

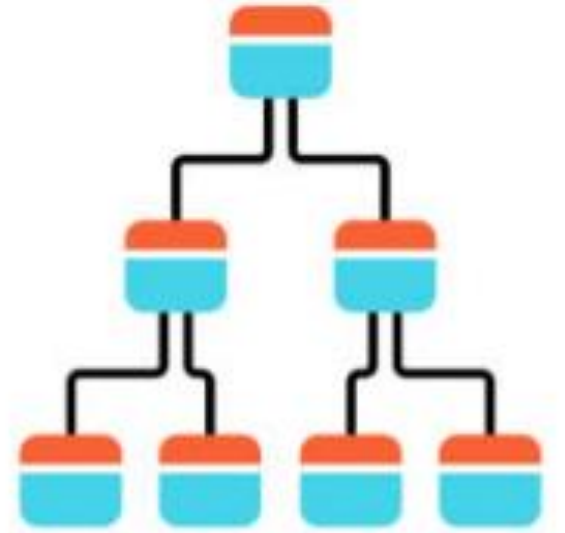
Örüntü Tanıma

Dr. Öğr. Üyesi Mehmet Zahid YILDIRIM

e-mail: m.zahidyildirim@karabuk.edu.tr

Karar Ağaçları (Decision Trees)

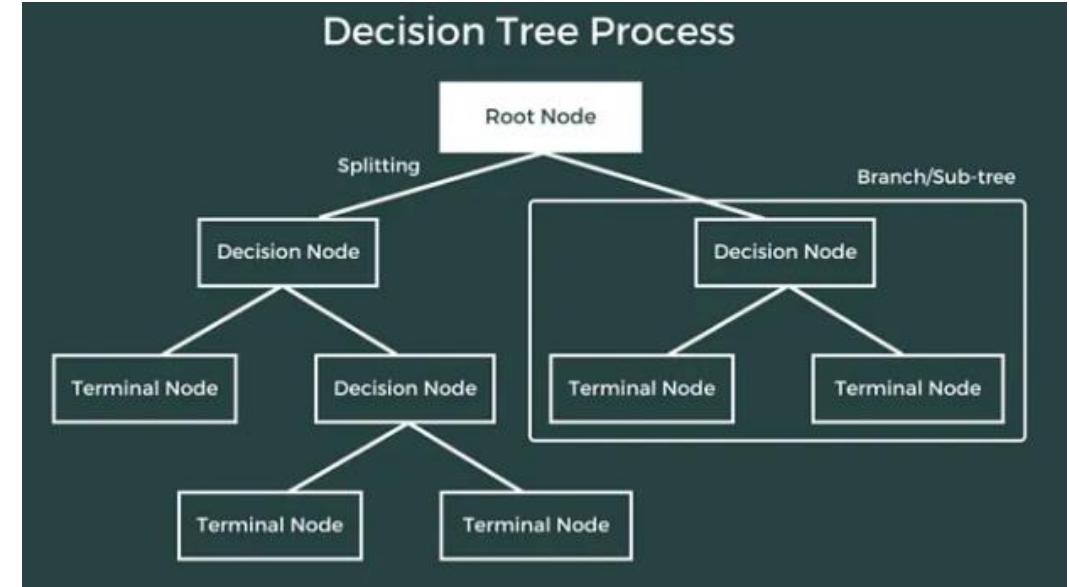
- Karar ağacı, hem regresyon hem de sınıflandırma için kullanılabilen bir tür denetimli makine öğrenimi modelidir.
- Girdi değişken(ler)i ile çıktı değişkenini tek bir ağaç formunda gösterebilen bir algoritmadır.
- Entropiye dayalı sınıflandırma ağaçları (ID3, C4.5) ve Regresyon ağaçları (CART) olmak üzere iki kategoride birçok algoritma önerilmiştir.
- Tümevarımlı mantığın programlama ortamına taşındığı basit fakat çok yaygın bir metottur.



