DEĞİŞKEN TÜRLERİ

⇒ Kategorik Değişkenler → Nominal (Sınıflayıcı)
→ Ordinal (Sıralayıcı)

⇒ Sürekli Değişkenler → Aralık
→ Oransal

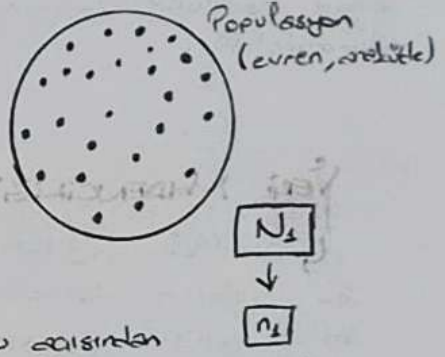
NOT
Kodların yeri değiştirildiğinde doğal sıralama bozulmuyorsa nominal, bozuluyorsa ordinaldir.

ÖRNEKLEM TEKNİKLERİ

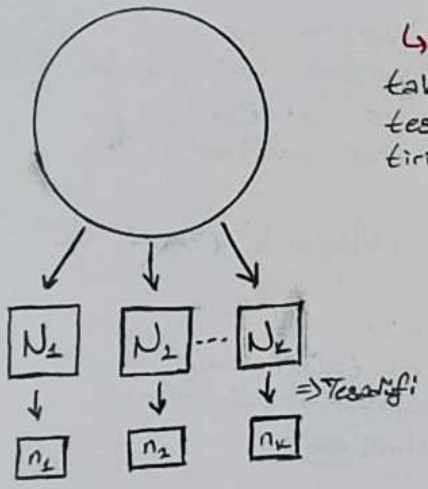
*) Tesadüfi Örnekleme Yöntemleri

1) Basit Tesadüfi Örnekleme

↳ Anakütledeki birimlerin tamamı her bir birime eşit seçilme şansı tanımak üzere bir torbaya atılır ve belirlenen örneklem sayısı kadar birim tesadüfi olarak seçilir.



2) Tabakalı Örnekleme



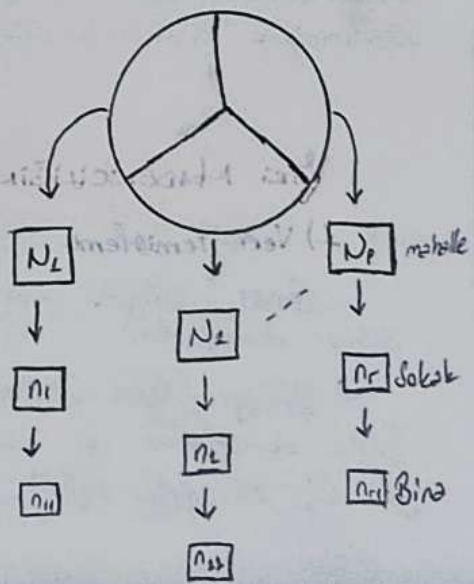
↳ Anakütle araştırılan konu adısından tabakalara ayrıldığıysa her bir tabakadan basit tesadüfi örnekleme metoduyla örneklem seçilip birleştirilerek analize tabi tutulur.

⇒ Örneğin bir sanayi bölgesindeki şirketler üzerine bir araştırma yapılmak istendiğinde şirketler küçük, orta ve büyük ölçekli şirketler olmak üzere tabakalara ayrılır ve her bir tabakadan tesadüfi örnekleme metoduyla birimler seçilir.

3) Küme Örneklemi

↳ Anakütle içerisinde doğal bir kümelenme söz konusuysa her bir kümeden basit tesadüfi örnekleme metoduyla birimler seçilir. Daha sonra birleştirilerek analiz edilir.

⇒ Örneğin Aksaray ilinde yapılacak bir araştırmada tesadüfi yöntemle mahalleler, mahallelerden tesadüfi yöntemle sokaklar ve binalar seçilir.



*) Tesadüfi Olmayan Örnekleme Yöntemleri

1) Kolayda Örnekleme

↳ Anakütleden ulaşılabilen herhangi birimlerin seçilmesiyle oluşturulan örnekleme tekniğidir.

2) Kartopu Örneklemesi

↳ Anakütle içerisindeki birimler birbirleriyle yüksek ilişki içerisindeyse genelde kartopu örneklemesi yapılır. Kartopu örneklemesinde araştırmaya bir kişi ile başlanır, araştırma o kişinin yönlendirdiği kişiler ve onlarında yönlendirdiği kişiler, olmak üzere örneklem sayısı çoğaltılır. İstenen örneklem sayısına ulaşılınca örneklem seçimi durdurulur.

VERİ MADENCİLİĞİ YÖNTEMLERİ

↳ Özellikle dijital ortamlarda sınırsız sayıda verinin yer alması bu verilerin nispeten daha ucuz yollarla veri tabanlarında hatta bulut bilişim gibi sistemlerin gelişmesiyle kişi, kurum ya da kuruluşların ellerinde bulunmaktadır. Bu veri yığınlarında matematiksel ve istatistiksel yöntemlerle anlamlı sonuçlar elde etmek işine veri madenciliği denilir.

- Uygulama Alanları => Pazarlama, bankacılık, elektronik ticaret vb. gibi çok sayıda alanda uygulanmaktadır.

→ Örneğin müşteri satın alma davranışının, kredi taleplerinin değerlendirilmesinde, sosyal alanda müşteri yorumlarının değerlendirilmesinde gibi pek çok sayıda alanda uygulanabilmektedir.

VERİ MADENCİLİĞİNİN SÜRECİ

1) Veri Temizleme

Bazı büyük verilerde veriler istenilen özelliklere sahip olmayabilir.

Örneğin bazı yöntemler dahada kayıp verinin olmasını istemez. Böyle durumlarda ya kayıp veriler genelde serinin ortalamasıyla tamamlanır ya da bütünüyle ilgili gözlem atılır.

Müşteri No	Yaş	Cinsiyeti	Gelir	Kredi
1	38	Kadın	3.000	Onaylandı
2	35	Erkek	5.000	Onaylanmadı
~~3~~	~~36~~	~~Kadın~~	~~8.000~~	~~Onaylandı~~
X	36		8.000	

2) Veri Bütünleştirme

Farklı veri tabanlarından elde edilen verilerin tek bir veri tabanı altında toplanmasıdır.

3) Veri Dönüştürme

Veri madenciliği algoritmaları uygulanmadan önce ortalama ve varyansları dolayısıyla aldıkları değerlerin birbirinden önemli ölçüde farklı olduğu değişkenlerle çalışırken veri dönüştürme işlemi uygulanır.

→ Temelde 2 yöntem kullanılır.

a) Minimum – maximum Normalleştirilmesi

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

X_i	X_i^*
3	0.005
30	0,156
45	0,244
150	0.854
175	1

$X_3^* = \frac{3-3}{175-3} = \frac{1}{172} = 0.005$

$X_{30}^* = \frac{30-3}{175-3} = \frac{27}{172} = 0,156$

$X_{45}^* = \frac{45-3}{175-3} = \frac{42}{172} = 0,244$

$X_{150}^* = \frac{150-3}{175-3} = \frac{147}{172} = 0,854$

$X_{175}^* = \frac{175-3}{175-3} = \frac{172}{172} = 1$

b) Z Score Standartlaştırılması

→ Formül $X^* = \frac{X - \bar{X}}{\sigma_X}$ $\quad \sigma_X = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$

X_i	$(x_i - \bar{x})^2$	X^*
3	$(3-80,6)^2 = 6021,76$	-1,010
30	$(30-80,6)^2 = 2560,36$	-0,659
45	$(45-80,6)^2 = 1267,36$	-0,463
150	$(150-80,6)^2 = 4816,34$	0,903
175	$(175-80,6)^2 = 8911,36$	1,229
403	23577,18	

$\bar{X} = \frac{403}{5} = 80,6$

$\sigma_X = \sqrt{\frac{23577,18}{5-1}} = \sqrt{5894,29} = 76,77$

$X_3^* = \frac{3-80,6}{76,77} = \frac{-77,6}{76,77} = -1,010$

$X_{30}^* = \frac{30-80,6}{76,77} = \frac{-50,6}{76,77} = -0,659$

$X_{45}^* = \frac{45-80,6}{76,77} = \frac{-35,6}{76,77} = -0,463$

$X_{150}^* = \frac{150-80,6}{76,77} = \frac{69,4}{76,77} = 0,903$

$X_{175}^* = \frac{175-80,6}{76,77} = \frac{94,4}{76,77} = 1,229$

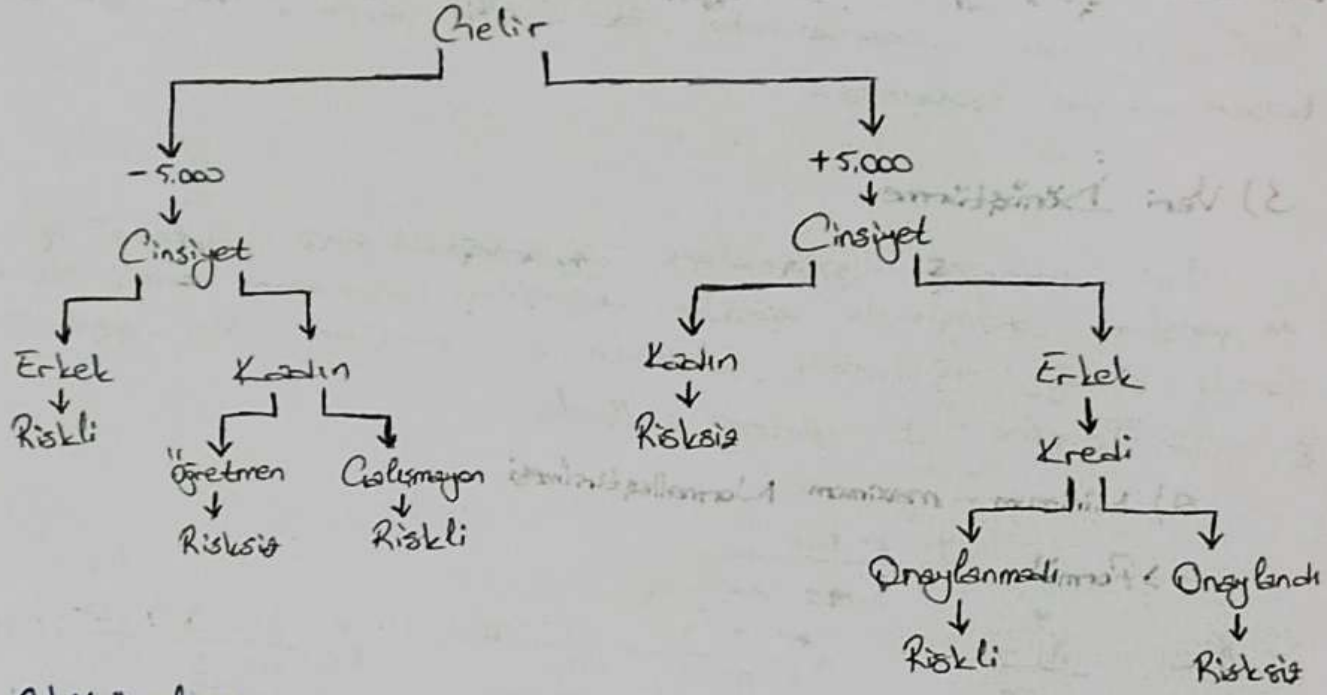
Veri Madenciliği Yöntemleri

Veri madenciliği yöntemleri temelde üçe ayrılır.

