Proje Dokümantasyonu: Fiziksel Tıp ve Rehabilitasyon Veri Seti Analizi

1. Proje Amacı

Bu projenin temel amacı, 2235 gözlem ve 13 özellik içeren fiziksel tıp ve rehabilitasyon veri setini, potansiyel bir makine öğrenmesi modeli için hazır hale getirmektir. Çalışmanın hedef değişkeni TedaviSuresi olarak belirlenmiştir. Bu süreçte, veri temizliği, tutarlılığının sağlanması ve keşifçi veri analizi (EDA) adımları uygulanmıştır.

2. Veri Seti ve İlk İnceleme

Proje için sağlanan veri seti, hasta kimlik numarası (`HastaNo`), yaş, cinsiyet, kan grubu, uyruk gibi hasta bilgilerini; kronik hastalıklar, alerjiler, tanılar gibi tıbbi geçmişi ve uygulanan tedavi adı, tedavi süresi, uygulama yerleri gibi tedavi detaylarını içermektedir.

İlk incelemelerde, veri setinde şu ana problemler tespit edilmiştir:

- TedaviSuresi ve UygulamaSuresi sütunları, içerdikleri metin ifadelerinden dolayı sayısal bir veri tipi olarak okunmamaktadır.
- Cinsiyet, KanGrubu, KronikHastalik, Bolum, Alerji, Tanilar ve UygulamaYerleri sütunlarında eksik değerler (NaN) bulunmaktadır.
- KronikHastalik, Alerji, Tanilar ve UygulamaYerleri gibi sütunlar, virgülle ayrılmış birden fazla değer içermektedir.

3. Uygulanan Veri Ön İşleme Adımları

Veri setini modellemeye hazır hale getirmek için aşağıdaki adımlar uygulanmıştır:

a. Sayısal Sütunların Temizlenmesi `TedaviSuresi` ve `UygulamaSuresi` sütunlarındaki " Seans" ve " Dakika" metin ifadeleri kaldırılmıştır. Bu işlem sonrasında her iki sütun da `int64` veri tipine dönüştürülerek sayısal analizler için uygun hale getirilmiştir.

b. Eksik Değerlerin Doldurulması Eksik değer içeren tüm sütunlar, veri kaybını önlemek amacıyla "Bilinmiyor" etiketiyle doldurulmuştur. Bu sayede her bir eksik değer, model tarafından ayrı bir kategori olarak değerlendirilmeye hazır hale getirilmiştir.

c. Çoklu Değer İçeren Sütunların Düzenlenmesi Virgülle ayrılmış birden fazla değer içeren sütunlar (`KronikHastalik`, `Alerji`, `Tanilar`, `UygulamaYerleri`), `str.get_dummies()` metodu kullanılarak her bir benzersiz değer için yeni bir ikili (binary) sütun oluşturulmuştur. Bu yöntem, orijinal veri setinde satır tekrarı yaşanmasının önüne geçerek daha tutarlı bir yapı sağlamıştır. Bu işlem tamamlandıktan sonra orijinal sütunlar veri setinden çıkarılmıştır.

d. Sütun Adlarının Düzenlenmesi Yeni oluşturulan sütunlar da dahil olmak üzere tüm sütun adlarındaki boşluklar ve özel karakterler temizlenerek, modelleme süreçlerinde sorun yaratmayacak, standart bir isimlendirme formatına getirilmiştir.

4. Keşifçi Veri Analizi (EDA) Bulguları

Veri temizleme adımları sonrasında, veri setinin yapısını anlamak ve temel özelliklerini keşfetmek amacıyla görsel analizler yapılmıştır. Bu analizler, veri setindeki kalıpları ve değişkenler arasındaki ilişkileri ortaya koymuştur.

- **Sayısal Değişkenlerin Dağılımı:** Hastaların yaş, tedavi süresi ve uygulama süresi dağılımları histogramlarla incelenmiştir.
    - **Yaş dağılımında**, en yüksek hasta yoğunluğunun 30-60 yaş aralığında olduğu görülmüştür.
    - **Tedavi Süresi** (`TedaviSuresi`) incelendiğinde, tedavilerin büyük bir kısmının 15 seans civarında yoğunlaştığı tespit edilmiştir.
    - **Uygulama Süresi** (`UygulamaSuresi`) ise en çok 20 dakika civarında birikim göstermektedir, ancak 5 dakikalık daha kısa süreli uygulamaların da önemli bir yer tuttuğu gözlemlenmiştir.
- **Kategorik ve Hedef Değişken İlişkisi:** Cinsiyet ve kan grubu gibi kategorik özelliklerin, hedef değişken olan `TedaviSuresi` ile ilişkisi kutu grafikleriyle analiz edilmiştir.
    - **Cinsiyet ve Tedavi Süresi** arasında ortalama değerler açısından belirgin bir fark bulunmamaktadır. Hem erkek hem de kadın hastaların tedavi süreleri benzer bir dağılım sergilemektedir.
    - **Kan Grubu ve Tedavi Süresi** ilişkisinde de benzer bir durum söz konusudur. Farklı kan gruplarındaki hastaların tedavi süreleri arasında istatistiksel olarak anlamlı bir farklılık gözlemlenmemiştir. Hatta "Bilinmiyor" olarak etiketlenen hastaların tedavi süresi dağılımı bile diğer kan gruplarına benzemektedir.

5. Sonuç

Bu projenin sonunda, fiziksel tıp ve rehabilitasyon veri seti, makine öğrenmesi modeli oluşturulması için gerekli olan tüm temizleme ve ön işleme adımlarından geçirilmiştir. Veri setindeki eksik değerler giderilmiş, veri tipleri standartlaştırılmış ve çok değerli kategorik bilgiler, her bir özelliği ayrı bir sütun olarak temsil eden bir formata dönüştürülmüştür. Elde edilen temizlenmiş veri seti, tahmin modellerinin eğitimi için hazır durumdadır.