

**İST405- İSTATİSTİKSEL VERİ MADENCİLİĞİ DERSİ VİZE ÖDEVİ**

**HAZIRLAYAN: HANDE NUR BANUŞ 2210329067**

**DERS SORUMLUSU: Dr. ONUR TOKA**

18.12.2024

**Veri seti hakkında bilgi**

* Anonimleştirilmiş hasta değişkenleri için astım ve KOAH tanısını sınıflandırmak ve test kümesinde tahminde bulunmak çalışmanın amacıdır. Bağımlı değişken TANI değişkenidir.

Gerekli kütüphaneler kullanıma hazır hale getirilir.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

LabelEncoder: Kategorik verileri sayısal değerlere dönüştürmek için kullanılır.

preprocessing: Veriyi ölçeklendirme, normalizasyon gibi ön işleme araçlarını içerir.

train\_test\_split: Veriyi eğitim ve test setlerine ayırır.

GridSearchCV: Model için en iyi hiperparametreleri belirlemek amacıyla çapraz doğrulama yapar.

cross\_val\_score: Modelin doğruluğunu çapraz doğrulama ile değerlendirir.

confusion\_matrix: Modelin tahmin performansını gösteren hata matrisi oluşturur.

accuracy\_score: Doğruluk oranını hesaplar.

classification\_report: Precision, recall ve F1-score gibi sınıflandırma performans metriklerini sağlar.

roc\_auc\_score: ROC eğrisi altında kalan alanı hesaplar (AUC).

roc\_curve: Modelin ROC eğrisini çizmek için kullanılacak değerleri hesaplar.

LogisticRegression: Lojistik regresyon modeli oluşturur.

filterwarnings: Uyarıları filtrelemek veya kapatmak için kullanılır.

os: Dosya ve dizin işlemlerini yapmak için kullanılır.

numpy: Sayısal hesaplamalar için kullanılan bir kütüphane.

pandas: Veri analizi ve işleme için kullanılan bir kütüphane.

seaborn: İstatistiksel veri görselleştirmeleri için kullanılır.

pyplot: Grafik çizim işlemleri için kullanılan matplotlib modülü.

missingno: Eksik veri analizi ve görselleştirmesi yapmak için kullanılan bir kütüphane.

msno: Eksik verileri analiz etmek için kullanılan missingno kütüphanesi.

date: Tarih işlemleri yapmak için kullanılan datetime modülü.

LocalOutlierFactor: Aykırı değerleri algılamak için kullanılan yöntemlerden biri.

MinMaxScaler: Verileri belli bir aralıkta ölçeklendirmek için kullanılır.

LabelEncoder: Kategorik değişkenleri sayısallaştırır.

StandardScaler: Verileri standart normal dağılıma dönüştürür.

RobustScaler: Aykırı değerlere karşı dayanıklı bir ölçeklendirme yapar.

xlrd: Excel dosyalarını okuma işlemleri için kullanılan bir kütüphane.

pd.set\_option('display.max\_columns', None): DataFrame’in tüm sütunlarını görüntülemeye olanak tanır.

pd.set\_option('display.max\_rows', None):DataFrame’in tüm satırlarını görüntülemeye olanak tanır.

pd.set\_option('display.width', 500): DataFrame çıktısının genişliğini artırarak daha okunabilir hale getirir.

filterwarnings('ignore')

Program sırasında çıkan uyarıları gizler.

VERİ ÖN İŞLEME VE VERİ MANİPÜLASYONU

Test ve eğitim veri kümelerini okuttuktan sonra bu kümelerdeki “NA", " na ", "N/A" veya boşluk gibi eksik veri göstergesi olarak düşünülen değerleri NaN ile değiştirelim.

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

Açıklama otomatik olarak oluşturuldu

Check\_df isimli fonksiyon kurup elimizdeki test ve train veri setlerine bir bakış atalım.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Sigara bırakma süresi (gün):** Ortalama **1396.95 gün** (~4 yıl) ancak standart sapma **2831.99 gün**, yani bireyler arasında büyük bir varyasyon var. Bazı bireyler yeni bırakmışken, bazıları çok uzun süre önce bırakmış.

* **Bırakmadan önce günde içilen sigara sayısı:** Ortalama **11.67 adet**, ancak yine standart sapma yüksek (**17.25**). Bu, bireyler arasında sigara tüketimi alışkanlıklarının oldukça farklı olduğunu gösteriyor.
* **Halen sigara içenler (günlük sigara sayısı):** Ortalama **3.07 adet**. Çoğu birey ya bırakmış ya da çok az sigara içiyor.

**Kan basıncı (sistolik ve diyastolik):**

* Ortalama sistolik basınç: **125.33 mmHg**, diyastolik basınç: **77.06 mmHg**.
* Sistolik basınç için standart sapma **17.86**, bu değer sağlıklı aralıklarda bir dağılım olduğunu gösteriyor. Ancak diyastolik basınçta **minimum değer 0**, bu büyük olasılıkla eksik veri veya hata.

**Acil servis yatışları:**

* Ortalama yatış sayısı: **1.42**. Çoğu bireyin nadiren acil servise başvurduğunu gösteriyor.
* Ortalama toplam yatış süresi: **3.58 saat**, yani bireyler genellikle kısa süreli kalmışlar.

**Yoğun bakım yatışları:**

* Ortalama yatış süresi **0.13 gün** (yaklaşık 3 saat). Bu, ciddi sağlık sorunlarının genelde az olduğunu gösteriyor.

**Servis yatışları:** Ortalama yatış süresi **4.41 gün**, bu diğer yatışlara kıyasla daha uzun ve bireylerin serviste tedaviye ihtiyaç duyduğunu gösteriyor.

**FEV1 (Zorlanmış Ekspirasyon Hacmi 1 saniye):** Ortalama **1.76 litre**, standart sapma: **0.76 litre**. Değerler genel olarak solunum fonksiyonlarında hafif ila orta dereceli bir bozulmaya işaret ediyor olabilir.

**PEF (Pik Ekspirasyon Akımı):** Ortalama **3.85 l/s**, ancak **standart sapma 7.85 l/s** oldukça yüksek. Bu, bireyler arasında büyük farklılıklar olduğunu gösteriyor.

**FEV1/FVC:** Ortalama **66.93%**, bu değer KOAH (Kronik Obstrüktif Akciğer Hastalığı) veya diğer solunum bozuklukları açısından izlem gerekliliğini işaret edebilir.

**Yaş:** Ortalama **53.11 yıl**, bireyler ağırlıklı olarak orta yaş grubunda.

**Boy ve ağırlık:** Ortalama boy: **163.83 cm**, ağırlık: **73.19 kg**. Bu değerler genel popülasyonun ortalamalarına yakın.

metin, ekran görüntüsü, doküman, belge, menü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Beklenildiği gibi tanı bağımlı değişkeni test kümesinde yok.

Tip bilgilerinde yanlışlık var. Elimizdeki bilgilere göre kategorik olması gerekenleri ‘object’ numerik olması gerekenleri ‘numeric’ veri tipine çevirelim.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, menü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Test ve train veri kümelerinin tip bilgileri düzeldi.

Kategorik değişkenler “object”, numerik değişkenler ise “float” ve “int” veri tipine dönüştü.

EKSİK DEĞER ANALİZİ

metin, ekran görüntüsü, yazılım, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bu kod, bir veri setindeki eksik değerleri (NaN) analiz eder ve eksik değer içeren sütunları ve oranlarını bir tablo olarak gösterir. Ayrıca, isteğe bağlı olarak eksik değer içeren sütunların isimlerini döndürebilir. Hangi değişkenlerde ne kadar eksik değer var ve bunların yüzdelerini görüyoruz. Eksik değerleri görselleştirelim:

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

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, çizgi, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, çizgi, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

EKSİK DEĞER DOLDURMA

Test ve train kümesinin kopyasını oluşturup doldurma işlemlerini buradan yapacağız.



Sigara kullanımı:

Hiç içmemiş:1

Bırakmış:2

Halen içiyor:3

metin, ekran görüntüsü, yazı tipi, doküman, belge içeren bir resim

Açıklama otomatik olarak oluşturulduSigara kullanımı kategorik değişkeninin bilgisine göre; “sıgarayıbırakannekadarGÜNıcmıs” , “sıgarabırakangundekacadetıcmıs”, “nezamanbırakmısGÜN”,“sıgarayadevamedengundekacadetıcıyo” sayısal değişkenleri şekillenir. Yani hiç sigara içmeyen biri için bu sayısal değişken bilgileri 0 olmalı. Sigarayı bırakmış ve hiç içmeyen birinin “sigaraya devam eden günde kaç adet içiyor “bilgisinde sıfır yazmalı. Verimizde de bu değişkenlerdeki eksiklikler bu mantıktan dolayı. Bu değişkenlerin her biri için özel olarak doldurma işlemleri yapmalıyız.

DataFrame’ den satır ve sütün seçme işlemleri için loc ve iloc kullanılır. Loc, hem satır hem de sütun için **isim veya etiketleri** kullanır.

“sıgarayıbırakannekadarGÜNıcmıs “: Hiç içmemiş(1) ve halen içen(3) kişiler için 0 olmalı. Bırakan(2) kişiler için değerler içerir.

“sıgarabırakangundekacadetıcmıs”: Hiç içmemiş(1) ve halen içen(3) kişiler için 0 olmalı. Bırakan(2) kişiler için değerler içerir.

“nezamanbırakmısGÜN”: Hiç içmemiş(1) ve halen içen(3) kişiler için 0 olmalı. Bırakan kişiler(2) için değerler içerir.

“sıgarayadevamedengundekacadetıcıyo”: Hiç içmemiş(1) ve bırakmış(2) kişiler için 0 olacak. Hala içen (3) kişiler için eğerler içerir.

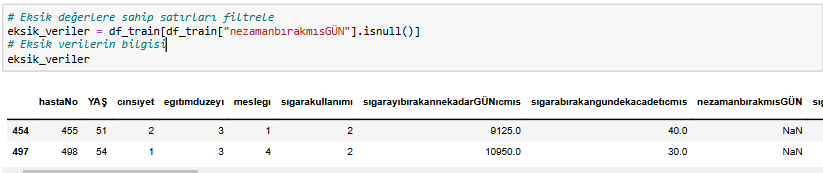
metin, ekran görüntüsü, sayı, numara, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

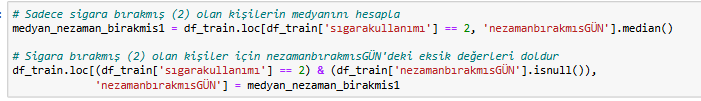
metin, elektronik donanım, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

"nezamanbırakmısGÜN" değişkenini doldurmamıza rağmen eksik 2 değer kalmış. Detaylı incelediğimizde sigarayı bırakan (2) durumundaki kişilere ait.



