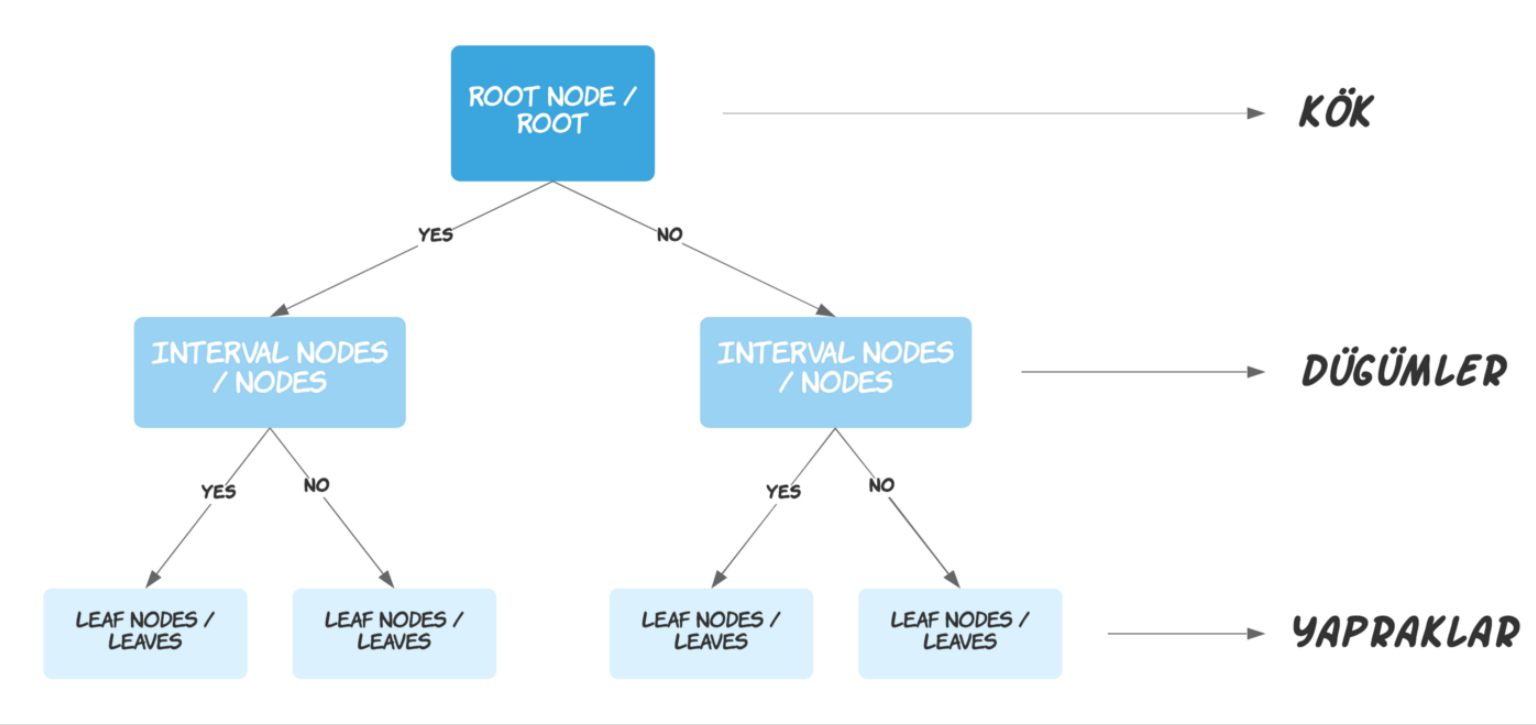
KARAR AĞAÇLARI

Dr.Günay TEMÜR

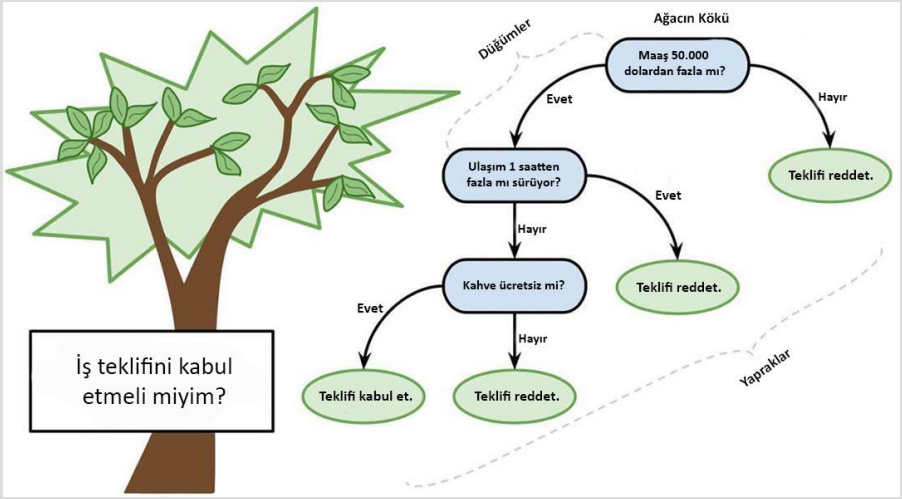
KARAR AĞAÇLARINI TANIYALIM

- Karar ağaçları, Sınıflandırma ve Regresyon problemlerinde kullanılan, ağaç tabanlı algoritmadan biridir. Karmaşık veri setlerinde kullanılabilir.



KARAR AĞAÇLARINI TANIYALIM

- Karar ağaçlarının ilk hücrelerine **kök** (root veya root node) denir. Her bir gözlem kökteki koşula göre “Evet” veya “Hayır” olarak sınıflandırılır.
- Kök hücrelerinin altında **düğümler** (interval nodes veya nodes) bulunur. Her bir gözlem düğümler yardımıyla sınıflandırılır. Düğüm sayısı arttıkça modelin karmaşıklığı da artar.
- Karar ağacının en altında **yapraklar** (leaf nodes veya leaves) bulunur. Yapraklar, bize sonucu verir.



KARAR AĞAÇLARI İLE SINIFLANDIRMA

- Sınıflandırma problemleri için yaygın kullanılan yöntemdir.
- Sınıflandırma doğruluğu diğer öğrenme metotlarına göre çok etkindir.
- Öğrenmiş sınıflandırma modeli ağaç şeklinde gösterilir ve karar ağacı (decision tree) olarak adlandırılır.
- Karar ağaçlarında sınıflandırma yöntemleri 2 çeşittir.
