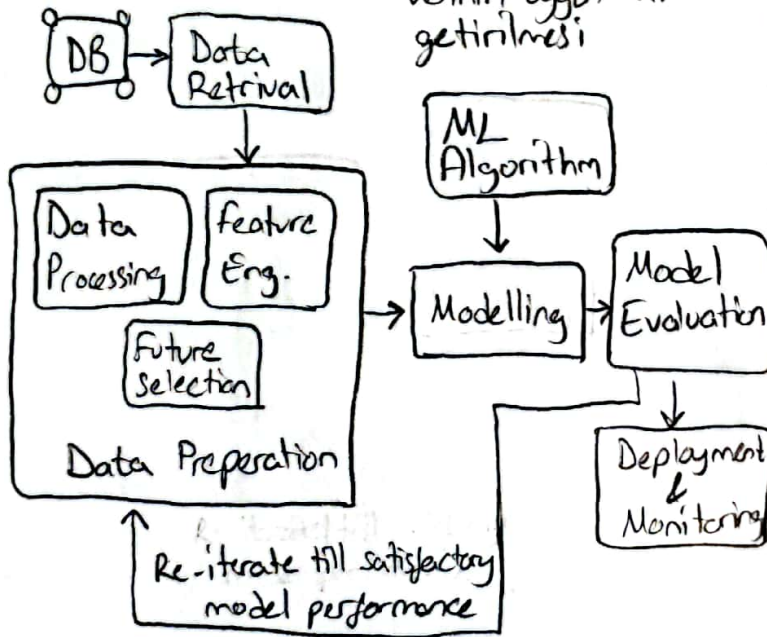


1- Outliers

↳ Feature Eng. & Data Processing

Özellik müh. ⇒ Hom veriden değişken üretmek

Veri ön işleme ⇒ Çalışmaya öncesi verinin uygun hale getirilmesi



↳ Aykırı Değerler (Outliers)

↳ Outlier ⇒ Verideki genel eğilimin oldukça dışına çıkan değer

↳ Outlier yakalama:

- 1) Sektör bilgisi
- 2) Standart Sapma yaklaşımı
- 3) Z-Skoru yaklaşımı
- 4) Boxplot yöntemi (tek değişken)
 $Q3 - Q1 = IQR$
 $Lower = Q1 - 1.5 IQR$
 $Upper = Q3 + 1.5 IQR$
- 5) LOF yöntemi (Çok değişken)

↳ Çok Değişkenli Aykırı Değer Analizi

↳ Tek başına anlamlı olmayan yapılar bir araya gelince anlamlı olabilir

↳ LOF: Local Outlier Factor

↳ Local yoğunluk ⇒ Komşuların yoğunluğu daha düşükse, daha seyrek bir bölgedir

↳ Inlier: Outlier olmayan

↳ LOF yöntemi local yoğunluğa göre uzaklık hesaplaması yapar. Bir skor atar ve bu 1'e ne kadar yakınsa o kadar inlier'dir. Ama sen de istersen bir threshold belirleyebilirsin

↳ PCA yöntemi: Temel Birklesim Analizi

↳ Thresholdu dörsek yöntemle belirlirsin

⚠ Ağac yöntemlerinde outlierlara dökünme, veünden trasla

2- Missing Values

↳ Eksik değerleri çözmek için

- 1) Silme
- 2) Değer atama yöntemleri
- 3) Tahmine dayalı yöntemler

↳ Eksik Değer Problemini Çözme

- 1) Hızlıca silmek → `dropna()`
- 2) Basit atama yöntemleriyle doldurmak → `fillna()`

↳ Tahmine Dayalı Atama İşlemi

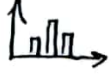
↳ Tahmin için ML kullanılır


Label encoder yapılır → Kat. değ. binary temsili edilmesi
Değişkenler standartlaştırılır

KNN uygulanır

↳ en yakın x komşusunun ortalaması alınır

↳ Eksik Verinin Yapısının İncelenmesi

↳ `msno.bar(df)` →  Değişkenlerdeki tüm sayıları gösterir

`msno.matrix(df)` → 
↓
Eksik verileri beyazla gösterir, değişkenler arasında bağlantı varsa gösterir

`msno.heatmap(df)` → 1s1 haritası

↳ Eksik Değerlerin Bağımlı Değişken ile Analizi

↳ `missing -vs- target()`

3- Encoding Scaling

↳ Encoding

↳ Label Encoder

sex → is-female
male 0
female 1

↳ `nunique()` → NaN'ı saymaz
`len(unique())` → sayar

⚠ Label encoder NaN'ları da dolduruyor
binary bir değişken 0, 1, 2, ... oluyo

↳ One Hot Encoder

⚠ Dummy değişken tuzağı

XYZ	X	Y	Z
Z	0	0	1
X	1	0	0
Y	0	1	0

 → X, Y, Z sınıfı yaratıldıktan sonra XYZ silinmeli

↳ Nominal değişken → Aralarında sayılarla belirtilen bir bağlantı yok

↳ Rare Encoding

↳ Bir verinin gözlenmesi çok düşükse onu tutmanın bir anlamı yok

1) Kategorik değ. azlık çokluk durumlarını analiz et

2) Rarelerle bağımlı değişken arasında ilişki var mı?

↳ Özellik Ökelleştirme

↳ Değişkenler arasındaki ölçüm farklılığını gidermek, standartlaştırmak gerekiyor ki kullanılacak modeller değişkenlere eşit şartlarda yaklaşsın

↳ Amaçlardan biri de algoritmaların train sürelerini azaltmaktır

↳ Errorların sayıları ve errorların giderilme süresi, değişkenler standart olduğunda daha kısa oluyo

↳ Standart Scaler

↳ Bütün gözlem birimlerinden standart ortalamayı çıkartıp standart sapmaya böler

⚠ Veri setindeki aykırı değerlerden etkilenir

↳ Robust Scaler

↳ Bütün gözlem birimlerinden medyanı çıkarır, IQR'a böler

⚠️ Fykirı deęerlere daha dayanıklı

↳ MinMax Scaler

↳ Dönüştürmek istediğin özel bir aralık varsa mantıklı

⚠️ Ölçeklendirirken deęiskenlerin yapılarını bozmadan, ifade edilis tarzını deęistirebilirsin

↳ Sayısal deęiskenleri kategorik deęiskenlere çevirme → Binning
(cut(), qcut())

4- Feature Extraction (Feature Engineering)

↳ Özellik Çıkarımı

"Yapısal ve yapısal olmayan (görüntü ses vs) verilerden deęisken üretmek"

"Ham veriden deęisken üretmek"

↳ Binary Features

0, 1 deęiskenler üzerinden yeni deęiskenler üretmek

↳ propositions - z test → pvalue 0 sa baęlantı var

↳ Text Features

Metinler üzerinden deęisken üretme

↳ Regex Features

Regular expressionlar ile deęisken üretmek → Mr, Ms, Doctor yakalama