Veri Madenciliği

Sınıflandırma

Dr. Öğretim Üyesi Alper Talha KARADENİZ

Veri madenciliği yöntemleri

Tahmin edici modeller (predictive): Sonuçları bilinen verilerden yola çıkarak bir model kurulur. Bu modelden faydalanılarak veri kümelerindeki değeri bilinmeyen sonuç değerleri tahmin edilmeye çalışılır.

- *Sınıflandırma*

Tanımlayıcı modeller (descriptive): Veriyi tanımlayan örüntülerin bulunması.

- *Kümeleme*
- *Birliktelik kuralları*

Öğrenme

Denetimli (supervised) öğrenme: (Sınıflandırma)

Sınıf sayısı ve bir grup örneğin hangi sınıfa ait olduğunu bilinir.

Denetimsiz (unsupervised) öğrenme: (Kümeleme)

Hangi nesnenin hangi sınıfa ait olduğu ve grup sayısı belirsizdir.

Pekiştirmeli (reinforcement) öğrenme:

Makinanın belirli bir görevi yapmak için eğitildiği, belirli bir görevi yaparken daha önceki deneyimlerine ve çıktılarına bağlı olarak kendi kendine öğrendiği bir öğrenme türüdür.

Sınıflandırma Süreci

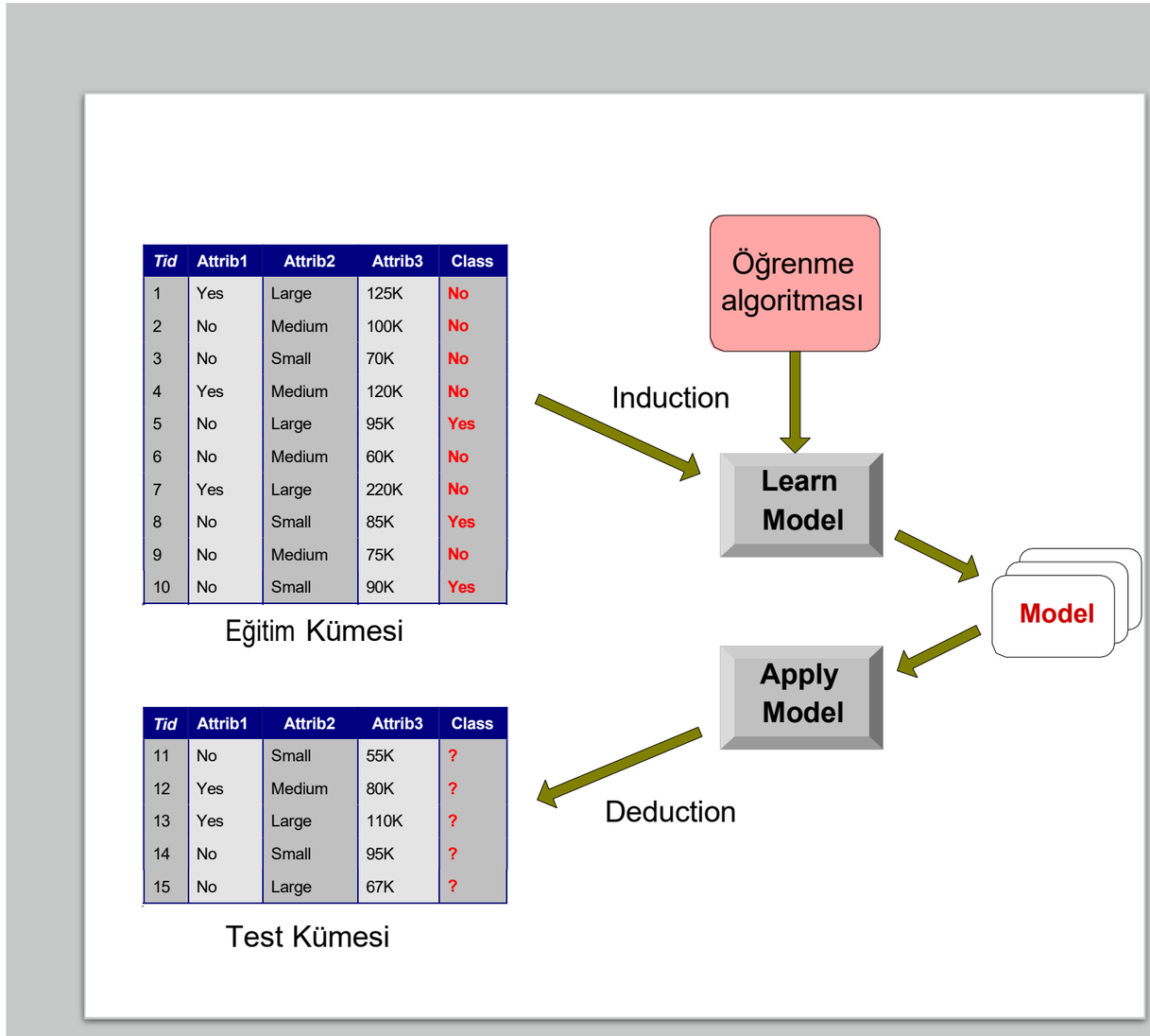
Sınıflandırma süreci iki aşamadan oluşur;

1. Model Oluşturma:

Model, veritabanındaki kayıtların nitelikleri veya alan isimleri kullanılarak gerçekleştirilir. Sınıflandırma modelinin elde edilmesi için verilerin bir kısmı eğitim verileri olarak kullanılır. Rastgele seçilen eğitim verileri üzerinde çalışma yapılarak sınıflama modeli elde edilir

2. Modelin Öngörü İçin Kullanılması:

Eğitim verileri üzerinde sınıflandırma kuralları belirlenir. Testler uygulanarak kurallar kontrol edilir



Sınıflandırma yöntemleri nerelerde kullanılır?

