Decision Tree Node Explanation (education\_num <= 13.5)

# 1. 조건 (Condition): education\_num <= 13.5

이론:

이 조건은 education\_num이라는 피처가 13.5보다 작거나 같은지 여부를 확인합니다. education\_num은 보통 개인이 받은 교육 연수를 나타내며, 이 조건을 통해 데이터가 두 그룹으로 분할됩니다: 교육 연수가 13.5년 이하인 사람과 13.5년을 초과한 사람.

수식:

이 조건은 데이터를 분할할 때, 주어진 피처에 대한 수학적 비교 연산을 수행합니다. 수식으로 표현하면:

if education\_num <= 13.5: True (left branch); else: False (right branch)

분할 기준을 결정할 때, 지니 계수(Gini Index)나 엔트로피(Entropy) 등의 수식을 사용하여 최적의 분할을 찾습니다.

사례:

예를 들어, education\_num이 13.5라면 이는 고등학교 졸업 후 몇 년의 추가 교육(예: 대학)을 받았는지를 나타낼 수 있습니다. 이 조건은 "이 사람은 고등학교 졸업 이후 어느 정도의 추가 교육을 받았는가?"를 묻고 있으며, 교육 연수가 13.5년 이하인 경우는 주로 고등학교 수준 또는 그 이하, 초과한 경우는 대학 이상을 나타낼 수 있습니다.

# 2. 샘플 (Samples): 53.3%

이론:

이 노드에서 전체 데이터의 53.3%가 처리된다는 것을 나타냅니다. 즉, 이 노드는 트리 구조 내에서 전체 데이터의 약 절반 정도를 다루고 있음을 의미합니다.

수식:

이 샘플 비율은 전체 데이터에서 해당 노드로 분할된 데이터의 비율로 계산됩니다:

Sample percentage = (Number of samples at node / Total number of samples) × 100

사례:

예를 들어, 1000개의 데이터 포인트가 있는 데이터셋에서 이 노드로 533개의 샘플이 분할되었다면, 샘플 비율은 53.3%가 됩니다.

# 3. 값 (Value): value = [0.049, 0.951]

이론:

이 값은 이 노드에 있는 데이터가 두 개의 클래스(예: low와 high)에 속하는 비율을 나타냅니다. [0.049, 0.951]은 4.9%의 데이터가 첫 번째 클래스(high)에 속하고, 95.1%의 데이터가 두 번째 클래스(low)에 속한다는 의미입니다.

수식:

클래스 비율을 계산하는 수식은 다음과 같습니다:

Class probability = (Number of samples in class / Total number of samples at node)

이 수식을 사용해 각 클래스의 비율을 계산합니다.

사례:

예를 들어, 특정 교육 수준 이하의 사람들 중 95.1%가 낮은 소득(low)을 가지고 있고, 4.9%만이 높은 소득(high)을 가지고 있다면, 이 비율은 value = [0.049, 0.951]로 나타낼 수 있습니다.

# 4. 클래스 (Class): class = low

이론:

이 노드의 최종 예측 클래스는 low입니다. 이는 노드에 도달한 데이터 중 대부분이 low 클래스에 속하기 때문에 결정된 것입니다.

수식:

클래스는 주로 노드에 포함된 데이터 중 가장 높은 비율을 차지하는 클래스로 결정됩니다. 이 예에서는 [0.049, 0.951]에서 95.1%가 low 클래스에 속하므로 최종 클래스는 low로 결정됩니다.

사례:

예를 들어, 이 노드에 도달한 데이터가 주로 낮은 소득(low)을 가진 사람들로 이루어져 있다면, 모델은 이 노드에서의 예측 클래스를 low로 설정하게 됩니다.