1) Sınıflama

Eğer veri tabanından elde edilen verilerle sınıflayıcı kural kuralları oluşturulmak isteniyorsa sınıflandırma teknikleri kullanılır.

Örneğin bir bankanın geçmiş müşterilerinin yaş, cinsiyet, gelir gibi özelliklerine bakarak kredi verip vermemeyi ya da riskli müşteri olup olmadıklarını sınıflandırabiliriz.



2) Kümeleme

Kümeleme, verilerin kendi aralarındaki benzerlikleri dikkate alarak gruplandırma işlemidir.

Örneğin üniversiteler öğrencilerin memnuniyet düzeyine göre, alınan TÜBİTAK destek sayısına göre, mezunlarının işe yerleştirilme oranlarına göre kümelenebilir.

3) Birliktelik Kuralları

Bir veri tabanında yer alan birbiriyle ilişkili verilerin incelenerek hangi olay ya da durumların birlikte ya da eş zamanlı ortaya çıktığını belirlemeye çalışan veri madenciliği yöntemleridir.

Örneğin bir marketten bir yılda yapılan bütün alış-verişlerin fişleri incelenerek hangi ürünlerin birlikte alındıkları tespit edilebilir. Böylece örneğin internet reklamlarında bu ürünler birlikte yer alır.

Sınıflandırma Algoritmaları

1) Karar Ağaçları

Sınıflandırma algoritmalarında belli değişkenlerin incelenmesiyle hedef niteliğin sınıfının belirlenmesi amaçlanmıştır. Çoğu veri madenciliği algoritmasında olduğu gibi eğitim ve test seti ya da kümesi bulunmaktadır.

Eğitim seti, sınıfları önceden belirlenmiş verinin yapıyı öğrenme aşamasını içerir. Verinin eğitildiği yer burasıdır.

Test setiyle ise algoritmaya sınıfı belli olmayan veriler gösterilip sınıflandırma performansı ölçülür.

	Cinsiyeti	Yaşı	Önceki kredi Miktarı	~~Kredi Onay~~
1	Kadın	38	10.000	Evet
2	Erkek	25	15.000	Hayır

Karar ağaçları bir ağaç gibi kök, dal ve sınıflama etiketleriyle gösterilen yapraklardan oluşan bir yapı gösterilir.

Oluinan tarafından 1980'li yılların sonunda gerçekleştirilmiştir. En sık kullanılan karar ağacı algoritmaları ID3, C4.5 (C5 geliştirildi) ve Classification and Repression Trees (CART) sınıflandırma ve regresyon ağaçlarıdır.

a) ID3 Algoritmaları

Entropi tabanlı algoritmalardır. Bir sistemdeki belirsizliğin ölçüsüne entropi denir.

"S" bir kaynak olmak üzere, kaynağın $m_1, m_2, \dots, m_n$ olarak mesajların olduğunu varsayalım. Bu mesajların üretilme olasılığına "P_i" dersek "S" kaynağının entropisi

$$H(S) = -\sum P_i \cdot \log_2(P_i)$$

şeklinde gösterilir.

Ör) Bir deneyin sonuçlarının olasılıkları;

1. deney için $\Rightarrow \frac{1}{2}$

2. deney için $\Rightarrow \frac{1}{3}$

3. deney için $\Rightarrow \frac{1}{6}$ olarak varsayılsın. Entropisi kaçtır?

P_i

$\frac{1}{2} = 0,50$

$\frac{1}{3} = 0,33$

$\frac{1}{6} = 0,16$

$$H(S) = -\sum P_i \cdot \log_2(P_i)$$

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

$$H(S) = -\left(0,50 \cdot \log_2 0,50 + 0,33 \cdot \log_2 0,33 + 0,16 \cdot \log_2 0,16\right)$$

$$= -\left(0,50 \cdot \frac{\log 0,50}{\log_2} + 0,33 \cdot \frac{\log 0,33}{\log_2} + 0,16 \cdot \frac{\log 0,16}{\log_2}\right)$$

$$= -\left(0,50 \cdot \frac{-0,30}{0,30} + 0,33 \cdot \frac{-0,48}{0,30} + 0,16 \cdot \frac{-0,79}{0,30}\right)$$

$$= -\left(0,50 \cdot (-1) + 0,33 \cdot (-1,60) + 0,16 \cdot (-2,63)\right)$$

$$= -\left(-0,50 + (-0,52) + (-0,42)\right) = -(-0,50 - 0,52 - 0,42)$$

$$= -(-1,44) = 1,44$$

Q4 10 elemanlı risk kümesi şu şekildedir;

(Var, Var, Var, Yok, Var, Yok, Yok, Var, Var, Yok)

entropisi kaçtır?

Var → 6 Yok → 4

$P_1 = Var = \frac{6}{10} = 0,6$

$P_2 = Yok = \frac{4}{10} = 0,4$

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

$H(Risk) = -(0,6 \cdot \log_2 0,6 + 0,4 \cdot (\log_2 0,4)$

$\log_2 0,6 = \frac{\log 0,6}{\log 2} = \frac{-0,22}{0,30} = -0,73$

$\log_2 0,4 = \frac{\log 0,4}{\log 2} = \frac{-0,39}{0,30} = -1,30$

$H(Risk) = -(0,6 \cdot (-0,73) + 0,4 \cdot (-1,30))$

$= -(-0,43 + (-0,52)) = -(-0,43 - 0,52) = -(-0,95) = 0,95$

Dallanma için Niteliklerin Seçilmesi ve Kazanç Ölçütü

bilgi kriteri : $H(X,T) = \sum \frac{|T_i|}{|T|} \cdot H(T_i)$

Kazanç Ölçütü : Kazanç$(X,T) = H(T) - H(X,T)$

Uygulama

Hava (0,26)	Isı (0,04)	Nem (0,15)	Rüzgar (0,05)	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

1. Adım: Kök niteliği bulunması

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

$$= -(0{,}35 \cdot \log_2 0{,}35 + 0{,}64 \cdot \log_2 0{,}64)$$

$$= -\left(0{,}35 \cdot \frac{\log 0{,}35}{\log 2} + 0{,}64 \cdot \frac{\log 0{,}64}{\log 2}\right)$$

$$= -\left(0{,}35 \cdot \frac{-0{,}45}{0{,}30} + 0{,}64 \cdot \frac{-0{,}19}{0{,}30}\right) = -(0{,}35 \cdot (-1.5) + 0{,}64 \cdot (-0{,}63))$$

$$= -(-0{,}52 + (-0{,}40)) = -(-0{,}52 - 0{,}40) = -(-0{,}92) = 0{,}92$$

$$Kazanç(Hava) = H(Oyun) - H(Hava, Oyun)$$

$$H(Hava, Oyun) = H(Hava_{güneşli}) + H(Hava_{bulutlu}) + H(Hava_{yağmurlu})$$

$$H(Hava_{güneşli}) = \frac{5}{14}\left(\frac{3}{5} \cdot \log_2\frac{3}{5} + \frac{2}{5} \cdot \log_2\frac{2}{5}\right)$$

$$= \frac{5}{14}(0{,}60 \cdot \log_2 0{,}60 + 0{,}40 \cdot \log_2 0{,}40) = \frac{5}{14}\left(0{,}60\frac{\log 0{,}60}{\log 2} + 0{,}40\frac{\log 0{,}40}{\log 2}\right)$$

$$= \frac{5}{14}\left(0{,}60 \cdot \frac{0{,}22}{0{,}30} + 0{,}40\frac{0{,}39}{0{,}30}\right) = \frac{5}{14}(0{,}60 \cdot 0{,}73 + 0{,}40 \cdot 1{,}3) = 0{,}35(0{,}43 + 0{,}52)$$

$$= 0{,}35 \cdot 0{,}95 = 0{,}33$$

$$H(Hava_{yağmurlu}) = \underbrace{\frac{5}{14}}_{0{,}35}\underbrace{\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)}_{0{,}95} = 0{,}35 \cdot 0{,}95 = 0{,}33$$

$$H(Hava_{bulutlu}) = \frac{4}{14}\left(\frac{4}{4}\log_2\frac{4}{4}\right) = 0$$

$$H(Hava, Oyun) = 0{,}33 + 0{,}33 + 0 = 0{,}66$$

$$Kazanç(Hava) = H(Oyun) - H(Hava, Oyun)$$

$$= 0{,}92 - 0{,}66 = 0{,}26$$

$\rightarrow 0,92$

$Kazanç(Isı) = H(Oyun) - H(Isı, Oyun)$

$H(Isı, Oyun) = H(Isı_{sıcak}) + H(Isı_{ılık}) + H(Isı_{soğuk})$

$H(Isı_{sıcak}) = \frac{4}{14}\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = \frac{4}{14}(0,50\cdot\log_2 0,50 + 0,50\cdot\log_2 0,50)$

$= \frac{4}{14}\left(0,50\frac{\log 0,50}{\log 2} + 0,50\frac{\log 0,50}{\log 2}\right) = \frac{4}{14}\left(0,50\frac{0,30}{0,30} + 0,50\frac{0,30}{0,30}\right)$

$= \frac{4}{14}(0,50\cdot 1 + 0,50\cdot 1) = \frac{4}{14}(0,50 + 0,50) = 0,28\cdot 1 = 0,28$

$H(Isı_{ılık}) = \frac{6}{14}\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = \frac{6}{14}(0,66\log_2 0,66 + 0,33\log_2 0,33)$

$= \frac{6}{14}\left(0,66\frac{\log 0,66}{\log 2} + 0,33\frac{\log 0,33}{\log 2}\right) = \frac{6}{14}\left(0,66\frac{0,18}{0,30} + 0,33\frac{0,48}{0,30}\right)$

$= \frac{6}{14}(0,66\cdot 0,60 + 0,33\cdot 1,60) = \frac{6}{14}(0,39 + 0,52) = 0,42\cdot 0,91 = 0,38$

$H(Isı_{soğuk}) = \frac{4}{14}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = \frac{4}{14}(0,75\cdot\log_2 0,75 + 0,25\cdot\log_2 0,25)$

$= \frac{4}{14}\left(0,75\frac{\log 0,75}{\log 2} + 0,25\frac{\log 0,25}{\log 2}\right) = \frac{4}{14}\left(0,75\frac{0,12}{0,30} + 0,25\frac{0,60}{0,30}\right)$

$= \frac{4}{14}(0,75\cdot 0,40 + 0,25\cdot 2) = \frac{4}{14}(0,30 + 0,50) = 0,28\cdot 0,80 = 0,22$

$H(Isı, Oyun) = 0,28 + 0,38 + 0,22 = 0,88$

$Kazanç(Isı) = 0,92 - 0,88 = 0,04$

$$Kazanç(Nem) = H(Oyun) - H(Nem, Oyun)$$
$$\hookrightarrow 0,92$$

$$H(Nem, Oyun) = H(Nem_{yüksek}) + H(Nem_{normal})$$

$$H(Nem_{yüksek}) = \frac{7}{14}\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right)$$

$$= \frac{7}{14}(0,42 \cdot \log_2 0,42 + 0,57 \cdot \log_2 0,57) = \frac{7}{14}\left(0,42\frac{\log 0,42}{\log 2} + 0,57\frac{\log 0,57}{\log 2}\right)$$

$$= \frac{7}{14}\left(0,42\frac{0,37}{0,30} + 0,57\frac{0,24}{0,30}\right) = \frac{7}{14}(0,42 \cdot 1,23 + 0,57 \cdot 0,80)$$

$$= 0,50(0,51 + 0,45) = 0,50 \cdot 0,96 = 0,48$$

$$H(Nem_{normal}) = \frac{7}{14}\left(\frac{6}{7}\log_2\frac{6}{7} + \frac{1}{7}\log_2\frac{1}{7}\right)$$

$$= \frac{7}{14}(0,85 \cdot \log_2 0,85 + 0,14 \cdot \log_2 0,14) = \frac{7}{14}\left(0,85\frac{\log 0,85}{\log 2} + 0,14\frac{\log 0,14}{\log 2}\right)$$

$$= \frac{7}{14}\left(0,85\frac{0,07}{0,30} + 0,14\frac{0,85}{0,30}\right) = \frac{7}{14}(0,85 \cdot 0,23 + 0,14 \cdot 2,83)$$

$$= 0,50(0,19 + 0,39) = 0,50 \cdot 0,58 = 0,29$$

$$H(Nem, Oyun) = 0,48 + 0,29 = 0,77$$

$$Kazanç(Nem) = 0,92 - 0,77 = 0,15$$

Kazanç (Rüzgar) = H (Oyun) − H (Rüzgar, Oyun)

↳ 0,92

$$H(\text{Rüzgar, Oyun}) = H(\text{Rüzgar}_{\text{hafif}}) + H(\text{Rüzgar}_{\text{kuvvetli}})$$

$$H(\text{Rüzgar}_{\text{hafif}}) = \frac{8}{14}\left(\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right)$$

$$= \frac{8}{14}(0{,}75\log_2 0{,}75 + 0{,}25\log_2 0{,}25) = \frac{8}{14}\left(0{,}75\frac{\log 0{,}75}{\log 2} + 0{,}25\frac{\log 0{,}25}{\log 2}\right)$$

$$= \frac{8}{14}\left(0{,}75\frac{0{,}12}{0{,}30} + 0{,}25\frac{0{,}60}{0{,}30}\right) = \frac{8}{14}(0{,}75\cdot 0{,}40 + 0{,}25\cdot 2)$$

$$= 0{,}57(0{,}30 + 0{,}50) = 0{,}57\cdot 0{,}80 = 0{,}45$$

$$H(\text{Rüzgar}_{\text{kuvvetli}}) = \frac{6}{14}\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right)$$

$$= \frac{6}{14}(0{,}50\cdot\log_2 0{,}50 + 0{,}50\cdot\log_2 0{,}50) = \frac{6}{14}\left(0{,}50\frac{\log 0{,}50}{\log 2} + 0{,}50\frac{\log 0{,}50}{\log 2}\right)$$

$$= \frac{6}{14}\left(0{,}50\frac{0{,}30}{0{,}30} + 0{,}50\frac{0{,}30}{0{,}30}\right) = \frac{6}{14}(0{,}50\cdot 1 + 0{,}50\cdot 1)$$

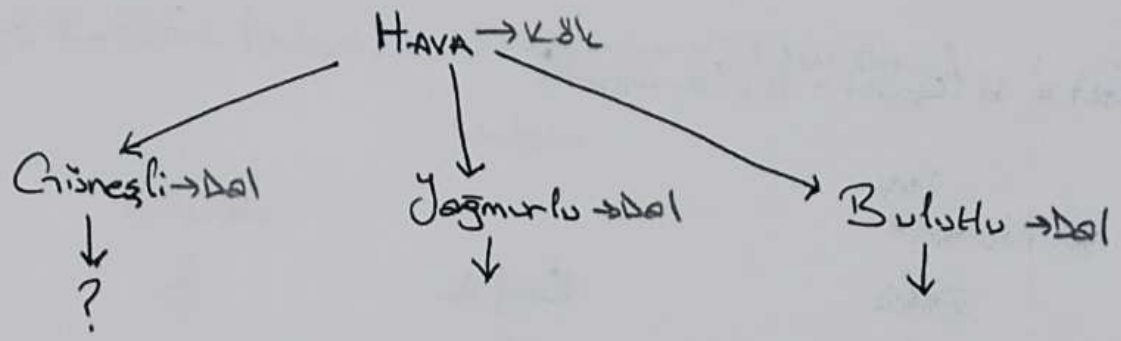
$$= 0{,}42(0{,}50 + 0{,}50) = 0{,}42\cdot 1 = 0{,}42$$

$$H(\text{Rüzgar, Oyun}) = 0{,}45 + 0{,}42 = 0{,}87$$

$$\text{Kazanç (Rüzgar)} = 0{,}92 - 0{,}87 = 0{,}05$$

Nitelik	Kazanç
Hava	0,26
Isı	0,04
Nem	0,15
Rüzgar	0,05

Kök
↳ Hava



1) Hava niteliğinin "güneşli" değeri için dallanma

Hava	Isı (0,55)	Nem (0,95)	Rüzgar (0,01)	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.

$$H(\text{Oyun}) = -\left(\frac{2}{5}\cdot\log_2\frac{2}{5} + \frac{3}{5}\cdot\log_2\frac{3}{5}\right)$$

$$= -\left(0{,}40\cdot\log_2 0{,}40 + 0{,}60\cdot\log_2 0{,}60\right)$$

$$= -\left(0{,}40\cdot\frac{\log 0{,}40}{\log 2} + 0{,}60\,\frac{\log 0{,}60}{\log 2}\right)$$

$$= -\left(0{,}40\,\frac{-0{,}39}{0{,}30} + 0{,}60\,\frac{-0{,}22}{0{,}30}\right) = -\left(0{,}40\cdot(-1{,}3) + 0{,}60\cdot(-0{,}73)\right)$$

$$= -\left(-0{,}52 + (-0{,}43)\right) = -(-0{,}52 - 0{,}43) = -(-0{,}95) = 0{,}95$$

→ Kazanç (Rüzgar) = H(Oyun) − H(Rüzgar, Oyun)

Rüzgar

Sıcak	Ilık	Soğuk
$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

evet	hayır	evet	hayır	evet	hayır
0	$\frac{2}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{1}$	0

$H(Oyun) = 0,95$

$$H(Rüzgar, Oyun) = \underbrace{\frac{2}{5}\left(\frac{2}{2}\cdot\log_2\frac{2}{2}\right)}_{0} + \frac{2}{5}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\cdot\log_2\frac{1}{2}\right) + \underbrace{\frac{1}{5}\left(\frac{1}{1}\cdot\log_2\frac{1}{1}\right)}_{0}$$

$$= 0 + 0,40\,(0,50\log_2 0,50 + 0,50\cdot\log_2 0,50) + 0$$

$$= 0,40\left(0,50\,\frac{\log 0,50}{\log 2} + 0,50\,\frac{\log 0,50}{\log 2}\right) = 0,40\left(0,50\,\frac{0,30}{0,30} + 0,50\,\frac{0,30}{0,30}\right)$$

$$= 0,40\,(0,50\cdot 1 + 0,50\cdot 1) = 0,40\,(0,50 + 0,50) = 0,40\cdot 1 = 0,40$$

Kazanç (Rüzgar) = 0,95 − 0,40 = 0,55

→ Kazanç (Nem) = H(Oyun) − H(Nem, Oyun)

Nem

Yüksek
$\frac{3}{5}$

Normal
$\frac{2}{5}$

evet — hayır | evet — hayır

0 — $\frac{3}{3}$ | $\frac{2}{2}$ — 0

$H(Oyun) = 0,95$

$H(Nem, Oyun) = \frac{3}{5}\left(\frac{3}{3}\cdot\log_2\frac{3}{3}\right) + \frac{2}{5}\left(\frac{2}{2}\cdot\log_2\frac{2}{2}\right)$