- Bunlar "Entropiye Dayalı Algoritmalar" ve "Sınıflandırma ve Regresyon Ağaçları(CART)"dır. ID3 Algoritması ve C4.5 Algoritması
- "Entropiye dayalı algoritmalar" arasındayken Twoing Algoritması ve Gini Algoritması "Sınıflandırma ve regresyon ağaçları" sınıfındadır.

KARAR AĞACI OLUŞTURMA YÖNTEMLERİ

- Karar ağacı oluşturma yöntemleri genel olarak iki aşamadan oluşur;
- **1.Ağaç oluşturma**
  - En başta bütün öğrenme kümesi örnekleri kökte seçilen niteliklere bağlı olarak örnek yinelemeli olarak bölünüyor.
- **2. Ağaç budama**
  - Öğrenme kümesindeki gürültülü verilerden oluşan ve sınama kümesinde hataya neden olan dalları silme (sınıflandırma başarımını artırır)

ID3 NEDİR?

- Karar Ağaçları yapısını oluşturan birçok algoritmanın en iyilerinden birisi ID3 Algoritması olarak adlandırılır. ID3,Iterative Dichotomiser 3 anlamına gelir.(Tekrarlı ikili ağaç).ID3 algoritması 3 adımı esas alır: Henüz ağaca dahil edilmeyen özellikler ele alınıp entropi (dağınım) değerleri hesaplanır. Entropi değerlerine göre sıralanır ve aralarından en düşük değerli özellik seçilir. Seçilen özelliğin kararı ağaca eklenir.

ENTROPİ NEDİR?

