Woman in Data Science

Lecture 1

18.11.2025

1. VERİ, ENFORMASYON VE AMAÇ

1.1. Data – Information – Knowledge (Veri – Enformasyon – Bilgi)

🔹 Veri (Data) nedir?

- Ham, işlenmemiş gerçeklerdir.
- Tek başına anlamlı olmak zorunda değil.
- Ölçümler, sayılar, metinler, log kayıtları, sensör verileri olabilir.

Örnekler:

- Sınıftaki öğrencilerin yaşları → 19, 20, 21, 23...
- Bir mağazanın günlük satış adedi → 78 sipariş
- Bir arabanın hız sensörü → 95 km/s

Bu veriler tek başına bir anlam ifade etmez. Sadece ham bilgi.

1. VERİ, ENFORMASYON VE AMAÇ

1.1. Data – Information – Knowledge (Veri – Enformasyon – Bilgi)

🔹 Enformasyon (Information) nedir?

Verinin işlenmiş, yorumlanabilir hale gelmiş hâlidir.

Örnek:

- **"Sınıftaki öğrencilerin yaş ortalaması 21."**
- **"Mağaza satışları geçen aya göre %15 arttı."**

Bu artık yorumlanmış veridir → enformasyon.

1. VERİ, ENFORMASYON VE AMAÇ

1.1. Data – Information – Knowledge (Veri – Enformasyon – Bilgi)

🔹 Bilgi (Knowledge) nedir?

Enformasyondan sonuç çıkarma, karar verme aşamasıdır.

Örnek:

- "Öğrencilerin yaş ortalamasına göre ders programını akşam yaparsak daha verimli olur."
- "Satışların %15 artması nedeniyle stok artırılmalıdır."

Yani bilgi → karar alma içindir.

1. VERİ, ENFORMASYON VE AMAÇ

1.1. Data – Information – Knowledge (Veri – Enformasyon – Bilgi)

📌 **Kısaca formül:**

Veri → Enformasyon → Bilgi → Karar → Aksiyon

1. VERİ, ENFORMASYON VE AMAÇ

1.2. Veri Neden Önemlidir?

🔹 1) Doğru karar almayı sağlar

İşletmeler, şirketler, devletler veriye bakarak karar alır.

Örneğin:

- Netflix hangi içeriklere yatırım yapacağını izleme verilerine bakarak belirler.
- Hastaneler yoğunluk planlamasını hasta verilerine göre yapar.

🔹 2) Tahmin yapabilmemizi sağlar

Veri olmadan tahmin modeli kurulamaz.

Örneğin:

Hava durumu tahmini

Satış tahmini

Hastalık risk tahmini

Makine öğrenmesi algoritmalarının tamamı veri ile beslenir.

🔹 3) Otomasyon ve verimlilik sağlar

- Üretim hatalarında hataların otomatik bulunması
- E-ticaret ürün önerileri
- ChatGPT'nin bile çalışması → dev bir veri setine dayanır

Veri → Otomasyon → Para, zaman ve iş gücünden tasarruf.

1. VERİ, ENFORMASYON VE AMAÇ

1.3. Günlük Hayattan Basit Data Örnekleri

🍔 McDonald's "Sipariş yoğunluğu" verisi

- Saat 12:00-14:00 en yoğun saat
- Bu veri zincirin eleman sayısını artırma kararını etkiler
- → Data → Information → Action

🚗 Araç hız göstergesi

- Sensörler ham veri sağlar
- Araç bu veriyi işler → hız ekranı
- Sürücü karar verir → yavaşla / hızlan

📱 Telefon adım sayısı

- Adım sayıları ham veridir
- Uygulama bunu grafiğe dökünce enformasyon olur
- "Son 7 gün adımlarım düşmüş" → bilgi
- "Yürüyüşe çıkayım" → aksiyon

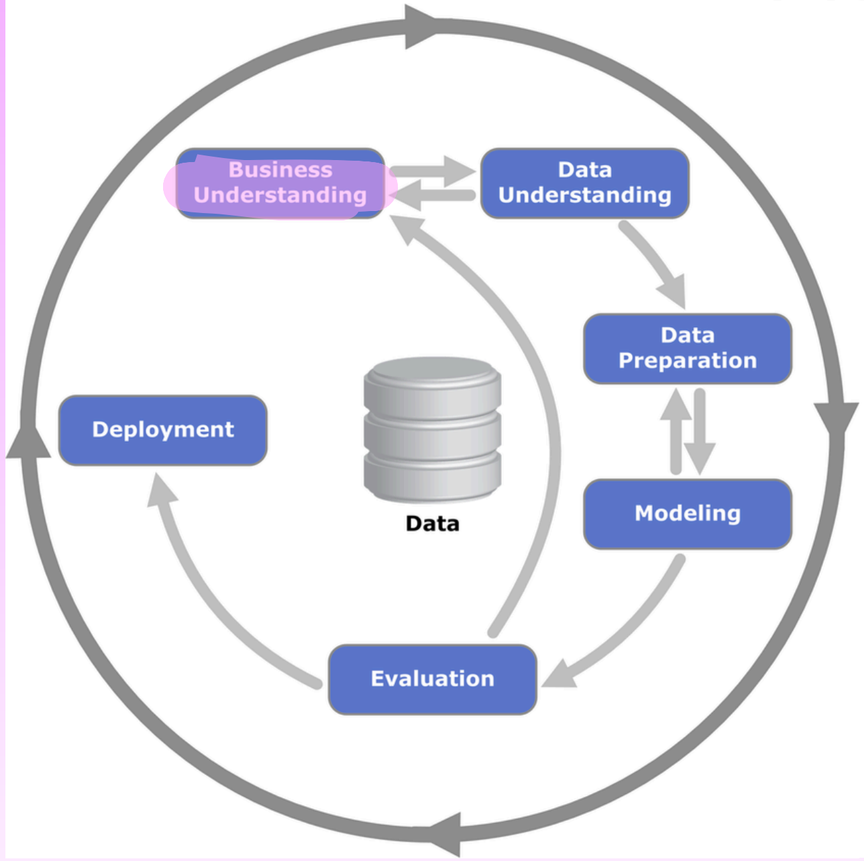
🛒 E-ticaret sipariş geçmişi

- Ham veri: "Ali 14 Mart'ta 2 ürün satın aldı."
- Enformasyon: "Ali, 3 ayda 5 kere elektronik ürün aldı."
- Bilgi: "Ali'ye elektronik ürün öner."

2. DATA SCIENCE SÜRECİ (CRISP-DM + EDA)

2.1. CRISP-DM Nedir?

CRISP-DM (Cross Industry Standard Process for Data Mining), veri bilimi projelerinde kullanılan en yaygın süreç modelidir. 6 aşamadan oluşur:



Aşama 1: Business Understanding (İş Problemini Anlamak)

Bu aşamada teknik veri yoktur; sadece problem anlaşılır.

Soru şudur:

"Ne çözmek istiyoruz?"

Örnekler:

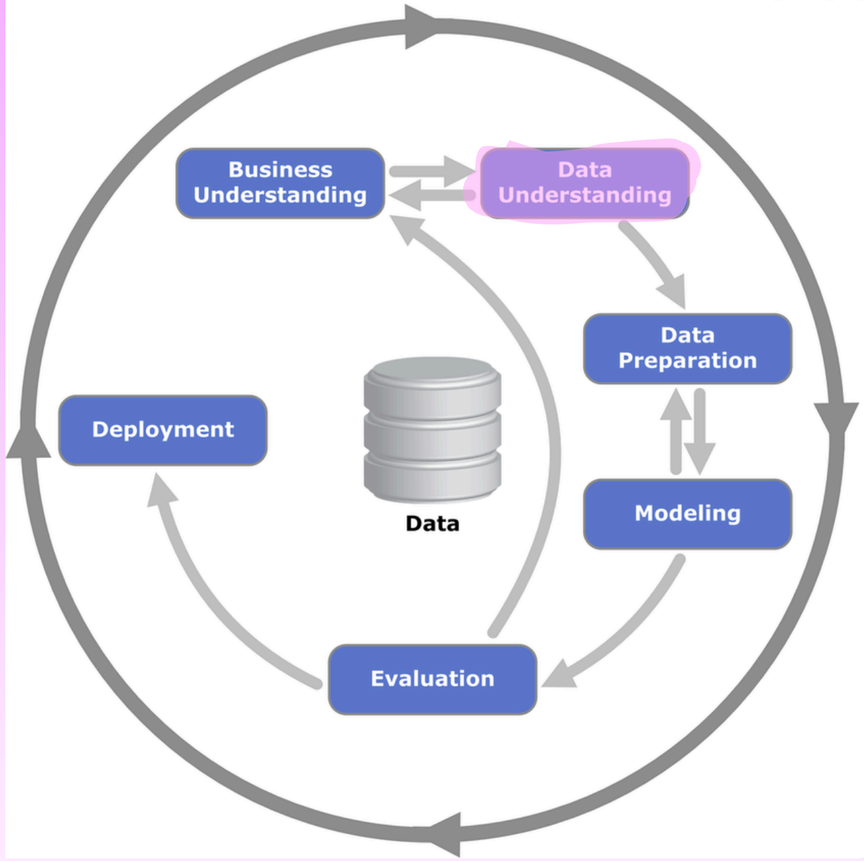
- Müşteri terk etmesini (churn) tahmin etmek
- Hastanın risk seviyesini tahmin etmek
- Satışları tahmin etmek

📌 Teknik çözümün başlaması bu aşamadan sonra olur.

