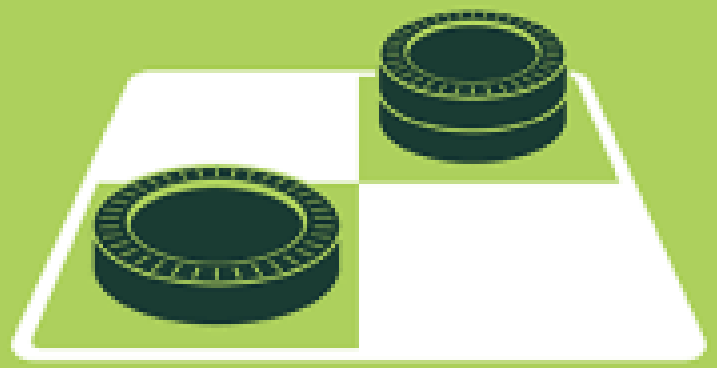
Karar Ağaçları ile Sınıflandırma

Dr. Caner Erden cerden@sakarya.edu.tr

Sakarya Üniversitesi, Endüstri Mühendisliği Bölümü

ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



MACHINE LEARNING

Machine learning begins to flourish.

DEEP LEARNING

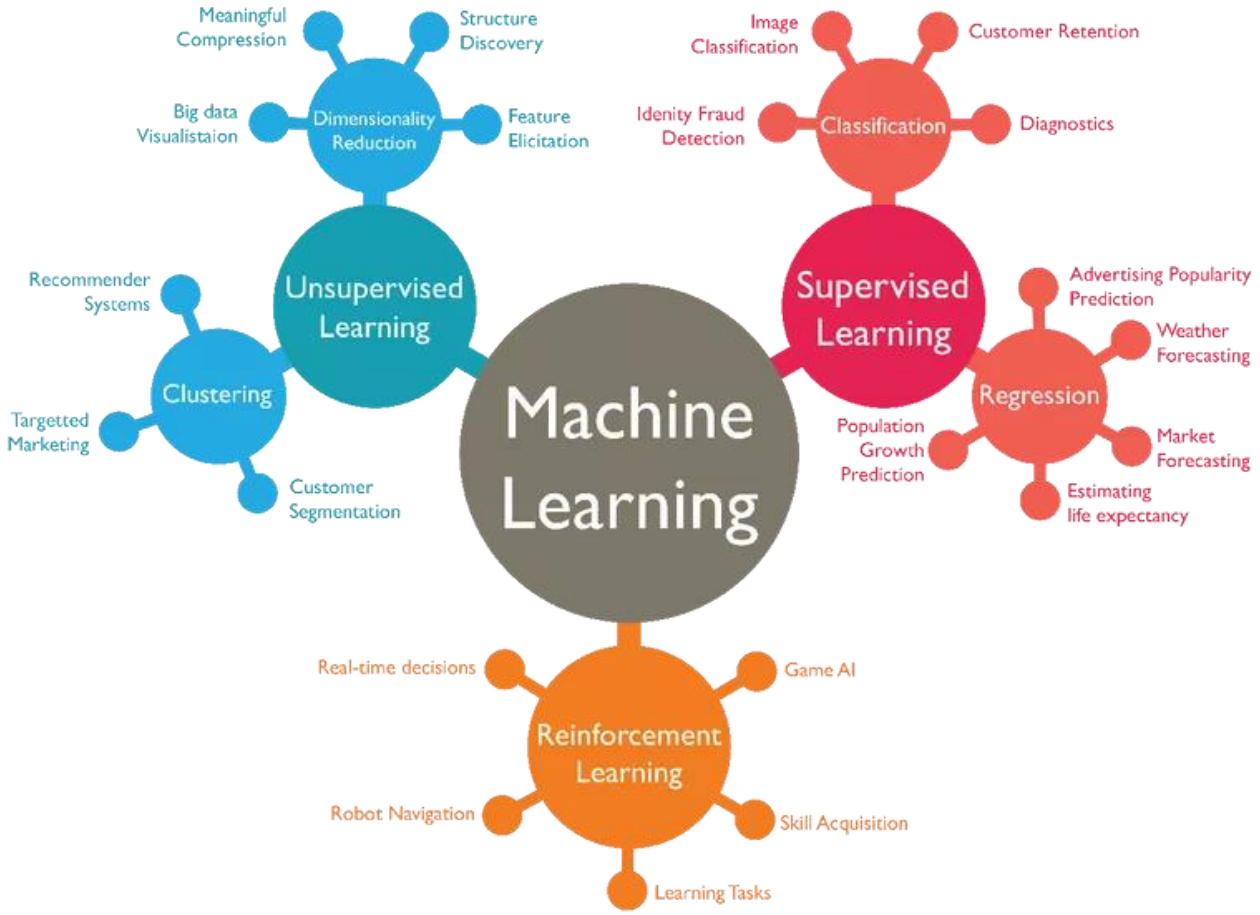
Deep learning breakthroughs drive AI boom.

1950's 1960's 1970's 1980's 1990's 2000's 2010's

Makine Öğrenmesi

1. Gözetimli(Supervised)
2. Gözetimsiz(Unsupervised)
3. Takviyeli(Reinforcement)

Algoritmaların Kullanım Alanları



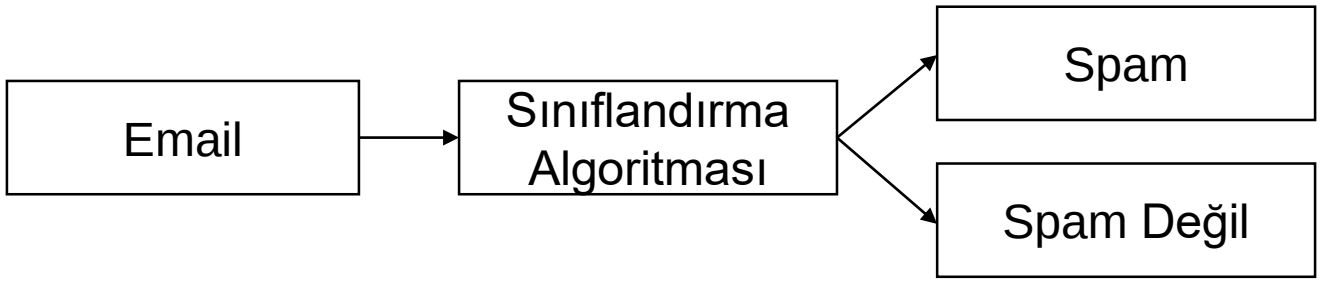
Sınıflandırma

- Verilen bir eğitim setindeki bağımlı değişkenlerinden (x) yola çıkılarak bir sınıf etiketine(y) atanması işlemi.

