KARAR AĞAÇLARI İLE SINIFLANDIRMA

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

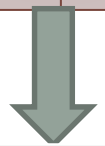
Sınırlayıcı Model

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

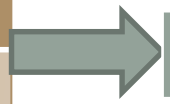
Müşteri	Borç	Gelir	Risk
Halit	Yüksek	Düşük	Kötü
Eser	Düşük	Yüksek	İyi
Selin	Düşük	Düşük	Kötü
Mehmet	Yüksek	Yüksek	Kötü



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;

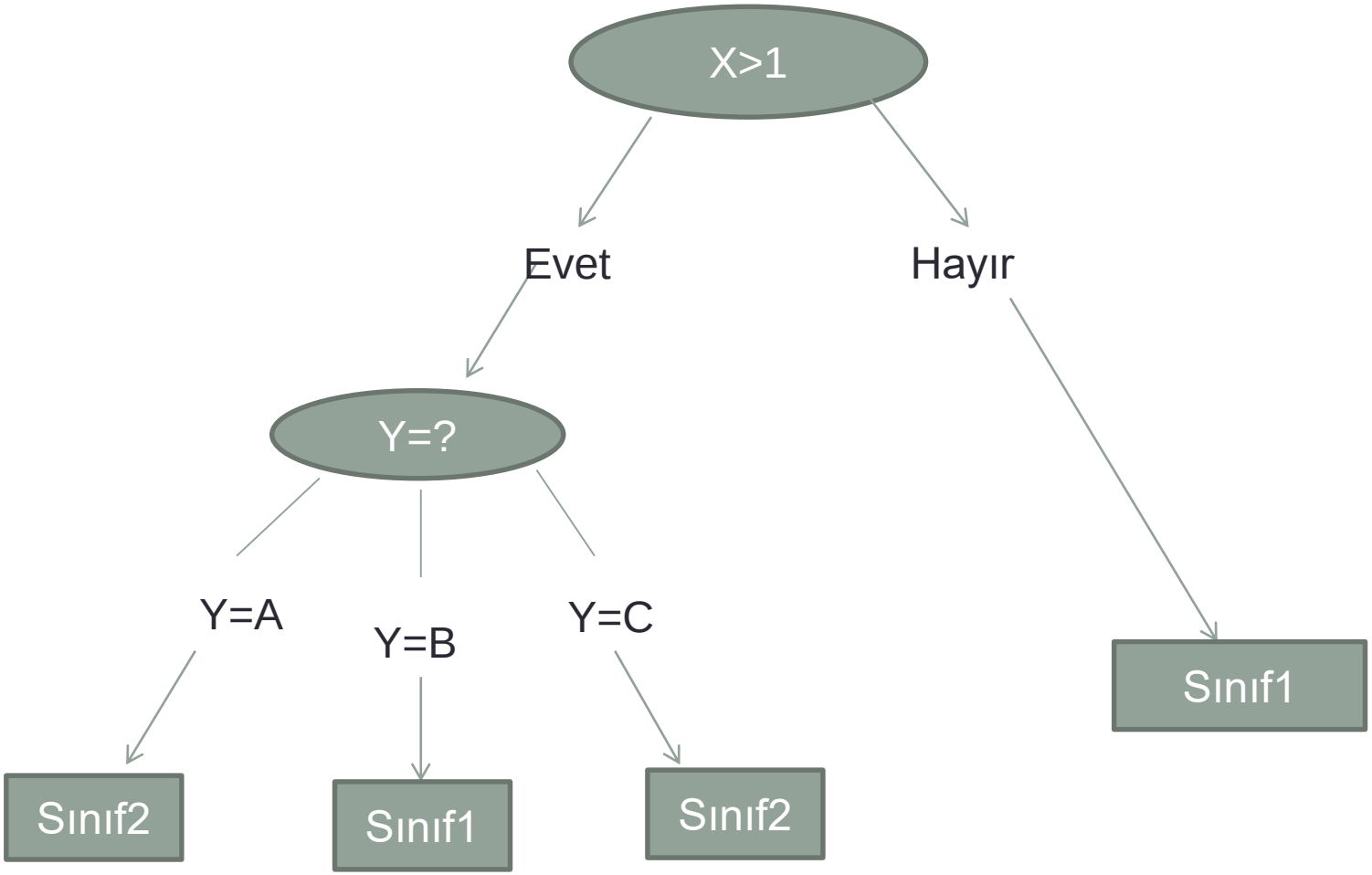
Müşteri	Borç	Gelir	Risk
Hakan	Düşük	YÜKSEK	?



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 $\{C_1, C_2, \dots, C_k\}$ 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

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

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.

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

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;

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

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

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

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

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;

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

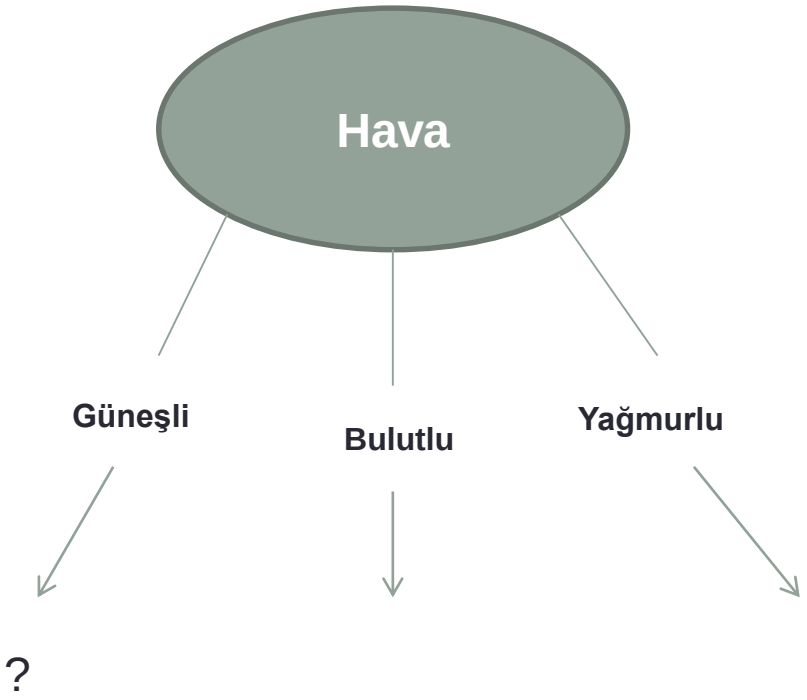
$$= 0.940\text{-}0.892$$

$$= 0.048$$

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

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

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

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

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

$$=0.970-0.951$$

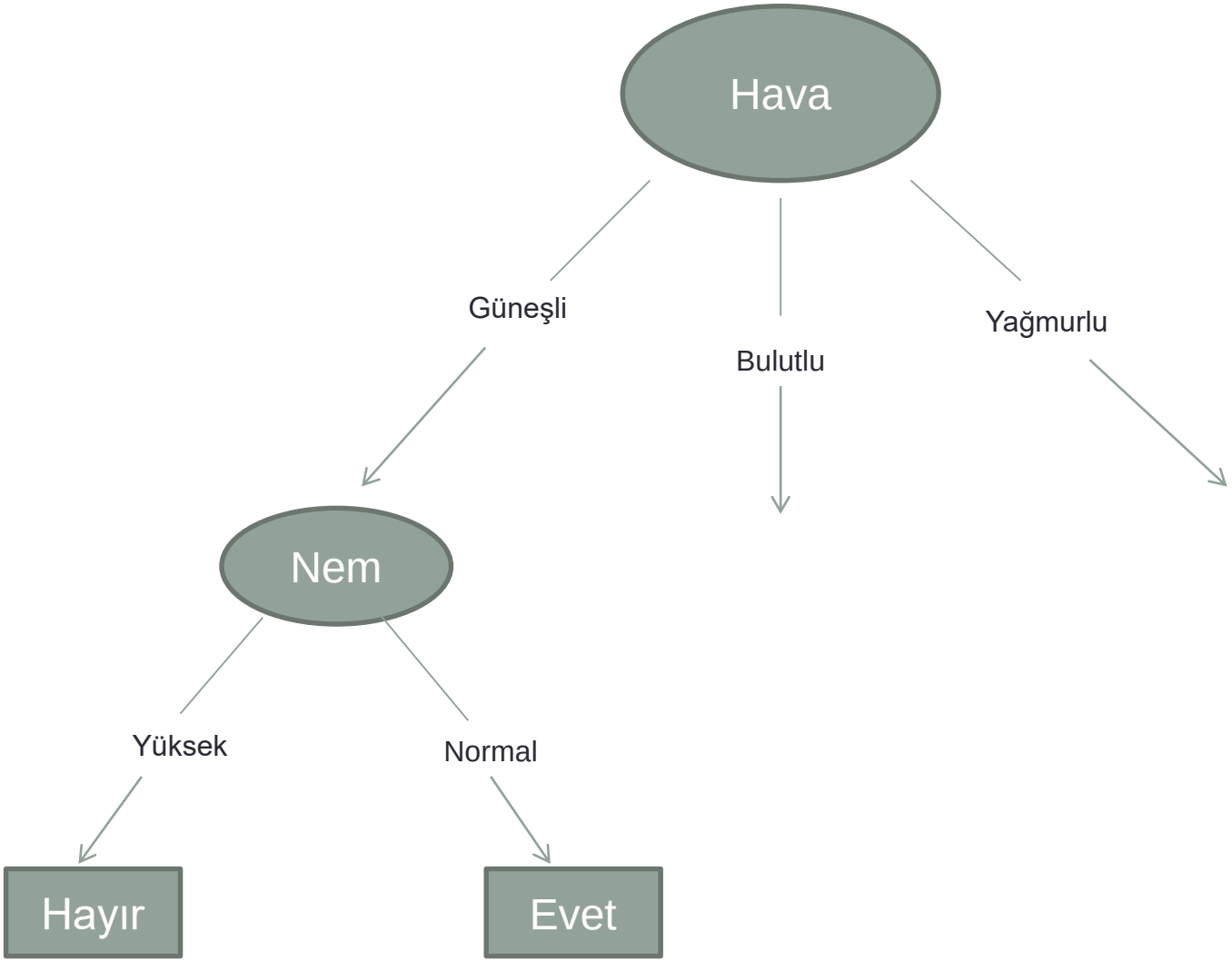
$$=0.019$$

- 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.

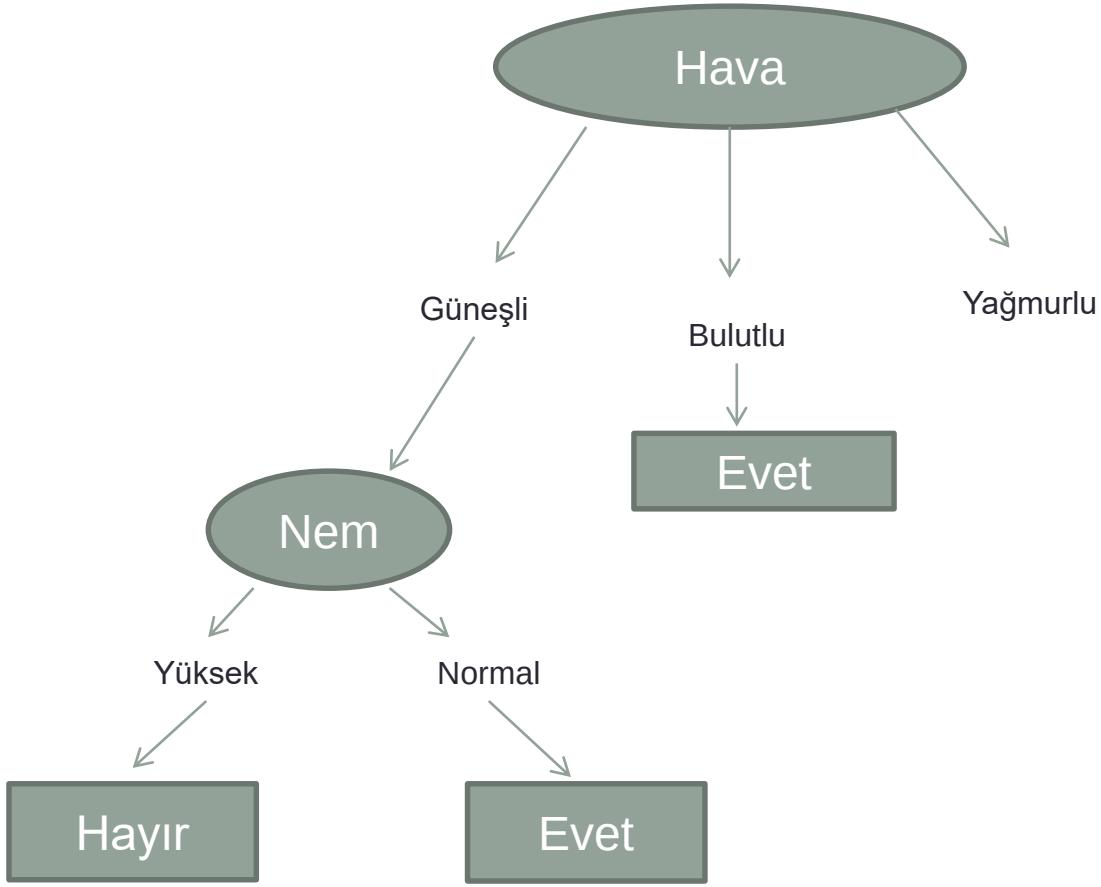
d) HAVA niteliğinin bulutlu değeri için dallanma:

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.

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

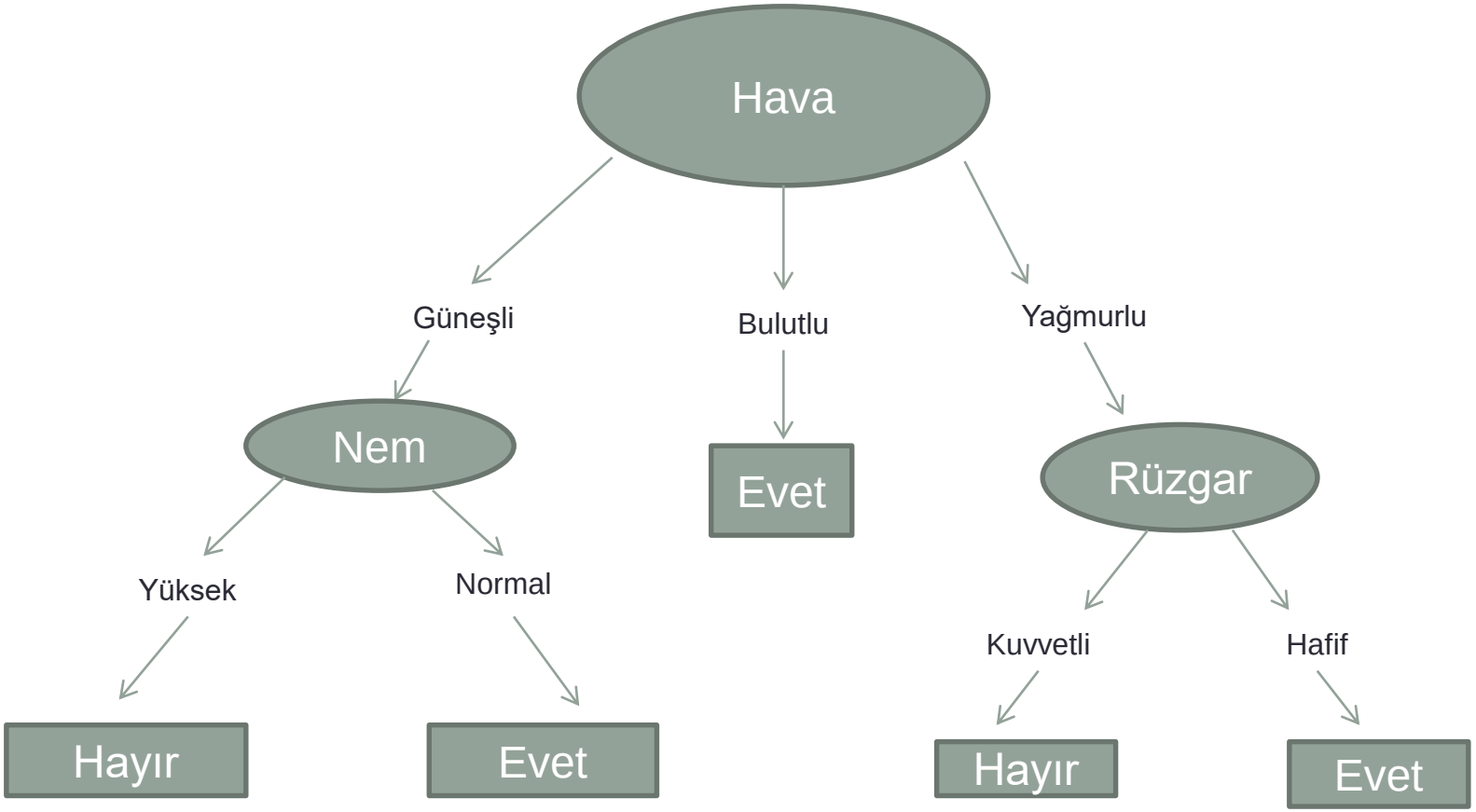
$$\left|R\ddot{U}ZGAR_{g\ddot{u}\ddot{c}l\ddot{u}}\right|=2$$

