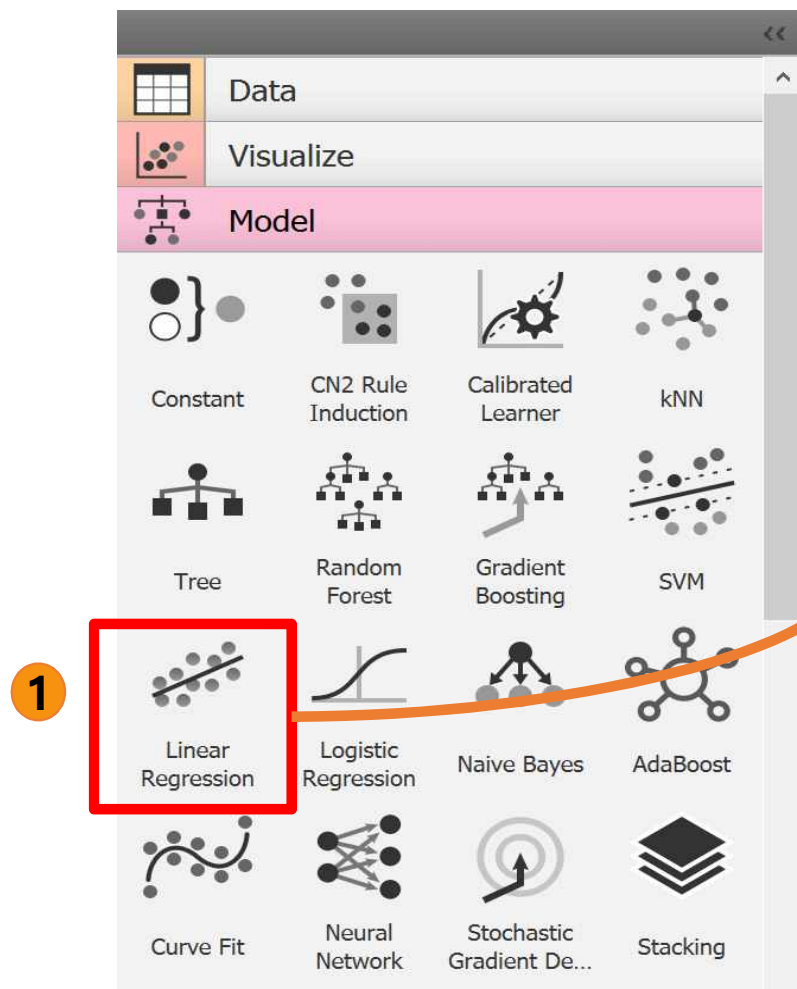
Target 예측

	A	B	C
1	x		
2		5	
3		6	
4		7	
5		8	
6		9	
7		10	
8			
9			

test.csv

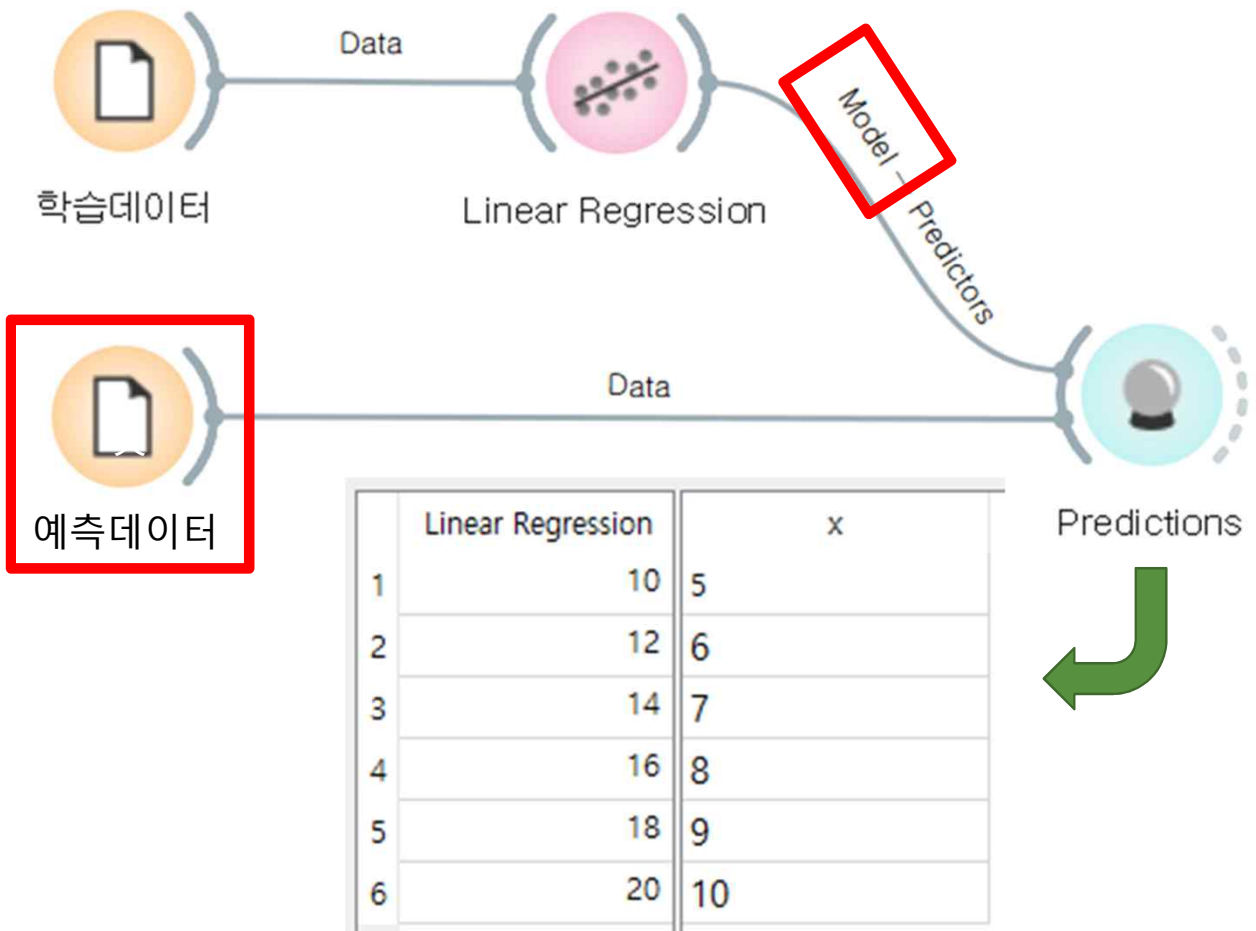
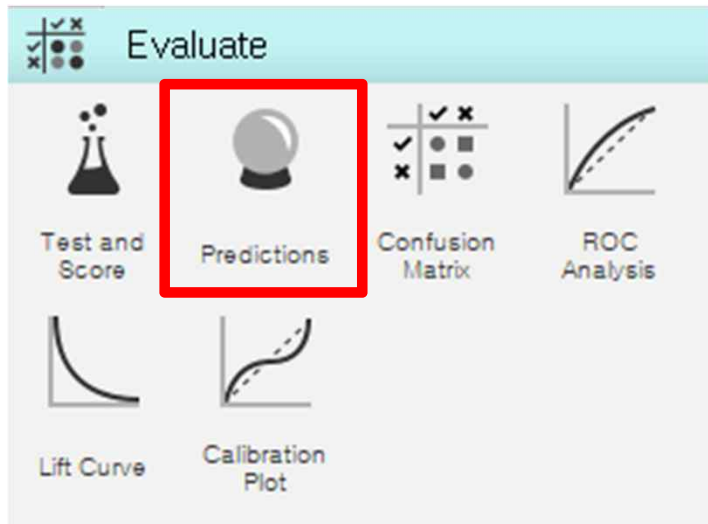
- feature(특성): 기계학습에 영향을 미치는 원인이 되는 속성, 독립변수
- target(목표값): 예측하고자 하는 결과가 되는 속성, 종속변수

학습알고리즘 적용

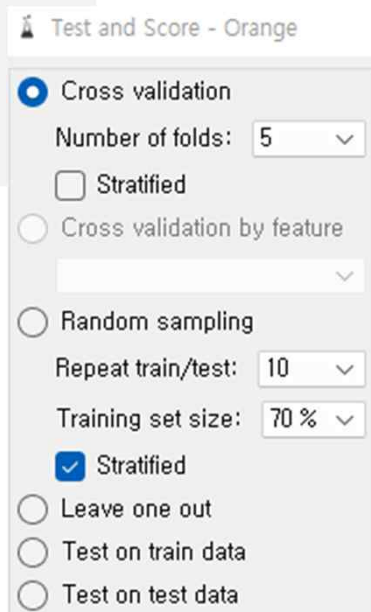
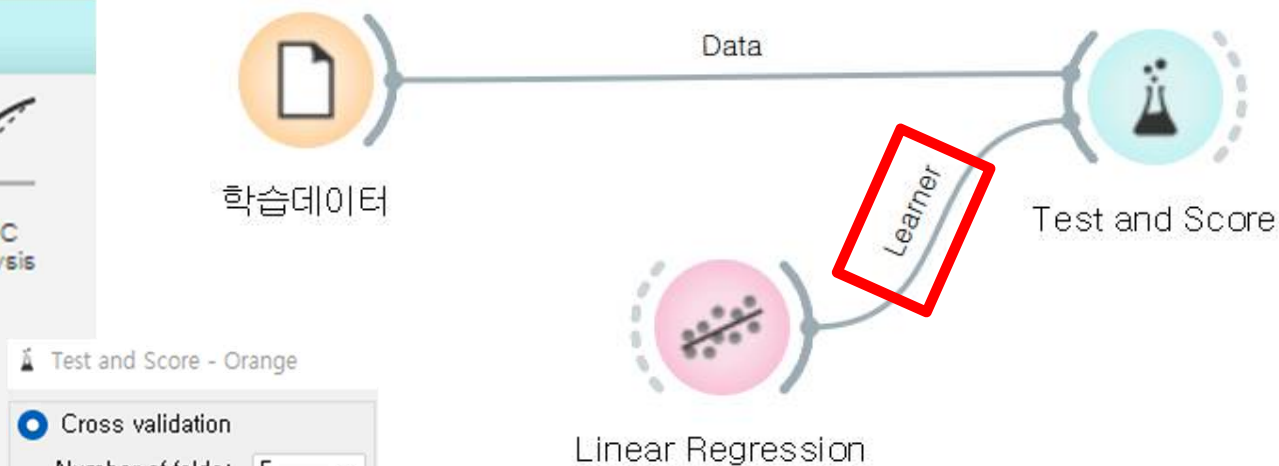
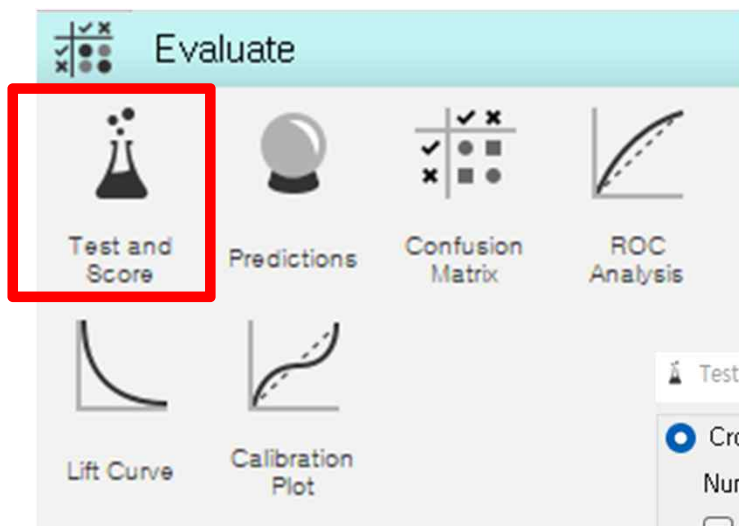


