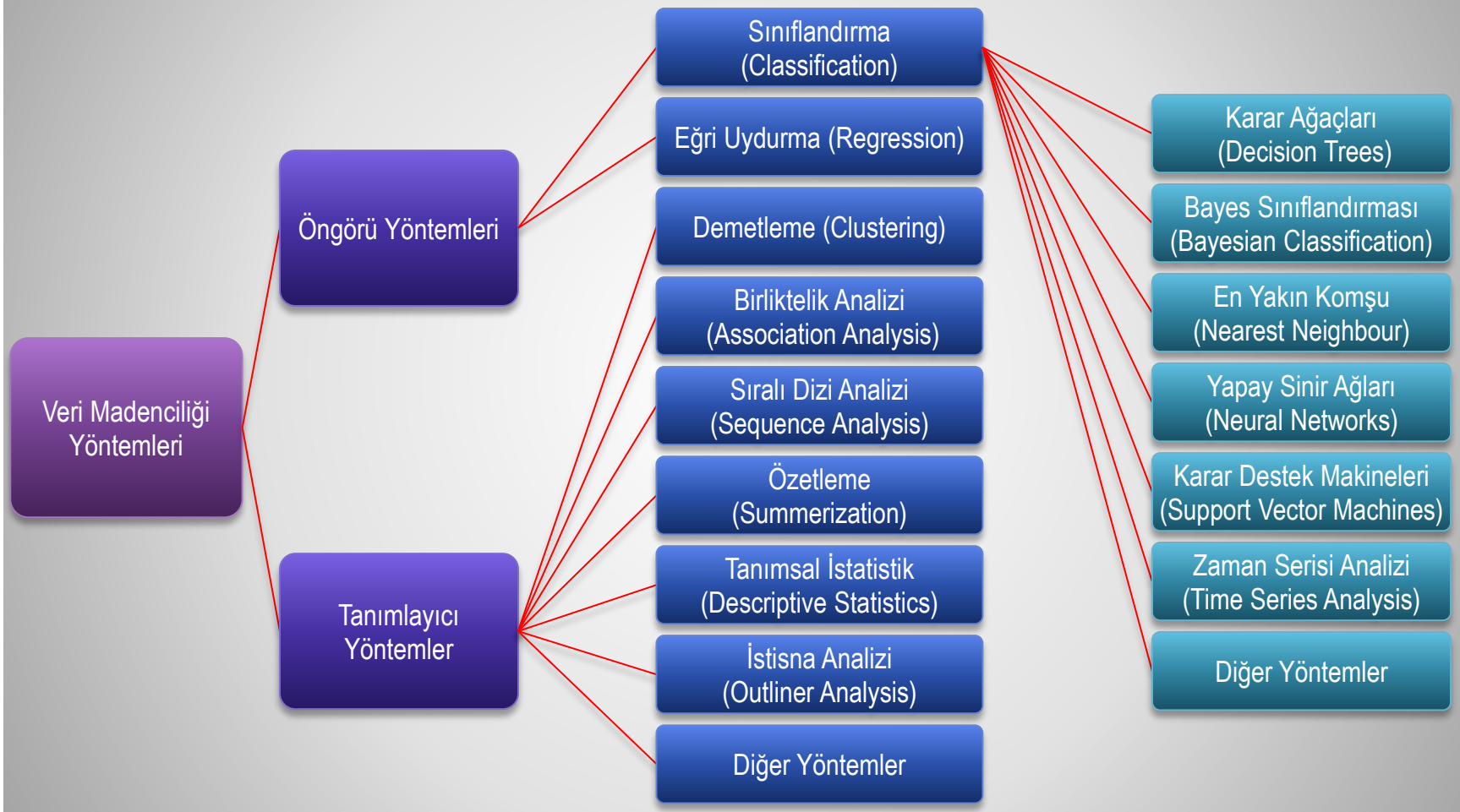
Veri Madenciliği

Karar Ağaçları ile Sınıflandırma

Genel olarak veri madenciliği yöntemleri iki sınıfa ayrılabilir:

1. Öngörü Yöntemleri (Prediction Methods)
   - Öngörü amacı ile var olan verilerden yorum çıkarılması
2. Tanımlayıcı Yöntemler (Description Methods)
   - Veriyi tanımlayan yorumlanabilir örüntülerin bulunması