- Önceden belirli sınıflara atanacak yeni özellik verileri belirlenmeye çalışılır.

Sınıflandırma Örnekleri

Görev	Özellik Seti (x)	Sınıf etiketi (y)
Email mesajlarının kategorizasyonu	Email mesajından alınan metinler	SPAM ya da SPAM değil
Çiçek Türünün Belirlenmesi	Çiçeğin petal ve sepal uzunlukları ve genişlikleri	Şekillerine göre sınıflanan zambak çiçekleri
…	…	…



Canlı türleri sınıflandırması

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has legs	Hibernates	Class Label
Human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

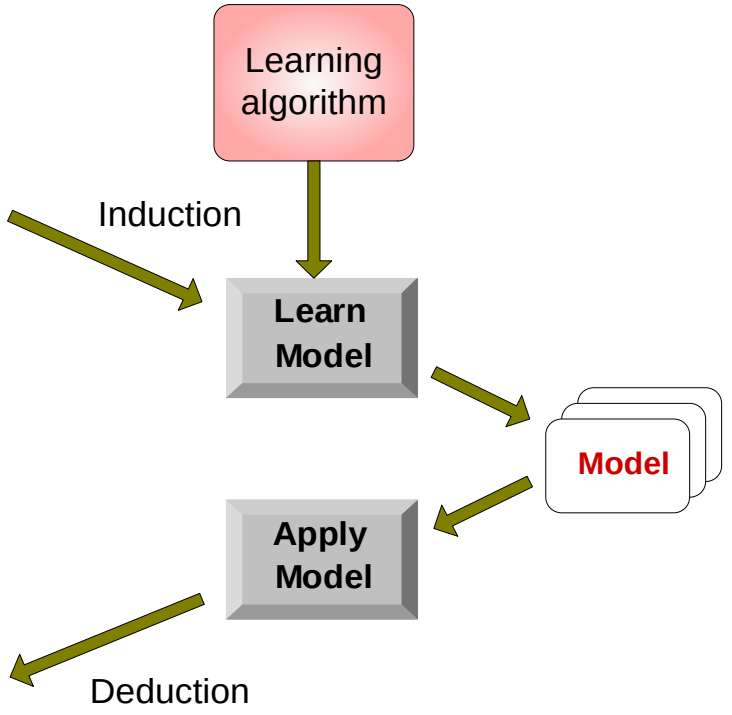
Sınıflandırma Modeli

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Sınıflandırma Teknikleri

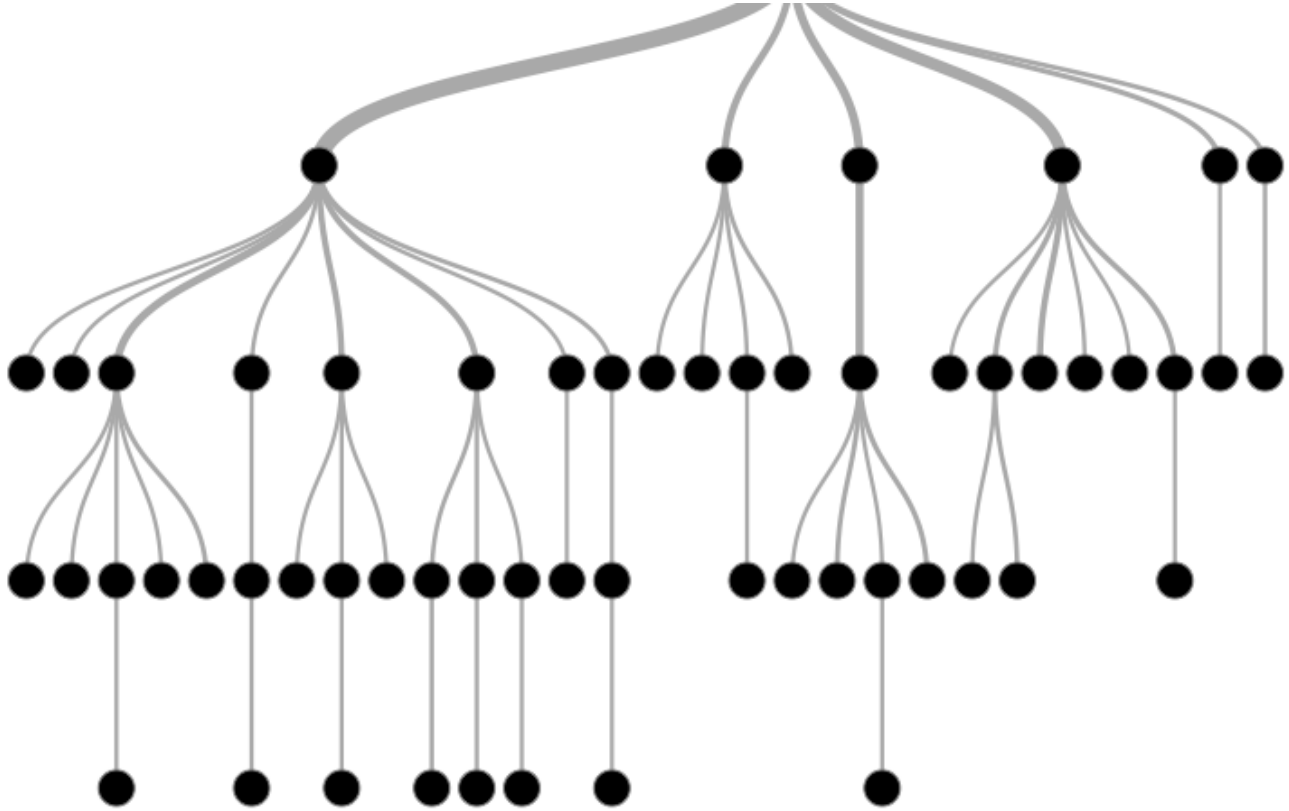
- Temel Sınıflandırıcılar
  - **Karar Ağaçları**
  - **Kural Tabanlı**
  - **En yakın komşu**
  - **Yapay Sinir Ağları**
  - Derin Öğrenme
  - **Naive Bayes**
  - **Destek Vektör Makineleri**