연결선을 연결하면 자동으로 위젯이 실행되어
모델이 훈련 데이터를 학습

예측하기

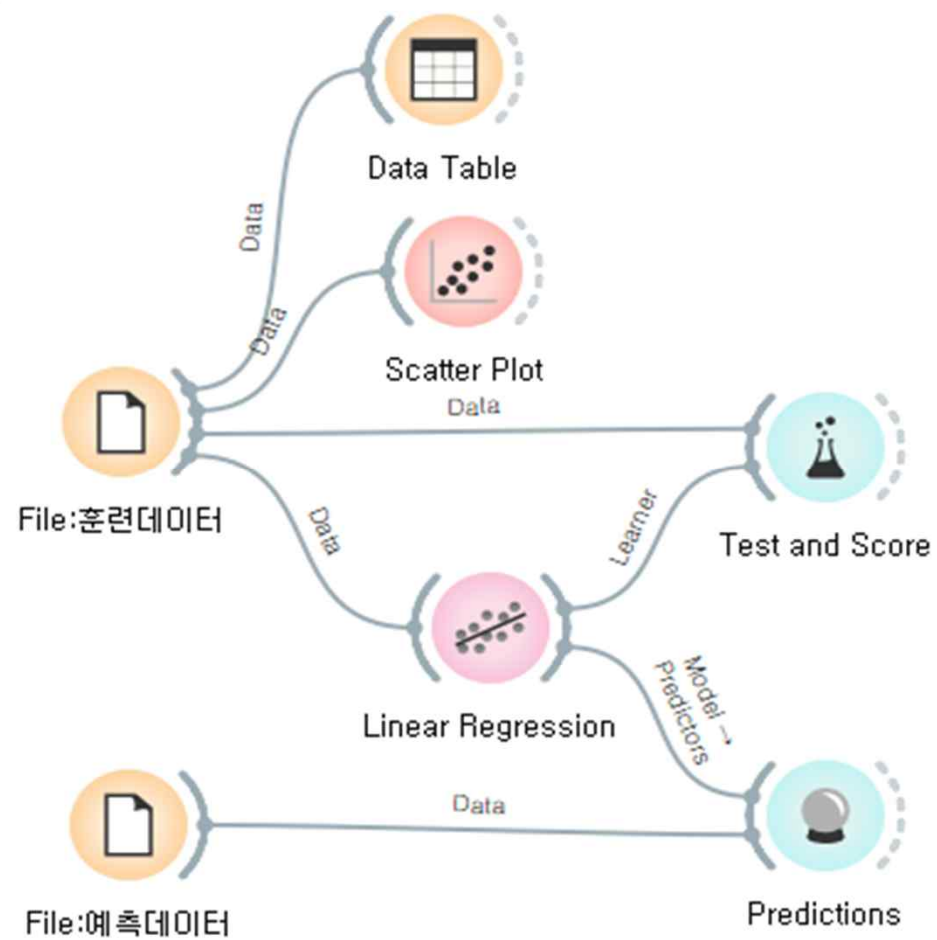


성능평가하기

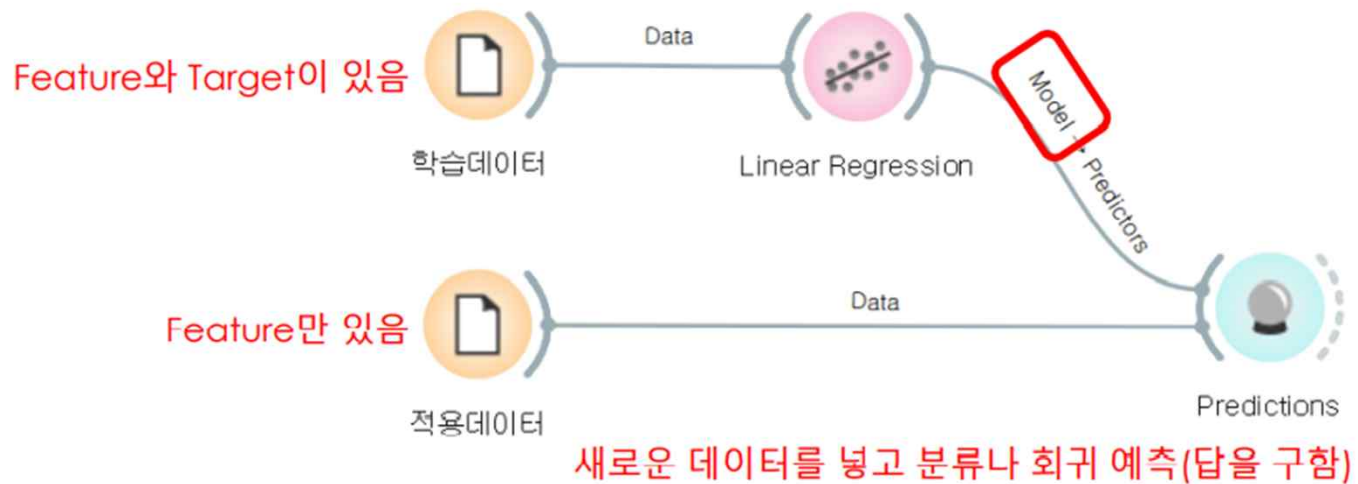
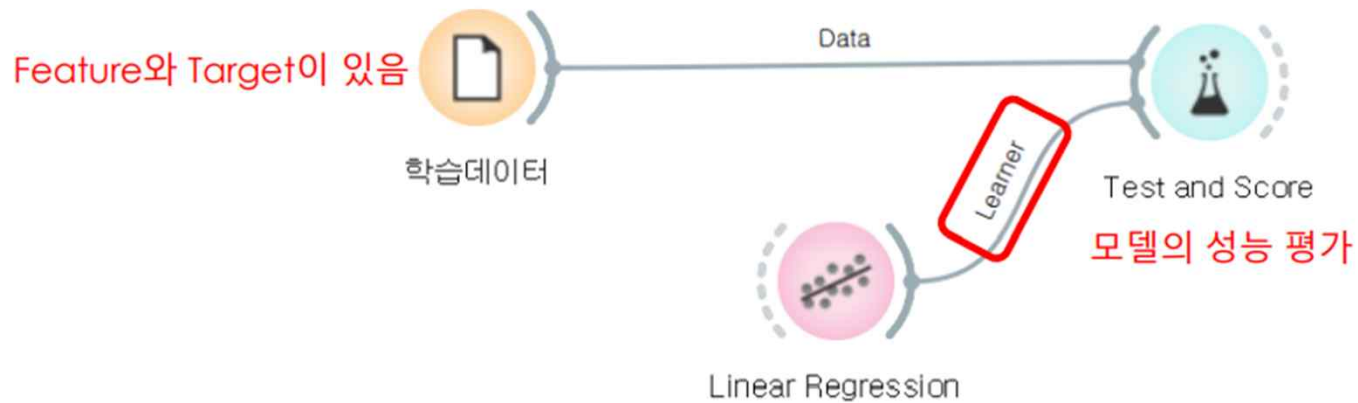


Model	MSE	RMSE	MAE	R2
Linear Regression	0.000	0.000	0.000	1.000

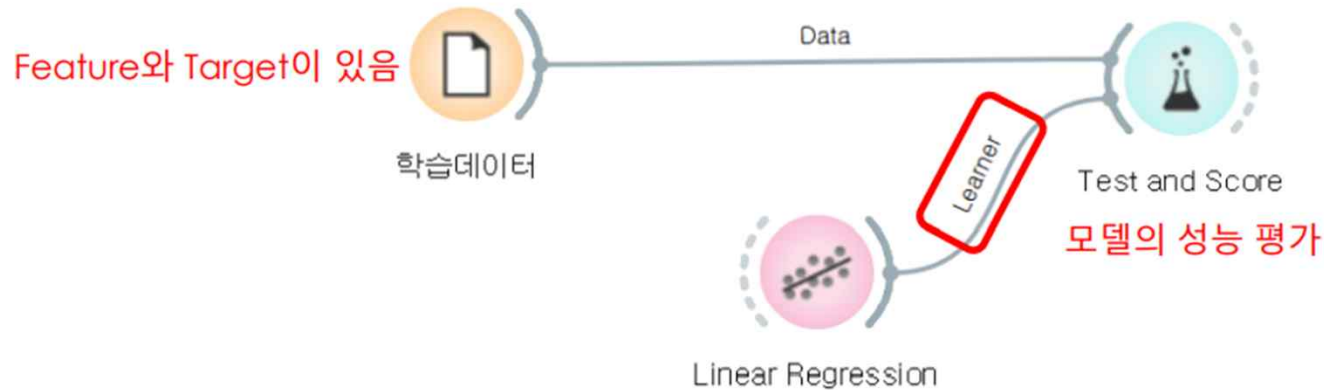
전체 워크플로



학습알고리즘과 데이터 모델링



평가(Test and Score)와 예측(Predictions)



Test and Score - Orange

☒ Cross validation
Number of folds: 5
☐ Stratified

☐ Cross validation by feature

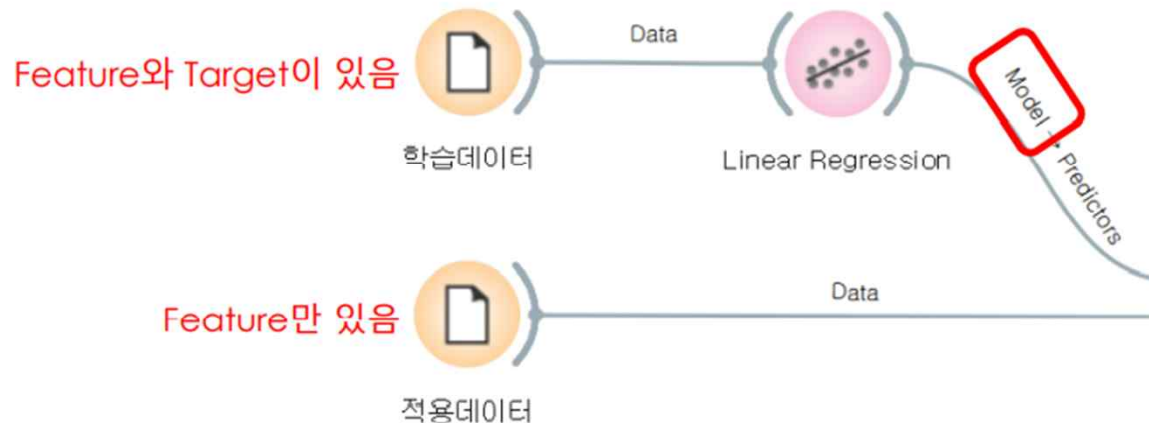
☐ Random sampling
Repeat train/test: 10
Training set size: 66 %
☒ Stratified
☐ Leave one out

Model	MSE	RMSE	MAE	R2
Linear Regression	3280.036	57.272	38.008	0.916

Compare models by: Mean s... ☐ Negligible diff.: 0.1

Linear Regression

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.