Buradaki eksiklikleri bırakan kişilerin ne zaman bıraktığı bilgisi ile dolduralım.



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

Açıklama otomatik olarak oluşturuldu

Kalan eksik değer bilgileri de bu şekilde. Ailede koah veya astım yok ise anne, baba, kardeş ve diğerlerinde yok (0) yazmalı. Aynı şekilde ailede koah veya astım varsa ama NaN yerleri kendimiz dolduramayız yani o değişkene ait özellik bilgisi yok.

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

Açıklama otomatik olarak oluşturuldu

Bütün eksik değerlerden kurtulmuş olduk. Ancak bazı satırlarda ailede koah veya astım var(2) olmasına rağmen ‘varsakimde’ değişken bilgilerinde yok(0) gözüküyor.

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

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

'varsakımde ANNE', varsakımde BABA', 'varsakımde KARDES', 'varsakımde DİGER' sütunlarının 0 olması ailede hasta kimse olmadığını gösterir. O yüzden bu değişkenler sıfır olunca 'ailedekoahveyaastımTanılıHastavarmı' değişkeni 1 olacak. Train kümesindeki eksiklikler bitti. Test kümesindeki eksik değerleri inceleyelim.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Train kümesinde yaptığımız mantıkla test kümesindeki eksiklikleri dolduralım.

metin, ekran görüntüsü, sayı, numara, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, doküman, belge, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

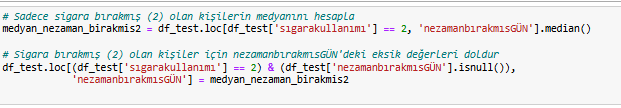
metin, ekran görüntüsü, sayı, numara, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldu

Sayısal değişkenleri medyanla dolduralım. "nezamanbırakmısGÜN" değişkenini doldurmamıza rağmen eksik 4 değer kalmış. Detaylı incelediğimizde sigarayı bırakan (2) durumundaki kişilere ait. Sigarayı bırakan kişi bilgisine göre doldurma yapalım.

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

Açıklama otomatik olarak oluşturuldu



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

Açıklama otomatik olarak oluşturuldu

metin, yazı tipi, çizgi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

varsakimde KARDEŞ VE DİĞER bilgisi 1 yani var olarak girildiyse bu eksik değer "2" yani ailede koah veya astım tanılı biri var. 59. indekse denk geliyor.



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

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bazı yanlışlıklar var. Bazılarının da bilgisi yok. Eğer ailede astım veya koah tanılı biri yoksa (1) varsakimde değişkenleri 0 olmalı. Önce verilerdeki tutarsızlıkları train kümesine yaptığımız mantıkla yapmaya çalışalım.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tutarsızlıkları düzeltmeye çalıştık. Ancak eksik değer bilgisi fazla. Bu çıktıda olduğu gibi ailede astım veya koah var (2) ama kimde var olduğu bilgisi yok. Ben bu bilgileri iş bilgisi olmadan kendim dolduramam. O yüzden bu eksiklikleri “bilgiyok” şeklinde dolduralım.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Hem tutarsızlıklardan kurtulduk hem de eksik gözlemlerden kurtulduk.

**Hedef değişken analizi:**

Hedef değişkenin dağılımı, veri setimizdeki "tanı" değişkeninin dağılımını analiz etmek, hem dengesiz veri problemi olup olmadığını tespit etmemizi sağlar, hem de modelimizin bu dağılımı ne kadar iyi tahmin edebileceğine dair bize bir fikir verir. Eğer hedef değişkenimiz dengesizse (örneğin, bir sınıf diğerlerine göre çok daha fazla gözleme sahipse), modelimiz bu dengesizliği öğrenebilir ve yanlış tahminlerde bulunabilir.

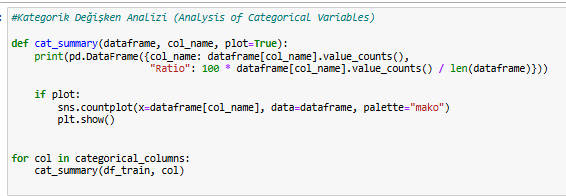
metin, ekran görüntüsü, yazı tipi içeren bir resim

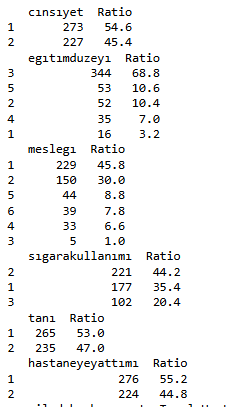
Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Hedef değişkenimiz dengeli dağılıyor.

**Kategorik Değişken Analizi**



metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Her bir kategorik değişken için grafiklerini görmek istersek plot=true yaparız. Bu şekilde her bir kategorik değişkenin hangi kategorilerinde kaç tane değer var ve bu değerlerin yüzdesel dağılımlarını görüyoruz. Örneğin, ailede koah hastası olan kişi sayısı (2) 163, yüzde olarak söylersek veri kümemizde ailesinde koah olanların oranı %32,6’dır.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Mesleği 1 olanlar yani işsiz olanlar veride daha çok var, mesleği 3 olanlar yani memur olanlar ise en az bulunan meslek grubu olmuş.

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Hastaneye yatmayanlar (1) hastanede yatanlara(2) göre daha çok görülmüş.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Annede astım veya koah görülmeme oranı (0) görülmesine göre daha fazla.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Babada astım veya koah olmaması olanlara göre daha fazla.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, diyagram, ekran, görüntüleme içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Test kümesinde ise bu şekilde yine her değişken için grafik çizilir.

**NUMERİK DEĞİŞKEN ANALİZİ**

Her bir numerik değişken için tanımlayıcı istatistik bilgileri ve histogram grafikleri çizilir. Birkaç tanesi için bakalım:

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Ortalama yaş **53.10** civarında. Veri setindeki yaşların çoğu orta yaş grubuna (50-60) odaklanmış. En genç birey **18 yaşında**, en yaşlı birey **87 yaşında**.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Grafik, tanı süresinin çoğunlukla kısa yıllar (0-10 yıl) arasında yoğunlaştığını ve yüksek tanı sürelerinin nadir olduğunu göstermektedir ve veri sağa çarpıktır.

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Burda ise veri full sıfırda toplanmış. Bütün değişkenlere bir arada bakalım.



metin, diyagram, plan, teknik çizim içeren bir resim

Açıklama otomatik olarak oluşturuldu

Değişkenlerimiz çarpık dağılmışlar ve bazılarıda hep sıfırda toplanmış. Sıfır olanları ileride veri setinden drop edeceğim. Çarpıklar için ise standartlaştırma yapacağım.

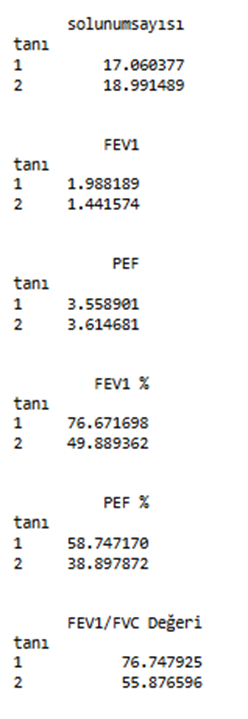
NUMERİK DEĞİŞKENLERİN HEDEF DEĞİŞKENE GÖRE ANALİZİ

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

Açıklama otomatik olarak oluşturuldu

Birkaç tanesine bakalım.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu****

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

Açıklama otomatik olarak oluşturuldu**

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Tanı 1 yani astım olup servise yatanların sayılarının ortalamaları 0.188miş.

Tanı 2 yani koah olup acil servise toplam yatış süresi saatinin ortalaması 5.13imiş, şeklinde çıkarımlar yapılabilir.

KATEGORİK DEĞİŞKENLERİN HEDEF DEĞİŞKENE GÖRE ANALİZİ

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

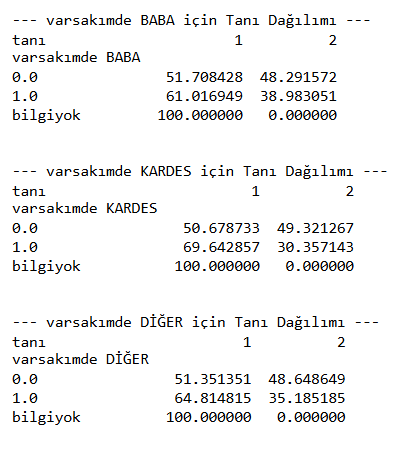
Birkaç çıktıyı inceleyelim.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

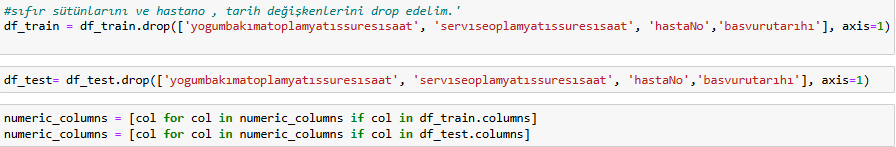
Açıklama otomatik olarak oluşturuldu

Cinsiyeti erkek (1) olup tanı astım (1) olanların oranı 25.64, koah olan erkek oranı ise 74.35’tür.

Sigarayı bırakan (2) kişilerden koah (2) olanların oranı 72.85miş.



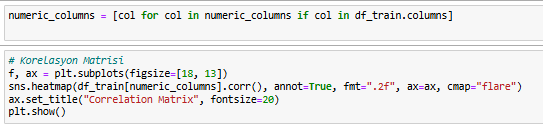
Varsakimde baba, kardeş, diğer değişkenlerinde bulunan “bilgiyok” çıkarımlarının tamamı astım olanlardanmış.



Full sıfırlardan oluşan değişkenleri, test ve trainde ortak olanları veri kümesinden düştüm. Kardinal değişken olan ‘hastaNo’ değişkeni ve tarih sütununu da ileride model kurarken zorlanmamak için sildim.