- Rastgeleliği, belirsizliği ve beklenmeyen durumun ortaya çıkma olasılığını gösterir. Eğer örnekler tamamı düzenli / homojen ise entropisi sıfır olur. Eğer değerler birbirine eşit ise entropi 1 olur. Örneğin Futbol Oyna hepsi "Evet" veya "Hayır" olsa entropi sıfır olurdu. Entropi formülasyonu:

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

ÖRNEK

Özellikler				Hedef
Hava Durumu	Sıcaklık	Nem	Rüzgar	Futbol Oyna
Yağmurlu	Sıcak	Yüksek	Yok	Hayır
Yağmurlu	Sıcak	Yüksek	Var	Hayır
Bulutlu	Sıcak	Yüksek	Yok	Evet
Güneşli	Ilık	Yüksek	Yok	Evet
Güneşli	Soğuk	Normal	Yok	Evet
Güneşli	Soğuk	Normal	Var	Hayır
Bulutlu	Soğuk	Normal	Var	Evet
Yağmurlu	Ilık	Yüksek	Yok	Hayır
Yağmurlu	Soğuk	Normal	Yok	Evet
Güneşli	Ilık	Normal	Yok	Evet
Yağmurlu	Ilık	Normal	Yok	Evet
Bulutlu	Ilık	Yüksek	Var	Evet
Bulutlu	Sıcak	Normal	Yok	Evet
Güneşli	Ilık	Yüksek	Var	Hayır

ÖRNEK

Özellikler				Hedef
Hava Durumu	Sıcaklık	Nem	Rüzgar	Futbol Oyna
Yağmurlu	Sıcak	Yüksek	Yok	Hayır
Yağmurlu	Sıcak	Yüksek	Var	Hayır
Bulutlu	Sıcak	Yüksek	Yok	Evet
Güneşli	Ilık	Yüksek	Yok	Evet
Güneşli	Soğuk	Normal	Yok	Evet
Güneşli	Soğuk	Normal	Var	Hayır
Bulutlu	Soğuk	Normal	Var	Evet
Yağmurlu	Ilık	Yüksek	Yok	Hayır
Yağmurlu	Soğuk	Normal	Yok	Evet
Güneşli	Ilık	Normal	Yok	Evet
Yağmurlu	Ilık	Normal	Yok	Evet
Bulutlu	Ilık	Yüksek	Var	Evet
Bulutlu	Sıcak	Normal	Yok	Evet
Güneşli	Ilık	Yüksek	Var	Hayır

Futbol Oyna	
Evet	Hayır
9	5

$$\mathrm{E}(S)\sum_{i=1}^{S} -p_i \log_2(pi)$$

E(FutbolOyna)=***E***(FutbolOyna=Evet)+***E***(FutbolOyna=Hayır)

Evet için olasılık değeri p1 = 9/14=0.643

Hayır için olasılık değeri p2 = 5/14 = 0.357

***E*(FutbolOyna)** = 0.940

toplam

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

Özellikler				Hedef
Hava Durumu	Sıcaklık	Nem	Rüzgar	Futbol Oyna
Yağmurlu	Sıcak	Yüksek	Yok	Hayır
Yağmurlu	Sıcak	Yüksek	Var	Hayır
Bulutlu	Sıcak	Yüksek	Yok	Evet
Güneşli	Ilık	Yüksek	Yok	Evet
Güneşli	Soğuk	Normal	Yok	Evet
Güneşli	Soğuk	Normal	Var	Hayır
Bulutlu	Soğuk	Normal	Var	Evet
Yağmurlu	Ilık	Yüksek	Yok	Hayır
Yağmurlu	Soğuk	Normal	Yok	Evet
Güneşli	Ilık	Normal	Yok	Evet
Yağmurlu	Ilık	Normal	Yok	Evet
Bulutlu	Ilık	Yüksek	Var	Evet
Bulutlu	Sıcak	Normal	Yok	Evet
Güneşli	Ilık	Yüksek	Var	Hayır

		Futbol Oyna		
		Evet	Hayır	Toplam
Hava Durumu	Güneşli	3	2	5
	Bulutlu	4	0	4
	Yağmurlu	2	3	5
				14

