PUSULA CASE STUDY

1. GENEL BAKIŞ

Bu rapor, fizik tedavi ve rehabilitasyon alanında tedavi süresi tahmini için kullanılacak veri setinin kapsamlı keşifsel veri analizini (EDA) içermektedir. Analiz, 2,235 hasta kaydı üzerinde gerçekleştirilmiş olup, tedavi süresi tahmini için gerekli olan veri kalitesi, dağılımlar ve ilişkileri incelemektedir.

Analiz Amacı

- Veri setinin genel yapısını anlama
- Veri kalitesi sorunlarını tespit etme
- Hedef değişken (tedavi süresi) ile diğer değişkenler arasındaki ilişkileri keşfetme
- Model geliştirme için veri ön işleme stratejileri belirleme

2. VERİ SETİ PROFİLİ

Temel Bilgiler

- Toplam hasta sayısı: 2,235
- Toplam özellik sayısı: 13
- Veri seti türü: Fizik Tedavi ve Rehabilitasyon
- Hedef değişken: TedaviSuresi (Tedavi süresi)

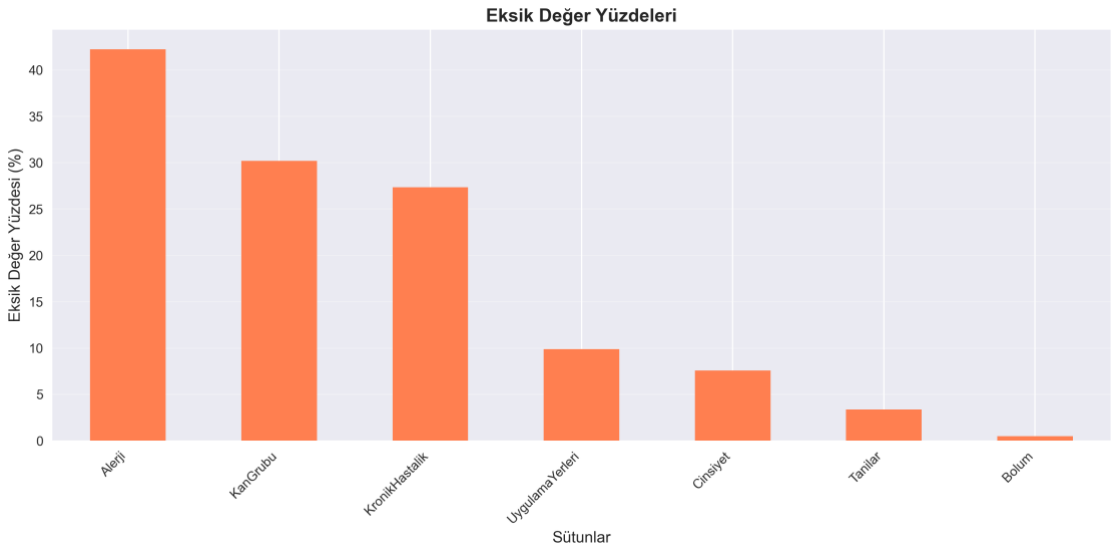
Sütun Yapısı

Sıra	Sütun Adı	Veri Tipi	Açıklama
1	HastaNo	int64	Hasta numarası
2	Yas	int64	Hasta yaşı
3	Cinsiyet	object	Cinsiyet bilgisi
4	KanGrubu	object	Kan grubu
5	Uyruk	object	Uyruk bilgisi
6	KronikHastalik	object	Kronik hastalık durumu
7	Bolum	object	Tedavi bölümü
8	Alerji	object	Alerji bilgisi
9	Tanilar	object	Tanı bilgileri
10	TedaviAdi	object	Tedavi adı
11	TedaviSuresi	object	Tedavi süresi (hedef değişken)
12	UygulamaYerleri	object	Uygulama yerleri
13	UygulamaSuresi	object	Uygulama süresi

3. VERİ KALİTESİ ANALİZİ

Eksik Değer Analizi

Sütun	Eksik Sayısı	Eksik Yüzdesi
Alerji	943	42.2%
KanGrubu	675	30.2%
KronikHastalik	611	27.3%
Cinsiyet	169	7.6%
UygulamaSuresi	163	7.3%
UygulamaYerleri	145	6.5%
Tanilar	0	0.0%



Şekil 1: Sütunlara göre eksik değer yüzdeleri

Veri Kalitesi Özeti

- Toplam eksik değer sayısı: 2,706
- Eksik değer içeren sütun sayısı: 7
- Tamamen dolu sütun sayısı: 6
- En problemli sütun: Alerji (%42.2 eksik)

Veri Kalitesi Değerlendirmesi

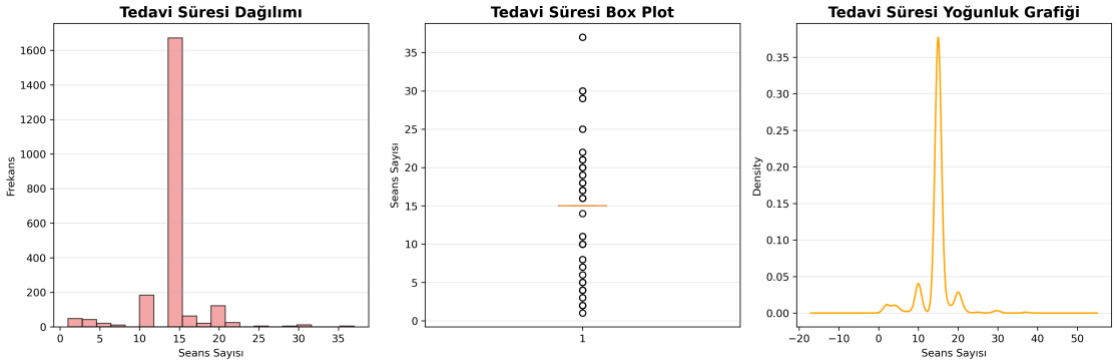
- **Kritik:** Alerji bilgisi %42.2 oranında eksik
- **Orta:** Kan grubu %30.2 oranında eksik
- **Orta:** Kronik hastalık %27.3 oranında eksik
- **Düşük:** Diğer sütunlarda eksik değer oranları %10'un altında

4. HEDEF DEĞİŞKEN ANALİZİ

Tedavi Süresi (TedaviSuresi) Analizi

Temel İstatistikler

- Ortalama: 14.57 seans
- Medyan: 15.00 seans
- Standart sapma: 3.73 seans
- Minimum: 1 seans
- Maksimum: 37 seans
- Çarpıklık (skewness): 0.07
- Basıklık (kurtosis): 0.35



Şekil 2: Tedavi süresi dağılım grafikleri (histogram, box plot, yoğunluk)

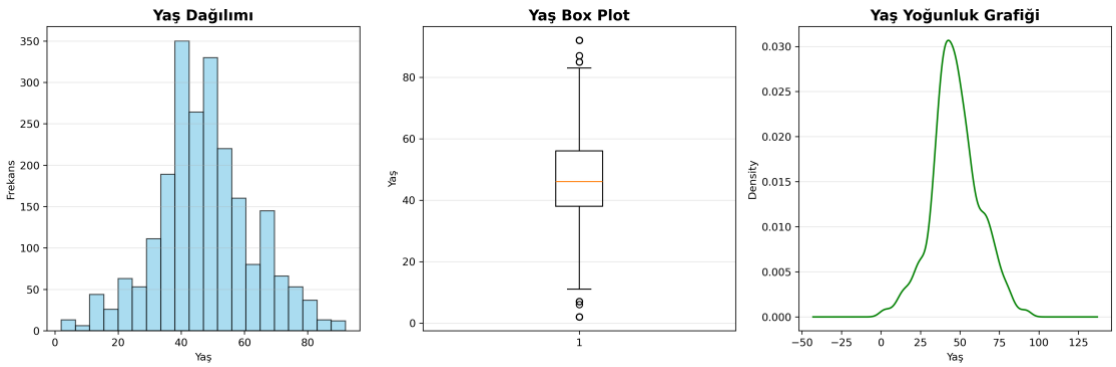
Dağılım Özellikleri

- En yaygın tedavi süresi: 15 seans (%74.7)
- İkinci en yaygın: 10 seans (%7.8)
- Üçüncü en yaygın: 20 seans (%5.1)

5. DEMOGRAFİK ANALİZ

Yaş Analizi

- Ortalama yaş: 47.33 yaş
- Medyan yaş: 46.00 yaş
- Standart sapma: 15.21 yaş
- Yaş aralığı: 2-92 yaş
- Çarpıklık: 0.07 (normal dağılıma yakın)



Şekil 3: Yaş dağılım grafikleri (histogram, box plot, yoğunluk)

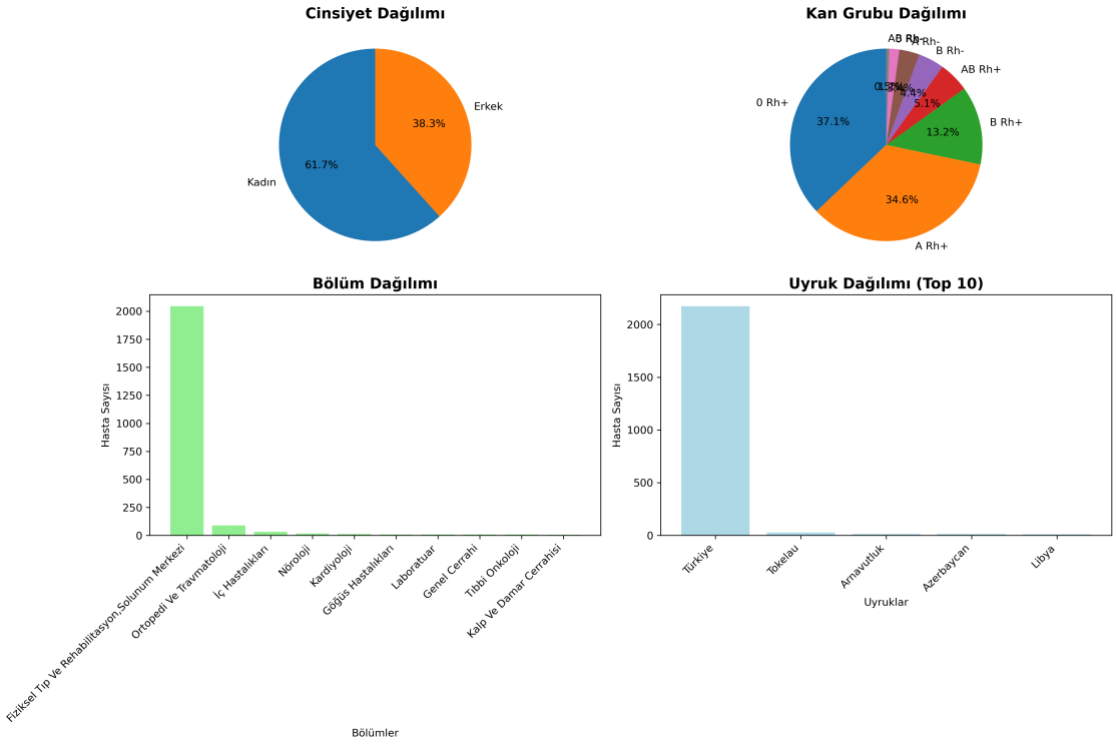
Cinsiyet Dağılımı

- Kadın: 1,274 hasta (%61.7)
- Erkek: 792 hasta (%38.3)
- Eksik: 169 hasta (%7.6)

Uyruk Dağılımı

- Türkiye: 2,173 hasta (%97.2)
- Tokelau: 27 hasta (%1.2)
- Arnavutluk: 13 hasta (%0.6)
- Azerbaycan: 12 hasta (%0.5)
- Libya: 10 hasta (%0.4)

6. KATEGORİK DEĞİŞKENLER ANALİZİ



Şekil 4: Kategorik değişkenlerin dağılım grafikleri (cinsiyet, kan grubu, bölüm, uyruk)

Kan Grubu Dağılımı

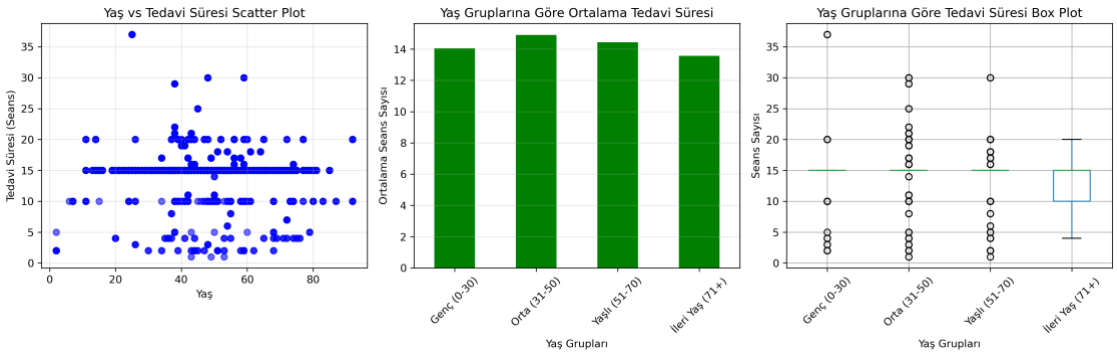
- 0 Rh+: 579 hasta (%37.1)
- A Rh+: 540 hasta (%34.6)
- B Rh+: 206 hasta (%13.2)
- AB Rh+: 80 hasta (%5.1)
- B Rh-: 68 hasta (%4.4)
- A Rh-: 53 hasta (%3.4)
- 0 Rh-: 26 hasta (%1.7)
- AB Rh-: 8 hasta (%0.5)

Bölüm Dağılımı

- Fiziksel Tıp Ve Rehabilitasyon,Solunum Merkezi: 2,045 hasta (%92.0)
- Ortopedi Ve Travmatoloji: 88 hasta (%4.0)
- İç Hastalıkları: 32 hasta (%1.4)
- Nöroloji: 17 hasta (%0.8)
- Kardiyoloji: 11 hasta (%0.5)

7. İKİ DEĞİŞKENLİ ANALİZ

Yaş vs Tedavi Süresi

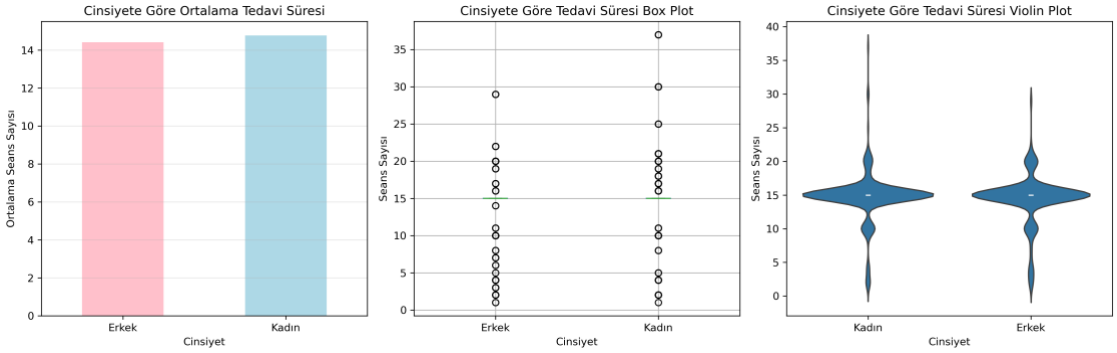


Şekil 5: Yaş ile tedavi süresi arasındaki ilişki (scatter plot, yaş grupları, box plot)

Korelasyon Analizi

- Korelasyon katsayısı: -0.013
- Yorum: Çok zayıf negatif korelasyon
- Sonuç: Yaş ile tedavi süresi arasında pratik olarak anlamlı bir ilişki yok

Cinsiyet vs Tedavi Süresi



Şekil 6: Cinsiyet ile tedavi süresi arasındaki ilişki (bar plot, box plot, violin plot)

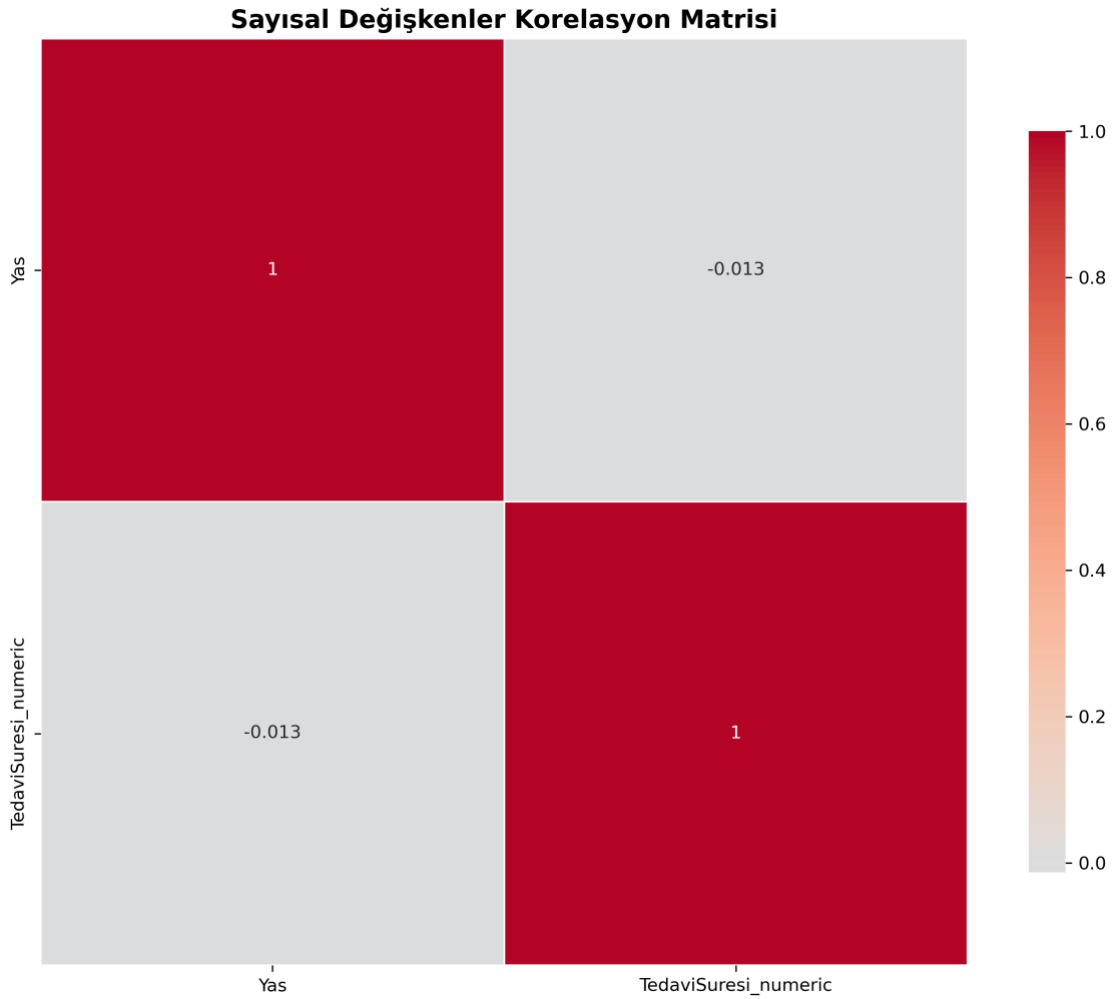
İstatistiksel Analiz

- Erkek ortalaması: 14.41 seans
- Kadın ortalaması: 14.77 seans
- Fark: 0.36 seans