2. DATA SCIENCE SÜRECİ (CRISP-DM + EDA)

2.1. CRISP-DM Nedir?

CRISP-DM (Cross Industry Standard Process for Data Mining), veri bilimi projelerinde kullanılan en yaygın süreç modelidir. 6 aşamadan oluşur:



Aşama 2: Data Understanding (Veriyi Anlamak)

Bu aşamada:

- Veri toplanır
- İlk defa veriye bakılır
- Dağılımlar, eksikler, tipler kontrol edilir

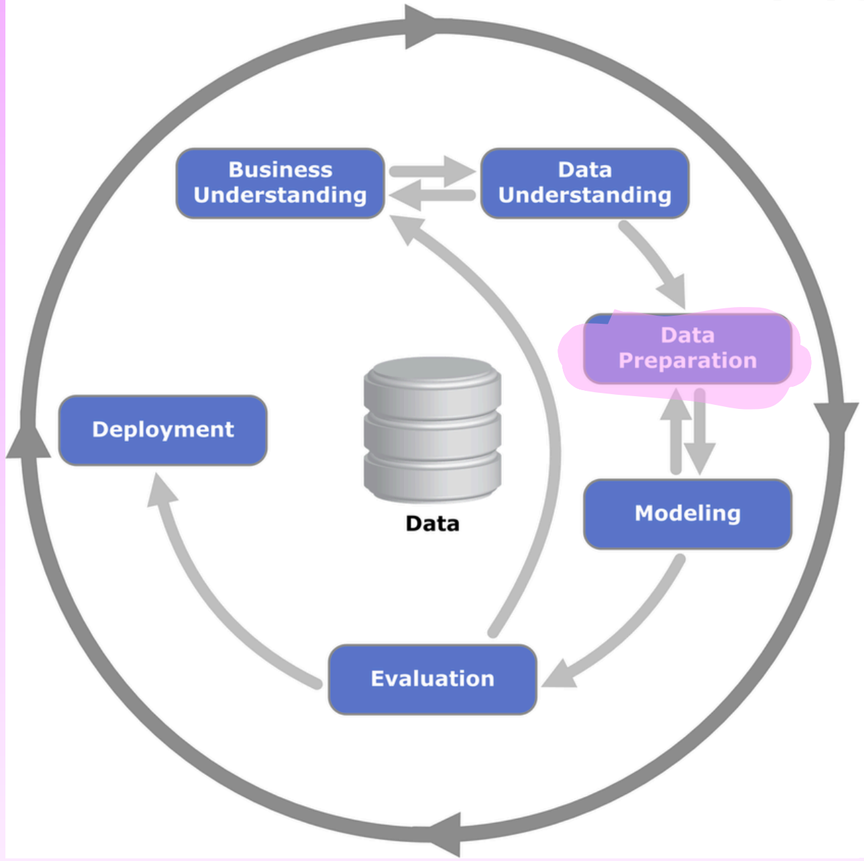
Bu aşama → EDA (Exploratory Data Analysis) ile yakından ilişkilidir.

"Veri ne anlatıyor? Sağlıklı mı?"

2. DATA SCIENCE SÜRECİ (CRISP-DM + EDA)

2.1. CRISP-DM Nedir?

CRISP-DM (Cross Industry Standard Process for Data Mining), veri bilimi projelerinde kullanılan en yaygın süreç modelidir. 6 aşamadan oluşur:



Aşama 3: Data Preparation (Veri Hazırlama)

En uzun süren aşamadır. Zamanın yaklaşık %60-70'i burada harcanır.

Yapılanlar:

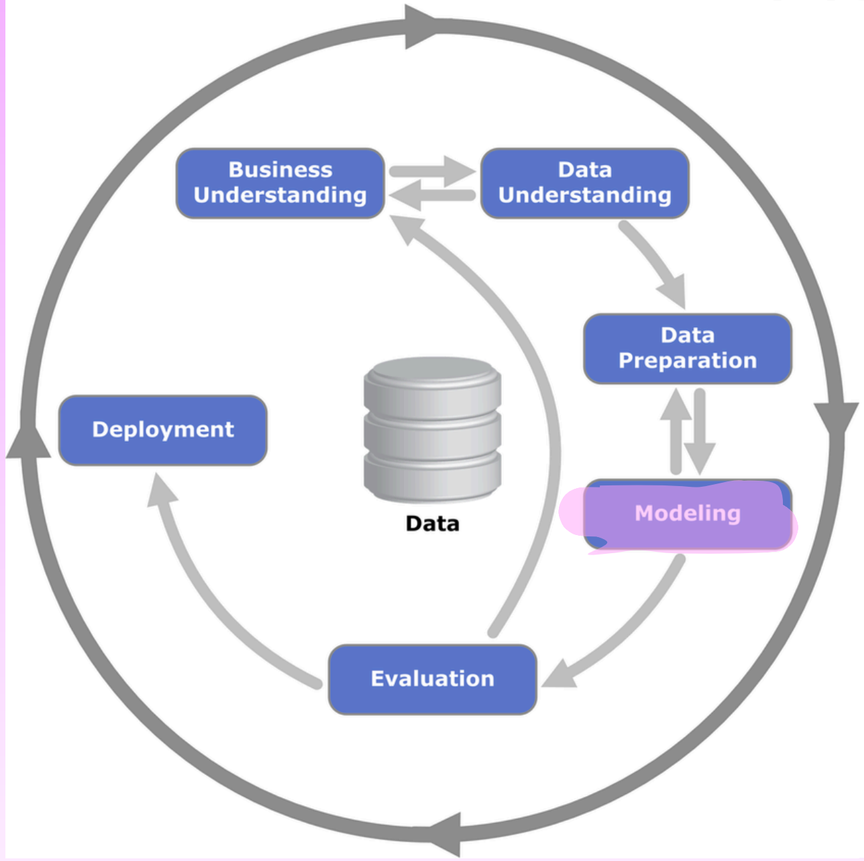
- Eksik değer doldurma
- Aykırı değer tespiti
- Normalizasyon, encoding
- Gereksiz sütun temizleme
- Yeni özellik üretimi (feature engineering)

📌 Modelin başarısını en çok etkileyen aşamadır.

2. DATA SCIENCE SÜRECİ (CRISP-DM + EDA)

2.1. CRISP-DM Nedir?

CRISP-DM (Cross Industry Standard Process for Data Mining), veri bilimi projelerinde kullanılan en yaygın süreç modelidir. 6 aşamadan oluşur:



Aşama 4: Modeling (Model Kurma)

Artık makine öğrenmesi modelleri seçilir.

Örnekler:

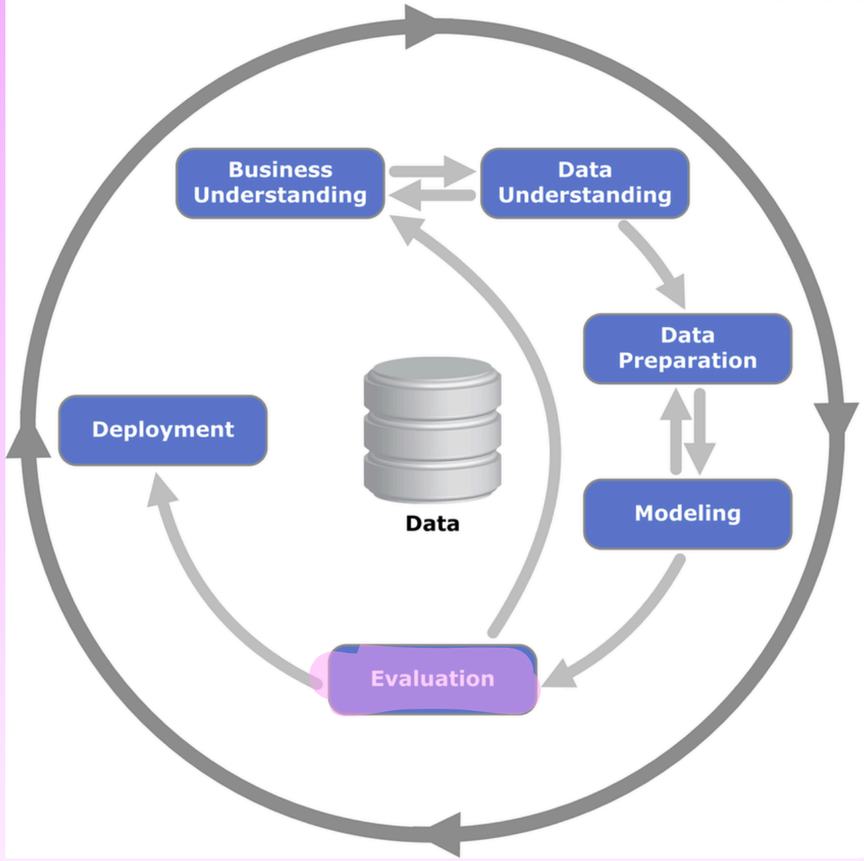
- Decision Tree
- Logistic Regression
- Random Forest
- Neural Networks
- KNN

Model eğitilir, optimize edilir.

