Published in Deep Learning Türkiye · Follow

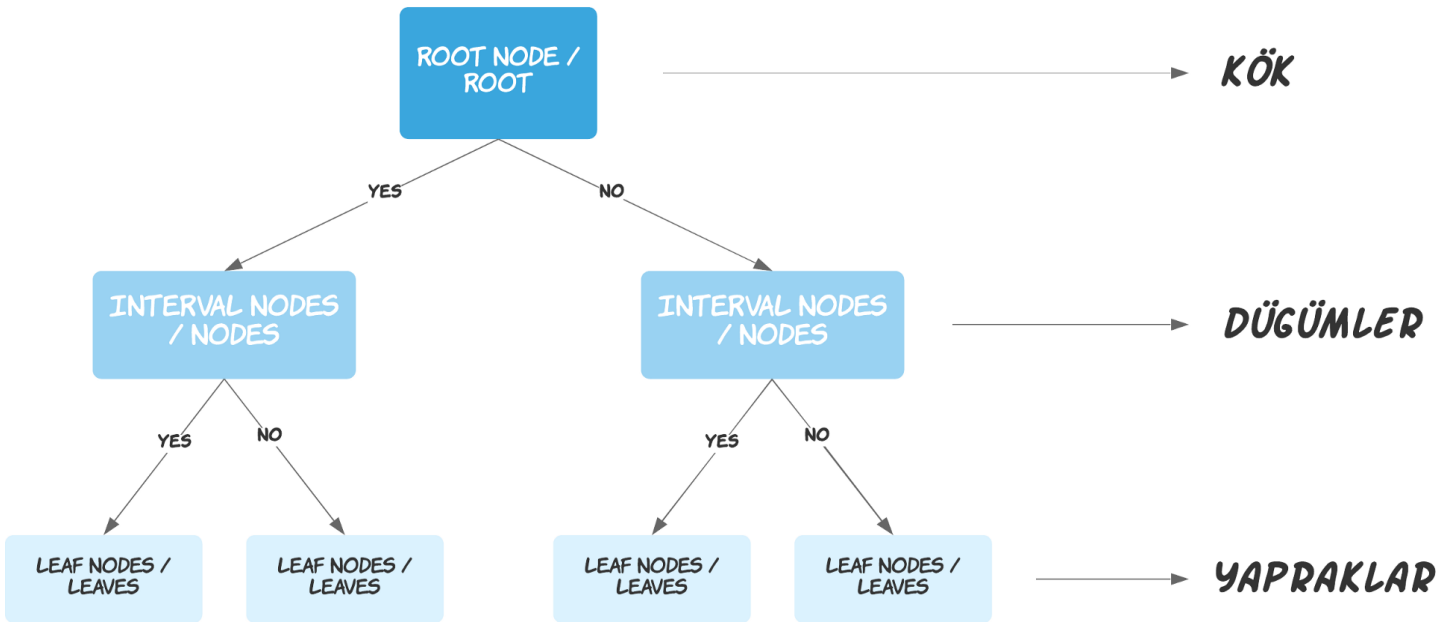
Mehmet Fatih AKCA · Following
Sep 7, 2020 · 4 min read

Karar Ağaçları (Makine Öğrenmesi Serisi-3)

Karar ağaçları, Sınıflandırma ve Regresyon problemlerinde kullanılan, ağaç tabanlı algoritmadan biridir. Karmaşık veri setlerinde kullanılabilir.

. . .

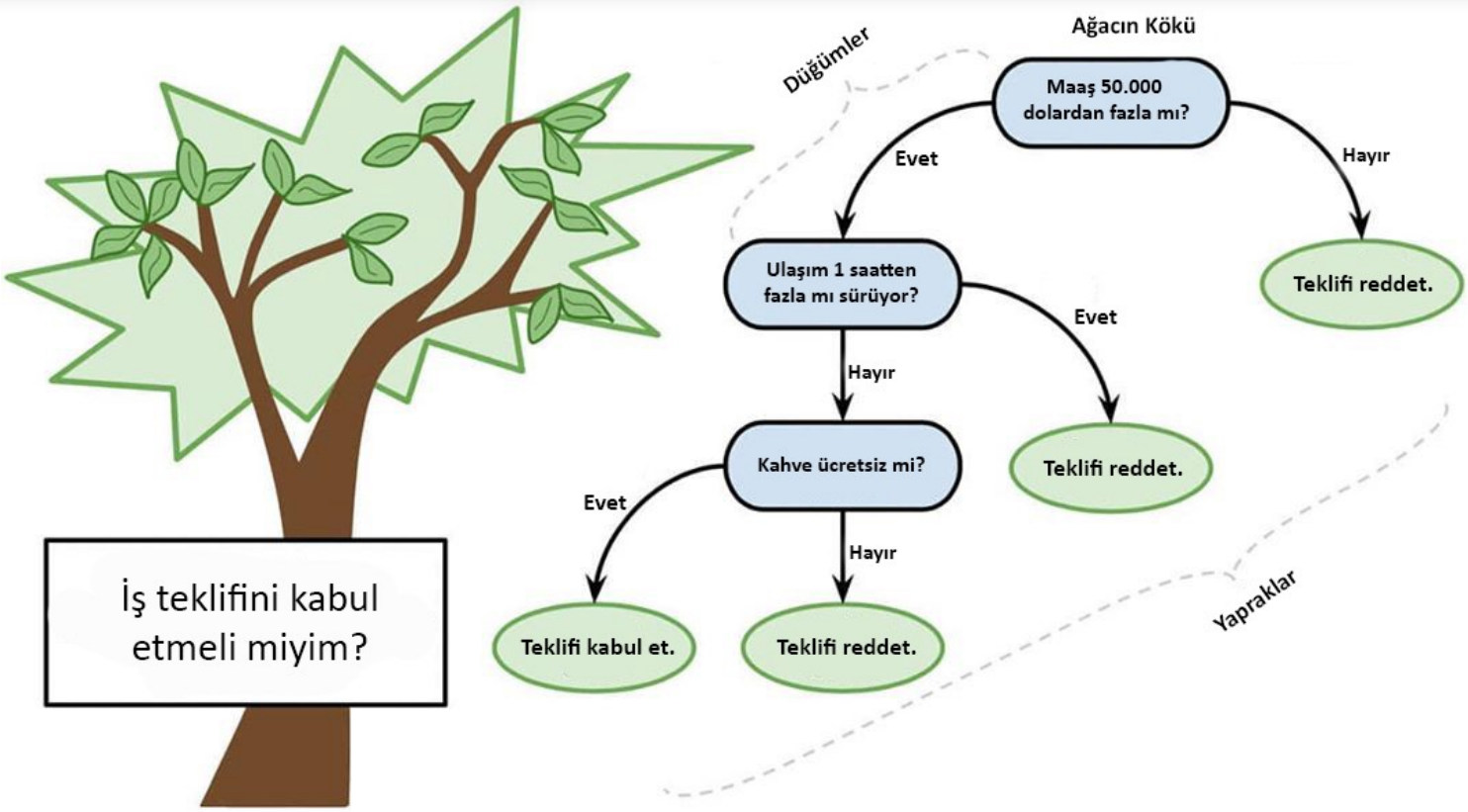
Karar Ağaçlarını Tanıyalım



Karar Ağacı (Decision Tree) (Görsel 1)

Karar ağaçlarının ilk hücrelerine **kök** (root veya root node) denir. Her bir gözlem kökteki koşula göre “Evet” veya “Hayır” olarak sınıflandırılır.

Kök hücrelerinin altında **düğümler** (interval nodes veya nodes) bulunur. Her bir gözlem



(Görsel 2)

Peki ya kök hücreyi nasıl seçiyoruz?

Burada seçeceğimiz kökün veri setimizi olabildiğince çok açıklamasını isteriz. Örneğin yukarıdaki örnek karar ağacına bakarsak, bu kişi için iş teklifinde en önemli etken maaşmış.

