EKONOMİ VE FİNANSTA MAKİNE ÖĞRENMESİ

HAFTA 2

# Makine Neyden Öğrenir?

## Veri, hedef ve öğrenme türleri

*Bir iktisat problemini makine öğrenmesinin diline çevirmek*

**Prof. Dr. Mehmet Songur**

Dicle Üniversitesi · İİBF · İktisat Bölümü

# Geçen haftadan

*İki farklı soru*



## Kestirim

x değişirse y ne olur?

*Nedensellik · ekonometri*



## Öngörü

x'i gördüğümde y ne olur?

*Tahmin · makine öğrenmesi*

**Bu hafta: Bir öngörü problemini kurmak için neye ihtiyacımız var? Hangi veri, hangi hedef, hangi hazırlık?**

# Bugün ne konuşacağız?

MEHMET Sonmez

1

## Öğrenme türleri

Denetimli, denetimsiz ve diğerleri

2

## Değişken ve hedef

Bir problemi kurmanın üç kararı

3

## Veri türleri

Kesit, zaman serisi, panel ve daha fazlası

4

## Veri hazırlama

Eksik veri, aykırı değer, ölçekleme, kodlama

# Bir tablo

*Bir bankanın altı müşterisi (kurgusal veri)*

| Müşteri | Yaş | İl | Aylık gelir (TL) | Kart harcaması (TL) | Önceki gecikme | Temerrüt |
|---|---|---|---|---|---|---|
| 1 | 34 | Diyarbakır | 28.000 | 9.500 | Hayır | Hayır |
| 2 | 51 | İstanbul | 85.000 | 31.000 | Hayır | Hayır |
| 3 | 23 | Mersin | 19.500 | 12.800 | Evet | Evet |
| 4 | 45 | Ankara | — | 7.200 | Hayır | Hayır |
| 5 | 29 | Diyarbakır | 32.000 | 268.000 | Evet | Hayır |
| 6 | 62 | İzmir | 41.000 | 5.100 | Hayır | Hayır |

**Bu tabloyla hangi soruları sorabiliriz?**

# Aynı tablo, üç soru

*Her soruda neyi tahmin etmeye çalışıyoruz?*



**Yeni bir başvuran temerrüde düşer mi?**



**Bir müşterinin kart harcaması ne kadar olur?**



**Müşterilerimiz kaç farklı profile ayrılıyor?**

# Aynı tablo, üç farklı öğrenme türü

MEHMET Sonmez



**Yeni bir başvuran temerrüde düşer mi?**

*Hedef: Temerrüt (evet/hayır)*

**Denetimli · Sınıflandırma**



**Bir müşterinin kart harcaması ne kadar olur?**

*Hedef: Harcama (TL)*

**Denetimli · Regresyon**



**Müşterilerimiz kaç farklı profile ayrılıyor?**

*Hedef sütunu yok*

**Denetimsiz · Kümeleme**

MEHMET Sonqur

# Öğrenme türünü veri değil, soru belirler.

*Aynı tablo; soruya göre sınıflandırma, regresyon ya da kümeleme problemi olabilir.*

