ÇUKUROVA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

İkili Lojistik Regresyon Analizinde Aykırı Değer Belirleme Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi

Melis ÇELİK GÜNEY

Zootekni Anabilim Dalı

Mart, 2023

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZ ONAYI

İkili Lojistik Regresyon Analizinde Aykırı Değer Belirleme Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi

---

Melis ÇELİK GÜNEY

Zootekni Anabilim Dalı

Bu Doktora Tezi 31/03/2023 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından Değerlendirilmiş ve Oy Birliği / Oy Çokluğu ile Kabul Edilmiştir.

Jüri : Prof. Dr. G. Tamer KAYAALP (Danışman) ……

: Prof. Dr. Zeynel CEBECİ ……

: Prof. Dr. Gülsen KIRAL ……

: Prof. Dr. Hasan ÖNDER ……

: Doç. Dr. Yalçın TAHTALI ……

Bu Tez Fen Bilimleri Enstitüsü, Zootekni Anabilim Dalında Hazırlanmıştır.

Tez No:

Prof. Dr. Sadık DİNÇER
Enstitü Müdürü

Bu çalışma Ç. Ü. Araştırma Projeleri Birimi Tarafından Desteklenmiştir.
Proje No: FDK-2018-10287

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge ve fotoğrafların kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

İÇİNDEKİLER

ÖZ I
ABSTRACT II
TEŞEKKÜR III
ÇİZELGELER DİZİNİ IV
ŞEKİLLER DİZİNİ V
SİMGELER VE KISALTMALAR VI
1. GİRİŞ 1
2. ÖNCEKİ ÇALIŞMALAR 3
2.1. Lojistik Regresyon Analizi ile İlgili Çalışmalar 3
2.2. Aykırı Değerler ile İlgili Çalışmalar 8
3. MATERYAL VE METOT 17
3.1. Materyal 17
3.2. Metod 17
3.2.1. Lojistik Regresyon Analizi 17
3.2.2. Aykırı Değerler 25
4. BULGULAR VE TARTIŞMA 31
4.1. Yapay Veri Setine Ait Bulgular 31
4.1.1. Yapay Veri Setinde Lojistik Regresyon Analizine Ait Bulgular 32
4.1.2. Yapay Veri Setinde Aykırı Değer Belirleme Yöntemlerine Ait Bulgular 38
4.1.3. Aykırı Değer İçermeyen Yapay Veri Setinde Lojistik Regresyon Analizine Ait Bulgular 44
4.1.4. Aykırı Değer İçeren/İçermeyen Yapay Veri Setinin Karşılaştırılması 47
4.2. Gerçek Veri Setine Ait Bulgular 49
4.2.1. Aykırı Değer İçeren/İçermeyen Gerçek Veri Setinin Karşılaştırılması 50
5. SONUÇLAR VE ÖNERİLER 55
KAYNAKLAR 57
ÖZGEÇMİŞ 69
EKLER 70

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

İkili Lojistik Regresyon Analizinde Aykırı Değer Belirleme Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi

Melis ÇELİK GÜNEY

Danışman: Prof. Dr. G. Tamer KAYAALP

Zootekni Anabilim Dalı

ÖZ

Bu tez çalışmasında, aykırı değer belirleme yöntemlerinden istatistik tabanlı, yakınlık tabanlı, kümeleme tabanlı, sapma tabanlı ve yalıtım tabanlı aykırı değer belirleme yöntemleri karşılaştırılmıştır. R yazılımı ile standart normal dağılıma sahip bir simülasyon yapılmıştır. Üretilen veri setinin örnek büyüklüğü 3000 olarak seçilmiştir. 3 bağımsız 1 bağımlı değişkene sahip iki kategorili lojistik regresyona uygun bir veri seti üretilmiştir. Yöntemleri karşılaştırmak amacıyla veri setine 30 adet aykırı değer eklenerek veri modifiye edilmiştir. Çalışmanın sonucunda, yalıtım tabanlı aykırı değer belirleme yöntemlerinden iForest algoritması, eklenen tüm aykırı değerleri bulmuştur ve performansı en yüksek çıkan algoritma olduğu görülmüştür. Gerçek bir veri setinde de iForest algoritması kullanarak aykırı değerler belirlenmiştir. Ayrıca hem yapay veri de hem de gerçek veride aykırı değer içeren ve içermeyen veri setlerinde model tahmin edilmiştir. Tahmin edilen modelin uyumunun iyiliği ve yeterlilik ölçütleri incelenmiştir. Böylelikle aykırı değerlerin modele etkisi saptanmıştır.

Anahtar Kelimeler: Aykırı değer, Lojistik regresyon, Model, Veri madenciliği.

CUKUROVA UNIVERSITY

INSTITUTE OF NATURAL AND APPLIED SCIENCES

PhD THESIS

Comparative Examination of Outlier Detection Methods in Binary Logistics Regression Analysis

Melis ÇELİK GÜNEY

Advisor: Prof. Dr. G. Tamer KAYAALP

Department of Animal Science

ABSTRACT

In this thesis, statistical based, proximity based, cluster based, deviation based and isolation based outlier detection methods were compared. The data set was simulated to have the standard normal distribution in R. The sample size of the data set was selected as 3000. The data set suitable for binary logistic regression with 3 independent and 1 dependent variable were produced. In order to compare the methods, the data was modified by adding 30 outliers to the data set. As a result of the study, the iForest algorithm from isolation based outlier detection methods has found all the added outliers and has the highest performance. In a real data set, outliers were determined using the iForest algorithm. In addition, the model was estimated in both artificial and real data sets for all observations and without outliers. The goodness of fit and adequacy of the model were examined. Thus, the effect of outliers on the model was determined.

Keywords: Outlier, Logistic regression, Model, Data mining.

TEŞEKKÜR

Tez konumun belirlenmesi, yürütülmesi ve yazımı aşamalarındaki yardımlarından dolayı danışman hocam Sayın Prof. Dr. G. Tamer KAYAALP'e, Sayın Prof. Dr. Zeynel CEBECİ'ye, Sayın Prof. Dr. Gülsen KIRAL'a ve emeği geçen diğer hocalarıma sonsuz teşekkürü bir borç bilirim.

Bugün bulunduğum noktada olmamı sağlayan, beni yetiştiren; babam Sayın Okay ÇELİK ve annem Sayın Nurten ÇELİK'e şükran ve minnetlerimi sunarım. Ayrıca bütün başarılarımda beni destekleyen ve daima yanımda olan sevgili eşim Hakan GÜNEY'e ve çocuklarım Kumsal ve Kutay GÜNEY'e teşekkürlerimi sunarım.

ÇİZELGELER DİZİNİ

Çizelge 3.1. Yanılgı matrisi 21
Çizelge 4.1. Üretilen veri setinin içeriği 31
Çizelge 4.2. Tanımlayıcı istatistikler 31
Çizelge 4.3. Y değişkeni seviyelerinin sayısı 32
Çizelge 4.4. Aykırı değer içeren yapay veri setinin değişkenleri ve ilgili test istatistikleri 33
Çizelge 4.5. Değişkenlerin tahmin değerleri ve güven aralıklarının odds oranları 33
Çizelge 4.6. Aykırı değer içeren yapay veri setinin yanılgı matrisi 34
Çizelge 4.7. Aykırı değer içeren yapay veri setinin Sahte R^2 değerleri 35
Çizelge 4.8. Bazı gözlemlerin Student türü hataları, Şapka değerleri ve Cook uzaklıkları 37
Çizelge 4.9. Mahalanobis uzaklıkları 38
Çizelge 4.10. En yakın komşu sayısına göre LOF değerleri 39
Çizelge 4.11. En yakın komşu sayısına göre aykırı değerler 39
Çizelge 4.12. Oluşan iki kümenin merkezleri 40
Çizelge 4.13. Sıralanmış ABOF değerleri 41
Çizelge 4.14. iForest yöntemi ile bulunan aykırı değer puanları 42
Çizelge 4.15. Aykırı değer içermeyen yapay veri setinin değişkenleri ve ilgili test istatistikleri 44
Çizelge 4.16. Aykırı değer içermeyen yapay veri setinin yanılgı matrisi 45
Çizelge 4.17. Aykırı değer içermeyen yapay veri setinin Sahte R^2 değerleri 46
Çizelge 4.18. Yapay veri setinde değişkenler ve ilgili test istatistiklerinin karşılaştırılması 47
Çizelge 4.19. Yapay veri setinde Sahte R^2 değerlerinin karşılaştırılması 48
Çizelge 4.20. Yapay veri setinde sınıflama performanslarının karşılaştırılması 48
Çizelge 4.21. Gerçek veri setinin içeriği 49
Çizelge 4.22. Gerçek veri setinde iForest yöntemi ile bulunan aykırı değer puanları 50
Çizelge 4.23. Gerçek veri setinde değişkenler ve ilgili test istatistiklerinin karşılaştırılması 51
Çizelge 4.24. Gerçek veri setinde Sahte R^2 değerlerinin karşılaştırılması 51
Çizelge 4.25. Gerçek veri setinde sınıflama performanslarının karşılaştırılması 52

ŞEKİLLER DİZİNİ

Şekil 3.1. Lojistik regresyon eğrisi 19
Şekil 4.1. Verinin yapısı 32
Şekil 4.2. Aykırı değer içeren yapay veri setinin ROC eğrisi 35
Şekil 4.3. Student türü hatalar 35
Şekil 4.4. Pearson hatalar grafiği 36
Şekil 4.5. Cook uzaklıkları ve Şapka değerleri grafiği 36
Şekil 4.6. Student türü hatalar ve Şapka değerleri grafiği 37
Şekil 4.7. Mahalanobis uzaklığı ile bulunan aykırı değerler grafiği 38
Şekil 4.8. LOF algoritmasıyla bulunan aykırı değerler grafiği 40
Şekil 4.9. K-ortalamalar algoritmasıyla bulunan aykırı değerler grafiği 41
Şekil 4.10. Sıralanmış ABOF değerleri grafiği 42
Şekil 4.11. iForest algoritmasıyla bulunan aykırı değerler grafiği 43
Şekil 4.12. Aykırı değer belirleme yöntemlerinin ROC eğrisi ile karşılaştırılması 44
Şekil 4.13. Aykırı değer içermeyen yapay veri setinin ROC eğrisi 46
Şekil 4.14. Yapay veri setinde ROC eğrilerinin karşılaştırılması 49
Şekil 4.15. iForest algoritması ile bulunan aykırı değerler 50
Şekil 4.16. Gerçek veri setinde ROC eğrilerinin karşılaştırılması 52

SİMGELER VE KISALTMALAR

$\hat{\beta}$: Tahmin edilen regresyon katsayısı

$S(\hat{\beta}_j)$: Regresyon katsayısının en çok olabilirlik tahmininin standart hatası

p : Önem seviyesi

OR : Odds oranı

χ^2 : Ki-kare istatistiği

EKK : En küçük kareler

ML : En çok olabilirlik

RILS : Yeniden ağırlıklandırılmış en küçük kareler

MLCS : Minimum lojit ki-kare

$Z(X_i)$: Olabilirlik oran fonksiyonu

$\pi(X)$: X biliniyorken Y'nin koşullu olasılığı

$g(X)$: X biliniyorken Y'nin koşullu olasılığının lojit dönüşümü

W : Wald istatistiği

ROC : Alıcı işletim karakteristiği eğrisi

AUROC : Alıcı işletim karakteristiği eğrisinin altındaki alan

R^2 : Belirtme katsayısı

Pseudo R^2 : Sahte belirtme katsayısı

R^2_{McF} : McFadden R^2

R^2_{CS} : Cox-Snell R^2

R^2_{NAG} : Nagelkerke R^2

e_i : i. ham hata

r_{pe} : Pearson hata

$r_{std_{pe}}$: Standartlaştırılmış Pearson hata

r_i : i. içsel student türü hata

r_i^* : i. dışsal student türü hata

H : Şapka matrisi

h_{ii} : Şapka matrisinin i. köşegen elemanı

CU_i : i. gözlemin Cook uzaklığı

M_i : i. gözlemin Mahalanobis uzaklığı

ABOD : Açı tabanlı aykırı değer belirleme yöntemi

LOF : Yerel aykırı değer faktörü

k : En yakın komşu sayısı

ABOF : Açı tabanlı aykırı değer faktörü

iForest : Yalıtım orman yöntemi

1. GİRİŞ

Yapay zeka, bilgi, algı, muhakeme, öğrenme, anlama ve benzeri bilişsel yetenekleri bilgisayarlarla yerine getirebilme kabiliyetidir (Saloky ve Šeminský, 2005). Bilgisayarların verilere dayalı tahminleri veya davranışları yapmak ve geliştirmek için kullandığı bir dizi yönteme ise makine öğrenimi denmektedir (Molnar, 2019). Makine öğrenimi, sisteme otomatik veya deneyimleyerek öğrenme kabiliyeti sağlayan bir yapay zeka uygulamasıdır. Makine öğreniminde birçok algoritma kullanılmaktadır. Lojistik regresyon, doğrusal regresyondan sonra en ünlü makine öğrenme algoritmasıdır. Değerleri tahmin etmek için doğrusal regresyon algoritmaları kullanılırken; sınıflandırma yapmak için lojistik regresyon algoritmaları kullanılmaktadır. Sınıflandırma yapmak, günlük hayatımızın kaçınılmaz bir eylemidir. Örneğin, bir e-postanın spam olup olmadığını sınıflandırmak, bir tümörün kötü huylu mu yoksa iyi huylu mu olduğunu sınıflandırmak, bir web sitesinin hileli olup olmadığını sınıflandırmak (Gandhi, 2018), bir sürüde bulunan hayvanların gebe olup olmadığını sınıflandırmak, bir bitkiyi veya hayvanı hasta veya hasta değil olarak sınıflandırmak. Bu gibi sınıflandırmalarda lojistik regresyon algoritması kullanmak, basit, geleneksel ve kullanışlıdır.

Basit veya çoklu doğrusal regresyon analizi, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi incelemek için kullanılan bir yöntemdir. Ancak bu analizlerde bağımlı değişken sürekli verilerden oluşmaktadır. Bağımlı değişken kategorik veya sıralama ölçekli ise lojistik regresyon analizine ihtiyaç duyulmaktadır. Lojistik regresyon analizinde bağımsız değişkenler kesikli, sürekli ve bunların karışımından olabilmektedir. Lojistik regresyon analizinde bağımlı değişken iki kategorili ise (var/yok, doğru/yanlış, evet/hayır gibi) iki kategorili (binary) lojistik regresyon olarak adlandırılmaktadır. Bu analizde, doğrusallık, hata dağılışının normalliği gibi varsayımların sağlanmasına gerek duyulmaması bu analizin sıkça tercih sebebidir.

Lojistik regresyon analizinin amacı, çok kategorili veya sıralama ölçeğinde olan bağımlı değişkenin değerini tahmin etmek, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi araştırmak ve sınıflandırma yapmaktır (Mertler ve Vannatta, 2005; Çokluk, 2010).

Lojistik regresyon analize uygun olan verilerde anormal gözlemler olabilmektedir. Bu gözlemlere aykırı değer denmektedir. Bu değerler, kayıtların hatalı girilmesi, hileli davranışlar, insanın veya aletin yanlış ölçümü, popülasyondaki doğal bir sapma nedeniyle olabilmektedir (Hodge ve Austin, 2004). Aykırı değer saptama yöntemleri, tek değişkenli ve çok değişkenli veri setinde aykırı değer saptama yöntemleri olarak ikiye ayrılmaktadır (Bedrick ve Hill, 1990). Bunun dışında parametrik, yarı parametrik ve parametrik olmayan yöntemler veya eğitimli, yarı eğitimli ve eğitimsiz yöntemler olarak da sınıflandırılabilmektedir. Aykırı değerler, hatanın varyansını arttırıp, testin gücünü zayıflatmaktadır. I. ve II. tip hata yapma ihtimallerini artırabilmektedir. Tahmin değerlerinde önemli olabilecek değişikliklere neden olabilmektedir (Osborne ve Amy,

2004). Aykırı değerlerin belirlenmesinin amacı, modelin yeterliliğinin, varsayımlardan sapmalarının, verinin düzeltilip düzeltilmeyeceğinin belirlenmesidir (Aydın, 2006).

Veri madenciliği, büyük hacimli veri yığınlarından geçerli, anlamlı ve kullanışlı bilgileri çıkarmak için kullanılan birçok metot ve tekniklerin tümüdür (Yıldız, 2018). Veri madenciliğinde aykırı değerler birçok alanda karşımıza çıkmaktadır. Örneğin, kredi kartı sahtekarlığı, cep telefonlarının hileli kullanımı, bilgisayar ağlarına izinsiz giriş, tıbbi teşhis, anormal iklim, boru hatlarında anormal akış sorunu, düşman hareketleri, hayvanların davranışlarında değişiklik, tarımda sel, kuraklık ve ticaret savaşları gibi (Chandola ve ark., 2009; Gogoi ve ark., 2011; Yücel Altay, 2014; Akdemir, 2016; Gurrapu ve ark., 2021). Kullanım alanının bu kadar geniş olması, veri madenciliğinde aykırı değerler belirleme yöntemlerine olan ilgiyi arttırmıştır.

Bu çalışmada, R yazılımında standart normal dağılıma sahip rastgele bir veri seti üretilmiştir. Veri setinin örnek büyüklüğü 3000 olarak seçilmiştir. Üretilen veri setinde Bernoulli dağılımına uyan bir bağımlı değişken bulunmaktadır. Üretilen veri setindeki bağımsız değişkenlere 30 adet aykırı değer eklenerek veri, modifiye edilmiştir. Modifiye edilen veri setinde lojistik regresyon analizi ile model tahmin edilmiştir. Tahmin edilen modelin uyumunun iyiliği ve yeterlilik ölçütleri incelenmiştir. Ardından özellikle veri madenciliğinde kullanılan aykırı değer belirleme yöntemlerinden istatistik tabanlı, yakınlık tabanlı, kümeleme tabanlı, sapma tabanlı ve yalıtım tabanlı aykırı değer belirleme yöntemleri ile eklenen aykırı değerler bulunarak yöntemlerin performansları karşılaştırılmıştır. Aykırı değerler, veri setinden çıkartıldıktan sonra tekrardan lojistik regresyon analizi ile model tahmin edilmiştir. Tahmin edilen modelin uyum iyiliği ve yeterlilik ölçütleri incelenmiştir. Ayrıca çalışmada Zhao ve ark. (2011) tarafından elde edilen gerçek bir veri seti kullanılmıştır. Bu veri setinde yaprak tüylenmeye (tüylenme yok:0, tüylenme var:1) etki eden faktörleri belirlemek amacıyla lojistik regresyon analizi kullanılmıştır. Yapay veri setinde bulunan en iyi algoritmayla gerçek veri setinde aykırı değerler bulunmuştur. Belirlenen aykırı değerler veri setinden çıkartıldıktan sonra gerçek veri setinde model tahmin edilip modelin uyum iyiliği ve yeterlilik ölçütleri incelenmiştir.

Bu çalışmanın amacı, veri madenciliğinde kullanılan aykırı değer belirleme yöntemlerinin karşılaştırmalı olarak incelenmesi, aykırı değerlerin modele etkisinin saptanmaya çalışılmasıdır.

2. ÖNCEKİ ÇALIŞMALAR

2.1. Lojistik Regresyon Analizi ile İlgili Çalışmalar

Halperin ve ark. (1971), çok değişkenli lojistik regresyon modelinde hastalık riskini tahmin etmek için bir çalışma yapmışlardır. Tahmin yaparken ayırma fonksiyonu ve maksimum olabilirlik yaklaşımını karşılaştırmışlardır. Çalışmanın sonucunda, bağımsız değişkenler nitel olduğunda maksimum olabilirlik yönteminin tercih edilmesi gerektiğini vurgulamışlardır.

Montgomery ve ark. (1987), süt sığırcılığında mastisis hastalığını tahmin etmek için ayırma analizi ve lojistik regresyon analizini karşılaştırmışlardır. Bu analizde, ineğin yaşı, sıcaklık, dakikada nabız, dakikada solunum hızı, memede apse, iştah kaybı, süt rengi berraklığı, süt pıhtıları, kulak sıcaklığı, depresyon, mastisisten sonra muayene, susuzluk, meme sıkılığı, cilt sıcaklığı, süt rengi (beyaz ve açıktan başka), ineğin zayıflığı, şişme, sütün sulu tutarlılığı ve rumen hareketliliği değişkenlerini kullanmışlardır. Çalışmanın sonucunda, her iki analizin doğruluk değerinin yüksek olduğunu bulmuşlardır.

Gardside ve Glueck (1995), kalp hastalığına etki eden faktörleri belirlemek amacıyla anket çalışmasından elde edilen verilere lojistik regresyon analizi uygulamışlardır. Çalışmanın sonucunda, alkol alımı, boş zaman egzersizi, fiziksel aktivite, cinsiyet olarak kadın, magnezyum serum alımı, demir ve B_2 vitaminin diyeti değişkenleri, kalp hastalığı sebebiyle ölüm ve hastaneye yatış ile ters orantılı ve istatistiksel olarak önemli; vücut ağırlığı, yaş, alyuvar çökelme hızı, vücut kitle endeksi ve sigara kullanımı değişkenleri ile doğru orantılı ve istatistiksel olarak önemli bulunmuştur.

O'Neill ve Barry (1995), trafik kazalarında ölüme sebebiyet veren faktörleri belirlemek amacıyla hem koşullu lojistik regresyon analizini hem de kesilmiş lojistik analizini kullanmışlardır. Her iki analizin teorisi verilmiş, trafik kazaları verisinde uygulaması yapılmış ve analizler karşılaştırılmıştır. Çalışmanın sonucunda, kesilmiş lojistik regresyon ile kaza seviyesi etkilerini tahmin etmede koşullu lojistik regresyon analizden daha verimli olduğu belirlenmiştir.

Angen ve ark. (1996), lojistik regresyon analizi kullanarak broiler yetiştiriciliğinde Salmonella hastalığına etki eden risk faktörlerini belirlemişlerdir. Bu risk faktörleri olarak kullanılan değişkenler mevsim, kuluçkahane ve yem fabrikasının üretim kümeslerine uzaklığı, önceki sürünün hastalığı geçirip geçirmemesi, farklı yaşlarda hayvanların aynı kümeste olup olmamasıdır. Çalışmanın sonucunda, Salmonella hastalığının oluşmasında en çok kuluçkahaneler ve yem fabrikalarının etkili olduğu bulunmuştur.

Peduzzi ve ark. (1996), lojistik regresyonda değişken başına olay sayısının etkisini incelemek için bir Monte Carlo simülasyonu yapmışlardır. 673 hastadan 500 hasta rastgele seçilmiştir ve değişken başına olay sayısını farklı alarak 500 örneğin her grubundaki her değişkenin sonuçları karşılaştırılmıştır. Çalışmanın sonucunda, değişken başına olay sayısının yetersizliğinin, lojistik modelin geçerliliğinde büyük sorunlara yol açabileceğini belirlemişlerdir.

Kinsel ve ark. (1998), 260 süt sığırının ikizliğe neden olan risk faktörlerini lojistik regresyon analizi ile incelemişlerdir. Araştırmanın sonunda, ikizliğe neden olan en önemli faktörün süt verimi olduğunu belirlemişlerdir.

Atakurt (1999), lojistik regresyon analizini tıp verilerine uygulamıştır. Koroner arter hastalığına etki eden 34 değişkeni analizde kullanmıştır. Çalışmanın sonucunda, modelde kalan değişkenler belirlenmiş, bu değişkenler yorumlanmış, koroner damar hastalığına sahip olanların duyarlılığı %93.82, özgüllüğü ise %80.56 olarak bulunmuştur.

Al-Ghamdi (2002), trafik kazalarının şiddetini etkileyen faktörleri belirlemek için lojistik regresyon analizini kullanmıştır. Çalışmanın sonucunda, kavşak olmayan bir yerde meydana gelen kazalar, kavşaktaki kazalara göre 2,64 kat daha yüksek olduğu belirlenmiştir. Ayrıca Riyad'daki trafik kazalarındaki güvenliğin sağlanması için önerilerde bulunmuştur.

Compton ve ark. (2002), lojistik regresyon analizini kullanarak Wood kaplumbağaların faaliyet alanları içindeki habitat seçimini modellemişlerdir. Çalışmanın sonucunda, Wood kaplumbağaların düşük gölgelik örtüsüyle kaplı, suya yakın ormanların olduğu yerleri seçtikleri belirlenmiştir. Ayrıca vücut ısısını dengelemek ve beslenme ihtiyaçlarını karşılamak için orman kenarlarını seçtikleri görülmüştür. Oluşturulan modelin doğruluğu %84 bulunmuştur.

Önder ve Cebeci (2002), sütten kesim ağırlığına etki eden faktörleri belirlemek amacıyla lojistik regresyon analizini kullanmışlardır. Bu faktörlerden, doğum ağırlığı, ırk, cinsiyet ve doğum tipi kayıtları ele alınmıştır. Çalışmada, değişken seçimi yöntemlerinden ileri doğru seçim, geriye doğru eleme, adımsal seçim ve en iyi alt kümeler yöntemleri karşılaştırılmıştır. Çalışmanın sonucunda, değişken seçimi yöntemleri arasında bir üstünlük olmadığı sonucuna varılmıştır.

Ohlmacher ve Davis (2003), Kansas'ın Kuzeydoğusu'nda heyelan tehlikesini öngörmek için çoklu lojistik regresyon analizini kullanmışlardır. Heyelan tehlikesini tahmin etmede kullanılan değişkenleri belirlemeyi ve yerel yamaçlar ve jeolojik birimler göz önüne alındığında, gelecekteki heyelan olasılığını gösteren bir harita üretmeyi amaçlamışlardır. Çalışmanın sonucunda, çalışılan alanda heyelan tehlikesinin tahmininde eğimin en önemli değişken olduğunu belirlemişlerdir.

Coşkun ve Kartal (2004), yaptıkları çalışmada protez kullanımını etkileyen risk faktörlerini belirlemek için lojistik regresyon analizini kullanmışlardır. Sonuç modelinde, ağızdan solunum yapma, eğitim seviyesi, sigara kullanımı ve yaş değişkenlerinin bulunduğunu belirtmişlerdir.

Bircan (2004), çocuklarda doğum ağırlığına etki eden faktörleri belirlemek amacıyla lojistik regresyon analizini kullanmıştır. Çalışmanın sonucunda, cinsiyet, gebelik haftası, beslenme durumu ve anne boyunun düşük doğum ağırlığına sebep olduğu belirlenmiştir.

Bayram (2004), lojistik regresyon analizini teorik olarak açıklamış, daha sonra ücretsiz çalışan ile gelir karşılığı çalışan kişileri bağımlı değişken olarak alıp cinsiyet, medeni durum, eğitim durumu, yaş gibi çeşitli değişkenleri bağımsız değişken alarak lojistik regresyon analizini kullanmıştır. Çalışmanın sonucunda, tüm bağımsız değişkenlerin önemli olduğu bulunmuştur.

Oğuzlar (2005), Bursa Emniyet Müdürlüğü ile yaptığı çalışmada, ahlak, yan kesicilik ve narkotik bürolarındaki verileri kullanarak suçluların profilini lojistik regresyon analizi ile belirlemiştir. Çalışmanın sonucunda, ahlak ve yan kesicilik bürosuna ait olay saati, cinsiyet, yaş, doğum yeri, öğrenim durumu ve meslek değişkenlerinin önemli; narkotik bürosunda ise cinsiyet hariç diğer değişkenlerin önemli olduğu sonucuna varılmıştır.

İyit ve Genç (2005), kadınların menopoz evresine geçişte fiziksel olarak hangi özelliklere sahip olduklarını belirlemek amacıyla lojistik regresyon analizini kullanmışlardır. Konya Bölgesi'nde yaşayan, 115 kadından alınan verilerle model kurulmuş ve bu verileri %93 oranla başarılı bir şekilde sınıflandırmışlardır.

Ürük (2007), yaptığı tez çalışmasında yoğun bakım ünitesindeki hastaların yaşamları ile ilgili bir veri setinde lojistik regresyon analizinde kullanılan üç farklı değişken seçimi yöntemi ile üç farklı model oluşturmuştur. Çalışmanın sonucunda, adımsal seçim yöntemi ile bulunan modelin en iyi model oluştuğuna karar vermiştir.

Kaşko (2007), FORTRAN programı ile farklı örnek büyüklüklerine sahip bağımlı ve bağımsız değişkenler içeren veri seti üretmiştir. Çalışmanın sonucunda, bağımsız değişkenler arasında yüksek korelasyon olarak tanımlanan çoklu bağlantı varlığının, tüm örnek genişliklerinde I. tip hata olasılığını etkilemezken, testin gücünü önemli ölçüde düşürdüğünü gözlemlemiştir.

Urgan ve Tez (2008), lojistik regresyon analizinde çoklu bağlantı sorunu olması durumunda en çok olabilirlik yöntemine alternatif olan Stein, Liu ve Ridge tahmin edicilerini önermişlerdir. Çalışmanın sonuncunda, bu tahmin edicilerin varyansları, en çok olabilirlik yöneminin varyansından daha küçük olduğu için çoklu bağlantı sorunundan daha az etkilendiği sonucuna ulaşmışlardır.

Girginer ve Cankuş (2008), Eskişehir'deki tramvayda yolcu memnuniyetini belirlemek için iki kategorili (binary) lojistik regresyon analizini kullanmışlardır. Kurulan model sonrasında bağımsız değişkenlere ait regresyon katsayılarının hepsinin negatif olduğu görülmüştür. Bu durum bu değişkenlerin memnuniyeti olumsuz yönde etkilediğini göstermektedir.

Akçay (2009), yaptığı çalışmada Türkiye'deki Broiler yetiştiriciliğinde Coccidiosis hastalığının ortaya çıkmasında etkili olan faktörleri lojistik regresyon analizi ile belirlemiştir. Bunun için çeşitli illerden 1200 kümes seçerek anket çalışması yapmıştır. Çalışmanın sonucunda, Coccidiosis hastalığı için önemli risk faktörleri; mevsim, kümes içi hayvan durumu, başka bir hastalığın olması, havalandırma, yaş, çiftlikteki kümes sayısı, çatı izolasyonu, altlık olup olmaması, altlığın güvenli bir uzaklığa atılması, iki yetiştirme dönemi arası süre, kümeste giriş odasının varlığı ve kümeste kemirgen varlığı olarak belirlenmiştir.

Cankurt ve ark. (2010), İzmir ilinde sığır eti tercihlerini etkileyen faktörlerin belirlenmesi amacıyla anket çalışması yapmışlardır. Sığır eti tüketmede; cinsiyet, yaş, eğitim, gelir, hanedeki kişi sayısı ve hanede hasta varlığının etkilediğini lojistik regresyon analizi ile belirlemişlerdir.

Nurunnabi ve ark. (2010), yaptıkları çalışmada lojistik regresyon analizinde çoklu etkili gözlemlerin tespiti için Genelleştirilmiş DFFITS ölçüsünü önermişlerdir. Bu çalışma, Genelleştirilmiş DFFITS ölçüsünün, yaygın olarak kullanılan yöntemlerde birkaç kez başarısız olduğunda çoklu etkili gözlemleri başarılı bir şekilde tanımlayabildiğini birkaç veri seti üzerinde göstermişlerdir.

Karagöz ve ark. (2010), yaptıkları çalışmada kriz ortamında iş tatminini etkileyen faktörlerin etkisini belirlemek amacıyla lojistik regresyon analizini kullanmışlardır. Çalışmanın sonucunda, iş tatminini etkilediği düşünülen faktörlerin çoğunluğunun, kriz ortamında etkisini kaybettiğini belirlemişlerdir.

Çokluk (2010), lojistik regresyon analizini eğitim bilimleri verilerine uygulamıştır. Bağımlı değişken olarak eleştirisel düşünme düzeyini, bağımsız değişkenler olarak da bilimsel düşünme becerisi ölçeği, epistomolojik inanç ölçeği ve fatalizm ölçeğini kullanmıştır. Çalışmanın sonucunda, bilimsel düşünme becerisi ve epistemolojik inanç ölçeklerinin yüksek eleştirisel düşünmeyi etkilediğini belirlemiştir.

Pereira ve ark. (2011), Broiler hayvan yetiştiriciliğinde hayvanların refah düzeyinin belirlenmesi amacıyla lojistik regresyon analizini kullanmışlardır. Çalışmanın sonucunda, hayvanların refah düzeyinin belirlenmesinde; hava sıcaklığı, amonyak konsantrasyonu, oturma sıklığı ve içme suyu sıklığı değişkenlerinin etkili olduğu belirlenmiştir.

Bozkurt (2011), Kredi ve Yurtlar Kurumu'nda kalan öğrencilerin memnuniyetini belirlemek amacıyla lojistik regresyon analizini kullanmıştır. Analizin sonucunda, öğrencilerin yurttan memnun olduğunu belirlemiş ve memnuniyetin arttırılabilmesi için neler yapılabileceği hakkında önerilerde bulunmuştur.

Aksaraylı ve Saygın (2011), yaptıkları çalışmada Buca Kız Öğrenci Yurdu'nda kalan öğrencilerin hizmet kalitesini algılamalarını ve bu algının hizmet tercihine etkilerini lojistik regresyon analizi ile araştırmışlardır. Çalışmanın sonucunda, algılanan hizmet kalitesinin, hizmet tercihine ve tavsiye etmede etkisi olduğunu belirlemişlerdir.

Akın ve Şentürk (2012), sıralı lojistik regresyon analizi ile bireylerin mutluluk düzeylerini incelemişlerdir. Çalışmada, cinsiyet, yaş, eğitim durumu, gelir gibi sosyo-demografik özellikleri içeren değişkenler ele alınmıştır. Çalışmanın sonucunda, evli ve erkeklerin daha mutlu olduğu belirlenmiştir.

Bahadır ve Özdemir (2013), İlköğretim Matematik Öğretmenliği'nde okuyan lisans öğrencilerinin ders başarısının, lisansüstü eğitim başarısına etkisini belirlemek amacıyla lojistik regresyon analizini kullanmışlardır. Öğrencilerin lisansüstü eğitim başarıları üzerinde Özel Öğretim Yöntemleri 2, Elementer Sayı Kuramı ve Cebire Giriş derslerinin etkisi olduğu bulunmuştur. Ayrıca kurulan modelin doğruluk oranı %92 olarak belirlenmiştir.

Özyurt (2013), Ceylanpınar Tarım İşletmesi'nde yetiştirilen Siyah Alaca süt sığırlarında doğum sonrası dönemde gelişen mastitis hastalığının üreme performansı üzerine etkilerini lojistik

regresyon analizi ile araştırmıştır. Çalışmada, doğum sonrası ilk 60 gün içerisinde gelişen klinik mastitis olgularının, doğum-ilk tohumlama arası süre ve gebelik başına tohumlama sayısı üzerine etkili olduğunu belirlemiştir.

Oktay ve ark. (2013), yaptıkları anket çalışması ile Atatürk Üniversitesi'ndeki lisans öğrencilerinin internet ve cep telefonu kullanımında etkili olan faktörleri belirlemişlerdir. İnternet kullanımı ile ilgili kurulan modelde, yaş ve anne eğitim düzeyi; cep telefonu ile ilgili kurulan modelde, ailedeki birey sayısı, anne eğitim düzeyi ve harç kredisi alıp almadığı değişkenlerinin etkili olduğu lojistik regresyon analizi ile belirlenmiştir.

Budak ve ark. (2013), yaptıkları çalışmada bilgisayar ağlarındaki düzensizlikleri tespit etmek için lojistik regresyon analizini kullanmışlardır. Saldırı tespit sistemlerinde kullanılan veri setinde model oluşturmuşlardır. Bu model ile yeni bir yazılım üretilip gerçek hayatta kullanılabileceği sonucuna varmışlardır.

Akçay ve ark. (2014), 126 baş Siyah Alaca sığırında gebelik oranını etkileyen faktörleri belirlemek amacıyla lojistik regresyon analizini kullanmışlardır. Çalışma sonucunda, senkronizasyon yöntemi, tohumlama sayısı, tohumlama anındaki süt verimi, vücut kondisyon skoru değişkenlerinin gebelik oranı üzerine etkili faktörler olduğunu belirtmişlerdir.

Kocabaş (2014), yaptığı çalışmada lojistik regresyon analizini banka sektöründe kullanmıştır. Müşterilerin kredi ödeme performansını etkileyen faktörleri belirlemek amacıyla bu çalışmayı yürütmüştür. Çalışmanın sonucunda, hesap dengesi değişkeninin kredi ödeme performansını en çok etkilediğini belirtmiştir.

Bejaei ve ark. (2015), yaptıkları anket çalışmasıyla, yumurta tüketicilerinin özelliklerine göre tüketicinin hangi tür yumurtayı alacağını lojistik regresyon analizi ile tahmin etmeye çalışmışlardır. Çalışmanın sonunda, normal yumurta tüketicileri daha az eğitimli, daha yaşlı, fiyatlara daha duyarlı, büyük mağaza zinciri alıcıları, farklı yumurta türleri hakkında daha az farkındalık duyan ve hayvan refahında daha az endişe duyan kişiler olduklarını belirtmişlerdir. Kafessiz yumurta tüketicileri ise hayvanın refahı, yumurtayı satın aldıkları yer, farklı yumurta türleri hakkında daha çok endişe duyan, fiyat konusunda ise daha az endişe duyan kişiler oldukları belirlenmiştir.

Demir ve ark. (2016), Gümüşhane ilinde konvansiyonel hayvancılık yapan üreticilere anket çalışması yapmışlardır. Anket verileri ile lojit model oluşturmuşlar ve üreticilerin organik hayvancılık yapma kararlarını etkileyen faktörleri belirlemişlerdir. Çalışmanın sonucunda, hayvancılık desteklerinden faydalanma oranı ve yem ihtiyacının işletmelerden karşılama oranı yüksek, fabrika yemi kullanma oranı düşük olan işletmelerde organik hayvancılığa geçiş olduğunu belirlemişlerdir.

2.2. Aykırı Değerler ile İlgili Çalışmalar

Chatterjee ve Hadi (1986), lineer regresyonda etkili gözlemleri (Cook uzaklığı, Welsch-Kuh uzaklığı, Welsch'in uzaklığı, Modifiye edilmiş Cook uzaklığı), kaldıraç noktalarını ve aykırı değerleri belirleme yöntemleri arasındaki karşılıklı ilişkileri tanımlamışlardır.

Rousseeuw ve Van Zomeren (1990), Sağlam LMS (Least Median of Squares) tahmin edicisinin aykırı değerleri bulmak için kullanılabileceğini önermişlerdir. Standartlaştırılmış LMS hatalar bulunarak Sağlam uzaklık grafiği ile aykırı değerlerin kolayca görülebileceğini belirtmişlerdir.

Hadi ve Simonoff (1993), yaptıkları çalışmada doğrusal modellerde birden fazla aykırı değer belirleme yöntemi geliştirmişlerdir. Bu yöntemde iki tür prosedür bulunduğunu ve bu iki prosedürün performanslarını bir Monte Carlo çalışması ile belirlemişlerdir.

Doğan (1995), Harran Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Hayvancılık İşletmesi'nde bulunan bıldırcınların yumurta ağırlık ve yumurta boy ölçüm verilerinde aykırı değerleri tespit etmek için Grafik yöntemi, Cook uzaklığı ve Katkı kareler toplamı testini kullanmıştır.

Kaya (1995), kimyasal bir üretim sürecinden alınan konsantrasyon ölçümleri ile çimento fabrikasından alınan aylık toplam ihracat rakamlarını kullanarak zaman serilerinde aykırı değerleri belirlemiştir.

Penny (1996), çok değişkenli veri setinde aykırı değer belirleme yöntemlerinden biri olan Mahalanobis uzaklığını kullanmıştır. Bu yöntemde kritik değerlerin uygunluğu konusundaki tutarsızlıklar sebebi ile çalışmada uygun ve uygun olmayan kritik değerleri klinik laboratuvar veri seti ile belirlemiştir.

Ahn ve James (1999), Güney Florida'da, atmosferde biriken fosfor miktarı ölçümündeki sorunu belirlemek amacıyla bir çalışma yapmışlardır. Çalışmadaki aykırı değerleri belirlerken Hadi ve Simonoff (1993)'un önerdiği yöntemi kullanmışlardır.

Breunig ve ark. (2000), aykırı değer belirleme yöntemi olarak yoğunluk tabanlı LOF algoritmasını geliştirmişlerdir. Bu algoritma ile anlamlı gibi görünen ama mevcut yaklaşımlarla tanımlanamayan aykırı değerleri bulmak için kullanılabileceğini göstermişlerdir.

Chen (2002), SAS/STAT'ın ROBUSTREG prosedürünü tanıtmış, Sağlam LTS (Least Trimmed Squares), M, S ve MM tahmin edicileri ile regresyon modeli kurup aykırı değerleri incelemiştir. İlgili komutları da çalışmasında belirtmiştir.

He ve ark. (2003), aykırı değer belirleme yöntemi olarak kümeleme tabanlı CBLOF algoritmasını geliştirmişlerdir. Bu yöntemin aykırı değer bulmada etkili olduğunu ispatlamışlardır.

Osborne ve Amy (2004), yaptıkları çalışmada aykırı değerlerin çeşitli nedenlerini özetleyip, aykırı değerlerin nasıl bulunacağını belirleyip, veri setinden kaldırılıp kaldırılmayacağına nasıl karar verileceğini araştırmışlardır. Çalışma, aykırı değerlerin küçük bir bölümünün basit analizleri bile ne kadar etkileyebileceğini göstermiştir.

Zioutas ve ark. (2005), aykırı değerleri tespit etmek PTS adında yeni bir yöntem geliştirmişlerdir. Simülasyon çalışmasıyla, PTS ve IPTS yöntemleri performansının LTS, S1S, MM yöntemlerinden daha üstün olduğunu bildirmişlerdir.

Kıral ve Billor (2005), yaptıkları çalışmada çok değişkenli ve çok büyük veri setinde aykırı değeri belirlemede kullanılabilecek olan Sağlam temel bileşenler analizine dayalı BACON yönteminin diğer yöntemlere alternatif olacağını iki örnek üzerinde göstermişlerdir.

Kontrimas ve Verikas (2006), aykırı değerleri belirlemek için kullanılan yöntemleri; temel bileşenler analizine dayalı teknikler, Sağlam regresyon yöntemleri, sınıflama yöntemlerine dayalı teknikler, hataların grafiksel analizi ve etkilerin ölçümü şeklinde gruplandırmışlardır.

Aydın (2006), yaptığı tez çalışmasında nitrik asit yükseltgenmesinde açığa çıkan amonyağa etki eden faktörleri incelerken grafiksel gösterimlerle etkili gözlemleri ve aykırı değerleri belirlemiştir.

Martin ve Roberts (2006), student türü hatalar ile aykırı değer belirlemek için kritik noktanın oluşumunda Bootstrap yaklaşımını önermişlerdir. Çalışmanın sonucunda, bu yaklaşımın olası çalışmalarda hatalar normal dağılım göstermezse kullanılması gerektiğini belirtmişlerdir.

Ortiz ve ark. (2006), kimyasal analiz verilerinde aykırı değerleri belirlerken Sağlam M, GM ve LMS tahmin edicilerini kullanmışlardır. Çalışmanın sonucunda, LMS tahmin edicinin matematiksel özelliğinden dolayı kimyasal analiz verilerinde daha ilgi gösterileceğini belirtmişlerdir.

Blatná (2006), yaptığı çalışmada Sağlam LTS tahmin edicisi ile gerçek bir verinin çoklu regresyon denklemini tahmin etmiş, ardından Mahalanobis uzaklığı, Sağlam MCD uzaklığı ve standartlaştırılmış hatalar ile kaldıraç noktaları ve aykırı değerleri grafiksel olarak belirlemiştir.

Tekin (2007), yaptığı tez çalışmasında, TÜİK'den alınan aylık ortalama Cumhuriyet Altını satış fiyatları ve TCMB döviz rezervleri verilerini kullanarak zaman serilerinde aykırı değerleri tespit etmiştir.

Vural (2007), yaptığı çalışmada yaş-aylık gelir verilerinde kutu grafiği, histogram grafiği, hataların incelenmesi gibi klasik yöntemler ile aykırı değerleri belirlemiştir.

Ekici ve Yorulmaz (2008), aykırı değerleri Bayesian sürece dayalı belirleme yöntemleri ile tespit etmişlerdir. TÜİK'te bulunan 1992-2000 yıllarını kapsayan "Seçilmiş Ülke ve Fasıllara Göre Dış Ticaret" verileri alınmıştır. Çalışmada iki farklı Bayesian yaklaşımı kullanıp karşılaştırmasını yapmışlardır.

Billor ve Kıral (2008), regresyonda çoklu aykırı değer belirleme yöntemlerini (HS93, PY, SMR, BR, RWLS, LMS, M) Monte Carlo simülasyonu ile karşılaştırmışlardır. Çalışmanın sonucunda, hangi yöntemin hangi koşullarda diğerlerinden daha üstün olduğu konusunda önerilerde bulunmuşlardır.

Elahi ve ark. (2008), veri akışı var olduğunda kümeleme tabanlı aykırı değer belirleme yöntemlerine ek olarak yeni bir yöntem geliştirmişlerdir. Farklı veri setleri üzerinde yapılan

çalışmada, kullandıkları yöntemin düşük hesaplama maliyetine sahip olduğunu ve veri akışı olduğunda uzaklığa dayalı diğer yöntemlerden daha iyi aykırı değer bulabildiğini bildirmişlerdir.

Kriegel ve ark. (2008), aykırı değer belirlemek için sapma tabanlı ABOD algoritmasını geliştirmişlerdir. Bu algoritmayı uzaklık tabanlı LOF algoritması ile gerçek ve yapay veri setinde karşılaştırmışlardır. Karşılaştırma yaparken kesinlik, hatırlama değerlerini ve çalışma zamanını kriter olarak almıştır.

Liu ve ark. (2008), aykırı değerleri belirlemek için yalıtım orman yöntemini geliştirmiştir. Özellikle büyük veri setlerinde AUROC ve işlem süresi bakımından LOF algoritması gibi iyi sonuçlar verdiğini gözlemlemişlerdir.

Zhang ve ark. (2009), aykırı değer belirleme yöntemlerinden uzaklık tabanlı LDOF algoritmasını kullanmışlardır. Aykırı değerleri belirlerken bir top-n tekniği önermişlerdir. Kesiklik değeri ile bu yöntem ve geleneksel yaklaşımlar (top-n KNN ve top-n LOF gibi) karşılaştırıldığında, top-n LDOF yöntemi aykırı değerlerin belirlenmesinde daha etkili olduğunu belirlemişlerdir.

Tekbir (2009), yaptığı tez çalışmasında R-P-DBSCAN (Recursive-Partitioned DBSCAN) adlı yeni bir algoritma geliştirmiştir. Bölümleme ve birleştirme temeline dayalı olan bu yöntemde, kümeleme için DBSCAN algoritmasını kullanmıştır. Aykırı değer belirleme yöntemlerinden biri olan yoğunluk tabanlı DBSCAN'a göre R-P-DBSCAN performansının daha iyi olduğunu belirlemiştir.

Verardi ve Croux (2009), EKK ve Sağlam LAD, S, M ve MM tahmin edicilerini Stata programı ile incelemiş ve aykırı değerleri gösteren grafiksel bir araç sunmuşlardır.

Hekimoğlu ve ark. (2009), Sağlam LMS, LTS, M, Theil-Sen ve RM tahmin edicileri ile aykırı değerleri incelemişlerdir. Çalışmanın sonucunda, LMS ve LTS tahmin edicileri ile aykırı değerleri bulmanın daha iyi sonuçlar verdiğini bildirmişlerdir.

Coşkuntuncel (2009), ÖSS ile Orta Öğretim ders başarı puanlarındaki aykırı değerleri belirlerken EKK yöntemi ve Sağlam M tahmin edicisini kullanmıştır. Çalışmanın sonucunda, özellikle eğitim alanında yapılan çalışmalarda aykırı değer belirlerken Sağlam tahmin edicilerini önermiştir.

Yıldırım (2010), yem tüketim miktarına etki eden ham yağ, ham protein, asit deterjan fiber, nötr deterjan fiber tüketim miktarlarını ele alarak EKK, Ridge ve Sağlam M tahmin edicisi ile parametre tahmini yapmıştır. Aykırı değerleri belirlemeden ve aykırı değer belirleme yöntemleriyle (standartlaştırılmış hatalar, student türü hatalar, R-student türü hatalar, Ortalama değişim aykırı değer modellemesi, Bonferroni testi) aykırı değerleri tespit edip veri setinden çıkarttıktan sonra modeller karşılaştırılarak aykırı değerlerin modele etkisi saptanmıştır.

Üçkardeş ve ark. (2010), yaptıkları çalışmada tek değişkenli veri setlerinde tekli aykırı değer belirleme yöntemlerinden Dixon, Rosner, Discordance, Grubbs ve Walsh testlerini incelemişlerdir.

Koğar (2010), yaptığı çalışmada Türkiye'nin yedi Üniversitesi'nde birinci ve dördüncü sınıfta olup Sınıf Öğretmenliği okuyan öğrencilere anket uygulamıştır. Tek değişkenli veri setinde aykırı değer belirleme yöntemlerinden Z puana dönüştürme, Grafik yöntemi (kutu grafiği, normal olasılık grafiği, histogram grafiği); çok değişkenli veri setinde aykırı değer belirleme yöntemlerinden ise Mahalanobis uzaklığı, Cook uzaklığı ve kaldıraç noktasını kullanmıştır.

Turgut (2010), çok değişkenli veri setinde çoklu aykırı değerlerin belirlenmesinde MVE, MCD, FMCD ve MVV yöntemlerini tanıtıp aykırı değerleri doğru tespit etme oranları ve sonuca ulaşma hızları açısından karşılaştırmasını yapmıştır. Çalışmanın sonucunda, MVV yönteminin daha kolay çözümlendiği ve daha fazla aykırı değer bulduğunu belirtmiştir.

Dağhan (2011), yaptığı tez çalışmasında materyal olarak kan ölçümleri alınan 23 gönüllünün kandaki etken madde konsantrasyonlarını kullanmıştır. Çalışmada, biyoeşdeğerlik çalışmalarında aykırı değerleri bulma yöntemlerini incelemiştir.

Gogoi ve ark. (2011), uzaklık tabanlı, yoğunluk tabanlı ve yumuşak bilişim tabanlı aykırı değer belirleme yöntemlerini ve bu yöntemlerdeki algoritmaları kapsamlı bir şekilde incelemişlerdir.

Amer (2011), Ağırlıksız CBLOF ve LDCOF algoritmalarını tanıtmış ve ROC eğrisi ile karşılaştırmıştır. Çalışmanın sonucunda, iki algoritmanın benzer sonuçlar verdiğini ancak yorumlama açısından LDCOF algoritmasının daha kolay olduğu bildirilmiştir.

Amer ve Goldstein (2012), en yakın komşu tabanlı (k-NN Global Anormallik Skoru, LOF, COF, INFLO, LoOP, LOCI) ve kümeleme tabanlı (CBLOF, LDCOF) algoritmalar kullanarak aykırı değerleri tespit etmişlerdir. Çalışmanın sonucunda, ROC eğrisi incelendiğinde kümeleme tabanlı algoritmalardan CBLOF'un önemli ölçüde aykırı değerleri tespit ettiğini belirlemişlerdir.

Türkan ve ark. (2012), regresyon analizinde aykırı değerleri belirlerken Cook uzaklığı, DFFITS ve Hadi'nin ölçütünü kullanmışlardır.

Park ve ark. (2012), bölme tekniği ve Sağlam M tahmin edicisini kullanarak yeni bir Sağlam tahmin edicisi olan DPM'yi önermişlerdir. DPM tahmin edicisinin performansını belirlemek için bir simülasyon çalışması yapmışlardır. Çalışmanın sonucunda, DPM tahmin edicisinin diğer Sağlam teknikler gibi aykırı değerleri belirlemek için kullanılabileceğini belirtmişlerdir.

Ovla ve Taşdelen (2012), tek değişkenli veri setinde tekli aykırı değerlerin bulunmasında kullanılan yöntemleri örneklerle tanıtmışlardır. Çalışmanın sonucunda, Dixon testinin diğer testlere göre daha az duyarlı olduğunu, hesaplama zamanı açısından ise en hızlı test olduğunu belirtmişlerdir.

Ding ve Fei (2013), iForestASD adında yeni bir algoritma önermişlerdir. Dört gerçek veri setindeki deney sonuçları, önerilen yöntemin verimli olduğunu göstermiştir.

Bahadır (2013), Bingöl ve Ahi Evran Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Kanatlı Üretim Birimi'nde bulunan bıldırcınların canlı ağırlık performasına ait verileri materyal

olarak kullanmıştır. Bu çalışmada, aykırı değerlerin tespitinde, DFBETAS ve DFFITS ölçüsü, Cook uzaklığı ve Atkinson ölçüsünü kullanmıştır.

Kuppusamy ve Kaliyaperumal (2013), aykırı değerleri belirlemek için Dixon, Hampel, Grubbs ve Rosner testini kullanmışlardır. Çalışmanın sonunda, Grubbs ve Dixon testinin daha düşük duyarlılık gösterdiğini belirlemişlerdir.

Morano ve ark. (2013), İtalya'nın Bari şehrindeki bir semtinde satılan gayrimenkul fiyatlarındaki aykırı değerleri Sağlam LMS ile belirlemişlerdir.

Kıral ve ark. (2013), iki bankada bulunan banknotlarla ilgili Swiss verileri, Akut Leukemia hastası olan kişilerden alınan 7129 gen verileri ile simülasyon çalışmasından üretilen veriler materyal olarak kullanmışlardır. Çalışmada, klasik yöntemler ile çok boyutlu veri kümelerinde aykırı değerleri tespit eden CBRD2 yöntemini kullanarak aykırı değerleri bulmuşlardır.

Abed Mohammed (2013), Sağlam M, MM, LTS tahmin edicileri ve OLS metoduyla hem model tahmini yapmış hem de aykırı değerleri bulmuştur. Çalışmada, S-plus, Minitab ve SPSS programlarını kullanmıştır.

Tuncer (2014), yaptığı tez çalışmasında, mide kanseri teşhisi konulan hastalardan elde edilen verilere yaşam analizi uygulayıp, yaşam analizinde kullanılan yöntemler ile aykırı değerleri bulmuştur.

Sreevidya (2014), istatistik tabanlı, derinlik tabanlı, sapma tabanlı, uzaklık tabanlı, kümeleme tabanlı, sürgülü pencere tabanlı ve otomatik regresyon tabanlı aykırı değer belirleme yöntemlerini incelemiştir. Çalışmanın sonucunda, kümeleme tabanlı aykırı değer belirleme yönteminde aykırı değer bulma yüzdesinin daha yüksek olduğunu bildirmiştir.

Nurunnabi ve ark. (2014), çoklu aykırı değerlerin belirlenmesi için R yazılımında BACON algoritmasını kullanmışlardır. Bu algoritmayı seçme nedeninin hesaplama verimliliğinden kaynaklandığını bildirmişlerdir.

Yücel Altay (2014), Atatürk Üniversitesi, Diş Hekimliği Fakültesi'nden aldığı verilerde aykırı değerleri tespit etmiştir. Aykırı değer belirleme yöntemlerinden uzaklık ve yoğunluk tabanlı yöntemleri birleştiren TRAOD ve TOD-SS algoritmalarını kullanmıştır.

Nayak ve Dash (2015), üç ay boyunca işlem gören borsa verilerinde hem uzaklık hem de yoğunluk tabanlı aykırı değer belirleme yöntemlerini kullanarak aykırı değerleri tespit etmişlerdir.

Goldstein ve Uchida (2016), 10 farklı veri setinde 19 farklı aykırı değer belirleme algoritmasını RapidMiner veri madenciliği yazılımı ile karşılaştırmışlardır. Çalışmanın sonucunda, ROC eğrisine bakarak çoğu durumda kümeleme algoritmalarına kıyasla en yakın komşu tabanlı algoritmaların daha iyi performans gösterdiğini bildirmişlerdir.

Zhang (2016), lojistik regresyon analizinde araştırılan hatalar, etkili gözlemler ve aykırı değerleri R yazılımı ile incelemiştir. Bazı paketler ile aykırı değerleri, Cook uzaklığını, Student türü hataları grafikler yardımıyla kolaylıkla görselleştirmiştir.

Akdemir (2016), perakende sektöründe faaliyet gösteren bir şirketin verilerinde en yakın komşu tabanlı (k-NN Global Anormallik Skoru, LOF, COF, INFLO, LoOP, LOCI) ve kümeleme tabanlı (CBLOF, LDCOF) algoritmaları kullanarak aykırı değerleri tespit etmiştir.

Kaur ve Garg (2016), istatistik tabanlı, derinlik tabanlı, uzaklık tabanlı, yoğunluk tabanlı, kümeleme tabanlı algoritmaları karşılaştırmışlardır. Çalışmanın sonucunda, kümeleme tabanlı algoritmaların daha iyi aykırı değer bulduklarını bildirmişlerdir.

Ali ve ark. (2017), PingER veri setinde k-ortalamalar algoritması kullanılarak veriyi kümelere ayırmışlar ve LOF ile CBLOF algoritmaları yardımıyla aykırı değerleri tespit etmişlerdir.

Aleng ve ark. (2017), Sağlam LTS tahmin edicisi ile aykırı değerleri bulmak için bir simülasyon çalışması yapmışlardır. Çalışmanın sonucunda, aykırı değer tespitinde LTS tahmin edicinin iyi sonuçlar verdiğini bildirmişlerdir.

Kannan ve Sampath (2017), lojistik regresyon analizinde Pearson Hatalar, Standartlaştırılmış Pearson hatalar, sapma hatalar ve Sağlam Sapma Bileşeni (robDEVC) yöntemlerinin performanslarını karşılaştırarak yüksek kaldıraçların iyi/kötü olup olmadığını veya aykırı değerler olup olmadığını incelemişlerdir. Çalışmanın sonucunda, veride kötü kaldıraç olması durumunda robDEVC yönteminin daha başarılı olduğunu belirlemişlerdir.

Deb ve Dey (2017), k-ortalamalar ve hiyerarşik kümeleme algoritmaları ile aykırı değerleri belirleyerek kümelemeyi daha iyi hale getirmeye çalışmışlardır.

Susto ve ark. (2017), yalıtım ormanlar yöntemi ile kontrol grafikleri tabanlı ve açı tabanlı (ABOD) aykırı değer belirleme yöntemlerini kesinlik ve hatırlama değerleri ile karşılaştırmışlardır. Çalışmanın sonucunda, hassasiyet açısından yalıtım ormanlar yönteminin ABOD algoritmasından daha iyi performans gösterdiğini belirtmişlerdir.

Calheiros ve ark. (2017), büyük ölçekli bulut veri merkezlerinin kaynak kullanımında aykırı değerlerin tespiti için yalıtım orman (iForest) algoritmasının uygulanabilirliğini araştırmışlardır. Ayrıca zaman serisi bilgilerini, zamansal aykırı değer tespitini mümkün kılan, zaman serilerindeki mevsimselliğe ve eğilimlere uyum sağlama, çevrimiçi ve gerçek zamanlı olarak uygulanacak fizibilitesini belirleyen ekstra özellikler olarak kodlamak için bir yöntem geliştirmişlerdir.

Kalaycı ve Ercan (2018), kablosuz algılayıcı ağlardan elde edilen verilerde aykırı değerleri belirlemek için histogram tabanlı aykırı değer skoru (HBOS) ile uzaklık tabanlı aykırı değer belirleme yöntemlerini doğruluk oranı ve ROC eğrisi ile karşılaştırmışlardır. Çalışmanın sonucunda, HBOS yönteminin uzaklık tabanlı aykırı değer belirleme yöntemlerine göre daha iyi sonuçlar verdiği gözlemlemişlerdir.

Domingues ve ark. (2018), yapay ve gerçek veri kümelerinde 14 farklı algoritmanın (GMM, DPGMM, DPMM, RKDE, PPCA, LSA, Mahalanobis, LOF, ABOD, SOD, KL, GWR, OneClass-SVM, iForest) ortalama hassasiyetini, sağlamlığını, hesaplama süresini ve bellek

kullanımını karşılaştırmışlardır. Çalışmanın sonucunda, yalıtım orman (iForest) yönteminin aykırı değerleri belirlemede mükemmel bir yöntem olduğunu ortaya koymuşlardır.

Shou ve ark. (2018), açı tabanlı aykırı değer belirleme yöntemlerinden EAOF-OD adında yeni bir yöntem geliştirmiştir. Bazı algoritmalarla karşılaştırıldığında EAOF-OD'un özellikle yüksek boyutlu veri setinde daha yüksek doğruluk oranı ve düşük sınıflama hatası ile daha iyi performans gösterdiğini bildirmişlerdir.

Bisren ve Alkan (2019), aykırı değer belirleme yöntemlerinden kümeleme tabanlı, en yakın komşu tabanlı ve istatistik tabanlı algoritmaları doğruluk oranı, gerçek pozitif, yanlış pozitif ve AUROC başarı kriterleri ile incelemişlerdir. Çalışmanın sonucunda, kümeleme tabanlı CBLOF algoritmasının performans bakımından daha iyi olduğunu gözlemlemişlerdir.

Cheng ve ark. (2019), iForest ile LOF algoritmalarının zayıflıklarının üstesinden gelmek için, aykırı değer tespitinde iki katmanlı bir topluluk yöntemi önermişlerdir. İki algoritmayı da kullanan bu yöntem mevcut yöntemlerle karşılaştırıldığında, topluluk yönteminin aykırı değer belirleme oranını önemli ölçüde iyileştirebileceğini ve zaman karmaşıklığını büyük ölçüde azaltabileceğini göstermektedir.

Kamaladevi ve ark. (2019), göğüs kanseri verilerinde aykırı değerleri belirlemek için LOF ve kNN algoritmasını incelemişlerdir. Çalışmanın sonucunda, LOF algoritmasının doğruluk oranının daha yüksek olduğunu belirlemişlerdir.

Gao ve ark. (2019), aykırı değerleri belirlemek için sınıflandırma tabanlı, yoğunluk tabanlı ve yalıtım tabanlı algoritmaları doğruluk oranı ve F1-skor başarı kriterleri ile karşılaştırmışlardır. Yalıtım tabanlı iForest algoritmasının üstün olduğunu bildirmişlerdir. Ancak yerel aykırı değerleri tespit etmede yetersiz olduğu için yeni bir algoritma olan CBIF algoritmasını geliştirmişlerdir.

Vijayakumar ve ark. (2020), kredi kartı dolandırıcılığının önüne geçmek için iForest ve LOF algoritmasını kullanarak aykırı değerleri tespit etmişlerdir. Çalışmanın sonucunda, iForest algoritmasının LOF algoritmasına göre daha yüksek doğruluk oranına sahip olduğunu bildirmişlerdir.

Santosuosso ve ark. (2020), kanadı noktalı sirke sineği (Drosophila suzukii) verilerinde aykırı değerleri belirlemek için yalıtım ormanlar yöntemini kullanmışlardır. Bu çalışma ile bir istilanın gelecekteki yönünü belirlemek amaçlanmıştır.

Sternby ve ark. (2020), tek sınıf öğrenme için optimize edilmiş yeni bir aykırı değer belirleme algoritması olan Anomali Algılama Ormanı'nı (ADF) geliştirmişlerdir. 14 genel veri setinden oluşan bir setteki deneysel sonuçlar incelendiğinde bu algoritmanın, tek sınıflı öğrenme ortamında aykırı değer belirleme yöntemlerinden yalıtım ormanı yöntemi ve tek sınıf rastgele orman yönteminden daha iyi performans gösterdiğini ROC eğrileriyle belirtmişlerdir.

Kiruthika ve Sowmyarani (2020), kredi kartı dolandırıcılığını azaltmak amacıyla aykırı değer belirlemede LOF ve iForest algoritmasını kullanmışlardır. Çalışmanın sonucunda, iForest

algoritmasının (%99.751 doğruluk oranı, 71 yanlış tahmin sayısı) LOF algoritmasından (%99.708 doğruluk oranı, 83 yanlış tahmin sayısı) daha iyi sonuçlar verdiğini vurgulamışlardır.

Gnat (2020), Polonya'nın gayrimenkul verilerinde lojistik regresyon analizi ile model tahmini yapmıştır. iForest, kNN, LOF ve PCA aykırı değer belirleme yöntemlerini kullanarak aykırı değerleri tespit etmiştir. Daha sonra, gayrimenkul verilerinde aykırı değerler olmadan model tahminini tekrarlamıştır. Kullanılan yöntemlerinden en iyi performansa sahip algoritmanın en yüksek doğruluk oranı ile iForest olduğunu belirlemiştir.

Negi (2020), yaptığı çalışmada kredi kartlarında dolandırıcılığı tespit etmek için aykırı değer belirleme yöntemlerini karşılaştırmıştır. Çalışmanın sonucunda, iForest algoritmasının LOF algoritmasından daha iyi performans sergilediğini bildirmiştir. Ancak iForest algoritmasının hesaplama maliyeti olduğunu ve geliştirilmesi gerektiğini belirtmiştir.

Mecheva (2022), LOF, iForest ve SVM (One Class Support Vector Machine) yöntemlerini trafik veri setinde karşılaştırmasını yapmıştır. Çalışmanın sonucunda, özellikle tatil günlerinde iForest algoritmasının daha iyi aykırı değerleri bulduğunu tespit etmiştir.

Negi ve ark. (2022), yaptıkları çalışmada kredi kartı dolandırıcılığı ispat etmek için LOF ve iForest algoritmasını kullanmışlardır ve doğruluk oranlarını tespit etmişlerdir. Çalışmanın sonucunda, aykırı değer tespitinde iForest algoritmasının (%99.74 doğruluk oranı) LOF algoritmasından (%99.65 doğruluk oranı) daha iyi olduğunu bulmuşlardır.

Rajeev ve Devi (2022), yaptıkları çalışmada kredi kartı sistemindeki dolandırıcılığı tespit etmek için iForest ve LOF algoritmasını kullanmışlardır. Çalışmada karşılaştırma yaparken doğruluk, kesinlik ve hatırlama değerlerini incelemişlerdir. Çalışmanın sonucunda, %27 dolandırıcılık tespit oranı ile iForest algoritmasının, %2 oran ile LOF algoritmasından daha iyi olduğunu bulmuşlardır.

3. MATERYAL VE METOT

3.1. Materyal

Veri madenciliğinde kullanılan veri setlerinde değişkenler farklı ölçü birimlerinde ve/veya farklı büyüklüklerde olduğu için bu çalışmada bağımsız değişkenler için standart normal dağılıma sahip bir simülasyon yapılmıştır. Veri setinin örnek büyüklüğü 3000 olarak seçilmiştir. Bu veri setinde 3 bağımsız (X_1, X_2, X_3), 1 bağımlı (Y) değişken bulunmaktadır. Bağımsız değişkenler sürekli değişkenlerden oluşmaktadır. Bağımlı değişken 0 ve 1 değerlerini içeren iki kategorili (binary) Bernoulli dağılımına uyan bir değişkendir. Aykırı değer belirleme yöntemlerini karşılaştırmak için üretilen veri setine her bir bağımsız değişkende 10 adet olacak şekilde toplam 30 adet aykırı değer eklenerek veri modifiye edilmiştir. Ayrıca çalışmada materyal olarak Zhao ve ark. (2011) tarafından elde edilen bir veri seti kullanılmıştır. Bu veri seti 282 gözlem değerinden oluşmaktadır. Bu veri setinden bağımsız değişken olarak tohum uzunluğu (X_1), tohum genişliği (X_2), tohum hacmi (X_3); bağımlı değişken olarak yaprak tüylenme (Y) kullanılmıştır. Bağımsız değişkenler sürekli değişkenlerden, bağımlı değişken ise iki kategorili değişkenden (tüylenme yok:0, tüylenme var:1) oluşmaktadır. Analizlerde, R istatistiksel hesaplama ve grafik analiz yazılımının 4.5.3 sürümü (R Core Team, 2019) kullanılmıştır.

3.2. Metod

3.2.1. Lojistik Regresyon Analizi

Lojistik regresyon analizi ilk olarak Cox (1958) tarafından geliştirilmiştir. Kategorik veya sıralama ölçekli bağımlı değişken içeren veri setini modellemek için kullanılmaktadır. Bu analiz, bağımlı değişkenin tipine göre iki kategorili (binary) lojistik regresyon, çok kategorili sıralama ölçeksiz (multinominal) lojistik regresyon, çok kategorili sıralama ölçekli (ordinal) lojistik regresyon gibi çeşitlere ayrılmaktadır.

Lojistik regresyon analizinde kullanılan terimlerden biri olan odds, bir olayın olma olasılığının, olayın olmama olasılığına bölümüdür. Odds oranı (OR) 0 ile sonsuz (∞) arasında değerler almaktadır.

$$OR = \frac{p}{1-p} = \frac{\pi}{1-\pi} \tag{3.1}$$

Eşitlikte;

p ve π: Bir olayın olma olasılığını,

1-p ve $1-\pi$: Bir olayın olmama olasılığını göstermektedir.

Bir diğer terim olan lojit, odds oranının doğal logaritmasıdır. Lojitler - ∞ ile + ∞ arasında değerler almaktadır.

Tek Değişkenli Lojistik Regresyon Modeli

Bağımsız değişken sayısı bir olduğunda lojistik regresyon modeli aşağıdaki şekilde yazılabilmektedir (Agresti, 1996).

$$\pi(X) = P(Y = 1|X = x) = \frac{e^{(\beta_0+\beta_1 X)}}{1+e^{(\beta_0+\beta_1 X)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \quad (3.2)$$

π, X bağımsız değişkeni bilindiğinde (X=x olduğunda), Y'nin 1 olma olasılığını göstermektedir. Bu model, y değişkeninin odds'u türünden aşağıdaki gibi de yazılabilmektedir.

$$\frac{\pi(X)}{1-\pi(X)} = e^{(\beta_0+\beta_1 X)} \quad (3.3)$$

Odds'un lojit dönüşümü yapılır ve model Eşitlik (3.4)'deki gibi doğrusal hale gelmektedir.

$$\text{lojit } \pi(X) = g(X) = \text{In}\left(\frac{\pi(X)}{1-\pi(X)}\right) = \ln e^{(\beta_0+\beta_1 x)} = \beta_0 + \beta_1 X + e_{ij} \quad (3.4)$$

Eşitlikte;

e: Euler sayısını,

β_0: Sabit terimi,

β_1: Regresyon katsayısını,

e_{ij}: Hata terimini göstermektedir.

Çok Değişkenli Lojistik Regresyon Modeli

k tane bağımsız değişken olduğunda bağımsız değişkenler vektörü $X=(x_1, x_2,...,x_k)$ olmak üzere çoklu lojistik regresyon modeli aşağıdaki şekilde yazılabilmektedir (Agresti, 1996).

$$\pi(X) = P(Y = 1|x) = \frac{e^{(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}}{1+e^{(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} \quad (3.5)$$

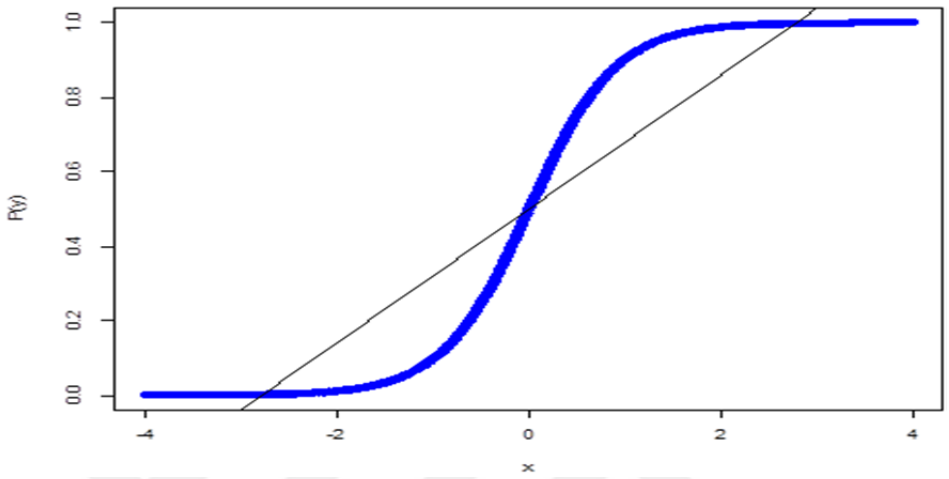
Bu model, y değişkeninin odds'u türünden aşağıdaki gibi yazılabilmektedir.

$$\frac{\pi(X)}{1-\pi(X)} = e^{(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)} \quad (3.6)$$

Odds'un lojit dönüşümü yapılır ve model Eşitlik (3.7)'deki gibi doğrusal hale gelmektedir.

$$\text{lojit } \pi(X) = g(X) = \text{In}\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \quad (3.7)$$

Lojistik regresyon eğrisi, S şeklinde olup Şekil 3.1'de verilmiştir.



Şekil 3.1. Lojistik Regresyon Eğrisi

Parametre Tahmini

Lojistik regresyon analizinde parametre tahmininde en çok olabilirlik yöntemi (ML), yeniden ağırlıklandırılmış en küçük kareler yöntemi (RILS), minumum lojit ki kare yöntemi (MLCS) kullanılmaktadır. Analizlerde, parametre tahmininde en çok olabilirlik yöntemi kullanılmıştır.

En Çok Olabilirlik Yöntemi (ML)

Lojistik regresyonda, en çok olabilirlik fonksiyonu aşağıdaki eşitlikle gösterilmektedir (Pampel, 2000).

$$Z(X_i) = \pi(X_i)^{Y_i}[1 - \pi(X_i)]^{1-Y_i} \quad (3.8)$$

Gözlemler birbirinden bağımsız olduğu varsayıldığı için en çok olabilirlik fonksiyonu aşağıdaki şekilde yazılabilmektedir (Chatfield ve Collins, 1992).

$$\prod_{i=1}^{n} Z(X_i) \quad (3.9)$$

İşlem kolaylığı açısından bu fonksiyonun logaritması alınarak log-olabilirlik fonksiyonu elde edilmektedir.

$$\text{Log-olabilirlik} = \sum_{i=1}^{n} Y_i \ln \pi(X_i) + (1 - Y_i)[\ln 1 - \ln \pi(X_i)] \quad (3.10)$$

Parametrelerin Önem Testi

Lojistik regresyon analizinde, parametrelerin önem testleri yapılırken Olabilirlik oran testi, Skor ve Wald testi kullanılmaktadır. Analizlerde, parametrelerin önem testlerinden Wald testi kullanılmıştır.

Wald testi

Wald testi, normal dağılıma (Z) uymaktadır ve aşağıdaki gibi hesaplanmaktadır (Agresti, 1996).

$$W = \frac{\widehat{\beta_j}}{S(\widehat{\beta_j})} \quad (3.11)$$

Eşitlikte;

β_j: Regresyon katsayısının en çok olabilirlik tahminini,

$S(\widehat{\beta_j})$: Regresyon katsayısının en çok olabilirlik tahmininin standart hatasını göstermektedir.

Eşitlik (3.11) ile verilen denklemin pay ve paydanın karesi alındığında Wald istatistiği, yaklaşık olarak 1 serbestlik dereceli χ^2 dağılımına uymaktadır.

$$W = \frac{\widehat{\beta_j}^2}{S(\widehat{\beta_j})^2} \quad (3.12)$$

Modelin Uyum İyiliği

Modelin uyum iyiliği, modelde olması gereken değişkenler belirlendikten sonra modelin, bağımlı değişkeni açıklayabilme etkinliğidir. Modelin uyum iyiliğinde kullanılan ölçütler; Pearson Ki-kare ve sapma istatistiği, Hosmer ve Lemeshow istatistiği ve Alıcı işletim karakteristiği eğrisidir (Receiver Operating Characteristic - ROC). Analizlerde modelin uyum iyiliği ölçütü olarak Alıcı işletim karakteristiği eğrisi (ROC) kullanılmıştır.

Alıcı İşletim Karakteristiği Eğrisi* (*Receiver Operating Characteristic -ROC)

Alıcı işletim karakteristiği eğrisi (ROC), yanılgı matrisi yardımıyla modelin uyum iyiliğini incelemektedir. Yanılgı matrisinde gözlenen ve tahmin edilen değerlerin, pozitif ve negatif olanlarının sayısı belirlenir ve aşağıdaki çapraz tablo oluşturulur.

Çizelge 3.1. Yanılgı matrisi

	Gözlenen pozitif	Gözlenen negatif
Tahmin edilen pozitif	a	b
Tahmin edilen negatif	c	d

Burada;

Duyarlılık (Doğru pozitif oranı) = $\frac{a}{a+c}$

Özgüllük (Doğru negatif oranı) = $\frac{d}{b+d}$

1-Duyarlılık (Yanlış negatif oranı) = $\frac{c}{a+c}$

1-Özgüllük (Yanlış pozitif oranı) = $\frac{b}{b+d}$ olarak hesaplanır (Metz, 1978).

Duyarlılık, gerçekte pozitif olan gözlemlerden ne kadarının pozitif olarak tahmin edildiğini; Özgüllük, negatif tahmin edilen gözlemlerden ne kadarının gerçekten negatif olduğunu; 1-Duyarlılık, gerçekte pozitif olanların yanlış şekilde negatif tahmin edilme olasılığını; 1-Özgüllük, gerçekte negatif olanların yanlış şekilde pozitif tahmin edilme olasılığını gösteren bir ölçüttür (Cebeci ve ark., 2022).

ROC eğrisinde, X ekseninde 1-özgüllük, Y ekseninde ise duyarlılık değerleri bulunmaktadır ve eğrinin altında kalan alan hesaplanmaktadır. Eğrinin altında kalan alan (AUROC), pozitif ve negatif sonuçların doğru sınıflama yüzdesini göstermektedir. Modele ait ROC eğrisinin sol üst köşeye en yakın, Y=X doğrusuna en uzak olması istenmektedir. Böylece AUROC daha yüksek değere sahip olacaktır (Dirican, 2001; Weinsteim ve ark., 2005; Obuchowski, 2005; Sasse, 2002; Tomak ve Bek, 2009). Bu da modelin uyumunun iyi olduğunu göstermektedir.

Ayrıca, performans değerlendirilmesinde kullanılan ve yanılgı matrisinden elde edilen doğruluk oranı aşağıdaki gibi bulunmaktadır (Menaka ve Booba, 2022).

$$\text{Doğruluk} = \frac{a+d}{a+b+c+d} \tag{3.13}$$

Sınıflama hatası ise 1- $\frac{a+d}{a+b+c+d}$ (1-Doğruluk) şeklinde bulunmaktadır.

Modelin Yeterliliği İle İlgili Ölçüler

Modelin yeterliliği belirlenirken belirtme katsayıları (McFadden R^2, Cox-Snell R^2, Nagelkerke R^2), hatalar (ham hatalar, lojit hatalar, Pearson hatalar, standartlaştırılmış Pearson hatalar, Student türü hatalar, Student türü silinmiş hatalar, sapma hatalar, standartlaştırılmış sapma hatalar) ve etkili gözlemler (uzaklık değerleri, DFBETAS ölçüsü, DFFITS ölçüsü, Cook uzaklığı) incelenmektedir.

Belirtme Katsayıları

Belirtme katsayısı, bağımlı değişkendeki değişimin ne kadarının regresyon modelindeki bağımsız değişkenler tarafından açıklandığını göstermektedir. Lojistik regresyon analizinde kullanılan R^2 değerleri, sahte R^2 (pseudo R^2) değerleri olarak da adlandırılmaktadır. Sahte R^2 değeri, EKK yönteminde hesaplanan R^2 değerinden yorumlama ve genel olarak küçük çıkması bakımından farklılıklar göstermektedir (Alpar, 2011).

McFadden R^2

McFadden (1973) bu yöntemi geliştirmiştir ve Eşitlik (3.14) ile verilmiştir.

$$R^2_{McF} = 1 - \frac{\ln \hat{L}(M_\beta)}{\ln \hat{L}(M_\alpha)} \tag{3.14}$$

Eşitlikte;

M_α: Sadece model sabitinin bulunduğu modelin log olabilirliğini,

M_β: Bağımsız değişkenlerin olduğu modelin log olabilirliğini göstermektedir.

Cox-Snell R^2

Maddala (1983) tarafından geliştirilen Cox-Snell R^2, en çok olabilirlik R^2 olarak da adlandırılmaktadır ve Eşitlik (3.15) ile verilmiştir.

$$R^2_{CS} = 1 - \left\{\frac{\hat{L}(M_\alpha)}{\hat{L}(M_\beta)}\right\}^{\frac{2}{n}} \tag{3.15}$$

Nagelkerke R^2

Cragg ve Uhler (1970) tarafından geliştirilen Nagelkerke R^2 Eşitlik (3.16) ile verilmiştir.

$$R^2_{NAG} = \frac{1 - \left\{\frac{\hat{L}(M_\alpha)}{\hat{L}(M_\beta)}\right\}^{\frac{2}{n}}}{1 - \hat{L}(M_\alpha)^{\frac{2}{n}}} \tag{3.16}$$

Hatalar

Hatalar, y değişkenine ait gözlemlerin gerçek ve tahmin değerleri arasındaki farktır (Arıcan, 2010). Hataların, ham (standartlaştırılmamış) hatalar, lojit hatalar, Pearson hatalar, standartlaştırılmış Pearson hatalar, Student türü hatalar, Student türü silinmiş hatalar, sapma hatalar, standartlaştırılmış sapma hatalar gibi çeşitleri vardır. Analizlerde ham hatalar, Pearson hatalar, Standartlaştırılmış Pearson hatalar ve Student türü hatalar kullanılmıştır.

