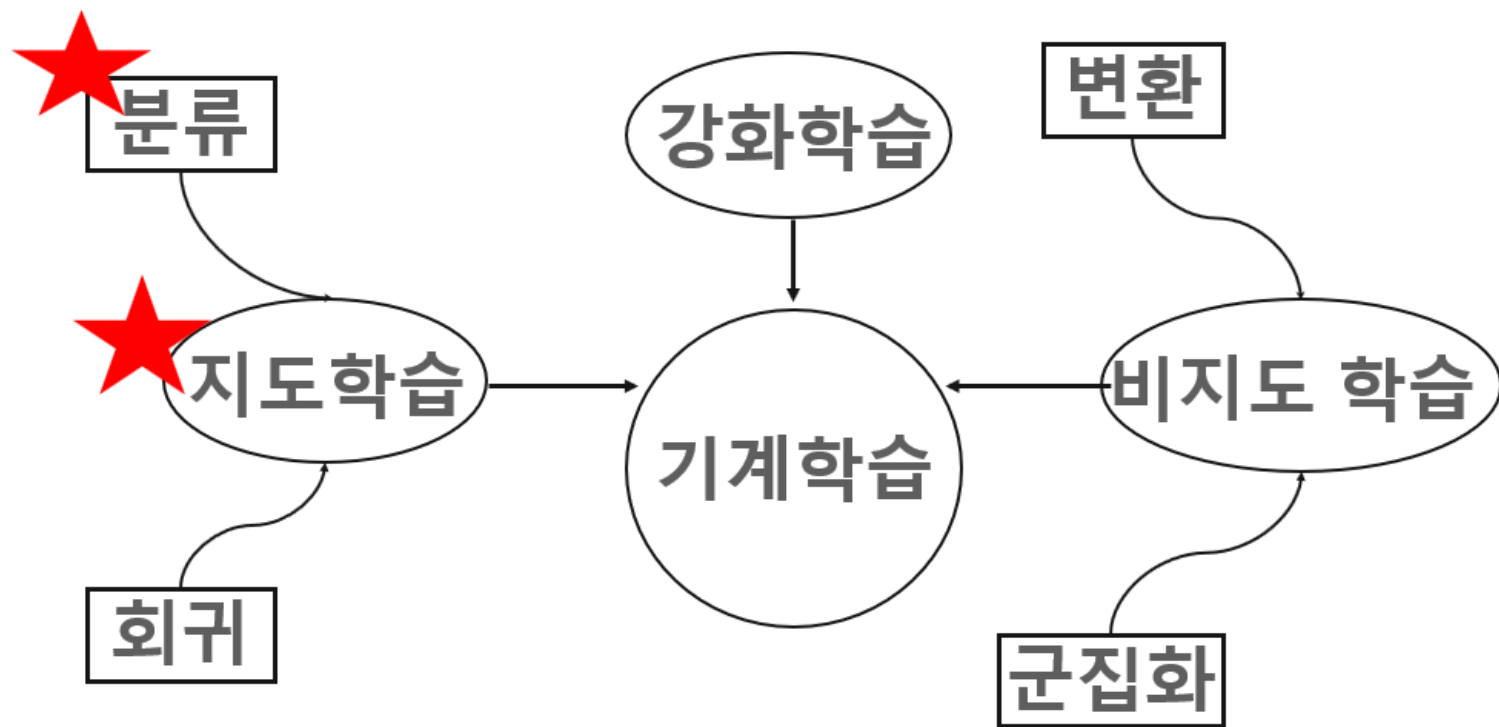


# 지도학습: 분류

홍익 대학교  
Hyun-Sun Ryu

# 머신러닝의 종류



# 분류(Classification)

독립변수와 종속변수를 가지고 있는 과거의 데이터를 학습하여  
‘범주’를 예측할 때 분류 모델을 사용

# 분류의 종류

이진분류

binary classification

다중분류

multinomial classification

# 분류의 종류



## 다중 분류의 전략

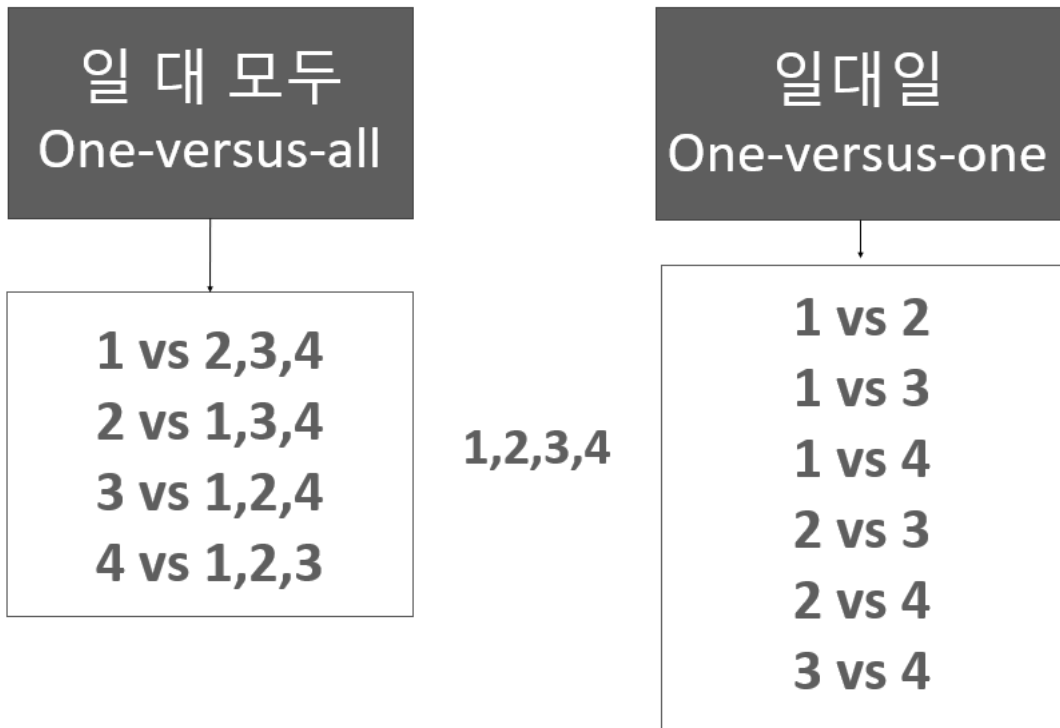
이진분류를 시행하는 이진분류기는  
대부분 다중분류 문제에 적용할 수 없다.

# 다중 분류의 전략

일 대 모두  
One-versus-all

일대일  
One-versus-one

# 다중 분류의 전략





# 분류 모델로 할 수 있는 일

분류 모델은 참/거짓, 저위험/중위험/고위험,  
동물의 종과 같은 레이블을 할당 할 수 있다.

# 분류 모델로 할 수 있는 일

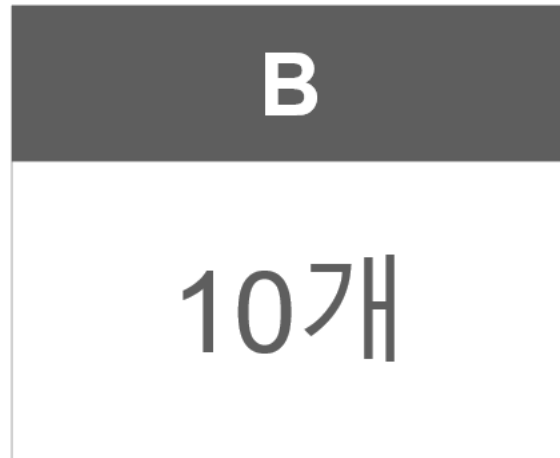
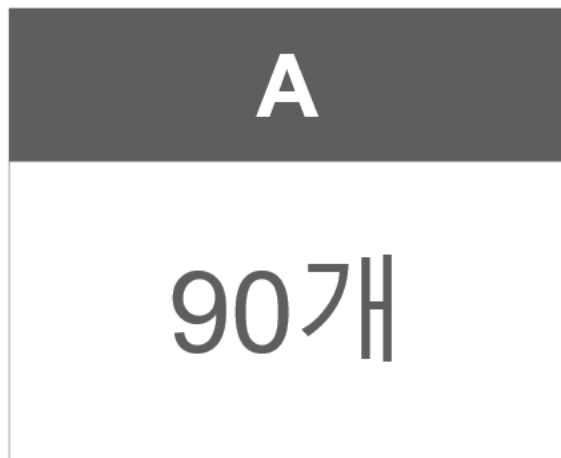
할 수 있는 일	독립변수	종속변수
시험 합격 여부 예측	공부시간, 학원 수강 여부 등	합격/불합격
소고기 등급 판단	<u>고기색</u> , 지방함량 등	1++, 1+, 1, 2
스팸메일 판별	제목, 발신인명 등	참/거짓
<u>암판별</u>	종양 사진, 크기, 두께 등	양성/음성

# 분류(Classification) 모델 평가

# 분류 성능 평가지표

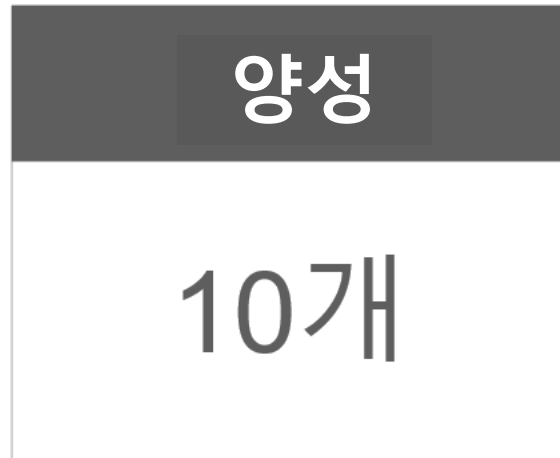
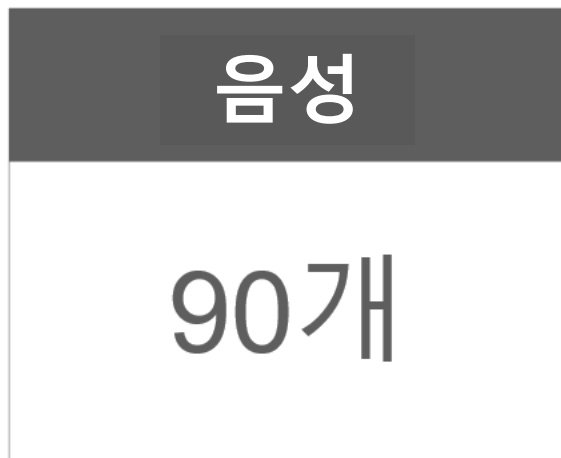
분류 모델의 예측 성능을 계량화하기 위해  
점수화 할 필요성이 있음

## 분류 성능 평가지표



전체 100개

## 분류 성능 평가지표



전체 100명

# 분류 성능 평가지표

어떤 모델이 가장 최적의 모델인지 선택하기 위해서는 분류모델의 성능을 평가하는 지표(score)를 알아야 됨 → 이때 사용하는 것이 **혼동행렬(confusion matrix)**

## 혼동행렬 Confusion Matrix

		Predicted(예측)	
		Negative	Positive
Actual (실제)	Negative	True Negative(TN)	False Positive(FP)
	Positive	False Negative(FN)	True Positive(TP)

# 분류 성능 평가지표

※ 의학에서 양성/음성은 있다/없다의 의미


A(음성)	B(양성)
90개	10개

혼동행렬  
Confusion Matrix

		Predicted(예측)	
		음성	양성
Actual (실제)	음성	TN : 음성으로 예측했는데 맞은 경우 <b>True Negative</b>	FP: 양성으로 예측했는데 틀린 경우 <b>False Positive</b>
	양성	FN : 음성으로 예측했는데 틀린 경우 <b>False Negative</b>	TP: 양성으로 예측했는데 맞은 경우 <b>True Positive</b>



# 분류 성능 평가지표

위젯	설명	입력	출력
 Confusion Matrix	분류자 평가 결과에서 생성된 혼동 행렬을 표시한다.	Evaluation Results	Selected Data, Data

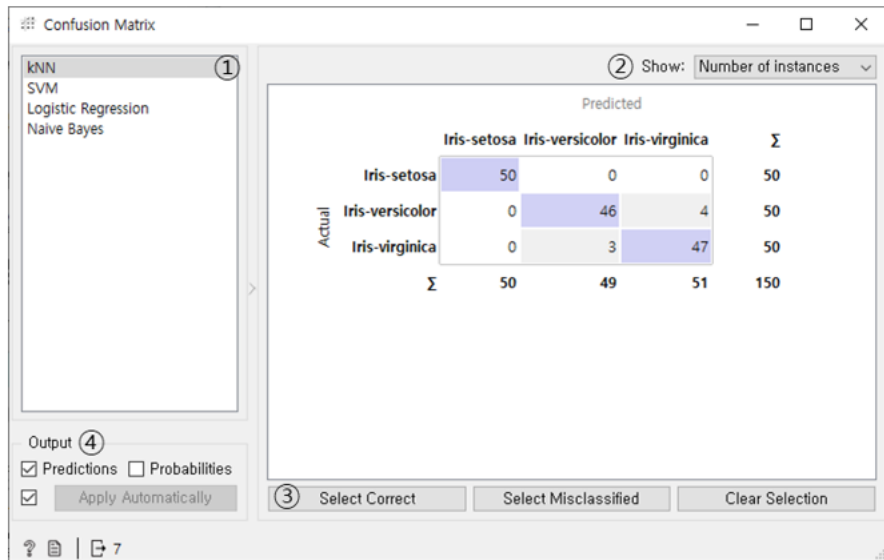
- Confusion Matrix 위젯은 예측 클래스와 실제 클래스 사이의 인스턴스 수/비율을 제공
- 매트릭스에서 요소를 선택하면 해당 인스턴스가 출력 신호에 공급
- 이렇게 하면 어떤 특정 사례가 잘못 분류되었고 어떻게 분류되었는지 관찰

# 분류 성능 평가지표



Confusion Matrix

## 혼동행렬 Confusion Matrix



① 평가 결과에 여러 학습 알고리즘에 대한 데이터가 포함된 경우 학습자 상자에서 하나를 선택해야 한다.

