

# BELLEK TABANLI SINIFLANDIRMA




# En Yakın k-komşu Algoritması

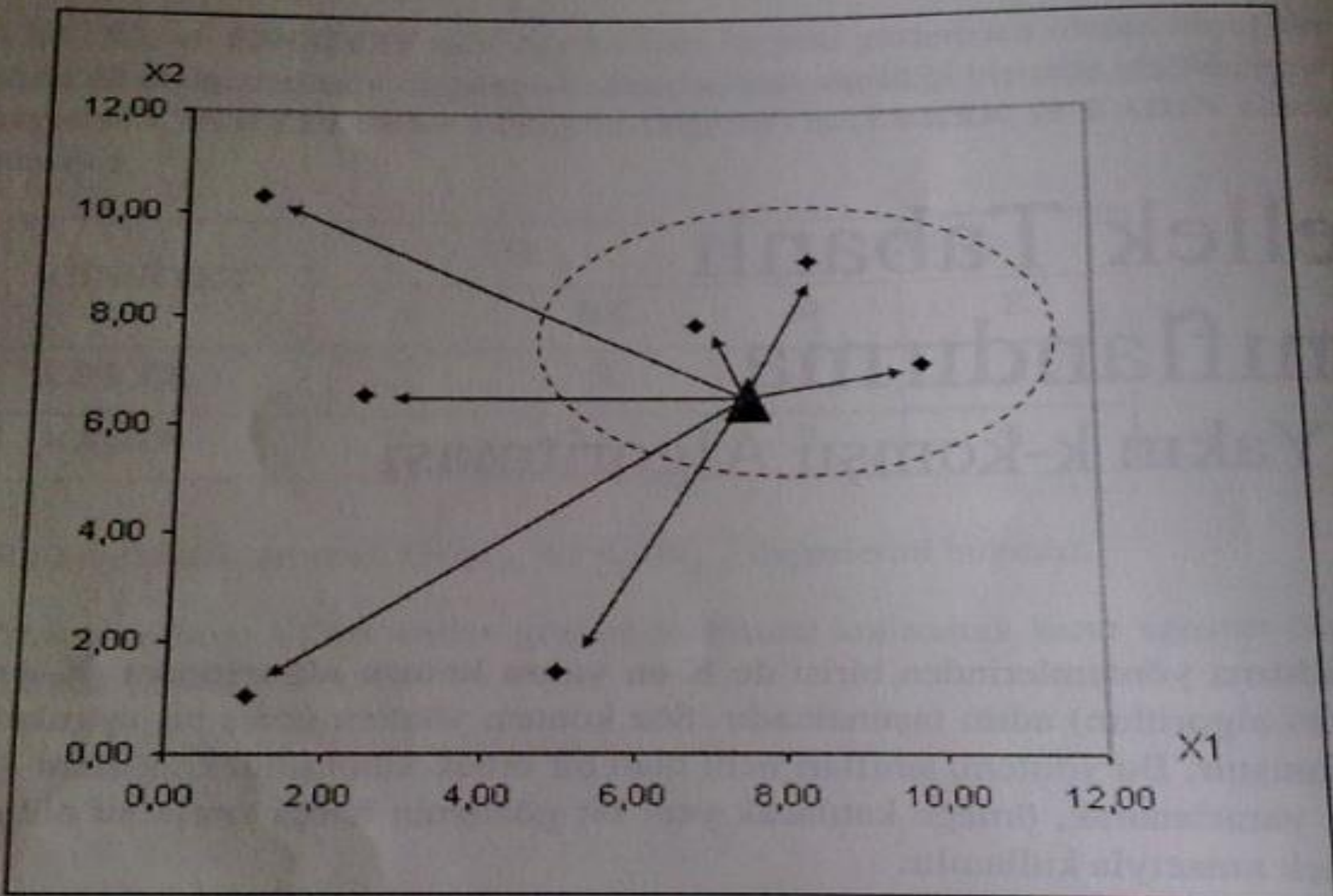
Bu yöntem, sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak, örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacıyla kullanılır.


Bu yöntem örnek kümedeki gözlemlerin her birinin , sonradan belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en küçük uzaklığa sahip sayıda gözlemin seçilmesi esasına dayanmaktadır.

