**Pusula Vaka Çalışması**

**Hazırlayan:** Ege Cankut Solkol  
**Tarih:** 22 Eylül 2024

**1. Keşifsel Veri Analizi (EDA)**

**Veri Kümesi Genel Bakış:**

* **Toplam Satır Sayısı:** 2357
* **Toplam Sütun Sayısı:** 19
* **Ana Özellikler:**
  + Demografik bilgiler (ör. "Kullanici\_id," "Cinsiyet," "Dogum\_Tarihi," "Il")
  + Sağlık ile ilgili veriler (ör. "Kronik Hastaliklarim," "Baba Kronik Hastaliklari," "Anne Kronik Hastaliklari")
  + İlaç kullanımı verileri (ör. "Ilac\_Adi," "Ilac\_Baslangic\_Tarihi," "Yan\_Etki")
  + İlaçlara bağlı yan etkiler (ör. "Yan\_Etki," "Yan\_Etki\_Bildirim\_Tarihi")

**Ana Bulgular:**

1. **Eksik Veriler:**
   * Aşağıdaki sütunlarda önemli eksik veriler bulundu:
     + **Cinsiyet:** 778 eksik değer
     + **Kilo:** 293 eksik değer
     + **Boy:** 114 eksik değer
     + **Kronik Hastalıklar:** 392 eksik değer
2. **Özelliklerin Dağılımları:**
   * Sayısal özelliklerin dağılımını görmek için histogramlar kullanıldı. Özellikler arasında **Kilo, Boy, Yaş** ve **İlaç Kullanım Süresi** bulunuyor.
   * Sayısal özellikler arasındaki ilişkileri incelemek için dağılım grafikleri oluşturuldu.
3. **Korelasyon Analizi:**
   * Korelasyon matrisi, özellikler arasındaki ilişkileri belirlemek için kullanıldı.
   * Özellikle 0.2’den büyük olan korelasyonlar filtrelendi ve kendileriyle olan korelasyonlar çıkarıldı.
   * İlginç korelasyon bulguları şunlardır:
     + **Alerjilerim\_Kedi** ve **Kan Grubu\_A RH+:** 0.42 korelasyon değeri, hayvan alerjileri ile kan grubu arasında ilginç bir ilişkiyi gösteriyor olabilir.
     + **Alerjilerim\_Süt** ve **Anne Kronik Hastaliklari:** 0.37 korelasyon değeri, süt alerjisinin anne tarafındaki kronik hastalıklarla olası bir bağlantısı olduğunu gösterebilir.
     + **Etki Süresi Gün** ve **İlaç Kullanım Süresi Gün:** 0.34 korelasyon değeri, ilaç kullanım süresi ile yan etkilerin görülme süresi arasında beklenen bir ilişkiyi ortaya koyuyor.
     + **Alerjilerim\_Muz** ve **Kronik Hastaliklarim\_Kemik Erimesi:** 0.31 korelasyon değeri, muz alerjisi olan kişilerin kemik erimesi yaşama olasılığının olabileceğini düşündürüyor.
     + **Alerjilerim\_Domates** ve **İl\_Ankara:** 0.31 korelasyon değeri, Ankara'da yaşayan kişilerin domates alerjisi olma olasılığının diğer illere göre daha yüksek olduğunu gösterebilir.

**Yan Etkiler Analizi:**

* Her ilaç için yan etkiler toplandı ve bunların yüzdesel dağılımları pasta grafikleriyle gösterildi.
* **Yüzde Hesaplamaları:**
  + %5’in altında kalan yan etkiler “Diğer” kategorisine alındı.
  + Her grafikte toplam yan etki sayısı başlıkta gösterildi.

**2. Veri Ön İşleme**

**1. Eksik Verilerin İşlenmesi:**

* **Kilo** ve **Boy** gibi önemli sütunlardaki eksik değerler, **Random Forest** yöntemi ile dolduruldu. Bu yöntem, basit ortalama ile doldurmaya göre daha iyi sonuçlar verdi.
* **Cinsiyet** gibi kategorik özelliklerde ise, Random Forest ile yapılan tahminler tatmin edici sonuçlar vermediğinden, **mod imputation (en sık görülen değeri kullanarak doldurma)** yöntemi uygulandı.

**2. Özellik Mühendisliği:**

* **Yaş Hesaplaması:** **Doğum Tarihi**'nden **Yaş** sütunu hesaplandı.
* **Süre Hesaplamaları:** Aşağıdaki sütunlar oluşturuldu:
  + **İlaç Kullanım Süresi (Gün):** İlaç başlama ve bitiş tarihleri arasındaki gün sayısı.
  + **Etki Süresi (Gün):** İlaç başlama tarihi ile yan etkinin bildirilme tarihi arasındaki gün sayısı.

**3. Standartlaştırma ve Kodlama:**

* **Standartlaştırılmış Girişler:** Çoklu değer içeren sütunlar (ör. "Kronik Hastalıklarım") standartlaştırıldı.
* **Sütunların Patlatılması:** Birden fazla değeri olan sütunlar, her değer ayrı bir satıra bölünerek analiz edilebilir hale getirildi.
* **One-Hot Encoding:** Kategorik sütunlar (ör. kronik hastalıklar, ilaç isimleri, yan etkiler, kan grubu) one-hot encode edilerek makine öğrenimi modelleri için uygun hale getirildi.

**4. Verilerin Gruplanması:**

* One-hot encoded veriler, her bireyi (indeks bazında) temsil edecek şekilde gruplandırıldı ve maksimum operasyonu uygulandı.

**5. Gereksiz Sütunların Çıkarılması:**

* **Uyruk** ve **Kullanıcı ID** gibi analize katkısı olmayan sütunlar çıkarıldı.

Bu belge, veri ön işleme ve keşifsel veri analizindeki temel adımları özetlemektedir. Korelasyon analizi sonuçları, özellikle alerji ve kronik hastalıklar arasındaki ilişkileri incelemek için daha derin araştırmaların yapılabileceğini önermektedir.