Karar Ağaçlarıyla Sınıflandırma

Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir.

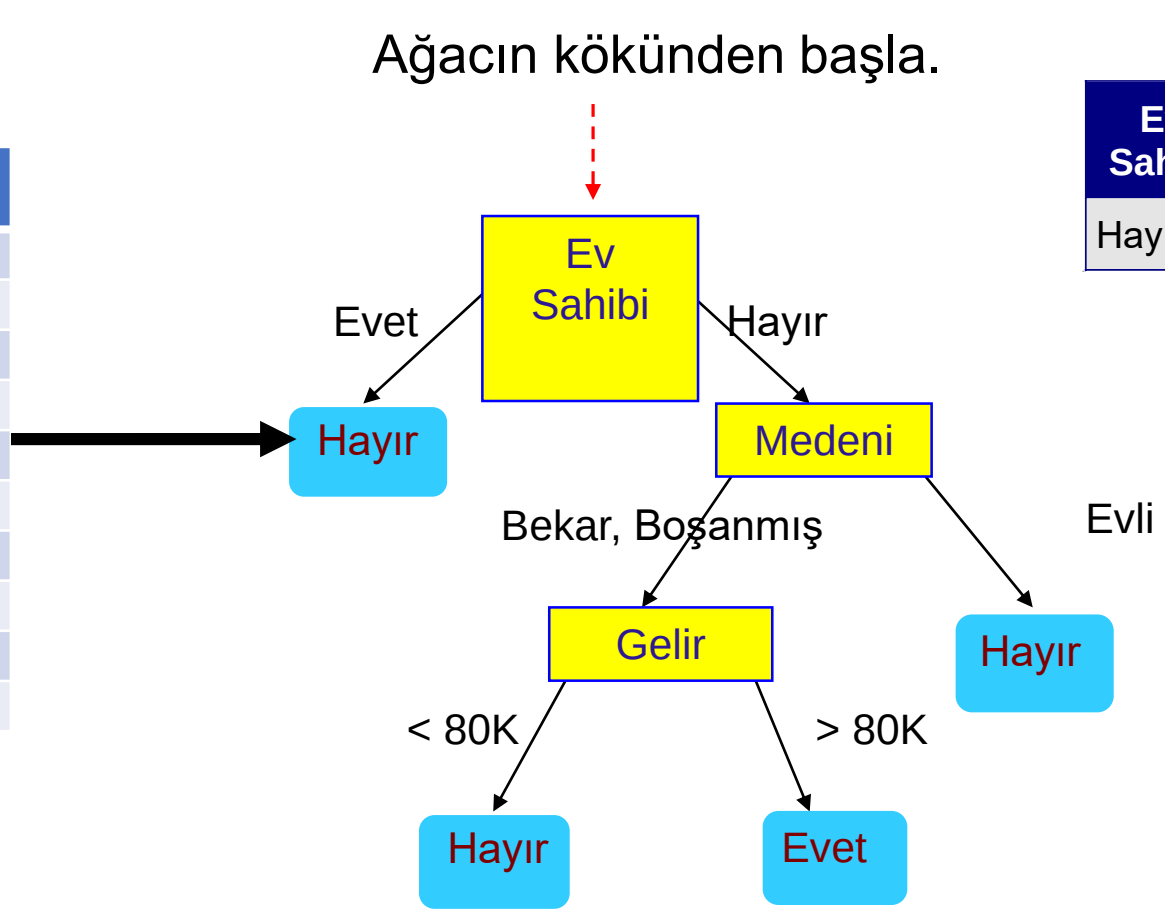
Dallar ve yapraklar ağaç yapısının elemanlarıdır.

En son yapı ”yaprak”, en üst yapı ”kök” ve bunların arasında kalan yapılar ise ”dal” olarak isimlendirilir.



Karar Ağaçlarının Uygulanması

ID	Ev Sahibi	Medeni Hal	Gelir	Borç
1	Evet	Bekar	125K	Hayır
2	Hayır	Evli	100K	Hayır
3	Hayır	Bekar	70K	Hayır
4	Evet	Evli	120K	Hayır
5	Hayır	Dul	95K	Evet
6	Hayır	Evli	60K	Hayır
7	Evet	Dul	220K	Hayır
8	Hayır	Bekar	85K	Evet
9	Hayır	Evli	75K	Hayır
10	Hayır	Bekar	90K	Evet

