

SAKARYA ÜNİVERSİTESİ

Veri Madenciliği Uygulamaları

Hafta 6

Yrd. Doç.Dr. Nilüfer YURTAY



Sınıflandırma- İstatistiğe dayalı algoritmalar (Regresyon Ağaçları)

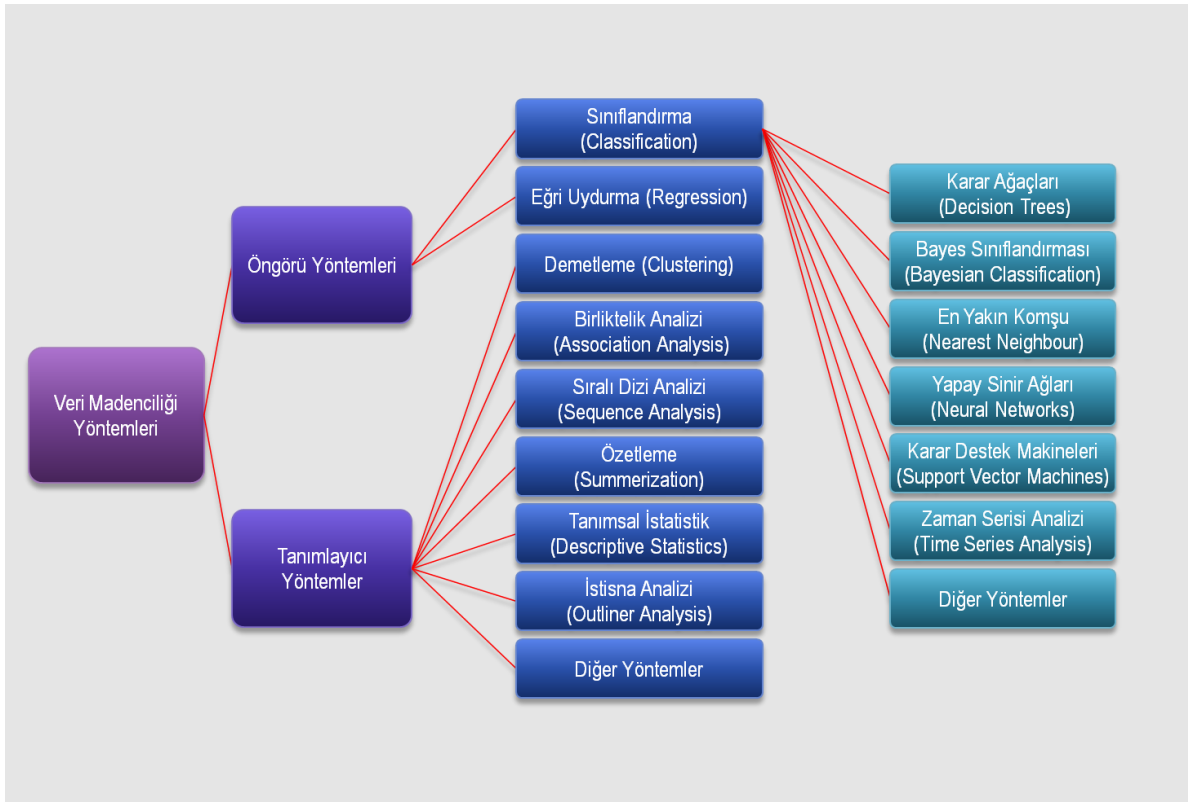
5.1 Giriş

Genel olarak veri madenciliği yöntemleri iki sınıfa ayrılabilir(Şekil 5.1):

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

Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi sınıflandırma olarak anılır. Karar ağaçları sınıflandırma yöntemlerinden biridir. Karar ağaçları oluşturmak için entropiye dayalı farklı algoritmalar geliştirilmiştir.

- ❖ Sınıflandırma ve regresyon ağaçları
- ❖ Bellek tabanlı sınıflandırma modeli
- ❖ ...



Şekil 5.1 Veri Madenciliği Yöntemleri

Sınıflandırma ve Sınıflandırma Sürecinde verilerin içerdiği ortak özellikler kullanılarak söz konusu verileri sınıflandırmak mümkündür.

