

T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
EKONOMETRİ PROGRAMI
YÜKSEK LİSANS TEZİ

SINIF DENGELEME YÖNTEMLERİNİN
MAKİNE ÖĞRENMESİ TEKNİKLERİ ÜZERİNE
ETKİSİ: KREDİ RİSK ÖRNEĞİ

Migraç Enes Furkan MİLLİ

Danışman
Prof. Dr. İpek DEVECİ KOCAKOÇ

2022

TEZ ONAY SAYFASI



ÖZET

Yüksek Lisans Tezi

Sınıf Dengeleme Yöntemlerinin Makine Öğrenmesi Teknikleri

Üzerine Etkisi: Kredi Risk Örneği

Migraç Enes Furkan MİLLİ

Dokuz Eylül Üniversitesi

Sosyal Bilimler Enstitüsü

Ekonometri Anabilim Dalı

Ekonometri Programı

Kredi riski, bankalar tarafından geri ödenmek üzere müşterilere verilen fonların belirlenen vade sonunda yükümlülüklerin yerine getirilmemesi sonucunda oluşmaktadır. Bu durum, kredi müşterisinin temerrüde düşmesine ve itibarını olumsuz etkilemesine sebep olmaktadır. Gelişen teknoloji ile birlikte farklı alanlarda kullanılan makine öğrenmesi yöntemleri, bankaların kredilendirme sürecine uyarlanan sistem ile riskli müşterilerin tespit edilmesini kolaylaştırmıştır. Böylece, kredi talebinde bulunan müşterilerin risk durumu, hızlı ve daha doğru bir şekilde belirlenebilmesi söz konusu olmuştur. Söz konusu riskli müşterilerin belirlenmesi hem kredi riskini azaltması hem de bankaların finansal faaliyetlerine devam etmesi açısından önem arz etmektedir.

Bu çalışmada, bankaların kredibilite açısından en uygun değerlendirmeyi yapabilmeleri için dengeli olmayan veri kümelerinde karşılaşılan sınıf dengesizliği sorununa çözüm bulmak için farklı yeniden örnekleme yöntemleri ile veri kümeleri dengeli hâle getirilerek makine öğrenmesi üzerindeki etkilerinin araştırılması amaçlanmıştır. Çalışmanın uygulama aşamasında, açık kaynaklı site üzerinden alınan Alman, Avustralya ve HMEQ gerçek hayat problemlerine ait kredi veri setleri kullanılmıştır. Çalışmada, riskli müşterilerin saptanması için K-En Yakın komşu, Naive Bayes, Lojistik Regresyon, Destek Vektör Makineleri, Çok Katmanlı Algılayıcı, Karar Ağaçları, Rassal Ormanlar, Gradyan Artırma Karar Ağaçları, Extra Ağaçlar, Sert ve Yumuşak Oylama gibi farklı makine

öğrenmesi sınıflandırma teknikleri kullanılmıştır. Sınıf dengesizliği problemi; **RUS, ROS, SMOTE-ENN, SMOTE-Tomek Links ve Balanced Bagging Classifier** gibi yeniden örnekleme ve hibrit yöntemleri ile dengeli hâle getirilmiştir. Bu doğrultuda, üç farklı kredi veri kümesi üzerinde dört farklı senaryo ile inceleme yapılmıştır. Sonuç olarak, sınıf dengeleme problemi için aşağı ve yukarı örnekleme yöntemlerinin bir arada kullanıldığı hibrit yöntemin makine öğrenme tekniklerinin en iyi sınıflandırma performansına ulaşmasında yardımcı olabileceği gözlemlenmiştir. Bu yöntem, bankaların riskli müşterileri saptanmasında büyük avantaj sağlayarak kredi riskini azaltmasına yardımcı olabilecektir.

Anahtar Kelimeler: Kredi Riski, Makine Öğrenmesi, Topluluk Öğrenmesi, Sınıflandırma Algoritmaları, Yeniden Örnekleme, Sınıf Dengeleme.

ABSTRACT

Master's Thesis

The Effect Of Class Balancing Methods On Machine Learning Techniques:

Example of Credit Risk

Migraç Enes Furkan MİLLİ

Dokuz Eylul University

Graduate School of Social Sciences

Department of Econometrics

Econometrics Program

Credit risk arises as a result of defaulting on liabilities at the end of the specified maturity of funds ceded to customers to be reimbursed by banks. This situation causes the credit customer to default and negatively affects their reputation. With technological development, machine-learning methods used in different areas facilitated the identification of risky customers with the system adapted to the lending process of banks. Thus, the risk situation of customers applying for loans can be determined quickly and more accurately. The Identification of these risky customers is important in terms of both reducing the credit risk and maintaining the financial activities of the banks. In this study, it is aimed to investigate the effects on machine learning by balancing the datasets with different resampling methods in order to find a solution to the class imbalance problem encountered in unbalanced datasets so that banks can make the most appropriate assessment in terms of credibility. In the implementation phase of the study, German, Australian and HMEQ real-life problem credit datasets obtained from an open source website were used. In the study, different machine learning classification methods such as K-Nearest Neighbor, Naive Bayes, Logistic Regression, Support Vector Machines, Multilayer Perceptron, Decision Trees, Random Forests, Gradient Boosting Decision Trees, Extra Trees, Voting Hard and Soft were used to detect risky customers. The class imbalance

problem has been balanced with resampling and hybrid methods such as RUS, ROS, SMOTE-ENN, SMOTE-Tomek and balanced bagging Classifier. In this direction, four different scenarios were examined in three different sets of credit data. As a result, it was observed that the hybrid method, in which over-and-under sampling methods are used together for the class balancing problem, can help machine learning techniques to achieve the best classification performance. This method will provide a great advantage in identifying risky customers and will help banks to reduce their credit risk.

Keywords: Credit Risk, Machine Learning, Ensemble Learning, Classification Algorithms, Resampling, Class Balancing.

SINIF DENGELEME YÖNTEMLERİNİN MAKİNE ÖĞRENMESİ TEKNİKLERİ ÜZERİNE ETKİSİ: KREDİ RİSK ÖRNEĞİ

İÇİNDEKİLER

TEZ ONAY SAYFASI	ii
YEMİN METNİ	iii
ÖZET	iv
ABSTRACT	vi
İÇİNDEKİLER	viii
KISALTMALAR	xii
TABLolar LİSTESİ	xiii
ŞEKİLLER LİSTESİ	xiv
GİRİŞ	1

BİRİNCİ BÖLÜM

KREDİ VE BANKACILIKTA RİSK KAVRAMINA GENEL BAKIŞ

1.1. KREDİ KAVRAMI	5
1.1.1. Kredi Unsurları	6
1.1.1.1. Zaman Unsuru	6
1.1.1.2. Güven Unsuru	6
1.1.1.3. Risk Unsuru	7
1.1.1.4. Gelir Unsuru	7
1.2. KREDİ FONKSİYONLARI VE AVANTAJLARI	8
1.3. KREDİ KAYNAKLARI	9
1.4. KREDİ TÜRLERİ	10
1.4.1. Bireysel Krediler	12
1.4.1.1. İhtiyaç Kredisi	13
1.4.1.2. Taşıt Kredisi	14
1.4.1.3. Eğitim Kredisi	14
1.4.1.4. Konut Kredisi	14

1.4.1.5. Kariyer Kredisi	15
1.4.1.6. Kişisel Gelişim ve Hobi Kredisi	15
1.4.1.7. Evlilik Kredisi	15
1.4.1.8. Tatil Kredisi	15
1.4.1.9. Tadilat ve Modernizasyon Kredisi	16
1.4.1.10. Alışveriş Kredisi	16
1.4.1.11. İnşaat Tamamlama Kredisi	16
1.4.1.12. Kredi Kartı	17
1.4.2. Ticari Krediler	18
1.5. BANKACILIKTA RİSK KAVRAMI VE TÜRLERİ	20
1.5.1. Bankacılıkta Risk Kavramı	20
1.5.2. Bankacılıkta Risk Türleri	20
1.5.2.1 Kredi Riski	20
1.5.2.2. Piyasa Riski	21
1.5.2.3. Likidite Riski	21
1.5.2.4. Faiz Oranı Riski	22
1.5.2.5. Operasyonel Risk	22
1.5.2.6. Döviz Kuru Riski	23

