사이킷런 소개와 첫번째 머신러닝 애플리케이션 만들어 보기 - 붓꽃(Iris) 품종 예측

[사이킷런 소개]

- 파이썬 기반의 다른 머신러닝 패키지도 사이킷런 스타일의 API를 지향할 정도로 쉽고 가장 파이썬스러운 API를 제공
- 머신러닝을 위한 매우 다양한 알고리즘과 개발을 위한 편리한 프레임워크와 API를 제공
- 오랜 기간 실전 환경에서 검증됐으며, 매우 많은 환경에서 사용되는 성숙한 라이브러리
- 주로 Numpy와 Scipy 기반 위에서 구축된 라이브러리

[머신러닝을 위한 용어 정리]

- 피처(Feature)
  - 피처는 데이터 세트의 일반 속성
  - 머신러닝은 2차원 이상의 다차원 데이터에서도 많이 사용되므로 타겟값을 제외한 나머지 속성을 모두 피처로 지정
- 레이블, 클래스, 타겟(값), 결정(값)
  - 타겟값 또는 결정값은 지도 학습 시 데이터 학습을 위해 주어지는 정답 데이터
  - 지도 학습 중 분류의 경우에는 이 결정값을 레이블 또는 클래스로 지칭

[지도학습 - 분류]

- 분류(classification)
  - 대표적인 지도학습 (Supervised Learning) 방법의 하나
- 지도학습
  - 학습을 위한 다양한 피처와 분류 결정값인 레이블(label) 데이터로 모델을 학습한 뒤, 별도의 테스트 데이터 세트에서 미지의 레이블 예측
  - 명확한 정답이 주어진 데이터를 먼저 학습한 뒤 미지의 정답을 예측하는 방식
- 학습 데이터 세트
  - 학습을 위해 주어진 데이터 세트
- 테스트 데이터 세트
  - 머신러닝 모델의 예측 성능을 평가하기 위해 별도로 주어진 데이터 세트

[사이킷런을 이용한 붓꽃 데이터 분류]

붓꽃 데이터 세트로 붓꽃의 품종을 분류(Classification) 하는 모델

- 꽃잎의 길이와 너비, 꽃받침의 길이와 너비 피처(Feature)를 기반으로 꽃의 품종 예측

[붓꽃 데이터 분류 예측 프로세스]

1. 데이터 세트 분리: 데이터를 학습 데이터와 테스트 데이터로 분리

```
X_train, X_test, y_train, y_test

        = train_test_split(iris_data, iris_label, test_size=0.2, random_state=11)
```

관습적으로 X: Feature, y: 결정값/타겟값을 사용

2. 모델 학습: 학습 데이터를 기반으로 ML 알고리즘을 적용해 모델을 학습시킴

```
dt_clf.fit(X_train, y_train)
```

학습용 피처 데이터 세트에 대해서 학습용 데이터 세트를 매핑해서 학습 수행

3. 예측 수행: 학습된 ML 모델을 이용해 테스트 데이터의 분류(즉, 붓꽃 종류)를 예측

```
pred = dt_clf.predict(X_test)
```

인자는 피처 데이터 세트

테스트 데이터 세트로 예측을 수행하고, 예측 값 반환

4. 평가: 이렇게 예측된 결과값과 테스트 데이터의 실제 결과값을 비교해 ML 모델 성능을 평가

```
accuracy_score(y_test, pred)
```

정확도를 구할 때 사용.

인자로 실제 타겟값과 예측한 타겟값을 넣으면 예측 정확도 반환