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

1 $$|Hava_{güneşli}| = 5$$
$$|Hava_{bulutlu}| = 4$$
$$|Hava_{yağmurlu}| = 5$$

2 $$E(Hava, Oyun) = \frac{5}{14}E(Hava_{güneşli}) + \frac{4}{14}E(Hava_{bulutlu}) + \frac{5}{14}E(Hava_{yağmurlu})$$

3 $$E(Hava_{güneşli}) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.971$$

4 $$E(Hava_{bulutlu}) = -\left(\frac{4}{4}log_2\frac{4}{4} + \frac{0}{4}log_2\frac{0}{4}\right) = 0$$

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

6 $$E(Hava, Oyun) = \frac{5}{14}0.971 + \frac{4}{14}0 + \frac{5}{14}0.971) = 0.694$$

ÖRNEK

Özellikler				Hedef
Hava Durumu	Sıcaklık	Nem	Rüzgar	Futbol Oyna
Yağmurlu	Sıcak	Yüksek	Yok	Hayır
Yağmurlu	Sıcak	Yüksek	Var	Hayır
Bulutlu	Sıcak	Yüksek	Yok	Evet
Güneşli	Ilık	Yüksek	Yok	Evet
Güneşli	Soğuk	Normal	Yok	Evet
Güneşli	Soğuk	Normal	Var	Hayır
Bulutlu	Soğuk	Normal	Var	Evet
Yağmurlu	Ilık	Yüksek	Yok	Hayır
Yağmurlu	Soğuk	Normal	Yok	Evet
Güneşli	Ilık	Normal	Yok	Evet
Yağmurlu	Ilık	Normal	Yok	Evet
Bulutlu	Ilık	Yüksek	Var	Evet
Bulutlu	Sıcak	Normal	Yok	Evet
Güneşli	Ilık	Yüksek	Var	Hayır

		Futbol Oyna		
		Evet	Hayır	Toplam
Nem	Yüksek	3	4	7
	Normal	6	1	7
				14

$$\mathrm{E}(T,X) = \sum_{c \in X} P(c)E(c)$$

E(FutbolOyna, Nem) = **P**(Yüksek)*__E__(3,4)+**P**(Normal)*__E__(6,1)

P(Yüksek) = 0.5, **E**(3,4) = 0.985

P(Normal) = 0.5, **E**(6,1) = 0.592

0.5*0.985 + 0.5*0.592

E(FutbolOyna, Nem) = 0.788

ÖRNEK

Özellikler				Hedef
Hava Durumu	Sıcaklık	Nem	Rüzgar	Futbol Oyna
Yağmurlu	Sıcak	Yüksek	Yok	Hayır
Yağmurlu	Sıcak	Yüksek	Var	Hayır
Bulutlu	Sıcak	Yüksek	Yok	Evet
Güneşli	Ilık	Yüksek	Yok	Evet
Güneşli	Soğuk	Normal	Yok	Evet
Güneşli	Soğuk	Normal	Var	Hayır
Bulutlu	Soğuk	Normal	Var	Evet
Yağmurlu	Ilık	Yüksek	Yok	Hayır
Yağmurlu	Soğuk	Normal	Yok	Evet
Güneşli	Ilık	Normal	Yok	Evet
Yağmurlu	Ilık	Normal	Yok	Evet
Bulutlu	Ilık	Yüksek	Var	Evet
Bulutlu	Sıcak	Normal	Yok	Evet
Güneşli	Ilık	Yüksek	Var	Hayır

		Futbol Oyna		
		Evet	Hayır	Topla
Sıcaklık	Sıcak	2	2	4
	Ilık	4	2	6
	Soğuk	3	1	4
				14

$$\mathrm{E}(T,X) = \sum_{c \in X} P(c)E(c)$$

E(FutbolOyna, Sıcaklık) = **P**(Sıcak)***E**(2,2) + **P**(Ilık)***E**(4,2) + **P**(Soğuk)***E**(3,1)