Uzaklıkların hesaplanmasında ve j noktaları için aşağıdaki Öklid uzaklık formülü kullanılabilir.

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

Yukarıdaki denklemde görüldüğü gibi, gözlem değerleri arasında bir  noktasına en yakın k=3 komşu belirlenmiştir.



Şekil-5.1. Verilen bir  noktasına en yakın  $k=3$  komşunun belirlenmesi

# En yakın k- komşu Algoritması



En yakın komşu algoritması, gözlem değerlerinden oluşan bir küme için aşağıdaki işlemler yapılır.

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



## Uygulama 1

Aşağıda verilen gözlem tablosunu göz önüne alalım. Bu gözlemler **X1** ve **X2** niteliklerinden ve **Y** sınıfından oluşmaktadır. Bu gözlem değerlerine bağlı olarak, yeni bir gözlem olan  $X1=8$ ,  $X2=4$  değerinin yani (8,4) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu yöntemiyle bulalım.

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

Tablo1. Gözlem değerleri

**a) K nın belirlenmesi:** Algoritmaya başlamadan önce, k-en yakın algoritması için  $k=4$  olduğunu kabul ediyoruz. Böylece bu problem çerçevesinde verilen  $(8,4)$  noktasına en yakın 4 komşuyu arayacağımızı belirttik.

**b) Uzaklıkların hesaplanması:**  $(8,4)$  noktası ile gözlem değerinin her birisi arasındaki uzaklıkları hesaplamamız gerekiyor. Uzaklık bağıntısı olarak Öklid uzaklık formülünü kullanıyoruz. Öklid bağıntısı,

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

biçiminde olduğuna göre birinci gözlem olan  $(2,4)$  noktası ile  $(8,4)$  noktası arasındaki uzaklık şu şekilde hesaplanır :

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

Benzer biçimde diğer gözlemlerin her birinin (8,4) noktasına olan uzaklıkları tek tek hesaplanır.

- ✓  $D(i,j)=\sqrt{(3-8)^2+(6-4)^2}=5,39$
- ✓  $D(i,j)=\sqrt{(3-8)^2+(4-4)^2}=5.00$
- ✓  $D(i,j)=\sqrt{(4-8)^2+(10-4)^2}=7.21$
- ✓  $D(i,j)=\sqrt{(5-8)^2+(8-4)^2}=5.00$
- ✓  $D(i,j)=\sqrt{(6-8)^2+(3-4)^2}=2.24$
- ✓  $D(i,j)=\sqrt{(7-8)^2+(9-4)^2}=5.10$
- ✓  $D(i,j)=\sqrt{(9-8)^2+(7-4)^2}=3.16$
- ✓  $D(i,j)=\sqrt{(11-8)^2+(7-4)^2}=4.24$
- ✓  $D(i,j)=\sqrt{(10-8)^2+(2-4)^2}=2.83$

Hesaplanan değerler tablo üzerine yerleştirilir.



X1	X2	Uzaklık
2	4	6.00
3	6	5.39
3	4	5.00
4	10	7.21
5	8	5.00
6	3	2.24
7	9	5.10
9	7	3.16
11	7	4.24
10	2	2.83

