
GAZİ ÜNİVERSİTESİ
Gazi
1926

ENERJİ SEKTÖRÜNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK MÜŞTERİLERİN TAHSİLAT POTANSİYELLERİNİN DEĞERLENDİRİLMESİ

Emine Ceren ÖZAY

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Kasım 2022

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

Bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Emine Ceren ÖZAY

24/11/2022

ENERJİ SEKTÖRÜNDE MAKİNE ÖĞRENMESİ ALGORİTMALARI KULLANILARAK MÜŞTERİLERİN TAHSİLAT POTANSİYELLERİNİN DEĞERLENDİRİLMESİ

(Yüksek Lisans Tezi)

Emine Ceren ÖZAY

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Kasım 2022

ÖZET

Günümüzde elektrik birim fiyatlarının artmasıyla birlikte kaçak elektrik kullanımı artmakta ve müşterilerin kaçak elektrik faturalarını ödeme oranları düşmektedir. Kaçak elektrik tüketimi yapan müşteriler genel olarak faturalarını ödememe eğiliminde olduğundan bu müşterilerden tahsilat yapılıp yapılamayacağını belirlemek güçtür. 2004 yılında başlayan özelleştirmeler ile Türkiye 21 elektrik dağıtım bölgesine ayrılmış olup özelleştirme sonrasında kaçak elektrik tüketimin tespitinin yapılması, faturalandırılması ve tahsilatı, elektrik dağıtım şirketlerinin sorumluluğu olarak belirlenmiştir. Bu sorumluluğun karşılığında dağıtım şirketleri kaçak elektrik faturalarının tahsilat miktarı üzerinden belirli oranda gelir sağlamaktadır. Bu nedenle dağıtım şirketleri için hem şirket kazancının arttırılması hem de nakit akışının iyileştirilmesi için kaçak elektrik kullanımı yapan müşterilerin ödeme potansiyellerinin tespit edilmesi önemli hale gelmiştir. Bu çalışmada, Türkiye'de faaliyet gösteren üç elektrik dağıtım şirketine ait kaçak elektrik kullanımı yapan müşterilerin, 2020 yılı verileri kullanılarak borçlarını ödeyip, ödemeyecekleri makine öğrenmesi algoritmaları ile analiz edilmiştir. Müşterilerin geçmiş ödeme alışkanlıkları, borç miktarları, demografik verileri, bölge kaçak oranları girdi değişkeni olarak kullanılmıştır. Çıktı değişkeni ise öder/ ödemez şeklinde ikili kategorik olarak dikkate alınmıştır. Veri kümesine uygun olan denetimli öğrenme sınıflandırma yöntemlerinden; lojistik regresyon, destek vektör makineleri, yapay sinir ağları, k-en yakın komşu algoritması, rastgele ormanlar, gradyan artırma makineleri, aşırı gradyan artırma, hafif gradyan artırma makineleri algoritmaları kullanılarak modeller oluşturulmuştur. Model performanslarını artırmak ve aşırı öğrenmenin önüne geçmek için k-katlamalı çapraz doğrulama ve hiper parametre ayarlama yapılmıştır. Ayrıca karar ağacı tabanlı topluluk yöntemleri olan rastgele ormanlar, gradyan artırma, aşırı gradyan artırma, hafif gradyan artırma algoritmaları için değişken önem düzeyleri belirlenmiştir. Model tahmin performans değerleri ve değişken önem düzeyi analizi sonuçları değerlendirildiğinde en iyi sonuç veren model olarak gradyan artırma makineleri modeli belirlenmiştir. Müşteri ödeme durumu analizinin sadece kredi sağlayan kuruluşların problemi olmaktan çıkıp diğer sektörlerde de uygulanabileceğini gösteren bu çalışma; bundan sonraki benzer çalışmalara yol göstererek literatüre bu anlamda katkı sağlayacaktır.

Bilim Kodu	:	90619
Anahtar Kelimeler	:	Borç tahsilatı, kredi risk analizi, makine öğrenmesi, denetimli öğrenme
Sayfa Adedi	:	81
Danışman	:	Doç. Dr. Erman ÇAKIT

ASSESSING THE COLLECTION POTENTIAL OF CUSTOMERS IN THE ENERGY SECTOR USING MACHINE LEARNING ALGORITHMS

(M. Sc. Thesis)

Emine Ceren OZAY

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

November 2022

ABSTRACT

Today, with the increase in electricity unit prices, the use of theft electricity increases and the rate of customers to pay their theft electricity bills decreases. Since customers who consume illegal electricity generally tend not to pay their bills, it is difficult to determine whether collections can be made from these customers. With the privatizations that started in 2004, Turkey was divided into 21 electricity distribution regions, and after the privatization, determining, invoicing and collection of theft electricity consumption was determined as the responsibility of electricity distribution companies. In return for this responsibility, distribution companies provide a certain amount of income over the collection amount of illegal electricity bills. For this reason, it has become important for distribution companies to determine the payment potential of customers who use theft electricity in order to increase the company's earnings and improve cash flow. In this study, it was analysed by machine learning algorithms whether the customers who use theft electricity belonging to three electricity distribution companies operating in Turkey, will pay their debts using 2020 data. Customers' previously payment habits, debt amounts, demographic data, and regional electricity theft rates were used as input variables. The output variable has been considered as binary categorical as pays or not pays. One of the supervised learning classification methods suitable for the data set; models were created using logistic regression, support vector machines, artificial neural networks, k-nearest neighbour algorithm, random forests, gradient boosting machine, extreme gradient boosting machine, light gradient boosting machine algorithms. K-fold cross validation and hyperparameter tuning were performed to increase model performances and prevent over-learning. In addition, variable importance levels were determined for random forest, gradient boosting machine, extreme gradient boosting machine, light gradient boosting machine, which are decision tree-based ensemble methods. When the model estimation performance metrics and variable significance analysis results were evaluated, the gradient boosting machine model was determined as the model that gave the best results. This study, which shows that customer payment status analysis is not only the problem of credit institutions and can be applied in other sectors, will contribute to the literature in this sense by guiding similar studies from now on.

Science Code : 90619

Key Words : Debt collection, credit risk analysis, machine learning, supervised learning

Page Number : 81

Supervisor : Assoc. Prof. Erman ÇAKIT

TEŞEKKÜR

Çalışmalarım boyunca değerli yardımları ve katkılarıyla beni yönlendiren, her konuda destek olan danışman hocam Sn. Doç. Dr. Erman ÇAKIT'a sonsuz teşekkürlerimi sunarım. Hayatım boyunca karşılıksız sevgi, emek ve fedakârlıkları ile her zaman yanımda olan, eğitim sürecim boyunca desteklerini hiçbir zaman esirgemeyen, sevgili annem Mukaddes ACAR'a, babam Ahmet ACAR'a, eşim Arda İdris ÖZAY'a çok teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

KISALTMALAR

Bu çalışmada kullanılmış kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklamalar
ABD	Amerika Birleşik Devletleri
CM-CoP	Birleşik madencilik tabanlı ödeme davranışı tahmini
KA	Karar ağaçları
DN	Doğru negatif
DP	Doğru pozitif
DVM	Destek vektör makineleri
EPDK	Elektrik Piyasası Denetleme Kurumu
GAM	Gradyan artırma makineleri
GARA	Gradyan artırılmış regresyon ağacı
KEYK	K en yakın komşu algoritması
KOBİ	Küçük ve orta ölçekli işletmeler
LDA	Lineer diskriminant analizi
HGAM	Hafif gradyan artırma makineleri
LR	Lojistik regresyon algoritması
CLAR	Cezalandırılmış lojistik ağaç regresyon
RO	Rastgele ormanlar algoritması
RAU	Rastgele alt uzaylar
TC/VKN	Türkiye Cumhuriyeti Kimlik ve Vergi Numarası
AGAM	Aşırı gradyan artırma makineleri
YN	Yanlış negatif
YP	Yanlış pozitif
YSA	Yapay sinir ağları

1. GİRİŞ

Türkiye'de elektrik piyasası Elektrik Piyasası Denetleme Kurumu (EPDK) tarafından denetlenmekte olup, enerji arzı elektrik dağıtım şirketleri tarafından sağlanmakta ve elektrik satışı ise perakende elektrik satış firmaları aracılığıyla yapılmaktadır. Türkiye'de 2004 yılında başlayan elektrik piyasasının özelleştirilmesi ile 21 dağıtım bölgesi oluşturulmuştur. Yönetmeliklerle belirlenen kurallar doğrultusunda elektrik dağıtım şirketleri enerji arzını sağlamak, şebeke yatırımlarını yapmak, arıza bakım faaliyetlerini gerçekleştirmek ve kayıp-kaçak enerji miktarını azaltmak için çalışmalar yapmaktadır.

Kayıp- kaçak miktarı enerji şebekesine giren enerjiden, satılan enerjinin çıkarılması ile hesaplanmaktadır. Kayıp elektrik enerjisi şebeke üzerinde elektriğin iletimi sırasında gerçekleşen teknik kayıplara karşılık gelmektedir. Teknik kayıplar elektrik iletiminin doğası gereği yaşanmaktadır. Kaçak elektrik enerjisi ise ilgili mevzuata aykırı olarak müşteriler tarafından tüketilen elektrik enerjisini tanımlamaktadır. Kaçak elektrik kullanımı 21 dağıtım bölgesinde farklı oranlarda gerçekleşmekte olup tüm dağıtım şirketlerinin EPDK tarafından belirlenen ilgili yıl hedef kayıp-kaçak oranları mevcuttur. Dağıtım şirketlerinin gelir elde edebilmek ve ilgili yıl hedeflerini tutturmak için kaçak elektrik kullanımlarını tespit etmek ve faturalandırmak en büyük amaç ve faaliyetlerinden biridir. Faturalanan kaçak tüketiminin belli oranında dağıtım şirketine gelir yazılmaktadır. Kaçak elektrik faturaları vadesinde ödenmediğinde yasal yollardan tahsil edilebilmektedir. Yasal yollardan tahsil edilen borçlar için ekstra kazanç elde edilmektedir. Hem süreci etkin yönetebilmek hem de tahsilat oranlarını artırmak için müşterilerin ödeme yapıp yapmayacaklarının bilinmesi önemli bir problem haline gelmektedir.

Elektrik birim fiyatlarında yaşanan artışlarla birlikte elektrik faturaları da artmaktadır. Bu da kaçak elektrik faturalarına yansımakta ve tahsilatını zorlaştırmaktadır. Vadesi geçen ödemeler, gecikmeler ve tahsil edilemeyen alacaklar firmaların finansal istikrarı açısından oldukça önemlidir. Son dönemlerde Türkiye'de ödeme gecikmelerinin sayısı ve vadesi geçen borçların yüzdesi artmaktadır (Onar, Öztayşi, Kahraman ve Öztürk, 2020).

Müşterilerin borç ödeme durumlarını konu alan çalışmalar incelendiğinde genellikle bankacılık ve finans alanlarına yönelik kredi risk skoru hesaplama çalışmalarına

rastlanmaktadır. Kredi risk skorlama çalışmaları mevcut veriler altında müşterinin ödeme yapıp yapmayacağını yansıtmakta ve finansal olarak kurumları gözetmektedir. Dağıtım şirketleri de müşterilerin ödeme yapıp, yapmayacaklarını tespit ederek;

- Etkin müşteri/icra takibi
- Avukatlara atanan dosyaların daha adil dağıtılması
- Avukat performans ölçümlerinin daha sağlıklı yapılması
- Tahsil edilmeyecek borçlar için yeni tahsilat stratejilerinin belirlenebilmesi
- İcra masrafları bütçesinin daha etkin yönetilmesi
- Operasyonel süreçlerin iyileştirilmesi için aksiyonlar almayı
- Nakit akışında artış sağlamayı hedeflemektedir.

Bu çalışma sonucunda eldeki veriler kullanılarak oluşturulan modeller ve çıktılarla müşterilerin tahsilat oranı ve operasyon süreçlerinin iyileştirilmesinde kullanılmak üzere ödeme durumu üzerinde girdi değişkenlerinin etkisinin sunulması hedeflenmektedir.

Kararların veriye dayandırılması fikri yeni değildir. Yöneylem araştırması, istatistik ve yönetim bilgi sistemleri gibi disiplinler bir süredir mevcut ve veri analitiği yardımıyla firmalara verimlilik sağlamaya çalışmaktadır (Sarkar, Bali ve Sharma, 2018: 4).

Müşteri ödeme durumu analizleri veri odaklı olup istatistiksel analiz yöntemleri ile başlayan çalışmalar, günümüzde makine öğrenmesi ve yapay zekâ analizleriyle çözümlenmeye devam etmektedir. İşletmelerin makine öğrenmesi ve yapay zekâ gibi nispeten yeni alanlara yoğun yatırım yapmasının temel nedeni verilerden önemli bilgiler veya fikirler elde edebilmektir. Elektrik dağıtım firmaları da zengin veri kümesine sahip ve regülasyona tabi firmalar oldukları için mevcut verilerini işleyerek hızlı aksiyonlar almayı amaçlamaktadırlar.

Bu çalışmada, makine öğrenmesi algoritmaları ile elektrik dağıtım şirketi müşterilerinin borçlarını ödeme/ödememe durumu tahmin edilmektedir. Çalışma kapsamında denetimli makine öğrenmesi algoritmalarından sınıflandırma modelleri kullanılmıştır. Çalışma beş bölüm olarak ele alınmıştır.

Birinci kısım Giriş bölümünde problemin tanımı, tezin amacı, kapsamı ve tez çalışmasında kullanılan yöntemlerden bahsedilmiştir. İkinci bölümde literatürde yer alan çalışmalar ile ilgili bilgi verilmiştir. Üçüncü bölümde mevcut veri detaylandırılarak, problem çözümünde kullanılan yöntemlerin detayları verilmiştir. Dördüncü bölümde önerilen modellerin elektrik dağıtım sektörü uygulamasına yer verilmiştir. Beşinci bölümde ise sonuçlar yorumlanmış ve gelecek çalışmalara dair önerilere yer verilmiştir.

2. LİTERATÜR ÇALIŞMASI

Tez çalışmasının bu adımında literatürde, müşterilerin ödeme durumunu tahmin etmek için yapılmış olan çalışmalara değinilmektedir. Literatürde, müşteri borç ödeme durumu tahmininde bankacılık ve finans sektöründe kredi risk puanı analizi için yapılan çalışmalar çoğunluktadır. Makine öğrenmesi yöntemlerinin yaygınlaşması ve ekonomik krizlerle birlikte müşterilerin ödeme durumları analizi firmalar için de önemli hale gelmeye başlamıştır. Bunun sonucunda firmalara ait müşterilerin borçlarını ödeyip, ödemeyeceklerinin belirlenmesi adına yapılan çalışmalar artmaktadır. Ayrıca finansal krizlerle birlikte işletmelerin veya kurumların kriz karşısındaki davranışlarının analiz edilmesi çalışmaları da literatürde yer almaktadır.

Literatür çalışmaları işledikleri konu bakımından 3 ayrı bölümde ele alınacaktır: i) kredi risk analizi, ii) borç tahsilatı analizi, iii) finansal risk analizi

2.1. Kredi Risk Analizi

Kredi risk analizi çalışmaları temelde müşterilerin ödeme yapıp, yapmayacağının belirlenmesi işlemidir. Kredi borçlarının ödeme ihtimalinin belirlenmesi, fatura ödeme ihtimali ile benzerdir. Literatürde bu alanda çoğunlukla bankalar için yapılmış çalışmalar bulunmaktadır.

West (2000) yaptığı bu çalışmada çeşitli sinir ağı modellerini birbirleri ile kıyaslamıştır. İki veri seti kullanılarak kredi puanlama modellerinin tahmin doğruluğu test edilmiştir. Alman veri seti; 700 krediye uygun başvuru, 300 krediye uygun olmayan örnek içermektedir. Avustralya veri seti; 307 uygun ve 383 uygun olmayan örnek ile daha dengeli veri içermektedir. Başvuru sahipleri için hesap bakiyeleri, kredi çekme amacı, kredi miktarı, kredi geçmişi, istihdam durumu, kişisel bilgileri, yaşı, mal varlığı ve iş durumu gibi 24 farklı değişken kullanılmıştır. Sonuçlar, çok katmanlı algılayıcının en doğru sinir ağı modeli olmayabileceğini ve hem uzman karışımı hem de radyal tabanlı sinir ağı modellerinin kredi modellerinin kredi puanlama uygulamaları için dikkate alınması gerektiğini göstermektedir. Bu araştırma aynı zamanda lojistik regresyonun, yapay sinir ağı modellerine iyi bir alternatif olduğunu öne sürmektedir. Araştırılan iki veri seti için de

lojistik regresyon, lineer diskriminant analizinden daha doğru sonuç vermiştir.

Abdou ve Pointon (2009) çalışmasında, 298 bankaya ait, t-1 dönemindeki veriler baz alınarak, t döneminde bankaların geri ödeme performansı tahminlenmiştir. Lojistik regresyon ve olasılıksal sinir ağları algoritmaları kullanılmıştır. Lojistik regresyon modelinin doğruluğu %98,5 olasılıksal sinir ağları yöntemin doğruluğu %98,3 olduğu görülmüştür.

Khandani, Kim ve Leo (2010) yaptıkları çalışmada, 3-12 ay öncesinde kredi risk tahmini yapabilmek için parametrik olmayan, doğrusal bir model kurmuştur. Veri seti Ocak 2005 ile Nisan 2009 dönemindeki müşteri hareketlerini ve kredi bürosu verilerini içermektedir. Makine öğrenmesi model sonuçlarına dayanarak maliyetten kazancının %6 ile %25 civarında olacağı tahminlenmektedir.

J. Kim ve Sohn (2010) çalışmalarında makine öğrenmesi algoritmalarını kullanarak teknoloji kredisi almak isteyen müşterilerine kredi verme kararını objektif olarak alabilmeyi hedeflemişlerdir. Lojistik regresyon, destek vektör makineleri, yapay sinir ağları yöntemleri ile modellenerek, model performansları karşılaştırılmıştır. Müşterilerinin mevcut finansal değerleri, teknoloji verileri ve ekonomi şartları parametre olarak kullanılarak, %66,2 doğruluk ile destek vektör makinesinin daha başarılı olduğu görülmüştür.

Xia, C. Lui, Y. Y. Li, N. Lui (2017) çalışmalarında, aşırı gradyan artırma makineleri (AGAM) algoritmasını kullanarak model oluşturmakta ve kredi puanlama işlemi yapmaktadır. Çalışma temel olarak üç adımda yapılmıştır. Birinci adım veri ön işleme ve eksik veri analizi, ikinci aşama gereksiz özniteliklerin elenmesi, üçüncü aşama ise model hiper parametrelerini analiz ederek en iyi sonuç veren modelin oluşturulmasıdır. Önerilen ayarlanmış model, kredi puanlama modelinin yorumlanabilirliğini artıran değişken önem düzeylerini ve karar ağacını da sunmaktadır.

Bao, Lianju ve Yue (2019) yaptıkları çalışmada kredi risk analizi yaparak kredi puanlaması belirleyebilmek adına finansal kurumlara başvuran “kötü” kredi müşterilerini “iyi” kredi müşterilerinden ayırt etmek için denetimsiz öğrenme ile denetimli öğrenme yöntemlerini entegre eden bir kombinasyon önermiştir. Model performansının karşılaştırmaları, dört

gruptaki üç kredi veri kümesine dayalı olarak gerçekleştirilir. Sonuç olarak hem fikir birliği aşamasında hem de veri kümesi kümeleme aşamasında entegrasyon, kredi puanlama modelleri performansını iyileştirmede etkili olduğu görülmüştür.

Zhu, L. Zhou, Xie, G. J. Wang ve Nguyen (2019) çalışmalarında KOBİ'lerin kredi risk tahmini doğruluğunu artırmak için iki temel makine öğrenmesi yöntemi, rastgele alt uzaylar (RAU) ve çoklu artırma algoritmasını birleştirerek RAU çoklu artırma adı verilen gelişmiş toplu öğrenme modeli geliştirmişlerdir. Küçük ve orta ölçekli işletmeler (KOBİ'ler) finansal faaliyetlerini sürdürebilmek için belirli durumlarda tedarik zinciri kredisini kullanmaktadır. KOBİ'lerin kredi riskinin tahminlenmesi, finansman kararların alınmasında kritik süreçlerden biri haline gelmiştir. 31 Mart 2014 ve 31 Aralık 2015 tarihleri arasında Çin piyasasından elde edilen veri seti, RAU çoklu artırma algoritmasını eğitmek için kullanılmıştır. Tahmin sonuçları, RAU çoklu artırmanın düşük sayılı veri setlerinde iyi bir performans gösterdiğini ortaya koymaktadır.

J. P. Li, Mirza, Rahat ve Xiong (2020) yaptıkları çalışmada, banka kredi puanlarını tespit etmek için yapay zekânın bir alt kümesi olan makine öğrenme tekniklerini kullanmıştır. Bankaya ait parametrelerin, 2010-2018 dönemindeki üçer aylık veri seti modellemede kullanılmıştır. Model sonuçları incelendiğinde performans değerlerinden en iyisi rastgele ormanlar yaklaşımına ait olduğu görünmektedir. Ancak, sınıflandırma ve regresyon ağacı sonuçlarının da önemli ölçüde iyileştiği görünmektedir. Rastgele ormanlar modeli ile sınıflandırma ve regresyon ağacı algoritmaları birlikte kullanılarak daha iyi bir sonuçlar elde edilebileceği belirtilmiştir.

Lappas ve Yannacopoulos (2021) çalışmalarında kredi risk tahmini algoritmasında uzman görüşlerinin etkisini göstermek için bir birleşim yöntemi önermiştir. Kredi risk tahmini algoritmaları otomatik olarak modellenerek sonuçlanacağı şekilde düşünülmektedir. İlgili çalışmada uzman görüşlerin kredi risk tahmini sürecindeki katılımının nasıl sağlanabileceği ve katkıları araştırılmaktadır. Önerilen iki adımlı çözüm ile birinci adımda analitik hiyerarşi süreci tekniği işletilerek öznitelik optimizasyonu ve önemleri belirlenmiştir. İkinci aşamada ise k en yakın komşu ve navie bayes algoritmalarının birleşimine dayanan modeller oluşturularak kredi risk tahmini gerçekleştirilir. Model etkinliğini doğrulamak için veri kümesi üzerinden oluşturulan test örnekleri kullanılmıştır. Çalışmanın sonucunda tekli modeller ile karşılaştırıldığında kombinasyon modelinin daha

iyi bir performans gösterdiği görülmüştür.

Moscato, Picariello ve Sperli (2021) çalışmalarında bir kredinin bir eşten eşe platformunda ödeme durumunu tahminlemek için kredi riski puanlama algoritmalarını karşılaştırmaktadır. Ayrıca çalışmada sınıf dengesizliği durumu incelenmiştir. Gerçek bir sosyal kredi platformundan alınan 877956 örneklem kullanılarak oluşturulan modeller performans ölçütleri bazında karşılaştırılmıştır. Rastgele ormanlar, lojistik regresyon, yapay sinir ağları algoritmaları kullanılarak yetersiz örneklem, aşırı örneklem ve hibrit yaklaşım altındaki sınıflandırma sonuçları değerlendirilmiştir.

Luong ve Scheule (2022) çalışmalarında 2000'den 2016'ya kadar Amerika Birleşik Devletleri'ndeki (ABD) birincil ipotek kredilerinden elde edilen gözlem değerlerini kullanarak, kredi risk analizi gerçekleştirilmiştir. Ortak borçlu olma durumu, kredi sözleşmesi tipi ve dış özelliklerin kredi riskini açıklamada etkin olduğu belirlenmiştir. Yaşam döngüsü etkileri ile ileriye dönük dönem değişken etkilerinin birleştirilmesinin, varsayılan tahminlerin doğruluğunu önemli ölçüde iyileştirdiği görülmüştür. Bu sebeple her iki yöntemi birleştiren hibrit bir model geliştirilmiştir. Model sonucundaki iyileşmelerin, daha doğru ekonomik sermaye, kredi zararı provizyonu ve ipotek fiyatlandırması ile ticari bankalarda daha verimli ve esnek bir kaynak tahsisi sağlayacağı öngörülmektedir.

Nazemi, Rezazadeh, Fabozzi ve Höchstötter (2022) yaptıkları çalışmada telekomünikasyon sektöründe Alman üçüncü taraf şirket tarafından alınan 65000 temerrüt kredisinin ödenme oranını tahmininde etmek için destek vektör regresyonu, artırma, derin öğrenme vb. birçok makine öğrenmesi yönteminden yararlanmışlardır. Ödeme oranı modelleri karşılaştırılırken temerrüt risk verilerine dayalı olarak ağırlıklı performans ölçüleri tanımlanmıştır. Derin öğrenme yönteminin kredi risk analizi için performans arttırmada başarılı bir yöntem olduğu görülmüştür.

Dumitrescu, Hurlin ve Tokpavi (2022) yaptıkları çalışma ile lojistik regresyon modelinin performansını iyileştirmek için karar ağacı modellerinden elde edilen verileri kullanan, cezalandırılmış lojistik ağaç regresyon (CLAR) adında kredi puanlama yöntemi önermiştir. Resmi olarak, orijinal değişkenlerle elde edilmiş karar ağaçlarından çıkarılan kurallar, öngörücü olarak cezalandırılmış lojistik regresyon algoritmasında kullanılır. CLAR,

lojistik regresyon modelinin içsel yorumlanabilirliğini korurken, kredi puanlama verilerinde ortaya çıkabilecek doğrusal olmayan etkilerin yakalanmasını sağlamaktadır. Dört gerçek kredi temerrüt veri seti kullanarak oluşturulan modeller sonucunda; CLAR'nin kredi risk sonucunun lojistik regresyondan daha iyi sonuç verdiği görülmüştür. Ayrıca, CLAR yöntemi kullanımı yanlış sınıflandırma maliyetlerini azaltmaktadır.

Kredi risk analizi konularında yapılmış olan literatür araştırması çalışmasının özet tablosu Çizelge 2.1'de verilmiştir.

