T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİYOİSTATİSTİK VE TIP BİLİŞİMİ ANA BİLİM DALI

KRONİK BÖBREK YETMEZLİĞİ (KBY) TANISINDA FARKLI MAKİNE ÖĞRENİMİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Yüksek Lisans Tezi

Tolga DEMİREL

Danışman

Doç. Dr. Leman TOMAK

Bu çalışma Ondokuz Mayıs Üniversitesi tarafından PYO.TIP.1904.22.002 proje numarası ile desteklenmiştir.

SAMSUN
2023

ÖZET

KRONİK BÖBREK YETMEZLİĞİ (KBY) TANISINDA FARKLI MAKİNE ÖĞRENİMİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Tolga DEMİREL
Ondokuz Mayıs Üniversitesi
Lisansüstü Eğitim Enstitüsü
Biyoistatistik ve Tıp Bilişimi Ana Bilim Dalı
Yüksek Lisans, Ocak/2023
Danışman: Doç. Dr. Leman TOMAK

Kronik Böbrek Yetmezliği (KBY), hastaların fiziksel durumunu ve sosyal yaşamını olumsuz etkileyen önemli bir sağlık sorunudur ve makine öğrenimi alanındaki çalışmalar, nefroloji alanında karar vericilere destek olacak önemli sonuçlar vermektedir. Bu çalışmada, KBY hastalığının tanısı için makine öğrenimi yöntemleri ile farklı tahmin modelleri oluşturulması ve bu tahmin modellerinin performanslarının karşılaştırması ve değerlendirilmesi amaçlanmaktadır.

Ondokuz Mayıs Üniversitesi Tıp Fakültesi Nefroloji Anabilim Dalına 2015-2021 yılları arasında başvuran hasta bilgilerinin kullanıldığı veri setinde 9.136 hastaya ilişkin 22 farklı değişken kullanılmıştır. Bu çalışmada, tek sınıf, Naive Bayes, lojistik regresyon, karar ağaçları, rastgele orman, en yakın k-komşu ve destek vektör makinesi algoritmaları için veri seti, WEKA yazılımı tarafından analiz edilen %50 ile %90 arasında değişen eğitim veri setlerine bölünmüştür ve tahmin modelleri oluşturulmuştur. Bu modellerin başarı oranları doğruluk oranı, hata oranı, kesinlik, duyarlılık, seçicilik, F ölçümü ve Matthews korelasyon katsayısı ölçütleri kullanılarak karşılaştırılmıştır.

Veri setinde 9.136 hastanın %54,28'nde KBY hastalığı bulunuyorken %45,72'nde bu hastalık bulunmamaktadır. Veri setinin tümü ile oluşturulan tahmin modellerinde en iyi performansı; doğruluk oranı için %97,1 ile rastgele orman ve en yakın k-komşu algoritmaları ile elde edilmiştir. Veri setinin, eğitim veri seti ve test veri seti olarak %50-90 arasında ayrıldığı tahmin modellerinin tümünde ise en iyi performansı; doğruluk oranı için lojistik regresyon algoritması göstermiştir. Diğer performans ölçütleri için rastgele orman ve destek vektör makineleri algoritmalarının incelenen diğer algoritmalardan daha başarılı olduğu gözlemlenmiştir. Ayrıca işlem karateristik eğrisi analizi sonucunda hasta yaşı değişkeni için kesim noktası olarak 54,5 değeri belirlenmiştir.

KBY tanısında kullanılan farklı makine öğrenimi yöntemlerinin, veri setinin farklı oranlarda eğitim ve test veri setlerine ayrılmasıyla ortaya çıkan sonuçları değerlendirilmiştir. Bu değerlendirmeler sonucunda %60-80 aralığında eğitim verisi, %20-40 aralığında ise test veri setinin kullanılmasının uygun olduğu ve farklı büyüklükteki veri setlerine uygulanan algoritmalar için farklı performans ölçütlerindeki başarı oranlarının değişiklik gösterdiği görülmüştür.

Anahtar Sözcükler: Makine öğrenmesi, Kronik böbrek yetmezliği, Tahmin modeli, Algoritma.

ABSTRACT

COMPARISON OF DIFFERENT MACHINE LEARNING METHODS IN THE DIAGNOSIS OF CHRONIC KIDNEY FAILURE (CKF)

Tolga DEMİREL
Ondokuz Mayıs University
Institute of Graduate Studies
Department of Biostatistics and Medical Informatics
Master, January/2023
Supervisor: Assoc. Prof. Dr. Leman TOMAK

Chronic Kidney Failure (CKF) is an important health problem that negatively affects the physical condition and social life of patients, and studies in the field of machine learning provide decision makers with important results in the field of nephrology. The objective of this study was to create different prediction models with machine learning methods for the diagnosis of CKF and to compare and evaluate the performances of these prediction models.

The material of this study was 22 different variables related to 9,136 patients which patient information was used, who applied to Ondokuz Mayıs University Faculty of Medicine, Department of Nephrology between the years 2015-2021. In this study, the dataset for OneR, Naive Bayes, logistic regression, decision trees, random forest, nearest k-neighbor and support vector machine algorithms is divided into training datasets ranging from 50% to 90% analyzed by WEKA software and forecasting models were created. The success rates of these models were compared using the criteria of accuracy, error rate, precision, sensitivity, specificity, F measurement and Matthews correlation coefficient.

The results of the study revealed that 54.28% of patients had CKF, while 45.72% of patients did not have this disease. The best performance in prediction models created with the entire data set was obtained by random forest and k-nearest k-neighbor algorithms (97.1%) with respect to accuracy rate. Besides, the best performance in all prediction models, where the data set is divided between 50-90% as the training data set and the test data set was obtained by the logistic regression algorithm with respect to the accuracy rate. For other performance measures, random forest and support vector machine algorithms were observed to be more successful than the other algorithms. The results of the receiver operating characteristic analysis introduced that, a cut-off point of 54.5 was determined for the patient age variable.

The results of different machine learning methods used in the diagnosis of CKF were evaluated by separating the data set into training and test data sets at different rates. The results highlighted that it was appropriate to use the training data in the range of 60-80% and the test dataset in the range of 20-40%, the success rates in different performance criteria varied for the algorithms applied to the datasets of different sizes.

Keywords: Machine learning, Chronic kidney failure, Prediction model, Algorithm.

ÖN SÖZ VE TEŞEKKÜR

Ondokuz Mayıs Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından desteklenen “Kronik Böbrek Yetmezliği Tanısında Farklı Makine Öğrenimi Yöntemlerinin Karşılaştırılması” çalışmasında günümüzde bir çok alanda yer alan farklı makine öğrenimi yöntemlerinin kronik böbrek yetmezliği tanısında kullanılması, bu yöntemlerin farklı ölçütler kullanılarak karşılaştırılması ve karşılaştırılan makine öğrenimi modelleri arasında verimlilik sıralaması yapılması amaçlanmaktadır.

Lisans mezuniyetimden uzun bir ara sonra başladığım yüksek lisans eğitimim boyunca bana her türlü imkânı sunan, beni her daim destekleyen, bilgi birikimini ve tecrübelerini benimle her zaman paylaşan çok kıymetli danışman hocam Doç. Dr. Leman TOMAK’a en derin saygı ve en içten teşekkürlerimi sunarım.

Ayrıca hayatıma girdiği günden beri her zaman bana destek olan sevgili eşim Özlem SOYLU DEMİREL ve biricik çocuklarım Emir ve Esma’ya bana gösterdikleri anlayış ve destek için çok teşekkür ederim.

Tolga DEMİREL

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

AO	: Aritmetik Ortalama
BKİ	: Beden Kitle İndeksi
BT	: Bilgisayarlı Tomografi
BUN	: Üre değeri
DO	: Doğruluk Oranı
DVM	: Destek Vektör Makineleri
DVMA	: Destek Vektör Makineleri Algoritması
EYKKA	: En Yakın K-Komşu Algoritması
HO	: Hata Oranı
KAA	: Karar Ağaçları Algoritması
KBY	: Kronik Böbrek Yetmezliği
LRA	: Lojistik Regresyon Algoritması
MKK	: Matthews Korelasyon Katsayısı
MR	: Manyetik Rezonans
NBA	: Naive Bayes Algoritması
OR	: Odds Oranı
PD	: Periton Diyalizi
RBC	: Kırmızı Kan Hücresi Sayımı
ROA	: Rastgele Orman Algoritması
ROC	: Receiver Operating Characteristic
TSA	: Tek Sınıf Algoritması
WBC	: Beyaz Kan Hücresi Sayımı
WEKA	: Waikato Environment for Knowledge Analysis

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

1. GİRİŞ

Kronik Böbrek Yetmezliği (KBY) hastalığı, dünya genelinde ve ülkemizde halk sağlığı sorunları açısından önemli bir yer tutmaktadır (Uysal ve ark., 2020). KBY hastalığının prevalansı dünya genelinde %11-13 arasındadır (Hill ve ark., 2016). KBY yaşamı olumsuz etkileyen, işgücü kaybına neden olan, her yaş grubundan kişileri etkileyebilen, ihmal edilmesi durumunda hayati risk oluşturan çok önemli bir hastalıktır (Şentürk ve Levant, 2000; Khan ve ark., 2020). Böbreklerin temel görevleri arasında üre, kreatinin ve ürik asit gibi metobolik atık maddelerin atılmasını sağlamak, vücut sıvı hacimlerinin dengelenmesini sağlamak, düzenleyici ve dışa atım fonksiyonlarının yürütülmesi yer almaktadır (Yalçın ve ark., 2019). KBY hastalığı bulunan kişilerde böbreğin bu fonksiyonları tam olarak yerine getirememesi söz konusudur. Buna bağlı olarak bu hastalığa sahip kişilerin, böbreğin yürütemediği ya da yeterli derecede gerçekleştiremediği fonksiyonların yürütülebilmesi için diyaliz ya da böbrek nakli gibi tedavi yöntemleri ön plana çıkmaktadır (Ceyhun ve Kırpınar, 2019). Türk Nefroloji Derneği'nin 2020 yılı Türk Böbrek Kayıt Sistemi Raporunda pediatrik vakalar dahil Türkiye'de 60.558 hemodiyaliz, 3.387 periton diyalizi (PD) ve yaklaşık 19.405 böbrek nakli olmak üzere toplamda 83.350 böbrek yetmezliği vakası olduğu belirtilmiştir (Süleymanlar ve ark., 2020; Demir ve Özer, 2022). Türkiye nüfusunu göz önüne alarak bir değerlendirme yaptığımızda bu hastalığa sahip kişilerin ülke nüfusunun yaklaşık binde birine karşılık geldiği açık bir şekilde görülmektedir (TÜİK, 2021). Bu konu sadece ülkemizde değil tüm dünyada insan sağlığı açısından öneme sahip hayati bir konudur.

Bu hastalığa sahip kişilerin sayısının fazla olması da tedavi için sağlık harcamaları açısından gerek hastalara gerek kamuya oldukça önemli finansal yükler oluşturabilmektedir (Ağır ve Tıraş, 2018). Bu nedenle bu hastalığın erken tanısı hem insan sağlığı hem maliyet açısından önem arz etmektedir (Mans ve ark., 2019). Farklı hastalıkların erken tanısı amacıyla farklı yöntemler geliştirilmektedir. Geliştirilen yöntemlerde teknolojik ilerlemelerden faydalanılmakta ve buna bağlı olarak insan ömrünün daha uzun ve konforlu olması amaçlanmaktadır (Hofmann ve ark., 2018). Belirtilen teknolojik yöntemler arasında makine öğrenmesi önemli bir rol oynamaktadır. Makine öğrenimi yöntemleri hastalıkların tanısı için günümüzde kullanılan teknolojik yöntemler arasında öne çıkmaktadır (Battineni ve ark., 2020).

Bu yöntem ile oluşturulan tahmin modelleri sayesinde hastaların farklı vücut değerleri ile ilgili hastalığa olan yatkınlıkları tahmin edilmek istenmektedir (İbrahim ve Abdulazez, 2021). Makine öğrenimi yöntemleri KBY hastalığının erken tanısında son dönemde kullanımı artan ve geliştirilen bir yöntem olarak karşımıza çıkmaktadır (Qin ve ark., 2019). Makine öğrenimi, makinelere herhangi bir programlama olmaksızın öğrenme yeteneği ve becerisi kazandırmak için kullanılan bilgisayar bilimi kavramlarını olarak tanımlanabilir (Pulat ve Kocakoç, 2021). Makine öğrenimi yöntemleri ile geçmişte elde edilen verileri kullanarak yeni veri için en uygun modelin bulunması hedeflenmektedir. Makine öğreniminde en iyi tahmini üretecek olan modeli oluştururken farklı algoritmalardan yararlanılır (Canedo ve Mendes, 2020). En çok bilinen ve kullanılan makine öğrenimi yöntemleri arasında "Tek sınıf, Naive Bayes, destek vektör makineleri, en yakın k komşu, karar ağaçları, rastgele ağaç, lojistik regresyon" algoritmaları yer almaktadır (Khan ve ark., 2020; Eroğlu ve Palabaş, 2019).

Bu çalışmada, KBY hastalığının tanısı için makine öğrenimi yöntemleri ile farklı tahmin modelleri oluşturmak ve oluşturulan makine öğrenimi tahmin modellerini karşılaştırarak ortaya çıkan sonuçların başarı düzeyini ölçmek amaçlanmaktadır.

2. GENEL BİLGİLER

2.1. Kronik Böbrek Yetmezliği

2.1.1. Kronik Böbrek Yetmezliği Nedir?

KBY böbreklerin görevlerini, geriye dönüşü olmaksızın yerine getirememesi olarak tanımlanmaktadır (Bilgen, 2021). Teknik olarak glomerüler filtrasyon değerinde azalmanın sonucu böbreğin sıvı-solüt dengesini ayarlama ve metobolik-endokrin fonksiyonlarında kronik ve ilerleyici bozulma halidir (Tanrıverdi ve ark., 2010). KBY tanısı glomerüler filtrasyon hızının (GFH), 15 ml/dakikanın altına düşmesiyle koyulur. Üremi KBY hastalığının neden olduğu tüm klinik ve biyokimyasal anormallikleri içeren bir deyimdir ve birçok kaynakta KBY ile eş anlamda kullanılmaktadır. KBY hastalarının büyük bir bölümünde böbrek boyutları küçülmüştür (Elmacı ve ark., 2019). Hastaların klinik belirti ve bulguları altta yatan patoloji, böbrek yetersizliğinin derecesi ve gelişme hızı ile yakından ilişkilidir (Jankowski ve ark., 2021). Böbreklerin ikisinde de nefron adı verilen ve her birinde sayısı milyonu bulan, böbreğin en küçük anatomik ve fonksiyonel birimi filtreler bulunmaktadır (Eyüpoğlu, 2020). Nefronlar zarar gördüklerinde çalışmayı bırakırlar. Böylelikle filtrasyon görevini tam anlamıyla yerine getiremediklerinden dolayı üre, kreatinin ve ürik asit gibi metobolik atık maddelerinin atılması görevini yerine getiremezler (Amaresan ve Geetha, 2008). Böbreklerde oluşan bu hasar kalıcı olarak devam ederse hastalık kronik hale gelmektedir. Bu şekilde KBY hastalığı oluşmaktadır (Schanstra ve ark., 2015).

KBY hastalığına neden olan önemli sağlık sorunları bulunmaktadır. Bunlar arasında arteriyel hipertansiyona bağlı diyabetik nefropati ve nefroanjiyoskleroz en yaygın olanlarıdır. Bunun dışında obezite gibi risk faktörleri de bu hastalığa olan yatkınlığı arttırabilmektedir (Karaman, 2022). Son yapılan araştırmalara göre KBY hastalarının %78,8'i hemodiyalize, %7,2'si periton diyalizine girmekte, %13,9'una ise renal transplantasyon yapılmaktadır (Bahtiyarca ve Çakcı, 2021).

2.1.2. Kronik Böbrek Yetmezliği Tanı Yöntemleri

KBY hastalığının tanısı için biyokimyasal değerlendirme, idrar incelemesi, görüntüleme yöntemleri ve böbrek biyopsisi kullanılan yöntemlerdir (Wu J ve ark., 2018). Biyokimyasal değerlendirmeler ve idrar incelemesi içerisinde hastanın dansite, albümin, kırmızı kan hücresi, bun, kreatinin, kalsiyum, fosfor, sodyum,

potasyum, kan şekeri değerleri, hemoglobin, hematokrit, beyaz kan hücresi sayımı, kırmızı kan hücresi sayımı, enfeksiyon varlığı gibi değerlerinin incelenmesi önemlidir (Sharma ve ark., 2016). Ayrıca hastanın hipertansiyon, diyabet, kalp yetmezliği gibi kronik hastalıklara sahip olup olmadığının incelenmesi de bu hastalığın tanısında öne çıkan faktörlerdendir (Yalçın ve ark., 2019).

Hastalığın değerlendirilmesinde noninvazif veya invazif görüntüleme yöntemleri kullanılabilir (Elsivers ve ark., 1995). Böbreklerin direkt grafisi, intravenöz piyelografi, ultrason, bilgisayarlı tomografi (BT), manyetik rezonans (MR), konvansiyonel anjiyografi ve nükleer görüntüleme (sintigrafi) kullanılan temel görüntüleme yöntemleri arasında yer almaktadır (Perazella, 2015).

Bu geleneksel yöntemlere ek olarak günümüz teknolojisindeki ilerlemelerin katkısıyla yapay zeka yazılımları da çeşitli hastalıkların tanısında karar vericilere yol gösterici olarak kullanılmaktadır (Demirhan ve ark., 2010). KBY hastalığının tanısında da gerek biyokimyasal değerler ve idrar incelemesinde gerek görüntüleme yöntemlerinde makine öğrenimi, derin öğrenme, yapay sinir ağları gibi yapay zeka modellemeleri kullanılmaya başlanmıştır (Alnazer ve ark., 2021). Biyomedikal araştırmaların veri açısından zenginleşmesiyle makine öğrenimi, büyük ölçekli biyolojik veri kümelerinden bilgi çıkarmak için güçlü bir araç seti haline gelmiştir (Bilgin, 2017). Kapsamlı böbrek omik bileşenleri (transkriptomik, proteomik, metabolomik ve genom dizileme), elektronik sağlık kayıtları, dijital nefropatoloji depoları ve radyoloji böbrek görüntüleri gibi diğer veri yöntemlerinin artan kullanılabilirliği, makine öğrenimi yaklaşımlarını insan analizi için giderek daha gerekli hale getirmektedir (Sealfon ve ark., 2020).

Kronik Böbrek Yetmezliği Tanısında Makine Öğrenimi Yöntemlerinin Kullanılması

Makine öğrenimi aklımıza gelebilecek hemen her alanda kullanıldığı gibi sağlık sektöründe ve buna bağlı olarak böbrek hastalıklarının tanısında kullanılmaktadır (Allugunti, 2022). Yapay zeka teknolojisinin gelişimiyle makine öğrenimi algoritmaları, gerek böbrek hastalığının tanısında başvurulan kan ve idrar tahlil sonuçları gerek ön muayene raporları gerekse görüntüleme sonuçlarını değerlendirerek karar vericilere atacakları adımlar için öncülük etmektedirler. Bu yöntemler ile böbrek hastalığının erken tanısını daha mümkün hale gelirken, tahmin modellerinin öğrenme yeteneklerini geliştirmesiyle riski azaltacak adımların

atılabilmesi de mümkün hale gelmektedir (Li Q ve ark., 2022).

Ancak bu alanda yapılan çalışmalarda sadece test sonuçları değil aynı zamanda görüntüleme sonuçlarının da değerlendirilmesi söz konusu olduğundan büyük veri ortaya çıkmaktadır. Bu büyük veriyi de yönetebilmek ve yorumlayabilmek için yapay zekâ temelli yazılımlara ihtiyaç duyulmaktadır. Bu yazılımların geliştirilmesiyle tahmin olasılıkları arttırılmakta ve böbrek hastalarının gerek tedavisinde gerek yaşam kalitelerinin arttırılmasında önemli roller üstlenmektedir (Lemley, 2019). Naive Bayes, lojistik regresyon, karar ağaçları, rastgele orman, en yakın k komşu ve destek vektör makineleri makine öğrenimi yöntemleri KBY tanısında kullanılmaktadır.

2.2. Makine Öğrenimi

2.2.1. Makine Öğrenimi Tanımı

Makine öğrenimi açıkça programlamaya gerek duymaksızın makinelere öğrenme yetisi kazandırmak için kullanılan bilgisayar bilimi tekniklerini ifade eder (Khan ve ark., 2020). Makine öğrenmesi, bir problemi o probleme ait veriye göre modelleyen bilgisayar algoritmalarının genel adıdır. Mevcut veri seti ve kullanılan algoritma ile oluşturulan model, en yüksek performansı vermek üzere kurulmaktadır (Roscher ve ark., 2020). Makine öğrenimi günümüzde sıklıkla kullanılan tahmin modelleri arasında yer almaktadır. Makine öğrenimi tarihi süreçleri incelendiğinde günümüze kadar birçok tahmin modeli geliştirilmiştir (Molnar ve ark., 2021). Bunlar arasında en yaygın kullanılan tahmin modelleri, Naive Bayes, lojistik regresyon, karar ağaçları, rastgele orman, en yakın k komşu ve destek vektör makineleri olarak karşımıza çıkmaktadır. (Roscher ve ark., 2020; Atalay ve Çelik, 2017).

2.2.2. Makine Öğreniminin Tarihsel Gelişimi

Son dönemde yapay zeka alanında yaşanan gelişmeler de dikkate alındığında makine öğrenimi, yapay zeka, derin öğrenme, yapay sinir ağları gibi bir çok farklı ancak iç içe geçmiş kavramla karşılaşılmaktadır (Çalışkan ve ark., 2020). Makine öğrenimi, yapay zeka arayışından ortaya çıkan bir bilimsel çabanın neticesinde şekillenmeye başlamıştır. Alan Turing, II (Turing, 1950). Dünya Savaşındaki şifre kırıcılığı ile tanınmaktadır. Alan Turing'in 1950 yılında yayınlamış olduğu "Bilgisayar Makineleri ve Zeka" başlıklı makalesinde "Bilgisayarların kendi zekaları var mı?" diye sorması ile bu arayışın ilk adımı atılmıştır (Turing, 1950; Koç, 2019).

