T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TOPLULUK ÖĞRENİMİ İLE KREDİ KARTI DOLANDIRICILIK TESPİTİ: KARŞILAŞTIRMALI BİR ANALİZ

Barış AKYÜZ

YÜKSEK LİSANS TEZİ

Endüstri Mühendisliği Anabilim Dalı

Sistem Mühendisliği Programı

Danışman

Prof. Dr. Selin SONER KARA

Haziran, 2022

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

TOPLULUK ÖĞRENİMİ İLE KREDİ KARTI DOLANDIRICILIK TESPİTİ: KARŞILAŞTIRMALI BİR ANALİZ

Barış AKYÜZ tarafından hazırlanan tez çalışması 30.06.2022 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı, Sistem Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Selin SONER KARA
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Selin SONER KARA, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Selçuk ÇEBİ, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Engin PEKEL, Üye
Hitit Üniversitesi _____

Danışmanım Prof. Dr. Selin SONER KARA sorumluluğunda tarafımca hazırlanan Topluluk Öğrenimi ile Kredi Kartı Dolandırıcılık Tespiti: Karşılaştırmalı Bir Analiz başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Barış AKYÜZ

İmza

Aileme

ve

sevgili eşime

TEŞEKKÜR

Bu çalışma literatürde makine öğrenmesi alanında ele alınan başlıaca konulardan biri olan dolandırıcılık tespitine yönelik kapsamlı bir karşılaştırma kazandırmaktadır. Çalışmada Türkiye'den güncel kredi kartı işlem verisi kullanılarak özellik ekleme yoluyla ve topluluk öğrenmesi yöntemleriyle dolandırıcılık tespitinde kullanılabilecek katkılar araştırılmıştır. Topluluk öğreniminde birleştirme yöntemlerinin etkisine ek olarak, model seçimi için alternatif öneriler getirilmiştir. Elde edilen bulgular farklı yaklaşımlar üzerinden karşılaştırmalı olarak değerlendirilmiştir.

Çalışma süresince beni yönlendiren ve desteğini esirgemeyen tez danışmanım Prof. Dr. Selin SONER KARA'ya ve uygulama kapsamında kullanılan verilerin toplanmasında yardımcı olan Sayın Furkan AKARSLAN Bey'e ve Mehmet ERDOĞAN Bey'e yardımları için teşekkür ederim.

Barış AKYÜZ

İÇİNDEKİLER

SİMGE LİSTESİ vii

KISALTMA LİSTESİ viii

ŞEKİL LİSTESİ ix

TABLO LİSTESİ x

ÖZET xi

ABSTRACT xiii

1 GİRİŞ 1

1.1 Literatür Özeti ... 1
1.1.1 Sınıflandırma Çalışmaları ... 4
1.1.2 Tahmin Modeli Birleştirme Uygulamaları ... 7
1.1.3 Tahmin Eşik Değeri Çalışmaları ... 8

1.2 Tezin Amacı ... 9

1.3 Orijinal Katkı ... 10

1.4 Dolandırıcılık Tespitindeki Zorluklar ... 10

2 KULLANILAN YÖNTEMLER 12

2.1 Veri Dengeleme Yöntemleri ... 12
2.1.1 Yeniden Örnekleme ... 13
2.1.2 Çapraz Doğrulama ... 14
2.1.3 Ağırlıklandırma ... 15

2.2 Sınıflandırma Yöntemleri ... 15
2.2.1 Lojistik Regresyon ... 16
2.2.2 K-En Yakın Komşu ... 17
2.2.3 Karar Ağaçları ve Rastgele Orman Algoritması ... 18
2.2.4 Destek Vektör Makineleri ... 20
2.2.5 Yapay Sinir Ağları ... 21
2.2.6 Model Birleştirme ... 22

3 UYGULAMA 24

3.1 Veri Örneğinin Açıklanması ... 24

3.2 Tahmin Modeli ve Performans Göstergelerine Bakış ... 27
3.2.1 Tahmin Modelinin Kurulması ... 28
3.2.2 Anahtar Performans Göstergelerinin Seçilmesi ... 29

3.3 Üretilen Yardımcı Özellikler ... 32
3.3.1 Müşteri RFM Analizi ... 32
3.3.2 Harcama Risk Puanı Analizi ... 33

3.4 Tahmin Modeli Uygulamaları ... 36
3.4.1 Tekil Makine Öğrenmesi Modelleri ... 38

3.4.2 Tekil Modellerin Birleştirilmesi 40
3.4.3 Tahmin Eşik Değeri Uygulaması 46
3.5 Temel Bileşen Analizi 52
4 SONUÇ VE ÖNERİLER **57**
4.1 En İyi Tahmin Modeli 58
4.1.1 Dengeli Yaklaşım 58
4.1.2 Yüksek Risk İştahlı Yaklaşım 60
4.1.3 Düşük Risk İştahlı Yaklaşım 62
4.2 Veri Özelliklerinin Yorumlanması 64
4.3 Öneriler 66
KAYNAKÇA **67**
TEZDEN ÜRETİLMİŞ YAYINLAR **72**

SİMGE LİSTESİ

k	Alt küme sayısı
E_i	Azınlık sınıfına ait gözlem vektörü
β_i	Bağımsız değişken
c	Ceza parametresi
F_1	F skoru
p	İstenilen durumun gerçekleşme olasılığı
γ	Kernel fonksiyonu katsayısı
d	Mesafe
A_{ij}	Öğrenme verisi aday gözlem grubu
n	Örnek gözlem boyutu
δ	Rassal sayı
β_0	Sabit sayı

KISALTMA LİSTESİ

ABD	Amerika Birleşik Devletleri
FN	False Negative
FP	False Positive
GBM	Gradient Boosting Machines
KNN	K–Nearest Neighbours
LASSO	Least Absolute Shrinkage And Selection Operator
MCC	Merchant Category Code
RFM	Recency Frequency Monetary
SMOTE	Synthetic Minority Over-Sampling Technique
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

Şekil 1.1 Literatürden alınan çalışmaların yıllara göre dağılımı 3
Şekil 2.1 Karşılaştırma matrisi 16
Şekil 2.2 Örnek KNN modeli gösterimi 18
Şekil 2.3 Örnek karar ağacı 19
Şekil 2.4 Destek vektör makinelerinde marjin çizimi 20
Şekil 2.5 Yapay sinir ağının bileşenleri 22
Şekil 3.1 Gözlemlerin dağılımı 26
Şekil 3.2 Değişken katkıları 35
Şekil 3.3 Uygulama özeti 37
Şekil 3.4 İyimser, kötümser ve çoğunluk oylama birleştirme yöntemleri 41
Şekil 3.5 Yumuşam oylama, ağırlıklı ortalama ve uzlaşma yöntemleri 42
Şekil 3.6 Üst katman birleştirme yöntemleri 43
Şekil 3.7 Birleştirme yöntemlerinin F1 skoru performansı 44
Şekil 3.8 Birleştirme yöntemlerinin Kesinlik performansı 45
Şekil 3.9 Birleştirme yöntemlerinin Belirlilik performansı 45
Şekil 3.10 Hibrit seçimde uzlaşma yönteminin eşik değerleri 49
Şekil 3.11 Performans seçimde ağırlı ortalama yönteminin eşik değerleri 50
Şekil 3.12 Temel seçimde yumuşak oylama yönteminin eşik değleri 51
Şekil 3.13 Özelliklerin temel bileşenler eksenlerinde dağılımı 54
Şekil 3.14 Örnek gözlemlerin temel bileşen eksenlerinde dağılımı 55
Şekil 4.1 Öğrenme veri boyutlarının test sonuçları 65

TABLO LİSTESİ

Tablo 1.1 Referanslarda kullanılan öğrenme yöntemleri 5
Tablo 3.1 Uygulama verisinin detayı 24
Tablo 3.2 Veriye ait temel özellikler 25
Tablo 3.3 Temel özelliklerin açıklaması 25
Tablo 3.4 Öğrenme veri boyutu kombinasyonları 27
Tablo 3.5 En iyi aday öğrenim modellerinin sonuçları 30
Tablo 3.6 F_1 skoruna göre test sonuçlarının karşılaştırılması 31
Tablo 3.7 Müşteri harcamalarına ait yardımcı özellikler 33
Tablo 3.8 İşyeri detayına ait yardımcı özellikler 34
Tablo 3.9 Kategori detayına ait yardımcı özellikler 34
Tablo 3.10 Yardımcı özelliklerin tahmin performansına etkisi 35
Tablo 3.11 Farklı öğrenme algoritmalarının tahmin sonuçları 38
Tablo 3.12 Olasılık eşik değerlerinin tahmin sonucuna etkisi 46
Tablo 3.13 Rastgele orman modelinde eşik değeri enterpolasyonları 47
Tablo 3.14 Temel bileşenlerin açıklama yüzdeleri 53
Tablo 3.15 Özelliklerin temel bileşenlere etkisi 56
Tablo 4.1 Dengeli yaklaşıma göre oluşan sıralama 59
Tablo 4.2 Yüksek risk iştahı yaklaşımında oluşan sıralama 61
Tablo 4.3 Düşük risk iştahlı yaklaşımda oluşan sıralama 63

ÖZET

Topluluk Öğrenimi ile Kredi Kartı Dolandırıcılık Tespiti: Karşılaştırmalı Bir Analiz

Barış AKYÜZ

Endüstri Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Selin SONER KARA

Kredi kartı hırsızlık vakalarını önleme çalışmaları giderek artan kart kullanım sayılarına bağlı olarak önemini sürdürmeye devam etmektedir. Bu çalışmanın amacı, kredi kartı müşterilerine ait mevcut işlemleri analiz ederek çok sayıda normal işlemin arasından az sayıdaki hırsızlık işlemlerini doğru tespit etmeyi sağlayan etkin tahmin modelini oluşturmaktır. Bu amaçla literatürde sıklıkla tercih edilen Makine Öğrenmesi algoritmaları kullanılarak veri hazırlığından tahmin modeli oluşturmaya, performans parametresi seçiminden model gruplamaya kadar farklı uygulama adımları ele alınmıştır. Öncelikle işlem verisini hazırlamak ve başlangıç noktasını görmek amacıyla bir tahmin modeli üretilmiştir. Böylece öğrenme verisinde farklı gözlem boyutu sayısı ve hırsızlık oranı değerlerinin test verisindeki tahmin sonuçlarına etkisi baz model aracılığıyla test edilerek ortaya çıkarılmıştır. Tahmin sonuçlarını yorumlamak amacıyla, problemin gözlem dengesizliğine uygun olduğundan dolayı F_1 skoru değeri birincil performans göstergesi olarak belirlenmiştir. Sonrasında uygulama verisine ait müşteri numarası, harcama tutarı, işlem tarihi gibi temel özellikler kullanılarak haftalık, aylık gibi çeşitli periyotları kapsayan müşteri ve iş yeri bazlı yardımcı özellikler

üretilmiştir. Üretilen yardımcı özelliklerin ardından literatürde öne çıkan makine öğrenmesi algoritmaları sınıflandırma tahmini amacıyla önce teker teker kullanılmış, ardından tahmin sonuçları incelendikten sonra önerilen risk algısı bazlı seçim kriterine göre üç farklı şekilde gruplama için seçim yapılmıştır. Seçilen tekil tahmin modellerinin gruplanması ise çeşitli yollarla yapılmıştır. İyimser, kötümser, ortalama gibi temel yaklaşımlara ek olarak ağırlıklı ortalama, uzlaşma (Hurwicz kriteri), yumuşak oylama (soft voting), üst katman öğrenme gibi ek yaklaşımlar değerlendirilmiştir. Gruplanmış modellerin tekil modeller ile elde edilen başarı seviyesini arttırdığı görülmüştür.

Anahtar Kelimeler: Makine Öğrenmesi, kredi kartı hırsızlık tespiti, sınıflandırma.

ABSTRACT

Credit Card Fraud Detection Using Ensemble Learning: A Comparative Analysis

Barış AKYÜZ

Department of Industrial Engineering

Master of Science Thesis

Advisor: Prof. Dr. Selin SONER KARA

Prevention of credit card fraud cases continues to be important due to the increasing number of card usage. The aim of this study is to create an effective classification model that enables to accurately detect a small number of fraud transactions among a large number of normal transactions by analyzing the transactions of credit card customers. For this purpose, different application steps like data preparation, model training, selection of performance parameters and ensemble models are discussed by using frequently preferred Machine Learning algorithms in the literature. First of all, a classification model is created to prepare the transaction dataset properly and to specify the starting point of the study. Furthermore, the effect of the number of different observation sizes and fraud rates of the training data was investigated by testing through the base model. Thus, test results of using different sizes and fraud rates in training data was revealed with the help of the base model. Since it is compatible with the class imbalance problem, F_1 score is selected as the primary performance indicator in order to evaluate the classification results. Then, additional periodic customer and merchant-based features are created by using the basic features from the

application data such as customer number, merchant number or transaction amount. Thus, the most popular machine learning algorithms are used individually and then three different ensembles are selected according to the new proposed risk appetite-based selection criteria. Also, the ensemble models are prepared with various ways. In addition to the basic approaches such as optimistic, pessimistic and average methods, additional approaches such as weighted average, realism (Hurwicz criterion), soft voting and top layer models are also used. It has been observed that ensemble models are increased the success of individual models.

Keywords: Machine Learning, credit card fraud detection, classification.

1
GİRİŞ

Kartla ödeme sisteminin küresel çapta pazar büyüklüğü 2021 yılı itibariyle 251 milyar doları geçerken, 2026 yılında ise 400 milyar doları bulması beklenmektedir. Türkiye'de 2021 yılında kartla yapılan ödeme tutarı bir önceki yıla göre %49 artarak 1.71 trilyon Türk Lirası olmuştur. Bununla birlikte toplam 291 milyon adet ile Avrupa'da en çok kart kullanılan ülke Türkiye'dir [1]. Diğer yandan artan pazar büyüklüğü ile birlikte kart dolandırıcılığı miktarı da dünya çapında 29 milyar dolar düzeyine ulaşmıştır. Bu durum sadece müşterilere değil, iş yerlerine ve bankalara da kayıp yaşatan önemli bir sorun olmaya devam etmektedir.

1.1 Literatür Özeti

Kredi kartı dolandırıcılığı ya da hırsızlığı olarak isimlendirilen bu sorun, internetin ve sosyal medyanın sağladığı ulaşılabilirlik sayesinde günden güne daha sık gündeme gelmektedir. ABD toplam vaka sayısında tüm dünyada ilk sırayı almıştır. Binden fazla katılımcı yapılan bir ankete göre 2021 yılında her 100 Amerikalıdan 60'ı kredi kartı dolandırıcılık vakası yaşamıştır [2]. Öte yandan gerek makro düzeydeki uluslararası otoriteler ve devletler gerek de finans kuruluşları dolandırıcılık vakalarına karşı çeşitli mücadele yolları sürdürmektedir. Bu mücadele yolları temel olarak önleme ve yakalama olmak üzere ikiye ayrılır [3].

Önleme faaliyetleri temelde hırsızların müşterilere ait kişisel verilere ve kart bilgilerine ulaşmasını engellemeye yöneliktir. Mevcut önleme faaliyetlerine örnek olarak, iki faktörlü doğrulama ve sanal kart uygulaması verilebilir. İki faktörlü doğrulama ana uygulama parolasına ek olarak genellikle e-posta veya mesaj yoluyla tek seferlik ikinci bir şifre istenmesi ya da mobil uygulama ile onay talep edilmesi yoluyla gerçekleştirilir. Böylelikle ikinci bir güvenlik katmanı elde edilmeye çalışılır. Sanal kart uygulaması ise, mevcut ana müşteri hesabın altında fiziksel olarak var olmayan ve limit miktarı kontrol altında tutulan yeni kart numaraları üretilmesidir. Böylelikle internet aracılığıyla alışveriş yaparken esas

kart bilgileri ve toplam limit kullanılmamış olur. Ancak burada iki örneği açıklanan önleme faaliyetleri teknolojinin verdiği imkanlar sayesinde günden güne gelişmekte olsa da temassız ödeme gibi pazara sunulan yeni uygulamaların potansiyel açıkları ya da insan faktörünün etkisiyle hiçbir zaman kesin başarıya ulaşamamaktadır.

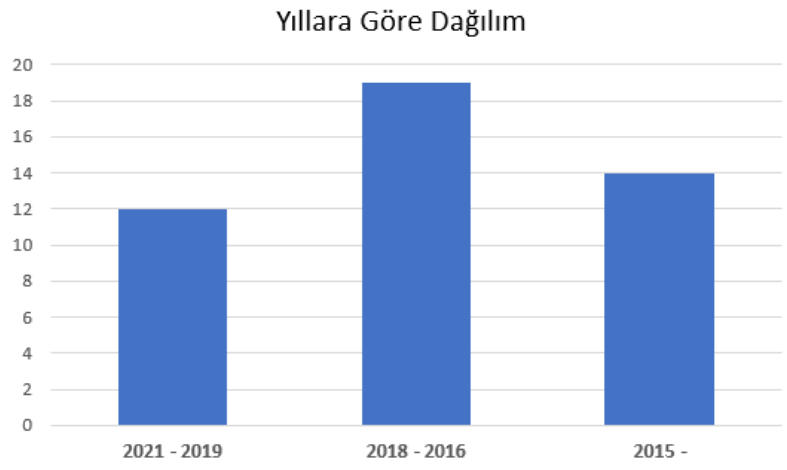
Kart dolandırıcılığını yakalama faaliyetleri ise, gerçekleşen veya gerçekleşmekte olan harcamanın müşteriye ait olup olmadığını önceden tahmin etmeye yöneliktir. Geleneksel tahmin şekli uzman kişilerce belirlenen kural bazlı modellerdir. Bu modellerde geçmişte karşılaşılan hırsızlık vakalarının benzerleri olabilecek senaryolara uyan işlemler uyarı üretir. Örneğin geçmiş durumlara bakıldığında başarılı gerçekleşen düşük tutarlı bir işlemin hemen arkasından yüksek miktarlı bir işlem talebi potansiyel bir dolandırıcılık şüphesini doğurmaktadır. Ancak diğer yandan bilinen senaryoların aksine hırsızların yakalanmaktan kaçınmak adına normale en yakın harcamaları taklit etmeye çalıştıkları bilinmektedir. Dolayısıyla kural bazlı modellerin uzman bilgisine bağımlı olması ve yeni yöntemlere ayak uyduramayan statik öğrenme yapıları kısa sürede yetersiz kalmaya başlamıştır. Dolandırıcıların hızla değişen stratejilerini takip etmeye çalışan uzman kontrolünde üretilen senaryolar ile güncel işlemleri yakalayana kadar fazlasıyla zaman kaybedilmektedir. Bu durum ise bankaları daha etkin tahmin modellerine yöneltmiş ve dolandırıcılık tespiti alanında Makine Öğrenimi tekniklerinin kullanımı yaygınlaşmıştır.

Dolandırıcılık tespiti amacıyla uygulanan Makine Öğrenmesi çalışmaları temelde geçmiş harcama işlemlerinin kullanılmasıyla yeni işlemlerin normal ve dolandırıcılık kategorilerinden hangisine daha çok benzediğini tahmin etmeye yöneliktir. Bu işlem birçok farklı teknik uygulanarak gerçekleştirilebilmektedir. Zaman Serisi Analizi, Kümeleme ve Sınıflandırma teknikleri öne çıkan tekniklere örnek olarak verilebilir. Bununla birlikte, dolandırıcılık tespiti probleminin bazı karakteristik özellikleri Makine Öğrenmesi uygulamalarını zorlaştırmaktadır.

Makine Öğrenmesi kavramı, dolandırıcılık tespiti konusunda olan uygulamaların dışında oldukça geniş ve detaylı teknikleri içeren öğrenme yöntemidir. Temelde istenen bilginin verinin içinde etiketli olup olmamasına bağlı olarak gözetimli ve

gözetimsiz öğrenme şeklinde ikiye ayrılır. Öğrenme yöntemleri ise kullanım amacına göre ise tahmin, kümeleme ve boyut indirme olmak üzere basitçe üç başlıkta özetlenebilir. Boyut indirme büyük veri havuza ait özellikleri minimum açıklama kaybıyla daha az özellik boyutuyla yansıtabilmeyi amaçlar. Kümeleme ise algoritmanın aranan özellik değerini görmeden birbirine benzer gözlemleri bir araya getirebilmesidir. Tahmin ise aranan değerin kategorik olup olmamasına göre sınıflandırma ve regresyon olarak ikiye ayrılmaktadır [4,6].

Literatürde dolandırıcılık tespiti amacıyla kullanılan Makine Öğrenmesi tekniklerine yakından bakıldığında hem gözetimli hem de gözetimsiz öğrenme yollarının kullanıldığı görülür [7]. Birbiriyle kıyaslandığında ise gözetimsiz öğrenme tekniklerinin daha az ele alındığı ve gözetimli öğrenme yöntemlerinin başta sınıflandırma teknikleri olmak üzere daha sık kullanıldığı söylenebilir. Gözetimsiz öğrenme tercihinde harcama işlemlerinin dolandırıcılık olup olmadığının ayrımı bilinmediğinden tahmin referansını normal olan işlemlerin özelliklerine dayanarak yapılır. Gözetimli öğrenme kullanıldığında ise dolandırıcılık işlemlerinin özelliklerine ait bilgiler de tahmin kararını oluşturmada kullanılmış olur. Aşağıda öne çıkan çalışmalar ışığında literatür bulguları açıklanmıştır. Ayrıca her iki öğrenme yönteminin konu özelinde detaylı taramasını sunan [5] ve [8] ile geniş kapsamlı bilgi edinilebilir.



Şekil 1.1 Literatürden alınan çalışmaların yıllara göre dağılımı

İncelenen çalışmaların yıllara göre dağılımı Şekil 1.1 ile gösterilmiştir. Dolandırıcılık tespiti alanı özelinde bakıldığında, hırsızların yakalanmamak için

dinamik olarak strateji değiştirmesinden dolayı güncel verilerin kullanıldığı çalışmaların geçmiş verileri kullananlara göre yeniden uygulanabilirliği daha yüksektir. Bu bağlamda, güncel çalışmaların elde ettiği bulgular 2019 yılından itibaren yayınlanmış 12 çalışma ile irdelenmiştir. Ayrıca seçilen çalışmalar arasında 2016 ile 2018 yılları arasında yapılanlar çoğunluktadır. Böylece güncel çalışmalara ağırlık verilmiştir.

1.1.1 Sınıflandırma Çalışmaları

Sınıflandırma amacıyla ortaya konulmuş çalışmalar genel anlamda aynı grupta toplanabilirse de detaya inildiğinde kullanılan öğrenme algoritmaları, veri kaynağı çeşitleri veya performans ölçütleri gibi temel noktalarda birbirlerinden tamamen ayrılabilmektedir. Tablo 1.1 incelendiğinde, kullanılan öğrenme algoritmaları özelinde literatürden seçilen çalışmalar gruplanmıştır. Buna göre, en sık başvurulan yöntemlerin arasında Destek Vektör Makineleri (SVM), Yapay Sinir Ağları, Rastgele Orman ve Lojistik Regresyon gelmektedir. Ayrıca, tekil öğrenme algoritmalarının kullanıldığı model birleştirme yöntemlerini içeren çalışmalar da incelenmiştir.

Destek Vektör Makineleri ile sınıflandırmada, kategoriler arasındaki uzaklığı en üst düzeye çıkarmak için destek noktaları aranır. Gözlemler vektörün hangi tarafına düştüklerine bağlı olarak ait oldukları kategori tahmin edilir. Sınıflandırma problemlerini çözmede etkilidir [8]. Yapay Sinir Ağları (YSA), birbirine bağlı düğümler arasında katmanları olan ve katmanlar arasında hesaplamaların birbirlerinin girdisi olarak kullanıldığı insan sinir sistemine benzeyen bir öğrenme modelidir. YSA, diğer algoritmalara kıyasla daha yüksek işlem kapasitesi gerektirir. Rastgele Orman algoritması, çok sayıda karar ağacı sonucunu gruplayarak tahmin sonucuna ulaşır. Dengesiz dağılma sahip verilerde çalışmaya uygundur ve karmaşık modellere göre daha hızlı çalışır. Lojistik Regresyon, kategorik değişkenin değerlerini 0-1 arasında olasılık değeri olarak hesaplayarak, sınıflandırma yapma imkanı veren bir yöntemdir.

Tablo 1.1 Referanslarda kullanılan öğrenme yöntemleri

Öğrenme Algoritması	Referanslar
Destek Vektör Makineleri (SVM)	[9,12,20,21,22,24,26,27,28,35,36]
Yapay Sinir Ağları	[10,11,13,19,21,31,32,33]
Rastgele Orman	[12,14,16,17,20,21,22,23,24,26,27,29,34]
Lojistik Regresyon	[14,15,16,17,18,20,24,22,23,26,28]
Model Birleştirme Yöntemleri	[12,19,20,21,22,23,24]

Wei ve Yuan 2012 yılında %2 dolandırıcılık oranına sahip 30,000 gözlemden oluşan gerçek işlem verisi kullanarak SVM algoritmasında parametre optimizasyonu önermiştir. Önerilen model karar ağacı bazlı ID3 algoritmasını ile karşılaştırılmış ve başarılı sonuç verdiği görülmüştür [9].

Srivastava vd., 2016 yılındaki çalışmalarında iş yeri bilgisi odaklı bir YSA modeli önermiştir [10]. Önerilen yaklaşımda müşteriye ait meslek bilgisi, kazanç ve fatura adresi, konum, mevcut ile bir önceki işlem arasındaki zaman farkı, para birimi türü ve tutar gibi bilgileri kullanılmıştır.

Zakaryazad ve Duman, 2015 yılındaki çalışmalarında YSA temelli maliyet katsayısı eklenmiş bir algoritma tanıtmıştır. Model öğrenme algoritmasındaki yenilik sayesinde pozitif gözlemler için daha yüksek ağırlıklı hale getirilmiştir [11]. Sonuçlar üç farklı test veri setinde incelenmiş ve dolandırıcılık işlemlerine ait maliyetlerin azaltılma başarısına göre kıyaslanmıştır. Sonuçta standart ölçütlere göre önerilen model öndeyken maliyet ölçütüne göre Karar Ağaçları ve diğer modellerin gerisinde kalmıştır.

Kültür ve Çağlayan 2016 yılında yayınlanan çalışmada Rastgele Orman algoritmasının da aralarında bulunduğu 6 farklı yöntemi kullanarak 152,706 gözlemden oluşan uygulama verisinde %0.5 dolandırıcılık oranı ile bir model

önerisi getirmiştir [12]. Elde edilen sonuçlar farklı risk stratejilerine göre yorumlanmış ve Alarm Oranı gibi farklı performans ölçütleriyle de desteklenmiştir.

2019 yılında yayınlanan çalışmada Kim vd. mevcut kural bazlı modele alternatif olarak YSA ile derin öğrenme önerisi sunmuştur [13]. Model öğrenimi için parametre optimizasyonu yapılmış, uygulanan farklı denemeler sonuçları ile birlikte paylaşılmıştır. Çalışmada model performansını test etmek adına gerçek hayattan temin edilen %0.04 dolandırıcılık oranını sahip test verisi kullanılmıştır.

Lojistik Regresyon ve Karar Ağaçları algoritmalarının kullanıldığı 2019 yılına ait bir çalışmada Dornadula ve Geetha, literatürde sıklıkla kullanılan örnek veri üzerinde elde edilen bulguları paylaşmıştır [14]. Çalışmada öne çıkan bir diğer nokta ise, veri özelliklerini zenginleştirebilmek adına müşteriler için ortalama işlem tutarı, minimum ve maksimum harcama tutarı gibi özellikler hesaplanmıştır. Sonuçları yorumlamak için MCC (Matthews Correlation Coefficient) göstergesi ile kullanılmıştır. Buna göre en yüksek sonuç veren algoritma 0.84 MCC değeri ile Karar Ağaçları olmuştur.

Thennakoon vd. 2019 yılında yaptıkları dolandırıcılık tespiti çalışmasında, 917,781 gözlemden oluşan uygulama verisi çeşitli özelliklerine göre dört gruba ayrılmıştır [15]. Bu özellikle şu şekildedir; riskli iş yeri olarak bilinen yerlerden yapılmış işlemler, 100$ üzerinde yapılan işlemler, bilinmeyen internet adreslerinden gerçekleştirilen işlemler ve riskli internet adreslerinden gerçekleştirilen işlemlerdir. Bu veri grupları, nümerik ve kategorik olmak üzere iki ayrı şekilde düzenlenmiştir. Sonuçta farklı veri gruplarına ait tahmin sonuçları F_1 ve MCC göstergeleri ile karşılaştırılmıştır.

2020 yılında yayınlanan çalışmada Khatri vd. performans ölçütlerinin arasına model kurma için harcanan zamanı da dahil etmiştir [16]. Uygulama verisi olarak literatürde sıklıkla çalışılan 2013 tarihine ait örnek veri kullanılmıştır. Buna göre en yavaş model Rastgele Orman algoritması ile kurulmuşken, en hızlı modeller ise Lojistik Regresyon ve Naive Bayes ile kurulmuştur.

Bahnsen vd. 2016 yılında yaptıkları çalışmada ek bir performans göstergesi olarak dolandırıcılık işlemlerinin tespiti sayesinde yapılan maliyet tasarrufu değerini

kullanmayı tercih etmiştir [17]. Çalışmada öne çıkan bir diğer durum ise müşteri harcama istatistiklerini içeren yardımcı özelliklere ek olarak, harcama detayında işlem saati bilgisine göre yeni özellikler üretilmiş kullanılmış olmasıdır. Sonuçta yeni eklenen özelliklerin tasarruf kriterine göre 16% iyileştirme sağladığı görülmüştür.

2017 yılında Awoyemi vd. tarafından yayınlanan araştırmada örnek veri kullanılmış ve performans ölçütleri arasında Dengelenmiş Doğruluk (Balanced Accuracy) değerine yer verilmiştir [18]. Örnek verideki gözlem dengesizliğinin üstesinden gelmek adına öne sürülen hibrit yöntemle, dolandırıcılık gözlemlerinin sayısını sentetik olarak arttırarak ve normal gözlemleri azaltarak iki farklı şekilde dengelenme yapılmaya çalışılmıştır.

1.1.2 Tahmin Modeli Birleştirme Uygulamaları

Forough ve Momtazi, 2020 yılında yayınlanan çalışmalarında farklı YSA algoritmalarını Çoğunluk Oylama (Majority Voting) ve Üst Katman Öğrenme (Top Layer Model) yollarıyla birleştirerek örnek veriler üzerinde test etmiştir. Elde edilen sonuçlara ait F_1 skorları incelendiğinde Üst Katman Öğrenme yoluyla yapılan birleştirmelerin Çoğunluk Oylamaya göre daha iyi sonuçlar verdiği görülmüştür [19].

Mishra ve Ghorpade, 2018 yılında yaptıkları çalışma ile SVM, Lojistik Regresyon ve Karar Ağacı gibi algoritmaların bulunduğu farklı öğrenme yöntemlerini kullanarak ayrı ayrı tekil modelleri ve bir birleştirilmiş modeli karşılaştırmıştır [20]. Gözlem dengesizliğine karşı normal işlemlerin oranı azaltılması denenmiştir. Elde edilen sonuçlara göre Rastgele Orman algoritması diğer tekil modelleri ve birleştirilmiş modeli geçerek en iyi sonucu veren yöntem olmuştur.

Randhawa vd. 2018 yılında gerçekleştirdikleri çalışmada model birleştirme uygulaması için Rastgele Orman, SVM ve Lojistik Regresyon algoritmalarını seçmiş, Çoğunluk Oylama ve Adaboost Üst Katman Modeli yöntemlerini kullanarak birleştirmiştir [21]. 2018 yılında yayınlanan çalışma, literatürde sıklıkla kullanılan örnek veri seti ile kurulmuş, bir önceki yıla ait Malezya'daki müşterilerin gerçek

verileri test edilmiştir. MCC göstergesi ile birleştirme yöntemleri kıyaslandığında Çoğunluk Oylama yöntemi en iyi sonucu vermiştir.

Güven tarafından 2021 yılında yapılan çalışmada farklı veri hazırlığı yöntemleri incelenmiştir [22]. Buna göre, standart olarak öğrenme algoritmasına tüm özelliklerin dahil edilmesine alternatif olarak özellik seçimi için LASSO ve özellik çıkarımı için Temel Bileşen Analizi (Principal Component Analysis) önerilmiştir. Tekil öğrenme modellerini oluşturmak adına Rastgele Orman, Lojistik Regresyon, SVM, K – En Yakın Komşu (KNN) ve Naive Bayes algoritmaları olmak üzere 6 farklı yöntem tercih edilmiştir. Tekil modellerin birleştirmesi için de Yumuşak Oylama ve Çoğunluk Oylama yöntemleri kullanılmıştır. Önerilen modelleri test etmek amacıyla %2 dolandırıcılık gözlem oranına sahip veri seti kullanılmıştır. Sonuçlar $F1$ skoruna göre kıyaslandığında Çoğunluk Oylama yöntemi ile yapılan birleştirme modeli hem Yumuşak Oylama birleştirme modelinden hem de tekil modellerden daha iyi sonuç vermiştir.

Bagga vd. tarafından 2020 yılında yapılan araştırmada model birleştirme amacıyla Üst Katman Rastgele Orman Modeli kullanılmıştır [23]. Tekil modeller arasında Rastgele Orman, Lojistik Regresyon, KNN ve Adaboost bulunmaktadır. Önerilen modelin performansı 2013 yılına ait sık kullanılan örnek veri ile test edilmiştir. Elde edilen sonuçlar F_1 skoru ile kıyaslandığında birleştirilmiş model tekil modellerden daha iyi sonuç vermiştir.

Zareapoor ve Shamsolmoali, 2015 yılında yaptıkları çalışmada uygulama verisini farklı dolandırıcılık oranlarına göre 4 parçaya ayırmıştır [24]. En yüksek %20 dolandırıcılık oranını ile en düşük %3 dolandırıcılık oranı arasında değişen 4 veri seti için oluşturdukları SVM ve KNN tekil modeller, topluluk karar ağaçları bazlı Bagging yöntemi ile birleştirilmiştir. Sonuçlar incelendiğinde dolandırıcılık oranının azalmasıyla performans göstergelerinin başarılarının azaldığı görülmüştür. Bu çalışmadaki bir diğer önemli bulgu ise, önerilen birleştirilmiş modelin tekil modellere göre tüm durumlarda daha iyi sonuç verebilmiş olmasıdır.

1.1.3 Tahmin Eşik Değeri Çalışmaları

Sa vd. tarafından 2018 yılında yayınlanan çalışmada, önerilen tahmin modelinin eşik değerine göre gösterdiği F_1 skoru performansı karşılaştırılmıştır [25]. Elde edilen test sonuçları Brezilya'dan gerçek işlem verileriyle ölçülmüştür. Uygulama verisinde kullanılan dolandırıcılık oranı %2 olarak seçilmiştir. Sonuçta önerilen tahmin modelinin varsayılan eşik değerine kıyasla 0.3 düzeyi ve altındaki tahmin eşik değerlerinde daha iyi sonuç verdiği gösterilmiştir. Eşik değeri uygulaması sayesinde elde edilen en yüksek F_1 skoru değeri %80 düzeyinde ölçülmüştür.

Kullanılan veriye ait özelliklerin katkı düzeylerinin paylaşıldığı 2017 yılına ait Figueira vd. tarafından yapılan çalışmada, işlemin tutarı ve müşterinin ilk işlemi ile mevcut işlemi arasında geçen süreyi gösteren özellikler öne çıkmıştır [26]. Diğer yandan sipariş saatini gösteren özellik ise en son sırada yer almıştır. Çalışmada gösterilen bir diğer önemli bulgu, tahmin sonuçlarının karar seviyesindeki olasılık eşik değerine göre değişmiştir. Yazarlar iki farklı yaklaşım üzerinden eşik değerini optimize etmeyi denemiştir.

Srivastava vd. tarafından yapılmış 2016 yılındaki çalışmada iki değerden oluşan klasik tahmin sonuçları yerine, tahmin olasılık değerine göre 0.0-0.2 arasında ise işlem normal, değer 0.21-0.44 arasında ise işlem şüpheli, 0.45-0.69 arasında ise riskli, 0.7'den büyükse işlem dolandırıcılık olarak etiketlenmiştir [11]. Başka bir deyişle, varsayılan 0.5 seviyesine alternatif olarak 0.7 değeri tahmin sonucunu belirleyen kritik düzey olarak seçilmiştir.

Khatri vd., 2020 yılında yaptıkları çalışmada varsayılan eşik değerine alternatif olarak 0.4 düzeyini kullanmayı denemiştir [17]. Elde edilen bulgular, eşik değerinin azalmasıyla Belirlilik göstergesinin iyileştiğini ancak Kesinlik düzeyinin kötüleştiğini göstermektedir. Yapılan karşılaştırma sonucunda en iyi dolandırıcılık tahmin algoritması olarak Karar Ağaçları önerilmiştir.

1.2 Tezin Amacı

Bu çalışmanın amacı, literatürden öğrenilen güncel araştırmaların üzerine gerçek veri örneği kullanarak kapsamlı bir analiz yapmak ve makine öğrenmesinde topluluk öğreniminin örneklerini çeşitlendirerek farklı yaklaşımları literature

kazandırmaktır. Böylelikle mevcut çalışmaların en büyük boşluklarından biri olan güncel veriye ait uygulama eksikliği giderilmesi hedeflenmektedir. Ayrıca güncel bir uygulama ortaya koymanın yanında, bir diğer kritik başlık olan dolandırıcılık tespitinde performans ölçümü amacıyla literatürde ele alınmış göstergeler kullanılarak karşılaştırması yapılacak ve güncel uygulamada öne çıkan yanlarıyla anahtar performans göstergeleri belirlenecektir.

İlave olarak bu tez çalışması boyunca hangi makine öğrenmesi uygulama adımının dolandırıcılık tespitine ne ölçüde fayda sağlayabileceği de incelenecektir. Bu amaçla, veri temizliği, dengesiz gözlem dağılımıyla mücadele yolları, model parametrelerini belirleme ve topluluk öğreniminde birleştirme yöntemleri gibi farklı uygulama adımları detaylı bir şekilde alınacaktır. Bu sayede, hızlı ve etkin bir tahmin modelinin en temel gereksinimleri ortaya çıkarılmaya çalışılacaktır.

1.3 Orijinal Katkı

Bu tez çalışmasının literatüre katkısı, veri hazırlığı, tahmin modeli, model birleştirme ve eşik değeri uygulamasını aynı anda uygulayan, klasik yaklaşımlardan farklı olarak risk odaklı alternatiflerle birbirleriyle etkileşimini ölçen ilk çalışma olmasıdır. Buna ilaveten en güncel gerçek örnek gözlemler üzerinde çalışarak hem önerilen yaklaşımın hem de mevcut uygulamaların geçerliliğini test etmiş olmasıdır.

1.4 Dolandırıcılık Tespitindeki Zorluklar

Dolandırıcılık tespiti amacıyla Makine Öğrenimine başvururken karşılaşılan zorlukların en başında müşterilere ait gerçek veriye ulaşmanın zor olması gelmektedir [4,5]. Başta kişisel veriler olmak üzere, harcamanın gerçekleştiği iş yeri bilgisi ve konum gibi değerli bilgilerin paylaşılması ve amacı dışında kullanılması hukuki olarak uygun değildir. Diğer yandan, uygun olarak elde edilen veriler mevcut olduğunda bu kez farklı zorluklar ortaya çıkar. Bunlardan ilki gözlem dağılımı sorunudur. Doğası gereği dolandırıcılık vakaları toplam işlemlerin arasında %1 oranına bile sahip değildir. Başka bir deyişle, dolandırıcılık tespitinde iki kategori arasında aşırı dengesiz bir oran bulunur. Bu durum sadece problemin zorluğunu göstermekle kalmayıp, model üretme için öğrenmede kullanılması

gereken yeterli miktarda dolandırıcılık işlemi gözlemi bulmayı da güçleştirmektedir.

Diğer yandan Makine Öğrenmesi sayesinde kural bazlı yapıya göre daha dinamik hale gelen yakalama çalışmaları karşı tarafta dolandırıcıların da bu modellere yakalanmamak adına stratejilerini daha sık güncellemelerine yol açmıştır [6]. Dolayısıyla bugün hırsızlık işlemine aitmiş gibi görünen bir durum sonraki gün sıklıkla normal işlem olabilmekteyken, normal işlemlerden ayırt edilemeyecek bir işlem de dolandırıcılara ait olabilir. Hırsızların neredeyse müşteri ile aynı özellikte harcama denemeleri yapması, en iyi modellerin bile kusursuz tahminleri olmasını olanaksız kılar.

Başka bir kritik zorluk ise, tahminlerin performansını değerlendirme sürecinde yaşanmaktadır. Genellikle bir tahmin başarısının ölçümü, toplam gözlemlerin ne kadar başarılı bilindiği anlamına gelen Doğruluk göstergesi ile ölçülmekteyken dolandırıcılık tespitinde gözlem dengesizliğinden dolayı bu göstergenin açıklayıcılığı çok zayıftır. Örneğin dolandırıcılık tahmin modeli, tüm gözlemler için normal etiketlemesi yapsa bile normal işlemlerin baskınlığı sebebiyle toplam Doğruluk değeri %100'e yakın olur. Oysa bu sonuç gerçekte hiçbir dolandırıcılık işleminin yakalanamamış olması anlamına gelmektedir. Dolayısıyla, klasik ölçütlerin aksine dolandırıcılık tespiti modellerinin tahmin performansını doğru bir şekilde ölçebilecek uygun göstergeler seçilmesi gereklidir.

Yukarıda kısaca özetlenen başlıklar uzatılarak daha birçok konu sıralanabilir. Ancak tüm bu zorluklar, hiçbir zaman araştırmacıların dolandırıcılık tespiti konusuna ilgisini çekmekten alıkoyamamıştır. Aksine araştırmalar ve uygulama örnekleri sayesinde konunun zorluklarıyla başa çıkmak üzere farklı öneriler getirilmiştir [7]. Konunun zorluklarını meydana getiren sosyal ve ekonomik boyutlarıyla gerçek hayatın içinden bir problem olması araştırmalarda ele alınmasını pozitif etkilemiştir. Bunun sonucunda dolandırıcılık tespiti alanında Makine Öğrenmesi ve ona bağlı olan çeşitli istatistiksel analiz yöntemleri literatürde sıklıkla tercih edilmiştir. Gelecek bölümde, öne çıkan yönleriyle dolandırıcılık tespiti için kullanılan Makine Öğrenmesi yöntemleri paylaşılmıştır.

2
KULLANILAN YÖNTEMLER

Bu bölümde, literatürde öne çıkan çalışmaların incelenmesinin ardından uygulama bölümüne geçmeden önce kullanılan yöntemler açıklanmıştır. Ayrıca kullanılan yöntemlere ek olarak alternatif yöntemlere de yer verilmiştir. Böylelikle farklı yöntemlerin arasından seçim yapılırken hangi sebeplere göre karar verildiği belirtilmiştir. Kullanılan yöntemler amaçlarına göre ayrılarak, veri hazırlığı üzerine kullanılan yöntemler veri dengeleme ana başlığında, tahmin sonucu elde etmeye yarayan yöntemler ise sınıflandırma ana başlığında ele alınmıştır.

2.1 Veri Dengeleme Yöntemleri

Gerçek hayattan doğan bir problem olan dolandırıcılık tespitinin zorlukları hakkında önceki bölümde açıklanan maddelerin başında veri elde etme gelmektedir. Normal gözlemlerin açık bir şekilde toplam gözlemlerin neredeyse tamamını oluşturması araştırmacıların veri dengeleme yöntemlerine iten temel unsur olmaktadır. Model kurulumu sürecinde yeterli dolandırıcılık gözlemine sahip olabilmek, öğrenme verisi ve test verisi kümelerinin dağılımını dengelemek başlıca veri dengeleme çalışmalarıdır. Aksi takdirde daha az gözleme sahip olan dolandırıcılık vakalarını normal durumdan ayırt etmek zorlaşır ve tahmin başarısı zayıflar. Eğer örnek bir uygulama verisinde dolandırıcılık gözlemleri gerçek hayatta rastlanan duruma ilaveten daha da az oranda bulunursa dolandırıcılık vakalarını doğru yakalama başarısı ciddi oranda düşecektir. Diğer yandan gerçek hayata göre daha sık dolandırıcılık gözlemi bulunursa da bu durum birçok normal gözleminde hatalı şekilde dolandırıcılık olarak tahmin edilmesine sebebiyet verebilecektir.

Ayrıca yukarıda bahsedildiği üzere tüm gözlemlerin homojen bir şekilde öğrenme ve test kümelerinde dağılmasını sağlamak veri dengeleme açısından önemli bir husustur [18]. Gerçek hayatta var olan rassallığın uygulama verisinde dengeli şekilde bulunmaması tahmin çalışmasının anlamlılığına zarar verir. Örneğin gözlemlerin sahip olduğu farklı özelliklerden biri olan harcama kategorisi ele

alındığında bir kategoriye ait tüm gözlemlerin mümkün olduğu kadar hem öğrenme hem de test kümelerinde bulunması tahmin başarısını arttıracaktır. Bu amaçla literatürde veri dengesizliği ile mücadele etmek için birçok farklı yöntem denenmiştir. Kullanılabilecek başlıca yöntemler aşağıda açıklanmıştır.

2.1.1 Yeniden Örnekleme

Yeniden örnekleme gerçek hayatı en iyi temsil eden durumların örnek uygulama verisinden nasıl elde edilebileceğini bulmaya yarar. En iyi başarıyı elde etmek için hangi gözlemlerin öğrenme ve test amacıyla kullanılacağını belirler. Yeniden örnekleme amacıyla birçok yöntem kullanılsa da, sık kullanılan yeniden örnekleme metotları Eksik Örnekleme (Undersampling), Aşırı Örnekleme (Oversampling), SMOTE (Synthetic Minority Over-Sampling Technique) ve Çapraz Doğrulama yöntemleridir [37].

Eksik Örnekleme, sayıca üstün olan sınıfa ait gözlemlerin sayıca daha az olan sınıfa yaklaştırılmasıdır. Çoğunlukta olan sınıfta yapılan bu azaltma işlemi rastgele bir şekilde yapılabileceği gibi istatistiksel yöntemlerden de faydalanılabilir. Eksik Örneklemenin en büyük dezavantajı, zaten yetersiz gözlem sayısı olan çalışmalarda tahmin sonucuna katkı sağlayabilecek gözlemleri de veri setinden kaldırmasıdır. Ayrıca, bu şekilde gözlem sayısı azaltıldığında örnek uzayının yapısı bozulabilir [37,38].

Aşırı Örneklemede ise, çoğunlukta olan sınıfa ait gözlem sayısını koruyarak azınlıkta olan sınıfa ait gözlemleri çoğaltmaya çalışılır. Aşırı Örnekleme rastgele olarak yapılabileceği gibi yapay olarak azınlık sınıfına ait yeni gözlemler üreterek de yapılabilir. Dolandırıcılık tespitinde olduğu gibi yüksek düzeyde dengesiz dağılıma sahip büyük bir veri kümesi ile çalışıldığında, toplam gözlem sayısı daha da artacağından çok maliyetli olabilir [37,38].

Aşırı Örnekleme yöntemlerinden biri olan SMOTE tekniği, azınlık durumunda olan sınıfa ait gözlem ve en yakın komşusunun farkını alarak [0,1] aralığından rastgele bir sayı ile çarpar. Elde edilen sonuç ile ilk gözleme ait değerler toplanır. Böylece azınlık sınıfı çoğaltılarak aşırı örnekleme yapılmış olur [39]. SMOTE yönteminin diğer yeniden örnekleme yöntemlerinden en temel farkı, gözlemlerin doğrudan

kopyalanarak çoğaltılması yerine en yakın komşularına bakılarak yeni örneklerin elde edilmesidir. Eşitlik 2.1 SMOTE yönteminde kullanılan basit formülü göstermektedir.

$$E_{yeni} = E_i + (E_i - E_j)\delta \tag{2.1}$$

2.1.2 Çapraz Doğrulama

Yeniden örnekleme bazlı yöntemlerden biri olan Çapraz Doğrulama ise, diğer metotlardan farklı olarak öncelikle öğrenme verisi ile test verisi arasındaki farklılığı önlemek amacıyla kullanılır [40,41]. Çünkü genellikle bu farklılık öğrenme verisinde başarılı sonuçlar varken test verisinde başarısızlığa sebep olur. Bu durumu önleme adına Çapraz Doğrulama öğrenme verisini alt gruplara ayırır. Bu işlem için birçok farklı yöntem olsa da en yaygın kullanılan Çapraz Doğrulama yöntemleri Leave-One-Out ve K-Fold metotlarıdır.

Leave-One-Out Çapraz Doğrulama, isminden de anlaşılacağı üzere gözlemlerin yalnızca tek bir adedini ayırarak test için kullanır [42]. Bu seçimi sırasıyla her bir gözlem için devam ettirir. Toplamda gözlem sayısı (n) kadar farklı öğrenme modeli oluşturulmuş olur. Bu yöntemde veri büyüklüğüne bağlı olarak model sayısı aşırı artabilir. (2.2) ile Leave-One-Out Çapraz Doğrulama formülü gösterilmiştir.

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \tag{2.2}$$

K-Fold Çapraz Doğrulama, farklı olarak örnekleri tekil ele almak yerine eşit boyutlu alt gruplara böler. Bu alt gruplardan bir tanesi modeli test etmek için doğrulama verisi olarak tutulur, kalan k − 1 alt örnek öğrenme verisi olarak kullanılır. Bu doğrulama işlemi daha sonra k kez tekrarlanır ve alt grupların her biri doğrulama verisi olarak bir kez kullanılır. Daha sonra sonuçlarının ortalaması alınır. Bu yöntemde k = n olduğunda her gözlem tek başına grup olacağından aslında Leave-One-Out metodu uygulanmış olur. Ancak genellikle k = 10 sayısı kullanılarak örnekler 10 alt kümeye ayrılarak doğrulama yapılır [43]. (2.3) eşitliği K-Fold Çapraz Doğrulama formülünü göstermektedir.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{n} MSE_i \quad (2.3)$$

2.1.3 Ağırlıklandırma

Dengesiz veri setlerinde tercih edilen dengeleme yöntemlerden biri de Ağırlıklandırma yöntemidir. Bu yöntem çıktı kategorilerini eşit önem derecesinde ele almak yerine aralarındaki dengesizliğe zıt oranda bir öncelik seviyesi belirleme yoluyla çalışır [44]. Ağırlıklandırma, önem seviyesi daha yüksek olan gruba ait yanlış sınıflandırma tahminlerini minimize eder. Başka bir deyişle, veriye doğrudan bir müdahalede bulunmak yerine karar verici tarafından belirlenecek katsayı oranında öğrenme fonksiyonu üzerinden sonuca ulaşır. Ancak verilen katsayı ya da ağırlıklandırma dengeleri bu kez de diğer yönde bozabilir. Bu sebepten dolayı sınıf ağırlıklandırma yönteminde küçük değişiklikler yaparak sonuca etkisi kontrol etmek daha doğru olur. Bunun yerine sınıf ağırlıklarıyla dolaylı yoldan değişiklik yapmanın yolu ise tahmin olasılık eşik değerlerini değiştirmektir. İki sınıftan oluşan çıktı durumunda varsayılan tahmin eşik değeri 0.5 noktasıdır. Sınıflardan biri diğerine göre daha azınlıkta ise eşik değerinde yapılacak bir değişiklik ile tahmin sonucu ciddi oranda farklılaştırılabilmektedir.

2.2 Sınıflandırma Yöntemleri

Makine Öğrenmesi kullanarak dolandırıcılık tespiti yapmanın en yaygın yolu sınıflandırma algoritmalarıdır. Gözetimli öğrenme yollarından biri olan sınıflandırmada amaç, öğrenme verisinden elde edilen sınıf bilgilerini kullanıp test verisindeki gözlemleri doğru sınıflarda etiketlemektir. Sınıflandırma, dolandırıcılık tespiti yapmanın dışında pek çok farklı amaçla da kullanılabilir. Bu amaçların arasında evet/hayır seçeneğinin olduğu ikili durumlar olabileceği gibi ikinden fazla kategoriyle de tahmin sonucu elde edilebilir. Literatürde yaygın olarak tercih edilen sınıflandırma yöntemleri aşağıdaki gibidir.

Sınıflandırma modelleri, belirlenen bir hata veya kazanç kriterine göre en iyi sonucu üretecek şekilde kurulur. Modelin kurulumu aşamasında bu kriteri farklılaştıracak parametre ayarları yapılması mümkündür. Dolandırıcılık tespitinde gözlem oranı dengesiz şekilde dağıldığından, sınıflandırma modelini

eğitirken ve tahmin sonuçlarını yorumlarken uygun performans çıktılarını kullanmak kritik önem arz eder [14,15,45].