Çizelge 2.1. Kredi risk analizi alanındaki literatür araştırması özet tablosu

Yazar, Yıl	Konu	Yöntem
West (2000)	Kredi risk analizi	Sinir Ağları, Sınıflandırma ve Regresyon Ağaçları, K-En Yakın Komşu Algoritması, Çekirdek Yoğunluğu Tahmini, Lojistik Regresyon, Lineer Diskriminant Analizi
Abdou ve Pointon (2009)	Kredi risk analizi	Lojistik Regresyon, Olasılıksal Sinir Ağları
Khandani, A. J. Kim ve Lo (2010)	Kredi risk analizi	Sınıflandırma ve Regresyon Ağaçları
J. Kim ve Sohn (2010)	Kredi risk analizi	Lojistik Regresyon, Yapay Sinir Ağları, Destek Vektör Makineleri
Xia, C. Liu, Y. Y. Li ve N. Liu (2017)	Kredi risk analizi	Gradyan Artırma Algoritması, Aşırı Gradyan Artırma Algoritması, Adaboost, Lojistik Regresyon, Rastgele Ormanlar, Destek Vektör Makineleri, Sinir Ağları, Karar Ağacı, Gradyan Artırma Karar Ağacı
Bao, Lianju ve Yue (2019)	Kredi risk analizi	Destek Vektör Makineleri, Lojistik Regresyon, Karar Ağacı, Yapay Sinir Ağları, Rastgele Ormanlar, Gradyan Artırma Karar Ağacı, K-En Yakın Komşu Algoritması, K- Ortalama
Zhu, L. Zhou, Xie, G. J. Wang ve Nguyen (2019)	Kredi risk analizi	Rastgele Alt uzay, Çoklu Artırma Modeli, Karar Ağacı
J. P. Li, Mirza, Rahat ve Xiong (2020)	Kredi risk analizi	Destek Vektör Makineleri, Yapay Sinir Ağları, Rastgele Ormanlar, Karar Ağacı
Lappas ve Yannacopoulos (2021)	Kredi risk analizi	Genetik Algoritma, K-En Yakın Komşu Algoritması, Naive Bayes, Analitik Hiyerarşi Prosesi
Moscato, Picariello ve Sperlí (2021)	Kredi risk analizi	Lojistik Regresyon, Rastgele Ormanlar, Çok Katmanlı Algılayıcı
Luong ve Scheule (2022))	Kredi risk analizi	Hibrit Regresyon Modeli

Çizelge 2.1. (devam) Kredi risk analizi alanındaki literatür araştırması özet tablosu

Nazemi, Rezazadeh, Fabozzi ve Höchstötter (2022)	Kredi risk analizi	Lineer Regresyon, Regresyon Ağaçları, Derin Öğrenme, Artırma, Destek Vektör Makineleri, Sinir Ağları
Dumitrescu, Hurlin ve Tokpavi (2022)	Kredi risk analizi	Cezalı Lojistik Ağaç Regresyonu, Rastgele Ormanlar, Lojistik Regresyon

2.2. Borç Tahsilatı Analizi

Borç tahsilat analizi müşterilerin faturalarını ödeme yapıp yapmayacaklarının tespit edilmesi problemleridir. Borç tahsilat analizi çalışmaları incelendiğinde literatürde genelde telekomünikasyon sektörüne yönelik çalışmaların ağırlıkta olduğu görülmektedir.

Deloffre ve Abdallah (2006) çalışmalarında markovian yöntemi kullanarak telekomünikasyon sektöründe borç ödeme süreci optimizasyonu için risk ve kayıp analizi modeli oluşturulmuştur. Model ile tahsilat sürecinde iptal etme tarihinin tespiti ve faturalama süreçlerinin optimizasyonu kararlarına odaklanılır. Tahsilat modellerinin performans ölçümü için yeni bir gösterge sunulmuştur.

C. H. Chen, Chiang, Wu ve Chu (2013) çalışmalarında etki alanına dayalı bir veri madenciliği stratejisi benimsenerek birleşik madencilik tabanlı müşteri ödeme davranışı tahmin sistemi (CM-CoP) oluşturmak için sabit hat kullanıcılarından gelen veriler birliktelik kuralları, kümeleme ve karar ağaçları kullanılarak analiz etmiştir. Tasarlanan geç ödeme tahmin sistemi ücreti zamanında ödeyemeyen potansiyel kullanıcıları göstermektedir. Önerilen sistemin uygulanmasında, müşteri ödeme davranışını analiz etmek için ilk olarak birliktelik kuralları kullanılmış ve analiz sonuçları türev niteliklerini oluşturmak için kullanılmıştır. Daha sonra müşteri segmentasyonu için kümeleme algoritması kullanılmıştır. Son olarak, telekomünikasyon firmaları tarafından sağlanan türev öznitelikler kullanarak verilerin geri kalanını tahmin ve analiz etmek için bir karar ağacı oluşturulmuştur. Değerlendirme sonuçlarında, CM-CoP modelinin ortalama doğruluğu, altı aylık bir doğrulamadan sonra ortalama %88,1 duyarlılık değeri %11,2 ortalama kazanç altında %78.5 olmuştur. Telekomünikasyon firmaları tarafından kullanılan mevcut yöntemin tahmin doğruluğu %65,6 iken önerilen modelin tahmin doğruluğu %13 daha fazla olmuştur. Başka bir deyişle, sonuçlar CM-CoP modelinin etkili olduğunu ve kullanılan mevcut yaklaşımdan daha iyi olduğunu göstermektedir.

J. Kim ve Kang (2016) çalışmalarında halihazırda ödenmemiş borcu olan müşteriler için ödeme olasılığını ve geç ödeme miktarını tahmin etmek amacıyla makine öğrenmesi tabanlı beş adet geç ödeme tahmin modeli ve on müşteri puanlama kuralı geliştirmiştir. Önerilen puanlama kuralları, aracı sayısı değiştirilerek on farklı bağlamda doğrulanır. Deneysel sonuçlar, tahmin modeline dayalı puanlama kurallarının, mevcut buluşsal tabanlı müşteri puanlama kurallarına kıyasla aracılar arasında daha adil müşteri tahsisi sonuçlarına yol açtığını doğrulamaktadır. Tahmin modelleri arasında hibrit bir yaklaşım, geç ödeyenleri etkili bir şekilde yakalayabilirken, ağaç tabanlı modeller diğer yöntemlere göre daha tarafsız müşteri dağılımı gösterdiği görülmüştür.

Höglund (2017) çalışmasında vergi ödeme temerrütlerini tahmin etmek için genetik algoritma tabanlı bir karar destek aracı geliştirilmiştir. Finlandiya vergi dairelerinden alınan istatistiklere göre, 2015 yılı sonunda Finlandiya'daki tüm aktif firmaların yaklaşık %12'si ödenmemiş vergiye sahiptir. İncelenen firmaları temerrüde düşen veya temerrüde düşmeyen olarak sınıflandıran bir lineer diskriminant analizi (LDA) modeli için optimal veya alt kümesini optimale yakın bir değişken belirlemek için ise genetik algoritma kullanılmıştır. Sistemin çıktısı, hem vergi temerrütlerini tahmin etmede çeşitli değişkenlerin önemi hakkında bilgi sağlayan bir liste hem de LDA tabanlı bir vergi temerrüt tahmin modelidir. Veri seti, işveren katkı vergileri veya katma değer vergilerinde temerrüde düşmüş Fin limitet şirketlerinden oluşmaktadır ve mevcut değişkenlerin toplam sayısı 72'dir. Sonuçlar, ödeme gücü, likidite ve ticari borçların ödeme süresini ölçen değişkenlerin tahminde önemli değişkenler olduğunu göstermektedir. En iyi performans gösteren model, doğrusal olmayan dönüştürülmüş üç değişken içerir ve %73,8'lik bir tahmin doğruluğuna sahiptir.

Ozan (2018) çalışmasında segmentleri manuel belirlenmiş müşterileri inceleyerek, makine öğrenmesi yöntemleri ile müşteri borç ödemelerini tahmin ederek şirket için veri segmentasyon problemini gidermektedir. Müşterilerin ödeme verileri girdi parametreleri olarak alınarak, müşteri segmentlerini standart veya özel olarak ayırt emek için sınıflandırma yöntemleri kullanılmıştır. Lojistik regresyon yöntemi en iyi sonuç veren model olmuştur.

Khansong, Kamjana, Laitrakun ve Usanavasin (2020) yaptıkları çalışmada müşteri hizmetlerini iyileştirmek ve geç ödenen veya ödenmemiş elektrik faturalarının sayısını

azaltmak amacıyla konut müşterilerinin elektrik ödeme davranışlarını tahmin etmek için veri madenciliği algoritmaları kullanılarak sonuçları karşılaştırmıştır. Tayland İl Elektrik Kurumu'nun 19,4 milyon konut müşterisi vardır ve bunların yaklaşık %73'ü elektrik borçlarını geç ödemekte veya ödenmemiş faturaları bulunmaktadır. Bu, Tayland İl Elektrik Kurumu gelirinin azalması da dahil olmak üzere çeşitli sorunlara neden olmaktadır. 1079 müşteriden toplanan müşterilerin ödeme davranışları ve demografik özelliklerinden oluşan veri seti kullanılmıştır. İki aşamalı çalışmada k-ortalama algoritması kullanılarak müşteri sınıflarının kümelenmesi belirlenmiş ve müşteri modeli tahmini için lojistik regresyon, karar ağacı, rastgele ormanlar, destek vektör makinesi ve aşırı gradyan artırma makineleri algoritmaları kullanılmıştır. Müşterilerin memnuniyetini ve Tayland İl Elektrik Kurumunun gelirini artırmak için uygun stratejiler geliştirmede F1 skoruna göre bakıldığında %97,9 ile rastgele orman algoritmasının en iyi seçim olduğu görülmüştür.

Bahrami, Bozkaya ve Balcısoy (2020) çalışmalarında fatura ödemelerine ilişkin müşteri davranışını anlamak ve ödeme davranışını tahmin etmek için analitik bir yaklaşım önermiştir. Çeşitli sektörlerden edinilen deneyimler, şirketlerin müşteri fatura ödemelerini tahsil etmekte sorun yaşayabileceğini göstermektedir. Araştırmalar ABD ve Birleşik Krallıktaki küçük ve orta ölçekli işletme ve işletmeler arası faturaların neredeyse yarısının geç ödendiğini bildirmektedir. Analizde 1,6 milyondan fazla müşteri ve onların fatura ve ödeme geçmişi ile faturayı düzenleyen şirket tarafından gerçekleştirilen çeşitli bilgilendirme verileri (örneğin: e-posta, kısa mesaj hizmeti, telefon görüşmesi) kullanılarak büyük bir veri seti üzerinde çalışılmıştır. Test edilen denetimli ve denetimsiz üç makine öğrenimi yaklaşımı arasında, %97'ye varan doğruluk sağlayan lojistik regresyon algoritmasının en iyi sonuç verdiği görülmüştür.

Ma ve Fildes (2020) yaptıkları çalışmalarında mağazaların toplam müşteri akışlarını tahmin etmek yerine, katılımcı mağazaların günlük müşteri akışlarındaki paylarını tahmin etmek için üçüncü taraf mobil ödeme verilerini kullanmıştır. Katılımcı perakendeciler arasındaki verilerdeki ortak noktalardan yararlanmak için yeni bir havuzlanmış yaklaşım olan gradyan artırılmış regresyon ağacı (GARA) yöntemine dayanan ve aynı anda birçok mağaza için çok adımlı günlük mobil ödeme tahmin çözümü önerilmiştir. 2000 mağazadan oluşan bir örneklem üç hata ölçümünde değerlendirildikten sonra önerilen çözümün geleneksel zaman serisi modellerinden, havuzlanmış regresyondan veya popüler rastgele ormandan daha iyi performans gösterdiği görülmüştür. Çeşitli kategorilerdeki

büyük bir mağaza havuzuna dayalı günlük mobil ödeme müşteri akışı tahmininin, bireysel mağazalara dayalı yöntemlerle oluşturulan tahminlerden daha doğru tahminler üretebildiği bulunmuştur. Bu sonuç mağazaların, yalnızca kendi verilerini kullanarak türetilen tahminlere güvenmektense, paylaşılan platforma katılarak daha doğru müşteri akışı tahminleri elde edebileceğini göstermektedir.

Bellotti, Brigo, Gambetti ve Vrins (2021) yaptıkları çalışmada Avrupa borç tahsilat kurumundan alınan özel bir veri tabanını kullanarak, tahsili gecikmiş alacakların tahsilat oranlarını tahmin etmek için bir dizi regresyon teknikleri ve makine öğrenimi algoritmaları kullanmışlardır. Tahsili gecikmiş kredilerin geri kazanım oranlarını tahmin etme yetenekleri açısından 20 makine öğrenimi algoritmasının performansı karşılaştırılmıştır. Kübist, güçlendirilmiş ağaçlar ve rastgele ormanlar gibi kural tabanlı algoritmaların diğer yaklaşımlardan önemli ölçüde daha iyi performans gösterdiğini görülmüştür.

Coenen, Verbeke ve Gruns (2022) yaktıkları çalışmada, bir faturanın kabul edilebilir bir zaman diliminde ödenme olasılığını tahmin etmek için makine öğrenimi algoritmaları kullanmıştır. Riski tahmin etmek için önceden belirlenmiş bir gecikmiş gün sınırı için ikili sınıflandırma, gecikmiş günlerin gerilemesi ve tüm durumlar için riskle ilgili sıralamayı optimize etmeyi öğrenen sıralamayı öğrenme üç olası makine öğrenimi görevi araştırılmıştır. Regresyon yöntemleri, değerlendirme ölçütlerinde ve tüm yöntemlerde iyi değerler elde etmiştir. Kâr temelli değerlendirme sonuçlarına göre aşırı gradyan artırma makineleri algoritması en iyi performans sergileyen algoritma olarak tanımlanmıştır.

Borç tahsilatı analizi konularında yapılmış olan literatür araştırması çalışmasının özet tablosu Çizelge 2.2'de verilmiştir.

Çizelge 2.2. Borç tahsilatı analizi alanındaki literatür araştırması özet tablosu

Yazar, Yıl	Konu	Yöntem
Deloffre ve Abdallah (2006)	Borç tahsilatı analizi	Markov zinciri
C. H. Chen, Chiang, Wu ve Chu (2013)	Borç tahsilatı analizi	Veri madenciliği
J. Kim ve Kang (2016)	Borç tahsilatı analizi	Destek Vektör Makineleri, Tekli Sınıflandırma Ağacı, Rastgele Ormanlar, Hibrit model, Yapay Sinir Ağları

Çizelge 2.2. (devam) Borç tahsilatı analizi alanındaki literatür araştırması özet tablosu

Höglund (2017)	Borç tahsilatı analizi	Lineer Diskriminant Analizi, Genetik Algoritma
Ozan (2018)	Borç tahsilatı analizi	Çok Değişkenli Doğrusal Regresyon, Lojistik Regresyon, Normal Eşitliği
Khansong, Karnjana, Laitrakun ve Usanavasin (2020)	Borç tahsilatı analizi	Lojistik Regresyon, Karar Ağacı, Rastgele Orman Destek Vektör Makineleri, Aşırı Gradyan Artırma Algoritması
Bahrami, Bozkaya ve Balcisoy (2020)	Borç tahsilatı analizi	Lojistik Regresyon, Destek Vektör Makineleri, Tek Sınıf Sınırlandırma
Ma ve Fildes (2020)	Borç tahsilatı analizi	Gradyan Artırma Regresyon Ağacı
Bellotti, Brigo, Gambetti ve Vrins (2021)	Borç tahsilatı analizi	Sıradan En Küçük Kareler, Çok Değişkenli Regresyon, K-En Yakın Komşu Algoritması, Ortalamalı Sinir Ağları, Destek Vektör Makineleri, Uygunluk Vektörü Regresyonu, Gauss, Regresyon Ağaçları, En Küçük Mutlak Küçültme ve Seçim Operatörü, Rastgele Ormanlar, Kübist, Artırılmış Ağaçlar, Torbalı Ağaçlar, Lineer Regresyon
Coenen, Verbeke ve Guns (2022)	Borç tahsilatı analizi	Lojistik Regresyon, Kritik Düzenlileştirilmiş Regresyon, Rastgele Ormanlar, Destek Vektör Makineleri, Aşırı Gradyan Artırma

2.3. Finansal Risk Analizi

Şirketlerin faaliyetlerini sürdürebilmeleri için finansal risklerini öngörüp aksiyon almaları gerekmektedir. Literatürde finansal risk analizi çalışmaları genellikle firmaların ekonomik kriz ortamında başarılı olup olmayacaklarının belirlenmesi, iflas edip/etmeyeceklerinin tahmin edilmesi konularını içermektedir.

Koh ve Low (2004) çalışmalarında karar ağaçları, lojistik regresyon ve yapay sinir ağları algoritmalarını kullanarak firma başarısını analiz etmiştir. Veri seti olarak yarısı başarılı, yarısı başarısız olan ve 1980-1987 döneminde faaliyette olan 330 firma verisi kullanılmıştır. Modellemeler sonucunda firma başarısı tahminlemede en iyi sonuç veren model karar ağacı iken, en kötü sonuç veren ise lojistik regresyon algoritması olmuştur.
Shin, T. S. Lee, H. J. Kim (2005) yaptıkları çalışmada, DVM iflas tahmin problemine uygulanmasının etkinliğini araştırmaktadır. Geri yayılımlı sinir ağının örüntü tanıma

görevlerinde iyi performans gösterdiği iyi bilinen bir gerçek ağın ağırlıklarını belirlemek için mümkün olduğunca çok sayıda eğitim setinin yüklenmesi gerekir. Öte yandan, DVM, eğitim verilerinden ağların ağırlıklarını türetmeden özellik uzayının geometrik özelliklerini yakaladığı için, küçük eğitim seti boyutu ile en uygun çözümü çıkarma yeteneğine sahiptir. Sonuçlar, eğitim seti boyutu küçüldükçe DVM'nin doğruluk ve genelleme performansının geri yayılımlı sinir ağından daha iyi olduğunu göstermektedir.

Yapraklı ve Erdal (2016) çalışmalarında, yapay sinir ağları ve destek vektör makineleri algoritmaları ile firma başarısını tahminlemeyi amaçlamaktadır. İnşaat sektöründe yer alan bir firmanın 157 müşterisine ait veri seti kullanılmıştır. Ayrıca veri setini oluşturan değişkenler için temel bileşen analizi gerçekleştirilmiştir. TBA-DVM hibrit modelinin tahmin performansı diğer modellere göre daha yüksek çıkmıştır.

Barboza, Kimura ve Altman (2017) yaptıkları çalışmada Salomon Center veri tabanından ve Compustat'tan alınan bilgileri birleştirerek, Kuzey Amerika firmalarına ilişkin 10000'den fazla firmanın 1985'ten 2013'e kadar olan verilerini kullanılarak olaydan bir yıl önce iflas durumunu tahmin etme modeli geliştirmiştir. Destek vektör makineleri, torbalama, artırma ve rastgele orman makine öğrenmesi modelleri test edilerek, performansları diskriminant analizi, lojistik regresyon ve sinir ağlarından elde edilen sonuçlarla karşılaştırılmıştır. Tüm tahmin değişkenleriyle en iyi modeller karşılaştırıldığında, rastgele ormanlar algoritması %87 doğruluk sağlarken, lojistik regresyon ve doğrusal diskriminant analizinin sırasıyla %69 ve %50 doğruluk sağladığı görülmüştür.

Fantulin, Lagravinese ve Resce (2021) yapmış oldukları çalışmada, 2009-2016 döneminde 7795 İtalyan belediyesinin iflasının öngörülebilirliği analiz etmiştir. Veri setinde az sayıda iflas vakası olması nedeniyle tahmin görevi son derece zor olmuştur. Her belediye için tarihsel mali verilerin yanı sıra, sosyo-demografik ve ekonomik bağlamla birlikte alternatif kurumsal veriler kullanılmaktadır. Gradyan artırma makineleri kullanılarak modelleme yapılmış olup doğrulamak için en küçük mutlak küçültme ve seçim operatörü, rastgele ormanlar ve sinir ağları kullanılmıştır. Model sonucunda belediyelerin temerrüdünü tahmin etmek için bazı finansal olmayan özelliklerin (örneğin coğrafi alan); birçok finansal özellikten daha önemli olduğu görülmektedir. Modellere dahil edilen yöneticilerin sosyo-demografik özellikleri arasında belediye meclisindeki üyelerin cinsiyeti ve yaşı, belediye

temerrütlerini tahmin etme açısından önem arz eden ilk 10 özellik arasında yer almaktadır.

Finansal risk analizi konularında yapılmış olan literatür araştırması çalışmasının özet tablosu Çizelge 2.3' te verilmiştir.

Çizelge 2.3. Finansal risk analizi alanındaki literatür araştırması özet tablosu

Yazar, Yıl	Konu	Yöntem
Koh ve Low (2004)	Finansal risk analizi	Sinir Ağları, Karar Ağacı, Lojistik Regresyon
Shin, T. S. Lee ve H. J. Kim (2005)	Finansal risk analizi	Destek Vektör Makineleri, Geri Yayılım Ağı
Yapraklı ve Erdal (2016)	Finansal risk analizi	Destek Vektör Makineleri, Yapay Sinir Ağları
Barboza, Kimura ve Altman (2017)	Finansal risk analizi	Destek Vektör Makineleri, Rastgele Ormanlar, Lojistik Regresyon, Sinir Ağları, Lineer Diskriminant Analizi
Fantulin, Lagravinese ve Resce (2021)	Finansal risk analizi	Gradyan Artırma Algoritması, En Küçük Mutlak Küçültme ve Seçim Operatörü, Rastgele Ormanlar, Sinir Ağları

3. MATERYAL VE METODOLOJİ

Bu bölümde, çalışmada kullanılacak olan materyaller ve makine öğrenmesi yöntemleri incelenecek ve değerlendirilecektir.

3.1. Veri

Türkiye'de 21 elektrik dağıtım bölgesi mevcuttur. Bu çalışmada üç elektrik dağıtım firmasından elde edilen veriler kullanılarak tanımlanan model sonucunda müşterilerin borçlarını ödeme/ödememe durumunun tahmin edilmesi amaçlanmaktadır. Ödeme durumu tahminine ilişkin olarak açık borcu olan müşterilerin verisi kullanılmıştır.

Elektrik dağıtım sektöründe müşteriler; kaçak elektrik kullanan tüketicilerdir. Bu müşterilerin ödeme potansiyeli düşük olup, kaçak kullanma eğilimi yüksektir (Özay ve Çakıt, 2022). Faturalarını ödeme durumunu etkileyen çeşitli parametreler mevcuttur. Bu parametreler modelin öğrenmesinde kullanılmak üzere bağımsız değişken olarak tanımlanmış ve Çizelge 3.1.'de gösterilmektedir. SAP veri tabanından çekilen veri setinin modelin işleyebileceği formata çevrilmesi için veri ön işleme adımları izlenmiştir. Model sonucunda elde edilecek olan bağımlı değişken 1 = Öder, 0 = Ödemez şeklinde ikili kategorik değişken olarak belirlenmiştir.

Çizelge 3.1. Bağımsız değişken adı, tanımı ve tipi

Değişken Adı	Değişken Tanımı	Değişken Tipi
Güncel durum borç	Müşterinin ödenmemiş toplam borç miktarı	Nominal
İcradan tahsilat	Müşteriden daha önceden icra yolu ile tahsil edilen toplam borç miktarı	Nominal
İcra öncesi tahsilat	Müşteriden icra öncesi tahsil edilen toplam borç miktarı	Nominal
İlçe kaçak oranı	Müşterinin bağlı olduğu ilçenin kaçak oranı	Nominal
Avukat performansı	Müşteri borcunun atandığı avukat performansı	Nominal

Çizelge 3.1. (devam) Bağımsız değişken adı, tanımı ve tipi

Kaçak sayısı	Müşterinin toplam kaçak kullanım sayısı	Nominal
Abone grubu	Müşteri abone grubu tipi	Kategorik
TCKN/VKN	Türkiye Cumhuriyeti kimlik ve vergi numarası	Kategorik
Adres	Adres bilgisi	Metin
Abone	Abonelik bilgisi	Kategorik
Kesik dönem	Müşterinin enerjisinin kesik olma durumu	Kategorik

3.1.1. Bağımsız değişkenlerin tanımlanması

Güncel durum borç

Dağıtım şirketi müşterilerinin ödenmemiş tüm kaçak elektrik faturala tutarlarının toplamıdır. Yapılan ödemeler güncel borç değerinden düşürülerek sürekli revize olmaktadır.

İcradan tahsilat

Kaçak elektrik faturalarının ortalama 10- 15 gün vadesi bulunmaktadır. Müşteriler kaçak elektrik faturalarını vadesinde ödemediği durumda dağıtım şirketleri yasal takip başlatarak icra kanalı ile tahsilat yoluna gidebilir. Ayrıca dağıtım şirketleri icradan tahsil edilen borçlar için ekstra kazanç elde etmektedir. Yasal takip işlemleri sözleşmeli avukatlar aracılığı ile yapılmakta olup, avukatların yeteneklerine bağlıdır. Bu sebeplerle icra takibinin daha etkin yapılabilmesi amaçlanmaktadır. İlgili müşterinin icra kanalı ile tahsil edilen geçmiş borç toplamı icradan tahsilat değişkeni ile ifade edilmektedir.

İcra öncesi tahsilat

Kaçak elektrik faturaları peşin veya taksitli olarak ödenebilmektedir. Müşteri elektrik dağıtım şirketine faturası ile ilgili taksit talebinde bulunduğunda fatura tutarına oranla

belirli sayıda taksit yapılmaktadır. Taksit prosedürü gereği borçların aylık ödemeli yeni vade tarihleri oluşmaktadır. Peşin veya taksitli ödemelerde vade tarihi gelmeden yapılan ödemelerin toplamı icra öncesi tahsilat olarak ifade edilmektedir.

İlçe kaçak oranı

Türkiye'de kaçak elektrik kullanım oranları bölgelere göre farklılık göstermektedir. İlgili çalışmada da 3 farklı elektrik dağıtım firmasının/bölgenin verileri kullanılmaktadır. Bu 3 elektrik dağıtım firması toplamda 14 ilde faaliyet göstermektedir. Farklı bölgelerin kaçak kullanımının ödeme üzerindeki etkisi modele yansıtılmak istenilmiştir. Bu sebeple müşterilerin elektrik tüketim noktalarının bulundukları ilçenin kaçak oranı alınmıştır. Aynı müşterinin farklı ilçelerdeki tüketimden kaynaklı borçları varsa ağırlıklı ortalama ile ilçe kaçak oranı değeri hesaplanmıştır.