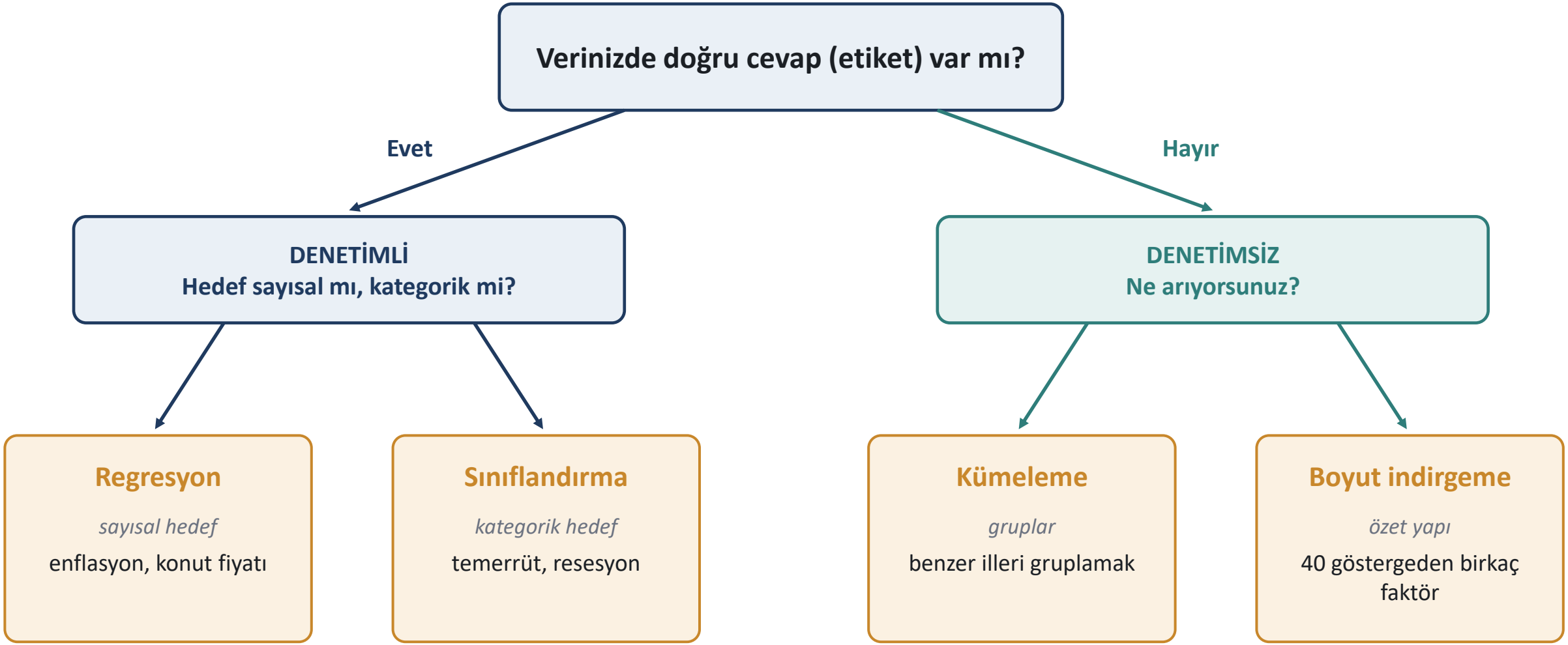
MEHMET Sonqur

# 1

# Öğrenme türleri

Problemi doğru kutuya yerleştirmek

# Hangi öğrenme türü? Bir karar ağacı



# Denetimli öğrenmenin iki yüzü

*Tek fark: hedef değişkenin türü*

| | Regresyon | Sınıflandırma |
|---|---|---|
| **Hedef** | Sayısal, sürekli | Kategorik (sınıflar) |
| **Örnek** | Gelecek ayın enflasyonu<br>Bir ilin işsizlik oranı | Temerrüt var / yok<br>Firma iflas eder mi? |
| **Modelin çıktısı** | Bir sayı: %2,4 | Her sınıf için bir olasılık: 0,18 |
| **Ekonometriden tanıdık** | Doğrusal regresyon | Logit ve probit |

# Üç yaygın yanılgı

MEHMET Sonjur



**“Lojistik regresyon bir regresyon yöntemidir.”**

Adına rağmen bir sınıflandırma yöntemidir: olasılık verir, sınıf tahmininde kullanılır. Ekonometrideki logit modeli.



**“Regresyon, doğrusal regresyon demektir.”**

ML’de regresyon yalnızca hedefin sayısal olduğunu söyler. Dönem içinde doğrusal olmayan regresyon yöntemleri de göreceğiz.



**“Hedef bir sayıysa problem regresyondur.”**

Kredi notunu 1–5 diye kodlamak onu sürekli yapmaz. Bu, sıralı bir kategoridir.

# Denetimsiz öğrenme: doğru cevap yok

MEHMET Sonjur

*Amaç, verideki gizli yapıyı bulmak*

## Kümeleme

**Birbirine benzeyen gözlemleri gruplamak.**

- 81 ili gelişmişlik göstergelerine göre gruplamak
- Banka müşterilerini harcama davranışına göre ayırmak

## Boyut indirgeme

**Çok sayıda değişkendeki ortak bilgiyi birkaç özet değişkende toplamak.**

- Faiz eğrisinin hareketinin büyük kısmı üç faktörle açıklanır: seviye, eğim, eğrilik

**Zorluk:** Başarıyı ölçecek doğru cevap yok. Kümelerin anlamlı olup olmadığına alan bilgisiyle karar verilir.

**Bir de pekiştirmeli öğrenme var:** Eylem yap, ödül al, zamanla öğren (ör. fiyatını gün içinde ayarlayan bir site). Bu derste kapsam dışı.