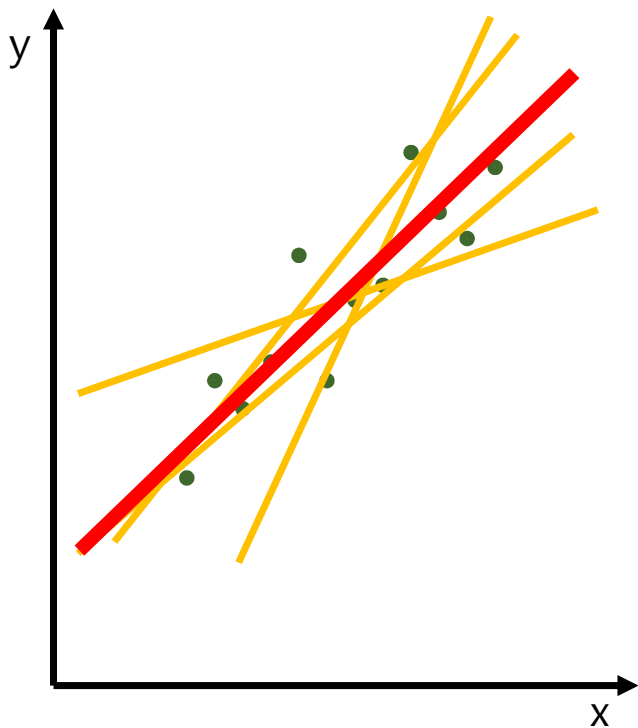
Predictions - Orange

Linear Regression	행정구역	학교수	학급수	학생수	교원수	다문화학생수
1	수원시	44	1475	36338	3545	239
2	성남시	36	1097	26272	2569	178
3	부천시	28	867	19733	2047	189
4	안양시	21	717	17267	1600	104
5	과천시	4	103	2193	236	5
6	안산시	24	886	20264	1975	481
7	용인시	31	969	27152	2235	72
8	군포시	8	296	6730	686	40
9	의왕시	5	148	3321	322	7
10	시흥시	17	509	13093	1173	148
11	평택시	21	573	14748	1288	162

새로운 데이터를 넣고 분류나 회귀 예측(답을 구함)

선형회귀 (Linear Regression)의 원리

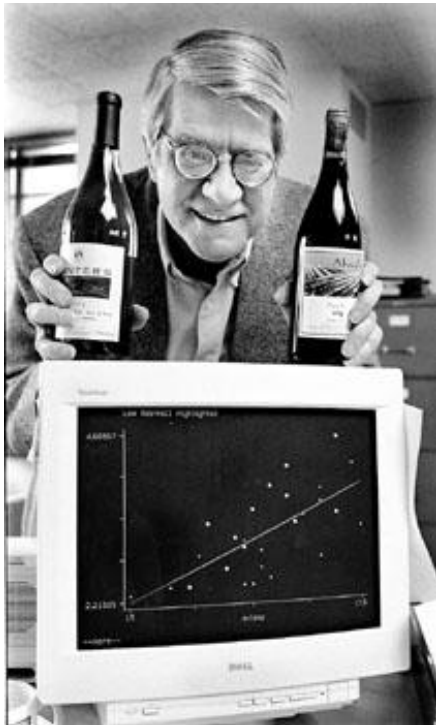
반복하며 w 와 b 값을 찾아 나간다



$$y = wx + b$$

$$y = w_1x_1 + w_2x_2 + \dots + w_nx_n + b$$

예) 선형 회귀를 활용한 와인 품질 감별

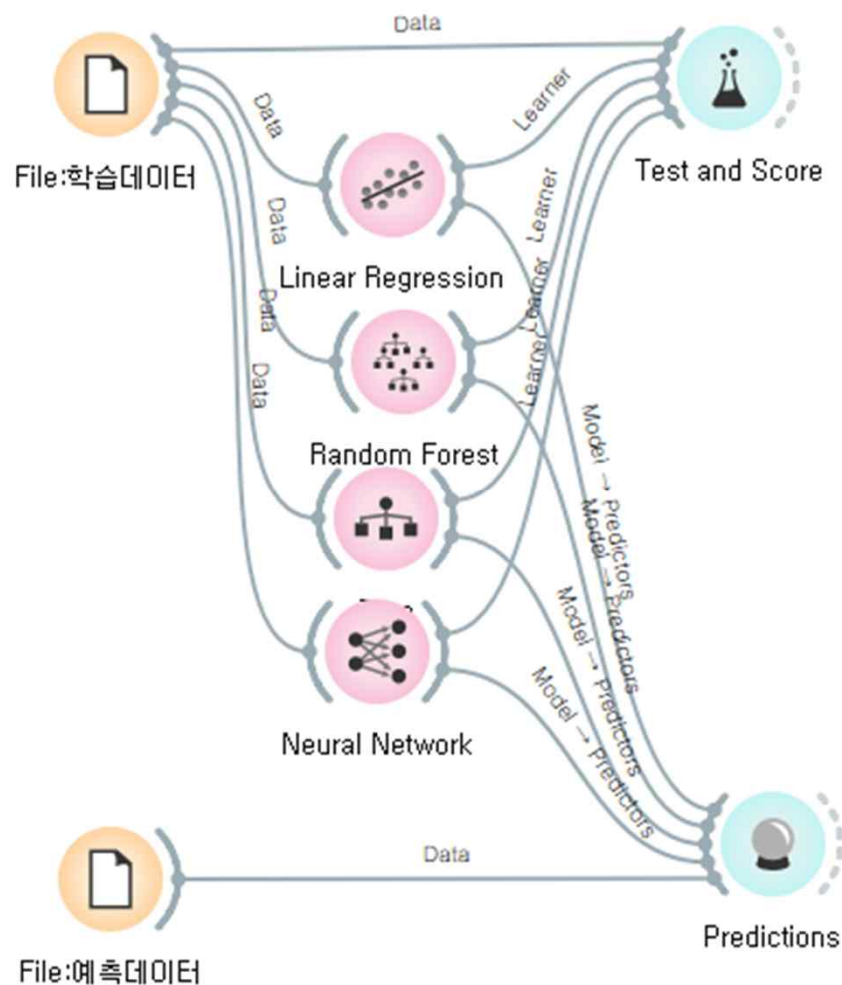


- | | |
|------------|-----------|
| 1. 고정 산도 | 7. 총 이산화황 |
| 2. 휘발성 산도 | 8. 밀도 |
| 3. 구연산 | 9. pH |
| 4. 잔류 설탕 | 10. 황산염 |
| 5. 염화물 | 11. 알코올 |
| 6. 유리 이산화황 | |

$$\text{Quality} = 4.4300987 + (-1.0127527) \times (\text{휘발산}) + (-2.0178138) \times (\text{염화물}) + (0.0050774) \times (\text{free.무수아황산}) + (-0.0034822) \times (\text{total.무수아황산}) + (-0.4826614) \times (\text{pH}) + (0.8826651) \times (\text{황산염}) + (0.2893028) \times (\text{알코올})$$

"좋은 품질"(7 점 이상) 또는 아닙니다 (7 점 미만)

다른 회귀알고리즘과의 성능비교



Model	\hat{MSE}	RMSE	MAE	R2
Linear Regression	0.000	0.000	0.000	1.000
Neural Network	4.820	2.195	1.870	0.397
Tree	12.500	3.536	3.000	-0.562
Random Forest	13.359	3.655	3.080	-0.670

평가 결과값을 확인하면 Linear Regression 모델의 성능이 가장 우수함을 알 수 있다.

Predictions - Orange

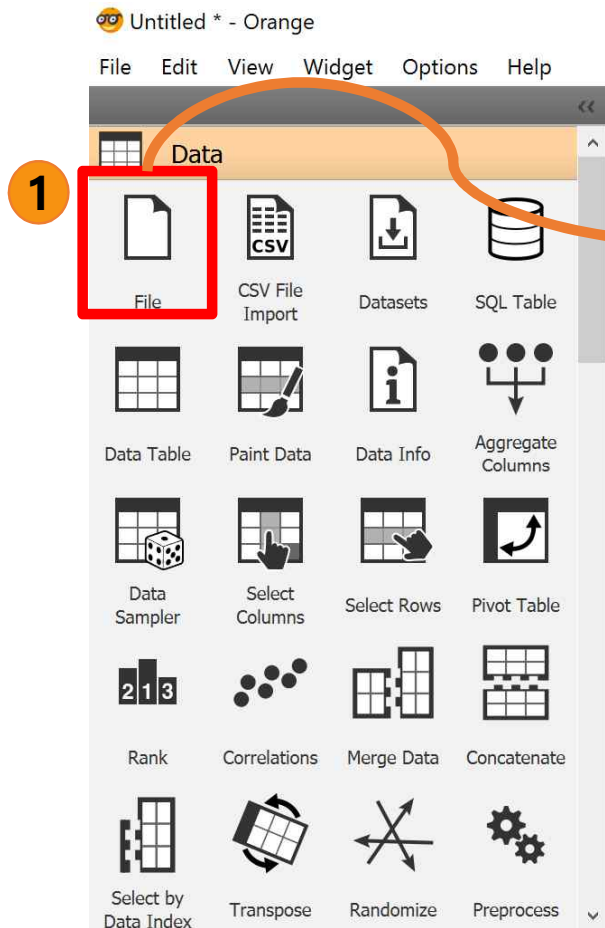
	Linear Regression	Random Forest	Tree	Neural Network	x
1	10	6	8	8	5
2	12	6	8	10	6
3	14	6	8	12	7
4	16	6	8	14	8
5	18	6	8	16	9
6	20	6	8	18	10

보스톤지역의 주택가격예측



feature	description
crim	자치시(town)별 1인당 범죄율
zn	25,000 평방 피트를 초과하는 거주지역의 비율
indus	비소매상업지역이 차지하고 있는 토지의 비율
chas	찰스강의 경계에 위치한 경우는 1, 아니면 0
nox	10ppm 당 농축 일산화질소
rm	주택 1가구당 평균 방의 수
age	1940년 이전에 건축한 소유주택 비율
dis	보스톤 직업센터까지의 접근성 지수
rad	방사형 도로까지의 접근성 지수
tax	10,000 달러당 재산세율
ptratio	자치시(town)별 학생/교사 비율
b	자치시(town)별 흑인의 비율
lstat	모집단의 하위 계층의 비율
medv	본인 소유의 주택가격 중앙값(단위 \$1000)

데이터의 입출력



File - Orange

Source

File: housing.tab

URL: <https://www.ika.si/podatki/141025/12F5B3CC/>

File Type

Automatically detect type

Info

Housing dataset
Data collected by the U.S Census Service concerning housing in Boston.

