ISE 302 –Veri Madenciliği

DR. ÖĞR. ÜYESİ ESİN AYŞE ZAİMOĞLU

esinzaimoglu@sakarya.edu.tr

Karar Ağaçları

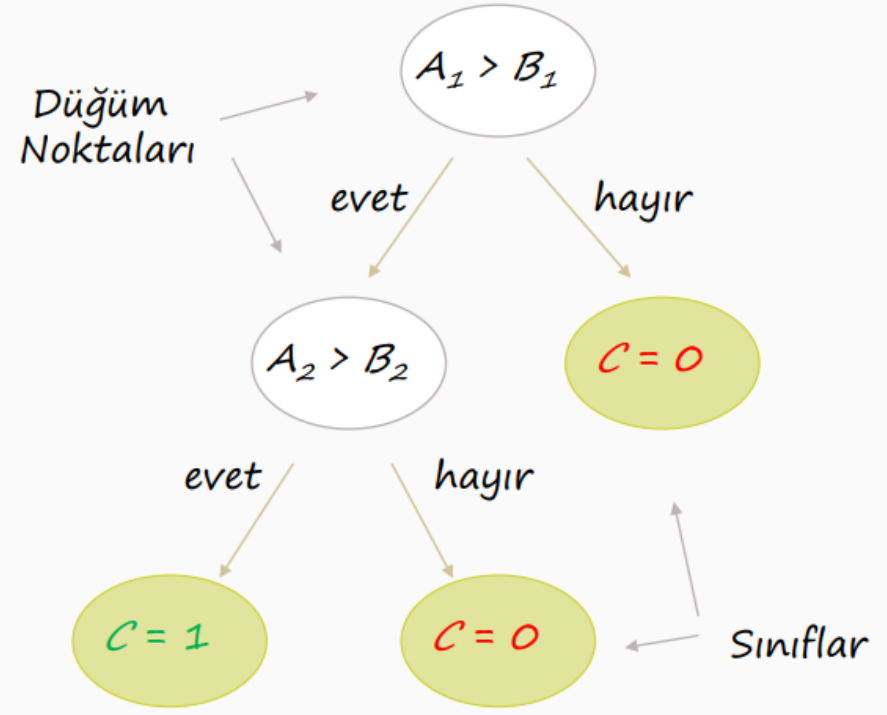
- Karar ağacı, veri madenciliği ve makine öğrenimi alanlarında kullanılan bir sınıflandırma ve regresyon yöntemidir.
- Veri setlerindeki desenleri ve ilişkileri analiz ederek, veri tabanlı kararlar almak için kullanılır.
- Ağaç mantığıyla bir model kurularak sınıflandırma işlemi yapılır.

Karar Ağaçları

Modelin eğitimi aşamasında;

- Ham veriden yukarıdan aşağıya doğru ters bir ağaç biçiminde akış diyagramları yardımıyla gösterilen bir ağaç inşa edilir.
- Bütün karar ağaçları, bir öznitelikten başlayıp gerek görülürse alt öz niteliklere bölünerek devam eden ve dalın sonunda sınıf değerine ulaşan yapıdadır.
- Karar ağacı algoritmalarında amaç ağaç dallarını budayarak, düğüm sayısını azaltarak daha hızlı ve etkin kurallara ulaşmaktır.

Test aşamasında ağaç üzerinde arama yapılarak sınıfı bilinmeyen elemanın sınıfı elde edilir.



Karar Ağaçları Yapısı

Öncelikle, veri setindeki sürekli değişken değerleri kesikli değerlere dönüştürülür.

Ağaç bütün verinin oluşturduğu tek bir düğümle başlar.

Eğer veri kayıtlarının tümü aynı sınıfa ait değil ise bu kayıtları en iyi şekilde sınıflandıracak olan öznitelik seçilir.

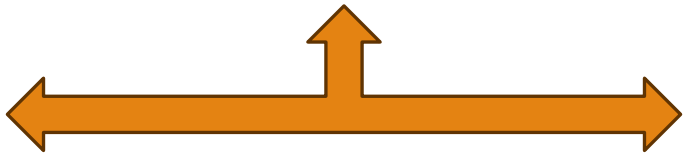
Karar ağacı algoritmaları alt düğümleri oluşturmak için bir tür kazanç fonksiyonundan faydalanır.

Karar Ağaçları ile sınıflandırmada aşağıdaki durumlarda işlem sonlandırılır.

- Veri kayıtlarının hepsi aynı sınıfa aittir.
- Kayıtları sınıflandıracak öznitelik kalmamıştır.
- Kalan özniteliklerin değerini taşıyan kayıt yoktur.

Karar Ağaçları Çeşitleri

Karar ağaçları temelde iki alt grupta incelenebilir:



Entropiye Dayalı Algoritmalar	Regresyon Ağaçları (CART)
• ID3 →Bilgi kazancı (Information Gain) • C4.5 →Kazanç Oranı (Gain Ratio)	• Twoig • Gini →Gini Indeks

Karar Ağaçları Kavramlar

Entropi, bilgi kazancı ve Gini indeksi, karar ağaçları gibi makine öğrenimi modellerinde kullanılan terimlerdir.

Bu terimler arasındaki ilişki şu şekildedir: Karar ağaçları, bilgi kazancı veyaGini indeksi gibi kriterlere dayanarak veri kümesini bölerek homojen alt kümeler elde etmeye çalışır.

Yani, karar ağaçları, entropi veGini indeksi gibi kavramları kullanarak veri kümesini bölme stratejisi belirler ve bu sayede daha homojen alt kümelere sahip olmaya çalışır.

Bilgi kazancı, entropi azaltımını veya Gini indeksi düşüşünü optimize etmeye yönelik bir kriterdir.

Karar Ağaçları Kavramlar

Entropi:

1. Entropi, bir sistemdeki belirsizliği ölçen bir kavramdır. Bir veri kümesindeki homojenlik veya heterojenlik düzeyini ifade eder.
2. Bir karar ağacının kök düğümünde, veri kümesi genellikle bir özellik (feature) üzerinde bölünür. Entropi, bu bölünmüş veri kümesindeki homojenliği veya düzensizliği ölçer.
3. Düşük entropi, daha homojen alt kümeleri gösterirken(sınıflar arasında daha az belirsizlik veya daha fazla düzen), yüksek entropi daha heterojen alt kümeleri gösterir.
4. Eğer kayıtlar belirlenen sınıflar arasında eşit dağılmış ise entropi değeri «1» olur «0» ile «1» arasındaki entropi değerleri için kayıtların sınıflar arasında rastgele dağıldığı farz edilir.

Karar Ağaçları Kavramlar

Bilgi Kazancı (Information Gain):

1. Bilgi kazancı, bir özellik (feature) üzerinde veri kümesini bölmenin, entropideki düşüşü ifade eder. Yani, bir özellik seçildiğinde, bu özelliği kullanarak veriyi böldüğümüzde entropide ne kadar bir azalma yaşandığını ölçer.
2. İdeal durumda, bilgi kazancı en yüksek olan özellik seçilir çünkü bu, veri kümesini en iyi şekilde ayıran özelliktir.
3. Yüksek bilgi kazancı, bu özelliğin sınıflandırma için önemli olduğunu gösterir.

Karar Ağaçları Kavramlar

Gini İndeksi:

1. Gini indeksi, bir veri kümesinin homojenliğini ölçen bir metriktir. Gini indeksi düşükse, veri kümesi homojen demektir.
2. Bir karar ağacında, bir özellik üzerinde veri kümesini bölme kriteri olarak kullanılan Gini indeksi, bölünmüş alt küme verilerinin ne kadar homojen olduğunu da ölçer.
3. **Kısaca ;bir düğümde rasgele bir örneğin yanlış sınıflandırılma olasılığını ölçer.**

Entropi Nasıl Hesaplanır?

Durum	Sınıf
A	1
B	1
C	0
D	1
E	0

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

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

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

Entropi Nasıl Hesaplanır?

Bu veri seti için entropi hesaplamak için şu adımları izleyebiliriz:

1. Toplam örnek sayısı: N=5
2. Sınıf 1'in sayısı: 3, Sınıf 0'ın sayısı: 2
3. Sınıf 1'in olasılığı: P1 = N1/N = 3/5
4. Sınıf 0'ın olasılığı: P0 = N0/N =2/5

Durum	Sınıf
A	1
B	1
C	0
D	1
E	0

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

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

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

Entropi formülüne bu değerleri yerine koyarak hesaplama yapalım:

$$\text{Entropi (S)} = -\left(\frac{3}{5} \cdot \log_2\left(\frac{3}{5}\right)\right) - \left(\frac{2}{5} \cdot \log_2\left(\frac{2}{5}\right)\right) = 0.971+0.529 = 1.5$$

Entropi Nasıl Hesaplanır?

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

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

ID	Yaş	Nabız	Durum	Hasta
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
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

Entropi Nasıl Hesaplanır?

Yüksek Nabız

$p_{\text{Evet}} = \frac{2}{4}$
$p_{\text{Hayır}} = \frac{2}{4}$

Entropi hesaplaması:
$\text{Entropi}_{\text{Yüksek}} = -\left(\frac{2}{4}\right) \cdot \log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right) \cdot \log_2\left(\frac{2}{4}\right)$

Orta Nabız

$p_{\text{Evet}} = \frac{4}{6}$
$p_{\text{Hayır}} = \frac{2}{6}$

Entropi hesaplaması:
$\text{Entropi}_{\text{Orta}} = -\left(\frac{4}{6}\right) \cdot \log_2\left(\frac{4}{6}\right) - \left(\frac{2}{6}\right) \cdot \log_2\left(\frac{2}{6}\right)$

Düşük Nabız

$p_{\text{Evet}} = \frac{3}{4}$
$p_{\text{Hayır}} = \frac{1}{4}$

Entropi hesaplaması:
$\text{Entropi}_{\text{Düşük}} = -\left(\frac{3}{4}\right) \cdot \log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right) \cdot \log_2\left(\frac{1}{4}\right)$

$\text{Ent}_{\text{Yaş}}(B) = 0{,}694$

$\text{Ent}_{\text{Nabız}}(B) = 0{,}911$

$\text{Ent}_{\text{Durum}}(B) = 0{,}892$

$\text{Gain} = \text{Ent}(A) - \text{Ent}_{\text{Yaş}}(B) =$

$0{,}940 - 0{,}694 = 0{,}246$

$\text{Gain} = \text{Ent}(A) - \text{Ent}_{\text{Nabız}}(B) = 0{,}029$

$\text{Gain} = \text{Ent}(A) - \text{Ent}_{\text{Durum}}(B) = 0{,}048$

Karar Ağacı'nın Avantajları Nelerdir?

Kolay Anlaşılabilir

Karar ağaçları, karmaşık veri ve karar süreçlerini açıklamak ve anlamak için idealdir.

Yüksek Hassasiyet

Doğru parametreler ve eğitim verileri ile karar ağaçları yüksek doğruluk sağlar.

Anlamak ve Eğitmek Kolaydır

Hem profesyoneller hem de eğitimciler için öğrenme süreçlerini kolaylaştırır.

Karar Ağacı'nın Dezavantajları Nelerdir?

- **Aşırı Basit Modeller**

  Bazı durumlarda karar ağacı modelleri gereğinden fazla basit olabilir ve karmaşıklığı yeterince yansıtmayabilir.

- **Overfitting Riski**

  Veri kümesine fazla uyum sağlayarak, yeni veri örneklerini yanlış sınıflandırma riski bulunur.

Ezberlemek

Entropi Düşük
Gini Düşük
Gain Yüksek

Entropiye Bağlı Algoritmalar

ID3

- ID3 algoritmasında karar ağaçları alt dallara bölünürken **entropi değerleri incelenerek** en az kayıp olan bölünme dikkate alınır.
- Kesikli değerler ile çalışır.
- Sade bir yapısı var-az kaynak kullanır.
- ID3, genellikle iki sınıflı (binary) sınıflandırma problemleri için daha uygundur.

C4.5 Algoritması

- C4.5, ID3'ün bazı zayıflıklarını gidermiş ve daha geniş bir kullanım alanına sahip olacak şekilde geliştirilmiştir.
- ID3 te **bilgi kazanımı** dikkate alınırken, C4.5 algoritmasında kazanım oranı ölçütü ile kuralların iyiliği sorgulanır.

GAIN

- C4.5, çok sınıflı problemlerle daha iyi başa çıkabilir.
- Temelde kesikli değerler ile çalışmasına rağmen sürekli değerler de çalışabilir.(C4.5, kategorik ve sayısal özelliklerle başa çıkabilme yeteneğine sahiptir.)
- C4.5, ağaç oluşturulduktan sonra gereksiz dalların kaldırılması (pruning) için bir mekanizma içerir. Bu, ağacın daha genelleştirilebilir ve daha iyi genelleme yeteneğine sahip olmasını sağlar. ID3, bu tür bir kırpma mekanizmasını içermez.

ID3 (**Iterative Dichotomiser 3**)Algoritması

Adım 1: Entropi Hesaplaması

Karar sütununa göre entropiyi hesaplayalım:

$$\text{Entropi} = -p_{\text{Yes}} \cdot \log_2(p_{\text{Yes}}) - p_{\text{No}} \cdot \log_2(p_{\text{No}})$$

$$p_{\text{Yes}} = \frac{9}{14}$$
$$p_{\text{No}} = \frac{5}{14}$$

$$\text{Entropi} = -\left(\frac{9}{14}\right) \cdot \log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \cdot \log_2\left(\frac{5}{14}\right)$$

Entropi hesaplandığında:

$$\text{Entropi} \approx 0.94$$

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

ID3 Algoritması için yardımcı tablolar

Outlook	No	Yes	Genel Toplam
Overcast	0	4	4
Rain	2	3	5
Sunny	3	2	5
Genel Toplam	**5**	**9**	**14**

Humidity	No	Yes	Genel Toplam
High	4	3	7
Normal	1	6	7
Genel Toplam	**5**	**9**	**14**

Temp.	No	Yes	Genel Toplam
Cool	1	3	4
Hot	2	2	4
Mild	2	4	6
Genel Toplam	5	9	14

Wind	No	Yes	Genel Toplam
Strong	3	3	6
Weak	2	6	8
Genel Toplam	**5**	**9**	**14**

ID3 Algoritması- Outlook

Outlook / Sunny:

$$\text{Entropi}_{\text{Sunny}} = -\left(\frac{2}{5}\right)\cdot\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\cdot\log_2\left(\frac{3}{5}\right)$$

$$\text{Entropi}_{\text{Sunny}} \approx 0.971$$

Outlook / Overcast:

$$\text{Entropi}_{\text{Overcast}} = -\left(\frac{4}{4}\right)\cdot\log_2\left(\frac{4}{4}\right) - 0$$

$$\text{Entropi}_{\text{Overcast}} = 0$$

Outlook / Rain:

$$\text{Entropi}_{\text{Rain}} = -\left(\frac{3}{5}\right)\cdot\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\cdot\log_2\left(\frac{2}{5}\right)$$

$$\text{Entropi}_{\text{Rain}} \approx 0.971$$

Bu şekilde hesaplanan entropi değerleri:

- $\text{Entropi}_{\text{Sunny}} \approx 0.971$
- $\text{Entropi}_{\text{Overcast}} = 0$
- $\text{Entropi}_{\text{Rain}} \approx 0.971$

Outlook	No	Yes	Genel Toplam
Overcast	0	4	4
Rain	2	3	5
Sunny	3	2	5
Genel Toplam	**5**	**9**	**14**

$$\text{Bilgi Kazancı}_{\text{Outlook}} = \text{Entropi} - \sum_i \left(\frac{N_i}{N}\cdot\text{Entropi}_{\text{Alt\_i}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Outlook}} = \text{Entropi} - \left(\frac{N_{\text{Sunny}}}{N}\cdot\text{Entropi}_{\text{Sunny}} + \frac{N_{\text{Overcast}}}{N}\cdot\text{Entropi}_{\text{Overcast}} + \frac{N_{\text{Rain}}}{N}\cdot\text{Entropi}_{\text{Rain}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Outlook}} = 0.94 - \left(\frac{5}{14}\cdot 0.971\right) - \left(\frac{4}{14}\cdot 0\right) - \left(\frac{5}{14}\cdot 0.971\right)$$

$$\text{Bilgi Kazancı}_{\text{Outlook}} \approx 0.247$$

ID3 Algoritması- Humidity

Humidity / High:

$p_{\text{Yes}} = \frac{3}{7}$

$p_{\text{No}} = \frac{4}{7}$

$$\text{Entropi}_{\text{High}} = -\left(\frac{3}{7}\right) \cdot \log_2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right) \cdot \log_2\left(\frac{4}{7}\right)$$

Humidity / Normal:

$p_{\text{Yes}} = \frac{6}{7}$

$p_{\text{No}} = \frac{1}{7}$

$$\text{Entropi}_{\text{Normal}} = -\left(\frac{6}{7}\right) \cdot \log_2\left(\frac{6}{7}\right) - \left(\frac{1}{7}\right) \cdot \log_2\left(\frac{1}{7}\right)$$

Bu şekilde hesaplanan entropi değerleri:

- $\text{Entropi}_{\text{High}} \approx 0.985$
- $\text{Entropi}_{\text{Normal}} \approx 0.592$

Humidity	No	Yes	Genel Toplam
High	4	3	7
Normal	1	6	7
Genel Toplam	**5**	**9**	**14**

$$\text{Bilgi Kazancı}_{\text{Humidity}} = \text{Entropi} - \left(\frac{N_{\text{High}}}{N} \cdot \text{Entropi}_{\text{High}} + \frac{N_{\text{Normal}}}{N} \cdot \text{Entropi}_{\text{Normal}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Humidity}} = 0.94 - \left(\frac{7}{14} \cdot 0.985\right) - \left(\frac{7}{14} \cdot 0.592\right)$$

$$\text{Bilgi Kazancı}_{\text{Humidity}} \approx 0.151$$

ID3 Algoritması- Temp.

Temp / Hot:

$p_{\text{Yes}} = \frac{2}{4}$
$p_{\text{No}} = \frac{2}{4}$

$$\text{Entropi}_{\text{Hot}} = -\left(\frac{2}{4}\right) \cdot \log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right) \cdot \log_2\left(\frac{2}{4}\right)$$

Temp / Mild:

$p_{\text{Yes}} = \frac{4}{6}$
$p_{\text{No}} = \frac{2}{6}$

$$\text{Entropi}_{\text{Mild}} = -\left(\frac{4}{6}\right) \cdot \log_2\left(\frac{4}{6}\right) - \left(\frac{2}{6}\right) \cdot \log_2\left(\frac{2}{6}\right)$$

Temp / Cool:

$p_{\text{Yes}} = \frac{3}{4}$
$p_{\text{No}} = \frac{1}{4}$

$$\text{Entropi}_{\text{Cool}} = -\left(\frac{3}{4}\right) \cdot \log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right) \cdot \log_2\left(\frac{1}{4}\right)$$

Bu şekilde hesaplanan entropi değerleri:

- $\text{Entropi}_{\text{Hot}} = 1.0$
- $\text{Entropi}_{\text{Mild}} \approx 0.918$
- $\text{Entropi}_{\text{Cool}} \approx 0.811$

Temp.	No	Yes	Genel Toplam
Cool	1	3	4
Hot	2	2	4
Mild	2	4	6
Genel Toplam	5	9	14

$$\text{Bilgi Kazancı}_{\text{Temp}} = \text{Entropi} - \left(\frac{N_{\text{Hot}}}{N} \cdot \text{Entropi}_{\text{Hot}} + \frac{N_{\text{Mild}}}{N} \cdot \text{Entropi}_{\text{Mild}} + \frac{N_{\text{Cool}}}{N} \cdot \text{Entropi}_{\text{Cool}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Temp}} = 0.94 - \left(\frac{4}{14} \cdot 1.0\right) - \left(\frac{6}{14} \cdot 0.918\right) - \left(\frac{4}{14} \cdot 0.811\right)$$

$$\text{Bilgi Kazancı}_{\text{Temp}} \approx 0.028$$

ID3 Algoritması- Wind.