**KORELASYON ISI HARİTASI**

metin, ekran görüntüsü, kalıp, desen, düzen, renklilik içeren bir resim

Açıklama otomatik olarak oluşturuldu

* YAŞ ve FEV1: Korelasyon katsayısı yaklaşık -0.49. Bu, yaş arttıkça FEV1 değerinin düştüğünü gösterir.
* FEV1/FVC Değeri ile YAŞ: -0.43 korelasyon, yaş arttıkça bu değerin azaldığını işaret ediyor.
* acilservistoplamyatışsüresigün ve acilservıseyatıssayısı: Korelasyon 0.79, yani acil servis yatış sayısı arttıkça toplam yatış süresi de artar.
* yogumbakımyatıssayısı ve acilservıseyatıssayısı: Pozitif bir ilişki (0.18 civarı) var ancak bu ilişki güçlü değil.
* sıgarayıbırakannekadarGÜNıcmıs ve sıgarabırakangundekacadetıcmıs: Pozitif korelasyon (0.75). Daha fazla sigara içenler, bırakmadan önceki alışkanlık süresi açısından da uzun süre sigara içmiş olabilir.
* sıgarayadevamedengundekacadetıcıyo ve nezamanbırakmısGÜN: Aralarında gözle görülür bir pozitif ilişki yok.
* kanbasıncısıstolık ile kanbasıncıdıastolık: Beklendiği gibi yüksek korelasyon (0.67). Sistolik ve diyastolik kan basınçları birlikte artış gösteriyor.
* solunumsayısı ve FEV1/FVC Değeri: Zayıf ve negatif korelasyon (-0.28). Solunum sayısı arttıkça FEV1/FVC değerinin hafifçe düştüğü gözlemlenebilir.
* PEF ve FEV1: Güçlü pozitif korelasyon (0.99). Bu iki değişken akciğer kapasitesini ölçen benzer parametrelerdir.



Minimum korelasyon bu değişkenler arasındaymış.



Beklenildiği gibi en yüksek korelasyon aynı değişkende yani 1 değerlerinde.

**GÖRSELLEŞTİRME**

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

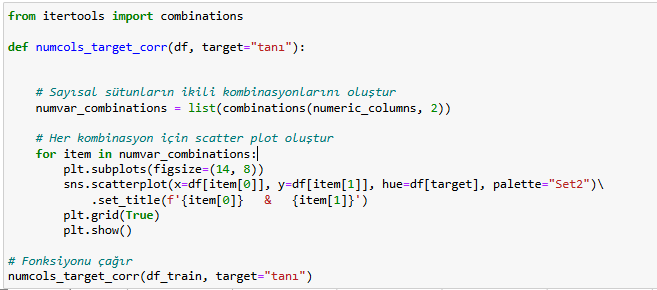
Birkaç tanesini inceleyelim.

metin, diyagram, ekran görüntüsü, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Değişkenlerde aykırı değerler var. Sigarayı bırakan günde kaç adet içmiş değişkenindeyse bırakan kişilerin bazıları sigarayı bıraktıkları dönemde aşırı miktarda sigara tüketmiş olduklarını gösteriyor.

Ne zaman bırakmış GÜN değişken grafiğindeki aykırı değerler **5.000 günün üzerinde** olup, özellikle **20.000 gün civarında** bazı ekstrem örnekler vardır. Bu durum, uzun süre önce sigarayı bırakmış az sayıda bireyin varlığını işaret etmektedir.

****

Birkaç tanesini yorumlayalım:

**metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Yorum:

•Veriler, yaş arttıkça sigara içilen miktarın değişkenlik gösterdiğini ortaya koyuyor.

•Sigara içme miktarının büyük bir kısmı 0-20 adet/gün aralığında toplanmış.

•Yaşı genç bireylerde (20-40 arası), daha düşük sigara tüketimi (0-20 arası) yaygın.

•Yaşı büyük bireylerde (50-80 yaş), bazı bireylerde daha yüksek sigara tüketimi (40-80 adet/gün) görülüyor.

•Genellikle daha düşük sigara tüketimi gösteriyor.

•Veriler çoğunlukla 0-20 adet/gün civarında yoğunlaşmış.

•Sigara içme miktarı daha geniş bir aralığa yayılmış.

•40 yaş üzerindeki bireylerde yüksek tüketim (20-80 adet/gün) yaygın.

•Y ekseninde 80-120 adet/gün gibi uç değerler mevcut.

•Bu kişiler KOAH grubunda (2) daha fazla yer alıyor.

•Astım hastaları daha genç yaşlarda ve daha düşük sigara tüketimi gösteriyor.

•KOAH hastalarında yaş ilerledikçe sigara tüketimi hem daha yüksek hem de daha yaygın hale geliyor.

•Bu grafik, sigara tüketiminin hastalık tanıları ile ilişkili olabileceğini ve yaş faktörünün KOAH tanısında önemli bir rol oynayabileceğini düşündürüyor.

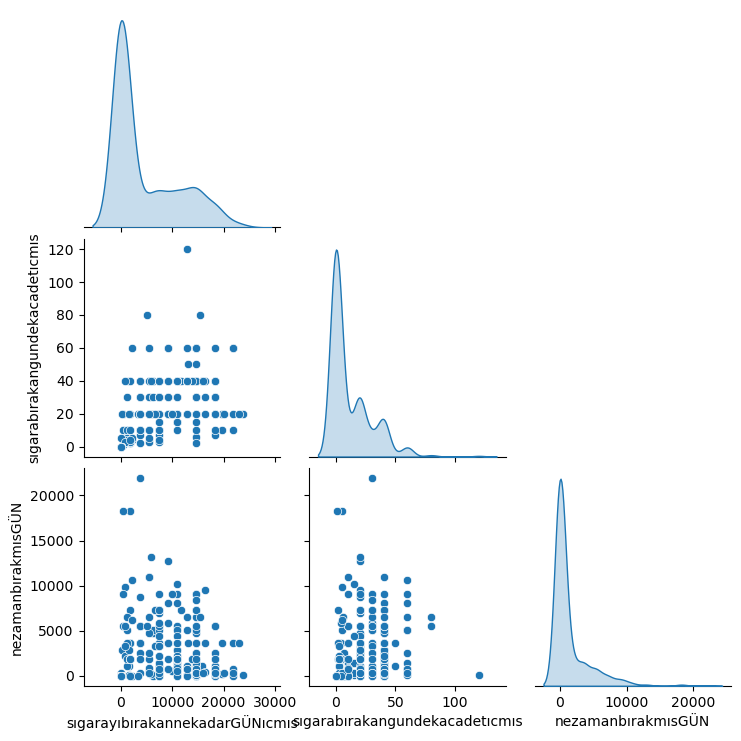
**metin, ekran görüntüsü, sayı, numara, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

* Yatay eksende (sigarayibırakankadarGÜNcmıs): Çok büyük değerler mevcut (5000’den 20000’e kadar). Bu, bazı bireylerin çok uzun süre sigara içip bıraktığını gösteriyor.
* Dikey eksende (yoğunbakımayatıssayısı): Büyük çoğunlukta 0 değerine yakın yoğunlaşma görülüyor. Bu da birçok kişinin yoğun bakımda yatışının olmadığını ya da çok düşük olduğunu gösteriyor.
* KOAH olan bireylerde yoğun bakımda yatış sayısının artma eğilimi daha fazla gibi görünüyor.
* Astım olan bireylerde, yatış sayısı daha çok 0 seviyesinde yoğunlaşmış durumda.
* "sigarayibırakankadarGÜNcmıs" değişkeninde çok yüksek değerler mevcut (>10.000 gün gibi). Bu değerler aykırı gözükebilir ve analizinize etki edebilir.
* Görselde net bir trend ya da korelasyon gözlemlenmiyor. Sigara bırakma süresi arttıkça yoğun bakım yatış sayısında belirgin bir artış ya da azalış gözükmüyor. Ancak, "tanı=2:koah" sınıfı bireyler için daha yüksek yatış sayılarına rastlanmış.

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

Açıklama otomatik olarak oluşturuldu**



Bu çapraz dağılım grafiği (pair plot) üç değişken arasındaki ilişkileri ve tek değişkenli dağılımları göstermektedir:

1.sıgarayıbırakannekadarGÜNıcmıs (Sigara bırakalı geçen gün sayısı)

2.sıgarabırakangundekacadetıcmıs (Sigara bırakmadan önce günde içilen sigara adedi)

3.nezamanbırakmısGÜN (Sigara ne zaman bırakılmış, gün cinsinden).

Tek Değişkenli Dağılımlar (Diyagonal Hücreler):

1.sıgarayıbırakannekadarGÜNıcmıs:

•Çoğu değer 0-10.000 gün arasında yoğunlaşmış.

•Az sayıda bireyde çok yüksek değerler (~20.000-30.000 gün) gözlemleniyor.

•Sağ çarpık bir dağılım mevcut.

2.sıgarabırakangundekacadetıcmıs:

•Sigara bırakmadan önce içilen sigara adedi 0-20 adet/gün aralığında yoğunlaşmış.

•50 adet ve üzeri yüksek değerler nadir fakat dikkat çekici.

3.nezamanbırakmısGÜN:

•Değerlerin çoğu 0-5000 gün aralığında yığılmış durumda.

•Sağ çarpık dağılım belirgin. Sigara bırakma süreleri birkaç yıl ile sınırlı olan bireyler çoğunlukta.

İkili Dağılımlar (Alt Üçgen Hücreler):

1.sıgarayıbırakannekadarGÜNıcmıs & sıgarabırakangundekacadetıcmıs:

•Bu iki değişken arasında net bir ilişki gözlenmiyor.

•Sigara bırakma süresi arttıkça, sigara bırakmadan önce içilen sigara sayısında bir değişim ya da tutarlı bir trend gözlenmiyor.

•Dağılım oldukça yaygın ve düzensiz.

2.sıgarayıbırakannekadarGÜNıcmıs & nezamanbırakmısGÜN:

•İki değişkenin yüksek bir doğrusal ilişkisi olabilir.

•Verilerin çoğu sol alt köşede yoğunlaşmış. Sigara yakın zamanda bırakılmış ya da hiç bırakılmamış bireyler çoğunlukta.

•Bazı bireylerin sigarayı çok uzun süre önce bırakmış olması nadir fakat dikkat çekici.

3.sıgarabırakangundekacadetıcmıs & nezamanbırakmısGÜN:

• Sigara bırakma zamanı ile sigara içilen adet arasında belirgin bir ilişki gözlenmiyor.

•Sigara ne kadar süre önce bırakılmış olursa olsun, sigara tüketim miktarı geniş bir aralıkta dağılmış durumda.

•Ancak sigara tüketiminin genellikle 0-20 adet/gün aralığında yoğunlaştığını görüyoruz.

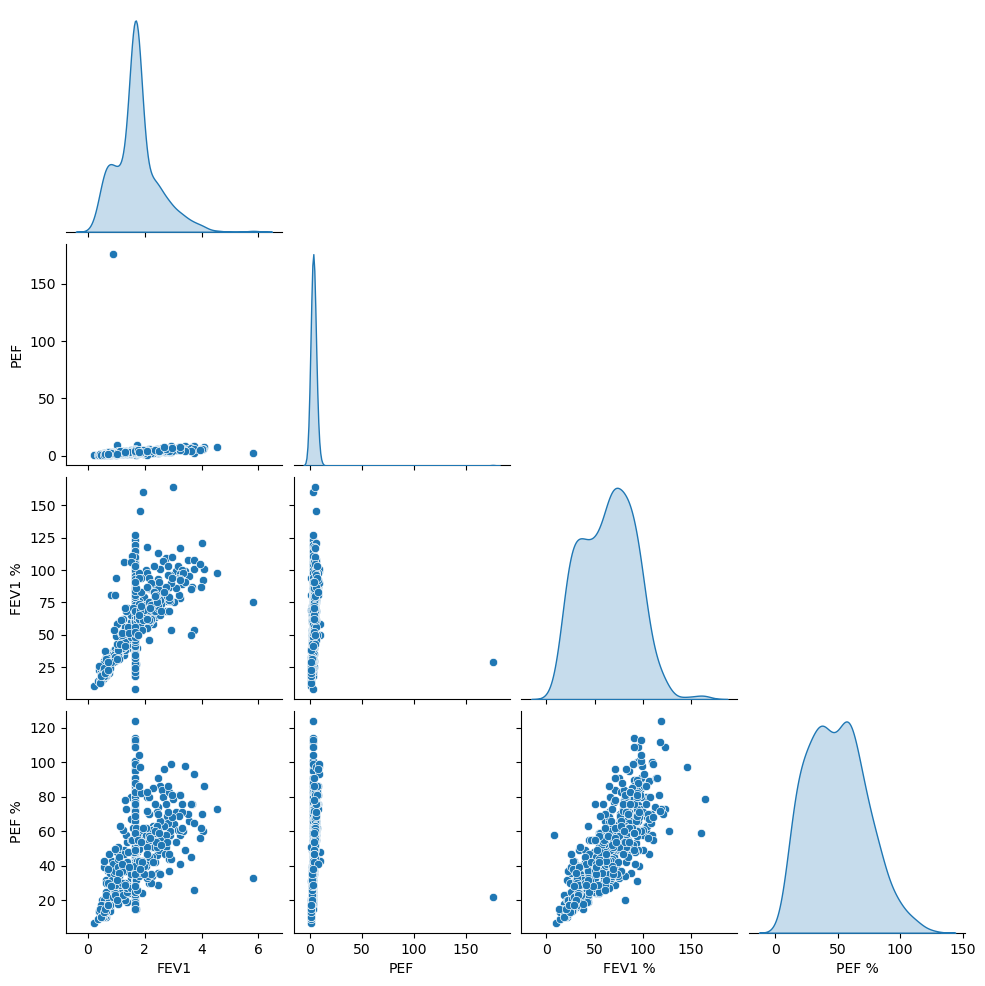
•Sigara bırakma süresi (sıgarayıbırakannekadarGÜNıcmıs) ve sigara bırakma zamanı (nezamanbırakmısGÜN) arasında doğal bir bağlantı mevcut. Ancak yüksek değerlerin az sayıda bireyde gözlemlenmesi sağ çarpıklığa yol açıyor.

•Sigara tüketim miktarı (sıgarabırakangundekacadetıcmıs) ile bırakma süresi veya bırakma zamanı arasında belirgin bir korelasyon gözlenmiyor.

•Çoğu birey, sigara bırakmadan önce düşük-orta düzeyde sigara tüketmiş ve yakın zamanlarda sigarayı bırakmış durumda.

•Uç değerler (yüksek sigara tüketimi veya çok uzun süre sigara bırakma süreleri) azınlıkta ve bu durum grafikte açıkça görülebiliyor.

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

Açıklama otomatik olarak oluşturuldu

1.Diyagonal Hücreler: Tek Değişkenli Dağılımlar

•FEV1, PEF, FEV1 %, PEF % değişkenlerinin her biri için dağılım grafikleri çizilmiş.

•FEV1 % ve PEF %: Bu iki değişkenin dağılımları sağa çarpık görünüyor. Veriler genelde 0-100 arasında yoğunlaşmış.

•PEF: PEF değişkeninin dağılımı büyük oranda küçük değerlere odaklanmış ancak birkaç uç değer (outlier) mevcut.

•FEV1: Değerler daha sıkışık bir aralıkta yoğunlaşmış, düşük ve orta seviyelerde daha fazla gözlem var.

2.İki Değişken Arası Scatter Plot (Alt Üçgen):

•FEV1 % ve PEF % Arasındaki İlişki: Güçlü ve doğrusal pozitif korelasyon gözleniyor. FEV1 % arttıkça PEF % de artıyor. Bu iki değişkenin benzer trend göstermesi beklenen bir durum çünkü her ikisi de akciğer fonksiyonlarını ölçen parametrelerdir.

•FEV1 ve FEV1 % Arasındaki İlişki: Doğrudan güçlü bir ilişki gözükmüyor. Dağılım daha yaygın ve düzensiz görünüyor.

•FEV1 ve PEF Arasındaki İlişki:İlişki zayıf ve yatay bir dağılım söz konusu. PEF için bazı uç değerler verinin yorumlanmasını zorlaştırıyor.

•PEF % ve FEV1 Arasındaki İlişki:PEF % değerleri ile FEV1 arasında orta düzeyde bir korelasyon bulunuyor. Ancak FEV1 düşükken PEF %’de dağınıklık daha belirgin.

3.Outlier (Uç Değerler):

•PEF ve FEV1 değişkenlerinde birkaç uç değer var.Bu uç değerler dağılımların yorumlanmasını zorlaştırabilir. İleride analiz yaparken bu uç değerlerin etkisi incelenecektir.

•FEV1 % ve PEF % arasında güçlü bir doğrusal pozitif ilişki gözlemleniyor.

•PEF değişkeninde uç değerler mevcut.

•Dağılımlar sağa çarpık olduğu için verilerin normal dağılım göstermediği söylenebilir. Bu durum, verilerin log dönüşümü veya standartlaştırma gibi yöntemlerle düzeltilmesini gerektirebilir.

•Grafikten anlaşıldığı üzere, FEV1 % ve PEF % değişkenleri özellikle dikkat edilmesi gereken değişkenlerdir çünkü birbirleriyle güçlü ilişkilidir.

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

Açıklama otomatik olarak oluşturuldu

metin, yazı tipi, ekran görüntüsü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, diyagram, kare içeren bir resim

Açıklama otomatik olarak oluşturuldu

* **tanı = 1 (astım) olan bireyler için:**
  + sigarakullanımı = 1(hiç içmemiş): 160 kişi ile en yoğun grup.
  + sigarakullanımı = 2(bırakmış): 60 kişi.
  + sigarakullanımı = 3(halen içen): 45 kişi.
  + Burada hiç içmemiş olan bireyler astım grubunda baskın.
* **tanı = 2(koah) olan bireyler için:**
  + sigarakullanımı = 2: 161 kişi ile en yoğun grup.
  + sigarakullanımı = 3: 57 kişi.
  + sigarakullanımı = 1: 17 kişi.
  + Bu grupta ise sigarayı bırakan kişiler baskın durumda.

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

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

1.Cinsiyet ve Tanı Arasındaki Dağılım:

•Erkekler (cinsiyet = 1):

•Astım Tanısı (tanı = 1): 70 kişi

•KOAH Tanısı (tanı = 2): 203 kişi

Erkeklerin büyük çoğunluğunun KOAH tanısı aldığı görülmektedir.

•Kadınlar (cinsiyet = 2):

•Astım Tanısı (tanı = 1): 195 kişi

•KOAH Tanısı (tanı = 2): 32 kişi

Kadınların büyük çoğunluğunun astım tanısı aldığı görülmektedir.

2.Renklerin Yorumu:

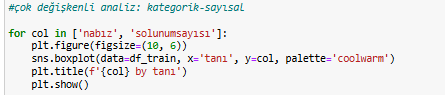
•Renk yoğunluğu sayısal değerlere göre değişiyor. Koyu renkler yüksek değerlere, açık renkler ise düşük değerlere karşılık geliyor.

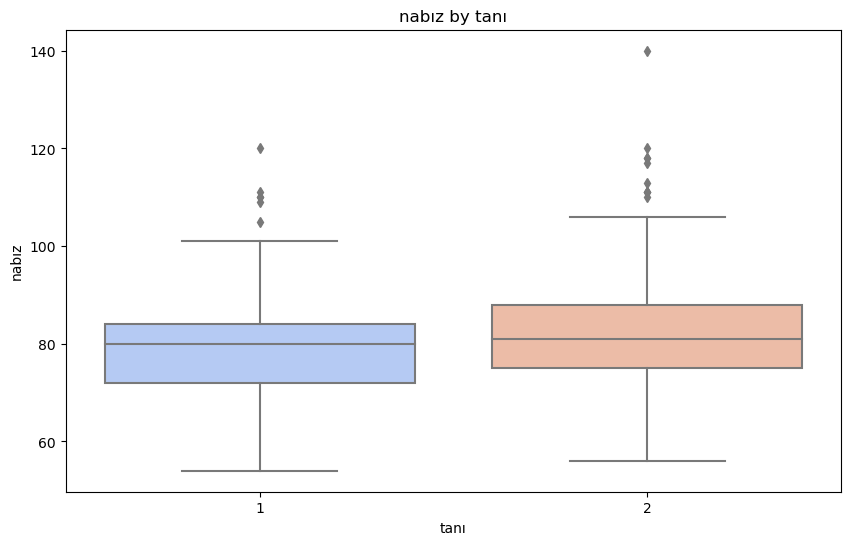
•KOAH tanısı alan erkekler (203 kişi) en yoğun grubu oluşturuyor (koyu renk).

•Astım tanısı alan kadınlar (195 kişi) ikinci en büyük grup olarak dikkati çekiyor.

•KOAH tanısı alan kadınlar (32 kişi) en düşük değere sahip grup olarak görünüyor.

•Erkeklerde KOAH daha yaygınken,Kadınlarda astım daha yaygındır.





1. Genel Karşılaştırma:

•Astım (tanı = 1):

•Nabız değerlerinin medyanı (ortalama değerin merkez çizgisi) 80 civarında.

•Veriler 75. yüzdelik dilimde 100’e kadar uzanıyor ve bazı aykırı değerler 100-120 aralığında bulunuyor.

•Minimum değer 50 civarında gözüküyor.

•KOAH (tanı = 2):

•Nabız değerlerinin medyanı 85 civarında ve Astım grubuna göre daha yüksek.

•Veriler 75. yüzdelik dilimde 120’ye kadar uzanıyor ve aykırı değerler 140’a kadar çıkıyor.

•Minimum değer 60 civarında gözüküyor.

2.Aykırı Değerler:

•Her iki tanı grubu için de yukarı doğru bazı aykırı değerler mevcuttur.

•Astım grubunda aykırı değerler genellikle 100-120 aralığında yoğunlaşmış.

•KOAH grubunda aykırı değerler 120 ve 140 gibi daha yüksek değerlere çıkıyor.

3.Merkezi Eğilim:

•KOAH grubunda nabız değerlerinin ortalaması ve medyanı Astım grubuna göre daha yüksek.

