**VERİ ÖN İŞLEME**

**MÜLEYKE ÖNAL**

**2200329096**

**CEM MOLLA**

**2200329082**

Ara sınav ödevi

IST405-İSTATİSTİKSEL VERİ MADENCİLİĞİ

* **metin, ekran görüntüsü, yazılım içeren bir resim

  Açıklama otomatik olarak oluşturulduİlk olarak veriyi okutalım ve değişkenlerin tipleri hakkında bilgi edinelim.**

**metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Burada ilk 10 değişkenin veri tipleri ve gözlem sayıları hakkında bilgi edinmiş olduk.

* **Bu bilgiler ile analiz için anlamsız olduğunu düşündüğümüz, %30’dan fazla eksik değer içeren ve ilgisiz değişkenleri veriden çıkaralım. Yeni oluşturduğumuz değişkeni ekleyelim.**

**metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

* **metin, ekran görüntüsü, yazı tipi içeren bir resim

  Açıklama otomatik olarak oluşturulduKategorik değişkenleri ‘category’ olarak tanıtalım ve düzeylerini belirleyelim.**
* **metin, ekran görüntüsü, yazı tipi içeren bir resim

  Açıklama otomatik olarak oluşturulduVeriye baktığımızda bazı değişkenlerin birkaç değerinin ‘na’ olduğunu görüyoruz. Bunu Python ‘string’ olarak algılıyor ve aslında sayısal olan bir değişkeni eksik gözlemler içerdiği için ‘object’ olarak tanımlıyor. Bunun olmasını engellemek için ‘na’ değerlerini ‘nan’ olarak değiştirelim ve eksik gözlemler olduklarını tanımlayalım.**

**metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Şimdi veri türlerini doğru şekilde belirlemiş olduk.

* metin, ekran görüntüsü, yazı tipi, menü içeren bir resim

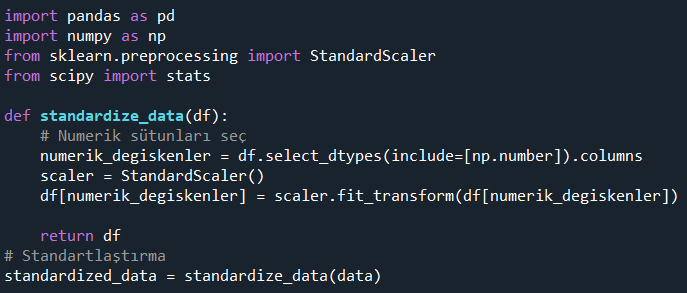
  Açıklama otomatik olarak oluşturuldu**Eksik gözlem olup olmadığını varsa kaç tane olduğunu kontrol edelim ve eksik gözlem bulunduran değişkenlerin eksik kısımlarını değişkenin medyan değerini kullanarak dolduralım.**

metin, ekran görüntüsü, yazı tipi içeren bir resim

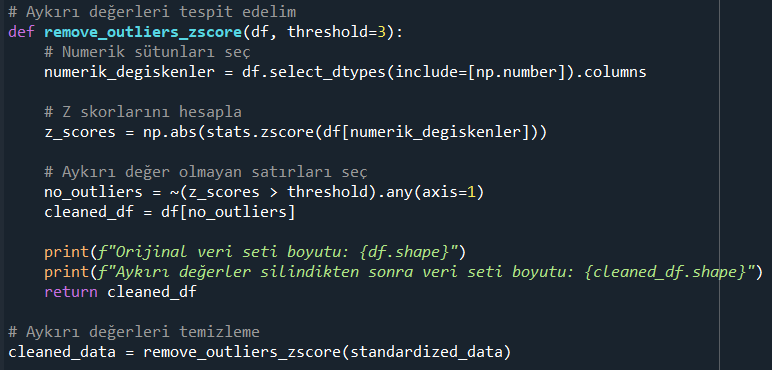
Açıklama otomatik olarak oluşturuldu

Burada eksik gözlem bulunduran değişkenleri belirleyip eksik kısımları medyan değerleri ile doldurduktan sonra hiçbir değişkende eksik gözlem kalmadığını görüyoruz.

* **Şimdi veriyi standartlaştıralım ve aykırı değer incelemesi yapalım.**



Sadece numerik değişkenleri seçerek standartlaştırma uyguladık.



Burada yeni oluşturduğumuz aykırı değerlerden arındırılmış olan cleaned\_data, 416 tane gözlemden oluşuyor.

Aykırı değerleri veriden çıkarmış olduk.

ekran görüntüsü, metin, renklilik, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturulduSonucunda elde edilen data bu şekilde gözüküyor.