Köke tabi ki biz karar vermiyoruz. Buna karar vermek için çeşitli değerler var. Bunlardan bazıları:

Gini: Alt kümenin saflık değeri

$$Gini = 1 - \sum_{j} p_j^2$$

(Görsel 3)

pj, j sınıfının gerçekleşme olasılığıdır. Her sınıf için hesaplanır ve çıkan sonuçların karelerinin toplamı birden çıkartılır. Gini değeri 0 ile 1 arasında bir sonuç alır ve sonuç 0'a ne kadar yakınsa

Elimizde Kırmızı ve Mavi renkte 4 topumuz olduğunu varsayalım;

Gini = 1 - (Kırmızı Topun Seçilme Olasılığı^2 + Mavi Topun Seçilme Olasılığı^2)

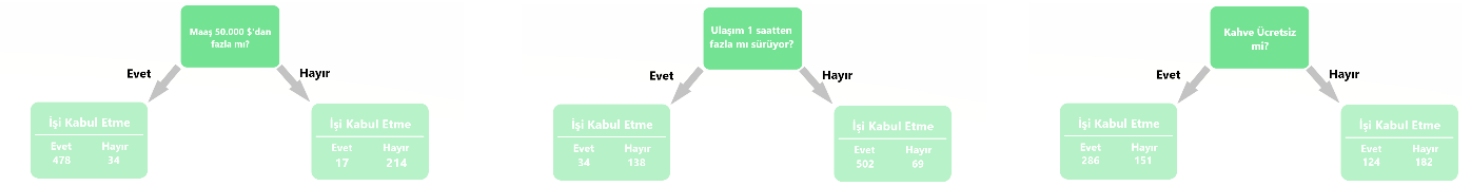
2 Kırmızı - 2 Mavi
Gini = 1 - ((1/2)^2 + (1/2)^2)
Gini = 0.5

0 Kırmızı - 4 Mavi
Gini = 1 - ((0)^2 + (1)^2)
Gini = 0

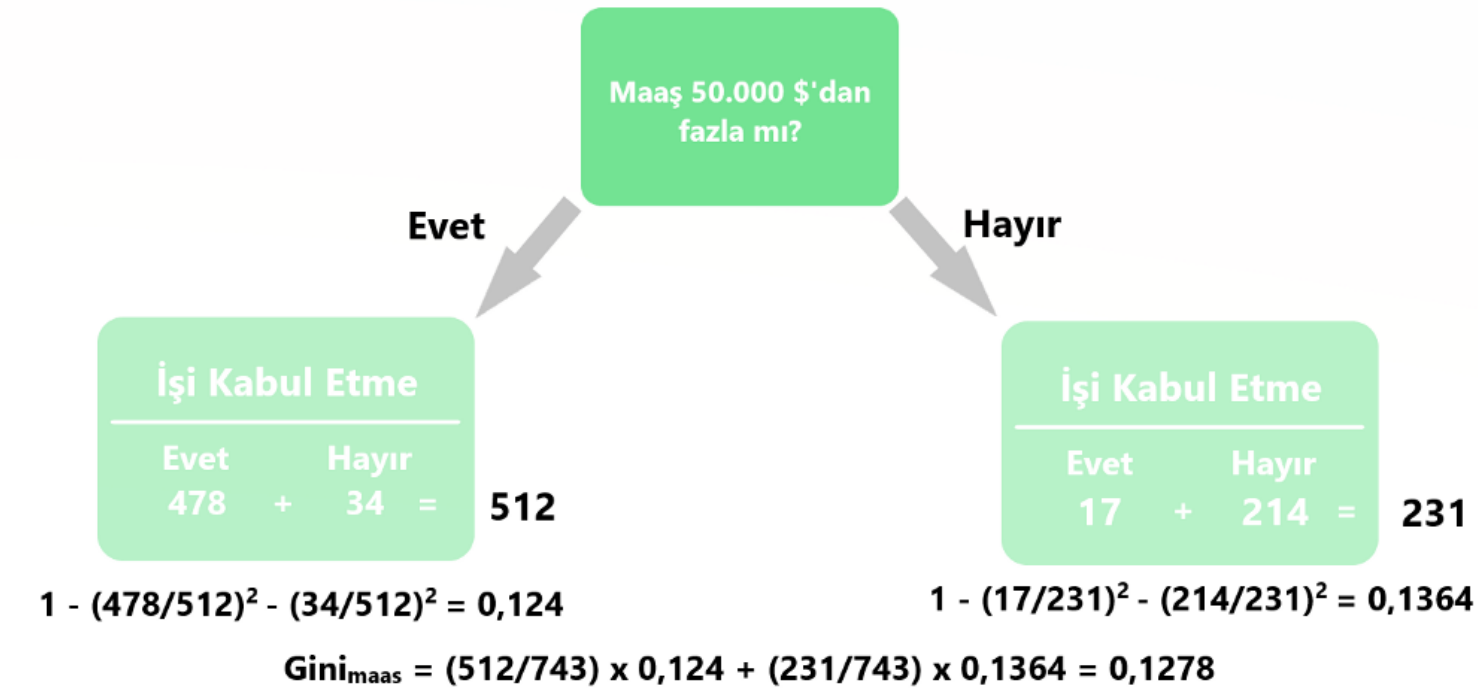
3 Kırmızı - 1 Mavi
Gini = 1 - ((3/4)^2 + (1/4)^2)
Gini = 0.375