E(3,2) = 1.000, **P**(Güneşli) = 0.286

E(4,0) = 0.918, **P**(Bulutlu) = 0.429

E(2,3) = 0.811, **P**(Yağmurlu) = 0.286

1.0*0.286 + 0.918*0.429 + 0.811*0.286

E(FutbolOyna, Sıcaklık) = 0.911

ÖRNEK

Özellikler				Hedef
Hava Durumu	Sıcaklık	Nem	Rüzgar	Futbol Oyna
Yağmurlu	Sıcak	Yüksek	Yok	Hayır
Yağmurlu	Sıcak	Yüksek	Var	Hayır
Bulutlu	Sıcak	Yüksek	Yok	Evet
Güneşli	Ilık	Yüksek	Yok	Evet
Güneşli	Soğuk	Normal	Yok	Evet
Güneşli	Soğuk	Normal	Var	Hayır
Bulutlu	Soğuk	Normal	Var	Evet
Yağmurlu	Ilık	Yüksek	Yok	Hayır
Yağmurlu	Soğuk	Normal	Yok	Evet
Güneşli	Ilık	Normal	Yok	Evet
Yağmurlu	Ilık	Normal	Yok	Evet
Bulutlu	Ilık	Yüksek	Var	Evet
Bulutlu	Sıcak	Normal	Yok	Evet
Güneşli	Ilık	Yüksek	Var	Hayır

		Futbol Oyna		
		Evet	Hayır	Toplam
Rüzgar	Yok	6	2	8
	Var	3	3	6
				14

$$\mathrm{E}(T,X) = \sum_{c \in X} P(c)E(c)$$

E(FutbolOyna, Rüzgar) = **P**(Yok)*__E__(6,2)+**P**(Var)*__E__(3,3)

E(3,4) = 0.811, **P**(Yüksek) = 0.571

E(6,1) = 1.000, **P**(Normal) = 0.429

0.811*0.571 + 1.0*0.429

E(FutbolOyna, Rüzgar) = 0.892

INFORMATİON GAİN (BİLGİ KAZANIMI)

- Bilgi kazanımı, bir veri setini bir özellik üzerinde böldükten (Örneğin **E**(FutbolOyna, HavaDurumu)) sonra tüm entropiden (**E**(FutbolOyna)) çıkarmaya dayanır. Entropinin küçük değer içermesi durumunda özelliğin önemi Decision Tree algoritması ID3 için artmaktadır. Diğer taraftan 1'e yaklaştıkça özelliğinin önemi azalır. Ancak information gain'de olay tam tersidir ve bu açıdan entropinin tersi gibi düşünülebilir. Decision Tree inşa edilirken en yüksek değerleri information gain'e sahip özellik seçilir.