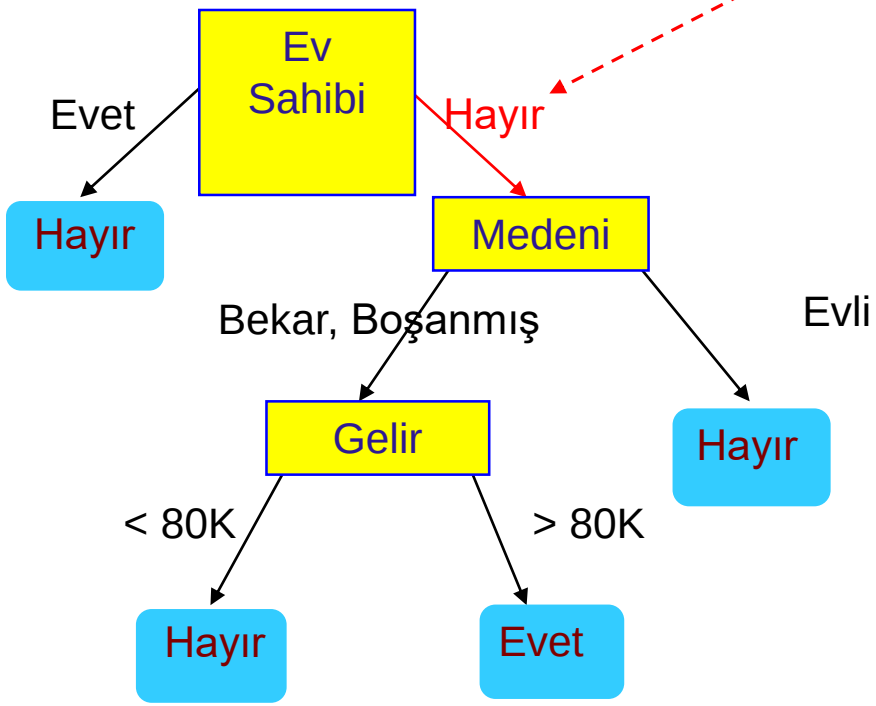


Test Verisi

Ev Sahibi	Medeni Hal	Gelir	Borç
Hayır	Evli	80K	?

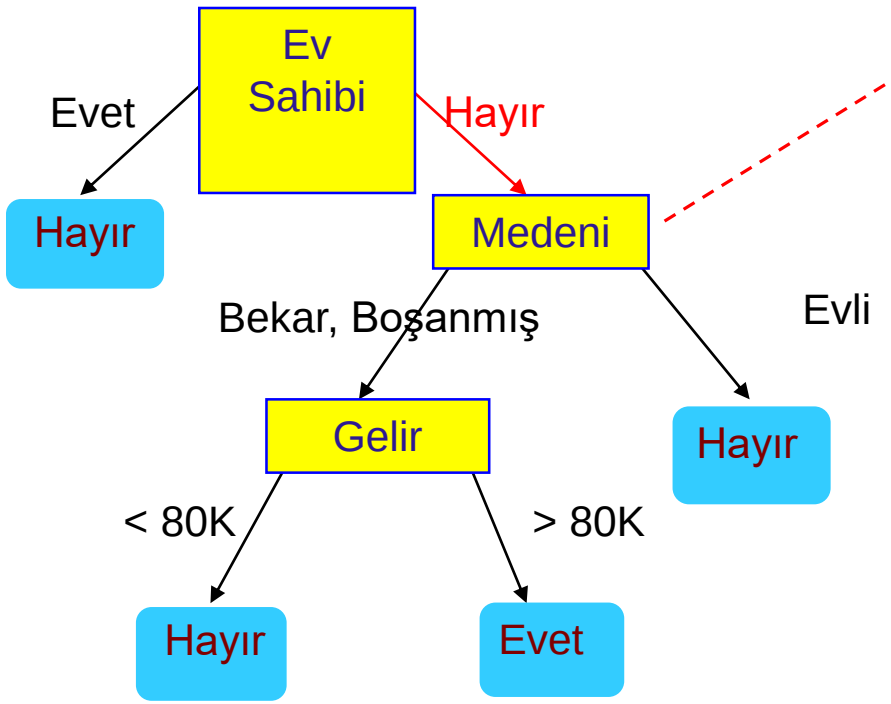
Test Veri

Ev Sahibi	Medeni Hal	Gelir	Borç
Hayır	Evli	80K	?



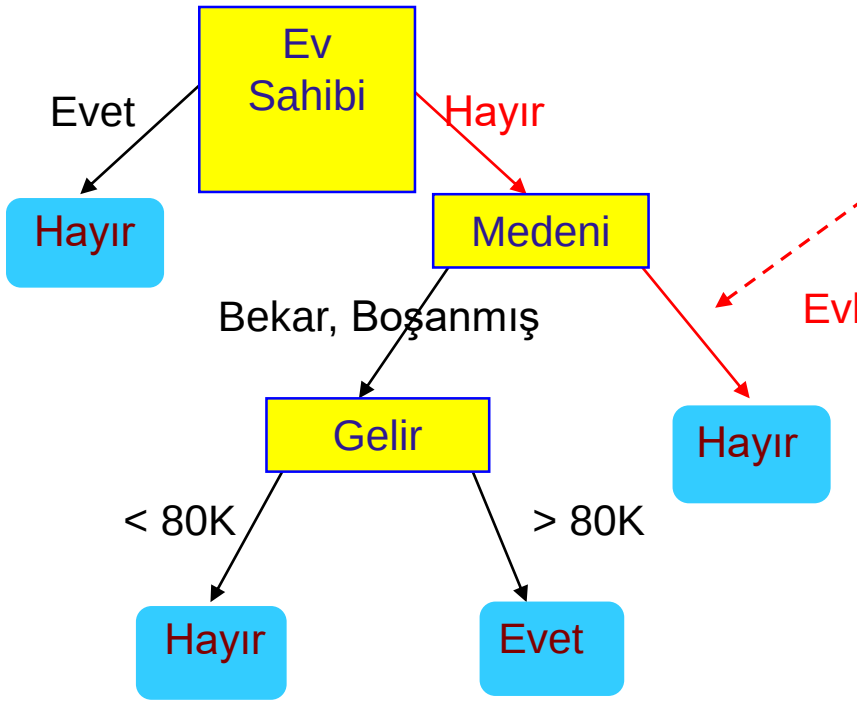
Test Veri

Ev Sahibi	Medeni Hal	Gelir	Borç
Hayır	Evli	80K	?