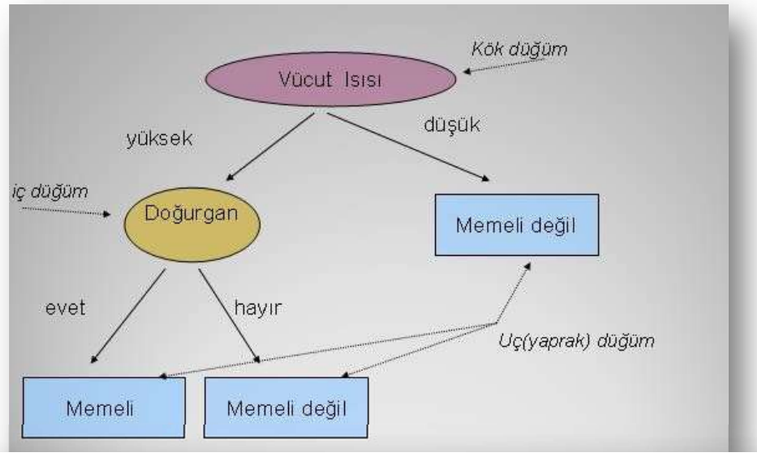
- Kredi başvurusu değerlendirme
- Sahtekarlık tespiti
- Hastalık teşhisi
- Ses tanıma
- Karakter tanıma
- Resim tanıma
- Kullanıcı davranışlarını belirleme
- Kalite kontrol çalışmaları
- Pazarlama

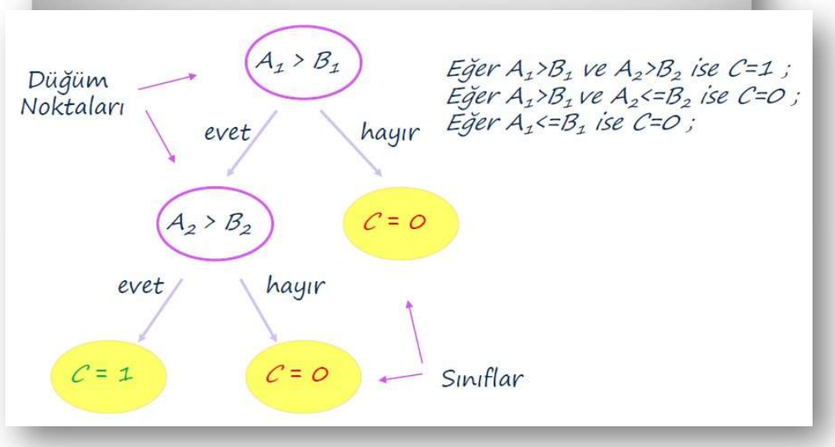
Sınıflandırma Teknikleri ve Algoritmaları

- **Karar Ağaçları**
  ID3
  C4.5
- **İstatistiğe Dayalı Algoritmalar**
  Regresyon ağaçları
  Bayes sınıflandırma
- **Mesafeye Dayalı Algoritmalar**
  k-en yakın komşu algoritması
- **Yapay Sinir Ağları**
- **Destek Vektör Makineleri (Support Vector Machine) - SVM**

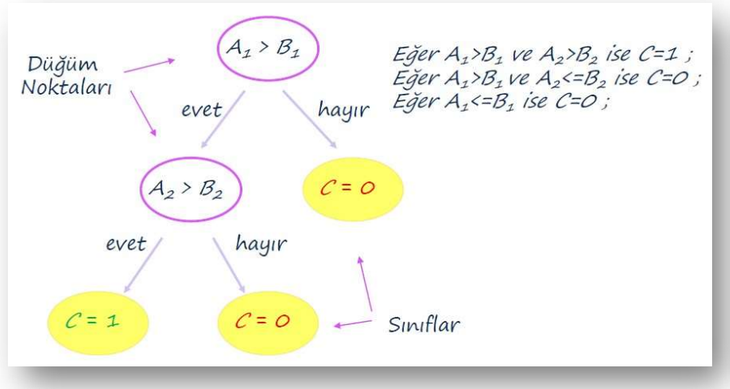
Karar Ağaçları

- Karar ağaçları, akış şemalarına benzeyen yapılandırmalardır. Bu yapıyı ağacın ters dönmüş haline benzetebiliriz.
- Yaygın kullanılan öngörü yöntemlerinden bir tanesidir.
- Her bir nitelik bir karar noktası(düğüm) tarafından belirlenir.
- Düğüm dalları testin sonucunu belirtir.
- Ağaç yaprakları sınıf etiketlerini içerir.





Karar Ağaçları



Karar ağacı çıkarımı iki aşamadan oluşur:

1. *Ağaç inşası*
   Başlangıçta bütün öğrenme örnekleri kök düğümdedir.
   Örnekler seçilmiş özelliklere tekrarlamalı olarak bölünür.
2. *Ağaç Temizleme*
   Gürültü ve istisna kararları içeren dallar belirlenir ve kaldırılır.