Makine öğrenimi terimi ise ilk kez 1952 yılında Arthur Samuel tarafından kullanıldı. Arthur Samuel geliştirdiği bir program ile dama oyununda yapılan hataların daha sonra tekrarlanmamasına dayanan bir öğrenme modeli ile ilk makine öğrenimi hayata geçirmiş oldu (Koç, 2019; Singh ve Baluja, 2017).

Atılan bu ilk adımlardan sonra farklı alanlarda geliştirilen tahmin modelleri ve algoritmalar ile makine öğrenimi hayatın içinde yer almaya başlamıştır (Artrith ve ark., 2021). Hatta zamanla sadece sayısal veriler üzerinde değil aynı zamanda resim ve benzeri görüntü çıktıları kullanılarak geliştirilen makine öğrenimi yöntemleri akademik camianın ilgi alanına girmiştir (Machado ve ark., 2021).

Makine öğrenimi ve yapay zeka ile kentlerin trafik sistemlerinin yönetilmesi, şoförsüz araçların trafikte bulunması, insanların günlük yaşamda karşılaştığı ağır ve tehlikeli işleri yürütecek yapay yardımcıların üretilmesi, uzuv kaybı bulunan insanların yaşamını kolaylaştıracak aparatlar ve yapay organların geliştirilmesi, COVID-19 gibi pandemiye dönüşen büyük hastalıkların geleceğe dönük tahminleri gibi birçok çalışma yürütülmektedir ve geliştirilmeye devam etmektedir (Campbell ve ark., 2021; Rustam ve ark., 2020).

2.2.3. Makine Öğreniminin Kullanım Alanları

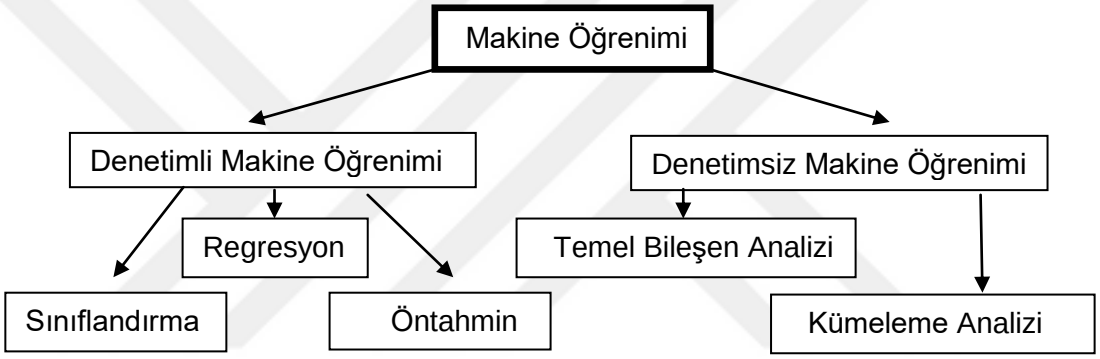
Makine öğrenimi tahmin modellerinin başarı oranı arttıkça kullanıldığı alanlarda genişlemekte ve farklı konularda çalışmalar yapılmaktadır (Raschaka ve ark., 2020). Günümüzde finans, sağlık, eğitim, ulaşım, lojistik, reklamcılık ve daha birçok sektör ve alanda makine öğrenimi kullanımı yaygın hale gelmiştir (Morency ve ark. 2022). Sağlık sektöründe makine öğrenimi kullanımı henüz gelişme aşamasında olduğundan araştırmacılar tahmin modellerinin geliştirilmesi üzerinde çok çeşitli ve farklı araştırmalar yapmayı sürdürmektedirler (Kaur ve ark., 2018). Ayrıca makine öğrenimi sadece matematiksel veri üzerinden değil aynı zamanda görüntü çıktıları üzerinden de tahmin modelleri üretebildiğinden, sağlık alanında manyetik rezonans (MR) görüntüleme, radyografi, ultrasonografi görüntüleri üzerinden de tahmin modelleri oluşturulabilmektedir (Bailey ve ark., 2018; Sharma ve ark., 2016).

2.2.4. Makine Öğrenimi Yöntemleri

Makine öğrenimi yöntemlerini genel olarak denetimli öğrenme (supervised learning) ve denetimsiz öğrenme (unsupervised learning) olarak ikiye ayrılmaktadır

(Gökalp, 2021). Denetimli yöntemler genellikle veriden bilgi ve sonuç çıkarma amacıyla kullanılırken, denetimsiz yöntemler daha çok veriyi tanımaya ve anlamaya yönelik olarak kullanmayı ve sonraki uygulanacak yöntemler için fikir vermeyi amaçlamaktadır. Denetimli öğrenme sınıflandırma, regresyon ve tahmin problemlerinde kullanılırken denetimsiz öğrenme yöntemi kümeleme problemlerine çözüm aramaktadır (Chang ve ark., 2020). Bazı akademisyenler yarı denetimli öğrenme (semi-supervised learning), takviyeli öğrenme (reinforcement learning), transdüktif çıkarsama (transductive inference), çevrim içi öğrenme (on-line learning) ve aktif öğrenme (active learning) yöntemlerine de çalışmalarında yer vermektedirler (Mahesh, 2018).

Makine öğreniminin aşamalarının görüldüğü şema Şekil 2.1'de verildi.

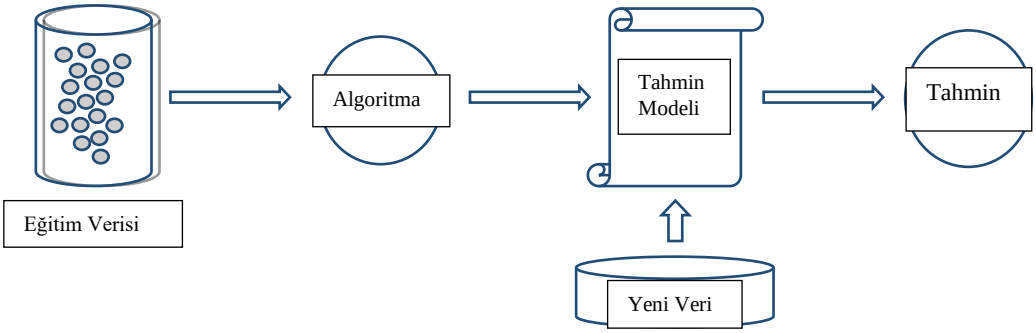


Şekil 2.1. Makine öğrenimi şeması

Denetimli Öğrenme

Gözetimli öğrenme olarak ta bilinen denetimli makine öğrenimi, belirsizlik durumunda kanıta dayalı tahminler yapan bir model oluşturur. Bu öğrenme yönteminde eğitim veri seti üzerinden sınıflamaya dair bir tahmin algoritması oluşturulmakta ve yeni gelecek olan örneğin veri seti değişkenleri üzerinden hangi sınıflamaya dahil olacağını öğretmektedir (Osisanwo ve ark., 2017). Örneğin daha önce yapılmış tahliller ile herhangi bir hastalığa sahip olan ya da olmayan bireylerden oluşan bir veri setini eğitim seti olarak kullanan bu öğrenme yöntemi yeni gelen bir hastanın tahlil sonuçlarını değerlendirerek belli bir hastalığa sahip olup olmadığını sınıflandırmaktadır. Bu öğrenme yönteminin akış şeması aşağıdaki gibi gösterilmektedir (Muhammad ve Yan, 2015).

Denetimli makine öğrenmesine ilişkin aşamaların bulunduğu şema Şekil 2.2'de verildi.



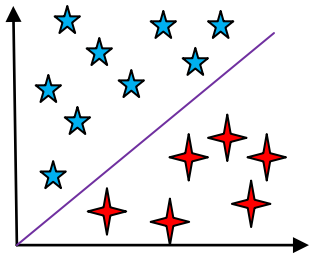
Şekil 2.2. Denetimli öğrenme şeması

Makine öğreniminde denetimli öğrenme yöntemlerini için sınıflandırma, regresyon ve tahmin teknikleri kullanılmaktadır (Xu H ve ark., 2019).

Sınıflandırma

Sınıflandırma denetimli öğrenme yöntemi bir örneğin hedef değişkende hangi kategoriye ait olduğunu araştırır. Bunu yaparken de geçmiş gözlemlerden faydalanarak ikili sınıflandırma, çoklu sınıflandırma ve çoklu etiket sınıflandırma yöntemlerini kullanır.

Bu çalışmada KBY (var, yok) şeklinde ikili sınıflandırma tekniği uygulanmıştır. Çoklu sınıflandırma tekniklerinde kategori sayısı ikiden fazla olmaktadır. Çoklu etiket sınıflandırmasında ise bağımsız değişkenin birden fazla kategoriye ait olabileceği olasılığı göz önünde bulundurulmaktadır. Örneğin bir filmin hem aksiyon, hem komedi hem de dram türünde sınıflandırılabileceği şeklinde bir uygulama olabilmektedir (Tütüncü, 2022).



Şekil 2.3. İkili sınıflandırma

En çok kullanılan sınıflandırma yöntemleri arasında karar ağaçları, lojistik regresyon, Naive Bayes, destek vektör makineleri, rastgele orman, en yakın k-komşu, tek sınıf gibi algoritmalar yer almaktadır.

Regresyon

Regresyon istenen çıktı değerlerinin sürekli değişkenlerden oluştuğu durumlarda kullanılan yöntemler arasındadır. Bu teknikte değişkenlerin boyut değerleri arasında bir ilişki aranmaktadır. Örneğin hasta yaşları ile hastaların kreatinin sayısal değerleri, ya da enflasyon değeri ile hasta muayene ücretleri arasındaki ilişkiyi öğrenmek istediğimizde bu teknikten faydalanmak mümkündür.

Günümüzde kullanılan regresyon yöntemleri arasında lineer regresyon, çoklu lineer regresyon ve polinomial regresyon en sık kullanılanlar arasındadır (Muhammad ve Yan, 2015).

Öntahmin

Bu yöntemde geçmişteki ve mevcutta bulunan verilere dayanılarak, gelecek hakkında tahminde bulunma süreci söz konusudur. Bu yöntem sıklıkla eğilimleri analiz etmekte kullanılmaktadır (Tütüncü, 2022). Örneğin Merkez Bankası'nın her ay yapmış olduğu "Tüketici Eğilim Anketleri"nin analizinde kullanılan bir yöntemdir. Bu araştırma ile tüketicilerin enflasyon başta olmak üzere çeşitli ekonomik göstergelerin bir sonraki dönemde ne olacağı üzerine belirttikleri tahminler analiz edilmekte ve sonuç tahminleri yayınlanmaktadır.

Denetimsiz Öğrenme

Denetimsiz öğrenme, elde etmek istenen çıktının nasıl göründüğü konusunda çok az ya da herhangi bir fikir sahibi olunmadığı durumlarda kullanılan bir makine öğrenimi yöntemidir. Bu öğrenme yönteminde sadece veri kümelerinden sonuçlar elde edilmeye çalışılmaktadır. Temel bileşen analizi (**P**rincipal **C**omponent **A**nalysis) ve kümeleme analizi (**C**luster **A**nalysis) denetimsiz makine öğrenimi tekniklerindendir (Dündar ve ark., 2021).

Temel bileşen analizi ilk olarak 1901 yılında Karl Pearson'un çalışmalarıyla başlamış ve 1933 yılında Hotelling tarafından geliştirilmiştir. Bu yöntem çok sayıda birbiri ile ilişkili değişkenlerden oluşan veri seti boyutlarının daha az boyuta indirgenmesi tekniğidir (Yazar ve ark., 2009). Kümeleme analizi ise gözlemler arasındaki benzerlik yapılarına göre verinin farklı kümelere ayrılması yöntemine dayanmaktadır. Bu şekilde benzer özelliklere sahip veriler aynı kümede toplanmaktadır (Ding ve ark., 2011).

3. GEREÇ VE YÖNTEM

3.1. Uygulama Verisi

Çalışmada kullanılacak olan veri seti Ondokuz Mayıs Üniversitesi Tıp Fakültesi Nefroloji Anabilim Dalı'na 2015-2021 yılları arasında başvuran 18 yaşından büyük hastalardan elde edildi. Bu çalışma Ondokuz Mayıs Üniversitesi Etik Kurulu tarafından onaylanmış (Tarih: 16.08.2021 Sayı: B.30.2.ODM.0.20.08/378-493) ve Ondokuzmayıs Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından PYO.TIP.1904.22.002 proje numarası ile desteklenmiştir.

Çalışma verisi için kullanılan değişkenlere ilişkin açıklamalar Tablo 4.1'de verildi.

Tablo 3.1. Çalışma verisine ait açıklamalar

	Değişken	**Veri Tipi**
1	Hastanın cinsiyeti (kadın/erkek)	Kategorik
2	Hastanın yaşı	Sayısal
3	Albumin (düşük/normal/yüksek)	Kategorik
4	Dansite(düşük/normal/yüksek)	Kategorik
5	Fosfor (düşük/normal/yüksek)	Kategorik
6	Şeker (normal/yüksek)	Kategorik
7	Açlık şekeri (düşük/normal/yüksek)	Kategorik
8	Hematokrit (düşük/normal/yüksek)	Kategorik
9	Hemoglobin (düşük/normal/yüksek)	Kategorik
10	Kalsiyum (düşük/normal/yüksek)	Kategorik
11	Kreatinin (düşük/normal/yüksek)	Kategorik
12	Potasyum (düşük/normal/yüksek)	Kategorik
13	RBC (Kırmızı kan hücresi sayımı) (düşük/normal/yüksek)	Kategorik
14	Sodyum (düşük/normal/yüksek)	Kategorik
15	WBC (beyaz kan hücresi sayımı) (düşük/normal/yüksek)	Kategorik
16	BUN (düşük/normal/yüksek)	Kategorik
17	Enfeksiyon varlığı (var/yok)	Kategorik
18	Hipertansiyon (var/yok)	Kategorik
19	Diyabet (var/yok)	Kategorik
20	Kalp hastalığı (var/yok)	Kategorik
21	Böbrek nakli (var/yok)	Kategorik
22	Kronik böbrek yetmezliği (var/yok)	Kategorik

Tablo 3.1 incelendiğinde toplam 9.136 hastaya ilişkin 21 farklı değişken üzerinden değerlendirmeler yapılmıştır. Bu hastalardan 4.477'si (%49) erkek, 4.659'u (%51) kadındır.

Verilerin düzenlenmesi için STATA ve R yazılımları kullanıldı. Veri üzerinde gerekli düzenlemeler yapıldıktan sonra WEKA yazılımı aracılığıyla tek sınıf, Naive Bayes, lojistik regresyon, karar ağaçları, rastgele orman, en yakın k komşu ve destek vektör makineleri algoritmalarının tahmin oranları karşılaştırılarak değerlendirildi.

Veri setinin çalışmaya uygun hale getirilebilmesi için aşağıdaki adımlar izlendi.

3.2. Adımlar

3.2.1. Verinin Toplanması

Makine öğreniminde uygulanacak verinin gerçek hastalara ait değerlerden oluşması çalışmanın sonucunu olumlu yönde etkilemektedir. Veri setinin büyüklüğü ve gerçek hastalara ilişkin sonuçların etkinliği de denetimli makine öğrenmesi yöntemlerinde istenilen özellikler arasında yer almaktadır (Nasteski, 2017). Bu çalışmada kurum izni ve etik kurul izinleri alınarak 2015-2021 yılları arasında Samsun Ondokuz Mayıs Üniversitesi Tıp Fakültesi Nefroloji Anabilim Dalına başvuran tüm hastaların cinsiyet, yaş, albumin, dansite, fosfor, şeker, açlık şekeri, hematokrit, hemoglobin, kalsiyum, kreatinin, potasyum, kırmızı kan hücresi sayımı, sodyum, beyaz kan hücresi sayımı ve bun değerleri elde edilmiştir. Ayrıca bu hastalara ilişkin enfeksiyon varlığı, hipertansiyon, diyabet, kalp hastalığı, böbrek nakil durumu ve KBY hastalığı olup olmadığı da incelenerek veri setine eklenmiştir.

3.2.2. Veri Hazırlama

Toplanan verilerin analize hazır hale getirilmesi gerekmektedir. Makine öğrenimi algoritmalarını uygulanabilmesi için veri setinin uygun hale getirilmesi çeşitli yöntemlerle uygulanmaktadır, bu yöntemler arasında verinin temizlenmesi, entegrasyonu, dönüşümü, indirgenmesi, filtrelenmesi, istenilen aralıkta kategorize edilmesi ön plana çıkanlardır (Brownlee, 2022).

Bu çalışmada toplanan ham veri STATA ve R programları aracılığıyla temizleme, filtreleme, kategorize etme ve birleştirme aşamalarından geçirilerek son analiz aşamasına gelinmiştir. Temizleme aşamasında verinin bulunduğu değişkene ait olmayan ve aykırı olan değerler veri setinden çıkartılmıştır. Ayrıca bir hastaya ait veri setinde herhangi bir ya da daha fazla değişken değeri eksik olanlar veri setinden tamamen çıkartılmıştır. Aynı hastanın aynı değişkenlerine ait birden çok değer farklı tarihlerde oluşabildiğinden filtreleme aşamasında bu değerlerden en son tarihte

ortaya çıkmış olan değer esas alınmıştır. Filtreleme aşamasında cinsiyet ve yaş durumlarına göre değişken değeri en son tarihte olan hastaya ilişkin veriler alınmıştır. Kategorize etme aşamasında veri değerleri laboratuvar sonuçlarında esas alınan alt sınır ve üst sınır değerlerine göre ikili ya da üçlü kategorilere dönüştürülmüştür. Bu kategorize etme işlemi sırasında laboratuvar değerlerinin sınırlarının belirlenmesinde bazı değişkenler için cinsiyet ayrımında farklı sınır değerleri kullanılmıştır. Son aşama olarak ayrı dosyalardan oluşan veri setleri için ilk üç aşamanın gerçekleştirilmesine müteakiben tüm dosyalar birleştirilerek tek bir veri tablosu oluşturulmuştur. Veri kaynaktan ilk elde edildiğinde milyonu aşkın veri yığını bulunurken bu dört aşamanın sonucunda 9.136 hastaya ilişkin 22 değişken ölçümü ile toplamda 200.992 veriye indirgenmiştir. Veri setinde eksik gözleme yer verilmemiştir. Son olarak veri seti excel formatından WEKA üzerinden denetimli makine öğrenimi algoritmaları için tahmin modeli oluşturmaya uygun format olan, ayraç olarak virgül kullanılan "csv" formatına dönüştürülmüştür.

3.2.3. Veriye İlişkin Açıklamalar

Çalışmada kadın ve erkeklere ait yaş dağılımı Tablo 3.2'de verildi.

Tablo 3.2. Hastaların cinsiyet dağılımı ve yaş ortalamaları

Cinsiyet	Sayı	%	AO±SS (Yaş) *
Erkek	4.477	49,00	54,51±17,28
Kadın	4.659	51,00	52,86±17,79
Toplam	9.136	100	53,67±17,56

*AO±SS(Aritmetik ortalama ± standart sapma)

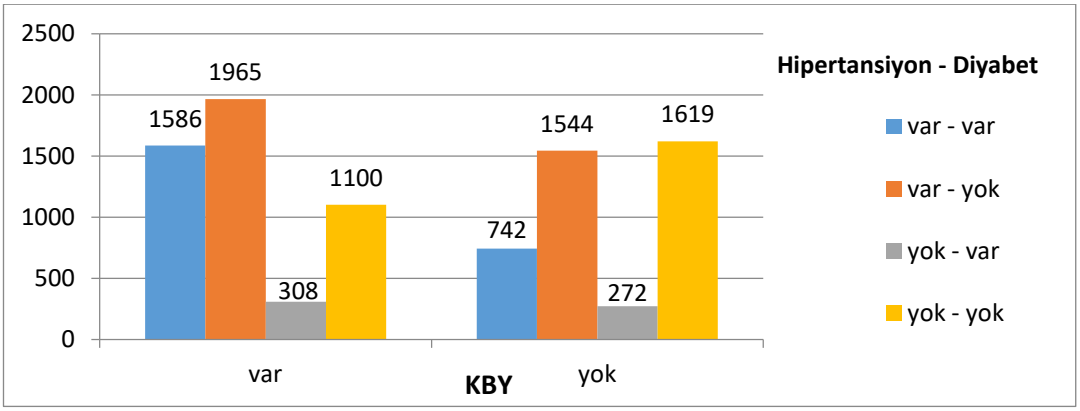
Çalışma verisine ait hastalık tanılarının cinsiyete göre dağılımı Tablo 3.3'de verildi.

Tablo 3.3. Hastaların cinsiyete göre hastalık tanı durumları

Değişken		Erkek		Kadın		TOPLAM
		Sayı	%	Sayı	%	
KBY	var	2.698	54,41	2.261	45,59	4.959
	yok	1.779	42,59	2.398	57,41	4.177
Böbrek nakli	var	131	56,22	102	43,78	233
	yok	4.346	48,82	4.557	51,18	8.903
Hipertansiyon	var	2.887	49,46	2.950	50,54	5.837
	yok	1.590	48,20	1.709	51,80	3.299
Diyabet	var	1.422	48,90	1.486	51,10	2.908
	yok	3.055	49,05	3.173	50,95	6.228
Kalp yetmezliği	var	866	57,01	653	42,99	1.519
	yok	3.611	47,41	4.006	52,59	7.617
Enfeksiyon varlığı	var	962	43,18	1.266	56,82	2.228
	yok	3.515	50,88	3.393	49,12	6.908

Tablo 3.3 incelendiğinde 4.959 (%54,28) KBY hastası ve 4.177 (%45,72) KBY hastalığı olmayan kişi bulunmaktadır. KBY hastalığı bulunan hastaların 1.586'sının (%31,98) hem hipertansiyon hem de diyabet hastalığı bulunmakta, 1.965'inde (%39,62) hipertansiyon varken diyabet hastalığı bulunmamakta, 308'inde (%6,21) hipertansiyon yokken diyabet hastalığı bulunmakta, 1.100'ünde (%22,19) ise hipertansiyon ve diyabet hastalığı bulunmamaktadır

KBY hastalarının hipertansiyon ve diyabet hastası olma durumları Şekil 3.1'de verildi.