506 instance(s)
13 feature(s) (no missing values)
Regression; numerical class (no missing values)
0 meta attribute(s)

Columns (Double click to edit)

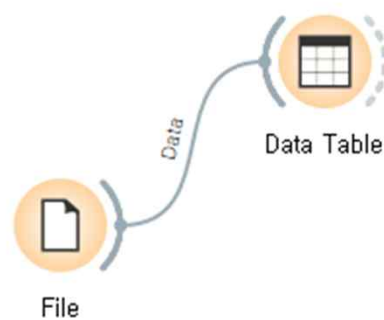
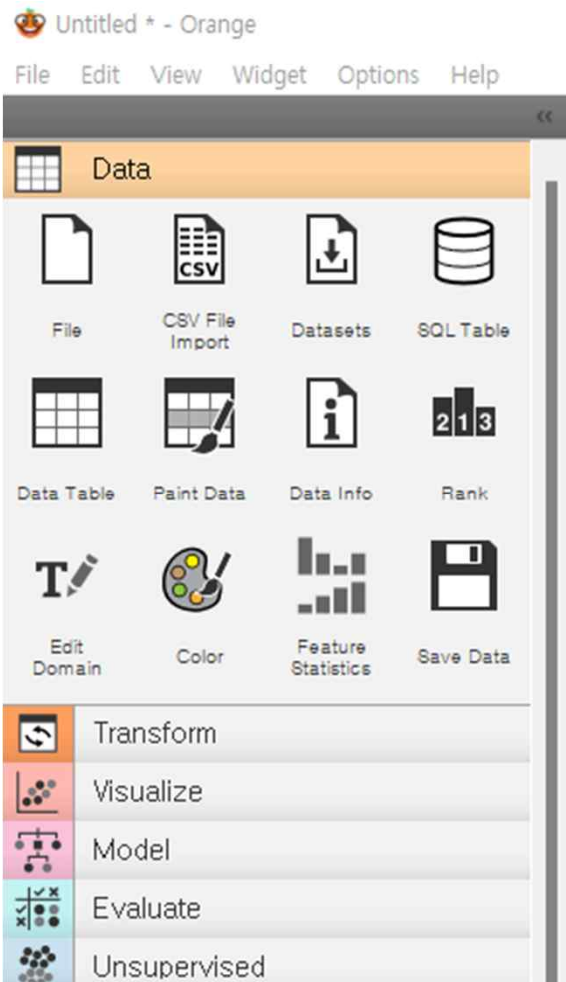
	Name	Type	Role	Values
2	ZN	N numeric	feature	
3	INDUS	N numeric	feature	
4	CHAS	N numeric	feature	
5	NOX	N numeric	feature	
6	RM	N numeric	feature	
9	RAD	N numeric	feature	
10	TAX	N numeric	feature	
11	PTRATIO	N numeric	feature	
12	B	N numeric	feature	
13	LSTAT	N numeric	feature	
14	MEDV	N numeric	target	

Reset

Apply

- feature(특성): 기계학습에 영향을 미치는 원인이 되는 속성, 독립변수
- target(목표값): 예측하고자 하는 결과가 되는 속성, 종속변수

입력된 데이터의 탐색



Data Table - Orange

Info
506 instances (no missing data)
13 features
Numeric outcome
No meta attributes

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

	MEDV	CRIM	ZN	INDUS	CI
1	24.0	0.00632	18.0	2.31	
2	21.6	0.02731	0.0	7.07	
3	34.7	0.02729	0.0	7.07	
4	33.4	0.03237	0.0	2.18	
5	36.2	0.06905	0.0	2.18	
6	28.7	0.02985	0.0	2.18	
7	22.9	0.08829	12.5	7.87	
8	27.1	0.14455	12.5	7.87	
9	16.5	0.21124	12.5	7.87	
10	18.9	0.17004	12.5	7.87	
11	15.0	0.22489	12.5	7.87	
12	18.9	0.11747	12.5	7.87	
13	21.7	0.09378	12.5	7.87	
14	20.4	0.62976	0.0	8.14	
17	23.1	1.05393	0.0	8.14	
18	17.5	0.78420	0.0	8.14	
19	20.2	0.80271	0.0	8.14	
20	18.2	0.72580	0.0	8.14	
21	13.6	1.25179	0.0	8.14	
22	19.6	0.85204	0.0	8.14	
23	15.2	1.23247	0.0	8.14	
24	14.5	0.98843	0.0	8.14	
25	15.6	0.75026	0.0	8.14	
26	13.9	0.84054	0.0	8.14	

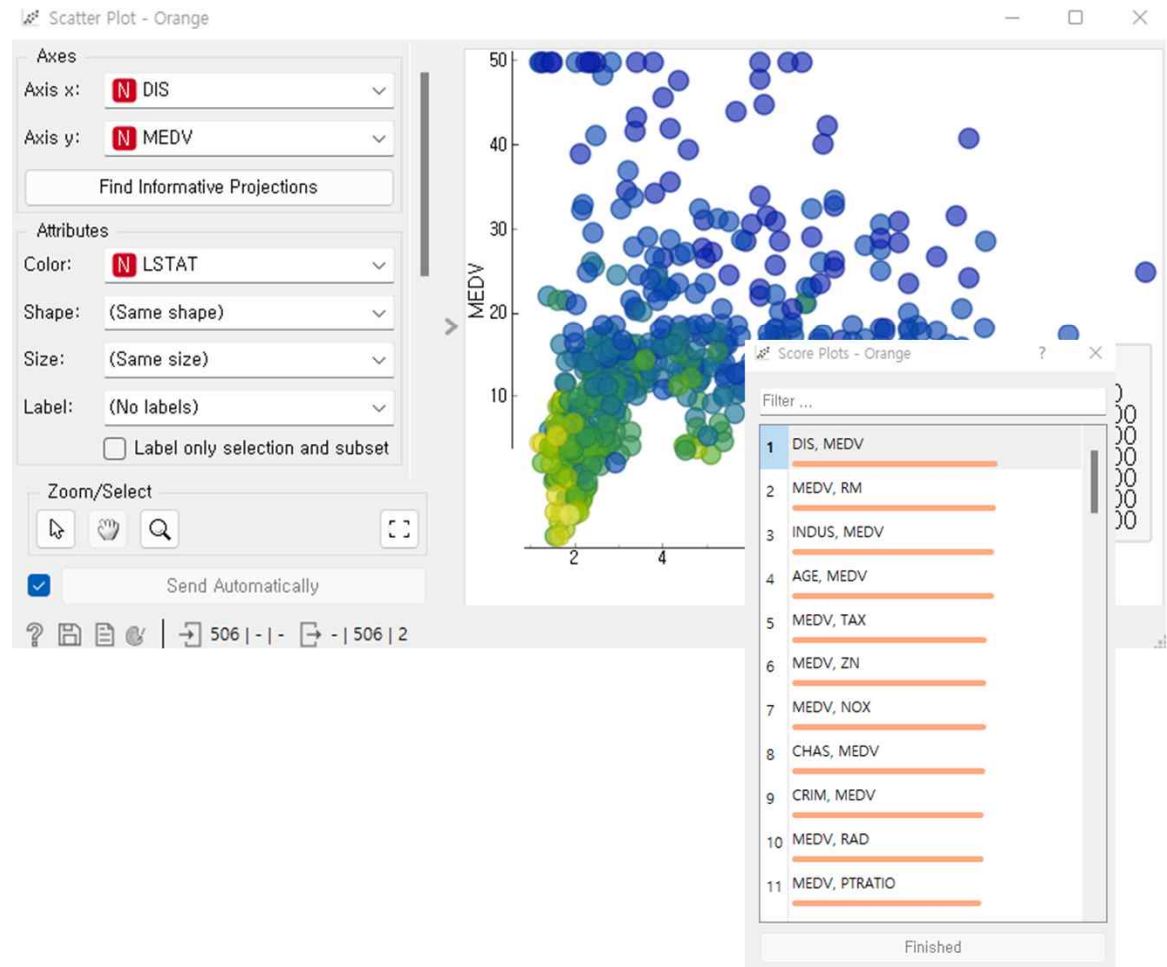
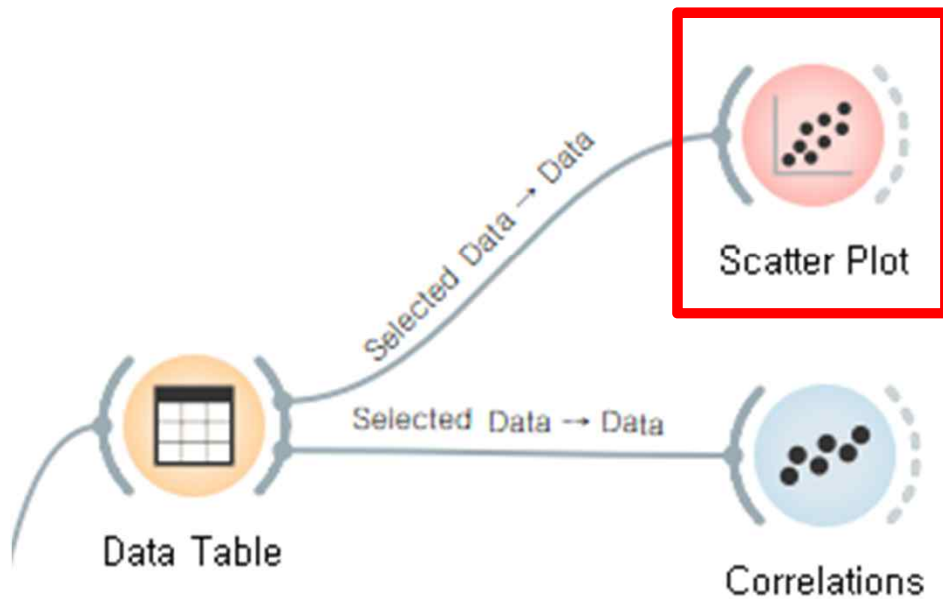
Restore Original Order