Karar ağacı kullanımı:

Sınıfı bilinmeyen yeni örneğin özellikleri karar ağacında test edilerek sınıfı bulunur.

Karar ağaçları yapısı:

- Karar ağaçlarında en önemli problem niteliklerin alt düğüm noktalarında hangi alt niteliklere bölüneceğinin belirlenmesidir.
- En iyi bölen nitelik nasıl belirlenir?
- İyilik fonksiyonları
  Bilgi kazancı (information gain): ID3
  Kazanç oranı (gain ratio): C4.5
  Gini index: CART

Entropi temelli algoritmalar

- Entropi belirsizliğin ölçütüdür
- 0 ile 1 arasında değer alır
- Eğer örneklerin tamamı aynı sınıfa ait ise entropi 0 olur
  - Herkesin aynı futbol takımını tuttuğu bir grupta herhangi birine tuttuğu takımı sorduğumuzda alacağımız cevap bizi şaşırtmayacak, yani entropi 0 olacaktır.
- Örnekler belirlenen sınıflar arasında eşit dağılmış ise entropi 1 olur

$$Ent(A) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Ent_B(A) = \sum_{j=1}^{v} \frac{|A_j|}{|A|} Ent(A_j)$$

$$Gain = Ent(A) - Ent_A(B)$$

Entropi Hesabı

R={+,+,*,*,*,*,*,*} veri seti için entropi hesabı yapalım

Veri setinde 2 tane +, 6 tane * var.

+ için olasılık değeri 2\8, * için olasılık değeri 6\8 olduğu görülmektedir.

Entropi hesabı:

$$Ent(R) = I(2,6) = -\frac{2}{8}\log_2(\frac{2}{8}) - \frac{6}{8}\log_2\left(\frac{6}{8}\right) = 0{,}81128$$

Entropi Hesabı

ID	Yaş	Gelir	Krd. P.n	PC Alma
1	Genç	Yüksek	Orta	Hayır
2	Genç	Yüksek	İyi	Hayır
3	Orta Yaşlı	Yüksek	Orta	Evet
4	Yaşlı	Orta	Orta	Evet
5	Yaşlı	Düşük	Orta	Evet
6	Yaşlı	Düşük	İyi	Hayır
7	Orta Yaşlı	Düşük	İyi	Evet

ID	Yaş	Gelir	Krd. P.n	PC Alma
8	Genç	Orta	Orta	Hayır
9	Genç	Düşük	Orta	Evet
10	Yaşlı	Orta	Orta	Evet
11	Genç	Orta	İyi	Evet
12	Orta Yaşlı	Orta	İyi	Evet
13	Orta Yaşlı	Yüksek	Orta	Evet
14	Yaşlı	Orta	İyi	Hayır

$$Ent(A) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0{,}940$$

$$Ent_{Yaş}(A) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0{,}694$$

$$Gain = Ent(A) - Ent_{Yaş}(B) = 0{,}246$$
$$Gain = Ent(A) - Ent_{Gelir}(B) = 0{,}029$$
$$Gain = Ent(A) - Ent_{Kredi}(B) = 0{,}048$$

Entropi temelli algoritmalar

ID3

ID3 algoritmasında karar ağaçları alt dallara bölünürken entropi değerleri incelenerek en az kayıp olan bölünme dikkate alınır.
Kesikli değerler ile çalışır.

C4.5

ID3 algoritmasının bir uzantısıdır
ID3 te bilgi kazanımı dikkate alınırken, C4.5 te kazanım oranı ölçütü ile kuralların iyiliği sorgulanır.
Temelde kesikli değerler ile çalışmasına rağmen sürekli değerler ile de çalışabilir.

ID3

- Genel entropi hesaplanır
- Her bir özniteliğin entropisi ayrı ayrı hesaplanır
- Hesaplanan entropi değerleri genel entropi değerinden çıkartılarak hangi özniteliğin genel entropi değerini en çok azalttığı belirlenir (kazanç).
- Entropiyi en çok azaltan öznitelik en çok kazanç sağlayandır.

ID3
Örnek

Öğrenci	Mezuniyet	Not Ort.	Cinsiyet	Tecrübe	Memnuniyet
1	İşletme	Yüksek	E	Var	Hayır
2	İşletme	Yüksek	E	Yok	Hayır
3	Endüstri	Yüksek	E	Var	Evet
4	YBS	Orta	E	Var	Evet
5	YBS	Düşük	K	Var	Evet
6	YBS	Düşük	K	Yok	Hayır
7	Endüstri	Düşük	K	Yok	Evet
8	İşletme	Orta	E	Var	Hayır
9	İşletme	Düşük	K	Var	Evet
10	YBS	Orta	K	Var	Evet
11	İşletme	Orta	K	Yok	Evet
12	Endüstri	Orta	E	Yok	Evet
13	Endüstri	Yüksek	K	Var	Evet
14	YBS	Orta	E	Yok	Hayır

