

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BANKACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ
YÖNTEMLERİNİ KULLANARAK POS CİRO TAHMİNİ

Tuğçe SAYIN

YÜKSEK LİSANS TEZİ

Endüstri Mühendisliği Anabilim Dalı
Sistem Mühendisliği Programı

Danışman
Doç. Dr. Tufan DEMİREL

Mart, 2023

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**BANKACILIK SEKTÖRÜNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİ
KULLANARAK POS CİRO TAHMİNİ**

Tuğçe SAYIN tarafından hazırlanan tez çalışması 02.03.2023 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı, Sistem Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Doç. Dr. Tufan DEMİREL
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. Tufan DEMİREL, Danışman
Yıldız Teknik Üniversitesi

Doç. Dr. Şükran ŞEKER, Üye
Yıldız Teknik Üniversitesi

Dr. Öğr. Üye. Cafer Erhan BOZDAĞ, Üye
İstanbul Teknik Üniversitesi

Danışmanım Doç. Dr. Tufan DEMİREL sorumluluğunda tarafımda hazırlanan Bankacılık Sektöründe Makine Öğrenmesi Yöntemlerini Kullanarak POS Ciro Tahmini başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Tuğçe SAYIN

İmza



Aileme
ve
Sevdiklerime

TEŞEKKÜR

Çalışma kapsamında değerli bilgilerini ve deneyimini aktaran kıymetli danışmanım Doç Dr. Tufan DEMİREL'e çok teşekkür ederim.

Tuğçe SAYIN



İÇİNDEKİLER

SİMGE LİSTESİ	vii
KISALTMA LİSTESİ	viii
ŞEKİL LİSTESİ	ix
TABLO LİSTESİ	xi
ÖZET	xii
ABSTRACT	xiv
1 GİRİŞ	1
1.1 Literatür Özeti	3
1.2 Tezin Amacı	11
1.3 Hipotez	12
2 MAKİNE ÖĞRENMESİ	13
2.1 Makine Öğrenimi Süreci	14
2.2 Makine Öğrenmesi Uygulama Alanları	15
2.2.1 Sınıflandırma	15
2.2.2 Regresyon	16
2.2.3 Kümeleme	17
2.2.4 Hibrit Sistemler	18
2.3 Makine Öğrenimi Türleri	19
2.3.1 Denetimli Makine Öğrenmesi	19
2.3.2 Denetimsiz Makine Öğrenmesi	20
2.3.3 Yarı Denetimli Makine Öğrenmesi	21
2.3.4 Pekiştirmeli Öğrenme	22
2.4 Makine Öğreniminde Karşılaşılan Sorunlar	23
2.4.1 Aşırı Öğrenme-Eksik Öğrenme	23
2.4.2 Yanlılık/Varyans İlişkisi	24
2.5 Model Doğrulama Yöntemleri	25
2.5.1 Train & Test Modeli	25
2.5.2 K-Katlı Çapraz Doğrulama Yöntemi	26
2.5.3 Leave-One-Out Çapraz Doğrulama Yöntemi	27
2.6 Modelde Kullanılan Yöntemler	27
2.6.1 Lineer Regresyon	28
2.6.2 Ridge Regresyon	29

2.6.3	Lasso Regresyon.....	29
2.6.4	K-En Yakın Komşu (KNN) Regresyon	29
2.6.5	Karar Ağaçları.....	29
2.6.6	Rastgele Ormanlar	30
2.6.7	XGBoost.....	30
2.7	Model Performansının Ölçülmesi	31
2.7.1	Korelasyon Katsayısı (r).....	31
2.7.2	Ortalama Kare Hatası	31
2.7.3	Ortalama Karekök Sapması.....	32
2.7.4	Ortalama Mutlak Değer.....	32
3	UYGULAMA	33
3.1	Problemin Tanımlanması.....	34
3.2	Verinin Hazırlanması.....	35
3.3	Veri Setinin Analizi	39
3.3.1	Veri Temizleme	40
3.3.2	Eksik Veri Analizi	40
3.3.3	Değişken Dönüşümü.....	47
3.4	Veri Setinin Modellenmesi.....	48
3.4.1	Lineer Regresyon (Tek Değişkenli/Çok Değişkenli)	49
3.4.2	Ridge Regresyon.....	53
3.4.3	Lasso Regresyon.....	53
3.4.4	K-En Yakın Komşu	54
3.4.5	Karar Ağaçları.....	56
3.4.6	Rastgele Ormanlar	58
3.4.7	XG- Boost	60
3.5	Modellerin Sonuç Değerleri	61
4	SONUÇ VE ÖNERİLER	63
	KAYNAKÇA	65
	TEZDEN ÜRETİLMİŞ YAYINLAR	70

SİMGELİSTESİ

β_k	Değişken Katsayıları
oi	Deneyisel Olarak Saptanmış Değerler
$\bar{o}$	Deneyisel Verilerin Ortalaması
x_k	Girdi Değişkenleri
N	Gözlem Değeri
$d_{(i,j)}$	i ve j Noktaları Arasındaki Uzaklık
ti	Regresyon Eşitliğinden Hesaplanmış Değerler
x_{ik}	x Ekseninde i. Nuktada y Ekseninde k. Nuktada Bulunan Lokasyon
x_{jk}	x Ekseninde j. Nuktada y Ekseninde k. Nuktada Bulunan Lokasyon

KISALTMA LİSTESİ

BKM	Bankalararası Kart Merkezi
CRISP-DM	Veri Madenciliği İçin Sektörler Arası Standart Süreç
DBN	Sınıflandırıcı Derin İnanç Ağı
EFT	Elektronik Para Transferi
GA	Genetik Algoritma
KNN	K – En Yakın Komşu
MAE	Ortalama Mutlak Değer
MLR	Çok Değişkenli Doğrusal Regresyon
MSE	Ortalama Kare Hatası
ÖKC	Ödeme Kaydedici Cihazlar
POS	Point of Sale
RMSE	Ortalama Karekök Sapması
SNS	Sosyal Ağ Servisi
SVR	Destek Vektör Regresyonu
XGBoost	Aşırı Gradyan Artırma Ağacı

ŞEKİL LİSTESİ

Şekil 2.1 Makine öğrenmesi sistemi.....	13
Şekil 2.2 Makine öğrenimi adımları (CRISP-DM).....	14
Şekil 2.3 İkili sınıflandırma sistemi.....	15
Şekil 2.4 Çoklu sınıflandırma sistemi.....	16
Şekil 2.5 Çoklu sınıflandırmanın ikili sınıflandırmaya ayrılmış hali	16
Şekil 2.6 Tek değişkenli regresyon	17
Şekil 2.7 Çok değişkenli regresyon	17
Şekil 2.8 Zaman serisi tahmini.....	17
Şekil 2.9 İkili sınıflandırma.....	18
Şekil 2.10 Çoklu sınıflandırma.....	18
Şekil 2.11 Regresyon ve sınıflandırma çalışmalarından oluşan hibrit sistem.....	19
Şekil 2.12 Kümeleme ve sınıflandırma çalışmalarından oluşan hibrit sistem.....	19
Şekil 2.13 Denetimli öğrenme.....	20
Şekil 2.14 Denetimsiz öğrenme.....	21
Şekil 2.15 Yarı denetimli öğrenme.....	22
Şekil 2.16 Pekiştirmeli öğrenme.....	23
Şekil 2.17 Aşırı-eksik-uygun öğrenme.....	24
Şekil 2.18 Yüksek yanlılık-varyans ilişkisi.....	24
Şekil 2.19 Train & Test modeli.....	26
Şekil 2.20 K-Katlı çapraz doğrulama yöntemi.....	26
Şekil 2.21 Çalışmada kullanılan makine öğrenmesi yaklaşımı ve algoritmalar	28
Şekil 2.22 İki özellik alan bir karar ağacı modelinin gösterimi	30
Şekil 2.23 Rastgele ormanlar algoritmasının işlevsel gösterimi	30
Şekil 3.1 Uygulamanın adımları.....	34
Şekil 3.2 Veri setinde değişkenler detayında örneklem sayısı.....	41
Şekil 3.3 Değişkenlerin boxplot grafikleri.....	42
Şekil 3.4 Tek değişkenli regresyonda residuals ×fitted grafiği (%80 Eğitim-%20 Test).....	50
Şekil 3.5 Tek değişkenli regresyonda residuals ×fitted grafiği (%70 Eğitim-%30 Test).....	50
Şekil 3.6 Çok değişkenli regresyonda residuals ×fitted grafiği (%70 Eğitim-%30 Test).....	52

Şekil 3.7 Çok değişkenli regresyonda residuals ×fitted grafiği (%70 Eğitim-%30 Test).....	52
Şekil 3.8 Katsayılar×L2 norm grafiği.....	53
Şekil 3.9 Katsayılar×L1 norm grafiği.....	54
Şekil 3.10 Komşuluk sayısı×RMSE değeri (%80 eğitim-%20 test).....	55
Şekil 3.11 Komşuluk sayısı×RMSE değeri (%70 eğitim-%30 test).....	56
Şekil 3.12 Ağaç derinliği×RMSE değeri (%80 eğitim-%20 test).....	57
Şekil 3.13 Ağaç derinliği×RMSE değeri (%70 eğitim-%30 test)	58
Şekil 3.14 Ağaç sayısı×hata değeri (%80 eğitim-%20 test).....	59
Şekil 3.15 Ağaç sayısı×hata değeri (%80 eğitim-%20 test).....	59
Şekil 3.16 x × test_y değeri (%80 eğitim-%20 test).....	60
Şekil 3.17 x × test_y değeri (%80 eğitim-%20 test).....	61

TABLO LİSTESİ

Tablo 1.1 Literatür incelemesi özet tablo	8
Tablo 3.1 Değişken açıklamaları.....	35
Tablo 3.2 Veri setinin istatistiksel bilgileri	37
Tablo 3.3 Veri seti hakkında genel bilgi içeren program çıktısı	38
Tablo 3.4 %80 eğitim-%20 test verisiyle elde edilen sonuç değerleri	62
Tablo 3.5 %70 eğitim-%30 test verisiyle elde edilen sonuç değerleri	62



Bankacılık Sektöründe Makine Öğrenmesi Yöntemlerini Kullanarak POS Ciro Tahmini

Tuğçe SAYIN

Endüstri Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Tufan DEMİREL

Günümüzde büyük tutarlara ulaşan ödeme alma ve ödeme verme işlemlerinin sorunsuz ve hızlı şekilde sonuçlandırılmasını sağlayan ödeme sistemleri hayatın her alanında yüksek öneme sahiptir. Açılımı “Point of Sale” olan ve ödeme alma şekline göre ayrılan POS cihazları ise ödeme sistemlerinin temelini oluşturmaktadır. Finans sektöründe, Elektronik Fon Transferinin (EFT) yapılmasına olanak sağlayan POS cihazlarından geçen ciro işletmeler arasında rekabete neden olmaktadır. Yapılan çalışmada Türkiye’nin önde gelen katılım bankalarından birinde bireysel müşterilerden çok tüzel müşterilerin POS cihazlarından geçen ciro değerlerine odaklanılmaktadır. Böylelikle, portföy sorumlularının potansiyel tüzel POS müşterilerinin tespiti, yeni müşteri kazanımı ve banka stratejilerinin belirlenmesi hedeflenmektedir. Çalışma kapsamında 2021 yılı ağustos ayı ile 2021 yılı temmuz ayı arasında gerçekleşen veriler kullanılmıştır. Veriler 16 farklı bölgenin 402 subesinden rastgele şekilde seçilmiştir. Veri seti 1500 gözlem biriminden ve 23 adet farklı değişkenden oluşmaktadır. Gözlem birimleri müşterileri ifade ederken değişkenler POS cirosunu etkileyebilecek müşteri bilgilerinden oluşmaktadır. Bu kapsamda sektörde önemli yere sahip tüzel banka müşterilerinin POS cihazlarından

geçen aylık ciro değerlerine ait veri seti %80 eğitim -%20 test ve %70 eğitim-%30 test örneklemelerine göre denetimli makine öğrenmesi algoritmalarıyla (Tek Değişkenli Lineer Regresyon, Çok Değişkenli Lineer Regresyon, Ridge Regresyon, Lasso Regresyon, K-En Yakın Komşu, Random Forest, Karar Ağaçları ve XGBoost) R-Studio yazılım dilinde tahminlenmektedir. Veri setinin analizi sonrasında kullanılan algoritmaların performansı farklı istatistiksel ölçütlerle (r^2 -korelasyon katsayısı, MSE- Ortalama Kare Hatası, RMSE- Ortalama Karekök Sapması, MAE- Ortalama Mutlak Değer) 10 Katlı Çapraz Doğrulama yöntemi kullanılarak kıyaslanmıştır. Kıyaslama sonucunda en iyi performans gösteren modelin XGBoost algoritması (%80 eğitim-%20 test) olduğu tespit edilmiştir.

Anahtar Kelimeler: Bankacılık sektörü, POS, ciro, tüzel müşteri, makine öğrenmesi

POS Revenue Estimation Using Machine Learning Methods in Banking Sector

Tuğçe SAYIN

Department of Industrial Engineering

Master of Science Thesis

Supervisor: Doç. Dr. Tufan DEMİREL

Today, payment systems, which ensure that large amounts of payment and payment transactions are concluded smoothly and quickly, have a high importance in all areas of life. POS devices, which stands for "Point of Sale" and are differentiated according to the payment method, form the basis of payment systems. In the finance sector, turnover passing through POS devices that allow Electronic Funds Transfer (EFT) causes competition among businesses. In this study, one of Turkey's leading participation banks focuses on the turnover values of corporate customers, which pass through POS devices rather than individual customers. Thus, it is aimed to identify potential corporate POS customers, to acquire new customers and to determine bank strategies of portfolio managers. Within the scope of the study, data from August 2021 to July 2021 were used. Data were randomly selected from 402 branches of 16 different regions. The data set consists of 1500 observation units and 23 different variables. While the monitoring units represent the customers, the variables consist of customer information that can affect the POS turnover. In this context, the data set of the monthly turnover values of corporate bank customers, which have an important place in the sector, passed through the POS devices, is used

with supervised machine learning algorithms (Single Variable Linear Regression, Multivariate Linear Regression, Ridge Regression, Lasso Regression, K-Nearest Neighbor, Random Forest, Decision Trees and XGBoost) are estimated in R-Studio. After the analysis of the data set, the performance of the algorithms used was compared with different statistical criteria (r^2 -correlation coefficient, MSE- Mean Squared Error, RMSE- Root Mean Squared Deviation, MAE- Mean Absolute Value) using the 10-fold Cross Validation method. As a result of the comparison, it was determined that the best performing model was the XGBoost algorithm (80% training-20% testing).

Keywords: Banking industry, POS, turnover, legal customer, machine learning



Her geçen gün insan popölasyonunun artmasıyla birlikte tüm sektörlerde yüksek veri artışı yaşanmaktadır. Bu verileri depolamak, işlemek, yönetmek ve analiz etmek için yeni teknolojilerin hızla gelişmesi, işletmelerin çalışma biçimlerini değiştirmektedir. Yüksek hacimli ve farklı formlarda toplanan bu verilerin şirketler için anlamlandırılması önemli hale gelmektedir. Bu nedenle de “Big data” yani “Büyük Veri” kavramı son yıllarda en sık duyduğumuz kavramlar arasına girmektedir. İşletmeler ise büyük veriyi karar vermeye yardımcı olan yapılandırılmış ve yapılandırılmamış verilerin analizi ile birlikte faydalı bilgiye dönüştürmektedir. Veri analizi, işletmeler için daha iyi karar almak, müşteri memnuniyetini sağlamak, üretimi arttırmak gibi ihtiyaç duydukları stratejik bilgileri elde edebilecekleri öneme sahiptir. Tarım, gıda, telekominikasyon, finans gibi bir çok sektörde kullanım alanına sahiptir [1].

Metin, rakam, fotoğraf, video, ses, konum bilgisi vb. gibi birçok çeşitte ve boyutta olan veri topluluğundan anlamlı ve nitelikli bilgiyi elde etmek önemlidir. Bu amaçla ise “Big Data Analysis-Büyük Veri Analizi” kavramını ortaya çıkmıştır. Büyük veri analizinde ise temel olarak altı farklı yöntem kullanılmaktadır. Bu yöntemlerden bir tanesi ise makine öğrenmesidir [2].

Makine öğrenmesi bir problemi mevcut veri seti ile modelleyen ve en yüksek performansı vermek için uygulanan bilgisayar algoritmalarının genel adıdır. Sınıflandırma, regresyon ve kümeleme olarak 3 farklı uygulama alanı vardır. Tahmin edilen çıktı değeri kategorik ise sınıflandırma, nümerik ise regreyon ve benzer gözlemleri aynı kümeye attıysa kümeleme yöntemi uygulanmaktadır. Denetimli, denetimsiz, yarı denetimli ve hibrit olmak üzere 4 farklı uygulama türü bulunmaktadır. Mevcut verinin etiketli olup olmamasına göre kullanılan uygulama türü değişiklik göstermektedir [3].

Finansal işletmelerin performanslarını ve sektöre kattıkları değerin tahmin edilmesi rakipler açısından oldukça önemlidir. Finansal kurumlar, Merkez bankası ve danışmanlık şirketleri örnek olarak verilmektedir. Finansal işletmelerde makine öğrenmesi yöntemleriyle birlikte bazı kullanımları şöyledir: churn (kayıp) tahmini, ciro tahmini, döviz tahmini, hisse senedi tahmini, müşteri-ürün tahmini [4].

Bankacılık sektörü finansal işletmelere finansman sağlamada kritik öneme sahiptir. Bankalar hem şubeden hem de dijital ortamlardan bireysel ve tüzel müşterilerin günlük verilerini sistemlerine kaydetmektedir. Kaydolan bu veriler farklı analiz yöntemleriyle işlenerek anlamalı hale getirilmektedir. Dolandırıcılık tespiti, kampanya müşteri kitlesinin tespiti, muhasebe ve yasal gereklilikler için bir çok amaçla kullanılmaktadır. Bu çerçevede makine öğrenmesi yöntemleri müşteri deneyimini arttırmak için sıklıkla kullanılmaktadır [5].

Ödeme sistemleri, ödeme veya tahsilat yapılmasını sağlayan en önemli bankacılık sistemlerin arasında yer almaktadır. Teknolojide yaşanan gelişmelerle birlikte ödeme sistemlerine olan talepte büyük artış yaşanmaktadır. Günümüzde bir çok tüketici istek ve ihtiyaçlarını hızlı ve zahmetsiz yollarla ödeme yapabilmeye başlamıştır. Ödeme sistemlerinin bir parçası olan Elektronik Fon Transferi yapabilmek için ödeme kartı ile iletişim sağlayan POS cihazları teknolojik gelişmelerin sonuçlarından biri olan dijitalleşme ile gelişme göstermektedir. Bankalar ve finansal kuruluşlar ise POS cihazı kullanan müşteri sayısını ve ciro değerlerini arttırmak için rekabet halindedirler. Bu noktada müşterilerin gelecek cirolarının tahmini ve potansiyel müşterilerin tespiti kritik öneme sahiptir.

Ödeme alma-Acquiring taraf olarak bilinen POS cihazları ödeme sistemlerinin temel taşlarından biridir. Çalışmanın amacı, katılım bankalarında öncü olan bir bankanın tüzel müşterilerinin makine öğrenmesi yöntemleri aracılığıyla POS ciro tahminin yapılmasıdır. Genel yaklaşım ise bireysel müşterilerden çok küçük ve orta ölçekli tüzel müşterilerin cirolarını ele almaktır. Yapılan bu geliştirme sayesinde stratejik müşteriler belirlenerek ciro arttırılıp pazar payında artış sağlanacaktır.

4 bölümden oluşan bu çalışmada, Bölüm 1’de çalışma konusundaki literatür özetini, tezin amacını ve hipotezi içermektedir. Bölüm 2’ de makine öğrenmesi, makine öğrenmesi yöntemleri ve algoritmalarına yer verilmiştir. Bölüm 3 tezin uygulama

kısmını oluşturmaktadır. Bölüm 4' te ise uygulamadan elde edilen sonuç ve öneriler paylaşılmıştır.

1.1 Literatür Özeti

Literatürde eğitim, sağlık, tekstil, bankacılık gibi tüm sektörlerde bir çok makine öğrenmesi çalışması bulunmaktadır. Gelir tahmini ve ciro tahmini çalışmalarının yapıldığı 19 makale özetlenmiştir. Tablo 1.1'de ilgili çalışmanın yayınlanan yılı, programlama dili, kullanılan algoritmalar ve en iyi sonuç veren algoritma belirtilmiştir.

Budak ve Sarvari, lojistik sektöründe sürdürülebilir karayolu yük taşımacılığında kar marjını tahminlemiştir. Python yazılım dilinde 3 farklı algoritma kullanılmıştır. Bu algoritmalar; Rastgele Ormanlar, Dayanıklı (Robust) Regresyon ve XGBoost 'dir. Uygulama sonucunda XGBoost algoritmasının 2 farklı işletmede yapılan verilerle kar tahmininde üstünlük sağladığı belirtilmiştir. Ayrıca stratejik ve sürdürülebilir yönetim hakkında bilgiler sunmaktadır [6].

Sepenu ve Eliassen, profesyonel saç ürünleri üreticisi olan işletmenin pazarlama ve satış stratejilerini belirlemede yardımcı olmak amacıyla gelir tahmini çalışması yapmışlardır. Uygulamada kümeleme ve regresyon modelleri kullanılmıştır. Bu algoritmalar; Lineer Regresyon, Destek Vektör Makineleri, Rastgele Ormanlar, Mini-Batch K-Means, BIRCH Sınıflandırma, K-Ortalama, Hiyerarşik Kümeleme and Gaussian Karışım Modeli'dir. Ocak 2015 ve Kasım 2021 yılları arasında üç ürünü satın alan yaklaşık 3000 müşteriye indirgenerek çalışma tamamlanmıştır. Uygulama sonucunda Rastgele Ormanlar, K-Ortalama modellerinin üstünlük sağladığı belirtilmiştir. Gelir tahminini artırmak için hangi özelliklerin önemli olduğu konusunda işletmeye öneriler verilmiştir [7].

Hajek ve Olej, mali stres nedeniyle belediyelerde ilginin arttığı gelir tahmini konusu üzerine çalışma gerçekleştirmiştir. Weka yazılım diliyle 3 farklı algoritma kullanılmıştır. Bu algoritmalar; İleri Beslemeli Sinir Ağları, Destek Vektör Makineleri ve Lineer Regresyon'dur. Ekonomik ve finansal parametrelerin belediye gelirine etkisi 4 yıllık periyot ile incelenmiştir. Uygulama sonucunda İleri Beslemeli Sinir Ağları, Destek Vektör Makineleri algoritmalarının regresyon modeline göre üstünlük sağladığı belirtilmiştir [8].

Li, makine öğrenmesi yöntemlerini kullanarak film gelir modelinin tahmini üzerine çalışma yapmıştır. Temmuz 2017 ve öncesinde vizyona giren 45.000'den fazla filmde oluşan filmlerin gelirlerini ve özelliklerini dikkate alarak 5000 film örneği seçilerek çalışma yapılmıştır. 4 farklı algoritma kullanılmıştır. Bu algoritmalar Lineer Regresyon, Bayesyen Regresyon, Gradyan Artırma, Naive Bayes'dir. Uygulama sonucunda Gradyan Artırma algoritmasının regresyon modellerine göre üstünlük sağladığı belirtilmiştir [9].

Lin, Pai, Lu ve Chang, Tayvan'daki MediaTek şirketinin gerçek gelir verileri kullanılarak makine öğrenmesi yöntemleriyle gelir tahmini çalışması yapılmıştır. Deneysel veriler üç gruba ayrılmıştır: eğitim, doğrulama ve test. 5 farklı algoritma kullanılmıştır. Bu algoritmalar; Mevsimsel Otoregresif Entegre Hareketli Ortalama, Genelleştirilmiş Regresyon Sinir Ağları, Genetik Algoritmalarla Vektör Regresyonunu, En Küçük Kareler Genetik Algoritmalarla Vektör Regresyonunu ve Bulanık En Küçük Kareler Genetik Algoritmalarla Vektör Regresyon'dur. Uygulama sonucunda Bulanık En Küçük Kareler Genetik Algoritmalarla Vektör Regresyon algoritmalarının diğer modellere göre üstünlük sağladığı belirtilmiştir [10].

Penpece ve Elma, Yapay Sinir Ağları yöntemini kullanarak perakende sektöründe satış gelir tahmini üzerine çalışma yapmıştır. 2010-2013 yılı ilk çeyreğinde gerçekleşen değerler kullanılarak 2013 yılının son çeyreğinin satış gelir tahmini yapılmıştır. Yapay Sinir Ağları yönteminin sonucu olarak bu yöntemin diğer sektörlerde de yüksek doğruluk değerine yol açacağı sonucuna varılmıştır [11].

