Elif Karadeniz • karadenizelif45@gmail.com

Keşifsel Veri Analizi (EDA)

- Proje için ilk olarak veri seti yüklendi. Bu aşamada ilk yapılan şey veri seti hakkında temel bir bilgiye sahip olmaktı. Bu nedenle veri setinin ilk 5 satırı, sütunların tipi, sayısal ve sayısal olmayan sütunlar, veri setini sütun ve satır yapısı incelendi.
- Daha sonrasında veri setindeki eksik değerlerin sayısı bulunup, grafikle çizdirildi. Bunun dışında ön işleme ve modelleme aşaması için önemli olan analizler yapılmaya başlandı;
  - Boy ve kilo arasındaki korelasyon hesaplandı ve negatif bir ilişki olduğu ortaya çıkarıldı.
  - Alerjilerin sıklığına bakıldığında ise kullanıcılar için; en çok domatese karşı bir alerji görülmekteyken, en az kolalı içeçek için alerji görünmekte olduğu anlaşıldı.
  - En yaygın 10 kronik hastalık incelendi ve en çok görülen hastalığın hipertansiyon olduğu ortaya çıkarıldı.
  - En çok kullanılan 10 ilaç ile en sık görülen 10 kronik hastalık için sıralama yaptığımızda ise kan hastalıkları ve astım için en çok tercih edilen ilacın hdyrocortisone cream olduğu gibi çıkarımlar yapıldı.
  - En çok kullanılan 10 alerji ile en sık görülen 10 kronik hastalık için sıralama yaptığımızda ise örn; Alzheimer, astım hastalığında en çok karaciğere alerji görülmekte olduğu anlaşıldı.
  - Yan etkiler incelendiğinde en sık görülen yan etkinin ağızda farklı bir tat iken en az görülen yan etkinin ise deride morarma olduğu anlaşıldı.
  - Hastalıklara özel yan etki incelendiğinde ise kalp hastalıkları için en çok görülen yan etkinin yorgunluk, hipertansiyon için tansiyon yükselme, diyabet için ise ağızda farklı bir tat olduğu görüldü.
  - Boy ve kilo için subplot çizildiğinde kiloların büyük çoğunluğunun 70-100 kg aralığında yoğunlaştığı, boyların ise 160-190 cm arasında daha sık görüldüğü anlaşıldı.

Veri Ön İşleme

- Özellik Mühendisliği
  - İlaç Kullanım Süresi: Bu aşamada, iki tarih sütunu olan Ilac_Baslangic_Tarihi ve Ilac_Bitis_Tarihi kullanılarak İlaç kullanım süresi hesaplandı. İlaç bitiş tarihi ile başlangıç tarihi arasındaki fark gün cinsinden yeni bir sütun olarak eklendi. Bu yeni özellik, kullanıcıların ilaçları ne kadar süre kullandıklarını anlamamıza ve bu bilginin sağlık durumu ile nasıl ilişkili olabileceğini analiz etmemize yardımcı olacağı için eklendi.
  - Vücut Kitle İndeksi (BMI): Kilo ve boy değerlerinden yola çıkarak BMI hesaplandı. BMI, kullanıcının sağlık durumu ile ilgili önemli bir gösterge olup, fazla kilolu ya da zayıf olmanın hastalık riski üzerindeki etkilerini analiz etmek için oluşturuldu.

    Bu 2 sütun, model performansını artırmak amacıyla özellik (öznitelik) mühendisliği sürecinde örnek olarak eklenmiştir. Bu tarz başka sütunların oluşturulması modelleme adımı için önemlidir. Özellik mühendisliği, modelin tahmin gücünü iyileştirmek için yeni ve anlamlı veriler üretmeye yardımcı olur.
    - Eksik Verilerin Temizlenmesi
    - Bazı sütunlarda eksik değerler bulunuyordu ve bu veriler modelleme için uygun değildi. Özellikle demografik ve sağlık verilerinin kritik öneme sahip olduğunu göz önünde bulundurarak, eksik değer içeren satırlar (Cinsiyet , Il, Alerjilerim, Kronik Hastalıklarım, Baba Kronik Hastaliklari , Anne Kronik Hastaliklari, Kiz Kardes Kronik Hastaliklari, Erkek Kardes Kronik Hastaliklari Kan Grubu, Boy ve Kilo) temizlendi. Bu adım, verilerin eksiksiz ve güvenilir olmasını sağladı.
    - Tarih Sütunlarının Dönüştürülmesi ve Kategorik Verilerin Kodlanması (Encoding)
    - Bu aşamada tarih verileri datetime formatına dönüştürülerek ardından Unix zaman damgasına (timestamp) çevirelerek sayısal bir formata dönüştürüldü, böylece analiz ve modelleme süreçlerinde kullanılabilir hale getirildi. Geriye kalan sayısal olmayan veriler, sayısal modellere uygun hale getirilmek üzere label encoding yöntemiyle sayısal değerlere dönüştürüldü. Bu işlem, özellikle makine öğrenimi algoritmaları için gereklidir.

    - Verilerin Standartlaştırılması
        - Sayısal sütunlar arasında dağılım farklılıklarını dengelemek için StandardScaler kullanılarak veriler standart hale getirildi. Bu sayede, farklı ölçeklerde olan veriler aynı düzeye getirildi.
    - Aykırı Değerlerin Tespiti
        - Aykırı değerlerin tespiti, veri kalitesini artırmak için önemlidir. Z-Score yöntemi kullanılarak her sütun için aykırı değerler belirlendi. Aykırı değerler, analizde olası hataların önüne geçmek ve modellerin performansını iyileştirmek için dikkate alındı. Veri setinde aykırı değer bulunmadığını teyit etmek amacıyla Boxplot kullanılarak verilerin dağılımı ve aykırı değerler görselleştirildi. Bu, veri setindeki uç noktaların belirlenmesi açısından önemli bir adımdı.
- Korelasyon Analizi
  - Korelasyon matrisi ile değişkenler arasındaki ilişkiler analiz edildi.

  Sonuç
  Veri analizi aşamasında veri seti hakkında birçok çıkarım yapıldı. Veri ön işleme ve özellik mühendisliği sürecinde ise, hem ham veriler temizlenmiş hem de modelleme için kullanılabilecek yeni özellikler türetilmiştir. Özellikle İlaç Kullanım Süresi ve BMI gibi özellikler, verilerin daha anlamlı hale getirilmesine katkıda bulunmuş ve modelleme aşamasında sağlık durumunu daha iyi analiz etmek için önemli değişkenler haline gelmiştir.