T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YENİ DOĞANLARIN YOĞUN BAKIM İHTİYACININ MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE DOĞUM ÖNCESİNDE TAHMİN EDİLMESİ

Serkan GESOĞLU

YÜKSEK LİSANS TEZİ
Matematik Mühendisliği Anabilim Dalı
Matematik Mühendisliği Programı

Danışman

Doç. Dr. Birol ASLANYÜREK

Eş Danışman

Doç. Dr. Emrah AYDIN

Ekim, 2022

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YENİ DOĞANLARIN YOĞUN BAKIM İHTİYACININ MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE DOĞUM ÖNCESİNDE TAHMİN EDİLMESİ

Serkan GESOĞLU tarafından hazırlanan tez çalışması 27.10.2022 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı, Matematik Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Birol ASLANYÜREK Yıldız Teknik Üniversitesi Danışman	Doç. Dr. Emrah Aydın Tekirdağ Namık Kemal Üniversitesi Eş-Danışman

Jüri Üyeleri

DOÇ. DR. Birol ASLANYÜREK, Danışman
Yıldız Teknik Üniversitesi _____

DOÇ. DR. HALE KÖÇKEN, Üye
Yıldız Teknik Üniversitesi _____

DR. ÖĞR. ÜYE. OĞUZHAN KIVRAK, Üye
Bandırma Onyedi Eylül Üniversitesi _____

Danışmanım Doç. Dr. Birol ASLANYÜREK sorumluluğunda tarafımca hazırlanan Yeni Doğanların Yoğun Bakım İhtiyacının Makine Öğrenmesi Teknikleri ile Gebelik Esnasında Tahmin Edilmesi başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Serkan GESOĞLU

İmza

Aileme

ve

biricik eşime

TEŞEKKÜR

Bu tez çalışmasında bilgi ve tecrübesiyle bana her zaman yardımcı olan saygıdeğer danışman hocam Doç. Dr. Birol Aslanyürek'e, veri setinin elde edilmesi, tıp alanındaki bilgi ve tecrübeleriyle bizleri yönlendiren saygıdeğer danışman hocam Doç. Dr. Emrah Aydın'a, verinin sağlanması konusunda ve kadın doğum verileri konusunda bizlere destek olan Doç. Dr. Asiye Uzun'a, yüksek lisans eğitimim süresince ders aldığım tüm hocalarıma ve öğretim hayatım boyunca destekleri ile her zaman yanımda olan sevgili eşim Gizem Gesoğlu ve aileme sonsuz teşekkürlerimi sunarım.

Serkan Gesoğlu

İÇİNDEKİLER

KISALTMA LİSTESİ vii

ŞEKİL LİSTESİ ix

TABLO LİSTESİ x

ÖZET xi

ABSTRACT xiii

1 GİRİŞ 1

1.1 Literatür Özeti 1
1.2 Tezin Hipotezi ve Amacı 3
1.3 Orijinal Katkı 3

2 MAKİNE ÖĞRENMESİ VE KAVRAMLARI 4

2.1 Verilerin Düzenlenmesi 5

2.1.1 Veri Temizleme 5
2.1.2 Veri Birleştirme 6
2.1.3 Veri Dönüştürme 7
2.1.4 Veri İndirgeme 7

2.2 Makine Öğrenmesi Yöntemleri 8

2.2.1 Denetimli Öğrenme Yöntemleri 9
2.2.2 Denetimsiz Öğrenme Yöntemleri 10

2.3 Denetimli Makine Öğrenmesi Algoritmaları 10

2.3.1 Lineer Regresyon 10
2.3.2 Destek Vektör Makineleri (DVM) 11
2.3.3 Karar Ağaçları 14
2.3.4 Naive Bayes Algoritması 16
2.3.5 En Yakın K Komşu Algoritması 18
2.3.6 Yapay Sinir Ağları 19
2.3.7 AdaBoost Algoritması 20

2.4 Denetimsiz Makine Öğrenmesi Algoritmaları 21

2.4.1 Kümeleme Algoritması 21
2.4.2 Hiyerarşik Kümeleme 21

2.4.3 K-Ortalama 22
2.5 Makine Öğrenmesi Algoritmasına Karar Verme..... 22
2.5.1 Regresyon Modelleri için Kullanılan Ölçüler 22
2.5.2 Sınıflandırma Modelleri için Kullanılan Ölçüler 23
3 YENİ DOĞANLARIN YOĞUN BAKIM İHTİYACININ TAHMİN EDİLMESİ 26
3.1 Verilerin Düzenlenmesi 26
3.2 İstatistiksel Analiz..... 32
3.3 Yeni Doğanların Yoğun Bakım İhtiyacının Makine Öğrenmesi Teknikleriyle Tahmin Edilmesi 39
4 SONUÇ VE ÖNERİLER 46
KAYNAKÇA **48**
TEZDEN ÜRETİLMİŞ YAYINLAR **50**

KISALTMA LİSTESİ

AUC	(Area Under Curve) Eğri altında kalan
BMI	(Body Mass Index) Vücut Kitle Endeksi
BW	(Birth Weight) Doğum Ağırlığı
DVM	Destek Vektör Makineleri
GA	(Gestational Age) Gebelik haftası
ICD	(International Statistical Classification of Diseases and Related Health Problems) Uluslararası İstatistiksel Hastalık Sınıflandırması ve İlgili Sağlık Sorunları
PI	(Ponderal Index) Ponderal Endeksi
ROC	(Receiver Operator Characteristic) Alıcı Karakteristik Operatör
YYBÜ	Yeni Doğan Yoğun Bakım Ünitesi

ŞEKİL LİSTESİ

Şekil 2.1 Veri Birleştirilmesi 8
Şekil 2.2 Makine Öğrenmesi Sınıflandırmaları 9
Şekil 2.3 Çizilebilecek Hiper-Düzlemler 11
Şekil 2.4 En Uygun Hiper-Düzlem 12
Şekil 2.5 Hiper-Düzlemin Belirlenmesi 12
Şekil 2.6 Doğrusal Ayrılamayan Veri Seti için Hiper-Düzlemin Belirlenmesi 13
Şekil 2.7 Optimum Hiper -Düzlemin Belirlenmesi 13
Şekil 2.8 Kernel Fonksiyonu ile Verinin Boyutunun Yükseltilmesi 14
Şekil 2.9 Karar Ağacı Yapısı 14
Şekil 2.10 En Yakın K Komşu Algoritması Kolay Gösterimi 18
Şekil 2.11 Biyolojik Sinir Hücresi ve Yapay Sinir Ağı 19
Şekil 2.12 Yapay Sinir Hücresi 20
Şekil 2.13 AdaBoost Algoritmasının Basit Gösterimi 20
Şekil 2.14 K-Ortalama Kümeleme Algoritması Gösterimi (K=3 için) 22
Şekil 3.1 Anne yaş ve Kan Parametrelerine ait Kutu Diyagramları 35
Şekil 3.2 Kuadratik çekirdeğe sahip destek vektör makineleri için ROC eğrisi ve AUC değeri 41
Şekil 3.3 Medium Gaussian çekirdeğe sahip destek vektör makineleri için ROC eğrisi ve AUC değeri 41
Şekil 3.4 Gentle AdaBoost yöntemi için ROC eğrisi ve AUC değeri 42
Şekil 3.5 Karar Ağacı Yöntemine göre Özniteliklerin Etki Analizi 43

TABLO LİSTESİ

Tablo 2.1 Gini Hesabı için Bölmeleme Örneği 15
Tablo 2.2 Naive Bayes Verisi 16
Tablo 2.3 Cinsiyet Özelliğine Göre Sayılar ve Olasılıklar 17
Tablo 2.4 Biyolojik Sinir Sistemi Elemanları ve Yapay Sinir Sisteminde Karşılıkları 19
Tablo 2.5 Karmaşıklık Matrisi 24
Tablo 3.1 Sonuç veri setinin içerdiği öznitelikler 31
Tablo 3.2 Kontrol ve Hasta Sınıfa ait Nümerik Öznitelikler için Ortalama ve Standart Sapma Değerleri 33
Tablo 3.3 Kontrol ve Hasta Sınıfa ait Kategorik Öznitelikler için Adet ve Oranları 34
Tablo 3.4 Kontrol ve Hasta Sınıfa ait Nümerik Öznitelikler için p değerleri 37
Tablo 3.5 Kontrol ve Hasta Sınıfa ait Kategorik Öznitelikler için p değerleri 38
Tablo 3.6 Makine Öğrenmesi Algoritmalarında Kullanılan Maliyet Tablosu 39
Tablo 3.7 Makine Öğrenmesi Yöntemleri için Duyarlılık-Özgünlük Değerleri Tablosu 40
Tablo 3.8 AUC Değerleri için Aralıklar 40
Tablo 3.9 En Başarılı Yöntemler için Karmaşıklık Matrisleri 42
Tablo 3.10 ICD Kodlarına Göre Elde Edilen Doğruluk Değerleri 44
Tablo 3.11 Farklı Şikayet Türlerine Göre Elde Edilen Doğruluk Değerleri 45

ÖZET

Yeni Doğanların Yoğun Bakım İhtiyacının Makine Öğrenmesi Teknikleri ile Doğum Öncesinde Tahmin Edilmesi

Serkan GESOĞLU

Matematik Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Birol ASLANYÜREK

Eş-Danışman: Doç. Dr. Emrah AYDIN

Gebelik döneminde annede, bebekte ya da her ikisinde birlikte meydana gelen çeşitli sağlık sorunları doğum sonrasında bebeklerin yenidoğan bakım ünitesine yatmalarına neden olabilmektedir. Doğum öncesi, anne veya bebekteki bazı belirtiler bebeğin doğum sonrası yoğun bakım ihtiyacı konusunda fikir verse de bazı durumlarda doğum öncesi belirgin bir belirti olmamasına rağmen bebeklerin yoğun bakım ihtiyacı olmaktadır. Bu belirtilerden birçoğunun kişinin kan parametrelerinde değişime neden olduğu uzun zamandır bilinmektedir. Ayrıca, anne karnında bebek ile anne arasındaki madde alışverişi bebekteki bazı sağlık sorunlarının annenin kan parametrelerini değiştirme potansiyelini ortaya koymaktadır. Bu çalışmanın amacı, doğum öncesi annenin tam kan sayım parametreleri, anne yaşı ve temel bazı şikayetleri ile bebeğin doğum sonrası yoğun bakım ihtiyacının olup olmayacağı arasındaki ilişkiyi istatistiksel olarak analiz etmek ve makine öğrenmesi yöntemleriyle yenidoğanların yoğun bakım ihtiyacının prenatal tahmin etmektir.

Yapılan çalışmanın ilk aşamasında; 2014-2019 yılları arasında İstanbul Medipol Hastanesine başvuran annelerin bilgileri, annelerden alınan tam kan sayım verileri, yoğun bakıma giren bebek bilgileri toplanmıştır. Bu ham verilere temizleme, dönüştürme ve birleştirme veri ön işlemeleri uygulandıktan sonra, 41.387 gebe anne ve 3.361 yoğun bakıma yatan yenidoğan verisi betimleyici ve çıkarımsal istatistik yöntemleriyle analiz edilmiştir. Yapılan çalışmada p değerinin 0,05'ten küçük olduğu değerler anlamlı olarak kabul edilmiştir. Bebeği yoğun bakıma yatan ve yatmayan annelerin yaş ortalaması ve kan sayımı parametrelerinden RBC, HTC, MCH, MCV, MPV, PCT ve RDW ortalamaları için anlamlı istatistiksel farklar bulunmuştur. Çalışmanın son aşaması ise makine öğrenmesi teknikleriyle yenidoğanların yoğun bakım ihtiyacının doğum öncesi tahmin edilmesine ayrılmıştır. Bu kapsamda, verilerin %80'i eğitim, %20'si test verisi olarak alınmıştır. Eğitim verisine 10 katlamalı çapraz doğrulama uygulanarak makine öğrenmesi modelleri eğitilmiştir. Karar ağaçları, destek vektör makineleri, Naive Bayes yöntemleri ve topluluk sınıflandırıcıları gibi makine öğrenmesi teknikleri kullanılarak elde edilen sınıflandırma sonuçları analiz edilmiştir. Bebeklerin yoğun bakıma yatma nedenlerinin çeşitliliği düşünüldüğünde veri sayısının artırılması ve diğer klinik verilerin dahil edilmesi başarı oranlarının artırılması hususunda ümit vermektedir.

Anahtar Kelimeler: Gebelik, Yenidoğan Yoğun Bakımı, Tıpta Makine Öğrenmesi

ABSTRACT

Predicting Intensive Care Needs of Newborns Prenatal by Machine Learning

Serkan GESOĞLU

Department of Mathematical Engineering

Master of Science Thesis

Supervisor : Assoc. Prof. Dr. Birol ASLANYÜREK

Co-Supervisor: Assoc. Prof. Dr. Emrah AYDIN

Various health problems that occur in the mother, baby or both of them during pregnancy cause babies to be admitted to the neonatal care unit after birth. Although some symptoms in the prenatal, mother or baby give an idea about the possibility of the baby's need for intensive care after birth, in some cases babies need intensive care despite the lack of a clear prenatal symptom. It has long been known that many health problems cause changes in a person's blood parameters. In addition, the exchange of substances between the baby and the mother in the womb reveals the potential for certain health problems in the baby to change the blood parameters of the mother. The aim of this study is to statistically analyze the relationship between the complete blood count parameters, maternal age, and some basic complaints of the prenatal mother and whether the baby will need postpartum intensive care and to predict the intensive care needs of newborns prenatally with various machine learning methods.

In the first stage of the study, between 2014 and 2019, the information of mothers admitted to Istanbul Medipol Hospital, complete blood count data from mothers,

and baby entering intensive care unit information were collected. After applying preliminary data operations such as cleaning, conversion and combining these raw data, 41.387 pregnant mothers and 3.361 intensive care unit neonatal data were analyzed by descriptive and inferential statistical methods. p values less than 0,05 were considered significant. When the differences in feature means belonging to the mothers whose babies were hospitalized and were not hospitalized in the intensive care unit were statistically tested, significant statistical differences were determined for maternal age and the parameters RBC, HCT, MCH, MCV, MPV, PCT and RDW in the blood count. The final stage of the study is devoted to the prenatal estimation of the intensive care needs of newborns via various machine learning techniques. In this context, 80% of the data was taken as training and 20% as test data. Machine learning models were trained by applying 10-fold cross-validation to the training data. The classification results obtained using machine learning techniques such as decision trees, support vector machines, Naive Bayes methods and ensemble classifiers were analyzed. When we consider the diversity of reasons for infants to be admitted to an intensive care unit, increasing the number of data or including other clinical features in the data set promises to increase the success rates.

Keywords: Pregnancy, Neonatal Intensive Care, Machine Learning in Medicine

1
GİRİŞ

1.1. Literatür Özeti

Bebeklerin doğum öncesi yaşamsal fonksiyonları anne tarafından sağlanır. Bu yaşamsal fonksiyonlar doğum sonrası yavaş yavaş bebeğin kendisi tarafından gerçekleştirilmeye başlar. Bu süreçte zaman zaman bu yaşamsal fonksiyonlarda sorunlar yaşandığı bilinmektedir. Bu sorunlar annenin hamilelik döneminde yaşadığı sorunlara bağlı olabildiği gibi bebeklerin gelişimlerinde ortaya çıkan sorunlara da bağlı olabilir. Yaşanan sorunlar sebebiyle doğum erken başlayabilir veya bebek anne karnından alınmak durumunda kalabilir. Bu şekilde doğan bebeklere de prematüre denir. Yaşamsal fonksiyonların bebeklerin kendileri tarafından gerçekleştirilmeye başlandığı bu geçiş döneminde bazı bebeklerin zorlandığı literatürde bildirilmiştir. Bebeklerin sağlıklı bir şekilde hayatına devam edebilmesi için yeni doğan yoğun bakım ünitelerinde (YYBÜ) takibi güncel uygulamada kabul edilen yöntemdir.

Yenidoğan yoğun bakım ünitelerinde bebeklerin hastalıklarına yönelik bakımların dışında günlük bakımları da gerçekleştirilmektedir. Bu ünitelerde bebeklerin sahip oldukları sorunlar nedeni ile tetkik ve tedavi süreleri farklılık gösterebilmektedir. Yeni doğan bebeklerin YYBÜ'ye alınmasına neden olan faktörler üzerine birçok araştırma yapılmıştır. YYBÜ'de yatış sürelerine göre yapılan bir incelemede doğum ağırlığı, vücut kitle indeksi ve ponderal indeks parametreleri ele alınmıştır [1]. Bebeğin YYBÜ'de yatma süresinde doğum ağırlığının diğer parametrelerden daha etkili olduğu gözlemlenmiştir.

Jennifer L. Barkin ve ekibinin yaptığı çalışmada; üçüncü seviye olarak YYBÜ'ye kabul edilmiş, evine canlı olarak taburcu edilen ve en az 6 gün YYBÜ'de kalan 146 anne-bebek ikilisi üzerinde inceleme yapılmıştır [2]. Annenin fonksiyonel durumu ile depresif belirti oranı arasındaki anlamlı ilişki literatür tarafından desteklenmektedir. Doğum sonrası artan depresyonu gestasyonel diyabetle ilişkilendiren mevcut literatür göz önüne alındığında, bebek hipoglisemisi teşhisi ve annenin bebekle ilişki düzeyine ait sonuçların ilgi çekici olduğu görülmüştür.

Gebelik sonrasında bebekleri yoğun bakıma yatan annelerle bebeklerin ilişki bozukluğu yaşadığı görülmüştür.

Şişli Etfal Hastanesi doktorları tarafından hastanenin yenidoğan ünitesinde yatan bebekler üzerinde yapılan incelemede YYBÜ'ye yatırılarak izlenen ve izleme sırasında kaybedilen bebeklerin demografik özellikleri sunulmuştur [3]. Çalışma süresince YYBÜ'ye yatan bebeklerin %3'ünün kaybedildiği saptanmıştır. Kaybedilen bebeklerin %15,6'sının ilk 24 saatte, %74,9'unun ise ilk 7 gün içinde kaybedildiği görülmüş ve anne yaşının 19 yaş altı olma oranının %2,3 ve 35 yaş üstü olma oranının %14,3 olduğu tespit edilmiştir. Bu inceleme sonucunda; immatürite (olgunlaşmamış) ve konjenital (doğumdan gelen) anomalilerin günümüzde en sık görülen yenidoğan ölüm nedenleri arasında olduğu tespit edilmiştir.

Japonya'da yapılan bir çalışmada [4]; 2013-2017 yılları arasında doğan, doğum ağırlığı 2000 gramdan yüksek, gebelik haftası 36 haftadan büyük olan ve doğuştan solunum sıkıntısı sebebiyle oksijene ihtiyaç duyan yenidoğanlarda perinatal faktörler (doğum öncesi ve doğumdan sonraki ilk dört haftadaki annenin sağlık durumu) ve prediktif faktörler (kanser hücrelerinin özellikleri) değerlendirilmiştir. Bu analizin sonucunda; maternal anemi, kısırlık tedavisi, erken doğum tehdidi, erken membran reptürü, rahim ameliyatı geçmişi ve ikiz bebek sahibi olmanın YYBÜ'ye kabullerin önemli belirleyicileri olduğu görülmüştür. Klinisyenlerin, solunum sıkıntısı yaşayan yenidoğanları tedavi ederken bu faktörleri göz önüne almaları önerilmiştir.

Özdemir ve arkadaşlarının İstanbul Medicine Hastanesinde yapmış olduğu araştırmada, 2012 – 2014 yılları arasında YYBÜ'ye yatırılan ve izlenimleri sırasında kaybedilen bebekler incelenmiştir. YYBÜ'ye yatırılan bu bebeklerin ölüm oranları, risk faktörleri ve ölüm nedenleri belirlenmiştir [5]. Yeni doğan bebeklerdeki ölüm sebeplerinin önemli bir kısmının prematüre doğum, enfeksiyon ve konjenital anomalilere bağlı olduğu izlenmiştir. Tüm ölümlerin %75'ini 2500 gramın altındaki bebeklerin oluşturduğu gözlenmiştir. Bu nedenle önlenebilir ölüm nedenlerini ve oranlarını azaltmak için gebelik takip programlarının yaygınlaştırılması, yeterli antenatal bakımın sağlanması, doğum ve doğum sonrası bakımın uygun koşullarda sağlanması gerekliliği görülmüştür.

Yoğun bakım ihtiyacı olan bebeklerin doğumunun mutlaka YYBÜ'ye sahip bir hastanede gerçekleşmesi teşhis ve tedavi açısından önemlidir. Annenin gebelik anında ya da bebek doğduktan sonra çok belirgin olan hastalıklara (erken doğum, yemek borusunun olmaması, vb.) sahip olması durumunda bebeğin YYBÜ'ye gireceği bilinmektedir. Ancak sağlıklı gittiği görülen bir gebelikte çocuğun yoğun bakıma girip girmeyeceğini belirleyebilecek parametrelerin hala net olmadığı görülmektedir.

1.2. Tezin Hipotezi ve Amacı

Sunulmakta olan tezin hipotezi gebelik esnasında anneden alınan tam kan sayım parametreleri ile doğum sonrasında bebeklerin yoğun bakım ünitesine yatma ihtiyacının öngörülebileceğidir. Bu hipotezi gerçekleştirmek için çalışmanın ilk amacı geriye dönük olarak toplanan verilerin temizlenmesi ve anonimleştirilmesidir. Bu sayede veriler eşleştirmeye uygun hale getirilebilecektir. Bu ön işlemeyi takiben çalışmanın ikinci amacı farklı dosyalar halinde alınan verilerin tek dosyada birleştirilmesidir. Bu birleştirme ile yapısal bir formata dönüştürülmüş olan veri seti analize uygun hale gelecektir. Veri setinin yapısallaştırılmasının ardından çalışmanın üçüncü amacı anne kan parametreleri ile bebeklerin yoğun bakım yatma ihtiyacı arasındaki ilişkiyi makine öğrenmesi teknikleriyle analiz edebilmektir.

1.3. Orijinal Katkı

Yapılan literatür araştırması sonucunda doğum öncesi annenin kan parametreleri, yaşı ve şikayetlerini kullanarak bebeğin YYBÜ'ye yatıp yatmayacağını tahmin etmeye çalışan bir araştırmanın bulunmadığı tespit edilmiştir. Bu çalışmada; yukarıda belirtilen parametreler ile bebeklerin doğum öncesi YYBÜ'ye yatma oranı arasındaki ilişki ele alınmıştır. Ayrıca makine öğrenmesi teknikleri kullanılarak bebeğin yoğun bakıma girip girmeyeceği doğum öncesinden tahmin edilmeye çalışılmıştır. Elde edilen başarı oranlarının analizi sonucunda; bebeklerin yoğun bakıma yatma nedenlerinin çeşitliliği düşünüldüğünde, veri sayısının artırılması ve diğer klinik bilgilerin veri setine dahil edilmesi başarı oranlarının artırılması hususunda ümit vermektedir.

2
MAKİNE ÖĞRENMESİ VE KAVRAMLARI

Veri, işlenmemiş bilgidir. Bu bilgiler aralarında bir ilişki olması durumunda birbiriyle anlamlandırılabilir ve birleştirilebilir. Bu anlamlandırılan verileri belirli amaçlar doğrultusunda işleyerek bilgiye erişmek mümkündür. Veri madenciliği ise, bilgiyi elde etmek ve bilgiyi üretmek için verileri bir araya getirme ve veriyi araştırma yöntemidir.

Günümüzde bilgisayar alanındaki gelişmelerle verilerin uzun süreli depolanması, böylece büyük kapasiteli veri tabanlarının oluşması başlamıştır. Bu veri tabanlarından anlamlı ve doğrudan görülmeyen bilgileri elde etmek için yeni bir sistem oluşturulmuştur. Verilerin istatiksel yöntemlerle analiz edilerek karar verme sürecinin etkinliğinin artırılması ve yeni stratejiler geliştirilmesine yol göstermesi hedeflenmektedir [5].

Veri madenciliği yöntemiyle elde edilen sonuçların analiz edilmesine yardımcı olan teknik makine öğrenmesi olarak karşımıza çıkar. Makine öğrenmesi yapay zeka alanında geliştirilen teknikler arasında yer alır. Bununla birlikte yapay zeka ile karışmaması gereken bir kavramdır. Bu teknikler sayesinde veri madenciliği çok daha önemli ve etkili hale gelir.

Makine Öğrenmesi, yapay zekânın bir dalıdır ve matematiksel ve istatistiksel işlemler ile veriler üzerinden çıkarımlar yaparak tahminlerde bulunan sistemlerin bilgisayarlar ile modellemesidir. Veri setlerinden örüntülerin bulunmasını ve çeşitli algoritmalarla belirli bir mantık oluşturulmasını sağlar. Mevcut veri kümeleri ve kullanılan algoritmalar ile oluşturulan model, en yüksek performansı vermek üzere kurulmaktadır. Bu nedenle birçok makine öğrenmesi tekniği geliştirilmiştir. Bunlardan bazıları; k-en yakın komşu algoritmaları, Naive Bayes sınıflandırıcıları, karar ağaçları, lojistik regresyon analizi, k-ortalama algoritmaları, destek vektör makineleri (DVM) ve yapay sinir ağlarıdır. Bu yaklaşımların bazılarıyla tahmin ve kestirim, bazılarıyla da kümeleme ve sınıflandırma yapabilmektedir.

Makine öğrenmesi stratejileri; denetimli ve denetimsiz olmak üzere iki grupta analiz edilir. Denetimli öğrenmede oluşturulan model ile, bir dizi girdi değeri oluşturulur. Bu girdi değerlerine karşılık hedef değerleri verilerek aralarındaki ilişkiyi öğrenmesi ve hedef değerlere en yakın çıktıların üretilmesi amaçlanmaktadır. Elde edilen en iyi model, yeni girdi değerinin bir sonraki çıktısını da verebilecektir. Denetimsiz öğrenmede, hedef değerler olmadan sadece girdi değerleri arasındaki ilişki ortaya çıkarılmaya çalışılır. Birbirine yakın değerler bu ilişkiler kullanılarak gruplandırılır, bu şekilde kümelenme oluşacaktır. Yeni girdi bu kümelerden hangisiyle ilişkili ise o kümeye aittir.

Aşağıda sırasıyla verilerin düzenlenmesi, makine öğrenmesi algoritmaları ve makine öğrenmesi algoritmasına karar verme işlemleri açıklanmaktadır.

2.1 Verilerin Düzenlenmesi

Verilerin düzenlenmesi aşaması kendi içerisinde temizleme, birleştirme, dönüştürme ve indirgeme adımlarından oluşmaktadır. Modelin kurulması aşamasında ortaya çıkabilecek sorunlara çözüm getirmek için veriler öncesinde modele uygun olarak düzenlenir.

2.1.1 Veri Temizleme

Veri temizleme; eksik verilerin giderilmesi, olmaması gereken verilerin bulunması ve tutarsızlıkların giderilmesi amacıyla yapılan işlemdir. Bu işlemin yapılabilmesi için bazı yöntemler bulunmaktadır. Örnek olarak; eksik değer içeren kayıtların silinmesi, eksik değerlerin yerine değişkenin ortalamasının eklenmesi, elimizdeki değerlere göre en uygun değerin eksik değişkenin yerine yazılması verilebilir.

Ayrıca veri temizlemenin kullanılması gereken bir diğer problem ise gürültülü verilerin düzeltilmesidir. Gürültü, ölçülen değişkendeki varyans veya rassal bir hatadır. En sık kullanılan veri temizleme tekniklerinde üçü aşağıda verilmiştir [7]:

2.1.1.1 Paketleme (Binning)

Paketleme yöntemleri sıralanmış verileri düzeltmek (gürültüden ayıklamak) için kullanılır. Bu yöntemde ilk olarak sıralanmış veriler eşit büyüklükte paketlere

(gruplara) ayrılır. Daha sonra paketler, paket ortalamaları veya paket sınırları yardımıyla düzeltilir. Örneğin değerler 5, 7, 15, 21, 21, 21, 25, 33 ve 35 olsun.

Paket 1: 5, 7, 15

Paket 2: 21, 21, 21

Paket 3: 25, 33, 35 olacaktır.

Paket ortalaması yardımı ile düzeltme yapılırsa,

Paket 1: 9, 9, 9

Paket 2: 21, 21, 21

Paket 3: 31, 31, 31 olacaktır.

Paket sınırları yardımı ile düzeltilme yapılırsa,

Paket 1: 5, 5, 15

Paket 2: 21, 21, 21

Paket 3: 25, 35, 35

2.1.1.2. Kümeleme

Aykırı değerler kümeleme yöntemi ile elde edilir. Aynı kümenin içine benzeyen değerler alınır. Aykırı değerler grubun dışında kalır.

2.1.1.3. Regresyon (Verilerin Düzeltmesi)

Regresyon verilerin düzeltmesi veya veri grubunun gürültüden arındırılması olarak ifade edilir. Veri seti regresyon yöntemi ile bir fonksiyon yazılarak düzeltilebilir. Bu fonksiyon veri değerleri kullanılarak belirlenir ve fonksiyona uymayan noktalar aykırı değerler olarak grubun dışında kalır. Veri setindeki tutarsızlıklar veri düzeltme ihtiyacının doğmasına neden olan bir diğer hususdur. Bu tutarsızlıklar verilerin birleştirilmesinden kaynaklı olabilmektedir.

2.1.2 Veri Birleştirme

Veri birleştirmede iki ya da daha fazla tablo ortak kolonlar üzerinden birleştirilebilir. Birleştirilirken şema birleştirme sorunları oluşabilir. Bu sorunları ortadan kaldırmak için meta veriler kullanılır. Veri tabanları ve veri ambarları genellikle meta veriye sahiptirler. Meta veri, veriye ilişkin veridir. Bir değişken başka bir tablodan üretilmişse fazlalık olabilir, bu yüzden veri birleştirmede indirgeme de yapılabilir [7].

2.1.3 Veri Dönüştürme

Veri dönüştürme verilerin uygun formlara dönüştürülmesini sağlar. Veri dönüştürme düzeltme, birleştirme, normalleştirme ve genelleştirme gibi işlemleri içerebilir. Veri normalleştirme en çok kullanılan işlemlerden birisidir ve üç başlık altında incelenebilir [7].

2.1.3.1. Min-Max Normalleştirmesi

Min-Max normalleştirmesi verileri doğrusal dönüşüm ile yeni bir veri aralığına yerleştirmeyi sağlar. Bu aralık genellikle [0,1] olarak alınmaktadır. Aşağıdaki formül aracılığıyla; $[\min_A, \max_A]$ aralığında değişen bir v değeri, $[\text{new\_min}_A, \text{new\_max}_A]$ aralığına göre normalize edilir.

$$v' = \frac{v - \min_A}{\max_A - \min_A}(\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A \tag{2.1}$$

2.1.3.2. Z – Puanı Normalleştirmesi

Z puanı bir veri noktasının, tüm veri setinin ortalamasından kaç standart sapma kadar uzak olduğunu gösterir. Z-puanı ayrıca standart puan olarak da bilinir ve normal bir dağılım eğrisine yerleştirilebilir. Temel Z skoru formülü;

$$z_i = \frac{x_i - \overline{x}}{s} \tag{2.2}$$

şeklindedir.