Sınıflandırma bir öğrenme algoritmasına dayanır. Amaç sınıflandırma modelinin oluşturulmasıdır. Sınıfı bilinmeyen herhangi bir verinin sınıfının belirlenmesi sürecidir denebilir.

Yöntem çok çeşitli alanlarda kullanılabilir,

- ❖ Bankada (Kredi taleplerinin değerlendirilmesinde)
- ❖ Finansal (Pazardaki eğilimlerin ayrıştırılmasında)
- ❖ ...

Giriş yapılan her kayıt kümesi bir öğrenme kümesi olarak düşünülebilir. Kayıt kümesinde her kayıt belirli özellikler (Attribute) taşır. Kayıta ait alan isimlerinden bir tanesi de sınıftır(Class). Kayıtlara ait diğer özelliklerden sınıf özelliğini öngörebilecek bir model fonksiyon geliştirilir.

Amaç; yeni bir kayıt geldiğinde, bu kayıt geliştirilen model kullanılarak mümkün olduğunca doğru bir sınıfa atanmasıdır.

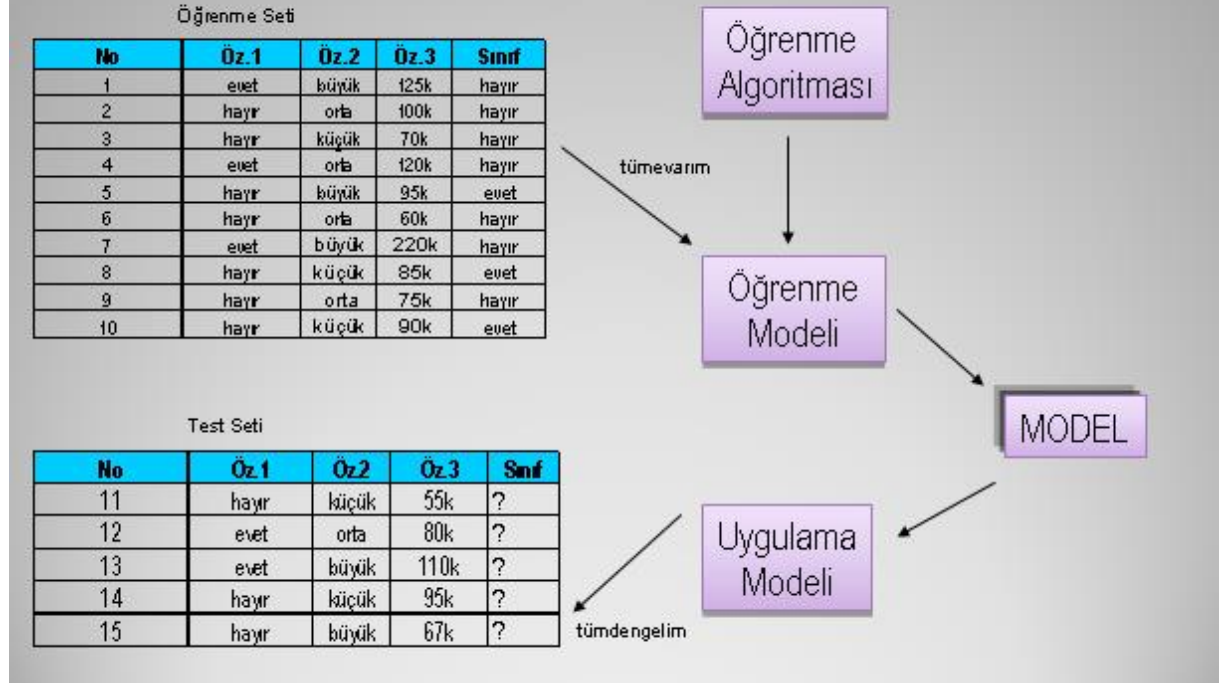
Bir deneme kümesi modelin doğruluğunu belirlemek için kullanılır. Genellikle verilen veri kümesi **öğrenme** ve **deneme kümesi** olarak ikiye ayrılır. Öğrenme kümesi modelin oluşturulmasında, deneme kümesi modelin doğrulanmasında kullanılır.

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. Veriler veri tabanı üzerinden rastgele seçilerek oluşturulurlar. Oluşturulan eğitim verileri üzerinde algoritma üzerine çalışma yapılarak sınıflama modeli elde edilir

2. Modelin Öngörü İçin Kullanılması: Eğitim verileri üzerinde sınıflandırma kuralları belirlenir. Testler uygulanarak kurallar kontrol edilir(Şekil 5.2).