Veri Madenciliği Yöntemleri
Öngörü Yöntemleri
Tanımlayıcı Yöntemler
Sınıflandırma (Classification)
Eğri Uydurma (Regression)
Demetleme (Clustering)
Birliktelik Analizi (Association Analysis)
Sıralı Dizi Analizi (Sequence Analysis)
Özetleme (Summerization)
Tanımsal İstatistik (Descriptive Statistics)
İstisna Analizi (Outliner Analysis)
Diğer Yöntemler
Karar Ağaçları (Decision Trees)
Bayes Sınıflandırması (Bayesian Classification)
En Yakın Komşu (Nearest Neighbour)
Yapay Sinir Ağları (Neural Networks)
Karar Destek Makineleri (Support Vector Machines)
Zaman Serisi Analizi (Time Series Analysis)
Diğer Yöntemler

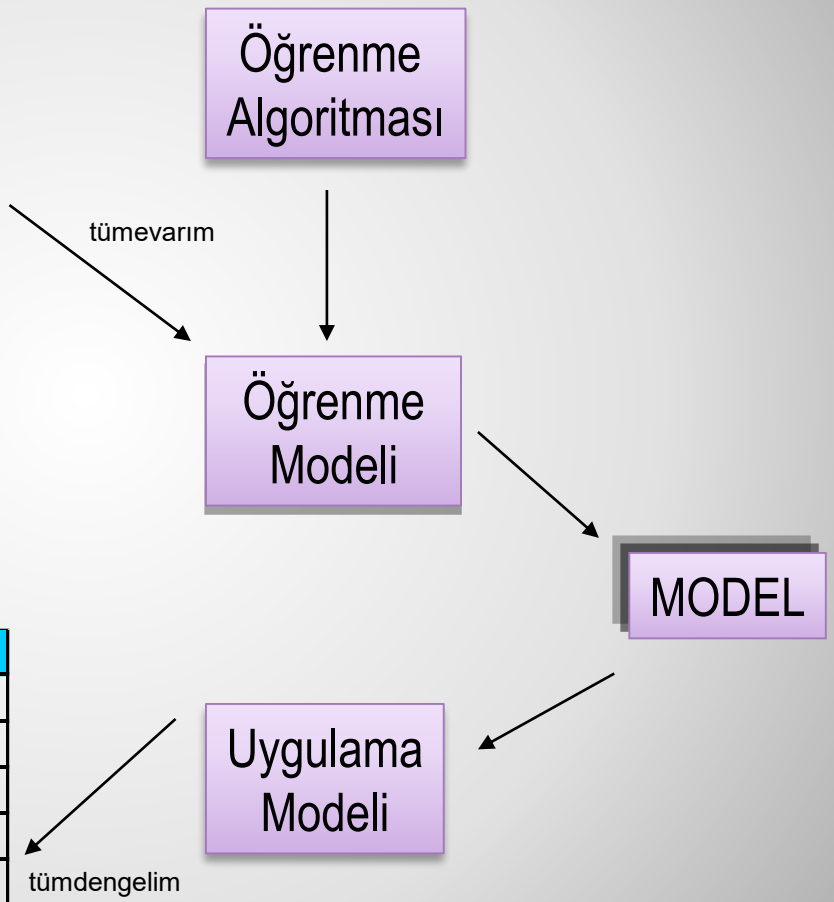
Sınıflandırma süreci

Öğrenme Seti

No	Öz.1	Öz.2	Öz.3	Sınıf
1	evet	büyük	125k	hayır
2	hayır	orta	100k	hayır
3	hayır	küçük	70k	hayır
4	evet	orta	120k	hayır
5	hayır	büyük	95k	evet
6	hayır	orta	60k	hayır
7	evet	büyük	220k	hayır
8	hayır	küçük	85k	evet
9	hayır	orta	75k	hayır
10	hayır	küçük	90k	evet

Test Seti

No	Öz.1	Öz.2	Öz.3	Sınıf
11	hayır	küçük	55k	?
12	evet	orta	80k	?
13	evet	büyük	110k	?
14	hayır	küçük	95k	?
15	hayır	büyük	67k	?