☒ Send Automatically

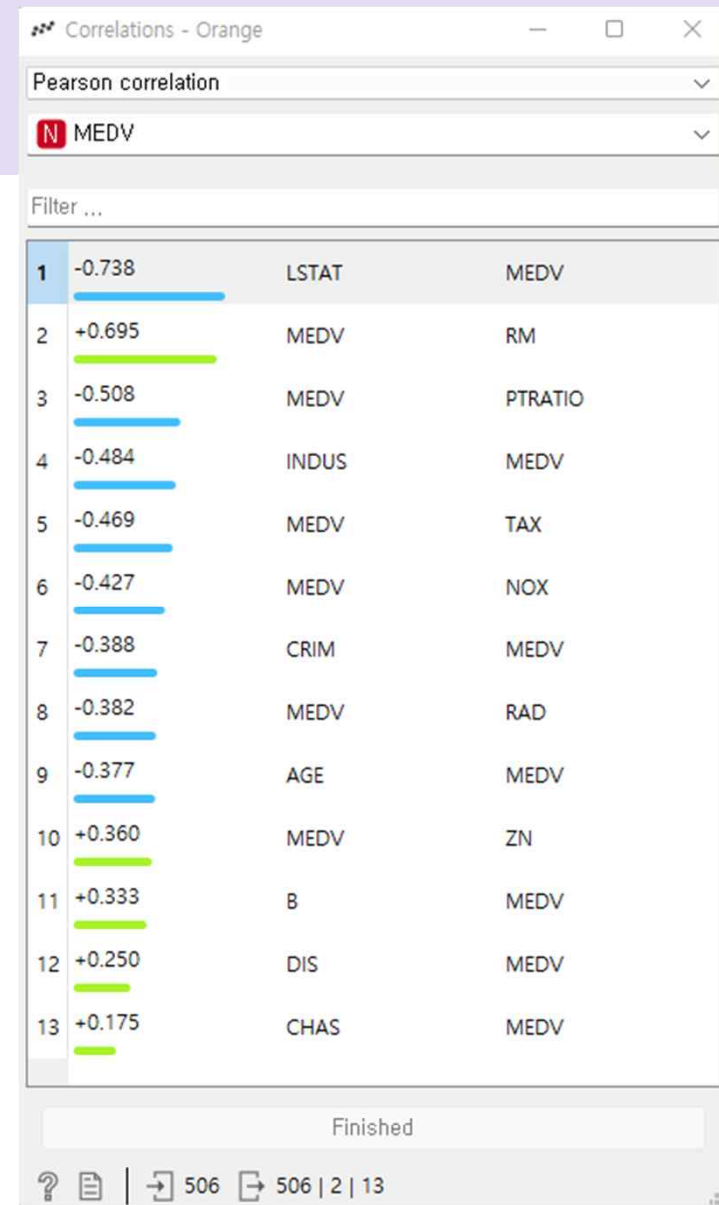
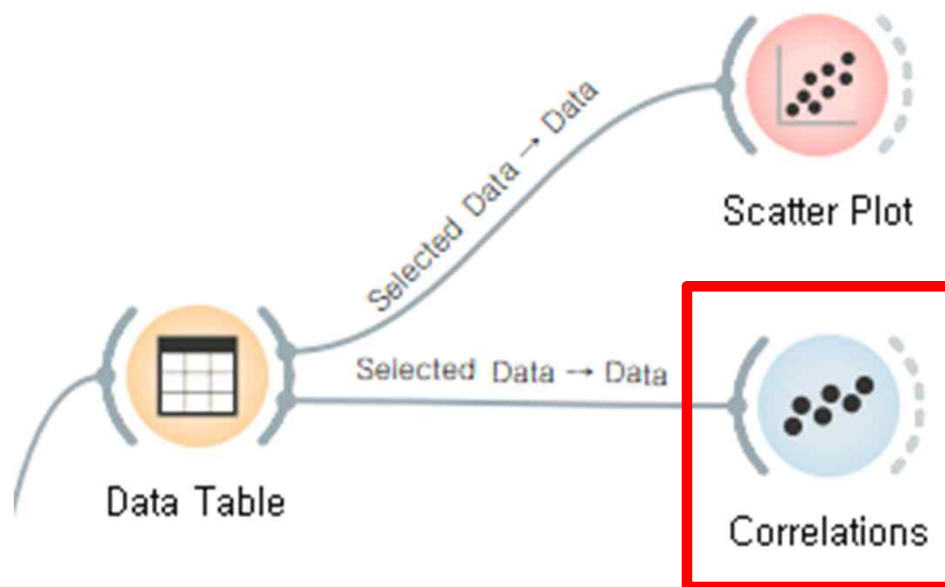
506 | 506 | 506

Target인 MEDV가 왼쪽에 표시됨

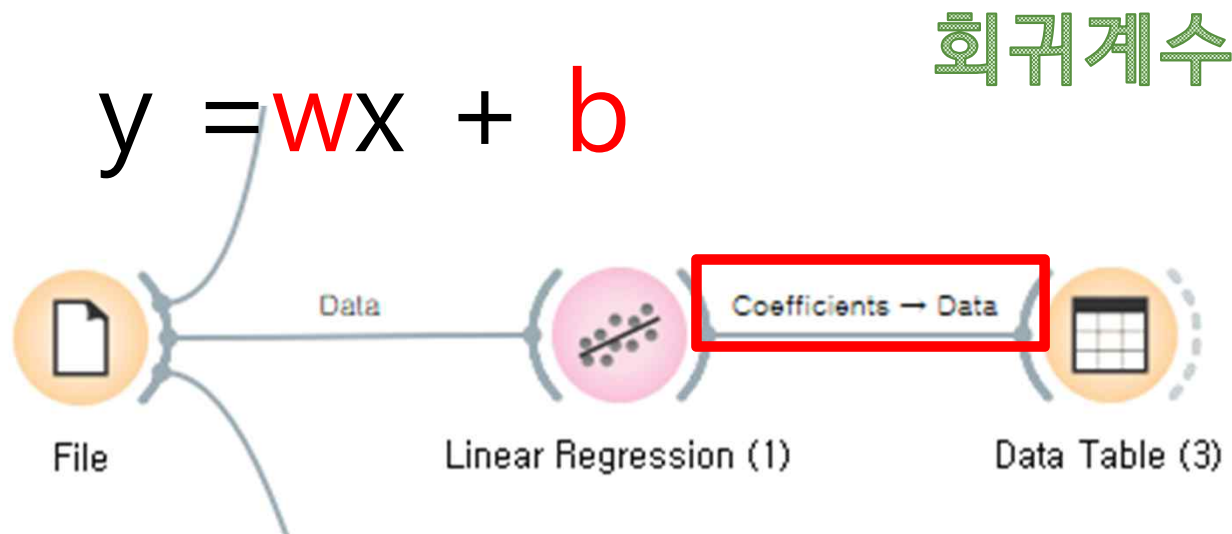
속성들 간의 상관관계 분석



속성들 간의 상관관계 분석



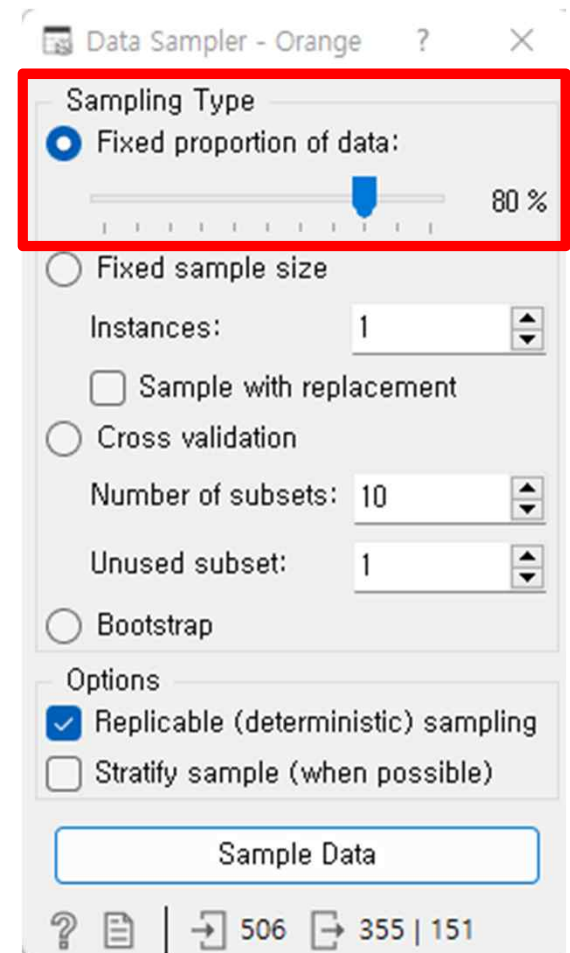
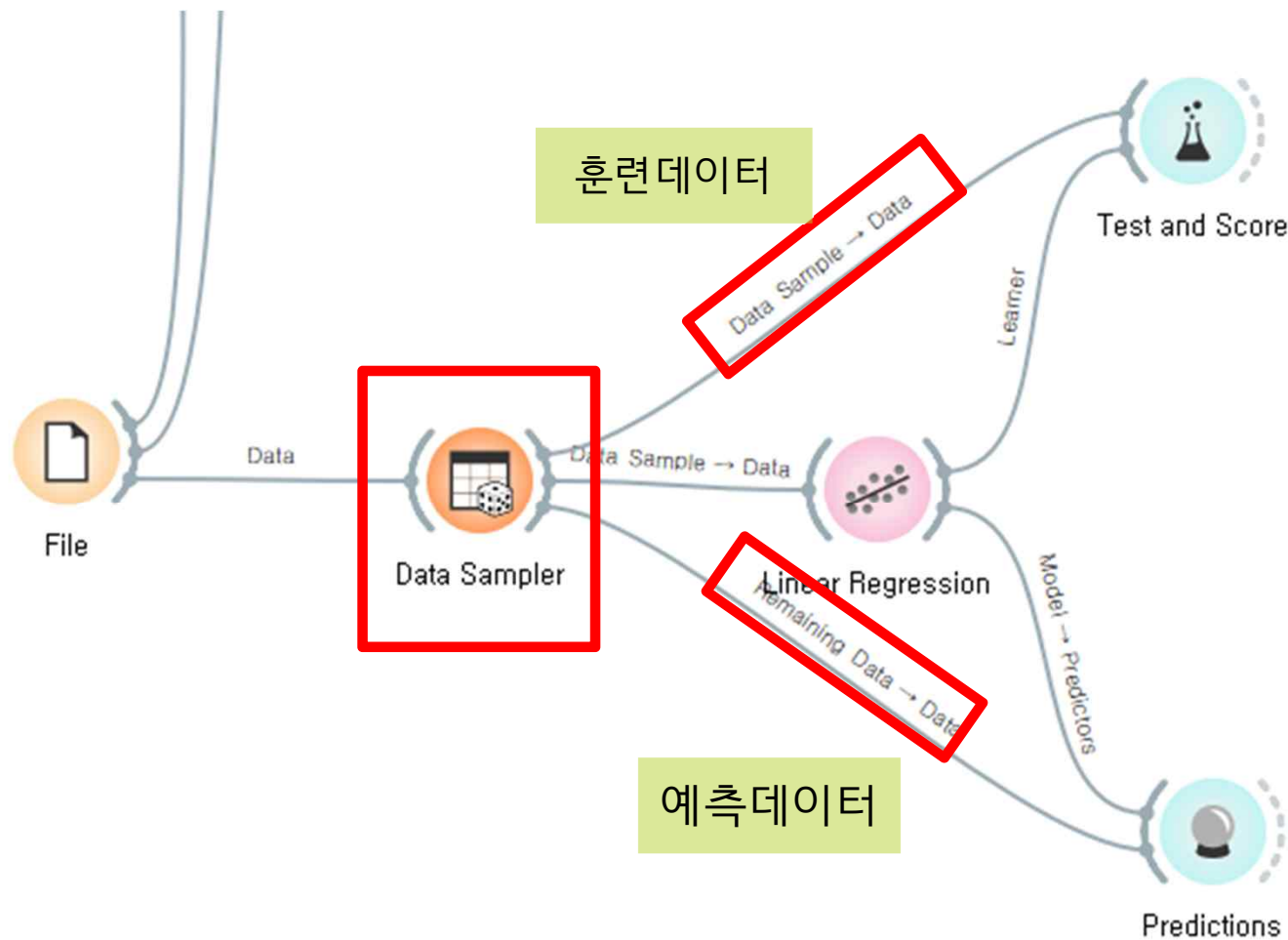
학습 알고리즘을 적용하여 모델을 생성해보자