		Tahmin Sonucu	
		Normal	Dolandırıcılık
Referans Değer	Normal	*Gerçek Negatif (TN)*	*Hatalı Pozitif (FP)*
	Dolandırıcılık	*Hatalı Negatif (FN)*	*Gerçek Pozitif (TP)*

Şekil 2.1 Karşılaştırma matrisi

Yukarıda Şekil 2.1 sınıflandırma tahminlerinde sıklıkla kullanılan karşılaştırma matrisinin alanlarını göstermektedir. Karşılaştırma matrisi iki değerli kategorik değişken olan dolandırıcılık durumunun gerçek gözlem değeri ile model tahmin sonucunu eşleştirir. Karşılaştırma matrisi elde edildikten sonra da performans göstergeleri aşağıda verilen eşitlikler yardımıyla elde edilir. Performans göstergelerinden hangilerinin tercih edileceği uygulama bölümünde verilen örnekle açıklanmıştır.

$$Doğruluk = \frac{Gerçek\ Pozitif\ (TP)+Gerçek\ Negatif\ (TN)}{TP+TN+FP+FN} \tag{2.4}$$

$$Hassasiyet = \frac{Gerçek\ Negatif\ (TN)}{TN+FP} \tag{2.5}$$

$$Belirlilik = \frac{Gerçek\ Pozitif\ (TP)}{TP+FN} \tag{2.6}$$

$$Kesinlik = \frac{Gerçek\ Pozitif\ (TP)}{TP+FP} \tag{2.7}$$

$$F_1 = 2 \times \frac{Kesinlik \times Belirlilik}{Kesinlik+Belirlilik} \tag{2.8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.9}$$

2.2.1 Lojistik Regresyon

Lojistik regresyon algoritması ikili değere sahip verilerin çıktılarını hesaplamak için lojistik fonksiyonların kullanıldığı makine öğrenimi metodudur [46]. İkili değerlere sahip verilere örnek olarak doğru veya yanlış, 0 veya 1, evet ya da hayır gibi çıktılar veren özellikler gösterilebilir. Lojistik Regresyon model kurulumunun

karmaşık modellere göre çok daha hızlı olması ve buna rağmen tahmin sonuçlarında genellikle başarılı olması sıklıkla tercih edilmesinde başlıca sebeplerdir.

Lojistik regresyon algoritması y değişkeni değerinin olasılığı ile verimizdeki x değişkeni öznitelikleri arasında bulunan ilişkiyi belirlemeye yardımcı olmaktadır. Lineer Regresyon algoritmasında [-∞, +∞] aralığında sonuçlar alınır. Bu sonuçların olasılık değeri elde etmeye çalıştığımız için [0 – 1] aralığında bir değere çevrilmesi gerekir. Bu işlem için sigmoid fonksiyonu kullanılır [47]. Başta sağlık alanındaki araştırmalar olmak üzere birçok alanda sıklıkla tercih edilir. Lojistik Regresyon karmaşık öğrenme algoritmalarına kıyasla çok daha hızlı çalıştırılabilir ve buna rağmen tahmin sonuçlarında da genellikle başarılı sayılır. Eşitlik 3.5'de Lojistik Regresyon formülünde p istenilen durumun gerçekleşme olasılığını, β_0 sabitin değerini, β_i (i=1,2,...n) ise her bir bağımsız değişkenin X_i (i=1,2,...n) katsayısını belirtir. Ayrıca p'nin $1 - p$ 'ye oranı da bahis oranı olarak adlandırılır. Dolayısıyla Lojistik Regresyon şu şekilde (2.10) yazılır:

$$\log \frac{p}{(1-p)} = \beta_0 + \beta_1 * X_1 + \cdots + \beta_n * X_n \tag{2.10}$$

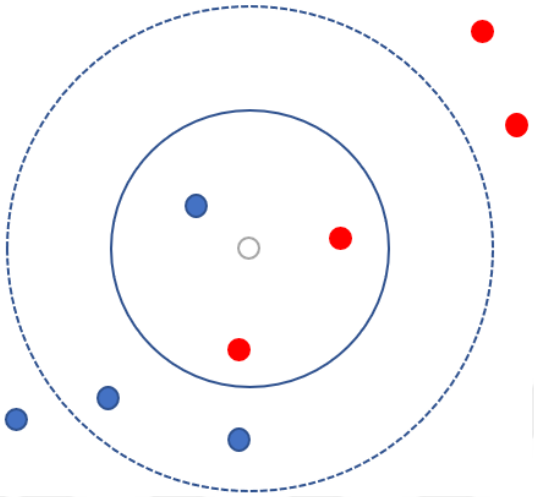
2.2.2 K-En Yakın Komşu

K-En Yakın Komşu algoritması gözlemleri benzerliğine göre sınıflandırır. Bu benzerlik, komşuların oylarının çokluğu ile sınıflandırılır ve mesafe fonksiyonu kullanılarak yapılan ölçümle en yakın komşular için ortak bir sınıfa atanır. Pozitif bir tamsayı olan K değeri genellikle küçük değerler alır, K = 1 olduğu durumda gözlem en yakındaki komşunun sınıfına atanır. K-En yakın komşular algoritması sınıflama hesaplamalarını farklı mesafe ölçüm fonksiyonlarını kullanarak gerçekleştirebilir [48].

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2.11}$$

Yukarıda (2.11) Öklid mesafesi hesaplamasını göstermektedir. Bu hesaplamaya örnek olarak Şekil 2.2 incelendiğinde KNN algoritmasıyla ortadaki yeni gözlemin

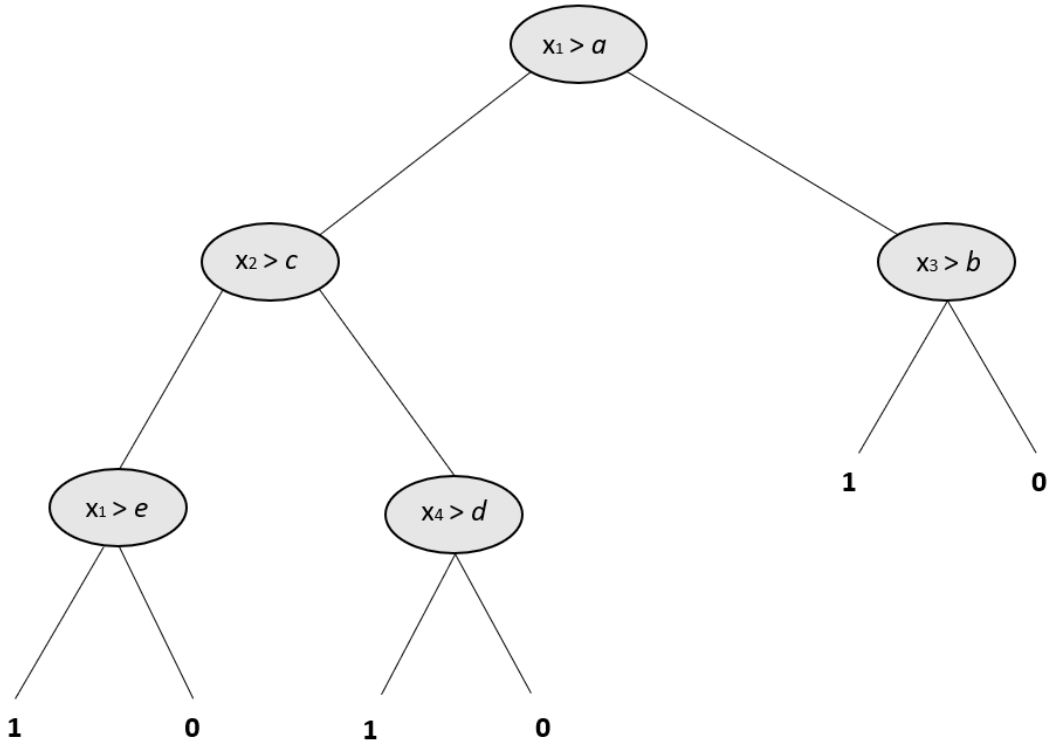
mavi ya da kırmızı kategorilerden hangisine dahil olacağını belirlenmesi gösterilmiştir. Buna göre, K=3 olması durumunda (düz çizgili daire) en yakın 3 gözlemden ikisi kırmızı olduğundan tahmin sonucu kırmızıdır. Eğer K=5 olursa, en yakın 5 gözlemden üç tanesi mavi olduğundan sonuç mavidir.



Şekil 2.2 Örnek KNN modeli gösterimi

2.2.3 Karar Ağaçları ve Rastgele Orman Algoritması

Karar Ağacı Algoritmaları veriye ait sayısal ya da kategorik nitelikler üzerinden hiyerarşik olarak karar noktaları (düğümler) belirleyerek örnekleri sınıflandırmaya yarayan, bu özelliğiyle şekil itibariyle yaprak ve dalları olan bir ağaca benzeyen temel bir makine öğrenmesi metodudur [49]. En sık tercih edilen Karar Ağacı algoritmalarından ID3 algoritması J.R. Quinlan, tarafından 1986 yılında geliştirmiştir. Bu algoritmada bilgi kazanımının en fazla olduğu düğüm kök olarak en başta yer alır. Bilgi kazanımı durana kadar her bir karar noktasında yeni dallar oluşarak dallanma devam eder [50]. Basitçe ifade edilen bu yöntem sonucunda karar ağacı oluşturulmuş olur. Ancak Karar Ağaçları genellikle öğrenme verisindeki özellikleri çok yakından takip eder ve bu durum aşırı öğrenmeye yol açarak test verisinde düşük tahmin başarısına sebebiyet verebilir. Şekil 2.3 Örnek bir karar ağacı çıktısını göstermektedir [61].



Şekil 2.3 Örnek karar ağacı

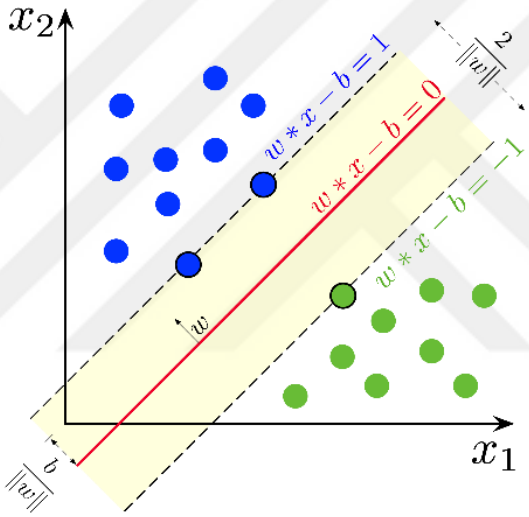
Makine öğrenmesinde en popüler yöntemlerden biri olan Rastgele Orman algoritması ise, orijinalinde karar ağaç bazlı öğrenme yönteminin geliştirilip rastgele olarak çok sayıda ağaç kullanan ve veriye ait özellikleri de rastgele olarak seçen bir versiyonudur. Rastgele Orman algoritması genellikle Karar Ağaçlarına göre aşırı öğrenme durumundan daha uzak kalarak tahmin başarısını yükseltebilmektedir [51]. Rastgele Orman algoritmasında, çok sayıda ayrı ayrı ağaçlardan elde edilen sonuçlar tek bir kararda birleştirilirken Çoğunluk Oylama (Majority Voting) kullanılır.

Sonuç olarak, Rastgele Orman algoritmasının tercih edilme sebepleri yüksek sınıflandırma başarısı ve büyük verilerde bile hızlı ve kolayca oluşturulabilmesidir. Diğer yandan Rastgele Orman algoritması Karar Ağaçlarındaki gibi ortaya bir ağaç modeli çıkartmaz, ancak bunun yerine veriye ait özelliklerin önem miktarını ölçmek için kullanılabilmektedir [52].

2.2.4 Destek Vektör Makineleri

Sınıflandırma uygulamalarında sıklıkla kullanılan destek vektör makineleri (SVM) [53], gözlemleri farklı sınıflara ait örnekleri ayırmak için bir hiper düzlem kullanır.

Fakat farklı sınıflara ait örnekleri ayırmak için çok sayıda hiper düzlem oluşturabilmek mümkündür. SVM, oluşturulabilecek bu hiper düzlemler içinden kendisine en yakın ve farklı sınıflarda bulunan örnekler arasında aralığı (Margin) maksimize eden düzlemi seçer ve bu düzlem karar sınırı olarak adlandırılır. Margin her zaman net şekilde bulunamayabilir. Bazen örnekler Margin bölgesine girebilir. Bu duruma Soft Margin denir. Hard Margin, örnekler tamamen doğrusal olarak ayrılabiliyorsa çalışır ve aykırı değerlere karşı çok duyarlıdır. İkisi arasındaki denge SVM içerisindeki C hiperparametresi ile kontrol edilir. C ne kadar büyükse Margin o kadar dardır. Aşırı öğrenme durumuyla karşı karşıya kalındığında C değeri azaltılarak Margin genişletilebilir. Şekil 2.4 Marjin çizimini göstermektedir [62].



Şekil 2.4 Destek vektör makinelerinde marjin çizimi

SVM, doğrusal (lineer) fonksiyonlarla sınıflandırılabilen problemleri çözmek için kullanılabildiği gibi farklı Kernel Fonksiyonları sayesinde doğrusal olarak ayrılamayan durumlar için de kullanılabilir. Bu durumlarda Polynomial Kernel ve Radial Kernel fonksiyonları devreye girer. Aşağıdaki eşitlerle sırasıyla Lineer, Polynomial ve Radial Kernel fonksiyonlarının yazımı gösterilmiştir.

$$K(x, y) = \sum_{j=1}^{p} x_{ij} y_{ij} \tag{2.12}$$

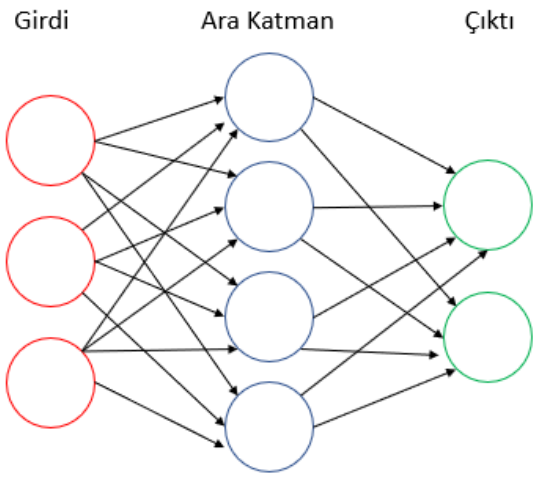
$$K(x,y) = (1 + \sum_{j=1}^{p} x_{ij} y_{ij})^d \quad (2.13)$$

$$K(x,y) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - y_{ij})^2) \quad (2.14)$$

2.2.5 Yapay Sinir Ağları

Yapay Sinir Ağı mimarisi insan beynine benzer çalışır. Nöron adı verilen insan beyni hücreleri, karmaşık ve yüksek oranda birbirine bağlı bir ağ oluşturarak birbirine elektrik sinyalleri gönderir ve böylece insanların bilgileri işlemesine yardımcı olur. Benzer şekilde, yapay sinir ağı, bir amaca ulaşmak için birlikte çalışan yapay nöronlardan meydana gelir. Yapay nöronlar, düğüm adı verilen algoritma adımlarıdır.

Temel bir sinir ağı, üç katmanda birbirine bağlı yapay nöronlara sahiptir [10,11,13]. Girdi katmanında, uygulamadaki bilgiler yapay sinir ağı algoritmasına girer. Düğümler verileri işler, sınıflandırır ve ardından bir sonraki katmana aktarır. Gizli (Ara) katmanlar, girdilerini girdi katmanından ya da diğer gizli katmanlardan alır. Yapay sinir ağlarında çok sayıda gizli katman bulunabilir. Her bir gizli katman, bir önceki katmandaki çıktıları analiz eder, daha ayrıntılı şekilde işler ve bir sonraki katmana aktarır. Çıktı katmanı, yapay sinir ağı tarafından işlenen tüm verilerin nihai sonucunu verir. Bu katman, bir veya daha fazla düğüme sahip olabilir. Örneğin, ikili (evet/hayır) bir sınıflandırma varsa çıktı katmanı, sonucu 1 veya 0 olarak verecek tek bir çıktı düğümüne sahip olur. Ayrıca, çoklu bir sınıflandırma söz konusu olduğunda çıktı katmanı, birden fazla çıktı düğümünden oluşabilir [54]. Yapay Sinir Ağı modelinde bulunan katmanlar Şekil 2.5 ile gösterilmiştir [63].



Şekil 2.5 Yapay sinir ağının bileşenleri

2.2.6 Model Birleştirme

Sınıflandırma amacıyla kullanılan makine öğrenmesi algoritmalarının aynı anda karara etki edebilmesi anlamına gelen model birleştirme yöntemi, toplam tahmin başarısını yükseltmek için tercih edilen ilave metotların başında gelir [20,21,22]. Birleştirme ile tekil makine öğrenmesi algoritmalarının dezavantajları, kurulan modellerin aşırı öğrenmeye veya eksik öğrenmeye yatkınlığı gibi olumsuz durumlar ortadan kaldırılmaya çalışılır. Model birleştirme, tahmin kararını tek bir modele göre belirlemek yerine, aynı ya da farklı algoritmaların kullanıldığı farklı modelleri karar verme sürecine birlikte dahil eder.

Modellerin nasıl birleştirileceği ise bir diğer önemli konudur. İyimserlik ve Kötümserlik gibi temel karar verme yaklaşımları kullanılabileceği gibi Ağırlıklı Ortalama gibi daha hassas yöntemler tercih edilebilir [23,24]. Karmaşık seçenekler arasında da üst katman makine öğrenimi gelmektedir. Bu yöntemde mevcut modellerin tahmin sonuçları girdi olarak bir sonraki modelin öğretilmesinde kullanılır. Sıklıkla kullanılan birleştirme yöntemleri şu şekildedir [8]:

- İyimserlik / Kötümserlik
- Ağırlıklı Ortalama
- Yumuşak Oylama (Soft Voting)
- Çoğunluk Oylama (Majority Voting)
- Üst Katman Tahmin Modeli

Bu bölümde kısaca açıklanan yöntemler ve daha fazlası dolandırıcılık tespiti amacıyla gerçekleştirilen makine öğrenmesi uygulamalarında sıklıkla yer almaktadır. Temel noktalarıyla öne çıkan veri dengeleme ve sınıflandırma yöntemlerin tanıtılmasının ardından uygulama bölümüne geçilmiştir.

3
UYGULAMA

Makine Öğrenmesi kullanarak kredi kartı hırsızlık tespitinde bulunmak amacıyla gerçekleştirilen uygulama çalışması, veri hazırlığından model eğitimine, performans parametresi seçiminden model birleştirmeye ve başka birçok detaya inen kapsamlı adımları içermektedir. Başlangıç noktasını belirleyebilme adına uygulama adımında ortaya konan başlangıç modeli de dahil olmak üzere toplam elliden fazla farklı tahmin modeli oluşturulmuştur.

3.1 Veri Örneğinin Açıklanması

Uygulama kapsamında kullanılmış olan gerçek kredi kartı işlem verisi, Türkiye'deki bankacılık sektöründe üst sıralarda yer alan, yabancı sermayeli bir bankadan temin edilmiştir. Uygulama verisi olarak adlandırılan bu veri örneği, Haziran 2021 ile Kasım 2021 tarihleri arasında 1,846 farklı müşterinin toplam 115,159 harcama işleminden oluşmaktadır. Başka bir deyişle, uygulama verisi her bir müşteri için ortalama 62 işlem içermektedir. Uygulama verisinde yer alan işlemlerden 1,166 tanesi dolandırıcılık işlemi olarak bilinmektedir. Bu durum, beklendiği gibi uygulama verisinde gözlemler arasında dengesizliği işaret etmektedir. Uygulama boyunca işte bu dengesiz veri durumu ile mücadele ederek nasıl etkin tahmin modeli kurulabilir sorusuna cevap aranacaktır. Tablo 3.1 Uygulama verisine ait genel özellikleri göstermektedir.

Tablo 3.1 Uygulama verisinin detayı

Müşteri Sayısı	Toplam İşlem Sayısı	Toplam Dolandırıcılık İşlemi	Ortalama İşlem Sayısı
1,846	115,159	1,166	62.4

Uygulama verisinde yer alan her bir gözlem ayrı bir harcama işlemini göstermektedir. Gözlemlerin içerisinde yer alan bilgiler, bağımlı değişken ve bağımsız değişkenler olarak iki gruba ayrılmıştır. Bağımlı değişken *y*, ilgili işlemin dolandırıcılık işlemi olup olmadığı bilgisini tutar. Bu bilgi, banka tarafından

müşteriye telefon yoluyla ulaşılarak teyit edilmiş gerçek ve kesin bilgidir. Tablo 3.2 Uygulama verisi içerisinden alınmış örnek dört gözlemi göstermektedir. Tablo içerisinde kolon isimleri ile gösterilen değişkenler hakkında açıklamalar ise Tablo 3.3 ile belirtilmiştir.

Tablo 3.2 Veriye ait temel özellikler

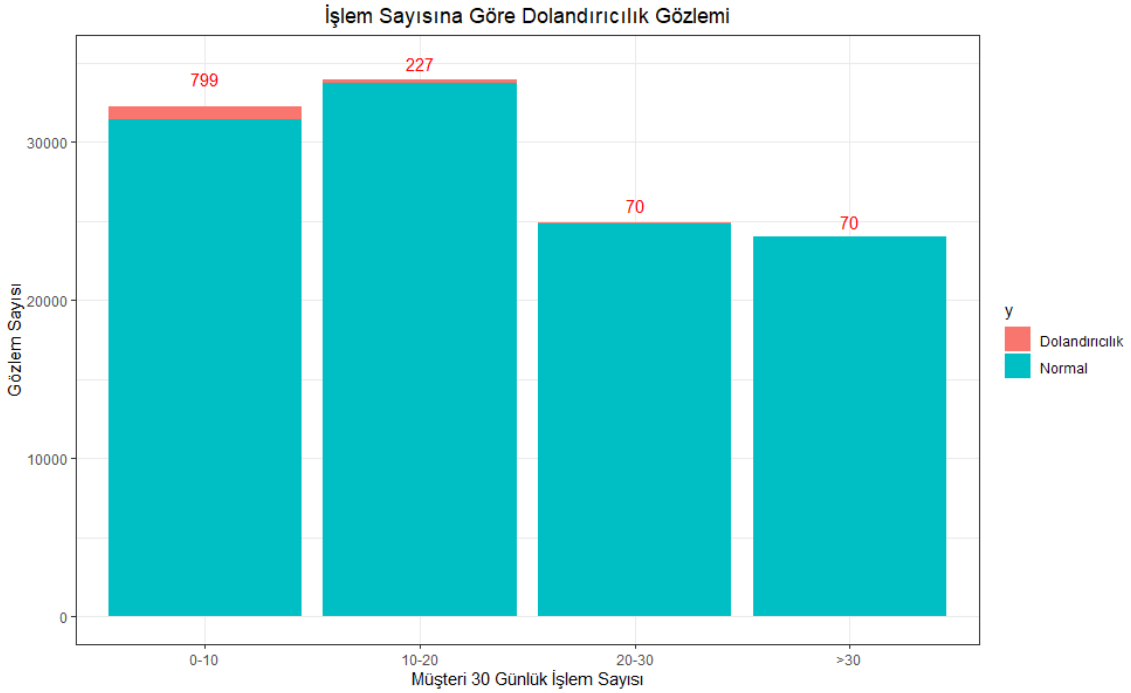
islem_no	musteri_no	tarih	tutar	doviz	kategori	isyeri_no	y
109560	5170	15.06.2021	4.89	1	4091	9347217293	0
109561	5134	08.07.2021	27.95	2	6591	2022524852	1
109562	5140	09.07.2021	273.2	2	1588	6563994466	0
109563	5196	14.09.2021	27.95	2	6591	143893621	0

Kategori bilgisi, ISO 18245 ile standardize edilmiş uluslararası ödeme sistemleri şirketleri tarafından kullanılan dört haneli nümerik harcama kodudur. Her iş yerinin tek bir kategori kodu bulunur. Harcama kodları zaman zaman şirketler için de özelleştirilebilmektedir. Örneğin 32XX şeklinde belirlenen kategori kodları çoğunlukla havayolu şirketlerinden oluşur. Bu sebeple bazı kategori kodlarının işletmelere özel olmasından dolayı veri hazırlığı öncesinde harcama gözlemlerine ait gerçek kategori kodları veri bütünlüğü bozulmadan maskelenmiştir.

Tablo 3.3 Temel özelliklerin açıklaması

Değişken Adı	Açıklama
islem_no	İşleme özel numara
musteri_no	Müşteriye özel numara
Tarih	İşlemin tarihi (gg/aa/yy)
tutar	İşlemin tutarı
doviz	İşlemin döviz cinsi
kategori	İşlemin harcama kategorisi
işyeri_no	İşlem yerine özel numara
y	Dolandırıcılık var mı?

Harcama gözlemlerine daha yakından bakarak öncül çıkarımlar elde etmek amacıyla Tablo 3.3 ile özetlenen değişkenler kullanılarak uygulama verisine ait temel göstergeler üretilmiştir. Şekil 3.1 müşterilerin 30 günlük işlem sayısı özelliğine göre uygulama verisinin dağılımını göstermektedir. İlk dikkat çeken nokta, müşteri işlem sayısının arttıkça dolandırıcılık gözleminin ciddi oranda azalmış olmasıdır. Aylık işlem sayısı 0-10 aralığında olan müşterilerin işlemlerinde yaklaşık %3 dolandırıcılık gözlemi bulunuyorken, ayda 30'dan fazla işlem sayısı olan müşterilerin işlemlerinde aynı oran %0.2 'dir.



Şekil 3.1 Gözlemlerin dağılımı

Tahmin modeli kurulurken, uygulama verisi içinden öğrenme ve test kümeleri belirlenmiştir. Öğrenme verisi, tahmin algoritmasının gözetimli olarak başka bir deyişle bağımlı değişkenin değerlerini de kullanarak model kurulmasında kullanılmıştır. Test verisi ise, öğrenilen modelin tahmin performansını ölçmek adına öğrenme verisinden farklı gözlemleri içeren veri kümesidir. Öğrenme verisi ve test verisi özelliklerine bağlı olarak tahmin modelinin davranışları değişir. Gözlem boyutları, özellik sayısı, dolandırıcılık oranı gibi kritik durumların nasıl karar verileceğini belirlemek adına tahmin modeli farklı senaryolarda tekrarlanmıştır.

3.2 Tahmin Modeli ve Performans Göstergelerine Bakış

Bu bölümde uygulama verisinde yer alan temel özelliklerin incelenmesinin ardından literatürde sıklıkla öne çıkan tahmin yöntemlerinden biri seçilerek ilk sonuçlar elde edilmiştir. Seçilen tahmin algoritması, literatür taraması bölümünde detayları açıklanan Rastgele Orman olmuştur [5,8,29]. Başlangıç modeli sayesinde uygulama kapsamında yapılacak araştırmaların literatürce bilinen noktadan ne kadar ileriye taşınabileceği değerlendirilebilecektir. Ek olarak, tahmin modelinin test sonuçlarına detaylıca bakılacak ve performans göstergelerinin arasında önce çıkan anahtar değerler tespit edilecektir. Böylece sonraki bölümlerde elde edilen sonuçlar sadece anahtar göstergeler üzerinden gidilerek doğru ve sade bir şekilde karşılaştırılabilecektir.

Başlangıç modelinin bir diğer faydası ise, öğrenme ve test veri setlerinin en iyi tahmin sonucunu elde etmek amacıyla birçok açıdan optimizasyonunu sağlayabilmektir. Bu fayda başta elde edilmesi zor ve uygulanması karmaşık gibi görünse de baz modelin içerdiği basitlik sayesinde veri boyutundaki değişimin sonuçlara etkisi kolaylıkla görülebilmektedir. İlk olarak, öğrenme veri seti için alternatif boyutlar ve dolandırıcılık oranları farklı senaryolarda enterpolasyonlar yapılarak oluşturulmuştur. Aday öğrenme verileri Tablo 3.4 ile gösterilmiştir.

Tablo 3.4 Öğrenme veri boyutu kombinasyonları

Adaylar	%D	Öğrenme Verisinde İşlem Sayısı	
		Toplam Gözlem	Dolandırıcılık Gözlem
A_{11}	%1	10,000	100
A_{21}	%2	10,000	200
A_{31}	%3	10,000	300
A_{41}	%4	10,000	400
A_{51}	%5	10,000	500
A_{12}	%1	20,000	200
A_{22}	%2	20,000	400
A_{32}	%3	20,000	600
A_{42}	%4	20,000	800
A_{52}	%5	20,000	1,000
A_{13}	%1	30,000	300
A_{23}	%2	30,000	600
A_{33}	%3	30,000	900
A_{15}	%1	50,000	500
A_{25}	%2	50,000	1,000
A_{16}	%1	60,000	600

Öğrenme verisi için optimal özellikler aranırken beş farklı dolandırıcılık oranı ile beş farklı veri boyutunun kombinasyonları kullanılmıştır. Adaylar isimlendirilirken önce dolandırıcılık oranı ardından toplam öğrenme gözlemi boyutu dikkate alınmıştır. Örneğin A_{32}, %3 dolandırıcılık oranına sahip 20,000 gözlemden oluşan adaydır. A_{23} ise %2 dolandırıcılık oranına sahip 30,000 gözlemden oluşan adaydır. Diğer yandan gözlem boyutu arttıkça dolandırıcılık oranına bağlı olarak dolandırıcılık gözlemi sayısı artar. Bu da aday kombinasyonlarını sınırlandırır. Örneğin elli bin gözleme sahip öğrenme verisinde %2 değerinden fazla dolandırıcılık oranı kullanılamaz. Aksi takdirde uygulama verisindeki toplam dolandırıcılık gözlem sayısı kısıdına takılmış olunur. Başka bir deyişle yeteri kadar dolandırıcılık gözlemi bulunamadığından yüksek toplam gözlem boyutlarında dolandırıcılık oranı düşük kalır.

3.2.1 Tahmin Modelinin Kurulması

Öncül göstergeleri elde edebilmek amacıyla önceki bölümde belirlenen öğrenme veri seti adayları kullanılmıştır. Modellerin tahmin sonuçlarıyla bir yandan öğrenme ve test verileri incelenebilecek diğer yandan da uygulama verisinin sunduğu temel özelliklerin tahmin performansı çeşitli göstergeler aracılığıyla ortaya konulabilecektir. Her bir aday veri setiyle kurulacak tahmin modellerini elde etmek içim literatür bölümünde öne çıkan makine öğrenmesi algoritmalarından olan Rastgele Orman algoritması seçilmiştir. Bu yöntemle her bir aday öğrenme veri setine ait tahmin modelleri R programlama dili kullanılarak oluşturulmuştur.

Ardından her bir tahmin modeli dolandırıcılık gözlemleme oranı aynı olan üç farklı test veri boyutunda test edilmiştir. Aşağıda Tablo 3.5 kurulan modellerden her bir test boyutu için en iyi sonuç veren modele ait Doğruluk, Hassasiyet, Belirlilik, Kesinlik, F_1 skoru ve MCC değerlerini göstermektedir. Tablo incelendiğinde, test verisi gözleminin artışının tüm göstergelerde aynı etkiyi oluşturmadığı belirlenmiştir. Hassasiyet ve Doğruluk değerlerinde değişim olmazken, Belirlilik, Kesinlik, F_1 ve MCC değerleri ayrışmıştır. Uygulama çalışması boyunca anlaşılabilirlik ve karşılaştırma kolaylığı sağlamak amacıyla performans göstergeleri arasından öne çıkan anahtar göstergeler seçilmiştir.

3.2.2 Anahtar Performans Göstergelerinin Seçilmesi

Anahtar Performans Göstergesi, gerçek değerler ile uygulama sonuçlarına bağlı ortaya çıkan çok çeşitli göstergeler arasında problemin gerekliliklerini en iyi ortaya çıkaran göstergedir. Tahmin modelleri açısından ele alındığında, karşılaştırma matrisi ile ayrıştırılan doğru ve yanlış tahminlerin sonucunda birçok farklı gösterge oluşur.

Doğruluk değeri, gerçek verideki tüm değerlerin toplamda ne kadarının doğru bilindiğini gösterir. Genelleme yapıldığında, her tahmin modeli için anlamlı bir sonuç olduğu inkar edilemese de dolandırıcılık gözleminin normal gözlemlere göre çok düşük oranda yer aldığı mevcut durumda Doğruluk göstergesinin açıklayıcılığı yeterli değildir.

Hassasiyet değeri, gerçek verideki normal durum gözlemlerinin doğru yakalanma oranı olarak basitleştirilebilir. Bu sebeple normal gözlemlerin aranma ihtiyacının olmadığı mevcut dolandırıcılık tespiti problemi için anahtar göster olarak kullanılması uygun olmaz.

Belirlilik değeri, doğru dolandırıcılık gözlem tahminleri sayısı ile toplam gerçek dolandırıcılık gözlem sayısının oranıdır. Kesinlik değeri ise, aynı değerin modelin toplam dolandırıcılık tahmin sayısına oranıdır. Belirlilik ve Kesinlik değerleri doğru dolandırıcılık tahminlerine odaklı olduğundan uygulama sonuçlarını değerlendirmede kullanılmaları uygundur.

F_1 skoru, Belirlilik ve Kesinlik arasında eşit önem derecesi benimseyen harmonik ortalama değeri olarak iki göstergeyi tek bir göstergede toplar [55]. Bu sebeple F_1 skoru anahtar gösterge olarak seçilmeye yakındır. Son olarak ise MCC değeri incelendiğinde F_1 skoruna yakın sonuçlarla karşılaşıldığı tespit edilmiştir. Bu durumun sebebi MCC değerinin F_1 skorunu ve normal gözlemler için tahmin başarısını kapsamasıdır. Ancak normal gözlemlerin tahmin başarısı dolandırıcılık tespiti için öncelikli bir gösterge olmadığından F_1 skoru ile elde edilen değerlerin anahtar gösterge olarak seçilmesine karar verilmiştir.

Öte yandan karar vericiler F_1 skoruna ek olarak risk iştahı seçiminde bulunabilirler. Risk iştahı yüksek veya düşük olduğunda F_1 skorundan farklı performans göstergeleri değerlendirilebilir. Yüksek risk iştahı gerçek dolandırıcılık verilerinin mümkün olduğu kadar tamamını doğru tespit etmeyi, düşük risk iştahı ise yapılan tahminlerde minimum sayıda hata yapmayı amaçlar. İlk bakışta birinin zıttı gibi gözükseler de tam anlamıyla zıt amaçlar değildir. Ancak aynı anda iki amacı birden optimize etmek gerçek hayat koşullarında çok zordur.

Tablo 3.5 En iyi aday öğrenim modellerinin sonuçları

Öğrenme Modeli	Test Verisi	Doğruluk	Hassasiyet	Belirlilik	Kesinlik	F_1	MCC
A_{51}	10,000	%98.6	%99.7	%68.0	%38.2	%49.0	%50.3
A_{42}	20,000	%98.7	%99.5	%47.5	%38.1	%42.0	%41.5
A_{42}	30,000	%98.8	%99.5	%55.3	%42.5	%47.4	%47.3

Bu değerlendirmeye göre tüm aday modeller ile test gözlem boyutları arasındaki karşılaştırma için F_1 skoruna göre Tablo 3.6 oluşturulmuştur. Tablo incelenirse, ilk göze çarpan durum öğrenme dolandırıcılık oranı arttıkça F_1 skorlarının yükselmesidir. Aday A_{11} ile A_{51} kıyaslanırsa %1 dolandırıcılık oranı ile %5 dolandırıcılık oranı arasında ortalama %50 iyileşme sağlandığı görülmüştür. Benzer bir sonuç aday A_{12} ile A_{52} arasında da öne çıkmaktadır. Toplamda baz modeller ile yapılan değerlendirmeye göre, ortalama olarak %40 ile en iyi F_1 skoru değerine sahip olan %4 dolandırıcılık oranı ve 20,000 gözlem boyutu ile A_{42} adayı öne çıkmaktadır. Test gözlem boyutu olarak A_{42} modelinin en yüksek F_1 skoru değeri ile 30,000 gözleme sahip test verisi seçilmiştir. Yapılabilecek bir diğer tespit, %50 seviyelerindeki F_1 skorunun iyi bir başlangıç noktası olabileceğidir. Uygulama boyunca F_1 skorunun nasıl iyileştirilebileceği incelenecektir.

Tablo 3.6 F_1 skoruna göre test sonuçlarının karşılaştırılması

Adaylar	**F_1 Skoru**			
	Test Verisi Gözlemi			
	10,000	20,000	30,000	Ortalama
A_{11}	%23	%33	%23	%25
A_{21}	%26	%35	%15	%25
A_{31}	%29	%29	%17	%26
A_{41}	%33	%37	%15	%28
A_{51}	%**49**	%40	%15	%35
A_{12}	%12	%29	%29	%23
A_{22}	%30	%31	%40	%34
A_{32}	%30	%35	%43	%36
A_{42}	%30	%**42**	%**47**	%**40**
A_{52}	%33	%39	%33	%35
A_{13}	%12	%6	%5	%8
A_{23}	%12	%26	%31	%23
A_{33}	%35	%37	%39	%37
A_{15}	%4	%24	%18	%15
A_{25}	%29	%27	%34	%30
A_{16}	%5	%31	%23	%20
Ortalama	%25	%31	%27	%28

Mevcut değerlendirmelerde test verisi için dolandırıcılık oranı tüm gözlem boyutlarında %1 tutulmuştur. Test verisinde toplam uygulama verisine ait oranı korumanın başlıca sebebi tahmin performansını net görebilmektir. Aksi durumda gerçek verinin sahip olduğu dolandırıcılık oranını değiştirmenin bir alternatif olmaktan öteye geçmesi doğru olmayacaktır, çünkü dolandırıcılık tespiti yapmanın amacı bu gözlem dengesizliğine karşı çözüm üretmekten geçer. Test veri setinde dolandırıcılık gözleminin bulunma oranını arttırmak sadece kâğıt üzerinde iyi sonuçlar almak anlamına geleceğinden gerçek manada bir optimizasyon olarak görülemez.

Uygulama boyunca test verisine ait dolandırıcılık oranı değiştirilmeden, farklı yöntemler izlenerek tahmin performansının iyileştirilebileceği gösterilmeye çalışılmıştır. Bu yöntemlerden ilki ve en öne çıkanı yardımcı değişkenler üreterek veriye daha kapsayıcı özellikler eklemektir.

3.3 Üretilen Yardımcı Özellikler

Önceki bölümde elde edilen sonuçlara bakıldığında uygulama verisi her ne kadar sahip olduğu temel değişkenler açısından kritik bir içeriğe sahip olsa da çıkarım üretilmesi açısından yeterli olduğu söylenemez. Bu durum gerçek hayatta da mevcut olan temel özelliklere ek olarak, çeşitli türetilmiş özellikler üretilmesi için gerekli bir sonuç doğurmuştur. Üretilen yardımcı değişkenlerden kasıt da, müşterinin ve harcamaların sahip olduğu temel istatistiklerdir. Belirli periyotlar ile hesaplanan bu istatistikler, başta dolandırıcılık tespitine doğrudan fayda sağlamıyor gibi görünse de yapılan detaylı analizler neticesinde belirgin miktarda açıklayıcılığa sahip oldukları görülmüştür. Bu bölümde ilk olarak müşteri kırılımında elde edilen istatistikler ile başlanıp ardından harcama noktası ve kategorisi ile devam edilecektir. Üretilen tüm özellikler 7 ile 60 arasında toplam beş farklı periyotta ayrı ayrı hesaplanmıştır. Çıkarımların otomatik olarak yapılması için Python dilinde Le Borgne vd. [56] tarafından önceden sağlanmış hesaplama fonksiyonları kullanılmıştır. Aşağıda yardımcı özelliklerin hesaplama yöntemlerine ait detaylar ayrı başlıklarda açıklanmıştır.

3.3.1 Müşteri RFM Analizi

Dolandırıcılık işlemlerini gerçekleştiren hırsızların, yakalanmama amacıyla özel olarak belirledikleri hırsızlık harcamaları ile kartın gerçek sahibi olan müşterinin harcamaları çoğunlukla birbirine benzemez. Üretilen özelliklerden ilki olan Müşteri Değişkenleri için de RFM Analizi baz alınarak bu durumdan faydalanılmaya çalışılmıştır. Kısaca özetlemek gerekirse, her bir müşterinin belirlenen periyottaki Türk Lirası (TL) cinsinden ortalama harcama tutarı ve yine belirli periyottaki harcama adedi çıkarılmış, böylece müşterilerin harcama alışkanlıklarına dair bilgi edinilmiştir. Hırsızlık işlemler ile müşterinin kendine ait tercihiyle yaptığı işlemleri arasında olacak farklılıklar yardımcı değişkenler aracılığıyla görülebilecektir. Örneğin bir müşterinin aylık ortalama işlem tutarı 100 TL iken, aynı müşterinin kredi kartında görülen 5,000 TL tek bir işlemdeyse kolaylıkla ayırt edilebilir. Her ne kadar bu işlemin müşterinin bizzat kendisi tarafından gerçekleştirilmiş olma ihtimali bulunsa da ortalama harcama tutarı

bilgisi diğer özellikler ile de bir araya getirildiğinde toplam açıklama becerisine önemli bir katkı sunar.

Tablo 3.7 Müşteri harcamalarına ait yardımcı özellikler

musteri_no	5199	5199	5189	5189	5151	5151
islem_tarihi	23.10.21	15.11.21	21.09.21	30.11.21	30.08.21	30.08.21
Musteri_NB_7	2	3	3	8	7	10
Musteri_NB_15	2	3	3	9	14	17
Musteri_NB_30	5	5	3	15	25	28
Musteri_NB_45	5	6	5	17	33	36
Musteri_NB_60	5	8	10	19	34	37
Musteri_AVG_7	12.42	8.90	187.13	496.23	54.54	44.84
Musteri_AVG_15	12.42	8.90	187.13	443.65	44.23	40.35
Musteri_AVG_30	14.43	10.31	187.13	472.86	38.03	36.33
Musteri_AVG_45	14.43	15.09	162.11	446.11	36.52	35.33
Musteri_AVG_60	14.43	12.36	222.27	405.73	35.51	34.43

Yukarıda Tablo 3.7 ile seçilen üç farklı müşteri için iki ayrı gözleme göre üretilmiş yardımcı özellikler görülmektedir. Periyodik olarak harcama sayısını ifade eden özellik Musteri_NB ile isimlendirilirken, ortalama harcama tutarının hesaplandığı özellikler ise Musteri_AVG ile isimlendirilmiştir. Tüm periyotlar için her hesaplama, mevcut işlem tarihine göre geriye yönelik yapılmıştır. Bu yöntem aynı zamanda büyük periyotların küçük periyotları kapsayarak çalışır. Sonuçta, daha uzun periyotları dahil eden özellikler müşteri harcama alışkanlığını daha doğru ifade eder. Bu nedenle de tek bir günlük de dahil olmak üzere bir haftadan küçük periyotların hesaplanmasına ihtiyaç duyulmamıştır.

3.3.2 Harcama Risk Puanı Analizi

Dolandırıcılık işlemlerini normal işlemlerden daha doğru ayırabilmek adına gerçekleştirilen diğer yardımcı özellik eklemesi ise harcama noktaları ve harcama kategorileri üzerindendir. Öncelikle bir önceki adıma benzer olarak her bir işlem noktası ve de kategori için periyodik harcama adetleri üretilmiştir. Tek fark önceki işlemde referans değerleri müşteriler olmuşken bu kez referans değerleri harcama noktaları ve kategoriler olmaktadır. Birinci hesapların ardından yine aynı referans değerleri ile bu kez dolandırıcılık gözlemleri kullanılarak risk puanı üretilmiştir. Risk Puanı bilgisi, temelde periyot içindeki dolandırıcılık gözlem sayısının ilk adımda hesaplanan periyot içindeki toplam işlem sayısına oranıdır. Eğer periyot

içinde hiç işlem yoksa risk puanı sıfır olur. Sonuçta hem işlem noktası hem de harcama kategorisi için ayrı ayrı işlem sayısı ve risk puanı hesaplandığında toplamda yirmi adet yeni özellik üretilmiş olur.

Tablo 3.8 İşyeri detayına ait yardımcı özellikler

musteri_no	5139	4853	5144	5180	5280	4879
Islem_tarihi	28.09.21	25.09.21	28.09.21	28.09.21	28.09.21	28.09.21
Isyeri_NB_7	26	46	3	19	19	7
Isyeri_NB_15	62	83	9	44	44	8
Isyeri_NB_30	62	195	15	87	87	15
Isyeri_NB_45	62	251	23	113	113	16
Isyeri_NB_60	62	333	32	148	148	23
Isyeri_RISK_7	0.35	0.37	0.33	0.00	0.00	0.14
Isyeri_RISK_15	0.39	0.25	0.33	0.16	0.16	0.25
Isyeri_RISK_30	0.39	0.26	0.27	0.15	0.15	0.13
Isyeri_RISK_45	0.39	0.20	0.17	0.12	0.12	0.13
Isyeri_RISK_60	0.39	0.17	0.13	0.09	0.09	0.09

Tablo 3.9 Kategori detayına ait yardımcı özellikler

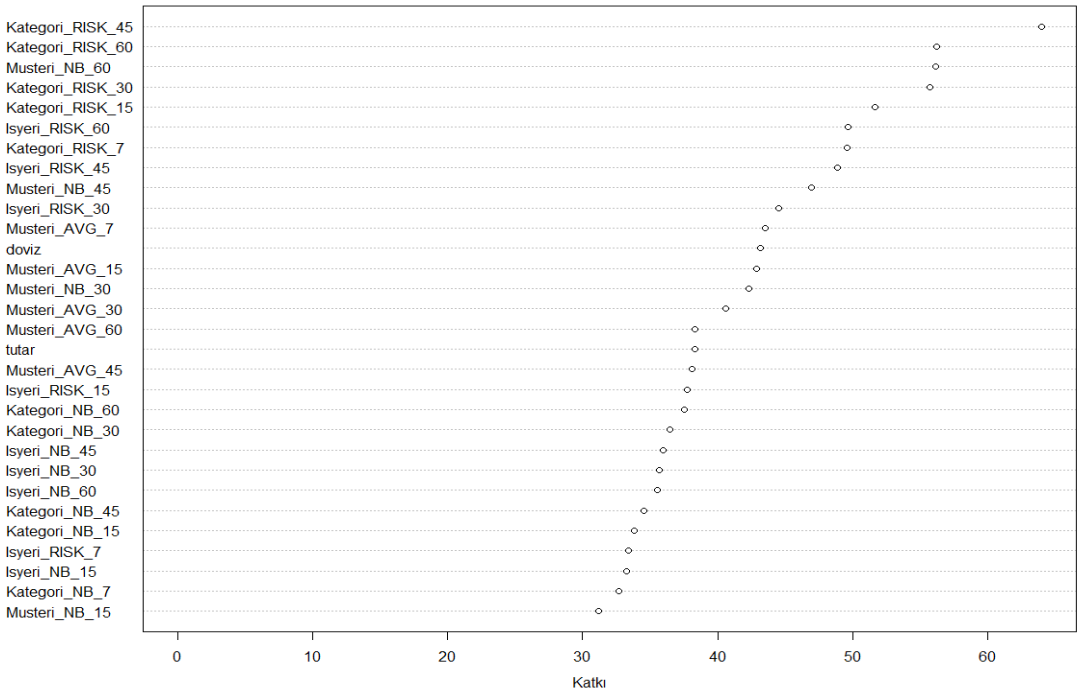
islem_no	107173	79422	109331	7022	7021	32734
musteri_no	5139	4853	5144	5180	5180	4879
islem_tarihi	28.09.21	25.09.21	28.09.21	28.09.21	28.09.21	28.09.21
Kategori_NB_7	63	14	63	63	63	63
Kategori_NB_15	141	25	141	141	141	141
Kategori_NB_30	209	81	209	209	209	209
Kategori_NB_45	260	123	260	260	260	260
Kategori_NB_60	334	164	334	334	334	334
Kategori_RISK_7	0.43	0.43	0.43	0.43	0.43	0.43
Kategori_RISK_15	0.40	0.24	0.40	0.40	0.40	0.40
Kategori_RISK_30	0.28	0.23	0.28	0.28	0.28	0.28
Kategori_RISK_45	0.23	0.15	0.23	0.23	0.23	0.23
Kategori_RISK_60	0.18	0.12	0.18	0.18	0.18	0.18

Yukarıda Tablo 3.8 ve Tablo 3.9 ile seçilen işyeri ve kategoriler için üretilmiş yardımcı özellikler görülmektedir. Periyodik olarak harcama sayısını ifade eden özellikler Isyeri_NB ve Kategori_NB ile isimlendirilirken, risk puanının hesaplandığı özellikler ise Isyeri_RISK ve Kategori_RISK ile isimlendirilmiştir. Müşteri özelliklerinde olduğu gibi her hesaplama, mevcut işlem tarihine göre geriye yönelik yapılmıştır.

Tablo 3.10 Yardımcı özelliklerin tahmin performansına etkisi

Gözlem İçeriği	F_1 Skoru	Belirlilik	Kesinlik
Sadece Temel Değişkenler	%47	%55	%42
Müşteri Harcama Alışkanlıkları	%30	%22	%44
İşyeri Değişkenleri	%58	%52	%65
Kategori Değişkenleri	%57	%54	%61
Müşteri ve İşyeri	%77	%77	%78
Müşteri ve Kategori	%70	%70	%69
İşyeri ve Kategori	%63	%66	%60
Yardımcı Değişkenlerin Hepsi	%78	%78	%78

Sonuç olarak yukarıda elde edilen müşteri, iş yeri ve kategori bazlı otuz farklı özellik ile tahmin sonuçlarının değişimini gözlemlemek amacıyla model tekrar çalıştırılmıştır. Elde edilen tahmin sonuçları Tablo 3.10 ile özetlenmiştir. Buna göre temel değişkenler kullanılarak %47 olarak sağlanan F_1 skoru %78 e kadar yükselmiştir. Bu başarı, üretilen yardımcı değişkenlerin dolandırıcılık işlem tespitinde etkin olarak kullanılabileceğini göstermektedir. Sonradan eklenen değişkenlerin tahminlemede hangi düzeyde katkı verdiği de ayrı ayrı incelendiğinde, İşyeri ve Kategori değişkenlerinin hem ayrı ayrı hem de Müşteri değişkenleriyle birlikte başarılı olduğu gözlemlenmiştir.

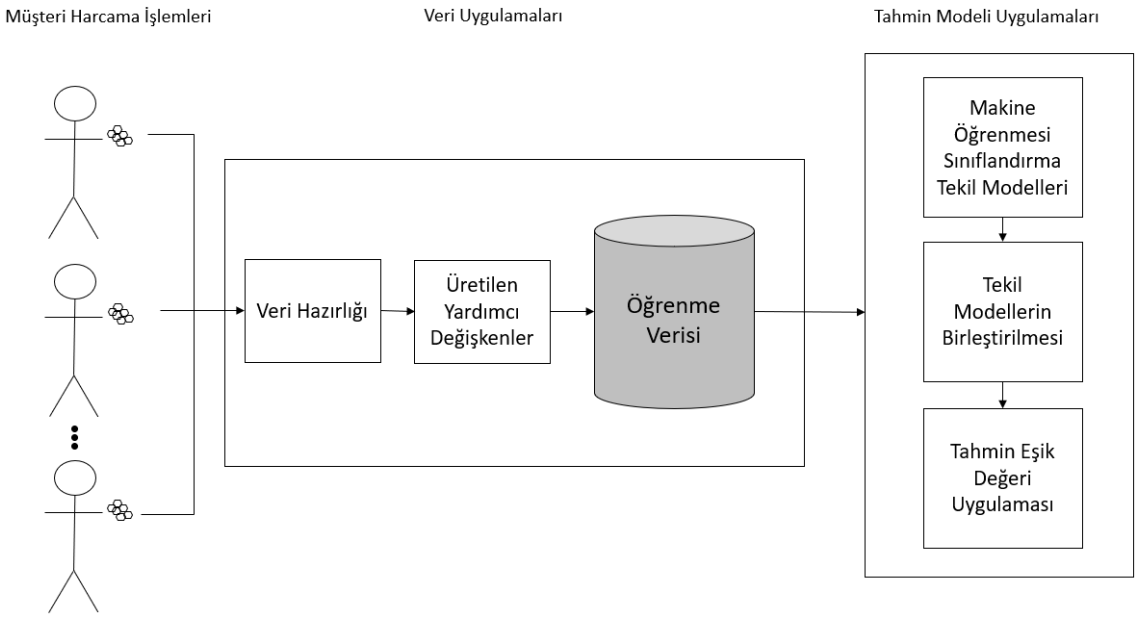


Şekil 3.2 Değişken katkıları

Detaylı katkı grafiğini gösteren Şekil 3.2 ile, risk puanı özelliklerinin net bir şekilde tahmin sonuçlarına yüksek katkısı olduğu ve ortalama işlem sayısı özelliklerinin de risk puanının ardından katkı sıralamasında üstlerde olduğu görülmektedir. Diğer yandan işyeri işlem sayısı, kategori işlem sayısı ve müşteri ortalama harcama tutarı özelliklerinin diğer yardımcı değişkenlere göre daha zayıf bir katkı verdiği anlaşılmıştır. Yine de tüm yardımcı değişkenlerin en az temel değişkenler kadar tahminlemede katkısı olduğu ortaya çıkmıştır. Özellikle de toplamda ulaşılan %78 düzeyindeki F_1 skoruna ek olarak %78 Belirlilik ve %78 Kesinlik elde edilmiş olması tahmin modelinin aynı anda hem toplam dolandırıcılık gözlemlerini yüksek başarı ile tespit ettiğini hem de tahminlerin normal gözlemleri dolandırıcılık olarak sınıflandırmadığını gösterir. Sonuçta yardımcı değişkenlerin tahmin başarısını büyük ölçüde iyileştirebildiği ispatlanmıştır. Ancak yine de daha ileri iyileştirmeler yapılabileceğinin ve farklı risk düzeylerine göre daha uygun modellerin bulunabileceğinin araştırılması gerekmektedir.

3.4 Tahmin Modeli Uygulamaları

Dolandırıcılık tespiti uygulamasının başarısını arttırmak amacıyla önceki bölümde veri odaklı gerçekleştirilen uygulamaların ardından tahmin modeli optimizasyonu üzerine de çalışmalar yapılabilir. Şekil 3.3 şu ana kadar yapılan veri uygulamaları sonrasında elde edilen öğrenme verisi kullanarak gerçekleştirilecek tahmin modeli uygulamalarını göstermektedir. Bu çalışmalar hem anahtar performans göstergesi olan F_1 skorunu iyileştirmeyi amaçlayabileceği gibi hem de daha yüksek ya da daha düşük risk yaklaşımlarının olduğu farklı amaçları da öne çıkarabilir. Örneğin, Belirlilik odaklı yaklaşım istenirse fazla gerçek dolandırıcılık gözlemi tespit edilebilir. Ancak karşılığında yanlış tahmin sayısının artmasına yani Kesinlik değerinin azalmasına sebep olur. Başlangıçtaki tahmin modeli yaklaşımında seçilen öğrenme metodu Rastgele Orman algoritması temelde F_1 skoru açısından başarılı sonuçlar vermiş olsa da modeller üzerinde çeşitli yöntemler denenerek iyileştirmeler yapılabilir.



Şekil 3.3 Uygulama adımlarının özeti

Bu denemelerden ilki, farklı öğrenme algoritmalarını tek tek denemek olmuştur. Bu kapsamda Rastgele Orman algoritması dışında Lojistik Regresyon, K-En Yakın Komşu, GBM (Gradient Boosting Machines), Destek Vektör Makineleri, Adaboost ve Yapay Sinir Ağları ayrı ayrı kullanılmıştır. Tüm öğrenme algoritmaları, önceki adımlarda baz model yardımıyla şekillendirilmiş olan, yardımcı değişkenlerin tamamını içeren nihai uygulama verisine ait 20,000 gözlem sayısı boyutuna sahip ve %4 dolandırıcılık oranı olan öğrenme verisi ile öğretilmiş, %1 dolandırıcılık oranına sahip 30,000 gözlemden oluşan test verisi ile test edilmiştir.

Ardından, tekil öğrenme modellerinin arasından çeşitli kriterlere göre seçilen başarılı modeller ile birleştirme yöntemleri denenmiştir. Bu kapsamda risk yaklaşımlarına göre farklı modeller birleştirme için seçilmiş, hangi yöntemler kullanılarak birleştirme yapılacağı da ek bir uygulama çalışması olmuştur. İyimser, kötümser ve ağırlıklı ortalama gibi temel karar verme yöntemlerine ek olarak üst katman öğrenme modelleri ile de birleştirmeler yapılmıştır.

Tahmin modeli uygulamalarında son olarak ise, öğrenme algoritmalarının tahmin sonuçlarının elde edilmesi sürecine odaklanılmıştır. Bu amaçla, varsayılan iki değerli sınıflandırma tahmin eşik değeri olan 0.5 değerinden daha farklı eşik değerleri kullanılarak öğrenme modellerinde test sonuçlarının tepkisi ölçülmüştür. Böylelikle varsayılandan daha yüksek eşik değeri ile daha düşük eşik değeri

arasında risk yaklaşımına göre nasıl sonuçlar elde edilebileceği araştırılmıştır. Sonuçta tüm yöntemler kıyaslanarak, sağladıkları katkılar yorumlanmıştır.

3.4.1 Tekil Makine Öğrenmesi Modelleri

Dolandırıcılık tespiti tahmininde kullanılan mevcut model ile elde edilen sonuçlar incelendiğinde Rastgele Orman öğrenme algoritmasının başarılı olduğu düşünülmüştür. Fakat kesin karara varmak adına farklı öğrenme algoritmalarının da sonuçlarını görmek gereklidir. Literatürde bilinen sınıflandırma tahmin algoritmaları içinden öne çıkan diğer algoritmalar şöyle sıralanabilir [5,8]:

- Lojistik Regresyon
- K - En Yakın Komşu
- GBM
- Adaboost
- Destek Vektör Makineleri
- Yapay Sinir Ağları

Yukarıda sıralanan öğrenme algoritmalarının tahmin sonuçlarında göstereceği performans araştırılırken önceki adımlarda kullanılan veri özellikleri seçilmiştir. Böylece Rastgele Orman algoritması ile aynı koşullar altında sonuçlar karşılaştırılmış olur.

Tablo 3.11 Farklı öğrenme algoritmalarının tahmin sonuçları

Öğrenme Yöntemi	F_1 Skoru	Belirlilik	Kesinlik
Rastgele Orman	**%77**	%76	%78
Adaboost	%72	%**85**	%62
Destek Vektör Makineleri	%69	%77	%63
Yapay Sinir Ağları	%66	%83	%55
Lojistik Regresyon	%66	%77	%57
GBM	%66	%62	%**71**
K - En Yakın Komşu	%46	%52	%42

Elde edilen sonuçlara Tablo 3.11 ile bakılırsa, ilk seçilen Rastgele Orman algoritmasının F_1 skoru bazında en iyi sonuca ulaştığı teyit edilmiş olur. Ardından gelen Adaboost ve Destek Vektör Makineleri ilse birbirlerine yakın sonuçlar vermiştir. Ancak bu iki algoritma ile Rastgele Orman kıyaslandığında, F_1 skoru dışında da farklılıklar olduğu görülmektedir. Belirlilik değerleri en az Rastgele

Orman kadar yüksek olmalarına rağmen Kesinlik değerleri net olarak daha zayıf ölçülmüştür. Bu da dolandırıcılık gözlemlerinin tespit edilmesine karşılık büyük yüzde ile normal işlemlerin de dolandırıcılık olarak tespit edildiğini gösterir. Dolayısıyla risk almayı kabul eden bir yaklaşıma göre Adaboost en iyi Belirlilik değeri ile öne çıkmaktadır. 300 gözlemden oluşan dolandırıcılık test gözlemlerinin 261'ünü doğru bilmiştir (%85 Belirlilik). Adaboost bu başarının karşılığında 179 adet normal gözlem değerini dolandırıcılık olarak işaretlemiştir (%62 Kesinlik). Sonuçta Adaboost öğrenme algoritması yaptığı dolandırıcılık tahminlerinin yarısından azını hatalı yaparak toplamda test dolandırıcılık gözlemlerinin çoğunu yakalayabilmiştir.

Diğer yandan F_1 skoru bazında daha alt sıralarda yer alan GBM algoritması ise kullanılan tüm yöntemler arasında %71 ile en yüksek ikinci Kesinlik değerine ulaşan yöntem olmuştur. Başka bir deyişle az hatalı tahmin yapılmış ancak buna karşılık olarak gerçek dolandırıcılık gözlemlerinin 187 tanesi doğru yakalanabilmiştir (%62 Belirlilik). Dolayısıyla riskten kaçınma yaklaşımı izlenmek istenirse GBM en iyi sonuç veren yöntemlerden biri olacaktır. Model, 300 gözlemden oluşan dolandırıcılık verilerinin 187'sini doğru yakalarken 78 tane normal gözlemi yanlış bir şekilde dolandırıcılık olarak belirlemiştir. Son sırada yer alan K – En Yakın Komşu algoritmasının sonuçlarına bakıldığında ise, ne Belirlilik yönünden ne de Kesinlik yönünden ayrışamadığı görülmüştür. Bu durumun sonucunda da F_1 skoru olarak son sırada yer almıştır. Dolayısıyla K – En Yakın Komşu algoritmasının dolandırıcılık tespiti amacıyla kullanılmasının faydalı olmayacağı görülmüştür.

Sonuçta toplamda yedi farklı öğrenme yönteminin performans göstergeleri kıyaslandığında, dengeli yaklaşım ile F_1 skoru değeri en iyi olan Rastgele Orman, riskli yaklaşımla Belirlilik değeri en iyi olan Yapay Sinir Ağları, risksiz yaklaşımla ise Rastgele Orman ve GBM yöntemleri öne çıkmıştır. Diğer yandan modellerin tahmin başarısı, kendi aralarında çeşitli yollarla birleştirmeler yaparak üst katman bir tahmin modeli oluşturulduğunda iyileştirilebilir. Bir sonraki bölümde mevcut modellerin birleştirilmesi konusu ile uygulama çalışması genişletilmiştir.

3.4.2 Tekil Modellerin Birleştirilmesi

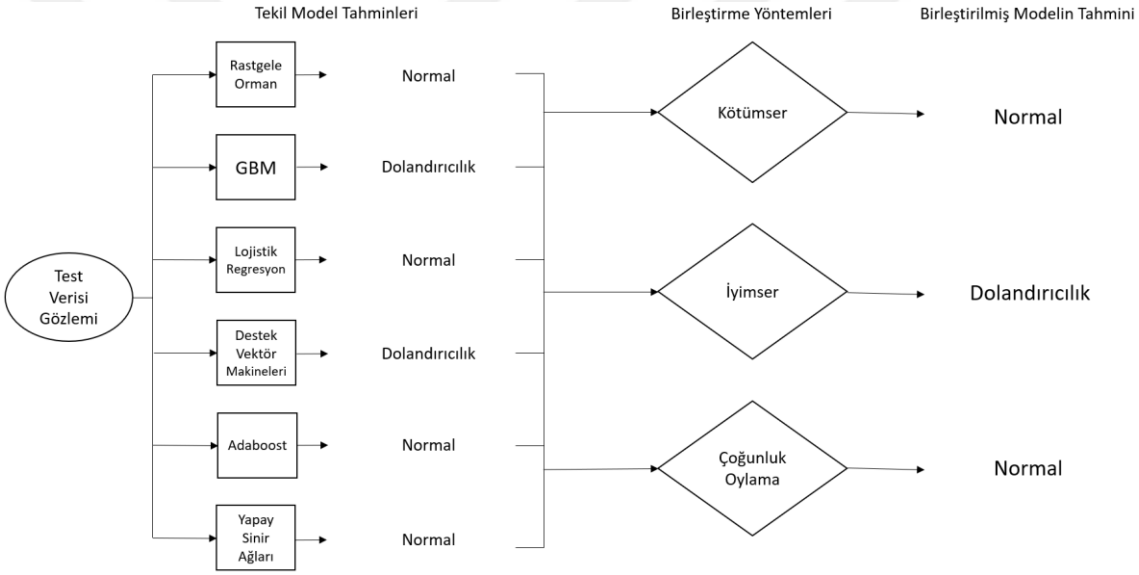
Dolandırıcılık tespiti amacıyla yapılan makine öğrenmesi uygulamalarında bir diğer yöntem de topluluk öğrenmesi oluşturma yöntemi, başka bir deyişle model birleştirmedir. Bu yöntemde birbirinden bağımsız tahmin modellerinin farklı yöntemlerle ortak bir tahmin modelinde birleştirilmesi amaçlanır [11,18,19,20]. Girdi olarak seçilen tekil modeller farklı yaklaşımlara göre belirlenebilir. Tüm tekil modeller kullanılabileceği gibi, sadece F_1 skoruna göre en iyi yöntemler ya da Belirlilik ve Kesinlik değerlerine göre en iyi modeller de seçilebilir. Ayrıca birleştirme işlemi, girdi olarak kategorik sınıflandırma değerleri üzerinden yapılabileceği gibi tahmin olasılıkları ile de yapılabilir ancak olasılık değerleri kullanıldığında birleştirme yöntemleri farklılaşır. Bu bölümde hangi tekil modellerin birleştirileceğini belirlemek adına üç farklı durum incelenmiştir:

- Temel Seçim: Tüm modelleri birleştirmek
- Performans Bazlı Seçim: F_1 skoruna göre en iyi modelleri birleştirmek
- Hibrit Seçim: Farklı göstergelerden en iyi modelleri birleştirmek

Tüm modellerin seçilmesi durumu, aşırı öğrenme ve eksik öğrenme sorunlarına karşı iyi bir alternatif olarak görülse de tekil modeller incelendiğinde F_1 skoru ve diğer göstergelerin zayıf sonuçlarla elde edilen avantajı geri vereceği tahmin edilebilir. Bu durumun önüne geçilmesi adına son sırada yer alan yöntem olan K – En Yakın Komşu algoritması birleştirme işleminden çıkarılmıştır. Diğer yandan performansa dayalı seçim yöntemiyle anahtar gösterge F_1 skoruna göre belirlenen en iyi üç model Rastgele Orman, Adaboost ve Destek Vektör Makineleri olur. Hibrit seçim yöntemi ise, F_1 skoru, Belirlilik ve Kesinlik göstergelerinde en iyi olan modeller seçilir. Böylece F_1 skoruna göre seçilen önceki gruptan farklı olarak Destek Vektör Makinelerinin yerine GBM modeli dahil edilmiş olur. Gruplanacak modelleri belirlemenin ardından birleştirme yöntemleri uygulanmaya başlanmıştır. Uygulama kapsamında ele alınan on iki birleştirme yöntemi üç ayrı başlıkta özetlenebilir:

- İyimserlik/Kötümserlik ve Uzlaşma Yaklaşımları
- Ortalama/Ağırlıklı Ortalama Yaklaşımı
- Üst Katman Makine Öğrenmesi Yaklaşımı

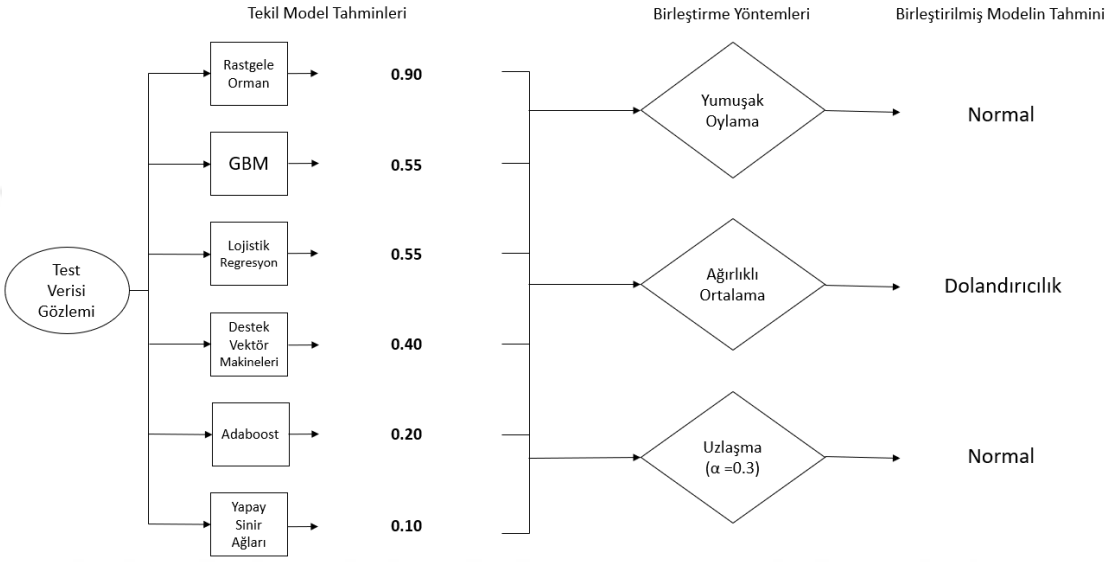
İyimserlik ve Kötümserlik yaklaşımları mevcut tahmin sonuçlarını bir araya getirirken maksimum ve minimum değerleri referans alır. İyimserlik yaklaşımı, harcama işlemi için modellerin yaptığı tahmin sonuçları arasında tek bir tane bile dolandırıcılık sonucu varsa nihai tahminini dolandırıcılık olarak belirler. Kötümserlikte ise zıt olarak, tahmin sonuçlarında yalnız bir tane normal işlem sonucu varsa dahi nihai tahmin normal işlem olarak belirlenir. Uzlaşma (Huwicz) yaklaşımında ise karar verici tarafından 0 ile 1 belirlenen bir katsayı ile iyimserlik ve kötümserlik sonuçlarının toplamı hesaplanır. Karar verici risk algısına göre ya da iyimserlik ve kötümserlik yaklaşımlarının performansına göre katsayıyı değiştirebilir. Şekil 3.4 ve Şekil 3.5 ile İyimser, Kötümser ve Uzlaşma birleştirme yöntemleri gösterilmiştir.



Şekil 3.4 İyimser, kötümser ve çoğunluk oylama birleştirme yöntemleri

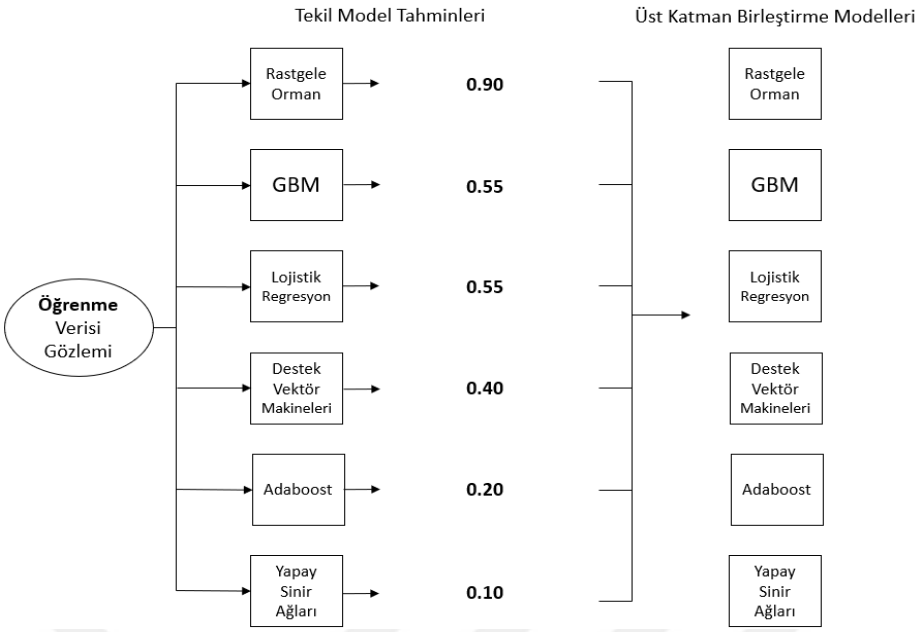
Diğer yandan ortalama yaklaşımları kendi içinde 3'e ayrılır. Çoğunluk oylama yönteminde tekil modellerin tahmin çıktıları ile karar verilir. Eğer modellerin en az yarısı dolandırıcılık sonucu vermişse nihai sonuç dolandırıcılık olarak belirlenir, aksi durumda işlem normal kabul edilir. Yumuşak Oylama (Soft Voting) yönteminde ise Çoğunluk Oylamadan farklı olarak tekil modellerin çıktı sonucu yerine tahmin olasılıklarının ortalaması hesaplanır. Örneğin üç model için, bir modelde 0.10 olasılıkla dolandırıcılık olarak belirlenen işlem başka iki modelde 0.60 ve 0.50 olasılık hesaplanmışsa, ortalama olasılık değeri olan 0.40 sonucu neticesinde işlemin normal olarak belirlenmesine kara verilir. Öte yandan aynı

işlem Çoğunluk Oylama yönteminde üç modelden ikisinin oyuyla dolandırıcılık işlemi olarak belirlenecektir. Ağırlıklı Ortalama yönteminde ise, tekil modellerin tahminleri kendi F_1 skoruna göre ağırlıklandırılıp ortalaması alınır. Model ağırlıkları hangi seçim yöntemleriyle modellerin birleştirileceğine göre ayrı ayrı hesaplanır. Şekil 3.4 ve Şekil 3.5 ile Çoğunluk Oylama, Yumuşak Oylama ve Ağırlıklı Ortalama birleştirme yöntemleri gösterilmiştir.



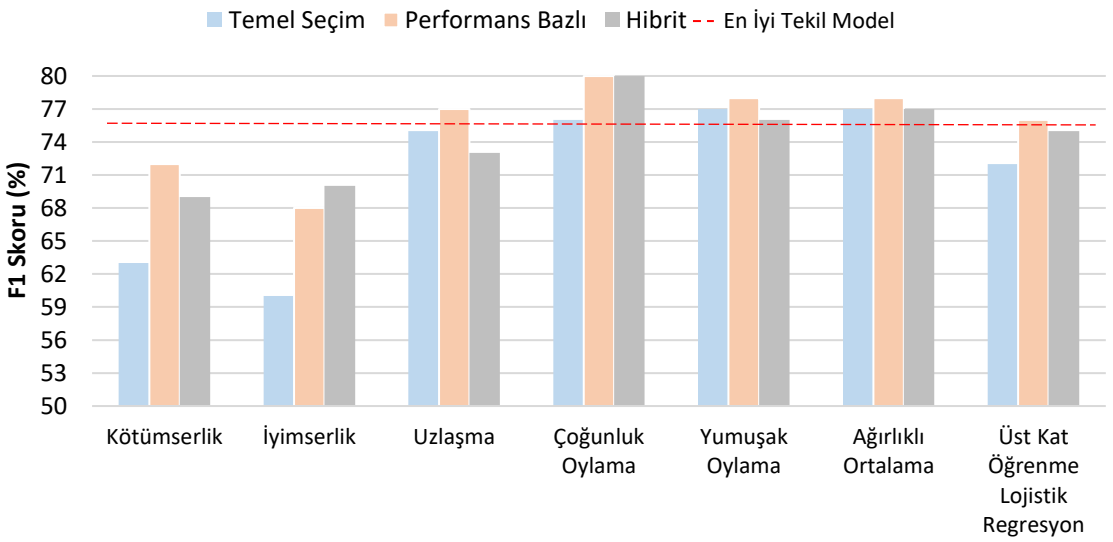
Şekil 3.5 Yumuşak oylama, ağırlıklı ortalama ve uzlaşma yöntemleri

Diğer yandan tekil modellerin birleştirilmesi için makine öğrenmesi yöntemleri de kullanılabilmektedir. Bu yaklaşımda tekil modeller için kullanılan öğrenme yöntemleriyle ayrı ayrı birleştirme yöntemi olarak üst kat modeller kurulmuştur. Öğrenme işlemi, tekil modellerin öğrenme verisindeki tahmin olasılıkları ile yapılırken test işlemi ise tekil modellerin test verisindeki tahmin olasılıkları ile yapılır. Yapılan üst kat öğrenme modelleri arasında en iyi performansı gösteren ise Lojistik Regresyon olmuştur. Şekil 3.6 ile üst katman öğrenme algoritmalarıyla yapılan birleştirme modelleri gösterilmiştir.



Şekil 3.6 Üst katman birleştirme yöntemleri

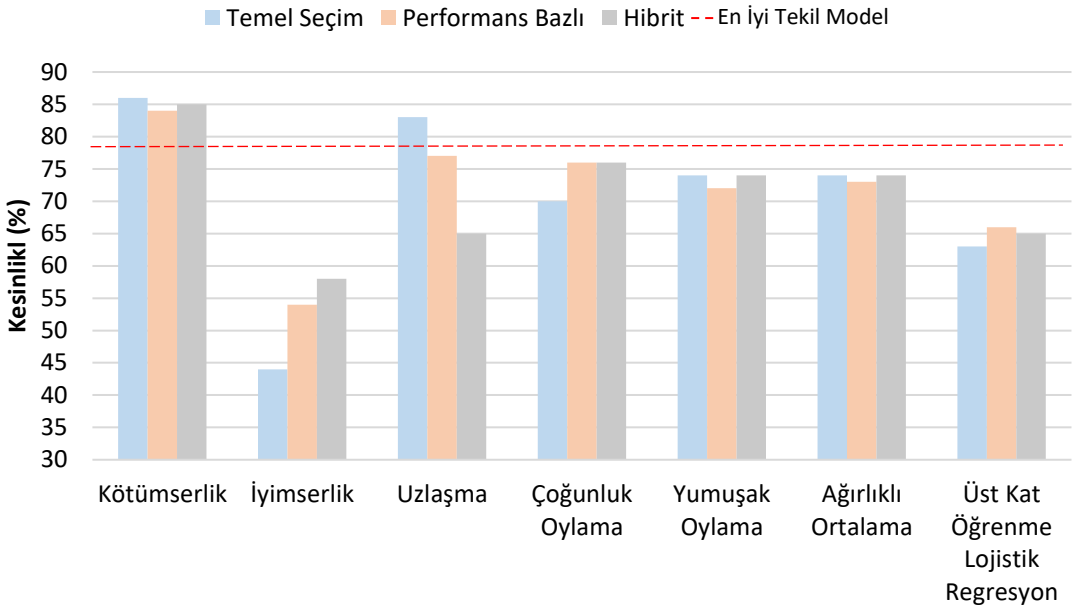
Sonuçlara bakıldığında Şekil 3.7 birleştirmelerin F_1 skorları açısından performansını göstermektedir. Toplamda 12 farklı birleştirme yönteminin, 3 farklı seçim yöntemi ile uygulanması sonucunda elde edilen F_1 skorları incelendiğinde, dört farklı örnekte önceki adımda tekil modellerle elde edilmiş olan en iyi skor değeri geçildiği görülmüştür. Buna göre en iyi sonuç veren birleştirme yöntemi çoğunluk oylama olurken, performans bazlı ve hibrit seçim yöntemleri temel seçime göre öne çıkmıştır. Ortalama bazlı birleştirme yöntemleri genel olarak daha iyi sonuçlar ortaya koymuş, üst katman makine öğrenmesi birleştirme yöntemleri de genel olarak kötü performans göstermiştir. Toplamda 5 birleştirme işleminde performans bazlı seçim öndeyken, 2 birleştirme yönteminde ise hibrit seçim daha iyi sonuç vermiştir. İyimserlik, Kötümserlik ve Uzlaşma yöntemleri kıyaslandığında en iyi sonuçların Uzlaşma yöntemi ile elde edildiği görülmüştür. Sonuçta da yapılan model birleştirme uygulaması sayesinde tekil tahmin modellerinin F_1 skoru performansı yukarıya çekilebilmiştir.



Şekil 3.7 Birleştirme yöntemlerinin F_1 skoru performansı

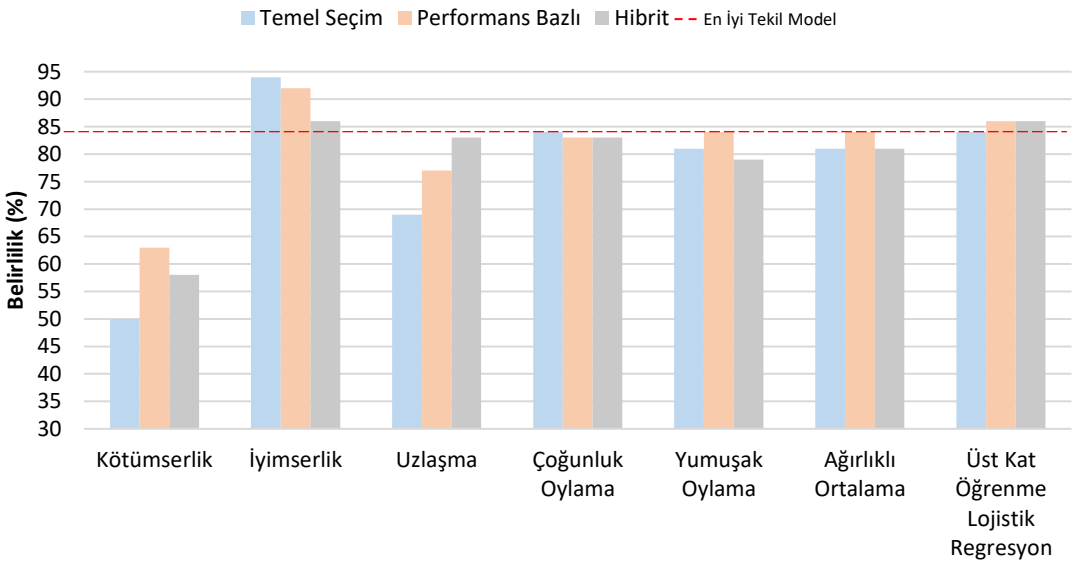
Kesinlik bazında birleştirme yöntemlerinin performans sonuçları Şekil 3.8 ile gösterilmiştir. Tekil modellerle elde edilebilen maksimum kesinlik değeri olan %78 değerini aşan dört farklı birleştirme örneği elde edilmiştir. F_1 skorunda görülen durumun aksine, Kesinlik değerleri açısından en yüksek sonuç veren birleştirme yöntemi Kötümserlik yaklaşımı olmuştur. Seçim yöntemleri üzerinden incelendiğinde ise, yine önceki durumun tersine Temel seçim yöntemi performans bazlı seçim ve hibrit seçime göre yüksek sonuçlar vermeye daha yatkındır. F_1 skorlarında görülen durumun bir benzeri olarak en zayıf performans gösteren birleştirme yöntemleri üst katman makine öğrenmesi birleştirme yöntemleridir.

Belirlilik performans göstergesi ile birleştirme yöntemleri incelendiğinde, Şekil 3.9 en yüksek sonuç veren birleştirme yönteminin Temel seçim yöntemiyle İyimserlik Yaklaşımı olduğunu göstermektedir. Diğer yandan Kötümserlik ve Uzlaşma yöntemleri önceki durumların aksine zayıf kalmıştır. Yine önceki durumlardan farklı olarak bu kez üst katman makine öğrenmesi birleştirme yöntemleri genel olarak iyi sonuçlar üretmiştir. Tekil tahmin modelleri ile elde edilebilen maksimum Belirlilik değeri olan %85, toplamda 5 farklı birleştirme örneği ile geçilmiştir.



Şekil 3.8 Birleştirme yöntemlerinin Kesinlik performansı

Sonuç olarak, tekil modellerin performansını iyileştirme amacıyla yapılan birleştirme uygulaması farklı seçim ve birleştirme yöntemleri ile istenen faydayı sağlayabilmiştir. F1 skoru performans bazlı ve hibrit seçimlerde %77 düzeyinden %80'e yükselmiş, Kesinlik değeri temel seçim yaklaşımı ve kötümserlik birleştirme yöntemiyle %78'den %86'ya çıkmış, Belirlilik değeri ise temel seçim yaklaşımı ile iyimserlik birleştirme yöntemiyle %85 düzeyinden %94'e kadar yükselmiştir.



Şekil 3.9 Birleştirme yöntemlerinin Belirlilik performansı

Birleştirme modellerinde tüm sonuçlar varsayılan olasılık eşik değeri ile ölçülmüştür. Bir sonraki uygulama çalışmasıyla, varsayılan eşik değerinin tekil ve birleştirme modellerde performansa etkisi ölçülecektir.

3.4.3 Tahmin Eşik Değeri Uygulaması

Dolandırıcılık tespiti amaçlanan sınıflandırma tahmin modellerinde aynı anda iki farklı zorluğun üstesinden gelinmeye çalışılır. Kurulan modelin, hem yüksek Belirlilik ile gerçek dolandırıcılık gözlemlerinin hepsini bulmaya çalışıp hem de yüksek Kesinlik ile az hata yapması beklenir. Ancak gerçek hayatta bu başarıyı aynı anda elde etmek çok zordur. Önceki adımlarda seçilen tekil öğrenme yöntemleri arasından pek azı dengeli bir test sonucu ortaya koyabilmişken, çoğunlukla iki amaçtan birini daha iyi sağlamış diğerinde ise daha zayıf kalmıştır [5,8]. Birleştirme modellerde durum çok farklı olmazken Kesinlik ve Belirlilik değerinin birbirine uzak olması elde edilen performansı düşürmüştür. Bu bölümde, Kesinlik ve Belirlilik göstergelerinde daha dengeli sonuçlar almaya yönelik farklı bir çalışma olarak eşik değerler üzerinde araştırma yapılacaktır.

Eşik değerlerine göre tahmin kararlarının değişimini gösteren Tablo 3.12 incelendiğinde, esas tahmin sonuçları olan olasılık çıktılarının sabit olduğu fakat tahmin kararının eşik değere göre değiştiği kolaylıkla görülebilmektedir. Test verisinde gerçek değerlerin bilinmemesinden dolayı doğrudan hangi gözlemde hangi eşik değerin uygun olduğunu söylemek mümkün değildir. Fakat tüm tahminler için varsayılan eşik değeri değiştirerek tahmin kararını belirlemek mümkündür. Ancak bu işlem tahmin başarısına fayda sağlayabileceği gibi zarar da verebilir [17,25,26].

Tablo 3.12 Olasılık eşik değerlerinin tahmin sonucuna etkisi

Test Gözlemleri	Tahmin Olasılık Değerleri		Eşik Değer	Tahmin Kararı	Gerçek Değer
	Normal	Dolandırıcılık			
1	0.6228	0.3772	**0.35**	**Dolandırıcılık**	**Dolandırıcılık**
			0.50	Normal	Dolandırıcılık
			0.65	Normal	Dolandırıcılık
2	0.4540	0.5460	0.35	Dolandırıcılık	Normal
			0.50	Dolandırıcılık	Normal
			0.65	**Normal**	**Normal**

Dolayısıyla eşik değerlerin tahmin başarısına etkisi üzerinden bir yorum yapmadan önce performans kriteri olarak ele alınan F_1 skoru, Belirlilik ve Kesinlik sonuçlarının tepkisine bakmak daha doğru olacaktır. Örnek olarak önceki uygulama adımlarında öne çıkan tahmin modeli Rasgele Orman için eşik değerlerine bağlı olarak performans çıktıları Tablo 3.13 ile gösterilmiştir. Buna göre, varsayılan değer olan 0.50 yükseldikçe Kesinlik değerlerinde önemli bir iyileşme olurken Belirlilik değerleri ise tam tersi yönde azalmıştır. Bu durumun ana sebebi tahmin modelinin yüksek eşik değerinde daha az dolandırıcılık sınıflandırması yapmasında yatar. Eşik değer düştüğünde de dolandırıcılık olarak sınıflandırılan tahmin sayısı artış gösterir. Sonuçta Belirlilik değeri, yani gerçek dolandırıcılık gözlemlerinin doğru tahmin edilme oranı yükselir fakat karşılığında hatalı dolandırıcılık tahminleri de artar ve Kesinlik değeri düşer. Sonuç olarak Belirlilik ve Kesinlik değerlerinin arasındaki fark ne kadar fazla olursa F_1 skorunun azaldığı görülmüştür.

Tablo 3.13 Rastgele orman modelinde eşik değeri enterpolasyonları

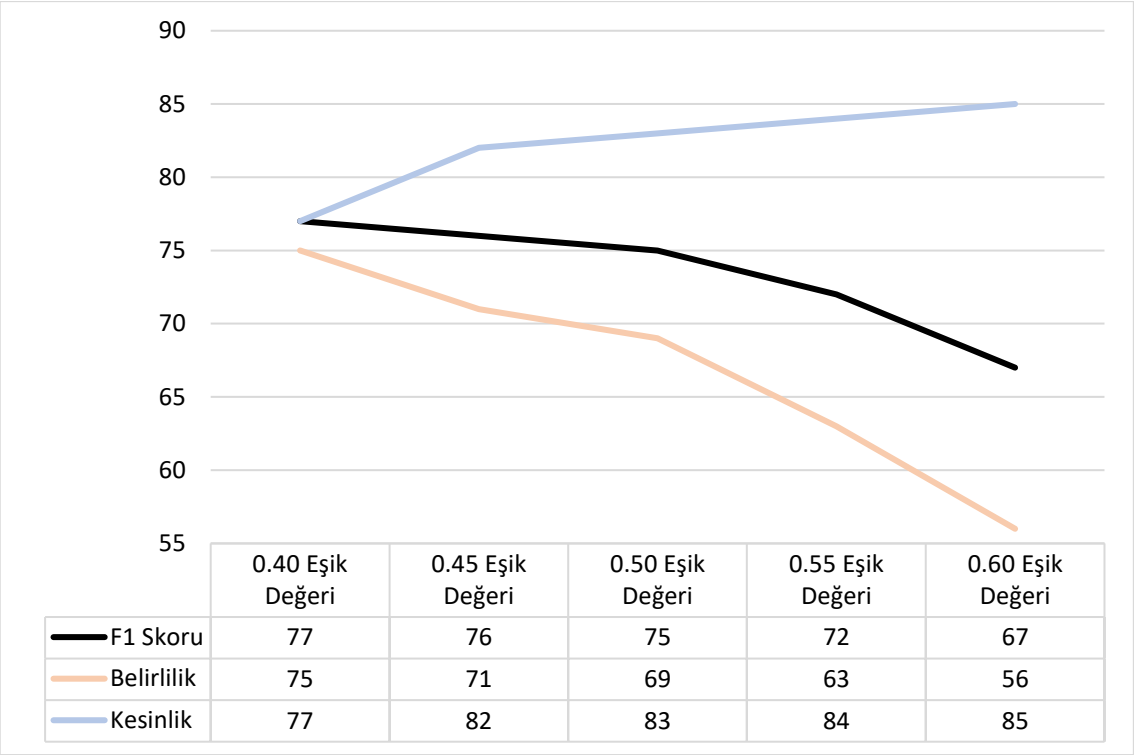
Tahmin Eşiği	Rastgele Orman Modeli		
	F_1 Skoru	Belirlilik	Kesinlik
0.35	%71	%**91**	%59
0.40	%73	%84	%64
0.45	%76	%82	%72
0.50	%**78**	%78	%78
0.55	%77	%72	%83
0.60	%72	%62	%85
0.65	%64	%52	%**89**

Tahmin eşiğinin varsayılandan faklı değerler alması F_1 skorunu her zaman negatif etkilemeyebilir. Diğer algoritmaların sonuçları incelendiğinde hem yüksek eşik değeri hem de düşük eşik değerinde F_1 skorunun iyileştiği durumlar olduğu görülmektedir. Örneğin Yapay Sinir Ağları modeli varsayılan eşik değerinde %68 F_1 skoru elde edebilmişken, eşik değeri yükseltilerek %65 olduğunda F_1 skorunun da yükseldiği tespit edilmiştir. Bu durumun sebebi yukarıda açıklanan dengesizliği dengeleme ile alakalıdır. Diğer yandan Belirlilik ve Kesinlik sonuçları Rastgele Orman yöntemindekiyle aynı davranışı göstermiştir. Eşik değerleri yükseldikçe

Kesinlik sonuçları artarken Belirlilik değerlerinin ise düşük eşik değerlerinde yükseldiği gözlemlenmiştir.

Tekil modelde eşik değeri uygulamasıyla F_1 skoru açısından daha önce elde edilenlerden daha iyi bir sonuç bulunamamış olsa da, Belirlilik ve Kesinlik değerlerinin ayrı ayrı iyileştirilebildiği tespit edilmiştir. Dolayısıyla, riskten kaçınma yolu izlendiğinde tahmin eşiğinin arttırılması Kesinlik değerini %78 den %89 a kadar çıkarabilir. Risk alma yaklaşımı benimsenirse de Belirlilik değeri %78 den %91 e kadar iyileştirilebilir. Elde edilen bu sonuçlar, eşik değeri uygulamasının risk odaklı farklı performans göstergeleri üzerinden yorumlandığında anlamlı bir katkı sağladığını göstermektedir. Ayrıca Belirlilik ve Kesinlik arasındaki fark azaltıldığında F_1 skorunun iyileştirilebildiği gösterilmiştir.

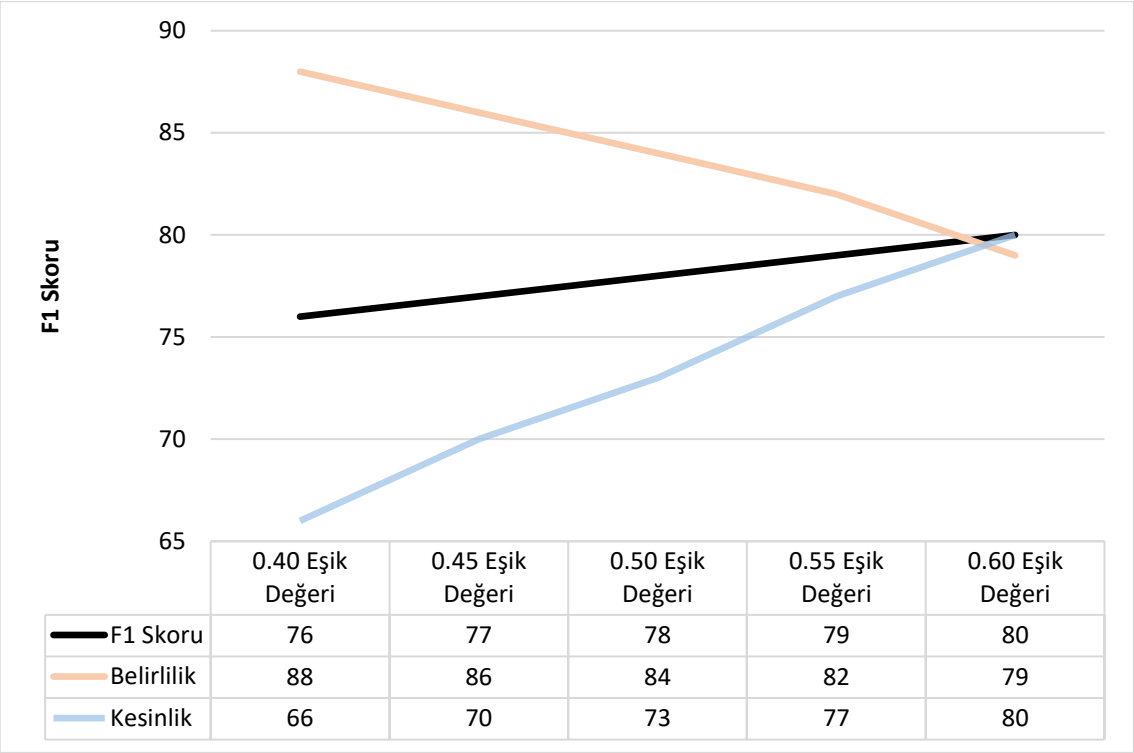
Birleştirilmiş modeller için eşik değeri uygulaması yapıldığında ise ortaya çıkan sonuçlar farklılık göstermektedir. Şekil 3.10 ile Hibrit Seçimde Huwicz (Uzlaşma) Birleştirme Modelinin farklı eşik değerlere göre tahmin başarısı F_1 skoru, Belirlilik ve Kesinlik değerleri üzerinden incelenmiştir. Buna göre, eşik değerinin azaldıkça, modelin Kesinlik ve Belirlilik düzeylerinin birbirine yaklaştığı görülür ve böylelikle F_1 skoru yükselir. Tahmin eşik değerinin 0.40 düzeyinde olduğu durumda en yüksek F_1 skoruna ulaşılabilmişken, eşik değeri arttıkça F_1 skorunun test orantılı olarak düştüğü açıkça ortaya çıkmıştır. Önceki örneklerde görülen durumun benzeri olarak eşik değerinin yükselmesi doğrudan Kesinlik değerinin artmasına ve Belirlilik değerinin azalmasına sebep olur. Bir diğer bulgu ise, Belirlilik değerinin mevcut aralığında eşik düzeyine Kesinlik değerinden daha duyarlı olduğudur. Belirlilik değeri tüm eşik değeri değişimin toplamda 8 puan yükselmişken, Kesinlik değeri daha keskin bir tepki vererek aynı aralıkta 20 puan düşmüştür.



	0.40 Eşik Değeri	0.45 Eşik Değeri	0.50 Eşik Değeri	0.55 Eşik Değeri	0.60 Eşik Değeri
F1 Skoru	77	76	75	72	67
Belirlilik	75	71	69	63	56
Kesinlik	77	82	83	84	85

Şekil 3.10 Hibrit seçimde uzlaşma yönteminin eşik değerleri

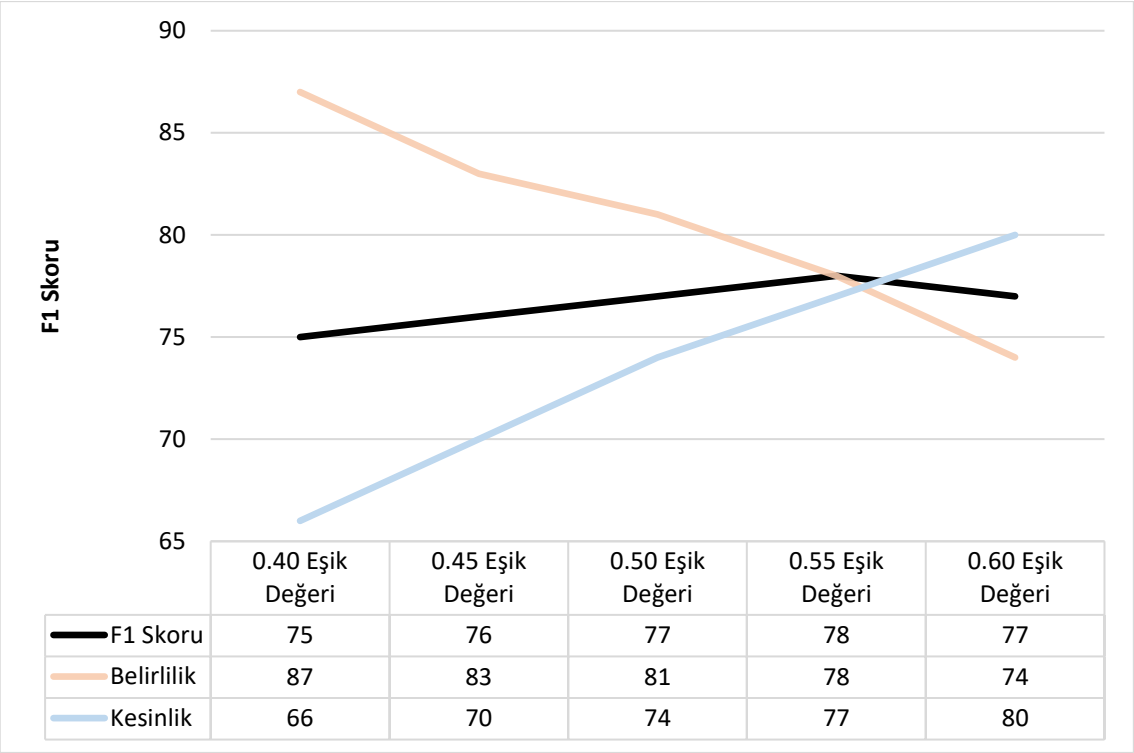
Performans Bazlı Seçim kriteri ve Ağırlıklı Ortalama Birleştirme yönteminin ortaya çıkardığı tahmin modelinde yapılan eşik değeri uygulamasına bakıldığında ise, Şekil 3.11 üzerinde F_1 skorunun eşik düzeyi ile aynı yönde ilerlediği görülmektedir. Tahmin eşik düzeyi 0.60 seviyesine ulaştığında birleştirme modelinde Belirlilik ve Kesinlik aynı değerde birbirine yakınsamıştır. Buna bağlı olarak F_1 skoru da yakın hesaplanmış ve tüm göstergelerde %80 düzeyi ölçülmüştür. Diğer yandan yukarıda incelene uzlaşma yönteminden farklı olarak Belirlilik ve Kesinlik değerleri arasında duyarlılık farklılığı olmadığı söylenebilir. Her iki gösterge de eşik değerinin değişim boyunca zıt yönde ancak yakın düzeylerde değişim göstermiştir. Sonuçta, Performans Bazlı Seçim kriteri ve Ağırlıklı Ortalama Birleştirme yönteminde uygulanan eşik değeri değişimi sayesinde F_1 skoru açısından başarılı olarak nitelendirilebilecek düzeylere kadar performansının iyileştiği görülmüştür. Ancak bu iki örneğe bakarak F_1 skorunun ne kadar daha artacağı anlaşılamamıştır.



	0.40 Eşik Değeri	0.45 Eşik Değeri	0.50 Eşik Değeri	0.55 Eşik Değeri	0.60 Eşik Değeri
F1 Skoru	76	77	78	79	80
Belirlilik	88	86	84	82	79
Kesinlik	66	70	73	77	80

Şekil 3.11 Performans seçimde ağırlı ortalama yönteminin eşik değerleri

Son örnekte ise Temel Seçim yöntemi ve Yumuşak Oylama Birleştirme modelinin sonuçları farklı eşik düzeylerine göre gösterilmiştir. Önceki iki örnekte anlaşılamayan önemli bir durum bu örnek sayesinde açıklığa kavuşturulmuştur. Bu durum basitçe şöyle ifade edilebilir, eşik düzeyinin F_1 skoruna pozitif etkisi ancak Belirlilik ve Kesinlik değerlerinin yakınlaştırılabilmesine bağlıdır. Öyle ki, bir düzeyden itibaren Belirlilik ve Kesinlik birbirinden tekrar uzaklaşmaya başladığında F_1 skoru da aşağı yönde değişmeye başlar. Şekil 3.12 incelendiğinde 0.40'dan başlayarak 0.55'e kadar F_1 skorunun yükselerek tavan yaptığı anlaşılmaktadır. Eşik düzeyi yükselmeye devam ettiğinde ise Belirlilik ve Kesinlik arasındaki fark tekrar artışa geçtiğinden dolayı F_1 skoru azalmaya başlamıştır. Sonuç olarak eşik düzeyi uygulaması yapılırken F_1 skorunun değişimini Belirlilik ve Kesinlik değerlerine bakarak tahmin etmek mümkündür. Bu değerlere göre F_1 skorunun iyileştirilebileceği maksimum seviye yaklaşık olarak görülmüş olur. Çünkü belirli bir noktadan itibaren eşik düzeyinin her iki yöne de hareketi Belirlilik ve Kesinlik değeri arasındaki farkı arttıracaktır.



	0.40 Eşik Değeri	0.45 Eşik Değeri	0.50 Eşik Değeri	0.55 Eşik Değeri	0.60 Eşik Değeri
F1 Skoru	75	76	77	78	77
Belirlilik	87	83	81	78	74
Kesinlik	66	70	74	77	80

Şekil 3.12 Temel seçimde yumuşak oylama yönteminin eşik değleri

Özetlemek gerekirse, bu bölümde tekil ve grup makine öğrenmesi modellerinde tahmin kararlarını belirlerken kullanılan eşik değerlerinin dolandırıcılık tespitinde nasıl kullanılabileceği üzerine yapılan araştırmalar açıklanmıştır. Her bir öğrenme yönteminde elde edilen sonuçlar incelendiğinde hem F_1 skoru hem de Belirlilik ve Kesinlik göstergelerinde çeşitli iyileştirmeler elde edilmiştir. Dolandırıcılık tespiti uygulamasının doğası gereği aynı anda tüm göstergelerde iyileşme elde etmek zor olsa da, her bir gösterge için eşik değerinin nasıl kullanılabileceğine yönelik çıkarımlar elde edilmiştir. F_1 skorunu iyileştirmek adına Belirlilik ve Kesinlik değerini yaklaştırmaya yönelik işlem yapılması fayda sağlamaktayken, salt eşik değerinin Kesinlik ile doğru orantılı Belirlilik ile ise ters orantılı olduğu görülmüştür. Eşik düzeyi uygulaması ile çalışma kapsamında ele alınan tahmin iyileştirme uygulamaları tamamlanmıştır. Ancak elde edilen nihai sonuçları yorumlamadan önce, tüm modelleri meydana getiren verideki bileşenlerin özelliklerini ve tahmin sonuçlarına katkılarını anlamak kritik önem taşımaktadır. Bu amaçla bir sonraki bölüm ile uygulamadaki çıktıların daha doğru yorumlanmasını kolaylaştırmak adına Temel Bileşen Analizi yapılmıştır.

3.5 Temel Bileşen Analizi

Temel Bileşen Analizi, esasen verinin sahip olduğu açıklama yeteneğinin hangi özelliklerden ne ölçüde sağlandığını bulmaya yarayan, veri görselleştirme veya sınıflandırma gibi farklı amaçlarla da kullanılabilen bir yöntemdir. Temel Bileşen Analizi ile daha az sayıda özellik seçerek kabul edilebilir en yüksek varyans aranmaya çalışılır [57,58,59]. Orijinal özelliklerin oluşturduğu düzlemsel eksenlerin haricinde çoğunlukla birden fazla özelliği barındıran yeni eksenler üzerinden temel bileşenleri tanımlar. Uygulama verisi ele alındığında görece az sayıdaki başlangıç özellik sayısı, uygulamanın sonraki adımlarında üretilen yardımcı özelliklerin dahil olmasıyla katlanarak artmıştır. Müşteri, işlem noktası ve kategori bazlı istatistikler kullanılarak yapılan işlemler neticesinde bu uygulamanın elde ettiği tahmin başarısı baz modele kıyasla çok daha yüksek olmuştur. Dolandırıcılık tespiti probleminin zorluk seviyesi de hesaba katıldığında birçok farklı uygulama neticesinde elde edilmiş başarı düzeyinin tekrar azaltılmasını gerektirecek bir sebep yoktur. Dolayısıyla bu bölümde Temel Bileşen Analizi yapılmasındaki amaç özellik sayısını azaltmaktan ziyade, en anlamlı özellikleri ortaya çıkartmak olacaktır.

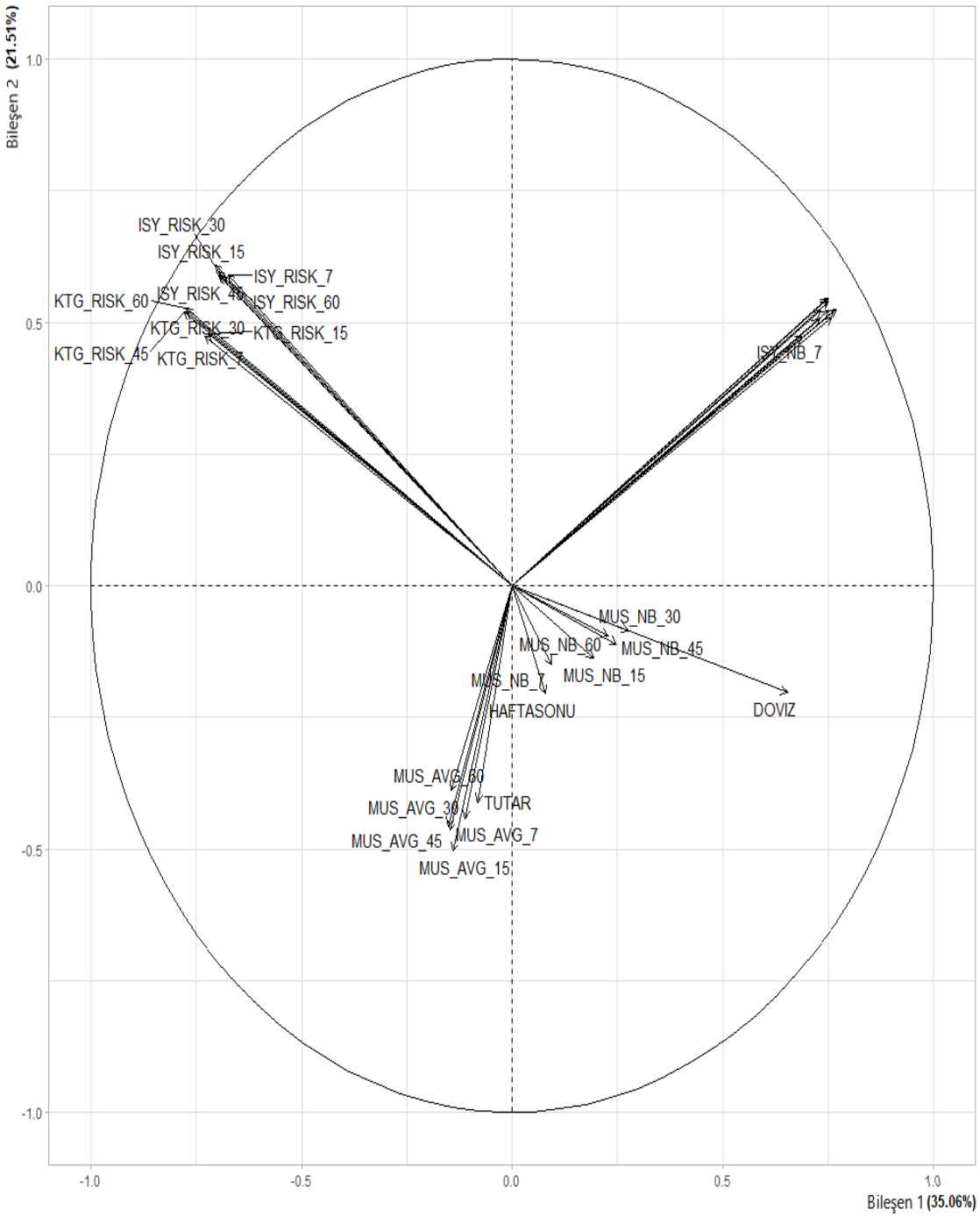
Tablo 3.14 Ölçeklenmiş veri kullanılarak hesaplanan temel bileşenleri açıklama oranına (% Varyans) göre göstermektedir. Buna göre 40 farklı özellikten oluşan tüm veri için birinci temel bileşen ile elde edilen açıklama oranı %35 olarak elde edilmiştir. Başka bir deyişle birinci bileşen verinin içerdiği toplam özelliklerin %35'ini gösterebilmektedir. Diğer temel bileşenleri durumu incelendiğinde de ikinci sıradaki bileşenin %21.51 ile açıklama gücüne sahip olduğu görülmüştür. Toplamda ise ilk iki bileşen ile %56.57 seviyesinde açıklama gücü elde edilmiş olur. Tüm verinin sahip olduğu anlam seviyesinin yarısını iki bileşen boyutunda görebilmek yapılan Temel Bileşen Analizinin önemli çıktılarından biridir. Devamında diğer bileşenlerin katkısı incelenirse, 6. Bileşende varyansın 1 değerine düştüğü görülmektedir. Bu düzeyde tüm verinin barındırdığı anlam altı temel bileşenin yardımıyla %88 oranında sağlanabilmektedir.

Tablo 3.14 Temel bileşenlerin açıklama yüzdeleri

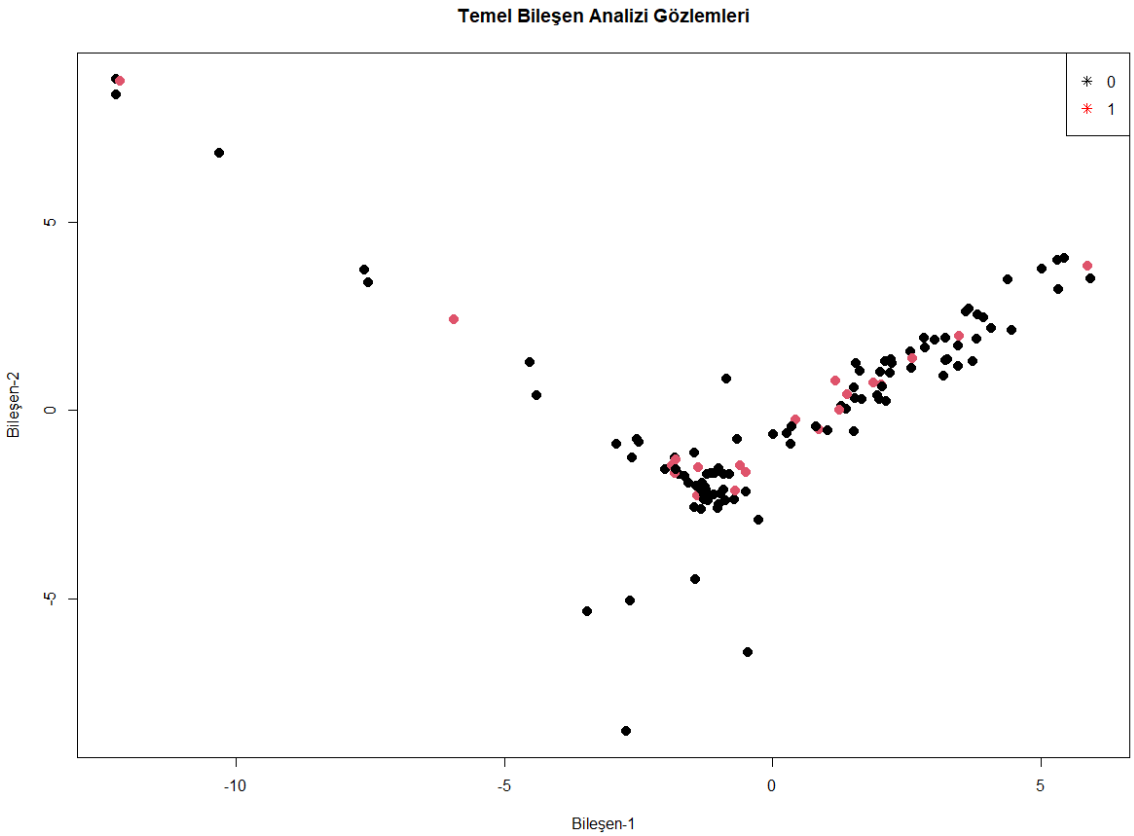
Temel Bileşen	TB-1	TB-2	TB-3	TB-4	TB-5	TB-6	TB-7	TB-8
Varyans	11.57	7.1	4.43	3.58	1.49	1	0.89	0.79
Açıklama Oranı (% Varyans)	35.06	21.51	13.43	10.83	4.52	3.03	2.7	2.38
Kümülatif Toplam	35.06	56.57	70.0	80.83	85.35	88.39	91.09	93.47

Aşağıda Şekil 3.13 ile bileşen eksenleri ile düzlem üzerinde gösterilmiştir. Şekil incelendiğinde yatayda 1. Bileşen ve dikeyde 2. Bileşen üzerinden özelliklerin etki düzeylerini görmek mümkündür. Şekil 3.14 ise örneklem alınan gözlem değerlerinin bileşenler düzleminde yayılımını göstermektedir. Gerçek verinin sahip olduğu eksenler haricinde temel bileşenlerin oluşturduğu bu iki boyutlu düzlemde, dolandırıcılık gözlemlerinin normal gözlemlerden ayırt etmek kabaca mümkün olmuştur. Ölçeklenmiş veri kullanıldığında birinci bileşenin negatif olduğu değerlerde dolandırıcılık gözleminin giderek seyrekleştiği görülebilir. Tek başına bu çıkarım bile %35 açıklama oranına sahip olduğu da bilinen bir durumdur.

Benzer şekilde ikinci bileşin de negatif değerinde dolandırıcılık gözlemine rastlamanın zor olduğu sonucuna varılabilir. Ancak ilk bileşenden farklı olarak, ikinci bileşenin pozitif yönde hareketleri ile dolandırıcılık gözlemleri arasındaki ilişki birinci bileşene göre daha zayıftır.



Şekil 3.13 Özelliklerin temel bileşenler eksenlerinde dağılımı



Şekil 3.14 Örnek gözlemlerin temel bileşen eksenlerinde dağılımı

Aşağıda Tablo 3.15 ise verideki gerçek özelliklerin bileşen eksenlerinde aldıkları değerler gösterilmiştir. En yüksek ifade gücüne sahip iki temel bileşende ortalama etki değerine göre sıralanan veri özellikleri incelendiğinde, genel olarak uygulama adımında üretilen yardımcı değişkenlerin önde olduğu ve İş Yeri Risk özelliklerinin başı çektiği görülmektedir. En yüksek etkiye sahip özellik 15 günlük periyotta İş Yeri Dolandırıcılık Riski olmuş, 30 ve 45 günlük periyotlar da yakından takip etmiştir. Ardından Harcama Kategorisinde Risk ve İş Yeri İşlem Sayısı özellikleri gelmektedir. İş Yeri Risk özelliklerinden farklı olarak Harcama Kategorisinde uzun periyotlara ait değerlerin kısa periyotlardan daha önde olduğu görülmüştür. Son sıralara bakıldığında ise, Müşteri özelliklerini taşıyan değişkenlerin İş Yeri ve Harcama Kategorisine kıyasla çok daha zayıf düzeyde etkileri olduğu görülmektedir. Bu nedenle toplamda %56.57 düzeyinde olan iki bileşenin açıklama becerisinde müşteri özelliklerinin katkısının yüksek olmadığı söylenebilir.

Tablo 3.15 Özelliklerin temel bileşenlere etkisi

Özellikler	Bileşen 1	Bileşen 2	Toplam	Ortalama
ISY_RISK_15	4.30	5.20	9.49	4.75
ISY_RISK_30	4.20	5.01	9.21	4.61
ISY_RISK_45	4.20	4.90	9.09	4.55
KTG_RISK_60	5.18	3.91	9.09	4.54
ISY_NB_45	4.87	4.20	9.06	4.53
KTG_RISK_45	5.25	3.81	9.06	4.53
ISY_NB_60	4.84	4.18	9.02	4.51
ISY_RISK_60	4.14	4.85	9.00	4.50
KTG_NB_15	5.11	3.86	8.98	4.49
KTG_RISK_30	5.16	3.81	8.97	4.48
ISY_NB_30	4.86	4.10	8.97	4.48
KTG_NB_30	5.06	3.86	8.92	4.46
ISY_RISK_7	3.93	4.89	8.82	4.41
KTG_NB_7	4.96	3.63	8.58	4.29
ISY_NB_15	4.62	3.85	8.46	4.23
KTG_NB_60	4.60	3.61	8.20	4.10
KTG_RISK_7	4.60	3.15	7.75	3.87
KTG_RISK_15	4.49	3.23	7.72	3.86
ISY_NB_7	4.09	3.16	7.25	3.63
DOVIZ	3.69	0.58	4.28	2.14
MUS_AVG_15	0.17	3.58	3.75	1.87
MUS_AVG_30	0.20	2.93	3.13	1.56
MUS_AVG_7	0.11	2.76	2.87	1.43
TUTAR	0.06	2.38	2.44	1.22
MUS_AVG_60	0.18	2.12	2.30	1.15
MUS_NB_30	0.66	0.10	0.76	0.38
MUS_NB_45	0.53	0.17	0.70	0.35
MUS_NB_15	0.32	0.27	0.59	0.29
MUS_NB_60	0.45	0.13	0.57	0.29
MUS_NB_7	0.08	0.31	0.39	0.20

Sonuç olarak Temel Bileşen Analizi ile veride yer alan tüm özellikler arasında sonradan üretilen yardımcı özelliklerin yüksek düzeyde açıklayıcılığa sahip olduğu görülmüştür. Açıklayıcılık olarak %56.57 seviyesindeki iki eksenli temel bileşen seçiminde İş Yeri ve Kategori bazlı özelliklerin baskın olarak yer aldığı tespit edilmiştir. Risk Puanı özellikleri işlem sayısı özelliklerinden daha etkilidir. Periyotlar açısından bakıldığında ise, beş periyodun arasında en az açıklama etkisine sahip olan 7 günlük periyottur. Bu da, kısa vadede ortaya çıkan müşteri harcama alışkanlıklarının ve risk puanlamalarının uzun vadeye göre dolandırıcılık tespitine daha az fayda sağladığını gösterir. Müşteri istatistiklerine bakıldığında ise, ortalama harcama miktarını gösteren istatistiklerin, harcama sayısı istatistiğine göre daha faydalı olduğunu görmek mümkündür.

4
SONUÇ VE ÖNERİLER

Çalışma boyunca yapılan makine öğrenmesi uygulamalarıyla dolandırıcılık tespiti amacını gerçekleştirmeye kayda değer katkı sağlanmıştır. Gerçek hayattan güncel olarak elde edilen örnek veri sayesinde tahmin sonuçlarının başarısı ve faydası ortaya konmuştur. Uygulamanın ilk adımlarından itibaren veri hazırlığında yapılan çalışmalar, topluluk öğrenimi adıyla irdelenen birleştirilmiş tahmin modelleri ve son olarak eşik değeri uygulaması ile anahtar göstergeler üzerinde iyileştirme elde edilmiştir. Böylece dolandırıcılık tespiti gibi dijital dünyada gittikçe öne çıkan sorunlardan biri olan bu zor problemde makine öğrenmesi yöntemlerinin katkısı öne çıkarılmıştır.

En iyi F_1 skoru olan %80 üzerinden literatürle kıyaslama yapıldığında, bu çalışmadakine yakın veri boyutu ve dolandırıcılık oranına (en fazla %1) sahip uygulamalardan çoğunlukla daha iyi sonuçlar elde edilmiştir [5]. Diğer yandan daha yüksek dolandırıcılık oranı tercih eden, başka bir deyişle görece daha kolay test senaryosuna sahip, çalışmalara ait en iyi bulgular mevcut çalışmayla ancak yakınlaşabilmiştir. Dolayısıyla bu çalışma kapsamında yardımcı özellik üretme, model birleştirme ve tahmin eşik değeri uygulamalarının performansa olumlu etkisi ortaya çıkarılmıştır.

Bu bölümün devamında, öne çıkan bulgular hem hepsi birlikte hem de farklı risk algıları ve model tipleri ayrımı yapılarak yorumlanmıştır. Dengeli risk yaklaşımı için F_1 skoru öne çıkarılmışken, risk yanlısı yaklaşımda Belirlilik, düşük riskli yaklaşımda ise Kesinlik öne çıkarılmıştır. Tekil modeller kendi arasında kıyaslandığı gibi gruplanmış modeller ile birlikte kıyaslamaları da bu bölümde gösterilmiştir. Tahmin eşik değeri sonucunda elde edilen değerler hem kendi aralarında kıyaslanmış hem de toplam uygulama içinde yerleri gösterilmiştir.

4.1 En İyi Tahmin Modeli

Sonuçlara bakılmadan önce şu çıkarımı belirtmek çok önemlidir, aşırı derecede gözlem dengesizliği olan dolandırıcılık tespiti probleminde bir en iyi tahmin modeli belirlemek hiçbir zaman tek boyutlu bir bakış açısı ile yapılamaz. Bunu yerine beklenen fayda ya da asıl amaç durumunu farklı yaklaşımlara ayırıp farklı göstergeler üzerinden yorumlamak gerekir. Diğer yandan, farklı amaçları eşit öncelikle ele almak da ayrı bir yaklaşım olabilir. Aşağıda önerilen dolandırıcılık tespiti seçiminde üç farklı yaklaşım vardır. Birincisi, tüm tahmin hatalarını eşit düzeyde gören ve F_1 skoru ile ölçülen dengeli yaklaşımdır. İkincisi gerçek dolandırıcılık gözlemlerini yakalama oranını maksimize etmeyi hedefleyen yüksek risk iştahlı yaklaşım, üçüncüsü ise yanlış dolandırıcılık tahminini minimize etmeyi hedefleyen düşük risk iştahlı yaklaşımdır. Böylelikle her risk yaklaşımı beraberinde farklı öncelikler getirir.

Bununla beraber uygulama kapsamında oluşturulan metodoloji ise farklı model türlerini ortaya çıkartmıştır. Geleneksel şekilde makine öğrenmesi algoritmalarının münferit olarak tahmin modeli kurmada kullanılmasının yanı sıra, bu çalışmada birden fazla münferit algoritmanın farklı seçim kriterleri ile seçilerek bir araya getirildiği ve farklı gruplama yöntemleri ile de birleştirildiği modellemeler de tahmin üretme amacıyla kullanılmıştır. Böylelikle en iyi dolandırıcılık tespiti tahmin modelinin belirlenmesi farklı risk yaklaşımı ve model çeşitleri ile birçok boyutta ele alınabilmiştir.

4.1.1 Dengeli Yaklaşım

Dengeli Yaklaşım, gerçek hayatta normal gündeki ortalama bir senaryo için en iyi seçenektir. Kredi kartı sahibi ile kartı veren finans kuruluşu arasında tam ortada yer alır. Kart sahipleri genelde bankalar tarafından mümkün olduğu kadar az rahatsız edilmek isterlerken, bankalar ise gerçekte çok düşük oranda gerçekleşen kart hırsızlığı olaylarını yönetirken bariz bir gösterge olmadığı müddetçe genellikle müşterinin kendi tedbirine bırakır. Diğer yandan, risk algısı dinamik bir olgudur. Hem banka tarafında hem de müşteri tarafında çeşitli gelişmeler sebebiyle zaman içerisinde değişkenlik gösterebilir. Bu sebeplerden dolayı karar

vericiler için risk algısı açısından bir uçtan diğer uca uzun vadede en kapsayıcı olan yaklaşım dengeli yaklaşımdır.

Aynı anda mümkün olduğu kadar yüksek oranda dolandırıcılık işlemini doğru tahmin ederken diğer yandan olabildiğince az sayıda normal işlemi dolandırıcılık olarak hatalı tahmin edebilmeye çalışmak Dengeli Yaklaşımın esas çıkış noktasıdır. Bu durum da aynı anda hem Belirlilik ve Kesinlik değerlerini yüksek tutmaya çalışmak, başka bir deyişle de F_1 skorunu arttırmak ile aynı anlama gelir. F_1 skoru ise doğası gereği Belirlilik ve Kesinlik ile beslenerek iki ayak üzerinde durması gereken bir göstergedir. Yalnızca Belirlilik ya da Kesinlik değerinin çok yüksek olması F_1 skoru için hiçbir zaman gerçekten yeterli olmaz. Tablo 4.1 Dengeli Yaklaşımda F_1 skoru baz alınarak seçilmiş en iyi performansı veren modelleri göstermektedir. Sıralama oluşturulurken her seçim kriterine göre tüm birleştirme yöntemleri temsil edilecek şekilde, kendi içinde eşik değerleri de hesaplandıktan sonra en yüksek performanslarıyla dahil edilmiştir. Bu da her model tipi için sadece en iyi sonuçların tabloda gösterildiği anlamına gelir.

Tablo 4.1 Dengeli yaklaşıma göre oluşan sıralama

Sıralama	Model	Tahmin Eşik Değeri	F_1 Skoru
1	Performans Seçim + Ağırlıklı Ortalama	0.6	%80
1	Performans Seçim + Çoğunluk Oylama	Varsayılan	%80
1	Hibrit Seçim + Çoğunluk Oylama	Varsayılan	%80
4	Performans Seçim + Yumuşak Oylama	0.6	%79
4	Performans Seçim + Uzlaşma	0.4	%79
6	Temel Seçim + Yumuşak Oylama	0.55	%78
6	Temel Seçim Ağırlıklı Ortalama	0.6	%78
6	Hibrit Seçim + Ağırlıklı Ortalama	0.55	%78
9	Rastgele Orman Tekil Model	Varsayılan	%77
9	Hibrit Seçim + Yumuşak Oylama	0.55	%77
11	Performans Seçim + Lojistik Regresyon	Varsayılan	%76
11	Temel Seçim + Çoğunluk Oylama	Varsayılan	%76
11	Temel Seçim + Uzlaşma	0.45	%76
14	Hibrit Seçim + Lojistik Regresyon	Varsayılan	%75
15	Temel Seçim + SVM	Varsayılan	%74
15	Hibrit Seçim + Uzlaşma	0.55	%74

Sıralama incelendiğinde, %80 F_1 skoru ile üç farklı birleştirme modelinin ilk sırayı paylaştığı görülür. Bu üç modelde seçim yöntemi olarak performans bazlı seçim ve hibrit seçim öne çıkmıştır. Birleştirme yöntemlerinde ise, ağırlıklı ortalama ve çoğunluk oylama en iyi sonuçları vermiştir. Eşik değerleri incelendiğinde, varsayılan (0.5) tahmin eşik değeri üç modelin ikisinde kullanılmış, 0.6 eşik değeri ise bir kez kullanılmıştır. İlk üç modeli takip eden %79 F_1 skoruna sahip iki model incelendiğinde ise yine performans bazlı seçimin kullanıldığı birleştirme modellerin en iyi sonuçları verdiği görülmektedir. Tekil modeller açısından bakıldığında, en iyi sonucun Rastgele Orman yönteminde olduğu görülmektedir. Diğer yandan en iyi tekil modelin birçok birleştirme modelinin arkasında kaldığı gözükmektedir. Dolayısıyla birleştirme modellerinin tekil modellerden daha üstün olduğu yorumu yapılabilir. Alt sıralarda dikkat çeken bir diğer konu ise, üst katman öğrenme yoluyla elde edilen birleştirme modellerinin tekil modellerden bile geride kalmış olmasıdır. En iyi üst katman modeli Performans Seçim ile oluşturulan Lojistik Regresyon birleştirme modelidir. Hibrit Seçim için de yine Lojistik Regresyon en iyi sonuç veren üst katman birleştirme yöntemi olurken. Temel Seçim için Destek Vektör Makineleri (SVM) en yüksek F_1 skorunu veren model olmuştur.

4.1.2 Yüksek Risk İştahlı Yaklaşım

Yüksek Risk İştahı, karar vericinin dolandırıcılık işlemlerini en iyi oranda yakalamayı istemesidir. Bankalar çeşitli nedenlerden dolayı risk iştahını arttırmaya yönelik politika izleyebilir. Kimi zaman ulusal veya uluslararası otoritelerden gelen güvenlik artırım talepleri, kimi zaman da piyasadaki hırsızlık haberleri risk iştahını arttırarak dolandırıcılık işlemleri üzerindeki baskıyı yükseltebilir. Diğer yandan Yüksek Risk İştahı kart sahibi müşteri için hem olumlu hem olumsuz etkileri olabilecek bir durumdur. Eğer gerçekten kart hırsızlık vakası söz konusu ise, müşteri bankanın ondan önce bu durumu fark etmesinden memnun olacaktır. Fakat eğer işlem müşteri teyidi sonucunda normal bir işlem çıkarsa işler tersine döner. Banka hem vakit harcayarak ek işgücü maliyeti yüklenmiş hem de müşteri açısından güvensizlik izlenimi oluşturmuştur. Daha da ötesi durum sadece bu kadarıyla sınırlı da değildir, yüksek risk iştahını yönetmek

için teyit ve kontrol kapasitesinin yeterli olması gerekir. Örneğin izlenen model günde yüzlerce tespit üretmişse, her birini teyit ederek ortaya çıkarmak sürdürülebilir bir durum olmaktan uzaktır.

Dolandırıcılık gözlemlerinin ne kadarının yakalandığı anlamına gelen Yüksek Risk İştahı Yaklaşımı, doğrudan Belirlilik göstergesi ile kontrol edilebilir. Dengeli Yaklaşımın aksine aynı anda iki farklı göstergeyi iyileştirmeyi önemsemez. Fakat benzer Belirlilik seviyelerinde en yüksek F_1 skorunu tercih eder. Başka bir deyişle, en iyi modeli belirlerken salt Belirlilik odaklı olmak ana gösterge olur sonrasında F_1 skoru ile de ikincil öncelik seviyesinde son karar verilir. Tablo 4.2 Yüksek Risk İştahlı Yaklaşımda Belirlilik değer baz alınarak seçilmiş en iyi performansı veren modelleri göstermektedir. Sıralama oluşturulurken her seçim kriterine göre tüm birleştirme yöntemleri temsil edilecek şekilde, kendi içinde eşik değerleri de hesaplandıktan sonra en yüksek Belirlilik performanslarıyla dahil edilmiştir. Aynı Belirlilik düzeyine sahip olan modeller F_1 skoru ile birbirinden ayrıştırılmıştır.

Tablo 4.2 Yüksek risk iştahı yaklaşımında oluşan sıralama

Sıralama	Model	Tahmin Eşik Değeri	F_1 Skoru	Belirlilik
1	Temel Seçim + İyimser	Varsayılan	%60	%94
2	Performans Seçim + İyimser	Varsayılan	%68	%92
3	Performans Seçim + YSA	Varsayılan	%65	%89
3	Hibrit Seçim + Adaboost	Varsayılan	%63	%89
5	Performans Seçim + Ağırlıklı Ortalama	0.4	%76	%88
5	Performans Seçim + Yumuşak Oylama	0.4	%74	%88
7	Temel Seçim + Yumuşak Oylama	0.4	%75	%87
8	Temel Seçin + Ağırlıklı Ortalama	0.4	%75	%86
8	Hibrit Seçim + İyimser	Varsayılan	%70	%86
10	Adaboost Tekil Model	Varsayılan	%72	%85
10	Hibrit Seçim + Uzlaşma	0.4	%70	%85
12	Hibrit Seçim + Ağırlıklı Ortalama	0.4	%75	%84
12	Temel Seçim + Lojistik Regresyon	Varsayılan	%72	%84
14	Hibrit Seçim + Yumuşak Oylama	0.4	%74	%83
15	Performans Seçim + Uzlaşma	0.4	%79	%82
16	Rastgele Orman Tekil Model	Varsayılan	%77	%76
17	Temel Seçim + Uzlaşma	0.4	%76	%75

Sıralama incelendiğinde, %94 düzeyinde yüksek Belirlilik değeri ile İyimser birleştirme yöntemi Temel Seçim yoluyla ilk sırayı almıştır. F_1 skoru açısından daha iyi olan ancak %92 Belirlilik düzeyine sahip olan Performans Seçim ikinci sıradadır. İlk iki modelin de ortak noktası İyimserlik birleştirme yöntemi olarak öne çıkmıştır. Uzlaşma, Ortalama ve Oylama birleştirme yöntemleri İyimserlik yöntemine göre F_1 skoru açısından daha güçlü olsa da, Belirlilik kriterinde arkada kalmıştır. Tahmin eşik değerleri incelendiğinde varsayılan (0.5) eşik değerinin çoğunlukla 0.4 eşik değerine göre daha üst sıralarda yer aldığı görülmüştür. Farklı eşik değerlerinin yerine Dengeli Yaklaşım ile ortaya çıkan tablo ile kıyaslandığında Temel Seçim Yönteminin, bu kez diğer seçim yöntemlerinden daha zayıf olmadığı tespit edilmiştir. Dengeli Yaklaşım ile oluşan diğer bir fark ise üst katman öğrenme yoluyla elde edilen birleştirme modellerinin üst sıralarda olmasıdır. Performans Seçim ile Yapay Sinir Ağları, Hibrit Seçim ile Adaboost üst katman öğrenme modelleri ilk üç sırada kendilerine yer bulmuştur. Diğer yandan Dengeli Yaklaşıma benzer olarak, tekil modellerin çoğunlukla birleştirme modellere geçildiği görülmüştür. Dolayısıyla Yüksek Risk İştahlı Yaklaşım için de birleştirme modellerin tekil modellerden daha üstün olduğu yorumu yapılabilir.

4.1.3 Düşük Risk İştahlı Yaklaşım

Düşük Risk İştahı, karar vericinin yaptığı dolandırıcılık tahminlerinde en az sayıda normal işlemi hatalı yakalamak istemesidir. Başka bir deyişle müşteriye ilettiği tüm dolandırıcılık bildirimlerinde işlemlerin gerçekten de dolandırıcılık olduğunu teyit edebilmesidir. Gerçek hayatta bu durumun istendiği gibi olması problemin gözlem dengesizliği sebebiyle modelde yapılan dolandırıcılık tahmin sayısı ile ters orantılıdır. Bankalar tıpkı Yüksek İştahlı Yaklaşımda olduğu gibi, düşük risk iştahını da çeşitli sebeplerden dolayı öne çıkarabilirler. Örneğin mevcutta piyasadaki müşteriler arasında bir güvensizlik ortamı oluştuysa, banka bu ortamı daha da güvensizleştirmemek için hatalı tahminlerini müşteriye bildirmekten kaçınır. Başka bir örnek olarak ise büyüyerek yeni müşteriler kazanmak isteyen bankalar verilebilir. Diğer yandan dolandırıcılık işlemlerinin aşırı seyrek gözlemlenmesi ve tespit edilmesinin zor olması da Düşük Risk İştahı için kabul

edilebilir bir bahane olabilir. Yüksek Risk İştahının aksine, daha az teyit gerektirdiğinden yönetilmesi daha kolaydır.

Dolandırıcılık tahminlerinin ne kadar isabetli olduğu anlamına gelen Düşük Risk İştahı Yaklaşımı, Kesinlik değeri ile ölçülmektedir. Dengeli yaklaşımdan farklı olarak aynı anda iki farklı gösterge öne çıkarılmaz. Ancak benzer Kesinlik değerine ulaşan modeller varsa, F_1 skoruna başvurur. Tablo 4.3 Düşük Risk İştahlı Yaklaşımda Belirlilik değer baz alınarak seçilmiş en iyi performansı veren modelleri göstermektedir. Sıralama oluşturulurken her seçim kriterine göre tüm birleştirme yöntemleri temsil edilecek şekilde, kendi içinde eşik değerleri de hesaplandıktan sonra en yüksek Kesinlik performanslarıyla dahil edilmiştir. Aynı Kesinlik düzeyine sahip olan modeller F_1 skoru ile birbirinden ayrıştırılmıştır.

Tablo 4.3 Düşük risk iştahlı yaklaşımda oluşan sıralama

Sıralama	Model	Tahmin Eşik Değeri	F_1 Skoru	Kesinlik
1	Temel Seçim + Kötümser	Varsayılan	%63	%86
2	Hibrit Seçim + Kötümser	Varsayılan	%69	%85
2	Temel Seçim + Uzlaşma	0.6	%67	%85
4	Performans Seçim + Kötümser	Varsayılan	%72	%84
5	Performans Seçim + Uzlaşma	0.6	%74	%83
5	Rastgele Orman Tekil Model	0.6	%70	%83
7	Temel Seçim Ağırlıklı Ortalama	0.6	%78	%81
7	Hibrit Seçim + Ağırlıklı Ortalama	0.6	%78	%81
9	Performans Seçin + Ağırlıklı Ortalama	0.6	%80	%80
9	Temel Seçim + Yumuşak Oylama	0.6	%77	%80
11	Performans Seçim + Yumuşak Oylama	0.6	%79	%79
11	Hibrit Seçim + Yumuşak Oylama	0.6	%77	%79
13	Rastgele Orman Tekil Model	Varsayılan	%77	%78
14	Hibrit Seçim + Uzlaşma	0.6	%74	%68
15	Performans Seçim + Lojistik Regresyon	Varsayılan	%74	%66
15	Temel Seçim + SVM	Varsayılan	%76	%66
15	Hibrit Seçim + SVM	Varsayılan	%74	%66

Sıralama incelendiğinde, 86 Kesinlik seviyesi ile Temel Seçim ve Kötümser birleştirme yöntemi ilk sırada yer almıştır. Ardından yakın Kesinlik değeri ile ve daha yüksek F_1 skorlarıyla Hibrit Seçim ile Kötümser birleştirme yöntemi ve Temel

Seçim ile Uzlaşma birleştirme yöntemi gelmektedir. Yüksek Risk İştahından farklı olarak Kötümserlik düşük risk iştahına daha uygundur. Önceki sonuçlara benzer olarak Uzlaşma, Ortalama ve Oylama birleştirme yöntemleri Kötümserlik yöntemine göre F_1 skoru açısından daha güçlü olsa da, Kesinlik kriterinde arkada kalmıştır. Dengeli yaklaşım ile benzer Yüksek Risk İştahı ile farklı olarak üst katman öğrenme yoluyla elde edilen birleştirme modellerinin üst sıralarda yer bulamamıştır. Bir diğer önemli bulgu, önceki yaklaşımların her ikisinden de farklı olarak, tekil öğrenme modeli Rastgele Orman için 0.6 tahmin eşik değerine sahip modelin Kesinlik sıralamasında üst sıralarda bulunmasıdır. Dolayısıyla Düşük Risk İştahlı Yaklaşım için birleştirme modellerin tekil modellerden her zaman daha üstün olduğu yorumu yapılamaz.

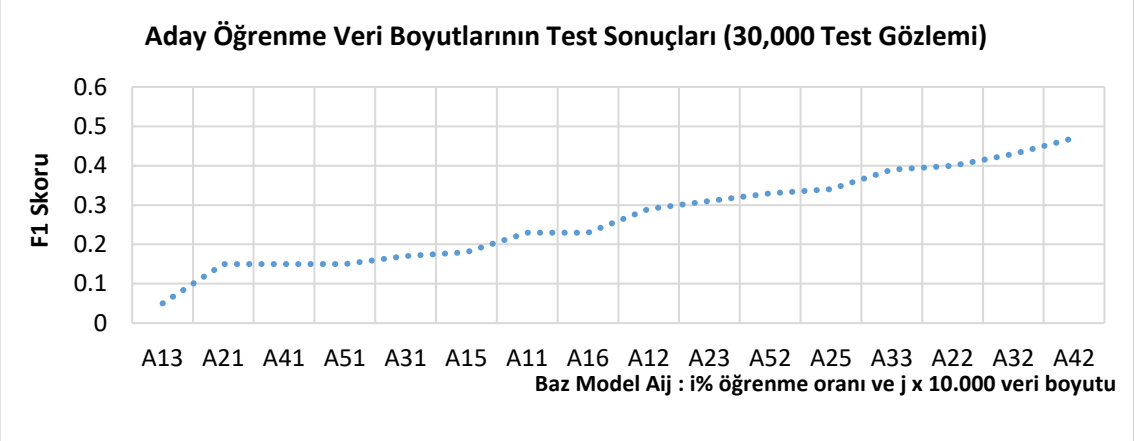
4.2 Veri Özelliklerinin Yorumlanması

Aşırı derecede dengesiz gözlem dağılımıyla öne çıkan kredi kartı işlem verisi uygulama çalışması boyunca birçok farklı işlemden geçirilmiştir. Dolandırıcılık tespiti için kullanılan makine öğrenmesi modelleri, tahmin başarısını iyileştirmek adına bağımlı değişken ile bağımsız değişkenler arasında ilişkilendirme kurmaya çalışır. Diğer konulardan farklı olarak dolandırıcılık gözlemini ayırt etme konusunun birden fazla yolu olabilir. Müşterilerin harcama alışkanlıkları ile hırsızların yaptığı harcamalar arasındaki farklılığın oluşturduğu ayrımdan yararlanılabileceği gibi, iş yerleri ve harcama kategorilerinin periyodik işlem istatistikleri de hırsızların izlediği dolandırıcılık yöntemi hakkında bilgi verebilmektedir [17,60]. Bu bağlamda yapılan veri hazırlığı uygulamalarının başında yardımcı özelliklerin üretilmesi sayılabilir. Öğrenme ve test veri boyutlarının belirlenmesi de ardından gelen bir başka önemli uygulamadır. Sonuçta uygulama verisine yeni eklenen yardımcı özellikler hakkında şu sonuçlar elde edilmiştir.

Birincisi, tüm yardımcı özellikler öğrenme ve test başarısını olumlu yönde etkilemiştir. Kendi aralarında sıralamak gerekirse, Kategori ve İş Yeri Risk puanı özellikleri en yüksek katkıya sahip özellikler olmuştur. Periyotlar incelendiğinde daha uzun periyotlarda üretilen yardımcı değişkenlerin daha kısa periyotlardan elde edilenlere göre daha anlamlı oldukları görülmüştür. İkinci olarak, Müşteri

ortalama işlem tutarı özellikleri diğer yardımcı özellikler arasında anlamlılık açısından son sırada yer almıştır. Bu durum işlem tutarlarının doğrudan müşteri harcama alışkanlıklarını yansıtamadığını, daha doğrusu hırsızların işlemlerinden ayrışamadığı gösterir. Diğer yandan genel çapta hızlı değişen ürün/hizmet fiyatları da, ortalama işlem tutarının değişken olmasına sebep olmuş olabilir. Üçüncü tespit, temel özelliklerden biri olan işlemin döviz türünün sağladığı fayda hakkındadır. Diğer temel özelliklerden farklı olarak işlemin döviz türü sadece temel özelliklere kıyasla değil, müşteri istatistiklerine göre de daha yüksek oranda katkı üretebilmeyi başarmıştır. Buna göre işlemin yabancı para cinsinden olup olmadığı dolandırıcılık işlemi olma tahmin sonucunu anlamlı seviyede etkileyebilmektedir.

Son olarak, öğrenme verisi ve test verisi boyutları üzerinde yapılan araştırma sonucu incelendiğinde salt öğrenme oranı (dolandırıcılık oranı) ile performans arasında doğrusal bir ilişki olmadığı görülmüştür. Şekil 4.1 ile görüldüğü gibi en iyi eşleşme baz model kullanılarak enterpolasyonlar yardımıyla %4 öğrenme oranının bulunduğu 20.000 gözlemden oluşan veri setinde elde edilmiştir. Diğer yandan en iyi üç sonucun da aynı gözlem boyutunda olması gözlem boyutları ile test boyutları eşleşmesi ele alındığında 30.000 gözleme sahip test verisi için 20.000 gözleme sahip öğrenme verisinin daha uygun olacağını göstermektedir.



Şekil 4.1 Öğrenme veri boyutlarının test sonuçları

Özetle, öğrenme ve test veri boyutlarında yapılan en iyi eşleşme araştırması sonucunda F_1 skoru ortalama %30 düzeylerinden %47 düzeyine yükseltilmiş, ardından üretilen yardımcı değişkenlerle ise tahmin performansının baz

modeldeki %47 F_1 skoru seviyesinden %70 lere kadar yükselebildiği ortaya çıkartılmıştır. Makine öğrenmesi modelleri üzerinde yapılan çeşitli uygulamalar sonucunda nihai en iyi F_1 skoru %80 düzeyinde ölçülmüştür. Bu sonuçlar, sağlanan fayda üzerinden bakıldığında veri özellikleri üzerinde yapılan çalışmaların en az model uygulamaları kadar önemli olduğunu göstermiştir.

4.3 Öneriler

Uygulama çalışmasını son bölümünde, gelecek çalışmalarda dikkate alınabilecek noktaların önerilerine yer verilmiştir. En başta şu husus unutulmamalıdır ki, gerçekte binlerce işlem arasından elle sayılabilecek kadar az sayıda gözlemde rastlanan dolandırıcılık işlemlerinin, önceden bilinmesi her zaman beklenenden daha zor olacaktır. Bu sebeple dolandırıcılık tespiti problemini farklı kılan nedenlerin kaynağına inmeye çalışmanın her zaman birinci öncelik olmasında yarar vardır. Bu çalışmada ortaya çıkan sonuçlar ışığında, uygulama verisinden üretilen yardımcı özelliklerin tahmin performansını iyileştirmede kritik öneme sahip olduğu görüldüğünden farklı özellik elde etme yöntemlerinin sonraki çalışmalarda ele alınmasının faydalı olabileceği öngörülmektedir.

Ek olarak, üretilen değişkenlerde uzun periyotların daha faydalı olduğu sonucundan yola çıkarak toplam veri içinde örneğin bir yıldan uzun süreyi kapsayan daha uzun tarihe yayılmış gözlemlerin bulunmasının anlam düzeyine etkisi ölçülebilir. Çalışma boyunca takip edilen risk odaklı performans değerlendirme yaklaşımları ve farklı model birleştirme yöntemleri çeşitlendirilebilir. Ayrıca makine öğrenmesi modelleri açısından bakıldığında ileri seviye öğrenme yolları arasında gösterilen Yapay Sinir Ağları algoritmalarının daha kapsamlı modellerinde tahmin performansları ölçülebilir.

Son olarak farklı ulusal bankalardan elde edilecek veri örnekleri ile tekrarlanabilecek sonraki çalışmalar vasıtasıyla bu çalışmanın yerel ölçekte genel geçerliliği kontrol edilebilir, doğruları ve eksikleri yorumlanabilir. Problemin dinamik olması durumunda dolayı, bu alandaki çalışmaların her zaman güncel veriler ve yöntemler ile yenilenerek devam ettirilmesi elbette daha faydalı olacaktır.

KAYNAKÇA

[1] "Raporlar ve Yayınlar", [Online]. Available: https://bkm.com.tr/raporlar-ve-yayinlar/donemsel-bilgiler/. [Accessed: 01-May-2022].

[2] "Credit Card Fraud 2021 Annual Report: Prevalence, Awareness, and Prevention", [Online]. Available: https://www.security.org/digital-safety/credit-card-fraud-report/. [Accessed: 01-May-2022].

[3] A. Abdallah, M. A. Maarof, A. Zainal, "Fraud detection system: a survey", *Journal of Network and Computer Applications*, 68, pp. 90–113, 2016.

[4] V. Dheepa, R. Dhanapal, "Behavior Based Credit Card Fraud Detection Using Support Vector Machines", *ICTACT Journal On Soft Computing*, 02 (04), pp. 391-397, 2012.

[5] N. F. Ryman-Tubb, P. Krause, and W. Garn, "How Artificial Intelligence and Machine Learning Research Impacts Payment Card Fraud Detection: A survey and industry benchmark," *Engineering Applications of Artificial Intelligence*, 76, pp. 130–157, 2018.

[6] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, B. Baesens, "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions," *Decision Support Systems*, 75, pp. 38–48, 2015.

[7] R. C. Cavalcante, R. C. Brasileiro, V. L. Souza, A. L. Oliveira, "Computational Intelligence and Financial Markets: A Survey and Future Directions", *Expert Systems With Applications*, 55, pp. 194–211, 2016.

[8] K. G. Al-Hashedi, P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019", *Computer Science Review*, 40, 100402, 2021.

[9] X. Wei, L. Yuan, "An Optimized SVM Model for Detection of Fraudulent Online Credit Card Transactions, "2012 *International Conference on Management of e-Commerce and e-Government*, 20-21 Oct. 2012.

[10] A. Srivastava, M. Yadav, S. Basu, S. Salunkhe, M. Shabad, "Credit card fraud detection at merchant side using neural networks", *2016 3rd International Conference on Computing for Sustainable Global Development*, 16-18 March 2016.

[11] A. Zakaryazad, E. Duman, "A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing", *Neurocomputing*, 175, pp. 121–131, 2016.

[12] Y. Kültür, M. U. Çağlayan, "Hybrid approaches for detecting credit card fraud", *Expert Systems*, 34 (2), e12191, 2016.

[13] E. Kim, J. Lee, H. Shin, H. Yang, S. Cho, S. Nam, Y. Song, J. Yoon, J. Kim, "Champion-challenger analysis for credit card fraud detection: Hybrid

ensemble and deep learning", *Expert Systems With Applications*, 128, pp. 214–224, 2019.

[14] V. N. Dornadula, S. Geetha, "Credit Card Fraud Detection using Machine Learning Algorithms", *Procedia Computer Science*, 165, pp. 631–641, 2019.

[15] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga, N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning", *9th International Conference on Cloud Computing, Data Science & Engineering (Confluence),* 10-11 Jan. 2019.

[16] S. Khatri, A. Arora, A. P. Agrawal, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison", *10th International Conference on Cloud Computing, Data Science & Engineering (Confluence),* 29-31 Jan. 2020.

[17] A. C. Bahnsen, D. Aouada, A. Stojanovic, B. Ottersten, "Feature engineering strategies for credit card fraud detection" *Expert Systems With Applications*, 51, pp. 134–142, 2016.

[18] J. O. Awoyemi, A. O. Adetunmbi, S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis", *International Conference on Computing Networking and Informatics,* 29-31 Oct. 2017.

[19] J. Forough, S. Momtazi, "Ensemble of deep sequential models for credit card fraud detection", *Applied Soft Computing Journal*, 99, 106883, 2021.

[20] A. Mishra, C. Ghorpade, "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques" *IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS),* 24-25 Feb. 2018.

[21] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting", *IEEE Access*, 6, pp. 14277–14284, 2018.

[22] Ö. Güven, "Makine Öğrenmesi Teknikleriyle Mobil Ödemede Sahtekarlık Tespiti", *Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi*, *Sosyal Bilimler Enstitüsü*, *İzmir*, 2021.

[23] S. Bagga, A. Goyal, N. Gupta, A. Goyal, "Credit Card Fraud Detection using Pipeling and Ensemble Learning", *Procedia Computer Science*, 173, pp. 104–112, 2020.

[24] M. Zareapoor, P. Shamsolmoali, "Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier", *Procedia Computer Science*, 48, pp. 679–685, 2015.

[25] A. G. C. de Sa, A. C. M. Pereira, G. L. Pappa, "A customized classification algorithm for credit card fraud detection", *Engineering Applications of Artificial Intelligence*, 72, pp. 21–29, 2018.

[26] N. Carneiro, G. Figueira, M. Costa, "A data mining based system for credit-card fraud detection in e-tail", *Decision Support Systems*, 95, pp. 91–101, 2017.

[27] M. Hejazi, Y. P. Singh, "Credit Data Fraud Detection using Kernel Methods with Support Vector Machine", *Journal of Computer Science and Technology*, 2, pp. 35–49, 2012.

[28] X. Niu, L. Wang, X. Yang, "A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised", *arXiv*, 1904.10604v1, 2019.

[29] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature", *Decision Support Systems*, 50, pp. 559–569, 2011.

[30] Y. Sahin, E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines", *Proceedings of the International Multi conference of Engineers and Computer Scientist*, 16-18 March 2011.

[31] E. Aleskerov, B. Freisleben, B. Rao, "CARDWATCH: a neural network based database mining system for credit card fraud detection", *Proceedings of the Computational Intelligence for Financial Engineering,* 24-25 March 1997.

[32] F. Ghobadi, M. Rohani, "Cost Sensitive Modeling of Credit Card Fraud Using Neural Network Strategy", *International Conference of Signal Processing and Intelligent Systems,* 14-15 Dec. 2016.

[33] N. F. Ryman-Tubb, "Understanding Payment Card Fraud Through Knowledge Extraction From Neural Networks Using Large-Scale Datasets", *Doktora Tezi, University of Surrey, Department of Computer Science,* Surrey, 2016.

[34] S. Patil, V. Nemade, P. K. Soni, "Predictive Modelling For Credit Card Fraud Detection Using Data Analytics", *Procedia Computer Science*, 132, pp. 385–395, 2018.

[35] M. Kamboj, S. Gupta, "Credit Card Fraud Detection and False Alarms Reduction using Support Vector Machines", *International Journal of Advance Research, Ideas and Innovations in Technology*, 2 (4), 2016.

[36] I. Rajak, K. J. Mathai, "Intelligent Fraudulent Detection System Based SVM And Optimized By Danger Theory", *IEEE International Conference on Computer, Communication and Control*, 10-12 Sept. 2015.

[37] B. Baesens, V. V. Vlasselaer, W. Verbeke, "Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection", *Wiley and SAS Business Series*, 2015.

[38] M. Shelke, D. P. R. Deshmukh, V. Shandilya, "A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique", *International Journal of Recent Trends in Engineering and Research*, 3 (4), pp. 444-449, 2017.

[39] N. V. Chawla, K. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, 16 (1), pp. 321-357, 2002.

[40] D. M. Allen, "The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction", *Technometrics*, 16 (1), pp. 125-127, 2012.

[41] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *International Joint Conferences on Artificial Intelligence*, pp. 1137-1143, 1995.

[42] A. M. Molinaro, R. Simon, R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods", *Bioinformatics*, 21 (15), pp. 3301-307, 2005.

[43] G. J. McLachlan, K. A. Do, C. Ambroise, "Analyzing Microarray Gene Expression Data", *Wiley*, 2005.

[44] "How to Improve Class Imbalance using Class Weights in Machine Learning", [Online].Available:https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/. [Accessed: 01-May-2022].

[45] P. Mills, "Efficient statistical classification of satellite measurements", *International Journal of Remote Sensing*, 32, 2011.

[46] B. Everitt, "The Cambridge dictionary of statistics", *Cambridge University Press*, 1998.

[47] A. V. Joshi, "Machine Learning and Artificial Intelligence", *Springer*, 2020.

[48] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", *The American Statistician*, 46 (3), pp. 175-185, 1992.

[49] L. Rokach, O. Maimon, "Data Mining with Decision Trees", *World Scientific*, 2014.

[50] J. R. Quinlan, "Induction of decision trees", *Machine Learning*, 1, pp. 81-106, 1986.

[51] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", *Springer,* 2009.

[52] R. Zhu, D. Zeng, M. R. Kosorok, "Reinforcement Learning Trees", *Journal of the American Statistical Association*, 110, pp. 1770-1784, 2015.

[53] B. E. Boser, I. M. Guyon, V. N. Vapnik, "A training algorithm for optimal margin classifiers", *Proceedings of the fifth annual workshop on Computational learning theory*, July 1992.

[54]"Sinir ağı nedir?" [Online]. Available: https://aws.amazon.com/tr/what-is/neural-network/. [Accessed: 01-May-2022].

[55] Y. Sasaki, *The truth of the F-measure*, 2007.

[56] Y. Le Borgne, W. Siblini, B. Lebichot, G. Bontempi, "Reproducible Machine Learning for Credit Card Fraud detection - Practical handbook", *Université Libre de Bruxelles, Belgium,* 2022.

[57] W. Dong, W. Quan-yu, Z. Shou-yi, L. Feng-xia, W. Da-zhen, "A feature extraction method for fraud detection in mobile communication networks", *Fifth World Congress on Intelligent Control and Automation*, 15-19 June 2004.

[58] H. Cao, R. Zhang, "Using PCA to Improve the Detection of Medical Insurance Fraud in SOFM Neural Networks", *Proceedings of the 2019 3rd International Conference on Management Engineering*, Jan. 2019.

[59] A. D. Pawar, P. N. Kalavadekar, S. N. Tambe, "A Survey on Outlier Detection Technique s for Credit Card Fraud Detection", *Journal of Computer Engineering*, 16(2), pp. 44–48, 2014.

[60] C. Jiang, J. Song, G. Liu, L. Zheng, W. Luan, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism", *IEEE Internet of Things Journal*, 5(5), pp. 3637–3647, 2018.

[61] T. Kavzaoglu, I. Colkesen, "Karar Ağaçları İle Uydu Görüntülerinin Sınıflandırılması: Kocaeli Örneği", *Electronic Journal of Map Technologies*, 2 (1), pp. 36–45, 2010.

[62]"SVM margin" [Online]. Available: https://commons.wikimedia.org/-wiki/File:SVM_margin.png [Accessed: 01-May-2022].

[63]"Artificial neural network" [Online]. Available: https://en.wikipedia.org/-wiki/Artificial_neural_network#/media/File:Colored_neural_network.svg [Accessed: 01-May-2022].

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. S. Soner Kara, B. Akyüz, "Topluluk Öğrenimini ile Kredi Kartı Dolandırıcılık Tespiti: Karşılaştırmalı Bir Analiz", *2nd International Multidısıplinary Digital Live Congress,* 20-22 Mayıs 2022.