Explanation of Variable Order in Decision Tree

# 1. 루트 노드의 선택

루트 노드는 트리의 최상위 노드로, 데이터 전체를 가장 잘 분할할 수 있는 변수가 선택됩니다. 이 트리에서는 marital\_status\_Married-civ-spouse <= 0.5가 루트 노드로 선택되었습니다. 이 변수는 결혼 여부를 나타내며, 이 조건을 기준으로 데이터를 크게 두 그룹으로 나눕니다. 루트 노드의 선택은 주어진 데이터를 가장 잘 분리할 수 있는 변수를 기준으로 하며, 지니 계수(Gini Index), 엔트로피(Entropy) 등의 기준을 사용합니다. 트리 모형이 이 변수를 루트 노드로 선택한 것은 결혼 여부가 이 데이터에서 소득 예측에 가장 중요한 요인이라는 것을 의미합니다.

# 2. 하위 노드로의 분기

루트 노드 아래에서 트리는 각각의 하위 노드로 분기됩니다. 예를 들어, 루트 노드에서 marital\_status\_Married-civ-spouse <= 0.5라는 조건을 충족하는 경우는 왼쪽 가지로, 충족하지 않는 경우는 오른쪽 가지로 이동합니다. 이 때, 각각의 가지에서 새로운 변수가 선택됩니다. 선택된 변수들은 해당 하위 집합에서 가장 분리 효과가 높은 변수들입니다.

왼쪽 분기:

첫 번째로 capital\_gain <= 7073.5가 선택됩니다. 이는 자본 이득이 7073.5 이하인 경우와 아닌 경우로 데이터를 나눕니다. 다음으로 education\_num <= 13.5와 age <= 20.0이 각각 선택되어 데이터를 더 세부적으로 나누어 줍니다.

오른쪽 분기:

education\_num <= 12.5가 선택되며, 이는 교육 연수가 12.5년 이하인 사람들을 더 세부적으로 나누는 데 사용됩니다. 이후 capital\_gain <= 5095.5라는 조건이 선택되어 자본 이득이 낮은지 높은지를 기준으로 데이터를 분할합니다.

# 3. 노드 선택의 원리

각 노드에서 선택되는 변수는 해당 변수로 데이터를 분할할 때 클래스 간의 불순도를 최대한 줄일 수 있는 방향으로 결정됩니다. 즉, 각 노드는 가능한 변수들 중에서 해당 노드의 데이터 불순도를 가장 잘 줄일 수 있는 변수를 선택합니다. 불순도를 줄이는 것은 주로 지니 계수(Gini Index)나 정보 이득(Information Gain)과 같은 지표를 사용해 계산됩니다.

지니 계수(Gini Index):

노드의 불순도를 측정하는 지표로, 0에 가까울수록 불순도가 낮고, 1에 가까울수록 불순도가 높습니다. 각 변수에 대해 지니 계수를 계산하고, 가장 불순도를 많이 줄일 수 있는 변수를 선택합니다.

정보 이득(Information Gain):

엔트로피(Entropy)를 기준으로 데이터의 불확실성을 줄이는 정도를 나타냅니다. 정보 이득이 가장 큰 변수가 선택됩니다.

# 4. 트리의 깊이와 가지치기(Pruning)

트리 모형은 기본적으로 변수를 선택해 분할을 진행하지만, 트리가 너무 깊어지면 과적합(overfitting)의 위험이 있습니다. 따라서 트리의 깊이를 제한하거나 가지치기(pruning)를 통해 불필요한 변수를 제거하여 모델의 복잡성을 줄이고, 일반화 성능을 높이는 방법을 사용합니다. 이 트리에서는 상대적으로 얕은 트리 깊이를 유지하면서도 주요 변수들(capital\_gain, education\_num, age)이 사용되어 효율적으로 분할이 이루어졌습니다.

# 5. 결론

트리 모형에서 변수 순서는 데이터의 분할 효과를 극대화하는 방향으로 결정됩니다. 첫 번째 변수는 데이터의 전체 분포를 크게 나눌 수 있는 변수가 선택되고, 이후 노드에서는 각 분기에서 불순도를 최대한 줄일 수 있는 변수들이 순차적으로 선택됩니다. 트리 모형은 이러한 과정을 반복하면서 데이터를 점점 더 세밀하게 분할하며, 최종적으로 각 리프 노드에서 예측 결과를 도출하게 됩니다.

이 트리 모형을 통해, marital\_status, capital\_gain, education\_num, age 등의 변수가 중요한 순서로 선택되었음을 확인할 수 있으며, 이는 각각의 변수가 데이터를 분할하는 데 얼마나 중요한 역할을 했는지를 보여줍니다.