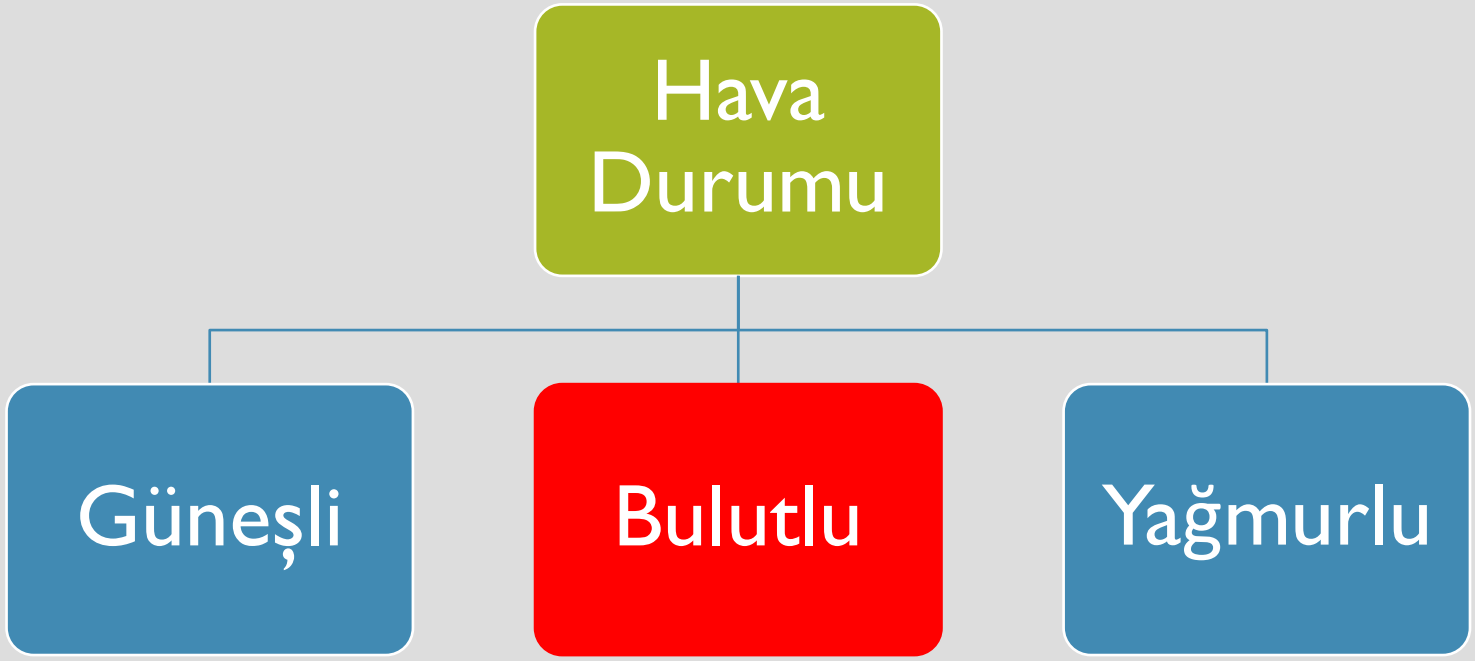
$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

INFORMATİON GAİN (BİLGİ KAZANIMI)

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

- **Gain**(FutbolOyna, HavaDurumu) = **E**(FutbolOyna) – **E**(FutbolOyna, HavaDurumu)
  - **Gain**(FutbolOyna, HavaDurumu) = 0.940 – 0.694 = 0.247 *
    - **Gain**(FutbolOyna, Nem) = 0.940 – 0.788 = 0.152
    - **Gain**(FutbolOyna, Sıcaklık) = 0.940 – 0.911 = 0.029
    - **Gain**(FutbolOyna, Rüzgar) = 0.940 – 0.892 = 0.048

İLK DALLANMA



ADIM 2: HAVA NİTELİĞİNİN "GÜNEŞLİ" DEĞERİ İÇİN DALLANMA

- İkinci aşamada Hava Durumu Güneşli olan durumlar seçilir. Güneşli olduğunda oynama ve oynama durumları vardır. Bu durumda tekrar information Gain hesaplanır ve

Özellikler				Hedef
Hava Durumu	Sıcaklık	Nem	Rüzgar	Futbol Oyna
Güneşli	Ilık	Yüksek	Yok	Evet
Güneşli	Soğuk	Normal	Yok	Evet
Güneşli	Soğuk	Normal	Var	Hayır
Güneşli	Ilık	Normal	Yok	Evet
Güneşli	Ilık	Yüksek	Var	Hayır

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

ADIM 2: HAVA NİTELİĞİNİN "GÜNEŞLİ" DEĞERİ İÇİN DALLANMA

Özellikler			Hedef
Sıcaklık	Nem	Rüzgar	Futbol Oyna
Ilık	Yüksek	Yok	Evet
Soğuk	Normal	Yok	Evet
Soğuk	Normal	Var	Hayır
Ilık	Normal	Yok	Evet
Ilık	Yüksek	Var	Hayır

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

1. $E(Sıcaklık_{ılık}) = -\left(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}\right) = 0.918$
2. $E(Sıcaklık_{Soğuk}) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1$
3. $E(Nem_{yüksek}) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1$
4. $E(Nem_{normal}) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1$
5. $E(Rüzgar_{yok}) = -\left(\frac{3}{3}log_2\frac{3}{3}\right) = 0$
6. $E(Rüzgar_{var}) = -\left(\frac{2}{2}log_2\frac{2}{2}\right) = 0$