Ham (Standartlaştırılmamış) Hatalar

Ham hatalar, Eşitlik (3.17) yardımıyla bulunmaktadır (Chatterjee ve Hadi, 2006).

$$\widehat{e_i} = Y_i - \widehat{Y_i} \quad \text{veya} \quad \widehat{e_i} = Y_i - \widehat{\pi_i} \qquad i=1,2,...,n \tag{3.17}$$

Eşitlikte;

e_i: i. ham hatayı,

Y_i: i. değişkenin gözlenen değerini,

$\widehat{Y_i}$ veya $\widehat{\pi_i}$: i. değişkenin tahmin (olasılık) değerini göstermektedir.

Pearson Hatalar

Pearson hatalar (r_{pe}), ham hataların kendi standart sapmalarına bölünmesi ile bulunmaktadır.

$$r_{pe} = \frac{e_i}{\sqrt{\widehat{Y_i}(1-\widehat{Y_i})}} \quad i=1,2,...,n \tag{3.18}$$

Pearson hatalar, mutlak değerce 3 ve 3'ten büyük ise aykırı değer olarak adlandırılmaktadır (Ryan, 1997).

Standartlaştırılmış Pearson Hatalar

Standartlaştırılmış Pearson hatalar, student türü pearson hatalar veya standartlaştırılmış hatalar olarak da adlandırılmaktadır (Dunteman ve Ho, 2006). Eşitlik (3.19) yardımıyla hesaplanmaktadır.

$$r_{std_{pe}} = \frac{r_{pe}}{\sqrt{1-h_{ii}}} \tag{3.19}$$

Eşitlikte;

$r_{std_{pe}}$: Standartlaştırılmış Pearson hataları,

h_{ii}: Şapka (H) matrisinin i. köşegen elemanını göstermektedir (H= $X(X'X)^{-1}X'$).

Standart Pearson hatalar, 3 ve 3'ten büyük ise o gözlem aykırı değerdir (Hadi ve Imon, 2008).

Student Türü Hatalar

Student türü hatalardan içsel student türü hatalar (r_i) Eşitlik (3.20) yardımıyla bulunmaktadır (Cook ve Weisberg, 1982).

$$r_i = \frac{e_i}{\sqrt{\widehat{\sigma^2}(1-h_{ii})}} \quad i=1,2,...,n \tag{3.20}$$

Student türü hatalardan dışsal student türü hatalar (r_i^*) Eşitlik (3.21) yardımıyla bulunmaktadır.

$$r_i^* = \frac{e_i}{\sqrt{\widehat{\sigma^2_{(i)}}(1-h_{ii})}} \tag{3.21}$$

Etkili Gözlemler

Etkili gözlemler belirlenirken uzaklık (şapka) değerleri, DFBETAS ölçüsü, DFFITS ölçüsü ve Cook uzaklığı incelenmektedir. Analizlerde, uzaklık değerleri ve Cook uzaklığı kullanılmıştır.

Uzaklık (Şapka) Değerleri

Pregibon (1981) tarafından geliştirilen Şapka (H) matrisi H= $X(X'X)^{-1}X'$ olarak tanımlanmıştır. Bu matrisin i. köşegen elemanları ise $h_{ii}=x_i(X'X)^{-1}x_i'$ şeklinde gösterilmektedir (Hoaglin ve Welsch, 1978). h_{ii} değerleri uzaklık (şapka) değerleri olarak adlandırılmaktadır. h_{ii} değeri $\frac{2(p+1)}{n}$ değerinden büyük ise gözlem, büyük gözlem uzaklığına sahiptir, etkili gözlemdir.

Cook Uzaklığı

Cook (1977) tarafından geliştirilmiştir ve Eşitlik (3.22) yardımıyla bulunmaktadır.

$$CU_i = \frac{r_i^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) \quad i=1,2,\ldots,n \tag{3.22}$$

Cook uzaklığı, $\frac{4}{n}$ değerinden büyük ise o gözlem etkili gözlemdir.

3.2.2. Aykırı Değerler

Aykırı değer, diğer gözlemlerden çok fazla sapan gözlemdir (Hawkins, 1980). Aykırı değerler, kayıtların hatalı girilmesi, insanın veya aletin yanlış ölçümü, doğal afetler veya popülasyondaki doğal bir sapma nedeniyle olabilmektedir (Hodge ve Austin, 2004). Bu değerler, sapan değer, aykırı değer, outlier, aşırı değer gibi isimler alırlar. Aykırı değerler, gerçek verilerde çok sık bulunurlar ve çoğu zaman çalışmalarda göz ardı edilirler. Veri setinde aykırı değerlerin varlığı, parametre tahminlerinin hatalı olmasına, modelin yanlış kurulmasına, yanlış sınıflandırma yapmaya ve yorumlama yaparken hata yapmaya sebep olurlar (Nurunnabi ve West, 2012). Bu nedenle çalışmalarda aykırı değerlerin incelenmesi çalışmanın hassasiyeti için çok önemlidir.

Aykırı değerler, kaldıraç noktaları, etkili gözlemler ve aykırı değerler olarak 3 sınıfa ayrılmaktadırlar.

Bunlar;

1. Kaldıraç noktaları olarak adlandırılan X (bağımsız değişken) alanındaki sapma / değişim ile
2. Dikey aykırı değer olarak adlandırılan Y (bağımlı değişken) alanındaki sapma/değişim ile
3. Hem X hem de Y alanında sapma/değişim ile elde edilmektedir (Nurunnabi ve West, 2012).

Aykırı değer belirleme yöntemleri tek değişkenli veya çok değişkenli yöntemler; parametrik, yarı parametrik ve parametrik olmayan yöntemler; eğitimli, yarı eğitimli ve eğitimsiz yöntemler olarak sınıflandırılmaktadır. Özellikle veri madenciliğinde kullanılan yöntemlerden yoğunluk tabanlı, kümeleme tabanlı, uzaklık tabanlı ve derinlik tabanlı aykırı değer belirleme yöntemleri olarak da sınıflandırılmaktadır (Ben-Gal, 2005; Gogoi ve ark., 2011; Yücel Altay, 2014; Cebeci, 2020). Bu sınıflandırmaya istatistik tabanlı, sapma tabanlı ve alt uzay tabanlı aykırı değer belirleme yöntemleri de eklenebilmektedir (Xu ve ark., 2018). Son zamanlarda aktif olarak kullanılan ve performansı yüksek olan yalıtım tabanlı aykırı değer belirleme yöntemi de bulunmaktadır (Liu ve ark., 2008). Tek değişkenli aykırı değer belirleme yöntemleri açıklayıcı istatistiklere dayalı yöntemler (Z değerleri, MAD kuralı, Tukey yöntemi, Sağlam tahmin edicilere dayanan yöntemler) ve istatistiksel testlere dayalı yöntemler (Thompson tau testi, Chauvenet'in ω testi, Pearson ve Sekar testi, Dixon Q testi, Cochran testi, Bartlett testi, Hartley testi, Levene testi, Brown-Forsythe testi, Walsh testi, Nalimov testi, Grubbs testi, Tietjen-Moore testi, Rosner testi, Iglewicz ve Hoaglin testi) olarak ikiye ayrılmaktadır. Çok değişkenli aykırı değer belirleme yöntemleri Mahalanobis uzaklığına dayalı yöntem, Sağlam tahmin edicilere dayalı yöntemler (MCD, MVE, FMCD, BACON, BOFOLS, MVV) olarak ayrılırlar.

Tek değişkenli aykırı değer belirleme yöntemlerinde her değişken bağımsız olarak dikkate alındığı için iki veya daha fazla değişken aynı anda dikkate alındığında aykırı değerler tespit edilememektedir (Sunderland ve ark., 2019). Bu nedenle tek değişkenli aykırı değer belirleme yöntemleri yerine çok değişkenli aykırı değer belirleme yöntemleri tercih edilmektedir.

Aykırı Değer Belirleme Yöntemleri

Aykırı değerleri belirlemenin veri madenciliğinde birçok açıdan faydası vardır. Bankalar, kredi kartı şirketleri, cep telefonu şirketleri, borsa, sigorta acenteleri gibi kuruluşlarda meydana gelebilecek suçların ortaya çıkartılması aykırı değerlerin belirlenmesi ile mümkündür. Tıbbi ve halk sağlık alanında sigortaların kötüye kullanımlarının ortaya çıkartılması, hastalıkların kontrol altında tutulması veya teşhis koyma ve etkili tedavi yöntemlerinin geliştirilmesi aykırı değerlerin tespiti ile mümkündür. Askeri alanda, düşman bölgesinde her zamankinden farklı bir uydu görüntüsünün alınması ve olacakları tahmin etmek aykırı değerlerin bulunması ile mümkündür. Endüstri alanında motor ve diğer fiziksel yapılardaki kusurlarla ilgili tespitin yapılması yine aykırı değerlerin belirlenmesi ile mümkündür. Hayvan davranışlarında ani değişikliklerin belirlenmesi, tarımda sel, kuraklık ve ticaret savaşları gibi durumların tespit edilip önlemler alınması aykırı değerlerin belirlenmesi ile mümkündür (Chandola ve ark., 2009; Gogoi ve ark., 2011; Yücel Altay, 2014; Akdemir, 2016; Gurrapu ve ark., 2021).

Bu çalışmada, aykırı değer belirleme yöntemlerinden model (istatistik) tabanlı, yakınlık tabanlı, kümeleme tabanlı, sapma tabanlı ve yalıtım tabanlı aykırı değer belirleme yöntemleri incelenmiştir.

Model (İstatistik) Tabanlı Aykırı Değer Belirleme Yöntemleri

Model tabanlı aykırı değer belirleme yöntemlerinde, veri kümesine uyan bir model belirlenmektedir ve bu modelin düşük olasılıklı bölgelerindeki nesneler aykırı değer olarak tanımlanmaktır (Han ve ark, 2012).

Mahalanobis Uzaklığı İle Aykırı Değer Belirleme

Model tabanlı aykırı değer belirleme yöntemlerinden biri Mahalanobis uzaklığıdır. Mahalanobis uzaklığı, Eşitlik (3.23) ile verilen formülle bulunmaktadır (Rousseeuw ve Van Zomeren, 1990).

$$M_i = \sqrt{(X_i - T(X_i))C(X)^{-1}(X_i - T(X_i))'} \qquad i=1,2,...,n \tag{3.23}$$

Eşitlikte;

M_i: i. gözlemin Mahalanobis uzaklığını,

$T(X_i)$: Veri setinin aritmetik ortalamasını,

$C(X)$: Veri setinin kovaryans matrisini,

n: Örnek büyüklüğünü göstermektedir.

Eşitlik (3.23) ile bulunan her gözlemin Mahalanobis uzaklıkları, p serbestlik dereceli χ^2 dağılışın cetvel değeriyle karşılaştırılır. Bulunan değerden büyük uzaklıklara sahip gözlemler aykırı değer olarak tanımlanmaktadır (Rousseeuw ve Van Zomeren, 1990; Cebeci, 2020).

Yakınlık Tabanlı Aykırı Değer Belirleme Yöntemleri

Yakınlık tabanlı aykırı değer belirleme yöntemleri en yakın komşuluğa dayalıdır. Bu yöntemde nesnenin en yakın komşularına yakınlığı diğer nesnelerin çoğunun komşularına yakınlığından önemli ölçüde uzaksa, bu nesnenin aykırı değer olduğunu varsaymaktadır. Yakınlık tabanlı aykırı değer belirleme yöntemleri kendi içerisinde yoğunluk ve uzaklık tabanlı aykırı değer belirleme yöntemleri olarak ayrılmaktadır (Han ve ark., 2012).

Yoğunluk Tabanlı Aykırı Değer Belirleme Yöntemleri

Yoğunluk tabanlı aykırı değer belirleme algoritmalardan biri Yerel Aykırı Değer Faktörüdür (Local Outlier Factor-LOF). Bu algoritma, en yakın k komşu temeline dayanmaktadır (Breunig ve ark., 2000). Alt popülasyonların olması durumunda, bir gözlemin yerel yoğunluğu ile komşularının yoğunluğu karşılaştırılmaktadır. Bu algoritmada, gözlemin aykırı değer olması için, yerel yoğunluğun, komşularının yoğunluğundan daha düşük olması gerekmektedir (Cebeci, 2020).

Bir O gözlemi ile en yakın k komşusu arasındaki uzaklık O'nun k-uzaklığı olup $\text{uzaklık}_k(O)$ ile gösterildiğinde O'nun k uzaklık komşuluğu aşağıdaki gibidir.

$$\mathrm{N}_k(O) = \{O' | O' \in \mathrm{D}, \text{uzaklık}_k(O, O') \leq \text{uzaklık}_k(O)\} \tag{3.24}$$

O gözleminden herhangi bir O' gözlemine uzaklık, k kullanıcı tanımlı komşu sayısı aşağıdaki gibi gösterilmektedir.

$$\text{Ulaşılabilir uzaklık}_k(O \leftarrow O') = \text{maks.}\{\text{uzaklık}_k(O), \text{uzaklık}(O, O')\} \tag{3.25}$$

O gözleminin yerel erişilebilirlik yoğunluğu Eşitlik (3.26) ile verilmiştir (Hofmann ve Klinkenberg, 2014).

$$\text{Yerel uzaklık}_k(O) = \frac{\|\mathrm{N}_k(O)\|}{\sum O' \in \mathrm{N}_k(O) \text{ ulaşılabilir uzaklık}_{\mathbf{k}}(O \leftarrow O')} \tag{3.26}$$

O gözleminin yerel aykırı değer faktörü (LOF) Eşitlik (3.27) ile verilmiştir.

$$\mathrm{LOF}_k(O) = \frac{\sum O' \in \mathrm{N}_k(O) \,.\, \text{yerel uzaklık}_{\mathbf{k}}(O')}{\|\mathrm{N}_k(O)\| \,.\, \text{yerel uzaklık}_k(O)} \tag{3.27}$$

O gözleminin yerel erişebilirlik yoğunluğunun düşük, *k* en yakın komşunun (kNN) yerel erişebilirlik yoğunluğunun yüksek olması, LOF'un yüksek olması anlamına gelmektedir. Bundan dolayı, yerel yoğunluğu komşularına göre düşük olan LOF değerli gözlemler aykırı değerler olarak tanımlanmaktadır (Cebeci, 2020).

Kümeleme Tabanlı Aykırı Değer Belirleme Yöntemleri

Kümeleme tabanlı aykırı değer belirleme yöntemleri nesneler ve kümeler arasındaki ilişkiyi inceleyerek aykırı değerleri belirlemektedir. Normal veri nesneleri, büyük ve yoğun kümelere aitken; aykırı değer, küçük ve seyrek kümelere veya herhangi bir kümeye ait olmadığını varsaymaktadır (Han ve ark., 2012).

Kümeleme yöntemleri hiyerarşik ve hiyerarşik olmayan (bölümleyici) kümeleme yöntemleri olarak ikiye ayrılmaktadır ve bu yöntemler ile aykırı değerler bulunabilmektedir.

Hiyerarşik Olmayan (Bölümleyici) Kümeleme ile Aykırı Değer Belirleme

Büyük veri setlerinde sıkça tercih edilen hiyerarşik olmayan (bölümleyici) kümeleme ile aykırı değer belirleme algoritmalarından biri k-ortalamalar (K-means) algoritmasıdır. Burada amaç, n adet verinin k sayısı kadar kümelerine bölerek kümeler arası benzerliği minimum, küme içi benzerliği maksimum yapmaktır (Yadav ve Sharma, 2013; Deb ve Dey, 2017).

Burada k istenen küme sayısı, D ise n adet veri içeren bir veri kümesi olmak üzere algoritmanın adımları aşağıda verilmiştir (Kaya ve Köymen, 2008; Han ve ark., 2012; Bertizlioğlu ve Özgönenel, 2012).

1. Başlangıçta küme merkezlerini belirlemek için D veri setinden k küme sayısını oluşturacak rastgele olarak n adet veri belirlenir.
2. Her verinin ortalamasını hesaplanır ve merkez noktaları belirlenir.
3. Her veri, ortalama değerlerine bağlı olarak en yakın merkez noktasının bulunduğu kümeye atanır.
4. Her bir veri öğesinin yeni ortalama değeri hesaplanır.
5. k değişmeyene kadar adım 2 ve adım 3 tekrar edilir.

Bölümleyici kümeleme ile aykırı değer belirleyebilmek için yukarıda belirtilen 3. adımdan sonra her bir veri elemanın ait olduğu kümenin merkezinden sapmalarının kareleri toplanıp karekökü alınarak her bir veri elemanının uzaklıkları bulunmaktadır. En fazla uzaklığa sahip olan nesneler aykırı değer olarak kabul edilmektedir (Cebeci, 2020).

Sapma Tabanlı Aykırı Değer Belirleme Yöntemleri

Sapma tabanlı aykırı değer belirleme yöntemlerinde, veri kümesindeki nesnenin ana özellikleri gözlemlenerek aykırı değerler tanımlanmaktadır. Bu özelliklerden çok fazla sapan nesne, aykırı değer olarak değerlendirilmektedir (Patel ve ark., 2015).

Açı Tabanlı Aykırı Değer Belirleme Yöntemleri

Sapma tabanlı aykırı değer belirleme yöntemlerinden olan açı tabanlı aykırı değer belirleme yöntemleri algoritmalarından birisi ABOD (Angle Based Outlier Detection-Açı Tabanlı Aykırı Değer Belirleme) algoritmasıdır. Bu yöntemde, veri nesnelerinin fark vektörleri arasındaki açıların varyansları dikkate alınmaktadır (Kriegel ve ark., 2008). Bu nesneler arasındaki açıların varyansları küçük ise aykırı değer olarak kabul edilmektedir.

D veri setinde, bir nokta $\vec{A} \in D$ ($\vec{A} = (A_1, A_2, \dots A_n)$) ve diğer iki nokta $\vec{B}, \vec{C} \in D$ ve $\vec{B}, \vec{C} \neq A$ olduğunda açı tabanlı aykırı değer faktörü (Angle-Based Outlier Factor, ABOF) Eşitlik (3.28) ile hesaplanmaktadır.

$$ABOF(\vec{A}) = Var_{\vec{B},\vec{C} \in D}\left(\frac{(\overline{AB},\overline{AC})}{\|\overline{AB}\|^2.\|\overline{AC}\|^2}\right) \quad (3.28)$$

Eşitlikte;

$\overline{AB}$: Fark vektörünü ($\vec{B} - \vec{A}$),

$\|\overline{AB}\|$: $\vec{A}$ ile $\vec{B}$ arasındaki öklid uzaklığını,

ABOF($\vec{A}$): $\vec{A}$'nın sapma vektörleri ile D'deki noktaların uzaklığına göre ağırlıklandırılmış tüm nokta çiftleri arasında bulunan açılardaki sapmayı göstermektedir (Kriegel ve ark., 2008).

Eşitlik (3.28) ile verilen ABOF değerlerinden diğerlerine kıyasla daha küçük olanlar aykırı değer olarak adlandırılırlar.

Yalıtım Tabanlı Aykırı Değer Belirleme Yöntemleri

Yalıtım tabanlı aykırı değer belirleme yöntemlerinde herhangi bir uzaklık veya yoğunluk ölçüsüne güvenmeden örnekleri izole ederek aykırı değerleri belirlemektedir (Liu ve ark., 2012).

Yalıtım Orman Yöntemi ile Aykırı Değer Belirleme

Yalıtım orman yöntemi (iForest) ile aykırı değer belirleme iki aşamalı aşamadan oluşmaktadır. İlk (eğitim) aşama, verilen eğitim setinin alt örneklerini kullanarak yalıtım ağaçları oluşturmaktır. İkinci (test) aşama, her örnek için bir aykırı değer puanı elde etmek üzere test örneklerini yalıtım ağaçlarından geçirmektir (Liu ve ark., 2008).

Bir veri noktasının aykırı değer puanını hesaplamak için kullanılan algoritma, yalıtım ağacı (iTrees) yapısının ikili arama ağaçları'na (BST) eşdeğer olduğu gözlemine dayanmaktadır. Karar ağacının harici bir düğümüne yapılan bir sonlandırma, BST'de başarısız bir aramaya karşılık gelmektedir (Liu ve ark., 2008). Sonuç olarak, harici düğüm sonlandırmaları için ortalama h(x) tahmini, BST'deki başarısız aramalarındaki gibidir (Gao ve ark., 2019; Shaffer, 2011).

$$c(m)=\begin{cases} 2H(m-1)-\frac{2(m-1)}{n}, & m>2 \\ 1, & m=2 \\ 0, & \text{diğer durumlar} \end{cases} \quad (3.29)$$

Eşitlikte;

n: Test verisi boyutunu,

m: Örnek kümesi boyutunu,

H: Harmonik bir sayıyı (H(i)=In(i)+γ, γ =0.5772156649 (Euler-Mascheroni sabiti)) göstermektedir.

Yalıtım ormanı yöntemi ile aykırı değer puanı Eşitlik (3.30) ile hesaplanmaktadır (Liu ve ark., 2008; Negi, 2020).

$$s(x,m) = 2^{-\frac{E(h(x))}{c(m)}} \quad (3.30)$$

Eşitlikte;

h(x): x gözleminin yol uzunluğunu,

E(h(x)): Bir yalıtım ağaçlarından elde edilen h(x)'in ortalamasını,

c(m): İkili arama ağacındaki başarısız aramanın ortalama yol uzunluğunu göstermektedir.

E(h(x)) değeri c(n)'e eşit ise aykırı değer puanı 0.5'e; E(h(x)) değeri 0'a eşit ise aykırı değer puanı 1'e; E(h(x)) değeri n-1'e eşit ise aykırı değer puanı 0'a eşit olmaktadır.

Aykırı değer puanına göre gözlemlerin aykırı değer olup olmadığı aşağıdaki şekilde kararlaştırılmaktadır.

- Aykırı değer puanı 1'e yakınsa o gözlem aykırı değerdir.
- Aykırı değer puanı 0.5'ten çok daha küçük ise normal değerdir.
- Aykırı değer puanı yaklaşık olarak 0.5'e eşitse tüm veri setinde herhangi bir aykırı değer bulunmamaktadır (Liu ve ark., 2008).

4. BULGULAR VE TARTIŞMA

4.1. Yapay Veri Setine Ait Bulgular

Çalışmada üretilen veri setine 30 adet aykırı değer eklenerek veri modifiye edilmiştir. Bu 30 adet aykırı değer her bir bağımsız değişkene 10 adet olacak şekilde rastgele eklenmiştir. Veri setinde bulunan değişkenler ve değişkenlerin içerikleri Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Üretilen veri setinin içeriği

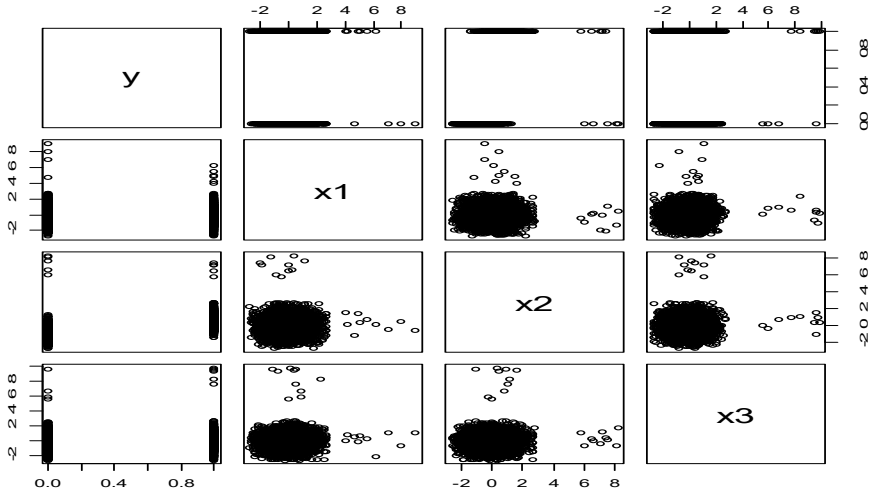
N:3000
X_1: Sürekli değişken
X_2: Sürekli değişken
X_3: Sürekli değişken
Y: İki kategorili değişken (0 ve1)

Aykırı değer eklenerek modifiye edilen veri setinde bulunan değişkenlere ait tanımlayıcı istatistikler Çizelge 4.2'de verilmiştir

Çizelge 4.2. Tanımlayıcı istatistikler

Tanımlayıcı İstatistikler	Y	X_1	X_2	X_3
Ortalama	0.53	0.01	0.03	-0.03
Medyan	1.00	0.01	0.01	-0.07
Minimum değer	0.00	-2.68	-2.64	-2.69
Maksimum değer	1.00	9.00	8.20	9.80
1. çeyreklik	0.00	-0.70	-0.68	-0.72
3. çeyreklik	1.00	0.67	0.69	0.63

Aykırı değerli verinin dağılışı ve verinin yapısı Şekil 4.1'de grafiksel olarak gösterilmiştir.



Şekil 4.1. Verinin yapısı

Makine öğrenimi, çeşitli algoritmalar ve yöntemler ile eldeki veri setinde bazı kalıpları aramaktadır ve bu kalıplara karşılık gelen etiketlere bakarak öğrenmektedir. Daha sonra öğrendiği benzer kalıplarla karşılaşırsa deneyimlerinden yararlanarak çıkarım yapabilen sistemler geliştirmektedir (Anonim, 2020). Makine öğrenimi algoritmalarıyla model kurmaya çalışırken eldeki veri setini eğitim (train) ve kontrol (test) veri seti olarak iki gruba ayırmak gerekmektedir. Eğitim veri setinde model, verilerden öğrenerek oluşturulmaktadır. Test veri setinde ise modelin ne kadar sağlıklı sonuçlar elde ettiği test edilmektedir (Akküçük, 2011). Bu çalışmada kullanılan veri seti, eğitim (%70) ve kontrol (%30) veri seti olmak üzere ikiye ayrılmıştır.

Kullanılan veri setindeki bağımlı değişkende (Y), başarı ve başarısızlık oranı birbirine yakın olmalıdır. Bu durum gerçekleşmezse sonuçlar yanlı olup güvenilir olmamaktadır. Bu çalışmada kullanılan veri setinde bağımlı değişkenin seviyelerinin sayısı incelenmiş ve Çizelge 4.3'de verilmiştir.

Çizelge 4.3. Y değişkeni seviyelerinin sayısı

	Başarı oranı	Başarısızlık oranı
Y değişkeni	1610	1390

Çizelge 4.3'de başarı/başarısızlık oranı yaklaşık olarak eşit olduğu için bu çalışmada iadesiz şansa bağlı örnekleme yöntemi kullanılmıştır.

4.1.1. Yapay Veri Setinde Lojistik Regresyon Analizine Ait Bulgular

Eğitim veri seti kullanarak bulunan değişkenlerin tahmin değerleri ve ilgili istatistikler Çizelge 4.4'de verilmiştir.

Çizelge 4.4. Aykırı değer içeren yapay veri setinin değişkenleri ve ilgili test istatistikleri

Değişkenler	Tahmin Değeri	Standart Hata	Wald	p	Güven Aralıkları	
					Alt Limit	Üst Limit
Sabit	0.2872	0.06385	4.499	6.82e-06 ***	0.16268	0.41310
X_1	-0.0689	0.05825	-1.183	0.237	-0.18352	0.04499
X_2	2.8125	0.12298	22.869	< 2e-16 ***	2.57855	3.06092
X_3	0.0574	0.06024	0.954	0.340	-0.06011	0.17594

Önem seviyeleri: *: p<0.05, **:p<0.01

Çizelge 4.4'de parametre önem testlerinden biri olan Wald testine göre X_2 değişkeni önemli bulunmuştur (p<0.001). X_2 değişkeninin Wald değeri aşağıdaki şekilde hesaplanmıştır.

