103

- Her özniteliğin bilgi kazancını hesapla
- Karar ağacı düğümüne seçilen özniteliğin değeriyle genişlet
- Başka sınıflandırılmamış örnek kaldı mı?
- Kalan özniteliklerle devam et

Bilgi kazancı (Gain):

• Bilgi kazancı ne kadar yüksekse, bir özellik o kadar iyi bir bölünme kriteri olarak kabul edilir.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

S: veri seti, A: Feature

Entropy = Düğümdeki veri noktalarının sınıflarının ne kadar karışık olduğunu ölçen bir metriktir.

E↓ Homojenlik ↑ H↑ Belirsizlik ↑

* Karışıklık ne kadar yüksek olursa bilgi kazancı yüksek olur.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(S.G) = $Y = \frac{9}{14} = 0,64$, $No = \frac{5}{14} = 0,36$

$= -(0.36 \log_2 0,36) - (0,64 \log_2 0,64)$

$= 0,94$ → Entropisi yüksek bir değer neredeyse 1 olacak

[min entro, max entro)

[0 1]

CART

- Her özniteliğin Gini impurity değerini, En düşük Gini impurity değerine sahip özniteliği seç
- Karar ağacı düğümüne seçilen özniteliğin değeriyle genişlet
- Başka sınıflandırılmamış örnek kaldı mı?
- Kalan özniteliklerle devam et

Gini Impurity (Gini belirsizliği) : Beşte bir karar ağacı algoritması olan CART tarafından kullanılan bir ölçümdür. Veriden, sınıfların dağılımının ne kadar karışık olduğunu [0, 0,5] arasında karışıklık olabilir (2 etiket varsa) [0, 1] Madem etiket varsa gini ordığı bu olur. Gini ne kadar düşükse o kadar homojenizedir.

$$\text{Gini Impurity} = 1 - \sum_{i=0}^{n} P_i^2$$

* Gini ve Entropy karar ağaçlarında node'ları belirlemek için kullandığımız karmaşıklık ölçütleridir. Karmaşıklığın en az olmasını istediğimiz ama karmaşıklığı en fazla olan node olarak seçeriz.

Örnek

Heart Disease	
Yes	No
105	39

144

$$\text{Gini} = 1 - \left[\left(\frac{105}{144}\right)^2 + \left(\frac{39}{144}\right)^2\right]$$

(P_0, P_1)

$= 0,4$ → belirsizlik yüzdesi