VERİ MADENCİLİĞİ

(Karar Ağaçları ile Sınıflandırma)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

İçerik

- Sınıflandırma yöntemleri
  - Karar ağaçları ile sınıflandırma
    - Entropi Kavramı
    - ID3 Algoritması
    - C4.5 Algoritması

    } Entropiye dayalı algoritmalar

    - Twoing Algoritması
    - Gini Algoritması

    } Sınıflandırma ve regresyon ağaçları (CART)

    - k-en yakın komşu algoritması } Bellek tabanlı algoritmalar

Sınıflandırma ve Regresyon Ağaçları (CART)

- Sınıflandırma ve regresyon ağaçları veri madenciliğinin sınıflandırma ile ilgili konuları arasında yer alır. Bu yöntem 1984'te Breiman tarafından ortaya atılmıştır. CART karar ağacı, herbir karar düğümünden itibaren ağacın iki dala ayrılması ilkesine dayanır. Yani bu tür karar ağaçlarında ikili dallanmalar söz konusudur.
- CART algoritmasında bir düğümde belirli bir kriter uygulanarak bölünme işlemi gerçekleştirilir. Bunun için önce tüm niteliklerin var olduğu değerler gözönüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilir. Bu bölünmeler üzerinde seçme işlemi uygulanır. Bu kapsamdaki iki algoritma bulunmaktadır.
  - Twoing Algoritması
  - Gini Algoritması

Twoing Algoritması

- Twoing algoritmasında eğitim kümesi her adımda iki parçaya ayrılarak bölümleme yapılır.
- Aday bölünmelerin sağ ve sol kısımlarının her birisi için nitelik değerinin ilgili sütundaki tekrar sayısı alınır.
- Aday bölünmelerin sağ ve sol kısımlarındaki her bir nitelik değeri için sınıf değerlerinin olma olasılığı hesaplanır.
- Her bölünme için uygunluk değeri en yüksek olan alınır.

$$\Phi(B \mid d) = 2 \frac{|B_{sol}|}{|T|} \frac{|B_{sag}|}{|T|} \sum_{j=1}^{n} abs\left(\frac{|\text{Tsinif}_j|}{|B_{sol}|} - \frac{|\text{Tsinif}_j|}{|B_{sag}|} \right)$$

- Burada, T eğitim kümesindeki kayıt sayısını, B aday bölünmeyi, d düğümü, Tsinif_j ise j.sınıf değerini gösterir.

Örnek (1/8)

- Tabloda çalışanların maaş, deneyim, görev niteliklerine göre hedef niteliği olan memnun olma durumlarına ait 11 gözlem verilmiştir. Twoing algoritmasını kullanarak sınıflandırma yapınız.

PERSONEL	MAAŞ	DENEYİM	GÖREV	MEMNUN
1	NORMAL	ORTA	UZMAN	EVET
2	YÜKSEK	YOK	UZMAN	EVET
3	DÜŞÜK	YOK	YÖNETİCİ	EVET
4	YÜKSEK	ORTA	YÖNETİCİ	EVET
5	DÜŞÜK	ORTA	YÖNETİCİ	EVET
6	YÜKSEK	İYİ	YÖNETİCİ	EVET
7	DÜŞÜK	İYİ	YÖNETİCİ	EVET
8	YÜKSEK	ORTA	UZMAN	HAYIR
9	DÜŞÜK	ORTA	UZMAN	HAYIR
10	YÜKSEK	İYİ	UZMAN	HAYIR
11	DÜŞÜK	İYİ	UZMAN	HAYIR

Örnek (2/8)

- Aday bölünmeler aşağıdaki gibidir.