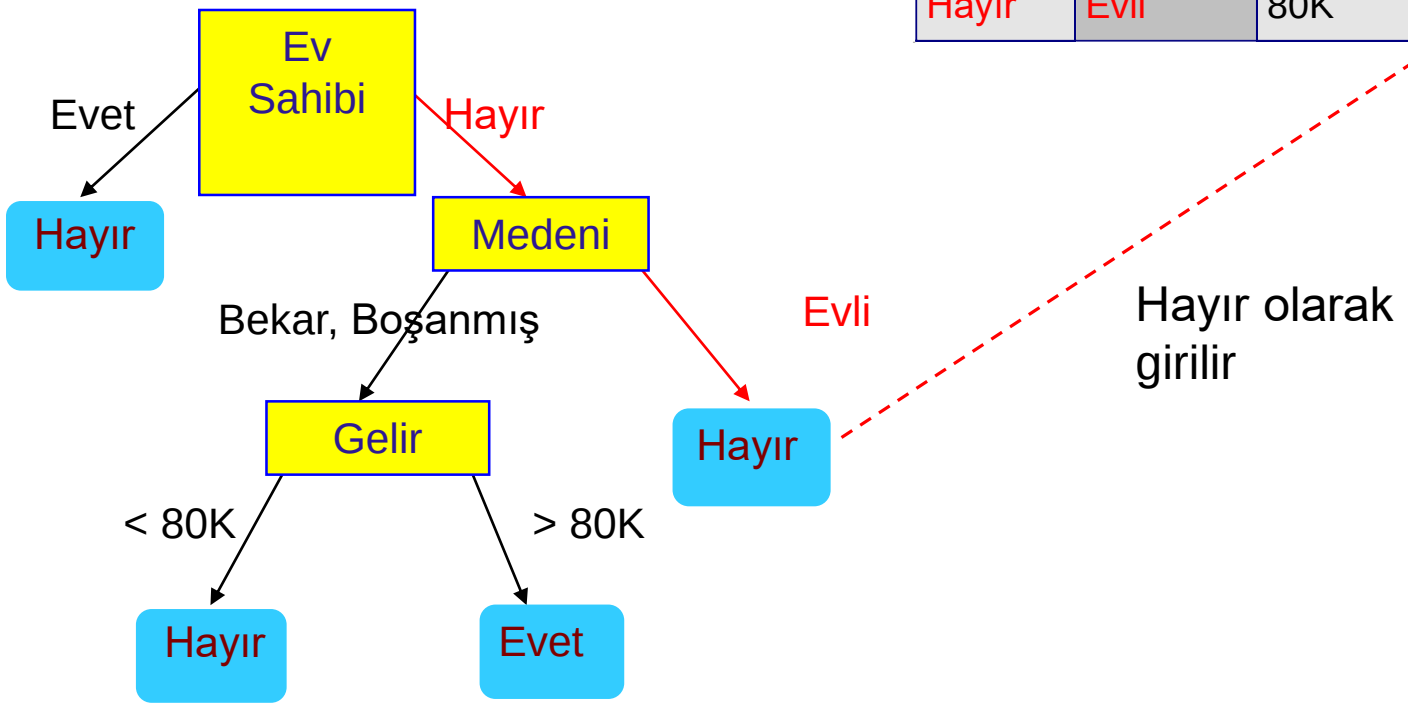
Test Veri

Ev Sahibi	Medeni Hal	Gelir	Borç
Hayır	Evli	80K	?



Test Veri

Ev Sahibi	Medeni Hal	Gelir	Borç
Hayır	Evli	80K	?



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

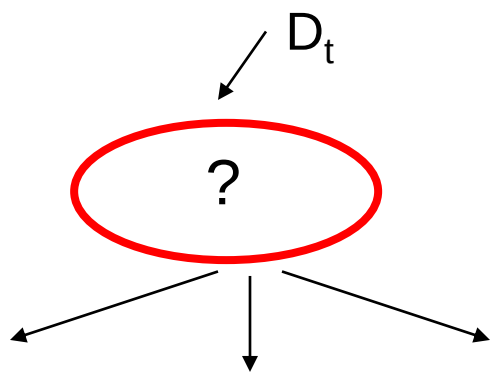
Karar Ağacı Algoritmaları

- Örnek Algoritmalar:
  - **Hunt's Algoritması**
  - CART
  - **ID3**, C4.5
  - SLIQ,SPRINT

Hunt Algoritması

- Dt t düğümüne ulaşan bir eğitim seti olsun.
- Genel prosedür:
  - Eğer Dt bir yt sınıfına sahipse t düğümüne yt etiketi verilir.
  - Eğer Dt birden fazla sınıfa sahipse veriyi küçük sınıflara bölecek şekilde algoritmayı devam ettir. En son tek bir sınıfa ait olana kadar algoritma devam eder.

ID	Ev Sahibi	Medeni Hal	Gelir	Borç Alma
1	Evet	Bekar	125K	Hayır
2	Hayır	Evli	100K	Hayır
3	Hayır	Bekar	70K	Hayır
4	Evet	Evli	120K	Hayır
5	Hayır	Dul	95K	Evet
6	Hayır	Evli	60K	Hayır
7	Evet	Dul	220K	Hayır
8	Hayır	Bekar	85K	Evet
9	Hayır	Evli	75K	Hayır
10	Hayır	Bekar	90K	Evet