ID3
Örnek

$$Ent(Genel) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0{,}940$$

$$Ent_{Mezun}(Genel) = \frac{5}{14}Ent_{IŞL} + \frac{4}{14}Ent_{END} + \frac{5}{14}Ent_{YBS}$$

$$Ent(Mezun - IŞL) = -\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0{,}971$$

$$Ent(Mezun - END) = -\frac{4}{4}\log_2(\frac{4}{4}) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) = 0$$

$$Ent(Mezun - YBS) = -\frac{3}{5}\log_2(\frac{3}{5}) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0{,}971$$

$$Ent_{Mezun}(Genel) = \frac{5}{14}0{,}971 + \frac{4}{14}0 + \frac{5}{14}0{,}971 = 0{,}694$$

$$Gain = Ent(Genel) - Ent_{Mezun}(Genel) = 0{,}940 - 0{,}694 = 0{,}247$$

$$Ent(A) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Ent_B(A) = \sum_{j=1}^{v} \frac{|A_j|}{|A|} Ent(A_j)$$

$$Gain = Ent(A) - Ent_A(B)$$

Mezuniyet	Memnuniyet
Endüstri	Evet
Endüstri	Evet
Endüstri	Evet
Endüstri	Evet
İşletme	Evet
İşletme	Evet
İşletme	Hayır
İşletme	Hayır
İşletme	Hayır
YBS	Evet
YBS	Evet
YBS	Evet
YBS	Hayır
YBS	Hayır

ID3
Örnek

$$Ent(Genel) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0{,}940$$
$$Ent_{Not}(Genel) = \frac{4}{14}Ent_D + \frac{6}{14}Ent_O + \frac{4}{14}Ent_Y$$

$$Ent(Not - Düşük) = -\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0{,}811$$

$$Ent(Not - Orta) = -\frac{4}{6}\log_2(\frac{4}{6}) - \frac{2}{6}\log_2\left(\frac{2}{4}\right) = 0\,,918$$

$$Ent(Not - Yüksek) = -\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1{,}00$$

$$Ent_{Not}(Genel) = \frac{4}{14}0{,}811 + \frac{6}{14}0{,}918 + \frac{4}{14}1{,}00 = 0{,}911$$

$$Gain = Ent(Genel) - Ent_{Not}(Genel) = 0{,}940 - 0{,}911 = 0{,}029$$

$$Ent(A) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Ent_B(A) = \sum_{j=1}^{v} \frac{|A_j|}{|A|} Ent(A_j)$$

$$Gain = Ent(A) - Ent_A(B)$$

Not Ort.	Memnuniyet
Düşük	Evet
Düşük	Evet
Düşük	Evet
Düşük	Hayır
Orta	Evet
Orta	Evet
Orta	Evet
Orta	Evet
Orta	Hayır
Orta	Hayır
Yüksek	Evet
Yüksek	Evet
Yüksek	Hayır
Yüksek	Hayır

ID3
Örnek

$$Ent(Genel) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0{,}940$$

$$Ent_{Cinsiyet}(Genel) = \frac{7}{14}Ent_E + \frac{7}{14}Ent_B$$

$$Ent(Cinsiyet - Erkek) = -\frac{3}{7}\log_2(\frac{3}{7}) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0{,}985$$

$$Ent(Cinsiyet - Bayan) = -\frac{6}{7}\log_2(\frac{6}{7}) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) = 0{,}592$$

$$Ent_{Cinsiyet}(Genel) = \frac{7}{14}0{,}985 + \frac{7}{14}0{,}592 = 0{,}789$$

$$Gain = Ent(Genel) - Ent_{Cinsiyet}(Genel) = 0{,}940 - 0{,}789 = 0{,}151$$

$$Ent(A) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$Ent_B(A) = \sum_{j=1}^{v} \frac{|A_j|}{|A|} Ent(A_j)$$

$$Gain = Ent(A) - Ent_A(B)$$

Cinsiyet	Memnuniyet
E	Evet
E	Evet
E	Evet
E	Hayır
E	Hayır
E	Hayır
E	Hayır
K	Evet
K	Evet
K	Evet
K	Evet
K	Evet
K	Evet
K	Hayır

ID3
Örnek

$$Ent(Genel) = I(9{,}5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0{,}940$$

$$Ent_{Tecrübe}(Genel) = \frac{8}{14}Ent_V + \frac{6}{14}Ent_Y$$

$$Ent(Tecrübe - Var) = -\frac{6}{8}\log_2(\frac{6}{8}) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0{,}811$$

$$Ent(Tecrübe - Yok) = -\frac{3}{6}\log_2(\frac{3}{6}) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) = 1{,}00$$

$$Ent_{Tecrübe}(Genel) = \frac{6}{14}0{,}811 + \frac{8}{14}1{,}00 = 0{,}892$$

$$Gain = Ent(Genel) - Ent_{Tecrübe}(Genel) = 0{,}940 - 0{,}892 = 0{,}048$$

Tecrübe	Memnuniyet
Var	Hayır
Yok	Hayır
Var	Evet
Var	Evet
Var	Evet
Yok	Hayır
Yok	Evet
Var	Hayır
Var	Evet
Var	Evet
Yok	Evet
Yok	Evet
Var	Evet
Yok	Hayır