ADIM 2: GAİN

1. $\mathrm{P}(Sıcaklık) = \frac{2}{5}(1) + \frac{3}{5}(0.918) = 0.9508$
2. $\mathrm{P}(Nem) = \frac{2}{5}(1) + \frac{3}{5}(1) = 1$
3. $\mathrm{P}(\mathrm{R}üzgar) = \frac{2}{5}(0) + \frac{3}{5}(0) = 0$

Entropi

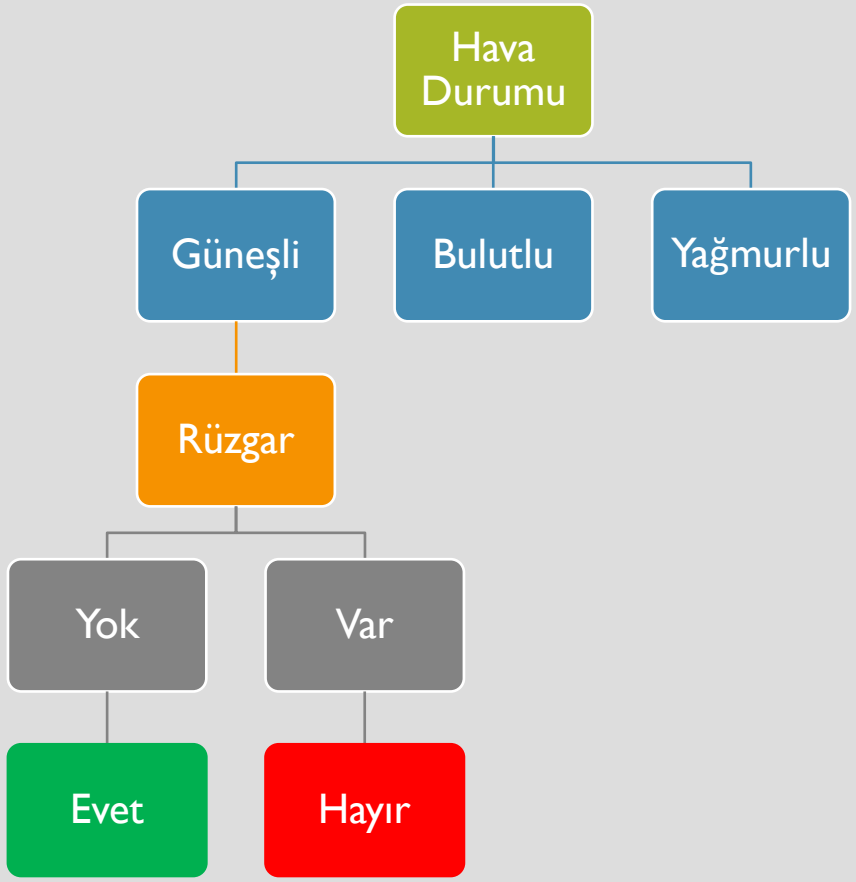
$$E(oyun) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.970$$

Kazançlar

$$Gain(oyun, sıcaklık) = 0.970 - 0.9508 = 0.0192$$
$$Gain(oyun, nem) = 0.970 - 1 = -0.03$$
$$Gain(oyun, rüzgar) = 0.970 - 0 = 0.970 *$$