Wind / Weak:

$$p_{\text{Yes}} = \frac{6}{8}$$
$$p_{\text{No}} = \frac{2}{8}$$

$$\text{Entropi}_{\text{Weak}} = -\left(\frac{6}{8}\right) \cdot \log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right) \cdot \log_2\left(\frac{2}{8}\right)$$

Wind / Strong:

$$p_{\text{Yes}} = \frac{3}{6}$$
$$p_{\text{No}} = \frac{3}{6}$$

$$\text{Entropi}_{\text{Strong}} = -\left(\frac{3}{6}\right) \cdot \log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right) \cdot \log_2\left(\frac{3}{6}\right)$$

Bu şekilde hesaplanan entropi değerleri:

- $\text{Entropi}_{\text{Weak}} = 0.811$
- $\text{Entropi}_{\text{Strong}} = 1.0$

Wind	No	Yes	Genel Toplam
Strong	3	3	6
Weak	2	6	8
Genel Toplam	**5**	**9**	**14**

$$\text{Bilgi Kazancı}_{\text{Wind}} = \text{Entropi} - \left(\frac{N_{\text{Weak}}}{N} \cdot \text{Entropi}_{\text{Weak}} + \frac{N_{\text{Strong}}}{N} \cdot \text{Entropi}_{\text{Strong}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Wind}} = 0.94 - \left(\frac{8}{14} \cdot 0.811\right) - \left(\frac{6}{14} \cdot 1.0\right)$$

$$\text{Bilgi Kazancı}_{\text{Wind}} \approx 0.048$$

Bilgi Kazancı

Outlook: GAIN

Bilgi Kazancı: 0.247

Humidity:

Bilgi Kazancı: 0.151

Temp:

Bilgi Kazancı: 0.028

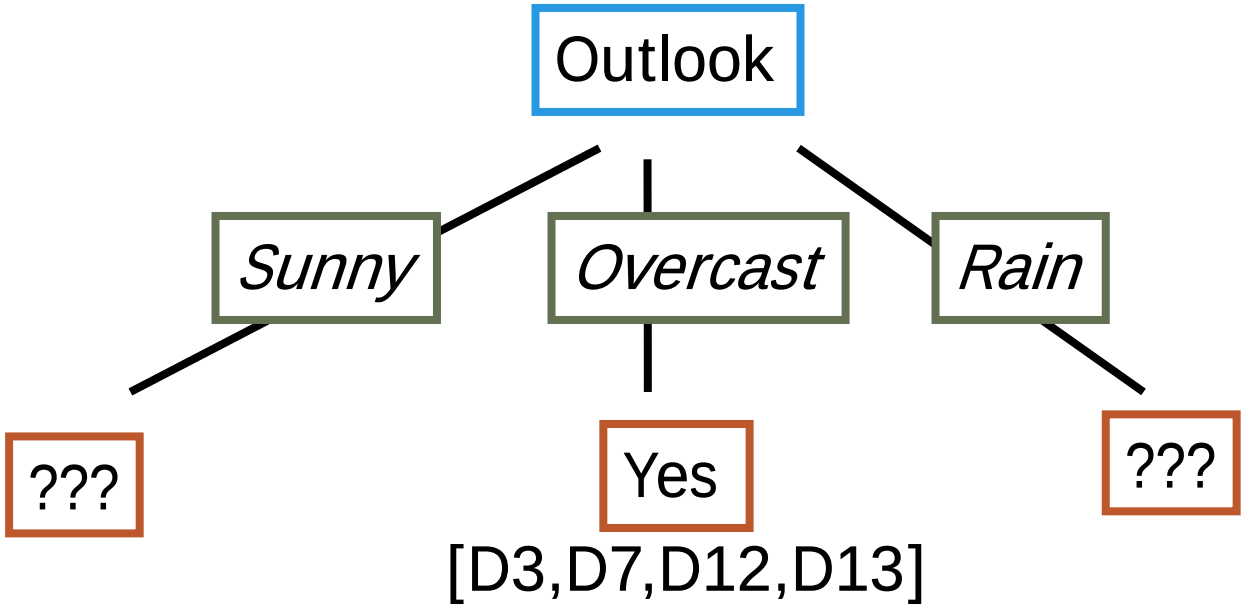
Wind:

Bilgi Kazancı: 0.048

Bu değerler arasında en yüksek bilgi kazancına sahip olan özellik, bir sonraki bölme kriteri olarak seçilebilir.(İlk –Kök Düğüm)

ID3 Algoritması

Day	Outlook	Temp.	Humidity	Wind	Decision
3	Overcast	Hot	High	Weak	Yes
7	Overcast	Cool	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes



- Outlook – Overcast olduğunda Sonuç Yes olduğundan karar ağacı bu düğümün altında sonlandırılır.
- Karar ağacından da anlaşılacağı üzere Rain ve Sunny altında tekrar dallanma olmalıdır.

ID3 Algoritması

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Örnek uzay $S = 5$ (Sunny durumu için geçerli günler).

Humidity / High:

$p_{\text{Yes}} = \frac{3}{3}$
$p_{\text{No}} = \frac{0}{3}$

$$\text{Entropi}_{\text{Humidity\_High}} = -\left(\frac{3}{3}\right) \cdot \log_2\left(\frac{3}{3}\right) - \left(\frac{0}{3}\right) \cdot \log_2\left(\frac{0}{3}\right)$$

$$\text{Entropi}_{\text{Humidity\_High}} = 0$$

Humidity / Normal:

$p_{\text{Yes}} = \frac{2}{2}$
$p_{\text{No}} = \frac{0}{2}$

$$\text{Entropi}_{\text{Humidity\_Normal}} = -\left(\frac{2}{2}\right) \cdot \log_2\left(\frac{2}{2}\right) - \left(\frac{0}{2}\right) \cdot \log_2\left(\frac{0}{2}\right)$$

$$\text{Entropi}_{\text{Humidity\_Normal}} = 0$$

$$\text{Entropi}_{\text{Sunny}} = -\left(\frac{3}{5}\right) \cdot \log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right) \cdot \log_2\left(\frac{2}{5}\right)$$

$$\text{Entropi}_{\text{Sunny}} \approx 0.971$$

Şimdi, "Humidity" için bilgi kazancını tekrar hesaplayalım:

$$\text{Bilgi Kazancı}_{\text{Humidity}} = \text{Entropi}_{\text{Sunny}} - \sum_i \left(\frac{N_i}{N} \cdot \text{Entropi}_{\text{Humidity\_i}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Humidity}} \approx 0.971 - \left(\frac{3}{5} \cdot 0\right) - \left(\frac{2}{5} \cdot 0\right)$$

$$\text{Bilgi Kazancı}_{\text{Humidity}} \approx 0.971$$

ID3 Algoritması

Diğer bilgi kazancını Humidity ölçtüğümüz yöntemin aynısını takip ederek ölçebiliriz.

Temp:

$\text{Entropi}_{\text{Temp\_Sunny\_High}} = 0$ (Çünkü bu durumda sadece "Yes" sınıfı var)
$\text{Entropi}_{\text{Temp\_Sunny\_Normal}} = 1$ (Çünkü bu durumda "Yes" ve "No" sınıfları var)

$$\text{Bilgi Kazancı}_{\text{Temp}} \approx 0.971 - \left(\frac{3}{5} \cdot 0\right) - \left(\frac{2}{5} \cdot 1\right)$$
$$\text{Bilgi Kazancı}_{\text{Temp}} \approx 0.571$$

Wind:

$\text{Entropi}_{\text{Wind\_Sunny\_Weak}} = 0$ (Çünkü bu durumda sadece "Yes" sınıfı var)
$\text{Entropi}_{\text{Wind\_Sunny\_Strong}} = 0.918$ (Çünkü bu durumda "Yes" ve "No" sınıfları var)

$$\text{Bilgi Kazancı}_{\text{Wind}} \approx 0.971 - \left(\frac{3}{5} \cdot 0\right) - \left(\frac{2}{5} \cdot 0.918\right)$$
$$\text{Bilgi Kazancı}_{\text{Wind}} \approx 0.019$$

Yukarıdaki hesaplamalara göre:

- "Temp" özelliği için bilgi kazancı yaklaşık olarak 0.571.
- "Wind" özelliği için bilgi kazancı yaklaşık olarak 0.019.

- **Temp:**

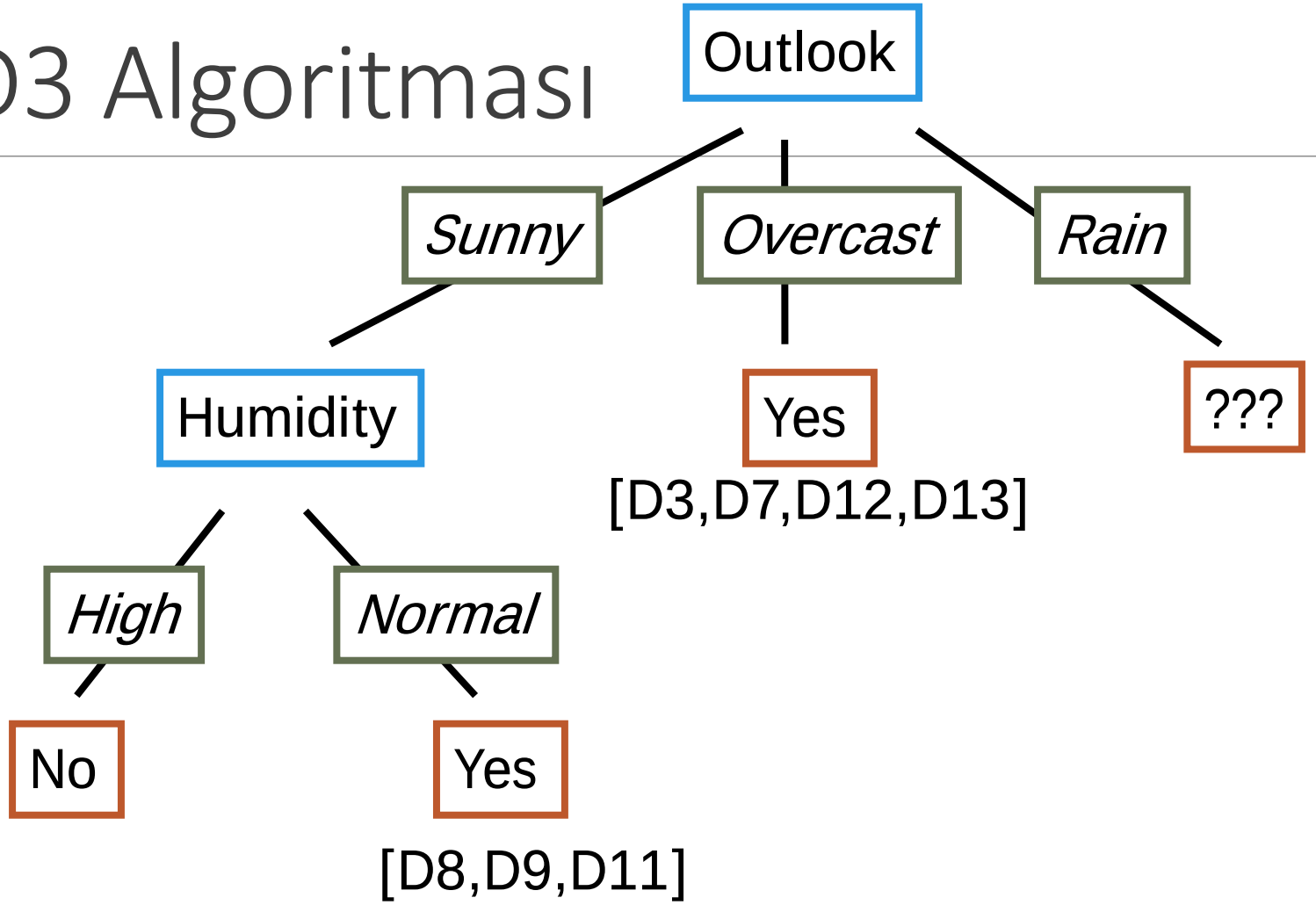
  Bilgi Kazancı ≈ 0.571
- **Wind:**

  Bilgi Kazancı ≈ 0.019
- **Humidity:**

  Bilgi Kazancı ≈ 0.971

Bu değerlere göre, "Humidity" özelliği en yüksek bilgi kazancına sahiptir, bu nedenle "Humidity" özelliği kök düğüm olarak seçilebilir.