# Alıştırma: Hangi kutu?

*Karar ağacını kullanın*

MEHMET Sonjur

1. Bir ilin gelecek yılki işsizlik oranı
2. Bir firmanın gelecek yıl iflas edip etmeyeceği
3. 81 ili sosyoekonomik göstergelere göre gruplamak
4. 40 göstergeden tek bir “aktivite endeksi”
5. Merkez bankası metinlerini şahin/güvercin ayırmak (uzman etiketli örnekler var)
6. Fiyatları gün içinde ayarlayan e-ticaret sitesi
7. Bir ülkenin kredi notu (AAA, AA, A…)
8. BIST 100’ün yarınki hareketi

# Cevaplar

*Son madde: “Yükselir mi?” sınıflandırma, “Yüzde kaç değişir?” regresyon*

| | | |
|---|---|---|
| 1 | Bir ilin gelecek yılki işsizlik oranı | **Regresyon** |
| 2 | Bir firmanın gelecek yıl iflas edip etmeyeceği | **Sınıflandırma** |
| 3 | 81 ili sosyoekonomik göstergelere göre gruplamak | **Kümeleme** |
| 4 | 40 göstergeden tek bir “aktivite endeksi” | **Boyut indirgeme** |
| 5 | Merkez bankası metinlerini şahin/güvercin ayırmak (uzman etiketli örnekler var) | **Sınıflandırma** |
| 6 | Fiyatları gün içinde ayarlayan e-ticaret sitesi | **Pekiştirmeli** |
| 7 | Bir ülkenin kredi notu (AAA, AA, A…) | **Sıralı sınıflandırma** |
| 8 | BIST 100’ün yarınki hareketi | **Soruya bağlı!** |

