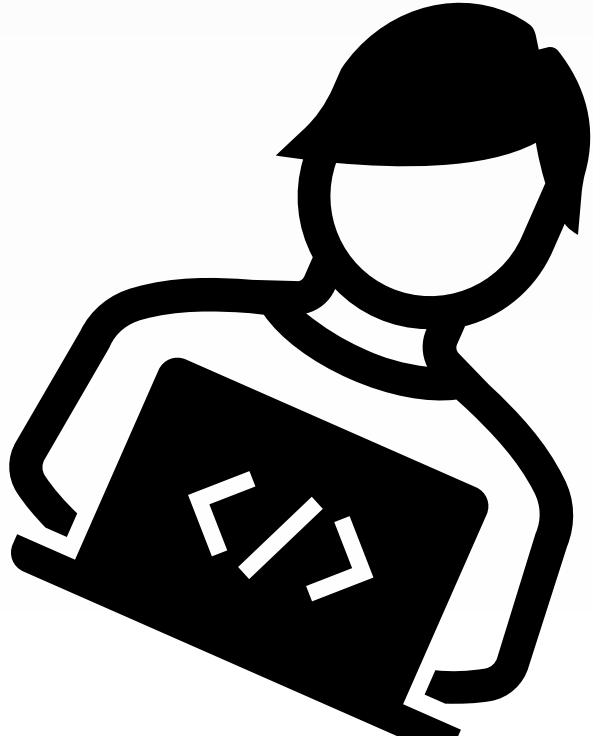
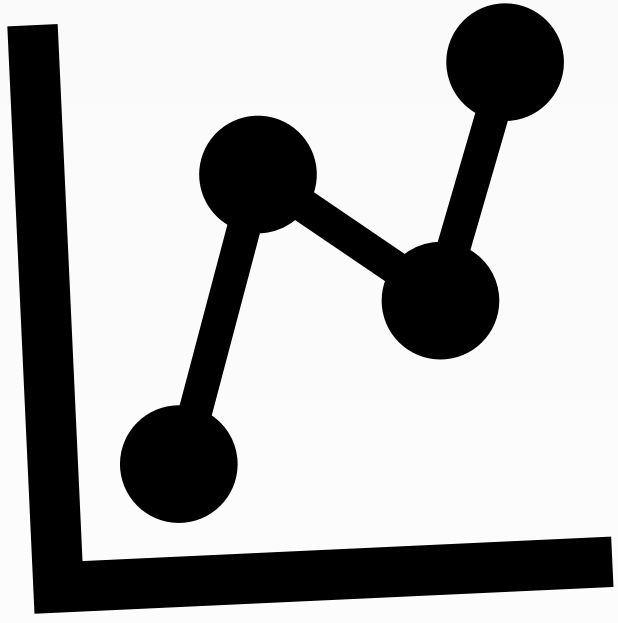




EKO 469-Veri Madenciliđi

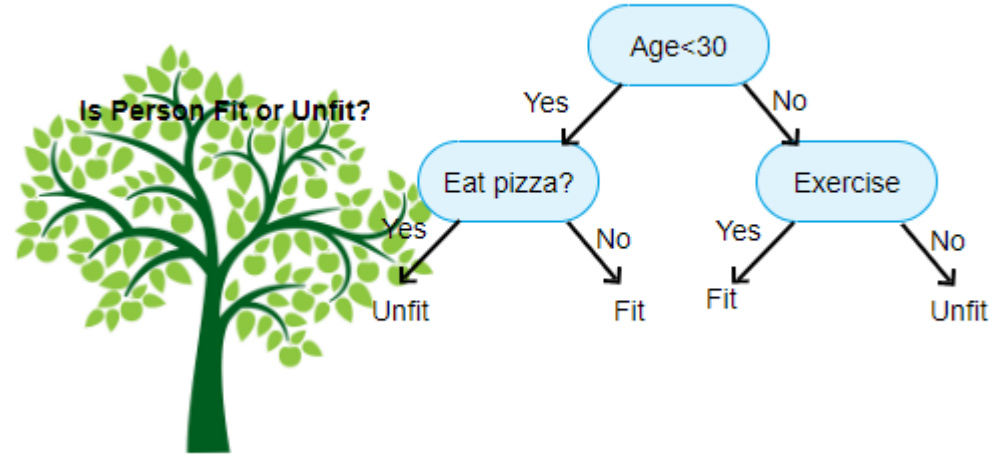
Hafta 11 – Sınıflandırma

Sınıflandırma ve Regresyon Ağaçları (CART)



- Sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin **kategorik** veya **süreklilik gösteren** bir değere sahip olmasıdır.
- Ancak çok terimli lojistik regresyon (multinomial logistic regression) gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerle, her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır.
- **Sınıflama** kategorik değerleri tahmin ederken, **regresyon** süreklilik gösteren değerlerin tahmin edilmesinde kullanılır.
- Örneğin bir **sınıflama modeli** banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, **regresyon modeli** geliri ve mesleği verilen potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir.