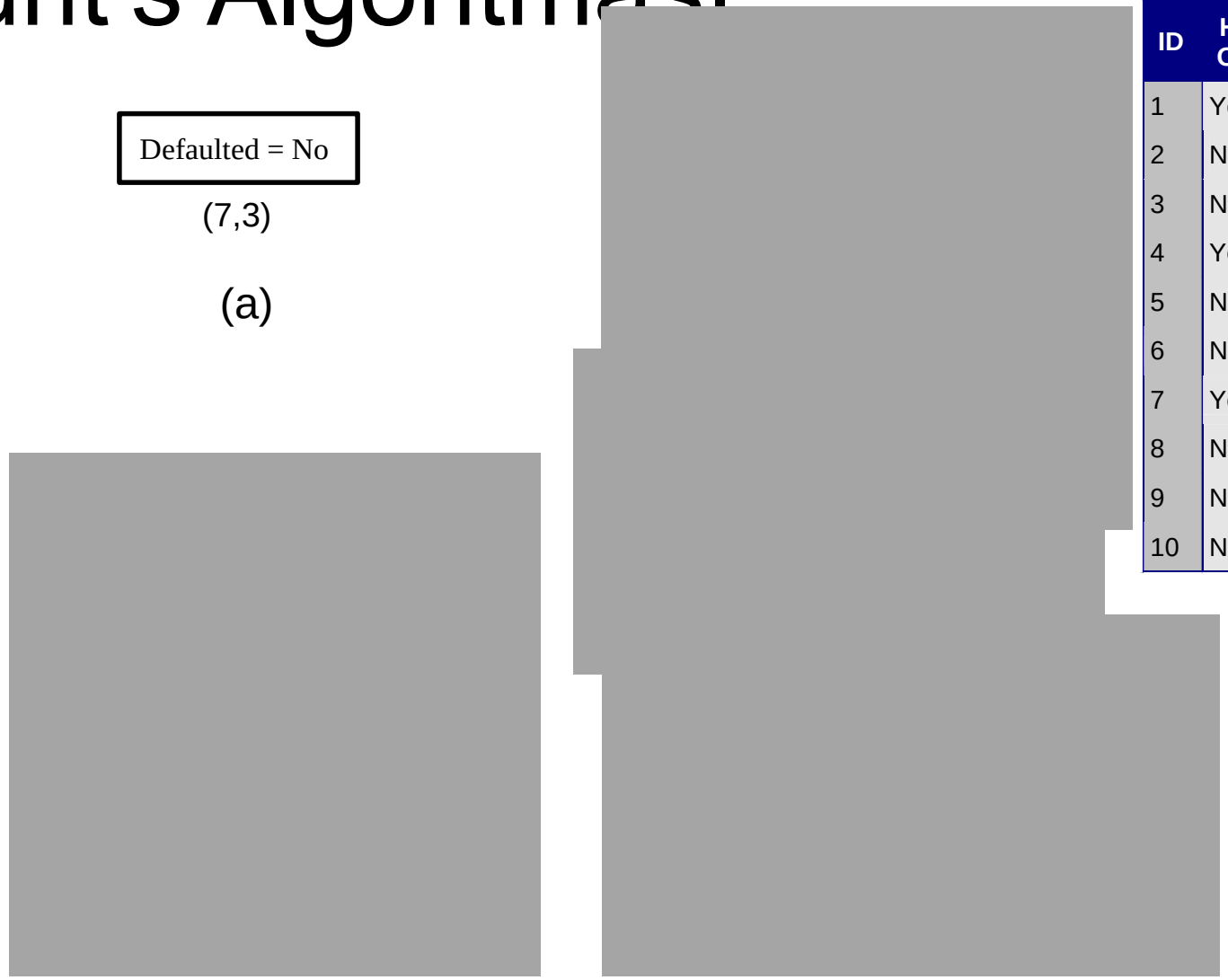


Hunt’s Algoritması

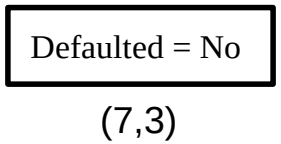
Defaulted = No

(7,3)

(a)

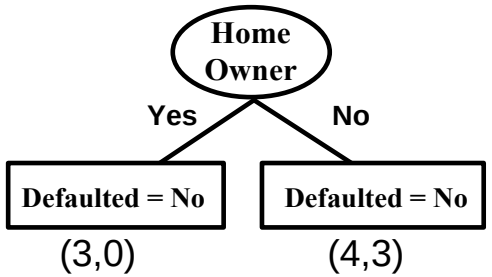


ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



(7,3)

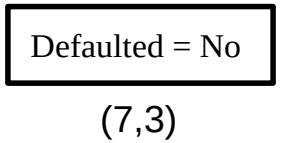
(a)



(3,0) (4,3)

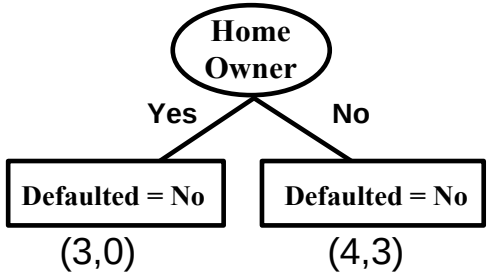
(b)

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



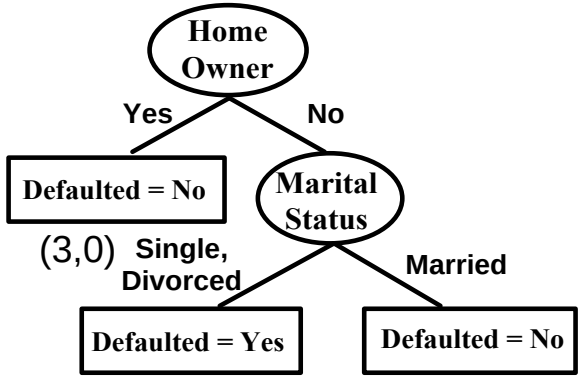
(7,3)

(a)



(3,0) (4,3)

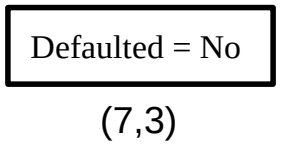
(b)



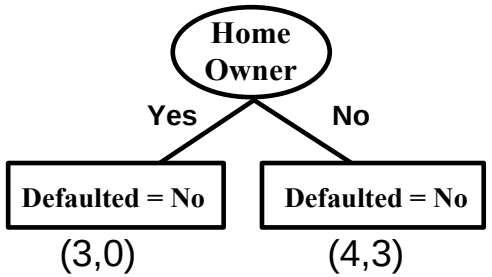
(1,3) (3,0)