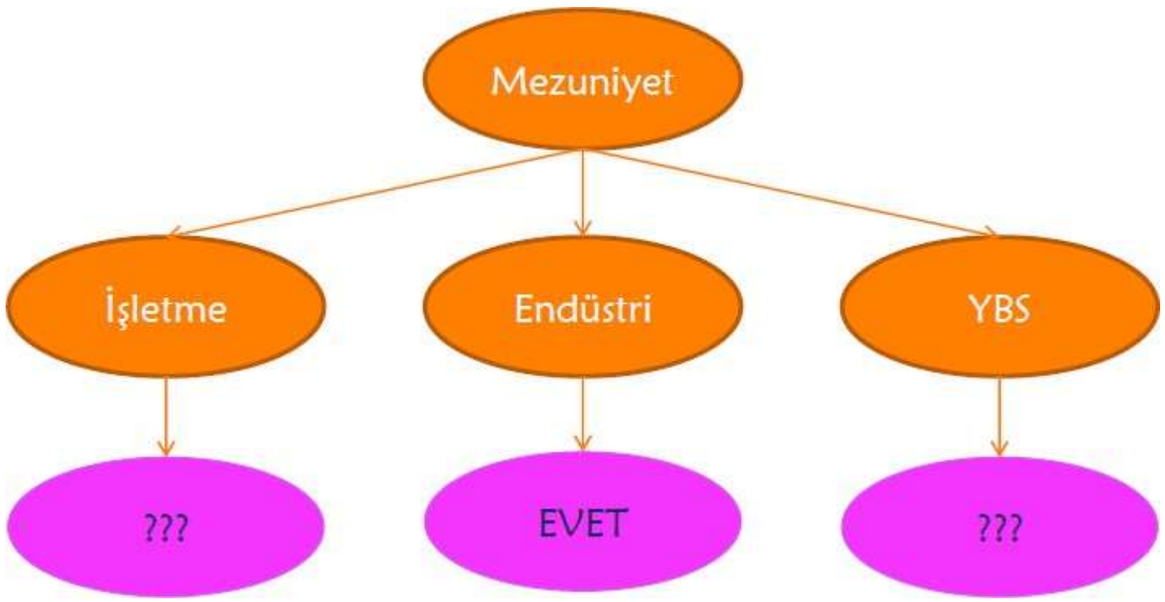
ID3
Örnek

- Tüm değişkenler bölünme için kontrol edildi.

- Elde edilen kazanım değerleri:
  - Mezuniyet: 0.247 en büyük kazanç
  - Not ortalaması: 0.029
  - Cinsiyet: 0.151
  - Tecrübe: 0.048

- En büyük kazanım mezuniyet için olduğundan ilk düğüm mezuniyet olacak.

ID3
Örnek



İşletme ve YBS altında tekrar bir bölünme olacak.

ID3
Örnek

- İşletme bölümü kazanç değerleri:
  - Not ortalaması: 0.570
  - Cinsiyet: 0.970
  - Tecrübe: 0.019

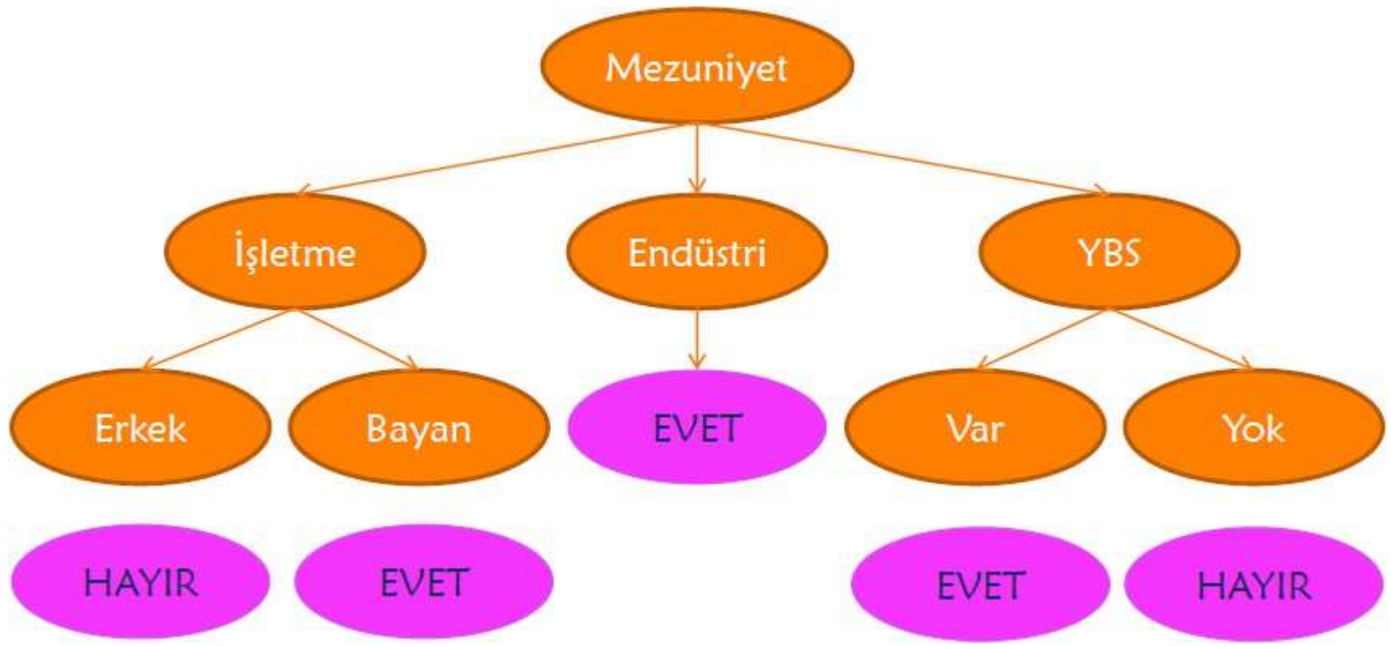
Öğrenci	Mezuniyet	Not Ort.	Cinsiyet	Tecrübe	Memnuniyet
1	İşletme	Yüksek	E	Var	Hayır
2	İşletme	Yüksek	E	Yok	Hayır
8	İşletme	Orta	E	Var	Hayır
9	İşletme	Düşük	K	Var	Evet
11	İşletme	Orta	K	Yok	Evet

