İSTATİSTİKSEL SINIFLANDIRMA MODELLERİ

Bayes Sınıflandırıcılar ve Bayes Ağları

Sınıflandırma işleminde istatistiksel teknikler de kullanılmaktadır. Bunlardan birisi de ***Bayes teoremine*** dayanmaktadır. Değişkenlere ait alt kümeler arasındaki koşullu bağımsızlıkları tanımlamak üzere ***"Bayes Ağları"*** kullanılabilir. Bu bölümde istatistiksel sınıflandırma ve *Bayes ağlar* ele alınmıştır.

8.1. Koşullu Olasılık

Bir olayın gerçekleşmesi bazı koşulların var olmasına bağlı ise *"koşullu olasılık"* dan söz edilir. A ve B gibi iki bağdaşan olayı göz önüne alalım. Yani A ve B olaylarının ortak noktaları vardır. Bu durum $A \cap B \neq \emptyset$ biçiminde ifade edilebilir. B olayı A olayının bilinmesi durumunda gerçekleşecektir. Bu olasılık P(B|A) biçiminde gösterilir ve şu şekilde ifade edilir.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Bu ifadeden yararlanarak bileşik olasılık bağıntısı elde edilir.

$$P(A \cap B) = P(A)P(B \mid A)$$

8.1. Koşullu Olasılık

O halde A ve B gibi iki olayın birbiri ardına gerçekleşme olasılığı, iki olaya ait olasılıkların çarpımına eşittir.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

yazılabildiğinden,

$$P(A \cap B) = P(B)P(A \mid B)$$

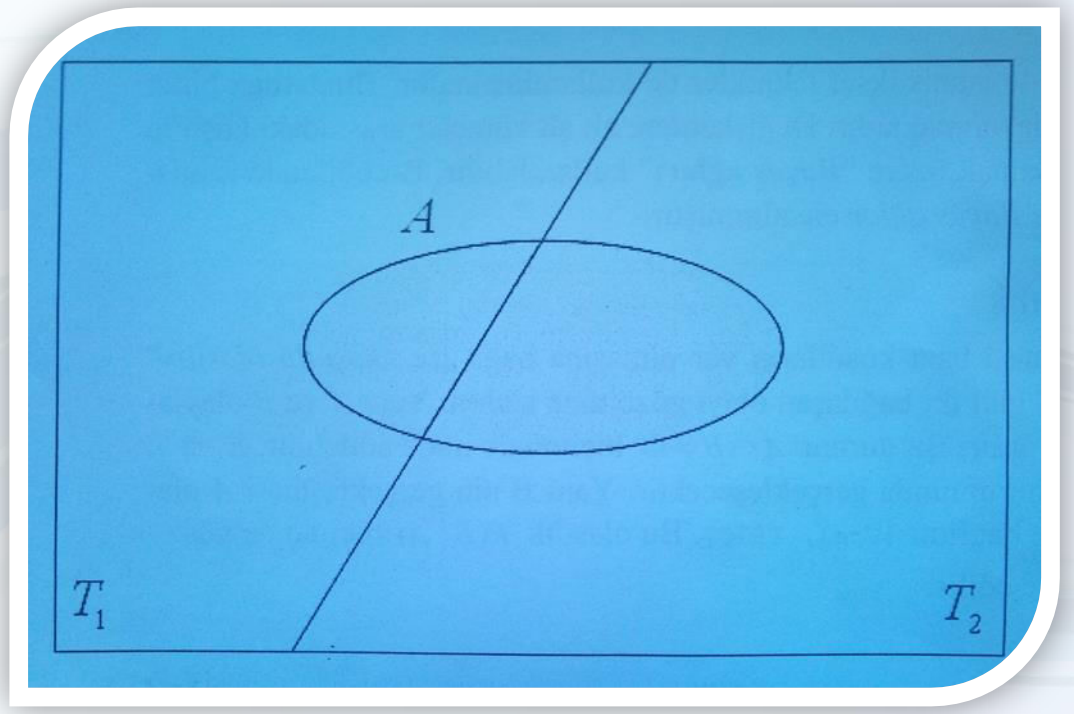
biçiminde ifade edilebilir. Bu bağıntı yerine yazılırsa;

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(A)}$$

elde edilir.

8.2. Bayes Teoremi

Bayes Teoremi olasılıklar hesabında önemli bir yere sahiptir. *Bayes* teoremine dayanarak sınıflandırma işlemi yapmak mümkündür. Bu teoremin esasını ortaya koymak üzere T_1 ve T_2 gibi iki olayı göz önüne alalım. Bu iki olayın bağdaşmaz olaylar olduğunu, yani $T_1 \cap T_2 = \emptyset$ olduğunu varsayalım.



8.2. Bayes Teoremi

Bir A olayını T_1 ve T_2 olayları cinsinden ifade etmemiz mümkündür. Bunun için koşullu olasılık bağıntısından yararlanılır.

Yani $P(T_1|A)$ olasılığı,

$$P(T_1 \mid A) = \frac{P(A \mid T_1)}{P(A)}$$

biçimindedir.

A olayı T_1 ve T_2 olaylarından birinde gerçekleşmektedir. A olayı için şu bağıntı yazılabilir:

$$A = (T_1 \cap A) \cup (T_2 \cap A)$$

8.2. Bayes Teoremi

Doğal olarak A olayının olasılığı,

$$P(A) = P(T_1 \cap A) + P(T_2 \cap A)$$

yazılabilir.

Denklemden yararlanılarak şu bağıntı elde edilir.

$$P(T_1 \mid A) = \frac{P(A \mid T_1)}{P(T_1 \cap A) + P(T_2 \cap A)}$$

8.2. Bayes Teoremi

Bu sonuç genelleştirilebilir. Temel küme $T_1,T_2,\ldots,T_n$ olaylarından oluşuyorsa ve olasılıkları sıfırdan farklı ise bir A olayının T_j olayı içinde gerçekleşme olasılığı şu şekilde hesaplanır:

$$P(T_j \mid A) = \frac{P(A \mid T_j)}{\sum_{i=1}^{n} P(A \cap T_i)}$$

Burada, $P(A \mid T_j) = P(A \mid T_j)P(T_j)$ olduğundan aşağıdaki gibi yazılabilir.

$$P(T_j \mid A) = \frac{P(A \mid T_j)P(T_j)}{\sum_{i=1}^{n} P(A \cap T_i)}$$

8.3. Bayes Sınıflandırıcısı

Bayes sınıflandırıcıları istatistiksel sınıflandırma teknikleri arasında yer alır. Bu sınıflandırma işlemine başlarken X kümesini sınıf üyeliği bilinmeyen veri kümesi olarak kabul edelim. H ise bu X veri örneğinin C sınıfına ait olduğu iddia edilen hipotez olsun. O halde, H'nin C sınıfına ait olduğu varsayımıyla P(H|X) olasılığını hesaplamamız söz konusudur. Burada P(H|X), H hipotezinin X üzerinde koşullandırılmasıyla ilgili *"sonrasal olasılık"* olarak bilinir.

8.3. Bayes Sınıflandırıcısı

Örneğin bir torbada bazı cisimlerin bulunduğunu varsayalım. Elimizdeki bilgiler X'i tanımlar. Cisimlerin yuvarlak ve kırmızı olduğunu bildiğimizi varsayalım. Hipotezimiz ise *"bu cisim bir elmadır"* olsun. Bu durumda P(H) bir *"önsel olasılık"* olarak karşımıza çıkacaktır. Yani başlangıçta bu olasılığın ne olduğunu biliyoruz. Ancak P(X|H) olasılığı ise H koşulu üzerine kurulduğundan bir *"sonrasal olasılık"* olarak değerlendirilir. Yani, bir X'in kırmızı ve yuvarlak olma olasılığı biliniyorsa "bu bir elmadır" sonucunu doğurur. Bu durumda *Bayes* bağıntısı şu şekli alır.

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

8.3.1. Sade Bayes Sınıflandırıcısı

Sade Bayes sınıflandırıcısı (Naive Bayes classifier) ya da kısaca *"Bayes sınıflandırıcısı"* adını verdiğimiz kavramı şu şekilde açıklayabiliriz:

X sınıf üyeliği bilinmeyen veri örneği olsun. Örnek $X=\{x_1,x_2....,x_n\}$ nitelik değerlerinden oluşsun. Bu örnek sınıfta m sınıf olduğunu varsayalım. $C_1,C_2...,C_N$ sınıf değerleri olsun. Sınıfı belirlenecek olan örneğe ilişkin olarak,

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}$$

8.3.1. Sade Bayes Sınıflandırıcısı

olasılıkları hesaplanır. Hesaplamalardaki işlem yükünü azaltmak üzere $P(X|C_i)$ olasılığı için basitleştirme yoluna gidilebilir. Bunun için, örneğe ait x_i değerlerinin birbirinden bağımsız olduğu kabul edilerek şu bağıntı kullanılabilir.

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

8.3.1. Sade Bayes Sınıflandırıcısı

Bilinmeyen örnek X'i sınıflandırmak için $P(C_i|X)$ içinde yer alan paydalar birbirine eşit olduğuna göre sadece pay değerlerinin karşılaştırılması yeterlidir. Bu değerler içinden en büyük olanı seçilerek bilinmeyen örneğin bu sınıfa ait olduğu belirlenmiş olur.

$$\arg\max_{C_i} \{P(X \mid C_i)P(C_i)\}$$

Sonrasal olasılıkları kullanan yukarıdaki ifade, en büyük sonrasal sınıflandırma yöntemi (Maximum A Posteriori classification = MAP) olarak da bilinir. O halde sonuç olarak, *Bayes* sınıflandırıcısı olarak aşağıdaki bağıntı kullanılabilir.

$$C_{MAP} = \arg\max_{C_i} \prod_{k=1}^{n} P(x_k \mid C_i)$$

Uygulama-1

Aşağıdaki verileri göz önüne alalım. Bu verileri *Gini algoritması* ile ilgili bölümde ele alarak incelemiştik.

Örnek veri kümesi

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 = ?

Bayes olasılıklarını hesaplamak amacıyla bir sonraki slayttaki tabloyu düzenliyoruz.

Uygulama-1

Olasılıkların yer aldığı tablo

NİTELİKLER	DEĞERİ	KABUL			
		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

Bayes sınıflandırmasını gerçekleştirmek için her bir hipotez için *Bayes* olasılıkları tek tek hesaplanır.

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.

Uygulama-1

a) 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.

Uygulama-1

b) 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

Uygulama-1

c) Sonuç

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

$$\arg\max_{C_i}\{P(X \mid C_i)P(C_i)\} = \max\{0.03,\ 0.027\} = 0.03$$

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

8.3.2. 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,

$P(x_3 \mid C_2)$ = P(CİNSİYET = KADIN | KABUL = HAYIR) = 0 elde edilir. Bu durumda sıfır ile çarpım sonunda,

$P(X \mid C_2)$ = 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,

8.3.2. Bayes Sınıflandırıcılarda Sıfır Değer Sorunu

(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,

$P(X \mid C_2) = P(X \mid KABUL = HAYIR) = (((1 + 0.5) / (3 + 1)) . ((1 + 0.5) / (3 + 1)) . ((2 + 0.5) / (3 + 1))) = 0.0878$

elde edilir.

8.3.3. 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.

$$P(x_k \mid C_i) = f(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi\sigma_{C_i}}} e^{-\frac{(x_k - \mu_{C_i})^2}{2\sigma_{C_i}^2}}$$

Burada μ_{C_i} ortalama, σ_{C_i} ise standart sapma değerleridir.

Uygulama - 2

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.

Tablo 8.3. Sayısal verileri de içeren gözlem kümesi

Başvuru	EĞİTİM	YAŞ	CİNSİYET	KABUL
1	ORTA	60	ERKEK	EVET
2	İLK	22	ERKEK	HAYIR
3	YÜKSEK	38	KADIN	HAYIR
4	ORTA	40	ERKEK	EVET
5	İLK	40	ERKEK	EVET
6	YÜKSEK	60	KADIN	EVET
7	İLK	20	KADIN	HAYIR
8	ORTA	42	KADIN	EVET

Uygulama - 2

Hangi sınıfa ait olduğunu bulmak istediğimiz örnek şu şekildedir:

x_1: EĞİTİM=YÜKSEK

x_2: YAŞ=44

x_3: CİNSİYET=KADIN

KABUL = ?

Burada,

C_1: KABUL=EVET

C_2: KABUL=HAYIR

$$P(X \mid C_1) = P(X \mid KABUL = EVET) = \left(\frac{1}{5}\right) P(YAŞ = 44 \mid KABUL = EVET) \left(\frac{2}{5}\right)$$

$$P(X \mid C_2) = P(X \mid KABUL = HAYIR) = \left(\frac{1}{3}\right) P(YAŞ = 44 \mid KABUL = HAYIR) \left(\frac{2}{3}\right)$$

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

Uygulama - 2

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

$u_{Ci} = 48.4 \quad o_{Ci} = 10.62$

Bu durumda şu hesaplama yapılır:

$$P(YAŞ = 44 \mid KABUL = EVET) = \frac{1}{\sqrt{2\pi}(10.62)} e^{\frac{(44-48.4)^2}{2(10.62)^2}} = 0.0344$$

Uygulama - 2

KABUL=HAYIR ile ilgili YAŞ değerleri için,

$u_{C2} = 26.66$ $o_{C2} = 9.86$ elde edilir.

Bu değerler bağıntıda kullanılırsa şu değerler elde edilir:

$$P(YAŞ = 44 \mid KABUL = HAYIR) = \frac{1}{\sqrt{2\pi}(9.86)} e^{\frac{(44-26.66)^2}{2(9.86)^2}} = 0.0086$$

Uygulama - 2

Bu durumda,

$$P(X \mid C_1) = P(X \mid KABUL = EVET) = \left(\frac{1}{5}\right)(0.0344)\left(\frac{2}{5}\right) = 0.0027$$

$$P(X \mid C_2) = P(X \mid KABUL = HAYIR) = \left(\frac{1}{3}\right)(0.0086)\left(\frac{2}{3}\right) = 0.011$$

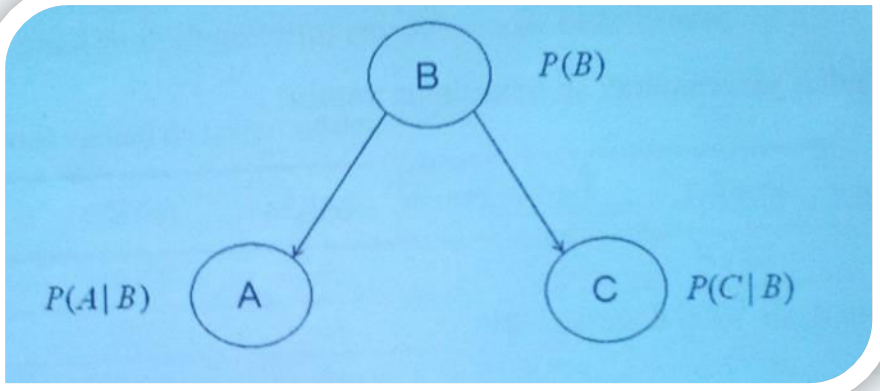
elde edilir. Bu sonuçlar değerlendirildiğinde,

$$\arg\max_{C_i}\{P(X \mid C_i)P(C_i)\} = \max\{0.0027,\ 0.0011\} = 0.0027$$

olduğundan, verilen örneğin EVET sınıfına ait olduğu kararına varılır.

8.4. Bayes Ağları

Değişkenlere ait koşullu olasılık dağılımlarını ortaya koymak ve değişkenlere ait alt kümeler arasındaki koşullu bağımsızlıkları tanımlamak üzere *'Bayes Ağları'* kullanılabilir. *Bayes* ağları grafik modeller arasında değerlendirilir. Bir *Bayes* ağı kurulduktan sonra, bu ağ sınıflandırma amacıyla kullanılabilir. Ağa bir takım sorular yöneltilerek sonuçlar üretilir. Basit bir *Bayes* ağı aşağıdaki şekil üzerinde yer almaktadır.



Yukarıdaki *Bayes* ağında A ve C değişkenlerinin B'den koşullu olarak bağımsız olduğu varsayılır. Bu durumda şu bağıntı yazılabilir:

$$P(A,B,C) = P(A \mid B)P(C \mid B)P(B)$$

8.4. Bayes Ağları

Önceki slayttaki bağıntıya *zincir kuralı* adı da verilmektedir. Bu bağıntı genelleştirilecek olursa, aşağıda verilen bağıntıya ulaşılır. Bu *bileşik olasılık dağılımı* olarak da bilinmektedir.

$$P(X_1, X_2, \ldots, X_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P\left(X_i = x_i \mid ebeveyn(X_i)\right) \quad (8.15)$$

Burada alt düğüm olasılıklarının sadece kendi ebeveynlerine bağlı olduğuna dikkat etmek gerekiyor.

Bayes ağları üzerinde işlemler yapabilmek için öncelikle bir grafik modelin belirlenmesi beklenir. Model oluşturulduktan sonra olasılıklar model üzerinde yerleştirilir. Söz konusu olasılıklar konuyla ilgili uzmanların görüşlerine göre belirlenebildiği gibi, göreli frekanslar kullanılarak da bulunabilir.

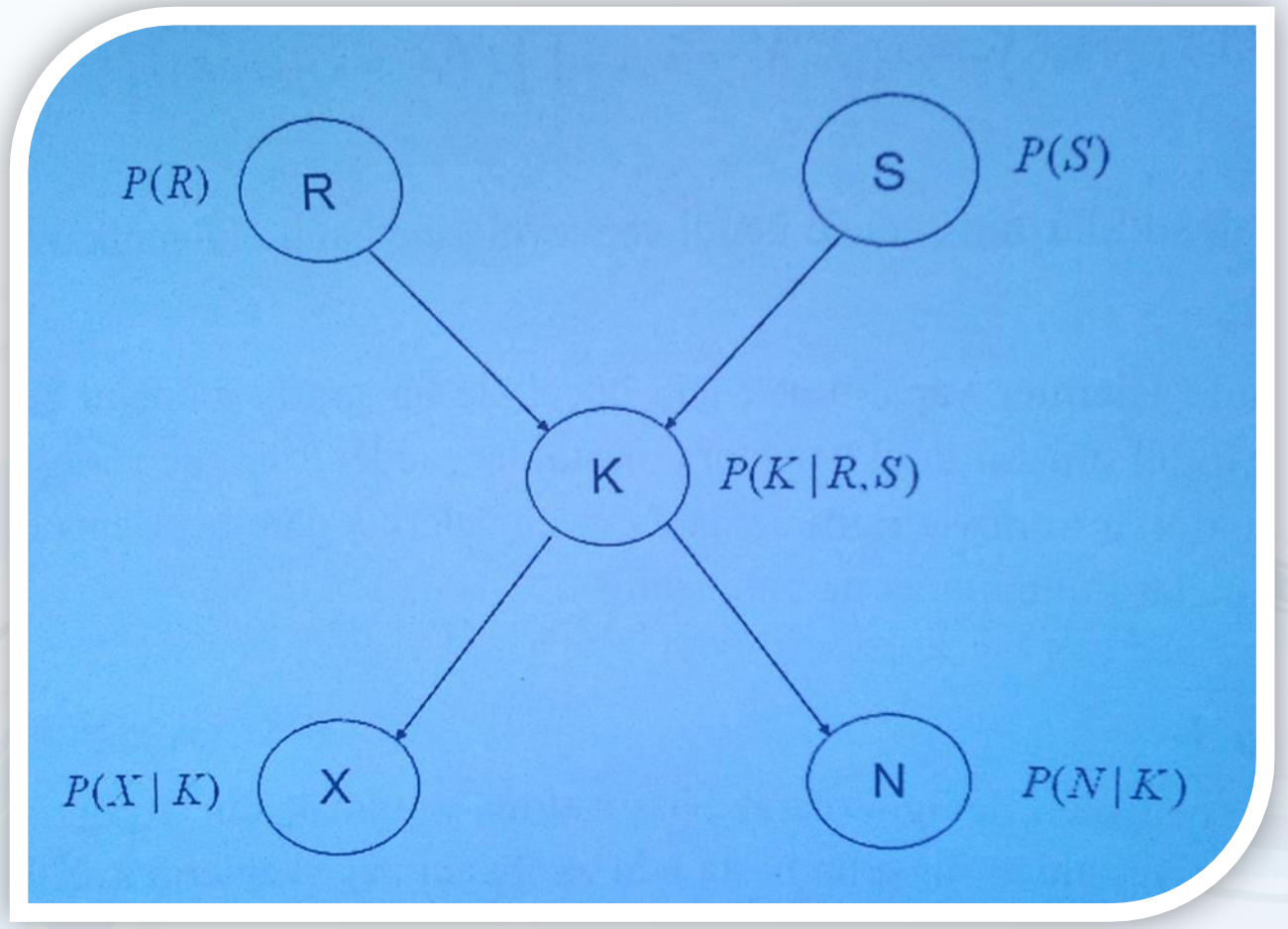
Uygulama - 3

Akciğer kanserinin nedenleri ile ilgili olarak bir araştırma yapılmaktadır. Aralarında kanserli olanların da yer aldığı bir grup hasta için aşağıdaki değişkenlerin kullanılmasına karar verildiğini varsayalım.

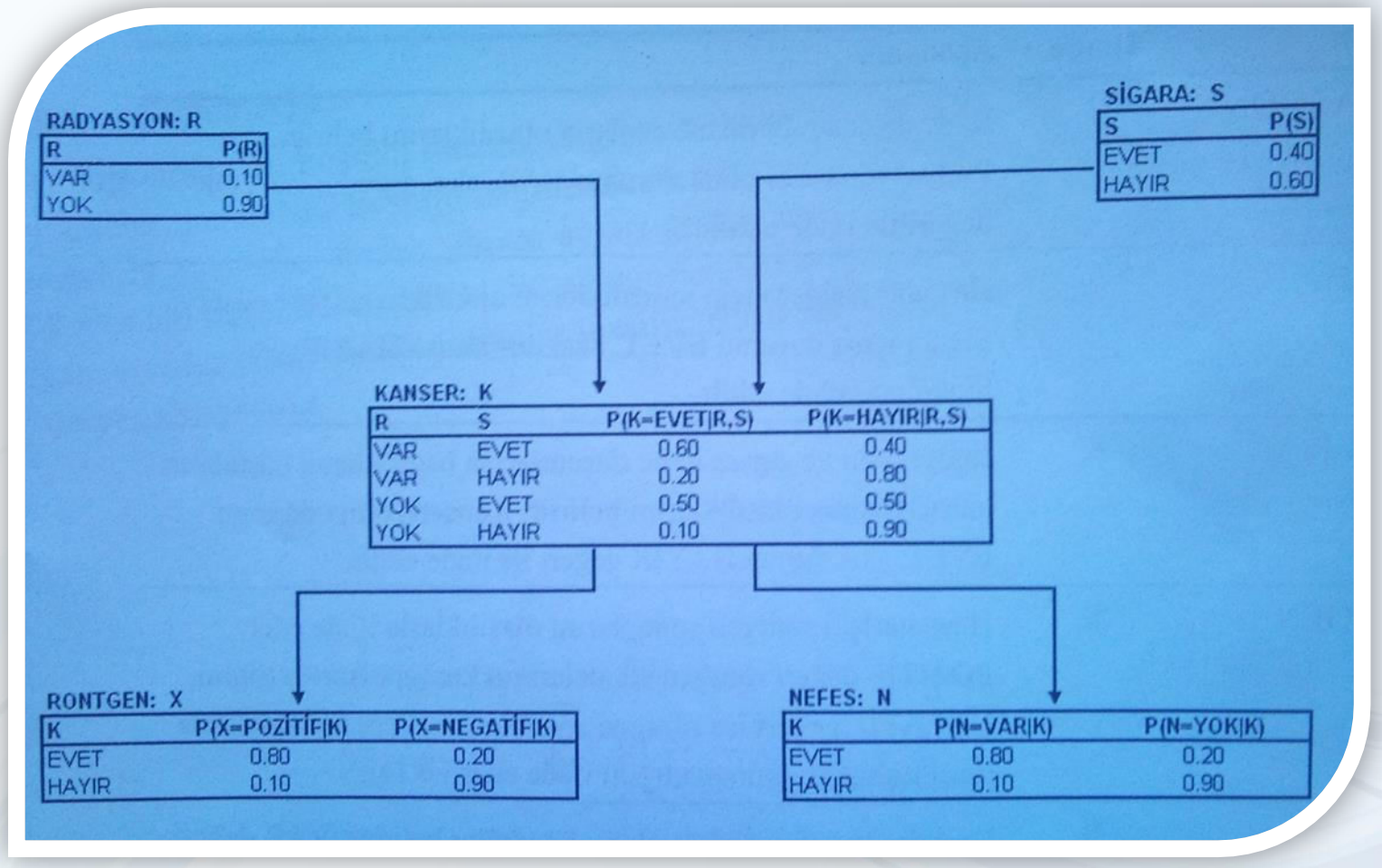
Değişken	Simge	Açıklama
RADYASYON	R	Radyasyona tabi olan hastaların olasılıklarını belirler. Radyasyona tabi olma durumu VAR, aksi durumu ise YOK değeri ile ifade edilir.
SİGARA	S	Hastaların sigara içip içmediklerini olasılıklarla ifade eder. Sigara içme durumu EVET, aksi durum ise HAYIR biçiminde ifade edilir.
KANSER	K	Radyasyon ve sigara içme durumlarına bağlı olarak hastaların kanserli olma olasılıklarını belirtir. Kanserli olma durumu EVET, aksi durum HAYIR değeri ile ifade edilir.
RÖNTGEN	X	Hastalardaki röntgen sonuçlarını olasılıklarla ifade eder. POZİTİF değeri röntgen sonuçlarının kansere işaret ettiğini, NEGATİF değeri ise röntgen sonuçlarının kanserli hastalarla ilgili teşhisi doğrulamadığını ifade eder.
NEFES	N	Hastalarda nefes darlığı olup olmadığını belirler. VAR değeri böyle bir sıkıntısı olanların olasılığını, YOK ise aksi durumu ifade eder.

Uygulama - 3

Önceki slayttaki şekil üzerinde beş değişkene ilişkin *Bayes* ağını görüyoruz. Söz konusu model konunun uzmanları tarafından belirlenmektedir. Söz konusu *Bayes* ağına ilişkin olasılıklar ise aşağıdaki şekil üzerinde yer almaktadır.



Uygulama - 3



Yukarıdaki *Bayes* ağını ele alarak bir çok soruya olasılıklara bağlı olarak yanıt vereceğiz.

Uygulama - 3

a) Radyasyonun etkisi altında kalmış, sigara içen, röntgeni pozitif çıkan ve nefes alma sorunları yaşayan kanserli hastaların olasılığı nedir? Bu sorunun yanıtını *Bayes* ağı üzerinden elde etmek istiyoruz. Bu sorudan, aşağıdaki nitelik değerlerini sorgulayacağımız anlaşılıyor.

Tablo 8.4. Nitelik ve değerleri

Nitelik	Değeri
RADYASYON (R)	VAR
SİGARA (S)	EVET
RÖNTGEN (X)	POZİTİF
NEFES (N)	VAR
KANSER (K)	EVET

Uygulama - 3

Böyle bir durumda P(R = VAR, S = EVET, R = POZİTİF, N = VAR, K = EVET) olasılığının hesaplanması gerektiği anlaşılıyor. Bağıntı içinde yer alan aşağıdaki olasılıklar tek te hesaplanır.

$$p_1 = P(R = VAR) = 0.10$$

$$p_2 = P(S = EVET) = 0.40$$

$$p_3 = P(K = EVET \mid R = VAR, S = EVET) = 0.60$$

$$p_4 = P(X = POZİTİF \mid K = EVET) = 0.80$$

$$p_5 = P(N = VAR \mid K = EVET) = 0.60$$

O halde sonuç olarak şu olasılık elde edilir.

$$P(R = VAR, S = EVET, R = POZİTİF, N = VAR, K = EVET) = p_1 p_2 p_3 p_4 p_5$$

$$= 0.10(0.40)(0.60)(0.80)(0.60) = 0.011$$

Uygulama - 3

b) Benzer biçimde *Bayes* ağı kullanılarak RADYASYON, SİGARA, KANSER, RÖNTGEN ve NEFES değişken değerlerinin çeşitli kombinasyonları seçilerek diğer olasılıklar hesaplanabilir.

c) Hastanın kanserli olduğu bilindiğine göre bu hastanın sigara içenler arasında yer alıp almadığını öğrenme istiyoruz. Bunun anlamı, kanserli bir hastanın SİGARA niteliğinin hangi sınıfına ait olduğunu ortaya çıkarmaktır. Bunu sağlamak için

P(S = EVET | K = EVET) ve P(S = HAYIR | K = EVET) olasılıkları hesaplanır.

MAP kuralına göre bu olasılıklardan en büyük olanı seçilir. Söz konusu olasılıkları elde etmek için bazı *önsel* ve *sonrasal* olasılıkların elde edilmesi söz konusudur.

Önsel Olasılıklar:

Bayes ağı kullanılarak her düğüm için önsel olasılıklar hesaplanabilir. Örneğin P(K=EVET) önsel olasılığı şu şekilde ifade edilebilir:

$$
\begin{aligned}
P(K=EVET) = {} & P(K=EVET \mid R=VAR, S=EVET)P(R=VAR, S=EVET) \\
& + P(K=EVET \mid R=VAR, S=HAYIR)P(R=VAR, S=HAYIR) \\
& + P(K=EVET \mid R=YOK, S=EVET)P(R=YOK, S=EVET) \\
& + P(K=EVET \mid R=YOK, S=HAYIR)P(R=YOK, S=HAYIR)
\end{aligned}
$$

Uygulama - 3

Burada RADYASYON ve SİGARA değişkenlerinin birbirinden bağımsız olduğuna dikkat etmek gerekiyor. O halde, bu bağımsızlık özeliğinden yararlanılarak,

$$P(R = VAR, S = EVET) = P(R = VAR)P(S = EVET)$$

yazılabilir. Bu durumda,

$$P(R = VAR, S = EVET) = 0.10(0.40) = 0.04$$

elde edilir. Benzer biçimde aşağıdaki sonuçlar bulunur:

$$P(R = VAR, S = HAYIR) = 0.10(0.60) = 0.06$$
$$P(R = YOK, S = EVET) = 0.90(0.40) = 0.36$$
$$P(R = YOK, S = HAYIR) = 0.90(0.60) = 0.54$$

Bunları P(K = EVET) olasılık hesaplamasında yerine yazılırsa şu sonuç elde edilir.

$$P(K = EVET) = 0.60(0.04) + (0.20)(0.06) + (0.50)(0.36) + (0.10)(0.54)$$
$$= 0.27$$

Uygulama - 3

Sonrasal Olasılıklar:

Bu aşamada yukarıda hesaplanan önsel olasılıkları kullanarak P(S = EVET | K = EVET) ve P(S = HAYIR | K = EVET) olasılıklarını elde etmemiz söz konusudur.

$$P(S = EVET \mid K = EVET) = \frac{P(S = EVET, K = EVET)}{P(K = EVET)}$$

Öncelikle P(S = EVET, K = EVET) olasılığını hesaplamak gerekiyor.

$$P(S = EVET, K = EVET) = P(S = EVET, R = VAR, K = EVET) + P(S = EVET, R = YOK, K = EVET)$$

olduğundan, *Bayes* ağı üzerindeki olasılıkları kullanarak aşağıdaki hesaplamaları yapmak gerekiyor.

$$P(S = EVET, R = VAR, K = EVET) = P(S = EVET)P(R = VAR) P(K = EVET \mid S = EVET, R = VAR)$$
$$= 0.40(0.10)(0.60) = 0.024$$

$$P(S = EVET, R = YOK, K = EVET) = P(S = EVET)P(R = YOK) P(K = EVET \mid S = EVET, R = YOK)$$
$$= 0.40(0.90)(0.50) = 0.18$$

Uygulama - 3

Önceki slayttan şu değerler elde edilir.

$$P(S = EVET, K = EVET) = 0.024 + 0.18 = 0.204$$

Böylece,

$$P(S = EVET \mid K = EVET) = \frac{P(S = EVET \mid K = EVET)}{P(K = EVET)} = \frac{0.204}{0.27} = 0.76$$

Hesaplanır. Benzer biçimde P(R = VAR, S = EVET) = 0.10(0.40) = 0.04 olasılığı bulunur. SİGARA değişkeni için iki sınıf var olduğuna göre bu olasılığın,

$$P(S = HAYIR \mid K = EVET) = 1 - P(S = EVET \mid K = EVET) = 1 - 0.76 = 0.24$$

Olması beklenir. Ancak biz hesaplamaları tek te yaparak aynı sonucu görmek istiyoruz.

$$P(S = HAYIR \mid K = EVET) = \frac{P(S = HAYIR, K = EVET)}{P(K = EVET)}$$

Uygulama - 3

Bağıntıda yer alan P(S = HAYIR, K = EVET) olasılığı şu şekilde hesaplanabilir:

$$P(S = HAYIR, K = EVET) = P(S = HAYIR, R = VAR, K = EVET) + P(S = HAYIR, R = YOK, K = EVET)$$

Burada yer alan olasılıklar şu şekilde hesaplanır:

$$P(S = HAYIR, R = VAR, K = EVET) = P(S = HAYIR)P(R = VAR) P(K = EVET \mid S = HAYIR, R = VAR) = 0.60(0.10)(0.20) = 0.012$$

$$P(S = HAYIR, R = YOK, K = EVET) = P(S = HAYIR)P(R = YOK) P(K = EVET \mid S = HAYIR, R = YOK) = 0.60(0.90)(0.10) = 0.054$$

O halde,

$$P(S = HAYIR, K = EVET) = 0.012 + 0.054 = 0.066$$

elde edilir.

Uygulama - 3

Bu ifadeden yararlanılarak değere ulaşılır.

$$P(S = HAYIR \mid K = EVET) = \frac{P(S = HAYIR, K = EVET)}{P(K = EVET)}$$

$$= \frac{0.066}{0.27} = 0.24$$

Sonuç olarak MAP kuralına göre,

$$\arg\max_{C_i} \{P(X \mid C_i)P(C_i)\} = \max\{0.76,\ 0.24\} = 0.76$$

elde edilir. O halde SİGARA = EVET sınıfının seçildiği anlaşılıyor.

ÖZET

Sınıflandırma işinde istatistiksel teknikler de kullanılmaktadır. Bunlardan birisi de *Bayes teoremine* dayanmaktadır. *Bayes* bağıntısı;

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

biçiminde ifade edilir. Bu bağıntıdan hareket edilerek *Bayes sınıflandırıcılarına* ulaşılır. X = {x_1, x_2, …., x_n} nitelik değerlerinden oluşsun. Bu örnekte m sınıf olduğunu kabul edecek olursak, C_1,C_2, …., C_n sınıf değerleri için;

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}$$

olasılıkları hesaplanır. Hesaplamalardaki işlem yükünü azaltmak üzere P(X | C_i) olasılığı için basitleştirme yoluna gidilebilir. Bunun için, örneğe ait x_i değerlerinin birbirinden bağımsız olduğu kabul edilir. Ve

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

bağıntısı kullanılır.

8.5. ÖZET

Bilinmeyen örnek X'i sınıflandırmak için;

$$\arg\max_{C_i}\{P(X \mid C_i)P(C_i)\}$$

seçimi yapılır. O halde *Bayes sınıflandırıcısı* olarak ;

$$C_{MAP} = \arg\max_{C_i} \prod_{k=1}^{n} P(x_k \mid C_i)$$

ifadesi kullanılabilir. Değişkenlere ait alt kümeler arasındaki koşullu bağımsızlıkları tanımlamak üzere *'Bayes ağları'* kullanılabilir. *Bayes* ağları sınıflandırma amacıyla da tercih edilebilir. *Bayes* ağlarında şöyle bir bileşik olasılık dağılımına başvurulur:

$$P(X_1, X_2, \ldots, X_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} P\left(X_i = x_i \mid ebeveyn(X_i)\right).$$

8.6. SORULAR

1) Beş gözleme ait aşağıdaki veriyi göz önüne alalım. Burada TEŞHİS hedef değişken olarak belirlenmiştir. *Bayes sınıflandırıcılarını* kullanarak aşağıda verilen örneğin hangi sınıfa ait olduğunu saptayınız.

CİNSİYET = ERKEK YAŞ = GENÇ KİLO = AĞIR TEŞHİS = ?

Gözlem	CİNSİYET	YAŞ	KİLO	TEŞHİS
1	Erkek	Genç	Hafif	Negatif
2	Erkek	Orta Yaşlı	Hafif	Negatif
3	Erkek	Genç	Ağır	Negatif
4	Kadın	Yaşlı	Hafif	Pozitif
5	Kadın	Orta Yaşlı	Ağır	Pozitif

8.6. SORULAR

2) *Quinlan'ın* sınıflandırma örneğini bu kez *Bayes* sınıflandırıcılarını ele alarak çözünüz.

HAVA	ISI	NEM	RÜZGAR	OYUN
güneşli	sıcak	yüksek	hafif	Hayır
güneşli	sıcak	yüksek	kuvvetli	Hayır
bulutlu	sıcak	yüksek	Hafif	Evet
yağmurlu	ılık	yüksek	Hafif	Evet
yağmurlu	soğuk	normal	Hafif	Evet
yağmurlu	soğuk	normal	kuvvetli	Hayır
bulutlu	soğuk	normal	kuvvetli	Evet
güneşli	ılık	yüksek	Hafif	Hayır
güneşli	soğuk	normal	Hafif	Evet

yağmurlu	ılık	normal	Hafif	Evet
güneşli	ılık	normal	kuvvetli	Evet
bulutlu	ılık	yüksek	kuvvetli	Evet
bulutlu	sıcak	normal	Hafif	Evet
yağmurlu	Ilık	yüksek	kuvvetli	Hayır

8.6. SORULAR

3) *Bayes sınıflandırıcılarını* kullanarak aşağıda verilen örneğin hangi sınıfa ait olduğunu saptayınız?