İKİNCİ BÖLÜM

MAKİNE ÖĞRENMESİ TEKNİKLERİ VE YENİDEN ÖRNEKLEME YÖNTEMLERİ

2.1. MAKİNE ÖĞRENMESİ	24
2.1.1. Makine Öğrenmesi Algoritması Türleri	28
2.1.1.1. Denetimli Makine Öğrenmesi	29
2.1.1.2. Denetimsiz Makine Öğrenmesi	30
2.1.1.3. Yarı Denetimli Makine Öğrenmesi	31
2.1.1.4. Pekiştirmeli Makine Öğrenmesi	32

2.2. TEKİL MAKİNE ÖĞRENMESİ TEKNİKLERİ	32
2.2.1. K-En Yakın Komşu	32
2.2.2. Lojistik Regresyon	34
2.2.3. Naïve Bayes	38
2.2.4. Yapay Sinir Ağları	40
2.2.5. Destek Vektör Makinesi	45
2.2.6. Karar Ağaçları Algoritması	53
2.3. TOPLULUK ÖĞRENME TEKNİKLERİ	57
2.3.1. Rassal Ormanlar Algoritması	59
2.3.2. Gradyan Artırma Karar Ağaçları	64
2.3.3. Ekstra Ağaçlar Sınıflandırıcısı	66
2.3.4. Oylama Sınıflandırıcısı	66
2.3.4.1. Sert Oylama (Hard Voting)	67
2.3.4.2. Yumuşak Oylama (Soft Voting)	67
2.4. PERFORMANS DEĞERLENDİRME ÖLÇÜTLERİ	68
2.4.1. Karmaşıklık Matrisi (Confusion Matrix)	68
2.4.2. Doğruluk	70
2.4.3. Kesinlik (Precision / Positive Predictive Value)	71
2.4.4. Hassasiyet (Recall / Sensitivity)	71
2.4.5. F1 Skoru (F1 Score)	72
2.4.6. Matthews Korelasyon Katsayısı	72
2.5. YENİDEN ÖRNEKLEME TEKNİKLERİ	73
2.5.1. Aşağı Örneklem	73
2.5.1.1. Rassal Aşağı Örneklem (Random Undersampling - RUS)	74
2.5.2. Yukarı Örneklem	74
2.5.2.1. Rassal Yukarı Örneklem (Random Oversampling - ROS)	75
2.6. ÖRNEKLEYİCİLER TOPLULUĞU	76
2.6.1. Dengeli Özyinelemeli Birleştirme Sınıflandırıcısı (BBC)	76
2.7. AŞAĞI VE YUKARI ÖRNEKLEME KOMBİNASYONLARI	77
2.7.1. SMOTE-Tomek Links	77
2.7.2. SMOTE-ENN	78

ÜÇÜNCÜ BÖLÜM

UYGULAMA

3.1. ÇALIŞMANIN AMACI	80
3.2. LİTERATÜR TARAMASI	81
3.3. DENEY TASARIMI AŞAMASI	88
3.3.1. Veri Setleri ve Açıklamaları	89
3.3.1.1. Alman Kredi Veri Seti	90
3.3.1.2. Avustralya Kredi Veri Seti	91
3.3.1.3. HMEQ Kredi Veri Seti	91
3.3.2. Veri Ön İşleme Süreci	92
3.3.2.1. Eksik Verilerin Tamamlanması	92
3.3.2.2. Değişkenlerin Özniteliklerine Göre Kodlanması	93
3.3.2.3. Verilerin Dönüştürülmesi Aşaması	95
3.3.3. Veri Kümelerinin Keşfi Aşaması	95
3.3.3.1. HMEQ Ev Kredi Veri Kümesinin Keşif Aşaması	96
3.3.4. Eğitim ve Test Veri Kümelerinin Hazırlanması Aşaması	104
3.3.5. Yeniden Örnekleme Tekniklerinin Uygulanması Aşaması	105
3.4. SONUÇLARIN DEĞERLENDİRİLMESİ VE YORUMLANMASI	106
3.4.1. Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları	106
3.4.2. Avustralya Kredi Veri Setine ait Performans Ölçüm Sonuçları	112
3.4.3. HMEQ Mortgage Veri Setine ait Performans Ölçüm Sonuçları	117
3.4.4. Senaryolar Arasından Optimal Seçim İşlemi	123
SONUÇ	127
KAYNAKÇA	132

KISALTMALAR

ABD	Amerika Birleşik Devletleri
ANN	Yapay Sinir Ağları
BDDK	Bankacılık Düzenleme ve Denetleme Kurulu
BKKK	Banka Kartları ve Kredi Kartları Kanunu
CART	Sınıflandırma ve Regresyon Ağacı
DT	Karar Ağaçları
DVM	Destek Vektör Makinesi
FN	Doğru Negatif
FP	Yanlış Pozitif
GBDT	Gradyan Artırma Karar Ağaçları
Hard Voting	Sert Oylama
HMEQ	Ev Kredisi (Home Equity)
K-NN	K-En Yakın Komşu
LR	Lojistik Regresyon
MCC	Matthews Korelasyon Katsayısı
MLP	Çok Katmanlı Algılayıcı
NB	Naive Bayes
OOB	Out Of Bag
RF	Karar Ağaçları
ROS	Rassal Yukarı Örnekleme
RUS	Rassal Aşağı Örnekleme
SMOTE	Sentetik Azınlık Yukarı Örnekleme Tekniği
Soft Voting	Yumuşak Oylama
TN	Doğru Negatif
TP	Doğru Pozitif
XTrees	Ekstra Ağaçlar

TABLÖLER LİSTESİ

Tablo 1: Karmaşıklık Matrisi (Confusion Matrix) Yapısı	s.68
Tablo 2: Uygulama Aşamasında Yer Alan Senaryolar	s.88
Tablo 3: Alman Kredi Veri Kümesindeki Değişkenlere İlişkin Bilgiler	s.90
Tablo 4: HMEQ Kredi Veri Kümesindeki Değişkenlere İlişkin Bilgiler	s.92
Tablo 5: Dengelenmemiş Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları	s.106
Tablo 6: Senaryo 1: Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları	s.107
Tablo 7: Senaryo 2: Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları	s.108
Tablo 8: Senaryo 3: Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları	s.109
Tablo 9: Senaryo 4: Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları	s.110
Tablo 10: Dengelenmemiş Avustralya Kredi Veri Setine ait Performans Ölçüm Sonuçları	s.112
Tablo 11: Senaryo 1: Avustralya Kredi Veri Seti Performans Ölçüm Sonuçları	s.113
Tablo 12: Senaryo 2: Avustralya Kredi Veri Seti Performans Ölçüm Sonuçları	s.114
Tablo 13: Senaryo 3: Avustralya Kredi Veri Seti Performans Ölçüm Sonuçları	s.115
Tablo 14: Senaryo 4: Avustralya Kredi Veri Seti Performans Ölçüm Sonuçları	s.116
Tablo 15: Dengelenmemiş HMEQ Kredi Veri Setine ait Performans Ölçüm Sonuçları	s.117
Tablo 16: Senaryo 1: HMEQ Veri Setine ait Performans Ölçüm Sonuçları	s.118
Tablo 17: Senaryo 2: HMEQ Veri Setine ait Performans Ölçüm Sonuçları	s.119
Tablo 18: Senaryo 3: HMEQ Veri Setine ait Performans Ölçüm Sonuçları	s.121
Tablo 19: Senaryo 4: HMEQ Veri Setine ait Performans Ölçüm Sonuçları	s.122
Tablo 20: F1-Skoru Açısından En İyi Senaryolar	s.124
Tablo 21: MCC Açısından En İyi Senaryolar	s.125

ŞEKİLLER LİSTESİ

Şekil 1: Kredi Türleri	s. 11
Şekil 2: Bireysel Kredilerin (Tüketici Kredi) Türleri	s. 13
Şekil 3: Tüketici ve Bireysel Kredi Kartlarının Gelişimlerine İlişkin Örnek	s. 17
Şekil 4: Kredi Türleri Gelişimi Örneği	s. 19
Şekil 5: Eğitim Verilerinin Etiket Niteliklerine Göre Kategorize Edilmesi	s. 27
Şekil 6: Makine Öğrenmesi Çeşitleri	s. 29
Şekil 7: Denetimli Makine Öğrenmesi: Sınıflandırma Örneği	s. 30
Şekil 8: Denetimsiz Makine Öğrenmesi: Kümeleme Örneği	s. 31
Şekil 9: K -En Yakın Komşu Örneği	s. 34
Şekil 10: S ve Ters S Şeklindeki Olasılık Fonksiyon Grafikleri	s. 35
Şekil 11: Doğrusal Regresyon Modeli ve Lojistik Regresyon Modeli	s. 37
Şekil 12: Yapay Nöron (Sinir Hücrelerinin) Matematiksel Form Örneği	s. 41
Şekil 13: İki Gizli Katmana Sahip Çok Katmanlı Algılayıcının Yapısı	s. 45
Şekil 14: Doğrusal Ayrılabilen Durum İçin Hiper-Düzlem Belirlenmesi	s. 47
Şekil 15: Doğrusal Ayrılamayan Durum İçin Hiper Düzlemin Belirlenmesi	s. 50
Şekil 16: İki Boyutlu Uzaydaki Verilerin Çekirdek Fonksiyonu ile Üç Boyutlu Uzaya Dönüştürülmesi ve Hiper Düzlemin Belirlenmesi Örneği	s. 52
Şekil 17: Karar Ağacı Yapısı Örneği	s. 55
Şekil 18: Rassal Ormanlar Algoritmasının Genel Yapısı	s. 62
Şekil 19: Bagging ve Boosting Algoritma Yapısı Örneği	s. 64
Şekil 20: Oylama Sınıflandırıcısı Soft ve Hard Yapısı Örneği	s. 67
Şekil 21: Aşağı Örnekleme ve Yukarı Örnekleme Yapısı Örneği	s. 75
Şekil 22: Dengeli Özyinelemeli Birleştirme Sınıflandırıcısı Örneği	s. 76
Şekil 23: HMEQ Ev Kredisi Korelasyon Matrisi	s. 96
Şekil 24: Talep Edilen Kredi Miktarı ve Mevcut Mülk Değerine ait Grafikler	s. 97
Şekil 25: Müşterinin Çalıştığı Toplam Yıl ve Ödenmesi Gereken Kredi Tutarı	s. 98
Şekil 26: Borç- Gelir Oranı ve Kredi Taksit Sayısına ait Dağılım Grafikleri	s. 98
Şekil 27: En Son Alınan Kredi Taksiti ve Son Kredi Sorgulama Sayısı	s. 99
Şekil 28: Mesleki Statülerine Göre Kredi Riskinin İncelenmesi	s. 99
Şekil 29: Müşterinin Krediyi Alma Gerekçesine Göre Kredi Risk Durumu	s. 100

Şekil 30: Mevcut Mülkün Değerine Göre Kredi Risk Durumunun İncelenmesi	s. 101
Şekil 31: Talep Edilen Kredi ile Ödenmesi Gereken Kredi Tutarı ve Sayısı	s. 102
Şekil 32: Müşterinin Mesleği ve Kredi Talep Tutarına Göre Kredi Riski	s. 103



GİRİŞ

Günümüzde ekonomik yapının gelişimi ve kalkınma amaçlarının gerçekleştirilmesine yönelik olarak ortaya çıkan en kayda değer finansman kaynakları arasında krediler yer almaktadır. Krediler, banka ve kredi veren diğer finans kuruluşları tarafından sağlanmaktadır. Bankalar, müşterilerin kredi taleplerini belirli koşullar altında değerlendirmekte ve uygunluğuna göre kredi verme işlemini gerçekleştirmektedir. Bu değerlendirmenin doğru yapılmaması durumunda belirli risk faktörleri oluşmakta ve kredi riskini ortaya çıkarmaktadır. Bu risklerden birisi, kredi talebi için başvuran müşterilerden alınan bilgi ve belgelerin eksik veya gerçeğe aykırı olması sonucu ortaya çıkmaktadır. Bu durumda, kredi kararı esnasında müşteri bilgileri, karar vericilerin yanlış karar almasına ve dolayısıyla uygun olmayan müşterilerin kredi almasına sebep olacak ve kredi riskini oluşturacaktır. Kredi riski, sadece müşteri kaynaklı olmamaktadır. Aynı zamanda, kredi tahsis eden personel belirlenmiş politika ve prosedürler çerçevesinde, müşterilerin vermiş olduğu bilgileri kapsamlı bir şekilde değerlendirmeli ve kredi için en uygun müşterileri belirlemelidir. Aksi takdirde, kredi tahsis edenlerin yanlış bir değerlendirme yapması durumunda, kredi için uygun olmayan müşteriye kredi verilmesine neden olacak ve bir kredi riski söz konusu olacaktır. Bankaların, kredi riski taşıyan müşterileri elle yapılan ve otomatik olmayan bir şekilde tespit etmesi durumunda iş yükü, maliyet, zaman ve kaynak kullanımını artırmakta ve süreç epey zorlu bir hâl alabilmektedir. Riskli müşterilerin daha kolay saptanması ve sürecin iyileştirilmesi amacıyla son yıllarda makine öğrenme tekniklerinden sıklıkla faydalandığı görülmektedir. Bu teknikler bankaların kredi riskini azaltarak kârının artmasına yardımcı olabilmektedir.

Kredi riski problemini makine öğrenme teknikleri ile modellerken karşılaşılan bir problem sınıf dengesizliğidir. Sınıf dengesizliği, bir sınıfın diğer sınıfa göre veri seti içerisinde çok küçük sayıda kalmasıyla ortaya çıkmakta ve kullanılan modelleme tekniklerinin çoğunluk sınıfını aşırı derecede öğrenmesine yol açabilmektedir. Bu çalışma kapsamında son yıllarda geliştirilmiş çeşitli sınıf dengeleme yöntemlerinin faydalanılan makine öğrenme teknikleri üzerindeki performans etkileri incelenmiştir. Bu amaç doğrultusunda, banka ve finans kuruluşlarının, kredi talebinde bulunan müşterilerin kredibilite açısından en uygun değerlendirmeyi yapabilmeleri için

kullanılan sınıf dengeleme yöntemleri farklı makine öğrenmesi teknikleri üzerinde çalıştırılmıştır. Bu yöntemle, kredi talebinde bulunan müşteriler içerisinde kredi almaya en uygun müşterileri belirlemek ve aynı zamanda bankalar açısından kredi riskini azaltarak kârlılığını artırmak amaçlanmıştır. Açık kaynaklı site üzerinden alınan Alman, Avustralya ve Ev kredisi (HMEQ) gerçek hayat problemlerine ait kredi veri seti kullanılmıştır. Bu veri setleri, standart olarak kullanılan rassal aşağı ve rassal yukarı örnekleme yöntemlerine ek olarak Dengeli Özyinelemeli Birleştirme Sınıflandırıcısı (Balanced Bagging Classifier) ve aşağı ve yukarı örneklemeye dayanan iki farklı hibrit yöntem ile dengeli hâle getirilmiştir. Bu veri setleri, dört farklı senaryo üzerinde test edilmiştir. Bu senaryolar kredi verilmesine en uygun müşterilerin belirlenmesi için hangi makine öğrenmesi türünün hangi dengeleme yöntemiyle kullanılacağına belirlenmesine yönelik olarak oluşturulmuştur. Makine öğrenme teknikleri tekil modeller ve topluluk modelleri olmak üzere iki kategoride incelenmesi senaryo oluşturmadaki birinci kriteri oluşturmaktadır. Veriyi dengeli hâle getirmede izlenecek yöntem ise ikinci kriter olarak analiz kapsamında incelenmiştir. Böylece her bir veri seti için en iyi tahmin performansını üreten senaryo tespit edilmiştir. Sonuç olarak, veri dengelemede kullanılan aşağı ve yukarı örneklemeye dayanan hibrit yöntem ve topluluk öğrenmesine dayanan makine öğrenme modellerinin oluşturduğu senaryo en iyi sınıflandırma performansını sergilemiş ve gelecekte banka ve finans kuruluşlarının karşılaşılabileceği benzer problemlerde kullanılabileceği önerisinde bulunulmuştur.