(Görsel 4)

Şimdi Karar Ağaçlarına bakalım. Bakalım düğümlerimiz veri setimizi nasıl ayırabiliyor.



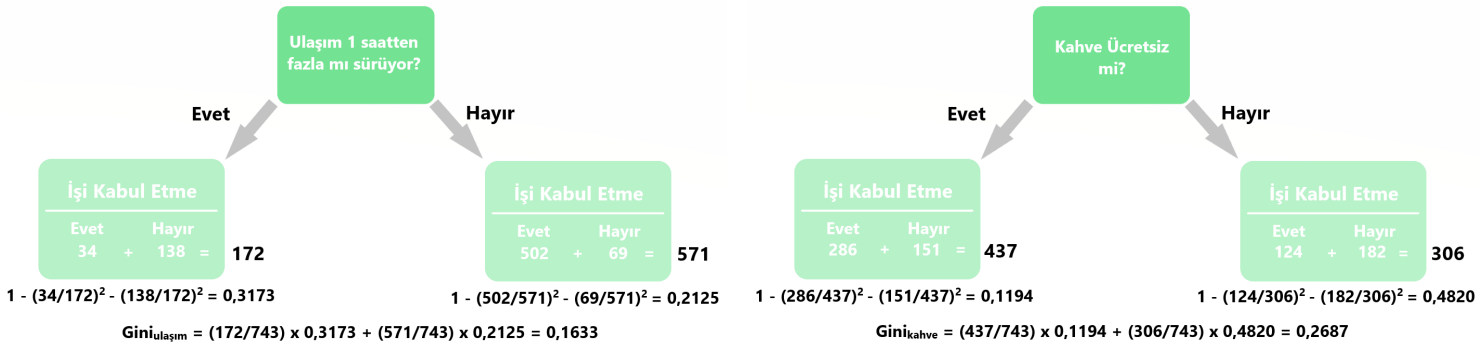
Kök (root) hücreyi bulabilmek için her düğüm için Gini değerini hesaplamamız gerekiyor.



Maaş için Gini değerinin hesaplanması

reddedilme olasılığı işlemini yaparak cevabı “Evet” olanların Gini değerini bulmuş oluyoruz. Aynı işlemi cevabı “Hayır” olanlar için tekrarlıyoruz. Son olarak (işi kabul edenlerin veri setindeki oranı * işi kabul edenlerin Gini değeri) + (işi reddedenlerin veri setindeki oranı * işi reddedenlerin Gini değeri) işlemini yapıyoruz ve maaş için 0,1278 değerini buluyoruz.

