**결정트리(Decision Tree) 요약**

**의사결정트리 개념 및 구조**

* 의사결정트리는 데이터의 속성(특성)으로부터 예측 가능한 규칙을 찾아 분류 또는 회귀를 수행하는 지도학습 모델이다. 범주형(분류)과 연속형(회귀) 데이터 모두에 적용할 수 있다
* 트리는 뿌리마디(루트 노드)에서 시작해, 각 분기점(부모/자식 노드)을 따라 조건에 따라 데이터를 분할하며, 마지막에 도달하는 끝마디(리프 노드)에서 최종 예측값을 반환한다. 깊이(depth)는 루트에서 리프까지의 노드 개수다

**결정트리의 학습 및 예측 과정**

* 트리 형성: 데이터와 목표에 따라 적절한 분리기준(불순도, 순수도 등)과 정지규칙(최대 깊이, 최소 샘플 수 등)을 정해 트리를 만든다. 부모노드의 불순도를 줄이는 방향으로 자식노드를 분할한다
* 가지치기(Pruning): 분기가 너무 많아 과적합(overfitting)이 발생할 수 있으므로, 불필요하거나 오류가 큰 가지를 제거해 일반화 성능을 높인다
* 예측: 새로운 데이터는 루트에서 조건에 따라 분기점을 따라가며 리프 노드에 도달, 그 노드의 대표값(분류: 최빈값, 회귀: 평균값)으로 예측한다

**CART 알고리즘**

* CART(Classification and Regression Tree)는 대표적인 결정트리 알고리즘으로, 각 노드에서 비용함수(지니 불순도 또는 평균제곱오차 등)를 최소화하는 특성과 임계값을 찾아 데이터를 분할한다. 이 과정을 반복해 트리를 완성한다

**지니 불순도와 엔트로피**

* 지니 불순도(Gini impurity)와 엔트로피(Entropy)는 데이터의 혼잡도를 측정하는 지표다.
  + 지니 불순도는 기본값으로 자주 쓰이며, 계산이 빠르다.
  + 엔트로피는 정보이득(Information Gain)을 기반으로 분할을 평가하며, 노드를 더 균형 있게 나누는 경향이 있다. 두 방식 모두 큰 차이는 없으나, 엔트로피는 약간 더 균형 잡힌 트리를 만들 수 있다

**결정트리의 규제(Regularization)**

* 결정트리는 데이터에 대한 가정이 없는 비파라미터 모델로, 과적합 위험이 크다.
* 규제 하이퍼파라미터(예: max\_depth, min\_samples\_split, min\_samples\_leaf, max\_leaf\_nodes, max\_features 등)를 조정해 트리의 복잡도를 제한할 수 있다. min\_ 접두사는 값을 늘리고, max\_ 접두사는 값을 줄이면 규제가 강화된다

**회귀 결정트리**

* 분류뿐 아니라 연속형 타깃값 예측(회귀)에도 결정트리를 사용할 수 있다. 이때는 리프 노드에서 평균값을 반환하며, 분할 기준은 평균제곱오차(MSE)를 최소화하는 방향으로 결정한다

**장점과 단점**

* 장점:
  + 해석이 쉽고, 변수 간 상호작용을 파악할 수 있으며, 비모수적(분포 가정 불필요)이다.
* 단점:
  + 연속형 변수를 비연속적으로 처리해 경계 부근 예측 오류가 커질 수 있고, 데이터의 작은 변화에도 민감(높은 분산)하며, 축에 수직인 분할만 가능해 데이터 회전에 약하다. 랜덤포레스트 등 앙상블 기법으로 단점을 보완할 수 있다