Fatma Tuğçe Mukul - mukultugce@gmail.com

FİZİKSEL TIP VE REHABİLİTASYON VERİLERİNİN KEŞİFSEL VERİ ANALİZİ (EDA)

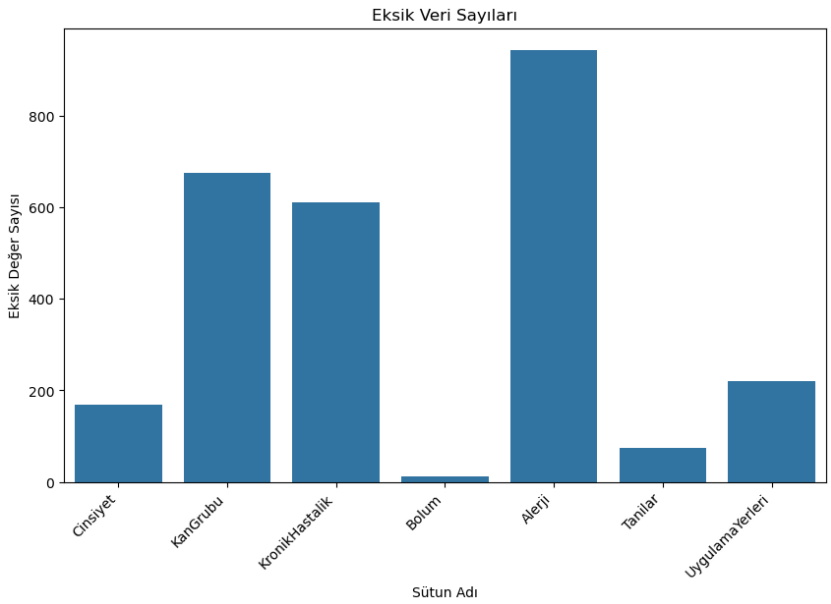
Bu görev kapsamında, fiziksel tıp ve rehabilitasyon verilerini içeren bir veri seti üzerinde EDA (Exploratory Data Analysis) çalışması gerçekleştirildi. Veri seti, 2235 satır ve 13 sütundan oluşmakta olup, temel amacı hasta bilgilerini ve tedavi süreçlerini anlamak ve veri ön işleme adımlarını belirlemektir.

1. VERİ SETİNİN İNCELENMESİ

Çalışmanın ilk adımında veri seti genel hatlarıyla incelenmiştir. info() fonksiyonu ile sütun tipleri ve eksik değer durumları kontrol edilmiştir.

Veri setinde 2 adet sayısal (HastaNo, Yas), 11 adet kategorik (object tipinde) değişken bulunmaktadır.

Eksik değer analizi sonucunda özellikle Cinsiyet (169), Kan Grubu (675), Kronik Hastalık (611), Alerji (944), Tanılar (75) ve Uygulama Yerleri (221) sütunlarında boş değerler olduğu görülmüştür.



TedaviSuresi ve UygulamaSuresi sütunlarının aslında sayısal değerler içermesi gerekirken object tipinde olduğu fark edilmiştir. Bunun sebebi, süre bilgilerine ek olarak “Dakika” ve “Seans” string ifadelerinin yer almasıdır. Bu durum, veri üzerinde matematiksel işlemler yapılmasını engellemektedir.

Bu nedenle söz konusu sütunlarda veri temizleme işlemi uygulanmış, süre bilgileri içerisinden metinsel kısımlar çıkarılarak yalnızca sayısal değerler bırakılmıştır. Ardından bu sütunlar integer (int) tipine dönüştürülmüştür. Böylece ilerleyen aşamalarda bu veriler üzerinden doğru istatistiksel analizler yapılabilmesi sağlanmıştır.

2. TANIMLAYICI İSTATİSTİKLER VE DEĞİŞKEN ANALİZİ

describe() çıktısı incelenerek sayısal değişkenlerin dağılımı analiz edilmiştir:

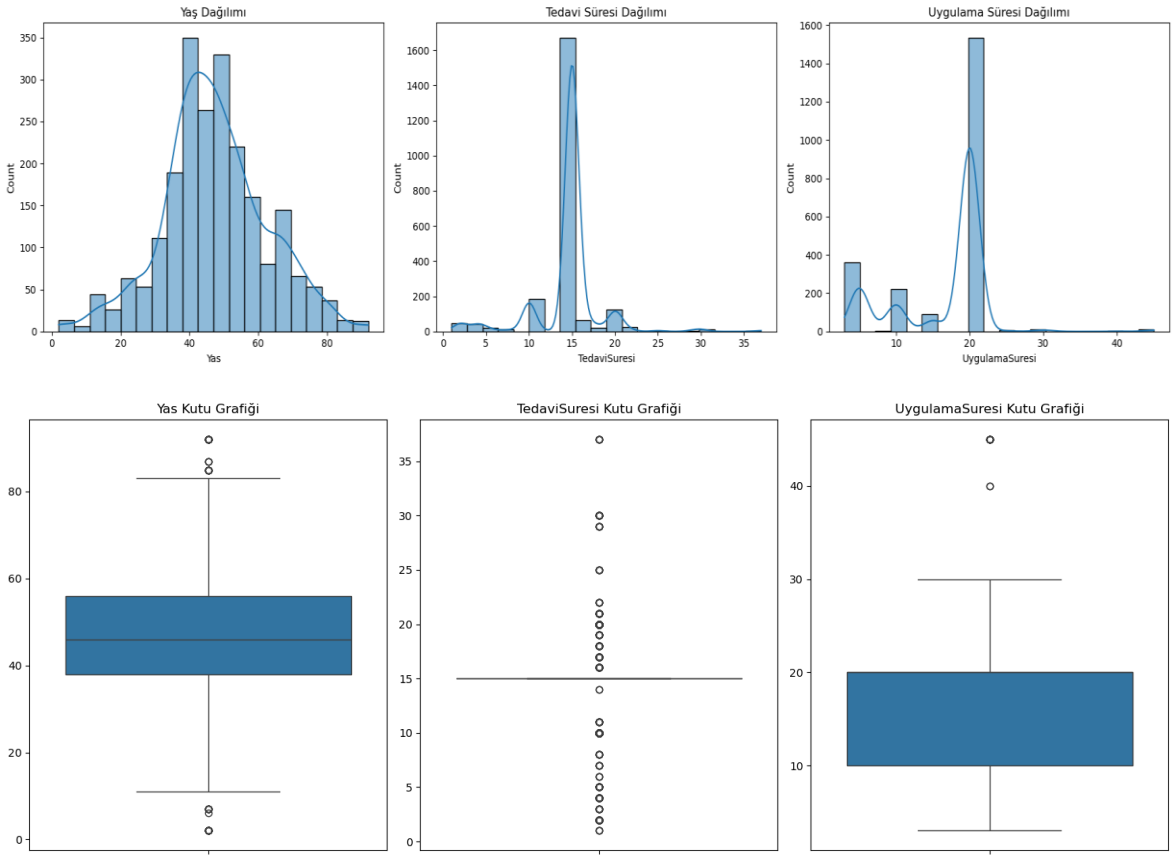
HastaNo: Kimlik numarası niteliğinde olup istatistiksel açıdan anlamlı değildir, analizlerde kullanılmayacaktır.

Yaş: 2–92 arasında değişmektedir. Ortalama 47, medyan 46'dır. Ortalama ve medyanın yakın olması dağılımın simetrik olduğunu göstermektedir.

Tedavi Süresi: Ortalama 14,6, medyan 15 seanstır. Çoğunluk 15 seans sürmüş, minimum 1, maksimum 37 seanstır. Bu durum bazı aykırı tedavi protokollerine işaret etmektedir.

Uygulama Süresi: Ortalama 16,6, medyan 20 dakikadır. En yaygın süreler 10 ve 20 dakikadır. Minimum 3, maksimum 45 dakika olup farklı uygulama çeşitliliğini göstermektedir.

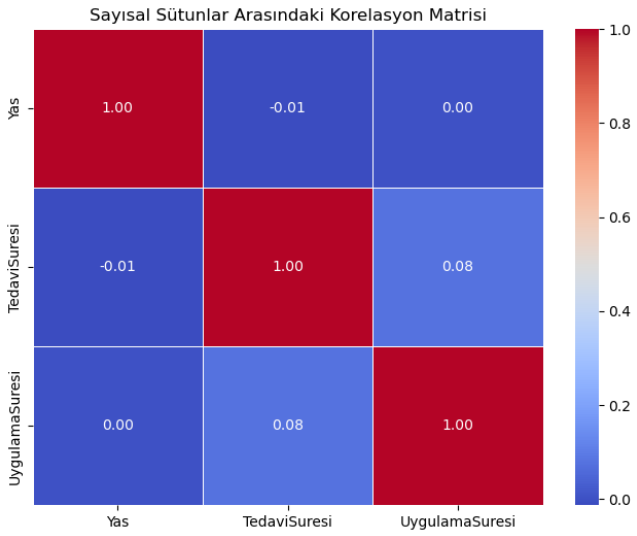
Sayısal değişkenlerin dağılımları histplot ve boxplot grafiklerle incelenmiştir:



Yaş: Çoğunlukla 40–60 aralığında yoğunlaşmaktadır. Boxplot simetrik bir dağılım göstermiş, 10 yaş altı ve 80 yaş üstü birkaç aykırı değer tespit edilmiştir. Bunlar pediatri ve geriatri hastalarını temsil ettiği için doğal değerler olarak değerlendirilmiş ve veri setinde tutulmuştur.

Tedavi Süresi: Veriler büyük ölçüde 15 seans etrafında toplanmıştır. 1–37 seans arasındaki farklılıklar boxplot'ta aykırı değer gibi görünse de aslında farklı tedavi protokollerini yansıtmaktadır. Bu nedenle silinmemiştir.

Uygulama Süresi: En yaygın süre 20 dakika olup 10–20 dakika aralığında yoğunluk gözlenmiştir. 5 dakikanın altındaki ve 30 dakikanın üzerindeki değerler aykırı görünse de özel uygulamaları temsil ettiği için korunmuştur.



Korelasyon Analizi:

Isı haritası, sayısal sütunlar arasında güçlü bir lineer ilişki olmadığını göstermiştir:

Yaş & Tedavi Süresi: -0.01

Yaş & Uygulama Süresi: 0.00

Tedavi Süresi & Uygulama Süresi: 0.08

Sayısal değişkenlerde gözlenen aykırı değerler anlamlı klinik farklılıklardan kaynaklanmakta olup veri setinde bırakılmıştır. Ayrıca değişkenler arasında güçlü bir korelasyon bulunmamaktadır.

3. VERİ TEMİZLEME

Eksik değerler ve tutarsızlıklar incelenmiş, uygun yöntemlerle düzeltilmiştir:

Tekrarlayan HastaNo: 2235 kayıttan 2223'ünde tekrar olduğu görüldü. Eksik değerlerin aynı HastaNo'ya sahip tüm kayıtlarda tutarlı olduğu gözlemlendi.

Hasta bazlı doldurma: Bazı sütunlar aynı hastaya ait mevcut kayıtlar kullanılarak dolduruldu. Bu işlem sonucunda sonucunda 31 KanGrubu, 5 KronikHastalik, 24 Cinsiyet bilgisi dolduruldu.

Bölüm (Bolum): 11 eksik değerin tamamının tek bir hastaya ait olduğu belirlendi (145157) . Bu hastanın diğer tedavi bilgileri dikkate alınarak eksik değerler "Fiziksel Tıp ve Rehabilitasyon, Solunum Merkezi" olarak dolduruldu.

Kan Grubu: Eksik değerler KNNImputer ile tahmin edilerek dolduruldu. Bu yöntem, hastaların benzer özellikleri (yaş, cinsiyet) üzerinden en uygun kan grubunu atamaktadır.

Tanılar & Uygulama Yerleri: Eksik değerler mode yöntemiyle dolduruldu.

Alerji & Kronik Hastalık & Cinsiyet: Eksik kayıtlar "Bilinmiyor" etiketiyle işaretlendi.

Bu işlemler sonucunda veri setinde eksik değer kalmamış, analiz ve modelleme için daha tutarlı bir yapı elde edilmiştir.

Eksik değer doldurma işlemlerinin ardından, metin tabanlı sütunlarda kapsamlı temizlik ve standartlaştırma işlemleri yapılmıştır:

Tanılar sütunu, aynı hastalığın farklı yazılışları ve birden fazla tanının aynı satırda yer alması nedeniyle dağınık bir yapıdaydı. Tüm metinler küçük harfe çevrilip boşluk ve noktalama işaretleri temizlendi. Böylece veri modelleme için daha anlaşılır ve kullanılabilir hâle geldi.

KronikHastalik sütunu, bir satırda birden fazla hastalık ve eksik değerler içeriyordu. Metinler standartlaştırıldı. Bu sayede karmaşık yapı daha düzenli hâle getirildi.

TedaviAdi sütunu yazım hataları, kısaltmalar ve tekrar eden ifadeler nedeniyle tutarsızdı. Harfler küçültüldü, gereksiz karakterler temizlendi, sık görülen hatalar sözlükle düzeltildi ve FuzzyWuzzy ile benzer yazımlar aynı kategoriye toplandı. Tüm tedavi adları standart ve tutarlı hâle getirildi.