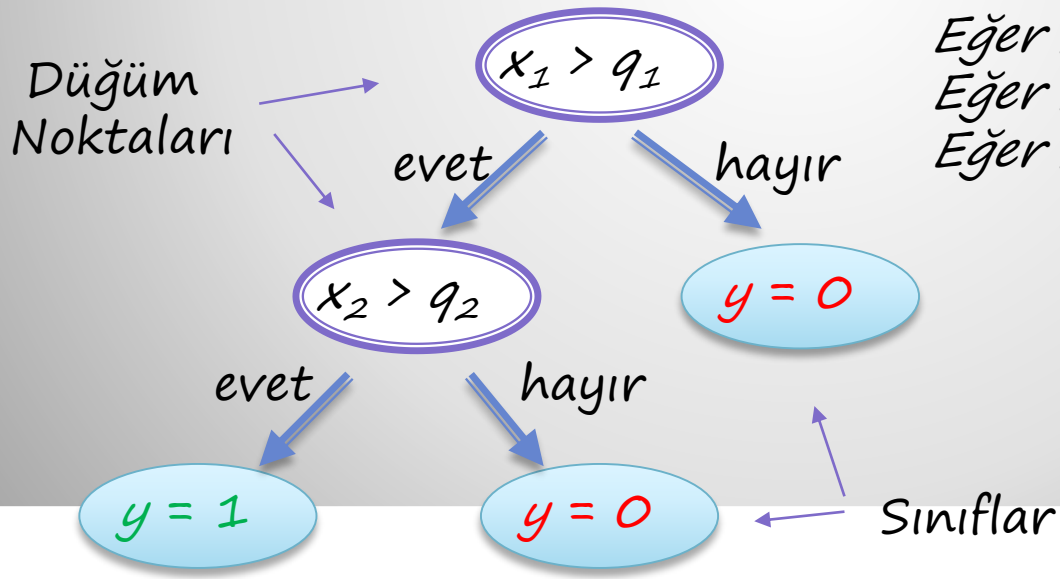
Karar Ağaçları :

Verilerin sınıflandırma yöntemlerinden biride karar ağaçlarıdır.

Karar ağaçlarının oluşturulmasında çok sayıda öğrenme yöntemi mevcuttur.

Akış şemalarına benzeyen yapılandırmalardır. Her bir nitelik bir karar noktası(düğüm) tarafından belirlenir. Bu yapıyı ağacın ters dönmüş haline benzetebiliriz.

Bu tür yaklaşımlar karar ağaçları sınıflandırma algoritmaları uygulayabilmek için uygun bir altyapı sağlamaktadır.



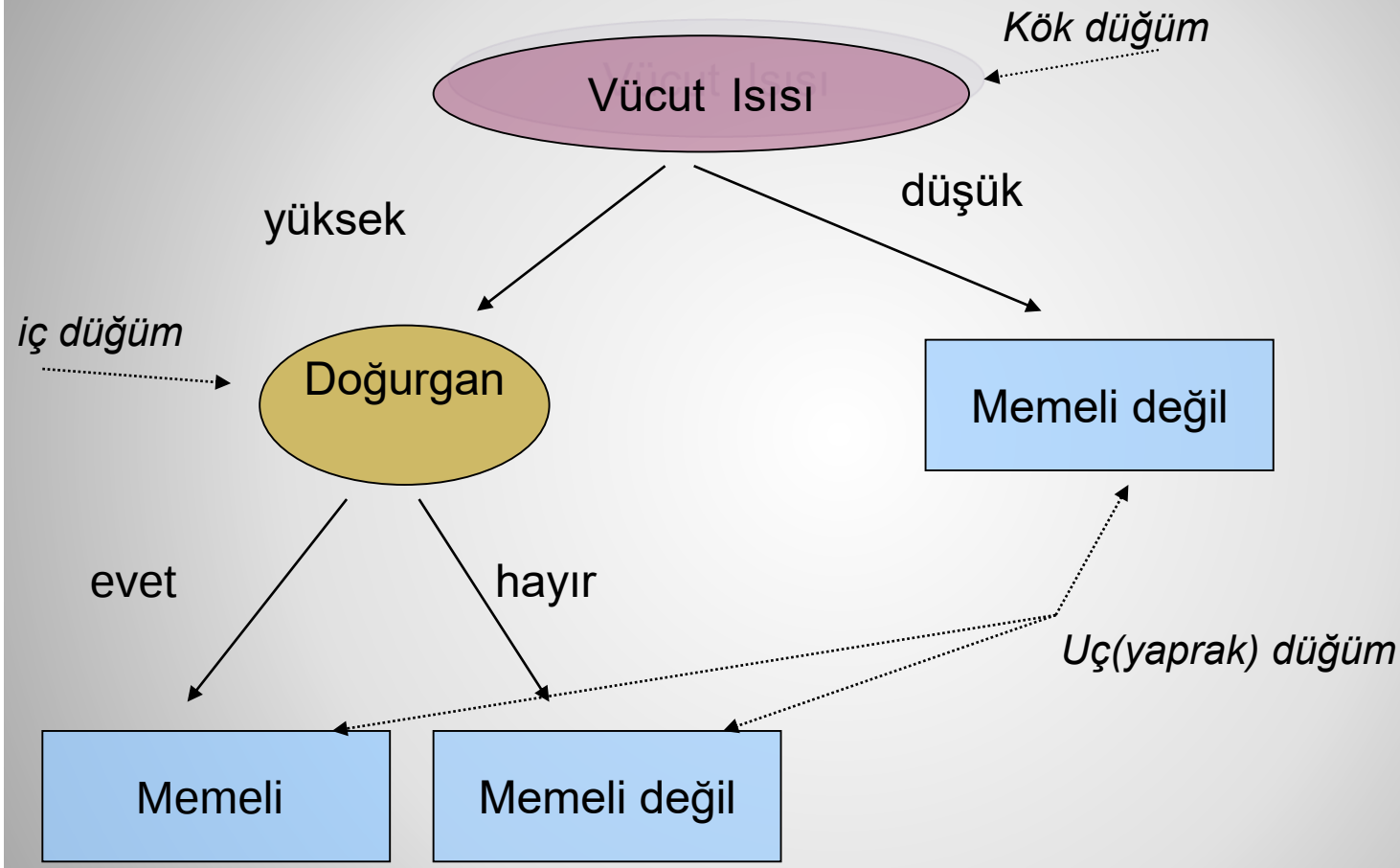
Eğer $x_1 > q_1$ ve $x_2 > q_2$ ise $y = 1$;
Eğer $x_1 > q_1$ ve $x_2 <= q_2$ ise $y = 0$;
Eğer $x_1 <= q_1$ ise $y = 0$;