* **Şimdi kategorik değişkenleri sayısallaştırmak için one hot encoding uygulayalım.**

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

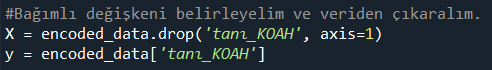
ekran görüntüsü, kare, kalıp, desen, düzen, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturulduBunun sonucunda elde ettiğimiz veri şöyle gözüküyor.

Şimdi numerik değişkenleri standartlaştırdık ve kategorik değişkenleri de sayısallaştırdık. Artık analize başlayabiliriz.

**ANALİZ**

* **Bağımlı değişkeni belirleyerek veriden çıkaralım.**

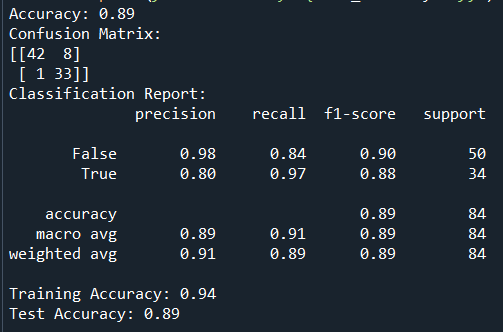
Böylelikle X sadece bağımsız değişkenleri içeriyor oldu. y ise bağımlı değişkenimiz olarak belirlendi.

* **Şimdi eğitim kümesi ile test kümesini ayıralım ve lojistik regresyon modelini oluşturalım.**

metin, ekran görüntüsü, yazı tipi, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

Eğitim kümesi ile test kümesini ayırdık ve modeli oluşturarak tahmin yaptık.

* **Oluşturduğumuz test kümesi için kurduğumuz lojistik regresyon modelini değerlendirelim**.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

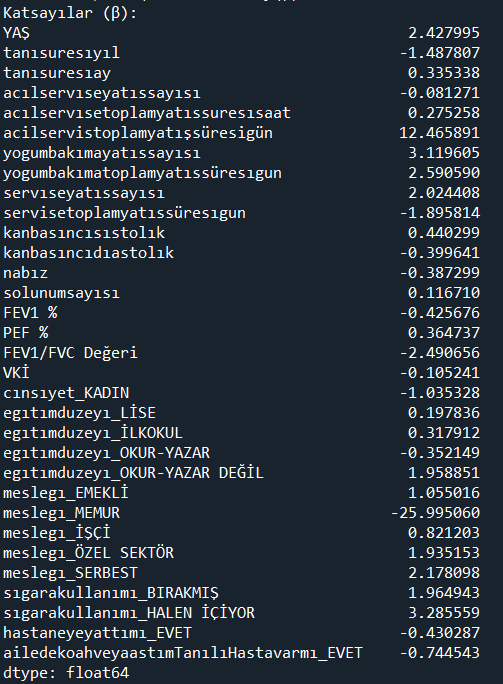
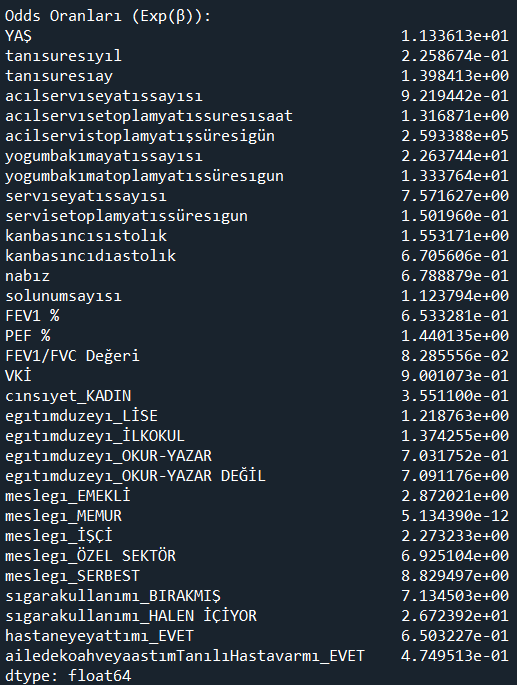
Confusion matrix’ e baktığımızda True-Pozitive 42 False-Negative 33 olarak bulunmuş. Yani modelin hata oranı düşük olduğunu söyleyebiliriz.

Accuracy değerine baktığımızda eğitim kümesinin %94 oluşturduğumuz test kümesinin ise %89 olduğunu söyleyebiliriz.

