Rapor: İlaç Yan Etkileri Veri Seti Analizi ve Modelleme

1. Giriş

Bu proje kapsamında ilaç yan etkileri ile ilgili bir veri seti üzerinde analiz ve modelleme yaptım. Amaç, veri seti üzerinde Keşifsel Veri Analizi (EDA) gerçekleştirip, veriyi modellemeye hazır hale getirmek ve farklı makine öğrenmesi algoritmalarını uygulayarak en iyi modeli seçmekti.

2. Veri Setinin İncelenmesi

Veri seti, farklı ilaçların kullanıcılar üzerindeki yan etkilerini ve demografik bilgileri içeriyor. İlk adımda, veri setinin genel yapısı inceledim, sayısal ve kategorik değişkenler belirledim. Verinin boyutu ve içerdiği eksik değerler gibi detaylar analiz ettim.

- *Sayısal Değişkenler:* Boy, kilo gibi ölçülebilir değişkenler.
- *Kategorik Değişkenler:* Cinsiyet, il, ilaç adı gibi sınıflandırılabilir değişkenler.

3. Keşifsel Veri Analizi (EDA)

EDA sürecinde verinin yapısı hakkında daha fazla bilgi edinmek için çeşitli görselleştirme teknikleri kullandım. Seaborn ve Matplotlib kütüphaneleri ile sayısal ve kategorik değişkenlerin dağılımları inceledim ve görselleştirmeleri yaptım.

- *Dağılım Grafikleri:* Sayısal değişkenlerin histogramları çizilerek veri dağılımları gözlemlendi..
- *Kategorik Verilerin Sayımları:* Özellikle cinsiyet, il ve ilaç adı değişkenleri için count plot'lar oluşturuldu.
- *Korelasyon Analizi:* Sayısal değişkenler arasındaki ilişkiler ısı haritası (heatmap) kullanılarak gösterildi.

4. Veri Ön İşleme

Veri setinde eksik değerler olduğu gözlemlendi. Bu eksik değerler, uygun imputation yöntemleri kullanılarak doldurulduu:

- *Kilo ve Boy:* Medyan değerlerle doldurulduu.
- *Cinsiyet:* Mod değeri kullanılarak eksik değerler tamamlandı.

Kategorik değişkenler LabelEncoder kullanılarak sayısal değerlere dönüştürdümve modelleme aşamasına geçtim. Sayısal değişkenler ise MinMaxScaler ile ölçeklendirdim.

5. Modelleme ve Değerlendirme

Veri, %75 eğitim ve %25 test olarak ikiye ayırdım. Aşağıda belirtilen makine öğrenmesi modellerini uyguladım ve sonuçları karşılaştırdım:

- *Logistic Regression*
- *Random Forest*
- *Decision Tree*
- *K-Nearest Neighbors (KNN)*
- *Naive Bayes*

Modellerin başarıları accuracy score ve classification report kullanılarak ölçtüm.

6. Hyperparametre Optimizasyonu

En iyi model performansını elde etmek için Decision Tree algoritmasında RandomizedSearchCV ile hiperparametre optimizasyonu gerçekleştirdim. En iyi parametreler seçilerek model yeniden eğitilmiş ve sonuçları değerlendirdim.

- *En İyi Parametreler:*
  - max_depth: {en iyi değer}
  - min_samples_split: {en iyi değer}
  - criterion: {en iyi değer}

7. Sonuçlar

Modellerin doğruluk oranları karşılaştırdığımda, en yüksek başarıyı sağlayan model RandomizedSearchCV ile optimize edilen *Decision Tree* modeli oldu. ,,

8. Sonuç ve Gelecek Çalışmalar

Proje kapsamında gerçekleştirilen analizler ve modeller başarılı sonuçlar verdi. En iyi model olarak Decision Tree seçtim ve bu model ile ilaç yan etkilerinin tahmininde iyi sonuçlar almayı başardım. Gelecekte, model performansını daha da artırmak için daha fazla veri kullanımı, farklı algoritmaların denenmesi ve model tuning işlemlerinin daha kapsamlı yapılması ile .modle daha iyi sonuç verebilir.

9. Kaynaklar

Proje boyunca kullanılan kütüphaneler: Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn.