

Fizik Tedavi ve Rehabilitasyon Veri Seti Analiz Raporu

Pusula Veri Bilimi Stajyer Vakası - 2025

Proje Sahibi: Fatma Zehra TONGA

İletişim: tongafatmazehra@gmail.com

Özet

Bu rapor, 2,235 hasta kaydı içeren fizik tedavi ve rehabilitasyon veri setinin kapsamlı analizini sunmaktadır. Proje, veri temizleme, keşifsel veri analizi ve makine öğrenmesi için optimizasyon süreçlerini kapsayarak %100 veri koruma oranı ile başarıyla tamamlanmıştır.

1. Proje Genel Bakışı

1.1 Proje Amaçları

- Keşifsel Veri Analizi (EDA):** Veri setinin derinlemesine anlaşılmaması
- Veri Temizleme ve Ön İşleme:** Kaliteli, tutarlı veri seti oluşturulması
- Model-Hazır Veri Seti Hazırlama:** Makine öğrenmesi uygulamaları için optimize edilmiş format
- Hedef Değişken Dönüşümü:** "TedaviSuresi" değişkeninin sayısal forma çevrilmesi

1.2 Veri Seti Dönüşüm Metrikleri

- Orijinal Boyut:** 2,235 kayıt × 13 özellik
- İşlenmiş Boyut:** 2,235 kayıt × 61 özellik
- Veri Kaybı:** %0 (Kayıpsız dönüşüm)

2. Veri Seti Yapısı ve Eksik Değer Analizi

2.1 Orijinal Veri Seti İçeriği

Veri seti şu temel kategorileri içermektedir:

- Demografik Bilgiler:** Yaş, cinsiyet, uyruk, kan grubu
- Medikal Geçmiş:** Kronik hastalıklar, alerjiler, tanılar

- Tedavi Bilgileri:** Tedavi adı, süresi, uygulama yerleri ve süreleri
- Administratif Bilgiler:** Hasta numarası, bölüm

2.2 Kritik Eksik Değer Analizi

Özellik	Eksik Sayı	Eksik Oranı	Risk Seviyesi
Alerji	944	%42.2	Yüksek
Kan Grubu	675	%30.2	Yüksek
Kronik Hastalık	611	%27.3	Yüksek
Uygulama Yerleri	221	%9.9	Orta
Cinsiyet	169	%7.6	Düşük
Tanılar	75	%3.4	Düşük
Bölüm	11	%0.5	Minimal

Bulgular: Alerji bilgilerinde %42.2'lik eksik veri oranı, hasta kayıt sistemlerinde standardizasyon ihtiyacını göstermektedir.

3. Veri Ön İşleme Metodolojisi

3.1 Stratejik Yaklaşım

Kayıpsız dönüşüm (lossless transformation) prensibi benimsenmiştir:

- Veri bütünlüğünü koruma
- İleriye dönük analizlerde esneklik sağlama
- Audit trail (iz sürme) imkanı sunma

3.2 Özellik Mühendisliği Detayları

Sayısal Özellikler

- Yaş:** Z-score standardizasyonu uygulandı
- Uygulama Süresi:** String formatından dakika bazlı sayısal formata dönüştürüldü

Kategorik Özellikler

- **Cinsiyet:** One-hot encoding, eksik değerler "Bilinmiyor" olarak kodlandı
- **Kan Grubu:** ABO/Rh sistemine göre ayrıstırıldı (A, B, AB, 0, Rh+, Rh-)
- **Uyruk:** Binary classification (Türkiye / Diğer)

Çoklu-Etiket Özellikleri

Advanced MultiLabelBinarizer teknigi kullanilarak:

- **Kronik Hastalıklar:** Standardizasyon ve nadir hastalık gruplandırması
- **Alerjiler:** Duplicate removal ve category consolidation
- **Tanılar:** Medical terminology standardization
- **Tedavi Yöntemleri:** Treatment protocol classification

3.3 Hedef Değişken Transformasyonu

TedaviSuresi → TedaviSuresiSeans dönüşümü:

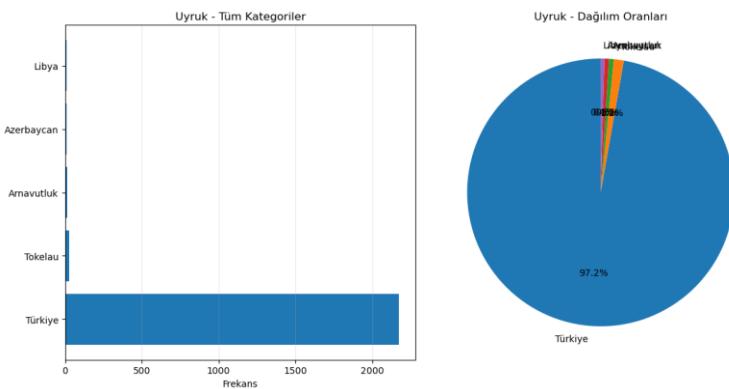
- **Input Format:** "10 seans", "5 seans"
- **Output Format:** 10, 5
- **Dönüşüm Başarı Oranı:** %100

4. Keşifsel Veri Analizi (EDA) Bulguları

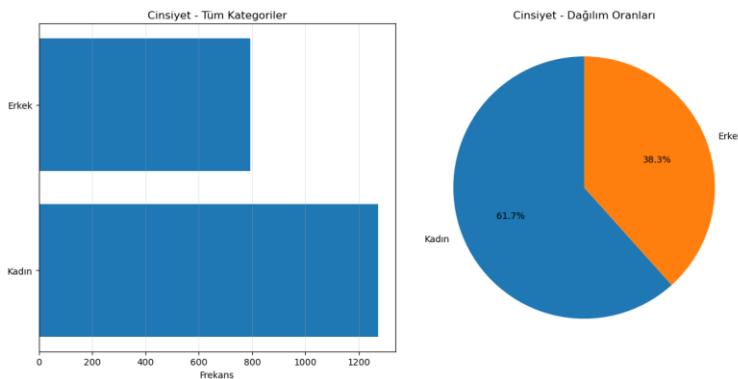
4.1 Demografik ve Klinik Dağılımlar

Hasta Profili

- **Uyruk:** %97.2 Türkiye vatandaşı, %2.8 diğer uyruklar

