T.C.
SELÇUK ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

SINIFLANDIRMA PROBLEMLERİNDE LOJİSTİK REGRESYON İLE MAKİNE ÖĞRENMESİ ALGORİTMALARININ PERFORMANSLARININ KARŞILAŞTIRILMASI: VETERİNERLİK ALANINDA BİR UYGULAMA

Mert DEMİRSÖZ

YÜKSEK LİSANS TEZİ

BİYOİSTATİSTİK ANABİLİMDALI

Danışman

Dr. Öğr. Üyesi Harun YONAR

KONYA-2023

ÖNSÖZ

Yüksek lisans eğitimi ve tez çalışmamda her zaman desteği ile yanımda olan, bana kıymetli zamanını ayıran saygı değer danışman hocam Dr. Öğr. Üyesi Harun YONAR'a yardımları için çok teşekkür ederim.

Hayatım boyunca bana desteklerini esirgemeyen, beni bu günlere getiren, çok kıymetli annem ve babama, bu çalışma süresince gösterdikleri sabır, anlayış ve sevgi için çok sevgili eşime sonsuz teşekkür ederim.

İÇİNDEKİLER

Sayfa

SİMGELER VE KISALTMALAR ..... **vii**
ÖZET ..... **viii**
SUMMARY ..... **ix**
1. GİRİŞ **1**
1.1 Literatür Taraması 4
1.2 Lojistik Regresyon 11
1.2.1 İkili (Binary) Lojistik Regresyon 11
1.2.2 Lojistik Regresyon Analizinde Model Uyum İyiliği 16
1.3 Makine Öğrenmesi Çalışma Prensibi 17
1.4 Makine Öğrenmesi Bölümleri 18
1.4.1 Öğrenme 18
1.4.2 Öğrenme Yöntemleri 19
1.4.3 Performans Metrikleri 26
2.GEREÇ VE YÖNTEM **28**
2.1 Gereç 28
2.2 Yöntem 28
3.BULGULAR **30**
3.1. Değişkenlerin Keşifçi Veri Analizi 30
3.2. Lojistik Regresyon Sonuçları 33
3.3. Makine Öğrenmesi Algoritmaları Sonuçları 38
3.4. Lojistik Regresyon ile Makine Öğrenmesi Algoritmaları Performans Sonuçları 39
4.TARTIŞMA **43**
5. SONUÇ VE ÖNERİLER **45**
6. KAYNAKLAR **46**
7. EKLER **51**
EK-A: Etik Kurul Kararı 51
EK-B: Turnitin Raporu 52
9. ÖZGEÇMİŞ **53**

SİMGELER VE KISALTMALAR

KNN : K-Nearest Neighbor (K-En Yakın Komşu)

RF : Random Forest

NB : Naive Bayes

JASP : Jeffrey's Amazing Statistics Program

LDA : Lineer Diskriminant Analizi

ROC : Receiver Operating Characteristic

AUC : Area Under Curve

pH : Hidrojen İyon Konsantrasyonu

pCO_2 : Kısmi Karbondioksit Basıncı

pO_2 : Kısmi Oksijen Basıncı

Na^+ : Sodyum

K^+ : Potasyum

WBC : Akyuvar

LYM : Lenfosit

RBC : Alyuvar

HCT : Hemotokrit

HGB : Hemoglobin

β : Regresyon Katsayısı (Lojit Katsayı)

ÖZET

T.C.

SELÇUK ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

Sınıflandırma Problemlerinde Lojistik Regresyon ile Makine Öğrenmesi Algoritmalarının Performanslarının Karşılaştırılması: Veterinerlik Alanında Bir Uygulama

Mert DEMİRSÖZ

Danışmanı: Dr. Öğr. Üyesi Harun YONAR

Biyoistatistik Anabilim Dalı

YÜKSEK LİSANS TEZİ / KONYA-2023

Bu çalışmada, buzağılara ait veriler kullanılarak oluşturulan iki farklı modelde, buzağı ishalinin tespit edilmesinde lojistik regresyon, Boosting, K En Yakın Komşu (KNN), Lineer Diskriminant ve Random Forest (RF) makine öğrenmesi sınıflandırma algoritmalarının performansları karşılaştırıldı.

Çalışmada, toplam 220 buzağıya ait klinik bulgu, kan gazı ölçümleri, hematolojik ve biyokimyasal 14 değişkene ait veriler kullanıldı. Buzağı ishalinde etkili olduğu düşünülen 14 değişken modelde iken lojistik regresyon analizinde enter yöntemi kullanılarak analizler gerçekleştirildi. Ardından stepwise yöntemi ile en uygun değişkenlerin bulunduğu model oluşturuldu. Her iki modelde bulunan değişkenler makine öğrenmesi algoritmaları için ayrı ayrı uygulandı. Lojistik regresyon ile makine öğrenmesi algoritmalarının tahmin performansları doğruluk, AUC, duyarlılık, seçicilik ve F ölçüsü metrikleri kullanılarak karşılaştırıldı.

Klinik bulgu, kan gazı ölçümleri, hematolojik ve biyokimyasal verilere ait 14 değişken kullanılarak oluşturulan modelde doğruluk değerleri; lojistik regresyon için %89,1, Boosting için %88,2, KNN için %85, Lineer Diskriminant için %84,1 ve RF için %95,5 olarak elde edildi. Stepwise yöntemi ile oluşturulan 7 değişkenli modelde ise doğruluk değerleri lojistik regresyon için %90,5, Boosting için %89,1, KNN için %87,3, lineer diskriminant için %86,8 ve RF için %97,7 olarak elde edildi. RF algoritmasının her iki modelde hem lojistik regresyondan hem de diğer sınıflandırma algoritmalarından daha yüksek doğruluk değerine sahip olduğu tespit edildi.

Bu çalışmada öncelikle makine öğrenmesi algoritmalarının sağlık ve veterinerlik alanında yapılacak modelleme çalışmalarında kullanılabilirliği incelenmiştir. Ayrıca modelleme çalışmalarında karşılaşılabilecek değişkenlerin istatistiksel yöntemlerden kaynaklı dışlanma sorununun ortadan kaldırılabileceği ve makine öğrenmesi algoritmalarının modellemede katkı sağlayacağı öngörülmektedir.

Anahtar Sözcükler: Lojistik Regresyon; Makine Öğrenmesi, Sınıflandırma

SUMMARY

REPUBLIC OF TURKEY
SELÇUK UNIVERSITY
HEALTH SCIENCES INSTITUTE

Comparison of the Performance of Logistic Regression and Machine Learning Algorithms in Classification Problems: An Application in the Field of Veterinary Medicine

Mert DEMİRSÖZ

Department of Biostatistics

MASTER THESIS / KONYA-2023

In this study, the performances of logistic regression, Boosting, K Nearest Neighbor (KNN), Linear Discriminant and Random Forest (RF) machine learning classification algorithms in detecting calf diarrhea were compared in two different models created using data from calves.

In the study, clinical findings, blood gas measurements, hematological and biochemical data of 14 variables of 220 calves were used. While 14 variables thought to be effective in calf diarrhea were in the model, analyzes were performed using the enter method in logistic regression analysis. Then, the model with the most suitable variables was created with the stepwise method. The variables in both models were applied separately for machine learning algorithms. The prediction performances of logistic regression and machine learning algorithms were compared using the metrics of accuracy, AUC, sensitivity, selectivity, and F measure.

Accuracy values in the model created by using 14 variables of clinical findings, blood gas measurements, hematological and biochemical data; 89,1% for logistic regression, 88,2% for Boosting, 85% for KNN, 84,1% for Linear Discriminant and 95,5% for RF. In the 7 variable model created by the Stepwise method, the accuracy values were 90,5% for logistic regression, 89,1% for Boosting, 87,3% for KNN, 86,8% for linear discriminant and 97,7% for RF. It was determined that the RF algorithm had higher accuracy than both logistic regression and other classification algorithms in both models.

In this study, first of all, the usability of machine learning algorithms in modeling studies in the field of health and veterinary medicine have been examined. In addition, it is predicted that the exclusion problem of variables that may be encountered in modeling studies can be eliminated from statistical methods and machine learning algorithms will contribute to modeling.

Key words: Logistic Regression; Machine Learning; Classification

1. GİRİŞ

Makine öğrenmesi, yapay zekâda kullanılan sayısal öğrenme ve model tanıma çalışmalarından faydalanarak geliştirilmiş ve yapay zekânın alt dalı olarak ortaya çıkan bir yöntemdir. Yapısal olarak öğrenebilen ve veri üzerinden tahminler yapabilen makine öğrenmesi, algoritmaların çalışma prensiplerini araştıran bir sistemdir. Algoritmalar, açık programlama yerine veri setinden faydalanarak öğrenme yeteneği kazanır. Öğrenme sonucunda elde edilen tahminleri ve kararları gerçekleştirebilecek bir model oluşturur. Yani makine öğrenmesi deneme yanılma yöntemi ile doğruyu yapmaya çalışan ve doğruyu yapmak için veri setinden faydalanan bir yöntemdir. Temelde bilgisayar, matematik ve istatistik bilimlerinden yararlanan makine öğrenmesi, günümüzde yaygın olarak kullanılmaktadır. Makine öğrenmesinin yapay zekâ açısından bulunduğu konum ise Şekil 1.1'de verilmiştir.



Şekil 1.1. Makine öğrenmesi yönteminin yapay zekâ ile ilişkisi

Massachusetts Teknoloji Enstitüsü Bilgisayar Bilimleri laboratuvar yöneticilerinden Edward Fredkin BBC ile söyleşisinde "Tarihte en büyük üç olayından birincisi kâinatın oluşumu, ikincisi yaşamın başlangıcının olması ve üçüncüsü de yapay zekânın ortaya çıkmasıdır" ifadesi ile yapay zekânın dünya tarihinde önemli bir yerde olduğunu belirtmiştir (Pirim 2006).

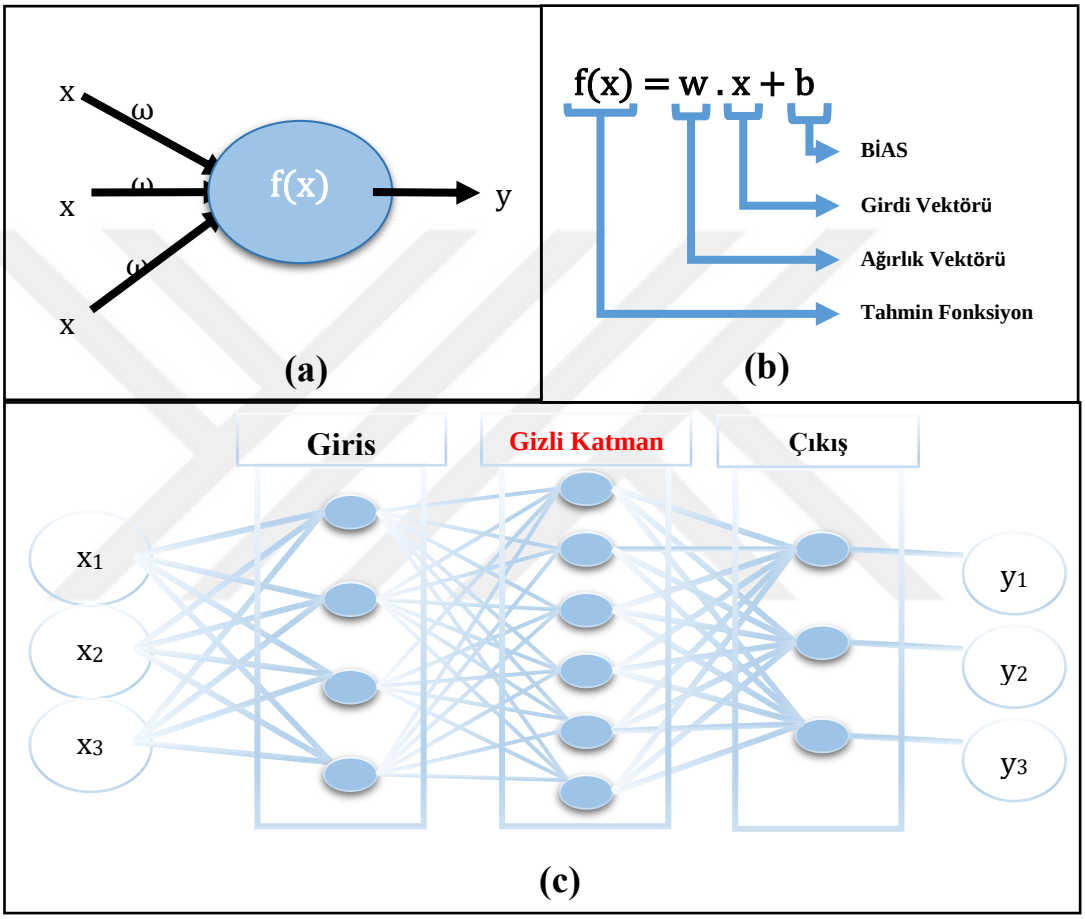
Yapay zekâ ve makine öğrenmesi İngiliz matematikçi, bilgisayar bilimcisi ve şifre bilimcisi olan Alan Turing tarafından 2. Dünya savaşı sırasında gizli mesajları çözmek için elektromekanik aygıtlar üretilmesi ile ortaya çıkmıştır. Literatürde ise yapay zekâ ve makine öğrenmesinin ortaya çıkış tarihi 1950 olarak kabul edilmektedir. Yine bu yılda yayınlanan "Ben Robot" isimli bilim kurgu romanı okuyucularına yapay zekânın gelişimi hakkında çeşitli fikirler vermiştir. 1951 yılında Manchester Victoria Üniversitesi ve bir özel firmanın iş birliği ile üretilen Ferranti-1 adlı makineyi kullanarak Christopher Strachey dama oynamayı, Dietrich Prinz ise satranç oynamayı programlama yardımı ile bu makineye öğretmişlerdir (Kıymaz 2021).

Newell ve Simon (1956), Mantık Teorisi Makinesi (Logic Theory Machine) adında bir programlama geliştirmişlerdir. Bu programlama, Whitehead ve Russell'in üç cilt halinde yayınladıkları Principia Mathematica kitabında bahsedilmekte olan Mantık Teorisini, sembolik mantık ile kanıtlayabilen bir bilgisayar programıdır. Yine 1956 yılında düzenlenmiş olan Dartmouth Konferansı için 2 Eylül 1955 tarihinde John McCarthy ve birlikte çalıştığı bilim insanları tarafından yapay zekâ terimini tanıtan bir proje önerilmiştir. Öneri sonucunda bu alan yapay zekâ olarak adlandırılmıştır (Kıymaz 2021).

Bilgisayar bilimci ve psikolog Frank Rosenblatt (1958), canlı sinir hücrelerinin çalışma prensiplerinden yola çıkarak Algılayıcı modelini oluşturmuştur. Algılayıcı, en basit tek katmanlı sinir ağı modelidir. Yapay sinir ağlarında yapılan temel işlem; modelin en iyi sonucunu, ağırlık vektörü ve bias (yanlılık) parametrelerini dâhil ederek ortaya çıkartmaktır. Algılayıcı modeli Şekil 1.2. (a), tahmin fonksiyonu ise Şekil 1.2. (b) görselinde verilmiştir.

1962 yılında Joseph Frederick Engelberger ve George Devol tarafından kurulan Unimation şirketi, dünyanın ilk robotik şirketidir. Bu şirket yapay zekâ ve makine öğrenmesi algoritmaları ile endüstriyel olarak kullanılacak akıllı makineler tasarlamış ve George Devol tarafından patentlenen robotik kollar üretmiştir . Marvin Minsky (1967), yapay zekâya yol gösterici çalışmalarından olan "Hesaplama: Sonlu ve sonsuz makineler" kitabını yayınlamış ardından Minsky ve Papert (1969) ile "Algılayıcılar: Hesaplamalı Geometriye Giriş" kitabını yazmışlardır. Bu kitapta XOR

problemine değinerek algoritmalarda çok katmanlı algılayıcılar (gizli katman) kullanımının önünü açmıştır. Sonrasında Rumelhart ve ark (1986) tarafından geliştirilen Hata Yayma Modeli denilen ara katmanlı modeller oluşturulmuştur. Bu modellerde öğrenme metotlarından biri olan Delta Öğrenme Kuralı kullanılmıştır. Çok katmanlı algılayıcılar, özellikle sınıflandırma problemlerinde etkin çalışmaktadır. Çok Katmanlı Ağların yapısı Şekil 1.2. (c)'deki gibidir.



Şekil 1.2. Algılayıcı **(a)**, tahmin fonksiyonu **(b)** ve çok katmanlı ağ **(c)** modeli (Rosenblatt 1958; Rumelhart ve ark 1986)

Hinton ve Salakhutdinov (2006), "Yapay Sinir Ağları ile veri boyutunun azaltılması" isimli makaleyi yayınlamışlardır. Bu çalışma literatürde derin evrişimsel sinir ağı olarak bulunmaktadır. Günümüzde birçok kaynak tarafından çeşitli ve yüksek boyutlu veriler kısa sürede üretilmektedir. 2014 yılından bu yana dünya genelindeki verilerin %90'ının son 2 senede üretildiği ayrıca her gün 2,5 eksabayt boyutunda veri oluştuğu açıklanmıştır. 2003 yılı boyunca üretilen veri miktarı 5 eksabayt iken,

günümüzde aynı boyuttaki bir veri ortalama 2 gün içerisinde üretilmektedir. Bu durum her geçen gün katlanarak artmaktadır (Aktan 2018; Sagiroglu ve Sinanc 2013).

Büyük veri söz konusu olduğunda makine öğrenmesi algoritmaları tahmin ve model oluşturmada istatistiksel yöntemlere alternatif oluşturmaktadır (Aktan 2018; Cihan ve ark 2017; Gök 2017).

Bu çalışmanın amacı, buzağılardan alınan hemogram, kan gazı, serum biyokimya ölçümleri ve hayvanlara ilişkin bilgiler kullanılarak buzağı ishali tespitinde lojistik regresyon, Boosting, KNN, lineer diskriminant ve RF makine öğrenmesi algoritmalarının sınıflandırma problemlerindeki performanslarını değerlendirmektir.

Bu amaç doğrultusunda ikili (binary) lojistik regresyonda enter ve stepwise metotları kullanılarak modeller oluşturulmuştur. Buzağı ishalinde etkili olduğu düşünülen değişkenler modelde iken enter yöntemi kullanılarak analizler gerçekleştirildi. Ardından stepwise yöntemi ile en uygun değişkenlerin bulunduğu model oluşturuldu. Daha sonra makine öğrenmesi algoritmaları, enter modelindeki değişkenler ve stepwise modelindeki değişkenler için ayrı ayrı uygulandı. Lojistik regresyon ile makine öğrenmesi algoritmalarında gerçek sonuçlarla tahmin sonuçları arasındaki performanslar doğruluk, AUC, duyarlılık, seçicilik ve F ölçüsü metrikleri ile incelendi.

1.1 Literatür Taraması

Alkhamis ve ark (2021), yaptıkları çalışmada, 2019-2020 yılları arasında geviş getiren hayvanlarda görülen Mavidil (Bluetongue) hastalığına etki eden serotipleri incelemişlerdir. Bu incelemede virüs görülen 24.245 hayvan üzerinde RF Destek Vektör Makineleri ve XGBoost sınıflandırma algoritmaları kullanılmıştır. Avrupa'da kaydedilen en yaygın serotipleri inceleyerek Mavidil virüsü salgın riskinin serotipin çevresel faktörlere özgü olup olmadığını test etmişlerdir. Sonuç olarak makine öğrenmesi algoritmaları, Mavidil virüsü (doğruluk > 0,87) için tüm serotiplerin etkilerini yüksek tahmin performansıyla ve güçlü doğrusal olmayan ilişkilerle ortaya koymuşlardır.

Romero ve ark (2021), yaptıkları çalışmada, İngiltere'deki sığır tüberkülozunun açıklayıcı modellerini oluşturmak için RF ve lojistik regresyon sınıflandırma algoritmalarını kullanmıştır. Sınıflamalar üç risk grubuna göre yapılmış ve RF algoritması düşük risk gruplarında, lojistik regresyon algoritması ise diğer iki risk grupları için mükemmel sınıflandırma ortaya koyduklarını bildirmişlerdir.

Fadul Pacheco ve ark (2021), yaptıkları çalışmada, ilk emzirmelerinde klinik mastit durumları pozitif olan ineklerin mastit başlangıcını tahmin etmek için RF, NB ve XGBoost sınıflandırma algoritmaları kullanmışlardır. Bu algoritmalardan en iyi tahmin performansını %72 doğruluk oranı ile RF yöntemi verdiği görülmüştür.

Pandit ve ark (2021), yaptıkları çalışmada, süt ineklerinden alınan dışkı örneklerinde bulunan üç farklı bakteri türünü (Salmonella, Commensals, Escherichia Coli) çoklu ilaç direncine göre test etmişlerdir. Çoklu ilaç direncine göre üç bakteri türünü en uygun şekilde sınıflandırmak ve değişken önemlerini belirlemek için Karar Ağaçları, RF ve XGBoost algoritmaları kullanılmıştır. Karar ağacı algoritması için tüm sürünün ortalama süt üretimi, RF algoritması için seçilen hayvan sayısı, aylık seçilme sıklığı ve yüzdesi, sürü büyüklüğü ve siyah alaca ineklerinin oranı, XGBoost algoritma için ise aylık seçilme sıklığı ve yüzdesi en önemli değişkenler olarak belirlenmiştir.

Esener ve ark (2021), yaptıkları çalışmada, 24 çiftlikte sığır mastit teşhisi konan 67 inekten toplanan 82 Staphylococcus aureus bakterisi üzerinde çalışmışlar ve NB, Destek Vektör Makineleri, Karar Ağacı, RF, Yapay Sinir Ağları, XGBoost ve lojistik regresyon algoritmaları kullanılmıştır. Bu çalışmada araştırmacılar makine öğreniminin sonuçları üzerine geliştirilen özel bir biyoinformatik boru hattı aracılığıyla, moleküler belirleyiciler ve antibiyotik direnç fenotiplerinin altında yatan mekanizma hakkında yeni bilgiler elde etmişlerdir.

Sykes ve ark (2022), yaptıkları çalışmada, biyolojik güvenlik uygulamaları, çiftlik demografisi ve önceki salgınlarla ilgili anket verilerini baz almışlardır. Bu verilerle çiftliklerdeki domuz üreme ve solunum sendromunu sınıflandırmak için Destek Vektör Makineleri, RF, XGBoost sınıflandırma algoritmaları kullanmışlardır. Makine öğrenimi algoritma sonuçlarına göre ciro ve çalışan sayısı, domuz

tesislerindeki domuzların yoğunluğu, domuz taşıma araçları, yola uzaklık ve çiftliklerdeki üretim türü salgınlar için önemli değişkenler olduğunu ortaya koymuşlardır.

Liu ve ark (2021), yaptıkları çalışmada, dondurularak kurutulmuş kalp kapakçıklarında oksidatif hasarın belirlenmesini amaçlamışlardır. Bu amaç doğrultusunda TensorFlow yöntemi ile algoritma eğitimi ve değerlendirmesi için Yapay Sinir Ağı uygulanırken performansı değerlendirmek için çapraz doğrulama yöntemini uygulamışlardır. Sonuç olarak, sükrozun dondurarak kurutmanın neden olduğu hücresizleştirilmiş kalp kapakçıklarındaki oksidatif hasarın boyutunu doza bağlı bir şekilde en aza indirdiği gösterilmiştir.

Carslake ve ark (2020), yaptıkları çalışmada, on üç sütten kesilmiş süt buzağısının davranışlarını video kameralarla izlemiş ve davranışları sınıflandırmak için bir XGBoost topluluğu öğrenme algoritmasını uygulamışlardır. Algoritma lokomotor oynamayı %99,73 doğruluk, kendi kendini temizlemeyi %98,18 doğruluk, geviş getirmeyi %94,47 doğruluk, besleyici olmayan emmeyi %94,96 doğruluk, besleyici emzirmeyi %96,44 doğruluk, aktif yatmayı %90,38 doğruluk ve aktif olmayan yalan söylemeyi %90,38 doğruluk ile tahmin etmiştir.

Brock ve ark (2021), yaptıkları çalışmada, hayvancılık sistemlerindeki sürü tiplerini sınıflandırmak için uzman bilgisi ile Self Organising Maps adı verilen bir makine öğrenimi algoritmasını birleştirerek yeni bir yaklaşım ortaya koymuşlardır.

Charan ve ark (2021), yaptıkları çalışmada, köpeklerin olgunluğunu belirlemek, kırıkları tarihlemek ve son olarak köpeklerde uzun kemik kırıklarını tespit etmek için sinir ağları algoritmaları olan AlexNet, ResNet-50 ve GoogLeNet ve karşılaştırma için destek vektör makinelerini kullanmışlardır. ResNet-50, AlexNet ve GoogLeNet, sırasıyla 0.75, 0.80 ve 0.88 F1 puanlarıyla üç görev için en başarılı algoritmalar olarak elde edilmiştir.

Mani ve ark (2021), yaptıkları çalışmada, 60 gün boyunca peletlenmiş yem, probiyotik peletlenmiş yem, probiyotik inci boncuklar ve probiyotik pirinç pufu ile beslenen sazan balıklarının besleme rejimlerine ait veri setleri ile RF, Doğrusal Model

ve Karar Ağacı algoritmaları kullanılarak en uygun yem türüne göre sazan balıklarını sınıflandırmışlardır. RF algoritması en yüksek doğruluk oranını (%98) verirken ardından en iyi sınıflandırmayı Karar Ağacı algoritmasının yaptığı rapor edilmiştir.

Banzato ve ark (2021), yaptıkları çalışmada, köpeklerde göğüs radyografilerinin sınıflandırılması için yapay sinir ağına dayalı yeni bir yöntem geliştirmişlerdir. ResNet-50 ve DenseNet-121 mimarilerine dayalı iki farklı yapay sinir ağı geliştirilmiş ve test edilmiştir. ResNet-50'ye dayalı yapay sinir ağı, tüm radyografik bulgular için %80'in üzerinde ROC eğrisi AUC değerine, DenseNet-121'e dayalı yapay sinir ağının daha düşük bir performans değerine sahip olduğunu bulmuşlardır.

Alves ve ark (2021), yaptıkları çalışmada, Nellore sığırlarında üreme özelliklerinin genomik tahmini için RF, NB ve Destek Vektör Makineleri algoritmalarını kullanarak regresyon modeli oluşturmuşlardır. Modellerin doğrulukları çapraz doğrulama, ortalama tahmin doğruluğu ve ortalama kareler hataları ile değerlendirilmiştir. Destek Vektör Makinelerinin tüm özellikler için diğer modellerden daha iyi sonuçlar verdiğini ortaya koymuşlardır.

Paploski ve ark (2021), yaptıkları çalışmada, SMOTE (Synthetic Minority Oversampling Technique) örnekleme tekniği ile XGBoost makine öğrenimi algoritmasını kullanarak, araştırmaya katılan çiftliklerde haftalık domuz ishali salgını tahminlerini neredeyse gerçek zamanlı olarak gösteren bir model geliştirmişlerdir. 2014 ve 2017 yılları arasında toplanan geriye dönük verilere göre modelin duyarlılık, seçicilik, pozitif ve negatif tahmin değerleri sırasıyla %19,9, %99,9, %70,5 ve %99,4 olarak bulmuşlardır.

Zimpel ve ark (2021), yaptıkları çalışmada, hayvan refahı risklerini tahmin etmek için RF, Destek Vektör Makineleri, NB ve lojistik regresyon makine öğrenmesi sınıflama algoritmalarını kullanmışlardır. Veri seti olarak 10 hayvan sağlığı riski ve 14 emzirme özelliği göstergesi olan 57.000'den fazla domuz kullanmışlardır. Domuzlarla ilgili olan veri setinde 10 değişken ayrı ayrı incelenerek tüm değişkenler için en iyi model seçilmiştir. Sonuç olarak destek vektör makinelerinin çoğu değişkende yüksek doğruluk oranına sahip olduğu tespit edilmiştir.

Tohalino ve Amancio (2022), yaptıkları çalışmada, Tıp, Diş Hekimliği ve Veterinerlik olmak üzere üç ana araştırma alanında makine öğrenimi yoluyla hibe üretkenliğini tahmin etmek için KNN, Destek Vektör Makineleri, NB, Yapay Sinir Ağları ve Karar ağaçları algoritmalarını kullanmışlardır. Sonuçlara göre en yüksek doğruluk oranlarına ulaşan sınıflandırma algoritmaları Destek Vektör Makineleri ve Yapay Sinir Ağları algoritmaları olurken, NB ve KNN en kötü performansı elde ettiği görülmüştür.

De Barros ve ark (2021), yaptıkları çalışmada, karar ağacı algoritmaları kullanılarak atların basitleştirilmiş akut fizyoloji puanı için EqSAPS olarak adlandırılan son puanlama modeli, hastaneye kaldırıldıktan sonraki 24 saat içinde ölüm olasılığı için %92 doğruluk oranı yakalamışlardır. Ölüm sonucu için ROC eğrisi AUC değeri 0.742 iken, hayatta kalma için 0.652 olarak bulmuşlardır.

Tinkov ve ark (2022), yaptıkları çalışmada, model oluşturmak için 1792 toksik maddeden oluşan bir veri seti kullanmışlardır. QSAR (Sayısal Yapı-Etkinlik İlişkisi) modelleri, Dragon, ISIDA, CDK, PyDescriptor, alvaDesc ve SIRMS moleküler tanımlayıcıları ve En Küçük Kareler Destek Vektör Makinesi ve Transformer Evrişimli Sinir Ağı gibi makine öğrenme yöntemleri kullanılarak geliştirmişlerdir. HYBOT tanımlayıcıları ve RF, karşılaştırmalı bir QSAR araştırması için kullanmışlardır. En iyi tahmin yeteneği Transformer Convolutional Neural Network modeli tarafından gösterildiğinden, bireysel avermektinlerin T. pyriformis'e karşı toksisitesini tahmin etmek için kullanmışlardır. Geliştirilen QSAR modelinin yapısal bir yorumu sırasında, organik bileşiklerin akut toksisitesini artıran ve azaltan önemli moleküler dönüşümleri belirlemişlerdir.

Bakoev ve ark (2021), yaptıkları çalışmada, domuz yavrularında fetal malformasyonlarla ilişkili genomik lokusların tanımlanması amacı ile XGBoost makine öğrenmesi algoritmasına dayanan genom çapında bir ilişki çalışması yapmışlardır. Sonuç olarak XGBoost algoritmasının, incelenen malformasyonların ve üretken özelliklerin önemli seçim göstergeleriyle ilişkili Tek Nükleotid Polimorfizmi ve bunların kombinasyonlarını tanımlayabildiğini göstermişlerdir.

Bol lig ve ark (2020), yaptıkları çalışmada, veteriner otopsi raporlarını kullanarak sendroma sürveyansa yürütmek için lojistik regresyon, destek vektör makinesi, karar ağaçları, RF ve XGBoost algoritmalarını kullanmışlardır. Tek bir algoritma üstün olmamasına rağmen, RF algoritmasının F1 puanları 0,923 (gastrointestinal), 0,960 (solunum) ve 0,888 (idrar yolu) ile en iyi performans gösteren algoritma olduğunu ortaya koymuşlardır.

Hyde ve ark (2020), yaptıkları çalışmada, mastitinin bulaşıcı ve çevresel etkilerini, çevresel etkilerin emzirme olmayan ve emzirme dönemlerine göre RF sınıflandırma algoritması kullanılarak sınıflandırma yapmışlardır. Sonuç olarak bulaşıcı ve çevresel etkileri %98, emzirme olmayan ve emzirme dönemlerine göre etkileri %78 doğruluk oranı ile tahmin etmişlerdir.

Cihan ve Coşkun (2021), yaptıkları çalışmada, hastalarda diyabet olasılığını maksimum doğrulukla erken bir aşamada tespit etmek için model tasarlamışlardır. Bunun için lojistik regresyon, KNN, destek vektör makinesi, NB, karar ağacı, RF ve yapay sinir ağı olmak üzere yedi makine öğrenimi sınıflandırma algoritması kullanmışlardır. Elde edilen sonuçlara göre RF algoritması, diyabet hastalığının doğru sınıflandırılmasında diğer yöntemlere göre daha başarılı olduğunu öne sürmüşlerdir.

Ergün ve ark (2021), yaptıkları çalışmada, köpeklere ait görüntülerinden uzun kemiğin çeşidini belirlemişlerdir. Biyomedikal görüntü işleme alanındaki pek çok çalışma gibi, bu çalışmada da Sinir Ağları yöntemleri kullanılmıştır. AlexNet, GoogLeNet ve VGG-19 derin öğrenme modelleri ile öğrenme gerçekleştirilmiş, destek vektör makineleri ile sınıflandırma performansı test edilmiştir. En iyi sınıflandırma başarısının %82 başarı ile AlexNet ile yakalandığı görülmüştür. Buna karşın; GoogLeNet ile VGG-19 yöntemlerine ait başarıların da AlexNet'den çok farklı olmadığını ortaya koymuşlardır.

Metlek ve Kayaalp (2021), arılar üzerinde yaptıkları çalışmada yapay sinir ağı algoritmalarını Softmax sınıflandırıcısı kullanmışlardır. 6 farklı arı hastalığını incelemede Softmax sınıflandırıcısı ile elde edilen doğruluk oranı %93,07 iken iyileştirilen sistem sonucunda bu oranı %95,04 olarak elde etmişlerdir.

Baydan ve Ünver (2021), yaptıkları çalışmada, derin öğrenme yöntemleri kullanarak 1488 kedi ve köpeğin bütün görüntüleri üzerinde kırık durumunu sınıflandırmışlardır. Kırık olarak tespit edilmiş tibialar üzerinde kırığın yerini sınırlandırmak için üç farklı çalışma yapmışlardır. Bu çalışmalarda Yapay Sinir Ağı algoritmaları kullanmışlardır. Çalışmalarda F1 skorları sırası ile birinci çalışmanın ilk fazı için %85, birinci çalışmanın ikinci fazı için %84,5, ikinci çalışma için %68,5 ve üçüncü çalışma için %46,2 olarak bulunmuştur.

Gökçe ve ark (2017), yapay sinir ağlarının çalışmasında normalizasyon yöntemlerinden minimum-maksimum, ondalık ölçeklendirme, z-skor ve sigmoid normalizasyon teknikleri ile normalize yöntemlerini kullanmışlardır. Ham veri ile elde edilen sonuçlarda F1 oranı 0,31 iken sigmoid normalizasyon uygulandıktan sonra elde edilen sonuçlarda F1 skoru 0,36 olarak bulmuşlarıdır.

Cihan ve ark (2020), yaptıkları çalışmada, makine öğrenmesi yöntemleri ile kuzularda erken hastalık teşhisi yapmayı hedeflemişlerdir. Bunun için çalışmada Karar Ağaçları, NB, KNN, Yapay Sinir Ağları ve RF algoritmaları kullanmışlardır. Çalışma sonucunda bilgisayar destekli tanı için NB yönteminin diğer yöntemlerden daha başarılı sonuçlar ürettiğini ortaya koymuşlardır.

Pelit ve ark (2019), Wistar albino türü sıçanlar üzerinde yapıtları çalışmada karar ağaçları algoritmalarından J48 yöntemini kullanmışlardır. 4 farklı grupta incelenen sıçanlar üzerinde diyabetik biyokimyasal değerleri üzerinde yapılan manyetik alan uygulamasıyla sonucunda doğruluk oranının %96'dan %98'e çıktığını bulmuşlardır.