표시에서 행렬에서 보려는 데이터를 선택한다.

Number of instances: 인스턴스 수가 올바르게 잘못 분류된 인스턴스를 숫자로 표시한다.

② Proportions of predicted: 예측된 비율에 따라 실제 클래스가 있는 분류된 인스턴스 수가 표시된다.

Proportions of actual: 실제의 비율은 반대 관계를 보여준다.

③ 선택에서 원하는 출력을 선택할 수 있다.

④ 선택한 인스턴스를 보낼 때 해당 옵션인 예측 또는 확률을 선택하면 위젯이 예측 클래스나 확률과 같은 새 속성을 추가할 수 있다.

# 분류 성능 평가지표

Q: 왜 이렇게 많은 평가지표가 필요한가요?

A: 한 가지 지표만 사용하기에는 한 가지씩 부족한 점이 있기 때문에 여러 개의 지표를 동시에 평가하여 전체적으로 모델을 평가

정밀도  
precision

재현율  
recall

F1 점수  
F1 score

정확도  
accuracy

ROC

AUC

# 분류 성능 평가지표

정확도  
accuracy

정밀도  
precision

재현율  
recall

F1 점수  
F1 score

ROC

AUC  
Area Under The Curve

		Predicted(예측)	
		음성	양성
Actual (실제)	음성	TN : 음성으로 예측했는데 맞은 경우	FP: 양성으로 예측했는데 틀린 경우
	양성	FN : 음성으로 예측했는데 틀린 경우	TP: 양성으로 예측했는데 맞은 경우

$$\text{정확도 (CA)} = \frac{TP+TN}{TP+FN+FP+TN}$$

- 전체 중 실제 TRUE 를 TRUE 라 하고, 실제 FALSE를 FALSE 라고 예측한 것의 비율
- 전체 중에서 정답을 맞춘 비율
- 참, 거짓에 상관없이 정답을 맞춘 비율로 1에 가까울수록 좋음.
- 데이터에 따라 부정확할 수 있음.

# 분류 성능 평가지표

정확도  
accuracy

정밀도  
precision

재현율  
recall

F1 점수  
F1 score

ROC

AUC  
Area Under The Curve

		Predicted(예측)	
		음성	양성
Actual(실제)	음성	TN : 음성으로 예측했는데 맞은 경우	FP: 양성으로 예측했는데 틀린 경우
	양성	FN : 음성으로 예측했는데 틀린 경우	TP: 양성으로 예측했는데 맞은 경우

$$\text{정밀도} = \frac{\text{참 양성}(TP)}{\text{참 양성}(TP) + \text{거짓 양성}(FP)}$$

- **TRUE** 라고 예측한 것 중에서 실제 **TRUE** 인 것의 비율
- 긍정적인 예에 집중하고 얼마나 모델이 긍정적인 점수를 잘 예측하는지 측정하는 지표
- 이진분류기에서 양성이라고 예측한 것 중 실제로 양성인 경우의 비율을 구하는 것
- **PPV(Positive predictive value)**라고도 불리며 1에 가까울 수록 좋음.

# 분류 성능 평가지표

정확도  
accuracy

정밀도  
precision

재현율  
recall

F1 점수  
F1 score

ROC

AUC  
Area Under The Curve

		Predicted(예측)	
		음성	양성
Actual(실제)	음성	TN : 음성으로 예측했는데 맞은 경우	FP: 양성으로 예측했는데 틀린 경우
	양성	FN : 음성으로 예측했는데 틀린 경우	TP: 양성으로 예측했는데 맞은 경우

$$\text{재현율} = \frac{\text{참 양성}(TP)}{\text{참 양성}(TP) + \text{거짓음성}(FN)}$$

- 실제 TRUE인 경우 중 TRUE 로 예측한 비율
- **Sensitivity** 또는 **high rate**라고도 불림
- 재현율의 값이 1에 가까울 수록 좋음.
- 암환자 판별, 보험사기 적발과 같이 실제 Positive 데이터를 **Negative**로 잘못 판단하면 큰 영향이 있는 경우 재현율이 중요
- 정밀도와 재현율은 서로 반비례하는 경향

# 분류 성능 평가지표

정확도  
accuracy

정밀도  
precision

재현율  
recall

F1 점수  
F1 score

ROC

AUC  
Area Under The Curve

$$F1 = 2 * \frac{1}{\frac{1}{\text{정밀도}} + \frac{1}{\text{재현율}}} = 2 * \frac{\text{정밀도} * \text{재현율}}{\text{정밀도} + \text{재현율}}$$

<재현율 = 1, 정밀도 = 0.01일 때, >

1. 산술평균

$$(1+0.01) / 2 = 0.505$$

2. 조화평균

$$2 * \frac{1 * 0.01}{1 + 0.01} = 0.019$$

F1: 정밀도(precision)과 재현율(recall)의 조화평균

# 분류 성능 평가지표

정확도 accuracy	정밀도 precision	재현율 recall	<b>F1 점수 F1 score</b>	ROC	AUC Area Under The Curve
-----------------	------------------	---------------	---------------------------	-----	-----------------------------

- **정밀도**와 **재현율**이 각각 가지고 있는 **단점을 보완**하기 위해 제시된 지표가 **F1 점수**
- F1 점수는 정밀도와 재현율의 조화평균 값. 즉, **F1 점수가 높아야 성능이 좋음**
- 산술평균을 사용하지 않고 조화평균을 사용하는 까닭은 **정밀도와 재현율 중 하나가 0에 가깝게 낮은 지표를 나타낼 때 그 지표를 잘 반영할 수 있기 때문**
- 예를 들어 재현율이 1이고 정밀도가 0.01일 때 산술평균은 0.505로 50%에 가까운 예측력을 표현하지만 조화평균에서는 0.019로 매우 낮은 것을 알 수 있음.



# 분류 성능 평가지표

		Positive	Negative	
Actual	Positive	True Positive (TP)	False Negative (FN) Type II Error	재현율(RECALL) $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		정밀도(PRECISION) $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	정확도(ACCURACY) $\frac{TP + TN}{(TP + TN + FP + FN)}$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$



정밀도와 재현율의 조화평균

# 분류 성능 평가지표

- ROC(Receiver Operating Characteristic) 곡선은 이진분류시스템에 대한 성능평가 기법
- ROC 곡선은 참 양성비율과 허위 양성비율 두개의 매개변수를 표시
- 참 양성비율은 실제 양성 중 양성으로 예측한 비율로 **재현율**과 같음.
- 허위 양성비율은 실제 음성 중 양성으로 잘못 예측한 비율

정확도 accuracy	정밀도 precision	재현율 recall	F1 점수 F1 score	ROC	AUC Area Under The Curve
-----------------	------------------	---------------	-------------------	-----	-----------------------------

		Predicted(예측)	
		음성	양성
Actual (실제)	음성	TN : 음성으로 예측했는데 맞은 경우	FP : 양성으로 예측했는데 틀린 경우
	양성	FN : 음성으로 예측했는데 틀린 경우	TP : 양성으로 예측했는데 맞은 경우

$$\text{참 양성비율(TPR)} = \frac{TP}{TP+FN}$$

$$\text{허위 양성비율(FPR)} = \frac{FP}{FP+TN}$$

Fall-out(False Positive Rate): 실제 False인 data 중에서 모델이 TRUE 라고 예측한 비율  
 ROC curve: 여러 임계치들을 기준으로 recall-fallout의 변화를 시각화 한 것  
 AUC: ROC 그래프 아래의 면적

# 분류 성능 평가지표

세가지 ROC 곡선에서 빨간색 ROC곡선은 예측 정확도가 50%에 가까운 곡선으로 예측 성능이 가장 나쁜 경우이고 곡선이 굽어지면 굽어질수록 AUC가 넓어지므로 더욱 정확한 모델임

정확도  
accuracy

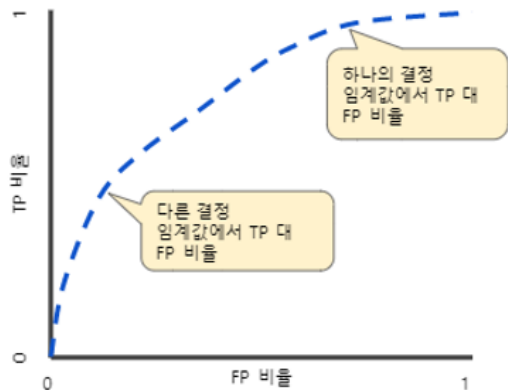
정밀도  
precision

재현율  
recall

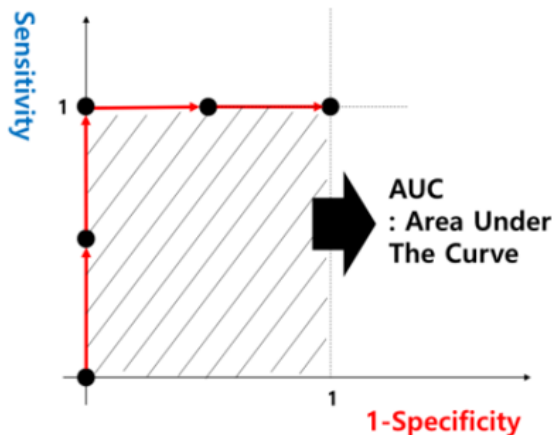
F1 점수  
F1 score

ROC

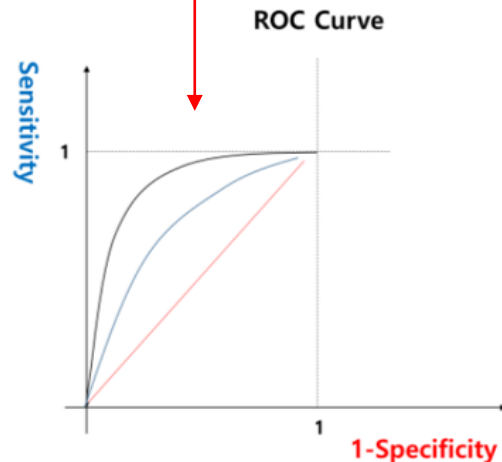
AUC  
Area Under The Curve



ROC 곡선



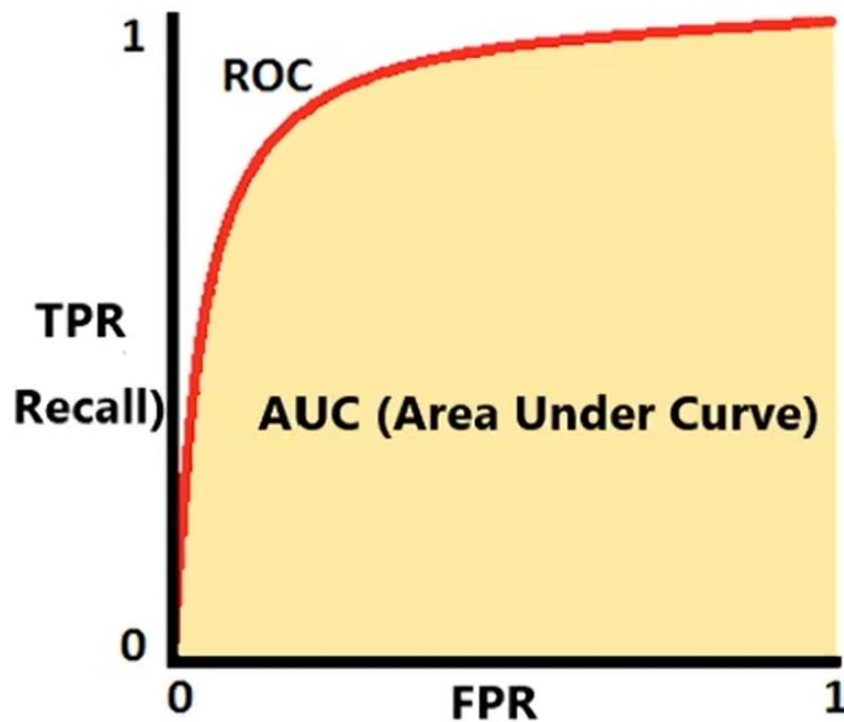
AUC



세가지 ROC 곡선

ROC 그래프의 아래의 면적. 1에 가까울 수록 정확도가 높음.

# 분류 성능 평가지표



# 오렌지3에서 분류 평가지표 확인하기



Datasets

titanic

Title	Size	Instances	Variables	Target	Tags
Titanic	44.1 KB	2201	4	categorical	

Description

**Titanic**

This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner Titanic, summarized according to economic status (class), sex, age and survival.

**See Also**

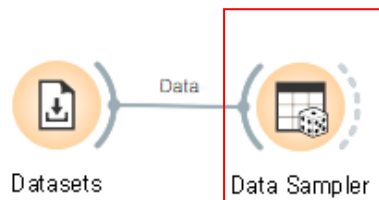
[Nomogram - Visualizing Probabilities.](#)

**References**

Dawson Robert J. MacG. (1995) The 'Unusual Episode' Data Revisited. Journal of Statistics Education 3(3).