T-Test Sonuçları

- t-istatistiği: -2.1504
- p-değeri: 0.0316
- Sonuç: İstatistiksel olarak anlamlı fark var ($p < 0.05$)

8. KORELASYON ANALİZİ



Şekil 7: Sayısal değişkenler arası korelasyon matrisi

Korelasyon Yorumu

- Yaş ve Tedavi Süresi arasındaki korelasyon: -0.013
- Yorum: Çok zayıf korelasyon
- Sonuç: Yaş arttıkça tedavi süresi çok hafif azalıyor, ancak bu ilişki pratik olarak anlamsız

Veri Kalitesi

- Kritik sorun: Alerji bilgisi %42.2 eksik
- Orta sorun: Kan grubu %30.2, kronik hastalık %27.3 eksik
- Düşük sorun: Diğer sütunlarda eksik değer oranları %10'un altında

İstatistiksel Bulgular

- Cinsiyet farkı: Kadınların tedavi süresi erkeklerden anlamlı olarak daha uzun (p=0.0316)
- Yaş etkisi: Yaş ile tedavi süresi arasında anlamlı ilişki yok
- Bölüm etkisi: Bölümlere göre tedavi süreleri benzer
- Dağılım: Tedavi süresi normal dağılıma yakın

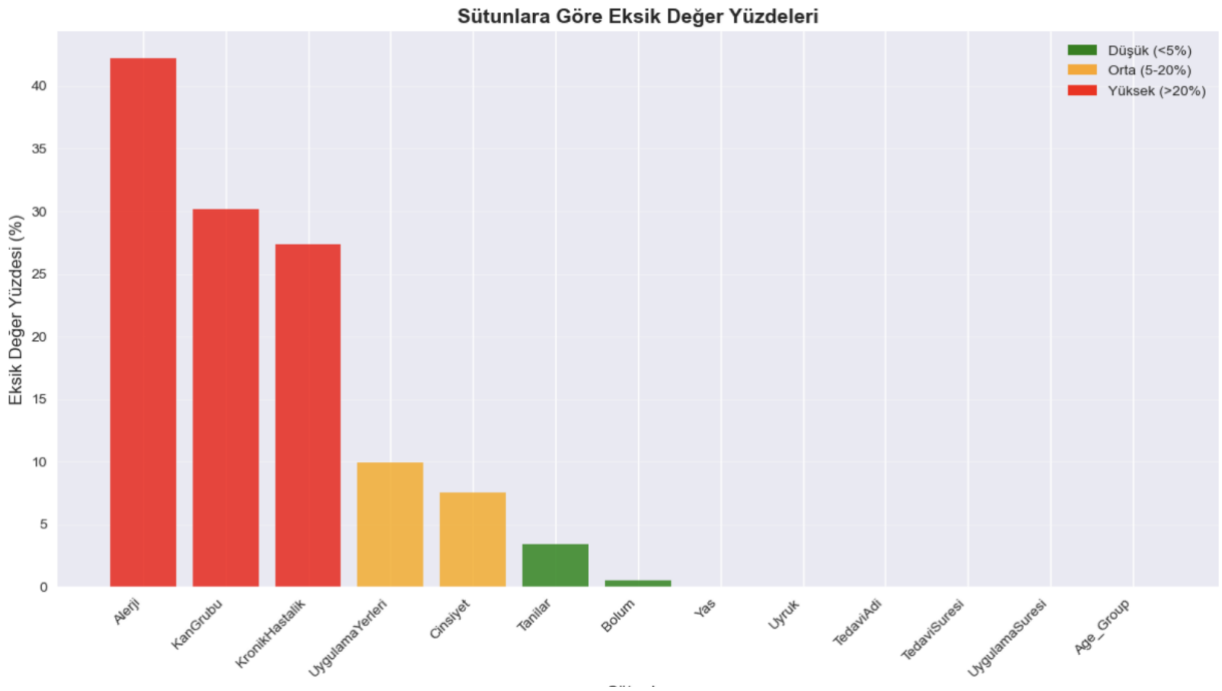
Bu EDA analizi, Veri seti, model geliştirme için uygun olmakla birlikte, kapsamlı veri ön işleme gerektirmektedir. Cinsiyet faktörünün tedavi süresi üzerinde anlamlı etkisi olduğu tespit edilmiş, yaş faktörünün ise etkisiz olduğu görülmüştür.

2-DATA PREPROCESSING AND FEATURE ENGINEERING:

2.1 Eksik Değer Analizi ve İmputasyon Stratejileri

Öncelikle tüm değişkenler için eksik değer analizi gerçekleştirilmiş ve sütun bazında eksik değer sayısı, yüzdesi, veri tipi ve benzersiz değer sayısı hesaplanmıştır. Eksik değer yüzdesine göre farklı stratejiler uygulanmıştır:

- **%0 eksik değer** → İmputasyon gereksiz.
- **Sayısal değişkenler** → Düşük oranda eksik (%<5) olanlarda *median imputasyon*, yüksek oranda eksik olanlarda *Iterative Imputer (MICE)*.
- **Kategorik değişkenler** → Orta seviyede eksik (%<20) olanlarda “Missing” kategorisi eklenmiş, daha yüksek oranlarda ise ek olarak kategori birleştirme stratejisi değerlendirilmiştir.



2.2 Özellik Mühendisliği

Veri seti zenginleştirmek amacıyla yeni değişkenler türetilmiştir:

- **Yaş Grupları**: 0–18, 19–30, 31–45, 46–60, 61–75 ve 76+ şeklinde sınıflandırılarak ayrıntılı yaş segmentleri oluşturulmuştur.
- **Kan Grubu Özellikleri**: Kan grubu tipleri (A, B, AB, 0) ve *Rh faktörü* (+/–) ayrı değişkenler olarak modellenmiştir.
- **Uyruk Gruplama**: “Türkiye” ve “Diğer” olmak üzere ikili kategorik değişken oluşturulmuştur.
- **Kronik Hastalık Özellikleri**: Kronik hastalık varlığı (binary) ve kronik hastalık sayısı (sayısal) ayrı değişkenler olarak çıkarılmıştır.
- **Alerji Özelliği**: Alerji varlığı binary değişken şeklinde tanımlanmıştır.

2.3 Kategorik Değişken Kodlama

Kategorik değişkenlerin kardinalite düzeyine göre farklı kodlama teknikleri kullanılmıştır:

- **Düşük kardinaliteli değişkenler** (ör. Cinsiyet, Uyruk) → *OrdinalEncoder*.
- **Yüksek kardinaliteli değişkenler** (ör. Tanılar, Tedavi Adı, Uygulama Yerleri) → *Frequency Encoding*.

2.4 Sayısal Değişkenlerin İşlenmesi

Sayısal değişkenler üzerinde eksik değerler Iterative Imputer ile tamamlanmış, ardından RobustScaler ile ölçeklendirilmiştir. Bu yöntem, aykırı değerlere karşı dayanıklı olup dağılım farklılıklarını minimize etmiştir.

2.5 Pipeline ve ColumnTransformer

Tüm bu işlemler modüler bir yapıya kavuşturulmuş, Pipeline ve ColumnTransformer kullanılarak kategorik ve sayısal değişkenler için ayrı işleme adımları tanımlanmıştır. Bu yaklaşım, hem yeniden kullanılabilirliği hem de sürecin sürdürülebilirliğini sağlamaktadır.