Karaman ve Çiçek (2020), yaptıkları çalışmada, Cryptosporidium spp bağırsak hastalığı durumunu sınıflandırmak için iki farklı yapay sinir ağı kullanmışlardır. Çok katmanlı yapay sinir ağı ile elde edilen doğruluk oranı %75, Radyal tabanlı yapay sinir ağı modeli için ise %71,4 olarak elde etmişlerdir.

1.2 Lojistik Regresyon

Lojistik regresyon, herhangi bir etkiye bağlı sonucun varlığını ya da yokluğunu tahmin etmek için kullanılır. Sonuçların uygun şekilde sınıflandırılabilmesi için matematiksel bir model oluşturur ve bu model yardımı ile sonuçları en doğru şekilde tahmin etmeyi amaçlar (DeMaris 1995).

Regresyon analizinde bağımlı değişkenin iki ya da daha fazla gruptan oluşan bir kategorik değişken olması durumunda lojistik regresyon analizi kullanılmaktadır. Kategorik değişkeninin türüne göre kullanılacak lojistik regresyon yöntemleri değişiklik göstermektedir. Bağımlı değişkenin iki değer aldığı durumlarda ikili (binary) lojistik regresyon; ikiden fazla değer aldığı ve sınıflama ölçme düzeyinde olduğu durumlarda çok düzeyli (multinomial) lojistik regresyon; ikiden fazla değer aldığı ve sıralama ölçme düzeyinde olduğu durumlarda sıralı (ordinal) lojistik regresyon adlarını almaktadır (Alpar 2013; Gamgam ve Altunkaynak 2017; Karagöz 2015).

Bunlara ek olarak olgu-kontrol çalışmalarında kontrol grubu, her bireyin eşleştirilmesiyle oluşturulduysa Koşullu (conditional) Lojistik Regresyon yöntemi kullanılır. Kesin (Exact) Lojistik Yöntemi ise gözlem sayısının az olduğu durumunda kullanılmaktadır (Stephenson 2008; Hosmer ve Lemeshow 2013; Arslanoğlu 2019).

Bu çalışma kapsamında ikili (binary) lojistik regresyon yöntemi ile analizler gerçekleştirilmiştir.

1.2.1 İkili (Binary) Lojistik Regresyon

İkili (binary) lojistik regresyon analizi, iki gruptan oluşan bağımlı değişkenin, bağımsız değişkenlere göre olasılık olarak elde edildiği bir regresyon yöntemidir. Analiz yapılırken ilgilenilen olayın gerçekleşmesi 1, gerçekleşmemesi ise 0 ile gösterilir. Bağımsız değişkenlerde herhangi bir sınırlılık bulunmamaktadır. Yöntemde, bir değer kestirimi yapmaktan ziyade olayın gerçekleşme riski olasılıksal olarak öngörülmeye çalışılır (Alpar 2013; Gamgam ve Altunkaynak 2017; Hosmer ve Lemeshow 2013; Kleinbaum ve Hedeker 1996; Karagöz 2015;)

Lojistik Regresyon Modeli;

$$P(Y=1|X=x_i)=\frac{e^{\left(\beta_0+\sum_{i=1}^{k}\beta_i x_i\right)}}{1+e^{\left(\beta_0+\sum_{i=1}^{k}\beta_i x_i\right)}}=\frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{k}\beta_i x_i\right)}} \quad (1)$$

şeklindedir. Burada, β regresyon katsayısını, k ise bağımsız değişken sayısını ifade etmektedir.

Odds Oranı (Odds Ratio)

Odds oranı, ikili lojistik regresyon yönteminde olayın gerçekleşmesinde etkili her bir bağımsız değişkenin göreceli riskini hesaplamamızı sağlar. Odds, bir olayın gerçekleşme olasılığının, gerçekleşmeme olasılığına oranıdır. Odds oranının 1'den küçük olması bağımsız değişkenin riski azaltıcı etkisi olduğunu, 1'den büyük olması ise bağımsız değişkenin riski arttırıcı etkisinin olduğunu gösterir (Akçay 2009; Kılıç 2015).

Lojit Model

Lojistik regresyon analizinde önemli noktalardan biri modelde meydana gelen iyileşmeleri ortaya koymaktır. Bu iyileşmeyi ortaya koyabilmek için bağımsız değişkenlerin bulunmadığı durumda elde edilen sonuçların performansını değerlendiren başlangıç modeli oluşturulur. Bağımsız değişkenler modele dâhil edildiğinde elde edilen sonuçlar ile başlangıç modelinin sonuçları incelenerek modeldeki iyileşme ya da gelişme ortaya konulur (Alpar 2013).

Odds oranı asimetrik bir yapıya sahiptir. Odds oranının logaritması alınarak,

$$g(x)=ln\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k \quad (2)$$

şeklinde logit model elde edilir ve doğrusal model formuna ulaşılır. Bu durum logit ile bağımsız değişkenler arasındaki ilişkinin doğrusal olması gerekliliği ortaya çıkmaktadır (Alpar 2013).

Logit -∞ ile +∞ arasında değer alabilmekte olup model bağımsız değişkende bir birimlik değişimin logitte ne kadar değişime neden olduğunu gösterir. Lojistik regresyonda artıkların ortalaması 0, varyansı $(\pi * (1 - \pi))$ şeklinde binom dağılıma sahiptir. (Alpar 2013; Boucher ve ark 1998; Akçay 2009).

Lojistik regresyonda model katsayılarının kestiriminde en çok tercih edilen yöntem En Çok Olabilirlik yöntemidir. Bu yöntem dışında yinelemeli olmayan ağırlıklandırılmış en küçük kareler ve diskriminant fonksiyon analizi yöntemleri de kullanılabilmektedir. Örneklem sayısı fazla olduğunda en çok olabilirlik yöntemi ve ağırlıklandırılmış en küçük kareler yöntemlerinin dağılımları benzerlik göstermektedir. Diskriminant fonksiyon yöntemi, normal dağılmayan bağımsız değişkenlerin tahminlerinde, olması gerekenden yüksek sonuçlar verdiği için çok sık tercih edilmemektedir (Alpar 2013; Güç, 2015).

Katsayıların Önemliliğinin Test Edilmesi

Modeldeki bağımsız değişkenlerin, bağımlı değişken üzerindeki etkisinin anlamlılığı Olabilirlik Oranı, Wald ya da Skor (Lagrange multiplier) testleri ile incelenebilir (Alpar 2013; Hosmer ve Lemeshow 2013).

Olabilirlik oranı testi, G istatistiği yardımı ile bağımsız değişkenler modele dâhil edildiğinde performansın artıp artmadığını inceler. Bu değerin incelenebilmesi için bağımsız değişkenlerin bulunmadığı bir model oluşturulur. Modelde sabit dışında herhangi bir bağımsız değişken bulunmamaktadır. Bu modele boş model adı verilir. Sonrasında bağımsız değişkenlerin bulunduğu model oluşturulur. Bu modele de doymuş model adı verilir (Yazar 2018).

Boş modele ait olasılık fonksiyonunun maksimize edilmiş değeri ile doymuş modele ait olasılık fonksiyonunun maksimize edilmiş değerinin oranı $-2ln$ ile

çarpılarak olabilirlik oranı hesaplanır. Daha açık ifade ile G istatistiği, denklem 3 ve 4'te verilmiştir (Alpar 2013; Sharma 1996; Akçay 2009; Çolak 2001).

$$LR = G = -2 \times ln\left(\frac{L(Boş\ model)}{L(\ Doymuş\ model)}\right) \tag{3}$$

$$LR = G = -2 \times lnL(Boş\ model) - lnL(\ Doymuş\ model) \tag{4}$$

Wald istatistiği, her bir değişken için En Çok Olabilirlik yöntemi ile elde edilen Lojit katsayısı anlamlılığını inceler. Wald testi, Lojit katsayısının, standart hatasına bölünmesi ile elde edilir (Alpar 2013).

Wald istatistiği,

$$Wald = \left(\frac{Lojit\ katsayısı_j}{SE(\ Lojit\ katsayısı_j)}\right) \tag{5}$$

şeklindedir.

Skor testi çok değişkenli bir test olup Lojit katsayısı anlamlılığını inceleyen ve test istatistiği standart normal dağılıma uyan bir yöntemdir. Adımsal yöntemlerde modele alınacak ve modelden çıkarılacak bağımsız değişkenlerin tespitinde hızlı ve etkili bir yöntemdir (Alpar 2013; Gamgam ve Altunkaynak 2017; Karagöz 2015; Yazar 2018).

Bağımsız Değişken Seçimi

Analiz esnasında, bağımsız değişken sayısal veya sıralanabilir kategorik değişken ise modele olduğu gibi alınır veya polinom yöntemi ile modele dâhil edilir. İki gruptan oluşan sınıflanabilir kategorik değişken olması durumunda 0 veya 1 olarak kodlanan grup referans olarak seçilip kukla (dummy) değişken olarak modele dâhil edilir. İkiden fazla gruptan oluşan sınıflanabilen kategorik değişkenler de kukla değişken olarak modele dâhil edilir. Değişkenler modele dâhil edilmesi için referans grubun modele uygun şekilde seçilmesi gerekmektedir. Basit yöntem ile ilk veya son grubun kodu referans olarak seçilerek risk hesaplanır. Fark yöntemi ile ilk grup dışında her bir grubun riski kendinden önce bulunan grupların ortalaması referans alınarak risk hesaplanır. Halmert yöntemi ise fark yönteminin tersi olarak risk hesaplar. Repeated

yönteminde, ilk grup dışında her bir grubun riski kendinden önce bulunan grup referans alınarak risk hesaplanır. Sapma yönteminde ise tüm grupların genel ortalaması referans alınarak risk hesaplanır (Hosmer ve Lemeshow 2013).

Çalışmanın amacına göre model belirlenirken regresyon yöntemlerinin doğasına uygun şekilde en az değişken kullanılarak en verimli modeli elde etmek hedeflenir. En uygun modeli oluşturmak için bağımsız değişkenlerin Lojit katsayılarının anlamlılığı incelenir ve bağımlı değişken üzerinde etkili değişkenler belirlenir (Alpar 2013).

Modelde bulunacak bağımsız değişkenler, kategorik ise ki kare testiyle, sayısal ise Mann-Whitney U testi kullanılarak belirlenir. Bu testler sonucunda bağımlı değişken ile ilişkili olan bağımsız değişkenler modele dâhil edilir. Buna ek olarak tek değişkenli lojistik regresyon sonucunda 0,25'ten küçük anlamlılık değerine sahip değişkenler çok değişkenli analizlerde kullanılabilir (Scott 2001). Bunun dışında pratik yöntemler olan ve farklı metotlarla değişken elemesi yapabilen adımsal yöntemler kullanılabilir.

Adımsal Yöntemler

Lojistik regresyon modelinde değişken seçimi için birçok farklı yöntem vardır. Tüm değişkenlerin aynı anda modele dâhil edilerek analiz yapılan yönteme enter yöntemi denir. Değişken seçimleri için ileri yönlü seçim (forward selection) ve geriye doğru seçim (backward elimination) en çok kullanılan yöntemlerdir. İleri yönlü seçim tekniğinde, yeni bir değişken uyumu arttırmayana kadar değişkenler modele eklenir. Her adımda uyumu en çok artıracak değişken seçilir. Geriye doğru seçim tekniğinde ise süreç, en karmaşık değişkenlerin tamamının modele dâhil edilmesiyle başlar ve değişkenler sırayla modelden çıkarılır. Her adımda model üzerinde en az etkiye sahip değişken modelden çıkarılarak en uygun model bulunmaya çalışılır (Lee ve Koval 1997).

İleri yönlü seçim yöntemlerinde değişkenler skor istatistiği kullanılarak modele dâhil edilir. Geriye doğru seçim tekniğinde ise önce tüm değişkenler modele dâhil edilir. İki yöntemde de Wald istatistiği, koşullu parametre tahminlerine dayanan

olabilirlik oranı ve kısmi maksimum olabilirlik tahminlerine dayanan olabilirlik oranı kullanılarak değişkenler modelden çıkartılabilir (Kara 2015).

1.2.2 Lojistik Regresyon Analizinde Model Uyum İyiliği

İstatistiksel yöntemlerde oluşturulan birçok modelin uygunluğunun test edilmesi, değerlendirilmesi ve bu modeller arasından en uygun modelin seçilebilmesi için tüm gözlem değerlerini temsil edecek bir istatistiksel değere ihtiyaç duyulur. Lojistik regresyon modeli için anlamlılık Pearson ki-kare istatistiği ile değerlendirilir. Bu durum Lojistik regresyon analizini makine öğrenmesi yöntemlerinden ayırır. Modelde ki-kare istatistiği için elde edilen anlamlılık değeri 0.05'in altında olması modelin istatistiksel olarak anlamlı olduğunu ifade eder. Modelin uyum iyiliğini incelemek amacıyla doğru sınıflandırma yüzdesi ve Hosmer-Lemeshow ölçütleri kullanılır. Hosmer-Lemeshow istatistiği, olabilirlik oran testidir ve modelde bağımsız değişkenlerin olmadığı $-2\log L_0$ istatistiği ile modelde bağımsız değişkenlerin yer aldığı -2logL istatistiği arasındaki fark alınarak hesaplanır. Hosmer-Lemeshow testi için ki-kare istatistiği kullanılır ve bu istatistik için elde edilen anlamlılık değeri 0.05'in üzerinde olması modelin istatistiksel olarak uyumlu olduğunu ifade eder. Modelin açıklayıcılığını Cox&Snell veya Nagelkerke R^2 değerleri ile değerlendirilir. Bu değerler 1'e ne kadar yakınsa model o kadar iyidir şeklinde yorum yapılır (Hosmer ve ark 1997; Peng ve So 2002).

Modelin uyumunu değerlendirmek için diğer bir yaklaşım da doğru sınıflanma oranlarının incelenmesidir. Bu amaçla, bireylerin gerçekte oldukları sınıflar ile lojistik regresyon modeli yardımıyla tahmin edilen sınıfların çapraz tablosu incelenir. Burada doğru sınıflama oranları (sırası ile duyarlılık, seçicilik, doğruluk yüzdeleri) yorumlanır (Alpar 2013). Doğruluk oranı, bilgiden modelin yüzde kaç doğru sınıflama oranına sahip olduğunu ya da 1-doğruluk oranı ile yanlış sınıflama oranının büyüklüğüne bağlı olarak modelin yeterliliğine karar verilir.

Model oluşturmak için bağımlı değişkenler önsel bilgilerden faydalanılarak modele dâhil edilir. Model çalıştırılarak bağımsız değişkenler için model katsayıları elde edilir. Bağımsız değişkenler için anlamlılık Wald istatistiği kullanılarak değerlendirilir. Katsayıların anlamlılığında Wald istatistiği için elde edilen p değeri

0.05’ten küçükse katsayının model için etkisi anlamlı olarak ifade edilir. Katsayıların anlamlılığı sonucunda risk (odds) oranı ile bağımsız değişkenin sınıflara ayırmadaki etkisi yorumlanır. Modelin daha verimli sonuçlar vermesi için anlamsız değişkenler modelden çıkarılarak modelin tahmin doğruluğu arttırılabilir (Austin ve Steyerberg 2012; Bland ve Altman 2000).

Lojistik regresyon analizi varsayımları aşağıdaki gibidir (Ottenbacher ve ark 2004).

- Uygun tüm bağımsız değişkenler modelde bulunmalı,
- Uygun olmayan tüm bağımsız değişkenler dışlanmalı,
- Bir bireyden birden fazla ölçüm alınmamalı,
- Yüksek hatalar, katsayı tahmininde yanlılığa ve modelin yetersizliğine sebep olduğundan minimum ölçüm hatası olmalı ve eksik veri olmamalı,
- Bağımsız değişkenler arasında yüksek ilişki (çoklu bağlantı) olmamalı,
- Aykırı gözlem değerleri olmamalıdır.

1.3 Makine Öğrenmesi Çalışma Prensibi

Son yıllarda birçok araştırma yöntemine konu olan makine öğrenmesi algoritmaları, yapay zekânın bir dalıdır. Makine öğrenmesi algoritmalarının amacı akıllı sistemler oluşturularak bu sistemleri verimli çalışmasını ve yeni veriler ile geliştirilmesini sağlamaktır. Makine öğrenmesi algoritmaları beş adımdan oluşur (Aksoy ve ark 2022; Lantz 2019).

1. Adım (Verilerin Toplaması): Bu süreçte ilgili konu hakkında gerekli bilgiler elde edilir. Elde edilen bilgilerin boyutu ve kalitesi ne kadar yüksek olursa makine öğrenmesinde öğrenme süreci o kadar iyi olacaktır.

2. Adım (Verilerin Hazırlanması): Gerçek dünya verileri sıklıkla düzensiz, eksik veya gürültülüdür. Bu nedenle, eğitim modelini oluşturmak için sisteme veriler verilmeden önce, ön işleme adımında düzenlenmeli ve temizlenmelidir.

3. Adım (Modelin Eğitilmesi): Bu adım, bir model oluşturmak için uygun algoritmanın seçimini içerir. Burada veriler, tahminleri öğrenmek ve geliştirmek için matematiksel modelleri kullanan seçilmiş algoritma ile çalıştırılır. Makine öğrenimi algoritması denetimli öğrenim ise, veriler eğitim ve test seti olmak üzere iki kısma ayrılır. Sistem eğitim seti ile çalıştırılır ve test aşamasında kullanılmak üzere verilere uygun model oluşturulur.

4. Adım (Modelin Değerlendirilmesi): Bu adımda sistemin performansını belirlemek için bir önceki aşamada oluşturulan model, modelin daha önce görmediği test setine uygulanır. Kullanılan model, performans değerlendirme metrikleri ile analiz edilir. Sonuçlar yeterince iyi değilse, model bir sonraki adımda iyileştirilmeli ve yeniden eğitilmelidir.

5. Adım (Performansın İyileştirilmesi): Önceki bölümde elde edilen sonuçlar yeterince iyi değilse, bu adım modelin tamamen farklı bir modelinin seçilmesini gerektirebilir. 3. adımdan itibaren aşamalar tekrar değerlendirilir.

1.4 Makine Öğrenmesi Bölümleri

Makine öğrenmesi, matematik ve istatistik bilimlerinden yararlanarak elde edilen verilerden sonuç çıkarımı yapan ve tahminlerde bulunan sistemlerin bilgisayarlar ile modellenmesidir. Makine öğrenmesi, öğrenme, öğrenme yöntemleri ve performans metrikleri olmak üzere üç aşamadan incelenir.

1.4.1 Öğrenme

Makine öğrenmesinin öğrenme aşamasında veri seti test, eğitim ve doğrulama olmak üzere üç gruba ayrılmaktadır. Bu ayırma işlemi dışarıda tutma, üçlü ayırma çapraz doğrulama ve yeniden örnekleme yöntemleri ile gerçekleştirilmektedir (Şahinbaş ve Gülseçen 2019).

Dışarıda Tutma (Holdout)

Veri eğitim ve test seti olarak bölünür. Modelin test edilmesinde kullanılan veri modelin eğitilmesi için kullanılan verilerden geri kalanlar olduğu için dışarıda tutma

yöntemi olarak adlandırılmıştır. Düşük örneklemde verilerde veri setinin verimli bölünemeyeceğinden değerlendirmelerde sorun oluşturabilmektedir. Ayrıca verinin tamamı kullanılarak eğitim seti ve test seti olarak bölünmesi yöntemin dezavantajı olarak belirtilebilir (Şahinbaş ve Gülseçen 2019).

Üçlü Ayırma

Veri seti modelin test edilmesi, eğitilmesi ve doğrulanması için üç bölüme ayrılır. Algoritmanın bölme işlemi sırasında verinin kullanım miktarlarının belirlenebilmesi parametre seçiminde önemli bir noktadır. Doğrulama veri seti en verimli sonuçları ortaya koymaya çalışır. Ortaya konulan sonuçlar test veri seti ile değerlendirilir. Bu yöntem model seçimi ve performans değerlendirilmesi yapılabildiği için avantajlı bir yöntemdir (Şahinbaş ve Gülseçen 2019).

Çapraz Doğrulama (Cross Validation)

Çapraz Doğrulama, bir makine öğrenimi sınıflandırma modelinin etkinliğini test etmek için istatistiksel bir yöntemdir. Bu yöntem veri setini k adet alt kümelere bölünür ve bu alt kümelerden biri test seti olarak kullanılır. Kalan k-1 tanesi ise eğitim veri seti olarak kullanılır. Test seti, eğitim seti üzerinde kullanılan modelin performansını değerlendirmek için kullanılmaktadır. Her alt küme için doğrulama seti ile test seti arasındaki hatalar elde edilir. Tespit doğruluğu her test için hesaplanır ve ortalama hata elde edilir (Aksoy ve ark 2022).

Yeniden Örnekleme (Bootstrap)

Yeniden örnekleme yönteminde veri setinden iadeli olarak veri setindeki birim sayısı kadar örneklem seçilerek eğitim seti oluşturulur. Bu seçimden sonra seçilmeyen birimler test veri setine sahil edilir. Yeniden örnekleme yöntemi, verilerin rasgele seçilmesi sebebi ile güçlü bir yöntemdir. (Şahinbaş ve Gülseçen 2019; Sacchi 1998).

1.4.2 Öğrenme Yöntemleri

Makine öğrenmesi yöntemleri temelde hedef (bağımlı) değişkenin değerlerini en iyi şekilde tahmin etmeye çalışır. Tahmini uygun şekilde gerçekleştirmek için

bağımsız değişkenleri ele alan ve analiz eden programlanmış algoritmalar kullanır. Bu algoritmalar yeni veriler eklendikçe, performansı iyileştirmek için öğrenir ve sonuçları optimize ederler. (Springenberg 2015). Hedef değişkeninin türüne göre öğrenme yöntemleri üç bölümde incelenir.

Denetimsiz Öğrenme:

Bağımlı değişkenlerin olduğu fakat hedef değişkenin öğrenme sonucunda oluşturulduğu öğrenme yöntemidir. Denetimsiz öğrenmenin amacı, veriler hakkında daha fazla bilgi edinmek için verinin altında yatan yapıyı veya dağılımı modellemektir. Denetimsiz öğrenme, tüm özelliklerin yapısını inceleyerek birbirine benzer özellikleri sınıflayıp boyut azaltmamızı sağlayabilen bir yöntemdir. Denetimsiz Öğrenme; Kümeleme ve Birliktelik (Association) olmak üzere ikiye ayrılır. Kümeleme problemi, ayrık grupların keşfedildiği yerdir. Birliktelik ise verinin geniş bir bölümünü tanımlayan kuralları keşfetme problemidir (Aksoy ve ark 2022; Springenberg 2015; Şahinbaş ve Gülseçen 2019).

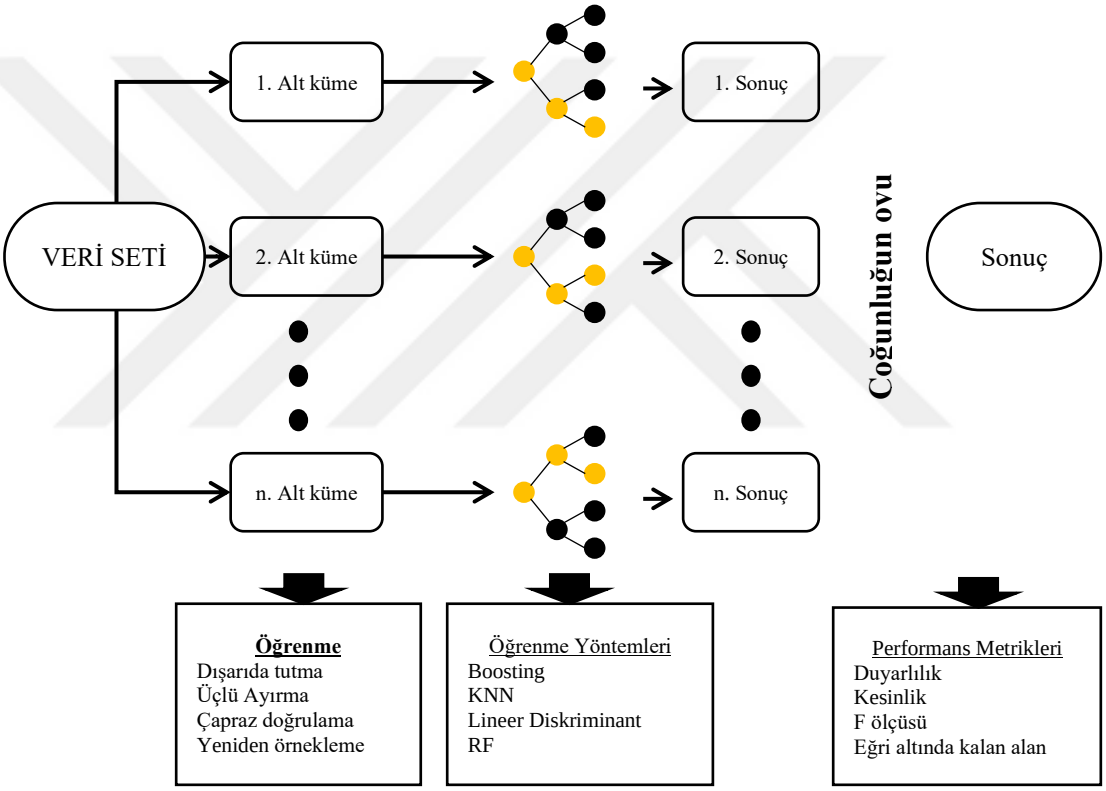
Yarı Denetimli Öğrenme:

Hedef değişkenin sadece bazılarının bilindiği öğrenme yöntemidir. Denetimli ve denetimsiz öğrenme yöntemlerinin her ikisini de kapsar. Bir veriye etiket vermenin ayrı bir maliyeti vardır. Etiketsiz veriler ise bu anlamda daha ucuz, toplanması ve saklanması ise daha kolaydır. Bu açılardan pratik olması nedeniyle çoğu makine öğrenmesi yöntemi bu kapsama girer. Bağımsız değişkenlerin yapısını keşfetmek ve öğrenmek istediğimiz durumlarda denetimsiz öğrenme, hedef değişken hakkında en iyi tahminleri yapmak istediğimiz durumlarda denetimli öğrenme yöntemleri kullanılmalıdır. Yarı denetimli öğrenme algoritmalarında bağımsız değişkenlerin, etiketlenmiş hedef değişkenler ile bir model kurarak, etiketlenmemiş hedef değişkenler için en uygun sonuçları ortaya koyar (Aksoy ve ark 2022; Springenberg 2015; Şahinbaş ve Gülseçen 2019).

Denetimli Öğrenme:

Hedef değişken ve bağımsız değişkenlerin belli olduğu öğrenme yöntemidir. Hedef değişkenin yapısına göre sınıflama ve regresyon olarak iki gruba ayrılabilir. Hedef değişkenin sürekli olması durumunda denetimli öğrenme regresyon algoritmaları kullanılırken kategorik bir yapıya sahip ise sınıflama algoritmaları kullanılır (Aksoy ve ark 2022; Springenberg 2015; Şahinbaş ve Gülseçen 2019).

Çalışmanın bundan sonraki bölümlerinde tezin amacına uygun olarak makine öğrenmesi, Boosting, KNN, lineer diskriminant ve RF sınıflama algoritmalarından bahsedilecektir. Bu yöntemlerin çalışma prensipleri Şekil 1.3'te verilmiştir.



Şekil 1.3. Sınıflama algoritmaları (Ferreira ve Figueiredo 2012; Freund 1995)

Boosting sınıflama algoritması

Boosting algoritması, Freund (1995) tarafından geliştirilen bir makine öğrenme algoritmasıdır. Boosting, çok sayıda zayıf öğreniciyi bir araya getirerek güçlü öğrenici oluşturan yöntemidir. Boosting algoritması, temelde karar ağaçlarına dayanan bir yöntemdir (Şekil 1.3).

Boosting yöntemindeki temel fikir, veri setine farklı ağırlıklar verilmesi sonucu elde edilen ağaçlar topluluğundan çıkarsamalar yapılmasıdır. Başlangıçta tüm gözlemler eşit ağırlığa sahiptir. Ağaçlar büyümeye başladıkça, oluşan modele bağlı olarak ağırlıklar düzenlenir. Yanlış sınıflandırılan gözlemlerin ağırlığı arttırılırken, yanlış sınıflandırılan gözlemlerin ağırlığı azaltılır. Bu sayede ağaçlar zor durumlar karşısında öğrenme yeteneği kazanır. Boosting algoritması şu şekilde çalışır (Freund 1995).

- Orijinal veri kümesinden bir alt küme oluşturulur.
- Başlangıçta, tüm veri noktalarına eşit ağırlık verilir.
- Bu alt kümede bir temel model oluşturulur.
- Bu model, tüm veri kümesi için tahminler yapmak için kullanılır.
- Hatalar gerçek değerler ve öngörülen değerler kullanılarak hesaplanır.
- Yanlış öngörülen gözlemlere daha fazla ağırlık verilir.
- Başka bir model oluşturulur ve veri setinde tahminler yapılır.
- Benzer şekilde, her biri önceki modelin hatalarını düzelten çoklu modeller oluşturulur.
- Son model (güçlü öğrenen), tüm modellerin ağırlıklı ortalamasıdır.

K en yakın komşu sınıflama algoritması

KNN algoritması, Cover ve Hart (1967) tarafından geliştirilen bir makine öğrenme algoritmasıdır. Boosting, çok sayıda zayıf öğreniciyi bir araya getirerek güçlü öğrenici oluşturan yöntem olup bu yöntem temelde karar ağaçlarına dayanmaktadır.

KNN sınıflama algoritması, uygulaması kolay gözetimli öğrenme algoritmalarındandır. Hem sınıflandırma hem de regresyon problemlerinin çözümünde kullanılıyor olmakla birlikte, endüstride çoğunlukla sınıflandırma problemlerinin çözümünde kullanılmaktadır. KNN parametrik olmayan bir algoritmadır, çünkü ilişkinin fonksiyonel formu hakkında açık bir varsayım yoktur. Ayrıca, belirli bir veri modelini uygulamaktan ziyade doğrudan eğitim örneklerine odaklandığı anlamda örnek tabanlı öğrenmenin bir örneğidir. Algoritma, sınıfları belli olan bir örnek

kümesindeki verilerden yararlanılarak kullanılmaktadır. Örnek veri setine katılacak olan yeni verinin, mevcut verilere göre uzaklığı hesaplanıp, k sayıda yakın komşuluğuna bakılır. Uzaklık hesapları için Öklid, Manhattan ve Minkowski uzaklık fonksiyonları kullanılmakta olup formülleri denklem 6, 7 ve 8'de verilmiştir (Rodrigues 2018; Taşcı ve Onan 2016).

$$\text{Öklid uzaklığı} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)} \qquad i \neq j: 1, 2, \dots, n \tag{6}$$

$$Manhattan\ uzaklığı = \sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j| \qquad i \neq j: 1, 2, \dots, n \tag{7}$$

$$Minkowski\ uzaklığı = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^q\right)^{\frac{1}{q}} \qquad i \neq j: 1, 2, \dots, n \tag{8}$$

KNN; eski, basit ve gürültülü eğitim verilerine karşı dirençli olması sebebiyle en popüler makine öğrenme algoritmalarından biridir. Fakat bunun yanında dezavantajı da mevcuttur. Örneğin, uzaklık hesabı yaparken bütün durumları sakladığından, büyük veriler için kullanıldığında çok sayıda bellek alanına gereksinim duymaktadır. KNN algoritması şu şekilde çalışır (Rodrigues 2018; Taşcı ve Onan 2016).

- İlk olarak k parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır. Örneğin: k=2 olsun. Bu durumda en yakın 2 komşuya göre sınıflandırma yapılacaktır.
- İlgili uzaklık fonksiyonları yardımı ile örnek veri setine katılacak olan yeni verinin, mevcut verilere göre uzaklığı tek tek hesaplanır.

- İlgili uzaklıklardan en yakın k komşu ele alınır. Öznitelik değerlerine göre k komşu veya komşuların sınıfına atanır.
- Seçilen sınıf, tahmin edilmesi beklenen gözlem değerinin sınıfı olarak kabul edilir. Yani yeni veri etiketlenmiş olur.

KNN, değerleri için bir tahmin yapmak için yeni gözlemlere en çok benzeyen k tahmini gözlem sayısına bakan bir regresyon yöntemidir. Küçük sayılar modelin esnekliğini artırdığından, en yakın komşuların sayısı içsel olarak model karmaşıklığıyla bağlantılıdır (Rodrigues 2018; Taşcı ve Onan 2016).

Lineer Diskriminant Sınıflama Algoritması

Lineer Discriminant, Fischer (1936) tarafından ortaya atılan bir sınıflama metodudur. Rao (1948) tarafından geliştirilerek ikiden fazla grubun doğru sınıflandırılabilme temeline dayanan bir yöntem haine getirilmiştir. Basit bir yapıya sahip bir yöntem olmasına karşın sınıflandırma problemleri için yüksek performanslı sonuçlar ortaya koymaktadır.

Lineer Diskriminant Sınıflama Algoritması, veri içerisinde bulunan farklı sınıflara ait grupların doğrusal sınıflana bilirliğini maksimize ederek boyut azaltması yapan bir yöntemdir. Her grup içerisindeki varyansı minimum ve grupların ortalamalarını birbirlerinden maksimum düzeyde uzak tutar (Jingve ark 2017; Martinez ve Kak 2001; Wilson ve ark 2018).

Lineer Diskriminant Analizi (LDA), bağımlı değişkendeki sınıfları en iyi şekilde ayırt eden bileşenlerini bulmayı amaçlayan bir sınıflandırma yöntemidir. LDA doğrusal bir sınıflandırıcıdır ve bazı varsayımları vardır. Bu varsayımlar aşağıdaki gibidir.

- Hedef değişken kategorik veya sıralı kategorik bir değişkendir.
- Tahmin edici değişkenler sürekli, kategorik veya sıralı kategorik değişkenlerden oluşur.

- Sınıfın eşitliği şu anlama gelir: Sınıf ortalamaları eşit olmalıdır, karşılık gelen tablo ile kontrol edilebilir.
- Varyans-Kovaryans matrislerinin homojen olması gerekir, ilgili tablodan kontrol edilebilir.
- Bağımsız değişkenler arasında yüksek bağlantı olmamalıdır.

Lineer Diskriminant sınıflama algoritması bazı varsayımları sağlaması açısından Lineer Regresyon analizi ile benzerlik göstermektedir (Balakrishnama ve Ganapathiraju 1998; Çağlıyan ve Köse 2021; James, Witten ve ark 2013; Lee ve Ou 2007).

RF Sınıflama Algoritması

RF, hiperparametre kestirimi yapılmadan da iyi sonuçlar vermesi hem regresyon hem de sınıflandırma problemlerine uygulanabilir olmasından dolayı popüler makine öğrenmesi modellerinden biridir. RF regresyon birden fazla karar ağacını kullanarak daha uyumlu modeller üreterek isabetli tahminlerde bulunmaya yarayan bir regresyon modelidir. Karar ağaçlarını kullandığı için kesiklidir. Yani belli bir aralıkta istenen tahminler için aynı sonuçları üretir. RF algoritması şu şekilde çalışır.

- Toplam "M" özellikten rastgele olarak "K" özellik seçilir (K<M olmalı).
- "K" özellik arasından en iyi ayırma noktasını kullanarak "d" düğümü hesaplanır.
- En iyi ayırmayı kullanarak düğümler alt düğümlere ayrılır.
- "L" düğüm sayısına ulaşılana kadar 1. ve 3. adımları tekrarlanır.
- "n" adet sayı ağacı oluşturmak için "n" kere 1. ve 4. adımları tekrarlanır.

RF, bir topluluk olarak çalışan çok sayıda bağımsız ağaçtan oluşan bir dizi karar ağacı oluşturan bir regresyon yöntemidir (Breiman 2001; James ve ark 2013; Pal 2005).

1.4.3 Performans Metrikleri

Model sınıflaması için performansı değerlendiren ve başarıyı ölçen birtakım ölçütler vardır. Bu ölçütlerin hesaplanmasında hata matrisi kullanılır. Hata matrisi, tahminlerin doğruluğu hakkında bilgi veren 2x2'lik bir matristir. Model performansını ölçmek için tahmin değerleri ile gerçek değerleri karşılaştıracak bir matris oluşturulur.

Hata matrisi, iki veya çok kategorili sınıflandırma problemi için makine öğrenimi modelini tahmin etmek ve görselleştirmek için kullanılan bir tablo düzenidir. Bu tablo düzeni Tablo 1.1'de verilmiştir.

Tablo 1.1. Hata Matrisi (Swets 1988; Konur 2020)

		Gerçek Sınıf	
		Pozitif	**Negatif**
Tahmin Edilen Sınıf	**Pozitif**	*DP*	*YP*
	Negatif	*YN*	*DN*

Hata matrisinde DP, pozitif olarak karar verilenlerin gerçekte de pozitif olması, YP pozitif olarak karar verilenlerin gerçekte negatif olması, YN negatif olarak karar verilenlerin gerçekte pozitif olması ve DN negatif olarak karar verilenlerin gerçekte negatif olmasıdır. Hata matrisinde elde edilen sonuçlar yardımı ile sınıflandırma algoritması için performans metrikleri elde edilir. Hesaplanan doğruluk, duyarlılık, seçicilik, kesinlik ve F ölçüsü denklem 9, 10, 11, 12 ve 13'te görüldüğü gibi hesaplanır.

$$Doğruluk\ (accuracy) = \frac{DP + DN}{DP + DN + YP + YN} \quad (9)$$

$$Duyarlılık\ (Sensitivity) = \frac{DP}{DP + YN} \quad (10)$$

$$Seçicilik\ (Specificity) = \frac{DN}{DN + YN} \quad (11)$$

$$Kesinlik\ (Precesion) = \frac{DP}{DP + DN} \quad (12)$$

$$F\ \text{Ölçüsü} = 2 \times \frac{Duyarlılık \times Kesinlik}{Duyarlılık + Kesinlik} \quad (13)$$

Bir sınıflandırma performansının sadece hata matrisi ölçüleri ile ifade edilmesi veya karşılaştırılması uygun ve anlaşılır değildir. Bu nedenle, sınıflandırma performanslarını değerlendirmek için çeşitli ölçütler önerilmiştir. ROC analizinin kökeni, seçim yapılması gereken bir durumda kararların alınma biçimini modellemek için teorik çerçeve olan ve alınan kararların stokastik veya olasılıksal temsiline dayanan sinyal algılama teorisine dayanmaktadır. ROC eğrisi, alınan kararların doğruluğu veya yanlışlığını modellemek ve görselleştirmek için oluşturulan bir grafiktir (Lusted 1984).

ROC analizi, modelden elde edilen tahmini sonuçlar ile gerçek sonuçlar arasındaki duyarlılık ve özgürlük arasındaki ilişkiye bağlıdır. ROC analizi grafiğinin oluşturulacağı koordinat sisteminde, Y ekseninde tanı testinin DP oranı (duyarlılık), X ekseninde ise yanlış pozitif oranı (1- Seçicilik) yer alır. Her kesim noktasındaki DP oranı ve YP oranına karşılık gelen noktalar birleştirilerek ROC eğrisi çizilir. ROC analizi sonucunda oluşturulan grafik sonucunda AUC değeri hesaplanır. AUC değeri 1'e ne kadar yakınsa modelin performansı o kadar iyidir. AUC en düşük 0,5 değeni alabilmektedir (Beşer ve ark 2021; Canbek ve ark 2017; Fan ve ark 2006; Narkhede 2018).

2.GEREÇ VE YÖNTEM

2.1 Gereç

Bu çalışmadaki veriler, Çanakkale ili Bayramiç ilçesinde bulunan özel Alper Parer Veteriner Kliniği yardımıyla elde edilmiştir. 220 buzağıya ait vücut ısısı, hidrojen iyon konsantrasyonu (pH), kısmi karbondioksit basıncı (pCO_2-mmHg), kısmi oksijen basıncı (pO_2-mmHg), sodyum (Na+-mmol/L), potasyum (K+-mmol/L), Glikoz (mg/dL) ve Laktat (mmol/L), akyuvar (WBC-103/µL), Lenfosit (LYM-109/L), alyuvar (RBC-106/µL), hemotokrit (HCT-%) ve hemoglobin (HGB-mg/dL) ölçümler kullanılmıştır. Analizler Lojistik Regresyon, Boosting, KNN, Lineer Diskriminant ve RF makine öğrenmesi algoritmaları kullanılarak gerçekleştirilmiştir. Buzağılara ait klinik bulgular, kan gazı ölçümleri, hematolojik ve biyokimyasal ölçümler bağımsız değişkenler olarak kabul edilmiştir.

Buzağı ishalini öngörmede klinik bulgular, kan gazı ölçümleri, hematolojik ve biyokimyasal ölçümlerin etkisini incelemede Lojistik regresyon analizinde enter yöntemi ile tüm değişkenlerin bulunduğu model değerlendirilmiştir. Ardından stepwise yöntemi sonucunda elde edilen en iyi model için tüm yöntemlerin hata matrisi ve performans metrikleri değerlendirilerek yöntemlerin performansları değerlendirilmiştir.

Tez projesi için, Selçuk Üniversitesi Veteriner Fakültesi Deney Hayvanları Üretim ve Araştırma Merkezi Etik Kurulu'ndan (SÜVDAMEK), 12.01.2022 gün ve 2/11 sayılı karar ile izin alınmıştır. Etik Kurul onayına Ek-A'da yer verilmiştir.

2.2 Yöntem

Çalışmada, ishal bulunan 69 baş, bulunmayan 151 baş olmak üzere toplam 220 buzağıya ait klinik bulgular, kan gazı ölçümleri, hematolojik ve biyokimyasal veriler kullanılmıştır. Klinik bulgular, kan gazı ölçümleri, hematolojik ve biyokimyasal değişkenler kullanılarak Lojistik Regresyon, Boosting, KNN, Lineer Diskriminant ve RF makine öğrenmesi algoritmaları ile buzağı ishalini en doğru şekilde tahmin etmek amaçlanmıştır. Hata matrislerine göre tüm yöntemler için doğruluk oranları belirlenmiştir. Analizler yapılırken ishal bulunan grup 1, bulunmayan grup ise 0 olarak

belirlenmiştir. Bağımsız değişkenler sayısal veriler olduğu için herhangi bir kısıtlama uygulanmamıştır.

Uygulanan analizde yapılan tahminlerin uygunluğu değerlendirmek ve etkili, uzak ve aykırı gözlem bulunup bulunmadığını değerlendirmek için Standardize edilmiş hatalara ilişkin Normal ve Detrended Q-Q plot grafikleri ile aykırı gözlemler incelendi. Cook Uzaklığı değerlerine ilişkin saçılım grafiği ile etkili gözlemler incelendi (Çokluk 2010).

Bağımsız değişkenler arasında 0,80 üzerinde ilişkilerin bulunması çoklu bağlantı sorunu olmasına sebep olmaktadır. Bağımsız değişkenler arasında çoklu bağlantı sorunu Spearman korelasyon katsayıları incelenerek değerlendirildi. Modelin açıklayıcılık oranlarını incelemede Cox-Snell ve Nagelkerke R^2 istatistikleri incelendi. R^2 istatistiklerinin 0,20 üzerinde olması yeterli kabul edildi (Kalaycı 2005). Model uyum iyiliklerini değerlendirmek için Hosmer-Lemeshow testi ve Pearson Ki-kare istatistiği incelendi.

Lojistik Regresyon yönteminde bağımsız değişkenlerin katsayıları incelenmiş ve en uygun model Stepwise yöntemi ile oluşturulmuştur. Bağımsız değişkenlerdeki azalmanın, yöntemlerin performansını nasıl etkilediği incelenmiştir. Lojistik regresyon incelemesinde anlamlılık değeri dikkate alınmıştır. Burada anlamlılık değeri 0.05'ten küçük olduğu durumda modelin anlamlı olduğu kabul edilmiştir.

Analizler JASP 0.16.4.0 istatistik paket programında yapılmıştır. Makine öğrenmesi algoritmaları için validasyon doğruluğu ve test doğruluğu değerleri incelenmiştir. Çalışmada 220 baş buzağıya ait veri seti kullanılmıştır. Veri setinin 176 (%80) baş buzağıdan 140'ı (%80) eğitim 36'sı (%20) validasyon için kullanılmış olup 44 (%20) baş buzağı ise test için kullanıldı.

3.BULGULAR

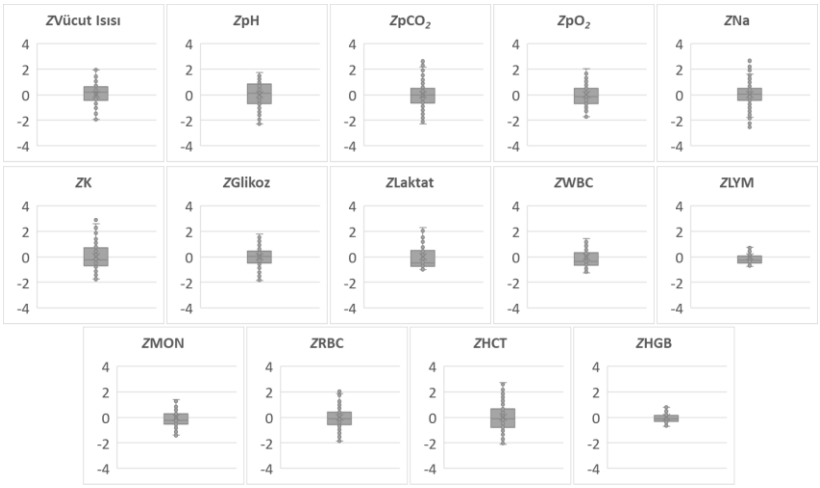
3.1. Değişkenlerin Keşifçi Veri Analizi

Çalışmada toplam 220 baş buzağı ya ait klinik bulgular, kan gazı ölçümleri, hematolojik ve biyokimyasal değişkenler bulunmaktadır. Bu değişkenler çalışmada bağımsız değişkenler olarak değerlendirilmiştir. Tüm bağımsız değişkenler sayısal yapıya sahiptir. 220 baş buzağı ya ait verilerde herhangi bir eksik gözlem bulunmamaktaydı. İncelenen değişkenlere ait betimleyici istatistikler Tablo 3.1'de verilmiştir.

Tablo 3.1. Bağımsız Değişkenlere Ait Betimsel İstatistikler

Değişken	Aritmetik Ortalama	Standart Sapma	Medyan	Minimum	Maximum	25. yüzdelik	75. yüzdelik
Vücut ısısı	37,49	1,13	37,70	34,00	39,70	37,00	38,20
pH	7,20	0,17	7,22	6,80	7,51	7,08	7,35
pCO_2	45,18	10,09	45,00	22,00	73,70	38,60	50,08
pO_2	26,55	8,79	25,10	11,50	66,00	20,30	30,88
Na	138,80	10,61	139,00	112,00	167,00	134,00	144,00
K	5,21	1,27	4,90	3,00	8,90	4,30	6,10
Glikoz	90,77	30,23	92,00	14,00	222,00	76,00	104,00
Laktat	3,68	3,53	2,00	0,10	15,00	1,10	5,50
WBC	16,79	10,58	13,12	3,64	73,54	9,92	20,10
LYM	9,06	12,11	5,73	0,26	76,88	3,24	9,78
MON	0,68	0,45	0,58	0,05	4,74	0,44	0,80
RBC	10,16	2,75	9,92	5,00	32,00	8,54	11,30
HCT	36,82	10,25	35,90	15,00	64,60	28,78	43,50
HGB	12,97	10,51	11,75	5,70	158,20	9,50	14,60

Bağımsız değişkenlere ait değerlerde etkili, uzak ve aykırı gözlem bulunup bulunmadığı kutu grafiği yardımı ile incelenmiştir. İncelemede vücut ısısı, pH, pCO_2, pO_2, Na, K, Glikoz, Laktat, WBC, MON, LYM, RBC, HCT ve HGB değişkenleri için standartlaştırılmış değerler kullanılmış olup aykırı gözlem olmadığı görülmüştür. Sonuçlar Şekil 3.1'de verilmiştir.



Şekil 3.1. Bağımsız Değişkenlerde Aykırı Gözlem İncelemesi

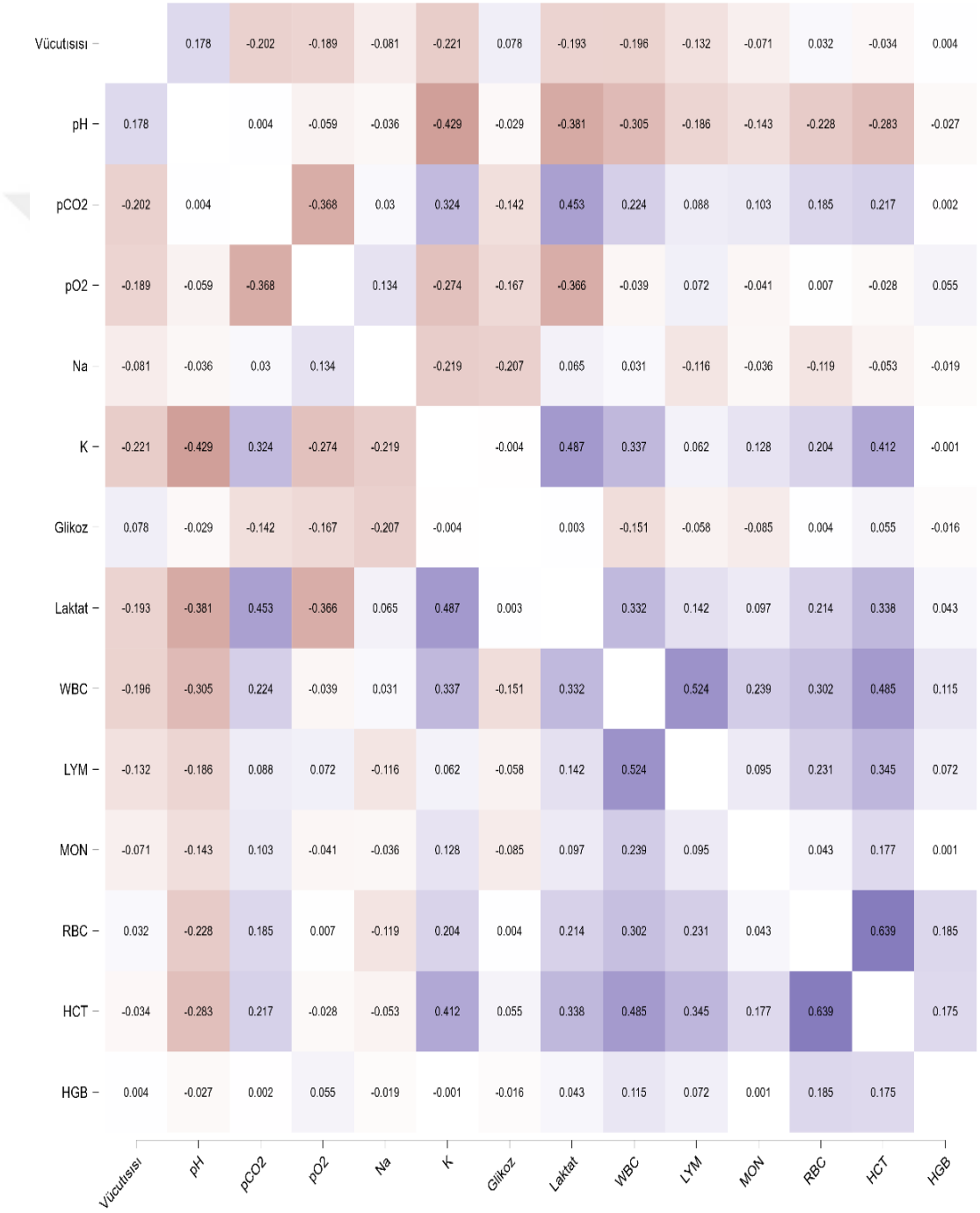
Buzağılardan elde edilen vücut ısısı, pH, pCO_2, pO_2, Na, K, Glikoz, Laktat, WBC, MON, LYM, RBC, HCT ve HGB ölçümlerinin normal dağılıma uygun olmadığı görüldü. Normallik inlemesi sonuçları Tablo 3.2'de verilmiştir.

Tablo 3.2. Bağımsız Değişkenlere Ait Normallik İncelemesi

Değişken	Çarpıklık	Çarpıklık Standart Hatası	Basıklık	Basıklık Standart Hatası	Shapiro Wilk	p değeri
Vücut ısısı	-0,97	0,16	1,10	0,33	0,940	<0,001
pH	-0,43	0,16	-0,62	0,33	0,961	<0,001
pCO_2	0,26	0,16	0,22	0,33	0,984	0,015
pO_2	1,03	0,16	1,68	0,33	0,942	<0,001
Na	0,04	0,16	0,94	0,33	0,957	<0,001
K	0,77	0,16	0,13	0,33	0,944	<0,001
Glikoz	0,64	0,16	2,51	0,33	0,955	<0,001
Laktat	1,31	0,16	1,06	0,33	0,828	<0,001
WBC	2,35	0,16	8,00	0,33	0,779	<0,001
LYM	3,74	0,16	15,26	0,33	0,536	<0,001
MON	4,35	0,16	32,14	0,33	0,679	<0,001
RBC	2,42	0,16	17,07	0,33	0,859	<0,001
HCT	0,34	0,16	-0,16	0,33	0,987	0,042
HGB	12,19	0,16	168,40	0,33	0,272	<0,001

Yapılan Spearman korelasyon analizi sonucunda vücut ısısı ile pH arasında pozitif yönlü, pO_2, K, WBC, LYM ve MON arasında negatif yönlü istatistiksel olarak anlamlı ilişki vardı. pH ile K, Laktat, WBC, RBC, HCT ve HGB arasında negatif

istatistiksel olarak anlamlı ilişki vardı. pCO_2 ile pO_2 arasında negatif yönlü, Na, K, Laktat, RBC, HCT ve HGB arasında pozitif yönlü istatistiksel olarak anlamlı ilişki vardı. pO_2 ile K, Glikoz ve Laktat arasında negatif yönlü istatistiksel olarak anlamlı ilişki vardı. Na ile K ve Glikoz arasında negatif yönlü istatistiksel olarak anlamlı ilişki vardı. Değişkenler arasındaki Spearman korelasyon katsayılarının bulunduğu Heatmap grafiği Şekil 3.2'de verilmiştir.



Şekil 3.2. Bağımsız Değişkenler Arasındaki İlişkiler için Heatmap Grafiği

Şekil 3.2'e göre bağımsız değişkenler arasında anlamlı ilişkiler bulunmada bu ilişkiler 0,80'in altındadır. Bu durum sonucunda bağımlı değişkenler arasında çoklu bağlantı sorunu olmadığı tespit edilmiştir.

3.2. Lojistik Regresyon Sonuçları

Lojistik Regresyon için oluşturulan modelde vücut ısısı, pH, pCO_2, pO_2, Na, K, Glikoz, Laktat, WBC, MON, LYM, RBC, HCT ve HGB değişkenlerinin buzağı ishalini öngörmede etkisi istatistiksel olarak anlamlıydı. Lojistik regresyon enter modeli kullanılarak tüm bağımsız değişkenler modele dâhil edildi. Buna göre amaçlanan model ile sabit model arasındaki fark istatistiksel olarak önemlidir. (χ^2=150,806; p<0,001). R^2 bağımlı değişkedeki değişimin bağımsız değişkenler tarafından açıklanma oranını gösterir. Buna göre, Cox - Snell R^2 değerine göre %49,6 oranında, Nagelkerke R^2 değerine göre %69,7 oranında bağımsız değişkenler bağımlı değişkenleri açıklamaktadır. Hosmer-Lemeshow testi ile modelin uyum iyiliği, bağımlı değişkeni açıklamada oluşturulan en iyi modelin verilere uygun olduğu görülmektedir (χ^2=1,406; p>0,05). Sonuçlar Tablo 3.3'te verilmiştir.

Tablo 3.3. Lojistik Regresyon Enter Yöntemi Sonuçları

Değişken	*β*	*Standart hata*	*Odds oranı*	*Wald*	*p*	*Odds oranı* için %95 GA	
						Alt sınır	*Üst sınır*
Sabit	-16,768	18,450	0,000	0,826	0,363	-52,930	19,393
Vücut ısısı	0,131	0,255	1,140	0,265	0,607	-0,369	0,631
pH	0,231	1,765	1,260	0,017	0,896	-3,228	3,690
pCO_2	0,012	0,031	1,012	0,153	0,696	-0,049	0,073
pO_2	0,113	0,038	1,120	**8,729**	**0,003**	0,038	0,189
Na	0,054	0,029	1,055	3,359	0,067	-0,004	0,112
K	0,361	0,284	1,434	1,613	0,204	-0,196	0,917
Glikoz	-0,014	0,008	0,986	3,303	0,069	-0,029	0,001
Laktat	-0,123	0,093	0,885	1,759	0,185	-0,304	0,059
WBC	-0,207	0,064	0,813	**10,430**	**0,001**	-0,333	-0,081
LYM	-0,088	0,074	0,916	1,418	0,234	-0,233	0,057
MON	-1,231	0,884	0,292	1,938	0,164	-2,963	0,502
RBC	1,264	0,293	3,538	**18,570**	**<0,001**	0,689	1,838
HCT	-0,361	0,078	0,697	**21,530**	**<0,001**	-0,513	-0,208
HGB	0,184	0,091	1,202	**4,063**	**0,044**	0,005	0,363

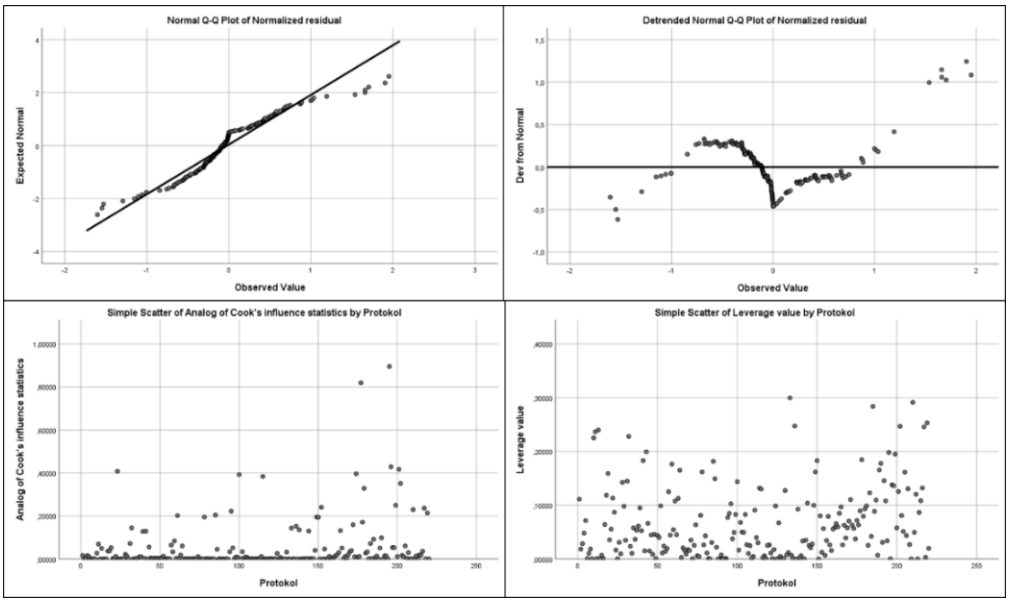
Tablo 3.3'de verilen bulgularda referans kategorimiz buzağı ishali olma durumudur. pO_2 ölçümünde elde edilen model katsayısı 0,113 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(0,113)=1,120 olup pO_2 değişkeninde bir birimlik artış buzağı ishali olma riskini 1,120 kat arttırmaktadır (*Wald* istatistiği=8,729; p=0,003).

WBC ölçümünde elde edilen model katsayısı -0,207 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(-0,207)=0,813 olup WBC değişkeninde bir birimlik artış buzağı ishalinin görülme riskini (1-0,813=0,187) yaklaşık %19 azaltmaktadır (*Wald* istatistiği=10,430; p=0,001).

RBC ölçümünde elde edilen model katsayısı 1,264 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(1,264)=3,538 olup RBC değişkeninde bir birimlik artış buzağı ishalinin görülme riskini 3,538 kat arttırmaktadır (*Wald* istatistiği=18,570; p<0,001).

HCT ölçümünde elde edilen model katsayısı -0,361 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(-0,361)= 0,697 olup HCT değişkeninde bir birimlik artış buzağı ishalinin görülme riskini (1-0,697=0,303) yaklaşık %30 azaltmaktadır (*Wald* istatistiği=21,530; p<0,001). HGB ölçümünde elde edilen model katsayısı 0,184 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(0,184)=1,202 olup HGB değişkeninde bir birimlik artış buzağı ishalinin görülme riskini 1,202 kat arttırmaktadır (*Wald* istatistiği=4,063; p=0,044).

Artıklar incelenerek etkili, uzak ve aykırı gözlemlerin olup olmadığı tespit edildi. Artıklar standardize edilmiş hata (standardized residuals) değerleriyle, etkili gözlemler ise Cook Uzaklığı yardımı ile incelendi. Standardize edilmiş artık değerlerinin -3 ile +3 arasında değer aldığı görüldü ve aykırı gözlem olmadığı tespit edildi. Cook Uzaklığı için elde edilen değerlerin ise 1'den küçük olduğu bulundu ve etkili gözlem olmadığı görüldü.



Şekil 3.4. Model Artıklarına İlişkin Bulgular

Modelde stepwise yöntemi ile en uygun değişkenler adımsal olarak modele dâhil edildi. En uygun model için WBC değişkeni seçildi ve ilk model ile sabit model arasındaki fark istatistiksel olarak önemlidir (χ^2=65,597; p<0,001). İkinci olarak pO_2 değişkeni dâhil edildi ve ilk model ile sabit model arasındaki fark istatistiksel olarak önemlidir (χ^2=16,921; p<0,001). Üçüncü olarak LYM değişkeni dâhil edildi ve ilk model ile sabit model arasındaki fark istatistiksel olarak önemlidir (χ^2=12,577; p<0,001). Dördüncü olarak Glikoz değişkeni dâhil edildi ve ilk model ile sabit model arasındaki fark istatistiksel olarak önemlidir (χ^2=8,865; p=0,003). Beşinci olarak HGB değişkeni dâhil edildi ve ilk model ile sabit model arasındaki fark istatistiksel olarak önemlidir (χ^2=5,369; p=0,021). Altıncı olarak HCT değişkeni dâhil edildi ve ilk model ile sabit model arasındaki fark istatistiksel olarak önemlidir (χ^2=7,878; p=0,005). Son olarak RBC değişkeni dâhil edildi ve ilk model ile sabit model arasındaki fark istatistiksel olarak önemlidir (χ^2=25,285; p<0,001).

Stepwise yöntemi ile kurulan Lojistik Regresyon modelinde WBC, pO2, LYM, Glikoz, HGB, HCT ve RBC değişkenlerinin bağımlı değişken üzerindeki etkisi istatistiksel olarak anlamlıdır (χ^2=25,285; p<0,001). Cox - Snell R^2 değerine göre %47,7 oranında Nagelkerke R^2 değerine göre %67 oranında bağımsız değişkenler bağımlı değişkenleri açıklanmaktadır. Hosmer-Lemeshow testi ile modelin uyum iyiliği, bağımlı değişkeni açıklamada oluşturulan en iyi modelin verilere uygun olduğu görülmektedir (χ^2=1,248; p>0,05). Sonuçlar Tablo 3.4'te verilmiştir.

Tablo 3.4. Lojistik Regresyon Stepwise Yöntemi Sonuçları

Model	**Değişken**	*β*	*Standart hata*	*Odds oranı Exp (β)*	*Wald*	*p*	*Odds oranı* için %95 GA	
							Alt sınır	*Üst sınır*
1	Sabit	-0,783	0,145	0,457	**29,048**	**<0,001**	-1,068	-0,498
2	Sabit	2,539	0,556	12,666	**20,853**	**<0,001**	1,449	3,629
	WBC	-0,249	0,045	0,780	**29,982**	**<0,001**	-0,338	-0,160
3	Sabit	0,455	0,756	1,577	0,363	0,547	-1,026	1,937
	WBC	-0,263	0,048	0,768	**30,582**	**<0,001**	-0,357	-0,170
	pO2	0,083	0,022	1,087	**14,106**	**<0,001**	0,040	0,127
4	Sabit	1,081	0,836	2,947	1,671	0,196	-0,558	2,720
	WBC	-0,245	0,049	0,783	**24,881**	**<0,001**	-0,342	-0,149
	pO2	0,087	0,023	1,091	**14,035**	**<0,001**	0,041	0,132
	LYM	-0,165	0,056	0,848	**8,545**	**0,003**	-0,276	-0,054
5	Sabit	3,178	1,154	23,996	**7,586**	**0,006**	0,917	5,439
	WBC	-0,266	0,053	0,766	**25,522**	**<0,001**	-0,369	-0,163
	pO2	0,080	0,025	1,083	**10,446**	**0,001**	0,031	0,128
	LYM	-0,160	0,058	0,852	**7,763**	**0,005**	-0,273	-0,048
	Glikoz	-0,018	0,007	0,982	**7,881**	**0,005**	-0,031	-0,006
6	Sabit	2,140	1,283	8,500	2,780	0,095	-0,375	4,656
	WBC	-0,289	0,056	0,749	**27,109**	**<0,001**	-0,398	-0,180
	pO2	0,083	0,026	1,086	**10,344**	**0,001**	0,032	0,133
	LYM	-0,165	0,059	0,848	**7,806**	**0,005**	-0,282	-0,049
	Glikoz	-0,019	0,007	0,981	**8,616**	**0,003**	-0,032	-0,006
	HGB	0,115	0,057	1,121	**4,010**	**0,045**	0,002	0,227
7	Sabit	2,811	1,353	16,621	**4,316**	**0,038**	0,159	5,463
	WBC	-0,246	0,057	0,782	**18,836**	**<0,001**	-0,357	-0,135
	pO2	0,094	0,026	1,099	**12,772**	**<0,001**	0,042	0,146
	LYM	-0,164	0,061	0,848	**7,376**	**0,007**	-0,283	-0,046
	Glikoz	-0,020	0,007	0,981	**7,823**	**0,005**	-0,033	-0,006
	HGB	0,241	0,080	1,272	**9,056**	**0,003**	0,084	0,398
	HCT	-0,086	0,034	0,917	**6,486**	**0,011**	-0,152	-0,020
8	Sabit	0,260	1,535	1,297	0,029	0,866	-2,750	3,269
	WBC	-0,229	0,059	0,795	**14,897**	**<0,001**	-0,345	-0,113
	pO2	0,100	0,030	1,105	**11,277**	**<0,001**	0,042	0,158
	LYM	-0,145	0,069	0,865	**4,472**	**0,034**	-0,280	-0,011
	Glikoz	-0,017	0,007	0,983	**5,368**	**0,021**	-0,031	-0,003
	HGB	0,181	0,090	1,199	**4,081**	**0,043**	0,005	0,358
	HCT	-0,325	0,071	0,723	**20,685**	**<0,001**	-0,465	-0,185
	RBC	1,089	0,254	2,972	**18,409**	**<0,001**	0,592	1,587

Tablo 3.4'te yapılan analizlerde referans kategorimiz buzağı ishali olma durumu olup stepwise yöntemi ile analizler gerçekleştirilmiştir.

WBC ölçümünde elde edilen model katsayısı -0,229 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(-0,229)=0,795 olup WBC değişkeninde bir birimlik artış buzağı ishalinin görülme riskini (1-0,795=0,205) yaklaşık %20 azaltmaktadır (*Wald* istatistiği=14,897; $p<0,001$). pO_2 ölçümünde elde edilen model katsayısı 0,100 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(0,100)=1,105 olup pO_2 değişkeninde bir birimlik artış buzağı ishali olma riskini 1,105 kat arttırmaktadır (*Wald* istatistiği=11,277; $p<0,001$). LYM ölçümünde elde edilen model katsayısı -0,145 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(-0,145)=0,865 olup LYM değişkeninde bir birimlik artış buzağı ishalinin görülme riskini (1-0,865=0,135) yaklaşık %14 azaltmaktadır (*Wald* istatistiği=4,472; $p=0,034$). Glikoz ölçümünde elde edilen model katsayısı -0,017 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(-0,017)=0,983 olup Glikoz değişkeninde bir birimlik artış buzağı ishalinin görülme riskini (1-0,983=0,017) yaklaşık %2 azaltmaktadır (*Wald* istatistiği=5,368; $p=0,021$). HGB ölçümünde elde edilen model katsayısı 0,181 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(0,181)=1,199 olup HGB değişkeninde bir birimlik artış buzağı ishalinin görülme riskini 1,199 kat arttırmaktadır (*Wald* istatistiği=4,081; $p=0,043$). HCT ölçümünde elde edilen model katsayısı -0,325 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(-0,325)= 0,723 olup HCT değişkeninde bir birimlik artış buzağı ishalinin görülme riskini (1-0,723=0,277) yaklaşık %28 azaltmaktadır (*Wald* istatistiği=20,689; $p<0,001$). RBC ölçümünde elde edilen model katsayısı 1,089 olarak hesaplanmıştır. Değişkenin Odds oranı Exp(1,089)=2,972 olup RBC değişkeninde bir birimlik artış buzağı ishalinin görülme riskini 2,972 kat arttırmaktadır (*Wald* istatistiği=18,409; $p<0,001$).

Uygulanan analizler sonucunda enter metodu ile 14 değişkenin bulunduğu ilk model (Model 1) oluşturulmuştur. Stepwise yöntemi kullanılarak en uygun model belirlenmiş ve 14 değişkenden 7 değişken modelden dışlanarak 7 değişkenli ikinci model (Model 2) oluşturulmuştur. Oluşturulan iki model Makine Öğrenmesi Algoritmaları ile bundan sonraki bölümlerde değerlendirilmiştir.

3.3. Makine Öğrenmesi Algoritmaları Sonuçları

Makine öğrenmesi algoritmalarının uygulanması aşamasında Lojistik regresyon modelleri ile elde edilen Model 1 ve Model 2'deki değişkenler kullanılarak incelemeler gerçekleştirilmiştir.

Model 1 ile elde edilen modelde Boosting yöntemi için validasyon doğruluğu 0,111 iken test doğruluğu 0,136 olarak bulundu. KNN yöntemi için validasyon doğruluğu 0,889 iken test doğruluğu 0,841 olarak bulundu. Lineer Diskriminant yöntemi için Test doğruluğu 0,818 olarak bulundu. RF yöntemi validasyon doğruluğu 0,917 iken test doğruluğu 0,909 olarak bulundu.

Model 2 ile elde edilen modelde ise Boosting yöntemi için validasyon doğruluğu 0,194 iken test doğruluğu 0,841 olarak bulundu. KNN yöntemi için validasyon doğruluğu 0,841 iken test doğruluğu 0,909 olarak bulundu. Lineer Diskriminant yöntemi için Test doğruluğu 0,864 olarak bulundu. RF yöntemi validasyon doğruluğu 0,900 iken test doğruluğu 0,955 olarak bulundu. Elde edilen Sonuçlar Tablo 3.5'te verilmiştir.

Tablo 3.5. Makine Öğrenmesi Model Uygunluk Sonuçları

Algoritma	**Model 1**		**Model 2**	
	Test Doğruluğu	Validasyon Doğruluğu	Test Doğruluğu	Validasyon Doğruluğu
Boosting	0,111	0,136	0,194	0,841
KNN	0,889	0,841	0,841	0,909
Lineer Diskriminant	0,864	-	0,864	-
RF	0,917	0,909	0,900	0,955

3.4. Lojistik Regresyon ile Makine Öğrenmesi Algoritmaları Performans Sonuçları

Lojistik Regresyonda Model 1 ile 14 değişken analiz edilmiş ardından Model 2 ile model için en uygun 7 değişken bulunduğu model elde edilmiştir. Bu iki farklı yöntemle oluşturulan modeller Lojistik Regresyon, Boosting, KNN, Lineer Diskriminant ve RF yöntemleri ile elde edilen sonuçların performansı değerlendirilmiştir. Gerçek değerler ile tahmin edilen değerler arasındaki uyum Tablo 3.6'da verilmiştir.

Tablo 3.6. Lojistik Regresyon ve Makine Öğrenmesi Algoritmaları Hata Matrisi

Algoritma		**Model 1**		**Model 2**	
		Doğru	Yanlış	Doğru	Yanlış
Lojistik Regresyon	Doğru	%28,5	%6	%28,5	%4,5
	Yanlış	%6	%69,5	%6	%71
Boosting	Doğru	%26,5	%5	%26	%3,5
	Yanlış	%8	%70,5	%8,5	%72
KNN	Doğru	%24	%6	%2	%5,5
	Yanlış	%10,5	%69,5	%8,5	%70
Lineer Diskriminant	Doğru	%23	%6	%24	%4
	Yanlış	%11,5	%69,5	%10,5	%71,5
RF	Doğru	%30	%0,5	%33	%1
	Yanlış	%4,5	%75	%1,5	%74,5

Tablo 3.6'ya göre 14 değişkenin analiz edildiği Model 1'e göre ishal bulunan buzağıları Lojistik Regresyon %28,5 doğrulukla, Boosting %26,5 doğrulukla, KNN %23 doğrulukla, Lineer Diskriminant %23 doğrulukla ve RF %30 doğrulukla tahmin etmiştir.

Model 2 göre ishal bulunan buzağıları Lojistik Regresyon %28,5 doğrulukla, Boosting %26 doğrulukla, KNN %24 doğrulukla, Lineer Diskriminant %23 doğrulukla ve RF %33 doğrulukla tahmin etmiştir.

Oluşturulan hata matrisi ile yöntemlerin performansını değerlendirmek için Doğruluk, AUC, Duyarlılık, Seçicilik ve F Ölçüsü kullanılmıştır. Lojistik Regresyon, Boosting, KNN, Lineer Diskriminant ve RF algoritmalarının performans sonuçları Tablo 3.7'de verilmiştir.

Tablo 3.7. Lojistik Regresyon ve Makine Öğrenmesi Algoritmaları Performans Metrikleri

Algoritma	Performans Metrikleri	Model 1	Model 2
Lojistik Regresyon	Doğruluk	%89,1	%90,5
	AUC	%94,7	%94,6
	Duyarlılık	%82,6	%82,6
	Seçicilik	%92,1	%94,0
	F Ölçüsü	%82,6	%84,4
Boosting	Doğruluk	%88,2	%89,1
	AUC	%85,1	%85,4
	Duyarlılık	%76,8	%75,4
	Seçicilik	%93,4	%95,4
	F Ölçüsü	%80,3	%81,3
KNN	Doğruluk	%85	%87,3
	AUC	%80,8	%84
	Duyarlılık	%69,6	%75,4
	Seçicilik	%92,1	%92,7
	F Ölçüsü	%74,4	%78,8
Lineer Diskriminant	Doğruluk	%84,1	%86,8
	AUC	%79,4	%82,1
	Duyarlılık	%66,7	%69,6
	Seçicilik	%92,1	%94,7
	F Ölçüsü	%72,4	%76,8
RF	Doğruluk	%95,5	%97,7
	AUC	%93,1	%97,2
	Duyarlılık	%87	%95,7
	Seçicilik	%99,3	%98,7
	F Ölçüsü	%92,3	%96,4

Tablo 3.7'ye göre Model 1'e göre Lojistik Regresyon algoritması için oluşturulan modelin performansı toplam doğruluk oranı %89,1, kontrol grubunun doğru bilinme oranı %92,1 ve ishal bulunan buzağıları doğru bilme oranı %82,6 olarak bulundu. AUC ve F1 oranları sırası ile %94,7 ve %82,6 olarak bulundu. Model 2'ye göre ise performansı toplam doğruluk oranı %90,5, kontrol grubunun doğru bilinme oranı %94 ve ishal bulunan buzağıları doğru bilme oranı %92,6 olarak bulundu. AUC ve F1 oranları sırası ile %94,6 ve %84,4 olarak bulundu.

Model 1'e göre KNN algoritması için oluşturulan modelin performansı toplam doğruluk oranı %85, kontrol grubunun doğru bilinme oranı %92,1 ve ishal bulunan buzağıları doğru bilme oranı %69,6 olarak bulundu. AUC ve F1 oranları sırası ile %80,8 ve %74,4 olarak bulundu. Model 2'ye göre ise performansı toplam doğruluk oranı %87,3, kontrol grubunun doğru bilinme oranı %92,7 ve ishal bulunan buzağıları doğru bilme oranı %75,4 olarak bulundu. AUC ve F1 oranları sırası ile %84 ve %78,8 olarak bulundu.

Model 1'e göre Lineer Diskriminant algoritması için oluşturulan modelin performansı toplam doğruluk oranı %84,1, kontrol grubunun doğru bilinme oranı %92,1 ve ishal bulunan buzağıları doğru bilme oranı %66,7 olarak bulundu. AUC ve F1 oranları sırası ile %79,4 ve %72,4 olarak bulundu. Model 2'ye göre ise performansı toplam doğruluk oranı %86,8, kontrol grubunun doğru bilinme oranı %94,7 ve ishal bulunan buzağıları doğru bilme oranı %69,6 olarak bulundu. AUC ve F1 oranları sırası ile %82,1 ve %76,8 olarak bulundu.

Model 1'e göre RF algoritması için oluşturulan modelin performansı toplam doğruluk oranı %95,5, kontrol grubunun doğru bilinme oranı %99,3 ve ishal bulunan buzağıları doğru bilme oranı %87 olarak bulundu. AUC ve F1 oranları sırası ile %93,1 ve %92,3 olarak bulundu. Model 2'ye göre ise performansı toplam doğruluk oranı %97,7, kontrol grubunun doğru bilinme oranı %98,7 ve ishal bulunan buzağıları doğru bilme oranı %95,7 olarak bulundu. AUC ve F1 oranları sırası ile %97,2 ve %96,4 olarak bulundu.

Elde edilen sonuçlar doğrultusunda Lineer Diskriminant algoritmasında ikinci modelde doğruluk %2,7 oranında artmıştı. KNN algoritmasında ikinci modelde

doğruluk %2,3 oranında artmıştı. RF algoritmasında ikinci modelde doğruluk %2,2 oranında artmıştı. Lojistik Regresyon algoritmasında ikinci modelde doğruluk %1,4 oranında artmıştı. Boosting algoritmasında ikinci modelde doğruluk %0,9 oranında artmıştı. Bu sonuçlara göre her iki modelde de RF yöntemi etkili sonuçlar ortaya koyarken daha az değişkenin bulunduğu modelde artış gösteren yöntemler sırası ile Lineer Diskriminant, KNN, RF, Lojistik Regresyon ve Boosting algoritmalarıydı.

4.TARTIŞMA

Bu çalışmada, buzağılara ait verilerin istatistik analizleri, lojistik regresyon ve makine öğrenmesi algoritmaları kullanılarak gerçekleştirilmiş olup yöntemlerin performans karşılaştırmaları yapılmıştır.

Çalışmada 14 değişkenin bulunduğu modelde doğruluk oranı lojistik regresyon ile %89,1, Boosting ile %88,2, KNN ile %85, lineer diskriminant ile %84,1 ve RF ile %95,5 doğrulukla tespit edilmiştir. Bu algoritmalardan en verimli çalışan yöntemler sırasıyla RF, Boosting, lojistik regresyon, KNN ve Lineer Diskriminant olarak belirlenmiştir. Romero ve ark (2021), yaptıkları çalışmada, İngiltere'deki sığır tüberkülozunu tahmin etmek için RF ve lojistik regresyon sınıflandırma algoritmalarını kullanmışlardır. RF algoritması düşük risk gruplarında, lojistik regresyon algoritması ise diğer risk gruplarında daha doğru sonuçlar vermesi bakımından bu çalışma ile benzerlik göstermektedir.

14 değişkenin bulunduğu modelde doğru tahminde en iyi sonucu %95,5 doğruluk oranı ile RF yöntemi ortaya koymuştur. Bu oranın %99,3'si ishal bulunmayan grubu %87'si ise ishal bulunan grubu tespit etmede kullanılmıştır. Fadul Pacheco ve ark (2021), yaptıkları çalışmada, ilk emzirmelerinde klinik mastit durumları pozitif olan ineklerin mastit başlangıcını tahmin etmek için RF, NB ve Boosting sınıflandırma algoritmaları kullanılmışlardır. %72 doğruluk oranı ile RF yönteminin en verimli yöntem olduğu bulunmuştur. Mani ve ark (2021), yaptıkları çalışmada ise peletlenmiş yem, probiyotik peletlenmiş yem, probiyotik inci boncuklar ve probiyotik pirinç pufu ile beslenen sazan balıklarının besleme rejimlerine ait veri setlerini kullanarak RF, doğrusal model ve karar ağacı algoritmalarını değerlendirmişlerdir. En uygun yem türüne göre sazan balıklarını sınıflandırmışlardır. RF algoritması en yüksek doğruluk oranını (%98) vermiştir. Bu çalışmada elde edilen sonuçlara göre tüm veriler kullanılarak değerlendirmeler yapıldığında benzer sonuçlar ortaya konulmuştur.

Lojistik regresyon stepwise yöntemiyle elde edilen 7 değişkenin bulunduğu modelde de en iyi sonucu %96,4 doğruluk oranı ile RF algoritması ortaya koymuştur. İlk model için elde edilen sonuçlar bu çalışmalar ile paralellik gösterirken daha az

değişkenin bulunduğu modelde de RF algoritması daha verimli sonuçlar ortaya koymuştur.

Çalışmada 220 buzağı ile yapılan değerlendirmeler sonucunda 14 değişkenin bulunduğu modelde en iyi performansı sırasıyla RF, Boosting, Lojistik Regresyon, KNN ve Lineer Diskriminant algoritmaları, 7 değişkenin bulunduğu modelde ise sırasıyla RF, Lineer Regresyon, Boosting, KNN ve Lineer Diskriminant algoritmaları en iyi performansı göstermişlerdir. Tohalino ve Amancio (2022) yaptıkları çalışmada KNN, Destek Vektör Makineleri, NB, Yapay Sinir Ağları ve Karar ağaçları algoritmalarını 31.000 örneklemde kullanmış en kötü performansı KNN algoritmasında elde etmiş olup sonuçlar bu çalışmadaki 7 değişkenin kullanıldığı KNN ile benzerlik göstermektedir.

Buzağılara ait kan gazı, hematolojik ve biyokimyasal ölçümlerin buzağı ishalinin öngörmede etkisi değerlendirildiğinde RF 14 değişken bulunduğu modelde %95,5 doğruluk oranı, 7 değişken bulunduğunda ise %97,5 doğruluk oranı ortaya koymuştur. Zimpel ve ark (2021), yaptıkları çalışmada, hayvan refahı risklerini tahmin etmek için RF, Destek Vektör Makineleri, NB ve Lojistik Regresyon makine öğrenmesi sınıflama algoritmalarını kullanmışlardır. Veri seti olarak 10 hayvan sağlığı riski ve 14 emzirme özelliği göstergesi olan 57.000'den fazla domuz kullanmışlardır. Domuzlarla ilgili olan veri setinde 10 değişken ayrı ayrı incelenerek tüm değişkenler için en iyi model seçilmiştir. Sonuç olarak emzirme döneminde ölümü belirlemede RF algoritmasının %83,7 ile en yüksek doğruluk oranına sahip olduğu tespit edilmiştir. Elde edilen sonuçlar bakımından paralellik olduğu ortaya konulmuştur.

Cihan ve Coşkun (2021), yaptıkları çalışmada, diyabet hastalığının tespitinde etkili değişkenler iki farklı yaklaşımla değerlendirildi. 39 değişken incelenerek ilk yöntemde anlamlı çıkan 18 değişken ikinci modelde ise uzmanlar tarafından belirlenen 21 değişken Lojistik Regresyon, KNN, Destek Vektör Makinesi, NB, Karar Ağacı, RF ve Yapay Sinir Ağı olmak üzere yedi makine öğrenimi sınıflandırma algoritmasının performansları değerlendirildi. İlk yaklaşım için %84,48 doğruluk oranı ve ikinci yaklaşımda %81,90 doğruluk oranı ile en verimli sonuçları ortaya koymuş olup bu çalışmadaki sonuçlarla benzerlik göstermektedir.

5. SONUÇ VE ÖNERİLER

Lojistik regresyondaki enter yönteminde araştırmacının modelde tutmak istediği değişkenler kullanılarak analizler gerçekleştirilirken; stepwise yönteminde değişkenlerin seçilmesi veya modelden çıkartılması değişkenlerin önemliliğini kontrol eden bir istatistiksel işleme göre yapılmaktadır. Regresyon yöntemlerindeki bu yaklaşımlar makine öğrenmesi algoritmalarında söz konusu değildir. Yani modeldeki değişkenlerin seçimi, değişkenlerin katkı düzeylerine göredir. Buradaki yaklaşım değişkenlerin öznitelik olarak değerlendirildiği bağımlı değişkenin boyutunu ifade etmektedir.

Sağlık ve veterinerlik alanında hastalığı ön görmede etkili değişkenler anlamsızlığa yol açması sebebiyle modelden dışlanmaktadır. Bu durum araştırmacının kontrolü dışında gerçekleşmekte olup uygun olmayan sonuçların ortaya çıkmasına sebep olabilmektedir. Makine öğrenmesi algoritmalarının sınıflandırma problemlerindeki kullanımının değerlendirildiği bu çalışmada sağlık ve veterinerlik alanında yapılacak analizlerde karşılaşılacak değişkenlerin dışlanma durumu ortadan kaldırılabilir. Bu alanda yapılmış çalışmaların oldukça az olduğu ve makine öğrenmesi yaklaşımlarının modelde değişkenlerin tutulması bakımından daha katkısal olduğu sonucu bu çalışmanın öne çıkan tarafı olup literatüre katkı sağlayacağı öngörülmektedir.

6. KAYNAKLAR

Akçay A, 2009. Broiler coccidiosis'inde risk faktörlerinin lojistik regresyon analizi ile belirlenmesi, Yayınlanmamış Doktora, Tezi Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Aksoy S, Özavşar M, Altındal A, 2022. Classification of VOC Vapors Using Machine Learning Algorithms. Journal of Engineering Technology Applied Sciences, 7, 97-107.

Aktan E, 2018. Büyük veri: Uygulama alanları, analitiği ve güvenlik boyutu. Bilgi Yönetimi, 1, 1-22.

Alkhamis MA, Fountain-Jones NM, Aguilar-Vega C, Sánchez-Vizcaíno JM, 2021. Environment, vector, or host? Using machine learning to untangle the mechanisms driving arbovirus outbreaks. Ecological Applications, 31, e02407.

Alpar R, 2013. Uygulamalı çok değişkenli istatistiksel yöntemler, Ankara: Detay Yayıncılık, 637-712.

Alves AAC, Espigolan R, Bresolin T, Costa RM, Fernandes Júnior GA, Ventura RV, Carvalheiro R, Albuquerque LG, 2021. Genome-enabled prediction of reproductive traits in Nellore cattle using parametric models and machine learning methods. Animal Genetics, 52, 32-46.

Arslanoğlu E, 2019. Bınary lojistik regresyon analizinde model uyumunda kullanılan testlerin performanslarının değerlendirilmesi (Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü).

Austin PC, Steyerberg EW, 2012. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC medical research methodology, 12, 1-8.

Bakoev S, Traspov A, Getmantseva L, Belous A, Karpushkina T, Kostyunina O, Usatov A, Tatarinova TV, 2021. Detection of genomic regions associated malformations in newborn piglets: a machine-learning approach. PeerJ, 9, e11580.

Balakrishnama S, Ganapathiraju A, 1998. Linear discriminant analysis-a brief tutorial. Institute for Signal information Processing 18, 1-8.

Banzato T, Wodzinski M, Burti S, Osti VL, Rossoni V, Atzori M, Zotti A, 2021. Automatic classification of canine thoracic radiographs using deep learning. Scientific Reports, 11, 1-8.

Baydan B, Ünvar HM, 2021. Detection of tibial fractures in cats and dogs with deep learning. Ankara Üniversitesi Veteriner Fakültesi Dergisi, 68, 283-290.

Beşer BÖ, Yaygın M, Sarıbaş ES, 2021. Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25, 112-120.

Bland JM, Altman DG, 2000. The odds ratio. British Medical Journal, 320, 1468.

Bollig N, Clarke L, Elsmo E, Craven M, 2020. Machine learning for syndromic surveillance using veterinary necropsy reports. Journal Plos one, 15, e0228105.

Breiman L, 2001. Random forests. Machine learning, 45, 5-32.

