Makine öğrenmesi algoritmalarıyla kalp hastalıklarının tespit edilmesine yönelik performans analizi

Elif ÇİL[1]
Ali GÜNEŞ[2]

Geliş tarihi / *Received:* 14.06.2021
Düzeltilerek geliş tarihi / *Received in revised form:* 23.12.2021
Kabul tarihi / *Accepted:* 12.01.2022
DOI: 10.17932/IAU.ABMYOD.2006.005/abmyod_v17i65004

Özet

Kişilerin karşı karşıya kaldıkları hastalıkların içinde kalp hastalıkları önemli bir yer tutmaktadır. Kalp hastalıklarının ortaya çıkmasının birçok nedeni olabilmekte ve kişiyi çok fazla etkileyebilmektedir. Ancak kişilerin kalp hastalığı nedeniyle yaşamın sonlamasının önüne geçilebilecek tedaviler mümkündür. Bu noktada erken teşhisin önemi büyüktür. Bu nedenle çalışmanın konusu makine öğrenmesi algoritmaları kullanılarak kalp hastalıklarının tespit edilmesinin performans analizi olarak seçilmiştir. Bu konuda yapılan araştırmaların sınırlı olmasıyla beraber makine öğrenme yöntemiyle birlikte bazı araştırmalar bulunmaktadır. Makine öğrenmesi yöntemlerinden; destek vektör makineleri, rastgele orman algoritmaları, yapay sinir ağları, Naive Bayes ve k-NN bu çalışma içinde karşılaştırılarak analizi yapılmıştır.

Anahtar Kelimeler: *Makine Öğrenmesi, Yapay Zekâ, Veri İşleme Teknikleri, Kalp Hastalıkları.*

[1] *İstanbul Aydın Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği, İSTANBUL/ TÜRKİYE*
elifcil@stu.aydin.edu.tr, https://orcid.org/0000-0003-3229-2020
[2] *İstanbul Aydın Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği, İSTANBUL/ TÜRKİYE*
aligunes@aydin.edu.tr, https://orcid.org/0000-0001-6177-3136

Performance analysis of machine learning algorithms used to detect heart diseases

Abstract

Heart diseases have an important place among the diseases that people face. There can be many reasons for the occurrence of heart diseases and it can affect the person very much. However, some treatments that can prevent people from ending life due to heart disease. At this point, early diagnosis is of great importance. For this reason, the subject of the study was chosen as the performance analysis of the detection of heart diseases using machine learning algorithms. Although the research on this subject is limited, there is some researches with machine learning method. From machine learning methods; support vector machines, random forest algorithms, artificial neural networks, Naive Bayes, and k-NN were compared and analyzed in this study.

Keywords:*Machine Learning, Artificial Intelligence, Data Processing Techniques, Heart Diseases.*

Giriş

Günümüzde karşılaşılan hastalıklar içinde kalp hastalıkları önemli bir yer tutmaktadır. Kalp rahatsızlıkları kontrol altına alınmadıklarında ölümcül olmak ile beraber dünyada ölüm nedenleri arasında başta yer almaktadır. Kalpte meydana gelen birçok rahatsızlık vardır; kalpten çıkan ana damarların hastalıkları, kalp kapakçık hastalıkları, koroner damar yetmezliği, kalbin etrafını saran zarın hastalıkları, doğuştan kalp hastalıkları, kalp delikleri, kalp yetmezliği, kalbin delici cisimlerle hasarı gibi hastalıklar. Koroner arter hastalığı ya da iskemik kalp damar hastalığı, genellikle bu damarların damar sertleşmesi sebebi ile tıkanması ya da daralmasıdır. Damar sertliği, damarın iç kısmında yağ birikintileri ile kolesterolün ortaya çıkmasıdır. Neticede mekanik şekilde bozulan ve

tıkanan damarların anormal halleri sebebi ile kan akımında kalp kasına azalma meydana gelir. Bu durumda, kalbe besinsel ihtiyaçlar ve oksijen normal olandan daha az gelmekte ve anjina denilen göğüste ağrıya neden olmaktadır. Şayet kan akışı tamamı ile kesilir ya da çok az olur ise kalp krizi meydana gelir ve kalp kasları hasar alır. Çocuklukta başlangıç gösteren damar sertliği kendini orta yaşlarda göstermektedir. Kadınlara göre erkeklerde daha sık görülmektedir. Kadınlarda koruyucu östrojen hormonun azalma gösterdiği zamanlarda yani menopoza girmesinin ardından altmış ile yetmiş yaşlarında erkeklerde ise en fazla elli ile altmış yaşları civarında sıklık göstermektedir. Kalp kişinin yaşamsal fonksiyonu için çok önemli olmasından dolayı tedavisi en ciddiye alınan ve teşhisi erken konulması gereken bir organdır. Günümüzde birçok yöntem ve araştırma yapılmaktadır. Kalp rahatsızlığı bulunan ya da bulunmayan bazı gruplar oluşturulmuş ve diyabet, kolesterol, yaş, cinsiyet ve hasta olup olmadığı gibi veriler makine öğrenmesi yöntemleri gerçekleştiğinde erken teşhis yapılabildiği görülmüştür. Dünya'da son zamanlarda hızlı bir artışla birçok alanda uygulanan makine öğrenme yöntemleri sağlık alanında aktif bir şekilde kullanılmaya başlanmıştır. Birçok yöntemden daha başarılı ve kullanışlı olması ile tercihi fazla olmaktadır.

Sağlık alanında birçok rahatsızlığın olmasına karşın dünyada ve Türkiye'de ölüm nedenleri arasında kalp hastalıkları ilk sıradadır. Bu çalışmada kalp hastalıklarının tespit edilmesinde kullanılabilecek makine öğrenmesi algoritmalarının performans analizi ve analiz yapılırken kullanılan büyük verinin işlenmesi ele alınmıştır.