Huang ve Yen, denetimli-denetimsiz ve yarı denetimli makine öğrenmesi yöntemleri kullanılarak finansal sıkıntıların tahmini üzerinde çalışma yapmıştır. Tayvanlı firmaların 2010-2016 yılları arasındaki finansal tablolarından toplamda 16 adet değişken belirlenmiştir. 6 farklı algoritma kullanılmıştır. Bu algoritmalar; Destek Vektör Makineleri, Bulanık kümeleme algoritmasına sahip Genetik Algoritma (GA), XGBoost, Sınıflandırıcı Derin İnanç Ağı (DBN) ve Derin İnanç Ağı - Destek Vektör Makineleri'dir. Uygulama sonucunda XGBoost algoritmalarının diğer modellere göre üstünlük sağladığı belirtilmiştir [12].

Kermani, Charmahineh ve Razavi&Ouarda, Jiroft Barajı çıkış suyunun buharlaşma, baraja giren su hacmi ve barajdan çıkan su hacmi bağımsız faktörleri kullanarak satış gelirlerinin ve ayrıca satış gelirlerine ilişkin geçmiş kayıtların tahmin edilmesi

üzerinde çalışma yapmıştır. 4 farklı algoritma kullanılmıştır. Bu algoritmalar; Sınıflandırma ve Regresyon Ağacı, Ki-Kare Otomatik Etkileşim Dedektörü, Çok Katmanlı Algılayıcı Sinir Ağı ve Radyal Tabanlı İşlev Sinir Ağı'dır. Uygulama sonucunda Çok Katmanlı Algılayıcı Sinir Ağı algoritmalarının regresyon modellerine göre üstünlük sağladığı belirtilmiştir [13].

Alkelani ve Awad, Filistin'de yıldan yıla değişiklik gösteren zeytinyağı verimliliğini yağmur ve diğer iklim değişikliklerini göz önünde bulundurarak tahmin edilmesini sağlamıştır. 25 yılın verisini kullandıkları bu çalışmada karar ağacı algoritması kullanılarak zeytinyağı verimliliği tahminlenmiştir. Çalışma sonucunda zeytinyağı verimliliğinin iklim faktörleri ile arasında yüksek korelasyon değerinin olduğu tespit edilmiştir. Çalışmadan elde edilen bulgular sayesinde yerel pazarın dengelenmesi, ithalat-ihracat miktarının belirlenmesi, gelecekteki üretim için stratejik planların belirlenmesi sağlanmıştır [14].

Rao, Shastri ve Sathyashree & Sahu temel bileşenler analizi, özellik seçimi ve etiket kodlama gibi çeşitli veri analizi yöntemleri seçilerek restoranların satış geliri tahmini üzerine çalışma sağlanmıştır. 2 farklı algoritma kullanılmıştır. Bu algoritmalar; Rastgele Orman ve Doğrusal Regresyondur. Uygulama sonucunda Rastgele Orman algoritmasının regresyon modeline göre üstünlük sağladığı belirtilmiştir [15].

Schmidt, Kabir ve Hoque 3 parça halinde topladıkları POS (Point of Sale) cihazında bulunan zaman, satılan ürünler, brüt satışlar, konum, müşteri adı gibi bilgileri kullanarak orta ve küçük restoranların satış tahmini üzerinde çalışma gerçekleştirmiştir. Toplamda 25 farklı algoritma kullanılmıştır. Bu algoritmalarından bazıları şöyledir; Lineer Regresyon, Stochastic Gradient Descent Regresyon, Ridge Regresyon, Lasso Regresyon, Elastic-Net Regresyon, Bayesian Ridge Regresyon, Kernel Ridge Regresyon, Karar Ağaçları, Destek Vektör Regresyon, Nu Destek Vektör Regresyon, Lineer Destek Vektör Regresyon, K-En Yakın Komşu Regresyon, Gauss Süreci Regresyonu, Yineleme Sinir Ağları, Zamansal Füzyon Transformatörü. Uygulama sonucunda Yineleme Sinir Ağları ve Zamansal Füzyon Transformatörü modeli diğer modellere göre üstünlük sağladığı belirtilmiştir [16].

Chen, Cho, Dou ve Lev makine öğrenmesi yöntemlerini ve finansal verileri kullanarak bir yıllık kazanç değişkenlerinin yönünü tahmin edilmesi üzerinde

alıřma yapmıřtır. alıřamada 2 farklı algoritma kullanılmıřtır. Bu algoritmalar Rastgele Ormanlar ve Stokastik Gradyan Artırma'dır. Geleneksel modellere gre hem algoritmaların hem de daha ayrıntılı finansal verilerin kullanılmasının sonucu olumlu ynde etkilediėi belirtilmiřtir [17].

Kim, Hong ve Kang Sosyal Aė Servisi (SNS) verilerini kullanarak karmařık makine ėrenimi tabanlı regresyon algoritmaları yardımıyla film giřelerinin ciro tahmini zerine 3 sıralı tahmin alıřması gerekleřtirmiřtir. Film vizyona girmeden, bir hafta sonra ve 3 hafta sonra olmak zere  farklı tahmin alıřması yapılmıřtır. 4 farklı algoritma kullanılmıřtır. Bu algoritmalar; ok Deėiřkenli Doėrusal Regresyon (MLR), Destek Vektr Regreyon (SVR), K-En Yakın Komřu Regresyon, Gauss Sreci Regresyonu algoritmaları kullanılmıřtır. Girdi ve hedef deėiřkenler arasındaki fonksiyonel iliřkiye lineer denklem řeklinde uyan ok deėiřkenli regreyon algoritması diėer algoritmalara gre stnlk gstermiřtir [18].

Pundir, Ganapati, Maheshwari ve Kumar ekonomik performans, fiyat endeksleri, yakıt fiyatları, fiyat indirimleri ve retim gibi sezgisel olmayan dıř faktrle birlikte perakende sektrnde olan bir firmanın gelir tahmini zerine alıřma yapmıřtır. 2 algoritma ve 1 geleneksel yntem kullanılmıřtır. Rastgele Orman Regresyonu ve Vektr Otoresgresyon Algoritmaları kullanılırken Geleneksel Hareketli Ortalamalar methodu uygulanmıřtır. Uygulamanın sonucunda makine ėrenmesi algoritmalarının geleneksel ynteme gre stnlk saėladıėı belirtilmiřtir [19].

řim ve Pourhomayoun bir filmin aılıř hafta sonuna kadar toplanan Twitter verilerini analiz ederek aılıř hafta sonları boyunca film gelirini tahmin etmek zerine alıřma gerekleřtirmiřtir. Verilerin hazır olmaması nedeniyle veriler ncelikle Twitter API'sı kullanılarak toplanmıřtır. Denetimli ve denetimsiz ėrenme modellerinin birleřimine dayalı birkaç tahmine dayalı model geliřtirilmiřtir. Doėru tahmini yapabilmek iin doėrusal bir regresyon modeli geliřtirilmiřtir. alıřma iin Python Scikit-Learn ktphanesi kullanılmıřtır [20].

Milosevi, en nemli yatırım stratejilerinden biri olan hisse senedinin gelecekteki fiyatının tahmini zerine alıřma gerekleřtirmiřtir. 2012 yılının ilk eyreėiyle 2015 yılının son eyreėinin sonuna kadar 1739 adet hisse senedinin fiyatları girdi olarak deėerlendirilmiřtir. 7 adet algoritma kullanılmıřtır. Bu algoritmalar; Karar Aėaları, Destek Vektr Makineleri, JRip, Random Forest, Logistic Regresyon, Naive Bayes,

Bayes Ağı'dır. Random Forest algoritması diğer algoritmalara göre üstünlük sağlamıştır [21].

Darapaneni, Muthuraj, Prabakar ve Sridhar, Newyork taksi verilerini kullanarak ulaşım talebi ve elde edilecek geliri tahmin etmek için çeşitli zaman serisi tahmin modelleriyle denetimli öğrenme algoritmalarının performansını karşılaştırmıştır. 7 farklı algoritma kullanılmıştır. Bu algoritmalar; Doğrusal Regresyon, Karar Ağaçları, Rastgele Orman, Bagging, Boosting, Çok Değişkenli Zaman Serileri Tahmini, Otoregrif Hareketli Ortalama Modeli'dir. Uygulama sonucunda gelir tahmininde Çok Değişkenli Zaman Serisi algoritması üstünlük sağlarken, Rastgele Orman algoritması talep tahmini çalışmasında diğer algoritmalara göre üstünlük sağlamıştır [22].

Adinyira, Akoi-Adjei, Agyekum ve Fugar, Gana'da bulunan bir inşaat projesinin karını tahmin etmek üzerine çalışma gerçekleştirmiştir. 2014-2018 yılları arasında türetilen toplam 150 projeden elde edilen veriler kullanılmıştır. Matlab programlama dilinde Destek Vektör Regresyonu algoritması kullanılarak çalışma tamamlanmıştır. Çalışma sonucunda %73,66 oranında tahmin doğruluğu elde edilmiştir [23].

Cheriyian, İbrahim, Mohanan ve Treesa e-moda mağazasının verilerini kullanarak gelecekteki satış tahminlerini iyileştirmek için tahmin modellerinin detaylı çalışmasını ve analizini içeren çalışma gerçekleştirmiştir. 2015-2017 yılları arasında üç yıllık geçmiş satış verileri toplanmıştır. Bazı veri başlıkları; miktar, öge sayısı, çeyrek, satış geliri, yıl, haftadır. 3 farklı algoritma kullanılmıştır. Bu algoritmalar; Genelleştirilmiş Doğrusal Model, Karar Ağacı ve Gradient Boost Ağacı'dır. Uygulama sonucunda max doğruluk değerine sahip olan Gradient Boost algoritmasının diğer algoritmalara göre üstünlük sağladığı belirtilmiştir [24].

Tablo 1.1 Literatür incelemesi özet tablo

Yazar(lar)(Yıl)	Sektör	Kullanılan Algoritmalar	En İyi Sonuç Veren Algoritma(lar)
Budak ve Sarvari (2021) [6]	Lojistik	Rastgele Ormanlar, Dayanıkl (Robust) Regresyon ve XGBoost	XGBoost
Sepenu ve Eliasen (2022) [7]	Kozmetik	Lineer Regresyon, Destek Vektör Makineleri, Rastgele Ormanlar, Mini-Batch K-Means, BIRCH sınıflandırma, K-Ortalama, Hiyerarşik Kümeleme ve Gaussian Karışım Modeli	Rastgele Ormanlar, K-Ortalama
HÁJEK ve OLEJ (2010) [8]	Kamu	İleri Beslemeli Sinir Ağları, Destek Vektör Makineleri, Lineer Regresyon	İleri Beslemeli Sinir Ağları, Destek Vektör Makineleri
Li (2021) [9]	Medya	Lineer Regresyon,Bayesyen Regresyon,Gradyan Artırma Sınıflandırması, Naïve Bayes	Gradyan Artırma Sınıflandırması
Lin vd. (2013) [10]	Medya	Mevsimsel Oto regresif Entegre Hareketli Ortalama , En Küçük Kareler Genetik Algoritmalarla Vektör Regresyonunu , Bulanık En Küçük Kareler Genetik Algoritmalarla Vektör Regresyon	Bulanık En Küçük Kareler Genetik Algoritmalarla Vektör Regresyon

Tablo 1.2 Literatür incelemesi özet tablo(devamı)

Penpece ve Elma (2014) [11]	Perakende	Yapay Sinir Ağları	Yapay Sinir Ağları
Huang ve Yen (2019) [12]	Finans	Destek Vektör Makineleri, Bulanık kümeleme algoritmasına sahip genetik algoritma (GA), XGBoost, sınıflandırıcı derin inanç ağı (DBN) ve Derin İnanç Ağı - Destek Vektör Makineleri	XGBoost
Kermani vd. (2020) [13]	Gıda	Sınıflandırma ve Regresyon Ağacı, Ki-Kare Otomatik Etkileşim Dedektörü, Çok Katmanlı Algılayıcı Sinir Ağı ve Radyal Tabanlı İşlev Sinir Ağı	Çok Katmanlı Algılayıcı Sinir Ağı
Alkelani ve Awad (2020) [14]	Gıda	Karar Ağaçları	Karar Ağaçları
Rao vd. (2020) [15]	Gıda	Rastgele Orman, Doğrusal Regresyon	Rastgele Orman
Schmidt vd.(2022) [16]	Gıda	Lineer regreyon, Stochastic Gradient Descent Regresyon, Lineer Destek Vektör Regreyon, K-en Yakın Komşu Regresyon, Gauss Süreci Regresyonu, Yineleme Sinir Ağları, Zamansal Füzyon Transformatörü	Yineleme Sinir Ağları ve Zamansal Füzyon Transformatörü

Tablo 1.3 Literatür incelemesi özet tablo(devamı)

Chen vd. (2022) [17]	Finansal	Rastgele orman ve Stokastik Gradyan Artırma	Stokastik Gradyan Artırma
Kim vd. (2015) [18]	Medya	Çok Değişkenli Doğrusal Regresyon, Destek Vektör Regresyon, K-En Yakın Komşu Regresyon, Gauss Süreci Regresyonu	Çok Değişkenli Doğrusal Regresyon
Pundir vd. (2020) [19]	Perakende	Rastgele Orman, Vektör Oto regresyon Algoritmaları ve Geleneksel Yöntemler	Rastgele Orman, Vektör Oto regresyon Algoritmaları
Şim ve Pourhomayoun (2017) [20]	Medya	Denetimli ve Denetimsiz öğrenme yöntemlerinin birleşimi olan bazı algoritmalar	K-Means
Miloseviç (2016) [21]	Finans	Karar Ağaçları, Destek Vektör Makineleri, JRip, Rastgele Orman, Lojistik Regresyon, Naive Bayes, Bayes Ağı	Rastgele Orman
Darapaneni vd. (2019) [22]	Lojistik	Doğrusal Regresyon, Karar Ağaçları, Rastgele Orman, Bagging, Boosting, Çok Değişkenli Zaman Serileri Tahmini, Oto regregrif Hareketli Ortalama Modeli	Çok Değişkenli Zaman Serisi, Rastgele Orman
Adinyira vd. (2021) [23]	İnşaat	Destek Vektör Regresyonu	Destek Vektör Regresyonu

Cheriyen vd. (2018) [24]	Tekstil	Genelleştirilmiş Doğrusal Model, Karar Ağacı ve Gradient Boost Ağacı	Gradient Boost Ağacı
-----------------------------	---------	--	----------------------

Tablo 1.1’de incelenen çalışmalar değerlendirildiğinde en sık kullanılan makine öğrenmesi algoritmasının Rastgele Ormanlar olduğu tespit edilmiştir. Ayrıca sektör açısından değerlendirildiğinde tahmin çalışmalarının farklı sektörlerde yapıldığı gözlemlenmektedir.

1.2 Tezin Amacı

Uygulamanın gerçekleştirildiği katılım bankasında illere göre sektör ve banka POS ciroları göz önünde bulundurularak pazar payı hesaplaması yapılmaktadır. Strateji departmanından 6 aylık ve senelik hedef POS ciro değerleri şube ve bölge detayında belirlenmektedir. Dolayısıyla portföy sorumlularının genel hedefe ulaşmaları için POS ciro hedeflerini de gerçekleştirmeleri gerekmektedir. Bu noktada en kritik adım ise potansiyel POS ciro müşterisinin tespitidir. Böylece doğru ürünü, doğru zaman ve doğru satış stratejisiyle müşteriye sunacak hale getirmeye katkı sağlanacaktır. Mevcut durumda çalışmanın yapıldığı katılım bankasında ciro tahmini yapılmamaktadır. Bu çalışmanın amacı ise tüzel banka müşterilerinin POS ürününde elde edecekleri ciroyu en iyi tahmin eden algoritmanın tespit edilerek modelin kurulmasıdır.

Çalışmada Türkiye’nin önde gelen katılım bankasının Ağustos 2021 -Temmuz 2022 tarihleri arasındaki 12 aylık verileri manipüle edilerek kullanılmıştır. Çalışma R Studio yazılım dilinde yapılmış olup 10 Katlı Çapraz Doğrulama yöntemi kullanılarak Tek Değişkenli Lineer Regresyon, Çok Değişkenli Lineer Regresyon, Ridge Regresyon, Lasso Regresyon, K-En Yakın Komşu, Random Forest, Karar Ağaçları ve XGBoost algoritmaları kullanılmıştır. POS ciro tahminlemesi ile birlikte stratejik müşterilerin belirlenmesi, ciro değerinde artış yaşanması ve pazar payı oranının artırılması hedeflenmektedir.

1.3 Hipotez

Literatürde makine öğrenmesi yöntemlerinin, gelir tahmini ve ciro tahmini konularında kullanıldığı birçok çalışma bulunmaktadır. Bu çalışmaların bir kısmını ise bankacılık ve finans sektörü oluşturmaktadır. Müşteri açısından bu çalışmalar değerlendirildiğinde birçoğunun bireysel segment müşteri üzerinde çalışmalar yaptığı görülmektedir. Türkiye'nin önde gelen katılım bankalarından birinde gerçekleştirilen bu çalışmada tüzel müşteri segmenti kullanılarak tahmin çalışması yapılmıştır. Ayrıca literatürdeki ciro tahmini çalışmaları incelendiğinde yapılan çalışmalarda POS ürününe nadir şekilde yer verildiği görülmektedir. Yapılan çalışmanın literatürden farkı ise hem tüzel segmentteki müşterilerin kullanılması hem de POS ürününün ciro tahmininin yapılmasıdır. Tüzel müşterilerin POS cirolarını en iyi tahmin eden modelin kurulması için Tek Değişkenli Lineer Regresyon, Çok Değişkenli Lineer Regresyon, Ridge Regresyon, Lasso Regresyon, K-En Yakın Komşu, Random Forest, Karar Ağaçları ve XGBoost algoritmaları kullanılmıştır. Algoritmalarından en başarılı sonucu elde edebilmek adına 10 Katlı Çapraz Doğrulama Yöntemi uygulanmıştır.

2 MAKİNE ÖĞRENMESİ

Günümüzde insan popülasyonunun hızla artması, teknolojik gelişmeler ve internet kullanımı veri hacmini ve veri çeşitliliğini hızla arttırmaktadır. Hayatımızın her alanına giren teknolojiyle birlikte bireyler ve akıllı cihazlar günlük faaliyetlerde veri üretmeye devam etmektedir. Verinin çoğalmasıyla birlikte ise hayatımıza “Büyük Veri (Big Data)” kavramı girmiştir. Büyük veri, 5 bileşenden oluşmaktadır. Bu bileşenler; çeşitlilik, hız, hacim, gerçeklik ve değerdir. Sanayi, gıda, inşaat, tekstil, bankacılık gibi endüstriyel alanlarda ise verinin üretimi, analizi ve anlamlandırılmış hali hem Büyük Veriye kaynak oluştururken hem de işletmeler için stratejik aksiyonlara imkan tanımaktadır [25].

Makine öğrenme sisteminin oluşturulabilmesi için Şekil 2.1’de görüldüğü gibi 3 temel adımın oluşturulması gerekmektedir. Bu 3 adım; veri toplama, özellik üretme ve model eğitimidir [26].



Şekil 2.1 Makine öğrenmesi sistemi [26]

İlk adım olan veri toplama aşamasında bilgisayarın öğrenimini sağlamak amacıyla yeterli miktarda eğitim verisine sahip olunması gerekmektedir. Genellikle sistemin performansını arttırabilmek için daha fazla eğitim verisine ihtiyaç duyulmaktadır. Bu aşamada toplanan veriye ise genellikle ham veri denilmektedir. İkinci adım olan özellik üretme aşamasında ise toplanan ham verilerin niteliğine ve çalışılan alana göre analiz edilerek alt küme oluşturulmaktadır. Son adım olan model eğitimi aşamasında elde ettiğimiz verilerle matematiksel model oluşturmak için öğrenme algoritması seçimi yapılmaktadır [26].

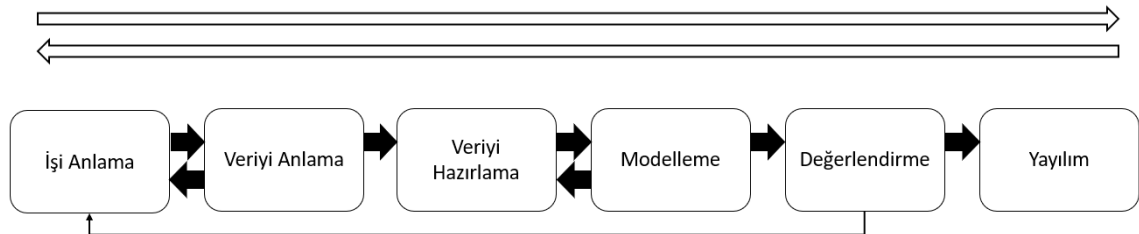
Veri setlerinin analizinde çok sayıda yöntem ve teknik kullanılmaktadır. Bu tekniklerden bazıları şöyledir: A/B Testi, Birliktelik Kuralı Yöntemi, Sınıflandırma,

Kümeleme Analizi, Genetik Algoritmalar, Makine Öğrenmesi, Doğal Dil İşleme, Regresyon Analizi [27].

Tez çalışmasında kullanılacak olan veri analiz tekniği ise “Makine Öğrenmesi”dir. Yapay zeka ve bilgisayar biliminin alt dallarından biri olan Makine Öğrenmesi insanların öğrenme şekillerini taklit ederek algoritmalar yardımıyla doğruluğun artırılmasıdır. Kullanılan algoritmalar matematiksel yöntemlerle tasarlanır ve geliştirilir. Örnek olarak görüntü işleme, hareket algılama, veri madenciliği, satış tahmini, hastalık tespiti verilebilir [28].

2.1 Makine Öğrenimi Süreci

Makine öğrenimi, deneyimle otomatik gelişen ve öğrenme sürecine sahip bilgisayar sistemleri oluşturma süreci olarak da tanımlanabilmektedir. Makine öğrenmesi çalışmalarının odak noktalarından biri otomatik olarak modeller üretmektir. Şekil 2.2’de “Veri Madenciliği İçin Sektörler Arası Standart Süreç (Cross-Industry Standard Process For Data Mining (CRISP –DM))” kısmen birbirine bağlı ve yinelemeli fazlardan oluşan altı adım gösterilmektedir. Tüm sürecin temelini oluşturan bu aşamalar; işi anlama, veriyi anlama, verinin hazırlanması, modelleme, değerlendirme ve yayımdır [29].



Şekil 2.2 Makine öğrenimi adımları (CRISP-DM) [29]

İlk adım olan işi anlama aşamasında projenin öğeleri ve gereksinimleri belirlenerek tüm proje için çözüm süreci tasarlanmaktadır. Veri anlama aşamasında verilerin toplanması ve veri kalitesinin artırılması sağlanmaktadır. Veri hazırlama aşamasında ise ana veriye dayalı olarak nihai veri kümesinin oluşturulması için gereken düzenlemeler tamamlanmaktadır. Verileri yüklemek, dönüştürmek ve temizlemek gibi adımlardan oluşmaktadır. Modelleme aşamasında matematiksel ve istatistiksel modellere dayanarak çeşitli modeller ve veri madenciliği tekniklerinin uygulanması gerçekleştirilmektedir. Değerlendirme aşamasında analizin kalitesini

yükseltmek için model gözden geçirilmektedir. Yapılan analizin sonucu işi anlama aşamasında belirlenen proje ve iş hedefleriyle karşılaştırılmaktadır. Uygulama aşamasında ise model hayata geçirilmektedir. Modelin çıktıları anlamlandırmak için düzenlenerek sunulmaktadır [29].

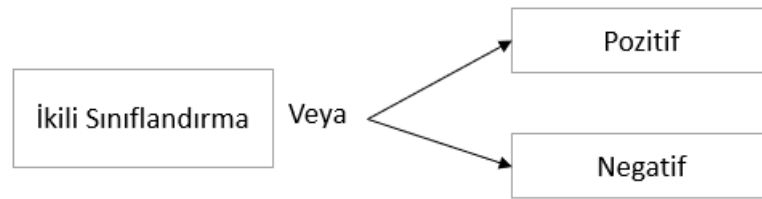
2.2 Makine Öğrenmesi Uygulama Alanları

Makine öğrenmesi alanları 4 farklı başlık altında incelenmektedir. Bu başlıklar; sınıflandırma, regresyon, kümeleme ve hibrit sistemlerdir [30].

2.2.1 Sınıflandırma

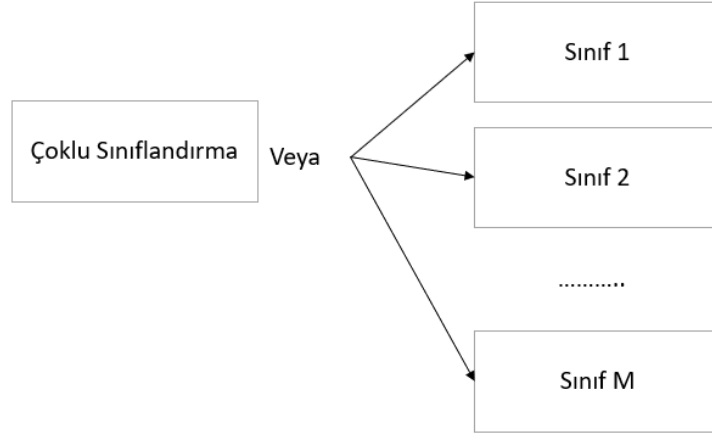
Önceden tanımlanmış kategorilerden bir veya daha fazlasının her bir maddeye atanma sürecine sınıflandırma denir. İkili ve çoklu sınıflandırma olarak ikiye ayrılır. İkili sınıflandırma, her bir maddenin iki kategoriden birine sınıflandırıldığı en basit sınıflandırma türüdür. İkili sınıflandırmaya daha fazla kategori eklenmesiyle birlikte çoklu sınıflandırma meydana gelmektedir. Çıktı değeri ise kesiklidir [30].

