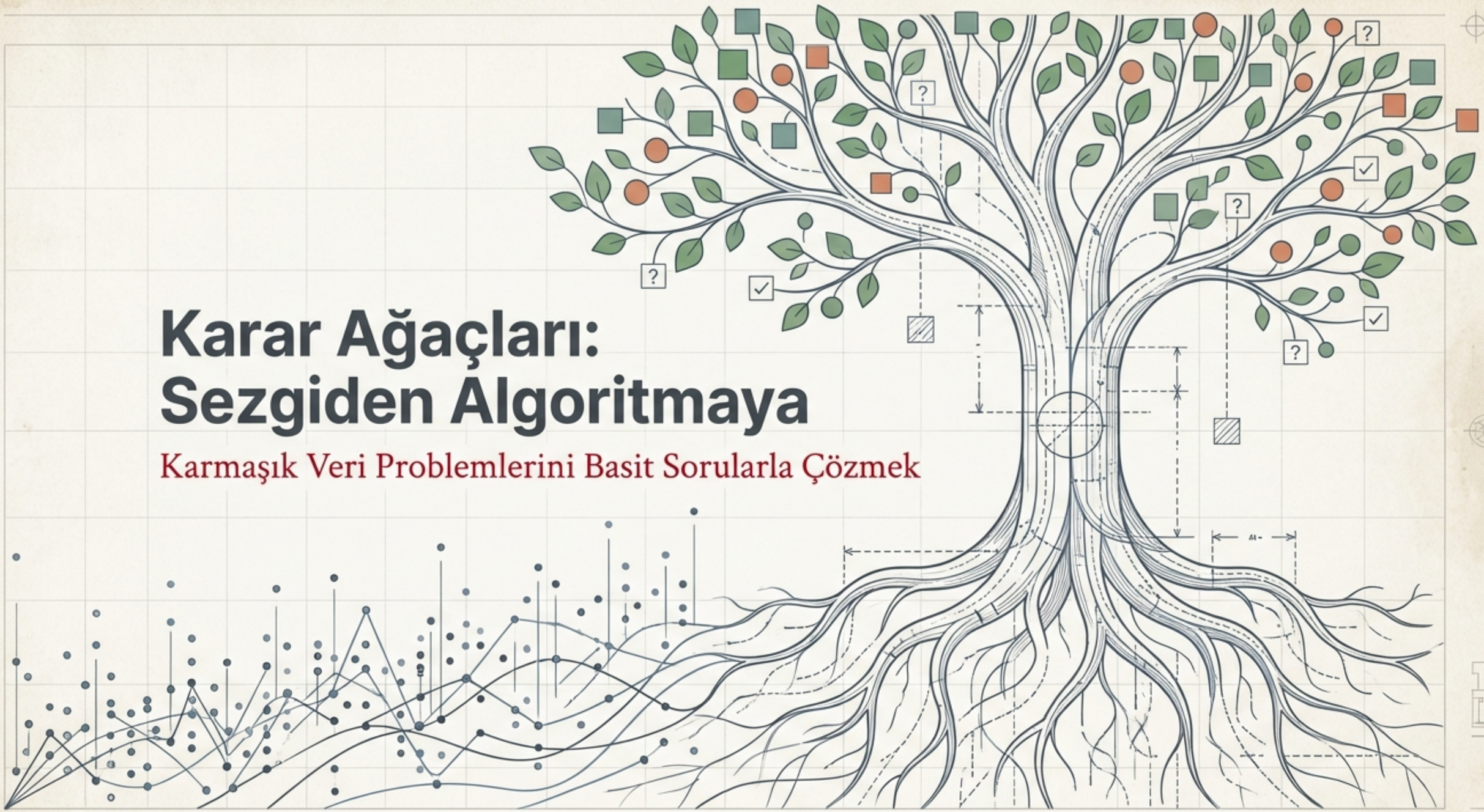


Karar Ağaçları: Sezgiden Algoritmaya

Karmaşık Veri Problemlerini Basit Sorularla Çözmek

GÖZETIMLI ÖĞRENME SERISI

NotebookLM

Bir Banka Krediyi Nasıl Onaylar?