BÖLÜNME	SOL	SAĞ
1	MAAŞ = NORMAL	MAAŞ = {DÜŞÜK, YÜKSEK}
2	MAAŞ = YÜKSEK	MAAŞ = {DÜŞÜK, NORMAL}
3	MAAŞ = DÜŞÜK	MAAŞ = {NORMAL, YÜKSEK}
4	DENEYİM = YOK	DENEYİM = {ORTA, İYİ}
5	DENEYİM = ORTA	DENEYİM = {YOK, İYİ}
6	DENEYİM = İYİ	DENEYİM = {YOK, ORTA}
7	GÖREV = UZMAN	GÖREV = YÖNETİCİ
8	GÖREV = YÖNETİCİ	GÖRE = UZMAN

Örnek (3/8)

- MAAŞ = NORMAL için

$$P_{sol} = \frac{|B_{sol}|}{|T|} = \frac{1}{11} = 0{,}09 \qquad P_{(EVET/t_{sol})} = \frac{|\text{Tsinif}_{EVET}|}{|B_{sol}|} = \frac{1}{1} = 1 \qquad P_{(HAYIR/t_{sol})} = \frac{|\text{Tsinif}_{HAYIR}|}{|B_{sol}|} = \frac{0}{1} = 0$$

BÖLÜNME	$\lvert B_{sol} \rvert$	P_{Sol}	$\lvert sinif_{EVET} \rvert$	$\lvert sinif_{HAYIR} \rvert$	$P(EVET \lvert t_{Sol})$	$P(HAYIR \lvert t_{Sol})$
1	1	0,09	1	0	1	0
2	5	0,45	3	2	0,6	0,4
3	5	0,45	3	2	0,6	0,4
4	2	0,18	2	0	1	0
5	5	0,45	3	2	0,6	0,4
6	4	0,36	2	2	0,5	0,5
7	6	0,55	2	4	0,33	0,67
8	5	0,45	5	0	1	0

Örnek (4/8)

- MAAŞ = {DÜŞÜK, YÜKSEK} için

$$P_{sag} = \frac{|B_{sag}|}{|T|} = \frac{10}{11} = 0{,}91$$

$$P_{(EVET|t_{sag})} = \frac{|\text{Tsinif}_{EVET}|}{|B_{sag}|} = \frac{6}{10} = 0{,}6$$

$$P_{(HAYIR|t_{sag})} = \frac{|\text{Tsinif}_{HAYIR}|}{|B_{sag}|} = \frac{4}{10} = 0{,}4$$

BÖLÜNME	$\|B_{sag}\|$	P_{sag}	$\|sinif_{EVET}\|$	$\|sinif_{HAYIR}\|$	$P(EVET\|t_{Sag})$	$P(HAYIR\|t_{Sag})$
1	10	0,91	6	4	0,6	0,4
2	6	0,55	4	2	0,67	0,33
3	6	0,55	4	2	0,67	0,33
4	9	0,82	5	4	0,56	0,44
5	6	0,55	4	2	0,67	0,33
6	7	0,64	5	2	0,71	0,29
7	5	0,45	5	0	1	0
8	6	0,55	2	4	0,33	0,67

Örnek (5/8)

Uygunluk değeri (1. aday bölünme için)

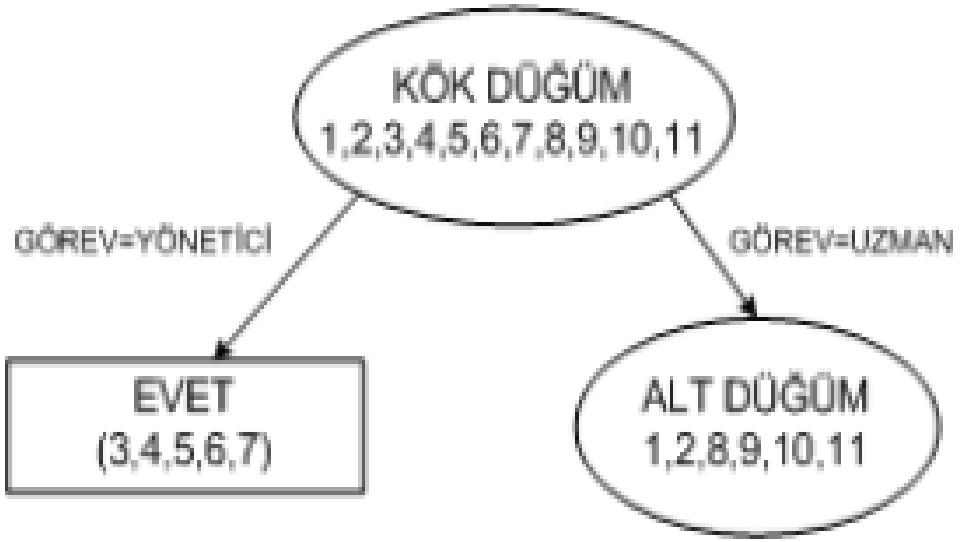
$$\Phi(1|d) = 2\frac{|B_{sol}|}{|T|}\frac{|B_{sag}|}{|T|}\sum_{j=1}^{n} abs\left(\frac{|\text{Tsinif}_j|}{|B_{sol}|} - \frac{|\text{Tsinif}_j|}{|B_{sag}|}\right)$$

