VERİ MADENCİLİĞİ

(Sınıflandırma Yöntemleri)

Yrd.Doç.Dr. Kadriye ERGÜN
kergun@balikesir.edu.tr

Genel İçerik

- Veri Madenciliğine Giriş
- Veri Madenciliğinin Adımları
- Veri Madenciliği Yöntemleri
  - Sınıflandırma
  - Kümeleme
  - İlişkilendirme/birliktelik kuralları
- Metin madenciliği
- WEB madenciliği
- Veri Madenciliği Uygulamaları

İçerik

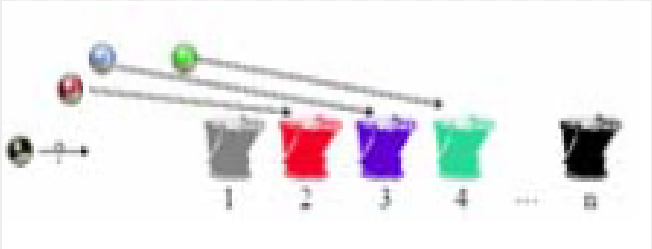
- Sınıflandırma işlemi
  - Sınıflandırma tanımı
  - Sınıflandırma uygulamaları
- Sınıflandırma yöntemleri
  - Karar ağaçları
  - Yapay sinir ağları
  - Bayes sınıflandırıcılar
  - Bayes ağları

Sınıflandırma (Classification)

- Sınıflandırma (classification) problemi:
  - nesnelerden oluşan veri kümesi (öğrenme kümesi): $D=\{t_1,t_2,\ldots,t_n\}$
  - her nesne niteliklerden oluşuyor, niteliklerden biri sınıf bilgisi
- Sınıf niteliğini belirlemek için diğer nitelikleri kullanarak bir model bulma
- Öğrenme kümesinde yer almayan nesneleri (test kümesi) mümkün olan en iyi şekilde doğru sınıflara atamak
- sınıflandırma=ayrık değişkenler için öngörüde (prediction) bulunmak.

Sınıflandırma (Classification)

- Amaç: Yeni bir kayıt geldiğinde, bu kaydı geliştirilen modeli kullanılarak mümkün olduğunca doğru bir sınıfa atamak.
  - verinin dağılımına göre bir model bulunur
  - bulunan model, başarımı belirlendikten sonra niteliğin gelecekteki
  - ya da bilinmeyen değerini tahmin etmek için kullanılır
    - model başarımı: doğru sınıflandırılmış sınama kümesi örneklerinin oranı
- Veri madenciliği uygulamasında:
  - ayrık nitelik değerlerini tahmin etmek: sınıflandırma
  - sürekli nitelik değerlerini tahmin etmek: öngörü



- Sınıflandırma: hangi topun hangi sepete koyulabileceği
- Öngörü: Topun ağırlığı

Danışmanlı & Danışmansız Öğrenme

- Danışmanlı (Gözetimli, Supervised) öğrenme= sınıflandırma
  - Sınıfların sayısı ve hangi nesnenin hangi sınıfta olduğu biliniyor.

- Danışmansız (Gözetimsiz, Unsupervised) öğrenme= kümeleme (clustering)
  - Hangi nesnenin hangi sınıfta olduğu bilinmiyor. Genelde sınıf sayısı bilinmiyor.

Sınıflandırma Uygulamaları

- Kredi başvurusu değerlendirme
- Kredi kartı harcamasının sahtekarlık olup olmadığına karar verme
- Hastalık teşhisi
- Ses tanıma
- Karakter tanıma
- Gazete haberlerini konularına göre ayırma
- Kullanıcı davranışları belirleme

Sınıflandırma için Veri Hazırlama

- Veri dönüşümü:
  - Sürekli nitelik değeri ayrık hale getirilir
  - Normalizasyon ([-1,...,1], [0,...,1])
- Veri temizleme:
  - gürültüyü azaltma
  - gereksiz nitelikleri silme

Sınıflandırma İşlemi

- Sınıflandırma işlemi üç aşamadan oluşur:
  1. Model oluşturma
  2. Model değerlendirme
  3. Modeli kullanma

Sınıflandırma İşlemi: Model Oluşturma

- 1. Model Oluşturma:
  - Her nesnenin sınıf etiketi olarak tanımlanan niteliğinin belirlediği bir sınıfta olduğu varsayılır
  - Model oluşturmak için kullanılan nesnelerin oluşturduğu veri kümesi öğrenme kümesi olarak tanımlanır
  - Model farklı biçimlerde ifade edilebilir
    - IF – THEN – ELSE kuralları ile
    - Karar ağaçları ile
    - Matematiksel formüller ile

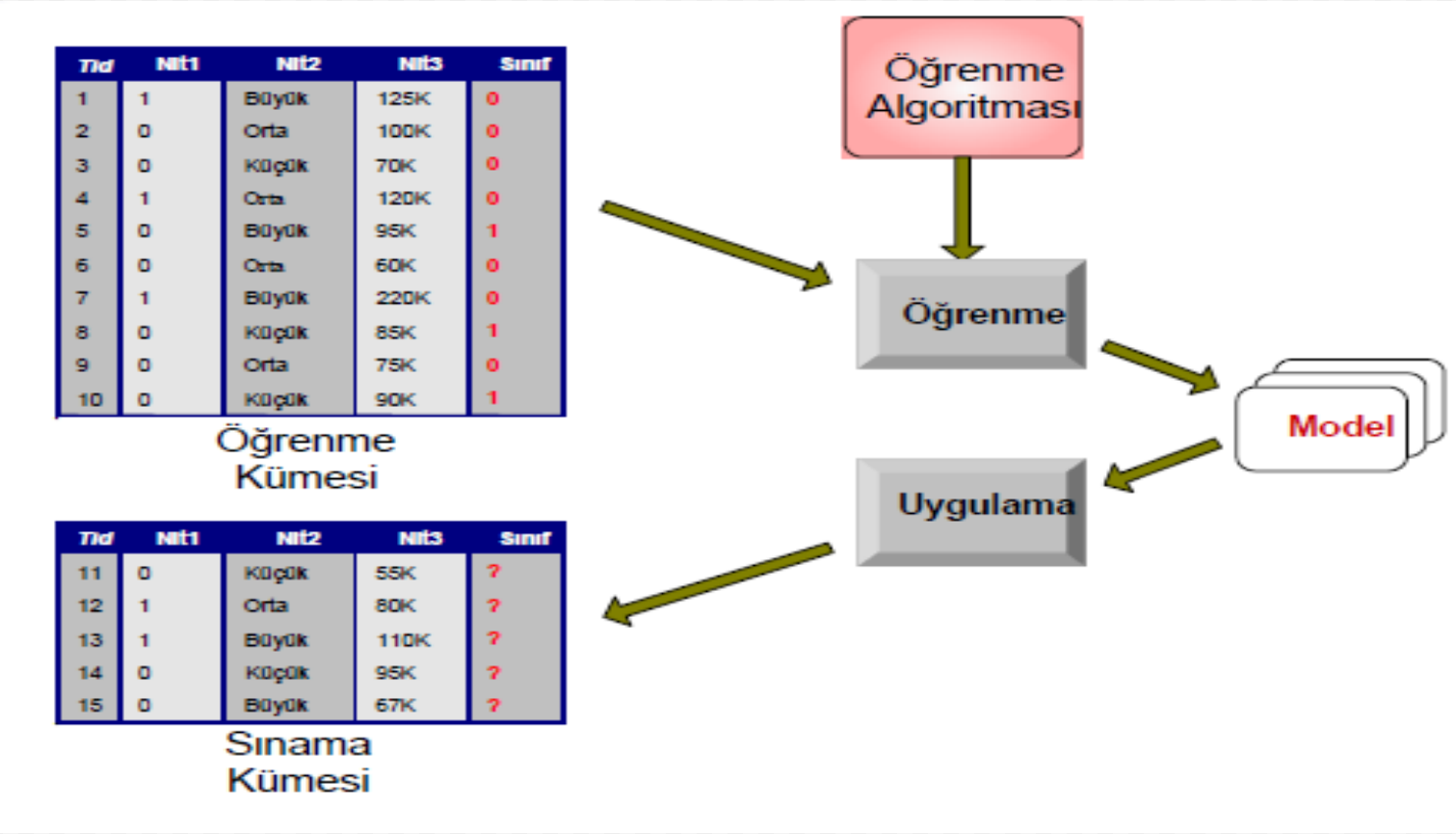
Sınıflandırma İşlemi: Model Değerlendirme

- 2. Model Değerlendirme:
- Modelin başarımı (doğruluğu) sınama kümesi örnekleri kullanılarak belirlenir.
- Sınıf etiketi bilinen bir sınama kümesi örneği model kullanılarak belirlenen sınıf etiketiyle karşılaştırılır.
- Modelin doğruluğu, doğru sınıflandırılmış sınama kümesi örneklerinin toplam sınama kümesi örneklerine oranı olarak belirlenir.
- Sınama kümesi model öğrenirken kullanılmaz.

Sınıflandırma İşlemi: Modeli Kullanma

- 3. Modeli kullanma:
  - Model daha önce görülmemiş örnekleri sınıflandırmak için kullanılır
    - Örneklerin sınıf etiketlerini tahmin etme
    - Bir niteliğin değerini tahmin etme

Örnek


Tid
Nit1
Nit2
Nit3
Sınıf
1 1 Büyük 125K 0
2 0 Orta 100K 0
3 0 Küçük 70K 0
4 1 Orta 120K 0
5 0 Büyük 95K 1
6 0 Orta 60K 0
7 1 Büyük 220K 0
8 0 Küçük 85K 1
9 0 Orta 75K 0
10 0 Küçük 90K 1
Öğrenme Kümesi
Tid
Nit1
Nit2
Nit3
Sınıf
11 0 Küçük 55K ?
12 1 Orta 80K ?
13 1 Büyük 110K ?
14 0 Küçük 95K ?
15 0 Büyük 67K ?
Sınama Kümesi
Öğrenme Algoritması
Öğrenme
Model
Uygulama

Sınıflandırıcı Başarımını Değerlendirme

Doğru sınıflandırma başarısı

- Hız
  - modeli oluşturmak için gerekli süre
  - sınıflandırma yapmak için gerekli süre
- Kararlı olması
  - veri kümesinde gürültülü ve eksik nitelik değerleri olduğu durumlarda da iyi sonuç vermesi
- Ölçeklenebilirlik
  - büyük miktarda veri kümesi ile çalışabilmesi
- Anlaşılabilir olması
  - kullanıcı tarafından yorumlanabilir olması
- Kuralların yapısı
  - birbiriyle örtüşmeyen kurallar

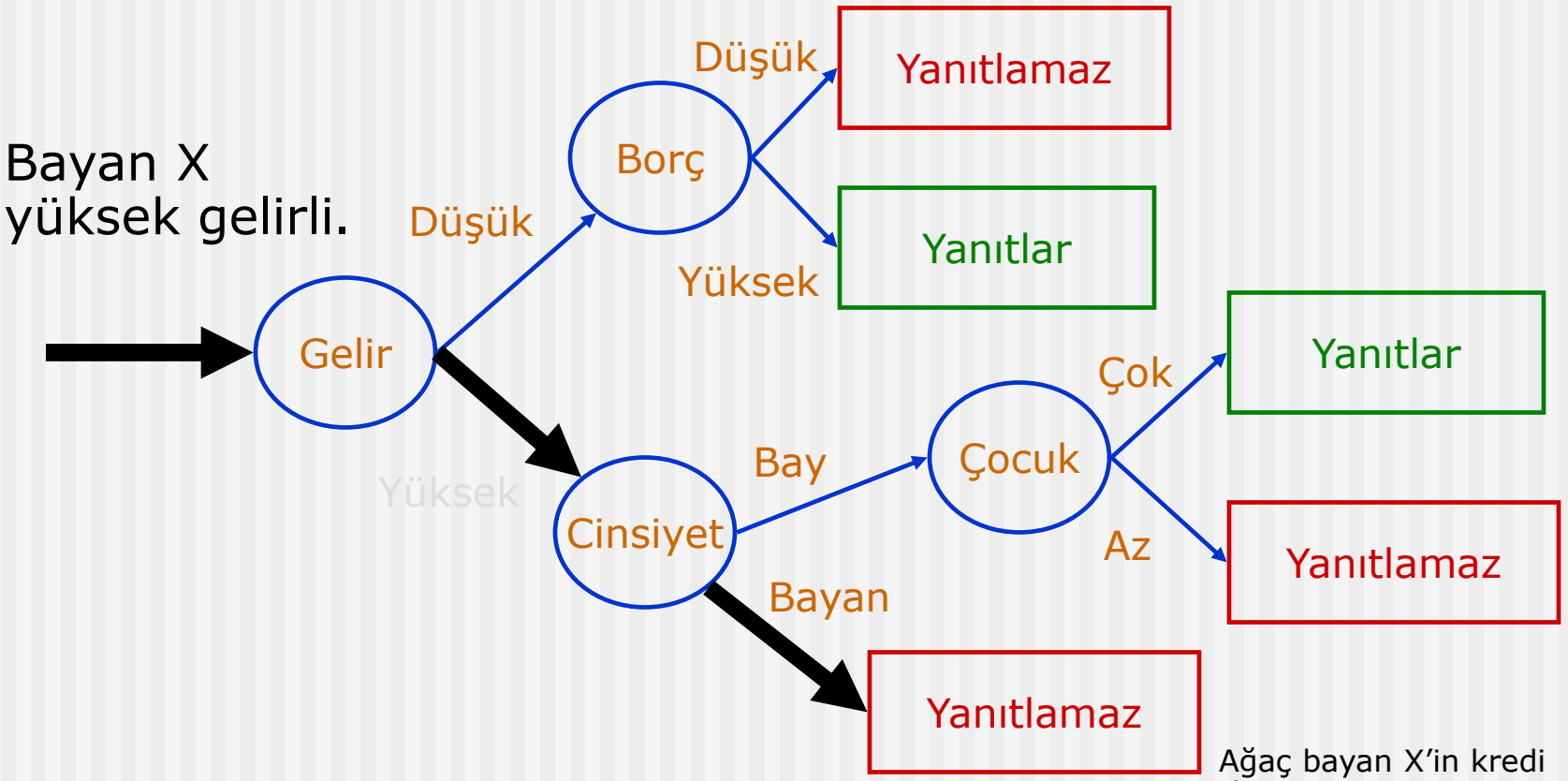
Sınıflandırma Yöntemleri

- Karar ağaçları (decision trees)
- Yapay sinir ağları (artificial neural networks)
- Bayes sınıflandırıcılar (Bayes classifier)
- İlişki tabanlı sınıflandırıcılar (association-based classifier)
- k-en yakın komşu yöntemi (k- nearest neighbor method)
- Destek vektör makineleri (support vector machines)
- Genetik algoritmalar (genetic algorithms)
- ...

Karar Ağaçları

- 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
  - Ağaç inşası
    - Başlangıçta bütün öğrenme örnekleri kök düğümdedir.
    - Örnekler seçilmiş özelliklere tekrarlamalı olarak göre bölünür.
  - Ağaç Temizleme (Budama) (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.

Bir Kredi Kartı Kampanyasında Yeni Bir Örneğin Sınıflandırılması



Ağaç bayan X'in kredi kampanyasına yanıt vermeyeceğini öngörür.

Karar Ağacı Yöntemleri

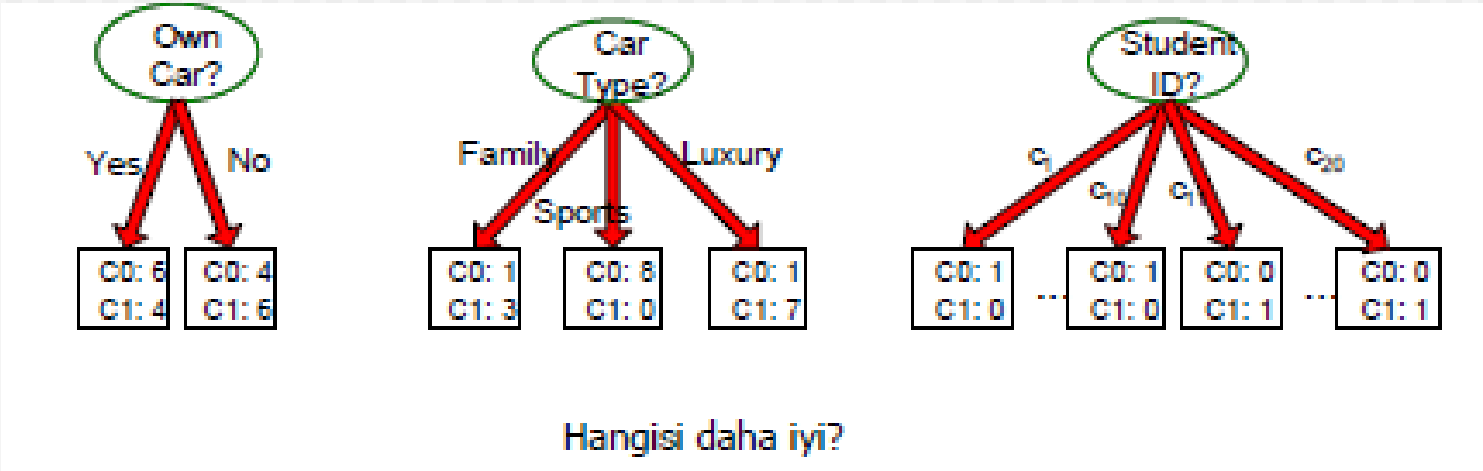
- Karar ağacı oluşturma yöntemleri genel olarak iki aşamadan oluşur:
  - 1. ağaç oluşturma
    - en başta bütün öğrenme kümesi örnekleri kökte
    - seçilen niteliklere bağlı olarak örnek yinelemeli olarak bölünüyor.
  - 2. ağaç budama
    - öğrenme kümesindeki gürültülü verilerden oluşan ve sınama kümesinde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

Karar Ağacı Oluşturma

- Yinelemeli işlem
  - ağaç bütün verinin oluşturduğu tek bir düğümle başlıyor
  - eğer örnekleri hepsi aynı sınıfa aitse düğüm yaprak olarak sonlanıyor ve sınıf etiketini alıyor
  - eğer değilse örnekleri sınıflara en iyi bölecek olan nitelik seçiliyor
  - işlem sona eriyor
    - örneklerin hepsi (çoğunluğu) aynı sınıfa ait
    - örnekleri bölecek nitelik kalmamış
    - kalan niteliklerin değerini taşıyan örnek yok

Örnekleri En İyi Bölen Nitelik Hangisi?

- Bölmeden önce:
  - 10 örnek C0 sınıfında
  - 10 örnek C1 sınıfında



En iyi Bölme Nasıl Belirlenir?

- "Greedy" (aç gözlü)yaklaşım
  - çoğunlukla aynı sınıfa ait örneklerin bulunduğu düğümler tercih edilir
- Düğümün kalitesini ölçmek için bir yöntem

En İyi Bölen Nitelik Nasıl Belirlenir?

- İyilik Fonksiyonu (Goodness Function)
- Farklı algoritmalar farklı iyilik fonksiyonları kullanabilir:
  - bilgi kazancı (information gain): ID3, C4.5
    - bütün niteliklerin ayrık değerler aldığı varsayılıyor
    - sürekli değişkenlere uygulamak için değişiklik yapılabilir
  - gini index (IBM IntelligentMiner)
    - her nitelik ikiye bölünüyor
    - her nitelik için olası bütün ikiyi bölünmeler sınanıyor

