Karar Ağaçları (Decision Trees)

Atıl Samancıoğlu

1 Giriş

Karar ağaçları (Decision Trees), hem sınıflandırma (classification) hem de regresyon (regression) problemlerinde kullanılabilen, sezgisel ve güçlü makine öğrenimi algoritmalarıdır. İf-else yapısına benzeyen çalışma mantığı sayesinde insanlar tarafından kolayca anlaşılabilir.

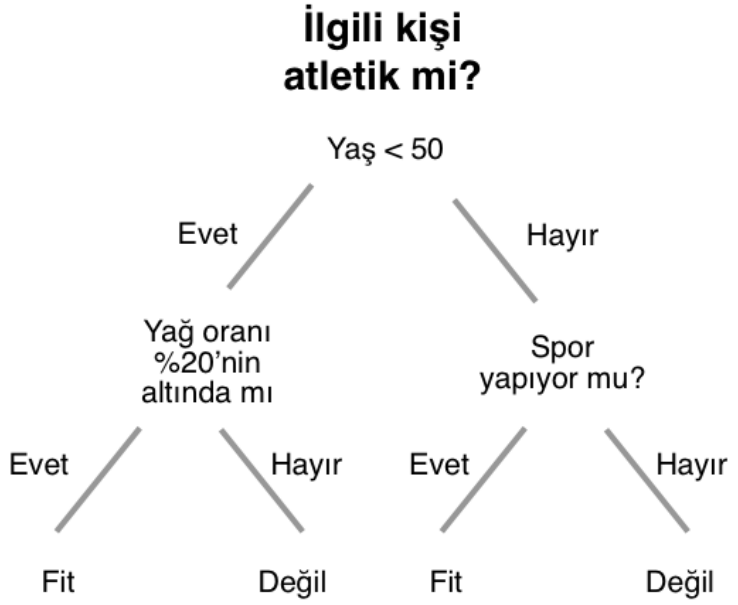


Figure 1: Tree Örnek Sınıflandırma

2 CART ve ID3 Farkı

Karar ağaçlarının farklı türleri vardır:

- **ID3:** Düğümler birden fazla dallanabilir. (örneğin outlook: sunny, rainy, overcast)
- **CART:** Sadece ikili (binary) dallanmalara izin verir.

3 Örnek Veri Seti

Aşağıda 7 satırlık bir kredi onay veri seti bulunmaktadır:

ID	GPA	Interview Score	Approval (Y)
1	<3.0	Düşük	Red
2	<3.0	Yüksek	Onay
3	<3.0	Yüksek	Onay
4	>3.0	Düşük	Red
5	>3.0	Yüksek	Onay
6	>3.0	Normal	Onay
7	<3.0	Normal	Red

Table 1: Karar ağacı için örnek veri seti

4 Karar Ağacında Amaç

Karar ağacında amaç; veri setini öyle bölmek ki, her bölme sonucunda oluşan alt kümeler mümkün olduğunca **homojen** (tek sınıflı) olsun. Bunun için saflık (purity) ölçütleri kullanılır. En yaygın iki ölçüt:

- **Entropy**
- **Gini Impurity**

Bu ölçütleri kullanarak hangi split'in daha iyi olduğunu belirleriz.

5 Entropy: Tanım ve Örnek

Tanım: Entropy, bilgi teorisinden gelen ve bir veri setindeki düzensizliği ölçen bir metriktir. Düşük entropy = daha saf küme.

Formülü:

$$H(S) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

Burada:

- C: Sınıf sayısı
- p_i: Sınıf i'nin olasılığı

Örnek: 7 örnekte 4 Onay (Yes), 3 Red (No):

$$H = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} \approx 0.985$$

6 Gini Impurity: Tanım ve Örnek

Tanım: Gini Impurity, rastgele seçilen bir örneğin yanlış sınıflandırılma olasılığıdır. Basit ve hızlı hesaplanır.