$$= 2(0{,}09)(0{,}91)[|1-0{,}6|+|0-0{,}4|] = 0{,}13$$

BÖLÜNME	P_{Sol}	$P_{Sağ}$	$2P_{Sol}P_{Sağ}$	$\Phi(B\|d)$
1	0,09	0,91	0,17	0,13
2	0,45	0,55	0,5	0,07
3	0,45	0,55	0,5	0,07
4	0,18	0,82	0,3	0,26
5	0,45	0,55	0,5	0,07
6	0,36	0,64	0,46	0,2
7	0,55	0,45	0,5	0,66
8	0,45	0,55	0,5	0,66

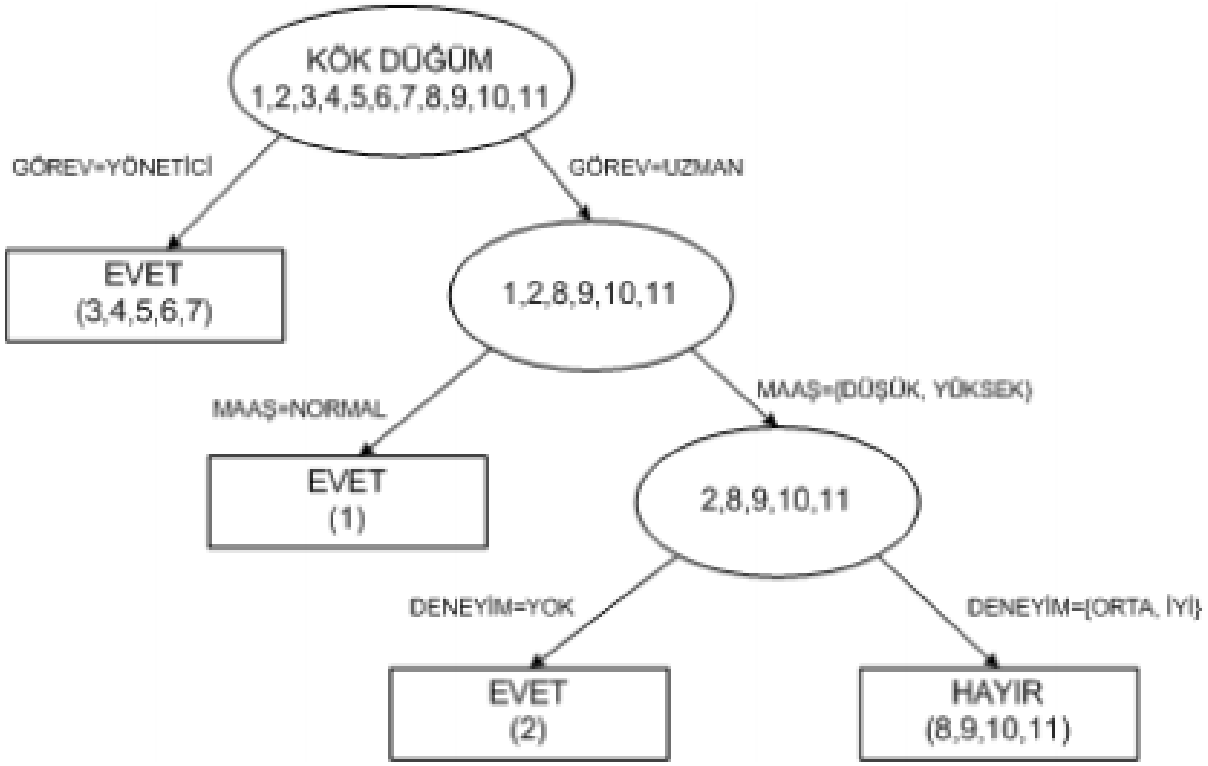
Örnek (6/8)

- Aynı işlemler ALT DÜĞÜM için tekrarlanır.



Örnek (7/8)

- Sonuç karar ağacı.



Örnek (8/8)

- Karar ağacından elde edilen kurallar
  - 1. EĞER (GÖREV = YÖNETİCİ) İSE (MEMNUN = EVET)
  - 2. EĞER (GÖREV = UZMAN) VE (MAAŞ = NORMAL) İSE (MEMNUN =EVET)
  - 3. EĞER (GÖREV = UZMAN) VE (MAAŞ = DÜŞÜK VEYA MAAŞ = YÜKSEK) VE (DENEYİM=YOK) İSE (MEMNUN = EVET)
  - 4. EĞER (GÖREV = UZMAN) VE (MAAŞ = DÜŞÜK VEYA MAAŞ = YÜKSEK) VE (DENEYİM = ORTA VEYA DENEYİM = İYİ) İSE (MEMNUN = HAYIR)

Gini Algoritması

- **Gini algoritmasında nitelik değerleri iki parçaya ayrılarak bölümleme yapılır.**
- Her bölünme için $Gini_{sol}$ ve $Gini_{sağ}$ değerleri hesaplanır.

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left(\frac{|\text{Tsinif}_i|}{|B_{sol}|} \right)^2 \qquad Gini_{sag} = 1 - \sum_{i=1}^{k} \left(\frac{|\text{Tsinif}_i|}{|B_{sag}|} \right)^2$$

- Burada, Tsinif_i soldaki bölümdeki her bir sınıf değerini, Tsinif_i sağdaki bölümdeki her bir sınıf değerini, $|B_{sol}|$ sol bölümdeki tüm değer sayısını, $|B_{sağ}|$ sağ bölümdeki tüm değer sayısını gösterir.

$$Gini_j = \frac{1}{n} \left(|B_{sol}| Gini_{sol} + |B_{sag}| Gini_{sag} \right)$$

- Her bölümlemeden sonra **Gini değeri en küçük olan seçilir.**

Örnek (1/8)

SIRA	EĞİTİM	YAŞ	CİNSİYET	SONUÇ
1	ORTA	YAŞLI	ERKEK	EVET
2	İLK	GENÇ	ERKEK	HAYIR
3	YÜKSEK	ORTA	KADIN	HAYIR
4	ORTA	ORTA	ERKEK	EVET
5	İLK	ORTA	ERKEK	EVET
6	YÜKSEK	YAŞLI	KADIN	EVET
7	İLK	GENÇ	KADIN	HAYIR
8	ORTA	ORTA	ERKEK	EVET

	EĞİTİM		YAŞ		CİNSİYET	
SONUÇ	İLK	ORTA,YÜKSEK	GENÇ	ORTA,YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

Örnek (2/8)

SONUÇ	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA,YÜKSEK	GENÇ	ORTA,YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

EĞİTİM için

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0{,}444$$

$$Gini_{sag} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0{,}320$$

Örnek (3/8)

SONUÇ	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA,YÜKSEK	GENÇ	ORTA,YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

YAŞ için

$$Gini_{sol} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$Gini_{sag} = 1 - \left[\left(\frac{5}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right] = 0{,}278$$

Örnek (4/8)

	EĞİTİM		YAŞ		CİNSİYET	
SONUÇ	İLK	ORTA,YÜKSEK	GENÇ	ORTA,YAŞLI	KADIN	ERKEK
EVET	1	4	0	5	1	4
HAYIR	2	1	2	1	2	1

CİNSİYET için

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0{,}444$$

$$Gini_{sag} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right] = 0{,}320$$