HAVA = YAĞMURLU ISI = ILIK NEM = NORMAL RÜZGAR = HAFİF OYUN = ?

BİRLİKTELİK KURALI

- Birliktelik kuralı, nesnelerin veya niteliklerin bir arada olma durumlarını belirlemede kullanılan bir yöntemdir.
- İşlemlerden oluşan ve her bir işlemin de ürünlerin birlikteliğinden oluştuğu düşünülen bir veri tabanında, bütün ürün birlikteliklerinin tarayarak, sık tekrarlanan ürün birlikteliklerinin veri tabanından ortaya çıkarılmasıdır.

BİRLİKTELİK KURALI

- Market sepet verisi üzerinde birliktelik kuralı problemi, ilk olarak 1993 yılında ele alınmıştır. Sepet analizinde amaç, nitelikler (ürün satışları) arasındaki ilişkiyi bulmaktır. Bu ilişkilerin bilinmesi şirketin kârını arttırmak için kullanılabilir. Eğer X malını alan müşterilerin Y malını da çok yüksek bir olasılıkla aldıkları biliniyorsa o müşteriler potansiyel bir Y müşterisidir. Sepet analizi günlük işlemler sonucu elde edilen verilerden anlamlı bağıntılar çıkarmada kullanılır.
- $\Phi(D)=<X \rightarrow Y,c,s>$ şeklinde ifade edilir.

Destek ve Güven

- Destek (Support) : $X \rightarrow Y =$

$$\frac{\text{X ve Y ürünlerini satın alan müşterilerin sayısı}}{\text{Toplam Müşteri Sayısı}}$$

- Güven (Confidence) : $X \rightarrow Y =$

$$\frac{P(X \text{ ve } Y)\ \{X \text{ ve Y ürünlerini satın alanların sayısı}\}}{P(X)\ \{X \text{ ürününü satın alanların Sayısı}\}}$$

Örnek :

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

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

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

İşlemler, satın alınan ürünler, destek değerleri ve güven değerleri

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

Apriori Algoritması

C1

Her bir aday sayısı D veritabanından sayılır.

Ürünler	Destek Sayısı
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

L1

Her adayın Destek değerleri minimum destek değeri ile karşılaştırılır.

Ürünler	Destek Sayısı
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Apriori Algoritması

L1'den C2 aday kümeleri üretilir →

C2

Ürünler
{I1,I2}
{I1,I3}
{I1,I4}
{I1,I5}
{I2,I3}
{I2,I4}
{I2,I5}
{I3,I4}
{I3,I5}
{I4,I5}

Her bir aday D'den sayılır. →

C2

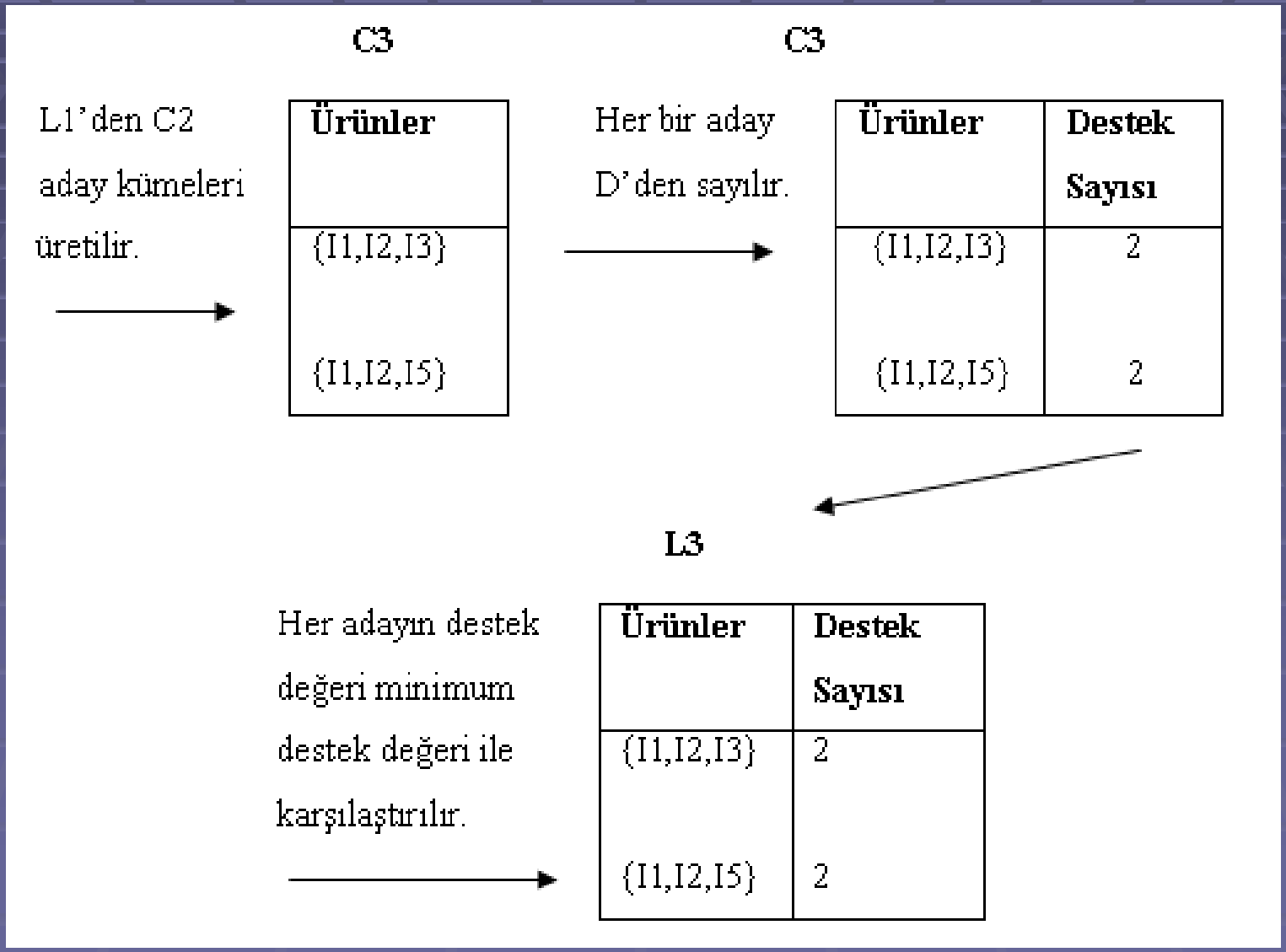
Ürünler	Destek Sayısı
{I1,I2}	4
{I1,I3}	4
{I1,I4}	1
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2
{I3,I4}	0
{I3,I5}	1
{I4,I5}	0

Her adayın destek değeri min. destek değeri ile karşılaştırılır. →

L2

Ürünler	Destek Sayısı
{I1,I2}	4
{I1,I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

Apriori Algoritması



L2 Kullanılarak C3 (Üç Elemanlı Aday Öğe Kümesi) Oluşturulması

- Birleştirme: C3=L2xL2 =

{{I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}}

X

{{I1,I2}, {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}}

=

{{I1,I2,I3}, {I1,I2,I5}, {I1,I3,I5}, {I2,I3,I4}, {I2,I3,I5}, {I2,I4,I5}}.

L2 Kullanılarak C3 (Üç Elemanlı Aday Öğe Kümesi) Oluşturulması

- Budama :
- {{I1,I2,I3}} ün iki elemanlı alt kümeleri {I1,I2}, {I1,I3} ve {I2,I3} tür. Bu iki elamanlı alt kümelerin tümü L2'nin bir öğesi olduğuna göre {{I1,I2,I3}} C3'ün bir aday öğesi olabilir.
- {{I1,I3,I5}} in iki elemanlı alt kümeleri {I1,I3}, {I1,I5} ve {I3,I5} tir. Bu iki elamanlı alt kümelerden {I3,I5}, L2'nin bir öğesi değildir ve bu nedenle sık tekrarlı değildir. Bundan dolayı {{I1,I3,I5}} C3'ün aday öğeliğinden çıkarılmalıdır.

Kuralların Elde Edilmesi

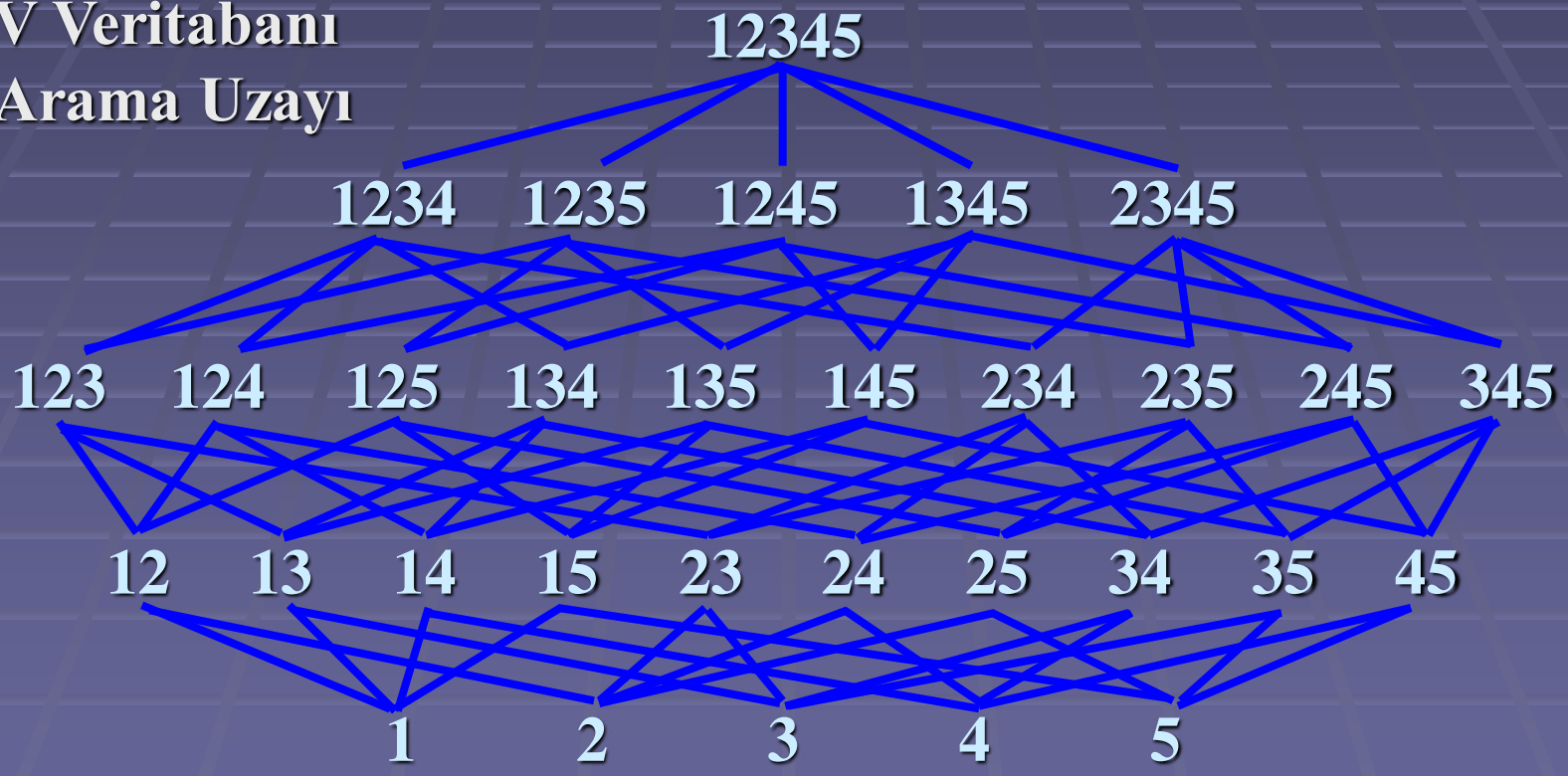
- {I1,I2,I5} den elde edilen kurallar :

$$\text{Güven } (A \Rightarrow B) : P(B \backslash A) = \frac{\textit{Destek değeri } (A \cup B)}{\textit{Destek değeri } (A)}$$

- (1) $I1 \wedge I2 \Rightarrow I5$, güven = 2 / 4 = % 50
- (2) $I1 \wedge I5 \Rightarrow I2$, güven = 2 / 2 = % 100
- (3) $I2 \wedge I5 \Rightarrow I1$, güven = 2 / 2 = % 100
- (4) $I1 \Rightarrow I2 \wedge I5$, güven = 2 / 6 = % 33
- (5) $I2 \Rightarrow I1 \wedge I5$, güven = 2 / 7 = % 29
- (6) $I5 \Rightarrow I1 \wedge I2$, güven = 2 / 2 = % 100

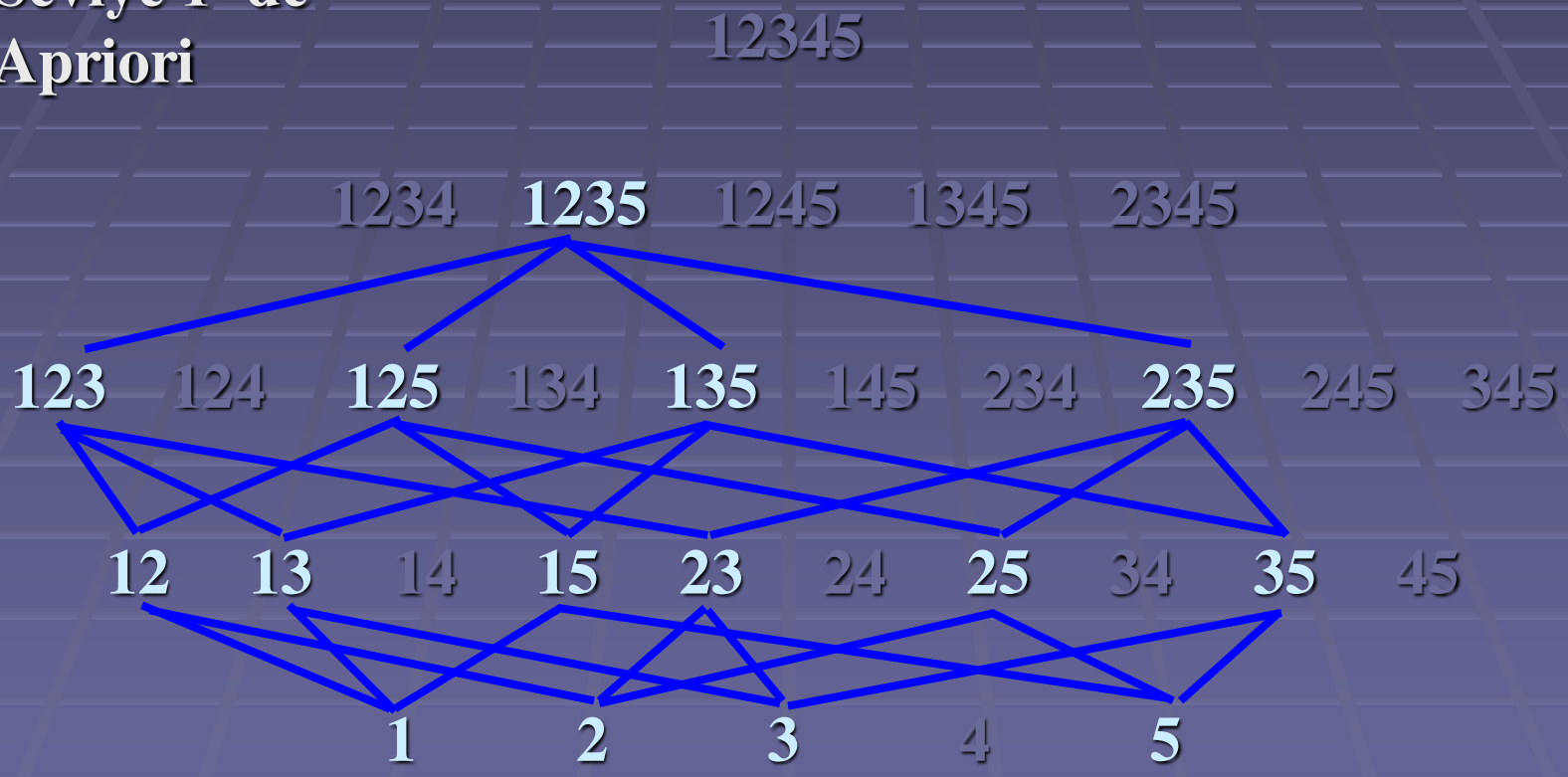
Apriori

V Veritabanı
Arama Uzayı



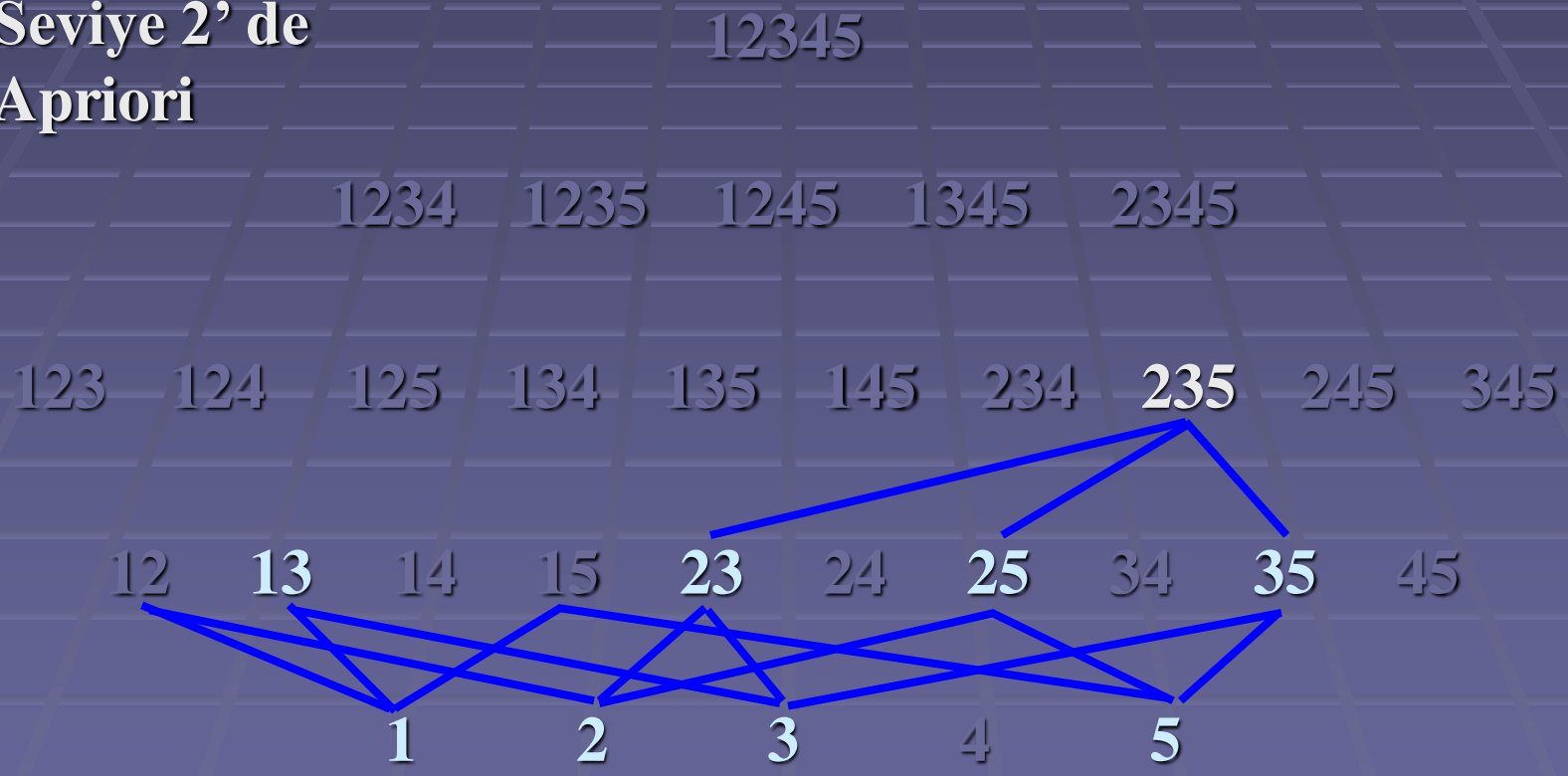
Apriori

Seviye 1' de Apriori



Apriori

Seviye 2' de Apriori



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. Ck çok 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	$\leq N$	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
- Sınırlandırılmış Birliktelik Kuralları
- Nicel Birliktelik Kuralları
- Sıralı Örüntüler
- Periyodik Kurallar
- Ağırlıklandırılmış Birliktelik Kuralları
- Negatif Birliktelik Kuralları

6.BÖLÜM

KÜMELEME

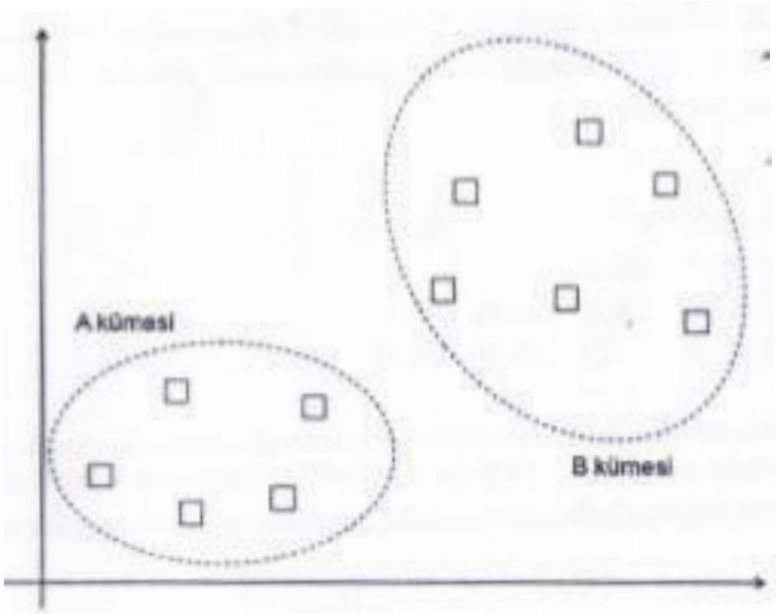
İÇERİK

KÜMELEME

Kümeleme birbirlerine benzeyen veri parçalarını ayırma işlemidir ve küme yöntemlerinin çoğu veri arasındaki uzaklıkları kullanır.

6.1.KÜMELEME ÇÖZÜMLEMESİ

Kümeleme Çözümlemesi, verilerin birbirleriyle benzer alt kümelere ayırma işlemidir.Çok sayıda kümeleme yöntemi kullanılmaktadır.Bu yöntemler değişkenler arasındaki benzerliklerden ya da farklılıklardan yararlanarak kümeyi alt kümelere ayırmakta kullanılmaktadır.



Şekil-6.1. A ve B gibi iki kümenin görünümü

6.2.UZAKLIK ÖLÇÜLERİ

- Kümeleme yöntemlerinin çoğu gözlem değerleri arasındaki uzaklıkların hesaplanması esasına dayanmaktadır.O nedenle iki nokta arasındaki uzaklığı hesaplayan bağıntılara gereksinim vardır.

- Çeşitli değişkenlerden oluşan gözlem değerlerini bir X matrisi biçiminde gösterebiliriz.Örneğin üç değişken 5 gözlemden oluşan matris aşağıdaki gibi ifade edilebilir:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{bmatrix}$$

$X_{11}, x_{12}, x_{13} \longrightarrow$ 1.gözlem noktasının konumu

$X_{21}, x_{22}, x_{23} \longrightarrow$ 2.gözlem noktasının konumu

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

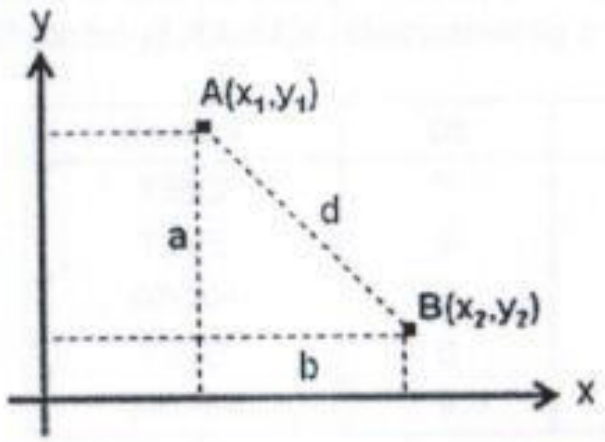
- X matrisinin her bir satırının diğerine olan uzaklığı d(i,j) olarak ifade edilecek olursa, simetrik D uzaklıklar matrisi şu şekilde yazılabilir:

$$D = \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & \textit{Simetrik} & \\ d(3,1) & d(3,2) & 0 & & \\ d(4,1) & d(4,2) & d(4,3) & 0 & \\ d(5,1) & d(5,2) & d(5,3) & d(5,4) & 0 \end{bmatrix}$$

Yukarıdaki matirisin üst kısmı alt kısmının simetriği olduğundan ayrıca yazılmamıştır.Bu durumda d(i,j)=d(j,i) olduğu kabul edilir.

- Kümelene çözümlemelerinde birçok uzaklık bağıntısı kullanılabilmektedir.Bu bölümde Öklit,Manhattan ve Minkowski uzaklıklarına değineceğiz.

6.2.1. ÖKLİT UZAKLIĞI



Şekil-6.2. İki nokta arasındaki d uzaklığı

- A ve B noktası arasındaki Öklit uzaklığı aşağıdaki gibidir:

$$d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

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

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

6.2.2. MANHATTAN UZAKLIĞI

- Manhattan uzaklığı, gözlemler arasındaki mutlak uzaklıkların toplamı alınarak hesaplanır

$$d(i,j) = \sum_{k=1}^{p} \left(\left| x_{ik} - x_{jk} \right| \right) \qquad i, j=1, 2...n;\ k=1, 2...p$$

6.2.3. MİNKOWSKİ UZAKLIĞI

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

$$d(i,j) = \left[\sum_{k=1}^{p} \left(\left| x_{ik} - x_{jk} \right|^{m} \right) \right]^{\frac{1}{m}} \qquad i,j=1, 2...n;\ k=1, 2...p$$

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:

Burada yer alan üç değişken için, i ve j gözlem noktaları ve p=3 olmak üzere Öklit uzaklık bağıntısını şu şekilde tanımlayabiliriz:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2}$$

İkinci gözlem ile birinci gözlem arasındaki uzaklık şu şekilde hesaplanır:

$$d(2,1) = \sqrt{(x_{21} - x_{11})^2 + (x_{22} - x_{12})^2 + (x_{23} - x_{13})^2}$$

$$= \sqrt{(4-2)^2 + (1-3)^2 + (3-1)^2} = 3.46$$

Üçüncü gözlem ile ikinci gözlem arasındaki uzaklık şu şekilde hesaplanır:

$$d(3,1) = \sqrt{(x_{31} - x_{11})^2 + (x_{32} - x_{12})^2 + (x_{33} - x_{13})^2}$$
$$= \sqrt{(5-2)^2 + (7-3)^2 + (3-1)^2} = 5.39$$

Herbir gözlem arasındaki Öklit uzaklıkları hesaplandığında aşağıdaki sonuçlar elde edilir:

Gözlem	1	2	3	4	5
1	0.00				
2	3.46	0.00			
3	5.39	6.08	0.00		
4	5.48	7.07	1.73	0.00	
5	7.28	8.31	3.46	3.32	0.00

Manhattan Uzaklığı

Örnekteki veriler kullanılarak 3 değişken için Manhattan uzaklığı aşağıdaki gibi hesaplanır:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + |x_{i3} - x_{j3}|$$

İkinci gözlem ile birinci gözlem arasındaki Manhattan uzaklığı aşağıdaki gibidir:

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

Üçüncü gözlem ile ikinci gözlem arasındaki Manhattan uzaklığı aşağıdaki gibidir:

$$d(3,1) = |5-2| + |7-3| + |3-1| = 9$$

Tüm gözlemlerin birbirlerine olan uzaklıkları hesaplandığında aşağıdaki sonuçlar elde edilir:

Gözlem	1	2	3	4	5
1	0.00				
2	6.00	0.00			
3	9.00	7.00	0.00		
4	8.00	8.00	3.00	0.00	
5	11.00	11.00	6.00	5.00	0.00

Minkowski Uzaklığı

Örnekteki veriler kullanılarak 3 değişken için Minkowski uzaklığı aşağıdaki gibi hesaplanır:

$$d(i,j) = \left[\left| x_{i1} - x_{j1} \right|^m + \left| x_{i2} - x_{j2} \right|^m + \left| x_{i3} - x_{j3} \right|^m \right]^{1/m}$$

m=3 varsayımı altında ikinci gözlem ile birinci gözlem arasındaki Minkowski uzaklığı:

$$d(2,1) = \left[\left| x_{21} - x_{11} \right|^3 + \left| x_{22} - x_{12} \right|^3 + \left| x_{23} - x_{13} \right|^3 \right]^{1/3}$$

$$= \left[\left| 4-2 \right|^3 + \left| 1-3 \right|^3 + \left| 3-1 \right|^3 \right]^{1/3} = 2.88$$

m=3 varsayımı altında üçüncü gözlem ile ikinci gözlem arasındaki Minkowski uzaklığı:

$$d(3,1) = \left[\left| x_{31} - x_{11} \right|^3 + \left| x_{32} - x_{12} \right|^3 + \left| x_{33} - x_{13} \right|^3 \right]^{1/3}$$

$$= \left[\left| 5-2 \right|^3 + \left| 7-3 \right|^3 + \left| 3-1 \right|^3 \right]^{1/3} = 4.63$$

Tüm gözlem arasındaki uzaklıklar hesaplandığında aşağıdaki sonuçlar elde edilir:

Gözlem	1	2	3	4	5
1	0.00				
2	2.88	0.00			
3	4.63	6.01	0.00		
4	5.12	7.01	1.44	0.00	
5	6.55	8.05	2.88	3.07	0.00

6.3.HIYERARŞIK KÜMELEME

Kümelerin bir ana küme olarak ele alınması ve sonra aşamalı olarak içerdiği alt kümelere ayrılması veya ayrı ayrı ele alınan kümelerin aşamalı olarak bir küme biçiminde birleştirilmesi esasına dayanır.

6.3.1.Birleştirici Hiyerarşik Yöntemler

Ayrı ayrı ele alınan kümelerin aşamalı olarak birleştirilmesini sağlayan yöntemlerdir. Bu yöntemlerden aşağıda belirtilenleri ele alarak inceleyeceğiz:

A) En yakın komşu algoritması

B) En uzak komşu algoritması

A) En Yakın Komşu Algoritması

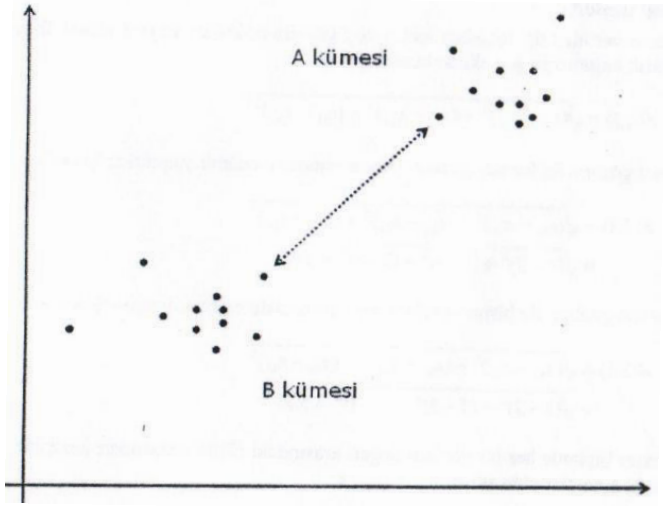
Başlangıçta tüm gözlem değerleri birer küme olarak değerlendirilir.Adım adım bu kümeler birleştirilerek yeni kümeler elde edilir.

Bu yöntemde öncelikle komşular arasındaki uzaklıklar belirlenir.i ve j gözlem noktaları arasındaki uzaklıkların hesaplanmasında öklit bağıntısı kullanılabilir.

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

Uzaklıklar göz önüne alınarak Min d(i,j) seçilir.Söz konusu uzaklıkla ilgili satırlar birleştirilerek yeni bir küme elde edilir.Bu duruma göre uzaklıkların yeniden hesaplanması gerekir

Birden fazla gözlem değerine sahip olan iki küme arasındaki uzaklığın belirlenmesi gerektiğinde farklı bir yol izlenir.İki kümenin içerdiği gözlemler arasında **"birbirine en yakın olanların uzaklığı"** iki kümenin birbirine olan uzaklığı olarak ifade edilir.



ÖRNEK

Aşağıdaki tabloda verilen 5 adet gözlemi göz önüne alalım.Bu veriler üzerinde **en yakın komşu algoritması**nı kullanarak kümeleme işlemlerini yapmak istiyoruz.

Gözlemler	X_1	X_2
1	4	2
2	6	4
3	5	1
4	10	6
5	11	8

ADIM 1 :Öncelikle uzaklık tablosunun (matrisinin) hesaplanması gerekiyor. Bunun için Öklit uzantı ölçüsünü kullanalım.

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

Bu formül yardımıyla aşağıdaki hesaplamalar yapılır daha sonra gözlemlere ilişkin uzaklıklar matrisi çıkartılır.

$$d(1,2) = \sqrt{(4-6)^2 + (2-4)^2} = 2.83$$

$$d(1,3) = \sqrt{(4-5)^2 + (2-1)^2} = 1.41$$

$$d(1,4) = \sqrt{(4-10)^2 + (2-6)^2} = 7.21$$

$$d(1,5) = \sqrt{(4-11)^2 + (2-8)^2} = 9.22$$

$$d(2,3) = \sqrt{(6-5)^2 + (4-1)^2} = 3.16$$

$$d(2,4) = \sqrt{(6-10)^2 + (4-6)^2} = 4.47$$

$$d(2,5) = \sqrt{(6-11)^2 + (4-8)^2} = 7.20$$

$$d(3,4) = \sqrt{(5-10)^2 + (1-6)^2} = 7.07$$

$$d(3,5) = \sqrt{(5-11)^2 + (1-8)^2} = 9.22$$

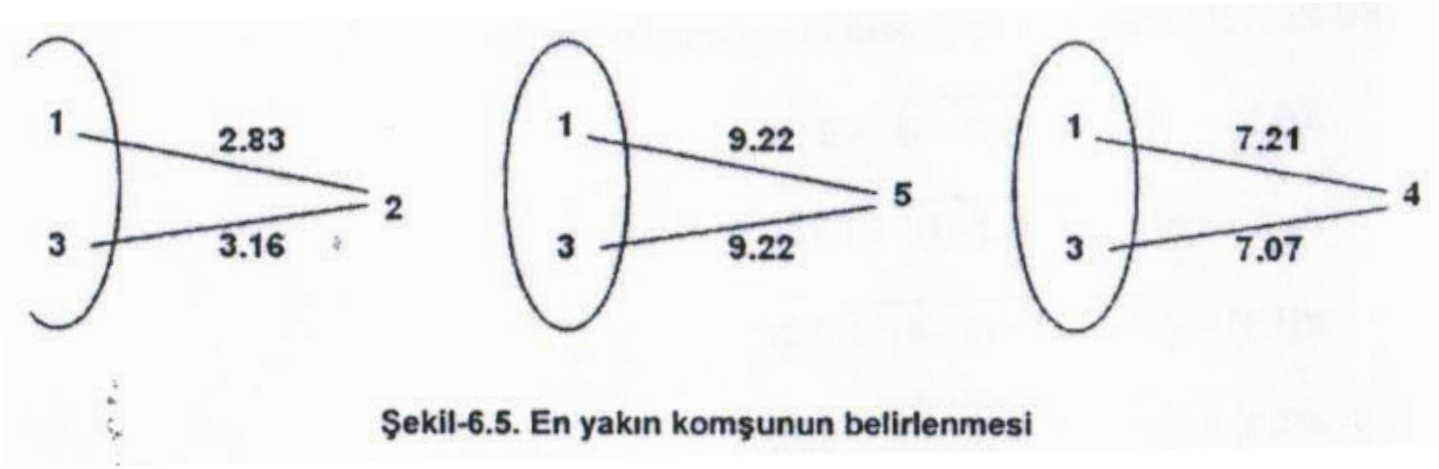
$$d(4,5) = \sqrt{(10-11)^2 + (6-8)^2} = 2.24$$

→

Tablo-6.6. Uzaklıklar tablosu

Gözlemler	1	2	3	4	5
1					
2	2.83				
3	1.41	3.16			
4	7.21	4.47	7.07		
5	9.22	7.20	9.22	2.24	

ADIM 2: Uzaklıklar tablosunda *Min d(i,j)* hücresinin belirlenmesi gerekiyor.Tablo incelendiğinde *Min d(i,j)*=1.41 olduğu görülür.O halde bu değerin ilgili olduğu 1 ve 3 numaralı gözlemler ele alınır.Bu iki değer birleştirilerek (1,3) kümesi elde edilir.Şimdi elde edilen bu kümeye göre uzaklıklar matrisini yeniden gözden geçirelim.Bu amaçla (1,3) kümesi ile 2,4 ve 5 numaralı gözlemler arasındaki uzaklıkları belirleyelim.



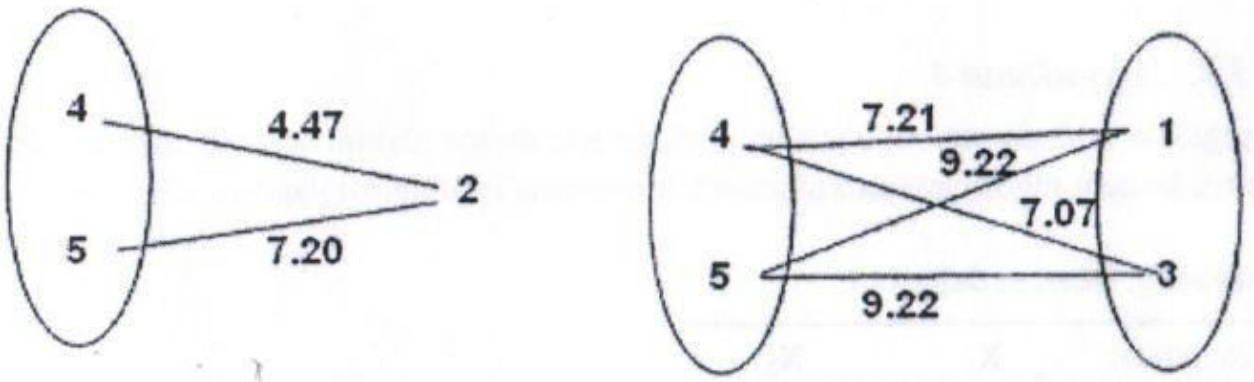
Şekil-6.5. En yakın komşunun belirlenmesi

Yapılan hesaplamalar ile yeni uzaklıklar tablosu aşağıdaki gibi olur

lo-6.7. Uzaklıklar tablosu

özlemler	(1,3)	2	4	5
(1,3)				
2	2.83			
4	7.07	4.47		
5	9.22	7.20	2.24	

ADIM 3: Uzaklıklar tablosu incelendiğinde *Min d(i,j)*=2,24 olarak hesaplanır.O halde bu değerin ilgili olduğu 4 ve 5 kümeleri birleştirilerek (4,5) biçiminde bir küme oluşturulacaktır.Elde edilen (4,5) kümesinin diğer (1,3) ve 2 gözlemi ile olan uzaklıklarını belirlemek gerekiyor.



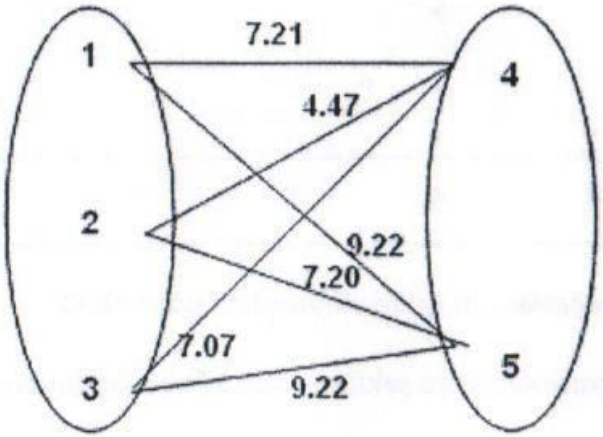
Şekil-6.6. En yakın komşunun belirlenmesi

Bu durumda uzaklıklar tablosu aşağıdaki gibi olur:

Tablo-6.8. Uzaklıklar tablosu

Gözlemler	(1,3)	2	(4,5)
(1,3)			
2	2.83		
(4,5)	7.07	4.47	

ADIM 4: Uzaklıklar tablosu incelendiğinde Min d(i,j)=2.83 olduğu görülür.O halde bu uzaklık ile ilgili 2 gözlemi ile (1,3) kümesi birleştirilecektir.Elde edilen (1,2,3) kümesi ile (4,5) kümesi arasındaki uzaklığı belirlemek için kümeler içindeki her bir değeri eşliyoruz ve aralarında en küçük olanı belirliyoruz.



Şekil-6.7. En yakın komşunun belirlenmesi

Yeni uzaklık değerini de içeren uzaklıklar tablosu şu şekildedir:

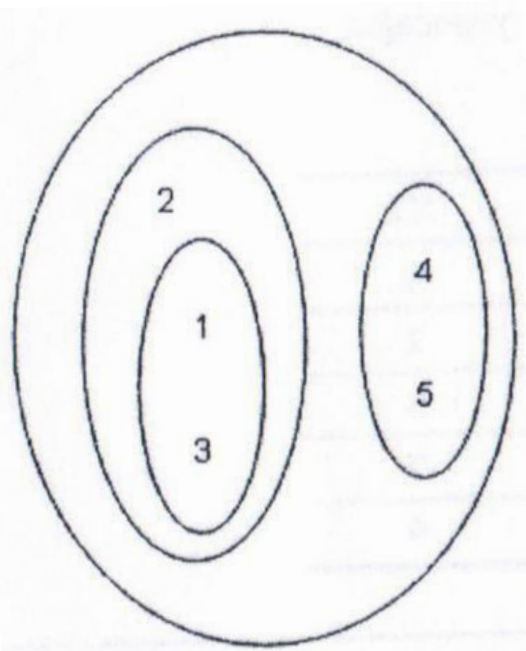
Tablo-6.9. Uzaklıklar tablosu

Gözlemler	(1,2,3)	(4,5)
(1, 2, 3)		
(4, 5)	4.47	

ADIM 5: 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.

Tablo-6.10.

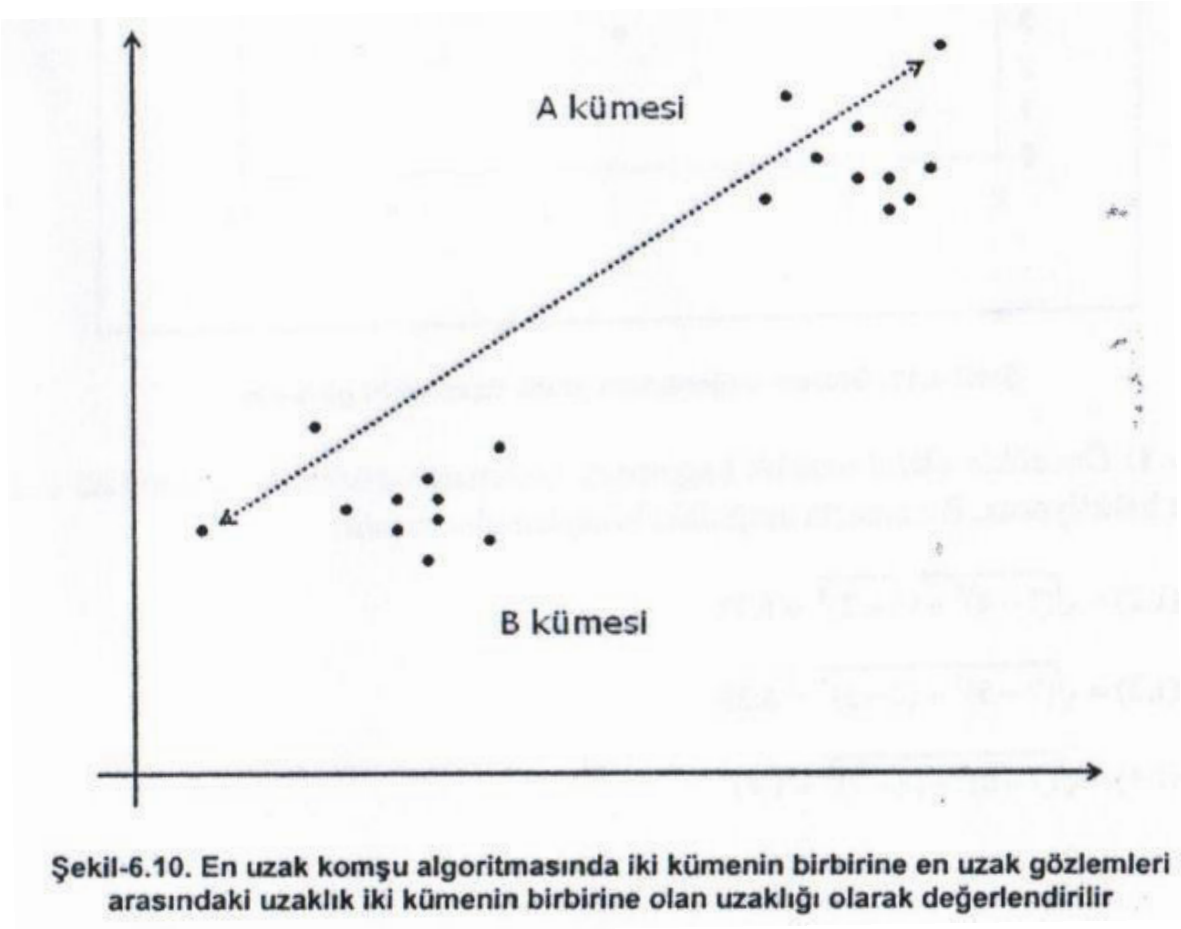
Uzaklık	Kümeler
1.41	(1, 3)
2.24	(4, 5)
2.83	(1, 2, 3)
4.47	(1, 2, 3, 4, 5)



Şekil-6.9. Kümeler

B) 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.



Şekil-6.10. En uzak komşu algoritmasında iki kümenin birbirine en uzak gözlemleri arasındaki uzaklık iki kümenin birbirine olan uzaklığı olarak değerlendirilir

ÖRNEK

Aşağıdaki gözlem değerlerini göz önüne alalım.Bu kez kümeleme analizini "en uzak komşu algoritmasına" göre yapacağız.

Io-6.11. Gözlem değerleri

özlemler	X1	X2
1	7	8
2	4	2
3	5	3
4	8	7
5	9	9

ADIM 1: Öncelikle Öklıt uzaklık bağıntısını kullanarak gözlemler arasındaki uzaklıkları belirliyoruz.Bu amaçla aşağıdaki hesaplamalar yapılır:

$$d(1,2)=\sqrt{(7-4)^2+(8-2)^2}=6.71$$

$$d(1,3)=\sqrt{(7-5)^2+(8-3)^2}=5.39$$

$$d(1,4)=\sqrt{(7-8)^2+(8-7)^2}=1.41$$

$$d(1,5)=\sqrt{(7-9)^2+(8-9)^2}=2.24$$

$$d(2,3)=\sqrt{(4-5)^2+(2-3)^2}=1.41$$

$$d(2,4)=\sqrt{(4-8)^2+(2-7)^2}=6.40$$

$$d(2,5)=\sqrt{(4-9)^2+(2-9)^2}=8.60$$

$$d(3,4)=\sqrt{(5-8)^2+(3-7)^2}=5.00$$

$$d(3,5)=\sqrt{(5-9)^2+(3-9)^2}=7.21$$

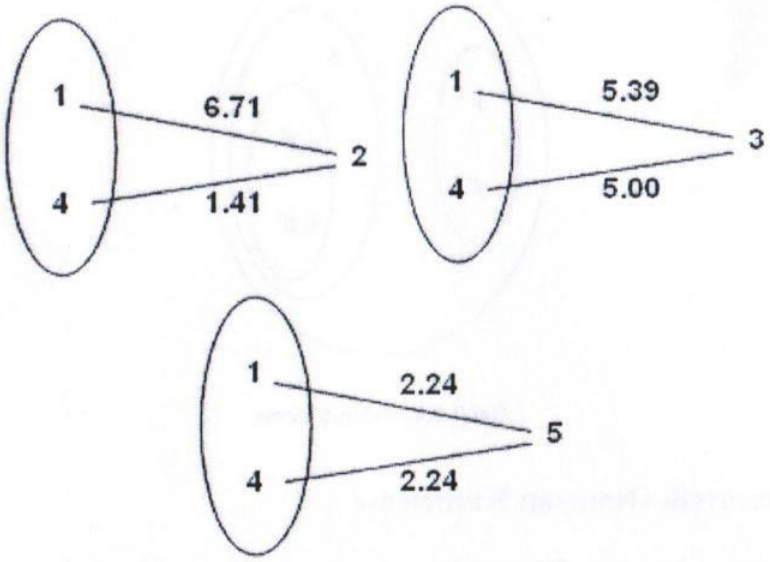
$$d(4,5)=\sqrt{(8-9)^2+(7-9)^2}=2.24$$

Gözlemlere ilişkin uzaklık matrisi şu şekilde olacaktır:

Tablo-6.12. Uzaklıklar tablosu

Gözlemler	1	2	3	4	5
1					
2	6.71				
3	5.39	1.41			
4	1.41	6.40	5.00		
5	2.24	8.60	7.21	2.24	

ADIM 2: Uzaklıklar tablosunda Min d(i,j)=1.41 olduğu görülür.Bu değere sahip iki hücre bulunmaktadır.Bu değere sahip (1,4) ve (2,3) hücreleri bulunmaktadır.Herhangi birini seçebiliriz. Biz 1 ve 4 numaralı gözlemleri ele alarak (1,4) kümesini oluşturalım.(1,4) kümesi ile 2,3 ve 5 numaralı gözlemler arasındaki uzaklıkları hesaplayarak **birbirine en uzak olan gözlemler** seçilir.



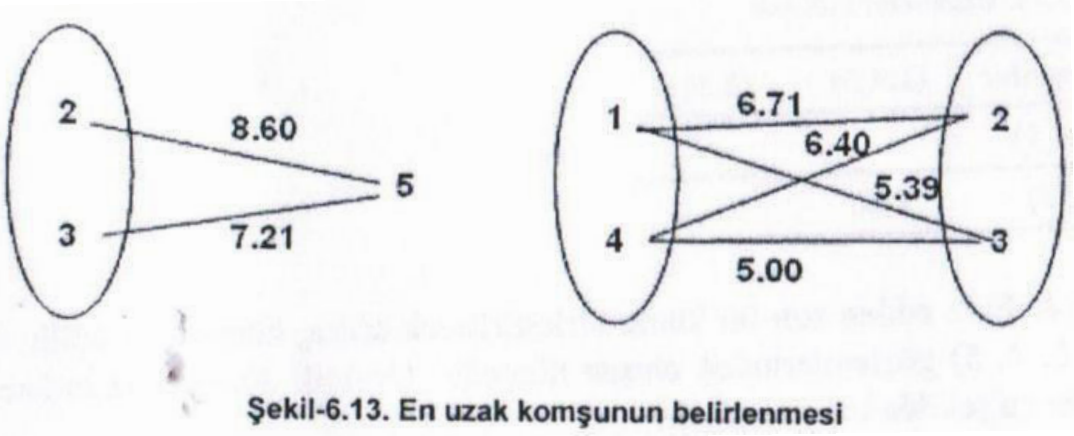
Şekil-6.12. En uzak komşunun belirlenmesi

Hesaplamalar yapıldığında uzaklık tablosu aşağıdaki gibi olur:

Tablo-6.13. Uzaklıklar tablosu

Gözlemler	(1,4)	2	3	5
(1,4)				
2	6.71			
3	5.39	1.41		
5	2.24	8.60	7.21	

ADIM 3: Uzaklıklar tablosunda Min d(i,j)=1.41 olduğu görülür.O halde 2 ve 3 gözlemleri birleştirilerek bir küme oluşturulur.Elde (2,3) kümesinin diğer (1,4) kümesi ve 5 gözlemleri arasındaki uzaklıklar hesaplandığında aşağıdaki gibi yeni uzaklıklar tablosu çıkartılır.

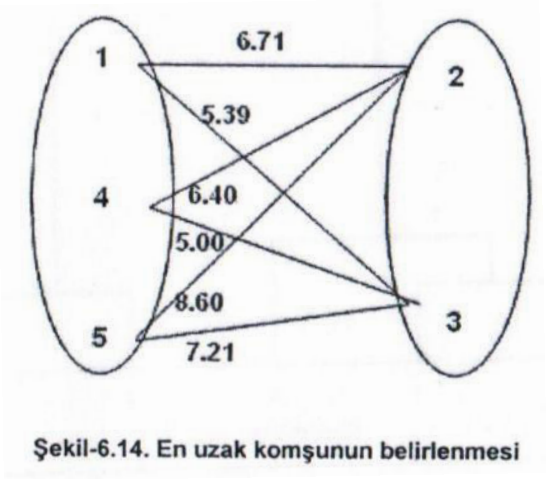


Şekil-6.13. En uzak komşunun belirlenmesi

Tablo-6.14. Uzaklıklar tablosu

Gözlemler	(1,4)	(2,3)	5
(1, 4)			
(2, 3)	6.71		
5	2.24	8.60	

ADIM 4: Uzaklıklar tablosunda Min d(i,j)=2.24 olduğu görülür. O halde 5 gözlemi ile (1,4) birleştirilerek (1,4,5) adında yeni bir küme oluşturulur.Elde (1,4,5) kümesi ile (2,3) kümesi arasındaki uzaklığı belirlemek için küme içindeki her bir değeri eşliyoruz ve aralarında en büyük olanı belirliyoruz.



Şekil-6.14. En uzak komşunun belirlenmesi

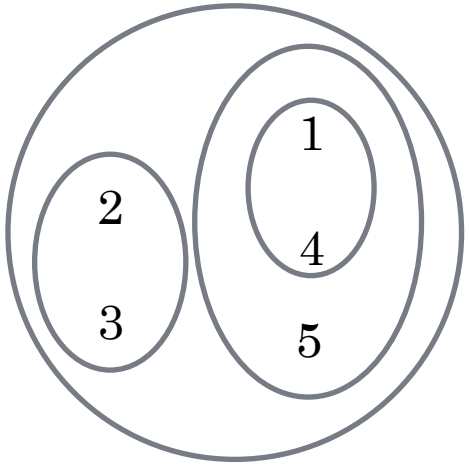
blo-6.15. Uzaklıklar tablosu

Gözlemler	(1,4,5)	(2,3)
(1,4,5)		
(2,3)	8.60	

ADIM 5: 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:

blo-6.16.

Uzaklık	Kümeler
1.41	(1, 4)
1.41	(2, 3)
2.24	(1, 4, 5)
8.60	(1, 2, 3, 4, 5)



6.4. HİYERARŞİK OLMAYAN KÜMELEME

6.4.1. 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 { C_1,C_2,…..,C_k} biçiminde K kümeye ayrılsın.

O zaman $\sum n_k = N \ (k=1,2,..,k)$ olmak üzere C_k kümesinin ortalama vektörü M_k şu şekilde hesaplanır.

$$M_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik}$$

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

$$e_i^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2$$

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

$$E_t^2 = \sum_{k=1}^{K} e_k^2$$

Kare-hata kümeleme yönteminin amacı, verilen K değeri için E_k^2 değerini minimize eden K kümelerini bulmaktır.O halde k-ortalama algoritmasında E_k^2 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 $\{C_1, C_2, \ldots, C_k\}$ kümeleri belirlenmiş olur.Ardından aşağıdaki işlemler gerçekleştirilir:

a) Her bir kümenin merkezi belirlenir.Bu merkezler $M_1, M_2, .., M_k$ biçimindedir.

b) $e_1, e_2, \ldots, e_k$ küme içi değişmeler hesaplanır.Bu değişmelerin toplamı olan E_k^2 değeri bulunur.

c) M_k merkez değerleri ile gözlem değerleri arasındaki uzaklıklar hesaplanır.Bu gözlem değeri hangi merkeze yakın ise, o merkez ile ilgili küme içine dahil edilir.

d) 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.

Tablo-6.17. Gözlem değerleri

Gözlemler	Değişken1	Değişken2
X_1	4	2
X_2	6	4
X_3	5	1
X_4	10	6
X_5	11	8

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

$C_1=\{X_1,X_2,X_4\}$

$C_2=\{X_3,X_5\}$

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

Tablo- 6.18. Gözlem değerleri

Gözlemler	Değişken1	Değişken2	Küme üyeliği
X_1	4	2	C_1
X_2	6	4	C_1
X_3	5	1	C_2
X_4	10	6	C
X_5	11	8	C_2

ADIM 1:

İki kümenin merkezleri şu şekilde hesaplanır:

$$M_1 = \left\{\frac{4+6+10}{3}, \frac{2+4+6}{3}\right\} = \{6.67,\ 4.0\}$$

$$M_2 = \left\{\frac{5+11}{2}, \frac{1+8}{2}\right\} = \{8.00,\ 4.50\}$$

Küme içi değişmeler şu şekilde hesaplanır:

$$e_1^2 = \left[(4-6.67)^2 + (2-4.00)^2\right] + \left[(6-6.67)^2 + (4-4.00)^2\right] + \left[(10-6.67)^2 + (6-4.00)^2\right] = 26.67$$

$$e_2^2 = \left[(5-8)^2 + (1-4.50)^2\right] + \left[(11-8)^2 + (8-4.50)^2\right] = 42.50$$

Toplam kare hata aşağıdaki gibi hesaplanır:

$$E^2 = e_1^2 + e_2^2 = 26.67 + 42.50 = 69.17$$

Gözlemlerin M_1 ve M_2 merkezlerinden olan uzaklıkların minimum olması istendiğinden aşağıdaki hesaplamalar yapılır:

$$d(M_1, X_1) = \sqrt{(6.67-4)^2 + (4-2)^2} = 3.33$$

$$d(M_2, X_1) = \sqrt{(8-4)^2 + (4.5-2)^2} = 4.72$$

$d(M_1,X_1) < d(M_2,X_1)$ olduğundan M_1 merkezinin X_1 gözlem değerine daha yakın olduğu anlaşılır.O halde $X_1 \in C_1$ olarak kabul edilir.

Tüm gözlem değerleri için hesaplanarak aşağıdaki tablo elde edilir:

Tablo-6.19.

Gözlemler	M_1 den uzaklık	M_2 den uzaklık	Küme üyeliği
X_1	$d(M_1, X_1) = 3.33$	$d(M_2, X_1) = 4.72$	C_1
X_2	$d(M_1, X_2) = 0.67$	$d(M_2, X_2) = 2.06$	C_1
X_3	$d(M_1, X_3) = 3.43$	$d(M_2, X_3) = 4.61$	C_1
X_4	$d(M_1, X_4) = 3.89$	$d(M_2, X_4) = 2.50$	C_2
X_5	$d(M_1, X_5) = 5.90$	$d(M_2, X_5) = 4.61$	C_2

$$\rightarrow \quad C_1 = \{X_1, X_2, X_3\}, \quad C_2 = \{X_4, X_5\}$$

ADIM 2:

Yukarıdaki iki kümenin merkezleri şu şekilde hesaplanır:

$$M_1 = \left\{\frac{4+6+5}{3}, \frac{2+4+1}{3}\right\} = \{5,\ 2.33\}$$

$$M_2 = \left\{\frac{10+11}{2}, \frac{6+8}{2}\right\} = \{10.5,\ 7\}$$

$\rightarrow$

Küme içi değişmeler şu şekilde hesaplanır:

$$e_1^2 = \left[(4-5)^2 + (2-2.33)^2\right] + \left[(6-5)^2 + (4-2.33)^2\right] + \left[(5-2.33)^2 + (1-2.33)^2\right] = 9.33$$

$$e_2^2 = \left[(10-10.5)^2 + (6-7)^2\right] + \left[(11-10.5)^2 + (8-7)^2\right] = 2.50$$

Toplam kare hata aşağıdaki gibi hesaplanır:

$$E^2 = e_1^2 + e_2^2 = 9.33 + 2.5 = 11.83$$

Bu değer bir önceki iterasyonda elde edilen $E^2=69.17$ değerinden daha küçük olduğu anlaşılır

M_1 ve M_2 merkezinden gözlem değerlerine olan uzaklıklar hesaplandığında aşağıdaki tablo elde edilir.

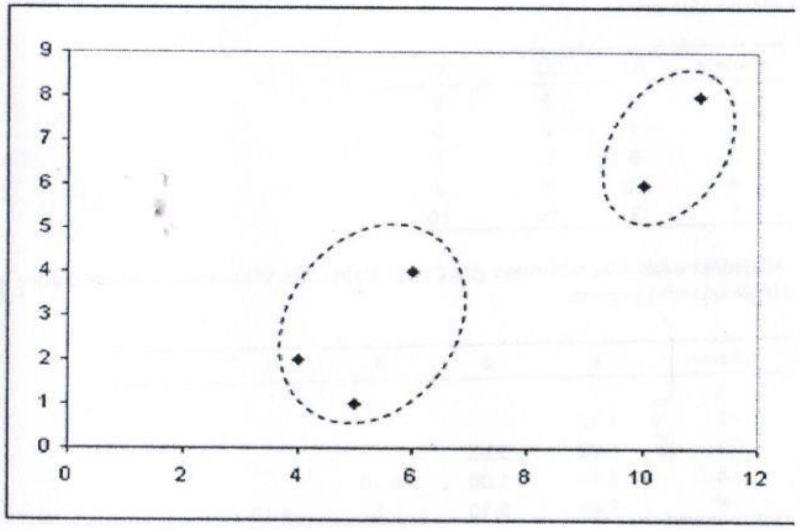
Tablo-6. 20.

Gözlemler	M_1 den uzaklık	M den uzaklık	Küme üyeliği
X_1	$d(M_1, X_1) = 1.05$	$d(M_2, X_1) = 8.20$	C_1
X_2	$d(M_1, X_2) = 1.94$	$d(M_2, X_2) = 5.41$	C_1
X_3	$d(M_1, X_3) = 1.33$	$d(M_2, X_3) = 8.14$	C_1
X_4	$d(M_1, X_4) = 6.20$	$d(M_2, X_4) = 1.12$	C_2
X_5	$d(M , X_5) = 8.25$	$d(M_2, X_5) = 1.12$	C_2

→ $C_1 = \{X_1, X_2, X_3\}$

$C_2 = \{X_4, X_5\}$

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



Şekil-6.18. Sonuç olarak elde edilen kümeler

UYGULAMA-1:

Gözlem	X1	X2	X3
1	9	2	3
2	1	2	9
3	8	9	0
4	10	9	8
5	3	10	10

Yukarıdaki gözlem noktalarını göz önüne alarak *Öklit*,*Manhattan* ve *Minkowski* uzaklıklar matrisini elde ediniz

UYGULAMA-2:

Gözlem	1	2	3	4	5
1					
2	4.12				
3	9.22	5.83			
4	3.16	1.00	6.40		
5	2.24	5.10	8.94	4.12	

Yukarıdaki uzaklıklar matrisini göz önüne alarak;

a)En yakın komşu algoritmasına,

b)En uzak komşu algoritmasına,

göre kümeleme işlemini yapınız

UYGULAMA-3:

Gözlem	Değişken1	Değişken2
x_1	7	6
x_2	5	9
x_3	1	9
x_4	6	3
x_5	3	9

Yukarıdaki tabloda 5 adet gözlem üzerinde k-ortalamalar algoritmasını uygulayarak kümeleme işlemini yapınız.Başlangıçta iki kümeyi tesadüfi olarak $C_1=\{X_1,X_2\}$ ve $C_2=\{X_3,X_4,X_5\}$ biçiminde seçtiğimizi varsayalım.

KARAR AĞAÇLARI İLE SINIFLANIRMA

DOÇ. DR. MURAT KARABATAK

Karar Ağaçları İle Sınıflandırma

- Verinin içerdiği ortak özelliklere göre bir veri veya veri grubunun hangi sınıfa dahil olduğunun belirlenmesi işlemi sınıflandırma olarak adlandırılır.
- Karar ağaçları oluşturmak için temel olarak entropiye dayalı algoritmalar , sınıflandırma ve regresyon ağaçları, bellek tabanlı sınıflandırma modelleri biçiminde birçok yöntem geliştirilmiştir.

Sınıflandırma

- Sınıflandırma bir öğrenme algoritmasına dayanır. Öğrenmenin amacı bir sınıflandırma modelinin yaratılmasıdır.
- Diğer bir ifadeyle sınıflandırma, hangi sınıfa ait olduğu bilinmeyen bir kayıt için bir sınıf belirleme sürecidir.
- Örnek olarak; ”Ödemeleri üç gün içinde yapanlar” ve ”Ödemeleri üç günden sonra yapanlar” gibi sınıflandırma yapılabilir.

Sınıflandırma Süreci

- Verilerin sınıflandırma süreci iki adımdan oluşur;

a) İlk adım, veri kümelerine uygun bir modelin ortaya konulmasıdır. Bu model ,veri tabanındaki kayıtların nitelikleri kullanılarak gerçekleştirilir. Sınıflandırma modelinin elde edilebilmesi için veri tabanın bir kısmı eğitim verileri olarak kullanılır.

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

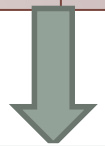
Sınırlayıcı Model

```
EĞER Borç= Yüksek ise
            Risk=KÖTÜ;
EĞER Borç= Düşük
            ve Gelir=DÜŞÜK ise Risk= KÖTÜ;
EĞER Borç= Düşük
            ve Gelir=YÜKSEK ise RİSK=İYİ;
```

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

Test verisi

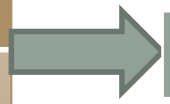
Müşteri	Borç	Gelir	Risk
Halit	Yüksek	Düşük	Kötü
Eser	Düşük	Yüksek	İyi
Selin	Düşük	Düşük	Kötü
Mehmet	Yüksek	Yüksek	Kötü



Sınıflayıcı Model

EĞER Borç= Yüksek ise
Risk=KÖTÜ;
EĞER Borç= Düşük
ve Gelir=DÜŞÜK ise Risk= KÖTÜ;
EĞER Borç= Düşük
ve Gelir=YÜKSEK ise RİSK=İYİ;

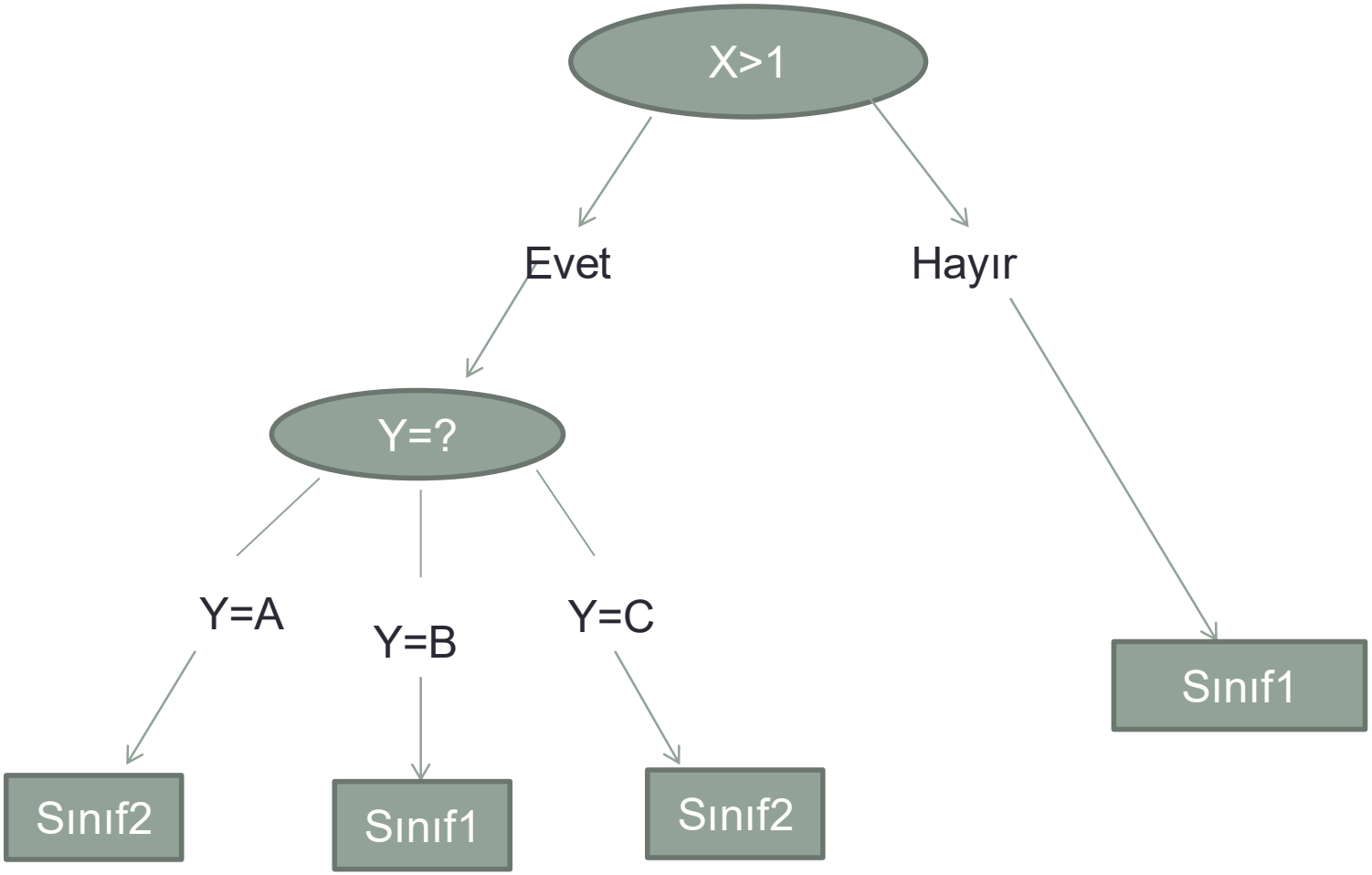
Müşteri	Borç	Gelir	Risk
Hakan	Düşük	YÜKSEK	?



Risk=İYİ

Karar Ağaçlarıyla Sınıflandırma

- Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir.
- Dallar ve yapraklar ağaç yapısının elemanlarıdır.
- En son yapı ”yaprak”, en üst yapı ”kök” ve bunların arasında kalan yapılar ise ”dal” olarak isimlendirilir.


X>1
Evet
Hayır
Y=?
Y=A
Y=B
Y=C
Sınıf2
Sınıf1
Sınıf2
Sınıf1

Karar Ağaçlarında Dallanma Kriterleri

- Karar ağaçlarında en önemli sorunlardan birisi herhangi bir kökten itibaren dallanmanın hangi kıstasa göre yapılacağıdır.
- Her farklı kriter için bir karar ağacı algoritması karşılık gelmektedir. Algoritmaları gruplandıracak olursak;

a. Entropiye Dayalı Algoritmalar
b. Sınıflandırma ve Regresyon ağaçları
c. Bellek tabanlı sınıflandırma algoritmaları

ID3 Algoritması

- Karar ağaçları yardımıyla sınıflandırma işlemlerini yerine getirmek üzere Quinlan tarafından birçok algoritma geliştirilmiştir.
- ID3 ve C4.5 algoritmaları entropi tabanlı algoritmalardır.

Entropi

- Bir sistemdeki belirsizliğin ölçüsüne entropi denir.
- Olasılık dağılımına sahip mesajları üreten S kaynağının entropisi şu şekildedir;

$$P=\left\{p_1,p_2,p_n\right\}$$

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Örnek:

Deney Sonuçları (S)	a_1	a_2	a_3
P_i	$1/2$	$1/3$	$1/6$

S= $\{a_1, a_2, a_3\}$ deney kümesini ifade etsin. a_1, a_2, a_3 olaylarının belirsizlikleri şöyle hesaplanır:

$$-\frac{1}{2}\log_2\frac{1}{2},\ -\frac{1}{3}\log_2\frac{1}{3},\ -\frac{1}{6}\log_2\frac{1}{6}$$

- Bu durumda toplam belirsizlik, yani entropi şu şekilde olacaktır:

$$H(S)=-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{3}\log_2\frac{1}{3}+\frac{1}{6}\log_2\frac{1}{6}\right)$$
$$=1.4591$$

Örnek:

- Aşağıda sekiz elemanlı S kümesini göz önüne alalım.

S=$\{evet, evet, hayır, hayır, hayır, hayır, hayır, hayır\}$

Olasılıklar, iki adet ”evet” değeri için,

$$P_1 = \frac{2}{8} = 0.25$$

Diğer altı adet ”hayır” değeri için,

$$P_2 = \frac{6}{8} = 0.75$$

S için toplam entropi şu şekilde elde edilir;

$$H(S) = -\{P_1 \log_2(P_1) + P_2 \log_2(P_2)\}$$

$$= -(0.25\log_2(0.25) + 0.75\log_2(0.75)) = 0.81128$$

Karar Ağaçlarında Entropi

- Karar ağaçlarının oluşturulması esnasında dallanmaya hangi nitelikten başlanacağı önem taşımaktadır. Çünkü sınırlı sayıda kayıttan oluşan bir eğitim kümesinden yararlanarak olası tüm ağaç yapılarını ortaya çıkarmak ve içlerinden en uygun olanı seçerek ondan başlamak kolay değildir.

- Veri tabanından eğitim için elde edilen kayıt kümesini ele alalım. Eğitim kümesi sınıf niteliğinin alacağı değerlere göre $\{C_1, C_2, \ldots, C_k\}$ olmak üzere k sınıfa ayrıldığını varsayalım. Bu sınıflarla ilgili olarak ortalama bilgi miktarına ihtiyaç duyulabilir.

- Burada T sınıf değerlerini içeren küme için P_T sınıfların olasılık dağılımıdır ve şu şekilde hesaplanır:

$$P_T = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|} \ldots, \frac{|C_k|}{|T|}\right)$$

$|C_i|$ ifadesi C_i kümesindeki elemanların sayısını vermektedir. Entropi şu şekilde ifade edilir;

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Örnek:

- Aşağıdaki on elemanlı RİSK kümesini göz önüne alalım.