- YBS bölümü kazanç değerleri:
  - Not ortalaması: 0.019
  - Tecrübe: 0.970

Öğrenci	Mezuniyet	Not Ort.	Cinsiyet	Tecrübe	Memnuniyet
4	YBS	Orta	E	Var	Evet
5	YBS	Düşük	K	Var	Evet
6	YBS	Düşük	K	Yok	Hayır
10	YBS	Orta	K	Var	Evet
14	YBS	Orta	E	Yok	Hayır

- Kazanç değerleri incelendiğinde işletme cinsiyet ile, YBS tecrübe ile dallanacaktır

ID3
Örnek


Mezuniyet
İşletme
Endüstri
YBS
Erkek
Bayan
EVET
Var
Yok
HAYIR
EVET
EVET
HAYIR

ID3
Örnek

- Ağacın dallanması tamamlandığında 5 yaprak (karar) oluşmuştur.
- Bu durumda 5 kural yazılabilir:

Eğer Mezuniyet=İşletme ve Cinsiyet =Erkek ise İşçiden memnun kalınmıyor

Eğer Mezuniyet=İşletme ve Cinsiyet =Bayan ise İşçiden memnun kalınıyor

Eğer Mezuniyet=Endüstri ise İşçiden memnun kalınıyor

Eğer Mezuniyet=YBS ve Ders verme tecrübesi=VAR ise İşçiden memnun kalınıyor

Eğer Mezuniyet=YBS ve Ders verme tecrübesi=YOK ise İşçiden memnun kalınmıyor

Çalışma Sorusu

ID3 algoritmasını kullanarak karar ağacı oluşturunuz

HAVA	ISI	NEM	RÜZGAR	OYUN
güneşli	sıcak	yüksek	hafif	hayır
güneşli	sıcak	yüksek	kuvvetli	hayır
bulutlu	sıcak	yüksek	hafif	evet
yağmurlu	ılık	yüksek	hafif	evet
yağmurlu	soğuk	normal	hafif	hayır
yağmurlu	soğuk	normal	kuvvetli	evet
bulutlu	soğuk	normal	kuvvetli	hayır
güneşli	ılık	yüksek	hafif	evet
güneşli	soğuk	normal	hafif	evet
yağmurlu	ılık	normal	hafif	evet
güneşli	ılık	normal	kuvvetli	evet
bulutlu	ılık	yüksek	kuvvetli	evet
bulutlu	sıcak	normal	hafif	evet
yağmurlu	ılık	yüksek	kuvvetli	hayır

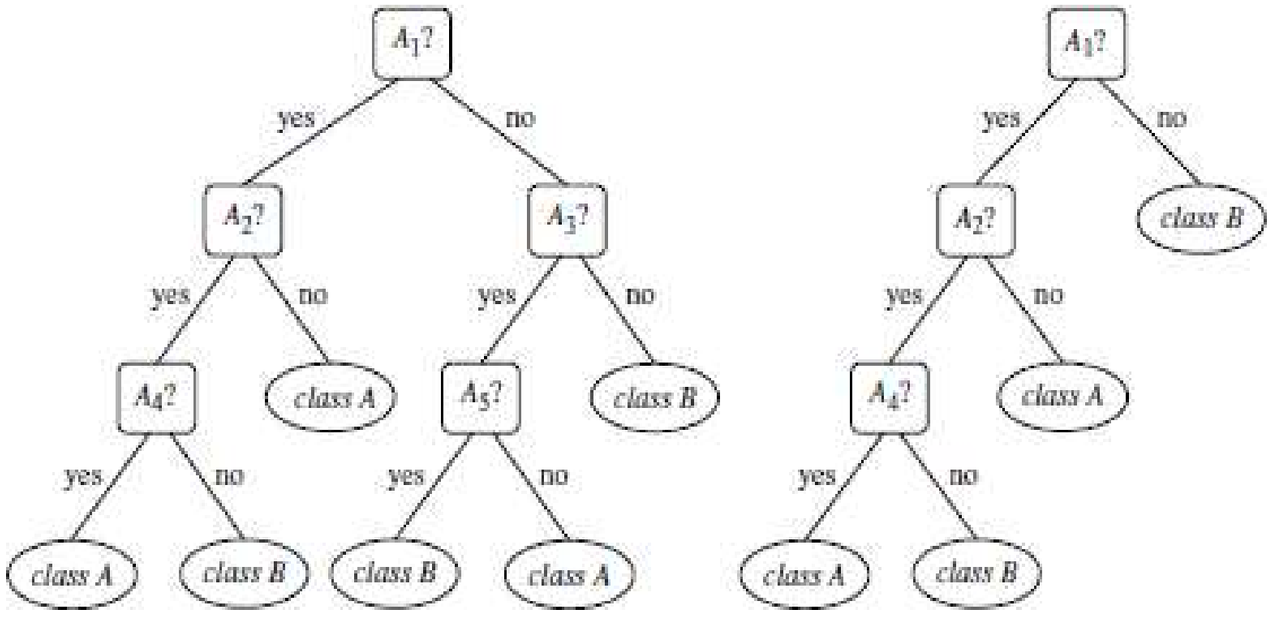
C4.5

- C4.5 algoritması ile sayısal değerlere sahip niteliklerin de karar ağaçları oluşturulabilir.
- Sayısal değerler için en büyük bilgi kazancını sağlayacak biçimde bir eşik değeri belirlenir.
- Nitelik değerler sıralanır, eşik değer ile nitelik değeri iki parçaya ayrılır.
- Eşik değeri dizinin orta noktası olabilir.