? | 2201

## 오렌지3에서 분류 평가지표 확인하기



**Data Sampler** ? X

**Sampling Type** ①

☒ Fixed proportion of data:  
 70 %

☐ Fixed sample size  
 Instances: 1  
☐ Sample with replacement

☐ Cross validation  
 Number of subsets: 10  
 Unused subset: 1

☐ Bootstrap

**Options** ②

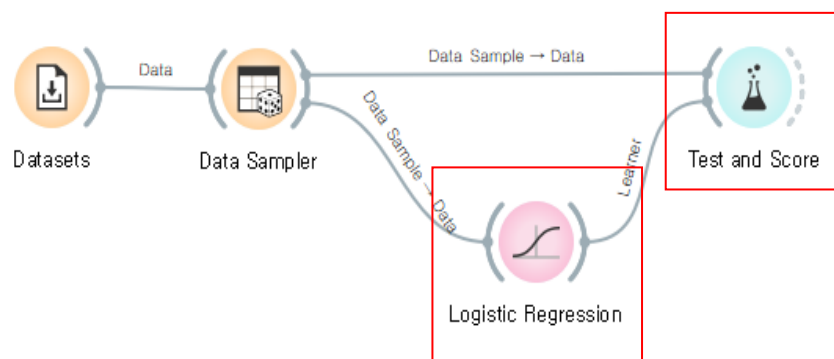
☒ Replicable (deterministic) sampling  
☐ Stratify sample (when possible)

**Sample Data** ③

150 105

① Sampling Type	Fixed proportion	데이터가 전체 데이터의 선택된 백분율 (예: 전체 데이터의 70%)을 반환한다.
	Fixed sample size	고정 샘플 크기는 선택한 수의 데이터 인스턴스를 반환하며, 항상 전체 데이터 세트에서 샘플을 추출한다(하위 집합에 이미 있는 인스턴스를 빼지 않음). 교체를 사용하면 입력 데이터 세트에서 사용할 수 있는 것보다 많은 인스턴스를 생성할 수 있다.
	Cross validation	데이터 인스턴스를 지정된 수의 하위 집합으로 분할한다. 일반적인 유효성 검사 스키마를 따라 사용자가 선택한 하위 세트를 제외한 모든 하위 집합이 데이터 샘플로 출력되고 선택한 하위 집합은 나머지 데이터로 이동한다.
	Bootstrap	모집단 통계에서 표본을 추출한다.
② Options	Replicable sampling	사용자 간에 전달할 수 있는 샘플링 패턴을 유지한다.
	Stratify sample	입력 데이터 세트의 구성을 모방한다.
③ Sample Data		작업을 마치고 데이터 샘플을 출력하려면 데이터 샘플을 누른다.

# 오렌지3에서 분류 평가지표 확인하기



Test and Score

Sampling

- ☒ Cross validation
  - Number of folds: 10
  - ☒ Stratified
- ☐ Cross validation by feature
- ☐ Random sampling
  - Repeat train/test: 100
  - Training set size: 50 %
  - ☒ Stratified
- ☐ Leave one out
- ☐ Test on train data
- ☐ Test on test data

Target Class

(Average over classes)

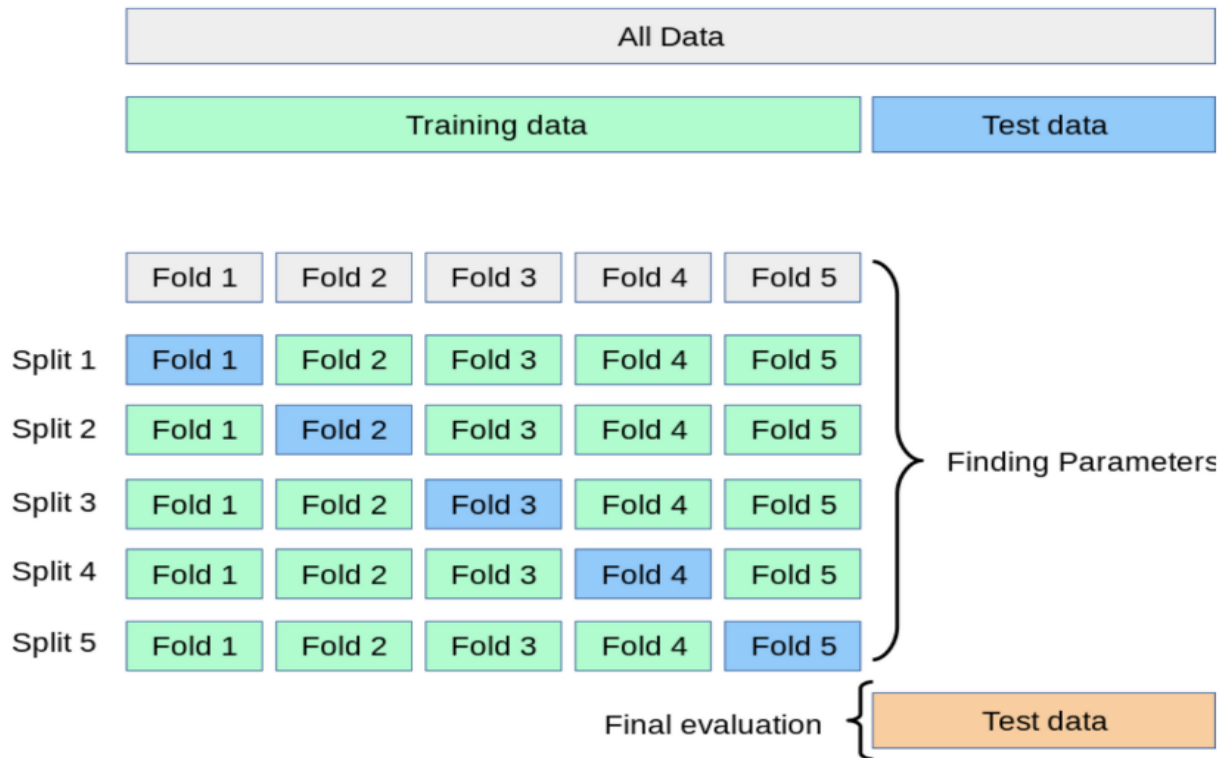
Model Comparison

Area under ROC curve

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.740	0.779	0.765	0.772	0.779

# K-겹 교차검증(K-Fold Cross Validation)

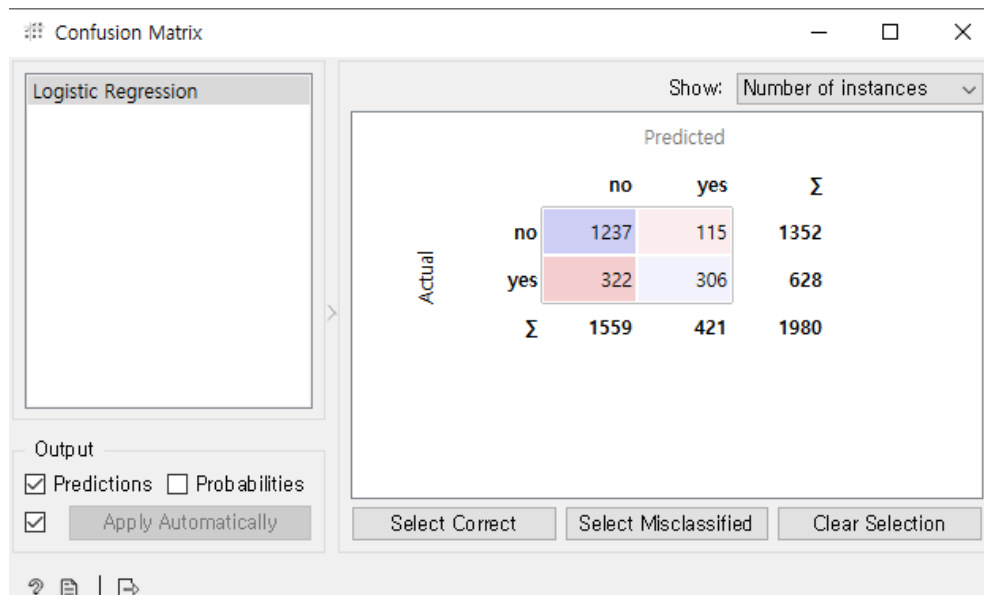
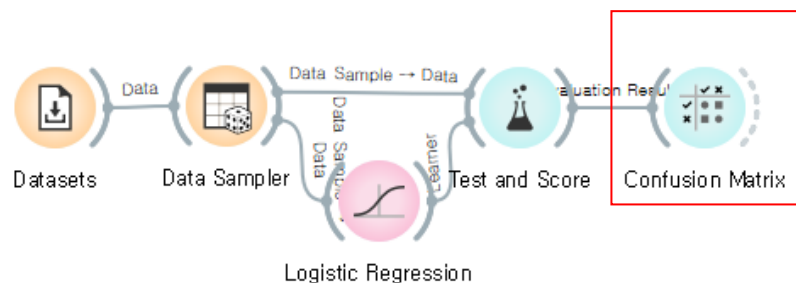




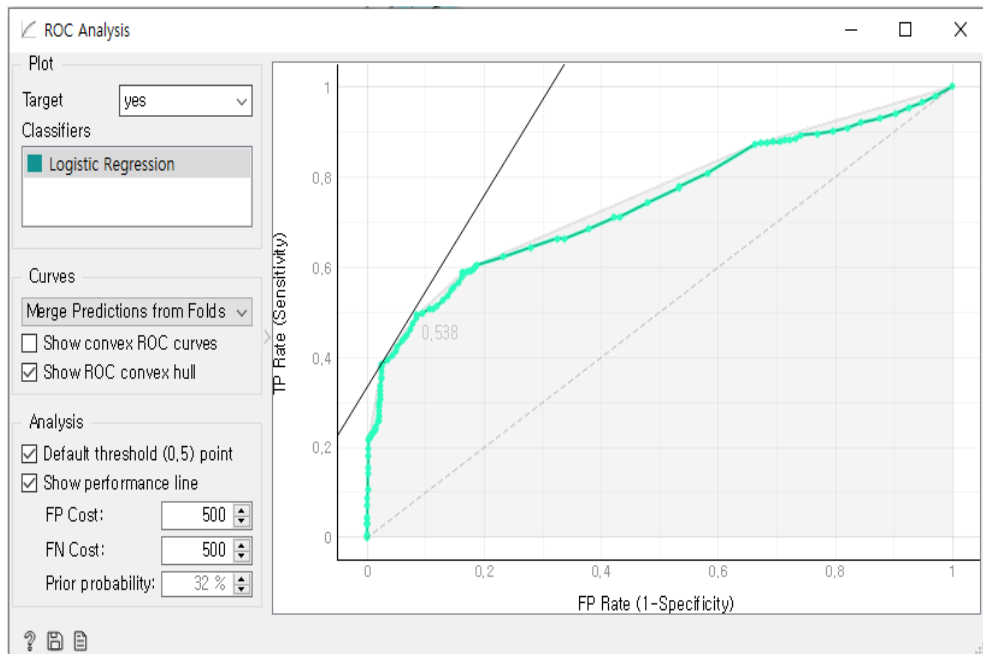
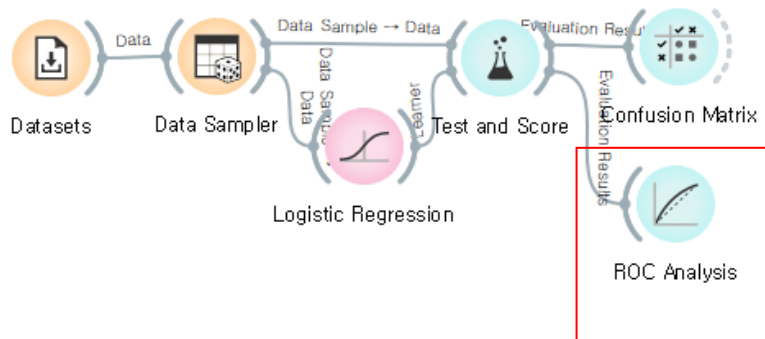
# K-겹 교차검증(K-Fold Cross Validation)

- K-Fold는 가장 일반적으로 사용되는 교차 검증 방법
- 보통 회귀 모델에 사용되며, 데이터가 독립적이고 동일한 분포를 가진 경우에 사용
- 일반적인 K-Fold 교차 검증 과정은 다음과 같음.
  - 전체 데이터셋을 Training Set과 Test Set으로 나눔
  - Training Set를 Training Set + Validation Set으로 사용하기 위해 k개의 폴드로 나눔
  - 첫 번째 폴드를 Validation Set으로 사용하고 나머지 폴드들을 Training Set으로 사용
  - 모델을 Training한 뒤, 첫 번째 Validation Set으로 평가
  - 차례대로 다음 폴드를 Validation Set으로 사용하며 3번을 반복
  - 총 k 개의 성능 결과가 나오며, 이 k개의 평균을 해당 학습 모델의 성능이라 함.

# 오렌지3에서 분류 평가지표 확인하기



# 오렌지3에서 분류 평가지표 확인하기



# 분류(Classification) 알고리즘

# 아이리스 데이터 세트

- 아이리스는 꽃잎의 모양과 길이에 따라 여러 가지 품종으로 나뉨
- 사진을 보면 품종마다 비슷해 보이는데 과연 딥러닝을 사용하여 이들을 구별해 낼 수 있을까?