RİSK=$\{var, var, var, yok, var, yok, yok, var, var, yok\}$

Burada C_1 sınıfı ”var” , C_2 sınıfı ”yok” değerlerini içersin. Bu durumda;

$$|C_1|=6 \quad |C_2|=4$$

Olduğuna göre, olasılıklar $P_1 = \frac{6}{10}$=0.6 ve $P_2 = \frac{4}{10}$ biçiminde hesaplanır ve olasılık dağılımı; $P_{RİSK}= \left(\frac{6}{10}, \frac{4}{10}\right)$

$$H(\text{RİSK}) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Eşitliği kullanarak RİSK kümesi için entropi şu şekilde hesaplanır:

$$H(\text{RİSK}) = -\left(\frac{6}{10}\log_2\frac{6}{10} + \frac{4}{10}\log_2\frac{4}{10}\right)$$

$$= 0.97$$

Dallanma İçin Niteliklerin Seçilmesi ve Kazanç Ölçütü

$$H(\mathrm{X},\mathrm{T}) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i)$$

T veri tabanı X testine göre bölmekle elde edilen bilgileri ölçmek için ”kazanç ölçütü” adı verilen bir ifadeye başvurulur. Bu ölçüt;

Kazanç(X,T)=H(T)-H(X,T)

Örnek: Tablo 3.1

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	Yüksek	Yüksek	İşveren	Kötü
2	Yüksek	Yüksek	Ücretli	Kötü
3	Yüksek	Düşük	Ücretli	Kötü
4	Düşük	Düşük	Ücretli	İyi
5	Düşük	Düşük	İşveren	Kötü
6	Düşük	Yüksek	İşveren	İyi
7	Düşük	Yüksek	Ücretli	İyi
8	Düşük	Düşük	Ücretli	İyi
9	Düşük	Düşük	İşveren	Kötü
10	Düşük	Yüksek	İşveren	İyi

- Hedef nitelik olan RİSK niteliğinin sınıf değerlerini şu şekilde gösterilir;

Risk=$\{kötü, kötü, kötü, iyi, kötü, iyi, iyi, iyi, kötü, iyi\}$

Risk kümesinde ‘’kötü’’ değerlerinin sayısı 5 olarak görülüyor. Buna karşılık ‘’iyi’’ değerlerinin de sayısı 5 olduğu görülüyor.

$$|C_1|=5 \quad |C_2|=5$$

$P_1 = \frac{5}{10} = \frac{1}{2}$ ve $P_2 = \frac{1}{2}$ olasılıkları hesaplanabilir. Risk kümesinin içerdiği ’’kötü’’ ve ’’iyi’’ değerleri için olasılık dağılımı;

$$P_{RİSK} = \left(\frac{1}{2}, \frac{1}{2}\right)$$

$$H(\text{RİSK}) = -\sum_{i=1}^{n} P_i \log_2(P_i)$$

olduğuna göre, RİSK niteliğinin entropisi;

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

Öncelikle BORÇ niteliğinin her bir değerden kaç tane içerdiği belirlenir.

$$|BORÇ_{Yüksek}| = 3$$
$$|BORÇ_{Düşük}| = 7$$

BORÇ niteliğinin RİSK hedef niteliğindeki karşılıklarına bakalım. BORÇ niteliğinin yüksek niteliği karşısında RİSK niteliğinin 3 adet kötü değeri karşılık olmaktadır. Buna karşılık, BORÇ niteliğinin 7 adet düşük değeri için RİSK üzerinden 5 adet iyi,2 adet kötü değeri karşılık gelmektedir.

$$H\left(BORÇ_{yüksek}\right) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

$$H\left(BORÇ_{düşük}\right) = -\left(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7}\right) = 0.863$$

Burada borç niteliğinin değerlerine göre bir dallanma yapmak istiyoruz. Böyle bir işlemin bize kazancı ne olacaktır? Söz konusu kazancı hesaplamak için bu niteliğin değerlerine göre entropilerini hesaplayacağız.

$$H(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|}\,\mathrm{H}(T_i)$$

Olduğuna göre ;

$H(BORÇ, RİSK)$=$\frac{3}{10}H\left(BORÇ_{yüksek}\right) + \frac{7}{10}H(BORÇ_{düşük})$=0.64

Kazanç ölçütü;

Kazanç(X,T)=H(T)-H(X,T) olduğuna göre,

$Kazanç(BORÇ, RİSK)$=1-0.64=0.36 elde edilir.

Uygulama (Tablo 3.2)

HAVA	ISI	NEM	RÜZGAR	OYUN
Güneşli	Sıcak	Yüksek	Hafif	Hayır
Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
Bulutlu	Sıcak	Yüksek	Hafif	Evet
Yağmurlu	Ilık	Yüksek	Hafif	Evet
Yağmurlu	Soğuk	Normal	Hafif	Evet
Yağmurlu	Soğuk	Normal	Kuvvetli	Hayır
Bulutlu	Soğuk	Normal	Kuvvetli	Evet
Güneşli	Ilık	Yüksek	Hafif	Hayır
Güneşli	Soğuk	Normal	Hafif	Evet
Yağmurlu	Ilık	Normal	Hafif	Evet
Güneşli	Ilık	Normal	Kuvvetli	Evet
Bulutlu	Ilık	Yüksek	Kuvvetli	Evet
Bulutlu	Sıcak	Normal	Hafif	Evet
Yağmurlu	Ilık	Yüksek	Kuvvetli	Hayır

- ID3 algoritması yardımıyla bir karar ağacı oluşturacağız. Burada OYUN niteliği hedef sınıf değerleri içermektedir. O halde OYUN nitelik değerlerinden oluşan küme T kümesi olarak kabul edilebilir.

$$OYUN = \{hayır, hayır, hayır, hayır, hayır, evet, evet, evet, evet, evet, evet, evet, evet\}$$

Burada C_1 sınıfı 'hayır', C_2 sınıfı ise 'evet' değerine uymaktadır. $|OYUN| = 14$, beş adet 'hayır' değeri için $|C_1|$=5 olmak üzere;

$$P_1 = \frac{5}{14}$$

Ve dokuz adet 'evet' değeri için olasılık, $|C_2|$=9 olmak üzere,

$$P_2 = \frac{9}{14}$$

Olasılık dağılımı şu şekilde yazılabilir;

$$P_{OYUN} = \left(\frac{5}{14}, \frac{9}{14}\right)$$

$$H(OYUN) = -\sum_{i=1}^{n} P_i \log_2(P_i)$$

Eşitliğini kullanarak OYUN kümesi için entropi hesaplanabilir.

$$H(OYUN) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right) = 0.940$$

Adım 1: Birinci dallanma

- Önce karar ağacının oluşturulması için hangi niteliğin seçileceği üzerinde duracağız. Bunun için elde edilen *H(OYUN)=0.940* entropi değeri göz önüne alınarak her bir nitelik için kazanç ölçütleri belirlenir.

a) ISI niteliği için kazanç ölçütü:

Burada önce ISI niteliğini ele alalım. Bu niteliğin her bir değeri için aşağıdaki değerleri yazabiliriz. Burada her değerin nitelik içinde tekrarlanma sayıları hesaplanmıştır.

$$|ISI_{soğuk}| = 4$$

$$|ISI_{ılık}| = 6$$

$$|ISI_{sıcak}| = 4$$

Bu durumda ISI niteliğine göre bir ayırma gerçekleştirildiğinde kazancın ne olacağını hesaplamak gerekiyor.

$$Kazanç = (X,T) = H(T) - H(X,T)$$

$$H(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i)$$

Bağıntısını ve öncelikle her bir nitelik değeri için entropiyi bulmak gerekiyor. Burada $|T_i|$ değeri, ISI niteliğinin her bir değerinden kaç tane olduğunu belirler. Yani $ISI_{soğuk}$ değerinin karşılığıdır.

$$H(ISI,OYUN) = \frac{4}{14} H\left(ISI_{soğuk}\right) + \frac{6}{14} H(ISI_{ılık}) + \frac{4}{14} H(ISI_{sıcak})$$

Tablo 3.3 Bir önceki (Tablo 3.2) de yer alan eğitim kümesinin ISI ve OYUN değerleri

ISI	OYUN
Soğuk	Evet
Soğuk	Hayır
Soğuk	Evet
Soğuk	Evet
Ilık	Evet
Ilık	Hayır
Ilık	Evet
Ilık	Evet
Ilık	Evet
Ilık	Hayır
Sıcak	Hayır
Sıcak	Hayır
Sıcak	Evet
Sıcak	Evet

- Tablo 3.3'e göre entropiler şöyle hesaplanır;

$$H(ISI_{soğuk}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right)=0.811$$

$$H(ISI_{ılık}) = -\left(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6}\right)=0.918$$

$$H(ISI_{sıcak}) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right)=1.00$$

Bu durumda $H(ISI, OYUN)$ entropisi şu şekilde hesaplanır;

$$H(ISI, OYUN) = \frac{4}{14}(0.811) + \frac{6}{14}(0.918) + \frac{4}{14}(1.00)=0.911$$

Önceki örnekte *H(OYUN)=0.940* hesaplamıştık. O halde kazanç ölçütü;

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN)$$
$$= 0.940\text{-}0.911$$
$$= 0.029$$

b) HAVA niteliği için kazanç ölçütü:

- İkinci adım olarak HAVA niteliği için benzer hesaplamalar yaparak kazanç ölçütüne ulaşmak istiyoruz. Öncelikle;

$$|HAVA_{güneşli}|=5$$

$$|HAVA_{yağmurlu}|=5$$

$$|HAVA_{bulutlu}|=4$$

HAVA niteliği için entropi değeri;

$$H(HAVA, OYUN) = \frac{5}{14} H(HAVA_{güneşli}) + \frac{4}{14} H(HAVA_{bulutlu}) + \frac{5}{14} H(HAVA_{yağmurlu})$$

$$H\left(HAVA_{g\ddot{u}ne\c{s}li}\right) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971$$

$$H\left(HAVA_{yağmurlu}\right) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971$$

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

Tablo 3.4. HAVA ve OYUN nitelik değerleri

HAVA	OYUN
Güneşli	Hayır
Güneşli	Hayır
Güneşli	Hayır
Güneşli	Evet
Güneşli	Evet
Yağmurlu	Evet
Yağmurlu	Evet
Yağmurlu	Hayır
Yağmurlu	Evet
Yağmurlu	Hayır
Bulutlu	Evet
Bulutlu	Evet
Bulutlu	Evet
Bulutlu	Evet

Bu durumda $H(HAVA, OYUN)$ entropisi şu şekilde elde edilir;

$$H(HAVA, OYUN) = \frac{5}{14} H(HAVA_{güneşli}) + \frac{4}{14} H(HAVA_{bulutlu}) + \frac{5}{14} H(HAVA_{yağmurlu})$$

$$= \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971)$$

$$= 0.694$$

Kazanç ölçütü aşağıda belirtildiği biçimde hesaplanır:

$$Kazanç(HAVA, OYUN) = H(OYUN) - H(HAVA, OYUN)$$

$$= 0.940\text{-}0.693$$

$$= 0.247$$

C) NEM niteliği için kazanç ölçütü:

$$|NEM_{yüksek}|=7$$
$$|NEM_{normal}|=7$$

NEM niteliği için entropi değeri şu şekilde belirlenir;

$$H(NEM, OYUN) = \frac{7}{14} H\left(NEM_{yüksek}\right) + \frac{7}{14} H(NEM_{normal})$$

Bu ifade içinde yer alan $H\left(NEM_{yüksek}\right)$ ve $H(NEM_{normal})$ entropileri şu şekilde hesaplanır;

$$H\left(NEM_{yüksek}\right)=-\left(\frac{4}{7}\log_2\frac{4}{7}+\frac{3}{7}\log_2\frac{3}{7}\right)=0.985$$

$$H(NEM_{normal})= -\left(\frac{1}{7}\log_2\frac{1}{7}+\frac{6}{7}\log_2\frac{6}{7}\right) = 0.592$$

Tablo 3.5. NEM ve OYUN nitelik değerleri

NEM	OYUN
Yüksek	Hayır
Yüksek	Hayır
Yüksek	Evet
Yüksek	Evet
Yüksek	Hayır
Yüksek	Evet
Yüksek	Hayır
Normal	Evet
Normal	Hayır
Normal	Evet
Normal	Evet
Normal	Evet
Normal	Evet
Normal	Evet

Bu durumda $H(NEM, OYUN)$ entropisi şu şekilde elde edilir:

$$H(NEM, OYUN) = \frac{7}{14} H\left(NEM_{yüksek}\right) + \frac{7}{14} H(NEM_{normal})$$

$$= \frac{7}{14}(0.985) + \frac{7}{14}(0.592)$$

$$= 0.789$$

Kazanç ölçütü şu şekilde hesaplanır;

$$Kazanç(NEM, OYUN) = H(OYUN) - H(NEM, OYUN)$$

$$= 0.940 - 0.789$$

$$= 0.151$$

D) RÜZGAR niteliği için kazanç ölçütü:

Son kez RÜZGAR niteliği için kazanç ölçütünü hesaplamak istiyoruz;

$$|R\ddot{U}ZGAR_{hafif}| = 8$$

$$|R\ddot{U}ZGAR_{kuvvetli}| = 6$$

RÜZGAR niteliği için entropi değeri;

$$H(R\ddot{U}ZGAR, OYUN) = \frac{8}{14} H\left(R\ddot{U}ZGAR_{y\ddot{u}ksek}\right) + \frac{6}{14} H(R\ddot{U}ZGAR_{kuvvetli})$$

Tablo 3.6. RÜZGAR ve OYUN nitelik değerleri

RÜZGAR	OYUN
Hafif	Hayır
Hafif	Evet
Hafif	Evet
Hafif	Evet
Hafif	Hayır
Hafif	Evet
Hafif	Evet
Hafif	Evet
Kuvvetli	Hayır
Kuvvetli	Hayır
Kuvvetli	Evet
Kuvvetli	Evet
Kuvvetli	Evet
Kuvvetli	Hayır

$$H\left(\text{RÜZGAR}_{hafif}\right) = -\left(\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right) = 0.811$$

$$H\left(\text{RÜZGAR}_{kuvvetli}\right) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1$$

Bu durumda $H(\text{RÜZGAR}, OYUN)$ $entropisi$;

$$H(\text{RÜZGAR}, OYUN) = \frac{8}{14}H\left(\text{RÜZGAR}_{hafif}\right) + \frac{6}{14}H\left(\text{RÜZGAR}_{kuvvetli}\right)$$

$$= \frac{8}{14}(0.811) + \frac{6}{14}(1)$$

$$= 0.892$$

Kazanç ölçütü;

$$Kazanç(\text{RÜZGAR}, OYUN) = H(OYUN) - H(\text{RÜZGAR}, OYUN)$$

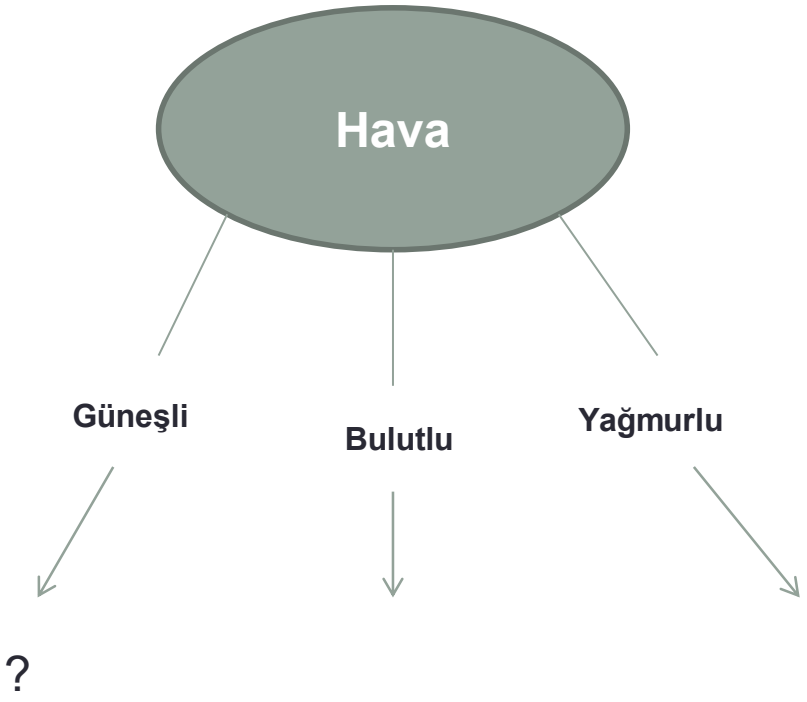
$$= 0.940\text{-}0.892$$

$$= 0.048$$

Tablo 3.7. Elde Edilen Kazanç Ölçütleri

Nitelik	Kazanç
HAVA	0.246
ISI	0.029
NEM	0.151
RÜZGAR	0.048

- Bu değerlere bakarak en büyük kazancı HAVA niteliğini seçerek elde edilebileceğini söyleyebiliriz. Elde edilen sonuç kullanılarak başlangıç karar ağacı şu şekilde çizilebilir;



Adım 2: HAVA niteliğinin ‘güneşli’ değeri için dallanma

- Bu aşamada HAVA niteliğinin ‘güneşli’ değeri için alt karar ağacını düzenleyeceğiz.

Tablo 3.8. HAVA=güneşli için gözlem değerleri

HAVA	ISI	NEM	RÜZGAR	OYUN
Güneşli	Sıcak	Yüksek	Hafif	Hayır
Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
Güneşli	Ilık	Yüksek	Hafif	Hayır
Güneşli	Soğuk	Normal	Hafif	Evet
Güneşli	Ilık	Normal	Kuvvetli	Evet

- Burada önce OYUN için entropiyi hesaplamak gerekiyor;

$$H(OYUN) = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5})$$

$$= 0.970$$

a) ISI niteliği için kazanç ölçütü:

Önce ISI niteliğini ele alalım. Bu niteliğin her bir değeri için aşağıdaki değeri yazabiliriz.

$$\left|ISI_{soğuk}\right| = 1$$

Tablo 3.9. ISI ve OYUN nitelik değerleri

ISI	OYUN
Soğuk	Evet
Sıcak	Hayır
Sıcak	Hayır
Ilık	Hayır
Ilık	Evet

Bu tabloya göre ISI niteliğini çeşitli değerleri için entropi hesaplaması yapılır;

$$H(ISI_{soğuk}) = -(\frac{1}{1}\log_2\frac{1}{1}) = 0$$

$$H(ISI_{sıcak}) = -(\frac{2}{2}\log_2\frac{2}{2}) = 0$$

$$H(ISI_{ılık}) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$$

$H(ISI, OYUN)$ *entropisi şu şekilde elde edilir*;

$$H(ISI, OYUN) = \frac{1}{5}(0) + \frac{2}{5}(0) + \frac{2}{5}(1) = 0.4$$

$H(ISI) = 0.970$ *olduğuna göre, kazanç* ölçütü ;

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN)$$
$$= 0.970\text{-}0.4$$
$$= 0.570$$

b) NEM niteliği için kazanç ölçütü:

NEM niteliği için kazanç ölçütü aşağıda belirtildiği gibi hesaplanır;

$$H\left(NEM_{yüksek}\right) = -\left(\frac{3}{3}\log_2\frac{3}{3}\right) = 0$$

$$H(NEM_{normal}) = -\left(\frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

$$H(NEM, OYUN) = -\frac{3}{5}(0) + \frac{2}{5}(0) = 0$$

$$Kazanç(NEM, OYUN) = H(OYUN) - H(NEM, OYUN)$$

$$=0.970\text{-}0$$

$$=0.970$$

Tablo 3.10. NEM ve OYUN nitelik değerleri

NEM	OYUN
Yüksek	Hayır
Yüksek	Hayır
Yüksek	Hayır
Normal	Evet
Normal	Evet

C) RÜZGAR niteliği için kazanç ölçütü:

RÜZGAR niteliği için kazanç ölçütü aşağıda belirtildiği biçimde hesaplanır;

Tablo 3.11. RÜZGAR ve OYUN nitelik değerleri

RÜZGAR	OYUN
Hafif	Hayır
Hafif	Hayır
Hafif	Evet
Kuvvetli	Hayır
Kuvvetli	Evet

$$H\left(R\ddot{U}ZGAR_{hafif}\right) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0.918$$

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

$$H(NEM, OYUN) = \frac{3}{5}(0.918) + \frac{2}{5}(1) = 0.951$$

$$Kazanç(R\ddot{U}ZGAR, OYUN) = H(OYUN) - H(R\ddot{U}ZGAR, OYUN)$$

$$=0.970-0.951$$

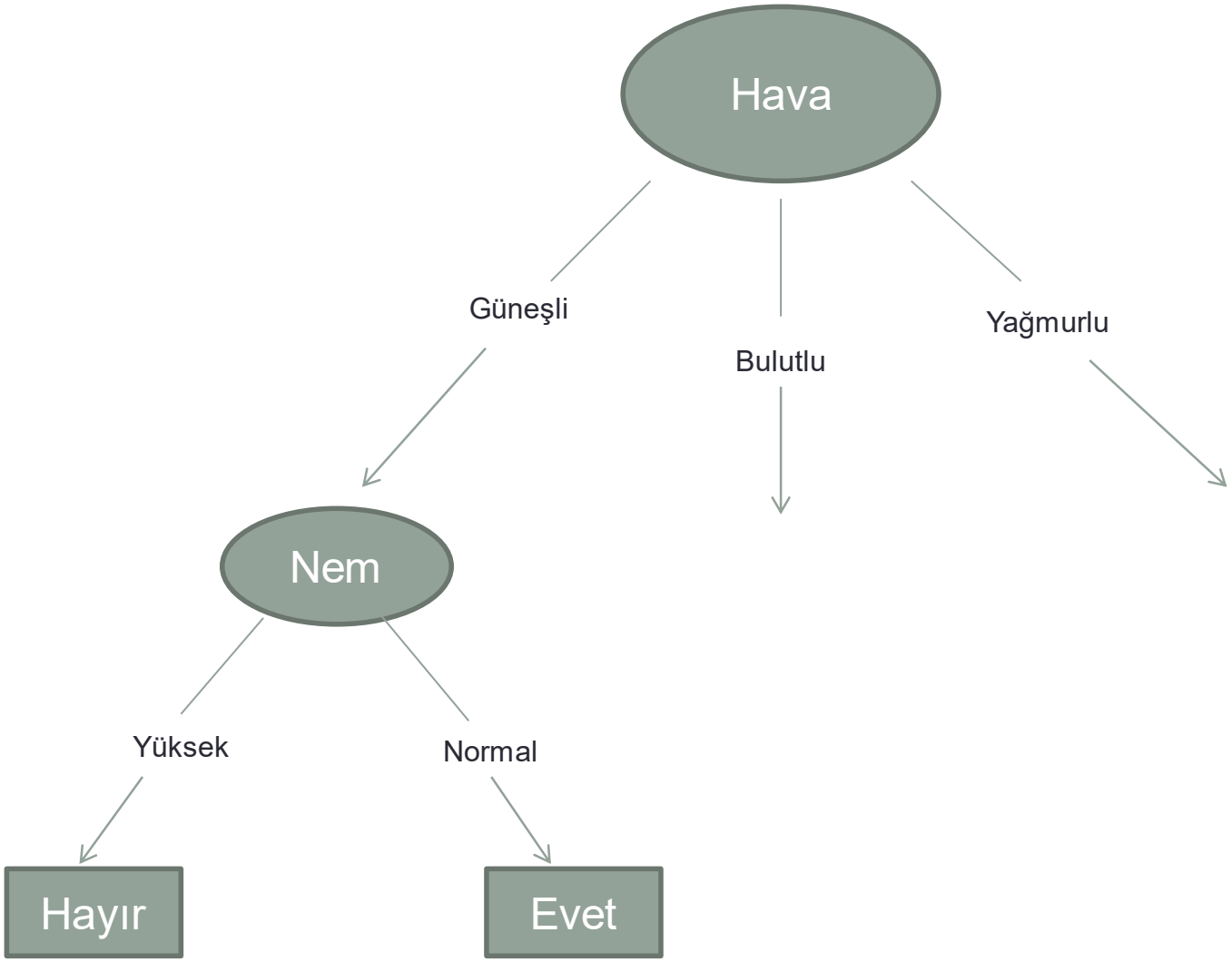
$$=0.019$$

- Elde edilen kazanç ölçütlerini aşağıdaki tabloda topluca veriyoruz:

Tablo 3.12. Kazanç ölçütleri

Nitelik	Kazanç
ISI	0.570
NEM	0.970
RÜZGAR	0.019

Bu değerlere bakarak en büyük kazancın NEM niteliğini seçerek elde edilebileceğini görüyoruz. Elde edilen sonuçlara bağlı olarak karar ağacını şu şekilde geliştiriyoruz;


Hava
Güneşli
Bulutlu
Yağmurlu
Nem
Yüksek
Normal
Hayır
Evet

- NEM ile ilgili ‘yüksek’ değerine sadece ‘hayır’ değeri karşılık geldiğinden, bu noktadan itibaren aşağıya doğru dalın ilerlemesi son bulur. ‘Hayır’ değeri artık bir yaprak değeridir. Benzer biçimde ‘normal’ değeri içinde ‘evet’ değeri yaprak değeridir.

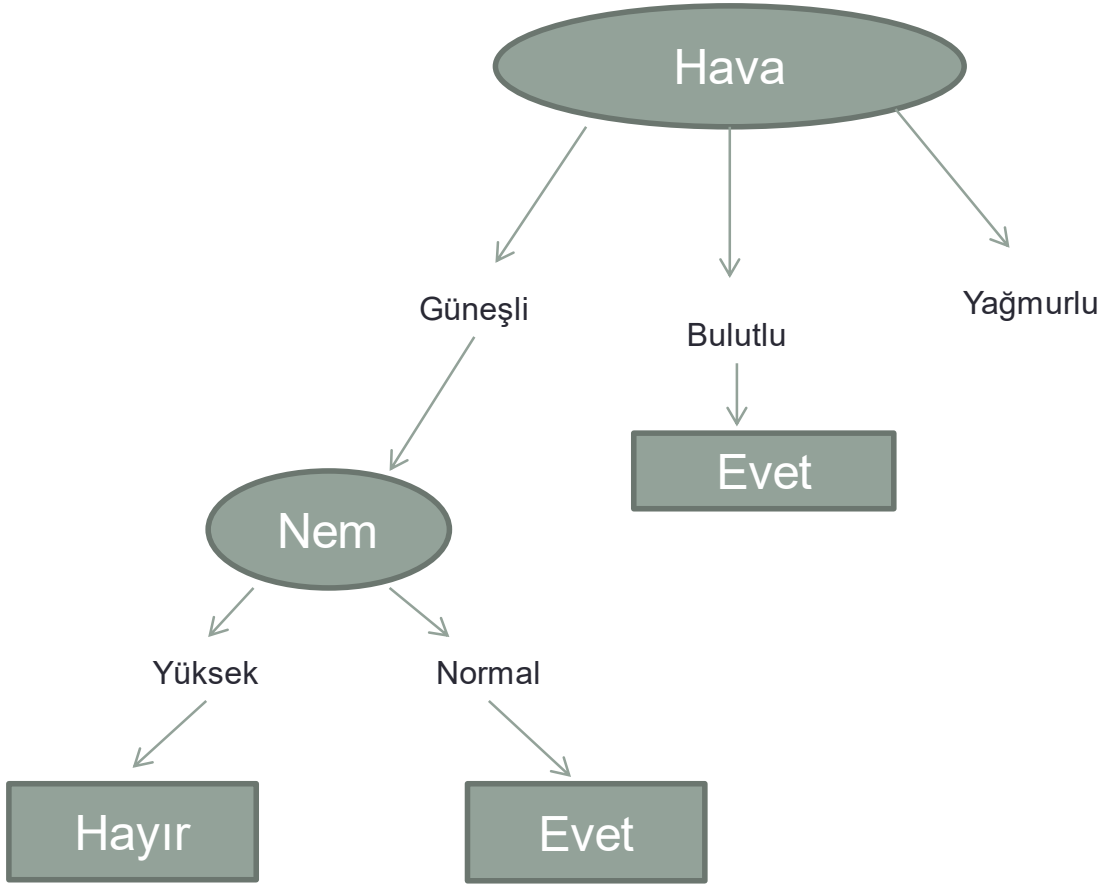
d) HAVA niteliğinin bulutlu değeri için dallanma:

Veri kümesini HAVA niteliğinin ‘bulutlu’ değeri için yeniden düzenleyelim.

Tablo 3.13. Niteliklerin değerleri

HAVA	ISI	NEM	RÜZGAR	OYUN
Bulutlu	Sıcak	Yüksek	Hafif	Evet
Bulutlu	Soğuk	Normal	Kuvvetli	Evet
Bulutlu	Ilık	Yüksek	Kuvvetli	Evet
Bulutlu	Sıcak	Normal	Hafif	Evet

- Görüldüğü gibi tüm karar değerleri ‘evet’ olduğu için herhangi bir analize gerek yoktur. Bu noktadan itibaren bir dallanma olmaz ve bu değer bir yaprağı belirlemiş olur.



e) HAVA niteliğinin yağmurlu değeri için dallanma:

Bu değerle ilgili olarak aşağıdaki tablo düzenlenebilir.

Tablo 3.14. HAVA niteliğinin 'yağmurlu' değeri göz önüne alınıyor

HAVA	ISI	NEM	RÜZGAR	OYUN
Yağmurlu	Ilık	Yüksek	Hafif	Evet
Yağmurlu	Soğuk	Normal	Hafif	Evet
Yağmurlu	Soğuk	Normal	Kuvvetli	Hayır
Yağmurlu	Ilık	Normal	Hafif	Evet
Yağmurlu	Ilık	Yüksek	Kuvvetli	Hayır

Burada önce OYUN için entropiyi hesaplamak gerekiyor.

$$H(OYUN) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)=0.970$$

g) ISI niteliği için kazanç ölçütü:

ISI niteliğini ele alalım. Bu niteliğin her bir değeri için aşağıdaki değeri yazabiliriz.

$$|ISI_{soğuk}|=2$$

$$|ISI_{ılık}|=3$$

ISI niteliğine göre bir ayırma gerçekleştirildiğinde kazancın ne olacağını hesaplayalım.

Tablo 3.15. ISI ve OYUN nitelik değerleri

ISI	OYUN
Soğuk	Evet
Soğuk	Hayır
Ilık	Evet
Ilık	Evet
Ilık	Hayır

Tablo 3.15.'e göre 'soğuk' değeri için aşağıdaki entropi hesaplaması yapılır:

$$H(ISI_{soğuk}) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$$

$$H(ISI_{ılık}) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 0.918$$

$$H(ISI, OYUN) = \frac{2}{5}(1) + \frac{3}{5}(0.918) = 0.951$$

$$Kazanç(ISI, OYUN) = H(OYUN) - H(ISI, OYUN)$$

$$=0.970-0.951$$

$$=0.019$$

f) RÜZGAR niteliği için kazanç ölçütü:

Bu kez RÜZGAR niteliğini ele alalım. Bu niteliğin her bir değeri için aşağıdaki değeri yazabilirsiniz.

$$|R\ddot{U}ZGAR_{hafif}|=3$$

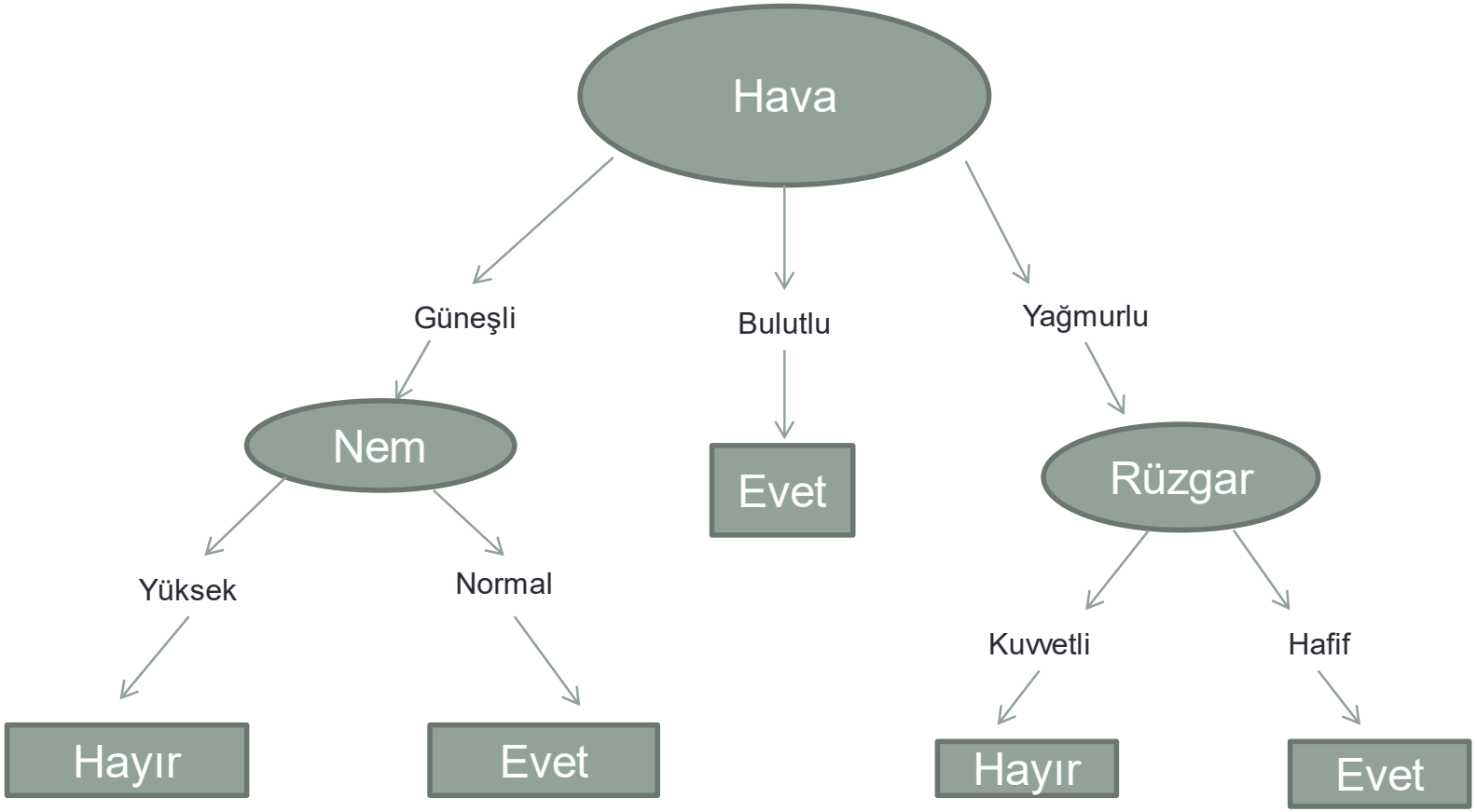
$$|R\ddot{U}ZGAR_{g\ddot{u}\c{c}l\ddot{u}}|= 2$$

RÜZGAR niteliğine göre bir ayırma gerçekleştirildiğinde kazancın ne olacağını bulmak istiyoruz.

Tablo 3.16. RÜZGAR ve OYUN nitelik değerleri

RÜZGAR	OYUN
Hafif	Evet
Hafif	Evet
Hafif	Evet
Kuvvetli	Hayır
Kuvvetli	Hayır

- Görüldüğü gibi, RÜZGAR niteliğinin ‘hafif’ değerleri için ‘evet’ değeri elde edilmektedir. Benzer biçimde tabloda ‘kuvvetli’ değeri için ‘hayır’ değerini aldığı görülüyor. O halde RÜZGAR düğümünden itibaren yeni bir niteliğe dallanamayız. Yukarıda tabloda yer alan değerlere bağlı olarak yaprak değerleri elde edilir. Böylece karar ağacı oluşturma süreci sona ermiş olur.



Sonuç karar ağacı

Kazanç Oranı

- Karar ağacının oluşturulması esnasında ‘kazanç ölçütü’ adı verilen bir değeri kullandık. Ancak uygulamada bu yöntemden daha iyi sonuçlar veren bir başka yöntem kullanılmaktadır. Bilgi bölünmesi adı verilen bu kavram Quinlan tarafından ortaya atılmıştır. T kümesi için X niteliğinin değerini belirlemek için gereken bilgi miktarının ortaya koymak için bu yol bulunmuştur. Bilgi miktarı $H(P_{X,T})$ biçiminde ifade edilebilir. Burada $P_{X,T}$ ifadesi X değerlerinin olasılık dağılımıdır ve şu şekilde hesaplanır;

$$P_{X,T}=\left(\frac{|T_1|}{|T|},\frac{|T_2|}{|T|},\dots,\frac{|T_k|}{|T|}\right)$$

Burada $H(P_{X,T})$ miktarı T kümesindeki X niteliği için bilgi bölünmesi olarak bilinmektedir. Bu değer şu şekilde hesaplanır;

$$H(P_{X,T}) = H\left(\frac{|T_1|}{|T|}, \frac{|T_2|}{|T|}, \dots, \frac{|T_k|}{|T|}\right)$$

Bunun yerine aşağıdaki ifade kullanılabilir;

$$H(P_{X,T}) = -\sum_{i=1}^{k} \frac{T_i}{T} \log_2 \frac{T_i}{T}$$

Yukarıda elde edilen $H(P_{X,T})$ değeri ve kazanç ölçütü yardımıyla kazanç oranı hesaplanır:

$$Kazanç\ oranı(X,T) = \frac{Kazanç(X,T)}{H(P_{X,T})}$$

Örnek:

- Önceki örnekte yer alan verileri yeniden göz önüne alalım. ISI niteliği ile ilgili olarak bilgi bölümlemesini ve kazanç oranını hesaplamak istiyoruz. Burada $\{T_1, T_2, \dots, T_k\}$ değerlerinin ISI niteliğinin her bir değerine karşılık gelen hedef nitelik değerleridir. Örneğin, ISI niteliği ile ilgili olarak Tablo 3.3. üzerinde görüldüğü gibi, $T_1 = \{evet, hayır, evet, evet\}$ kümesi ele alınır. O halde $H(P_{ISI,OYUN})$ bağıntısı şu şekilde ifade edilebilir:

$$H(P_{X,T}) = -\sum_{i=1}^{3} \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|}$$

Olduğuna göre;

$$H\left(P_{ISI,OYUN}\right) = -\left[\frac{\left|ISI_{soğuk}\right|}{|ISI|}\log_2\frac{\left|ISI_{soğuk}\right|}{|ISI|} + \frac{|ISI_{ılık}|}{|ISI|}\log_2\frac{|ISI_{ılık}|}{|ISI|} + \frac{|ISI_{sıcak}|}{|ISI|}\log_2\frac{|ISI_{sıcak}|}{|ISI|}\right]$$

Hesaplamalar yapılırsa;

$$H\left(P_{ISI,OYUN}\right) = -\left(\frac{4}{14}\log_2\frac{4}{14} + \frac{6}{14}\log_2\frac{6}{14} + \frac{4}{14}\log_2\frac{4}{14}\right)$$

$$= 1.577$$

Kazanç(ISI,OYUN)=0.029 hesaplanmıştı.