Tablo2. Gözlem değerlerinin verilen bir(8,4) noktasına olan uzaklıkları

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

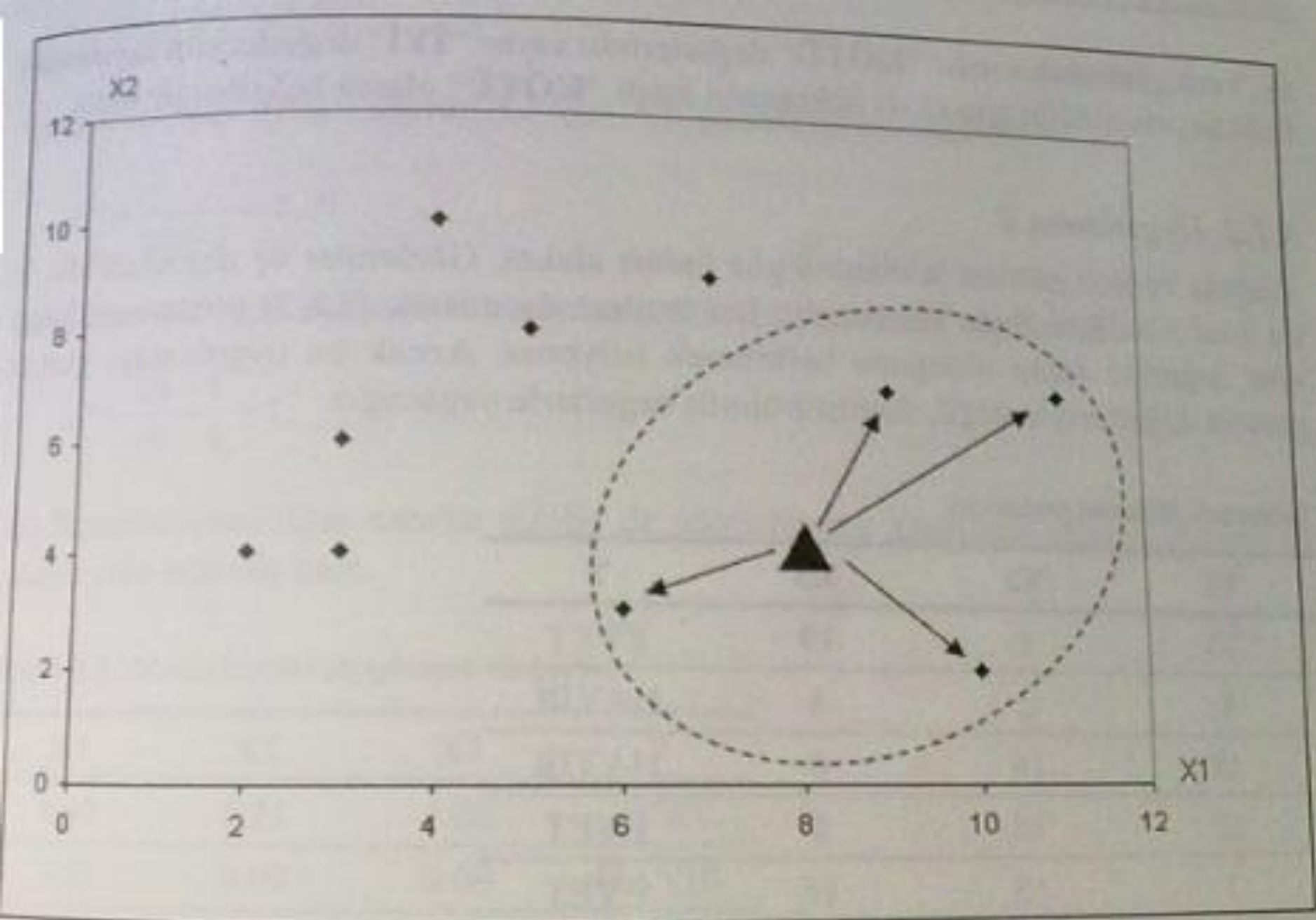


X1	X2	Uzaklık	Sıra
2	4	6.00	9
3	6	5.39	8
3	4	5.00	6
4	10	7.21	10
5	8	5.00	5
6	3	2.24	1
7	9	5.10	7
9	7	3.16	3
11	7	4.24	4
10	2	2.83	2



Tablo3.Uzaklık göz önüne alınarak k=4 komşu gözlemin belirlenmesi





Sekil-5.4. (8,4) noktasına en yakın dört komşu

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

X1	X2	Uzaklık	Sıra	K komşusunun Y değeri
2	4	6.00	9	
3	6	5.39	8	
3	4	5.00	6	
4	10	7.21	10	
5	8	5.00	5	
6	3	2.24	1	İYİ
7	9	5.10	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 değerinin sayısından fazla sayıda olduğu için (8,4) noktasının sınıfı KÖTÜ olarak belirlenmiştir.

## Uygulama 2

Aşağıda verilen gözlem tablosunu göz önüne alalım. Gözlemler üç değişkenlidir. **Y** ise sınıf niteliğini ifade etmektedir. Bu verilere dayanarak (7,8,5) noktasının hangi sınıf değerine sahip olduğunu belirlemek istiyoruz. Ancak bu uygulamayı gerçek gözlem değerleriyle değil dönüştürülmüş değerlerle yapacağız.

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

Tablo5. Gözlem değerleri

Gözlem değerlerini (0,1) aralığına göre dönüştürmek için min-max normalleştirme yöntemini uygulayacağız. Bu amaçla tablo5 deki gözlem değerleri için,

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

Bağıntısı uygulanır. Söz konusu bağıntıyı kitabımızın ikinci bölümünde ele alınarak incelemiştik. Burada  $X^*$  dönüştürülmüş değerleri,  $X$  gözlem değerlerini,  $X_{min}$  en küçük değerini ve  $X_{max}$  en büyük gözlem değerini ifade etmektedir. Bu değerler aşağıdaki tablo üzerinde yer almaktadır.

	X1	X2	X3
Xm	3	2	4
Xmax	18	16	19

Tablo6. Dönüştürme işleminde kullanılan bazı değerler

Birinci satırda X1 için X\* şu şekilde elde edilir:

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

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

Benzer biçimde birinci satırda X2 için X\* şu şekilde bulunur.




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

Benzer biçimde birinci satırda X3 için X\* şu şekilde bulunur.

$$= \frac{19-4}{19-4} = 1$$

Bu hesaplamalar diğer satırlar içinde de tekrarlanırsa **tablo7** elde edilir.












X1	X2	X3	Y
0.47	0.21	1.00	EVET
0.33	0.00	0.00	HAYIR
1.00	1.00	0.13	HAYIR
0.60	0.93	0.27	EVET
0.00	0.93	0.73	EVET

Tablo7. Dönüştürülmüş gözlem değeri

Bu durumda sınıflandırmaya tabi tutulacak (7,8,5) gözlemi de aynı dönüşüm formülüyle yeni değerlere dönüştürülür. Bununla ilgili yeni gözlem noktası (0.26, 0.43, 0.07) biçiminde elde edilir. Yeni gözlemler elde edildiğine göre artık k-en yakın komşu algoritmasını uygulayabiliriz.

- 
- 
- 
- 
- a) **K nın belirlenmesi:** K-en yakın komşu algoritması için k=3 kabul ederek çözülmeye başlıyoruz.
  - b) **Uzaklıkların hesaplanması:** (0.26, 0.43, 0.07) noktası ile dönüştürülmüş gözlem değerlerinin her birisi arasındaki Öklid uzaklıklarını hesaplandığında **tablo8** elde edilir.

