**2. 사이킷런으로 시작하는 머신러닝**

**- 머신러닝을 위한 용어정리**

**피처(Feature) :** 타겟 값을 제외한 나머지 속성

**레이블, 클래스, 타겟(값), 결정(값) :** 타겟값/결정값 지도학습 시, 정답데이터

분류 지도학습 시, 결정값을 레이블/클래스로 지칭

**- 지도학습(Supervised Learing) – 분류(Classification)**

**지도학습 :**  정답이 주어진 데이터를 학습시킨 뒤 미지의 정답을 예측하는 방식

**학습 데이터 세트 / 테스트 데이터 세트** (예측 성능 평가를 위해)

**데이터 세트 분리 🡪 모델학습 🡪 예측 수행 🡪 평가**

**- Sklearn기반 프레임워크 – Estimator , fit() , predict()**

**Estimator 학습 : fit() // 예측 : predict()**

**Classifier(분류) : [ 분류 구현 클래스 ]**

DecisionTreeClassifier / RandomForestClassifie / GradientBoostingClassifier / GaussianNB / SVC

**Regressor(회귀) :** **[ 회귀 구현 클래스 ]**

LinearRegression / Ridge / Lasso / RandomForestRegressor / GradientBoostingRegressor

**- Model Selection 소개 – 학습 데이터 / 테스트 데이터**

**학습데이터 :** 데이터 속성과 결정 값의 **패턴**을 인지하고 학습

**테스트데이터 :** 속성데이터만 제공

**- 학습 데이터와 테스트 데이터 분리 – train\_test\_split()**

**sklearn.model\_selection 의 train\_test\_split() 함수**

X\_train, X\_test, y\_train, y\_test = train\_test\_split( feature 데이터, target 데이터, **test\_size** = 0.3 **, random\_state** =121 )

**test\_size :** 몇 %를 테스트 데이터세트로 정할 것인지

**shuffle :** 데이터를 분리하기 전에 섞을 것인지? defalut는 True

데이터를 분산시켜 좀 더 효율적인 학습 및 테스트 데이터 세트 생성

**random\_state :** random\_state를 지정하지 않으면 호출할 때마다 다른 학습/테스트 데이터 생성

( seednumber )

**- 교차검증**

**검증 데이터 세트 :** 기존의 학습데이터 세트에서 다시 분할 (학습 dataset / 검증 dataset)

**- K폴드 교차검증**

인자 K 번 만큼 교차검증 진행 🡪 **교차검증 최종평가** = **평균** ( K 번 교차검증 평가 )

**일반 K 폴드**

**Stratified K 폴드 :**  레이블(결정클래스) 데이터가 불균형 ( 즉 한쪽 데이터 편향성이 큰 경우)

학습데이터 / 검증 데이터 세트의 **레이블 분포가 유사**하도록,,

**분류**를 할 경우 Stratified K 폴드 거의 필수적

**cross\_val\_score() :** 폴드 세트 추출, 학습/예측, 평가를 한번에 수행

**GridSearchCV :**  교차검증과 최적 하이퍼 파라미터 튜닝을 한 번에

Classifier나 Regressor 알고리즘에 사용되는 Hyper-Parameter를 순차적으로 입력하며 최적의 파라미터를 도출할 수 있는 방안을 제공

**Hyper-Parameter :** 사용자가 직접 세팅해 주는 값

성능을 최적화하거나 편향(bias)과 분산(variance) 사이의 균형을 맞출 때 사용

**- 데이터 전처리(Preprocessing)**

**데이터 클린징**

**결손값 처리(Null/NaN 처리)**

**데이터 인코딩 :** 머신러닝 알고리즘의 모든 데이터는 **숫자형으로**

**( 레이블 ) :** 문자열 데이터를 숫자형으로 🡪 알고리즘이 원치 않은 방향으로 갈 수도

**( 원-핫 인코딩 ) :** 0또는 1(고유값에 해당)인 새로운 피처를 추가 **pd.get\_dummies(DataFrame)**

**데이터 스케일링**

**이상치 제거**

**Feature 선택, 추출 및 가공**

**- 피처 스케일링(Feature Scaling)**

**표준화 :** 데이터 피처 각각이 평균 0 분산 1인 가우시안 정규분포를 가진 값으로 변환하는 것

**정규화 :** 서로 다른 피처의 크기를 **통일**하기 위해 크기를 변환해주는 개념

**StandardScaler :** 평균이 0 이고 분산이 1인 정규 분포 형태로 변환

**MinMaxScaler :** 데이터 값을 0과 1사이의 범위 값으로 변환 (음수 있으면 -1에서 1값으로 변환)

테이블이(가) 표시된 사진

자동 생성된 설명