Şekil 2.3'te ikili sınıflandırma işlevsel olarak örneklendirilmiştir. Önceden tanımlanmış olan pozitif ve negatif sınıflar mevcuttur. Her bir öge pozitif veya negatif sınıfa ayrılır. Makine öğrenimi algoritmalarını incelerken ilk adımda verilen problemi ikili sınıflandırma olduğu varsayılır. Çoklu sınıflandırma veya regresyonda ikili sınıflandırmalara ayırma işlemi yapılabilir [30].



Şekil 2.3 İkili sınıflandırma sistemi [30]

Şekil 2.4'te ise çoklu sınıflandırma işlevsel olarak örneklendirilmiştir. Önceden tanımlanmış olan ikiden fazla olan M sayıda sınıf bulunmaktadır. Her bir öge M sayıda sınıfa ayrılır. Çoklu sınıflandırmayı ikili sınıflandırmaya ayırmak mümkündür [30].



Şekil 2.4 Çoklu sınıflandırma sistemi [30]

Şekil 2.5'te ise çoklu sınıflandırmanın ikili sınıflandırmaya ayrılmış hali görülmektedir. Ana problemde önceden tanımlanmış M sayıda kategori bulunmaktadır. Her bir öge kategori detayında sınıflandırılarak ilgili kategoriye atanmaktadır [30].



Şekil 2.5 Çoklu sınıflandırmanın ikili sınıflandırmaya ayrılmış hali [30]

2.2.2 Regresyon

Birden fazla faktöre dayalı olarak bir çıktı değerinin tahmin edilmesi sürecine regresyon denir. Tek değişkenli regresyon ve birden fazla çıktı değerinin bulunduğu çok değişkenli regresyon olmak üzere iki tür regresyon bulunmaktadır. Çıktı değeri sürekli dir. Bazı regresyon methodları şu şekildedir; lineer regresyon, non-lineer regresyon, lasso regresyon, lojistik regresyon [30].

Şekil 2.6’da tek bir girdi ve tek bir çıktı değişkenini tahmin eden tek değişkenli regresyon işlevsel olarak örneklendirilmiştir [30].



Şekil 2.6 Tek değişkenli regresyon [30]

Şekil 2.7’de tek bir girdi değişkeninin çoklu çıktı değişkenini tahmin eden çok değişkenli regresyon yaklaşımı işlevsel olarak örneklendirilmiştir. Çok değişkenli regresyon modelleri tek değişkenli regresyon modellerine ayrılabilir [30].



Şekil 2.7 Çok değişkenli regresyon [30]

Şekil 2.8’de ise zaman serisi tahmini işlevsel olarak örneklendirilmiştir [26]. Sürekli olarak kaydedilen veya belirli aralıklarla elde edilen verileri içeren serilere zaman serisi denmektedir. Regresyonda olduğu gibi tek ve çok değişkenli zaman serileri olarak ikiye ayrılır. Faiz oranı, üretim miktarı, elektrik sinyali vb. bu verilere örnek olarak verilebilir [31].



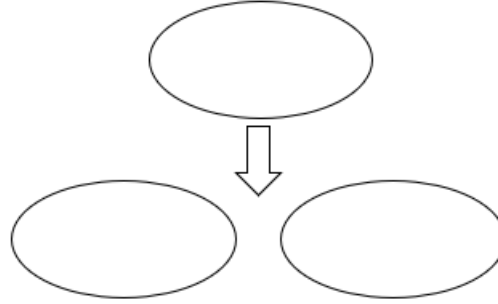
Şekil 2.8 Zaman serisi tahmini [30]

Sınıflandırma çalışmalarında çıktı değeri kesikli iken regresyon çalışmalarında çıktı değeri sürekli [30].

2.2.3 Kümeleme

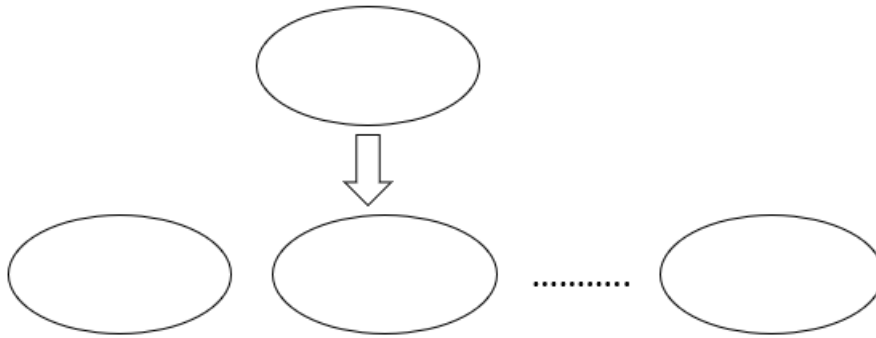
Bir grup öğeyi her biri benzer öğeler içeren alt gruplara ayırma işlemine kümeleme denir. Kümeleme çalışmasını yapabilmek için öğeler arasında benzerlikler olmalıdır.

Şekil 2.9 'da öğeler arasındaki benzerliklere göre iki alt gruba ayrılmış ikili kümeleme yaklaşımı işlevsel olarak örneklendirilmiştir. Başlangıçta tek bir öğe grubunun girdi olarak verildiği varsayılmaktadır [30].



Şekil 2.9 İkili sınıflandırma [30]

Şekil 2.10'da öğeler arasındaki benzerliklere göre k adet alt gruba ayrılmış çoklu kümeleme yaklaşımı işlevsel olarak örneklendirilmiştir. İkili kümeleme çalışmasında olduğu gibi başlangıçta tek bir grup vardır. k adet alt gruba ayrılır ve her bir öğe benzerliklerine göre bu alt gruplara atanır [30].

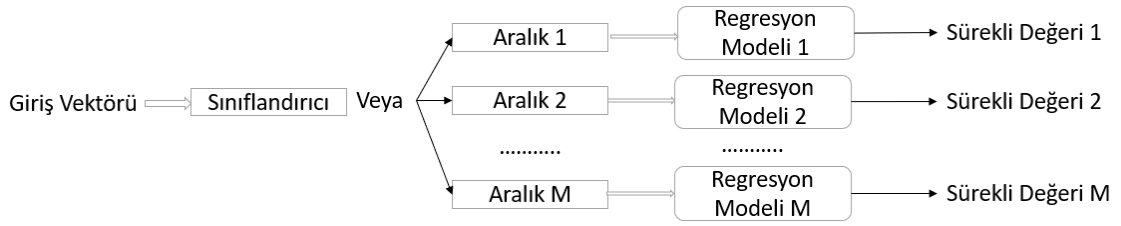


Şekil 2.10 Çoklu sınıflandırma [30]

2.2.4 Hibrit Sistemler

Hibrit sistemler, sınıflandırma, regresyon ve kümeleme çalışmalarının birleşimi olarak tanımlanmaktadır. Şekil 2.11'de regresyon ve sınıflandırma çalışmalarının karışımı olan hibrit sistem işlevsel olarak örneklendirilmiştir.

Girdi olarak bir öğeyi temsil eden vektör M aralıkta sınıflandırılır ve karşılık gelen aralıktaki değerler M adet regresyon modelinde tahmin edilir. Böylelikle M adet sürekli veri sınıfı oluşmaktadır. Burada regresyon ana görev iken sınıflandırma hatanın azaltılması için kullanılan ek görevdir [30].



Şekil 2.11 Regresyon ve sınıflandırma çalışmalarından oluşan hibrit sistem [30]

Şekil 2.12 'de kümeleme ve sınıflandırma çalışmalarının karışımı olan hibrit sistem işlevsel olarak örneklendirilmiştir. Etiketlenmemiş öğeler benzerliklerine bakılarak kümelenir. Bu sayede etiketlenen veriler kategori detayında sınıflandırılır. Sınıflandırma ana görev iken kümeleme ek görev niteliğindedir [30].



Şekil 2.12 Kümeleme ve sınıflandırma çalışmalarından oluşan hibrit sistem [30]

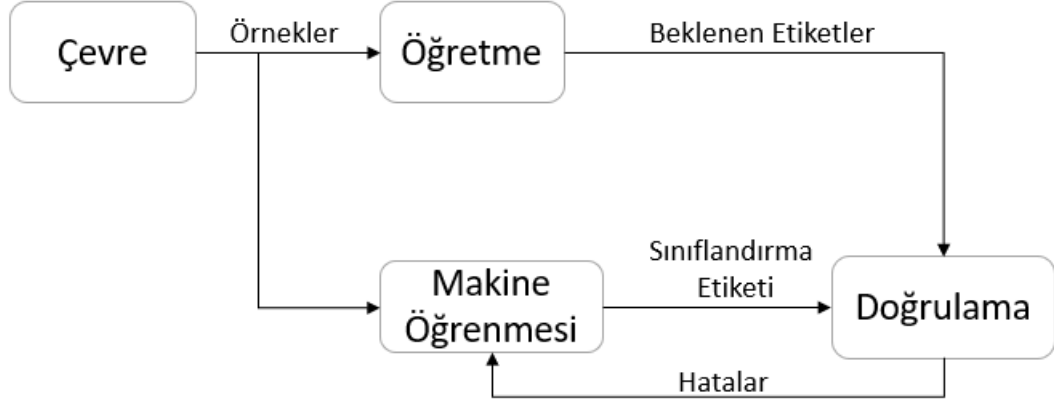
2.3 Makine Öğrenimi Türleri

Toplamda dört ana başlıkta incelenmektedir. Bu türler; Denetimli Öğrenme, Denetimsiz Öğrenme, Yarı Denetimli Öğrenme ve Pekiştirmeli Öğrenme' dir [32].

2.3.1 Denetimli Makine Öğrenmesi

Makine öğreniminin en yaygın türü olan Denetimli Öğrenme, bir bilgisayar programının bilinen örnek verilerin eğitildiği ve bilinen çıktı değerine ulaşıldığı süreçtir. Bu öğrenme türü girdi verilerini çıktı verileriyle ilişkilendirir. Öğrenilen kuralları yeni verilere uygulayarak bir bağlantı kurmayı amaçlar [33].

Hedef çıktı ile hesaplanan çıktı arasındaki farkı en aza indirmek için parametrelerin optimize edildiği öğrenme paradigması sınıflandırma ve regresyon uygulama alanlarında ve K En Yakın Komşu (KNN), Naive Bayes ve Karar Ağacı gibi algoritmalara sahiptir. Her eğitim örneğinde başlangıçta etiketlenen çıktıya hedef çıktı, algoritmanın hesapladığı çıktıya ise hesaplanan çıktı denir. Amaç ise hedeflenen çıktı ile hesaplanan çıktı arasındaki farkı en aza indirmektir. Şekil 2.13'de denetimli öğrenme süreci işlevsel olarak gösterilmektedir [30].

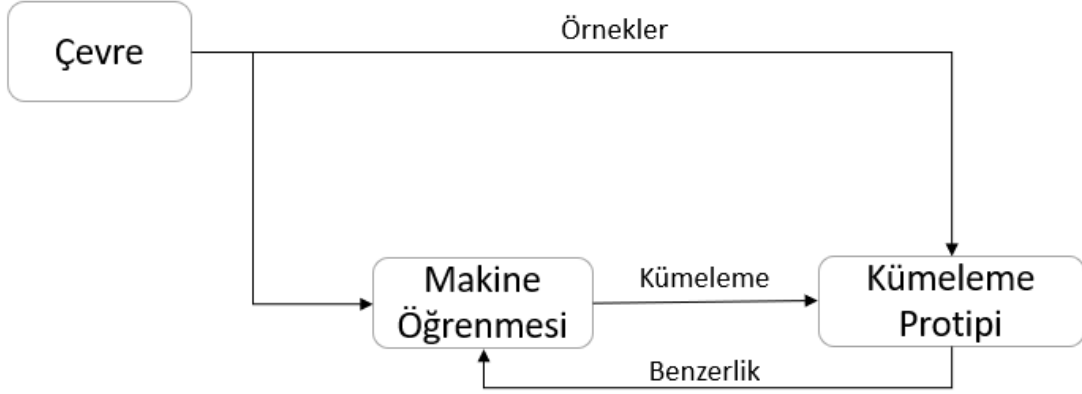


Şekil 2.13 Denetimli öğrenme [30]

2.3.2 Denetimsiz Makine Öğrenmesi

Denetimsiz öğrenme bilginin keşfedildiği bir sistemdir. Ayrıca öğretmensiz öğrenme olarak da tanımlanmaktadır. Önceden hedeflenmiş hedef çıktı değeri bulunmamaktadır. Kümeleme çalışmalarında kullanılmaktadır. Girdiler benzer özelliklere göre tanımlanarak kategorize edilmektedir [33].

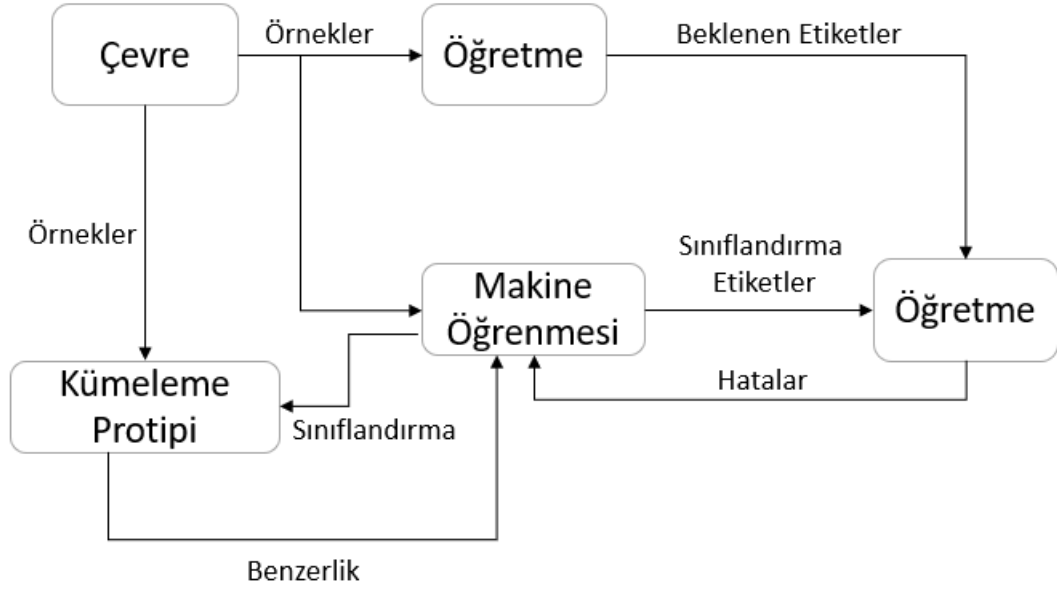
Bir başka deyişle denetimsiz öğrenme eğitim örnekleri arasındaki benzerlikleri göz önünde bulundurarak küme prototiplerini optimize etme süreci olarak tanımlanmaktadır. Çalışmada küme sayısına karar verilir ve prototipleme işlemi başlatılır. Gruptaki her veri ögesi prototipte olan en benzer kümeye atanır ve düzenlenmiş öğelerle daha benzer olacak şekilde güncellenir. Küme prototiplerinin düzenlenmesi ve güncellenmesi yakınsama gerçekleşene kadar devam etmektedir. Denetimli öğrenmede hesaplanan çıktı ve hedef çıktı arasındaki fark en aza indirmeye çalışılırken denetimsiz öğrenmede küme prototipleri arasındaki benzerlikleri en üst seviyeye çıkarmak amaçlanmaktadır. Şekil 2.14'te denetimsiz öğrenme süreci işlevsel olarak gösterilmektedir [30].



Şekil 2.14 Denetimsiz öğrenme [30]

2.3.3 Yarı Denetimli Makine Öğrenmesi

Yarı denetimli öğrenme, denetimli ve denetimsiz öğrenmenin kombinasyonudur. Öğrenme algoritmalarını eğitmek için hem etiketli hem de etiketsiz veriler kullanılmaktadır. Denetimli öğrenmede olduğu gibi etiketli verileri hedef çıktı ve hesaplanan çıktı arasındaki farkı en aza indirme çalışması yapılırken denetimsiz öğrenmede olduğu gibi etiketlenmemiş verilerin küme prototiplerinin arasındaki benzerliklerin artırılması hedeflenmektedir [30]. Etiketli verilerin elde edilmesi pahalı iken etiketsiz verilerin elde edilmesi ucuzdur. Bu nedenle de yarı-denetimli öğrenme sıklıkla tercih edilmektedir. Sınıflandırma ve regresyon çalışmalarının performansını arttırmak için etiketli verilerin yanı sıra etiketlenmemiş verilerin eklenmesi de sağlanmaktadır. Şekil 2.15'te yarı denetimli öğrenme süreci işlevsel olarak gösterilmektedir [33].

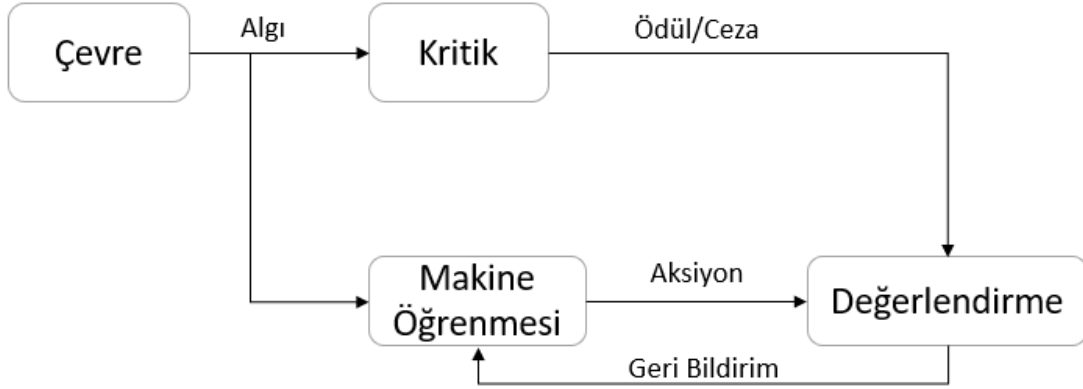


Şekil 2.15 Yarı denetimli öğrenme [30]

2.3.4 Pekiştirmeli Öğrenme

Ödülü en yüksek seviyeye cezayı en az seviyeye indirmeyi amaçlayan öğrenme türüne pekiştirmeli öğrenme denir. Girdi dış ortamdan alınırken çıktı eylem olarak üretilmektedir. Dış ortamın önemli olduğu bu öğrenme türünde çevre ve öğrenme aracı arasındaki etkileşim öğrenme olarak görülmektedir. Öğrenme aracı dış ortamdan gelen algıyı kabul eder ve çıktı olarak rastgele veya tablodan eylemi üretir. Ödül veya ceza çıktı olan eyleme verilir. Ödül verilirse çıktı istenen çıktı olarak belirlenirken ceza verilirse çıktı güncellenir [33].

Denetimli, denetimsiz ve yarı denetimli öğrenmede eğitim örnekleri sabit iken pekiştirmeli öğrenmede eğitim örnekleri değişkendir. Zamana bağlı olarak aynı girdi aynı çıktı sonucunu vermemektedir. Şekil 2.16'da pekiştirmeli öğrenme süreci işlevsel olarak gösterilmektedir [30].



Şekil 2.16 Pekiştirmeli öğrenme [30]

2.4 Makine Öğreniminde Karşılaşılan Sorunlar

Makine öğrenmesi yöntemleri eğitim ve test verilerinden oluşmaktadır. Eğitim verileri modelle sağlanan öğrenme sonrasında sinyal ve gürültü bileşenlerine ayrılmaktadır. Sinyal bileşeni modelle sağlanan öğrenme sonucunda elde edilen değerleri belirtmektedir. Gürültü bileşeni ise veri ile sinyal arasında oluşan farkı temsil etmektedir. Sinyal bileşeni verilerde bulunan düzenlilikleri temsil ederken gürültü bileşeni öngörülemez artıkları temsil etmektedir [26].

2.4.1 Aşırı Öğrenme-Eksik Öğrenme

Model, eğitim verisini gereğinden fazla ezberleyerek çok düşük bir tahmin hatasıyla yüksek başarı elde etmektedir. Test verisine uygulandığında ise çok yüksek tahmin hatasıyla düşük başarı elde etmektedir. Bu yaşanan öğrenme sürecine aşırı öğrenme (over-fitting) denilmektedir. Yaşanan soruna karşılık olarak eğitim verisi arttırılmaktadır veya model karmaşıklığı azaltılmaktadır [26].

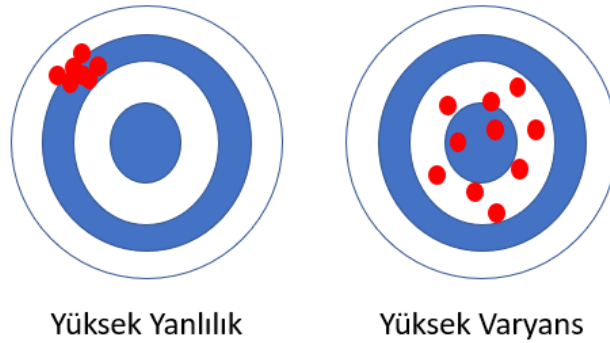
Model, eğitim verisini yetersiz öğrenemeyerek çok yüksek tahmin hatasıyla düşük başarı elde etmektedir. Girdi ile çıktı arasındaki bağlantı çözülememektedir. Girdi özelliklerinin çıktı özelliklerini yeterince açıklayamadığı durumlarda yaşanmaktadır. Yaşanan soruna karşılık olarak yeni girdi değişkenleri eklenerek modelin karmaşıklığı arttırılmaktadır. Şekil 2.17’de aşırı öğrenme ve eksik öğrenme sorunları işlevsel olarak gösterilmektedir [26].



Şekil 2.17 Aşırı-eksik-uygun öğrenme [26]

2.4.2 Yanlılık/Varyans İlişkisi

Makine öğrenmesi yöntemlerinde toplam beklenen hata iki gruba ayrılmaktadır: yanlılık ve varyans. Modellerde beklenen durum ise iki hatanın da minimum değerde olmasıdır [26].



Şekil 2.18 Yüksek yanlılık-varyans ilişkisi [26]

- **Yanlılık:** Model eğitim verisini yeterince öğrenememektedir. Bu nedenle de gerçek değer ile tahmin değeri arasında yüksek farklılıklar oluşmaktadır. Çözüm olarak girdi değişkenleri eklenmektedir ve model farklılaştırılmaktadır. Şekil 2.18’de görüldüğü gibi kırmızı noktalar rastgele bir eğitim veri seti üzerinde çalıştırılmış modeli temsil etmektedir. Öğrenilen modelin, öğrenme hedefinden olan uzaklığı yüksek yanlılık hatasına neden olmaktadır [26].
- **Varyans:** Model eğitim verisinde iyi performans gösterirken test verisinde kötü performans göstermektedir. Model tahmininin değişkenliğini ve hedef

öğrenme değerlerinden ne değerde saptığını göstermektedir. Şekil 2.18’de görüldüğü gibi tüm öğrenme sonuçları gerçek hedeften farklı bir şekilde rastgele sapmaktadır [26].

Varyans ve yanlılık hatası kullanılarak gerçek hata değeri aşağıdaki gibi hesaplanmaktadır:

$$\text{Öğrenme Hatası} = \text{bias}^2 + \text{variance} \quad (2.1)$$

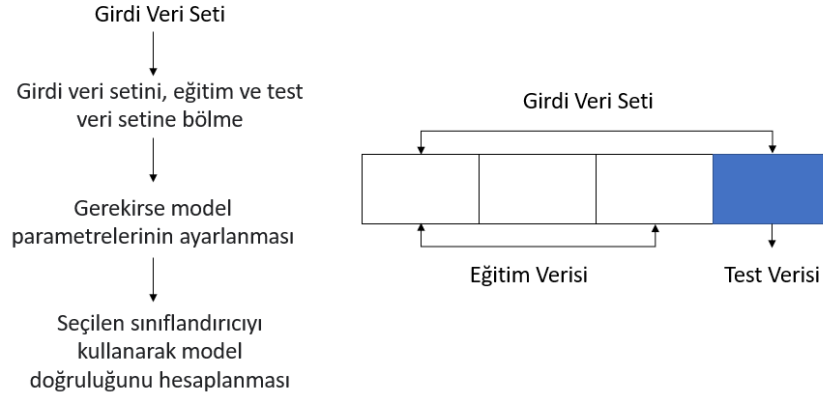
Belirli bir problem ve model seçimi yapıldığında hem yanlılık hatası hem de varyans hatasını aynı anda azaltılamamaktadır. Basit bir model seçimi yapıldığında düşük varyans, yüksek yanlılık hatasına neden olurken karmaşık bir yöntem seçildiğinde yüksek varyans, düşük yanlılık hatasına neden olmaktadır. Genellikle yanlılık ve varyans hataları dengede tutularak minimum öğrenme hatası elde edilmeye çalışılmaktadır. Bu duruma makine öğreniminde yanlılık- varyans çelişkisi denilmektedir [26].

2.5 Model Doğrulama Yöntemleri

Makine öğrenmesi yöntemlerinde modelin performansını ölçmek için model doğrulama yöntemleri kullanılmaktadır. Bu yöntemlerden bazıları şunlardır: Train&Test modeli, K-Katlı Çapraz Doğrulama Yöntemi, Leave-One-Out Çapraz Doğrulama Yöntemi’dir.

2.5.1 Train & Test Modeli

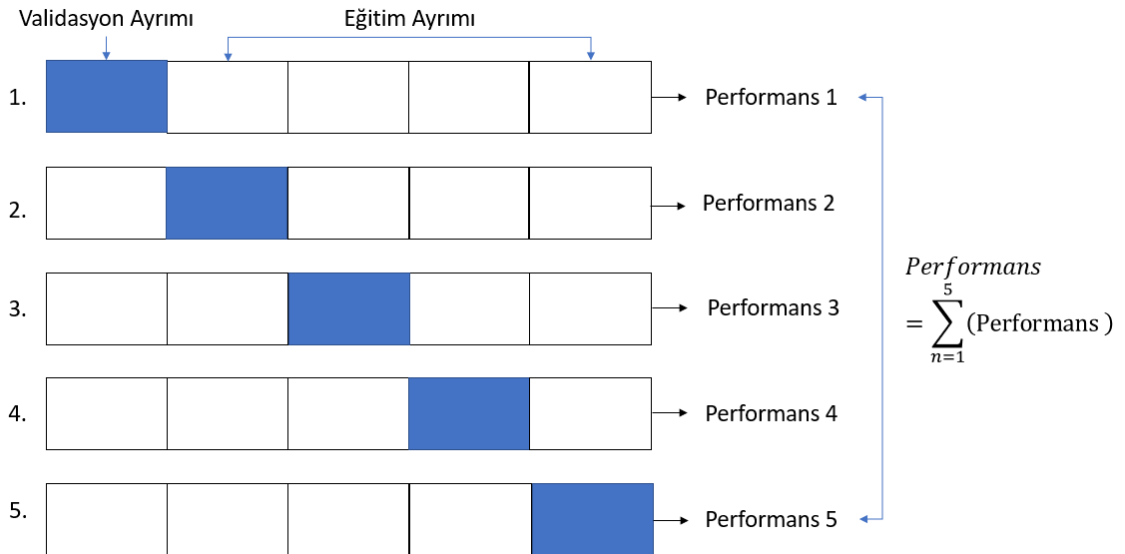
Bu yöntemde girdi veri seti rastgele 2 veya 3 kümeye ayrılmaktadır. Oluşan bu veri setleri ise şöyledir: eğitim veri seti, girdi veri seti, test veri seti. Şekil 2.19 ‘da Train & Test doğrulama yönteminin akışı gösterilmektedir [34].



Şekil 2.19 Train & Test Modeli [34]

2.5.2 K-Katlı Çapraz Doğrulama Yöntemi

Makine öğrenimi uygulamalarında model değerlendirilmesi açısından en sık kullanılan yöntemler arasına girmektedir. Çapraz doğrulama yönteminde veri kümesindeki her girdi test edilme şansına sahip olmaktadır. K katlı çapraz doğrulama yönteminde ise veri kümesini k parçaya ayrılmaktadır. Ayrılan k parçadan kalan k-1 parça ise eğitim verisini oluşturmaktadır. Şekil 2.20’de 5 katlı çapraz doğrulama yöntemi uygulanmış örneklem gösterilmektedir [35].



Şekil 2.20 5-Katlı Çapraz Doğrulama Yöntemi [35]

2.5.3 Leave-One-Out Çapraz Doğrulama Yöntemi

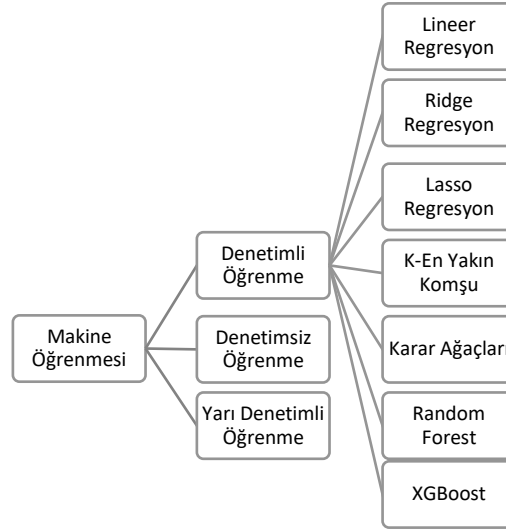
Bu yöntemde modeldeki girdi sayısının boyutuna göre n sayıya ayrılmaktadır. Her seferinde 1 parçası test için $n-1$ parçası ise eğitim verisi için kullanılmaktadır. $K=n$ şeklinde gösterilmektedir. K -katlı çapraz doğrulama yöntemine benzemektedir. Veri kümesinin boyutuna göre seçim yapılmaktadır. Boyut küçük olduğunda genelde Leave -One-Out Çapraz Doğrulama Yöntemi kullanılmaktadır [36].

Yapılan çalışmada Train & Test Modeli ve K -Katlı Çapraz Doğrulama Yöntemi yöntemleri kullanılmıştır.

2.6 Modelde Kullanılan Yöntemler

Makine öğreniminde problemleri çözmek için çok farklı algoritmalar kullanılmaktadır. Bu farklılıkların nedeni ise problemin çözümünde tek bir algoritmanın yüksek performans göstermeyeceğidir. Her problemde kullanılan algoritma problemin türüne, değişkenlerin sayısına, değişkenlerin boyutuna bağlı olarak değişiklik göstermektedir [28].

Araştırma algoritmaları R Studio yazılım dili kullanılarak gerçekleştirilmiştir. Makine öğrenmesi yaklaşımıyla denetimli öğrenmede kullanılan algoritmalar uygulanmıştır. Modelde kullanılan algoritmalar Şekil 2.21’de gösterilmektedir.



Şekil 2.21 Çalışmada kullanılan makine öğrenmesi yaklaşımı ve algoritmalar

2.6.1 Lineer Regresyon

İstatistikteki adı Doğrusal Regresyon olan Lineer Regresyon, bir veya birden fazla değişkenin arasındaki ilişkiyi modellemek için kullanılan popüler bir yöntemdir. Çalışmanın prensibi ise tahmin etmek için kullanılan x girdi değişkeninin, tahmin etmek istediğimiz y çıktı değişkeniyle doğrusal bir eşleşme oluşturmaktır. Literatürde x girdi değişkeni bağımsız değişken, y çıktı değişkeni ise bağımlı değişken olarak adlandırılmaktadır. Tek değişkenli ve çok değişkenli lineer regresyon olmak üzere ikiye ayrılmaktadır. Algoritmanın avantajı ise model tahmininin diğer modellere göre nispeten daha basit olması ve kapalı formulu bir çözümle çözümlenebilmesidir [26].

Modelde bulunan girdi değişkenlerini $x_1, x_2, \dots, x_k$, şeklinde gösterilirse y çıktı değişkeni ile birlikte oluşan denklem aşağıdaki gibidir:

$$E[Y|X_1=x_1, X_2=x_2, \dots, X_k=x_k] = \phi(x_1, x_2, \dots, x_k) \quad (2.2)$$

Modelde bulunan k değişkeni modelde bulunan girdi değişkenlerinin sayısını ifade ederken β_k değişkeni ise katsayıları ifade etmektedir. Oluşan lineer regresyon denklemi ise aşağıdaki gibidir [26].

$$\phi(x_1, x_2, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (2.3)$$

2.6.2 Ridge Regresyon

Çok değişkenli lineer regresyondaki eş doğrusallık probleminin ele alan Ridge regresyon, parametre tahmini yöntemidir. Çalışmada en küçük kareler toplamını minimize etmek için katsayılara ceza işlemi uygulanmaktadır. Değişkenlerin tamamını modele dahil eden bu yöntem, çıktıya etkisi olmayan değişkenlerin katsayısını sıfıra yaklaştırmaktadır [35].

2.6.3 Lasso Regresyon

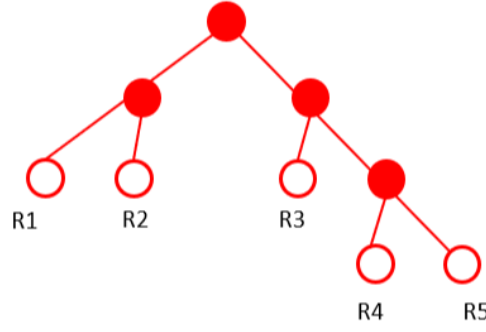
Ridge regresyonla aynı amaca sahip olan Lasso regresyon tahmin hatasını en aza indirerek modele en uygun katsayılara ceza işlemi uygulamaktadır. Modelle ilgisi olmayan katsayıları sıfıra eşitlemektedir. Sıfıra eşitlenen katsayıya sahip olan değişkenler modelden çıkarılmaktadır. Böylece modelin karmaşıklığı azalmaktadır [36].

2.6.4 K-En Yakın Komşu (KNN) Regresyon

KNN (K-Nearest neighbors' algorithm) ya da K-En yakın komşu olarak bilinen bu algoritma parametrik olmayan denetimli öğrenme türüdür. Çalışma, k adet aynı komşuluktaki gözlemlerin ortalamasını alarak bağımsız değişkenler arasındaki ilişkiyi tahmin etmektedir. Çalışmayı yapan kişinin belirlediği k değeri komşuluk sayısını ifade eder ve pozitif tam sayıdır. Hem sınıflandırma hem de regresyon çalışmalarında kullanılmaktadır [30].

2.6.5 Karar Ağaçları

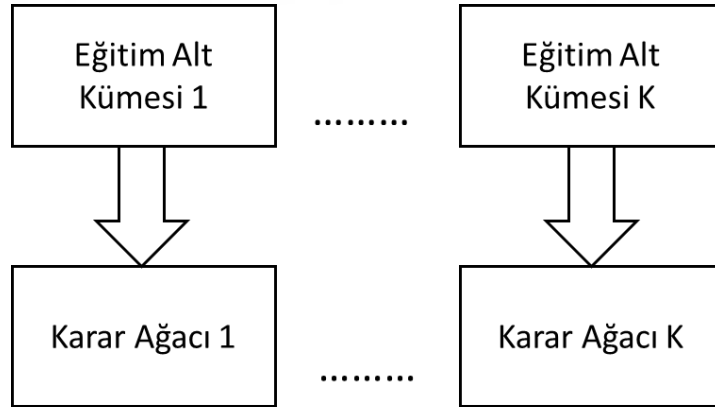
Seçimleri ve sonuçları ağaç yapısı şeklinde olan karar ağaçları, hem sınıflandırma hem de regresyon çalışmalarında kullanılmaktadır. Karar ağaçları düğüm ve dallardan oluşmaktadır. Düğümler bir olayı veya seçimi, dallar ise karar kurallarını veya koşullarını temsil etmektedir. Başlangıçtaki veri kümesinin daha küçük alt kümelere ayırarak karar ağacını oluşturmaktadır. Hem kategorik hem de nümerik verilerle işlenebilmektedir. Şekil 2.22 'de ise x'i girdi olarak kabul etmiş, y değerini üreten bir sistem işlevsel olarak örneklendirilmiştir [28].



Şekil 2.22 İki özellik alan bir karar ağacı modelinin gösterimi [26]

2.6.6 Rastgele Ormanlar

Hem sınıflandırma hem regresyon çalışmalarında kullanılan Rastgele Ormanlar algoritması, birden fazla karar ağacı oluşturmaktadır. Oluşan bu karar ağaçları, her biri birbirinden bağımsız ve eğitim verisinden rastgele elde edilmiş örneklemelere dayanmaktadır. Rastgele orman algoritmasının karar ağacı algoritmasından farkı ise düğümlerinin bölünmesi rastgele yapılmaktadır. Ayrıca bu algoritmanın Karar Ağacı algoritmasına göre avantajı aşırı uyum(fitting) sorununa çözüm getirmesidir [37].



Şekil 2.23 Rastgele ormanlar algoritmasının işlevsel gösterimi [30]

2.6.7 XGBoost

Tianqi Chen tarafından oluşturulan XGBoost(Extreme Gradient Boosting Tree-Aşırı Gradyan Artırma Ağacı) algoritması gradient boosting algoritmasının bir uygulamasıdır. Hem regresyon hem de sınıflandırma çalışmalarında kullanılmaktadır [38]. Çalışmada regresyon ağaçları tek tek oluşturulmaktadır. Böylece her model önceki modelin artıklarını minimize ederek kurgulanmaktadır.

Verilerin parçalanarak regresyon ağaçlarının oluşturulması algoritmanın daha küçük aralıklara bakmasıyla tahmin skorunu arttırmaktadır [39].

2.7 Model Performansının Ölçülmesi

Kullanılan algoritmaların performansı farklı istatistiksel ölçütler kullanılarak kıyaslanmıştır. Bu ölçütler; *Korelasyon katsayısı* (r), Ortalama Kare Hatası (MSE), Ortalama Karekök Sapması (RMSE), Ortalama Mutlak Değer (MAE)'dir.

2.7.1 Korelasyon Katsayısı (r)

Korelasyon katsayısı olarak adlandırılan r değeri iki değişken arasındaki ilişkinin düzeyini ve yönünü belirlemeye yardımcı olur. Bu değer -1 ile +1 arasında değer almaktadır. r^2 değeri ise 0 ile 1 arasında değer alırken iki değişkenin birbirine olan uyumunu göstermektedir. Bu değer 1 olması verilerin kusursuz bir doğrusal çizgiye uyduğunu göstermektedir. Bu değer 0 olması ise verilerin doğrusal çizgiye uyumsuz olduğunu göstermektedir. r^2 değerinin hesaplanma şekli aşağıda verilmiştir [40].

$$r^2 = 1 - \frac{\sum(o_i - t_i)^2}{\sum(o_i - \bar{o})^2} \quad (2.4)$$

o_i = deneysel olarak saptanmış değerler

t_i = regresyon eşitliğinden hesaplanmış değerler

$\bar{o}$ = deneysel verilerin ortalaması

r^2 değerinin 0-0.25 arasında olması ilişki düzeyi çok zayıf, 0.20-0.49 arasında olması ilişki düzeyi zayıf, 0.50-0.69 arasında olması ilişki düzeyi orta, 0.70-0.89 arasında olması ilişki düzeyi yüksek, 0.90-1.00 arasında olması ilişki düzeyi çok yüksek şeklinde yorumlanmaktadır [40].

2.7.2 Ortalama Kare Hatası

Ortalama Kare Hatası (MSE - Mean Squared Error) değeri ölçüm değerleri ile model tahminleri arasındaki farkın kareleri toplamının örneklem sayısına bölümüne eşittir. Bu ölçümde kare alındığından yön belirtmemektedir [41].

$$MSE = \frac{1}{N} \sum(o_i - t_i)^2 \quad (2.5)$$

2.7.3 Ortalama Karekök Sapması

Ortalama Karekök Sapması (RMSE - Root Mean Squared Error) değeri ölçüm değerleri ile model tahminleri arasındaki farkların hesaplanmasında kullanılan bir ölçüdür. Bu farklar artık olarak nitelendirilmektedir. RMSE değeri 0 ile ∞ arasında değer alabilmektedir. RMSE değerinin 0'a yakın olması model tahmin yeteneğinin yüksek olduğunu gösterirken, değerın büyümesi modelin tahmin yeteneğinin düşük olduğunu göstermektedir. RMSE değerinin hesaplanma şekli aşağıda verilmiştir [40].

$$RMSE = \sqrt{\frac{\sum (oi - ti)^2}{N}} \quad (2.6)$$

N =gözlem değeri

2.7.4 Ortalama Mutlak Değer

MAE (Mean Absolute Error-Ortalama Mutlak Değer) değeri ölçüm değerleri ile model tahminleri arasındaki farkın yönünü dikkat almadan ortalama büyüklüğü ölçmektedir. RMSE değeri 0 ile ∞ arasında değer alabilmektedir. Bu değerin 0'a yaklaşması modelin tahmin yeteneğinin yüksek olduğunu, değerin büyümesi modelin tahmin yeteneğinin düşük olduğunu göstermektedir. MAE değerinin hesaplanma şekli aşağıda verilmiştir [40].

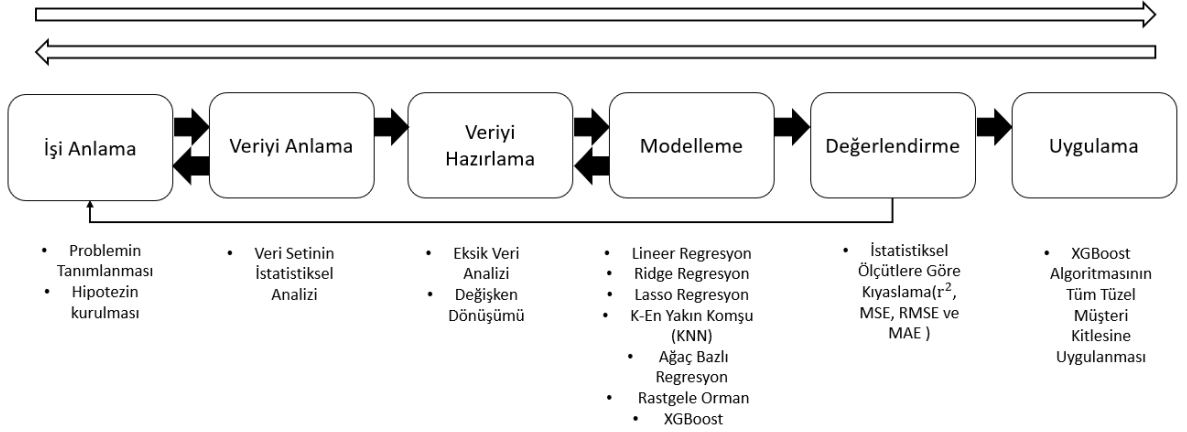
$$MAE = \frac{1}{N} \sum |oi - ti| \quad (2.7)$$

Ödeme sistemleri ödeme alma ve ödeme verme olarak temelde ikiye ayrılmaktadır. Günümüzde ise artan tüketim bilinciyle birlikte hem ödeme alma hem de ödeme verme araçlarına olan talepte yüksek artış meydana gelmektedir. Ödeme almayı sağlayan POS cihazları ise bu noktada müşteriler için farklılaşmaktadır. POS cihazları Elektronik Para Transferi (EFT) sağlayan ve sanal-fiziki olmak üzere ikiye ayrılan ödeme alma yöntemidir. Uygulamanın yapıldığı bankada ise kullanılan sektör ve kullanım şekline göre 7 farklı POS çeşidi bulunmaktadır.

Kullanıcılar ödeme verme araçlarından biri olan kart veya telefonlarındaki mobil şubeden elde ettikleri QR kodu aracılığıyla ödeme alma aracı olan POS cihazlarından işlem yapmaktadır. POS cihazlarında yapılan işlemlerin tutarı ise ciro değerini oluşturmaktadır. Ciro değeri yapılan işleme göre değişkenlik göstermektedir.

POS cihazlarında müşteriye ve bankaya göre çeşitlilik gösteren çalışma koşulları bulunmaktadır. Çalışma koşulları 3 'e ayrılmaktadır. Bu koşullar: blokeli, komisyonlu ve hem blokeli hem komisyonlu şeklindedir. Bankalar ise müşterinin çalışma koşuluna göre POS üzerinden geçen işlem tutarına göre kazanç sağlamaktadır. Bu nedenle bankalar arasında müşterilerin POSları üzerinden yaptıkları ciro değerleri oldukça önemli hale gelmektedir. Bankalararası Kart Merkezi (BKM) tarafından her ay tüm bankaların POS cihazlarından geçen ciro bilgisi, banka sıralaması ve pazar payı oranları paylaşılmaktadır. Bankalar ve ödeme kuruluşları için hedeflenen durum ise en yüksek ciro değerine ve en yüksek pazar payı oranına sahip olmaktır.

Çalışma kapsamında CRISP-DM (Cross-Industry Standard Process for Data Mining) yönteminin adımları gerçekleştirilmiştir. Şekil 3.1'de uygulamanın adımları özet olarak gösterilmektedir.



Şekil 3.1 Uygulamanın adımları

3.1 Problemin Tanımlanması

Uygulamanın gerçekleştirildiği bankada tüzel müşterilerinin POS cihazlarından geçen ciro değerini arttırmak ve pazar payında artış sağlamak için bölge ve portföy sorumlularına 6 aylık ve yıllık POS ciro hedefleri verilmektedir. Bu noktada portföy sorumlularının potansiyel müşterileri tespit ederek müşteriye özel çalışma koşulu belirlemesi, kampanya oluşturması ya da müşteriye uygun ürünü sunması gerekmektedir.

Banka müşterilerinin aylık şekilde POS cihazlarından geçireceği ciro değerinin tahmini ise portföy sorumlularının potansiyel müşteriye bulmasına olanak sağlayacaktır. Mevcut bankada böyle bir çalışma bulunmamaktadır. Bunun sonucu olarak portföy sorumluları potansiyel müşteriye tespit etmekte zorlanırken POS ciro hedeflerini tutturamamaktadır. Çalışmanın hedefi ise denetimli makine öğrenmesi yöntemleri kullanılarak müşterilerin POS cihazlarından geçen ciroyu nokta atışı şeklinde tahmin etmektedir. Yapılan tahmin sonucunda portföy sorumlusu kendi müşterisi olan ve gelecek ay yüksek POS ciro değerine sahip potansiyel müşterileri tespit edecektir. Bu sayede portföy sorumlularına verilen 6 aylık ve yıllık POS ciro hedefleri tutturularak diğer bankalar arasında POS ciro artışı ve pazar payı oranında artış meydana gelecektir.

Çalışma kapsamında beş ayrı aşama uygulanmıştır. Bu aşamalar; verinin anlamlandırılması, verinin hazırlanması, modelleme, değerlendirme ve yayılımdır. Verinin hazırlanma aşamasından sonra düzenlenen verilerle bankanın her bir kurumsal müşteriden elde edeceği POS cirosunu tahmin etmek için makine

öğrenmesi algoritmalarından Lineer Regresyon (Tek Değişkenli- Çok Değişkenli), Ridge Regresyon, Lasso Regresyon, K-En Yakın Komşu (KNN), Ağaç Bazlı Regresyon, Rastgele Orman, XG boost algoritmaları kullanılmıştır.

Makine öğrenmesi algoritmalarında en çok kullanılan açık kaynaklardan biri olan R-Studio programlama dili kullanılmıştır. Notlar, kodlar, çıktı ve görseller bir arada değerlendirilmiştir.

3.2 Verinin Hazırlanması

Türkiye'nin önde gelen katılım bankalarının birisinden alınan veriler Kişisel Verilerin Korunması Kanunu gereği belirli bir oranda manipüle edilerek kullanılmıştır. Kanun kapsamında veriler paylaşılmayacak olup algoritma ve modeller konusunda oluşan yorumlar paylaşılacaktır.

Çalışma kapsamında 2021 yılı ağustos ile 2022 yılı temmuz ayları arasında gerçekleşen veriler kullanılmıştır. Veriler 16 farklı bölgenin 402 şubesinden rastgele şekilde seçilmiştir. Veri seti 1500 gözlem biriminden ve 23 adet farklı değişkenden oluşmaktadır. Gözlem birimleri müşterileri ifade ederken değişkenler POS cirosunu etkileyebilecek müşteri bilgilerinden oluşmaktadır. Değişkenlerin 22 tanesi nümerik olup, 1 tanesi "EVET-HAYIR" değerlerinden oluştuğu için kategoriktir. Değişkenlerin açıklamaları aşağıdaki Tablo 3.1'de verilmektedir.

Tablo 3.1 Değişken Açıklamaları

Değişken	Açıklama
Müşteri No	Müşteriye özel tanımlanan minimum 4 hane maksimum 8 haneli rakamdan oluşan ID'dir.
Bölge	Müşterinin kayıtlı olduğu şubenin bulunduğu bölgedir.
Şube	Müşterinin kayıtlı olduğu şubedir.
MCC Kod	Müşterinin çalıştığı sektörün bulunduğu koddur.
MCC Bilgi	Müşterinin çalıştığı sektördür.

Tablo 3.1 Değişken Açıklamaları(devamı)

Ciro Yıllık Değişim Oranı	TÜİK verilerine göre 2021 yılının tamamında sektör bazlı gerçekleşen büyüme yüzdesidir.
Müşteri Segmenti	Müşterinin ürün, mevduat ve müşteri tipine bakılarak değerlendirilen segmenttir.
Müşteri Tipi	Müşterinin çalışan sayısına göre belirlenen "Kurumsal-Business" etiketidir.
Ciro	Müşterinin 2021 yılı ağustos ile 2022 yılı temmuz ayları arasında 12 aylık dönemde gerçekleştirdiği POS cirosudur.
İşlem Adedi	Müşterinin 2021 yılı ağustos ile 2022 yılı temmuz ayları arasında 12 aylık dönemde gerçekleştirdiği POS işlem adedidir.
POS Kayıt Tarihi	Müşterinin bankadan POS'unu aldığı ilk yıldır.
BKM TechPOS	Kart kabul kuruluşlarının ve POS/ÖKC üreticilerinin BKM üzerinden karşılıklı çalışabildiği bir altyapıdır.
MPOS	Üye işyerlerinin cep telefonlarını POS cihazı gibi kullanıp ödeme alabilecekleri bir uygulamadır.
Mobil POS	Müşteri yanında ödeme alan, iş yeri dışında da tahsilata imkan veren, GSM/GPRS ağı üzerinden çalışan POS cihazlarıdır.
Sabit POS	Sabit telefon hattı üzerinden belirli bir noktada sabit şekilde kullanılan POS sistemidir.
Sanal POS	E-ticaret sitelerinden yapılan alışverişlerin gerçekleşmesinde kullanılan POS sistemidir.

Tablo 3.1 Değişken Açıklamaları(devamı)

Standart Ökc	ÖKC (Ödeme Kaydedici Cihazlar) cihazları, Maliye Bakanlığı Gelir İdaresi Başkanlığına yapılan işlemlerin verisini aktaran bilgisayar tabanlı POS sistemidir.
Tahsilat POS	İnternet aracılığıyla tahsilat yapılmasını sağlayan POS sistemidir.
Toplam	Müşterinin toplam POS adedini belirtmektedir.
Verimlilik	Müşterinin kullandığı ürün, cari hesap bilgileri, segmenti vb. bilgilerle hesaplanan değerdir.
Cari Hesap	Müşterinin cari hesabında bulundurduğu tutardır.
Yıllık POS Katılma Hesabı	Müşterinin yıllık POS katılma hesabında bulundurduğu tutardır.
Toplam Mevduat	Müşterinin cari mevduat, katılım mevduat gibi tüm mevduatların toplamını içeren tutardır.

Tablo 3.2’de modelin girdi değişkenlerinin tanımlayıcı bilgileri mevcuttur. Bu bilgiler maksimum değer, minimum değer, ortalama, medyan, mod, standard sapma ve varyans değerleridir.

Tablo 3.2 Veri setinin istatistiksel bilgileri

Değişken	Max	Min	Mean	Median	Mode	Std	Variance
Müşteri No	1500,00	1,00	751,00	751,00	1,00	433,00	187489,00
Müşteri Segmenti	4,00	1,00	4,00	4,00	4,00	1,46	2,13
Kurumsal Mı ?	1,00	0,00	0,00	0,00	0,00	0,12	0,01
Ciro202108	7411672,00	10,00	159419,00	30867,00	5250,00	562426,00	316323005476,00
İşlem Adedi202108	58437,00	1,00	625,00	79,00	1,00	2451,00	6007401,00
Ciro202109	2147483647,00	0,00	7023068,00	25104,00	0,00	204709422,00	41905947455574100,00
İşlem Adedi202109	124356,00	0,00	742,00	78,00	0,00	4186,00	17522596,00
Ciro202110	2147483647,00	0,00	38623966,00	24587,00	0,00	528297660,00	279098417561476000,00
İşlem Adedi202110	153583,00	0,00	702,00	67,00	0,00	4702,00	22108804,00
Ciro202111	2147483647,00	0,00	29620780,00	22444,00	0,00	397635689,00	158114141166505000,00
İşlem Adedi202111	64940,00	0,00	616,00	61,00	0,00	2820,00	7952400,00
Ciro202112	2147483647,00	0,00	30318850,00	18340,00	0,00	410514753,00	168522362430651000,00
İşlem Adedi202112	53656,00	0,00	568,00	50,00	0,00	2423,00	5870929,00

Tablo 3.2 Veri setinin istatistiksel bilgileri(devamı)

Ciro202201	2147483647,00	0,00	38518099,00	19685,00	0,00	466932385,00	218025852161788000,00
İşlem Adedi202201	91923,00	0,00	590,00	44,00	0,00	3244,00	10523536,00
Ciro202202	2147483647,00	0,00	30632104,00	23033,00	0,00	377223582,00	142297630816911000,00
İşlem Adedi 202202	93044,00	0,00	654,00	51,00	0,00	3484,00	12138256,00
Ciro202203	2147483647,00	0,00	34724702,00	24108,00	0,00	393138927,00	154558215922711000,00
İşlem Adedi202203	55810,00	0,00	609,00	47,00	0,00	2573,00	6620329,00
Ciro202204	2147483647,00	0,00	34724702,00	24108,00	0,00	393138927,00	154558215922711000,00
İşlem Adedi 202204	55810,00	0,00	609,00	47,00	0,00	2573,00	6620329,00
Ciro202205	2147483647,00	0,00	39153696,00	24491,00	0,00	449670967,00	202203978562715000,00
İşlem Adedi202205	50485,00	0,00	598,00	45,00	0,00	2563,00	6568969,00
Ciro202206	2147483647,00	0,00	41540001,00	24260,00	0,00	446698841,00	199539854550743000,00
İşlem Adedi202206	48350,00	0,00	554,00	37,00	0,00	2246,00	5044516,00
Ciro202207	2147483647,00	0,00	47537050,00	27486,00	0,00	499032235,00	249033171569095000,00
İşlem Adedi202207	51856,00	0,00	601,00	41,00	0,00	2548,00	6492304,00
Ciro202208	2147483647,00	0,00	38152591,00	29745,00	0,00	483604587,00	233873396567441000,00
İşlem Adedi202208	59421,00	0,00	651,00	41,00	0,00	2746,00	7540516,00
POS KAYIT TARİHİ	2022,00	2001,00	1807,00	2019,00	2019,00	618,00	381924,00
BKM TechPOS	5,00	0,00	0,00	0,00	0,00	0,28	0,08
Cebim POS	1,00	0,00	0,00	0,00	0,00	0,47	0,23
Diğer Cihaz	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Mobil POS	19,00	0,00	1,00	0,00	0,00	1,36	1,86
Sabit POS	29,00	0,00	0,00	0,00	0,00	0,92	0,85
Sanal POS	10,00	0,00	0,00	0,00	0,00	0,58	0,33
Standart Ökc	30,00	0,00	1,00	0,00	0,00	1,68	2,81
Tahsilat POS	4,00	0,00	0,00	0,00	0,00	0,15	0,02
Toplam	32,00	0,00	2,00	1,00	1,00	2,41	5,79
Verimlilik	9452824,00	-80143,00	44287,00	4532,00	0,00	311604,00	97097052816,00
Cari Hesap	2147483647,00	0,00	1144311605,00	114367,00	0,00	2147483647,00	4611686014132420000,00
Yıllık POS Katılma Hesabı	2147483647,00	0,00	25578530,00	0,00	0,00	316796183,00	100359821563369000,00
Toplam Mevduat	2147483647,00	0,00	1171425225,00	191667,00	0,00	2147483647,00	4611686014132420000,00

Aşağıdaki Tablo 3.3’de model girdi değişkenlerinin boş olmayan sayısını ve verinin türünü göstermektedir. Data type sütununda bulunan “int” değeri tamsayı veri tipini, “float” ondalık sayı veri tipini ve “object” ise kategorik değişkeni ifade etmektedir.

Tablo 3.3 Veri seti hakkında genel bilgi içeren program çıktısı

	Değişken	Non-Null Count	Data Type
1	Müşteri No	1500 non-null	int
2	Ciro Yıllık Değişim Oranı	1500 non-null	float
3	Müşteri Segmenti	1500 non-null	object
4	Kurumsal Mı ?	1500 non-null	object
5	Ciro202108	1500 non-null	float
6	İşlem Adedi202108	1500 non-null	int
7	Ciro202109	1500 non-null	float
8	İşlem Adedi202109	1500 non-null	int
9	Ciro202110	1500 non-null	float
10	İşlem Adedi202110	1500 non-null	int
11	Ciro202111	1500 non-null	float
12	İşlem Adedi202111	1500 non-null	int

Tablo 3.3 Veri seti hakkında genel bilgi içeren program çıktısı (devamı)

13	Ciro202112	1500	non-null	float
14	İşlem Adedi202112	1500	non-null	int
15	Ciro202201	1500	non-null	float
16	İşlem Adedi202201	1500	non-null	int
17	Ciro202202	1500	non-null	float
18	İşlem Adedi 202202	1500	non-null	int
19	Ciro202203	1500	non-null	float
20	İşlem Adedi202203	1500	non-null	int
21	Ciro202204	1500	non-null	float
22	İşlem Adedi 202204	1500	non-null	int
23	Ciro202205	1500	non-null	float
24	İşlem Adedi202205	1500	non-null	int
25	Ciro202206	1500	non-null	float
26	İşlem Adedi202206	1500	non-null	int
27	Ciro202207	1500	non-null	float
28	İşlem Adedi202207	1500	non-null	int
29	Ciro202208	1500	non-null	float
30	İşlem Adedi202208	1500	non-null	int
31	POS KAYIT TARİHİ	1500	non-null	int
32	BKM TechPOS	1500	non-null	int
33	Cebim POS	1500	non-null	int
34	Diğer Cihaz	1500	non-null	int
35	Mobil POS	1500	non-null	int
36	Sabit POS	1500	non-null	int
37	Sanal POS	1500	non-null	int
38	Standart Ökc	1500	non-null	int
39	Tahsilat POS	1500	non-null	int
40	Toplam	1500	non-null	int
41	Verimlilik	1500	non-null	float
42	Cari Hesap	1500	non-null	float
43	Yıllık POS Katılma Hesabı	1500	non-null	float
44	Toplam Mevduat	1500	non-null	float

3.3 Veri Setinin Analizi

Veri analizi çalışmalarında doğru sonucu elde edebilmekte ve yüksek başarı oranına sahip olmakta girdi veri seti çok önemli role sahiptir. İşlenmemiş verinin toplanması ve istatistiki açıdan anlamlandırılması sonucunda girdi veri seti anlamlı hale gelmektedir. Çalışma kapsamında ilk olarak girdi veri setini oluşturmak için veri analizi çalışmaları yapılmıştır.

Veri seti oluşturulurken CRISP-DM (Cross-Industry Standard Process for Data Mining) yöntemi kullanılmıştır. Şekil 3.1’de ise adımlar gösterilmektedir. Toplamda 6 adımdan oluşmaktadır. Bu adımlar; işin anlamlandırılması, verinin

anlamlandırılması, verinin hazırlanması, modelleme, değerlendirme ve yayılımdır. Çalışmada öncelikle problem tanımı yapılmış olup gerekli değişkenler belirlenmiştir. Bölüm 3.2’de ise verinin anlamlandırılabilmesi için istatistiksel ölçütlere (max, min, mean, median, mode, std, variance) ve veri tipine göre sınıflandırma yapılmıştır. Verinin hazırlanması aşamasında ise veri temizleme, veri dönüşümü ve veri standardizasyonu/normalizasyonu yapılmıştır.

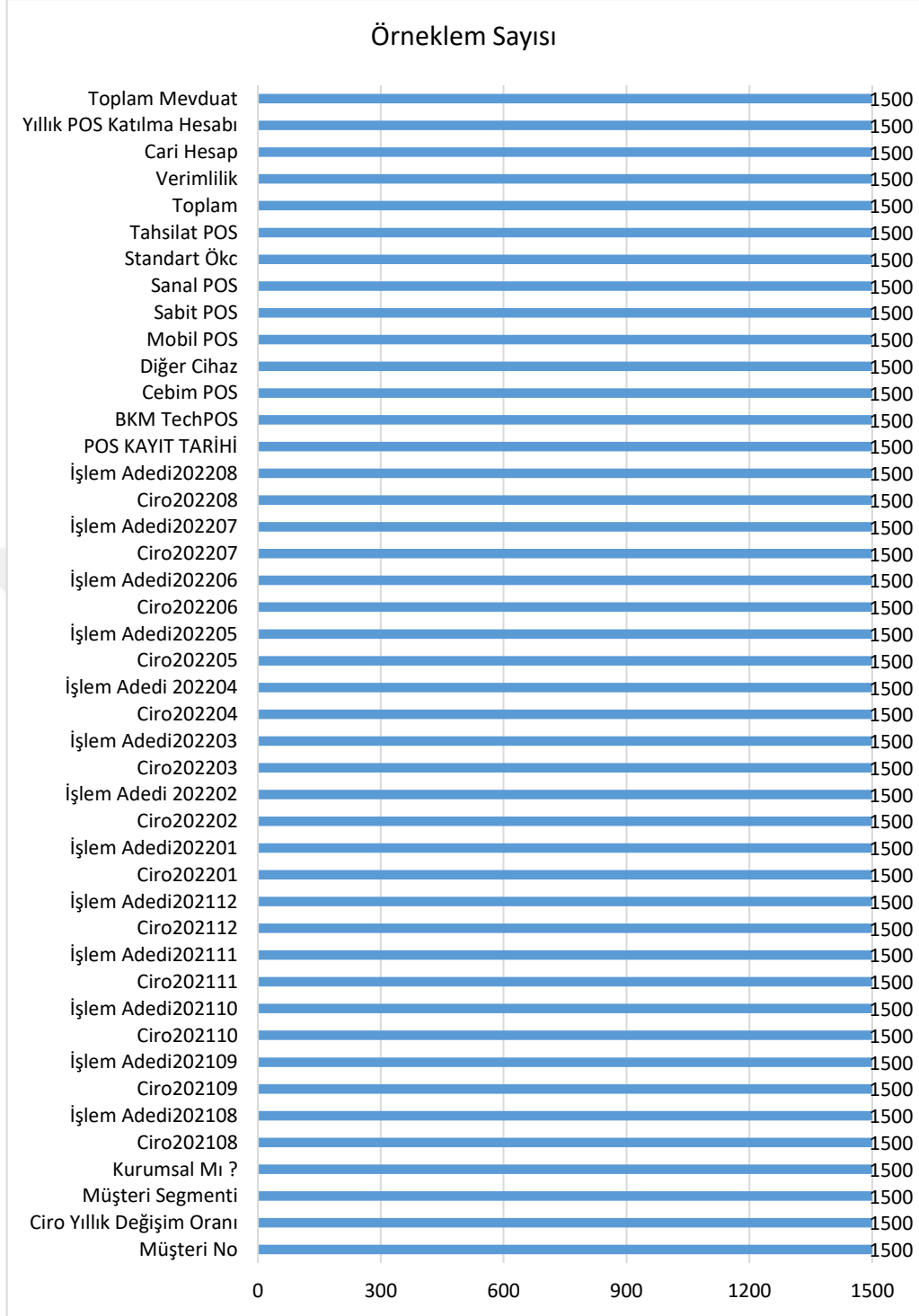
3.3.1 Veri Temizleme

Veri temizleme aşamasında gürültülü, eksik veya tanımsız veriler ayıklanarak veri seti anlamlı hale getirilmeye çalışılmaktadır. Bu adımlardan bazıları şöyledir: gürültülü, eksik veya yinelenen verilerin çıkarılması, eksik değerlerin yerine farklı değerlerin tanımlanması, en sık kullanılan ortak verinin girilmesi [42].

Uygulamada kullanılan veri seti Tablo 3.2 ve Tablo 3.3’te incelenmiş olup gürültülü, bozuk ve yinelenen değerlere ulaşılmamıştır.

3.3.2 Eksik Veri Analizi

Uygulamada kullanılan veri setinde bulunan örneklem sayısına bakıldığında eksik değerlerin bulunmadığı tespit edilmiştir. Şekil 3.1’de ise her bir değişken detayında örneklem sayısı görselleştirilmiştir. Girdi veri setinde 1500 örneklem bulunurken her bir değişken detayında örneklem sayısının bu değere eşit olduğu görülmektedir.



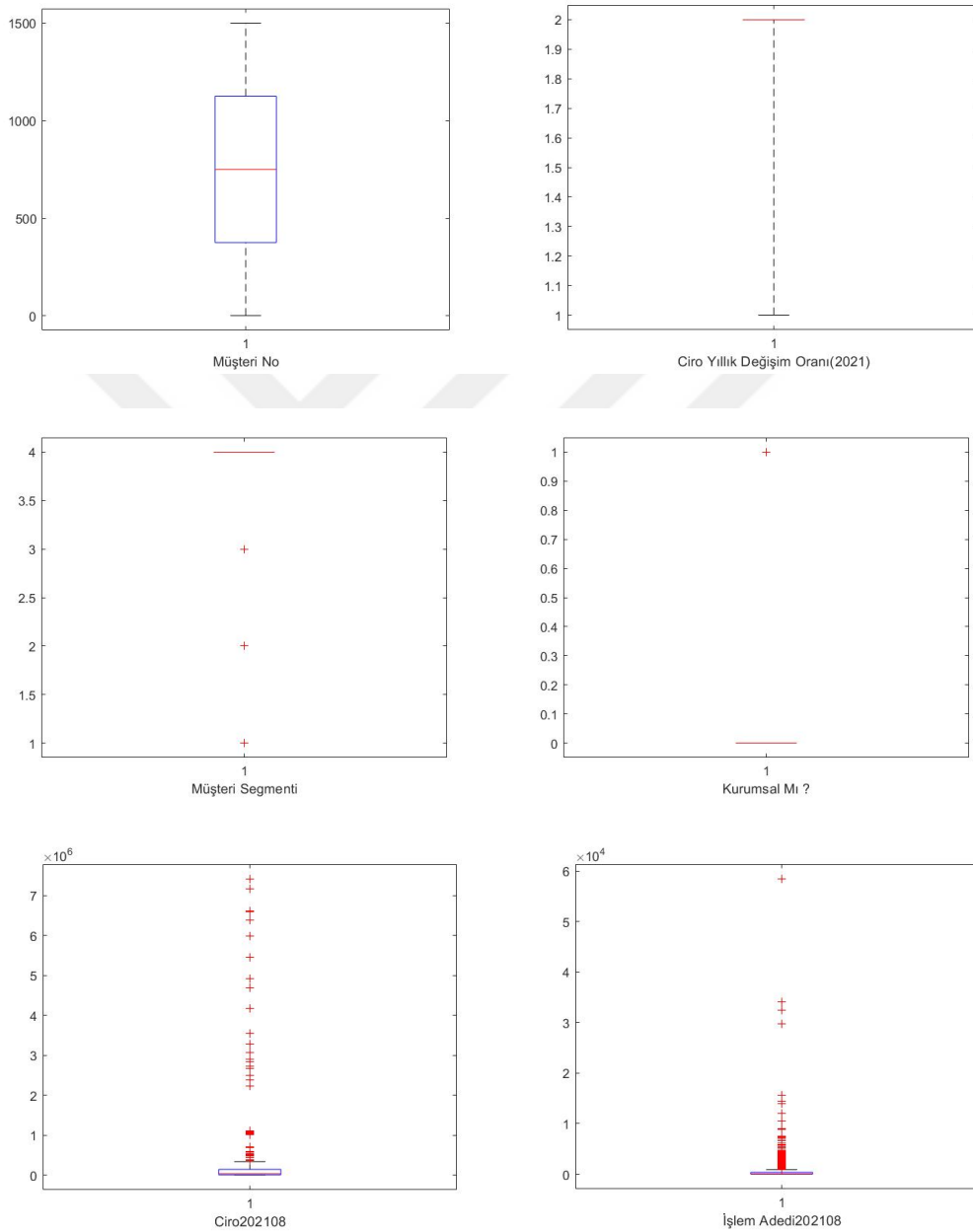
Şekil 3.2 Veri setinde değişkenler detayında örneklem sayısı

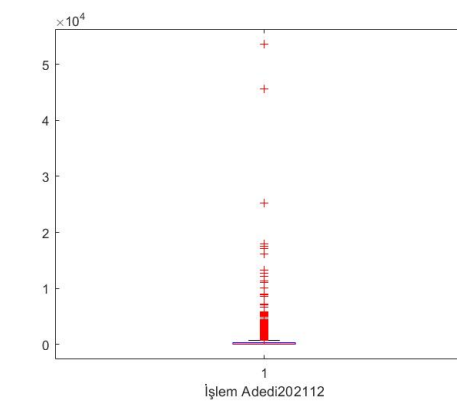
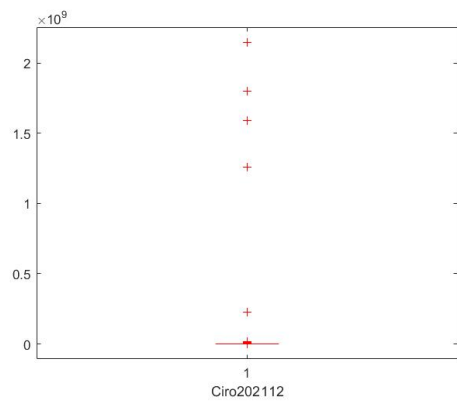
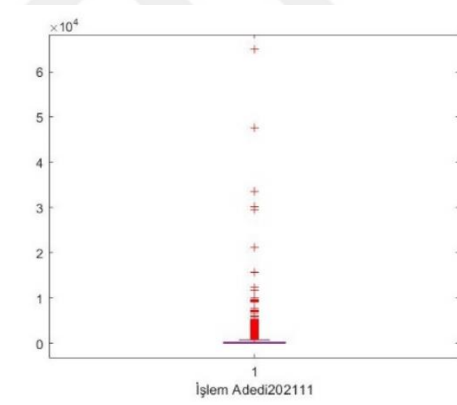
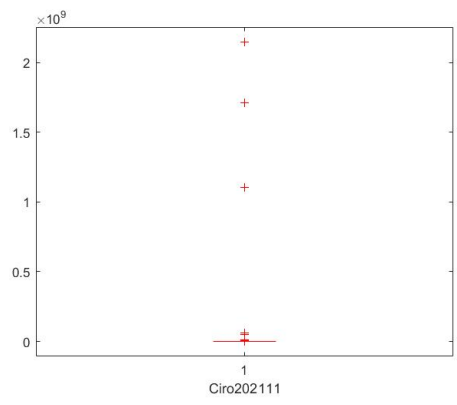
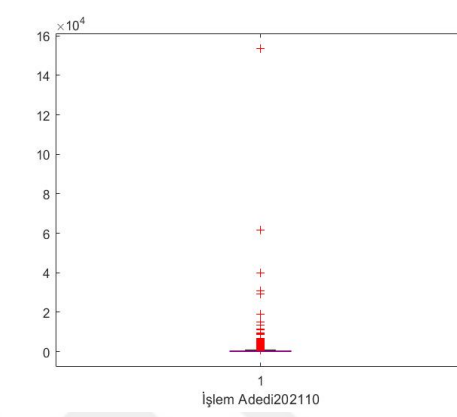
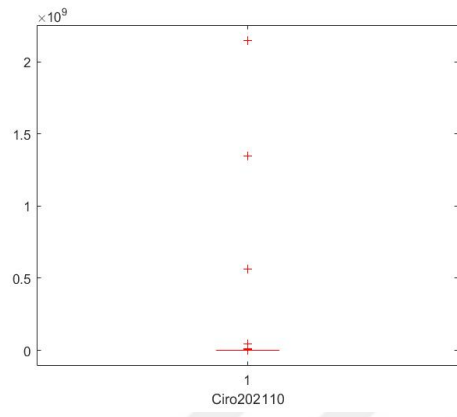
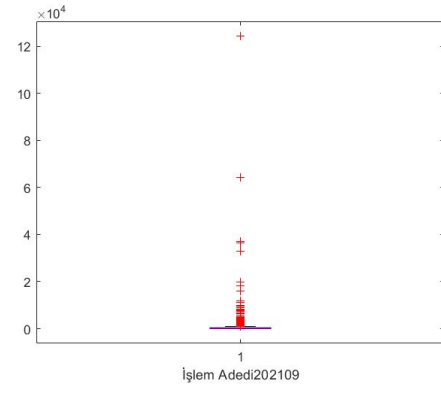
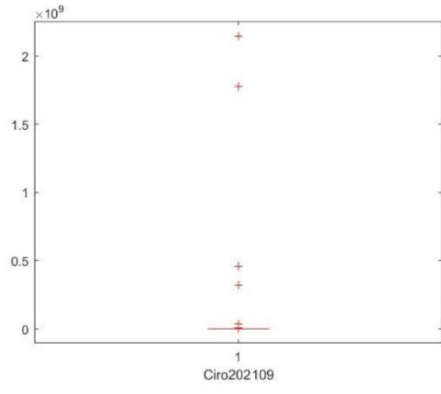
3.3.2.1 Aykırı Gözlem Analizi

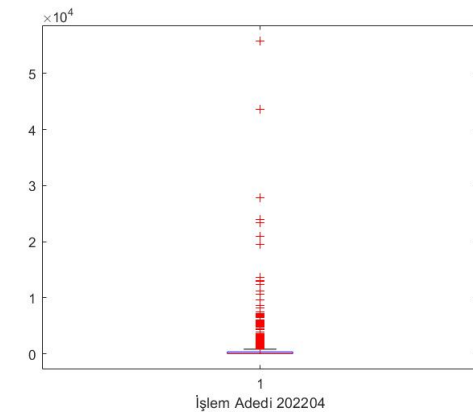
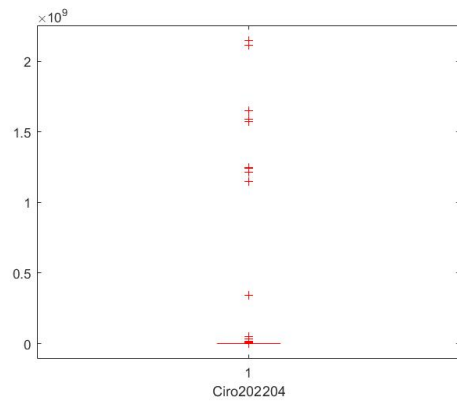
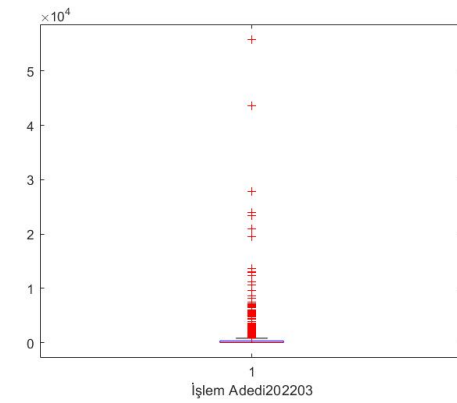
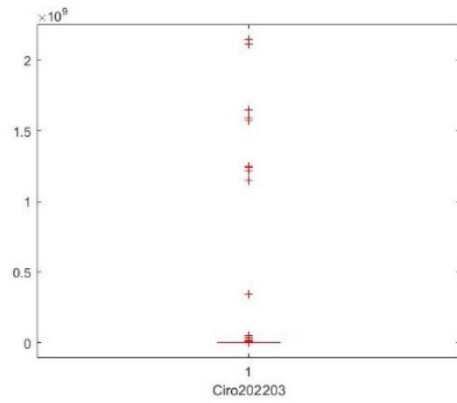
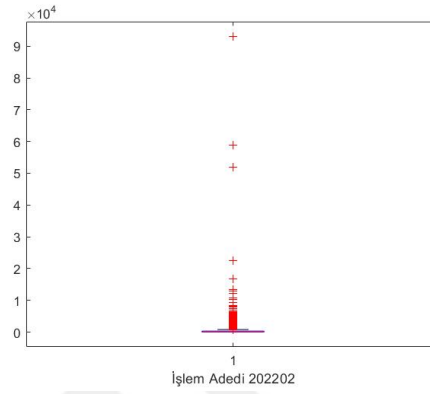
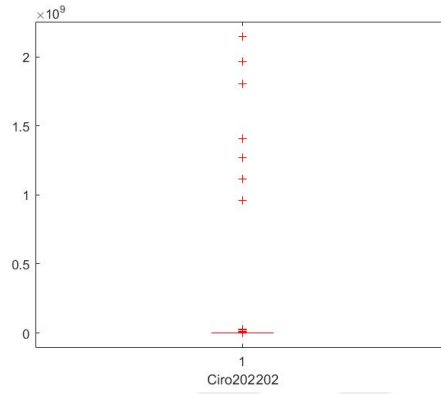
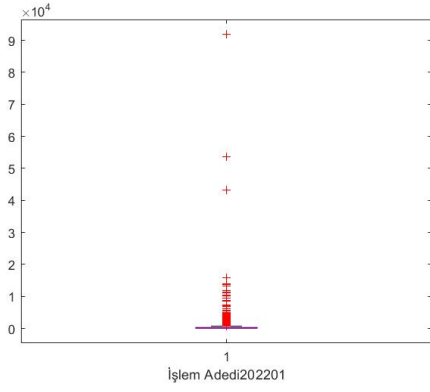
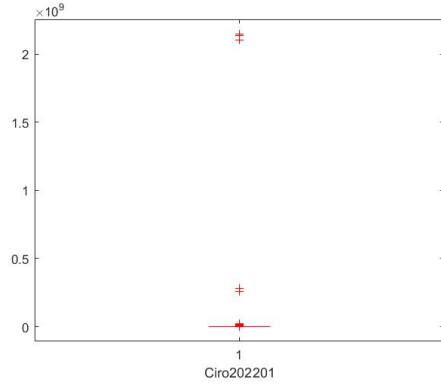
Veri madenciliği ve istatistik alanlarında veride anormalliklerin, uyumsuzlukların ve sapmaların olduğu veriye aykırı değer denilmektedir. Aykırı değerlerin oluşması ise modeli olumsuz yönde etkilemektedir. Bu nedenle aykırı değerlerin tespiti ve

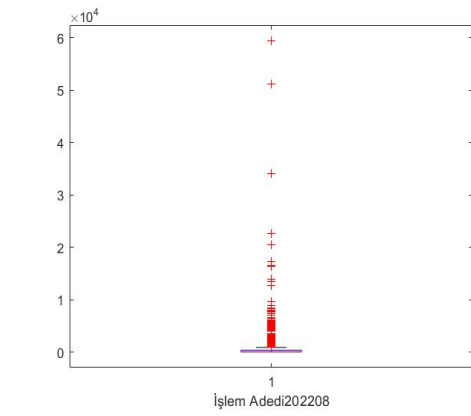
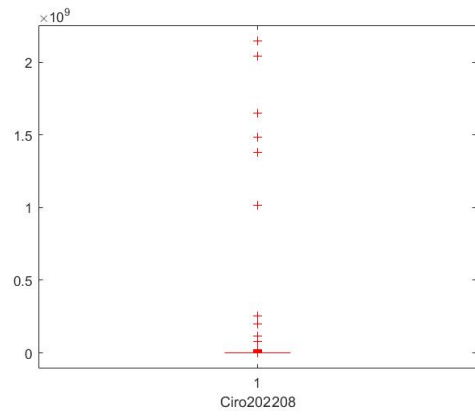
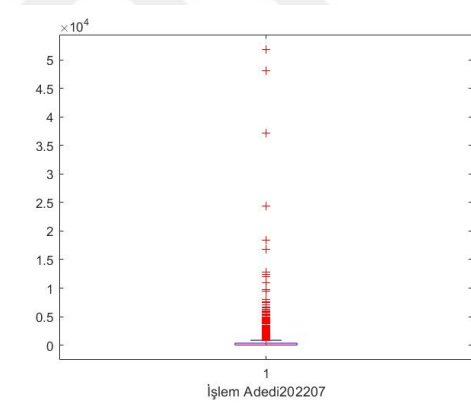
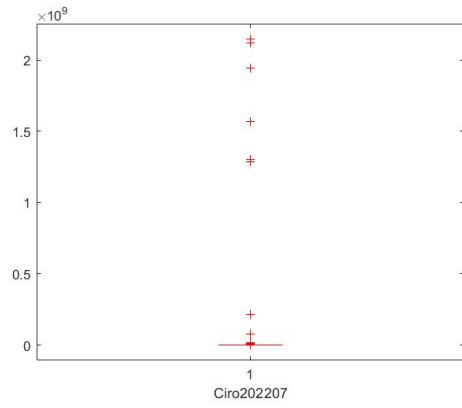
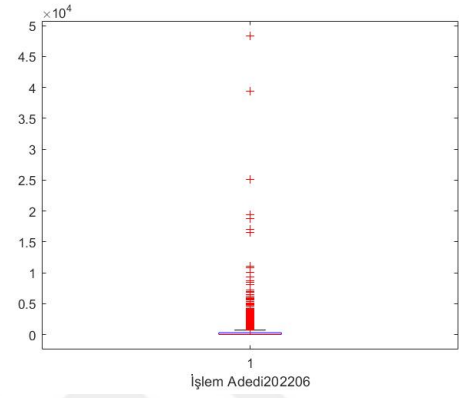
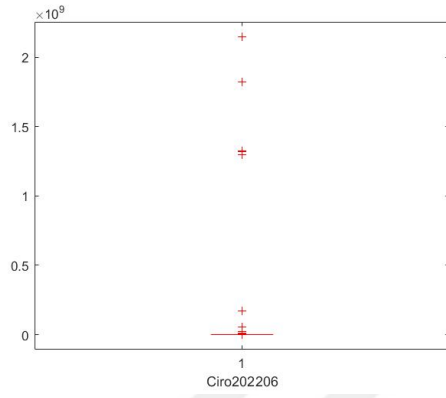
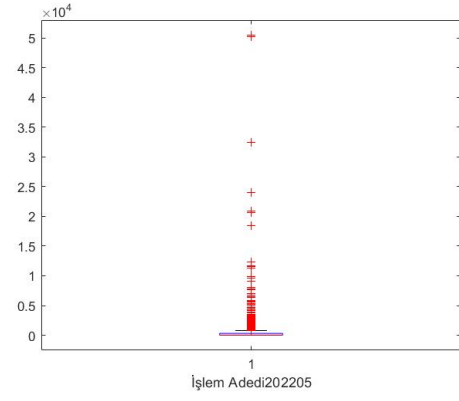
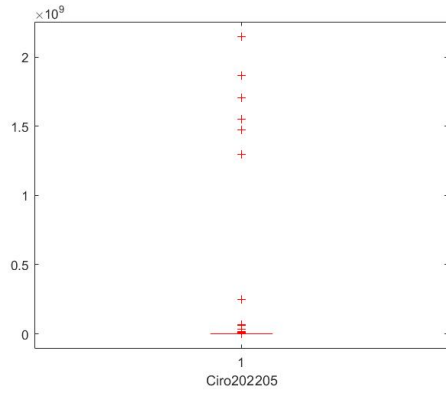
analizi büyük önem taşımaktadır. Bazı kullanım alanları şu şekildedir: saldırı tespit sistemleri, kredi kartı dolandırıcılığı, tıbbi teşhis [43].

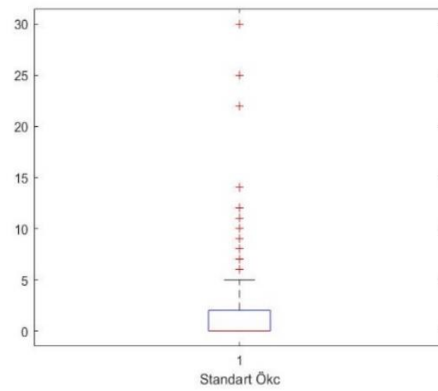
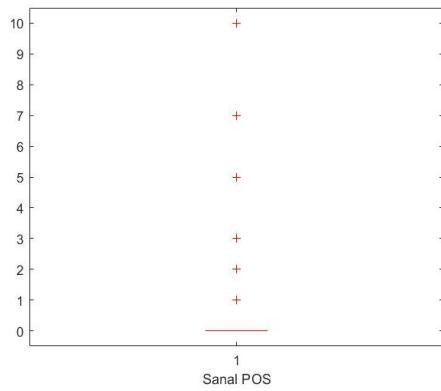
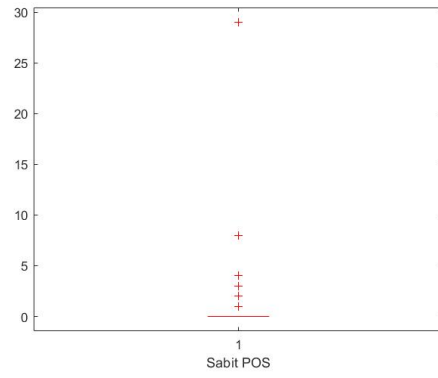
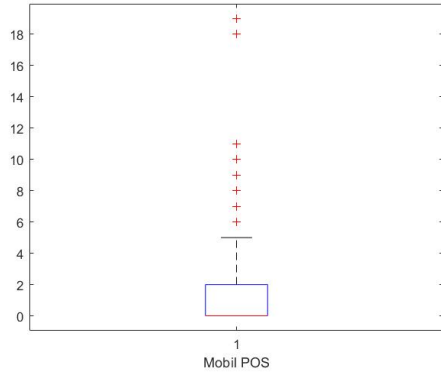
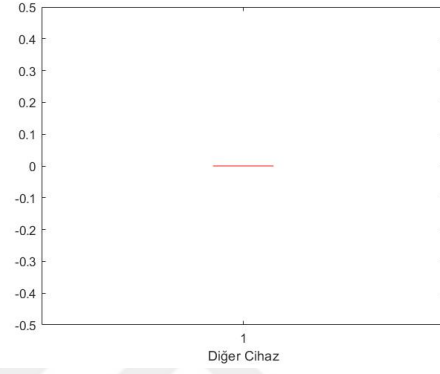
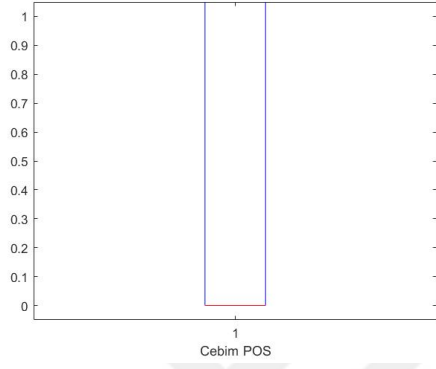
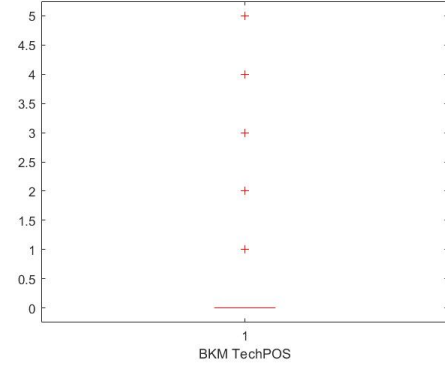
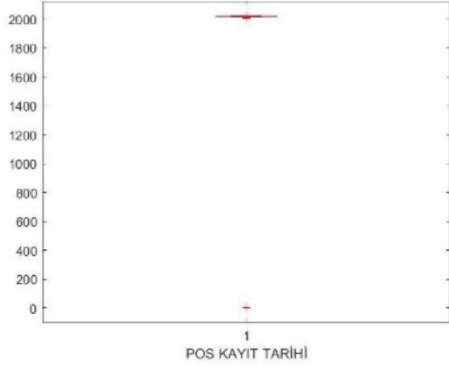
Şekil 3.3’de ise Aykırı Gözlem Analizinde en sık kullanılan yöntemlerden biri olan Boxplot Yöntemi ile görselleştirme çalışması verilmektedir. Girdi veri setinde bulunan 44 adet değişken boxplot aracılığıyla uzman görüşü alınarak incelenmiştir.

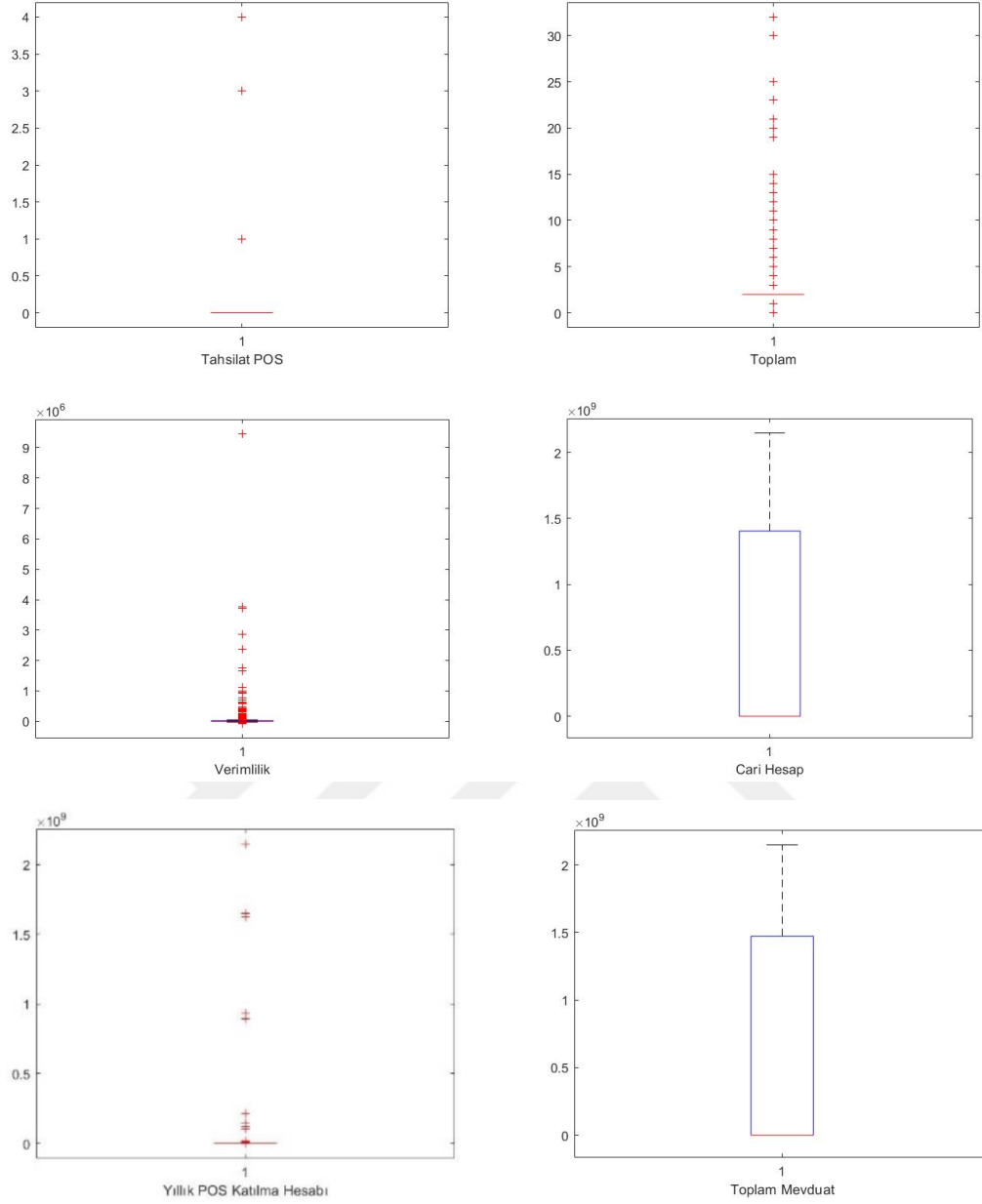












Şekil 3.3 Değişkenlerin boxplot grafikleri

3.3.3 Değişken Dönüşümü

Çalışmada bulunan girdi seti 23 farklı değişken ve toplamda 44 adet değişkenden oluşmaktadır. Bu değişkenlerin türleri Tablo 3.3'te gösterilmektedir. Bu değişkenlerden 43 tanesi nümerik iken 1 değişken kategorik değerdedir. “Kurumsal mı” değişkeninin cevabı ise “EVET” ve “HAYIR” şeklindedir. Bu kategorik olan değişkeni nümeriğe çevirmek amacıyla 0-1 dönüşümü yapılmaktadır. “EVET” cevabına karşılık 1 değeri kullanılırken “HAYIR” değerine karşılık 0 değeri ile

dönüşüm tamamlanmıştır. Yapılan dönüşüm işlemi sayesinde algoritmaların performansı artarak tahminleme çalışmasına katkı sağlamıştır.

1500 gözlem biriminden oluşan veri setinde sektör bilgisi yer almaktadır. Sektör bilgisini anlamlı hale getirmek için TÜİK'den alınan veriler kullanılarak "Ciro Yıllık Değişim Oranı" sütunu tanımlanmıştır [44].

3.3.1 Veri Standardizasyonu ve Normalizasyonu

Çalışma kapsamında kullanılan veri seti katılım bankasından alınmış olup Kişisel Verilerin Korunması Kanunu gereği manipüle edilerek kullanılmıştır. Kullanılan bu veriler 23 farklı değişkenden oluşmaktadır. Bu değişkenler farklı türde olup farklı ölçek değerlerine sahiptir. Veri setinin standart halde olması ise öğrenme sürecini olumlu yönde etkilemektedir. Çalışma kapsamında veri setinin standart hale getirmek içinse normalizasyon çalışması yapılmıştır. Bu sayede veri tekrarı, veri kaybı ve veri yetersizliğide ortadan kalkmaktadır.

Veri setinde bulunan tüm değerler 0 ile 1 arasındaki değerlere dönüştürülmektedir. 0 en küçük değeri, 1 ise en yüksek değeri ifade ederken arada kalan değerler 0-1 aralığına yayılmaktadır.

3.4 Veri Setinin Modellenmesi

İşin anlamlandırılması, verinin anlamlandırılması ve verinin hazırlanması adımlarından sonra gelen modelleme aşamasında katılım bankasının tüzel müşterilerinin POS cihazlarından geçirdikleri ciro tahmini için 10 Katlı Çapraz Doğrulama Yöntemi ile makine öğrenmesi algoritmalarından Lineer Regresyon (Tek değişkenli-Çok değişkenli), Ridge Regresyon, Lasso Regresyon, K-En Yakın Komşu (KNN), Ağaç Bazlı Regresyon, Rastgele Orman, XGBoost algoritmaları kullanılmıştır. Uygulamada birden çok algoritma kullanılmasının nedeni ise en yüksek doğruluk oranına ulaşmak ve en optimal algoritmanın seçiminin yapılmasıdır.

Yukarıda belirtilen algoritmalarda etiketli veriler kullanıldığından denetimli makine öğrenmesi algoritmalarının içerisinde yer verilmiştir. Çalışmada denetimli makine öğrenmesi kullanılmasının nedeni ise bilinen veri setiyle bilinen çıktı değerinin tahmin edilmesidir. Ayrıca çıktı verisi bağımlı değişkendir. Çalışmamızda bilinen çıktı değerimiz ise POS ciro değeridir.

Literatürde yapılan çalışmalara bakıldığında genellikle bir çok algoritmanın kullanıldığı ve en optimal sonucu veren algoritma seçiminin yapıldığı görülmektedir. Ciro tahmini ve gelir tahmini çalışmaları incelendiğinde ise denetimli makine öğrenmesi algoritmalarının sıklıkla tercih edildiği görülmektedir.

İlgili çalışmada yukarıda belirtilen algoritmalar 1500 gözlem biriminden ve 23 adet farklı değişkenden oluşan veri setine uygulanarak modelleme yapılmıştır. Çalışma kapsamında model doğrulama yöntemlerinden Train&Test Modeli ve K-Katlı Çapraz Doğrulama Yöntemi kullanılmıştır. İlgili modeller %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde ayrılarak model performansları ölçülmüştür. Modelin başarı oranını arttırmak için 10 Katlı Çapraz Doğrulama yapılarak doğruluk derecesi en yüksek sonuçlar tespit edilmiştir.

3.4.1 Lineer Regresyon (Tek Değişkenli/Çok Değişkenli)

3.4.1.1 Tek Değişkenli Lineer Regresyon

Lineer regresyon algoritması kullanılırken “stats” kütüphanesi çalıştırılmıştır. Veri değerleri analiz edildiğinde ciro değerini %72 oranında en çok etkileyen değişkenin “Verimlilik” olduğu tespit edilmiştir. Bu nedenle Tek Değişkenli Lineer Regresyon modelinde “Verimlilik” girdisi kullanılmıştır. Formülasyonu ise aşağıdaki gibidir:

$$\emptyset(Ciro202208) = \beta_0 + \beta_1(Verimlilik) \quad (3.1)$$

İlgili modelin parametreleri aşağıdaki gibidir:

Lm(Ciro202208~Verimlilik, copy_X=True, n_jobs=None, fit_intercept=True, Positive=False)

Parametreler belirlendikten sonra veri seti %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde ayrılmıştır. 10 Katlı Çapraz Doğrulama yöntemi kullanılarak optimum tahmin değeri elde edilmiştir.

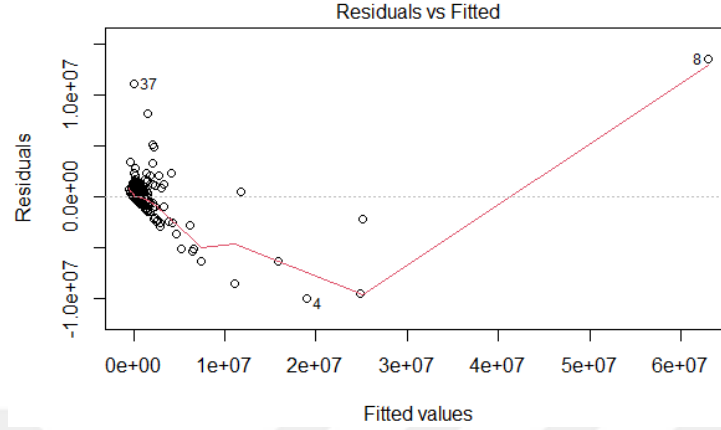
- %80 Eğitim-%20 Test Modeli

β_0 ve β_1 değerleriyle birlikte oluşan denklem aşağıdaki gibidir:

$$\emptyset(Ciro202208) = 24193.42 + 6.65(Verimlilik) \quad (3.2)$$

Şekil 3.4’de modelden elde edilen kalıntı (artık) değerlerini göstermektedir.

Residuals&Fitted grafiğinde bulunan residuals değeri tahmini değer ile regresyon çizgisi arasındaki farkların toplamını ifade etmektedir. Fitted değeri ise tahmin edilen uygun değerleri göstermektedir [45].



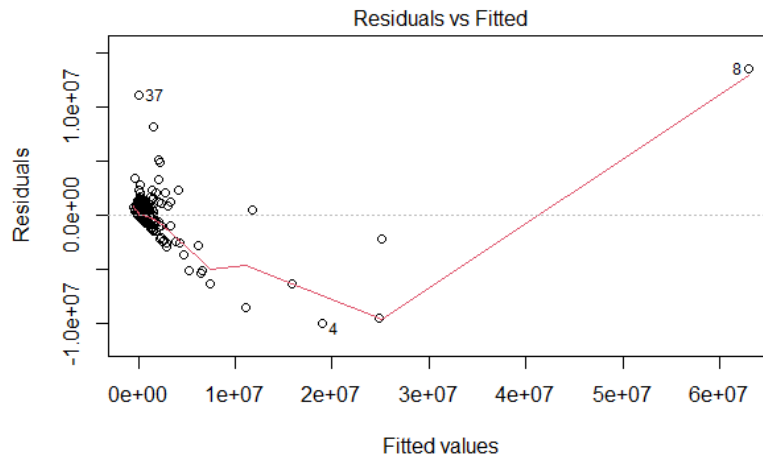
Şekil 3.4 Tek değişkenli regresyonda residuals ×fitted grafiği (%80 eğitim-%20 test)

- %70 Eğitim-%30 Test Modeli

β_0 ve β_1 değerleriyle birlikte oluşan denklem aşağıdaki gibidir:

$$\phi(Ciro202208) = 23940.91 + 6.58(Verimlilik) \quad (3.3)$$

Şekil 3.5’de modelden elde edilen kalıntı(artık) değerlerini göstermektedir. Residuals&Fitted grafiğinde bulunan residuals değeri tahmini değer ile regresyon çizgisi arasındaki farkların toplamını ifade etmektedir. Fitted değeri ise tahmin edilen uygun değerleri göstermektedir [45].



Şekil 3.5 Tek değişkenli regresyonda residuals ×fitted grafiği (%70 eğitim-%30 test)

3.4.1.2 Çok Değişkenli Lineer Regresyon

Lineer regresyon algoritması kullanılırken “stats” kütüphanesi çalıştırılmıştır. Veri değerleri analiz edildiğinde ciro değerini en çok etkileyen değişkenler tespit edilmiştir. Bu değişkenler şöyledir: Verimlilik, *Ciro202207*, *Ciro202206*, *Ciro202205*, Yıllık POS Katılma Hesabı, Toplam Mevduat, *Ciro Yıllık Değişim Oranı*’dır. Çok Değişkenli Lineer Regresyon algoritmasının formülasyonu aşağıdaki gibidir:

$$\emptyset(x_1, x_2, \dots, x_k) = \beta_0 + \beta_1(\text{Verimlilik}) + \beta_2(\text{Ciro202207}) + \beta_3(\text{Ciro202206}) + \beta_4(\text{Ciro202205}) + \beta_5(\text{Yıllık Pos Katılma Hesabı}) + \beta_6(\text{Toplam Mevduat}) + \beta_7(\text{Ciro Yıllık Değişim Oranı}) \quad (3.4)$$

İlgili modelin parametreleri aşağıdaki gibidir:

`Lm(Ciro202208~Verimlilik+Ciro202207+Ciro202206+Ciro202205+Yıllık POS Katılma Hesabı+Toplam Mevduat+Ciro Yıllık Değişim Oranı, fit_intercept=True, copy_X=True, n_jobs=None, Positive=False)`

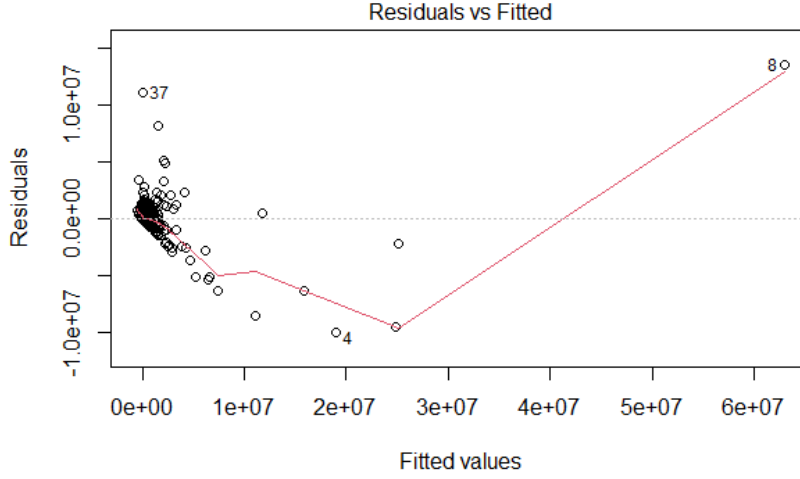
Parametreler belirlendikten sonra veri seti %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde 2 farklı modele ayrılmıştır. 10 Katlı Çapraz Doğrulama yöntemi kullanılarak optimum tahmin değeri elde edilmiştir.

- %80 Eğitim-%20 Test Modeli

$\beta_0, \beta_1, \dots, \beta_7$ değerleriyle birlikte oluşan denklem aşağıdaki gibidir:

$$\emptyset(\text{Ciro202208}) = 24178.11 + 6.57(\text{Verimlilik}) + 3.18(\text{Ciro202207}) + 2.6(\text{Ciro202206}) + 1.2(\text{Ciro202205}) + 0.56(\text{Yıllık Pos Katılma Hesabı}) + 0.41(\text{Toplam Mevduat}) + 0.01(\text{Ciro Yıllık Değişim Oranı}) \quad (3.5)$$

Şekil 3.6’da modelden elde edilen kalıntı (artık) değerlerini göstermektedir. Residuals&Fitted grafiğinde bulunan residuals değeri tahmini değer ile regresyon çizgisi arasındaki farkların toplamını ifade etmektedir. Fitted değeri ise tahmin edilen uygun değerleri göstermektedir [45].



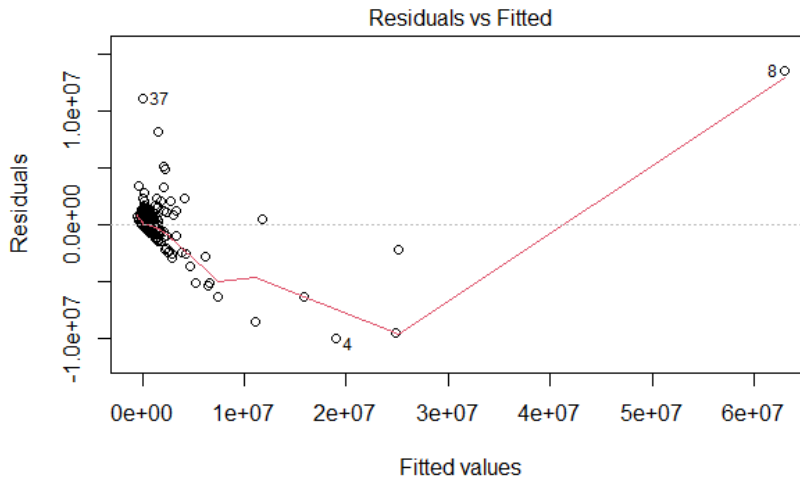
Şekil 3.6 Çok değişkenli regresyonda residuals ×fitted grafiği (%70 eğitim-%30 test)

- %70 Eğitim-%30 Test Modeli

$\beta_0, \beta_1, \dots, \beta_7$ değerleriyle birlikte oluşan denklem aşağıdaki gibidir:

$$\begin{aligned} \emptyset(Ciro202208) = & 23982.34 + 6.38(Verimlilik) + 3.05(Ciro202207) + \\ & 2.49(Ciro202206) + 1.02(Ciro202205) + 0.49(Yıllık Pos Katılma Hesabı) + \\ & 0.37(Toplam Mevduat) + 0.009(Ciro Yıllık Değişim Oranı) \end{aligned} \quad (3.6)$$

Şekil 3.7’de modelden elde edilen kalıntı (artık) değerlerini göstermektedir. Residuals&Fitted grafiğinde bulunan residuals değeri tahmini değer ile regresyon çizgisi arasındaki farkların toplamını ifade etmektedir. Fitted değeri ise tahmin edilen uygun değerleri göstermektedir [45].

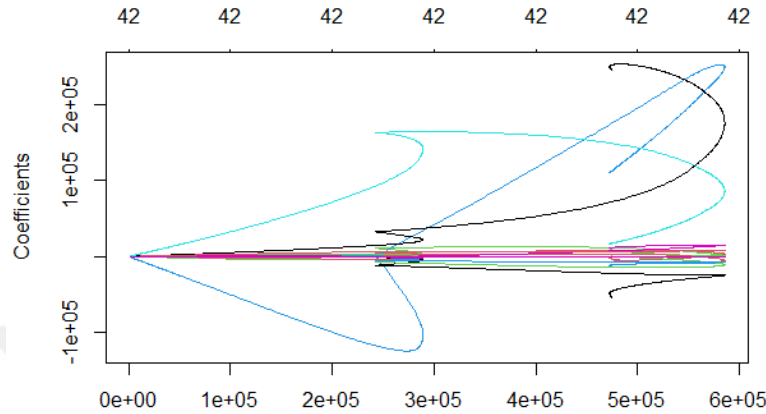


Şekil 3.7 Çok değişkenli regresyonda residuals ×fitted grafiği (%70 eğitim-%30 test)

3.4.2 Ridge Regresyon

Lineer regresyon modelinin L2 normunun minimizasyonuna ve en küçük kareler hatasına dayanan yöntemdir [35]. Aşağıda ise formülasyonu gözükmektedir.

$$\text{Ridge Regresyon} = (\text{lineer model}) \times (\text{en küçük kareler hatası} + \text{min L2 norm}) \quad (3.7)$$



Şekil 3.8 Katsayılar x L2 norm grafiği

Şekil 3.8’de bulunan grafik L2 norm değerlerinin katsayılara göre değişimini göstermektedir. İlgili modelin parametreleri aşağıdaki gibidir:

R Studio yazılım dilinde Ridge regresyon modelini uygulayabilmek için öncelikle” glmnet” paketi indirilmektedir. İlgili modelin parametreleri aşağıdaki gibidir:

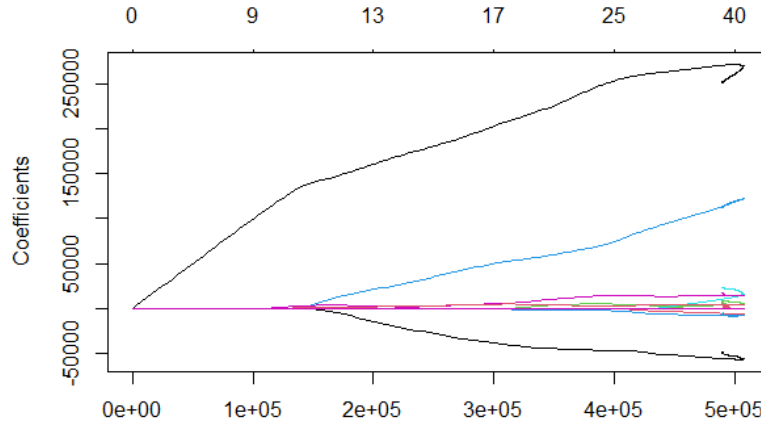
```
Glmnet(alpha=0,fit_intercept=True, copy_X=True,lambda=grid, solver='auto',max_
iter=None, tol=0.001 , random_state=None ,Positive=False)
```

Parametreler belirlendikten sonra veri seti %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde 2 farklı modele ayrılmıştır. Kurulan modeller 10 Katlı Çapraz Doğrulama yapılarak r^2 (korelasyon katsayısı), MSE (Mean Squared Error-Ortalama Kare Hatası), RMSE (Root Mean Squared Error- Ortalama Karekök Sapması) ve MAE (Mean Absolute Error- Ortalama Mutlak Değer) metriklerine göre en iyi performans gösteren model seçilir. Modelin performans değerleri Tablo 3.4 ve Tablo 3.5 ‘de gösterilmektedir.

3.4.3 Lasso Regresyon

Lineer regresyon modelinin L1 normunun minimizasyonuna en küçük kareler hatasına dayanan yöntemdir. Aşağıda ise formülasyonu gözükmektedir.

$$\text{Ridge regresyon} = (\text{lineer model}) \times (\text{en küçük kareler hatası} + \text{min L2 norm}) \times (\text{gradient descent}) \quad (3.8)$$



Şekil 3.9 Katsayılar x L1 norm grafiği

Şekil 3.9 'da bulunan grafik L1 norm değerlerinin katsayılar göre değişimini göstermektedir. R Studio yazılım dilinde Ridge regresyon modelini uygulayabilmek için öncelikle "glmnet" paketi indirilmektedir. İlgili modelin parametreleri aşağıdaki gibidir:

```
Glmnet(alpha=1,fit_intercept=True, max_iter=None,solver='auto'copy_X=True, tol=0.001, random_state=None ,Positive=False )
```

Parametreler belirlendikten sonra veri seti %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde 2 farklı modele ayrılmıştır. Kurulan modeller 10 Katlı Çapraz Doğrulama yapılarak r^2 (korelasyon katsayısı), MSE (Mean Squared Error-Ortalama Kare Hatası), RMSE (Root Mean Squared Error-Ortalama Karekök Sapması) ve MAE (Mean Absolute Error-Ortalama Mutlak Değer) metriklerine göre en iyi performans gösteren model seçilir. Modelin performans değerleri Tablo 3.4 ve Tablo 3.5 'de gösterilmektedir.

3.4.4 K-En Yakın Komşu

KNN-K En Yakın Komşu algoritması olarak bilinen bu yöntem hem sınıflandırma hem de regresyon çalışmalarında kullanılmaktadır. Herhangi bir özellik seçimi yapıldıktan sonra bu özelliğe en yakın olan k adet farklı özellik seçimi yapılarak aralarında bulunan mesafe öklid, kosinüs ve Manhattan hesaplama yöntemleri kullanılarak ölçülmektedir. Aşağıda ise öklid hesaplaması verilmektedir [44].

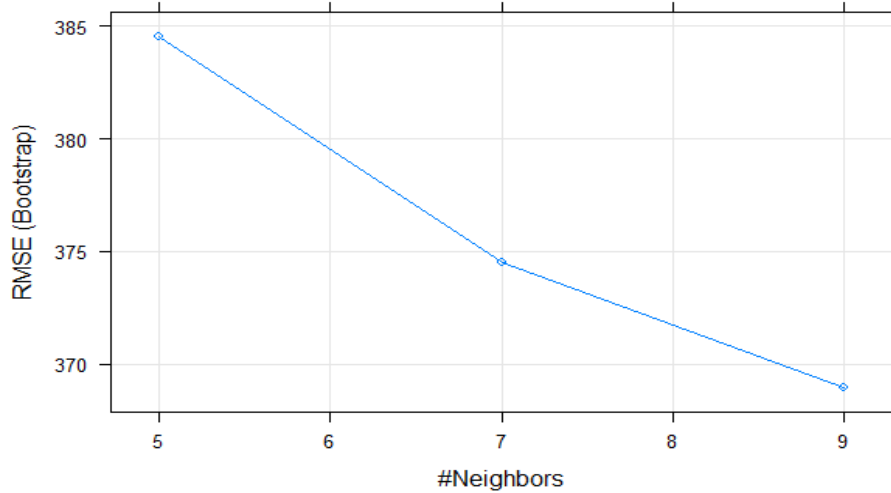
$$d_{(i,j)} = \sqrt{\sum\{(x_{\{ik\}} - x_{\{jk\}})^2\}} \quad (3.9)$$

R Studio yazılım dilinde KNN modelini uygulayabilmek için öncelikle "caret" paketi indirilmektedir. Bu paket sınıflandırma ve regresyon işlevlerini içermektedir. Parametreler belirlendikten sonra veri seti %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde 2 farklı modele ayrılmıştır.

- %80 Eğitim-%20 Test Modeli

İlgili modelin parametreleri aşağıdaki gibidir:

```
knn(n_neighbors=9,algorithm='auto',leaf_size=20,p=2,n_jobs=None,metric_params=None)
```



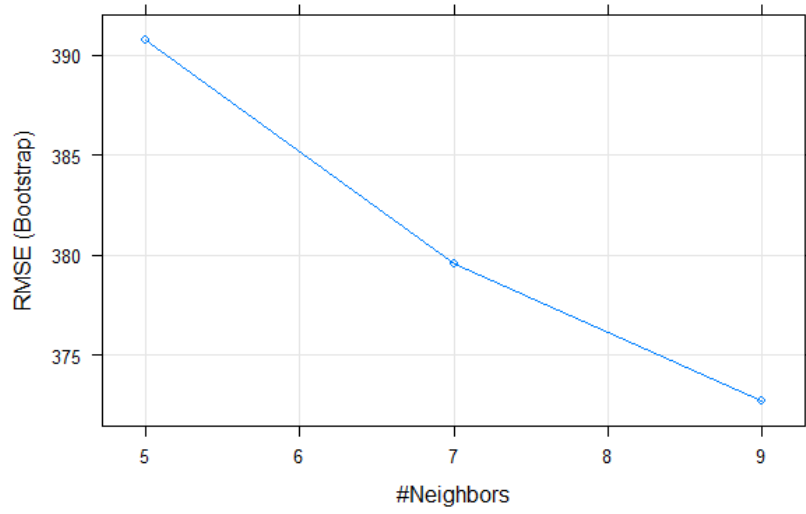
Şekil 3.10 Komşuluk sayısı×RMSE değeri (%80 eğitim-%20 test)

Şekil 3.10 'da komşu sayısına karşılık gelen RMSE hataları grafik halinde gösterilmektedir. Model incelendiğinde komşu sayısı arttığında RMSE hata değerinin azaldığı görülmektedir.

- %70 Eğitim-%30 Test Modeli

İlgili modelin parametreleri aşağıdaki gibidir:

```
knn(n_neighbors=9,algorithm='auto',leaf_size=20,p=2,n_jobs=None,metric_params=None)
```



Şekil 3.11 Komşuluk sayısı×RMSE değeri (%70 eğitim-%30 test)

Şekil 3.11 'de komşu sayısına karşılık gelen RMSE hataları grafik halinde gösterilmektedir. Model incelendiğinde komşu sayısı arttığında RMSE hata değerinin azaldığı görülmektedir.

Şekil 3.10 ile Şekil 3.11 kıyaslandığında %80 eğitim-%20 test ayrımıyla oluşturulan model daha iyi performans göstermektedir. Kurulan modeller 10 katlı çapraz doğrulama yapılarak r^2 (korelasyon katsayısı), MSE (Mean Squared Error-Ortalama Kare Hatası), RMSE (Root Mean Squared Error-Ortalama Karekök Sapması) ve MAE (Mean Absolute Error-Ortalama Mutlak Değer) metriklerine göre en iyi performans gösteren model seçilir. Modelin performans değerleri Tablo 3.4 ve Tablo 3.5 'de gösterilmektedir.

3.4.5 Karar Ağaçları

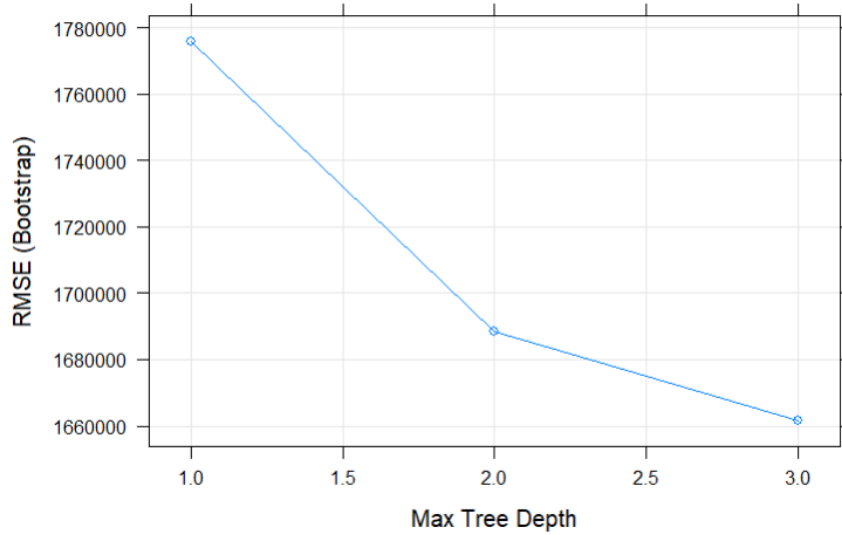
Hem sınıflandırma hem de regresyon çalışmalarında sıklıkla kullanılan Karar Ağacı algoritması başta bulunan veri setini alt kümelerle ayırmaktadır. Ayrılan kümeler sayesinde ise ağaç yapısı oluşmaktadır [26].

R Studio yazılım dilinde KNN modelini uygulayabilmek için öncelikle "rpart" paketi indirilmektedir. Bu paket sınıflandırma ve regresyon işlevlerini içermektedir. Parametreler belirlendikten sonra veri seti %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde 2 farklı modele ayrılmıştır.

- %80 Eğitim-%20 Test Modeli

İlgili modelin parametreleri aşağıdaki gibidir:

```
rpart(Ciro202208~.,depth=3,min_samples_split=2,max_depth=None,min_samples_
leaf=1, min_weight_fraction_leaf=0.0, class_weight=None,max_features=None, max
_leaf_nodes=None, min_impurity_decrease=0.0 )
```



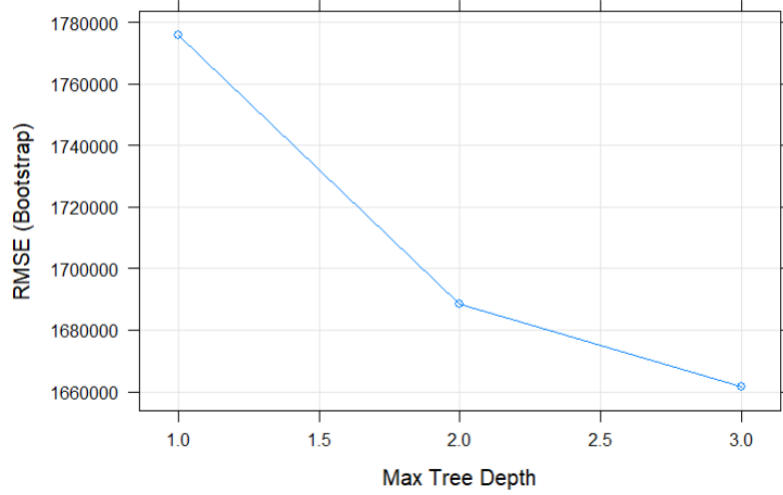
Şekil 3.12 Ağaç derinliği×RMSE değeri (%80 eğitim-%20 test)

Şekil 3.12 'de ağaç derinliğine karşılık gelen RMSE hataları grafik halinde gösterilmektedir. Model incelendiğinde ağaç derinliği sayısı arttığında RMSE hata değerinin azaldığı görülmektedir.

- %70 Eğitim-%30 Test Modeli

İlgili modelin parametreleri aşağıdaki gibidir:

```
rpart(Ciro202208~.,depth=3,min_samples_split=2,max_depth=None,min_samples_
leaf=1, min_weight_fraction_leaf=0.0, class_weight=None,max_features=None, max
_leaf_nodes=None, min_impurity_decrease=0.0 )
```



Şekil 3.13 Ağaç derinliği×RMSE değeri (%70 eğitim-%30 test)

Şekil 3.13 'de ağaç derinliğine karşılık gelen RMSE hataları grafik halinde gösterilmektedir. Model incelendiğinde ağaç derinliği sayısı arttığında RMSE hata değerinin azaldığı görülmektedir.

Kurulan modeller 10 katlı çapraz doğrulama yapılarak r^2 (korelasyon katsayısı), MSE (Mean Squared Error-Ortalama Kare Hatası), RMSE (Root Mean Squared Error-Ortalama Karekök Sapması) ve MAE (Mean Absolute Error-Ortalama Mutlak Değer) metriklerine göre en iyi performans gösteren model seçilir. Modelin performans değerleri Tablo 3.4 ve Tablo 3.5 'de gösterilmektedir.

3.4.6 Rastgele Ormanlar

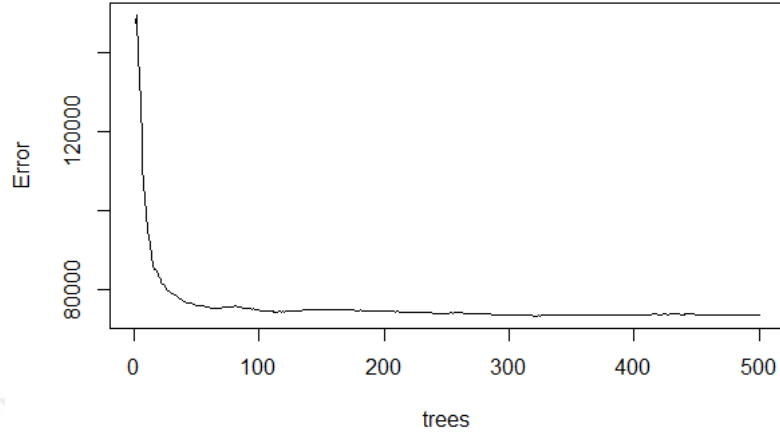
Temel modeller açısından incelendiğinde Rastgele Ormanlar algoritması Karar Ağaçlarının en sık kullanıldığı torbalama yöntemidir. Bu yöntemin başarılı olabilmesi için girdi veri setinin çeşitliliği yüksek önem taşımaktadır. Modelin performansını yükseltmek için şu yöntemler kullanılabilir: satır eklemesi, sütun örnekleme, yetersiz bölme [26].

R Studio yazılım dilinde KNN modelini uygulayabilmek için öncelikle "randomForest" paketi indirilmektedir. Bu paket sınıflandırma ve regresyon işlevlerini içermektedir. Parametreler belirlendikten sonra veri seti %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde 2 farklı modele ayrılmıştır.

- %80 Eğitim-%20 Test Modeli

İlgili modelin parametreleri aşağıdaki gibidir:

RandomForest(Ciro202208~,n_estimators=500,max_depth=None,min_samples_split=2,min_samples_leaf=1,max_features=1.0,min_weight_fraction_leaf=0.0,max_leaf_nodes=None)



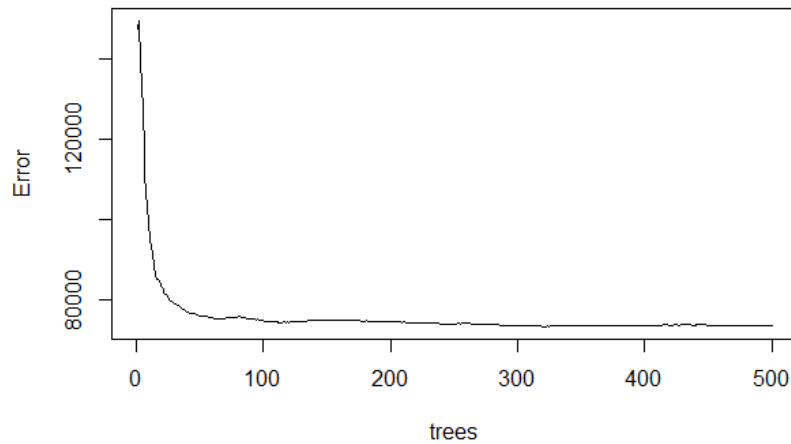
Şekil 3.14 Ağaç sayısı×hata değeri (%80 eğitim-%20 test)

Şekil 3.14’de ağaç sayısına karşılık gelen RMSE hata değeri grafik halinde gösterilmektedir. Model incelendiğinde ağaç sayısı arttığında RMSE hata değerinin azaldığı görülmektedir.

- %70 Eğitim-%30 Test Modeli

İlgili modelin parametreleri aşağıdaki gibidir:

RandomForest(Ciro202208~,n_estimators=500,max_depth=None,min_samples_split=2,min_samples_leaf=1,max_features=1.0,min_weight_fraction_leaf=0.0,max_leaf_nodes=None)



Şekil 3.15 Ağaç sayısı×hata değeri (%80 eğitim-%20 test)

Şekil 3.15’de ağaç sayısına karşılık gelen RMSE hata değeri grafik halinde gösterilmektedir. Model incelendiğinde ağaç sayısı arttığında RMSE hata değerinin azaldığı görülmektedir.

Kurulan modeller 10 katlı çapraz doğrulama yapılarak r^2 (korelasyon katsayısı), MSE (Mean Squared Error-Ortalama Kare Hatası), RMSE (Root Mean Squared Error-Ortalama Karekök Sapması) ve MAE (Mean Absolute Error-Ortalama Mutlak Değer) metriklerine göre en iyi performans gösteren model seçilir. Modelin performans değerleri Tablo 3.4 ve Tablo 3.5 ‘de gösterilmektedir.

3.4.7 XG- Boost

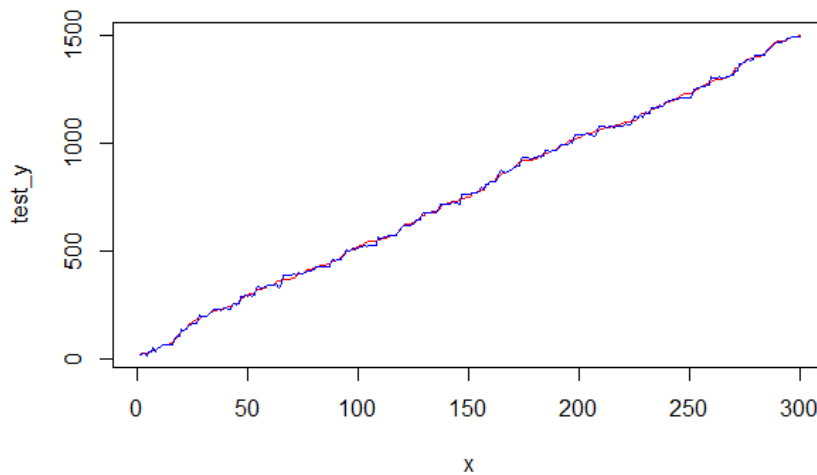
Hem sınıflandırma hem de regresyon çalışmalarında kullanılan XGBoost algoritması Gradient Boosting algoritmasının yüksek performanslı halidir. Literatürdeki çalışmalar incelendiğinde ise sıklıkla kullanıldığı görülmektedir [26].

R Studio yazılım dilinde KNN modelini uygulayabilmek için öncelikle "xgboost ve caret" paketi indirilmektedir. Bu paket sınıflandırma ve regresyon işlevlerini içermektedir. Parametreler belirlendikten sonra veri seti %80 Eğitim-%20 Test ve %70 Eğitim-%30 Test şeklinde 2 farklı modele ayrılmıştır.

- %80 Eğitim-%20 Test Modeli

İlgili modelin parametreleri aşağıdaki gibidir:

```
xgboost(data = xgb_train, max.depth = 3, nrounds = 86, verbose = 0)
```



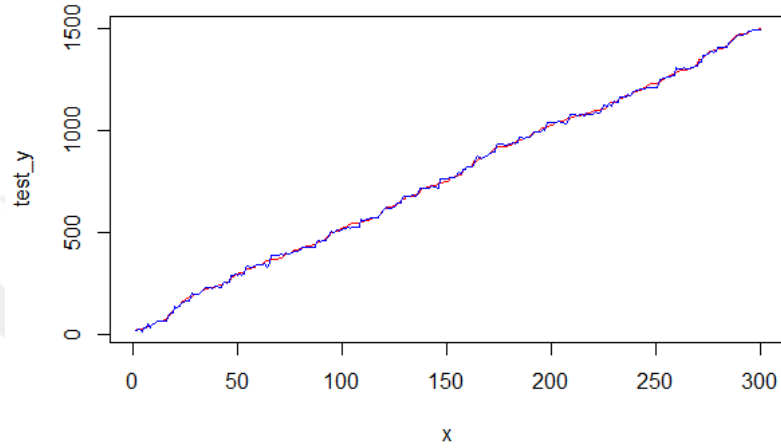
Şekil 3.16 $x \times test_y$ değeri (%80 eğitim-%20 test)

Şekil 3.16'da test veri setine karşılık gelen POS ciro değerinin, tahmini POS ciro değeri (test_y) arasındaki ilişki gösterilmektedir. Model incelendiğinde ağaç sayısı arttığında RMSE hata değerinin azaldığı görülmektedir.

- %70 Eğitim-%30 Test Modeli

İlgili modelin parametreleri aşağıdaki gibidir:

```
xgboost(data = xgb_train, max.depth = 3, nrounds = 86, verbose = 0)
```



Şekil 3.17 $x \times \text{test_y}$ değeri (%80 eğitim-%20 test)

Şekil 3.17'de test veri setine karşılık gelen POS ciro değerinin, tahmini POS ciro değeri (test_y) arasındaki ilişki gösterilmektedir. Model incelendiğinde ağaç sayısı arttığında RMSE hata değerinin azaldığı görülmektedir.

Kurulan modeller 10 katlı çapraz doğrulama yapılarak r^2 (korelasyon katsayısı), MSE (Mean Squared Error-Ortalama Kare Hatası), RMSE (Root Mean Squared Error-Ortalama Karekök Sapması) ve MAE (Mean Absolute Error-Ortalama Mutlak Değer) metriklerine göre en iyi performans gösteren model seçilir. Modelin performans değerleri Tablo 3.4 ve Tablo 3.5 'de gösterilmektedir.

3.5 Modellerin Sonuç Değerleri

Çalışmada Denetimli Makine Öğrenmesi algoritmalarından Lineer Regresyon (Tek değişkenli-Çok değişkenli), Ridge Regresyon, Lasso Regresyon, K-En Yakın Komşu (KNN), Ağaç Bazlı Regresyon, Rastgele Orman, XGboost algoritmaları kullanılmıştır. Veri setleri %80 eğitim-%20 test ve %70 eğitim-%30 test şeklinde ayrılmıştır. Model performanslarını değerlendirmek için r^2 , MSE, RMSE ve MAE değerleri hesaplanmıştır. Optimal model seçiminde r^2 değerinin 1'e yakın olması ve hatayı

ölçen MSE, RMSE, MAE değerlerinin 0'a yakın olması beklenmektedir. Hesaplanma sonucunda elde edilen değerler Tablo 3.4 ve Tablo 3.5 'de gösterilmektedir.

Tablo 3.4 %80 eğitim-%20 test verisiyle elde edilen sonuç değerleri

Model: %80 Eğitim-%20 Test	r^2	MSE	RMSE	MAE
Tek Değişkenli Lineer Regresyon	0.7063	164,1523	221,0959	150,9387
Çok Değişkenli Lineer Regresyon	0.7200	168,9067	216,9582	148,1094
Ridge Regresyon	0.5876	201,9830	252,4995	185,9384
Lasso Regresyon	0.5798	198,0192	251,1547	183,0498
K-En Yakın Komşu	0.1296	312,4721	370,9230	296,7835
Karar Ağaçları	0.4606	210,9394	282,2463	200,5458
Rastgele Ormanlar	0.5284	244,0124	269,8985	196,4489
XGBoost	0.9995	91,0345	9,5265	90,7553

Tablo 3.4 'de elde edilen sonuçlar incelendiğinde r^2 değerinde 1'e en yakın olan algoritmanın XGBoost algoritması olduğu görülmektedir. MSE, RMSE, MAE değerlerinde 0'a en yakın olan algoritmanın ise yine XGBoost algoritması olduğu görülmektedir.

Tablo 3.5 %70 eğitim-%30 test verisiyle elde edilen sonuç değerleri

Model: %70 Eğitim-%30 Test	r^2	MSE	RMSE	MAE
Tek Değişkenli Lineer Regresyon	0.7102	161,9365	219,7465	148,8727
Çok Değişkenli Lineer Regresyon	0.7170	164,2435	216,8601	146,9750
Ridge Regresyon	0.5910	198,2454	250,0209	184,0314
Lasso Regresyon	0.5889	191,3438	247,0647	179,9409
K-En Yakın Komşu	0.1229	320,9358	369,5394	299,7835
Karar Ağaçları	0.4604	215,1024	283,2868	202,8906
Rastgele Ormanlar	0.5200	250,5893	270,8858	197,8938
XGBoost	0.9995	94,5546	9,5488	91,1804

Tablo 3.5 'de elde edilen sonuçlar incelendiğinde r^2 değerinde 1'e en yakın olan algoritmanın XGBoost algoritması olduğu görülmektedir. MSE, RMSE, MAE değerlerinde 0'a en yakın olan algoritmanın ise yine XGBoost algoritması olduğu görülmektedir. Tablo 3.4 ve Tablo 3.5 'den elde edilen sonuçlar karşılaştırıldığında %80 eğitim- %20 test ayrımı olan XGBoost algoritması en başarılı performansı vermektedir.

Ödeme sistemlerinin temel taşlarından biri olan POS ürününden geçen ciro değeri bankalar arasında yüksek rekabete neden olmaktadır. Bu nedenle mevcutta bankaların hem bireysel hem de tüzel segmentte olan müşterilerinin POS cirolarını tahmin ederek potansiyel müşterinin tespiti oldukça önem kazanmaktadır. Bu kapsamda çalışmanın yapıldığı katılım bankasında portföy sorumlularına 6 aylık ve yıllık POS ciro hedefi verilmektedir. En üstün ciro tahminlemesini yapan algoritmanın tespitiyle birlikte portföy sorumlularının hedeflediği ciroya ulaşabilmeleri, ciro arttırımı ve pazar payının artışı sağlanacaktır.

Yapılan çalışma ise sektörde aktif olan katılım bankasının 2021 yılı ağustos ayı ile 2022 yılı temmuz ayı tarihleri arasında gerçekleşen verileri kullanılarak tüzel POS müşterilerinin ciro tahminlemesi için en uygun modelin bulunması üzerine yürütülmüştür. Bu kapsamda Şekil 3.1’de belirtilen işin anlamlandırılması, verinin anlamlandırılması, verinin hazırlanması, modelleme, değerlendirme ve yayılım aşamaları tamamlanmıştır.

İşin anlamlandırılması kısmında çalışmanın amacı ve hipotez kurulmuştur. Veri seti, 16 farklı bölge ve 402 şubeden rastgele seçilen 1500 gözlem biriminden ve 44 adet değişkenden oluşmaktadır. Verinin anlamlandırılması aşamasında ise istatistiksel ölçütlere (max, min, mean, median, mode, std, variance) ve veri tipine göre sınıflandırma yapılmıştır. Verinin hazırlama aşamasında ise veri temizleme, veri dönüşümü ve veri standardizasyonu/normalizasyonu yapılmıştır.

Verinin hazırlanması adımının sonucunda düzenlenen veriler bankanın tüzel müşterilerinden elde edilen POS ciro tahmini için Train&Test modeli ve 10 Katlı Çapraz Doğrulama yaklaşımı kullanılarak makine öğrenmesi yöntemlerinden Tek Değişkenli Lineer Regresyon, Çok Değişkenli Lineer Regresyon, Ridge Regresyon, Lasso Regresyon, K-En Yakın Komşu, Random Forest, Karar Ağaçları ve XGBoost algoritmaları kullanılmıştır. Her model %80 eğitim-%20 test ve %70 eğitim-%30 test şeklinde ayrılarak iki farklı eğitim ve test büyüklüklerinde çalışma yapılmıştır.

Algoritmaların performans ölçümü için r^2 , MSE, RMSE ve MAE değerleri esas alınarak değerlendirilmiştir. %80 eğitim-%20 test şeklinde ayrılan veri setinin genel olarak başarılı olduğu tespit edilmiştir. XGBoost (%80 Eğitim-%20 Test) algoritmasının ise diğer algoritmalara göre üstünlük sağladığı görülmüştür. Literatürde yapılan çalışmalarla kıyaslandığında XGBoost algoritmasının farklı sektörlerin ciro tahmininde de iyi performans gösterdiği görülmektedir.

Yapılan çalışmalar sonucunda kurulan model bankanın tüm tüzel müşterilerine uygulanarak potansiyel müşteri tespiti, ciro artışı ve pazar payı artışı sağlanabilir. Böylelikle portföy sorumlularının 6 aylık ve yıllık POS ciro hedeflerine ulaşmalarında kolaylık sağlanabilir. Farklı algoritmalarla müşteri sayısının ve girdi değişkenlerinin arttırımı üzerine çalışmalar yapılabilir. Ayrıca bankaların ve ödeme kuruluşlarının diğer bankacılık ürünlerinden elde edilecek ciro veya gelir değerleri üzerine çalışmalar yapılabilir.

Çalışmanın sonucu olarak, makine öğrenmesi yöntemleriyle POS ciro tahmini yüksek başarı oranıyla gerçekleştirilebilmektedir. Gözlem biriminin arttırılmasıyla birlikte eğitim ve test aşamalarında daha fazla verinin olması, başarı sonucunu olumlu yönde etkileyecektir.

- [1] J. L.-M. L.-C.-M. IsraelGonzález-Carrasco, *Information Sciences*, pp. 319-346, June 2019.
- [2] E. Ç. Muhammet ATALAY, «BÜYÜK VERİ ANALİZİNDE YAPAY ZEKÂ VE MAKİNE ÖĞRENMESİ,» *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, pp. 155-172, Aralık 2017.
- [3] Ö. M. Gökalp, «Makine Öğrenmesi - Machine Learning,» *ResearchGate*, pp. 1-14, Temmuz 2022.
- [4] M. C. Pedro Guerra, «Machine Learning Applied to Banking Supervision,» *Risks*, pp. 1-24, July 2021.
- [5] K.-M. T.-Ç. Y.-T. D. N. Chia-Nan Wang, «Forecasting and Measuring Financial Performance in Banking Sector,» *ResearchGate*, pp. 1-20, June 2022.
- [6] P. A. Ayşenur Budak, «Profit margin prediction in sustainable road freight transportation using machine learning,» *Journal of Cleaner Production*, pp. 1-18, September 2021.
- [7] L. E. Alexander K. Sepenu, «A Machine Learning Approach to Revenue Generation within the Professional Hair Care Industry,» *SMU Data Science Review*, pp. 1-29, 2022.
- [8] V. O. Petr Hajek, «Municipal Revenue Prediction by Ensembles of Neural Networks and,» *WSEAS TRANSACTIONS on COMPUTERS*, pp. 1255-1264, November 2010.
- [9] H. Li, «Using machine learning forecasts movie revenue,» *Institute of Electrical and Electronics Engineers (IEEE)*, 24 June 2022.

- [10] P.-F. Y.-M. P.-T. Kuo-PingLin, «Revenue forecasting using a least-squares support vector regression model in a fuzzy environment,» *Information Sciences*, pp. 196-209, 20 January 2013.
- [11] D. P. v. O. E. Elma, «Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey,» *International Journal of Trade, Economics and Finance*, pp. 435-440, October 2014.
- [12] M.-F. Y. Yu-Pei Huang, «A new perspective of performance comparison among machine learning algorithms for financial distress prediction,» *Applied Soft Computing*, October 2019.
- [13] A. R.-C. R. R. M. A. & T. B. O. Mohammad Zounemat-Kermani, «Machine Learning and Water Economy: a New Approach to Predicting Dams Water Sales Revenue,» *Water Resources Management*, p. 1893–1911, 2020.
- [14] M. Z. A. v. M. Awad, «Prediction of Olive Oil Productivity using Machine Learning Decision Tree Algorithm,» *International Journal on Emerging Technologies*, pp. 13-18, November 2020.
- [15] K. A. S. S. R. S. & S. S. G. P. Sanjana Rao, «Machine Learning based Restaurant Revenue Prediction,» *Evolutionary Computing and Mobile Sustainable Networks*, pp. 363-371, August 2020.
- [16] M. W. U. K. v. M. T. H. Austin Schmidt, «Machine Learning Based Restaurant Sales Forecasting,» *Machine Learning and Knowledge Extraction*, pp. 105-130, January 2022.
- [17] Y. H. (. C. Y. D. B. L. XI CHEN, «Predicting Future Earnings Changes Using Machine Learning and Detailed Financial Data,» *Journal of Accounting Research*, pp. 467-515, March 2022.

- [18] J. H. P. K. Taegu Kim, «Box office forecasting using machine learning algorithms based on SNS data,» *International Journal of Forecasting*, pp. 364-390, June 2015.
- [19] L. G. M. N. K. Ashok Kumar Pundir, «Machine Learning for Revenue Forecasting: A Case Study in Retail business,» *Institute of Electrical and Electronics Engineers (IEEE)*, 22 December 2020.
- [20] S. Ş. v. M. Pourhomayoun, «Predicting Movie Market Revenue Using Social Media Data,» *Institute of Electrical and Electronics Engineers (IEEE)*, November 2017.
- [21] N. Milosevic, «Equity forecast: Predicting long term stock price movement using machine learning,» *Computer Science , Machine Learning*, November 2018.
- [22] S. M. K. P. M. S. Narayana Darapaneni, «Demand and Revenue Forecasting through Machine Learning,» *Institute of Electrical and Electronics Engineers (IEEE)*, April 2019.
- [23] E. A.-G. A. K. A. F. D. K. F. Emmanuel Adinyira, «Application of machine learning in predicting construction project profit in Ghana using Support Vector Regression Algorithm (SVRA),» *Engineering, Construction and Architectural Management*, June 2021.
- [24] S. I. S. M. S. T. Sunitha Cheriyan, «Intelligent Sales Prediction Using Machine Learning Techniques,» *Institute of Electrical and Electronics Engineers (IEEE)*, March 2019.
- [25] D. E. AKTAN, «Big Data: Application Areas, Analytics and Security,» *Bilgi Yönetimi*, pp. 1-22, Haziran 2018.
- [26] H. Jiang, *Machine Learning Fundamentals*, USA, 2021.

- [27] M. U. Barış Karabay, «Büyük Veri İşlemede Yaygın Kullanılan Araçların Karşılaştırılması,» *ResearchGate*, Ocak 2017.
- [28] B. Mahesh, «Machine Learning Algorithms -A Review,» *International Journal of Science and Research (IJSR)*, pp. 381-386, January 2019.
- [29] M. X. B. M. A. K. M. H. Kevin Berwind, «Towards a Cross Industry Standard Process to support Big Data Applications in Virtual Research Environments (Long Paper),» %1 içinde *Collaborative European Research Conference (CERC) 2016*, Ireland, 2017.
- [30] T. Jo, *Machine Learning Foundations-Supervised, Unsupervised, and Advanced Learning*, Switzerland: Springer Nature, 2021.
- [31] M. Can, «İşletmelerde Zaman Serileri Analizi ile Tahmin,» İstanbul, 2019.
- [32] Y. Zhang, *New Advances in Machine Learning*, India, 2010.
- [33] T. B. a. C. M. R. (. Wolfgang Kersten, «A literature review on machine learning in supply chain management,» %1 içinde *Proceedings of the Hamburg International Conference of Logistics (HICL)* –, Berlin, 2019.
- [34] L. H. v. D. S. Tomislav Horvat, «The Impact of Selecting a Validation Method in Machine Learning on Predicting Basketball Game Outcomes,» *Symmetry*, March 2020.
- [35] G. C. McDonald, «Ridge Regression,» *WIREs Computatinal Statistics*, pp. 93-100, 13 July 2009.
- [36] J. v. J. Cook, «LASSO Regression,» *BJS Statistical Editors*, pp. 1338-1348, 7 August 2018.
- [37] T. E. Kalaycı, «Kimlik Hırsız WEB Sitelerinin Sınıflandırılması için Makine Öğrenmesi,» *Pamukkale Üniveristesi Mühendislik Bilimleri Dergisi*, pp. 870-878, 18 Şubat 2018.

- [38] J. Brownlee, Gradient Boosted Trees with XGBoost and Scikit-Learn, 2018.
- [39] M. G. A. Jessica Pesantez-Narvaez, «Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression,» *Machine Learning Insurance*, 20 June 2019.
- [40] Y. Gültepe, «Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine,» *European Journal of Science and Technology*, pp. 8-15, August 2019.
- [41] Z. Wang ve A. C. Bovik, «Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures,» *IEEE Signal Processing Magazine*, pp. 98-117, January 2009.
- [42] F. Ö. Koçoğlu, «Veri Madenciliğinde Veri Ayırıklaştırma Yöntemlerinin Karşılaştırılması ve Bir Uygulama,» *İSTANBUL ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ Yüksek Lisans Tezi*, Haziran 2012.
- [43] C. C. Aggarwal, An Introduction to Outlier Analysis, New York, 2016.
- [44] T. İ. Kurumu, «Ciro Endeksleri, Kasım 2021,» Ankara, 2022.
- [45] J. Lederer, «Linear Regression,» *Fundamentals of High-Dimensional Statistics*, pp. 37-79, November 2021.
- [46] E. B. F. Y. V. T. M. Ş. A. Ö. Deniz KILINÇ, «KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makele Tasnifi,» *Marmara Fen Bilimleri Dergisi*, pp. 89-94, 2016.

Konferans Bildirileri

T. Sayın ve T. Demirel, «Bankacılık Sektöründe Makine Öğrenmesi Yöntemlerini Kullanarak POS Ciro Tahmini,» *10. ULUSLARARASI MÜHENDİSLİK MİMARLIK VE TASARIM KONGRESİ*, İstanbul, 2022.