MEHMET Sonjur

2

# Değişken ve hedef

Bir problemi kurmanın üç kararı

# Her ML problemi üç kararla kurulur

*Kredi örneği üzerinden*

1



## Gözlem birimi

**Her satır neyi temsil ediyor?**

*Bir müşteri mi, bir başvuru mu, bir müşteri-ay mı?*

2



## Hedef (y)

**Neyi öngörmek istiyoruz?**

*Temerrüt: 12 ay içinde 90 günü aşan gecikme*

3



## Öznitelikler (X)

**Neyi kullanarak öngöreceğiz?**

*Yaş, gelir, harcama, önceki gecikme, il…*

# Hedefi tanımlamak bir karardır

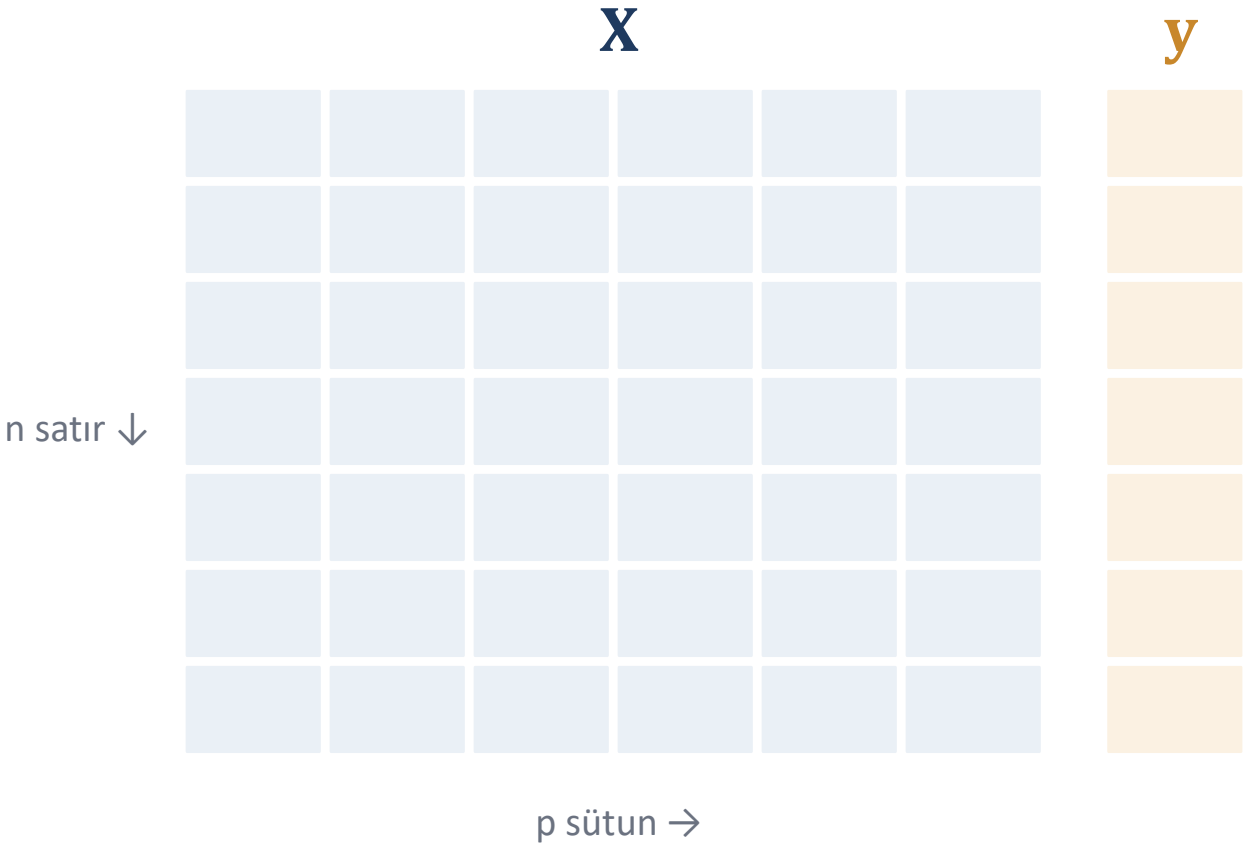
*Hedef değişirse model başka bir şey öğrenir*

| Hedef | Saklı karar |
|---|---|
| **Temerrüt** | Kaç günlük gecikme? Hangi zaman ufkunda? (uygulamada yaygın ölçüt: 90 gün) |
| **Resesyon** | Art arda iki çeyrek küçülme mi, komite kararı mı? |
| **Enflasyon** | Aylık mı yıllık mı? Manşet mi çekirdek mi? |
| **Konut fiyatı** | Seviye mi, değişim mi? Nominal mi reel mi? |

**Hatırlayın:** “Sağlık ihtiyacı” yerine “sağlık harcamasını” öngören algoritma mükemmel çalıştı, ama yanlış problemi çözdü.

# Veri bir matristir

*n gözlem (satır) × p öznitelik (sütun) ve bir hedef sütunu*




## Açılış tablosunda

n = 6 müşteri
p = 5 öznitelik: yaş, il, gelir, harcama, önceki gecikme
y = temerrüt

## Hangi sütun hedef?

Soruya bağlı. “Harcama ne olur?” diye sorsaydık harcama hedef, temerrüt bir öznitelik olurdu.

# Öznitelikl erin türleri

*Tür, hangi hazırlık işleminin gerektiğini belirler*

| Tür | Örnek | Nasıl işlenir? |
|---|---|---|
| **Sürekli** | Gelir, harcama, faiz oranı | Gerekirse ölçeklenir; çarpıksa log alınır |
| **Sayım** | Çocuk sayısı, şube sayısı | Sayı olarak kalır |
| **İkili** | Önceki gecikme (evet/hayır) | 0 / 1 |
| **Nominal kategorik** | İl, sektör, meslek | Kukla değişkenler |
| **Sıralı kategorik** | Eğitim düzeyi, kredi notu | Sıra kodu (1, 2, 3…) ya da kukla |
| **Tarih** | Başvuru tarihi | Ay, çeyrek, geçen süre gibi özniteliklere çevrilir |
| **Metin, görüntü** | Haber, uydu görüntüsü | Önce sayıya çevrilir |

