Pusula Talent Data Science Case Study - Rapor

Ad Soyad: Ramazan Karakılınç

Mail Adresi: ramazankarakilinc06@gmail.com

📋 Proje Özeti

Proje Adı: Physical Medicine & Rehabilitation Dataset Analysis **Tarih:** 06/09/2025 **Veri Seti:** 2235 gözlem, 13 özellik **Hedef Değişken:** TedaviSuresi (Seans cinsinden) **Amaç:** Kapsamlı EDA ve model-ready veri hazırlama

Veri Seti Hakkında:

Bu çalışmada, fizik tedavi ve rehabilitasyon alanındaki 2235 hasta kaydı üzerinde kapsamlı bir keşifsel veri analizi (EDA) ve veri ön işleme süreci gerçekleştirilmiştir. Veri seti, hasta demografik bilgileri, tıbbi geçmiş, tedavi detayları ve hedef değişken olan tedavi süresini içermektedir.

Ana Bulgular:

- **2,706 eksik değer** başarıyla dolduruldu
- **53 feature** ile model-ready veri seti oluşturuldu
- **392 hasta** birden fazla kayda sahip
- **Tedavi süresi** kategorilere ayrıldı (Kısa/Orta/Uzun)

📊 Veri Seti Genel Bilgileri

Veri Seti Özellikleri

- **Toplam Gözlem:** 2,235
- **Toplam Özellik:** 13
- **Sayısal Özellikler:** 2 (Yaş, HastaNo)
- **Kategorik Özellikler:** 11
- **Hedef Değişken:** TedaviSuresi (Seans)

Özellik Açıklamaları

Özellik	Açıklama	Veri Tipi
HastaNo	Anonim hasta kimliği	Sayısal
Yas	Yaş	Sayısal
Cinsiyet	Cinsiyet	Kategorik

Özellik	Açıklama	Veri Tipi
KanGrubu	Kan grubu	Kategorik
Uyruk	Uyruk	Kategorik
KronikHastalik	Kronik hastalıklar	Kategorik
Bolum	Bölüm/Klinik	Kategorik
Alerji	Alerjiler	Kategorik
Tanilar	Tanılar	Kategorik
TedaviAdi	Tedavi adı	Kategorik
TedaviSuresi	Tedavi süresi (seans)	Hedef
UygulamaYerleri	Uygulama yerleri	Kategorik
UygulamaSuresi	Uygulama süresi	Kategorik

🔍 Exploratory Data Analysis (EDA) Sonuçları

1. Veri Kalitesi Analizi

Eksik Değer Durumu (Başlangıç)

Özellik	Eksik Değer Sayısı	Eksik Değer Oranı (%)
Cinsiyet	169	7.6%
KanGrubu	675	30.2%
KronikHastalik	611	27.3%
Bolum	11	0.5%
Alerji	944	42.2%

Özellik	Eksik Değer Sayısı	Eksik Değer Oranı (%)
Tanilar	75	3.4%
UygulamaYerleri	221	9.9%

Tekrarlanan Kayıt Analizi:

- **Toplam hasta sayısı:** 404
- **Birden fazla kaydı olan hasta sayısı:** 392
- **Tekrarlanan kayıt sayısı:** 1,831
- **Ortalama hasta başına kayıt:** 5.5

2. Hedef Değişken Analizi

TedaviSuresi Dağılımı

Kategori	Hasta Sayısı	Oran (%)
Kısa (≤10 seans)	296	13.2%
Orta (11-20 seans)	1,887	84.4%
Uzun (>20 seans)	52	2.3%

İstatistiksel Özet

- **Ortalama:** 14.57 seans
- **Standart Sapma:** 3.73
- **Minimum:** 1 seans
- **Maksimum:** 37 seans
- **Medyan:** 15 seans

3. Kategorik Değişken Analizi

Cinsiyet Dağılımı

Cinsiyet	Hasta Sayısı	Oran (%)
Kadın	1,298	58.1%
Erkek	792	35.4%
Bilinmiyor	145	6.5%

Kan Grubu Dağılımı

Kan Grubu	Hasta Sayısı	Oran (%)
0 Rh+	590	26.4%
A Rh+	546	24.4%
B Rh+	211	9.4%
AB Rh+	82	3.7%
Diğerleri	806	36.1%

En Yaygın Kronik Hastalıklar

Hastalık	Hasta Sayısı
Myastenia gravis	38
Aritmi	36
Fascioscapulohumeral Distrofi	36
Hipertansiyon	36

Hastalık	Hasta Sayısı
Limb-Girdle Musküler Distrofi, Astım	34

4. Aykırı Değer Analizi

Aykırı Değer Özeti

Özellik	Aykırı Değer Sayısı	Aykırı Değer Oranı (%)
Yas	41	1.83%
UygulamaSuresi	12	0.54%
TedaviSuresi	565	25.28%

