PUSULA TALENT ACADEMY CASE STUDY SUMMARY

Hazırlayan: Akın İbrahim Keleş
E-posta: akin.ibrahim190@gmail.com

1 - EDA

- read() , describe() ve info() fonksiyonları ile veriye genel bakış
- .isna().sum() , duplicated().count() fonksiyonları ile veriyi tanıma
- Veri kontrolü sırasında numerik ve kategorik özellikleri ayrı bir değişken olarak ele alma.
    - numeric_features (kdeplot ile görselleştirme)
    - categorical_features (countplot ile görselleştirme)
- Duplike kayıtlarda aynı hastaların birden fazla kayıtları olabildiğini anlama
    - subplot() kullanarak en çok kayda sahip 10 hastayı belirleme
- Target definition (TedaviSuresi) özelliğinin unique değerleri kontrol edildi
- İki farklı kategorik değerin object tipinden int tipine dönüştürme işlemi yapıldı (TedaviSuresi ve UygulamaSuresi)
- TedaviSuresi'nin dağılımı incelendi**. %70 oranında 15 seans yapılmış.**
- Outlier tespiti için grafik oluşturuldu fakat 15 çok yoğun olduğu için diğer tüm değerleri outlier gibi gördü.
- Kategorik bir değere göre tedavi süresinin dağılımı incelendi.
    - Büyük bir farkla **Fiziksel Tıp Ve Rehabilitasyon,Solunum Merkezi** bölümü için kayıt yapılmış. **2045** kayıt.
    - Cinsiyet ele alındığında **Kadınların Erkeklere göre daha fazla** kaydının olduğu gözlemlendi. (1274/792)
- Numerik bir değere göre tedavi süresinin dağılımı incelendi.
    - Yaş ve Uygulama süresinin dağılımları grafik üstünde incelendi.
    - **En çok 48 yaşındakilerin** kayıt yaptığı gözlemlenirken, Uygulama sürelerinin büyük bir kısmı **20 dakika** olduğu görüldü.
- Ortalama olarak en çok seans gerektiren tedavinin **'Bel'** olduğu gözlemlendi.

2 - DATA PREPROCESSİNG

- **Metin Temizleme**
  - Verilerde anormallik olduğuna rastlanıldı.
  - **\xa0 , \u200b** gibi unicode dummy dataların olduğu görüldü.
  - **'Volteren,GriPiN': 'Volteren,Gripin'** örneği gibi aynı anlamlara gelen ama farklı yazılan verilere rastlanıldı.
  - Veri temizliği **dictionary** yapısı kullanılarak, **replace()** fonksiyonu yardımı ile gerçekleştirildi.
  - Verilerdeki missing value oranı %0'a indirilerek tamamen doldurulmuştur.

- **Kayıp Veri İmputasyonu**
  - 13 özellikten **7** tanesinin **NaN** datalarının olduğu gözlemlendi.
  - Bazı özelliklerin büyük çoğunluğu bazılarının ise çok az bir kısmının eksik olduğu görüldü.
    - **Alerji -> 938 NaN** değer var. Veriye göre çok büyük oran. Dolayısıyla herhangi bir metotla doldurmak yerine nan değerler **'Yok'** şeklinde dolduruldu.
    - **KanGrubu -> 675 NaN** değer var. Veriye göre büyük bir kısmı nan olduğu için **'Bilinmiyor'** şeklinde dolduruldu.
    - **KronikHastalik -> 611 NaN** değeri var. Mantık olarak ya kronik hastalığın vardır ya da yoktur. Kayıt sırasında yapılmamış olabilir. Fakat yine yüksek oranda boş olduğu için **'Bilinmiyor'** şeklinde dolduruldu.
    - **Bolum -> 11 NaN** değeri var. Veriyi inceleme sonrasında KronikHastalik ile bağlantılı olduğunu düşünüldü. KronikHastalik sütununa göre **Bolum** sütunundan en sık tekrar eden değeri buluyor. Ve onu NaN değerine atandı.
    - **Tanilar -> 69 NaN** değeri var. Tanilar özelliği **TedaviAdi** ile ilişkili olduğu görüldü. **TedaviAdi** sütununa göre en sık geçen **Tanilar** ile dolduruldu. Fakat **TedaviAdi'nda** tamamen **boş** değer olanlara rastlanıldı. O değerler **'Yok**' ile dolduruldu.
    - **UygulamaYerleri -> 220** NaN değer var. Tanilar ile yapılan imputasyonun aynısı uygulandı.

- **Feature Engineering**
  - **Alerji** özelliğindeki bilgilerde alerjisi olup olmadığına dair **AlerjiVarMı** şeklinde yeni kolon oluşturuldu.
  - **Tanilar** özelliğindeki verilerde birden fazla tani olduğu ve ',' ile ayırıldığı görüldü. Tüm tanı sayılarını count etmek için **Tani_Sayisi** adında yeni kolon oluşturuldu.

    - **UygulamaYerleri** özelliği de tanilar ile aynı şekilde olduğu için count etmek için **UygulamaYeri_Sayisi** adında yeni kolon oluşturuldu.
    - **KronikHastalik** özelliği de count edilebilir olduğu için **KronikHastalik_Sayisi** şeklinde yeni kolon oluşturuldu.
    - Dakika/Seans oranı (yoğunluk) model aşamasında yardımcı olabileceği için **TedaviYogunlugu** adında yeni kolon oluşturuldu.
    - Hastaların duplike değerlerinden dolayı ne kadar ziyaret ettiklerini görmek için **ZiyaretSayisi** adında yeni kolon oluşturuldu.
    - Tedavi süresine etki edebileceği için yaşları gruplamak amacıyla **YasGrubu[Genc,Yetiskin,Yasli]** şeklinde kolonlar oluşturuldu.

- **Encoding Data**
    - **Cinsiyet ->** LabelEncoder() ile encode edildi.
    - **YasGrubu kolonları ->** OneHotEncoder tekniği ile encode edildi.
    - **TedaviAdi** ve **Tanilar kolonları ->** Target Encoding ile encode edildi. Sebebi ise tedavi süresi ile doğrudan ilişkili olabilmeleri.
    - **Bolum ->** Dağılımı incelenmişti. Çok büyük oranda FTR değeri olduğu için, FTR_mi adında yeni kolon açılıp binary encode edildi

- **Final**
    - Kategorik kolonlar drop edildi.
    - Son bir kez veri seti özellikleri kontrol edildi.

Bu çalışmada veri seti üzerinde kapsamlı bir keşif analizi (EDA) yapılmış, eksik değerler uygun yöntemlerle doldurulmuş ve hem sayısal hem de kategorik değişkenler temizlenip standart bir forma getirilmiştir. Ayrıca, modelin başarısını artırmak için bazı ek açıklayıcı özellikler üretilmiştir. Son aşamada encoding işlemleri uygulanarak veri seti, tutarlı, temiz ve analiz edilebilir bir hâle getirilmiştir. Böylece, hedef değişken olan **TedaviSuresiNumeric** üzerinden makine öğrenmesi modelleri geliştirmeye uygun, hazır bir veri seti oluşturulmuştur.