Merhaba,

Veri setini ilk incelediğimde ilk gözüme çarpan oldukça fazla kategorik değişken olması ve cinsiyet değişkeninin çok fazla eksik değere sahip olmasıydı. Veri setini python ile öncelikle tanımaya çalıştım. Her özelliğin dağılımına, ne tür bilgiler içerdiğine, bunlardan nasıl bilgiler çıkarılabileceğini düşündüm. Özellikle zaman serisi için kullanılabilecek 3 tane "datetime" tipinde veri vardı. Bunlar ilaç başlangıç ve bitiş tarihleri ve yan etki tarihleriydi. Bunlarla ilgili mevsimsellik ve trend analizleri gerçekleştirilmeli diye düşündüm. Fakat bizden istenen sadece Keşifsel Veri Analizi ve Veri Ön İşleme adımlarıydı. Ayrıca verilerde eksikler vardı, zaman verilerinin eksik verileri nasıl ele alınmalı diye yaptığım araştırmada farklı bulgulara ulaştım fakat bu tarafta yeterli tecrübem olmadığı için kodları çalıştıramadım. Bende bu aklımda oluşan fikirleri dökümantasyona eklemeyi ve yaptığım çalışmada sadece görselleştirme ile yetinmek zorunda kaldım.

Daha sonra Cinsiyet verilerini nasıl ele almam gerekli diye düşündüm. Eksik verileri çıkarmak çok büyük bir bilgi kaybına sebep olacaktı bu yüzden bu yola sapmayı seçmedim. Veri setinden aralarında cinsiyeti belirlemesine yardımcı olabileceğini düşündüğüm kolomları seçerek bir tahmin algoritması kullandım ve eksik değerleri bunun ile doldurdum.

Boy ve Kilo değişkenlerindeki eksik değerleri ele alırken bunları görmezden gelmeyi düşündüm çünkü çok yüksek sayıda eksik değer yoktu. Fakat boy ve kilo ortalamalarının cinsiyete göre doldurulursa genel eğilimi yansıtabileceğini düşünerek, eksik değerleri ortalama ile doldurum.

Kategorik değişkenlerde kronik hastalıkların ele alındığı birden fazla değişken vardı. Bunların da eksik değerleri vardı. Kronik hastalıkları en çok tekrar eden veri ile, ya da en yakın komşuları ile doldurmayı düşündüm fakat bunların bizi yanıltabileceğini düşündüm. Kronik hastalık gibi önemli bir sağlık verisi çok güvenilir bir şekilde doldurulmalı, ve eğer hata olasılığı yüksek ise bence farklı bir yöntem kullanılmalı. Tahmin yöntemlerini daha önce cinsiyet değişkeninde kullandığım için tekrar kullanmak istemedim. Ayrıca çok fazla kronik hastalık değişkeni vardı (Anne Kronik Hast., Baba Kronik Hast. Vb.). Bu sebeple bu eksik değerlerin analizini boş bıraktım ve amaca yönelik olması gerektiğini çalışmamda bahsettim. Bana en uygun gelen yöntem uygun bir tahmin metodu ile doldurmaktı fakat bunu gerçekleştirmedim. Hatalı veri üretme riskinin bu değişkenlerde tehlikeli olacağını düşündüğüm için.

Diğer kategorik değişkenlerin eksik değerlerini incelerken en sık tekrar eden değişken ile doldurdum çünkü çok fazla eksik değerleri yoktu ve değişkenlerin içeriğinde bir yığılma yoktu.

Nicel verilerin aykırı değerlerini incelediğimde aykırı değerleri IQR metodu ile belirlemeye karar verdim. IQR için bir fonksiyon yazdım ve bu metodu veri setime uygulayarak aykırı değerlerden kurtuldum.

Son olarak değişkenler arasında yüksek korelasyonlu birer ilişki olup olmadığını ısı haritası yardımıyla gösterdim. Bunu gösterebilmek için verileri oldukça düzenlemem gerekti, bazı verileri mesela ilaç başlangıç ve bitiş tarihi verilerini “Tedavi Süresi” olarak yeni veri üretmekte kullandım. Zaten yüksek korelasyonlu bir durum olmadığını değişkenleri en başta tek tek incelerken olmadığını biliyordum. En sonunda da eğer hatalı bir işlem yapmadıysam olmadığı gözüktü.

Değerlendirmeniz ve vaktiniz için şimdiden teşekkür ederim,

Saygılarımla

Emrecan Bilgin