|  |  |
| --- | --- |
| AKÜ FEMÜBİD **XX** (202X) xxxxxx (xx s) | AKU J. Sci. Eng. **XX** (202X) xxxxxx (xx pp) |

**DOI:**

**Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi**

**Afyon Kocatepe University Journal of Science and Engineering**

**Araştırma Makalesi / Research Article**

**Ölüm Sayılarını Etkileyen Faktörlerin Analizi: Hastalıkların Ötesinde Diğer Etkenlerin Rolü**

**Ayşegül Sinem Karaman**

**1** *İstanbul Arel Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul.*

*e-posta: sinemkaramannn@gmail.com. ORCID ID:* [*http://orcid.org/0000-0001-1111-2222*](http://orcid.org/0000-0001-1111-2222)

Geliş Tarihi: ; Kabul Tarihi:

|  |  |
| --- | --- |
| ***Anahtar kelimeler***  Kronik Hastalıklar; Ölüm; CatBoost; RandomForest + GradientBoosting + K-Nearest Neighbors; K-Nearest Neighbors + XGBoost; | **Öz** |
| Dünyada küresel sağlık üzerinde derin etkilere, ölümlere sebep olan bazı hastalıklar vardır. Bunlar Özellikle kalp hastalıkları ve kanser gibi kronik hastalıklar olurken son zamanlarda artan bulaşıcı hastalıklar örneğin covid -19 gibi ciddi anlamda ölüm sonuçlarına yol açmıştır. Bu ölümlerin araştırılması, ölümlerin arkasında yatan farklı sebeplerin olup olmadığının belirlenmesi insan popülasyonu için üzerinde durulması gerekilen bir konudur. Bu makale de belirtilen hastalıkların Amerika’da ölümler ile ilişki ele alınmış, orada yaşayan insanların geçmiş kökenlerine bakılarak ölüm oranları üzerinde ne kadar etkili olduğu sonuçları değerlendirilmek istenmiştir. Bu çalışma için üç tane farklı model kullanılmış olup bunlar (CatBoost), (RandomForest + GradientBoosting + K-Nearest Neighbors), ve (K-Nearest Neighbors + XGBoost) modelleridir. |

**Analysis of Factors Affecting Death Numbers: The Role of Other Factors Beyond Diseases**

|  |  |
| --- | --- |
| ***Keywords***  CatBoost; RandomForest + GradientBoosting + K-Nearest Neighbors; K-Nearest Neighbors + XGBoost; Chronic Diseases, Death | **Abstract** |
| There are some diseases in the world that have profound effects on global health and cause deaths. While these are chronic diseases such as heart diseases and cancer, the recent increase in infectious diseases such as Covid-19 has caused serious death consequences. Investigating these deaths and determining whether there are different reasons behind the deaths is an issue that needs to be emphasized for the human population. In this article, the relationship between the diseases mentioned and deaths in America is discussed, and it is aimed to evaluate the effects of the diseases on death rates by looking at the past origins of the people living there. Three different models were used for this study: (CatBoost), (RandomForest + GradientBoosting + K-Nearest Neighbors), and (K-Nearest Neighbors + XGBoost) models. |

© Afyon Kocatepe Üniversitesi

# 

# 1. Giriş

Dünya genelinde sağlık, insanın hayatının devam ettirebilmesi, yaşamsal faaliyetlerini gösterebilmesi için gerekli olan en kritik alanlardan biridir. Günümüzde ülkeler arasında etnik köken, bölgesel farklılıklar, cinsiyet, yaş grubu gibi çeşitli faktörler sağlıksal problemin yanında ölümü de getirebiliyor.

(Bozkurt vd.2023). Bu makale de kullanılan

datasette Amerika’daki insanların geçirdiği

hastalıklardan ölen insanların sayısını vermektedir.

Bu hastalıklardan bahsedecek olursak eğer ilk olarak sepsis hastalığıdır. Sepsis hastalığı; organ fonksiyon bozukluğuna neden olduğu bilinen bir hastalıktır. Dünya çapındaki ölümlerin %19.7’sini kapsar. (Zhang vd.2022). Malignant neoplasms; normal işlev yapan hücrelerin artık normalinden fazla çoğalmasıyla işlevini yerine getirememe durumudur. Bu hücrelere kötü huylu hücreler de denir.(Solid vd.2023).

Diabetes Mellitus; pankreas yeterli miktarda insülin üretemediğinde vücut ise bunu kullanamadığında ortaya çıkan bir hastalıktır.(Chang vd.,2023).

Alzheimer's disease; Dünya çapında demansın önde gelen bir hastalığıdır. Amiloid β peptidlerin hücre dışı birikimi ve hücre içi nörofibriler yumaklar halinde oluşumudur.( Uwişema vd.2022).

İnfluenza and pneumonia; iki farklı solunum yolu hastalığıdır. Yüksek ateş, boğaz ağrısı, öksürük gibi belirtileri vardır.( Owolabi vd.2022).

Chronic lower respiratory tract diseases; yaşlı ve yetişkinlerde bulunan ciddi solunum yolu hastalıklarıdır. ABD’de ise bu hastalık mevsimsel solunum yolu rahatsızlığı olarak geçer.(Melgar vd.2023).

Nephritis nephrotic syndrome And nephrosis; NPHS1 (nefrin) genindeki mutasyonların neden olduğu CNS'dir.Genellikle yenidoğan döneminde ağır proteinüriye neden olur. Heart diseases; Kalbin fonksiyonkarının işlevlerini yerine getirememesi durumudur. Genelde damarları etkiler.

(TR vd.2022).

Beyindeki kan damarları fonksiyonunu etikeleyen bir hastalıktır.( Nguyen vd.2022).

Bir diğer inceleyeceğimiz hastalık ise COVID-19; 2019 yılında çıkan özellike solunum yolunu etkileyen bulaşıcı bir hastalık türüdür.(Owolabi vd. 2022).Günümüzde gelişen teknolojiyle birlikte makine öğrenmesi artık hemen hemen her alanda kullanılmaktadır. Hastalıkların incelenmesi, ölüm sebeplerinin teşhis edilmesi gibi bir çok nedenden ötürü sağlık alanında popülerleşmiştir.

Bu çalışmada makine öğrenim metodları kullanılarak Amerika’daki ölüm sayılarının hastalıklar üzerinde etkisi ve bunun yanında çeşitli faktörlerin ölüm sayıları üzerindeki etkisi ele alınmıştır. Daha öncesinde bu konuya benzer bir kaç çalışmadan bahsedecek olursak eğer Makine öğrenim modeli kullanarak diyabet hastalığının teşhisi hakkında bir çalışma yapılmıştır. Burada K-EYK,C4.5 algoritması ve RO modelleri kullanılmıştır. (Bölükbaşı 2023)Bir diğer çalışma ise Optimize Edilmiş denetimli öğrenme algoirtmaları ile obezite hastalığına teşhis burada knn ve rastgele orman modelleri kullanılmıştır. Hatalık teşhis etmek amacıyla yazılmıştır(Turan 2023) Genel olarak çalışmalar tek hastalık üzerine yapılmıştır. Bu makaledeki çalışmada ise Öncelikle veri ön işleme aşamaları kullanılmıştır. Daha sonra çeşitli modeller uygulanmış olup bu modeller hibrit olarak kullanılmıştır.Bunlar: 'Sex' sütunu, 'M' değeri için 1, 'F' değeri için 0 olarak kodlanıyor, 'AgeGroup' sütunu kategorik olarak işlenip kodlanıyor, 'Race/Ethnicity' sütunu Label Encoding yöntemiyle sayısal değerlere dönüştürülüyor,

Min-Max normalizasyonu ile sayısal sütunlar normalize ediliyor. Hastalık ölüm sayılarının birkaç değişkene göre modellemeleri yapılmıştır. İlk olarak hastalıkların ölüm sayıları üzerinde etnik köken değişkeni ele alınmıştır. Irk/Etnik kökenin hastalıklar üzerindeki etkisine bakılmıştır. Bağımlı değişken etnik köken olarak seçilip bağımsız değişkenler ise tüm hastalıklar olarak ayarlanmıştır.

Burada catboost kütüphanesini kullanarak bir sınıflandırma modeli oluşturmayı ve eğitmeyi sağlar.Doğruluk sonucu burada 0.82 olarak bulunmuştur.

Bir sonraki ele alınan konu ise cinsiyet değişkeninin hastalıkların ölüm sayıları üzerindeki etkisini amaçlamaktadır. Burada ölüm sebeplerinin cinsiyet ile etkisi nedir? Bu sorunun cevabı bulunmak istenmiştir.Burada bağımlı değişken olarak cinsiyet değişkeni verilmiş olup bağımsız değişkenler ise ölüme sebep olan hastalıklar olarak ayarlanmıştır. Burada Random Forest, Gradient Boosting, K-Nearest Neighbors bir araya getirerek bir hibrit model oluşturulmuştur. Burada cinsiyetin ölüm sayıları üzerindeki etkisine bakılmıştır ve doğruluk sonucu 0.87 olarak bulunmuştur. Daha sonra alınan sonuçlara göre kadın ve erkeklerde en çok ölüme sebep olan hastalığın ne olduğu sonucu ele alınmıştır ve her iki cinsiyet değişkeni için de Diabetes mellitus hastalığı en çok ölüme sebep olan hastalık olarak bulunmuştur. Bir sonraki ele alınan konu ise yaş gruplarının hastalıklardaki ölüm sayıları üzerinde etkisine bakılmıştır. Yaş grupları kullanılan dataset üzerinde kategorik olarak ele alınmıştır ancak uygulama için sayısal değere çevrilmiştir.Burada KNN ve XGBoost modelleri bir arada kullanılmış olup Voting Classifier modeli oluşturulmuştur. Burada 0.84 oranında doğruluk alınmıştır. Daha sonra ölüm sayılarının yaş grupları üzerinde etsini görebilmek için çubuk grafik oluşturulmuştur. Daha sonra zaman serisi grafikleri, confisuon matrix ve korelasyon analizleri yapılarak bulunan sonuçlar desteklenmek istenmiştir.