Müşteri Verisi Örneği

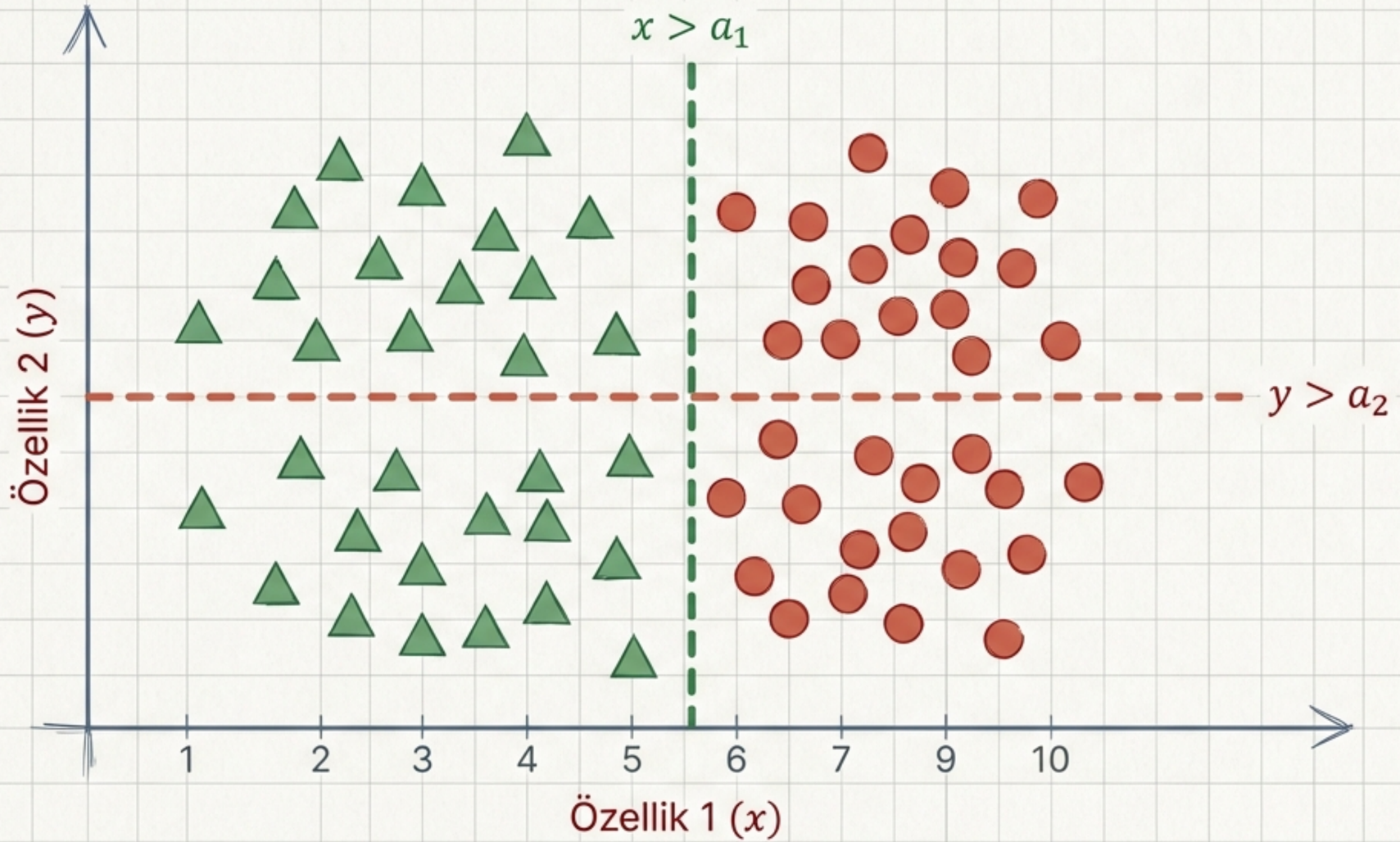
Kredi Skoru	Gelir	Vade	Karar (y_i)
Düşük	Yüksek	3 Yıl	?
Yüksek	Düşük	5 Yıl	?
Orta	Orta	3 Yıl	?

1. Hangi özellik kredi riskini en iyi tanımlar?
2. Önce gelire mi bakmalıyız, yoksa kredi geçmişine mi?
3. Bu müşteri ‘Güvenli’ mi yoksa ‘Riskli’ mi?

Anno: Karar Ağaçları, bu insan mantığını (soru-cevap zinciri) taklit ederek çalışır.