X_2 değişkeni için $W=\frac{2.8125}{0.12298}=22.869$

Bulunan lojistik regresyon modeli, $\ln(\frac{\pi(X)}{1-\pi(X)}) = 0.2872 - 0.0689X_1 + 2.8125X_2 + 0.0574X_3$ veya $\frac{\pi(X)}{1-\pi(X)} = e^{0.2872-0.0689X_1+2.8125X_2+0.0574X_3}$ şeklindedir.

Değişkenlerin tahmin değerleri ve güven aralıklarının Odds oranları Çizelge 4.5'de verilmiştir.

Çizelge 4.5. Değişkenlerin tahmin değerleri ve güven aralıklarının odds oranları

Değişkenler	Tahmin değerleri	Güven Aralıkları	
		Alt limit	Üst limit
Sabit	1.33279	1.17666	1.51150
X_1	0.93340	0.83233	1.04602
X_2	16.6517	13.1780	21.347299
X_3	1.05917	0.94165	1.19237

Çizelge 4.5'de verilen değerler Çizelge 4.4'te bulunan değişkenlerin tahmin değerleri ve güven aralıklarının e üssü alınarak bulunmuştur. Bu oranlar, modeldeki diğer değişkenler sabit tutulduğunda ilgili değişkenin bir birim arttırıldığında/azaltıldığında üstünlük oranında hangi düzeyde bir artış/azalış olacağını söylemektedir.

Modelin uyum iyiliği testlerinden biri olan alıcı işletim karakteristiği eğrisi (ROC) çizebilmek için öncelikle kesme noktası belirlenmiştir ve bu kesme noktası için tahmin değerleri kullanarak yanılgı matrisi bulunmuştur. Modele ait optimum kesme noktası 0.59 olarak hesaplanmıştır ve bulunan yanılgı matrisi Çizelge 4.6'da verilmiştir.

Çizelge 4.6. Aykırı değer içeren yapay veri setinin yanılgı matrisi

	Gözlenen pozitif	Gözlenen negatif	Toplam
Tahmin edilen pozitif	385 (%42.73)	48 (%5.33)	433 (%48.06)
Tahmin edilen negatif	98 (%10.87)	370 (%41.07)	468 (%51.94)
Toplam	483 (%53.60)	418 (%46.40)	901 (%100)

Çizelge 4.6'da verilen değerler kullanılarak duyarlılık, özgüllük, 1-duyarlılık, 1-özgüllük, doğruluk oranı ve sınıflama hatası aşağıdaki gibi hesaplanmıştır.

Duyarlılık (Doğru pozitif oranı) = $\frac{385}{385+98}$ = 0.797

Özgüllük (Doğru negatif oranı) = $\frac{370}{370+48}$ = 0.885

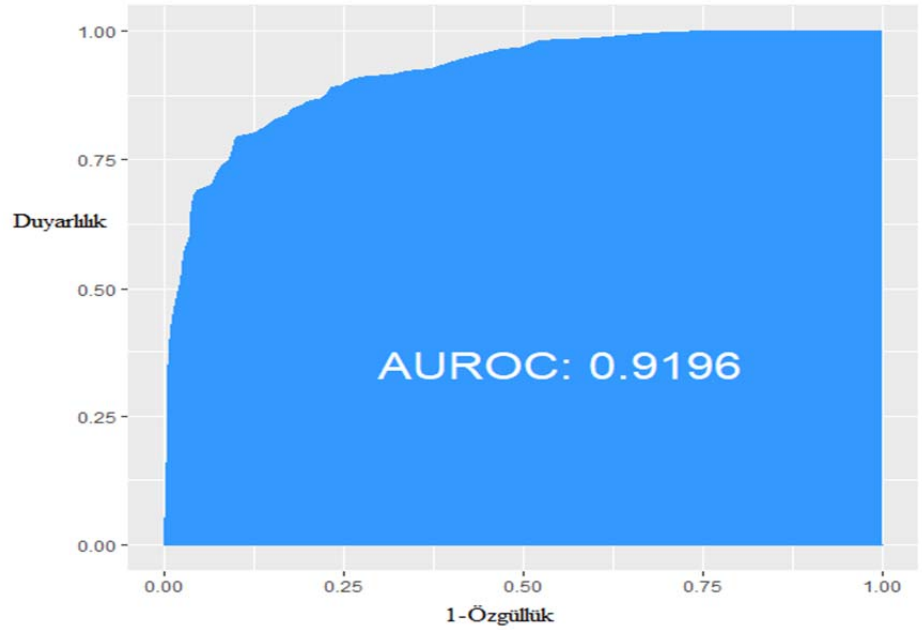
1-Duyarlılık (Yanlış negatif oranı) = $\frac{98}{98+385}$ = 0.202

1-Özgüllük (Yanlış pozitif oranı) = $\frac{48}{48+370}$ = 0.114

Doğruluk oranı= $\frac{385+370}{385+48+98+370}$ = 0.8380

Sınıflama hatası = 1-0.8380 = 0.1620

Test veri seti kullanarak çizdirilen ROC eğrisi Şekil 4.2'de verilmiştir. ROC eğrisi çizdirilirken yukarıda bulunan 1-özgüllük ve duyarlılık değerleri kullanılmıştır.



Şekil 4.2. Aykırı değer içeren yapay veri setinin ROC eğrisi

Şekil 4.2'de ROC eğrisi, sol üst köşeye yakındır. Böylelikle eğrinin altında kalan alan (AUROC=0.9196) fazladır. Bu da modelin uyumunun iyi olduğunu göstermektedir.

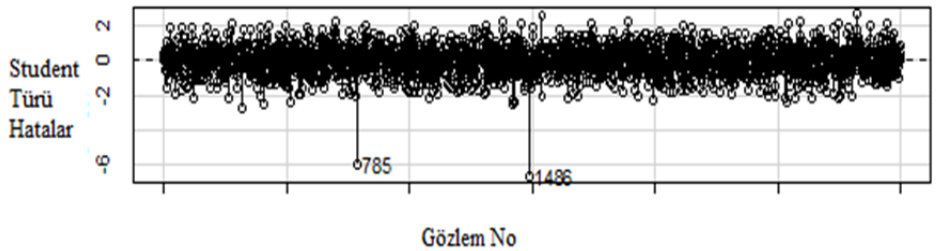
Modelin yeterliliğinde kullanılan sahte R^2 değerleri Çizelge 4.7'de verilmiştir.

Çizelge 4.7. Aykırı değer içeren yapay veri setinin sahte R^2 değerleri

McFadden R^2	Cox-Snell R^2	Nagelkerke R^2
0.4567747	0.4676844	0.6248164

Çizelge 4.7'de verilen McFadden R^2 değerine göre bağımlı değişkendeki toplam değişimin %45.67'sini, Cox-Snell R^2 değerine göre bağımlı değişkendeki toplam değişimin %46.76'sını, Nagelkerke R^2 değerine göre bağımlı değişkendeki toplam değişimin %62.48'ini bağımsız değişkenler ile açıklanmaktadır.

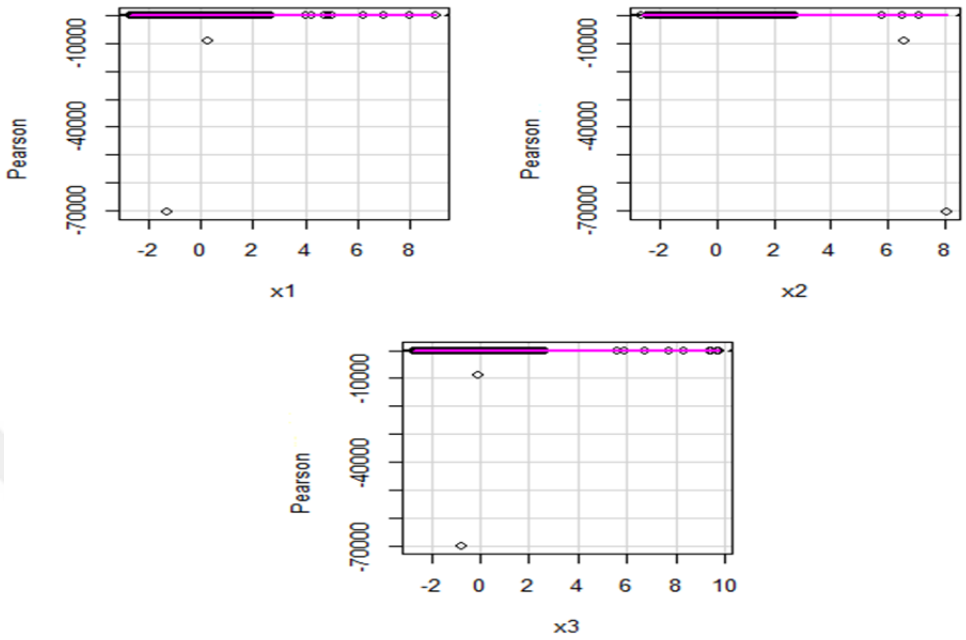
Kurulan modelin yeterliliğinde kullanılan hatalardan biri olan Student türü hataların grafiği Şekil 4.3'de verilmiştir.



Şekil 4.3. Student türü hatalar

Şekil 4.3 incelendiğinde 785 ve 1486 nolu gözlemler dikkat çekmektedir. Bu gözlemlerin aykırı değer olduğu düşünülmektedir. Ancak kesinlik için daha ayrıntılı inceleme gerekmektedir.

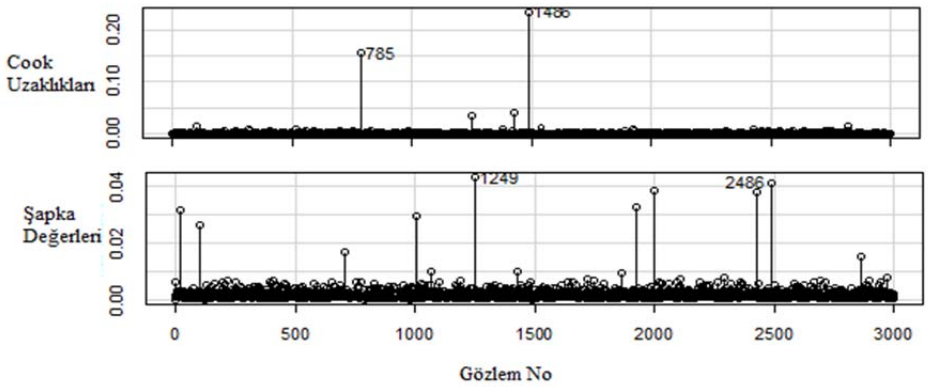
Her bir bağımsız değişken için Pearson hatalar grafiği Şekil 4.4'de verilmiştir.



Şekil 4.4. Pearson hatalar grafiği

Şekil 4.4 incelendiğinde aykırı değerler dikkat çekmektedir ancak tüm aykırı değerler görülememektedir.

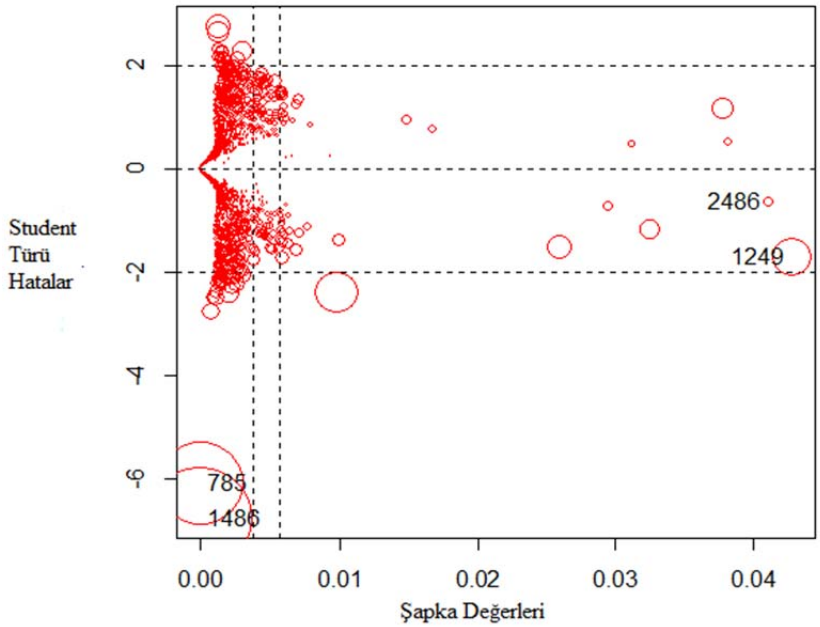
Etkili gözlemlerin incelenmesinde kullanılan Cook uzaklıkları ve şapka değerleri grafiği Şekil 4.5'de verilmiştir.



Şekil 4.5. Cook uzaklıkları ve Şapka değerleri grafiği

Cook uzaklığı grafiğinde kritik değer 0.001'dir (4/n). Şekil 4.5'e göre kritik değeri aşan gözlemler olup, en çok aşan gözlem 785 ve 1486 nolu gözlemlerdir. Şapka değerleri grafiğinde ise 1249 ve 2486 nolu gözlemlerin şapka değeri fazladır. Bu gözlemler, etkili gözlemlerdir.

Şapka değerlerine karşı student türü hatalar grafiği Şekil 4.6'da çizdirilmiştir.



Şekil 4.6. Student türü hatalar ve Şapka değerleri grafiği

Şekil 4.6'da dairenin büyüklüğü, Cook uzaklığı ile doğru orantılıdır. Burada en büyük daire 1486 nolu gözlemdedir. Bu gözlemi 785, 1249 ve 2486 nolu gözlemler izlemektedir. Bu gözlemlerin Cook uzaklıkları en fazladır. Bu gözlemler etkili gözlemlerdir.

Çizelge 4.8. Bazı gözlemlerin Student türü hataları, Şapka değerleri ve Cook uzaklıkları

Gözlem No	Student Türü Hatalar	Şapka değerleri (h_{ii})	Cook Uzaklığı
1486	-6.7514187	1.906469e-10	0.235336641
1249	-1.6856391	4.274094e-02	0.033606148
2486	-0.6321865	4.101668e-02	0.002403714
785	-6.0847953	7.756996e-09	0.155679363

Çizelge 4.8'de Şekil 4.5 ve Şekil 4.6'da çizilen grafiklerde dikkat çeken gözlemlerin Student türü hata, Şapka ve Cook uzaklığı değerleri verilmiştir. Bu gözlemler 1249, 1486, 2486 ve 785 nolu gözlemlerdir.

4.1.2. Yapay Veri Setinde Aykırı Değer Belirleme Yöntemlerine Ait Bulgular

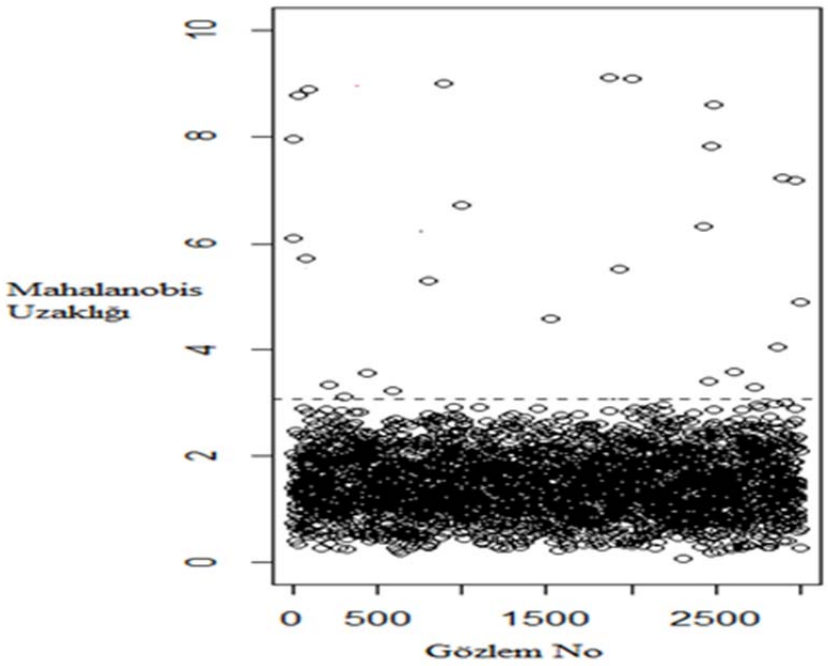
Model (İstatistik) Tabanlı Aykırı Değer Belirleme Yöntemlerine Ait Bulgular

Model tabanlı aykırı değer belirleme yöntemlerinden biri olan Mahalanobis uzaklığı çalışmada kullanılmıştır. Mahalanobis uzaklıkları bulunurken "chemometrics" (Filzmoser ve Varmuza, 2017) paketinden yararlanılmıştır. Mahalanobis uzaklıkları Çizelge 4.9'da verilmiştir.

Çizelge 4.9. Mahalanobis uzaklıkları

Gözlem No	Mahalanobis Uzaklıkları
1	7.9647
2	6.0911
3	0.7405
.	.
.	.
2999	0.6286
3000	4.8951

Çizelge 4.9'da verilen uzaklıklara göre aykırı değer belirlemek için kesme noktası 3.0575 (%97.5'luk düzeyde χ^2 dağılışı eşik değerinin karaköküdür) alınmıştır. 26 gözlemin Mahalanobis uzaklığı kesme noktasından büyüktür. Bulunan Mahalanobis uzaklıklarının grafiksel gösterimi Şekil 4.7'deki gibidir.



Şekil 4.7. Mahalanobis uzaklığı ile bulunan aykırı değerler grafiği

Şekil 4.7'de görüldüğü gibi kesme noktasını aşan 26 adet gözlem aykırı değer olarak adlandırılmıştır.

Yakınlık Tabanlı Aykırı Değer Belirleme Yöntemlerine Ait Bulgular

Yakınlık tabanlı aykırı değer belirleme yöntemlerinden yoğunluk tabanlı aykırı değer belirleme yönteminde kullanılan algoritmalardan biri olan Yerel Aykırı Değer Faktörü (LOF) çalışmada kullanılmıştır. LOF değerleri bulunurken "Rlof" (Hu ve ark., 2015) paketinden yararlanılmıştır.

(3.27) nolu eşitlik kullanılarak *O* gözleminin LOF değerleri bulunmuştur. En yakın komşu sayısı olan k= { 3, 4, 5, 6 } alınarak LOF değerleri Çizelge 4.10'da verilmiştir.

Çizelge 4.10. En yakın komşu sayısına göre LOF değerleri

Gözlem No	LOF Değerleri			
	k=3	k=4	k=5	k=6
1	1.2512125	1.3196123	1.2279284	1.1983593
2	1.2345586	1.3072740	1.2213093	1.1973427
3	0.9967768	1.0083777	1.0033600	1.0148959
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2999	0.9728727	0.9175470	0.9683469	0.9496927
3000	1.3135602	1.4894955	1.3124251	1.2008124

LOF yönteminde aykırı değer sayısını belirlemek için LOF değerlerinin 3. çeyreklik (Q3) ölçüsüne çeyrekler arası değişim genişliği (IQR=Q3−Q1) eklenerek bir kesme değeri hesaplanmaktadır (Cebeci, 2020). k=3 için kesme noktası 1.2277, k=4 için kesme noktası 1.204, k=5 için kesme noktası 1.1910, k=6 için kesme noktası 1.1825 olarak bulunmuştur. Bu kesme noktalarından daha büyük olan LOF değerleri, aykırı değer olarak adlandırılmıştır.

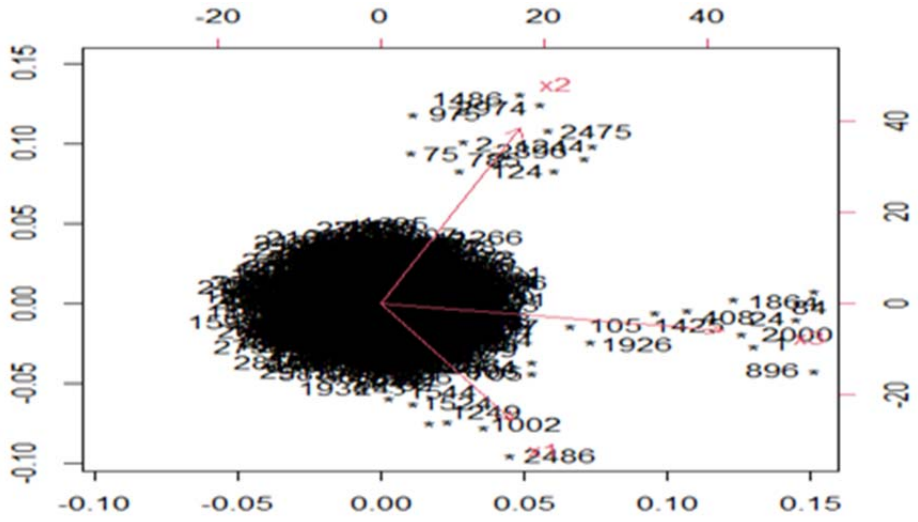
Her bir en yakın komşu sayısına göre belirlenen aykırı değerler Çizelge 4.11'de verilmiştir.

Çizelge 4.11. En yakın komşu sayısına göre aykırı değerler

En Yakın Komşu Sayısı	Aykırı Değerler*
k=3,4,5,6	2896 1002 1524 2864 75 2431 785 2486 105 896 705 408 84 975 1249 2974 2 1244 1425 806 2475 124 3000 1 1486 1926 2000 24 1864

*: Gözlem Numaraları

Çizelge 4.11 incelendiğinde her bir k seviyesinde bulunan aykırı değerler aynıdır. LOF algoritmasıyla eklenen 29 gözlem aykırı değer olarak belirlenmiştir. Bu aykırı değerlerin grafiksel olarak gösterimi Şekil 4.8'de verilmiştir.



Şekil 4.8. LOF algoritmasıyla bulunan aykırı değerler grafiği

Şekil 4.8'de görüldüğü gibi X_1, X_2 ve X_3 değişkenine eklenen 29 adet aykırı değer LOF algoritmasıyla bulunmuştur.

Kümeleme Tabanlı Aykırı Değer Belirleme Yöntemlerine Ait Bulgular

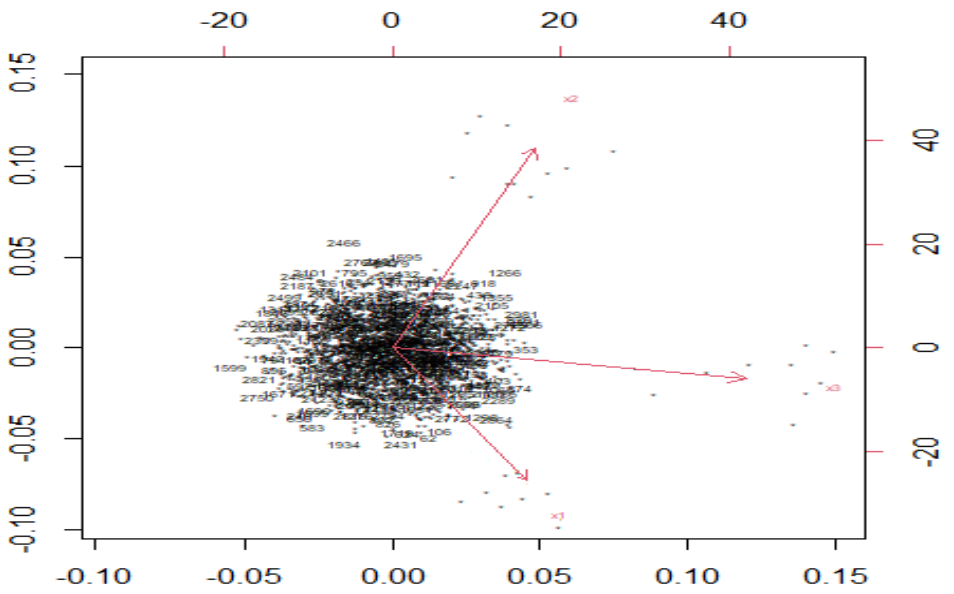
Kümeleme tabanlı aykırı değer belirleme yöntemlerinden hiyerarşik olmayan (bölümleyici) kümeleme yönteminde bulunan k-ortalamalar (K-means) algoritması çalışmada kullanılmıştır.

Bölümleyici kümeleme analizi sonucunda oluşan iki kümenin merkezleri Çizelge 4.12'de verilmiştir.

Çizelge 4.12. Oluşan iki kümenin merkezleri

Küme/Değişken	X_1	X_2	X_3
1	0.5981283	-0.4000211	0.3639226
2	-0.5760218	0.4618445	-0.4411762

Gözlemlerin ait oldukları küme merkezleri belirlendikten sonra gözlemlerin uzaklıkları bulunmuştur. Merkeze uzaklığı fazla olan gözlem, aykırı değer olarak kabul edilmiştir. Bu aykırı değerler Şekil 4.9'da gösterilmiştir.



Şekil 4.9. K-ortalamalar algoritmasıyla bulunan aykırı değerler grafiği

Şekil 4.9'da görüldüğü gibi eklenen 27 adet aykırı değer k-ortalamalar algoritmasıyla bulunmuştur.

Sapma Tabanlı Aykırı Değer Belirleme Yöntemlerine Ait Bulgular

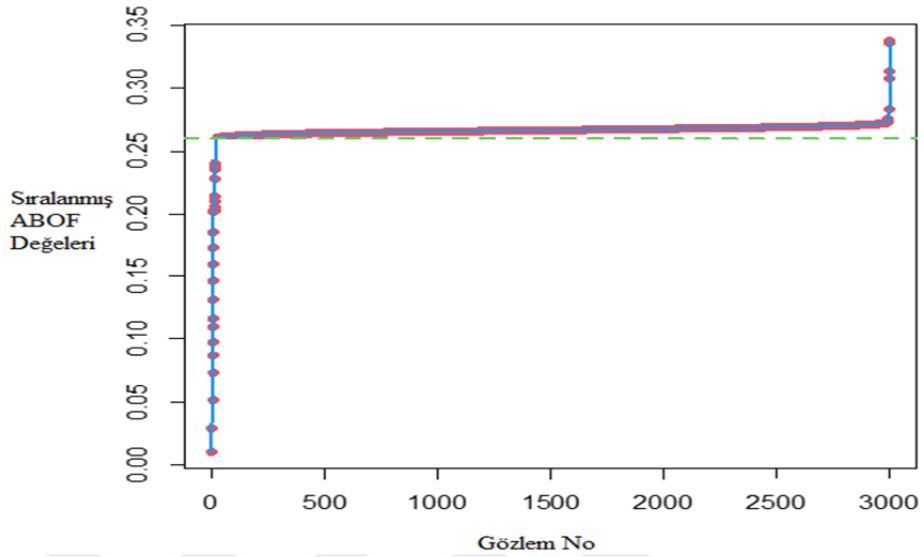
Sapma tabanlı aykırı değer belirleme yöntemlerinden açı tabanlı aykırı değer belirleme yönteminde bulunan ABOD algoritması çalışmada kullanılmıştır. ABOF değerleri bulunurken "abodOutlier" (Jimenez, 2015) paketinden yararlanılmıştır.

Eşitlik (3.28) kullanılarak ABOF değerleri bulunmuştur. Ardından ABOF değerleri küçükten büyüğe doğru sıralanmıştır ve Çizelge 4.13'de bu değerler verilmiştir.

