Pusula Data Science Intern Case – Bulgu Özeti

Hazırlayan: Buse Melike Dalbaş

E-posta: bmelikedalbas@gmail.com

Veri Seti:

2235 tane veri, 13 kolondan oluşuyor. Hedef değişken TedaviSuresi 'dir.

Hasta No, Yaş, Cinsiyet, Kan Grubu, Uyruk, Kronik Hastalık, Bölüm, Alerji, Tanılar, Tedavi Adı, Tedavi Süresi, Uygulama Yerleri, Uygulama Süresi kolonlarından oluşmuştur.

Kullanılan Kütüphaneler:

Pandas, numpy, matplotlib, seaborn

Scikit- learn:

- LabelEncoder, StandardScaler: Kategorik değişkenlerin dönüştürülmesi ve sayısal değişkenlerin ölçeklenmesi için kullanıldı.
- Modelleme (train_test_split): Veriyi eğitim ve test setlerine ayırmak için kullanıldı.
- Lineer Regresyon, Ridge, Lasso: Lineer tabanlı regresyon modellerini kurmak için kullanıldı.
- Svm (SVR): Destek vektör makinesi algoritması ile modelleme için kullanıldı.
- Rastgele Orman: Rastgele Orman algoritması ile tahmin modeli kurmak için kullanıldı.
- Metrikler (MAE, MSE, R^2): Modellerin performansını değerlendirmek için kullanıldı.
- TfidfVectorizer : Metin içerikli kolonların sayısal vektörlere dönüştürülmesi için kullanıldı.

Keşifsel Veri Analizi (EDA):

Histogram görselinde, veri setindeki Yaş, Tedavi Süresi ve Uygulama Süresi değişkenleri incelenmiştir. Yaş geniş bir aralığa yayılmış ve en çok 40-50 yaş aralığı ön plana çıkmıştır. Tedavi Süresi sağa çarpık bir görsel vardır ve 15-20 seans arasında yoğunlaşmıştır. Uygulama Süresi ise 15-20 dakika arasında yoğunlaşmıştır. Sağa çarpık bir yapıdadır.

Tedavi Süresi boxplot görseli yardımıyla incelenmiştir. Medyan değer yaklaşık 15 seans civarındadır. 30 seans ve üzeri veride çok sayıda aykırı değer bulunmaktadır.

Kronik hastalık ve Tedavi süresi arasındaki ilişki boxplot ile incelenmiştir. Farklı kronik hastalık gruplarına sahip hastalarda tedavi süresinin 15 seans civarında yoğunlaştığı gözlenmektedir.

Ancak bazı kronik hastalık türlerinde 25 seans üzeri tedavi görülmekte, bu da belirli hastalık gruplarının tedavi süresini uzatabileceğini düşündürmektedir.

En çok uygulanan ilk 10 tedavi için Tedavi Süresi boxplot görseliyle incelenmiştir. Genel olara tedavi süreleri 15 seans da yoğunlaşmıştır. Dorsalji-Boyun+Trapez gibi tedavilerde daha farklı boxplot görülmüştür, daha geniş bir yayılıma sahiptir ve aykırı değerler bulunmaktadır.

Veri Önişleme :

- TedaviSuresi kolonunda "seans" kelimesi kaldırıldı ve sayısal formata dönüştürüldü. UygulamaSuresi kolonunda ise "dakika" kelimesi kaldırıldı ve sayısal formata dönüştürüldü.
- Cinsiyet, UygulamaYerleri, Tanılar, Bölüm adlı kolonlar mod yöntemiyle dolduruldu. KronikHastalık ve Alerji kolonlarının eksiklikleri yok yöntemiyle dolduruldu. KanGrubu ise unknown yöntemiyle dolduruldu.
- Cinsiyet, Uyruk, KanGrubu Label Encoding ile sayısal formata dönüştürüldü. Tanılar ve TedaviAdı One-Hot Encoding ile dönüştürüldü (çok kategorili ve sırasız sütunlar). Aynı zamanda metin içerikli kolonlar ise TF-IDF vektörleştirme ile sayısallaştırıldı.
- Yaş, TedaviSüresi, UygulamaSüresi sayısal kolonlarına StandardScaler uygulandı. Böylece tüm sayısal değişkenler aynı ölçeğe getirilerek modelleme için hazır hale getirildi.

Modelleme:

Veri seti üzerinde farklı modeller denenerek MAE, RMSE, R^2 değerleri karşılaştırılmış ve en iyi model bulunmuştur. Denenen modeller Lineer regresyon, Ridge, Lasso, XGBOOST, Rastgele Orman, Destek Vektör Makinesi.

En iyi başarıyı Rastgele Orman modeli göstemiştir (R^2= 0.86).