머신러닝

✅ 1. 데이터 준비

주요 작업

- CSV, Excel, SQL 등에서 불러오기 (pandas 사용)

✅ 2. 데이터 전처리 (Preprocessing)

scikit-learn 사용

결측치 처리

- 평균/중앙값/최빈값으로 대체 **(imputer 사용)**
- 또는 해당 행 제거

SimpleImputer

- 하나의 열을 기준으로 전체 결측치를 단순한 통계값으로 대체

`"mean"`	평균으로 대체 (숫자형 데이터)
`"median"`	중앙값으로 대체
`"most_frequent"`	가장 많이 등장한 값으로 대체
`"constant"`	지정한 상수로 대체 (`fill_value=` 필요)

KNNImputer (최근접 이웃)

- 결측치가 있는 행과 **비슷한 다른 행들의 값 평균**으로 결측치 채움

IterativeImputer(선형 회귀 기반 반복 예측)

- 각 결측값을 다른 피처를 통해 예측하는 방식(조금 느리지만 정확도 높을 수 있음)

범주형 데이터 처리

- Label Encoding: 순서가 있는 경우
- One-Hot Encoding: 순서가 없는 경우

특성 스케일링

- `StandardScaler` : 평균 0, 표준편차 1

- `MinMaxScaler`: 0~1 정규화

✅ 3. 학습/테스트 분할

- Train-set, Test-set 분리

✅ 4. 모델 선택 및 학습

자주 쓰이는 모델

- **분류(Classification)**: `LogisticRegression`, `RandomForestClassifier`, `XGBoostClassifier`
- **회귀(Regression)**: `LinearRegression`, `RandomForestRegressor`

✅ 5. 예측 및 평가

분류

- 평가지표 : accuracy_score

회귀

- 평가지표 : mean_squared_error

✅ 6. 하이퍼파라미터 튜닝 (선택)

- GridSearchCV 사용

✅ 7. 전체 파이프라인 구성

- sklearn.pipeline