Sınıflandırma ve Regresyon Ağaçları (CART)



- CART karar ağacı, her bir karar düğümünden itibaren ağacın iki dala ayrılması ilkesine dayanır.
- Bir düğümden sadece iki dal ayrılabilir.
- Bölünme için önce tüm niteliklerin var olduğu değerler göz önüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilir.
- Bu bölünmeler üzerinde seçme işlemi uygulanır.
- **Twoing** ve **Gini** algoritmaları CART algoritmalarına örnek verilebilir.
- Karar ağacı tekniğini kullanarak verinin sınıflanması iki basamaklı bir işlemdir:
 - ❑ İlk basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından analiz edilir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir.
 - ❑ İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır.

Gini algoritması

- İkili bölünmeler şeklinde gerçekleşen bir sınıflandırma yöntemidir.
- Algoritma nitelik değerlerinin sol ve sağda olmak üzere ikili bölünmeler şeklinde ayrılması temeline dayanır.

Uygulama Şekli :

- Nitelik değerlerinin her biri ikili bölünmeler olacak şekilde sınıflanır. Elde edilen sol ve sağ bölünmelere karşılık gelen sınıf değerleri gruplandırılır.
- Her bir düğümde ilgili sol ve sağ bölünmeler için ayrı ayrı hesaplamalar yapılır.

Herbir nitelik ile ilgili sol ve sağ bölünmeler için $Gini_{sol}$ ve $Gini_{sağ}$ ifadeleri

L_i sol daldaki i grubundaki örnek(lerin) sayısı,

R_i sağ daldaki i grubundaki örnek(lerin) sayısı,

k sınıfların sayısı,

T düğümdeki örnekler

$|T_{sol}|$ Sol daldaki örnek(lerin) sayısı,

$|T_{sağ}|$ Sağ daldaki örnek(lerin) sayısı.

tanımlamaları ile aşağıdaki bağıntılar hesaplanabilecektir.

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2 \quad Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2$$

Her bir j niteliği için n eğitim kümesindeki kayıt sayısı olmak üzere aşağıdaki bağıntı hesaplanır.

$$Gini_j = \frac{1}{n} (|T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ})$$

Her j niteliği için hesaplanan $Gini_j$ ifadelerinden en küçük olanı seçilir ve bölünme bu nitelik üzerinden yapılır.

Sonraki aşamada işlemler tekrar edilir.

Gini algoritması-Örnek

- Tabloyu dikkate alarak Gini algoritması ile sınıflandırma yapalım.

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2$$

$$Gini_j = \frac{1}{n} (|T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ})$$

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
2	İlkokul	Genç	Erkek	Hayır
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet
7	İlkokul	Genç	Kadın	Hayır

Gini algoritması-Örnek

- Nitelik değerlerini ikili gruplandır

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2$$

$$Gini_j = \frac{1}{n} (|T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ})$$

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
2	İlkokul	Genç	Erkek	Hayır
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet
7	İlkokul	Genç	Kadın	Hayır

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Genç	Orta Yaşlı	Kadın	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1

Gini algoritması-Örnek

- Her nitelik için $Gini_{sol}$ ve $Gini_{sağ}$ hesapla

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
2	İlkokul	Genç	Erkek	Hayır
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet
7	İlkokul	Genç	Kadın	Hayır

Eğitim için

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2 = 1 - \left[\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right] = 0,444$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2 = 1 - \left[\left(\frac{3}{4} \right)^2 + \left(\frac{1}{4} \right)^2 \right] = 0,375$$

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Genç	Orta Yaşlı	Kadın	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1

Gini algoritması-Örnek

- Her nitelik için $Gini_{sol}$ ve $Gini_{sağ}$ hesapla

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
2	İlkokul	Genç	Erkek	Hayır
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet
7	İlkokul	Genç	Kadın	Hayır

Yaş için

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2 = 1 - \left[\left(\frac{0}{2} \right)^2 + \left(\frac{2}{2} \right)^2 \right] = 0$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2 = 1 - \left[\left(\frac{4}{5} \right)^2 + \left(\frac{1}{5} \right)^2 \right] = 0,320$$

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Genç	Orta Yaşlı	Kadın	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1

Gini algoritması-Örnek

- Her nitelik için $Gini_{sol}$ ve $Gini_{sağ}$ hesapla

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
2	İlkokul	Genç	Erkek	Hayır
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet
7	İlkokul	Genç	Kadın	Hayır

Cinsiyet için

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2 = 1 - \left[\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right] = 0,444$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2 = 1 - \left[\left(\frac{3}{4} \right)^2 + \left(\frac{1}{4} \right)^2 \right] = 0,375$$

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Genç	Orta Yaşlı	Kadın	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1

Gini algoritması-Örnek

- $Gini_j$ değerlerini hesapla

$$Gini_j = \frac{1}{n} (|T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ})$$

$$Gini_{Eğitim} = \frac{3(0,444) + 4(0,375)}{7} = 0,405$$

$$Gini_{Yaş} = \frac{2(0) + 5(0,320)}{7} = 0,229$$

$$Gini_{Cinsiyet} = \frac{3(0,444) + 4(0,375)}{7} = 0,405$$

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Genç	Orta Yaşlı	Kadın	Erkek
Evet	1	3	0	4	1	3
Hayır	2	1	2	1	2	1
$Gini_{sol}, Gini_{sağ}$	0,444	0,375	0,000	0,320	0,444	0,375
$Gini_j$	0,405		0,229		0,405	

Gini algoritması-Örnek

- En küçük $Gini_j$ değerini seç
 - En küçük $Gini_j$ değeri **Yaş** için olduğundan, kökten itibaren bölünme **Yaş = Genç ve Yaş $\in$ {Orta, Yaşlı}** biçiminde olacaktır.
 - Tabloda Yaş = Genç değerleri (2,7) satırları üzerinde olduğu için bölünme (2,7) ve geri kalan satırlardan (1,3,4,5,6) oluşacaktır.

Birinci Bölünme



Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
2	İlkokul	Genç	Erkek	Hayır
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet
7	İlkokul	Genç	Kadın	Hayır

Gini algoritması-Örnek

- Eğitim kümesinden (2,7) satırlarını çıkartıp aynı işlemleri tekrarla

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet

Eğitim için

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2 = 1 - \left[\left(\frac{1}{1} \right)^2 + \left(\frac{0}{1} \right)^2 \right] = 0$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2 = 1 - \left[\left(\frac{3}{4} \right)^2 + \left(\frac{1}{4} \right)^2 \right] = 0,375$$

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Orta	Yaşlı	Kadın	Erkek
Evet	1	3	2	2	1	3
Hayır	0	1	1	0	1	0

Gini algoritması-Örnek

- Eğitim kümesinden (2,7) satırlarını çıkartıp aynı işlemleri tekrarla

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet

Yaş için

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2 = 1 - \left[\left(\frac{2}{3} \right)^2 + \left(\frac{1}{3} \right)^2 \right] = 0,444$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2 = 1 - \left[\left(\frac{2}{2} \right)^2 + \left(\frac{0}{2} \right)^2 \right] = 0$$

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Orta	Yaşlı	Kadın	Erkek
Evet	1	3	2	2	1	3
Hayır	0	1	1	0	1	0

Gini algoritması-Örnek

- Eğitim kümesinden (2,7) satırlarını çıkartıp aynı işlemleri tekrarla

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet

Cinsiyet için

$$Gini_{sol} = 1 - \sum_{i=1}^k \left(\frac{L_i}{|T_{sol}|} \right)^2 = 1 - \left[\left(\frac{1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 \right] = 0,500$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left(\frac{R_i}{|T_{sağ}|} \right)^2 = 1 - \left[\left(\frac{3}{3} \right)^2 + \left(\frac{0}{3} \right)^2 \right] = 0$$

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Orta	Yaşlı	Kadın	Erkek
Evet	1	3	2	2	1	3
Hayır	0	1	1	0	1	0

Gini algoritması-Örnek

- $Gini_j$ değerlerini hesapla

$$Gini_j = \frac{1}{n} (|T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ})$$

$$Gini_{Eğitim} = \frac{1(0) + 4(0,375)}{5} = 0,300$$

$$Gini_{Yaş} = \frac{3(0,444) + 2(0)}{5} = 0,267$$

$$Gini_{Cinsiyet} = \frac{2(0,500) + 3(0)}{5} = 0,200$$

Kabul	Eğitim		Yaş		Cinsiyet	
	İlkokul	Ortaokul Yüksek	Orta	Yaşlı	Kadın	Erkek
Evet	1	3	2	2	1	3
Hayır	0	1	1	0	1	0
$Gini_{sol}, Gini_{sağ}$	0,000	0,375	0,444	0,000	0,500	0,000
$Gini_j$	0,300		0,267		0,200	