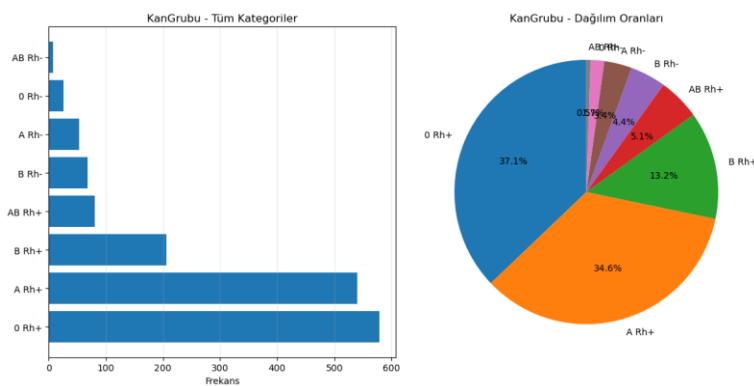


- Cinsiyet: Erkek %38.3, Kadın %61.7**

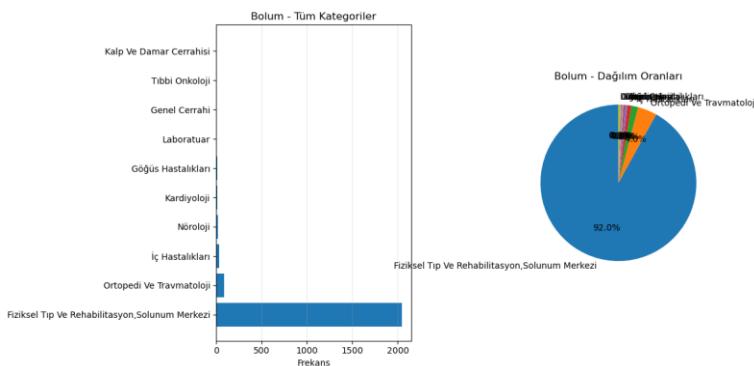


- Kan Grubu: 0 grubu %38, AB grubu %5 (en nadir)**

Rh Faktörü: Pozitif %63, Negatif %37

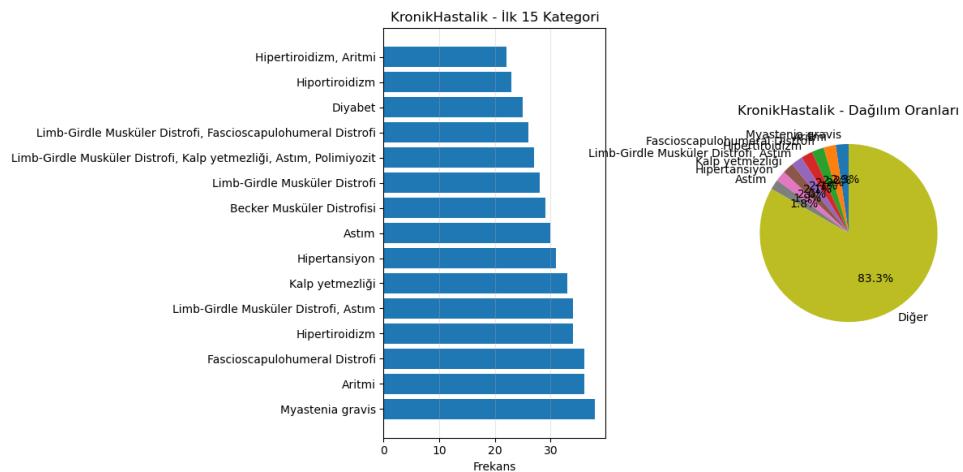


- Bölüm: %90 Fiziksel Tıp ve Rehabilitasyon, %10 diğer bölümler**