Kalp rahatsızlıkları üzerine birçok araştırma yapılmış olsa da makine öğrenme yöntemleri kalp hastalıklarının erken teşhis edilmesinde çok az ele alınmıştır. Çalışma kalp rahatsızlıkların önemli olmasından en iyi performansı gösterecek olan yöntemin hangi olduğuna dair bir analiz gerçekleşmiştir. Analizden önce büyük veri, veri işleme teknikleri

kullanılarak optimize edilmiştir. Kalpte meydana gelen tüm rahatsızlıklara sebep olan nedenler, belirtilen çalışma içerisinde yer almaktadır. Bu neden ve belirtileri makine öğrenme yöntemleri ile tespit etmek erken teşhisin daha hızlı yapabilmesine olanak sağlayabilir. Bu durumda makine öğrenme yöntemlerinden bazı ele alınarak incelenmiş ve karşılaştırılmıştır. Kalp rahatsızlıklarının makine öğrenme yöntemiyle yeterli sayıda yapılmaması bu araştırmayı daha önemli hale getirmekte ve yöntemleri de kendi içlerinde avantajlarını ortaya koymaktadır.

Makine öğrenme yöntemleri birçok alanda kullanılmaya başlamıştır. Kullanımı gerçekleşen alanlardan biri de tıp alanı olup, birçok çalışma gerçekleşmiştir. Ancak bu yöntemin kullanılmasıyla yapılan çalışmaların sayılarının artışı yakın tarihlere dayanmaktadır. Başer, Yangın ve Sarıdaş 2021 senesinde "Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması" üzerine bir çalışma gerçekleştirmiştir. Çalışmada ABD'de 1999-2008 yılları arasında 130 hastanedeki 70000 kayıt bulunan sağlık olaylarını veri seti düzenlenerek, diyabet durumuna göre kişileri gruplandırmayı amaçlamışlardır. Sonuçlarda performansı iyi olan 5 adet gruplandırma algoritması; Rastgele orman, Naive Bayes, Lojistik regresyon, k-NN ve Karar ağaçları kullanılmıştır. En iyi performans ise Rastgele orman algoritmasıyla ulaşılmıştır.

Yahyaoui 2017 senesindeki çalışmasında "Göğüs hastalıklarının teşhis edilmesinde makine öğrenmesi algoritmalarının kullanılmasını" ele almıştır. Göğüs rahatsızlıklarını tespit etmek amacıyla Basit Bayes sınıflandırma (NB), k-NN ve destek vektör makinaları (SVM) yöntemleri kullanılmıştır. Ayrıca göğüs hastalıkları teşhisinde destek vektör makina metodunun (ASVM) performansı da değerlendirmeye alınmıştır. Gerçekleştirilen araştırmada ASVM metodu, doğruluk ve öngörülebilir oranı en yüksek sonuçları vermiştir. Polatgil 2020 yılında "Anfis ve Bulanık K Ortalamalar ile Kalp Hastalığının Tespit edilmesi" üzerine bir çalışma yapmıştır. K kat

çaprazlama yöntemi çalışmada kullanılmış ve testi gerçekleşen sistemin başarısı %98.35 şeklinde saptanmıştır. Görgün 2020 yılında "Makine öğrenmesi yöntemleriyle kalp hastalıklarının tahmin edilmesi" başlıklı bir çalışma gerçekleştirmiştir. Çalışmada Bagging Model, Ridge Model, XGBoost Model, LightGBM Model, Rastgele Orman, Karar Ağacı, Naive Bayes, Destek Vektör Makineleri, K-NN ve Lojistik Regresyon kıyaslanmıştır. Çalışmanın neticesinde %90,16 oranında doğruluk değeri ile Rastgele Orman Algoritması başarılı bulunulmuştur. Aydın 2011 yılında "Kalp ritim bozukluğu olan hastaların tedavi süreçlerini desteklemek amaçlı makine öğrenmesine dayalı bir sistemin geliştirilmesi" adlı araştırmasında makine öğrenmesi algoritması şeklinde k-NN algoritmasını kullanmıştır. Bu yönteminin kullanılmasının nedeni düşük bias'ı olan nonlineer bir etmeni olmasından kaynaklanmıştır. Bu nedenle gerçekleştirilen tahminlerin doğruluk oranları yüksek oranda ortaya çıkmıştır. Cihan 2018 senesinde "Koroner arter hastalığı riskinin makine öğrenmesi ile analiz edilmesi" üzerine bir çalışma yapmıştır. Kardiyoloji alanında uzman olan hekimin gözetiminde veri analizi, istatistiksel ile grafiksel yöntemler ile gerçekleşmiştir. Oluşturulan veri kümlerinde doğruluk oranı %80 çıkmıştır. Yıldız ve Hasan 2019 senesinde "Segmantasyon yapmadan patolojik kalp sesi kayıtlarının tespiti için bir örüntü sınıflandırma algoritması" isimli çalışmalarında 6 adet veri bankasından alınmış kalp sesi kayıtlarına segmentasyon yapmadan sınıflandırıcı metotları topluluğu, Destek Vektör Makinesi (DVM) ve k-NN kullanıp gruplandırmaya faydalı bir algoritma gerçekleştirmeyi hedeflemişlerdir.

Temel kavramlar

Kalp hastalıkları

Kalpte ortaya çıkan ve kalbe etki eden bir sorunu içine alan terim kalp hastalığıdır. Kalp hastalığının içerisinde doğuştan gelen kalp kusurları, kan damar rahatsızlıkları ve kalp ritmi sorunları gibi rahatsızlıklar yer

almaktadır. Çoğunlukla kalp hastalığı kavramı kardiyovasküler hastalığıyla dönüşümlü şekilde kullanılmaktadır.