Bu tez üç bölümden oluşmaktadır. Çalışmanın birinci bölümünde, kredi ve bankacılıkta risk kavramı ele alınmış, kredinin kapsamı ve sınırları belirlenerek tanımlanmış ve bankacılık açısından önemi vurgulanmıştır. Ek olarak, kredinin unsurları ve türleri, kredi fonksiyonları ve avantajları, kredi çeşitleri (bireysel ve tüketici kredileri), kredi kaynaklarını oluşturan etmenlerden bahsedilmiştir. Bu bölümün sonunda ise bankacılıkta karşılaşılan risk kavramı ilişkin genel bir tanım yapılmış ve riskin türleri detaylandırılmıştır. Ayrıca, bankacılık ve bu çalışmanın ana amacı için önem arz eden kredi riski bu bölümde kapsamlı bir şekilde ele alınmıştır.

Çalışmanın ikinci bölümünde, makine öğrenmesinin genel bir tanımına yer verilmiş ve çeşitli kullanım alanları ile birlikte kapsamı, işlevselliği ve avantajlarından bahsedilmiştir. Aynı zamanda, makine öğrenmesi algoritması türleri tanıtılmıştır.

Daha sonra, uygulama aşamasında kullanılan tekil modeller ve topluluk öğrenmesi modelleri kapsamlı bir şekilde anlatılmıştır. Ayrıca, kullanılan modellerin performanslarını doğru bir biçimde değerlendirmek için doğruluk, kesinlik, hassasiyet, F1-skoru ve MCC olmak üzere beş farklı performans değerlendirme ölçütünden bahsedilmiştir. Bu bölümün sonunda ise, uygulama aşamasında karşılaşılan veri dengesizliği problemini ortadan kaldırarak verileri dengeli hâle getirmek için rassal aşağı ve rassal yukarı yeniden örnekleme yöntemleri, örnekleyiciler topluluğu; dengeli özyinelemeli birleştirme sınırlandırıcısı ve aşağı ve yukarı örneklemeyle dayanan hibrit yöntemlerden ise Sentetik Azınlık Yukarı Örnekleme Tekniği (SMOTE) ile Düzenlenmiş En Yakın Komşu (Edited Nearest Neighbor veya ENN) yöntemlerinin kombinasyonu olan SMOTE-ENN ve Sentetik Azınlık Yukarı Örnekleme Tekniği (SMOTE) ile Tomek Bağlantıları (Tomek-Links) yöntemlerinin birleşimi olan SMOTE-Tomek Links detaylandırılarak anlatılmıştır.

Çalışmanın üçüncü bölümünde, çalışmanın amacı hakkında genel bir giriş yapılmıştır. Daha sonra, kapsamlı bir literatür taraması gerçekleştirilmiştir. Uygulama aşaması için kullanılan üç veri seti içerisinde HMEQ veri seti hakkında görsel araçlar kullanılarak detaylı bilgi verilmiştir. Dengesiz bir yapıya sahip olan veri kümelerinin dengelenmeden önceki sonuçlarına yer verilmiştir. Daha sonra, dengeli hâle getirilen veri kümelerinin performanslarını daha iyi bir şekilde tespit edebilmek için deney tasarımı aşamasında senaryolar oluşturulmuştur. Oluşturulan dört farklı senaryo tanımlanarak bu senaryolara ait gerekli açıklamalar verilmiştir. Bu senaryolardan elde edilen sonuçlar doğruluk, kesinlik, hassasiyet, F1-skoru ve MCC ölçütlerine göre değerlendirilse de özellikle F1-skoru ve MCC ölçütü açısından en optimal olan senaryo, örnekleme yöntemi ve seçilen model kombinasyonları hakkında detaylı bilgi verilmiştir. Böylece, bankaların kredibilite açısından en uygun değerlendirmeyi yapabilmeleri için dengeli olmayan veri kümelerinde karşılaşılan sınıf dengesizliği sorununa çözüm bulmak için farklı yeniden örnekleme yöntemleri kullanılmış ve en iyi performansı gösteren makine öğrenmesine ait sınıflandırma modellerinin saptanması amaçlanmıştır. Üçüncü bölümde son olarak, uygulamadan elde edilen bulgular yorumlanmış ve gelecekteki çalışmalara katkı sağlamak için değişken seçim yöntemleri, farklı performans değerlendirme ölçütleri ve sınıf dengesizliğinin

giderilmesi için örnekleyiciler topluluđu içerisinde yer alan çeşitli yöntemlerle ilgili önerilere yer verilmiştir.