|  |  |
| --- | --- |
| Veriseti Değişkenler | Değişken Açıklaması |
| AnalysisDate | Analizin tarih |
| Date Of Death Year | ölüm yılı |
| Date Of Death Month | ölüm ayı |
| Start Date | Araştırma başlangıç tarihi |
| End Date | Araştırma bitiş tarih |
| Jurisdiction of Occurrence | Nerede Yapıldığı |
| Sex | Cinsiyet |
| Race/Ethnicity | Etnik Köken/Irk |
| AgeGroup | Yaş grubu |
| AllCause | Tüm nedenden ölen kişi sayısı |
| NaturalCause | Doğal yolla ölen kişi sayısı |
| Septicemia (A40-A41) | Septisemi ölen kişi sayısı |
| Malignant neoplasms (C00-C97) | Malign neoplazmlar ölen sayı |
| Diabetes mellitus (E10-E14) | Diabetes Mellitus ölen sayısı |
| Alzheimer disease (G30) | Alzheimer ölen kişi sayısı |
| Influenza and pneumonia (J09-J18 | alt solunum yolu ölen sayısı |
| Chronic lower respiratory diseases (J40-J47 Other diseases of respiratory system (J00-J06,J30-J39,J67,J70-J98) | Kronik solunum sistemi hastalığından ölen sayısı |
| Nephritis, nephrotic syndrome and nephrosis (N00-N07,N17-N19,N25-N27) | Nefrit, nefrotik sendrom ve nefroz hastalığından ölen kişi sayısı |
| Symptoms, signs and abnormal clinical and laboratory findings not elsewhere classified (R00-R99) | Semptomlar, belirtiler ve anormal klinik ve laboratuvar bulguları hastalığından ölen kişi sayısı |
| Diseases of heart (I00-I09,I11,I13,I20-I51) | Kalp hastalığından ölen kişi sayısı |
| Cerebrovascular diseases (I60-I69) | Serebrovasküler hastalığından ölen kişi sayısı |
| COVID-19 (U071, Multiple Cause of Death) | COVID-19 hastalığından ölen kişi sayısı |
| COVID-19 (U071, Underlying Cause of Death) | COVID-19 hastalığından ölen kişi sayısı |

**Çizelge 1.** Veri Seti Tablosu

**2. Çalışmada Kullanılan Veri Seti**

Çalışma için kullanılan veri seti data.gov sitesinden alınmış olup Hastalıkların ölüm sayıları ile ilişkisini bize vermektedir. Veri setinde 23 tane değişken kullanılmış olup bunlar; Analiz Tarihi, Ölüm Tarihi Yılı, Ölüm Tarihi Ayı, Başlangıç Tarihi, Bitiş Tarihi, Oluşum Yeri, Cinsiyet, Irk/Etnik Köken, Yaş Grubu, Tüm Nedenler, Doğal Neden, Septisemi (A40-A41), Malign neoplazmlar (C00-C97) ), Diabetes Mellitus (E10-E14),Alzheimer hastalığı (G30),Grip ve zatürre (J09-J18),Kronik alt solunum yolu hastalıkları (J40-J47),,"Solunum sisteminin diğer hastalıkları (J00 J06,J30) -J39,J67,J70))-J98)" Nefrit,nefrotik sendrom ve nefroz (N00-N07,N17-N19,N25-N27)""Semptomlar, belirtiler ve anormal klinik ve laboratuvar bulguları, başka yerde sınıflandırılmamış (R00-R99) )" Kalp hastalıkları (I00-I09,I11,I13,I20-I51)",Serebrovasküler hastalıklar (I60-I69),"COVID-19 (U071, Çoklu Ölüm Nedenleri)","COVID-19 (U071, Ana Neden) Ölüm) )" değişkenleridir. Oluşum Yeri, Cinsiyet, Irk/Etnik Köken, Yaş Grubu değişkenleri kategorik değişkenler olup diğer kalan değişkenler ise sayısal türde veri içermektedir. Veri setinde toplamda 3961 veri içermektedir. Bu verileri toplamak için belirli zaman aralığında çalışmalar yapılmıştır.2019 yılında belirli sürelerde her bir hastalığa bağlı ölümün sonuçları toplanmış analizler ise 2021 yılında yapılmıştır. Bu veriler ABD Sağlık Ve İnsani Hizmetler Bakanlığı Data tarafından toplanmış olup doğruluk oranı oldukça fazla bir veri setidir. Bu verilere daha kolay ulaşımın sağlanması adına makalenin ekinde tablo olarak sunulmuştur.

**3) Uygulama Öncesi Veri İşleme**

Model eğitimine başlamadan önce veriler ön işleme süreçlerine koyulmuştur.

***3.1 Tarih Bilgilerini düzenleme***

AnalysisDate, Date Of Death Year, Date Of Death Month, Start Date, ve End Date sütunlarını kullanarak tarih verilerini işleriz. Start Date ve End Date sütunları kullanılarak analizin başlangıç ve bitiş tarihleri arasındaki gün sayısı bulunuyor. DeathDuration adlı bir sütun oluşturulup gün sayılarını bu sütuna yazıyoruz bunu yapmaktaki amaç ise korelasyon analizinde geçen sürenin ölüm sayıları üzerinde etkisin olup olmadığını araştırmak, hangi hastalıklarda etkili onu ölçeklendirmektir

***3.2 Kategorik değerleri Sayısala Çevirme***

Yapılan projede datasette kategorik olan ‘Sex’, ‘AgeGroup’ ve ‘Race/Ethnicity’ sütunları işleme alınması için sayısal değerlere çevirlir. Sex sütununda kadınlara 1 erkeklere 0 sayısal değerleri verilir, AgeGroup sütununda ise benzersiz sayılar atanır. Race/Ethnicity sütunu ise label encoding işlemi uygulanır ve yeni sütun oluşturulup eski sütun data setten silinir.

**Şekil 1**. Korelasyon Analizi

***3.4 Korelasyon Analizi***

Ölüm olaylarının bulduğumuz gün sütunundaki gün sayısı ile çeşitli hastalıkların neden olduğu ölüm sayıları arasındaki korelasyonu incelemek için kullanılmıştır.

***3.4 Veri Setindeki Eksik Ve Gereksiz Veri Kontrolü***

Kullandığımız veri setindeki eksik veriler kontrol edilir ve eksik veri var ise buralar ortalama değeriyle doldurulur.Daha sonra gereksiz veriler; Jurisdiction of Occurrence', 'AnalysisDate' sütunları kaldırılır.

