Şerafettin Gazibaba serafettingazibaba7@gmail.com

Pusula Proje Dökümantasyonu

Toplamda task olarak atanan 2 adım bulunmakta.

1. adımda Keşifsel Veri Analizi(EDA) bulunuyor.
2. adımda ise Verinin Modelleme aşaması için hazırlanması bulunuyor.

Analiz aşamasını öznitelik bazlı yapmayı tercih ediyorum.
Uyguladığım adımları dökümantasyon içerisinde paylaşacağım.
Dikkat edilmesi gereken noktalara değineceğim.
İpynb dosyası, veri seti ve dökümantasyon GitHub üzerinden paylaşılmıştır.

A)Verilerin İçe Aktarılması

```
data_path = '../Pusula/side_effect_data 1.xlsx'
data = pd.read_excel(data_path)
```

Dikkat

Bilgisayarınızda veri setinin bulunduğu dizini kontrol edin ve düzeltin.

B)Ham Veri Setinin Analizi

Toplamda 19 kolona sahip veri setine sahibiz. Kullanici_id kolonu ilaç kullanan hastaların birbirinden farklı(Unique) özniteliği.

Yan Etki hedef özniteliğimiz.

Geriye kalan 17 kolon kullanıcıya ait farklı öznitelikler.
Öznitelikler;

1. Cinsiyet : Hastanın cinsiyeti. Bu öznitelik bazı hastaların cinsiyetleri belirtilmemiş. Bu özniteliğin diğer özniteliklere bakılarak eksik verilerinin tamamlanması sağlandı.
2. Doğum Tarihi : Hastanın doğduğu tarihi belirtmekte. Eksik verisi yok.
3. Uyruk : Hastanın uyruğunu belirtiyor. Özniteliğin değerine bakılırsa sadece Türk olan hastaların verileri bulunuyor. Modelleme kısmında fayda sağlamayacaktır.
4. İl : Hastanın ikamet ettiği yer olarak düşünüyorum(Doğum ili de olabilir).
5. İlaç Adi : Hastaya verilen ilacın adı. Araştırmalarım sonucunda bazı ilaçların aktif maddesinin aynı olduğu fakat kullanım şekli farklı olduğu için farklı ilaç olarak bulunuyor. NOT: Daha detaylı analiz yapılıp aktif maddelerinin aynı olan ilaçların yan etkileri de araştırılabilir.
6. İlaç başlangıç tarihi : Hastanın ilaca başladığı tarih.
7. İlaç bitiş tarihi : Hastanın ilacı kullanmayı bıraktığı tarih.
8. Yan etki : Hastada görülen yan etki.
9. Yan etki bildirim tarihi : Hastada görülen yan etkinin bildirildiği tarih.
10. Alerjilerim : Hastaya ait alerjiler. Eksik veriler içermekte verilerin nasıl toplandığı bilgisine sahip değilim.Eksik verilerin hastanın alerjisinin olmadığı varsayılarak ‘Yok’ şeklinde doldurulduu.
11. Kronik Hastalıklarım : Hastaya ait kronik rahatsızlıklar. Eksik veriler hastanın kronik rahatsızlığının bulunmadığı varsayılarak ‘Yok’ şeklinde doldurulduu.
12. Baba Kronik Hastalıkları : Hastanın babasının kronik rahatsızlıkları.
13. Anne Kronik Hastalıkları : Hastanın annesinin kronik rahatsızlıkları.

14. Kız Kardeş Kronik Hastalıkları : Hastanın kız kardeşinin kronik rahatsızlıkları.
15. Erkek Kardeş Kronik Hastalıkları : Hastanın erkek kardeşinin kronik rahatsızlıkları.
16. Kan Grubu : Hastanın kan grubu.
17. Kilo : Hastanın kilosu(kg)
18. Boy : Hastanın boyu(cm)

