Decision Tree Node Explanation (capital\_gain <= 7073.5)

# 1. 조건 (Condition): capital\_gain <= 7073.5

이론:

이 조건은 capital\_gain이라는 피처의 값이 7073.5 이하인지 여부를 판단하는 것입니다. capital\_gain은 자본 이득을 나타내며, 특정 금액 이하의 자본 이득을 얻은 사람들과 그 이상을 얻은 사람들로 데이터를 분할합니다.

수식:

이 조건을 수식으로 표현하면:

if capital\_gain <= 7073.5: True (left branch); else: False (right branch)

이 조건은 지니 계수(Gini Index)나 엔트로피(Entropy) 등의 수식을 사용하여 최적의 분할을 찾는 과정에서 선택됩니다.

사례:

예를 들어, capital\_gain이 7073.5 이하라면, 이 사람은 상대적으로 적은 자본 이득을 얻은 것으로 간주됩니다. 이 조건을 통해, 자본 이득이 적은 사람과 많은 사람으로 데이터를 나눌 수 있습니다. 이 노드는 자본 이득이 낮은 사람들(<= 7073.5)을 분류하는 역할을 합니다.

# 2. 샘플 (Samples): 54.2%

이론:

이 노드에서 전체 데이터의 54.2%가 처리된다는 것을 나타냅니다. 이 노드는 전체 데이터의 절반 이상을 포함하고 있음을 의미합니다.

수식:

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 542개의 샘플이 분할되었다면, 샘플 비율은 54.2%가 됩니다.

# 3. 값 (Value): value = [0.064, 0.936]

이론:

이 값은 이 노드에 있는 데이터가 두 개의 클래스(예: low와 high)에 속하는 비율을 나타냅니다. [0.064, 0.936]은 6.4%의 데이터가 첫 번째 클래스(high)에 속하고, 93.6%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 자본 이득 이하의 사람들 중 93.6%가 낮은 소득(low)을 가지고 있고, 6.4%만이 높은 소득(high)을 가지고 있다면, 이 비율은 value = [0.064, 0.936]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 노드에 포함된 데이터 중 가장 높은 비율을 차지하는 클래스로 결정됩니다. 이 예에서는 [0.064, 0.936]에서 93.6%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.