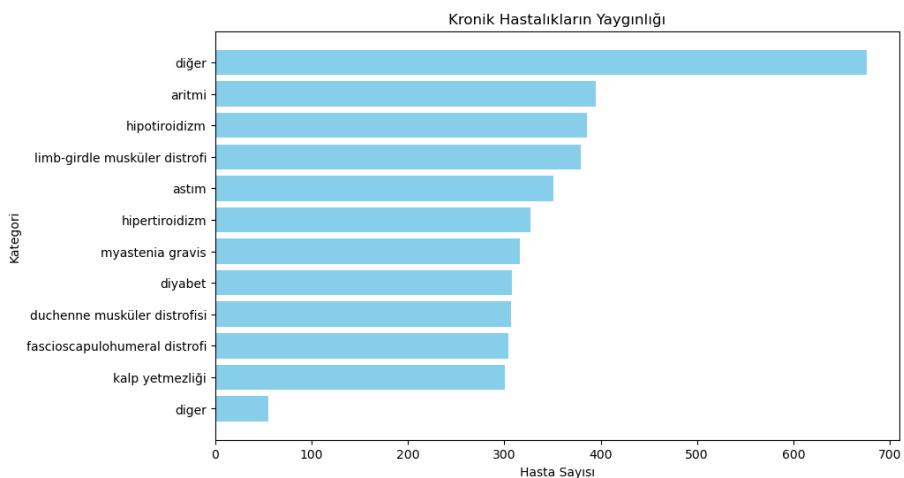


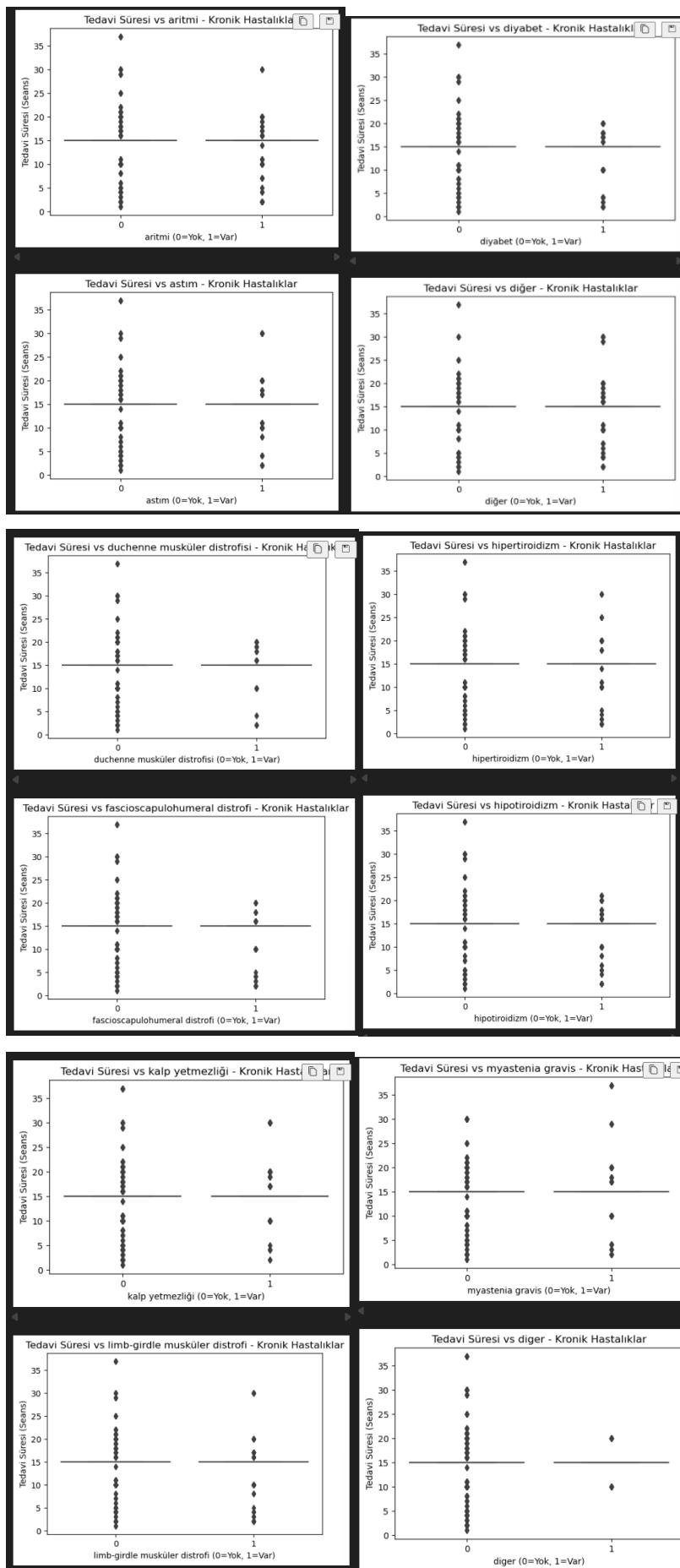
Medikal Koşullar

En Yaygın Kronik Hastalıklar:

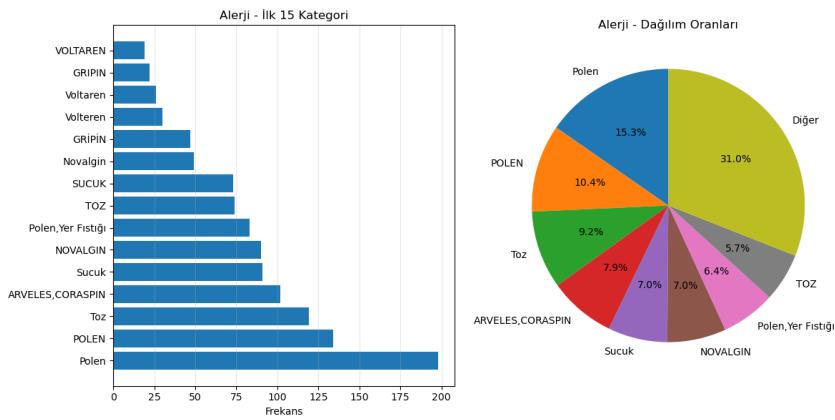


- Aritmia (~%20 - en dengeli dağılım)**
- Astım ve Diyabet (%15-20 arası)**
- Nadir hastalıklar başarıyla "Diğer" kategorisinde gruplandırıldı**