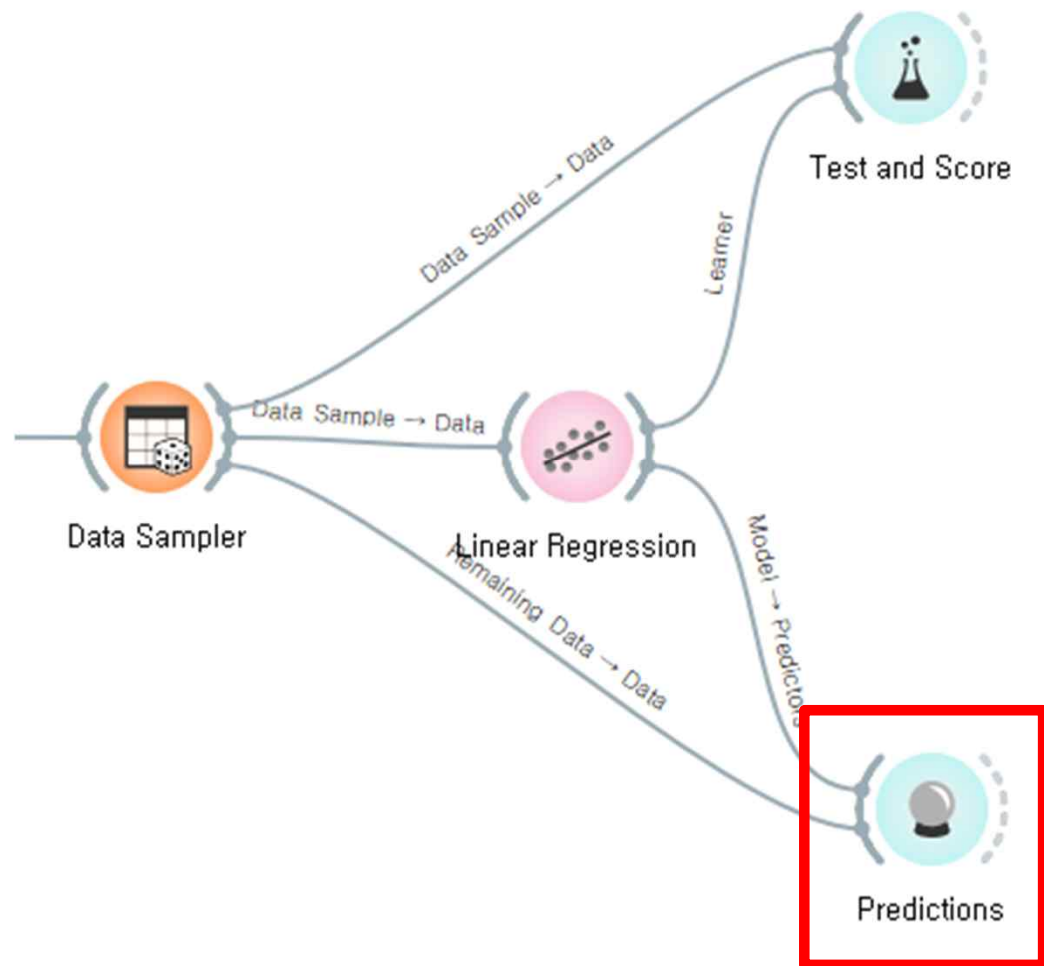
	name	coef
1	intercept	36.4371
2	CRIM	-0.107994
3	ZN	0.0464248
4	INDUS	0.0204177
5	CHAS	2.68491
6	NOX	-17.7328
7	RM	3.80991
8	AGE	0.000666336
9	DIS	-1.47504
10	RAD	0.305974
11	TAX	-0.0123375
12	PTRATIO	-0.952383
13	B	0.00931343
14	LSTAT	-0.524815

주택가격 = $-0.107994 \times \text{CRIM} + 0.0464248 \times \text{ZN} + 0.0204177 \times \text{INDUS} + 2.68491 \times \text{CHAS} + \dots + 36.4371$

예측하기 - 예측 데이터 준비



집값예측



Predictions - Orange

Restore Original

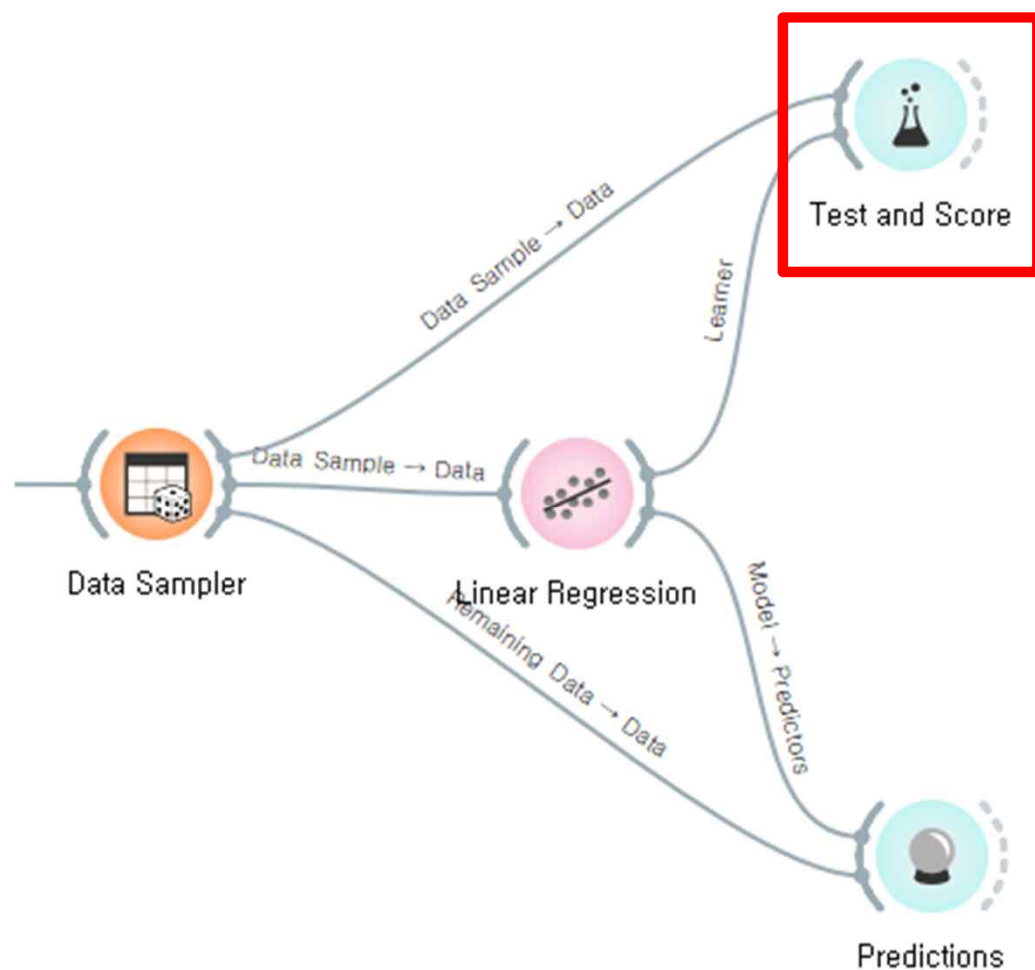
	Linear Regression	MEDV	CRIM	ZN	INDUS
1	27.7	36.2	0.06905	0.0	2.18
2	38.0	44.0	0.01538	90.0	3.75
3	22.1	17.8	8.24809	0.0	18.10
4	24.7	27.5	0.14866	0.0	8.56
5	37.6	37.6	0.38214	0.0	6.20
6	18.9	14.1	10.06230	0.0	18.10
7	24.8	28.1	0.14052	0.0	10.59
8	16.8	10.2	12.24720	0.0	18.10
9	24.3	19.1	2.31390	0.0	19.58
10	34.3	43.8	0.08187	0.0	2.89
11	32.7	27.9	0.03615	80.0	4.95
12	24.3	25.0	0.19802	0.0	10.59
13	19.5	16.0	0.17171	25.0	5.13
14	18.1	16.6	0.22927	0.0	6.91
15	8.8	13.2	1.38799	0.0	8.14
16	40.3	50.0	0.57834	20.0	3.97
17	19.1	22.2	0.24103	0.0	7.38
18	30.5	32.9	0.01778	95.0	1.47
19	19.4	15.2	5.44114	0.0	18.10
20	15.1	14.8	0.95577	0.0	8.14

✓ Show performance scores

Model	MSE	RMSE	MAE	R2
Linear Regression	22.226	4.714	3.359	0.725

? | 151 | 151 | 1x151

평가하기



Test and Score - Orange

☒ Cross validation
Number of folds: 5
☐ Stratified
☐ Cross validation by feature
☐ Random sampling
Repeat train/test: 10
Training set size: 70 %
☒ Stratified
☐ Leave one out
☐ Test on train data
☐ Test on test data

Model	MSE	RMSE	MAE	R2
Linear Regression	24.673	4.967	3.486	0.713

Compare models by: | ☐ Negligible diff.: 0,1

	Linear Regression
Linear Regression	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

355 | - | 355 | 1x355



지금까지 간단한 선형모델을 활용하여 선형회귀 모델의 개념에 대해 알아보았고, 보스턴 집값을 예측하는 모델을 생성해보았습니다. 다음시간에는 회귀모델의 성능평가지표를 활용하여 선형회귀모델 이외의 다양한 회귀모델을 생성 비교해보고 실생활에서 적용할 수 있는 회귀분석을 실습해보도록 하겠습니다. 감사합니다.