2. DATA SCIENCE SÜRECİ (CRISP-DM + EDA)

2.1. CRISP-DM Nedir?

CRISP-DM (Cross Industry Standard Process for Data Mining), veri bilimi projelerinde kullanılan en yaygın süreç modelidir. 6 aşamadan oluşur:



Aşama 5: Evaluation (Değerlendirme)

Modelin başarı metrikleri incelenir:

- Accuracy
- F1-Score
- RMSE
- Confusion Matrix

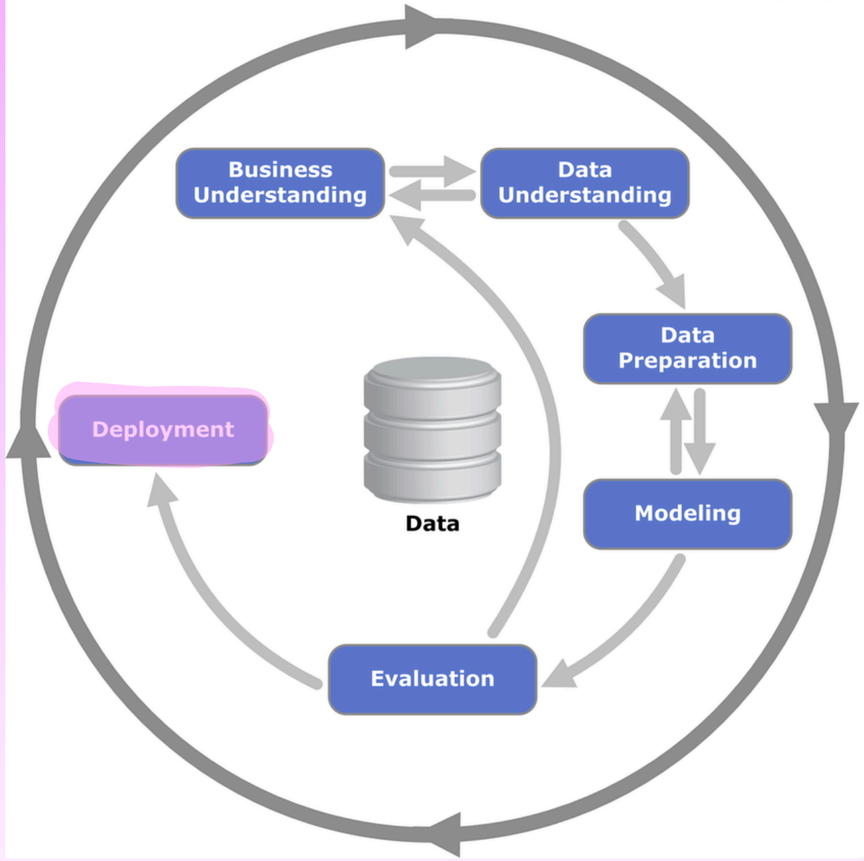
Soru:

"Model iş problemini çözüyor mu?"

2. DATA SCIENCE SÜRECİ (CRISP-DM + EDA)

2.1. CRISP-DM Nedir?

CRISP-DM (Cross Industry Standard Process for Data Mining), veri bilimi projelerinde kullanılan en yaygın süreç modelidir. 6 aşamadan oluşur:



Aşama 6: Deployment (Dağıtım ve Kullanıma Alma)

Model gerçek bir sisteme entegre edilir:

- API olarak kullanılır
- Web/Mobil uygulamaya eklenir
- Dashboard'a bağlanır
- Otomatik takip sistemi kurulur

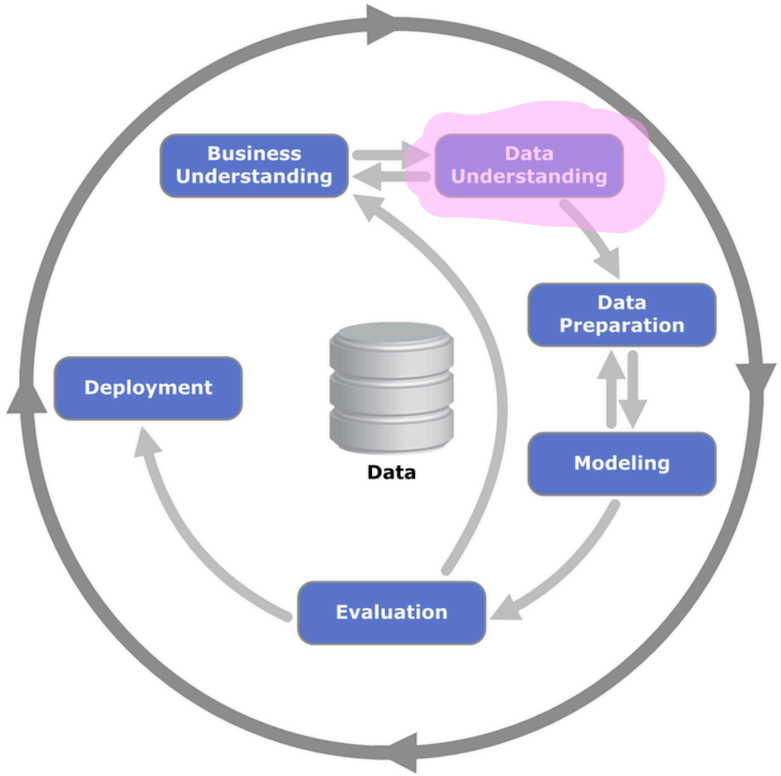
Ve proje kapanmaz — model düzenli güncellenir.