Formülü (iki sınıf için):

$$G(S) = 1 - (p_{\text{Yes}}^2 + p_{\text{No}}^2)$$

Örnek: Yine 4 Onay, 3 Red:

$$G = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 \approx 0.4898$$

7 Adım 1: Split Adayları

- **Split 1:** Interview Score = Yüksek mi?
- **Split 2:** GPA >3.0 mı?

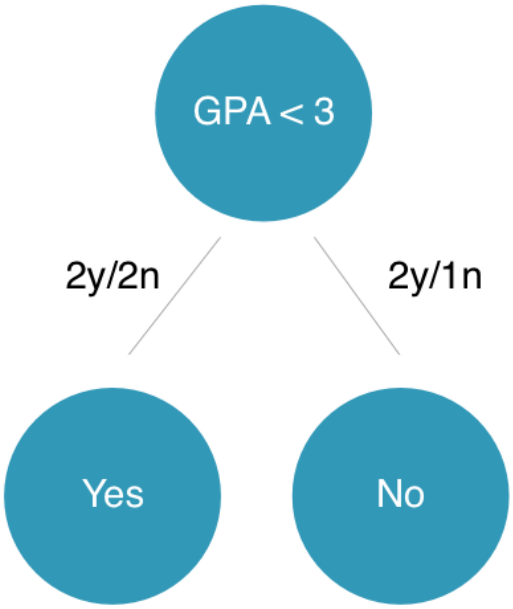


Figure 2: GPA Split Karar Ağacı

8 Adım 2: Entropy Hesaplamaları

Kök Düğüm

Daha önce hesapladığımız gibi:

$$H(\text{root}) \approx 0.985$$

Interview Score = Yüksek

- **Yes grubu:** (ID 2,3,5) → Hepsi Onay → Entropy = 0 - **No grubu:** (ID 1,4,6,7): 1 Onay, 3 Red

$$H = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} \approx 0.811$$

Bilgi kazancı:

$$\text{Gain} = 0.985 - \left(\frac{3}{7} \times 0 + \frac{4}{7} \times 0.811\right) \approx 0.985 - 0.464 = 0.521$$

GPA >3.0

- **Yes grubu:** (ID 4,5,6): 2 Onay, 1 Red

$$H = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \approx 0.918$$

- **No grubu:** (ID 1,2,3,7): 2 Onay, 2 Red

$$H = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$$

Bilgi kazancı:

$$\text{Gain} = 0.985 - \left(\frac{3}{7} \times 0.918 + \frac{4}{7} \times 1\right) \approx 0.985 - (0.393 + 0.571) = 0.021$$

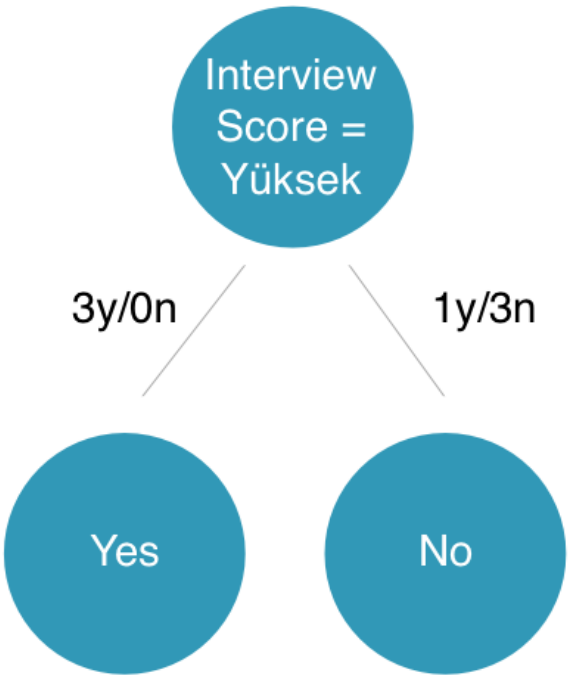


Figure 3: Interview Score Split Karar Ağacı

9 Sonuç: En İyi Split

Interview Score split'inin bilgi kazancı daha yüksek (0.521), bu nedenle seçilir.