Gini algoritması-Örnek

- İkinci adımda karar ağacı görünümü



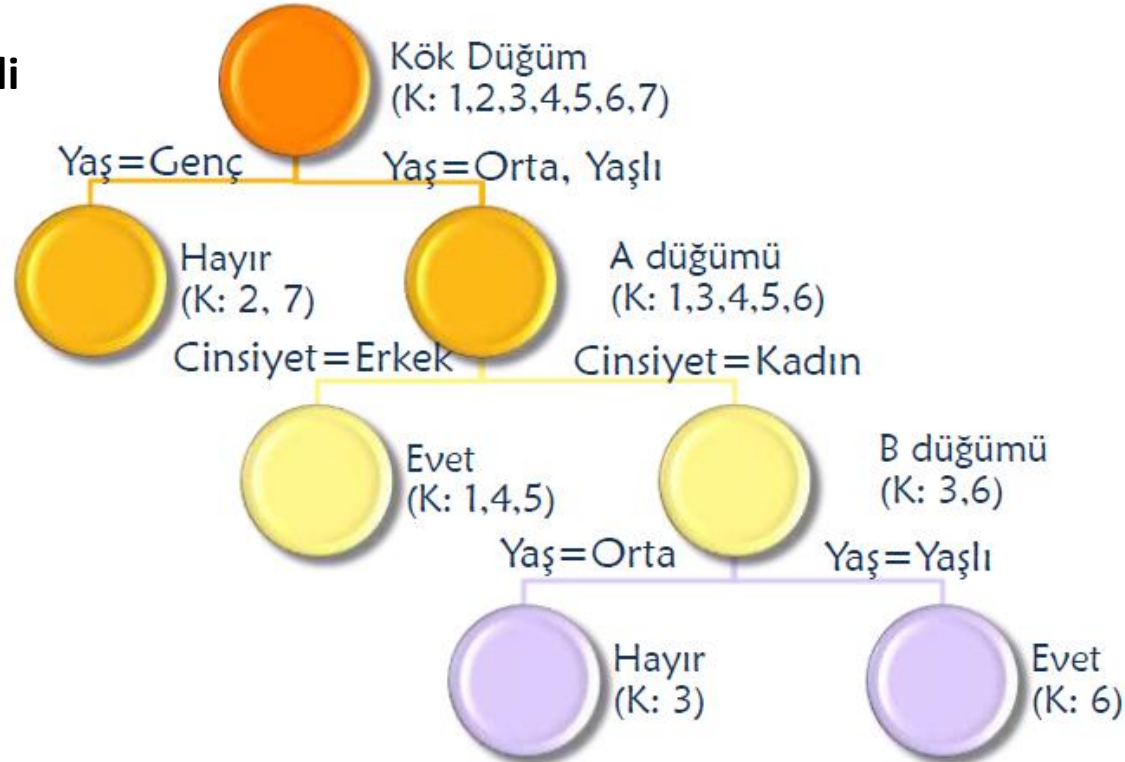
Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet

Gini algoritması-Örnek

- (1,4,5) satırlarını eğitim kümesinden çıkart

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
3	Yüksek	Orta	Kadın	Hayır
6	Yüksek	Yaşlı	Kadın	Evet

- Ağacın son hali



Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	Ortaokul	Yaşlı	Erkek	Evet
2	İlkokul	Genç	Erkek	Hayır
3	Yüksek	Orta	Kadın	Hayır
4	Ortaokul	Orta	Erkek	Evet
5	İlkokul	Orta	Erkek	Evet
6	Yüksek	Yaşlı	Kadın	Evet
7	İlkokul	Genç	Kadın	Hayır

Gini algoritması-Örnek

- Kural Tablosu

Kural 1:

Eğer Yaş=Genç ise Kabul=Hayır

Kural 2:

Eğer Yaş=Orta veya Yaşlı ise ve

Eğer Cinsiyet=Erkek ise Kabul=Evet

Kural 3:

Eğer Yaş=Orta veya Yaşlı ise ve

Eğer Cinsiyet=Kadın ise ve

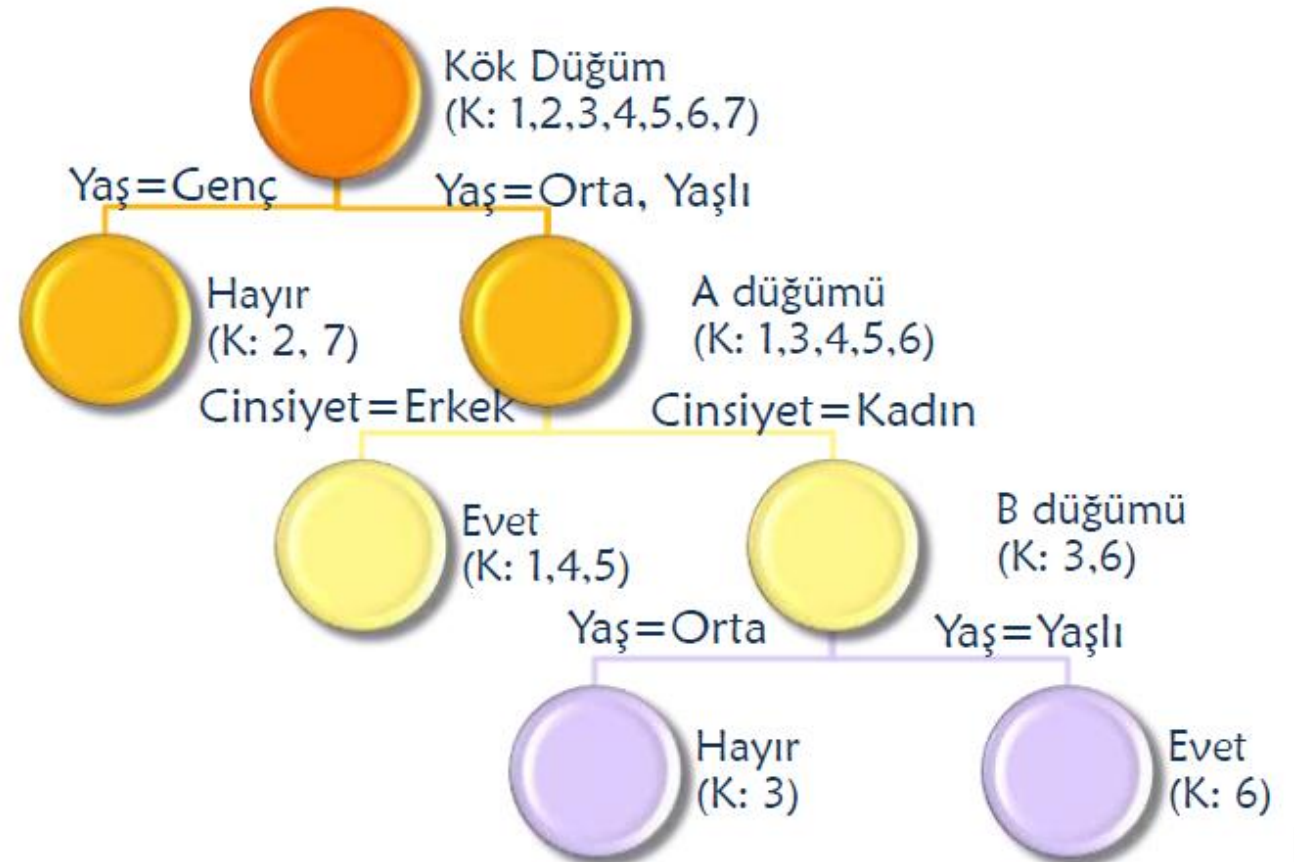
Eğer Yaş=Orta ise Kabul=Hayır

Kural 4:

Eğer Yaş=Orta veya Yaşlı ise ve

Eğer Cinsiyet=Kadın ise ve

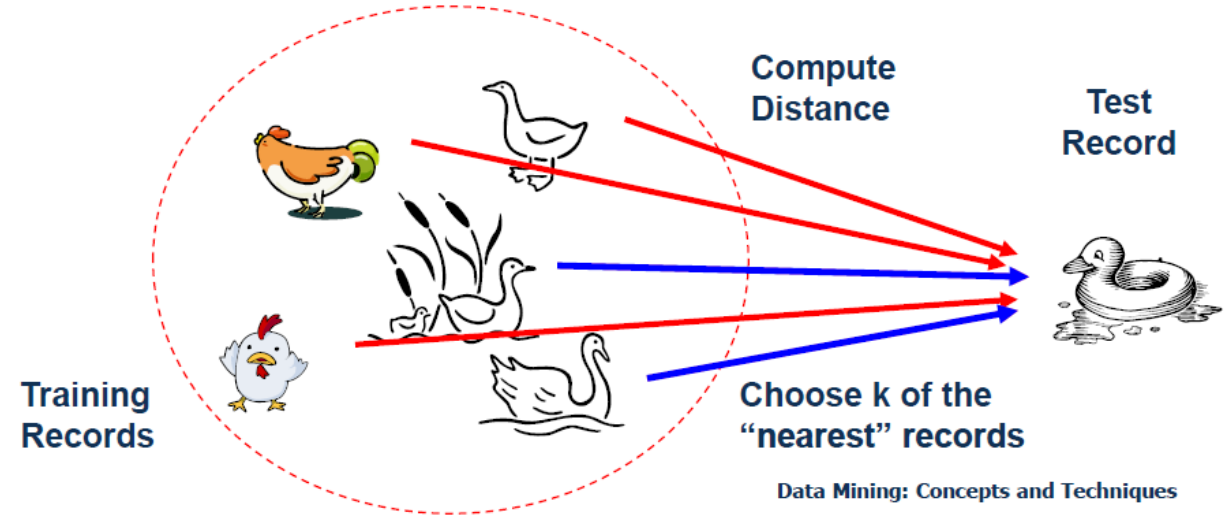
Eğer Yaş=Yaşlı ise Kabul=Evet



Mesafeye Dayalı Algoritmalar

K-en yakın komşu algoritması (kNN)

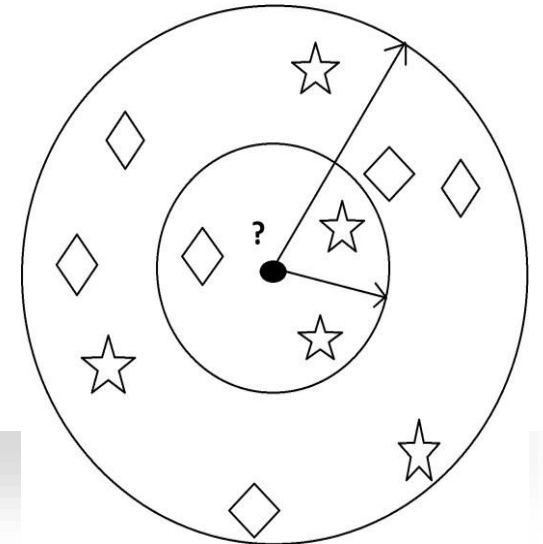
- Uygulama kolaylığı nedeni ile tercih edilen, yaygın kullanıma sahip bir algoritmadır.
- Sınıfları belli olan bir veri kümesine sınıfı bilinmeyen bir veri dahil olduğunda, bu verinin kendisine en yakın k adet sınıftan frekansı fazla olan sınıfa atanması prensibine dayanır.