Kalp hastalıklarına etki eden faktörler: Kalp hastalıklarını etki eden faktörler; kilo, diyabet, sigara kullanımı, tansiyon ve kolesteroldür.

Kilo: Kişilerin sağlıkların için vücut kitle indeksi oldukça önem arz etmektedir. Vücutta olan kolestrol ve şekere kilo çok büyük bir faktördür. Dolayısı ile yüksek vücut kitle indeksi olan kişilerin şeker ile kolesterolünün çok olması durumuna da bakılırsa kalp rahatsızlığına yakalanma olasılığı da yüksek olmaktadır (Onat ve diğerleri, 1991).

Diyabet: Kan şekeri kanda yer salan şekerin seviyesini ifade etmektedir. En fazla diyabetin etki gösterdiği alan damardır. Damarların için dolaşmakta olan kan şekeri ile beraber yoğunluk kazanarak plaklaşmaya sebep olur. Buda damar tıkanıklarına sebebiyet vermektedir.

Tansiyon: Kanın vücutta dolaşım göstermesi esnasında atardamarlarda olan basınca tansiyon denilmektedir. Kan basıncında azalma olursa düşün tansiyon, artış olursa da yüksek tansiyon meydana gelmektedir. Stres, alkol kullanma, çok fazla tuz ile şeker tüketimi ve yaşın ilerlemesi vb., etmenler tansiyon tehlikesini arttırır (Güleç, 2009).

Sigara kullanımı: Her toplumu yakın alakadar eden sigara kullanımı yaşamsal bir problemdir. Sigara etki ettiği her organa zarar gösteren bir maddedir.Gerçekleştirilen çalışmalara bakıldığında, bütün kronik akciğer rahatsızlıklarını sigara kullanmak %80'inden, kanser ile kalp hastalığına bağlı yaşam kaybının üçte biri sorumlu olmaktadır (URL-7).

Kolestrol: İnsanların vücutlarında bütün hücrelerde yer alan ve dolaşımı meydana getiren kanın, içindeki yağlanmasına kolesterol denmektedir. Kalpte oluşan sorunlardan en ciddi olanı kabul edilmektedir. Vücudumuzda yüksek kolesterolü taşımakta olan ve damarın duvarlarında yağ birikmesine neden olan kötü kolesterol ile yüksek kolesterolü karaciğere taşımakta olan iyi kolesterol bulunmaktadır (Kayıkçıoğlu, 2014).

Kalp hastalıklarının belirtileri: Kalp rahatsızlığının teşhisinin konulması kolay hale getiren göstergeleri mevcuttur. Bunlar; ciltte ortaya çıkan dökülmeler ve deri döküntüleri, kalıcı ve kuru öksürük, ateş, göz çevresinde, karında ya da bacaklarda şişlik, mavi ya fa soluk gri ten rengi, kalbin hissedilmeyecek derecede yavaş atması ya da tam aksine çok hızlı atması, bayılma, baş dönmesi, sırtta, üst karında, boğazda, çenede ya da boyun bölgesinde ağrı, kalpte yer alan damarların daralması ya da sertleşmesi neticesinde kol ya da bacakta ağrı, halsizlik ya da uyuşukluk, kısa faaliyetlerin ardından ya da ortada hiçbir neden bulunmadan meydana gelen nefes darlığı, göğüs bölgesinde sıkışma, basınç ya da ağrıdır.

Kalp hastalıklarının çeşitleri: Kalpte ortaya çıkan sorunların çevresel ile genetik etmenlerin etkilenerek neden olduğu değişik çeşitleri vardır. Bu çeşitler; konjenital hastalıklar, ritim bozuklukları, anevrizma, kapakçık hastalıkları, koroner hastalıklar şeklindedir.

Materyal ve metod

Makine öğrenmesi bilgisayar bilimi, yapay zekâ ve istatistiğin karşımı olan bir çalışma alanı olup istatistiksel ya da analitik öğrenme gibi yöntemlerle bir veriyi tahmin etmeye veya sınıflamaya yöneliktir.(URL-8). Hayatın her yerinde son senelerde makine öğrenmesi yöntemlerinin kullanımının olduğu alan bulunmaktır. Hangi ürünlerin satın alınacağına, hangi yiyeceklerin sipariş edileceğine ya da hangi filmlerin izleneceğine dair otomatik tavsiyelerden, fotoğraflarınızda arkadaşlarınızı tanımaya veya kişiselleştirilmiş çevrimiçi radyoya dek çoğu cihaz ile k modern web sitesinin temelinde makine öğrenmesi yöntemi bulunmaktadır.

Boyutsal küçültme teknikleri: Boyutsal Küçültme: Modellerin ortaya konulması sırasında boyutu yüksek olan yani çok fazla öznitelik sayısı bulunan veri setlerinin kullanımında yeterli miktarda permütasyonun girdi değişkenlerine dair keşfinin yapılması amacıyla ve veri setinde olan örnek-

lerin sayılarının da çok olması gereklidir. Bu sorunla başa çıkabilmek için genel olarak basitliğin bir miktar doğruluk seviyesin karşılık amaç edinildiği boyutsal küçültme teknikleri ile onların merkezi olan birbirleri ile korelasyon kapsayan özniteliklerin ele alınması hususu gündem olmaktadır. Boyutsal küçültme tekniklerinin uygulanması ile modellerin daha kolay genelleştirilmesi, modeller için gerekli belleğin azaltılması, daha güvenilir hala gelmesi, hesaplama karmaşıklığının azaltılması, performanslarının geliştirilmesi sağlanmaktadır. Çünkü belirtildiği gibi veri setinin boyutunun aza indirgenmesi neticesinde yaratılan yeni öznitelik uzayında alakası en fazla olan öznitelikler yer almaktadır. Ayrıyeten boyutu daha küçük olan veri setlerinin görsellik kazanabilmesi ve makine öğrenmesi yöntemi arayıcılığıyla çok daha hızlı şekilde analizi gerçekleşmektedir.