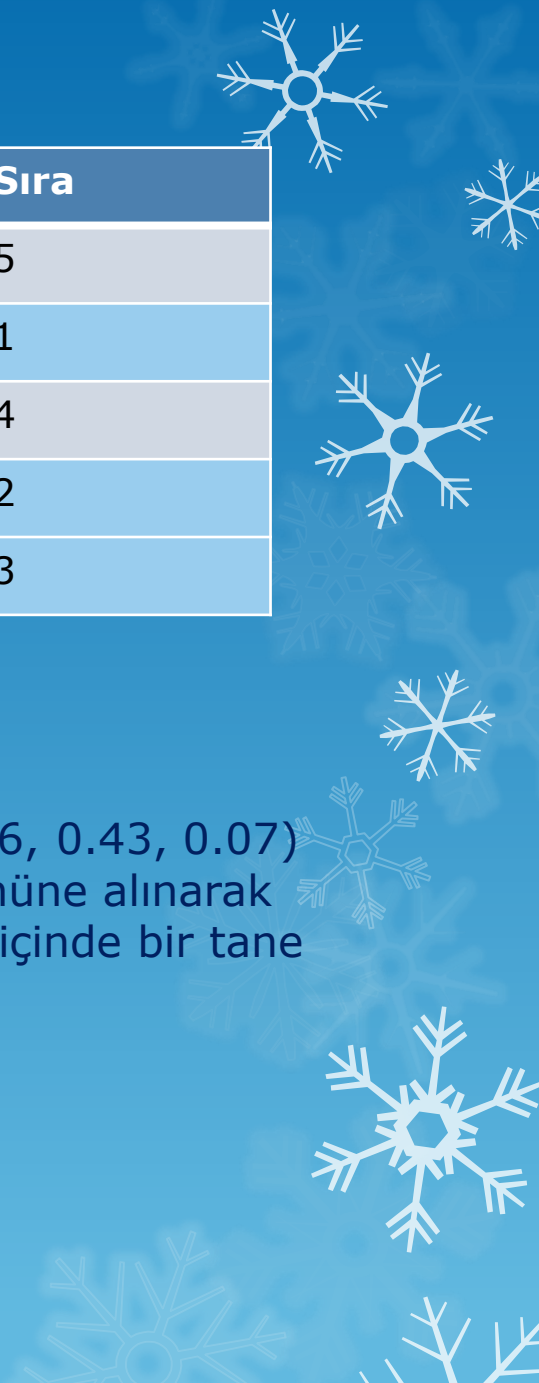




X1	X2	X3	Uzaklık
0.47	0.21	1.00	0.98
0.33	0.00	0.00	0.44
1.00	1.00	0.13	0.93
0.60	0.93	0.27	0.63
0.00	0.93	0.73	0.87

Tablo8. Gözlem değerlerinin (0.26, 0.43, 0.07) noktasına olan uzaklıkları

c) **En küçük uzaklıkların belirlenmesi:** Satırlar sıralanarak, en küçük  $k=3$  tanesi belirlenir. Bu üç nokta verilen (0.26, 0.43, 0.07) noktasına en yakın gözlem değerleridir.



X1	X2	X3	Uzaklık	Sıra
0.47	0.21	1.00	0.98	5
0.33	0.00	0.00	0.44	1
1.00	1.00	0.13	0.93	4
0.60	0.93	0.27	0.63	2
0.00	0.93	0.73	0.87	3

Tablo9.Uzaklık göz önüne alınarak k=3 komşu gözlemin belirlenmesi

d) **Seçilen satırlara ilişkin sınıflarının belirlenmesi:**(0.26, 0.43, 0.07) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınarak hangisinin daha çok tekrarlandığını belirliyoruz. Bu üç sonuç içinde bir tane **HAYIR**, üç tane **EVET** sonucu vardır.








X1	X2	X3	Uzaklık	Sıra	K komşunun Y değeri
0.47	0.21	1.00	0.98	5	
0.33	0.00	0.00	0.44	1	HAYIR
1.00	1.00	0.13	0.93	4	
0.60	0.93	0.27	0.63	2	EVET
0.00	0.93	0.73	0.87	3	EVET



Tablo10.Y sınıfına ilişkin ilk 3 değerin belirlenmesi

e)**Yeni gözlemin sınıfı:** Seçilenler arasında **EVET'** lerin sayısı diğerinden daha fazladır. O halde (7,8,5) gözleminin, yani dönüştürülmüş değerlerle ifade edilir.(0.26, 0.43, 0.07) gözleminin de sınıfı **EVET** olarak kabul edilir.



# Ağırlıklı Oylama

K- algoritması, verilen bir gözleme en yakın komşunun belirlenmesi ve sınıfı yeni bir gözlem değeri için, k gözlem içindeki en fazla tekrar eden sınıfın seçilmesi esasına dayanıyordu. 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 yöntemi uygulanır.

Bu yöntem gözlem değerlerini için aşağıdaki bağıntıya göre ağırlıklı uzunlukların hesaplanması esasına dayanır.