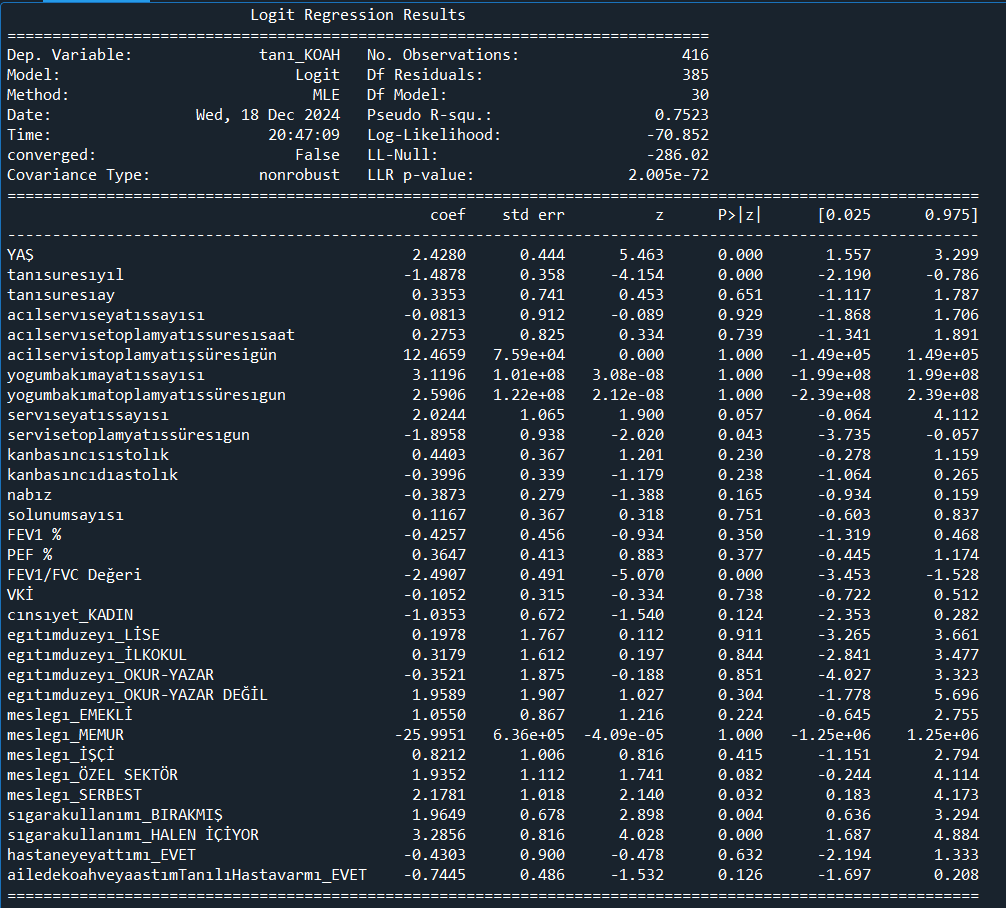
* **Model varsayımlarını inceleyelim.**

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu



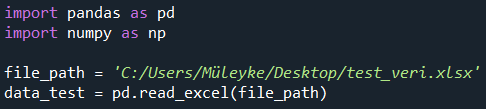
Lojistik regresyon modeline ait katsayıları ve odds oranlarını bu şekilde elde etmiş olduk.



Buradan da modeldeki bağımsız değişkenlerin bağımlı değişkeni açıklama oranının (R2) %75 olduğunu söyleyebiliriz.

* **Oluşturduğumuz bu modeli kaydedelim.**
* **metin, ekran görüntüsü, yazı tipi içeren bir resim

  Açıklama otomatik olarak oluşturulduTest kümesini yükleyelim ve ön işleme adımlarını uygulayalım.**

****

* **Anlamsız değişkenleri veriden çıkaralım ve yeni oluşturduğumuz değişkeni veriye ekleyelim.**

**metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

* **metin, ekran görüntüsü, yazı tipi içeren bir resim

  Açıklama otomatik olarak oluşturulduKategorik değişkenleri tanımlayalım ve düzeylerini belirleyelim.**
* **metin, ekran görüntüsü içeren bir resim

  Açıklama otomatik olarak oluşturulduVerideki bazı değişkenler numerik olsa bile bu değişkenlerin bazı değerleri string olarak girilmiş. Bu değerleri belirleyerek düzelttikten sonra ‘na’ olarak okutulan eksik gözlemleri ‘nan’ a çevirelim ve eksik kısımları dolduralım.**

Burada yaş değişkenindeki eksik gözlemler çok fazla olduğu için eğitim kümesinin medyan değeri ile doldurduk.

* **Veriyi standartlaştıralım.**

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Burada test kümesindeki numerik değişkenleri standartlaştırmış olduk.

* **Kategorik değişkenleri belirleyip one-hot encoding uygulayalım.**

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

* **Test kümesi için tahmin yapalım.**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturulduBunun sonucunda bize verilen test kümesine modelimizi uyguladık ve tanı değişkenini tekrardan veriye ekledik.**