***3.5 Normalleştirme İşlemleri***

Burada Veri setindeki sayısal değerlere min-max normalizasyon işlemi uygulanır ki sayısal değerler 0 ile 1 aralığına denk gelmelidir .Geliştirilen modeller daha doğru sonuç versin çünkü ölüm sayıları çok yüksek sayılardan oluşuyor.

metin, ekran görüntüsü, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldu

**4. Makine Öğrenim Algoritmaları**

Bu bölümde çalışmada kullanılan makine öğrenim modellerinden bahsedilmiştir.

***4.1 CatBoost Classifier***

CatBoost, kategorik değişkenleri verimli bir şekilde işleyebilen ve hızlı eğitim süreleriyle bilinen bir

modeldir. Kategorik değişkenleri etkili bir şekilde işleme yeteneği ve paralel işleme yetenekleri ile

büyük veri kümeleri üzerindeki performansını artırır. Öğrenme oranını otomatik olarak belirlediğinden ötürü aşırı uydurmayı büyük oranda azaltır(Rathod vd.2022) .CatBoost, hem regresyon hem de sınıflandırma problemleri için kullanılabilir.Sayısal olarak açıklarsak eğer;

Model : Burada, M toplam ağaç sayısını, (x) ise her bir ağacın tahminini temsil eder.

(1)

Ağaç Tahmini: *w* ağırlık vektörünü, *q*(*x*) ise girdi özellikleri *x* üzerinden seçilen ağaç düğümlerini ifade eder.

(2)

**Kayıp Fonksiyonu (Loss Function): Gerçek değer ile tahmini değer arasındaki farkı çözer.**

(3)

Ağaç Karmaşıklığı : Burada maaliyet kontrol edilir.

(4)

Bu yapılan çalışmada ise target olarak 'Race/Ethnicity' , feature olarak ise tüm hastalık tipleri kullanılmıştır. Daha sonra train\_test\_split fonksiyonu kullanılarak veri setini, eğitim ve test setlerine ayrıldı. CatBoostClassifier sınıfı kullanılarak bir model oluşturuldu. Modelin temel parametreleri, ağaç sayısı, ağaç derinliği, öğrenme oranı gibi parametreler olarak tanımlandı.

fit metodu ile oluşturduğumuz model eğitilir. eval\_set parametresi, eğitim sürecinde modelin performansını değerlendirmek için kullanıldı. Modelin doğruluk oranı (acc değeri) 0.82 olarak ölçülmüştür.

***4.2 Ensemble Modeli***

Topluluk yöntemleri, bir dizi sınıflandırıcı oluşturup ve daha sonra tahminlerinin sonuçlarını alarak yeni verileri sınıflandıran öğrenme algoritmalarıdır Bu model karmaşıklıkları ve çeşitlilikleri artırdıkları için daha yüksek tahmin performans sonuçları verebilirler.( Rezaei 2023)

Bu modeller iki temel kategoride incelenirler bunlardan ilki;

Bagging (Bootstrap Aggregating): Farklı alt örneklemler üzerinde eğitilir ve sonra bir araya getirilir.

Boosting: Her temel örnekte bir önceki hataları düzelterek ilerlemeye çalışır.

Matematiksel olarak ifade edersek model tahmini, oluşturulan modellerin tahmininin bir kombinasyonu olarak ifade edilir.

(5)

Bu çalışmada ensemble model olarak; Rastgele Orman, K-En Yakın Komşu, Gradient Boosting modelleri uygulanmıştır. Bu modellerden bahsedecek olursak eğer;

***4.2.1 Rastgele Orman***

Ensemble modelleme tekniğinde kullanılan bir sınıflandırma ve regresyon algoritmasıdır. Rastgele Orman, güçlü bir model oluşturmak için birçok karar ağacının bir araya getirilmesini sağlar. (Özkan 2023).

Bu makale deki ensemble modelde ilk olarak

n\_estimators değişkeni oluşturulacak ağaç sayısı,

random\_state değişkeni ile rastgelelik kontrolü için kullanılan tohum değeri ve class\_weight='balanced' değeri ile dengesiz sınıflara sahip veri setleri için sınıflar arasındaki dengesizliği ele almak için kullanılmıştır.

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

Açıklama otomatik olarak oluşturuldu**

**Şekil 2.**

***4.2.2 K-En Yakın Komşu***

Kolay ve etkili bir modeldir.. Temel prensibi, bir veri noktasının değerinin, komşularının değerlerine göre belirlendiği yöntemdir.En yakın komşu baz alınır.(Filiz 2023).

Matematiksel olarak ifade edersek eğer;

(6)

Şeklinde ifade edilebilir.

Bu proje de ise KNeighborsClassifier sınıflandırıcı modeli, knn\_model parametresiyle kullanılmıştır.

Daha sonra knn modeli diğer modellerle beraber VotingClassifier içinde kullanılmıştır.

metin, diyagram, yazı tipi, daire içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Şekil 3.**

***4.2.3 Gradient Boosting***

Her karar ağacı önceki karar ağacının hatasını tahmin edip  hatayı gradyan olarak *artırır ve iyileştirir.(* Malik vd.2022).

(7)

Bu formülde;

F(x) = toplam tahmini temsil eder.