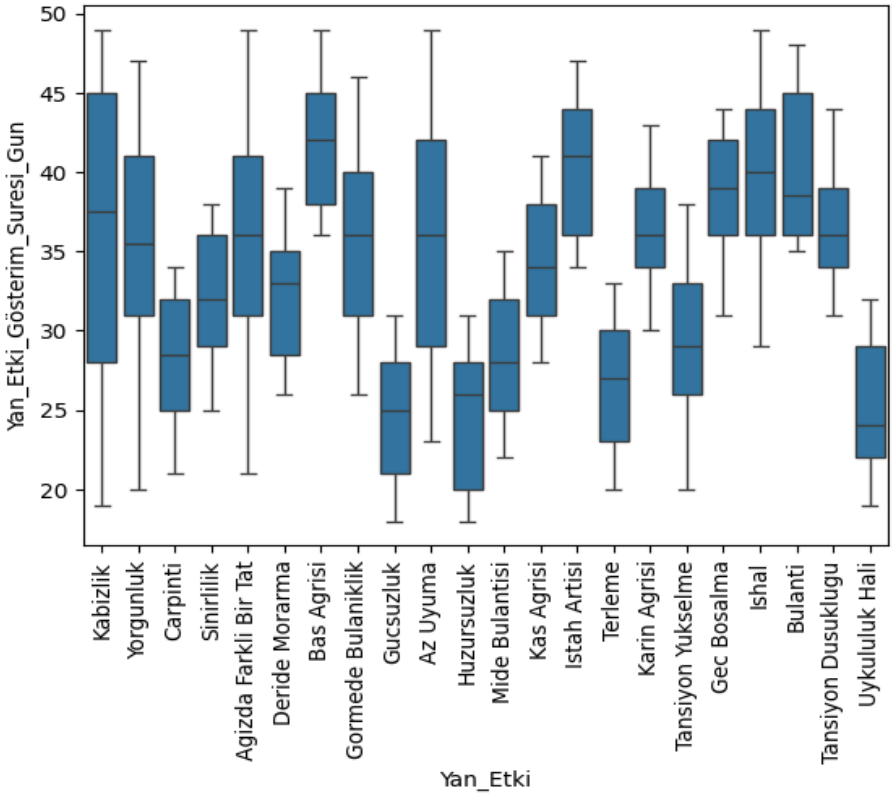
C)Veri setindeki özniteliklere bağlı yeni özniteliklerin türetilmesi;

1. Yaş : Doğum tarihi özniteliği kullanılarak hastanın yaş bilgisine ulaşıldı.
2. İlaç Kullanım Süresi(Gün) : Hastanın ilaca başladığı tarih ile ilacı kullanmayı bıraktığı tarih arasında geçen süre hesaplandı.
3. Yan Etki Gösterim Süresi(Gün) : Hastanın ilaca başladığı süre ile yan etki bildirim tarihi arasında geçen süre hesaplandı.
4. Vke(Vücut Kitle Endeksi) : Hastanın boy ve kilo değerleri kullanılarak vücut kitle endeksi hesaplandı.

   NOT: VKE = Kilo(kg) / Boy(metre)^2

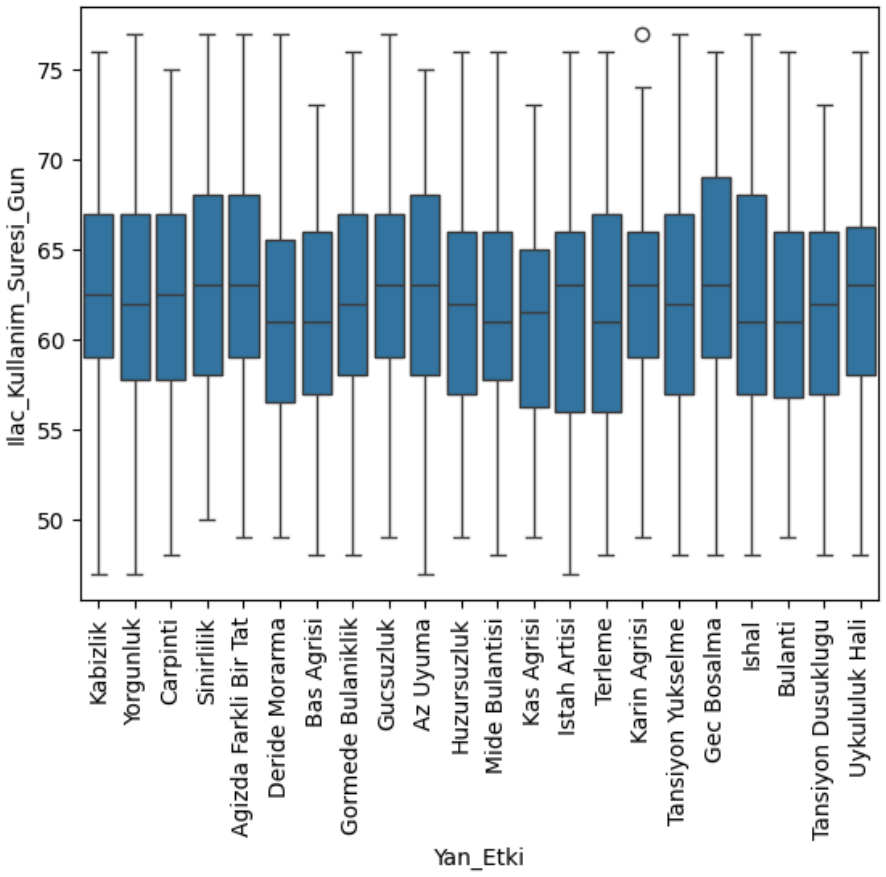
D)Yeni eklenen Özniteliklerle beraber Hedef Özniteliğin İlişkilerinin İncelenmesi;

Bir kaç tane örneği dökümantasyona ekledim.

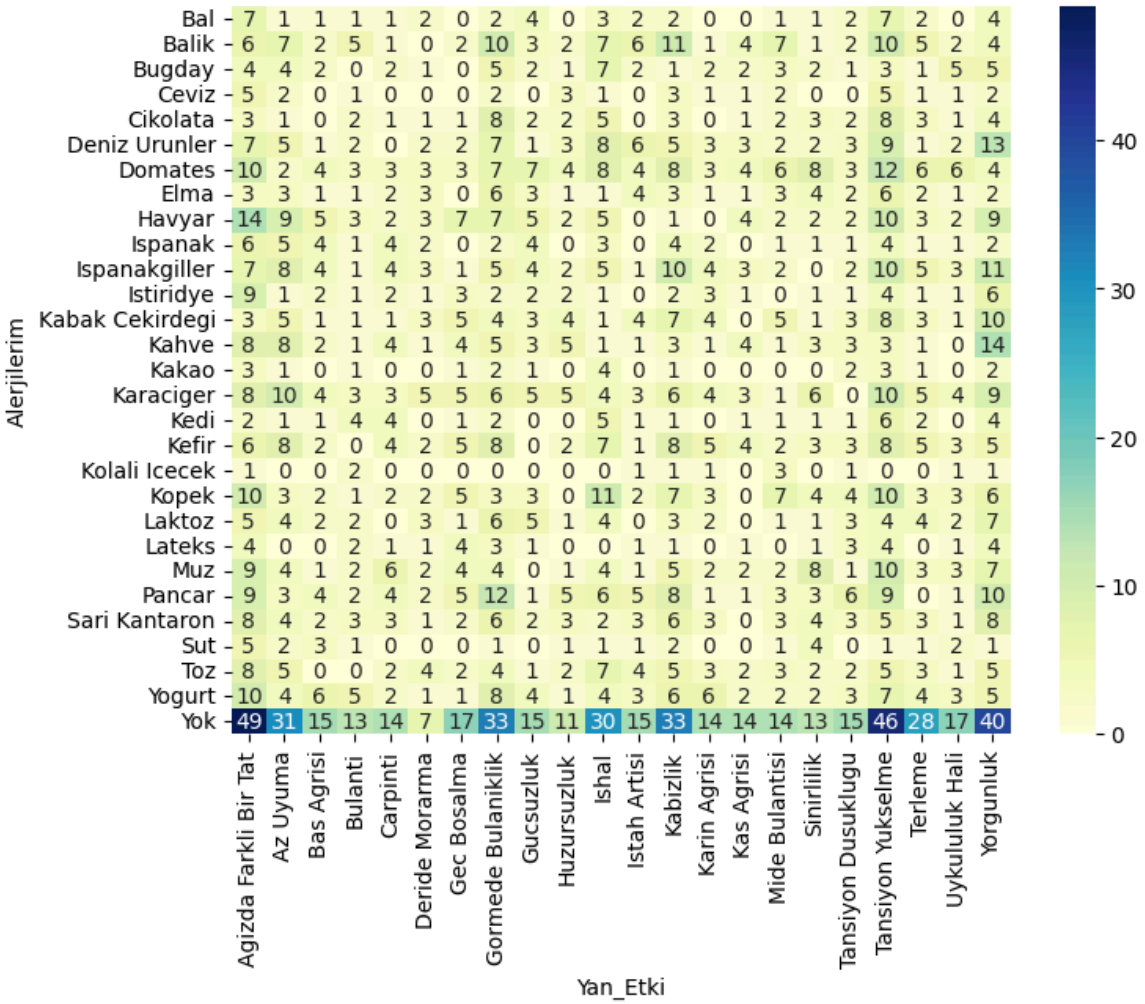


Hedef öznitelik nominal kategorik değişken olduğundan sayısal değerlerle ilişkini gösterebilmek için Box plot kullandım.

Örnekte görüldüğü üzere ilacın kullanım süresine bağlı olarak farklı yan etkiler gösterebildiği genellenebiliyor. Modelleme aşamasında Yan Etki Gösterim Özniteliği önemli rol oynayabilir.



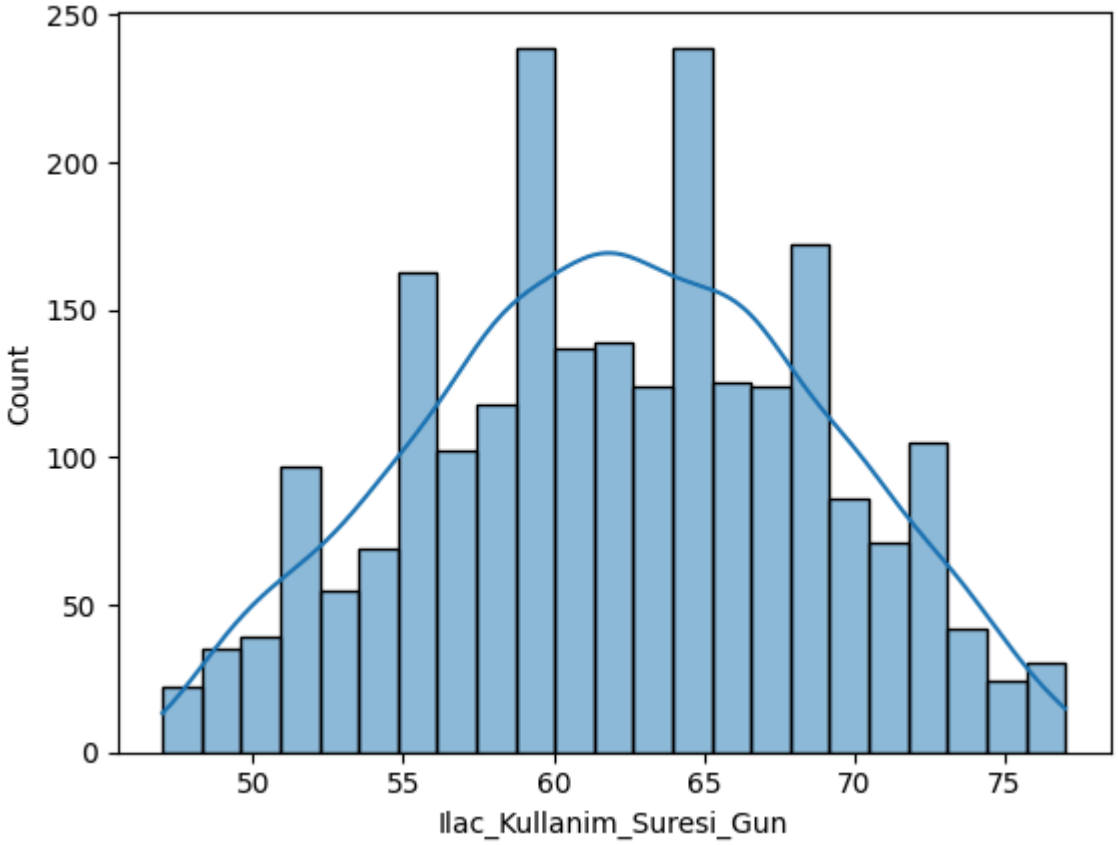
Şekildeki grafikte ise ilaç kullanım süresinin yan etkiler üzerinde gözle görülür bir etkisinin olmadığını görebiliyoruz.



Şekildeki plotta ise 2 tane nominal kategorik özniteliğin arasındaki ilişkinin incelenebilmesi için Heatmap kullanılarak hedef öznitelik olan Yan Etki ile Alerji özniteliğinin kaç defa görüldüğü gözlenmeye çalışıldı.

Gözlem : Yorgunluk yan etkisine sahip hastaların çoğunlukla kahveye alerjisi olduğu görülüyor. Belki bi kahve içebilselerdi böyle bir yan etki söz konusu olmayabilirdi.😅

E)Sayısal Verilerinin Dağılımlarının Kontrolü ve Normalizasyon;



Örnek olarak İlaç kullanım süresi görselde verilmiştir. Her sayısal veri için ayrı ayrı gerçekleştirilip çarpıklık kontrol edilmiştir. Gözle görülür çarpıklık fark edilmedi. Normale yakın dağılım gözlemlendi. Veri setinde bolca kategorik öznitelik olduğu için bu değerler ilerleyen adımları etkili kılması için (Feature Selection, Dimensionality Reduction vb.) normalize edildi.

F)Kategorik Verilerin Encode Edilmesi;

Ordinal kategorik değişken bulunmadığı için Hedef öznitelik haricinde One Hot Encoding uygulandı.

Hedef Öznitelikte Yan Etkilerin Sınıfsal tahmini gerçekleştirileceği için Label Encoding uyguladım.

NOT: İlaç Adı gibi çokça benzersiz(Unique) değere sahip kategorik değişkenler boyutu fazlasıyla arttıracağı için farklı encoding teknikleri uygulanabilir.(Target Encoding, Frequency Encoding).

NOT: Feature Selection, Dimensionality Reduction gibi boyut azaltacak tekniklerin kullanılabilmesi göz önüne alınarak veri seti Encode edilip ham hali modellemeye hazır şekilde bırakıldı.

DİKKAT: Görünmeyen test verisi üzerinde de aynı adımlar izlenmeli!

Özellikle dikkat edilmesi gereken durum One Hot Encoding ve Label Encoding uygulandığı durumlarda özniteliklerin aldığı değerlerin aynı olması gerekmektedir.

EK: İpynb dosyası içerisinde izlenen adımları dökümantasyon içerisinde genel hatlarıyla paylaştım. Dikkat edilmesi gereken kısımlara değindim. İpynb dosyası içerisinde adımların neden öyle gerçekleştirildiğine dair açıklamalar bulunmaktadır. Adımlar uygulanmadan önce incelenmesini tavsiye ederim.
EK: Veri Seti her modele uyacak şekilde hazırlanmıştır. Farklı modeller eğitilecekse bazı adımlar atlanabilir.

Örnek: Kategorik verileri işleyebilen CatBoost modeli encoding adımını atlamayı sağlayabilir.

Örnek: Decision Tree veya Decision Tree temelli modeller(Veri değerleri arasındaki uzaklık değerlerini kullanmayan modeller) veri setinin normalize edilmesine ihtiyaç duymamaktadır.