Karar Ağaçları (Decision Trees)

Karar Ağaçları ile ilgili Önemli Terminolojiler

1. Kök Düğüm (Root Node): Tüm popülasyonu veya örneği temsil eder ve bu daha sonra iki veya daha fazla homojen kümeye ayrılır.
2. Bölme (Splitting): Bir düğümü iki veya daha fazla alt düğüme bölme işlemidir.
3. Karar Düğümü (Decision Node): Bir alt düğüm başka alt düğümlere bölündüğünde, buna karar düğümü denir.
4. Yaprak / Terminal Düğüm (Leaf / Terminal Node): Bölünmeyen düğümlere Yaprak veya Terminal düğüm denir.



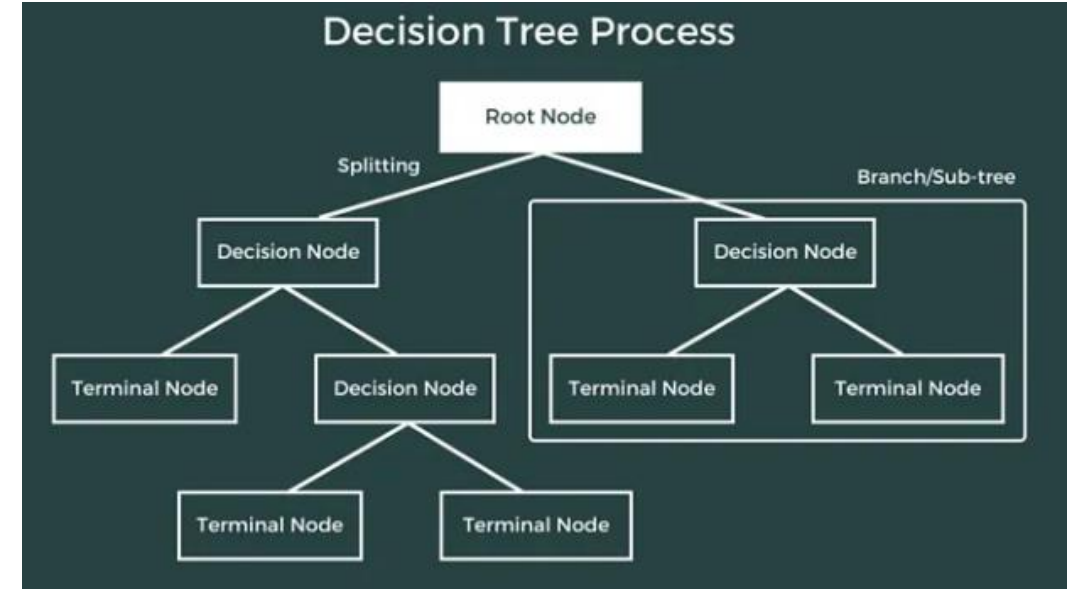
Karar Ağaçları (Decision Trees)

Karar Ağaçları ile ilgili Önemli Terminolojiler

5. Budama (Pruning): Bir karar düğümünün alt düğümlerini çıkardığımızda, bu işleme budama denir. Bölme işleminin tam tersi diyebiliriz.