Kazanç oranı(ISI,OYUN)=$\frac{0.029}{1.577}$=0.018

- Benzer biçimde RÜZGAR niteliği için şu şekilde hesaplamalar yapılır:

$$H\left(P_{RÜZGAR,OYUN}\right) = -\sum_{i=1}^{2} \frac{|T_i|}{|T|} \log_2 \frac{|T_i|}{|T|}$$

$$= -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = 0.985$$

Kazanç(RÜZGAR,OYUN)=0.048 hesaplanmıştı,

Kazanç oranı(RÜZGAR,OYUN)=$\frac{0.048}{0.985}$=0.049

C4.5 Algoritması

- Bu algoritma, sayısal değerlere sahip niteliklerin de karar ağaçlarını oluşturma olanağı sağlamıştır. Ayrıca bilinmeyen nitelik değerlerine sahip örnek kümeleri için karar ağacının nasıl oluşturulabileceği konusunda bir yol sunmaktadır.

Sayısal Değerlere Sahip nitelikler:

- Sayısal niteliklere ilişkin testlerin formüle edilmesinde bazı zorluklar görünebilir. Değerleri iki aralığa bölmek için gelişi güzel eşikler bulunmaktadır. Ancak en uygun t eşik değerini hesaplamak için çeşitli yollar bulunmaktadır. Eşik değerinin belirlenmesi amacıyla, en büyük bilgi kazancını sağlayacak biçimde bir eşik değer belirlenir. Bunun için nitelik değerleri sıralanır ve $\{v_1, v_2, \dots, v_3\}$ biçimini alır.

Bu eşik değeri kullanılarak nitelik değeri iki parçaya ayrılır. Eşik değeri olarak $[v_i, v_{i+1}]$ aralığının orta noktası alınabilir.

$$t_i = \frac{v_i + v_{i+1}}{2}$$

- C4.5 'de eşik olarak $[v_i, v_{i+1}]$ aralığının en küçük değeri eşik olarak alınır.
- Bir veri tabanında herhangi bir niteliği sürekli, yani sayısal değerlere sahip ise genellikle ikili test uygulanır. Bu testte niteliğin bir t eşik değeri ile karşılaştırılır.

Uygulama: Tablo 3.17 Eğitim kümesinin görünümü

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	Doğru	Sınıf1
a	90	Doğru	Sınıf2
a	85	Yanlış	Sınıf2
a	95	Yanlış	Sınıf2
a	70	Yanlış	Sınıf1
b	90	Doğru	Sınıf1
b	78	Yanlış	Sınıf1
b	65	Doğru	Sınıf1
b	75	Yanlış	Sınıf1
c	80	Doğru	Sınıf2
c	70	Doğru	Sınıf2
c	80	Yanlış	Sınıf1
c	70	Yanlış	Sınıf1
c	96	Yanlış	Sınıf1

Eşik değerin belirlenmesi:

NİTELİK2'nin içerdiği değerler göz önüne alındığında en fazla bilgi kazancını sağlayacak değerin 80 olduğu anlaşılır. NİTELİK2 niteliği $\{65,70,75,80,85,90,95,96\}$ değerlerine sahiptir. Kümenin orta noktaları olan (80,85) aralığının orta noktası olan $t_i=\frac{v_i+v_{i+1}}{2}=\frac{80+85}{2}\sim 83$ noktası eşik değer olarak alınabilir. O halde NİTELİK2'nin içerdiği değerlere (NİTELİK≤ 83 veya NİTELİK2>83) testi uygulanarak ona göre düzenleme yapılır. Bu teste göre eğitim kümesinin görünümü Tablo 3.18'de görüldüğü biçimi alır.

Tablo 3.18

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	Eşit veya küçük	Doğru	Sınıf1
a	Büyük	Doğru	Sınıf2
a	Büyük	Yanlış	Sınıf2
a	Büyük	Yanlış	Sınıf2
a	Eşit veya küçük	Yanlış	Sınıf1
b	Büyük	Doğru	Sınıf1
b	Eşit veya küçük	Yanlış	Sınıf1
b	Eşit veya küçük	Doğru	Sınıf1
b	Eşit veya küçük	Yanlış	Sınıf1
c	Eşit veya küçük	Doğru	Sınıf2
c	Eşit veya küçük	Doğru	Sınıf2
c	Eşit veya küçük	Yanlış	Sınıf1
c	Eşit veya küçük	Yanlış	Sınıf1
c	Büyük	Yanlış	Sınıf1

Burada SINIF niteliği için,

$$P_{SINIF} = \left(\frac{5}{14}, \frac{9}{14}\right)$$

olduğuna göre ,SINIF kümesi için entropi şu şekilde hesaplanabilir.

$$H(SINIF) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right) = 0.940$$

Her değerin nitelik içinde tekrarlanma sayıları hesaplanmıştır.

$$|NİTELİK2_{\leq 83}| = 9$$

$$|NİTELİK2_{>83}| = 5$$

Bu durumda NİTELİK2 niteliğine göre bir ayırma gerçekleştirildiğinde kazancın ne olacağını hesaplamak için entropiyi bulmak gerekiyor. NİTELİK2 niteliğinin '≤ 83' değeri için aşağıdaki tabloyu göz önüne alalım.

Tablo 3.19. NİTELİK2 ve SINIF nitelikleri

NİTELİK2	SINIF
Eşit veya küçük	Sınıf1
Eşit veya küçük	Sınıf1
Eşit veya küçük	Sınıf1
Eşit veya küçük	Sınıf1
Eşit veya küçük	Sınıf1
Eşit veya küçük	Sınıf2
Eşit veya küçük	Sınıf2
Eşit veya küçük	Sınıf1
Eşit veya küçük	Sınıf1

Bu tabloya göre aşağıdaki entropi hesaplaması yapılır:

$$H(\text{NİTELİK2}_{\leq 83}) = -\left(\frac{7}{9}\log_2\frac{7}{9} + \frac{2}{9}\log_2\frac{2}{9}\right)=0.764$$

NİTELİK2 niteliğinin ‘>83’ değeri için aşağıdaki tabloyu göz önüne alalım.

Tablo 3.20. NİTELİK2 ve SINIF nitelikleri

NİTELİK2	SINIF
Büyük	Sınıf2
Büyük	Sınıf2
Büyük	Sınıf2
Büyük	Sınıf1
Büyük	Sınıf1

Bu tabloya göre aşağıdaki entropi hesaplaması yapılır:

$$H(\text{NİTELİK}_{>83}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right)=0.970$$

Sonuç olarak NİTELİK2 ile ilgili aşağıdaki toplam entropi elde edilir.

$$H(\text{NİTELİK2}, SINIF) = \frac{9}{14}(0.764) + \frac{5}{14}(0.970)=0.837$$

Kazanç(NİTELİK2,SINIF)=H(SINIF)-H(NİTELİK2,SINIF)

=0.940-0.837

=0.103

Bilinmeyen Nitelik Değerleri

- Veri tabanındaki bilgilerde bazı kayıplar olabilir. Veri toplanırken bazı değerler kaydedilmemiş olabilir veya veri tabanına kaydedilirken bir sorun ortaya çıkmış olabilir. Kayıp değerlerin yaratacağı sorunları önlemek için aşağıda belirtilen yollardan biri izlenebilir:

a) Kayıp veriye ilişkin tüm değerler veri tabanından çıkarılır

b) Kayıp verilerle de çalışabilecek yeni bir algoritma kullanılır.

- C4.5'de kayıp verilere sahip örneklerde kazanç ölçütünü hesaplamak için bir düzeltme faktöründen yararlanılır.

$$F = \frac{Veri\ tabanında\ bilinen\ niteliğe\ sahip\ örneklerin\ sayısı}{Veri\ tabanındaki\ tüm\ örneklerin\ sayısı}$$

Yeni kazanç ölçütü şu şekilde olacaktır;

$$Kazanç(X)=F(H(T)-H(X,T))$$

ÖRNEK:

Aşağıdaki örnek üzerinde C4.5 algoritmasının bilinmeyen nitelik değerlerine nasıl yaklaştığını inceleyeceğiz. Burada eğitim kümesinin NİTELİK1 niteliği üzerinde bir kayıp değeri vardır ve ? İşareti ile gösterilmektedir.

Tablo 3.21. Kayıp değerlere sahip veri tabanı

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	Doğru	Sınıf1
a	90	Doğru	Sınıf2
a	85	Yanlış	Sınıf2
a	95	Yanlış	Sınıf2
a	70	Yanlış	Sınıf1
?	90	Doğru	Sınıf1
b	78	Yanlış	Sınıf1
b	65	Doğru	Sınıf1
b	75	Yanlış	Sınıf1
c	80	Doğru	Sınıf2
c	70	Doğru	Sınıf2
c	80	Yanlış	Sınıf1
c	70	Yanlış	Sınıf1
c	96	Yanlış	Sınıf1

NİTELİK1'in kazanç parametresi, önceki örnekte yaptığımız biçimde hesaplanır. Kayıp değerin yer aldığı satır çıkarılarak SINIF hedef niteliği için entropi hesaplanır:

$$H(P_{SINIF}) = -\frac{8}{13}\log_2\frac{8}{13} - \frac{5}{13}\log_2\frac{5}{13} = 0.961$$

Söz konusu kayıp değere sahip olan satır çıkarıldıktan sonra geri kalan NİTELİK1 değerleri göz önüne alınır ve NİTELİK1 değeri için şu şekilde bir hesaplama yapılır:

$$H(\text{NİTELİK1}, SINIF) = \frac{5}{13}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right] + \frac{3}{13}\left[-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}\right] + \frac{5}{14}\left[-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right] = 0.747$$

$$Kazanç(\text{NİTELİK1}, SINIF) = \frac{13}{14}(0.961 - 0.747)$$

$$= 0.199$$

Karar Ağaçlarının Budanması

- Karar ağaçları çoğu kez karmaşık bir görünüme sahip olabilir. Bir karar ağacında, bir alt ağacı atarak yerine bir yaprak yerleştirmek söz konusu olabilir. Bu şekilde yapılan işleme 'karar ağacının budanması' adı verilmektedir.
- Alt ağacın yerine yaprak yerleştirmekle, algoritma 'öngörülü hata oranını' azaltmayı ve sınıflandırma modelinin kalitesini arttırmayı amaçlar.
- Öngörülü hata oranını belirlemek için şu yol izlenir;

1) İlave test örneklerinden oluşan yeni bir küme kullanmak
2) Bu teknik önceden var olan örnekleri eşit boydaki bloklara böler ve her bir blok için bu bloğu oluşturan tüm örneklerden ağaç oluşturulur.
3) Ardından bu ağaç verilmiş örnekler bloğu ile test edilir.

Özyinelemeli bölme yönteminde düzenlemeler yapmak için iki yol vardır:

a) Bazı durumlarda örnekler kümesini daha fazla bölmemek kararı alınır. Bölme işlemine son verme, yani durdurma ölçütü olarak $\times^2$ gibi istatistiksel testlere dayanır. Bölünme öncesinde ve sonrasında kayda değer bir fark yoksa , o zaman söz konusu düğüm bir yaprak olarak gösterilir. Bu tekniğe 'ön budama' adı verilir.

b) Seçilen bir 'doğruluk ölçütü' kullanılarak bazı ağaçlar budanabilir. Bu yöntemde budama işlemi ağaç oluşturulduktan sonra yapılır. C4.5 sınıflandırma yöntemi bu tekniği kullanır.

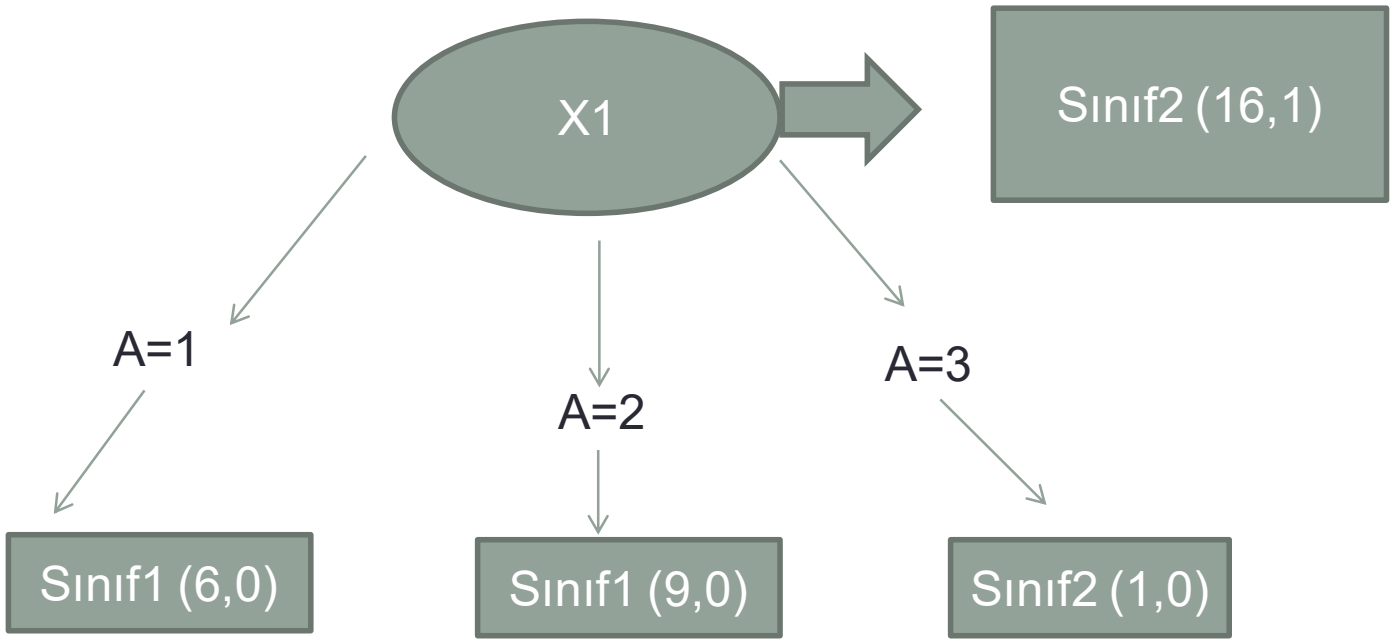
C4.5'de Budama

- C4.5'de beklenen hata oranını tahmin etmek için belirli bir yöntem kullanır. Ağaçtaki her bir düğüm için U_{cf} üst güven sınırı iki terimli dağılımların istatistiksel tablolarını kullanarak elde edilebilir.

- Verilen düğümde U_{cf} parametresi T_i ve *E*'nin bir fonksiyonudur. C4.5 algoritması %25 güven sınırını kullanır ve verilen her bir düğümdeki T_i , $U_{\%25}\left(\frac{|T_i|}{E}\right)$ düğüm yapraklarının güven aralığı ile karşılaştırılır.

Örnek:

Basit ve küçük bir örnekle budama işlemini açıklamak istiyoruz. Aşağıdaki şekilde bir karar ağacının alt ağacı verilmiştir. Burada X1 kök düğümüdür ve bundan A niteliğine ait $\{1,2,3\}$ gibi üç değer çıkmaktadır.



Kök düğümün alt düğümleri, kendilerine uyan sınıflardan ve $\left(\frac{T_i}{E}\right)$ parametreleriyle belirlenmiş birer yapraktır. Bütün düğümlerin üst güven sınırları %25 güven değeri kullanılarak istatistik tablolardan elde edilir:

$$U_{\%25}(6,0)=0.206$$
$$U_{\%25}(9,0) = 0.143$$
$$U_{\%25}(1,0)=0.750$$
$$U_{\%25}(16,1) = 0.157$$

Bu değerleri kullanarak başlangıç ağacının yanı sıra bu ağacın yerine konulan düğümün beklenen hataları hesaplanır;

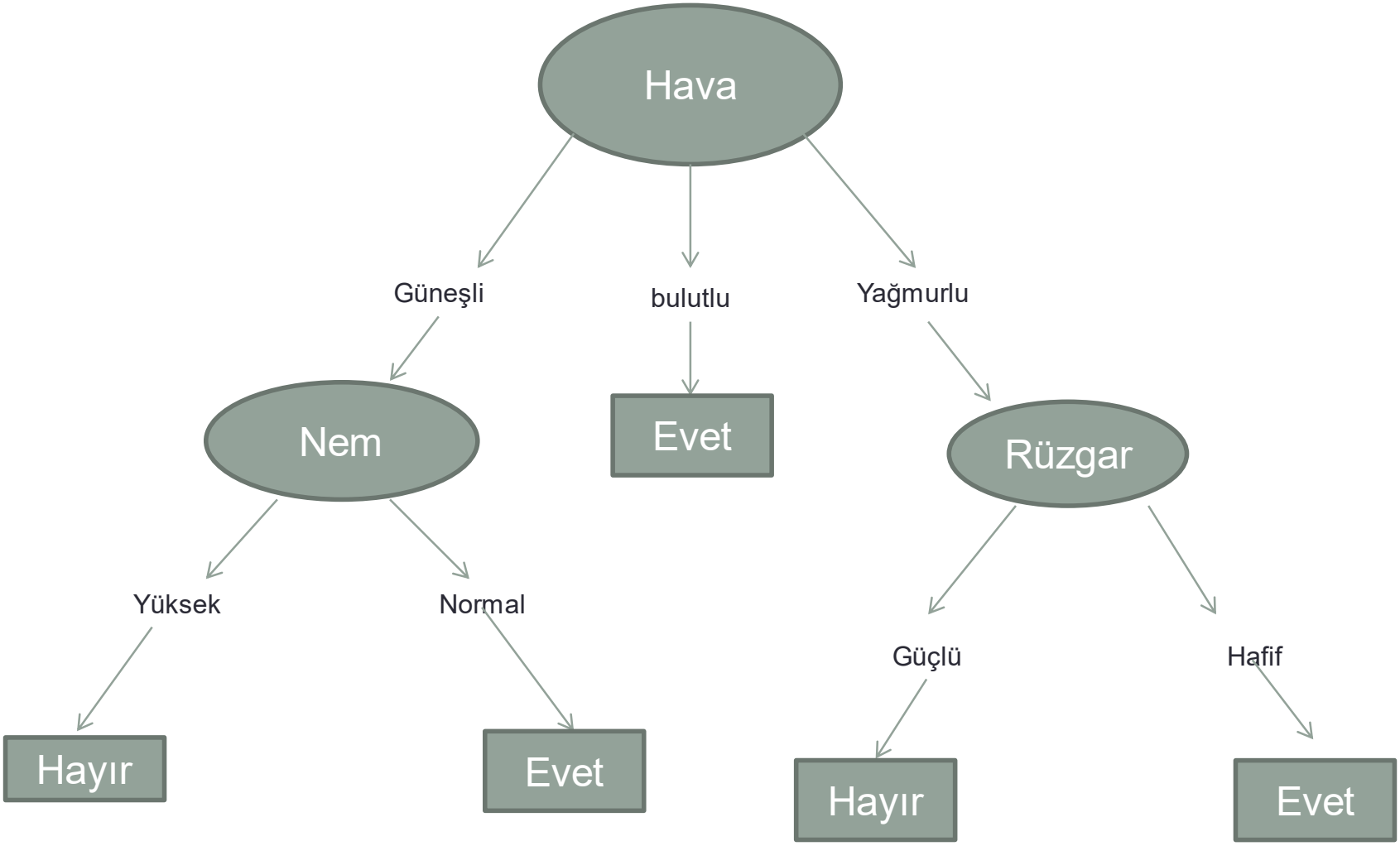
$$PE_{ağaç}=6(0.206)+9(0.143)+1(0.750)=3.257$$

$$PE_{ağaç}=16(0.157)= 2.517$$

Karar Kuralları Oluşturmak

- Eğitim kümesine bağlı olarak elde edilen karar ağacından yararlanarak karar kuralları oluşturulabilir. Karar kuralları aynen programlama dillerindeki IF..THEN…ELSE yapılarına benzer.

Örnek: Karar Ağacı



Bu karar ağacına dayanarak aşağıdaki kural tablosunu çıkarabiliriz:

KURAL1:
Eğer HAVA=güneşli **ise ve**
    Eğer NEM=yüksek ise OYUN=hayır;
KURAL2:
Eğer HAVA=güneşli **ise ve**
    Eğer NEM=normal **ise** OYUN=evet;
KURAL3:
Eğer HAVA=bulutlu **ise** OYUN=evet;
KURAL4:
Eğer HAVA=yağmurlu **ise ve**
    Eğer RÜZGAR=güçlü ise OYUN=hayır;
KURAL5:
Eğer HAVA=yağmurlu **ise ve**
    Eğer RÜZGAR=hafif **ise** OYUN=evet;

BELLEK TABANLI SINIFLANDIRMA

En Yakın k-komşu Algoritması

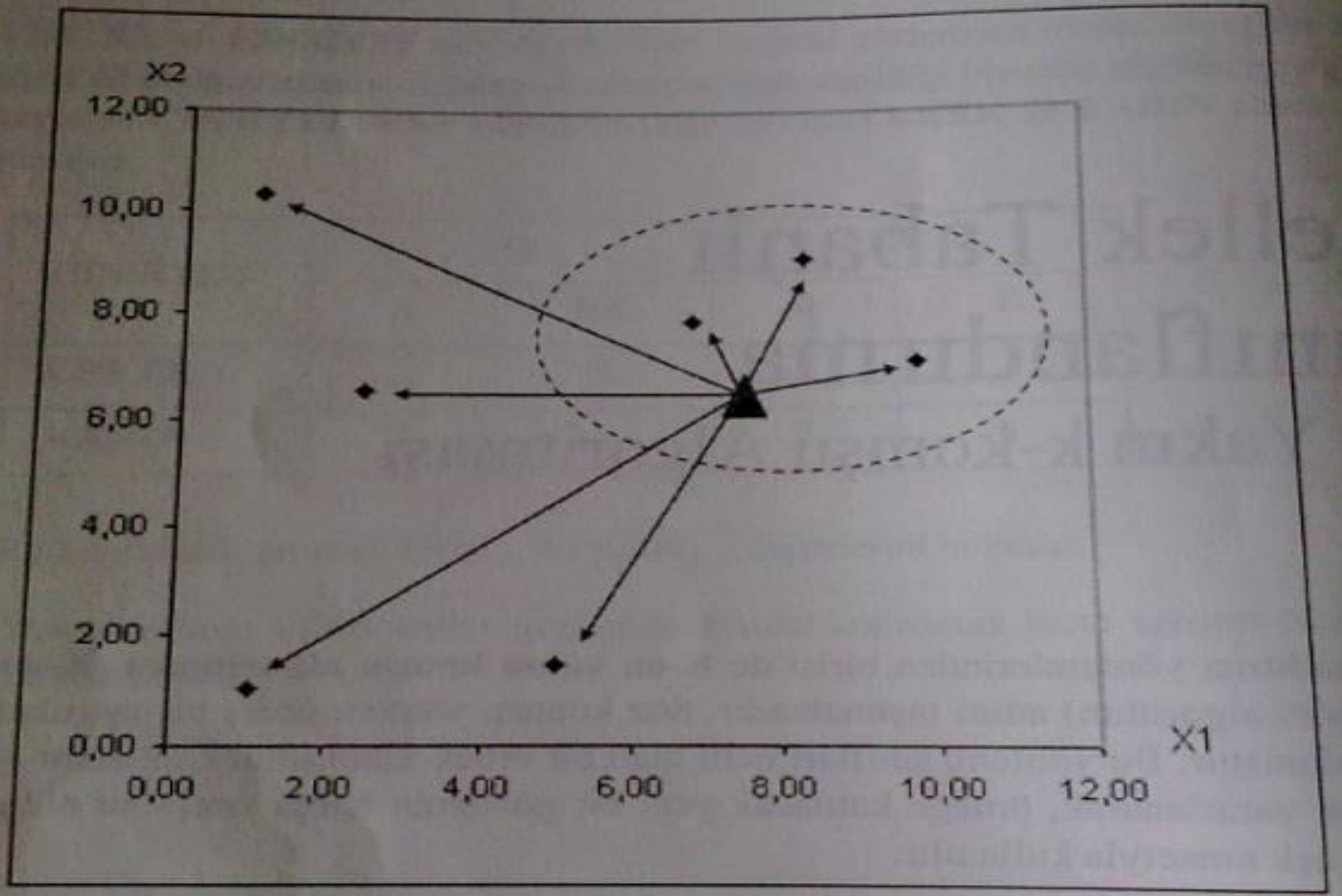
Bu yöntem, sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak, örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacıyla kullanılır.

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.

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

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,

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

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.

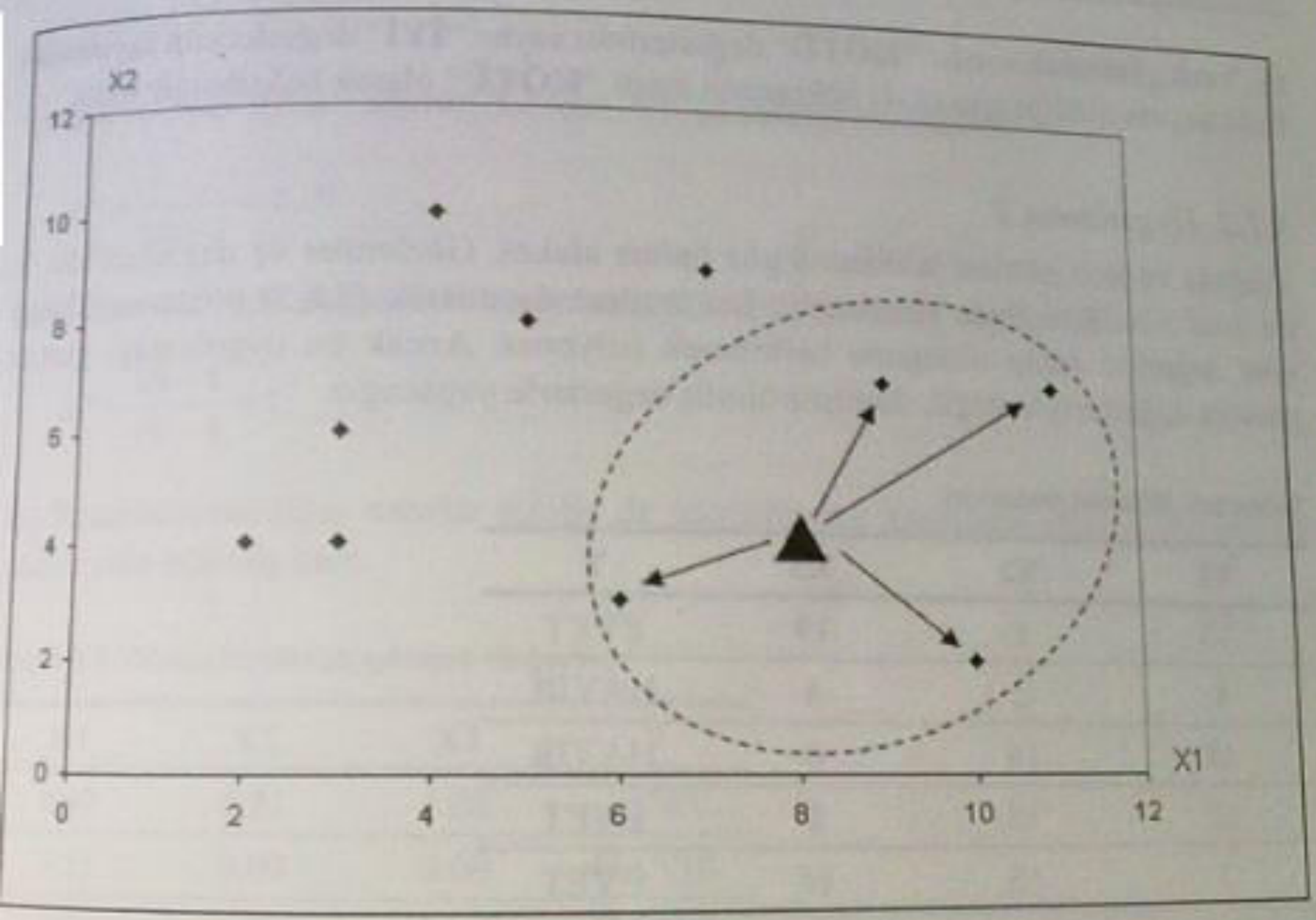
X1	X2	Uzaklık
2	4	6.00
3	6	5.39
3	4	5.00
4	10	7.21
5	8	5.00
6	3	2.24
7	9	5.10
9	7	3.16
11	7	4.24
10	2	2.83

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



Şekil-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,

$$X^* = \frac{X - Xmin}{X_{max} - Xm_{in}}$$

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:

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

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

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.

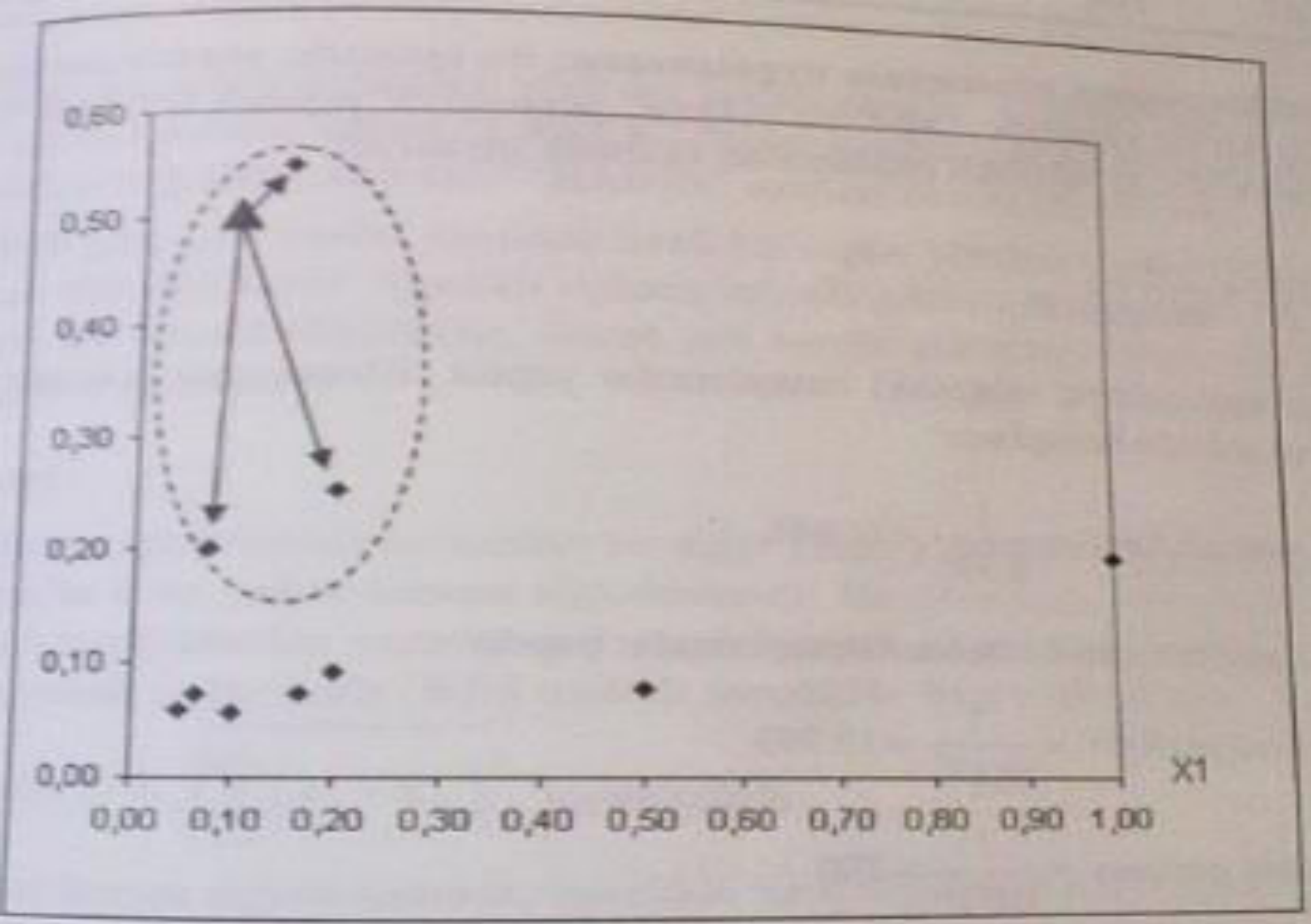
X1	X2	Uzaklık
0.08	0.20	0.30
0.07	0.07	0.43
0.20	0.09	0.42
1.00	0.20	0.95
0.05	0.06	0.44
0.20	0.25	0.27
0.17	0.07	0.43
0.15	0.55	0.07
0.50	0.08	0.58
0.10	0.06	0.44

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 değerin 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	Uzaklık	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 kümelerde işlemi yavaşlatır.

SINIFLANDIRMA VE REGRESYON AĞAÇLARI (CART)

- Twoning Algoritması
  - Uygulama
- Gini Algoritması
  - Uygulama

Sınıflandırma ve Regresyon Ağaçları (Classification And Regression Trees- CART)

- Verimadenciliğinin sınıflandırma ile ilgili konuları arasında yer alır.
- Yöntem 1984'te *Breiman* tarafından ortaya atılmıştır.
- Her bir karar düğümünden itibaren ağacın iki dala ayrılması ilkesine dayanır. Yani bu tür karar ağaçlarında ikili dallanmalar söz konusudur. Bir düğümde seçme işlemi yapıldığında, düğümlerden sadece iki dal ayrılabilir.
- CART algoritmasında, bir düğümde belirli bir kriter uygulanarak bölünme işlemi gerçekleştirilir. Bunun için önce tüm niteliklerin var olduğu değerler göz önüne alınır ve tüm eşleşmelerden sonra iki bölünme elde edilir

Twoing algoritması

Adımlar;

Adım **1**:

a) 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 iki ayrı dal bulunur. Bu bölümlenen kümeler t_{Sol} ve $t_{Sağ}$ biçimindedir.

b) Aday bölünmelerin her biri için P_{Sol} ve $P(j|_{Sol})$ olasılıkları hesaplanır. Söz konusu olasılıklar aşağıda verilmektedir. Burada $P(j|t_{Sol})$ ifadesi bir j sınıf değerinin sol taraftaki bölünmede olma olasılığını verir. Söz konusu j değerleri sınıf değerlerinin yer aldığı nitelik olarak göz önüne alınır.

$$P_{Sol} = \frac{t_{sol}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$P(j|t_{sol}) = \frac{t_{sol}\text{'daki kayıtların j sayısı}}{t_{sol}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}$$

c) Aday bölünmelerin her biri için $P_{Sağ}$ ve $P(j|t_{sağ})$ olasılıkları hesaplanır. Buruda $P(j|t_{sağ})$ ifadesi her bir j sınıf değerinin sağ taraftaki bölünme olma olasılığını verir.

$$P_{sağ}= \frac{t_{sağ}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$P(j|t_{sağ})= \frac{t_{sağ}\text{'daki kayıtların j sayısı}}{t_{sağ}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}$$

d) $\Phi(s|t)$,t düğümündeki s aday bölünmelerinin uygunluk ölçüsü olsun. Söz konusu uygunluk ölçüsü şu şekilde hesaplanır:

$$\Phi(s|t)=2P_{sol}P_{sağ}\sum_{j=1}^{n}|P(j|t_{sol})-P(j|t_{sağ})|$$

e) Φ(s|t) değerleri hesaplandıktan sonra içlerinden *en büyük olanı* seçilir. Bu değerin ilgili olduğu aday bölünme satırı bize dallanmanın yapılacağı satırı bildirecektir.

f) Dallanma bu şekilde yapıldıktan sonra bu adıma ilişkin olarak karar ağacı çizilir

Adım2:

Algoritmanın birinci adımına dönülerek ağacın alt kümesine aynı işlemler uygulanır.