Çizelge 4.13. Sıralanmış ABOF değerleri

Sıra No	ABOF Değerleri
1	0.001
2	0.032
3	0.057
.	.
.	.
.	.
2999	0.3127
3000	0.3425

Çizelge 4.13'de verilen ABOF değerlerinden ilk 28 gözlemin diğer gözlemlerden daha küçük olduğu görülmektedir. Dolayısıyla 28 gözlem, aykırı değer olarak kabul edilmektedir. ABOF değerlerinin grafiksel gösterimi Şekil 4.10'da verilmiştir.



Şekil 4.10. Sıralanmış ABOF değerleri grafiği

Şekil 4.10'da görüldüğü gibi 0.27 kesme noktası olarak belirlenmiştir. Bu kesme noktasından küçük olan gözlemler 28 adettir ve bunlar aykırı değerlerdir.

Yalıtım Tabanlı Aykırı Değer Belirleme Yöntemlerine Ait Bulgular

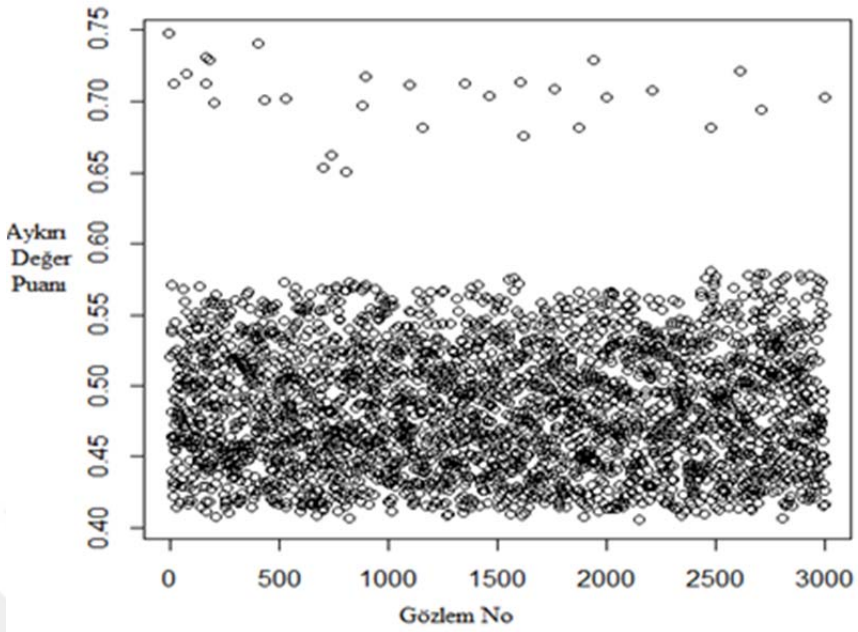
Yalıtım tabanlı aykırı değer belirleme yöntemlerinden biri olan yalıtım orman yöntemi (iForest) çalışmada kullanılmıştır. Aykırı değer puanları bulunurken “isofor" (Graves ve Drozdov, 2019) paketinden yararlanılmıştır.

(3.30) nolu eşitlik kullanılarak aykırı değer puanları bulunmuştur ve Çizelge 4.14'de bu puanlar verilmiştir.

Çizelge 4.14. iForest yöntemi ile bulunan aykırı değer puanları

Gözlem No	Aykırı Değer Puanları
1	0.7433011
2	0.7196945
3	0.4395645
.	
.	
.	
2999	0.4262840
3000	0.7031633

Bu aykırı değerlerin grafiksel gösterimi Şekil 4.11'de verilmiştir.

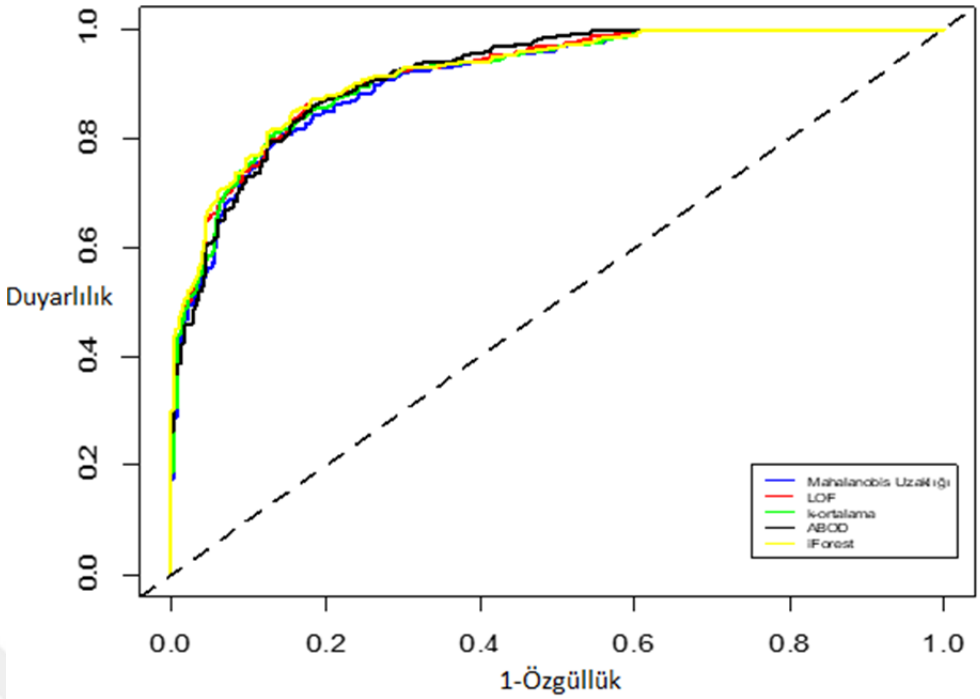


Şekil 4.11. iForest algoritmasıyla bulunan aykırı değerler grafiği

iForest algoritmasında aykırı değer puanlarına bakarak aykırı değerleri belirlemek için doğru eşik, her uygulama için farklıdır (Larionov, 2022). Liu ve ark. (2008) aykırı değer puanının 0.50'ye yakın olması durumunda, o gözlemin normal değer olduğunu bildirmişlerdir. Şekil 4.11'de görüldüğü gibi çoğu aykırı değer puanları 0.50 veya 0.50'nin yakınında toplanmıştır. Diğer gözlemlerin aykırı değer puanı bu gözlemlere göre oldukça uzaktır. Bu nedenle bu veri setinin eşik değeri 0.65 olarak belirlenmiştir. Bu eşik değeri aşan 30 gözlem vardır. Bu durumda eklenen 30 aykırı değer iForest algoritması ile bulunmuştur.

Yapay Veri Setinde Aykırı Değer Belirleme Yöntemlerinin Karşılaştırılması

Standart normal dağılıma sahip veri setinde (n=3000) Mahalanobis, LOF, k-ortalamalar, ABOD ve iForest yöntemlerinin karşılaştırılması için ROC eğrisi Şekil 4.12'de verilmiştir.



Şekil 4.12. Aykırı değer belirleme yöntemlerinin ROC eğrisi ile karşılaştırılması

Şekil 4.12'de iForest algoritmasının eğrisi sol üst köşeye en yakındır. Ayrıca bu algoritmanın daha yüksek AUROC değerine sahip olduğu görülmektedir. Bunu LOF algoritması izlemektedir. iForest algoritmasıyla bulunan aykırı değerlerin veri setinden temizlenmesinden sonra oluşturulan modelin en iyi uyumu sağladığı görülmektedir.

4.1.3. Aykırı Değer İçermeyen Yapay Veri Setinde Lojistik Regresyon Analizine Ait Bulgular

Aykırı değerler temizlendikten sonra aykırı değer içermeyen veri seti, eğitim (train) ve kontrol (test) veri seti olarak iki gruba ayrılmıştır. Eğitim veri seti kullanarak bulunan değişkenlerin tahmin değerleri ve ilgili istatistikler Çizelge 4.15'de verilmiştir.

Çizelge 4.15. Aykırı değer içermeyen yapay veri setinin değişkenleri ve ilgili test istatistikleri

Değişkenler	Tahmin Değeri	Standart Hata	Wald	p
Sabit	0.37367	0.06636	5.631	1.79e-08 ***
X_1	-0.01784	0.06621	-0.269	0.7876
X_2	2.98231	0.13070	22.818	< 2e-16 ***
X_3	0.15265	0.06755	2.260	0.0238 *

Önem seviyeleri: *: $p<0.05$, **:$p<0.01$

Çizelge 4.15'de görüldüğü gibi parametre önem testlerinden biri olan Wald testine göre X_2 ve X_3 değişkeni önemli bulunmuştur ($p<0.05$).

Bulunan lojistik regresyon modeli, $\ln(\frac{\pi(X)}{1-\pi(X)}) = 0.37367 - 0.01784X_1 + 2.98231X_2 + 0.15265X_3$ veya $\frac{\pi(X)}{1-\pi(X)} = e^{0.37367-0.01784X_1+2.98231X_2+0.15265X_3}$ şeklindedir.

Modelin uyum iyiliği testlerinden biri olan alıcı işletim karakteristiği eğrisini (ROC) çizebilmek için öncelikle kesme noktası belirlenmiştir ve bu kesme noktası için tahmin değerleri kullanarak yanılgı matrisi bulunmuştur. Modele ait optimum kesme noktası 0.4398 olarak hesaplanmıştır ve bulunan yanılgı matrisi Çizelge 4.16'da verilmiştir.

Çizelge 4.16. Aykırı değer içermeyen yapay veri setinin yanılgı matrisi

	Gözlenen pozitif	Gözlenen negatif	Toplam
Tahmin edilen pozitif	432 (%48.43)	76(%8.52)	508 (%56.95)
Tahmin edilen negatif	46 (%5.15)	338 (%37.90)	384 (%43.05)
Toplam	478(%53.58)	414 (%46.42)	892 (%100)

Çizelge 4.16'da verilen değerler kullanılarak duyarlılık, özgüllük, 1-duyarlılık, 1-özgüllük, doğruluk oranı ve sınıflama hatası aşağıdaki gibi hesaplanmıştır.

Duyarlılık (Doğru pozitif oranı) = $\frac{432}{432+46} = 0.903$

Özgüllük (Doğru negatif oranı) = $\frac{338}{338+76} = 0.813$

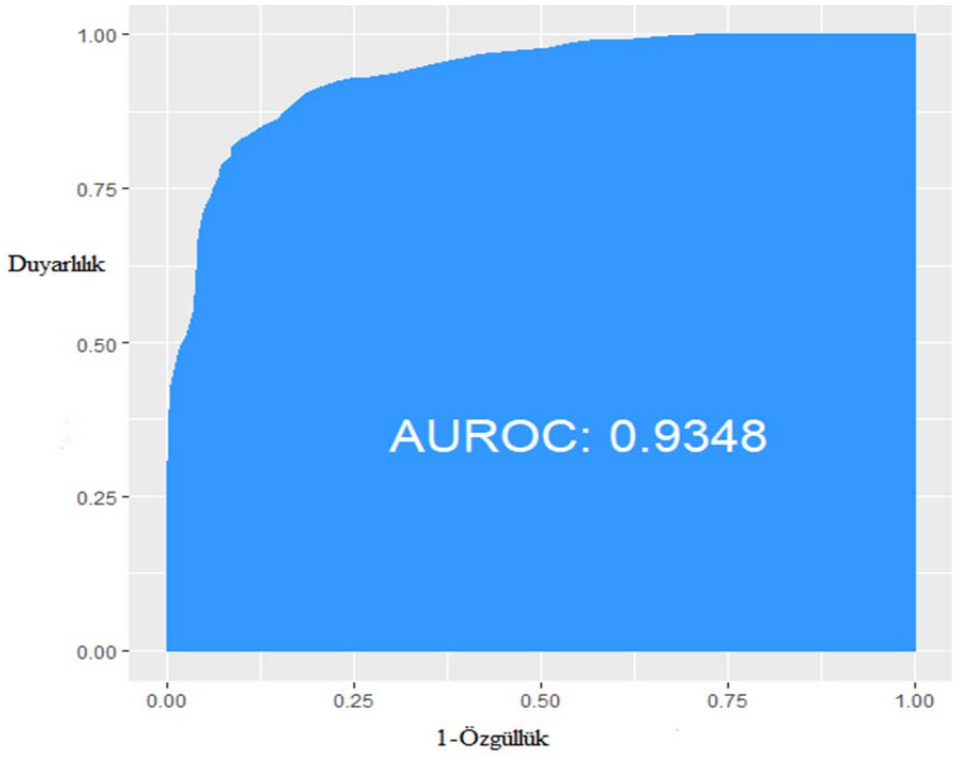
1-Duyarlılık (Yanlış negatif oranı) = $\frac{46}{46+432} = 0.096$

1-Özgüllük (Yanlış pozitif oranı) = $\frac{76}{76+338} = 0.183$

Doğruluk oranı= $\frac{432+338}{432+76+46+338} = 0.8633$

Sınıflama hatası = 1-0.8633 = 0.1367

Test veri seti kullanarak çizdirilen ROC eğrisi Şekil 4.13'de verilmiştir. ROC eğrisi çizdirilirken yukarıda bulunan 1-özgüllük ve duyarlılık değerleri kullanılmıştır.



Şekil 4.13. Aykırı değer içermeyen yapay veri setinin ROC eğrisi

Şekil 4.13'de ROC eğrisi, sol üst köşeye yakındır. Böylelikle eğrinin altında kalan alan (AUROC=0.9348) daha fazladır. Bu da modelin uyumunun iyi olduğunu göstermektedir.

Modelin yeterliliğinde kullanılan sahte R^2 değerleri Çizelge 4.17'de verilmiştir.

Çizelge 4.17. Aykırı değer içermeyen yapay veri setinin Sahte R^2 değerleri

McFadden R^2	Cox-Snell R^2	Nagelkerke R^2
0.4811433	0.4854702	0.6484216

Çizelge 4.17'de verilen McFadden R^2 değerine göre bağımlı değişkendeki toplam değişimin %48.11'i, Cox-Snell R^2 değerine göre bağımlı değişkendeki toplam değişimin %48.54'ü, Nagelkerke R^2 değerine göre bağımlı değişkendeki toplam değişimin %64.84'ü bağımsız değişkenler ile açıklanmaktadır.

4.1.4. Aykırı Değer İçeren ve İçermeyen Yapay Veri Setinin Karşılaştırılması

Bölüm 4.1.1 ile Bölüm 4.1.3'de incelenen aykırı değerli ve aykırı değer içermeyen veri setinde bulunan sonuçlar bu bölümde özetlenmiştir.

Aykırı değer içeren ve aykırı değer içermeyen veri setine ait parametre tahmin değerleri ve ilgili istatistikler Çizelge 4.18'de verilmiştir.

Çizelge 4.18. Yapay veri setinde değişkenler ve ilgili test istatistiklerinin karşılaştırılması

Aykırı Değer İçeren Veri Seti					**Aykırı Değer İçermeyen Veri Seti**				
Parametre Tahmini					Parametre Tahmini				
Değişkenler	Tahmin Değeri	Standart Hata	Wald	p	Değişkenler	Tahmin Değeri	Standart Hata	Wald	p
Sabit	0.28728	0.06385	4.499	6.82e-06 ***	Sabit	0.37367	0.06636	5.631	1.79e-08 ***
X_1	-0.06891	0.05825	-1.183	0.237	X_1	-0.01784	0.06621	-0.269	0.7876
X_2	2.81252	0.12298	22.869	< 2e-16 ***	X_2	2.98231	0.13070	22.818	<2e-16 ***
X_3	0.05749	0.06024	0.954	0.340	X_3	0.15265	0.06755	2.260	0.0238 *

Önem seviyeleri: *: p<0.05, **:p<0.01

Çizelge 4.18'de görüldüğü gibi iki veri setinde parametre tahmin edilirken tahmin değerlerinde farklılıklar görülmektedir. Bu da modelin kurulmasında farklılığa sebep olmaktadır ve yorumlamayı değiştirmektedir. X_3 değişkeni aykırı değer içeren veri setinde istatistiki olarak önemsizken; aykırı değer içermeyen veri setinde 0.05 önem seviyesinde istatistiki olarak önemli bulunmuştur.

Aykırı değer içeren ve aykırı değer içermeyen veri setine ait Sahte R^2 değerleri Çizelge 4.19'da verilmiştir.

Çizelge 4.19. Yapay veri setinde Sahte R^2 değerlerinin karşılaştırılması

Aykırı Değer İçeren Yapay Veri Seti			**Aykırı Değer İçermeyen Yapay Veri Seti**		
Sahte R^2			Sahte R^2		
McFadden R^2	Cox-Snell R^2	Nagelkerke R^2	McFadden R^2	Cox-Snell R^2	Nagelkerke R^2
0.4567747	0.4676844	0.6248164	0.4811433	0.4854702	0.6484216

Çizelge 4.19'da görüldüğü gibi iki veri setinde kurulan modelin uyum iyiliğinin karşılaştırılmasında kullanılan Sahte R^2 değerleri değişiklik göstermiştir. McFadden R^2, Cox-Snell R^2 ve Nagelkerke R^2 değerleri aykırı değer içermeyen veri setinde daha yüksektir. Bunun anlamı, aykırı değer içermeyen veri seti ile kurulan model daha iyi uyum göstermiştir.

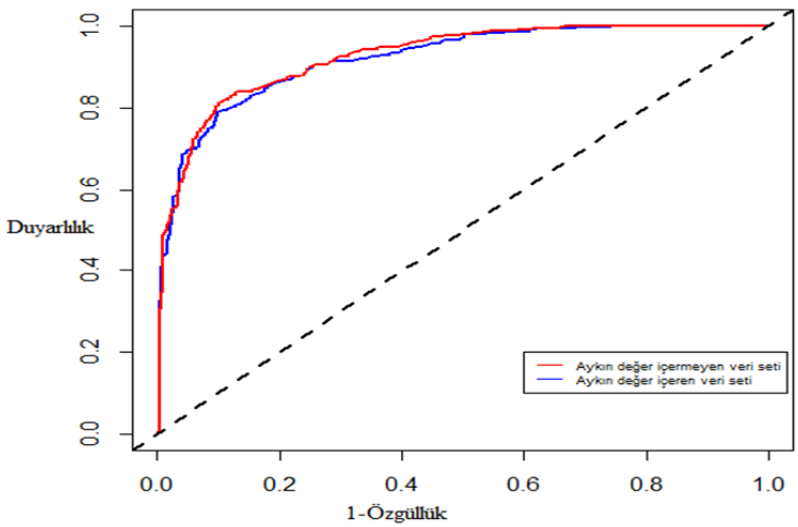
Aykırı değer içeren ve içermeyen veri setlerine ait yanılgı matrisleri Çizelge 4.20'de verilmiştir.

Çizelge 4.20. Yapay veri setinde sınıflama performanslarının karşılaştırılması

Aykırı Değer İçeren Yapay Veri Seti				**Aykırı Değer İçermeyen Yapay Veri Seti**			
	Gözlenen pozitif	Gözlenen negatif	Toplam		Gözlenen pozitif	Gözlenen negatif	Toplam
Tahmin edilen pozitif	385 (%42.73)	48 (%5.33)	433 (%48.06)	Tahmin edilen pozitif	432 (%48.43)	76 (%8.52)	508 (%56.95)
Tahmin edilen negatif	98 (%10.87)	370 (%41.07)	468 (%51.94)	Tahmin edilen negatif	46 (%5.15)	338 (%37.90)	384 (%43.05)
Toplam	483 (%53.60)	418 (%46.40)	901 (%100)	Toplam	478 (%53.58)	414 (%46.42)	892 (%100)
Doğruluk	%83.80			Doğruluk	%86.33		

Çizelge 4.20'de görüldüğü gibi aykırı değer içermeyen veri seti ile kurulan modelin doğruluk oranı daha yüksektir (%86.33).

Aykırı değer içeren ve içermeyen yapay veri seti ile kurulan modellerin ROC eğrisi ile karşılaştırılması Şekil 4.14'de verilmiştir.



Şekil 4.14. Yapay veri setinde ROC eğrilerinin karşılaştırılması

Şekil 4.14'de de görüldüğü gibi aykırı değer içermeyen veri setinin ROC eğrisi, sol üst köşeye daha yakındır. Böylelikle eğrinin altında kalan alan (AUROC) daha fazladır. Bu da aykırı değer içeren veri setine kıyasla, aykırı değer içermeyen veri seti ile kurulan modelin uyumunun daha iyi olduğunu göstermektedir.

4.2. Gerçek Veri Setine Ait Bulgular

Çalışmada, Zhao ve ark. (2011) tarafından elde edilen bir veri seti kullanılmıştır. Bu veri setinde bulunan 4 değişkenin içerikleri Çizelge 4.21'de verilmiştir.

Çizelge 4.21. Gerçek veri setinin içeriği

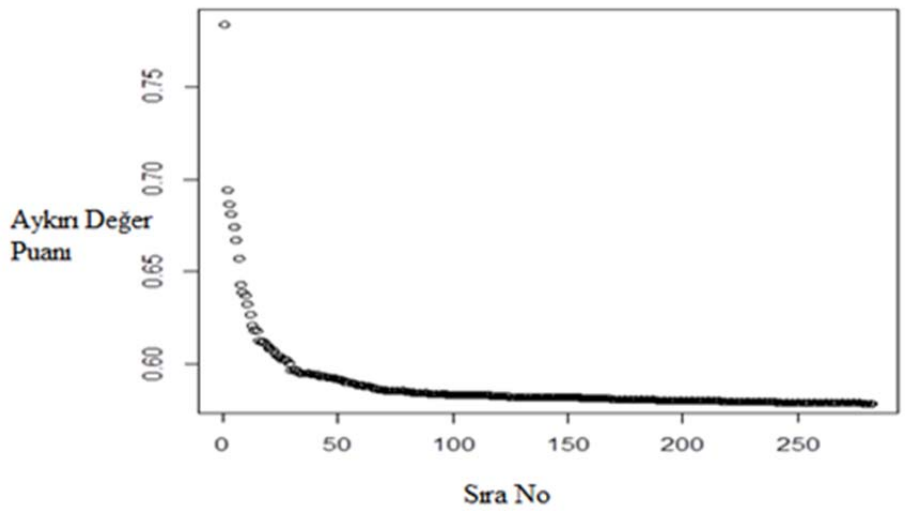
N: 282
X_1: Tohum uzunluğu (Sürekli değişken)
X_2: Tohum genişliği (Sürekli değişken)
X_3: Tohum hacmi (Sürekli değişken)
Y: Yaprak tüylenme (İki kategorili değişken, tüylenme yok:0, tüylenme var:1)

Yapay veri setinde iForest algoritmasıyla bulunan aykırı değerler veri setinden temizlenmesinden sonra oluşturulan modelin en iyi uyumu sağladığı görülmüştür (Şekil 4.12). Bu çalışmada kullanılan algoritmalar içerisinden iyi performans gösteren algoritma iForest algoritması olduğu için gerçek veri setinde iForest algoritması kullanarak aykırı değerler bulunmuştur. Aykırı değer puanları Çizelge 4.22'de verilmiştir.

Çizelge 4.22. Gerçek veri setinde iForest yöntemi ile bulunan aykırı değer puanları

Sıra No	Gözlem No	Aykırı Değer Puanları
1	164	0.783
2	125	0.693
3	144	0.686
.	.	.
.	.	.
.	.	.
281	140	0.582
282	160	0.582

Çizelge 4.22'de verilen aykırı değer puanlarının grafiği Şekil 4.15'de verilmiştir.



Şekil 4.15. iForest algoritması ile bulunan aykırı değerler

Şekil 4.15'de çoğu aykırı değer puanı 0.50 ile 0.65 arasında yoğunlaşmıştır. Bu nedenle bu veri setinin eşik değeri 0.65 olarak belirlenmiştir. Bu eşik değeri aşan 7 gözlem aykırı değer olarak adlandırılmıştır. Bu gözlemler 164, 125, 144, 269, 219, 249 ve 88 nolu gözlemlerdir.

4.2.1. Aykırı Değer İçeren/İçermeyen Gerçek Veri Setinin Karşılaştırılması

Aykırı değer içeren ve aykırı değer içermeyen (iForest algoritmasıyla bulunan aykırı değerler veri setinden çıkarıldıktan sonra) gerçek veri setinde lojistik regresyon analizi ile model tahmin edilmiştir. Tahmin edilen modelin uyum iyiliği ve yeterlilik ölçütleri incelenmiştir.

Aykırı değer içeren ve aykırı değer içermeyen gerçek veri setine ait parametre tahmin değerleri ve ilgili istatistikler Çizelge 4.23'de verilmiştir.

Çizelge 4.23. Gerçek veri setinde değişkenler ve ilgili test istatistiklerinin karşılaştırılması

Aykırı Değer İçeren Veri Seti					**Aykırı Değer İçermeyen Veri Seti**				
Parametre Tahmini					Parametre Tahmini				
Değişkenler	Tahmin Değeri	Standart Hata	Wald	p	Değişkenler	Tahmin Değeri	Standart Hata	Wald	p
Sabit	-11.730	7.730	-1.600	0.1095	Sabit	-10.906	8.014	-1.361	0.1736
X_1	1.531	1.065	1.438	0.1504	X_1	1.428	1.156	1.236	0.2165
X_2	11.071	5.463	2.027	0.0427 *	X_2	11.474	5.787	1.983	0.0474 *
X_3	-14.730	8.305	-1.774	0.0761 .	X_3	-15.211	8.801	-1.728	0.0839 .

Önem seviyeleri: *: $p<0.05$, **:$p<0.01$

Çizelge 4.23'de görüldüğü gibi iki veri setinde parametre tahmin edilirken tahmin değerlerinde farklılıklar görülmektedir. Bu da modelin kurulmasında farklılığa sebep olmaktadır ve yorumlamayı değiştirmektedir.

Aykırı değer içeren ve aykırı değer içermeyen gerçek veri setine ait Sahte R^2 değerleri Çizelge 4.24'de verilmiştir.

Çizelge 4.24. Gerçek veri setinde Sahte R^2 değerlerinin karşılaştırılması

Aykırı Değer İçeren Veri Seti			**Aykırı Değer İçermeyen Veri Seti**		
Sahte R^2			Sahte R^2		
McFadden R^2	Cox-Snell R^2	Nagelkerke R^2	McFadden R^2	Cox-Snell R^2	Nagelkerke R^2
0.1022	0.077	0.1419	0.1248	0.095	0.1727

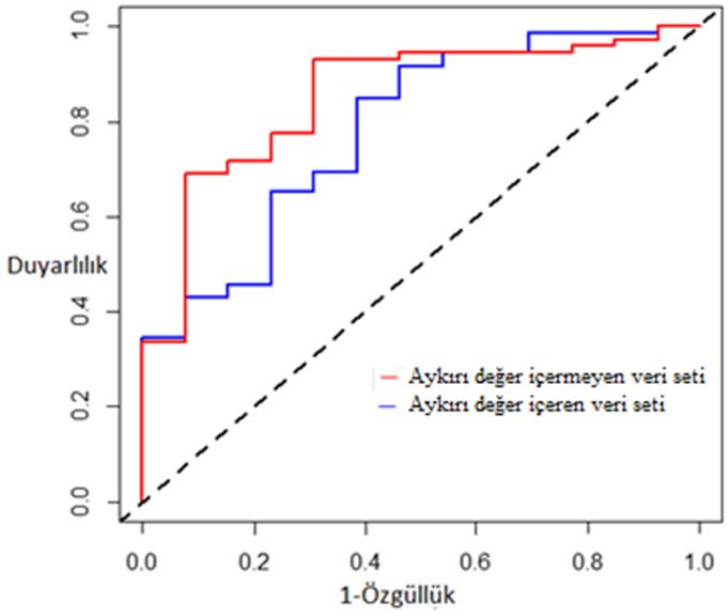
Çizelge 4.24'de görüldüğü gibi iki veri setinde kurulan modelin uyum iyiliğinin karşılaştırılmasında kullanılan Sahte R^2 değerleri değişiklik göstermiştir. McFadden R^2, Cox-Snell R^2 ve Nagelkerke R^2 değerleri aykırı değer içermeyen veri setinde daha yüksektir. Aykırı değer içeren ve içermeyen gerçek veri setine ait yanılgı matrisleri Çizelge 4.25'de verilmiştir.

Çizelge 4.25. Gerçek veri setinde sınıflama performanslarının karşılaştırılması

Aykırı Değer İçeren Veri Seti				**Aykırı Değer İçermeyen Veri Seti**			
	Gözlenen pozitif	Gözlenen negatif	Toplam		Gözlenen pozitif	Gözlenen negatif	Toplam
Tahmin edilen pozitif	71 (%83.53)	9 (%10.59)	80 (%94.12)	Tahmin edilen pozitif	67 (%79.77)	6 (%7.14)	73 (%86.91)
Tahmin edilen negatif	1 (%1.18)	4 (%4.70)	5 (%5.88)	Tahmin edilen negatif	4 (%4.76)	7 (%8.33)	11 (%13.09)
Toplam	72 (%84.71)	13 (%15.29)	85 (%100)	Toplam	71 (%84.53)	13 (%15.47)	84 (%100)
Doğruluk	%88.23			Doğruluk	%88.80		

Çizelge 4.25'de görüldüğü gibi aykırı değer içermeyen veri seti ile kurulan modelin doğruluk oranı daha yüksektir (%88.80).

Aykırı değer içeren ve içermeyen gerçek veri seti ile kurulan modellerin ROC eğrisi ile karşılaştırılması Şekil 4.16'da verilmiştir.



Şekil 4.16. Gerçek veri setinde ROC eğrilerinin karşılaştırılması

Şekil 4.16'da görüldüğü gibi aykırı değer içermeyen gerçek veri setinin ROC eğrisi, sol üst köşeye daha yakındır. Böylelikle eğrinin altında kalan alan (AUROC) daha fazladır. Bu da aykırı değer içeren veri setine kıyasla, aykırı değer içermeyen veri seti ile kurulan modelin uyumunun daha iyi olduğunu göstermektedir.

5. SONUÇLAR VE ÖNERİLER

Lojistik regresyon analizinde aykırı değerleri ve etkili gözlemleri belirlerken bazı klasik yöntemler kullanılmaktadır. Bu yöntemler bu çalışmada kullanılan yapay veri setinde incelendiğinde Pearson hatalara göre birkaç gözlem aykırı değer olarak görülmektedir (Şekil 4.4). Cook uzaklığı ve Şapka değerleri incelendiğinde ise 785, 1249, 1486, 2486 nolu gözlemler etkili gözlem olarak belirlenmiştir (Şekil 4.5). Student türü hatalar, şapka değerleri ile Cook uzaklığının birlikte verildiği grafikte de yine bu 4 gözlemin Cook uzaklığının en fazla olduğu görülmektedir (Şekil 4.6). Bu değerler, eklenen aykırı değerlerdendir. Ancak diğer aykırı değerler bu yöntemlerle bulunamamıştır. Görülmektedir ki bu yöntemler aykırı değerleri bulmada yetersiz kalmaktadır. Bu nedenle özellikle veri madenciliğinde kullanılan aykırı değer belirleme yöntemleri ile aykırı değerleri belirlemek, bu konuda çalışan araştırıcılara önerilmektedir.

Bu çalışmada, veri madenciliğinde kullanılan model tabanlı, yakınlık tabanlı, kümeleme tabanlı, sapma tabanlı ve yalıtım tabanlı aykırı değer belirleme yöntemlerinin performansları karşılaştırıldığında yalıtım tabanlı aykırı değer belirleme yöntemlerinden biri olan iForest algoritmasının en iyi performansa sahip olduğu belirlenmiştir. Bu algoritma ile eklenen tüm aykırı değerler bulunmuştur. iForest algoritmasını performans açısından LOF algoritması izlemiştir. Gao (2019), Gnat (2020), Negi (2020), Kiruthika ve Sowmyarani (2020), Vijayakumar ve ark. (2020) yaptıkları çalışmada iForest algoritmasının LOF algoritmasından daha iyi olduğunu bildirmişlerdir. Susto ve ark. (2017), hassasiyet açısından iForest algoritmasının ABOD algoritmasından daha iyi performans gösterdiğini belirtmişlerdir. Bu çalışmada da iForest algoritması ile 30 adet aykırı değer bulunmuşken; ABOD algoritmasıyla 28 adet aykırı değer bulunmuştur. Bu çalışmadan iForest algoritmasının ABOD algoritmasından daha iyi performans gösterdiği sonucuna varılabilmektedir. Ayrıca ABOD algoritmasıyla eklenen tüm aykırı değerlerin bulunamaması, bu yöntemin maskeleme etkisine karşı duyarlı olduğunu ve iyileştirme çalışmaları gerektiğini göstermektedir (Cebeci, 2020). Buna ek olarak Kriegel ve ark. (2008), büyük veri setlerinde ABOD algoritması ile doğrudan aykırı değerleri hesaplamak zaman açısından oldukça maliyetli olduğunu bildirmiştir. Bu çalışmada da ABOD algoritmasında kullanılan "complete" metodu ile sonuçlara ulaşmak iki saatten fazla sürmüştür. Bu durum zaman açısından maliyetin yüksek olduğunu göstermektedir. Ayrıca tüm yöntemlerin performanslarının karşılaştırmasında kullanılan ROC eğrisine göre, sol üst köşeye en yakın, AUROC değeri en yüksek olan iForest algoritmasıdır (Şekil 4.12). Bu da iForest algoritması ile bulunan aykırı değerlerin veri setinden temizlenmesinden sonra kurulan modelin en iyi uyumu sağladığını göstermektedir. Domingues ve ark. (2018) içerisinde LOF ve ABOD algoritmasının da bulunduğu birçok algoritmanın karşılaştırıldığı çalışmada, iForest algoritmasının verimliliğinin en yüksek olduğunu bildirmişlerdir.

Bu çalışmada ayrıca aykırı değerlerin model tahminine ve model uyumunun iyiliğine etkisi incelenmiştir. Aykırı değer içeren ve içermeyen veri setinde model kurulurken bulunan tahmin

değerlerinde farklılıklar görülmüştür (Çizelge 4.18). Aykırı değer içeren veri setinde sadece X_2 bağımsız değişkeni istatistiki olarak önemli bulunmuşken ($p<0.001$), aykırı değer içermeyen veri setinde X_3 bağımsız değişkeni de önemli bulunmuştur ($p<0.05$). ROC eğrisi ile kurulan modellerin uyumu incelendiğinde aykırı değer içermeyen veri setinin AUROC değerinin daha yüksek olduğu görülmüştür (Şekil 4.14). Aykırı değer içermeyen veri seti ile kurulan modelin doğruluğu %86.33 iken, aykırı değer içeren veri seti ile kurulan modelin doğruluğu %83.80'dir (Çizelge 4.20). Aykırı değer içermeyen veri setinde Sahte R^2 değerlerinden McFadden, Cox-Snell ve Nagelkerke R^2 değerleri aykırı değer içeren veri setine göre daha yüksek çıkmıştır (Çizelge 4.19). Kısacası, aykırı değer içeren ve aykırı değer içermeyen veri setinde yapılan analizlerde farklılıklar görülmektedir. Osborne ve Amy (2004), çalışmasında da aykırı değerlerin basit analizleri bile ne kadar etkileyebileceğini benzer şekilde göstermiştir.

Yapay veri setinde en iyi performansa sahip algoritma iForest olarak bulunduğu için Zhao ve ark. (2011) tarafından elde edilen veri setinde iForest algoritması ile aykırı değerler belirlenmiştir. Bu algoritma ile 7 adet aykırı değer bulunmuştur.

Ayrıca aykırı değer içeren ve içermeyen gerçek veri setinin karşılaştırılması yapılmıştır. Buna göre model kurulurken bulunan tahmin değerlerinde farklılıklar görülmüştür (Çizelge 4.23). Aykırı değer içermeyen gerçek veri setinde bulunan tüm Sahte R^2 değerlerinin aykırı değer içeren veri setine göre daha yüksek çıktığı görülmüştür (Çizelge 4.24). Ayrıca aykırı değer içermeyen veri seti ile kurulan modelin doğruluk oranı, aykırı değer içeren veri seti ile kurulan modele göre daha yüksek çıktığı görülmüştür (Çizelge 4.25).

Sonuç olarak, aykırı değerler, modelin kurulmasında, katsayıların tahminlerinde ve yorumlamalarda, modelin uyum iyiliğinde farklılıklara sebep olmasından dolayı analizlerden önce aykırı değerlerin incelenmesi araştırıcılara önerilmektedir. Çünkü bu çalışmada, aykırı değerleri incelemeden araştırmaya başlanması durumunda lojistik regresyon analizi kullanarak modelleme ve sınıflama yapmanın güvenilir olmadığı ortaya çıkmıştır. Ayrıca aykırı değerleri belirlerken iForest algoritmasının daha iyi performans göstermesi sebebiyle araştırıcılara bu algoritmayı kullanmaları tavsiye edilmektedir.

KAYNAKLAR

Abed Mohammed, S., 2013. Comparing Some Robust Methods With OLS Method in Multiple Regression With Application. Journal of Concrete & Applicable Mathematics, 11(1):230-245.

Agresti, A., 1996. An Introduction to Categorical Data Analysis (Vol. 135). Wiley: New York.

Ahn, H. and James, R.T., 1999. Outlier Detection in Phosphorus Dry Deposition Rates Measured in South Florida. Atmospheric Environment, 33:5123-5131.

Akçay, A., 2009. Broiler Coccidiosis'inde Risk Faktörlerinin Lojistik Regresyon Analizi ile Belirlenmesi. Doktora Tezi, Ankara Üniversitesi, Ankara.

Akçay, A., Abay, M. ve Bekyürek, T., 2014. Süt Sığırlarında Gebelik Oranı Üzerine Etkili Bazı Faktörlerin Lojistik Regresyon Analizi ile Değerlendirilmesi. Hayvansal Üretim, 55(2):16-21.

Akdemir, Ç., 2016. Hilenin Veri Madenciliği ile Ortaya Çıakrtılması ve Perakende Sektöründe Bir Uygulama. Doktora Tezi, Marmara Üniversitesi, İstanbul.

Akın, H.B. ve Şentürk, E., 2012. Bireylerin Mutluluk Düzeylerinin Ordinal Lojistik Regresyon Analizi ile İncelenmesi. Öneri Dergisi, 10(37): 83-193.

Akküçük, U., 2011. Veri Madenciliği: Kümeleme ve Sınıflama Algoritmaları. Yalın Yayıncılık, İstanbul.

Aksaraylı, M. ve Saygın, Ö., 2011. Algılanan Hizmet Kalitesi ve Lojistik Regresyon Analizi ile Hizmet Tercihine Etkisinin Belirlenmesi. Dokuz Eylül Üniversitesi, Sosyal Bilimler Enstitüsü Dergisi, 13(1):21-37.

Aleng, N.A., Naing, N.N., Mohamed, N. and Mokhtar, K., 2017. Outlier Detection Based on Robust Parameter Estimates. International Journal of Applied Engineering Research, 12(23):13429-13434.

Al-Ghamdi, A.S., 2002. Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity. Accident Analysis & Prevention, 34(6):729-741.

Ali, S., Wang, G., Cottrell, R.L. and Anwar, T., 2017. Detecting Anomalies from End-to-end Internet Performance Measurements (PingER) Using Cluster Based Local Outlier Factor. IEEE International Symposium on Parallel and Distributed Processing with Applications and IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), pp. 982-989.

Alpar, R., 2011. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. Detay Yayıncılık, 3. Baskı, Ankara.

Amer, M., 2011. Comparison of Unsupervised Anomaly Detection Techniques. Media Engineering and Technology German University in Cairo and Multimedia Analysis and Data Mining Competence Center German Research Center for Artificial Intelligence, Bachelor Thesis, Germany.

Amer, M. and Goldstein, M., 2012. Nearest-Neighbor and Clustering Based Anomaly Detection Algorithms for Rapidminer. In Proc. of the 3rd RapidMiner Community Meeting and Conference (RCOMM 2012), pp. 1-12.

Angen, Q., Skov, M.N., Chriel, M., Agger, J.F. and Bisgoard, M., 1996. A Retrospective Study On Salmonella Infection in Danish Broiler Flocks. Preventive Veterinary Medicine, 26:223-227.

Anonim, 2020. Yeni başlayanlar için A'dan Z'ye Makine Öğrenmesi. https://www.kizgibikodla.com/news/yeni-baslayanlar-icin-adan-zye-makine-ogrenmesi/ (Erişim Tarihi:06.01.2020).

Arıcan, E., 2010. Nitel Yanıt Değişkene Sahip Regresyon Modellerinde Tahmin Yöntemleri. Yüksek Lisans Tezi, Çukurova Üniversitesi, Adana.

Atakurt, Y., 1999. Lojistik Regresyon Analizi ve Tıp Alanında Kullanımına İlişkin Bir Uygulama. Ankara Üniversitesi, Tıp Fakültesi Mecmuası, 52(4):191-199.

Aydın, Y., 2006. Grafik Yöntemlerle Etkin Gözlemlerin ve Aykırı Değerlerin Tespiti. Yüksek Lisans Tezi, Ondokuz Mayıs Üniversitesi, Samsun.

Bahadır, 2013. Japon Bıldırcınlarında (Coturnix Coturnix Japonica) Canlı Ağırlık Performansına Ait Verilerin Aykırı Değerlerinin Belirlenmesi. Yüksek Lisans Tezi, Bingöl Üniversitesi, Bingöl.

Bahadır, E. ve Özdemir, A.Ş., 2013. Öğretmen Adaylarının Akademik Başarılarının Sınıflandırılmasında Lojistik Regresyon Analizi Yaklaşımı. Turkish Studies, 8(8):203-214.

Bayram, N., 2004. Multinominal Lojistik Regresyon Analizinin İstihdamdaki İşgücüne Uygulanması. İktisat Fakültesi Mecmuası, 54(2):61-75.

Bedrick, E.J. and Hill, J.R., 1990. Outlier Tests For Logistic Regression: A Conditional Approach. Biometrika, 77(4):815-827.

Bejaei, M., Wiseman, K. and Cheng, K.M., 2015. Developing Logistic Regression Models Using Purchase Attributes and Demographics to Predict The Probability of Purchases of Regular And Specialty Eggs. British Poultry Science, 56(4):425-435.

Ben-Gal, I., 2005. Outlier Detection. In Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA.

Bertizlioğlu, İ.N. ve Özgönenel, O., 2012. K-Ortalamalar Kümesi Yöntemi ile Sistem Çökmesi Olaylarının Sezinlenmesi. ELECO'2012 Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu, Bursa. https://www.emo.org.tr/ekler/1223dc143237f64_ek.pdf. (Erişim tarihi: 06.12.2021).

Billor, N. and Kıral, G., 2008. A Comparison of Multiple Outlier Detection Methods for Regression Data. Communications in Statistics-Simulation and Computation, 37(3):521-545.

Bircan, H., 2004. Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama. Kocaeli Üniversitesi, Sosyal Bilimler Enstitüsü Dergisi, 2:185-208.

Bisren, İ. ve Alkan, S., 2019. Denetimsiz Anomali Belirleme Yöntemleri ve Uygulamaları. Bitirme Projesi, Dokuz Eylül Üniversitesi, İzmir. https://istatistik.deu.edu.tr/tr/bitirme-projeleri/. (Erişim Tarihi:12.10.2020)

Blatná, D., 2006. Outliers in Regression. Trutnov, 30:1-6.

Bozkurt, B., 2011. Kredi ve Yurtlar Kurumunda Kalan Öğrencilerin Memnuniyet Derecelerinin Lojistik Regresyon Yöntemi ile Araştırılması: Edirne İli Örneği. Yüksek Lisans Tezi, Trakya Üniversitesi, Edirne.

Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J., 2000. LOF: Identifying Density-Based Local Outliers. In ACM Sigmod Record, 29(2):93-104.

Budak, İ., Şen, B. ve Yıldırım, M., 2013. Lojistik Regresyon ile Bilgisayar Ağlarında Anomali Tespiti. XV Akademik Bilişim Konferansı, pp. 813-818.

Calheiros, R.N., Ramamohanarao, K., Buyya, R., Leckie, C. and Versteeg, S., 2017. On the Effectiveness of Isolation-based Anomaly Detection in Cloud Data Centers. Concurrency and Computation: Practice and Experience, 29(18):1-12.

Cankurt, M., Miran, M. ve Şahin, A., 2010. Sığır Eti Tercihlerini Etkileyen Faktörlerin Belirlenmesi Üzerine Bir Araştırma: İzmir İli Örneği. Hayvansal Üretim, 51(2):16-22.

Cebeci, Z., 2020. Veri Biliminde R ile Veri Önişleme. Nobel Akademik Yayıncılık, Ankara.

Cebeci, Z., Tekeli, E., Tahtalı, Y., 2022. Tarım, Gıda ve Yaşam Bilimlerinde R ile Makine Öğrenmesi ve Veri Madenciliği. Atlas Akademik Basım Yayın Dağıtım, Yayın No: 4370. ISBN: 978-625-427-286-8. pp:152-157.

Chandola, V., Banerjee, A. and Kumar, V., 2009. Anomaly Detection: A Survey. ACM Computing Surveys (CSUR), 41(3):1-72.

Chatfield, C. and Collins, A., 1992. Introduction to Multivariate Analysis. Chapman & Hall, London.

Chatterjee, S. and Hadi, A.S., 1986. Influential Observations, High Leverage Points, and Outliers in Linear Regression. Statistical Science, 1(3):379-393.

Chatterjee, S. and Hadi, A.S., 2006. Regression Analysis by Example. Fourth Edition, Wiley-Interscience.

Chen, C., 2002. Robust Regression and Outlier Detection with the ROBUSTREG Procedure. In Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference, pp.265-27.

Cheng, Z., Zou, C. and Dong, J., 2019. Outlier Detection Using Isolation Forest and Local Outlier Factor. In Proceedings of The Conference on Research in Adaptive and Convergent Systems, pp. 161-168.

Compton, B.W., Rhymer, J.M. and McCollough, M., 2002. Habitat Selection By Wood Turtles (Clemmys İnsculpta): An Application Of Paired Logistic Regression. Ecology, 83(3):833-843.

Cook, R.D. and Weisberg, S., 1982. Residuals and Influence in Regression. Chapman and Hall, New York.

Cook, R.D., 1977. Detection of Influential Observation in Linear Regression. Technometrics, 19(1):15-18.

Coşkun, S. ve Kartal, M., 2004. Lojistik Regresyon Analizinin İncelenmesi ve Diş Hekimliğinde Bir Uygulaması. Cumhuriyet Üniversitesi, Diş Hekimliği Fakültesi Dergisi, 7(1):41-50.

Coşkuntuncel, O., 2009. Eğitimle İlgili Sapan Değer İçeren Veri Kümelerinde En Küçük Kareler ve Robust M Tahmin Edicilerin Karşılaştırılması. Mersin Üniversitesi, Eğitim Fakültesi Dergisi, 5(2):251-262.

Cox, D. R., 1958. The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2):215-232.

Cragg, J.G. and Uhler, R.S., 1970. The Demand for Automobiles. Canadian Journal of Economics/Revue Canadienne d'Economique, 3(3):386-406.

Çokluk, Ö., 2010. Lojistik Regresyon Analizi: Kavram ve Uygulama. Educational Science: Theory&Practice, 10(3):1357-1407.

Dağhan, G., 2011. Biyoeşdeğerlik Çalışmalarında Aykırı Değerlerin İncelenmesi. Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara.

Deb, A.B. and Dey, L., 2017. Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering. World Journal of Computer Application and Technology, 5(2):24-29.

Demir, N., Sancar, C., Demir, O., Aksakal, V. and Aksoy, A., 2016. Determining the Factors Affecting Farmers' Decision on Organic Livestock. Turkish Journal of Agriculture-Food Science and Technology, 4(4):313-317.

Ding, Z. and Fei, M., 2013. An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data Using Sliding Window. IFAC Proceedings Volumes, 46(20):12-17.

Dirican, A., 2001. Tanı Testi Performansının Değerlendirilmesi ve Kıyaslanması. Cerrahpaşa Tıp Dergisi, 32:25-30.

Doğan, Z., 1995. Basit Doğrusal Regresyon Modellerindeki Tek Bir Anormal Gözlemin Tespiti. Yüksek Lisans Tezi, Harran Üniversitesi, Şanlıurfa.

Domingues, R., Filippone, M., Michiardi, P. and Zouaoui, J., 2018. A Comparative Evaluation of Outlier Detection Algorithms: Experiments and Analyses. Pattern Recognition, 74:406-421.

Dunteman, G.H. and Ho, M.H.R., 2006. An Introduction to Generalized Linear Models. SAGE: USA.

Ekici, O. ve Yorulmaz, Ö., 2008. Sapan Gözlem İle Yapısal Kırılma Noktası İlişkisi ve Bunun Bayesyen Otoregresif Süreçle Tespiti. Doğuş Üniversitesi Dergisi, 9(2):146-157.

Elahi, M., Li, K., Nisar, W., Lv, X. and Wang, H., 2008. Efficient Clustering-Based Outlier Detection Algorithm for Dynamic Data Stream. In Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 5: 298-304.

Filzmoser, P. and Varmuza, K., 2017. Chemometrics: Multivariate Statistical Analysis in Chemometrics. Retrieved from https://CRAN.R-project.org/package=chemometrics.

Gandhi, R., 2018. Introduction to Machine Learning Algorithms: Logistic Regression. https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression cbdd82d81a36.(Erişim tarihi: 5.09.2022).

Gao, R., Zhang, T., Sun, S., and Liu, Z., 2019. Research and improvement of isolation forest in detection of local anomaly points. In Journal of Physics: Conference Series, 1237(5): 1-6. IOP Publishing. https://doi.org/10.1088/1742-6596/1237/5/052023.

Gardside, P.S. and Glueck, C.J., 1995. The Important Role of Modifiable Dietary and Behaviour Characteristic in the Causation and Prevention of Coronary Heart Disease Hospitalization and Mortality. Journal of American College of Nutrition, 14:71-79.

Girginer, N. ve Cankuş, B., 2008. Tramvay Yolcu Memnuniyetinin Lojistik Regresyon Analiziyle Ölçülmesi: Estram Örneği. Yönetim ve Ekonomi: Celal Bayar Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 15(1):181-193.

Gnat, S., 2020. Testing the Effectiveness of Outlier Detecting Methods in Property Classification. Real Estate Management and Valuation, 28(4):81-92.

Gogoi, P., Bhattacharyya, D., Borah, B., and Kalita, J.K., 2011. A Survey of Outlier Detection Methods in Network Anomaly Identification. The Computer Journal, 54(4):570–588.

Goldstein, M. and Uchida, S., 2016. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PloS One, 11(4):1-31.

Graves, E. and Drozdov, I., 2019. Zelazny7/isofor: Isolation Forest Anomaly Detection. Retrieved from https://github.com/Zelazny7/isofor.

Gurrapu, S., Batarseh, F. A., Wang, P., Sikder, M. N. K., Gorentala, N., & Gopinath, M., 2021. Deepag: Deep learning approach for measuring the effects of outlier events on agricultural production and policy. In 2021 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1-8). IEEE.

Hadi, A.S. and Imon, A.H.M.R, 2008. Identification of Multiple Outliers in Logistic Regression. Communications in Statistics-Theory and Methods, 37:1697–1709.

Hadi, A.S. and Simonoff, J.S., 1993. Procedures for the Identification of Multiple Outliers in Linear Models. Journal of the American Statistical Association, 88(424):1264-1272.

Halperin, M., Blackwelder, W.L. and Verter, J.I., 1971. Estimation of the Multivariate Logistic Rish Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches. Journal of Chronic Disease, 24:125-158.

Han, J., Pei, J. and Pei, J., 2012. Data Mining: Concepts and Techniques. Third Edition, Morgan Kaufmann Publishers Elsevier, USA.

Hawkins, D.M., 1980. Identification of Outliers (Vol. 11). Chapman and Hall, London.

He, Z., Xu, X., and Deng S., 2003. Discovering Cluster-Based Local Outliers. Pattern Recognition Letters, 24:1641-1650.

Hekimoğlu, S., Erenoğlu, R.C., and Kalina, J., 2009. Outlier Detection by Means of Robust Regression Estimators for Use in Engineering Science. Journal of Zhejiang University-Science A, 10(6):909-921.

Hoaglin, D.C. and Welsch, R.E., 1978. The Hat Matrix in Regression and ANOVA. The American Statistician, 32(1):17-22.

Hodge, V. and Austin, J., 2004. A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22(2):85-126.

Hofmann, M. and Klinkenberg, R., 2014. RapidMiner: Data Mining Use Cases and Business Analytics Applications. CRC Press.

Hu,Y., Murray,W. and Australia, Y.S., 2015. Rlof: R Parallel Implementation of Local Outlier Factor (LOF). Retrieved from https://CRAN.R-project.org/package=Rlof**.**

İyit, N. ve Genç, A., 2005. Lojistik Regresyon Analizi Yardımıyla Denekte Menopoz Evresine Geçişe İlişkin Bir Sınıflandırma Modelinin Elde Edilmesi. Selçuk Üniversitesi, Fen Edebiyat Fakültesi, Fen Dergisi, 1(25):19-28.

Jimenez, J., 2015. abodOutlier: Angle-Based Outlier Detection. Retrieved from https://CRAN.R-project.org/package=abodOutlier.

Kalaycı, İ. ve Ercan, T., 2018. Kablosuz Algılayıcı Ağ Verilerinde Histogram Temelli Aykırı Değer Skor Yöntemi Kullanarak Anomali Saptama. 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1-6.

Kamaladevi, M., Sekar, K.R., Venkataraman V. and Kannan, K., 2019. Outlier Detection in Imbalanced Data Classification. International Journal of Recent Technology and Engineering (IJRTE), 7(6):972-975.

Kannan, K.S. and Sampath, G., 2017. Identification of Outliers in Logistic Regression Based On Robust Deviance Component. Global and Stochastic Analysis, 4(1):71-76.

Karagöz, Y., Kıngır, S. ve Yıldız, M.S., 2010. İş Tatminini Etkileyen Faktörlerin Kriz Ortamındaki Etkisinin Lojistik Regresyon Analizi ile Belirlenmesi. Sosyal Ekonomik Araştırmalar Dergisi, 13:341-362.

Kaşko, Y., 2007. Çoklu Bağlantı Durumunda İkili (Binary) Lojistik Regresyon Modelinde Gerçekleşen I. Tip Hata ve Testin Gücü. Yüksek Lisans Tezi, Ankara Üniversitesi, Ankara.

Kaur, K. and Garg, A., 2016. Comparative Study of Outlier Detection Algorithms. International Journal of Computer Applications, 147(9):21-26.

Kaya, A., 1995. Kalite Kontrol Amaçlı Zaman Serilerinde Sapan Değer Analizi. Yüksek Lisans Tezi, Ege Üniversitesi, İzmir.

Kaya, H. ve Köymen, K., 2008. Veri Madenciliği Kavramı ve Uygulama Alanları. Fırat Üniversitesi, Doğu Araştırmaları Dergisi, 6(2):159-164.

Kıral, G. ve Billor, N., 2005. Bacon Yöntemine Dayalı Temel Bileşenler Analizi. Çukurova Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Dergisi, 9(1):1-21.

Kıral, G., Billor, N. ve Türkmen, A., 2013. Robust Kümeleme Yöntemi ile Grup Sapan Değerlerin Belirlenmesi. Ç.Ü. Sosyal Bilimler Enstitüsü Dergisi, 22(1):189-212.

Kinsel, M.L., Marsh, W.E., Ruegg, P.L. and Etherington, W.G., 1998. Risk Factors for Twinning in Dairy Cows. Journal of Dairy Science, 81:989-993.

Kiruthika, S. and Sowmyarani, C.N., 2020. Credit Card Fraud Detection using Machine Learning and Deployment of Model in Public Cloud as a Web Service. International Journal of Recent Technology and Engineering (IJRTE), 9(2):548-552. DOI:10.35940/ijrte.B3800.079220.

Kocabaş, E., 2014. Lojistik Regresyon ve Bankacılık Verileri Üzerine Bir Uygulama. Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, İstanbul.

Koğar, H., 2010. Farklı Örneklem Büyüklüklerinde Uç Değerlerle Baş Etme Yöntemlerinin Puanların Geçerlik ve Güvenirlik Kanıtları Üzerindeki Etkisi. Yüksek Lisans Tezi, Ankara Üniversitesi, Ankara.

Kontrimas, V. and Verikas, A., 2006. Tracking of Doubtful Real Estate Transactions by Outlier Detection Methods: A Comparative Study. Information Technology and Control, 35(2): 94-105.

Kriegel, H.P., Schubert, M. and Zimek, A., 2008. Angle-based Outlier Detection in High-dimensional Data. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 444-452.

Kuppusamy, M. and Kaliyaperumal, S.K., 2013. Comparison of Methods for Detecting Outliers. International Journal of Scientific & Engineering Research, 4(9):709-714.

Larionov, G., 2022. Detect Outliers Using Isolation Forest. https://www.vertica.com/blog/detect-outliers-using-isolation-forest/ (Erişim Tarihi: 16.01.2023).

Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining, Pisa, pp. 413-422, doi: 10.1109/ICDM.2008.17.

Liu, F.T., Ting, K.M. and Zhou, Z.H., 2012. Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1):1-39.

Maddala, G.S., 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Martin, M.A. and Roberts, S., 2006. An Evaluation of Bootstrap Methods for Outlier Detection in Least Squares Regression. Journal Of Applied Statistics, 33(7):703- 720.

McFadden, D., 1973. Conditional Logit Analysis of Qualitative Choice Behavior. In Frontiers in Econometrics (Edited by P. Zarembka), Academic Press, New York.

Mecheva, T., 2022. Outlier detection in traffic data set. In AIP Conference Proceedings (Vol. 2449, No. 1, p. 040014). AIP Publishing LLC. https://doi.org/10.1063/5.0093554.

Menaka, M. and Booba, B., 2022. Analysis to Improve Classifier Accuracy in Crime Data Prediction. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 721-725. ISBN: 978-1-6654-1028-1.

Mertler, C.A. and Vannatta, R.A., 2005. Advanced and Multivariate Statistical Methods: Practical Application and Interpretation (3rd edition). Glendale, Pyrczak Publishing, Los Angeles.

Metz, C.E., 1978. Basic Principles of ROC Analysis. In Seminars in Nuclear Medicine, 8(4):283-298.

Molnar C., 2019. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/. (Erişim Tarihi: 25.11.2022).

Montgomery, M.E., White, M.E. and Martin, S.W., 1987. A Comparison of Discriminant Analysis and Logistic Regression for The Prediction of Coliform Mastitis in Dairy Cows. Canadian Journal of Veterinary Research, 51(4):495-498.

Morano, P., De Mare, G. and Tajani, F., 2013. LMS for Outliers Detection in the Analysis of a Real Estate Segment of Bari. In International Conference on Computational Science and Its Applications, Springer, Berlin, Heidelberg, pp. 457-472.

Nayak, M. and Dash, P., 2015. Distance-based and Density-based Algorithm for Outlier Detection on Time Series Data. Applied Science and Advanced Materials International, 1(4-5):139-146.

Negi, S.S., 2020. Early Prediction of Credit Card Fraud Detection Using Isolation Forest Tree And Local Outlier Factor Machine Learning Algorithms. A Project Report of Capstone Project-2. Galgotias University, Uttar Paradesh ActNo:14. http://103.47.12.35/bitstream/handle/1/1729/16SCSE101154_SUBHASH%20SINGH%20NEGI%20%20Subhash%20Negi.pdf?sequence=1&isAllowed=y. (Erişim Tarihi:5.12.2022).

Negi, K., Kumar, G. P., Raj, G., Sahana, S., & Jain, V., 2022. Degree of Accuracy in Credit Card Fraud Detection Using Local Outlier Factor and Isolation Forest Algorithm. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 240-245). IEEE. https://doi.org/ 10.1109/Confluence52989.2022.9734123

Nurunnabi, A.A.M., Imon, A.H.M.R. and Nasser, M., 2010. Identification of Multiple Influential Observations in Logistic Regression. Journal of Applied Statistics, 37(10):1605-1624.

Nurunnabi, A. and West, G., 2012. Outlier Detection in Logistic Regression: A Quest for Reliable Knowledge from Predictive Modeling and Classification. IEEE 12th International Conference on Data Mining Workshops, pp. 643-652.

Nurunnabi, A.A.M., Hadi, A.S. and Imon A.H.M.R., 2014. Procedures for The Identification of Multiple Influential Observations in Linear Regression. Journal of Applied Statistics, 41(6):1315-1331.

O'Neill, T.J. and Barry, S.C., 1995. Truncated Logistic Regression. Biometrics, 51:533-541.

Obuchowski, N.A., 2005. ROC Analysis. American Journal of Roentgenology, 184(2):364-372.

Oğuzlar, A., 2005. Lojistik Regresyon Analizi Yardımıyla Suçlu Profilinin Belirlenmesi. Atatürk Üniversitesi, İktisadi ve İdari Bilimler Dergisi, 19(1):21-35.

Ohlmacher, G.C. and Davis, J.C., 2003. Using Multiple Logistic Regression and GIS Technology to Predict Landslide Hazard in Northeast Kansas, USA. Engineering Geology, 69(3):331-343.

Oktay, E., Çelik, A.K., Akbaba, A.İ. ve Küçükergüler, K., 2013. İnternet ve Cep Telefonu Kullanımının Lojistik Regresyon Yardımıyla Analizi: Atatürk Üniversitesi Lisans Öğrencileri Örneği. International Journal of Social Science, 6(1):479-496.

Ortiz, M.C., Sarabia, L.A. and Herrero, A., 2006. Robust Regression Techniques A Useful Alternative for The Detection of Outlier Data in Chemical Analysis. Talanta, 70:499–512.

Osborne, J.W. and Amy, O., 2004. The Power of Outliers (And Why Researchers Should Always Check For Them). Practical Assessment, Research & Evaluation, 9(6):1-12.

Ovla, H.D. ve Taşdelen, B., 2012. Aykırı Değer Yöntemi. Mersin Üniversitesi, Sağlık Bilimleri Dergisi, 5(3):1-8.

Önder, H. ve Cebeci, Z., 2002. Lojistik Regresyonlarda Değişken Seçimi. Çukurova Üniversitesi, Ziraat Fakültesi Dergisi, 17(2):105-114.

Özyurt, A., 2013. Siyah Alaca Süt Sığırlarında Doğum Sonrası Dönemde Gelişen Mastitis Olgularının Üreme Performansı Üzerine Etkisi. Hayvansal Üretim, 54(1):21-29.

Pampel, F.C., 2000. Logistic Regression: A primer (Vol. 132). Sage Publications.

Park, Y., Kim, D. and Kim, S., 2012. Robust Regression Using Data Partitioning and M-Estimation. Communications in Statistics - Simulation and Computation, 41(8):1282-1300.

Patel, S.P., Shah, V. and Vala, J., 2015. A Survey of Outlier Detection in Data Mining. National Conference on Recent Research in Engineering and Technology, International Journal of Advance Engineering and Research Development, 3(01):1-4. e-ISSN: 2348 - 4470, print-ISSN:2348-6406.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R., 1996. A Simulation Study of the Number of Events Per Variable in Logistic Regression Analysis. Journal of Clinical Epidemiology, 49(12):1373-1379.

Penny, K.I., 1996. Appropriate Critical Values When Testing for A Single Multivariate Outlier by Using The Mahalanobis Distance. Journal of the Royal Statistical Society, Series C (Applied Statistics), 45(3):73-81.

Pereira, D.F., Oliveira, S.C.D. and Penha, N.L., 2011. Logistic Regression to Estimate the Welfare of Broiler Breeders in Relation to Environmental and Behavioral Variables. Engenharia Agrícola, 31(1):33-40.

Pregibon, D., 1981. Logistic Regression Diagnostics. The Annals of Statistics, 705-724.

Rajeev, H., Devi, U., 2022. Detection of Credit Card Fraud Using Isolation Forest Algorithm. In Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021 (pp. 23-34). Springer Singapore.

R Core Team., 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rousseeuw, P.J. and Van Zomeren, B.C., 1990. Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association, 85(411):633-639.

Ryan, T. P., 1997. Modern Regression Methods. New York: Wiley.

Saloky, T. and Šeminský, J., 2005. Artificial Intelligence and Machine Learning. http://uni-obuda.hu/conferences/SAMI2005/SALOKY.pdf. (Erişim Tarihi: 5.12.2022).

Santosuosso, U., Cini, A. and Papini, A., 2020. Tracing Outliers In the Dataset of *Drosophila suzukii* Records with the Isolation Forest Method. Journal of Big Data, 7(14):2-11.

Sasse, E.A. 2002. Objective Evaluation of Data in Screening for Disease. Clinica Chimica Acta, 315(1-2):17-30.

Shaffer, C.A., 2011. Data Structures & Algorithm analysis in Java (3rd Dover ed.). Mineola, NY: Dover Publications.

Shou, Z., Tian, H., Li, S. and Zou, F., 2018. Outlier Detection With Enhanced Angle-Based Outlier Factor in High-Dimensional Data Stream. International Journal of Innovative Computing Information and Control, 14(5):1633-1651.

Sreevidya, S.S., 2014. A Survey on Outlier Detection Methods. International Journal of Computer Science and Information Technologies, 5(6):8153-8156.

Sternby, J., Thormarker, E. and Liljenstam, M., 2020. Anomaly Detection Forest. 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spain.

Sunderland, K. M., Beaton, D., Fraser, J., Kwan, D., McLaughlin, P. M., Montero-Odasso, M., ... and Binns, M. A., 2019. The Utility of Multivariate Outlier Detection Techniques for Data Quality Evaluation in Large Studies: An Application Within the ONDRI Project. BMC Medical Research Methodology, 19(1):1-16.

Susto, G.A., Beghi, A. and McLoone, S., 2017. Anomaly Detection through on-line Isolation Forest: An Application to Plasma Etching. 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), pp. 89-94.

Tekbir, M., 2009. Aykırı Değer Tespitinde Yoğunluk Tabanlı Kümeleme Yöntemleri. Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi, İstanbul.

Tekin, A.G., 2007. Zaman Dizilerini Kestiriminde Aykırı Değerlerin Etkisinin İncelenmesi. Yüksek Lisans Tezi, Gazi Üniversitesi, Ankara.

Tomak, L. ve Bek, Y., 2009. İşlem Karakteristik Eğrisi Analizi Ve Eğri Altında Kalan Alanların Karşılaştırılması. Journal of Experimental and Clinical Medicine, 27(2):58-65.

Tuncer, N., 2014. Yaşam Çözümlemesinde Aykırı Değerler. Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara.

Turgut, K., 2010. Çok Değişkenli Veride Aykırı Değerlerin Tespiti İçin MVV Yöntemi ve Diğer Yöntemlerle Karşılaştırılması. Yüksek Lisans Tezi, Gazi Üniversitesi, Ankara.

Türkan, S., Çetin, M.C. ve Toktamış, Ö., 2012. Outlier Detection by Regression Diagnostics Based on Robust Parameter Estimates. Hacettepe Journal of Mathematics and Statistics, 41(1): 147-155.

Urgan, N.N. ve Tez, M., 2008. Verilerin Lineer İç İlişkili Olduğu Lojistik Regresyonda Bazı Yanlı Parametre Kestiricileri ve Hata Karaler Ortalamasına Göre Karşılaştırmaları. İstatistikçiler Dergisi: İstatistik ve Aktüerya, 1:1-8.

Üçkardeş, F., Şahinler, S. ve Ercan, E., 2010. Aykırı Gözlemlerin Belirlenmesinde Kullanılan Bazı İstatistikler. KSÜ Doğa Bilimleri Dergisi, 13(1): 42-45.

Ürük, E., 2007. İstatistiksel Uygulamalarda Lojistik Regresyon Analizi. Yüksek Lisans Tezi, Marmara Üniversitesi, İstanbul.

Verardi, V. and Croux, C., 2009. Robust Regression in Stata. The Stata Journal, 9(3):439-453.

Vijayakumar, V., Divya, N. S., Sarojini, P., & Sonika, K., 2020. Isolation Forest and Local Outlier Factor for Credit Card Fraud Detection System. International Journal of Engineering and Advanced Technology (IJEAT), 9(4):261-265.

Vural, A., 2007. Aykırı Değerlerin Regresyon Modellerine Etkileri ve Sağlam Kestiriciler. Yüksek Lisans Tezi, Marmara Üniversitesi, İstanbul.

Weinsteim, S., Obuchowski, N.A. and Lieber, M.L., 2005. Clinical Evaluation of Diagnostic Tests. American Journal of Roentgenology, 184(1):14-19.

Xu, X., Liu, H., Li, L. and Yao, M., 2018. A Comparison of Outlier Detection Techniques for High-Dimensional Data. International Journal of Computational Intelligence Systems, 11(1): 652-662.

Yadav, J. and Sharma, M., 2013. A Review of K-mean Algorithm. International Journal of Engineering Trends and Technology, 4(7):2972-2976.

Yıldırım, N., 2010. En Küçük Kareler, Ridge Regresyon ve Robust Regresyon Yöntemlerinde Analiz Sonuçlarına Aykırı Değerlerin Etkilerinin Belirlenmesi. Yüksek Lisans Tezi, Çukurova Üniversitesi, Adana.

Yıldız, F., 2018. Tarımsal Veri Madenciliğinde Apriori Birliktelik Kuralının Uygulanması. Doktora Tezi, Çukurova Üniversitesi, Adana.

Yücel Altay, Ş., 2014. Sağlık Sektöründe Aykırı Verilerin Algılanması ve Yorumlanması İçin Mekânsal-Zamansal Veri Madenciliği Kullanımı. Yüksek Lisans Tezi, Atatürk Üniversitesi, Erzurum.

Zhang, K., Hutter, M. and Jin, H., 2009. A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, 813-822, Berlin.

Zhang, Z., 2016. Residuals and Regression Diagnostics: Focusing on Logistic Regression. Annals of Translational Medicine, 4(10):1-8.

Zhao, K., Tung, C.W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R:, Reynolds, A., Mezey, J., McClung, A.M.,. Bustamante, C.D. and McCouch, S.R., 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa., 2(1):1-10. DOI: 10.1038/ncomms1467.

Zioutas, G., Avramidis, A. and Pitsoulis, L.A., 2005. Penalized Trimmed Squares Method for Deleting Outliers in Robust Regression.

ÖZGEÇMİŞ

Melis ÇELİK GÜNEY, ilk, orta ve lise öğrenimini Adana'da tamamladı. 2010 yılında başladığı Çukurova Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü'nden, 2013 yılında mezun oldu. 2013 yılında Çukurova Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Biyometri ve Genetik Anabilim dalında yüksek lisansa başladı. 2014 yılında aynı bölümde araştırma görevlisi olarak çalışmaya başladı. 2015 yılında Çukurova Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Biyometri ve Genetik Anabilim dalında doktoraya başladı. Evli ve iki çocuk annesidir.

EKLER

EK 1. Analizlerde Kullanılan R betiği

```
#Veri seti üretme
>set.seed(666)
>x1 = rnorm(3000)
 >x2 = rnorm(3000)
 >x3 = rnorm(3000)
 >z = 3 + 20*x1 + 32*x2 + 1.5*x3   # İlişki Modeli
> pr = 1/(1+exp(-z))         #  inv-logit fonksiyonu ile olasılıkları üretme
> y = rbinom(3000, 1, pr)      # bernoulli tepki değişkeni oluştur (1 ve 0'lar)
> veri = data.frame(y=y, x1=x1, x2=x2, x3=x3) #veri çerçevesi yapma
> head(veri)
> table(y)
> plot(veri)
>boxplot(veri)
>q3x1 = quantile(x1, 0.75)
>bx1 <- q3x1+1.5*IQR(x1)
>bx1
>q3x1

#Veri setine aykırı değer ekleme
>x1[5] <- 4
>x1[705] <-5
>x2[2] <- 6.5
>x2[75] <- 6
>x3[1] <- 8.3
>x3[24] <- 9.4
.
.
.
>veri <- cbind(x1, x2, x3, y)
>veri  # Aykırı değerli veri seti
>summary(veri)
>plot(veri)

# Veri setinin eğitim ve test veri seti olarak ayrılması
>y1=veri [which(veri$y==1),]
```

```
>y0=veri [which(veri$y==0),]
>set.seed(1923)
>sy.1=sample(1:nrow(y1),0.7*nrow(y1))
>sy.0=sample(1:nrow(y0),0.7*nrow(y0))
>t.1=y1[sy.1,]
>t.0=y0[sy.1,]
>trn.veri=rbind(t.1, t.0)
>t.1=y1[-sy.1,]
>t.0=y0[-sy.0,]
>tst.veri=rbind(t.1,t.0)

# Lojistik regresyon
>log.model <- glm(y~ x1+x2+x3, family=binomial(link="logit"), data=trn.veri)
>summary(log.model)
>confint(log.model) #güven aralığı

#Katsayıların ve güven aralıklarının odds oranı
>exp(coef(log.model))
>exp(confint(log.model))

#Hatalar
 >est.veri=predict(log.model, tst.veri, type="response")
>head(est.veri,5)
>est.veri=plogis((predict(log.model, tst.veri)) )
>head(est.veri, 5)
 >res.veri=residuals(log.model,tst.veri,type="deviance")
 >head(res.veri, 5)

#ROC eğrisi
>if(!require(ROCR)) {install.packages("ROCR"); library(ROCR)}
>prob.log.model <- predict(log.model, newdata=tst.veri, type="response")
 >pred.log.model <- prediction(prob.log.model, tst.veri$y)
 >perf.log.model <- performance(pred.log.model, measure = "tpr",  x.measure = "fpr")
>plot(perf.log.model)

#AUROC
>perf.log.model=performance(pred.log.model, measure="auc")
```

```
>auc=perf.log.model@y.values[[1]]
> auc
>if(!require(InformationValue))
+ {install.packages("InformationValue");
+ library(InformationValue)}
>plotROC(tst.veri$y, est.veri)

# Yanılgı matrisi
>opt.kesme=optimalCutoff(tst.veri$y, est.veri)
>opt.kesme
>confusionMatrix(tst.veri$y, est.veri, threshold=opt.kesme)
>Concordance(tst.veri$y, est.veri)
>misClassError(tst.veri$y, est.veri, threshold=opt.kesme)
>sensitivity(tst.veri$y, est.veri, threshold=opt.kesme)
>specificity(tst.veri$y, est.veri, threshold=opt.kesme)

# Sahte R2 Değerleri
>if(!require(pscl)){install.packages("pscl")}
>pscl::pR2(log.model)

# Cook uzaklığı
>library (base)
>cooksd =cooks.distance(log.model)
> print(cooksd)
# Pearson hatalar grafiği
>library(car)
>residualPlots(log.model)
#Cook, student  türü hatalar ve şapka değerleri grafiği
>influenceIndexPlot(log.model, id.n=3)
#Student türü hatalar ve şapka değerleri ve cook uzaklığını birlikte gösteren grafik
>influencePlot(log.model,col="red",id.n=3, xlab=("Şapka değerleri"),ylab=("Student türü
hatalar"))

# Mahalanobis Uzaklığı
# x: bağımsız değişkenleri içeren veri seti
>if(!require(chemometrics)){
+install.packages ("chemometrics ") ; library(chemometrics)}
```

```
>mdx=Moutlier(x, quantile=0.975, plot=TRUE)
>mdx     #mah. uzaklıkları
>pcv2 <- qchisq(0.975, df=p)    #kesme noktası
>cutoff <- sqrt(pcv2)
 >adx <- which(mdx > cutoff)
 >adx       #aykırı değerler
>no=1:3000      #Mahalanobis grafiği
>plot(no,mdx,xlab="Gözlem No", ylab="Mahalanobis Uzaklığı")
> abline(h=3.0575)

# LOF algoritması
>install.packages("Rlof")
>library(Rlof)
>rlofres2 <- lof(x, k=c(3:6))  #k=3,4,5,6 için lof değerleri
>rlofres2
>adlof <- function(x){
+p <- ncol(x)
+for(i in 1:p){
+q3 <- summary(x[,i])[5]
+cutoff <- q3 +IQR(x[,i])
+ad <- which(x[,i] > cutoff)
+print(ad)
+}
+}
>adlof(x=rlofres2)
>n=nrow(x)  #grafik çizme
>labels=1:n
>labels [-adx]="*"
>biplot(prcomp(x),cex=2,xlabs=labels)

# k-ortalamalar
>k <- 2
>kmres <- kmeans(x, k)
>kmres$centers
>kmres$cluster
>merkezler <- kmres$centers[kmres$cluster, ]
>uzakliklar <- sqrt(rowSums((x - merkezler)^2))
```

```
>adx = order(uzakliklar, decreasing=TRUE)[1:30]
>adx
>plot(x[,c("x3", "x1")], pch=19, col=kmres$cluster, cex=1, main="K-ortalamalar İle Aykırı Değerler")
>points(kmres$centers[,c("x3", "x1")], col=1:k, pch="+", cex=2)
>points(x[adx, c("x3", "x1")], pch=1, col=4, cex=2)

#ABOD algoritması
>install.packages("abodOutlier")
>library(abodOutlier)
>abofx <- abod(x, method = "complete")
>sabofx <- sort(abofx)
>sabofx      #sıralanmış abof değerleri
>plot(sabofx, pch=19, col=2)
> lines(sabofx, lwd=2, col=4)
>abline(h=0.27,lty=2, lwd=2, col=3)
>abofx <= 0.27

#iForest algoritması
>install.packages("remotes")
>remotes::install_github("Zelazny7/isofor")
>mod = iForest(X = x, 100, 32)  # yalıtım ağaç sayısı: 100, her ağaç için alt örneklenecek kayıt sayısı:32
>p = predict(mod, x)
>col = ifelse(p > 0.65, "outlier", "normal")  # eşik değer:0.65
>no=1:3000      #Grafik için
>plot(no,col,xlab="Gözlem No", xlab="Gözlem No", ylab="Aykırı değer puanı")

#Her yöntem ile bulunan aykırı değerler veri setinden temizlenmesinden sonra veri setlerinin okutulması
>mah = read.table("C:/Users/melis/Documents/mah.txt",header=T)
>lof = read.table("C:/Users/melis/Documents/lof.txt",header=T)
>kort = read.table("C:/Users/melis/Documents/kort.txt",header=T)
>abod = read.table("C:/Users/melis/Documents/abod.txt",header=T)
>iforest = read.table("C:/Users/melis/Documents/iforest.txt",header=T)
```

```
# Aykırı değer belirleme yöntemlerinin ROC eğrisi ile karşılaştırılması
>install.packages("rsample")
>library(dplyr)
>library(rsample)
>set.seed(123)
>split_strat  <- initial_split(mah, prop = 0.7, strata = "y")
>trn.mah  <- training(split_strat)
>tst.mah  <- training(split_strat)
>log.modelm <- glm(y~ x1+x2+x3, family=binomial(link="logit"), data=trn.mah)
>summary(log.modelm)
>set.seed(123)
>split_strat  <- initial_split(lof, prop = 0.7, strata = "y")
>trn.lof  <- training(split_strat)
>tst.lof  <- training(split_strat)
>log.modell <- glm(y~ x1+x2+x3, family=binomial(link="logit"), data=trn.lof)
>summary(log.modell)
>set.seed(123)
>split_strat  <- initial_split(kort, prop = 0.7, strata = "y")
>trn.kort <- training(split_strat)
>tst.kort <- training(split_strat)
>log.modelk<- glm(y~ x1+x2+x3, family=binomial(link="logit"), data=trn.kort)
>summary(log.modelk)
>set.seed(123)
>split_strat  <- initial_split(abod, prop = 0.7, strata = "y")
>trn.abod <- training(split_strat)
>tst.abod  <- training(split_strat)
>log.modela<- glm(y~ x1+x2+x3, family=binomial(link="logit"), data=trn.abod)
>summary(log.modela)
>set.seed(123)
>split_strat  <- initial_split(iforest, prop = 0.7, strata = "y")
>trn.iforest <- training(split_strat)
>tst.iforest <- training(split_strat)
>log.modelb<- glm(y~ x1+x2+x3, family=binomial(link="logit"), data=trn.iforest)
>summary(log.modelb)
>if(!require(ROCR)) {install.packages("ROCR"); library(ROCR)}
>est.mah <- plogis(predict(log.modelm, tst.mah))
>est.lof <- plogis(predict(log.modell, tst.lof))
```

```
>est.kort <- plogis(predict(log.modelk, tst.kort))
>est.abod <- plogis(predict(log.modela, tst.abod))
>est.iforest <- plogis(predict(log.modelb, tst.iforest))
>pred.mah <- prediction(est.mah, tst.mah$y)
>pred.lof <- prediction(est.lof, tst.lof$y)
>pred.kort<- prediction(est.kort, tst.kort$y)
>pred.abod <- prediction(est.abod, tst.abod$y)
>pred.iforest <- prediction(est.iforest, tst.iforest$y)
>perf.log.modelm <- performance(pred.mah, "tpr", "fpr" )
>perf.log.modell <- performance(pred.lof, "tpr", "fpr" )
>perf.log.modelk <- performance(pred.kort, "tpr", "fpr" )
>perf.log.modela <- performance(pred.abod, "tpr", "fpr" )
>perf.log.modelb <- performance(pred.iforest, "tpr", "fpr" )
>plot(perf.log.modelm, lwd=2, col="blue", main="ROC Eğrisi", xlab="1-özgüllük",
ylab="duyarlılık")
>plot(perf.log.modell, add = TRUE, lwd=2, col="red")
>plot(perf.log.modelk, add = TRUE, lwd=2, col="green")
>plot(perf.log.modela, add = TRUE, lwd=2, col="black")
>plot(perf.log.modelb, add = TRUE, lwd=2, col="yellow")
 >abline(a=0, b=1, lty=2, lwd=2, col="black")
>legend(0.75, 0.2, legend=c("Mahalanobis Uzaklığı", "LOF", "k-ortalama", "ABOD","iForest"),
+col=c("blue", "red","green","black", "yellow"), lty=1, cex=0.5)
```