**VERİ MADENCİLİĞİ**

**1.Hafta**

1. **Veri ambarı :** Veri ambarı zaman içerisinde olabildiğince birikmiş verilerin oluşturduğu bir veri yığınıdır. Bir işletmenin sahip olduğu verilerin karar destek amacıyla kullanılmasına olanak sağlar. Veri madenciliği yapmak için tasarlanan bir ortam.
   1. **OLTP Sistemler :** Bir kurumun günlük verilerinin işlendiği ortamlara OLTP(Online Transaction Processing) (veritabanı) sistemler adı verilmektedir.
   2. **Karar Destek Sistemleri :** Karar Destek Sistemleri (Desicion Support Systems) ve Üst Yönetici Sistemleri (Executive Information Systems) bu amaçla ortaya atılmıştır.
   3. **Veri Ambarı Nedir? :** Veri ambarı en basit anlamda, karar destek uygulamaları için tasarlanan bir ortam olarak tanımlıyoruz.

Veri ambarı , karar verme sürecinde yöneticilere destek vermek üzere hazırlanmış;

* + Konuya yönelik
  + Bütünleşik
  + Zaman boyutu olan
  + Sadece okunabilen

Veri topluluğudur.

NOTT: video 1 6:45

* 1. **Veri Ambarının İçerdiği Veri :** Veri ambarı, içerdiği veri açısından da göz önüne alındığında farklı bir yapıya sahip olduğu anlaşılacaktır. Aşağıda veri ambarının içerdiği veriyi sınıflandırıyoruz;
  + Metadata
  + Ayrıntı veri
  + Eski ayrıntı veri
  + Düşük düzeyde özetlenmiş veri
  + Yüksek düzeyde özetlenmiş veri

**2.Hafta**

**Veri Madenciliği**

Basit tanımı, veri madenciliği, büyük ölçekli veriler arasında bir bilgiyi elde etme işidir. Bu sayede veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik kestirimlerde de bulunmak mümkün görünmektedir.

Veri madenciliği süreçleri;

* + Veri temizleme
  + Veri bütünleştirme
  + Veri indirgeme
  + Veri dönüştürme
  + Veri madenciliği algoritmasını uygulama
  + Sonuçları sunum ve değerlendirme

Veri madenciliği modellerini temel olarak şu şekilde gruplandırabilriz;

* + Sınıflandırma
  + Kümeleme
  + Birliktelik kuralları

**3-4.Hafta**

Karar Ağaçları ile Sınıflandırma;

Verinin içerdiği ortak özelliklere göre bir veri veya veri grubunun hangi sınıfa dahil olduğunun belirlenmesi işlemi sınıflandırma olarak adlandırılır.

Verilerin sınıflandırma süreci iki adımdan oluşur ;

1. İlk adım, veri kümelerine uygun bir modelin ortaya konulmasıdır. Bu model, veri tabanında ki kayıtların nitelikleri kullanılarak gerçekleştirilir. Sınıflandırma modellinin elde edilebilmesi için veri tabanının bir kısmı eğitim verileri olarak kullanılabilir.

|  |  |  |  |
| --- | --- | --- | --- |
| Müşteri | Borç | Gelir | Risk |
| Ali | Yüksek | Yüksek | Kötü |
| Ayşe | Yüksek | Yüksek | Kötü |
| Kenan | Yüksek | Düşük | Kötü |
| Burak | Düşük | Yüksek | İyi |
| Begüm | Düşük | Düşük | Kötü |
| Seray | Düşük | Yüksek | İyi |

Sınıflandırma Algoritması

Sınıflandırıcı Model = Eğer Borç=Yüksek => Risk=Kötü, Eğer Borç=Düşük ve Gelir=Düşük => Risk=Kötü, Eğer Borç=Düşük ve Gelir=Yüksek => Risk=İyi

1. Test verileri üzerinde sınıflandırma kuralları belirlenir. Ardından söz konusu kurallar bu kez test verilerine uygulanarak sınanır.

|  |  |  |  |
| --- | --- | --- | --- |
| Müşteri | Borç | Gelir | Risk |
| Ali | Yüksek | Yüksek | Kötü |
| Ayşe | Yüksek | Düşük | Kötü |
| Kenan | Düşük | Düşük | Kötü |
| Burak | Düşük | Yüksek | İyi |

Sınıflandırma Algoritması

Sınıflandırıcı Model = Eğer Borç=Yüksek => Risk=Kötü, Eğer Borç=Düşük ve Gelir=Düşük => Risk=Kötü, Eğer Borç=Düşük ve Gelir=Yüksek => Risk=İyi

|  |  |  |  |
| --- | --- | --- | --- |
| Müşteri | Borç | Gelir | Risk |
| Ali | Düşük | Yüksek | ??? |

RİSK=İYİ

Karar Ağaçlarında Dallanma Kriterleri

Her farklı kriter için bir karar ağacı algoritması karşılık gelmektedir. Algoritmaları sınıflandıracak olursak;

1. Entropiye Dayalı Algoritmalar
2. Sınıflandırma ve Regresyon Ağaçları
3. Bellek Tabanlı Sınıflandırma Algoritmaları

ID3 Algoritması

ID3 ve C4.5 Algoritmaları entropi tabanlı algoritmalardır.

H(S) =

Hesap makinesinde yaparken pi=1/2 diyelim ; -( (log101/2)1/2/(log102)

Karar Ağaçlarında Entropi

H(X,T)=

Kazanç(X,T)=H(T)-H(X,T)

Örnek:

Aşağıda sekiz elemanlı S kümesini göz önüne alalım.

S={𝑒𝑣𝑒𝑡,𝑒𝑣𝑒𝑡,ℎ𝑎𝑦𝚤𝑟,ℎ𝑎𝑦𝚤𝑟,ℎ𝑎𝑦𝚤𝑟,ℎ𝑎𝑦𝚤𝑟,ℎ𝑎𝑦𝚤𝑟,ℎ𝑎𝑦𝚤𝑟}

Olasılıklar, iki adet ’’evet’’ değeri için,

P1=2/8=0.25

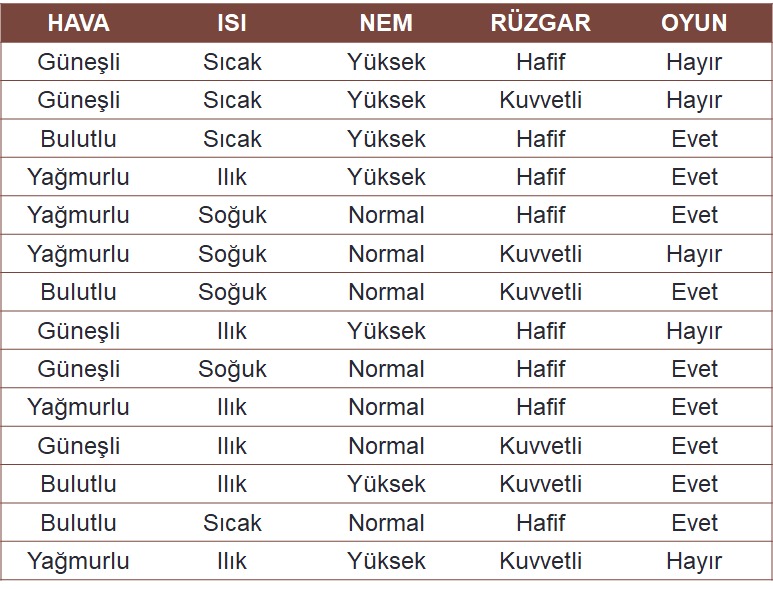
Diğer altı adet ’’hayır’’ değeri için,

P2=6/8=0.75

S için toplam entropi şu şekilde elde edilir;

𝐻(𝑆)=−{𝑃1 log2(𝑃1 )+𝑃2 log2(𝑃2) }

= -(0.25\*log2(0.25)+0.75\*log2 (0.75))=0.81128

Örnek: 

1.adım: H(OYUN) bulunur yani H(S) formülü uygulanır.

H(OYUN)=-(5/14\*log2(5/14)+9/14\*log2(9/14))=0.940

2.adım: HAVA, ISI, NEM ve RÜZGAR için H(X,T) formülü uygulanır. Daha sonra Kazanç(X,T) formülü uygulanır.

HAVA; güneşli=5, yağmurlu=5, bulutlu=4,

H(HAVAgüneşli)=-(3/5\*log23/5+2/5\*log22/5)=0.971

H(HAVAyağmurlu)=-(3/5\*log23/5+2/5\*log22/5)=0.971

H(HAVAbulutlu)=-(4/4\*log24/4+0/4\*log20/4)=0

H(HAVA,OYUN)=(5/14(0.971)+5/14\*(0.971)+4/14\*0)=0.693

KAZANÇ(HAVA,OYUN)=H(OYUN)-H(HAVA,OYUN)=0.940-0.693= 0.247

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ISI için; sıcak=4, ılık=6,soğuk=4

H(ISIsıcak)=-(2/4\*log2(2/4)+2/4\*log2(2/4))=1

H(ISIılık)= -(2/6\*log2(2/6)+4/6\*log2(4/6))=0.918

H(ISIsoğuk)= -(3/4\*log2(3/4)+1/4\*log2(1/4))=0.811

H(ISI,OYUN)=(4/14)\*1+(4/14)\*0.811+(6/14)\*0.918=0.911

KAZANÇ(ISI,OYUN)=H(OYUN)-H(ISI,OYUN)=0.940-0.911=0.029

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Nem için; yüksek=7, normal=7;

H(NEMyüksek)=-(3/7\*log2(3/7)+4/7\*log2(4/7))=0.985

H(NEMnormal)= -(6/7\*log2(6/7)+1/7\*log2(1/7))=0.592

H(NEM,OYUN)=(7/14)\*(0.985)+(7/14)\*(0.592)=0.789

KAZANÇ(NEM,OYUN)=H(OYUN)-H(NEM,OYUN)=0.940-0.789=0.151

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Rüzgar için; hafif=8, kuvvetli=6;

H(RÜZGARhafif)=-(6/8\*log2(6/8)+2/8\*log2(2/8))=0.811

H(RÜZGARkuvvetli)= -(3/6\*log2(3/6)+3/6\*log2(3/6))=1

H(RÜZGAR,OYUN)=(8/14)\*(0.811)+(6/14)\*(1)=0.892

KAZANÇ(RÜZGAR,OYUN)=H(OYUN)-H(RÜZGAR,OYUN)=0.940-0.892=0.048

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

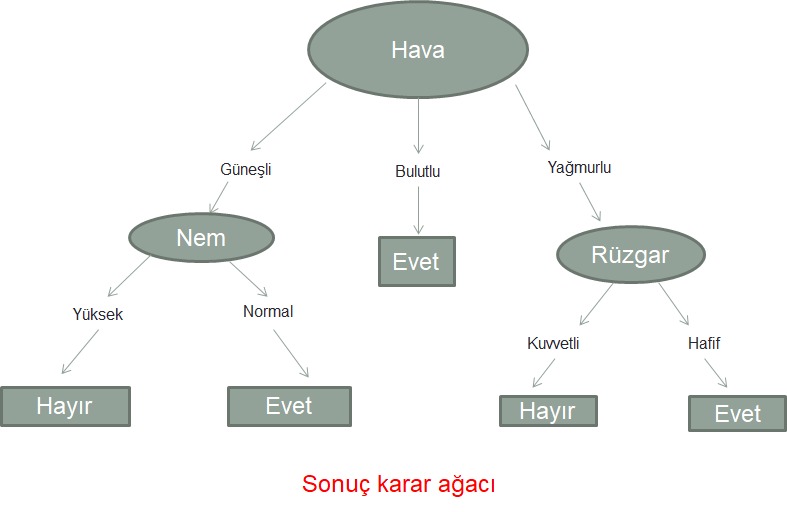
HAVA=0.246

ISI=0.029

NEM=0.151

RÜZGAR=0.048

Bu işlemlerden sonra hava için güneşliyi ele alırız. Aynı işlemleri yeniden yaparız en yüksek çıkan güneşlinin kararına etkilenir. Daha sonra yağmurlu ve bulutlu içinde yapıp karar ağacı oluşturulur.



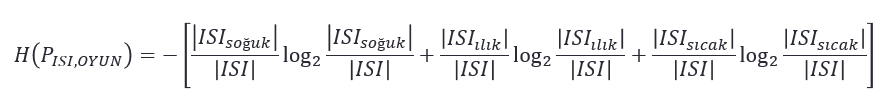
KAZANÇ ORANI

𝐻(𝑃(𝑋,𝑇) )=

KAZANÇ ORANI(X,T)=

ÖRNEK:

Önceki tablomuzdan ısı niteliği ile ilgili olarak kazanç oranını hesaplamak istiyoruz. T1={evet,hayır,evet,hayır} kümesi ele alınır. O halde H(Pısı,oyun) bağlantısı şu şekilde ifade edilir.



H(Pısı,oyun)= -(4/14log2(4/14)+6/14log2 (6/14)+ 4/14log2(4/14))=1.557

KAZANÇ(ISI,OYUN)=0.029 hesaplanmıştı.

KAZANÇ ORANI(ISI,OYUN)=0.029/1.557=0.018

C4.5 Algoritması

EĞİTİM KÜMESİ GÖRÜNÜMÜ

|  |  |  |  |
| --- | --- | --- | --- |
| NİTELİK1 | NİTELİK2 | NİTELİK3 | SINIF |
| A | 70 | DOĞRU | SINIF1 |
| A | 90 | DOĞRU | SINIF2 |
| A | 85 | YANLIŞ | SINIF2 |
| A | 95 | YANLIŞ | SINIF2 |
| A | 70 | YANLIŞ | SINIF1 |
| B | 90 | DOĞRU | SINIF1 |
| B | 78 | YANLIŞ | SINIF1 |
| B | 65 | DOĞRU | SINIF1 |
| B | 75 | YANLIŞ | SINIF1 |
| C | 80 | DOĞRU | SINIF2 |
| C | 70 | DOĞRU | SINIF2 |
| C | 80 | YANLIŞ | SINIF1 |
| C | 70 | YANLIŞ | SINIF1 |
| C | 96 | YALIŞ | SINIF1 |

Vİ = ortanca sayı(küçükten büyüğe sıralandığında) Tİ=eşik değeri

Ti=(Vİ+ Vİ+1)/2

{65,70,75,80,85,90,95,96}

Vİ=80 Vİ+1=85

Tİ=(80+85)/2=82.5=~83

Artık tablomuz bu hale geldi;

|  |  |  |  |
| --- | --- | --- | --- |
| NİTELİK1 | NİTELİK2 | NİTELİK3 | SINIF |
| A | Eşit veya küçük | DOĞRU | SINIF1 |
| A | Büyük | DOĞRU | SINIF2 |
| A | Büyük | YANLIŞ | SINIF2 |
| A | Büyük | YANLIŞ | SINIF2 |
| A | Eşit veya küçük | YANLIŞ | SINIF1 |
| B | Büyük | DOĞRU | SINIF1 |
| B | Eşit veya küçük | YANLIŞ | SINIF1 |
| B | Eşit veya küçük | DOĞRU | SINIF1 |
| B | Eşit veya küçük | YANLIŞ | SINIF1 |
| C | Eşit veya küçük | DOĞRU | SINIF2 |
| C | Eşit veya küçük | DOĞRU | SINIF2 |
| C | Eşit veya küçük | YANLIŞ | SINIF1 |
| C | Eşit veya küçük | YANLIŞ | SINIF1 |
| C | Büyük | YALIŞ | SINIF1 |

Buradan sonra sınıfa göre ID3 algoritması uygulanmaya başlar.