$\underbrace{\quad}_{0} \qquad \underbrace{\quad}_{0}$

$H(Nem, Oyun) = 0$ Kazanç (Nem) = 0,95 − 0 = 0,95

→ Kazanç (Rüzgar) = H(Oyun) − H(Rüzgar, Oyun)

Rüzgar

Hafif
$\frac{3}{5}$

Kuvvetli
$\frac{2}{5}$

evet — hayır | evet — hayır

$\frac{1}{3}$ — $\frac{2}{3}$ | $\frac{1}{2}$ — $\frac{1}{2}$

$H(Oyun) = 0,95$

$H(Rüzgar, oyun) = \frac{3}{5}\left(\frac{1}{3}\cdot\log_2\frac{1}{3} + \frac{2}{3}\cdot\log_2\frac{2}{3}\right) + \frac{2}{5}\left(\frac{1}{2}\cdot\log_2\frac{1}{2} + \frac{1}{2}\cdot\log_2\frac{1}{2}\right)$

$= 0,60\,(0,33\cdot\log_2 0,33 + 0,66\cdot\log_2 0,66) + 0,40\,(0,50\cdot\log_2 0,50 + 0,50\cdot\log_2 0,50)$

$= 0,60\left(0,33\cdot\frac{\log 0,33}{\log_2} + 0,66\,\frac{\log 0,66}{\log_2}\right) + 0,40\left(0,50\,\frac{\log 0,50}{\log_2} + 0,50\,\frac{\log 0,50}{\log_2}\right)$

$= 0,60\left(0,33\,\frac{0,48}{0,30} + 0,66\,\frac{0,18}{0,30}\right) + 0,40\left(0,50\,\frac{0,30}{0,30} + 0,50\,\frac{0,30}{0,30}\right)$

$= 0,60\,(0,33\cdot 1,60 + 0,66\cdot 0,60) + 0,40\,(0,50\cdot 1 + 0,50\cdot 1)$

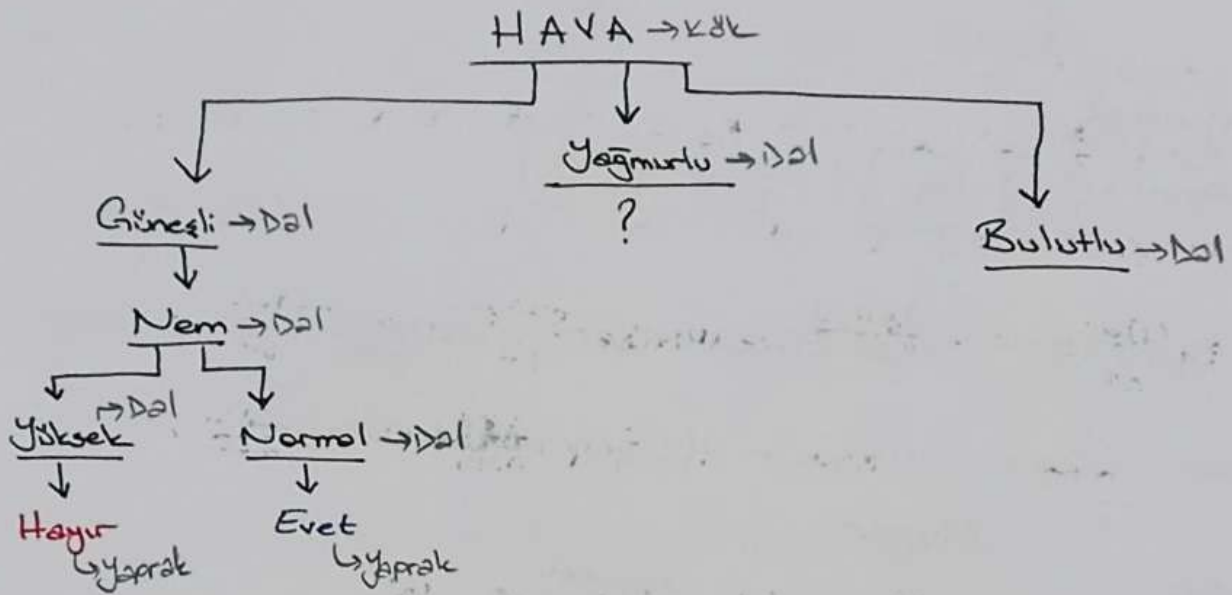
$= 0,60\,(0,52 + 0,39) + 0,40\,(0,50 + 0,50) = 0,60\,(0,91) + 0,40\cdot 1$

$= 0,54 + 0,40 = 0,94$

Kazanç (Rüzgar) = 0,95 − 0,94 = 0,01

Öznitelik	Kazanç
Isı	0,55
Nem	0,95
Rüzgar	0,01

→ En büyük kazanç.



2) Hava niteliğinin "yağmurlu" değeri için dallanma

Hava	Isı (0,01)	Nem (0,01)	Rüzgar (0,95)	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.

$$H(\text{Oyun}) - \left(\frac{3}{5}\cdot\log_2\frac{3}{5} + \frac{2}{5}\cdot\log_2\frac{2}{5}\right) = -(0{,}60\cdot\log_2 0{,}60 + 0{,}40\cdot\log_2 0{,}40)$$

$$= -\left(0{,}60\,\frac{\log 0{,}60}{\log 2} + 0{,}40\,\frac{\log 0{,}40}{\log 2}\right) = -\left(0{,}60\,\frac{-0{,}22}{0{,}30} + 0{,}40\,\frac{-0{,}39}{0{,}30}\right)$$

$$= -(0{,}60\cdot(-0{,}73) + 0{,}40\cdot(-1{,}30)) = -(-0{,}43 + (-0{,}52)) = -(-0{,}43 - 0{,}52)$$

$$= -(-0{,}95) = 0{,}95$$

→ Kazanç (Isı) = H(Oyun) − H(Isı, Oyun)

Isı

Soğuk	Sıcak	Ilık
2/5	0	3/5
evet 1/2		evet 2/3
hayır 1/2		hayır 1/3

H(Oyun) = 0,95

$H(Isı, Oyun) = \frac{2}{5}\left(\frac{1}{2}\cdot\log_2\frac{1}{2} + \frac{1}{2}\cdot\log_2\frac{1}{2}\right) + \frac{3}{5}\left(\frac{2}{3}\cdot\log_2\frac{2}{3} + \frac{1}{3}\cdot\log_2\frac{1}{3}\right)$

$= 0,40\,(0,50\cdot\log_2 0,50 + 0,50\cdot\log_2 0,50) + 0,60\,(0,66\cdot\log_2 0,66 + 0,33\cdot\log_2 0,33)$

$= 0,40\left(0,50\,\frac{\log 0,50}{\log 2} + 0,50\,\frac{\log 0,50}{\log 2}\right) + 0,60\left(0,66\,\frac{\log 0,66}{\log 2} + 0,33\,\frac{\log 0,33}{\log 2}\right)$

$= 0,40\left(0,50\,\frac{0,30}{0,30} + 0,50\,\frac{0,30}{0,30}\right) + 0,60\left(0,66\,\frac{0,18}{0,30} + 0,33\,\frac{0,48}{0,30}\right)$

$= 0,40\,(0,50\cdot 1 + 0,50\cdot 1) + 0,60\,(0,66\cdot 0,60 + 0,33\cdot 1,60)$

$= 0,40\,(0,50+0,50) + 0,60\,(0,39 + 0,52) = 0,40\cdot 1 + 0,60\cdot 0,91$

$= 0,40 + 0,54 = 0,94$

Kazanç (Isı) = 0,95 − 0,94 = 0,01

→ Kazanç (Nem) = H(Oyun) − H(Nem, Oyun)

Nem

Yüksek
$\frac{2}{5}$

evet $\frac{1}{2}$ hayır $\frac{1}{2}$

Normal
$\frac{3}{5}$

evet $\frac{2}{3}$ hayır $\frac{1}{3}$

$H(Oyun) = 0,95$

$H(Nem, Oyun) = \frac{2}{5}\left(\frac{1}{2}\cdot\log_2\frac{1}{2} + \frac{1}{2}\cdot\log_2\frac{1}{2}\right) + \frac{3}{5}\left(\frac{2}{3}\cdot\log_2\frac{2}{3} + \frac{1}{3}\cdot\log_2\frac{1}{3}\right)$

$= 0,40\,(0,50\cdot\log_2 0,50 + 0,50\cdot\log_2 0,50) + 0,60\,(0,66\cdot\log_2 0,66 + 0,33\cdot\log_2 0,33)$

$= 0,40\left(0,50\,\frac{\log 0,50}{\log 2} + 0,50\,\frac{\log 0,50}{\log_2}\right) + 0,60\left(0,66\,\frac{\log 0,66}{\log_2} + 0,33\,\frac{\log 0,33}{\log_2}\right)$

$= 0,40\left(0,50\,\frac{0,30}{0,30} + 0,50\,\frac{0,30}{0,30}\right) + 0,60\left(0,66\,\frac{0,18}{0,30} + 0,33\,\frac{0,48}{0,30}\right)$

$= 0,40\,(0,50\cdot 1 + 0,50\cdot 1) + 0,60\,(0,66\cdot 0,60 + 0,33\cdot 1,60)$

$= 0,40\,(0,50+0,50) + 0,60\,(0,39 + 0,52) = 0,40\cdot 1 + 0,60\cdot 0,91 = 0,40 + 0,54$

$= 0,94$

Kazanç (Nem) = $0,95 - 0,94 = 0,01$

→ Kazanç (Rüzgar) = H(Oyun) − H(Rüzgar, Oyun)

Rüzgar

Hafif
$\frac{3}{5}$

evet $\frac{3}{3}$ hayır 0

Kuvvetli
$\frac{2}{5}$

evet 0 hayır $\frac{2}{2}$

$H(Oyun) = 0,95$

$H(Rüzgar, Oyun) = \frac{3}{5}\underbrace{\left(\frac{3}{3}\cdot\log_2\frac{3}{3}\right)}_{0} + \frac{2}{5}\underbrace{\left(\frac{2}{2}\cdot\log_2\frac{2}{2}\right)}_{0}$