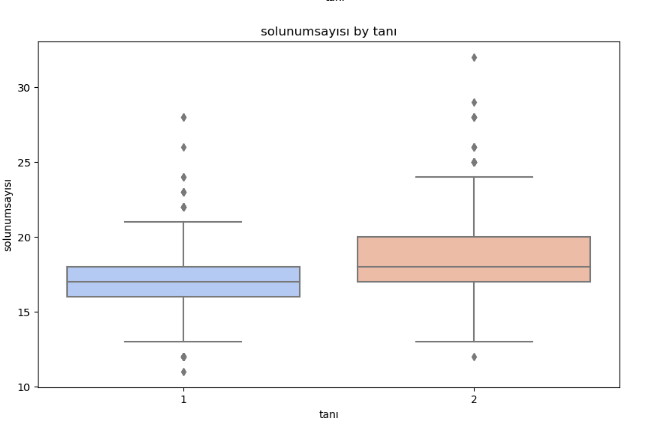
•Bu durum KOAH tanılı bireylerde nabız değerlerinin daha fazla yükselme eğiliminde olduğunu göstermektedir.

4.Değişkenlik:

•KOAH grubunda nabız değerlerinin yayılımı (IQR - interquartile range) daha geniştir.

•Astım grubunda daha dar bir yayılım görülmektedir, yani nabız değerleri bu grupta daha tutarlıdır.

Bu boxplot, KOAH tanılı bireylerde nabız değerlerinin Astım tanılı bireylerden daha yüksek bir medyana sahip olduğunu ve daha fazla değişkenlik gösterdiğini ortaya koymaktadır. KOAH grubunda nabız değerlerindeki yükselme ve aykırı değerler, hastalığın etkileri veya ilgili komplikasyonların bir yansıması olabilir.

****

1.Genel Karşılaştırma:

•Astım (tanı = 1):

•Solunum sayısının medyanı (merkez çizgi) 17 civarında.

•Solunum sayısı çoğunlukla 16-18 aralığında yoğunlaşmış.

•Minimum değer 12 civarında, maksimum değer 21 civarında.

•Aykırı değerler 25’in üzerine çıkmaktadır.

•KOAH (tanı = 2):

•Solunum sayısının medyanı 18 civarında ve bu değer Astım grubundan biraz daha yüksektir.

•Solunum sayısı 16-20 aralığında yoğunlaşmış.

•KOAH grubunda maksimum değer daha yüksektir (23-24 civarında).

•Aykırı değerler daha belirgindir ve 30’a kadar uzanmaktadır.

2.Aykırı Değerler:

•Her iki grupta da aykırı değerler bulunmaktadır:

•Astım grubunda 25 ve üzerinde birkaç aykırı değer mevcut.

•KOAH grubunda ise aykırı değerler 30’un üzerine çıkmaktadır.

•KOAH grubundaki aykırı değerlerin sayısı ve büyüklüğü Astım grubuna kıyasla daha fazladır.

3.Merkezi Eğilim:

•KOAH tanılı bireylerin solunum sayısı medyanı Astım grubuna göre daha yüksektir.

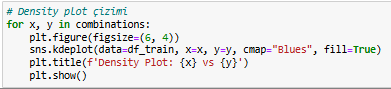
•Bu durum KOAH tanılı bireylerde solunum sayısının daha fazla arttığını göstermektedir.

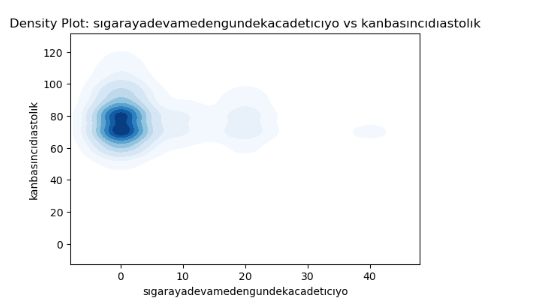
4.Değişkenlik:

•KOAH grubundaki solunum sayısı daha geniş bir yayılıma sahiptir.

•Astım grubunda solunum sayısı daha dar bir aralıkta toplanmış ve daha tutarlı görünmektedir.

Bu grafik, KOAH tanılı bireylerin solunum sayısının Astım tanılı bireylere göre daha yüksek bir medyana ve daha fazla değişkenliğe sahip olduğunu göstermektedir. KOAH grubunda solunum sayısındaki artış, hastalığın akciğer fonksiyonları üzerindeki olumsuz etkisinin bir sonucu olabilir. Aykırı değerler ise solunum yetmezliği gibi komplikasyonlara işaret edebilir.



****

Bu yoğunluk grafiği (Density Plot), sigarayadevamedengundekacadetıcıyo ile kanbasıncıdıastolık arasındaki ilişkiyi görselleştirmektedir.

•Grafikteki koyu mavi bölgeler, verilerin yoğun olarak toplandığı bölgeleri göstermektedir.

•Sıfıra yakın değerlerde (x ekseni: 0-5 arasında) sigara içmeye devam edenlerin yoğun bir şekilde toplandığı görülüyor.

•Kan basıncı (diyastolik) değerleri ise bu yoğun bölgede genellikle 80-100 aralığında yer alıyor.

•X ekseninde sigaraya devam eden günde kaç adet içiyor değişkeninin yoğunluğuna baktığımızda, en fazla sigara içme miktarı 5-10 adet/gün civarında yoğunlaşmaktadır.

•Y ekseninde kan basıncı diastolik değerleri için ise 80-100 diyastolik değerlerinin en yoğun aralık olduğu gözleniyor.

•Grafiğin sağ tarafında (x ekseni: 20-40 arası) sigara içme miktarı daha yüksek olsa da bu bölgede veri yoğunluğu düşüktür.

•Benzer şekilde, kan basıncı diastolik değerleri burada daha geniş bir aralığa yayılsa bile belirgin bir yoğunluk oluşmamıştır.

•Sigara içmeye devam eden bireylerin çoğu günde az sayıda sigara içenler grubundadır.

•Bu grupta kan basıncı diyastolik değerleri genellikle 80-100 arasında yoğunlaşmaktadır.

•Sigara içmeye devam eden kişilerin kan basıncı üzerinde belirgin bir artış trendi gözlemlenmiyor.

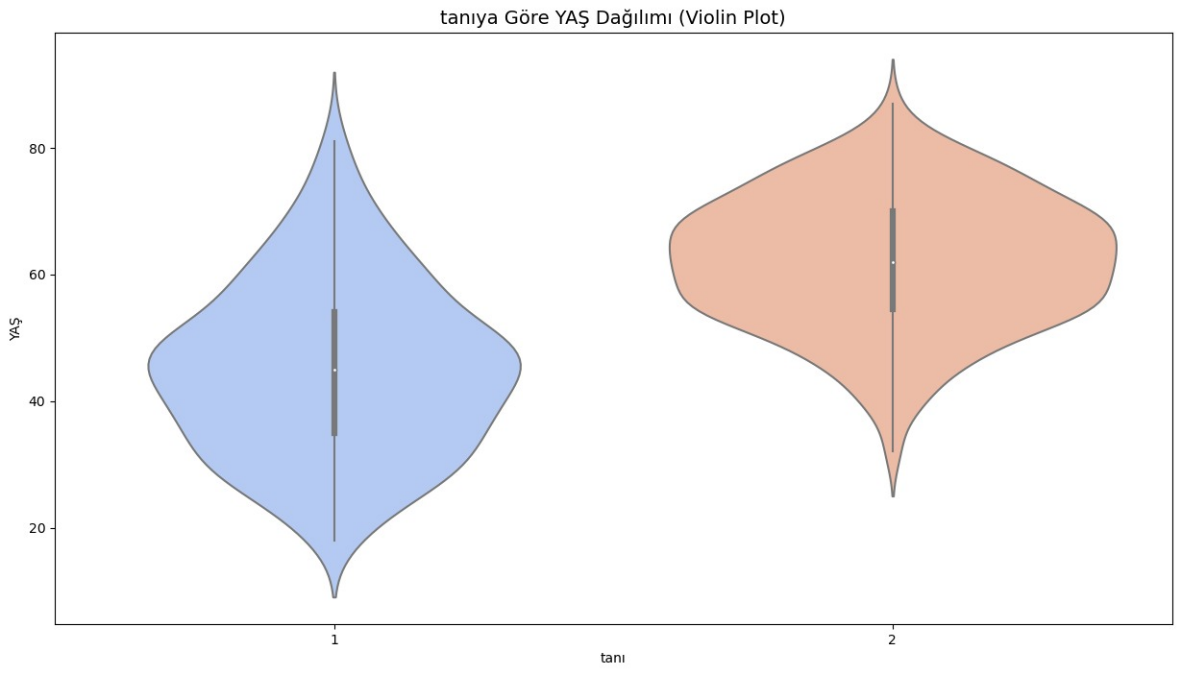
•Ancak sigara içme miktarı arttıkça (grafiğin sağa kayması), kan basıncı değerlerinde de farklı bölgelerde dağılmalar söz konusu olabilir

•Yoğunluk grafiği, sigara içme miktarının çoğunlukla 5-10 adet/gün arasında yoğunlaştığını, buna karşılık kan basıncı diyastolik değerlerinin genelde 80-100 aralığında olduğunu göstermektedir.

•Daha yüksek sigara içme miktarları için veriler seyrektir ve belirgin bir ilişki çıkarımı yapılamaz.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu



Violin Grafiği, tanı değişkenine göre yaş değerlerinin dağılımını göstermektedir. Violin plot, hem verilerin dağılımını (yoğunluğu) hem de istatistiksel özetlerini (medyan, çeyrek değerler) sunar.

1.(Tanı = 1 ,Astım):

•Yaş dağılımı daha geniş bir aralıkta dağılmıştır.

•Medyan yaş 40-45 civarındadır.

•Yaş değerleri 20-80 arasında yoğunlaşmış, ancak 20’den düşük yaşlarda ve 80 üzeri değerlerde de gözlemler bulunmaktadır.

•Daha genç bireyler bu tanı grubunda daha fazla temsil edilmektedir.

2.(Tanı = 2, KOAH):

•Yaş dağılımı daha dar ve yüksek yaş aralığında yoğunlaşmıştır.

•Medyan yaş 55-65 civarındadır.

•Yaş değerleri genel olarak 50 yaş ve üzeri bireylerde yoğunlaşmıştır.

•Bu grupta genç bireyler oldukça azdır, dağılım genellikle yaşlı bireyler üzerinde toplanmıştır.

•Astım (Tanı = 1) daha geniş bir yaş aralığında görülmekte, genç yaş gruplarında da oldukça sık karşılaşılmaktadır.

•KOAH (Tanı = 2) ise daha çok orta ve ileri yaş grubunda gözlenmektedir. Bu durum KOAH’ın yaşlanma ve uzun süreli zararlı etkenlere maruz kalma ile ilişkilendirilebileceğini göstermektedir.•Bu bulgular, tanı grupları arasındaki yaş farklılıklarını net bir şekilde ortaya koymaktadır ve ilerleyen analizlerde risk faktörleri açısından önemli bir bulgu olabilir.

**BASE MODEL**

Elimizdeki verilerin tekrar kopyasını oluşturalım.



Verimizdeki hedef değişken olan “tanı” dengeli dağılıyordu. Ama numerik verilerin çoğu çarpık dağılıyordu. Bu yüzden onlara logaritma dönüşümü uyguladım ama model performansı çok düştü. Standartlaştırmayla devam edeceğim. Log dönüşüm yaptığım adımı kalabalık olmaması adına eklemedim.

Standartlaştırma ve one-hot encoding yapalım. Temel olarak modeller ne durumda bakalım.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, çizgi, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Standartlaştırma**, değişkenlerin aynı ölçek üzerinde değerlendirilmesi gerektiği durumlarda yapılan bir işlemdir. Bu ölçek farklılıkları bazı algoritmaların (özellikle mesafeye duyarlı yöntemler) belirli değişkenlere daha fazla ağırlık vermesine neden olabilir. Örneğin:

* K-Nearest Neighbors (KNN)
* K-means
* Destek Vektör Makineleri (SVM) vb.

**Çözümü**: Standartlaştırma, tüm değişkenlerin **ortalamasını 0** ve **standart sapmasını 1** yaparak ölçek farkını ortadan kaldırır.

Standartlaştırılmış veriler, optimizasyon algoritmalarının daha kararlı çalışmasını sağlar. Örneğin:Gradient Descent algoritması, farklı ölçeklerde değişkenlerle çalışırken daha yavaş yakınsama gösterebilir. Standartlaştırma, daha düzgün bir kayıp yüzeyi sağlar. Özellikle sinir ağları (neural networks) ve lineer modellerde bu durum önemlidir.



Standartlaştırma da yaptığımızda modeller;

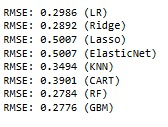


metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Base modeller düşük geldi. Z-skore yerine min-max scaler yaptığımızda ;





Ridge düşmüş, lineer regresyon çok az artmış, CART, KNN, random forest çok az artmış, lasso , elastic değişmemiş. Şimdi aykırı değerleri baskılayıp model performanslarına bakalım.

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

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu



Aykırı değer eşiklerini hesaplamak istiyoruz.

Çeyreklikler hesaplanır:

* **quartile1 (q1)**: Değerlerin %5’i (alt sınır).
* **quartile3 (q3)**: Değerlerin %95’i (üst sınır).
* Çeyrekler arası fark (IQR) hesaplanır: interquantile\_range = quartile3 - quartile1.
* Alt ve üst sınırlar belirlenir:
  + **Alt sınır**: low\_limit = quartile1 - 1.5 \* IQR.
  + **Üst sınır**: up\_limit = quartile3 + 1.5 \* IQR.

Alt ve üst sınır değerlerini döndürür.

Bir değişkende aykırı değer olup olmadığını kontrol eder. outlier\_thresholds fonksiyonundan alt ve üst sınırlar alınır.

Veride bu sınırların dışında kalan değerler aranır:

* **> up\_limit**: Üst sınırdan büyük.
* **< low\_limit**: Alt sınırdan küçük.

Eğer herhangi bir aykırı değer varsa, True döner; aksi halde False.

Aykırı değerleri alt ve üst sınırlarla değiştirmek. Alt ve üst sınırları alır(outlier\_thresholds).

**Alt sınırın altında kalan değerler**, alt sınır değeri ile değiştirilir.

**Üst sınırın üzerinde kalan değerler**, üst sınır değeri ile değiştirilir.

**num\_cols**: Sayısal değişkenlerin listesi.

Her bir sayısal değişken için:

* + **Aykırı değer kontrolü**: check\_outlier.
  + Eğer aykırı değer varsa, bunları **sınırlarla değiştirir**: replace\_with\_thresholds.

**"False"** değerleri, ilgili değişkende aykırı değer (outlier) olmadığını göstermektedir. Aykırı değer analizi sırasında her bir değişkenin alt ve üst eşik değerlerinin dışında kalan herhangi bir veri bulunmamıştır.

Baskılama var , z-skore yok , one-hot encoding var , base model:

Model genel olarak daha da düştü.

YENİ ÖZELLİK ÜRETME

One-hot encoding yaptıktan sonra yeni özellikler üretip standartlaştırma yapalım. Sonra model performanslarına kabaca bakalım.



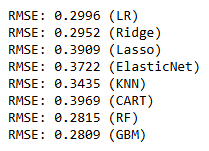
metin, ekran görüntüsü, yazılım, bilgisayar simgesi içeren bir resim

Açıklama otomatik olarak oluşturuldu

One-hot encoding yaparken tüm kategoriklere yaptığımızda model performansı: metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

bu şekilde kötü geldi. Ama hepsine değilde “varsakimde” değişkenlerine yaparsak ki zaten bazı eksiklikleri “bilgiyok” şeklinde stringle doldurmuştum. O yüzden sadece bu değişkenlere one hot yaptığımızda :



Model performansları, yeni değişken eklendikçe iyi olmuş.

Base model sayesinde fikir sahibi olduk. Min-max değil, z-skore kullanacağım. Log dönüşüm uygulamayacağım. Aykırı değerlere baskılama yapmayacağım. Aykırı değerler kalsın.one-hot encoding her kategorik değişkene değil, belirlediğim değişkenlere yapacağım. Yeni ürettiğimiz değişkenleri kullanacağım.

**MODEL SEÇİMİ**

Kodlar biraz karıştığı için veri setlerini tekrardan kopyalayıp, yeniden isimlendirelim.



Yeni değişkenler:

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldu

One-hot uygulayalım:

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, çizgi, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Standartlaştıralım:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Kullanacağım fonksiyonlar:

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

Veri setimizi modellemeye hazır hale getirdik. Eğitim kümesinin %20’sini test için yani doğrulama için ayırdım. İlk algoritmamızla modellemeye başlayalım.

**KNN:** **K-En Yakın Komşu (KNN) Algoritması:**

• KNN, verilerin komşuluk ilişkilerine dayalı çalışan, gözetimli bir sınıflandırma algoritmasıdır. Bir veri noktasının sınıfını belirlemek için en yakın “k” komşusunun sınıflarını dikkate alır.

•Çalışma Prensibi:

1.Bir veri noktasının komşularına olan mesafesi hesaplanır (genellikle Öklid mesafesi kullanılır).

2.En yakın k tane komşu belirlenir.

3.Komşuların sınıfına göre, çoğunluk oyu ile veri noktasının sınıfı tahmin edilir.

•Temel Özellikler:

•Parametre: k değeri (kaç komşuya bakılacağı).

•Eğitim aşaması yoktur; model tahmin sırasında hesaplama yapar (lazy learning).

•Mesafe ölçümleri: Öklid, Manhattan, Minkowski vb. kullanılabilir.

•Avantajları:

•Basit ve sezgisel bir algoritmadır.

•Küçük veri setlerinde iyi çalışır.

•Eğitim aşaması olmadığı için hızlıdır.

•Dezavantajları:

•Büyük veri setlerinde hesaplama maliyeti yüksektir.

•Gürültüye ve aykırı değerlere duyarlıdır.

•K değeri doğru seçilmezse model performansı düşebilir.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bu kod, \*K-Nearest Neighbors (KNN)\* algoritmasını kullanarak, farklı komşu sayılarıyla (1'den 19'a kadar) model eğitiyor ve her bir komşu sayısı için doğruluk (accuracy) değerlerini hesaplıyor. Ardından, her komşu sayısının doğruluk skorlarını bir pandas DataFrame'ine yerleştiriyor ve en yüksek doğruluk skorlarına göre sıralıyor.

1. \*give\_me\_knn(neighbor)\* fonksiyonu, verilen komşu sayısı (neighbor) ile KNN modelini oluşturur, eğitim verileriyle (x\_train, y\_train) modeli eğitir ve doğrulama verileriyle (x\_valid, y\_valid) doğruluk skorunu döndürür.

2. Döngü (for i in range(1, 20):) 1'den 19'a kadar komşu sayıları için bu fonksiyonu çağırır ve doğruluk skorlarını bir listeye ekler.

3. Bu listeyi \*accuracy\_df\* adında bir pandas DataFrame'ine dönüştürür ve bu DataFrame'i doğruluk skorlarına göre azalan sırayla sıralar. Sonuç olarak, farklı komşu sayılarının doğruluk değerleri ile sıralanmış bir tablo elde edilir.

Accuracy (Doğruluk): Modelin genel olarak doğru tahmin yapma oranını ölçer.

F1 skore: F1-Score, modelin hem doğruluğunu hem de kapsayıcılığını birleştirerek tek bir değerle ifade eder.

metin, ekran görüntüsü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

**En iyi k değeri 11’dir ve skoru 0.86’dır. Test hatası ise 0.305’miş.**

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

Açıklama otomatik olarak oluşturuldu**

**ekran görüntüsü, kare, renklilik, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu**

48: Astım olan hastaların 48'i doğru bir şekilde astım olarak sınıflandırılmış (True Positive- TP).

5: Astım olan hastaların 5'i yanlış bir şekilde KOAH olarak sınıflandırılmış (False Negative- FN).

9: KOAH olan hastaların 9'u yanlış bir şekilde astım olarak sınıflandırılmış (False Positive- FP).

38:KOAH olan hastaların 38'i doğru bir şekilde KOAH olarak sınıflandırılmış (TruePositive - TP).

**metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Bu grafik, ROC (Receiver Operating Characteristic) eğrisini göstermektedir ve modelin sınıflandırma performansını değerlendirmek için kullanılır.

ROC Eğrisi:Modelin farklı eşik değerleri için doğru pozitif oranı (TPR) ve yanlış pozitif oranı (FPR) arasındaki ilişkiyi gösterir.

AUC (Area Under Curve):ROC eğrisinin altında kalan alanı temsil eder. Modelin genel performansını değerlendirir.

AUC değeri 0.5: Rastgele tahmin.

AUC değeri 1.0: Mükemmel tahmin.

Grafik Yorumu:

AUC Değeri: 0.9156 olarak gösterilmiştir. Bu değer, modelin sınıflandırma performansının oldukça iyi olduğunu belirtir. Model, pozitif ve negatif sınıfları birbirinden ayırmada yüksek bir doğruluk göstermektedir.

ROC Eğrisinin Şekli: Eğri, sol üst köşeye doğru yükselmiş ve ardından plato yapmıştır. Bu, modelin yanlış pozitif oranını düşük tutarken doğru pozitif oranını hızla artırabildiğini gösterir.