Bilinmeyen Nitelik Değerleri

F=(Veri tabanında bilinen niteliğe sahip önceliklerin sayısı)/ (Veri tabanındaki tüm örneklerin sayısı)

Kazanç(X)=F\*(H(T)-H(X,T))

ÖRNEK:

|  |  |  |  |
| --- | --- | --- | --- |
| NİTELİK1 | NİTELİK2 | NİTELİK3 | SINIF |
| A | 70 | DOĞRU | SINIF1 |
| A | 90 | DOĞRU | SINIF2 |
| A | 85 | YANLIŞ | SINIF2 |
| A | 95 | YANLIŞ | SINIF2 |
| A | 70 | YANLIŞ | SINIF1 |
| ? | 90 | DOĞRU | SINIF1 |
| B | 78 | YANLIŞ | SINIF1 |
| B | 65 | DOĞRU | SINIF1 |
| B | 75 | YANLIŞ | SINIF1 |
| C | 80 | DOĞRU | SINIF2 |
| C | 70 | DOĞRU | SINIF2 |
| C | 80 | YANLIŞ | SINIF1 |
| C | 70 | YANLIŞ | SINIF1 |
| C | 96 | YALIŞ | SINIF1 |

Psınıf=-8/13\*log2(8/13)+ 5/13\*log2(5/13) =0.961

Nitelik1 için; a=5, b=3,c=5

H(NİTELİK1a)= -2/5\*log2(2/5)+3/5\*log2(3/5) =0.971

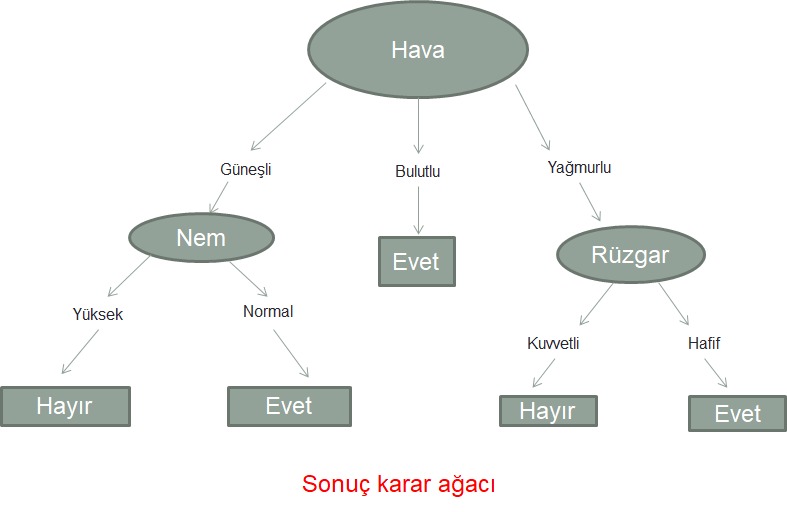
H(NİTELİK1b)= -3/3\*log2(3/3)+0/3\*log2(0/3) =0

H(NİTELİK1c)= -2/5\*log2(2/5)+3/5\*log2(3/5) =0.971

H(NİTELİK1,SINIF)=5/13\*(0.971)+5/13\*(0.971)+3/13\*0=0.747

KAZANÇ(NİTELİK1,SINIF)=13/14\*(0.961-0.747)=0.199

Karar Kuralları Oluşturmak



KURAL1:

Eğer hava=güneşli => ve Eğer nem=yüksek =>

OYUN=HAYIR

KURAL2:

Eğer hava=güneşli => ve Eğer nem=normal =>

OYUN=EVET

KURAL3:

Eğer hava=bulutlu =>

OYUN=EVET

KURAL4:

Eğer hava=yağmurlu => ve Eğer rüzgar=güçlü =>

OYUN=HAYIR

KURAL5:

Eğer hava=yağmurlu => ve Eğer rüzgar=düşük =>

OYUN=EVET

**5-6.Hafta**

Sınıflandırma ve Regresyon Ağaçları ( Cart Algoritması )

* + Twoing Algoritması
  + Gini Algoritması

Her bir karar düğümünden itibaren ağacın ii dala ayrılması ilkesine dayanır.

TWOİNG ALGORİTMASI

ADIM1:

1. Niteliklerin içerdiği değerler göz önüne alınarak eğitim kümesi iki ayrı dala ayrılır. Bunlara aday bölünme adı veriliyor. Bir t düğümünde sağ ve sol olmak üzere 2 ayrı dal bulunur. Bu bölünen kümeler tsağ ve tsol biçimindedir.
2. Aday bölünmelerinin her biri için Psol ve P(j|sol) olasılıkları hesaplanır. Söz konusu olasılıklar aşağıda verilmektedir. Burada P(j|tsol) ifadesi bir j sınıf değerinin sol taraftaki bölünmede olma olasılığını verir. Söz konusu j değerleri sınıf değerlerinin yer aldığı nitelik olarak göz önüne alınır.

Psol = tsol daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı

Eğitim kümesindeki kayıtların sayısı

P(j|tsol)= tsol daki kayıtların j sayısı

tsol daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı

1. Aday bölünmelerinin her biri için Psağ ve P(j|sağ) olasılıkları hesaplanır. Söz konusu olasılıklar aşağıda verilmektedir. Burada P(j|tsağ) ifadesi bir j sınıf değerinin sağtaraftaki bölünmede olma olasılığını verir.

Psağ = tsağ daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı

Eğitim kümesindeki kayıtların sayısı

P(j|tsağ)= tsağ daki kayıtların j sayısı

tsağ daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı

1. Φ(s|t),t düğümündeki s aday bölümlerinin uygunluk ölçüsü olsun. Söz konusu uygunluk ölçüsü şu şekilde hesaplanır.

Φ(s|t)=2\*Psol\*Psağ \*

1. Φ(s|t) değerleri hesaplandıktan sonra içlerinden en büyük olanı seçilir. Bu değerin ilgili olduğu aday bölünme satırı bize dallanmanın yapılacağı satırı bildirecektir.
2. Dallanma bu şekilde yapıldıktan sonra bu adıma ilişkin karar ağacı çizilir.

ADIM2: Algoritmanın 1. Adımına dönülerek ağacın alt kümesine aynı işlemler uygulanır.

Örnek : Müşterilerini cinsiyetlerine göre ve alışveriş miktarına göre sınıflandırmak isteyen bir mağazada müşteri 5 müşteri için tabloda 1’de verilen sonuçların elde edildiğini varsayalım. Bu tabloda sınıflandırmada yapıldığı sınıf niteliği hedef nitelik olarak kabul edilir.

CİNSİYET = KADIN, DEĞERİ= SOL TARAFTA ,

CİNSİYET = ERKEK, DEĞERİ=SAĞ TARAFTA olacak biçimde bölündüğünü varsayalım. Bu durumda cinsiyet = erkek değeri için Psol ve P(A|tsol) olasılığını hesaplayalım.

|  |  |  |  |
| --- | --- | --- | --- |
| MÜŞTERİ | CİNSİYET | SATIŞ | SINIF |
| 1 | ERKEK | DÜŞÜK | A |
| 2 | KADIN | YÜKSEK | A |
| 3 | ERKEK | YÜKSEK | B |
| 4 | ERKEK | NORMAL | A |
| 5 | KADIN | DÜŞÜK | B |

Tablo1

|  |  |
| --- | --- |
| Aday bölünme Tsol | T~~sağ~~ |
| Cinsiyet= ERKERK | Cinsiyet =KADIN |
| Cinsiyet=KADIN | Cinsiyet=ERKEK |
| Satış=DÜŞÜK | Satış {NORMAL, YÜKSEK} |
| Satış=NORMAL | Satış {DÜŞÜK, YÜKSEK} |
| Satış=YÜKSEK | Satış {DÜŞÜK, NORMAL |

Tablo2

İlk öncelik cinsiyet olacak şekilde ve kadınlar solda olacak şekilde düşünürsek;

Psağ = 3/5 = 0.6 Psol=2/5 = 0.40

P(A|tsağ)=2/3 = 0.67 P(A|tsol)=1/2 =0.50

P(B|tsağ)=1/3 = 0.33 P(B|tsol)=1/2 =0.50

Φ(s|t)=2\*0.40\*0.6\*[|0.50-0.33|+|0.50-0.67|]=0.16



1. Önceliğimiz gelir ve sol tarafta normaller olacak

Psol = 1/11 = 0.09 Psağ=10/11 = 0.91

P(Evet|tsol)=1/1 = 1 P(Evet|tsağ)=6/10 =0.6

P(Hayır|tsol)=0/1 = 0 P(Hayır|tsağ)=4/10 =0.4

Φ(s|t)=2\*0.09\*0.91\*[|1-0.6|+|0-0.4|]=0.13

1. Önceliğimiz gelir ve sol tarafta büyük olacak

Psol = 5/11 = 0.45 Psağ=6/11 = 0.54

P(Evet|tsol)=3/5 = 0.6 P(Evet|tsağ)=4/6 =0.67

P(Hayır|tsol)=2/5 = 0.4 P(Hayır|tsağ)=2/6 =0.33

Φ(s|t)=2\*0.45\*0.54\*[|0.6-0.67|+|0.4-0.33|]=0.07

1. Önceliğimiz gelir ve sol tarafta küçük olacak

Psol = 5/11 = 0.45 Psağ=6/11 = 0.54

P(Evet|tsol)=3/5 = 0.6 P(Evet|tsağ)=4/6 =0.67

P(Hayır|tsol)=2/5 = 0.4 P(Hayır|tsağ)=2/6 =0.33

Φ(s|t)=2\*0.45\*0.54\*[|0.6-0.67|+|0.4-0.33|]=0.07

1. Önceliğimiz eğitim ve sol tarafta ilk olacak

Psol = 2/11 = 0.18 Psağ=9/11 = 0.82

P(Evet|tsol)=2/2 = 1 P(Evet|tsağ)=5/9 =0.56

P(Hayır|tsol)=0/2 = 0 P(Hayır|tsağ)=4/9 =0.44

Φ(s|t)=2\*0.18\*0.82\*[|1-0.56|+|0-0.44|]=0.26

1. Önceliğimiz eğitim ve sol tarafta orta olacak

Psol = 5/11 = 0.45 Psağ=6/11 = 0.54

P(Evet|tsol)=3/5 = 0.6 P(Evet|tsağ)=4/6 =0.67

P(Hayır|tsol)=2/5 = 0.4 P(Hayır|tsağ)=2/6 =0.33

Φ(s|t)=2\*0.45\*0.54\*[|0.6-0.67|+|0.4-0.33|]=0.07

1. Önceliğimiz eğitim ve sol tarafta lise olacak

Psol = 4/11 = 0.36 Psağ=7/11 = 0.64

P(Evet|tsol)=2/4 = 0.5 P(Evet|tsağ)=5/7 =0.71

P(Hayır|tsol)=2/4 = 0.5 P(Hayır|tsağ)=2/7 =0.28

Φ(s|t)=2\*0.36\*0.64\*[|0.5-0.71|+|0.5-0.28|]=0.20

1. Önceliğimiz sektör ve sol bilişim olacak

Psol = 6/11 = 0.55 Psağ=5/11 = 0.45

P(Evet|tsol)=2/6 = 0.33 P(Evet|tsağ)=5/5 =1

P(Hayır|tsol)=4/6 = 0.67 P(Hayır|tsağ)=0/5 =0

Φ(s|t)=2\*0.55\*0.45\*[|0.33-1|+|0.67-0|]=0.66

1. Önceliğimiz sektör ve sol tarafta inşaat olacak

Psol = 5/11 = 0.45 Psağ=6/11 = 0.54

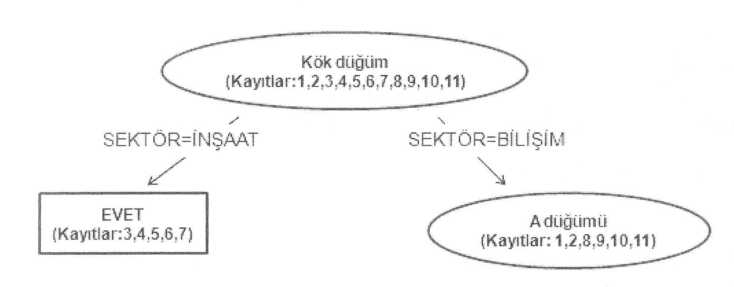
P(Evet|tsol)=5/5 = 1 P(Evet|tsağ)=2/6 =0.33

P(Hayır|tsol)=0/5 = 0 P(Hayır|tsağ)=4/6 =0.67

Φ(s|t)=2\*0.45\*0.54\*[|1-0.33|+|0-0.67|]=0.66



En büyük sektör bilişim ve sektör inşaat oldu. O halde İnşaat 🡪(3,4,5,6,7) , bilişim🡪(1,2,8,9,10,11)



İnşaat hepsi evet e gittiğinden dolayı A düğümünden işlemlere devam edeceğiz.



1. Önceliğimiz gelir ve sol tarafta normaller olacak

Psol = 1/6 = 0.17 Psağ=5/6 = 0.83

P(Evet|tsol)=1/1 = 1 P(Evet|tsağ)=1/5 =0.2

P(Hayır|tsol)=0/1 = 0 P(Hayır|tsağ)=4/5 =0.8

Φ(s|t)=2\*0.17\*0.83\*[|1-0.2|+|0-0.8|]=0.44

1. Önceliğimiz gelir ve sol tarafta büyük olacak

Psol = 3/6 = 0.5 Psağ=3/6 = 0.5

P(Evet|tsol)=1/3 = 0.33 P(Evet|tsağ)=1/3 =0.33

P(Hayır|tsol)=2/3 = 0.67 P(Hayır|tsağ)=2/3 =0.67

Φ(s|t)=2\*0.5\*0.5\*[|0.33-0.33|+|0.67-0.67|]=0.00

1. Önceliğimiz gelir ve sol tarafta küçük olacak

Psol = 2/6 = 0.33 Psağ=4/6 = 0.67

P(Evet|tsol)=0/2 = 0 P(Evet|tsağ)=2/4 =0.5

P(Hayır|tsol)=2/2 = 1 P(Hayır|tsağ)=2/4 =0.5

Φ(s|t)=2\*0.33\*0.67\*[|0-0.5|+|1-0.5|]=0.44

1. Önceliğimiz eğitim ve sol tarafta ilk olacak

Psol = 1/6 = 0.17 Psağ=5/6 = 0.83

P(Evet|tsol)=1/1 = 1 P(Evet|tsağ)=1/5 =0.2

P(Hayır|tsol)=0/1 = 0 P(Hayır|tsağ)=4/5 =0.8

Φ(s|t)=2\*0.17\*0.83\*[|1-0.2|+|0-0.8|]=0.44

1. Önceliğimiz eğitim ve sol tarafta orta olacak

Psol = 3/6 = 0.5 Psağ=3/6 = 0.5

P(Evet|tsol)=1/3 = 0.33 P(Evet|tsağ)=1/3 =0.33

P(Hayır|tsol)=2/3 = 0.67 P(Hayır|tsağ)=2/3 =0.67

Φ(s|t)=2\*0.5\*0.5\*[|0.33-0.33|+|0.67-0.67|]=0.00

1. Önceliğimiz eğitim ve sol tarafta lise olacak

Psol = 2/6 = 0.33 Psağ=4/6 = 0.67

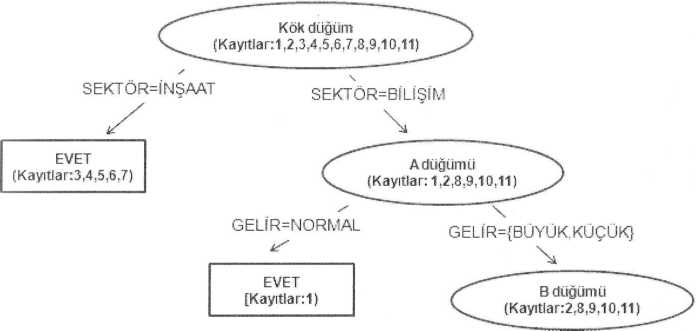
P(Evet|tsol)=0/2 = 0 P(Evet|tsağ)=2/4 =0.5

P(Hayır|tsol)=2/2 = 1 P(Hayır|tsağ)=2/4 =0.5

Φ(s|t)=2\*0.33\*0.67\*[|0-0.5|+|1-0.5|]=0.44



En büyük sayılarımız eğitimden lise ve ilk , gelir den küçük ve normal oldu. İstediğimizi seçebiliriz biz gelir=normal den devam edeceğiz.



Ağacımız bu hale geldi.

1. Önceliğimiz gelir ve sol tarafta büyük olacak

Psol = 3/5 = 0.6 Psağ=2/5= 0.4

P(Evet|tsol)=1/3 = 0.33 P(Evet|tsağ)=0/2 =0

P(Hayır|tsol)=2/3 = 0.67 P(Hayır|tsağ)=2/2 =1

Φ(s|t)=2\*0.4\*0.6\*[|0.33-0|+|0.67-1|]=0.32

1. Önceliğimiz gelir ve sol tarafta küçük olacak

Psol = 2/5 = 0.4 Psağ=3/5 = 0.6

P(Evet|tsol)=0/2 = 0 P(Evet|tsağ)=1/3 =0.33

P(Hayır|tsol)=2/2 = 1 P(Hayır|tsağ)=2/3 =0.67

Φ(s|t)=2\*0.4\*0.6\*[|0-0.33|+|1-0.67|]=0.32

1. Önceliğimiz eğitim ve sol tarafta ilk olacak

Psol = 1/5 = 0.2 Psağ=4/5 = 0.8

P(Evet|tsol)=1/1 = 1 P(Evet|tsağ)=0/4 =0

P(Hayır|tsol)=0/1 = 0 P(Hayır|tsağ)=4/4 =1

Φ(s|t)=2\*0.2\*0.8\*[|1-0|+|0-1|]=0.64

1. Önceliğimiz eğitim ve sol tarafta orta olacak

Psol = 2/5 = 0.4 Psağ=3/5 = 0.6

P(Evet|tsol)=0/2 = 0 P(Evet|tsağ)=1/3 =0.33

P(Hayır|tsol)=2/2 = 1 P(Hayır|tsağ)=2/3 =0.67

Φ(s|t)=2\*0.4\*0.6\*[|0-0.33|+|1-0.67|]=0.32

1. Önceliğimiz eğitim ve sol tarafta lise olacak

Psol = 2/5 = 0.4 Psağ=3/5 = 0.6

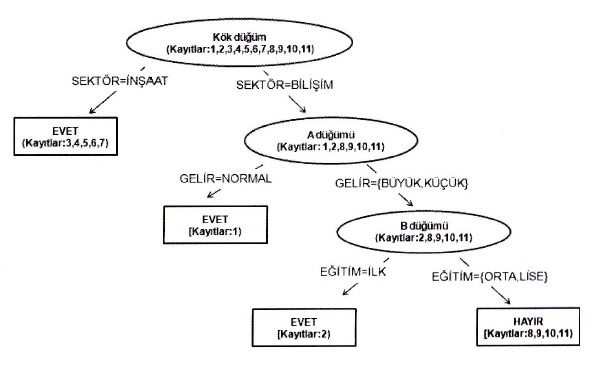
P(Evet|tsol)=0/2 = 0 P(Evet|tsağ)=1/3 =0.33

P(Hayır|tsol)=2/2 = 1 P(Hayır|tsağ)=2/3 =0.67

Φ(s|t)=2\*0.4\*0.6\*[|0-0.33|+|1-0.67|]=0.32



En büyüğümüz eğitim ilk olduğundan ona göre ağacı düzenliyoruz.