Özellik çıkarımı: Özellik çıkarımı bir veri setini ele alan bir alt kümenin saptanması ve veriyi en doğru yansıtacak değişkenlerin ayrıştırılması işlemidir. İşlem kullanımı olan algoritmaya en doğru biçimde özellikleri tarayıp en iyi k tanesini n adet özellikten seçmektedir. Bu sayede niteliklerin sayısı indirgenmiş ve sorunun çözülmesinde türlü yararlar sağlamaktadır.

Özellik çıkarımı, veri analizi için kullanımı olan algoritmanın özellik kümesinin boyutu düşer ve algoritma çok daha hızlı çalışmaya başlar. Eksik ya da gürültülü verileri ayırarak verilerin kalitesi artmaktadır. Veri kümesini basit bir duruma getirir ve karmaşık olmasını engeller. Ayrıyeten verilerin boyutunda küçülme olduğundan depolamada kazanç sağlanır.

Makine öğrenmesi algoritmaları: Çok fazla veri yüklendikçe daha başarılı bir performans göstermek amacıyla kendini ayarlayan algoritmalara makine öğrenmesi algoritmaları denmektedir. Literatür içinde makine öğrenmesi algoritması sayısı yüzden fazla şekilde yer almaktadır. Bu çalışma içerisinde birkaç makine öğrenmesi algoritması ele alınmıştır.

Yapay sinir ağları: Yapay sinir ağları, biyolojik olan sinir ağlarından esinlenen matematiksel bir model olmaktadır. Birbirlerine bağlantıları olan yapay nöronlar sınıfından oluşmaktadır ve bilgiyi bağlantısal bir yaklaşımla hesaplar (Pendharkar, 2009). Yapay sinir ağları meydana getiren ağır hale getirilerek birbirine bağlanmış olan sinir hücreleri nöronlardır. Bu işlemde sinyalleri nöronlar alır, aldığı sinyalleri birleştirip, dönüştürerek, sayısal bir neticeye ulaşır. Çoğunlukla YSA'yı öğrenme evresinde yapıyı farklılaştıran uyarlanabilir bir sistem olmaktadır. Çıkış ile girişler arasında olan verilerdeki örüntüleri bulmak veya karmaşık bağları modellemek amacıyla kullanılmaktadır (Joshi, 2019).

Destek vektör makineleri (SVM): Genelleme yanlışının üst sınırını azaltma amacıyla ayırıcı olan örnekler ile hiper düzlemler arasında olabilecek çok fazla mesafeyi yaratarak, üst noktaya çıkarmaya dayalı çok güçlü denetimli olan makine öğrenme tekniklerinden bir tanesi destek vektör makineleridir (Kim vd, 2005). Eğitim veri setinde yer alan birtakım örnekler, sınıflandırma için en faydalı bilgileri sağlar ve ayırıcı hiper düzlemine yakın olurlar. Destek vektörleri bunlara denmektedir. 1963 senesinde Alexey Chervonenkis ile Vladimir Vapnik arayıcılığıyla destek vektör makinelerinin temellerini oluşturmuşlardır. Gözetimli öğrenme algoritması destek vektör makineleridir. Isabelle Guyon ve Alexey Chervonenkis ile Vladimir Vapnik aracılığıyla geliştirilmiştir ve istatiksel öğrenme teorisine dayanmaktadır (Akpınar, 2014). Grupları birbirlerinden ayıran hiper sistemin çıkarılması amaçtır. Bu sayede destek vektörleri arasında olan uzaklık maksimuma ulaşmış olur.

k-NN: Örnek tabanlı öğrenmede en temel algoritmalar içinde K-NN algoritması da yer almaktadır. Eğitim setinde yer alan verilere dayalı şekilde öğrenme işlemi, Örnek tabanlı öğrenme algoritmalarında meydana gelmektedir. Eğitim setinde bulunan örnekler ve yeni karşılaşılan bir örnek arasında olan benzerlik durumuna bakılarak gruplandırılmaktadır

(Mitchell, 1997). Boyutlu sayısal nitelikler ile eğitim setinde bulunan örnekler n K-NN algoritmasında belirtilmektedir. Yer alan her örnek bir noktayı n boyutlu uzayda vekalet edecek şekilde bütün eğitim örnekler boyutu n olan bir örnek uzayında tutulmaktadır. Örnek bilinmiyor ise, ilgili örneğe eğitim setinde en yakın olan k tane örnek tespit edilerek k NN sınıf etiketlerinin çokluk oylama durumuna bakılarak yeni olan örneğin sınıf etiketleri atanmaktadır (URL-2).

Rastgele orman: Kimhon'un geliştirdiği teknik ile 1996 senesinde Brieman'ın geliştirdiği Bagging tekniğinin birleşmesi şeklinde Rastgele Orman Algoritması meydana gelmiştir. Leo Breiman 2001 senesinde geliştirmiştir (Sawant vd., 2018). Gerek regresyon gerek de sınıflandırma problemlerini çözme amacıyla tahmin modellerini ortaya koymayı hedefler. Topluluk yöntemleri birçok öğrenme modellerini çok daha doğru tahmin neticeleri üretmek amacıyla kullanmaktadır. Söz konusu Rastgele Orman modeli olduğunda, mümkün olabilecek en başarılı cevaba model ulaşabilmek amacıyla herhangi bağlantısız Karar Ağaçlarını içine alan bir orman oluşturmaktadır (Xie vd., 2009).