2.1.3.3. Ondalık Ölçekleme Normalleştirmesi

Ondalık ölçekleme ile normalleştirmede ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme işlemi yapılır. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçeklemenin formülü aşağıdaki şekildedir:

$$v' = \frac{v}{10^j}, \qquad j: Max(|v'|) < 1 \text{ olacak şekilde en küçük tam sayı} \tag{2.3}$$

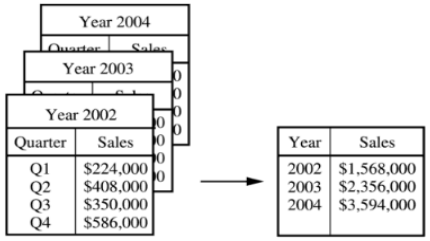
2.1.4 Veri İndirgeme

Veri madenciliği uygulamalarında çözümlemeden elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir. Veri

indirgeme yöntemleri; veri birleştirme, boyut indirgeme, veri sıkıştırma ve kesikli hale getirmedir.

2.1.4.1. Veri Birleştirme

Daha karmaşık modeller oluşturmak ve bir proje hakkında daha fazla bilgi edinmek için birden çok kaynaktan veri alma işlemidir. Genellikle tek bir konuda veri toplamak ve merkezi analiz için birleştirmek anlamına gelir.



Şekil 2.1 Verilerin Birleştirilmesi [8]

2.1.4.2. Boyut İndirgeme

Çok boyutlu verilerde sistemi beslemek daha zor olduğu için Temel Bileşen Analiziyle fazla kayıp verilmeden boyut ikiye indirilebilir. Özellikle verilerin görselleştirilmesi isteniyorsa çok boyutlu verilerde boyut indirgeme kullanılır.

2.1.4.3. Veri Sıkıştırma

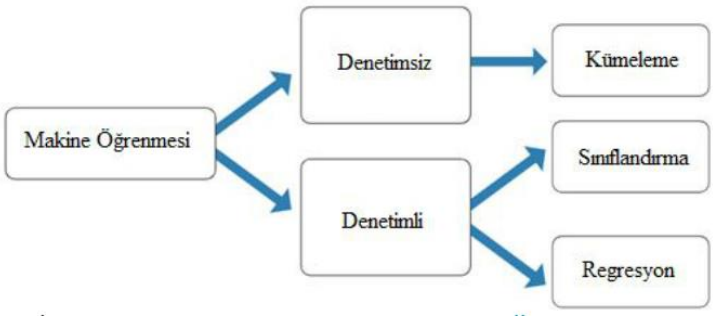
Bu yöntemde veriler veri şifreleme veya dönüşümü ile sıkıştırılır. Bilgi kaybı olan veri sıkıştırmanın yanı sıra bilgi kaybı olmadan da sıkıştırmak için bazı algoritmalar vardır.

2.1.4.4. Kesikli Hale Getirme

Yalnızca kategorik değerleri ele alan makine öğrenmesi algoritmaları sürekli verilerin kesikli değerler haline getirilmesini sağlar. Elde edilen bu değerler orijinal verinin yerine kullanılırlar.

2.2 Makine Öğrenmesi Yöntemleri

Makine öğrenmesi denetimli (gözetimli) ve denetimsiz (gözetimsiz) olarak ikiye ayrılır.



Şekil 2.2 Makine Öğrenmesi Sınıflandırmaları

Bu tezde yapılan çalışmada denetimli öğrenme yöntemleri kullanıldığı için, bu yöntemler ana hatlarıyla aşağıda ele alınmıştır. Ayrıca, denetimsiz öğrenme yöntemlerinden kısaca bahsedilmiştir.

2.2.1 Denetimli Öğrenme Yöntemleri

Denetimli makine öğrenimi, belirsizlik dahilinde kanıta dayalı tahminler yapan bir model oluşturur. Denetimli öğrenme algoritması bilinen bir girdi verisi ve etiketlenmiş yanıtları alır, ardından yeni verilere etiketleme için makul tahminler oluşturmak üzere bir modeli eğitir.

Denetimli Öğrenme için Temel Tanımlar

- Eğitim Verisi:

Makine öğrenmesi sürecinde modelin eğitilmesi için ayrılan veri seti eğitim veri seti olarak adlandırılır. Daha sonra bir tahminleme yapılacak olması durumunda bu veriler kullanılmaktadır.

- Test Verisi:

Veri setinin bir kısmı eğitim verisi olarak seçilip, makine öğrenmesi modeli eğitildikten sonra verinin diğer kısmı tahminlerin doğruluğunu test etmek için kullanılır ve bu nedenle test verisi olarak adlandırılır. Genelde, veri setinin %70, %80 ya da % 90 gibi bir oranı eğitim seti için ayrılırken, test için %30, %20 ya da %10'luk bir kısım ayrılır. Makine öğrenimi tamamladıktan sonra bu test verisiyle test yapılarak doğruluk oranının belirlenmesi amaçlanır.

Denetimli öğrenme problemleri ikiye ayrılır;

1. Regresyon: Çıktı değerleri belli aralıkta değişen sürekli sayısal değerlerdir. Bu nedenle bu tipteki makine öğrenmesi yöntemleri sürekli değerleri tahmin edebilecek algoritmalara sahiptir.
2. Sınıflandırma: Çıktı değerleri belli sayıda kategorik değerden oluşur ve bu tipteki makine öğrenmesi yöntemleri ayrık değerler üretebilecek algoritmalara sahiptir.

2.2.2 Denetimsiz Öğrenme Yöntemleri

Bu öğrenme de sadece girdi bulunmaktadır. Yani etiket bilgisi içermez. Bu öğrenmedeki amaç verileri daha iyi anlamak ve dağılımları modellemek, yapıyı öğrenmektir. Örnek olarak kümeleme ve ilişkilendirme kuralları verilebilir.

2.3 Denetimli Makine Öğrenmesi Algoritmaları

Bu bölümde en yaygın kullanılan denetimli makine öğrenmesi algoritmaları anlatılmaktadır.

2.3.1. Lineer Regresyon

Genel olarak farklı bilim dallarının da kullandığı lineer (doğrusal) regresyon yöntemi istatistiksel yöntemlerdendir. Lineer regresyon değişkenlerin arasında bulunan ilişkinin modellenmesini sağlayan bir tekniktir. 'Çoklu regresyon' ve 'basit regresyon' olarak ikiye ayrılır. Basit regresyonda bir bağımlı, bir bağımsız değişkene sahip problemler modellenir. Bağımsız olanlar x vektörü ile, bağımlı olanlar ise y vektörü ile gösterildiğinde, değişkenler arasındaki ilişki aşağıdaki denklemle ifade edilir.

$$y = \beta_0 + \beta_1 x + e \tag{2.4}$$

Burada e hata değeri, β_0 kesişim noktası, β_1 ise x'in regresyon katsayısıdır.

Çoklu regresyonda ise birden çok bağımsız değişken bulunur ve

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + e \tag{2.5}$$

şeklinde ifade edilebilir.

Doğrusal regresyon daha basit ifadeyle noktalar arasındaki ilişkiye bakılarak bu noktalardan geçen en uygun doğruyu bulma yöntemidir. Bu doğruya regresyon doğrusu denir [10].

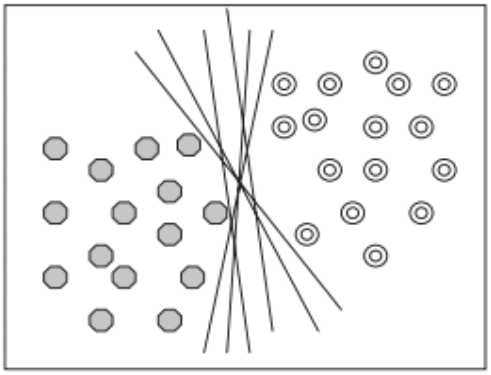
2.3.2. Destek Vektör Makineleri (DVM)

Destek vektör makineleri ile sınıflandırma problemlerinin çözümü başarılı bir şekilde gerçekleştirilebilir. Genelleme performansı yüksek ve etkin makine öğrenimi algoritmalarındandır. Sınıflandırma problemini kareli optimizasyon problemine dönüştürüp çözer. Bu durum öğrenme kısmındaki işlem sayısını azaltır ve algoritmanın hızlanmasını sağlar [10].

Destek vektör makineleri temel olarak iki sınıfı ayırabilen hiper düzlemin belirlenmesi prensibine dayanmaktadır [11].

2.3.2.1. Doğrusal Ayrılabilen Veriler için DVM

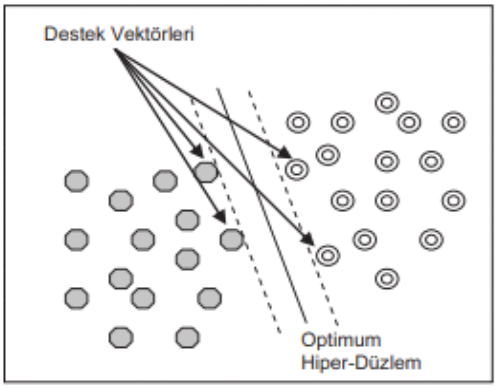
Eğitim verisiyle bir karar fonksiyonu belirlenir. Genelde {+1,-1} sınıf etiketleri yardımıyla iki sınıfa ait örnekleri bu karar fonksiyonu ile birbirinden ayırır. Bu ayırma için en uygun hiper-düzlem bulunur [12].



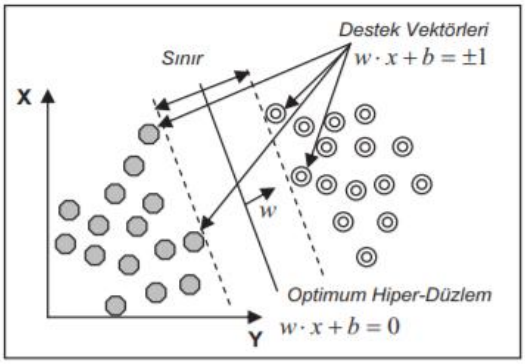
Şekil 2.3 Çizilebilecek Hiper-Düzlemler [12]

DVM'nin amacı kendisine en yakın noktaların arasındaki uzaklığın maksimum olması için kullanılacak hiper-düzlemin belirlenmesidir. En uygun ayrımın yapıldığı hiper-düzlem yani optimum hiper düzlem Şekil 2.4'te görülmektedir. Sınır genişliğini sınırlandıran noktalara da destek vektörleri adı verilir [12].

Burada x ∈ R^N olup N-boyutlu bir uzayı, y ∈ {-1, +1} ise sınıf etiketlerini, w ağırlık vektörünü (hiper-düzlemin normali) ve b eğilim değerini göstermektedir. Optimum hiper-düzlemin belirlenebilmesi için bu düzleme paralel ve sınırlarını oluşturacak iki hiper-düzlemin belirlenmesi gerekir (Şekil 2.5). Bu hiper-düzlemleri oluşturan noktalar destek vektörleri olarak adlandırılır ve bu düzlemler $w \cdot x + b = \pm 1$ şeklinde ifade edilirler.



Şekil 2.4 En Uygun Hiper-Düzlem [12]



Şekil 2.5 Hiper-Düzlem Belirlenmesi [12]

$\|w\|$ değerinin minimum olarak belirlenmesi hiper-düzlemin sınırının maksimum olmasını sağlar. Bu durum kısıtlı optimizasyon problemi ile çözülür.

$$min\left[\frac{1}{2}\|w\|^2\right] \tag{2.6}$$

Kısıtlar;

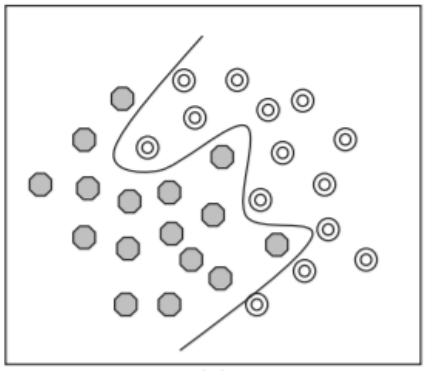
$$y_i(w.x_i + b) - 1 \geq 0 \text{ ve } y_i \in \{1, -1\} \tag{2.7}$$

Lagrange Denklemi ile çözülür ve bir karar fonksiyonu elde edilir. Bu karar fonksiyonu (2.8)'de gösterilmiştir.

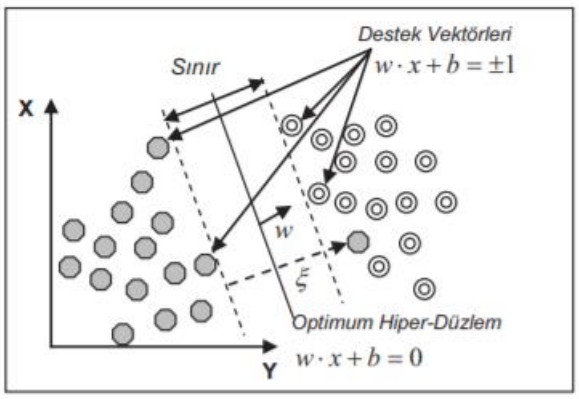
$$f(x) = sign\left(\sum_{i=1}^{k} \lambda_i y_i \,(x.x_i) + b\right) \tag{2.8}$$

2.3.2.2. Doğrusal Ayrılamayan Veriler için DVM

Verilerin doğrusal olarak ayrılamadığı durumlarda kullanılır. Bu durumda, eğitim verilerinden bazılarının optimum hiper-düzlemin diğer tarafında olmasından dolayı gerçekleşen bu problem pozitif yapay bir değişkenin (ξ_i) tanımlanması ile ortadan kaldırılır.



Şekil 2.6 Doğrusal Ayrılamayan Veri Seti için Hiper-Düzlemin Belirlenmesi [12]



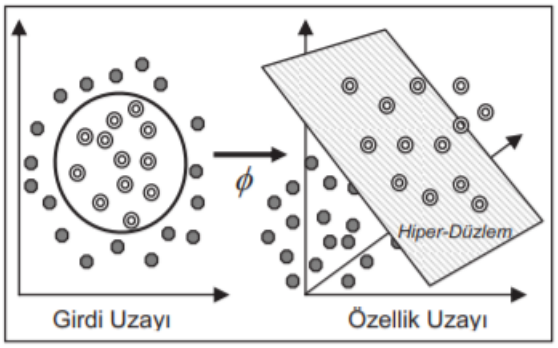
Şekil 2.7 Optimum Hiper-Düzlemin Belirlenmesi [12]

Bir Kernel fonksiyonu ile verilerin yüksek boyutta doğrusal ayrılabilmesi sağlanır.

$$K(x_i, x_j) = \varphi(x).\varphi(x_j) \tag{2.9}$$

Bu durumda karar fonksiyonu 2.10'daki gibidir.