Iris-virginica



Iris-setosa



Iris-versicolor

# 아이리스 데이터 세트

		속성				클래스
		정보 1	정보 2	정보 3	정보 4	품종
샘플	1번째 아이리스	5.1	3.5	4.0	0.2	Iris-setosa
	2번째 아이리스	4.9	3.0	1.4	0.2	Iris-setosa
	3번째 아이리스	4.7	3.2	1.3	0.3	Iris-setosa
	...	...	...	...	...	...
	150번째 아이리스	5.9	3.0	5.1	1.8	Iris-virginica

<표> 아이리스 데이터의 샘플, 속성, 클래스 구분

# 아|이|스 데|이|터 세|트

- 샘플 수: 150
- 속성 수: 4
  - 정보 1: 꽃받침 길이 (sepal length, 단위: cm)
  - 정보 2: 꽃받침 너비 (sepal width, 단위: cm)
  - 정보 3: 꽃잎 길이 (petal length, 단위: cm)
  - 정보 4: 꽃잎 너비 (petal width, 단위: cm)
- 클래스: Iris-setosa, Iris-versicolor, Iris-virginica

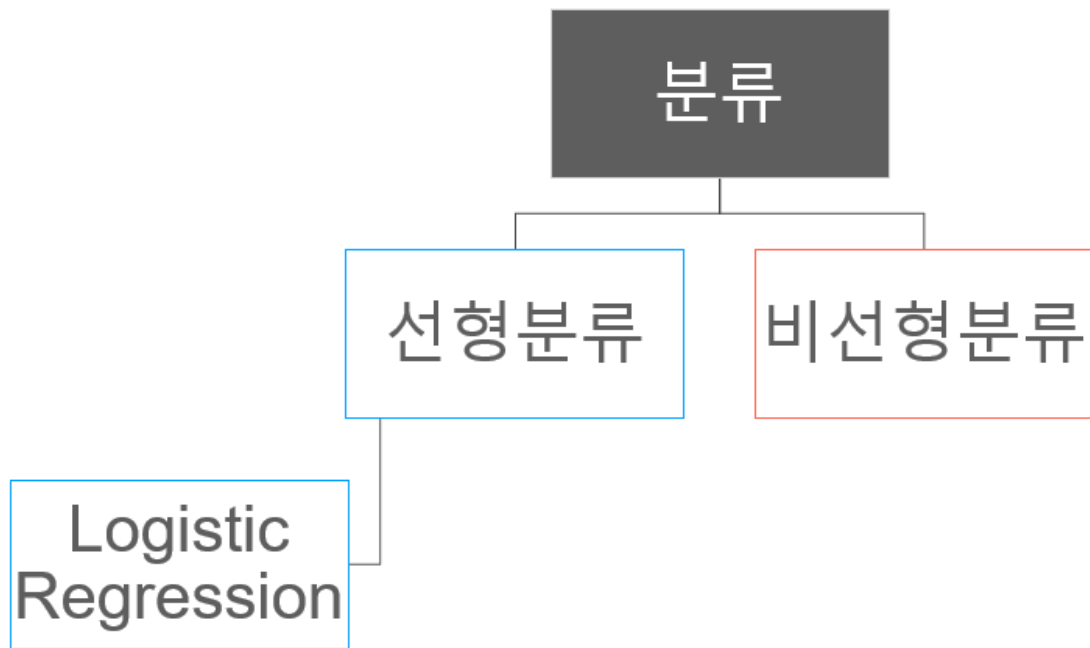
[illegible]

# 분류 모델





# 분류 모델



# 분류 모델

Logistic  
Regression

SVM

KNN

Decision tree

Random forest

회귀를 사용하여 데이터가 **어떤 범주에 속할 확률을 0에서 1 사이의 값으로 예측**하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해주는 지도학습 알고리즘

# 분류 모델

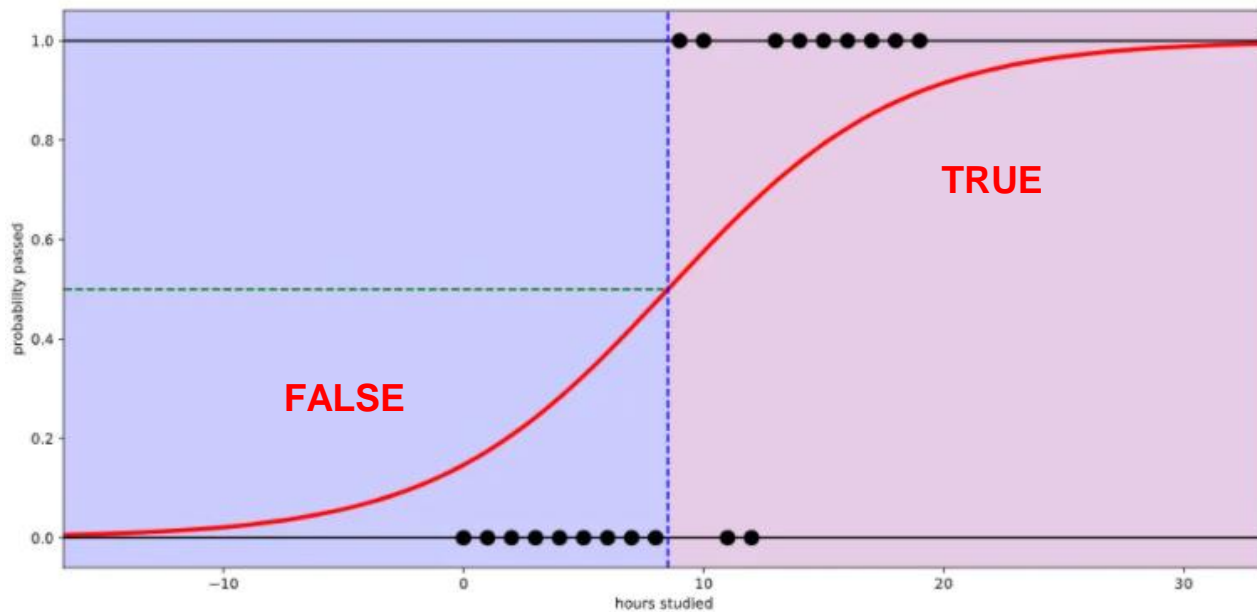
Logistic  
Regression

SVM

KNN

Decision tree

Random forest



# 분류 모델


Logistic  
Regression

SVM

KNN

Decision tree

Random forest

위젯	설명	입력	출력
 Logistic Regression	L1(LASSO) 또는 L2(리지) 정규화를 사용한 로지스틱 회귀 분류 알고리즘이다.	Data, Preprocessor	Learner, Model, Coefficients

- Logistic Regression 위젯은 데이터에서 로지스틱 회귀 모형을 학습
- 분류 작업**에만 사용

# 분류 모델

Logistic  
Regression

SVM

KNN

Decision tree

Random forest

The screenshot shows the 'Logistic Regression' widget interface. It has a title bar with a question mark and a close button. The main area contains a 'Name' field with the text 'Logistic Regression' and a circled '1' next to it. Below this is a 'Regularization type' dropdown menu set to 'Ridge (L2)' with a circled '2' next to it. Underneath is a 'Strength' slider ranging from 'Weak' to 'Strong', with a blue bar indicating the current position and the text 'C=1' below it. At the bottom, there is a checked checkbox and a button labeled 'Apply Automatically'. The bottom status bar contains icons for help, a document, and a refresh button.

① Name

다른 위젯에 표시할 이름으로 기본 이름은 Logistic Regression이다.

② Regularization type

정규화 유형(L1 또는 L2)과 cost 강도를 설정한다(기본값: C=1).

# 분류 모델

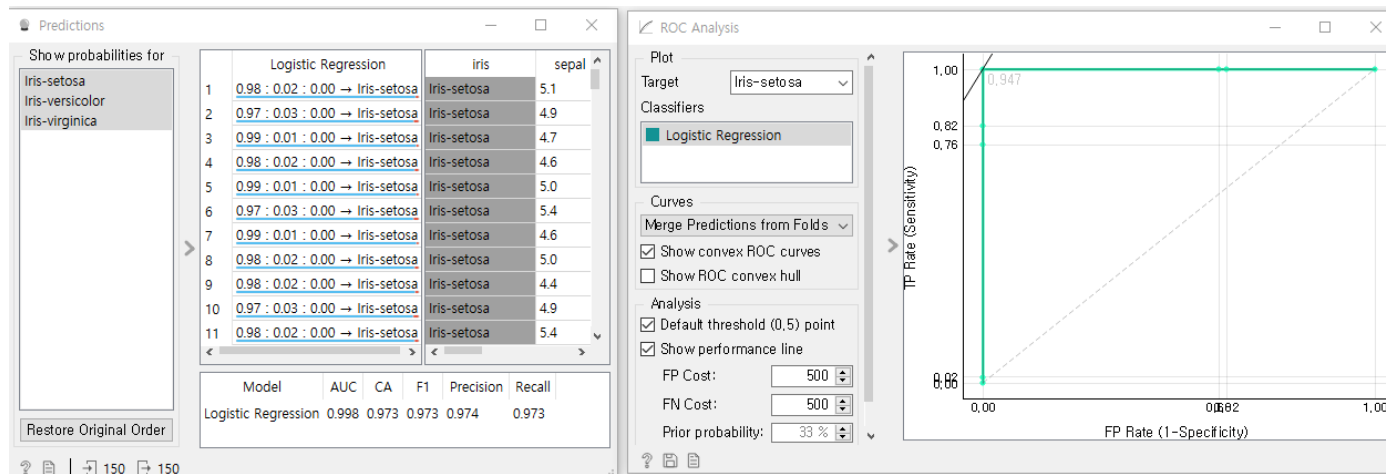
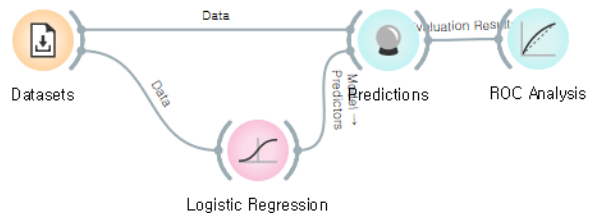
Logistic  
Regression

SVM

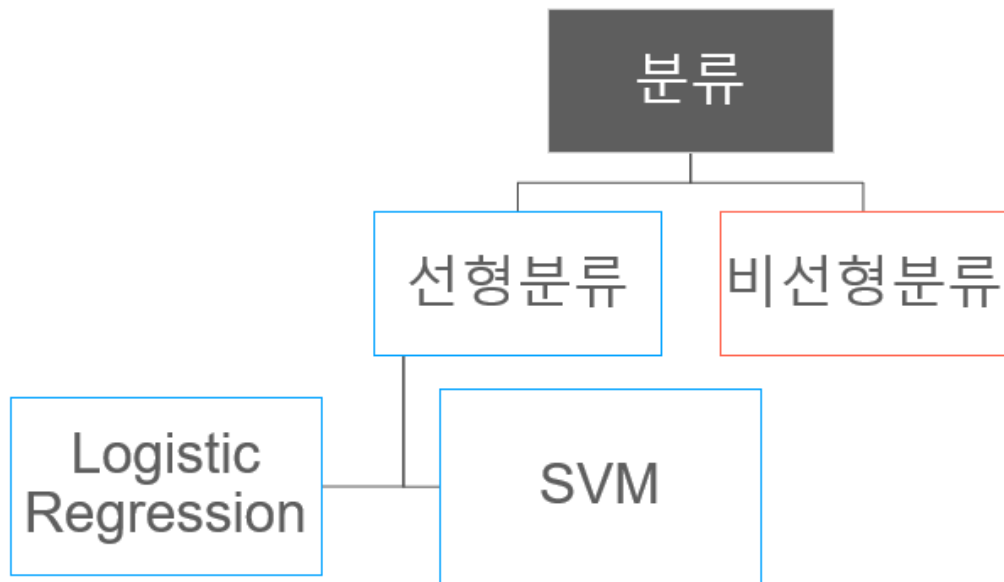
KNN

Decision tree

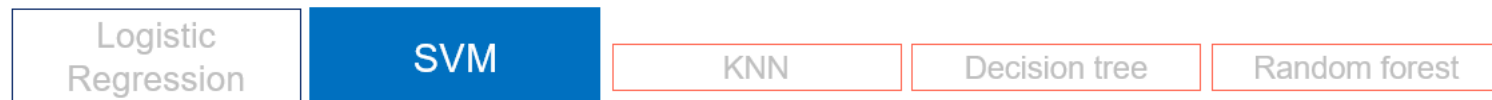
Random forest



# 분류 모델



# 분류 모델



- Support Vector Machine은 **분류나 회귀 모두 사용 가능한 지도학습 알고리즘**
- SVM은 결정경계(decision Boundary)를 어떻게 정의하고 계산하는지가 중요