Naive bayes: Thomas Bayes arayıcılığıyla on sekizinci yüzyılda ortaya konulmuştur. Bilinmeyen olayların olasılığını bilinen olaylardan saptamak amacıyla temel olan matematiksel kuralları geliştirilmiştir. Bu sınıfta olan belli niteliklerin başkalarına bakılarak bağımsızlığını varsaymaktadır. Bağımsızlık varsayımı, sınıflandırıcıyı kelime dağarcığının yüksek olduğunu düşünerek kelimelere dayanan e-posta gruplandırması gibi belli işler amacıyla hesaplamayı kolay hale getirmek bakımından etkili olan algoritma şeklinde kılınmaktadır. Pratik olan çalışmalarda Naive Bayes sınıflandırıcı şaşırtıcı olacak şekilde iyi performans ortaya koymaktadır. En başarılı neticenin ihtimalini bulabilmek amacıyla fazla sayıda öznitelikten gelmekte olan bilgilerin eş zamanda dikkat edildiği sorunlara uygulanmaktadır. Mevcut olan bütün bilgileri tahmin etmek için Bayes

algoritması kullanılmaktadır. Algoritmanın ana mantığı, kısmen etkisi küçük olan fazla sayıdaki niteliğin, birleşik olan etkisinin beraber alındığı kuvvetli sınıflandırıcılar yaratacağı ihtimaline dayanmaktadır (URL-3, URL-4).

Veri

Veri temizleme ve ölçekleme: Genel olarak veri temizleme, başka veri hazırlama işlemi öncesinde yapılan bir işlemdir. Verilerde yer alan sistem kaynaklı hataları ya da problemleri düzeltmeyi içermektedir. Veri temizliğinin en yararlısı, yanlış olabilecek belirli gözlemlerin belirlenmesi ile ele alınmasını ve derin alan uzmanlığını içermektedir. Verilerin çoğaltılması, sistematik nedenlerle bozulması, yanlış yazılması gibi sebepler ile yanlış olan değerleri olmasının çok fazla sebebi bulunmaktadır (URL-1). Hatalı, bozuk, gürültülü ve dağınık gözlemlerin tespit edilmesinin ardından bunlar incelenmelidir. Bu, bir sütunun ya da satırın ortadan kaldırılmasını içerebilmektedir. Seçenek şekilde, yeni değerler ile gözlemlerin değiştirilmesini içerebilmektedir. Be sebep ile yapılabilecek genel olan veri temizleme eylemleri bulunmaktadır. Örnek olarak (URL-1) aykırı değerleri belirlemek ve normal olan değerleri için istatistikleri kullanma, varyansı olmayan ya da aynı değeri bulunan sütunların tespiti ile kaldırılması, tekrar edilen veri satırlarını tespit ederek bunları yok etmek, eksik olarak boş olan değerleri işaretleme, öğrenilmiş ya da istatistik bir model kullanımı gerçekleştirerek eksiliği olan değerin tanımın yapılması.

Veri Azaltma: Boyutu yüksek olan verinin boyutu daha küçük olan bir uzayda manalı biçimde ifadesidir. Veri madenciliği, makine öğrenmesi, istatistik ile alakalı alanlarda kullanımı olan çok fazla veri azaltma yöntemleri bulunmaktadır. Lineer olmayan ile lineer olarak bu yöntemler ikiye ayrılmaktadır. Çalışma zamanı ve işlem yükü bakımından lineeer veri azaltma yöntemleri, lineer olmayan yöntemlere bakarak performans seviyeleri daha yüksektir. Bu sebep ile lineer olmayan yöntemlere bakarak

çalışması daha basit ve kolay olması amacıyla boyutu yüksek olan veri analizinde tercih edilmesi daha fazladır.

Bulgular

Çalışmada kullanılan veri seti Amerika Birleşik Devletleri'nin Hastalık Kontrol ve Korunma Merkezleri (CDC) tarafından yürütülen, Davranışsal Risk Faktörü Gözetim Sistemi (BRFSS) anketinden elde edilmiştir. Veri anket üzerinden elde edildiği için öncelikle veri ön işleme sürecinden geçirilmiştir. Burada eksik gözlemler ile "cevap vermek istemiyorum" ve "bilmiyorum" cevaplarına sahip bireyler örneklemden düşürülmüş ardından kategorik değişkenler ikili değişken haline getirilmiştir. Bu bağlamda özellik seçimi sürecinden önce başlangıçta çalışmada kullanılan değişkenlerin tablosu Tablo 1'de verilmiştir.

Tablo 1: *Kullanılan değişkenler*

Değişken	Tanım
_MICHD	Daha önce koroner kalp hastalığı (KKH) veya miyokard enfarktüsü (MI) olduğunu bildirmiş olan katılımcılar için 1 diğerleri için 0 değerini alan ikili değişken
_RFHYPE5	Bir doktor, hemşire veya başka bir sağlık uzmanı tarafından yüksek tansiyonu olduğu söylenen yetişkinler için 1 diğerleri için 0 değerini alan ikili değişken
TOLDHI2	Bir doktor, hemşire veya başka bir sağlık uzmanı tarafından kan kolesterolünüzün yüksek olduğu söylenen kişiler için 1diğerleri için 0 değerini alan ikili değişken
_CHOLCHK	Kişi son 5 yıl içinde kolesterol kontrolü yaptırdıysa 1 aksi halde 0 değeri alan ikili değişken
SMOKE100	Kişi hayatı boyunca en az 100 sigara içtiyse 1 aksi takdirde 0 değerini alan ikili değişken [Not: 5 paket = 100 sigara]
_TOTINDA	Son 30 gün içinde normal işleri dışında fiziksel aktivite veya egzersiz yaptığını bildiren yetişkinler için 1 diğerleri için 0 değerini alan ikili değişken
_FRTLT1	Günde 1 veya daha fazla kez meyve tüketen bireyler için 1 diğerleri için 0 değerini alan ikili değişken
_VEGLT1	Günde 1 veya daha fazla kez sebze tüketen bireyler için 1 diğerleri için 0 değerini alan ikili değişken