x_1 : yıllık gelir
x_2 : birikim
$y = 0$: Onaysız
$y = 1$: Onaylanmış

- **Karar Ağacı**
  - Yaygın kullanılan öngörü yöntemlerinden bir tanesidir
  - Ağaçtaki her düğüm bir özellikteki testi gösterir.
  - Düğüm dalları testin sonucunu belirtir.
  - Ağaç yaprakları sınıf etiketlerini içerir.

- **Karar ağacı çıkarımı iki aşamadan oluşur**
  1. Ağaç inşası
     - Başlangıçta bütün öğrenme örnekleri kök düğümdedir.
     - Örnekler seçilmiş özelliklere tekrarlamalı olarak bölünür.
  2. Ağaç Temizleme (Tree pruning)
     - Gürültü ve istisna kararları içeren dallar belirlenir ve kaldırılır.

- **Karar ağacı kullanımı:** Yeni bilinmeyen örneğin sınıflandırılması
  - Bilinmeyen örneğin özellikleri karar ağacında test edilerek sınıfı bulunur.


Kök düğüm
Vücut Isısı
yüksek
düşük
iç düğüm
Doğurgan
Memeli değil
evet
hayır
Uç(yaprak) düğüm
Memeli
Memeli değil

Karar ağaçlarında en önemli aşamalarından biriside düğüm noktalarına ait kriterlerin belirlenmesidir.

Her düğüm noktası için bir karar ağacı algoritması tasarlanır.

Algoritmalar gruplanırsa ;

- Sınıflandırma ve regresyon ağaçları
- Entropiye dayalı algoritmalar
- Bellek tabanlı sınıflandırma algoritmaları

Sınıflandırma ve regresyon ağaçları konusunda Twoig ve Gini algoritması entropiye dayalı algoritmalara örnek olarak ise ID3 ve C4.5 algoritmaları verilebilir.

ID3 algoritmasının temeli:

Karar ağacında,

Her bir düğüm hedef-olmayan bir niteliğe, Düğümler arasındaki her yay (arc) ise niteliğin olası bir değerine karşılık gelir.

Ağacın bir yaprağı, bu yapraktan köke kadar ki yolda tanımlanan kayıtlar için hedef niteliklerin beklenen değerini belirler.

Karar ağacında her bir düğüm kökten başlayarak yol üzerinde henüz dikkate alınmamış olan nitelikler arasından en çok bilgi sağlayan hedef-olmayan nitelikle ilişkilendirilebilir.

Bu durum "İyi" bir karar ağacın nasıl olduğunu gösterir.

Entropi bir düğümün ne kadar bilgi verici olduğunu ölçmede kullanılır. Bu "İyi" ile ne kastedildiğini belirtir.

ID3 algoritması : ID3, verilen hedef-olmayan nitelik kümesi C1, C2,.....,Cn, hedef nitelik C, ve bir öğrenme kümesi ile bir karar ağacı kurmak için kullanılır.