10 Karar Ağacı Şeması

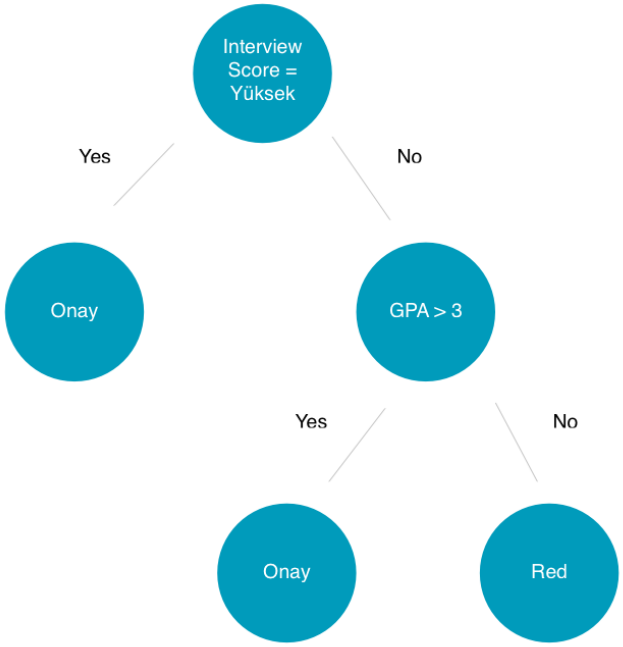


Figure 4: Final Karar Ağacı buna benzer olacaktır

11 Test Örneği

Test veri:

- GPA = 3.2
- Interview Score = Normal

Karar ağacında:

- Interview Score = Yüksek? → Hayır
- GPA >3.0? → Evet → Onay (Yes)

12 Gini ve Entropy Karşılaştırması

- Entropy daha sezgisel ve bilgi teorisine dayanır, genelde daha hesaplama yoğundur.
- Gini Impurity hesaplaması daha hızlıdır ve genelde benzer sonuçlar verir.

13 Pruning: Aşırı Öğrenmeyi Önlemek

Ağaç büyüdükçe overfitting riski artar. Pruning işlemi budama yaparak ağacı sadeleştirir.

