Veri Bilimi – Hızlı Bir Bakış

Kaan Öztürk
Yolcu 360

Ka|Ve 2019 Karmaşık Sistemler ve Veri Bilimi Yaz Okulu
19.7.2019

Karmaşık Sistemler ve Veri Bilimi: Kişisel bir bakış

- Bilimde en hakiki mürşit veridir.

Karmaşık Sistemler ve Veri Bilimi: Kişisel bir bakış

- Bilimde en hakiki mürşit veridir.
- Matematiksel modeller gerçek hayatla karşılaştırılmalıdır.

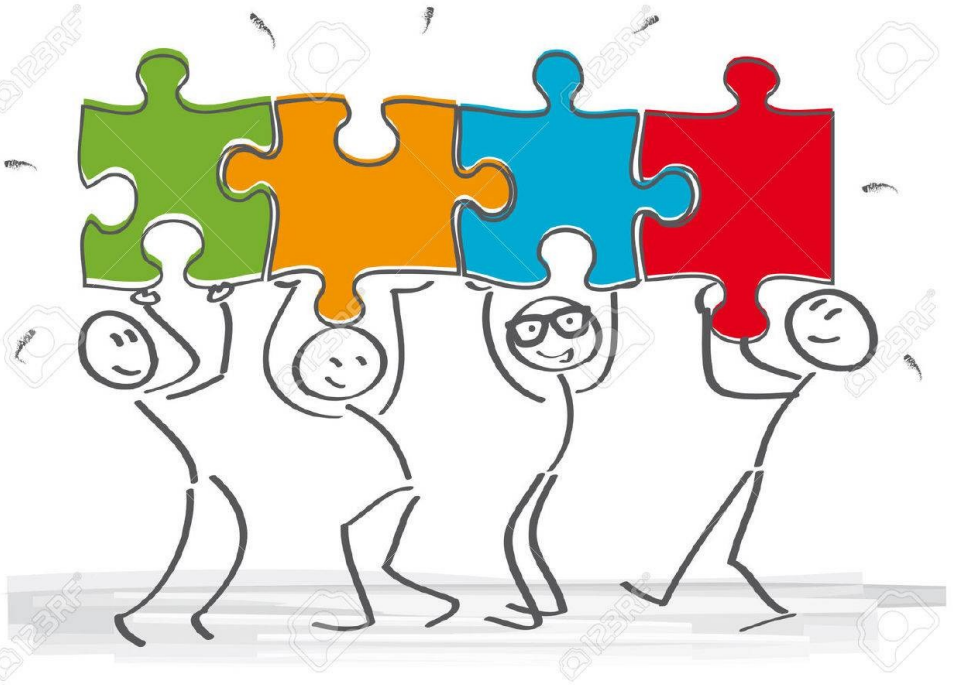
Karmaşık Sistemler ve Veri Bilimi: Kişisel bir bakış

- Bilimde en hakiki mürşit veridir.
- Matematiksel modeller gerçek hayatla karşılaştırılmalıdır.

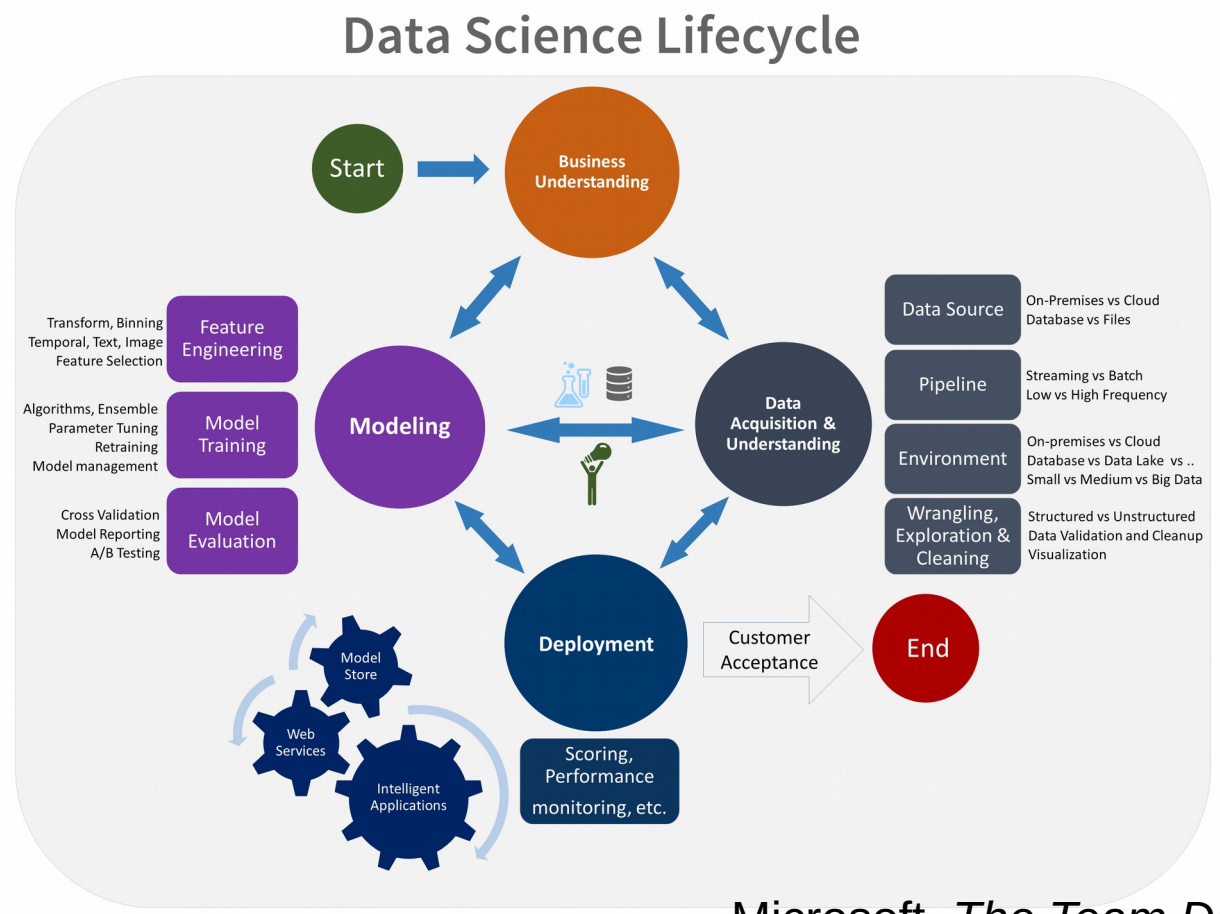
“The subtle tongue, the sophist guile, they fail when the broadswords sing.”
-- Robert E. Howard