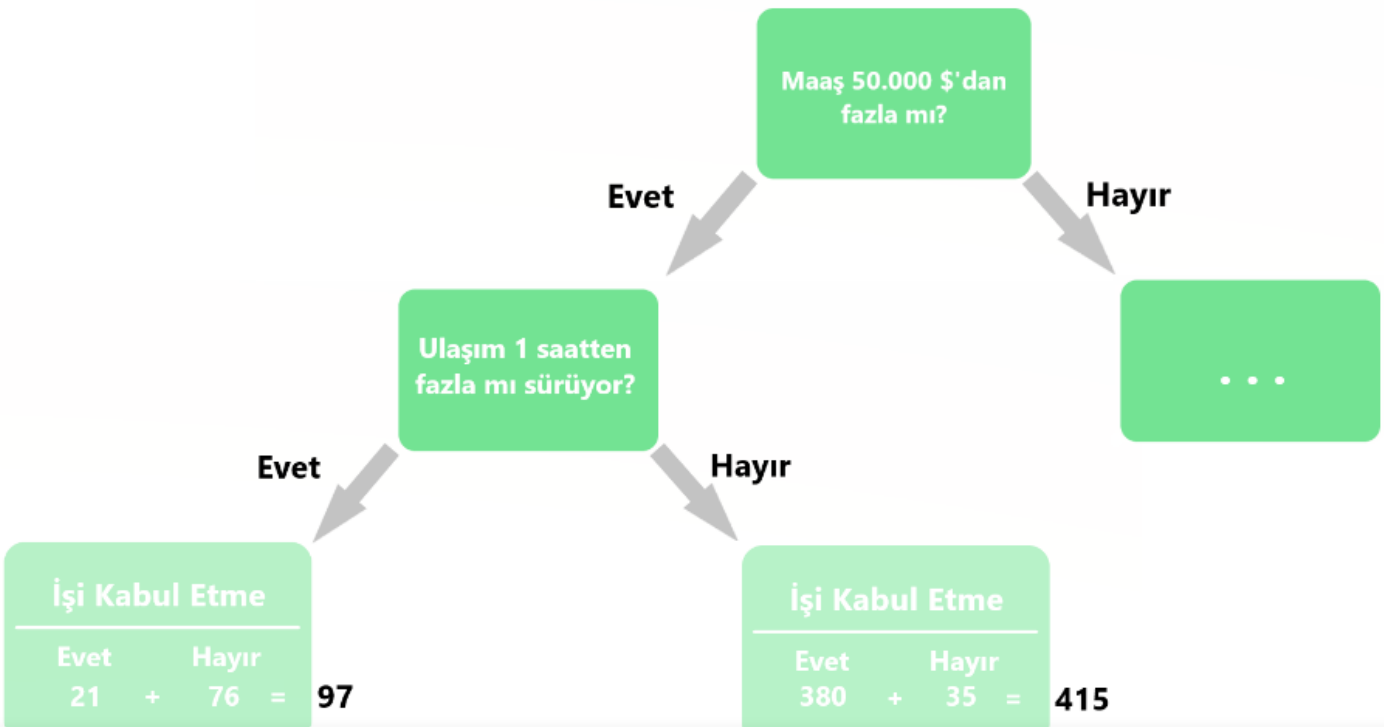
Diğerleri içinde aynı işlemleri yapıyoruz.



Ulaşım ve Kahve için Gini değerinin hesaplanması

Değerlere baktığımızda 0'a en yakın olanın Maaş olduğunu görüyoruz. Bu bize Maaş düğümü ile yapılan ayrımın veri setini diğerlerine göre daha iyi ayrıştırdığını gösteriyor. Bu yüzden kök (root) hücre olarak Maaş düğümünü seçiyoruz.

Kök hücreden sonra hangisinin geleceğini de aynı şekilde karar veriyoruz. Örnek olması için:



Entropy: Temel fikir, bir gruplamanın bozukluğunu hedef değişkene göre ölçmektir ama bunu log2 tabanında yapar.

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

Gini ile arasında çok büyük bir fark yoktur. Entropi daha dengeli bir ağaç çıkarmaya meyilli iken Gini, frekansı fazla olan sınıfı ayrıştırmaya meyillidir.