Model, doğrulukla tahmin yapabilen güçlü bir sınıflandırıcıdır. AUC değeri 0.9156 olduğundan, modelin performansı çok iyi seviyededir ve sınıflar arasında etkili bir ayrım yapabilmektedir.

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

Açıklama otomatik olarak oluşturuldu**

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Class 0(tanı:1)'da yüksek recall (0.91), ancak precision biraz daha düşük (0.84). Bu, modelin sınıf 0'ı doğru tespit etme oranının yüksek olduğunu ancak bazı yanlış pozitif tahminler yaptığını gösteriyor.

Class 1(tanı:2)'de ise yüksek precision (0.88) ve iyi bir recall (0.81) değeri, modelin bu sınıfı doğru tanıma oranının ve doğru pozitif tahminlerinin yüksek olduğunu gösteriyor.

F1-Score, her iki sınıf için de oldukça dengeli ve iyi seviyelerde. Yüksek F1 skoru, modelin hem doğru tahmin yapma hem de gerçek pozitifleri yakalama yeteneğinin iyi olduğunu gösteriyor.

Modelin genel doğruluğu 0.86, yani %86 doğrulukla makul bir başarı sağlıyor.

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

Açıklama otomatik olarak oluşturuldu

İleride daha rahat görmek için model oluşturdukça bu dataframe’in içine atacağız.

NAİVE BAYES

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

ekran görüntüsü, renklilik, kare, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

45: Astım olan hastaların 45'i doğru bir şekilde astım olarak sınıflandırılmış (True Positive- TP).

8: Astım olan hastaların 8'i yanlış bir şekilde KOAH olarak sınıflandırılmış (False Negative- FN).

9: KOAH olan hastaların 9'u yanlış bir şekilde astım olarak sınıflandırılmış (False Positive- FP).

38:KOAH olan hastaların 38'i doğru bir şekilde KOAH olarak sınıflandırılmış (TruePositive - TP).

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Test hatası 0.41’miş. Sınıf 0 (Precision, Recall, F1-Score):

* Precision: 0.83 — Sınıf 0 olarak tahmin edilenlerin %83'ü gerçekten sınıf 0(astım)
* Recall: 0.85 — Gerçek sınıf 0(astım)'ların %85'i doğru şekilde tahmin edilmiş.
* F1-Score: 0.84 — Precision ve recall'un dengeli olduğu iyi bir F1 skoru.

Sınıf 1 (Precision, Recall, F1-Score):

* Precision: 0.83 — Sınıf 1(koah) olarak tahmin edilenlerin %83'ü gerçekten sınıf 1.
* Recall: 0.81 — Gerçek sınıf 1'lerin %81'i doğru şekilde tahmin edilmiş.
* F1-Score: 0.82 — Yine iyi bir F1 skoru, precision ve recall arasında makul bir denge.

Genel Sonuçlar:

* Accuracy (Doğruluk): 0.83 — Model, genel olarak %83 doğrulukla doğru tahmin yapmış.
* Macro Avg (Ortalama): 0.83 — Sınıfların ortalama precision, recall ve F1 skoru.
* Weighted Avg (Ağırlıklı Ortalama): 0.83 — Sınıfların destek oranlarına göre ağırlıklı ortalama, yine %83.

metin, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Modelin AUC değeri 0.8986 olduğu için, modelin genel performansı oldukça iyidir.

ROC eğrisi, yanlış pozitif oranı düşükken doğru pozitif oranının hızla arttığını gösteriyor. Bu da modelin iyi bir ayrıştırma gücüne sahip olduğunu ortaya koyar.

Özellikle, hastalık ve sağlıklı bireyler arasındaki ayrımı doğru yapmada model başarılı görünmektedir.

AUC değeri ile yaklaşık %90 doğrulukla sınıfları ayırt edebilmektedir.

**LOGISTIC REGRESSION**

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

Açıklama otomatik olarak oluşturuldu

Skor 0.87.

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

Açıklama otomatik olarak oluşturuldu

ekran görüntüsü, renklilik, kare, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

45: Astım olan hastaların 45'i doğru bir şekilde astım olarak sınıflandırılmış (True Positive- TP).

8: Astım olan hastaların 8'i yanlış bir şekilde KOAH olarak sınıflandırılmış (False Negative- FN).

5: KOAH olan hastaların 5'u yanlış bir şekilde astım olarak sınıflandırılmış (False Positive- FP).

42:KOAH olan hastaların 42'i doğru bir şekilde KOAH olarak sınıflandırılmış (TruePositive - TP).

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Test hatası 0.360.

Model, sınıf 0(astım)'da yüksek precision (0.90) ve sınıf 1(koah)'de yüksek recall (0.89) sağlıyor.

Her iki sınıf için F1 skoru 0.87 civarında, bu da modelin precision ve recall arasında iyi bir denge kurduğunu gösteriyor.

Accuracy %87 ile iyi bir genel doğruluk sağlanmış.

metin, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

AUC değeri 0.94, modelin rastgele bir pozitif sınıf örneğini, negatif sınıf örneğinden %94 olasılıkla doğru bir şekilde ayırt edebileceğini belirtir.0.5 olan bir AUC değeri, rastgele tahmin anlamına gelirken; 1.0 ise mükemmel bir model anlamına gelir.

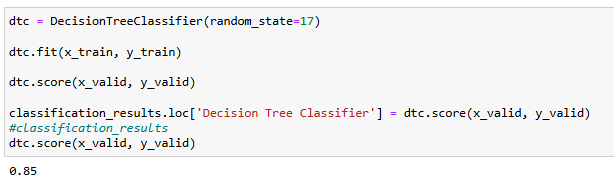
ROC Eğrisinin Şekli: Eğri, grafiğin sol üst köşesine doğru yaklaşarak yüksek bir TPR (doğru pozitif oranı) ve düşük bir FPR (yanlış pozitif oranı) elde edildiğini gösteriyor.Bu, modelin tahminlerinin oldukça güvenilir olduğunu ve yanlış pozitiflerin minimal düzeyde olduğunu vurgular.Burada Lojistik Regresyon (logreg) kullanılmış ve predict\_proba fonksiyonu ile tahmin olasılıkları alınmıştır.

Y ekseni: True Positive Rate (Hassasiyet)

X ekseni: False Positive Rate

Yorum:Bu sonuçlar, lojistik regresyon modelinizin doğruluk açısından oldukça iyi çalıştığını ve verilerdeki sınıfları doğru bir şekilde ayırabildiğini gösteriyor. AUC değeri 0.94 olduğu için modelin performansı başarılı olarak değerlendirilebilir.

KARAR AĞAÇLARI(Decision Tree Classifier)



Karar ağacının performansı 0.85’miş.

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

Açıklama otomatik olarak oluşturuldu

ekran görüntüsü, renklilik, kare, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

46: Astım olan hastaların 46'i doğru bir şekilde astım olarak sınıflandırılmış (True Positive- TP).

7: Astım olan hastaların 7'i yanlış bir şekilde KOAH olarak sınıflandırılmış (False Negative- FN).

8: KOAH olan hastaların 8'u yanlış bir şekilde astım olarak sınıflandırılmış (False Positive- FP).

39:KOAH olan hastaların 39'i doğru bir şekilde KOAH olarak sınıflandırılmış (TruePositive - TP).

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

Açıklama otomatik olarak oluşturuldu

Hata: 0.387

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

F1 skoru her iki sınıf için de dengeli ve 0.84-0.86 arasında. Bu, modelin precision ve recall arasında iyi bir denge sağladığını gösteriyor.

Accuracy %85 ile makul bir doğruluk sağlanmış ve model her iki sınıfı da dengeli bir şekilde sınıflandırmış gibi görünüyor.

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Grafikte AUC değeri 0.848olarak verilmiştir.

•AUC (Area Under the Curve): ROC eğrisinin altındaki alanı temsil eder ve modelin sınıflandırma performansını ölçer.

•AUC değeri 0.848, modelin sınıflandırma performansının “çok iyi” seviyede olduğunu göstermektedir.

•Model, pozitif ve negatif sınıfları ayırt etmede başarılıdır.

•ROC eğrisinin yukarı doğru hızlı bir şekilde yükselmesi, modelin yanlış pozitif oranını düşük tutarken doğru pozitif oranını artırabildiğini ifade eder.

RASTGELE AĞAÇLAR(Random Forest Classifier)

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

Açıklama otomatik olarak oluşturuldu

Modelin performansı %88.

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

Açıklama otomatik olarak oluşturuldu

ekran görüntüsü, renklilik, kare, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

48: Astım olan hastaların 48'i doğru bir şekilde astım olarak sınıflandırılmış (True Positive- TP).

5: Astım olan hastaların 5'i yanlış bir şekilde KOAH olarak sınıflandırılmış (False Negative- FN).

7: KOAH olan hastaların 7'si yanlış bir şekilde astım olarak sınıflandırılmış (False Positive- FP).

40:KOAH olan hastaların 40'i doğru bir şekilde KOAH olarak sınıflandırılmış (TruePositive - TP).

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

Açıklama otomatik olarak oluşturuldu

Hata:0.346

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

* **Precision**: 0.87 — Sınıf 0 (Astım) olarak tahmin edilenlerin %87'si gerçekten Astım. Yani, model astım hastalarını doğru bir şekilde ayırmada oldukça başarılı.
* **Recall**: 0.91 — Gerçek astım hastalarının %91'i doğru şekilde tahmin edilmiş. Bu, modelin astım hastalarını doğru tespit etme oranının yüksek olduğunu gösteriyor.
* **F1-Score**: 0.89 — Precision ve recall arasında iyi bir denge sağlanmış. Astım sınıfı için dengeli ve başarılı bir performans var.

**2. Sınıf 1 (KOAH) (Precision, Recall, F1-Score):**

* **Precision**: 0.89 — Sınıf 1 (KOAH) olarak tahmin edilenlerin %89'u gerçekten KOAH. Yani, model KOAH hastalarını doğru bir şekilde ayırmada da başarılı.
* **Recall**: 0.85 — Gerçek KOAH hastalarının %85'i doğru şekilde tahmin edilmiş. Bu, modelin KOAH hastalarını tespit etme oranının astım sınıfına göre biraz daha düşük olduğunu gösteriyor.
* **F1-Score**: 0.87 — Yine iyi bir denge sağlanmış, ancak astım sınıfına göre biraz daha düşük.

**3. Genel Sonuçlar:**

* **Accuracy (Doğruluk)**: 0.88 — Model, genel olarak %88 doğrulukla doğru tahminler yapmış. Hem astım hem de KOAH sınıflarında başarılı sonuçlar elde edilmiş.
* **Macro Avg (Ortalama)**: 0.88 — Sınıfların ortalama precision, recall ve F1 skoru da oldukça dengeli.
* **Weighted Avg (Ağırlıklı Ortalama)**: 0.88 — Sınıfların destek oranlarına göre ağırlıklı ortalama da %88, bu da modelin her iki sınıfı dengeli bir şekilde sınıflandırdığını gösteriyor.