Fonksiyon ID3
(**R:** Hedef-olmayan nitelikler kümesi, **C**: Hedef niteliği, **S**: Bir eğitim kümesi) // *returns karar ağacı*
Başla
Eğer (S == boş) {
kök="yanlış"; Döndür kök };
Eğer (S, hedef nitelik için aynı değere sahip kayıtlardan oluşuyorsa) {
kök= aynı olan bu değer; Döndür kök };
Eğer (R boşsa) {
kök=S'nin kayıtlarında hedef niteliğin değerlerinde en sık bulunan değer; Döndür kök};

D, R'deki nitelikler içinden en yüksek Kazanç(D,S) OLSUN;
$\{d^j / j=1,2,....,m\}$ D niteliğinin değerleri OLSUN;
$\{s^j / j=1,2,....,m\}$ D özelliği için d^j değerli kayıtları sırasıyla içeren S'nin altkümeleri OLSUN;
Döndür (D etiketli köke ve sırasıyla
ID3 (R - {D}, C, S^1), ID3 (R-{D}, C, S^2),....., ID3 (R-{D}, C, S^m)
ağaçlarına giden $d^1,d^2,....,d^m$ etiketli yayları olan ağacı)
Bitir ID3;

ID3 algoritması entropi(belirsizlik) zemininde oluşturulmuş bir algoritmadır.

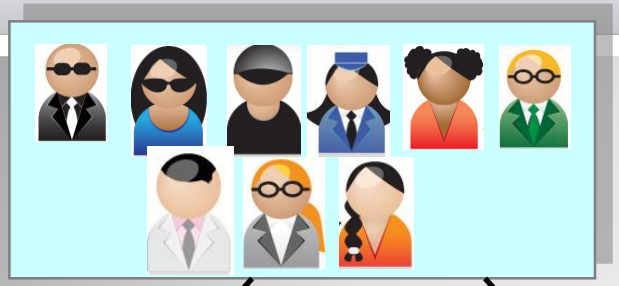
Karar ağaçlarında hangi niteliğe karşı dallanmanın yapılacağını belirlemek üzere entropi kavramına başvurulur.

Entropi : R bir kaynak olsun. Bu kaynağın $\{m_1, m_2, m_3, \ldots, m_n\}$ olmak üzere n mesaj üretilebildiğini varsayalım. Tüm mesajlar birbirinden bağımsız olarak üretilmektedir ve m_j mesajların üretilme olasılıkları p_j 'dir .

P={$p_1, p_2, p_3, \ldots, p_n$} olasılık dağılımına sahip mesajları üreten **R** kaynağın entropisi **H(R)** şu şekildedir.

$$H(R) = -\sum_{i=1}^{n} p_i \log_2(p_i) \qquad H(R) = \sum_{i=1}^{n} pi \log(1/pi)$$

Kişiler	Saç Uzunlukları (inç)	Ağırlık	yaş	Sınıf
 Hasan	0	250	36	E
 Meral	10	150	34	K
 Bahadır	2	90	10	E
 Lale	6	78	8	K
 Melike	4	20	1	K
 Ali	1	170	70	E
 Selma	8	160	41	K
 Osman	10	180	38	E
 Kemal	6	200	45	E
 Cemal	8	290	38	?

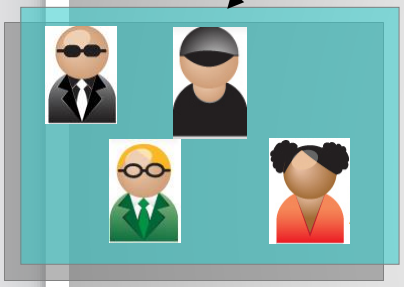


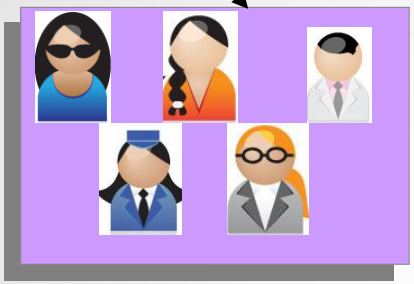
$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

evet

hayır

Saç uzunluğu <= 5?





$$Kazanç(A) = E(genelküme) - \sum E(tümaltkümeler)$$

Kazanç(Saç Uzunluğu <= 5) = **0.9911** – (4/9 * **0.8113** + 5/9 * **0.9710**) = **0.0911**