# 분류 모델

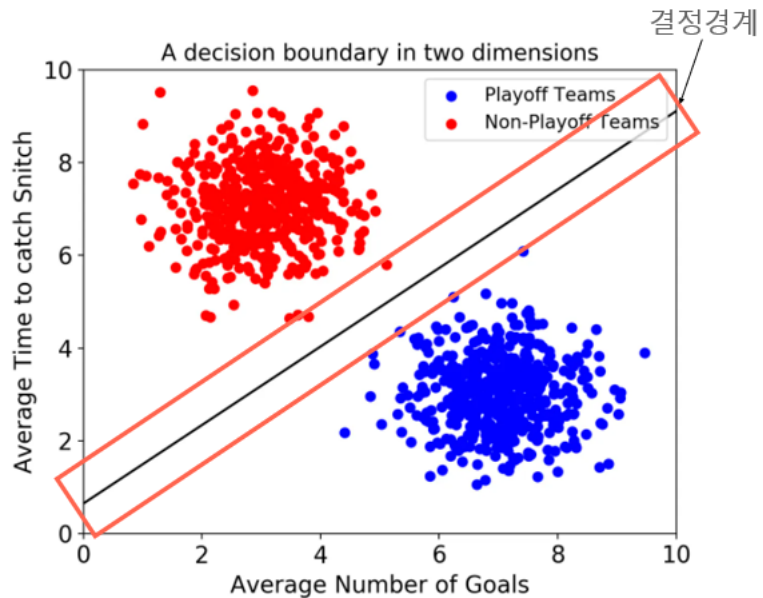
Logistic  
Regression

SVM

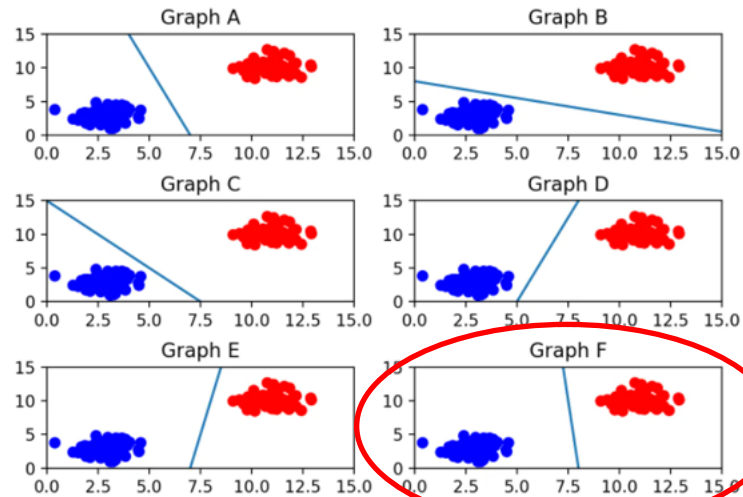
KNN

Decision tree

Random forest



Different Decision Boundaries



# 분류 모델

Logistic  
Regression

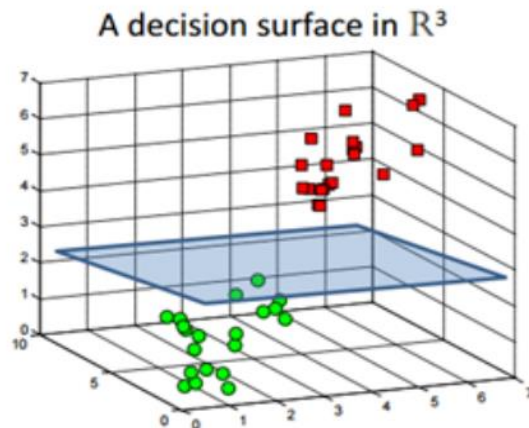
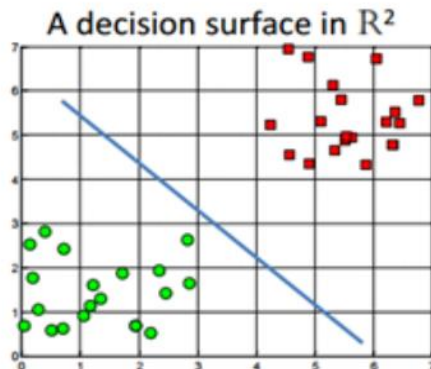
SVM

KNN

Decision tree


Random forest

- 하이퍼플레인(Hyperplane): 2차원 데이터 분류 – Line(선)  
3차원 – 2차원의 면



# 분류 모델



위젯	설명	입력	출력
 SVM	서포트 벡터 머신은 입력을 고차원 feature 공간에 매핑한다.	Data, Preprocessor	Learner, Model, Support Vectors

- 서포트 벡터 머신(support vector machine, SVM)은 속성 공간을 하이퍼플레인으로 분리하여 서로 다른 클래스 또는 클래스 값의 인스턴스 사이의 여백을 최대화하는 머신러닝 기법
- 이 기술은 종종 최고의 예측 성능 결과를 산출
- 오렌지3에는 LIBSVM 패키지에서 널리 사용되는 SVM 구현이 포함되어 있음

## 분류 모델

## Logistic Regression

## SVM

KNN

## Decision tree

Random forest

**SVM**

Name ①

SVM Type ②

☒ SVM Cost (C): 1.00

☐ v-SVM Regression loss epsilon ( $\epsilon$ ): 0.10

Regression cost (C): 1.00

Complexity bound ( $\nu$ ): 0.50

Kernel ③

☒ Linear Kernel:  $x \cdot y$

☐ Polynomial

☐ RBF

☐ Sigmoid

Optimization Parameters ④

Numerical tolerance: 0.0010

☒ Iteration limit: 100

☒ Apply Automatically

① Name	다른 위젯에 표시할 이름으로 기본 이름은 svm이다.
② SVM Type	<p>테스트 오류 설정이 있는 svm 유형이다. svm과 v-svm은 오류 기능의 서로 다른 최소화를 기반으로 한다. 오른쪽에서 검정 오차 한계를 설정할 수 있다.</p> <p>SVM</p> <p>Cost: 손실에 대한 벌칙 용어이며 분류 및 회귀 작업에 적용된다.</p> <p><math>\epsilon</math>: 엡실론-svr 모델에 대한 파라미터로, 회귀 작업에 적용된다. 예측값과 결점이 연관되지 않는 참값으로부터의 거리를 정의한다.</p> <p>v-SVM</p> <p>Cost: 손실에 대한 벌칙 용어이며 회귀 작업에만 적용된다.</p> <p>v: v-SVR 모델에 대한 매개변수는 분류 및 회귀 작업에 적용된다. 훈련 오류의 분수에 대한 상한과 지원 벡터의 분수에 대한 하한이다.</p>
③ Kernel	<p>커널은 속성 공간을 최대 여백 <u>하이퍼플레인</u>에 맞게 새로운 feature 공간으로 변환하는 함수이며, 따라서 알고리즘이 선형, 다항식, RBF 및 <u>시그모이드</u> 커널을 사용하여 모델을 만들 수 있다. 커널을 선택할 때 커널을 지정하는 함수가 표시되며, 관련된 상수는 다음과 같다.</p> <p>g: 커널 함수의 감마 상수(<u>권장값</u>은 <math>1/k</math>이며, 여기서 k는 속성의 수이지만 위젯에 대한 훈련 세트가 없을 수 있으므로 기본값은 0이고 사용자가 수동으로 이 옵션을 설정해야 함)</p> <p>c: 커널 함수의 상수 c(<u>기본값</u> 0)</p> <p>d: 커널의 정도(<u>기본값</u> 3)</p>
④ Optimization Parameters	수치 공차의 예상 값에서 허용되는 편차를 설정한다. 허용되는 최대 반복 횟수를 설정하려면 반복 제한 옆에 있는 상자를 선택한다.

# 분류 모델

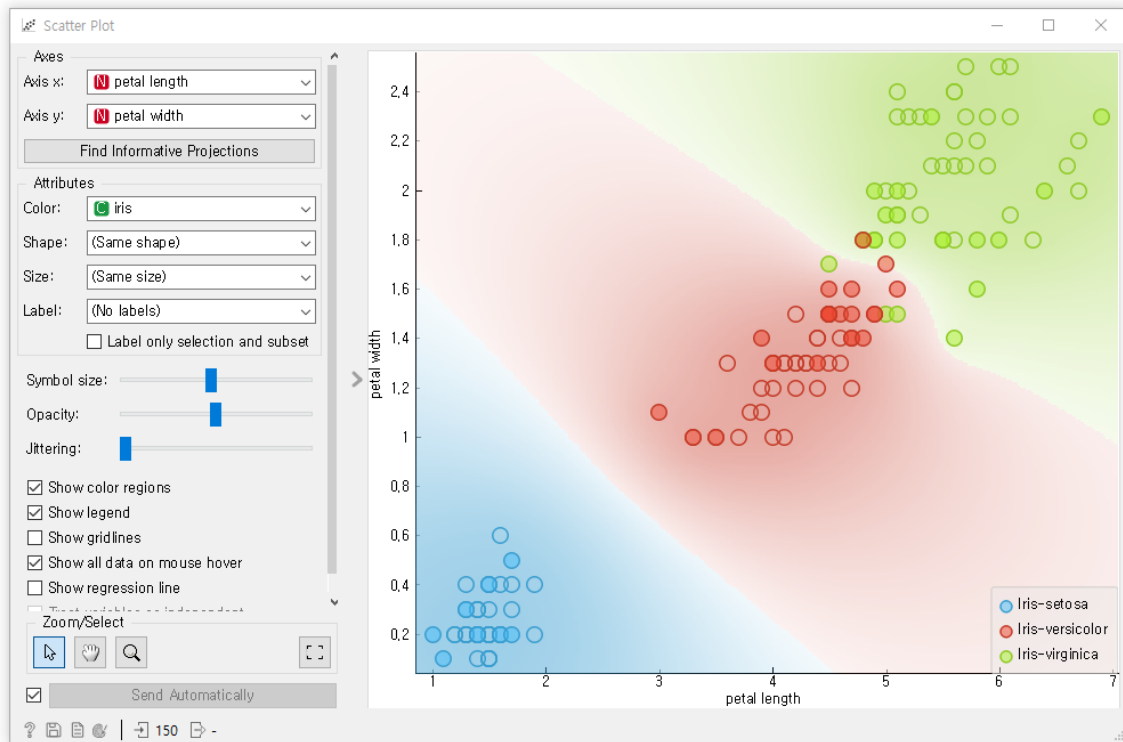
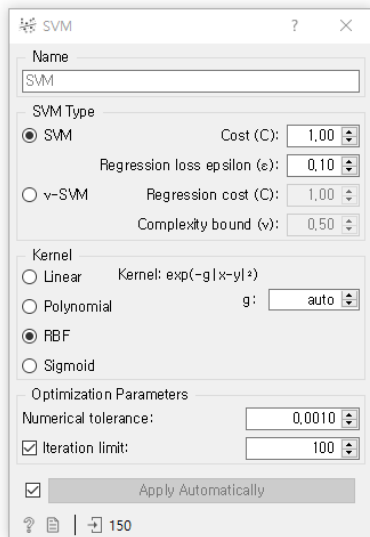
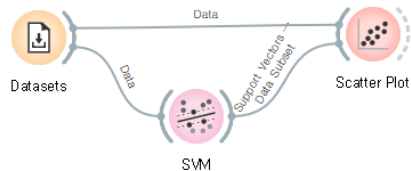
Logistic  
Regression

SVM

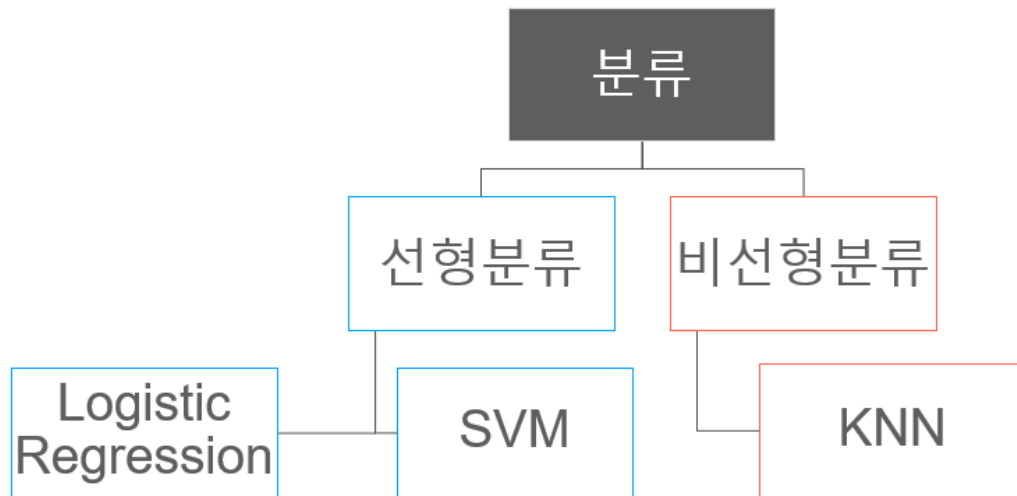
KNN

Decision tree

Random forest

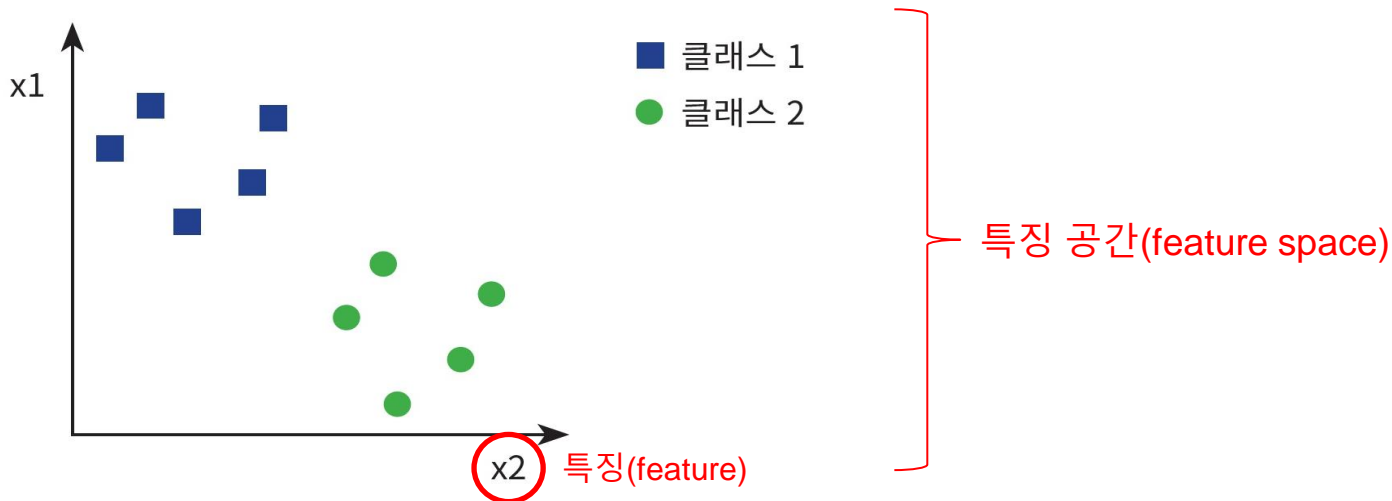


# 분류 모델



# kNN(k-Nearest Neighbor)

- k-Nearest Neighbor(kNN) 은 모든 기계 학습 알고리즘 중에서도 가장 간단하고 이해하기 쉬운 분류 알고리즘
  - 클래스: 서로 다른 종류의 도형
  - 특징공간: 모든 데이터가 투영되는 공간



# kNN 알고리즘

- 새로운 데이터가 입력되어서 그래프 상에 별표로 표시
- 별표는 파랑색 사각형과 녹색 원 중에서 하나에 속해야 함.