Şekil 3.1. KBY hastalarının hipertansiyon ve diyabet hastası olma durumları

Şekil 3.1 incelendiğinde KBY hastalığı bulunmayan hastaların 742'sinin (%17,77) hem hipertansiyon hem de diyabet hastalığı bulunmakta, 1.544'ünde

(%36,96) hipertansiyon varken diyabet hastalığı bulunmamakta, 272'sinde (%6,51) hipertansiyon yokken diyabet hastalığı bulunmakta, 1.619'unda (%38,76) ise hipertansiyon ve diyabet hastalığı bulunmamaktadır.

3.2.4. Model Oluşturulması ve Modellerin Çalışma Esasları

Veri seti makine öğrenimi algoritmalarında kullanılmak üzere hazır hale getirildikten sonra tahmin modellerinin oluşturulması aşamasına geçilir. Burada modelin en iyi tahmini yapabilme kapasitesi önem arz etmektedir. Oluşturulan makine öğrenimi modellerinin tahmininin yüksek başarı oranına sahip olması için veri setinin içinden eğitim verisi olarak adlandırılan bölüme ayrılması temel alınmaktadır. Buradaki amaç sınıf değişkeni olarak bilinen sonucun eğitim veri setinde tanımlı olmasından dolayı diğer değişiklerin aldığı değerlere göre modele sonuç tahmininin öğretilmesidir. Bu işlem sonucunda veri setinden test verisi oluşturularak sonucun tahmin yüzdesi test edilmektedir. Test verisinin temel özelliği sınıf değişkeninin sonucunu bilmemesidir. Bu değişken sonucunu eğitim verisinde öğrendiği model ile tahmin etmektedir. Eğitim veri seti bu tür çalışmalarda genellikle tüm veri setinin %60-80'ni kapsıyorken, test veri setinde bu oran %20-40 arasında olmaktadır.

Makine öğreniminde kullanılan yazılımlarda başarı yüzdesinin arttırılması için veri seti onlarca eğitim ve test verisine ayrılır ve modeli sürekli öğrenime açık hale getirir (Hsieh ve ark., 2020).

3.2.5. Modellerin Karşılaştırılması

Bu aşamada kullanılan farklı makine öğrenimi algoritmaları çalıştırılarak performansları değerlendirildi. Kullanılan modeller ile gerçeğe en yakın tahmin yapmak hedeflenmektedir. Makine öğrenimi algoritmaları oluşturdukları modeller ile gerçeğe en yakın tahmini ortaya koymaya çalışmaktadırlar (Tunthanathip ve ark., 2021). Veri setinin özelliklerine göre farklı algoritmalar farklı performanslar ortaya koyabilmektedirler.

3.3. Makine Öğrenimi Algoritmaları

3.3.1. Tek Sınıf (OneR) Algoritması

Tek sınıf algoritmasında (TSA) değişkenlerden en iyi sonucu veren değişken bulunması hedeflenmektedir. TSA yönteminde her bir değişken için bir çapraz tablo

oluşturulmakta ve bu matrislerdeki doğruluk oranları hesaplanarak en iyi sonucu veren değişken kullanılması yoluyla sınıflandırma yapılmaktadır. Doğruluk oranı hesaplanırken değişken içerisindeki sınıftan hedef değişkende en çok kategorisi olan esas alınır (Bhalodiya ve Parsania, 2014).

Aşı olma, maske kullanma ve toplu taşıma kullanma durumuna göre COVID-19'a yakalanma durumunun incelenmek istendiği bir örnek için ilgili veriler Tablo 3.4' de verildi.

Tablo 3.4. COVID- 19'a yakalanma durumu

	Aşı olma	**Maske takma**	**Toplu taşıma kullanma**	**Covid-19 olma**
1	Evet	Evet	Hayır	Hayır
2	Evet	Hayır	Evet	Evet
3	Hayır	Evet	Hayır	Evet
4	Hayır	Evet	Evet	Evet
5	Evet	Evet	Evet	Hayır
6	Evet	Evet	Hayır	Hayır
7	Hayır	Hayır	Evet	Evet
8	Hayır	Evet	Evet	Evet
9	Hayır	Hayır	Hayır	Evet

Tablo 3.4'de bulunan bilgiler ışığında hedef değişken olarak COVID-19' a yakalanma durumu alınarak TSA'nın hesaplanabilmesi için her bir değişken için ayrı çapraz tablo hesaplanmaktadır.

Değişkenlerde aşı olma durumu için çapraz tablo Tablo 3.5'de verildi.

Tablo 3.5. Aşı olma durumu için çapraz tablo

		Aşı olma	
		Evet	Hayır
COVID-19	Evet	1	5
	Hayır	3	0

TSA için ilk önce çapraz tablo üzerindeki her bir sütun için hangi satırın seçileceğine karar verilir. Tablo 3.5 incelendiğinde aşı olma durumu "Evet" iken COVID-19 olma durumu 1 adet "evet", 3 adet "hayır" olduğundan sayısı fazla olan "hayır" seçilir. Aşı olma durumu "hayır" iken COVID-19 olma durumu 5 adet "evet" olduğundan "evet" seçilir. Bu durumda bu sınıf için toplam 8 doğru tahmin ve 1

yanlış tahmin yapılmış olur. Doğru tahmin yapma oranı 8/9 = 0,889 olarak hesaplanır.

Maske takma ve toplu taşıma kullanma değişkenlerine göre de çapraz tablo oluşturulup doğru tahmin etme sayıları da bu yöntem ile hesaplanır. Hesaplama sonucunda her iki değişken için doğru tahmin sayısı 6 çıkmaktadır. Buna bağlı olarak doğru tahmin etme oranları 6/9 = 0,667 bulunmaktadır. Bu hesaplamalardan sonra TSA en yüksek doğruluk oranını veren "aşı olma durumu" değişkenini tahmin için esas alır ve diğer değişkenler ile ilgilenmez. En yüksek doğru tahmini birden fazla değişken gerçekleştirdiği durumda algoritma rastgele bir tanesini seçmektedir (Tan ve Gilbert, 2003).

3.3.2. Naive Bayes Algoritması

Bayes Teoremi kapsamında kullanılan olasılıksal bir sınıflandırma metodu olarak bilinmektedir. Naive Bayes Algoritması (NBA) diğer tahmin modellerinde olduğu gibi daha önce sınıflandırılmış olan verileri kullanarak veri setine yeni eklenen verinin hangi sınıfa ait olduğunu bu tahminin olasılığını hesaplayarak tahmin eder (Gökdemir ve Çalhan, 2022). NBA'nın en önemli özelliklerinden biri de olasılık hesaplaması sırasında niteliklerin birbirinden bağımsız olduğu varsayımı ile hesaplama yapılmasıdır. Bu tahmin modeli eksik gözlem durumunda da iyi sonuçlar ortaya koyduğundan sağlık alanında sık kullanılan bir modeldir. Ayrıca bu model karar ağaçları ve sinir ağları gibi farklı tahmin edici modellerde sınıflandırıcı olarak da kullanılmaktadır (Li Z ve ark., 2020).

$$P(A\backslash B\ ) = \frac{P(B\backslash A)P(A)}{P(B)} \quad \textbf{(3.1)}$$

P(A|B) : B olayı gerçekleştiği durumda A olayının meydana gelme olasılığı,

P(B|A) : A olayı gerçekleştiği durumda B olayının meydana gelme olasılığı,

P(A) ve P(B) : A ve B olaylarının olasılıklarıdır.

Cinsiyet, beden kitle indeksi (BKİ) ve yaş bilgileri ile bir kişinin diyabet hastası olup olmadığının tahminine ilişkin örnek eğitim veri seti Tablo 3.6'da verildi.

Tablo 3.6. Diyabet hastası olma durumu

	Cinsiyet	BKİ	Yaş	Diyabet
1	Erkek	20-30	<30	Hayır
2	Kadın	>30	30-50	Evet
3	Kadın	20-30	>50	Hayır
4	Erkek	20-30	30-50	Hayır
5	Erkek	>30	30-50	Evet
6	Kadın	20-30	30-50	Evet
7	Kadın	>30	30-50	Hayır
8	Kadın	20-30	>50	Evet
9	Kadın	>30	30-50	Evet
10	Erkek	>30	>50	Evet

BKİ'si 20-30 arasında olan ve yaşı 50 den fazla olan bir X kadın kişisinin diyabet hastası olup olmadığını tahmin edilmesi için Tablo 3.6'daki verilerden faydalanılarak NBA modeli oluşturuldu.

NBA modelini oluşturmanın ilk adımı olan hedef değişkeninin (diyabet hastalığı) için olasılıkların hesaplanmasıdır:

P(Diyabet hastalığı=Evet) = 6/10 = 0,6

P(Diyabet hastalığı=Hayır) = 4/10 = 0,4

Daha sonra diyabet hastası olup olmadığı öğrenilmek istenen kişiye ait tüm özelliklerin diyabet hastası olup olmama durumuna göre olasılıklar ayrı ayrı hesaplanır.

P(Cinsiyet=Kadın | Diyabet hastası = Evet) = 4/6= 0,67

P(Cinsiyet=Kadın | Diyabet hastası) Hayır) = 2/4 = 0,5

P(BKİ= 20-30 | Diyabet hastası =Evet) = 2/6 = 0,33

P(BKİ= 20-30 | Diyabet hastası =Hayır) = 3/4 = 0,75

P(Yaş>50 | Diyabet hastası =Evet) = 2/6 = 0,33

P(Yaş>50 | Diyabet hastası =Hayır) = 1/4 = 0,25

Son olarak diyabet hastalığının olduğu tüm durumların olasılıkları birbiri ile olmayanların tüm durumlarının olasılıkları da birbiri ile çarpılır. Hangi ihtimalin sonucu daha yüksek ise bu kişi o sınıfa dahil edilir.

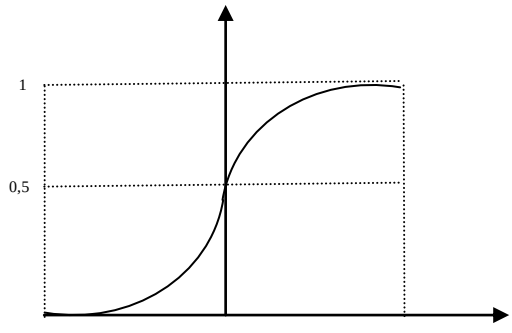
P(X | Diyabet hastası = Evet) = 0,6*0,67*0,33*0,33 = 0,0444

P(X | Diyabet hastası = Hayır) = 0,4*0,5*0,75*0,25 = 0,038

Bu ihtimaller sonucunda 0,044 > 0,038 olduğundan X kişisi diyabet hastası olarak sınıflandırılır. Bu algoritmanın en önemli olumsuz özelliği olasılıklardan herhangi birinin sıfır çıkması durumudur. Tüm olasılık çarpımlar sonucu hesaplandığından olasılıklardan herhangi birisi sıfır çıkarsa sonuç olasılığı da sıfır olacaktır. Bu durumu ortadan kaldırmak için en yaygın kullanılan yöntem Laplacian düzeltmesidir. Bu düzeltmelerde sıfırdan kurtulmak için tüm ihtimal hesaplarının payına 1 eklenir bu şekilde sonuç hiçbir zaman sıfır çıkmamaktadır (Rizki ve ark., 2021).

3.3.3. Lojistik Regresyon Algoritması

Lojistik regresyon verilerin sınıflandırılmasında kullanılan bir regresyon modeli olarak bilinmektedir. Lojistik Regresyon Algoritması (LRA) modeli bağımlı değişkeninin iki ya da ikiden fazla kategoriden oluştuğu durumlarda niteliksel veri türünde ve doğrusal sınıflandırma problemlerinde yaygın olarak kullanılmaktadır. Ayrıca bu algoritmada bağımsız değişkenler niteliksel ya da niceliksel veri türünde olabilmektedir (Alpar, 2017). LRA'nn kullanılabilirliğini arttıran önemli etkenlerden birisi de normal dağılım ve süreklilik varsayımı önkoşulu olmayışıdır (Alotaibi ve ark., 2020). Bu algoritma logaritmik fonksiyonu kullanarak verileri 0 ile 1 arasında konumlandırmakta ve 0,5 ve daha büyük olanları 1 olarak sınıflandırırken 0,5 den küçük olanları 0 olarak sınıflandırma şeklinde bir modelleme oluşturmaktadır. Lojistik regresyon algoritmasına ait grafik Şekil 3.2'de verildi (Osmanoğlu ve ark., 2020).



Şekil 3.2. Lojistik regresyon

Lojistik regresyon formülü Eşitlik 3.2'de verildi. Bu formül ile veri 0 ile 1 arasında bir değer ile sonuçlanmaktadır;

$$\sigma(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}} \quad \textbf{(3.2)}$$

Formülü değişkenlerin olduğu bir fonksiyon aşağıdaki gibi gösterilir (Rymarczyk ve ark., 2019);

$$F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad \textbf{(3.3)}$$

Lojistik regresyon modelinde en çok olabilirlik yöntemi kullanılarak değişkenlere ait katsayıların kestirimi yapılır. Kestirimi yapılan katsayıların yorumlanması için odds oranından (OR) faydalanılır. Odsa oranı formülü eşitlik 3.4'e verildi;

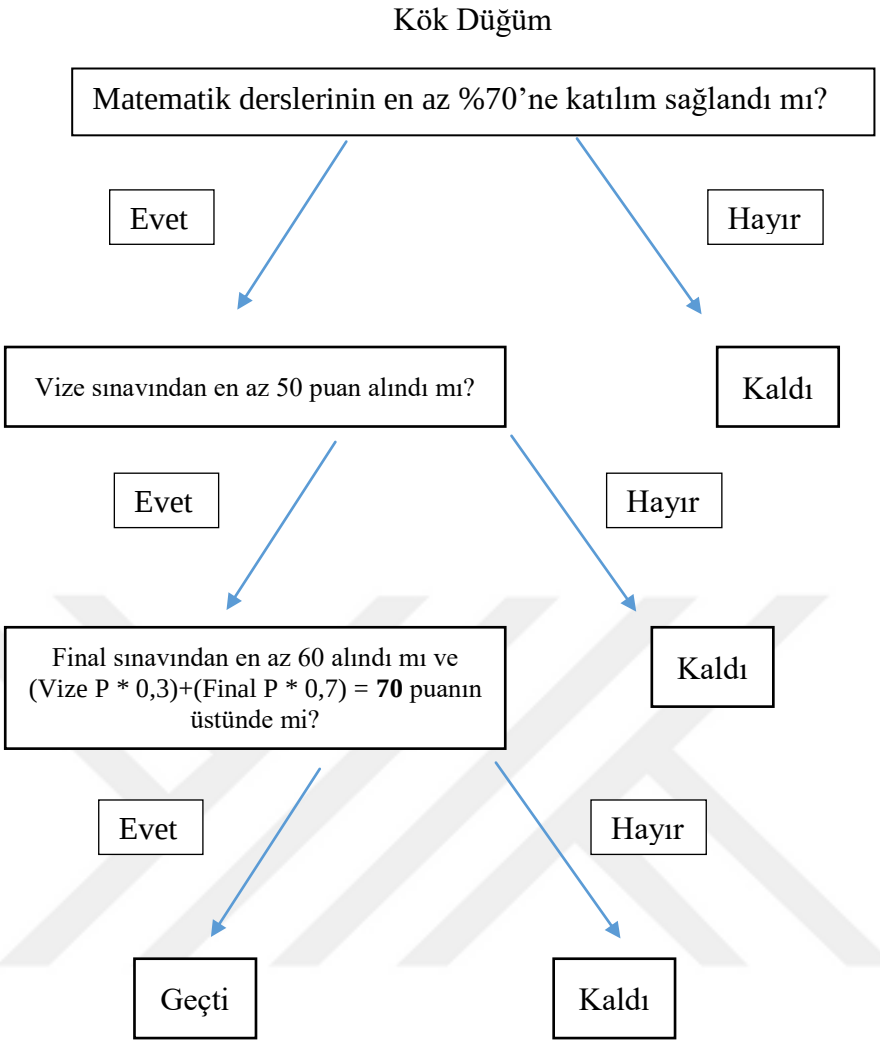
$$OR = \frac{\pi(x_i)}{(1-\pi(x_i))}. \quad \textbf{(3.4)}$$

OR değeri 0 ile 1 değeri arasında olursa risk faktörünün sonuç değişkeni "koruyucu", 1 olursa risk faktörü ile sonuç değişkeni arasında fark olmadığı, 1'den büyük olursa risk faktörü ve sonuç değişkeni arasında fark olduğu söylenebilmektedir (Yavuz ve Çilengiroğlu, 2020).

3.3.4. Karar Ağaçları Algoritması

Karar ağaçları algoritması (KAA) için ana hedef, büyük ölçülü veri kümelerini bir dizi karar kurallarıyla daha küçük veri kümelerine dönüştürmek ve bu küçük kümeler vasıtasıyla doğru sınıflamayı yapabilmektir. Bu algoritma özellikler ve potansiyel çıktılar arasındaki bağlantıyı kurallı bir modele dönüştürmek için ağaç şeklinde bir yapı ile algoritma oluşturan güçlü sınıflandırıcılardan birisidir. KAA'da ağaç şeklinde bir yapı kullanılmakta, ağacın kök düğümünden başlayarak her bir dalında doğru sınıflama için seçenekler oluşturulmakta ve sınıflama yöntemi aşağı doğru dallanarak devam etmektedir. Burada en üstte bulunan birim kök, en altta bulunan birim yaprak, kök ve yaprak arasında kalan birimler ise dal olarak adlandırılır. Karar ağacı yönetiminde bir olayın sonuçlandırılmasında sorunun cevabına göre hareket edilir (Burkart ve Huber, 2021).

Örneğin bir öğrencinin matematik dersinden geçip geçmediğini sınıflamak için oluşturulan karar ağacı modeli Şekil 3.3. ile verildi.



Şekil 3.3. Karar ağacı model örneği

Şekil 3.3.'de gösterilen karar ağacı modeli için matematik dersinden geçme koşulları, tüm derslerin en az %70'ne katılım sağlanma zorunluluğu, birinci sınav olan vize sınavından en az 50 puan alma zorunluluğu, 2.sınav olan final sınavından en az 60 alma zorunluluğu ve vize sınavının %30 ile final sınavının %70 nin toplamının en az 70 puan olma zorunluluğu olarak değerlendirilmiştir.

Bu çalışmada olduğu gibi hedef değişkenin kategorik olduğu sınıflamalarda en çok entropi, gini, sınıflandırma hatası yöntemleri ile karar ağacı modelleri oluşturulmaktadır (Matuszny, 2020).

Entropi verilerin belirsizliğine ilişkin bir ölçüdür. Bir veri kümesinde ne kadar az entropi bulunursa o kadar belirsizlik azalacağından modelin performansı da yüksek olacaktır (Damanik ve ark., 2019). Eşitlik 3.5 ile Claude Shannon'un entropi denklemi veridi;

$$H = -\sum p(x).\log p(x) \quad \textbf{(3.5)}$$

Burada p(x) belirli bir sınıfa ait grubun yüzdesel ifadesini gösterirken H entropiyi ifade etmektedir.

En iyi bölünmeyi ortaya koymak için bilgi kazancı hesaplanmaktadır. Bigi kazancı aşağıdaki gibi elde edilir;

$$Bilgi\ kazancı\ (S, D) = H(S) - \sum_{V \in D} \frac{|V|}{|S|} H(V) \quad \textbf{(3.6)}$$

Burada S orijinal veri kümesini ifade ederken, D bölünmüş alt parçayı göstermektedir. Her V ise S'nin bir alt kümesi olarak değerlendirilmektedir. Bilgi kazancı, orijinal veri setinin bölünmeden önceki entropisi ile her özniteliğin entropi değeri arasındaki fark olarak ifade edilmektedir (Ellerman, 2022).

KAA'da veri seti, eğitim verisi ve test verisi olarak ikiye ayrılmaktadır. Burada eğitim verileri kullanılarak oluşturulan model test verisi üzerinde uygulanarak başarı yüzdesi hesaplanmaktadır. KAA için ID3, C4.5 ve CART algoritmaları en yaygın kullanılan modellerdir. Bu çalışmada KAA olarak C4.5 algoritması kullanıldı. C4.5 algoritmasında kazanç oranı ölçütü kullanılmaktadır ve Eşitlik 3.7'de verildi.

$$Kazanç\ Oranı\ (A) = \frac{Bilgi\ Kazancı\ (A)}{Bölme\ Bilgisi(A)} \quad \textbf{(3.7)}$$

Burada A özelliği X''i $\{X_1, X_2, \ldots, X_v\}$ şeklinde v parçaya ayırır.
Eşitlik 3.8 ile bölme bilgisi verildi (Aldino ve Sulistiani, 2020);

$$Bölme\ Bilgisi_A(X) = -\sum_{j=1}^{v} \frac{|X_j|}{|X|} \times log_2 \left(\frac{|X_j|}{|X|}\right)$$
(3.8)

3.3.5. Rastgele Orman Algoritması

Rastgele orman algoritması (ROA) karar ağaçları topluluğu olarak bilinen birden fazla karar ağacını birleştirerek isabetli ve doğru tahminler yapabilmeyi hedeflemektedir. ROA geleneksel yöntemlerle tahmin edilmesi zor olan modellerin oluşturulmasında sıklıkla kullanıldığından en yaygın kullanılan modeller arasında yer almaktadır. Bu modelin önemli avantajlarından biri de hem sınıflama hem de regresyon problemleri için kullanışlı bir model olmasıdır. ROA'da karar ağaçlarının sayısının artması modele ek rastgelelik katmaktadır. Modelde bir düğüm bölümlere

ayrılırken en önemli özelliği bulmaktan ziyade rastgele bir özellik alt kümesinde en iyi özelliği arar. Bu durum da daha etkin sonuca ulaşan model ve çeşitlilikle sonuçlanmaktadır (Javeed ve ark., 2019). ROA topluluk modelleri içerisinde bulunmaktadır. Çoğunlukla rastgeleliği ile ön plana çıkan topluluk modelleri temel prensibi, tek bir öğrenme setinden birden fazla farklı modelin üretilmesi ve daha sonra bu modellerden elde edilen tahminleri birleştirerek topluluk modelin tahmin edilmesidir.