🛠️ Veri Ön İşleme Süreci

1. Eksik Değer Doldurma Stratejileri

Strateji 1: Aynı Hasta Kayıtlarından Doldurma

- **Uygulanan Özellikler:** Cinsiyet, KanGrubu, KronikHastalik, Tanilar
- **Doldurulan Değer Sayısı:** 73
- **Mantık:** Aynı hastanın sabit özellikleri

Strateji 2: TedaviAdi Bazlı Doldurma

- **Uygulanan Özellikler:** Bolum, UygulamaYerleri, Tanilar
- **Doldurulan Değer Sayısı:** 307
- **Mantık:** Aynı tedavi türü için standart değerler

Strateji 3: Mantıklı Varsayımlar

- **Alerji:** Eksik değerler "Yok" ile dolduruldu
- **Kalan eksikler:** "Bilinmiyor" ile dolduruldu

2. Veri Temizleme İşlemleri

Alerji Sütunu Standardizasyonu

- Büyük/küçük harf tutarsızlıkları düzeltildi
- Yazım hataları giderildi
- Virgülle ayrılmış değerler standardize edildi

Sayısal Dönüşümler

- **TedaviSuresi:** "15 Seans" → 15
- **UygulamaSuresi:** "20 Dakika" → 20

3. Encoding İşlemleri

OneHotEncoder (≤20 kategori)

- Cinsiyet (3 kategori)
- KanGrubu (9 kategori)
- Uyruk (5 kategori)
- Bolum (11 kategori)
- Alerji (20 kategori)
- TedaviSuresi_Kategori (3 kategori)

LabelEncoder (>20 kategori)

- KronikHastalik (221 kategori)
- Tanilar (370 kategori)
- TedaviAdi (244 kategori)
- UygulamaYerleri (38 kategori)

4. Normalizasyon

- **StandardScaler** ile sayısal özellikler normalize edildi
- Hedef değişken normalizasyondan çıkarıldı

📈 Model-Ready Veri Seti

Final Veri Seti Özellikleri

- **Toplam Gözlem:** 2,235
- **Feature Sayısı:** 52
- **Hedef Değişken:** TedaviSuresi (Seans)
- **Eksik Değer:** 0

- **Veri Tipleri:** float64 (45), int64 (4), int32 (4)

Train/Test Split Önerisi

- **Train Set:** 1,788 gözlem (%80)
- **Test Set:** 447 gözlem (%20)

🎯 Önemli Bulgular ve İçgörüler

1. Hasta Profili

- **Yaş Dağılımı:** 2-92 yaş arası, ortalama 47.3
- **Cinsiyet:** Kadın hastalar daha fazla (%58.1)
- **Uyruk:** Çoğunluk Türkiye (%97.2)

2. Tedavi Özellikleri

- **En Yaygın Tedavi Süresi:** 11-20 seans (%84.4)
- **En Yaygın Bölüm:** Fiziksel Tıp Ve Rehabilitasyon (%91.6)
- **En Yaygın Alerji:** Yok (%42.2)

3. Veri Kalitesi

- **Tekrarlanan Kayıtlar:** 392 hasta birden fazla kayda sahip
- **Eksik Değerler:** Başarıyla dolduruldu
- **Aykırı Değerler:** Tespit edildi ve analiz edildi

🔧 Teknik Detaylar

Kullanılan Kütüphaneler

import pandas **as** pd

import numpy **as** np

import seaborn **as** sns

import matplotlib.pyplot **as** plt

import missingno **as** msno

from sklearn.preprocessing **import** OneHotEncoder, LabelEncoder, StandardScaler

import re

import unicodedata

import warnings

📋 Sonuçlar ve Öneriler

Başarılı Tamamlanan Görevler

✅ **EDA:** Kapsamlı keşifsel veri analizi

✅ **Eksik Değer Doldurma:** 2,706 eksik değer dolduruldu

✅ **Veri Temizleme:** Tutarsızlıklar giderildi

✅ **Encoding:** Kategorik değişkenler encode edildi

✅ **Normalizasyon:** Sayısal değişkenler normalize edildi

✅ **Model-Ready Veri:** 52 feature ile hazır veri seti

Gelecek Adımlar

1. **Model Geliştirme:** Regression modelleri test edilebilir
2. **Feature Selection:** Önemli feature'lar seçilebilir
3. **Cross-Validation:** Model performansı değerlendirilebilir
4. **Hyperparameter Tuning:** Model parametreleri optimize edilebilir

Önerilen Modeller

- **Linear Regression:** Baseline model
- **Random Forest:** Feature importance analizi
- **XGBoost:** Yüksek performans
- **Neural Networks:** Karmaşık ilişkiler