MEHMET Sonjur

# 3

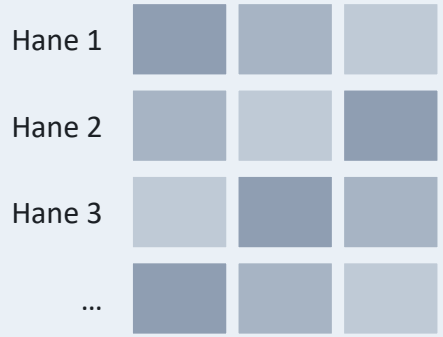
# Veri türleri

Bildiğimiz yapılar ve yeni veri kaynakları

# Üç temel veri yapısı

## Kesit

*Çok birim, tek zaman*

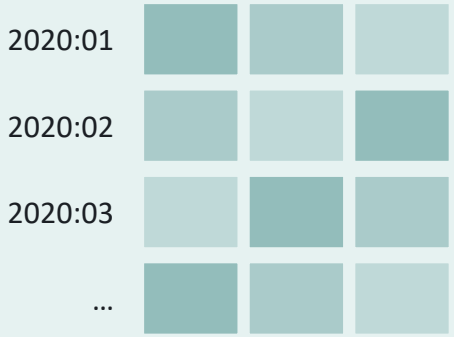



**Türkiye'den örnek**

Hanehalkı Bütçe Araştırması

## Zaman serisi

*Tek birim, çok zaman*




**Türkiye'den örnek**

EVDS'den aylık TÜFE, kur

## Panel

*Çok birim × çok zaman*




**Türkiye'den örnek**

İl bazında GSYH, 2004–

**Yeni olan:** Zaman serisi ve panelde satırların sırası bir bilgidir, karıştırılamaz. Model başarısını ölçerken bunun ne anlama geldiğini Hafta 3'te göreceğiz.

# Soru, veri yapısını belirler

*“Bir müşterinin gelecek ay kart harcaması ne olur?”*

## Elimizdeki: kesit

| Müşteri | Gelir | Harcama |
|---|---|---|
| 1 | 28.000 | 9.500 |
| 2 | 85.000 | 31.000 |
| 3 | 19.500 | 12.800 |

## Gereken: panel (müşteri-ay)

| Müşteri | Ay | Gelir | Harcama |
|---|---|---|---|
| 1 | Ocak | 28.000 | 9.500 |
| 1 | Şubat | 28.000 | 10.200 |
| 1 | Mart | 29.500 | 8.900 |
| 2 | Ocak | 85.000 | 31.000 |
| … | … | … | … |

Tek zaman noktası olan bir tabloyla “gelecek ay” sorusu cevaplanamaz. Her müşterinin birden fazla ayı gözlenmelidir.