Alerji sütunu serbest metin formatında olup birden fazla alerjen ve hatalı girişler içeriyordu. Metinler küçültüldü, boşluklar temizlendi ve yazım hataları düzeltildi. Yinelenen değerler çıkarıldı ve alerjenler alfabetik olarak birleştirilerek standart bir formata dönüştürüldü.

4. KATEGORİK DEĞİŞKENLERİN KODLANMASI

Veri setindeki kategorik değişkenler, modelin anlayabileceği sayısal formata dönüştürülmüştür:

Cinsiyet: One-Hot Encoding ile Cinsiyet_Kadin, Cinsiyet_Erkek ve Cinsiyet_Bilinmiyor olmak üzere üç sütuna dönüştürüldü. Bu sayede cinsiyet bilgisi modelde sayısal olarak ve ayrı sütunlar halinde temsil edilebildi.

Alerji: Bu sütun, birden fazla alerjenin aynı hücrede yer alması ve farklı yazım biçimleri içermesi nedeniyle doğrudan model için kullanılamaz durumdaydı. Bu sorunu çözmek için önce her hücredeki alerjenler virgül ile ayrılarak liste hâline getirildi. Ardından MultiLabelBinarizer kullanılarak her benzersiz alerjen için ayrı sütunlar oluşturuldu. Bu yeni sütunlarda alerjenin varlığı 1, yokluğu 0 ile gösterildi. Böylece, örneğin “penisilin” ve “voltaren” alerjisi olan bir hastada ilgili sütunlar 1 değerini alırken, diğer sütunlar 0 oldu. Son olarak orijinal Alerji sütunu kaldırıldı ve veri seti, modelin işleyebileceği sayısal formata dönüştürüldü.

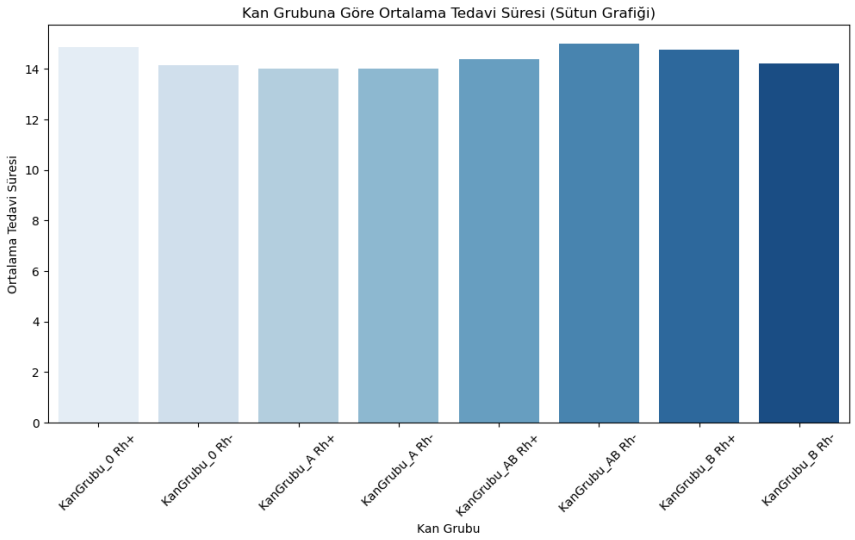
One-Hot Encoding işlemleri sonrası veri seti hem kategorik bilgileri koruyacak hem de makine öğrenmesi algoritmalarında kullanılabilir bir yapıya dönüştü.

5. GÖRSELLEŞTİRME VE ANALİZ

Veri setinde sayısal ve kategorik değişkenlerin etkileşimini anlamak için çeşitli görselleştirmeler yapılmıştır.

5.1 Kan Grubuna Göre Ortalama Tedavi Süresi

One-hot encoding ile kodlanan kan grubu sütunları uzun formata çevrildi ve her kan grubu için ortalama tedavi süresi hesaplandı. Sütun grafiği ile görselleştirildi.



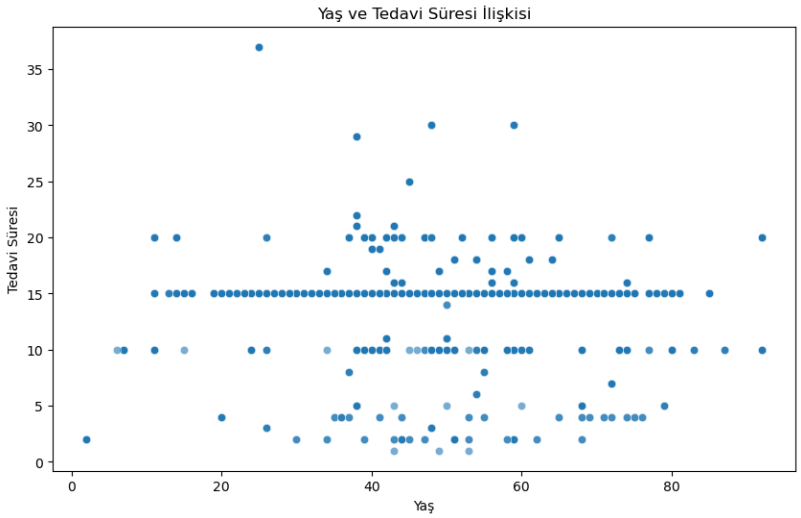
Tüm kan gruplarının ortalama tedavi süresi 14–15 seans arasında, farklar çok küçük.

AB Rh+ ve bilinmiyor grupları biraz daha yüksek görünse de, hata payları dikkate alındığında anlamlı bir fark bulunmamaktadır.

Kan grubu, tedavi süresi üzerinde belirleyici bir faktör olmadığı için modelde sayısal formata dönüştürülmemiştir. Bu bulgu literatürdeki genel gözlemlerle de uyumludur.

5.2 Yaş ve Tedavi Süresi İlişkisi

Scatter plot ile yaş ve tedavi süresi arasındaki ilişki görselleştirildi.

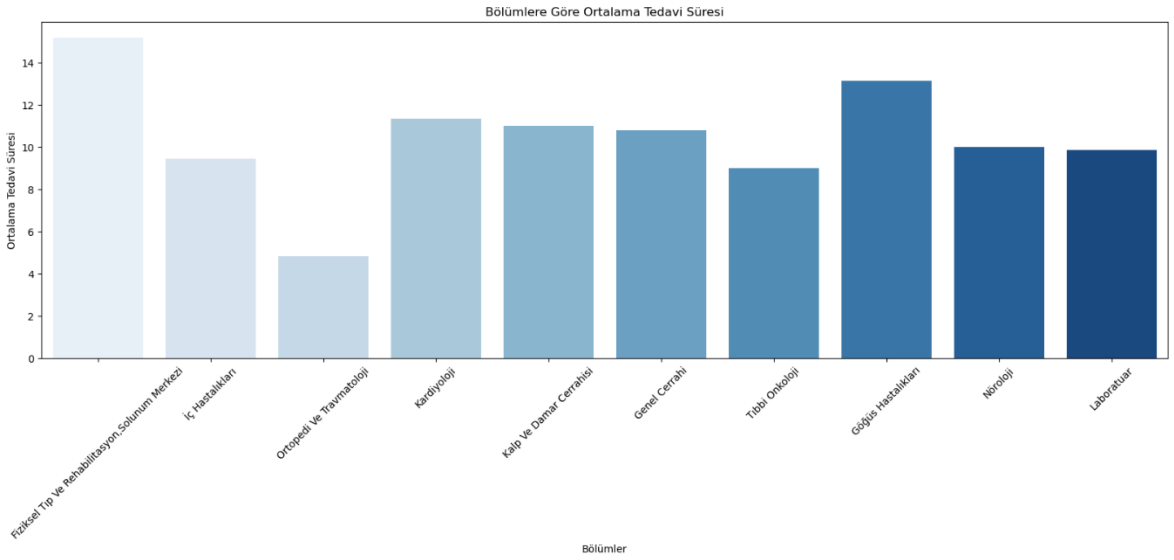


İleri yaşlarda tedavi süresinde hafif bir artış gözlemlense de genel olarak çoğu tedavi süresi 15 seans civarındadır.

Yaşın tedavi süresi üzerinde belirgin bir etkisi yoktur.

5.3 Bölümlere Göre Ortalama Tedavi Süresi

Bölüm sütunu ile tedavi süresi arasındaki ilişki bar plot ile görselleştirildi.



Tedavi süresi en yüksek bölüm Fiziksel Tıp ve Rehabilitasyon iken, en düşük bölüm Ortopedi ve Travmatoloji'dir.

Bölüm bazlı farklılıklar, tedavi türü ve protokol farklılıklarından kaynaklanıyor olabilir