

**İST405 - İSTATİSTİKSEL VERİ MADENCİLİĞİ ARA SINAV ÖDEVİ**

DR. ÖĞR. ÜYESİ ONUR TOKA

ŞİMAL VİZE - 2210329027

İREM TARIM - 2210329047

ANONİMLEŞTİRİLMİŞ HASTA VERİLERİYLE ASTIM VE KOAH TANISI SINIFLANDIRMA VE TEST KÜMESİ TAHMİNİ

ÇANKAYA, ANKARA

2024

İçindekiler Tablosu

**Hiçbir içindekiler tablosu öğesine rastlanmadı.**

# GİRİŞ

Astım, solunum yollarını tıkayan ve nefes almayı zorlaştıran kronik bir hastalıktır. Astım atakları genellikle alerjenlere, duman, soğuk hava, egzersiz gibi tetikleyicilere yanıt olarak ortaya çıkar. Astımın nedeni tam olarak bilinmemekle birlikte, genetik faktörler, çevresel etkenler, enfeksiyonlar ve kişisel tıbbi durumlar etkili olabilir. Astım belirtileri arasında nefes darlığı, hırıltı veya ıslık sesi, öksürük ve göğüste ağrı veya sıkışma gibi durumlar yer alır. Astım riskini artıran faktörler arasında sigara içme, alerji, ailede astım öyküsü, hava kirliliği gibi etkenler bulunmaktadır. Astımın tedavisi genellikle ilaçlar ve yaşam tarzı değişiklikleri ile yönetilir. Kronik Obstrüktif Akciğer Hastalığı (KOAH), solunum yollarında kalıcı tıkanıklık ve nefes darlığına yol açan ilerleyici bir hastalıktır. KOAH’ın belirtileri arasında sürekli ve şiddetli öksürük, balgam ve nefes darlığı bulunur. Bu hastalık genellikle uzun süreli sigara kullanımı ve diğer solunum yolu tahriş edicilerine maruz kalma sonucunda gelişir. Bu hastalıkların erken teşhisi ve doğru sınıflandırılması, tedavi sürecinin etkinliği açısından kritik bir öneme sahiptir. Gelişen tıbbi veri analitiği ve makine öğrenimi yöntemleri, hastalık sınıflandırmasında ve tahmin modellerinin geliştirilmesinde yeni fırsatlar sunmaktadır.

Bu çalışmada, anonimleştirilmiş hasta verileri üzerinden astım ve KOAH tanılarının sınıflandırılması ve tahmini hedeflenmiştir. Çalışma, hastaların demografik, klinik, yaşam tarzı ve genetik bilgilerini içeren kapsamlı bir veri seti kullanılarak gerçekleştirilmiştir. Veriler üzerinde gerçekleştirilen analizlerle, sınıflandırma sürecine katkı sağlayacak kritik değişkenlerin belirlenmesi ve bu bilgiler ışığında etkili bir tahmin modeli geliştirilmesi amaçlanmıştır. Model belirlenmesi için Python programı kullanılacaktır.

# Veri Ön İşleme

Bu kısımda, KOAH ve astım tanılı hastaların sağlık veri seti üzerinde gerçekleştirilen veri temizleme, dönüştürme ve özellik mühendisliği süreçleri detaylandırılmaktadır. Amaç, veri setini analiz ve modelleme için uygun hale getirmektir.

İlk adım olarak veri programa okutulduktan sonra veriyi doğru incelemek adına her sütundaki verilerin veri tipi incelemesi yapılmıştır.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Kategorileştirme için Kodlar

Veri tipi incelemesi aşamasında kategorik verilerin okutulması gerçekleştirildi. Cinsiyet, eğitim düzeyi, meslek, sigara kullanımı gibi değişkenler kullanıldı. Bu dönüşüm bellek verimliliğini artırması ve kategorik verilerle işlem yapmayı kolaylaştırması için tercih edildi.

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

Açıklama otomatik olarak oluşturuldu

Şekil . Eksik Değerleri Doldurma

Ardından eksik veri analizi sonucunda, bazı eksiklikler bulunduğundan bu eksiklikler her bir değişken için medyan hesaplaması yapılıp bulunan değer ile dolduruldu.

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

Açıklama otomatik olarak oluşturuldu

Şekil . Yaş Grubu Oluşturma

Yaş değişkeninin gruplandırılması, bireylerin belirli yaşam evrelerine göre sınıflandırılmasını sağlar. Bu işlem; farklı yaş gruplarının hastalık eğilimleri, risk faktörleri veya davranış kalıplarını daha net bir şekilde anlamamıza olanak tanır. Sürekli bir değişkenin kategorik hale getirilmesi, modelin genelleme kapasitesini artırabilir ve aşırı öğrenme (overfitting) riskini azaltabilir. Kategorik yaş grupları, analiz sonuçlarını daha anlaşılır hale getirir. Bu avantajları sağladığından bu şekilde bir gruplandırma yapılmıştır.

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

Açıklama otomatik olarak oluşturuldu

Şekil . Nabız Değişkeni Gruplandırma

Nabız grupları, bireyin sağlık durumunu değerlendirmek için hızlı ve etkili bir yol sağlar. Belirli risk gruplarını tanımlamak için kullanışlıdır. Nabız grubu oluşturma makine öğrenimi modellerinde genelleme kapasitesini arttırır. Kategorik gruplar modelin daha basit ve anlamlı hale gelmesine yardımcı olur. Bu gruplandırma 0-60 düşük, 60-100 normal, 100+ yüksek olarak gruplandırılmıştır.

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

Açıklama otomatik olarak oluşturuldu

Şekil . Gün Değişkenlerinin Verilerini Saate Çevirme

Gün değişkenlerini saate çevirmek veri analizinde birimler arasındaki tutarlılığı sağlamak ve ayrıntılı analizler yapmak için gerekli görülmüştür. Analizlerde karmaşıklığı azaltması ve veri setinin daha düzenli hale gelmesini sağlar. Saat birimini kullanmak daha hassas bir analiz sağlayacağı düşünülmüştür.

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

Açıklama otomatik olarak oluşturuldu

Şekil . Tanı Yılını Aylara Çevirme

Aynı şekilde saate çevirme de olduğu gibi veride aynı değişkenin hem yıl hem de ay olarak sayısal değerinin olması model kurulumunda yanlışlıklar doğurabileceğinden tek bir zaman dilimine çevrilmiştir.

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Kitle Endeksi Değişkeni Oluşturulması

Kitle endeksi, boy ve kilo verilerine dayalı olarak hızlı ve kolay bir şekilde bir bireyin sağlık durumunu değerlendirmek için kullanılabilir. Kitle endeksi, obezite veya aşırı zayıflık gibi durumları tespit etmede etkili bir araçtır. Buna bağlı olarak hastalıklarının tanısına bir etkisi olup olmadığı birleştirilerek daha iyi sonuç elde edileceği düşünüldüğünden tek değişkene indirilmiştir.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Kan Basıncı Değişkeni Oluşturma

PCA uygulaması, iki değişkenin birleşiminden tek bir bileşen oluşturulmasını sağlar. Veri setinin boyutunu küçülterek daha verimli ve hızlı analizler yapılmasını sağlar. Ayrıca ilişki bakımından yüksek anlamlı çıkabilecek iki değişken olduğundan PCA ile verinin boyutu küçültülerek, bazı makine öğrenimi algoritmalarının performansı artırılması amaçlanmıştır. Çünkü daha az bileşen ile modelin aşırı uyum (overfitting) yapma riski azalır.

Özetle, veri düzenlemesinde yaş değişkeni çıkarılarak yaş grubuna çevrildi. Nabız grupları oluşturuldu. Gün değişkenleri saate, yıl değişkeni ise aya dönüştürülerek model için daha temiz bir veri seti kullanılması sağlandı. Boy ve vücut ağırlığı kullanılarak kitle endeksi hesaplandı ve böylece iki değişken tek değişkene indirildi. Başvuru tarihi modele bir katkısı olmayacağı düşünüldüğünden veriden silindi. Sistolik ve diastolik değerleri kan basıncı olarak PCA uygulanarak tek değişken haline getirildi. Ana veride FEV1 ve PEF değerlerinin yüzdeliklerinin bulunmasından dolayı sağlık verileri analizinde genellikle yüzdelikler tercih edildiğinden bu ikisi silindi.

## Tanı ile Bağımsız Numerik Değişkenlerinin İlişki İncelemesi

Mann-Whitney U testi, iki bağımsız örneklem arasında ortalama sıralama farklarını incelemek için kullanılan bir non-parametrik testtir. Bu test, özellikle verilerin normal dağılıma uymadığı durumlarda kullanılır. Test tanı ile bağımsız değişkenlerin aralarında ilişkilere bakılarak modele hangi değişkeni koymak daha anlamlıdır için cevap aranması sonucu yapılmıştır.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Tanı Değişkenin Etiketlendirme ve Örnek Mann-Whitney Testi-Sonucu

Astım hastalarının solunum sayısı ile KOAH hastalarının solunum sayısı arasında istatistiksel olarak anlamlı bir fark yoktur.

Astım hastalarının solunum sayısı ile KOAH hastalarının solunum sayısı arasında istatistiksel olarak anlamlı bir fark vardır. (1)

Hipotez (1)’de olduğu gibi kuruduğunda test sonucunda elde edilen p değeri 6.63e-13 olarak elde edilmiştir ve bu değer 0’a çok yakın bir değer olduğundan yokluk hipotezi reddedilir. Sonuç olarak solunum sayısının her iki tanı için farklı işlemektedir. Bu modele koyulması için anlamlı bir değişken olduğunun belirtisidir. Test sonucu görselleştirmesi için box-plot grafiği tercih edilmiştir.

metin, diyagram, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Box-Plot Grafiği

Astım için Medyan solunum sayısı yaklaşık **18** olarak görünüyor. Bu, verilerin yarısının 18'in altında, diğer yarısının ise 18'in üstünde olduğu anlamına gelir. Dağılımı KOAH’a göre normal gibi durmaktadır. KOAH tanısında, özellikle üst sınırda daha geniş bir yayılım bulunmaktadır. Bu durum, KOAH hastalarının solunum sayılarının daha değişken olduğunu gösteriyor. Daha heterojen bir grup olduğu söylenebilir. Bu grafikle aykırı değere sahip olduğu görülmektedir.

Tanı ile diğer numerik bağımsız değişkenlerin Mann-Whitney U testi yapıldığında bırakalı kaç gün olduğu, tanı süresi(ay), acil servise yatış (saat), yoğun bakıma yatış sayısı ve (saat) gibi değişkenlerde tanı açısından bir fark bulunamadığından bu değişkenler modele konmamak üzere veriden çıkarılmıştır.

## Tanı ile Kategorik Değişkenlerin İlişki İncelemesi

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil . Tanı ile Ki-Kare Testi

Bu Ki Kare testi ile tanı ve bağımsız kategorik değişkenlerin arasındaki ilişkileri araştırmak amacıyla yapılmıştır. Sonuç özeti olarak, tanı ile cinsiyet, meslek grubu, sigara kullanımı durumu, ailede tanılı biri var mı, anne tanılı mı, kardeş tanılı mı, nabız grubu, yaş grubu, hastaneye yattı mı değişkenleri kullanıldığında tanı arasında fark vardır.

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 12. Ki Kare Testi

İki kategorik değişken arasında anlamlı bir ilişki varsa, bu durum analizde dikkate alınabilir ve modelleme sürecine katkı sağlar. Ki-kare testi, iki kategorik değişkenin birbirinden bağımsız olup olmadığını test eder. Buna göre kategorik değişkenlerin kendi aralarındaki ilişkileri incelenecektir. Önem öncelikle tanı ile anlamlı çıkanlara verilecektir. Sonuçlara genel olarak bakıldığında ailede tanılı biri olup olmamasıyla anne, baba, kardeş ve diğer değişkenleri ile anlamlı çıkmıştır. Fakat baba ve diğer değişkeni tanıyla anlamlı çıkmadığından veriden silineceklerdir, aynı zamanda anne ve kardeş de tanı olup olmadığı değişkeni ile yüksek seviyede ilişkili olduğundan modelde çoklu bağlantı sıkıntısı olmasını istenmediğinden onlar da modelden çıkarılacaktır.

Özetle bu testler sonucunda kategorik değişkenlerde vazgeçilenler; anne, baba, kardeş, diğerde tanı var mı, eğitim düzeyi olmuştur.

## Standartlaştırma

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 13. Robust ile Standartlaştırma

**RobustScaler,** medyan ve çeyrekler arası mesafeye dayalı bir ölçeklendirme işlemi uygular. Bu, verilerdeki uç değerlerden (outliers) etkilenmeden daha sağlam bir standartlaştırma sağlar. Robust kullanılmasının sebebi verinin normal dağılıma uymamasındandır. Bu standartlaştırma yöntemi ile hem standartlaştırma hem de aykırı değerlerin öneminin kalmaması sağlanmıştır. Bu yöntem diğer yöntemlere göre (örneğin MinMaxScaler veya StandardScaler) farkı, **aşırı uç değerler** gibi anormal verilerin etkisini minimize etmesidir. Birkaç çok büyük değer bulunuyorsa **MinMaxScaler** veya **StandardScaler** bu uç değerleri dikkate alır ve bunlar nedeniyle, tüm veri setinin ölçeklenmesi bozulabilir (örneğin, tüm veriler çok küçük bir aralığa sıkışabilir). Ancak **RobustScaler** kullanıldığında, bu uç değerler medyan ve IQR hesaplamalarında etkili olamayacağı için, normal dağılan değerler üzerinde daha uygun bir ölçekleme yapılır. Bu, modelin performansını iyileştirebilir.

## One-Hot Encoding İşlemi

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 14. One-Hot Encoding İşlemi

Makine öğrenmesi algoritmaları, genellikle sadece sayısal verilerle çalışabilir, kategorik veriler doğrudan model tarafından işlenemez. Bu tür kategorik verileri sayısal verilere dönüştürmek için **One-Hot Encoding** kullanılır. One-Hot Encoding, modelin her kategori için ayrı bir sütun oluşturmasını sağlar. Örneğin ana veride cinsiyet değişkeni cinsiyet\_1 ve cinsiyet\_2’ye dönüşür (1: Erkek, 2: Kadın). Modelin kategoriler arasında yanlış ilişkiler kurmasının önüne geçer ve daha doğru sonuçlar elde edilmesine yardımcı olur. Bu, modelin hangi kategoriye ait olduğunu öğrenmesini ve hangi özelliğin modelin sonucunda etkili olduğunu belirlemesini kolaylaştırır. Her kategoriyi eşit derecede değerlendirir.

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

Açıklama otomatik olarak oluşturuldu

Şekil 15. One-Hot Encoding İşlemi Çıktısı

# Model Seçimi

Makine öğrenmesi (ML) modelleri, verilerden öğrenmek ve tahminler yapmak için kullanılan algoritmalardır. Farklı problemler için çeşitli makine öğrenmesi teknikleri ve algoritmaları bulunmaktadır. Analizde lojistik regresyon, random forest, gradientboosting classifier, KNN, XGBoost, karar ağaçları, destek vektör makineleri, yapay sinir ağları yöntemleri denenmiştir. Oluşturulan değişkenler ile doğruluk hesaplaması, karışıklık matrisi, sınıflandırma raporu, test hatası gibi modeli diğer modellerle karşılaştırılabilecek değerler elde edilecektir.

Doğruluk skoru, test verileri ile modelin tahmin ettiği değerler arasındaki doğruluğu verir. Doğruluk, doğru tahmin edilen örneklerin toplam örneklere oranıdır. Karışıklık matrisi modelin tahminlerinin gerçek etiketlerle karşılaştırıldığı matris oluşturur. 4 farklı durum gösterir:

TP, modelin doğru olarak pozitif (Astım hastası) tahmin ettiği sayılar,

FP, modelin yanlışlıkla pozitif (Astım hastası) tahmin ettiği sayılar.

TN, modelin doğru olarak negatif (Astım hastası olmayan) tahmin ettiği sayılar.

FN, Modelin yanlışlıkla negatif (Astım hastası olmayan) tahmin ettiği sayılar.

Sınıflandırma raporu ise, modelin performansını daha ayrıntılı bir şekilde gösteren metriklerin yer aldığı rapordur. İçerisinde hassasiyet, duyarlılık, F1 skoru, destekleri gösterir.

Hassasiyet, modelin pozitif sınıflandırma yaptığı örneklerden ne kadarının doğru olduğunu gösterir. Yüksek hassasiyet, yanlış pozitiflerin düşük olduğunu belirtir.

Duyarlılık, gerçek pozitiflerden ne kadarının doğru tahmin edildiğini gösterir. Yüksek duyarlılık, yanlış negatiflerin düşük olduğunu belirtir.

F1 skoru, duyarlılık ve hassasiyetin harmonik ortalamasıdır ve genel başarıyı ölçer.

Destek, her sınıftaki test verilerindeki örnek sayısıdır.

Test hatası, modelin test verileri üzerinde doğru tahmin yapma oranını gösteren bir hata metriği olarak kullanılır.

Tüm bu değerler kullanılarak tüm modeller denendikten sonra en iyi model seçilecektir.

## Random Forest

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

Şekil 19. KOAH ve Astım için Random Forest Modeli Kurulumu

Daha önceki modellerde olduğu gibi bağımsız ve bağımlı değişkenler belirlendikten sonra eğitim ve test setlerine ayrılmıştır. Eğitim ve test setlerinin ayıran rastgeleliğin tutarlı olabilmesi için random\_state=42 kullanılmıştır. stım ve KOAH için ayrı ayrı iki model oluşturulmuştur. Model eğitilirken 100 adet karar ağacı kullanılmış, her bir ağacın maksimum derinliği 10 olarak sınırlandırılmıştır. Tahminler test seti üzerinde yapılmıştır.

Doğruluk değerleri Astım için 0.887, KOAH için ise 0.8866olarak bulunmuş ve her iki modelin de sınıflandırma başarısı değerlendirilmiştir.

metin, ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

(2)

(1)

Şekil 20. Karışıklık Matrisleri

Astım için (1) 65 adet astım hastası olmayan doğru tahmin yapılmış, 67 tane ise astım hastası olup doğru tahmini yapılmıştır. 4 adet astım hastası olmayan ama modelin hasta olarak bulduğu hastalar var. 13 adet de astım hastası olup modelin hasta değil olarak numaralandırdığı hastaların sayısıdır.

KOAH için (2) 67 adet KOAH hastası olmayan doğru tahmin yapılmış, 66 tane ise KOAH hastası olup doğru tahmini yapılmıştır. 13 adet astım hastası olmayan ama modelin hasta olarak bulduğu hastalar var. 4 adet de astım hastası olup modelin hasta değil olarak numaralandırdığı hastaların sayısıdır.

Tablo . Astım için Sınıflandırma Raporu

Hassasiyet için astım ve KOAH örneği verilecekse örneğin 0’lar ele alınsın: astım 0.83 ile 100 sınıftan 83’ünü doğru tahmin ederken, KOAH 0.94 ile 100 taneden 94’ünü doğru tahmin etmiştir. Duyarlılık, TP’yi ne kadar doğru yakaladığını gösterir. Astım, 0 için 0.94 modelin 0’ı doğru tahmin ettiği oran, gerçek sınıf 0 olan 70 örnekten 66’ini doğru etmiştir. KOAH için de 0.83 ile gerçek sınıf 0 olan 80 örnekten 66’sını doğru tahmin etmiştir.

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

Açıklama otomatik olarak oluşturuldu

Tablo . KOAH için Sınıflandırma Raporu

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

Açıklama otomatik olarak oluşturuldu

F1 skoru, 0.88 olması hem duyarlılığın hem de kesinliğin yüksek olduğunu gösterir. Destek, astımda 0’da 70 örnek, 1’de 80 örnek bulunduğunu gösterir. Makro ortalama, her sınıfın **kesinlik, duyarlılık** ve **F1 skoru** gibi metriklerinin aritmetik ortalamasını alır. Her sınıfı eşit önemde değerlendirir. **Ağırlıklı ortalama**, her sınıfın metriklerinin, o sınıfın destek (örnek sayısı) ile çarpılmasını sağlar. Bu, dengesiz veri setlerinde daha doğru bir değerlendirme sağlar.

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

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 21. Her İki Model için ROC Eğrisi

X ekseni yanlış pozitif oranı, Y ekseni ise doğru pozitif oranını verir. AUC eğri altında kalan alandır, bu değer modelin sınıflar arasındaki ayrımı ne kadar iyi yaptığına dair bir gösterge sunar. AUC değeri 1'e ne kadar yakınsa, model o kadar iyi performans gösterir. Her iki modelin de 0 ve 1 sınıfları arasında oldukça iyi bir ayrım yaptığını gösterir. Kırmızı diagonal çizgi, rastgele tahmin yapan bir modelin performansını temsil eder. Bu çizginin üzerinde yer alan bir model, rastgele tahmin yapmaktan daha iyi performans gösteriyor demektir. AUC değeri 1'e yaklaştıkça, ROC eğrisi bu diagonal çizgiden olabildiğince uzaklaşır. Başarılı bir model kurulduğu söylenebilir. Mavi çizgi ise ROC eğrisinin kendisidir. **Doğru Pozitif Oranı (TPR)** ile **Yanlış Pozitif Oranı (FPR)** arasındaki ilişkiyi görselleştirir. Mavi çizgi ne kadar yukarıda ve sola yakınsa, modelin o kadar iyi performans gösterdiği anlamına gelir.

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

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 22. Hata Görselleştirilmesi

Y ekseni modelin hata değerlerini gösterir. Bu değer gerçek ve tahmin edilen değerler arasındaki farktır. X ekseni tahmin edilen değerleri gösterir. Mavi noktalar her bir gözlem için hesaplanan hata değeridir. Gerçek değerle farkını gösterir. Kırmızı kesikli çizgi ise ideal hata durumunu temsil eder. Hatalar sıfıra yakınsa, modelin tahminleri çok doğru demektir. Eğer noktalar çizginin etrafında düzensiz bir şekilde dağılmıyorsa modelin daha fazla iyileştirilmesi gerektiği anlamına gelebilir.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 23. Overfitting Kontrolü

Overfitting, modelin eğitim verisine çok iyi uyum sağlaması ancak test verisinde kötü performans göstermesidir. Bu durumu tespit etmek için eğitim ve test doğruluğu arasındaki farkı kontrol edilir. Eğer eğitim doğruluğu çok yüksek ancak test doğruluğu düşükse, bu overfitting belirtisidir. Model eğitim verisindeki gürültüye aşırı uyum sağlamış olabilir. Çapraz doğrulama, daha tutarlı ve genellenebilir bir değerlendirme sağlar. Eğer çapraz doğrulama doğruluğu test doğruluğuna yakınsa, modelin genelleme kapasitesi iyi demektir. Elde edilen değerler KOAH için eğitim doğruluğu 1 iken test doğruluğu 0.88’dir. Arada büyük bir fark yoktur. Ortalama çapraz doğrulama doğruluğu 0.87 bulunmuş olup yakın değer gösterdiğine göre modelin genelleme kapasitesi iyi diyebiliriz. Astım için de benzer sonuçlar elde edilmiştir. Böylece modeller de overfitting olmadığı kanıtlanmıştır.

## KNN (K-Nearest Neighbors)

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 16. Lojistik Regresyon Modeli Kurulumu

Bağımsız ve bağımlı değişkenler belirlendikten sonra eğitim ve test setlerine ayrılmıştır. Eğitim ve test setlerini ayıran rastgeleliğin tutarlı olabilmesi için random\_state=42 kullanılmıştır. KNN (K-Nearest Neighbors) modeli için en uygun komşu sayısı GridSearchCV ile belirlenmiş ve Astım modeli için **6** olarak seçilmiştir. Model eğitildikten sonra test seti üzerindeki tahminler yapılmıştır.

KNN modelinin eğitim doğruluğu 1.0, test doğruluğu ise 0.8 olarak hesaplanmıştır.

Doğruluk değerlerinin Astım için 0.80 olduğu, sınıflandırma raporunda ise Sınıf 0 için 0.81 ve Sınıf 1 için 0.79 F1 skoru elde edildiği gözlemlenmiştir. Bu sonuçlar, KNN modelinin astım sınıflandırması için uygun bir performans gösterdiğini ortaya koymaktadır.

KNN (K-Nearest Neighbors) KOAH modeli için en uygun komşu sayısı, GridSearchCV ile **3** olarak belirlenmiştir.

Astım için genel doğruluk **0.80,** KOAH için ise **0.79** olarak bulunmuştur. Bu sonuçlar, KNN modelinin astım ve KOAH sınıflandırmasında belirli bir doğruluk sağladığını göstermektedir.

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

(2)

(1)

Şekil 18. Karışıklık Matrisleri

Astım için (1) 63 adet astım hastası olmayan doğru tahmin yapılmış, 57 tane ise astım hastası olup doğru tahmini yapılmıştır. 7 adet astım hastası olmayan ama modelin hasta olarak bulduğu hastalar var. 23 adet de astım hastası olup modelin hasta değil olarak numaralandırdığı hastaların sayısıdır.

KOAH için (2) 62 adet KOAH hastası olmayan doğru tahmin yapılmış, 57 tane ise KOAH hastası olup doğru tahmini yapılmıştır. 18 adet astım hastası olmayan ama modelin hasta olarak bulduğu hastalar var. 13 adet de astım hastası olup modelin hasta değil olarak numaralandırdığı hastaların sayısıdır.

Tablo 1. Astım için Sınıflandırma Raporu

Hassasiyet için astım ve KOAH örneği verilecekse örneğin 0’lar ele alınsın: astım 0.73 ile 100 sınıftan 73’ünü doğru tahmin ederken, KOAH 0.82 ile 100 taneden 82’sini doğru tahmin etmiştir. Duyarlılık, TP’yi ne kadar doğru yakaladığını gösterir. Astım, 0 için 0.9 modelin 0’ı doğru tahmin ettiği oran, gerçek sınıf 0 olan 70 örnekten 63’ünü doğru etmiştir. KOAH için de 0.77 ile gerçek sınıf 0

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

Açıklama otomatik olarak oluşturuldu

Tablo 2.KOAH için Sınıflandırma Raporu

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

Açıklama otomatik olarak oluşturuldu

olan 80 örnekten 61’ini doğru tahmin etmiştir. F1 skoru, 0.8 olması hem duyarlılığın hem de kesinliğin orta derece iyi olduğunu gösterir. Destek, astımda 0’da 70 örnek, 1’de 80 örnek bulunduğunu gösterir. Makro ortalama, her sınıfın **kesinlik, duyarlılık** ve **F1 skoru** gibi metriklerinin aritmetik ortalamasını alır. Her sınıfı eşit önemde değerlendirir. **Ağırlıklı ortalama**, her sınıfın metriklerinin, o sınıfın destek (örnek sayısı) ile çarpılmasını sağlar. Bu, dengesiz veri setlerinde daha doğru bir değerlendirme sağlar.

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

Açıklama otomatik olarak oluşturuldu metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

X ekseni yanlış pozitif oranı, Y ekseni ise doğru pozitif oranını verir. AUC eğri altında kalan alandır, bu değer modelin sınıflar arasındaki ayrımı ne kadar iyi yaptığına dair bir gösterge sunar. AUC değeri 1'e ne kadar yakınsa, model o kadar iyi performans gösterir. Her iki modelin de 0 ve 1 sınıfları arasında oldukça iyi bir ayrım yaptığını gösterir. Kırmızı diagonal çizgi, rastgele tahmin yapan bir modelin performansını temsil eder. Bu çizginin üzerinde yer alan bir model, rastgele tahmin yapmaktan daha iyi performans gösteriyor demektir. AUC değeri 1'e yaklaştıkça, ROC eğrisi bu diagonal çizgiden olabildiğince uzaklaşır. Başarılı bir model kurulduğu söylenebilir. Mavi çizgi ise ROC eğrisinin kendisidir. Doğru Pozitif Oranı (TPR) ile Yanlış Pozitif Oranı (FPR) arasındaki ilişkiyi görselleştirir. Mavi çizgi ne kadar yukarıda ve sola yakınsa, modelin o kadar iyi performans gösterdiği anlamına gelir.

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

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 22. Hata Görselleştirilmesi

Y ekseni modelin hata değerlerini gösterir. Bu değer gerçek ve tahmin edilen değerler arasındaki farktır. X ekseni tahmin edilen değerleri gösterir. Mavi noktalar her bir gözlem için hesaplanan hata değeridir. Gerçek değerle farkını gösterir. Kırmızı kesikli çizgi ise ideal hata durumunu temsil eder. Hatalar sıfıra yakınsa, modelin tahminleri çok doğru demektir. Eğer noktalar çizginin etrafında düzensiz bir şekilde dağılmıyorsa modelin daha fazla iyileştirilmesi gerektiği anlamına gelebilir.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 23. Overfitting Kontrolü

Overfitting, modelin eğitim verisine çok iyi uyum sağlaması ancak test verisinde kötü performans göstermesidir. Bu durumu tespit etmek için eğitim ve test doğruluğu arasındaki farkı kontrol edilir. Eğer eğitim doğruluğu çok yüksek ancak test doğruluğu düşükse, bu overfitting belirtisidir. Model eğitim verisindeki gürültüye aşırı uyum sağlamış olabilir. Çapraz doğrulama, daha tutarlı ve genellenebilir bir değerlendirme sağlar. Eğer çapraz doğrulama doğruluğu test doğruluğuna yakınsa, modelin genelleme kapasitesi iyi demektir. Elde edilen değerler KOAH için eğitim doğruluğu 0.88 iken test doğruluğu 0.79’dur. Arada büyük bir fark yoktur. Ortalama çapraz doğrulama doğruluğu 0.81 bulunmuş olup yakın değer gösterdiğine göre modelin genelleme kapasitesi iyi diyebiliriz. Astım için de benzer sonuçlar elde edilmiştir. Böylece modeller de overfitting olmadığı kanıtlanmıştır.

## DestekVektörAğları

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

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

Açıklama otomatik olarak oluşturuldu

Şekil 16. Lojistik Regresyon Modeli Kurulumu

Destek Vektör Makineleri (SVM) algoritması, verileri farklı sınıflara ayıran bir hiper düzlem bulmaya çalışır ve burada, RBF (Radial Basis Function) kernel kullanılarak Astım ve KOAH sınıflandırma modelleri oluşturulmuştur. SVM, modelin doğruluğunu artırmaya yönelik C parametresi ve gamma parametresi gibi hiperparametrelerle ayarlanabilir. Astım için oluşturulan modelde, doğruluk %74 olarak bulunmuş ve karışıklık matrisi, 0 sınıfı (Astım Yok) ve 1 sınıfı (Astım Var) arasındaki tahmin hatalarını göstermektedir. Sınıflandırma raporunda, Astım Var sınıfının yüksek bir recall değeri (0.93) olduğu, ancak Astım Yok sınıfının daha düşük bir recall değerine (0.53) sahip olduğu gözlemlenmiştir. Bu sonuçlar, modelin sınıflar arasındaki dengesizliği iyi bir şekilde ele alamadığını, özellikle Astım Yok sınıfında daha fazla yanlış negatif sonuç verdiğini gösteriyor. SVM modeli, Astım için uygulanmış ve benzer şekilde, KOAH için de aynı işlem yapılabilir.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

(2)

(1)

Şekil 18. Karışıklık Matrisleri

Astım için (1) 37 adet astım hastası olmayan doğru tahmin yapılmış, 74 tane ise astım hastası olup doğru tahmini yapılmıştır. 33 adet astım hastası olmayan ama modelin hasta olarak bulduğu hastalar var. 6 adet de astım hastası olup modelin hasta değil olarak numaralandırdığı hastaların sayısıdır.

KOAH için (2) 74 adet KOAH hastası olmayan doğru tahmin yapılmış, 37 tane ise KOAH hastası olup doğru tahmini yapılmıştır. 6 adet astım hastası olmayan ama modelin hasta olarak bulduğu hastalar var. 33 adet de astım hastası olup modelin hasta değil olarak numaralandırdığı hastaların sayısıdır.

Tablo 1. Astım için Sınıflandırma Raporu

Hassasiyet için astım ve KOAH örneği verilecekse örneğin 0’lar ele alınsın: astım 0.86 ile 100 sınıftan 86’sını doğru tahmin ederken, KOAH 0.69 ile 100 taneden 69’sini doğru tahmin etmiştir. Duyarlılık, TP’yi ne kadar doğru yakaladığını gösterir. Astım, 0 için 0.52 modelin 0’ı doğru tahmin ettiği oran, gerçek sınıf 0 olan 70 örnekten 36’sını doğru tahmin etmiştir. KOAH için de 0.92 ile gerçek sınıf 0

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

Açıklama otomatik olarak oluşturuldu

Tablo 2.KOAH için Sınıflandırma Raporu

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

Açıklama otomatik olarak oluşturuldu

olan 80 örnekten 73’ünü doğru tahmin etmiştir. F1 skoru, 0.79 olması hem duyarlılığın hem de kesinliğin orta derece iyi olduğunu gösterir. Destek, astımda 0’da 70 örnek, 1’de 80 örnek bulunduğunu gösterir. Makro ortalama, her sınıfın **kesinlik, duyarlılık** ve **F1 skoru** gibi metriklerinin aritmetik ortalamasını alır. Her sınıfı eşit önemde değerlendirir. **Ağırlıklı ortalama**, her sınıfın metriklerinin, o sınıfın destek (örnek sayısı) ile çarpılmasını sağlar. Bu, dengesiz veri setlerinde daha doğru bir değerlendirme sağlar.

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

Açıklama otomatik olarak oluşturuldu metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

X ekseni yanlış pozitif oranı, Y ekseni ise doğru pozitif oranını verir. AUC eğri altında kalan alandır, bu değer modelin sınıflar arasındaki ayrımı ne kadar iyi yaptığına dair bir gösterge sunar. AUC değeri 1'e ne kadar yakınsa, model o kadar iyi performans gösterir. Her iki modelin de 0 ve 1 sınıfları arasında oldukça iyi bir ayrım yaptığını gösterir. Kırmızı diagonal çizgi, rastgele tahmin yapan bir modelin performansını temsil eder. Bu çizginin üzerinde yer alan bir model, rastgele tahmin yapmaktan daha iyi performans gösteriyor demektir. AUC değeri 1'e yaklaştıkça, ROC eğrisi bu diagonal çizgiden olabildiğince uzaklaşır. Başarılı bir model kurulduğu söylenebilir. Mavi çizgi ise ROC eğrisinin kendisidir. Doğru Pozitif Oranı (TPR) ile Yanlış Pozitif Oranı (FPR) arasındaki ilişkiyi görselleştirir. Mavi çizgi ne kadar yukarıda ve sola yakınsa, modelin o kadar iyi performans gösterdiği anlamına gelir.

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

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil 22. Hata Görselleştirilmesi

Y ekseni modelin hata değerlerini gösterir. Bu değer gerçek ve tahmin edilen değerler arasındaki farktır. X ekseni tahmin edilen değerleri gösterir. Mavi noktalar her bir gözlem için hesaplanan hata değeridir. Gerçek değerle farkını gösterir. Kırmızı kesikli çizgi ise ideal hata durumunu temsil eder. Hatalar sıfıra yakınsa, modelin tahminleri çok doğru demektir. Eğer noktalar çizginin etrafında düzensiz bir şekilde dağılmıyorsa modelin daha fazla iyileştirilmesi gerektiği anlamına gelebilir.

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

Açıklama otomatik olarak oluşturuldu

Şekil 23. Overfitting Kontrolü

Overfitting, modelin eğitim verisine çok iyi uyum sağlaması ancak test verisinde kötü performans göstermesidir. Bu durumu tespit etmek için eğitim ve test doğruluğu arasındaki farkı kontrol edilir. Eğer eğitim doğruluğu çok yüksek ancak test doğruluğu düşükse, bu overfitting belirtisidir. Model eğitim verisindeki gürültüye aşırı uyum sağlamış olabilir. Çapraz doğrulama, daha tutarlı ve genellenebilir bir değerlendirme sağlar. Eğer çapraz doğrulama doğruluğu test doğruluğuna yakınsa, modelin genelleme kapasitesi iyi demektir. Elde edilen değerler KOAH için eğitim doğruluğu 0.72 iken test doğruluğu 0.74’tür. Arada büyük bir fark yoktur. Ortalama çapraz doğrulama doğruluğu 0.71 bulunmuş olup yakın değer gösterdiğine göre modelin genelleme kapasitesi iyi diyebiliriz. Astım için de benzer sonuçlar elde edilmiştir. Böylece modeller de overfitting olmadığı kanıtlanmıştır.