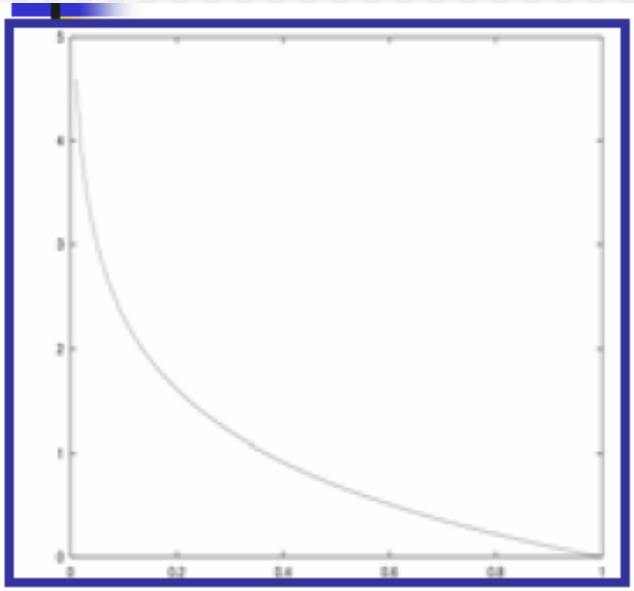
Bilgi / Entropi

- p_1, p_2,.., p_s toplamları 1 olan olasılıklar. Entropi (Entropy)

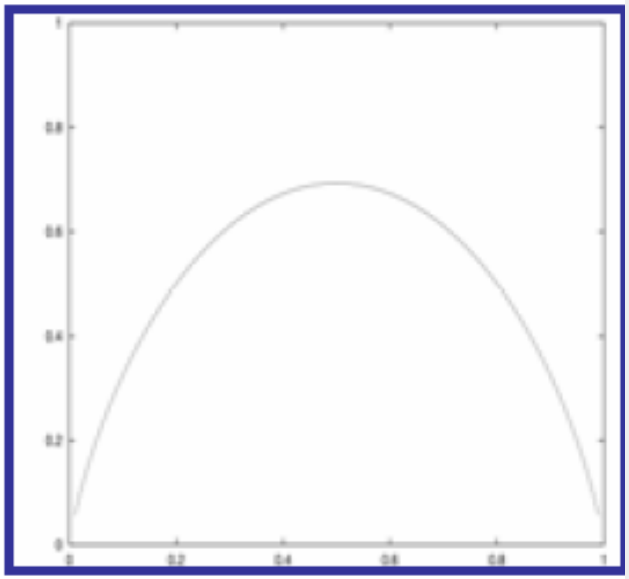
$$H(p_1, p_2, \ldots, p_s) = -\sum_{i=1}^{s} p_i \log(p_i)$$

- Entropi rastgeleliği, belirsizliği ve beklenmeyen durumun ortaya çıkma olasılığını gösterir.
- Sınıflandırmada
  - olayın olması beklenen bir durum
  - entropi=0

Entropi



log (p)



H(p,1-p)

- örnekler aynı sınıfa aitse entropi=0
- örnekler sınıflar arasında eşit dağılmışsa entropi=1
- örnekler sınıflar arasında rastgele dağılmışsa 0<entropi<1

Örnek

- S veri kümesinde 14 örnek: C0 sınıfına ait 9, C1 sınıfına ait 5 örnek.
- Entropi $$H(p_1, p_2, \ldots, p_s) = -\sum_{i=1}^{s} p_i \log(p_i)$$
- $H(p_1,p_2)$= - (9/14) Log_2 (9/14) - (5/14) Log_2 (5/14)
  = 0.940

Bilgi Kazancı (ID3 / C4.5)

- Bilgi kuramı kavramlarını kullanarak karar ağacı oluşturulur. Sınıflandırma sonucu için en az sayıda karşılaştırma yapmayı hedefler.
- Ağaç bir niteliğe göre dallandığında entropi ne kadar düşer?
- A niteliğinin S veri kümesindeki bilgi kazancı

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(S_v)$$

Values(A), *A* niteliğinin alabileceği değerler, S_v, *A=v* olduğu durumda S'nin altkümesi.