Figure 5: Budama Öncesi Karar Ağacı

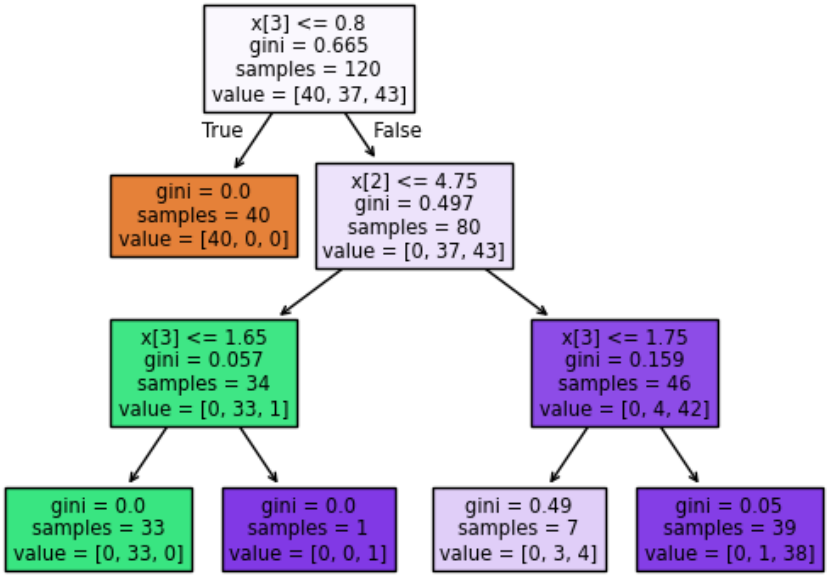


Figure 6: Budama Sonrası Karar Ağacı

14 Decision Tree Regressor ve Varyans Azaltımı

Karar ağaçları yalnızca sınıflandırma değil, aynı zamanda regresyon problemlerinde de kullanılabilir. Eğer hedef değişken sürekli (örneğin maaş, fiyat, sıcaklık) bir değer alıyorsa, **Decision Tree Regressor** kullanılmalıdır.

14.1 Sınıflandırmadan Farkı

Sınıflandırma problemlerinde dallanma saflaştırma metrikleri (entropy, Gini impurity) ile yapılırken, regresyonda bu metrikler uygun değildir. Bunun yerine:

- **Ortalama Kare Hata (Mean Squared Error - MSE)** ya da
- **Varyans Azaltımı (Variance Reduction)**

kullanılarak hangi split'in daha iyi olduğu hesaplanır.

14.2 Varyans Azaltımı Formülü

Aşağıdaki formül ile varyans azaltımı hesaplanır:

$$\text{Variance Reduction} = \text{Var}_{\text{root}} - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \cdot \text{Var}(S_i)$$

Burada:

- Var_{root}: kök düğümdeki varyans
- S_i: dallanmadan sonra oluşan i. alt küme
- $|S_i|/|S|$: alt kümenin oranı (ağırlığı)
- $\text{Var}(S_i)$: alt kümenin varyansı

14.3 Tahmin Nasıl Yapılır?

Decision Tree Regressor bir test örneğini yaprak düğüme ulaştırdığında, bu yapraktaki hedef değerlerin ortalamasını alır ve bu ortalamayı tahmin sonucu olarak döner.

Örnek:

- Bir yaprakta hedef değerler: {52, 60, 56}
- Tahmin edilen değer: $(52 + 60 + 56)/3 = 56$

15 Sürekli Değişkenlerle Çalışmak: GPA Örneği

Karar ağaçları yalnızca kategorik değil, sürekli değişkenlerle de çalışabilir. Örneğin **GPA** gibi sürekli bir değişken ile karar ağacı bölmesi yapmak için tüm olası eşikler denenir.

15.1 Örnek Veri Seti

Aşağıdaki veri seti, öğrencilerin GPA puanları ve onlara verilen burs miktarlarını (hedef değişken) içermektedir:

ID	GPA	Burs Miktarı (bin TL)
1	2.5	10
2	2.7	12
3	3.0	18
4	3.2	22
5	3.5	30
6	3.7	35
7	4.0	40

Table 2: Decision Tree Regressor için örnek veri seti

15.2 Olası Split Eşikleri

Karar ağacı bu veri için GPA değerlerini sıralar ve olası split noktalarını şu şekilde belirler:

$$\frac{2.5+2.7}{2}=2.6, \quad \frac{2.7+3.0}{2}=2.85, \quad \frac{3.0+3.2}{2}=3.1, \quad \frac{3.2+3.5}{2}=3.35, \quad \frac{3.5+3.7}{2}=3.6, \quad \frac{3.7+4.0}{2}=3.85$$

Her eşik için varyans azaltımı hesaplanır ve en iyi split noktası seçilir.