ID3 Algoritması



ID3 Algoritması-
Outlook-Sunny-Humidity Düğümü

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No

ID3 Algoritması Outlook-Rain-Wind Düğümü

$$\text{Bilgi Kazancı}_{\text{Temp}} = \text{Entropi}_{\text{Rain}} - \sum_i \left(\frac{N_i}{N} \cdot \text{Entropi}_{\text{Temp\_Rain\_i}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Temp}} = 0.971 - \left(\frac{4}{5} \cdot 0.811\right) - \left(\frac{1}{5} \cdot 1\right)$$

$$\text{Bilgi Kazancı}_{\text{Temp}} \approx 0.01997309402197489$$

$$\text{Bilgi Kazancı}_{\text{Humidity}} = \text{Entropi}_{\text{Rain}} - \sum_i \left(\frac{N_i}{N} \cdot \text{Entropi}_{\text{Humidity\_R}}\right.$$

$$\text{Bilgi Kazancı}_{\text{Humidity}} = 0.971 - \left(\frac{2}{5} \cdot 0\right) - \left(\frac{1}{5} \cdot 0\right) - \left(\frac{2}{5} \cdot 1\right)$$

$$\text{Bilgi Kazancı}_{\text{Humidity}} \approx 0.01997309402197489$$

$$\text{Bilgi Kazancı}_{\text{Wind}} = \text{Entropi}_{\text{Rain}} - \sum_i \left(\frac{N_i}{N} \cdot \text{Entropi}_{\text{Wind\_Rain\_i}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Wind}} = 0.971 - \left(\frac{3}{5} \cdot 0.918\right) - \left(\frac{2}{5} \cdot 0\right)$$

$$\text{Bilgi Kazancı}_{\text{Wind}} \approx 0.9709505944546686$$

- $\text{Bilgi Kazancı}_{\text{Temp}} \approx 0.01997309402197489$
- $\text{Bilgi Kazancı}_{\text{Humidity}} \approx 0.01997309402197489$
- $\text{Bilgi Kazancı}_{\text{Wind}} \approx 0.9709505944546686$

Bu değerler, "Wind" özelliğinin "Rain" durumu için diğer iki özelliğe göre daha yüksek bilgi kazancına sahip olduğunu gösterir. Bu nedenle, "Wind" özelliği, "Rain" durumunu belirleme konusunda daha fazla bilgi sağlamaktadır.

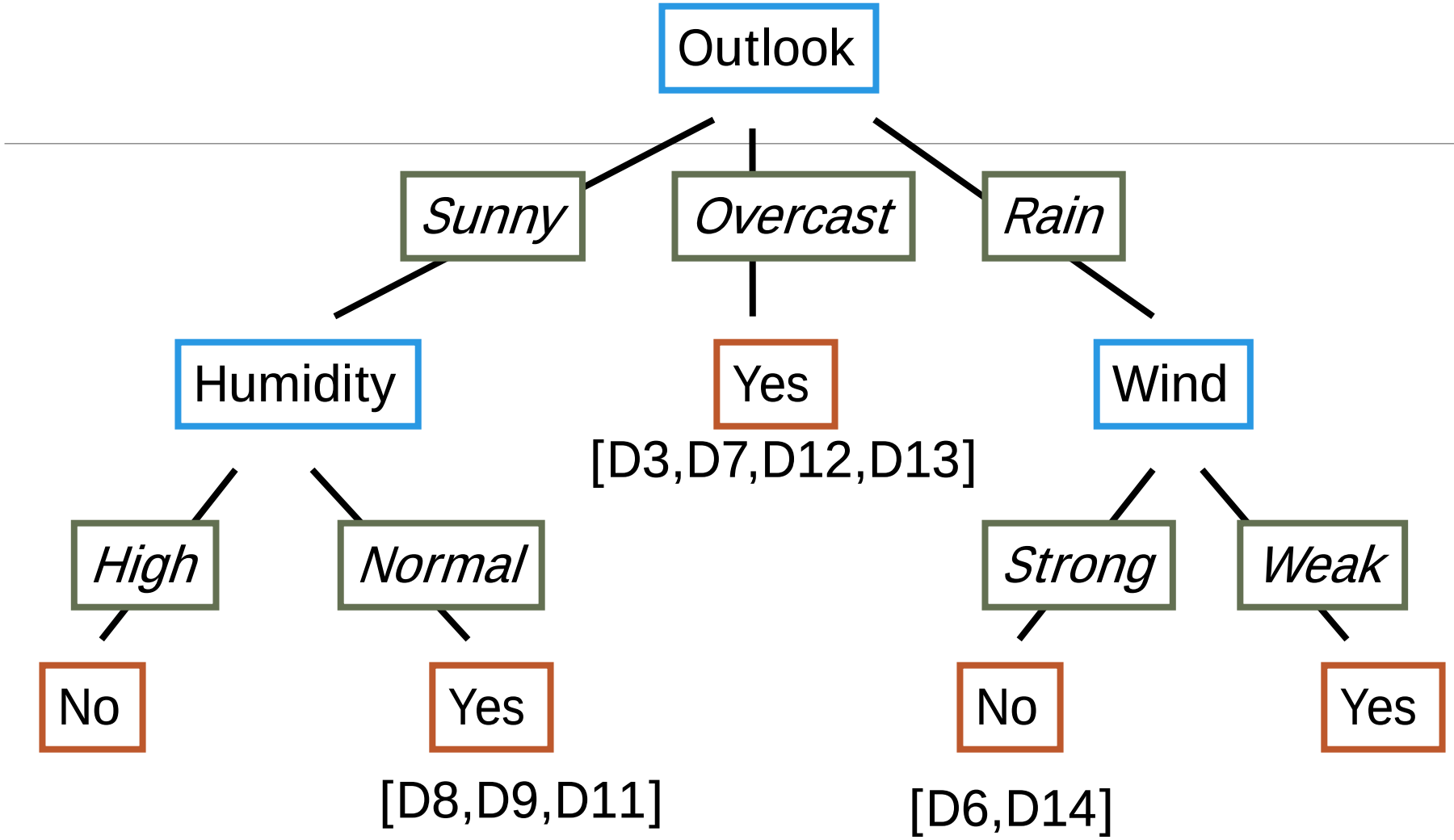
ID3 Algoritması-Rain Düğümü

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes

Day	Outlook	Temp.	Humidity	Wind	Decision
6	Rain	Cool	Normal	Strong	No
14	Rain	Mild	High	Strong	No

ID3 Algoritması



ID3 Algoritması

Elde edilen karar ağacından anlaşılacağı üzere dallanmanın tamamlandığı 5 yaprak (nihai karar) mevcuttur. Bu durumda 5 farklı kural yazılabilir.

