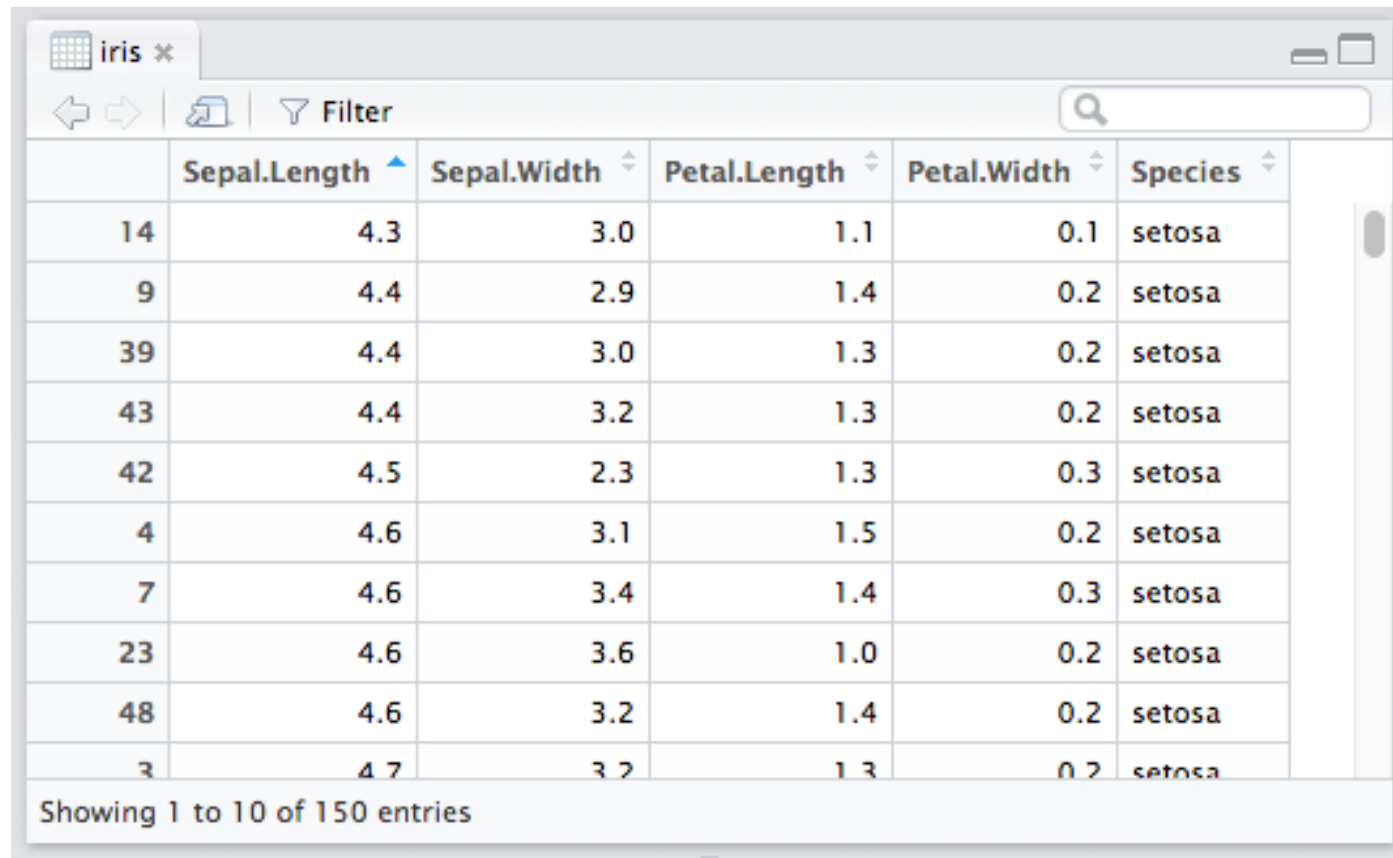


# Supervised learning

# Supervised learning

- Predictor variables/features and a target variable
- 목적: 예측 변수가 주어지면 target 변수를 예측함
  - Classification: target 변수는 categories로 구분됨
  - Regression: target 변수는 값으로 정의됨



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
14	4.3	3.0	1.1	0.1	setosa
9	4.4	2.9	1.4	0.2	setosa
39	4.4	3.0	1.3	0.2	setosa
43	4.4	3.2	1.3	0.2	setosa
42	4.5	2.3	1.3	0.3	setosa
4	4.6	3.1	1.5	0.2	setosa
7	4.6	3.4	1.4	0.3	setosa
23	4.6	3.6	1.0	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa

Showing 1 to 10 of 150 entries

# Naming conventions

- Feature = predictor variables = independent variables
- Target variable = dependent variable = response variable

# Supervised learning

- 시간이 오래 걸리거나 비용이 많이 드는 수동 작업을 자동화
  - 예) 의사의 진단
- 특징 및 상황에 대한 예측
  - 예) 고객이 광고를 클릭 할 것인가 또는 말 것인가?
- 레이블이 있는 데이터가 필요한 경우
  - 레이블이 있는 기록 데이터
  - 레이블이 지정된 데이터를 가져 오는 실험
  - 레이블이 지정된 데이터 크라우드소싱(crowdsourcing)

# Supervised learning(Python)

- scikit-learn/sklearn
  - Integrated well with the SciPy stack
- Other libraries
  - TensorFlow
  - keras

# EDA(Exploratory data analysis)

- Iris dataset
  - Feature
    - Petal length
    - Petal width
    - Sepal length
    - Sepal width
  - Target variables: Species
    - Versicolor
    - Virginica
    - Setosa

IRIS dataset



Iris Versicolor

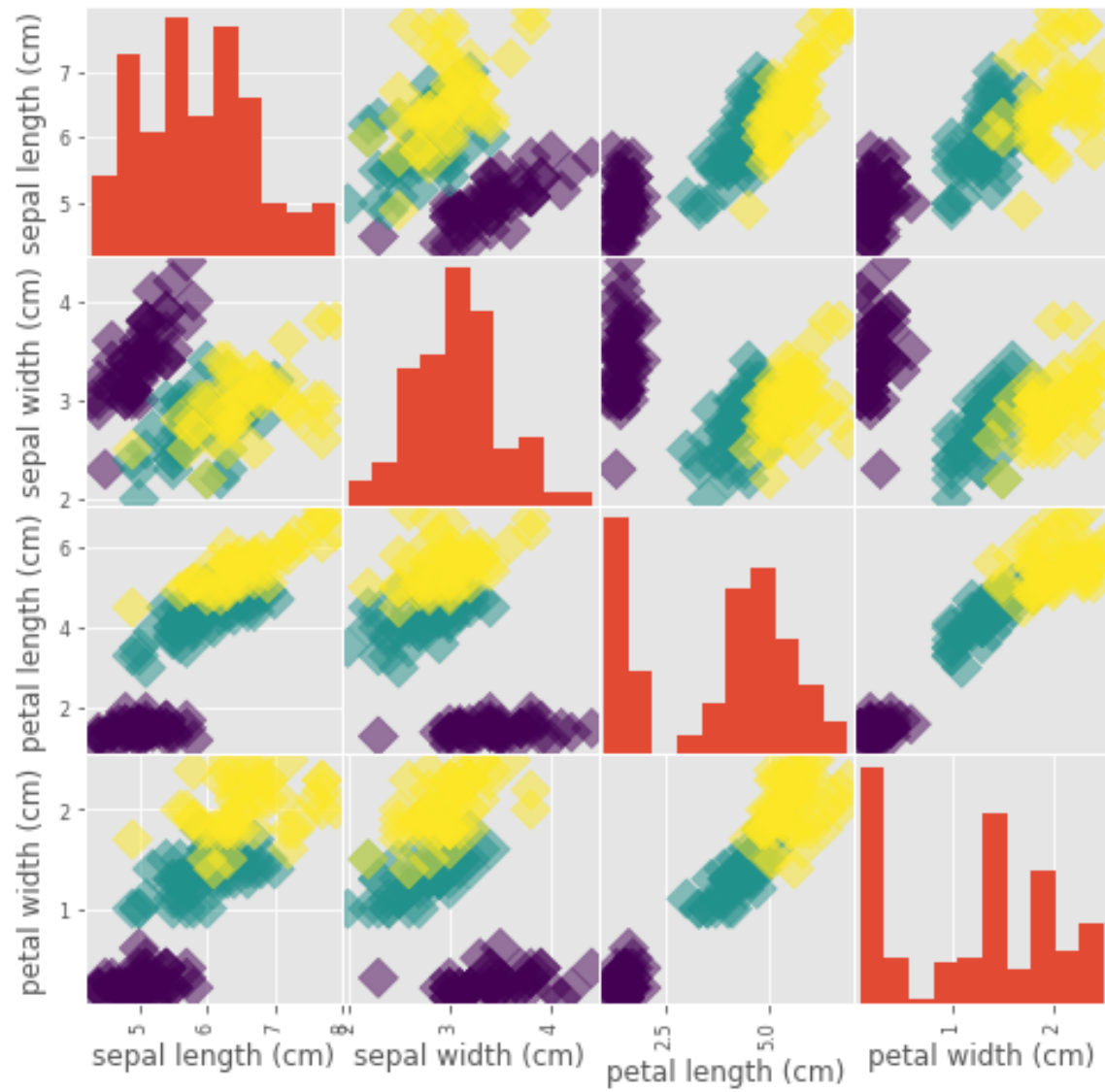


Iris Virginica



Iris Setosa

# Visual EDA



# Supervised learning종류

- 예측(prediction)
  - regression
- 분류(classification)
  - k-Nearest Neighbors



# Measuring model performance

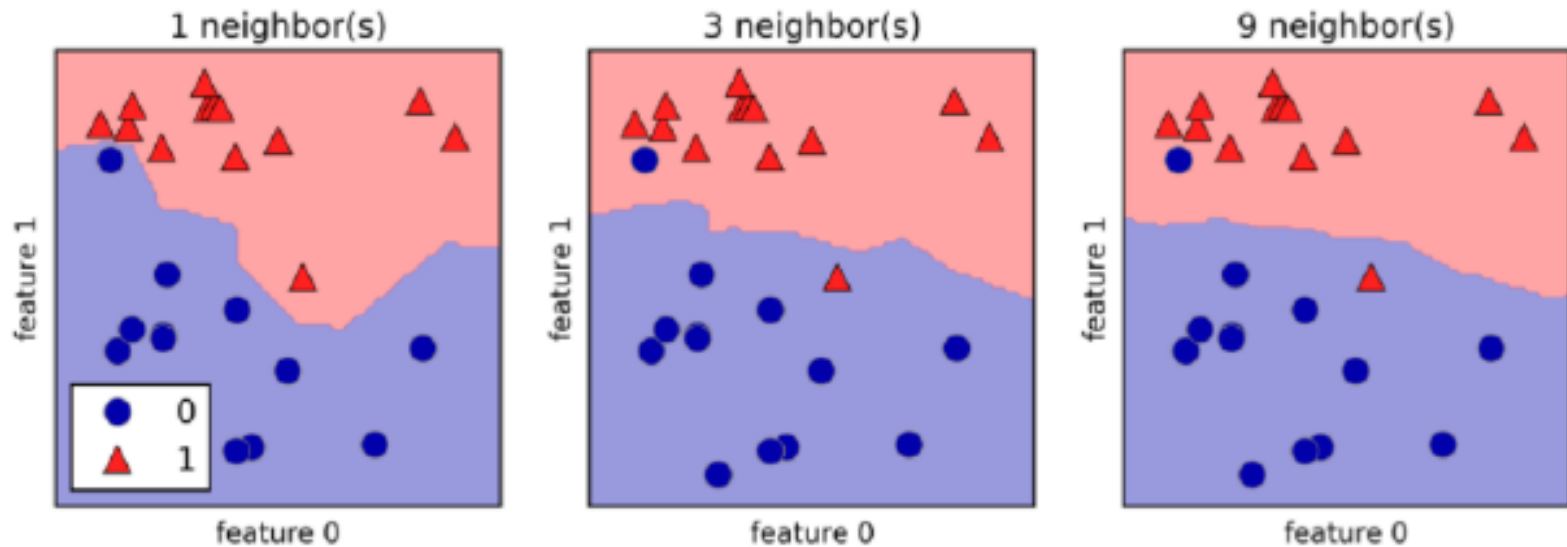
- 분류에서 정확도는 일반적으로 사용되는 측정 항목입니다.
- 정확도 = 정확한 예측의 비율
- 어떤 데이터가 정확도를 계산하는 데 사용해야합니까?
- 모델이 새로운 데이터를 얼마나 잘 수행합니까?

# Measuring model performance

- 분류 기준에 맞는 데이터의 정확도를 계산할 수 있음
  - 일반화 능력을 나타내지 않음
- 데이터를 교육 및 테스트 세트로 분할
  - 훈련 세트에 분류기 맞추기 / 훈련
  - 테스트 세트에 대한 예측을 하십시오.
  - 알려진 라벨과 예측 비교

# Model complexity

- Larger  $k$  = 보다 매끄러운 결정 경계 = 덜 복잡한 모델
- Smaller  $k$  = 보다 복잡한 모델 = overfitting으로 이어질 수 있음



# Model complexity and over/underfitting

