RAPOR

EDA ve Preprocessing Raporu

1.Veri Seti Bilgisi

- Satır/sütun sayısı: 2235 x 13
- Sayısal değişkenler: Yas, TedaviSuresi, UygulamaSuresi
- Kategorik değişkenler: Cinsiyet, KanGrubu, Uyruk, Bolum, Alerji, Tanilar, TedaviAdi, UygulamaYerleri

2.Eksik Değer Analizi

- Cinsiyet: %7.5 eksik
- KanGrubu: %30 eksik
- KronikHastalik: %27 eksik
- Alerji: %42 eksik
- Tanilar: %3.3 eksik
- UygulamaYerleri: %9.8 eksik

3. Aykırı Değerler

- Yas: 41 aykırı değer → Winsorize edildi
- TedaviSuresi: 565 aykırı değer → clip ile sınırlandırıldı
- UygulamaSuresi: 12 aykırı değer → clip ile sınırlandırıldı

4. Ön İşleme Adımları

- TedaviSuresi ve UygulamaSuresi sayısala çevrildi
- Eksik değerler uygun stratejilerle doldurulduu
- Metin verilerinden:
  - KronikHastalik_sayisi
  - Alerji_sayisi
  - Tanilar_sayisi
- NLP Feature Engineering: Tanılar ve KronikHastalık kelime sıklığı çıkarıldı
- Kategorik değişkenler OneHotEncode edildi
- Sayısal değişkenler StandardScaler ile standartlaştırıldı

5.EDA ve Görselleştirmeler

- Sayısal değişken histogram ve boxplotları
- Korelasyon heatmap: TedaviSuresi ile en ilişkili değişkenler
- Encode edilmiş kategorik değişkenlerin frekans grafikleri
- NLP özetleri: En sık geçen 10 tanı ve kronik hastalık kelimesi
- Tedavi Süresi ilişkili görselleştirmeler:
  - Cinsiyet vs TedaviSuresi
  - Yaş vs TedaviSuresi

6.Sonuç

- Veri temiz ve modele hazır hale getirildi
- NLP tabanlı özellikler ile metin verisi sayısala dönüştürüldü
- Görselleştirmeler ile TedaviSuresi ilişkili değişkenler belirlendi