Entropy(4K,5E) = -(4/9)log2(4/9) - (5/9)log2(5/9)
= 0.9911

Entropy(1K,3E) = -(1/4)log2(1/4) - (3/4)log2(3/4)
= 0.8113

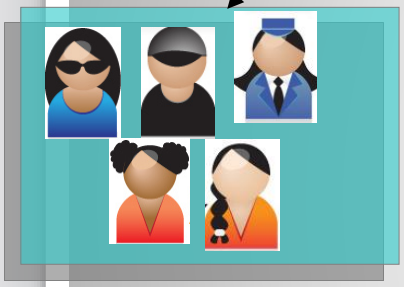
Entropy(3K,2E) = -(3/5)log2(3/5) - (2/5)log2(2/5)
= 0.9710

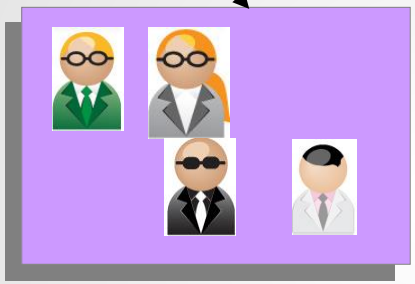
$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

evet

hayır

Ağırlık <= 160?





$$Kazanç(A) = E(genelküme) - \sum E(tümaltkümeler)$$

$Kazanç$(Ağırlık <= 160) = **0.9911** – (5/9 * **0.7219** + 4/9 * **0**) = **0.5900**

Entropy(4K,5E) = -(4/9)log2(4/9) - (5/9)log2(5/9)
= 0.9911

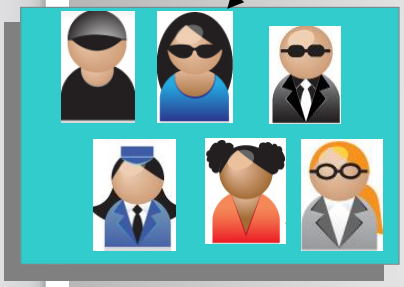
Entropy(4K,1E) = -(4/5)log2(4/5) - (1/5)log2(1/5)
= 0.7219

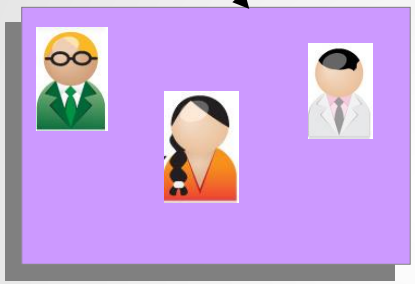
Entropy(0K,4E) = -(0/4)log2(0/4) - (4/4)log2(4/4)
= 0

evet

Yaş<= 40?

