학습과 테스트 데이터 세트의 분리

[Model Selection 소개 - 학습 데이터와 테스트 데이터]

- 학습 데이터 세트

  - 머신러닝 알고리즘의 학습을 위해 사용

  - 데이터의 속성들과 결정값(레이블)값 모두를 가지고 있음

  - 학습 데이터를 기반으로 머신러닝 알고리즘이 데이터 속성과 결정값의 패턴을 인지하고 학습

- 테스트 데이터 세트

  - 테스트 데이터 세트에서 학습된 머신러닝 알고리즘을 테스트

  - 테스트 데이터는 속성 데이터만 머신러닝 알고리즘에 제공하며, 머신러닝 알고리즘은 제공된 데이터를 기반으로 결정값을 예측

  - 테스트 데이터는 학습 데이터와 별도의 데이터 세트로 제공되어야 함

[학습 데이터와 테스트 데이터 분리 - train_test_split()]

- sklearn.model_selection의 train_test_split() 함수

```
X_train, X_test, y_train, y_test

  = train_test_split(iris_data.data, iris_data.target, test_size=0.3, random_state=121)
```

  - test_size: 전체 데이터에서 테스트 데이터 세트 크기를 얼마로 샘플링할 것인지 결정. 디폴트는 0.25, 즉 25%
  - train_size: 전체 데이터에서 학습용 데이터 세트 크기를 얼마로 샘플링 할 것인지 결정. test_size parameter를 통상적으로 사용하기 때문에 train_size는 잘 사용되지 않음
  - shuffle: 데이터를 분리하기 전에 데이터를 미리 섞을지 결정. 디폴트는 True. 데이터를 분산시켜서 좀 더 효율적인 학습 및 데이터 세트를 만드는데 사용됨
  - random_state: 호출할 때마다 동일한 학습/데이터 세트를 생성하기 위해 주어지는 난수 값. train_test_split()는 호출 시 무작위로 데이터를 분리하므로 random_state를 지정하지 않으면 수행할 때마다 다른 학습/테스트용 데이터 생성함

- 넘파이 ndarray 뿐만 아니라 판다스 DataFrame/Series도 train_test_split( )로 분할 가능