Muharrem Gedik [gdk.muharrem@gmail.com](mailto:gdk.muharrem@gmail.com)

İlk olarak veri setini import ederek inceledim ve ilk 10 satır gibi son 10 satır gibi inceleyebileceğim değerleri ekrana bastım.

Veri türleri ve genel bilgi almak amaçlı info ile data frame’i incelediğimde object türünde veriler gördüm fakat bu veriler model için karlı değil. Sebepleri:

**Object türündeki sütunları categorical türüne çevireceğiz.**

Neden Object Verileri Categorical Türüne Çevirmelisiniz?

***Bellek Kullanımını Optimize Etme***

Object türündeki veriler genellikle string (metin) verileri içerir ve bellekte daha fazla yer kaplayabilir. Pandas'ın categorical veri türü, bu metinleri daha kompakt bir şekilde temsil eder ve bellekte yer tasarrufu sağlar.

***Performans İyileştirmeleri***

Kategorik veriler üzerinde yapılan işlemler (örneğin, gruplama, sıralama) genellikle daha hızlıdır çünkü categorical veri türü, verileri sayısal kodlara dönüştürür ve bu sayede işlem süresi azalır.

***Modelleme İhtiyaçları***

Çoğu makine öğrenimi algoritması, sayısal verilerle çalışır. object türündeki kategorik verileri categorical türüne dönüştürmek, bu verileri sayısal formatlara dönüştürmek için daha uygun hale getirir. Ayrıca, bazı modeller bu tür verileri doğrudan işleyebilir.

***Kategorik Özellikleri Daha İyi Yönetme***

categorical veri türü, her kategori için bir etiket ve bu etiketlerin sıralı veya sırasız olup olmadığını belirlemeye olanak tanır. Bu, veri analizi ve işleme sürecinde daha fazla kontrol sağlar.

Daha sonra eksik verilerin sayısına baktım ve category türündeki ve sayısal türdeki değerleri tek tek inceledim.

Category türünde hastalıklarda bir kutuda 2,3 hastalık olabiliyordu bunun için hastalıkları sayısal olarak kodladım ve yeni sütun oluşturup onlara taşıdım.

Kodlama:  
0 Yok

1 Sen hasta

2 Baba hasta

3 Sen ve baba hasta

4 Anne hasta

5 Sen ve anne hasta

6 Baba ve anne hasta

7 Sen, baba ve anne hasta

8 Kız kardeş hasta

9 Sen ve kız kardeş hasta

10 Baba ve kız kardeş hasta

11 Sen, baba ve kız kardeş hasta

12 Anne ve kız kardeş hasta

13 Sen, anne ve kız kardeş hasta

14 Baba, anne ve kız kardeş hasta

15 Sen, baba, anne ve kız kardeş hasta

16 Erkek kardeş hasta

17 Sen ve erkek kardeş hasta

18 Baba ve erkek kardeş hasta

19 Sen, baba ve erkek kardeş hasta

20 Anne ve erkek kardeş hasta

21 Sen, anne ve erkek kardeş hasta

22 Baba, anne ve erkek kardeş hasta

23 Sen, baba, anne ve erkek kardeş hasta

24 Kız kardeş ve erkek kardeş hasta

25 Sen, kız kardeş ve erkek kardeş hasta

26 Baba, kız kardeş ve erkek kardeş hasta

27 Sen, baba, kız kardeş ve erkek kardeş hasta

28 Anne, kız kardeş ve erkek kardeş hasta

29 Sen, anne, kız kardeş ve erkek kardeş hasta

30 Baba, anne, kız kardeş ve erkek kardeş hasta

31 Sen, baba, anne, kız kardeş ve erkek kardeş hasta

Bu kodlama ile benzersiz hastalıkları sayısal olarak modelin anlayabileceği türde ve çok fazla sütun & satır oluşturmadan veri setini de yormadan düzenlemiş oldum.

Her sütun için benzersiz değerleri eksik ve uyumsuz değerleri inceledim.

Aykırı değer için görselleştirme işlemleri gerçekleştirdim. Daha sonra veri setine en uygun olabileceğini düşündüğüm şekilde, veri türlerine ve sürekli süreksiz çok fazla kategorik değer olup olmamasına bağlı olarak SimpleImputer, OnehotEncoder, LabelEncoder, StandartScaler kullandım ve data frame’in son halini excel dosyası olarak kaydettim.

Modelin anlayabileceği ve işleyebileceği şekilde veri setini analiz ettim ve ön işleme aşamalarını tamamladığımı düşünüyorum.