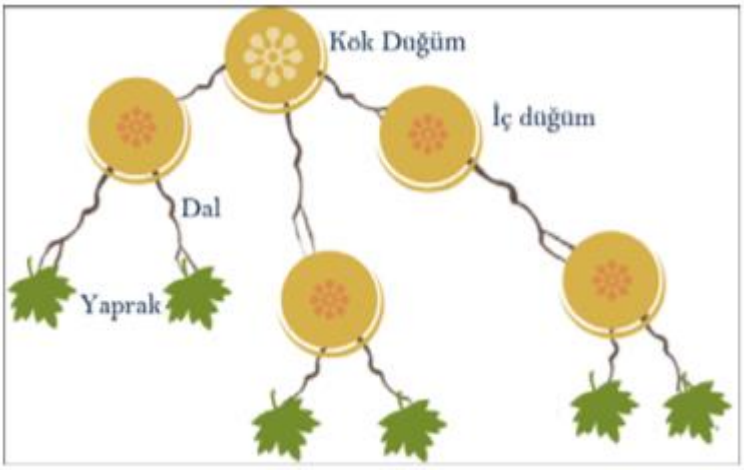
$$f(x) = sign\left(\sum_i \alpha_i y_i \, \varphi(x).\varphi(x_i) + b\right) \tag{2.10}$$



Şekil 2.8 Kernel Fonksiyonu ile Verinin Boyutunun Yükseltilmesi [12]

2.3.3 Karar Ağaçları

Karar ağaçları genelde sınıflandırma problemlerinin çözümünde kullanılan bir denetimli öğrenme algoritmasıdır. Aynı zamanda karar ağaçları regresyon problemlerinde de kullanılabilir. Özniteliklerdeki veriler sürekli ise regresyon ağacı oluşturulabilir. Basit bir yapısı vardır. Yaprak, dal ve düğüm elemanlarından oluşurlar. Karar ağaçlarında ID3, C4.5, Random Forest gibi birçok algoritma vardır. Random Forest diğerlerinden farklı olarak büyük verilerle çalıştırıldığında oldukça iyi performans gösterir. Bu algoritma farklı eğitim kümelerinin eğitilmesiyle oluşturulmuş birden çok karar ağacının birleştirilmesiyle oluşturulur.



Şekil 2.9 Karar Ağacı Yapısı [16]

Karar ağaçlarında bölmeleme yaparken bazı araçlar kullanılır. Bunlar;

- Entropi
- Gini
- Sınıflandırma Hatası
- En Küçük Kareler Yöntemi

2.3.3.1. Entropi

Verilerdeki belirsizlik durumunun ölçüsü olarak tanımlanır. $H = -\sum P(x)\ logi\ P(x)$ formülü ile hesaplanır. P ile olasılıklar gösterilir. En iyi bölünmeyi belirlemek için Gain hesabı yapılır.

$$Gain(S, D) = H(S) - \sum_{V \in D} \frac{|V|}{|S|} H(V) \tag{2.11}$$

Bu formülde S orijinal veri kümesi, D ise bölünmüş parçadır. V'ler ise S'nin alt kümesidir.

2.3.3.2 Gini Hesaplama

Gini hesaplama yöntemi her bir karar düğümünü bulduktan sonra düğümün bulunduğu dalı ikiye ayırarak ilerler. Gini değeri düşük olan düğümden bölmeleme yapılır.

Tablo 2.1 Gini Hesabı için Bölmeleme Örneği

• Açık • Parçalı Bulutlu • Yağışlı	• Parçalı Bulutlu ve Açık • Yağışlı

Bir özelliğin Gini hesabından önce Gini Sol ve Gini Sağ değerlerine bakılır ve buradan Gini değerine ulaşılır.

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left[\frac{L_i}{T_{sol}}\right]^2 \tag{2.12}$$

$$Gini_{sağ} = 1 - \sum_{i=1}^{k} \left[\frac{R_i}{T_{sağ}}\right]^2 \tag{2.13}$$

Aşağıda 2.12 ve 2.13 formülleri içerisinde kullanılan semboller açıklamıştır.
k: sınıf sayısı

T: Bir düğümdeki örnekler

T_{sol}: Sol taraftaki örnek sayısı

$T_{sağ}$: Sağ taraftaki örnek sayısı

L_i: Sol taraftaki i kategorisinde olan örnek sayısı

R_i: Sağ taraftaki i kategorisinde olan örnek sayısı

$$Gini_j = \frac{1}{n}\left((T_{sol} \times Gini_{sol}) + \left(T_{sağ} \times Gini_{sağ}\right)\right) \tag{2.14}$$

Sağ ve sol Giniler kullanılarak (2.14) ifadesinden Gini değeri hesaplanır [15].

2.3.4 Naive Bayes Algoritması

Naive Bayes algoritması, Bayes teoremini kullanarak istatistiksel tahminlerde bulunan bir sınıflandırıcıdır.

$$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)} \tag{2.15}$$

i: Veri kümesindeki sınıf adeti

$P(C_i)$: Sınıf i'nin ilk olasılığı

$P(X)$: Herhangi bir örneğin X olma olasılığı

$P(C_i|X)$: X olan bir örneğin sınıf i'den olma olasılığı

Naive Bayes algoritması hesaplamaların daha anlaşılabilir yapılmasını ve yorumlamayı sağlayan bir algoritmadır.

Örnek 2.1:

Aşağıda Naive Bayes algoritmasıyla bir sınıflandırma probleminin nasıl çözüleceği gösterilmektedir. Buna göre verilen tablo kullanılarak bir sınıflandırma grubu için cinsiyet bilgisinin ne olacağının tahmin edilmesi sağlanacaktır.

Tablo 2.2 Naive Bayes Verisi

Dergi Promosyonu	Televizyon Promosyonu	Hayat Sigortası Promosyonu	Kredi Kartı Sigortası	Cinsiyet
Evet	Hayır	Hayır	Hayır	Erkek
Evet	Evet	Evet	Evet	Kadın

Tablo 2.2 devamı Naive Bayes Verisi

Hayır	Hayır	Hayır	Hayır	Erkek
Evet	Evet	Evet	Evet	Erkek
Evet	Hayır	Evet	Hayır	Kadın
Hayır	Hayır	Hayır	Hayır	Kadın
Evet	Evet	Evet	Evet	Erkek
Hayır	Hayır	Hayır	Hayır	Erkek
Evet	Hayır	Hayır	Hayır	Erkek
Evet	Evet	Evet	Hayır	Kadın

Sınıflandırılacak Örnek:

Dergi Promosyonu = Evet

Televizyon Promosyonu = Evet

Hayat Sigortası Promosyonu = Hayır

Kredi Kartı Sigortası = Hayır

Cinsiyet = ?

Tablo 2.3 Cinsiyet Özelliğine Göre Sayılar ve Olasılıklar

	Magazin Promosyonu		**Televizyon Promosyonu**		**Hayat Sigortası Promosyonu**		**Kredi Kartı Sigortası**	
Cinsiyet	Erkek	Kadın	Erkek	Kadın	Erkek	Kadın	Erkek	Kadın
Evet	4	3	2	2	2	3	2	1
Hayır	2	1	4	2	4	1	4	3
Oran : Evet / Toplam	4/6	3/4	2/6	2/4	2/6	3/4	2/6	1/4
Oran : Hayır / Toplam	2/6	1/4	4/6	2/4	4/6	1/4	4/6	3/4

Cinsiyet = Kadın için olasılık hesaplama

$$P(cinsiyet = kadın|E) = \frac{P(E|ciniyet=kadın)\times P(cinsiyet=kadın)}{P(E)}$$

Cinsiyet = Kadın için koşullu olasılıkları;

$$P(magazin\ promosyonu = evet|cinsiyet = kadın) = 3/4$$

$P(televizyon\ promosyonu = evet|cinsiyet = kadın) = 2/4$

$P(hayat\ sigortası\ promosyonu = hayır|cinsiyet = kadın) = 1/4$

$P(kredi\ kartı\ promosyonu = hayır|cinsiyet = kadın) = 3/4$

$P(E|cinsiyet = kadın) = (3/4)(2/4)(1/4)(3/4) = 9/128$

$P(cinsiyet = kadın|E) \approx (9/128)(4/10)/P(E)$

Aynı işlemler cinsiyet = kadın için de yapıldığında işlemlerin son hali;

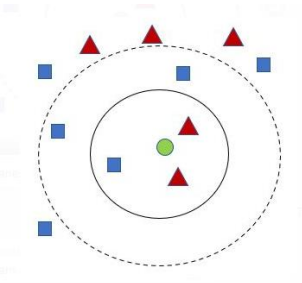
$P(cinsiyet = erkek|E) \approx 0{,}0593/P(E)$

$P(cinsiyet = kadın|E) \approx 0{,}0281/P(E)$

Bayes sınıflandırıcıya göre bu davranışı gösteren cinsiyet erkektir.

2.3.5 En Yakın K Komşu Algoritması

KNN algoritması, en temel örnek tabanlı öğrenme algoritmaları arasındadır. Örnek tabanlı öğrenme algoritmalarında, öğrenme işlemi eğitim setinde tutulan verilere dayalı olarak gerçekleştirilmektedir. Yeni karşılaşılan bir örnek, eğitim setinde yer alan örnekler ile arasındaki benzerliğe göre sınıflandırılmaktadır. KNN algoritmasında, eğitim setinde yer alan örnekler n boyutlu sayısal nitelikler ile belirtilir. Her örnek n boyutlu uzayda bir noktayı temsil edecek biçimde tüm eğitim örnekleri n boyutlu bir örnek uzay da tutulur. Bilinmeyen bir örnek ile karşılaşıldığında, eğitim setinden ilgili örneğe en yakın k tane örnek belirlenerek yeni örneğin sınıf etiketi, k en yakın komşusunun sınıf etiketlerinin çoğunluk oylamasına göre atanır [16].



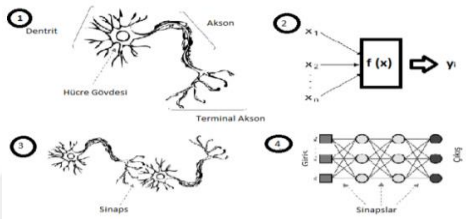
Şekil 2.10 En Yakın K Komşu Algoritması Kolay Gösterimi

KNN hem sınıflandırma hem de regresyon tahmin problemleri için kullanılabilir. Ancak endüstride sınıflandırma problemlerinde daha yaygın olarak kullanılmaktadır. Kırmızı, yeşil ve mavi noktaların her biri kendi içlerindeki kümeleri ifade etmektedir.

2.3.6 Yapay Sinir Ağları

Yapay sinir ağları insan beyninin öğrenme şeklini taklit ederek topladığı verilerden yeni veri üretebilen makine öğrenmesi yöntemlerinden biridir.

Biyolojik sinir hücresi ve yapay sinir ağı örnek görüntüleri Şekil 2.11'de verilmiştir.



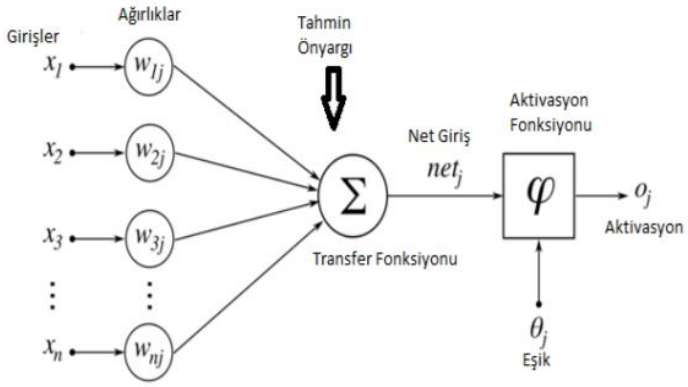
Şekil 2.11 Biyolojik Sinir Hücresi ve Yapay Sinir Ağı [17]

Biyolojik Sinir sistemi elemanları ve yapay sinir sisteminde yer alan karşılıkları ise Tablo 2.4'de verilmiştir. Burada biyolojik sinir sistemi parçalara ayrılmış ve her bir elemana karşılık yapay sinir ağı sisteminde bir karşılığı verilmiştir.

Tablo 2.4 Biyolojik Sinir Sistemi Elemanları ve Yapay Sinir Sisteminde Karşılıkları

Biyolojik Sinir Sistem	Yapay Sinir Sistemi
Nöron	İşlemci Elemanı
Dentrit	Toplama Fonksiyonu
Hücre Gövdesi	Transfer Fonksiyonu
Aksonlar	Yapay Nöron Çıkışı
Sinapslar	Ağırlıklar

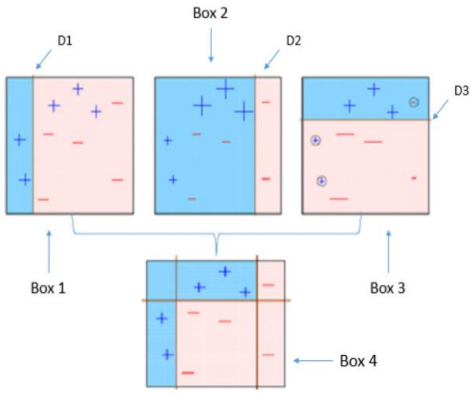
Şekil 2.12'de görüldüğü gibi bir hücreye n tane veri girişi gerçekleşmektedir (X_n veri girişi). Girilen veriler ağırlıklar ile çarpılırak tüm veriler toplanır. Elde edilen bu toplama tahmin eklenir bunun sonucunda net tahmin elde edilir. Net girdi aktivasyon fonksiyonundan geçirilir ve bir veri çıktısı elde edilmiş olur.



Şekil 2.12 Yapay Sinir Hücresi [17]

2.3.7. AdaBoost Algoritması

Boosting terimi, zayıf öğrenenleri güçlü öğrenicilere dönüştüren bir algoritma ailesi anlamına gelir. Algoritma önce her gözlem için eşit ağırlık verir. Bu yöntemle tahmin ettiğinde sonuç yanlışsa o zaman yanlış tahmin edilen gözlem için daha yüksek ağırlık verir. Model belirli bir sınıra ulaşıncaya kadar öğreniciler eklemeye devam eder. AdaBoost algoritması hem regresyon hem de sınıflandırma problemlerinde kullanılabilen bir algoritmadır.



Şekil 2.13 AdaBoost Algoritmasının Basit Gösterimi

Şekil 2.13 Adaboost algoritmasının basit bir gösterimini temsil etmektedir. Bu gösterimdeki ayrıntılar aşağıda ifade edilmiştir.

D1: Her veri noktasına eşit ağırlıklar atandığını varsayalım ve bunları + (artı) veya - (eksi) olarak sınıflandırmak için bir karar kütüğü uygulayalım. Karar kütüğü (D1), veri noktalarını sınıflandırmak için sol tarafta dikey bir çizgi olarak oluşturuldu. Bu dikey çizginin üç tane + değerini yanlış olarak tahmin ederek - ifade olarak tahmin

ettiği görülmektedir. Böyle bir durumda, bu üç + 'ya daha yüksek ağırlıklar atanacak ve başka bir karar kütüğü uygulanacaktır.

D2: Burada, yanlış tahmin edilen üç + boyutunun diğer veri noktalarına kıyasla daha büyük olduğu görülmektedir. Bu durumda ikinci karar kütüğü (D2) onları doğru bir şekilde tahmin etmeye çalışacaktır. Bu kutunun sağ tarafındaki dikey D2 çizgisi yanlış sınıflandırılmış üç + verisini doğru şekilde sınıflandırmıştır. Ancak bu sefer üç – ifadesi için yanlış sınıflandırma ortaya çıkmıştır. Bunun da önüne geçebilmek için bu sefer üç - verisine daha yüksek ağırlık ataması yapılarak başka bir karar kütüğü uygulanacaktır.

D3: Burada, üç - ifadesine daha yüksek ağırlıklar atanmıştır.. Bu yanlış sınıflandırılmış gözlemleri doğru bir şekilde tahmin etmek için yeni bir karar kütüğü (D3) uygulanır. Bu sefer + ve - yanlış sınıflandırılmış gözlemin daha yüksek ağırlığına dayalı olarak sınıflandırmak için yatay bir çizgi oluşturulur.

D4: Burada, bireysel zayıf öğrenciye kıyasla karmaşık kuralı olan güçlü bir tahmin oluşturmak için D1, D2 ve D3 karar kütükleri birleştirilerek yeni bir karar kütüğü oluşturulmuştur. Bu algoritmanın, bu gözlemleri herhangi bir zayıf öğrenenlere kıyasla oldukça iyi sınıflandırdığını görebilirsiniz.