_RFDRHV5	Ağır içiciler (haftada 14'ten fazla içki içen yetişkin erkekler ve haftada 7'den fazla içki içen yetişkin kadınlar) için 1 diğerleri için 0 değerini alan ikili değişken
HLTHPLN1	Sağlık sigortası, HMO'lar gibi ön ödemeli planlar veya Medicare veya Indian Health Service gibi hükümet planları dahil olmak üzere herhangi bir sağlık sigortası olanlar için 1 diğerleri için 0 değerini alan ikili değişken
MENTHLTH	Bireyin 30 gün içinde zihinsel sağlığının iyi olmadığı gün sayısı
PHYSHLTH	Bireyin 30 gün içinde fiziksel sağlığının iyi olmadığı gün sayısı
DIFFWALK	Yürümekte zorluk çeken bireyler için 1 diğerleri için sıfır değerini alan ikili değişken
SEX	Kadınlar için 0 erkekler için 1 değerini alan ikili değişken
_AGEG5YR	On dört seviyeli yaş kategorisi
EDUCA	Bireyin en yüksek eğitim seviyesini temsil eden değişken
CHECKUP1	Bireyin son doktor kontrolünden sonra geçen süre
BLOODCHO	Kişi kan kontrolü yaptırdıysa 1 aksi takdirde 0 değerini alan ikili değişken
ASTHMA3	Astımı olan bireyler için 1 diğerleri için sıfır değerini alan ikili değişken
CHCSCNCR	Daha önce deri kanseri geçiren bireyler için 1 aksi takdirde 0 değerini alan ikili değişken
CHCOCNCR	Herhangi bir tür kanser geçiren bireyler için 1 diğerleri için 0 değerini alan ikili değişken
HAVARTH3	Bireyin bir çeşit artrit, romatoid artrit, gut, lupus veya fibromiyaljini varsa 1 aksi takdirde 0 değerini alan ikili değişken
ADDEPEV2	Bireyin depresyon, majör depresyon, distimi veya minör depresyon gibi depresif bir bozukluğu varsa 1 aksi halde 0 değerini alan ikili değişken
CHCKIDNY	Kişinin böbrek hastalığı varsa 1 aksi takdirde 0 değerini alan ikili değişken
DIABETE3	Birey diyabet hastalığına sahipse 1 aksi takdirde 0 değerini alan ikili değişken

Veri ön işleme süreci tamamlandıktan sonra özellik seçimi sürecine geçilmiştir. Özellik seçimi, makine öğrenimi modelleri kullanılan çalışmalarda açıklayıcı değişken sayısını bir diğer deyişle girdi sayısını azaltmak için kullanılan bir yöntemdir. Bu yöntemin kullanılma sebebi girdi sayısının fazla olması sebebiyle öğrenme sürecinin yavaş olması ya da yanıt (bağımlı değişken) değişkeniyle korelasyonsuz değişkenlerin modelin gücünü düşürmesi olabilir. Regresyona dayalı makine öğrenmesi algoritmalarında, bağımlı değişken ile ilgisiz değişkenlerin modele dahil edilmesi sonucunda ekonometrik analizde sıkça görülen "gereksiz değişkenin modele dahil edilmesi" problemi ortaya çıkabilir.

Özellik seçimi tümdengelim ya da tümevarım yoluyla yapılabilir. Sıralı ileri seçim tümevarıma dayanır ki bu yöntemin başlangıcında herhangi bir girdi değişkenine sahip olmayan bir model kurulur ve sonrasında tüm girdi değişkenleri sırasıyla modele dahil edilir. Her bir adımda model seçim kriterlerine bakılarak (AIC:, düzeltilmiş R kare:, eklenen değişkenin katsayısının t istatistiğinin olasılık değeri: , hata kareler toplamı:) model performansına katkısı olan girdi değişkenleri modelde kalır diğerleri model dışı bırakılır sonuç olarak nihai model elde edilir. Sıralı geri seçim ise tümdengelime dayanır, başlangıçta tüm girdi değişkenlerinin bulunduğu bir model tahmin edilir ve sırasıyla her bir adımda istatistiksel olarak anlamsız olan değişkenler model dışı bırakılarak nihai model elde edilir.

Bu çalışmada bağımlı (yanıt değişkeni) değişken ikili bir değişken olduğu için özellik seçimi klasik doğrusal regresyon ile değil lojistik regresyon ile yapılmaktadır. Bağımlı değişkenin ikili olduğu durumlarda, doğrusal regresyon kullanılırsa hata teriminde ortaya çıkan değişen varyans sebebiyle katsayıların standart hataları yanlı tahmin edilir ve dolayısıyla t istatistikleri ve p-değerleri olduğundan büyük ya da küçük bulunur. Sıralı geriye seçim yapılırken aslında katsayısı istatistiksel olarak anlamlı olan bir değişkeni -yanlış tahmin edilen p-değerinden dolayı- modelden çıkartmak

ya da katsayısı istatistiksel olarak anlamlı olmayan bir değişkeni -yanlış tahmin edilen p-değerinden dolayı- modelden tutmaktan kaçınmak için özellik seçiminde lojistik regresyon kullanılmıştır.

Geriye doğru eleme yöntemi kullanılarak yapılan özellik seçimi sonucunda %5 önem düzeyinde istatistiksel olarak anlamsız olan FRTL1, PYSHLTH, BMI5, EDUCA, HLTHPLN1, CHCOCNCR, CHECKUP1 ve CHCSCNCR değişkenleri modelden dışlanarak nihai model elde edilmiştir.

Kalp hastalıklarının tespitinde hangi makine öğrenim algoritmasının başarılı olduğunu tespit etmek amacıyla yapılan bu çalışmada nihai model sınıflandırma algoritmalarından, sağlık konusunda çalışılırken sıklıkla kullanılan k-NN, Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Rassal Ormanlar, Naive Bayes ve Yapay Sinir Ağları aracılığıyla analiz edilmiştir. Bu algoritmaların kalp hastalıklarını tespit etmedeki gücü başarı oranı, kesinlik skoru, duyarlılık skoru ve F1 skoru açısından kıyaslanmıştır. Başarı oranı, toplam doğru sınıflandırmaların, toplam gözlemlere oranı olarak tanımlanır. Başarı oranı yükseldikçe modelin uyum iyiliğinin de arttığı söylenebilir. Kesinlik skoru, doğru pozitiflerin toplam pozitifler içindeki payı olarak tanımlanırken, Hassasiyet skoru, doğru pozitiflerin doğru sınıflandırmalar içindeki payıdır. F1 skoru ise, kesinlik ve hassasiyetin çarpımının 2 katının, kesinlik ve hassasiyet içindeki payı olarak tanımlanmaktadır. Tüm bu sınıflandırma kriterlerinin her bir algoritma için sonuçları Tablo 2'de verilmiştir.

Tablo 2: *Sınıflandırma kriterlerinin algoritma sonuçları*

	Başarı oranı	Kesinlik	Duyarlılık	F1 Skoru
k-NN	0.8625	0.59	0.58	0.58
Lojistik Regresyon	0.9077	0.74	0.56	0.58
Destek Vektör Makineleri	0.9052	0.45	0.50	0.48
Karar Ağaçları	0.8725	0.60	0.58	0.59
Rassal Ormanlar	0.8922	0.63	0.56	0.57
Naive Bayes	0.8722	0.61	0.69	0.63
Yapay Sinir Ağları	0.9054	0.76	0.53	0.54

Modellerin başarılı tahmin oranlarına bakıldığında en yüksek başarıya sahip modellerin Lojistik Regresyon, Destek Vektör Makineleri ve Yapay Sinir Ağları olduğu görülmektedir. Modellerin kesinlik skorlarına bakıldığında ise en yüksek başarıyla tahmin yapan algoritmaların Lojistik Regresyon ve Yapay Sinir Ağları olduğu görülmektedir. Duyarlılık skoru açısından bakıldığında ise en iyi model Naive Bayes'tir. F1 skorlarına bakıldığında ise en başarılı modeller Karar Ağaçları ve Naive Bayes'tir.

Tartışma ve sonuç

Dünya yer alan birçok ülkede kalp rahatsızlıkları gerek kadınlar gerekse de erkekler arasında çok yaygın görülmektedir. Bu sebep ile bireyler kalpte meydana gelecek risk etmenlerini dikkate almalıdır. Bazı etmenlerini yaşam tarzı etmenleri, bazı etmenler genetik bir rol oynamaktadır ve bu durumlar kalpte ciddi derecede etki etmektedir. Bütün bunlara bakıldığında çalışmada makine öğrenme yöntemleri ele alınmış ve incelenmiştir. Dünyada kullanımı git gide artış gösteren bu yöntemi birçok alt başlığı yer aldığı gibi kullanılacak alana göre başlıkların avantajları değişkenlik göstermektedir. Araştırma içerisinde destek vektör makineleri, rastgele orman algoritması, yapay sinir ağları, k-NN ve Naive Bayes yöntemleri karşılaştırılmıştır.

Çalışma açısından model başarı kriterlerine bakıldığında, uç gözlemlere sahip veri kümelerinde başarı oranı, kesinlik ve duyarlılığına bakılması doğru değildir ancak bu çalışmada veri ön işleme sürecinde değişkenler ikili değişken haline getirildiği için veri setinde uç gözlemler bulunmamaktadır. Bu sebeple kalp hastalıklarının makine öğrenmesi yöntemleri ile belirlenmesinde başarı oranı, kesinlik ve duyarlılığa bakılarak en başarılı modellerin Lojistik Regresyon ve Yapay Sinir Ağları modelleri olduğuna karar verilmiştir. Makine öğrenmesi algoritmalarının daha büyük veri

setleriyle daha iyi çalıştığı göz önünde bulundurulduğunda daha geniş bir veri setiyle çalışılarak, çalışmanın daha ileri noktalara taşınabileceği düşünülmektedir. Ayrıca, eğer kalp rahatsızlığına sahip kişilerin kalplerine ait görüntüler elde edilebilirse görüntü işleme algoritmaları kullanılarak daha başarılı sonuçlar elde edilebileceği düşünülmektedir. Bu çalışmada kullanılan bağımlı değişkenin ikili değişken olmasından dolay sınıflandırma algoritmaları kullanılmıştır. Bu sebeple kalp rahatsızlığını temsil edebilecek sürekli bir bağımlı değişken elde edilebildiği takdirde çalışmanın yalnızca sınıflandırma algoritmalarına sahip olma kısıtının ortadan kaldırılabileceği düşünülmektedir.

Kaynakça

[1] Akpınar, H., (2014), Data Veri Madenciliği Veri Analizi, Yayınevi: Papatya Bilim, Basım: İstanbul.

[2] Aydın, F. (2011). *Kalp ritim bozukluğu olan hastaların tedavi süreçlerini desteklemek amaçlı makine öğrenmesine dayalı bir sistemin geliştirilmesi,* Trakya Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi.