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

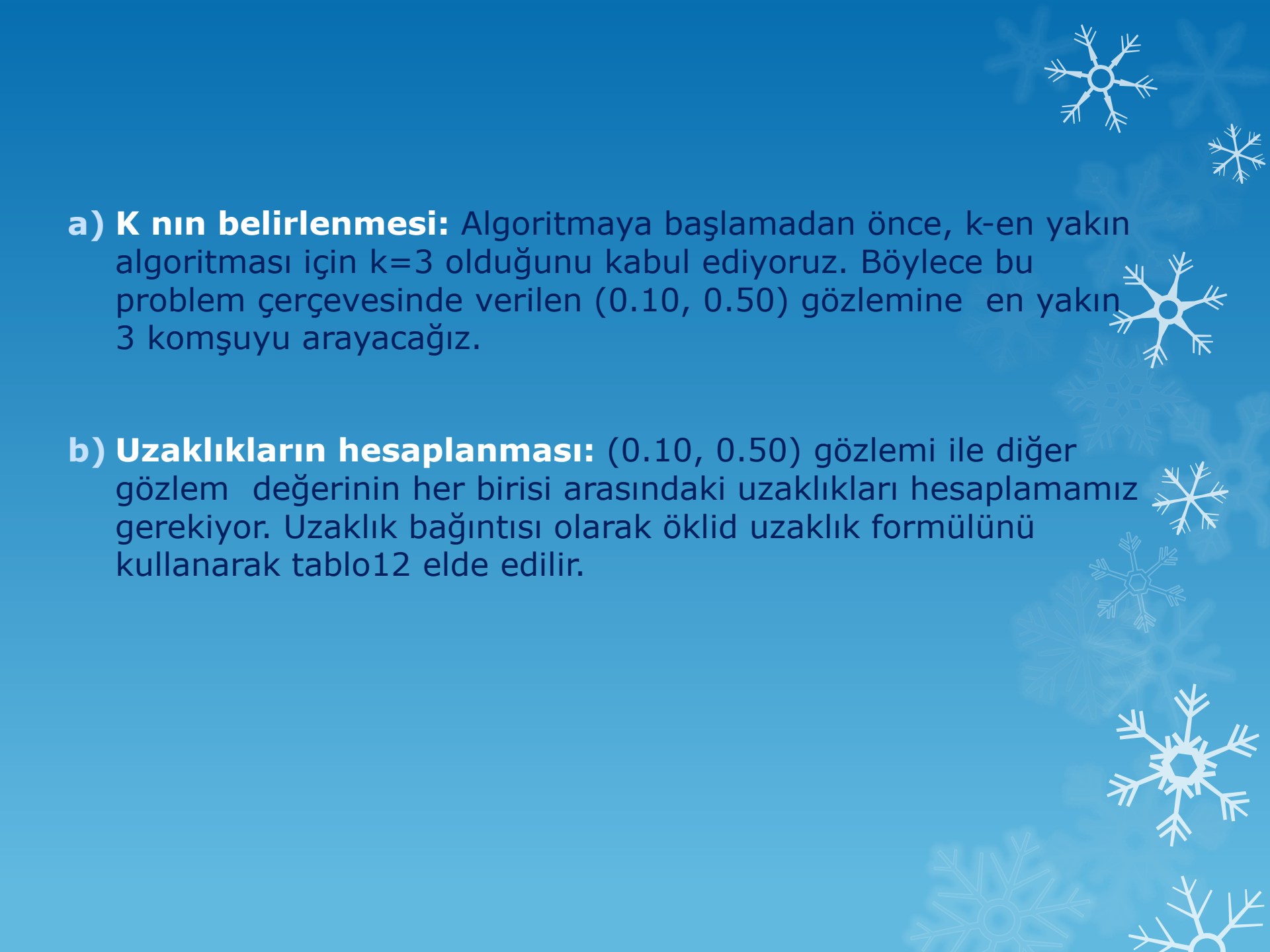
$d(i,j)$  ifadesi i ve j gözlemleri arasındaki Öklid uzaklığıdır. Her bir sınıf değeri için bu uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri elde edilir. En büyük ağırlıklı oylama değerine sahip olan sınıf değeri yeni gözlemin ait olduğu sınıf kabul edilir.

## Uygulama3

Aşağıda verilen gözlem tablosunu göz önüne alalım. Bu gözlemler **X1** ve **X2** niteliklerinden ve **CINS** sınıfından oluşmaktadır. Bu gözlem değerlerine bağlı olarak, yeni bir gözlem olan (0.10, 0.50) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu yöntemiyle bulalım


X1	X2	CINS
0.08	0.20	ERKEK
0.07	0.07	ERKEK
0.20	0.09	ERKEK
1.00	0.20	KADIN
0.05	0.06	ERKEK
0.20	0.25	ERKEK
0.17	0.07	ERKEK
0.15	0.55	KADIN
0.50	0.08	ERKEK
0.10	0.06	KADIN

Tablo11.Gözlem değerleri



**a) K'nın belirlenmesi:** Algoritmaya başlamadan önce, k-en yakın algoritması için  $k=3$  olduğunu kabul ediyoruz. Böylece bu problem çerçevesinde verilen  $(0.10, 0.50)$  gözlemine en yakın 3 komşuyu arayacağız.









**b) Uzaklıkların hesaplanması:**  $(0.10, 0.50)$  gözlemi ile diğer gözlem değerinin her birisi arasındaki uzaklıkları hesaplamamız gerekiyor. Uzaklık bağıntısı olarak öklid uzaklık formülünü kullanarak tablo12 elde edilir.



X1	X2	Uzaklık
0.08	0.20	0.30
0.07	0.07	0.43
0.20	0.09	0.42
1.00	0.20	0.95
0.05	0.06	0.44
0.20	0.25	0.27
0.17	0.07	0.43
0.15	0.55	0.07
0.50	0.08	0.58
0.10	0.06	0.44

Tablo12.Gözlem değerlerinin verilen bir (0.10, 0.50) noktasına olan uzaklık

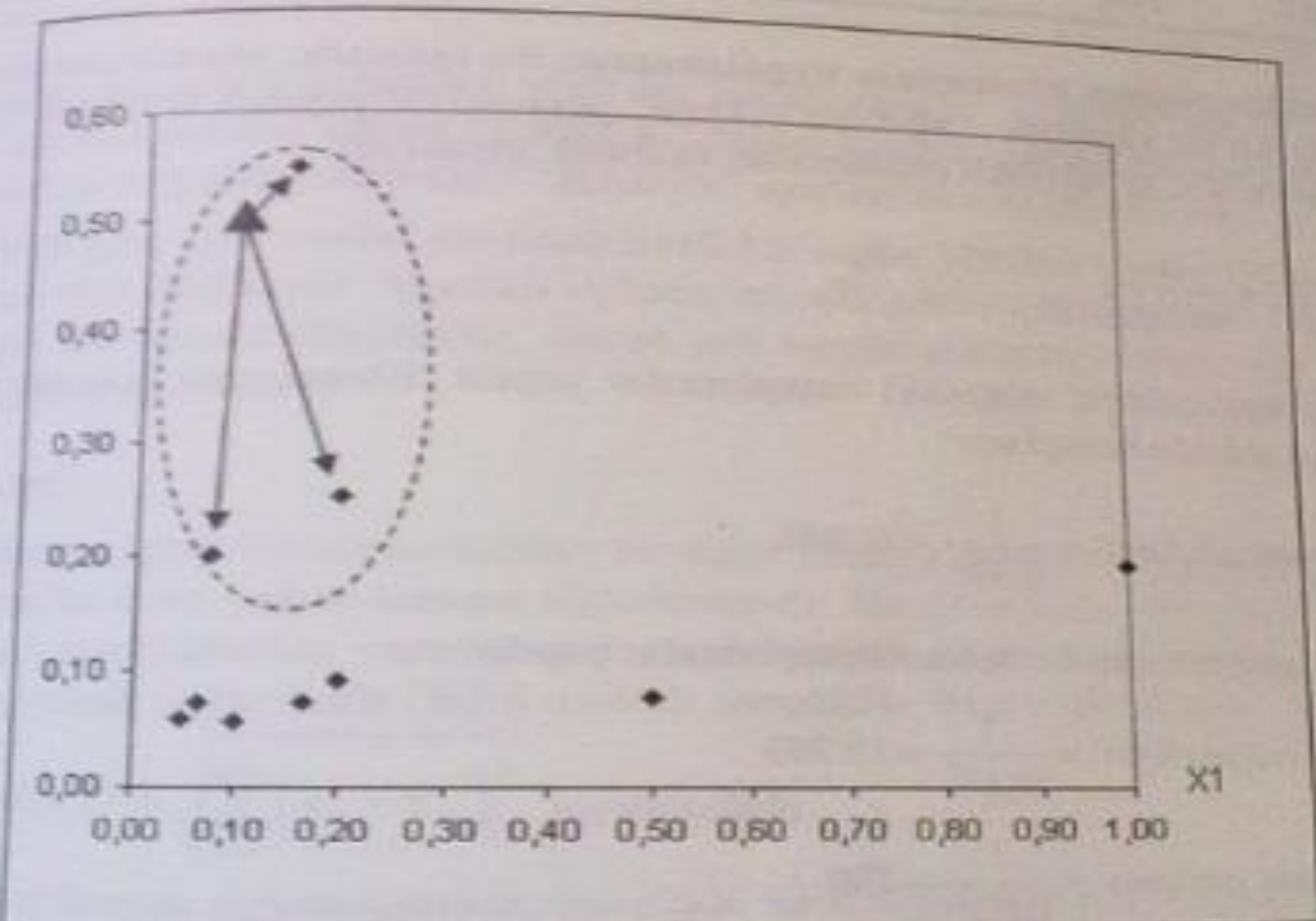
c) **En küçük uzaklıkların belirlenmesi:** Satırlar sıralanarak, en küçük k=3 tanesi belirleniyor. Bu üç nokta yeni gözlem noktasına en yakın noktalardır.