Eğer ördek gibi yürüyor ve ördek sesi çıkarıyorsa muhtemelen ördektir.

K-en yakın komşu algoritması (kNN)

- Öncelikle k değeri belirlenir, k değeri en yakın komşu sayısını ifade eder.
- Sınıfı belirlenecek olan gözlem değeri ile sınıfı önceden bilinen tüm değerlerin mesafeleri hesaplanır.
- En küçük mesafeye sahip k adet veri belirlenir ve bunların çoğunluğunun dahil olduğu sınıf seçilir.
- Seçilen sınıf, gözlem değerinin ait olduğu sınıf olarak kabul edilir.
- kNN algoritmasında k değerinin belirlenmesi önemlidir; k çok büyük seçilirse özellikleri çok benzemeyen elemanlar aynı kümeye dahil edilebilir.
- Şekilde $k=3$ ve $k=11$ seçildiğinde farklı sınıf değerlerinin karşımıza çıktığı görülmektedir.



K-en yakın komşu algoritması (kNN)

X ve Y noktaları arasındaki Uzaklık hesabı

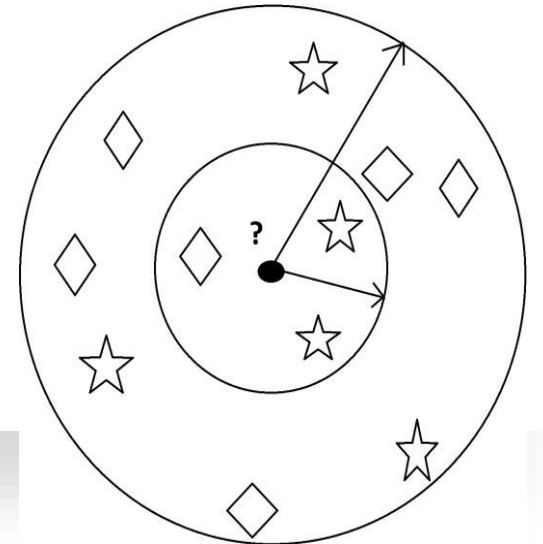
Öklid uzaklığı : $D(x,y) = \sqrt{\sum_{k=1}^p (X_k - Y_k)^2}$

Manhattan uzaklığı: $D(x,y) = \sum_{k=1}^p |X_k - Y_k|$

Minkowski uzaklığı: $D(x,y) = (\sum_{k=1}^p (|X_k - Y_k|)^q)^{1/q}$

Minkowski uzaklık ölçütünde $q=1$ ise Manhattan, $q=2$ ise Öklid uzaklığı elde edilir

Hamming uzaklığı: $D(x,y) = \sum_{k=1}^p \delta_k, \quad \delta_k = \begin{cases} 1, & X_k \neq Y_k \text{ ise} \\ 0, & X_k = Y_k \text{ ise} \end{cases}$



K-en yakın komşu algoritması (kNN) - Örnek

X ve Y noktaları arasındaki Uzaklık hesabı

Verilen gözlem tablosunda X ve Y niteliklerinden Z sınıfı oluşmaktadır.

Bu gözlem değerlerine bağlı olarak yeni bir gözlem olan $X=7$ ve $Y=3$ yani $(7,3)$ gözleminin hangi sınıfa dahil olduğunu $k=4$ için kNN algoritması ile bulalım.

X	Y	Z
1	3	Negatif
2	5	Pozitif
2	3	Pozitif
3	9	Negatif
4	7	Negatif
5	2	Pozitif
6	8	Pozitif
8	6	Negatif
10	6	Negatif
9	1	Negatif

K-en yakın komşu algoritması (kNN) - Örnek

(7,3) noktasının tüm gözlem değerleri ile uzaklıklarını hesaplayalım

$$d(1) = \sqrt{(1-7)^2 + (3-3)^2} = 6.00$$

$$d(2) = \sqrt{(2-7)^2 + (5-3)^2} = 5,39$$

$$d(3) = \sqrt{(2-7)^2 + (3-3)^2} = 5,00$$

⋮

X	Y	Uzaklık
1	3	6
2	5	5,39
2	3	5
3	9	7,21
4	7	5
5	2	2,24
6	8	5,10
8	6	3,16
10	6	4,24
9	1	2,83

K-en yakın komşu algoritması (kNN) - Örnek

En küçük $k=4$ tane uzaklığı belirleyelim

Belirlenen dört nokta **(7,3)** noktasına en yakın değerlerdir. Gözlem değerlerin içinde **bir pozitif** ve **üç negatif** değer olduğundan **(7,3)** noktasının sınıfı **negatif** olarak belirlenir.

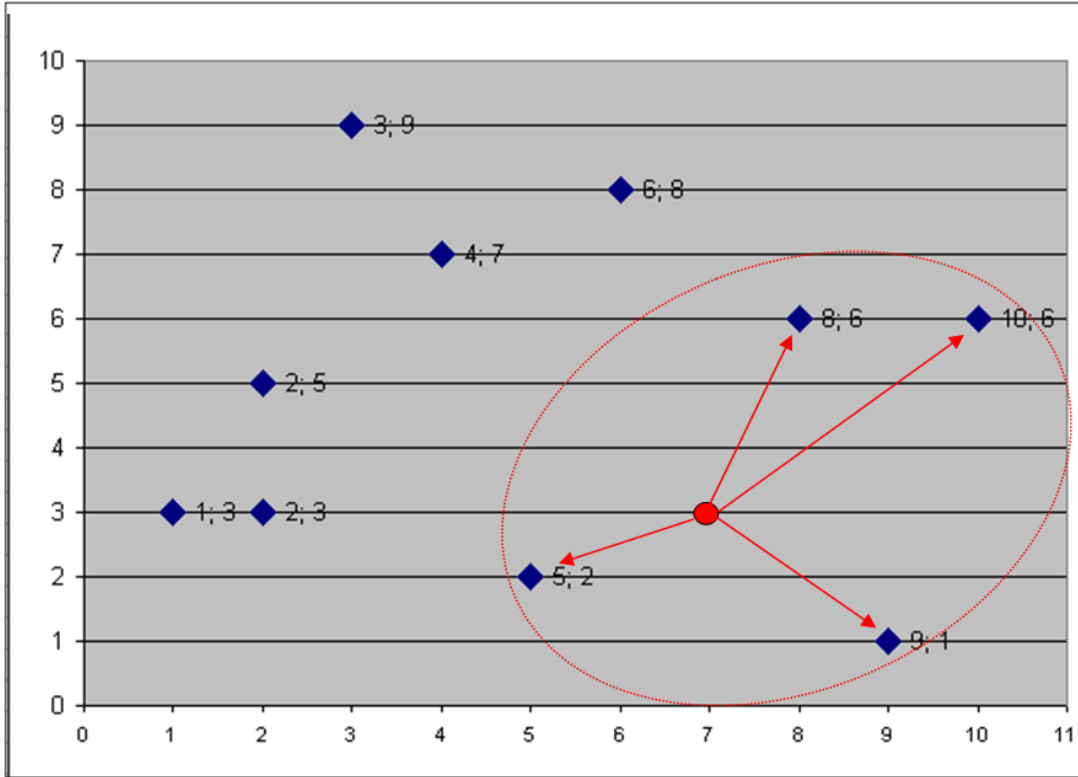
X	Y	Uzaklık	Sıralama	Z
1	3	6	9	Negatif
2	5	5,39	8	Pozitif
2	3	5	6	Pozitif
3	9	7,21	10	Negatif
4	7	5	5	Negatif
5	2	2,24	1	Pozitif
6	8	5,10	7	Pozitif
8	6	3,16	3	Negatif
10	6	4,24	4	Negatif
9	1	2,83	2	Negatif

K-en yakın komşu algoritması (kNN) - Örnek

En küçük $k=4$ tane uzaklığı belirleyelim

Belirlenen dört nokta **(7,3)** noktasına en yakın değerlerdir. Gözlem değerlerinin içinde **bir pozitif** ve **üç negatif** değer olduğundan **(7,3)** noktasının sınıfı **negatif** olarak belirlenir.

Grafik:



K-en yakın komşu algoritması (kNN) – Örnek 2

Verilen tabloya göre (7,8,5) noktasının hangi sınıf değerine sahip olduğunu belirleyelim. (k=3)
Min-max normalizasyonu ile verileri (0-1) aralığına dönüştürdükten sonra sınıflandırma yapalım.

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$$

x1	x2	x3	Y
10	5	19	Evet
8	2	4	Hayır
18	16	6	Hayır
12	15	8	Evet
3	15	15	Evet

K-en yakın komşu algoritması (kNN) – Örnek 2

Birinci satırdaki değerler için dönüşüm:

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

$$X^* = \frac{10-3}{18-3} = 0,47$$

$$X^* = \frac{5-2}{16-2} = 0,21$$

$$X^* = \frac{19-4}{19-4} = 1$$

	x1	x2	x3
$X_{\min}$	3	2	4
$X_{\max}$	18	16	19

x1	x2	x3	Y
10	5	19	Evet
8	2	4	Hayır
18	16	6	Hayır
12	15	8	Evet
3	15	15	Evet

Dönüştürülmüş veri ile uzaklık hesabı

(7,8,5) noktası da min max dönüşümü ile (0.26, 0.43, 0.07) noktasına dönüştürülür.