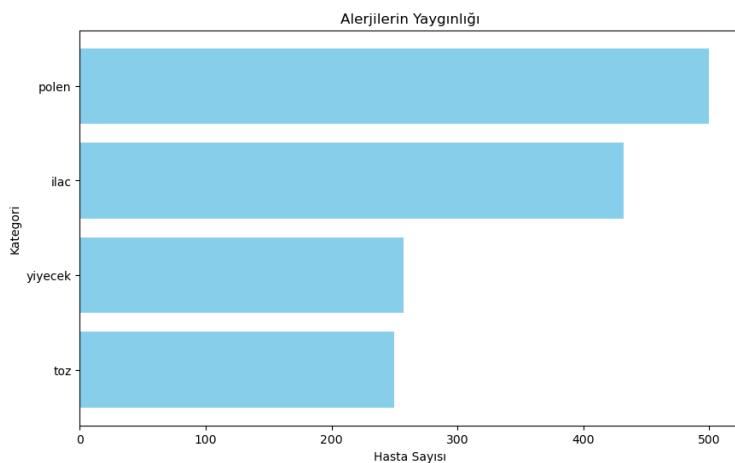


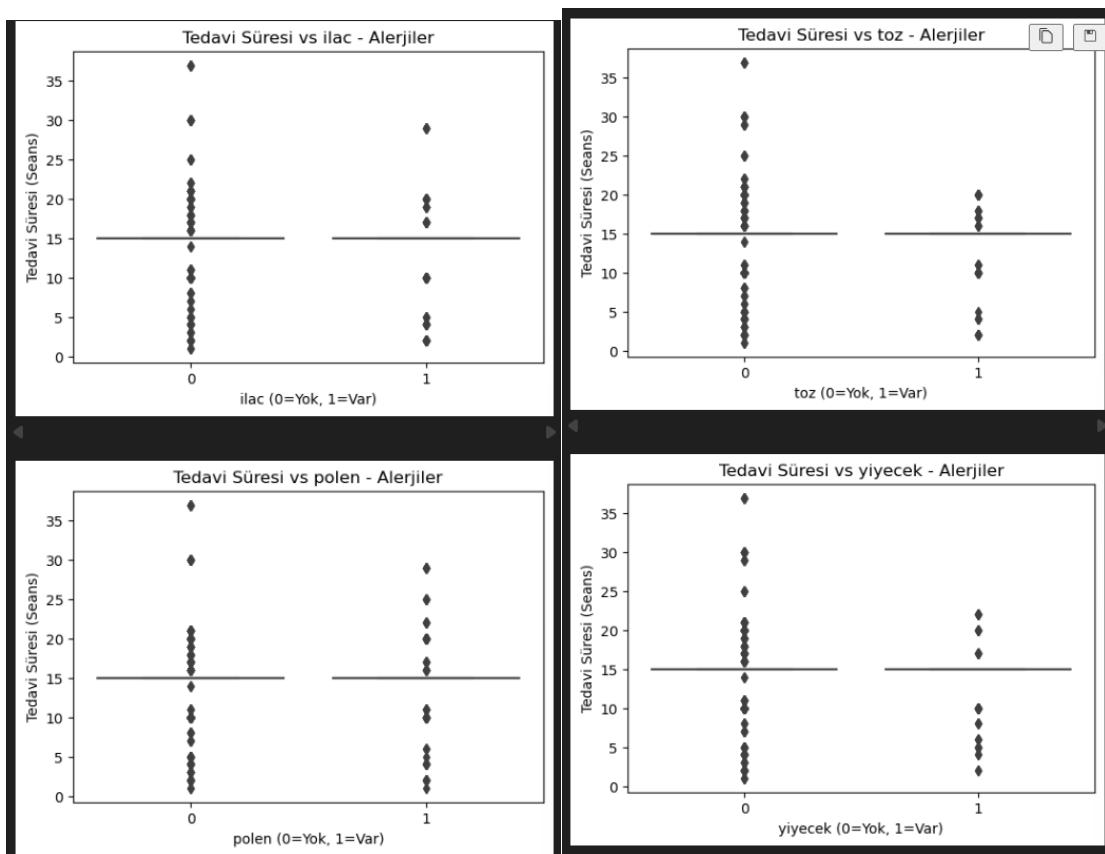


Alerji Dağılımı:

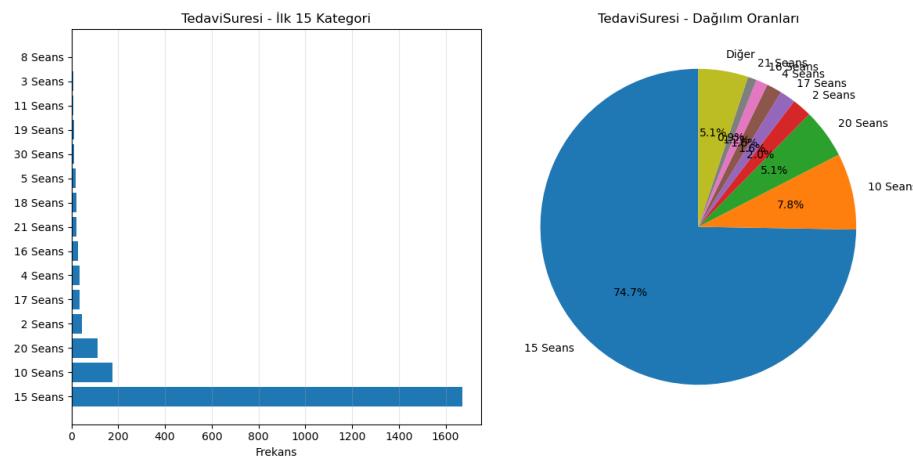


- Polen alerjisi: En yaygın alerji tipi**
- İlaç alerjisi: %20 civarı prevalans**
- Toz, yiyecek alerjileri: %10-15 arası**