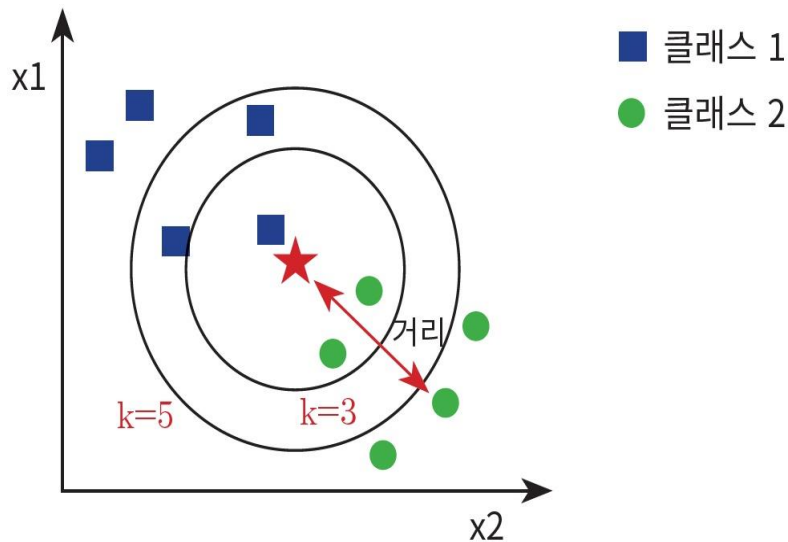
→ 이것을 **분류(classification)**

- 별표와 가장 가까운 k개의 이웃을 기준으로 소속을 결정

→ **kNN(k Nearest Neighbor) 방법**

→ 이때 **k**는 홀수로 결정

→ **k** 값에 따라 결과가 달라질 수 있음





# 분류 모델

Logistic  
Regression

SVM

KNN

Decision tree

Random forest

K-최근접이웃(K-Nearest Neighbor, KNN) 알고리즘은  
새로운 데이터가 주어졌을 때 기존 데이터 가운데 가장 가까운  
K개 이웃의 정보로 새로운 데이터를 예측하는 방법

# 분류 모델

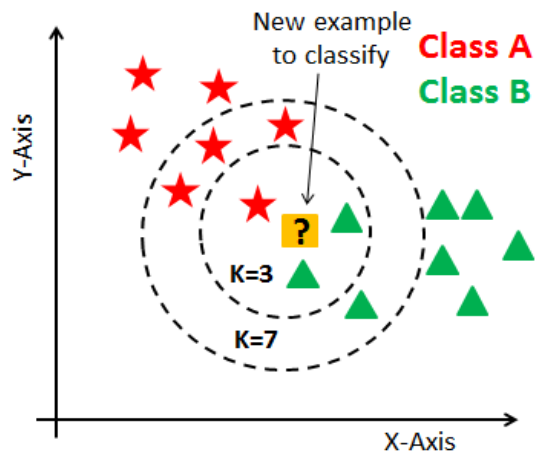
Logistic  
Regression

SVM

KNN

Decision tree


Random forest



새로운 데이터가 주어졌을 때 Class A,  
Class B인지 판단

# 분류 모델



위젯	설명	입력	출력
 kNN	가장 가까운 훈련 인스턴스에 따라 예측한다.	Data, Preprocessor	Learner, Model

kNN 위젯은 feature 공간에서 k개의 가장 가까운 훈련 예를 검색하고 그 평균을 예측으로 사용하는 kNN 알고리즘을 사용

# 분류 모델

Logistic  
Regression

SVM

KNN

Decision tree

Random forest

The screenshot shows the configuration window for the KNN widget. At the top, there is a title bar with a KNN icon, a question mark, and a close button. Below the title bar, the 'Name' field (labeled ①) contains the text 'kNN'. The 'Neighbors' section (labeled ②) includes a 'Number of neighbors' spinner set to 5, a 'Metric' dropdown menu set to 'Euclidean', and a 'Weight' dropdown menu set to 'Uniform'. At the bottom of the configuration area, there is a checked checkbox and a button labeled 'Apply Automatically'. The bottom status bar shows a help icon, a document icon, and the number '150'.

① Name

다른 위젯에 표시할 이름으로 기본 이름은 kNN이다.

② Neighbors

가장 가까운 이웃의 수, 거리 모수(측정지표) 및 가중치를 모형 기준으로 설정한다.

Matric : Euclidean("직선", 두 점 사이의 거리)

Manhattan(모든 속성의 절대적 차이의 합계)

Maximal(속성 간 절대 차이 중 가장 큰 차이)

Mahalanobis(점과 분포 사이의 거리)

Weight : Uniform(각 이웃의 모든 점은 동등하게 가중됨)

Distance(쿼리 포인트의 가까운 이웃이 멀리 있는 이웃보다 더 큰 영향을 미침)

# 분류 모델

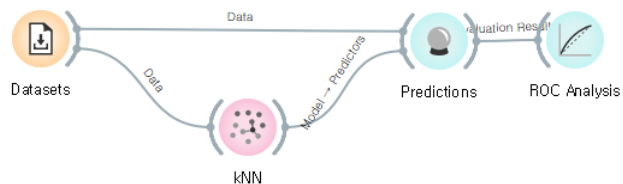
Logistic  
Regression

SVM

KNN

Decision tree

Random forest



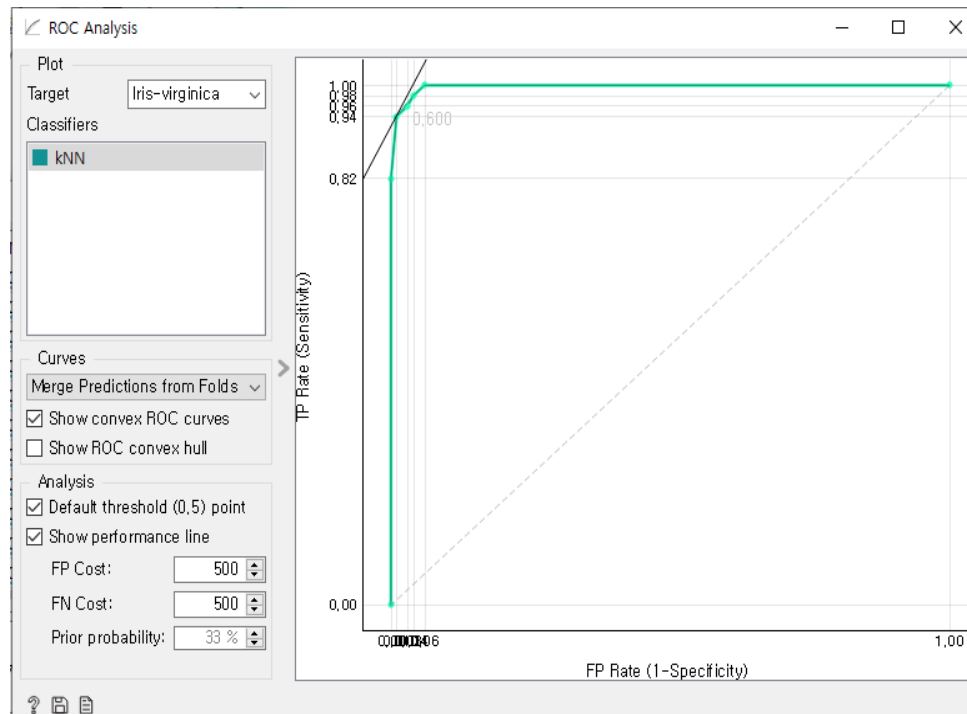
Predictions

Show probabilities for  
Iris-setosa  
Iris-versicolor  
Iris-virginica

	kNN	iris	sepal
1	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.1
2	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.9
3	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.7
4	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.6
5	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.0
6	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.4
7	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.6
8	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.0
9	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.4
10	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.9
11	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.4

Model	AUC	CA	F1	Precision	Recall
kNN	0.998	0.967	0.967	0.967	0.967

Restore Original Order



# 분류 모델

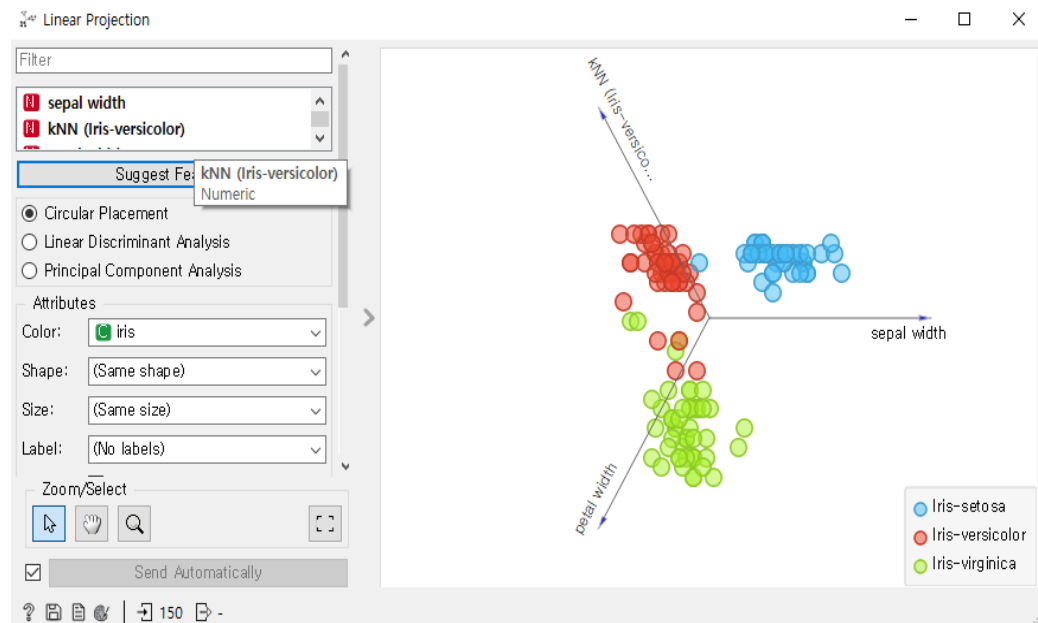
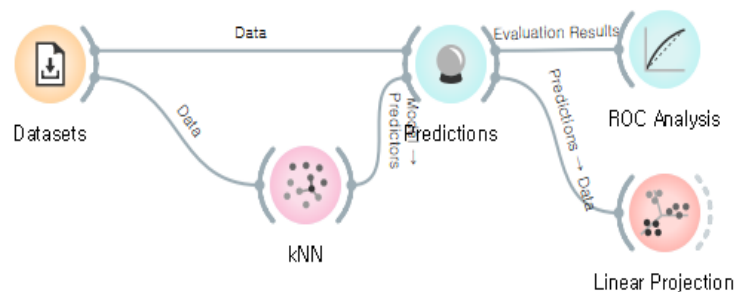
Logistic  
Regression

