Pusula Talent Academy – Veri Bilimi Görevi

Şevval Topkan

19/09/2024

1. Giriş

Bu proje, ilaç yan etkilerine dair bir veri setinin analiz edilmesi ve modellemeye uygun hale getirilmesini amaçlamaktadır. Projede temel olarak üç ana adımlı bir yol izlenmiştir, bunlar: Keşifsel Veri Analizi (EDA), Veri Ön İşleme ve Veri Görselleştirme.

İlk aşamada veri setinin yapısını anlamak ve sorunları tespit etmek amacıyla keşifsel analiz yapılmıştır. Eksik veriler ve veri türleri incelenmiş, sayısal ve kategorik değişkenlerin dağılımları analiz edilmiştir. İkinci aşamada ise, eksik verilerin doldurulması, kategorik verilerin uygun formatlara dönüştürülmesi ve sayısal verilerin ölçeklendirilmesi gibi veri ön işleme adımları gerçekleştirilmiştir. Son olarak, verinin görsel analizini sağlayan grafikler ve ısı haritaları ile veri içindeki önemli ilişkiler ortaya konmuştur.

Bu süreç sonunda elde edilen veri seti, makine öğrenmesi ve diğer analiz yöntemlerine uygun hale getirilmiştir. Proje boyunca kullanılan yöntemler, veri analizi için standart ve gerekli adımları içerirken, elde edilen sonuçlar da veri setinin genel yapısı ve ilişkileri hakkında derinlemesine bilgi sağlamaktadır.

2. Kullanılan Yöntemler ve Teknikler

Bu bölümde, projede kullanılan veri bilimi yöntemleri ve teknikleri detaylı olarak açıklanacaktır.

Veri seti üzerinde yapılan ilk analiz, veri setinin genel yapısını anlamak ve olası sorunları tesit etmek amacıyla gerçekleştirilmiştir. Bu aşamada:

2.1 Keşifsel Veri Analizi (EDA)

Veri seti üzerinde gerçekleştirilen ilk analiz, veri setinin genel yapısını anlamak ve olası sorunların tespit edilmesi amacıyla gerçekleştirilmiştir.

Pandas kullanılarak veri seti yüklenmiş ve incelenmiştir.

Eksik veriler tespit edilmiş ve sayısal veriler için ortalama, kategorik veriler için en sık kullanılan değer ile doldurulmuştur.

Veri setinde bulunan sayısal değerler için histogramlar ve scatter plot'lar oluşturulmuştur.

Kategorik değişkenler için dağılım grafikleri (bar chart) çizilmiştir.

2.2 Veri Ön İşleme

Veriyi modelleme ve analiz adımlarına uygun hale getirebilmek adına aşağıdaki işlemler oluşturulmuştur.

Kategorik Verilerin Kodlanması: Kategorik değişkenler (örneğin cinsiyet, ilaç adı) OneHotEncoder ile kodlanmıştır. Bu işlem, modelleme sırasında kateorik verilerin daha kullanılabilir bir hale gelmesine olanak tanımaktadır.

Standardizasyon: Sayısal değişkenler (kilo ve boy) StandartScaler kullanımıyla standardize edilmiştir. Bu verilerin ortalamasının 0 olarak, standart sapmasının ise 1 olarak ölçeklendirilmesi işlemine denir. Bu teknik, farklı ölçülerde olan sayısal verilerin karşılaştırılabilir hale gelmesine olanak tanır.

2.3 Veri Görselleştirme

Projede bulunan veri görselleştirme adımları, veri setinin dağılımlarını anlamak ve ilişkilerini kavrayabilmek adına kullanılmıştır. Kullanılan geliştirme araçları ise şunlardır:

Cinsiyet Dağılımı: Verilen veri setinde bulunan cinsiyet değişkenlerinin dağılımı bir bar grafiği kullanılarak görselleştirilmiştir.

İlaç Adı Dağılımı: İlaçların kullanım sıklıklarının oranları bir bar grafiği ile görselleştirilmiştir. İlaç adlarının uzun olmasından kaynaklı grafikte yazı boyutu küçültülmüş ve yönü yeniden konumlandırılmıştır.

Korelasyon Isı Haritası: Sayısal değişkenler arasındaki korelasyon bir ısı haritası kullanılarak görselleştirilmiştir. Bu sayısal değişkenler, boy ve kilo olmaktadır.

3. Kullanılan Kütüphaneler

Pandas: Veri setini yükleme, işleme ve manipülasyon işlemleri için kullanılmıştır.

Scikit-learn: Eksik verilerin doldurulması, kategorik verilerin kodlanması ve standardizasyon işlemleri için kullanılmıştır.

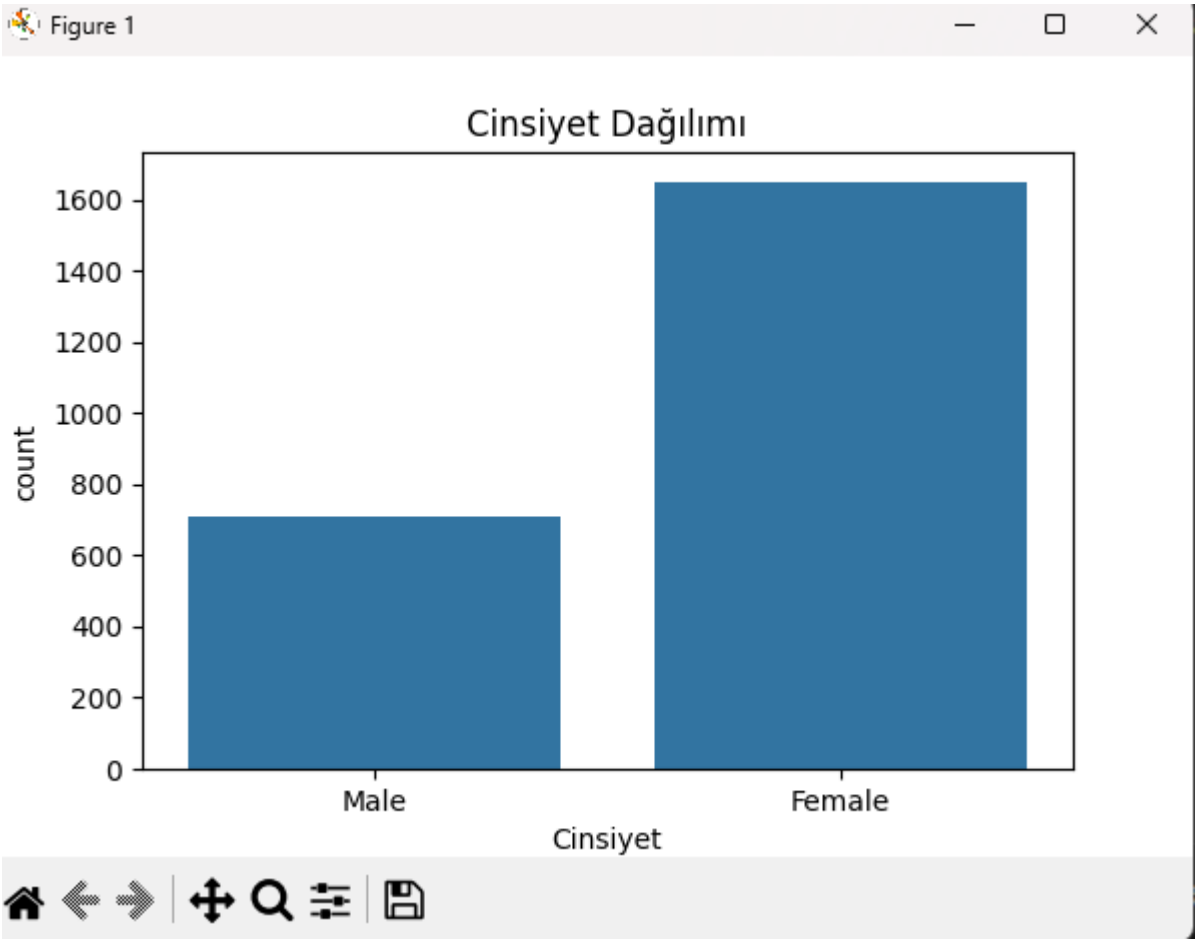
Matplotlib ve Seaborn: Veri görselleştirmeleri için kullanılmıştır.

4. Sonuçlar

Bu bölümde proje boyunca elde edilen bulguların görselleri ve analizleri yer alacaktır.

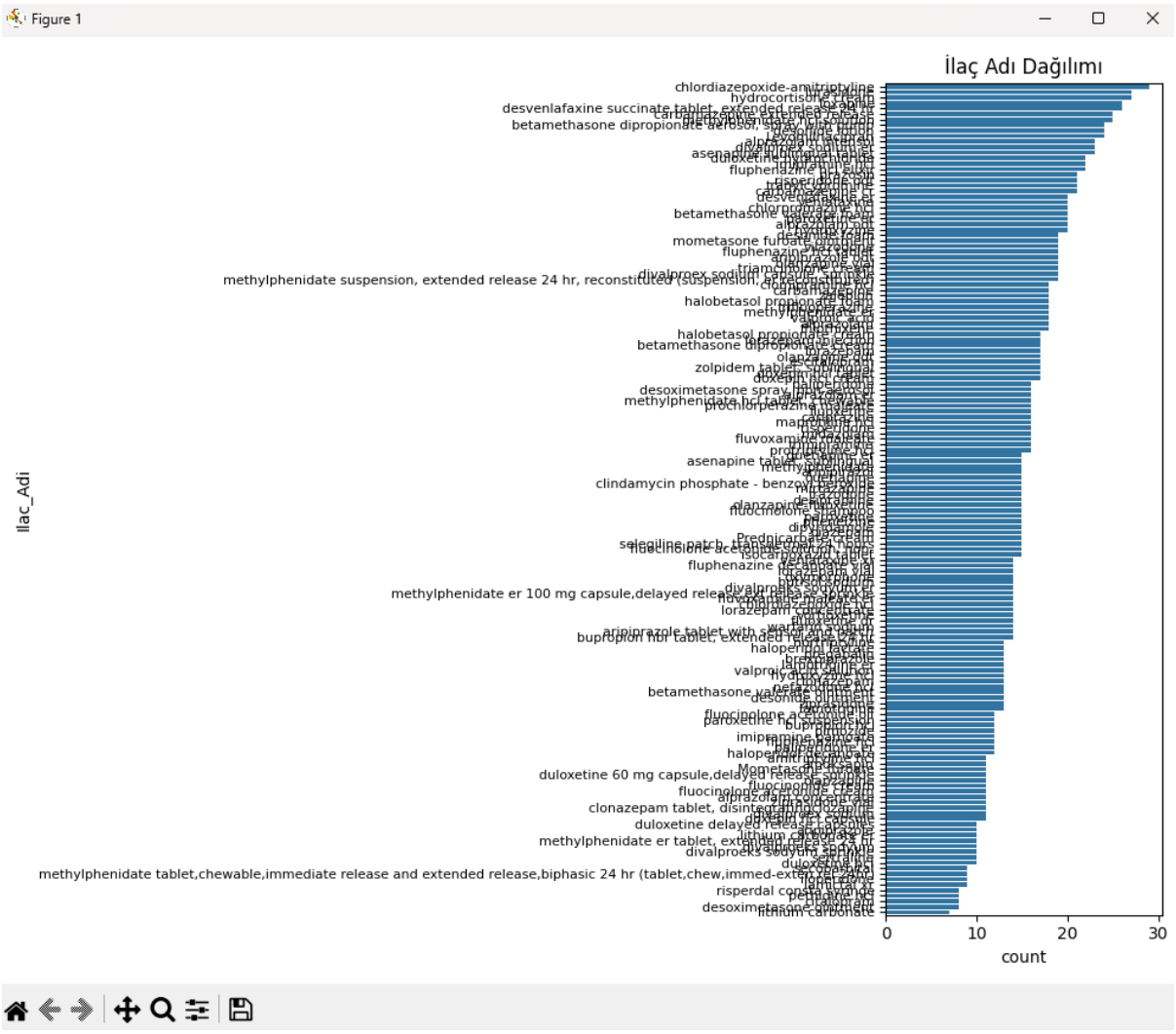
4.1 Cinsiyet Dağılımı

Cinsiyet dağılımı, veri setinde bulunan kadın bireylerin sayısı ve erkek bireylerin sayısının görselleştirilmesi amacıyla kullanılmıştır. Ortaya çıkan bulgulara göre kadın bireylerin erkek bireylerin daha fazla olduğu gözlenmektedir.



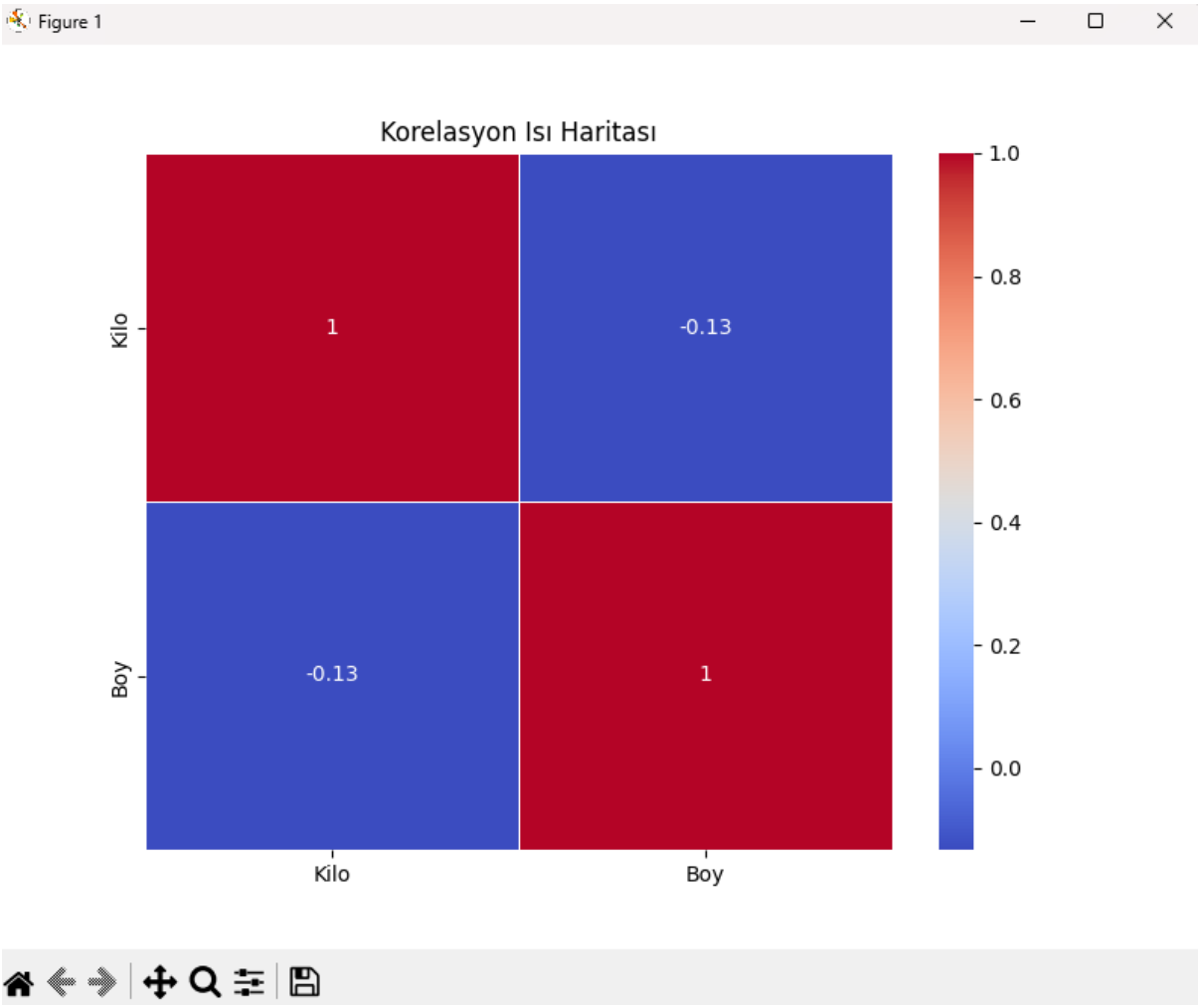
4.2 İlaç Adı Dağılımı

İlaç adı dağılımı, veri setinde bulunan ilaçların isimlerine göre kullanım sıklığını oranlamaktadır. Elde edilen bulgular sonucunda bazı ilaçların daha sık kullanıldığı gözlemlenmiştir. Tabloda, ilaç adlarının uzunluğundan kaynaklı bir karmaşıklık görünse de bunu önlemek amacıyla yazı boyutu ve yönü yeniden ayarlanmıştır.



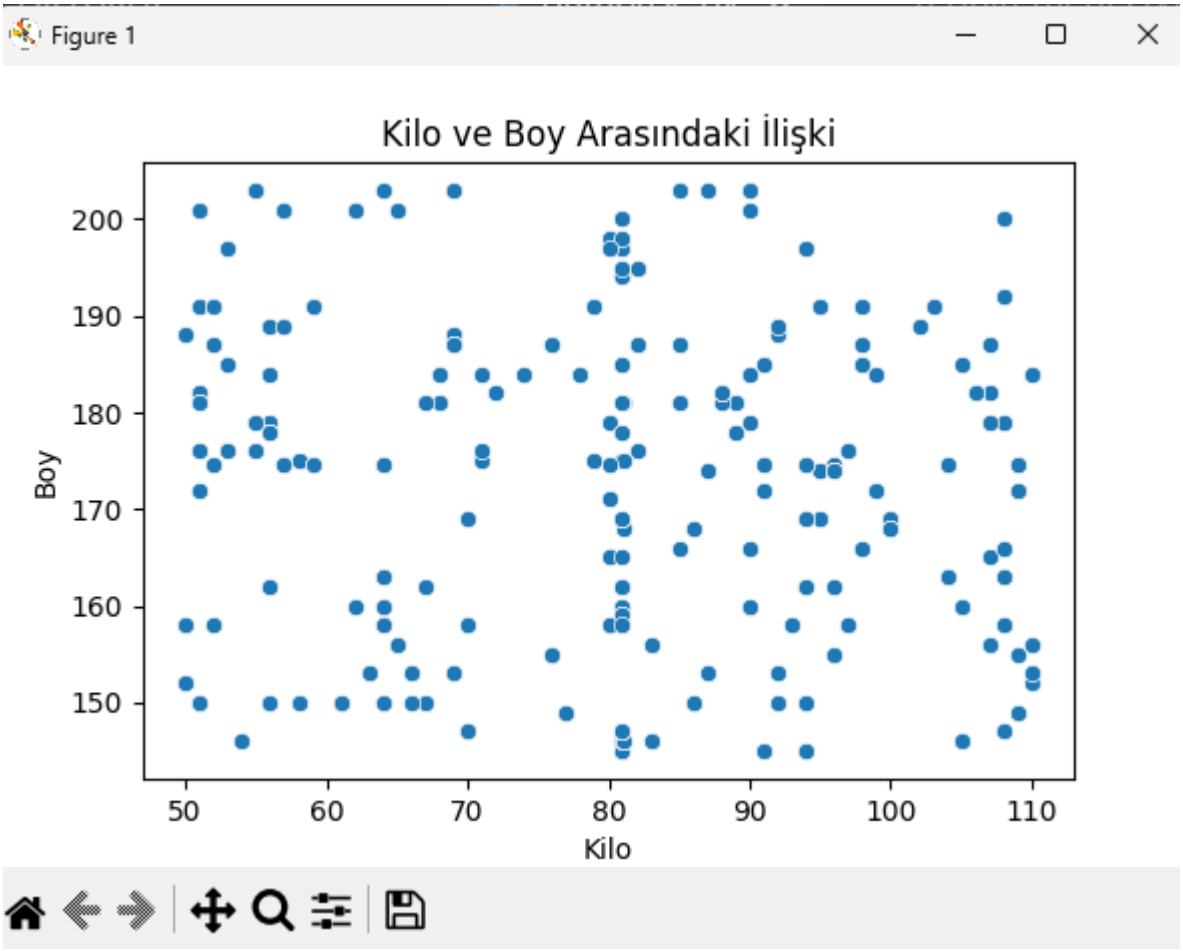
4.3 Korelasyon ısı Haritası

Isı haritası, kilo ve boy arasındaki korelasyonu görselleştirmiştir. Elde edilen bulgular sonucunda kilo ve boy arasında negatif korelasyon gözlemlenmiştir.



4.4 Kilo ve Boy Scatter Plot

Scatter plot, kilo ve boy arasında bir ilişkinin bulunup bulunmadığını anlamak adına görselleştirilmiştir. Elde edilen bulgular sonucunda kilo ve boy arasında bir ilişki bulunmadığı sonucuna varılmıştır.



5. Sonuç ve Gelecek Çalışmalar

Bu projede, ilaç yan etkileriyle ilgili bir veri seti üzerinde keşifsel veri analizi gerçekleştirilmiş, veriler başarılı bir şekilde ön işleme tabi tutulmuş ve görselleştirme adımları uygulanmıştır. Standardizasyon ve kategorik verilerin kodlanması ile veri seti, modelleme adımlarına uygun hale getirilmiştir.

Gelecekte, bu veri seti üzerinde çeşitli makine öğrenmesi algoritmalarını uygulayarak yan etkiler ve ilaçlar arasındaki ilişkiler daha derinlemesine analiz edilebilir. Ek olarak, veri seti genişletilerek daha fazla değişken eklenebilir ve analizler daha kapsamlı hale getirilebilir.