hayır





$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

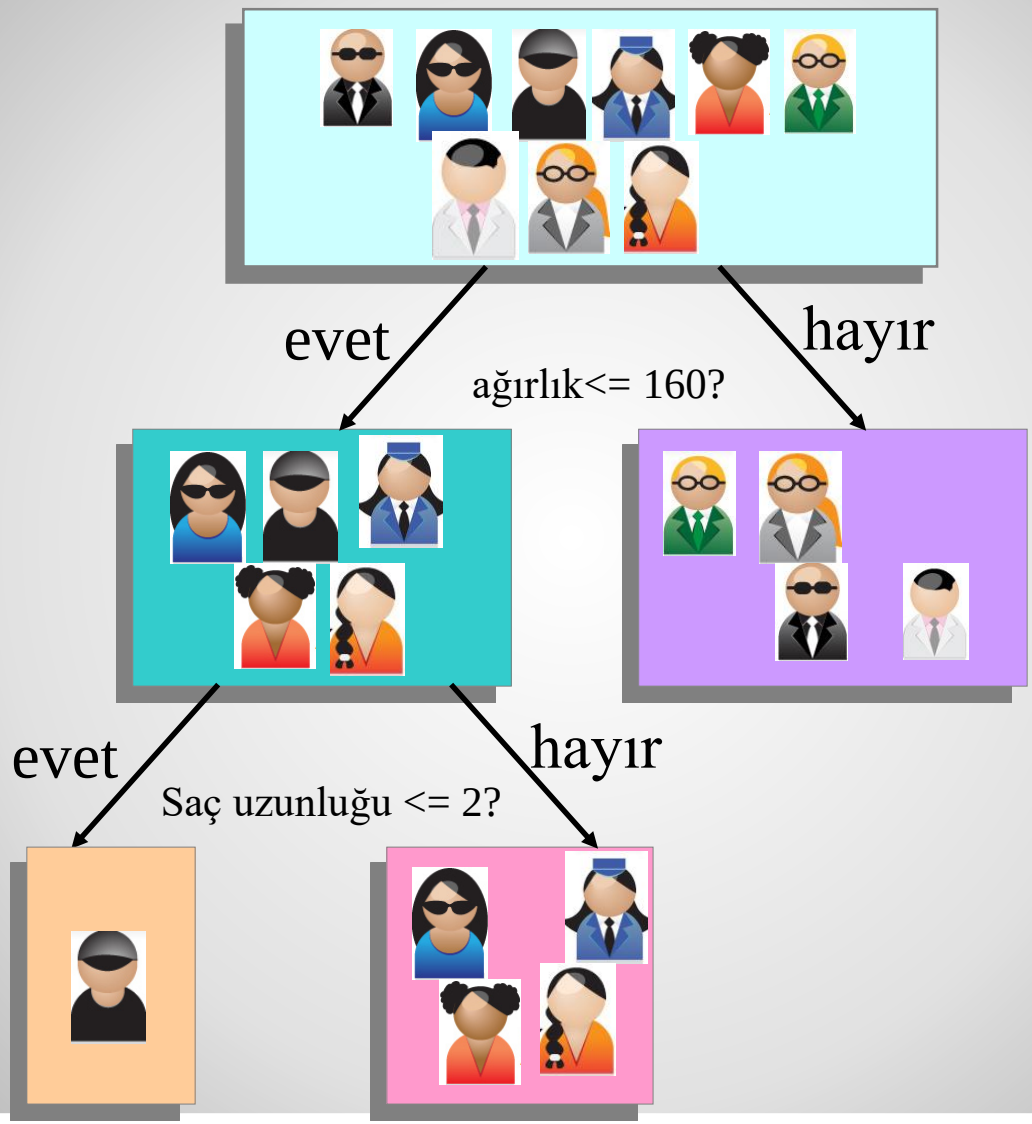
$$Kazanç(A) = E(genelküme) - \sum E(tümaltkümeler)$$

$Kazanç$(Yaş<= 40) = **0.9911** – (6/9 * **1** + 3/9 * **0.9183**) = **0.0183**

Entropy(4K,5E) = -(4/9)log2(4/9) - (5/9)log2(5/9)
= 0.9911

Entropy(3K,3E) = -(3/6)log2(3/6) - (3/6)log2(3/6)
= 1

Entropy(1K,2E) = -(1/3)log2(1/3) - (2/3)log2(2/3)
= 0.9183


evet
hayır
ağırlık<= 160?
evet
hayır
Saç uzunluğu <= 2?

Karar ağacı şu şekilde kesinleşir

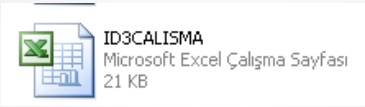
If *ağırlık* > 160, sınıfı **Erkek**
Elseif *Saç Uzunluğu*<= 2, sınıfı **Erkek**
Else sınıfı **Kadın**

	Cemal	8	290	38	?

Excelde entropi hesabı

ID3CALISMA
Microsoft Excel Çalışma Sayfası
21 KB

Excelde entropi hesabı



Karar ağaçlarının budanması :

Amaç : Karmaşık olmayan ağaçlar oluşturmak.

Ağacın budanması, bütün bir alt ağacın yerine bir yaprak düğümünün yerleştirilmesiyle yapılır. Yerleştirme ancak bir alt ağaçtaki beklenen hata tek yapraktakinden daha büyükse ancak yapılır.

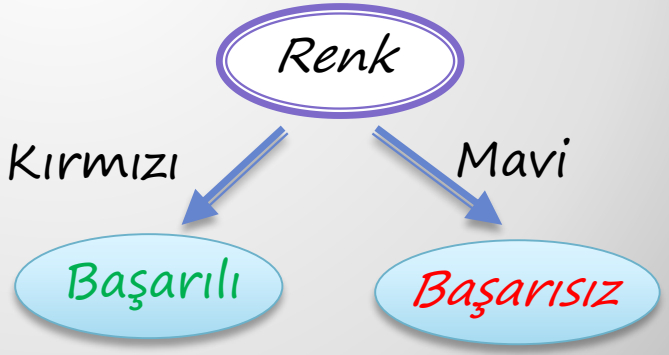
Alt ağacın yerine yaprak yerleştirmekle ,algoritma "öngörülü hata oranını" azaltmayı ve sınıflandırma modelinin kalitesini arttırmayı amaçlar.