F0(x) = başlangıç modelinin tahminidir.

h1(x),h2(x),…,hr(x)= sırasıyla eklenen modellerin tahminleridir.

metin, yazı tipi, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Şekil 4.** Gradient boosting algoirtması

Bu çalışma da ise GradientBoostingClassifier sınıflandırıcı modeli, gb\_model parametresiyle tanımlanmıştır. Ve daha sonra diğer modellerle bilrikte (randomforest, knn) VotingClassifier içinde kullanılıp bu üç model eğitilmiştir ve eğitilen veri seti üzerinde tahminler yapılmış gerçek değerle karşılaştırılma performansı değerlendirilmiştir. Bu proje de kategorik değer olan sex sütunu hedef değişken olarak seçildi ve ölüme neden olan tüm hastalıkların sayısını veren değişkenler ise bağımsız değişken olarak kullanıldı. Cinsiyetin ölüm sayılarının üzerinde bir etkisinin olup olmadığı ve varsa bu oranın kadınlar da ve erkeklerdeki etkisin ne olduğu, kadın ve erkeklerde ölüme en çok sebep olan hastalığın ne olduğu sonuçlarını elde etmeye çalışıldı. Bu modeller de ölçülen bazı değerler:

* Kadınlar için en çok ölünen hastalık: NaturalCause
* Erkekler için ise en çok ölünen hastalık: NaturalCause

Olarak ölçülmüştür.

***4.3 VotingClassifier*** ***Modeli***

Bu model çalışmamızda kullanılan bir diğer ve son modeldir. Burada da iki model birbiriyle hibrit hale getirilmiş şekilde kullanılmıştır.

Bu strateji, sınıflandırma modellerinin tahminlerini belirli bir ağırlıklandırma ile birlikte birleştirir. Birleştirdiği modellerin ürettikleri olasılık değerlerini kullanır ve bu olasılık değerlerini işleme alarak toplam bir olasılık skoru elde eder. Daha sonra, bu olasılık skoruna dayanarak bir tahmin yapılır.( Rai vd. 2022) Bu çalışmada ise knn ve xgboost modelleri birleştirilerek bir votingclassifier modelininin içinde birleştirilip bir sonuç elde edilmeye çalışılmıştır.

***4.3.1 K-En Yakın Komşu***

n\_neighbors=5 ile belirtilen yerde 5 en yakın komşuyu dikkate alan bir sınıflandırıcı model oluşturulmuştur. Daha sonra XGBoot modeliyle birlikte bir topluluk oarak değrlendirilmek için VootingClassifier modeli içine eklenmiştir.

***4.3.2 XGBoost Modeli***

Chen ve Guestrin tarafından geliştirilen bir topluluk ağacı algoritması olan XGBoost, tahmin performansı tek başına kullanılan tekniklerden daha iyi olan, hibrit bir model üretmek için kullanılan verimli bir sonuç üreten modeldir.( Tokmak 2023)

Matematiksel olarak temel formülü aşağıdaki gibi gösterilmedilir.

(8)

Burada iki değişkeni tahmin edilen modeli ifade etmektedir.

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

Açıklama otomatik olarak oluşturuldu

**Şekil 5.** XGBOOST model

Bu makale de ise knn ile birlşetirilip bir VotingClassifier Modeli içine eklenmiştir ve eğitim aşamasına geçilmiştir. Eğitilen değerler üzerinden tahminler yapılmıştır. Hedef değişken AgeGroup değişkeni olarak belirlenmiştir ve bağımsız değişkenler ise aynı şekilde hastalığa sebep olan ölüm sayılarının verdiği tüm hastalıklar seçilmiştir. Böylelikle yaş grubunun hastalıklardaki ölüm sayıları üzerinde bir sınıflandırma tahmini yapılmak istenmiştir. Alınan doğruluk oranı ise 0.84 olarak ölçülmüştür.

**5.Uygulanan Modellerden Elde Edilen Sonuçlar**

Yapılan araştırmalar ve uygulamalar sonucu Uygulanan modeller de precision, recall, F-1 core, support ve doğruluk oranları aşağıdaki gibidir.

***5.1 Catboost Classifier Sonuçları***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1-score | support |
| 0 | 0.87 | 0.87 | 0.87 | 129 |
| 1 | 0.70 | 0.80 | 0.75 | 128 |
| 2 | 0.80 | 0.80 | 0.80 | 143 |
| 3 | 0.92 | 0.86 | 0.89 | 125 |
| 4 | 096 | 0.95 | 0.96 | 124 |
| 5 | 0.72 | 0.66 | 0.69 | 143 |

***Çizelge 2.*** Catboost classifier sonuçları tabloda gözüktüğü gibidir***.***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy |  |  | 0.82 | 792 |
| Macro avg | 0.83 | 0.83 | 0.83 | 792 |
| Weighted avg | 0.82 | 0.82 | 0.82 | 792 |

***Çizelge 3.*** Catboost classifier sonuçları tabloda gözüktüğü

gibidir.

***5.2 Ensemble Model Sonuçları***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | support |
| 0.0 | 0.87 | 0.98 | 0.92 | 628 |
| 1.0 | 0.87 | 0.41 | 0.56 | 164 |
| Accuracy |  |  | 0.87 | 792 |
| Macro avg | 0.87 | 0.70 | 0.74 | 792 |
| Weighted avg | 0.87 | 0.87 | 0.85 | 792 |

***Çizelge 4.*** Ensemble Modelsonuçları tabloda gözüktüğü

gibidir.

|  |  |
| --- | --- |
| Ensemble Model Acc | 0.8661616161616161 |
| Kadınları en çok öldüren Hastalık | NaturalCase |
| Erkekleri en çok öldüren hastalık | NaturalCase |

***Çizelge 5.*** Ensemble Modelsonuçları tabloda gözüktüğü

gibidir.