Örnek (5/8)

Gini değerleri

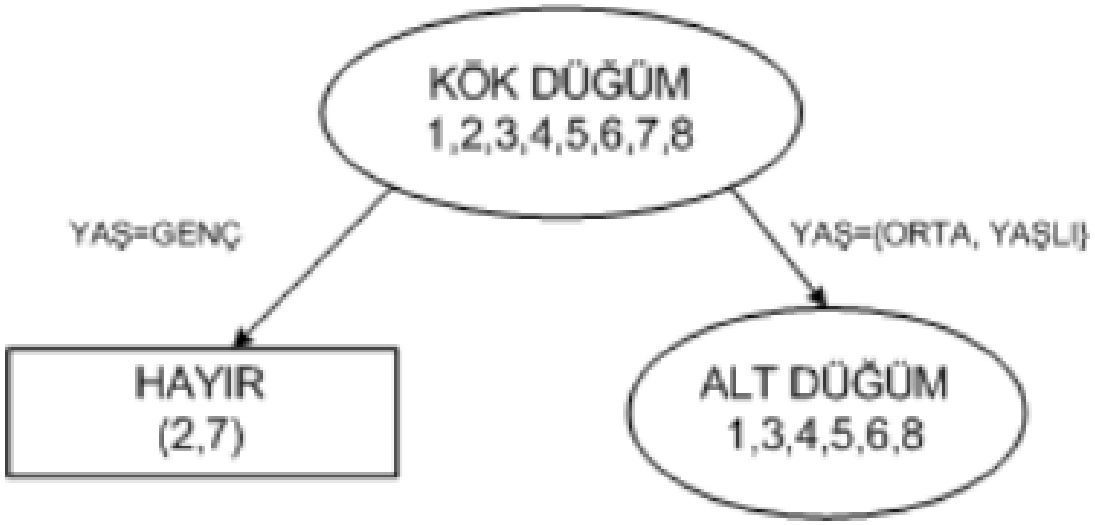
$$Gini_{EGITIM} = \frac{3(0,444) + 5(0,320)}{8} = 0,367$$

$$Gini_{YAS} = \frac{2(0) + 6(0,278)}{8} = 0,209$$

$$Gini_{CINSIYET} = \frac{3(0,444) + 5(0,320)}{8} = 0,367$$

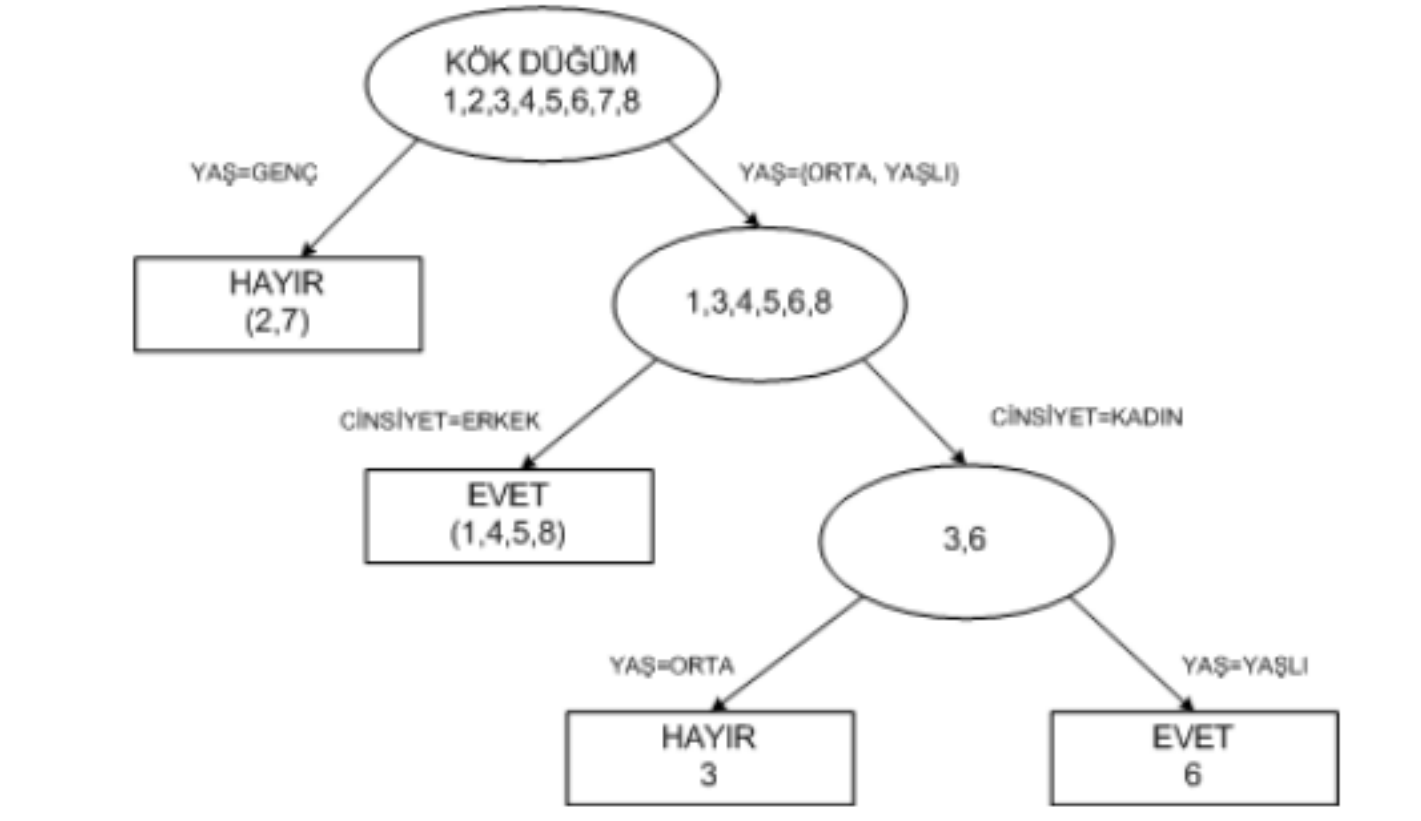
İlk bölünme YAŞ niteliğine göre yapılacaktır.

Örnek (6/8)



Aynı işlemler ALT DÜĞÜM için tekrarlanır.

Örnek (7/8)


KÖK DÜĞÜM
1,2,3,4,5,6,7,8
YAŞ=GENÇ
YAŞ={ORTA, YAŞLI}
HAYIR
(2,7)
1,3,4,5,6,8
CİNSİYET=ERKEK
CİNSİYET=KADIN
EVET
(1,4,5,8)
3,6
YAŞ=ORTA
YAŞ=YAŞLI
HAYIR
3
EVET
6

Örnek (8/8)

- Karar ağacından elde edilen kurallar
  - 1. EĞER (YAŞ = GENÇ) İSE (SONUÇ = HAYIR)
  - 2. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = ERKEK) İSE (SONUÇ = EVET)
  - 3. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = KADIN) VE (YAŞ = YAŞLI) İSE (SONUÇ = EVET)
  - 4. EĞER (YAŞ = ORTA VEYA YAŞ = YAŞLI) VE (CİNSİYET = KADIN) VE (YAŞ = ORTA) İSE (SONUÇ = HAYIR)

Bellek Tabanlı Algoritmalar

- K-en yakın komşu algoritması (K-nearest neighbor algorithm).

K-en yakın komşu algoritması

- Sınıflandırma yöntemlerinden birisi de K-en yakın komşu algoritmasıdır.
- Bu yöntem sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacıyla kullanılır.
- Bu yöntem örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının ve en küçük uzaklığa sahip k sayıda gözlemin seçilmesi esasına dayanmaktadır. Uzaklıkların hesaplanmasında i ve j noktaları için örneğin Öklid uzaklık formülü kullanılabilir. (Diğer uzaklıklar veri önişleme kısmında açıklanmıştı)

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

K-en yakın komşu algoritması

- K-en yakın komşu algoritması, gözlem değerlerinden oluşan bir küme için aşağıdaki adımları içerir.
  - a) K parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır.
  - b) Bu algoritma verilen bir noktaya en yakın komşuları belirleyeceği için söz konusu nokta ile diğer tüm noktalar arasındaki uzaklıklar tek tek hesaplanır.
  - c) Yukarıda hesaplanan uzaklıklara göre satırlar sıralanır ve bunlar arasından en küçük olan k tanesi seçilir.
  - d) Seçilen satırların hangi kategoriye ait oldukları belirlenir ve en çok tekrarlanan kategori değeri seçilir.
  - e) Seçilen kategori, tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir.

Örnek 1.

- Aşağıda verilen gözlem tablosu X1 ve X2 nitelikleri ve Y sınıfından oluşmaktadır. Bu gözlem değerine bağlı olarak yeni bir gözlem değeri olan X1=8, X2=4 değerlerinin yani (8,4) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu algoritması ile bulunuz.

X1	X2	Y
2	4	KÖTÜ
3	6	İYİ
3	4	İYİ
4	10	KÖTÜ
5	8	KÖTÜ
6	3	İYİ
7	9	İYİ
9	7	KÖTÜ
11	7	KÖTÜ
10	2	KÖTÜ

Örnek 1.

- a) **K'nın belirlenmesi:** k=4 kabul edilir.
- b) **Uzaklıkların hesaplanması:** (8,4) noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığına göre hesaplanır.

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

Biçiminde birinci gözlem olan (2,4) noktası ile (8,4) noktası arasındaki uzaklık,

$$d(i,j) = \sqrt{(2-8)^2 + (4-4)^2} = 6.00$$

Benzer şekilde uzaklıklar hesaplandığında tablodaki sonuç ortaya çıkacaktır.

Örnek 1.

- (8,4) noktasının gözlem değerlerine olan uzaklıkları,

X1	X2	Uzaklık
2	4	6
3	6	5,39
3	4	5
4	10	7,21
5	8	5
6	3	2,24
7	9	5,1
9	7	3,16
11	7	4,24
10	2	2,83

- c) **En küçük uzaklıkların belirlenmesi:** Satırlar sıralanarak en küçük k=4 tanesi belirlenir. Bu dört nokta verilen (8,4) noktasına en yakın gözlem değerleridir.

X1	X2	Uzaklık	Sıra
2	4	6	9
3	6	5,39	8
3	4	5	6
4	10	7,21	10
5	8	5	5
6	3	2,24	1
7	9	5,1	7
9	7	3,16	3
11	7	4,24	4
10	2	2,83	2

Örnek 1.

- d) **Seçilen satırların ilişkin sınıfların belirlenmesi:** (8,4) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu dört gözlem içinde bir tane **İYİ** 3 tane **KÖTÜ** sınıfı vardır.

X1	X2	Uzaklık	Sıra	k komşunun Y değeri
2	4	6	9	
3	6	5,39	8	
3	4	5	6	
4	10	7,21	10	
5	8	5	5	
6	3	2,24	1	İYİ
7	9	5,1	7	
9	7	3,16	3	KÖTÜ
11	7	4,24	4	KÖTÜ
10	2	2,83	2	KÖTÜ

- e) **Yeni gözlemin sınıfı: KÖTÜ** değerlerinin sayısı **İYİ** değerlerinin sayısından fazla olduğu için (8,4) noktasının sınıfı **KÖTÜ** olarak belirlenir.

Örnek 2.

- Aşağıda verilen gözlem tablosunda Y sınıf niteliğini ifade etmektedir. Bu verilere dayanarak (7,8,5) noktasının hangi sınıf değerine sahip olduğunu belirleyelim. Gözlemlerin gerçek değerleri değil normalize edilmiş değerleri kullanılacaktır. Gözlem değerlerini (0,1) aralığına çekmek için min-max normalleştirmesi kullanılacaktır.

X1	X2	X3	Y
10	5	19	EVET
8	2	4	HAYIR
18	16	6	HAYIR
12	15	8	EVET
3	15	15	EVET

Örnek 2.

- Min-max normalleştirmesi sonucu dönüştürülen değerler aşağıdadır.
- $X^* = \frac{X - X_{min}}{X_{max} - X_{min}}$ (min-max normalizasyonu

X1	X2	X3	Y
0,47	0,21	1	EVET
0,33	0	0	HAYIR
1	1	0,13	HAYIR
0,6	0,93	0,27	EVET
0	0,93	0,73	EVET

- Aday noktanın normalizasyon değeri (0.27,0.43, 0.07)

Örnek 2.

- a) **K'nın belirlenmesi:** k=3 kabul edilir.
- b) **Uzaklıkların hesaplanması:** (0,27, 0,43, 0,07) noktası ile gözlem değerlerinin her biri arasındaki uzaklıklar Öklid uzaklığına göre hesaplanır.

$$d(i,j) = \sqrt{(0{,}47 - 0{,}27)^2 + (0{,}21 - 0{,}43)^2 + (1 - 0{,}07)^2} = 0{,}98$$

X1	X2	X3	Uzaklık
0,47	0,21	1	0,98
0,33	0	0	0,44
1	1	0,13	0,93
0,6	0,93	0,27	0,63
0	0,93	0,73	0,87

Örnek 2.

- c) **En küçük uzaklıkların belirlenmesi:** Satırlar sıralanarak en küçük k=3 tanesi belirlenir.

X1	X2	X3	Uzaklık	Sıra
0,47	0,21	1	0,98	5
0,33	0	0	0,44	1
1	1	0,13	0,93	4
0,6	0,93	0,27	0,63	2
0	0,93	0,73	0,87	3

Örnek 2.

- d) **Seçilen satırların ilişkin sınıfların belirlenmesi:** (0,27, 0,43, 0,07) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içinde hangi değerin baskın olduğu araştırılır. Bu üç gözlem içinde bir tane **HAYIR** 2 tane **EVET** sınıfı vardır.

X1	X2	X3	Uzaklık	Sıra	k komşunun Y değeri
0,47	0,21	1	0,98	5	
0,33	0	0	0,44	1	HAYIR
1	1	0,13	0,93	4	
0,6	0,93	0,27	0,63	2	EVET
0	0,93	0,73	0,87	3	EVET

- e) **Yeni gözlemin sınıfı: EVET** değerlerinin sayısı **HAYIR** değerlerinin sayısından fazla olduğu için (7,8,5) gözleminin sınıfı **EVET** olarak kabul edilir.

Ağırlıklı Oylama

- K-en yakın komşu algoritması sınıfı bilinmeyen gözlem değeri için k gözlem içindeki en fazla tekrar eden sınıfın seçilmesi esasına dayanmaktadır. Ancak seçilen bu sınıf sadece k komşunun göz önüne alınması nedeniyle her zaman uygun olmayabilir. Bu son aşamada k komşu arasında en çok tekrarlanan sınıfı seçme yöntemi yerine **ağırlıklı oylama** (weighted voting) denilen bir yöntem uygulanabilir.
- Söz konusu ağırlıklı oylama yöntemi gözlem değerleri için aşağıdaki bağıntıya göre ağırlıklı uzaklıkların hesaplanmasına dayanır.

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

- *d(i,j)* ifadesi *i* ve *j* gözlemleri arasındaki Öklid uzaklığıdır. Her bir sınıf değeri için bu uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri elde edilir. En büyük ağırlıklı oylama değerine sahip olan sınıf değeri yeni gözlemin ait olduğu sınıf olarak kabul edilir.

Örnek 2. Ağırlıklı Oylama Sonucu

- Ağırlıklı Oylama sonucunda da Örnek 2.'deki değerin sınıfının EVET olduğu görülür.

X1	X2	X3	Uzaklık	Sıra	k komşunun Y değeri	Ağırlıklı Oylama
0,47	0,21	1	0,98	5		
0,33	0	0	0,44	1	HAYIR	5,17
1	1	0,13	0,93	4		
0,6	0,93	0,27	0,63	2	EVET	2,52
0	0,93	0,73	0,87	3	EVET	3,84

(Evet)Toplam=2,52+3,84=6,66