. . .

Eğitim Algoritmaları:

Scikit-learn Karar Ağaçlarını eğitirken CART (Classification and Regression Tree) algoritmasını kullanır. Bu algoritma veri setini ikiye ayırarak ayrıştırmaya çalışır. Diğer algoritmalar ile karşılaştırmak için şöyle bir tablo oluşturdum:

	ID3	C4.5	CART
Kullandığı Veri Tipleri	Kategorik	Sürekli ve Kategorik	Sürekli ve nominal
Hız	Düşük	ID3'den daha hızlı	Ortalama
Boosting	Desteklemiyor	Desteklemiyor	Destekliyor
Budama	Hayır	Önceden Budama	Sonra Budama
Kayıp değerler (nan)	Desteklemiyor	Desteklemiyor	Destekliyor
Formül	Entropi ve Kazanç değeri	Kazanç oranı ve bölünme	Gini değeri

Hiperparametreler:

Aşırı uyum (overfitting) veya öğrenememe(underfitting) gibi sorunlarla karşı karşıya iseniz hiperparametre değerleri ile biraz oynamanız gerekebilir.

max_depth= Karar Ağacının maksimum derinliğini ifade eder. Değer girilmezse limitsiz olur. Model overfit(aşırı uyum) olmuşsa düşürülmesi gerekir.

min_weight_fraction= min_samples_leaf'e benzer.Ağırlıklı örneklerin, toplam örnekler içerisindeki oranı. Ağacın dengeli gitmesi için kullanılır.

max_leaf_nodes= Maksimum yaprak sayısı.

max_features= En iyi bölünmeyi ararken göz önünde bulundurulması gereken featureların sayısı.

. . .

Özet:

1-) Sınıflandırma ve Regresyon problemlerinde çok çıktılı bir şekilde çalışabilir.

2-)Karmaşık veri setlerinde kullanılabilir.

3-) Scale etmeye ve çok fazla Veri Ön İşleme'ye gerek duymaz.

4-) Kök hücreyi seçerken veri setini mümkün olduğunca anlamlı şekilde ayrıştırabilen sütun seçilmeye çalışır. (Görsel 2'deki maaş gibi)

5-) Karar Ağacı, önceden tanımlanmış olan ***max_depth*** hiperparametresine ulaştığında veya daha saf bir alt küme elde edemediğinde durur.

6-) Gini ve Entropi arasında çok büyük bir fark yoktur. Entropi daha dengeli bir ağaç çıkarmaya meyilli iken Gini frekansı fazla olan sınıfı ayrıştırmaya meyillidir.

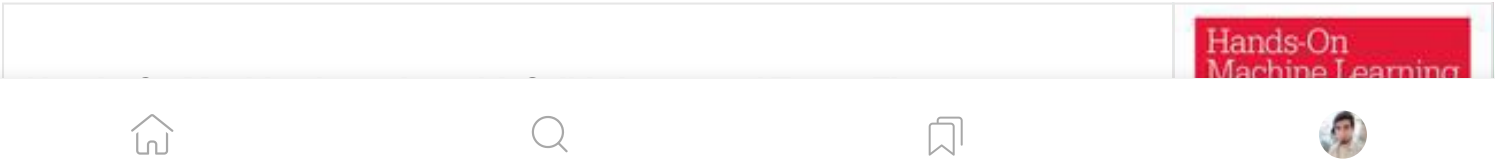
7-) Model overfit (aşırı uyum) olmuşsa genellikle önce ***max_depth*** hiperparametresi düşürülür.

. . .

Karar Ağaçları bu kadardı. Okuduğunuz için teşekkür ederim.

Esen Kalın.

Kaynaklar:



StatQuest with Josh Starmer

Statistics, Machine Learning and Data Science can sometimes seem like very scary topics, but since each technique is...

www.youtube.com

scikit-learn

"We use scikit-learn to support leading-edge basic research [...]" "I think it's the most well-designed ML package I've...

scikit-learn.org