15.3 Varyans Hesaplama (Örnek: GPA <3.1)

Sol Grup (GPA <3.1):

$$\{10, 12, 18\}$$

Ortalama: $(10+12+18)/3 = 13.33$

Varyans:

$$\frac{(10-13.33)^2+(12-13.33)^2+(18-13.33)^2}{3} = \frac{(11.11)+(1.77)+(21.78)}{3} \approx 11.55$$

Sağ Grup (GPA >3.1):

$$\{22, 30, 35, 40\}$$

Ortalama: $(22+30+35+40)/4 = 31.75$

Varyans:

$$\frac{(22-31.75)^2+(30-31.75)^2+(35-31.75)^2+(40-31.75)^2}{4} = \frac{(95.06)+(3.06)+(10.56)+(68.06)}{4} \approx 44.19$$

Kök Varyans: Tüm veri için ortalama: $(10+12+18+22+30+35+40)/7 = 23.86$

Varyans:

$$\frac{(10-23.86)^2+(12-23.86)^2+(18-23.86)^2+(22-23.86)^2+(30-23.86)^2+(35-23.86)^2+(40-23.86)^2}{7} \approx 124.83$$

15.4 Varyans Azaltımı

Ağırlıklı varyans:

$$\frac{3}{7} \times 11.55 + \frac{4}{7} \times 44.19 \approx 4.95 + 25.24 = 30.19$$

Varyans azaltımı:

$$124.83 - 30.19 = 94.64$$

15.5 Diğer Split'lerin Kısa Karşılaştırması

Karar ağacı tüm eşikleri dener ve varyans azaltımını hesaplar. Yukarıda detaylıca hesapladığımız **GPA <3.1** split'i yaklaşık **94.64** varyans azaltımı sağlamıştı. Diğer olası eşikler için özet hesaplamalar:

Eşik	Varyans Azaltımı	Açıklama
2.6	37.5	Çok küçük bölme; sağ grup geniş.
2.85	68.2	İyi ama optimum değil.
3.1	94.6	En yüksek varyans azaltımı.
3.35	80.3	Orta.
3.6	72.1	Daha zayıf.
3.85	60.4	En sağdaki değerler ayrışıyor ama varyans yüksek.

Table 3: Farklı eşiklerdeki varyans azaltımları

Görüldüğü gibi, **3.1 eşiği** en yüksek varyans azaltımını sağlamaktadır. Bu yüzden karar ağacı bu split'i seçmiştir. Split seçimi sırasında amaç, veri setini ikiye ayırırken **en homojen grupları** oluşturmaktır; varyans azaltımı bu homojenliği ölçer.

16 Decision Tree Regressor: Tahmin Mantığı

Decision Tree Regressor, test örneğini yaprak düğüme ulaştırdığında, bu yapraktaki hedef değerlerin ortalamasını döner.

Örnek: Bir test öğrencisi:

$$\text{GPA} = 3.3$$

Bu örnek, örneğin GPA > 3.1 olan dala gider. Bu dalda değerler:

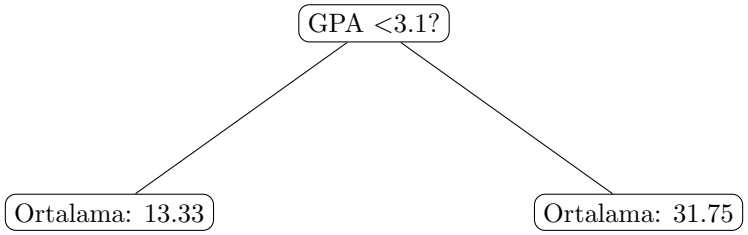
$$\{22, 30, 35, 40\}$$

Bu yapraktaki ortalama:

$$(22 + 30 + 35 + 40)/4 = 31.75$$

Dolayısıyla tahmin sonucu **31.75 bin TL** olur.

17 Karar Ağacı Şeması (Regresyon)



Bu basit örnek, Decision Tree Regressor'ın mantığını ve nasıl çalıştığını açık bir şekilde göstermektedir.

18 Sonuç

Karar ağaçları, hem matematiksel temelleri olan hem de sezgisel olarak güçlü yöntemlerdir. Entropy, Gini Impurity ve Information Gain gibi kavramlar sayesinde optimal bölünmeler yapılabilir ve etkili sınıflandırmalar gerçekleştirilebilir.