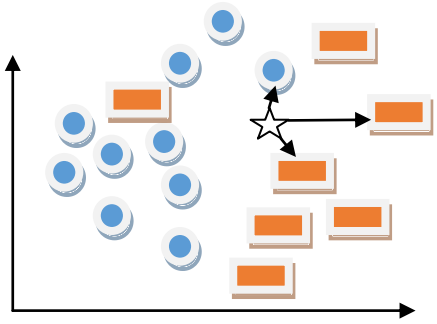
Örneğin yeni bir araba almak isteyen bir kişi bu durumu bir arkadaşına anlattığında arkadaşı ona daha önce kullandığı ya da beğendiği araçlarla ilgili sorular soracak ve o kişinin verdiği cevaplar doğrultusunda ona alacağı arabayla ilgili tavsiyelerde bulunacaktır bu durum bir karar ağacı algoritması oluşturmak gibi değerlendirilirken, bu kişi daha fazla arkadaşına bu konuda danışarak onların benzer soruları karşısında verdikleri tavsiyeleri değerlendirmesi sonucu en çok tavsiye edilen aracı almak istemesi ROA'nın bir örneği olarak değerlendirilebilmektedir.

3.3.6. En Yakın K-Komşu Algoritması

En yakın k-komşu algoritması (EYKKA) olasılık yoğunluk fonksiyonlarının bulunmasının zor olduğu durumlar için geliştirilen bu algoritma, verinin dağılımına ilişkin herhangi bir bilgi bulunmadığı durumlarda kullanılan en temel ve uygulaması oldukça basit olan bir algoritmadır.

Bu algoritma sınıflama yapılacak olan veriye en yakın olan k tane verinin bulunduğu sınıflardan en çok hangi sınıf verilmişse o sınıfın verilmesi şeklinde bir yapıya sahiptir. Burada k genellikle tek sayı olarak alınır zira çift sayı olduğunda sınıfların eşit çıkma ihtimali bulunmaktadır (Fan ve ark., 2019; Xing ve Bei, 2019).

En yakın k-komşu algoritmasına ilişkin bir örnek Şekil 3.4'de verildi.



Şekil 3.4. En yakın k=3 komşu hesaplaması

Şekilde görüldüğü üzere ulaşılmak istenen mavi yuvarlak ve turuncu dikdörtgen sınıflar mevcuttur. Burada amaç yıldız şeklinde hangi sınıfa ait olduğunu bilmediğimiz verinin mavi yuvarlak mı yoksa turuncu dikdörtgen mi olduğuna karar vermektir. Bu şekilde k=3 alındığından hangi sınıfa ait olduğunu bulmak istediğimiz veriye en yakın olan 3 veri tespit edilir bu 3 veriden 2'si turuncu dikdörtgen olduğundan yıldız verimiz için turuncu dikdörtgen sınıfındadır denilmektedir. Bu algoritmada en yakın komşuları tespit edebilmek için minkowsky, chebychev, camberra, manhattan, mahalanobis, öklit gibi uzaklık ölçüleri kullanılmaktadır (Bhavsar ve Ganatra, 2012). Bunlar arasında en yaygın ve etkin olarak kullanılanı öklit mesafesidir.

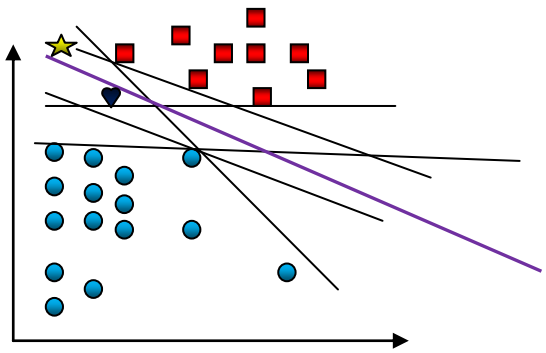
Öklit mesafesine için n değişkene göre formül Eşitlik 3.9'da verildi (Arora ve ark., 2021);

$$Öklit\ mesafesi = \sqrt{(x_2 - x_{1)}{}^2 + (y_2 - y_{1)}{}^2 + \ \ldots + \ (n_2 - n_{1)}{}^2} \qquad \textbf{(3.9)}$$

3.3.7. Destek Vektör Makineleri Algoritması

Destek vektör makineleri algoritması (DVMA) sınıflamada ve regresyonda kullanılan yaygın algoritmalardan birisidir. Bu algoritma hedef değişkendeki sınıfı tahmin etmek için bu sınıfın kategorileri arasında ve birbirine en uzak mesafede olan vektörü çizmeyi hedefler. Bu şekilde oluşturulan tahmin modeli kullanılarak tahmin edilmek istenen verinin oluşturulan vektörün hangi kategori yönünde olduğuna göre bir sonuç oluşturur (Mosavi ve ark., 2018).

Destek vektör makineleri algoritmasına ait grafik Şekil 3.5 ile verildi.



Şekil 3.5. Destek vektör makineleri algoritması

Hangi sınıfa ait olduğu bilinmeyen bir veriyi tahmin edebilmek için Şekil 3.5'de çizilen mor renkli kalın vektör her iki sınıfa da en uzak mesafede olduğundan destek vektör olarak alınır ve bu vektörün üst tarafında (kırmızı karelerin olduğu tarafta) kalan veri kırmızı kare sınıfı olarak değerlendirilir. Şayet veri mor vektörün alt tarafında yer alıyorsa mavi yuvarlak sınıfına dahil edilir. Bu durumda yıldız şekli kırmızı kare sınıfında değerlendirilirken kalp şekli mavi yuvarlak sınıfına dahil edilir (Venkatesan ve ark., 2018).

Girdi ve çıktı değişken çiftleri $Z = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$ olmak üzere, girdi vektörleri, $x \in X$ ve$y \in Y$ etiketleri ile eşlenmektedir. İkili sınıflamada etiketler $Y = \{-1,1\}$ şeklindedir. Amaç yeni (x, y) örneklerini doğru sınıflayan $f \in F$ sınıflayıcılarını hesaplamaktır (Howley ve Madden, 2004). Farklı sınıflar arasındaki birimleri ayırmak için ayırma çizgilerinin çizilmesi ile oluşan sınıflandırma, çoklu düzlem olarak bilinmektedir. Çoklu düzlemlere göre belirlenen DVM farklı sınıf üyeliklerine sahip nesneleri birbirinden ayırmaktadır (Statsoft, 2018).

Çoklu düzlem söz konusu olduğunda iki sorun ortaya çıkmaktadır. Birinci sorun, örnek veri seti için birden fazla çoklu düzlemin meydana gelmesidir. Sınıfları en iyi ayıran çoklu düzlemler içerisinden en yüksek marjine sahip olan düzlem, ideal olan çoklu düzlemdir.

Eşitlik 3.10' da yeni bir x_i sınıflamak üzere çoklu düzlem çözümü verildi (Howley ve Madden, 2004);

$$f(x) = \langle w.x_i \rangle + b \quad \textbf{(3.10)}$$

Burada $\langle w.x_i \rangle$ ağırlıklandırılmış vektör w ile girdi örneğinin iç çarpımı, b ise yanlılık (bias) miktarı olup w'nin her elemanın değeri sınıflama için sahip oldukları nisbi değerinin bir ölçüsünü vermektedir.

İdeal çoklu düzlem için ikinci dereceden kısıtlı optimizasyon problemi Eşitlik 3.11 ve 3.12 ile gösterildi (Howley ve Madden, 2004);

$$Minimize = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \quad \textbf{(3.11)}$$

$$k.s \begin{cases} y_i(w^T x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, \dots, l \end{cases} \quad \textbf{(3.12)}$$

Optimizasyon probleminin amacı bir x değişkenine göre f fonksiyonunun minimize ya da maksimize edilmesidir. Doğrusal olarak ayrılabilen veriler olduğunda DVMA'da ideal çoklu düzlemin bulunması için geometrik marjin (M) en büyük olan olarak ortaya çıkmaktadır (Kowalczyk, 2017). Optimizasyon problemi w'nin normunu minimize ederek yassılığı artırmaktadır. Bu şekilde genelleşme özelliği de artırılmaktadır. Sert marjin ile $\langle w.w \rangle$ için minimum değerinin bulunması gerçekleşirken $f(x)$ hiper düzlemi eğitim verisindeki l örnek veriyi doğru bir şekilde sınıflamış olacaktır. Gevşek değişken ξ_i olmasının sebebi ise bazı örnekleri yanlış sınıflayan hiper düzlemdir. Burada bazı veri setlerinin doğrusal olarak ayrılması mümkün olamamaktadır.

Dual problemin çözümü Eşitlik 3.13'de verildi;

$$W(a) = \sum_i^l \alpha_i - \frac{1}{2}\sum_i^l \alpha_i \alpha_j y_i y_j \langle x_i . x_j \rangle \quad \textbf{(3.13)}$$

Problemin kısıtı ise Eşitlik 3.14'de verildi (Howley ve Madden, 2004);

$$k.s\ C \geq \alpha_i \geq 0, i = 1, \dots, l \quad \sum_i^l \alpha_i y_i = 0 \quad \textbf{(3.14)}$$

Burada α_i ikili problemin çözümü iken karar fonksiyonu Eşitlik 3.15'de verildi;

$$f(x) = \sum_{i=1}^l \alpha_i y_i \langle x . x_i \rangle + b \quad \textbf{(3.15)}$$

Burada b yanlılık miktarı, y_i için $f(x) = 1$ herhangi i için $C > \alpha_i > 0$ olacaktır. Bu durumda yeni bir x_s, $f(x_s)$ sıfırdan küçük olduğunda negatif olarak sınıflanırken $f(x_s)$ sıfırdan büyük ya da eşit olduğunda pozitif olarak sınıflanmaktadır. Örneklemdeki x_i'ler ve eşlenikleri olan α_i sıfırdan farklı olduğunda bunlar destek vektörleri olarak bilinmektedir ve hiper düzleme çok yakın bir konumda yer almaktadırlar. Destek vektör olmayan örneklerde, karar fonksiyonu üzerinde herhangi bir etkileri bulunmamaktadır (Howley ve Madden, 2004).

İkinci problem ise verinin doğrusal olarak ayrılamaması ve buna bağlı olarak girdi uzayındaki nesnelerin matematiksel fonksiyonların kullanılarak yeniden düzenlenmesidir. Matematik fonksiyonları yardımıyla yapılan bu işleme dönüşüm denilmektedir. Özellik uzayında düzenlenen nesneler doğrusal olarak ayrılabilir hale getirilmekte ve dönüşüm çekirdek (kernel) adını almaktadır (Statsoft, 2018). DVM doğrusal özellik uzayındaki veriyi ayırma özelliğine sahip olduğu için çekirdek

fonksiyonun görevi verilerin doğrusal olarak ayrılabildiği daha yüksek bir boyuta eşlemektir (Hofmann, 2006).

3.4. Model Değerlendirilme Ölçütleri

Makine öğreniminin modellerinin karşılaştırılmasındaki esas amaç modelin başarısını ölçmektir. Bu karşılaştırmada kullanılan farklı performans ölçütleri bulunmaktadır. Bu performans ölçütlerinin ile farklı makine öğrenimi yöntemlerinin belirlenen temel esaslar üzerinden karşılaştırılması ve başarı yüzdesi yüksek olan yaklaşımın ortaya çıkartılmasının sağlanması amaçlanmaktadır (Ahmad ve ark., 2020).

Bu çalışmada sınıflama performans ölçümünde kullanılan doğruluk oranı, hata oranı, kesinlik (pozitif kestirim oranı), duyarlılık, seçicilik, F ölçümü, Matthews korelasyon katsayısı ölçütleri ile karşılaştırmalar yapılmıştır.

Bu ölçütlerin hesaplanabilmesi için çapraz tablonun oluşturulması gerekmekte olup, Tablo 3.7 ile verilmiştir. Çapraz tablodaki temel bileşenler, kestirim esnasında gerçekte doğru iken doğru ve yanlış tahmin edilen veri sayısı ve gerçekte yanlış iken doğru ve yanlış tahmin edilen gözlem sayıları ile oluşturulmaktadır (Hasnain ve ark., 2020; Omary ve Mitenzi, 2010).

Tablo 3.7. Çapraz tablo

		Gerçek Durum	
		Pozitif	**Negatif**
Tahmin	**Pozitif**	Doğru Pozitif	Yanlış Pozitif
	Negatif	Yanlış Negatif	Doğru Negatif

Doğru Pozitif (DP): Gerçekte pozitif olan ve sınıflandırıcı tarafından da pozitif olarak sınıflandırılmış gözlemlerin sayısı.

Doğru Negatif (DN): Gerçekte negatif olan ve sınıflandırıcı tarafından da negatif sınıflandırılmış gözlemlerin sayısı.

Yanlış Pozitif (YP): Gerçekte negatif olan ve sınıflandırıcı tarafından pozitif sınıflandırılmış gözlemlerin sayısı.

Yanlış Negatif (YN): Gerçekte pozitif olan ve sınıflandırıcı tarafından negatif sınıflandırılmış gözlemlerin sayısı.

Doğruluk Oranı

Bu performans ölçütü geniş bir araştırmacı kitlesi tarafından kabul görmüş ve yaygın olarak kullanılan basit bir formül temeline dayanan bir ölçüttür. DO'nun hesaplanmasında veri setindeki gözlemlerden gerçek durumu için hesaplanan (DP + DN) tahmin sayısının toplam gözlem sayısına bölünmesi formülü kullanılır ve Eşitlik 3.16'da verilmiştir (Mirza ve ark., 2018);

$$DO = \frac{(DP)+(DN)}{(DP)+[YN]+[YP]+\{DN)} \quad \textbf{(3.16)}$$

Burada modelin performansının yüksek oluşu doğruluk oranının yüksek olması ile doğru orantılıdır. Yani DO ne kadar yüksek olursa modelin performansının da o derece başarılı olduğu söylenebilmektedir.

Hata Oranı

Hata oranı, doğruluk oranının tam tersidir yani bu performans ölçütünde veri setindeki gözlemlerden gerçek durumunun hatalı tahmin edilenlerinin sayısının (YN+YP), toplam gözlem sayısına bölümü formülü kullanılmaktadır ve Eşitlik 3.17'de verilmiştir;

$$HO = \frac{[YN]+[YP]}{(DP)+[YN]+[YP]+\{DN)} \quad \textbf{(3.17)}$$

Burada modelin performansının yüksek oluşu HO'nın yüksek olması ile ters orantılıdır. Yani HO ne kadar düşük olursa modelin performansının da o derece başarılı olduğu söylenebilmektedir (Bithas, 2019).

Kesinlik (Pozitif Kestirim Oranı)

Bu performans ölçütü veri setinde gerçekte hastalık olan gözlemlerden doğru tahmin edilenlerin (DP) sayısının, gerçekte hem hastalık olan hem de olmayan gözlemleri hastalık var olarak tahmin edilen toplam gözlem sayısına (DP+YP) oranı şeklinde formüle edilir ve Eşitlik 3.18'de verildi;

$$Kesinlik = \frac{DP}{DP+YP} \quad \textbf{(3.18)}$$

Burada modelin performansının yüksek oluşu kesinlik oranının yüksek olması ile doğru orantılıdır. Yani kesinlik oranı ne kadar yüksek olursa modelin performansının da o derece başarılı olduğu söylenebilmektedir (Ghorbanzadeh ve ark., 2019).

Duyarlılık

Veri setinde gerçekte hastalık var iken doğru tahmin edilen gözlem sayısının (DP), gerçekte hastalık var iken doğru ve yanlış tahmin edilen toplam gözlem sayısına (DP+YN) oranı ile hesaplanan performans ölçütüdür ve Eşitlik 3.19'da verildi (Başer ve ark., 2021);

$$Duyarlılık = \frac{DP}{DP+YN} \quad \textbf{(3.19)}$$

Burada modelin performansının yüksek oluşu duyarlılık oranının yüksek olması ile doğru orantılıdır. Yani duyarlılık oranı ne kadar yüksek olursa modelin performansının da o derece başarılı olduğu söylenebilmektedir. Bu ölçüt ile gerçekte hastalığı olanların yakalanma oranı da ortaya konulmaktadır.

Seçicilik

Bu performans ölçütü gerçekte hastalık yok iken doğru tahmin ile yok olarak kodlanan gözlemlerin (DN), gerçekte hastalık olmayan tüm gözlemlere (DN+YP) oranı olarak hesaplanmaktadır ve Eşitlik 3.20'de verildi (Pradhan ve Kim, 2020);

$$Seçicilik = \frac{DN}{DN+YP} \quad \textbf{(3.20)}$$

Burada modelin performansının yüksek oluşu seçicilik oranının yüksek olması ile doğru orantılıdır. Yani seçicilik oranı ne kadar yüksek olursa modelin performansının da o derece başarılı olduğu söylenebilmektedir. Bu ölçüt ile gerçekte hastalığı olmayanların yakalanma oranı da ortaya konulmaktadır.

F Ölçümü

F1 Skoru olarakta bilinen bu performans ölçütü, geri çağırma ölçütlerinin kesinlik performans ölçütü ile beraber temsil edildiği bir ölçüttür. F ölçümü Eşitlik 3.21'de verildi (Roth ve ark., 2020);

$$F1 = \frac{2*(Kesinlik*Duyarlılık)}{Kesinlik+Duyarlılık} \quad \textbf{(3.21)}$$

Bu skor değişim oranları hakkında bilgi sahibi olmayı hedefleyen bir ölçüttür. Harmonik ortalama mantığı ile formül oluşturulmuştur. Skor değeri 1'e yaklaştıkça ölçütün performansı da o derece yüksek olarak görülmektedir (Sabita ve ark., 2021).

Matthews Korelasyon Katsayısı

Bu ölçüt bir korelasyon katsayısıdır. Burada gerçek gözlem sonuçları ile tahmin edilen sonuçlar arasında kurulan bir korelasyon söz konusudur ve Eşitlik 3.22'de verildi;

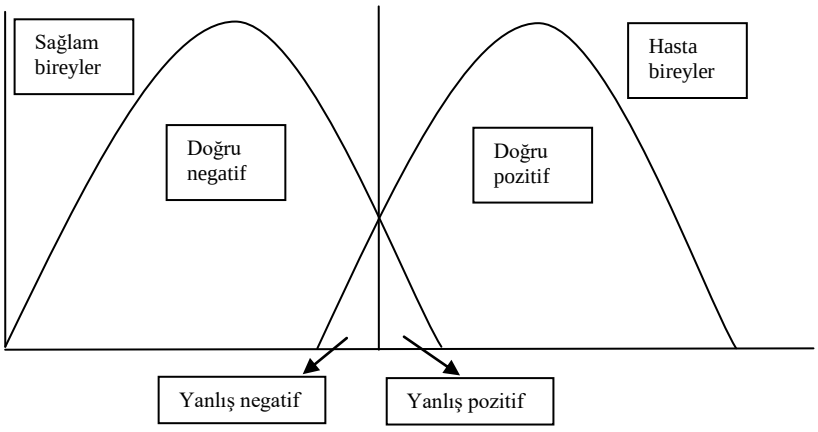
$$MKK = \frac{(DP*DN)-(YP*YN)}{\sqrt{(DP+YP)(DP+YN)(DN+YP)(DN+YN)}} \quad \textbf{(3.22)}$$

Bu katsayı 1'e yaklaştıkça model performansının yüksek olduğu görülmektedir (Gündüz, 2019).

İşlem Karakteristik Eğrisi Analizi

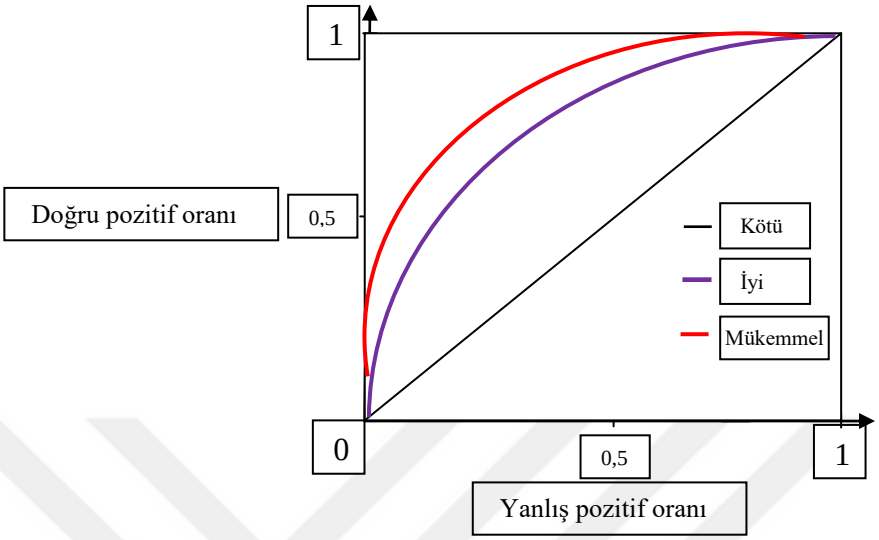
Klinik çalışmalarda, farklı tanı yöntemlerinden ve çeşitli testlerin sonuçlarından faydalanılarak hasta ve sağlıklı bireyleri birbirinden ayırdetmek gerekmektedir. Bu durumda bir testin, hasta bireyleri sağlıklılardan hangi düzeyde ve doğrulukta ayırt edebildiğinin bilinmesi çok önemlidir. Bu karar verme sürecinde kullanılan önemli istatistiksel yöntemlerden birisi de İşlem Karakteristik Eğrisi (Receiver Operating Characteristic) analizidir. ROC analizinin uygulaması ROC eğrisi olarak adlandırılan grafiksel çizimin oluşturulması ile gerçekleşmektedir.

ROC eğrisi, ikili sınıflama sistemlerinin performanslarını farklı eşik değerlere göre grafik şeklinde gösterim olarak bilinmektedir. ROC eğrisi doğru pozitif oranına karşılık yanlış pozitif oranı için farklı eşik değerlerine göre biçimlenmektedir (Şekil 3.6) (Flach, 2016).



Şekil 3.6. ROC analizinde hasta ve sağlam bireylerin test değerlerinin dağılımı

Şekil 3.7. ile ROC eğrisi çizim gösterildi. Tek bir sınıflayıcı için performans hesaplanmasının yanı sıra, ROC ile farklı sınıflayıcıları birbirleri ile karşılaştırma imkânı oluşmaktadır.



Şekil 3.7. Performanslarına göre ROC eğrileri

Optimal sınıflayıcı (0,1) noktasında olmaktadır (%100 doğru pozitif, %0 yanlış pozitif), tüm sınıflamaların yanlış olduğu nokta ise %0 doğru pozitif, %100 yanlış pozitif olan (1,0)'dır. Grafikte eğrinin sol üst köşeye yakın olması sınıflayıcının performansının iyi olduğunun göstergesidir. Köşegen üzerindeki (0,0) – (1,1) doğru ise şans doğrusu olarak adlandırılmaktadır. Şans doğrusunda durum %50 doğru pozitif ve %50 yanlış pozitif şeklindedir (Marsland, 2015).

Bu analizin kullanılmasının önemli avantajlarından birisi de duyarlılık ve seçicilik ölçütleri kullanması ve sınıflar arasındaki dengesizliğe karşı duyarsız olmasıdır.

4. BULGULAR

4.1. Çapraz Tabloya İlişkin Bulgular

Veri seti incelendiğinde 9.136 kişinin 4.959'unda (%54,28) KBY hastalığı bulunmakta iken 4.177'sinde (%45,72) bu hastalık bulunmamaktaydı. Veri seti üzerinde tek sınıf, Naive Bayes, lojistik regresyon, karar ağaçları, rastgele orman, en yakın k-komşu ve destek vektör makineleri algoritmaları ile modeller oluşturuldu. Makine öğrenimi algoritmaları için oluşturulan çapraz tablo, Tablo 4.1'de verildi.

Tüm veri setinin kullanıldığı algoritmalar çalıştırılarak çapraz tablolar oluşturuldu, 9.136 kişinin DP, YN, YP ve DN oranları Tablo 4.1'de gösterildi.

Tablo 4.1. Tüm veri seti ile oluşturulan çapraz tablo

Algoritmalar	DP		YN		YP		DN	
	Sayı	%	Sayı	%	Sayı	%	Sayı	%
Tek Sınıf	3113	34,07	1846	20,21	543	5,94	3634	39,78
Naive Bayes	3605	39,46	1354	14,82	839	9,18	3338	36,54
Lojistik Regresyon	3691	40,40	1268	13,88	796	8,71	3381	37,01
Karar Ağaçları	3870	42,36	1089	11,92	714	7,82	3463	37,90
Rastgele Orman	4765	52,16	194	2,12	68	0,74	4109	44,98
En Yakın K-Komşu	4833	52,90	126	1,38	136	1,49	4041	44,23
Destek Vektör Makineleri	3258	35,66	1701	18,62	543	5,94	3634	39,78

Tablo 4.1 incelendiğinde gerçekte 4.959 KBY hastalığı bulunan kişiden, modelde 4.833'nü doğru tahmin eden algoritma EYKKA oldu ve tahmin yüzdesi %97,46 olurken tüm kişiler içindeki DP tahmin oranı %52,90 olarak gerçekleşti. KBY hastalığı bulunmayan 4.177 kişi için ise en yüksek doğru tahmini 4.109 kişi ile ROA gerçekleştirdi ve bu algoritmanın KBY hastalığı bulunmayanlar arasındaki doğru tahmin oranı %98,37 olarak gerçekleşti. Tüm kişiler arasında DN oranı ise %44,98 oldu. Burada veri setinin tamamı kullanıldığından veri setine eklenecek olan yeni verilerde nasıl birçapraz tabloi oluşacağı konusu belirsizliğini korumaktadır.

Bu belirsizliğe cevap aramak için veri seti sırasıyla rastgele olarak %10, %20, %30, %40 ve %50 oranında test veri setine ayrıldı. Kalanlar ise eğitim veri seti olarak modelde kullanıldı. Eğitim verisi kullanılarak oluşturulan modeller ile test verisi üzerinden yapılan tahminlere ilişkin çapraz tablolar oluşturuldu.

Eğitim veri seti %90 alınarak oluşturulan tahmin modelleri, test veri setinin %10 olarak kullanıldığı algoritmalar için çalıştırılarak çapraz tablolar oluşturuldu, 914 (%10) kişinin DP, YN, YP ve DN oranları Tablo 4.2'de gösterildi.

Tablo 4.2. Çapraz tablo (Eğitim verisi %90, test verisi %10)

Algoritmalar	DP		YN		YP		DN	
	Sayı	%	Sayı	%	Sayı	%	Sayı	%
Tek Sınıf	304	33,26	191	20,90	60	6,56	359	39,28
Naive Bayes	362	39,61	133	14,55	90	9,85	329	35,99
Lojistik Regresyon	366	40,04	129	14,11	84	9,19	335	36,66
Karar Ağaçları	375	41,02	120	13,13	98	10,72	321	35,13
Rastgele Orman	389	42,56	106	11,60	128	14,00	291	31,84
En Yakın K-Komşu	354	38,73	141	15,43	148	16,19	271	29,65
Destek Vektör Makineleri	335	36,65	160	17,51	65	7,11	354	38,73

Tablo 4.2 incelendiğinde test veri setinde bulunan 914 kişiden 495'nde (%54,16) KBY hastalığı bulunuyorken 419'unda (%45,84) KBY hastalığı bulunmamaktadır. Test veri seti üzerinde çalıştırılan algoritmalar sonucunda en yüksek doğru tahmini KBY hastalığı bulunan 495 hasta için 389'nu doğru tahmin ederek KBY hastalığı olanlar arasındaki doğru tahmin oranı %78,59 olan ROA'sı gerçekleştirdi. KBY hastalığı bulunmayan 419 hasta için en yüksek doğru tahmini 359 kişi ve %85,68 oran ile TSA elde etti.

Eğitim veri seti %80 alınarak oluşturulan tahmin modelleri, test veri setinin %20 olarak kullanıldığı algoritmalar için çalıştırılarak çapraz tablo oluşturuldu, 1.827 (%20) kişinin DP, YN, YP ve DN oranları Tablo 4.3'de gösterildi.

Tablo 4.3. Çapraz tablo (Eğitim verisi %80, test verisi %20)

Algoritmalar	DP		YN		YP		DN	
	Sayı	%	Sayı	%	Sayı	%	Sayı	%
Tek Sınıf	616	33,71	388	21,24	103	5,64	720	39,41
Naive Bayes	726	39,74	278	15,22	158	8,65	665	36,39
Lojistik Regresyon	736	40,28	268	14,67	148	8,10	675	36,95
Karar Ağaçları	744	40,72	260	14,23	166	9,09	657	35,96
Rastgele Orman	778	42,58	226	12,37	226	12,37	597	32,68
En Yakın K-Komşu	710	38,86	294	16,09	276	15,11	547	29,94
Destek Vektör Makineleri	669	36,62	335	18,34	110	6,02	713	39,02

Tablo 4.3 incelendiğinde test veri setinde gerçekte hasta olan 1.004 (%54,95) kişi bulunurken gerçekte hastalığı bulunmayan 823 (%45,05) kişi bulunmaktadır. Test verisine eğitim verisi üzerinde oluşturulan tahmin modelleri uygulandığında, KBY hastalığı bulunan 1.004 kişi arasında en yüksek doğru tahmini 778 (%77,49) kişi ile ROA'sı gerçekleştirdi. KBY hastalığı bulunmayan 823 hasta için ise en yüksek doğru tahmini 720 (%87,48) ile TSA elde etti.

Eğitim veri seti %70 alınarak oluşturulan tahmin modelleri, test veri setinin %30 olarak kullanıldığı algoritmalar için çalıştırılarak çapraz tablo oluşturuldu, 2.741 (%30) kişinin DP, YN, YP ve DN oranları Tablo 4.4'de gösterildi.

Tablo 4.4. Çapraz tablo (Eğitim verisi %70, test verisi %30)

Algoritmalar	**DP**		**YN**		**YP**		**DN**	
	Sayı	**%**	**Sayı**	**%**	**Sayı**	**%**	**Sayı**	**%**
Tek Sınıf	917	33,45	569	20,76	154	5,62	1101	40,17
Naive Bayes	1070	39,04	416	15,18	250	9,12	1005	36,66
Lojistik Regresyon	1083	39,51	403	14,70	220	8,03	1035	37,76
Karar Ağaçları	1109	40,46	377	13,75	260	9,49	995	36,30
Rastgele Orman	1147	41,85	339	12,37	358	13,06	897	32,72
En Yakın K-Komşu	1036	37,80	450	16,42	429	15,65	826	30,13
Destek Vektör Makineleri	960	35,02	526	19,19	151	5,51	1104	40,28

Tablo 4.4 incelendiğinde tüm algoritmalar arasında verilerin tamamı için en yüksek başarı oranı ROA'nın %41,85 ile DP' üzerindeki tahmini olduğu görüldü. Test veri setinde bulunan 2.741 kişiden gerçekte 1.486 (%54,21) kişide KBY hastalığı bulunmaktayken 1.255 (%45,89) kişide KBY hastalığı bulunmamaktadır. Bu şartlar altında KBY hastalığı bulunan 1.486 kişiden 1.147'sini (%77,19) doğru tahmin eden ROA en başarılı tahmini yapan algortima oldu. KBY hastalığı bulunmayan 1.255 kişi için ise en yüksek doğru tahmini 1.104 (%87,97) ile DVMA elde etti.

Eğitim veri seti %60 alınarak oluşturulan tahmin modelleri, test veri setinin %40 olarak kullanıldığı algoritmalar için çalıştırılarak çapraz tablo oluşturuldu, 3.654 (%40) kişinin DP, YN, YP ve DN oranları Tablo 4.5'de gösterildi.

Tablo 4.5. Çapraz tablo (Eğitim verisi %60, test verisi %40)

Algoritmalar	DP		YN		YP		DN	
	Sayı	%	Sayı	%	Sayı	%	Sayı	%
Tek Sınıf	1232	33,72	753	20,61	211	5,77	1458	39,90
Naive Bayes	1440	39,41	545	14,92	330	9,03	1339	36,64
Lojistik Regresyon	1468	40,18	517	14,15	303	8,29	1366	37,38
Karar Ağaçları	1484	40,61	501	13,71	324	8,87	1345	36,81
Rastgele Orman	1528	41,82	457	12,51	474	12,97	1195	32,70
En Yakın K-Komşu	1380	37,77	605	16,56	574	15,71	1095	29,96
Destek Vektör Makineleri	1453	39,77	532	14,55	298	8,16	1371	37,52

Tablo 4.5 incelendiğinde test veri seti içerisinde gerçekte KBY hastalığı bulunan kişi sayısı 1.985 (%54,32), KBY hastalığı bulunmayan kişi sayısı ise 1.669 (%45,68) oldu. Eğitim veri seti üzerinde oluşturulan tüm tahmin modelleri test veri seti üzerinde çalıştırıldığında KBY hastalığı bulunan 1.985 hasta için en yüksek başarıyı 1.528'ni (%76,98) doğru tahmin eden ROA olurken, KBY hastalığı bulunmayan 1.669 hasta için en yüksek doğru tahmini 1.458 (%87,36) ile TSA elde etti.

Eğitim veri seti %50 alınarak oluşturulan tahmin modelleri, test veri setinin %40 olarak kullanıldığı algoritmalar için çalıştırılarak çapraz tablo oluşturuldu, 4.568 (%50) kişinin DP, YN, YP ve DN oranları Tablo 4.6'da gösterildi.

Tablo 4.6. Çapraz tablo (Eğitim verisi %50, test verisi %50)

Algoritmalar	DP		YN		YP		DN	
	Sayı	%	Sayı	%	Sayı	%	Sayı	%
Tek Sınıf	1539	33,69	941	20,60	270	5,91	1818	39,80
Naive Bayes	1789	39,16	691	15,13	422	9,24	1666	36,47
Lojistik Regresyon	1827	39,99	653	14,30	392	8,58	1696	37,13
Karar Ağaçları	1862	40,76	618	13,53	435	9,52	1653	36,19
Rastgele Orman	1890	41,38	590	12,92	604	13,22	1484	32,48
En Yakın K-Komşu	1714	37,52	766	16,77	728	15,94	1360	29,77
Destek Vektör Makineleri	1689	36,97	791	17,32	292	6,39	1796	39,32

Tablo 4.6 incelendiğinde test verisinde gerçekte KBY hastakığı bulunan 2.480 (%54,29) kişi, KBY hastalığı bulunmayan 2.088 (%45,71) kişi vardır. Eğitim veri seti üzerinde oluşturulan tüm tahmin modelleri test verisi üzerinde çalıştırıldığında en yüksek doğru tahmin sayılarını KBY hastalığı bulunan 2.480 hasta için 1.890'nı

(%76,21) doğru tahmin eden ROA olurken, KBY hastalığı bulunmayan 2.088 hasta için en yüksek doğru tahmini 1.818 (%87,07) ile TSA tahmin etti.

4.2. Performans Ölçütlerine İlişkin Bulgular

Veri seti üzerinde yapılan tahmin modeli çalışmalarında modelin performansını ölçmek için doğruluk oranı, hata oranı, kesinlik, duyarlılık, seçicilik, F ölçümü ve Matthews korelasyon katsayısı ölçütleri hesaplandı.

Tüm verinin eğitim kullanıldığı modeller için hesaplanan bu performans ölçütleri Tablo 4.7'de verildi.

Tablo 4.7. Tüm veri seti kullanıldığında modellere ilişkin performans ölçütleri

Tüm Veri Kullanıldığında							
Algoritmalar	**Doğruluk oranı**	**Hata oranı**	**Kesinlik**	**Duyarlılık**	**Seçicilik**	**F Ölçümü**	**Matthews korelasyon katsayısı**
Tek Sınıf	0,739	0,261	0,851	0,628	0,870	0,736	0,506
Naive Bayes	0,760	0,240	0,811	0,727	0,799	0,760	0,524
Lojistik Regresyon	0,774	0,226	0,823	0,744	0,809	0,774	0,552
Karar Ağaçları	0,803	0,197	0,844	0,780	0,829	0,803	0,607
Rastgele Orman	0,971	0,029	0,986	0,961	0,984	0,971	0,943
En Yakın K-Komşu	0,971	0,029	0,973	0,975	0,967	0,971	0,942
Destek Vektör Makineleri	0,754	0,246	0,857	0,657	0,870	0,769	0,533

Tablo 4.7 incelendiğinde doğruluk oranı ve hata oranı açısından en yüksek performansı ROA ve EYKKA gösterdi. Kesinlik performans ölçütü incelendiğinde en yüksek performansı 0,986 ile ROA gerçekleştirdi. Duyarlılık için en yüksek performansı 0,975 ile EYKKA ortaya koyarken, seçicilik performans ölçütüne göre en yüksek performansı 0,984 ile ROA gerçekleştirdi. F ölçümü performans kriterine göre ROA ve EYKKA 0,971 ile en yüksek performansı gösteren iki algoritma olurken, Matthews korelasyon katsayısı açısından en yüksek performansı 0,943 ile ROA hesapladı.

Burada performans ölçütlerinin ortaya koyduğu sonuçların değerlendirilmesinde tüm veri setinin modelde kullanılmasından dolayı yeni gelecek hastalara ilişkin modellerin yapacağı tahminlerin ne olacağı konusunda belirsizlik bulunmaktadır. Bu belirsizliğin giderilmesi için çapraz tabloda yapıldığı gibi performans ölçütleri için de veri seti, %10, %20, %30, %40 ve %50 oranında test

veri setine ayrıldı. Kalanlar ise eğitim veri seti olarak modelde kullanıldı. Eğitim verisi kullanılarak oluşturulan modeller için performans ölçütleri ayrı ayrı hesaplandı.

Veri setinin %90'ı eğitim verisi olarak kullanıldı ve bu eğitim verisi üzerinden ilgili algoritmalara ilişkin tahmin modelleri oluşturuldu. Oluşturulan tahmin modellerinin test verisi (%10) için hesapladığı performans ölçütleri Tablo 4.8'de gösterildi.

Tablo 4.8. Performans ölçütleri (Eğitim verisi %90, test verisi %10)

%90 Eğitim Verisi - %10 Test Verisi							
Algoritmalar	**Doğruluk oranı**	**Hata oranı**	**Kesinlik**	**Duyarlılık**	**Seçicilik**	**F Ölçümü**	**Matthews korelasyon katsayısı**
Tek Sınıf	0,725	0,275	0,835	0,614	0,857	0,723	0,479
Naive Bayes	0,756	0,244	0,801	0,731	0,785	0,756	0,515
Lojistik Regresyon	0,767	0,233	0,813	0,739	0,800	0,767	0,537
Karar Ağaçları	0,761	0,239	0,793	0,758	0,766	0,762	0,522
Rastgele Orman	0,744	0,256	0,752	0,786	0,695	0,743	0,483
En Yakın K-Komşu	0,684	0,316	0,705	0,715	0,647	0,684	0,362
Destek Vektör Makineleri	0,754	0,246	0,838	0,677	0,845	0,753	0,524

Tablo 4.8 incelendiğinde doğruluk oranı, hata oranı, F ölçümü ve Matthews korelasyon katsayısı performans ölçütleri açısından en yüksek performansı LRA gösterdi. Kesinlik performans ölçütüne göre en yüksek performansı DVMA ortaya koyarken, seçicilik için en yüksek performansı TSA gösterdi. Duyarlılık performans ölçütü için ise ROA en iyi sonucu verdi.

Veri setinin %80'i eğitim verisi olarak kullanıldı ve bu eğitim verisi üzerinden ilgili algoritmalara ilişkin tahmin modelleri oluşturuldu. Oluşturulan tahmin modellerinin test verisi için hesapladığı performans ölçütleri Tablo 4.9'da gösterildi.

Tablo 4.9. Performans ölçütleri (Eğitim verisi %80, test verisi %20)

%80 Eğitim Verisi - %20 Test Verisi							
Algoritmalar	**Doğruluk oranı**	**Hata oranı**	**Kesinlik**	**Duyarlılık**	**Seçicilik**	**F Ölçümü**	**Matthews korelasyon katsayısı**
Tek Sınıf	0,731	0,269	0,857	0,614	0,875	0,729	0,497
Naive Bayes	0,761	0,239	0,821	0,723	0,808	0,762	0,529
Lojistik Regresyon	0,772	0,228	0,833	0,733	0,820	0,773	0,551
Karar Ağaçları	0,767	0,233	0,818	0,741	0,798	0,767	0,537
Rastgele Orman	0,753	0,247	0,775	0,775	0,725	0,753	0,500
En Yakın K-Komşu	0,688	0,312	0,720	0,707	0,665	0,688	0,371
Destek Vektör Makineleri	0,756	0,244	0,859	0,666	0,866	0,756	0,536

Tablo 4.9 incelendiğinde doğruluk oranı, hata oranı, F ölçümü ve Matthews korelasyon katsayısı performans ölçütleri açısından en yüksek performansı LRA gösterdi. Kesinlik performans ölçütüne göre en yüksek performansı DVMA ortaya koyarken, seçicilik için en yüksek performansı TSA gösterdi. Duyarlılık performans ölçütü için ise ROA en iyi sonucu verdi.

Veri setinin %70'i eğitim verisi olarak kullanıldı ve bu eğitim verisi üzerinden ilgili algoritmalara ilişkin tahmin modelleri oluşturuldu. Oluşturulan tahmin modellerinin test verisi için hesapladığı performans ölçütleri Tablo 4.10'da gösterildi.

Tablo 4.10. Performans ölçütleri (Eğitim verisi %70, test verisi %30)

%70 Eğitim Verisi - %30 Test Verisi							
Algoritmalar	**Doğruluk oranı**	**Hata oranı**	**Kesinlik**	**Duyarlılık**	**Seçicilik**	**F Ölçümü**	**Matthews korelasyon katsayısı**
Tek Sınıf	0,736	0,264	0,856	0,617	0,877	0,734	0,505
Naive Bayes	0,757	0,243	0,811	0,720	0,801	0,757	0,519
Lojistik Regresyon	0,773	0,227	0,831	0,729	0,825	0,773	0,552
Karar Ağaçları	0,768	0,232	0,810	0,746	0,793	0,768	0,537
Rastgele Orman	0,746	0,254	0,762	0,772	0,715	0,746	0,487
En Yakın K-Komşu	0,679	0,321	0,707	0,697	0,658	0,680	0,355
Destek Vektör Makineleri	0,753	0,247	0,864	0,646	0,880	0,751	0,533

Tablo 4.10 incelendiğinde doğruluk oranı, hata oranı, F ölçümü ve Matthews korelasyon katsayısı performans ölçütleri açısından en yüksek performansı LRA

gösterdi. Kesinlik ve seçicilik performans ölçütlerine göre en yüksek performansı DVMA ortaya koyarken, duyarlılık performans ölçütü için ise ROA en iyi sonucu verdi.

Veri setinin %60'ı eğitim verisi olarak kullanıldı ve bu eğitim verisi üzerinden ilgili algoritmalara ilişkin tahmin modelleri oluşturuldu. Oluşturulan tahmin modellerinin test verisi için hesapladığı performans ölçütleri Tablo 4.11'de gösterildi.

Tablo 4.11. Performans ölçütleri (Eğitim verisi %60, test verisi %40)

%60 Eğitim Verisi - %40 Test Verisi							
Algoritmalar	**Doğruluk oranı**	**Hata oranı**	**Kesinlik**	**Duyarlılık**	**Seçicilik**	**F Ölçümü**	**Matthews korelasyon katsayısı**
Tek Sınıf	0,736	0,264	0,854	0,621	0,874	0,734	0,504
Naive Bayes	0,761	0,239	0,814	0,725	0,802	0,761	0,526
Lojistik Regresyon	0,776	0,224	0,829	0,740	0,818	0,776	0,556
Karar Ağaçları	0,774	0,226	0,821	0,748	0,806	0,775	0,551
Rastgele Orman	0,745	0,255	0,763	0,770	0,716	0,745	0,486
En Yakın K-Komşu	0,677	0,323	0,706	0,695	0,656	0,678	0,351
Destek Vektör Makineleri	0,773	0,227	0,830	0,732	0,821	0,773	0,552

Tablo 4.11 incelendiğinde doğruluk oranı, hata oranı, F ölçümü ve Matthews korelasyon katsayısı performans ölçütleri açısından en yüksek performansı LRA gösterdi. Kesinlik ve seçicilik performans ölçütlerine göre en yüksek performansı TSA ortaya koyarken, duyarlılık performans ölçütü için ise ROA en iyi sonucu verdi.

Veri setinin %50'si eğitim verisi olarak kullanıldı ve bu eğitim verisi üzerinden ilgili algoritmalara ilişkin tahmin modelleri oluşturuldu. Oluşturulan tahmin modellerinin test verisi için hesapladığı performans ölçütleri Tablo 4.12'de gösterildi.

Tablo 4.12. Performans ölçütleri (Eğitim verisi %50, test verisi %50)

%50 Eğitim Verisi - %50 Test Verisi							
Algoritmalar	**Doğruluk oranı**	**Hata oranı**	**Kesinlik**	**Duyarlılık**	**Seçicilik**	**F Ölçümü**	**Matthews korelasyon katsayısı**
Tek Sınıf	0,735	0,265	0,851	0,621	0,871	0,733	0,500
Naive Bayes	0,756	0,244	0,809	0,721	0,798	0,757	0,518
Lojistik Regresyon	0,771	0,229	0,823	0,737	0,812	0,772	0,547
Karar Ağaçları	0,769	0,231	0,811	0,751	0,792	0,770	0,540
Rastgele Orman	0,739	0,261	0,758	0,762	0,711	0,739	0,473
En Yakın K-Komşu	0,673	0,327	0,702	0,691	0,651	0,673	0,342
Destek Vektör Makineleri	0,763	0,237	0,853	0,681	0,860	0,762	0,544

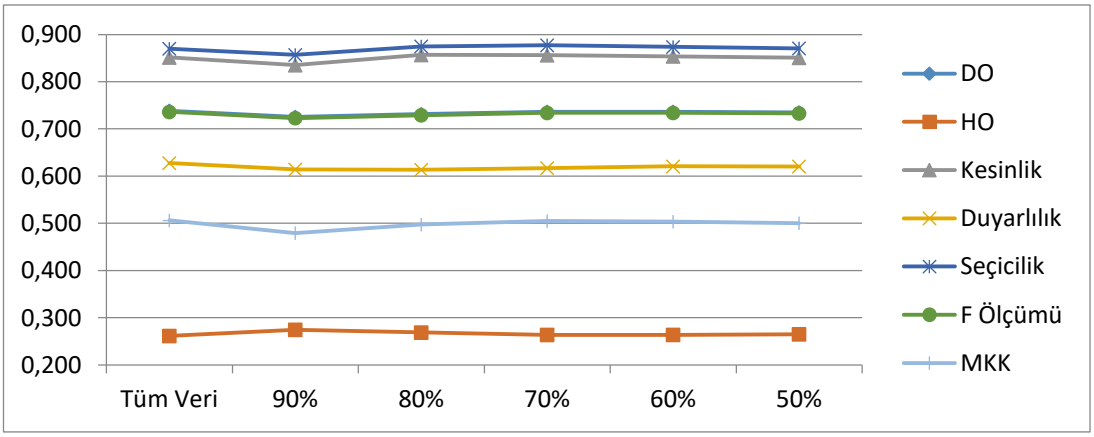
Tablo 4.11 incelendiğinde doğruluk oranı, hata oranı, F ölçümü ve Matthews korelasyon katsayısı performans ölçütleri açısından en yüksek performansı LRA gösterdi. Kesinlik performans ölçütüne göre en yüksek performansı DVMA ortaya koyarken, seçicilik için TSA, duyarlılık performans ölçütünde ise ROA gösterdi.

4.3. Algoritmaların Eğitim Verisi Büyüklüğüne Göre Performans Bulguları

Çalışmada kullanılan 7 farklı algoritmanın eğitim verisi büyüklüğü değiştikçe ortaya koyduğu performans sırasıyla incelendi. TSA'sı için eğitim verisi büyüklüğüne göre ortaya çıkan performans ölçütlerine ait sonuçlar Tablo 4.13 ve Şekil 4.1'de verildi.

Tablo 4.13. Tek sınıf algoritmasının eğitim verisi büyüklüğüne göre performans bilgileri

Tek Sınıf Algoritması						
	Tüm Veri	**%90**	**%80**	**%70**	**%60**	**%50**
Doğruluk oranı	0,739	0,725	0,731	0,736	0,736	0,735
Hata oranı	0,261	0,275	0,269	0,264	0,264	0,265
Kesinlik	0,851	0,835	0,857	0,856	0,854	0,851
Duyarlılık	0,628	0,614	0,614	0,617	0,621	0,621
Seçicilik	0,870	0,857	0,875	0,877	0,874	0,871
F Ölçümü	0,736	0,723	0,729	0,734	0,734	0,733
Matthews korelasyon katsayısı	0,506	0,479	0,497	0,505	0,504	0,500



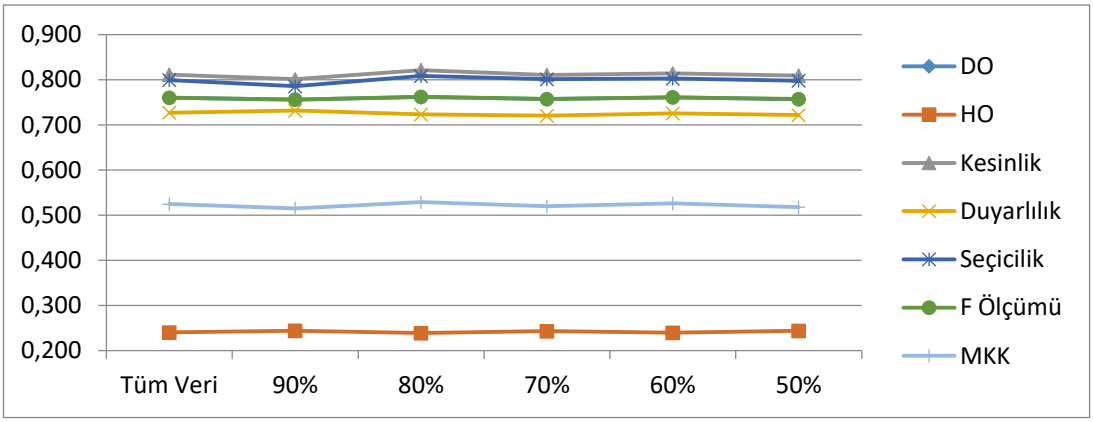
Şekil 4.1. Tek sınıf algoritmasının eğitim verisi büyüklüğüne göre performansı

Gerçekleştirilen analizler sonucunda Tablo 4.13 ve Şekil 4.1 incelendiğinde TSA'nın eğitim verisi büyüklüğünde oluşan değişiklikten fazla etkilenmediği ancak performans ölçütleri açısından oldukça farklılıklar gösterdiği ortaya çıkartıldı. En yüksek performansın seçicilik performans ölçütünde 0,877 ile Eğitim veri setinin %70, test veri setinin ise %30 olarak alındığı ayrılmada görüldü. Ayrıca bu modelin keskinlik ve seçicilik performans ölçütlerinin oldukça yüksek seviyelerde olduğu görüldü.

NBA için eğitim verisi büyüklüğüne göre ortaya çıkan performans ölçütlerine ait sonuçlar Tablo 4.14 ve Şekil 4.2'de verildi.

Tablo 4.14. Naive Bayes algoritmasının eğitim verisi büyüklüğüne göre performans bilgileri

	Naive Bayes Algoritması					
	Tüm Veri	**%90**	**%80**	**%70**	**%60**	**%50**
Doğruluk oranı	0,760	0,756	0,761	0,757	0,761	0,756
Hata oranı	0,240	0,244	0,239	0,243	0,239	0,244
Kesinlik	0,811	0,801	0,821	0,811	0,814	0,809
Duyarlılık	0,727	0,731	0,723	0,720	0,725	0,721
Seçicilik	0,799	0,785	0,808	0,801	0,802	0,798
F Ölçümü	0,760	0,756	0,762	0,757	0,761	0,757
Matthews korelasyon katsayısı	0,524	0,515	0,529	0,519	0,526	0,518



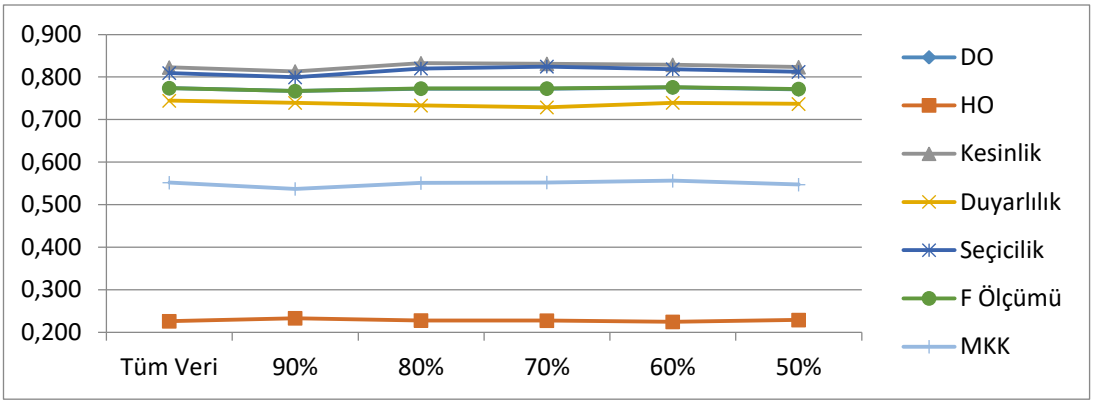
Şekil 4.2. Naive Bayes algoritmasının eğitim verisi büyüklüğüne göre performansı

Tablo 4.14 ve Şekil 4.2 incelendiğinde NBA'nın eğitim veri setinin büyüklüğünün değişiminden çok fazla etkilenmediği görüldü. DO, kesinlik, duyarlılık, seçicilik ve F ölçümü %70'in üzerinde bir değer gösterdi. Matthews korelasyon katsayısı performans ölçütü ise %50'nin üzerinde bir değer ile sonuçlandı.

LRA için eğitim verisi büyüklüğüne göre ortaya çıkan performans ölçütlerine ait sonuçlar Tablo 4.15 ve Şekil 4.3'de verildi.

Tablo 4.15. Lojistik regresyon algoritmasının eğitim verisi büyüklüğüne göre performans bilgileri

Lojistik Regresyon Algoritması						
	Tüm Veri	**%90**	**%80**	**%70**	**%60**	**%50**
Doğruluk oranı	0,774	0,767	0,772	0,773	0,776	0,771
Hata oranı	0,226	0,233	0,228	0,227	0,224	0,229
Kesinlik	0,823	0,813	0,833	0,831	0,829	0,823
Duyarlılık	0,744	0,739	0,733	0,729	0,740	0,737
Seçicilik	0,809	0,800	0,820	0,825	0,818	0,812
F Ölçümü	0,774	0,767	0,773	0,773	0,776	0,772
Matthews korelasyon katsayısı	0,552	0,537	0,551	0,552	0,556	0,547



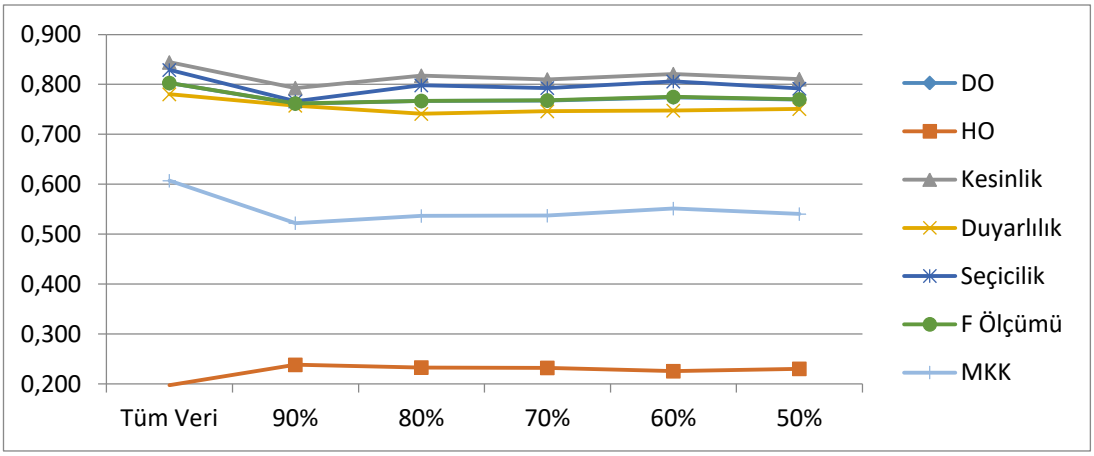
Şekil 4.3. Lojistik regresyon algoritmasının eğitim verisi büyüklüğüne göre performansı

Tablo 4.15 ve Şekil 4.3 incelendiğinde LRA için kesinlik ve seçicilik performans ölçütleri tüm eğitim veri seti büyüklüklerinde %80'nin üzerinde gerçekleşti. Doğruluk oranı, duyarlılık ve F ölçümü performans ölçütleri ise daha düşük bir başarı elde etti ve tüm eğitim veri seti büyüklüklerinde %70'in üzerinde ancak %80'nin altında bir sonuç ortaya koydu.

KAA için eğitim verisi büyüklüğüne göre ortaya çıkan performans ölçütlerine ait sonuçlar Tablo 4.16 ve Şekil 4.4'de verildi.

Tablo 4.16. Karar ağaçları algoritmasının eğitim verisi büyüklüğüne göre performans bilgileri

	Karar Ağaçları Algoritması					
	Tüm Veri	**%90**	**%80**	**%70**	**%60**	**%50**
Doğruluk oranı	0,803	0,761	0,767	0,768	0,774	0,769
Hata oranı	0,197	0,239	0,233	0,232	0,226	0,231
Kesinlik	0,844	0,793	0,818	0,810	0,821	0,811
Duyarlılık	0,780	0,758	0,741	0,746	0,748	0,751
Seçicilik	0,829	0,766	0,798	0,793	0,806	0,792
F Ölçümü	0,803	0,762	0,767	0,768	0,775	0,770
Matthews korelasyon katsayısı	0,607	0,522	0,537	0,537	0,551	0,540



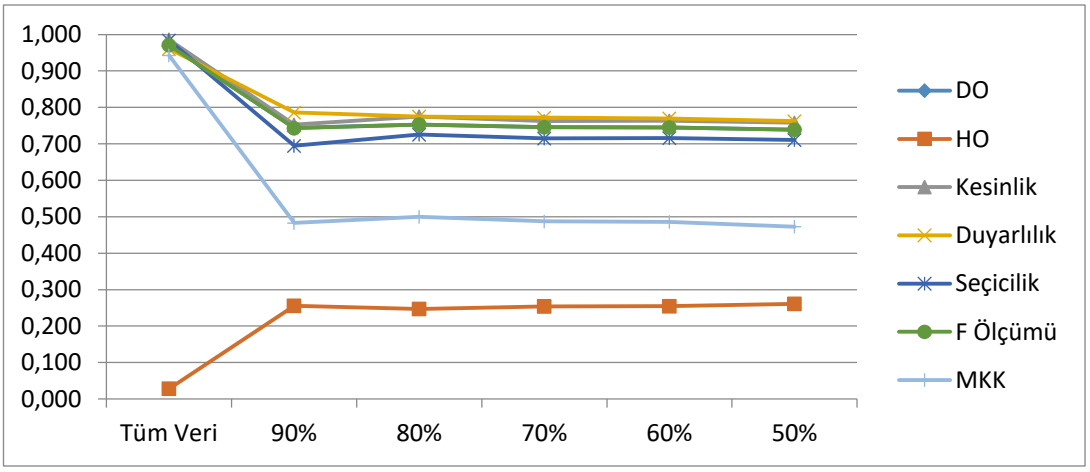
Şekil 4.4. Karar ağaçları algoritmasının eğitim verisi büyüklüğüne göre performansı

Tablo 4.16 ve Şekil 4.4 incelendiğinde KAA için doğruluk oranı ve F ölçümü performans ölçütleri tüm veri setinin kullanıldığı modelde %80'in üzerinde bir başarı sağlarken, eğitim veri setinin tüm veri setinden belli bir oranda çekildikten sonra kalan test veri seti üzerinde oluşturulan modellerde ise %70'in üzerinde ancak %80'nin altında bir performans gerçekleştirdi. Kesinlik performans ölçütü sadece eğitim veri setinin %90 olarak alındığı modelde %80'nin altında kaldı. Duyarlılık performans ölçütü ise tüm eğitim veri seti büyüklükleri için %70'in üzerinde ancak %80'nin altında bir performans gerçekleştirdi. Matthews korelasyon katsayısı performans ölçütünde tüm veri setinin kullanıldığı modelde %60'ın üzerinde bir performans oluşurken diğer tüm büyüklükteki eğitim veri setleri için oluşturulan modellerde ise %60'ın altında ancak %50'nin üzerinde bir performans gerçekleştirdi.

ROA için eğitim verisi büyüklüğüne göre ortaya çıkan performans ölçütlerine ait sonuçlar Tablo 4.17 ve Şekil 4.5'de verildi.

Tablo 4.17. Rastgele orman algoritmasının eğitim verisi büyüklüğüne göre performans bilgileri

	Rastgele Orman Algoritması					
	Tüm Veri	**%90**	**%80**	**%70**	**%60**	**%50**
Doğruluk oranı	0,971	0,744	0,753	0,746	0,745	0,739
Hata oranı	0,029	0,256	0,247	0,254	0,255	0,261
Kesinlik	0,986	0,752	0,775	0,762	0,763	0,758
Duyarlılık	0,961	0,786	0,775	0,772	0,770	0,762
Seçicilik	0,984	0,695	0,725	0,715	0,716	0,711
F Ölçümü	0,971	0,743	0,753	0,746	0,745	0,739
Matthews korelasyon katsayısı	0,943	0,483	0,500	0,487	0,486	0,473



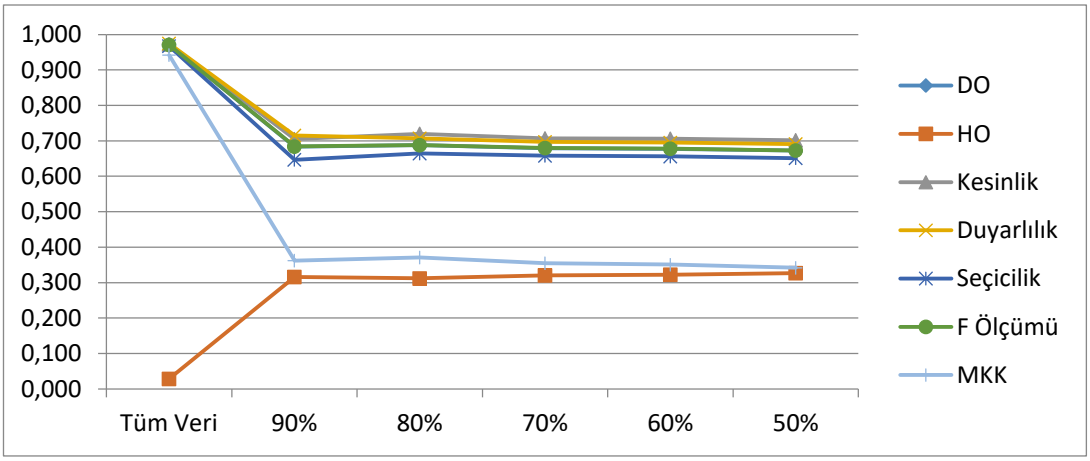
Şekil 4.5. Rastgele orman algoritmasının eğitim verisi büyüklüğüne göre performansı

Tablo 4.17 ve Şekil 4.5 incelendiğinde ROA için tüm veri setinin kullanıldığı modelde performans ölçütlerinin tümünde %90'ın üzerinde önemli bir başarı gösterdi. Ancak tüm veri setinin bir kısmının eğitim veri seti olarak ayrılıp kalanının test verisi olduğu ve test verisi üzerinde yapılan modellemelerde tüm performans ölçütlerinde önemli düşüşler yaşandı. DO eğitim veri setinin %100 olduğu modelde %97,1 gibi yüksek bir başarı elde etti ancak eğitim veri setinin %80 alındığı modelde test verisi üzerinde yapılan çalışmada bu başarı %75,3' e kadar geriledi. Bu durum diğer performans ölçütleri içinde benzer şekilde gerçekleşti.

EYKKA için eğitim verisi büyüklüğüne göre ortaya çıkan performans ölçütlerine ait sonuçlar Tablo 4.18 ve Şekil 4.6'da verildi.

Tablo 4.18. En yakın k-komşu algoritmasının eğitim verisi büyüklüğüne göre performans bilgileri

En Yakın K-Komşu Algoritması						
	Tüm Veri	**%90**	**%80**	**%70**	**%60**	**%50**
Doğruluk oranı	0,971	0,684	0,688	0,679	0,677	0,673
Hata oranı	0,029	0,316	0,312	0,321	0,323	0,327
Kesinlik	0,973	0,705	0,720	0,707	0,706	0,702
Duyarlılık	0,975	0,715	0,707	0,697	0,695	0,691
Seçicilik	0,967	0,647	0,665	0,658	0,656	0,651
F Ölçümü	0,971	0,684	0,688	0,680	0,678	0,673
Matthews korelasyon katsayısı	0,942	0,362	0,371	0,355	0,351	0,342



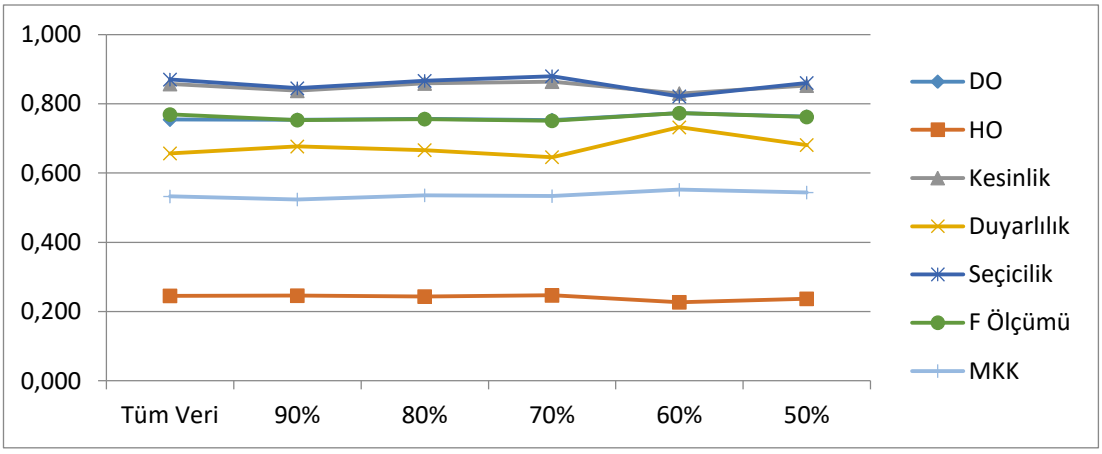
Şekil 4.6. En yakın k-komşu algoritmasının eğitim verisi büyüklüğüne göre performansı

Tablo 4.18 ve Şekil 4.6 incelendiğinde bu algoritmanın ROA'na benzer bir yapı sergilediği görüldü. Tüm veri setinin kullanıldığı model ile diğer modeller arasındaki başarı oranı arasındaki fark bu algoritmada çok daha fazla olarak gerçekleşti. Tüm veri setinin kullanıldığı modelde tüm performans ölçütleri %90'ın üzerindeyken eğitim ve test veri setine ayrılan modellerde ise tüm performans ölçütleri için başarı seviyeleri %70 hatta %60 seviyelerine kadar geriledi. Matthews korelasyon katsayısı performans ölçütü %40'ın da altına düştü.