orange 활용 데이터 분석 및 머신 러닝



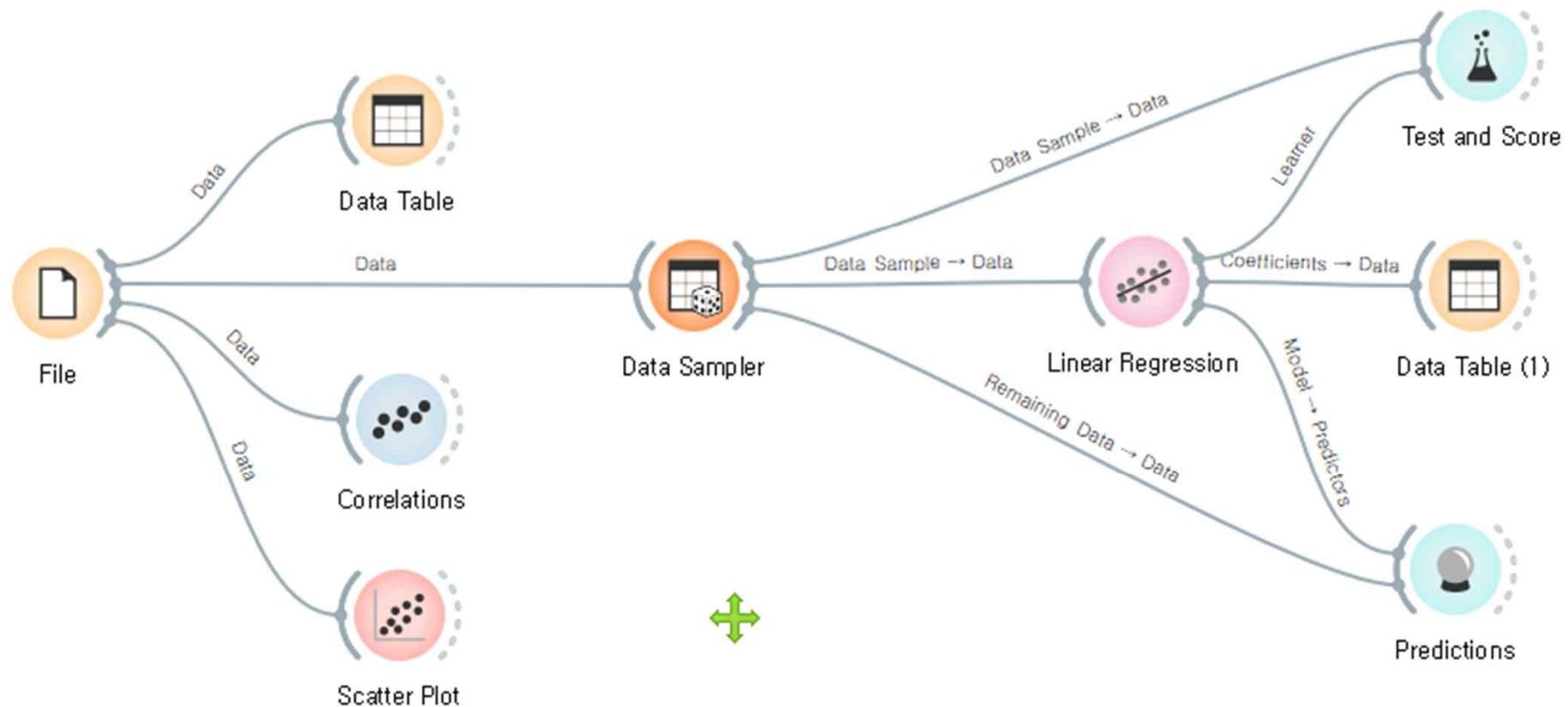
3차시

지도학습 - 회귀2

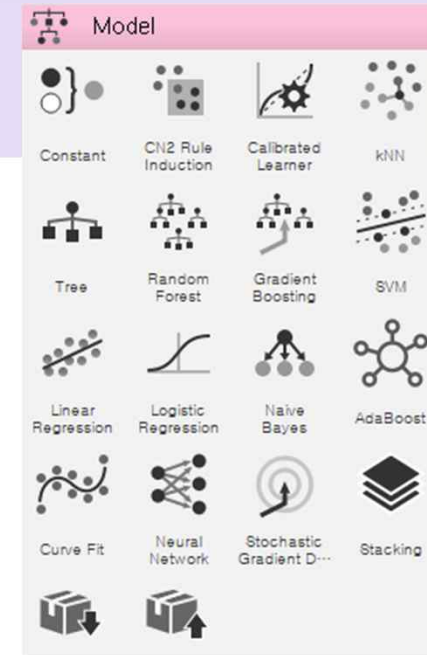
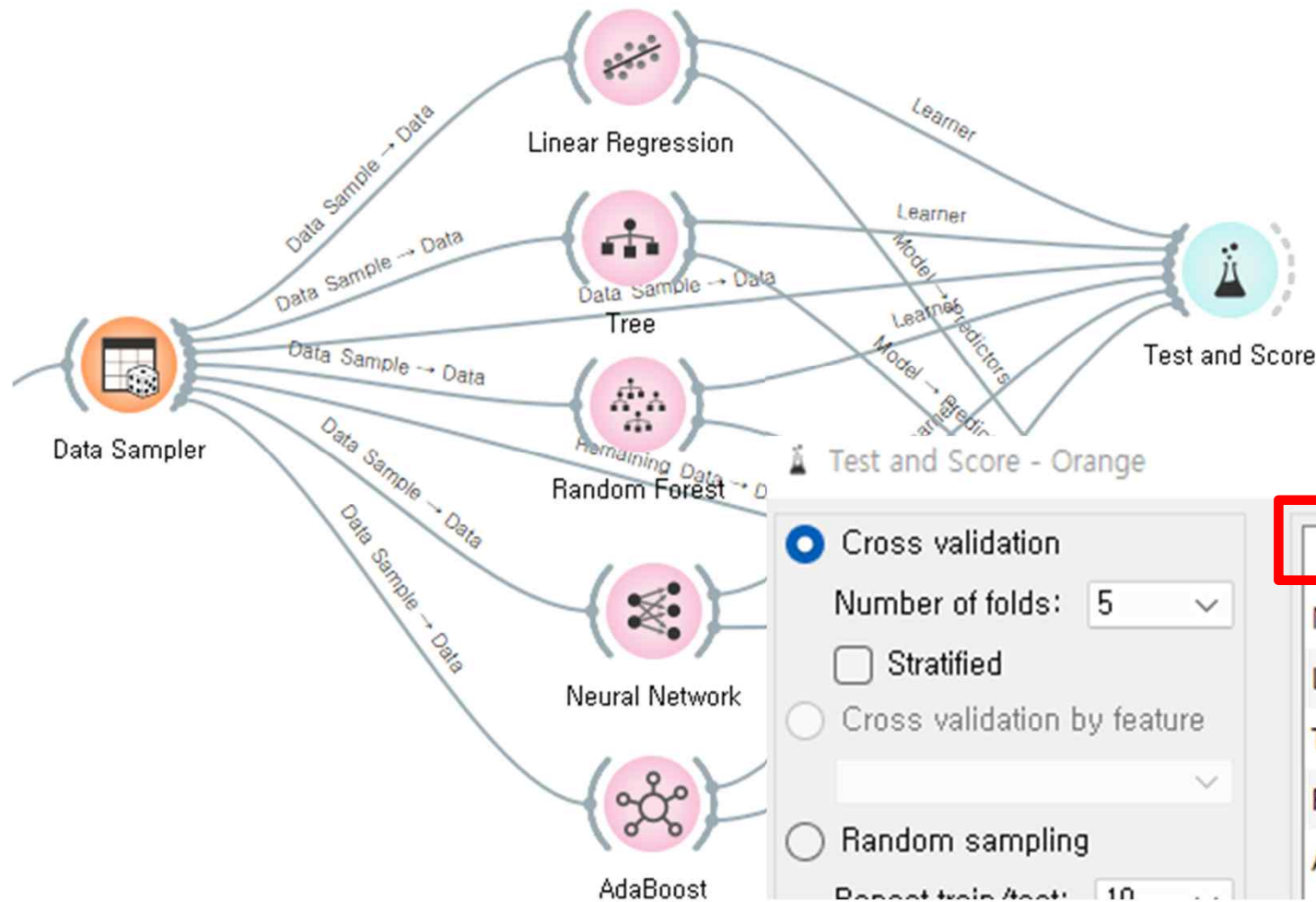


1. 다양한 회귀모델 활용하기
2. 회귀의 성능평가지표
3. 실생활에서 회귀분석 활용

보스턴 집값데이터셋을 활용한 회귀 워크플로



여러 모델을 평가하고 비교하기



Model	MSE	RMSE	MAE	\hat{R}^2
Neural Network	31.485	5.611	4.058	0.633
Linear Regression	24.673	4.967	3.486	0.713
Tree	24.212	4.921	3.297	0.718
Random Forest	14.300	3.782	2.531	0.833
AdaBoost	14.174	3.765	2.407	0.835

평가지표 (회귀)

Predictions - Orange

Restore Original Order

	Linear Regression	Tree	Random Forest	Neural Network	AdaBoost	MEDV	CR
1	27.7	36.7	34.5	35.6	36.1	36.2	0.06905
2	38.0	42.9	45.1	37.1	43.5	44.0	0.01538
3	22.1	13.0	17.1	16.6	16.1	17.8	8.24809
4	24.7	25.1	24.3	22.0	24.1	27.5	0.14866
5	37.6	46.2	47.4	37.3	48.3	37.6	0.38214
6	18.9	19.5	20.5	11.3	19.1	14.1	10.06230
7	24.8	25.6	23.7	22.6	23.7	28.1	0.14052

☒ Show performance scores

Model	MSE	RMSE	MAE	R2
AdaBoost	9.105	3.017	2.171	0.887
Random Forest	9.325	3.054	2.232	0.885
Tree	15.430	3.928	2.995	0.809
Linear Regression	22.226	4.714	3.359	0.725
Neural Network	30.965	5.565	4.066	0.616

? | 151 | 151 | 5x151

MSE(Mean Squared Error)

$$\frac{1}{n} \sum_{i=1}^n (H(x_i) - y_i)^2$$

RMSE(Root Mean Squared Error)

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (H(x_i) - y_i)^2}$$

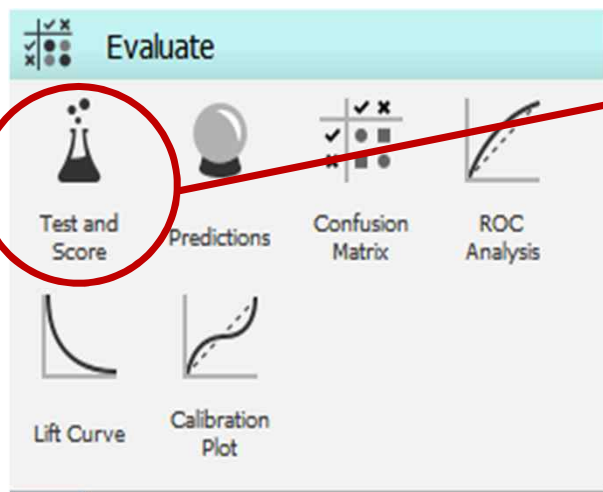
MAE(Mean Absolute Error)

$$\frac{1}{n} \sum_{i=1}^n |H(x_i) - y_i|$$

R2(R Squared, R^2 , 결정계수)

$$\frac{\sum_{i=1}^n (H(x_i) - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

회귀에서 성능평가



Test and Score - Orange

Sampling

- ☒ Cross validation
 - Number of folds: 5
 - ☐ Stratified
- ☐ Cross validation by feature
- ☐ Random sampling
 - Repeat train/test: 10
 - Training set size: 70 %
 - ☐ Stratified
- ☐ Leave one out
- ☐ Test on train data
- ☐ Test on test data

Model Comparison

Model

Model	MSE	RMSE	MAE	R ²
Tree	6.333	2.516	1.460	0.648
Linear Regression	7.781	2.789	1.927	0.568
kNN	10.979	3.313	2.592	0.390

1에 가까울수록 성능이 우수함

0에 가까울수록 성능이 우수함

Model Comparison by MSE

	Tree	Linear Regre...	kNN
Tree		0.287	0.100
Linear Regression	0.713		0.140
kNN	0.900	0.860	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

* 성능 평가

- MSE : 예측값과 실제값 차이 제곱하여 평균
- RMSE : MSE 를 ROOT 처리한 값(실제값)
- MAE : 예측값 - 실제값 차이의 절대값의 평균
- R2 : (예측값-실제값평균)의 제곱합 / (실제값-실제값평균)의 제곱합

[활용] 우리동네 미세먼지 농도 예측하기

다른 동네 미세먼지 농도 데이터를 보고 우리동네의 농도를 추정해보자.

➔ 구월동을 target으로 해서 결측치를 예측하자.