Sınıflandırma süreci



Şekil 5.2 Sınıflandırılmış bir modelin oluşturulma süreci

Verilerin sınıflandırma yöntemlerinden biride karar ağaçlarıdır. Karar ağaçlarının oluşturulmasında çok sayıda öğrenme yöntemi mevcuttur.

5.2 Karar Ağaçları

Karar ağaçları, 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.

❖ 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.

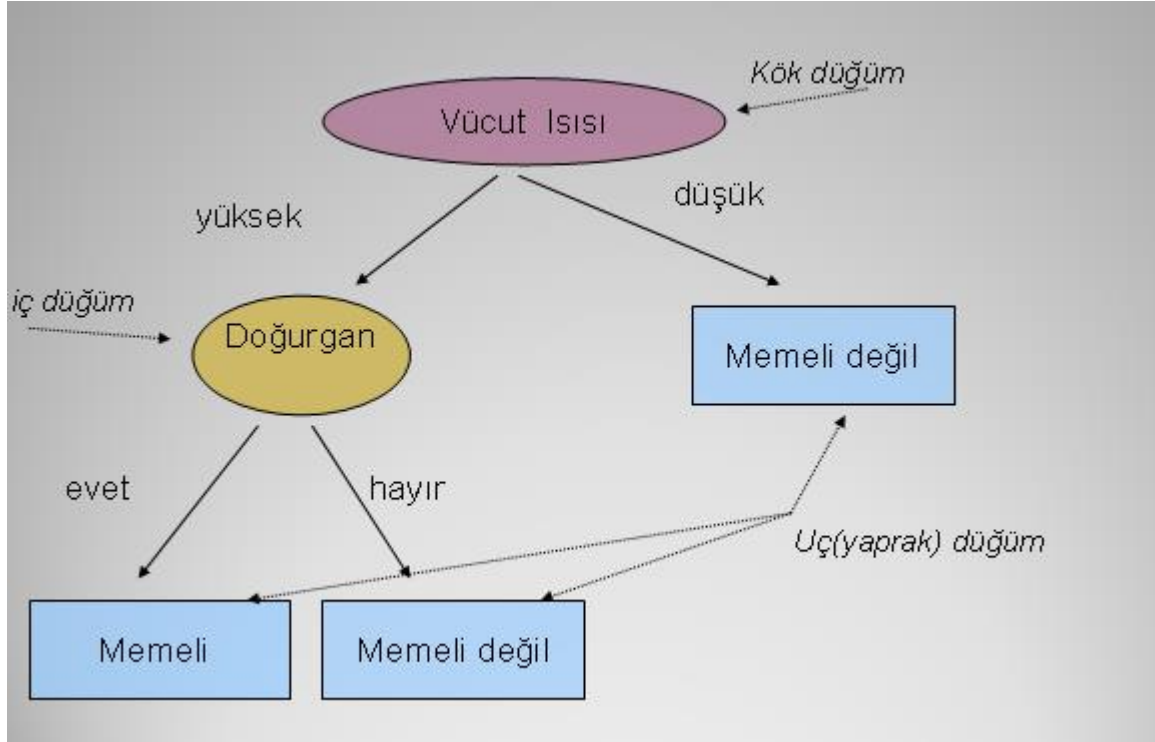
❖ Karar ağacı çıkarımı iki aşamadan oluşur

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.

Şekil 5.3 de örnek bir karar ağacı verilmektedir.



Şekil 5.3 Örnek Karar Ağacı

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.

5.3 ID3 ve C4.5 karar ağaçlarının kurulması

Karar ağaçları olarak da adlandırılan ID3 ve C4.5 algoritmaları, sınıflandırma modellerini işlemek için Quinlan (1993) tarafından geliştirilmiştir.

C4.5, ID3'ün geliştirilmiş halidir. C4.5 eksik ve sürekli nitelik değerlerini ele alabilmekte, karar ağacının budanması ve kural çıkarımı gibi işlemleri yapabilmektedir.

Karar ağacının kurulması için kullanılacak girdi olarak bir dizi kayıt verilirse bu kayıtlardan her biri aynı yapıda olan birtakım nitelik/değer çiftlerinden oluşur.