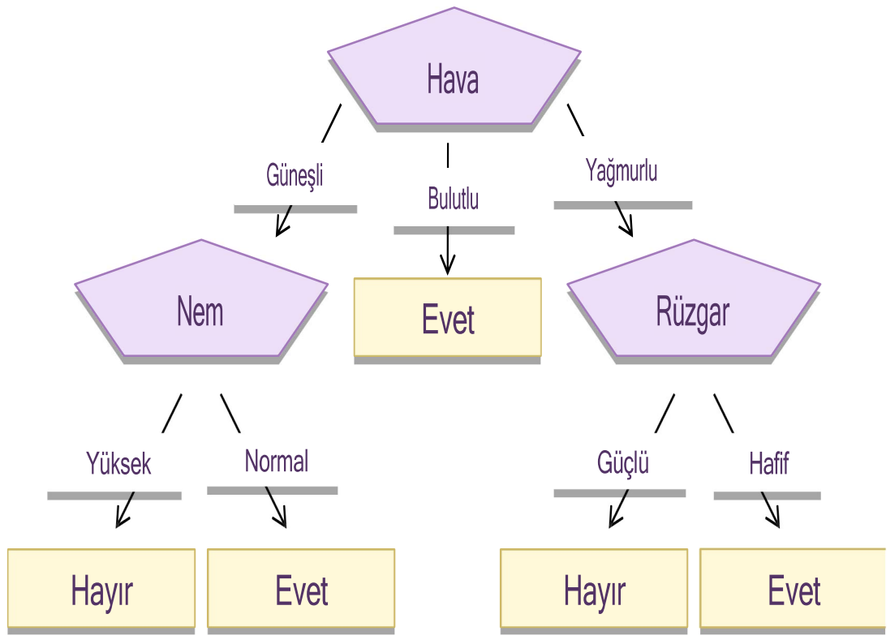
Karar ağaçlarının budanması

- Amaç, karmaşık olmayan ağaçlar oluşturmaktır.
- Ağacın budanması, bütün bir alt ağacın yerine bir yaprak düğümünün yerleştirilmesiyle yapılır.
- Yerleştirme ancak bir alt ağaçtaki beklenen hata tek yapraktakinden daha büyükse yapılır.
- Alt ağacın yerine yaprak yerleştirmekle ,algoritma “öngörülü hata oranını” azaltmayı ve sınıflandırma modelinin kalitesini arttırmayı amaçlar.

Karar ağaçlarının budanması



Karar ağaçlarından kuralların yazılması



1. Kural :
   Eğer Hava=Güneşli ise ve
   Eğer Nem=Yüksek ise Oyun=Hayır ;

2. Kural :
   Eğer Hava=Güneşli ise ve
   Eğer Nem=Normal ise Oyun=Evet ;

3. Kural :
   Eğer Hava=Bulutlu ise Oyun=Evet ;

4. Kural :
   Eğer Hava=Yağmurlu ise ve
   Eğer Rüzgar=Güçlü ise Oyun=Hayır ;

5. Kural :
   Eğer Hava=Yağmurlu ise ve
   Eğer Rüzgar=Hafif ise Oyun=Evet

KNN Algoritması

- Veri setinin olasılık dağılımı hakkında yeterli bilgi yoksa, KNN ilk seçenek olabilir.
- KNN, yeni örneklerin veri noktalarına olan mesafelerini belirlemek için eğitim setindeki örnekleri kullanılır.
- Eğitim setinde, yeni örneğe en yakın k adet komşu arasında en fazla ortak etikete sahip örneklerin ait olduğu sınıf, yeni örneğin de sınıfı olarak belirlenir.
- Yeni örneklerin, ait olduğu sınıfların tahmin edilmesi amacıyla, eğitim veri seti ile karşılaştırılması gerektiğinden, eğitim seti boyut (dimension) ve kapsam (size) olarak genişse, KNN maliyetli olabilir.

- Bununla birlikte KNN, eğitilmesi ve kesin sonuçlar alması kolay olduğu için yaygın olarak kullanılmaktadır.
- KNN'nin sınıflandırma performansı büyük ölçüde k (sınıf adedi) parametresine ve mesafe metriğine bağlıdır.
- En uygun k ve mesafe metriği deneme-yanılma yoluyla belirlenir.
- Minimum hata oranını veren k ve mesafe metriği kombinasyonu seçilir.
- Veri seti büyüdükçe genellikle k değerinin de artması beklenir.
- KNN veri setindeki gürültülere duyarlıdır. Diğer bir ifadeyle gürültülerin varlığı KNN'in sınıflandırma performansını olumsuz etkiler.

- Bu yüzden verilerin (KNN algoritmasına verilmeden önce) min-max, z-score gibi yöntemlerle normalize edilmesi önemlidir.
- KNN, bir veri noktasını, k değeri ile temsil edilen kendisine en yakın veri kümesine atama yapmak suretiyle sınıflandırma yapar. En yaygın kullanılan uzaklık ölçüleri “minkowski”, “euclidean”, “manhattan” olarak sayılabilir.

Örnek

Basitlik adına, iki boyutlu bir uzayda (x, y) koordinatlarına sahip **8 veri noktası** ve bu noktaların **iki farklı sınıfa** ait etiketleri olsun:

Nokta	x	y	Sınıf
A	1	1	**Kırmızı**
B	2	2	**Kırmızı**
C	2	1	**Kırmızı**
D	3	2	**Mavi**
E	6	6	**Mavi**
F	7	7	**Mavi**
G	5	6	**Kırmızı**
H	7	5	**Mavi**

Şimdi, **yeni bir noktanın** $(X_{new} = (4, 3))$ **hangi sınıfa** ait olduğunu **K = 3** kullanarak belirlemeye çalışalım.

> **Not: Elimizde 2 farklı sınıf var: "Kırmızı" ve "Mavi".**

KNN'de en önemli kısım, yeni noktanın mevcut veri noktalarına uzaklıklarını ölçmektir. Genellikle **Öklid (Euclidean)** uzaklık kullanılır:

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

Yeni noktamız: $X_{new} = (4, 3)$. Her noktaya tek tek bakalım:

Nokta A $(1, 1)$

$$d(X_{new}, A) = \sqrt{(4-1)^2 + (3-1)^2} = \sqrt{3^2 + 2^2} = \sqrt{9 + 4} = \sqrt{13} \approx 3.605$$

Nokta B $(2, 2)$

$$d(X_{new}, B) = \sqrt{(4-2)^2 + (3-2)^2} = \sqrt{2^2 + 1^2} = \sqrt{4 + 1} = \sqrt{5} \approx 2.236$$

Nokta C $(2, 1)$

$$d(X_{new}, C) = \sqrt{(4-2)^2 + (3-1)^2} = \sqrt{2^2 + 2^2} = \sqrt{4 + 4} = \sqrt{8} \approx 2.828$$

Nokta D $(3, 2)$

$$d(X_{new}, D) = \sqrt{(4-3)^2 + (3-2)^2} = \sqrt{1 + 1} = \sqrt{2} \approx 1.414$$

Nokta E $(6, 6)$

$$d(X_{new}, E) = \sqrt{(4-6)^2 + (3-6)^2} = \sqrt{(-2)^2 + (-3)^2} = \sqrt{4 + 9} = \sqrt{13} \approx 3.605$$

Nokta F $(7, 7)$

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

Nokta G $(5, 6)$

$$d(X_{new}, G) = \sqrt{(4-5)^2 + (3-6)^2} = \sqrt{(-1)^2 + (-3)^2} = \sqrt{1 + 9} = \sqrt{10} \approx 3.162$$

Nokta H $(7, 5)$

$$d(X_{new}, H) = \sqrt{(4-7)^2 + (3-5)^2} = \sqrt{(-3)^2 + (-2)^2} = \sqrt{9 + 4} = \sqrt{13} \approx 3.605$$

Şimdi tüm uzaklıkları tablo hâlinde görelim:

Nokta	Koordinat	Sınıf	Uzaklık $(4,3)$'e	Yaklaşık Değer
D	(3,2)	Mavi	$\sqrt{2}$	1.414
B	(2,2)	Kırmızı	$\sqrt{5}$	2.236
C	(2,1)	Kırmızı	$\sqrt{8}$	2.828
G	(5,6)	Kırmızı	$\sqrt{10}$	3.162
A	(1,1)	Kırmızı	$\sqrt{13}$	3.605
E	(6,6)	Mavi	$\sqrt{13}$	3.605
H	(7,5)	Mavi	$\sqrt{13}$	3.605
F	(7,7)	Mavi	5	5.000

En küçük uzaklıktan en büyüğe doğru sıralarsak:

1. $D \approx 1.414$ (Mavi)
2. $B \approx 2.236$ (Kırmızı)
3. $C \approx 2.828$ (Kırmızı)
4. $G \approx 3.162$ (Kırmızı)
5. $A \approx 3.605$ (Kırmızı)
6. $E \approx 3.605$ (Mavi)
7. $H \approx 3.605$ (Mavi)
8. $F = 5.000$ (Mavi)

Bizim belirlediğimiz $K = 3$. Dolayısıyla yeni noktanın hangi sınıfa ait olduğunu belirlemek için **en yakın 3 nokta**ya bakıyoruz:

1. **D** (Mavi)
2. **B** (Kırmızı)
3. **C** (Kırmızı)

Bu 3 noktanın sınıfları:

- 1 nokta **Mavi**
- 2 nokta **Kırmızı**

En çok oy alan sınıf = **Kırmızı**.

Dolayısıyla $(4, 3)$ noktası, **K=3** için **Kırmızı** sınıfına atanır.

```
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
# Veri noktaları (x, y)
X = np.array([
    [1, 1],  # A - Kırmızı
    [2, 2],  # B - Kırmızı
    [2, 1],  # C - Kırmızı
    [3, 2],  # D - Mavi
    [6, 6],  # E - Mavi
    [7, 7],  # F - Mavi
    [5, 6],  # G - Kırmızı
    [7, 5],  # H - Mavi
])
```

```
# Sınıf etiketleri ('K' = Kırmızı, 'M' = Mavi)
y = np.array(['K','K','K','M','M','M','K','M'])
# Yeni nokta
X_new = np.array([[4, 3]])
# K=3 için KNN modeli
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X, y)
# Tahmin
prediction = knn.predict(X_new)
   print("Yeni nokta (4,3) sınıfı:", prediction[0])
```