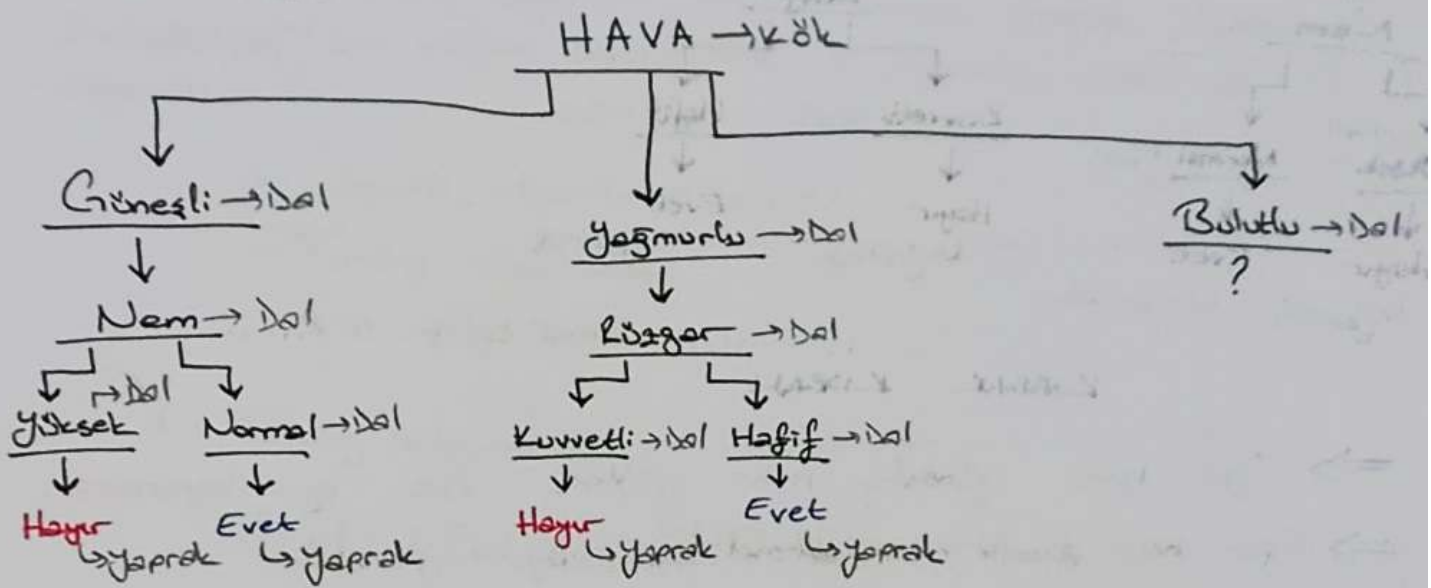
$0 + 0 = 0$

$H(Rüzgar, Oyun) = 0$

Kazanç (Rüzgar) = $0,95 - 0 = 0,95$

Öznitelik	Kazanç
Isı	0,01
Nem	0,01
Rüzgar	0,95

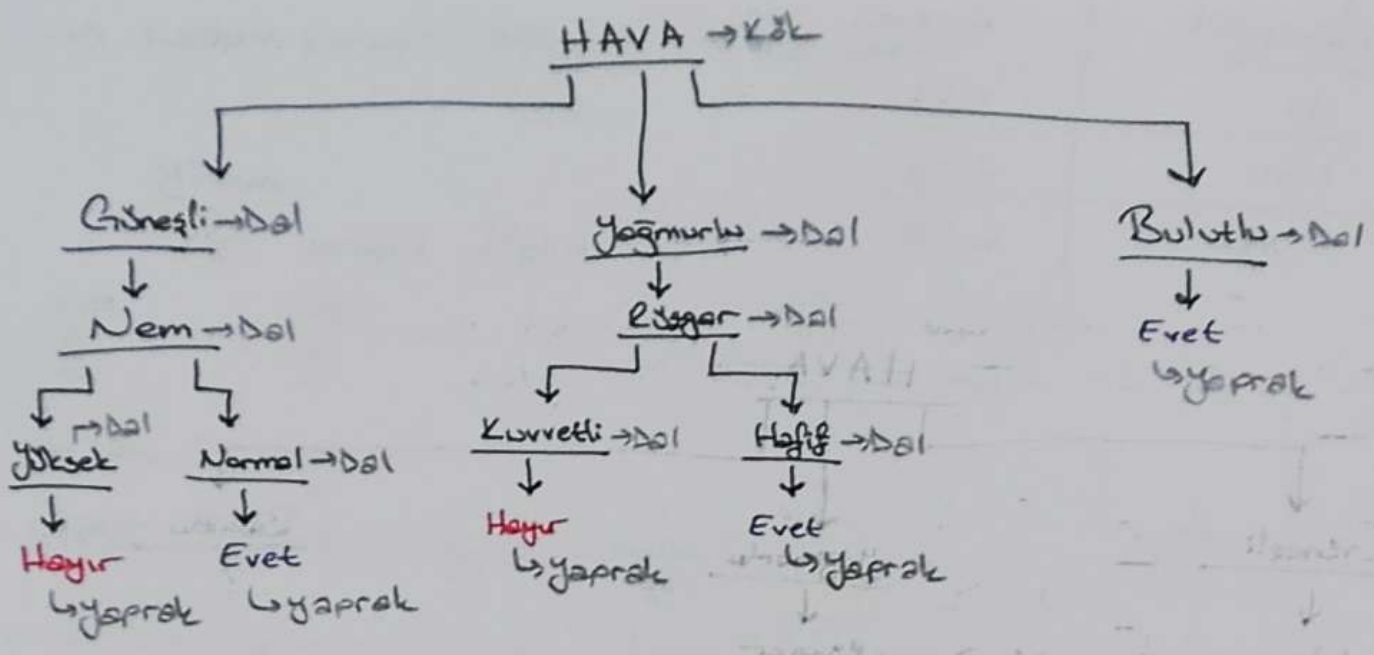
→ En büyük kazanç



3) Hava niteliğinin "bulutlu" değeri için dallanma

Hava	Isı	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.



KARAR KURALI

⇒ Eğer hava güneşli, nem yüksek ise oyun oynanamaz.
⇒ Eğer hava güneşli, nem normal ise oyun oynanabilir.
⇒ Eğer hava yağmurlu, rüzgar kuvvetli ise oyun oynanamaz.
⇒ Eğer hava yağmurlu, rüzgar hafif ise oyun oynanabilir.
⇒ Eğer hava bulutlu ise oyun oynanabilir.

b) C4,5 Algoritması

Yeni versiyonu C5 algoritmasıdır.

ID3 algoritmasının önemli bir dezavantajını yok etmek için ortaya çıkmıştır.

Önemli ⚠ Bilindiği gibi ID3 algoritmasında hem hedef hem de nitelik lerinde çok sayıda değişken süreklidir. Ancak gerçek dünya örnekmasında sürekli olan değişkenler kategorik hâle dönüştürülür.

Bunun için eşit bir değer belirlenir. Bu eşit değer serinin medyanı ya da aritmetik ortalaması olabilir. Genelde serinin aritmetik ortalaması kullanılır. Buna göre serinin terimleri aritmetik ortalamaya göre "≤" ve ">" şeklinde sınıflandırılır. Daha sonra işlemlere ID3 algoritmasındaki gibi devam edilir.

Bu algoritmanın WEKA programındaki adı J48'dir.

Sınıflandırma ve Regresyon Ağaçları
(CART)
(Classification and Regression Trees)

CART, karar ağacının ikili dallanması ilkesine dayanır. Dolayısıyla bir düğüm seçildiğinde o düğümden sadece iki dal çıkarılır.

İki önemli algoritması vardır:
- Twoing algoritması
- Gini algoritması

1) Twoing Algoritması

$$\phi(s|t) = 2 \cdot P_{sol} \cdot P_{sağ} \cdot \sum |P(j|t_{sol}) - P(j|t_{sağ})|$$

- Uygulama

Müşteri	Gelir	Eğitim	Sektör	Memnun
1	Normal	Orta	Bilişim	Evet
2	Büyük	İlk	Bilişim	Evet
3	Küçük	İlk	İnşaat	Evet
4	Büyük	Orta	İnşaat	Evet
5	Küçük	Orta	İnşaat	Evet
6	Büyük	Lise	İnşaat	Evet
7	Küçük	Lise	İnşaat	Evet
8	Büyük	Orta	Bilişim	Hayır
9	Küçük	Orta	Bilişim	Hayır
10	Büyük	Lise	Bilişim	Hayır
11	Küçük	Lise	Bilişim	Hayır

Aday Bölünme (s)	t_{sol}	$t_{sağ}$
1	Gelir = Normal	Gelir = Büyük, küçük
2	Gelir = Büyük	Gelir = Normal, küçük
3	Gelir = Küçük	Gelir = Normal, büyük
4	Eğitim = İlk	Eğitim = Orta, lise
5	Eğitim = Orta	Eğitim = İlk, lise
6	Eğitim = Lise	Eğitim = İlk, orta
7	Sektör = Bilişim	Sektör = İnşaat
8	Sektör = İnşaat	Sektör = Bilişim

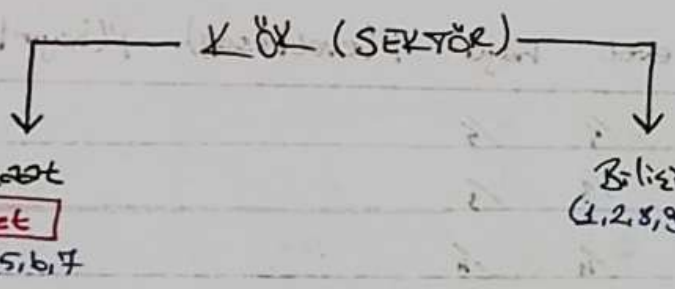
↳ Sola göre işlemler

Aday Bölünme	t_{sol}deki kayıt sayısı	P_{sol}	evet	hayır	P(evet|t_{sol})	P(hayır|t_{sol})
1	1/11	0.09	1/1	0	1.00	0.00
2	5/11	0.45	3/5	2/5	0.60	0.40
3	5/11	0.45	3/5	2/5	0.60	0.40
4	2/11	0.18	2/2	0	1.00	0.00
5	5/11	0.45	3/5	2/5	0.60	0.40
6	4/11	0.36	2/4	2/4	0.50	0.50
7	6/11	0.54	2/6	4/6	0.33	0.66
8	5/11	0.45	5/5	0	1.00	0.00