2.4 Denetimsiz Makine Öğrenmesi Algoritmaları

Denetimsiz öğrenme için kullanılan yöntemler aşağıda açıklanmaktadır.

2.4.1. Kümeleme Algoritmaları

Veri setindeki verilerin sınıfları bilinmediğinde denetimsiz algoritmalar kullanılmaktadır. Kümeleme algoritmaları eldeki verilerin benzerliklerini kullanarak ilerler. Bu benzerlikler birbirlerinde olan uzaklıkların belirlenmesiyle bulunur. Çoğunlukla uzaklıklar bulunurken Manhattan, Minkowski ve Euclid uzaklıklarını kullanırlar.

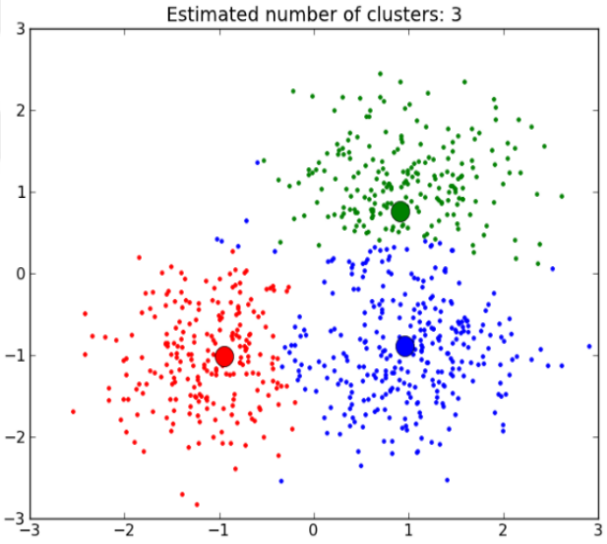
2.4.2. Hiyerarşik Kümeleme

Hiyerarşik kümeleme, örnekleri birbirlerine yakınlıklarına göre değişik aşamalarda bir araya getirerek kümeleri ardışık bir biçimde oluşturmaktadır. Temelde birleştirici ve ayrıştırıcı / bölücü olmak üzere iki yaklaşıma sahiptir. Birleştirici

hiyerarşik kümelemede, her bir örnek başlangıçta ayrı bir küme olarak kabul edilir, her adımda kümelerin birbirlerine uzaklıkları hesaplanır ve en yakın iki küme birleştirilerek bir üst küme oluşturulur. Böylece, her adımda küme sayısı bir azaltılır ve işlem bütün örneklerin dahil olduğu tek bir büyük küme oluşuncaya kadar devam ettirilir. Bu süreç, dendogram olarak adlandırılan hiyerarşik bir ağaç diyagramı ile görselleştirilebilmektedir. Ayrıştırıcı hiyerarşik kümelemede ise, tam tersi bir biçimde, tüm örnekler başlangıçta tek bir küme olarak kabul edilir ve her aşamada benzer olmayan gözlemler belirlenerek daha küçük kümeler oluşturulur. Bu süreç, her gözlem tek başına küme oluşturuncaya kadar sürdürülür [18].

2.4.3. K-Ortalama

En çok kullanılan kümeleme algoritmasıdır. Bu algoritmada örneklem k adet kümeye bölünür. Benzerlik bulunan veriler aynı kümeye dâhil edilir. Küme merkezlerinde herhangi bir değişiklik olmayana kadar işlem devam eder.



Şekil 2.14 K-Ortalama Kümeleme Algoritması Gösterimi (K=3 için)

Kırmızı, mavi ve yeşil noktaların her biri kümeleri ifade ederken büyük olan noktalar merkezi göstermektedir.

2.5 Makine Öğrenmesi Algoritmasına Karar Verme

Bu bölümde makine öğrenmesi modeline karar verirken göz önünde bulundurulması gereken ölçülerden bahsedilecektir.

2.5.1 Regresyon Modelleri için Kullanılan Ölçüler

2.5.1.1 Açıklanan Varyans Puanı

Varyans, bir örnek veya veri kümesi içindeki veri noktalarının ne kadar yayılmış olduğunun istatistiksel bir ölçüsüdür. En iyi sonuç 1.0'dır. Değer düştükçe algoritmanın başarısı düşer.

2.5.1.2 Ortalama Mutlak Hata (MSE)

Ortalama mutlak hata (mean square error) bir regresyon eğrisinin bir dizi noktaya ne kadar yakın olduğunu söyler. Gerçek değer ile tahmin değeri arasındaki hataların karesinin ortalamasıdır ve aşağıdaki şekilde hesaplanır.

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \tag{2.16}$$

2.5.1.3 Kare Ortalamalarının Karekökü (RMSE)

Değerlerin karelerinin alınıp bu değerlerin ortalamasının alınmasından sonra karekökünün alınmasıyla bulunur.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} = \sqrt{MSE} \tag{2.17}$$

2.5.1.4 R^2-Score

R^2, doğrusal regresyon modelleri için bir uygunluk ölçüsüdür. R^2, bir regresyon modelinin eldeki verilere yeterli bir uyum sağlayıp sağlamadığını göstermez. İyi bir model düşük bir R^2 değerine sahip olabildiği gibi taraflı bir model yüksek bir R^2 değerine sahip olabilir. R^2'nin yüksek tahminler yapması ihtiyaç duyulan hassasiyete ve verilerde bulunan varyasyon miktarına bağlıdır. Yüksek R^2 değerine sahip bir veri seti başarılı olabilirken diğer yandan aşırı öğrenme durumundan dolayı çok sayıda sorun da içeriyor olabilir. Bunun en iyi örneği aşırı öğrenme (overfitting) olarak verilebilir. Yüksek R^2 değerine sahip olsa da verinin başarı oranı için aynı şeyi söylemek mümkün olmayacaktır.

2.5.2 Sınıflandırma Modelleri için Kullanılan Ölçüler

2.5.2.1 Karmaşıklık Matrisi

Tahminlerin doğruluğunu anlamak için oluşturulmuş bir matristir. Test veri setinin sınıflandırılmasının doğruluğu hakkında bilgilere erişmemizi sağlar. Tablo 2.5'de bir karmaşıklık matrisinin yapısı verilmiştir. Burada; TP gerçek pozitif (true positive) sayısını, TN gerçek negatif (true negative), FP yanlış pozitif (false positive), FN ise yanlış negatif (false negative) sayılarını temsil etmektedir.

Tablo 2.5 Karmaşıklık Matrisi

	Tahmin		
Gerçek		Pozitif	Negatif
	Pozitif	Gerçek Pozitif (TP)	Yanlış Negatif (FN)
	Negatif	Yanlış Pozitif (FP)	Gerçek Negatif (TN)

Karışıklık matrisinden hesaplanan bazı oranlar vardır. TP, TN, FP ve FN'nin birbirleriyle ilişkisini gösteren bu terminolojiler aşağıda verilmiştir.

$$\text{TOPLAM} = \text{TP} + \text{TN} + \text{FP} + \text{FN} \tag{2.18}$$

$$\text{GERÇEK POZİTİFLER} = \text{TP} + \text{FN} \tag{2.19}$$

$$\text{GERÇEK NEGATİFLER} = \text{TN} + \text{FP} \tag{2.20}$$

2.5.2.2 Doğruluk (Accuracy)

Modelin doğruluk değeridir.

$$\text{Doğruluk} = \frac{TP + TN}{TOPLAM} \tag{2.21}$$

2.5.2.3 Kesinlik (Precision)

Modelin duyarlılığı anlamına gelir.

$$Kesinlik = \frac{TP}{TP + FN} \tag{2.22}$$

2.5.2.4 F1-Skoru

Kesinlik ve duyarlık değerlerinin harmonik ortalamasıdır.

$$F_1\ Skoru = \frac{2 * Kesinlik * Duyarlılık}{Kesinlik * Duyarlılık} \tag{2.23}$$

2.5.2.5 Duyarlılık (Recall)

Pozitif olarak tahmin edilenlerin ne kadarının doğru tahmin edildiğinin bir ölçüsüdür. Mümkün olduğu kadar yüksek olmalıdır. Pozitif tahmin edici değer olarak da bilinir.

$$Duyarlılık = \frac{TP}{TP + FP} \tag{2.24}$$

3
YENİ DOĞANLARIN YOĞUN BAKIM İHTİYACININ TAHMİN EDİLMESİ

IRB# 2018.234.IRB1.027 numaralı etik kurulu onayıyla, 2014-2019 yılları arasında İstanbul Medipol hastanesine doğum nedeniyle yatmış anne ve bebeklerine ait veriler geriye dönük olarak toplanmıştır. Hastane kayıtlarından temin edilen veriler her yıl için üç farklı grup olacak şekilde 14 adet Excel dosyalarından oluşmaktaydı.

Birinci grup dosyalarda hastaneye başvuran annelere ait bilgiler yer almaktadır. Bunlar özet olarak, annenin adı soyadı, doğum tarihi, uyruğu, hastaneye başvurma tarihleri, ICD kodları (başvuru kodları), başvuru sebeplerini içermektedir. İkinci grup dosyalar; tam kan sayımı alınan annenin adı soyadı, doğum tarihi, tam kan sayım parametrelerinin sonuçları, numunenin alındığı ve sonuçlandığı tarihleri içermektedir. Yoğun bakım bilgilerini içeren üçüncü grup dosyalarda ise; bebeğin adı soyadı, cinsiyeti, doğum tarihi, anne-baba adı, yoğun bakıma girme ve çıkma tarihiyle beraber yoğun bakıma girme sebebini içeren veriler yer almaktadır.

Bu tezde; doğum öncesi annenin tam kan sayım parametreleri, anne yaşı ve başvuru şikayetleri ile bebeğin doğum sonrası yoğun bakım ihtiyacının olup olmayacağı arasındaki ilişki incelenmektedir. Bu amaç kapsamında; öncelikle, temin edilen dosyalar üzerinde gerekli veri temizleme ve birleştirme işlemleri yapılmıştır. Aşağıda sırasıyla verilerin düzenlenmesi, istatistiksel analiz ve makine öğrenmesi teknikleriyle yoğun bakım ihtiyacının tahmin edilmesi aşamaları ayrıntılı bir biçimde ele alınmaktadır.

3.1 Verilerin Düzenlenmesi

Yapılan çalışmada eldeki tablolarda ortak olan parametreler üzerinden eşleştirme ve veri temizleme işlemi uygulanarak, veri seti yapılandırılmıştır. Verilerin birleştirilmesi yapılırken tanımlayıcı kişisel veriler çalışma dosyalarından ayıklanarak anonim hale getirilmiştir. Veriler içerisinde kaydı açan kullanıcı tarafından tek bir kolona noktalama işaretleriyle ayrılarak birden fazla şekilde yazılan ICD kodu ve şikayet alanları tekilleştirilmiştir.

Elde edilen son veri setinde veriler, yoğun bakıma giren (hasta) ve yoğun bakıma girmeyen (kontrol) veriler olarak iki gruba ayrılmıştır. Verilerin işleme aşamaları aşağıda özetlenmiştir.

Birbirinden farklı bu dosyaları birleştirme işlemi yapabilmek için öncelikle veri setleri üzerinde temizlik ve düzenlemeler gerçekleştirilmiştir.

- Adım 1: İkinci grup dosyalarda dikey bir şekilde yer alan kan sonuçlarını her bir anne için tek bir satırda göstermek, gerekli veri temizliği ve düzenlemelerini yapmak için Excel Visual Basic yardımıyla aşağıdaki adımlar gerçekleştirilmiştir.

  Adım 1.1: Her bir kan tahlili için dokümanda dikey bir şekilde yer alan Eritrosit (HBC), HCT (Hematokrit), HGB (Hemoglobin), Lökosit (WBC), MCH, MCHC, MCV, MPV, PCT, PDW, PLT (Trombosit) ve RDW adlı kan parametreleri satır şeklinde gösterilmiştir.

  Adım 1.2: Bu düzenleme her bir yıla ait dosya için ayrı ayrı yapılarak, tek bir dosya halinde birleştirilmiştir.

  Adım 1.3: Kan değerleri içerisinde ilgili kolona ait veriler incelendiğinde aykırı olarak tespit edilen değerleri olan veriler silinmiştir.

  Adım 1.4: Bunun ardından data içerisindeki kolonlar için önem ve ihtiyaç çalışması yapılmıştır. Daha net ifade etmek gerekirse; verilerin önem seviyelerine göre belirlenen ve ihtiyaç olmadığı tespit edilen kolonların (L/H, anne cinsiyeti, kurum, bölüm, doktor, geliş tipi, sut kodu, onay tarihi) temizleme işlemi gerçekleştirilmiştir.

- Adım 2: Anne kan tahlillerini içeren ikinci grup dosyalarla bebek verilerini içeren üçüncü grup dosyalar tekil anahtarlara sahip değildir. Bu nedenle anne ile bebe arasında kayıtları doğrudan eşleştirmek mümkün değildir. Bu sorunu aşmak için her iki dosyadaki anne ad-soyad ve kaydın açılış tarihleri dikkate alınarak her iki dosyadaki eşleştirmeler yapılmıştır. Buna göre hastanede doğan bebek verileri üzerinde aşağıdaki işlemler gerçekleştirilmiştir:

Adım 2.1: Anne adı soyadı ile bebeğin annesinin adı soyadı birleştirilmiştir.

Adım 2.2: Birleştirilen bu veri ile kan sayım verileri içerisindeki hasta adları eşleştirilmiştir.

Adım 2.3: Eşleşen hastaların kan sayımı için açılan kaydının açılış tarihi verisi yoğun bakım datasına (üçüncü grup dosyaya) yeni bir kolon olarak eklenmiştir.

Adım 2.4: Bebeğin açılış tarihi ile annenin hastane kaydının açılış tarihi arasındaki gün farkı bulunmuştur.

Adım 2.5: Anne isim soyadları eşleşen ve hastane açılış kayıtları 10 günün altında fark olan kayıtlar eşleştiği kabul edilerek annelere ait bebek verilerine ulaşılmıştır.

Adım 2.6: eşleşen tüm kayıtlar için anne ismine göre alfabetik sıralama yapılmıştır.

Böylece kan sayım verileri ile yoğun bakıma alınan bebeklerin eşleşmesi sağlanmıştır.

- Adım 3: Farklı yıllarda hastaneye başvuran anne bilgilerini içeren birinci grup dosyalarla yoğun bakımdaki bebek verilerini içeren üçüncü grup dosyalar tekil anahtarlara sahip olmadığı için aralarındaki kayıt eşleştirme işlemleri aşağıda belirtilen prosedürle belirlenmiştir:

  Adım 3.1: Yoğun bakım dosyasında daha önce birleştirilen anne adı ile bebeğin soyadı bilgisi kullanılarak birinci grup dosya içerisinde yer alan hasta adları eşleştirilmiştir.

  Adım 3.2: Eşleşen hastaların hastanede açılan kaydının açılış tarihi verisi yoğun bakım datasına yeni bir kolon olarak eklenmiştir.

  Adım 3.3: Bebeğin açılış tarihi ile annenin hastane kaydının açılış tarihi arasındaki gün farkı bulunmuştur.

Adım 3.4: Anne isim soyisimleri eşleşen ve hastane açılış kayıtları 10 günün altında fark olan kayıtların eşleştiği kabul edilerek annelere ait bebek verilerine ulaşılmıştır.

Adım 3.5: Yoğun bakım datası ve yıllar içerisinde hastaneye başvuran annelerin bilgileri ile yapılan bu eşleşmenin ardından anne ad soyad bilgisine göre alfabetik sıralama gerçekleştirilmiştir.

- Adım 4: Yıllar içerisinde hastaneye başvuran hasta kayıtları üzerinde aşağıda belirtilen düzenlemeler gerçekleşmiştir.

  Adım 4.1: Kolonlar üzerinde analiz yapılarak önem ve ihtiyaç çalışması yapılmıştır.

  Adım 4.2: Yoğun bakım ihtiyacının belirlenmesinde önemsiz görülen kolonlar temizlenmiştir.

- Adım 5: Yıllar içerisinde hastaneye başvuran annelere ait kayıtlar ile yoğun bakım datalarının eşleşmesi sağlanmıştır.

  Bu eşleşme için;

  Adım 5.1: Yoğun bakım datası içerisinde birleştirilen anne adı ile bebek soyadı bilgisi ile hastane başvuran kayıtların hasta ad-soyad eşleşmesi yapılmıştır.

  Adım 5.2: Yoğun bakım datasında yer alan açılış tarihi eşleşen kayıtlar için hasta kayıt datalarına eklenmiştir.

  Adım 5.3: İkinci adımda gelen yoğun bakım açılış tarihi ile hasta kayıt açılış tarihi arasındaki gün farkı hesaplanmıştır.

  Adım 5.4: 0'dan küçük olan kayıtlar birbirine uymayacağından gün farkı 0'dan büyük ve 10'dan küçük olan kayıtların eşleştiği kabul edilmiştir.

  Adım 5.5: Eşleşen kayıtlar için hasta adına göre alfabetik sıralama yapılmıştır.

- Adım 6: Yoğun bakım-yıl eşleşen datası (üçüncü adımda elde edilen) ile yıl – yoğun bakım eşleşen (beşinci adımda elde edilen) datalar arasında anne ad-soyad ve hasta adı eşleştirmesi kullanılarak iki tablo birleştirilmiş ve hastalara ait bebek bilgileri yanlarına getirilerek yıl-yoğun bakım ortak datası elde edilmiştir.
- Adım 7: Tüm yıl verileri için altıncı adımda yer alan eşleşme sağlanmıştır. Bu verilerin tamamı tek bir dosya olacak şekilde bir araya getirilerek 2014-2019 yılları için yıl-yoğun bakım ortak datası elde edilmiştir.

  Elde edilen bu veriler ile yapılan tespitler şunlardır:

  - Bir annenin bir çocuğunun birden fazla yoğun bakım kaydı olabilir.
  - Bir annenin birden fazla çocuğu olabilir.

- Adım 8: Yedinci adımda elde edilen verilerle yoğun bakım verilerinin kolon analizleri yapılmıştır;

  Adım 8.1: Ortak olan kolonların silinmesi sağlanmıştır.

  Adım 8.2: Dosya içerisinde aykırı veriler için veri temizlemesi yapılmıştır.

- Adım 9: Sekizinci adımda elde edilen veriler ile annelerin tam kan sayım verileri arasında eşleştirme yapılmıştır.

  Adım 9.1: Dokümanlar içerisindeki annenin adı soyadı ile doğum tarihi alanları kullanılmıştır.

  Adım 9.2: Birleştirilen dokümanda ortak veri içeren kolonların temizlenmesi işlemi yapılmıştır.

- Adım 10: Yaptığımız çalışmada gebelik anındaki verilere istinaden sonuç elde etmek istenildiğinden anneden alınan kan sayım tarihi ile bebeğin doğum tarihi arasında ilişki incelenmiştir.

  Adım 10.1: Anneden alınan kan verisi bebeğin doğum tarihinden sonraysa ilgili kayıt silinmiştir.

- Adım 11: Üretilen son veri seti içerisindeki kişisel verileri ortadan kaldırmak için elde edilen tüm kayıtlara eşsiz ID bilgisi atanmıştır.

  Adım 11.1: Her bir veriye tekil bir ID atanmıştır (*evrensel_id*).

Adım 11.2: Bir annenin birden fazla bebeği ya da bir bebek birden fazla kez yoğun bakıma girebileceğinden dolayı anne verilerinin çoklandığı gözlenmiştir. Bu sebeple her bir anneye ad-soyad ve doğum tarihi üzerinden tekil bir ID atanmıştır (*anne_id*).

Adım 11.3: Her bir bebek bilgisi için tekil bir id atanmıştır (*bebek_id*).

Adım 11.4: *evrensel_id*, *anne_id* ve *bebek_id* birleştirilerek tekil bir id elde edilmiştir.

Ham veriler üzerinden yapılan tüm veri işlemlerinin ardından, istatistik analiz ve makine öğrenmesi için kullanılacak veri setinin öznitelikleri listelenmiştir (Tablo 3.1).

Tablo 3.1 Sonuç veri setinin içerdiği öznitelikler

Öznitelik Adı	**Veri Açıklaması**	**Veri Tipi**
ID	Kişisel veri kullanmamak adına anne verileri numaralandırılmıştır. Her bir numara bir anneyi temsil etmektedir.	Nümerik
AnneUyruk	Annenin uyruk bilgisidir.	Kategorik
O69	Doğum süreci ve doğum, umbilikal kordon komplikasyonları ile birlikte olan	Kategorik (Evet-Hayır)
O44	Plasentanın rahim içinde doğum kanalını kapatacak şekilde yerleşmesi	Kategorik (Evet-Hayır)
Z34	Normal gebeliğin gözlemi	Kategorik (Evet-Hayır)
O45	Plasentanın erken ayrılması	Kategorik (Evet-Hayır)
O84	Çoğul doğum	Kategorik (Evet-Hayır)
D64	Anemi, diğer	Kategorik (Evet-Hayır)
O68	Doğum süreci ve doğum, fetal stresle komplike	Kategorik (Evet-Hayır)
O66	İlerlemeyen doğumlar, diğer	Kategorik (Evet-Hayır)
O80	Tek spontan doğum	Kategorik (Evet-Hayır)

Tablo 3.1 devamı Sonuç veri setinin içerdiği öznitelikler

O82	Tek doğum, sezeryan ile	Kategorik (Evet-Hayır)
AnneYas	Annenin yaş bilgisi	Nümerik
RBC	Eritrosit (kırmızı kan hücreleri)	Nümerik
HCT	Hematokrit (kırmızı kan hücrelerinin dolaşımdaki kan miktarına göre hacminin oranını)	Nümerik
HGB	Hemoglobin (kansızlık ya da demir eksikliği problemleri için bakılan değer)	Nümerik
WBC	Lökopeni (Beyaz kan hücresi)	Nümerik
MCH	Ortalama Hemoglobin miktarı (Kırmızı kan hücrelerinin her birinde bulunan ortalama hemglobin miktarının oranı)	Nümerik
MCHC	Ortalama Eritrosit Hemoglobin Konsantrasyonu	Nümerik
MCV	Ortalama Eritrosit Hacmi	Nümerik
MPV	Ortalama Trombosit Hacmi	Nümerik
PCT	Prokalsitonin (Bakteriyel enfeksiyon şüphesinde bakılır)	Nümerik
PDW	Trombosit dağılım aralığı	Nümerik
PLT	Trombosit (kan pulcukları)	Nümerik
RDW	Alyuvarların hacmini gösterir	Nümerik
Class	Bebeğin yoğun bakıma girip girmediğini gösteren alan	Kategorik (Hasta-Kontrol)
Şikayet	Annenin hastaneye başvururken bir şikayet de bulundu mu	Kategorik (Evet-Hayır)

3.2 İstatiksel Analiz

Ön işleme tamamlandıktan sonra analiz edilecek veri seti 41.387 kayıttan oluşmaktadır ve bu kayıtların 3.361 (%8,1) tanesi bebeği yoğun bakıma yatan annelere aittir. Yoğun bakıma giren bebeklerin 2037 (%61) tanesinin erkek, 1324 (%39) tanesinin kadın olduğu tespit edilmiştir. Yoğun bakıma giren bebeklerin ortalama 4,55 gün yoğun bakım ünitesinde kaldığı görülmüştür. Kontrol kategorisindeki veriler için ortalama gebelik haftası 38,69 olarak, standart sapma

1,15 olarak hesaplanmıştır. Hasta kategorisindeki veriler için ise ortalama gebelik haftası 38,16 olarak, standart sapma 2,31 olarak hesaplanmıştır.

Kontrol ve hasta sınıfına ait nümerik öznitelikler için hesaplanan ortalama ve standart sapma değerleri Tablo 3.2 de verilmiştir.

Tablo 3.2 Kontrol ve Hasta Sınıfa ait Nümerik Öznitelikler için Ortalama ve Standart Sapma Değerleri

Anne	**Kontrol**	**Hasta**
Yaş ortalaması	31,68 ± 5,19	32,74 ± 5,48
WBC	13,48 ± 4,18	13,28 ± 4,53
RBC	3,98 ± 0,42	3,93 ± 0,43
HGB	11,55 ± 1,39	11,49 ± 1,39
HCT	34,76 ± 3,83	32,74 ± 3,86
MCH	29,10 ± 2,81	29,31 ± 2,76
MCHC	33,21 ± 1,13	33,21 ± 1,08
MCV	87,54 ± 7,03	88,18 ± 7,02
RDW	14,21 ± 2,59	13,78 ± 2,42
PLT	209,40 ± 60,54	211,75 ± 63,25
MPV	9,73 ± 1,34	9,43 ± 1,31
PCT	0,20 ± 0,06	0,20 ± 0,06
PDW	15,72 ± 3,43	15,83 ± 3,67

Kategorik verilerin hastaneye başvuran annelerde bulunup bulunmama durumları Evet-Hayır olarak işaretlenmiştir. İlgili kategorik veri için annenin şikayeti bulunması durumunda ‘Evet’ olarak, şikayeti bulunmaması durumunda ‘Hayır’ olarak işaretlenmesi sağlanmıştır. Her bir kategorik veri için Evet-Hayır değerlerinin sayıları ve oranları Tablo 3.3 de verilmiştir.

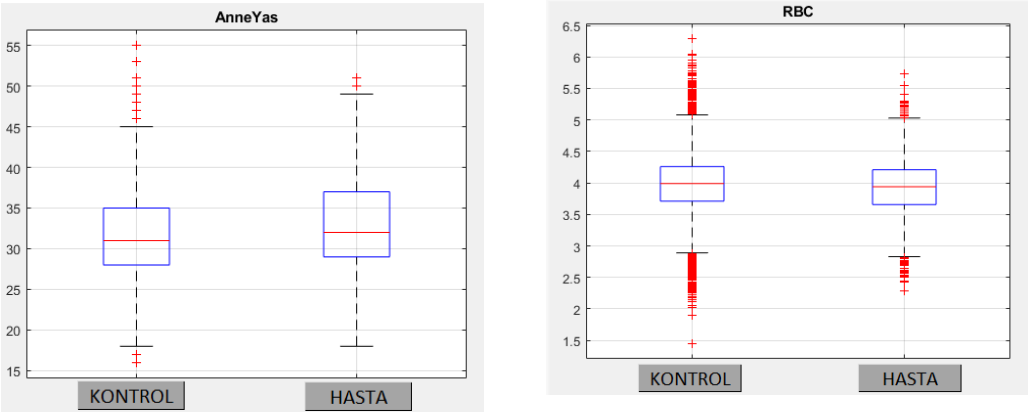
Tablo 3.3 Kontrol ve Hasta Sınıfa ait Kategorik Öznitelikler için Adet ve Oranları

	Kontrol		**Hasta**	
Öznitelikler	**Evet**	**Hayır**	**Evet**	**Hayır**
Şikayet	30.600 (%80,47)	7.426 (%19,53)	2.768 (%82,36)	593 (%17,64)
Doğum süreci ve doğum, kordon komplikasyonları	33 (%0,09)	37.993 (%99,91)	11 (%0,33)	3.350 (%99,67)
Plasentanın rahim içinde doğum kanalını kapatacak şekilde yerleşmesi	47 (%0,12)	37.979 (%99,88)	10 (%0,30)	3.351 (%99,70)
Normal gebeliğin gözlemi	93 (%0,24)	37.933 (%99,76)	3 (%0,09)	3.358 (%99,91)
Plasentanın erken ayrılması	201 (%0,53)	37.825 (%99,47)	67 (%1,99)	3.294 (%98,01)
Çoğul doğum	301 (%0,79)	37.725 (%99,21)	152 (%4,52)	3.209 (%95,48)
Anemi, diğer	565 (%1,49)	37.461 (%98,51)	58 (%1,73)	3.303 (%98,27)
Doğum süreci ve doğum, fetal stresle komplike	1.101 (%2,90)	36.925 (%97,10)	172 (%5,12)	3.189 (%94,88)
İlerlemeyen doğumlar, diğer	1.267 (%3,33)	36.759 (%96,67)	74 (%2,20)	3.287 (%97,80)
Tek spontan doğum	2.992 (%7,87)	35.034 (%92,13)	78 (%2,32)	3.283 (%97,68)
Tek doğum, sezeryan ile	15.406 (%40,51)	22.620 (%59,49)	1.712 (%50,94)	1.649 (%49,06)
Sancı	21.995 (%57,84)	16.031 (%42,16)	1.616 (%48,08)	1.745 (%51,92)
EMR	6.962 (%18,31)	31.064 (%81,69)	895 (%26,63)	2.466 (%73,37)
Fetüs	755 (%1,99)	37.271 (%98,01)	91 (%2,71)	3.270 (%97,29)
Baş ağrısı	4 (%0,01)	38.022 (%99,99)	2 (%0,06)	3.359 (%99,94)

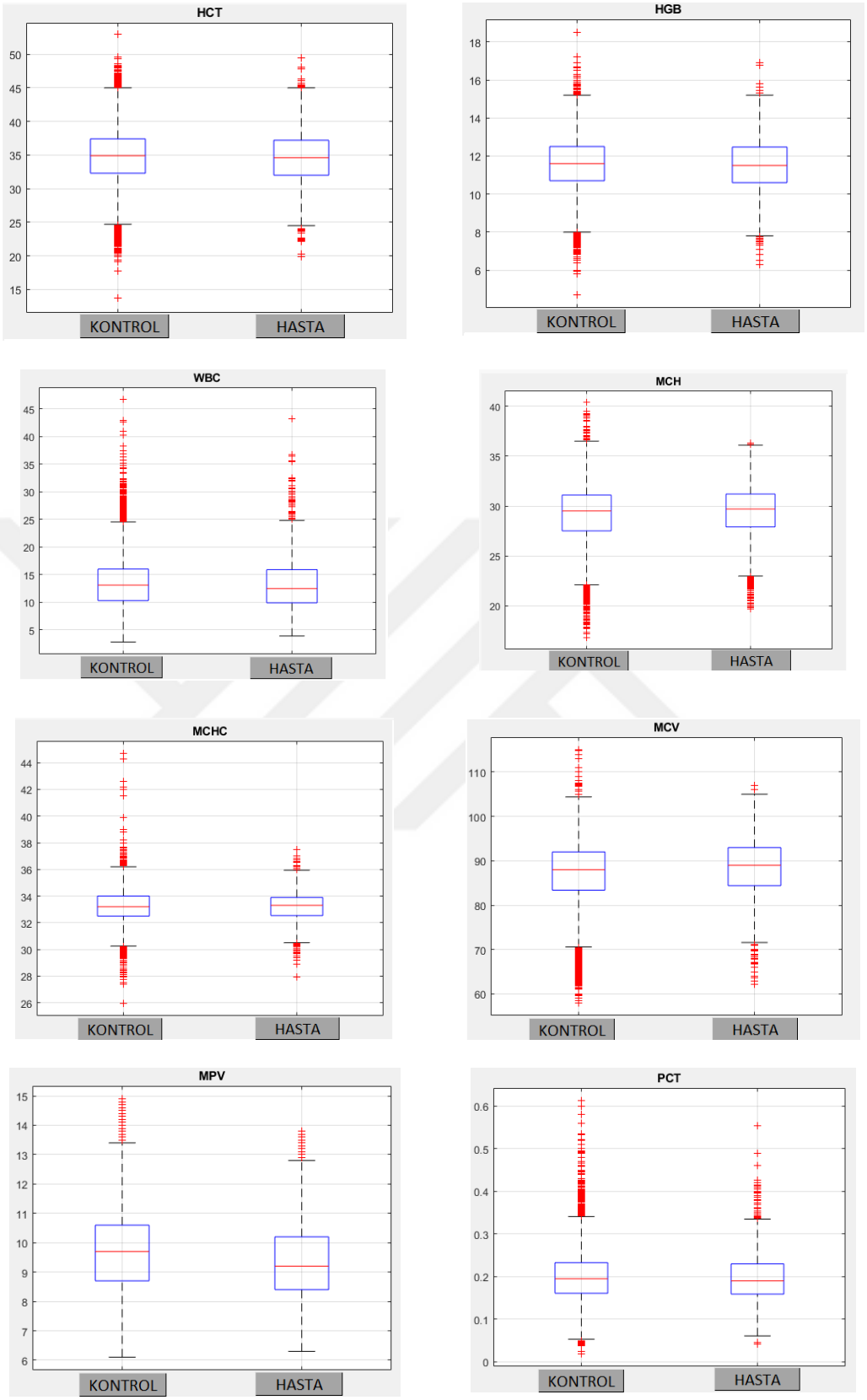
Tablo 3.3 devamı Kontrol ve Hasta Sınıfa ait Kategorik Öznitelikler için Adet ve Oranları

