**SÜMEYYE BULUT – sumeyyebulut222@gmail.com**

Projemizde kullandığımız veri seti ilaç yan etkileri. Bu veri seti ilaç yan etkilerini, kullanıcının tıbbi ve kişisel bilgilerini içeriyor.

Veri setiyle yapılabilecekler:

İlaç adı ile yan etki sütunları arasındaki ilişkiye bakılarak bir ilacın hangi yan etkiye sahip olduğu anlaşılabilir.

İlaç kullanıcılarının cinsiyet, kilo, boy, kan grubu, kronik hastalıklar, il, uyruk gibi özelliklerinin yan etki üzerindeki etkisi analiz edilebilir.

Ailede bulunan kronik hastalıklar(anne, baba, kız kardeş, erkek kardeş kronik hastalıkları) ile yan etki arasında bir ilişki olup olmadığına bakılabilir.

İlaca başlama tarihi, ilacın bitiş tarihi ve yan etki bildirim tarihine bakılarak yan etkilerin görülme zamanı ve belli zamanda görülen yan etkiler analiz edilebilir.

**1-Kütüphanelerin içe aktarılması**

Pandas: Veri manipülasyonu ve analizi için kullanılır.

Matplotlib ve Seaborn: Verileri görselleştirmek için kullanılır.

Sklearn: Makine öğrenimi araçları için kullanılır.

**2-Veri setini yükleme**

Veri, bir Excel dosyasından yüklenip DataFrame oluşturulur.

**3-Verilerin genel yapısını inceleme**

**df.head()** ile verilerin genel yapısının anlaşılması için ilk 5 satır görüntülenir.

**df.columns()** ile sütun isimleri listelenir.

**df.info()** ile DataFrame hakkında detaylı bilgi edinilir.(toplam satır sayısı, veri türü vb.)

**df.describe().T** ile sayısal sütunların temel istatistikleri döndürülür.(ortalama, standart sapma vb.)

**4-Eksik verileri tespit etme**

**df.isnull()** ile hücrelerde eksik değer(NaN) olup olmadığı kontrol edilir. Eksik değerlerde true, eksik olmayan değerlerde false yazar.

**print(df.isnull().sum())** ile sütunlardaki eksik değerlerin sayısı hesaplanır.

**5-Kategorik değişkenlerin dağılımının görselleştirilmesi**

cat\_vars adında bir kategorik değişkenler listesi oluşturulur. For döngüsü ile her bir değişken için countplot oluşturulur ve başlık ile döndürülür.

**6-Sayısal değişkenlerin dağılımının görselleştirilmesi**

Kilo ve boy sütunlarını içeren bir DataFrame oluşturulur. Bu sütunlar için histogram çizdirilir.

Kullanıcı id, sayısal bir değişken olmasına rağmen benzersiz değerlerdir ve her biri sadece bir kaydı temsil ettiği için görselleştirilmesi pek anlamlı olmayabilir.

**7-Tarih verilerinin görselleştirilmesi**

date\_vars adında içinde ‘Dogum\_Tarihi, Ilac\_Baslangic\_Tarihi, Ilac\_Bitis\_Tarihi, Yan\_Etki\_Bildirim\_Tarihi’ sütunları bulunan bir liste oluşturulur.

For döngüsü ve pd.to\_datetime() fonksiyonu kullanılarak yoğunluk grafikleri çizdirilir.

**8-Eksik verilerin doldurulması**

SımpleImputer kullanılarak kategorik değişkenler için en çok kullanılan değerle ve sayısal değişkenler için ortalamayla eksik veriler doldurulur.

**9-Kategorik değişkenleri encode etmek**

OneHotEncoder kullanılarak kategorik veriler sayısal verilere dönüştürülür ve modelleme için hazır hale getirilir.

**10-Sayısal değişkenleri standartlaştırma**

StandardScaler ile sayısal değişkenler standartlaştırılarak yani ortalama 0, standart sapma 1 olacak şekilde modelleme için hazır hale getirilir.

Bu işlemleri yaptıktan sonra **print(df.isnull().sum())** çalıştırdığımızda eksik değer sayısı bütün sütunlar için 0 olarak döndürüldü. Bu sayede eksik verilerin doldurulduğu anlaşıldı.

Standartlaştırılmış sayısal sütunları kontrol etmek için **print(df[['Kilo', 'Boy']].describe())** komutu çalıştırılır ve ortlamanın 0, standart sapmanın 1 olması kontrol edilir.