↳ Sağa göre işlemler

Aday Bölünme	$t_{sağ}$deki kayıt sayısı	$P_{sağ}$	evet	hayır	P(evet|$t_{sağ}$)	P(hayır|$t_{sağ}$)
1	10/11	0.90	6/10	4/10	0.60	0.40
2	6/11	0.54	4/6	2/6	0.66	0.33
3	6/11	0.54	4/6	2/6	0.66	0.33
4	9/11	0.81	5/9	4/9	0.55	0.44
5	6/11	0.54	4/6	2/6	0.66	0.33
6	7/11	0.63	5/7	2/7	0.71	0.28
7	5/11	0.45	5/5	0	1.00	0.00
8	6/11	0.54	2/6	4/6	0.33	0.66

Aday Bölünme	Φ
1	$\Rightarrow 2 \cdot 0{,}09 \cdot 0{,}90 \cdot (\lvert 1{,}00 - 0{,}60 \rvert + \lvert 0{,}00 - 0{,}40 \rvert) = 0{,}12$
2	$\Rightarrow 2 \cdot 0{,}45 \cdot 0{,}54 \cdot (\lvert 0{,}60 - 0{,}66 \rvert + \lvert 0{,}40 - 0{,}33 \rvert) = 0{,}06$
3	$\Rightarrow 2 \cdot 0{,}45 \cdot 0{,}54 \cdot (\lvert 0{,}60 - 0{,}66 \rvert + \lvert 0{,}40 - 0{,}33 \rvert) = 0{,}06$
4	$\Rightarrow 2 \cdot 0{,}18 \cdot 0{,}81 \cdot (\lvert 1{,}00 - 0{,}55 \rvert + \lvert 0{,}00 - 0{,}44 \rvert) = 0{,}09$
5	$\Rightarrow 2 \cdot 0{,}45 \cdot 0{,}54 \cdot (\lvert 0{,}60 - 0{,}66 \rvert + \lvert 0{,}40 - 0{,}33 \rvert) = 0{,}06$
6	$\Rightarrow 2 \cdot 0{,}36 \cdot 0{,}63 \cdot (\lvert 0{,}50 - 0{,}71 \rvert + \lvert 0{,}50 - 0{,}28 \rvert) = 0{,}19$
7	$\Rightarrow 2 \cdot 0{,}54 \cdot 0{,}45 \cdot (\lvert 0{,}33 - 1{,}00 \rvert + \lvert 0{,}66 - 0{,}00 \rvert) = 0{,}63$
8	$\Rightarrow 2 \cdot 0{,}45 \cdot 0{,}54 \cdot (\lvert 1{,}00 - 0{,}33 \rvert + \lvert 0{,}00 - 0{,}66 \rvert) = 0{,}63$



Müşteri	Gelir	Eğitim	Sektör	Memnun
1	Normal	orta	Bilişim	Evet
2	Büyük	ilk	Bilişim	Evet
8	Büyük	orta	Bilişim	Hayır
9	Küçük	orta	Bilişim	Hayır
10	Büyük	Lise	Bilişim	Hayır
11	Küçük	Lise	Bilişim	Hayır

Aday Bölünme	t_{sol}	$t_{sağ}$
1	Gelir = Normal	Gelir = Büyük, Küçük
2	Gelir = Büyük	Gelir = Normal, Küçük
3	Gelir = Küçük	Gelir = Büyük, Normal
4	Eğitim = ilk	Eğitim = orta, Lise
5	Eğitim = orta	Eğitim = ilk, Lise
6	Eğitim = Lise	Eğitim = ilk, orta

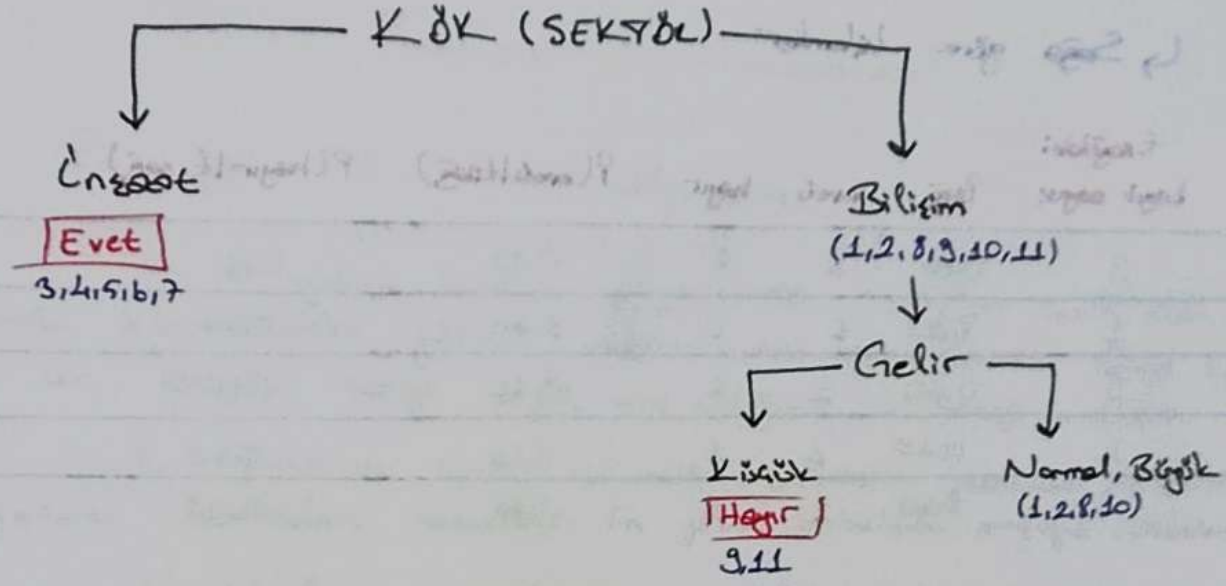
↳ Sola göre işlemler

Aday Bölünme	t sol deki kayıt sayısı	P_{sol}	evet	hayır	P(evet | t sol)	P(hayır | t sol)
1	1/6	0,16	1/1	0	1,00	0,00
2	3/6	0,50	1/3	2/3	0,33	0,66
3	2/6	0,33	0	2/2	0,00	1,00
4	1/6	0,16	1/1	0	1,00	0,00
5	3/6	0,50	1/3	2/3	0,33	0,66
6	2/6	0,33	0	2/2	0,00	1,00

↳ Sağa göre işlemler

Aday Bölünme	t sağdaki kayıt sayısı	$P_{sağ}$	evet	hayır	P(evet | t sağ)	P(hayır | t sağ)
1	5/6	0,83	1/5	4/5	0,20	0,80
2	3/6	0,50	1/3	2/3	0,33	0,66
3	4/6	0,66	2/4	2/4	0,50	0,50
4	5/6	0,83	1/5	4/5	0,20	0,80
5	3/6	0,50	1/3	2/3	0,33	0,66
6	4/6	0,66	2/4	2/4	0,50	0,50

Aday Bölünme	Φ
1	$\Rightarrow 2 \cdot 0,16 \cdot 0,83 \cdot (\lvert 1,00 - 0,20 \rvert + \lvert 0,00 - 0,80 \rvert) = 0,41$
2	$\Rightarrow 2 \cdot 0,50 \cdot 0,50 \cdot (\lvert 0,33 - 0,33 \rvert + \lvert 0,66 - 0,66 \rvert) = 0$
3	$\Rightarrow 2 \cdot 0,33 \cdot 0,66 \cdot (\lvert 0,00 - 0,50 \rvert + \lvert 1,00 - 0,50 \rvert) = 0,43$
4	$\Rightarrow 2 \cdot 0,16 \cdot 0,83 \cdot (\lvert 1,00 - 0,20 \rvert + \lvert 0,00 - 0,80 \rvert) = 0,41$
5	$\Rightarrow 2 \cdot 0,50 \cdot 0,50 \cdot (\lvert 0,33 - 0,33 \rvert + \lvert 0,66 - 0,66 \rvert) = 0$
6	$\Rightarrow 2 \cdot 0,33 \cdot 0,66 \cdot (\lvert 0,00 - 0,50 \rvert + \lvert 1,00 - 0,50 \rvert) = 0,43$



Müşteri	Gelir	Eğitim	Sektör	Memnun
1	Normal	Orta	Bilişim	Evet
2	Büyük	İlk	Bilişim	Evet
8	Büyük	Orta	Bilişim	Hayır
10	Büyük	Lise	Bilişim	Hayır

Adaylık Bölünme	tsol	tsağ
1	Gelir = Normal	Gelir = Büyük
2	Gelir = Büyük	Gelir = Normal
3	Eğitim = İlk	Eğitim = Orta, Lise
4	Eğitim = Orta	Eğitim = İlk, Lise
5	Eğitim = Lise	Eğitim = İlk, Orta

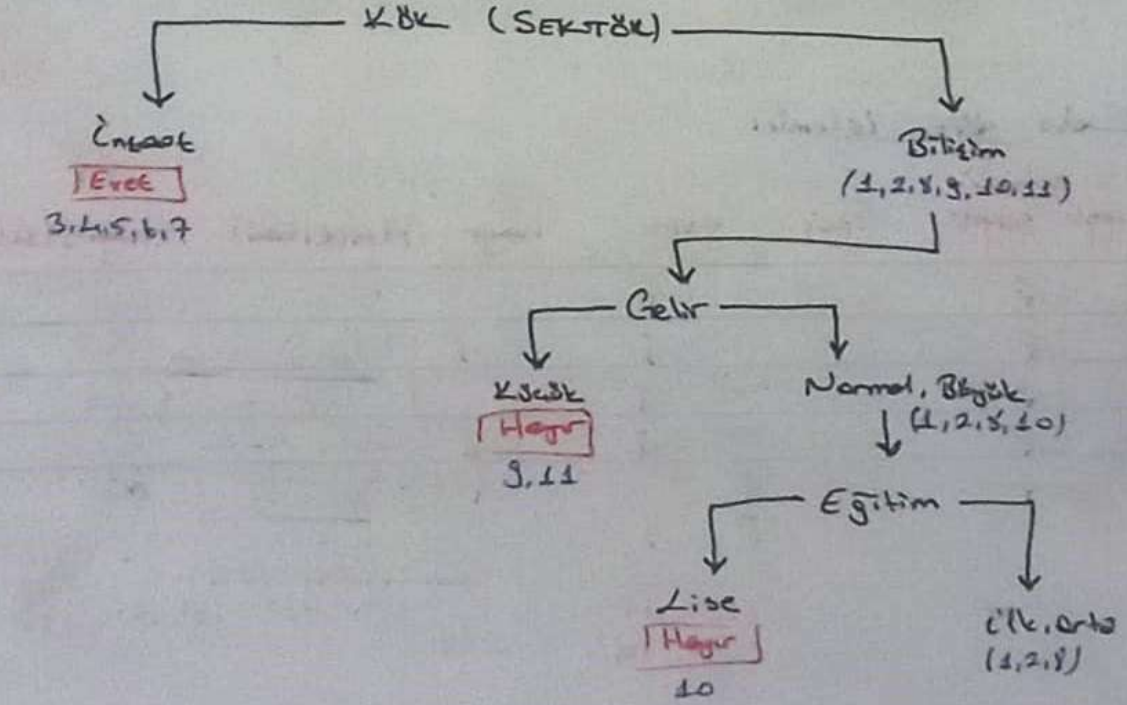
↳ Sola göre işlemler

Adaylık Bölünme	t soldaki kayıt sayısı	Psol	evet	hayır	P(evet|tsol)	P(hayır|tsol)
1	1/5	0,20	1/1	0	1,00	0,00
2	3/5	0,60	1/3	2/3	0,33	0,66
3	1/5	0,20	1/1	0	1,00	0,00
4	2/5	0,40	1/2	1/2	0,50	0,50
5	1/5	0,20	0	1/1	0,00	1,00

↳ Sağa göre işlemler

Aday Bilgisi	Kayıt sayısı	P(sağ)	evet	hayır	P(evet | sağ)	P(hayır | sağ)
1	3	0,60	1	2	0,33	0,66
2	1	0,20	1	0	1,00	0,00
3	3	0,60	1	2	0,33	0,66
4	2	0,40	1	1	0,50	0,50
5	3	0,60	2	1	0,66	0,33

Aday Bilgisi	Φ
1	$\Rightarrow 2 \cdot 0,20 \cdot 0,60 \cdot (\lvert 1,00 - 0,33 \rvert + \lvert 0,00 - 0,66 \rvert) = 0,31$
2	$\Rightarrow 2 \cdot 0,60 \cdot 0,20 \cdot (\lvert 0,33 - 1,00 \rvert + \lvert 0,66 - 0,00 \rvert) = 0,31$
3	$\Rightarrow 2 \cdot 0,20 \cdot 0,60 \cdot (\lvert 1,00 - 0,33 \rvert + \lvert 0,00 - 0,66 \rvert) = 0,31$
4	$\Rightarrow 2 \cdot 0,40 \cdot 0,40 \cdot (\lvert 0,50 - 0,50 \rvert + \lvert 0,50 - 0,50 \rvert) = 0$
5	$\Rightarrow 2 \cdot 0,20 \cdot 0,60 \cdot (\lvert 0,00 - 0,66 \rvert + \lvert 1,00 - 0,33 \rvert) = 0,31$



Bellek Tabanlı Sınıflandırma

En Yakın K – Komşu Algoritması

K en yakın komşu algoritması sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemeyi amaçlar.

Bu yöntemde örnek kümedeki birimlerin sonradan belirlenen değerlere uzaklıkları hesaplanır. En yakın birimlerin sınıfıyla etiketlenir.

Uzaklıkların hesaplanmasında öklid uzaklık formülünden yararlanılır.

$$d(i,j) = \sqrt{\sum (X_{ik} - X_{jk})^2}$$

Algoritma adımları şunlardır:

⇒ K (Verilen bir noktaya en yakın komşuların sayısı) belirlenir.
⇒ Uzaklıklar hesaplanır
⇒ Uzaklıklar sıralanır ve en küçük "K" tanesi seçilir.
⇒ Seçilen satırlar en çok hangi sınıfa ait belirlenen satırda o sınıfıyla etiketlenir.

$X_1 = 8$, $X_2 = 4$ yeni gözlem biriminin $K = 4$ olmak üzere sınıflandırmaya belirleyiniz.

X_1	X_2	y	uzaklık $d(i,j)$
2	4	Kötü	6,00
3	6	İyi	5,38
3	4	İyi	5,00
4	10	Kötü	7,21
5	8	Kötü	5,00
6	3	İyi	2,23 → 1
7	9	İyi	5,09
9	7	Kötü	3,16 → 3
11	7	Kötü	4,24 → 4
10	2	Kötü	2,82 → 2
8	4		

$d(i,j) = \sqrt{(2-8)^2 + (4-4)^2} = \sqrt{36} = 6,00$
$d(i,j) = \sqrt{(3-8)^2 + (6-4)^2} = \sqrt{29} = 5,38$
$d(i,j) = \sqrt{(3-8)^2 + (4-4)^2} = \sqrt{25} = 5,00$
$d(i,j) = \sqrt{(4-8)^2 + (10-4)^2} = \sqrt{52} = 7,21$
$d(i,j) = \sqrt{(5-8)^2 + (8-4)^2} = \sqrt{25} = 5,00$
$d(i,j) = \sqrt{(6-8)^2 + (3-4)^2} = \sqrt{5} = 2,23 \Rightarrow 1$
$d(i,j) = \sqrt{(7-8)^2 + (9-4)^2} = \sqrt{26} = 5,08$
$d(i,j) = \sqrt{(9-8)^2 + (7-4)^2} = \sqrt{10} = 3,16 \Rightarrow 3$
$d(i,j) = \sqrt{(11-8)^2 + (7-4)^2} = \sqrt{18} = 4,24 \Rightarrow 4$
$d(i,j) = \sqrt{(10-8)^2 + (2-4)^2} = \sqrt{8} = 2,82 \Rightarrow 2$

O hâlde $X_1 = 8$ ve $X_2 = 4$ iken

Sonuç: Kötü

Ağırlıklı Sınıflandırma

Bir önceki yöntemde sınıfı bilinmeyen yeni bir gözlem değeri için en sık tekrar eden sınıf seçiliyordu.

Ancak burada sadece K adet gözlem dikkate alındığında her zaman gerçekçi sonuç elde edilmeyebilir.

Bu yüzden bunun yerine sınıf K adet komşu arasından ağırlandırılarak seçilir.

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

Ör: 10 bireyin gen sayıları ve hemoglobin değerlerine göre kadın ve erkek olarak sınıflandırılmışlardır.

Buna göre X_1 değişkeni 0,10, X_2 değişkeni 0,50 değerini aldığına göre, $K=3$ yeni gözleminin cinsiyeti nedir?

X_1	X_2	Cins	$d(i,j)$ Uzaklık	(Ağırlık) Sıra
0,08	0,20	Erkek	0,30	3 → 11,11
0,07	0,07	Erkek	0,42	
0,20	0,09	Erkek	0,41	
1,00	0,20	Kadın	0,94	
0,05	0,06	Erkek	0,43	
0,20	0,25	Erkek	0,26	2 → 16,66
0,17	0,07	Erkek	0,42	
0,15	0,55	Kadın	0,06	1 → 333,33
0,50	0,08	Erkek	0,57	
0,10	0,06	Kadın	0,43	
0,10	0,50			

Uzaklık:

$$d(i,j) = \sqrt{(0,08-0,10)^2 + (0,20-0,50)^2} = \sqrt{(-0,02)^2 + (-0,30)^2} = \sqrt{0,09} = 0,30$$

$$d(i,j) = \sqrt{(0,07-0,10)^2 + (0,07-0,50)^2} = \sqrt{(-0,03)^2 + (-0,43)^2} = \sqrt{0,18} = 0,42$$

$$d(i,j) = \sqrt{(0,20-0,10)^2 + (0,09-0,50)^2} = \sqrt{(0,10)^2 + (-0,41)^2} = \sqrt{0,17} = 0,41$$

$$d(i,j) = \sqrt{(1,00-0,10)^2 + (0,20-0,50)^2} = \sqrt{(0,90)^2 + (-0,30)^2} = \sqrt{0,90} = 0,94$$

$$d(i,j) = \sqrt{(0,05-0,10)^2 + (0,06-0,50)^2} = \sqrt{(-0,05)^2 + (-0,44)^2} = \sqrt{0,19} = 0,43$$

$$d(i,\dot{1}) = \sqrt{(0,10-0,10)^2+(0,25-0,50)^2} = \sqrt{(0,10)^2+(-0,25)^2} = \sqrt{0,07} = 0,26$$

$$d(i,\dot{2}) = \sqrt{(0,17-0,10)^2+(0,07-0,50)^2} = \sqrt{(0,07)^2+(-0,43)^2} = \sqrt{0,18} = 0,42$$

$$d(i,\dot{3}) = \sqrt{(0,15-0,10)^2+(0,55-0,50)^2} = \sqrt{(0,05)^2+(0,05)^2} = \sqrt{0,004} = 0,06$$

$$d(i,\dot{4}) = \sqrt{(0,50-0,10)^2+(0,08-0,50)^2} = \sqrt{(0,40)^2+(-0,42)^2} = \sqrt{0,33} = 0,57$$

$$d(i,\dot{5}) = \sqrt{(0,10-0,10)^2+(0,06-0,50)^2} = \sqrt{(0)^2+(-0,44)^2} = \sqrt{0,19} = 0,43$$

Ağırlık:

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

1) $\Rightarrow \frac{1}{(0,06)^2} = \frac{1}{0,003} = 333,33 \rightarrow$ kadın

2) $\Rightarrow \frac{1}{(0,26)^2} = \frac{1}{0,06} = 16,66 \rightarrow$ erkek

3) $\Rightarrow \frac{1}{(0,30)^2} = \frac{1}{0,09} = 11,11 \rightarrow$ erkek

(+ Toplam) $\rightarrow 27,77$

O hâlde $X_1 = 0,10$ ve $X_2 = 0,50$ iken

Sonuç Cinsiyeti : Kadın'dır.

#NOT : Kaç adet "K" gözlem değeri verilmişse o kadar gözlem değeri sorulıyordur ve verilen "K" değeri kaç ise "ağırlık" hesaplaması o verilen değer kadar en küçüğünden başlanılarak yapılır.

KÜMELEME

* Kümeleme birbirine benzeyen veri parçaları ayırma işlemidir. Bunu yaparken gözlemler arasındaki uzaklıklara odaklanır. İki temel algoritması vardır. En yakın komşu algoritması ve en uzak komşu algoritmasıdır.

1) Hiyerarşik Kümeleme

Hiyerarşik kümeleme yöntemleri kümelerin bir ana küme olarak ele alınması ve sonra aşamalı olarak bir küme biçiminde birleştirilme esasına dayanır.

a) En yakın komşu algoritması

Bu algoritmada tüm gözlem değerleri birer küme kabul edilir. Daha sonra yakınlıklarına göre kümeler birleştirilir.

Aşağıdaki gözlemler dikkate alınarak bu birimleri en yakın komşu algoritmasına göre kümelerine ayırınız.

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

1. ADIM: Uzaklıkların hesaplanmasını Öklid formülüyle yaparız.

$$d(i,j) = \sqrt{\sum (X_{ij} - X_{jk})^2}$$

$$d(1,2) = \sqrt{(4-6)^2 + (2-4)^2} = \sqrt{4+4} = \sqrt{8} = 2,82$$

$$d(1,3) = \sqrt{(4-5)^2 + (2-1)^2} = \sqrt{1+1} = \sqrt{2} = 1,41$$

$$d(1,4) = \sqrt{(4-10)^2 + (2-6)^2} = \sqrt{36+16} = \sqrt{52} = 7,21$$

$$d(1,5) = \sqrt{(4-11)^2 + (2-8)^2} = \sqrt{49+36} = \sqrt{85} = 9,21$$

$d(2,3) = \sqrt{(6-5)^2 + (4-1)^2} = \sqrt{1+9} = \sqrt{10} = 3,16$

$d(2,4) = \sqrt{(6-10)^2 + (4-6)^2} = \sqrt{16+4} = \sqrt{20} = 4,47$

$d(2,5) = \sqrt{(6-11)^2 + (4-8)^2} = \sqrt{25+16} = \sqrt{41} = 6,40$

$d(3,4) = \sqrt{(5-10)^2 + (1-6)^2} = \sqrt{25+25} = \sqrt{50} = 7,07$

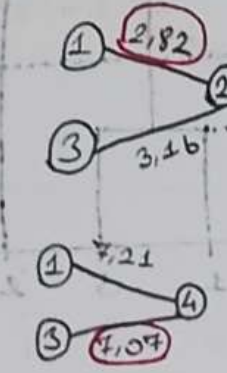
$d(3,5) = \sqrt{(5-11)^2 + (1-8)^2} = \sqrt{36+49} = \sqrt{85} = 9,21$

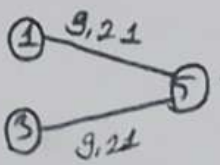
$d(4,5) = \sqrt{(10-11)^2 + (6-8)^2} = \sqrt{1+4} = \sqrt{5} = 2,23$

	✓1	2	✗3	4	5
1	—				
2	2,82	—			
✓3	1,41	3,16	—		
4	7,21	4,47	7,07	—	
5	9,21	6,40	9,21	2,23	—

1 ile 3 ortak kümeyi oluşturur.

	(1,3)	2	✓4	5
(1,3)	—			
2	2,82	—		
4	7,07	4,47	—	
✓5	9,21	6,40	2,23	—

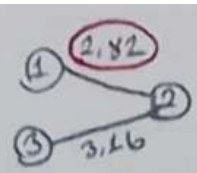


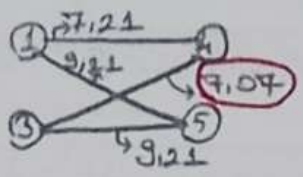


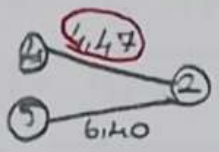
→ ikisi de aynı sonucu verdiği için aynısı yapılabilir.

4 ile 5 ortak kümeyi oluşturur.

	✓(1,3)	2	(4,5)
(1,3)	—		
✓2	2,82	—	
(4,5)	7,07	4,47	—

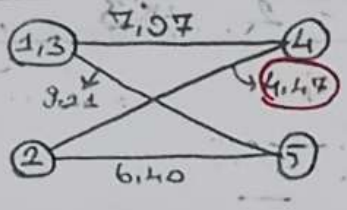




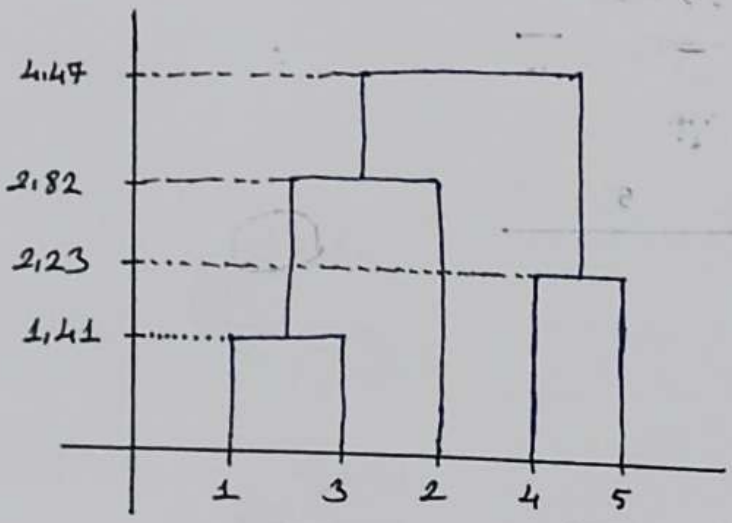


(1,3) ile 2 ortak küme oluşturulur.

	✓(1,3,2)	(4,5)
(1,3,2)	—	
✓(4,5)	4,47	—



(1,3,2) ile (4,5) ortak küme oluşturulur ve işlem bittiği için dendrogram çizilir.



Dendrogram

Sınavda ister. Çizmeyi unutma.

b) En Uzak Komşu Algoritması

Gözlemler	X_1	X_2
1	7	8
2	4	2
3	5	3
4	8	7
5	9	9

$d(1,2) = \sqrt{(7-4)^2+(8-2)^2} = \sqrt{3^2+6^2} = \sqrt{9+36} = \sqrt{45} = 6,70$

$d(1,3) = \sqrt{(7-5)^2+(8-3)^2} = \sqrt{2^2+5^2} = \sqrt{4+25} = \sqrt{29} = 5,38$

$d(1,4) = \sqrt{(7-8)^2+(8-7)^2} = \sqrt{1^2+1^2} = \sqrt{1+1} = \sqrt{2} = 1,41$

$d(1,5) = \sqrt{(7-9)^2+(8-9)^2} = \sqrt{2^2+1^2} = \sqrt{4+1} = \sqrt{5} = 2,23$

$d(2,3) = \sqrt{(4-5)^2+(2-3)^2} = \sqrt{1^2+1^2} = \sqrt{1+1} = \sqrt{2} = 1,41$

$d(2,4) = \sqrt{(4-8)^2+(2-7)^2} = \sqrt{4^2+5^2} = \sqrt{16+25} = \sqrt{41} = 6,40$

$d(2,5) = \sqrt{(4-9)^2+(2-9)^2} = \sqrt{5^2+7^2} = \sqrt{25+49} = \sqrt{74} = 8,60$

$d(3,4) = \sqrt{(5-8)^2+(3-7)^2} = \sqrt{3^2+4^2} = \sqrt{9+16} = \sqrt{25} = 5,00$

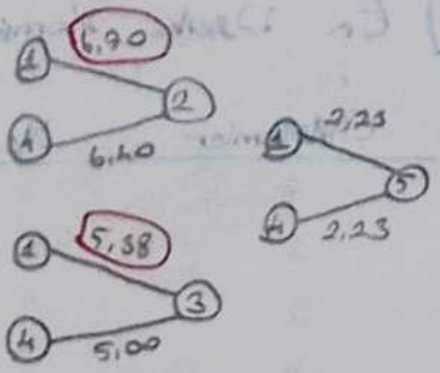
$d(3,5) = \sqrt{(5-9)^2+(3-9)^2} = \sqrt{4^2+6^2} = \sqrt{16+36} = \sqrt{52} = 7,21$

$d(4,5) = \sqrt{(8-9)^2+(7-9)^2} = \sqrt{1^2+2^2} = \sqrt{1+4} = \sqrt{5} = 2,23$

	✓1	2	3	4	5
1	—				
2	6,70	—			
3	5,38	1,41	—		
✓4	1,41	6,40	5,00	—	
5	2,23	8,60	7,21	2,23	—

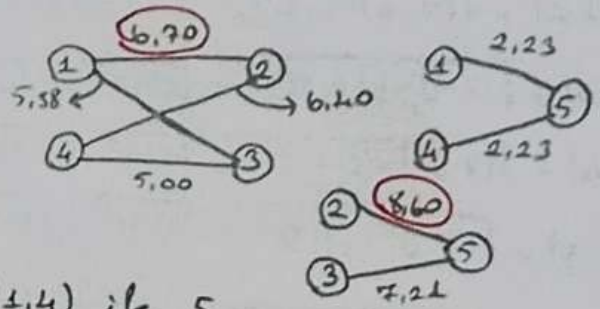
1 ile 4 ortak kümeyi oluştururlar.

	(1,4)	✓2	3	5
(1,4)	—			
2	6,70	—		
✓3	5,38	1,41	—	
5	2,23	8,60	7,21	—



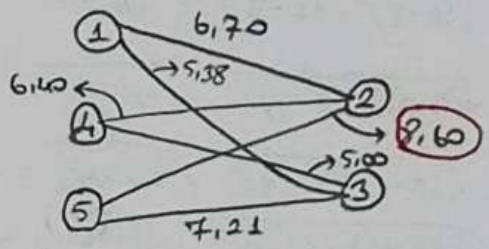
2 ile 3 ortak küme oluştururlar

	✓(1,4)	(2,3)	5
(1,4)	—		
(2,3)	6,70	—	
✓5	2,23	8,60	—



(1,4) ile 5 ortak küme oluştururlar.

	✓(1,4,5)	(2,3)
(1,4,5)	—	
✓(2,3)	8,60	—



(1,4,5) ile (2,3) ortak küme oluştururlar.