Kural Tablosu: Elde edilen karar ağacına uygun olarak aşağıdaki gibi düzenlenebilir.

KURAL 1:

Eğer SEKTÖR=İNŞAAT ise MEMNUN=EVET;

KURAL 2:

Eğer SEKTÖR=BİLİŞİM ise ve

Eğer GELİR=NORMAL ise MEMNUN=EVET;

KURAL 3:

Eğer SEKTÖR=BİLİŞİM ise ve Eğer GELİR=BÜYÜK veya GELİR=KÜÇÜK ise ve Eğer EGİTİM=İLK ise MEMNUN=EVET;

KURAL 4:

Eğer SEKTÖR=BİLİŞİM ise ve

Eğer GEI.İR=BÜYÜK veya GELİR=KÜÇÜK ise ve

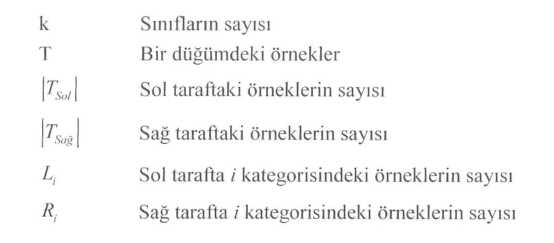
Eğer EGİTİM=ORTA veya EĞİTİM=LİSE ise MEMNUN=HAYIR;

GİNİ ALGORİTMASI

1. Her nitelik değerleri ikili olacak biçimde gruplanır. Bu şekilde elde edilen sol ve sağ bölünmelere karşılık gelen sınıf değerleri gruplandırılır
2. Her bir nitelikle ilgili sol ve sağ taraftaki bölünmeler için Ginisol ve Ginisağ değerleri hesaplanır.

GİNİsol=1-

GİNİsağ=1-



1. Her j niteliği için, n eğitim kümesindeki satır sayısı olmak üzere aşağıdaki bağıntının değeri hesaplanır:

GİNİj= (1/n)(|Tsol|Ginisol+|Tsağ|+Ginisağ)

1. Her j niteliği için hesaplanan *Ginij* değerleri arasından en küçük olanı seçilir ve bölünme bu nitelik üzerinden gerçekleştirilir.
2. En baştaki (a) adımına dönülerek işlemlere devam edilir.

Örnek: BORÇ, GELİR, STATÜ ve RİSK isimli dört nitelikten ve 8 gözlemden oluşan bir eğitim kümesinin, *Gini* algoritmasını uygulamak üzere aşağıda gösterildiği biçimdi gruplandırıldığını varsayalım. RİSK hedef niteliğinin değerleri İYİ ve KÖTÜ olarak belirlenmiştir.



Bu bilgileri kullanarak *GiniSol* ve *Ginisağ* ile *Giniborç* değerlerini hesaplarsak;

Sınıfı niteliği 2 değer içerdiğinden k=2 oalrak kabul edilir.

Borç;

Tsol=2+3=5 L1=2 ve L2=3 olarak belirlenmiştir.

Tsağ=1+2=3 R1=1 ve R2=2 olarak belirlenmiştir. O halde;

Ginisol=1-[(2/5)^2+(3/5)^2]=0.48

Ginisağ=1-[(1/3)^2+(2/3)^2]=0.44

Ginij=[(5\*048)+(3\*0.44)]/8=0.46

Örnek:

İlk olarak nitelik değerlerinin ikili gruplandırılması uygulanır;



Eğitim için;

Tsol=1+2=3 L1=1 ve L2=2 olarak belirlenmiştir.

Tsağ=3+1=4 R1=3 ve R2=1 olarak belirlenmiştir. O halde;

Ginisol=1-[(1/3)^2+(2/3)^2]=0.444

Ginisağ=1-[(3/4)^2+(1/4)^2]=0.375

Ginij=[3\*0.44)+(4\*0.375)]/7=0.405

Yaş için;

Tsol=0+2=2 L1=0 ve L2=2 olarak belirlenmiştir.

Tsağ=4+1=5 R1=1 ve R2=4 olarak belirlenmiştir. O halde;

Ginisol=1-[(0/2)^2+(2/2)^2]=0

Ginisağ=1-[(4/5)^2+(1/5)^2]=0.320

Ginij=[(2\*0)+(5\*0.32)]/7=0.229

Cinsiyet için;

Tsol=1+2=3 L1=2 ve L2=1 olarak belirlenmiştir.

Tsağ=3+1=4 R1=1 ve R2=3 olarak belirlenmiştir. O halde;

Ginisol=1-[(1/3)^2+(2/3)^2]=0.444

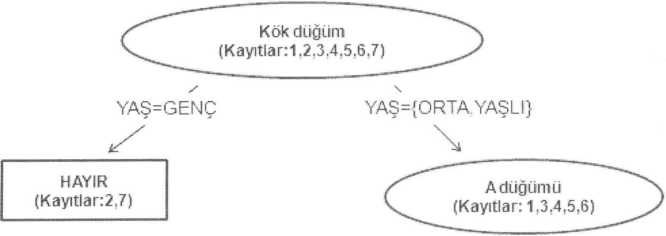
Ginisağ=1-[(3/4)^2+(1/4)^2]=0.375

Ginij=[(3\*0.44)+(4\*0.375)]/7=0.405

Sonuç olarak;



En küçük düğüm yaş olduğundan dolayı kök düğüm yaş olur.



Genç sonuçları hayıra gittiğinden dolayı genç silinerek aynı işlemler uygulanır.





Eğitim için;

Tsol=1+0=1 L1=1 ve L2=0 olarak belirlenmiştir.

Tsağ=3+1=4 R1=3 ve R2=1 olarak belirlenmiştir. O halde;

Ginisol=1-[(1/1)^2+(0/1)^2]=0

Ginisağ=1-[(3/4)^2+(1/4)^2]=0.375

Ginij=[(1\*0)+(4\*0.375)]/5=0.3

Yaş için;

Tsol=1+2=3 L1=1ve L2=2 olarak belirlenmiştir.

Tsağ=2+0=2 R1=2 ve R2=0 olarak belirlenmiştir. O halde;

Ginisol=1-[(1/3)^2+(2/3)^2]=0.444

Ginisağ=1-[(2/2)^2+(0/2)^2]=0

Ginij=[(3\*0.444)+(2\*0)]/5=0.266

Cinsiyet için;

Tsol=1+1=2 L1=1 ve L2=1 olarak belirlenmiştir.

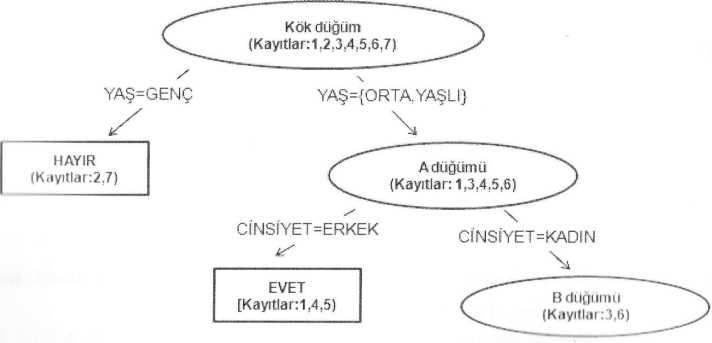
Tsağ=3+0=3 R1=3 ve R2=0 olarak belirlenmiştir. O halde;

Ginisol=1-[(1/2)^2+(1/2)^2]=0.5

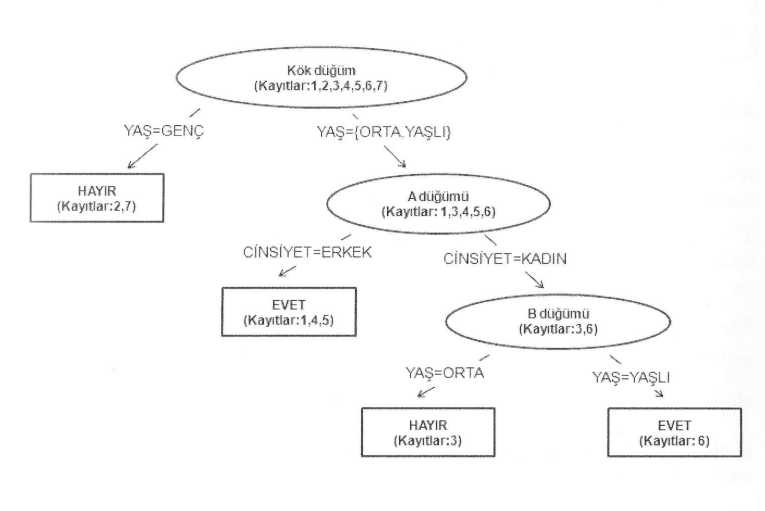
Ginisağ=1-[(3/3)^2+(0/3)^2]=0

Ginij=[(2\*0.5)+(3\*0)]/5=0.2

Sonuç olarak;



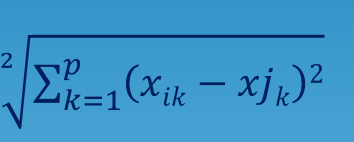




**7-8.Hafta**

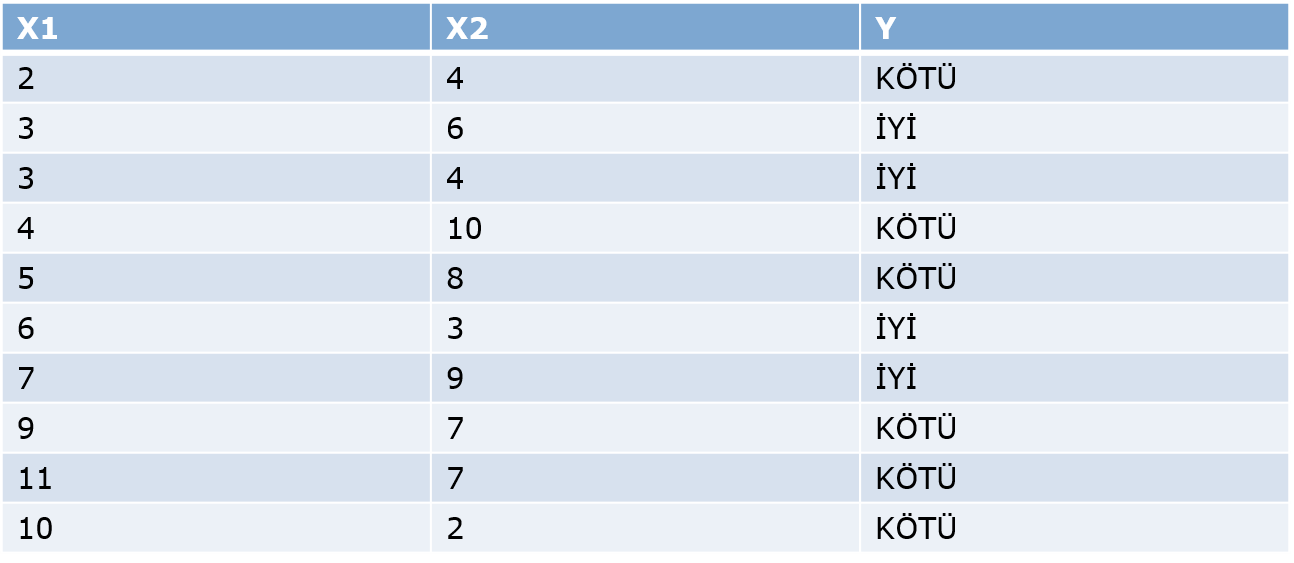
KNN K En Yakın Komşu Algoritması

En yakın k-komşu Algoritması ( k – komşu )

D ( i , j ) = 

Örnek:

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.



Adım1: 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.

Adım2: Uzaklıkların hesaplanması: (8,4) noktası ile gözlem değerinin her birisi arasındaki uzaklıkları hesaplamamız gerekiyor. Uzaklık bağıntısı olarak Öklid uzaklık formülünü kullanıyoruz.

Adım3: biçiminde olduğuna göre birinci gözlem olan (2,4) noktası ile (8,4) noktası arasındaki uzaklık şu şekilde hesaplanır :

D(i,j)=√((2−8)2+(4−4)2) =6.00

D(i,j)=√((3−8)2+(6−4)2) =5.39

D(i,j)=√((3−8)2+(4−4)2) =5.00

D(i,j)=√((4−8)2+(10−4)2) =7.21

D(i,j)=√((5−8)2+(8−4)2) =5.00

D(i,j)=√((6−8)2+(3−4)2) =2.24

D(i,j)=√((7−8)2+(9−4)2) =5.10

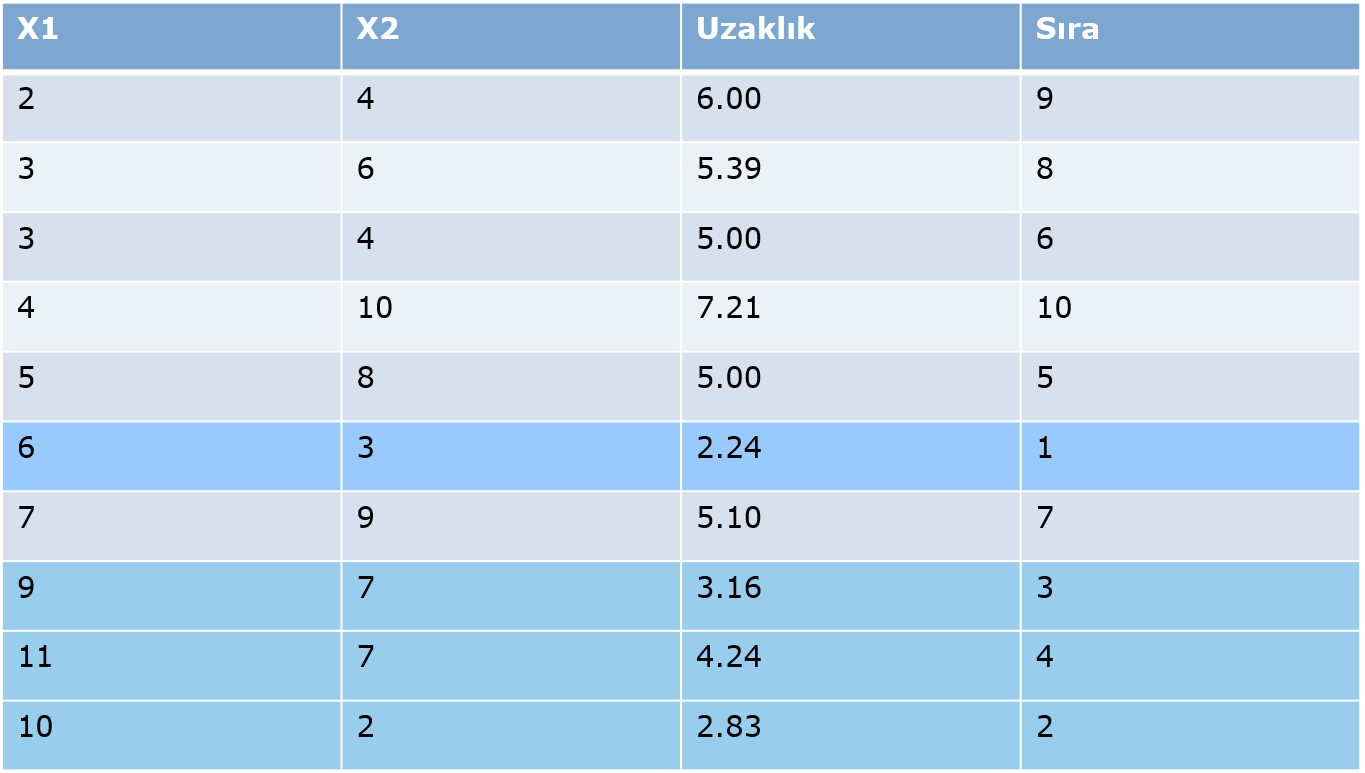
D(i,j)=√((9−8)2+(7−4)2) =3.16

D(i,j)=√((11−8)2+(7−4)2) =4.24

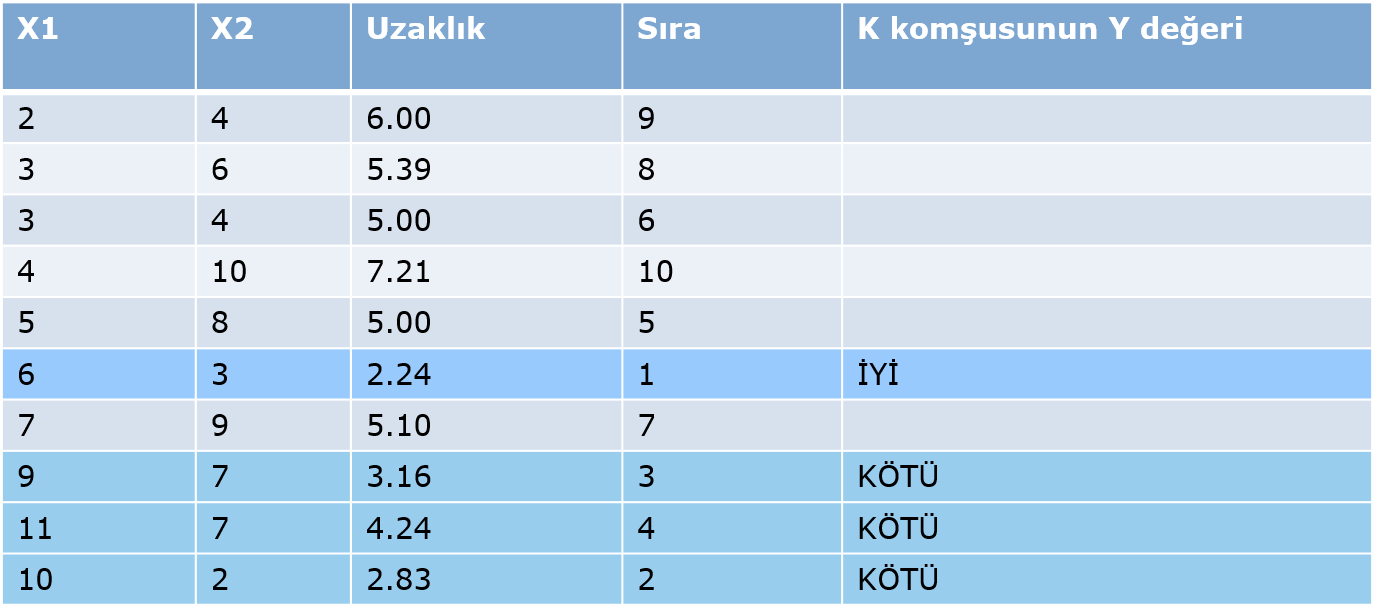
D(i,j)=√((10−8)2+(2−4)2) =2.83

Hesaplanan değerler tablo üzerine yerleştirilir.

Adım4: ) En küçük uzaklıkların belirlenmesi: Satırlar sıralanarak, en küçük k=4 tanesi belirleniyor. Bu dört nokta verilen (8,4) noktasına en yakın gözlem değerleridir.

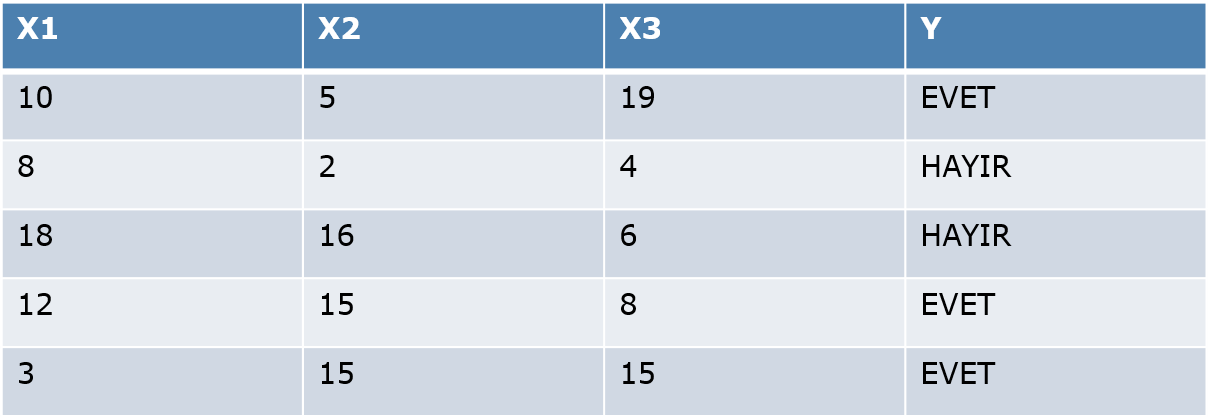


Adım5: Seçilen satırlara ilişkin sınıfların belirlenmesi:(8,4) noktasına en yakın olan gözlem değerlerinin Y sınıfları göz önüne alınır ve içine hangi değerin baskın olduğu araştırılır. Bu dört sonuç içinde bir tane İYİ, dört tane KÖTÜ sonucu vardır.



Adım 6: Yeni gözlemin sınıfı: KÖTÜ değerlerinin sayısı İYİ değerinin sayısından fazla sayıda olduğu için (8,4) noktasının sınıfı KÖTÜ olarak belirlenmiştir.

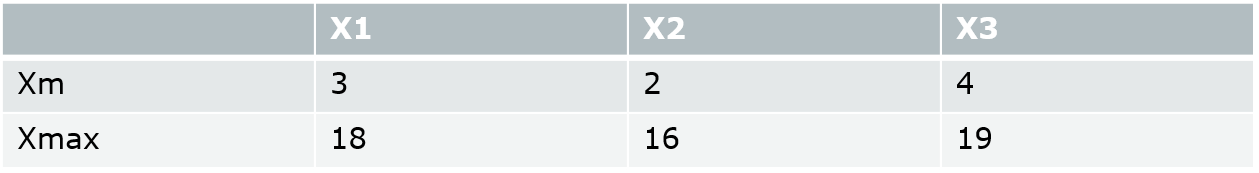
Örnek: Aşağıda verilen gözlem tablosunu göz önüne alalım. Gözlemler üç değişkenlidir. Y ise sınıf niteliğini ifade etmektedir. Bu verilere dayanarak (7,8,5) noktasının hangi sınıf değerine sahip olduğunu belirlemek istiyoruz. Ancak bu uygulamayı gerçek gözlem değerleriyle değil dönüştürülmüş değerlerle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



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

**X\*=(X-Xmin)/(Xmax-Xmin)**

Bağıntısı uygulanır. Söz konusu bağıntıyı kitabımızın ikinci bölümünde ele alınarak incelemiştik. Burada X\* dönüştürülmüş değerleri, X gözlem değerlerini, Xmin en küçük değerini ve Xmax en büyük gözlem değerini ifade etmektedir. Bu değerler aşağıdaki tablo üzerinde yer almaktadır.



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

**X\*=(X-Xmin)/(Xmax-Xmin)**

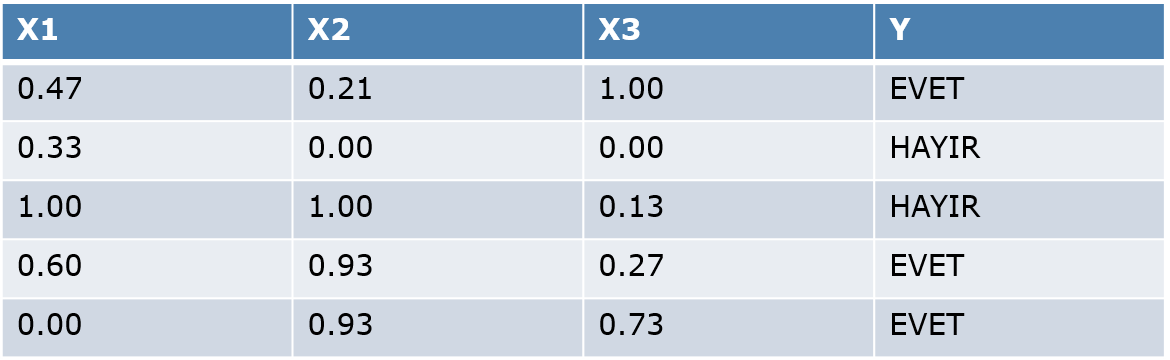
=**(10-3)/(18-3)**=0,47

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

=**(5-2)/(16-2)**=0,21

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

=**(19-4)/(19-4)**=1



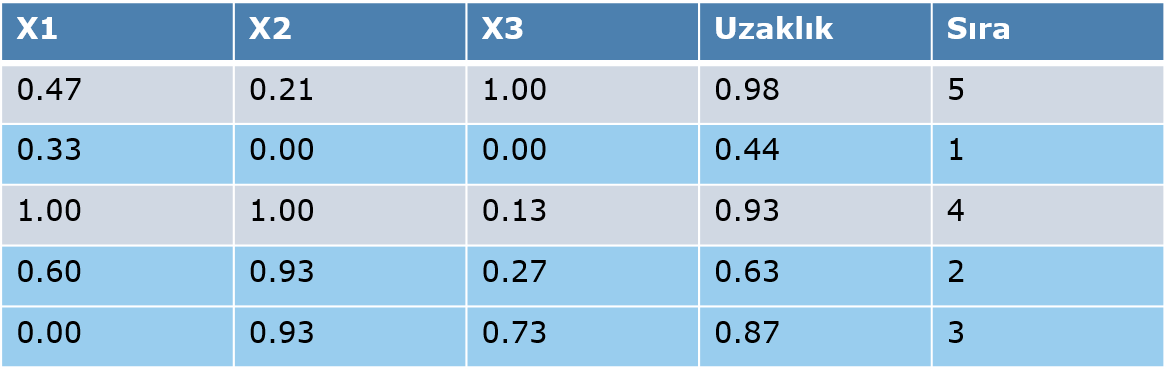
Bu durumda sınıflandırmaya tabi tutulacak (7,8,5) gözlemi de aynı dönüşüm formülüyle yeni değerlere dönüştürülür. Bununla ilgili yeni gözlem noktası (0.26, 0.43, 0.07) biçiminde elde edilir. Yeni gözlemler elde edildiğine göre artık k-en yakın komşu algoritmasını uygulayabiliriz.

Adım 1: K nın belirlenmesi: K-en yakın komşu algoritması için k=3 kabul ederek çözülmeye başlıyoruz.

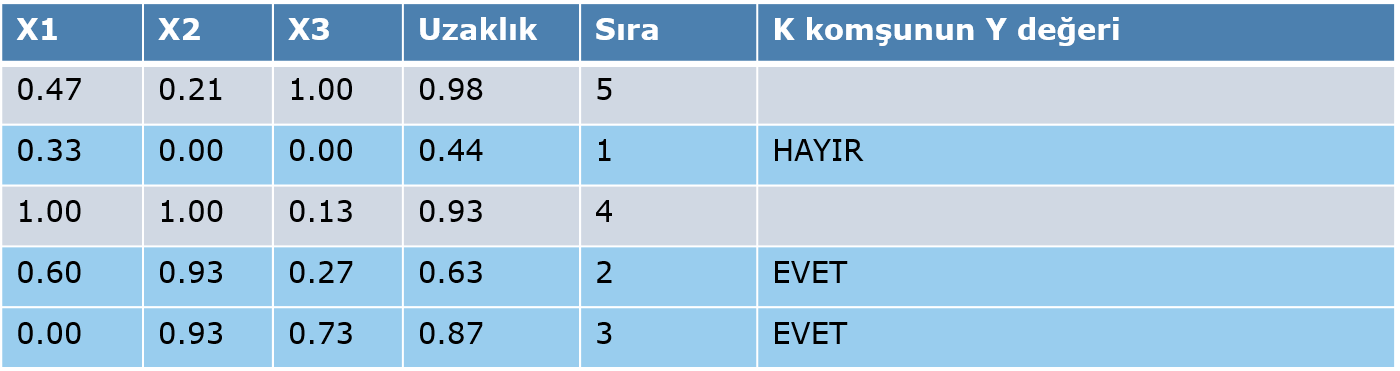
Adım 2: Uzaklıkların hesaplanması:(0.26, 0.43, 0.07) noktası ile dönüştürülmüş gözlem değerlerinin her birisi arasındaki Öklid uzaklıklarını hesaplandığında tablo8 elde edilir.



Adım 3 : 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.



Adım 4: 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.

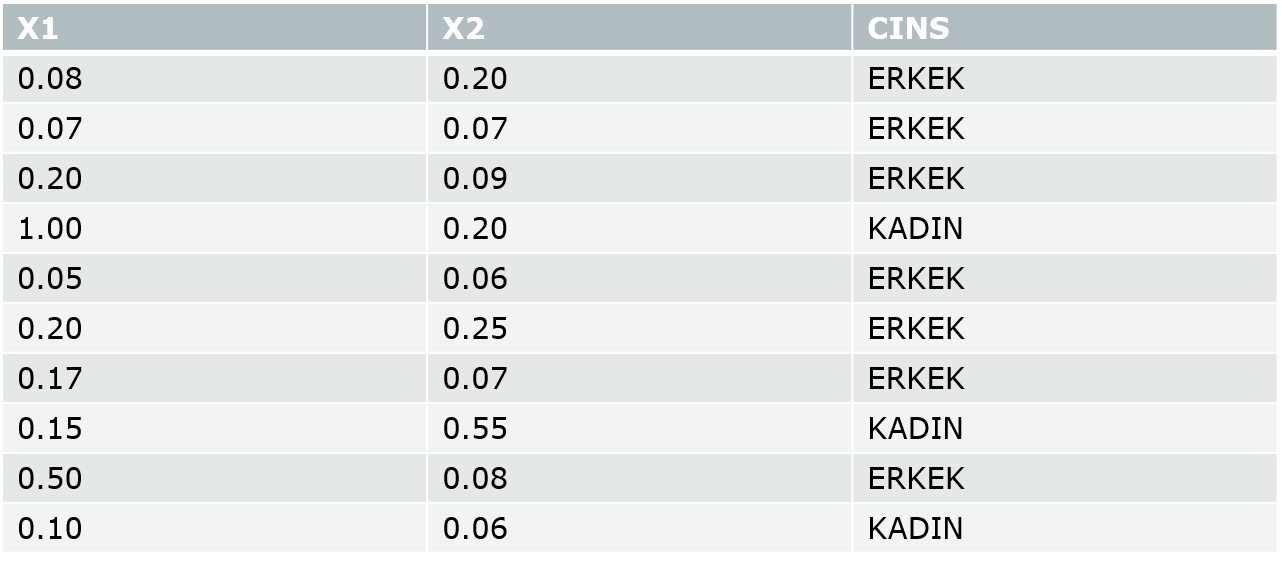


Adım 5: Yeni gözlemin sınıfı: Seçilenler arasında EVET’ lerin sayısı diğerinden daha fazladır. O halde (7,8,5) gözleminin, yani dönüştürülmüş değerlerle ifade edilir.(0.26, 0.43, 0.07) gözleminin de sınıfı EVET olarak kabul edilir.

**Ağırlıklı Oylama**

d(i,j)’=1/d(i,j)2

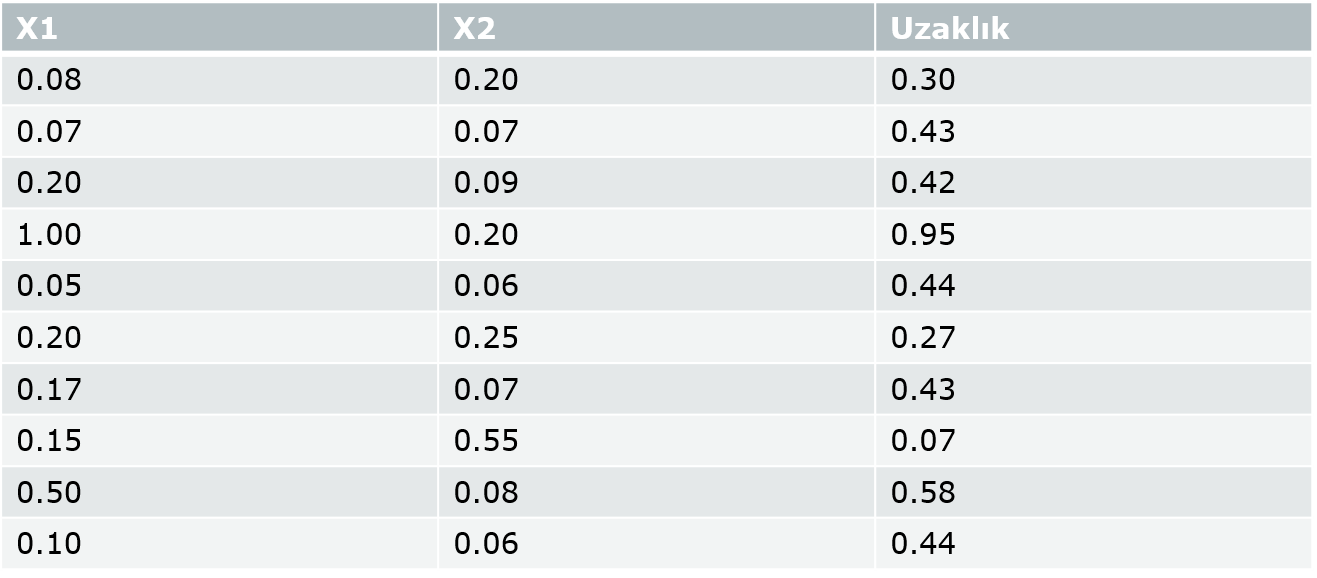
Örnek: Aşağıda verilen gözlem tablosunu göz önüne alalım. Bu gözlemler X1 ve X2 niteliklerinden ve CINS sınıfından oluşmaktadır. Bu gözlem değerlerine bağlı olarak, yeni bir gözlem olan (0.10, 0.50) gözleminin hangi sınıfa dahil olduğunu k-en yakın komşu yöntemiyle bulalım