6. Dal / Alt Ağaç (Branch / Sub-Tree): Tüm ağacın bir alt bölümüne dal veya alt ağaç denir.

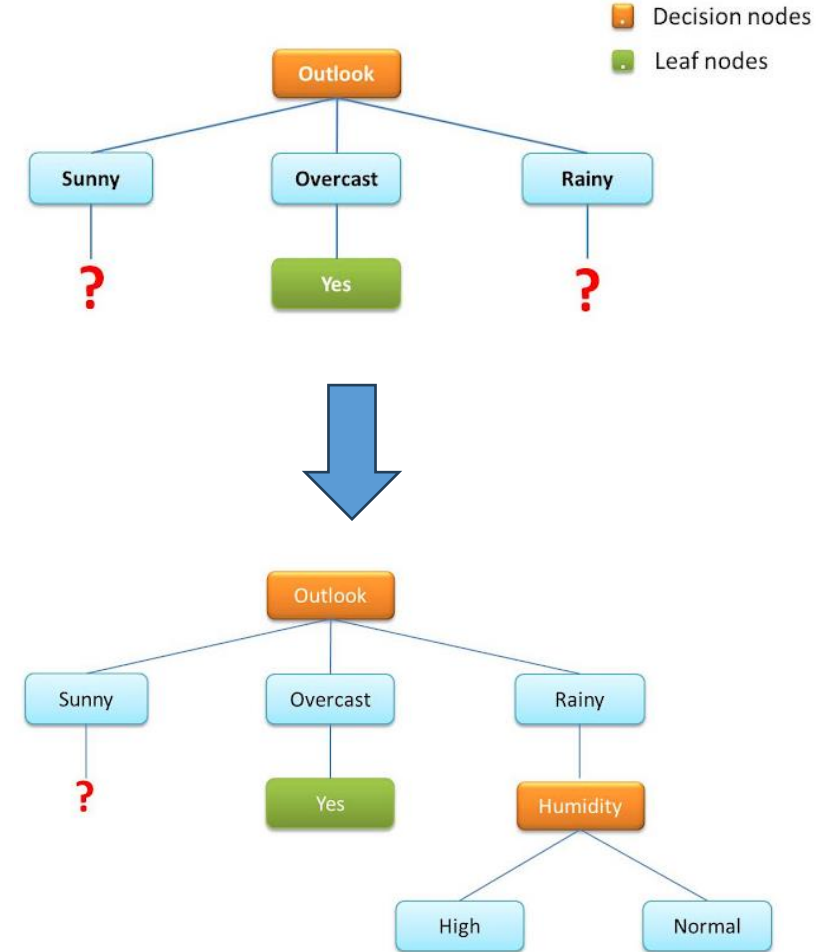
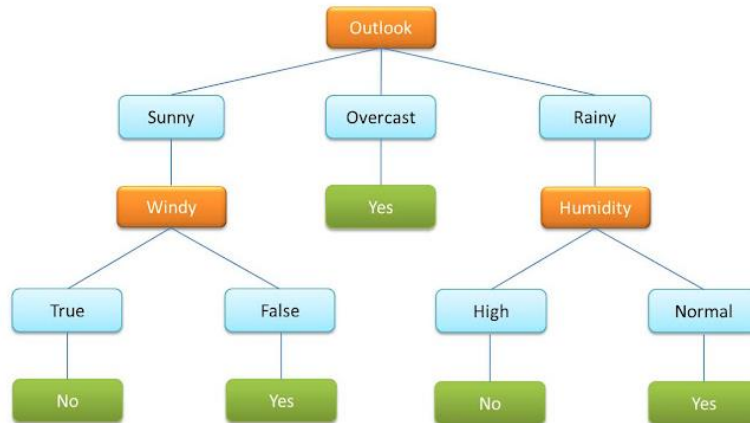
7. Ebeveyn ve Çocuk Düğüm (Parent and Child Node): Alt düğümlere ayrılan bir düğüm, alt düğümlerin ana düğümü olarak adlandırılırken, alt düğümler bir ana düğümün çocuğudur.



Karar Ağaçları (Decision Trees)

Algoritma Adımları

1. Her öznitelik üzerinde bölme için bilgi kazanımını (Information Gain) hesapla.
2. Yeni düğüm için en fazla bilgi kazanımını veren özneliği seç.
3. Seçilen özneliğe göre veri setini böl.
4. Alt düğümler için her düğüm için özyinelemeli olarak ilk 3 adımı tekrarla.



Karar Ağaçları (Decision Trees)

Karar Düğümlerini oluşturacak En İyi Niteliğin Seçimi

- Kök düğümden itibaren ağacın dengeli dağılması ve sınıflandırma algoritmasının verimli olması için öznitelikler arasından en uygun olanı seçilmelidir.
- En uygun özniteliğin seçimi için **Entropi** değerinin hesaplanması gerekir.

$$\text{Entropy} = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$



$$\text{Entropy} = \sum_{i=1}^C -p_i * \log_2(p_i)$$

Karar Ağaçları (Decision Trees)

Entropi

- p_i toplamı 1 olan olasılıklar.
- Örnekler aynı sınıfa ait ise **entropi = 0**.
- Örnekler sınıflara eşit olarak dağılmış ise **entropi = 1**;
- Örnekler sınıflar arasında rasgele dağılmış ise **$0 < \text{entropi} < 1$** olacaktır.

$$\text{Entropy} = \sum_{i=1}^C -p_i * \log_2(p_i)$$

Karar Ağaçları (Decision Trees)

Entropi – Örnek

V1	V2	S
A	C	E
B	C	F
B	D	E
B	D	F

Sınıf Entropisi

$$H(S) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

V1 Entropisi

$$\begin{aligned}H(V1) &= \frac{1}{4}H(A) + \frac{3}{4}H(B) \\&= \frac{1}{4}0 - \frac{3}{4}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \\&= 0 + \frac{3}{4}0,9183 = 0,6887\end{aligned}$$

V2 Entropisi

$$H(V2) = \frac{1}{2}H(C) + \frac{1}{2}H(D) = \frac{1}{2} + \frac{1}{2} = 1$$

V1 seçilir..

$$Entropy = \sum_{i=1}^C -p_i * \log_2(p_i)$$

Karar Ağaçları (Decision Trees)

Bilgi Kazancı

- Bir özneliliğin ne kadar bilgi sakladığının ölçütü olarak kullanılır.
- Herhangi bir özelliğe ait bilgi kazanımını hesaplamak için, sınıf entropi değerinden o özelliğe ait entropi değeri arasındaki farka bakılır.

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Karar Ağaçları (Decision Trees)

Bilgi Kazancı - Örnek

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

V1	V2	S
A	C	E
B	C	F
B	D	E
B	D	F

Sınıf Entropisi

$$H(S) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

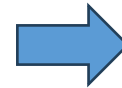
V1 Entropisi

$$\begin{aligned} H(V1) &= \frac{1}{4}H(A) + \frac{3}{4}H(B) \\ &= \frac{1}{4}0 - \frac{3}{4}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \\ &= 0 + \frac{3}{4}0,9183 = 0,6887 \end{aligned}$$

V2 Entropisi

$$H(V2) = \frac{1}{2}H(C) + \frac{1}{2}H(D) = \frac{1}{2} + \frac{1}{2} = 1$$

V1 seçilir..



$$Gain(V1) = H(S) - H(V1) = 1 - 0.6887 = \mathbf{0.3113}$$

$$Gain(V2) = H(S) - H(V2) = 1 - 1 = \mathbf{0}$$

Karar Ağaçları (Decision Trees)

Örnek: Veri setindeki bilgileri kullanarak hava durumuna bağlı olarak golf oynanıp oynanamayacağına karar vermek için kullanılacak Karar Ağacı yapısını oluşturunuz.

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Karar Ağaçları (Decision Trees)

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Play Golf(14)	
Yes	No
9	5

→ $Entropy = \sum_{i=1}^C -p_i * \log_2(p_i)$

$Entropy(PlayGolf) = E(5,9)$

$$E(PlayGolf) = E(5,9)$$

$$\begin{aligned} &= -\left(\frac{9}{14} \log_2 \frac{9}{14}\right) - \left(\frac{5}{14} \log_2 \frac{5}{14}\right) \\ &= -(0.357 \log_2 0.357) - (0.643 \log_2 0.643) \\ &= 0.94 \end{aligned}$$

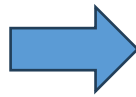
Karar Ağaçları (Decision Trees)

Entropi, iki veya daha fazla özellik için aşağıdaki formül ile hesaplanır.

$$Entropy(S, T) = \sum_{c \in T} P(c)E(c)$$

Karar Ağaçları (Decision Trees)

		PlayGolf(14)		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5



$$E(\text{PlayGolf}, \text{Outlook}) = \frac{5}{14}E(3,2) + \frac{4}{14}E(4,0) + \frac{5}{14}E(2,3)$$

$$E(\text{Sunny}) = E(3,2)$$

$$= -\left(\frac{3}{5} \log_2 \frac{3}{5}\right) - \left(\frac{2}{5} \log_2 \frac{2}{5}\right)$$

$$= -(0.60 \log_2 0.60) - (0.40 \log_2 0.40)$$

$$= -(0.60 * 0.737) - (0.40 * 0.529)$$

$$= 0.971$$

$$E(\text{Rainy}) = E(2,3)$$

$$= -\left(\frac{2}{5} \log_2 \frac{2}{5}\right) - \left(\frac{3}{5} \log_2 \frac{3}{5}\right)$$

