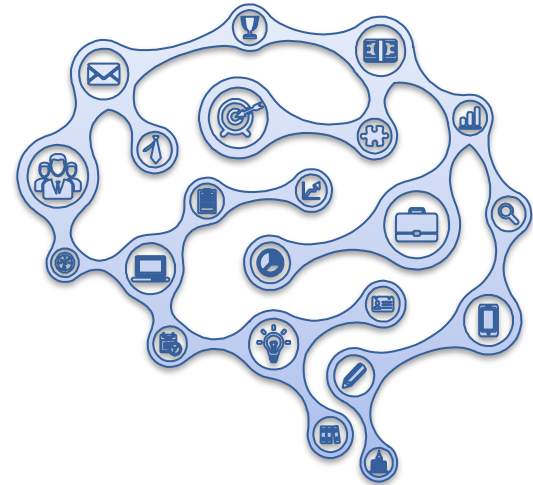


의료 Artificial Intelligence

데이터 마이닝

2022.05.12



오늘 배울 내용 ...

1. 데이터 마이닝
2. 인공지능 실습
3. mblock 실습
4. 팀 프로젝트

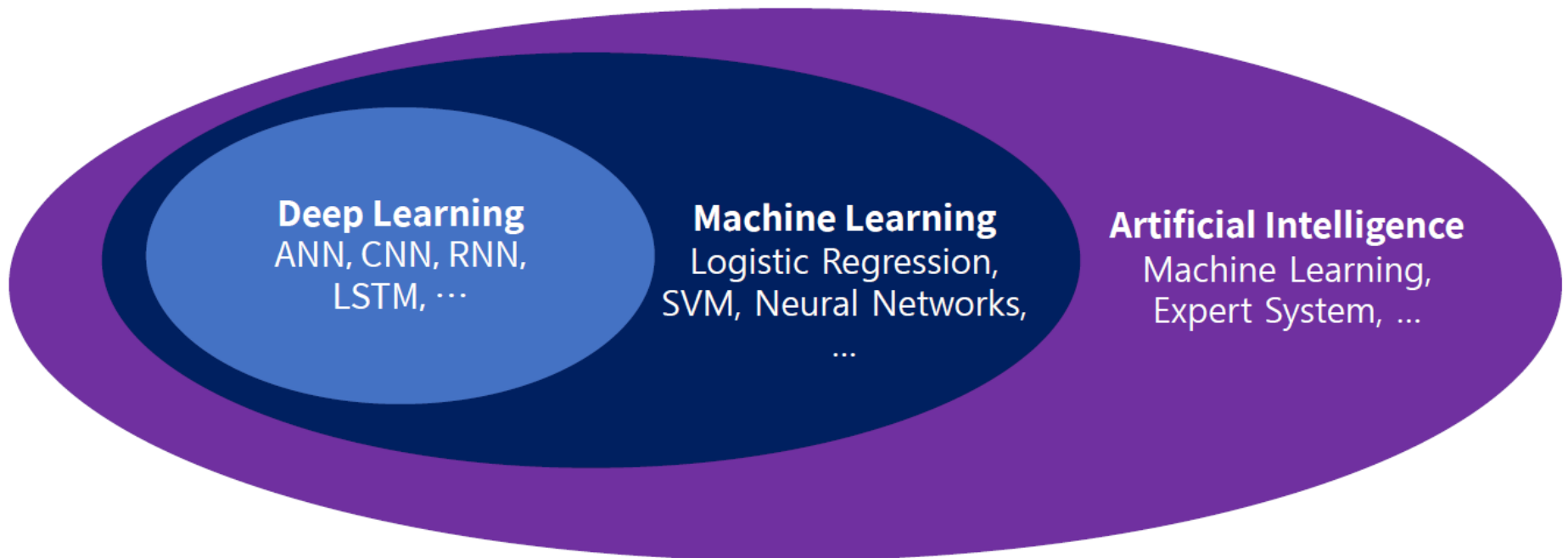
어렵지 않다
쉬운 것도 아니다



인공지능 이론

인공지능, 머신러닝, 딥러닝

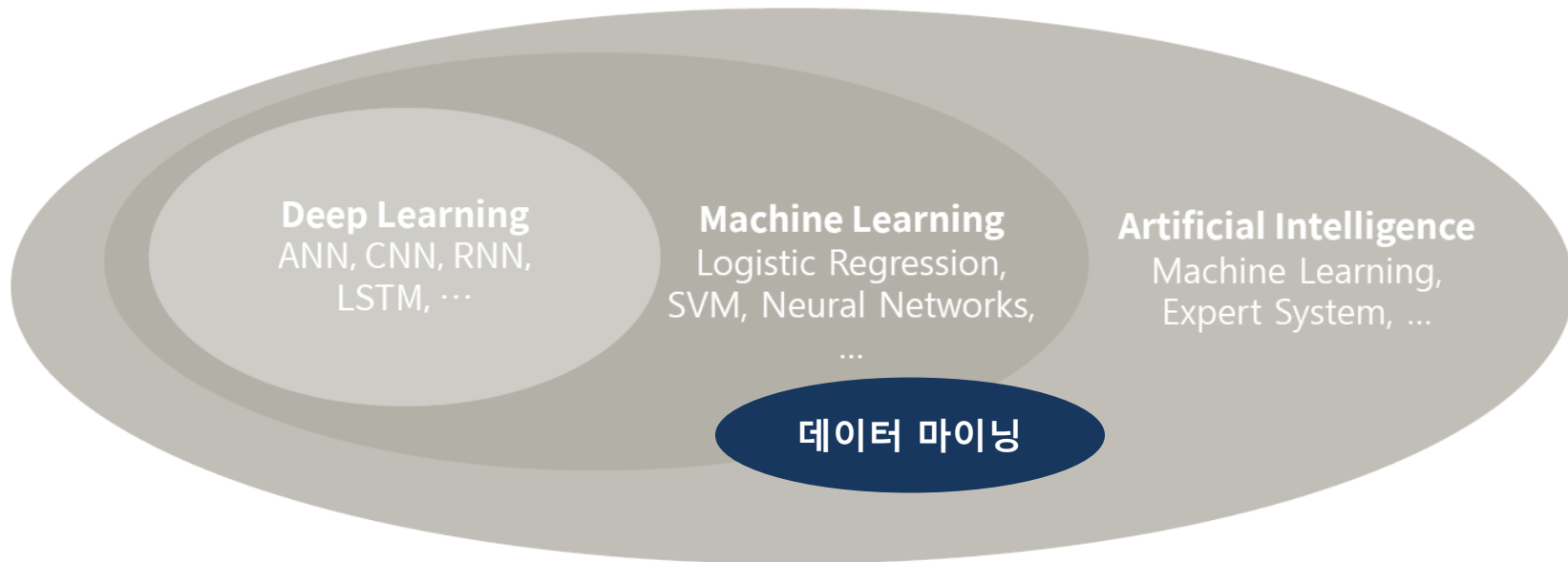
- 인공지능(Artificial Intelligence) : 컴퓨터가 인간과 같이 생각할 수 있도록 만드는 기법 연구
- 머신러닝(Machine Learning) : 데이터에 기반한 학습을 통해 인공지능을 구현하는 기법
- 딥러닝(Deep Learning) : 머신러닝 기법 중 하나인 인공 신경망(Artificial Neural Networks) 기법의 은닉층(Hidden Layer)을 깊게 쌓은 구조를 이용해 학습하는 기법



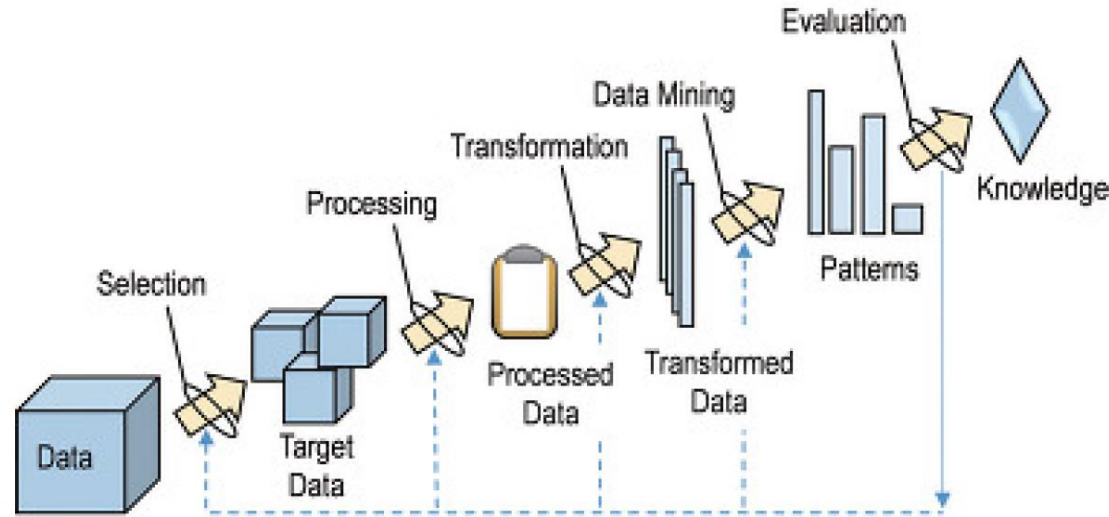
데이터 마이닝

데이터 마이닝(data mining)은 대규모로 저장된 데이터 안에서 체계적이고
자동적으로 통계적 규칙이나 패턴을 분석하여 가치있는 정보를 추출하는 과정

- 데이터를 탐색하고 모델을 구축하는데 주안점을 둠
- 자료의 수집, 분류, 가공 등의 전처리 작업이 필요함



데이터 마이닝 과정



1. 데이터 정제(Data Cleaning) : 불필요하거나 일치하지 않는 데이터를 제거
2. 데이터 통합(Data Integration) : 다수의 데이터 소스들을 결합
3. 데이터 선택(Data Selection) : 필요한 데이터들을 데이터베이스로부터 검색
4. 데이터 변환(Data Transformation) : 요약이나 집계 등을 수행해 데이터마이닝을 위한 적합한 형태로 데이터 가공
5. 데이터 마이닝(Data Mining) : 지능적 방법들을 적용하여 데이터 패턴이나 지식을 추출
6. 데이터 검증(Data Evaluation) : 데이터 마이닝으로 찾아낸 패턴이나 지식을 검증
7. 데이터 시각화(Data Presentation) : 발견한 패턴이나 지식을 사용자에게 효과적으로 보여주기 위해 시각화

데이터 형태

- 데이터의 형태 - 구조 유무 (의미와 데이터가 있는 가?)
 - 정형(structured) 데이터
 - 일정한 구조 보유
 - 예) 데이터베이스 테이블(table, relation)
 - 예) 시장바구니 데이터(market basket data)
 - 매출별 구매 항목 목록에 대한 데이터
 - 행(row)이 항목(item)의 리스트 구성
 - 비정형(unstructured) 데이터
 - 구조가 일정하지 않은 데이터
 - 예) 텍스트(text) 데이터 : 신문기사, SNS 메시지 등
 - 예) 스트림(stream) 데이터 : 지속적으로 관측되어 생성되는 데이터
 - 예) 서열(sequence) 데이터 : 염기 서열, 아미노산 서열 데이터
 - 예) 클릭(click) 데이터 : 홈페이지 방문자들의 순차적인 클릭
 - 예) 시스템 로그(log) 데이터
 - 예) 그래프(graph) 데이터
 - 반정형(semi-structured) 데이터
 - 구조화되어 있지만 관계형 데이터베이스의 테이블과 같은 형태로 저장되기 곤란한 데이터
 - XML(eXtensible Markup Language), JSON(JavaScript Object Notation) 등으로 표현

데이터 형태

데이터의 의미와 데이터가 구조화 된 데이터

XML 데이터



```
<?xml version="1.0" encoding="iso-8859-8" standalone="yes" ?>
<CURRENCIES>
  <LAST_UPDATE>2004-07-29</LAST_UPDATE>
  <CURRENCY>
    <NAME>dollar</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>USD</CURRENCYCODE>
    <COUNTRY>USA</COUNTRY>
    <RATE>4.527</RATE>
    <CHANGE>0.044</CHANGE>
  </CURRENCY>
  <CURRENCY>
    <NAME>euro</NAME>
    <UNIT>1</UNIT>
    <CURRENCYCODE>EUR</CURRENCYCODE>
    <COUNTRY>European Monetary Union</COUNTRY>
    <RATE>5.4417</RATE>
    <CHANGE>-0.013</CHANGE>
  </CURRENCY>
</CURRENCIES>
```

JSON 데이터



```
{ "users": [
  {
    "firstName": "Ray",
    "lastName": "Villalobos",
    "joined": {
      "month": "January",
      "day": 12,
      "year": 2012
    }
  },
  {
    "firstName": "John",
    "lastName": "Jones",
    "joined": {
      "month": "April",
      "day": 28,
      "year": 2010
    }
  }
]
}
```


데이터 마이닝 모델

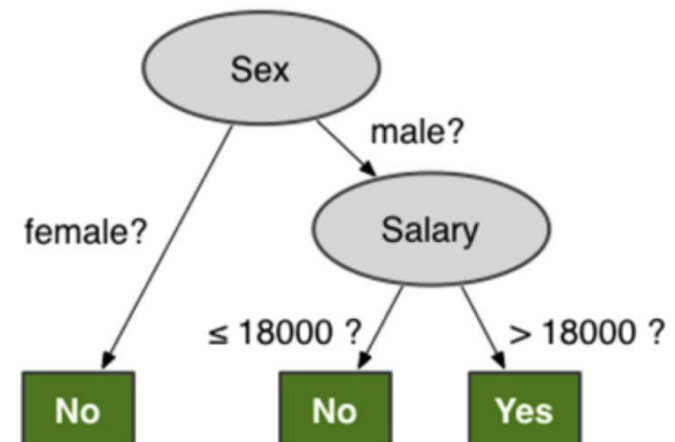
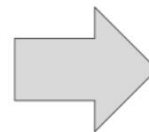
- **classification** : 분류
- **regression** (a.k.a. value estimation) : 회귀 (value 추정)
- **similarity matching** : 유사도 매칭
- **clustering** : 군집화
- **co-occurrence grouping** (a.k.a. association rule discovery) : 동시발생 (관계 규칙 발견)
- **profiling** (a.k.a. behavior description) : 프로파일링 (행동 묘사)
- **link prediction** : 연관성 예측 (ex. recommendation)
- **data reduction** : 데이터 사이즈 줄이기 (불필요한 데이터 제거, 형태 변환 등)
- **causal modeling** : 인과관계 모델링

Classification 분류

- 모집단의 각 개체가 어떠한 클래스의 집합에 속할지 예측하는 방법이다.
- 주로 상호배타적인 클래스로 분류할 때 사용한다.
분류에서 끝나는 것이 아니고, 분류를 학습하여 "예측" 하는 데 활용한다. → 지도학습에 활용
- 각각의 개체를 instance라고 하고, 속성을 attribute라고 한다.
예측의 대상이 되는 속성(클래스)를 classification target 이라고 한다.

Name	Salary	Sex	Age	Buy widget
Bloggs	15000	male	19	No
Jones	25000	male	33	Yes
Smit	23000	female	50	No
Smit	16000	male	40	No
Smit	200	male	10	No
Patel	30000	female	30	No
Steel	25000	male	23	Yes
Higgs	18000	female	55	No
Puggs	50000	male	57	Yes
Puggs	51000	female	57	No

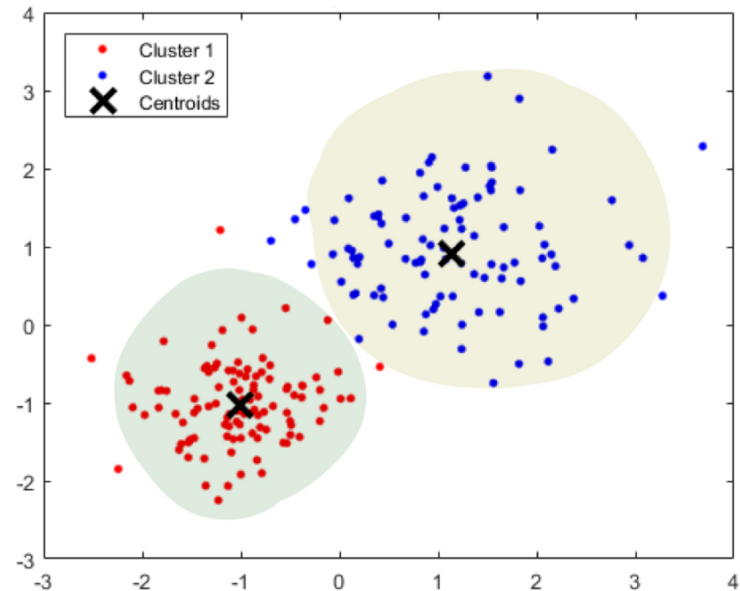
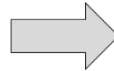
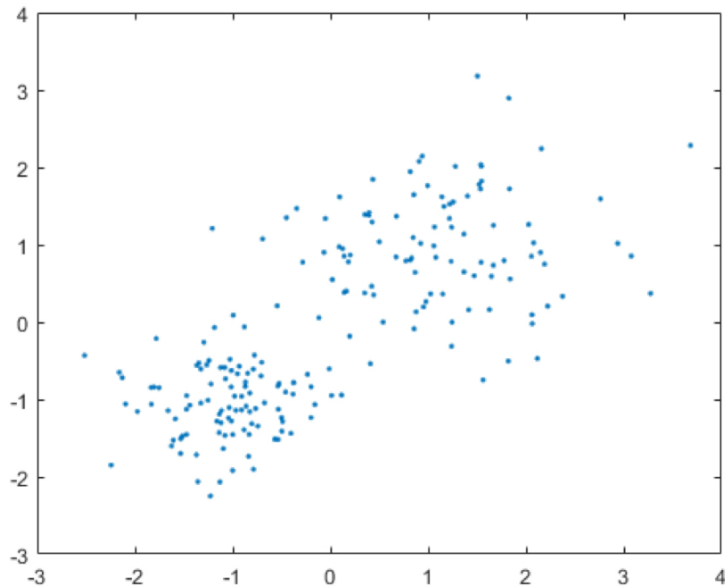
Training dataset



Model

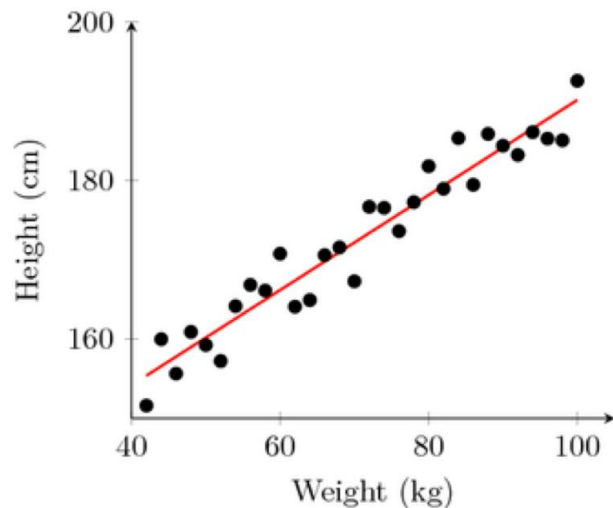
Clustering 군집화

- 비슷한 개체를 하나의 군집으로 묶는 모델이다.
예를 들면, 주로 어떤 종류의 고객을 보유하고 있는지 파악할 때 사용된다.
- 유사도를 측정하기 위해 "거리"를 사용한다.
- 산점도에서 자연스럽게 형성되는 그룹을 분석할 때 유용하다. → 비지도학습에 활용

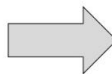


Regression 회귀

- 각 개체가 특정 변수에 대해 어떠한 숫자 값을 가질 것을 예측하는 모델이다.
value estimation 이라고도 불린다. 예를 들어, 특정한 키의 사람은 어떤 값의 몸무게를 가질지 예측하는 것이다.
- 일반적인 과정 : 훈련 데이터셋이 주어질 때, 각 개인에 대해 특정 변수 값을 묘사하는 모델 설정
→ 새로운 개체에 모델을 적용하여 측정된 값 생성
- classification과의 차이점
 - . classification은 인스턴스의 "클래스"를 예측하는 것이다. (ex. YES / NO)
 - . regression은 인스턴스와 관련된 "숫자 값"을 예측하는 것이다. (ex. 178).



Training dataset









$$\text{Height} = 0.9 \cdot \text{Weight} + 105.2$$

Model

Similarity Matching 유사도 매칭

- 알고 있는 데이터를 기반으로 유사한 데이터를 찾는 모델이다.
- 일반적인 과정 : 두 개체 사이의 거리 측정 → 한 개체에 대해서 가장 작은 거리를 갖는 개체 탐색
- 데이터의 종류가 다양하기 때문에 거리 측정에서는 유클리드 거리, 코사인 거리 등 다양한 종류의 거리를 사용한다.

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1 	8	1	?	2	7
User 2 	2	?	5	7	5
User 3 	5	4	7	4	7
User 4 	7	1	7	3	8
User 5 	1	7	4	6	?
User 6 	8	3	8	3	7

What users are similar to User 1?

What items are similar to Item 3?

(유클리드 거리) 예를 들어서,

User 3 = (5, 4, 7, 4, 7), User 4 = (7, 1, 7, 3, 8) → distance $\cong 3.87$

다른 데이터 마이닝 모델

Co-occurrence grouping 동시발생

: 개체들 사이의 동시 발생을 통해 관계를 찾는 모델이다.

- 예를 들어서, 마트 판매 상품을 분석해보니 기저귀를 살 때 맥주도 같이 사는 손님의 비율이 높았다. 이와 같이 분석을 통해 특별 프로모션, 상품 진열, 세트 판매, 추천 등 마케팅에 활용할 수 있다.

Profiling 행동 특성 묘사

: 개인이나 집단의 전형적인 행동을 특징짓는 모델이다.

- normal한 행동특성에서 벗어난 행동을 탐지할 때 매우 유용하다.
- 프로파일은 normal한 행동을 묘사하기 때문에, 갑자기 그것에 벗어난 행동을 할 때 알림을 주는 것이다.

Link Prediction 연관성 예측

: 데이터 아이템들 사이의 연관성을 예측하는 모델이다.

- 일반적으로 연결고리가 존재한다고 제안하고, 그 강도를 추정함으로써 사용된다.
- "추천" 알고리즘에 매우 유용하다. SNS에서 친구를 추천해주거나, 넷플릭스 등에서 영화를 추천할 때 주로 사용된다.

Data Reduction 데이터 사이즈 감소

: 큰 데이터셋을 중요한 정보를 많이 포함하는 작은 데이터셋으로 사이즈를 줄이는 것이다.

- 작은 데이터셋은 다루기 쉽고, 정보나 향상된 인사이트를 보다 더 잘 드러낸다.
- 하지만, 정보의 손실 또한 일어나기 쉽다는 단점이 있다.

Causal Modeling 인과관계 모델링

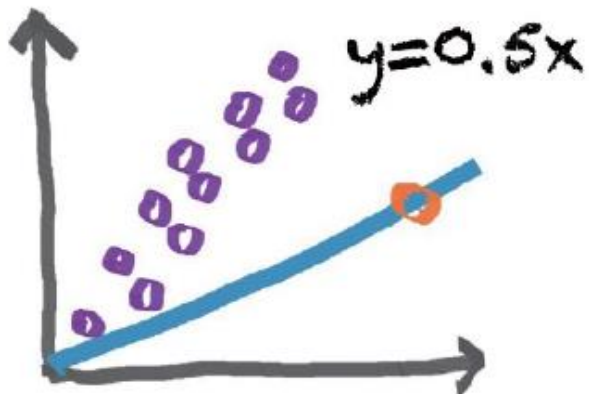
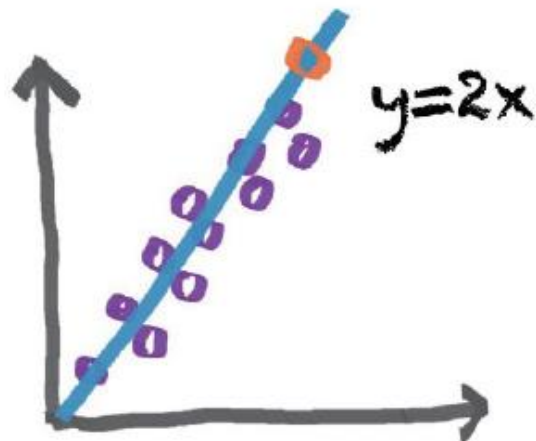
: 다른 사건에 실질적으로 영향을 주는 사건을 찾는 모델이다.

- 예를 들어 담배피는 사람들 중에 이가 누런 사람과 폐암에 걸린 사람들이 있다고 할 때, 누래짐과 폐암이 담배때문에 발생한 건지, 아니면 그냥 그런 사람들이 담배를 피는 건지 인과관계를 파악하는 것이다.

회귀 분석

★ 회귀분석 (Regression Test)

- 회귀분석은 독립변인이 종속변인에 영향을 미치는지 알아보고자 할 때 실시하는 분석방법
- 단순 선형 회귀분석은 독립변수 X(설명변수)에 대하여 종속변수 Y(반응변수)들 사이의 관계를 수학적 모형을 이용하여 규명하는 것
- 규명된 함수식을 이용하여 설명변수들의 변화로부터 종속변수의 변화를 예측하는 분석
- 선형 회귀 모델은 가장 단순하고 학습 속도가 빠르며 인간이 이해하기 쉬워 분류하는 요인의 수가 적을 때는 활용하기가 용이



회귀 분석

'키(Height)에 따른 몸무게(Weight)' 를 예로 들면,

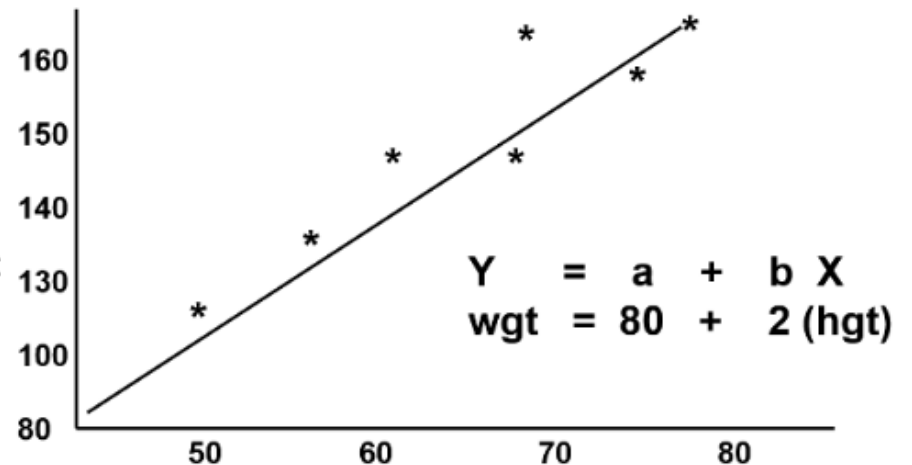
$$\text{Weight} = a + b * \text{Height}$$

가 되며, 결국 Height에 따라 Weight가 결정되므로, Height는 독립변수, Weight는 종속변수.

키(height) inch	몸무게(weight) pound
50	120
58	130
61	145
69	145
70	165
75	155
78	165

Y-axis:

Body Weight
(pounds)

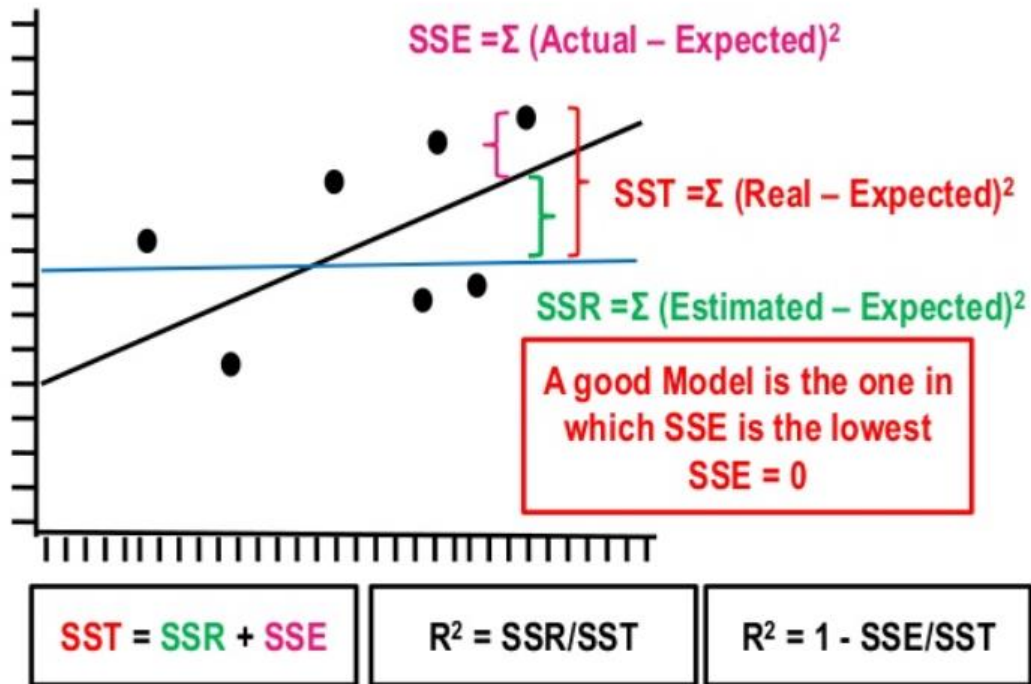


X-axis: Height (inches)

회귀 분석의 통계적 검정

모형이 믿을 만 한가?

- 최소제곱법(Method of Least Squares Estimation) 사용



$$\sum_{i=1}^n (y_i - \bar{y})^2: \text{총제곱합 (SST)}$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2: \text{잔차제곱합 (SSE)}$$

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2: \text{회귀제곱합 (SSR)}$$

(관찰치의 편차) = (회귀선에 의해 설명되는 편차) + (회귀선에 의해 설명되지 않는 편차)

$$\Rightarrow \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

모델을 평가하는 지표

MAE(Mean of Absolute Error) : 모델의 예측값과 실제값의 차이의 절대값의 평균
- 절대값을 취하기 때문에 가장 직관적으로 알 수 있는 지표이다. (해석에 용이하다.)

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x|$$

MSE(Mean of Squared Error) : 모델의 예측값과 실제값의 차이의 제곱값의 평균
- 제곱을 하기 때문에 MAE와는 다르게 모델의 예측값과 실제값 차이의 면적의(제곱)합이다.
(평균제곱오차)

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

RMSE(Root Mean Squared Error) : MSE에 루트를 씌워 사용한다.
- RMSE를 사용하면 오류 지표를 실제값과 유사한 단위로 다시 변환하여 해석을 쉽게한다.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum(\hat{y}-y)^2}{n}}$$

R-squared (Coefficient of determination, 결정계수) : R-squared는 현재 사용하고 있는 x변수가 y변수의 분산을 얼마나 줄였는가이다.

- y평균값 모델(기준모델)을 사용했을 때 대비 우리가 가진 x변수를 사용함으로써 얻는 성능 향상의 정도
- 값이 1에 가까우면 데이터를 잘 설명하는 모델이고 0에 가까울수록 설명을 못하는 모델이라고 생각할 수 있다.

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

* 머신러닝에서는 손실함수(Loss Function), 비용함수(Cost Function) 라고 함


인공지능 실습

인공지능 실습

<https://animalface.site/ko/index.html>

성별을 선택하세요

여자 ☒ 남자



포근한 매력의 곰상

첫 인상은 무서워 보이지만 알고 보면 귀여운 매력의 당신! 꼼꼼하고 섬세한 성격으로 연인을 헌신적으로 챙겨주는 당신은 연인에게 듬직한 존재! 포근한 매력에 듬직할까지 갖춘 최고의 남자대!

곰상 연예인: 마동석, 조진웅, 조세호, 안재홍

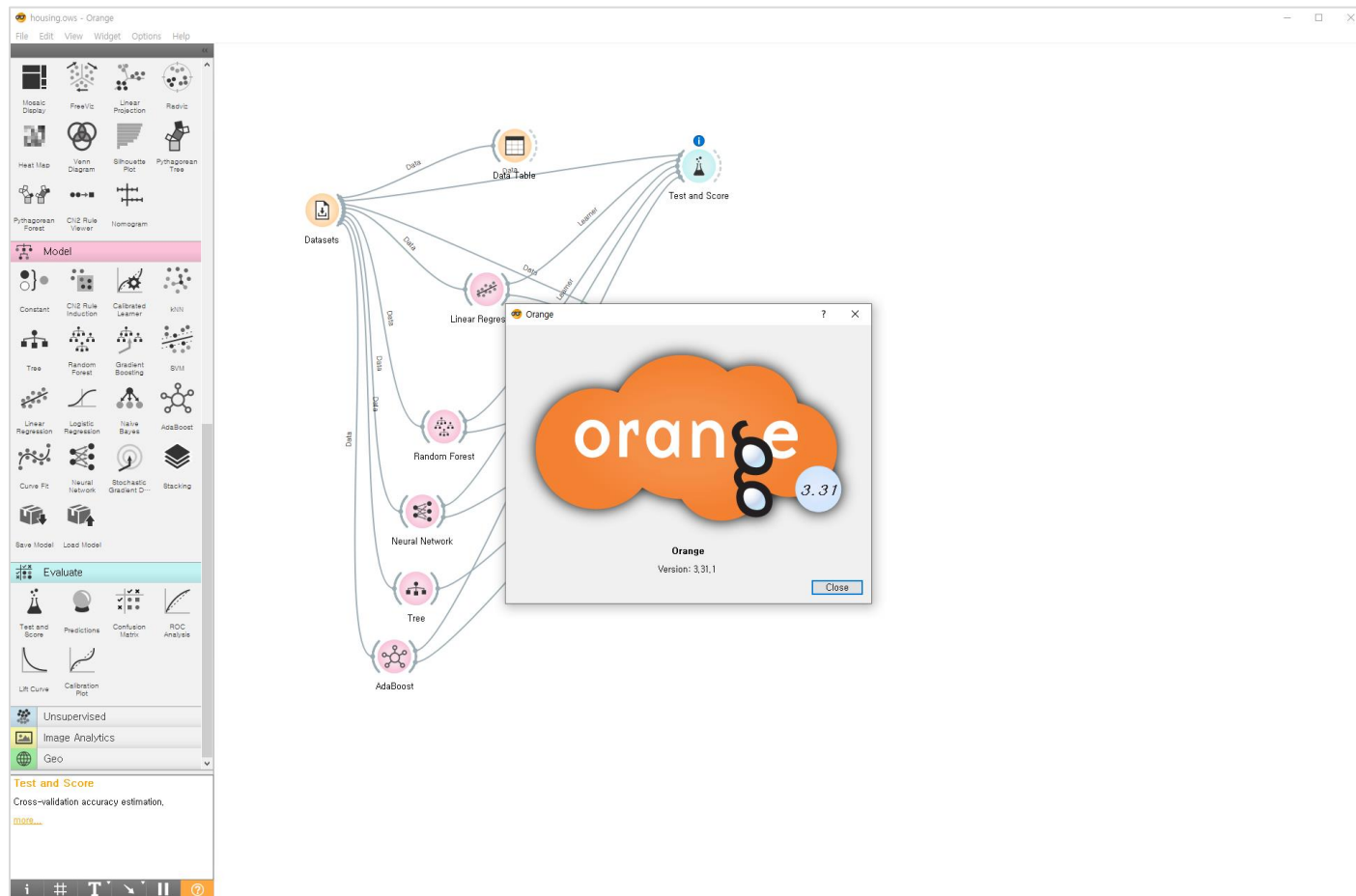
곰상	99%
곰룩상	1%
강아지상	0%
토피상	0%
고양이상	0%

🔔 🌐 🏠 + 63.2K

다른 사진으로 재시도

인공지능 실습

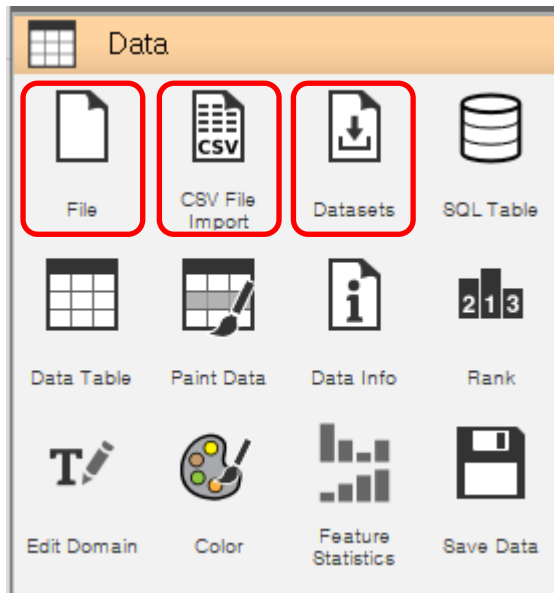
<https://orangedatamining.com/download/#windows>



인공지능 실습

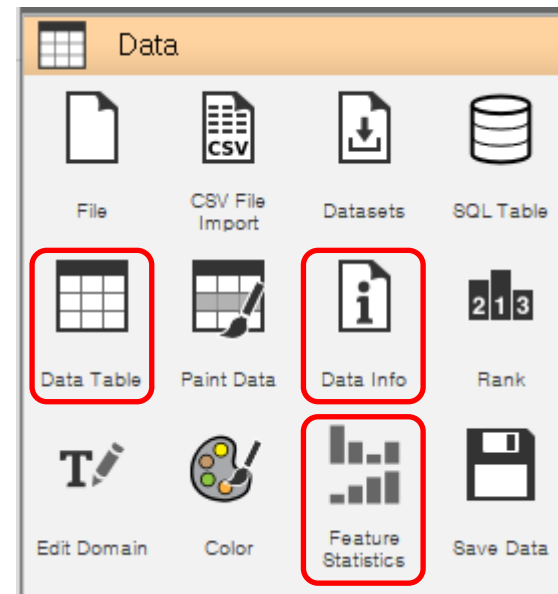
1. 데이터 입력

- Datasets
- File / CSV File



2. 데이터 보기

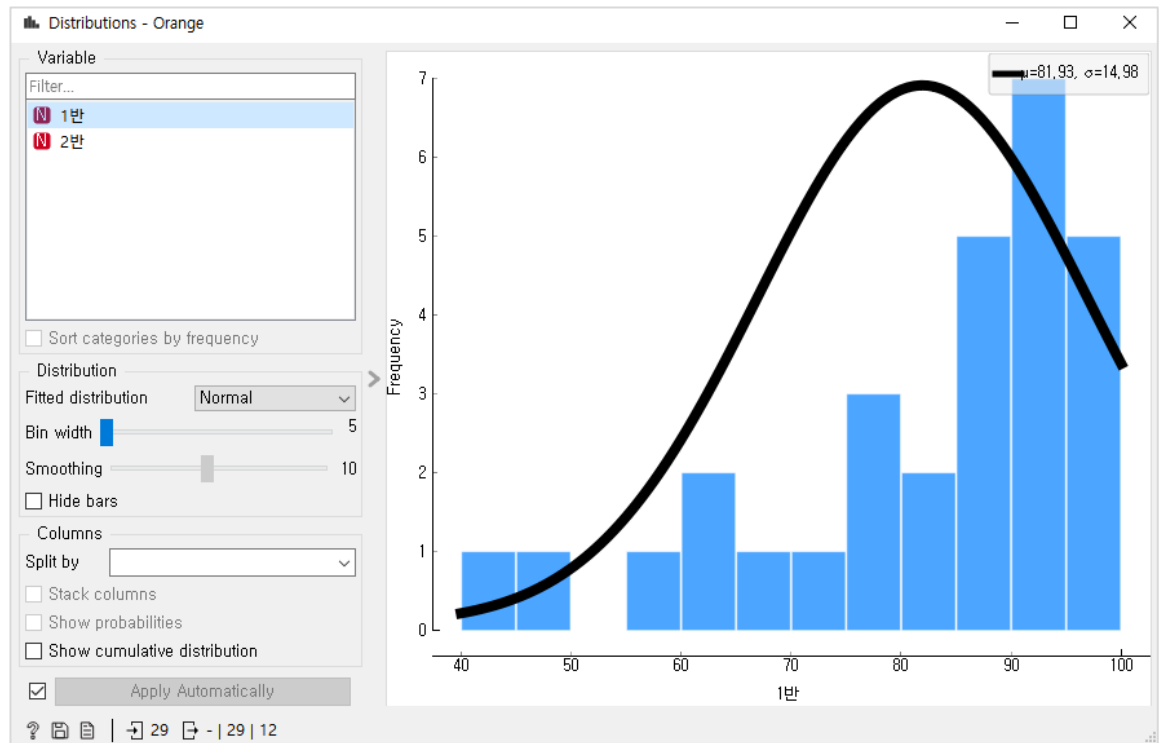
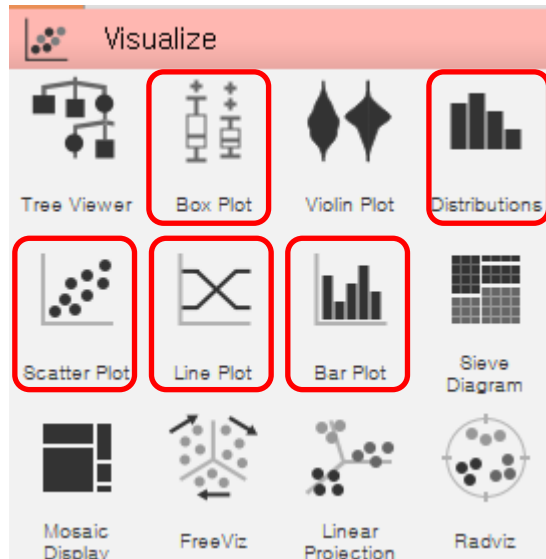
- Data Table
- Data Info
- Feature Statistics



인공지능 실습

3. 데이터 차트

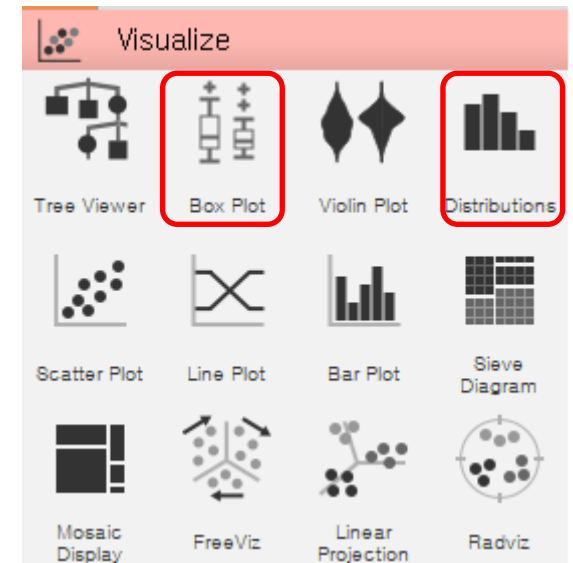
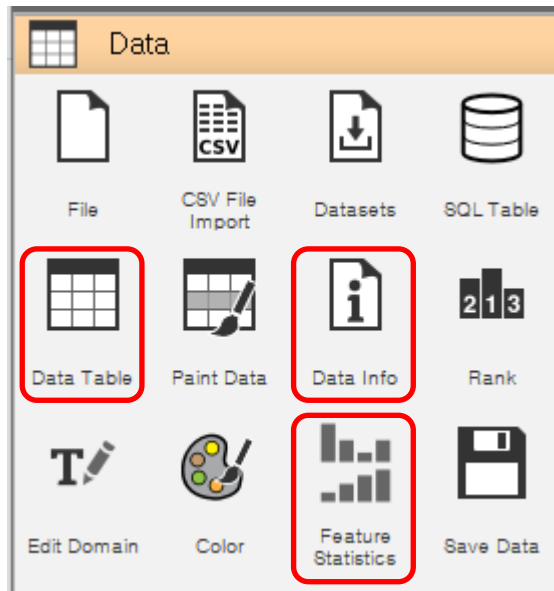
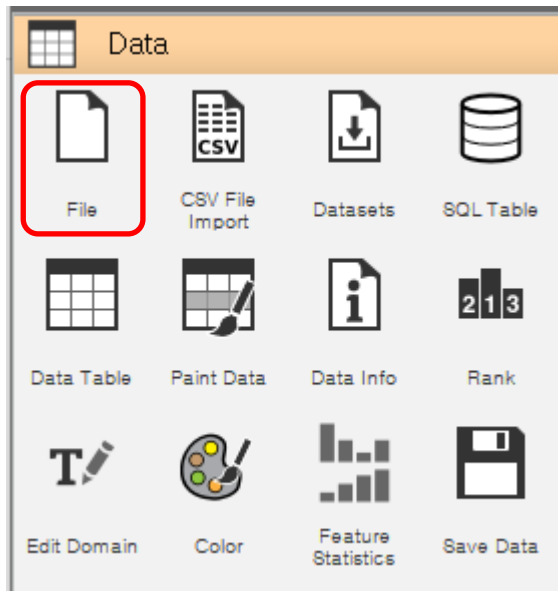
- Visualize 위젯