[3] Başer, B. Ö., Yangın, M., Sarıdaş, E. S. (2021). Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 25(1), 112-120.

[4] Cihan, Ş. (2018). *Koroner arter hastalığı riskinin makine öğrenmesi ile analiz edilmesi* (Master's thesis, Kırıkkale Üniversitesi).

[5] Görgün, M. (2020). *Makine öğrenmesi yöntemleriyle kalp hastalıklarının tahmin edilmesi* (Master's thesis, Lisansüstü Eğitim Enstitüsü).

[6] Sawant, R., Jangid, Y., Tiwari, T., Jain, S., Gupta, A. (2018, August). Comprehensive analysis of housing price prediction in pune using multi-featured random forest approach. In 2018 *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-5). IEEE.

[7] Güleç, S. (2009). Kalp damar hastalıklarında global risk ve hedefler. *Türk Kardiyol Dern. Arş,* 37, 3-5.

[8] Joshi, P. (2019). Predicting customers churn in telecom industry using centroid oversampling method and KNN classifier. *Int. Res. J. Eng. Technol.,Vol.* 6, No 4,3708-3712

[9]Kayikçioğlu, M. (2014). Homozygous familial hypercholesterolemia. *Turk Kardiyoloji Dernegi Arsivi: Turk Kardiyoloji Derneginin Yayin Organidir*, 42, 47-55.

[10] Kim, S., Shin, K. S., Park, K. (2005). An application of support vector machines for customer churn analysis: Credit card case. In *International Conference on Natural Computation* (pp. 636-647). Springer, Berlin, Heidelberg.

[11] Mitchell, T. M., Mitchell, T. M. (1997). *Machine learning* (Vol. 1, No. 9). New York: McGraw-hill.

[12] Onat, A., Örnek, E., Şenocak, M., vd., (1991). Survey on Prevalence of Cardiac Disease and its Risk Factors Adults in Turkey: 6. Diabetes and Obesity. *Archives of the Turkish Society of Cardiology*, 19(3), 178-185.

[13] Pendharkar, P. C. (2009). Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services. Expert Systems with *Applications,* 36(3), 6714-6720.

[14] Polatgil, M. (2020). Anfis ve Bulanık K Ortalamalar ile Kalp Hastalığının Tespit edilmesi. *Bilişim Teknolojileri Dergisi*, 13(4), 443-449.

[15] Xie, Y., Li, X., Ngai, E. W. T., Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3), 5445-5449.

[16] Yahyaoui, A. (2017). Göğüs hastalıklarının teşhis edilmesinde makine öğrenmesi algoritmalarının kullanılması.Yang X. Introduction to algorithms for data mining and machine learning. Published online.

[17] Yıldız, A., Hasan, Z. A. N. (2019). Segmantasyon yapmadan patolojik kalp sesi kayıtlarının tespiti için bir örüntü sınıflandırma algoritması. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 10(1), 77-91.

İnternet kaynakları

URL-1: Brownlee J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Published online. https://books.google.com/books?hl=tr&lr=&id=uAPuDwAAQBAJ&oi=fnd&pg=PP1&dq=Brownlee,+J.+(2020).+Data+preparation+for+machine+learning:+data+cleaning,+feature+selection,+and+data+transforms+in+Python.+Machine+Learning+Mastery.&ots=Cl3JxkcNtU&sig=l2XsLJ- BZrrQTHmw5lNJBoI Erişim Tarihi: 08.06.2022

URL-2: Han, J., Kamber, M., (2006), "Data mining: concepts and techniques",Morgan Kaufmann Publishers, Burlington. Introduction to Machine Learning with Python: A Guide for Data Scientists - Andreas C. Müller, Sarah Guido -Google Kitaplar.https://books.google.com.tr/books?hl=tr&lr=&id=1-

4lDQAAQBAJ&oi=fnd&pg=PP1&dq=Introduction+to+machine+learning+with+Python:+a+guide+for+data+scientists.&ots=28jRIQMFW_&sig=fRhfMUMM8RcBo_HsaypjE3oRAeU&redir_esc=y#v=onepage&q=Introduction to machine learning with Python%3A a guide for data scientists.&f=false
Erişim Tarihi: 09.06.2022

URL-3: Jayant, A., (2020), Data Science and Machine Learning Series: Naive Bayes Classifier Advanced Concepts, Technics Publications.
https://books.google.com.tr/books?id=Q0A_zQEACAAJ&dq=naive+bayes&hl=tr&sa=X&redir_esc=y
Erişim Tarihi: 09.06.2022

URL-4: Perez, C., (2019), Statıstıcs And Data Analysıs With Matlab. Naıve Bayes,Knn And Pattern Recognıtıon, Amazon Digital Services LLC- KDP Print US.https://books.google.com.tr/books?id=BV1_xQEACAAJ&dq=naive+bayes&hl=tr&sa=X&redir_esc=y
Erişim Tarihi: 08.06.2022

URL-7: US Department of Health and Human Services (1982). A Report of the Surgeon General: The health consequences of smoking. Washington (DC), US Department of Health and Human Services.

https://www.hhs.gov/surgeongeneral/reports-and-publications/tobacco/consequences-smoking-factsheet/index.html#:~:text=Since%20the%20first%20Surgeon%20General's%20Report%20in%201964%2C%20evidence%20has,obstructive%20pulmonary%20disease%20(COPD).
Erişim Tarihi: 09.06.2022

URL-8: Yılmaz, D. Ö. Ü. A., Yayın, K. (2021). *Yapay Zeka*, Kodlab Yayın Dağıtım Yazılım Ltd. Şti
https://www.google.com/search?tbm=bks&q=makine+öğrenme+yöntemleri Erişim Tarihi: 09.06.202.