Veri Biliminin Parçaları

- Bilgisayar bilimi
- Veri Mühendisliği
- İstatistik
- Yapay Öğrenme



Veri Bilimi İş Akışı



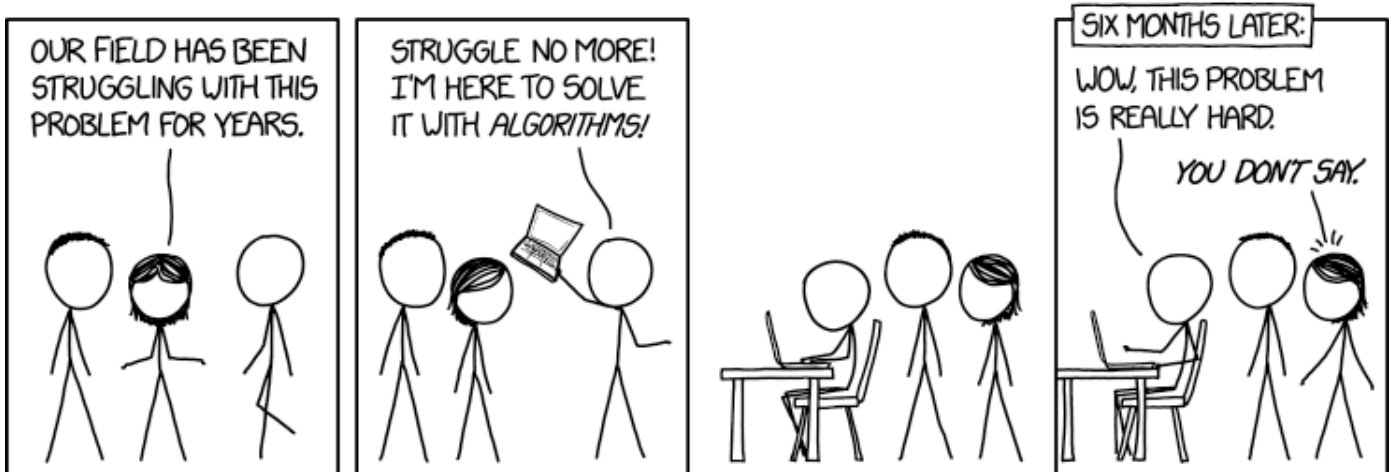
Microsoft, *The Team Data Science Process*

Yapay Öğrenme nedir?

- "Machine Learning" veya "istatistiksel öğrenme"
  - Kısmen istatistik, kısmen bilgisayar bilimi.
- "Açıkça programlanmadan öğrenebilme yeteneği."

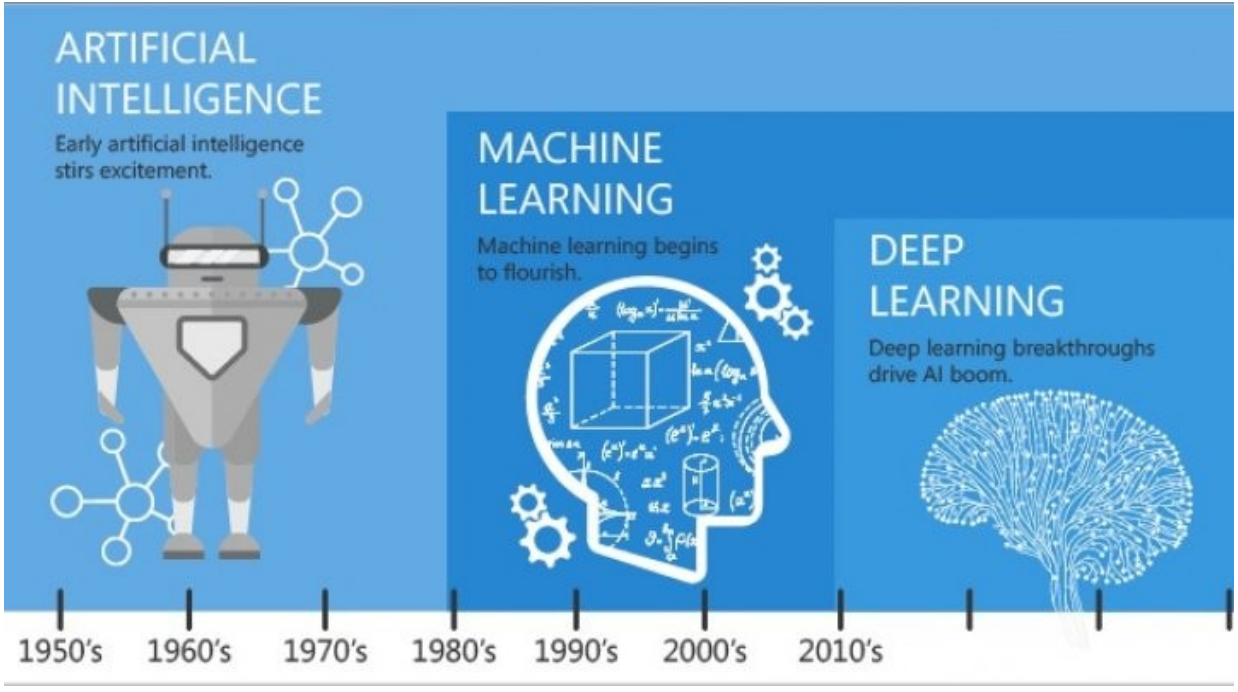
Yapay Öğrenme ne değildir?

- Hazır kurallar dizisi değildir. Kuralları keşfeder.
- Sihirli değnek değildir, sınırları vardır.
- Uzmanların yerine geçmez; işbirliği yapar.



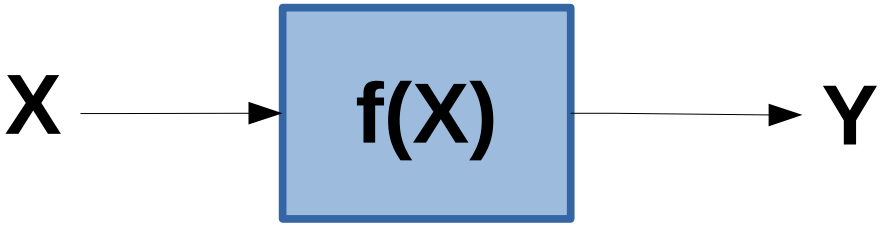
Yapay Öğrenme ve Derin Öğrenme

- Yapay Zeka > Yapay Öğrenme > Derin Öğrenme



Öğrenme

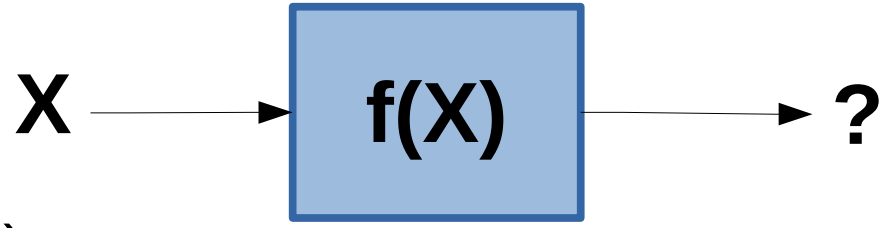
- "Model": Doğrusal bağlanım, lojistik bağlanım, karar ağacı, sinir ağı, Bayes ağları, …



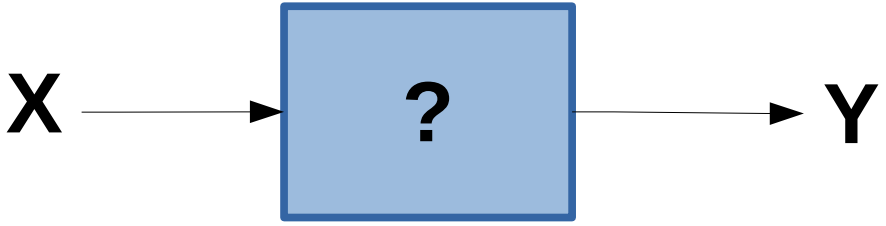
- Bir modeli "öğrenmek" = modeli veriye en iyi uyacak biçimde ayarlamak.