2. DATA SCIENCE SÜRECİ (CRISP-DM + EDA)

2.2. CRISP-DM'in İlk Derste Kullanım Mantığı

Bugün veri biliminin tamamını yapmayacağız.

Ama bugün CRISP-DM sürecinin "Data Understanding + EDA" kısmını işleyeceğiz.



Yani:

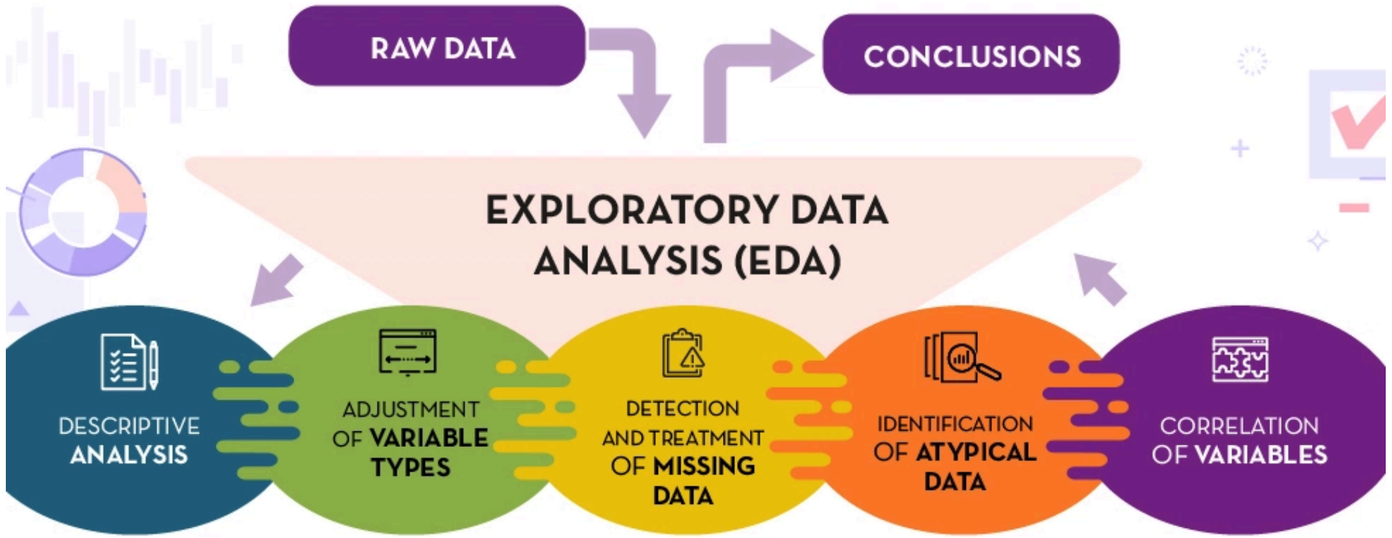
- **Veri bulma**
- **Veriyi ilk kez inceleme**
- **Grafiklerle keşfetme**
- **Korelasyonlara bakma**

Bu ders → sürecin başlangıcıdır.

2. DATA SCIENCE SÜRECİ (CRISP-DM + EDA)

2.3. EDA (Exploratory Data Analysis) Nedir?

EDA = Veriyi anlamak + gizli problemleri görmek.



Amaçlar:

- **Dağılımları görmek**
- **Outlier tespit etmek**
- **Eksik değerleri bulmak**
- **Korelasyonlara bakmak**
- **Veriyi "tanımak"**

EDA şudur:

"Model kurmadan önce adeta doktorun hastayı muayene etmesi."

3. VERİ SETİ BULMA VE İNCELEME

3.1. Veri Seti Kaynakları (Kaggle, UCI, Google Dataset Search, Open Data)

🔹 1. Kaggle Datasets

Dünyanın en büyük veri bilimi platformu.

Özellikleri:

- Hesap açmak kolay
- Etiketlenmiş veri setleri
- Popüler konular için çok veri var
- Kullanıcı yorumları ve açıklamalar var
- Veri seti çeşitliliği yüksek

🔹 2. UCI Machine Learning Repository

Data science'ın en eski ve en klasik veri seti deposu.