Kaşıntı	64 (%0,17)	37.962 (%99,83)	8 (%0,24)	3.353 (%99,76)
Kanama	477 (%1,25)	37.549 (%98,75)	78 (%2,32)	3.283 (%97,68)
Tansiyon	304 (%0,80)	37.722 (%99,20)	71 (%2,11)	3.290 (%97,89)
Ödem	31 (%0,08)	37.995 (%99,92)	7 (%0,21)	3.354 (%99,79)
Ateş	6 (%0,02)	38.020 (%99,98)	0 (%0)	3.361 (%100)
Diyabet	2 (%0,01)	38.024 (%99,99)	0 (%0)	3.361 (%100)

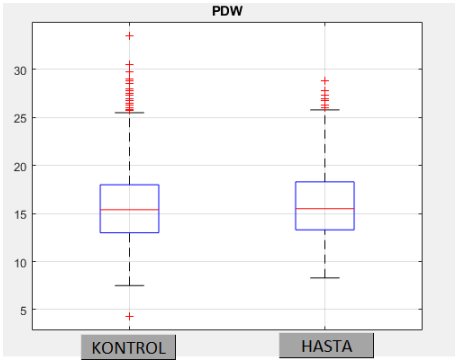
Şekil 3.1'de annenin yaşı ve kan parametrelerine ait kutu diyagramları verilmiştir. Şekiller içerisinde kırmızı renk ile işaretlenen veriler aykırı değerleri göstermektedir. RBC, WBC, MCV, PLT, RDW öznitelikleri için aykırı değerlerin fazla olduğu görülmektedir. Aykırı verilerden medikal uygunsuz olanlar çıkarılmıştır.

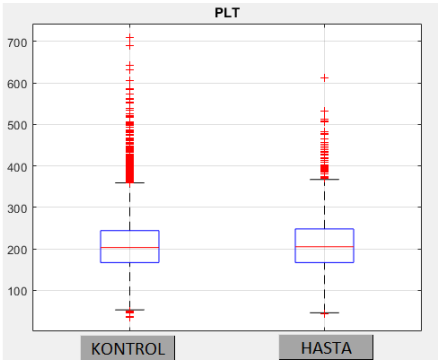


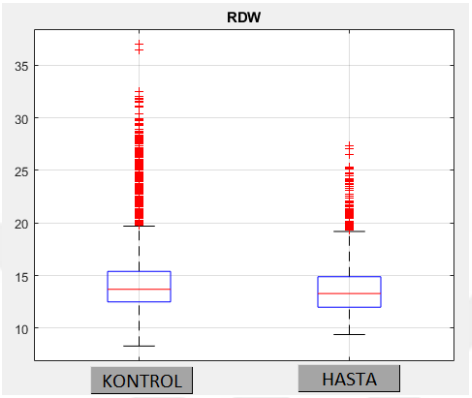
Şekil 3.1 Anne yaş ve Kan Parametrelerine ait Kutu Diyagramları



Şekil 3.1 devamı Anne yaş ve Kan Parametrelerine ait Kutu Diyagramları







Şekil 3.1 devamı Anne yaş ve Kan Parametrelerine ait Kutu Diyagramları

Bebeği yoğun bakıma yatan ve yatmayan annelere ait öznitelikler için ayrı ayrı istatistiksel testler uygulanmıştır. Bu kapsamda; nümerik özniteliklerin ortalama farkı için z testi ve Mann-Whitney U testi, kategorik özniteliklerin oran farkı için z testi uygulanmıştır. Hesaplanan p değerleri Tablo 3.4'te verilmiştir.

Tablo 3.4 Kontrol ve Hasta Sınıfa ait Nümerik Öznitelikler için p değerleri

Öznitelikler	**p değeri (z-testi)**	**p değeri (Mann-Whitney U testi)**
Yaş	0,000	0,000
WBC	0,016	0,000
RBC	0,000	0,000
HGB	0,012	0,004
HCT	0,005	0,001
MCH	0,000	0,000
MCHC	0,759	0,725
MCV	0,000	0,000

Tablo 3.4 devamı Kontrol ve Hasta Sınıfa ait Nümerik Öznitelikler için p değerleri

RDW	0,000	0,000
PLT	0,039	0,073
MPV	0,000	0,000
PCT	0,000	0,000
PDW	0,068	0,037

p değerlerinin 0,05'den küçük olduğu değerler istatistiki olarak anlamlı kabul edilmiştir. Buna göre; her iki testin sonuçları dikkate alındığında MCHC, PLT ve PDW dışındaki kan sayım parametreleri için bebeği yoğun bakıma yatan ve yatmayan annelerin parametrelerinin ortalamaları arasında anlamlı istatistiksel farklar bulunmuştur.

Tablo 3.5 Kontrol ve Hasta Sınıfa ait Kategorik Öznitelikler için p değerleri

Öznitelikler	**p değeri (z-testi)**	**p değeri (ki kare-testi)**
Şikayet	0,006	0,008
Doğum süreci ve doğum, kordon komplikasyonları (O69)	0,016	0,000
Plasentanın rahim içinde doğum kanalını kapatacak şekilde yerleşmesi (O44)	0,069	0,009
Normal gebeliğin gözlemi (Z34)	0,007	0,072
Plasentanın erken ayrılması (O45)	0,000	0,000
Çoğul doğum (O84)	0,000	0,000
Anemi, diğer (D64)	0,030	0,274
Doğum süreci ve doğum, fetal stresle komplike (O68)	0,000	0,000
İlerlemeyen doğumlar, diğer (O66)	0,000	0,000
Tek spontan doğum (O80)	0,000	0,000
Tek doğum, sezeryan ile (O82)	0,000	0,000

Tablo 3.5'te görüldüğü üzere O69, O44 ve D64 dışındaki ICD kodları için bebeği yoğun bakıma yatan ve yatmayan annelerin parametrelerinin 'Evet' oranları arasında anlamlı farklar bulunmuştur. Kontrol ve hasta gruplarında Z34 ve D64 dışındaki kategorik özniteliklerin homojen dağılmadığı tespit edilmiştir.

3.3 Yeni Doğanların Yoğun Bakım İhtiyacının Makine Öğrenmesi Teknikleriyle Tahmin Edilmesi

Elde edilen veri seti 38026 kontrol kategorisindeki annenin (bebeği yoğun bakıma girmeyen) ve 3361 hasta kategorisindeki annenin (bebeği yoğun bakıma giren) bilgilerini içermektedir. Matlab Sınıflandırma Öğreticisi (Classification Learner) uygulaması kullanılarak veri seti karar ağacı, Naive Bayes sınıflandırması, destek vektör makineleri (SVM), topluluk öğrenmesi (ensemble classifiers) teknikleri için tahminlemeler yapılmıştır. Bu teknikler uygulanırken, Matlab'de varsayılan parametre değerleri kullanılmıştır. Makine öğrenmesi tekniklerini uygulayabilmek için elde edilen verinin 4/5'i eğitim için ayrılırken 1/5'i test için ayrılmıştır. Eğitim verisine 10 katlamalı çapraz doğrulama uygulanarak makine öğrenmesi modelleri eğitilmiştir. Sınıflarda bulunan veri sayılarında dengesizlik bulunduğundan dolayı, makine öğrenmesi algoritmaları uygulanırken aşağıdaki maliyet tablosu dikkate alınmıştır.

Tablo 3.6 Makine Öğrenmesi Algoritmalarında Kullanılan Maliyet Tablosu

	Tahmin		
Gerçek		Kontrol	Hasta
	Kontrol	0	3.361
	Hasta	38.026	0

Makine öğrenimi tahminlemelerine göre elde edilen duyarlılık ve özgünlük değerleri Tablo 3.7'de gösterilmiştir.

Tablo 3.7 Makine Öğrenmesi Yöntemleri için Duyarlılık-Özgünlük Değerleri Tablosu

Yöntem	Duyarlılık	Özgünlük
Kuadratik çekirdeğe sahip destek vektör makineleri	%62,60	%60,70
Medium Gaussian çekirdeğe sahip destek vektör makineleri	%65,80	%56,90
Gentle AdaBoost (Ensemble Method)	%65,30	%59,00
Coarse Gaussiaan çekirdeğe sahip destek vektör makineleri	%62,50	%55,30
Boosted Trees	%61,40	%54,30
Kernel Naive Bayes	%61,40	%54,00
Gaussian Naive Bayes	%59,50	%52,70
Karar Ağacı (Fine Tree)	%60,30	%52,30

Gebelikte alınan tam kan sayımı ile bebeğin yoğun bakıma alınmasını tahmin etme başarısı en yüksek olan üç yöntemin doğruluk değerleri;

- Kuadratik çekirdeğe sahip destek vektör makineleri için %61,55
- Medium Gaussian çekirdeğe sahip destek vektör makineleri için %61,85
- Gentle AdaBoost yöntemi için %62,15

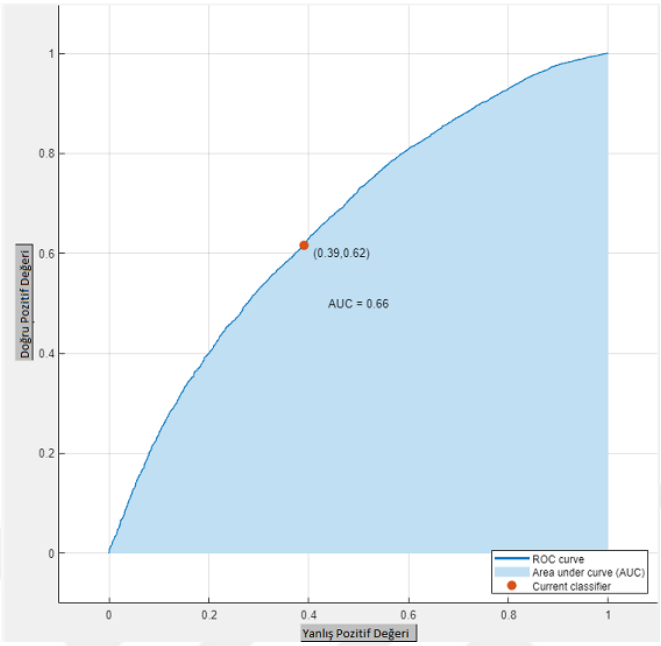
olarak tespit edilmiştir. Bu üç yönteme ait karmaşıklık matrisleri ise Tablo 3.7'de verilmiştir.

1982 yılında yapılan araştırmada ROC eğrisinin altında kalan alanın (AUC) tanı testlerinin performans değerlendirmesinde kullanılabileceği gösterilmiştir [19]. ROC eğrisinin yapısından dolayı koordinat değerleri 0 – 1 arasındaki değerleri alabilmektedir. Hatasız bir tahminlemede eğrinin altında kalan alan 1 iken, AUC değeri 0,5 ve altında olduğunda ise tanı testinin tahminleme performansının başarısız olduğu anlaşılır.

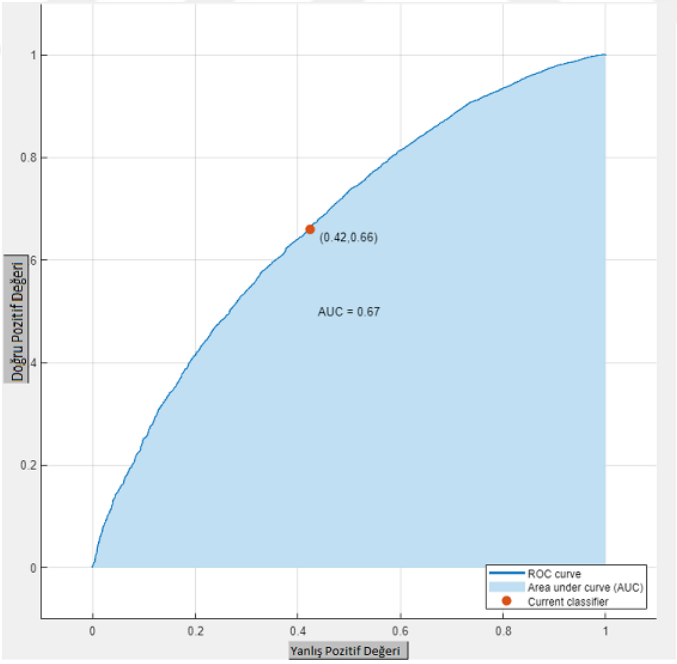
Tablo 3.8 AUC Değerleri için Aralıklar [20]

Mükemmel	İyi	Orta	Zayıf	Başarısız
0,9-1	0,8-0.9	0,7-0,8	0,6-0,7	0,5-0,6

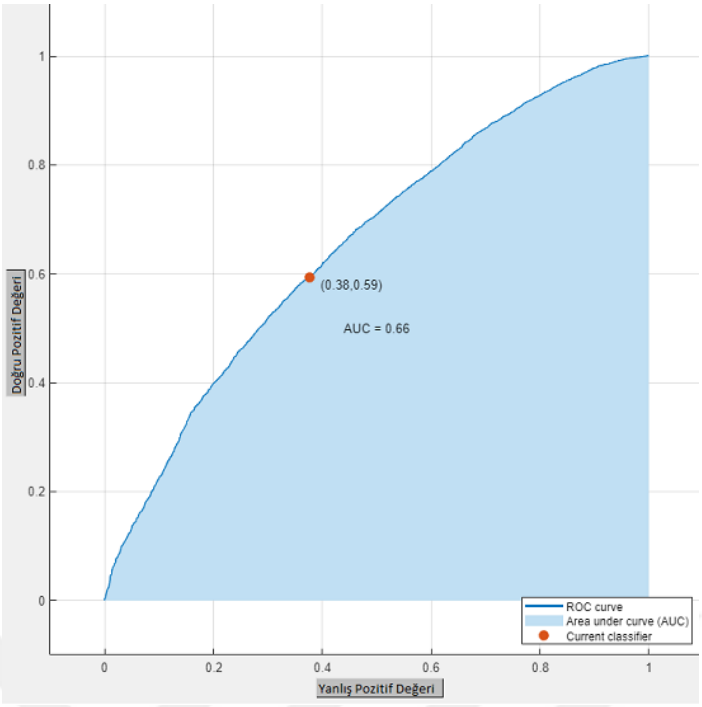
AUC değeri yükseldikçe performansının mükemmele yaklaştığı sonucuna varılır. AUC değerleri için Tablo 3.8'deki derecelendirme aralığı kullanılabilir [20].