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;

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

Hesaplamalar yapılırsa;

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

$$= 1.577$$

Kazanç(ISI,OYUN)=0.029 hesaplanmıştı.

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

- 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.

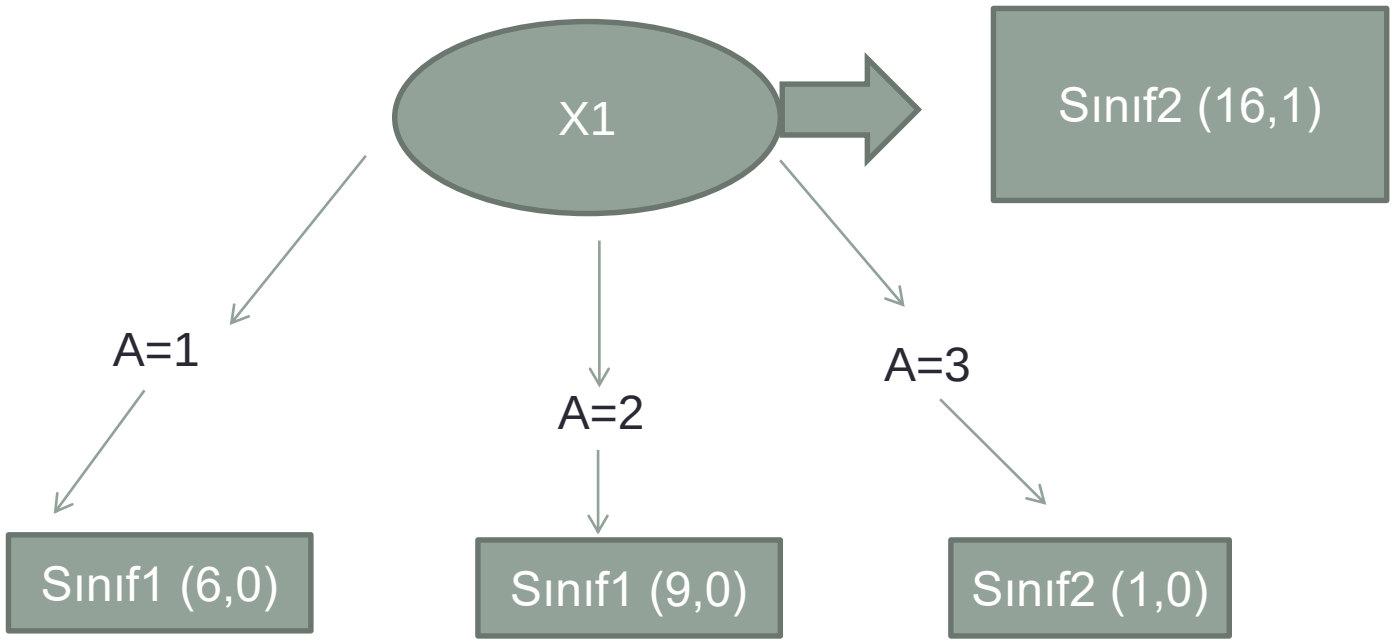
C4.5'de Budama

- 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;