Özellikleri:

- Akademik veri setleri
- Temel ML örneklerinde sürekli kullanılan klasik veri setleri (Iris, Wine, Breast Cancer vb.)

Avantaj: Kolay indirilebilir, küçük boyutlu

Dezavantaj: Arayüz eski

3. VERİ SETİ BULMA VE İNCELEME

3.2. Kaggle Üzerinden Veri Bulma

Adım 1 — Kaggle'a giriş yap

https://www.kaggle.com/

Yeni hesap → Google ile giriş en kolayıdır.

Adım 2 — "Datasets" bölümüne gir

Üst menü > Datasets

Adım 3 — Arama Kutusunu Kullan

Örnek aramalar:

- "house prices"
- "heart disease"
- "student performance"
- "restaurants"
- "movies"

Adım 4 — Bir veri seti sayfasını incele

Öğrenciler için çok önemli olan bölümleri anlat:

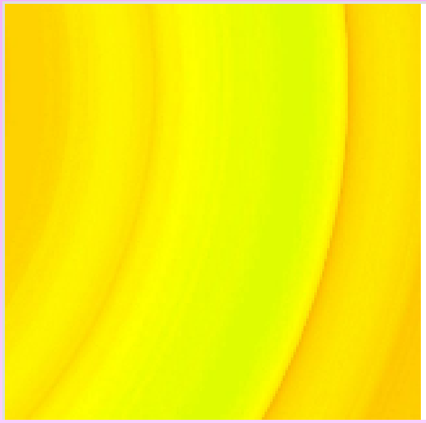
Kaggle veri seti sayfasında şunlar var:

- Description: Veri setinin açıklaması
- Context: Bu veri nereden geliyor?
- Content: Sütunların anlamı
- Columns:
  - column name
  - data type
  - explanation
- Download: Veri setini indir
- Notebooks: Diğer insanların yaptığı analiz örnekleri

Adım 5 — Veri setini indir

"Download" → zip dosyası → içinden CSV çıkıyor.

CSV → veri bilimindeki en standart dosya türü.



Forest Fires Data Set

predict the burned area of forest fires using meteorological and other data

kaggle.com

3. VERİ SETİ BULMA VE İNCELEME

3.3. Veri Setini İndirdikten Sonra İlk Bakılacak Şeyler

🔹 1. Dosya yapısı

CSV mi? XLSX mi? JSON mu? TXT mi?

🔹 2. Kaç satır (row) var?

→ Problem küçük mü, büyük mü?

🔹 3. Kaç sütun (column) var?

→ Veri karmaşık mı?

🔹 4. Sütunlar neyi temsil ediyor?

Veri seti açıklamasından öğrenilebilir.

🔹 5. Hedef değişken (target) hangisi?

Makine öğrenmesi yapacaksan tahmin edilecek değeri belirlemelisin.

3. VERİ SETİ BULMA VE İNCELEME

3.3. Veri Setini İndirdikten Sonra İlk Bakılacak Şeyler

- Kaynak: UCI Machine Learning Repository
- Amaç: Orman yangınlarında yanan alanı (area) tahmin etmek
- Tip: Regresyon problemi

Değişkenler

- X, Y: Coğrafi koordinatlar
- month, day: Zaman bilgisi
- FFMC, DMC, DC, ISI: Yangın tehlike indeksleri
- temp, RH, wind, rain: Hava durumu faktörleri
- area: Hedef değişken (yanan alan)

EDA'da bulduklarımız

- Area değişkeni dengesiz, çoğu değer 0
- Hava sıcaklığı, rüzgar ve indeksler yangın büyüklüğüyle ilişkili
- Korelasyonlar düşük ama anlamlı örüntüler var

3. VERİ SETİ BULMA VE İNCELEME

3.3. Veri Setini İndirdikten Sonra İlk Bakılacak Şeyler

Modelleme

- **Basit Lineer Regresyon uygulanabilir**
- **Biraz düşük R^2 → gerçek hayatta yangın büyüklüğünü tahmin etmek zor**
- **Farklı modeller (Decision Tree, Random Forest) daha iyi sonuç verebilir**

3. VERİ SETİ BULMA VE İNCELEME

3.3. Veri Setini İndirdikten Sonra İlk Bakılacak Şeyler

Mini Assignment — Forest Fire Analysis

Görev:

Bu veri setiyle aşağıdaki soruları cevaplayın:

1. **Hangi ayda daha fazla yangın görülüyor?**
2. **Area değişkeninin dağılımını yorumlayın.**
3. **Hangi meteorolojik değişken area ile daha çok ilişkili?**
4. **Linear Regression modeli kurun ve R^2 değerini raporlayın.**
5. **Model performansını artırmak için 1 öneride bulunun.**

4. VERİ SETİ MANTIĞI

(Forest Fires Dataset'e Göre)

4.1. Satır (Observation) nedir?

Bu veri setinde her satır bir orman yangını olayını temsil eder.

- Toplam 517 satır vardır.
- Her satır bir yangına ait ölçülen değerleri içerir.
- Örnek:
  - 10 Ağustos günü çıkan bir yangın, sıcaklık 22°C, rüzgar 4.3 km/h, yanan alan 12.4 hektar gibi.

4. VERİ SETİ MANTIĞI

(Forest Fires Dataset'e Göre)

4.2. Sütun (Feature / Attribute) nedir?

Her sütun yangına ait bir özelliği (feature) gösterir.

Forest Fires veri setinde 13 sütun vardır.

Bunlar; coğrafi konum, tarih bilgisi, meteorolojik indeksler ve hava durumu ölçümleridir.

Örnek sütunlar:

- temp → sıcaklık
- RH → nem
- wind → rüzgar
- rain → yağış
- FFMC, DMC, DC, ISI → yangın tehlikesi indeksleri
- month, day → ay ve gün bilgisi
- area → yanan alan

4. VERİ SETİ MANTIĞI

(Forest Fires Dataset'e Göre)

4.3. Hedef değişken (Label / Target) nedir?

Bu veri setinde hedef değişken area sütunudur.

- area → yanan alan miktarı (hektar)
- Sürekli (numerik) bir değişkendir.
- Bu veri seti bu yüzden regresyon problemi olarak kullanılır.

Bu veri setiyle 'bir yangında kaç hektar alan yanar?' sorusunu tahmin etmeye çalışıyoruz. Tahmin etmek istediğimiz sütuna hedef (label) denir. Bu veri setinde hedef değişken area'dır

4. VERİ SETİ MANTIĞI

(Forest Fires Dataset'e Göre)

4.4. Veri Tipleri (Categorical – Numerical – Text)

Forest Fires veri setinde veri tipleri şunlardır:

Kategorik (Category / Object)

- month (jan–dec)
- day (mon–sun)

Bu alanlar kategoriktir, matematiksel işlem yapılmaz → encode edilir (label encoding, dummy variable).

4. VERİ SETİ MANTIĞI

(Forest Fires Dataset'e Göre)

Sayısal (Numeric)

Sürekli / ölçülebilir değişkenler:

- X, Y
- FFMC, DMC, DC, ISI
- temp
- RH
- wind
- rain
- area

Bu değişkenler üzerinde:

- ortalama, medyan, minimum, maksimum
- korelasyon
- histogram, scatter plot
- gibi analizler yapılabilir.

4. VERİ SETİ MANTIĞI

(Forest Fires Dataset'e Göre)

Sayısal (Numeric)

Sürekli / ölçülebilir değişkenler:

- X, Y
- FFMC, DMC, DC, ISI
- temp
- RH
- wind
- rain
- area

Bu değişkenler üzerinde:

- ortalama, medyan, minimum, maksimum
- korelasyon
- histogram, scatter plot
- gibi analizler yapılabilir.

4. VERİ SETİ MANTIĞI
(Forest Fires Dataset'e Göre)

4.5. Veri Sözlüğü (Data Dictionary)

Bu veri seti 517 yangından oluşuyor; her satır bir yangını temsil ediyor. Sütunlar yangına ait sıcaklık, rüzgar, nem, indeks gibi bilgiler. Biz bu bilgileri kullanarak bir yangında ne kadar alan yanacağını (area) tahmin etmeye çalışıyoruz.

Sütun	Açıklama	Tür
X	Yangının X koordinatı (1–9)	Sayısal
Y	Yangının Y koordinatı (2–9)	Sayısal
month	Ay (jan–dec)	Kategorik
day	Gün (mon–sun)	Kategorik
FFMC	İnce yakıt nem kodu	Sayısal
DMC	Orta derinlik nem kodu	Sayısal
DC	Kuraklık kodu	Sayısal
ISI	Yangının yayılma potansiyeli	Sayısal
temp	Sıcaklık (°C)	Sayısal
RH	Bağıl nem (%)	Sayısal
wind	Rüzgar hızı (km/h)	Sayısal
rain	Yağış (mm/m^2)	Sayısal
area	Yanan alan (hektar) **(hedef değişken)**	Sayısal

5) Python ile Veri Yükleme ve İlk İnceleme

6) EDA – Exploratory Data Analysis (Keşifsel Veri Analizi)

Bu bölüm tamamen grafiklerle veri anlamayı öğretmek için.

6.1. Tek Değişkenli Analiz (Univariate)

6.2. Kategorik Veri Analizi

6.3. Çift Değişkenli Analiz (Bivariate)

6.4. Korelasyon Analizi

6.5. Sorun Tespiti

Thanks For Listening