# Tablonun ötesinde: yapılandırılmamış veri

*Makine öğrenmesinin iktisatçıya açtığı yeni veri kaynakları*



## Metin

Gazete haberleri, merkez bankası metinleri, KAP bildirimleri

*Hafta 1'den: Haberlerden Ekonomik Politika Belirsizliği endeksi*



## Görüntü

Uydu fotoğrafları, gece ışıkları

*Hafta 1'den: Gece ışıklarından ekonomik büyüme ölçümü*



## Konum

Şube, mağaza ve hanelerin koordinatları



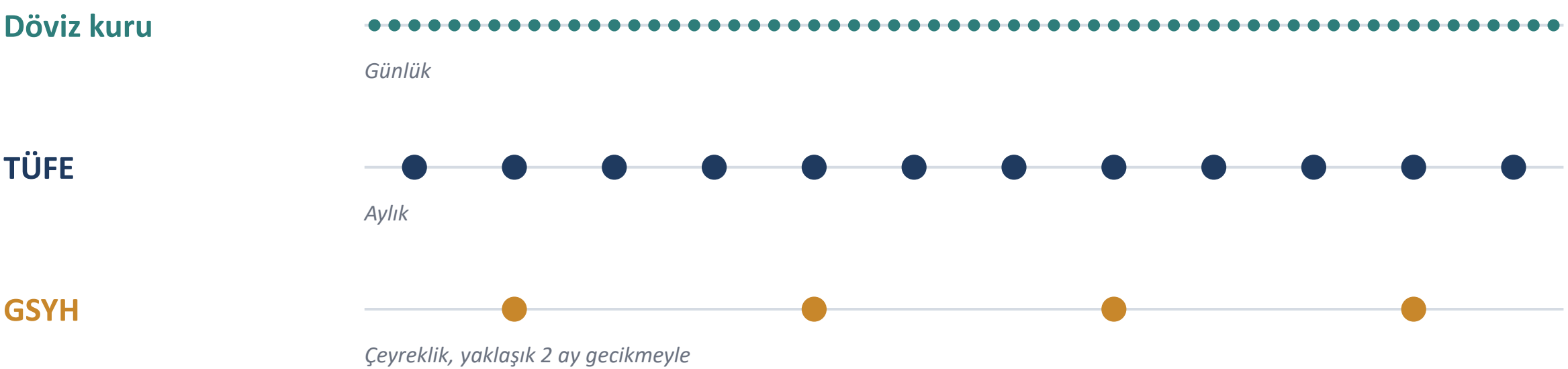
## Ses

Kazanç açıklaması toplantı kayıtları

**Hepsi, modele girmeden önce sayılardan oluşan bir tabloya dönüştürülür. Nasıl yapıldığını metin haftalarında göreceğiz.**

# Veriler farklı hızlarda gelir

*Frekans ve yayımlanma zamanı*




**Bir veri hazırlama kararı:** Günlük kuru aylık TÜFE ile aynı tabloya koymak için ay ortalaması mı, ay sonu değeri mi alınacak?

# Panel veriyi saklamanın iki yolu

MEHMET Sonjur

## Geniş format

*Her il bir satır, her yıl bir sütun*

| İl | GSYH_2021 | GSYH_2022 |
|---|---|---|
| Adana | a | b |
| Diyarbakır | c | d |

## Uzun format

*Her il-yıl bir satır, yıl ayrı bir sütun*

| İl | Yıl | GSYH |
|---|---|---|
| Adana | 2021 | a |
| Adana | 2022 | b |
| Diyarbakır | 2021 | c |
| Diyarbakır | 2022 | d |

TÜİK ve OECD veriyi çoğu zaman geniş formatta verir; makine öğrenmesi araçları genellikle uzun format bekler. Dönüşümü 9. haftada yapacağız.

MEHMET Sonqur

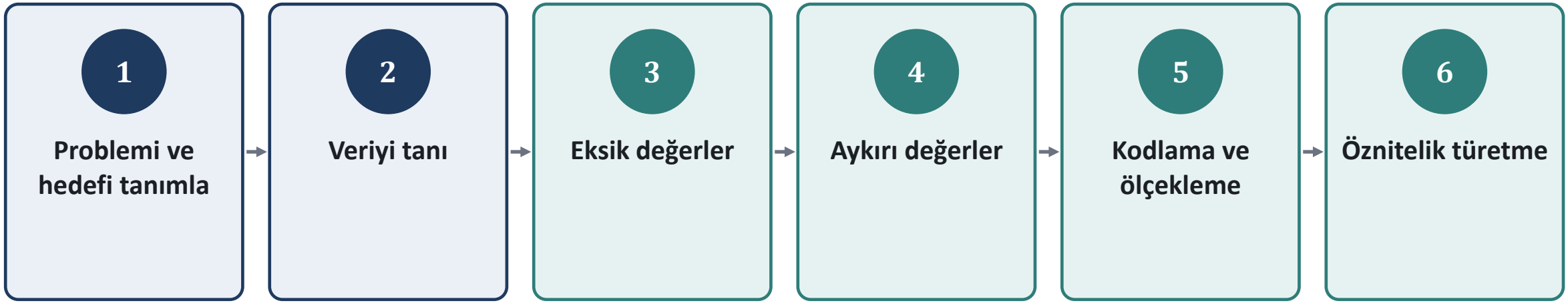
# 4

# Veri hazırlama

Model kurmadan önceki en uzun adım

# Veri hazırlamanın temel adımları

MEHMET Sonmez




**Adım 1–2:** Veriyi anlamak. Ne ölçülmüş, ne eksik, ne tuhaf?

**Adım 3–6:** Veriyi modelin kullanabileceği hâle getirmek.

# Tabloya geri dönelim

*Bu veriyle hemen model kurabilir miyiz?*

| Müşteri | Yaş | İl | Aylık gelir (TL) | Kart harcaması (TL) | Önceki gecikme | Temerrüt |
|---|---|---|---|---|---|---|
| 1 | 34 | Diyarbakır | 28.000 | 9.500 | Hayır | Hayır |
| 2 | 51 | İstanbul | 85.000 | 31.000 | Hayır | Hayır |
| 3 | 23 | Mersin | 19.500 | 12.800 | Evet | Evet |
| 4 | 45 | Ankara | — | 7.200 | Hayır | Hayır |
| 5 | 29 | Diyarbakır | 32.000 | **268.000** | Evet | Hayır |
| 6 | 62 | İzmir | 41.000 | 5.100 | Hayır | Hayır |