Bu noktaya en yakın 3 gözlemin sınıfı hayır, evet, evet olduğu için (7,8,5) gözleminin, yani dönüştürülmüş değerle ifade edilen (0.26, 0.43, 0.07) gözleminin sınıfı “evet” olarak kabul edilecektir.

x1	x2	x3	Y	x1	x2	x3	Uzaklık	Sıra
10	5	19	Evet	0,47	0,21	1,00	0,98	
8	2	4	Hayır	0,33	0,00	0,00	0,44	1
18	16	6	Hayır	1,00	1,00	0,13	0,93	
12	15	8	Evet	0,60	0,93	0,27	0,63	2
3	15	15	Evet	0,00	0,93	0,73	0,87	3

Ağırlıklı Oylama

- kNN algoritmasında, k adet en yakın komşu arasından en fazla tekrar eden sınıfın seçilmesi, sadece k adet komşunun dikkate alınması nedeni ile her zaman uygun olmayabilir. Bu nedenle son aşamada en fazla tekrar eden sınıfı seçme yerine **ağırlıklı oylama** adı verilen bir yöntem uygulanabilir.
- Ağırlıklı oylama yöntemi gözlem değerleri için aşağıdaki bağıntıya göre ağırlıklı uzaklıkların hesaplanması yöntemine dayanır. $d(i,j)$ Öklid uzaklığıdır.

$$d(i, j)' = \frac{1}{d(i, j)^2}$$

- Sınıf değerlerinin her biri için uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri bulunur.
- En büyük ağırlıklı oylama değerine sahip olan sınıf değeri yeni gözlem değerinin ait olduğu sınıf olarak belirlenir.

K-en yakın komşu algoritması (kNN) – Örnek 3

(0.10, 0.50) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu algoritmasından ve aşağıdaki tablodan yararlanarak bulunuz. (k=3)

$$d(i, j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

$$d(i, j)' = \frac{1}{d(i, j)^2}$$

x1	x2	Cins
0,08	0,2	Erkek
0,07	0,07	Erkek
0,2	0,09	Erkek
1	0,2	Kadın
0,05	0,06	Erkek
0,2	0,25	Erkek
0,17	0,07	Erkek
0,15	0,55	Kadın
0,5	0,08	Erkek
0,1	0,06	Kadın

K-en yakın komşu algoritması (kNN) – Örnek 3

Öklid uzaklık formülü ile (0.10, 0.50) gözleminin diğer gözlemlere uzaklıklarının ve ağırlıklı uzaklıkların hesaplanması

$$d(i, j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

$$d(i, j)' = \frac{1}{d(i, j)^2}$$

x1	x2	Cins	Uzaklık	Sıra	Ağırlıklı uzaklık
0,08	0,2	Erkek	0,30	3	11,046
0,07	0,07	Erkek	0,43		
0,2	0,09	Erkek	0,42		
1	0,2	Kadın	0,95		
0,05	0,06	Erkek	0,44		
0,2	0,25	Erkek	0,27	2	13,79
0,17	0,07	Erkek	0,44		
0,15	0,55	Kadın	0,07	1	200,00
0,5	0,08	Erkek	0,58		
0,1	0,06	Kadın	0,44		

K-en yakın komşu algoritması (kNN) – Örnek 3

Ağırlıklı uzaklıkların yorumlanması

Tabloda en yakın 3 komşu için 2 erkek, 1 kadın sınıfı vardır.

Erkek değerleri için ağırlıklı ortalama = $11,046 + 13,79 = 24,836$

Kadın değeri için ağırlıklı ortalama = 200

Kadın değeri için elde edilen ağırlıklı ortalama daha büyük olduğu için yeni gözlem değerinin sınıfı “kadın” olacaktır.

Ağırlıklı ortalama kullanılmasaydı yeni gözlemin sınıfı “erkek” olacaktı.

x1	x2	Cins	Uzaklık	Sıra	Ağırlıklı uzaklık
0,08	0,2	Erkek	0,30	3	11,046
0,07	0,07	Erkek	0,43		
0,2	0,09	Erkek	0,42		
1	0,2	Kadın	0,95		
0,05	0,06	Erkek	0,44		
0,2	0,25	Erkek	0,27	2	13,79
0,17	0,07	Erkek	0,44		
0,15	0,55	Kadın	0,07	1	200,00
0,5	0,08	Erkek	0,58		
0,1	0,06	Kadın	0,44		

Referanslar

- ***Veri madenciliği ders notları, Dr. Burcu Çarklı YAVUZ, SAÜ (Temel İçerik)***
- Data Mining: Concepts and Techniques, J. Han, M. Kamber, J. Pei
- Veri madenciliği Kavram ve Algoritmaları, Gökhan Silahtaroglu, Papatya Yayıncılık
- Veri Madenciliği Yöntemleri, Yalçın Özkan, Papatya Yayıncılık
- Veri madenciliği ders notları, Doç. Dr. Nilüfer YURTAY, SAÜ
- Veri madenciliği ders notları, Dr. Tuğrul TAŞCI, SAÜ
- <https://mathanrajsharma.medium.com/fundamentals-of-classification-and-regression-trees-cart-e9af0b152503>