$$= -(0.40 \log_2 0.40) - (0.6 \log_2 0.60)$$

$$= 0.971$$

$$E(\text{Overcast}) = E(4,0)$$

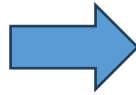
$$= -\left(\frac{4}{4} \log_2 \frac{4}{4}\right) - \left(\frac{0}{4} \log_2 \frac{0}{4}\right)$$

$$= -(0) - (0)$$

$$= 0$$

Karar Ağaçları (Decision Trees)

		PlayGolf(14)		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5



$$E(\text{PlayGolf}, \text{Outlook}) = \frac{5}{14}E(3,2) + \frac{4}{14}E(4,0) + \frac{5}{14}E(2,3)$$

$$E(\text{PlayGolf}, \text{Outlook}) = P(\text{Sunny}) E(3,2) + P(\text{Overcast}) E(4,0) + P(\text{rainy}) E(2,3)$$

$$E(\text{PlayGolf}, \text{Outlook}) = \frac{5}{14}E(3,2) + \frac{4}{14}E(4,0) + \frac{5}{14}E(2,3)$$

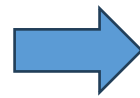
$$= \frac{5}{14}0.971 + \frac{4}{14}0.0 + \frac{5}{14}0.971$$

$$= 0.357 * 0.971 + 0.0 + 0.357 * 0.971$$

$$= 0.693$$

Karar Ağaçları (Decision Trees)

		PlayGolf(14)		
		Yes	No	
Temperature	Hot	2	2	4
	Cold	3	1	4
	Mild	4	2	6



$$E(\text{PlayGolf}, \text{Temperature}) = P(\text{Hot}) E(2,2) + P(\text{Cold}) E(3,1) + P(\text{Mild}) E(4,2)$$

$$E(\text{PlayGolf}, \text{Temperature}) = 4/14 * E(\text{Hot}) + 4/14 * E(\text{Cold}) + 6/14 * E(\text{Mild})$$

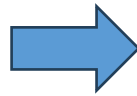
$$E(\text{PlayGolf}, \text{Temperature}) = 4/14 * E(2, 2) + 4/14 * E(3, 1) + 6/14 * E(4, 2)$$

$$\begin{aligned} E(\text{PlayGolf}, \text{Temperature}) = & 4/14 * -(2/4 \log 2/4) - (2/4 \log 2/4) \\ & + 4/14 * -(3/4 \log 3/4) - (1/4 \log 1/4) \\ & + 6/14 * -(4/6 \log 4/6) - (2/6 \log 2/6) \end{aligned}$$

$$\begin{aligned} E(\text{PlayGolf}, \text{Temperature}) = & 5/14 * 1.0 \\ & + 4/14 * 1.811 \\ & + 5/14 * 0.918 \\ = & 0.911 \end{aligned}$$

Karar Ağaçları (Decision Trees)

		PlayGolf(14)		
		Yes	No	
Humidity	High	3	4	7
	Normal	6	1	7



$$E(\text{PlayGolf, Humidity}) = 7/14 * E(\text{High}) + 7/14 * E(\text{Normal})$$

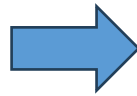
$$E(\text{PlayGolf, Humidity}) = 7/14 * E(3, 2) + 7/14 * E(4, 0)$$

$$\begin{aligned} E(\text{PlayGolf, Humidity}) &= 7/14 * -(3/7 \log 3/7) - (4/7 \log 4/7) \\ &\quad + 7/14 * -(6/7 \log 6/7) - (1/7 \log 1/7) \end{aligned}$$

$$\begin{aligned} E(\text{PlayGolf, Humidity}) &= 7/14 * 0.985 \\ &\quad + 7/14 * 0.592 \\ &= 0.788 \end{aligned}$$

Karar Ağaçları (Decision Trees)

		PlayGolf(14)		
		Yes	No	
Windy	TRUE	3	3	6
	FALSE	6	2	8



$$E(\text{PlayGolf, Windy}) = 6/14 * E(\text{True}) + 8/14 * E(\text{False})$$

$$E(\text{PlayGolf, Windy}) = 6/14 * E(3, 3) + 8/14 * E(6, 2)$$

$$\begin{aligned} E(\text{PlayGolf, Windy}) &= 6/14 * -(3/6 \log 3/6) - (3/6 \log 3/6) \\ &\quad + 8/14 * -(6/8 \log 6/8) - (2/8 \log 2/8) \end{aligned}$$

$$\begin{aligned} E(\text{PlayGolf, Windy}) &= 6/14 * 1.0 \\ &\quad + 8/14 * 0.811 \\ &= 0.892 \end{aligned}$$

Karar Ağaçları (Decision Trees)

$$E(\text{PlayGolf}, \text{Outlook}) = \mathbf{0.693}$$

$$E(\text{PlayGolf}, \text{Temperature}) = \mathbf{0.911}$$

$$E(\text{PlayGolf}, \text{Humidity}) = \mathbf{0.788}$$

$$E(\text{PlayGolf}, \text{Windy}) = \mathbf{0.892}$$

$$\text{Gain}(\text{PlayGolf}, \text{Outlook}) = \text{Entropy}(\text{PlayGolf}) - \text{Entropy}(\text{PlayGolf}, \text{Outlook})$$

$$= 0.94 - 0.693 = \mathbf{0.247}$$

$$\text{Gain}(\text{PlayGolf}, \text{Temperature}) = \text{Entropy}(\text{PlayGolf}) - \text{Entropy}(\text{PlayGolf}, \text{Temperature})$$

$$= 0.94 - 0.911 = \mathbf{0.029}$$

$$\text{Gain}(\text{PlayGolf}, \text{Humidity}) = \text{Entropy}(\text{PlayGolf}) - \text{Entropy}(\text{PlayGolf}, \text{Humidity})$$

$$= 0.94 - 0.788 = \mathbf{0.152}$$

$$\text{Gain}(\text{PlayGolf}, \text{Windy}) = \text{Entropy}(\text{PlayGolf}) - \text{Entropy}(\text{PlayGolf}, \text{Windy})$$