Şekil 3.2 Kuadratik çekirdeğe sahip destek vektör makineleri için ROC eğrisi ve AUC değeri



Şekil 3.3 Medium Gaussian çekirdeğe sahip destek vektör makineleri için ROC eğrisi ve AUC değeri



Şekil 3.4 Gentle AdaBoost yöntemi için ROC eğrisi ve AUC değeri

Tablo 3.8'de verilen başarı derecelendirmesine göre, en başarılı üç yönteme ait ROC eğrisi ve AUC değerleri için sınıflandırma başarısının zayıf olduğu görülmektedir.

Tablo 3.9 En Başarılı Yöntemler için Karmaşıklık Matrisleri

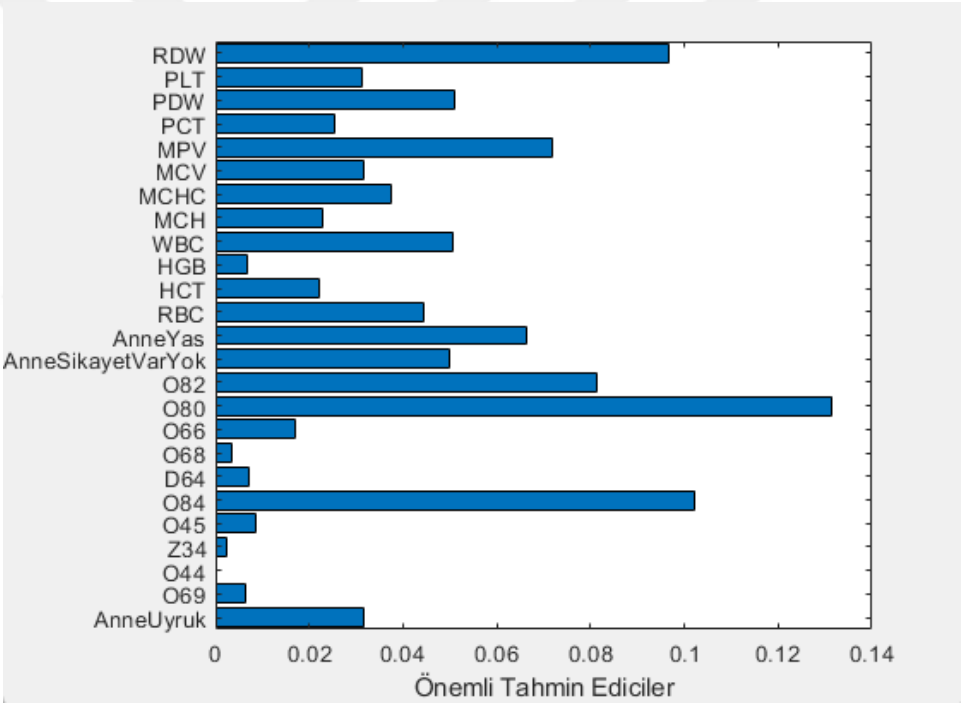
	Tahmin		
Gerçek	Kuadratik Çekirdeğe sahip Destek Vektör Makinesi Tahminlemesi	**Kontrol**	**Hasta**
	Kontrol	%62,70	%37,30
	Hasta	%38,20	%61,80

	Tahmin		
Gerçek	Medium Gaussian çekirdeğe sahip destek vektör makineleri tahminlemesi	**Kontrol**	**Hasta**
	Kontrol	%66,40	%33,60
	Hasta	%41,90	%58,10

Tablo 3.9 devamı En Başarılı Yöntemler için Karmaşıklık Matrisleri

	Tahmin		
Gerçek	Gentle AdaBoost (Ensemble Method) Tahminlemesi	**Kontrol**	**Hasta**
	Kontrol	%71,00%	%29,00
	Hasta	%39,30%	%60,70

Karar ağacı yöntemiyle sınıflandırma yapılırken özniteliklerin belirlenen tahmin edici önemleri Şekil 3.2'de gösterilmektedir. Buna göre; bebeğin doğum sonrası yoğun bakıma girme tahminlemesi için O80, O84 ve RDW özniteliklerinin görece daha önemli etkenler olduğu görülmektedir.



Şekil 3.5 Karar Ağacı Yöntemine göre Özniteliklerin Etki Analizi

Bebeklerin yoğun bakıma yatmasının çok farklı sebepleri bulunmaktadır. Bu nedenle ICD kodlarına ve anne şikayetlerine göre elde edilen sınıflandırma sonuçları ayrıca ele alınmıştır.

Tablo 3.10 ICD Kodlarına Göre Elde Edilen Doğruluk Değerleri

Kuadratik Kernel Destek Vektör Makineleri				
	Doğru Tahmin	Özgünlük (FPR)	Duyarlılık (TPR)	Dengesel Doğruluk
Doğum süreci ve doğum, kordon komplikasyonları (O69)	4 (%40)	%100	%0	%50
Plasentanın rahim içinde doğum kanalını kapatacak şekilde yerleşmesi (O44)	12 (%60)	%100	%50	%75
Normal gebeliğin gözlemi (Z34)	22 (%100)	%100	%100	%100
Plasentanın erken ayrılması (O45)	22 (%41)	%100	%16	%58
Çoğul doğum (O84)	38 (%38)	%100	%3	%52
Anemi (D64)	81 (%73)	%100	%70	%85
Doğum süreci ve doğum, fetal stresle komplike (O68)	97 (%36)	%79	%30	%54
İlerlemeyen doğumlar (O66)	236 (%82)	%25	%84	%55
Tek spontan doğum (O80)	615 (%98)	%8	%99	%54
Tek doğum, sezaryen ile (O82)	1.403 (%41)	%85	%36	%60

Farklı ICD kodlarına göre, kuadratik çekirdeğe sahip destek vektör makineleri için elde edilen sınıflandırma sonuçları Tablo 3.9'de verilmiştir. Z34 ICD kodu ile etiketlenen 22 adet örneğin tümü doğru tahmin edilmiştir. Ayrıca tabloda görüldüğü üzere, D64 ve O44 kodu ile etiketlenen örneklerin dengesel doğruluk değerleri (sırasıyla %85 ve %75) daha yüksek çıkmıştır.

Tablo 3.11 Farklı Şikayet Türlerine Göre Elde Edilen Doğruluk Değerleri

Kuadratik Kernel Destek Vektör Makineleri				
	Doğru Tahmin	**Özgünlük (FPR)**	**Duyarlılık (TPR)**	**Dengesel Doğruluk**
Sancı	3.606 (%78)	%42	%81	%62
EMR	479 (%29)	%93	%21	%57
Fetal hareketin azalması	121 (%65)	%79	%64	%71
Kaşıntı	7 (%64)	%100	%60	%80
Kanama	72 (%62)	%89	%57	%73
Tansiyon	27 (%36)	%100	%71	%62

Kuadratik çekirdeğe sahip destek vektör makineleri için elde edilen sınıflandırma sonuçları farklı şikayet türlerine göre incelendiğinde hesaplanan sonuçlar Tablo 3.10'da verilmiştir. Kaşıntı, kanama ve fetüs şikayetleri için daha yüksek doğrulukta sınıflandırma sonuçları elde edilmiştir. Kaşıntı için doğruluk yüzdesi %64, dengesel doğruluk oranı %80 olarak; kanama için doğruluk yüzdesi %62, dengesel doğruluk oranı %73 olarak ve fetüs tahminlemesi için doğruluk yüzdesi %65, dengesel doğruluk oranı %71 olarak bulunmuştur.

4
SONUÇ VE ÖNERİLER

Bu çalışmada anneden alınan tam kan sayım ölçüleri, annenin şikayetleri ve yaşından oluşan veri seti kullanılarak, bebeklerin doğumdan önce YYBÜ'ye kabul edilip edilmeme durumları incelenmiş ve çeşitli makine öğrenmesi teknikleriyle ilgili sınıflandırma problemi ele alınmıştır. Bu kapsamda, öncelikle 2014-2019 yılları arasındaki hastane verileri toplanmış ve toplanan veriler veri temizleme, dönüştürme ve birleştirme gibi gerekli ön işlemlerden geçirilmiştir. Sonrasında, kan parametreleri, annenin yaşı ve anne şikayetleriyle yoğun bakıma yatma oranları arasındaki ilişki istatiksel olarak analiz edilmiştir. Son olarak ise, çeşitli makine öğrenmesi yöntemleri kullanılarak sınıflandırma problemi incelenmiştir.

Ön işleme tabi tutulan veri setinde, bebeği yoğun bakıma yatmayan annelere ait örnekler "*kontrol"* olarak etiketlenirken, bebeği yoğun bakıma yatan annelere ait örnekler ise "*hasta"* olarak etiketlenmiştir. Anne yaşı ve annenin kan sayımı parametreleri gibi nümerik öznitelikler için, iki sınıfın ortalama değerleri arasında fark olduğu hipotezi z testi ve Mann-Whitney U testi ile sınanmıştır. P değerinin 0.05'ten küçük olduğu durumlar istatistiksel olarak anlamlı kabul edilmiştir. Buna göre bebeği yoğun bakıma yatan ve yatmayan annelerin yaş ortalaması ve kan sayımı parametrelerinden RBC, HCT, MCH, MCV, MPV, PCT ve RDW ortalamaları için anlamlı istatistiksel farklar bulunmuştur. Ayrıca, şikayet varlığı/yokluğu ve çeşitli ICD kodlarına göre oluşturulan kategorik özniteliklerin oranlarının kontrol ve hasta olarak etiketlenen sınıflar için farklı olduğu hipotezi z testi ile sınanmıştır. Kategorik veriler incelendiğinde şikayet, Z34, O45, O84, O68, O66, O80 ve O82 ICD kodları için bebeği yoğun bakıma yatan ve yatmayan annelerin parametrelerin 'Evet' oranları arasında anlamlı farklar bulunmuştur.

Bebeklerin yoğun bakıma girip/girmemesinin prenatal dönemde tahmin edilmesi için Matlab Sınıflandırma Öğreticisi uygulamasında bulunan destek vektör makinelerine, karar ağaçlarına, Naive Bayes yöntemine ve topluluk sınıflanrıcılarına dayanan çeşitli algoritmalar kullanılmıştır. Diğer algoritmalara görece daha başarılı olan kuadratik ve Gaussian çekirdeğe sahip destek vektör makineleri ve Gentle

AdaBoost yöntemi yaklaşık olarak %62 oranında doğru sınıflandırma yapmıştır. Sınıflandırma sonuçları; kaşıntı (%80), kanama (%73) ve fetüs (%71) gibi şikayetlerle veya Z34 (%100), D64 (%85) ve O44 (%75) gibi ICD kodları ile etiketlenen örnekler için incelendiğinde daha yüksek dengeli doğruluk sonuçları elde edilmiştir.

Veri setindeki bir çok öznitelik için iki sınıf arasında istatistiksel olarak anlamlı farklar tespit edilmiştir. Bebeklerin yoğun bakıma yatma nedenlerindeki çeşitlilik ve veri setindeki sınıflara ait örnek sayısındaki dengesizlik makine öğrenmesi yöntemleriyle elde edilen başarı oranının yüksek seviyelere çıkmasını engellemiştir. Veri sayısındaki örnek sayısının artırılması ve özellikle şikayetlere bağlı daha fazla sayıda özniteliğin veri setine eklenmesi ile başarı oranının artması öngörülmektedir.

KAYNAKÇA

[1] H. Tamim, H. Beydoun, "Predicting neonatal outcomes: birthweight, body mass index or ponderal index", 2004, s. 509-13. doi: 10.1515/JPM.2004.120.

[2] J. L. Barkin, "Evaluation of Maternal Functioning in Mothers of Infants Admitted to the Neonatal Intensive Care Unit", Jul. 2019, ss.941-950, doi: 10.1089/jwh.2018.7168

[3] S. Arslan, A. Bülbül, A. Ş. Aslan, E. K. Baş, M. Dursun, S. Uslu, A. Nuhoğlu, "Yenidoğan yoğun bakım ünitesinde beş yıllık sürede (2007-2011) neonatal ölüm nedenleri", Oct. 2012, doi: 10.5350/SEMB2013470104

[4] T. Kitano, K. Takagi, "Prenatal predictors of neonatal intensive care unit admission due to respiratory distress", Jun. 2018, ss. 560-564, doi: 10.1111/ped.13574.

[5] A. A. Özdemir, Y. Elgörmüş, "Yenidoğan Yoğun Bakım Ünitesinde Yatan Hastalarda Ölüm Nedenlerinin Değerlendirilmesi", Ülke: Türkiye, 2016, ss. 46-50

[6] H. Kaya, K. Köymen, "Veri Madenciliği Uygulama Alanları", Ülke: Türkiye, 2008, ss. 159-164

[7] Veri Ön İşleme, Sayı 21, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, A. Oğuzlar, Türkiye, 2003, ss. 67-76

[8] J. Han, J. Pei ve M. Kamber, Data Mining, SA, 2006

[9] D. M. Lane, Introduction to Linear Regression, An Interactive Multimedia Course of Study ch. 14

[10] Destek Vektör Makineleriyle Sınıflandırma Problemlerinin Çözümü İçin Çekirdek Fonksiyonu Seçimi, Sayı 1, Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi, S. Ayhan ve Ş. Erdoğmuş, Türkiye, 2014, ss: 175-201

[11] M. Üstüner, F. B. Sanli, F. B. Balcik, M. T. Esetlili, Destek vektör Makineleri Tekniği ile Sınıflandırma, TUFUAB 2013

[12] T. Kavzaoğlu, İ. Çölkesen, Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi, 2010, Harita Dergisi. 76. 73-82.

[13] O. Öztanır, Predictive Maintenance By Using Machine learning, M.S. thesis, Hacettepe University, Ankara, Turkiye, 2018

[14] Heyelan Duyarlılığının İncelenmesinde Regresyon Ağaçlarının Kullanımı, Harita Dergisi T. Kavzoğlu, E. K. Şahin, İ. Çölkesen, Türkiye, 2012

[15] Gini Algoritmasını Kullanarak Karar Ağacı Oluşturmayı Sağlayan Bir Yazılımın Geliştirilmesi, Sayı 3, Bilişim Teknolojileri Dergisi, M. F. Adak, N. Yurtay, Türkiye, 2013, ss. 1-6

[16] E. TASCI, E. ONAN, K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi, pp. 2, 2018

[17] Yapay Sinir Ağları ve Yapay Zekâ'ya Genel Bir Bakış, Sayı 2, Takvim-i Vekayi, K. Öztürk, M. E. Şahin, Türkiye, ss. 25-36, 2018

[18] D. Birant, Farklı Bağlantı Yöntemleri ile Hiyerarşik Kümeleme Topluluğu, S.Ü. Müh. Bilim ve Tekn. Derg., c.7, s.1, ss. 154-164, 2019

[19] J. A. Hanley ve B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36, 1982

[20] N. A. Obuchowski, Computing sample size for receiver operating characteristic studies. Investigative Radiology, 29(2), 238-243, 1994

TEZDEN ÜRETİLMİŞ YAYINLAR

Uluslararası Bildiri

1. S. GESOGLU, B. ASLANYÜREK, E. AYDIN, A. UZUN, "Yeni Doğanların Yoğun Bakım İhtiyacının Makine Öğrenmesi Teknikleri ile Prenatal Tahmin Edilmesi., Uluslarası Ege Sağlık Alanları Sempozyumu, Dec. 2022. Online