DVMA için eğitim verisi büyüklüğüne göre ortaya çıkan performans ölçütlerine ait sonuçlar Tablo 4.19 ve Şekil 4.7'de verildi.

Tablo 4.19. Destek vektör makineleri algoritmasının eğitim verisi büyüklüğüne göre performans bilgileri

	Destek Vektör Makineleri Algoritması					
	Tüm Veri	**%90**	**%80**	**%70**	**%60**	**%50**
Doğruluk oranı	0,754	0,754	0,756	0,753	0,773	0,763
Hata oranı	0,246	0,246	0,244	0,247	0,227	0,237
Kesinlik	0,857	0,838	0,859	0,864	0,830	0,853
Duyarlılık	0,657	0,677	0,666	0,646	0,732	0,681
Seçicilik	0,870	0,845	0,866	0,880	0,821	0,860
F Ölçümü	0,769	0,753	0,756	0,751	0,773	0,762
Matthews korelasyon katsayısı	0,533	0,524	0,536	0,533	0,552	0,544

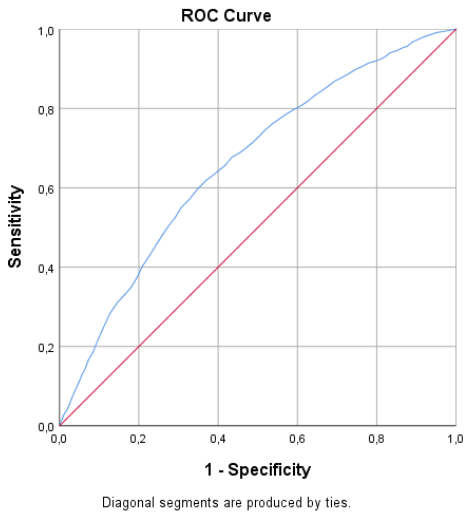


Şekil 4.7. Destek vektör makineleri algoritmasının eğitim verisi büyüklüğüne göre performansı

Tablo 4.19 ve Şekil 4.7 incelendiğinde bu algoritmanın başarı oranlarının eğitim veri setinin büyüklüğünün değişiminden çok fazla etkilenmediği görüldü. Tüm performans ölçütleri için eğitim veri setinin %70 ve %80 olarak alındığı modellerde en yüksek başarı oranları elde edildi. Bu algoritma kesinlik ve seçicilik performans ölçütleri açısında oldukça başarılı sonuçlar ortaya koydu.

4.4. İşlem Karakteristik Eğrisi Analizi ile Elde Edilen Bulgular

ROC analizi hedef değişkenin kategorik olduğu ve bağımsız değişkenlerin sayısal veriler olduğu durumlarda kullanılan performans ölçütlerinden olduğundan dolayı veri setinde tek nicel değişken olan ve 18-101 arasında değer alan yaş değişkeninin KBY hastalığına olan etkisi ROC analizi ile ölçüldü. ROC eğrisi Şekil 4.8'de verildi.



Şekil 4.8. Hasta yaşı ve KBY hastalığına ilişkin ROC eğrisi

ROC analizi sonuçları tablo 4.20'de verilmiş olup bu değişken için kesim noktası (cut off) 54,5 olarak saptanmıştır.

Tablo 4.20. ROC analizi sonuç bilgisi

	AUC	**Standart hata**	**p değeri**	**Güven aralığı**	
Test Sonuç Değerleri	0,661	0,006	< 0,001	0,648	0,673

5. TARTIŞMA

Makine öğrenimi yöntemleri ile bilinmeyen bir gözlemin ait olduğu konumun diğer bazı özellikleri değerlendirilerek hangi sınıfa ait olduğu tahmin edilir. Özellikle denetimli makine öğrenimi yöntemleri elde bulunan ipuçlarını en verimli şekilde değerlendirerek bu tahminleri en yüksek düzeyde yapmaktadır (Varshney, 2022). Bu denetimli makine öğrenimi yöntemlerinden TSA, NBA, LRA, KAA, ROA, EYKKA ve DVMA kullanılarak sonuçlar elde edildi. KAA'da kreatinin değişkeni kök değişken olarak alındı. ROA'da yüz iterasyon kullanıldı. EYKKA'da k=1 olarak alındı ve uzaklık ölçütü olarak öklit mesafesi kullanıldı. DVMA'da kategori ayrımı belirlenirken polinom ayrımı esas alındı. Bu algoritmalar ile gerçek hastalara ait veri setinin tamamının veri seti olarak alındığı ve sırasıyla %90 eğitim veri seti %10 test veri seti, %80 eğitim veri seti %20 test veri seti, %70 eğitim veri seti %30 test veri seti, %60 eğitim veri seti %40 test veri seti, %50 eğitim veri seti %50 test veri seti olarak alındığı 6 farklı durum çerçevesinde kullanılan algoritmalar ile tahmin modelleri oluşturuldu. Hangi veri setinin en yüksek performans ile tahmin oluşturduğunu bulmak için tüm veri grupları doğruluk oranı, hata oranı, kesinlik, duyarlılık, seçicilik, F ölçümü ve Matthews korelasyon katsayısı performans ölçütleri hesaplanarak karşılaştırıldı. Eğitim veri setinin hangi büyüklükte alınması gerektiği makine öğreniminde karşılaşılan problemlerden birisi olduğundan bu soruya cevap arandı.

Çalışmada öncelikli olarak gerçek hastalara ait olan veri setinin tamamının kullanıldığı, eğitim ve test veri setlerine yer verilmeyen tahmin modelleri ile çalışılan algoritmalar oluşturuldu ve 7 farklı performans ölçütünün sonuçları ortaya çıktı. TSA için yapılan tahmin modelinde "Kreatinin" değerleri tahmin edici sınıf olarak seçildi. Tüm veri seti modelde için kullanıldığından buradaki modellerin başarısını ölçecek bir test verisi kullanılmadı. Sadece tüm veri seti üzerinde, algoritmalar mevcut durumuyla tahminler oluşturdu ve bu tahminler için performans ölçütleri hesaplandı. Bu durum veri setine yeni gözlemler eklendiğinde sonucun başarısını ölçmek için bir belirsizlik oluşturduğundan ve denetimli makine öğrenimi algoritmalarının temelinde veriyi, eğitim verisi ve test verisi olmak üzere ikiye ayrılması gerektiğinden çalışmaya bu yöntemler ile devam edildi. Test verisi olmadan yapılan bu çalışmada algoritmaların performansları arasında da oldukça farklı sonuçlar elde edildi. ROA, doğruluk oranında %97,1, kesinlik için %98,6, seçicilik için %98,4, F ölçüm değeri

için %97,1 ve Matthews korelasyon katsayısı için %94,3 ile en iyi performansı gösterdi. EYKKA ise doğruluk oranı için %97,1 ve F ölçüm değerinde %97,1 ile diğerlerine göre yüksek düzeyde sonuçlar üretti. KBY hastalığı bulunan hastaların tahmininde EYKKA iyi bir sonuç gösterirken, KBY hastalığı bulunmayan hastaların tahmininde ise ROA oldukça başarılı bir performans gösterdi.

Test veri seti kullanılmadan gerçekleştirilen tahmin modellerinin performansını görmek için %50 ile %90 arasında değişen farklı büyüklüklerde eğitim ve test veri setleri oluşturularak sonuçlar karşılaştırıldı.

Veri setinin %90 eğitim veri seti, %10 test veri seti olarak alındığında tüm algoritmalar için tüm performans ölçütlerinin başarı yüzde düzeylerinde düşüşler gözlendi. TSA için doğruluk oranının %73,9'dan %72,5'e, kesinliğin %85,1'den %83,5'e, duyarlılığın %62,8'den %61,4'e, seçiciliğin %87'den %85,7'e, F ölçüm değerinin %73,6'dan %72,3'e ve Matthews korelasyon katsayısının %50,6'dan %47,9'a düştüğü gözlendi. NBA için doğruluk oranının %76'dan %75,6'ya, kesinliğin %8,1'den %80,1'e, seçiciliğin %79,9'dan %78,5'e, F ölçüm değerinin %76'dan %75,6'ya ve Matthews korelasyon katsayısının %52,4'den %51,5'e düştüğü görüldü. LRA için kesinliğin %82,3'den %81,3'e, duyarlılığın %74,4'den %73,9'a, seçiciliğin %80,9'dan %80'e düştüğü ve Matthews korelasyon katsayısının %55,2'den %53,7'ye düştüğü izlendi. KAA için doğruluk oranının %80,3'den %76,1'e, kesinliğin %84,4'den %79,3'e, duyarlılığın %78'den %75,8'e, seçiciliğin %82,9'dan %76,6'ya, F ölçüm değerinin %80,3'den %76,2'ye ve Matthews korelasyon katsayısının %60,7'den %52,2'ye düştüğü gözlendi. ROA için doğruluk oranının %97,1'den %74,4'e, kesinliğin %98,6'dan %75,2'ye, duyarlılığın %96,1'den %78,6'ya, seçiciliğin %98,4'den %69,5'e, F ölçüm değerinin %97,1'den %74,3'e ve Matthews korelasyon katsayısının %94,3'den %48,3'e düştüğü izlendi. EYKKA için doğruluk oranının %97,1'den %68,4'e, kesinliğin %97,3'den %70,5'e, duyarlılığın %97,5'den %71,5'e, seçiciliğin %96,7'den %64,7'ye, F ölçüm değerinin %97,1'den %68,4'e ve Matthews korelasyon katsayısının %94,2'den %36,2'ye düştüğü görüldü. Son olarak DVMA için kesinliğin %85,7'den %83,8'e, seçiciliğn %87'den %84,5'e, F ölçüm değerinin %76,9'dan %75,3'e ve Matthews korelasyon katsayısının %53,3'den %52,4'e düştüğü gözlendi. Ortaya çıkan değerler ile eğitim ve test veri setlerinin büyüklüğünün ne olması gerektiği tartışmalarında %90'a %10 olmasının uygun olmadığı görüldü. Bu sonuçlar literatürü desteklemektedir. Zira

Chakraborty (2019) çalışmasında, eğitim veri setinin %90 olduğu bölünmelerin sadece çok büyük veri setleri için uygun olduğunu ifade etmiştir.

Veri seti %80 eğitim verisi, %20 test verisi olarak ayrıldığında DO performans ölçütü açısından EYKKA %68,8 ile en düşük performansı ortaya koydu. EYKKA için elde edilen performans sonuçlarının diğer algoritmalardan elde edilen performans sonuçlarına göre az miktarda düşük olduğu görüldü. DO açısından diğer 6 algoritma birbirlerine yakın sonuçlar ortaya çıkardı en yüksek performansı %77,2 ile LRA ortaya koydu. Bu veri seti çalışmasında kesinlik performans ölçütü ROA ve EYKKA hariç diğer tüm algoritmalarda %80'in üzerinde bir performans gösterdi. En yüksek performansı ise %85,9 ile DVMA ortaya koydu. Daha sonra en iyi performanslar sırasıyla TSA (%85,7), LRA (%83,3), NBA (%82,1), KAA (%81,8), ROA (%77,5) ve EYKKA'dan (%72) elde edildi. Kesinlik performans ölçütünde DVMA (%85,9), duyarlılıkta ROA (%77,5), seçicilikte TSA (%87,5), F ölçümünde LRA (%77,3) ve Matthews korelasyon katsayısında yine LRA (%55,1) en iyi performansı sergiledi. Burada farklı performans ölçütlerinde farklı algoritmaların daha etkin olduğu ve algoritmaların birbirlerine yakın sonuçlar ortaya çıkardığı görüldü. Genel olarak bu eğitim veri setinde LRA ve DVMA'nın tüm performans ölçütlerinin ortalamasına bakıldığında diğerlerinden daha başarılı olduğu görüldü. Özellikle DVMA'nın KBY ve diğer önemli hastalıkların özelliklede meme kanserinin tanısında etkin olarak kullanıldığı sonucu literatürdeki çalışmaları da desteklemektedir. S. Kaur, J. Singla, L. N. ve diğer araştırmacılar (2020), çalışmalarında meme kanserinin tanısı için KAA, DVMA ve EYKKA algoritmaları arasında yapılan karşılaştırmada en iyi sonucun DVMA tarafında gösterildiği sonucu ortaya çıkmıştır (Kaur ve ark., 2020). Ayrıca Noi ve Kappas (2017), çalışmalarında farklı büyüklükteki veri setleri için analizler yapmıştır. Bu analizler sonucunda %90-%95 doğruluk elde edilmiştir. En yüksek performans, %60 bölünen eğitim veri seti için DVM ile elde edilmiştir (Noi ve Kappas, 2017; El Houssainy ve ark., 2019). Ayrıca bu veri setinde KBY hastalığı bulunanlar için en iyi tahmini ROA yaparken, KBY hastalığı bulunmayanlar için en iyi performansı TSA gösterdi.

Veri seti %70 eğitim verisi, %30 test verisi olarak ayrıldığında doğruluk oranında %77,3 ile F ölçümünde %77,3 ile ve Matthews korelasyon katsayısında %55,2 ile performans ölçütlerinde en yüksek performansı LRA gösterdi. Kesinlik için %86,4 ile ve seçicilik için %88 ile en yüksek performansı DVMA ortaya koydu.

Duyarlılık için ise en iyi performans %77,2 ile ROA tarafından gerçekleştirildi. Algoritmaların başarı yüzdeleri birbirine yakın sonuçlar ortaya koymasına rağmen tüm performans ölçütlerinin ortalamasına bakıldığında daha önceki veri setinde olduğu gibi LRA ve DVMA en iyi sonuçları gösterdi.

Veri seti %60 eğitim verisi, %40 test verisi ve %50 eğitim verisi, %50 test verisi olarak ayrıldığında da benzer sonuçlar elde edildi. Daha önceki veri setinden farklı olarak seçicilik performans ölçütünde TSA en iyi performansı gösteren algoritma oldu. Ayrıca KBY hastalığı bulunanlar için en yüksek tahmini ROA gösterirken, KBY hastalığı bulunmayanlar için en iyi tahminleri TSA ve DVMA gösterdi. Tüm veri setleri büyüklüğünde genel olarak en yüksek performansı LRA ve DVMA gösterdi. Eğitim veri setinin %90 olduğunda tüm algoritmaların performanslarında düşüşler olduğu görüldü. Bu sonuçlar doğrultusunda literatürün desteklendiği denetimli makine öğrenmesi üzerine eğitim setinin kullanılarak yapılan Darulova, Troyer ve Cassidy'nin (2021) çalışmalarında eğitim verisi ile yapılan uygulamalarda daha yüksek performans sergilendiği görülmüştür.

Hasta yaşları ile KBY hastası olunma durumu arasındaki ilişkinin tespiti için kullanılan ROC analizi neticesinde ROC eğrisinin altında kalan alanın 0,661 olduğu görülmüştür. Bu analizde p değeri 0,05 değerinden küçük olduğu için analiz sonucunda ortaya konulan değerleri için kabul edilebilir değerler olduğu ortaya çıkmıştır. Analiz neticesinde 54,5 yaşın üzerinde olan bireylerin KBY hastası olma olasılığı %66,1 olarak hesaplanmıştır.

Tüm bu sonuçlar değerlendirildiğinde eğitim veri seti için %60-80 ve test veri seti için %20-40 oranlarının en uygun ayırma oranı olduğu görüldü. Eğitim veri seti %60-80 arasında alındığında oluşturulan modeller arasında performans ölçütlerinin çoğunda DVMA en iyi performansı göstermiştir. Ayrıca EYKKA ve ROA hariç diğer algoritmaların oluşturduğu tahmin modellerinin bu aralıkta alınan test ve eğitim veri seti ayırmalarında performanslarında artışlar oldu. Elde edilen sonuçlar ile denetimli makine öğrenmesinde veri setinin uygun büyüklükteki eğitim ve test veri setine ayrılmasının tahmin modellerinin başarısı üzerinde önemli olduğu ve performans ölçütlerini önemli derecede etkilediği görüldü.

Bu çalışmada makine öğrenimi algoritmaları ile oluşturulan tahmin modellerinin performans ölçütlerini karşılaştırmada birden fazla durumun değerlendirilmesi gerektiği ortaya konuldu. Tahmin modelleri oluşturulurken farklı

algoritma modelleri üzerinde çalışıldığı gibi farklı eğitim ve veri seti büyüklüklerinin önemi farklı formüller ile oluşturulan performans ölçütleri ile değerlendirildi. Makine öğrenimi yöntemleri karşılaştırıldığında birden fazla performans ölçütü kullanıldığında en başarılı makine öğrenimi yöntemini tespit etmek mümkün görünmedi. Ancak performans ölçütlerinin çoğunda daha etkin tahminler oluşturan DVMA, ROA ve LRA'nın diğerlerine göre daha iyi tahmin modelleri oluşturduğu görüldü.

6. SONUÇ VE ÖNERİLER

Çalışmadan elde edilen sonuçlar neticesinde denetimli makine öğrenimi algoritmalarında veri setinin, eğitim veri seti ve test veri seti olarak bölünmesi, performans ölçütü sonuçlarına etki etmektedir. Eğitim veri setinin %60-80 aralığında alınması ilgili algoritmalarda en yüksek performansı ortaya koymaktadır. Eğitim ve test veri setleri kullanılmadan uygulanan algoritmaların bazıları çok yüksek performans gösterse de veri setine eklenecek olan yeni gözlemlerde nasıl bir başarı sağlayacağı belirsizdir. Bu nedenle eğitim veri seti uygulanarak yapılan analizlerde ilgili algoritmaların performanslarında önemli ölçüde düşüşler görülmektedir. Bu durum denetimli makine öğrenimi için eğitim veri setinin kullanılmasının önemini göstermektedir.

Literatür araştırmalarında çok fazla sayıda algoritma olduğu görülmektedir. Hangi veri seti için hangi algoritmanın en uygun olduğunun kararını vermek zaman isteyen bir uzmanlık konusudur. Bu çalışmada kullanılan kısıtlı algoritmaların dışında farklı algoritmalar da kullanılarak farklı modellerin oluşturulması mümkündür.

Gerçekleştirilen çalışmada farklı algoritmaların farklı büyüklükte seçilen eğitim veri setleri için değişken performanslar sergilediğini gösterdi. Hedef değişkenin farklı sınıfları için de farklı algoritmaların daha etkin olduğunu göstermektedir.

Gerçek hasta verileri ile gerçekleştirilen bu çalışmada hastaların cinsiyet, yaş ve kan değerlerinin değişken olarak alındığı 7 farklı algoritma ile denetimli makine öğrenimi algoritmaları kullanılarak hastalığın tanısı için tahmin modelleri oluşturulmuş ve bu modellerin başarısı farklı performans ölçütleriyle karşılaştırılmıştır. Yüksek performansların görüldüğü algoritmalar ile KBY hastalığının tanısı için karar vericilere yardımcı olacak sonuçlar çıkartılmaya çalışılmıştır. Gelecek çalışmalarda hastaların KBY'nin tanısından sonra sağ kalım sürelerinin tahminine ilişkin farklı algoritmaların oluşturulması planlanmaktadır. Ayrıca bu çalışmada oluşturulan tahmin modellerini, gelecek çalışmalarda elde edilecek aynı özellikteki yeni hastalara ilişkin veriler ile etkin hale getirerek bu çalışmanın model performansının değerlendirilmesi de planlanmaktadır.

KAYNAKLAR

Ağır H, Tıraş H.H. (2018). "Türkiye'de Sağlık Harcama Türlerinin Değerlendirilmesi". *KSÜSBD*, Cilt 15, Sayı 2. 643-670.

Ahmad Z et al. (2020). "Network intrusion detection system: A Systematic Study Ofmachine Learning and Deep Learning Approaches". *wileyonlinelibrary.com/journal/ett*.

Allugunti V.R. (2022). "Breast Cancer Detection Based on Thermographic Images Using Machine Learning and Deep Learning Algorithms". *International Journal of Engineering in Computer Scieence*, 4(1): 49-56.

Alnazer I, Bourdon P, Urruty T, et al. (2021). "Recent Advances in Medical İmage Processing for the Evaluation of Chronic Kidney Disease". *Medicine İmagine Analyses*.

Alotaibi S, Mehmood R, Katib I, et al. (2020). "Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter". *Apache Spark, and Machine Learning, MDPI*.

Alpar R. (2017). "Uygulamalı Çok Değişkenli İstatistiksel Yöntemler". *Detay Yayıncılık*, pp: (591-592).

Amaresan M.S, Geetha R. (2008). "Early Diagnosis of Ckd and Its Prevention". *JAPI*, pp: (41-46).

Arora I, Khanduja N, Bansal M. (2021). "Effect of Distance Metric and Feature Scaling on KNN Algorithm while Classifying X-rays". *The 10th Seminary of Computer Science Research at Feminine*.

Artrith N, Butler K.T, Coudert F.X, et al. (2021). "Best Practices in Machine Learning for Chemistry". *Nature Chemistry*, Vol 13, 505–508.