(c)

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

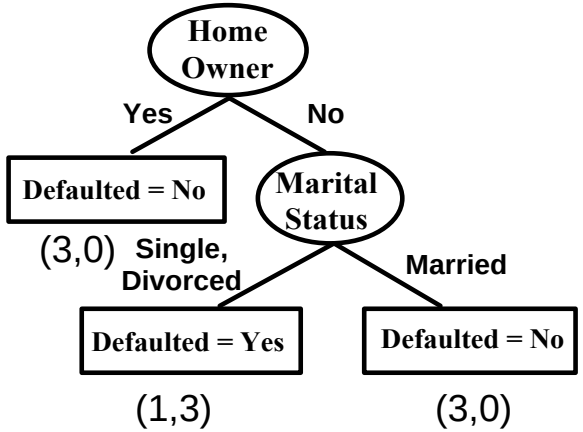


(a)

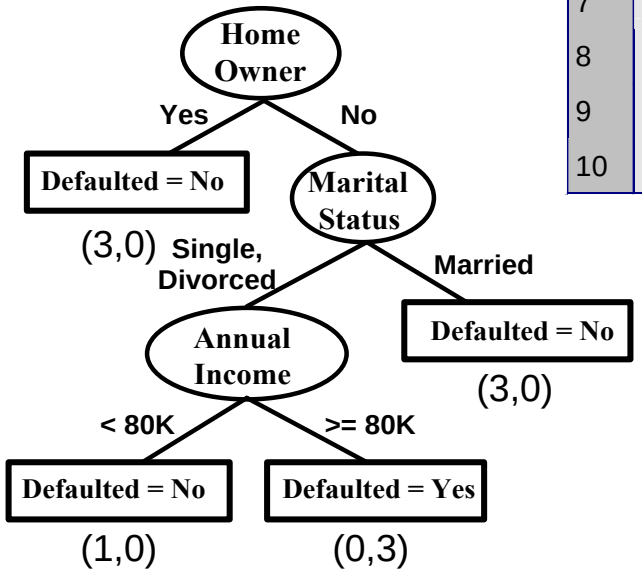


(b)

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



(c)



(d)

ID3 Algoritması

Karar ağaçları yardımıyla sınıflandırma işlemlerini yerine getirmek üzere Quinlan tarafından birçok algoritma geliştirilmiştir. Bu algoritma aşağıdan yukarı (top-down : kökten alt dallara doğru) ve greedy search (sonuca en yakın durum) teknikleri kullanılır. Decision Tree konusunda sıklıkla göreceğiz C4.5 algoritması ID3 algoritmasının bir uzantısıdır. ID3 algoritması Entropy ve Information Gain üzerine inşa edilmiştir.

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;

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

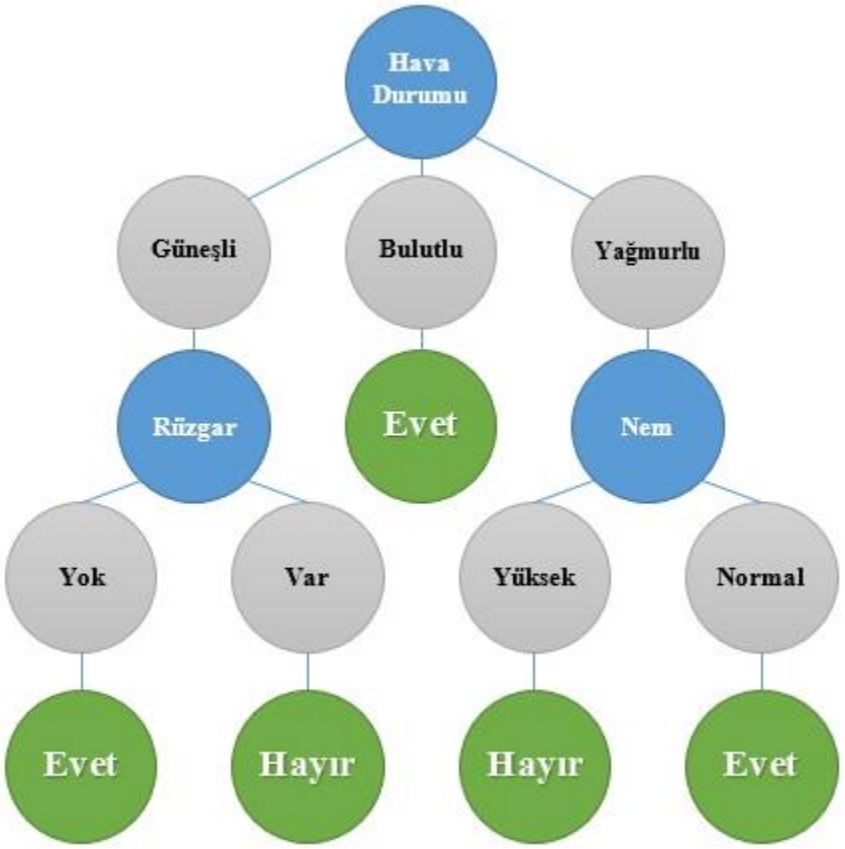
$$= -(0.25\log_2(0.25) + 0.75\log_2(0.75)) = 0.81128$$

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 Seti

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

- Entropi sadece hedef üzerine hesaplanmaz. Ayrıca özellikler üzerine entropi hesaplanabilir. Fakat özellikler üzerine entropi hesaplanırken hedefte göz önüne alır. Bu durumda entropi formülü:

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

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

E(FutbolOyna, HavaDurumu)
= **P**(Güneşli)*__E__(3,2)+**P**(Bulutlu)*__E__(4,0)+**P**(Yağmurlu)*__E__(2,3)
P(Güneşli) = 5/14 = 0.3571, **E**(3,2) = 0.971
P(Bulutlu)= 4 / 14 = 0.286, **E**(4,0) = 0
P(Yağmurlu)= 5 / 14 = 0.357, **E**(2,3) = 0.971
E(FutbolOyna, HavaDurumu) = 0.694

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

E(FutbolOyna, Nem) =
P(Yüksek)***E**(3,4)+**P**(Normal)***E**(6,1)
E(3,4) = 0.985, **P**(Yüksek) = 0.5
E(6,1) = 0.592, **P**(Normal) = 0.5
E(FutbolOyna, Nem) = 0.788

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

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

E(FutbolOyna, Sıcaklık) = 0.911

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

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

E(FutbolOyna, Rüzgar) = 0.892

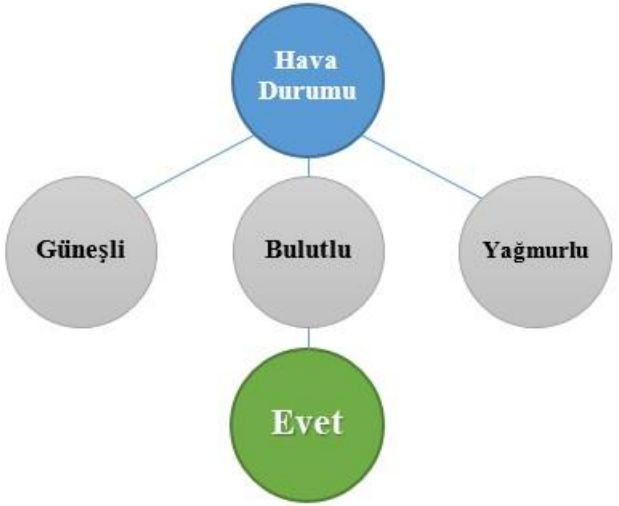
Information Gain (Bilgi Kazanımı)

- 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.
- Kazanç(X,T)=H(T)-H(X,T)

- **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, HavaDurumu) = 0.940 – 0.892 = 0.048

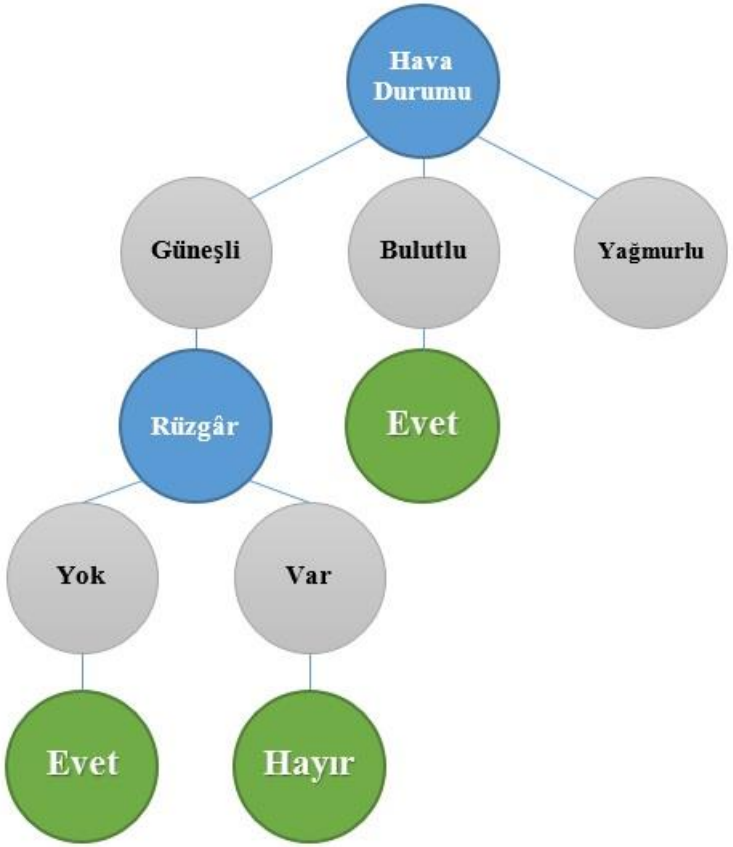
- Gain(FutbolOyna, HavaDurumu) özelliği en yüksek information gain değerine sahiptir. Bu özellik seçildikten sonra özelliğin değerlerine bakılarak en yüksek information gain'e sahip alan seçilir. HavaDurumu Bulutlu olduğunda tüm FutbolOyna değerleri Evet'tir. Şekil olarak aşağıdaki gibidir:

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



- Hava Bulutlu ise Futbol oynuyoruz. İkinci aşamada HavaDurumu Güneşli olan durumlar seçilir. Güneşli olduğunda oynama ve oynama durumları vardır. Bu durumda tekrar information Gain hesaplanır ve Rüzgar durumunun karar verici bir özellik olduğu görülür.
- Rüzgar var ise oynayamıyoruz. Rüzgar yok ise oynayabiliyoruz. Özyinelemeli bir şekilde yaprak kalmayıncaya kadar ID3 algoritması devam eder.

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



Örnek Çalışma

- https://towardsdatascience.com/decision-tree-intuition-from-concept-to-application-530744294bb6?source=rss----7f60cf5620c9---4

Kaynaklar

- Akküçük, Ulaş. “Veri madenciliği: kümeleme ve sınıflama algoritmaları”. *İstanbul: Yalın Yayıncılık* 18 (2011).
- Han, Jiawei, Jian Pei, ve Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
- Kantardzic, Mehmed. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.
- Sumathi, Sai, ve S. N. Sivanandam. *Introduction to data mining and its applications*. C. 29. Springer, 2006.
- Tan, Pang-Ning, Michael Steinbach, ve Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.
- Towards Data Science. “Towards Data Science”. Erişim 29 Mart 2020. https://towardsdatascience.com/.
- VanderPlas, Jake. *Python Data Science Handbook. OReilly Media*. Inc, 2017.