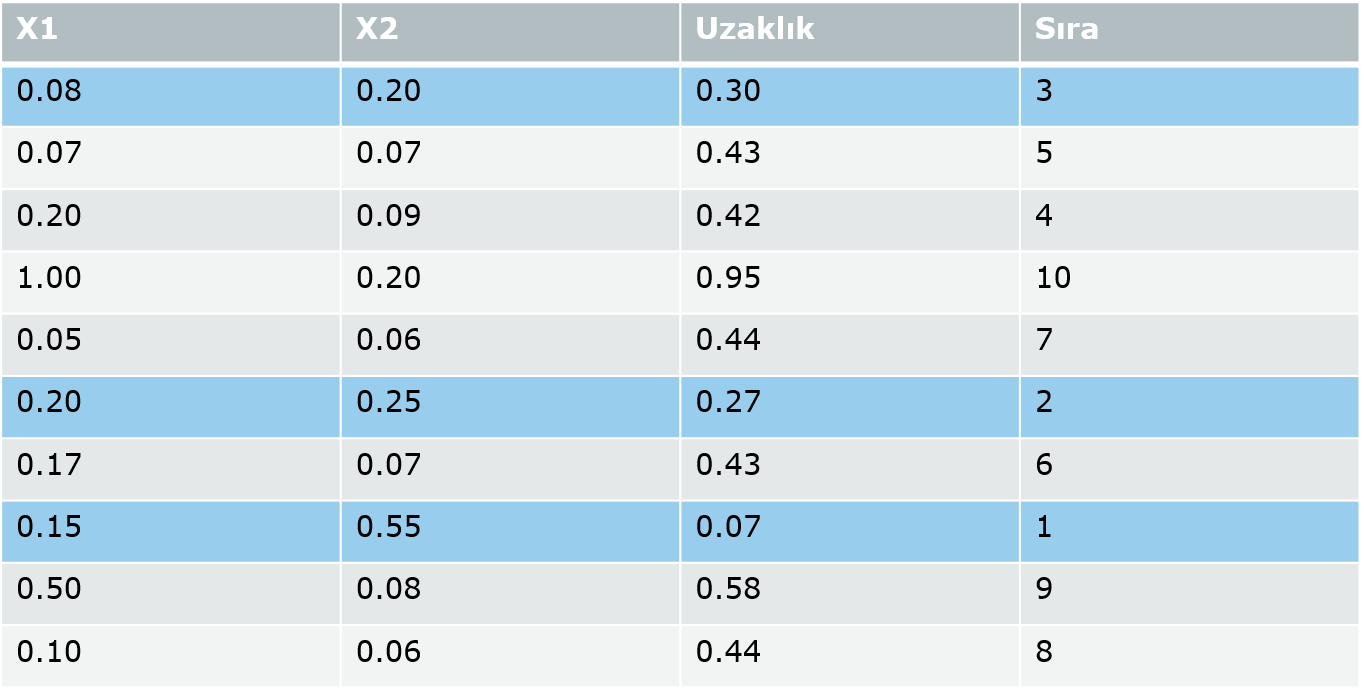
Adım1 : K nın belirlenmesi: Algoritmaya başlamadan önce, k-en yakın algoritması için k=3 olduğunu kabul ediyoruz. Böylece bu problem çerçevesinde verilen (0.10, 0.50) gözlemine en yakın 3 komşuyu arayacağız.

Adım 2: 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.

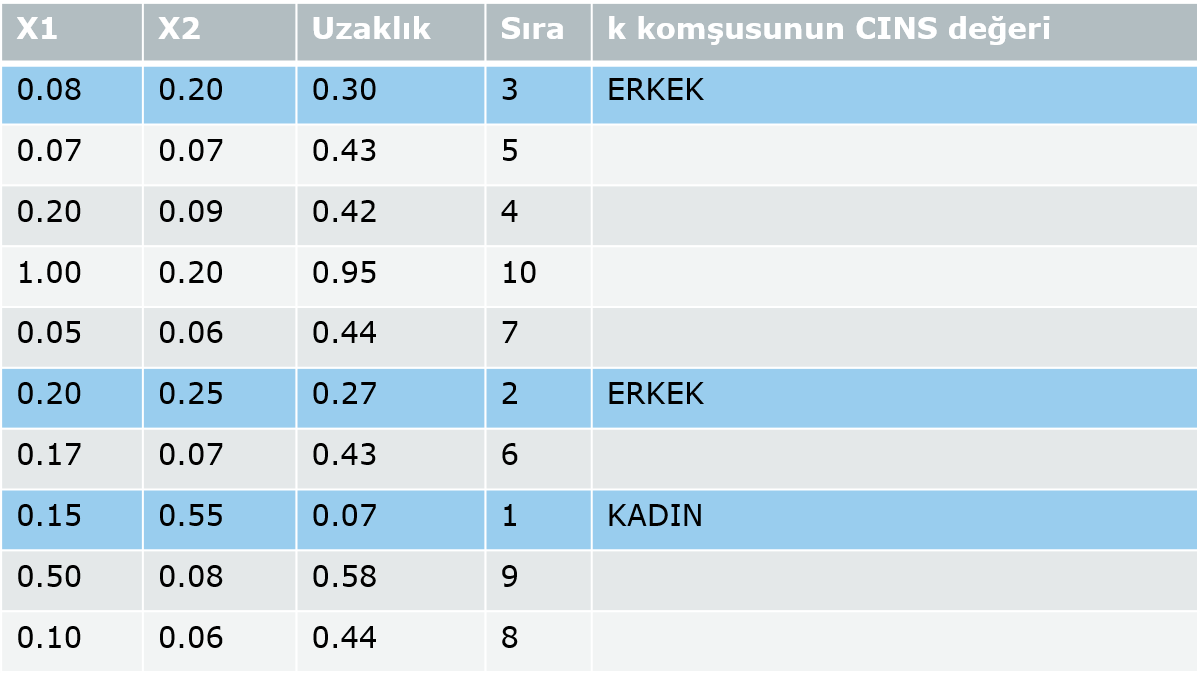
=((0.10-0.08)^2+(0.50-0.20)^2)^(1/2)=0.30



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



Adım 4: Seçilen satırlara ilişkin sınıfların belirlenmesi: Yeni gözlem noktasına en yakın olan gözlem değerlerinin CINS sınıfları göz önüne alınır ve içine hangi değerin baskın olduğu araştırılır. Bu üç sonuç içinde iki tane ERKEK, bir tane KADIN değeri vardır.

Adım 5: 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)’=𝟏/𝒅(𝒊,𝒋)𝟐**

Bağıntısı kullanılarak aşağıdaki hesaplamalar yapılır. Birinci satır için

d(1,yeni gözlem)’=𝟏/(𝟎.𝟑𝟎^𝟐)=11.046

Diğer iki gözlem için benzer hesaplamalar yapılır.

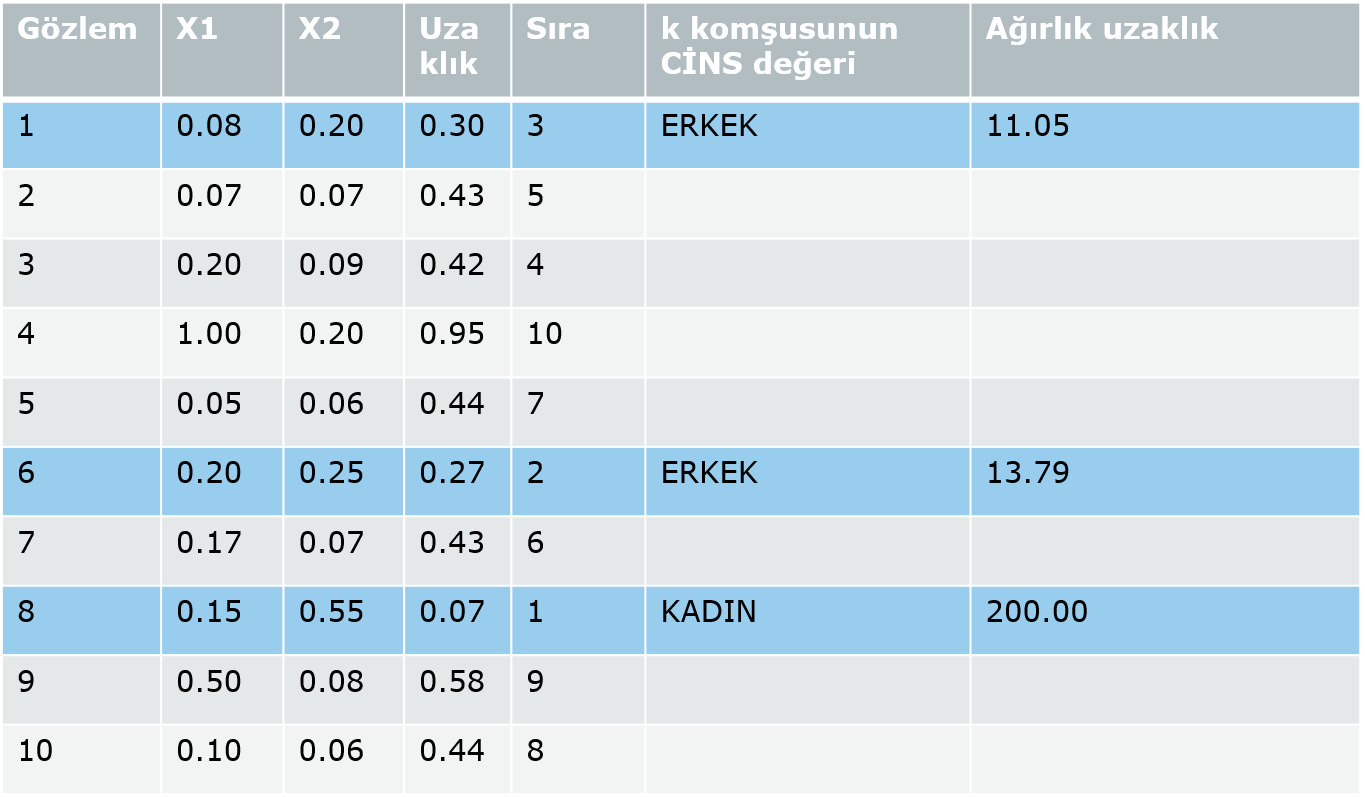
d(6,yeni gözlem)’=𝟏/(𝟎.𝟐𝟕^𝟐)=13.793

d(1,yeni gözlem)’=𝟏/(𝟎.𝟎𝟕^𝟐 )=200

Hesaplanan değerler tablo15 ‘tedir.

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

d(1,yeni gözlem)’+d(6,yeni gözlem)’=11.046+13.793=24.84



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 kümelerde işlemi yavaşlatır.

Sınıflandırma

TEST

2 Parçaya Bölme

Percentage Split ( Fixed or Holdout ) yönteminde veri, eğitim ve test verisi olacak şekilde ikiye ayrılır.

Örnek: Eğitim:%80 Test:%20

Çapraz Doğrulama Testi

Çapraz Doğrulama, k-kat çapraz doğrulama ( k-fold cross validation ) ve birini dışarıda bırak çapraz doğrulama (leave one out cross validation – LOOCV ) olmak üzere iki farklı şekilde kullanılmaktadır.

Model Performansını Değerlendirme Metrikleri

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Tahmin Edilen Sınıf | | |
| Sınıf=Pozitif | Sınıf=Negatif | Toplam |
| Gerçek  Sınıf | Sınıf=Pozitif | DP | YN | P |
| Sınıf=Negatif | YP | DN | N |
| Toplam | Pi | Ni | P+N |

Sınıflandırma Metrikleri

Doğruluk = (DP+DN)/(DP+DN+YP+DN) Hata Oranı = (YP+YN)/(P+N)

Duyarlılık = DPO = DP/(DP+YN) Kesinlik = DP/(DP+YP)

Özgüllük = DNO = DN/(DN+YP) Anma = DN/(DN+YN)

Örnek:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Tahmin Edilen Sınıf | | |
| Erkek | Kadın | Toplam |
| Gerçek  Sınıf | Erkek | 8 | 3 | 11 |
| Kadın | 7 | 2 | 9 |
| Toplam | 15 | 5 | 20 |

Doğruluk =0.5 Hata Oranı=0.5

Duyarlık=0.73 Kesinlik=0.53

Özgüllük=0.22 Anma=0.4

**9-10.Hafta**

Kümeleme

Kümeleme Çözümlemesi

1. Uzaklık Ölçüleri

Öklit Uzaklığı

Manhattan Uzaklığı

Minkowski Uzaklığı

2. Hiyerarşik Kümeleme

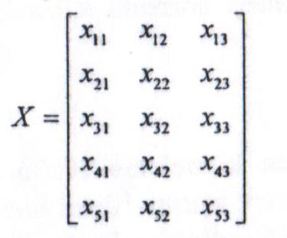
En Yakın Komşu Algoritması

En Uzak Komşu Algoritması

3. Hiyerarşik Olmayan Kümeleme

k-Ortalamalar Yöntemi

1. UZAKLIK ÖLÇÜLERİ

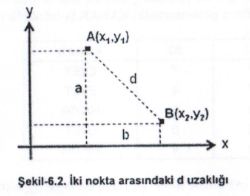


X11,x12,x13 1.gözlem noktasının konumu

X21,x22,x23 2.gözlem noktasının konumu

Bu iki nokta arasındaki uzaklık d(1,2) şeklinde ifade edilir

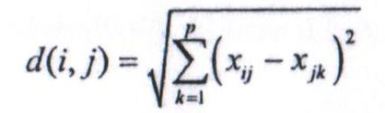
ÖKLİT UZAKLIĞI



A ve B noktası arasındaki Öklit uzaklığı aşağıdaki gibidir:

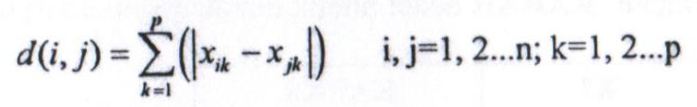
C:\Users\GEVAŞLI-065\Desktop\Ekran Alıntısı\73.JPG

Bu bağıntı genelleştirilecek olursa i ve j noktaları için şu şekilde bir bağıntıya ulaşılır:



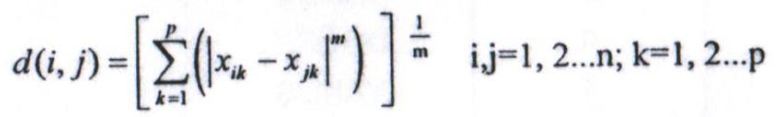
MANHATTAN UZAKLIĞI

Manhattan uzaklığı, gözlemler arasındaki mutlak uzaklıkların toplamı alınarak hesaplanır



MİNKOWSKİ UZAKLIĞI

p sayıda değişken göz önüne alınarak gözlem değerleri arasındaki uzaklığın hesaplanması söz konusu ise Minkowski uzaklık bağıntısı kullanılabilir.



Örnek : A,B ve C gibi üç değişkenden oluşan aşağıdaki gözlemleri göz önüne alalım.Bu gözlem noktalarının her birinin birbirine olan uzaklığını farklı uzaklık ölçüleriyle elde etmek istiyoruz.

|  |  |  |  |
| --- | --- | --- | --- |
| Gözlem | A | B | C |
| 1 | 2 | 3 | 1 |
| 2 | 4 | 1 | 3 |
| 3 | 5 | 7 | 3 |
| 4 | 4 | 8 | 2 |
| 5 | 3 | 9 | 5 |

Öklit uzaklığı :

Karekök içinde

d (2,1)=**√**(4-2)2+(1-3)2+(3-1)2=3.46

Manhattan uzaklığı:

d (2,1)= |(4-2)|+|(1-3)|+|(3-1)|=6

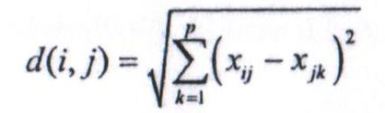
Minkowski uzaklığı:

M=3 olduğu için üsler 3 ve sonuncusu 1/3

d (2,1)= (|(4-2)3|+|(1-3)3|+|(3-1)3|)1/3=2.88

2 Hiyerarşik Kümeleme

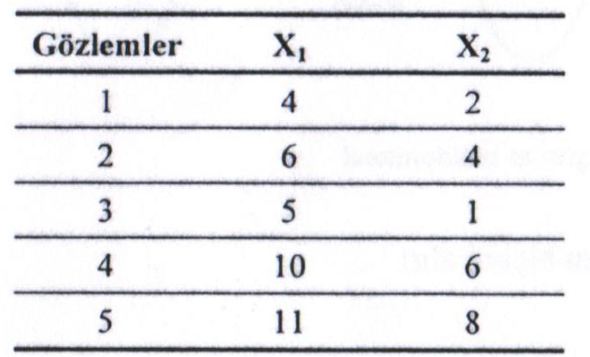
En Yakın Komşu Algoritması

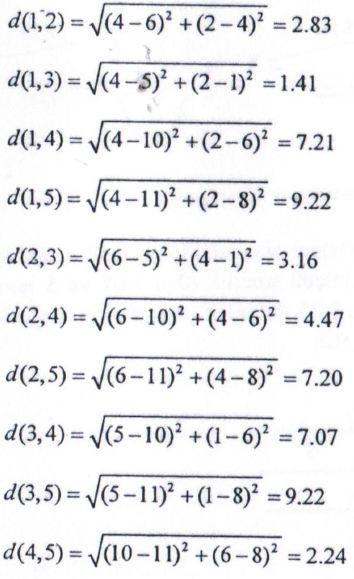


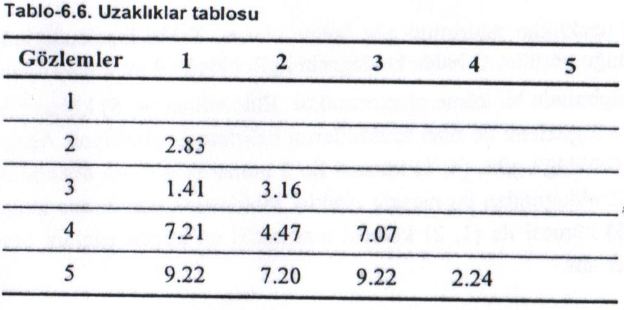
Örnek : Aşağıdaki tabloda verilen 5 adet gözlemi göz önüne alalım.Bu veriler üzerinde

en yakın komşu algoritmasını kullanarak kümeleme işlemlerini yapmak

istiyoruz.

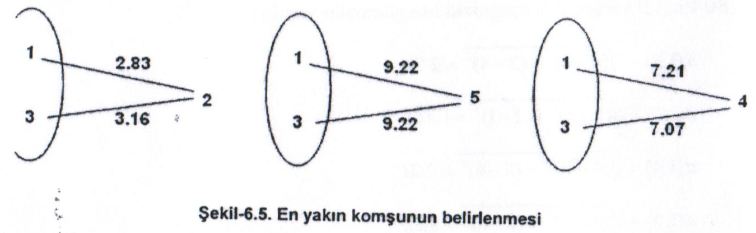


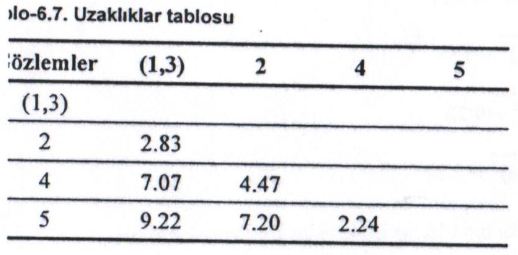




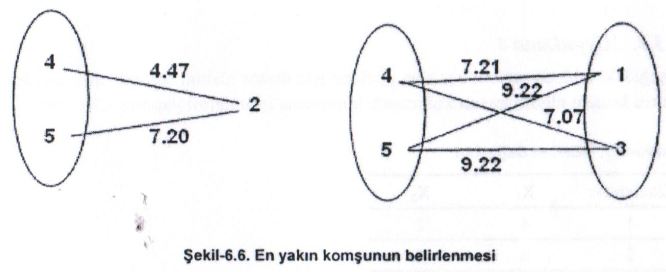
1 ve 3 birbirine en yakın olduğundan dolayı onları birleştiriyoruz.

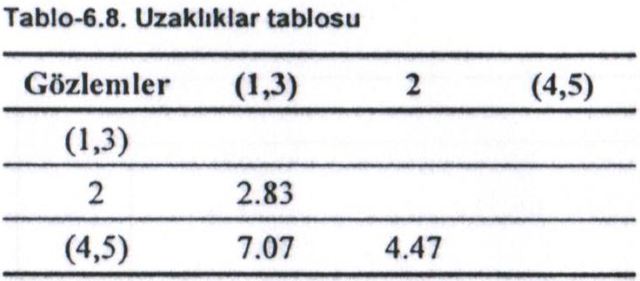
Daha sonra d(1,2) ve d(2,3) olan uzaklığına bakıp hangisi daha yakın ise onu d((1,3),2) kabul ederiz. Ve bunu 4 ve 5 içinde yaparız.



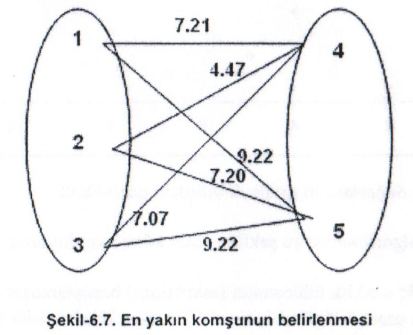


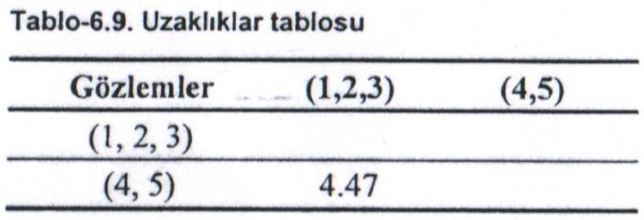
Daha sonra en küçük 2.24 olduğundan 4 ve 5 i de birleştiririz.



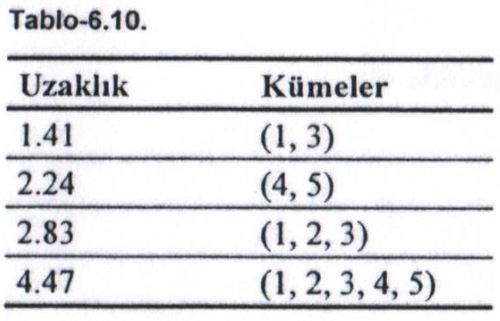


En küçük 2.83 olduğundan 1, 2 ve 3 birleştirilir.





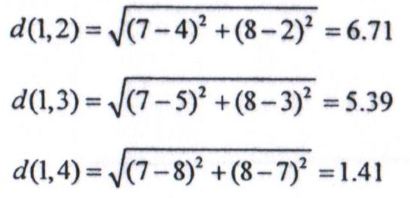
Elde edilen iki küme birleştirilerek sonuç küme bulunur.Bu küme (1,2,3,4,5) gözlemlerinden oluşan kümedir.Uzaklık düzeyi göz önüne alınarak kümeler şu şekilde belirlenmiştir.

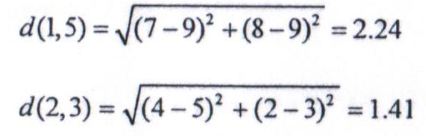
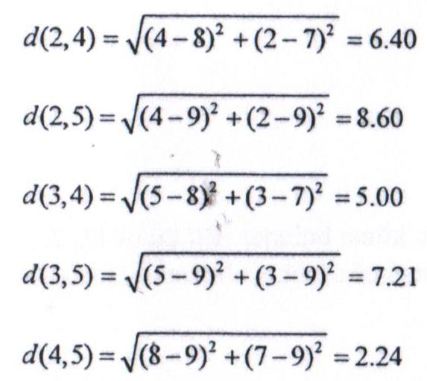


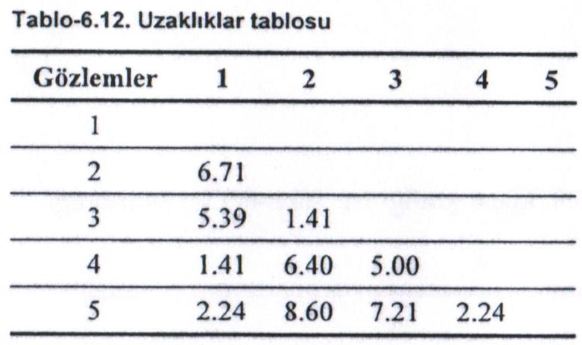
En Uzak Komşu Algoritması

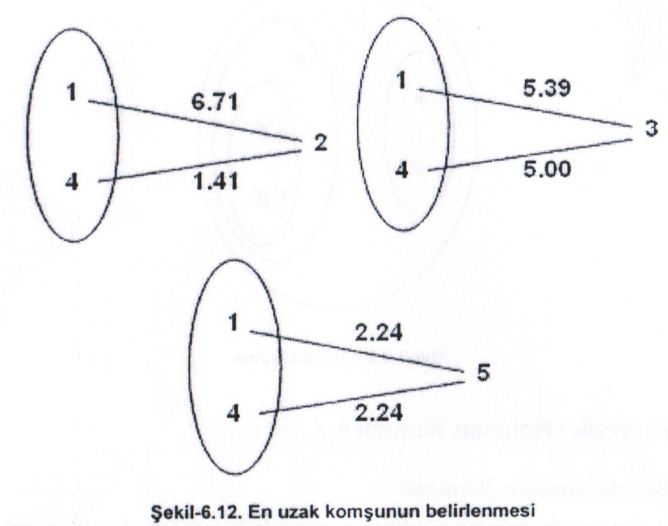
Yöntem yakın komşu algoritmasına çok benzer;ancak bu kez kümeler arasındaki uzaklık belirlenirken iki kümenin **birbirine en uzak elemanları** arasındaki mesafe küme arasındaki uzunluk olarak tayin edilir

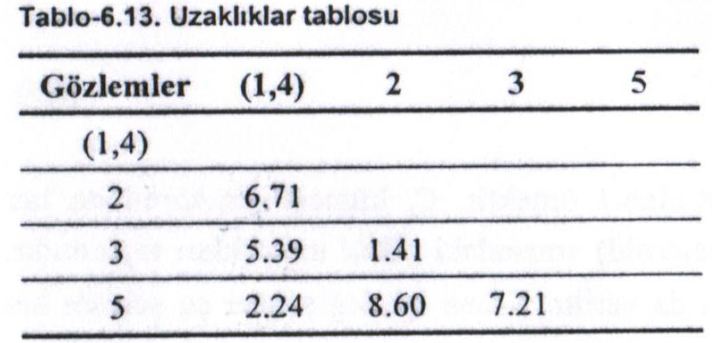
Örnek : Aşağıdaki gözlem değerlerini göz önüne alalım.Bu kez kümeleme analizini “en uzak komşu algoritmasına” göre yapacağız.

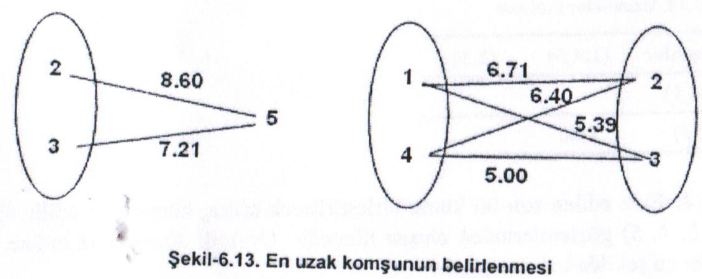


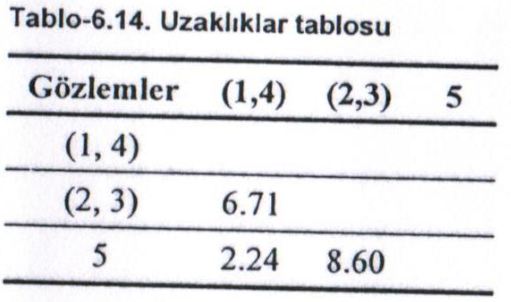


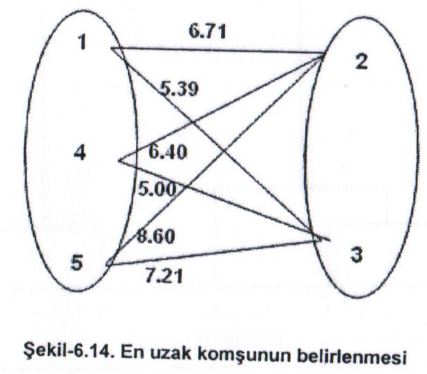


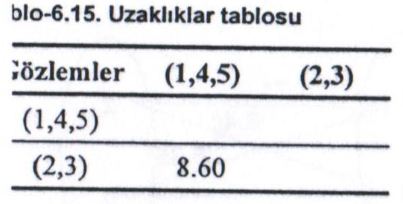




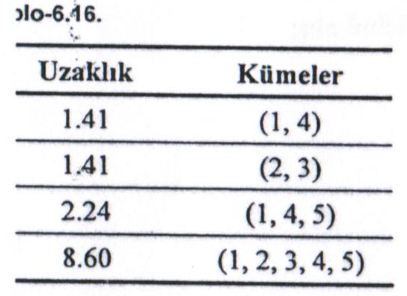








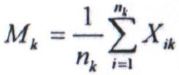
Elde edilen son iki küme birleştirilerek sonuç küme elde edilir.Bu küme (1,2,3,4,5) gözlemlerinden oluşan kümedir.Uzaklık düzeyi göz önüne alınarak kümeler şu şekilde belirlenmiştir:



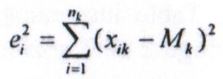
3 Hiyerarşik Olmayan Kümeleme

k-Ortalamalar Yöntemi

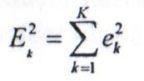
Bu yöntemde daha başlangıçta belli sayıdaki küme için toplam ortalama hatayı minimize etmek amaçlanır.N boyutlu uzayda N örnekli kümelerin verildiğini varsayalım.Bu uzay { C1,C2,…..,Ck} biçiminde K kümeye ayrılsın. O zaman C:\Users\GEVAŞLI-065\Desktop\Ekran Alıntısı\51.JPGolmak üzere Ck kümesinin ortalama vektörü Mk şu şekilde hesaplanır.



Burada Xk değeri Ck kümesine ait i. örnektir.Ck kümesi için kare-hata, her bir Ck değeri ile onun merkezi arasındaki Öklit uzaklıkların toplamıdır.Bu hataya **“küme içi değişme”** adı da verilir.Küme içi değişmeler şu şekilde hesaplanır:



K kümesini içeren bütün kümeler uzayı için kare-hata,küme içindeki değişmelerin toplamıdır.O halde söz konusu kare-hata değeri şu şekilde hesaplanır

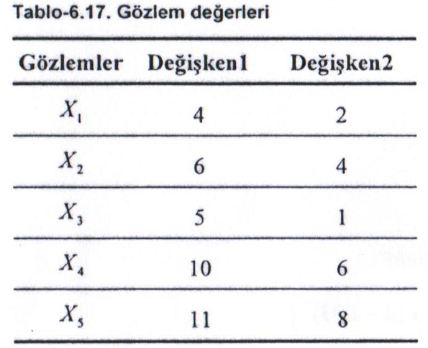


Kare-hata kümeleme yönteminin amacı, verilen K değeri için Ek2 değerini minimize eden K kümelerini bulmaktır.O halde k-ortalama algoritmasında Ek2 değerinin bir önceki iterasyona göre azalması gerekir.

Algoritma

* K-ortalama algoritmasına başlamadan önce k küme sayısının belirlenmesi gerekir. Söz konusu k değeri belirlendikten sonra her bir kümeye gözlem değerleri atanır ve böylece {C1,C2,….,Ck} kümeleri belirlenmiş olur.Ardından aşağıdaki işlemler gerçekleştirilir:
* Her bir kümenin merkezi belirlenir.Bu merkezler M1,M2,..,Mk biçimindedir.
* e1,e2,…,ek küme içi değişmeler hesaplanır.Bu değişmelerin toplamı olan Ek2 değeri bulunur.
* Mk merkez değerleri ile gözlem değerleri arasındaki uzaklıklar hesaplanır.Bu gözlem değeri hangi merkeze yakın ise, o merkez ile ilgili küme içine dahil edilir.
* Yukarıdaki b ve c adımları, kümelerde herhangi bir değişiklik olmayıncaya dek sürdürülür.

Örnek : Aşağıdaki gözlem değerlerini göz önüne alalım.Bu gözlem değerlerine k-ortalamalar yöntemini uygulayarak kümelemek istiyoruz.

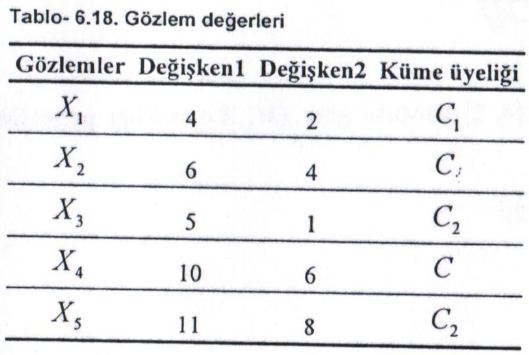


Kümelerin sayısına başlangıçta k=2 biçiminde karar veriyoruz.Başlangıçta tesadüfi olarak aşağıdaki iki kümeyi belirliyoruz.

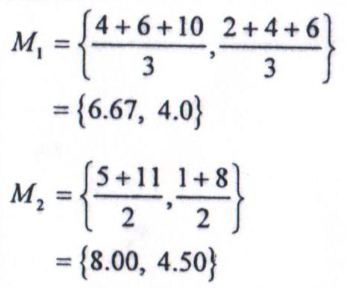
C1={X1,X2,X4}

C2={X3,X5}

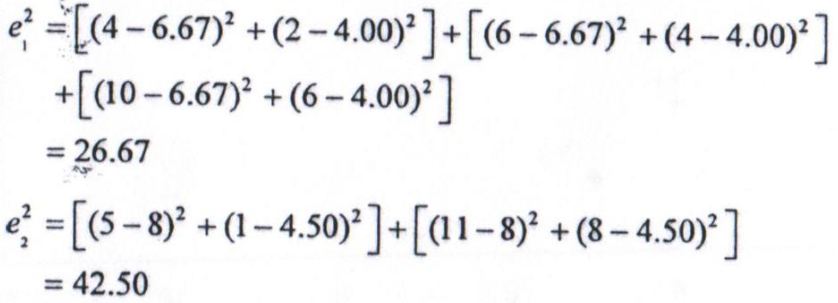
Bu kümeleri de içeren gözlem değerlerini aşağıdaki tablo üzerinde topluca gösteriyoruz.



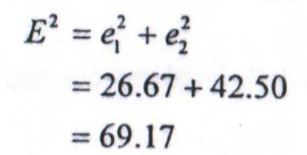
İki kümenin merkezleri şu şekilde hesaplanır:



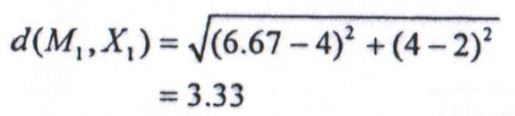
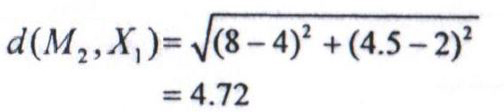
Küme içi değişmeler şu şekilde hesaplanır:



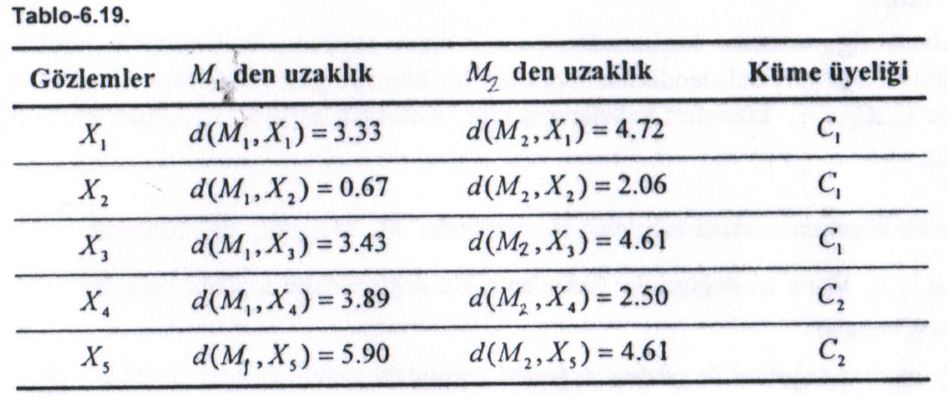
Toplam kare hata aşağıdaki gibi hesaplanır:



Gözlemlerin M1 ve M2 merkezlerinden olan uzaklıkların minimum olması istendiğinden aşağıdaki hesaplamalar yapılır:

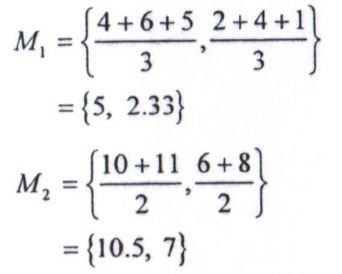
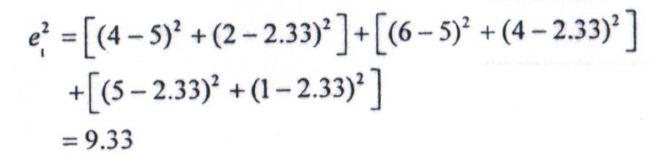


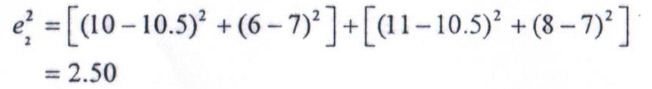
d(M1,X1) < d(M2,X1) olduğundan M1 merkezinin X1 gözlem değerine daha yakın olduğu anlaşılır.O halde X1  C1 olarak kabul edilir.



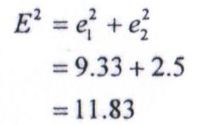
Yukarıdaki iki kümenin merkezleri Küme içi değişmeler şu şekilde

şu şekilde hesaplanır: hesaplanır:



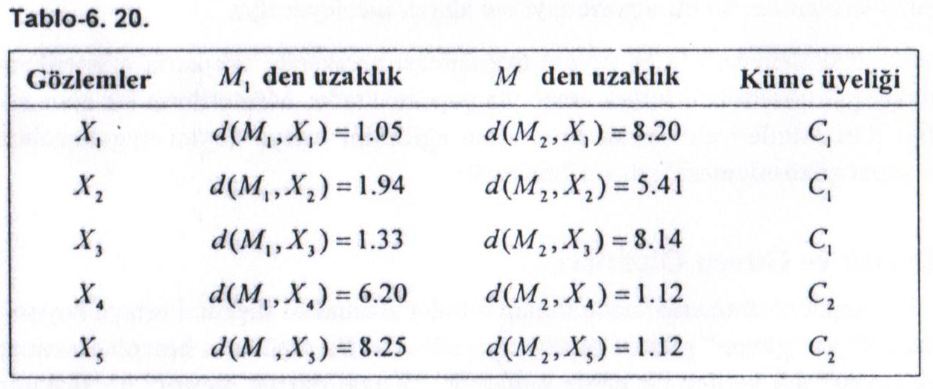


Toplam kare hata aşağıdaki gibi hesaplanır

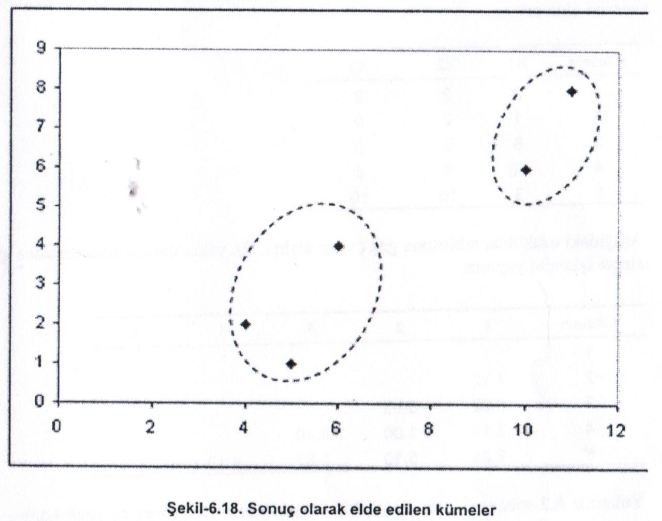


Bu değer bir önceki iterasyonda elde edilen E2=69.17 değerinden daha küçük olduğu anlaşılır

M1 ve M2 merkezinden gözlem değerlerine olan uzaklıklar hesaplandığında aşağıdaki tablo elde edilir



Kümelerde önceki adıma göre herhangi bir değişme olmadığına göre iterasyona burada son verilir.Elde edilen kümeler aşağıdaki şekilde gösterilmiştir.



**11-12.HAFTA**

**BİRLİKTELİK KURALI**

* Φ(D)=<X → Y,*c*,*s*> şeklinde ifade edilir.

**Destek ve Güven**

* Destek (Support) : X → Y=

X ve Y ürünlerini satın alan müşterilerin sayısı

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
| **İşlemler** | **Satın Alınan** **Ürünler** |
| 1 | Süt, Ekmek, Yumurta |
| 2 | Süt, Yumurta |
| 3 | Süt, Şeker |
| 4 | Ekmek, Yağ |

Toplam Müşteri Sayısı

* Güven (Confidence) : X → Y=

P(X ve Y) {X ve Y ürünlerini satın alanların sayısı}

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

P(X) {X ürününü satın alanların Sayısı}

|  |  |
| --- | --- |
| **Tekrarlanan** **Ürün** | **Destek Değeri** |
| Süt | % 75 |
| Ekmek | % 50 |
| Yumurta | % 50 |
| Şeker | % 25 |
| Yağ | % 25 |
| Süt,Yumurta | % 50 |

Örnek :

|  |  |  |
| --- | --- | --- |
| **Tekrarlanan Ürün** | **Destek Değeri** | **Güven Değeri** |
| Süt → Yumurta | % 50 | % 66 |
| Yumurta → Süt | % 50 | % 100 |

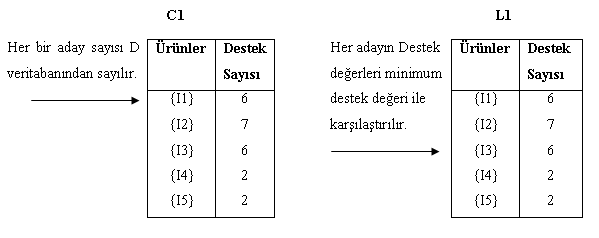
İşlemler, satın alınan ürünler, destek değerleri ve güven değerleri

Lift=0.66/0.5=1.32

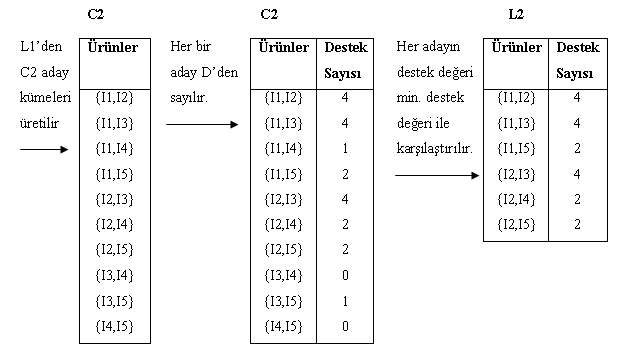
**Apriori Algoritması**

Örnek : Şekil 6’da bir firmada satın alınmış ürünlerle ilgili bir veritabanı görülmektedir. Bu veritabanını D olarak adlandıralım. Veritabanında 9 adet işlem görüldüğüne göre |D| =9 denilebilir. Bu durumda D veritabanı üzerinde apriori algoritması kullanılarak sık kullanılan nesnelerin nasıl bulunabileceğini aşağıdaki şekil den görebiliriz.

|  |  |
| --- | --- |
| **İşlemler** | **Satın Alınan** **Ürün Listesi** |
| T1 | I1, I2, I5 |
| T2 | I2, I4 |
| T3 | I2, I3 |
| T4 | I1, I2, I4 |
| T5 | I1, I3 |
| T6 | I2, I3 |
| T7 | I1, I3 |
| T8 | I1, I2, I3, I5 |
| T9 | I1, I2, I3 |

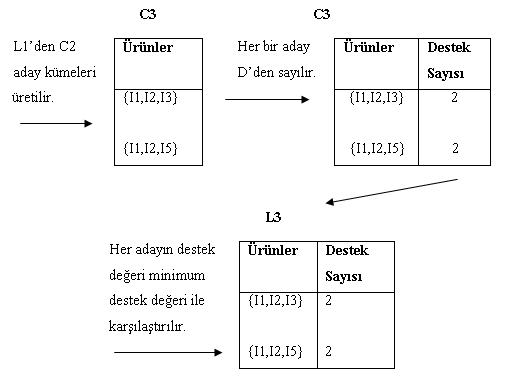


Kaç adet olduklarını ilk olarak C1 e yazarız min=2 kabul etmiş hoca L1 de 2 den küçük olanlar elenecek fakat 2 den küçük olmadığı için silinme işlemi yapılmamıştır bu aşamada.



Tüm kombinasyonlarımızı çıkarırız(C2), kaç adet olduğunu buluruz (C3), 2 den küçük olanlar silinir(L2)

Yeniden tüm kombinasyonları çıkarırız.



Birinci değer I1 iken I2,I3 ve I2,I5 ikilisi tabloda mevcut olduğu için bu üçlüyü alabiliyoruz. Oluşan tablo (C3) destek sayıları elde edilir bu olaya budama da denir L2🡪(C3) geçiş. 2 den küçükler silinir (L3)

**Diğer Birliktelik Kuralı Algoritmaları**

|  |  |  |
| --- | --- | --- |
| **Algoritma** | **Veri Tabanını**  **Tarama Sayısı** | **Algoritmanın Özelliği** |
| **AIS** | N | Adaylar, veri tabanı taranarak elde edilir. Küçük ölçekli veritabanlarında uygundur. |
| **Apriori** | N | Adaylar, L(k-1) ile L(k-1) in birleştirilmesiyle elde edilir ve veri tabanı taranarak sayılır.  Orta ölçekli veritabanlarında uygundur. AIS in dezavantajlarından arındırılmıştır. |
| **Apriori\_TID** | N | Adaylar, L(k-1) ile L(k-1) in birleştirilmesinden elde edilir. Ckçok ise yavaş aksi halde  Apriori’den daha performanslıdır. |
| **Apriori\_Hybrid** | N | Apriori ve Apriori\_TID’e göre daha iyi çalışır. Her iki algoritmayı da kullanan melez bir  yapıdır. |
| **OCD** | N | Adaylar, veri tabanı taranarak elde edilir. Büyük veritabanlarında ve küçük destek  Değerlerinde uygulanabilir. |
| **SETM** | N | SQL komutları uyumluluğu mevcuttur. |
| **DHP** | N | Adaylar, L(k-1) ile L(k-1) in birleştirilmesi ile elde edilir ve veri tabanı taranarak sayılır.  Hash tablosu kullanır. |
| **Partition** | 2 | Geniş veritabanlarında uygulanabilir. Homojen veriler üzerinde etkilidir. |
| **MONET** | N | Adaylar, L(k-1) ile L(k-1) in birleştirilmesiyle elde edilir ve kolonların kesiştirilmesi ile  sayılır. |
| **Sampling** | ≤ 2 | Adaylar, L(k-1) ile L(k-1) in birleştirilmesi ile elde edilir ve veri tabanı taranarak sayılır.  Geniş veritabanlarında ve küçük destek değerlerinde uygulanabilir. |
| **DIC** | ≤ N | Adaylar, veri tabanı taranarak sayılır. Yoğun olabilecek tüm nesne kümeleri kontrol edilir. |
| **MaxClique** | 1 | Olası yoğun nesne kümeleri incelenir. *tidlist* liste yapısı kullanır ve kesiştirme işlemi ile  adaylar sayılır. |
| **Max-Miner** | N | Adaylar, L(k-1) ile L(k-1) in birleştirilmesiyle elde edilir ve veri tabanı taranarak sayılır. |
| **Carma** | 2 | Veri tabanı verileri ağ üzerinden elde edildiğinde uygulanabilir. Destek ve güven değerleri  çevrimiçi olarak değiştirilebilmektedir. |

**Birliktelik Kuralı Türleri**

* Hiyerarşik Birliktelik Kuralları
  + Verilerin tek başına ürün olarak değerlenmeyip her bir ürünü kendi başına değerledirmek yerine belli bir hiyerarşi grubunda değerlendirmek. Mesela tüm içecekleri ve içecekleri kendi içerisinde ; alkollü/alkolsüz, gazlı/gazsız diye veya meyveler/tahıllar/sebzeler gibi gruplandırarak veri tabanını hazırlamaya denir. İçecek alan meyve alır gibi kanı belirler.
* Sınırlandırılmış Birliktelik Kuralı
  + Sadece içecekler veya sadece meyveler içerisinde bir bağlantı kurmak istiyorsanız sınırlandırılmış birliktelik kuralları kullanılır.
* Nicel/Nitel Birliktelik Kuralları
  + Havanın yağmurlu,bulutlu güneşli olması veya yaş nicel bir değer. Yaşı 35-40 arası olanlar kesinlikle top oynamamalı gibi.
* Sıralı Örüntüler
  + Ardaşık olarak gerçekleşen olaylar. Bu ürünü alanlar gelecekte şu şu ürünü alacaktır sepetindeki ürünleri değil zaman içerisinde alacakalrı ürünleri analiz eder. Mesela bebek bezi alan belli bir süre sonra mama alacaktır demesi.
* Periyodik Kurallar
  + Bazı kuralların belli zamanlarda gerçekleştirilebilecek olması. Kolonya alan herkesin şeker de alabileceği mantığı bayramlardan bayrama geçerli olacak bir kural olabilir. Veya her haftasonu gazete süt ekmek alınıyor gibi.
* Ağırlıklandırılmış Birliktelik Kuralları
  + Ekmek çok alındı diyelim karalahana az alındı diyelim bu karalahananın satılmadığı anlamına gelmez ekmeğin temel gıda olduğu anlamına gelir. Biz bu kuralda karalahananın diğer ürünlerle olana bağlantısına yine bakarız. Hepsine ağırlıklandırılmış birliktelik kuralı buluruz.
* Negatif Birliktelik Kuralları
  + A ürününü alan B ürününü almaz.

**İlginçlik Ölçütleri**

Destek

Güven veya güç

artış veya ilgi veya korelasyon

mahkumiyet

kaldıraç veya piatetsky – shapiro (leverange)🡪x=destek(A,B)-destek(A)\*destek(B)

kapsama

Lift(A🡪B) = sup(A,B)/(sup(A)-sup(B)) = P(B|A)/P(B) (Güven değeri/ destek )

|  |  |
| --- | --- |
| **İşlemler** | **Satın Alınan** **Ürün Listesi** |
| T1 | I1, I2, I5 |
| T2 | I2, I4 |
| T3 | I2, I3 |
| T4 | I1, I2, I4 |
| T5 | I1, I3 |
| T6 | I2, I3 |
| T7 | I1, I3 |
| T8 | I1, I2, I3, I5 |
| T9 | I1, I2, I3 |

Destek (I2🡪I3) = 4/9 = 0.44 //I2 ve I3 ün birlikte olduğu sayılar/Tüm işlemler

Güven(I2🡪I3) = 4/7=0.57 //I2 ve I3 ün birlikte olduğu sayılar/I2 ün tüm yer aldığı sayılar

Lift(I2🡪I3) = 0.57/0.44= 1.30 //Güven/Destek

**12.HAFTA**

BAYES SINIFLANDIRMA

Koşullu Olasılık;

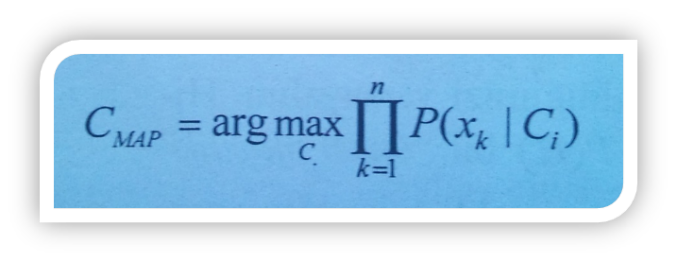
P(B|A)=P(AnB)/P(A) //P(B|A)🡪güven değeri, AnB != boş küme

P(AnB) = P(A)\*P(B|A)

P(B|A) = P(B)\*P(A|B)/P(A)

**Sade Bayes Sınıflandırıcısı**

P(Ci|A) = P(Ci)\*P(A| Ci)/P(A)



Örnek

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Başvuru** | **Eğitim** | **Yaş** | **Cinsiyet** | **Kabul** |
| 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 | KADIN | EVET |

Yukarıdaki eğitim kümesini ele alarak, Bayes sınıflandırıcısını kullanmak suretiyle aşağıdaki örneğin hangi sınıfa ait olduğunu belirlemek istiyoruz.

x1 : EĞİTİM = YÜKSEK

x2 : YAŞ = ORTA

x3 : CİNSİYET = KADIN

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | KABUL | | | |
| NİTELİKLER | DEĞERİ | EVET | | HAYIR | |
| SAYISI | OLASILIK | SAYISI | OLASILIK |
| EĞİTİM | İLK | 1 | 1/5 | 2 | 2/3 |
| ORTA | 3 | 3/5 | 0 | 0 |
| YÜKSEK | 1 | 1/5 | 1 | 1/3 |
| YAŞ | GENÇ | 0 | 0 | 2 | 2/3 |
| ORTA | 3 | 3/5 | 1 | 1/3 |
| YAŞLI | 2 | 2/5 | 0 | 0 |
| CİNSİYET | ERKEK | 3 | 3/5 | 1 | 1/3 |
| KADIN | 2 | 2/5 | 2 | 2/3 |

KABUL = ?

C1 : KABUL = EVET

C2 : KABUL = HAYIR olmak üzere P(X | C1)P(C1) ve P(X | C2)P(C2) ifadelerini hesaplamamız gerekiyor. Söz konusu ifadeler içinde en büyük olanı bize örneğin sınıfını verecektir.

P(X|C1)P(C1) olasılığının hesaplanması;

Burada P(X|KABUL=EVET) koşullu olasılığını hesaplamak gerekiyor. Söz konusu olasılığı bulmak için X={x1,x2,….,xn} değerleri için ayrı ayrı koşullu olasılıkları bulmak gerekiyor. Olasılıkların yer aldığı tablodan faydalanılarak bu olasılıklar hesaplanır.

P(x1|C1) = P(EĞİTİM=YÜKSEK|KABUL=EVET)=1/5

P(x2|C1) = P(YAŞ=ORTA|KABUL=EVET)=3/5

P(x3|C1) = P(CİNSİYET=KADIN|KABUL=EVET)=2/5

O halde;

P(x|C1)=P(X|KABUL=EVET)= (1/5)(3/5)(2/5) = 6/125 hesaplanır. Diğer taraftan P(KABUL=EVET) olasılığı şu şekilde elde edilir.

P(C1)=P(KABUL=EVET) = 5/8

Böylece,

P(X|C1)P(C1)=P(X|KABUL=EVET)\*P(KABUL=EVET) = (6/125)\*(5/8) = 0.03 elde edilmiş olur.

**P(X|C2)P(C2) olasılığının hesaplanması;**

burada önce P(X|C2) olasılığını hesaplamak gerekiyor. Yani P(X|KABUL=HAYIR) olasılığı hesaplanacaktır. X’in her bir değeri için olasılıkların yer aldığı tablodan yararlanılarak aşağıdaki hesaplamalar yapılır.

P(x1|C2) = P(EĞİTİM = YÜKSEK | KABUL = HAYIR) = 1/3

P(x2|C2) = P(YAŞ = ORTA | KABUL = HAYIR) = 1/3

P(x3|C2) = P(CİNSİYET = KADIN | KABUL = HAYIR) = 2/3 bu değerler kullanılarak şu hesaplama yapılır:

P(X|C2) = P(X| KABUL = HAYIR) = (1/3)(1/3)(2/3) = 2/27

Bunun dışında P(KABUL=HAYIR) olasılığı şu şekilde elde edilir:

P(C2) = P(KABUL=HAYIR) = 3/8 olduğundan şu hesaplama yapılabilir:

P(X|C2)P(C2) = P(X|KABUL=HAYIR)P(KABUL=HAYIR) = (2/27)(3/8) = 0.027

**MAP Yöntemine göre sınıflandırmayı yapmak üzere arg max P(P(X|Ci)P(Ci)) değerini bulabiliriz.**

**O halde örneğin 0.03 olasılığı ile ilgili olan sınıfa, yani “EVET” sınıfına ait olduğu anlaşılır.**

**(maksimum değeri alırız)**

**Bayes Sınıflandırıcılarda Sıfır Değer Sorunu**

Bayes sınıflandırma yöntemini uygularken bir sorunla karşılaşabiliriz. Önceki slaytta verilen son örneği ele alacak olursak, KABUL=HAYIR olan kadınların sayısı 0 ise,

P(x3 | C2) = P(CİNSİYET = KADIN | KABUL = HAYIR) = 0 elde edilir. Bu durumda sıfır ile çarpım sonunda,

P(X | C2) = P(X | KABUL = HAYIR) = (1/3)(1/3)(0) = 0 olacaktır. Yani diğer sonuçların etkisi yok olmuş ve sonuç anlamlı olmamıştır. Bunu önlemek için k gibi küçük bir değeri her bir orana ekleyebiliriz. Böylece n/d için,

(n + kp) / (d + k) bağıntısı kullanılır. Burada k, 0 ile 1 arasında bir sayıdır. Genellikle 1 tercih edilir. Burada p değeri ise her bir nitelik değerinin muhtemel toplam sayısının bir belirli kısmı olarak seçilir. Eğer bir niteliğin iki muhtemel değeri varsa,

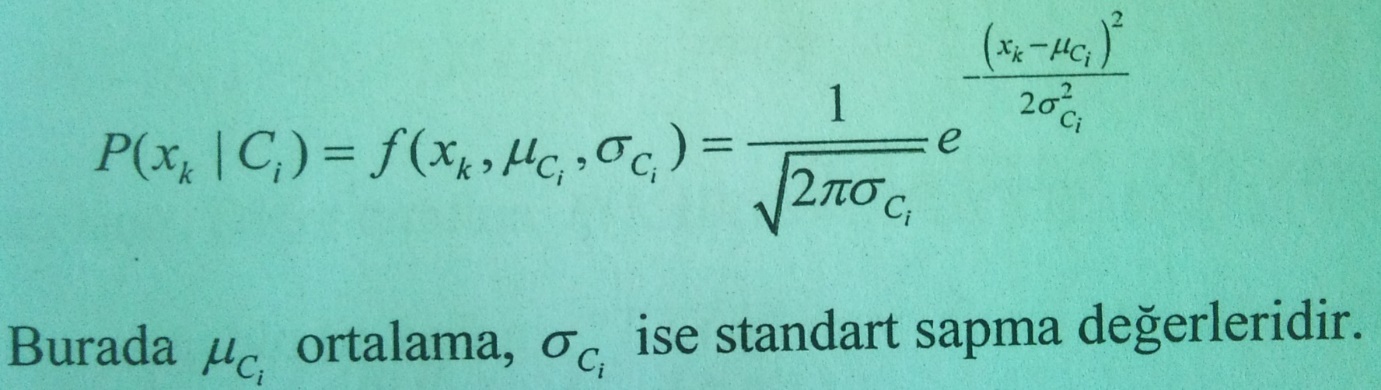
p = 1 / 2 = 0.5 olarak kabul edilebilir. Şimdi P(X | KABUL = HAYIR) koşullu olasılığını yeniden hesaplayalım. K = 1, p = 0.5 olacak biçimde,

P(X | C2) = P(X | KABUL = HAYIR) = ( ( (1 + 0.5) / (3 + 1) ) . ( (1 + 0.5) / (3 + 1) ) . ( (2 + 0.5) / (3 + 1) ) ) = 0.0878

elde edilir.

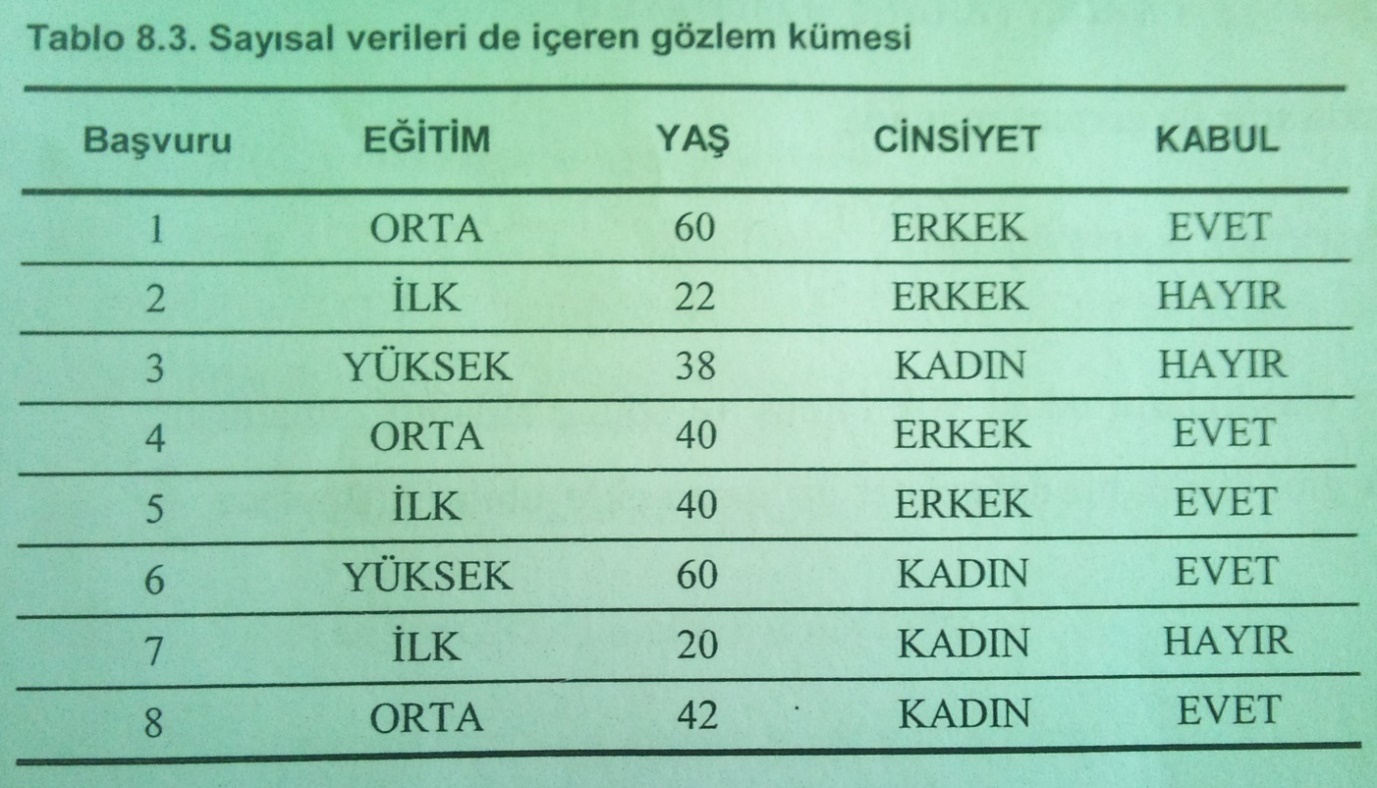
**Sayısal Nitelik Değerleri**

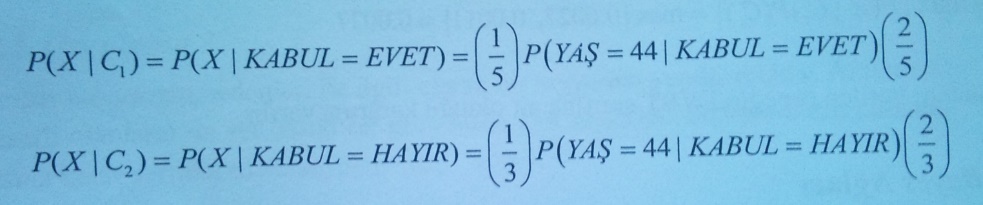
Nitelik değerleri sayısal ise, önceki slaytlarda yaptığımız işlemleri bu durum için tekrarlayamayız. Sayısal verilerin dağılımının normal dağıldığı varsayılarak aşağıdaki standart olasılık yoğunluk fonksiyonu kullanılır.



Örnek:

Önceki örneğimizde tüm değişkenler kategorik verilere sahipti. Uygulamalarda çoğu kez sürekli değişkenlerle, yani sayısal verilerle çalışmak gerekebilir. Bu uygulamada YAŞ niteliğinin değerlerini sayısal olarak belirliyoruz.





Hangi sınıfa ait olduğunu bulmak istediğimiz örnek şu şekildedir:

x1: EĞİTİM=YÜKSEK

x2: YAŞ=44

x3: CİNSİYET=KADIN

KABUL = ?

Burada,

C1: KABUL=EVET

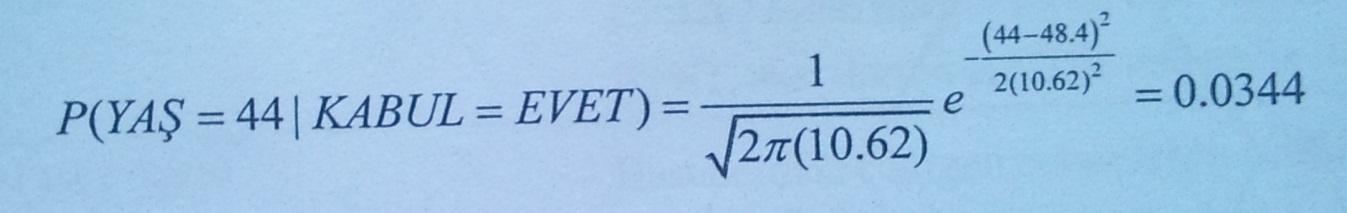
C2: KABUL=HAYIR

Olarak belirlenmiştir. Önceki örnekte aynı verileri kullanarak elde edilen koşullu olasılıkları yeniden hesaplamadan aynen kullanacağız.

Önceki slaytta yer alan iki olasılığı hesaplamak için normal dağılım fonksiyonundan yararlanılır. Bunun için öncelikle uCi ortalama değeri ile oCi standart sapma değerinin elde edilmesi gerekmektedir. KABUL=EVET ile ilgili YAŞ değerlerini göz önüne alalım.

uCi = 48.4 oCi = 10.62

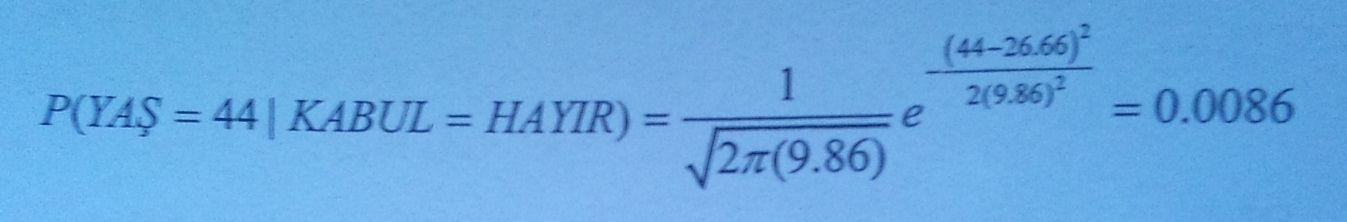
Bu durumda şu hesaplama yapılır:

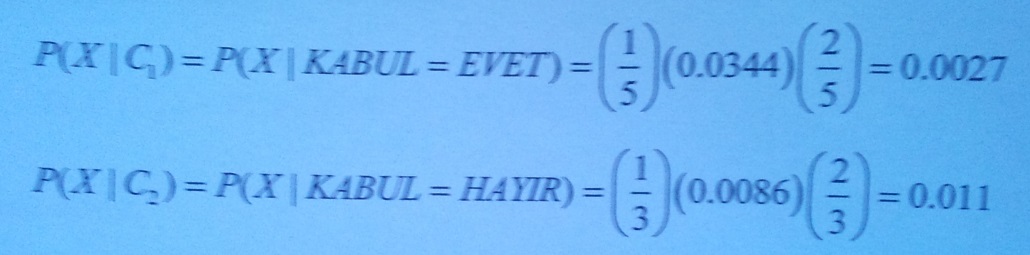


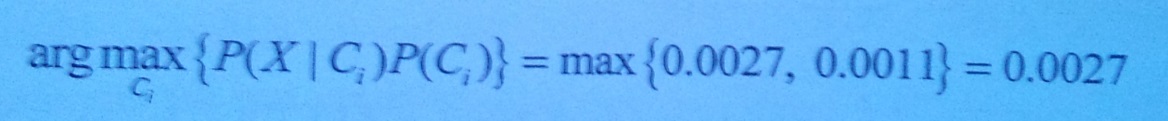
KABUL=HAYIR ile ilgili YAŞ değerleri için,

uC2 = 26.66 oC2 = 9.86 elde edilir.

Bu değerler bağıntıda kullanılırsa şu değerler elde edilir:







olduğundan, verilen örneğin EVET sınıfına ait olduğu kararına varılır