Bu niteliklerden biri kaydın hedefini belirtir. Problem, hedef-olmayan nitelikler kullanılarak hedef nitelik değerini doğru kestiren bir karar ağacı belirlemektir.

Hedef nitelik çoğunlukla sadece {evet, hayır}, veya {başarılı, başarısız} gibi ikili değerler alır.

5.3.1 ID3 Algoritması

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 nitelik ile ilişkilendirilebilir.

Bu durum “İyi” bir karar ağacının nasıl olduğunu gösterir.

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

ID3, verilen hedef-olmayan nitelik kümesi C_1, C_2, \dots, C_n , 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 ($S == \text{boş}$) {

$kök = \text{“yanlış”}$; Döndür $kök$;

Eğer (S , hedef nitelik için aynı değere sahip kayıtlardan oluşuyorsa) {

$kök = \text{aynı olan b-bu değer}$; 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;

{ $d_j / j=1, 2, \dots, m$ } D niteliğinin değerleri OLSUN;

{ $s_j / j=1, 2, \dots, m$ } D özelliği için d_j değerli kayıtları sırasıyla içeren S 'nin altkümeleri OLSUN;

Döndür (D etiketli köke ve sırasıyla

$ID3 (R - \{D\}, C, S_1), ID3 (R - \{D\}, C, S_2), \dots, ID3 (R - \{D\}, C, S_m)$

 ağaçlarına giden d_1, d_2, \dots, d_m etiketli yayları olan ağacı)

Bitir ID3;

C4.5 algoritması ID3 algoritmasının bir uzantısıdır. Bir karar ağacı kurarken, kazanç hesaplamasıyla eğitim kümesindeki değerler bilinmeden sadece nitelikler bilinerek işlem yapılır. Bir karar ağacı kullanımında bilinmeyen nitelik değerlerine sahip olan kayıtlar, mümkün olan sonuçların olasılıklarını tahmin ederek sınıflandırılabilir.

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 $\{m_1, m_2, m_3, \dots, m_n\}$ olmak üzere n mesaj üretilbildiğini varsayalım. Tüm mesajlar birbirinden bağımsız olarak üretilmektedir ve m_j mesajların üretilme olasılıkları p_j 'dir .

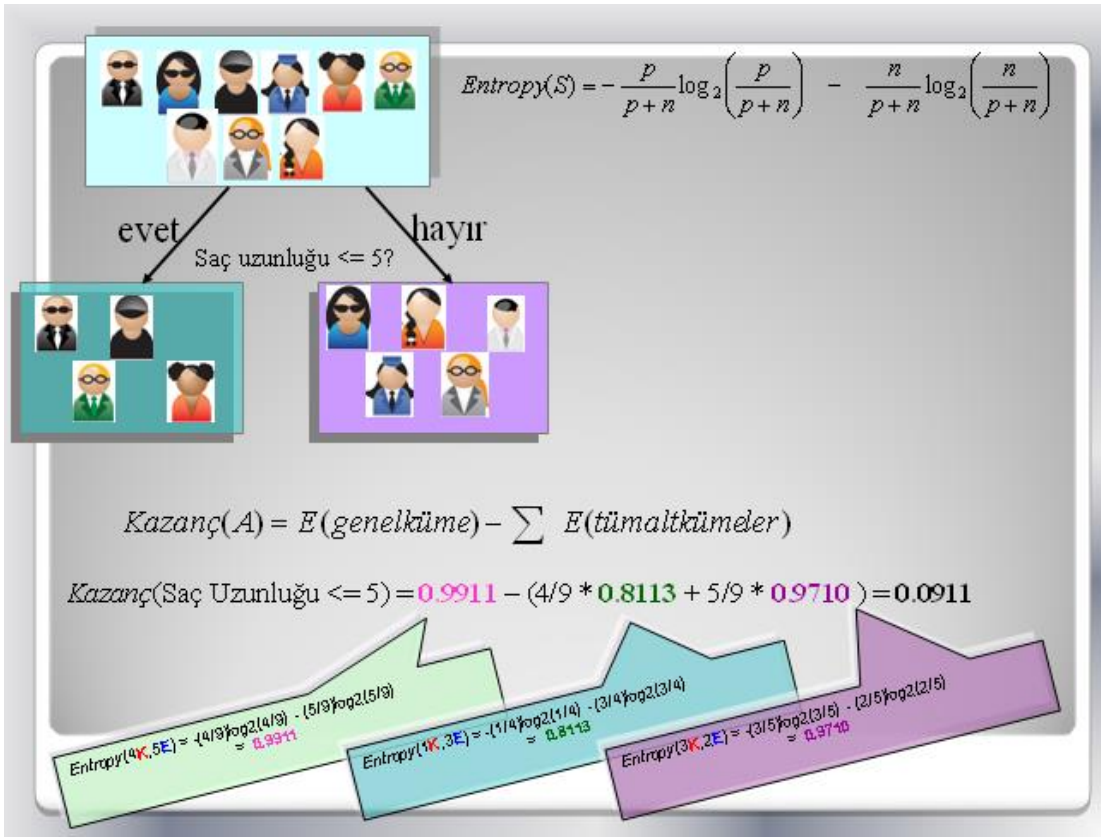
$P=\{p_1, p_2, p_3, \dots, p_n\}$ olasılık dağılımına sahip mesajları üreten R kaynağın entropisi $H(R)$ şu şekildedir.

$$H(R) = -\sum_{i=1}^n p_i \log_2(p_i) \text{ ya da } H(R) = \sum_{i=1}^n p_i \log(1/p_i) \text{ formülleri ile hesaplanabilir.}$$

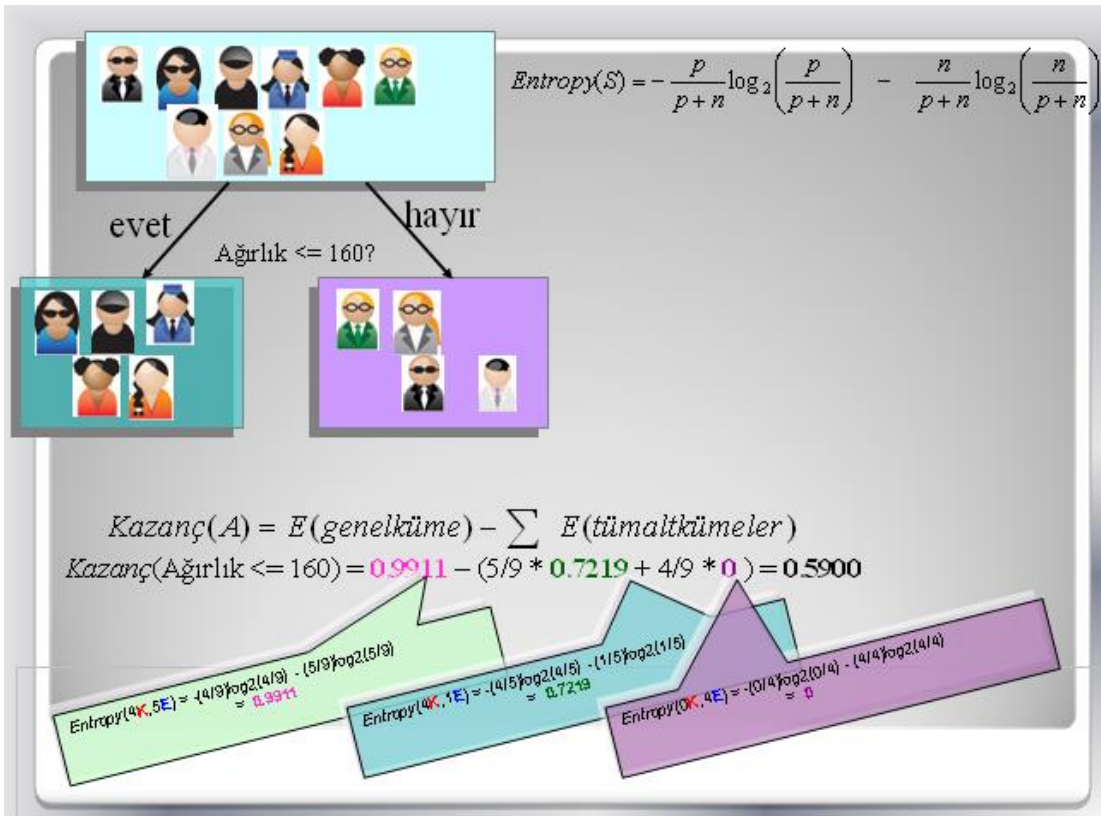
Örnek çalışma

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	?