	A	B	C	D	E	F	G	H	I	J
1	날짜	송림	연희	원당	송의	부평	구월	신흥	고잔	석남
2	2022-07-01	17	18	24	14	17	12	16	17	14
3	2022-07-02	31	28	36	29	29	26	28	31	23
4	2022-07-03	36	30	39	38	31		37	36	25
5	2022-07-04	32	35	42	31	33	28	34	32	26
6	2022-07-05	47	45	63	42	49		50	47	42
7	2022-07-06	41	39	47	35	39			42	33
8	2022-07-07	25	24	41	20	24			28	18
9	2022-07-08	27	27	48	24	30	19		24	21
10	2022-07-09	27	24	41	25	26	23	27	23	19
11	2022-07-10	39	43	70	36	46	34	36	37	39
12	2022-07-11	25	24	27	23	25	23	24	29	22
13	2022-07-12	22	19	19	17	19	16	22	25	15
14	2022-07-13	24	22			20	15	22		17

[활용] 우리동네 미세먼지 농도 예측하기

File - Orange

Source

File: 02_회귀_인천시... Reload

URL:

File Type

Automatically detect type

Info

31 instance(s)
25 feature(s) (1.3% missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role
1	날짜	datetime	meta
2	송림	numeric	feature
3	연희	numeric	feature
4	원당	numeric	feature
5	송의	numeric	feature
6	부평	numeric	feature
7	구월	numeric	target

Reset Apply

Browse documentation datasets

Select Rows - Orange

Conditions

구월 is greater than 0.000000

Add Condition Add All Variables Remove All

☐ Remove unused features ☒ Send Automatically

☐ Remove unused classes

31 21 10 31

Select Columns - Orange

Ignored

Filter

구월

Features

Filter

송림 연희 원당 수인

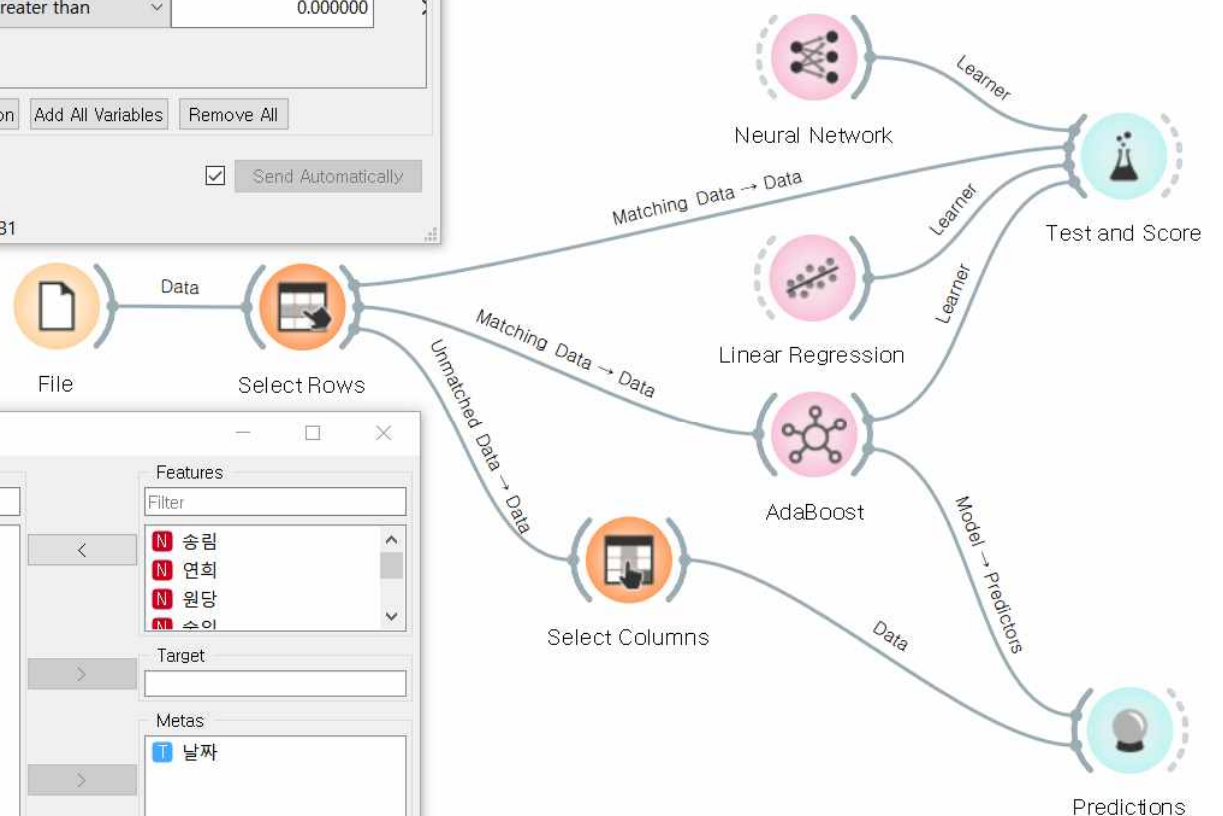
Target

Metas

날짜

Reset ☐ Ignore new variables by default ☒ Send Automatically

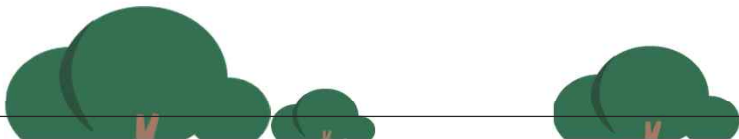
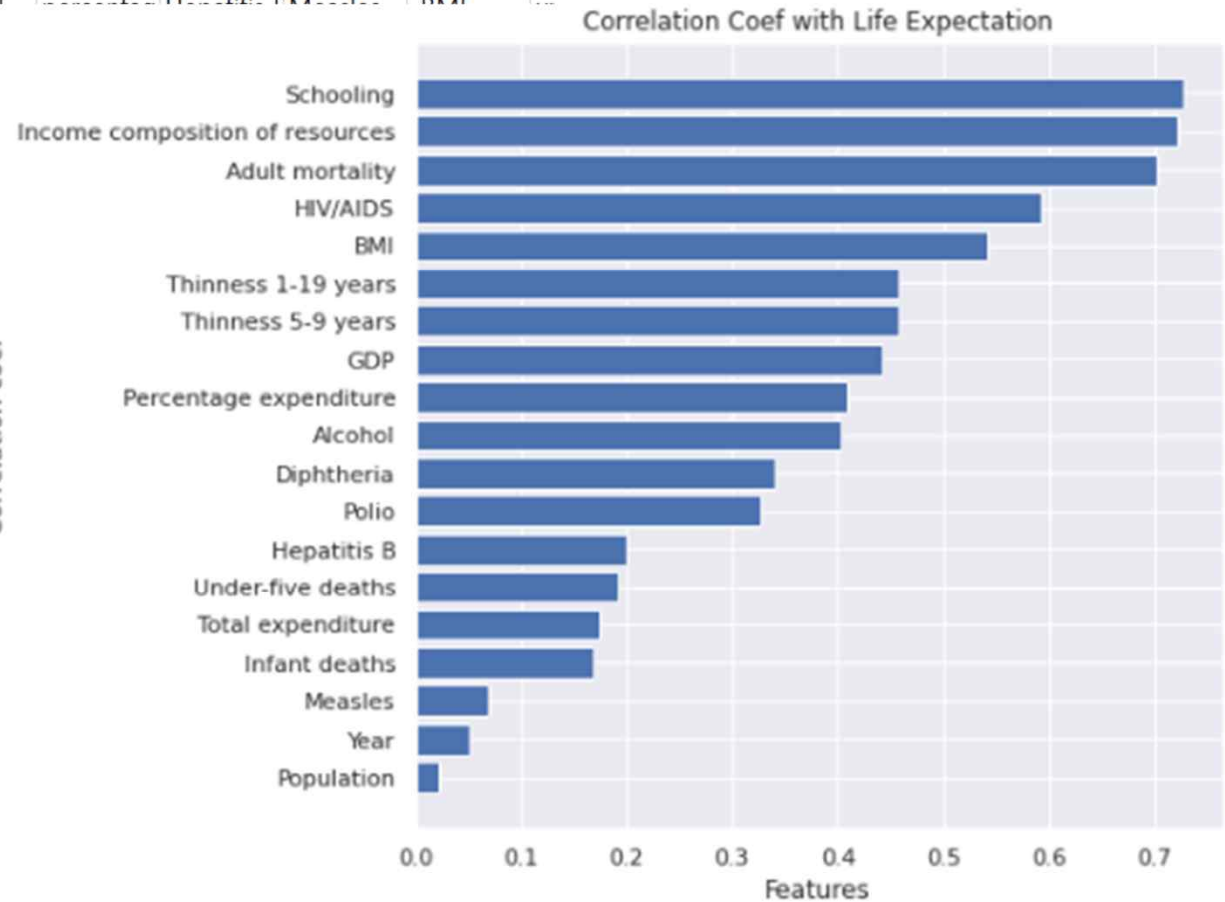
10 10 23



[활용] 국가별 기대수명 예측

	A	B	C	D	E	F	G	H	I	J	K	
1	Country	Year	Status	Life expect	Adult Mor	infant deat	Alcohol					
2	Afghanista	2015	Developing	65	263	62						
3	Afghanista	2014	Developing	59.9	271	64						
4	Afghanista	2013	Developing	59.9	268	66						
5	Afghanista	2012	Developing	59.5	272	69						
6	Afghanista	2011	Developing	59.2	275	71						
7	Afghanista	2010	Developing	58.8	279	74						
8	Afghanista	2009	Developing	58.6	281	77						
9	Afghanista	2008	Developing	58.1	287	80						
10	Afghanista	2007	Developing	57.5	295	82						
11	Afghanista	2006	Developing	57.3	295	84						

Correlation coef



기대수명 데이터셋의 속성들

- Country: 국가명
- Year: 2000년부터 2015년까지의 연도
- Status: Developed(선진국) or Developing(개발도상국) status
- Life expectancy: 기대수명(나이)
- Adult Mortality: 15세~60세사이의 성인 1000명당 사망자수
- infant deaths: 유아 1000명당 사망자수
- Alcohol: 1인당 알콜 소비량
- percentage expenditure: GDP 대비 보건 예산 지출비율(%)
- Hepatitis B: 1세 아동의 B형 간염 예방 접종률(%)
- Measles: 인구 1000명당 홍역 예방 접종률(%)
- BMI: 전인구 평균 체질량 지수
- Under-five deaths: 5세이하 아동 1000명당 사망자수
- Polio: 1세 아동의 소아마비 면역률(%)
- Total expenditure: 정부 총예산 대비 보건 분야 예산(%)
- Diphtheria: 1세 아동의 디프테리아 예방 접종률(%) HIV/AIDS: HIV/AIDS 감염상태로 태어남 0-4세 인구 1000명당 사망자수
- GDP: 1인당 GDP
- Population: 국가 총인구
- thinness 1-19 years: 1-19 세 청소년 중 저체중 비율
- thinness 5-9 years: 5-9세 사이의 아동의 저체중 비율
- Income composition of resources: 소득 구성에 따른 인간개발지수
- Schooling: 학교 재학 연수

Feature의 개수와 성능과의 관계?

- 상관도가 낮은 속성을 제거 한 후 예측했을 경우 평가지표는 어떻게 변할까?
- 특정한 두가지 속성이 연관관계가 있거나 상관도가 매우 높은 경우 둘 중 하나의 속성을 제거하는 것이 모델의 복잡도를 줄이고 성능을 높이는 데에 더 도움이 되지 않을까?

공공 데이터 제공 사이트

공공 데이터 : 나라에서 만들거나 취득하여 관리하는 데이터 (중앙정부, 지방자치단체, 공기업, 공공기관)

사이트명	URL
공공데이터 포털(행정안전부)	http://www.data.go.kr/
국가통계포털(통계청)	http://cosis.kr
고속도로 데이터포털(한국도로공사)	http://data.ex.co.kr
서울 열린 데이터 광장(서울시)	http://data.seoul.go.kr
경기데이터드림(경기도)	http://data.gg.go.kr
부동산 실거래가(국토교통부)	http://rt.molit.go.kr/
마이크로데이터 통합 서비스	https://mdis.kostat.go.kr/index.do
보건복지 데이터포털	https://kdx.kr/main
국민건강보험(NHISS)	https://nhiss.nhis.or.kr/bd/ay/bdaya001iv.do
구글 트렌드	https://trends.google.co.kr/trends/?geo=KR
네이버 데이터랩	https://datalab.naver.com/
한국데이터거래소(민간)	https://data.kihasa.re.kr/
캐글(해외)	https://www.kaggle.com/
Out World in Data(해외)	https://ourworldindata.org

다음 시간에는
지도학습 - '정형데이터를 활
용한 분류1' 를 알아보시다.