$$= 0.94 - 0.892 = \mathbf{0.048}$$

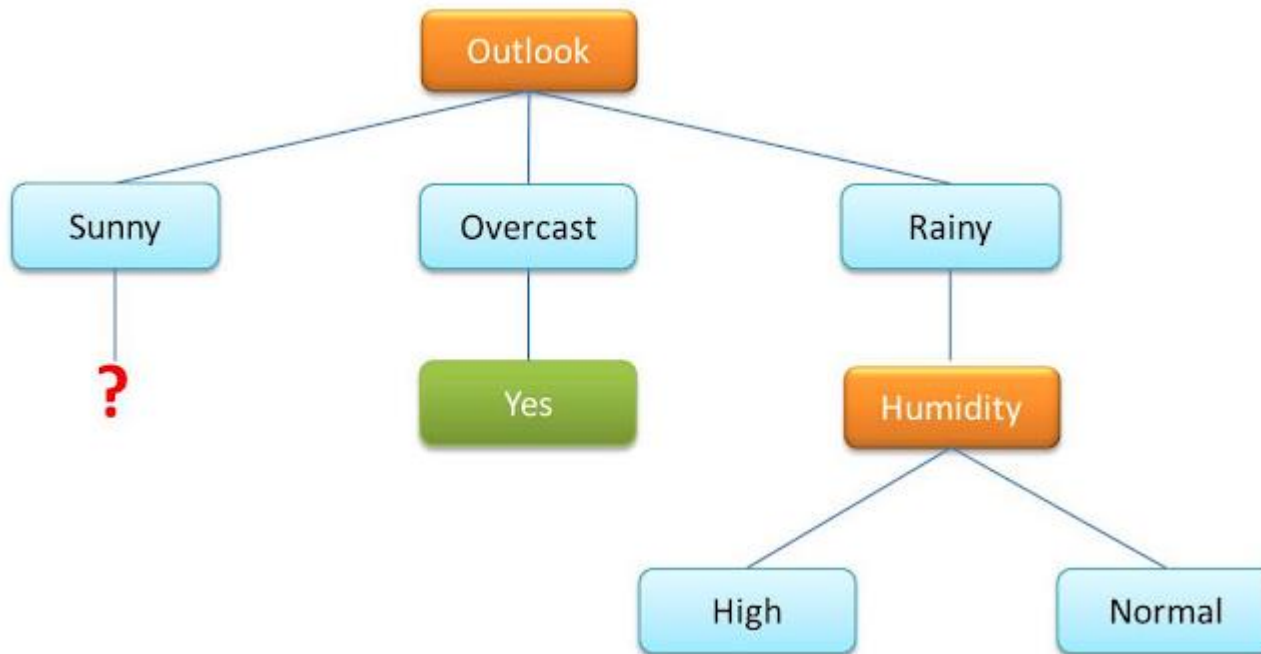


Karar Ağaçları (Decision Trees)



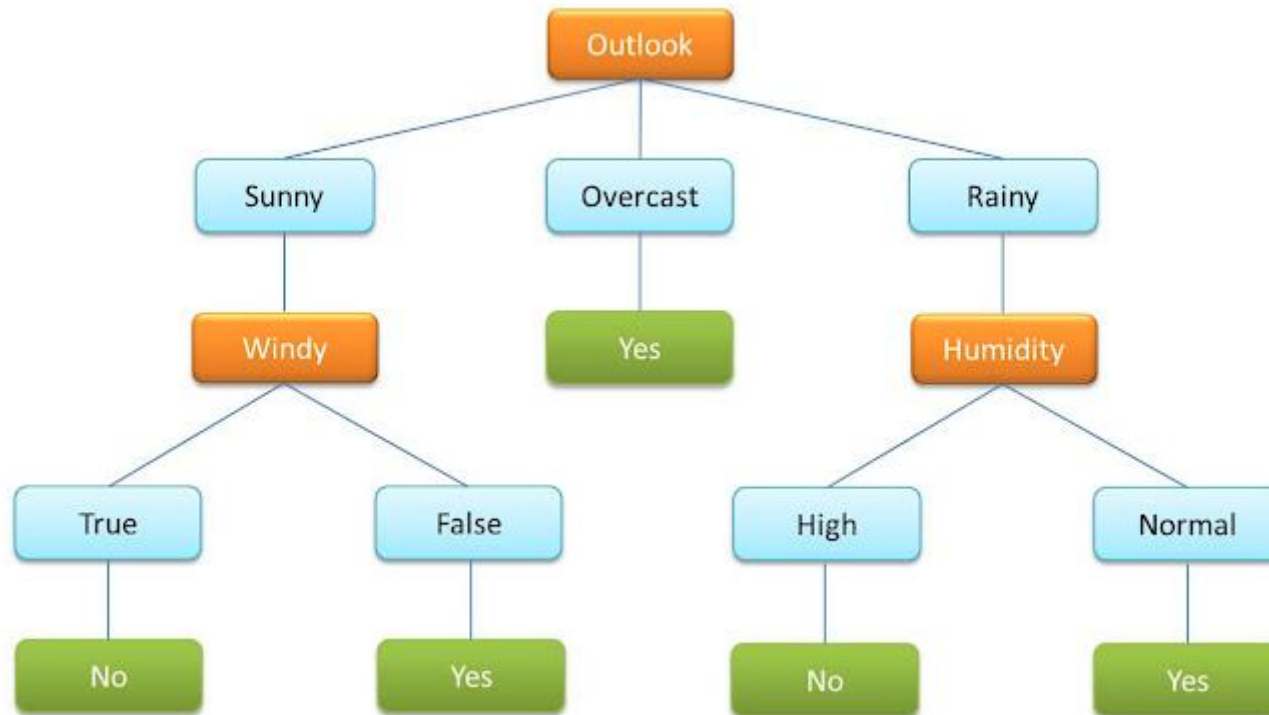
Outlook	Temperature	Humidity	Windy	Play Golf
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Karar Ağaçları (Decision Trees)



Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Karar Ağaçları (Decision Trees)



Outlook	Temperature	Humidity	Windy	Play Golf
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	High	TRUE	No