의사결정트리

의사결정 트리- 기준이 명확

: 데이터 패턴을 학습한 후, 트리 형태로 규칙을 나열

: 하향식 (if-else)

최적 = 키가 작고 가지가 x 모델

- 가지가 많으면 = 과대적합
- 치우친 모델 = 높은 분산(데이터 변화가 전체 트리에 큰 변화)

  → 랜덤 포레스트 (의사결정 트리 여러 개로 숲을 이룬 형태)

  : 데이터의 작은 변화 → 구조적 변화로 이어지는 것을 방지

  : 빈도가 높은 것을 결과로 반환

  : 앙상블의 한 예시(여러 머신러닝 알고리즘이 결합된 것)

  : 데이터가 적을 경우 → 응용x

ex) 타이타닉 생존자 예측

1. 전처리
   a. embarked 결측치 드랍

      ```
      dropna(subset = '이름', axisi= '축', inplace=True)
      ```

   b. 나이 평균으로 채우고

      ```
      fillna( ‘무엇을’, inplace=True)
      ```

   c. 이름에서 title 추출 / 범주형 데이터 정리

      ```
      df['Name'].str.extract('([A-Za-z]+)\.')
      replace('무엇을', '무엇으로')
      ```

   d. 필요 x 속성 삭제
   e. 범주형 → 숫자

2. 의사결정 트리
   a. 훈련/테스트 분활 → 학습 → 평가(예측, 정확도, 혼동행렬)→시각화→.png내보내기

3. 랜덤포레스트

a. 학습 → 평가 (6.7% 더 정확)

4. Gini계수

0에 수렴할수록 잘 분산되었다는 뜻으로, 데이터를 더 잘 나눌 수 있다.