머신러닝 시스템 훈련 과정

1. 데이터 구하기

2. 데이터 탐색과 시각화

1) 데이터프레임과 데이터 탐색

- head(): 데이터프레임에 포함된 처음 5 개의 샘플 확인
- info(): 데이터셋 정보 요약
- value_counts(): 범주형 특성 탐색
- describe(): 수치형 특성 탐색

2) 훈련셋과 테스트셋

- train_test_split()
    - 무작위 샘플링
    - 계층 샘플링: stratify = housing[“income_cat”] 키워드 인자 활용

3) 데이터 시각화

3. 데이터 준비: 정제와 전처리

1) 데이터 정제

- Isnull(): 결측치 있으면 True, 아니면 False 반환
- 결측치 처리 방법
    - 결측치 특성 포함 샘플 삭제
    - 결측치 포함한 특성 삭제
    - 결측치를 중앙값/평균값 등으로 대체

2) 사이킷런 API

- 추정기(estimator), 변환기(transformer), 예측기(predictor) 세 클래스의 인스턴스로 생성

3)SimpleImputer 변환기(Transformer)

- 결측치 다른 값으로 대체(strategy=”mean” | ”median” | ”most_frequent” | ”constant” 속성 사용)
- 1. fit(): 계산된 특성별 평균값, 중앙값, 최빈값 등을 statistics_ 속성에 저장
- 2. transform(): 결측치를 statistics_ 속성에 저장된 값으로 대체

4) 입력 데이터셋 & 타깃 데이터셋 지정

5) OneHotEncoder 변환기(Transformer)

- 범주형 특성 전처리

- 범주 수 만큼의 새로운 특성 추가
- 해당되는 범주와 관련된 특성값은 1, 나머지 특성값은 0
- transform(): 희소 행렬(sparse matrix) 반환
- toarray(): 희소 행렬을 밀집 배열(dene matrix)로 변환
- categories_ 속성: 변환에 사용된 범주들 저장
- feature_names_in_ 속성: 변환된 특성의 이름 저장
- get_feature_names_out(): 변환된 특성들에 대한 새로운 특성명 확인

6) MinMaxScaler, StandardScaler 변환기(Transformer)

- 수치형 특성 스케일링
- MinMaxScaler: 정규화
- StandardScaler: 표준화

7) FunctionTransformer 변환기(Transformer)

- 로그 변환기
- 비율 계산 변환기

8) 군집 변환기

- 사용자 정의 변환기
- BaseEstimator, TransformerMixin 클래스를 상속해야 함

4. 파이프라인

1) Pipeline 클래스

- make_pipeline()
- 변환기/예측기와 동일하게 활용

2) ColumnTransformer 클래스

- 특성별로 파이프라인 지정(ex)수치형, 범주형 구별하여 파이프라인 구성)
- make_column_transformer(): 지정된 자료형을 사용하는 특성들만 뽑아줌

5. 모델 선택과 훈련

1) 모델 훈련

- 선형 회귀 모델: LinearRegression()
- 결정트리 회귀 모델: DecisionTreeRegressor()
- 랜덤 포레스트 회귀 모델: RandomForestRegressor()

2) 모델 평가

- MSE: mean_squared_error()
- RMSE: np.sqrt(MSE)

3) 교차 검증

- cross_val_score(): 훈련 과정 중의 모델 성능 평가 진행

6. 모델 미세 조정

1) 그리드 탐색

- GridSearchCV()
- best_params_ 속성: 최적의 하이퍼파라미터 조합 저장
- best_estimator_ 속성: 최적의 모델 저장
- cv_results_ 속성: 훈련된 모델 각각의 평가지표

2) 랜덤 탐색

7. 최적 모델 저장 및 활용

1) joblib 모듈

- joblib.dump(): 최적의 모델을 이름을 지정하여 저장
- joblib.load(): 저장된 모델 불러오기
- predict(): 불러온 모델 이용하여 예측