Atalay M, Çelik E. (2017). "Büyük Veri Analizinde Yapay Zeka ve Makine Öğrenmesi Uygulamaları". *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, pp: (155-172).

Bahtiyarca Z.T, Çakcı F.A. (2021). "Kronik Hemodiyaliz Hastasında Kas İskelet ve Periferik Sinir Sistemi Tutulumu: Olgu Sunumu ve Literatürün Gözden Geçirilmesi". *Turk J Osteoporos*, (27):173-178.

Bailey G, Joffrion A, Pearson M. (2018). "A Comparison of Machine Learning Applications Across Professional Sectors". *School of Information, University of Texas at Aust; SSRN*.

Başer B.Ö, Yangın M, Sarıdaş E.S. (2021). "Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması". *Süleyman Demirel University Journal of Natural and Applied Sciences*, pp: (112-120).

Battineni G, Sagaro G.G, Chinatalapudi N, Amenta F. (2020). "Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis". *Journal of Personalized Medicine MDPI*.

Bhalodiya N, Parsania V.S. (2014). "Applying Naive Bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis". *International Journal of Darshan Institute on Engineering Research and Emerging Technologies*, pp: (60-64).

Bhavsar H, Ganatra A. (2012). "A Comparative Study of Training Algorithms for Supervised Machine Learning". *International Journal of Soft Computing and Engineering (IJSCE)*, pp: (74-81).

Bilgen T. (2021). "Covid-19 Pandemisi Sürecinde Kronik Böbrek Yetmezliği Tanısı ile Hemodiyaliz Tedavisi Alan Hastaların Psikolojik Dayanıklılık ve Etkileyen Faktörler Açısından İncelenmesi". (Uzmanlık Tezi). *Bursa Uludağ Üniversitesi Tıp Fakültesi*.

Bilgin M. (2017). "Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi". *Breast*.

Bithas P.S, Michailidis E.T. et al. (2019). "A Survey on Machine-Learning Techniques for UAV-Based Communications". *MDPI*.

Brownlee J. (2022). " Data Preparation for Machine Learning". *MLA*.

Burkart N, Huber M.F. (2021). "A Survey on the Explainability of Supervised Machine Learning". *Journal of Artificial Intelligence Research*, pp: (245-317).

Campbell S.I, Allan D.B, Barbour A.M, et al. (2021). "Outlook for Artificial İntelligence and Machine Learning at the NSLS-II". *Science and Technology*.

Canedo E.D, Mendes B.C. (2020). "Software Requirements Classification Using Machine Learning Algorithms". *Entropy MDPI*.

Ceyhun H.A, Kırpınar İ. (2019). "Son Dönem Böbrek Yetmezliği Nedeni ile Böbrek Nakli Yapılan veya Diyaliz Uygulanan Hastalarda Psikiyatrik Tanı". *Anatolian Journal of Psychiatry*, pp: (426-435).

Chakraborty C, Advanced Classification Techniques for Healthcare Analysis, IGI Global, 2019.

Chang Z, Du Z, Zhang F, et al. (2020). "Landslide Susceptibility Prediction Based on Remote Sensing Images and GIS: Comparisons of Supervised and Unsupervised Machine Learning Models". *Remote Sensing MDPI*, (12), 502.

Çalışkan S, Yazıcıoğlu S.A, Demirci U, Kuş Z. (2020). "Yapay Sinir Ağları, Kelime Vektörleri ve Derin Öğrenme Uygulamaları". *FSM Vakıf Üniversitesi Mühendislik ve Fen Bilimleri Enstitüsü Bilgisayar Anabilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı*.

Damanik I. S, Windarto A.P. et al. (2019). "Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm". *The International Conference on Computer Science and Applied Mathematic*.

Darulova J, Troyer M, Cassidy M.C. (2021). "Evaluation of Synthetic and Experimental Training Data in Supervised Machine Learning Applied to Charge-State Detection of Quantum Dots". *Machine Learning: Science and Technology*.

Demir C.A, Özer Z. (2022). "The Relationship of Symptoms and Comfort in Patients Receiving Hemodialysis". *Journal of Nephrology Nursing*.

Demirhan A, Kılıç Y.A, Güler I. (2010). "Tıpta Yapay Zeka Uygulamaları". *Yoğun Bakım Dergisi*, 9(1): 31-41.

Ding S, Jia W, Su C, et al. (2011). "Research of Neural Network Algorithm Based on Factor Analysis and Cluster Analaysis". *Neural Comput and Applic*, pp: (297-302).

Dündar T.R, Sarıçiçek İ, Çınar E, Yazıcı A. (2021). "Kestirimci Bakımda Makine Öğrenmesi: Literatür Araştırması". *ESOGÜ Müh. Mim. Fak Dergisi*, pp: (256-276).

El-Houssainy A. Radya, Ayman S. Anwar. (2019). "Prediction of kidney disease stages using data mining algorithms". *Science Direct*.

Ellerman D. (2022). "Introduction to Logical Entropy and Its Relationship to Shannon Entropy". *arxiv.org*.

Elmacı A.M, Dönmez M.İ. (2019). "Böbrek ve Üriner Sistemin Doğumsal Anomalileri: 806 Olgunun Analizi". *Journal of Contemporary Madicine*, pp: (284-287).

Elsivers M.M, Schepper A.D, Bosmans J.L, et al. (1995). "High Diagnostic Performance of CT Scan for Analgesic Nephropathy in Patients with Incipient to Severe Renal Failure". *Kidney International*, Volume 48, 1316-1323.

Eroğlu K, Palabaş T. (2019). "The Impact on the Classification Performance of the Combined Use of Different Classification Methods and Different Ensemble Algorithms in Chronic Kidney Disease Detection". *emo.org.tr*, pp: (512-516).

Eyüpoğlu C. (2020). "Kronik Böbrek Hastalığının Erken Tanısı İçin Yeni Bir Klinik Karar Destek Sistemi". *Avrupa Bilim ve Teknoloji Dergisi*, pp: (448-455).

Fan G.F, Guo Y.H, Zheng J.M, Hong W.C. (2019). "Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting". *MDPI*.

Flach P. (2016). "ROC Analysis". *University of Bristol.*

Ghorbanzadeh O, Blaschke T, Gholamnia K, et al. (2019). "Evaluation of Different Machine Learning Methods and Deep-Learning Convolutional Neural Networks for Landslide Detection". *MDPI*.

Gökalp Ö.M. (2021). "Makine Öğrenmesi, Gazi Üniversitesi, Gazi Bilişim Enstitüsü". *Adli Bilişim Bölümü.*

Gökdemir A, Çalhan A. (2022). "Deep Learning and Machine Learning Based Anomaly Detection in Internet of Things Environments". *Journal of the Faculty of Engineering and Architecture of Gazi University*, pp: (1945-1956).

Gündüz H. (2019). "Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets". *IEEE ACCESS*, pp: (15540-15551).

Hasnain M, Pasha M.F. et al. (2020). "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking". *IEEE ACCESS*, pp: (90847-90861).

Hill N.R, Fatoba S.T, Oke J.L. (2016). "Global Prevalence of Chronic Kidney Disease A Systematic Review and Meta-Analysis". *Plos One DOI:10.1371/journal.pone.015876*, pp: (1-18).

Hofmann B, Svenaeus F. (2018). "How medical technologies shape the experience of illness, Life Sciences". *Society and Policy*.

Hofmann, M. (2006). "Support Vector Machines - Kernels and the Kernel Trick". Stud.uni-b*amberg.de*.

Howley, T, Madden, M. G. (2004). "The Genetic Evolution of Kernels for Support Vector Machine Classifiers". *15th Irish Conference on Artificial Intelligence, Cognitive Science*.

Hsieh K, Phanishayee A, Mutlu O, Gibbons P.B. (2020). "The Non-IID Data Quagmire of Decentralized Machine Learning". *International Conference on Machine Learning Online*.

İbrahim İ.M, Abdulazeez A.M. (2021). "The Role of Machine Learning Algorithms for Diagnosing Diseases". *Journal of Applied Science and Technology Trends*, pp: (10-19).

Jankowski J, Floege J, Fliser D, Böhm M, Marx N. (2021). "Cardiovascular Disease in Chronic Kidney Disease". *Pathophysiology and Therapeutic Options*, pp: (1157-1172).

Javeed A, Zhou S, Yongjian L et al. (2019). "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection". *IEEE ACCESS*, pp: (180235-180243).

Karaman O. (2022). "Sağlık&Bilim". *Medikal Araştırmalar II, I.Basım, Efe Akademi* (43).

Kaur P, Sharma M, Mittal M. (2018). "Big Data and Machine Learning Based Secure Healthcare Framework". *International Conference on Computational Intelligence and Data Science (ICCIDS)*.

Kaur S, Singla J, Nkenyereye L, et al. (2020). "Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives". *IEEE ACCESS*, pp: (228049-228069).

Khan B, Naseem R, Muhammed F, Abbas G, Kim S. (2020). "An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy". *IEEE ACCESS*, pp: (55012-55022).

Koç S. (2019). "Genetik Alanında Elde Edilen Verilerin Makine Öğrenimi Algoritmaları Yardımıyla Karşılaştırılarak En Etkin Yöntemin Belirlenmesi". (Doktora Tezi). *OMÜ Sağlık Bilimleri Enstitüsü*.

Kowalczyk, B. (2017). "Support Vector Machines. Succinctly e- book Morrisville". *USA*, pp: (48-53).

Lemley K.V. (2019). "Machine Learning Comes to Nephrology". *jasn.org*.

Li Q, Fan Q.L, Han Q.X, et al. (2022). “Machine Learning in Nephrology: Scartching the Surface”. *Chinese Madical Journal*, pp: (687-698).

Li Z, Li R, Jin G. (2020). “Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary”. *IEEE ACCESS*, pp: (75073-75084).

Machado G.R, Silva E, Goldschmidt R.R. (2021). “Adverserial Machine Learning in İmage Classification: A Survey Toward the Defender’s Perspective”. *ACM Computing Surveys*, pp: (1-38).

Mahesh B. (2018). “Machine Learning Algorithms – A Review”. *International Journal of Science and Research (IJSR)*, pp: (381-386).

Marsland S. (2015). “Machine Learning An Algorithmic Perspective”. *CRC Press, Boca Raton, FL*, pp: (250-257).

Matuszny M. (2020). “Building Decision Trees Based on Production Knowledge as Support in Decision-Making Process”. *Production Engineering Archives*, pp: (36-40).

Mirza S, Mittal S, Zaman M. “Decision Support Predictive Model for Prognosis of Diabetes Using SMOTE and Decision Tree”. *International Journal of Applied Engineering Research ISSN*, pp: (9277-9282).

Molnar C, Casallicchio G, Bischi B. (2021). “Interpretable Machine Learning – A Brief History”. *State of the Art and Challengs, Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp: (417-431).

Morency L.P, Liang P.P, Zadeh A. (2022). “Tutorial on Multimodal Machine Learning”. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pp: (33–38).

Mosavi A, Öztürk P, Chau K.W. (2018). “Flood Prediction Using Machine Learning Models: Literature Review”. *MDPI*.

Muhammad I, Yan Z. (2015). “Supervised Machine Learning Approaches: A Survey”. *Ictact Journal on Soft Computing*, pp: (946-952).

Nasteski V. (2017). “An Overview of the Supervised Machine Learning Methods”. *researchgate.net*.

Noi P.T, Kappas M. (2017). “Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery”. *MDPI*.

Omary Z, Mitenzi F. (2010). “Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised Learning”. *International Journal for Infonomics (IJI)*, pp: (314-325).

Osisanwo F.Y, Akinsola J.E.T, Awodele O, et al. (2017). “Supervised Machine Learning Algorithms: Classification and Comparison”. *International Journal of Computer Trends and Technology*, pp: (128-138).

Osmanoğlu U.O, Atak O.N, Cağlar K ve ark. (2020). "Sentiment Analysis for Distance Education Course Materials: A Machine Learning Approach". *Journal of Educational Technology & Online Learning*, pp: (31-48).

Qin j, Chen l, Liu Y, Liu C, Feng C, Chen B. (2019). "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease". *IEEE ACCESS*, pp: (20991-21002).

Perazella M.A. (2015). "The Urine Sediment as a Biomarker of Kidney Disease". *Am J Kidney Dis*, 66(5):748-755.

Pradhan A.M.S, Kim Y.T. (2020). "Rainfall-Induced Shallow Landslide Susceptibility Mapping at Two Adjacent Catchments Using Advanced Machine Learning Algorithms". *International Journal of Geo-Information*.

Pulat M, Kocakoç İ.D. (2021). "Türkiye'de Makine Öğrenmesi ve Karar Ağaçları Alanında Yayınlanmış Tezlerin Bibliyometrik Analizi". *Yönetim ve Ekonomi*, Cilt 28 sayı 2: 287-308.

Raschka S, Patterson J, Nolet C. (2020). "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence". *Information MDPI*.

Rizki M, Arhami M, Huzeni D. (2021). "Algoritma Naive Bayes Classifier Menggunakan Teknik Laplacian Correction". *Jurnal Teknologi*, pp: (39-45).

Roscher R, Bohn B, Duarte M.F, et al. (2020). "Explainable Machine Learning for Scientific Insights and Discoveries". *IEEE ACCESS*.

Roth J.A, Radevski G, Marzolini C, et al. (2020). "Cohort-Derived Machine Learning Models for Individual Prediction of Chronic Kidney Disease in People Living With Human Immunodeficiency Virus: A Prospective Multicenter Cohort Study". *The Journal of Infectious Diseases*.

Rustam F, Reshi A.A, Mehmood A, et al. (2020). "Covid-19 Future Forecasting Using Supervised Machine Learning Models". *IEEE ACCESS*, pp: (101489-101499).

Rymarczyk T, Kozłowski E, Kłosowski G, Niderla K. (2019). "Logistic Regression for Machine Learning in Process Tomography". *MDPI*.

Sabita H, Fitria, Herwanto R. (2021). "Analisa dan Prediksi Iklan Lowongan Kerja Palsu Dengan Metode Natural Language Programing dan Machine Learning". *Jurnal Informatika*.

Schanstra J.P, Zürbig P, Alkhalaf A, et al. (2015). "Diagnosis and Prediction of CKD Progression by Assessment of Urinary Peptides". *Clinical Research*, pp: (1999-2010).

Sealfon R.S.G, Mariani L.H, Kretzler M, Troyanskaya O.G. (2020). "Machine learning, the kidney, and genotype–phenotype analysis, Kidney International". pp: (1141-1149).

Sharma S, Sharma V, Sharma A. (2016). "Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis". *Cornell Univercity Arxiv*.

Singh M.K, Baluja G.S, Sahu D.P. (2017). "Analysing Machine Learning Algorithms & their Areas of Application". *International Journal for Research in Applied Science and Engineering Technology*, pp: (653-657).

Statsoft. (2018). "Support Vector Machines". *www.statsoft.com/textbook/support-vector-machines*,

Süleymanlar G, Ateş K, Seyahi N. (2021). "Türkiye'de Nefroloji, Diyaliz ve Transplantasyon 2020". *Türk Nefroloji Derneği Yayınları*.

Şentürk A, Levent B. A. (2000). "Hemodiyalize Giren Kronik Böbrek Yetmezliği Olan Hastalarda Psikopatoloji". *O.M.Ü Tıp Dergisi*, 17(3): 163-172.

Tan A.C, Gilbert D. (2003). "An empirical comparison of supervised machine learning techniques in bioinformatics". *Preprint Version. To appear in the Proceedings of the First Asia Pacific Bioinformatics Conference*.

Tanrıverdi M.H, Karadağ A, Hatipoğlu E.Ş. (2010). "Kronik Böbrek Yetmezliği". *Konuralp Tıp Dergisi*, pp: (27-32).

Tunthanathip T, Duangsuwan J, Wattanakitrungroj N, et al. (2021). "Comparison of Intracranial Injury Predictability Between Machine Learning Algorithms and the Nomogram in Pediatric Traumatic Brain Injury". *Neurosurg Focus*, Volume 51.

Turing A. (1950). "Computing Machinery and İntelligence". *Mind* 49.

TÜİK. (2021), "Adrese Dayalı Nüfus Kayıt Sistemi Sonuçları". *www.tuik.gov.tr*

Tütüncü T.E. (2022). "Makine Öğrenmesi Algoritmaları ile Kredi Temerrüt Riskini Tahmin Etme". (Yüksek Lisans Tezi). *Uludağ Üniversitesi Sosyal Bilimler Enstitüsü*.

Uysal C, Kır S, Arık N. (2020). "Hemodiyaliz Tedavisi Gören Diyabetik Hastalarda Ölçülen HbA1c Değerlerindeki Değişkenliğin Sağkalım ile İlişkisi". *Uludağ Üniversitesi Tıp Fakültesi Dergisi*, pp: (189-194).

Varshney K.R. (2022). "Trustworthy Machine Learning". *IEEE*.

Venkatesan C, Karthigaigumar P, Paul A, et al. (2018). "ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare Applications". *IEEE ACCESS*, pp: (9767-9773).

Wu J, Dong M, Rigatto C, et al. (2018). "Lab-On-Chip Technology for Chronic Disease Diagnosis". *npj Dijital Medicine*.

Yalçın A.U, Yıldız A, Odabaş A.R ve ark. (2019). "Temel Nefroloji". *Güneş Tıp Kitapevleri*, pp: (17-19).

Yavuz A, Çilengiroğlu V.Ç. (2020). "Lojistik Regresyon ve CART Yöntemlerinin Tahmin Edici Performanslarının Yaşam Memnuniyeti Verileri İçin Karşılaştırılması". *Avrupa Bilim ve Teknoloji Dergisi*, sayı 18, 719-727.

Yazar I, Yavuz H.S, Çay M.A. (2009). "Temel Bileşen Analizi Yönteminin ve Bazı Klasik ve Robust Uygulamalarının Yüz Tanıma Uygulamaları". *Eskişehir Osmangazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, pp: (49-63).

Xu H, Shou J, Asteris P.G, et al. (2019). "Supervised Machine Learning Techniques to the Prediction of Tunnel Boring Machine Penetration Rate". *Applied Sience MDPI*, 9, 3715.

Xing W, Bei Y. (2019). "Medical Health Big Data Classification Based on KNN Classification Algorithm". *IEEE ACCESS*, pp: (28808-28819).

EKLER

Veri setinin çalışmaya uygun hale getirilebilmesi için gerçekleştirilen aşamalar

1- Albumin değeri <3,5 ise “düşük”, 3,5-5 arasında ise “normal”, >5 ise “yüksek” olacak şekilde kategorize edildi.

2- Dansite değeri <1015 ise “düşük”, 1015-1025 arasında ise “normal”, >1025 ise “yüksek” olacak şekilde kategorize edildi.

3- Fosfor değeri <2,3 ise “düşük”, 2,3-4,7 arasında ise “normal”, >4,7 ise “yüksek” olacak şekilde kategorize edildi.

4- Şeker değeri 0-45 arası ise “normal”, >45 ise “normal değil” olarak kategorize edildi.

5- Açlık şekeri değeri <70 ise “düşük”, 70-110 arasında ise “normal”, >70 ise “yüksek” olacak şekilde kategorize edildi.

6- Hematokrit değeri erkekler için <38,9 kadınlar için <34,8 ise “düşük”, erkekler için 38,9-50,9 arasında kadınlar için 34,8-45 arasında ise “normal”, erkekler için >50,9 kadınlar için >45 ise “yüksek” olacak şekilde kategorize edildi.

7- Hemoglobin değeri erkekler için <13,3 kadınlar için <12 ise “düşük”, erkekler için 13,3-17,2 arasında kadınlar için 12-15 arasında ise “normal”, erkekler için >17,2 kadınlar için >15 ise “yüksek” olacak şekilde kategorize edildi.

8- Kalsiyum değeri <8,8 ise “düşük”, 8,8-10,2 arasında ise “normal”, >10,2 ise “yüksek” olacak şekilde kategorize edildi.

9- Kreatinin değeri <0,4 ise “düşük”, 0,4-1,4 arasında ise “normal”, >1,4 ise “yüksek” olacak şekilde kategorize edildi.

10- Potasyum değeri <3,5 ise “düşük”, 3,5-5,5 arasında ise “normal”, >5,5 ise “yüksek” olacak şekilde kategorize edildi.

11- RBC değeri erkekler için <4,54 kadınlar için <3,85 ise “düşük”, erkekler için 4,54-5,78 arasında kadınlar için 3,85-5,16 arasında ise “normal”, erkekler için >5,78 kadınlar için >5,16 ise “yüksek” olacak şekilde kategorize edildi.

12- Sodyum değeri <135 ise “düşük”, 135-145 arasında ise “normal”, >145 ise “yüksek” olacak şekilde kategorize edildi.

13- WBC değeri erkekler için <3,7 kadınlar için <3,9 ise “düşük”, erkekler için 3,7-9,7 arasında kadınlar için 3,9-11,7 arasında ise “normal”, erkekler için >9,7 kadınlar için >11,7 ise “yüksek” olacak şekilde kategorize edildi.

BUN değeri <5 ise “düşük”, 5-24 arasında ise “normal”, >24 ise “yüksek” olacak şekilde kategorize edildi.

ÖZ GEÇMİŞ

Tolga Demirel, Samsun Ondokuz Mayıs Lisesi'ni bitirdikten sonra Ondokuz Mayıs Üniversitesi Fen-Edebiyat Fakültesi, İstatistik bölümünden 2001 yılında mezun oldu. 2019 yılında OMÜ LEE Biyoistatistik ve Tıp Bilişimi Yüksek Lisans programına girdi. Türkiye İstatistik Kurumu Gaziantep Bölge Müdürü olarak görev yapan Tolga Demirel, iyi derecede İngilizce bilmektedir (YDS: 68,75, YÖKDİL: 83,75). Temel ilgi alanları, proje geliştirme ve yazma, istatistik okur-yazarlığının yaygınlaştırılması çalışmalarıdır (06/01/2023).

İletişim Bilgileri

ORCID ID : *0000-0002-8912-221X*

Yayınlar:

1. Demirel T, Tomak L, Comparison of Different Machine Learning Methods in the Diagnosis of Chronik Renal Failure (CRF), XII. Ulusal ve V. Uluslararası Biyoistatistik Kongresi, Biyoistatistik Derneği, 2021.