**Eksik değer:** 4. müşterinin geliri yok.

**Şüpheli değer:** 268.000 TL harcama, gelirin 8 katı. Hata mı?

**Metin ve ölçek:** "İl" bir metin; gelir binlerle, yaş onlarla ölçülüyor.

# Veri neden eksik?

*Ne yapacağımız, eksikliğin nedenine bağlı (Rubin, 1976)*



## Tamamen rastlantısal

Eksiklik hiçbir şeye bağlı değil.

*Örnek: Sistem hatası rastgele bazı formları kaydetmemiş.*



## Gözlenen bir özelliğe bağlı

Eksiklik, elimizdeki başka bir değişkenle ilişkili.

*Örnek: Gençler gelir sorusunu daha sık boş bırakıyor; yaşı biliyoruz.*



## Eksik değerin kendisine bağlı

Eksiklik, eksik olan değerle ilişkili.

*Örnek: Yüksek gelirliler gelirini söylemek istemiyor.*

# Eksik veriyle ne yapabiliriz?

MEHMET Sonjur

| Strateji | Avantaj | Risk |
|---|---|---|
| **Satırı silmek** | Basit | Örneklem küçülür; eksiklik rastgele değilse yanlılık |
| **Ortalama / medyanla doldurmak** | Basit, satır kaybı yok | Varyansı yapay olarak küçültür |
| **Grup medyanıyla doldurmak** | Daha isabetli | Grup tanımı bir varsayımdır |
| **Diğer değişkenlerden tahmin ederek** | Değişkenler arası ilişkiyi kullanır | Karmaşık; tahmin de hata içerir |
| **Eksiklik göstergesi eklemek** | “Eksik olması” bilgi taşıyabilir | Bir doldurmayla birlikte kullanılır |

*Gelirini beyan etmeyen başvuranın bu tercihi bile temerrüt riski hakkında bilgi taşıyabilir.*