Örnek:

Müşterilerini cinsiyetlerine göre ve alışveriş miktarlarına göre sınıflandırmak isteyen bir mağazada müşteri beş müşteri için tabloda 1'de verilen sonuçların elde edildiğini varsayalım. Bu tabloda sınıflandırmanın yapıldığı SINIF niteliği hedef nitelik olarak kabul edilir.

Bu verileri kullanarak aday bölünmeleri elde edelim.

CİNSİYET=ERKEK değeri sol tarafta, CİNSİYET=KADIN sağ tarafta olacak biçimde bölündüğünü varsayalım. Bu durumda CİNSİYET=ERKEK değeri için P_{Sol} ve $P(A|\ t_{Sol})$ olasılığını hesaplayalım.

Müşteri	CİNSİYET	SATIŞ	SINIF
1	ERKEK	DÜŞÜK	A
2	KADIN	YÜKSEK	A
3	ERKEK	YÜKSEK	B
4	ERKEK	NORMAL	A
5	KADIN	DÜŞÜK	B

Tablo 1

Çözüm:

a) Twoing algoritmasını kullanarak bu verileri sınıflandırmak için öncelikle aday bölünmelerin ortaya konulması gerekmektedir. Bu bölünme ikili olacaktır; yani sol ve sağda olmak üzere iki ayrı küme elde edilecektir. Söz konusu bölünme, CİNSİYET niteliğinin ve ALIŞVERİŞ niteliğinin her bir değeri için, nitelik değerleri iki parçaya ayrılacak biçimde yapılır. Örneğin, CİNSİYET niteliği CİNSİYET= ERKEK ve CİNSİYET=KADIN biçiminde iki parçaya ayrılabilir. Bunlardan birincisi soldaki kısımda, diğeri ise sağdaki kısımda yer alacaktır. Benzer biçimde SATIŞ nitelik değerleri bölünebilir. Ancak bu nitelik üç farklı değere sahip olduğu için SATIŞ=DÜŞÜK ve SATIŞ $\in$ {NORMAL,YÜKSEK} biçiminde bir bölünme söz konusu olur. Sonuç olarak sol ve sağ aday bölünmeler Tablo 2 deki gibi elde edilir.

ADAY BÖLÜNME	T_{Sol}	$T_{Sağ}$
1	CİNSİYET=ERKEK	CİNSİYET=KAD1N
2	CİNSİYET=KADIN	CİNSİYET=ERKEK
3	SATIŞ=DÜŞÜK	SATIŞ ϵ {NORMAL, YÜKSEK}
4	SATIŞ=NORMAL	SATIŞ ϵ {DÜŞÜK, YÜKSEK}
5	SATIŞ=YÜKSEK	SATIŞ ϵ (DÜŞÜK, NORMAL}

Tablo 2

b) P_{Sol} ve P(A| t_{Sol}) olasılıkları şu şekilde hesaplanmaktadır.

$$P_{sol} = \frac{t_{sol}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$$

$$P(A|t_{sol}) = \frac{t_{sol}\text{'daki kayıtların A sınıfları sayısı}}{t_{sol}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}$$

Eğitim kümesindeki kayıtların sayısının 5 olduğu anlaşılıyor. Tablo2üzerinde t_{Sol} kümesinin elemanlarını görüyoruz. Hesaplamalarımızı 1. aday bölünme için, yani CİNSİYET=ERKEK satırı için yapacağız. Bu değerin Tablo 1 de CİNSİYET nitelik değerleri arasında kaç tane tekrarlandığını bulalım. Bu değerin 3 müşteriye ait olduğunu görüyoruz. O halde P_{sol} ifadesi;

$P_{sol}= 3/5= 0.6$

biçiminde hesaplanır.

$P(A \setminus t_{Sol})$ ifadesini hesaplayabilmek için önce t_{Sol} bölünmesi içindeki $j=A$ sınıfına ait satırların sayısını belirlemek gerekiyor. Tablo1de CİNSİYET=ERKEK satırlarının kaç tanesinin karşısında A sınıfı vardır? Bu sorunun yanıtının 2 olduğu anlaşılıyor. O halde $P(A \setminus t_{Sol})$ olasılığı şu şekilde hesaplanır:

$$P(A|t_{sol}) = 2/3 = 0.67$$

Örnek2:

Bir eğitim kümesinde hedef nitelik sınıf değerlerinin GEÇERLİ ve GEÇERSİZ olduğunu varsayalım. *Twoing* yöntemini kullanarak sınıflandırma işlemini yapmak amacıyla aşağıdaki değerlerin elde edildiğini varsayalım:

$Psol$	0.40
$P(GEÇERLİ \mid t_{Sol})$	0.50
$P(GEÇERLİ \mid t_{Sağ})$	0.33

Bu değerleri kullanarak Φ(s|t) uygunluk ölçüsünü hesaplayalım. t düğümündeki *s* aday bölünmelerinin *uygunluk* ölçüsü şu şekilde hesaplanır:

$$\Phi(s|t) = 2P_{sol}P_{sağ}\sum_{j=1}^{n}|P(j|t_{sol})-P(j|t_{sağ})|$$

Burada j sınıf değerleri GEÇERLİ ve GEÇERSİZ biçiminde olduğuna göre yukarıdaki bağıntı şu şekilde ifade edilebilir:

$$\Phi(s|t)=2P_{sol}P_{sağ}\left[\left|P(GEÇERLİ|t_{sol})-P\left(GEÇERLİ\middle|t_{sağ}\right)\right|+\left|P(GEÇERSİZ|t_{sol})-P\left(GEÇERSİZ\middle|t_{sağ}\right)\right|\right]$$

Olasılık teorisine göre P_{sol} ve $P_{sağ}$ olasılıkları için $P_{sol} + P_{sağ}=1$ kuralı geçerlidir. Bu durumda $P_{sağ}$ şu şekilde hesaplanabilir.

$P_{sağ} = 1 - P_{sol} \Rightarrow 1-0.40 = 0.60$

Benzer biçimde $P(GEÇERLİ|t_{sağ}) + P(GEÇERSİZ|t_{sağ}) = 1$ olduğundan $P(GEÇERSİZ|t_{sağ})$ ifadesi şu şekilde hesaplanır:

$P(GEÇERSİZ|t_{sağ}) = 1 - P(GEÇERLİ|t_{sağ}) \Rightarrow 1-0.33=0.67$

Benzer biçimde aşağıdaki hesaplama yapılabilir:

$P(GEÇERSİZ|t_{sol}) = 1 - P(GEÇERLİ|t_{sol}) \Rightarrow 1-0.50=0.50$

Bu durumda; $\Phi(s|t) = 2(0.40)(0.60)\left[\left|0.50-0.33\right| + \left|0.50-0.67\right|\right]$

$=0.16$

Uygulama

Tablo 3 de verilen eğitim verilerini göz önüne alalım. *Twoing* algoritmasını kullanmak sınıflandırma işlemini gerçekleştirmek istiyoruz. Bu veri kümesi bir firmanın 11 adet müşteri bilgisini içermektedir. Söz konusu veriler GELİR, EĞİTİM ve SEKTÖR nitelikleri ile MEMNUN isimli nitelikten oluşmaktadır. MEMNUN niteliği sınıf değerlerini içeren hedef niteliğidir. Hedef niteliği müşterilerin firmadan memnun olup olmadıklarını belirlemektedir.

Müşteri	GELIR	EGITIM	SEKTOR	MEMNUN
1	NORMAL	ORTA	BİLİŞİM	EVET
2	BÜYÜK.	İLK	BİLİŞİM	EVET
3	KÜÇÜK	İLK	İNŞAAT	EVET
4	BÜYÜK	ORTA	İNŞAAT	EVET
5	KÜÇÜK	ORTA	İNŞAAT	EVET
6	BÜYÜK	LİSE	İNŞAAT	EVET
7	KÜÇÜK	LİSE	İNŞAAT	EVET
8	BÜYÜK	ORTA	BİLİŞİM	HAYIR
9	KÜÇÜK	ORTA	BİLİŞİM	HAYIR
10	BÜYÜK	LİSE	BİLİŞİM	HAYIR
1 1	KÜÇÜK	LİSE	BİLİŞİM	HAYIR

Tablo 3

Adım1;

a) *Twoing* algoritmasını uygulamak için niteliklerin her bir değeri için iki ayrı dizi oluşturulur. Burada s aday bölünmenin her bir satırını ifade etmektedir. Örneğin GELİR=NORMAL olarak alınırsa bu sol taraf dizisinin elemanı olacaktır. Geriye kalan BÜYÜK ve KÜÇÜK nitelik değerleri için GELİR ϵ {BÜYÜK, KÜÇÜK} sağ taraf dizi elemanını oluşturur. İki diziden sol tarafta bulunanı t_{Sol}, sağ tarafta yer alanı ise $t_{sağ}$ dizisi olarak değerlendirilir.

Aday bölünme (s)	t_{sol}	$t_{sağ}$
1	GELİR=NORMAL	GELİR ϵ {BÜYÜK,KÜÇÜK}
2	GELİR=BÜYÜK	GELİR ϵ {NORMAL,KÜÇÜK}
3	GELİR=KÜÇÜK	GELİR ϵ {BÜYÜK,NORMAL}
4	EĞİTİM=İLK	EĞİTİM ϵ {ORTA,LİSE}
5	EĞİTİ M=ORTA	EĞİTİM ϵ {İLK, LİSE}
6	EĞİTİM=LİSE	EĞİTİ M ϵ {İLK,ORTA}
7	SEKTÖR=BİLİŞİM	SEKTÖR= İNŞAAT
8	SEKTÖR=İNŞAAT	SEKTÖR= BİLİŞİM

Aday bölünmeler/Tablo4

b) GELİR=NORMAL için P_{Sol} ve $P(j|t_{Sol})$ olasılıklarının hesaplanması:

Tablo 4'deki her bir nitelik değerinin Tablo 3 de GELİR niteliği içindeki tekrar sayılarını belirlememiz gerekiyor. Örneğin GELİR=NORMAL değerini Tablo3de GELİR sütununda 1 kez tekrar edildiği görülmektedir. Eğitim kümesinde 11 satır yer almaktadır. Bu durumda GELİR=NORMAL elde etme olasılığı olan P_{sol} değeri şu şekilde hesaplanır:

P_{sol}= 1/11= 0.09

Şimdi $P(j|t_{Sol})$ değerini hesaplayalım. Burada j sınıfları gösterir. MEMNUN isimli sınıf niteliğinin EVET ve HAYIR biçiminde iki değeri vardır. O halde P(E*VET*| t_{sol}) ve $P(HAYIR \mid t_{Sol})$ değerlerinin hesaplanması gerekmektedir.

Şimdi Tablo 3 de GELİR sütununda NORMAL değerlerinin hangi satırlarda olduğuna bir bakalım. Tabloda sadece birinci satırda GELİR=NORMAL değerlerine sahiptir. Bu satırın karşısında EVET değerinin olup olmadığına bakıyoruz. Söz konusu konumda EVET yer aldığına göre, bu değer ile ilgili koşullu olasılık,

$P(EVET|t_{Sol}) = 1/1 = 1$

$P(HAYIR|t_{Sol}) = 0/1 = 0$ elde edilir.

Bu hesaplamalar Tablo 4 deki tüm aday bölünmeler için tekrarlanırsa aşağıdaki tablo elde edilir:

Aday Bölünme	t_{sol}'daki kayıt sayısı	P_{sol}	EVET sayısı	HAYIR sayısı	$P(EVET\|t_{sol})$	$P(HAYIR\|t_{sol})$
1	1	0.09	1	0	1.00	0.00
2	5	0.45	3	2	0.60	0.40
3	5	0.45	3	2	0.60	0.40
4	2	0.18	2	0	1.00	0.00
5	5	0.45	3	2	0.60	0.40
6	4	0.36	2	2	0.50	0.50
7	6	0.55	2	4	0.33	0.67
8	5	0.45	5	0	1.00	0.00

Tablo5

c) GELİR ϵ {BÜYÜK,KÜÇÜK} için $P_{Sağ}$ ve $P(j \mid t_{Sağ})$ olasılıklarının hesaplanması:

GELİR=BÜYÜK ve GELİR=KÜÇÜK değerlerinin eğitim kümesi içindeki tekrar sayılarını belirlememiz gerekiyor. Bu tekrar sayısının 10 olduğu anlaşılıyor. O halde $P_{Sağ}$ değeri şu şekilde hesaplanır:

$P_{Sağ}$=10/11=0.91

Eğitim kümesinde GELİR=BÜYÜK ve GELİR=KÜÇÜK değerlerinin yer aldığı satırları göz önüne alalım. Bu satırlardan kaç tanesinde EVET kaç tanesinde HAYIR sınıf değerlerinin var olduğunu belirleyelim. Yani $P(EVET \mid t_{Sağ})$ ve $P(HAYIR|t_{Sağ})$ koşullu olasılık değerini

$P(EVET|t_{Sağ}) = 6/10 = 0.6$

$P(HAYIR|t_{Sağ}) = 4/10 = 0.4$

Benzer biçimde diğer satırlarda hesaplamalar yapılırsa aşağıdaki tablonun elde edildiği görülür.

Aday Bölünme	$t_{sağ}$'daki kayıt sayısı	$P_{sağ}$	EVET sayısı	HAYIR sayısı	P(EVETItsag)	P(HAYIRItsag)
1	10	0.91	6	4	0.60	0.40
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	1.00	0.00
8	6	0.55	2	4	0.33	0.67

Tablo6

d) $\Phi(s|t)$ uygunluk ölçütünün hesaplanması:

Uygunluk Ölçüsü Formülü:

$$\Phi(s|t) = 2P_{sol}P_{sağ}\sum_{j=1}^{n}|P(j|t_{sol})-P(j|t_{sağ})|$$

Tablo 5 ve Tablo 6 de elde edilen değerleri burada yerine yazarak her satır için uygunluk ölçütü hesaplanır. Bu *t* düğümünde GELİR=NORMAL ve GELİR ϵ {BÜYÜK,KÜÇÜK} biçimindeki ilk aday bölünme için söz konusu hesaplamayı sadece birinci satır için yapıyoruz. Burada s=l olarak kabul edilir.

$$\Phi(1|t) = 2(0.09)(0.91)\,[|1 - 0.6)| + |(0 - 0.4)|]$$

$$= 0.13$$

Diğer bütün satırlar için aynı hesaplamalar yapıldığında aşağıdaki tablo elde edilir.

Aday Bölünme	P_{sol}	$P_{sağ}$	$2\,P_{sol}\,P_{sağ}$	$\Phi(s\|t)$
1	0.09	0.91	0.17	0.13
2	0.45	0.55	0.50	0.07
3	0.45	0.55	0.50	0.07
4	0.18	0.82	0.30	0.26
5	0.45	0.55	0.50	0.07
6	0.36	0.64	0.46	0.20
7	0.55	0.45	0.50	0.66
8	0.45	0.55	0.50	0.66

Tablo7

e) En büyük uygunluk ölçütünün seçilmesi:

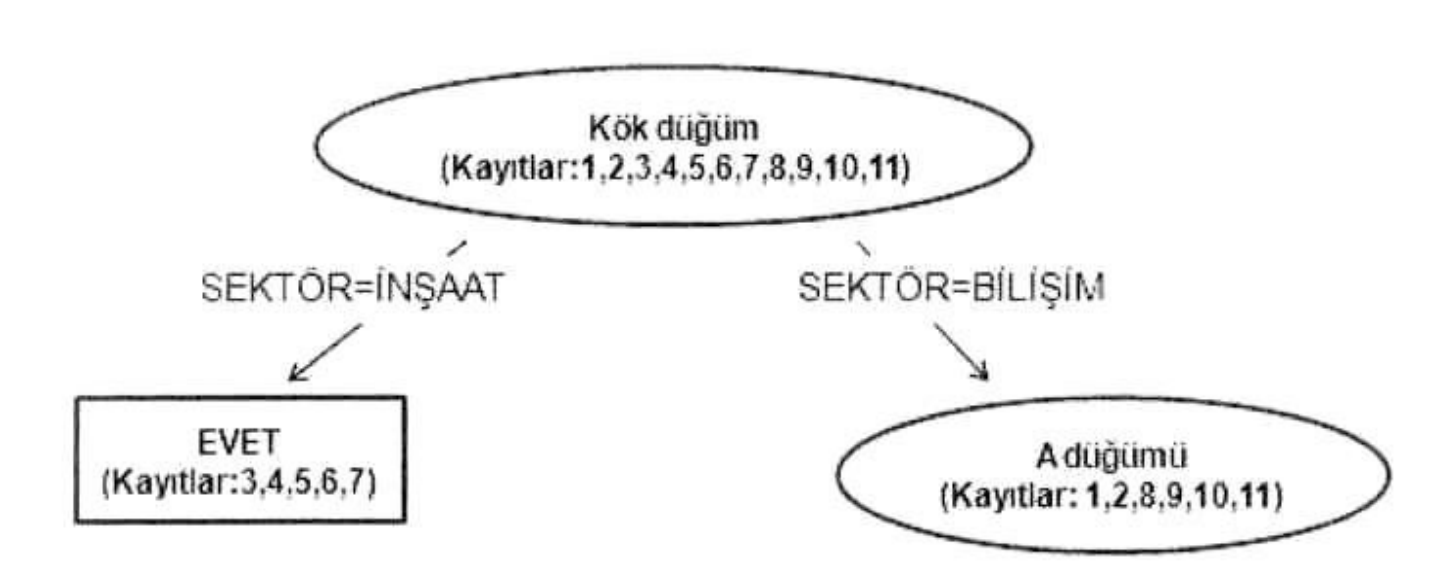
Tablo7 üzerinde Φ (s|t) sütununda en büyük değer 7 ve 8. satır üzerinde yer alan

0.66 değeridir. Tablo 4 de 8. satırda birinci bölünmede SEKTÖR=İNŞAAT yer aldığına göre, Tablo 3 de SEKTÖR niteliği içinde İNŞAAT değerleri araştırılır. Bu değerler 3, 4, 5, 6, 7 satırlarındadır. Bu durumda, eğitim serisinde kök düğümden itibaren nasıl bir ayrım yapılacağı belli olmuştur.

(3, 4, 5, 6, 7) satırları ve geri kalan (1, 2, 8, 9, 10, 11) kayıtları biçiminde bir bölünme yapılır. (3, 4, 5, 6, 7) kayıtlarının tümü EVET sınıf değerine sahip olduğuna göre birinci ayrım sonlanmıştır. Geri kalanlar ise yeni bir A karar düğümü oluşturur.

f) Karar Ağacının Oluşturulması:

Bulunan değerler göz önüne alınarak karar ağacı şu şekilde çizilebilir.



Şekil 1:Karar Ağacı

Adım 2;

Birinci adımda (3, 4, 5, 6, 7) satırları için karar verilmiştir. O halde ikinci adımda (1, 2, 8, 9, 10, 11) satırlarından oluşan A düğümü göz önüne alınarak birinci adımdaki işlemler tekrarlanır. Bunun için önce eğitim setinde sonlanan (3, 4, 5, 6, 7) satırları çıkarılarak Tablo 3 yeniden düzenlenir ve aşağıdaki Tablo 8 elde edilir.

Müşteri	GELİR	EĞİTİM	SEKTÖR	MEMNUN
1	NORMAL	ORTA	BİLİŞİM	EVET
2	BÜYÜK	İLK	BİLİŞİM	EVET
8	BÜYÜK	ORTA	BİLİŞİM	HAYIR
9	KÜÇÜK	ORTA	BİLİŞİM	HAYIR
10	BÜYÜK	LİSE	BİLİŞİM	HAYIR
11	KÜÇÜK	LİSE	BİLİŞİM	HAYIR

Tablo8

a) Tablo 8 den yaralanılarak aday bölünmeler şu şekilde elde edilir:

Aday bölünme	Tsol	tsağ
1	GELİR=NORMAL	GELİR∈{BÜYÜK,KÜÇÜK}
2	GELİR=BÜYÜK	GELİR ∈ {NORMAL,KÜÇÜK}
3	GELİR=KÜÇÜK	GELİR ∈ (BÜYÜK,NORMAL}
4	EGİTİM=İLK	EĞİTİM∈{ORTA,LİSE }
5	EĞİTİM=ORTA	EĞİTİM ∈ {İLK, LİSE}
6	EĞİTİM LİSE	EĞİTİM ∈ {İLK,ORTA}

Tablo 9

b) GELİR=NORMAL için $\boldsymbol{P_{Sol}}$ ve $\boldsymbol{P(j|t_{Sol})}$ olasılıklarının hesaplanması:

GELİR=NORMAL değerinin Tablo 8 de GELİR sütununda 1 kez tekrar edildiği görülmektedir. Eğitim kümesinde 6 satır yer almaktadır. Bu durumda GELİR=NORMAL elde etme olasılığı olan P_{Sal} değeri şu şekilde hesaplanır:

$P_{sol}=1/6 = 0.17$

$P(EVET\ |t_{Sol})$ ve $P(HAYIR\ |t_{sol})$ değerlerinin hesaplanması için Tablo 8 de GELİR sütununda NORMAL değerlerinin hangi satırlarda olduğuna bir bakalım. Tabloda sadece birinci satırda GELİR=NORMAL değerlerine sahiptir. Söz konusu satırın karşısında EVET değeri yer aldığına göre bu değer ile ilgili koşullu olasılık,

$P(EVET\ |t_{Sol}) = 1/1 = 1$

$P(HAYIR\ |t_{sol}) = 0/1 = 0$ *elde edilir.*

Bu hesaplamalar Tablo 9daki tüm aday bölünmeler için tekrarlanırsa aşağıdaki tablo elde edilir:

Aday Bölünme	T_{sol} **daki kayıt sayısı**	P_{sol}	EVET **sayısı**	HAYIR **sayısı**	$P(EVET\|t_{sol})$	$P(HAYIR\|t_{sol})$
1	1	0.17	1	0	**1.00**	**0.00**
2	3	0.50	**1**	2	0.33	0.67
3	2	0.33	0	2	**0.00**	**1.00**
4	1	0.17	**1**	**0**	**1.00**	**0.00**
5	3	0.50	1	2	0.33	0.67
6	2	0.33	0	2	**0.00**	**1.00**

Tablo 10

c) GELİR ϵ{BÜYÜK,KÜÇÜK} için P_{Sag} ve $P(j|t_{sağ})$ olasılıkları hesaplanması:

GELİR=BÜYÜK ve GELİR=KÜÇÜK değerlerinin eğitim kümesi içindeki tekrar sayılarını belirlememiz gerekiyor. Tekrar sayısı 5 olduğuna göre $P_{sağ}$ değeri şu şekilde hesaplanır:

$P_{sağ}$=5/6= 0.83

Eğitim kümesinde GELİR=BÜYÜK ve GELİR=KÜÇÜK değerlerinin yer aldığı satırları göz önüne alalım. Bu satırlardan 1 tanesinde EVET, 4 tanesinde HAYIR sınıf değerinin var olduğu görülür. O halde $P(EVET|t_{Sağ})$ ve $P(HAYIR|t_{Sağ})$ koşullu olasılığı şu şekilde hesaplanır:

$P(EVET \mid t_{Sağ}) = 1/5 = 0.2$

$P(HAYIR \mid t_{Sağ}) = 4/5 = 0.8$

Benzer biçimde diğer satırlarda hesaplamalar yapılırsa aşağıdaki tablonun elde edildiği görülür

Aday Bölünme	$T_{sağ}$ daki kayıt sayısı	$P_{sağ}$	EVET sayısı	HAYIR sayısı	$P(EVET \mid t_{sağ})$	$P(HAYIR \mid t_{sağ})$
1	5	0.83	1	4	0.20	0.80
2	3	0.50	1	2	0.33	0.67
3	4	0.67	2	2	0.50	0.50
4	5	0.83	1	4	0.20	0.80
5	3	0.50	1	2	0.33	0.67
6	4	0.67	2	2	0.50	0.50

Tablo 11

d) **Φ(s|t)** uygunluk ölçütünün hesaplanması:

Bu t düğümünde GELİR=NORMAL ve GELİR ϵ {BÜYÜK,KÜÇÜK} biçimindeki ilk aday bölünme için söz konusu hesaplamayı sadece birinci satır için yapıyoruz:

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

Hesaplamalar diğer tüm satırlar için yapıldığında Sonuç olarak tablo 12 elde edilir.

Aday Bölünme	*Psol*	*Psağ*	*2PsolPsağ*	*Φ (s|t)*
1	*0.17*	*0.83*	*0.28*	*0.44*
2	*0.50*	*0.50*	*0.50*	*0.00*
3	*0.33*	*0.67*	*0.44*	*0.44*
4	*0.17*	*0.83*	*0.28*	*0.44*
5	*0.50*	*0.50*	*0.50*	*0.00*
6	*0.33*	*0.67*	*0.44*	*0.44*

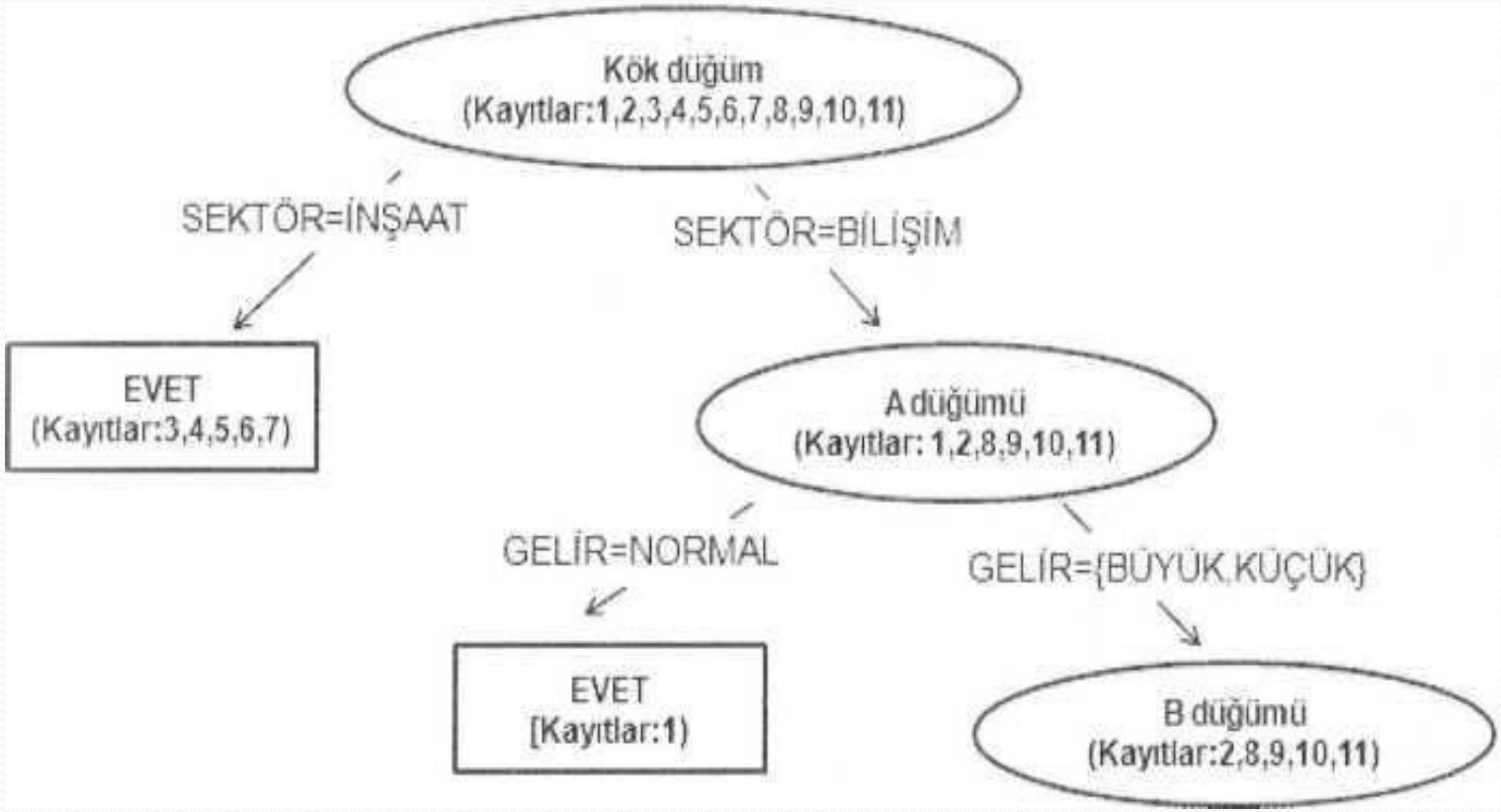
Tablo12

e) En Büyük uygunluk ölçütünün seçilmesi:

Tablo 12 üzerinde Φ (s|t) sütununda en büyük değer 0.44 olup bu değer 1, 3, 4 ve 8. satırda yer almaktadır. Bu değerler birbirine eşit olduğundan en büyük uygunluk ölçütü olarak herhangi biri seçilebilir. Biz birinci satırdaki değeri seçiyoruz. Bu durumda, aday bölünmelerin yer aklığı Tablo 9 üzerinde birinci satırda yer alan GELİR=NORMAL ve GELİR ϵ {BÜYÜK,KÜÇÜK} değerlerine göre bir dallanma olacaktır. Tablo 8 de GELİR=NORMAL değerinin nerelerde olduğu araştırılır. Sadece 1. satırda bu değerin yer aldığı görülüyor. O halde dallanma (1) ve (2, 8, 9, 10, 11) biçiminde olacaktır.

f) Karar Ağacı:

Bulunan değerler göz önüne alınarak karar ağacı şu şekilde çizilebilir.



Şekil 2

Adım3: Bu adımda (2,8,9,10,11) satırları için karar verilmesi gerekmektedir. Yeni eğitim kümesi şu şekilde olacaktır.

Müşteri	GELİR	EGİTİM	SEKTÖR	MEMNUN
2	BÜYÜK	İLK	BİLİŞİM	EVET
8	BÜYÜK	ORTA	BİLİŞİM	HAYIR
9	KÜÇÜK	ORTA	BİLİŞİM	HAYIR
10	BÜYÜK	LİSE	BİLİŞİM	HAYIR
11	KÜÇÜK	LİSE	BİLİŞİM	HAYIR

Tablo 13

a) Aday bölünme tablosu ise, yukarıdaki eğitim kümesine bağlı olarak şu şekilde düzenlenebilir:

Aday bölünme	t_{sol}	$t_{sağ}$
1	GELİR=BÜYÜK	GELİR=KÜÇÜK
2	GELİR=KÜÇÜK	GELİR=BÜYÜK
3	EĞİTİM=İLK	EĞİTİM ϵ {ORTA,LİSE}
4	EĞİTİM=ORTA	EĞİTİM ϵ {İLK, LİSE}
5	EĞİTİM=LİSE	EĞİTİM ϵ [İLK,ORTA}

Tablo 14

b) GELİR=BÜYÜK için $\boldsymbol{P_{sol}}$ ve $\boldsymbol{P(j|t_{Sol})}$ olasılıklarının hesaplanması:

GELİR=BÜYÜK değerinin Tablo 13 de GELİR sütununda 3 kez tekrar edildiği görülmektedir. Eğitim kümesinde ise 5 satır yer almaktadır. O halde GELİR=BÜYÜK elde etme olasılığı olan P_{Sol} değeri şu şekilde hesaplanır:

$\boldsymbol{P_{sol}}$=3/5=0.6

$P(EVET|t_{sol})$ ve $P(HAYIR|t_{sol})$ değerlerinin hesaplanması için Tablo 13 de GELİR sütununda BÜYÜK değerinin hangi satırlarda olduğu belirlenir. Tabloda üç satırda bu değer yer almaktadır. O halde EVET ve HAYIR değerleri için koşullu olasılık,

$P(EVET\ |t_{Sol}) = 1/3 = 1$

$P(HAYIR\ |t_{sol}) = 2/3 = 0.67$ *elde edilir.*

Tüm aday bölünmeler için tekrarlanırsa aşağıdaki tablo elde edilir.

Aday Bölünme	t_{sol} daki kayıt sayısı	Psol	EVET sayısı	HAYIR sayısı	$P(EVET\|t_{sol})$	$P(HAYIR\|t_{sol})$
1	3	0.60	1	2	0.33	0.67
2	2	0.40	0	2	0.00	1.00
3	1	0.20	1	0	1.00	0.00
4	2	0.40	0	2	0.00	1.00
5	2	0.40	0	2	0.00	1.00

Tablo 15

c) GELİR=KÜÇÜK için $P_{Sağ}$ ve $P(j \mid t_{Sağ})$ olasılıklarının hesaplanması:

GELİR=KÜÇÜK değerlerinin eğitim kümesi içindeki tekrar sayısı 2 olduğuna göre $P_{sağ}$ değeri şu şekilde hesaplanır:

$P_{Sağ}=2/5=0.4$

Bu kez GELİR=KÜÇÜK değerlerinin yer aldığı satırları göz önüne alalım. Söz konusu satırlar 2 tanedir. Bu satırlardan hiçbirinde EVET yoktur. Buna karşılık 2 tane HAYIR değeri bulunduğu görülür. O halde $P(EVET \mid t_{Sağ})$ ve $P(HAYIR \backslash t_{Sağ})$ koşullu olasılık değerleri şu şekilde hesaplanır:

$P(EVET \mid t_{sağ}) = 0/2 = 0$

$P(HAYIR \mid t_{sağ}) = 2/2 = 1$

Benzer biçimde diğer satırlarda hesaplamalar yapılırsa aşağıdaki tablonun elde edilir.

Aday Bölünme	$T_{sağ}$ daki kayıt sayısı	$P_{sağ}$	EVET sayısı	HAYIR sayısı	$P(EVET|t_{sağ})$	$P(HAYIR|t_{sağ})$
1	2	0.40	**0**	2	**0.00**	**1.00**
2	3	0.60	**1**	2	0.33	0.67
3	4	0.80	**0**	4	**0.00**	**1.00**
4	3	0.60	**1**	2	0.33	0.67
5	3	0.60	**1**	2	0.33	0.67

Tablo16

d) Φ(s|t) uygunluk ölçütünün hesaplanması:

Bu t düğümünde GELİR=KÜÇÜK ve GELİR=BÜYÜK biçimindeki ilk aday bölünme için söz konusu hesaplamayı sadece birinci satır için yapıyoruz:

$$\Phi(1|t) = 2(0.6)(0.4)\,[|0.33 - 0)| + |(0.67 - 1)|] = 0.32$$

Hesaplamalar bütün satırlar için yapıldığında:

Aday Bölünme	P_{sol}	$P_{sağ}$	$2\ P_{sol}\ P_{sağ}$	$\Phi(s\|t)$
1	0.60	0.40	0.48	0.32
2	0.40	0.60	0.48	0.32
3	0.20	0.80	0.32	0.64
4	0.40	0.60	0.48	0.32
5	0.40	0.60	0.48	0.32

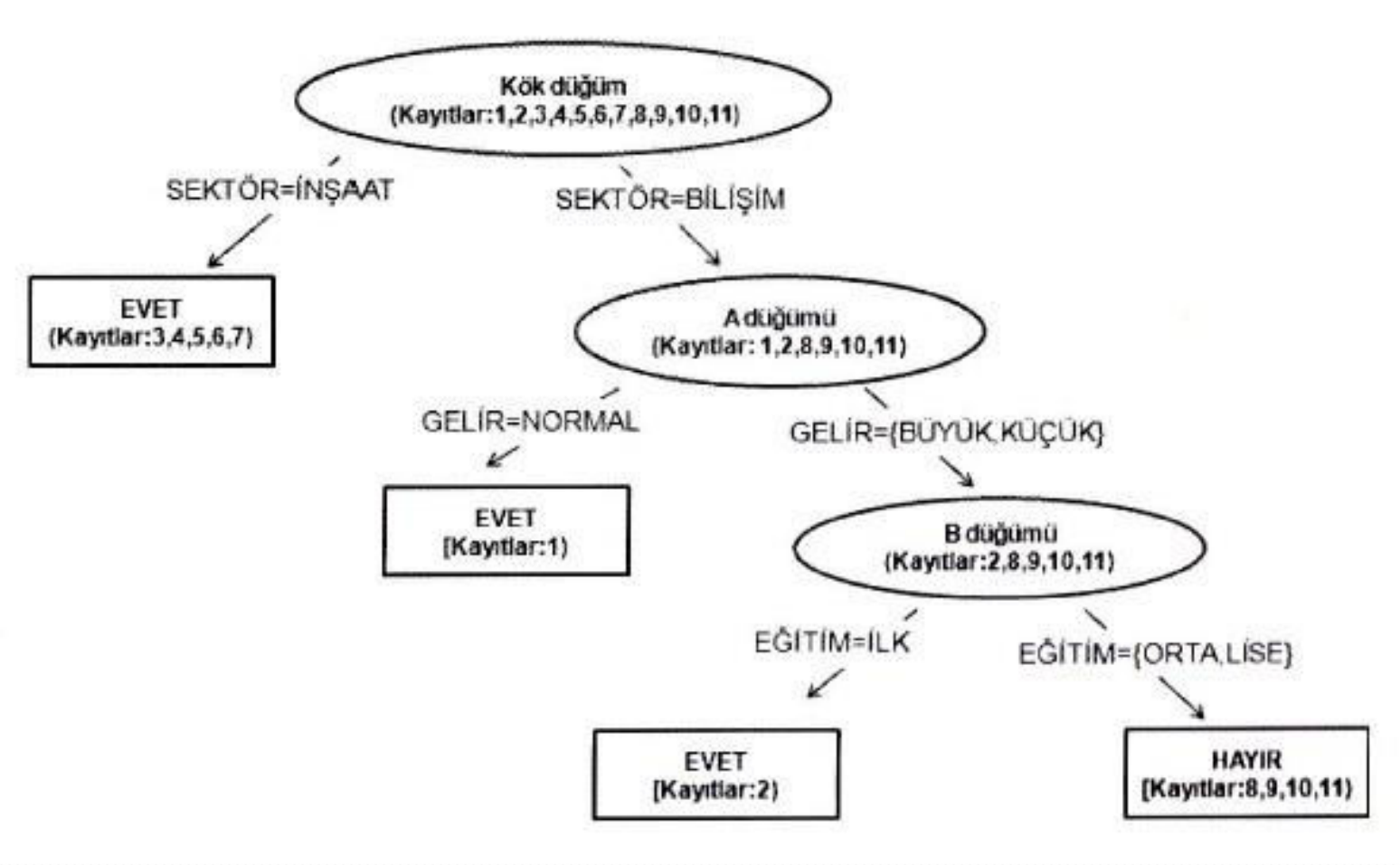
Tablo 17

En büyük uygunluk ölçütünün seçilmesi:

Tablo17 üzerinde **$\Phi(s|t)$** sütununda en büyük değer 0.64 olduğu görülmektedir.

Bu değer **3.** satırda yer almaktadır. Tablo **14** deki aday bölünme tablosunda aynı satır EGİTİM İLK değerine işaret etmektedir. Bu değer Tablo **13** de 2 numaralı müşteriye aittir. O halde bu noktadan itibaren bir bölünme yapılacaktır. Yani (2) ve (8, 9, 10, 11) biçiminde bir dallanma olacaktır. Söz konusu dallanma incelendiğinde (2) numaralı satırın EVET, (8, 9, **10, 11**) numaralı satırların ise HAYIR sınıf değerine sahip olduğu anlaşılır. O halde bölünme işlemi sona ermiştir.

f) Karar Ağacı: Bulunan değerler göz önüne alınarak karar ağacı şu şekilde çizilebilir:



Şekil 3

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 EĞİTİM=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;

Gini Algoritması:

İkili bölünmemelere dayalı bir diğer sınıflandırma yöntemi *Gini algoritması* olarak isimlendirilmektedir. Bu algoritma, nitelik değerlerinin sol ve sağda olmak üzere iki bölüme ayrılması esasına dayanmaktadır. *Gini algoritması* şu şekilde uygulanır:

a) Her nitelik değerleri ikili olacak biçimde gruplanır. Bu şekilde elde edilen sol ve sağ bölünmelere karşılık gelen sınıf değerleri gruplandırılır.

b) Her bir nitelikle ilgili sol ve sağ taraftaki bölünmeler için $Gini_{sol}$ ve $Gini_{sağ}$ değerleri hesaplanır.

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

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

Bu bağlantıda yer alan ifadeler şu şekildedir:

k	Sınıfların sayısı
T	Bir düğümdeki örnekler
$\|T_{Sol}\|$	Sol taraftaki örneklerin sayısı
$\|T_{Sağ}\|$	Sağ taraftaki örneklerin sayısı
L_i	Sol tarafta *i* kategorisindeki örneklerin sayısı
R_i	Sağ tarafta *i* kategorisindeki örneklerin sayısı

c) Her j niteliği için, *n* eğitim kümesindeki satır sayısı olmak üzere aşağıdaki bağıntının değeri hesaplanır:

$$Gini_j = \frac{1}{n}\left(\left|T_{Sol}\right| Gini_{Sol} + \left|T_{Sağ}\right| Gini_{Sağ}\right)$$

d) Her j niteliği için hesaplanan $Gini_j$ değerleri arasından en küçük olanı seçilir ve bölünme bu nitelik üzerinden gerçekleştirilir.

e) 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.

RISK	BORÇ		GELİR		STATU	
	YÜKSEK	DÜŞÜK	YÜKSEK	DÜŞÜK	İŞVEREN	ÜCRETLİ
IYI	2	1	4	2	2	3
KOTU	3	2	1	1	2	1

Bu bilgileri kullanarak $Gini_{Sol}$ ve $Gini_{sağ}$ ile $Gini_{borç}$ değerlerini hesaplarsak;

Çözüm

Tablo üzerinde görüldüğü gibi, BORÇ niteliği YÜKSEK değeri sol tarafta, DÜŞÜK değeri ise sağ tarafta olacak biçimde bölünmüştür. Diğer nitelikler de benzer biçimde bölünmüşlerdir. Bölünen her bir değerin kaç tanesinin İYİ, kaç tanesinin KÖTÜ sınıf değerine sahip olduğu belirlenmiştir. Örneğin, BORÇ niteliğinin 2 adet YÜKSEK değeri İYİ, 3 tanesi ise KÖTÜ sınıf değerine sahiptir. Burada RİSK niteliği hedef niteliktir. Bir başka deyişle sınıfları içeren niteliktir.

Sınıf niteliği iki değer içerdiğinden k=2 olarak kabul edilir.

T_{Sol} değeri sol taraftaki eleman sayısı olduğuna göre $T_{Sol} = 2 + 3 = 5$ olduğu anlaşılır. Sol taraf için $L_1 = 2$ ve $L_1 = 3$ olarak belirlenir. Benzer biçimde, $T_{Sag} = 1 + 2 = 3$, $R_1 = 1$, $R_2 = 2$ olduğuna göre aşağıda belirtilen hesaplamalar yapılabilir:

$$Gini_{Sol} = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_{Sol}|} \right)^2 = 1 - \left[\left(\frac{2}{5} \right)^2 + \left(\frac{3}{5} \right)^2 \right] = 0.48$$

$$Gini_{Sağ} = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_{Sağ}|} \right)^2 = 1 - \left[\left(\frac{1}{3} \right)^2 + \left(\frac{2}{3} \right)^2 \right] = 0.44$$

$$Gini_j = \frac{1}{n} \left(|T_{Sol}| Gini_{Sol} + |T_{Sağ}| Gini_{Sağ} \right) = \frac{(5)(0.48)+(3)(0.44)}{8} = 0.46$$

UYGULAMA

Aşağıdaki eğitim verilerini göz önüne alalım. Bu verilere dayanarak *Gini* algoritmasını kullanarak sınıflandırma işlemini yapacağız.

Başvuru	EĞİTİM	YAŞ	CİNSİYET	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

Tablo 19

Adım1:

a) **Nitelik değerlerinin ikili gruplandırılması:**

Bu eğitim verisi üzerinde *Gini* algoritmasını uygulayabilmek için önce aşağıda belirtilen hesaplamalar yapılır. Bu tabloya göre EVET sınıfına ilişkin olarak EĞİTİM niteliğinin İLK değerinden 1 tane bulunmaktadır. Benzer biçimde (ORTA, YÜKSEK) değerlerinden ise 3 tane bulunmaktadır. Bu şekilde diğer değerler de hesaplanır.

KABUL	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA YÜKSEK	GENÇ	ORTA YAŞLI	KADIN	ERKEK
EVET	1	3	0	4	1	3
HAYIR	2	1	2	1	2	1

Tablo 20

b) $Gini_{sol}$

EĞİTİM için $Gini_{sağ}$ değerlerinin hesaplanması: Tablo değerlerini kullanarak şu hesaplamalar yapılabilir:

EĞİTİM İçin:

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.444$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$$

YAŞ İçin:

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

$$Gini_{sağ} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{6}\right)^2\right] = 0.320$$

CİNSİYET İçin:

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.444$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$$

c)$Gini_j$ değerlerinin hesaplanması: Elde edilen sonuçlar kullanılarak her bir nitelik için Gini değerleri elde edilir.

$$Gini_{eğitim} = \frac{3(0.444) + 4(0.375)}{7} = 0.405$$

$$Gini_{yaş} = \frac{2(0) + 5(0.320)}{7} = 0.229$$

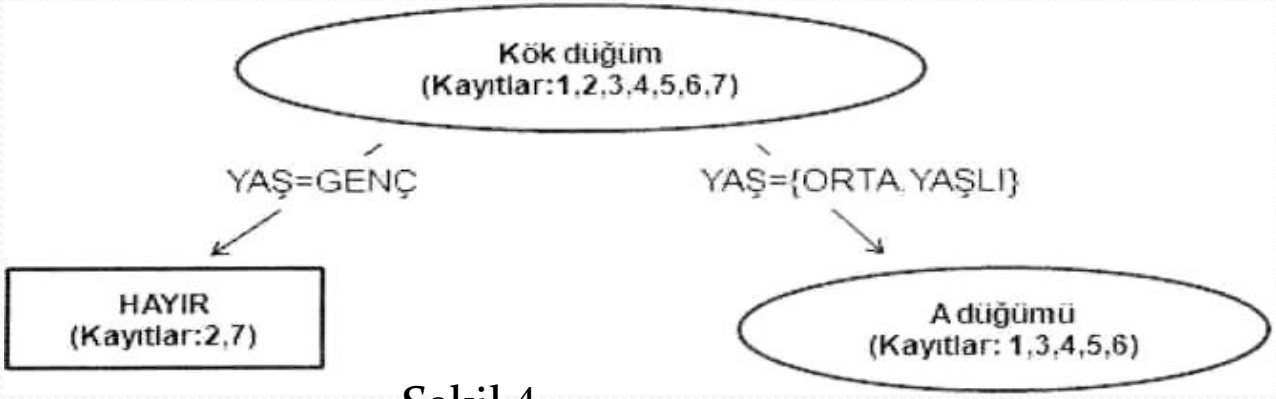
$$Gini_{cinsiyet} = \frac{3(0.444) + 4(0.375)}{7} = 0.405$$

Sonuç olarak şu şekilde bir tablo elde edilir:

KABUL	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA, YÜKSEK	GENÇ	ORTA, YAŞLI	KADIN	ERKEK
EVET	1	3	0	4	1	3
HAYIR	2	1	2	1	2	1
$Gini_{sol}$ $Gini_{sağ}$	0.444	0.375	0.000	0.320	0.444	0.375
$Gini_j$	0.405		0.229		0.405	

Tablo 21

En küçük *Gini* değerinin seçilmesi: Yukarıdaki tabloda hesaplanan değerler göz önüne alındığında, $Gini_{yaş}$ = 0.229 değerinin $Gini_j$ değerleri içinde en küçüğü olduğu anlaşılır. O halde kök düğümünden itibaren bölünme YAŞ=GENÇ ve *YAŞ* ϵ {ORTA,YAŞLI] biçiminde olacaktır. Bölünmeyi elde etmek için Tablo 19 üzerinde YAŞ=GENÇ değerleri aranır. Bu değer (2, 7) satırları üzerindedir. O halde bölünme (2, 7) ve geri kalan satırlardan (1, 3, 4, 5, 6) oluşacaktır. Söz konusu bölünmeyi aşağıdaki şekil üzerinde görüyoruz



Şekil 4

Adım 2:

Benzer işlemleri ikinci adımda tekrarlayacağız. Bunun için önce eğitim kümesinden (2, 7) satırlarını çıkarıyoruz. Bu adımdaki hesaplamaları bu yeni tabloya göre yapacağız.

Başvuru	EĞİTİM	YAŞ	CİNSİYET	KABUL
1	ORTA	YAŞLI	ERKEK	EVET
3	YÜKSEK	ORTA	KADIN	HAYIR
4	ORTA	ORTA	ERKEK	EVET
5	İLK	ORTA	ERKEK	EVET
6	YÜKSEK	YAŞLI	KADIN	EVET

Tablo 21

a) Nitelik değerlerinin ikilik gruplandırılması:

Yukarıdaki tablodan yararlanarak, her bir nitelik değerinden her bir sınıfa kaç adet olduğu belirlenir ve şu şekilde bir tablo hazırlanır.

	EĞİTİM		YAŞ		CİNSİYET	
KABUL	İLK	ORTA YÜKSEK	ORTA	YAŞLI	KADIN	ERKEK
EVET	1	3	2	2	1	3
HAYIR	0	1	1	0	1	0

Tablo 22

b) $Gini_{sol}$ **ve** $Gini_{sağ}$ **değerlerinin hesaplanması:** Tablo değerlerini kullanarak aşağıdaki hesaplamalar yapılabilir.

EĞİTİM İçin:

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

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

YAŞ İçin:

$$Gini_{sol} = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.444$$

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

CİNSİYET İçin:

$$Gini_{sol} = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = 0{,}500$$

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

c) $Gini_j$ **değerlerinin hesaplanması:** Bu sonuçlar kullanılarak her bir nitelik için aşağıdaki $Gini_j$ değerleri elde edilir.

$$Gini_{eğitim} = \frac{1(0) + 4(0.375)}{5} = 0.300$$

$$Gini_{yaş} = \frac{3(0.444) + 2(0)}{5} = 0.267$$

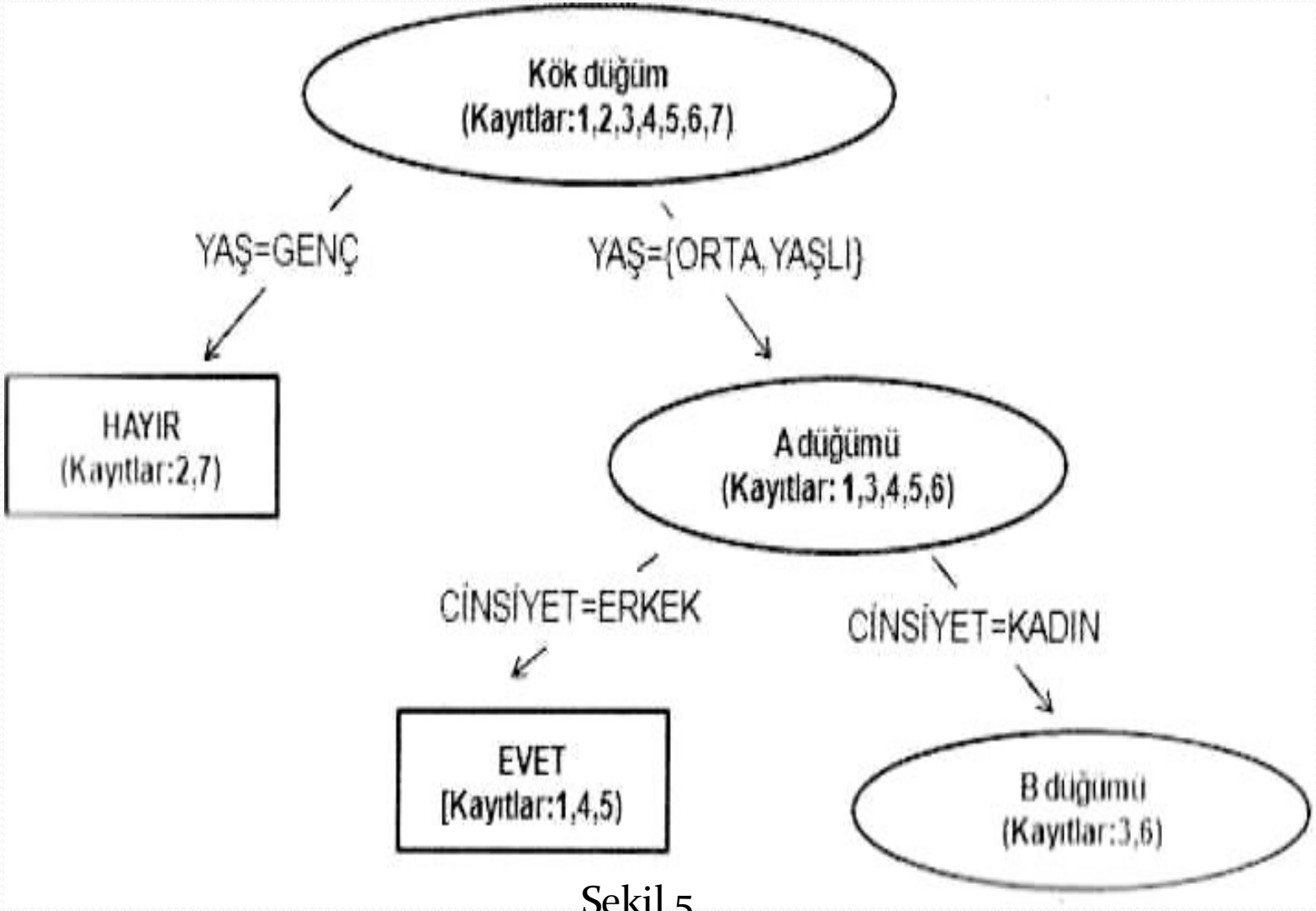
$$Gini_{cinsiyet} = \frac{2(0.500) + 3(0)}{5} = 0.200$$

Sonuç olarak şu şekilde bir tablo elde edilir

KABUL	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA YÜKSEK	ORTA	YAŞLI	KADIN	ERKEK
EVET	1	3	2	2	1	3
HAYIR	0	1	1	0	1	0
$Gini_{sol}$, $Gini_{sağ}$	0.000	0.375	0.444	0.000	0.500	0.000
$Gini_j$	0.300		0.267		0.200	

Tablo 23

d) En küçük $Gini_j$ **değerinin seçilmesi:** Yukarıdaki tabloda hesaplanan bu değerler göz önüne alındığında, $Gini_{cinsiyet}$ = 0.200 değerinin *Ginij* değerleri içinde en küçüğü olduğu anlaşılır. O halde bu niteliğe göre bir bölünme söz konusu olacaktır. Bölünme CİNSİYET niteliğinin KADIN ve ERKEK değerlerine göre düzenlenir. O halde CİNSİYET için KADIN değeri tablo üzerine araştırılırsa (3, 6) satırlarda yer aldığı anlaşılır. Bu durumda bölünmenin (3, 6) ve (1, 4, 5) satırları biçiminde gerçekleşmesi gerekir. Elde edilen sonuçlara göre karar ağacı aşağıda gösterildiği biçimi alır.


Kök düğüm
(Kayıtlar:1,2,3,4,5,6,7)
YAŞ=GENÇ
YAŞ={ORTA,YAŞLI}
HAYIR
(Kayıtlar:2,7)
A düğümü
(Kayıtlar: 1,3,4,5,6)
CİNSİYET=ERKEK
CİNSİYET=KADIN
EVET
[Kayıtlar:1,4,5)
B düğümü
(Kayıtlar:3,6)

Şekil 5

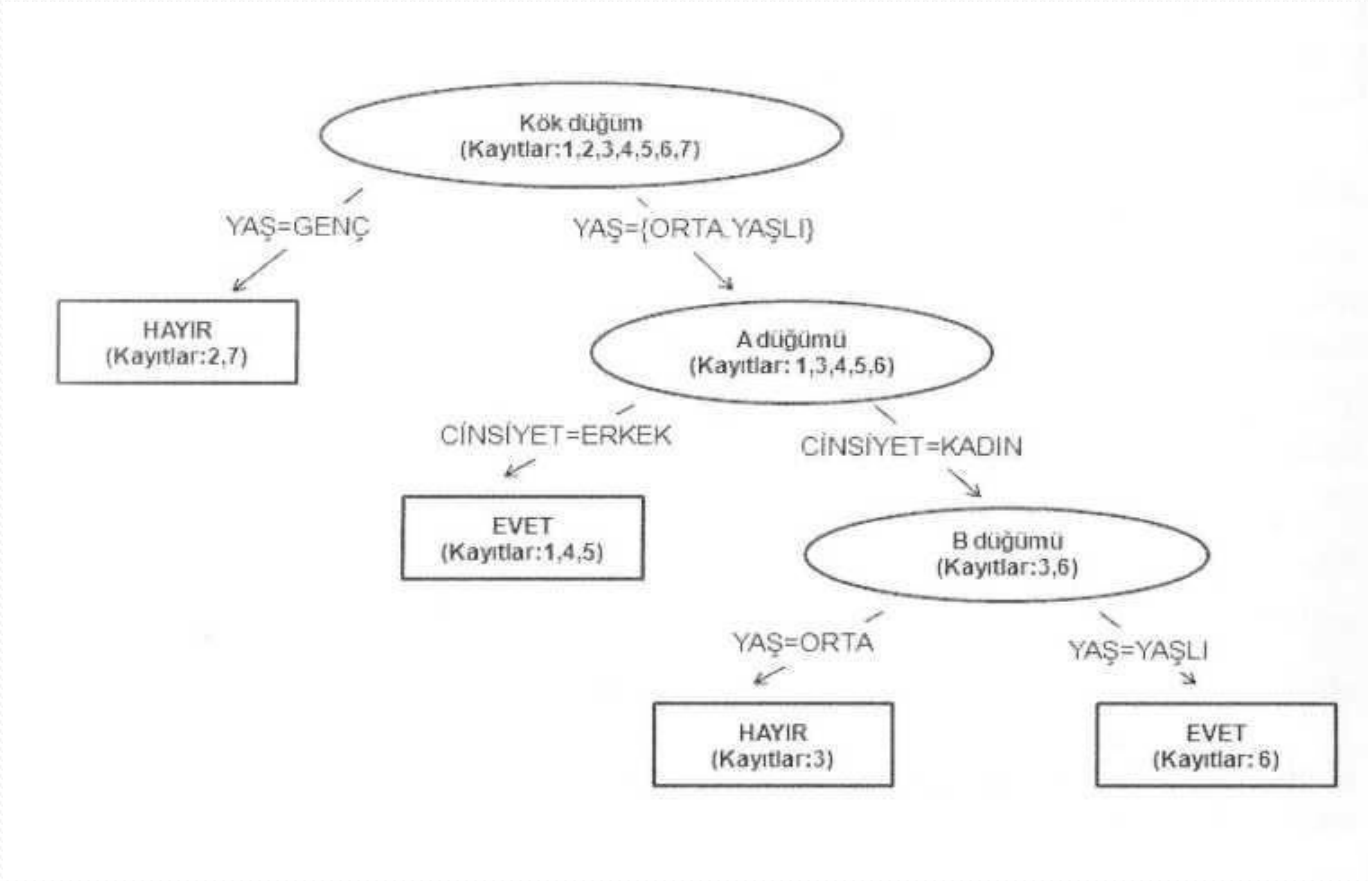
Adım 3:

Şekil üzerinde görüldüğü gibi (1,4,5) satırları EVET ile sonlanmıştır. (O halde bu satırları Tablo 22 den çıkaracak olursak, eğitim kümesinin aşağıdaki satırları elde edilir.

Başvuru	EĞİTİM	YAŞ	CİNSİYET	KABUL
3	YÜKSEK	ORTA	KADIN	HAYIR
6	YÜKSEK	YAŞLI	KADIN	EVET

Tablo 24

Elde edilen son tablo iki satırdan oluşmakta ve iki ayrı sınıfı tanımlamaktadır. O halde sonuç olarak aşağıdaki ağaç elde edilir:



Şekil 6

Karar Ağacı:

KURAL 1:
Eğer **YAŞ=GENÇ** ise **KABUL=HAYIR;**
KURAL 2:
Eğer **YAŞ=ORTA** veya **YAŞLI** ise ve
Eğer **CİNSİ YET=ERKEK** ise **KABUL=EVET;**
KURAL 3:
Eğer **YAŞ=ORTA** veya **YAŞLI** ise ve
Eğer **CİNSİ YET=KADIN** ise ve
Eğer **YAŞ=ORTA** ise **KABUL=HAYIR;**
KURAL 4:
Eğer **YAŞ=ORTA** veya **YAŞLI** ise ve
Eğer **CİNSİYET=KADIN** ise ve
Eğer **YAŞ=YAŞLI** ise **KABUL=EVET;**

BÖLÜM SONU

Veri Madenciliğine Giriş

Bölüm 2

- Kurumlarda biriken veri içerisinden kurum için yararlı olanlarını bulup ortaya çıkarma işlemine ***veri madenciliği*** adı verilmektedir.

2.1. Veriyi Bilgiye Dönüştürmenin Yolu

- Bilişim teknolojilerinin hızla gelişimi beraberinde bir sorunu da getirmiştir.

- Bilişim sistemleri sayesinde artık her bilgi sayısal ortamlara kaydedilmektedir.

- Örneğin bir mağazada satışlar ve müşterilerle ilgili her türlü bilgi sayısal ortamda yerini almaktadır.
- Üstelik günlük tüm veriler sayısal ortamda saklanmaktadır.
- Binlerce müşterisi olan bir mağaza her gün çok sayıda veri üretmek zorunda kalmaktadır.

- Bilişim teknolojisi bu devasa verileri saklamaya yeterli olabilir.
- Ancak bu veriler ne işe yarayacaktır?
- Bu verilerden firma bazı avantajlar kazanabilecek midir?
- Biriken veri gerçek anlamda «bilgiye » dönüştürülebilecek midir?

- Veriler üzerinde çözümlemeler yapmak amacıyla çeşitli istatistiksel ve matematiksel yöntemler kullanılabilir.
- Ancak veri sayısı arttıkça sorunlar ortaya çıkacaktır.
- Özellikle ilişkisel veri tabanları üzerinde bu çözümlemeleri yapmak zorlaşacaktır.

- Bu tür veriler üzerinde çözümlemeleri yapabilmek için hem yeni veri tabanı kavramlarına hem de yeni çözümleme yöntemlerine gereksinim duyulmaktadır.

- Veriyi yönetmek için «veri ambarı » ve verileri çözümleyerek «yararlı bilgiye » erişilmesini sağlayan «veri madenciliği» kavramları ortaya atılmıştır.

2.2.Veri Madenciliği

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

- Veri madenciliği bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi su yüzüne çıkarma süreci olarak değerlendirilebilir.

- Bu açıdan bakıldığında, veri madenciliği işinin kurumların karar destek sistemleri için önemli bir yere sahip olabileceğini söylenebilir.

- Veri madenciliği aslında klasik istatistiksel uygulamalara çok benzer.
- Ancak klasik istatistiksel uygulamalar yeterince düzenlenmiş ve çoğunlukla özet veriler üzerinde çalıştırılır.
- Veri madenciliğinde ise milyonlarca ve hatta milyarlarca veri ve çok daha fazla değişken ile ilgilenir.

2.3.Uygulama Alanları

- Veri madenciliğinin günümüzde yaygın bir kullanım alanı bulunmaktadır.
- Örneğin pazarlama, bankacılık ve sigortacılık gibi alanlarda ve elektronik ticaret ile ilgili alanlarda yaygın şekilde kullanılmaktadır.

Pazarlama

- Müşterilerin satın alma alışkanlıklarının belirlenmesi
- Müşterilerin demografik özellikleri arasındaki bağlantıların ortaya konulması
- Pazar sepet analizi
- Müşteri ilişkileri yönetimi
- Müşteri değerlendirme
- Satış tahmini

Bankacılık

- Farklı finansal göstergeler arasında gizli ilişkilerin ortaya konulması,
- Kredi kartı dolandırıcılıklarının ve sahtekarlıkların belirlenmesi
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi
- Kredi taleplerinin değerlendirilmesi

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespiti
- Riskli müşteri gruplarının belirlenmesi

Elektronik Ticaret

- Saldırıların çözümlenmesi
- e-CRM uygulamalarının yönetimi
- Web sayfalarına yapılan ziyaretlerin çözümlenmesi

2.4.Veri Madenciliği Süreci

- Veri madenciliğini bir süreç olarak değerlendirmek gerekir. Söz konusu süreç aşağıda belirten adımları içermektedir.
  - Veri temizleme
  - Veri bütünleştirme
  - Veri indirgeme
  - Veri dönüştürme
  - Veri madenciliği algoritmasını uygulama
  - Sonuçları sunum ve değerlendirme

2.4.1.Veri Temizleme

- Bazı uygulamalarda, üzerinde çözümleme yapılacak verilerin istenen özelliklere sahip olmadığı görülebilir.
- Örneğin eksik verilerle ve uygun olmayan verilerin oluşturduğu tutarsız verilerle karşılaşılabilir.

- Veri tabanında yer alan tutarsız ve hatalı veriler *gürültü* olarak değerlendirilmektedir.
- Bu gibi durumlarda verinin söz konusu sorunlardan temizlenmesi gerekecektir.
- Eksik verilerin yerine yenileri belirlenerek konulmalıdır.

- Bunun için aşağıdaki yöntemlerden biri kullanılabilir,
- A) Eksik değer içeren kayıtlar veri kümesinden atılabilir.
- B) Kayıp değerlerin yerine bir genel sabit kullanılabilir. Bütün kayıp değerler için aynı sabit kullanılabilir.

- Örneğin «bilinmiyor» değeri bu eksik veri yerine kullanılabilir.
- Ancak bütün değişkenlere kayıp değerler yerine aynı sabit değerin kullanımı sorun yaratacaktır.

- C) Değişkenlerin tüm verileri kullanılarak ortalaması hesaplanır ve eksik değer yerine bu değer konulabilir.
- D) Değişkenlerin tüm verileri yerine, sadece bir sınıfa ait örneklerin değişken ortalaması hesaplanarak eksik değer yerine kullanılabilir

- E) Verilere uygun bir tahmin yapılarak, örneğin regresyon ya da karar ağacı modeli kurularak eksik değer tahmin edilebilir ve eksik değer yerine kullanılabilir.

2.4.2.Veri Bütünleştirme

- Farklı veri tabanlarından ya da veri kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi yani bütünleştirilmesi söz konusu olacaktır.

2.4.3.Veri İndirgeme

- Veri madenciliği uygulamalarında bazen çözümleme işlemi uzun süre alabilir.
- Eğer çözümlemeden elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir.

- Veri indirgeme çeşitli biçimlerde yapılabilir.
- a. Veri birleştirme veya veri küpü
- b. Boyut indirgeme
- c. Veri sıkıştırma
- d. Örnekleme
- e. Genelleme

2.4.4.Veri Dönüştürme

- Veriyi bazı durumlarda veri madenciliği çözümlemelerine aynen katmak uygun olmayabilir.
- Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu taktirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rolleri önemli ölçüde azaltır.

- Ayrıca değişkenlerin sahip olduğu çok büyük ve çok küçük değerler de çözümlemelerin sağlıklı biçimde yapılmasını engeller
- Bu nedenle bir dönüşüm yöntemi uygulanarak söz konusu değişkenlerin normalleştirilmesi veya standartlaştırılması uygun bir yol olacaktır.

2.4.4.1.Min-Max Normalleştirilmesi

- Verileri 0 ile 1 arasındaki sayısal değerlere dönüştürmek için min-max normalleştirme yöntemi uygulanır.

- Bu yöntem, veri içindeki en büyük ve en küçük sayısal değerin belirlenerek diğerlerini buna uygun biçimde dönüştürme esasına dayanmaktadır.

- Söz konusu dönüştürme bağıntısı şu şekilde ifade edilmektedir:

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

- Burada X^* dönüştürülmüş değerleri,
- X gözlem değerlerini,
- X_{min} en küçük gözlem değerini
- X_{max} en büyük gözlem değerini ifade etmektedir.

Örnek:

- Aşağıdaki tabloda yer alan X değişkeni değerlerine min-max normalleştirme bağıntısını uygulayarak dönüştürmek istiyoruz.
- Bunun için, veriler için önce aşağıdaki değerler belirlenir:
- X_{min}=30
- X_{max}=62

- Bu değerlere dayanarak X örneğini birinci elemanı için şu şekilde bir hesaplama yapılır:

- $X^* = \frac{X - X_{min}}{X_{max} - X_{min}} = \frac{30-30}{62-30} = 0$

- Tabloda Min-Max normalleştirme dönüşümü sonucu elde edilen değerler

x	X^*
30	0,0000
36	0,1875
45	0,4688
50	0,6250
62	1,0000

2.4.4.2.Z-score Standartlaştırması

- İstatistik çözümlemelerde sıkça kullanılan bir diğer dönüşüm biçimi **z-score** adıyla anılmaktadır.
- Bu yöntem, verilerin ortalaması ve standart hatası göz önüne alınarak yeni değerlere dönüştürülmesi esasına dayanmaktadır.

- Söz konusu dönüşümlerde şu şekilde bir bağlantıya yer verilir:
- $X^* = \frac{X - \bar{\bar{X}}}{\sigma_x}$
- Burada X^* dönüştürülmüş değerleri,
- X gözlem değerlerini,
- $\bar{X}$ verilerin aritmetik ortalamasını,
- σx gözlem değerlerinin standart sapmasını ifade etmektedir.

Örnek:

- Önceki örnekte ele alınan verileri bu kez Z- score standartlaştırılmasını uygulayarak dönüştüreceğiz.
- Yapılacak ilk işlem $\bar{X}$ aritmetik ortalamanın bulunmasıdır.
- $\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i=44.6$

- Z - score standartlaştırma işlemi için X serisinin standart hatasının bulunması gerekmektedir.
- Söz konusu hata şu şekilde hesaplanır:
- $\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})\ \text{^}2}{n-1}} = 12.44$

- Bu durumda birinci satır için Z- score dönüşümü şu şekilde olabilir:

- $X^* = \frac{X - \bar{\bar{X}}}{\sigma_x} = \frac{30 - 44.6}{12.44} = -1.1735$

- Benzer biçimde diğer gözlemler içinde aynı hesaplamalar yapılır.

2.4.5Veri Madenciliği Algoritmasını Uygulama

- Veri madenciliği yöntemlerini uygulaya-bilmek için önceden bahsedilen işlemlerin uygun görünenleri yapılır.
- Veri hazır hale getirildikten sonra konuyla ilgili veri madenciliği algoritmaları uygulanır.

2.4.6. Sonuçları Sunum ve Değerlendirme

- Veri madenciliği algoritması veriler üzerinde uygulandıktan sonra, sonuçlar düzenlenerek ilgili yerlere sunulur.
- Sonuçlar çoğu kez grafiklerle desteklenir.
- Örneğin bir hiyerarşik kümeleme modeli uygulanmış ise sonuçlar *dendrogram* adı verilen özel grafiklerle sunulur.

2.5. Veri Madenciliği Yöntemleri

- Veri madenciliği konusunda çok sayıda yöntem ve algoritma geliştirilmiştir. Bu yöntemlerin bir çoğu istatistiksel tabanlıdır.
- Veri madenciliği modellerini temel olarak şu şekilde gruplandırabiliriz.
  - A) Sınıflandırma
  - B) Kümeleme
  - C) Birliktelik Kuralları

2.5.1.Sınıflandırma

- Sınıflama veri madenciliğinde sıkça kullanılan bir yöntem olup veri tabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılır.
- Verilerin sınıflandırılması için belirli bir süreç izlenir.

- Öncelikle var olan veri tabanının bir kısmı eğitim amacıyla kullanarak sınıflandırılma kurallarının oluşturulması sağlanır.
- Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir.

Örnek:

- Bir bankanın kredi verdiği müşterilerinin risk durumunu karar ağaçları yardımıyla ortaya koymak istediğini varsayalım.
- Bu sayede belirli özelliklere sahip müşterilerden kredi talebi geldiğinde karar ağacı bilgilerine dayanarak kredi verip vermeme konusunda karar verilecektir.

MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	Yüksek	Yüksek	İşveren	Kötü
2	Yüksek	Yüksek	Ücretli	Kötü
3	Yüksek	Düşük	Ücretli	Kötü
4	Düşük	Düşük	Ücretli	İyi
5	Düşük	Düşük	İşveren	Kötü
6	Düşük	Yüksek	İşveren	İyi
7	Düşük	Yüksek	Ücretli	İyi

Tablodaki veriler karar ağacının oluşturulması amacıyla eğitim verisi olarak kullanılacaktır. Söz konusu verileri kullanarak karar ağaçlarını oluşturmak üzere veri madenciliğinin çok sayıda yöntemi bulunmaktadır.


A Düğümü
Borç=Yüksek
Borç=Düşük
Kötü
B Düğümü
Gelir=Yüksek
Gelir=Düşük
İyi
C Düğümü
Statü=İşveren
Statü=Ücretli
Kötü
İyi

- Elde edilen karar ağacı karar kuralları oluşturulmasında kullanılabilir.
- Bir önceki slayttaki karar ağacı yorumlanarak şu şekilde karar kuralları oluşturulabilir.

- Kural 1:
- Eğer BORÇ=Yüksek ise RİSK=Kötü
- Kural 2:
- Eğer BORÇ=Düşük ise ve
- Eğer GELİR=Yüksek ise RİSK=İyi
- Kural 3:
- Eğer BORÇ=Düşük ise ve
- Eğer GELİR=Düşük ise ve
- Eğer STATÜ=İşveren ise RİSK=Kötü

- Kural 4:
- Eğer BORÇ=Düşük ise ve
- Eğer GELİR=Düşük ise ve
- Eğer STATÜ=Ücretli ise RİSK=İyi

Eğitim kümesinden elde edilen bu kurallar kullanılarak yeni bir müşterinin risk durumu hakkında karar verilebilir.

2.5.2.Kümeleme

- Kümeleme verilerin kendi aralarındaki benzerliklerin göz önüne alınarak gruplandırılması işlemidir. Bu özelliği nedeniyle pek çok alanda uygulanabilmektedir. Örneğin, pazarlama araştırmalarında yaygın biçimde kullanılmaktadır.
- Bunun dışında desen tanımlama, resim işleme, uzaysal harita verilerinin analizinde kullanılmaktadır.

Örnek

- Aşağıdaki gözlem değerlerini göz önüne alalım.
- Bu gözlem değerinin X1 ve X2 gibi iki değişkeni bulunmaktadır.
- Bu gözlem değerlerine dayanarak verilerdeki kümelenmeleri belirlemek istiyoruz.

Gözlem	X1	X2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

Kümeleri ortaya koymak üzere bir çok veri madenciliği ve istatistiksel yöntem bulunmaktadır.

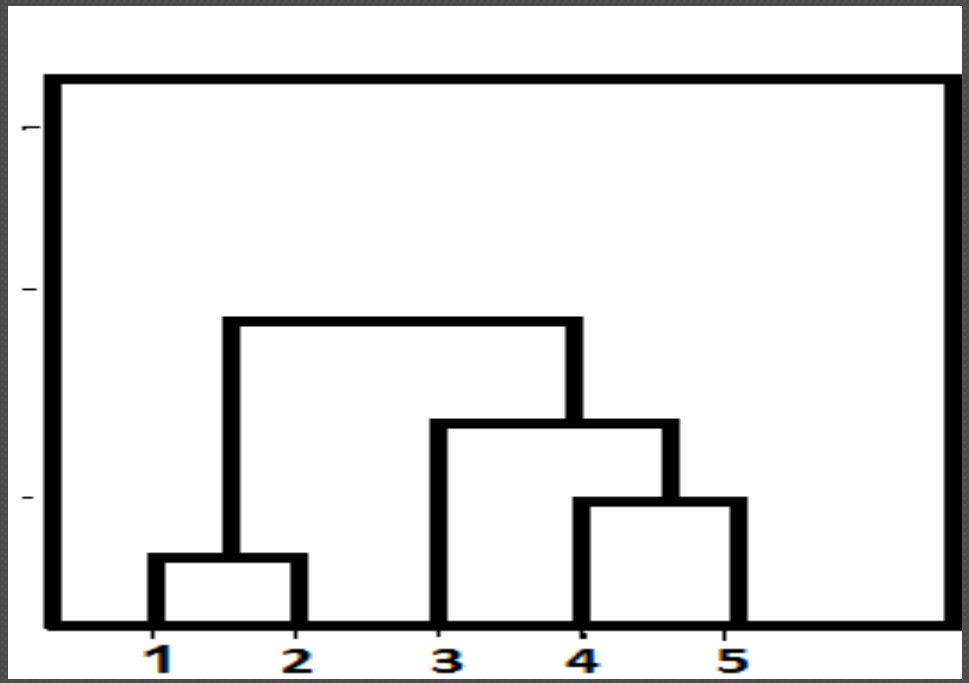
Söz konusu verilere hiyerarşik kümeleme yöntemlerinden en yakın komşu algoritmasını uygulandığında bir sonraki tabloda belirtilen kümeler elde edilir.

	Kümeler
Küme 1	(1,2)
Küme 2	(4,5)
Küme 3	(3,4,5)
Küme 4	(1,2,3,4,5)

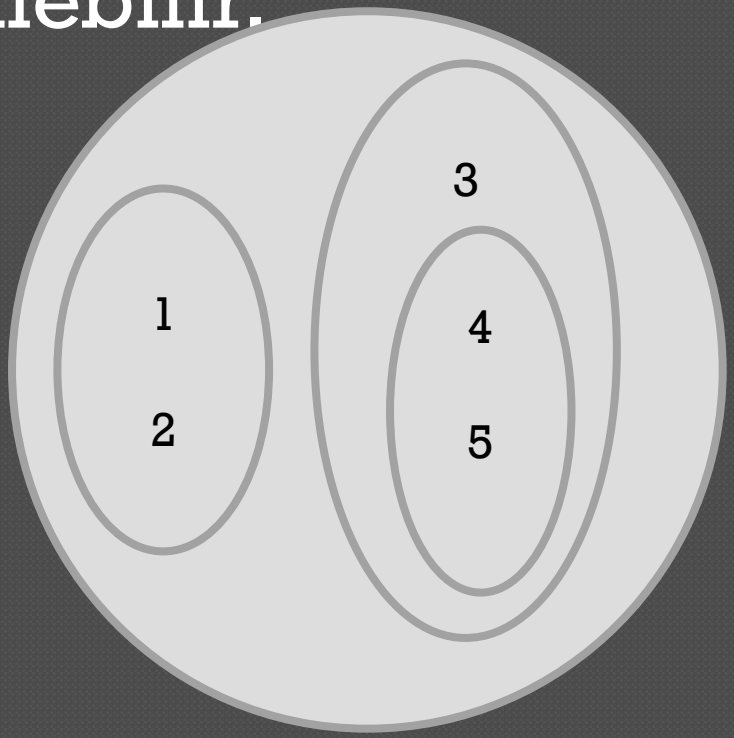
Söz konusu kümelere uygun olarak kümeleme grafiği çizilebilir.

Kümeleri gösteren grafiğe *dendrogram* adı verilmektedir.

- Dendrogramın görünüşü



- Söz konusu kümelere daha açık biçimde aşağıda belirtildiği biçimde de gösterilebilir.



2.5.3. Birliktelik Kuralları

- Veri tabanı içinde yer alan kayıtların birbirleriyle olan ilişkilerini inceleyerek hangi olayların eş zamanlı olarak birlikte gerçekleşebileceklerini ortaya koymaya çalışan veri madenciliği yöntemleri bulunmaktadır.

- Bu ilişkilerin belirlenmesi ile birliktelik kuralları (association rules) elde edilir.
- Birliktelik kuralları özellikle pazarlama alanında uygulama alanı bulmuştur.
- Pazar sepet analizleri adı verilen uygulamalar bu tür veri madenciliği yöntemlerine dayanmaktadır.

- Bu tür çözümlemelerden hareketle müşterilerin alışveriş alışkanlıkları belirlenmeye çalışılır.
- Pazar sepet analizleri yardımıyla bir müşteri herhangi bir ürünü aldığında sepetine başka hangi ürünleri de koyduğu belirli bir olasılığa göre konulur.

- Birlikte satın alınan ürünler belirlendiğinde mağazalarda raflar ona göre düzenlenerek müşterilerin bu tür ürünlere daha kolayca erişmeleri sağlanabilir.

Örnek:

- Bir mağazada alışveriş yapan müşterilerin alışveriş alışkanlıklarını belirlemek istediğimizi varsayalım.
- 5 müşterinin alışveriş sepetlerine hangi ürünleri koyduğunu bir sonraki slaytta görüyoruz.

Müşteri	Alışveriş sepetindeki ürünler
1	Makarna , yağ , meyve suyu , peynir
2	Makarna ketçap
3	Ketçap , yağ , meyve suyu, bira
4	Makarna, ketçap, yağ, meyve suyu
5	Makarna, ketçap, yağ, bira

- Bu verilerden yararlanarak birliktelik çözümlemeleri yapılır. *Apriori* algoritması yardımıyla aşağıdaki sonuçlar elde edilir.

- {Ketçap, Meyve suyu}→ {Yağ} (s=0.4, c=1.0)
- {Ketçap, Yağ}→ {Meyve suyu} (s=0.4, c=0.67)
- {Yağ, Meyve suyu}→ {Ketçap} (s=0.4, c=0.67)
- {Meyve suyu}→ {Ketçap, Yağ} (s=0.4, c=0.67)
- {Yağ}→ {Ketçap, Meyve suyu} (s=0.4, c=0.5)
- {Ketçap}→ {Yağ, Meyve suyu} (s=0.4, c=0.5)

Bölüm 1

VERİ AMBARI

Veri Ambarı

- Veri ambarı zaman içerisinde olabildiğince birikmiş verilerin oluşturduğu bir veri yığınıdır.
- Bir işletmenin sahip olduğu verilerin karar destek amacıyla kullanılmasına olanak sağlar.
- Bir zaman boyutu içinde analitik işlemlerin yapılmasını sağlamak için gerekli bilgi temelini sağlar.

1.Temel Kavramlar

- **1.1.Klasik Dosya Yapıları**
- **1.2.Kayıt ve Alan**
- Veri saklama birimlerinde depolanan veri topluluklarına «dosya » adını veriyoruz.
- Dosyalar ise kendi içerisinde kayıtlara bölünmüştür.

- Örneğin bir sınıftaki öğrenci listesini göz önüne alalım; bu liste çok sayıda veri içerebilir.
- Söz konusu listenin ana bellekte tutulması söz konusu olamaz.
- Ana bellekte saklandığı takdirde, bilgisayarın kapatılası durumunda bu bilgiler yok olacaktır.
- O halde bu verinin kalıcı bir ortamda , örneğin sabit disk üzerinde saklanması gerekecektir.

- Öğrenci listesindeki her bir öğrenci bilgisi bir mantıksal kayıt oluşturur.
- Her kayıt farklı bilgiler içerebilir.
- Örneğin öğrenci kaydı öğrencinin adı, baba adı, doğum yeri vb. gibi bilgileri içerebilir.
- Bu bilgilerin her birine alan (field) adı veriyoruz.

Adı	Baba Adı	Doğum Yeri	....

1.3.Sıralı Dosyalar

- Klasik bilgisayar dosyaları uygulamalarda kullanılmak üzere hazırlanır. Bu dosyalar, sıralı ya da doğrudan erişim yöntemi kullanılarak işlenir.
- Sıralı erişimde dosyanın tüm kayıtları ardı ardına taranarak istenilen kayıtlara erişilir.
- Doğrudan erişim yönteminde ise kayıtlar tek tek sırayla okutulmaz, istenilen kayıta doğrudan erişerek işlenir.

1.4.Dizinli Dosyalar

- Bu tür dosyalarda , her bir arama işlemi dosyanın başında itibaren yapılmaz .
- Belirlenen kayıtlara doğrudan erişilerek üzerinde işlem yapılır.
- Doğrudan erişimli dosyaların en tanınmışı dizinli dosyalar olarak bilinir.

- Bir dosya için oluşturulan dizin, söz konusu dosyanın anahtarları ile bu anahtarların disk üzerinde bulunduğu adresi içerir.
- Anahtar alan, erişimde kullanılmak üzere seçilen alan olarak değerlendirilir.

1.5.Hesaba Dayalı Dosyalar

- Bir diğer doğrudan erişimli dosya türü hesaba dayalı dosyalar (hashes files) olarak bilinir.
- Bu tür dosyalar dizinli dosyalar gibi ayrı bir dizinin tutulmasını gerektirmez.
- Dosyanın herhangi bir kaydına doğrudan doğruya erişebilmek için bir hesaplama (çırpı) algoritması kullanılır.

1.6. Veritabanı Sistemleri

- Karmaşık dosya yapıları, çok sayıda dosya arası ilişki ve kullanıcıların dosyalara erişimi söz konusu olduğunda geleneksel dosya sisteminin yetersiz kaldığı görülmüştür.
- Bu sorunu çözmek üzere veriyi saklama ve erişim konusunda yeni yazılım teknolojilerine yönelme başlamış ve veritabanı sistemlerini oluşturmak ve veriyi yönetmek üzere Veri Tabanı Yönetim Sistemleri(VTYS) yaklaşımı ortaya çıkmıştır.

1.7. Veritabanı Sistemlerinin Üstünlükleri

- a) Verinin tekrarlanmasını önler.
- b) Verinin tutarlı olmasını sağlar.
- c) Aynı andaki erişimlerde tutarsızlıkların ortaya çıkmasını önler.
- d) Verinin güvenliğini sağlar

Veri Modelleri

- Sıradüzensel Veri Modeli
- Ağ Veri Modeli
- İlişkisel Veri Modeli
- Nesneye Yönelik Veri Modeli

2.Veri Ambarı

2.1.OLTP Sistemler

Bir kurumun günlük verilerinin işlendiği ortamlara OLTP (OnLine Transaction Processing) sistemler adı verilmektedir.

Örneğin bir işletmenin sahip olduğu stok sistemi ile depoya giren ve çıkan ürünleri ve ödemeleri izlenebilir. Bu tür işlemler her gün yapılır.

Stoklarla ilgili tüm işlemler veritabanına kaydedilir. Bu kayıtlara dayalı çeşitli raporlar üretilir.

- OLTP sistemlerine ilişkin veritabanlarına veri kaydedilebilir, veriye erişilerek raporlanabilir ve istendiğinde veri silinebilir.

2.2. Karar Destek Sistemleri

- 1990'lı yıllara kadar bilgisayarların karar alma süreci üzerindeki etkisini artırmak üzere çok çaba harcanmıştır .
- Karar Destek Sistemleri (Desicion Support Systems) ve Üst Yönetici Sistemleri (Executive Information Systems) bu amaçla ortaya atılmıştır.

- Karar destek sistemleri, yöneticilerin programlanamayan türden karar verme işlemlerine yardımcı olmak üzere geliştirilir.
- Yöneticinin herhangi bir anda, daha önceden öngörülmemiş bir bilgiye aniden gereksinimi olabilir.
- Böyle bir durumda, hemen yanıt verebilecek bir sistemin varlığı gerekecektir.
- Karar destek sistemleri bu gibi durumlar için tasarlanır.

- Üst düzey yönetici sistemleri temel olarak karar destek sistemlerine benzer.
- Ancak bu tür sistemler sadece stratejik düzeydeki yönetici personel için tasarlanır.
- Bu sistemler yapısal olmayan, yani önceden programlanamayan karar türlerine destek veren sistemlerdir.

2.3. Veri Ambarı Nedir?

- Veri ambarı için çeşitli tanımlamalar yapılabilir.
- Veri ambarını en basit anlamda, karar destek uygulamaları için tasarlanan bir ortam olarak tanımlıyoruz.
- Bir başka deyişle veri ambarı, karar destek sistemlerinin teknik altyapısını oluşturur.

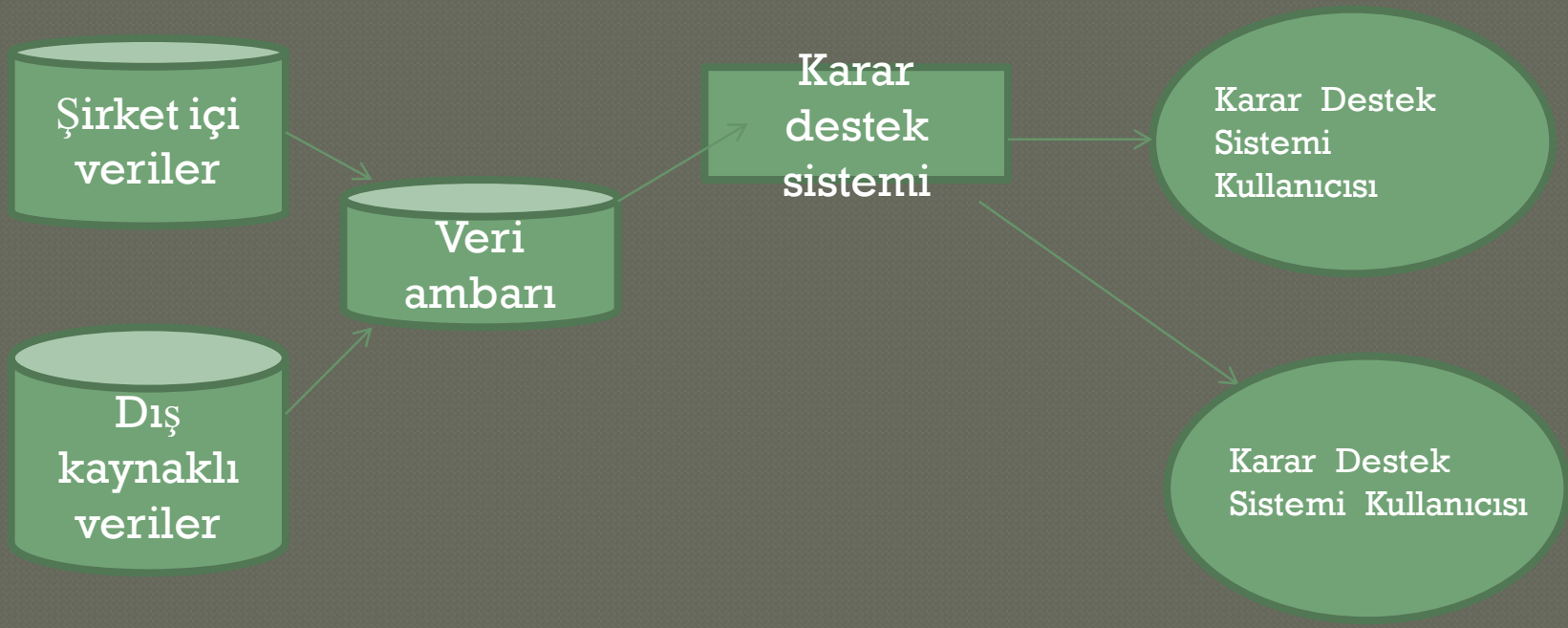
- Veri ambarı ,bir işletmenin sahip olduğu verinin, eskileri de dahil olmak üzere, karar destek amacıyla kullanılmasına olanak sağlar.
- Bunun anlamı, var olan ancak kullanılamayan veri de artık kullanılabilir ve çözümlenebilir hale gelecektir.

- Veri ambarı, birbiriyle bütünleşik olmayan uygulamaların bütünleştirilmesi açısından bir olanak sağlar.
- Veri ambarı bir zaman boyutu içinde analitik işlemlerin yapılması için ihtiyaç duyulan bilgi temelini sağlar.

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

- a) konuya yönelik
- b) bütünleşik
- c) zaman boyutu olan
- d) sadece okunabilen

veri topluluğudur.


Şirket içi veriler
Dış kaynaklı veriler
Veri ambarı
Karar destek sistemi
Karar Destek Sistemi Kullanıcısı
Karar Destek Sistemi Kullanıcısı

2.3.1. Konuya Yöneliktir

- Konuya yönelik olmasının anlamı, veri ambarının işletmedeki yüksek seviyeli varlıklar üzerinde odaklanmış olmasıdır.
- Bu varlıklar bir okul ortamı için öğrenciler, dersler, notlar vb. olabilir.

2.3.2. Bütünleşiktir

- Veri ambarı ortamdaki verinin en belirgin görünümü, bütünleşik durumda olasıdır.
- Veri ambarı içindeki veri mutlaka bütünleşik olmalıdır. İstisnası yoktur.
- Bütünleşme farklı biçimlerde olabilir. Verinin kodlanmasında görüş birliğine varılması, ölçü birimlerinin seçiminde tutarlılık, sayısal değerlerin fiziksel gösterimindeki tutarlıklık vb gibi bütünleştirme kavramlarından söz edilebilir.

2.3.3. Zaman Boyutu Vardır

- Veri ambarındaki tüm veriler zamanın belirli bir anına aittir. Veri ambarındaki verinin bu temel karakteristiği, işlemsel sistemlerdeki veriden oldukça farklıdır.
- İşlemsel sistemlerde veri o anda var olan değerdir.
- İşlemsel sistemlerde bir veriye ulaşıldığında çoğunlukla onun şu andaki değeriyle ilgilenir.
- İşlemsel veri de zamana dağılmış olabilir.
- Ancak bu süre örneğin 80 gün olabilir.

- Veri ambarındaki veri ise, sadece o andaki değerlerle değil, geçmişteki değerlerle de ilgilidir.
- Veri zaman içinde bir noktayla birleştirilerek değerlendirilir.
- Örneğin sömestre, mali yıl, ödeme dönemi gibi unsurlar göz önüne alınır.
- Anahtar alan, zaman değerini de kapsayacaktır.

2.3.4.Sadece okunabilir

- Veri ambarı veri yönetiminin gereksinimine yanıt vermek üzere tasarlandığı için günlük işlemlere tabi tutulmaz, yani silinemez veya güncelleştirilemez.
- Veri ambarında iki tür işlemden söz etmek mümkündür:

a)Verinin yüklenmesi
b)Veriye erişilmesi

2.3. Veri Ambarının Temel Özellikleri

- A)İşlemsel çevrede yer alan veri bir süzme işlemi sonucunda veri ambarı çevresine aktarılır.
- İşlemsel ortamdaki verinin tümüyle veri ambarına aktarılması beklenmez.
- Sadece Karar Destek Sistemlerinin gereksinim duyacağı veri aktarılır.

- B) Zaman yelpazesi her iki sistemde farklılık gösterir.
- İşlemsel ortamdaki veri çok tazedir.
- Veri ambarındaki veri ise daha eskidir.
- Ancak zaman açısından bakıldığında , işlemsel ve veri ambarı ortamındaki veri arasında kısa bir örtüşme vardır.

- C)Veri ambarı özet bilgileri içerebilir.
- İşlemsel sistemlerde ise özet veriye yer verilmez.

- D)Veri ambarının en önemli özelliklerinden biri olan bütünleştirmeyi sağlamak üzere, verinin önemli bir kısmı belirli bir dönüşümden sonra veri ambarına aktarılır.
- Böylece işlemsel veri ile veri ambarındaki verinin içeriği açısından farklılıklar oluşabilecektir.

2.5. Veri Ambarını 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

2.5.1.Metadata

- Veri ambarının en önemli bileşenlerinden biri *metadata* dır.
- Metadata doğrudan işlemsel çevreden gelen veriyi içermez.
- Metadata veri ambarı içindeki veriden ayrı bir yerdedir.

Metadata şu özelliklere sahiptir;

- a) Karar Destek Sistemleri analistlerine yardım etmek üzere yaratılan bir dizindir ve veri ambarı içeriğinin neler olduğunu belirtir. Kullanılan verinin yapısını ortaya koyan bilgileri içerir.

- b)İşlemsel çevreden veri ambarına dönüştürülen verinin konumları hakkında bilgileri içeren bir klavuzdur.
- c) İşlemsel çevreden alınana verinin hangi algoritmaya göre düşük ya da yüksek seviyede özetlendiği hakkında bilgileri içeren bir klavuzdur.

2.5.2. Ayrıntı Veri

- Veri ambarı şu andaki ayrıntı veriyle ilgilendiğini varsayalım. Bu veri en son olayları içermektedir ve henüz işlenmediği için diğerlerine oranla daha büyük hacimlidir.
- Bu tür veri disk üzerinde saklandığından bunlara erişim ve yönetimi pahalıdır.

- Ayrıntı veri denilince, şu andaki en son ayrıntı veri kastedilmektedir.
- Bu ayrıntılar belirli bir dönemi kapsayabilir.
- Örneğin, satışlara ilişkin veri son beş yılın ayrıntılarını içerebilir.

2.5.2 Eski Ayrıntı Veri

- Ayrıntı verinin dışında kalanlar, yani daha eski tarihe ait olanlar eski ayrıntı bilgileri oluşturacaktır.
- Bu veri şu andaki ayrıntı veriye göre daha düşük bir ayrıntı düzeyine indirilerek saklanır.
- Bu tür veriye çok sık biçimde erişilemediği için, disk üzerinde saklanabileceği gibi , genellikle başka saklama birimlerine yerleştirilebilir.

2.5.4. Düşük Düzeyde Özetlenmiş Veri

- Şu andaki ayrıntı veriden süzülerek elde edilen düşük seviyede özetlenmiş veri de veri ambarının bir parçası olabilir.
- Bu tür veri disk üzerinde saklanır.
- Veri ambarının tasarımı esnasında, hangi verinin özetleneceği ve özetleme işleminin ne düzeyde olacağı belirlenir.

2.5.5 Yüksek Düzeyde Özetlenmiş Veri

- Şu andaki ayrıntı veri daha yüksek düzeyde özetlenerek, kolayca erişilebilir hale getirilebilir.

- Bu tür veri de veri ambarının bir bileşeni olarak yer alabilir.

2.6. Veri Ambarı Veri Modeli

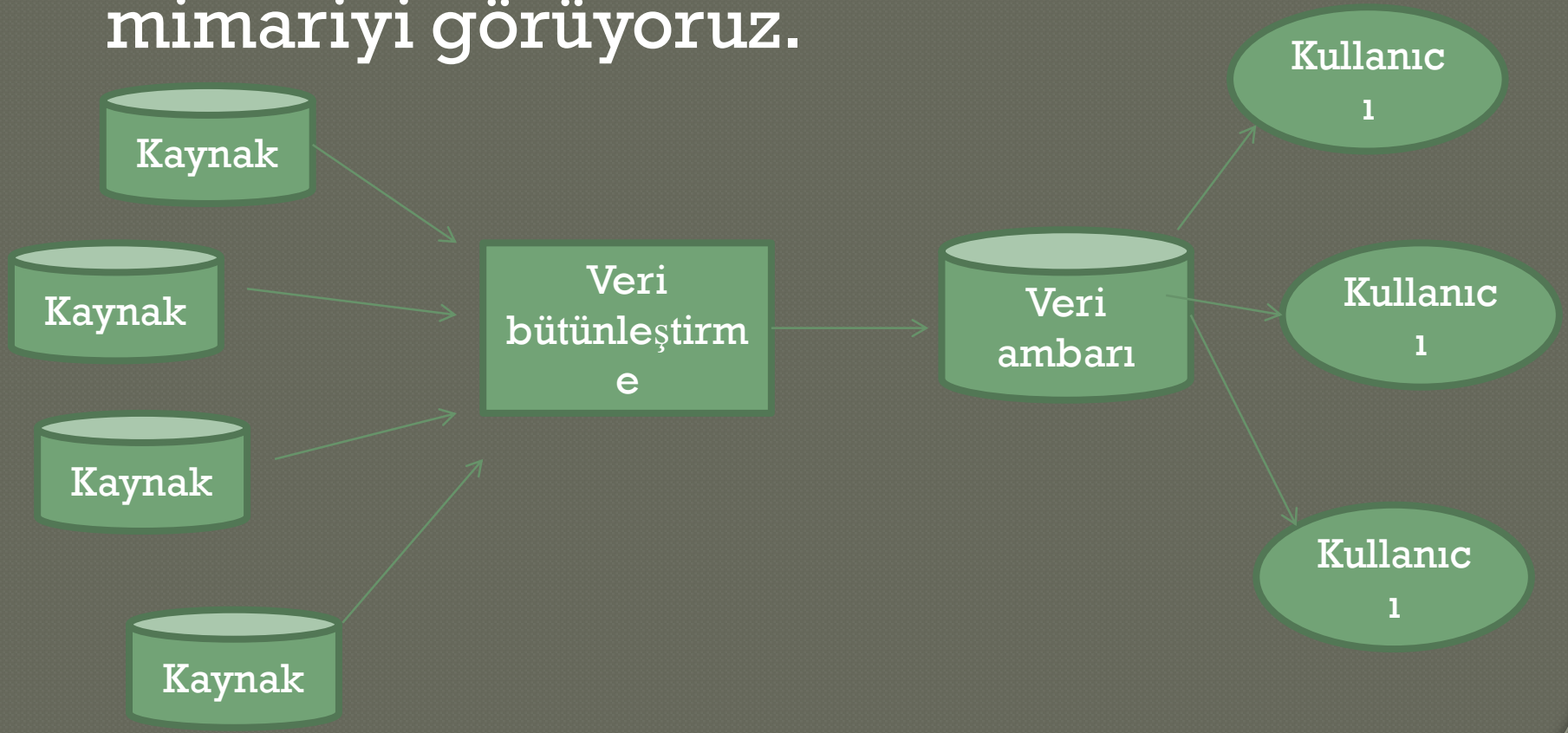
- Veri ambarı, veri modeli açısından OLTP sistemlerinden farklılık gösterir.
- Şöyle ki, veri ambarı gündelik işlemleri yürütmek için değil, toplanmış olan veriyi hızlı biçimde çözümlemek amacını taşır.
- Böyle olunca doğal olarak veri modelinde önemli farklılıklar ortaya çıkacaktır.
- Kabaca bir veri ambarının bir veya birkaç ana tablodan oluştuğunu söyleyebiliriz.
- Bu ana tabloya bağlı olarak birçok boyut tablosu veri modeli içinde yer alır.

- Veri ambarının veri modeli, işletmenin gereksinimlerine dayalı olarak bir «boyutsal model» olarak düşünülür.
- Bir veri ambarında birden fazla boyut yer alabilir.
- Bu nedenle söz konusu modele «çok boyutlu model» adı da verilir.
- Bu model «veri küpü» yada «yıldız şema » olarak da adlandırılabilir.
- Veri ambarını bu tür bir veri modeli seçmeye zorlayan birçok neden bulunmaktadır.

2.7. Veri Ambarı Mimarisi

- Veri ambarı mimarisinin genel karakteristikleri şu şekilde sıralanabilir:
- a)Kaynaklardan alınan veri dönüştürülür.
- b)Veri ambarı oluşturulur.
- c)Kullanıcıların veri ambarına erişimi sağlanır.

- Şekilde veri ambarını sahip olduğu mimariyi görüyoruz.



2.7.1. Verinin Dönüştürülmesi

- Üretildiği kaynaklardan veri teminin edilmesi gerekmektedir.
- Veri bu kaynaklarda farklı biçimde yer alabilir.
- Örneğin farklı veritabanlarında saklanıyor olabilir.

- Ayrıca aynı veri farklı biçimlerde temsil ediliyor olabilir.
- Ayrıca veri üzerinde bozukluklar söz konusu olabilir.
- Bu gibi veriyi veritabanında doğrudan kaydetmek sorunlara neden olabilir.
- Söz konusu verinin öncelikle düzenlemesi gerekir.

2.7.2 Veri Ambarının oluşturulması

- Veri dönüştürüldükten sonra veri ambarına aktarılır.
- Veri ambarını istenen amaçlara uygun biçimde çalışabilmesi için özel bir tasarıma gereksinim duyulur.

- Bunun yanı sıra veri ambarı veritabanı OLTP veritabanlarından fiziksel olarak ayrı konumda yaratılır.
- Çünkü OLTP sistemlerinden ve veri ambarlarından beklenen yarar farklıdır.
- Veri ambarları sadece sorgulara yöneliktir ve sadece karar destek amacıyla kullanılabilecek bilgileri içerir.

- Veri ambarında model olarak çok boyutlu model tercih edilir.
- Bu veri modeli adından da anlaşılacağı gibi veriye bazı boyutlar kazandırarak sorgulamaları bu boyutlar yardımıyla oluşturmak amacıyla düzenlenir.

- Bir işletmenin satış verilerini içeren bir veri ambarında yıl, ürün grubu ve mağazalar birer boyut olarak düşünülebilir.
- Yönetici satış rakamlarını bu boyutlara göre düzenleyebilir.
- Örneğin Bakırköy mağazasının yıllar bazında ve ürün grubuna göre satışlarını sorgulamak gerektiğinde söz konusu çok boyutlu model yardımcı olacaktır.

2.7.3. Kullanıcıların Veri Ambarına Erişimi

- Veri ambarı oluşturulduktan sonra kullanıcılar farklı biçimde erişerek kullanabilirler.
- Kullanıcı doğrudan veri ambarına erişerek sorgulama yapabilir.

- Bazı ilişkisel veri tabanları çok boyutlu analizlere olanak tanıyan analiz araçlarına sahiptir.
- İlişkisel veri tabanlarının dili olan SQL içinde bu tür analizleri yapmaya yarayan komutlar yer almaktadır.

- Bazı uygulamalarda veri ambarı doğrudan doğruya kullanılmaz.
- Veriler bir OLAP ortamına taşınarak onun üzerinde gerekli çözümlemeler yapılır.
- Bu tür araçlar çok boyutlu çözümlemelere olanak sağlar.

Sistemlerin Genel Karşılaştırılması

- Veri ambarı denilince doğal olarak «karar destek sistemleri», «karar destek sistemleri uygulamaları» ya da «karar destek çözümleri» adı verilen uygulamaların alt yapısı olarak anlaşılmaktadır.

- a) Her şeyden önce OLTP sistemlerin, kuruluşların gündelik işlerini bilgisayar ortamında yürütmek üzere yaratıldığını unutmamak gerekiyor.
- Karar destek sistemleri ise, verinin analizi amacıyla yaratılır.

- Böylece bu ikisi arasındaki en önemli fark ortaya çıkmış oluyor.
- OLTP uygulamaları «kuruluşu çalıştıran» , veri ambarı uygulamaları ise «kuruluşa yol gösteren » bir oluşumdur.

- b)OLTP sistemler arasında muhasebe, personel,stok vb. uygulamaları sayılabilir.
- Söz konusu sistemler kuruluşun normal işlemlerini yerine getirmek amacıyla kullanılır.
- Her sistem kendi amacına yönelik tablolara kayıt ekleme, silme ya da güncelleştirme işlemini uygular.
- Uygulamalar çok sayıda birbirleriyle ilişkili tablodan oluşur.
- Sorgu yada raporlar bu ilişkili tablolar taranarak oluşturulur.

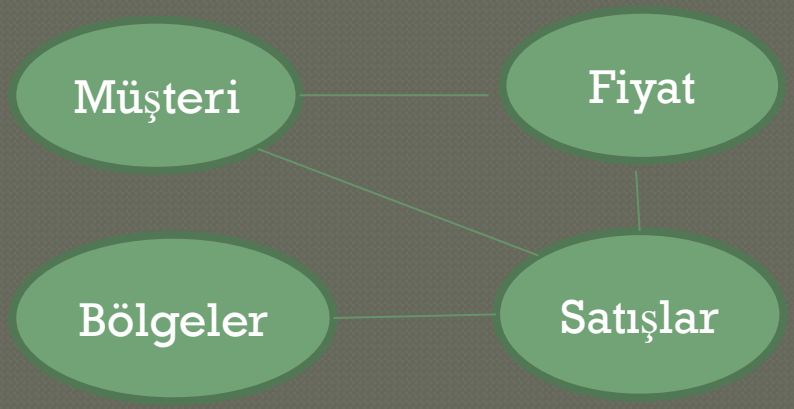
- OLTP sistemleri gündelik işlemlerin hızlı ve güvenilir biçimde yürütülmesi amacıyla optimize edilmiştir.
- Veri ambarı ise veri yapıları açısından OLTP sistemlerden farklıdır.

- Her şeyden önce veri ambarının amacı faklıdır.
- Şöyle ki, veri ambarı gündelik işlemleri yürütmek için değil; toplanmış olan veriyi hızlı biçimde çözümlemek amacıyla oluşturulur.

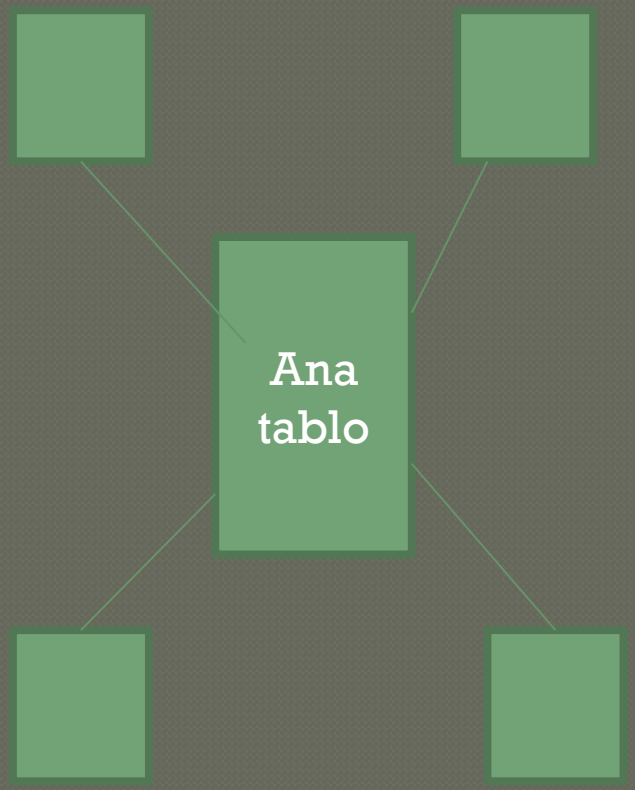
- Ana hatlarıyla veri ambarına bakacak olursak, bir veya birkaç ana tablodan oluşması gerektiğini söyleyebiliriz.
- Örneğin, bir ana tabloya bağlı olarak zaman, müşteriler ve bölgeler birer boyut tablosu olabilir.

- c) Veri ambarı veri modeli, işletmenin gereksinimlerine bağlı olarak bir «boyutsal model» olarak düşünülür.
- Bu model «veri küpü » ya da «yıldız şema» olarak da adlandırılabilir.

OLTP ağ şeması



Veri ambarı yıldız şeması



- d) OLTP uygulamaları girdi/çıktı yoğunluklu işlemleri yapmak üzere odaklandığından, karar destek sistemlerinin en çok gereksinim duyduğu sorgulama ve çözümleme işlemlerinde yetersiz kalacaktır.

- Karar destek uygulamaları genellikle karmaşık sorguları gerektirir.
- Üstelik bu tür sorgular, çoğunlukla çok sayıda tabloya erişimi gerektirecektir.
- OLTP veritabanları bu tür gereksinimi karşılayamaz.
- OLTP uygulamaları çoğunlukla aynı sistem içindeki az sayıda tablo ile çalışır.

- e)Veri ambarı üzerinde yapılabilecek sorgulamaların özelliği OLTP sistemlerinkinden farklıdır.

- OLTP sistemlerdeki sorgular çoğunlukla önceden belirlenmiş standart sorgu ya da raporlar biçimindedir.

- Veri ambarında ise, sorgular çoğunlukla önceden planlanmış ve aniden ortaya çıkabilecek türdendir.

- f) Veri ambarının OLTP sistemlerinden önemli farklarından birisi, tarihsel veriyi içermesidir.
- Eski verinin bir veri ambarı içinde yer alması karar destek sistemlerinin uygulanması açısından büyük önem taşımaktadır.

- Bunun anlamı, yönetici sadece günümüzün verisine bakarak değil, geçmişin verisine de bakarak eğilim çözümlemeleri yapabilecektir.
- En önemlisi geleceğe yönelik öngörü çözümlemelerinde bu hazır veri kullanılabilecektir.

- g) Sonuç olarak ,OLTP sistemlerin veri ambarı ya da bir başka değişle karar destek uygulamaları için veri kaynağı oluşturan sistemler olduğunu söyleyebiliriz.

- OLTP sistemler girdi/çıktı yoğunluklu, veri ambarları ise sorgulama yoğunluklu işleri kapsar.
- Bu nedenle her iki sistemin amacı ötesinde, yapıları ve işleyiş biçimleri de farklıdır.