İKİNCİ DALLANMA

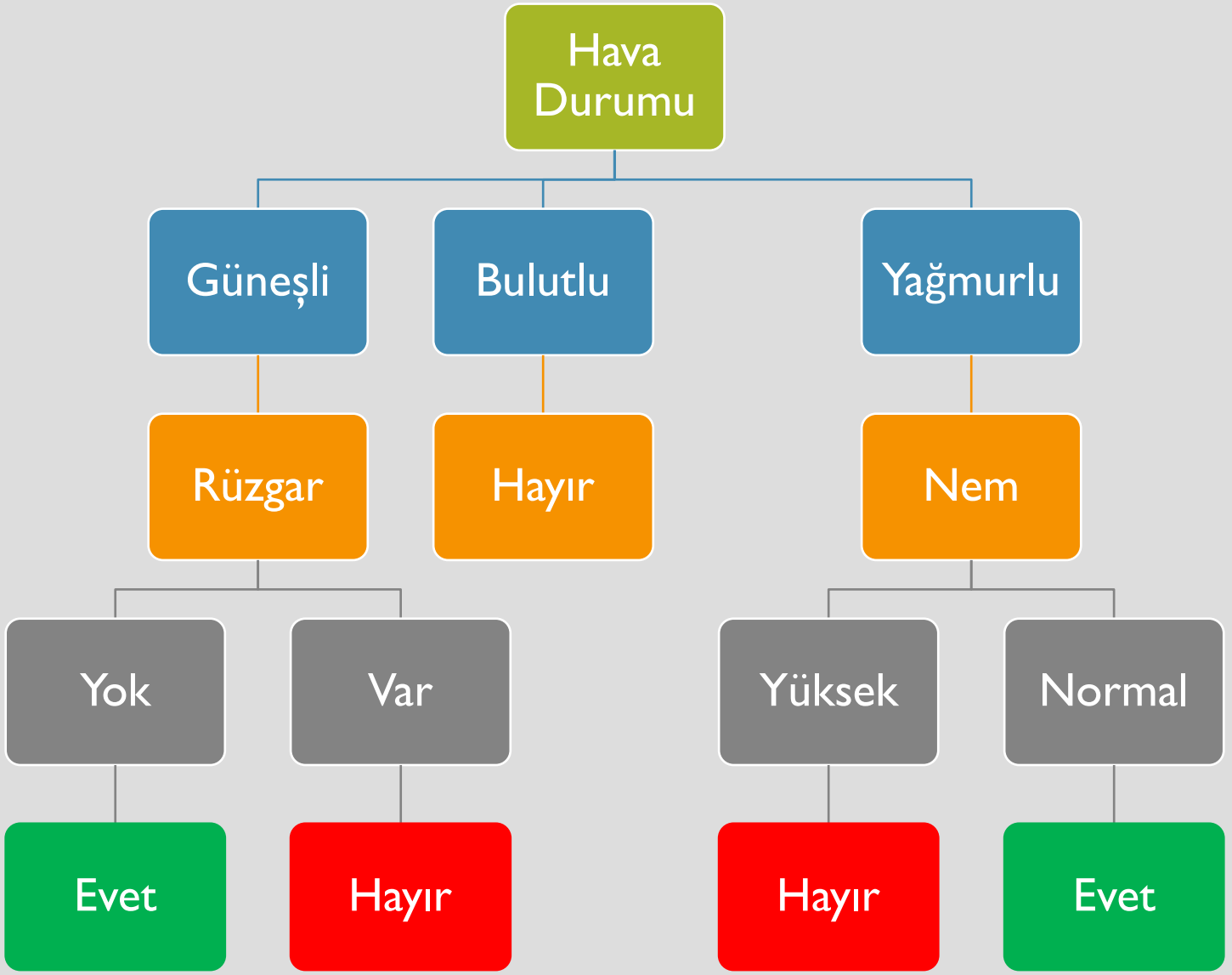
	Hedef
Rüzgar	Futbol Oyna
Yok	Evet
Yok	Evet
Var	Hayır
Yok	Evet
Var	Hayır



DİĞER AŞAMA

- Adım 3 sizin tarafınızdan yapılacaktır.

KARAR AĞACI


Hava Durumu
Güneşli
Bulutlu
Yağmurlu
Rüzgar
Hayır
Nem
Yok
Var
Yüksek
Normal
Evet
Hayır
Hayır
Evet

ÖDEV

- En iyi bildiğiniz bildirdiğiniz program ile karar ağacını kodlayınız.
- Mümkün ise görsel programlama

Özellikler				Hedef
Hava Durumu	Sıcaklık	Nem	Rüzgar	Futbol Oyna

BİTTİ ☺