

코딩없이 클릭으로 하는 머신러닝/딥러닝 분석법

머신러닝 학습하기

강의와 교재를 평생소장하고 공부하세요

코딩 없이, 클릭만으로 할 수 있는 머신러닝과 딥러닝



데이터분석 분야 최고 실력자의 직강

23년간의 1,000건 이상의 데이터 분석 프로젝트 경험을 토대로 만든 고품질의 강의 제공

평생 소장 가능한 강의

평생 옆에 두고 전문가의 고품질 강의를 소장할 수 있는 기회!
(전과목 USB+교재 제공)

데이터를 활용한 실습

데이터를 활용한 실습을 통해 AI분석의 이해도 향상!

클릭으로 완성하는 머신러닝과 딥러닝 올인원 패키지

USB(강의영상·실습파일) + 교재 제공!

1권. ORANGE와 핵심 마이닝

총 20강

2권. ORANGE 지도학습 마스터

총 35강

3권. ORANGE 비지도학습 마스터

총 19강

4권. ORANGE 텍스트와 이미지분석 마스터

총 30강

평생소장 가능 머신러닝 툴 ORANGE 강의와 함께라면
머신러닝과 딥러닝 마스터 가능합니다😊



구매 가격

	정가	할인율	판매가
1과목 구매시	300,000	0%	300,000
2과목 구매시	600,000	10%	540,000
3과목 구매시	900,000	15%	765,000
4과목 구매시	1,200,000	20%	960,000

입금 및 문의 안내

상품구성

- 본 상품은 USB+교재로 구성되어 있습니다
- 상품은 입금 확인 후 택배로 배송됩니다

계좌이체 안내

- 우리은행, 1005-402-421172, (주)와이즈인컴퍼니

문의사항 및 계산서/견적서 요청

- 연락처: 070-8676-1312
- 이메일: hs9177@wiseinc.co.kr

대학의 경우 각 대학 도서관에 구매요청을 하실 수 있습니다



김 원 표

現 (주)와이즈인컴퍼니 대표
한양대 겸임교수

【 주요 경력 】

- 20년간 [2,000건](#) 이상의 통계분석/ 빅데이터 프로젝트 수행
- [20권](#) 이상 통계분석/ 빅데이터 관련 서적 출간
- 연간 [2만명](#) 이상의 수강생이 검증한 전문 강사
- AI 통계, 데이터분석 솔루션 " [데이터인\(DataIN\)](#) " 개발 기획 총괄
- 국책연구소 인공지능 (딥러닝) 프로젝트 다수 수행

CONTENTS



머신러닝 학습하기

1. 머신러닝의 핵심지식 담기
2. 노코딩 머신러닝 솔루션 WEKA 실습

Part 01

머신러닝의 핵심지식 담기



1. 무료 머신러닝 툴

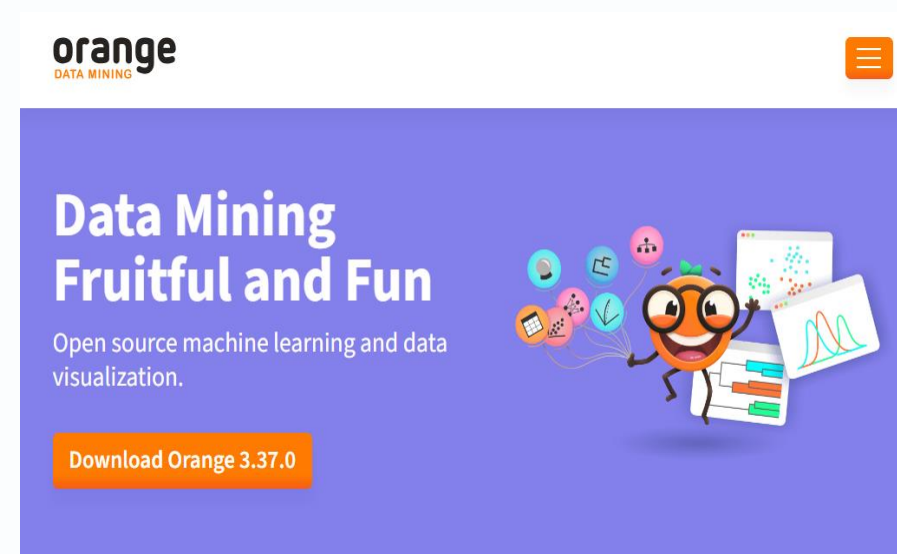
WEKA



다운로드

<https://sourceforge.net/projects/weka/>

ORANGE3



다운로드

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

1. 무료 머신러닝 툴 (공식 동영상)

<https://www.youtube.com/@WekaMOOC>



Data Mining with Weka



WekaMOOC

@WekaMOOC · 구독자 2.55만명 · 동영상 91개

"Data Mining with Weka", "More Data Mining with Weka" and "Advanced Data Mining with W..." >

weka.waikato.ac.nz 외 링크 4개

구독

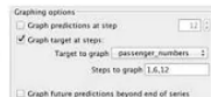
홈 동영상 재생목록 🔍

추천

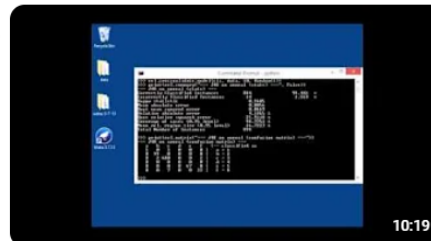
Looking at forecasts

Multi-step forecasts

- Graph predictions at step 12
- Graph target at step 12
- Compare 1-step-ahead, 6-steps-ahead, and 12 steps-ahead predictions
- Change base learner to SMOReg and see the difference
- Get better predictions by reducing attributes (see last lesson's Activity):
 - minimum lag of 12
 - turn off powers of time, products of time and lagged vibs
 - customize to no periodic attributes



9:40



10:19

Lag creation, and overlay data

appleStocks2011: daily High, Low, Open, Close, Volume

- Target selection
 - data contains more than one thing to predict
 - most days from 31 Jan 2011 – 10 Aug 2011
 - forecast Close
 - generates lags up to 12 (Monthly??); set to Daily (lags up to 7)
 - no instances for Jan 8/9, 15/16/17, 22/23, 29/30 ... weekends + a few holidays
 - these "missing values" are interpolated – but perhaps they shouldn't be!
- Skip list:
 - e.g. weekend, sat, tuesday, mar, october, 2011-07-04@yyyy-MM-dd
 - specify weekend, 2011-01-17@yyyy-MM-dd, 2011-02-21, 2011-04-22, 2011-05-30, 2011-07-01
 - set max lag of 10 (2 weeks)

10:30

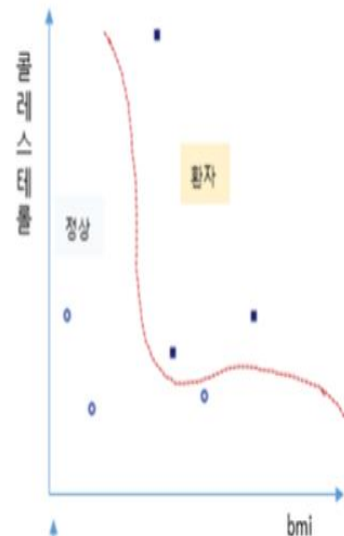
Infrared data from soil samples

- Now let's see where we get X and Y from
- Soil samples have traditionally been analysed in order to determine their organic carbon, etc.). These techniques to our Y values or targets.
- The soil from a "soil bank" is re-used record a unique identifier for each up with the right target(s) established. To actually get the input we put each in near-infrared spectrometer. If you Google lots of machines and also lots of uses for

2. 지도학습과 비지도학습

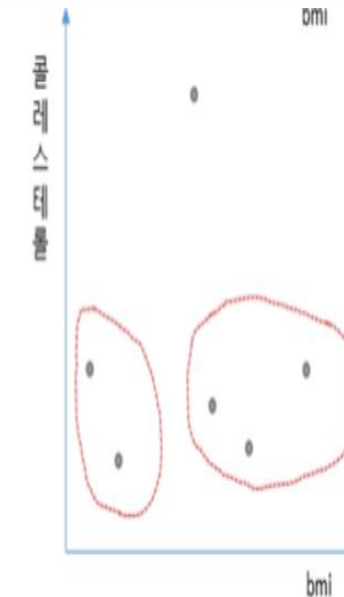
Supervised Learning

<ul style="list-style-type: none"> 특성(feature) 독립변수 (independent variable) 				<ul style="list-style-type: none"> 레이블(label) 종속변수 (dependent variable)
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



구분	목적	설명
지도학습	회귀(regression)	연속형 종속변수 특정 값 예측
	분류(classification)	범주형 종속변수 범주 분류 예측
비지도학습	군집(clustering)	유사한 특성 집단으로 묶음
	연관(association)	연관성 있는 사건 규칙 파악

2. 지도학습과 비지도학습

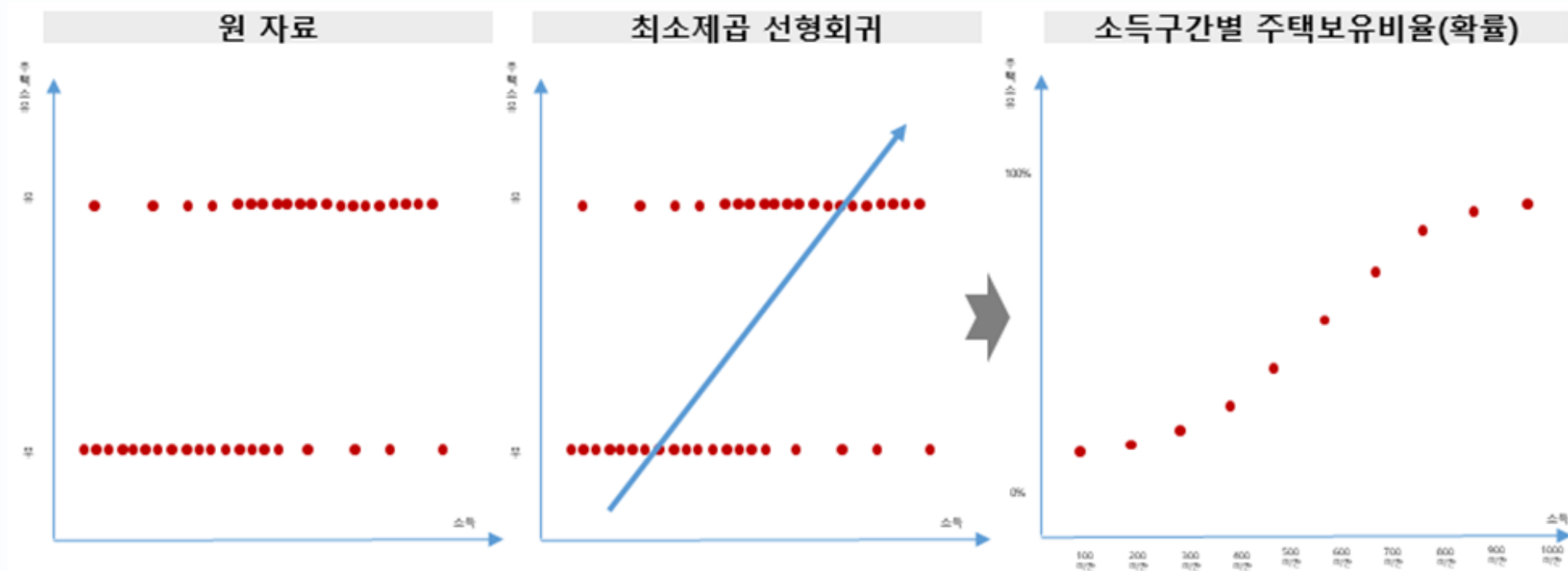
머신러닝 과제	방법	알고리즘
자동차 사양에 따른 중고차 매매가 예측	<u>지도학습</u> vs 비지도학습	선형회귀
스팸메일 자동 분류기	<u>지도학습</u> vs 비지도학습	로지스틱회귀
의료 진단에 따른 종양 판단 분석	<u>지도학습</u> vs 비지도학습	로지스틱회귀
의심되는 신용카드 거래 감지 분석	지도학습 vs <u>비지도학습</u>	비정상탐지
유사한 구매 및 취향 패턴을 보이는 고객 집단 분류	지도학습 vs <u>비지도학습</u>	군집/클러스터링
넷플릭스의 영화추천 시스템	지도학습 vs <u>비지도학습</u>	연관/추천
CT/MRI 이미지 분석을 통한 진단	<u>지도학습</u> vs 비지도학습	로지스틱 회귀/신경망
자동 번역 알고리즘	<u>지도학습</u> vs 비지도학습	딥러닝(RNN)
음성인식 알고리즘	<u>지도학습</u> vs 비지도학습	딥러닝(RNN)
감성분석을 활용한 주가예측 시스템	<u>지도학습</u> vs 비지도학습	딥러닝
귀하가 생각하시는 비즈니스 서비스는 ?	지도학습 vs 비지도학습	

3. 대표 알고리즘

1. 로지스틱 회귀

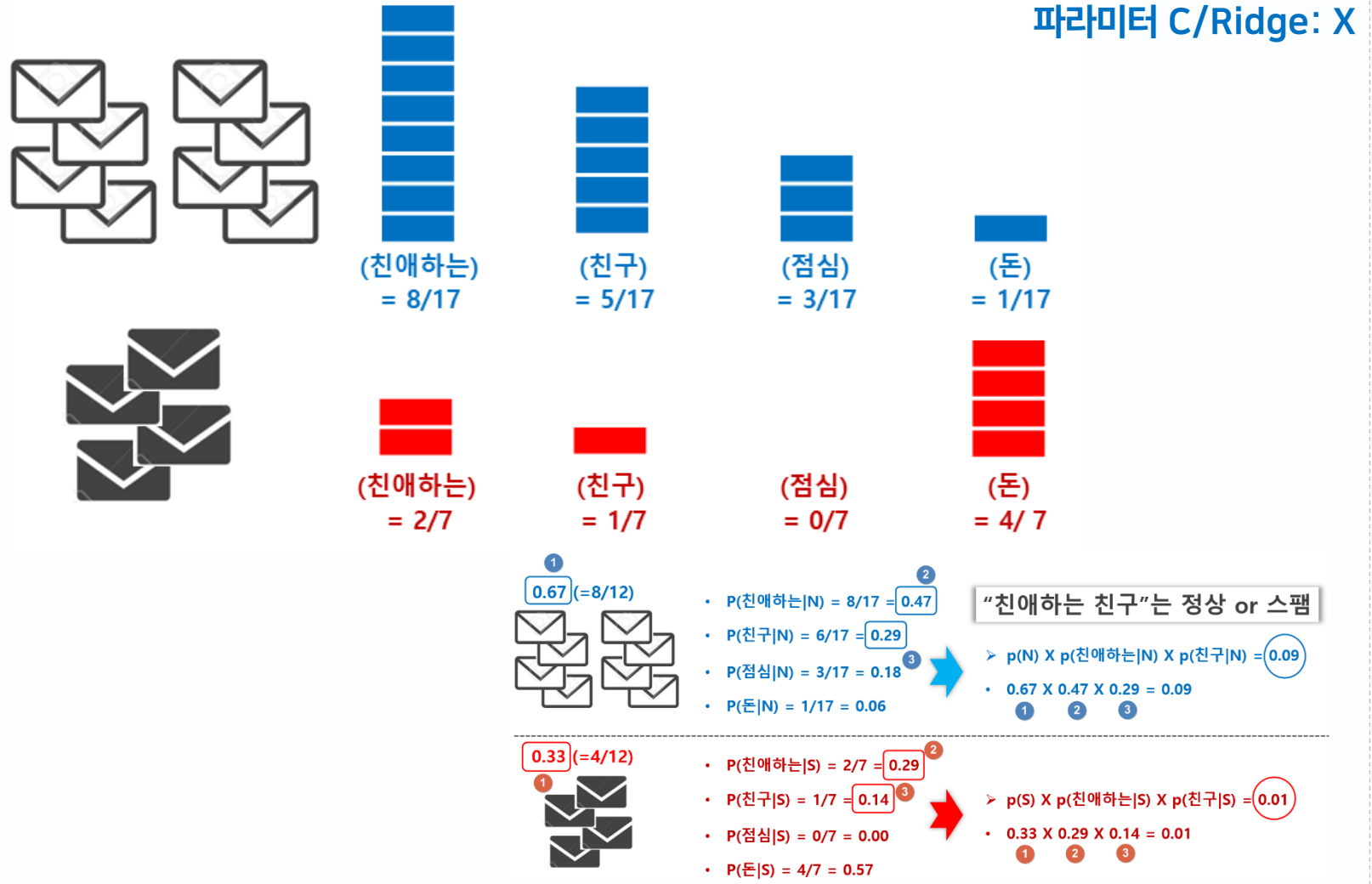
2. 나이브 베이즈
3. 의사 결정 트리
4. K최근접 이웃
5. 서포트 벡터 머신
6. 랜덤포레스트

파라미터 C/Ridge: 0~ 무한대



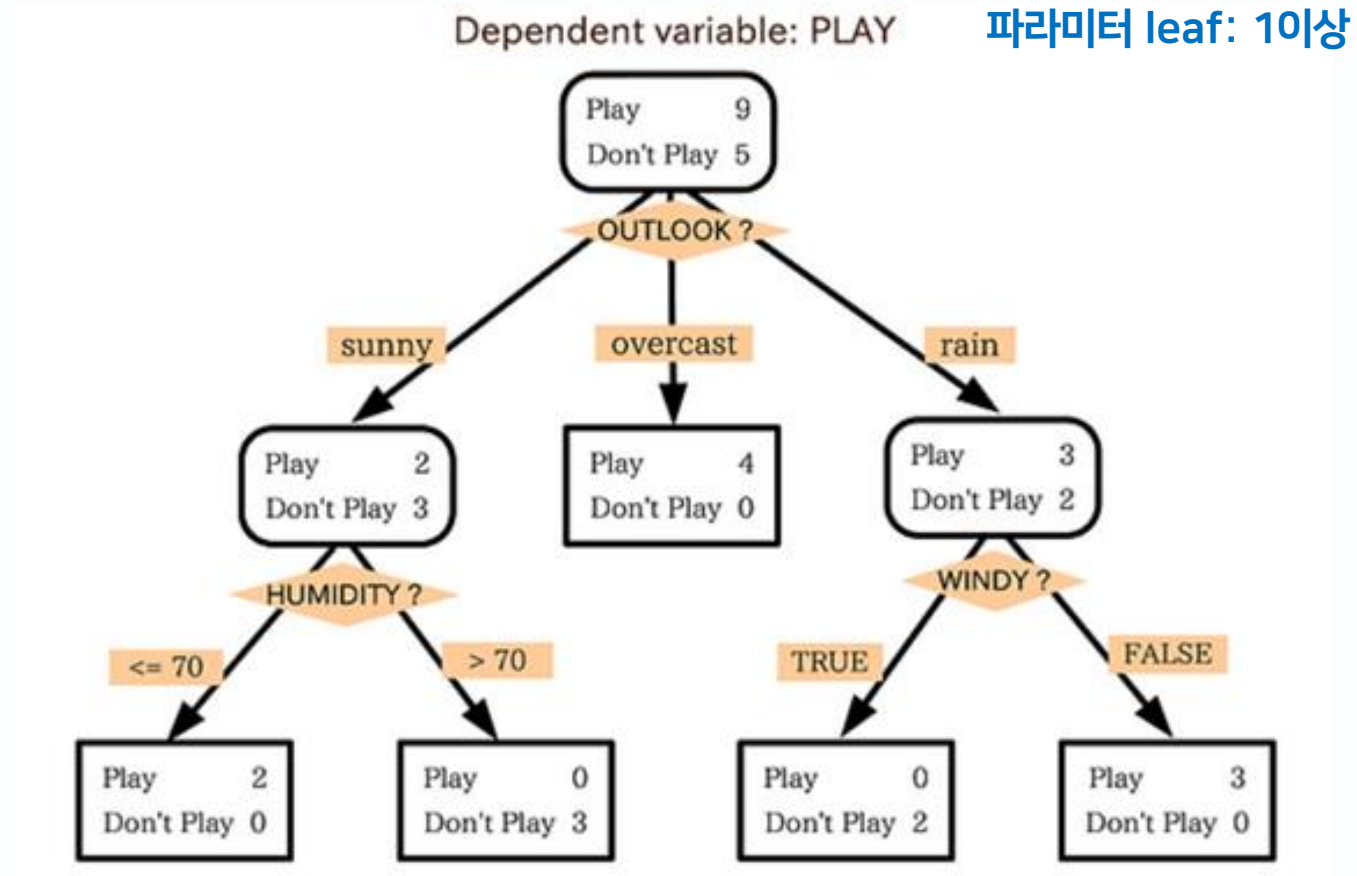
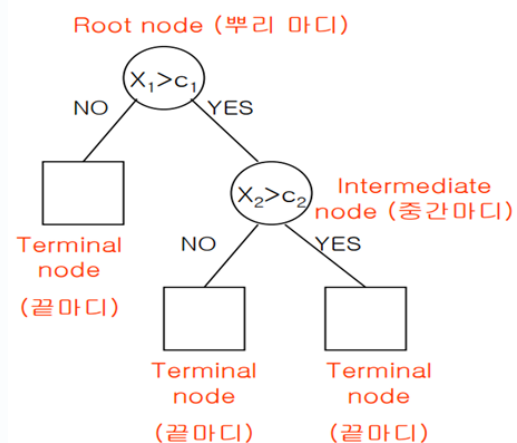
3. 대표 알고리즘

1. 로지스틱 회귀
2. 나이브 베이즈
3. 의사 결정 트리
4. K최근접 이웃
5. 서포트 벡터 머신
6. 랜덤포레스트



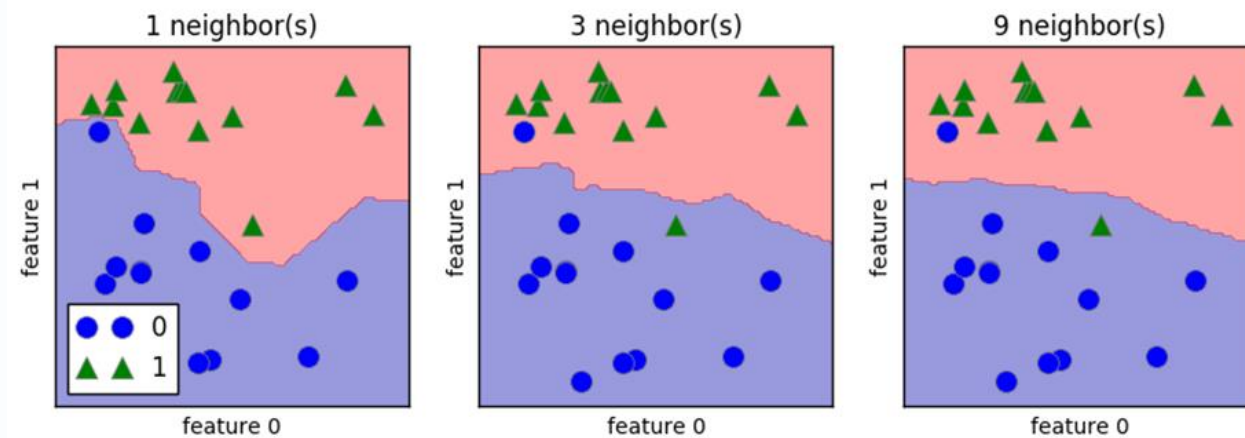
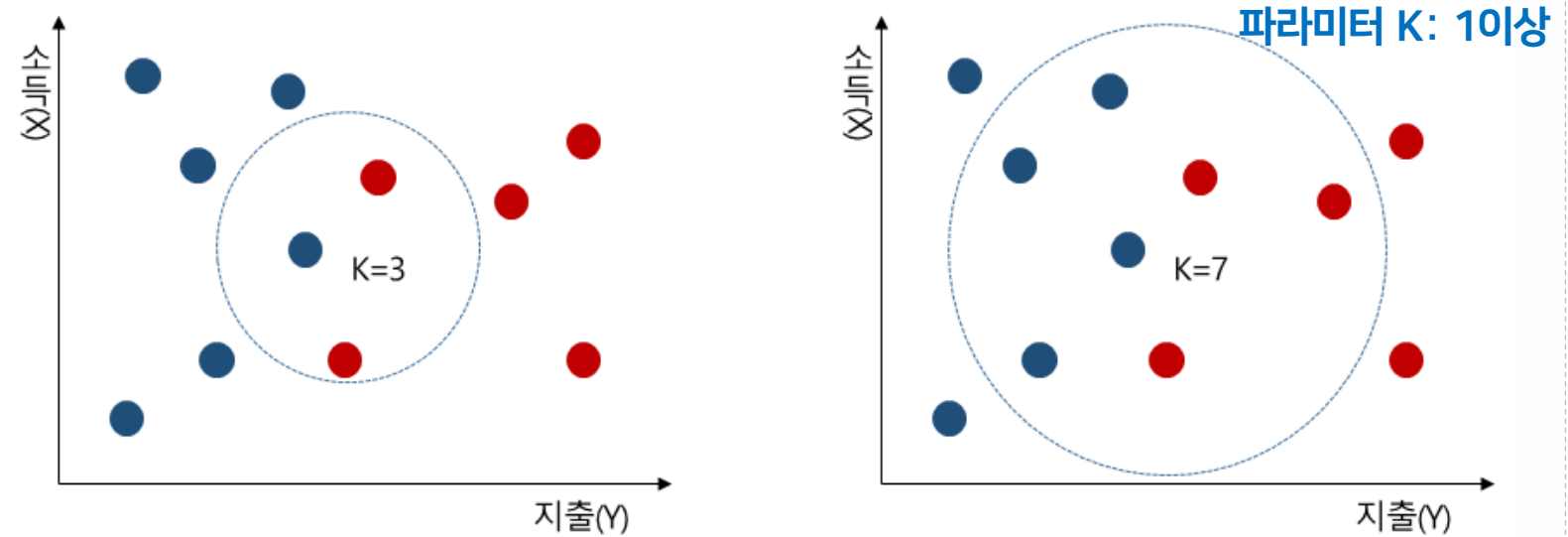
3. 대표 알고리즘

1. 로지스틱 회귀
2. 나이브 베이즈
- 3. 의사 결정 트리**
4. K최근접 이웃
5. 서포트 벡터 머신
6. 랜덤포레스트



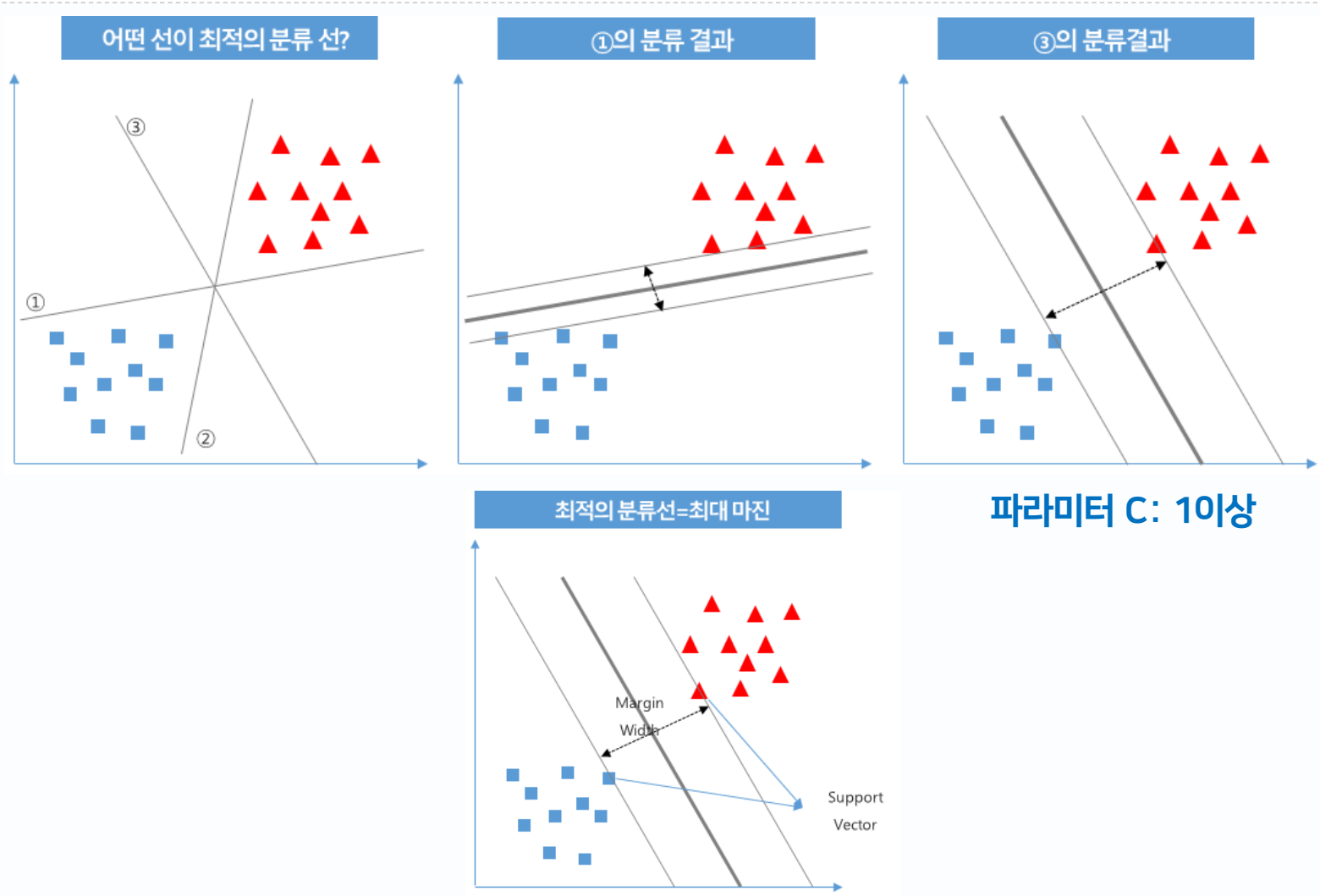
3. 대표 알고리즘

1. 로지스틱 회귀
2. 나이브 베이즈
3. 의사 결정 트리
4. K최근접 이웃
5. 서포트 벡터 머신
6. 랜덤포레스트



3. 대표 알고리즘

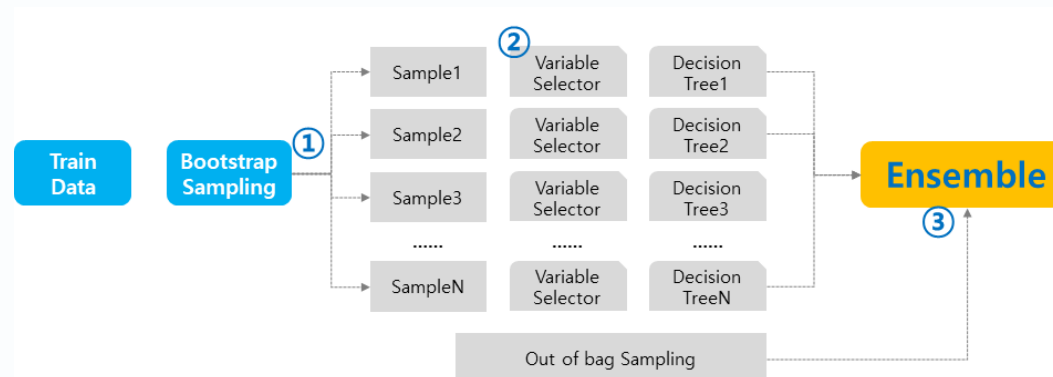
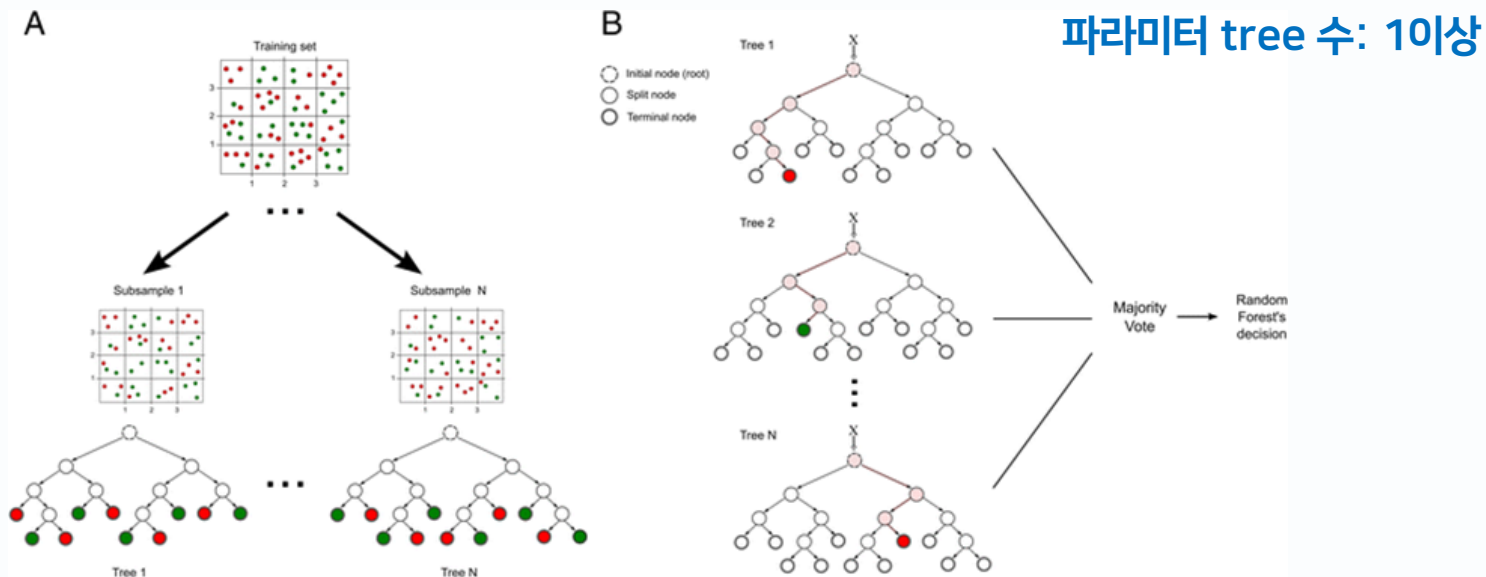
1. 로지스틱 회귀
2. 나이브 베이즈
3. 의사 결정 트리
4. K최근접 이웃
- 5. 서포트 벡터 머신**
6. 랜덤포레스트



3. 대표 알고리즘

1. 로지스틱 회귀
2. 나이브 베이즈
3. 의사 결정 트리
4. K최근접 이웃
5. 서포트 벡터 머신

6. 랜덤포레스트



Part 02

노코딩 머신러닝 솔루션 WEKA 실습

1. 머신러닝 프로세스

1) 개요



주요 과제

- 학습 데이터를 랜덤으로 학습/검증 셋(train/ validation) 분할
- 테스트 셋(test)도 준비
- 데이터의 표준/정규화
- 범주자료 one-hot-encoding
- 특성변수의 축약
- 과제 해결에 적합한 머신러닝 알고리즘 적용
- 평가지표를 통한 모델 평가
- 다양한 하이퍼 파라미터 적용
- 최적의 Hyper Parameter 및 모델 결정

실전 전략

- 학습데이터: 70~90%
- 검증데이터: 10~20%
- 테스트데이터: 10~20%
- 학습 데이터를 그룹으로 나누어서 교차 검증을 하는 방법도 추천
- 표준화(평균 0 / 표준편차 1) 또는 Min-max정규화
- 범주형 특성변수를 0과 1의 값으로 변환
- 고차원의 경우 PCA 방법 등으로 차원 축소
- 회귀/분류/비지도 알고리즘에 데이터 학습
- 학습된 모델에 검증 데이터로 평가
- 정확도 및 과소/과대추정 여부 판단
- 파라미터 조정을 통한 최적 모델 결정
- 최종 분류기에서 검증 셋은 사용하지 않는 것이 좋음
- 최종 모델을 테스트 셋에 대해 성능을 평가
- 테스트 셋에 대한 정확도를 현재 데이터로 학습한 알고리즘 성능으로 제시

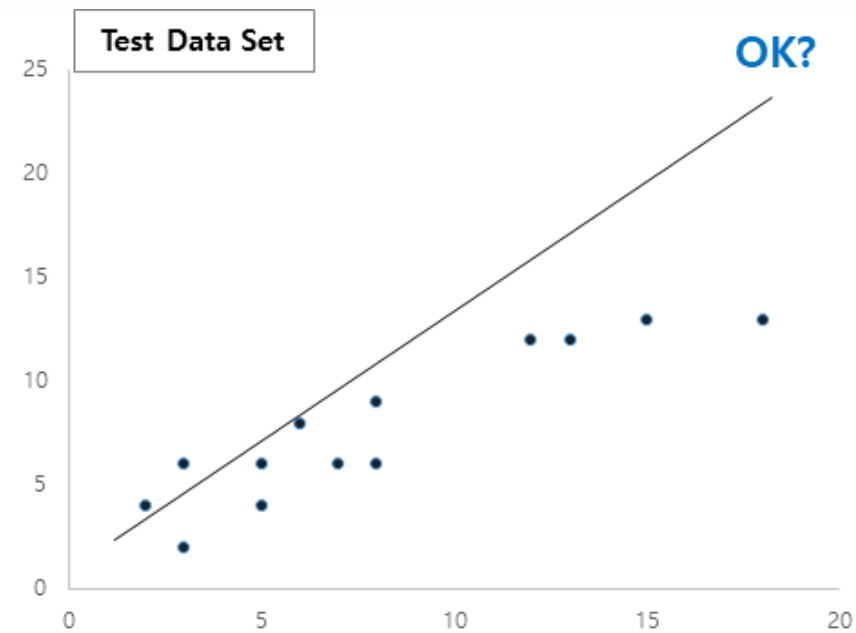
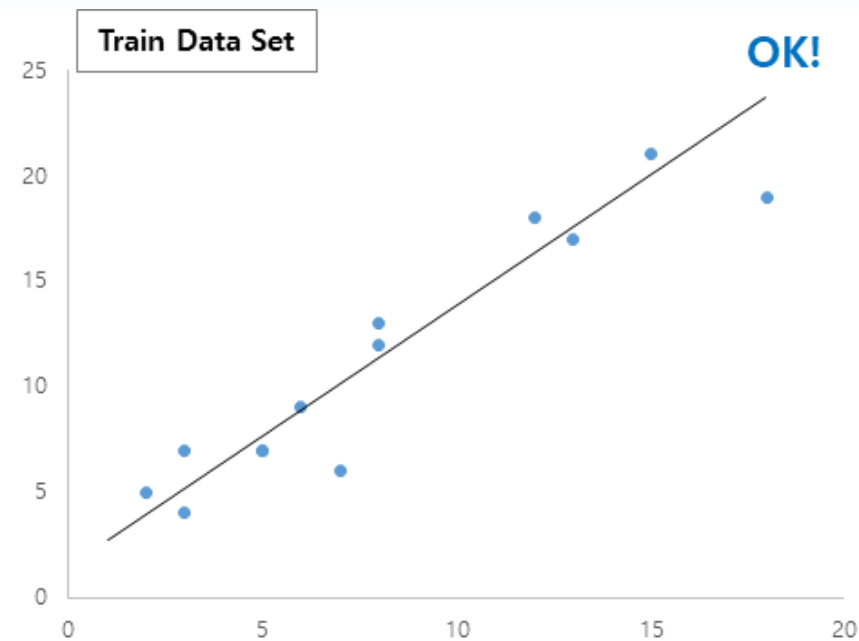
서비스 런칭

신규 유입
데이터에 대한
자동화 솔루션
구축

1. 머신러닝 프로세스

2-1) 데이터셋 분할

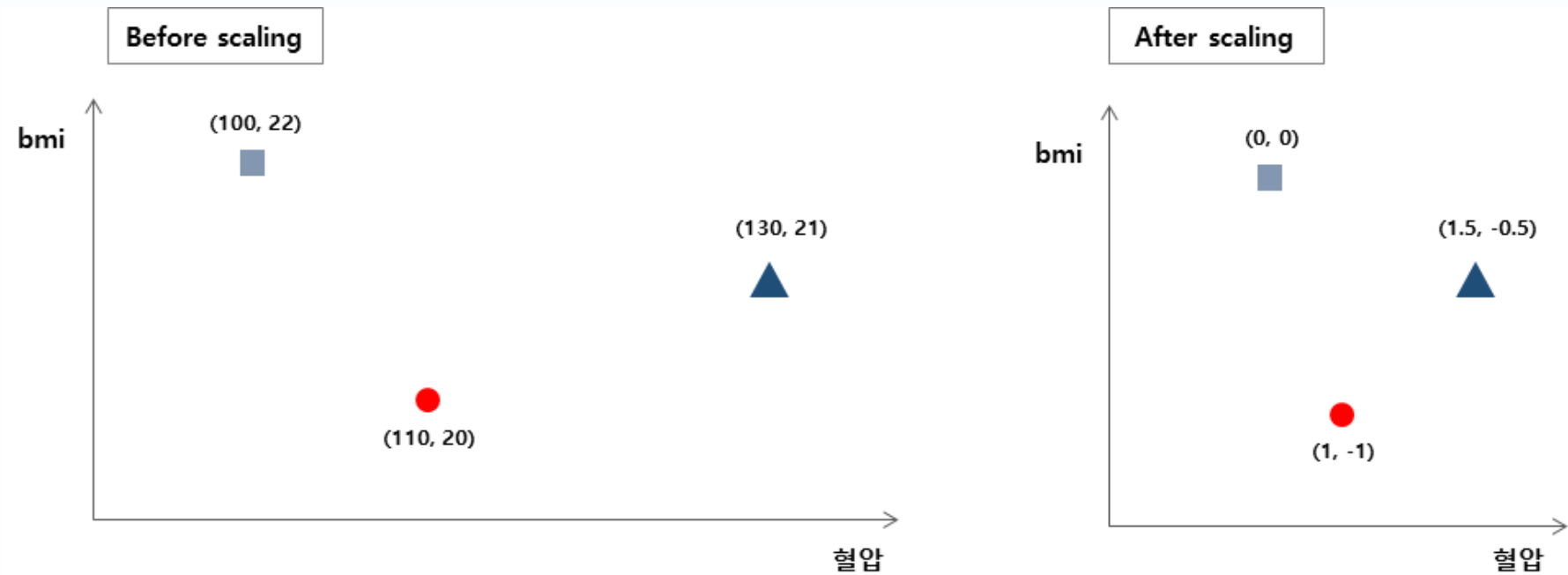
- 전체 분석 데이터 중 머신러닝 모델 알고리즘을 학습시키기 위한 학습데이터(train data)는 70~80%, 학습된 모델이 다른 데이터에도 맞는지를 확인하기 위한 테스트 데이터(test data)는 20~30%가량으로 나누는 것이 일반적



1. 머신러닝 프로세스

2-2) 데이터 전처리

- 목적변수인 레이블(y)은 건드리지 않고, 특성치(X)는 전처리 중 정규화를 반드시 진행해야 함
- 표준화 혹은 민맥스, min-max 방법을 주로 사용함



1. 머신러닝 프로세스

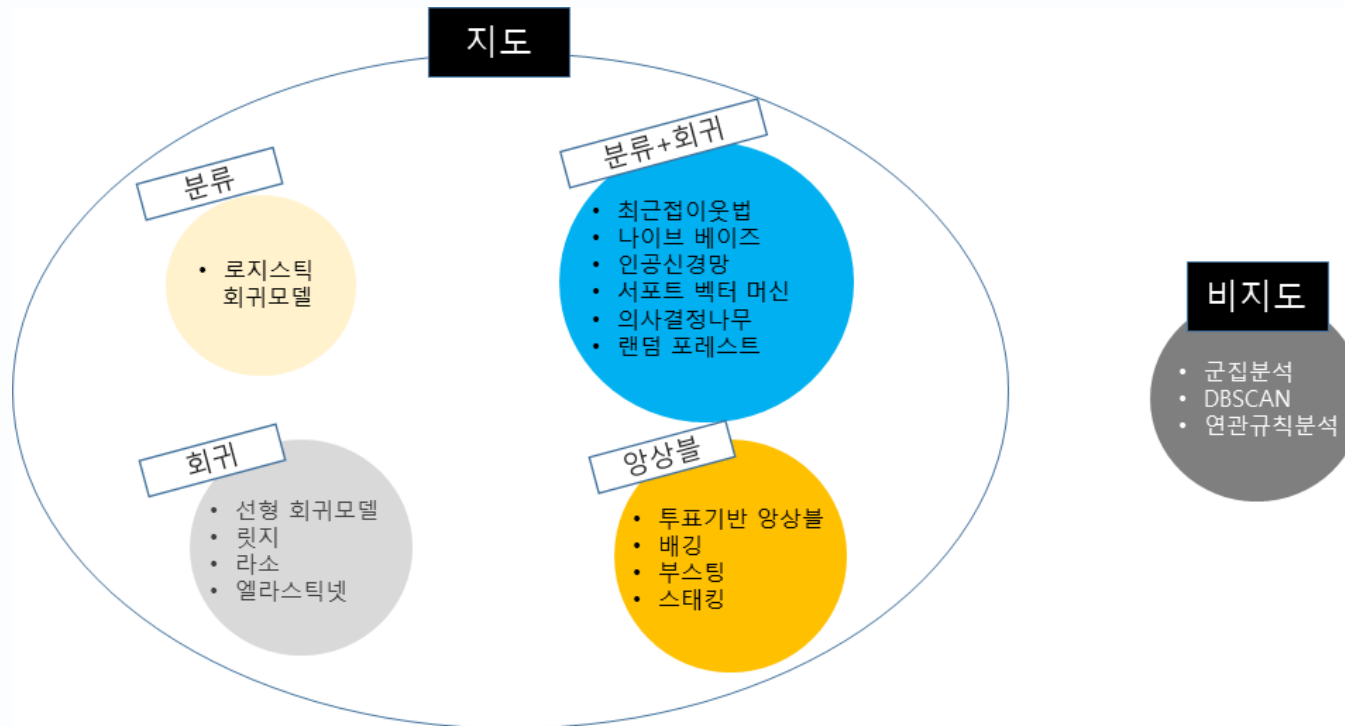
2-2) 데이터 전처리

- 원핫인코딩(one-hot-encoding)
- 특성치에 범주형 자료가 있을 경우 무조건 0과 1로 만드는 과정을 원핫인코딩이라고 함

성별	거주지		성별_남성	성별_여성	거주지_서울	거주지_경기	거주지_지방
1	1		1	0	1	0	0
2	2		0	1	0	1	0
1	3		1	0	0	0	1
2	1		0	1	1	0	0
1	3		1	0	0	0	1
2	2		0	1	0	1	0

1. 머신러닝 프로세스

2-3) 모델 학습



1. 머신러닝 프로세스

2-4) 성능 평가

분류

오차행렬
(confusion matrix)

	0	1
0	90	10
1	20	80

- 정확도/정분류율 (accuracy)
- 오류율(error rate)
- 민감도(sensitivity) / 재현율(recall)
- 특이도(specificity)
- 정밀도(precision)
- F1 Score

회귀

평가지표	계산식	설명
SSE	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	• 오차제곱합. 예측값과 실제값의 차이의 제곱합. 대표적 회귀모형 평가지표
AE	$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$	• 평균오차. 예측값과 평균과의 차이. 예측값들이 평균적으로 미달/초과하는지 확인
MSE	$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$	• 평균제곱오차. 예측오차 제곱합의 평균.
MAE	$\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	• 평균절대오차. 예측오차 절대값들의 평균
RMSE	$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	• 평균제곱근오차. 종속변수와 단위가 같아 설명이 쉽고, 표준편차처럼 예측에 대한 오차 정보 제공 가능
MAPE	$\frac{100}{n} \sum_{i=1}^n \frac{ y_i - \hat{y}_i }{y_i}$	• 평균절대백분율오차. 실제값에 대한 오차의 백분율. 시계열분석에서 예측 성능지표로 자주 활용

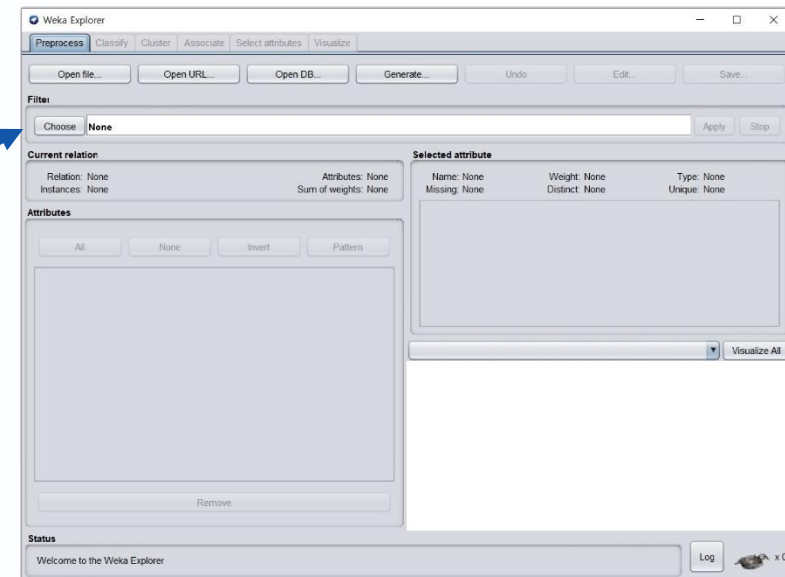
2. WEKA 둘러보기

1) WEKA 구성

WEKA GUI



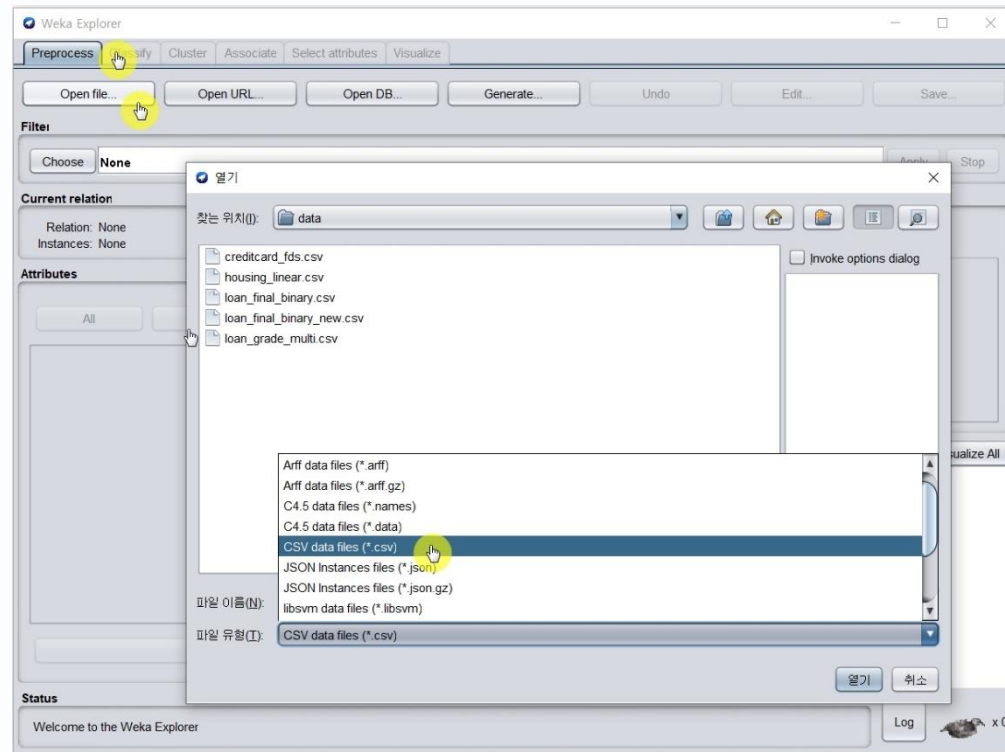
Explorer



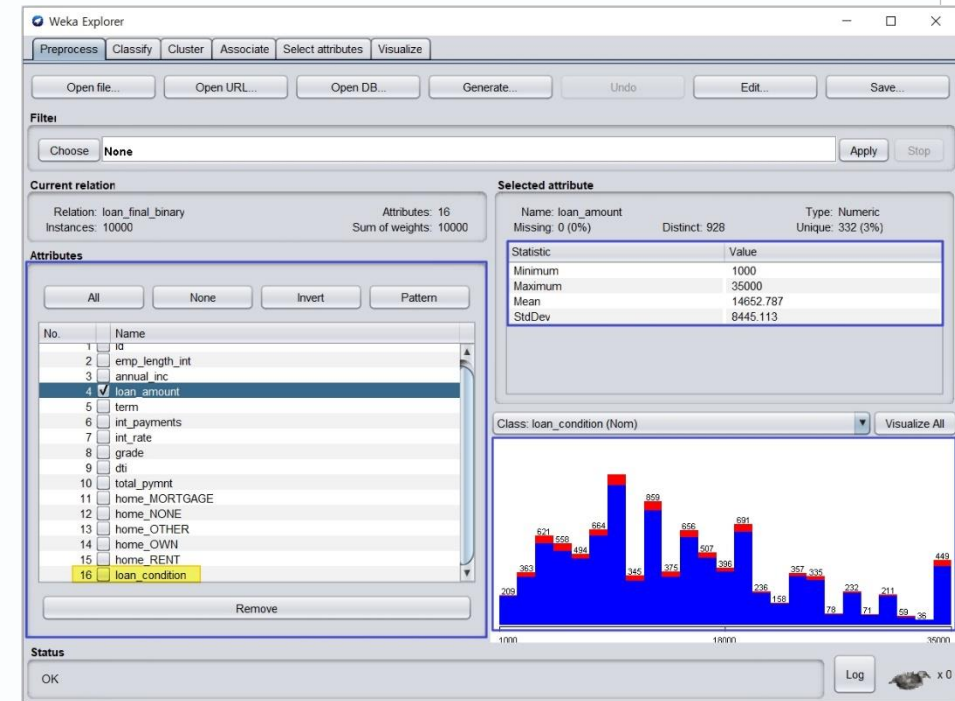
2. WEKA 둘러보기

2) 데이터 불러오기

Preprocess > open file



변수별 탐색



2. WEKA 둘러보기

3) 분석하기

Classification

Classifier

Choose: ZeroR

Test options

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds: 10
- ☐ Percentage split %: 66

(Nom) loan_condition

Start Stop

Result list (right-click for options)

14.03.20 - rules.ZeroR

Classifier output

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      9223      92.23 %
Incorrectly Classified Instances    777      7.77 %
Kappa statistic                    0
Mean absolute error                 0.1434
Root mean squared error            0.2677
Relative absolute error             100 %
Root relative squared error        100 %
Total Number of Instances         10000

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
1.000	1.000	0.922	1.000	0.960	?	0.459	0.922	
0.000	0.000	?	0.000	?	?	0.459	0.077	
Weighted Avg.	0.922	0.922	?	0.922	?	?	0.459	0.856

Confusion Matrix

```

a b <-- classified as
9223 0 | a = Good
777 0 | b = Bad

```

Status: OK

모델 선택

Classifier

Choose: J48 -C 0.25 -M 2

Test options

- ☐ Use training set
- ☐ Supplied test set
- ☒ Cross-validation Folds: 10
- ☐ Percentage split %: 66

(Nom) loan_condition

Start Stop

Result list (right-click for options)

15.37.25 - trees.J48

Classifier output

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      9401      94.01 %
Incorrectly Classified Instances    599      5.99 %
Kappa statistic                    0.9532
Mean absolute error                 0.0969
Root mean squared error            0.2324
Relative absolute error             60.6056 %
Root relative squared error        60.6056 %
Total Number of Instances         10000

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
0.966	0.977	0.933	0.966	0.966	0.966	0.923	0.971	
0.423	0.015	0.715	0.423	0.532	0.523	0.771	0.624	
Weighted Avg.	0.942	0.533	0.935	0.942	0.935	0.923	0.771	

Confusion Matrix

```

a b <-- classified as
9092 131 | a = Good
448 359 | b = Bad

```

Status: OK

분석 결과

3. WEKA 실전

암 예측을 진행해봅시다



THANK YOU

(주)와이즈인컴퍼니 / 서울시 강남구 언주로 309, 기성빌딩 3층 / T 02.558.5144 / F 02.558.5146