# Aykırı değer: hata mı, gerçek mi?

MEHMET Sonqur

**268.000 TL**

5. müşterinin aylık kart harcaması

**Büyük olasılıkla veri girişi hatası (26.800 TL?). Düzeltilir ya da çıkarılır.**

**%85+**

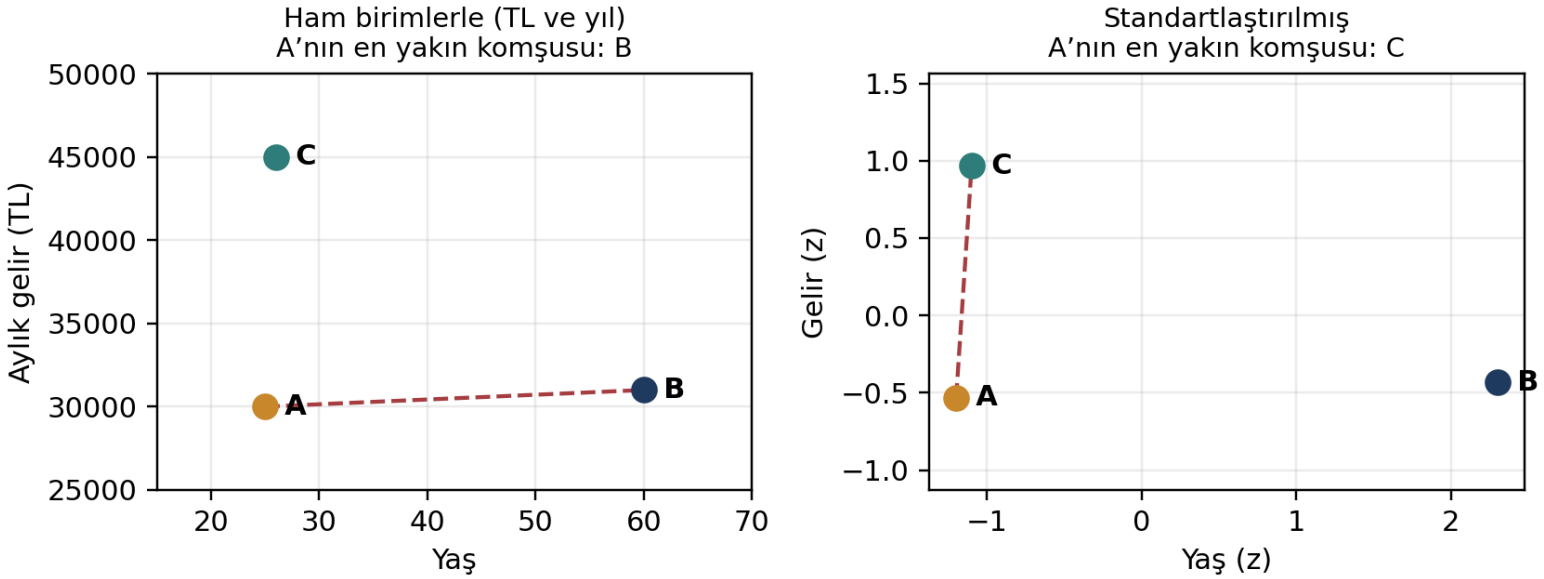
Türkiye'de yıllık TÜFE, Ekim 2022

**Gerçek ve son derece bilgilendirici. Silinirse model krizlere karşı kör kalır.**

Seçenekler: düzelt · sil · kırp (%1 ve %99) · log al · dokunma

# Ölçekleme: A kime benziyor?

*A (25 yaş, 30.000 TL) · B (60 yaş, 31.000 TL) · C (26 yaş, 45.000 TL)*







$$z = (x - \text{ortalama}) / \text{standart sapma}$$

**Ham birimlerle**

A–B ≈ 1.000
A–C ≈ 15.000
Gelir her şeyi belirliyor.

**Standartlaştırınca**

A–B ≈ 3,5
A–C ≈ 1,5
A'nın komşusu C.

# Ekonometride neden ölçeklemiyorduk?

MEHMET Sonjur

## Ekonometride (EKK)

Geliri TL yerine bin TL ile ölçün:

- **Katsayı 1.000 kat değişir.**
- **Öngörüler değişmez.**

*Birim seçimi sonucu etkilemez; ölçeklemeye gerek yok.*

## Makine öğrenmesinde

Gözlemler arasındaki benzerliği hesaplayan yöntemlerde birim sonucu değiştirir.

**A–B–C örneğinde gördüğümüz gibi: birimi büyük olan değişken kararı ele geçirir.**

Hangi yöntemlerde ölçekleme gerektiğini, o yöntemleri öğrendikçe tek tek göreceğiz.

# Kategorik değişkenleri sayıya çevirmek

*Ekonometrideki kukla değişkenler (ML literatüründe: one-hot kodlama)*

| Müşteri | İl |
|---|---|
| 1 | Diyarbakır |
| 2 | İstanbul |
| 3 | Mersin |
| 5 | Diyarbakır |



| Müşteri | İl_Diyarbakır | İl_İstanbul | İl_Mersin |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 5 | 1 | 0 | 0 |

### Kukla tuzağı

Ekonometrideki kural aynen geçerli: doğrusal regresyonda bir kategori dışarıda bırakılır.

### Sıralı değişkenler

Eğitim: ilkokul 1, ortaokul 2, lise 3… Sıra korunur ama aralıkların eşit olduğu varsayılır.

### Çok kategori

900'ü aşkın ilçe için 900 sütun mu? Bölgelere ya da gelişmişlik gruplarına toplanabilir.

# Öznitelik türetme: iktisatçının avantajı

*Ham değişkenlerden daha bilgili değişkenler türetmek*

### Oran

**Borç / gelir**

*Riski seviyeler değil oranlar belirler*

### Logaritma

**log(gelir)**

*Çarpık dağılımı dengeler*

### Gecikme

**Geçen ayın enflasyonu**

*Ekonomik değişkenler ataletlidir*

### Takvim

**Ay, bayram göstergesi**

*Mevsimsellik ve takvim etkileri*

### Reel değer

**Gelir / TÜFE**

*Yüksek enflasyonda nominal TL yanıltır*

2020'de 10.000 TL ile 2025'te 10.000 TL aynı şey değildir. Nominal veriyle eğitilen model, "herkesin zenginleştiğini" öğrenir.