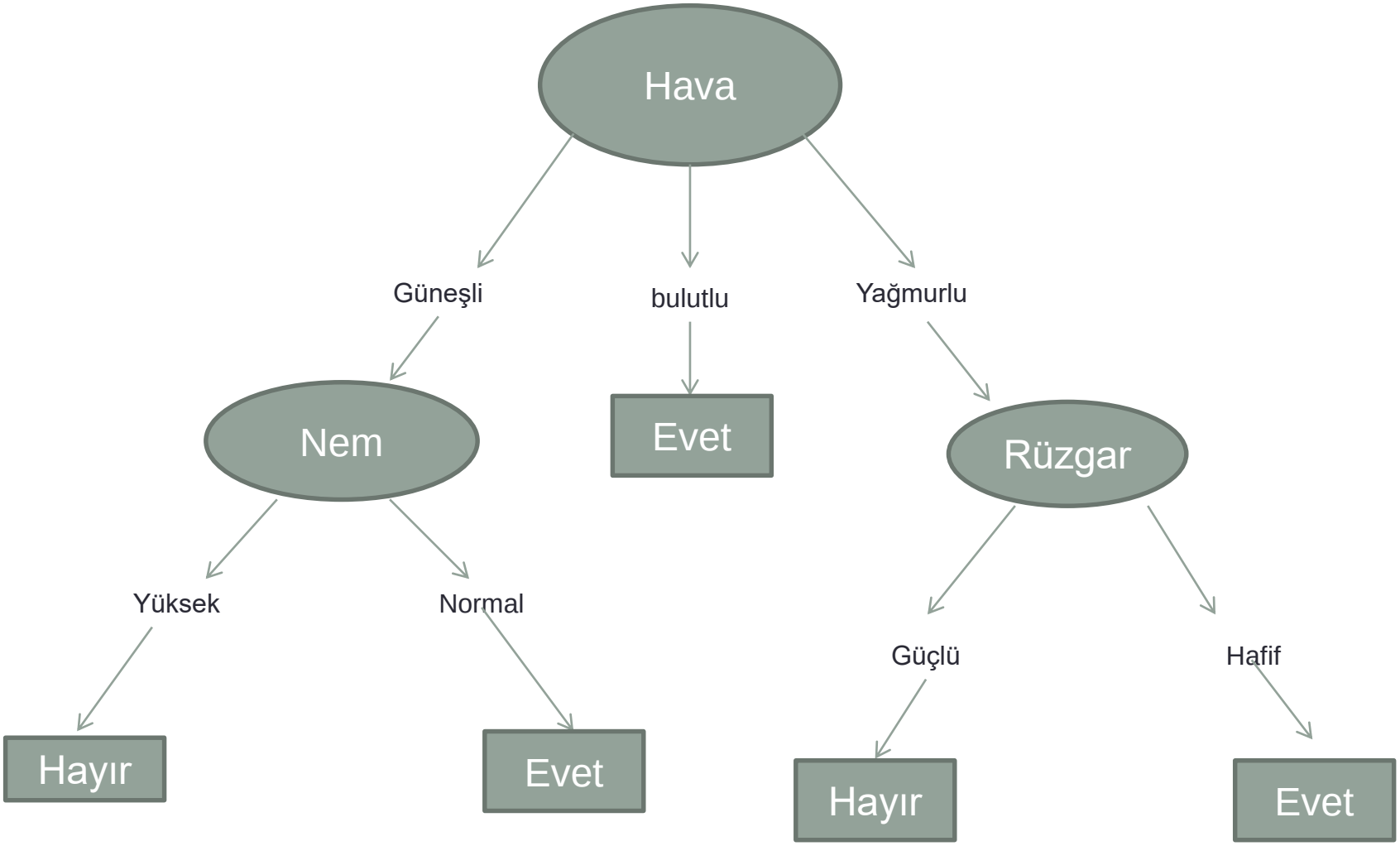
$$PE_{ağaç}=6(0.206)+9(0.143)+1(0.750)=3.257$$

$$PE_{ağaç}=16(0.157)= 2.517$$

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;