Örnek:

Aşağıdaki verilen basit karar ağacı; 1 kırmızı başarılı öğrenme kaydı ile 2 mavi başarısız öğrenmeden elde edilir ve sonra test dizininde 3 kırmızı başarısız ve 1 mavi başarılı bulunursa, şekildeki ağaç tek bir başarısız düğüm ile değiştirilir.

Değişimden sonra dört hata yerine yalnızca iki hata yapılmış olunacaktır.

Basit Karar Ağacı

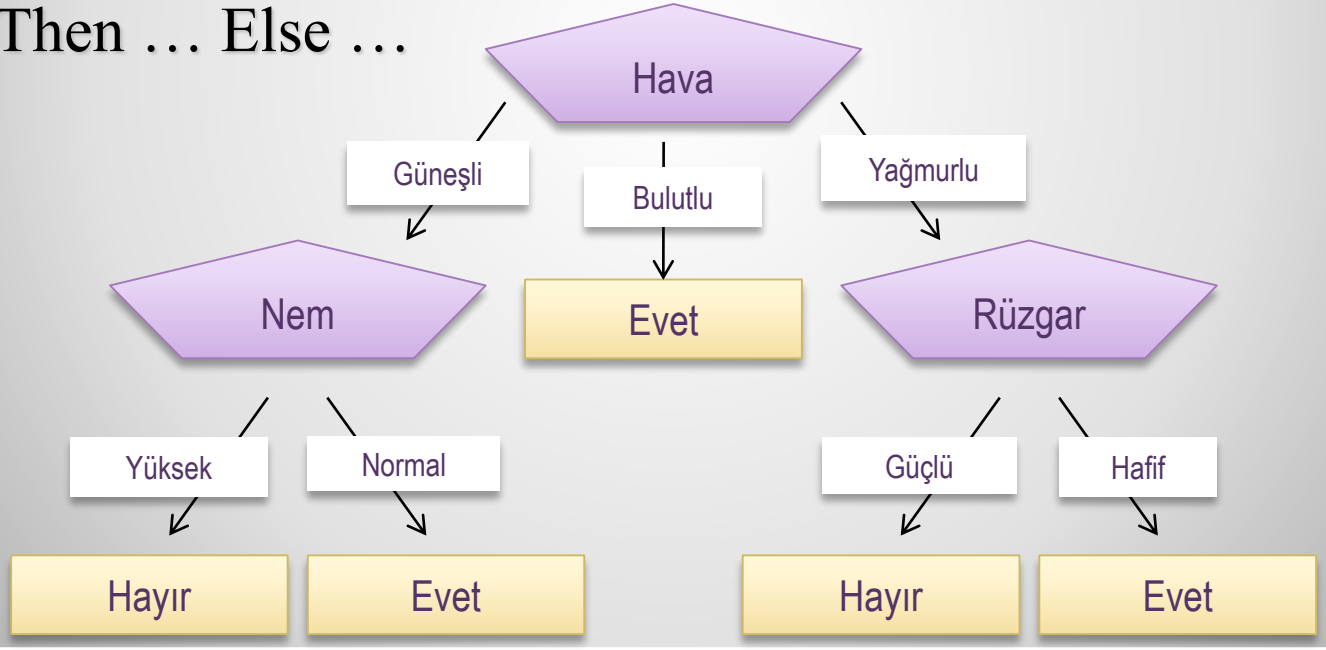


Karar Kuralları Oluşturma :

Eğitim kümesine bağlı olarak elde edilen karar ağacından yararlanarak karar kuralları oluşturulabilir.

Kurallar karşılaştırma işlemlerine benzerler;

If … Then … Else …



Karar ağacından yararlanarak aşağıdaki kuralları yazabiliriz.

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 ;

Kaynaklar :

- Veri Madenciliği DR Gökhan Silahtaroğlu 06'2008
- Veri Madencilği Yöntemleri Dr. Yalçın Özkan 06'2008
- Fatih Aydoğan H.Ü. YLTezi 2003
- M.A.Duchaineau, M.Wolinsky, D.E.Sigeti, M.C. Miller, C. Aldrich and M.B.Mineev - Weinstein,
- "ROAMingTerrain: Real-time Optimally Adapting Meshes". IEEE Visualization'97, 81–88. Nov. 1997
- Kitap : Introduction to Data Mining, Pang-Ning Tan, Michigan State University, Michael Steinbach, University of Minnesota, Vipin Kumar, University of Minnesota
- Business Intelligence and Data Mining, Prof. Dr. Haldun Akpınar, Dönence Basın ve Yayın Hizmetleri, İstanbul, 2004