Öğrenmenin iki amacı

- Tahmin (prediction)
  - Yeni bir veri noktası verildiğinde, çıktı değerini kestirmek.



- Çıkarım (inference)
  - Çıktının kestiricilere nasıl bağlı olduğunu anlamak

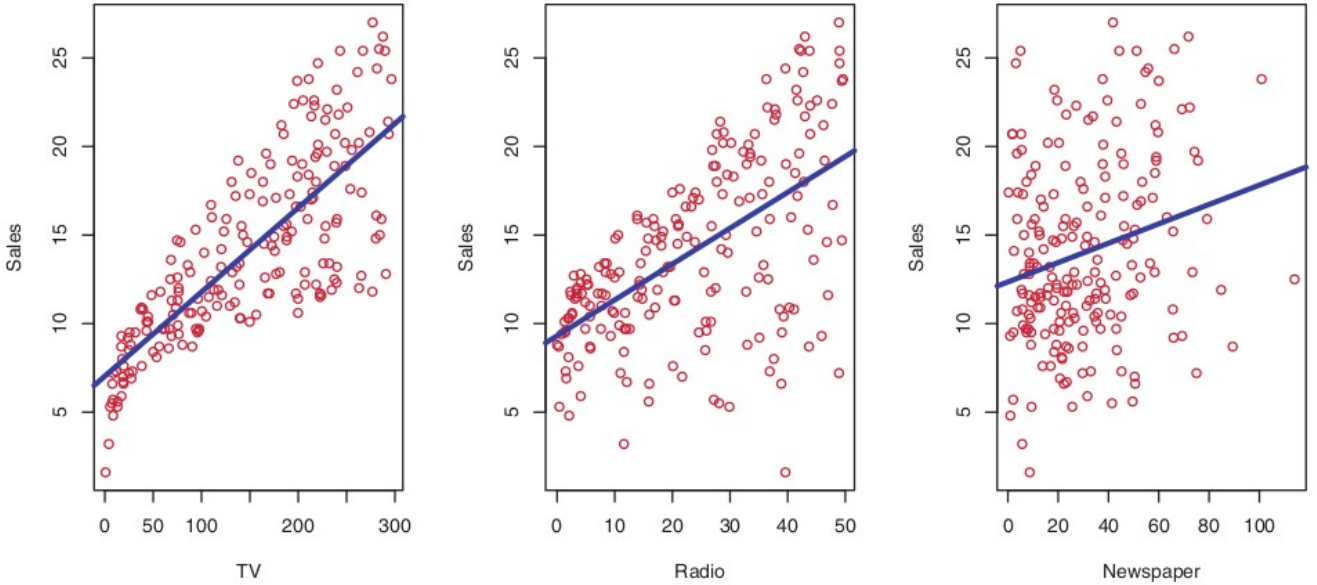


Öğrenme tipleri

- Güdümlü öğrenme
- Güdümsüz öğrenme
- Yarı güdümlü öğrenme
- Pekiştirmeli öğrenme

Değişkenler

- Giriş değişkenleri X1, X2, … Xn
  - Kestirici, bağımsız değişken, öznitelik,…
- Çıkış değişkeni Y
  - Çıktı, bağımsız değisken. hedef