Avukat performansı

Ödenmemiş kaçak borçları yasal takip ile tahsil edilebilmektedir. Yasal takip sürecinde ilgili 3 bölgenin hukuk müşavirlikleri; sözleşmeli avukatlara borçları atamakta ve bundan sonraki süreci sözleşmeli avukatlar yönetmektedir. Bir sözleşmeli avukata birden fazla müşterinin faturası atanabilirken, bir müşterinin farklı elektrik faturası borçları farklı avukatlara atanmış olabilir. Burada n-n ilişki mevcuttur. Bir müşteriden A avukatı tahsilat yaparken, B avukatının tahsilat yapamadığı durumlar görülmektedir. Avukat performansı insani bir kriter olup model üzerindeki etkisi analiz edilmek istenmektedir. Ayrıca model sonuçları avukat atama sürecinin yeniden şekillenmesinde de kullanılabilecektir.

Sözleşmeli avukat performansları yıllık olarak tahsil ettikleri dosyalar baz alınarak hesaplanmaktadır. Veri setinde müşteri bazında avukat performansı belirlenirken sözleşmeli avukatların önceki yılki performans oranlarının, borç tutarına göre ağırlıklı ortalaması alınarak hesaplama yapılmıştır.

Kaçak sayısı

Kaçak elektrik kullanımı yapan müşteriler genellikle birden fazla kaçak kullanma eğilimde olmaktadırlar. Müşterilerin kaçak kullanımı arttıkça, ödeme oranı da düşmektedir. Kaçak

sayısı belirlenirken müşterilerin tüketim noktası bağımsız kaçak elektrik kullanım sayıları toplamı alınmıştır. Özellikle kurumsal müşterilerin farklı illerde kaçak kullanımları ile karşılaşılmakta olup, tüm kaçak kullanım durumları dikkate alınmıştır.

Abone grubu

Elektrik sektöründe müşterilerin elektrik kullanımları tarife tipine göre fiyatlandırılmaktadır. Mesken, Ticarethane, Sanayi, Tarımsal Sulama ve Aydınlatma olmak üzere 5 temel tarife tipi mevcuttur. Kaçak kullanımında abone olan müşteriler için abonelik tarifesi, abone olmayan müşteriler için ise Ticarethane tarifesinden faturalandırma yapılmaktadır. Tüketim miktarı ve vergi mükellefi olma durumuna göre abone grupları değerlendirildiğinde mesken tarifesi diğerlerinden ayrışmaktadır. Bu ayrımın etkisini model üzerinde analiz edebilmek için tarife tipi veri setine abone grubu olarak eklenmiştir.

TCKN/VKN

Elektrik dağıtım şirketi bölgesindeki tüm tüketicilere enerji sağlamakla yükümlüdür. Abonelik olmadan elektrik kullanımı da bir kaçak elektrik kullanımıdır ve abonelik olmadığı durumlarda tüketici bilgileri veri tabanında eksik olabilmektedir. Yasal takip sürecinin başlaması için TCKN/VKN bilgisinin bilinmesi gerekmektedir. Bu gibi durumların ödeme üzerindeki etkisini analiz edebilmek için TCKN/VKN var/yok bilgisi veri setine eklenmiştir.

Adres

Veri tabanında tüketici adres verileri dolu olmayabilir veya güncel olmayabilir. Veri seti kullanılan dağıtım şirketlerinde %80 oranında faturalar yerinde oluşturulup, müşteriye teslim edilmektedir. Fakat yine de kaçak elektrik faturasının veya icra tebligatının posta ile yapılması gerektiği durumlarda adres bilgisine ihtiyaç duyulmaktadır. Adres verisi kapı numarasına kadar dolu olduğu durumda yeterli adres verisi var denilebilmektedir. Aksi takdirde yok sayılmaktadır.

Abone

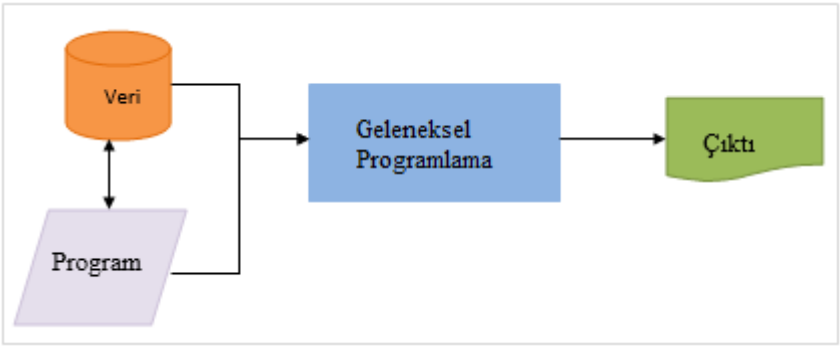
Daha önce bahsedildiği gibi kaçak elektrik kullanımı yapan tüketiciler abonelik yaptırmamış olabilir. Abone olmayan müşterilerin, abone olanlara göre faturalarını ödeme eğilimleri düşük olabilir. Bu durumun ödeme üzerindeki etkisini modele yansıtabilmek için abonelik bilgisi olan müşteriler için var, olmayanlar için yok girilmiştir.

Kesik dönem

Elektrik tüketim noktalarında enerjinin verilebilmesi için yerine getirilmesi gereken yasal mevzuat kuralları mevcuttur. Müşteri abonelik yaptırmadığında, borcunu ödemediğinde veya kaçak elektrik kullanımı yaptığında enerjisi kesilerek sayacı mühürlenmektedir. Kesme işlemi ilgili müşteri için ne kadar çoksa mevzuata uygun olmayan durumu da o kadar fazla denilebilir. İlgili müşterinin son kaçak kullanımı dışında; daha önceki dönemlerde enerjisi kesilmiş mi diye kontrol edilerek kesildi ise evet, kesilmedi ise hayır işaretlenmiştir.

3.2. Geleneksel Programlama

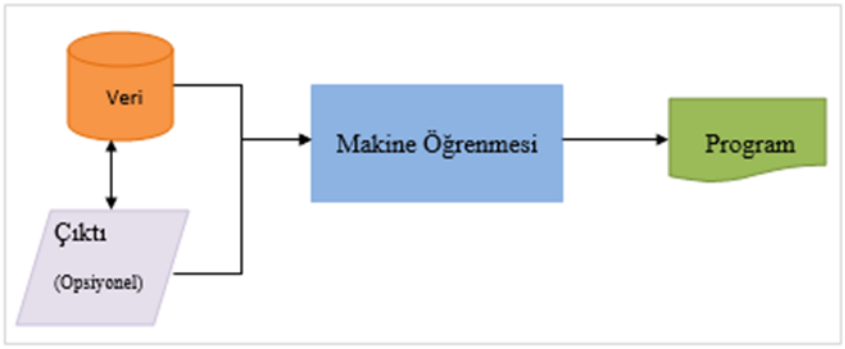
Geleneksel programlama, kullanıcı veya programcının istenen sonuçları elde edebilmek için kodlamaları kullanarak verilere hesap yaptıran bir dizi talimat veya işlemlerin oluşturulmasını içerir (Sarkar ve diğerleri, 2018: 5). Geleneksel programlama şeması Şekil 3.1'de görüldüğü gibi veriler, program ile işlenerek çıktılar elde edilir.



Şekil 3.1. Geleneksel programlama şeması

3.3. Makine Öğrenmesi

Makine öğrenmesi, Şekil 3.2'de gösterildiği gibi verileri ve varsa beklenen çıktıları dikkate alarak, model olarak da bilinen programı oluşturur. Bu program veya model gelecekte gerekli kararları vermek ve yeni girdilerden beklenen çıktıları elde etmek için kullanılabilir (Sarkar ve diğerleri, 2018: 7).



Şekil 3.2. Makine öğrenmesi programlama şeması

1959 yılında Arthur Samuel tarafından makine öğrenimi terimi "bilgisayarlara açık bir şekilde programlanmadan öğrenme yeteneği veren bir çalışma alanı" şeklinde tanımlanmıştır. Makineler tarafından verilerin daha etkili şekilde kullanımının sağlanmasının öğrenilmesine makine öğrenmesi denilmektedir. Veriler elde edildikten sonra, verilerin içerdiği bilgileri yorumlamak her zaman mümkün olmayabilir. Mevcut veri kümelerindeki artışla birlikte makine öğrenmesi kullanımı talebi de artmaktadır. Birçok matematikçi ve programcı, çok büyük veri kümelerine sahip olan bu problemin çözümünü bulmak için çeşitli yaklaşımlar uygular. Makine öğrenmesinin hedefi verinin içerdiği yapıyı öğrenmektir. Açıkça hesaplamaları içeren programı yazmadan kendi kendine öğrenmeyi sağlamak için yapılan çalışmalar artmaktadır (Batta, 2020).

Makine öğrenmesinin amacı, sistemlerin etraflarında ortaya çıkan değişikliklere uyum sağlamak için ampirik verileri, tecrübeyi ve eğitimi kullanan algoritmalar tasarlamak ve geliştirmektir. Makine öğrenmesinin asıl hedefi, analiz ettiği eğitim verisi içerisindeki kural ve kalıpları içeren modelleri otomatik olarak türetmektir (Hu ve Hao, 2012: 3).

Makine öğrenmesi algoritması, çıktıya ulaşmak için girdi verisini kullanarak programlanmadan istenen sonucu elde edecek yapıyı hesaplamaktadır.

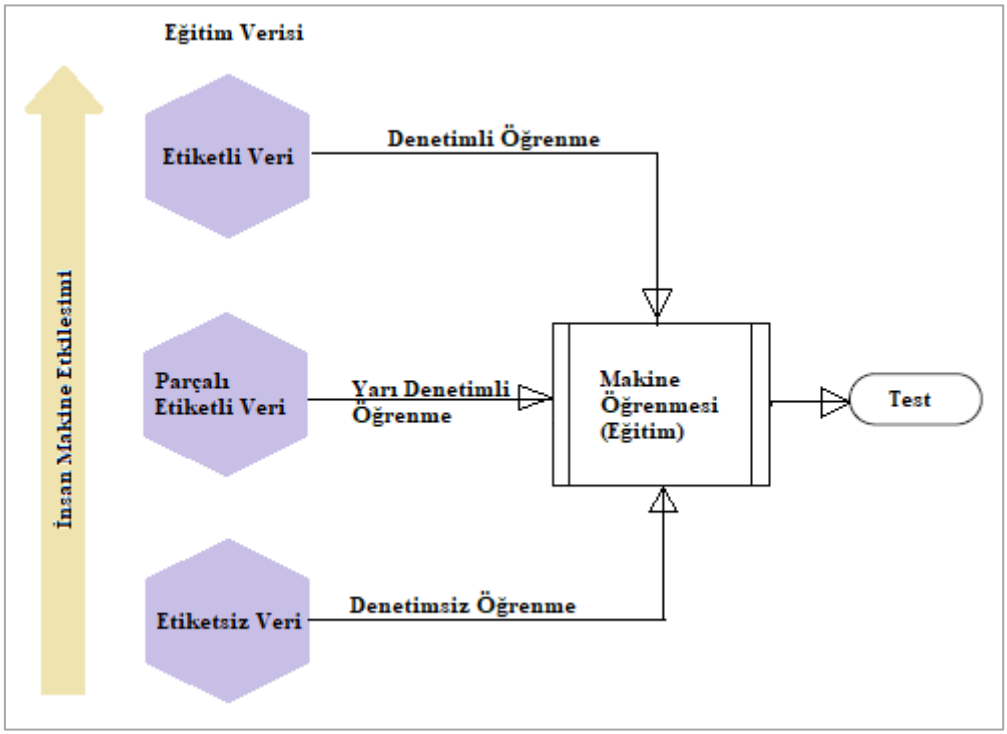
Makine öğrenmesi teknolojisi, istatistik, olasılık, yapay zekâ, psikoloji, nörobiyoloji ve diğer birçok disiplinde sunulan geniş bir alanı temsil eder. Problemlerin çözümü için seçilen bir veri kümesinin modeli oluşturulur. Makine öğrenmesi, öğrenme süreçlerine ilişkin temel istatistiksel hesaplama teorileri kullanarak bilgisayarlara insan beynini taklit etmeyi öğretmekten ileri bir alan haline gelmektedir (Nasteski, 2017).

Makine öğrenmesi algoritmalarında kullanılan verilerdeki her bir gözlem aynı değişkenler ile ifade edilir. Değişken tipleri kategorik veya sürekli olabilir (Kotsiantis, 2007).

3.3.1. Öğrenme yöntemleri

Öğrenme, mevcut veriler kullanılarak girdi ve çıktı arasındaki bağımlılıkların tespit edilmesi şeklinde ifade edilebilir (Cherkassky ve Mulier, 2001: 12).

Makine öğrenmesi, veri etiketlemelerine göre Şekil 3.3' de gösterildiği gibi denetimli, yarı denetimli ve denetimsiz öğrenme olarak ayrılabilir (Naqa ve Murphy, 2015).



Şekil 3.3. Makine öğrenmesi öğrenme çeşitleri (Naqa ve Murphy, 2015)

Denetimli öğrenme

Etiketlenmiş girdi değerlerini kullanarak istenen çıktıların elde edilmesini sağlayan model ortaya koyar. x_i'nin bir giriş değeri, y_i'nin onunla ilişkili doğru etiketlenmiş çıkış değeri olduğu; x_i, y_i biçimindeki giriş/çıkış değerleri ile temsil edilir. Amaç, girdiden çıktıya haritalamayı çıkarmaktadır. Makine öğrenmesinin, görülen her i için girdi/çıktı çiftlerini ($f(x_i) = y_i$) olarak öğrenmesi beklenmektedir. f fonksiyonu, sonuç değerleri kesikli ise sınıflandırma, sonuç değerleri sürekli ise regresyon fonksiyonu olarak ifade edilir. Sınıflandırma/regresyon fonksiyonunun, daha önce görmediği veri kümesindeki girdi değerleri için çıktıları doğru bir şekilde tahmin etmesi amaçlanmaktadır (Hu ve Hao, 2012: 5).

Denetimli öğrenme yöntemleri, sınıflandırma ve regresyon olarak iki ana sınıftan oluşur.

Sınıflandırma

Denetimli makine öğrenmesinin bir alt alanı olan sınıflandırma yöntemlerinin temel amacı, modelin eğitim aşamasında öğrendiği girdi verileri için kategorik çıktı etiketlerini veya sonuçlarını tahmin etmektir. Çıktı etiketleri sınıflar veya sınıf etiketleri olarak da adlandırılır. Sınıflandırma yöntemlerinde etiket değerleri kategorik ve ayrık değerlerdir. Bu sayede, çıktılar belirli bir kategori veya sınıfa dahil olur.

Popüler sınıflandırma algoritmaları, lojistik regresyon, destek vektör makineleri, yapay sinir ağları, rastgele ormanlar ve gradyan artırma makineleri, k-en yakın komşular, karar ağaçları ve daha fazlasını içerir.

Bu çalışmada çıktı değerleri ikili kategorik veriler ile ifade edildiği için denetimli öğrenme sınıflandırma algoritmaları kullanılmıştır.

Regresyon

Regresyon yöntemlerinde modeller sürekli sayısal değerler içeren veri örnekleri üzerinden eğitilirler. Çıktı değerleri de yine sürekli sayısal sonuçlar içerir.

Denetimsiz öğrenme

Etiketlenmemiş girdi değerleri kullanılır. Amaç, yeni veri örnekleri için çıktı değerini belirlerken verilerdeki doğal kalıpları kullanmaktır. Denetimsiz öğrenme varsayımı, girdi uzayında belirli kalıpların diğerlerinden daha sık meydana geldiği bir yapının olduğu ve genel olarak bu yapıların çıktı olarak elde edilmek istenmesidir. Bu istatistikteki yoğunluk tahminine karşılık gelir (Hu ve Hao, 2012: 17).

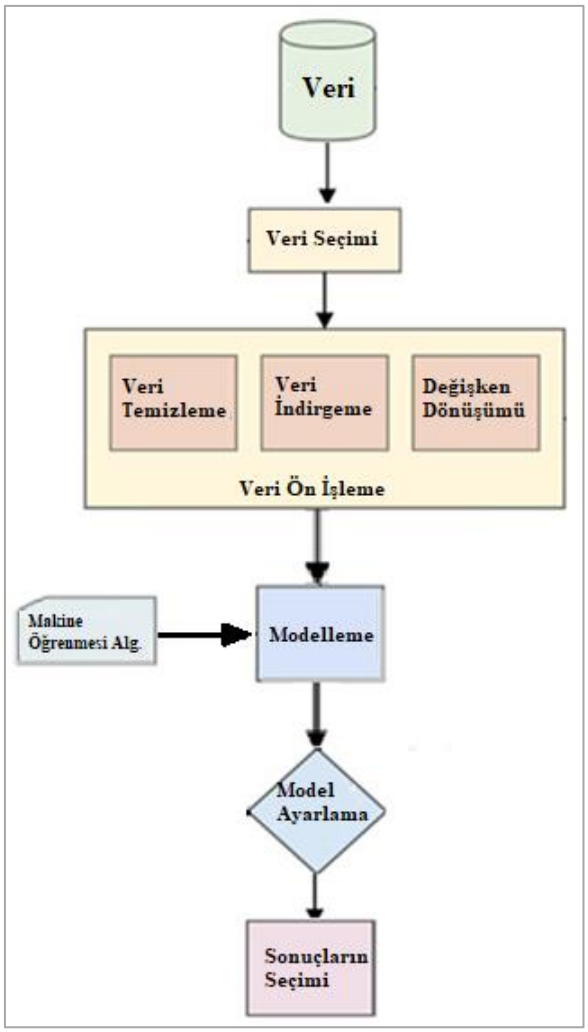
Denetimsiz öğrenme yöntemleri aşağıdaki gibi sıralanabilir.

- Kümeleme
- Boyutsal küçülme
- Anomali tespiti
- Birliktelik kuralı madenciliği

Yarı denetimli öğrenme

Eğitim için hem etiketli hem de etiketsiz veriler kullanır. Etiketli veriler eğitim veri kümesinin düşük bir kısmını oluşturur. Amaç etiketli ve etiketsiz verilerin birleştirilmesinin öğrenme davranışına olan etkisini analiz etmek ve bu birleşimden yararlanan algoritmalar geliştirmektir. Etiketli verileri elde etmek maliyetli ve uzmanlık gerektirdiği için yarı denetimli öğrenme, denetimli öğrenmeye göre daha umut verici bir yaklaşımdır (Hu ve Hao, 2012: 24).

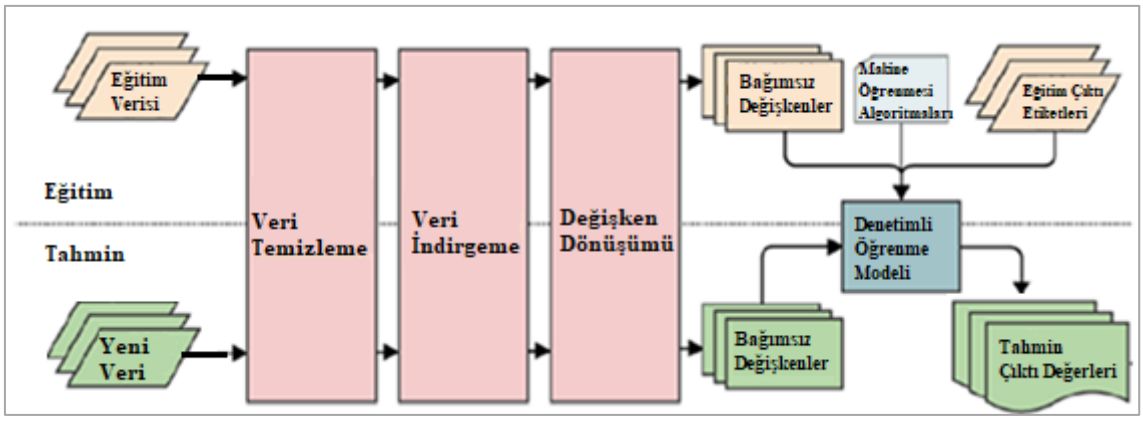
Temel makine öğrenmesi adımları Şekil 3.4'te ifade edilmiştir. Akış genel olarak; veri seçimi, veri ön işleme, model seçimi ve modelleme, model hiper parametrelerini ayarlama ve sonuçların izlenerek model performanslarının karşılaştırılması şeklinde ilerlemektedir.



Şekil 3.4. Makine öğrenmesi yol haritası (Sarkar ve diğerleri, 2018: 53)

Temel bir denetimli makine öğrenmesi modelinde öğrenme süreci eğitim ve test olarak iki adıma ayrılır: Eğitim süreci, eğitim veri kümesi baz alınarak öğrenme algoritması veya öğrenen tarafından özelliklerin öğrenildiği ve öğrenme modelinin oluşturulduğu aşamadır. Test sürecinde, test veri kümesi üzerinde öğrenme modeli kullanılarak tahminleme yapılır (Nasteski, 2017).

Bu çalışmada kullanılacak olan materyal ve literatürdeki uygulamalar incelendiğinde veri setinde kategorik değişkenlerin yer alması ve çıktı değişkeninin ikili kategorik değişken olarak ifade edilmesi sebebiyle denetimli öğrenme yöntemlerinden sınıflandırma modellerinin kullanılmasının uygun olduğu görülmektedir. Şekil 3.5'te denetimli öğrenme yöntemi işlem adımları görülmektedir.



Şekil 3.5. Denetimli makine öğrenmesi işlem adımları

Bu bölümün devamında denetimli makine öğrenmesi algoritmaları incelenecektir.

3.3.2. Lojistik regresyon (LR)

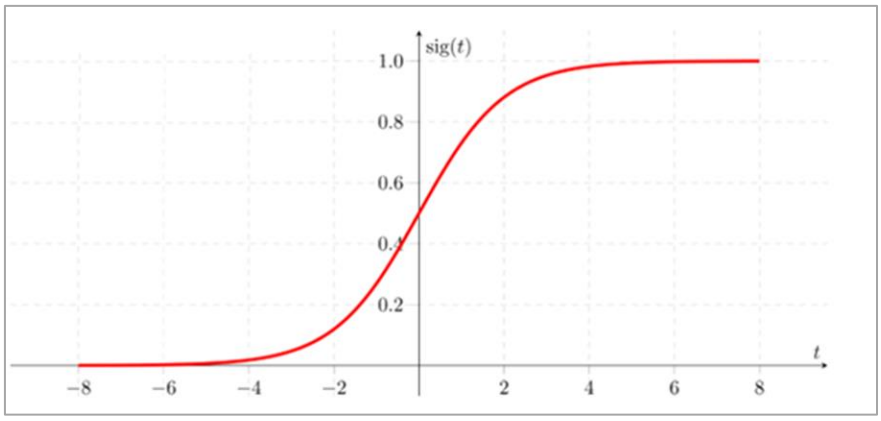
Lojistik Regresyon, belirli bir sınıfa ait olma olasılığını tahmin etmeyi sağlayan matematiksel bir modeldir (Ksiazek, Gandor ve Plawiak, 2021). Lojistik regresyon ayrıca ikili veya sıralı bir hedefin ilgili olayı bir veya birden fazla bağımsız değişkenin fonksiyonu olarak elde etme olasılığını tahmin etmeye çalışır (Y. Zhao ve Zhang, 2008).

Matematiksel olarak, lojistik regresyon Eş. 3.1'deki gibi tanımlanmış çoklu doğrusal regresyon fonksiyonunu tahmin eder:

$$\text{Logit}(p) = \ln\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad k = 1,2,..,n \qquad (3.1)$$

Bu çalışmada, ikili sınıflandırma modeli için LR kullanılmıştır. İkili değişkenleri tahmin etmede LR Sigmoid (Lojistik) Fonksiyonu kullanarak sınıflandırma yapar. Sigmoid fonksiyonu "S" şeklinde bir eğridir ve verileri 0-1 arasında sıkıştır.

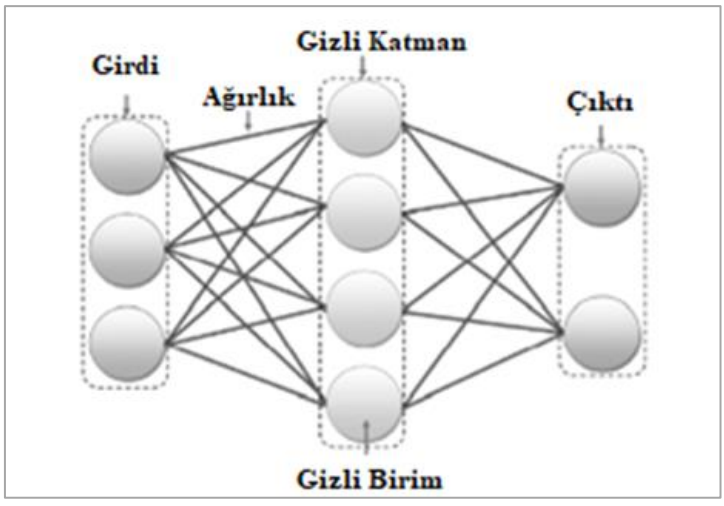
$$\text{sig}(t) = 1/(1+ e\text{^}(-t)\ ) \qquad (3.2)$$



Şekil 3.6. Sigmoid fonksiyon grafiği (Oprea ve Bâra, 2021)

3.3.3. Yapay sinir ağları (YSA)

Yapay sinir ağları insan veya hayvan gibi biyolojik merkezi sinir sisteminde yer alan sinir hücresi (nöronları) ağlarının karar verme sürecini kaba bir şekilde simüle etmeye çalışan hesaplamalı ağlardır (Graupe, 2013: 1). 1943 yılında W. McCulloch ve W. Pitts'in çalışmalarında yapay sinir ağının ilk modeli ortaya çıkmıştır. F. Rosenblatt, 1957 yılında insan beyninin işleyişini modellemek için eğitilen, tek katmanlı ve tek çıkışlı "Perceptron" modelini ortaya koymuştur (Yapraklı ve Erdal, 2016). En küçük yapay sinir ağı yapısı olan Perceptron ağırlık ile girdi değerlerini ölçekleyerek çıktı elde eder.