SVM

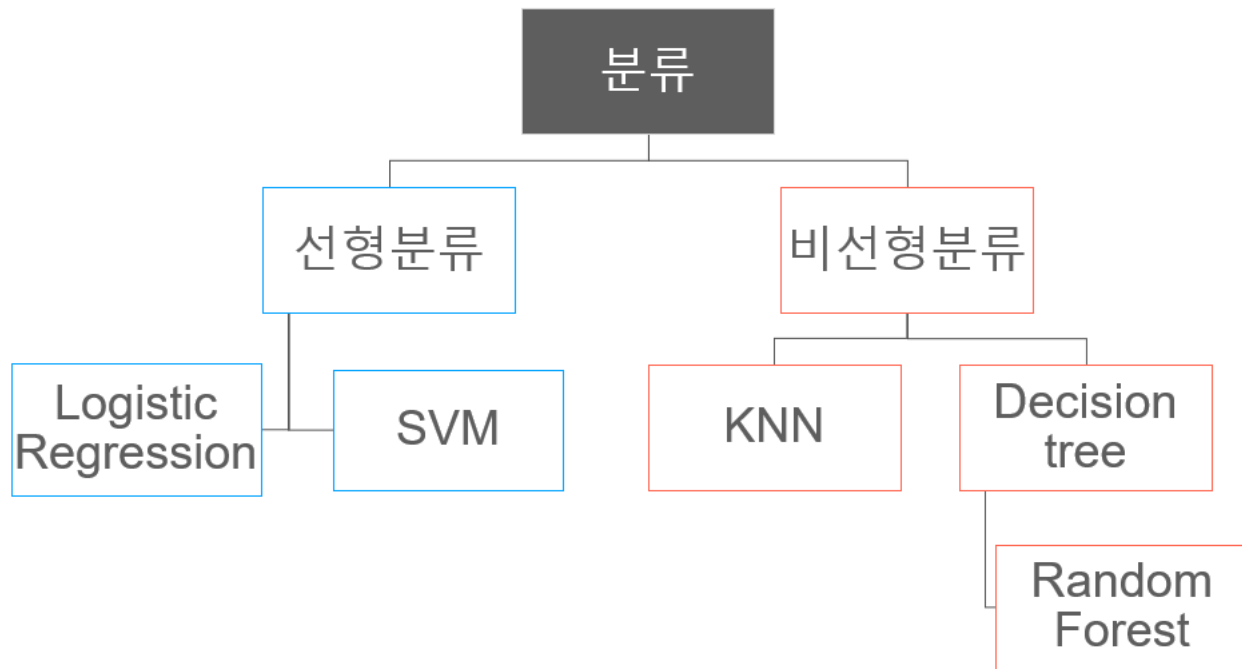
KNN

Decision tree

Random forest



# 분류 모델



# 분류 모델

Logistic  
Regression

SVM

KNN

Decision tree

Random forest

- 결정트리(Decision tree, 의사결정트리)는 데이터를 분석하여 이들 사이에 존재하는 패턴을 찾는 머신러닝 모델
- 질문을 던져서 대상을 좁혀 나가는 스무고개 놀이와 비슷한 개념



# 분류 모델

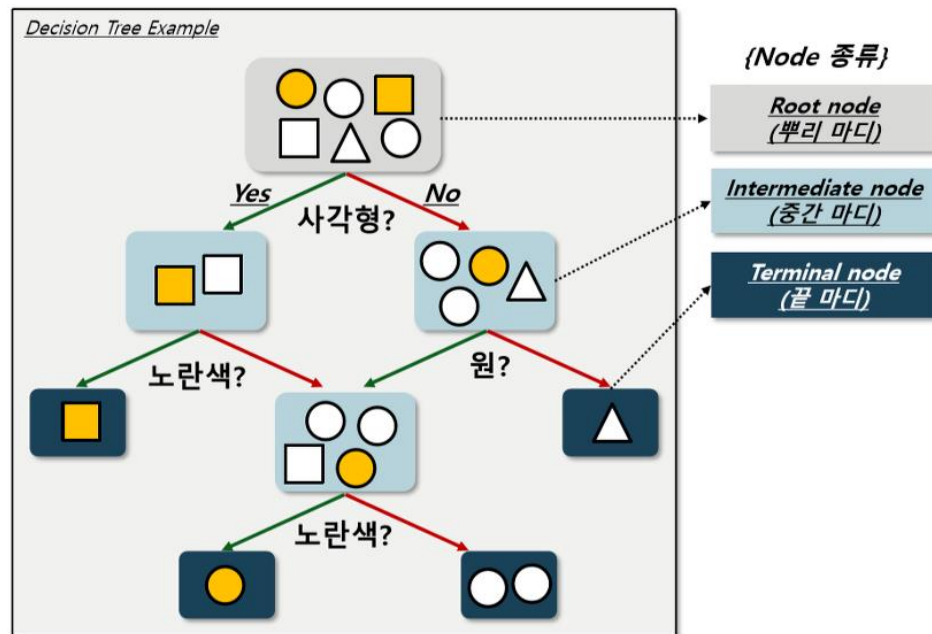
Logistic  
Regression

SVM

KNN


Decision tree

Random forest



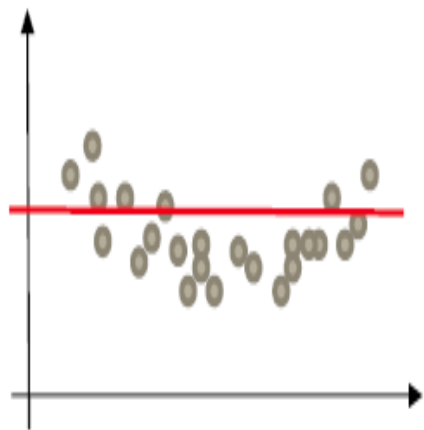
# 분류 모델



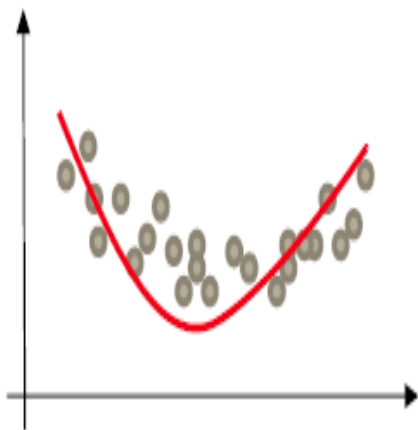
위젯	설명	입력	출력
 Tree	<u>정방향</u> 가지치기 기능이 있는 트리 알고리즘이다.	Data, Preprocessor	Learner, Model

- Tree 위젯은 클래스 순도에 따라 데이터를 노드로 분할하는 간단한 알고리즘으로 랜덤 포레스트의 선구자
- 오렌지3에서는 **이산형 데이터 세트**와 **연속형 데이터 세트를 모두 처리**할 수 있음. **분류**와 **회귀 작업** 모두에 대해 작동
- Decision tree 모델은 **과적합이 발생할 가능성이 높다**는 단점 → 가지치기

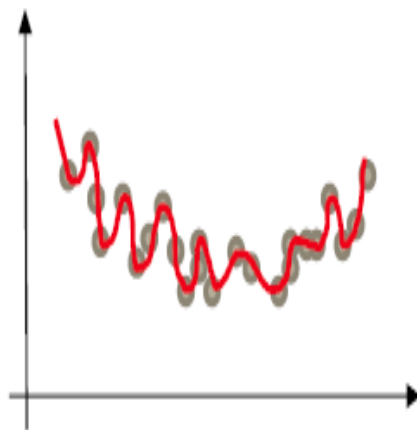
# 과적합(Overfitting)



부적합(underfitting)

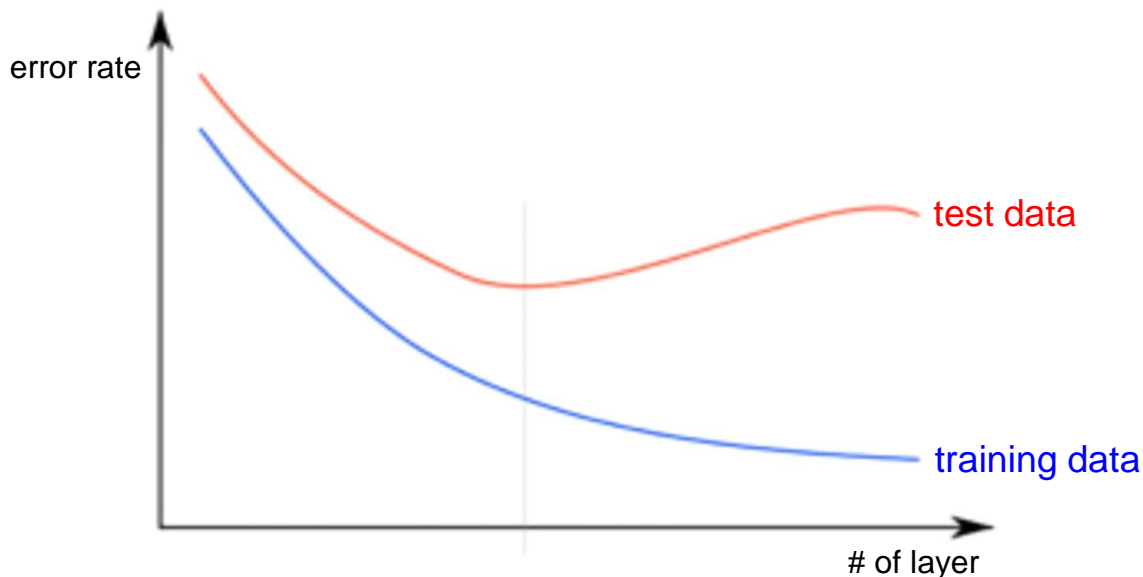


적합(good fitting)



과적합(overfitting)

내 모델이 과잉 적합(Overfitting) 인 것은 어떻게 판단할까?



- Very high accuracy on the training dataset (eg: 0.99)
- Poor accuracy on the test data set (0.85)

# 과잉 적합(Overfitting)의 해결법

- 더 많은 training data 를 사용
  - 데이터 증강(data augmentation): 소량의 훈련 데이터에서 많은 훈련 데이터를 뽑아내는 방법
- 특징(features)의 개수(#)를 감소
- 정규화(Regularization)

# 분류 모델


Logistic  
Regression

SVM

KNN

Decision tree

Random forest

 Tree ? X

Name ①

Parameters ②

☒ Induce binary tree

☒ Min. number of instances in leaves:




☒ Do not split subsets smaller than:

☒ Limit the maximal tree depth to:

Classification ③

☒ Stop when majority reaches [%]:

☒

  |  506

① Name

다른 위젯에 표시할 이름으로 기본 이름은 Tree이다.

② Parameters

Induce binary tree: 두 개의 하위 노드로 분할한다.

Min. number of instances in leaves: 알고리즘이 지정된 수의 훈련 예제를 분기 내에 넣는 분할을 구성하지 않는다.

Do not split subsets smaller than: 알고리즘에서 지정된 인스턴스 수보다 작은 노드를 분할 할 수 없다.

Limit the maximal tree depth: 분류 트리의 깊이를 지정된 노드 수준 수로 제한한다.

③ Classification

지정된 최대 임계값에 도달한 후 노드 분할을 중지한다.

# 분류 모델

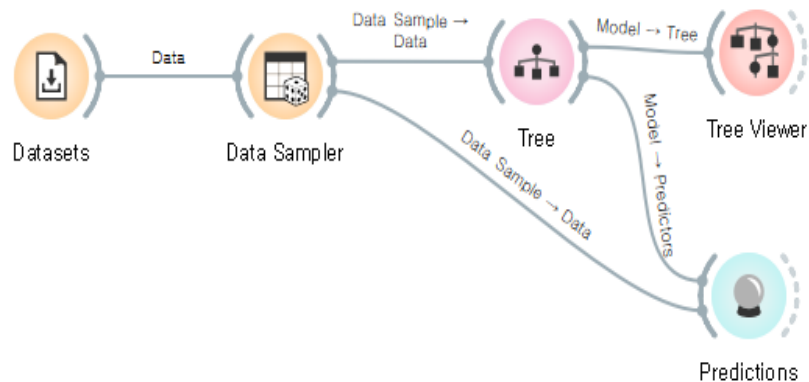
Logistic  
Regression

SVM

KNN

Decision tree

Random forest



Predictions

Show probabilities for

Iris-setosa  
Iris-versicolor  
Iris-virginica

Tree

	Tree	iris	sepal length	sepal width	petal length	petal width
34	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.8	3.4	1.6	0.2
35	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.4	3.2	1.3	0.2
38	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.2	3.5	1.5	0.2
39	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.0	3.6	1.4	0.2
40	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.2	4.1	1.5	0.1
44	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.4	3.9	1.3	0.4
45	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.4	3.7	1.5	0.2
54	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.5	1.4	0.2
57	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.2	3.4	1.4	0.2
58	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.0	3.5	1.3	0.3
59	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.8	1.9	0.4
62	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.8	3.4	1.9	0.2
63	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.0	3.0	1.6	0.2
64	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.3	1.7	0.5
66	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.4	1.5	0.2
67	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.6	3.3	1.4	0.2

Model AUC CA F1 Precision Recall

Tree 0.994 0.971 0.971 0.972 0.971

Restore Original Order

105 105

# 분류 모델

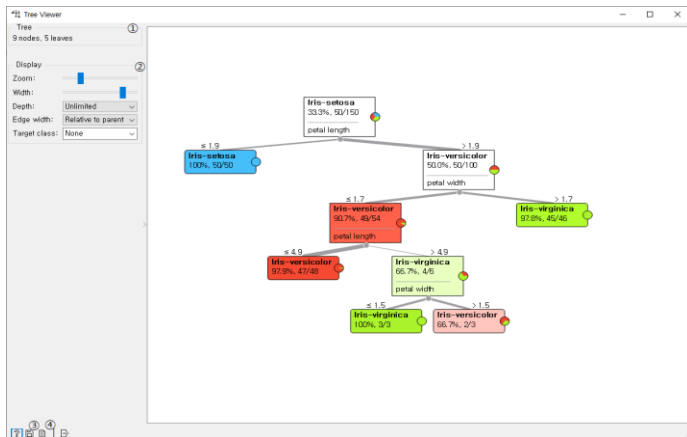
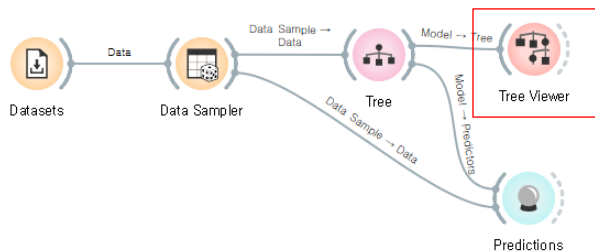
Logistic  
Regression

SVM

KNN

Decision tree

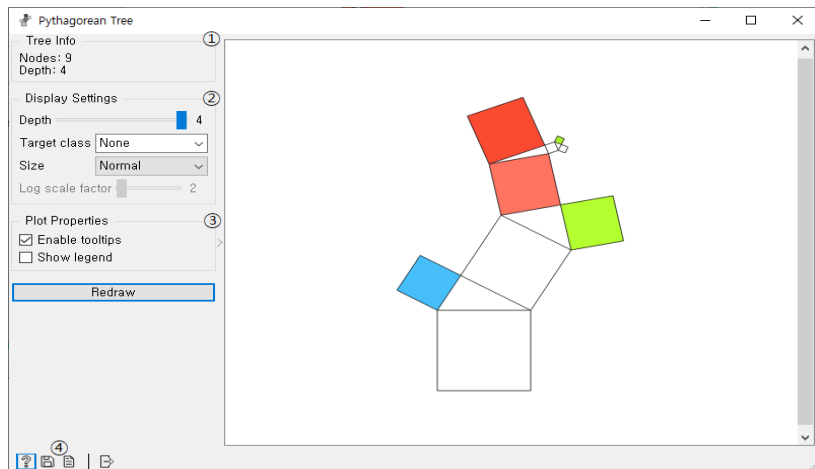
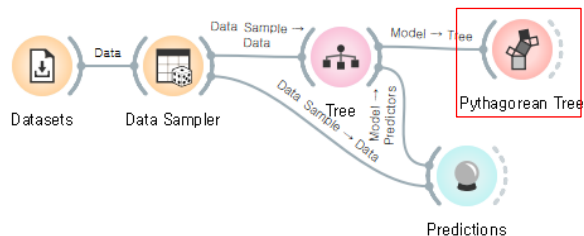
Random forest



① Information	입력값에 대한 정보를 나타낸다. 현재 9개의 노드, 5개의 잎으로 돼 있다.	
② Display	Zoom	좌우로 움직여 확대 및 축소를 할 수 있다.
	Width	좌우로 움직여 트리 너비를 선택할 수 있다.
	Depth	트리의 깊이를 선택할 수 있으며 Unlimited부터 9 levels까지 선택할 수 있다.
	Edge width	모서리 너비를 선택할 수 있다. 'Fixed'를 선택하면 모든 가장자리의 너비를 동일하게 한다. 'Relative to root'를 선택하면 가장자리의 너비는 데이터의 모든 인스턴스에 대한 해당 노드의 인스턴스의 비율에 따라 그려진다. 상위 노드에서 하위 노드로 이동하며 너비가 얇아진다. 'Relative to parent'를 선택하면 가장자리 너비는 상위 노드의 인스턴스에 대한 해당 노드의 인스턴스 비율에 따라 정해진다.
	Target class	목표가 되는 클래스에 따라 대상 클래스를 구분하여 볼 수 있다.
③ Save image	생성된 트리 그래프를 컴퓨터에 .svg 또는 .png 파일로 저장하기 위해 사용된다.	
④ Report	보고서를 만든다.	



# 분류 모델



① Tree info	입력 트리 모델에 대한 정보를 확인한다.	
② Display settings	Depth	표시되는 나무의 깊이를 설정한다.
	Target class	트리의 노드에 대한 색의 강도는 대상 클래스의 확률과 일치한다. '없음'을 선택하면 노드의 색상이 가장 가능성이 높은 클래스를 나타낸다.
	Size	노드 크기가 노드의 교육 데이터 부분 집합의 크기와 일치하도록 유지한다. 제곱근과 로그는 노드 크기의 각 변환이다.
	Log scale factor	로그변환을 선택한 경우에만 활성화되며 로그 팩터를 1에서 10 사이로 설정할 수 있다.
③ Plot Properties	Enable tooltips	호버링 시 노드 정보를 표시한다.
	Show legend	플롯의 색상 범례를 나타낸다.
④ 도움말	이미지를 컴퓨터에 .svg 또는 .png 형식으로 저장하거나 보고서를 작성한다.	

# 분류 모델

Logistic  
Regression

SVM

KNN

Decision tree

Random  
forest

- Random forest는 **앙상블 기계학습 모델**
- 여러 개의 Decision tree를 형성하고 새로운 데이터 포인트를 각 트리에 통과시키며 각 트리가 분류한 결과에서 투표를 실시하여 가장 많이 득표한 결과를 최종 분류 결과로 선택
- **여러 개의 모델을 조화롭게 학습시켜 그 모델들의 예측 결과들을 이용한다면 더 정확한 예측값을 구할 수 있다는 논리 → 앙상블 모델**

# 분류 모델

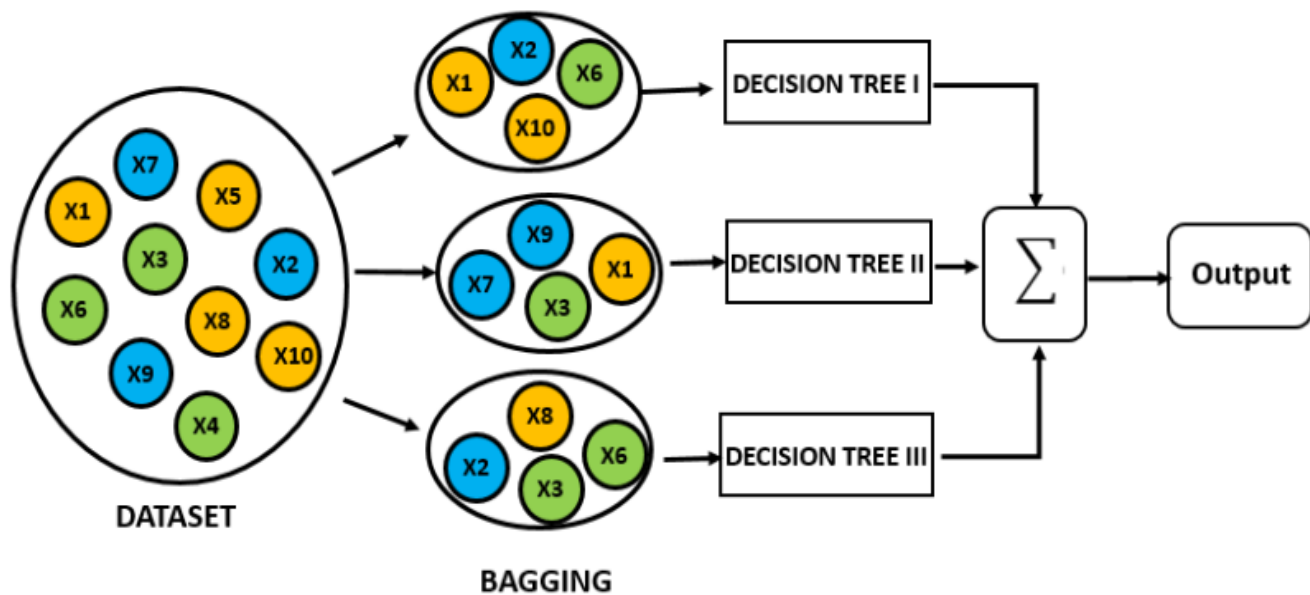
Logistic  
Regression

SVM

KNN


Decision tree

Random  
forest



# 분류 모델



위젯	설명	입력	출력
 Random Forest	의사 결정 트리의 앙상블을 사용하여 예측한다.	Data, Preprocessor	Learner, Model

- Random Forest 위젯은 일련의 **의사 결정 트리**를 만듦.
- 각 트리는 훈련 데이터에서 **부트스트랩 샘플**로 개발
- 개별 트리를 개발할 때 임의적인 속성 부분 집합이 그려지고 여기서 분할에 가장 적합한 속성이 선택
- 최종 모델은 숲에서 개별적으로 개발된 나무들의 다수결에 기초함.

# 분류 모델

Logistic  
Regression

SVM

KNN

Decision tree

Random  
forest

The screenshot shows the 'Random Forest' configuration window. It has a title bar with a question mark and a close button. The window is divided into several sections:

- Name (1):** A text field containing 'Random Forest'.
- Basic Properties (2):**
  - 'Number of trees:' is set to 10.
  - 'Number of attributes considered at each split:' is set to 5.
  - 'Replicable training' is unchecked.
- Growth Control (3):**
  - 'Limit depth of individual trees:' is set to 3.
  - 'Do not split subsets smaller than:' is checked and set to 5.
- Buttons:** At the bottom, there is a checked checkbox and a button labeled 'Apply Automatically'.
- Footer:** A status bar at the bottom left shows a help icon, a document icon, and the number '506'.

① Name

다른 위젯에 표시할 이름으로 기본 이름은 Random Forest이다.

② Basic Properties

Number of trees: 포레스트에 포함할 결정 트리 수를 지정한다.

Number of trees considered at each split: 각 노드에서 고려할 임의로 그릴 속성 수를 지정한다. 후자가 지정되지 않은 경우(옵션 속성 수가 선택되지 않은 상태) 이 숫자는 데이터에 있는 속성 수의 제공근과 같다.

Replicable training: 결과를 복제할 수 있는 트리 생성을 위한 시드를 수정한다.

③ Growth Control

Limit depth of individual trees: 사용자가 나무가 자랄 깊이를 정한다.

Do not split subsets smaller than: 분할 할 수 있는 가장 작은 부분 집합을 선택한다.



# 분류 모델

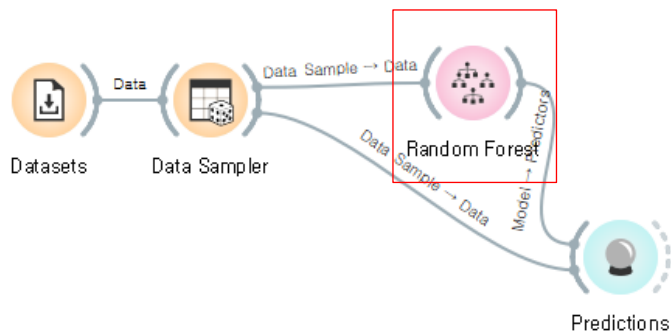
Logistic  
Regression

SVM

KNN

Decision tree

Random  
forest



Predictions

Show probabilities for  
Iris-setosa  
Iris-versicolor  
Iris-virginica

	Random Forest	iris	sepal length	sepal width	petal length	petal width
1	0.00 : 1.00 : 0.00 → Iris-versicolor	Iris-versicolor	6.1	2.8	4.7	1.2
2	0.90 : 0.10 : 0.00 → Iris-setosa	Iris-setosa	5.7	3.8	1.7	0.3
3	0.00 : 0.00 : 1.00 → Iris-virginica	Iris-virginica	7.7	2.6	6.9	2.3
4	0.00 : 1.00 : 0.00 → Iris-versicolor	Iris-versicolor	6.0	2.9	4.5	1.5
5	0.00 : 0.89 : 0.11 → Iris-versicolor	Iris-versicolor	6.8	2.8	4.8	1.4
6	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.4	3.4	1.5	0.4
7	0.00 : 1.00 : 0.00 → Iris-versicolor	Iris-versicolor	5.6	2.9	3.6	1.3
8	0.00 : 0.10 : 0.90 → Iris-virginica	Iris-virginica	6.9	3.1	5.1	2.3
9	0.00 : 0.88 : 0.12 → Iris-versicolor	Iris-versicolor	6.2	2.2	4.5	1.5
10	0.00 : 1.00 : 0.00 → Iris-versicolor	Iris-versicolor	5.8	2.7	3.9	1.2
11	0.00 : 0.00 : 1.00 → Iris-virginica	Iris-virginica	6.5	3.2	5.1	2.0
12	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.8	3.0	1.4	0.1
13	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.5	3.5	1.3	0.2
14	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	4.9	3.1	1.5	0.1
15	1.00 : 0.00 : 0.00 → Iris-setosa	Iris-setosa	5.1	3.8	1.5	0.3

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.999	0.981	0.981	0.981	0.981

Restore Original Order

105

# 분류 모델

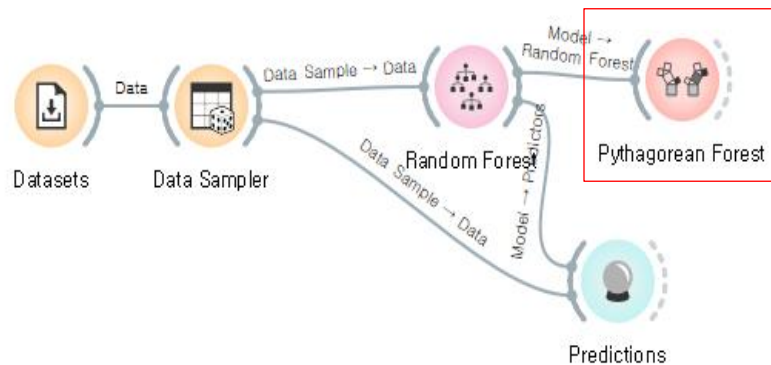
Logistic  
Regression

SVM

KNN

Decision tree

Random  
forest



Pythagorean Forest



# 질문 있나요?

hsryu13@hongik.ac.kr