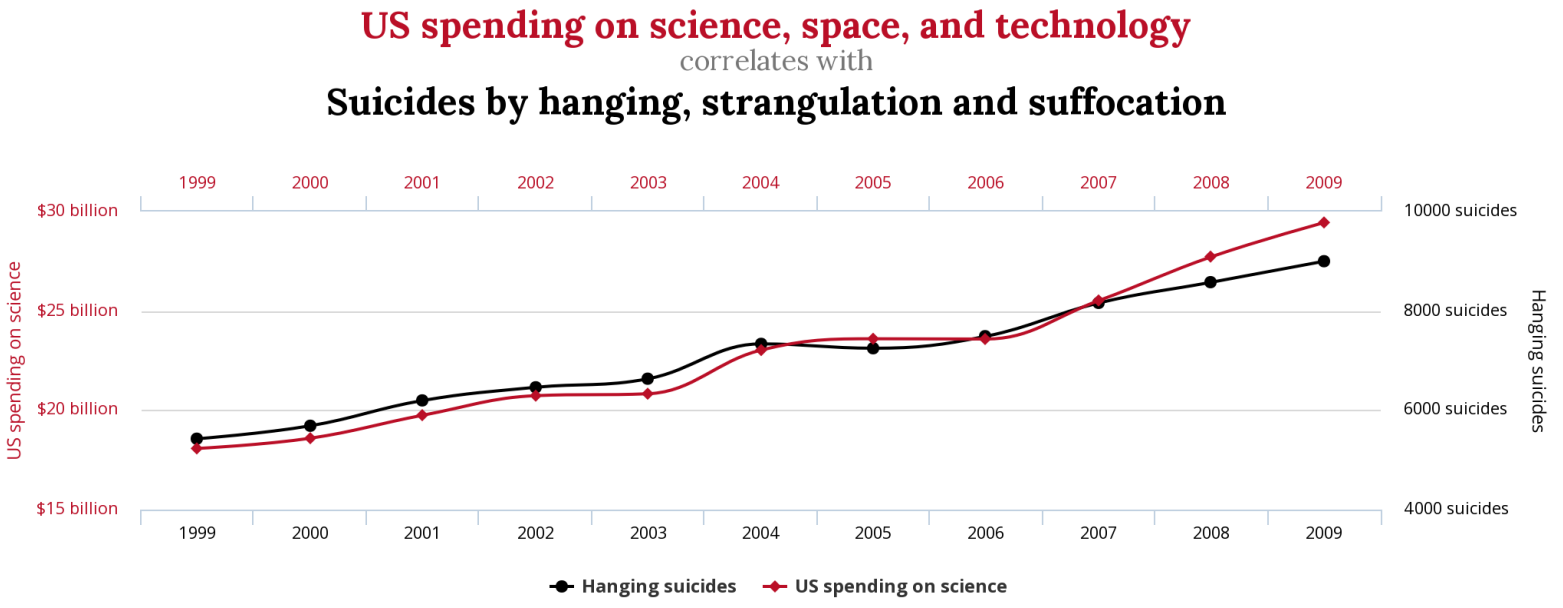


Öznitelik seçimi

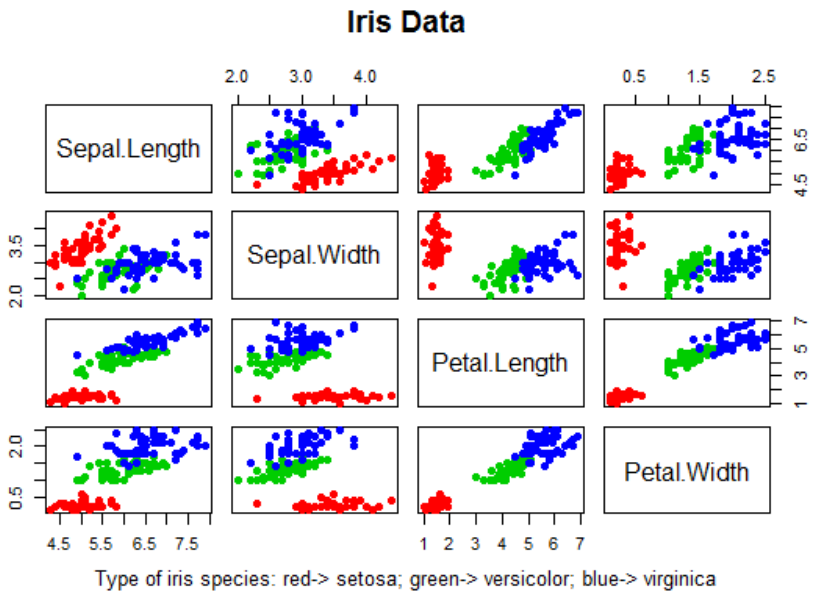
- Çok sayıda bağımsız değişken varsa
  - İyi: Model tam.
  - Kötü: Aşırı öğrenme, yorumlama zorluğu.
- Boyut indirgeme
- Düzenlileştirme
- Otomatik seçim

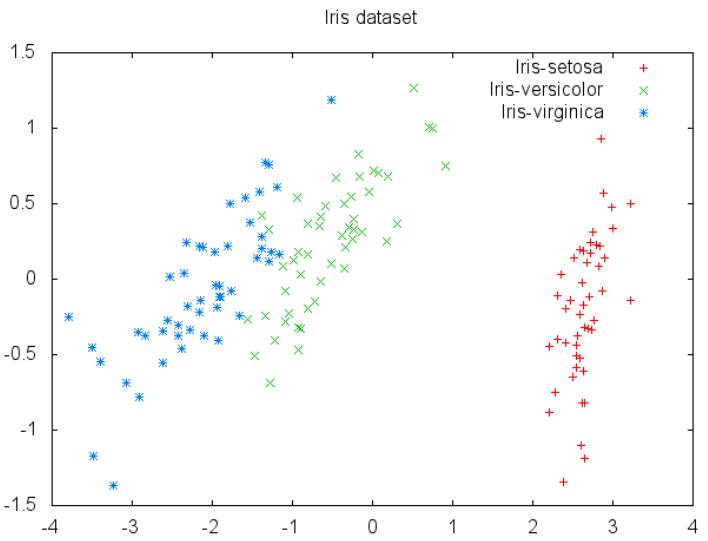
Tesadüfi bağlantılara dikkat!

- Çok sayıda değişken varsa, bazıları arasında yanlış ilişki bulunması kaçınılmaz.



Boyut indirgeme





Model seçimi

Güdümlü

Bağlanım

Çıktı değişkeni sürekli değer.
- Bir kişinin geliri
- Krediyi batırma olasılığı
- Bir evin değeri

Sınıflandırma

Çıktı değişkeni bir kategori.
- Kedi – kedi değil
- Sağlıklı – selim tümör – habis tümör
- 0 – 9

Güdümsüz

Çıktı değişkeni verilmemiş
- Müşteri katmanlaması
- Belge sınıflandırma
- İlişkilendirme
- Tavsiye sistemi

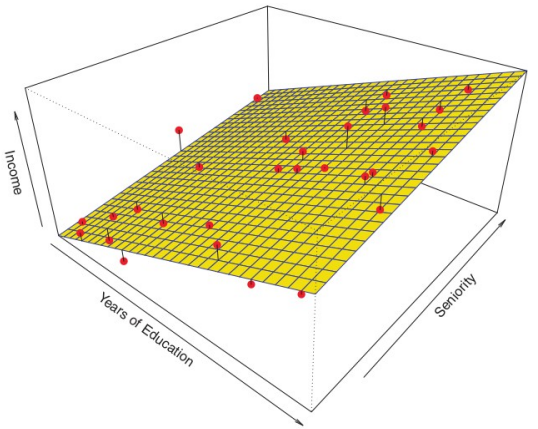
Doğrusal bağlanım

- Model: Hedef değişken, kestiricilere doğrusal şekilde bağlı.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

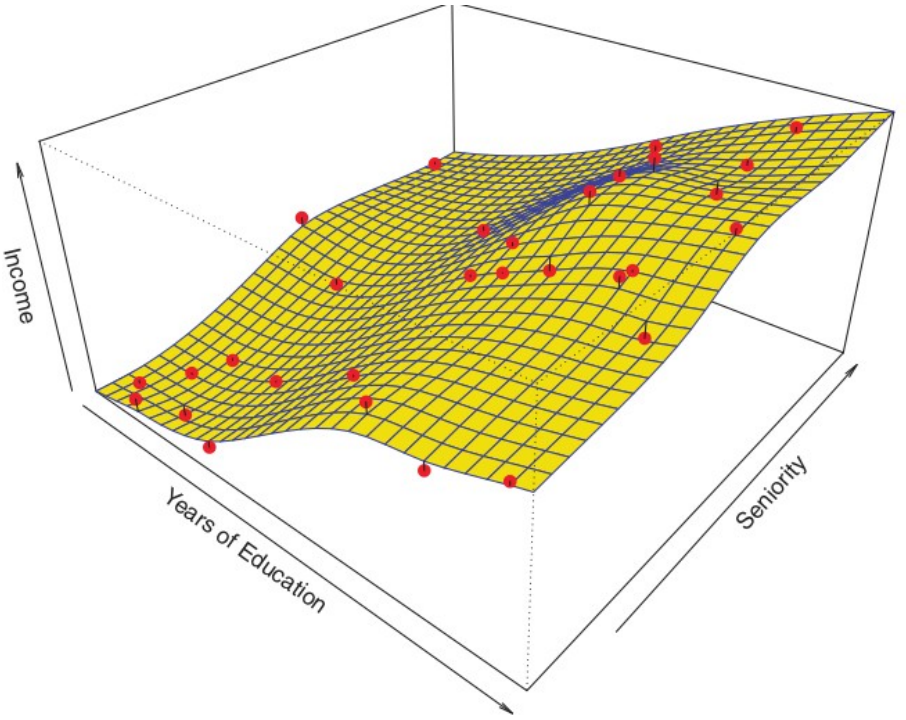
- Örnek:

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}$$



Spline bağlanım

- Parametrik olmayan model
  - Fonksiyonun biçimine dair açık varsayım yok.
  - Daha esnek; hatası düşük.
  - Ancak, yüksek doğruluk için çok sayıda veri gerekir.