*Eğer Outlook = sunny ve Humidity = High ise Karar = No

** ....

C4.5 Algoritması

Gain ratio ve Split Info, C4.5 algoritmasında kullanılan iki önemli kriterdir.

Gain ratio(Kazanç Oranı), özellikle bir düğümün bölünmesinin bilgi kazancını oransal olarak ifade eder.

Split Info(Dallanma Değeri) ise bir özellikle yapılan bölünmenin ne kadar "düzensiz" olduğunu ölçer.

C4.5 algoritması, her özelliğin gain ratio ve split info değerlerini kullanarak en iyi bölünmeyi bulur.

Örnek bir özellik (Outlook) üzerindengain ratio ve split info hesaplayalım.

C4.5 Algoritması

Outlook / Sunny:

$$\text{Entropi}_{\text{Sunny}} = -\left(\frac{2}{5}\right)\cdot\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\cdot\log_2\left(\frac{3}{5}\right)$$

$$\text{Entropi}_{\text{Sunny}} \approx 0.971$$

Outlook / Overcast:

$$\text{Entropi}_{\text{Overcast}} = -\left(\frac{4}{4}\right)\cdot\log_2\left(\frac{4}{4}\right) - 0$$

$$\text{Entropi}_{\text{Overcast}} = 0$$

Outlook / Rain:

$$\text{Entropi}_{\text{Rain}} = -\left(\frac{3}{5}\right)\cdot\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\cdot\log_2\left(\frac{2}{5}\right)$$

$$\text{Entropi}_{\text{Rain}} \approx 0.971$$

Bu şekilde hesaplanan entropi değerleri:

- $\text{Entropi}_{\text{Sunny}} \approx 0.971$
- $\text{Entropi}_{\text{Overcast}} = 0$
- $\text{Entropi}_{\text{Rain}} \approx 0.971$

Outlook	No	Yes	Genel Toplam
Overcast	0	4	4
Rain	2	3	5
Sunny	3	2	5
Genel Toplam	**5**	**9**	**14**

$$\text{Bilgi Kazancı}_{\text{Outlook}} = \text{Entropi} - \sum_i \left(\frac{N_i}{N}\cdot\text{Entropi}_{\text{Alt\_i}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Outlook}} = \text{Entropi} - \left(\frac{N_{\text{Sunny}}}{N}\cdot\text{Entropi}_{\text{Sunny}} + \frac{N_{\text{Overcast}}}{N}\cdot\text{Entropi}_{\text{Overcast}} + \frac{N_{\text{Rain}}}{N}\cdot\text{Entropi}_{\text{Rain}}\right)$$

$$\text{Bilgi Kazancı}_{\text{Outlook}} = 0.94 - \left(\frac{5}{14}\cdot 0.971\right) - \left(\frac{4}{14}\cdot 0\right) - \left(\frac{5}{14}\cdot 0.971\right)$$

$$\text{Bilgi Kazancı}_{\text{Outlook}} \approx 0.247$$

C4.5 Algoritması

$\text{Gain Ratio}_{(\text{Outlook})} = \text{Gain}_{(\text{Outlook})} / \text{Split Info}_{(\text{Outlook})}$

Outlook	No	Yes	Genel Toplam
Overcast	0	4	4
Rain	2	3	5
Sunny	3	2	5
Genel Toplam	**5**	**9**	**14**

$$\text{Split Info}_{(\text{Outlook})} = -\frac{4}{14}\left(log_2 \frac{4}{14}\right) - \frac{5}{14}\left(log_2 \frac{5}{14}\right) - \frac{5}{14}\left(log_2 \frac{5}{14}\right) = 1{,}5717$$

$$\text{Bilgi Kazancı}_{\text{Outlook}} \approx 0.247$$

$\text{Gain Ratio}_{(\text{Outlook})} = 0{,}247 / 1{,}5717 = 0.1566$

C4.5 Algoritması

Gain Ratio$_{(Humidity)}$ = Gain$_{(Humidity)}$ / Split Info$_{(Humidity)}$

$$\text{Bilgi Kazancı}_{\text{Humidity}} \approx 0.151$$

Split Info$_{(Humidity)}$ = $-\frac{7}{14}\left(log_2 \frac{7}{14}\right) - \frac{7}{14}\left(log_2 \frac{7}{14}\right) = 1{,}0$

Gain Ratio$_{(Humidity)}$ = 0,151 / 1 = 0,151

Gain Ratio$_{(Temp)}$ = Gain$_{(Temp)}$ / Split Info$_{(Temp)}$

Split Info$_{(Temp)}$ = $-\frac{4}{14}\left(log_2 \frac{4}{14}\right) - \frac{6}{14}\left(log_2 \frac{6}{14}\right) - \frac{6}{14}\left(log_2 \frac{6}{14}\right) = 1.5567$

$$\text{Bilgi Kazancı}_{\text{Temp}} \approx 0.028$$

Gain Ratio$_{(Temp)}$ = 0,028/ 1,5567 = 0,018

C4.5 Algoritması

$\text{Gain Ratio}_{(\text{Temp})} = \text{Gain}_{(\text{Temp})} / \text{Split Info}_{(\text{Temp})}$

$\text{Bilgi Kazancı}_{\text{Wind}} \approx 0.048$

$$\text{Split Info}_{(\text{Temp})} = -\frac{6}{14}\left(log_2 \frac{6}{14}\right) - \frac{8}{14}\left(log_2 \frac{8}{14}\right) = 0.9852$$

$\text{Gain Ratio}_{(\text{Temp})} = 0{,}048 / 0.9852 = 0{,}048$

$\text{Gain Ratio}_{(\text{Outlook})} = 0.1566$ ile en yüksek kazanç oranını Outlook değişkeni olduğundan karar ağacının ilk düğümü Outlook olacaktır.
Alt düğümlerin tespiti için algoritma işletilmeye devam edilecektir.

https://www.kaggle.com/code/hadibakhsh/decision-tree-play-tennis

CART Algoritması
GINI İndeksinin Hesaplanması

Outlook	No	Yes	Genel Toplam
Overcast	0	4	4
Rain	2	3	5
Sunny	3	2	5
Genel Toplam	**5**	**9**	**14**

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

pi o sınıftaki örneklerin toplam veri setine oranını temsil eder.