Brock J, Lange M, Tratalos JA, More SJ, Graham DA, Guelbenzu-Gonzalo M, Thulke H-H, 2021. Combining expert knowledge and machine-learning to classify herd types in livestock systems. Scientific Reports, 11, 1-10.

Boucher KM, Slattery ML, Berry TD, Quesenberry C, Anderson K, 1998. Statistical methods in epidemiology: a comparison of statistical methods to analyze dose–response and trend analysis in epidemiologic studies. Journal of clinical epidemiology, 51, 1223-1233.

Canbek G, Sagiroglu S, Temizel TT, Baykal N, 2017. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. Paper presented at the 2017 International Conference on Computer Science and Engineering (UBMK).

Carslake C, Vázquez-Diosdado JA, Kaler J, 2020. Machine learning algorithms to classify and quantify multiple behaviours in dairy calves using a sensor: Moving beyond classification in precision livestock. Sensors, 21, 88.

Charan PVS, Anand PM, Shukla SK, 2021. Study of Data Mining and Machine Learning Techniques in Advanced Persistent Threat Attribution and Detection. In Data Mining Concepts and Applications. Eds: Thomas C, Third Ed. London, UK. In (pp. s. 1-16).

Cihan P, Coşkun H, 2021. Performance comparison of machine learning models for diabetes prediction. 29th Signal Processing and Communications Applications Conference s. 1-4, 9 June 2021.

Cihan P, Gokce E, Kalipsiz O, 2017. A review of machine learning applications in veterinary field. Kafkas Universitesi Veteriner Fakultesi Dergisi, 23.

Cihan P, Kalıpsız O, Gökçe E, 2020. Yenidoğan kuzularda bilgisayar destekli tanı. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 26, 385-391.

Çokluk Ö, 2010. Lojistik regresyon analizi: Kavram ve uygulama. Kuram ve Uygulamada Eğitim Bilimleri, 10, 1357-1407.

Cover T, Hart P, 1967. Nearest neighbor pattern classification. IEEE transactions on information theory, 13, 21-27.

Çağlıyan B, Köse U, 2021. Epilepsi EEG Verilerinin Makine Öğrenmesi Teknikleriyle Sınıflandırılması. Avrupa Bilim ve Teknoloji Dergisi, 23, 163-172.

Çolak C, 2001. Lojistik regresyon analizi ve sağlık bilimlerinde bir uygulama, Yüksek Lisans Tezi, İnönü Üniversitesi Sağlık Bilimleri Enstitüsü, Malatya.

De Barros AdMC, Silva AFR, Zibordi M, Spagnolo JD, Corrêa RR, Belli CB, de Camargo MM, 2021. Equine simplified acute physiology score: Personalised medicine for the equine emergency patient. Veterinary Record, 189, no-no.

DeMaris A, 1995. A tutorial in logistic regression. Journal of Marriage the Family, 1, 956-968.

Ergün GB, Günay S, Ergün TG, 2021. Köpeklerdeki Uzun Kemiklerin Evrişimsel Sinir Ağları Kullanılarak Sınıflandırılması. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 33, 125-132.

Esener N, Maciel-Guerra A, Giebel K, Lea D, Green MJ, Bradley AJ, Dottorini T, 2021. Mass spectrometry and machine learning for the accurate diagnosis of benzylpenicillin and multidrug resistance of Staphylococcus aureus in bovine mastitis. PLoS computational biology, 17, e1009108.

Fadul-Pacheco L, Delgado H, Cabrera VE, 2021. Exploring machine learning algorithms for early prediction of clinical mastitis. International Dairy Journal, 119, 105051.

Fan J, Upadhye S, Worster A, 2006. Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine, 8, 19-20.

Ferreira A, Figueiredo M, 2012. Boosting algorithms: A review of methods, theory, and applications. Ensemble machine learning, 35-85.

Fisher RA, 1936. The use of multiple measurements in taxonomic problems. Annals of eugenics, 7, 179-188.

Freund Y, 1995. Boosting a weak learning algorithm by majority. Information computation, 121, 256-285.

Gamgam H, Altunkaynak B, 2017. Regresyon Analizi, Seçkin Yayıncılık, Ankara.

Gök M, 2017. Makine öğrenmesi yöntemleri ile akademik başarinin tahmin edilmesi. azi University Journal of Science Part C: Design and Technology, 5, 139-148.

Gökçe E, Kalıpsız O, Cihan P, 2017. Hayvan Hastalığı Teşhisinde Normalizasyon Tekniklerinin Yapay Sinir Ağı ve Özellik Seçim Performansına Etkisi. . Turkish Studies (Elektronik), 12, 59-70.

Güç K, 2015. Türkiye'de resmi kurumlara duyulan güvenin kategorik regresyon ve lojistik regresyon analizi ile incelenmesi. Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Hinton GE, Salakhutdinov RR, 2006. Reducing the dimensionality of data with neural networks. Science, 313, 504-507.

Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S, 1997. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine, 16, 965-980.

Hosmer DW, Lemeshow S, 2013. Applied logistic regression, John Wiley & Sons, Canada, 398.

Hyde RM, Down PM, Bradley AJ, Breen JE, Hudson C, Leach KA, Green MJ, 2020. Automated prediction of mastitis infection patterns in dairy herds using machine learning. Scientific Reports, 10, 1-8.

James G, Witten D, Hastie T, Tibshirani R, 2013. Classification. In An introduction to statistical learning. FirstEd. Witten, Hastie & Tibshirani.

Jing Z, Wang G, Zhang S, Qiu C, 2017. Building Tianjin driving cycle based on linear discriminant analysis. Transportation Research Part D: Transport Environment, 53, 78-87.

Kalaycı S, 2005. SPSS Applications, Multi-variables Statistics Techniques. Asil Yayin Dagitim, Ankara, s: 273-305.

Kara ÖS, 2015. Lojistik regresyon analizi ve kadın işgücü üzerine bir uygulama. Yüksek Lisans Tezi, Uludağ Üniversitesi Sosyal Bilimler Enstitüsü, Bursa.

Karagöz Y, 2015. SPSS 22 Uygulamalı Biyoistatistik, Güncellenmiş 2. Basım, Nobel Yayınevi, Ankara.

Karaman U, Çiçek İB, 2020. Determination Of Cryptosporidium Spp. Risk Factors Using Multilayer Perceptron Neural Network And Radial Based Functional Artificial Neural Network Method. The Journal of Cognitive Systems, 5, 83-87.

Kılıç S, 2015. Binary logistic regression analysis. Journal of Mood Disorders, 5, 191-4.

Kıymaz YE, 2021. Güneş enerjisi santrallerinde derin öğrenme kullanılarak elektrik üretimi tahmininin yapılması. (Yüksek Lisans Tezi), Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü, Konya.

Kleinbaum DG, Hedeker D, 1996. Logistic regression: a self-learning text. Statistical Methods in Medical Research, 5,103.

Konur U, 2020. Sınıflandırma Başarımını Ölçme ve Seyreklik İşleme Üzerine. EMO Bilimsel Dergi, 10, 43-56.

Lantz B, 2019. Introducing Machine Learning. In Expert techniques for predictive modeling: Machine learning with. Second Ed. Birmingham, UK: Packt publishing ltd.

Lee KC, Ou JS, 2007. Radar target recognition by using linear discriminant algorithm on angular-diversity RCS. Journal of Electromagnetic Waves Applications, 21, 2033-2048.

Lee KI, Koval JJ, 1997. Determination of the best significance level in forward stepwise logistic regression. Commun Stat Simul Comput, 26, 559–75

Liu D, Caliskan S, Rashidfarokhi B, Oldenhof H, Jung K, Sieme H, Hilfiker A, Wolkers WF, 2021. Fourier transform infrared spectroscopy coupled with machine learning classification for identification of oxidative damage in freeze-dried heart valves. Scientific reports,, 11, 1-13.

Lusted LB, 1984. ROC recollected. Medical Decision Making, 4, 131-135.

Mani SR, Kandhasamy P, Vijayan K, Palanichamy M, Jacob JP, Kandhasamy S, 2021. Haematological parameters of Cyprinus carpio with reference to probiotic feed: A machine learning approach. The Israeli Journal of Aquaculture-Bamidgeh, 73, 1-10.

Martinez AM, Kak AC, 2001. Pca versus lda. IEEE transactions on pattern analysis machine intelligence, 23, 228-233.

Metlek S, Kayaalp K, 2021. Arı hastalıklarının hibrit bir derin öğrenme yöntemi ile tespiti. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 36, 1715-1732.

Minsky M, Papert S, 1969. An introduction to computational geometry. Cambridge tiass., HIT, 479, 480.

Minsky ML, 1967. Finite and Infinite Machines. In Computation. First British Ed. Prentice-Hall Englewood Cliffs, N.J.

Narkhede S, 2018. Understanding auc-roc curve. Towards Data Science, 26, 220-227.

Newell A, Simon H, 1956. The logic theory machine-A complex information processing system. IRE Transactions on information theory, 2, 61-79.

Ottenbacher KJ, Ottenbacher HR, Tooth L, Ostir GV, 2004. A review of two journals found that articles using multivariable logistic regression frequently did not report commonly recommended assumptions. Journal of clinical epidemiology, 57, 1147-1152.

Pal M, 2005. Random forest classifier for remote sensing classification. International journal of remote sensing, 26, 217-222.

Pandit PS, Williams DR, Rossitto P, Adaska JM, Pereira R, Lehenbauer TW, Byrne BA, Li X, Atwill ER, Aly SS, 2021. Dairy management practices associated with multi-drug resistant fecal commensals and Salmonella in cull cows: a machine learning approach. PeerJ, 9, e11732.

Paploski IAD, Bhojwani RK, Sanhueza JM, Corzo CA, VanderWaal K, 2021. Forecasting viral disease outbreaks at the farm-level for commercial sow farms in the US. Preventive Veterinary Medicine, 196, 105449.

Pelit A, İbrikçi T, Özalp P, Taştekin B, 2019. Classification of Biochemical and Biomechanical Data of Diabetic Rats Treated with Magnetic Field By PCA-Supported J48 Algorithm. Kafkas Universitesi Veteriner Fakultesi Dergisi, 25, 741-747.

Peng CYJ, So TSH, 2002. Logistic regression analysis and reporting: A primer. Understanding Statistics: Statistical Issues in Psychology, Education, the Social Sciences, 1, 31-70.

Pirim H, 2006. Yapay zeka. Yaşar Üniversitesi E-Dergisi, 1, 81-93.

Rao CR, 1948. The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society. Series B, 10, 159–203

Rodrigues ÉO, 2018. Combining Minkowski and Chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier. Pattern Recognition Letters, 110, 66-71.

Romero MP, Chang Y-M, Brunton LA, Prosser A, Upton P, Rees E, Tearne O, Arnold M, Stevens K, Drewe JA, 2021. A comparison of the value of two machine learning predictive models to support bovine tuberculosis disease control in England. Preventive Veterinary Medicine, 188, 105264.

Rosenblatt F, 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65, 386–408.

Rumelhart DE, Hinton GE, Williams RJ, 1986. Learning representations by back-propagating errors. Nature, 323, 533-536.

Sacchi MD, 1998. A bootstrap procedure for high-resolution velocity analysis. Geophysics, 63, 1716–1725.

Sagiroglu S, Sinanc D, 2013. Big data: A review. 2013 International Conference on Collaboration Technologies and Systems. s. 42-47, 20 April 2013.

Scott M, 2001. Applied Logistic Regression Analysis, Sage University Paper, 67-90.

Sharma S, 1996. Applied Multivariate Techniques. John Wiley and Sons,Inc.

Springenberg JT, 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. 4th International Conference on Learning Representations, ICLR 2016. s. 1-20, 2 May 2016.

Stephenson B, 2008. Binary response and logistic regression analysis.

Swets JA, 1988. Measuring the accuracy of diagnostic systems. Science, 240, 1285-1293.

Sykes AL, Silva GS, Holtkamp DJ, Mauch BW, Osemeke O, Linhares DC, Machado G, 2022. Interpretable machine learning applied to on-farm biosecurity and porcine reproductive and respiratory syndrome virus. Transboundary Emerging Diseases, 69, e916-e930.

Şahinbaş K, Gülseçen S, 2019. Dengesiz Klinik Veriler İçin Karar Destek Önerisi: Akut Apandisit Örneği. (Doktora Tezi), İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul,

Taşcı E, Onan A, 2016. K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi. Akademik Bilişim, 1, 4-18.

Tinkov OV, Grigorev VY, Grigoreva LD, 2021. QSAR analysis of the acute toxicity of avermectins towards Tetrahymena pyriformis. SAR and QSAR in Environmental Research, 32, 541-571.

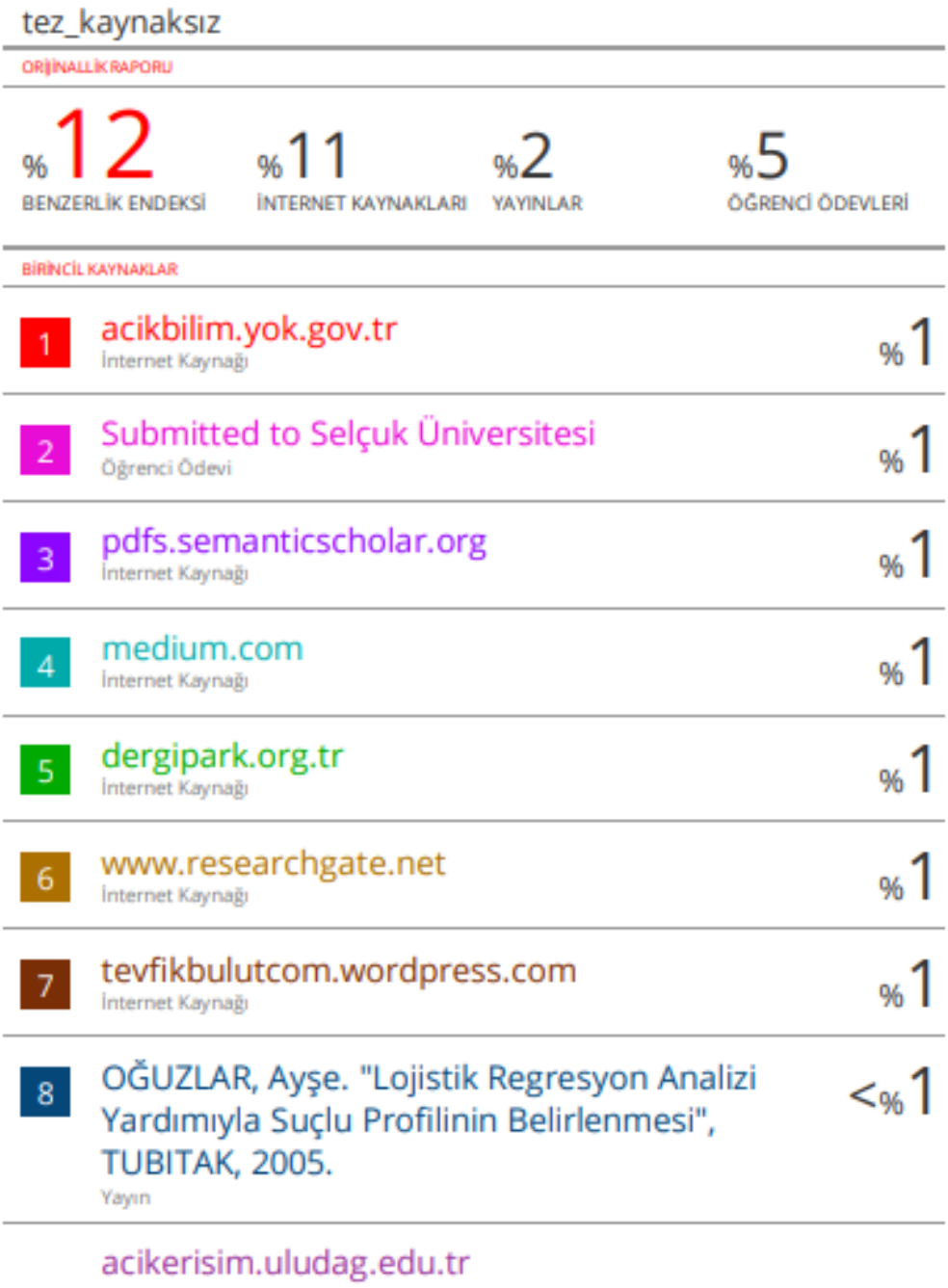
Tohalino JA, Amancio DR, 2022. On predicting research grants productivity via machine learning. Journal of Informetrics, 16, 101-260.

Wilson S, Close M, Abraham P, 2018. Applying linear discriminant analysis to predict groundwater redox conditions conducive to denitrification. Journal of Hydrology, 556, 611-624.

Yazar GN, 2018. Aile içi kadına yönelik şiddeti etkileyen faktörlerin lojistik regresyon ile analizi. Yüksek Lisans Tezi, Uludağ Üniversitesi Sosyal Bilimler Enstitüsü, Bursa.

Zimpel T, Riekert M, Klein A, Hoffmann C, 2021. Machine learning for predicting animal wel-fare risks in pig farming. Landtechnik, 76, 24-35.

EK-B: Turnitin Raporu



tez_kaynaksız

ORİJİNALLİK RAPORU

%12	%11	%2	%5
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

1	acikbilim.yok.gov.tr İnternet Kaynağı	%1
2	Submitted to Selçuk Üniversitesi Öğrenci Ödevi	%1
3	pdfs.semanticscholar.org İnternet Kaynağı	%1
4	medium.com İnternet Kaynağı	%1
5	dergipark.org.tr İnternet Kaynağı	%1
6	www.researchgate.net İnternet Kaynağı	%1
7	tevfikbulutcom.wordpress.com İnternet Kaynağı	%1
8	OĞUZLAR, Ayşe. "Lojistik Regresyon Analizi Yardımıyla Suçlu Profilinin Belirlenmesi", TUBITAK, 2005. Yayın	<%1
	acikerisim.uludag.edu.tr	