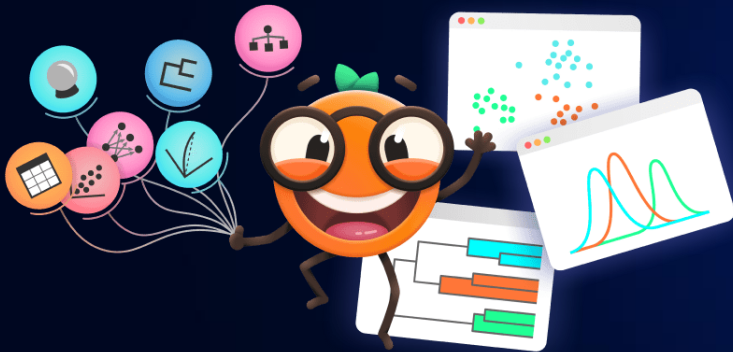


클릭으로 하는 AI분석 웨비나

# ORANGE와 데이터마이닝



# 2025년 신학기 **대폭 할인** 이벤트

**통계분석+빅데이터+데이터마이닝 3종 세트를 두고 두고 평생 학습하세요**

W 데이터캠퍼스

통계분석

## SPSS-STATA

핵심 통계분석에서 고급 통계분석까지

김원표 지음



SPSS-STATA, FROM CORE TO ADVANCED  
FROM CORE STATISTICAL ANALYSIS TO ADVANCED STATISTICAL ANALYSIS

W 데이터캠퍼스

빅데이터

## 빅데이터 2.0

파이썬, 머신러닝, 딥러닝, 텍스트마이닝

김원표 지음



BIG DATA ANALYSIS 2.0  
PYTHON, MACHINE LEARNING, DEEP LEARNING and TEXT MINING

W 데이터캠퍼스

클릭으로 완성하는 머신러닝 딥러닝  
(with ORANGE)

## ORANGE와 핵심 마이닝

김원표 지음



No Code AI ANALYSIS  
Data Mining, Machine Learning / Deep Learning / Text & Image Mining

# 1종: 통계분석 USB+교제 세트

모든 통계분석을 마스터할 수 있는 소장용 강의! 7개 과목, 231강좌

선착순 30명 한정판매

이미 500여명의 교수님/연구진 300개 대학에서  
구매하여 입증한 명강의!

75% 할인

**600,000원** ← **2,400,000원**

- SPSS 기본분석
- SPSS 고급회귀분석
- AMOS 구조방정식모델분석
- STATA 패널데이터분석
- STATA 메타분석
- STATA 시계열분석
- HLM 다층선형모델분석



구매 사이트





# 2종: 빅데이터분석 USB+교제 세트

모든 빅데이터분석을 마스터할 수 있는 소장용 강의! 4개 과목, 90강좌

선착순 30명 한정판매

파이썬을 활용한 빅데이터 분석의 모든 것!  
AI시대 연구자의 필수!

65% 할인

560,000원 ← 1,600,000원

- Python 핵심: 파이썬 데이터 다루기 마스터
- 머신러닝: 머신러닝 마스터
- 딥러닝: 딥러닝 마스터
- 텍스트마이닝: 텍스트마이닝 마스터



구매 사이트



# 3종: 클릭으로 완성하는 머신러닝과 딥러닝 USB+교제 세트

무료 데이터마이닝툴 ORANGE를 활용하여 쉽게 연구물 작성하기!

선착순 30명 한정판매

1개월 안에 AI를 이용하여 내 논문과 연구물을  
작성하실 수 있습니다

50% 할인

600,000원 ← 1,200,000원

- ORANGE와 핵심 마이닝
- 지도학습 마스터
- 비지도학습 마스터
- 텍스트와 이미지분석 마스터



구매 사이트



# 통계분석 분야 최고의 고수가 직접 강의한 최고의 명강의라 자신합니다!



- ✓ 김원표 교수(와이즈인컴퍼니 대표)
- ✓ 교수를 가르치는 교수
- 1,000여명의 교수 대상 통계분석, 빅데이터분석 강의
- 서울대병원 의사 통계분석 교육(7년간 강의, 만족도 4.5/5점)

• **경제부총리상 수상** (지식서비스 산업발전 유공)

• **24권의 통계·빅데이터분석 서적 출간**

• **분석자동화 솔루션 개발 총괄** (조달등록제품)



## 구매 문의

- 기간: 각 종 30명 한정판매
- 할인: 50~75% 할인
- 문의: 02-558-5144 / [hs9177@wiseinc.co.kr](mailto:hs9177@wiseinc.co.kr) (고현서 연구원)
- 우리은행 / 1005-402-421172 / (주)와이즈인컴퍼니

2종 구매자 10% 추가할인

3종 구매자 15% 추가할인

대학도서관에서 구매 신청 후, 활용하실 수도 있습니다!

\*) 2종 이상 구매자는 반드시 미리 전화/이메일로 상담 후, 결제 혹은 입금을 하셔야 합니다.

## INDEX

### PART1. ORANGE소개와 설치하기

---

1. ORANGE와 머신러닝
2. ORANGE 기능과 위젯 둘러보기
3. 데이터 연동과 분석 스토리 설계하기

### PART2. CASE분석1: ORANGE로 지도학습 해보기

---

1. ORANGE의 지도학습 알고리즘 소개
2. 회귀의 지도학습 시나리오 설계와 실습
3. 분류의 지도학습 시나리오 설계와 실습

### PART3. CASE분석2: ORANGE로 비지도학습 해보기

---

1. ORANGE의 비지도학습 알고리즘 소개
2. 군집으로 유사한 집단 묶어보기

### PART4. CASE분석3: ORANGE로 텍스트마이닝 해보기

---

1. 텍스트마이닝 기본이해와 활용
2. 텍스트마이닝 시나리오 설계와 기초분석
3. 텍스트 클러스터링과 시각화
4. 텍스트 임베딩과 활용

### PART5. CASE분석4: ORANGE로 이미지분석 해보기

---

1. 이미지분석 기본이해와 활용
2. 이미지 임베딩과 시각화
3. 이미지 분류와 전이모델



# I

P A R T

## ORANGE 소개와 설치하기

---

1. ORANGE와 머신러닝
2. ORANGE 기능과 위젯 둘러보기
3. 데이터 연동과 분석 스토리 설계하기

# 1. ORANGE와 머신러닝

---



## 개념 설명

- ORANGE는 Python 기반 드래그 & 드롭 방식의 분석도구로서 연구 및 실무에 활용도가 매우 높음
- 최근 들어 다양한 분야에 사용이 확장되고 있음

### 특징과 장점

#### ORANGE

- 1996년부터 슬로베니아의 류블랴나 대학(University of Ljubljana)에서 개발을 시작한 오픈 소스
- 데이터 마이닝, 머신러닝, 시각화의 도구가 지속적으로 업데이트 되고 있음
- 탐색, 학습, 시각화, 텍스트, 이미지까지 Python 기반으로 위젯을 연결하는 방식으로 다루기 매우 쉬움

#### 장점

- Easy to Use
- Various Algorithms for Datamining
- Free for Use

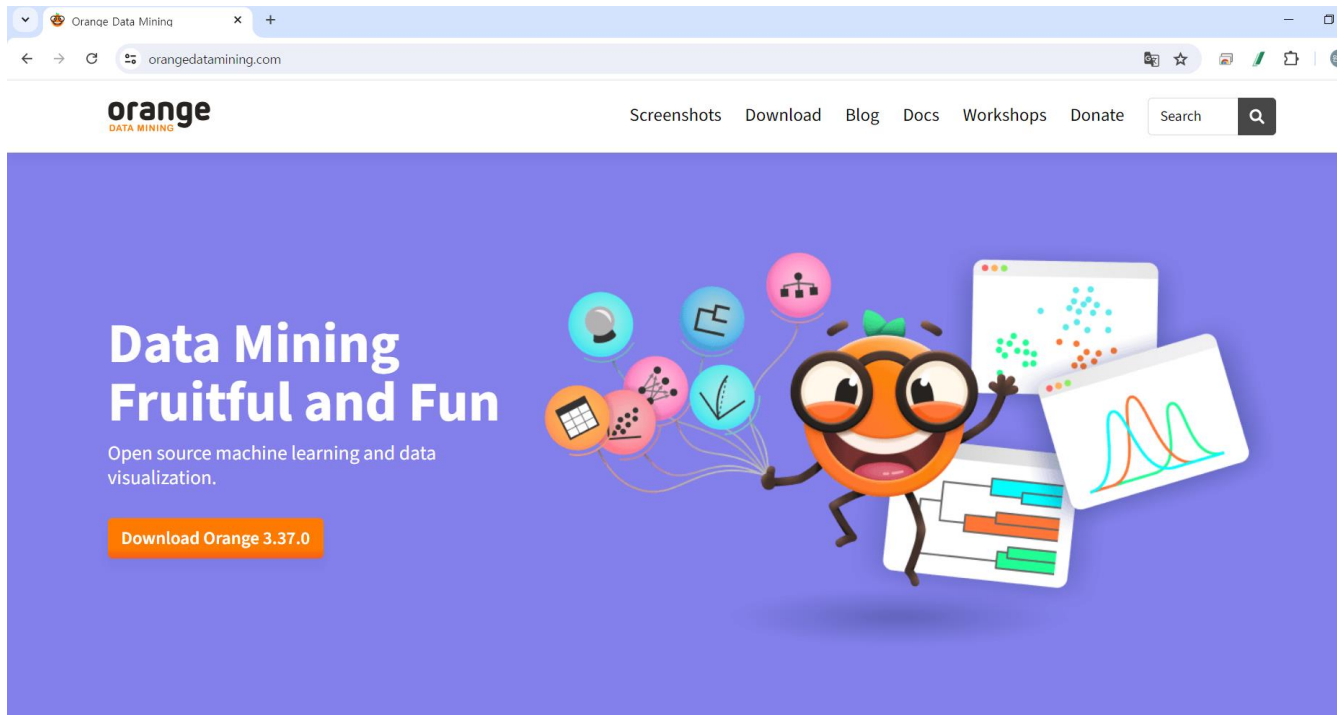
#### Key Point

분석 결과의 시각화를 통한 이해성과 가독성 향상

# 설치하기

- <https://orangedatamining.com/>
- Download Orange 버튼을 눌러 상세 설치 페이지로 이동
- 윈도우, 맥 등 사용환경에 적합한 버전 설치

## > 설치하기



## Windows

### Standalone installer (default)

↓ [Orange3-3.37.0-Miniconda-x86\\_64.exe](#)  
Can be used without administrative privileges.

### Portable Orange

↓ [Orange3-3.37.0.zip](#)  
No installation needed. Just extract the archive and open the shortcut in the extracted folder.

## macOS

### Orange for Apple silicon

↓ [Orange3-3.37.0-Python3.11.8-arm64.dmg](#)

### Orange for Intel

↓ [Orange3-3.37.0-Python3.10.11-x86\\_64.dmg](#)

## Key Point

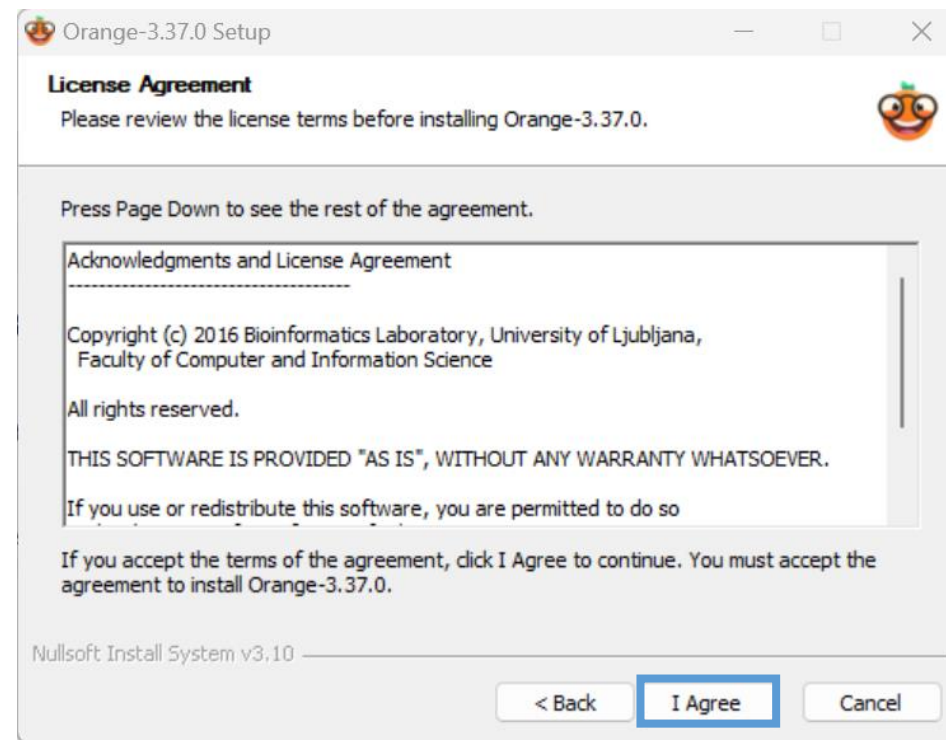
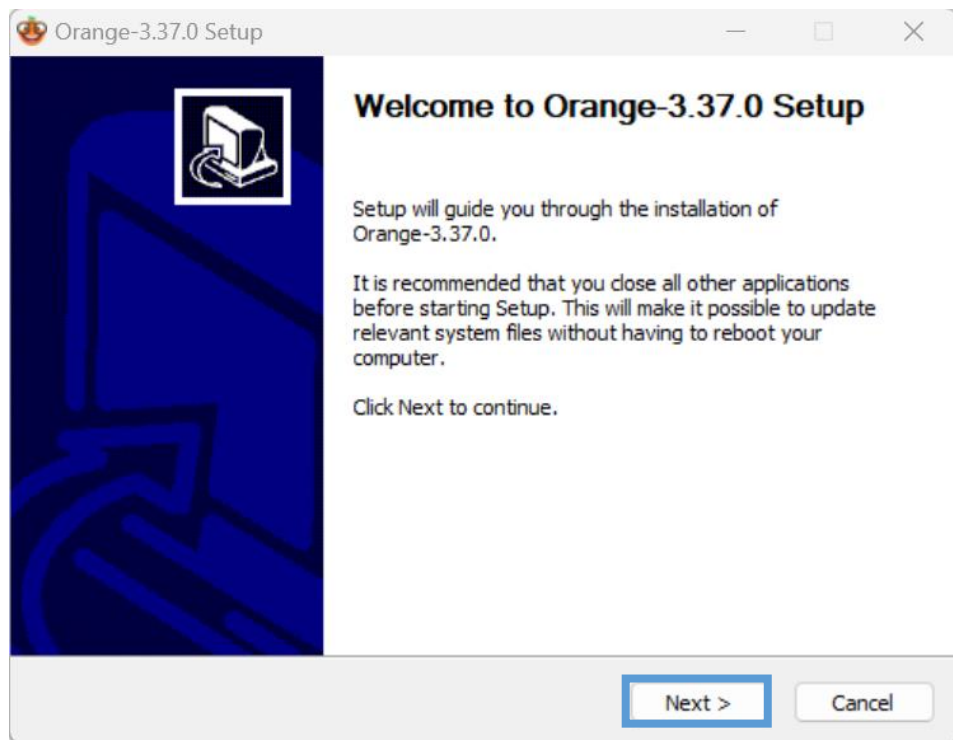
분석 결과의 시각화를 통한 이해성과 가독성 향상



# 설치하기

- 설치파일을 실행하여 안내되는 절차에 따라 설치를 진행

## > 설치파일 실행 1



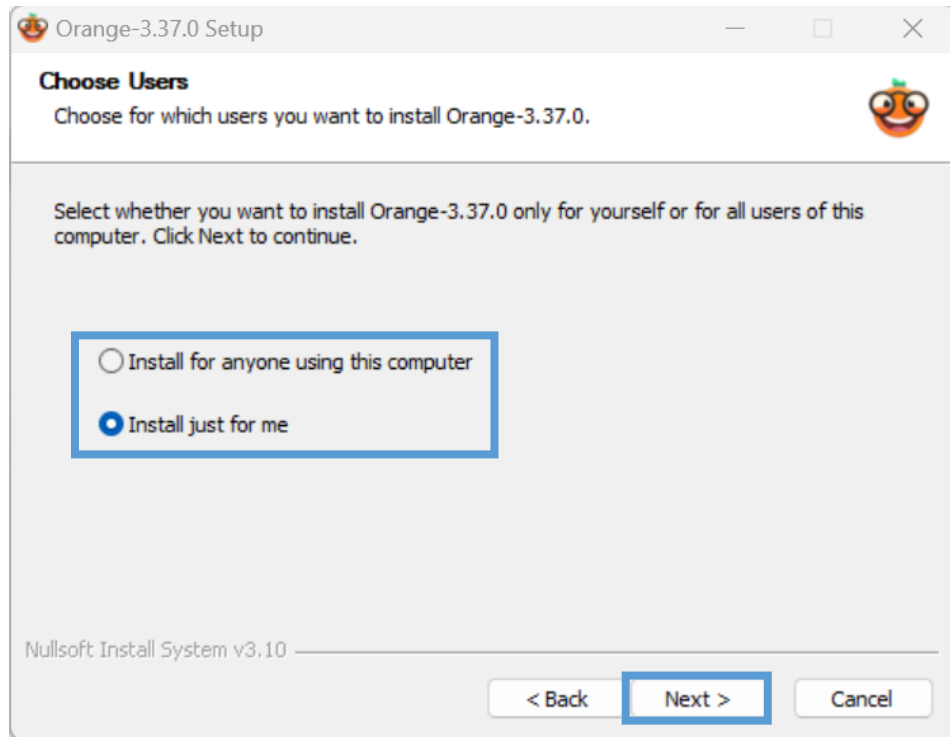
### Key Point

사용자 환경에 따라 적합한 ORANGE 버전 설치

# 설치하기

- 설치파일을 실행하여 안내되는 절차에 따라 설치를 진행

## > 설치파일 실행 2



## > 이미지 관련 설명 (필요시)

- 프로그램 사용 범위 설정
- Install for anyone using this computer  
: 컴퓨터에 설정된 모든 사용자 계정이 프로그램 사용
- Install just for me  
: 현재 사용자 계정만 프로그램 사용

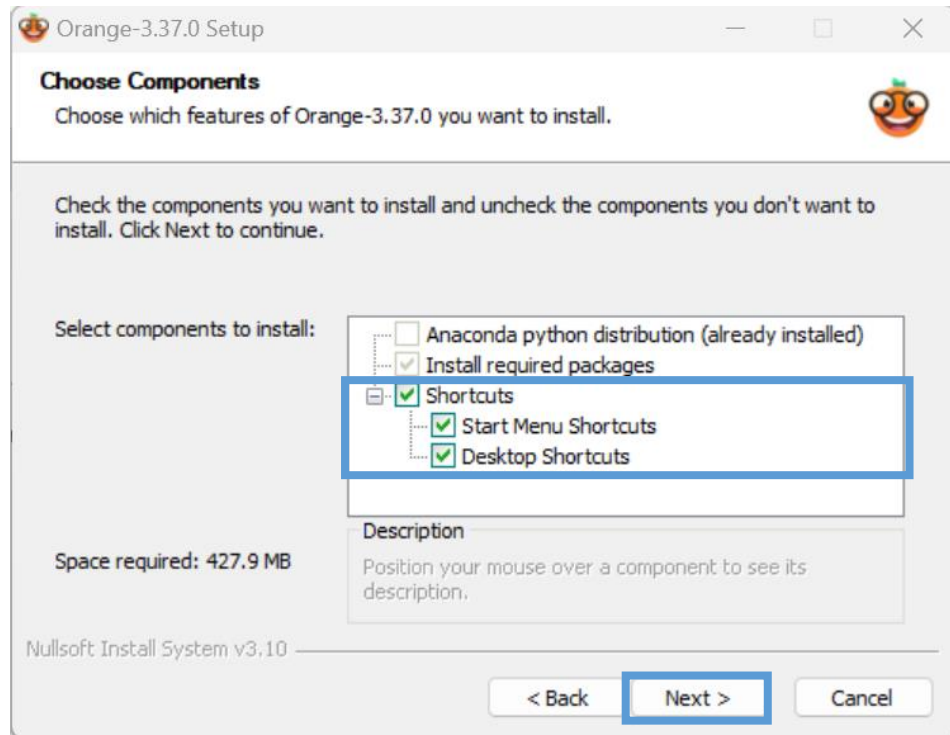
### Key Point

사용자 환경에 따라 적합한 ORANGE 버전 설치

# 설치하기

- 설치파일을 실행하여 안내되는 절차에 따라 설치를 진행

## > 설치파일 실행 3



## > 이미지 관련 설명 (필요시)

- Start Menu Shortcuts  
: 시작 메뉴에 바로가기 추가
- Desktop Shortcuts  
: 바탕화면에 바로가기 추가

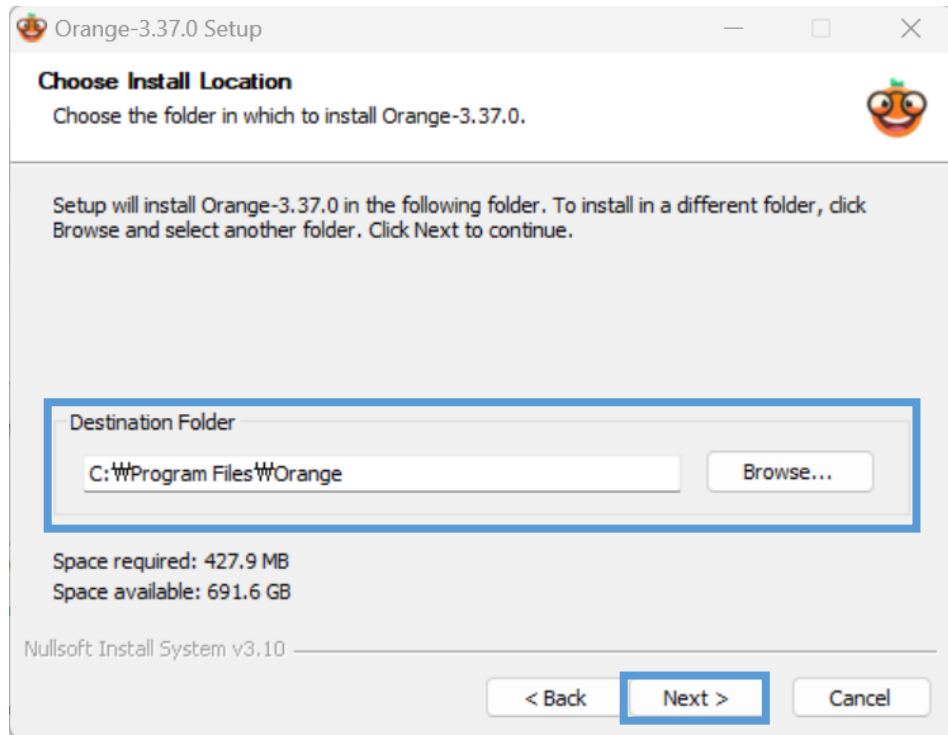
### Key Point

사용자 환경에 따라 적합한 ORANGE 버전 설치

# 설치하기

- 설치파일을 실행하여 안내되는 절차에 따라 설치를 진행

## > 설치파일 실행 4



## > 이미지 관련 설명 (필요시)

- Destination Folder  
: orange 프로그램을 설치할 경로 설정

### Key Point

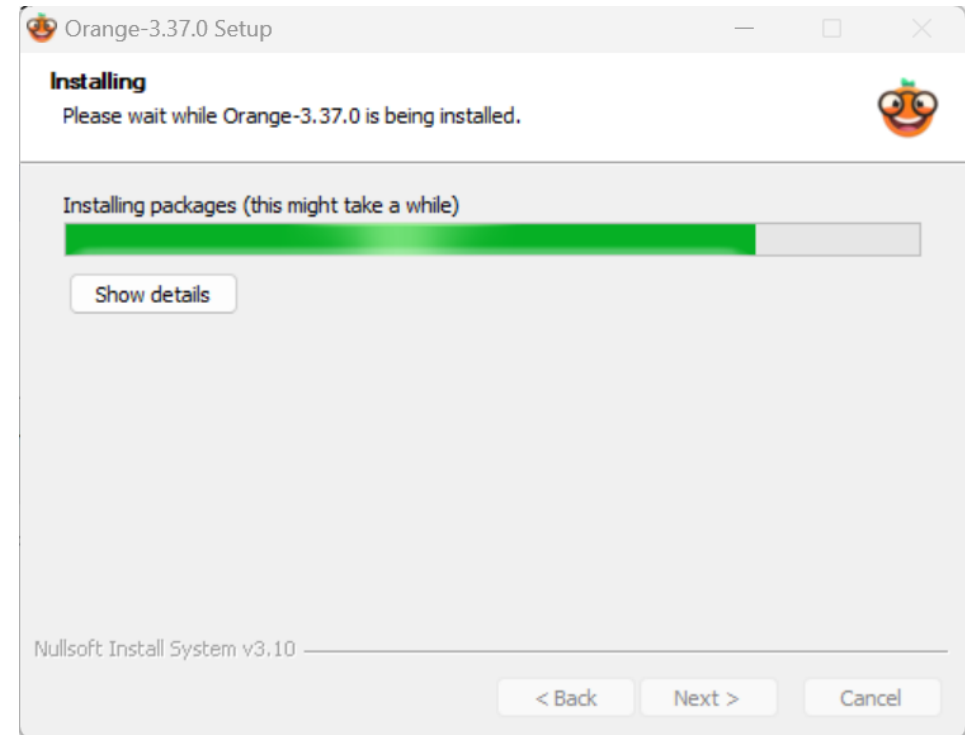
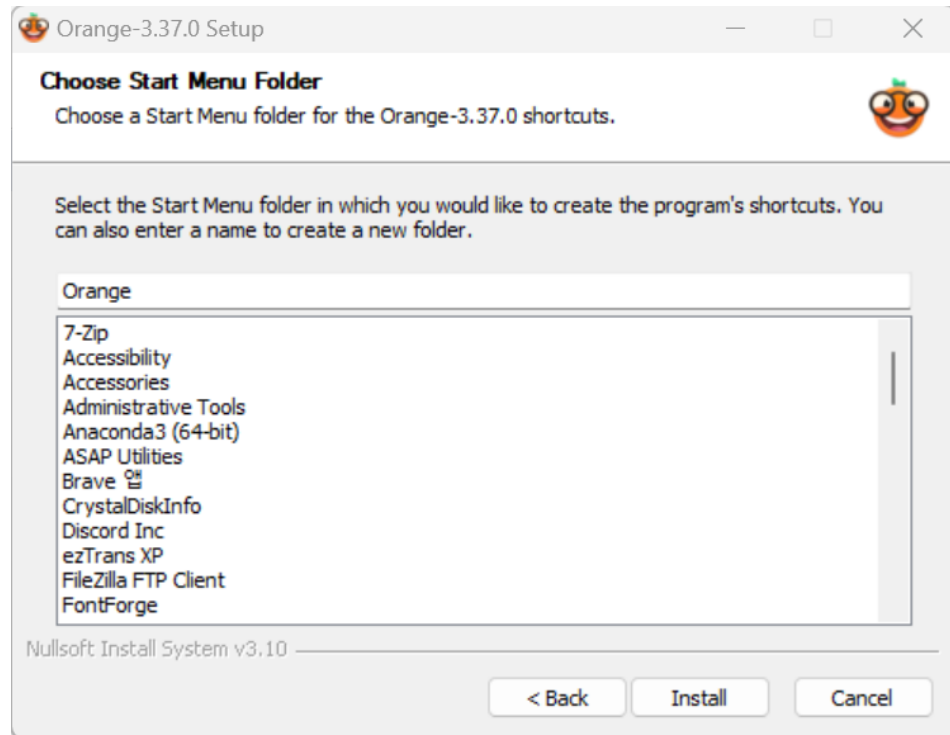
사용자 환경에 따라 적합한 ORANGE 버전 설치



# 설치하기

- 설치파일을 실행하여 안내되는 절차에 따라 설치를 진행

## > 설치파일 실행 5



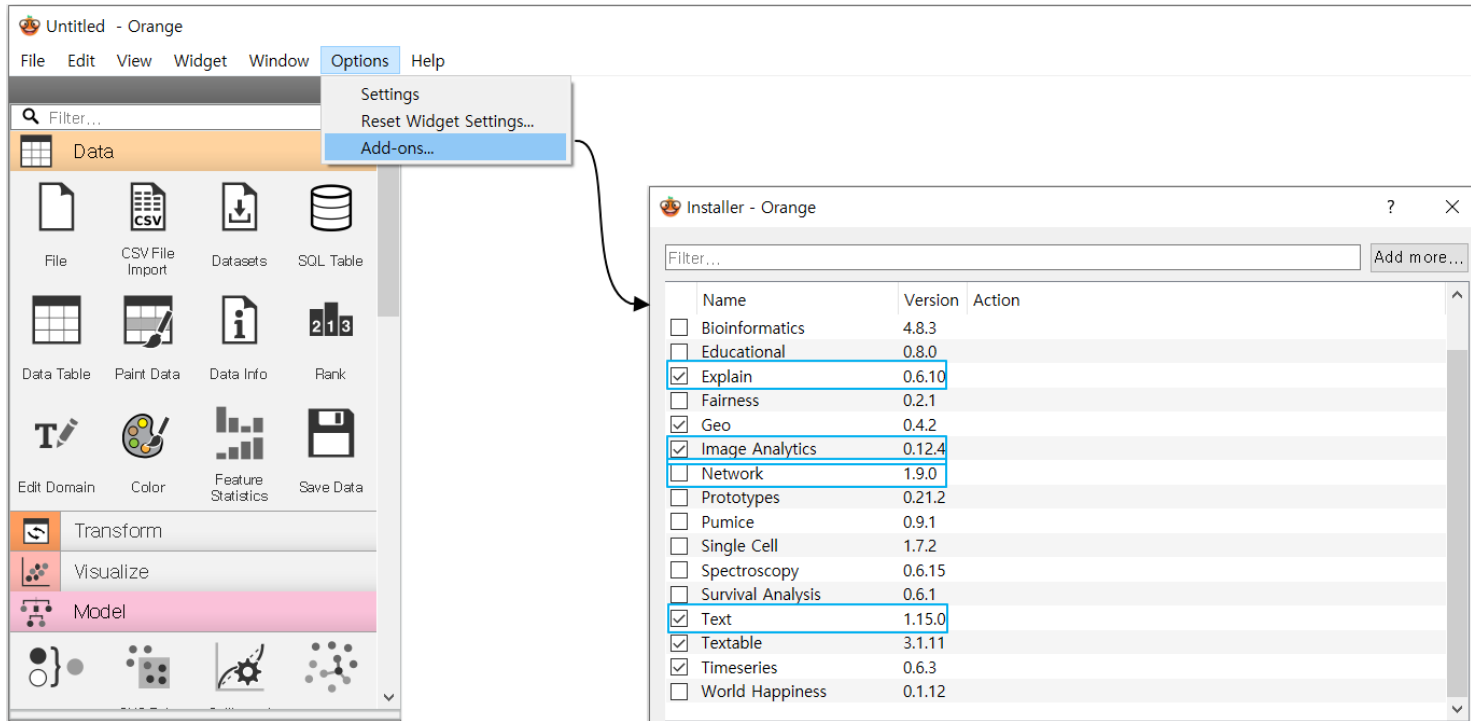
### Key Point

사용자 환경에 따라 적합한 ORANGE 버전 설치

# 설치하기

- Add-ons 설치로 추가 분석 기능 설치. Add-ons는 “관리자 권한으로 실행”으로 ORANGE를 실행해야 함
- Image Analytics, Text, Network, Explain 등 4가지는 고급 분석을 위해 필수 추가 설치
- 설치시간이 상당히 오래 걸리기 때문에 미리 설치하는 것이 좋음

## > Add-on 기능 설치



## > 이미지 관련 설명 (필요시)

- Explain  
: 분류/회귀 모형의 변수별 예측 기여도 설명 기능
- Image Analytics  
: 이미지 분석을 위한 불러오기, 벡터화 기능
- Networks  
: 데이터 기반의 네트워크 생성, 분석, 시각화, 검정 기능
- Text  
: 텍스트 분석을 위한 데이터 탑재 및 온라인 공개 데이터 불러오기, 전처리, 토픽 분석, 감성 분석, 시각화 기능

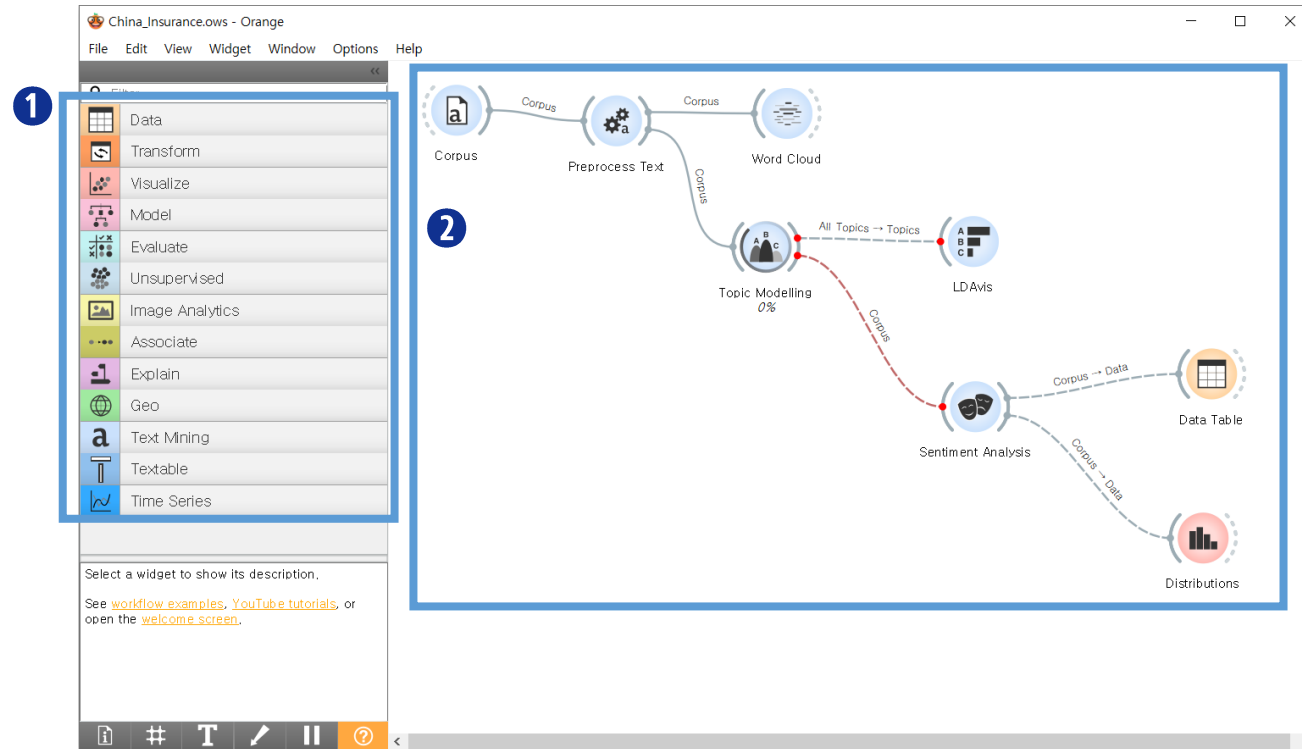
## Key Point

분석/학습 용도에 따른 Add-on 선택 및 설치

# 둘러 보기

- Add-ons 설치로 추가 분석 기능 설치. Add-ons는 “관리자 권한으로 실행”으로 ORANGE를 실행해야 함
- Image Analytics, Text, Network, Explain 등 4가지는 고급 분석을 위해 필수 추가 설치
- 설치시간이 상당히 오래 걸리기 때문에 미리 설치하는 것이 좋음

## > ORANGE 분석 플로우 화면 구성



## > 이미지 관련 설명 (필요시)

- ❶ 위젯 모듈
  - 데이터 분석에 필요한 위젯들을 분류별로 모아둔 곳.
  - 기본 모듈에서 추가(Add ons)를 하면 Image Analytics 부터 추가 설치된 위젯 모듈이 생성됨
- ❷ 작업판
  - 위젯을 이용하여 데이터 불러오기 > 전처리 > 탐색 > 모델링 > 성능 및 시각화 등을 드래그 & 드롭 방식으로 하나의 프로세스를 도식화하여 분석할 수 있음
  - 추후 작업에도 용이하고 쉽게 분석을 수행할 수 있음

## Key Point

분석 플로우 이미지를 바탕으로 한 분석 및 모형 생성

## 2. ORANGE 기능과 위젯 둘러보기

---

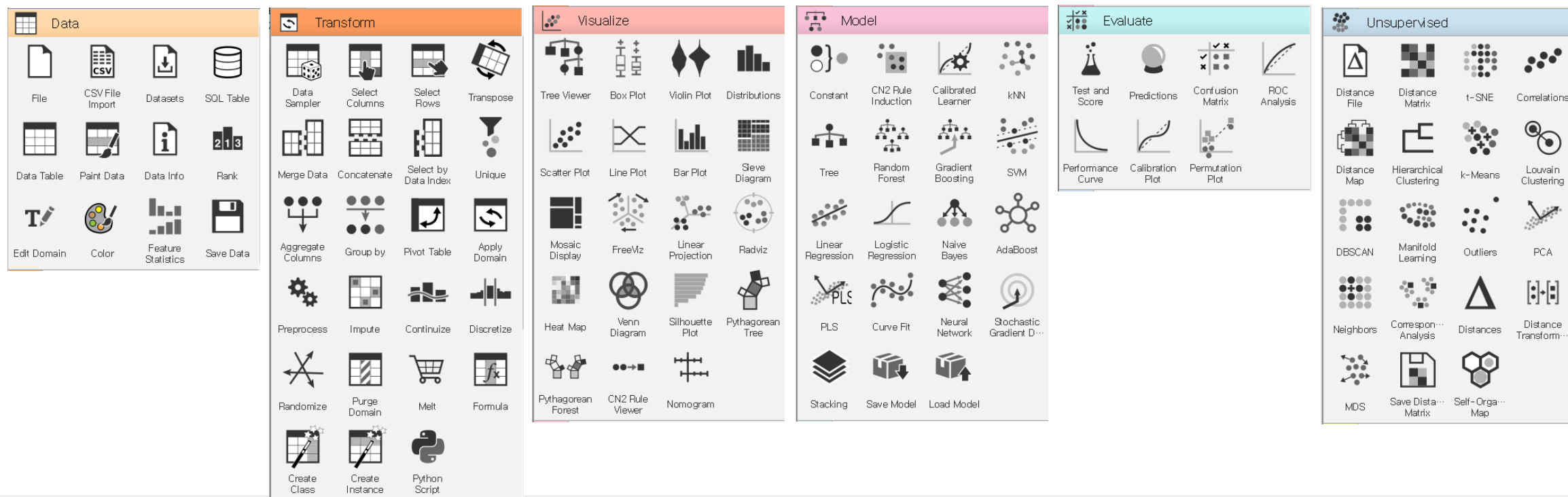




# 기본 위젯

- 기본 위젯은 다음과 같음

## ▶ 기본 위젯의 종류



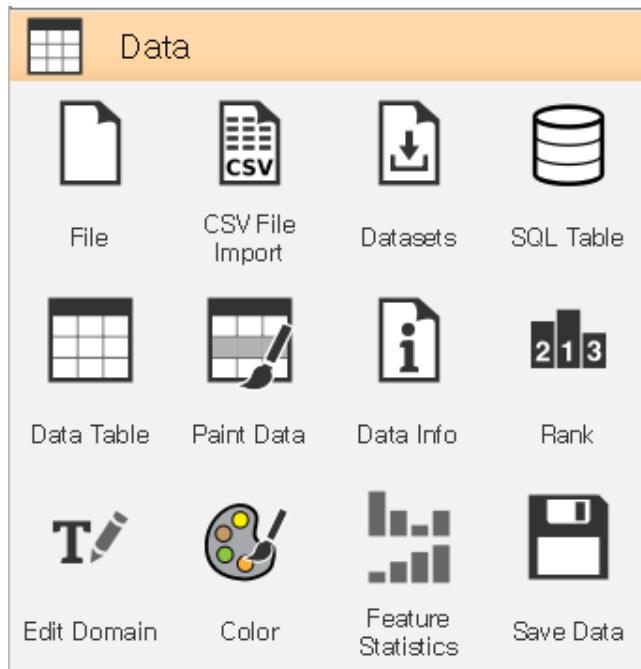
## Key Point

기본 위젯의 기능 및 용도에 대한 이해

## 위젯 설명

- 기본 설치 위젯 중 Data에 있는 세부 위젯은 다음과 같음

### > Data Widget



### > Widget 설명

- File : 학습에 사용할 데이터 불러오기 (로컬 파일, 온라인 데이터 포함)
- CSV File Import : CSV 형식 데이터 불러오기
- Datasets : ORAGNE에서 자체 제공하는 용도별 데이터 불러오기
- SQL Table : SQL 형식 네트워크 데이터 불러오기
- Data Table : 불러온 데이터를 표 형식으로 열람
- Paint Data : 2차원 데이터 분포도 작성 및 편집
- Data Info : 데이터의 수량, 데이터의 자료 형식, 결측치 등 데이터 특성 확인
- Rank : 양적 자료에 대한 순위점수 계산
- Edit Domain : 이산형 자료의 명명 및 값 변환
- Color : 데이터 변수를 구분하는 색상 정의
- Feature Statistics : 변수별 분포, 평균, 중위값, 최빈값, 최소/최대값 등 기술통계 특성 확인
- Save Data : 탑재된 데이터를 로컬 파일로 다운로드

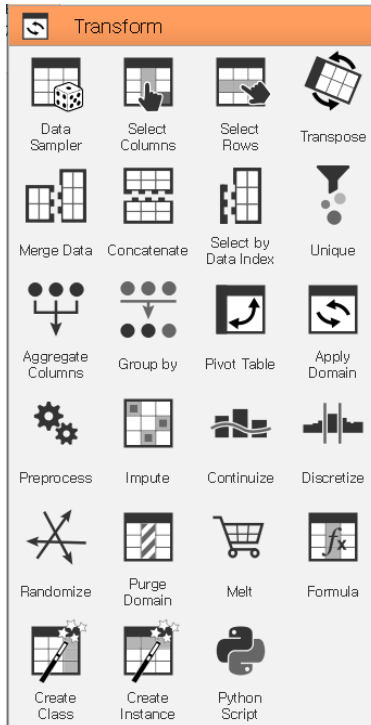
### Key Point

Data 위젯은 자료를 불러오거나 기초적인 특성을 파악하는 데에 사용

## 위젯 설명

- 기본 설치 위젯 중 Transform에 있는 세부 위젯은 다음과 같음

### ▶ Transform Widget



### ▶ Widget 설명

- Data Sampler : 데이터 케이스 일부를 무작위 추출
- Select Columns : 변수별 자료 특성 정의, 사용할 변수 지정
- Select Rows : 케이스 사용 조건 설정
- Transpose : 데이터 행/열 변경
- Merge Data : 데이터를 수평으로 병합 (변수 추가)
- Concatenate : 데이터를 수직으로 병합 (케이스 추가)
- Select by Data Index : 인덱스를 기준으로 매칭 데이터 탐색
- Unique : 중복 데이터 제거
- Aggregate Columns : 데이터 열별 집계값 출력(합계, 평균, 분산 등)
- Group by : 그룹 정의 변수 선택 및 하위 그룹별 집계값 출력
- Pivot table : 행, 열, 값 내용을 정의하여 피벗 테이블 출력
- Apply Domain : 정의된 데이터 양식과 동일한 처리를 다른 데이터에도 적용
- Preprocess : 표준화, 결측값 대체, 구간화/연속화 등 데이터 전처리 방식 설정
- Impute : 결측값 대체 규칙 설정
- Continuize : 범주형 자료를 연속형 자료로 변환
- Discretize : 연속형 자료를 범주형 자료로 구간화
- Randomize : 데이터 행을 무작위 배열
- Purge Domain : 사용하지 않을 자료 및 속성 정보 제거
- Melt : 데이터를 id, 변수명, 값의 3열 자료로 변경
- Formula : 입력된 계산식 규칙에 의해 새로운 변수 생성
- Create Class : 이산형 또는 문자열 자료에 클래스 속성 생성
- Create Instance : 평균, 중앙값 등 입력한 규칙에 따라 무작위 데이터 생성
- Python Script : Python 스크립트로 규칙을 입력하여 데이터 편집

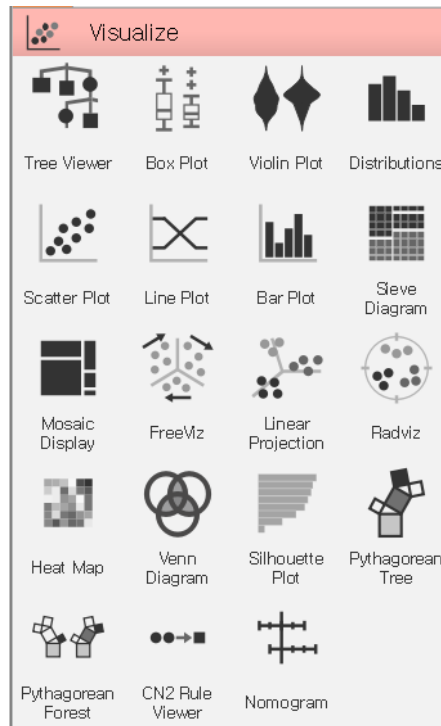
### Key Point

Transform 위젯은 분석에 필요한 형태로 자료를 가공하거나 전처리하는 데에 사용

## 위젯 설명

- 기본 설치 위젯 중 Visualize에 있는 세부 위젯은 다음과 같음

### > Visualize Widget



### > Widget 설명

- Tree Viewer : 의사결정나무 분류 결과 시각화
- Box Plot : 자료 범위 및 이상값이 포함된 상자수염 그래프
- Violin Plot : 자료 값 구간별 빈도가 포함된 피라미드형 도식
- Distributions : 히스토그램 그래프
- Scatter Plot : 산점도 그래프
- Line Plot : 꺾은선 그래프
- Bar Plot : 막대 그래프
- Sieve Diagram : 독립적인 변수 간의 예측 빈도와 실제 빈도를 비교하는 시각화
- Mosaic Display : 정사각형을 2개 변수의 빈도로 분할하는 교차표 시각화
- FreeViz : 케이스 속성별 벡터량을 포함하는 시각화 그래프
- Line Projection : 지정된 클래스 레이블별 축을 도입하는 시각화 그래프
- Radviz : 자료 속성별 원형 공간을 배치하는 시각화 그래프
- Heat Map : 숫자형 2개 변수의 행렬을 시각화하는 그래프
- Venn Diagram : 기준점에 대한 공통 빈도를 탐색하는 벤 다이어그램
- Silhouette Plot : 케이스와 집단 내 다른 케이스들의 유사도를 평가하는 시각화
- Pythagorean Tree : 의사결정나무 분류 결과 시각화
- Pythagorean Forest : 랜덤 포레스트 분류 결과 시각화
- CN2 Rule Viewer : 데이터에서 규칙을 추출하는 분류 알고리즘 결과
- Nomogram : 특성에 따른 예측 목표변인의 예상 결과값을 시각화

### Key Point

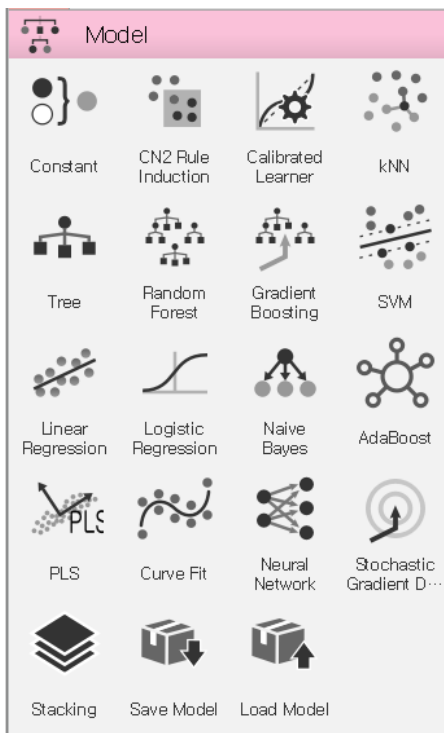
Visualize 위젯은 분석 결과를 이해하기 쉽도록 시각화하는 기능으로 구성



## 위젯 설명

- 기본 설치 위젯 중 Model에 있는 세부 위젯은 다음과 같음

### > Model Widget



### > Widget 설명

- Constant : 분류 또는 회귀모형에 따른 예측값 생성
- CN2 Rule Induction : 분류 예측을 위한 알고리즘
- Calibrated Learner : 이진 분류 예측을 위한 알고리즘
- kNN : 최근접 이웃 알고리즘
- Tree : 의사결정나무 알고리즘
- Random Forest : 의사결정나무 모형을 반복 생성하여 가중치를 평균
- Gradient Boosting : 의사결정나무 모형을 반복 생성하여 순차적으로 가중치를 조정
- SVM : 초평면으로 분리되는 서포트 벡터 거리를 최대화하는 분류 알고리즘
- Linear Regression : 선형회귀모형을 통한 연속변수 예측 알고리즘
- Logistic Regression : 로지스틱 회귀모형 기반의 이산형 목표변수 예측 알고리즘
- Naive Bayes : 변수 간 독립을 가정하는 나이브 베이즈 모형을 활용한 분류 알고리즘
- AdaBoost : 약한 분류기를 조합해 순차적으로 분류 성능을 강화하는 AdaBoost 알고리즘
- PLS : 편최소제곱 회귀모형을 활용하는 예측 알고리즘
- Curve Fit : 입력 데이터와 함수의 유사도를 측정
- Neural Network : 다층 퍼셉트론 기반의 노드 간 가중치를 조정하는 인공신경망 예측 모형
- Stochastic Gradient Descent : 모형의 성능 최적화를 위한 반복 방법
- Stacking : 여러 모형의 예측값을 최종 모형의 학습 데이터로 사용하는 예측 알고리즘
- Save Model : 형성한 모형을 파일로 저장
- Load Model : 모형 파일을 ORANGE에 불러옴

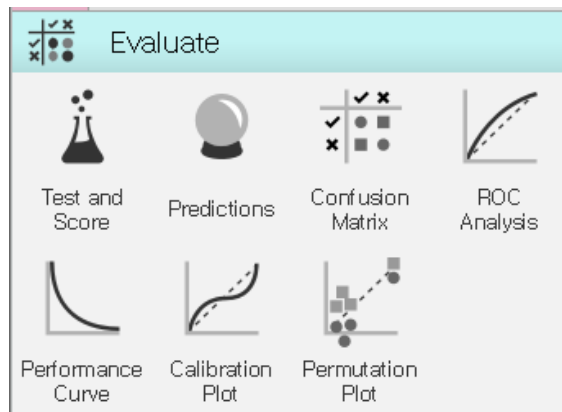
### Key Point

Model 위젯은 자료 특성에 따라 적합한 학습 방법을 적용하는 지도학습 알고리즘으로 구성

## 위젯 설명

- 기본 설치 위젯 중 Evaluate에 있는 세부 위젯은 다음과 같음

### > Evaluate Widget



### > Widget 설명

- Test and Score : 학습/테스트 데이터를 통해 모델 정확성 평가치 출력
- Predictions : 예측 모델에 따라 목표 변수의 예측치 생성
- Confusion Matrix : 테스트 데이터 기준의 혼동 행렬로 모델 정확성 확인
- ROC Analysis : 민감도와 특이도를 기준으로 하는 ROC 곡선 기반으로 모델 정확성 확인
- Performance Curve : 임계값 이상 사례의 예측빈도와 실제 빈도 비교
- Calibration Plot : 예측 모델과 실제 결과를 그래프로 시각화하여 비교
- Permutation Plot : 예측 모델의 적합도와 과적합 수준 측정

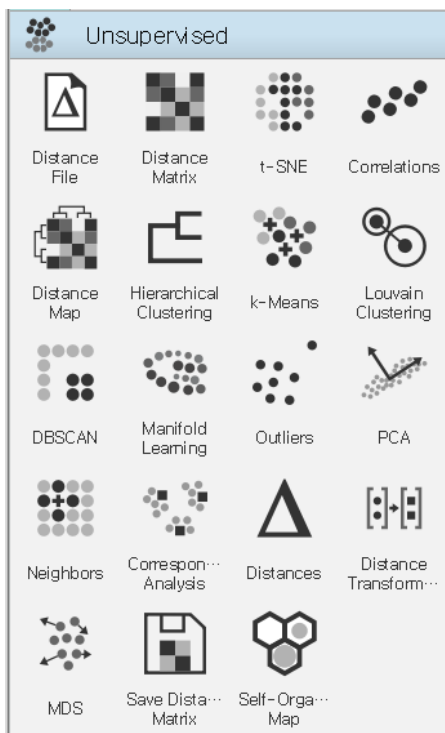
### Key Point

Evaluate 위젯은 생성된 모형의 적합성을 평가하는 데에 사용

## 위젯 설명

- 기본 설치 위젯 중 Unsupervised에 있는 세부 위젯은 다음과 같음

### > Unsupervised Widget



### > Widget 설명

- Distance File : Distance Matrix가 포함된 자료 파일 불러오기
- Distance Matrix : 데이터를 활용하여 2차원 거리 행렬 생성
- T-SNE : 다차원 데이터를 평면 차원으로 축소하여 표현하는 시각화
- Correlations : 변수 간 상관계수 산출
- Distance Map : 투입한 개체 간의 거리를 시각화
- Hierarchical Clustering : 유사 특성을 가진 데이터 간 위계적 군집화
- K-Means : 데이터를 k 개의 군집으로 묶는 알고리즘
- Louvain Clustering : 군집 간 중첩이 발생하지 않도록 하는 군집 알고리즘
- DBSCAN : 다차원 데이터를 밀도 기반의 거리를 최소화하는 군집 알고리즘
- Manifold Learning : 고차원 데이터를 최적으로 표현하는 2차원 Manifold를 탐색
- Outliers : 기준에 따른 이상값 점수 평가
- PCA : 상관계수 기반의 차원축소
- Neighbors : 여러 레이블 간의 거리가 가장 가까운 케이스를 이용해 속성을 예측하는 알고리즘
- Correspondence Analysis : 다변량 범주형 자료 간의 관계를 시각화
- Distances : 자료 내 행과 열 간의 거리를 측정
- Distance Transformation : 자료 간 거리 행렬을 표준화
- MDS : 다차원 공간상의 객체 거리를 보존하는 차원축소 좌표계 탐색
- Save Distance Matrix : 생성 및 편집된 거리 행렬을 파일로 저장
- Self-Organizing Map : 저차원 격자 형태의 행렬에 사례를 대응시키는 인공신경망 기반 군집화

### Key Point

Unsupervised 위젯은 데이터 학습 모델을 제시하지 않는 비지도 학습을 위한 기능으로 구성

### 3. 데이터 연동과 분석 스토리 설계하기

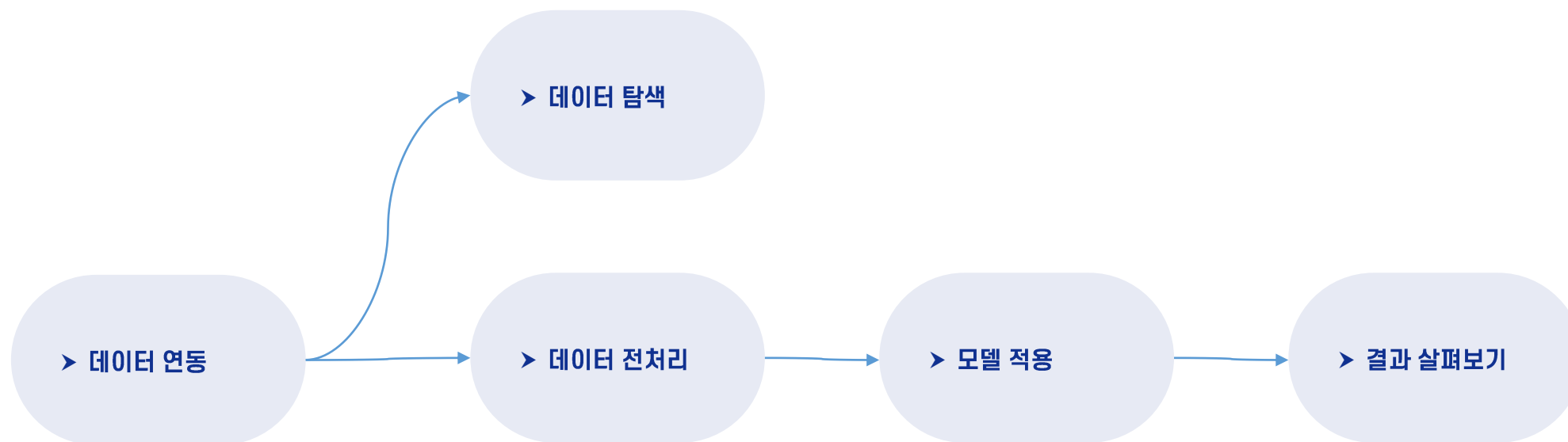
---



## 분석 스토리

- 분석 스토리는 다음과 같은 프로세스를 일반적으로 가짐
- 데이터 연동 – 데이터 전처리 – 데이터 탐색 – 모델 적용 – 결과 살펴보기

### ➤ 분석 스토리(분석 플로우) 개괄



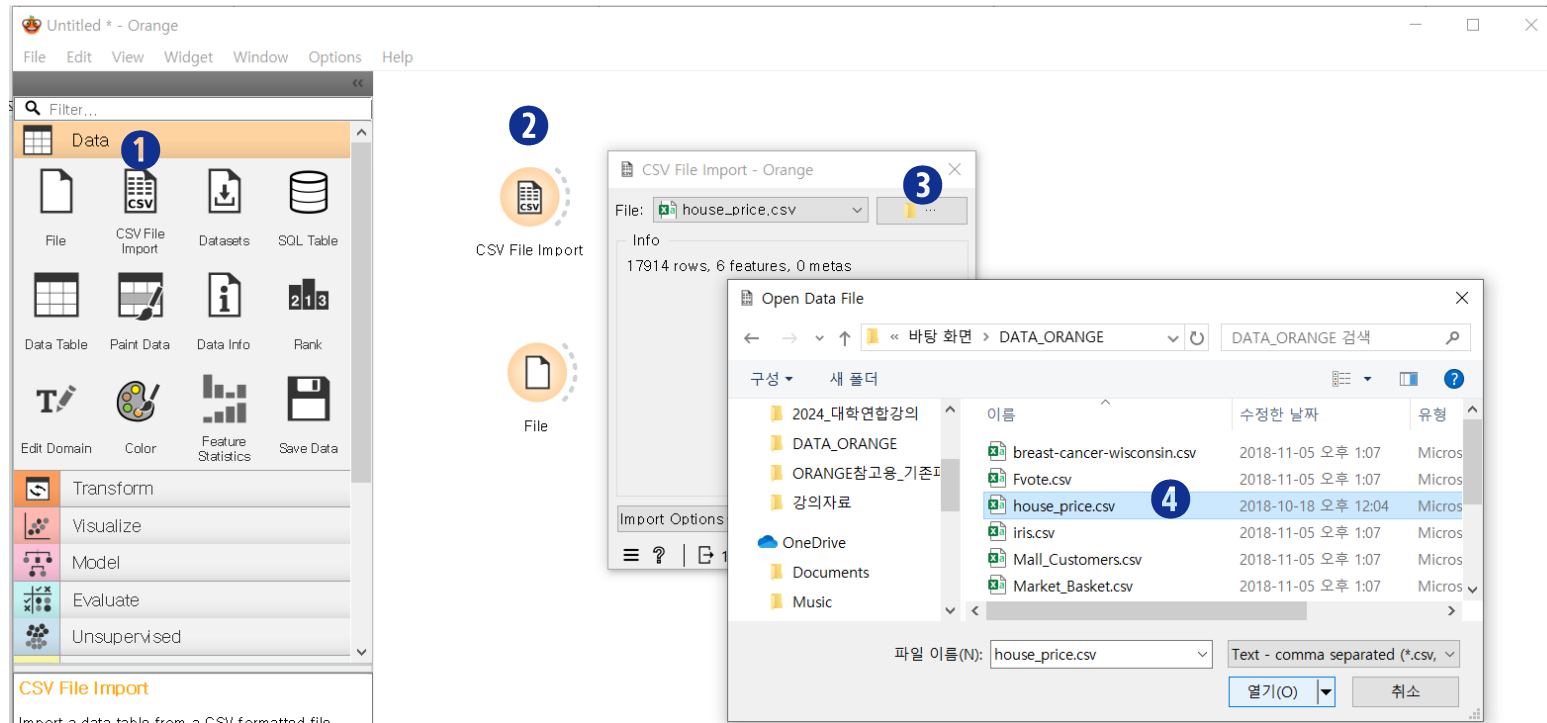
### Key Point

ORANGE의 기본 분석 FLOW

## 데이터 불러오기

- 다양한 데이터 소스를 불러올 수 있음
- Csv파일은 csv File Import로 불러오는 것이 용량이 적고 좋음. 그 외 파일은 File에서 불러올 수 있음
- Csv 파일은 Data > csv File Import (작업창에 위젯 생성) > 더블 클릭 후 폴더에서 원하는 파일 선택 후 열기를 누르면 됨

### > 데이터 불러오기1. csv File Import



### > 작업 순서

- ❶ Data에서 csv File Import 위젯을 선택하면 작업창에 위젯이 생성
- ❷ 생성된 csv File Import 위젯을 더블 클릭하면 파일 불러오기 창이 뜬
- ❸ 폴더 선택에서 원하는 csv 파일이 있는 경로를 선택
- ❹ 여기서는 house\_price.csv 파일을 선택하여 열기를 누름

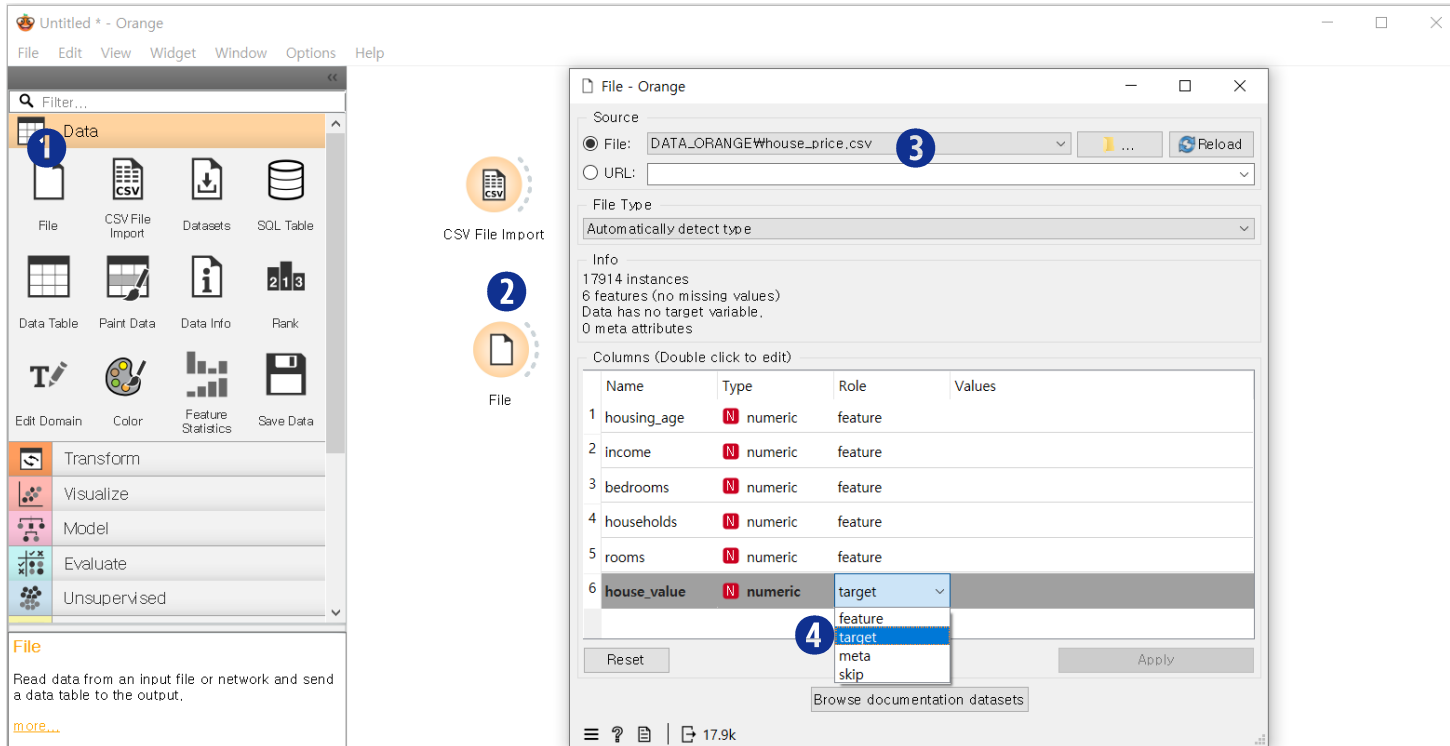
### Key Point

CSV File Import 기능을 활용하여 데이터 소스를 불러옴

# 데이터 불러오기

- File 기능은 엑셀 등 다양한 형태의 데이터셋을 불러올 수 있으며, 또한 url 등 웹상의 링크(구글시트 등) 데이터도 연동 가능
- 또한 데이터 중 Feature변수와 Target 변수를 설정하여 다음 모델링 작업이 편리함

## > 데이터 불러오기2. File



## > 작업 순서

- ❶ Data에서 File 위젯을 선택하면 작업창에 위젯이 생성
- ❷ 생성된 File 위젯을 더블 클릭하면 파일 연동 창이 뜬
- ❸ 폴더 선택에서 원하는 파일이 있는 경로 및 파일을 선택
- ❹ 파일 선택 후 변수들 중 feature와 target을 설정할 수 있음 (필수는 아님)

### Key Point

File 위젯을 활용하여 다양한 형식의 데이터 파일을 불러오거나 속성을 정의



# II

PART

## CASE분석1: ORANGE로 지도학습 해보기

---

1. ORANGE의 지도학습 알고리즘 소개
2. 회귀의 지도학습 시나리오 설계와 실습
3. 분류의 지도학습 시나리오 설계와 실습

## II CASE분석1: ORANGE로 지도학습 해보기

# 1. ORANGE의 지도학습 알고리즘 소개

---



# 개념 설명

- 머신러닝은 크게 지도학습과 비지도학습으로 구분됨
- 지도학습: 분류나 예측의 목적이 명확함
- 비지도학습: 경향과 관계성을 파악하는 목적

## ▶ 지도학습과 비지도학습 비교

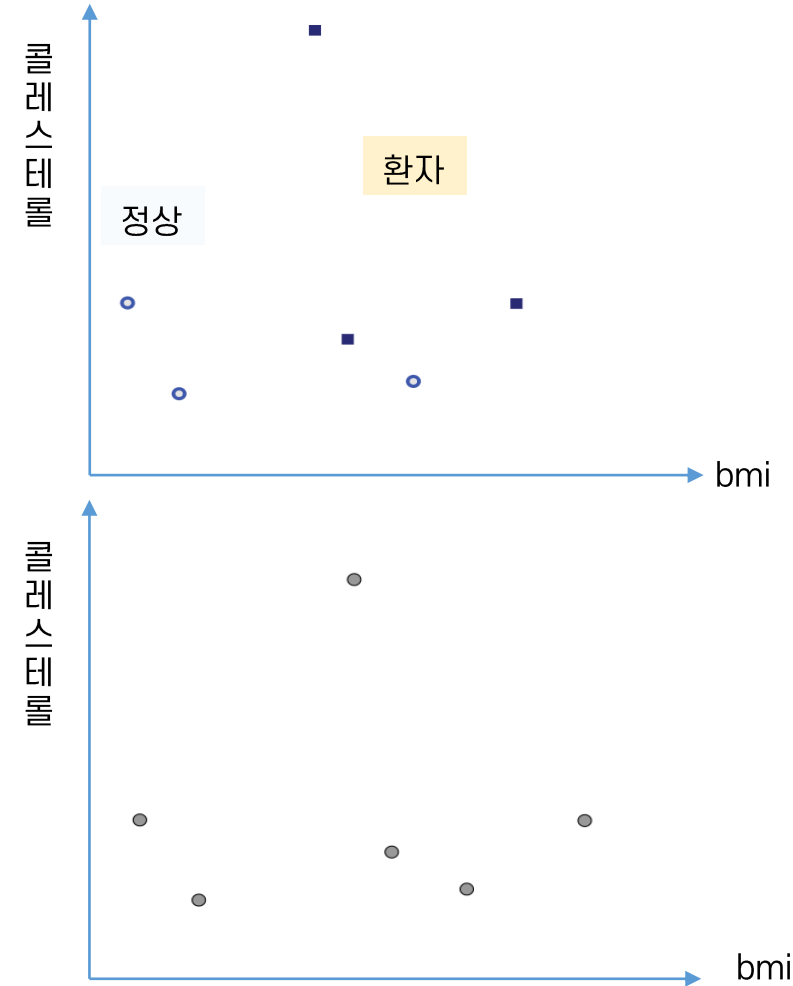
- 특성(feature)
- 독립변수  
(independent variable)
- 레이블(label)
- 종속변수  
(dependent variable)

*Supervised Learning*

bmi	가족력	콜레스테롤	.....	환자 여부
23.1	있음	113.4	.....	환자
18.4	없음	123.4	.....	정상
22.4	있음	198.4	.....	환자
19.5	없음	98.4	.....	정상
26.7	있음	123.2	.....	환자
24.5	없음	101.8	.....	정상

*Unsupervised Learning*

bmi	가족력	콜레스테롤	.....
23.1	있음	113.4	.....
18.4	없음	123.4	.....
22.4	있음	198.4	.....
19.5	없음	98.4	.....
26.7	있음	123.2	.....
24.5	없음	101.8	.....

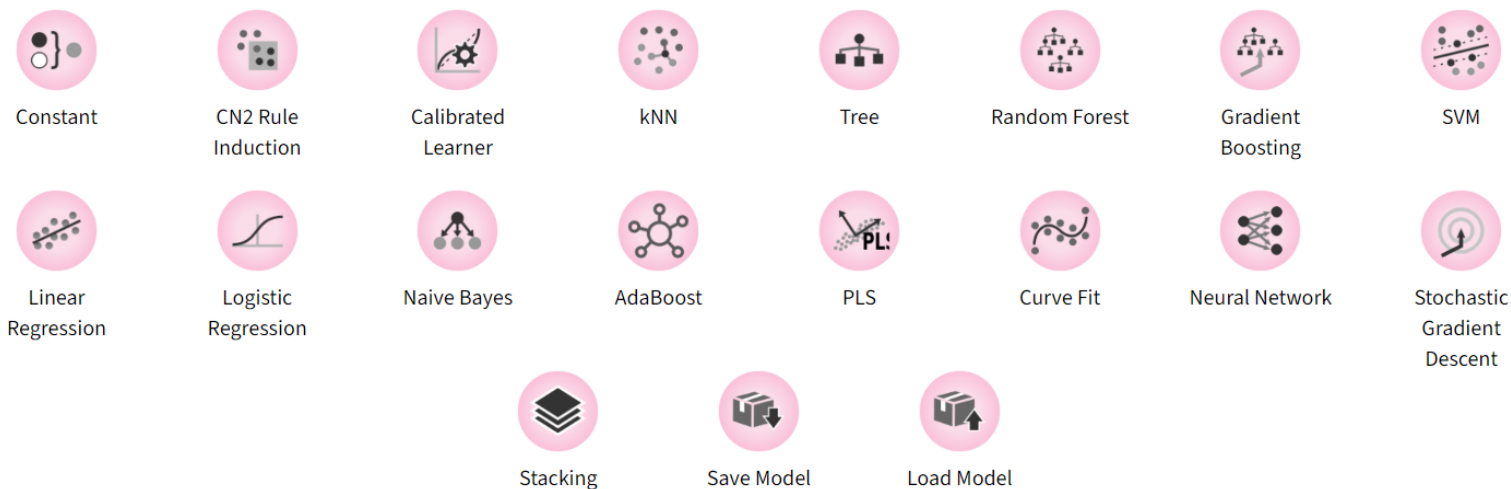


## 개념 설명

- Orange에서는 지도학습이 Model이라는 영역에 있으며, 총 12개의 알고리즘을 제공함

### > Orange의 지도학습 알고리즘

#### Model



### > Orange의 지도학습 알고리즘 종류

- 1. KNN
- 2. Tree
- 3. Random Forest
- 4. Gradient Boosting
- 5. SVM
- 6. Linear Regression
- 7. Logistic Regression
- 8. Naïve Bayes
- 9. AdaBoost
- 10. Neural Network
- 11. Stochastic Gradient Descent
- 12. Stacking

#### Key Point

12개의 알고리즘 제공으로 분석하고자 하는 알고리즘 대부분 사용 가능

## 개념 설명

- Orange에서는 지도학습을 아래와 같이 재분류할 수 있음
- 단일 모델, 앙상블모델, 딥러닝 모델로 구분하는 것이 적합함

### > Orange 지도학습 알고리즘 분류

분류	알고리즘	개요
회귀	Linear Regression	타겟(라벨)이 수치(예: 집값, 주식 등). 응용모델로 릿지, 라소, 엘라스틱모델 등이 있음
분류	Logistic Regression	타겟(라벨)이 범주(예: 합격여부, 환자여부 등). 로지스틱 회귀모델은 분류 중 기저모델(base model)
분류+회귀 (단일모델)	KNN	이웃한 유사 데이터로 예측 혹은 분류를 수행
	Naïve Bayes	발생확률(조건부 확률)로 활용. 회귀보다는 분류에 적합
	SVM	서포트벡터머신. 모델 정확도가 높아 실전에서 많이 활용됨
	Tree	예측/분류 기준을 나무의 가지치기처럼 나누어 진행하여 Tree분석이라고 명명함. 시각화에 좋은 모델
분류+회귀 (앙상블)	Random Forest	Tree를 수 백 개 수행하면서 일반화를 진행함. 정확도가 높음
	Gradient Boosting	부스트(boost) 방법 중 하나. 표본을 무작위 재추출하면서 일반화의 성능을 높임
	AdaBoost	부스트(boost) 방법 중 하나.
	Stacking	2개 이상의 모델과 1개의 최종 모델로 예측을 수행.
딥러닝	Stochastic Gradient Descent	오차/손실을 최소화 하면서 최적 결과를 찾는 방법. 딥러닝 알고리즘의 초기 핵심 방법론
	Neural Network	신경망분석. 이에 반복학습+초기가중치설정+다양한 함수를 추가하여 딥러닝으로 발전

### Key Point

분석 목적에 해당하는 분석 모델을 사용하는 것이 중요

## II CASE분석1: ORANGE로 지도학습 해보기

# 2. 회귀의 지도학습 시나리오 설계와 실습

---



## 2. 회귀의 지도학습 시나리오 설계와 실험

회귀  
모델

- 학습데이터: house\_price\_new.csv (8개 변수, 17,627개의 데이터)
- 예측데이터: house\_price\_predict.csv (7개 변수, 9개의 데이터)
- 작업파일: regression.ows

## ▶ 학습 데이터(house\_price\_new.csv)

	A	B	C	D	E	F	G	H
1	housing_age	income	bedrooms	households	rooms	house_value	bed_per_rooms	room_per_household
2	23	6.777	0.141112	2.44224	8.10396	500000	0.0174127	3.31824
3	49	6.0199	0.160984	2.72669	5.75241	500000	0.0279854	2.10967
4	35	5.1155	0.249061	1.90268	3.88808	500000	0.0640577	2.04348
5	32	4.7109	0.231383	1.91367	4.50839	500000	0.0513227	2.35589
6	21	4.5625	0.255583	3.09266	4.66795	500000	0.0547527	1.50936
7	38	4.3403	0.26835	1.768	5.068	500000	0.05295	2.86652
8	34	3.7306	0.272993	1.50727	3.8643	500000	0.070645	2.56377
9	29	3.625	0.258112	1.62673	4.21352	500000	0.061258	2.59018
10	29	3.6121	0.27736	1.71926	4.00696	500000	0.0692194	2.33063
11	31	3.5744	0.251751	2.09535	4.16312	500000	0.0604717	1.98684
12	25	3.5556	0.258405	1.86006	4.02201	500000	0.0642477	2.1623
13	44	2.6103	0.29553	2.0303	4.06734	500000	0.0726592	2.00332
14	26	2.3536	0.353787	3.97344	2.82656	500000	0.125165	0.711365
15	20	2.2444	0.375408	2.35809	2.97701	500000	0.126102	1.26247
16	26	1.9891	0.27771	1.90279	4.6067	500000	0.060284	2.42102
17	19	1.2656	0.304878	1.43353	3.79191	500000	0.0804023	2.64516
18	28	6.7861	0.146739	2.25182	7.38686	499100	0.0198649	3.28039
19	18	8.1489	0.151703	2.22616	6.60082	499000	0.0229824	2.96512
20	29	8.248	0.138303	3.00364	7.07273	498800	0.0195545	2.35472
21	40	7.2692	0.156981	2.19082	6.81643	498700	0.0230298	3.11136
22	33	5.738	0.183805	1.98462	5.98462	497600	0.0307129	3.0155
23	40	3.5637	0.246415	1.71223	4.41439	497400	0.0558208	2.57815
24	21	7.9135	0.14018	2.44571	7.92857	496400	0.0176804	3.24182
25	33	6.597	0.135957	2.6318	7.01674	496400	0.0193761	2.66614
26	17	5.9285	0.160832	2.34204	6.51306	496000	0.0246937	2.78093
27	44	6.7058	0.156519	2.5891	6.33543	495900	0.0247053	2.44696
28	36	8.0499	0.171736	1.96106	6.08053	495600	0.0282436	3.10063
29	38	3.6181	0.2596	1.66515	4.05187	495600	0.0640693	2.43333
30	33	5.9683	0.172631	2.63942	6.11298	495500	0.02824	2.31603

## ▶ 학습 데이터 설명

- 8개 변수, 17,627개 데이터
- housing\_age: 주택 연식
- income: 소득
- bedrooms: 침실
- households: 가구 수
- rooms: 방 개수
- house\_value: 집 가격
- bed\_per\_rooms: 방 1개당 침대 수
- room\_per\_household: 가구 1개당 방 개수

## Key Point

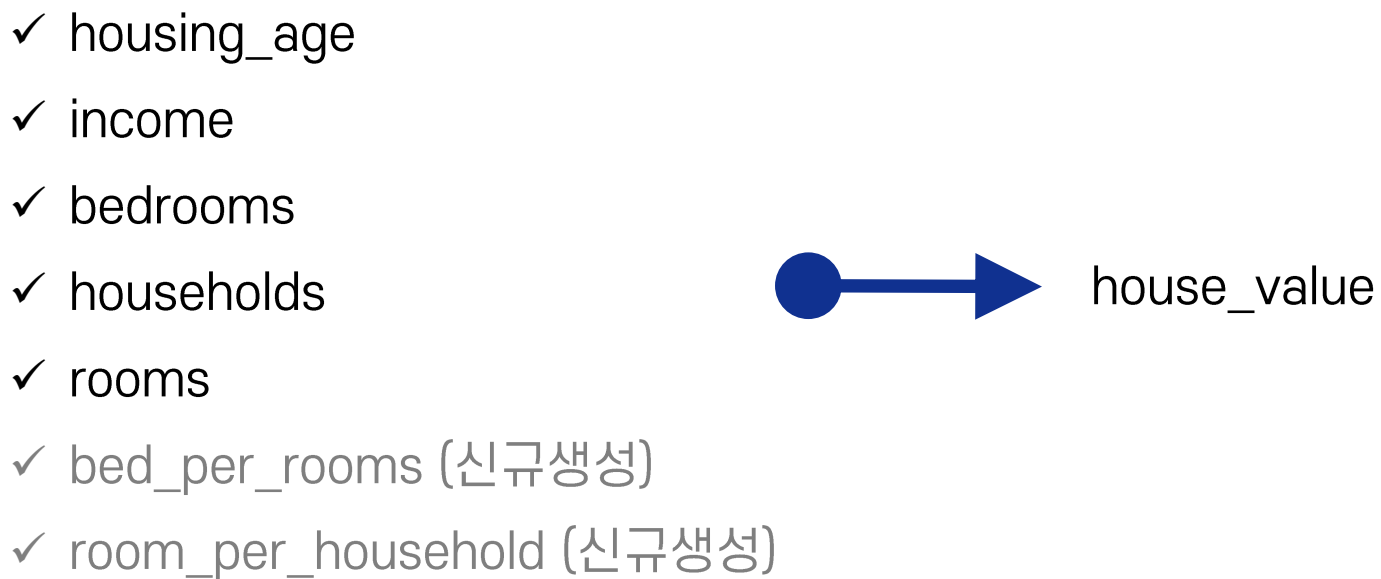
집값에 영향을 주는 주요한 변수는 무엇인가?



## 회귀 모델

- 학습데이터: house\_price\_new.csv (8개 변수, 17,627개의 데이터)
- 예측데이터: house\_price\_predict.csv (7개 변수, 9개의 데이터)
- 작업파일: regression.ows

### ➤ 분석 모형



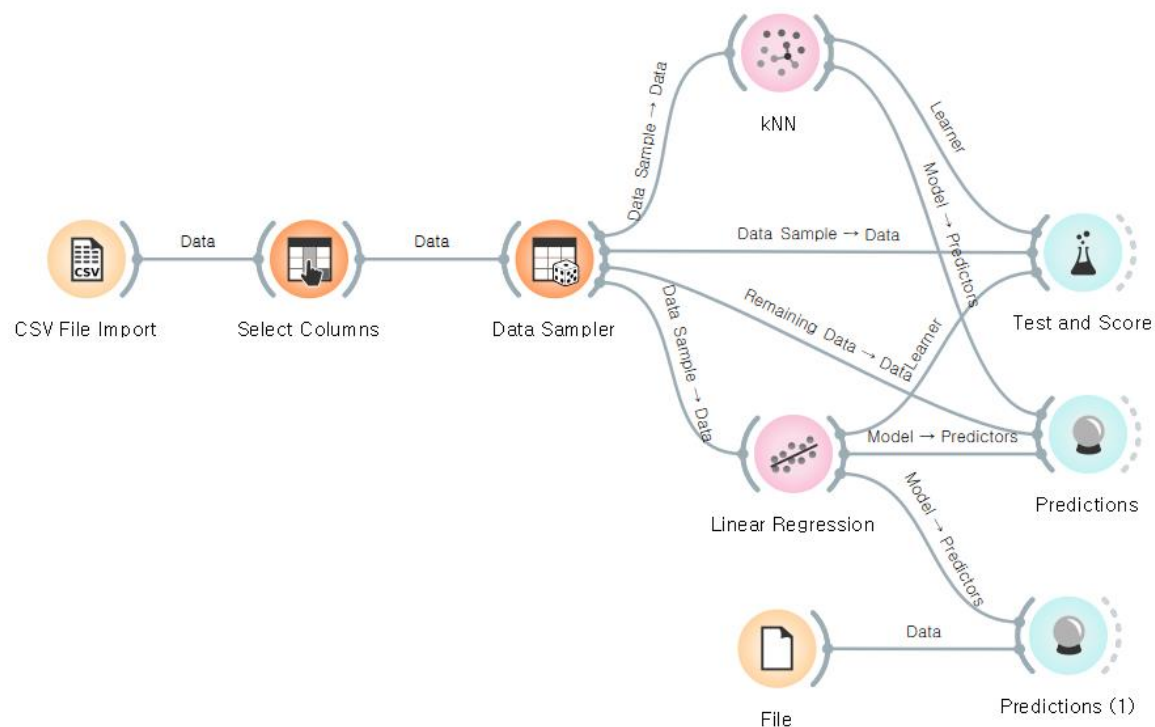
### Key Point

“house\_value”는 Target 변수, 그 외 7개의 변수는 Feature 변수

## 회귀 모델

- 학습데이터: house\_price\_new.csv (8개 변수, 17,627개의 데이터)
- 예측데이터: house\_price\_predict.csv (7개 변수, 9개의 데이터)
- 작업파일: regression.ows

### ▶ 분석 전체 과정



### Key Point

데이터 불러오기부터 모델 설정, 모델 성능 확인까지 순차적으로 진행

## II CASE분석1: ORANGE로 지도학습 해보기

### 3. 분류의 지도학습 시나리오 설계와 실습

---



## 3. 분류의 지도학습 시나리오 설계와 실습

분류  
모델

- 학습데이터: breast\_cancer.csv(10개 변수, 683개 데이터)
- 예측데이터: breast\_cancer\_predict.csv(9개 변수, 10개 데이터)
- 작업파일: classification.ows

## ➤ 학습 데이터(breast\_cancer.csv)

	A	B	C	D	E	F	G	H	I	J
1	Clump_Thickness	Cell_Size	Cell_Shape	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	Class
2	5	1	1	1	2	1	3	1	1	0
3	5	4	4	5	7	10	3	2	1	0
4	3	1	1	1	2	2	3	1	1	0
5	6	8	8	1	3	4	3	7	1	0
6	4	1	1	3	2	1	3	1	1	0
7	8	10	10	8	7	10	9	7	1	1
8	1	1	1	1	2	10	3	1	1	0
9	2	1	2	1	2	1	3	1	1	0
10	2	1	1	1	2	1	1	1	5	0
11	4	2	1	1	2	1	2	1	1	0
12	1	1	1	1	1	1	3	1	1	0
13	2	1	1	1	2	1	2	1	1	0
14	5	3	3	3	2	3	4	4	1	1
15	1	1	1	1	2	3	3	1	1	0
16	8	7	5	10	7	9	5	5	4	1
17	7	4	6	4	6	1	4	3	1	1
18	4	1	1	1	2	1	2	1	1	0
19	4	1	1	1	2	1	3	1	1	0
20	10	7	7	6	4	10	4	1	2	1
21	6	1	1	1	2	1	3	1	1	0
22	7	3	2	10	5	10	5	4	4	1
23	10	5	5	3	6	7	7	10	1	1
24	3	1	1	1	2	1	2	1	1	0
25	1	1	1	1	2	1	3	1	1	0
26	5	2	3	4	2	7	3	6	1	1
27	3	2	1	1	1	1	2	1	1	0
28	5	1	1	1	2	1	2	1	1	0
29	2	1	1	1	2	1	2	1	1	0

## ➤ 학습 데이터 설명

- 위스콘신 주 유방암 관련 데이터
- 10개 변수, 683개 데이터
- Clump\_Thickness: 뭉침 두께 정도
- Cell\_Size: 세포 크기의 균일도
- Cell\_Shape: 세포 모양의 균일도
- Marginal\_Adhesion: 밀착도
- Single\_Epithelial\_Cell\_Size: 단일 상피 세포 크기
- Bare\_Nuclei: 세포 핵
- Bland\_Chromatin: 염색질 건조도
- Normal\_Nucleoli: 핵소체 정상도
- Mitoses: 분열도
- Class: 유방암 유무(양성, 음성)

## Key Point

유방암 유무에 영향을 주는 주요한 변수는 무엇인가?

## 분류 모델

- 학습데이터: breast\_cancer.csv(10개 변수, 683개 데이터)
- 예측데이터: breast\_cancer\_predict.csv(9개 변수, 10개 데이터)
- 작업파일: classification.ows

### ▶ 분석 모형

- ✓ Clump\_Thickness
- ✓ Cell\_Size
- ✓ Cell\_Shape
- ✓ Marginal\_Adhesion
- ✓ Single\_Epithelial\_Cell\_Size
- ✓ Bare\_Nuclei
- ✓ Bland\_Chromatin
- ✓ Normal\_Nucleoli
- ✓ Mitoses



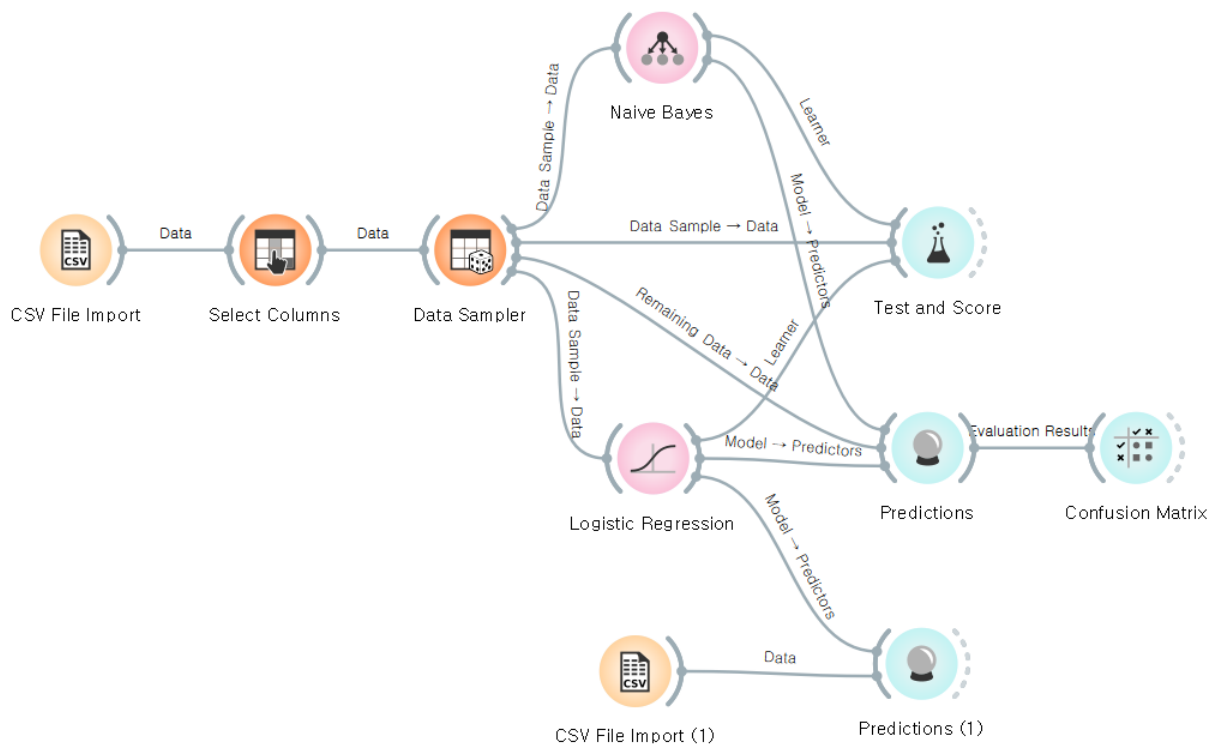
### Key Point

“Class”는 Target 변수, 그 외 9개의 변수는 Feature 변수

## 분류 모델

- 학습데이터: breast\_cancer.csv(10개 변수, 683개 데이터)
- 예측데이터: breast\_cancer\_predict.csv(9개 변수, 10개 데이터)
- 작업파일: classification.ows

### ▶ 분석 전체 과정



### Key Point

데이터 불러오기부터 모델 설정, 모델 성능 확인까지 순차적으로 진행

# III

PART

## CASE분석2: ORANGE로 비지도학습 해보기

---

1. ORANGE의 비지도학습 알고리즘 소개
2. 군집으로 유사한 집단 묶어보기



### III CASE분석2: ORANGE로 비지도학습 해보기

## 1. ORANGE의 비지도학습 알고리즘 소개

---



# 개념 설명

- 머신러닝은 크게 지도학습과 비지도학습으로 구분됨
- 지도학습: 분류나 예측의 목적이 명확함
- 비지도학습: 경향과 관계성을 파악하는 목적

## ▶ 지도학습과 비지도학습 비교

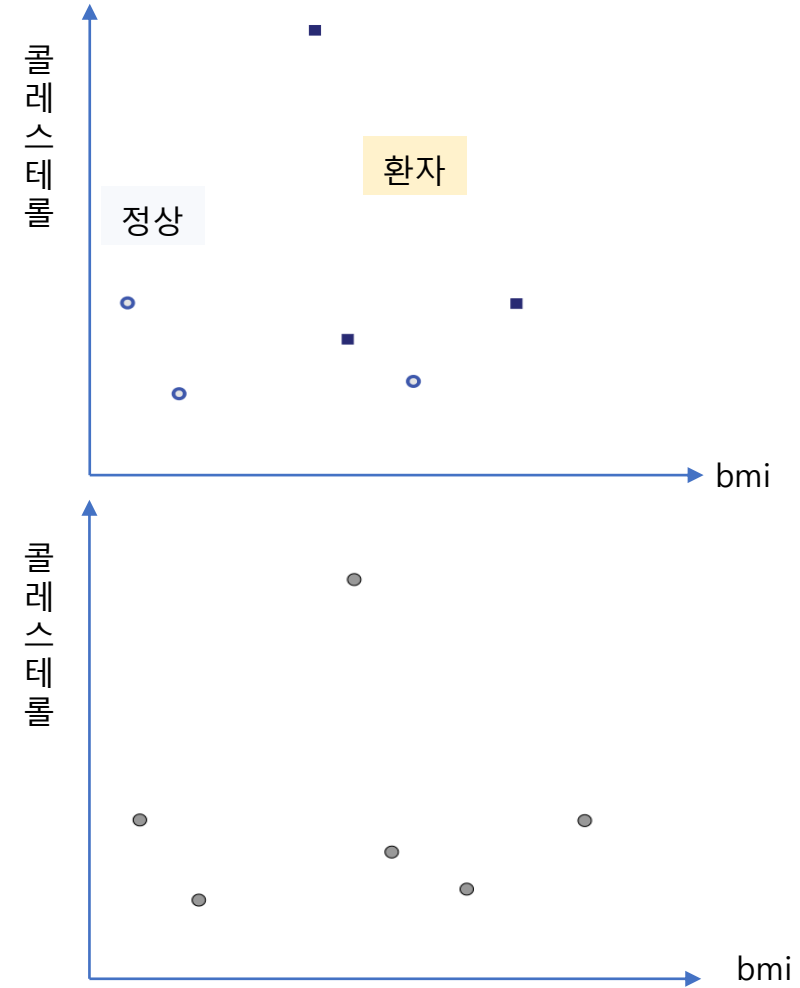
- 특성(feature)
- 독립변수  
(independent variable)
- 레이블(label)
- 종속변수  
(dependent variable)

*Supervised Learning*

bmi	가족력	콜레스테롤	.....	환자 여부
23.1	있음	113.4	.....	환자
18.4	없음	123.4	.....	정상
22.4	있음	198.4	.....	환자
19.5	없음	98.4	.....	정상
26.7	있음	123.2	.....	환자
24.5	없음	101.8	.....	정상

*Unsupervised Learning*

bmi	가족력	콜레스테롤	.....
23.1	있음	113.4	.....
18.4	없음	123.4	.....
22.4	있음	198.4	.....
19.5	없음	98.4	.....
26.7	있음	123.2	.....
24.5	없음	101.8	.....



## 개념 설명

- Orange에서는 비지도학습이 Unsupervised 영역에 있으나, 연관성 기반 Associate, Network도 비지도학습으로 볼 수 있음
- 먼저 Unsupervised에는 군집화의 군집분석 알고리즘들과, 시각화의 대응분석, 다차원척도법, t-SNE 등이 있음

### > Orange의 비지도학습 알고리즘

## Unsupervised



### > 핵심 기능

- 1. t-SNE
- 2. Hierarchical Clustering
- 3. k-Means
- 4. Louvain Clustering
- 5. DBSCAN
- 6. Corresponding Analysis
- 7. MDS

### Key Point

비지도 학습을 위한 19개의 알고리즘 제공으로 분석하고자 하는 알고리즘 대부분 사용 가능

## 개념 설명

- Associate는 Association Rules(연관규칙)을 파악하여 항목 간 연관성을 파악하는 알고리즘임
- Network는 네트워크 관계를 통해 전체 및 그룹/개별 관계성을 주로 시각화를 통해 파악하는 비지도적 학습법

### Orange의 Associate와 Network 알고리즘

#### Associate



Frequent  
Itemsets



Association  
Rules

#### Networks



Network File



Network  
Explorer



Network  
Generator



Network  
Analysis



Network  
Clustering



Network Of  
Groups



Network From  
Distances



Single Mode



Save Network

### 핵심 기능

#### 가. Associate

- 1. Associate Rules

#### 나. Networks

- 2. Network Explorer
- 3. Network Analysis
- 4. Network Clustering
- 5. Network of Groups

### Key Point

연관성 기반 Associate과 Network도 비지도적 학습법이며, 위젯을 통해 알고리즘 활용 가능

개념  
설명

- 비지도학습은 경향과 관계성을 파악하는 데 목적을 두고 있음
- Orange에서는 비지도학습 알고리즘을 Unsupervised, Associate, Network로 구분하고 있으며, 각 영역에서 제공하는 비지도 학습 알고리즘을 참고하여 원하는 형태의 분석 수행이 가능함

▶ 비지도 학습의 종류

분류	구분	알고리즘
Unsupervised	군집화	Hierarchical Clustering
		k-Means
		Louvain Clustering
		DBSCAN
	지각도	Corresponding Analysis
		MDS
		t-SNE
Associate	연관규칙분석	Associate Rules
Networks	네트워크분석	Network Explorer
		Network Analysis
		Network Clustering
		Network of Groups

Key Point

분석 목적에 해당하는 분석 모델을 사용하는 것이 중요

### III CASE분석2: ORANGE로 비지도학습 해보기

## 2. 군집으로 유사한 집단 묶어보기

---



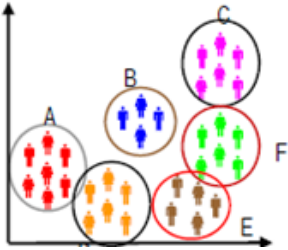
개념  
설명

- 군집분석은 개체들을 다양한 변수를 기준으로 다차원 공간에서 유사한 특성을 가진 개체로 묶는(Clustering) 방법
- 통계적으로는 개체들의 유사성(Similarity)과 상이성(Dissimilarity)에 근거하여 군집을 찾고 자료를 요약하는 탐색적인 자료분석방법

클러스터링 알고리즘



인구학적 특성, 지역적 특성, 라이프스타일, 거래특성 등의 데이터를 기준으로 다차원 관점의 개체분석과 분류



군집 C의 고객특성

- 연령은 30대이며, 소득은 월 300만원 이상이고, 기혼이며, 유아기의 자녀가 있음.
- 교육수준은 대졸이상이며, 주로 아파트에 거주하고, 맛벌이를 하며, 외식빈도가 높음.
- 구매금액은 월 평균 20만원, 구매회수는 월 평균 4.2회, 주 지불수단은 카드
- ....

	Demo/Geo	Lifestyle	Transaction	
▪ 군집 A	-	-	-	-
▪ 군집 B	-	-	-	-
▪ 군집 C	-	-	-	-
▪ 군집 D	-	-	-	-
▪ 군집 E	-	-	-	-
▪ 군집 F	-	-	-	-

개체 군집별  
특성파악

분류	구분	알고리즘
Unsupervised	군집화	Hierarchical Clustering
		k-Means
		Louvain Clustering
		DBSCAN

Key Point

클러스터링을 통해 개체 군집별 특성 파악 가능



## 개념 설명

- 군집분석 알고리즘은 Hierarchical Clustering, k-Means, Louvain Clustering, DBSCAN 등이 있음

### > ORANGE의 군집분석

분류	알고리즘	개념	장단점
군집 분석	Hierarchical Clustering	<ul style="list-style-type: none"> <li>• 군집의 형성에 위계(계층)가 존재</li> <li>• 일단 한 군집에 속하게 된 두 개체는 헤어지지 않음</li> <li>• 군집화 속도가 오래 걸림</li> <li>• 군집수를 정하지 않고 탐색적 적합 군집수 파악에 사용</li> </ul>	<ul style="list-style-type: none"> <li>• 군집수를 확정하기 위한 탐색적 분석으로 사용됨</li> <li>• 연속형 변수만 사용할 수 있고 범주형 자료는 사용할 수 없음[표준화 필수]</li> <li>• 계층적 군집분석은 대규모 데이터의 처리에 적합하지 않아 고객 세분화 분석에서 잘 사용되지 않음</li> </ul>
	k-Means	<ul style="list-style-type: none"> <li>• 군집이 형성된 이후에도 일정 기준에 따라 개체들이 이합집단을 되풀이</li> <li>• 군집화 속도가 빠름 개체수가 많은 경우에 적합</li> <li>• 최종 군집의 수를 미리 정해주어야 함</li> </ul>	<ul style="list-style-type: none"> <li>• 타 군집방법에 비해 연산시간이 상대적으로 짧음</li> <li>• 연속형 변수만 사용할 수 있고 범주형 자료는 사용할 수 없음[표준화 필수]</li> <li>• 군집수 결정이 어려움</li> </ul>
	Louvain Clustering	<ul style="list-style-type: none"> <li>• 네트워크 이론에 기반한 알고리즘</li> <li>• 대규모 데이터에서 적합하며, Bottom-up 방식의 군집 탐색</li> </ul>	<ul style="list-style-type: none"> <li>• 군집분석에서는 군집수를 임의적으로 지정해야 하는 반면, Louvain Clustering은 클러스터의 수를 정하지 않아도 됨</li> <li>• 알고리즘의 초기 조건이나 무작위성 때문에 다른 결과가 도출될 수 있어 적절한 파라미터 설정이 필요함</li> </ul>
	DBSCAN	<ul style="list-style-type: none"> <li>• DBSCAN은 Density-based spatial clustering of applications with noise의 약자로서 밀도 기반 클러스터링 기법</li> <li>• 밀도 기반의 클러스터링은 케이스가 집중되어 있는 밀도(density)에 초점을 두어 밀도가 높은 그룹을 클러스터링하는 방식임</li> <li>• 중심점을 기준으로 특정한 반경 이내에 케이스가 n개 이상 있을 경우 하나의 군집을 형성하는 알고리즘</li> </ul>	<ul style="list-style-type: none"> <li>• 군집분석에서는 군집수를 임의적으로 지정해야 하는 반면, DBSCAN은 클러스터의 수를 정하지 않아도 됨</li> <li>• Noise point를 통하여 outlier 검출이 가능하며 이상거래 및 이상징후 탐지 분야에 활용도가 높아 이상치나 노이즈가 존재하는 현실적인 데이터 적합</li> <li>• 케이스간 거리 계산시 시작되는 데이터의 순서에 따라 결과가 달라짐</li> <li>• 고차원의 경우 거리측정법에 따라 차원의 저주에 걸려 적절한 엡실론을 찾는 데 어려울 수 있음</li> <li>• 다양한 밀도를 다루기 어려움</li> <li>• 파라미터에 민감하여 적정 파라미터를 결정하기 어려움</li> </ul>

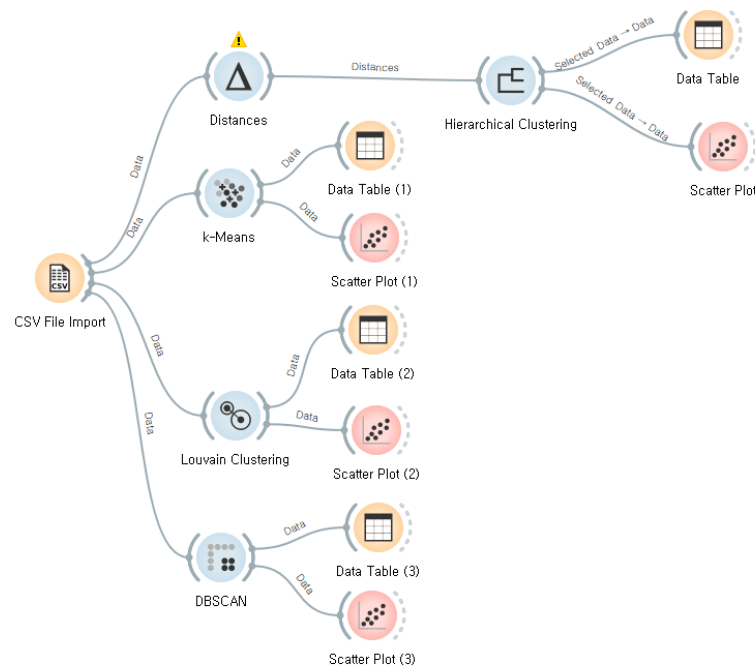
### Key Point

데이터에 적합한 클러스터링 방법을 적용

## 분석 모델

- 학습데이터: Cluster\_iris.csv
- 작업파일: clustering.ows

### ▶ 전체 분석 과정



### Key Point

클러스터링 방법에 따른 결과 차이 탐색

## 데이터 설명

- 군집분석(Clustering) 실습에 활용할 데이터는 “Cluster\_iris.csv”로 5개의 변수, 151개의 데이터로 이루어짐

### > 데이터 예제

	A	B	C	D	E
1	sepal_leng	sepal_widt	petal_leng	petal_widt	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa
15	4.3	3	1.1	0.1	Iris-setosa
16	5.8	4	1.2	0.2	Iris-setosa
17	5.7	4.4	1.5	0.4	Iris-setosa
18	5.4	3.9	1.3	0.4	Iris-setosa

### > 데이터 특징

- 151개 데이터, 5개 변수
- sepal\_leng : 꽃받침 길이
- sepal\_width : 꽃받침 너비
- petal\_leng : 꽃잎 길이
- petal\_width : 꽃잎 너비
- Class : 품종

### Key Point

붓꽃의 품종 데이터

# IV

PART

## CASE분석3: ORANGE로 텍스트마이닝 해보기

---

1. 텍스트마이닝 기본이해와 활용
2. 텍스트마이닝 시나리오 설계와 기초분석
3. 텍스트 클러스터링과 시각화
4. 텍스트 임베딩과 활용

## IV CASE분석3: ORANGE로 텍스트마이닝 해보기

# 1. 텍스트마이닝 기본이해와 활용

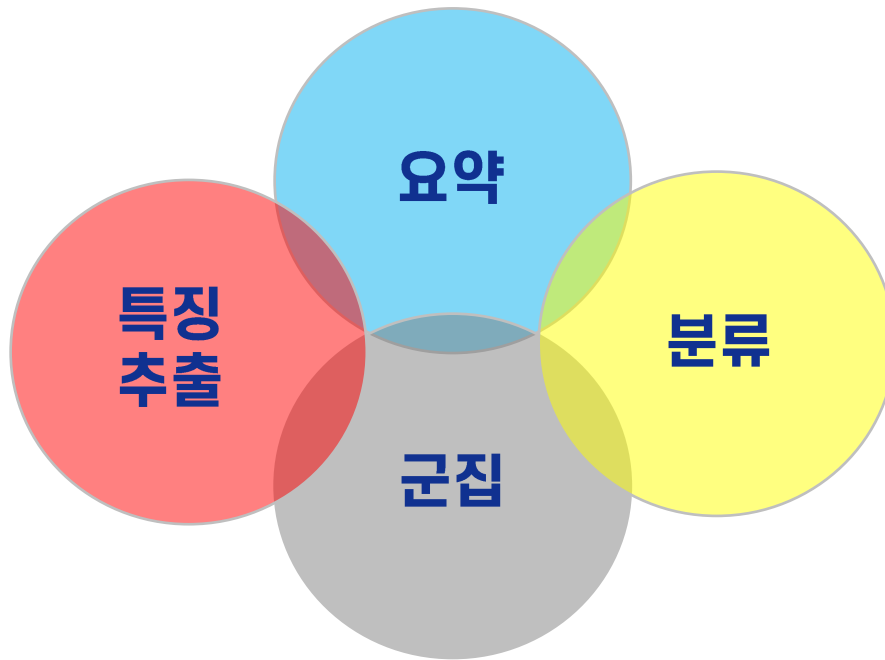
---



## 개념 설명

- 텍스트마이닝은 문서 요약, 분류, 군집, 특징 추출에 활용할 수 있음

### ▶ 텍스트마이닝의 활용범위



### ▶ 텍스트마이닝의 활용범위 설명

- 문서 요약 (Summarization)  
: 논문/신문/보고서 요약
- 문서 분류 (Classification)  
: 자동 범주화 (예: 뉴스 기사 분석 -> 사회/경제/금융 등으로 분류)
- 문서 군집 (Clustering)  
: 유사 단어 또는 유사 문서 간의 군집 분석
- 특징 추출 (Feature Extraction)  
: 주요 키워드 추출

### Key Point

데이터를 분석하고 유용한 정보를 추출하는데 활용됨

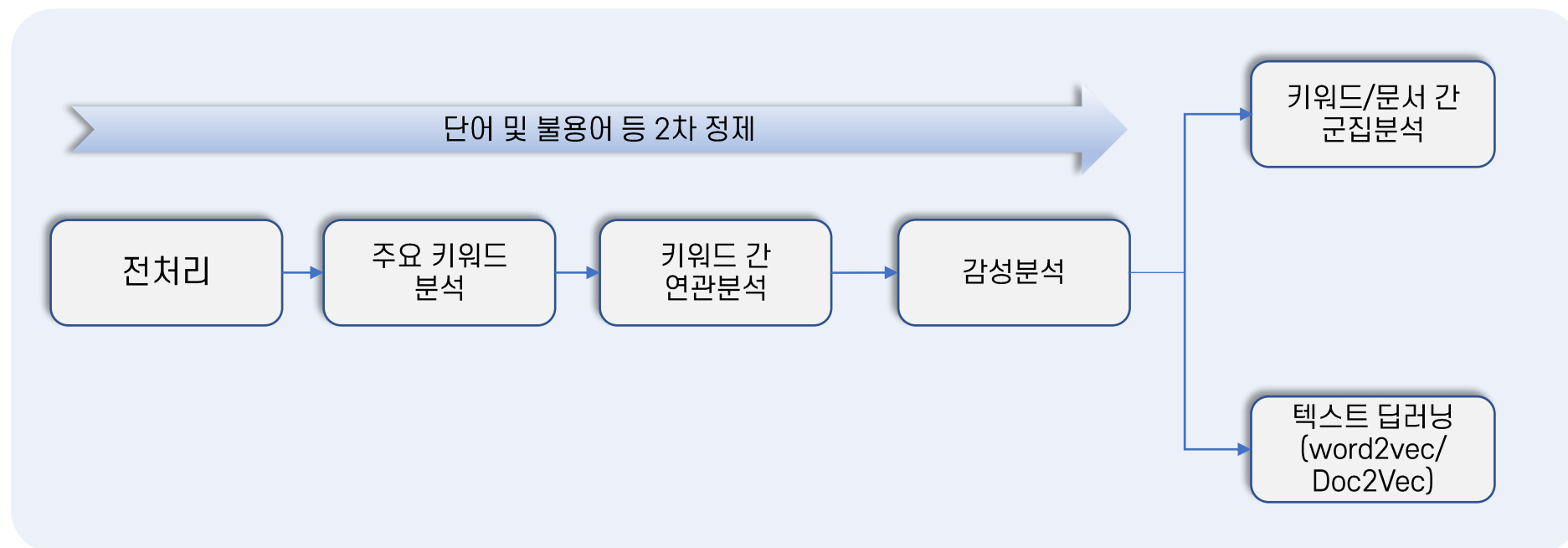
## 개념 설명

- 텍스트마이닝의 전반적인 프로세스는 다음과 같음

### ▶ 텍스트 분석 프로세스



Documents



## 실습 과정

- Orange에서는 Text Mining 영역에서 총 31개 위젯을 제공함

### ▶ Orange의 텍스트마이닝 위젯



### Key Point

31개의 위젯 제공으로 다양한 텍스트 마이닝 분석 가능



## IV CASE분석3: ORANGE로 텍스트마이닝 해보기

## 2. 텍스트마이닝 시나리오 설계와 기초분석

---



## 개념 설명

- 텍스트마이닝 기초분석은 전처리 – 주요 키워드 분석 – 감성분석 순으로 진행

### ▶ 텍스트마이닝 프로세스

#### 데이터 전처리

- Bag of Words(BOW)
  - 문서에 존재하는 단어와 그 출현 빈도를 벡터공간으로 매핑하여 단어 벡터를 생성
  - 단어 벡터 간의 유사도를 통해 문서의 유사도를 측정하는 방식
- 불용어(Stopwords)
  - 문서에서 빈번하게 사용되어 검색에서 무시해 버리는 문자열을 의미
  - 전치사, 관사, 접속사, 한국어의 조사, 특수기호 등

#### 주요 키워드 분석

- 텍스트에서 가장 중요한 단어들을 식별하고 이들의 빈도와 관련성 등을 분석하는 과정
- TF(Term Frequency) 자료변환과 wordcloud 시각화

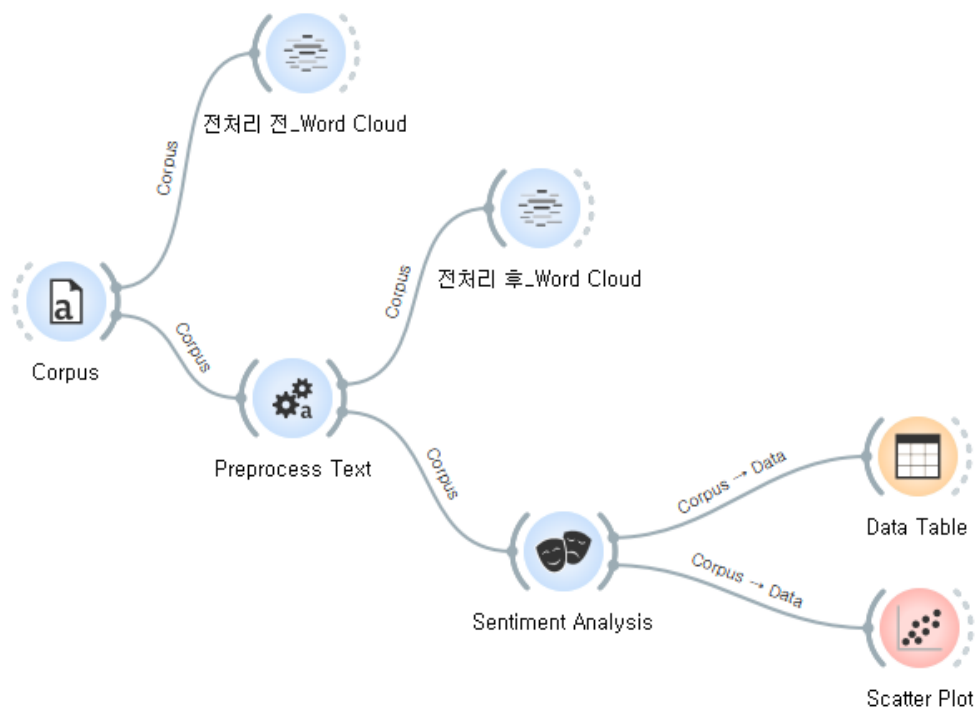
#### 감성분석

- 텍스트 마이닝 및 오피니언 마이닝의 한 기법으로써 데이터로부터 소비자의 감성 관련 정보를 추출하는 방식
- 주로 데이터를 작성한 사용자의 감정, 태도, 의견, 성향 같은 주관적인 데이터를 추출하고자 하는 방법으로써 컴퓨터 언어 및 자연어를 사용
- Opinion extraction, Opinion mining, Sentiment mining, Subjectivity analysis 등으로도 불림

## 분석 모델

- 학습데이터: wos\_ai\_.csv
- 작업파일: 기초분석.ows

### ▶ 전체 분석 과정



### Key Point

텍스트 데이터를 활용한 감성분석

# 데이터 설명

- 실습 데이터는 “wos\_ai.csv”로 SCI(E)/SSCI/AHCI 수록 논문 및 인용정보를 동시에 검색할 수 있는 웹 데이터베이스 Web of Science에서 ‘artificial intelligence’ 키워드를 바탕으로 수집한 논문 데이터임
- 연도, 제목, 저자, 초록 등의 상세 필드 구축하고 논문 초록을 바탕으로 텍스트 마이닝 실습 진행

## ▶ 데이터 예제

The screenshot displays the Web of Science search results page. The search term 'artificial intelligence' is entered in the search bar. The results are displayed in a table with columns: YEAR, TITLE, AUTHOR, and ABSTRACT. The table shows a list of research papers related to artificial intelligence, including titles like 'Computational intelligence (CI) involves using a computer algorithm to capture hidden knowledge from data and to use the', 'Smart production >> is a new model of industrial development of the 21st century on the basis of digital technologies', and 'This article is based on the New Horizons lecture delivered at the 2016 Radiological Society of North America Annual Meeting'.

## ▶ 데이터 특징

- 119개 데이터
- 4개 변수
- Year: 연도
- Title: 제목
- Author: 저자
- Abstract: 초록

## Key Point

논문 초록 데이터

## IV CASE분석3: ORANGE로 텍스트마이닝 해보기

### 3. 텍스트 클러스터링과 시각화

---



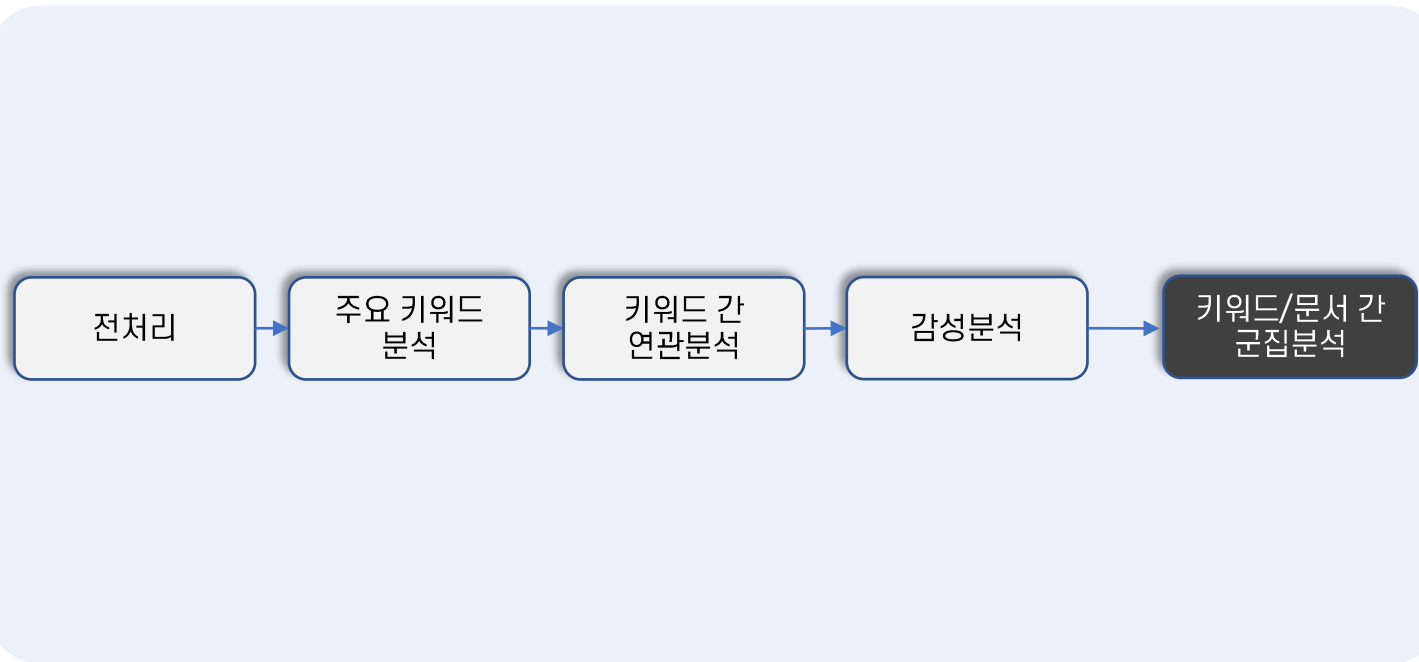
## 개념 설명

- 개체들을 다양한 변수를 기준으로 다차원 공간에서 유사한 특성을 가진 개체로 묶는(Clustering) 방법.
- 키워드 클러스터링: 단어간 거리를 계산하기 위해서는 문서가 축(기준)이 되고 단어의 좌표 간의 거리를 계산함. TDM 형태의 자료가 분석의 단위임
- 문서 클러스터링: 문서간 거리를 계산하기 위해서는 단어가 축(기준)이 되고 문서의 좌표 간의 거리를 계산함. DTM 형태의 자료가 분석의 단위임

### ▶ 텍스트 클러스터링 프로세스



Documents



### ▶ 자료 형태

- 키워드 클러스터링 TDM

	문서1	문서2	문서3
Model	24	21	12
system	32	10	16

- 문서 클러스터링 DTM

	model	System	algorithm
문서1	24	21	9
문서2	32	10	5

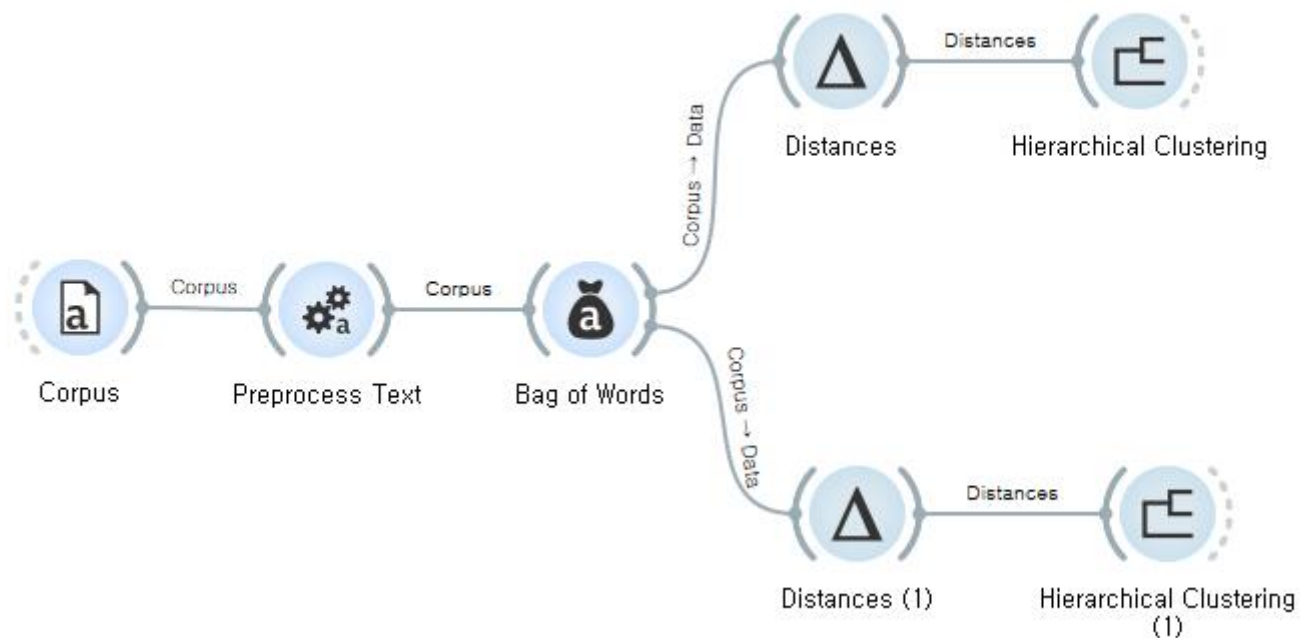
### Key Point

개체들의 유사성과 상이성에 근거하여 군집을 찾고 자료를 요약

## 분석 모델

- 학습데이터: movie\_review.csv
- 작업파일: text\_clustering.ows

### ▶ 전체 분석 과정



### Key Point

영화 리뷰를 활용한 텍스트 클러스터링

## 데이터 설명

- 실습 데이터는 “movie\_reivew.csv”로 2개의 변수, 10개의 데이터로 이루어짐

### > 데이터 예제

	A	B
1	genre	Review
2	Action	The plot of the movie was predictable, but the special effects were stunning.
3	Drama	An outstanding performance by the lead actor, truly a must-watch.
4	Mystery	The storyline was confusing and hard to follow.
5	Romance	A beautiful film with breathtaking cinematography.
6	Drama	The movie was too long and felt dragged out.
7	Family	An inspiring and heartwarming story, perfect for a family night.
8	Musical	The soundtrack was amazing and complemented the film perfectly.
9	Drama	I didn't enjoy the movie; the acting was subpar.
10	Adventure	A thrilling adventure from start to finish.
11	Drama	The characters were well-developed and relatable.

### > 데이터 특징

- 10개 데이터
- 2개 변수
- genre: 장르
- Review: 리뷰

### Key Point

영화 리뷰 데이터



## IV CASE분석3: ORANGE로 텍스트마이닝 해보기

### 4. 텍스트 임베딩과 활용

---



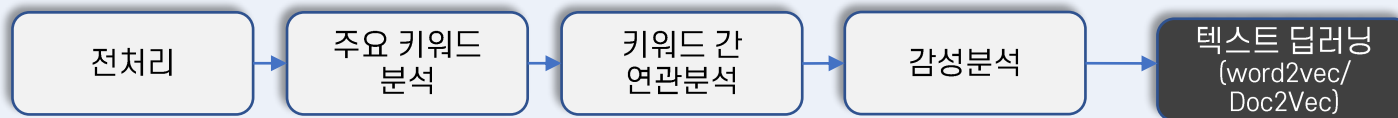
## 개념 설명

- 텍스트를 딥러닝 등 알고리즘으로 적용하기 위해서는 벡터화가 필요함. 다차원에서 텍스트를 벡터화하는 방법을 임베딩(embedding)이라고 함
- 통상 300차원의 내외의 다차원을 임의로 형성한 후 텍스트의 위치를 좌표값으로 변환

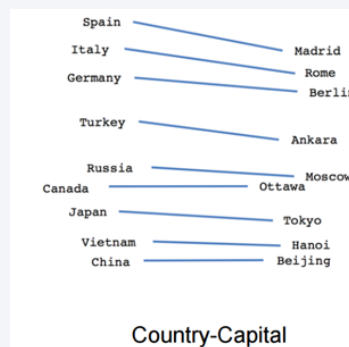
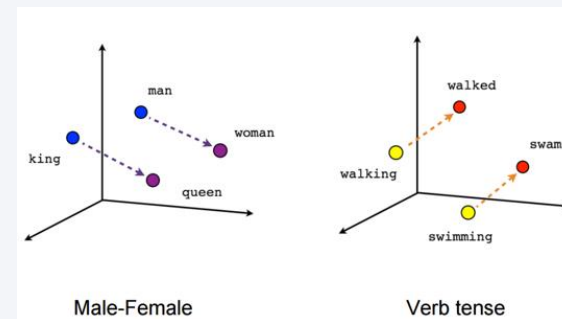
### ▶ 텍스트 임베딩 프로세스



Documents



### ▶ 텍스트 임베딩과 활용 예시



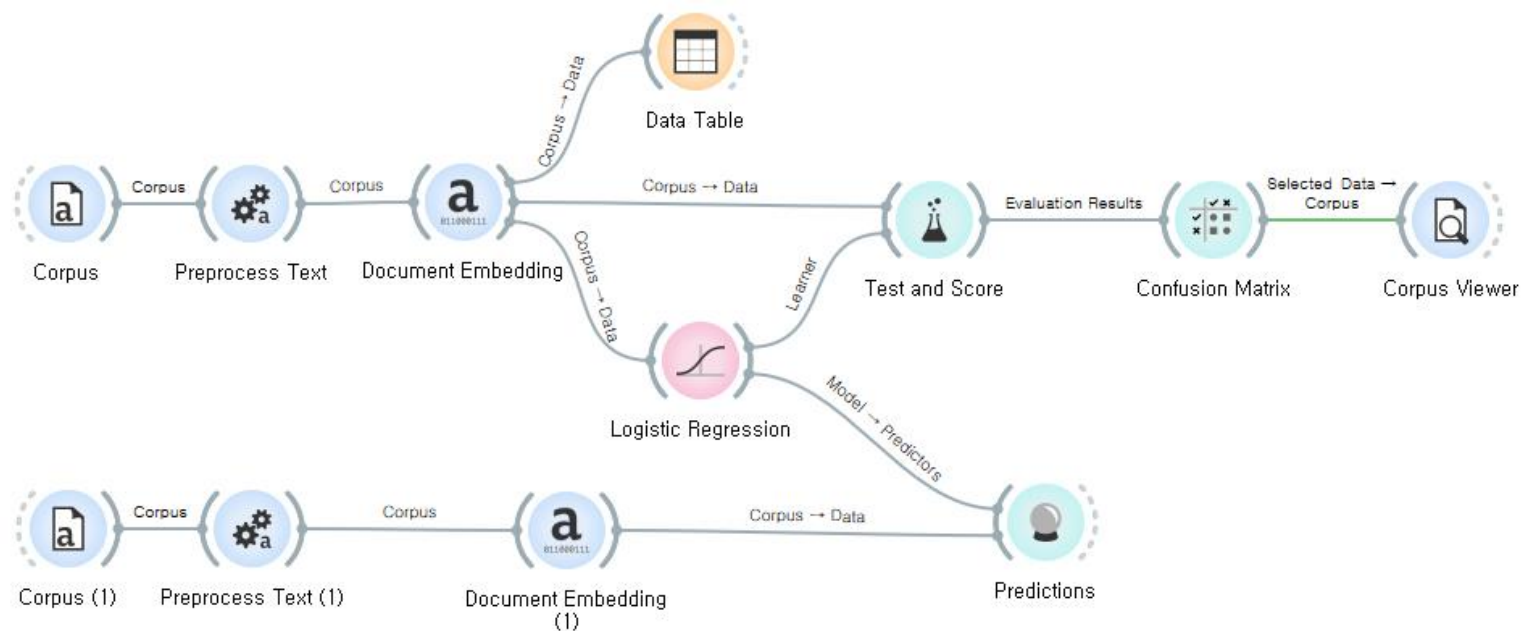
### Key Point

타겟이 있는 경우 지도학습, 없는 경우 비지도학습으로 다양하게 활용 가능

## 분석 모델

- 학습데이터: grimm-tales-selected.tab, Andersen.tabmovie\_review.csv
- 작업파일: text\_embedding.ows

### ▶ 전체 분석 과정



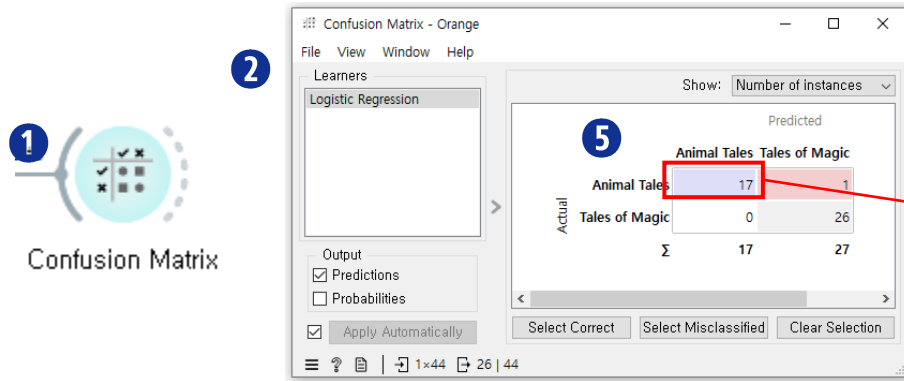
### Key Point

동화 데이터를 활용한 텍스트 임베딩

# 분석 결과

- Confusion Matrix 설정
- 결과 확인

## > Confusion Matrix 설정



Predictions - Orange

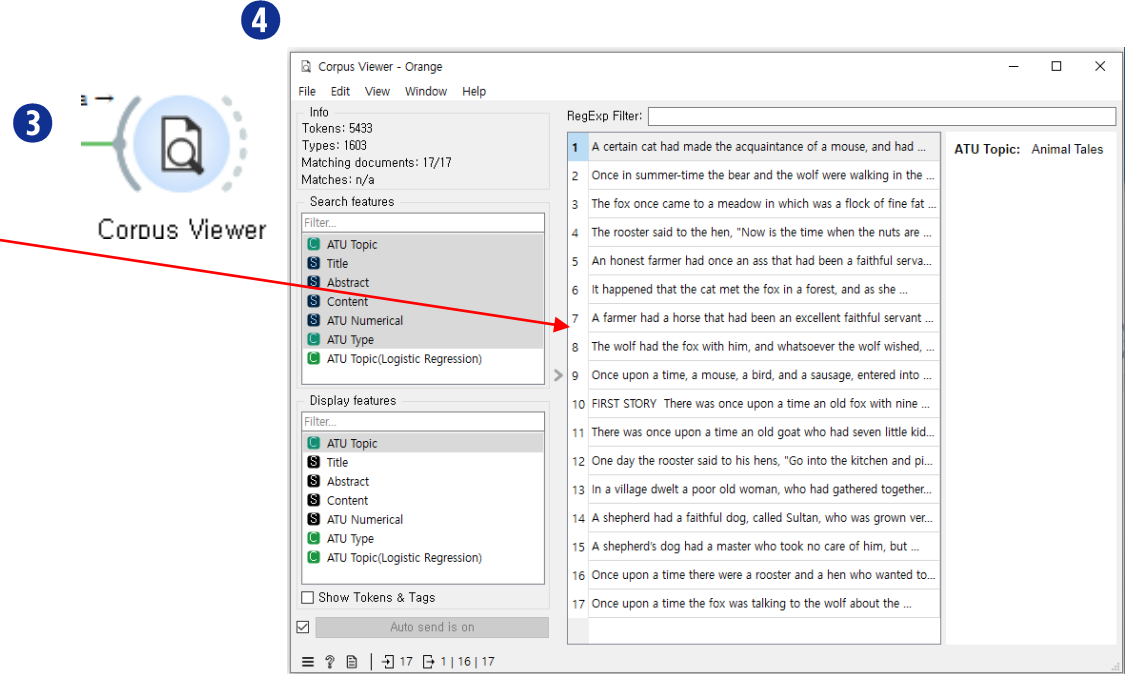
File View Window Help

Show probabilities for: (None)

Restore Original Order

Logistic Regression	Title	Content	Dim1	Dim2	Dim3	Dim4
1 Tales of Magic	The Little Matc...	it was terribly c...	-0.179991	0.309565	0.181175	0.500217
2 Tales of Magic	The Philosophie...	Far away towar...	0.204133	-0.0740185	0.0733724	0.287322
3 Animal Tales	The Ugly Duckl...	it was lovely su...	0.33734	0.389494	0.302773	0.0983606

## > 결과 확인



- 1 Evaluate – Confusion Matrix를 클릭한 후 Test and Score와 이어준다
- 2 Confusion Matrix 옵션을 선택한 후 결과값을 확인한다
- 3 Text Mining – Corpus Viewer를 클릭한 후 Confusion Matrix와 이어준다

- 4 Confusion Matrix 결과창을 화면에 띄운 채로 Corpus Viewer를 더블 클릭한다
- 5 Confusion Matrix 결과창에서 보고싶은 결과를 클릭한 후 Corpus Viewer에서 결과를 확인한다



PART

## CASE분석4: ORANGE로 이미지분석 해보기

---

1. 이미지분석 기본이해와 활용
2. 이미지 임베딩과 시각화
3. 이미지 분류와 전이모델

# 1. 이미지분석 기본이해와 활용

---



## 개념 설명

- Orange에서는 이미지 분석이 Image Analytics라는 영역에 있으며, 총 5개의 위젯을 제공함

### ➤ Orange의 이미지 분석 위젯

## Image Analytics



Import Images



Image Viewer



Image  
Embedding



Image Grid



Save Images

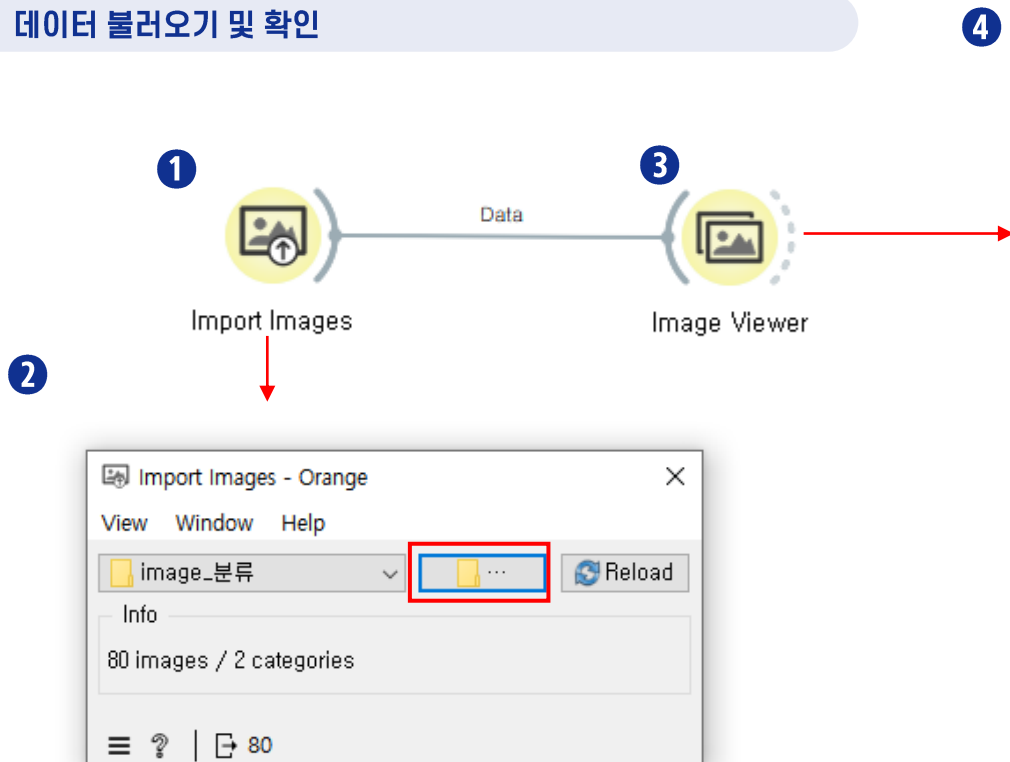
### Key Point

5개의 위젯 제공으로 다양한 이미지 분석 가능

# 실습 과정

- 분석에 사용할 데이터를 불러온 후 확인

## > 데이터 불러오기 및 확인



- 1 Image Analytics – Import Images를 클릭한다
- 2 분석할 데이터를 불러온다(image\_분류)
- 3 Image Analytics – Image Viewer를 클릭한 후 Import Images와 이어준다

- 4 Image Viewer를 더블 클릭하여 데이터를 확인한다



## 2. 이미지 임베딩과 시각화

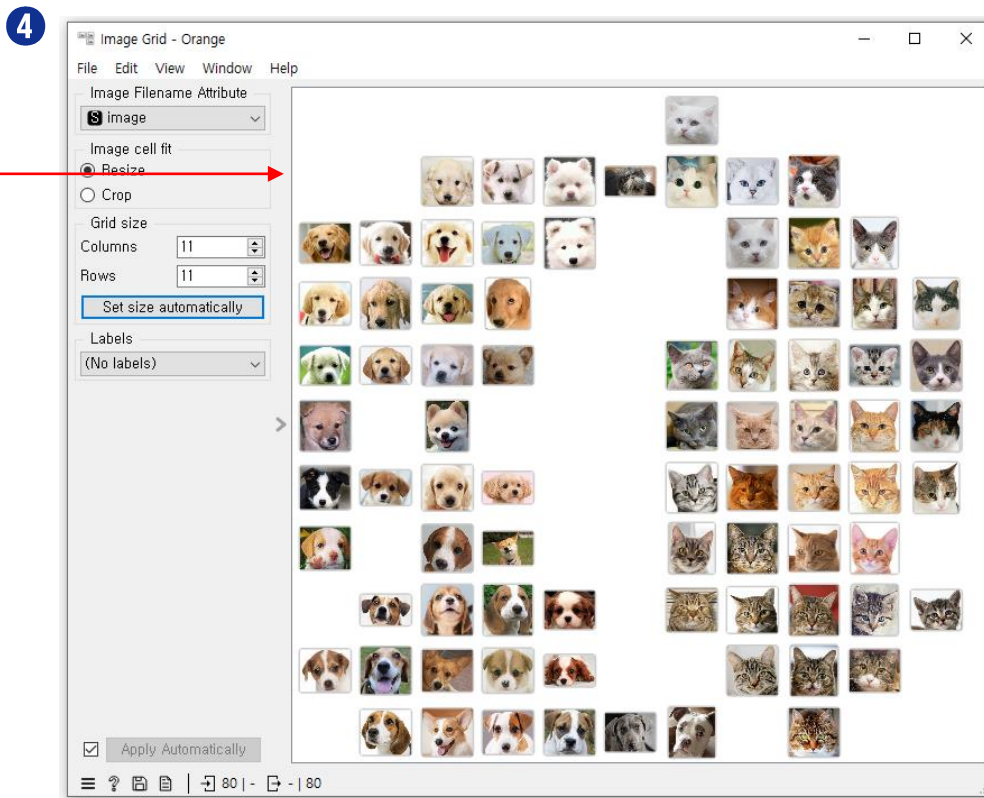
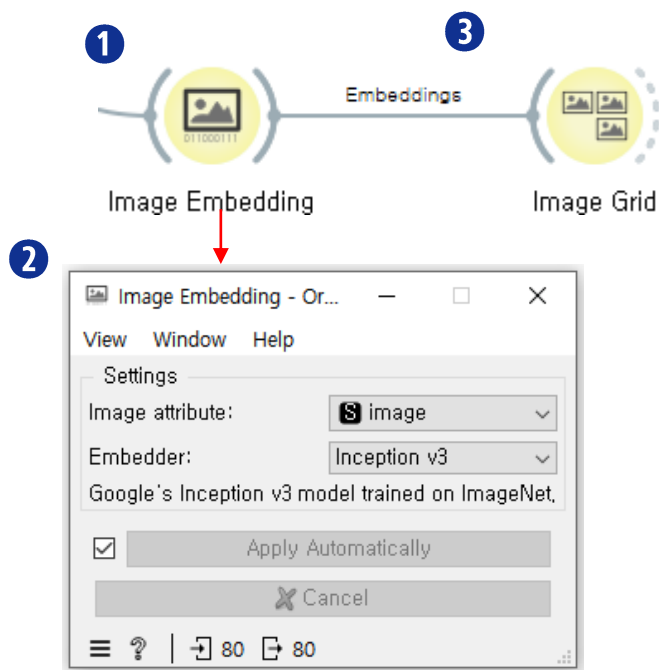
---



# 실습 과정

- Image Embedding 설정 및 시각화 이미지 확인

## > 이미지 임베딩 및 시각화



- 1 Image Analytics – Image Embedding을 클릭한다
- 2 Import Embedding 옵션을 선택한다
- 3 Image Analytics – Image Grid를 클릭한 후 Image Embedding과 이어준다

- 4 Image Grid를 더블 클릭하여 시각화 이미지를 확인한다

### 3. 이미지 분류와 전이모델

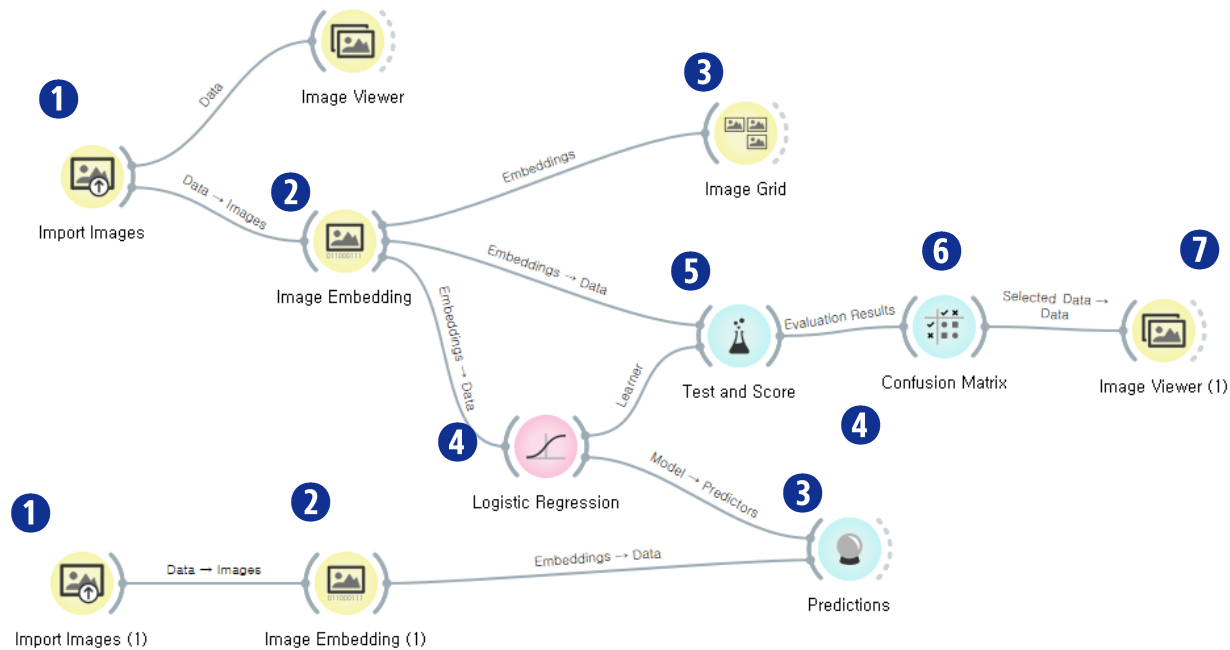
---



## 분석 모델

- 학습데이터: image\_data 폴더 중 image\_분류 / 예측데이터: image\_test
- 작업파일: Image\_analysis.ows

### ▶ 전체 분석 과정



### ▶ 진행순서

#### 1) 학습결과

- ❶ 분석 데이터 불러오기
- ❷ 이미지 임베딩 데이터 전처리
- ❸ 이미지 그리드
- ❹ 선형회귀 알고리즘 설정
- ❺ Test and Score 설정
- ❻ Confusion Matrix 설정
- ❼ 결과 확인

#### 2) 예측결과

- ❶ 예측 데이터 불러오기
- ❷ 예측 데이터 이미지 임베딩
- ❸ 예측값 확인

### Key Point

동물 데이터를 활용한 이미지 임베딩