1. **Sunny için Gini:**

$$Gini(\text{sunny}) = 1 - (P(\text{no}))^2 - (P(\text{yes}))^2$$
$$= 1 - (\tfrac{3}{5})^2 - (\tfrac{2}{5})^2$$
$$= 1 - \tfrac{9}{25} - \tfrac{4}{25}$$
$$= 1 - 0.36 - 0.16$$
$$= 0.48$$

2. **Overcast için Gini:**

$$Gini(\text{overcast}) = 1 - (P(\text{no}))^2 - (P(\text{yes}))^2$$
$$= 1 - (\tfrac{0}{4})^2 - (\tfrac{4}{4})^2$$
$$= 1 - 0 - 1$$
$$= 0$$

3. **Rainy için Gini:**

$$Gini(\text{rainy}) = 1 - (P(\text{no}))^2 - (P(\text{yes}))^2$$
$$= 1 - (\tfrac{2}{6})^2 - (\tfrac{4}{6})^2$$
$$= 1 - \tfrac{1}{9} - \tfrac{4}{9}$$
$$= \tfrac{4}{9}$$

Outlook	Gini
Sunny	0.48
Overcast	0.00
Rainy	0.44

Gini her bir hava durumu koşulunun veri setinin homojenliğini ölçer. Daha düşük Gini değerleri, daha homojen alt kümeleri temsil eder.

Şimdi toplam Gini değerini hesaplayalım:

$$Gini_{\text{total}} = \tfrac{5}{14} \times 0.48 + \tfrac{4}{14} \times 0.00 + \tfrac{5}{14} \times 0.44$$

$$Gini_{\text{total}} = 0.3571 \times 0.48 + 0.2857 \times 0.00 + 0.3571 \times 0.44$$

$$Gini_{\text{total}} = 0.17143 + 0.0 + 0.15748$$

$$Gini_{\text{total}} = 0.32891$$

CART Algoritması

Temp.	No	Yes	Genel Toplam
Cool	1	3	4
Hot	2	2	4
Mild	2	4	6
Genel Toplam	5	9	14

Temperature (Temp):

$$Gini(\text{Temp}) = \sum_i \frac{N_i}{N} \cdot Gini(\text{Temp=i})$$

1. **Hot için Gini:**

$$Gini(\text{hot}) = 1 - (P(\text{no}))^2 - (P(\text{yes}))^2$$
$$= 1 - (\tfrac{2}{4})^2 - (\tfrac{2}{4})^2$$
$$= 1 - 0.25 - 0.25$$
$$= 0.50$$

2. **Mild için Gini:**

$$Gini(\text{mild}) = 1 - (P(\text{no}))^2 - (P(\text{yes}))^2$$
$$= 1 - (\tfrac{2}{6})^2 - (\tfrac{4}{6})^2$$
$$= 1 - \tfrac{1}{9} - \tfrac{4}{9}$$
$$= \tfrac{4}{9}$$

3. **Cool için Gini:**

$$Gini(\text{Cool}) = 1 - (P(\text{No}))^2 - (P(\text{Yes}))^2$$
$$= 1 - (\tfrac{1}{4})^2 - (\tfrac{3}{4})^2$$
$$= 1 - 0.0625 - 0.5625$$
$$= 1 - 0.625$$
$$= 0.375$$

$$Gini_{\text{total}} = \tfrac{4}{14} \times 0.5 + \tfrac{6}{14} \times 0.444 + \tfrac{4}{14} \times 0.375$$

$$Gini_{\text{total}} = 0.2857 \times 0.5 + 0.4285 \times 0.444 + 0.2857 \times 0.375$$

$$Gini_{\text{total}} = 0.14285 + 0.19053 + 0.10718$$

$$Gini_{\text{total}} = 0.44056$$

CART Algoritması

$P(\text{No}) = \frac{2}{8} = \frac{1}{4}$
$P(\text{Yes}) = \frac{6}{8} = \frac{3}{4}$

Şimdi "wind=weak" durumu için Gini impurity değerini hesaplayalım:

$$Gini(\text{Weak}) = 1 - (P(\text{No}))^2 - (P(\text{Yes}))^2$$
$$= 1 - (\tfrac{1}{4})^2 - (\tfrac{3}{4})^2$$
$$= 1 - \tfrac{1}{16} - \tfrac{9}{16}$$
$$= 1 - 0.0625 - 0.5625$$
$$= 0.375$$

$$Gini_{\text{total}} = \tfrac{8}{14} \times 0.375 + \tfrac{6}{14} \times 0$$
$$Gini_{\text{total}} = 0.5714 \times 0.375 + 0.4286 \times 0$$
$$Gini_{\text{total}} = 0.214$$

Wind	No	Yes	Genel Toplam
Strong	3	3	6
Weak	2	6	8
Genel Toplam	**5**	**9**	**14**

CART Algoritması

$$Gini(\text{High}) = 1 - (P(\text{No}))^2 - (P(\text{Yes}))^2$$
$$= 1 - (\tfrac{3}{7})^2 - (\tfrac{4}{7})^2$$
$$= 1 - \tfrac{9}{49} - \tfrac{16}{49}$$
$$= 1 - 0.1837 - 0.3265$$
$$= 1 - 0.5102$$
$$= 0.4898$$

$$Gini(\text{Normal}) = 1 - (P(\text{No}))^2 - (P(\text{Yes}))^2$$
$$= 1 - (\tfrac{1}{7})^2 - (\tfrac{6}{7})^2$$
$$= 1 - \tfrac{1}{49} - \tfrac{36}{49}$$
$$= 1 - 0.0204 - 0.7347$$
$$= 1 - 0.7551$$
$$= 0.2449$$

Humidity	No	Yes	Genel Toplam
High	4	3	7
Normal	1	6	7
Genel Toplam	**5**	**9**	**14**

$$Gini_{\text{total}} = \tfrac{7}{14} \times 0.4898 + \tfrac{7}{14} \times 0.2449$$
$$Gini_{\text{total}} = 0.5 \times 0.4898 + 0.5 \times 0.2449$$
$$Gini_{\text{total}} = 0.2449$$

Wind değişkeni en küçük GINI indeks değerini sağladığından karar ağacının ilk düğümü Wind olacaktır. Alt düğümlerin tespiti için algoritma işletilmeye devam edilecektir.