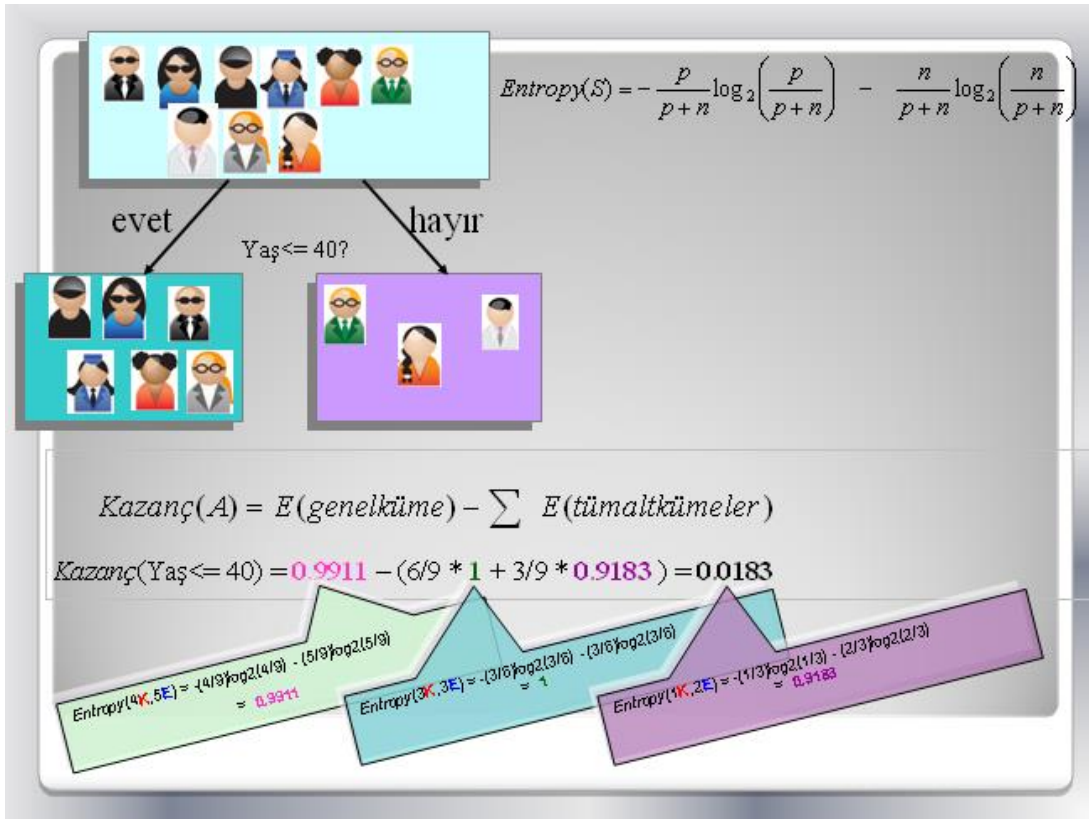
Şekil 5.4 Örnek veriler



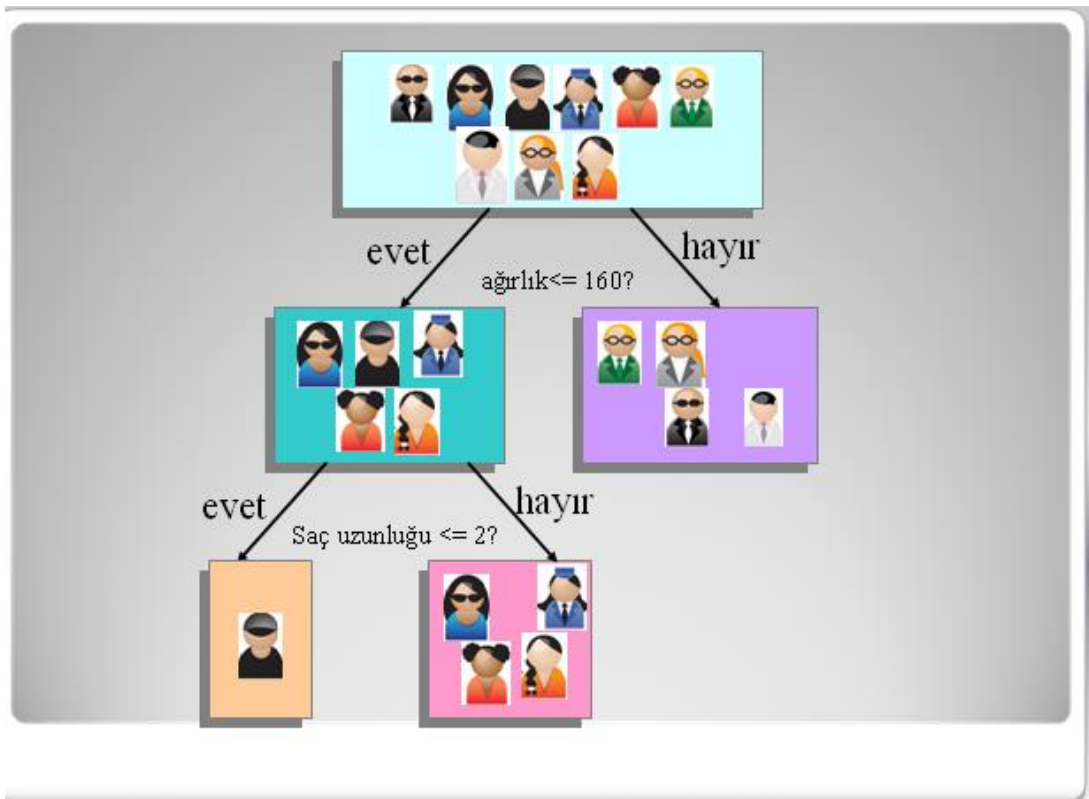
Şekil 5.5 Saç uzunluğuna göre değerlendirme



Şekil 5.6 Ağırlığa göre değerlendirme



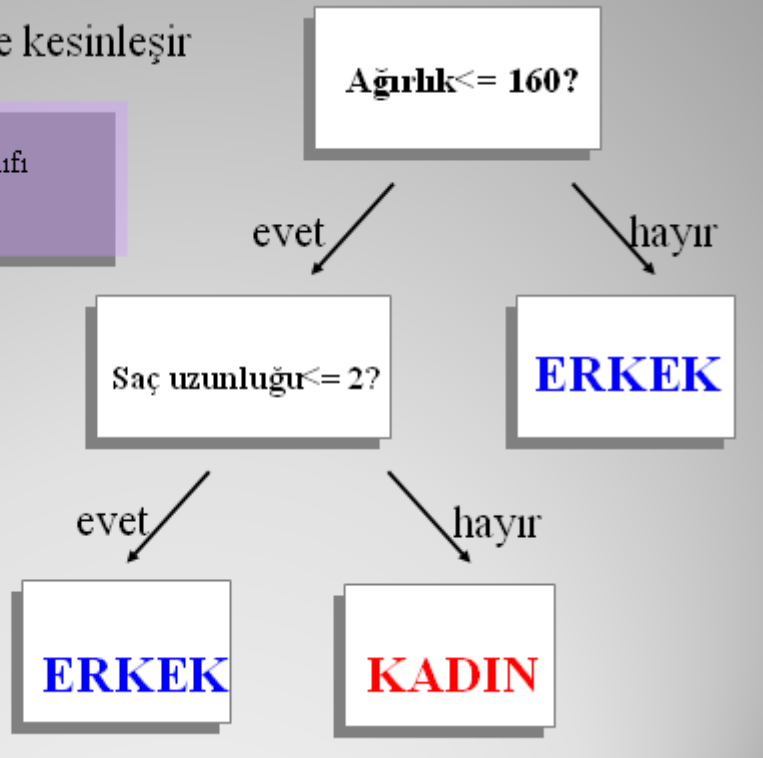
Şekil 5.7 Yaşa göre değerlendirme



Şekil 5.8 Karar ağacı

Karar ağacı şu şekilde kesinleşir

If $ağırlık > 160$, sınıfı **Erkek**
 Elseif $Saç\ Uzunluğu \leq 2$, sınıfı
Erkek
 Else sınıfı **Kadın**