인공지능 실습

실습 1

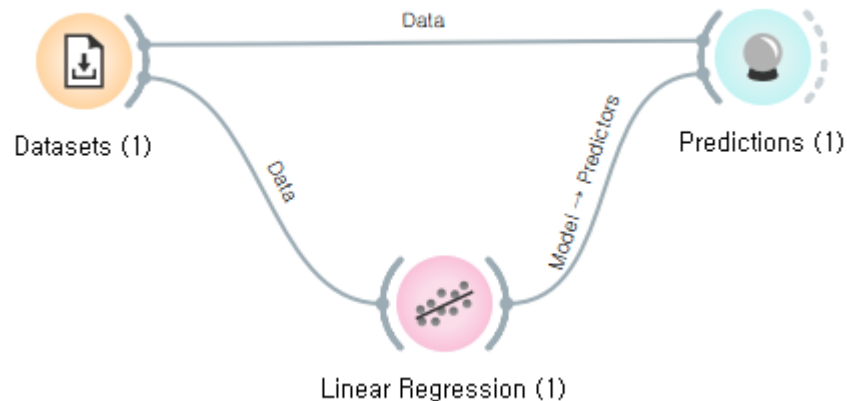
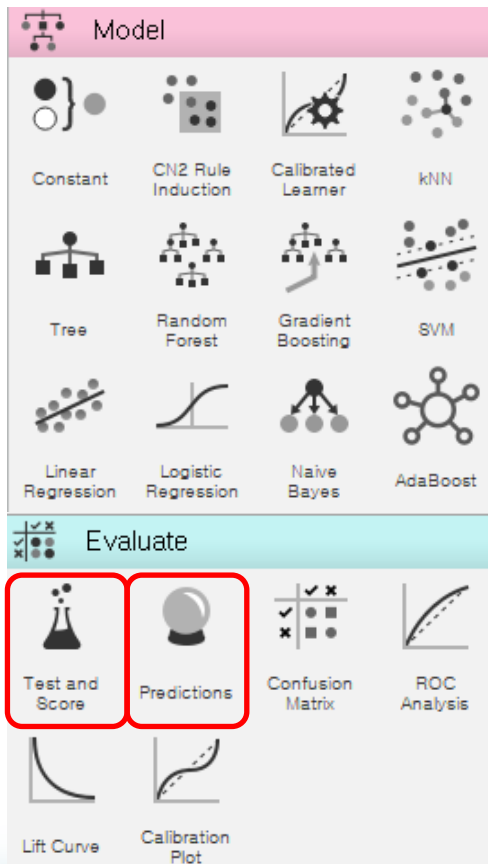
- 데이터 불러오기 : '시도별 출생 성비' 읽어 오기
- 데이터 보기 : Data Table, Data Info, Feature Statistics
- 데이터 차트 : Box plot, Distributions



인공지능 실습

4. 데이터 분석 : 모델

- Model 위젯
- Evaluate 의 'Predictions'나 'Test and Score' 와 연결
- Data는 따로 연결해 주어야 함



인공지능 실습

데이터 불러오기 : housing

보스턴의 506개 타운(town)의 13개 독립변수값로부터 해당 타운의 주택가격 중앙값을 예측하는 문제

독립변수

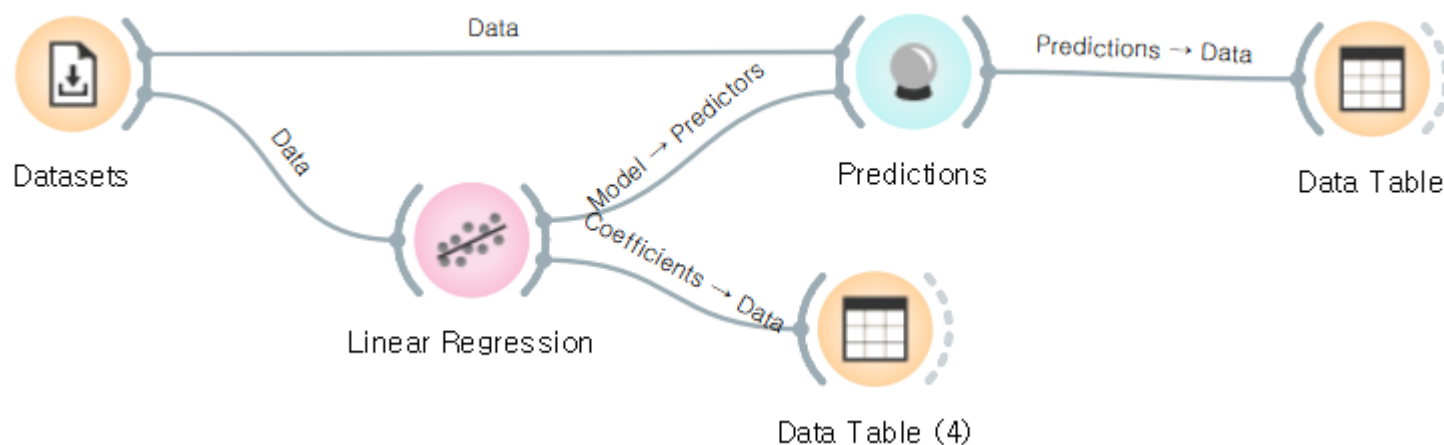
- **CRIM** : 범죄율
- **INDUS** : 비소매상업지역 면적 비율
- **NOX** : 일산화질소 농도
- **RM** : 주택당 방 수
- **LSTAT** : 인구 중 하위 계층 비율
- **B** : 인구 중 흑인 비율
- **PTRATIO** : 학생/교사 비율
- **ZN** : 25,000 평방피트를 초과 거주지역 비율
- **CHAS** : 찰스강의 경계에 위치한 경우는 1, 아니면 0
- **AGE** : 1940년 이전에 건축된 주택의 비율
- **RAD** : 방사형 고속도로까지의 거리
- **DIS** : 직업센터의 거리
- **TAX** : 재산세율

종속변수

- **MEDV** : 보스턴 506개 타운의 1978년 주택 가격 중앙값 (단위 1,000 달러)

인공지능 실습

- 회귀분석 모델링



Datasets - Orange					
Search for data set ...					
	Title	Size	Instances	Variables	Tab
●	Housing	33.9 KB	506	14	N
●	HDI	65.7 KB	188	52	
●	Kickstarter projects	214.1 KB	1163	15	C
●	Sailing	455 bytes	20	4	C

인공지능 실습

Predictions (1) - Orange

예측 결과

Restore Original Order

	Linear Regression (1)	MEDV	CRIM	ZN	INDUS
1	30.0	24.0	0.00632	18.0	2.31
2	25.0	21.6	0.02731	0.0	7.07
3	30.6	34.7	0.02729	0.0	7.07
4	28.6	33.4	0.03237	0.0	2.18
5	27.9	36.2	0.06905	0.0	2.18
6	25.3	28.7	0.02985	0.0	2.18
7	23.0	22.9	0.08829	12.5	7.87
8	19.5	27.1	0.14455	12.5	7.87
9	11.5	16.5	0.21124	12.5	7.87
10	18.9	18.9	0.17004	12.5	7.87
11	19.0	15.0	0.22489	12.5	7.87
12	21.6	18.9	0.11747	12.5	7.87
13	20.9	21.7	0.09378	12.5	7.87

☒ Show performance scores

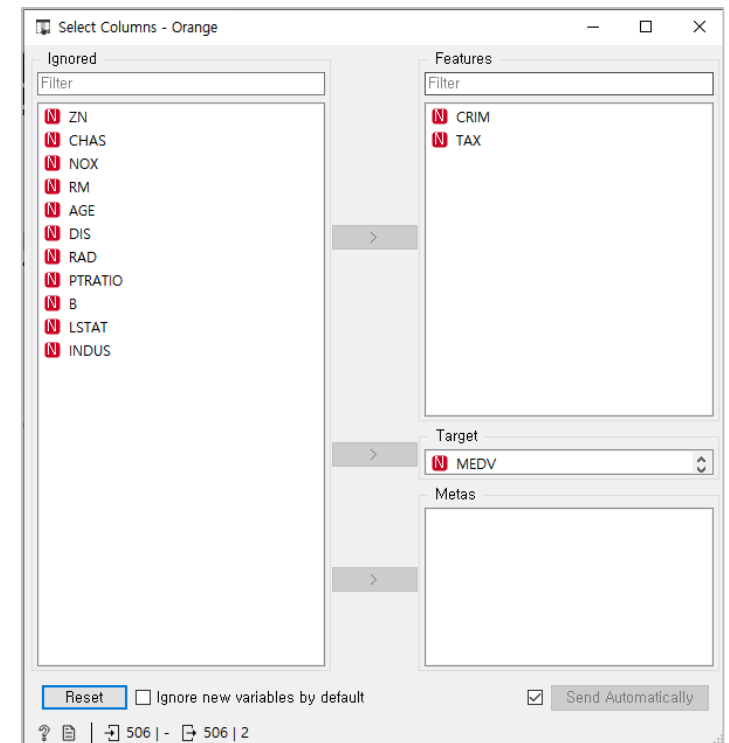
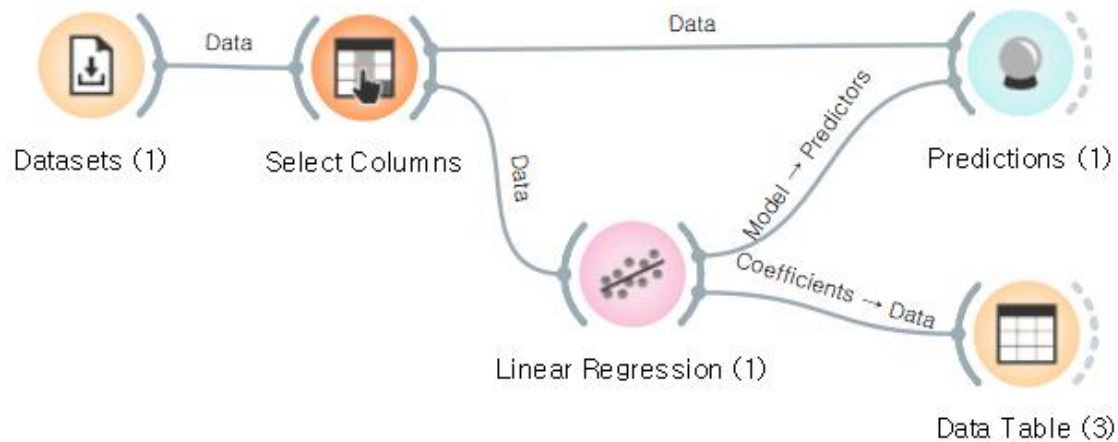
Model	MSE	RMSE	MAE	R2
Linear Regression (1)	21.895	4.679	3.271	0.741

모델 지표

? | 506 | 506 | 1x506

인공지능 실습

- 회귀분석 모델링



인공지능 실습

실습 2

- housing 데이터에서 흑인 비율과 주택당 방수를 독립변수로 회귀분석
- R2 값은 얼마인가? 범죄율과 재산세율에 비해 더 상관성이 높은 가?

독립변수

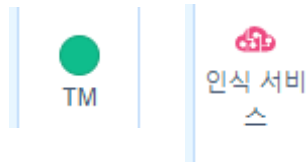
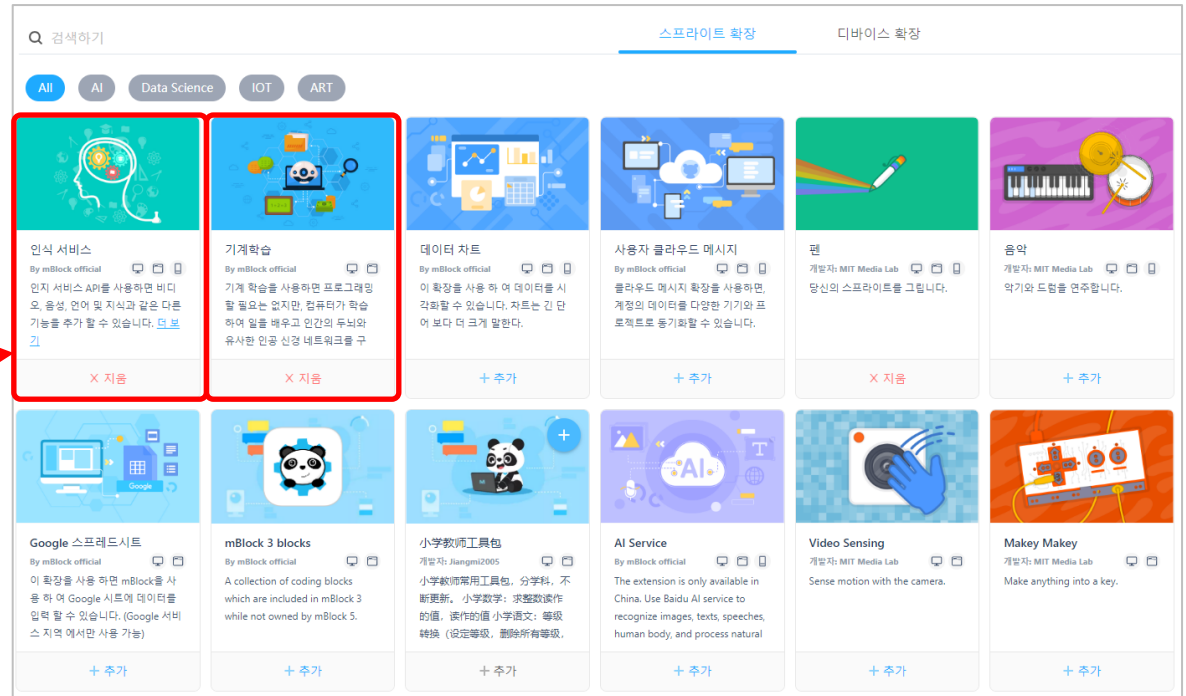
- CRIM : 범죄율
- INDUS : 비소매상업지역 면적 비율
- NOX : 일산화질소 농도
- RM : 주택당 방 수
- LSTAT : 인구 중 하위 계층 비율
- B : 인구 중 흑인 비율
- PTRATIO : 학생/교사 비율
- ZN : 25,000 평방피트를 초과 거주지역 비율
- CHAS : 찰스강의 경계에 위치한 경우는 1, 아니면 0
- AGE : 1940년 이전에 건축된 주택의 비율
- RAD : 방사형 고속도로까지의 거리
- DIS : 직업센터의 거리
- TAX : 재산세율

종속변수

- MEDV : 보스턴 506개 타운의 1978년 주택 가격 중앙값 (단위 1,000 달러)

MBlock 실습

인공지능 MBlock 실습



인공지능 MBlock 실습

카메라를 연결하고 로그인을 한 다음 블록을 편집합니다.



이 스프라이트를 클릭했을 때

클릭했을 때

1 초 후, 사람 나이 인식하기
나이 인식 결과 을(를) 말하기

1 초 후, 감정 인식하기
감정 행복 강도 을(를) 말하기

2 초 동안 적혀진 영어 인식하기
문자 인식 결과 을(를) 말하기

1 초 후, 성별 인식하기
성별 인식 결과 을(를) 말하기

1 초 후 이미지의 이미지 인식
이미지 인식 인식 결과 을(를) 말하기

2 초 후, 미소 점수 인식하기
미소 인식 결과 을(를) 말하기

1 초 후, 안경 유형 인식하기
입고 독서 안경 ? 을(를) 말하기

1 초 후, 감정 인식하기
감정이 행복 입니까? 을(를) 말하기

1 초 후, 머리 동작 인식하기
머리 요각 각도(°) 을(를) 말하기