Şekil 3.7. Tek katmanlı YSA modeli (Hu ve Hao, 2012: 583)

Tek katmanlı YSA'da girdi doğrudan bir sonraki katmana iletilir. x_i girdi vektörünün, çıktı değeri bir transfer fonksiyonundan geçirilerek aşağıdaki eşitlik ile elde edilir.

$$y_i = f(\sum_{i=1}^{n} w_{ij} x_i + b) \quad (3.3)$$

y_i : çıktı vektörü

f: aktivasyon fonksiyonu

n: girdi miktarı

w_{ji}: i. elemandan j. elemanına olan ağırlık miktarı

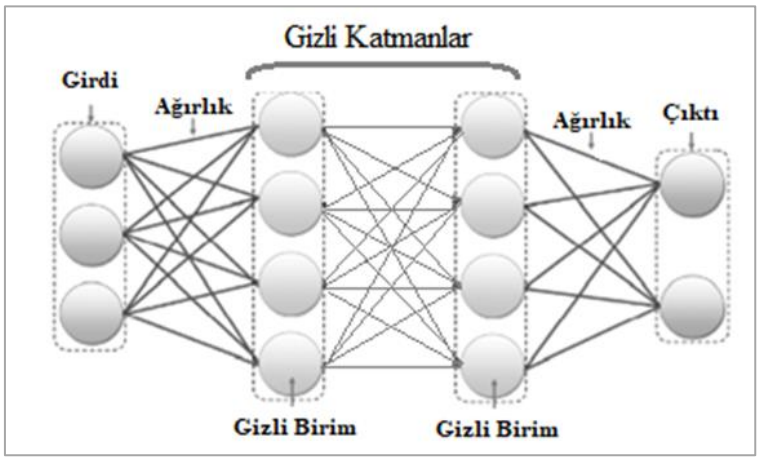
b: eşik değeri olarak belirlenir.

Tek katmanlı YSA sınırlı sayıda eşleştirme sağladığı için karmaşık problemlerde çok katmanlı YSA tercih edilmektedir.

Çok katmanlı YSA birden fazla nöronun birbirlerine bağlanması veya birden fazla gizli katmanın oluşturduğu yapılardır. Çok katmanlı YSA modelinde besleme yöntemine göre yöntemler:

- İleri besleme
- Geri besleme

Bu çalışmada ileri besleme yöntemi kullanılmış olup; ileri besleme yönteminde akış ileri doğru katmanlardan geçerek gerçekleştirilir.



Şekil 3.8. Çok katmanlı YSA modeli

İleri yönlü beslemede girdi verilerinin ağırlıkları ağdan çıkana kadar değişmez. Ağın giriş katmanındaki i. elemanın hesaplaması;

$$Ç_i^n = G_i \tag{3.4}$$

Ara/ gizli katmanlarda elemanlara gelen net girdi miktarı:

$$NET_j^i = \sum_{i=1}^{k} A_{ij}\, Ç_i^n \tag{3.5}$$

Eş. 3. A_{ij} 5'te i. giriş elemanının j. ara katman elemanı ile bağlantısının ağırlık değeri olarak ifade edilir. j. elemanının çıktısı Eş. 3.5'teki hesaplanma sonucu elde edilir (Kabalcı, 2014).

Aktivasyon fonksiyonu doğrusal veya doğrusal olmayan şeklinde ifade edilebilir. En çok kullanılan doğrusal olmayan fonksiyonlar hiperbolik tanjant veya sigmoid gibi fonksiyonlardır (Konakoğlu, 2020). Aktivasyon fonksiyonu olarak Eş. 3.6'da verilen sigmoid fonksiyonu kullanılmıştır.

$$Sigmoid(z) = \frac{1}{1+e^{-z}} \tag{3.6}$$

j. ara katman elemanı için aktivasyon fonksiyonu Sigmoid seçildiğinde hesaplama Eş. 3.7'deki gibi yapılır:

$$Ç = \frac{1}{1+ e^{-(NET_j^i + \beta_j^a)}} \tag{3.7}$$

β_j^a ara veya gizli katmanda yer alan j. elemana bağlı eşik değerli elemanın ağırlığıdır. Aktivasyon işlemi ile ileri doğru beslemeli ağ akışı tamamlanmıştır.

3.3.4. K-en yakın komşu algoritması (KEYK)

İstatistikte, k en yakın komşu algoritması (KEYK), ilk olarak 1951'de Evelyn Fix ve Joseph Hodges tarafından geliştirilen ve parametrik olmayan denetimli öğrenme

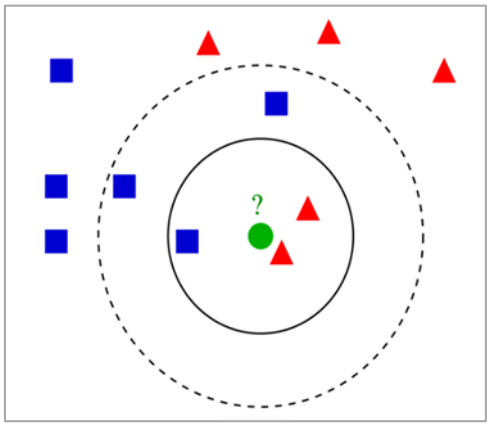
yöntemidir (Fix ve Hodges, 1989).

KEYK sınıflandırma algoritmasının çalışma yöntemi, test setindeki her bir değer ile eğitim setindeki değer arasındaki mesafeyi hesaplamak ve ardından elde edilen mesafe değerine göre test verisi kategorisinin k en yakın komşusu t'yi belirlemektir. Algoritmanın avantajları basit çalışma prensibi ve modeli etkileyen az sayıda faktör olmasıdır, ancak aynı zamanda çok fazla zaman harcaması, alan yükü ve k değerini seçmedeki zorluk gibi birçok eksikliği de vardır (H. Wang, Xu, ve J. Zhao, 2022). Bir anlamda, KEYK yöntemi k tarafından önyargılıdır (Guo, H. Wang, Bell, Bi ve Greer, 2003). En yakın komşu belirlenirken genelde Öklid uzaklık fonksiyonu kullanılmakla birlikte; Manhattan, Minkowski ve Hamming fonksiyonları da kullanılabilir. Bu çalışmada t sayıda değişkene bağlı bir uzaklık hesaplamak istendiği için Minkowski yöntemi kullanılır. Minkowski mesafesi Eş. 3.8'deki gibi belirlenir.

$$(\sum_{i=1}^{n} |x_i - y_i|^t)^{1/t} \tag{3.8}$$

$t = (x_1, x_2, x_3 \ldots, x_n)$, $Q = (y_1, y_2, y_3 \ldots, y_n) \in R_n$

Mesafe değeri kullanılarak Şekil 3.9'da gösterildiği gibi test verisinin hangi sınıfa ait olduğu belirlenmeye çalışılır.



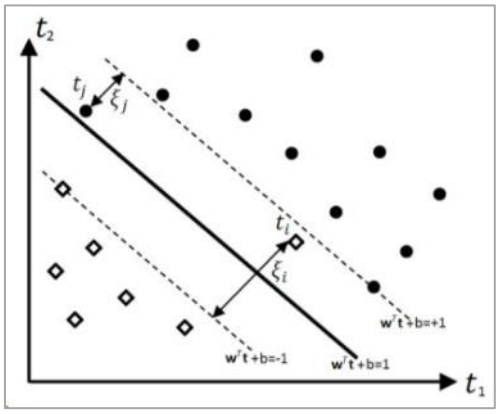
Şekil 3.9. K en yakın komşu algoritması gösterimi (Altunkaynak, Başakın ve Kartal, 2020)

3.3.5. Destek vektör makineleri (DVM)

Destek vektör makinesi, Vapnik ile meslektaşlarının Bell laboratuvarlarında geliştirdiği bir ikili sınıflandırma yöntemidir (Madzarov, Gjorgjevikj ve Chorbev, 2009). DVM sınıflandırma veya regresyon için kullanılan küçük ve orta ölçekli problemler için uygun denetimli öğrenme modelidir.

DVM, girdi verilerini yüksek boyutlu bir alana dönüştürerek sınıflar arasında doğrusal olmayan sınırlar uygulayan bir algoritmadır. Bu eşleme, girdi veri setini doğrusal olarak yeni bir uzaya ayrılabilir yapan çekirdek fonksiyonlarının bir görevidir. Yeni alanda, DVM, çıktı sınıfları arasında maksimum ayrım sağlayan bir hiper düzlemi oluşturur. Maksimal kenar hiper düzlemine en yakın olan eğitim gözlemlerine destek vektörleri denir. Standart DVM sınıflandırıcısı yalnızca ikili tip sınıflandırma problemlerine uygulanabilir (Golbayani, Florescu ve Chatterjee, 2020).

İkili sınıflandırma probleminde, DVM için model Eş. 3.9 – Eş. 3.11 ile ifade edilmiştir.



Şekil 3.10. Destek vektör makineleri algoritması modeli (Karagül, 2014)

$$\min_{w,b} \quad \frac{1}{2} w^t w + C \sum_{i=1}^{N} \zeta_i \tag{3.9}$$

$$y_i(w^t t_i + b) \geq 1 - \zeta_i \, , i = 1, \ldots, N \tag{3.10}$$

$$\zeta_i \geq 0, \; i = 1, \ldots, N \tag{3.11}$$

b ve w noktalarının hiper düzleme olan uzaklıkları hesaplanır. Lagrange değişkenin alabileceği en yüksek değer C ceza değişkeni olarak ifade edilir. DVM'leri için aşağıdaki denklemler kullanılarak ikili model belirlenir (Karagül, 2014).

$$\min_{\alpha} \ \mathcal{L}(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j t_i^T t_j - \sum_{i=1}^{N} \alpha_i \quad (3.12)$$

$$\sum_{i=1}^{N} \alpha_i y_i, \ i = 1,.., \text{N} \quad (3.13)$$

$$0 \leq \alpha_i \leq \text{C}, \text{ i= 1,..,N} \quad (3.14)$$

Burada, girdi ti, çıktı yi değişkenleri ile ifade edilir. Lagrange değişkeni α ile gösterilir. Giriş bölgesini, nitelik alanı ile çekirdek işlevi eşlemektedir. Destek vektör makinelerinde Polinom, Gauss, Sigmoid, Lineer ve Radyal Temel Fonksiyonu en çok tercih edilen fonksiyonlardır (Karagül, 2014). İkili uzayda bulunan sınıflandırma modeli Eş. 3.15 ile hesaplanmaktadır.

$$\hat{y} = \sum_{i \epsilon SV}^{\#SV} a_i y_i K(t, t_i), \ i = 1, .., \#SV \quad (3.15)$$

Burada, Kij çekirdek matrisidir. S destek vektör makineleri kümesini #SV sayısını, ŷ ise sınıflandırma tahmin sonucunu gösterir (Karagül, 2014).

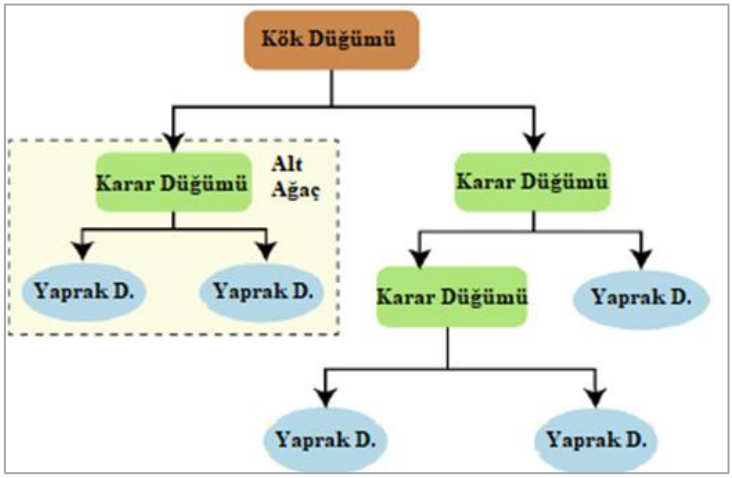
3.3.6. Karar ağaçları (KA)

Karar ağacı, ilgilendiğimiz bir olayı veya değişkeni sınıflandırmak veya tahmin etmek için kuralları tanımlamaya yardımcı olan ağaç tabanlı algoritmaların temelini oluşturur. Ayrıca, regresyon veya sınıflandırma için optimize edilmiş doğrusal veya lojistik regresyondan farklı olarak, karar ağaçları, her ikisini de gerçekleştirir (Ayyadevara, 2018: 69).

Karar ağacı kök, düğüm ve yapraklardan oluşur. Kök düğüm, tüm düğümlerin ebeveynidir ve adından da anlaşılacağı gibi, ağaçtaki en üstteki düğümdür. Karar ağacında her düğüm bir özelliği, her bağlantı bir kararı veya kuralı ve her yaprak kategorik veya sürekli olabilen bir sonucu göstermektedir (Patel ve Prajapati, 2018). Karar ağacı genellikle soldan sağa doğru çizilerek, kökten başlayıp aşağı doğru gider. "Yaprak" düğüm zincirin

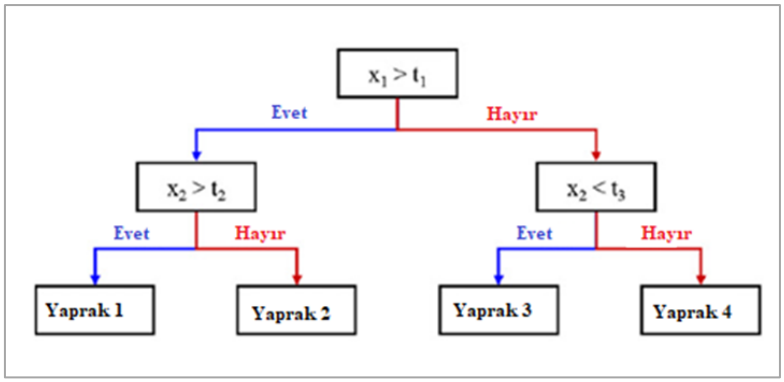
bittiği düğümdür. Kök düğümden yaprak düğüme kadar tüm düğümlerden iki veya daha fazla dal genişletilebilir. Düğüm özellikleri, dallar ise değerleri temsil etmektedir. Özelliğin değer kümesi için değer aralıkları bölüm noktası olarak görev yapar (Ali, Khan, Ahmad ve Maqsood, 2012).

Karar Ağaçları, denetimli sınıflandırma modelleridir (Y. Zhao ve Zhang, 2008). Şekil 3.11 karar ağacı yapısını göstermektedir.



Şekil 3.11. Karar ağacı kök ve düğüm yapısı (Charbuty ve Abdulazeez, 2021)

Karar ağacı öğrenme algoritmasının amacı, her bir tek boyutlu alt uzaya (yani, bir seferde tek özellik) odaklanmak, en iyi özelliği (en iyi tek boyutlu alt uzay) ve en iyi bölme konumunu seçmek ve en iyi bölmede özellik değerini çıkarmaktır (Suthaharan, 2016). Karar ağaçlarında her karar düğümde bir karar verilir ve modeli iyileştirmeyen değere sahip dallar budanır. Şekil 3.2'de karar ağacının karar yapısı yer almaktadır.



Şekil 3.12. Karar ağacı karar yapısı gösterimi (J. Kim ve Kang, 2016)

Karar ağaçları kullanıcıların karar vermesine yardımcı olan sezgisel çıktı ve görselleştirme aracıdır. Karar ağaçları sınıflandırma durumlarında aykırı değerlere tipik bir regresyon tekniğinden daha az duyarlıdır.

3.3.7. Topluluk yöntemleri

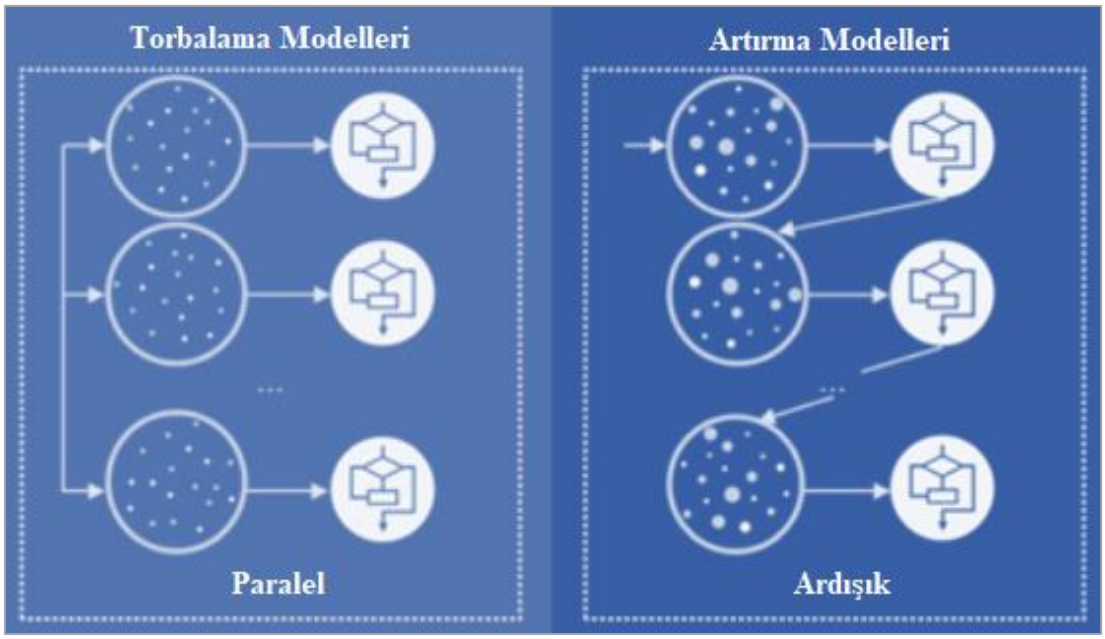
Sınıflandırma algoritmalarının dezavantajlarını gidermek için oluşturulan topluluk yöntemleri model performanslarını daha iyi hale getirmektedir. Birden fazla model birlikte kullanılarak daha başarılı, yeni ve birleşik bir model elde edilmektedir.

Sınıflandırma algoritmalarında hatayı belirleyen sınıflandırıcının doğruluğu ve eğitim sapması-varyansı şeklinde iki etmen bulunmaktadır. Topluluk yöntemleri bu iki bileşeni iyileştirmek için çalışır. Bu bileşenler arasında ters orantı bulunmaktadır. Düşük hatalı sınıflandırıcılar yüksek sapma eğilimi göstermektedir (Zhang ve Ma, 2012: 2).

Çoklu makine öğrenmesi kullanılarak girdi verilerinin eğitimi sağlandığında; gerçekleşen tahminler bir makine öğrenmesi algoritması sonucuna göre daha iyi performansa sahip olabilir. Topluluk yöntemleri algoritmaların birleştirilmesinden ziyade algoritmalardan elde edilen tahminlerin birleştirilerek ortak bir tahmin üretilmesi esasına dayanmaktadır (Z. H. Zhou, 2012: 15).

Ortalama almak sapmayı azaltmaktadır. Bu sebeple topluluk yöntemleri verileri gruplayarak ortalamalarını tespit edip, sapmaları azaltmayı hedeflemektedir (Zhang ve Ma, 2012: 2).

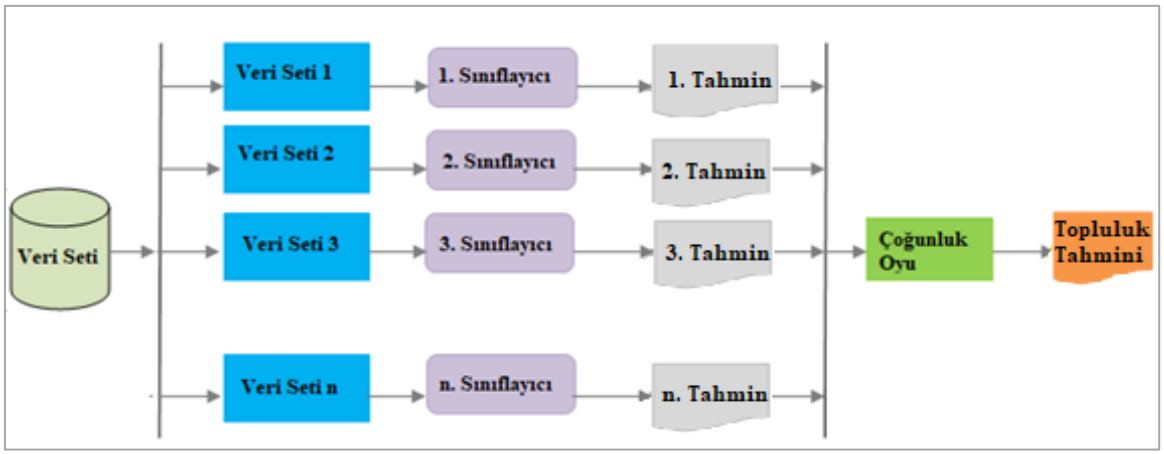
Maximum oylama, ortalama, ağırlıklı ortalama, yığma, harmanlama, rastgele alt uzaylar, torbalama, artırma gibi birleştirme teknikleri, örneklem seçimine ve işlem adımlarına göre farklı topluluk öğrenme metotları vardır. Bu çalışmada en çok kullanılan torbalama ve artırma yöntemleri ele alınacaktır. Torbalama ve artırma modellerinin çalışma mantıkları Şekil 3.13'te verilmiştir.



Şekil 3.13. Torbalama ve artırma modellerinin çalışma mantığının gösterimi (Terko, Zunic, Donko ve Dzelihodzic, 2019)

Torbalama (Bagging) yöntemi

Kısaltması Bootstrap Aggregating olan bagging (torbalama) topluluk öğrenme yöntemi 1996 yılında Breiman tarafından geliştirilmiştir. Torbalama yöntemi, farklı veri setleri üzerinden eğitilen öğrenme algoritmaları birleştirilerek sınıflandırıcı topluluk öğrenme modeli tasarlanmasıdır. Amaç doğruluğu artırmak için varyansı azaltmaktır (Breiman, 1996). Her sınıflandırıcı, eğitim kümesinden bir değiştirme ile alınan bir örnek örneği üzerinde eğitilir. Her örnek boyutu, orijinal eğitim kümesinin boyutuna eşittir. Veri seti üzerinden farklı eğitim setlerini oluşturmak için torbalama yöntemi kullanılır. Torbalama kullanılarak oluşturulan eğitim setleri ile öğrenen model çıktıları, çoğunluk oylaması ile birleştirilir (Onan, 2018). Böylece model varyansı düşürülerek, aşırı öğrenme önlenmiş olmaktadır. Şekil 3.14'te torbalama topluluk öğrenme yöntemi işlem adımları verilmiştir.



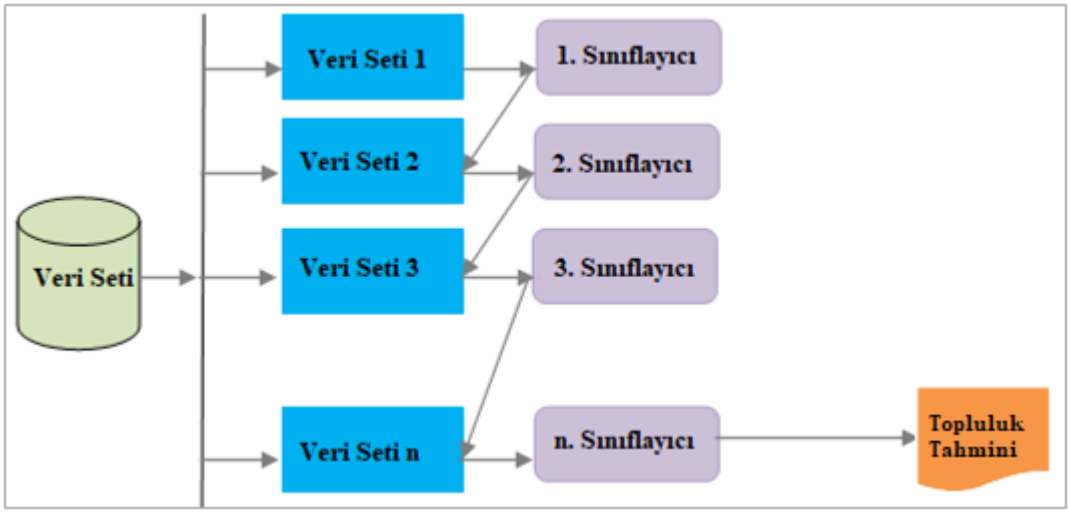
Şekil 3.14. Torbalama topluluk öğrenme yöntemi işlem adımları

En popüler torbalama yöntemi rastgele ormanlar algoritmasıdır.

Artırma yöntemi (Boosting)

Artırma yönteminin temelinde zayıf öğrenenleri iyileştirerek güçlü öğrenenler elde etmek yatmaktadır. Artırma yöntemi bir maliyet fonksiyonunun optimizasyon algoritması olarak yorumlanabilir (Breiman, 1996). Artırma yönteminde işlemler sıralı olarak uygulanmaktadır. Her işlem adımında bir önceki tahminde oluşan hatalara odaklanarak işleri hızlandırmaktadır (Z. H. Zhou, 2012: 24).

Artırma algoritması, zayıf öğrenme algoritmasını yeniden çağırır, her seferinde farklı bir alt eğitim kümesini modele öğretir. İşlemler sıralı olarak yapılır. Her iterasyonda algoritma yeni bir zayıf tahmin ortaya koyar, bu zayıf öğrenenlere en fazla ağırlık verilerek en zor örneğe odaklanılmak istenir. Tahminlerinin ağırlıklı ortalaması alınarak sonuçlar birleştirilir. Şekil 3.15'te artırma topluluk öğrenme yöntemi işlem adımları verilmiştir.



Şekil 3.15. Artırma topluluk öğrenme yöntemi işlem adımları

Bu çalışmada en popüler topluluk yöntemlerinden; rastgele ormanlar, gradyan artırma (GAM), aşırı gradyan artırma makineleri (AGAM), hafif gradyan artırma (HGAM) ele alınacaktır.

3.3.8. Rastgele ormanlar (RO)

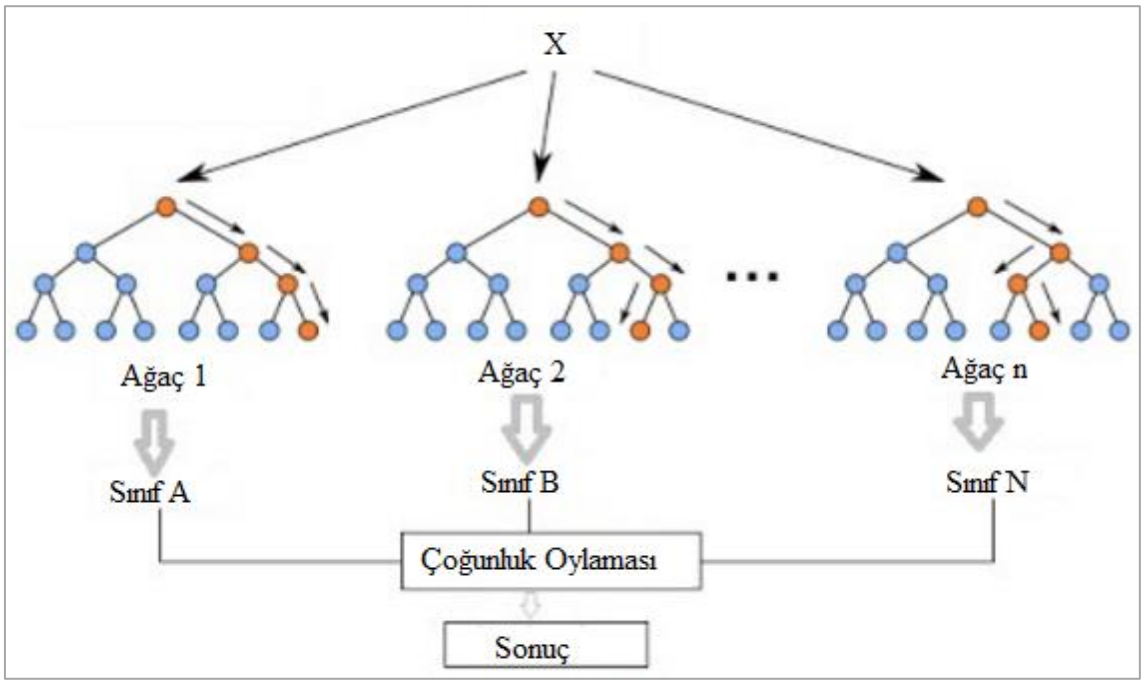
Rastgele ormanlar verilerin rastgele alt kümesini almakla beraber her düğümde rastgele bir X alt kümesini seçer ve yalnızca verilen X alt kümesi içindeki o düğümdeki en iyi bölünmeyi hesaplar. Bu yapı, ilişkisiz veya zayıf korelasyonlu tahminlemeyi sağlar. Ayrıca budama basamağı yoktur, bu da ormanın tüm ağaçlarının derinlere büyüdüğü anlamına gelir. RO'nin yalnızca iki hiper parametresi vardır. Bunlar: her düğümde yer alan rastgele alt kümedeki değişken sayısı ve ormandaki ağaçların sayısıdır. Ayrıca, RO, değişkenler arasındaki etkileşimi dikkate alırken, sınıflandırma doğruluğuna dayalı olarak bir değişkenin önemine göre değişkenleri sıralar (Ali ve diğerleri, 2012).

Rastgele ormanlar hem sınıflandırma hem de regresyon amaçlı kullanılabilen topluluk yaklaşımıdır. Topluluk yöntemleri bir grup zayıf öğrenenin güçlü öğrenen oluşturmak için birleştirilmesidir. Topluluk öğrenme algoritmaları birleştirilen algoritmaların herhangi birinden daha iyi bir performans ortaya koymaktadır. Rastgele ormanlar, ağaç tabanlı tahminleme algoritmalarının kombinasyonudur. Birbirinden bağımsız örneklemler içeren her dal rastgele bir vektöre bağlıdır. Rastgele ormanlar, ağacı oluşturmanın yanında nasıl oluşturulduğunu da gösterir. Karar ağaçlarında düğümler, tüm değişkenler arasındaki en

iyisi seçilerek bölünürken, rastgele ormanlar algoritmasında düğümler, o düğümde rastgele belirlenen tahmin edicilerin alt kümesi arasından en iyisi seçilerek bölünür (Liaw ve Wiener, 2002). Bu strateji diğer birçok sınıflandırma algoritması ile karşılaştırıldığında çok iyi performans gösterdiği ve aşırı öğrenmeye karşı başarılı olduğu görülmektedir (Ali ve diğerleri, 2012). Şekil 3.16'da rastgele ormanlar algoritması gösterim şeması gösterilmektedir.

Rastgele bir orman oluşturma algoritması adımları aşağıdaki şekildedir.

1. Orijinal veri kümesi karar ağacı tarafından alt kümelere ayrılır.
2. Karar ağacı oluşturulurken bağımsız değişkenler (özellikler) alt kümelere ayrılır.
3. Alt küme verilerine dayalı bir karar ağacı oluşturulur.
4. Oluşturulan model ile test veya doğrulama veri seti tahminlenir.
5. 1,2 ve 3. adımlar n ağaç sayısı kadar tekrarlanır.
6. Test veri kümesindeki son tahmin, tüm n tane ağacın tahminlerinin ortalamasıdır.



Şekil 3.16. Rastgele ormanlar algoritması gösterim şeması (Serengil, 2020)

Rastgele ormanlar yalnızca iki hiper parametrenin (düğümlerdeki rastgele alt kümelerin değişken sayısı ile rastgele ormanda yer alan ağaç sayısı) belirtilmesini gerektirir. Özellikle nihai model bir dizi karar ağacından oluşmaktadır ve bir insan için yorumlanması kolay

değildir. Yani rastgele orman, tanımlayıcı bir araç değil, tahmine dayalı bir modelleme aracıdır (Ali ve diğerleri, 2012).

3.3.9. Gradyan artırma makineleri (GAM)

Gradyan, model sonucunda elde edilen hatayı veya kalıntıyı ifade etmektedir. Gradyan artırma makineleri kademeli olarak hatayı iyileştirme yöntemidir.

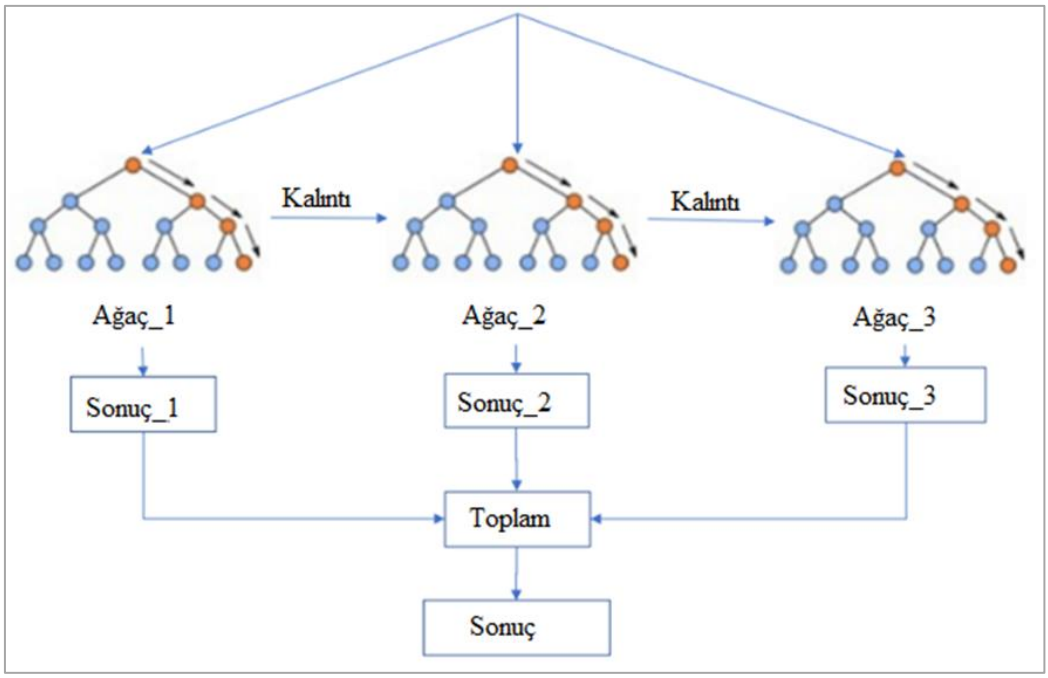
Tipik olarak, bir torbalama algoritmasında ağaçlar, her ağacın bir orijinal veri örneği üzerine inşa edildiği tüm ağaçlar arasında ortalama tahmin elde etmek için paralel olarak büyütülür. Gradyan artırma ise tahminleri elde etmek için paralel ağaç oluşturmak yerine, sıralı yaklaşım kullanılmaktadır. Böylece gradyan artırma yönteminde, tüm karar ağaçları bir önceki karar ağacının hatasını tahmin ederek, hatayı iyileştirir (Ayyadevara, 2018: 117).

Aşağıdaki adımlar izlenir;

1. Basit bir karar ağacı oluşturularak tahmin değeri belirlenir.
2. Gerçek değerden tahmin değeri çıkarılarak kalıntı değeri belirlenir.
3. Bağımsız değerlere bağlı kalıntıları tahmin eden yeni bir karar ağacı oluşturulur.
4. Orijinal tahmin yeni tahmin öğrenme oranı ile çarpılarak güncellenir.
5. Ağaç sayısı kadar yineleme için 2 ila 4 arasındaki adımlar tekrarlanır.

3.3.10. Aşırı gradyan artırma makineleri (AGAM)

AGAM T. Chen ve Guestrin, (2016) tarafından gradyan artırma modeli baz alınarak geliştirilmiştir. Temelinde karar ağacı yer alan daha yüksek performanslı ve hızlı bir algoritmadır. AGAM karar ağaçlarını paralel çalışmalarla oluşturduğu için çalışma süresi nispeten kısadır. Algoritma öncelikle ağacın derinliğini belirlemektedir. Sonrasında en fazla derinliğe ulaşınca geriye budama yapmaktadır. Yan ağaçların hiper parametrelerini artırarak modelin aşırı öğrenmesi engellenmektedir. Veri kümesindeki veri eksikliklerini tespit ederek boş verilerin yönetilmesini sağlamaktadır (Kuş, Bozkurt Keser ve Yolaçan, 2021). Şekil 3.17'de aşırı gradyan artırma makineleri algoritması şeması gösterilmektedir.

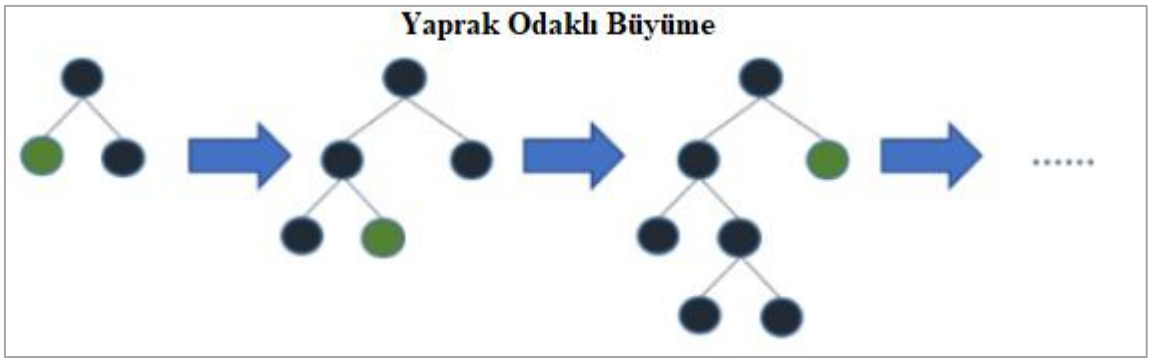


Şekil 3.17. Aşırı gradyan artırma makineleri örnek gösterimi (W. Wang, G. Chakraborty ve B. Chakraborty, 2020)

3.3.11. Hafif gradyan artırma makineleri (HGAM)

HGAM algoritması histogram tabanlıdır. HGAM algoritmasında karar ağaçlarının eğitim süresi, hesaplama ve bölünme sayısı ile doğru orantılıdır. Böylece kaynak kullanımı düşerken eğitim süresi de kısalmaktadır. Diğer artırma algoritmaları ağacı derinlik veya seviye bazında bölerken, HGAM algoritması karar ağacı algoritması temeline dayanmakta olup ağacı en uygun yaprağa göre bölmektedir. Bu nedenle, HGAM algoritması kaybı daha çok azaltmaktadır. Bunun yanında büyük boyutlu verilerle çalışabilen bir öğrenme algoritmasıdır (Kuş ve diğerleri, 2021).

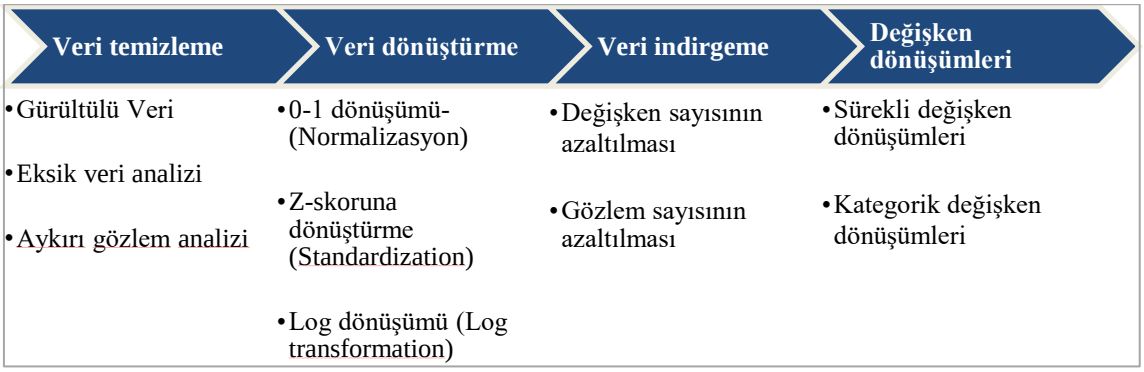
HGAM, bir sinir ağına benzer herhangi bir keyfi türevlenebilir kayıp işlevi ve gradyan iniş optimizasyon hedefine uyar. Özniteliklerin hedefle korelasyonu düşük olduğundan, HGAM'in diğer makine öğrenmesi sınıflandırma algoritmalarından daha iyi performans göstermesine izin veren kendi öznitelik seçim mekanizması vardır. Böylece HGAM, özellikleri otomatik olarak seçip, eğitim sürecini hızlandırır ve tahmin verimliliğini artırarak gradyan artırma algoritmasını genişletir (Oprea ve Bâra, 2021). Şekil 3.18'de hafif gradyan artırma makineleri makineleri büyüme yöntemi gösterilmektedir.



Şekil 3.18. Hafif gradyan artırma makineleri yaprak odaklı büyüme gösterimi (Cypriano, 2018)

3.4. Veri Ön İşleme

Tahminleme modelleri veri üzerinden öğrenen yapılardır. Veriler çözülmek istenen problemi karakterize eder. Bu sebeple veri ne kadar sağlıklı olursa model de problemi o kadar iyi yansıtacaktır. Denetimli öğrenme algoritmalarında model girdi ve çıktı değerlerini kullanarak öğrenmektedir. Model kurulmadan önce eksik, aykırı, yanlı, farklı ölçeklendirilmiş veya gürültülü verilerin ön işleme prosesleri ile giderilmesi gerekmektedir. Veri ön işleme adımları Şekil 3.19'da şematize edilmiştir.



Şekil 3.19. Veri ön işleme adımları gösterim şeması

3.4.1. Veri temizleme

Veri temizleme bozuk veya hatalı verileri tespit etme ve düzeltme (veya kaldırma) işlemidir; yani eksik değerlerin doldurulması, aykırı değerlerin tespit edilmesi ve yönetilmesi gibi işlemlerle veriler temizlenir (Kang ve Tian, 2018).

Gürültülü Veri

Verilerdeki uyumsuz veya olma ihtimali bulunmayan verilere gürültülü veri denir. Örneğin; veri setinde cinsiyet erkek iken hamilelik durumunun evet olması. Veriyi düzenlemek için; bozuk verileri düzeltme veya filtreleme/silme işlemi yapılabilir.

Eksik veri analizi

Eksik veri; veri seti içindeki boş olan değerlerdir. Eksik veriyi düzenlemek için;

1. Eksik değer içeren kayıt veya kayıtlar yok sayılabilir.
2. Boş olan veriler için değişken ortalaması kullanılabilir.
3. Sınıf bazında değişken ortalaması alınarak eksik verilerin yerine kullanılabilir.
4. Mevcut verilerden tahminleme yolu ile en uygun değer belirlenerek kullanılabilir. (Örnek: Regresyon, Navie Bayes, Hot-Deck yöntemleri ile)

Aykırı gözlem analizi

Aykırı gözlem; diğer verilerden oldukça farklı bir dağılım gösteren değerlerdir. Makine öğrenmesi yöntemlerinin çoğu verilerin belirli bir aralığa sıkıştırıldığı ve aykırı gözlemlerin olmadığı durumda daha iyi sonuçlar vermektedir. Aykırı gözlemleri tespit etmek için birçok yöntem mevcuttur.

İstatistiksel metodolojiler;

- Çeyrekler açıklığı
- Ortanca mutlak sapma yöntemi
- Hampel filtresi
- Z skoru

Grafiksel metodolojiler;

- Kutu grafiği,
- Histogram

Aykırı gözlemleri düzenlemek için 3 yöntem kullanılabilir.

- Silme:
  Tespit edilen aykırı gözlem veri setinden silinebilir.
- Ortalama ile doldurma:
  Tespit edilen aykırı gözlem değeri veri ortalamasına etki etmemek için ortalama değer ile doldurulabilir.
- Baskılama:
  Tespit edilen aykırı gözlem; gözlem değer aralığının alt ve üst sınırlarına çekilerek baskılanabilir. Alt ve üst sınır değer aralığı ortalamadan çok farklı olduğu veya aykırı gözlem etkisinin hissedilmek istendiği durumlarda baskılama yapılarak gözlem değerinin gerçek değerinden sapma oranı düşürülür.

3.4.2. Veri dönüştürme

Veri seti içindeki değerlerin kendi içinde varyans ve bilgi yapısını etkilemeden, değerleri belirli bir formatta değiştirerek standart hale getirme işlemidir. 3 farklı veri standardizasyon yöntemi mevcuttur.

- 0-1 dönüşümü- Normalizasyon:
  En yaygın normalizasyon yöntemi min-max normalleştirmesi olup değişken değerleri doğrusal dönüşüm ile 0-1 veri aralığına dönüştürülür.
- Z-skoruna dönüştürme (Standardization):
  Veri setindeki parametreler Z dönüşümü ile değişkenin ortalaması ve standart sapması ile bağlantılı olarak normalleştirilir.
- Log dönüşümü (Log transformation):
  Veri setindeki her x değeri log(x) ile belirli bir dönüşüm tabanında değiştirilebilir.

3.4.3. Veri indirgeme

Özellikle denetimli öğrenme modellerinde aşırı öğrenmenin önüne geçebilmek için veri indirgeme yapılabilir. Veri indirgeme yöntemleri;

- Değişken sayısının azaltılması

- Gözlem sayısının azaltılması

3.4.4. Değişken dönüşümleri

Makine öğrenmesi modelleri veri üzerinden anlamlı desenler çıkarmaya çalışırken sayısal veriler üzerinden işlemlerini gerçekleştirmektedir. Veri setinde yer alan değişkenlerde dönüşüm yaparak model geliştirilebilir.

- Sürekli değişkenlerde dönüşümler:
  Sayısal bir değerin yine sayısal bir değere dönüştürülmesidir. Değerlerin ölçeğini dönüştürmek veya çarpıklıklarını gidermek için yapılmalıdır.
- Kategorik değişkenlerde dönüşümler:
  Kategorik değişkenleri sıralayabilmek, karşılaştırabilmek ve işleyebilmek için sayısal değişkene dönüştürmektir. Bazı makine öğrenmesi modelleri için bu dönüşümü yapmak şarttır.

3.5. Hiper Parametre Ayarlama

Hiper parametreler model eğitilirken modelin denetlenmesine yarayan iyileştirilebilir değişkenlerdir. Modele müdahale edilmez ise hiper parametreler için varsayılan değerler atanmaktadır, bu varsayılan değerler veri setleri için her zaman en iyi sonuçların üretilmesine olanak sağlamayabilir. Bu sebeple hiper parametre ayarlaması yapılarak model sonuçlarının iyileştirilmesi amaçlanmaktadır.

Hiper parametre belirlemesi yaparken kullanılan yöntemlerden ilki ızgara arama yöntemidir. Girilen değerleri bir aralık olarak alarak alır ve aralıkta yer alan değerlerin tümü için modelin eğitilmesini sağlayarak en iyi sonuç veren değeri belirler.

İkinci yöntem rastgele arama yöntemidir. Bu yöntemde belirlenen aralıktaki değerler arasında rastgele değer seçimi yapılarak model eğitilir. Her zaman en iyi değerin belirlenmesini garanti etmez fakat aralıkta yer alan tüm değerleri denemek yerine seçimli deneme işlemi yaptığı için ızgara arama yöntemine göre hızlıdır.

Üçüncü yöntem bayes teoremi ile hesaplama yaparak hiper parametre belirlemesinin

yapılmasıdır. Bayes teoremi ile hiper parametre belirlenmesi de yaygın yöntemlerden biridir.

Modellemede kullanılan algoritmalar için ayarlanması gereken farklı hiper parametre değerleri mevcuttur ve bu durum her bir algoritma için ayrıca hesaplama yapılmasını gerektirmektedir.

3.6. Model Performans Ölçümü

Bu çalışmada denetimli öğrenme sınıflandırma modelleri kullanılacaktır. Sınıflandırma modellerinde model performansının ölçülmesi için karışıklık matrisi kullanılmaktadır. Karışıklık matrisi, sınıflar için doğru veya yanlış sınıflandırılmış öğe sayısını içerir. Ana köşegeninde doğru sınıflandırılmış gözlemlerin sayısını görebiliriz. Her sınıf için; köşegen dışı öğeler, yanlış sınıflandırılmış gözlemlerin sayısını ifade eder. Karışıklık matrisi, sistemin iki sınıfı doğru şekilde karıştırıp karıştırmadığını kolayca görülmesini sağlar.

Doğru ve yanlış değeri modele dair gerçek sonuçları, pozitif ve negatif değeri ise modele dair tahminleri göstermektedir. Çizelge 3.2'de karışıklık matrisi verilmiştir.

Çizelge 3.2. Sınıflandırma modelleri karışıklık matrisi

		Tahminlenen	
		Hayır	Evet
Gerçekleşen	Hayır	Doğru Negatif	Yanlış Pozitif
	Evet	Yanlış Negatif	Doğru Pozitif

Örnek üzerinden değerlendirecek olursak; çıktı değişkenimiz 0 ise başarısız, 1 ise başarılı olarak kabul edelim.

DN: 0 olarak tahmin ettiğimiz değerin (negatif), gerçekte de 0 olması (doğru).
YP: 1 olarak tahmin ettiğimiz değerin (pozitif), gerçekte 0 olması (yanlış). - > 1. hata tipi

YN: 0 olarak tahmin ettiğimiz değerin (negatif), gerçekte 1 olması (yanlış). - > 2. hata tipi
DP: 1 olarak tahmin ettiğimiz değerin (pozitif), gerçekte de 1 olması (doğru).

Karışıklık matrisi değerlerine göre performans metrikleri belirlenir. Bu metrikler:

Doğruluk: Bir modelde doğru tahminlerin, tüm veri kümesine oranı olarak hesaplanır.

$$\text{Doğruluk} = \frac{DP+DN}{DP+YP+YN+DN} \tag{5.1}$$

Kesinlik: Pozitif tahmin edilmesi gereken değişkenlerin gerçekte ne kadarının pozitif olduğudur.

$$\text{Kesinlik} = \frac{DP}{DP+YP} \tag{5.2}$$

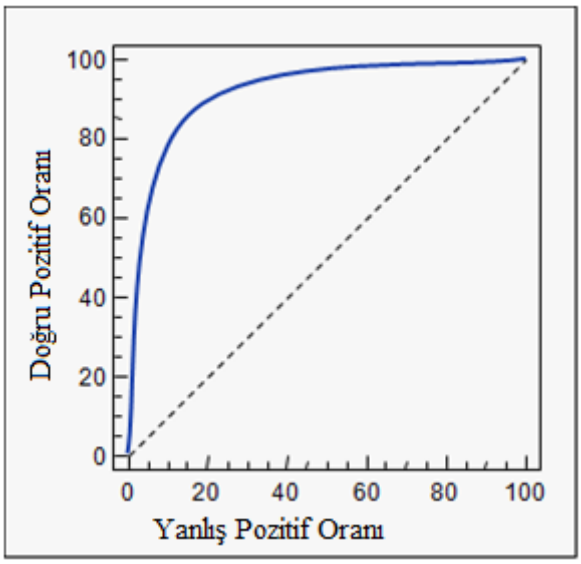
Duyarlılık: Pozitif tahmin edilmesi gereken değişkenlerin gerçekte ne kadarının doğru tahmin edildiğidir. Gerçek pozitiflerin ne kadarının doğru tahmin edildiğidir.

$$\text{Duyarlılık} = \frac{DP}{DP+YN} \tag{5.3}$$

F1 Skoru: Kesinlik ve Duyarlılık değerlerinin harmonik ortalamasıdır.

$$F1 = 2* \frac{Kesinlik*Duyarlılık}{Kesinlik+Duyarlılık} \tag{5.4}$$

Çalışma karakteristiği eğrisi: Elde edilen yanlış pozitif ve doğru pozitif oranı kullanılarak; x eksenine yanlış pozitif oranı, y eksenine doğru pozitif oranı değerleri yerleştirilerek bir eğri çizilir. Oluşturulan eğri altında kalan alan değerinin büyüklüğü modelin başarısını gösterir. Şekil 3.20'de yer alan grafikteki mavi çizgi ne kadar fazla alan kaplıyorsa modelin tahmin başarısı o derece yüksek, ortadaki kesikli çizgiye ne kadar yakınsa modelin tahmin başarısı o derece düşük diye değerlendirilebilir.



Şekil 3.20. Çalışma karakteristiği eğrisi grafiği

4. ELEKTRİK DAĞITIM SEKTÖRÜNDE MÜŞTERİ ÖDEME DURUMU TAHMİNİ UYGULAMASI

Bu bölüm Türkiye'de elektrik dağıtım sektöründe müşterilerin ödeme yapıp/yapmayacaklarının makine öğrenmesi algoritmalarıyla tahmini için yapılan tez çalışmasının uygulama bölümüdür. Verilerin elde edilmesi , veri ön işleme çalışmaları, makine öğrenmesi modellerinin oluşturulması, iyileştirilmesi ve model performanslarının karşılaştırılması aşamalarından oluşmaktadır. Bu kapsamda, tez çalışmasında uygulanan yöntemin aşamaları Şekil 4.1'de detaylı olarak sunulmuştur.

4.1. Verilerin Elde Edilmesi

Çalışmada kullanılan veriler Türkiye'de yer alan 3 elektrik dağıtım şirketinden temin edilmiştir. Çalışmaya konu olan elektrik dağıtım firmaları 14 ilde faaliyet göstermekte olup, 2020 yılı verilerine göre 11,3 milyon müşteriye hizmet vermektedir. Firmaların 2020 yılında dağıtım hattı 236064 km uzunluğa ulaşırken, toplamda 90829 trafoyla elektrik dağıtım hizmetlerini gerçekleştirmiştir.

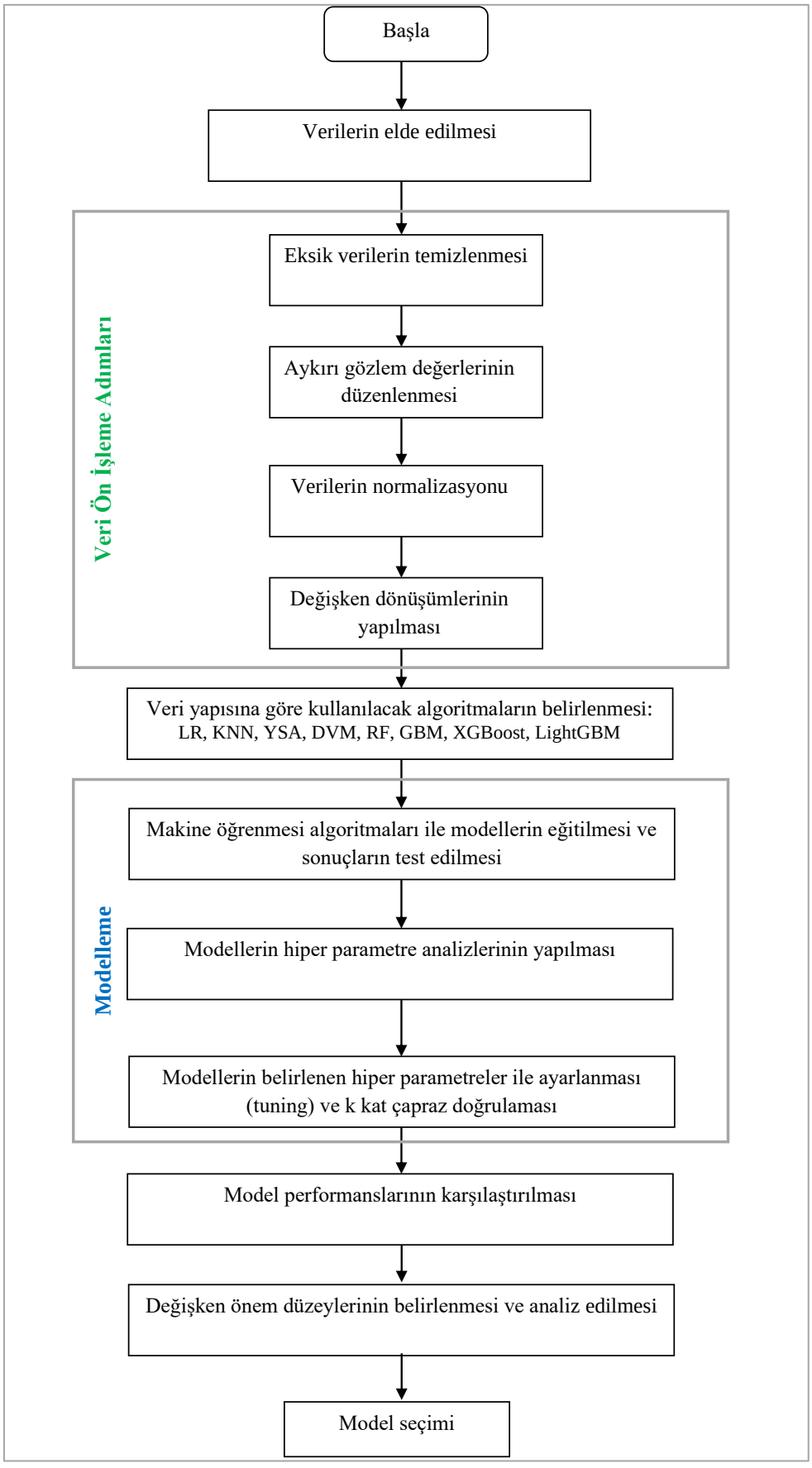
Bu çalışmada dağıtım şirketi faaliyetlerinden kaçak elektrik kullanan müşterilerin, borçlarının tahsilatı ele alınmıştır. 2020 yılında borcu bulunan 13124 müşteriden elde edilen veri seti kullanılmıştır. Veri setinde 11 farklı parametre ele alınmıştır. Değişken tiplerine göre veri ön işleme analizleri yapılarak veriler düzenlenmiştir. Çıktı değişkeni öder/ödemez şeklinde 1 ve 0'dan oluşan ikili kategorik değişken olarak belirlenmiştir.

4.2. Veri Ön İşleme Uygulaması

Çalışma sırasında veri setine aşağıdaki veri ön işleme işlemleri gerçekleştirilmiştir.

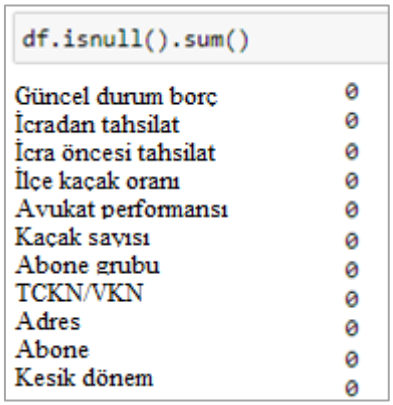
4.2.1. Veri temizleme uygulaması

Veri setinde yer alan değişkenlerde eksik veri analizi yapılmış olup, İcradan tahsilat değeri boş olan kayıtlar için "0" değeri atanmıştır.



Şekil 4.1. Uygulama adımları akış diyagramı

Veri kümesinde yer alan tüm öznitelikler Şekil 4.2'deki gibi kontrol edildiğinde veri kümesinde eksik (boş) değer kalmadığı görülmektedir.

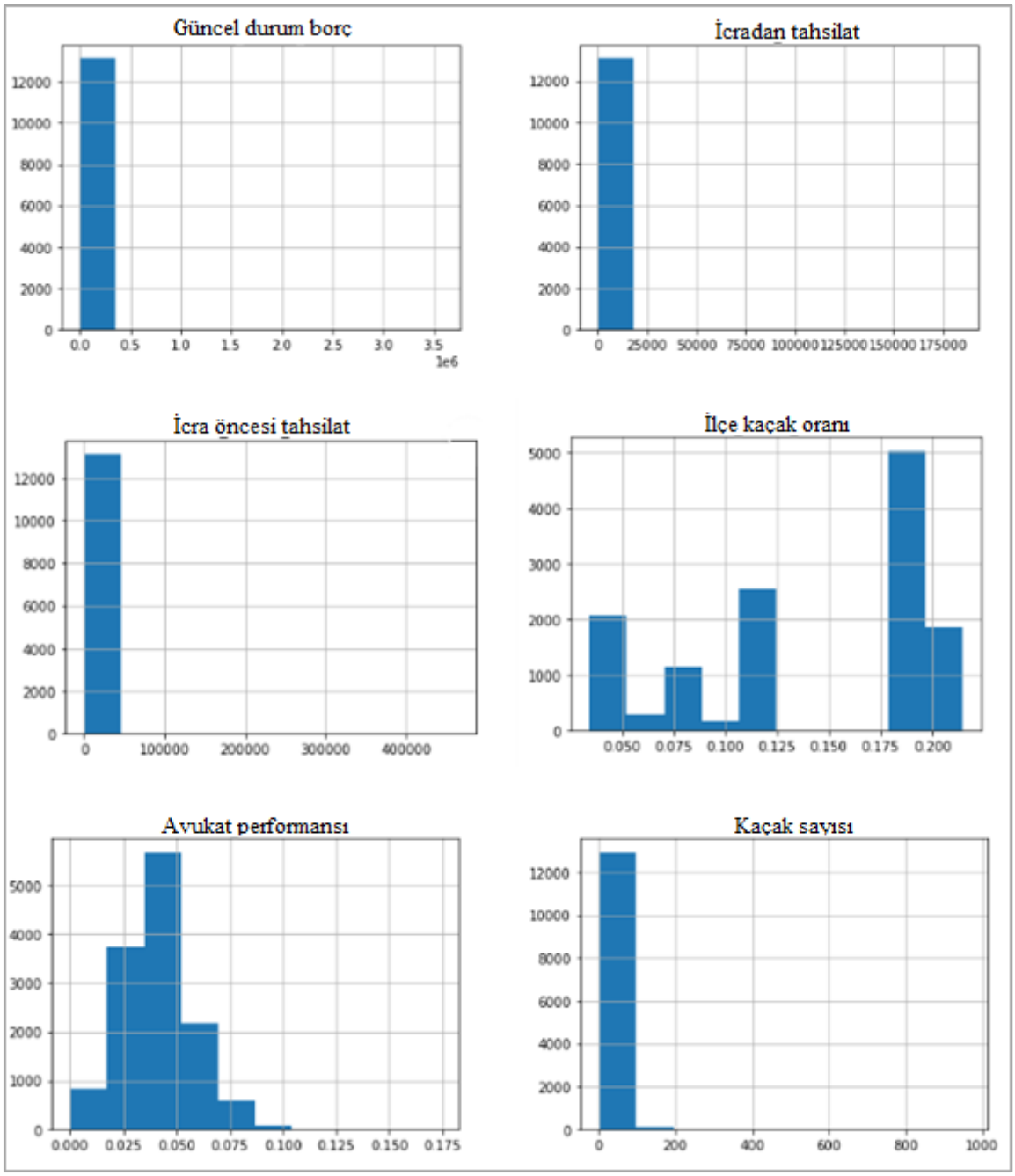


Şekil 4.2. Bağımlı değişkenlerin eksik veri analizi sonuçları

Verilerin histogram grafikleri Şekil 4.3'te verilmiştir. Kategorik olmayan değişkenlerin değişken düzeyleri ve veri dağılımları grafikler üzerinden görünmektedir.

Verilerin histogram grafikleri incelendiğinde güncel durum borç değişkeni 0,0- 3,5 milyon arasında değer almaktadır, veri yoğunluğu 0,0 ile 0,4 milyon arasındadır. İcradan tahsilat değişkeninin ise 0- 175000 arasında değerler aldığı görülmektedir. İcra öncesi tahsilat değişkeni 0- 4 milyon aralığında değer almaktadır. İlçe kaçak oranı ise 0,0 ile 0,2 arasında değerler almaktadır, yoğunluk 0,18 ile 0,2 arasında görünmektedir. Avukat performansı değişkeni 0,0 – 0,1 arasında dağılım göstermiştir. Kaçak sayısı değişkeni 0,0- 1000 arasında değer almakta olup, yoğunluğu 0,0 ile 200 arasındadır.

Histogram grafiği sonuçları incelendikten sonra aykırı gözlem analizi çalışmaları gerçekleştirilmiştir.



Şekil 4.3. Kategorik olmayan bağımsız değişkenlerin histogram grafikleri

4.2.2. Aykırı gözlem analizi uygulaması

Aykırı gözlem veri kümesinin genel dağılımı dışında kalan verilerdir. Aykırı gözlemleri gidermek ve veri yanlılığını önlemek için çeyrekler açıklığı tekniğinden yararlanılmıştır.

Çeyrekler açıklığı tekniği aşağıdaki şekilde çalışmaktadır.

Q1: parametre değerlerinin ilk çeyreklik kısmı (0,25)
Q3: parametre değerlerinin üçüncü çeyreklik kısmı (0,75)

$$ÇA = Q3-Q1 \quad (4.1)$$

ÇA değeri kullanılarak alt ve üst sınır değerleri belirlenir.

$$\text{Alt sınır} : 1–1,5*ÇA \quad (4.2)$$

$$\text{Üst sınır} : 3 + 1,5*ÇA \quad (4.3)$$

Çeyrekler açıklığı tekniği ile her bir değişkenin alt/üst sınır değerleri hesaplanmıştır. Ayrıca değişkenlerin ortalama değeri hesaplanmıştır. Ortalama değer ile alt/üst sınır değerleri karşılaştırılarak veri ön işleme yöntemine karar verilmiştir.

Verilerin alt sınır değerleri sıfır veya daha küçük çıktığı ve ilgili parametrelerde eksi değerli gözlem mümkün olamayacağı için, sadece üst sınır değerlerinden büyük olan aykırı gözlemler için sınırlandırma yoluna gidilmiştir.

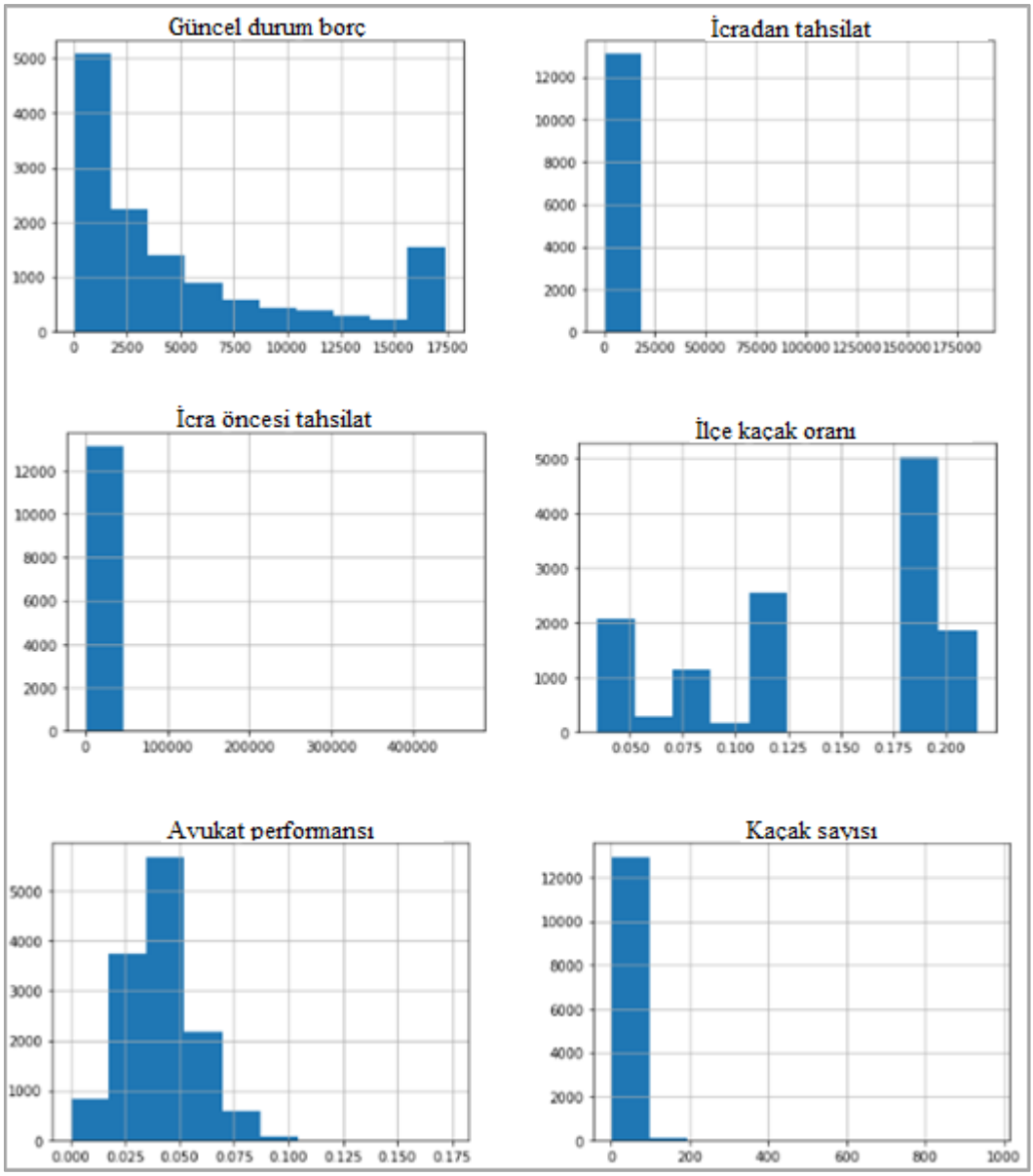
Güncel durum borç, icra öncesi tahsilat, kaçak sayısı için baskılama yöntemi kullanılarak; üst sınır değerinden büyük olan değerler, üst sınır değerine eşitlenmiştir. Baskılama yöntemi sayesinde aykırı gözlemlerin etkisi veride hissedilmeye devam edilecektir.

Yapılan işlemleri inceleyecek olursak: güncel durum borç için üst sınır değeri 17373,22 çıkmıştır, üst sınır değerinden yüksek olan 1337 aykırı adet gözlem değeri üst sınır değerine eşitlenmiştir. İcra öncesi tahsilat değişkeni için çeyrekler açığı yöntemi ile belirlenen üst sınır değeri 1971,41 olup 1560 adet aykırı gözlem değeri üst sınır değerinden büyük olduğu için üst sınır değeri ile değiştirilmiştir. Kaçak sayısı üst sınır değeri 28,5 çıkmış olup, 1496 adet aykırı gözlem değeri üst sınır değerinden büyük olduğu için 28,5'e eşitlenmiştir.

İcradan tahsilat değişkeni için ortalama değer, üst sınır değerinden büyük olduğu için; aykırı değerlerin etkisini hissedebilmek adına ortalamaya eşitlenmiştir. 1084 adet aykırı gözlem değeri ortalama değer olan 417,25'e eşitlenmiştir.

İlçe kaçak oranı ve avukat performansı değerleri için çeyrekler açıklığı yöntemine göre aykırı gözlem bulunmadığında işlem yapılmamıştır.

Verilerin aykırı değişken dönüşümü yapıldıktan sonraki durumlarını gösteren histogram grafikleri Şekil 4.4'te görülmektedir.



Şekil 4.4. Aykırı gözlem analizi sonrasında verilerin histogram grafikleri

4.2.3. Veri normalizasyonu uygulaması

Makine öğrenmesinde veri ölçeklerindeki farklılığın yanlı öğrenmeye sebep olmasını

engellemek adına normalizasyon işlemi yapılmalıdır.

Borç ve ödeme bilgilerini içeren değişkenler; güncel durum borç, icradan tahsilat, icra öncesi tahsilat, aykırı gözlem analizinden sonra 0-20000 aralığında değerler içermektedir. Kaçak sayısı verileri sayısal sürekli değerler olup aykırı gözlem analizinden sonra 0-30 aralığında değer içermektedir. İki parametre değerinin aralıkları çok farklı olduğu için makine öğrenmesi modellerinde yanlılığa sebep olmamak ve tüm özniteliklerin çıktıya/ ödemeye etkisini kullanabilmek adına normalizasyon işlemi gerçekleştirilmiştir.

Normalizasyon sonrası değerler 0-1 aralığına çekilerek tüm veri setinin aynı değer aralığında yer alması sağlanmıştır.

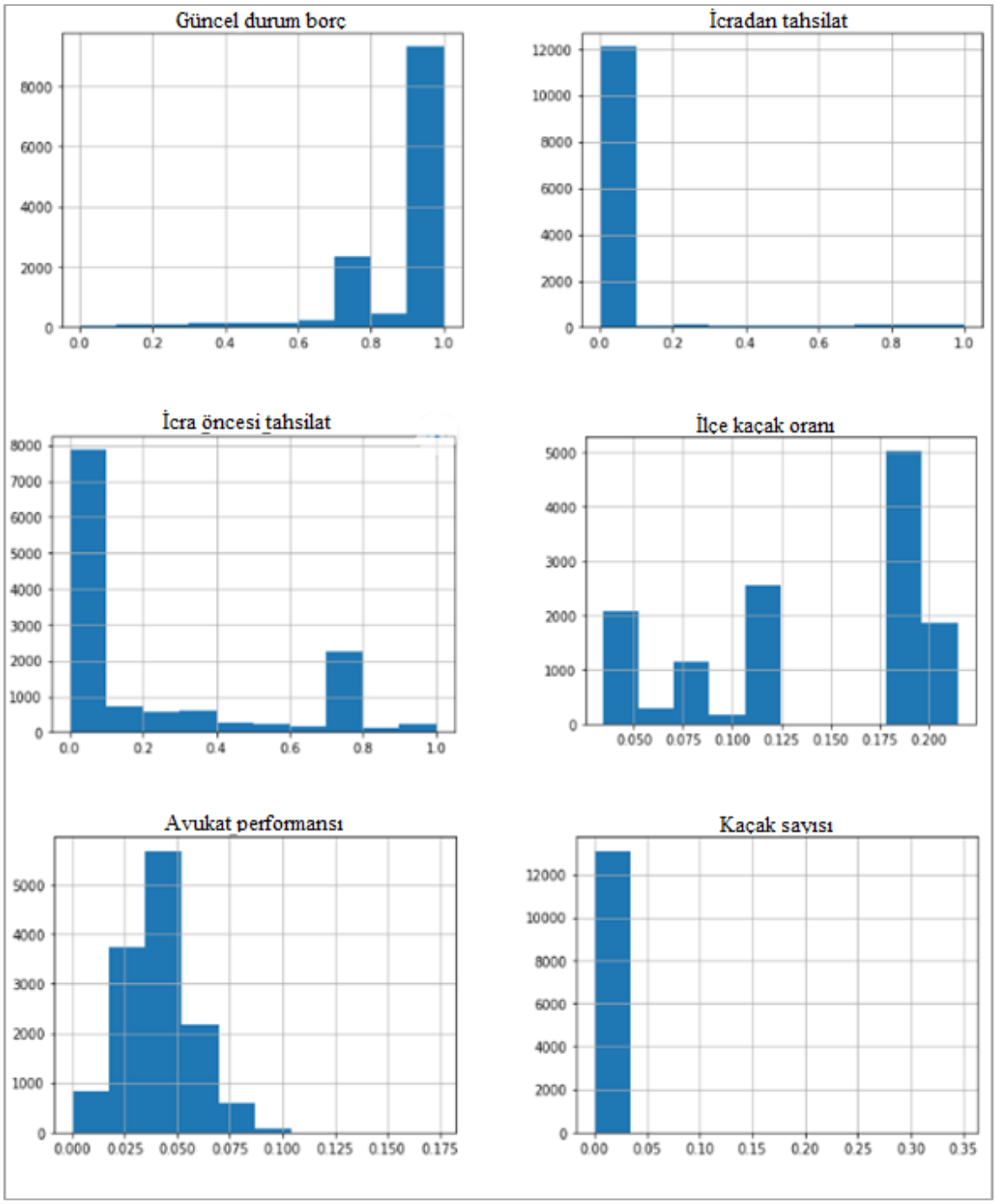
İlk 5 gözlemin tüm veri ön işleme işlemlerinden önceki durumu Şekil 4.5'te, veri ön işleme işlemlerinden sonraki görünümü Şekil 4.6'da verilmiştir. Veri ön işleme işlemlerinden sonra bağımsız değişkenlerin histogram grafikleri Şekil 4.7'de görülmektedir.

Güncel durum borç	İcradan tahsilat	İcra öncesi tahsilat	İlçe kaçak oranı	Avukat performansı	Kaçak sayısı
567.75	10251.36	43786.08	0.045262	0.024713	14
268.27	17613.76	4913.10	0.214504	0.040888	48
486.90	0.00	27965.98	0.214504	0.031936	2
205.03	7600.42	3742.22	0.214504	0.000000	2
1301.49	33030.35	25570.37	0.183065	0.000000	2

Şekil 4.5. Bağımsız değişkenlerin veri ön işleme öncesi ilk 5 satır görünümü

Güncel durum borç	İcradan tahsilat	İcra öncesi tahsilat	İlçe kaçak oranı	Avukat performansı	Kaçak sayısı
0.271186	0.199298	0.941644	0.045262	0.024713	0.006687
0.131954	0.205231	0.969676	0.214504	0.040888	0.014018
0.239775	0.000000	0.970828	0.214504	0.031936	0.000985
0.101225	0.205998	0.973302	0.214504	0.000000	0.000987
0.542549	0.173937	0.821818	0.183065	0.000000	0.000834

Şekil 4.6. Bağımsız değişkenlerin veri ön işleme sonrası ilk 5 satır görünümü



Şekil 4.7. Aykırı gözlem ve normalizasyon sonrası verilerin histogram grafikleri

Tüm veri ön işleme işlemlerinden sonra verilerin histogram grafikleri değerlendirildiğinde Güncel durum borç değişkenin değerlerinin 0,6- 1 arasında yığıldığı görülmektedir. İcradan tahsilat değişkeni veri kümesi 0- 0,1 arasında yoğunlaşmaktadır. İcra öncesi tahsilat değeri 0– 0,1 arasında yoğunlaşmakta olup, diğer değişkenlere göre farklı aralıklarda daha fazla dağılım göstermektedir. İlçe kaçak oranı ve avukat performansında bir değişiklik yapılmamıştır. Kaçak sayısı 0- 0,02 arasında dağılmaktadır.

4.2.4. Değişken dönüşümleri uygulaması

Makine öğrenimi modelleri, tüm girdi ve çıktı değişkenlerinin sayısal olmasını gerektirmektedir. Kategorik değerler içeren veri setleri, model kurulmadan önce sayılar ile ifade edilmelidir. Bu sebeple veri setinde yer alan kategorik değişkenler için değişken dönüşümü işlemi gerçekleştirilmiştir.

Abone grubu değişken dönüşümü

Müşterilerin abone gruplarına göre ödeme olasılığı farklılıklarının ödemeye olan etkisinin görülebilmesi için modelde yer alması istenilmektedir. Mesken ise evet, değilse hayır işaretlenmişti. Kategorik değişkenleri sayısal değerlere çevirmek için evet ise 1, değilse 0 olarak kodlanmıştır.

TCKN/VKN değişken dönüşümü

TCKN/VKN bilgisinin var olması model için önemli bir değişken olup var/yok olarak veri tutulmaktaydı. TCKN/VKN'si var olan müşteriler için 1, olmayanlar için 0 olarak dönüşüm yapılmıştır.

Adres değişken dönüşümü

Müşterilere ait adres verisi incelenerek kapı numarasına kadar adres bilgisi mevcut ise var, yoksa yok olarak veri tutulmaktaydı. Var ise 1, yoksa 0 olarak kodlanmıştır.

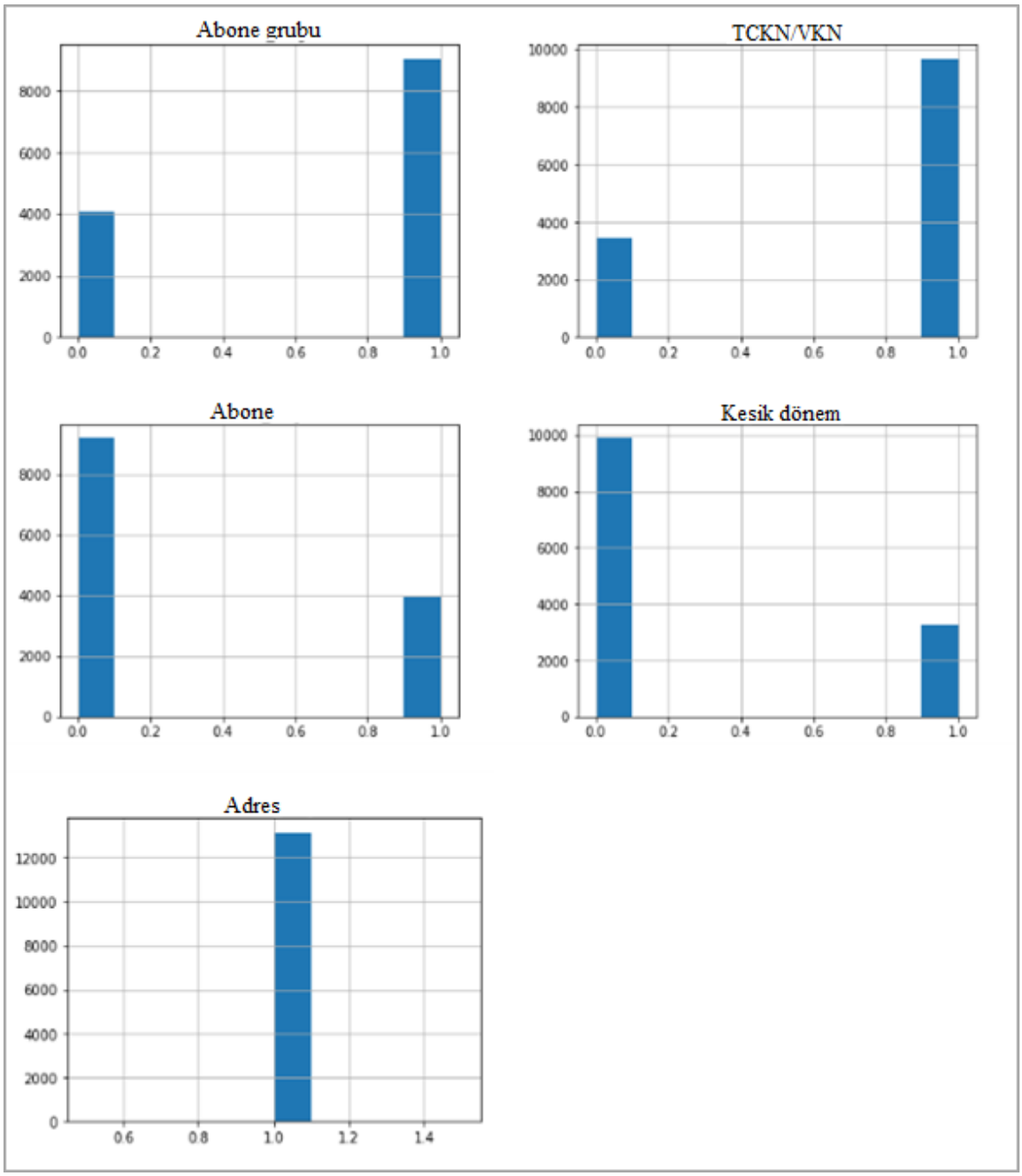
Abone değişken dönüşümü

Abone olan müşteriler için 1, olmayan müşteriler için 0 girilmiştir.

Kesik dönem değişken dönüşümü

Kesik dönem verisi evet ise 1, hayır ise 0 olarak işaretlenmiştir.

Kategorik değişkenlerin histogram grafiği Şekil 4.8'de görünmektedir.



Şekil 4.8. Kategorik değişkenlerin histogram grafikleri

Abone grubu, TCKN/VKN değişkeni gözlem değerleri daha çok 1 değerini içermektedir. Ortalamaları 0,5'ten büyüktür. Abone ve kesik dönem değişkeni gözlem değerleri ise daha çok 0 değeri içermekte olup ortalama 0,5'ten küçüktür. Adres değişkeni ortalaması 1 olup tüm müşterilerin adresleri mevcuttur. Bu sebeple bu veri seti için Adres değişkeni anlamlı değildir. Veri ön işleme çalışmalarından sonra değişken tipleri Çizelge 4.1'deki şekilde düzenlenmiştir.

Çizelge 4.1. Veri ön işleme sonrası bağımsız değişkenlerin adı, tanımı ve tipi

Değişken Adı	Tanımı	Değişken Tipi
Güncel durum borç	Müşterinin ödenmemiş toplam borç miktarı	Normalize
İcradan tahsilat	Müşteriden daha önceden icra yolu ile tahsil edilen toplam borç miktarı	Normalize
İcra öncesi tahsilat	Müşteriden icra öncesi tahsil edilen toplam borç miktarı	Normalize
İlçe kaçak oranı	Müşterinin bağlı olduğu ilçenin kaçak oranı	Nominal
Avukat performansı	Müşteri borcunun atandığı avukat performansı	Nominal
Kaçak sayısı	Müşterinin toplam kaçak kullanım sayısı	Normalize
Abone grubu	Müşteri Abone grubu tipi	Kategorik
TCKN/VKN	TCKN/VKN bilgisi	Kategorik
Adres	Adres bilgisi	Kategorik
Abone	Abone bilgisi	Kategorik
Kesik dönem	Müşterinin enerjisinin kesik olma durumu	Kategorik

4.3. Model Sonuçları

Çıktı değişkeni ikili kategorik değişken olan veri setimize uygun olan makine öğrenmesi yöntemleri değerlendirildiğinde; denetimli sınıflandırma modelleri kullanılması gerektiği görülmüştür.

Bu çalışmada 13124 farklı müşteri için 11 bağımsız, 1 bağımlı değişkenden oluşan veri seti üzerinde Anaconda Jupyter Notebook 6.4.5 sürümü Python kodlaması kullanılarak deneyler gerçekleştirilmiştir.

Lojistik regresyon, k en yakın komşu algoritması, destek vektör makineleri, yapay sinir ağları, rastgele ormanlar, gradyan artırma, aşırı gradyan artırma, hafif gradyan artırma algoritmaları kullanılmış ve 8 farklı model kurulmuştur.

Makine öğrenmesi algoritmaları eğitilirken, hiper parametrelerin belirlenmesi önem arz etmektedir. Hiper parametreler algoritmaya göre farklılaşarak, öğrenme işlemi sırasındaki davranışları yönlendirirler. Veri seti öncelikle %70 eğitim, %30 test olacak şekilde ayarlanmıştır. Modelleri ayarlamak için ızgara arama yöntemi ile hiper parametre seçimi yapılmış ve 5 katlı çapraz doğrulama kullanılarak aşırı öğrenmenin önüne geçilmiştir.

4.3.1. Hiper parametre analizi

Izgara arama yöntemi kullanılarak hiper parametre analizi yapılırken; girilen aralıktaki tüm değerler için modeller eğitilir ve en iyi sonucu sağlayan değerler hiper parametre olarak seçilir. Modellemede kullanılan algoritmalar için farklı hiper parametre değerleri mevcut olup, her bir algoritma için ayrıca hesaplama yapılmıştır.

Lojistik regresyon C düzenlileştirme hiper parametresi analiz edilmiştir. Düzenlemenin gücü, C ile ters orantılıdır. Kesinlikle pozitif olmalıdır. Model sonucunu en iyileyen C değeri 1 olarak hesaplanmıştır.

KEYK algoritması için n_neighbors sorguda kullanılacak olan komşu sayısı hiper parametresi analiz edilmiş olup, en iyi sonuç veren değer 2 olarak hesaplanmıştır.

DVM algoritması için C ve Kernel hiper parametre değerleri belirlenmiştir. C düzenlileştirme hiper parametresidir ve pozitif değer olmalıdır. Kernel, algoritmada kullanılacak çekirdek türünü belirtir. C: 1, kernel: linear olarak belirlenmiştir.

YSA modeli için hidden_layers ve alpha hiper parametreleri analiz edilmiştir. Hidden_layers, gizli katmandaki nöronların sayısını temsil eder. Analiz sonucunda en iyi parametre değeri (100, 100) olarak belirlenmiştir. Alpha düzenleme hiper parametresidir. Ondalık değerler alır ve en iyi parametre değeri 0,1 olarak belirlenmiştir.

Rastgele ormanlar modelini ayarlamak için 3 hiper parametre analizi yapılmıştır. Bunlar; n_estimators, max_features, min_samples_split. Ormandaki ağaç sayısı n_estimators ile ifade edilmektedir ve en iyi parametre değeri 1000 olarak hesaplanmıştır. En iyi bölünmeye ulaşmak için kullanılması gereken özelliklerin sayısı max_features hiper parametresi ile ifade edilir. En iyi max_features değeri 3 olarak belirlenmiştir. Yaprak

düğümünde yer alması gereken en düşük örnek sayısı min_samples_split hiper parametresi ile ifade edilir. Model için en iyi parametre değeri 20 olarak hesaplanmıştır.

Gradyan artırma modeli learning_rate, max_dept, n_estimators hiper parametreleri analiz edilmiştir. Öğrenme oranı, her ağacın katkısını 'learning_rate' oranında küçültür. Gerçekleştirilecek güçlendirme aşamalarının sayısı n_estimators hiper parametresi ile ifade edilir. Gradyan artırma, aşırı öğrenmeye karşı oldukça dayanıklıdır, bu nedenle n_estimators'ın çok olması genellikle daha iyi performans sağlar. Maksimum derinlik, max_dept, ağaçtaki düğüm sayısını sınırlar. Hiper parametrelerin en iyi değerleri max_depth: 5 çıkmıştır, learning_rate: 0,01 ve n_estimators: 1000 olarak hesaplanmıştır.

AGAM algoritması modelinde learning_rate, max_depth, n_estimators, subsample hiper parametreleri için en iyi parametre değerleri araştırılmıştır. Hiper parametreleri incelediğimizde; n_estimators: yükseltme turlarının sayısı, max_depth: maksimum ağaç derinliği, learning_rate: artan öğrenme oranı, subsample: eğitim örneğinin alt örnek oranını ifade etmektedir. Parametre analizi sonucunda learning_rate: 0,1, max_depth: 3, n_estimators: 500 ve subsample: 0,8 olarak hesaplanmıştır.

HGAM algoritması modelinde AGAM modelindeki gibi learning_rate, max_depth, n_estimators hiper parametreleri analiz edilmiştir. En iyi hiper parametre değerleri learning_rate: 0,01, max_depth: 35, ve n_estimators: 500 olarak belirlenmiştir.

Çizelge 4.2'de tüm modeller için belirlenmiş olan en iyi hiper parametre değerleri listelenmiştir. Tüm modeller için belirlenen hiper parametreler kullanılarak modeller tekrar kurulmuştur. Böylece model ayarlamaları gerçekleştirilerek, model performansları iyileştirilmiştir.

Çizelge 4.2. Hiper parametre analizi sonucu belirlenen en iyi hiper parametre değerleri

Algoritma	Hiper Parametre	En İyi Değer
LR	C	1
KEYK	n_neighbors	2
DVM	C	1
	Kernel	linear

Çizelge 4.2. (devam) Hiper parametre analizi sonucu belirlenen en iyi hiper parametre değerleri

YSA	Hidden_layers	(100, 100)
	Alpha	0,1
RO	n_estimators	1000
	max_features	3
	min_samples_split	20
GAM	learning_rate	5
	max_dept	0,01
	n_estimators	1000
AGAM	learning_rate	3
	max_dept	0,1
	n_estimators	0,8
	subsample	500
HGAM	learning_rate	0,01
	max_dept	35
	n_estimators	500

4.3.2. Model performanslarının karşılaştırılması

Bölüm 4.3.1'de belirlenen hiper parametreler ile modeller tekrar kurularak nihai sonuçlar elde edilmiştir. Uygulamada 8 farklı algoritma için model kurulmuştur. Bu modellerin sonuçlarının başarısını ölçmek için test verisi üzerinde doğrulama yapılmıştır. Model sonuçlarını ölçümleyebilmek için modellerin test verisini tahminleme durumlarını gösteren karışıklık matrisi değerleri hesaplanmıştır. Model tahmin performanslarını değerlendirmek için elde edilen karışıklık matrisi değerleri Şekil 4.9'da görünmektedir. Şekil 4.9'da yer alan karışıklık matrisinde sırasıyla doğru negatif, yanlış pozitif, yanlış negatif ve doğru pozitif değerleri gösterilmektedir.

Denetimli öğrenmede etiketli eğitim verisi ile öğrenen model, test verisi ile doğrulanırken; test verisinin gerçek etiket değerleri ile model sonucunda tahminlenen çıktı değerleri karşılaştırılır. Problemimiz üzerinde DN, YP, YN, DP değerlerini değerlendirecek olursak; çıktı değişkenimiz Öder/Ödemez şeklinde etiketlenmişti.

DN: Modelin ödemez olarak tahmin ettiği müşterinin (negatif), gerçekte de ödemez olması
YP: Modelin öder olarak tahmin ettiği müşterinin (pozitif), gerçekte ödemez olması
YN: Modelin ödemez olarak tahmin ettiği müşterinin (negatif), gerçekte öder olması
DP: Modelin öder olarak tahmin ettiğimiz müşterinin (pozitif), gerçekte de öder olması

Modeller	Karışıklık Matrisi
Lojistik Regresyon	[[1512, 94], [361, 1971]]
En Yakın Komşu Alg.	[[1511, 95], [444, 1888]]
Destek Vektör Makineleri	[[1547, 59], [476, 1856]]
Yapay Sinir Ağları	[[1488, 118], [322, 2010]]
Rastgele Ormanlar	[[1495, 111], [253, 2079]]
Gradyan Artırma	[[1485, 121], [234, 2098]]
Aşırı Gradyan Artıma	[[1482, 124], [235, 2097]]
Hafif Gradyan Artırma	[[1481, 125], [233, 2099]]

Şekil 4.9. Sınıflandırma modelleri karışıklık matrisi

Karışıklık matrisi değerlerini kullanarak lojistik regresyon üzerinden performans metriklerini hesaplayalım. LR için DN sayısı 1512, YP sayısı ise 94'tür. YN sayısı 361, DP sayısı ise 1971'dir.

Doğruluk değeri modelin gerçekte öder veya ödemez şeklinde etiketli olan müşterileri ne kadar doğru tahminlediğinin göstergesidir. Modelin öder dediği ve gerçekte öder olan DP sayısı ise 1971, modelin ödemez dediği ve gerçekte de ödemez olan DN sayısı 1512'dir. DP ve DN değerlerinin toplamının tüm test verisine bölünmesi ile elde edilir.

Doğruluk: (1512+1971)/ (1512+94+361+1971) = 0,88445912'dir.

Kesinlik değeri modelin öder olarak tahminlediği müşterilerin gerçekte ne kadarının ödeme yapar olarak etiketli olduğunun göstergesidir. Modelin öder dedikleri YP ve DP'tir, Bunların ne kadarının gerçekte öder olduğu da DP değeridir.

Kesinlik: (1971)/ (94+1971) = 0,95447942

Duyarlılık gerçekte öder olarak etiketli olan müşterilerinin ne kadarını modelin öder olarak tahminlendiğinin göstergesidir.

Duyarlılık: (1971) / (361+1971) = 0,84519726

F1 Skoru kesinlik ve duyarlılık değerinin harmonik ortalamasıdır.

F1 Skoru: (2* 0,95447942* 0,84519726)/ (0,95447942 + 0,84519726) = 0,89652035

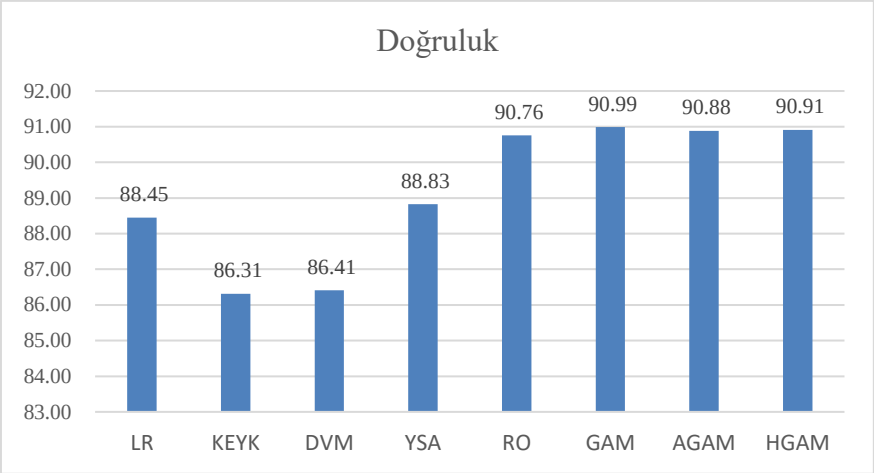
Çalışma karakteristiği eğrisi değeri hesaplanırken x eksenine YP, y eksenine de DP değerleri yerleştirilerek grafik çizdirilmiş olup altında kalan alan hesaplanmıştır. Çalışma karakteristiği değeri 0,89333337 çıkmıştır.

Aynı hesaplama mantığı diğer tüm algoritmalar için de uygulanarak performans metrikleri belirlenmiştir. Şekil 4.10'da tüm modeller için performans metriği değerleri görünmektedir.

Modeller	Doğruluk	Kesinlik	Duyarlılık	F1 Skoru	Çalışma Karakteristiği
Lojistik Regresyon	88.445912	95.447942	84.519726	89.652035	89.333337
En Yakın Komşu Alg.	86.312849	95.209279	80.960549	87.508691	87.522616
Destek Vektör Makineleri	86.414424	96.919060	79.588336	87.402873	87.957306
Yapay Sinir Ağları	88.826816	94.454887	86.192110	90.134529	89.422331
Rastgele Ormanlar	90.756729	94.931507	89.150943	91.950464	91.119681
Gradyan Artırma	90.985272	94.547093	89.965695	92.199517	91.215724
Aşırı Gradyan Artırma	90.883697	94.416929	89.922813	92.115089	91.100883
Hafif Gradyan Artırma	90.909091	94.379496	90.008576	92.142230	91.112632

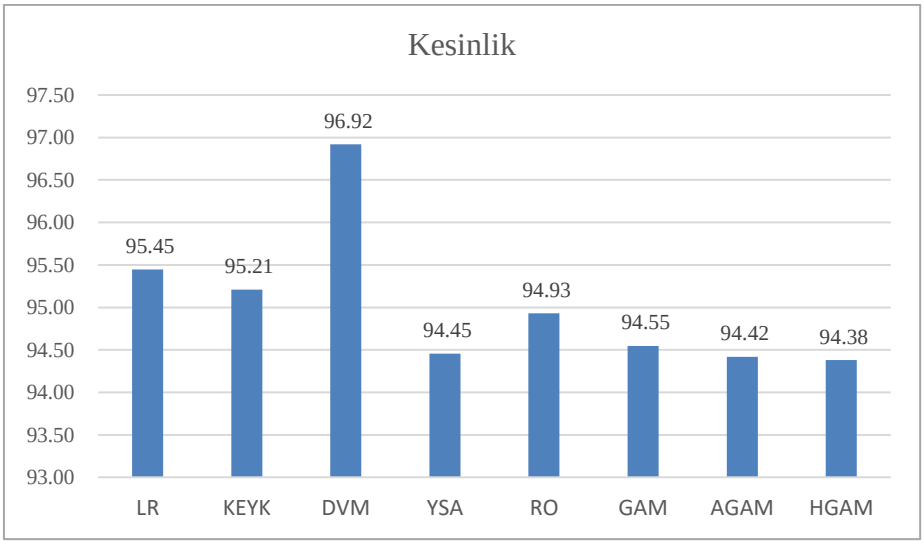
Şekil 4.10. Sınıflandırma modelleri performans ölçümü sonuçları

Tüm modeller için belirlenen performans metrikleri grafikler üzerinde gösterilmiştir. (Şekil 4.11- Şekil 4.15)



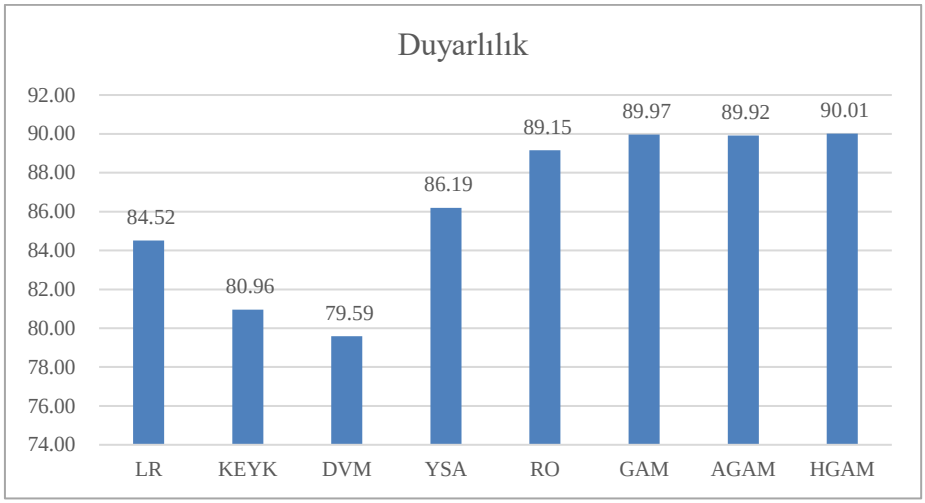
Şekil 4.11. Sınıflandırma modelleri doğruluk değerlerinin grafik gösterimi

Şekil 4.11'de yer alan doğruluk değerleri incelendiğinde tüm modellerin genel olarak test verisi üzerinde müşterilerin öder veya ödemez olduğunu tahminleme oranı yüksektir. GAM, HGAM, AGAM, RO modellerinin doğruluk değerleri %90 üzeri hesaplanmıştır. En iyi doğruluk değeri ise gradyan artırma makineleri modeli sonucunda elde edilmiştir.



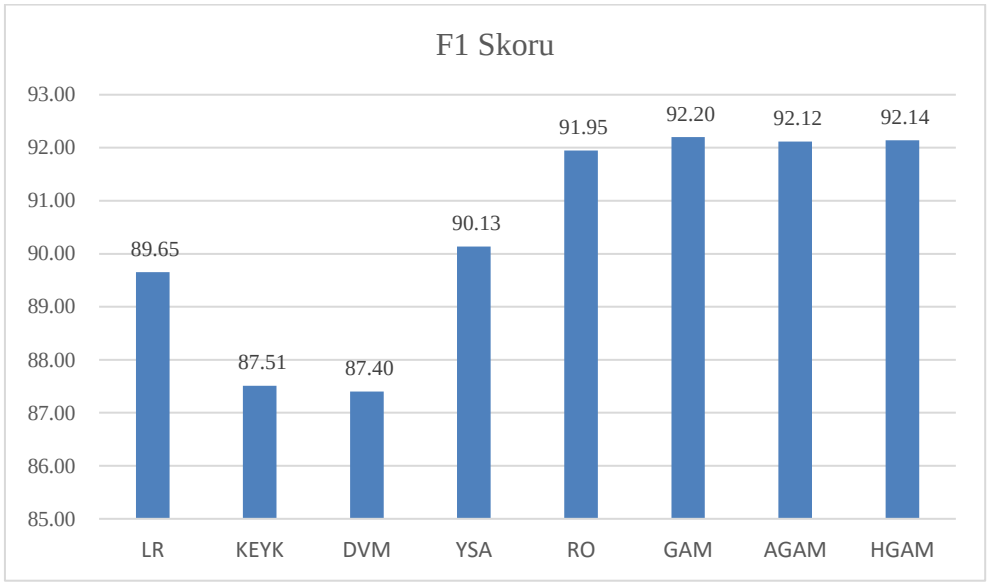
Şekil 4.12. Sınıflandırma modelleri kesinlik değerlerinin grafik gösterimi

Şekil 4.12'de yer alan kesinlik değerleri sonuçları incelendiğinde tüm modellerin öder olarak tahminlediği müşterilerin %94'ü gerçekte de öder olarak etiketlidir. Bu da aslında modellerin test veri seti üzerinde öder diye tahmin yaptığı müşterilerde yanılmadığının bir göstergesidir. Tüm modeller içerisinde en iyi kesinlik değeri %96,92'lik oranla DVM modelinin olmuştur.



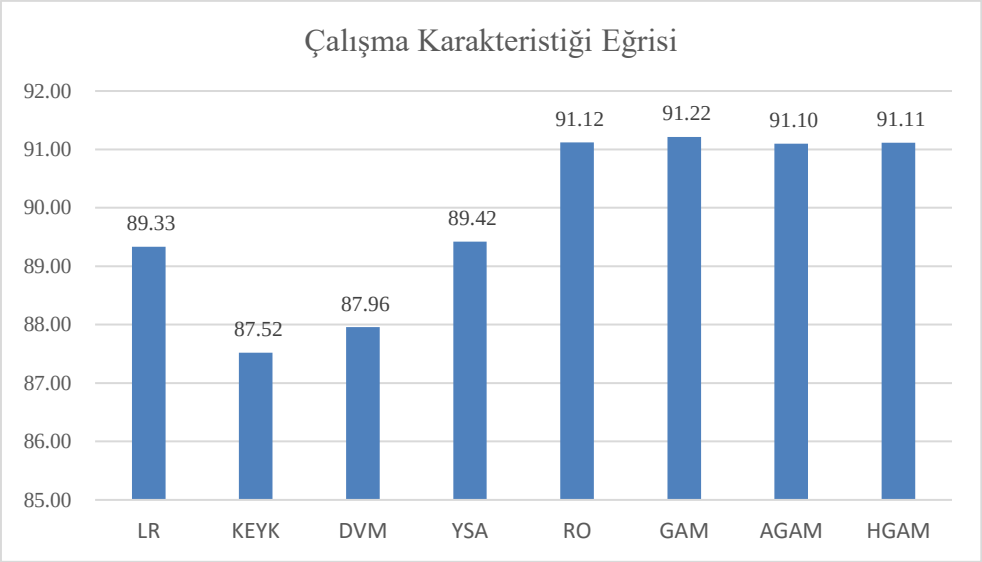
Şekil 4.13. Sınıflandırma modelleri duyarlılık değerlerinin grafik gösterimi

Şekil 4.13'te yer alan duyarlılık değeri sonuçlarına baktığımızda modellerin gerçekte öder olarak etiketli olan müşterileri doğru yakaladığı görülmektedir. Diğer performans değerlerine paralel şekilde topluluk modelleri duyarlılık değerleri daha yüksek olup, en iyi duyarlılık değeri 90,01 HGAM modeli sonucunda hesaplanmıştır.



Şekil 4.14. Sınıflandırma modelleri F1 skoru değerlerinin grafik gösterimi

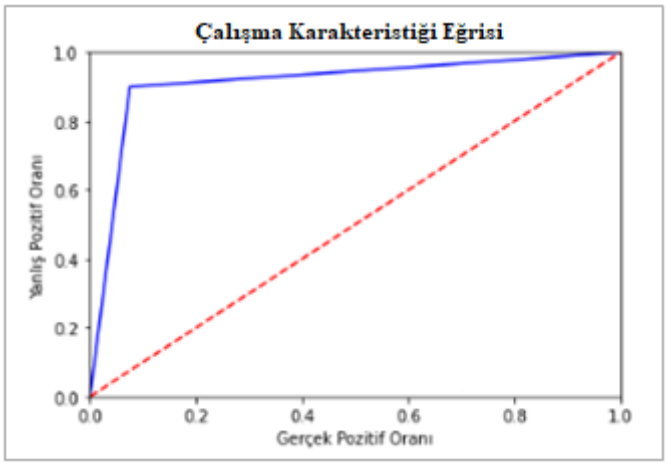
Şekil 4.14'te yer alan F1 skoru duyarlılık ve kesinlik değerlerine paralel olarak topluluk modellerinde yüksek çıkmıştır. F1 skoru veri setindeki dengesiz dağılım olan durumlarda etkili bir model seçim kriteridir. Sonuçlar incelendiğinde F1 skoru en yüksek model GAM olarak hesaplanmıştır.



Şekil 4.15. Sınıflandırma modelleri çalışma karakteristiği eğrisi değerlerinin gösterimi

Çalışma karakteristiği eğrisi değerleri doğru pozitif ve yanlış pozitif değerleri kullanılarak oluşturulan eğri altında kalan alanı ifade etmektedir. Bu eğrinin altında kalan alanın büyük olmasının yolu doğru pozitif değerinin yüksek, yanlış pozitif değerinin düşük olmasıdır. Böylece modelin öder olarak etiketli müşterileri doğru yakalama oranı değerlendirilmiş olmaktadır. Modeller için hesaplanmış olan çalışma karakteristiği eğrisi değerleri Şekil 4.15'te görülmektedir.

Çalışma karakteristiği eğrisi değeri en iyi sonuç veren GAM modelinin grafiği Şekil 4.16'da görünmektedir.



Şekil 4.16. GAM sonuçlarının çalışma karakteristiği eğrisi grafiği

Model performansları karşılaştırılırken performans metriklerinden yararlanılmıştır. İlgili veri seti için topluluk algoritması RO, GAM, AGAM, HGAM modellerinin tahmin performans değerleri birbirine yakın çıkmış olup, LR, KEYK, DVM ve YSA'ya göre topluluk modellerinin daha iyi performans gösterdiği görülmüştür. Doğruluk, Kesinlik, Duyarlılık, F1 Skoru ve Çalışma karakteristiği eğrisi performans metrikten 3'ünde GAM modeli değerlerinin daha yüksek olduğu görülmekte ve doğruluk oranı: 90,99, F1 skoru: 92,20, çalışma karakteristiği eğrisi değeri: 91,22'dir. GAM modelinden sonra en başarılı model tahmin performansı olan HGAM modelinin doğruluk oranı: 90,91, F1 skoru: 92,14, çalışma karakteristiği eğrisi değeri: 91,11 olduğu görülmektedir. Bu sonuçlar doğrultusunda en iyi performans gösteren modeli seçerken değişken önem düzeyleri analizinden de yararlanılarak model çalışma performansının yanında değişken durum analizi de değerlendirilmiştir.

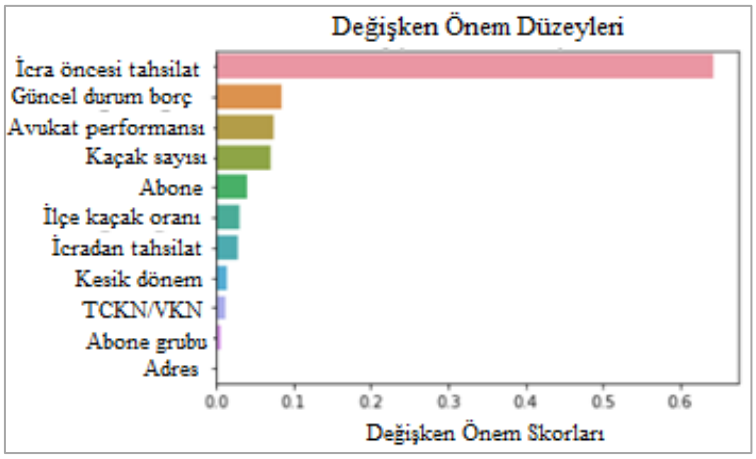
4.3.3. Değişken önem düzeyleri

Ağaç temelli modellerde değişken önem düzeyleri analiz edilmiş olup birbirleriyle karşılaştırdığımızda modelden modele farklı parametrelerin daha etkili olduğu görülmüştür. Değişken önem düzeyleri analiz sonuçları kullanılarak operasyonel süreçlerin şekillenmesi ve stratejik kararların alınması sağlanabilir.

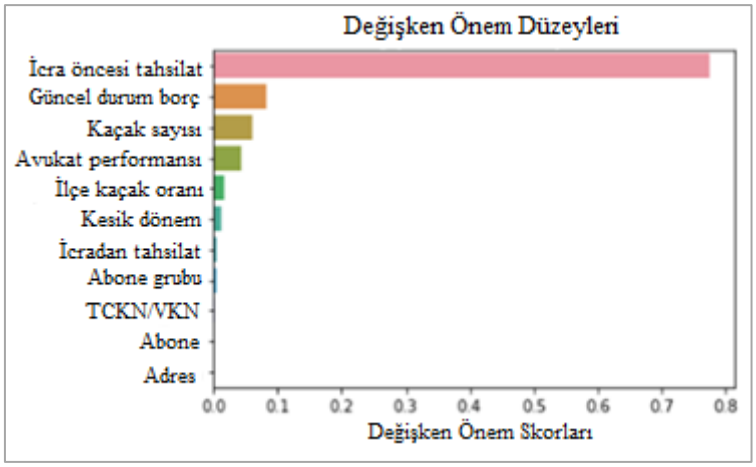
İcra öncesi tahsilat verisi 3 modelde açık ara en önemli parametre olarak görünmektedir. İcra öncesi tahsilat değişkeni modeller için en yüksek önem düzeyine sahip diyebiliriz. Onu güncel durum borç, kaçak sayısı, avukat performansı, içe kaçak oranı, kesik dönem, icradan tahsilat ve abone grubu izlemektedir. TCKN/VKN, Abone ve Adres ise genelde modele girmeyen parametre olmuştur.

Adres değişkeninin önem düzeyi tüm modellerde düşük çıkmış olup ödeme üzerindeki etkisi düşük olduğu görülmüştür. Bunda adres verisinin tüm müşteriler için 1 olması bir etken olabilir.

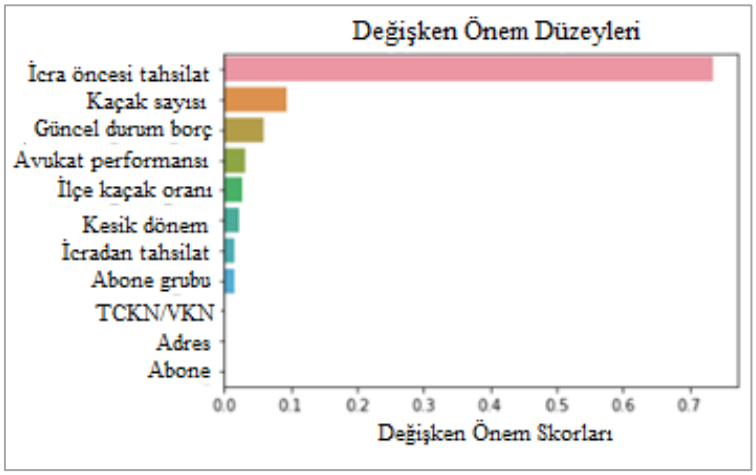
Görüldüğü gibi modellerin çalışmasında farklı parametreler farklı etkiler doğurmaktadır. Modellerin değişken önem düzeyleri Şekil 4.17 - Şekil 4.20'de verilmiştir.



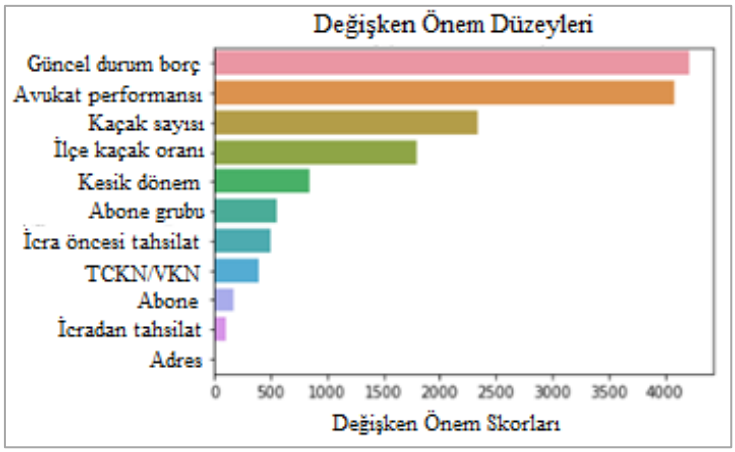
Şekil 4.17. Rastgele ormanlar modeli sonucunda çıkan değişken önem düzeyleri



Şekil 4.18. Gradyan artırma modeli sonucunda çıkan değişken önem düzeyleri



Şekil 4.19. Aşırı gradyan artırma modeli sonucunda çıkan değişken önem düzeyleri



Şekil 4.20. Hafif gradyan artırma modeli sonucunda çıkan değişken önem düzeyleri

Değişken önem düzeyleri sonuçları kullanılarak şirketlerin tahsilat stratejilerine yön verilebilir. Tahsilat kampanyası düzenlenirken model sonuçları kullanılarak kampanyaya dahil edilecek müşterileri belirlemek için ödeme yapar/ yapmaz bilgisinin yanında, önem düzeyi yüksek olan; icra öncesi tahsilat, güncel borç durumları ve kaçak sayısı değişkenleri kullanılarak, tahsilat potansiyeli yüksek olan müşteriler seçilebilir. Böylece operasyonel süreçlerde harcanacak efor, tahsilat yapılabilecek müşterilere yönlendirilmiş olur. Tahsilat potansiyeli düşük olan müşteriler de elde edilebileceği için bu müşteriler için de özel bir çalışma yöntemi geliştirilebilir.

İcradan tahsilat gerçekleştiğinde dağıtım şirketleri ekstra kazanç sağlamaktadır. Bu sebeple avukat performansı gibi insani etkenlerin yönetilmesi oldukça önemlidir. Modellerin kurulmasında numerik değerlerin yanında avukat performansı gibi insani değişkenler bu sebeple veri setine eklenmiştir. Değişken önem düzeyleri analizi de göstermektedir ki tahsilat sürecinde avukat performansı metriği genel olarak etkili görünmektedir. Yöneticiler karar aşamasında bu metriği dikkate alarak stratejilerini oluşturabilirler. Ayrıca avukat performansları ölçümlenirken de müşteri tahsilat potansiyeli bilgisi ve tahsilat gerçekleştirmelerini karşılaştırabilir.

TC/VKN değişkeni müşterilerin kimlik bilgilerini tutmaktadır. Bu bilginin var olması operasyonel olarak bakıldığında icra takibinin başlatılmasında önemli bir kriter olarak görülmektedir. Fakat değişken önem düzeyi analizlerine baktığımızda ödeme üzerinde etkisinin olmadığı değerlendirilebilir.

Model tahmin performansı değerlendirmesinde iyi sonuç veren GAM ve HGAM algoritmalarının değişken önem düzeyi analizi sonuçları karşılaştırılmıştır. Sonuçlar incelendiğinde gerçek hayatta kurulacak olan model değerlendirildiğinde daha az parametre ile daha hızlı çalışan bir model kurmak tercih edilebilir. Bu gözle baktığımızda GAM modeli daha az parametre ile daha hızlı bir şekilde sonuç elde etmektedir. Bu sebeple sadece model tahmin performans değerlerine göre değil, değişken önem düzeyi analizi sonuçlarına göre de GAM modeli en iyi sonuç veren model olarak değerlendirilmiştir.

5. SONUÇLAR VE ÖNERİLER

Bu çalışmada Türkiye'de üç elektrik dağıtım şirketine ait müşterilerinin borç ödeme durumu tahmini yapılmıştır. Veri seti hazırlanırken firmanın 2020 yılındaki kaçak elektrik kullanımı yapan müşterilerinin borçları baz alınmıştır. Denetimli makine öğrenmesi algoritmalarından; lojistik regresyon, k- en yakın komşu, yapay sinir ağları, destek vektör makineleri, rastgele ormanlar, gradyan artırma makineleri, aşırı gradyan artırma ve hafif gradyan artırma makineleri kullanılarak müşterilerin ödeme durumlarını tahminleme modelleri geliştirilmiş, model tahmin performansları ve değişken önem düzeyleri karşılaştırılarak en iyi sonuç veren model seçilmiştir.

Geçmişten günümüze borç ödeme durumunu tahmin etme üzerine yapılmış olan literatür çalışmaları incelendiğinde genelde finans ve bankacılık sektöründe kredi risk tahmini üzerine gerçekleştirilen çalışmalar ile karşılaşılmaktadır. Bunun yanında müşteri borç tahsilatı tahmini ve firma başarı tahmini çalışmalarına da rastlanmaktadır. Türkiye'de elektrik dağıtım sektöründe müşterilerin ödeme durumunu tahmin etmek için oluşturulan bir çalışmaya rastlanmamıştır.

Günümüzde borç ödeme tahmini problemi sadece finans kuruluşlarını değil tüm sektörlerdeki firmaları ilgilendirir hale gelmiştir. Firmalar finansal faaliyetlerini yürütebilmek, nakit akışını artırmak ve kar elde edebilmek için müşterinin ödeme davranışlarını inceleyerek, stratejiler belirlemeyi ve tahsilatı artırmayı amaçlamaktadır.

Şirketlerin sahip oldukları veri boyutu arttıkça veriyi anlamlandırmak ve şirket stratejilerine yön vermek için kullanabilmek önem arz etmeye başlamaktadır. Veri anlamlandırılmadığı sürece depolanan ve maliyet yaratan bir kümedir. Ancak anlamlandırıldığı noktada faydalı olup kazanç sağlamaya başlayacaktır. İstatistiksel yöntemler veya yapay zekâ teknikleri büyük verilerin içerdiği bilgileri ortaya çıkarabilmek için önemli katkılar sağlamaktadır.

Bu çalışmada makine öğrenmesi algoritmaları kullanılarak; model ayarlama ve k-kat çapraz doğrulama ile model sonuçları iyileştirilmiş ve aşırı öğrenmenin önüne geçilmiştir. Veri seti eğitim ve test olarak ayrılarak; test verisi üzerinden model performansları değerlendirilmiştir. Ağaç temelli algoritmalar için değişken önem düzeyleri analiz edilmiş

olup birbirleri ile karşılaştırıldığında; modelden modele farklı parametrelerin daha etkili olduğu görülmüştür.

Türkiye'de 21 dağıtım bölgesi bulunmakta olup kaçak kullanan müşterilerin ödeme durumlarının analizi çalışması yaygınlaştırılarak tüm dağıtım şirketleri için kullanılabilir. Bu yaygınlaştırma işlemi sayesinde kaçak faturalarının tahsilatını artırmak hem dağıtım şirketlerine hem de devlete kazanç sağlamış olacaktır. Müşteri ödeme durumu analizinin sadece kredi sağlayan kuruluşların problemi olmaktan çıkıp diğer sektörlerde de uygulanabileceğini gösteren bu çalışma; bundan sonraki benzer çalışmalara yol göstererek literatüre bu anlamda katkı sağlayacaktır.

Gelecekteki çalışmalarda;

- Müşterilere ait 21 elektrik dağıtım bölgesindeki kaçak elektrik kullanım bilgileri, geçmiş ödeme alışkanlıkları ve toplam güncel borçları alınabilir.
- Findeks gibi kredi notu bilgisine ulaşılabilecek veri tabanlarına erişim sağlanabilir.
- İcra süreçlerindeki verilerin kullanımını artırabilmek için UYAP veri entegrasyonu ile müşterinin farklı dosya bilgilerine erişilebilir.
- NVI aracılığıyla kişisel verilerin doluluğunun artırılması sağlanabilir.
- Model hiper parametrelerini belirlerken daha fazla aralık ve deneme yapılabilmesi için daha iyi işlemcili bilgisayarlar üzerinde deneme yapılabilir.
- Modelde kullanılan veri seti farklı oranlarda eğitim ve test verisi şeklinde ayrılabilir (örneğin %80 eğitim verisi, %20 test verisi) ve model tahmin performansındaki olası farklılıklar kıyaslanabilir

KAYNAKLAR

Abdou, H. A., Pointon, J. (2009). Credit scoring and decision making in Egyptian public sector banks. *International Journal of Managerial Finance*, 5(4), 391–406.

Ali, J., Khan, R., Ahmad, N., Maqsood, I. (2012). Random forests and decision trees. *IJCSI International Journal of Computer Science Issues*, 9(5), 272–278.

Altunkaynak, A., Başakın, E. E., Kartal, E. (2020). Dalgacık k-en yakın komşuluk yöntemi ile hava kirliliği tahmini. *Uludağ University Journal of The Faculty of Engineering*, 25(3), 1547-1556.

Ayyadevara, V. K. (2018). *Pro machine learning algorithms*. Hindistan: Apress, 372.

Bahrami, M., Bozkaya, B., Balcısoy, S. (2020). Using behavioral analytics to predict customer invoice payment. *Big Data*, 8(1), 25–37.

Bao, W., Lianju, N., Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128, 301–315.

Barboza, F., Kimura, H., Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83, 405–417.

Batta, M. (2020). Machine learning algorithms - a review. *International Journal of Science and Research (IJSR*), 9(1), 381–386.

Bellotti, A., Brigo, D., Gambetti, P., Vrins, F. (2021). Forecasting recovery rates on non-performing loans with machine learning. *International Journal of Forecasting*, 37(1), 428–444.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123–140.

Charbuty, B., Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(1), 20-28.

Chen, C. H., Chiang, R. D., Wu, T. F., Chu, H. C. (2013). A combined mining-based framework for predicting telecommunications customer payment behaviors. *Expert Systems with Applications*, 40(16), 6561–6569.

Chen, T., Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, San Francisco, California, 785–794.

Cherkassky, V., Mulier, F. (2007). *Learning from data: concepts, theory, and methods.* (İkinci Baskı). Kanada: John Wiley & Sons, 538.

Coenen, L., Verbeke, W., Guns, T. (2022). Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods. *Journal of the Operational Research Society*, 73(1), 191–206.

Deloffre, A., Abdallah, R. N. (2006). Modelling of the debts collection process for service companies. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 12(PART 1), 349–354.

Dumitrescu, E., Hué, S., Hurlin, C., Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178–1192.

F. Hu, Q. Hao. (Editörler). (2012). *Intelligent sensor networks: the integration of sensor networks, signal processing and machine learning*. Amerika: Taylor & Francis, 649.

Fantulin, N. A., Lagravinese, R., Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior and Organization*, 183, 681–699.

Fix, E., Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3), 238-247.

Golbayani, P., Florescu, I., Chatterjee, R. (2020). A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees. *North American Journal of Economics and Finance*, 54, 101251.

Graupe, D. (2013). *Principles of artificial and neural networks.* (3. baskı). Singapur: World Scientific Publishing Co. Pte. Ltd, 364.

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K. (2003). KNN model-based approach in classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2888, 986–996.

Höglund, H. (2017). Tax payment default prediction using genetic algorithm-based variable selection. *Expert Systems with Applications*, 88, 368–375.

İnternet: Cypriano, G. (Şubat, 2018). XGBoost & LightGBM. Web: https://www.slideshare.net/GabrielCyprianoSaca/xgboost-lightgbm#, Son Erişim Tarihi: 07.05.2022.

İnternet: Kabalcı, E. (2014). Yapay sinir ağları (artificial neural networks). Web: https://docplayer.biz.tr/109226146-Yapay-sinir-aglari-artificial-neural-networks.html, Son Erişim Tarihi: 10.01.2022.

İnternet: Serengil, İ.,S., (Ekim, 2020). Rastgele ormanlar sizi karar ağaçlarından nasıl korur. Web: https://bilisim.io/2017/12/29/rastgele-ormanlar-sizi-karar-agaclarindan-nasil-korur/, Son Erişim Tarihi: 20.05.2022.

Kang, M., Tian, J. (2018). *Machine learning : data pre-processing.* In G. P. Michael and K. Myeongsu (Editors), Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things. Hoboken: IEEE Press, pp: 111–130.

Karagül, K. (2014). The classification of the firms traded in İstanbul stock exchange by using support vector machines. *Pamukkale University Journal of Engineering Sciences*, 20(5), 174–178.

Khandani, A. E., Kim, A. J., Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance*, 34(11), 2767–2787.

Khansong, P., Karnjana, J., Laitrakun, S., Usanavasin, S. (2020). *Customer service ımprovement based on electricity payment behaviors analysis using data mining approaches*. 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Phuket, 114–117.

Kim, H. S., Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*, 201(3), 838–846.

Kim, J., Kang, P. (2016.) Late payment prediction models for fair allocation of customer contact lists to call center agents. *Decision Support Systems*, 85, 84–101.

Koh, H. C., Low, C. K. (2004). Going concern prediction using data mining techniques. *Managerial Auditing Journal*, 19(3), 462–476.

Konakoğlu, B. (2020). Çok katmanlı algılayıcı yapay sinir ağı ile jeodezik elipsoidal koordinatların (φ, λ, h) 3 boyutlu global kartezyen koordinatlara (x, y, z) dönüşümü. *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 10(3), 702–710.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), 3–24.

Ksiazek, W., Gandor, M., Pławiak, P. (2021). Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. *Computers in Biology and Medicine*, 134, 104431.

Kuş, İ., Bozkurt Keser, S., Yolaçan, E. (2021). Saldırı tespit sistemlerinde topluluk öğrenme yöntemlerinin kıyaslanması. *European Journal of Science and Technology*, 31, 725–734.

Lappas, P. Z., Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. *Applied Soft Computing*, 107, 107391.

Li, J. P., Mirza, N., Rahat, B., Xiong, D. (2020). Machine learning and credit ratings prediction in the age of fourth industrial revolution. *Technological Forecasting and Social Change*, 161, 120309.

Luong, T. M., Scheule, H. (2022). Benchmarking forecast approaches for mortgage credit risk for forward periods. *European Journal of Operational Research*, 299(2), 750–767.

Ma, S., Fildes, R. (2020). Forecasting third-party mobile payments with implications for customer flow prediction. *International Journal of Forecasting*, 36(3), 739–760.

Madzarov, G., Gjorgjevikj, D., Chorbev, I. (2009). A multi-class SVM classifier utilizing binary decision tree. *Informatica (Ljubljana)*, 33(2), 233–242.

Moscato, V., Picariello, A., Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165, 113986.

Naqa, I. El, Murphy, M. J. (Editors). (2015). *Machine learning in radiation oncology*, Switzerland: Springer Cham, 3–11.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, 4, 51–62.

Nazemi, A., Rezazadeh, H., Fabozzi, F. J., Höchstötter, M. (2022). Deep learning for modeling the collection rate for third-party buyers. *International Journal of Forecasting*, 38(1), 240–252.

Onan, A. (2018). A clustering based classifier ensemble approach to corporate bankruptcy prediction. *Alphanumeric Journal*, 6(2), 365-376.

Onar, S. C., Oztaysi, B., Kahraman, C., Ozturk, E. (2020). Evaluation of legal debt collection services by using Hesitant Pythagorean (Intuitionistic Type 2) fuzzy AHP. *Journal of Intelligent and Fuzzy Systems*, 38(1), 883–894.

Oprea, S. V., Bâra, A. (2021). Machine learning classification algorithms and anomaly detection in conventional meters and Tunisian electricity consumption large datasets. *Computers and Electrical Engineering*, 94, 107329.

Ozan, Ş. (2018). *A case study on customer segmentation by using machine learning methods*. 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, 1–6.

Özay, E. C., Çakıt, E. (2022). M*akine öğrenmesi algoritmalarıyla müşteri ödeme durumu tahmini: elektrik dağıtım sektörü uygulaması*, 2nd International Conference on Applied Engineering and Natural Sciences, Konya, 411.

Patel, H. H., Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, 6(10), 74–78.

Sarkar, D., Bali, R., Sharma, T. (2018). *Machine learning basics*. *Practical machine learning with python*. Hindistan: Apress, 63.

Shin, K. S., Lee, T. S., Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1), 127–135.

Suthaharan, S. (2016). *Machine learning models and algorithms for big data classification*, Switzerland: Springer Science & Business Media, 237-269.

Wang, H., Xu, P., Zhao, J. (2022). Improved KNN algorithms of spherical regions based on clustering and region division. *Alexandria Engineering Journal*, 61(5), 3571–3585.

Wang, W., Chakraborty, G., & Chakraborty, B. (2020). Predicting the risk of chronic kidney disease (CKD) using machine learning algorithm. *Applied Sciences*, 11(1), 202.

West, D. (2000). Neural network credit scoring models. *Computers and Operations Researc*h, 27(12), 1131–1152.

Xia, Y., Liu, C., Li, Y. Y., Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78, 225–241.

Terko, A., Zunic, E., Donko, D., & Dzelihodzic, A. (2019, October). *Credit scoring model implementation in a microfinance context*. In 2019 XXVII International Conference on Information, Communication and Automation Technologies (ICAT), Sarajevo, 1-6.

Yapraklı, T. Ş., Erdal, H. (2016). Firma başarısızlığı tahminlemesi: makine öğrenmesine dayalı bir uygulama. *Bilişim Teknolojileri Dergisi*, 9(1), 21.

Zhang, C., Ma, Y. (Editörler). (2012). *Ensemble machine learning: methods and applications.* New York: Springer Science & Business Media, 329.

Zhao, Y., Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research*, 41(12), 1955–1959.

Zhou, Z. H. (2012). *Ensemble methods foundations and algorithms*. New York: CRC Press, 222.

Zhu, Y., Zhou, L., Xie, C., Wang, G. J., Nguyen, T. V. (2019). Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. *International Journal of Production Economics*, 211, 22–33.

Gazili olmak ayrıcalıktır