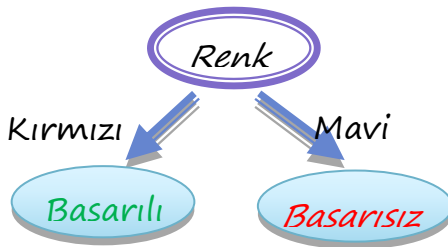
Şekil 5.9 Karar ağacı

5.4 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 yapraktan daha büyükse ancak yapılır. Alt ağacın yerine yaprak yerleştirmekle ,algoritma "öngörülü hata oranını" azaltmayı ve sınıflandırma modelinin kalitesini arttırmayı amaçlar.

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.

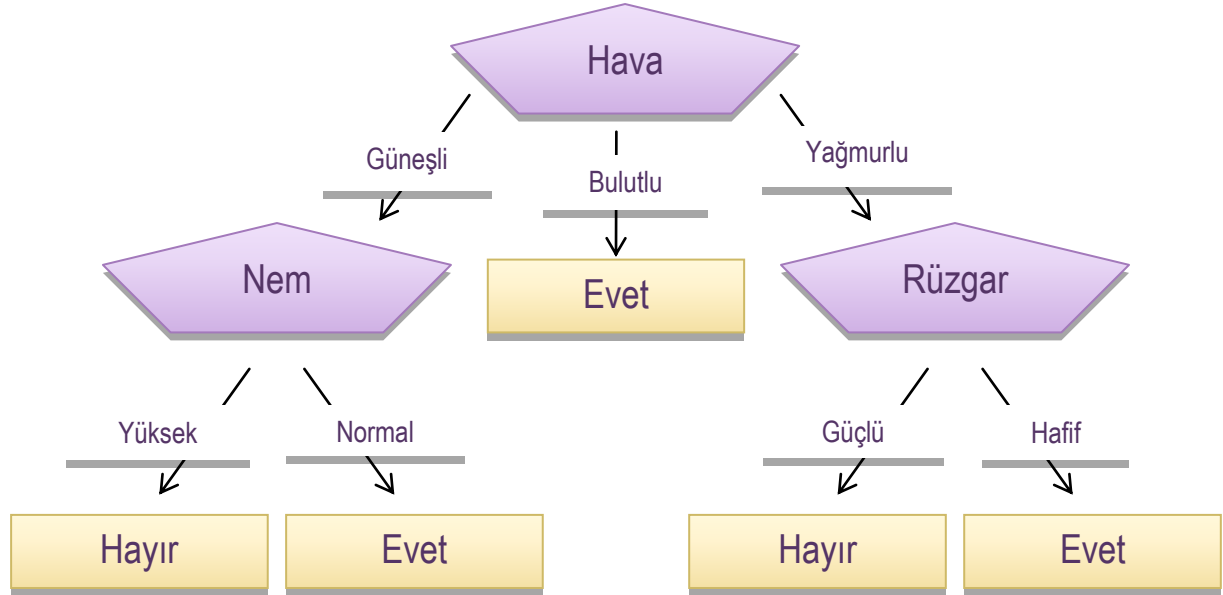
Değişimden sonra dört hata yerine yalnızca iki hata yapılmış olunacaktır.



Eđitim kümesine bađlı olarak elde edilen karar ađacından yararlanarak karar kuralları oluřturulabilir.

Kurallar karřılařtırma iřlemlerine benzerler;

If ... Then ... Else ...



řekil 5.10 Karar ađacı

Karar ađacından yararlanarak ařađıdaki kuralları yazabiliriz.

1.Kural :

Eđer Hava=Güneřli ise ve
Eđer Nem=Yüksek ise Oyun=Hayır ;

2.Kural :

Eđer Hava=Güneřli ise ve
Eđer Nem=Normal ise Oyun=Evet ;

3.Kural :

Eđer Hava=Bulutlu ise Oyun=Evet ;

4.Kural :

Eđer Hava=Yađmurlu ise ve
Eđer Rüzgar=Güçlü ise Oyun=Hayır ;

5.Kural :

Eđer Hava=Yađmurlu ise ve
Eđer Rüzgar=Hafif ise Oyun=Evet ;

Kaynaklar

- Veri Madenciliği DR Gökhan Silahtaroglu 06'2008
- Veri Madenciliği 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