NotebookLM

Böl ve Yönet: Veri Uzayının Geometrisi

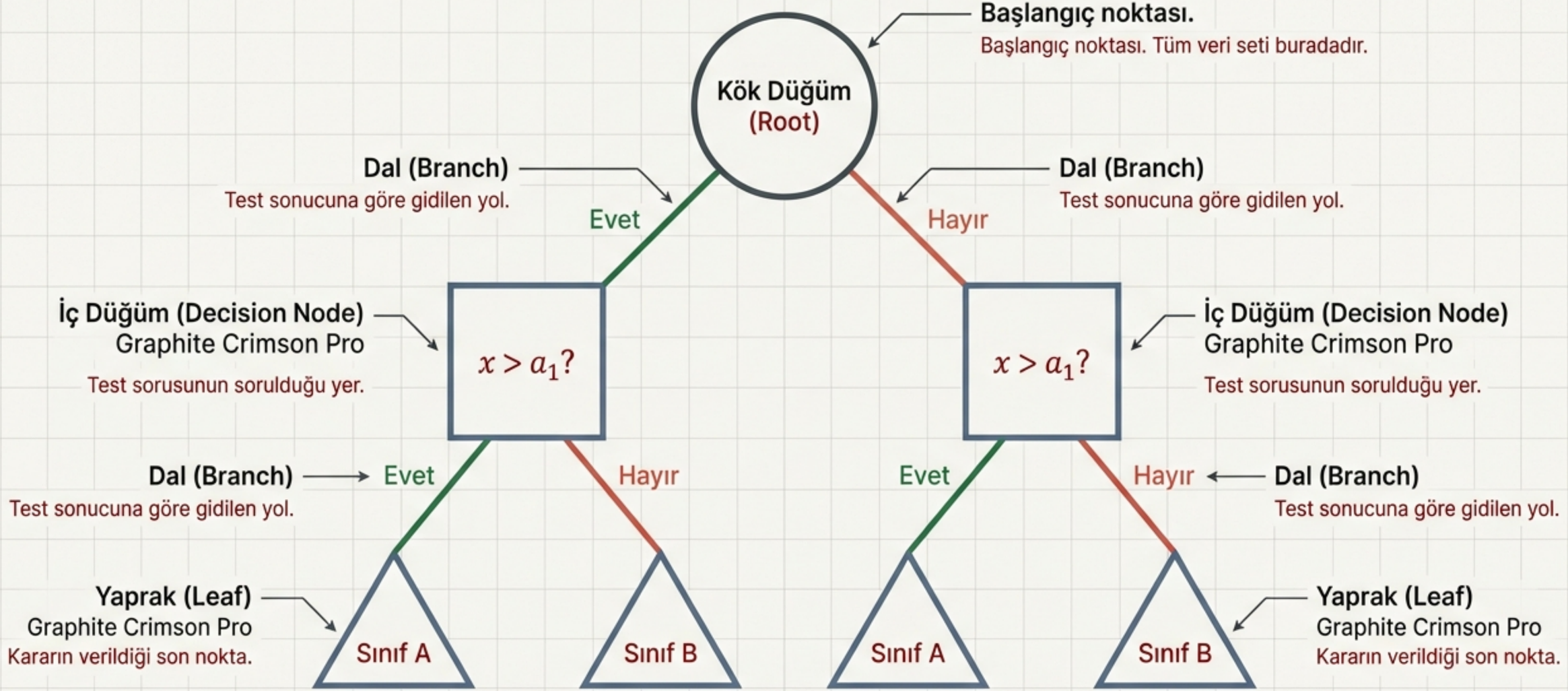


Eksenlere Paralel Bölünme

- Karar Ağaçları, karmaşık bir veri uzayını, her biri tek bir sınıfa ait olana kadar daha küçük dikdörtgenlere böler.
- Her kesik, ağaçtaki bir karar düğümüne karşılık gelir.