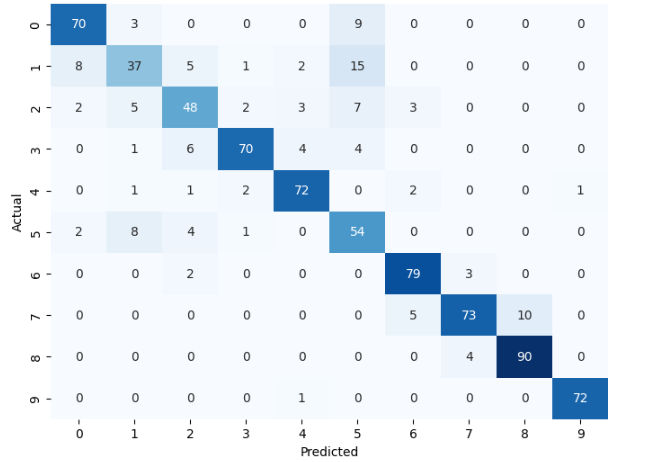
## *5.3. VotingClassifier* *Model Sonuçlar*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | support |
| 0 | 0.85 | 0.85 | 0.85 | 82 |
| 1 | 0.67 | 0.54 | 0.60 | 68 |
| 2 | 0.73 | 0.69 | 0.71 | 70 |
| 3 | 0.92 | 0.82 | 0.87 | 85 |
| 4 | 0.88 | 0.91 | 0.89 | 79 |
| 5 | 0.61 | 0.78 | 0.68 | 69 |
| 6 | 0.89 | 0.97 | 0.91 | 84 |
| 7 | 0.91 | 0.83 | 0.87 | 88 |
| 8 | 0.90 | 0.96 | 0.93 | 94 |
| 9 | 0.99 | 0.99 | 0.99 | 73 |
| Accuracy |  |  | 0.84 | 792 |
| Macro avg | 0.83 | 0.83 | 0.83 | 792 |
| Weighted avg | 0.84 | 0.84 | 0.84 | 792 |

***Çizelge 5.*** VotingClassifier Modelsonuçları tabloda gözüktüğü gibidir.

|  |  |
| --- | --- |
| Voting Model Accuracy | 0.8396464646464646464 |

***Çizelge 6.*** VotingClassifier Modelacc değeri tabloda gözüktüğü gibidir.



**Şekil 9.** Votingclassifier modelinin confision matrix sonucu.

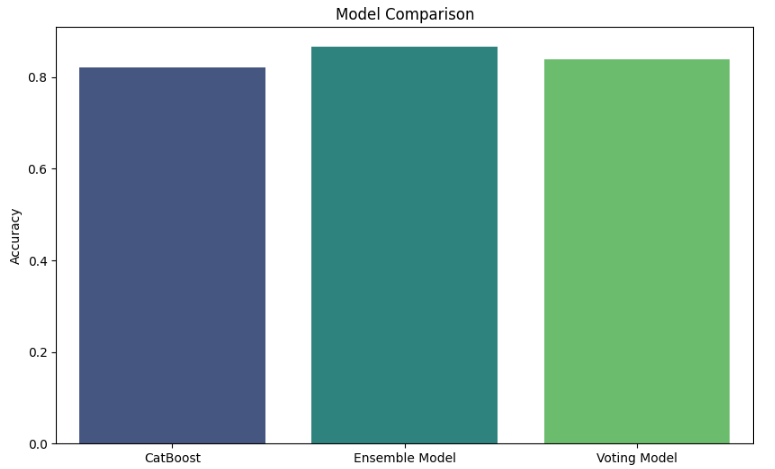
**ekran görüntüsü, metin, dikdörtgen, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Şekil 10.** Agegroup değişkeninin graiksel olarak gösterimi verilmiştir. X ekseninde yaş grupları, y ekseninde ise tahmin edilen yaşlar mevcuttur.

**6.Model Karşılaştırmaları**

Bir önceki bölümde farklı sınıflandırma modelleri kullanarak elde ettiğimiz üç farklı sonucun doğruluk oranlarını grafiksel olarak karşılaştırması yapılmıştır. Kullanılan sınıflandırma modelleri belirtilmiştir her bir model i bir dizi oluşturup sırasıyla değerler okunmuş ve aşağıdaki grafikte gösterilmiştir

****

**Şekil 11.** Model doğruluk karşılaştırılması

Bu grafikte elde edilen sonuçlara göre ensemble model üzerinde daha fazla doğruya yakın sonuç vermiştir.

**7.Değerlendirme Kriterleri**

Modelimizi eğittikten ve tahminleri yaptıktan sonra sonucunu aldığımız ve tahminin ne kadar doğru olduğunu öğrendiğimiz değerlendirme kriterleridir.

***7.1 Precision***

Precision, bir sınıflandırma modelinin pozitif olarak tahmin ettiği değerlerin gerçekten pozitif olan değerlerinin oranını ifade eden bir değerdir. (Oikonomou vd. 2023)

(9)

True Positives (TP): Modeli doğru bir şekilde pozitif olarak tahmin ettiği durumlarıdır.

False Positives (FP): Modeli yanlış bir şekilde pozitif olarak tahmin ettiği durumlarıdır.

Precision değeri 0 ile 1 arasında bir sayıdır. 1'e ne kadar yakınsa, modelin pozitif tahminlerinin doğru olduğu göstergesidir.(Miao vd. 2023)

***7.2 Recall***

Recall (geri çağırma), bir sınıflandırma modelinin gerçek pozitif değerleinin ne kadarını doğru tespit ettiğini gösteren bir performans metriğidir.

Burada da değerler 0 ile 1 arasında değer alır.1’e ne kadar yakınsa tahminlerin o kadar doğru olduğunu gösterir.( . Banerjee vd.2023)

***7.3 F-1 Score***

F1-score, hassasiyet (precision) ve geri çağırmanın (recall) harmonik ortalamasının alınmış halidir.. F1-score, bu iki kriteri birleştirerek bir modelin performansını tek bir değer olarak özetlemeyi sağlar.

metin, yazı tipi, beyaz, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu (10)

***7.4 Support***

Herhangi bir sınıfın gerçek veri setinde kaç kez görüntülendiğini temsil eder. Sınıflandırma raporlarında genellikle her sınıf için ayrı ayrı raporlanan örnek sayısını verir. Precision, recall, ve F1-score gibi metrikler hesaplanırken bu support değeri de göz önüne alınır.( Kufel vd 2023)

***7.5 Confision Matrix***

True Positive (TP): Gerçek pozitif modeldir. Modelin doğru bir şekilde pozitif olarak tahmin ettiği durumdur

True Negative (TN): Gerçek negatif model denir. Modelin doğru bir şekilde negatif olarak tahmin ettiği durumdur

False Positive (FP): Yanlış pozitif modeldir. Modelin pozitif olarak tahmin ettiği ancak gerçekte negatif olan durumdur.

False Negative (FN): Yanlış negatifbir modeldir. Modelin negatif olarak tahmin ettiği ancak gerçekte pozitif olan durumunu verir. (Febrian vd.2023)

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

Açıklama otomatik olarak oluşturuldu

***Şekil 12.*** Confision matrix değerinin görsel olarak gösterimi

***7.6 Zaman Serisi Analizi***

En son aldığımız sonuçlarla Verileri yıllara göre analiz ederek toplam ölüm sayılarını bulmak için bir zaman grafisi oluştururuz. Burada y değişkenleri gruplandırılmış ölüm sayılarını, x değişkeni ise ölüm yıl göstergesidir.

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

Açıklama otomatik olarak oluşturuldu

**Şekil 13**. Ölüm yıllarına göre sayıları gruplayan zaman serisi analizi

**8. Tartışma ve Sonuç**

Bu makalede Amerikadaki çeşitli hastalıklardan dolayı ölen insanların verileri üzerinden işlem yapılmıştır.Belirli kategorik değerler ile sayısal veriler üzerinde bir classification işlemi uygulanmıştır. Modeller eğitilip daha sonra tahminler yapılmıştır. Yapılan tahminler ile belirli kategorik değerlerin(AgeGroup, Sex, Race/Ethnicity sütunları hedef değişken olarak seçilip veri setinde bahsedilen diğer hastalıkların ölüm sayılarını veren sütunların tümü bağımsız değişken olarak seçilmiştir ve hedef değişkenlerin bağımsız değişkenler üzerinde ne kadar etkili olduğu ele alınmıştır. Her bir eğitim ve tahmin için farklı modeller kullanılıp ve hibrit modeller kullanılmaya özen gösterilmiştir. Çıkan sonuçlara göre de doğruluk oranları 0.80-0.90 arasında ölçülmüştür. Bu konu hakkında daha öncesinde hastalıkların tümüyle değil de sadece tek bir hastalık üzerinde özelleşmiş belirli çalışmalar yapılmış; diyabet,obezite bunlardan sadece birkaçıdır. Ancak bu çalışmada ise tüm hastalıkların ölüm sayılarının hangi değişkenlere bağlı olduğu üzerinde durulmuştur. Bu çalışma bir devletin ölümlerle ilgilenen bakanlıkları ,eyaletleri vs. tarafından kullanılması o bölge de ölen insanlar hakkında detaylı bilgi verebilmektedir. Hatta günümüzde de etkisini gösteren bulaşıcı hastalıklar gibi etkenlerden ölen insanların sağlık demografisi, ölümlerinin altında hangi sebeplerin olduğu hakkında detaylı bir bilgi verebilmektedir. Gelecekte daha derinlemesine analizler yapılarak hastaneler, ilgili devlet bakanlıkları, bulaşıcı hastalıklar ile ilgilenen kurum ve kuruluşlara gelişmiş mühendislik fırsatları açarak ölüme neden olan hastalıklar hakkında, belki de daha fazlası için bir gelişe fırsatı sunar.

**9. Kaynaklar**

BOZKURT, İ. (2023). GEÇMİŞTEN GÜNÜMÜZE ÜLKEMİZİN PANDEMİLERE YÖNELİK SAĞLIK POLİTİKALARINA GENEL BİR BAKIŞ. Atlas Sosyal Bilimler Dergisi, 1(13), 51-66.

Zhang, W., Fang, X., Gao, C., Song, C., He, Y., Zhou, T., Yang, X., Shang, Y., Xu, J. (2023). MDSCs in sepsis-induced immunosuppression and its potential therapeutic targets. Journal Name, 5 June 2022,

Soldi, LR, Silva, VLC, Rabelo, DH ve diğerleri. Malign neoplazmların mikroçevresinde doğal öldürücü hücrelerin monoklonal antikorlarla yeniden aktivasyonu. J Cancer Res Clin Oncol 149 , 6661–6673 (2023).

Chang, V., Bailey, J., Xu, QA ve diğerleri. Makine öğrenimi (ML) algoritmalarına dayalı Pima Kızılderilileri diyabet sınıflandırması. Nöral Hesaplama ve Uygulama 35 , 16157–16173 (2023).

Tokito T, Kido T, Muramatsu K, Tokutsu K, Okuno D, Yura H, Takemoto S, Ishimoto H, Takazono T, Sakamoto N, ve diğerleri. İnfluenza Virüsü Pnömonisi İçin Hastaneye Yatıştan Sonraki 7 Gün İçinde İntravenöz Azitromisinin Uygulanmasının Etkisi: Ülke Çapında İdari Veri Tabanı Kullanan Bir Eğilim Skoru Analizi. Virüsler . 2023; 15(5):1142.

Uwişema, Ö. , Mahmud, Bir güneş , J. , Correia, IFS , Bejjani, N. , Alwan, M. , Nicholas, A. , Ölüyemi, A. ve Dost, B. ( 2022 ). Alzheimer hastalığı bulaşıcı bir nörolojik hastalık mıdır? Literatürün gözden geçirilmesi . Beyin ve Davranış , 12 , e2728.

Owolabi, M., Ali, R., Dacosta, J., Muhanna, A., & Slim, J. (2022). When Influenza, Bacterial Pneumonia, and COVID-19 Co-exist. Cureus, 14(12).

Melgar M, Britton A, Roper LE, et al. Yaşlı Yetişkinlerde Solunum Sinsityal Virüs Aşılarının Kullanımı: Bağışıklama Uygulamaları Danışma Komitesinin Önerileri — Amerika Birleşik Devletleri, 2023. MMWR Morb Mortal Wkly Rep 2023;72:793–801.

Jalanko, H., Jahnukainen, T., Ng, KH (2022). Konjenital Nefrotik Sendrom. İçinde: Emma, F., Goldstein, SL, Bagga, A., Bates, CM, Shroff, R. (eds) Pediatrik Nefroloji. Springer, Cham.

TR, R. ., Lilhore, UK, M, P. ., Simaiya, S. ., Kaur, . A. . ve Hamdi, M. . (2022). MAKİNE ÖĞRENME YAKLAŞIMLARIYLA KALP HASTALIKLARININ ÖNGÖRÜLEBİLİR ANALİZİ. Malezya Bilgisayar Bilimleri Dergisi , 132–148.

Nguyen, AT, Kouri, N., Labuzan, SA ve diğerleri. MRI'daki difüzyon değişiklikleriyle ilişkili serebrovasküler hastalığın nöropatolojik ölçekleri. Acta Neuropathol 144 , 1117–1125 (2022).

KABA, G. ve BAĞDATLI KALKAN, S. (2022). KARDİYOVASKÜLER HASTALIK TAHMİNİNDE MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARININ KARŞILAŞTIRILMASI. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21(42), 183-193.

Bölükbaşı, İ. B. (2023). *Dengesiz Bir Diyabet Veri Setinde Makine Öğrenmesi Yöntemlerini Kullanarak Diyabet Hastalığının Teşhisi* (Doctoral dissertation, Bursa Uludag University (Turkey)).

Y. Rathod ve diğerleri , "CatBoost Algoritmasını Kullanarak Polikistik Over Sendromunun Tahmin Edici Analizi", 2022 IEEE Bölge 10 Sempozyumu (TENSYMP) , Mumbai, Hindistan, 2022, s. 1-6, doi: 10.1109/TENSYMP54529.2022.9864439.

Wang, Y., Wang, R., Wang, J. ve diğerleri. Bayesian Optimizasyon Catboost Algoritmasına Dayalı Dalga Hızı ve Operasyonel Parametreleri Bütünleştiren Bir Kaya Kütlesi Dayanımı Tahmin Yöntemi. KSCE J Civ Eng 27 , 3148–3162 (2023).

Baradaran Rezaei, H., Amjadian, A., Sebt, MV *ve diğerleri.*Mide kanserinin prognozunu tahmin etmek için makine öğreniminin birleştirilmiş bir yöntemi. *Ann Oper Res*328 , 151–192 (2023).

ÖZKAN, B., PARİM, C. ve ÇENE, E. (2023). Ülkelerin Gelişmişlik Düzeylerinin Karar Ağacı ve Rastgele Orman Yöntemleriyle Tahmin Edilmesi. EKOIST Ekonometri ve İstatistik Dergisi ( 38 ), 87-104.

FİLİZ, E. (2022). Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasına yönelik tahmininin destek vektör makineleri ve k-en yakın komşu algoritmaları ile gerçekleştirilmesi. Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 12(1), 370-379.

Malik, A., Saggi, M. K., Rehman, S., Sajjad, H., Inyurt, S., Bhatia, A. S., ... & Yaseen, Z. M. (2022). Deep learning versus gradient boosting machine for pan evaporation prediction. Engineering Applications of Computational Fluid Mechanics, 16(1), 570-587.

Rai, N., Kaushik, N., Kumar, D., Raj, C. ve Ali, A. (2022). Yumuşak oylama sınıflandırıcısını kullanarak COVID-19 hastalarının ölüm tahmini. Uluslararası Mühendislikte Bilişsel Hesaplama Dergisi , 3 , 172-179.

TOKMAK, M. (2023). Öğrencilerin Siber Güvenlik Farkındalık Düzeylerinin Makine Öğrenmesi Yöntemleri ile Belirlenmesi. Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 28(2), 451-466.

*Oikonomou, EK, Khera, R. Hassas diyabet bakımı ve kardiyovasküler risk tahmininde makine öğrenimi. Cardiovasc Diabetol 22 , 259 (2023).*

*Miao, J., Zhu, W. Hassas hatırlama eğrisi (PRC) sınıflandırma ağaçları. Evrim. Intel. 15 , 1545–1569 (2022).*

*Banerjee, A., Mutlu, O. C., Kline, A., Surabhi, S., Washington, P., & Wall, D. P. (2023). Training and profiling a pediatric facial expression classifier for children on mobile devices: machine learning study. JMIR formative research, 7, e39917.*

*Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., ... & Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. Diagnostics, 13(15), 2582.*

*Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. Procedia Computer Science, 216, 21-30.*

TURAN, T. Optimize Edilmiş Denetimli Öğrenme Algoritmaları ile Obezite Analizi ve Tahmini. *Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *14*(2), 301-312.