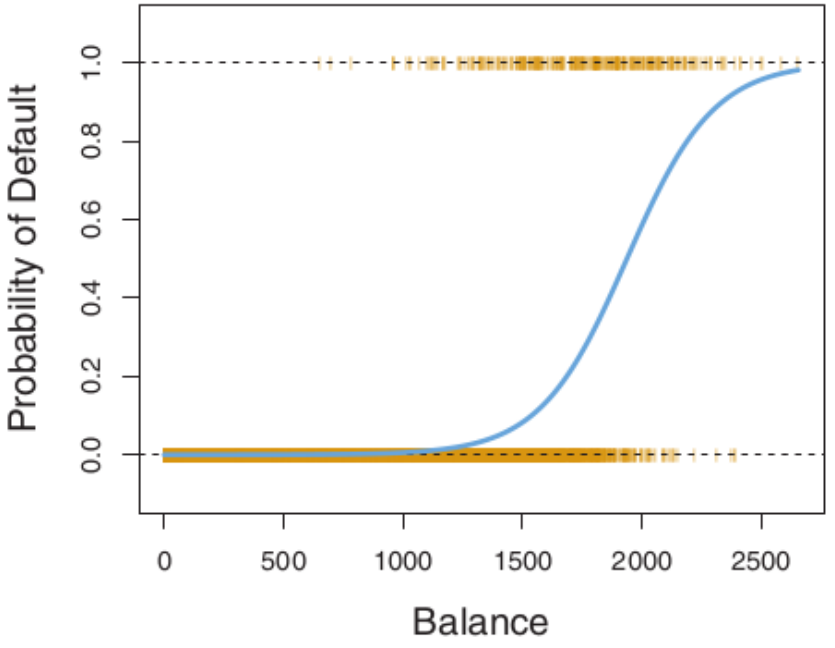
Lojistik bağlanım

- İkili sınıflandırma için olasılık modeli

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

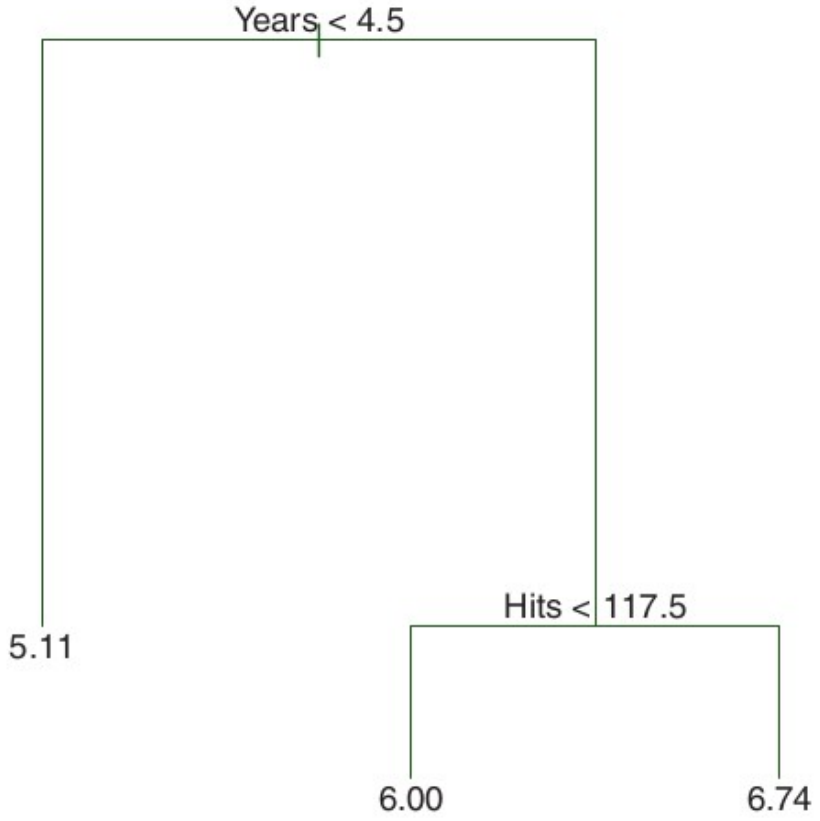
- Doğrusal.
- Genelleştirilebilir

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$



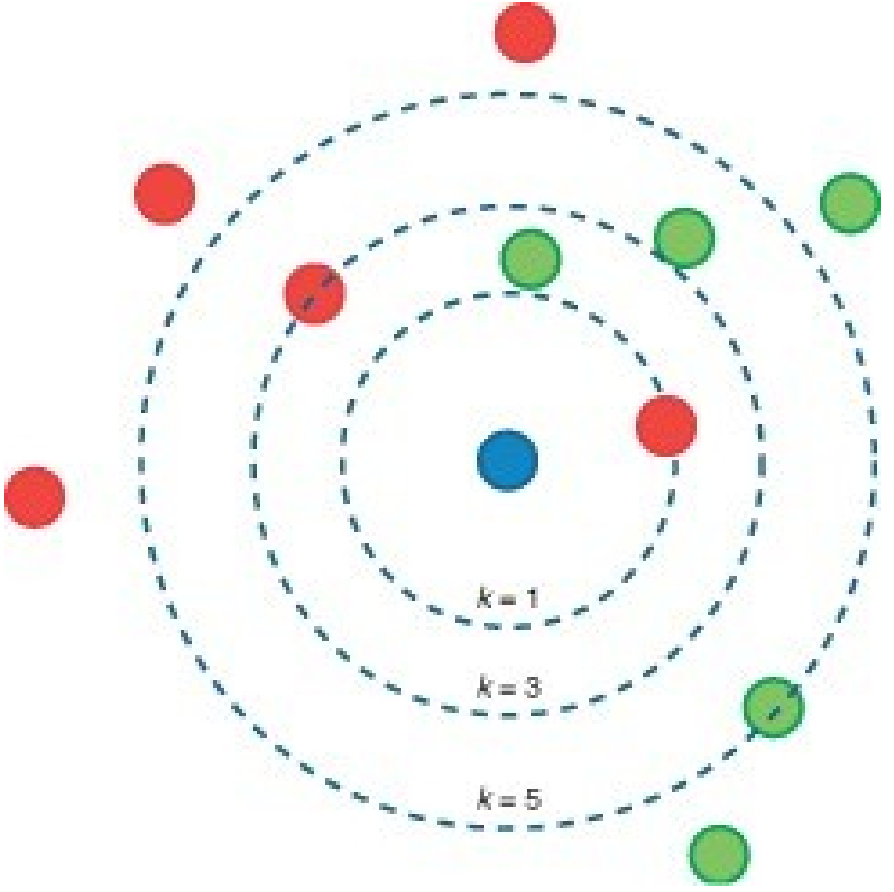
Karar ağaçları

- Ardışık evet/hayır soruları.
- Öznitelik uzayını defalarca ikiye böler.
- Nonparametrik.



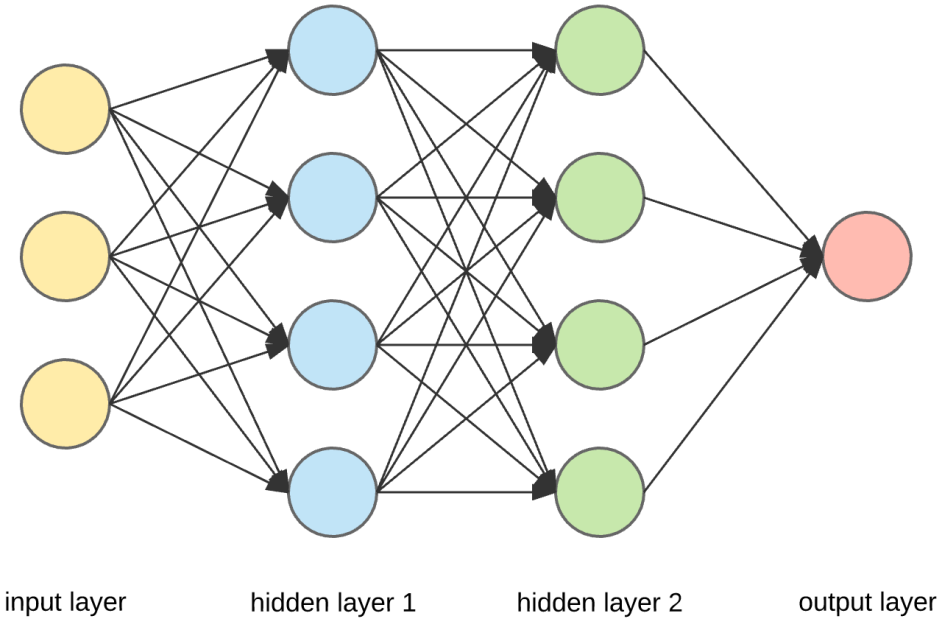
En yakın k komşu (k-NN)

- Yeni bir veri noktası verildiğinde en yakındaki *k* noktanın değerine bakılır.
- Çoğunluğun değeri kabul edilir.
- Nonparametrik.

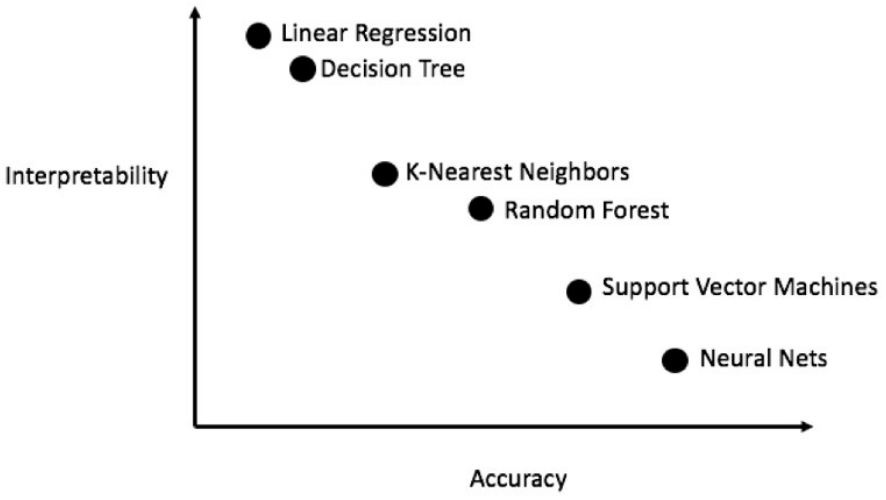


Yapay sinir ağları

- Girdi değerleri birleştirilerek üst seviye yapılar oluşturulur.
- Aradaki katman sayısına göre karmaşıklık artar.
- Etkileşimli demo: https://playground.tensorflow.org



Modellerin açıklama gücü



- Basit, katı, parametrik modelleri yorumlaması kolaydır, ama daha fazla hata yaparlar.
- Parametrik olmayan, karmaşık ve esnek modeller daha doğru sonuç verirler ama yorumlanmaları zordur.

Optimizasyon

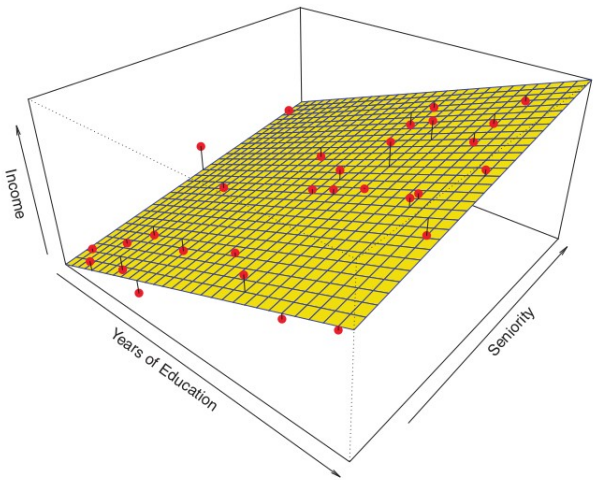
- Elimizdeki veriye **en iyi uyan** model parametreleri nedir?

$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}$$

?

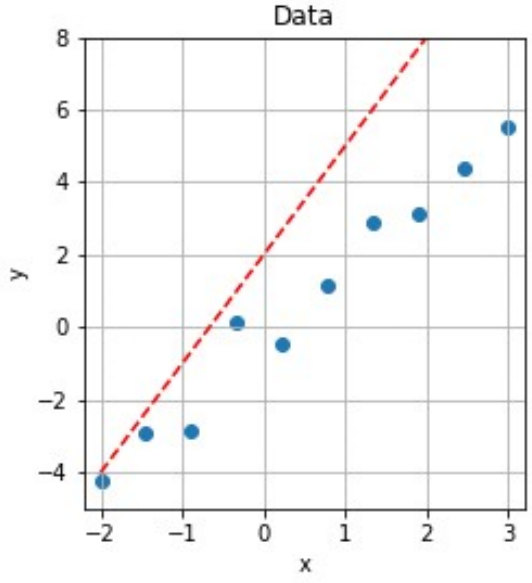
- Doğrusal bağlanım: *en küçük kareler*.

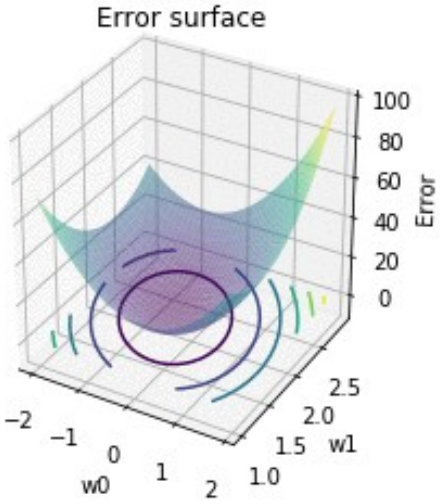
$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$



Bayır inişi

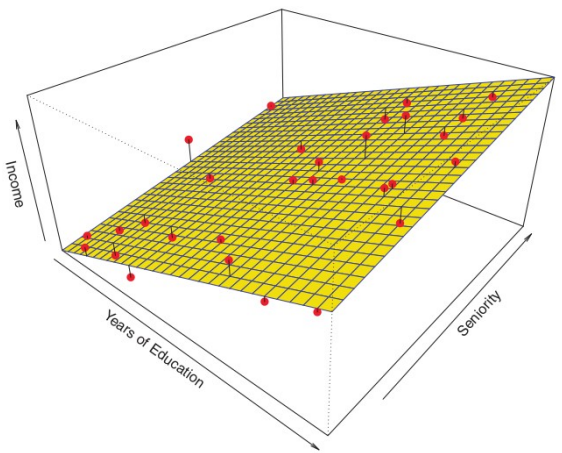
- Bir tahminle başla.
- Hatayı azaltacak yönde bir adım at.
- Gerektiği kadar tekrarla.

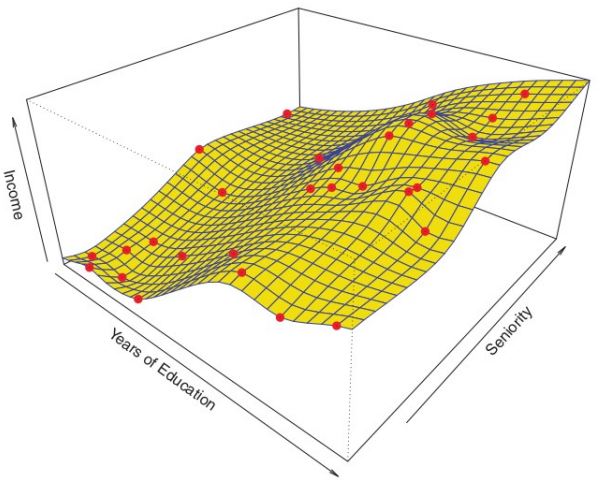




Eğitim hatası ve sınama hatası

- Eğitim hatasını asgariye indirmek yeterli değil.
- Sıfır eğitim hatası = ezberleme
- Eğitim dışı verilerdeki başarı daha önemli.

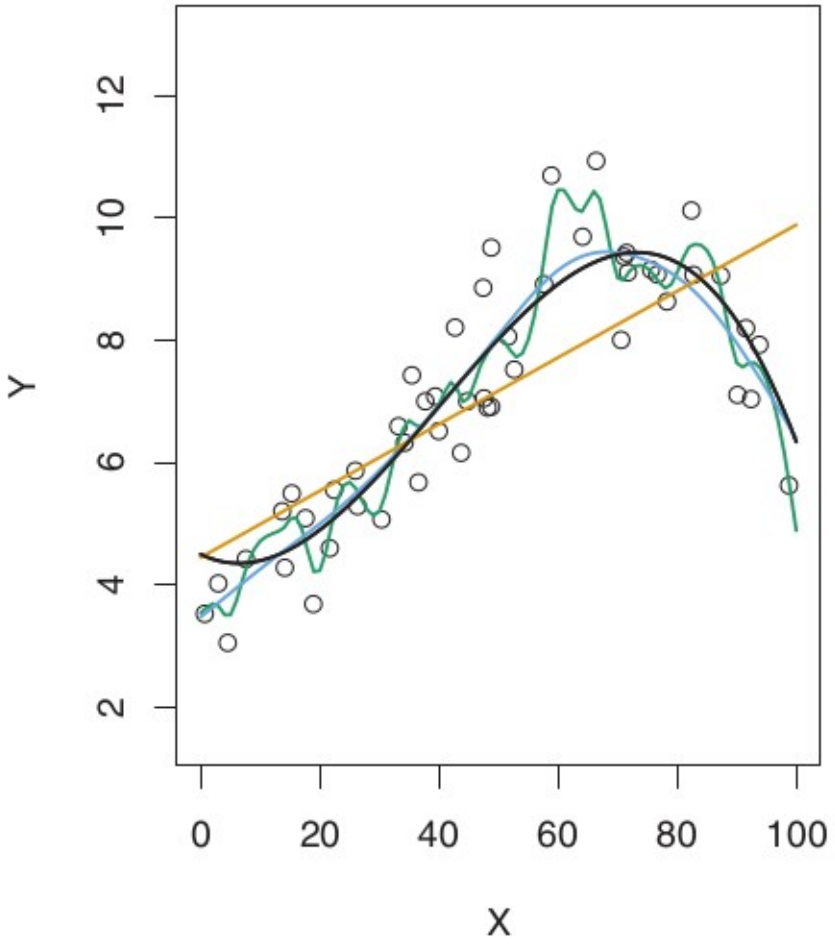


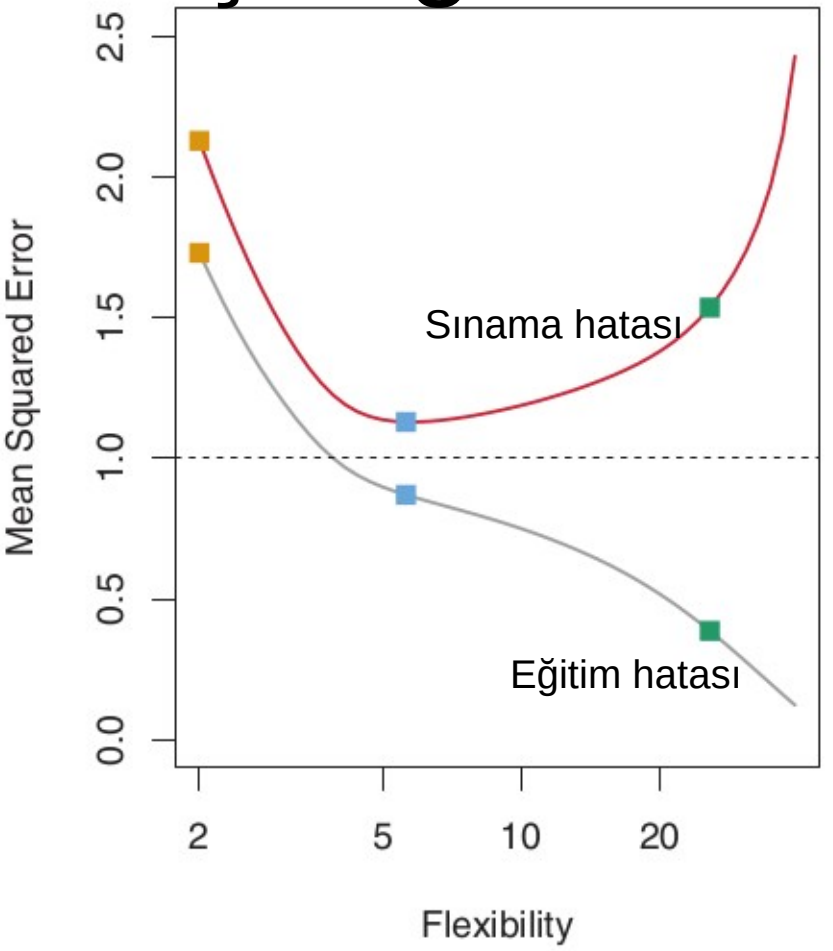


Sınama kümesi

- Veri **önceden** ikiye ayrılır
  - Eğitim kümesi
  - Sınama kümesi
- Sınama kümesi eğitimde **hiç kullanılmaz**.
- Modelin başarısını ölçmek için en sonda kullanılır.

Model karmaşıklığı



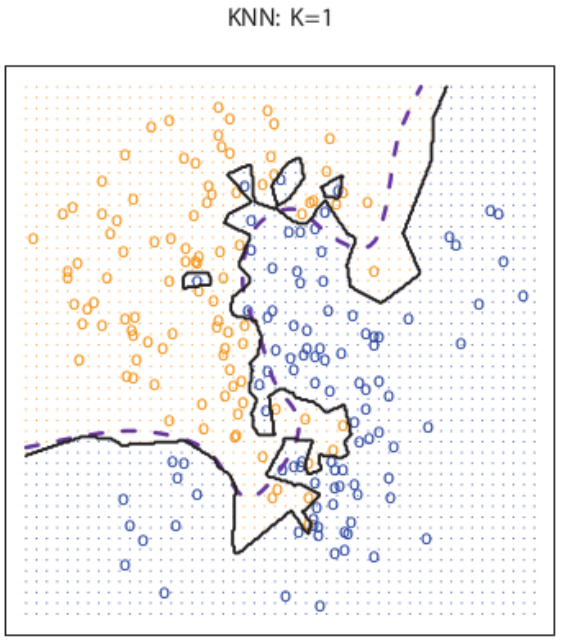


Aşırı Öğrenme

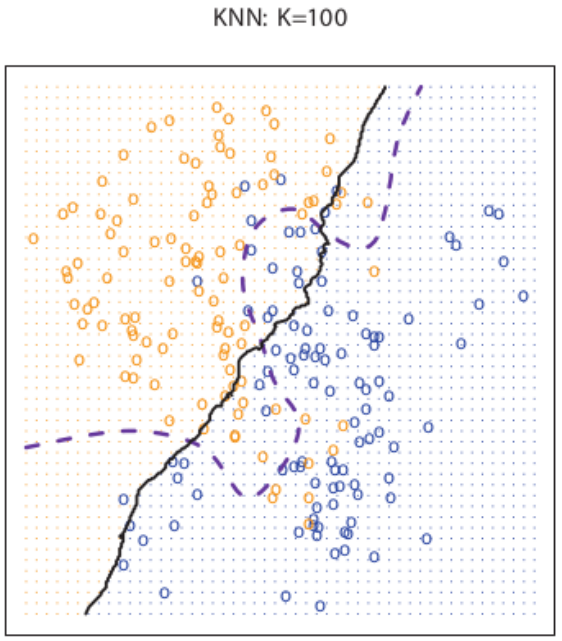
- Eğitim hatası küçük, sınama hatası büyük.
- Veri *ezberlenmiş*.
- Yeni verileri tahminde kullanılamaz.

Model karmaşıklığı – Sınıflandırma

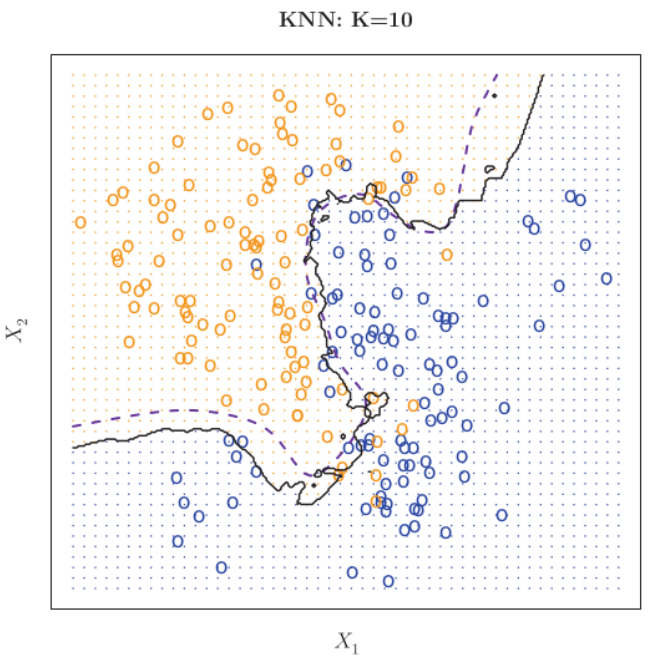
Aşırı öğrenme



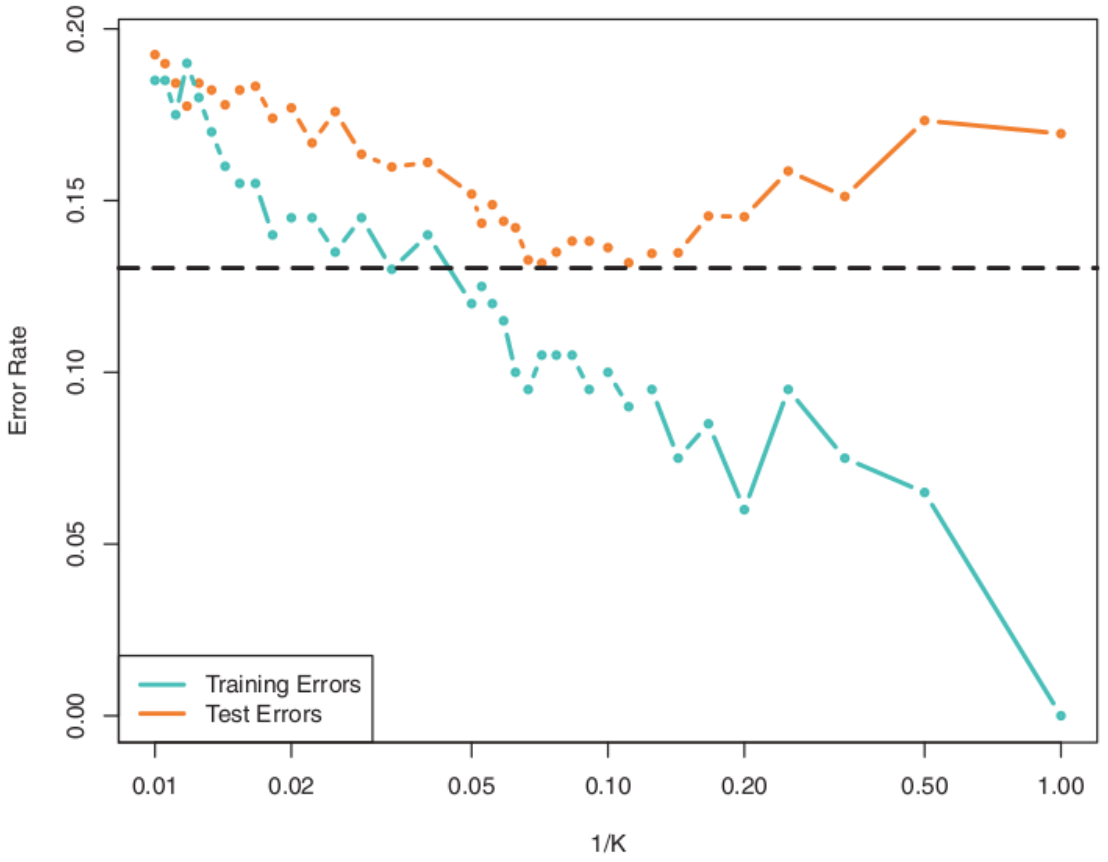
Eksik öğrenme



Uygun karmaşıklık



kNN - Eğitim ve sınama hatası



Aşırı öğrenmeyi engellemek

- Düzenlileştirme (regularization): Hata fonksiyonuna ek terim.

$$J(\beta_0, \beta_1) = \sum_{i=1}^{N} \left(\beta_1 x^{(i)} + \beta_0 - y^{(i)}\right)^2 + C\left(\beta_0^2 + \beta_1^2\right)$$

- Topluluk (ensemble) yöntemleri: Eğitim hatası yüksek birçok model eğit, ortalama al.
  - Rastgele Orman
  - XGBoost

Aşırı öğrenmeyi engellemek

- Topluluk (ensemble) yöntemleri
- Eğitim hatası yüksek birçok model eğit, ortalama al.
  - Rastgele Orman
  - XGBoost