X1	X2	Uzaklık	Sıra
0.08	0.20	0.30	3
0.07	0.07	0.43	5
0.20	0.09	0.42	4
1.00	0.20	0.95	10
0.05	0.06	0.44	7
0.20	0.25	0.27	2
0.17	0.07	0.43	6
0.15	0.55	0.07	1
0.50	0.08	0.58	9
0.10	0.06	0.44	8

Tablo13. Uzaklık göz önüne alınarak k=3 komşu gözlemin belirlenmesi





Şekil-5.3. (0,10, 0,50) noktasına en yakın üç komşu

**d)Seçilen satırlara ilişkin sınıfların belirlenmesi:** Yeni gözlem noktasına en yakın olan gözlem değerlerinin CINS sınıfları göz önüne alınır ve içine hangi değer baskın olduğu araştırılır. Bu üç sonuç içinde iki tane **ERKEK**, bir tane **KADIN** değeri vardır.

X1	X2	Uzaklık	Sıra	k komşusunun CINS değeri
0.08	0.20	0.30	3	ERKEK
0.07	0.07	0.43	5	
0.20	0.09	0.42	4	
1.00	0.20	0.95	10	
0.05	0.06	0.44	7	
0.20	0.25	0.27	2	ERKEK
0.17	0.07	0.43	6	
0.15	0.55	0.07	1	KADIN
0.50	0.08	0.58	9	
0.10	0.06	0.44	8	

Tablo14. Y sınıfına ilişkin ilk 3 değerin belirlenmesi

**e)Ağırlıklı oylama yönteminin uygulanması:** Bu aşamada söz konusu seçme işlemini bir yöntemle, ağırlıklı oylama yöntemiyle yapmak istiyoruz o halde son tabloya  $d(i,j)$  ağırlıklı ortalamaları eklemek gerekiyor.

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

Bağıntısı kullanılarak aşağıdaki hesaplamalar yapılır. Birinci satır için

$$d(1,\text{yeni gözlem})' = \frac{1}{0.302} = 11.046$$

Diğer iki gözlem için benzer hesaplamalar yapılır.

$$d(6,\text{yeni gözlem})' = \frac{1}{0.27^2} = 13.793$$

$$d(1,\text{yeni gözlem})' = \frac{1}{0.07^2} = 200$$

Hesaplanan değerler **tablo15** 'tedir.

Bu durumda **CINS** sınıfının **ERKEK** değerleri için ağırlıklı oylama değeri hesaplanır.

$$d(1,\text{yeni gözlem})' + d(6,\text{yeni gözlem})' = 11.046 + 13.793 = 24.84$$



Gözlem	X1	X2	Uza klık	Sıra	k komşusunun CİNS değeri	Ağırlık uzaklık
1	0.08	0.20	0.30	3	ERKEK	11.05
2	0.07	0.07	0.43	5		
3	0.20	0.09	0.42	4		
4	1.00	0.20	0.95	10		
5	0.05	0.06	0.44	7		
6	0.20	0.25	0.27	2	ERKEK	13.79
7	0.17	0.07	0.43	6		
8	0.15	0.55	0.07	1	KADIN	200.00
9	0.50	0.08	0.58	9		
10	0.10	0.06	0.44	8		








Tablo15.Ağırlık uzaklık değerleri

- Elde edilen sonuçlara göre şöyle bir yorum yapılır: **KADIN** değeri için elde edilen ağırlıklı oylama değeri **ERKEK** değeri için elde edilenden daha büyük olduğundan, yeni gözlem değerinin **KADIN** sınıfına ait olduğu anlaşılır.
- Görüldüğü gibi, aynı veriler üzerinde farklı bir seçme yöntemi uygulanmış ve farklı bir sonuç elde edilmiştir. Ağırlıklı oylama aslında gözlem değerlerinin tümüne uygulanarak bir sonuca ulaşılabilir. Ancak çok sayıda veri kümelerinde işlemi yavaşlatır.