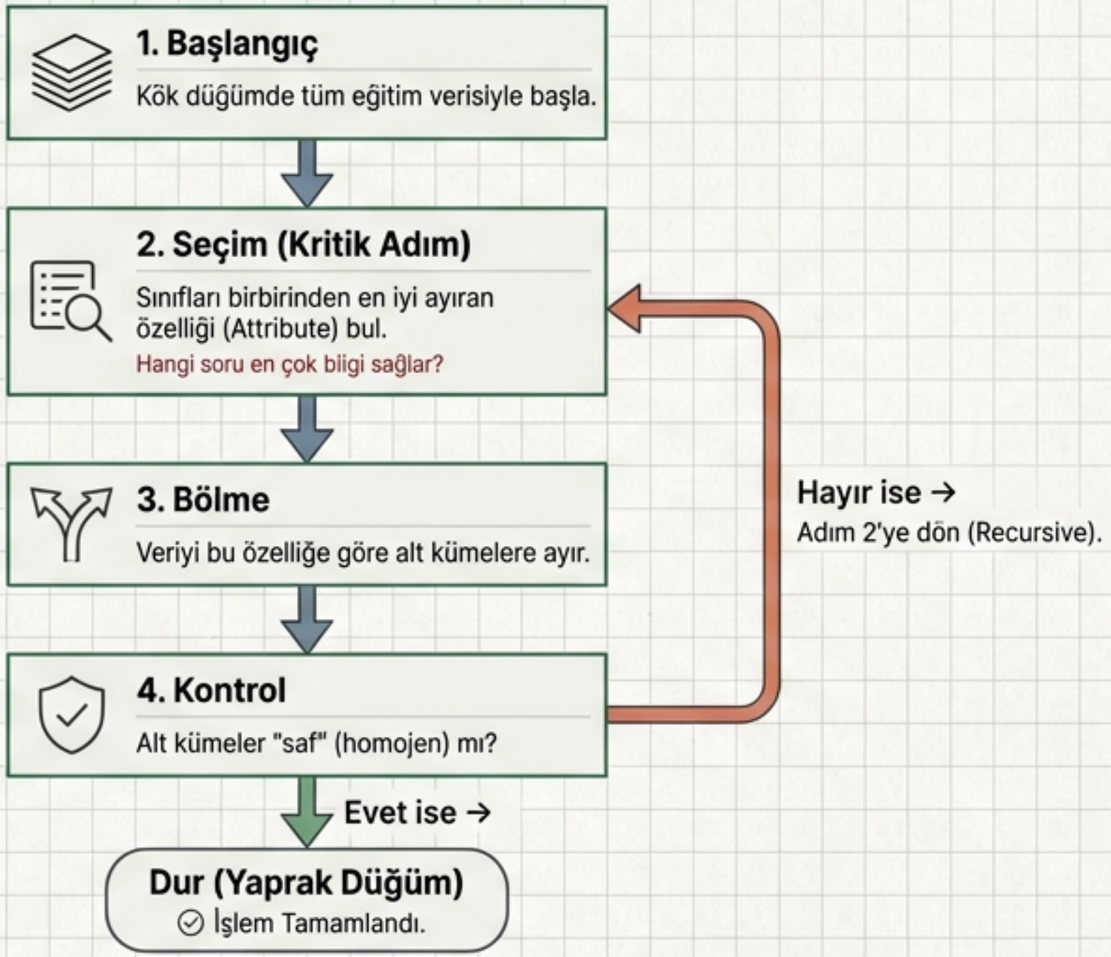
NotebookLM

Bir Kararın Anatomisi



NotebookLM

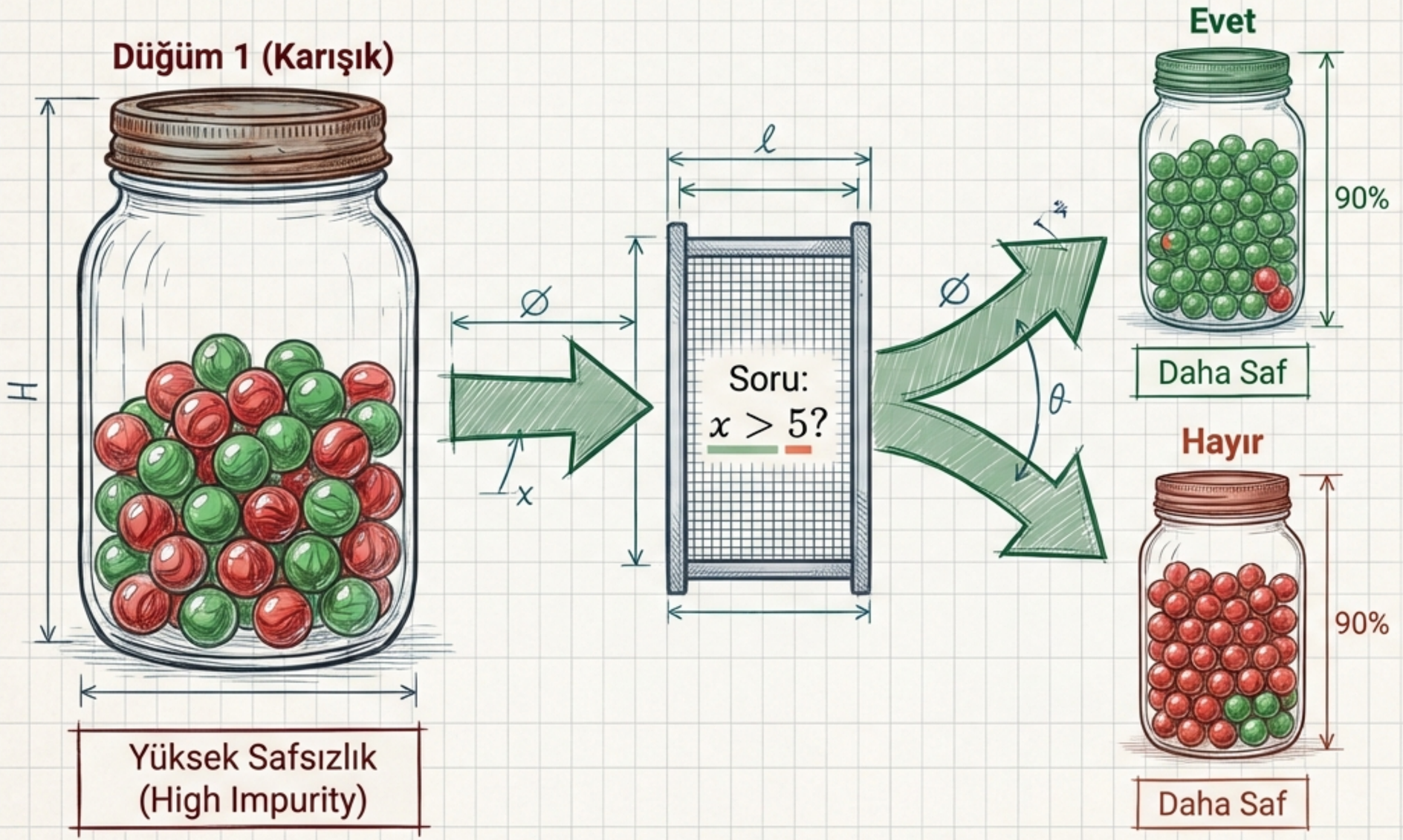
Ağaç Nasıl Öğrenir?



Temel Soru: Sınıfları “en iyi” ayıran özellik hangisidir? Bu bizi matematiksel optimizasyona götürür.

NotebookLM

Amaç: Veriy Veriyi Temizlemek ve Saflaştırmak



Tanım:
Safsızlık (Impurity): Bir düğümdeki verilerin ne kadar karışık olduğunu ölçer.

Hedef:
Her yaprakta mümkün olduğunca homojen (tek tip) veri elde etmek.

Prensip:
Benzer gözlemleri benzer sınıflarda gruplandır.

NotebookLM

Matematiği Anlamak: Gini İndeksi

$$Gini(S) = 1 - \sum (p_i)^2$$

S: Veri seti

p_i: 'i' sınıfının olasılığı (göreceli frekansı)

Gini İndeksi, rastgele seçilen bir öğenin yanlış sınıflandırılma olasılığını ölçer. Ağaç, bölünme sonrası ağırlıklı ortalama Gini skorunu en aza indiren soruyu seçer.

0.0 (Mükemmel Saflık)

Tüm veri tek sınıfta.

0.5 (Maksimum Safsızlık)

Sınıflar eşit dağılmış (50/50).

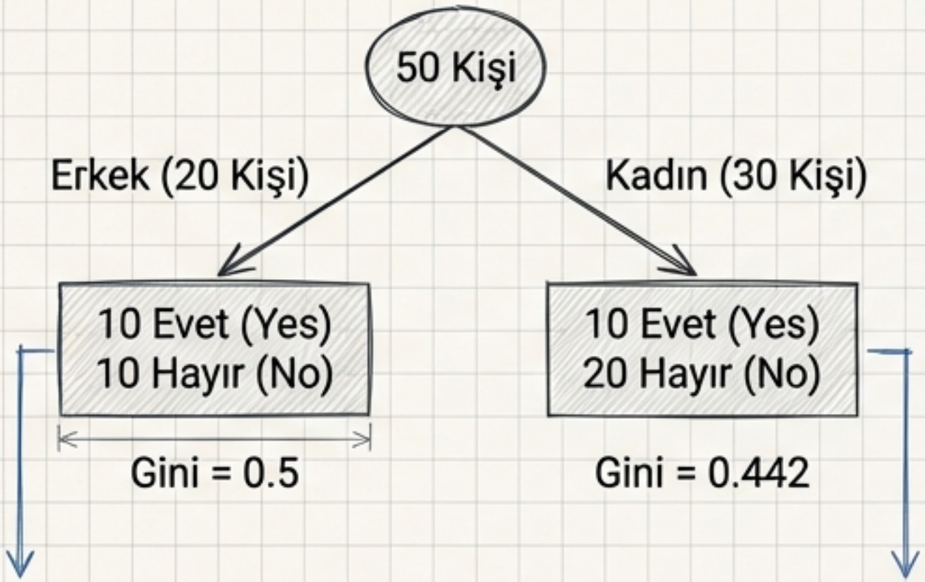
NotebookLM

Hesaplama Örneği: Cinsiyet mi, Yaş mı?

Hedef: Bilgisayar Kullanımı (50 Kişi)

A

Bölünme: CİNSİYET (Gender)

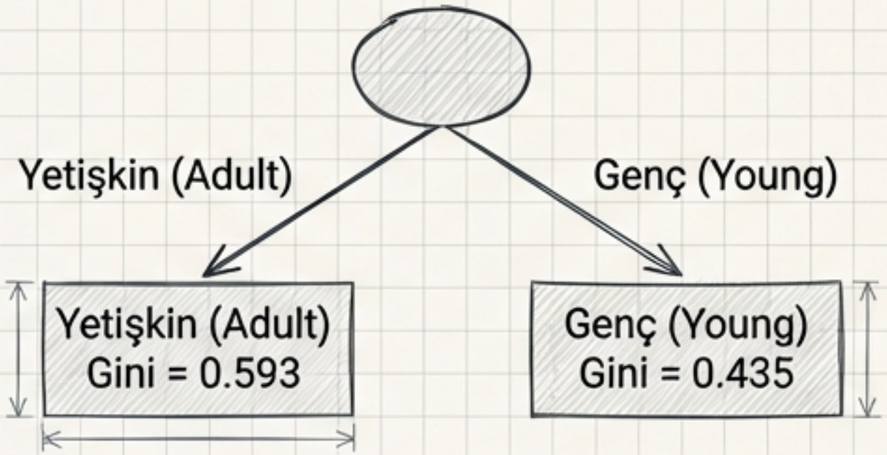


Ağırlıklı Ortalama Gini = (20/50) × 0.5 + (30/50) × 0.442

= 0.465 Kazanan (Winner)

B

Bölünme: YAŞ (Age)



Ağırlıklı Ortalama Gini = 0.514

$0.465 < 0.514. Gini İndeksi daha düşük olduğu için algoritma 'Cinsiyet' sorusunu seçer.

NotebookLM

Alternatif Yaklaşım: Bilgi Kazancı (Information Gain)



Entropi: Verideki belirsizliğin veya düzensizliğin ölçüsüdür.

Bilgi Kazancı: Bir bölünme öncesi ve sonrası entropi arasındaki farktır.

$$ Bilgi Kazancı = Entropi(Önce) - Entropi(Sonra) $$

ID3 ve C4.5 algoritmaları, Bilgi Kazancını **maksimize** eden bölünmeyi arar.

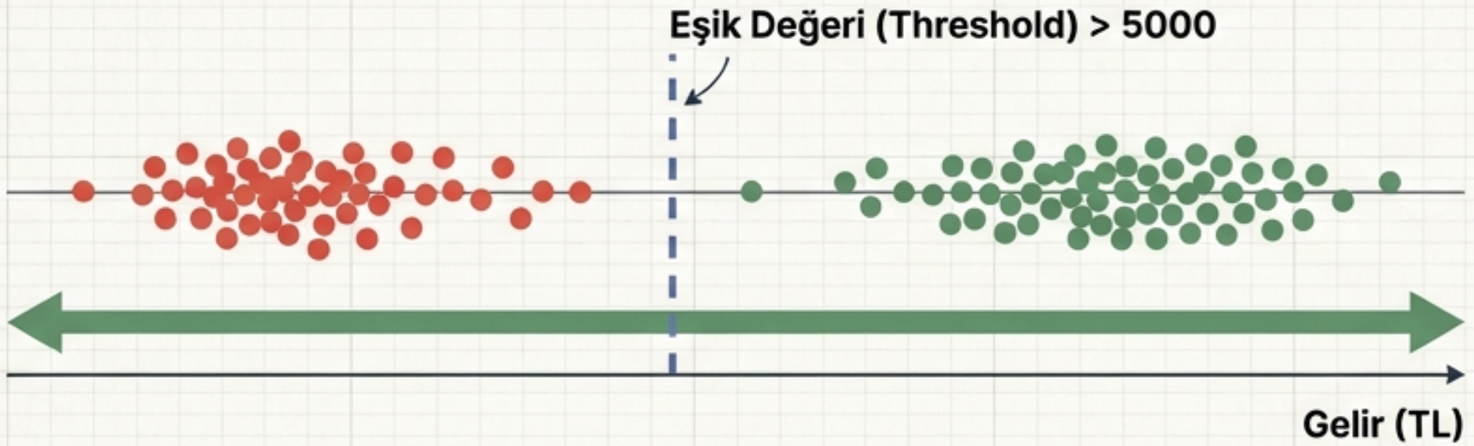
NotebookLM

Veri Tipleriyle Çalışmak

Kategorik Veri

Doğal gruplar (Örn: Hava = Güneşli, Yağmurlu). Kategori seçimi doğrudandır.

Sayısal Veri ve Eşik Değerleri



Sayısal verilerde algoritma, saflığı en iyi sağlayan kesme noktasını (breakpoint) tüm olası değerleri tarayarak bulur.

Breakpoint for a numerical attribute

NotebookLM

Algoritmalar Savaşı: ID3, C4.5 ve CART

Feature	ID3	C4.5	CART
Veri Tipi	Sadece Kategorik	Sayısal & Kategorik	Sayısal & Kategorik
Kriter (Metric)	Bilgi Kazancı (Entropi)	Bilgi Kazancı (Entropi)	Gini İndeksi
Ağaç Yapısı	Çoklu Dallar	Çoklu Dallar	İkili Ağaç (Binary) - Daima 2 Dal
Eksik Veri	İşleyemez	İşleyebilir	İşleyebilir

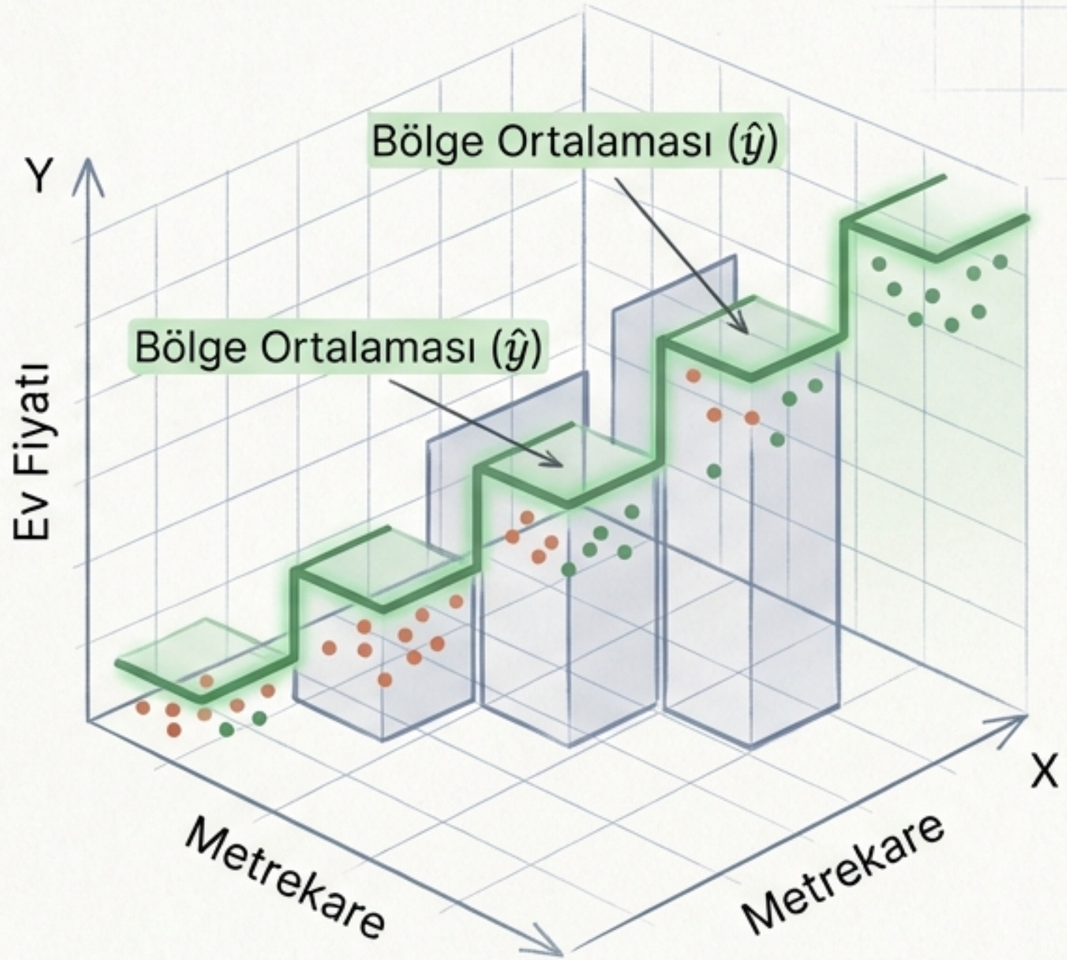
NotebookLM

Sınıflandırmanın Ötesinde: Regresyon Ağaçları

- **Problem:** Çıktı bir sınıf (**Evet/Hayır**) değil, sürekli bir sayı (Fiyat) ise?

  Çözüm: Veri uzayı yine bölgelere ayrılır. Ancak yaprak düğümde bir sınıf etiketi yerine, o bölgeye düşen eğitim verilerinin ortalaması alınır.

$$\hat{y} = \frac{1}{m} \sum_{i} y_i$$



EDITORIAL BLUEPRINT
NotebookLM

Neden Karar Ağaçları?



Beyaz Kutu (White-Box)

Yorumlanabilirlik. Kararın nedeni açıkça izlenebilir. Bir 'Kara Kutu' (Black Box) değildir.



Veri Hazırlığı

Minimum ön işleme. Verinin normalize edilmesine gerek yoktur. Sayısal ve kategorik veriyle aynı anda çalışır.



Uygulama Alanları

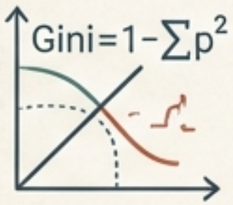
Tıbbi teşhis, Kredi risk analizi, Müşteri segmentasyonu.

NotebookLM

Özet ve Çıkarımlar



- **İnsan Odaklı:**
  İnsan mantığını taklit eden 'Böl ve Yönet' stratejisi.

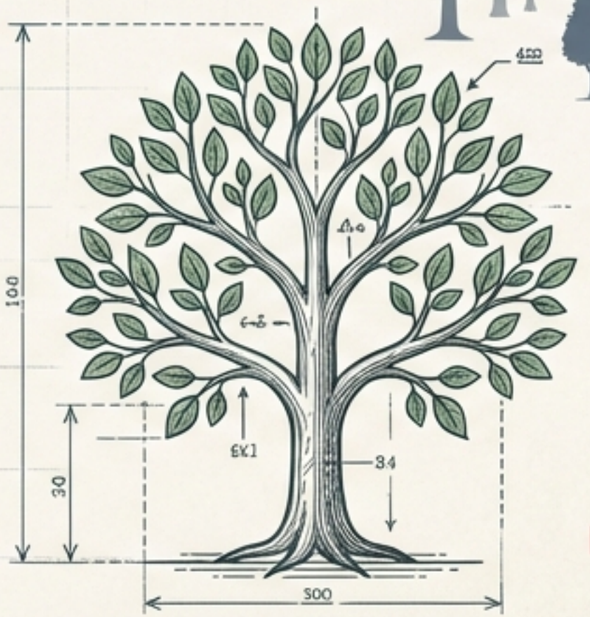


- **Matematiksel Kesinlik:**
  Gini veya Entropi ile optimize edilmiş karar sınırları.



- **Şeffaflık:**
  Kararların neden alındığını açıklayabilen 'Beyaz Kutu' modelleri.





Sırada ne var?
Topluluk Öğrenmesi (Random Forests).

Tek bir ağaç veriyi ezberlemeye (overfitting) meyillidir. Modern veri biliminde genellikle yüzlerce ağacın bir araya geldiği 'Ormanlar' kullanılır.

NotebookLM