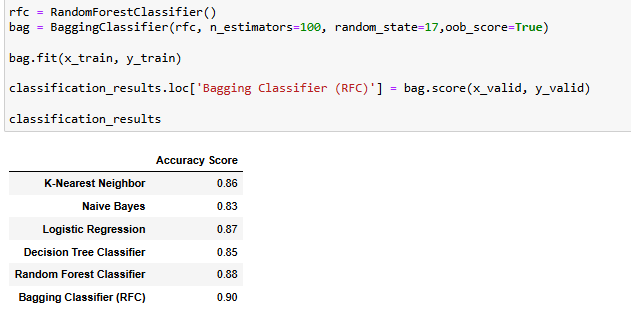
metin, diyagram, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Grafikte verilen AUC değeri: 0.9516.

• AUC, ROC eğrisinin altında kalan alanı ifade eder.AUC değeri 0.95 (95%) oldukça yüksek bir değerdir ve modelin sınıflandırma performansının çok iyi olduğunu gösterir. Eğri, sol üst köşeye yakın bir şekilde ilerlediği için yanlış pozitif oranı düşük, doğru pozitif oranı yüksektir.

BAGGING CLASSIFIER (RFC)



Model performansı %90.

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

Açıklama otomatik olarak oluşturuldu

ekran görüntüsü, renklilik, kare, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

48: Astım olan hastaların 48'i doğru bir şekilde astım olarak sınıflandırılmış (True Positive- TP).

5: Astım olan hastaların 5'i yanlış bir şekilde KOAH olarak sınıflandırılmış (False Negative- FN).

5: KOAH olan hastaların 5'i yanlış bir şekilde astım olarak sınıflandırılmış (False Positive- FP).

42:KOAH olan hastaların 42'si doğru bir şekilde KOAH olarak sınıflandırılmış (TruePositive - TP).

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Hata:0.316

Modelin her iki sınıfı da çok başarılı bir şekilde sınıflandırdığı ve dengeleli bir performans sergilediği söylenebilir. Astım (sınıf 0) ve KOAH (sınıf 1) için hem precision hem de recall oldukça iyi, dolayısıyla model yanlış pozitif (false positive) ve yanlış negatif (false negative) oranlarını düşük tutmuş. Bu, oldukça güçlü bir sınıflandırma modeline işaret eder.

metin, diyagram, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

AUC değeri 0.956, modelin mükemmele yakın bir performans gösterdiğini belirtir.

Bu, modelin rastgele bir pozitif örneği negatif örnekten %95.6 olasılıkla doğru bir şekilde ayırt edebileceğini ifade eder.

Eğri, grafiğin sol üst köşesine daha yakın ilerlemekte. Bu, modelin yüksek doğru pozitif oranını (TPR) korurken, düşük yanlış pozitif oranına (FPR) sahip olduğunu gösterir.

Eğrinin bu şekilde sol üst köşeye yakın olması, modelin başarılı bir sınıflandırıcı olduğuna işaret eder.Bagging yöntemlerinin, lojistik regresyon gibi lineer modellere kıyasla daha karmaşık yapıları öğrenme ve daha iyi performans gösterme kapasitesi vardır.

Genel Yorum:Bagging modeli, veriler üzerinde daha başarılı bir performans göstermiştir. Bu, özellikle modelin daha karmaşık örüntüleri öğrenebilme yeteneğine ve düşük yanlış tahmin oranına sahip olduğunu gösterir. Bu nedenle, AUC değerine göre bu model tercih edilebilir.

GRADIENT BOOSTING CLASSIFIER

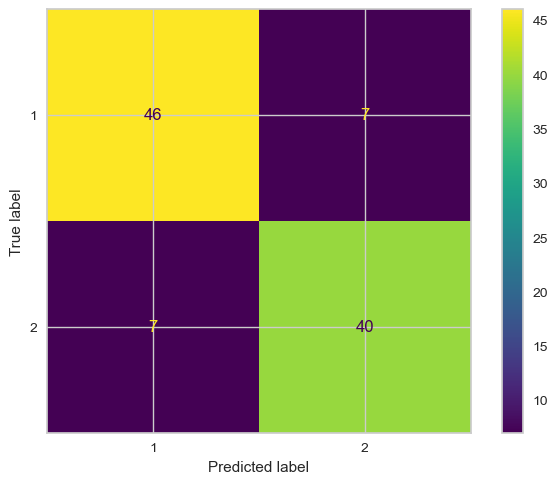
metin, ekran görüntüsü, yazı tipi, yazılım içeren bir resim

Açıklama otomatik olarak oluşturuldu

Modelin performansı %86.

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

Açıklama otomatik olarak oluşturuldu



46: Astım olan hastaların 46'si doğru bir şekilde astım olarak sınıflandırılmış (True Positive- TP).

7: Astım olan hastaların 7'si yanlış bir şekilde KOAH olarak sınıflandırılmış (False Negative- FN).

7: KOAH olan hastaların 7'si yanlış bir şekilde astım olarak sınıflandırılmış (False Positive- FP).

40:KOAH olan hastaların 40'ı doğru bir şekilde KOAH olarak sınıflandırılmış (TruePositive - TP).



Hata: 0.374

Model, astım (sınıf 0) ve KOAH (sınıf 1) sınıflarını başarıyla sınıflandırmış ve dengeyi sağlamış.

Precision ve recall değerleri her iki sınıf için de oldukça yüksek. Yani, model yanlış pozitif (false positive) ve yanlış negatif (false negative) oranlarını düşük tutmuş.

F1 skoru, her iki sınıf için de 0.87 ve 0.85 ile dengeyi gösteriyor, yani model hem doğru tespit hem de yanlış pozitifleri azaltma açısından başarılı.

Accuracy %86, bu da modelin genel olarak iyi bir performans sergilediğini gösteriyor, ancak biraz daha yüksek doğruluk elde edilebilir.

metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu  
Eğri, sol üst köşeye yakın ilerliyor. Bu, modelin False Positive Rate (Yanlış Pozitif Oranı)’nı düşük tutarken True Positive Rate (Doğru Pozitif Oranı)’nı yüksek tuttuğunu ifade eder.

Eğrinin düz bir çizgi yerine dik ve yukarıya doğru ilerlemesi, modelin doğru tahmin oranının yüksek olduğunu gösterir.

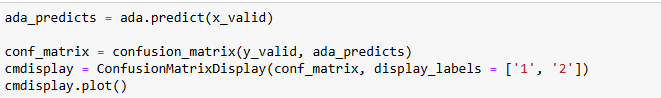
AUC = 0.9490: Model, sınıfları birbirinden çok iyi ayırıyor.Eğri, sol üst köşeye yakın bir şekilde ilerliyor.Bu sonuç, modelin yüksek doğruluk ve duyarlılığa sahip olduğunu ifade eder.

ADABOOST CLASSIFIER

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

Açıklama otomatik olarak oluşturuldu

Model performansı %87’miş.



ekran görüntüsü, kare, renklilik, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

47: Astım olan hastaların 47'si doğru bir şekilde astım olarak sınıflandırılmış (True Positive- TP).

6: Astım olan hastaların 6'sı yanlış bir şekilde KOAH olarak sınıflandırılmış (False Negative- FN).

7: KOAH olan hastaların 7'si yanlış bir şekilde astım olarak sınıflandırılmış (False Positive- FP).

40:KOAH olan hastaların 40'i doğru bir şekilde KOAH olarak sınıflandırılmış (TruePositive - TP).

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

Açıklama otomatik olarak oluşturuldu

Hata: 0.3605

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

Açıklama otomatik olarak oluşturuldu

* **Accuracy (Doğruluk)**: 0.87 — Model genel olarak %87 doğrulukla doğru tahminler yapmış. Yani model, **%87 doğru sınıflandırma** yaparak başarılı bir performans sergilemiş.

**Sınıf 0 (Precision, Recall, F1-Score)**:

* **Precision**: 0.87 — Modelin **sınıf 0 (Astım)** olarak tahmin ettiği örneklerin %87'si gerçekten astım (doğru pozitif). Yani model, astım olarak tahmin ettiği hastaların çoğunu doğru şekilde sınıflandırmış.
* **Recall**: 0.89 — Gerçek **astım hastalarının %89'u** doğru şekilde tahmin edilmiş. Bu, modelin astım hastalarını tespit etme oranının **yüksek** olduğunu gösteriyor.
* **F1-Score**: 0.88 — Precision ve recall arasında bir dengeyi gösteren bu skoru yüksek sayılır, yani model astım sınıfını **dengeli bir şekilde sınıflandırmış**.

**Sınıf 1 (KOAH) (Precision, Recall, F1-Score):**

* **Precision**: 0.87 — Model, **KOAH** olarak tahmin ettiği örneklerin %87'sinin gerçekten KOAH olduğunu doğru bir şekilde tahmin etmiş. Yani modelin **true positives** oranı oldukça yüksek.
* **Recall**: 0.85 — Gerçek **KOAH hastalarının %85'i** doğru şekilde tahmin edilmiş. Bu da modelin KOAH hastalarını **iyi şekilde tespit ettiğini** gösteriyor.
* **F1-Score**: 0.86 — Precision ve recall arasında bir denge bulunuyor. Modelin KOAH sınıfı için de **başarılı** bir performansı olduğunu gösteriyor

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

AUC (Eğri Altındaki Alan) değeri 0.94 civarında. Bu değer, modelin sınıflandırma performansının oldukça iyi olduğunu gösteriyor.

AUC değeri 1’e yaklaştıkça modelin doğruluk oranı artar.

0.94 gibi bir değer, modelin hem pozitifleri hem de negatifleri doğru bir şekilde ayırt edebilme yeteneğinin yüksek olduğunu belirtir.

Eğri, sol üst köşeye doğru kıvrılmıştır. Bu durum, modelin yanlış pozitif oranı düşükken doğru pozitif oranı yüksek tuttuğunu gösterir.

Eğrinin ideal bir sınıflandırıcıya yakın bir eğri olduğu söylenebilir.

Bu grafik, modelin yüksek performans gösterdiğini (AUC = 0.94) açıkça belirtmektedir. Model, tahminlerinde doğru sınıflandırma yapma konusunda başarılıdır.

**CATBOOST**

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

Modelin performansı %97 iken hatası 0.272’dir.

metin, ekran görüntüsü, yazılım, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

0.36055512754639896-ADA

0.2725341568439064-catboost

0.37416573867739417-gradient

0.31622776601683794-bagging

0.34641016151377546-rfc

0.3872983346207417-dtc

0.36055512754639896-logreg

0.41231056256176607-bayes

0.30559520570857474-knn

Hem en küçük hataya sahip olduğundan hem de en iyi skora sahip olduğundan catboost seçtim.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu