**Pusula Talent Academy**

Data Science Intern Case Study

Ebru Küçük - Bilgisayar Mühendisi

[kucukkebru@gmail.com](mailto:kucukkebru@gmail.com)

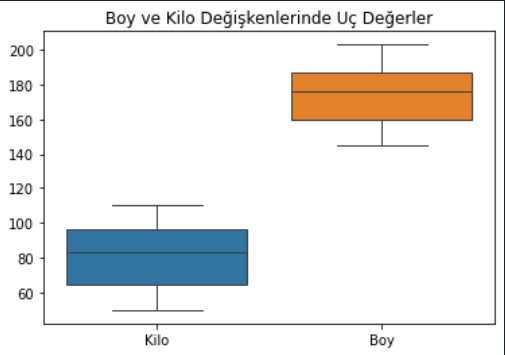
Bu rapor, bir hastane verisi üzerinde yapılan analiz ve pre-process aşamalarını incelemektedir. Dataset hastalarla ilgili kişisel bilgiler, ilaç kullanımı ve aile hastalık geçmişi gibi tıbbi bilgiler yer almaktadır. Çalışmada veriyi anlamak, eksik değerleri ele almak, kategorik ve sayısal değişkenleri işlemek, ve veriyi görselleştirmek amacıyla çeşitli adımlar gerçekleştirilmiştir.

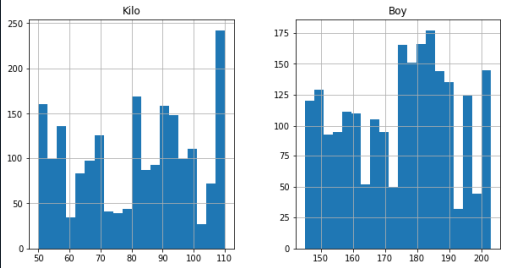
**a.Giriş**

Bu çalışma, hastane datasetindeki potansiyel anomalileri ve veri yapısını anlamaya yönelik yapılan bir keşifsel veri analizi (EDA) ve ardından veriyi modellemeye hazır hale getirmek için yapılan pre-process adımlarını içermektedir. Datasette **Cinsiyet, Boy, Kilo**, ve **İlaç Adı** gibi değişkenler yer almakta olup, eksik değerler, kategorik değişkenler ve sayısal değişkenlerin normalleştirilmesi gibi geliştirmeler yapılmaya çalışılmıştır.

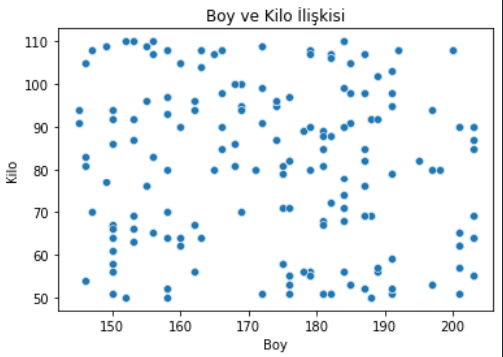
**b.Keşif Veri Analizi (EDA)**

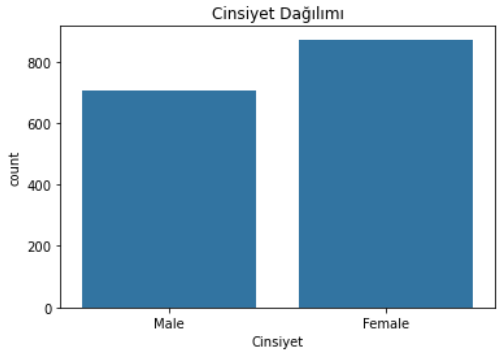
Verideki uç değerleri incelemek için **Kilo** ve **Boy** sütunları için bir **boxplot** oluşturulmuştur. Bu grafik, iki değişkende de aşırı yüksek ya da düşük değerlerin varlığını görselleştirmiştir. Çıktı şu şekildedir:



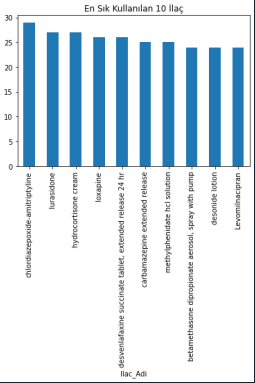
Kilo ve boy dağılımı için özellik için histogram çizilmiştir, böylece veri dağılımı gözlemlenmiştir. Çıktı şu şekildedir:  
 

Scatterplot kullanılarak **Boy** ve **Kilo** değişkenleri arasındaki ilişki incelenmiştir. Bu ilişki, iki değişken arasındaki doğrusal bir bağıntıyı ortaya koymaktadır. Çıktı şu şekildedir:



**Cinsiyet** değişkeni için bir **countplot** oluşturulmuş, böylece veri setindeki cinsiyet dengesinin görselleştirilmesi sağlanmıştır. Çıktı şu şekildedir:  
 

En sık kullanılan 10 ilacın dağılımı çubuk grafik ile görselleştirilmiştir. Çıktı şu şekildedir:



**c.Pre-processing**

Veriyi modellemeye hazır hale getirmek için çeşitli ön işleme adımları uygulanmıştır.

Veride bulunan eksik değerler iki farklı yöntemle doldurulmuştur:

* **Kategorik Değişkenler İçin SimpleImputer:** Kategorik değişkenlerdeki eksik değerler, en sık görülen (mod) değer ile doldurulmuştur. Bu işlem, **Cinsiyet**, **Il**, **Alerjilerim**, **Kronik Hastalıklarım** gibi sütunlara uygulanmıştır.
* **Sayısal Değişkenler İçin KNNImputer:** **Kilo** ve **Boy** gibi sayısal değişkenlerdeki eksik değerler ise KNNImputer ile doldurulmuştur. Bu yöntem, benzer veri noktalarına bakarak eksik değerlerin tahmin edilmesini sağlamaktadır.

Kategorik değişkenler modelleme işlemlerinde kullanılabilmesi için sayısal değerlere dönüştürülmüştür. Bu işlem için **OneHotEncoder** kullanılmış ve kategorik değişkenler birer dummy değişkene dönüştürülmüştür. Bu aşamada **drop='first'** parametresi ile her kategoriden bir sınıf düşürülerek tekil bir sınıfı temsil eden değişkenler oluşturulmuştur.

Sayısal değişkenlerde standart bir ölçekleme yapılmıştır. **StandardScaler** kullanılarak **Kilo** ve **Boy** değişkenleri normalize edilmiştir. Bu işlem, bu iki değişkenin ortalamasının 0 ve standart sapmasının 1 olacak şekilde dönüştürülmesini sağlamıştır.

### **d.Sonuç**

Bu çalışma kapsamında, hastane veri seti üzerinde keşifsel veri analizi gerçekleştirilmiş ve veriyi temizleyip modellemeye uygun hale getirmek için ön işleme adımları uygulanmıştır. Eksik değerler başarılı bir şekilde doldurulmuş, kategorik değişkenler sayısal hale getirilmiş ve sayısal değişkenler ölçeklenmiştir. Bu aşamalar, daha ileri analizler ve makine öğrenmesi modelleri için uygun bir veri seti sunmaktadır.