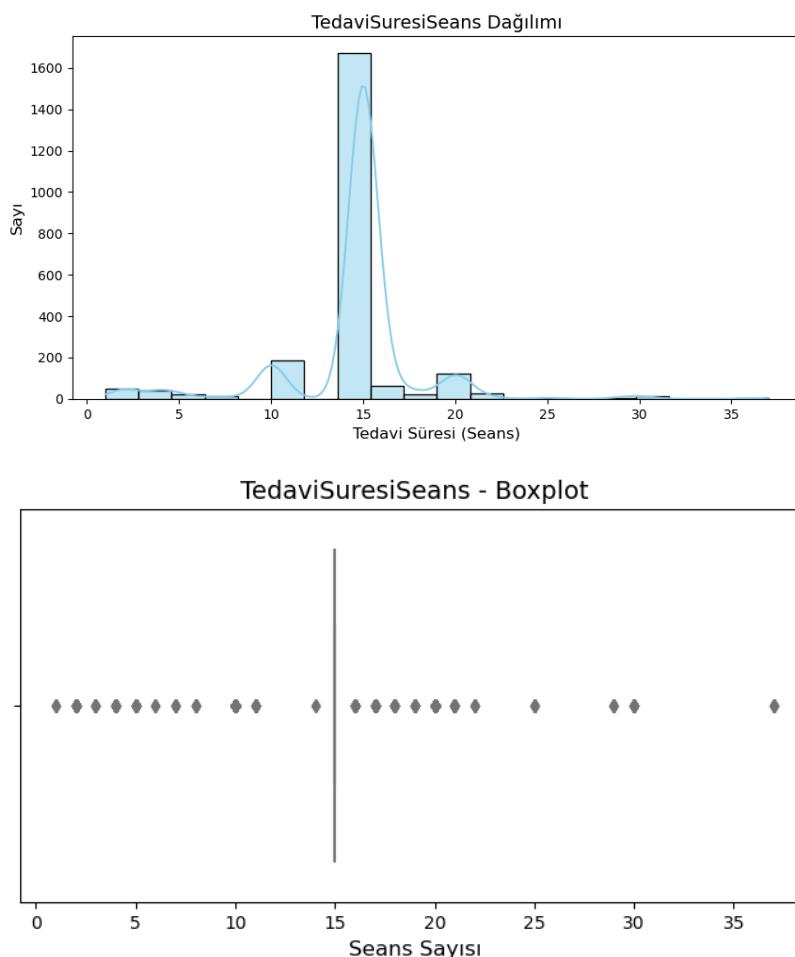
4.2 Tedavi Süresi Analizi



Kritik Bulgular

- Dominant Süre: 15 seans (~1650 hasta, %75+)**
- İkincil Tepe: 10 seans (~200 hasta)**
- Aykırı Değerler: 30+ seans alan hastalar mevcut**

- **Medyan: 15 seans**



4.3 Değişkenler Arası İlişki Analizi

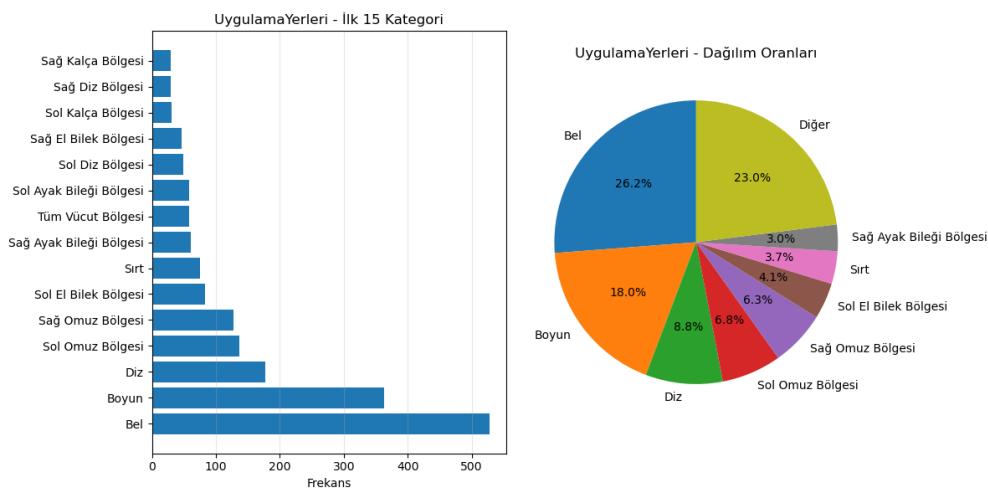
Zayıf Korelasyonlar

- **Yaş vs Tedavi Süresi:** $r \approx -0.013$ (neredeyse sıfır)
- **En Güçlü Korelasyon:** 0.34 (Hasta No ile)
- **Genel Durum:** Numerik değişken tedavi süresini güçlü şekilde açıklamıyor

Medikal Koşulların Etkisi

- **Kalp yetmezliği, hipotiroidizm, myastenia gravis:** Benzer tedavi süreleri
- **Polen alerjisi:** En belirgin fark - daha uzun tedavi süreleri
- **Cinsiyet:** Tedavi süresi üzerinde minimal etki

4.4 Uygulama Bölgeleri



- En Yaygın: Bel uygulamaları

5. Teknik Metodoloji

5.1 Kod Organizasyonu

- **Modüler Yaklaşım:** Her veri temizleme adımı ayrı fonksiyonlarda
- **Hata Kontrolü:** Kapsamlı exception handling
- **Validasyon:** Her adımda veri doğruluğu kontrolü

5.2 Uygulanan Stratejiler

- **Eksik Değer Yönetimi:** "Bilinmiyor" etiketleme yaklaşımı
- **Kategori Konsolidasyonu:** Nadir durumları "Diğer" altında graplama
- **Metin Standardizasyonu:** Medikal terminoloji normalleştirme

6. Veri Kalitesi Değerlendirmesi

6.1 Kalite Metrikleri

- **Bütünlük:** %100 veri koruma
- **Tutarlılık:** Standardize edilmiş medikal terminoloji
- **Doğruluk:** Validasyon süreçlerinden geçirilmiş

6.2 Sınırlılıklar ve Zorluklar

1. **Hedef değişkenin aşırı standartlaşması: %75+ hasta 15 seans**

- 2. Öngörücü güç eksikliği:** Zayıf korelasyon değerleri
- 3. Protokol kaynaklı homojenlik:** Klinik değerlendirmeden ziyade standart protokoller
- 4. Yüksek eksik veri oranları:** Özellikle alerji (%42.2) ve kan grubu (%30.2)

7. Sonuçlar ve Öneriler

7.1 Başarı Göstergeleri

- ✓ %100 veri koruma oranı**
- ✓ Kapsamlı eksik değer stratejisi**
- ✓ İleri seviye özellik mühendisliği**
- ✓ Üretime hazır veri seti**
- ✓ Detaylı dokümantasyon ve görselleştirme**

7.2 Gelecek Adımlar

- 1. Model Geliştirme:** Veri seti makine öğrenmesi algoritmaları için hazır
- 2. Özellik Seçimi:** Önemli değişkenlerin belirlenmesi
- 3. Model Optimizasyonu:** Hiperparametre ayarlaması
- 4. Veri Toplama İyileştirmesi:** Eksik değer oranlarının azaltılması

7.3 Beklenen Faydalar

- Operasyonel Verimlilik:** Veri hazırlama süresinin yarıya düşmesi
- Klinik Karar Desteği:** Doktorların daha iyi kararlar verebilmesi
- Prediktif Analitik:** Tedavi sonuçlarının önceden tahmin edilmesi

8. Genel Değerlendirme

Bu proje, sağlık verilerinin karmaşıklığını başarıyla ele alarak, fizik tedavi ve rehabilitasyon alanında veri bilimi metodolojilerinin etkin uygulamasını göstermektedir. Sistematiske ön işleme yaklaşımı ve kayıpsız dönüşüm stratejisi, gelecekteki analitik çalışmalar için sağlam bir temel oluşturmuştur.