1. 문제 정의 -> 데이터 전처리 -> 학습 -> 모델 평가

문제 정의 : 해당 데이터를 통해 찾고자 하는 문제를 파악

데이터 전처리 : 탐색적 분석을 전체적인 데이터 분포를 살펴보고(EDA) 통계값이나 그래프를 통한 시각화 기법을 사용한다.

학습 : 데이터 종류와 문제에 따라 알고리즘을 선택하고 전처리 한 데이터를 집어넣어 학습

모델 평가 : 모델이 성능이 얼마나 좋은지 나타내는 지표를 통해 모델의 성능을 예측

1. 분류 데이터 셋 : 레이블이 있는 데이터로 각 샘플이 특정 클랫에 속함

회귀 데이터 셋 : 레이블이 연속적인 값으로 입력 데이터에 대해 연속적인 출력을 예측하는데 사용함.

군집화 데이터 셋 : 레이블이 없는 데이터로, 비슷한 특징을 가진 데이터를 그룹하하는데 사용

차원 축소 데이터 셋 : 고차원의 데이터를 저차원으로 변환하여 시각화하거나 분석

강화 학습 데이터 셋 : 행동을 취하고 그에 대한 보상을 받음?

1. 지도 학습의 특징 : 모델이 주어진 데이터와 그에 대한 정답(레이블)을 기반으로 학습하여 새로운 데이터에 대해 정확한 예측을 하도록 도와주는 모델을 만들 때 사용되는 학습법.
2. 결측치 처리, 중복값 처리, 이상치 탐지 및 처리, 데이터 형식 변환, 단위 변환, 정규화, 표준화, 특징 선택, 특징 추출 등
3. 데이터로부터 학습하여 특정 작업을 수행하거나 예측하는 알고리즘 또는 수학적 구조를 의미.
4. 회귀(Regression)
5. 모델이 특정 데이터를 학습하고 해당 모델에서 생성되는 매개변수 (사용자가 직접 바꾸진 못함,.)
6. KNN 알고리즘이란 주로 분류와 회귀 문제에서 사용된다. 주어진 데이터 포인트와 가장 가까운 K개의 이웃을 찾아서 예측을 수행한다.
7. 하이퍼 파라미터란 모델에서 사용자가 직접 설정할 수 있는 매개변수로 하이퍼 파라미터튜닝을 통하여 해당 모델의 성능을 높일 수 있다.
8. MSE : 예측 값과 실제 값 간의 차이의 제곱을 평균한 값

MAE : 예측 값과 실제 값 간의 차이의 절대값을 평균한 값

RMSE : MSE의 제곱근

R2 : 모델의 설명력 또는 결정계수, 값이 1에 가까울수록 모델의 성능이 좋음

15. 과대 적합 : 모델이 학습을 너무 과하게 해서 다른 데이터가 들어오면 해당 데이터를 제대로 처리하지 못하는 상태.

과소 적합 : 자기 자신의 데이터도 제대로 학습하지 못하여 모델의 성능이 낮은 상태.