BİRİNCİ BÖLÜM

KREDİ VE BANKACILIKTA RİSK KAVRAMINA GENEL BAKIŞ

Bu bölümde, kredi tanımı, unsurları, fonksiyonları ve avantajları, kredi kaynakları, tedarik eden kuruluşlar, kredi türleri ve bankacılık açısından risk kavramı ve türleri ele alınarak detaylandırılacaktır.

1.1. KREDİ KAVRAMI

Kredi kavramı; kelime kökeni olarak “*inanmak, güvenmek ve itibar*” anlamlarına gelen “*Credere – Credibilitas*” kelimesinden türetilmiş olan Latince bir kelimedir (Edwards, 2004: 10). “Kredi” kelimesi uzun yıllar geçmesine rağmen ilk türetildiği köke ve anlamına olan yakınlığını korumuştur. Kredi kelimesi anlamı itibariyle; borç verenler veya alacaklı durumdakiler, anapara tutarının yanı sıra, anlaşma yükümlülük esasına göre belirli bir faiz oranı ile birlikte borçlunun borcunu vadesi geldiğinde geri ödeme yükümlülüğünü yerine getireceğine dair güvene dayalı olarak “kredi” verilmesidir.

Kredi, temel olarak iki prensibe dayanmaktadır:

- Kredi alan borçlu, borcunu geri ödemeye isteklidir.
- Kredi alan borçlu, borcunu geri ödeme yeteneğine sahiptir.

Kredi alan borçlunun, krediyi geri ödemesi konusunda istekli olmasından kasıt, alacaklının borcu geri ödemesi konusunda güvenilir olduğunu ifade etmektedir. İkinci öncül, kredi alan borçlunun, alacaklının borcunu geri ödeyebilmesi için gerekli mali durumu olduğuna ilişkin güvenilir bir değerlendirme yapılması esasına dayanır (Golin ve Delhaise, 2013: 1).

19.10.2005 tarihli Resmî Gazete’de yayımlanan 5411 sayılı Bakanlık Kanunu’nun 48. maddesinde kredinin doğrudan yasal bir tanımına yer verilmemekle birlikte, kredi sayılabilecek işlemlerin kapsamı ve sınırları belirtilmiştir. İlgili maddeye göre;

Bankalarca verilen nakdî krediler ile teminat mektupları, kontrgarantiler, kefaletler, aval, ciro, kabul gibi gayrinakdî krediler ve bu niteliği haiz taahhütler, satın alınan tahvil ve benzeri sermaye piyasası araçları, tevdiatta bulunmak suretiyle ya da herhangi bir şekil ve surette verilen ödünçler, varlıkların vadeli satışından doğan alacaklar, vadesi geçmiş nakdî krediler, tahakkuk etmekle birlikte tahsil edilmemiş faizler, gayrinakdî kredilerin nakde tahvil olan bedelleri, ters repo işlemlerinden alacaklar, vadeli işlem ve opsiyon sözleşmeleri ile benzeri diğer sözleşmeler nedeniyle üstlenilen riskler, ortaklık payları ve Kurulca kredi olarak kabul edilen işlemler izlendikleri hesaba bakılmaksızın bu Kanun uygulamasında kredi sayılır (T.C. Resmî Gazete, 2005: madde 48).

1.1.1. Kredi Unsurları

Kredi tanımı itibari ile dört temel unsurdan oluşmaktadır. Bunlar; zaman, risk, güven ve gelir unsurlarıdır.

1.1.1.1. Zaman Unsuru

Bankalar tarafından geri ödenmek üzere verilen paranın belirlenmiş bir vade içerisinde krediyi kullanan borçlu kişi tarafından iade edilmesi gerektiğinden, kredide zaman unsuru oldukça önemli bir faktördür. Belirlenen ödeme süresindeki gecikme riskini de beraberinde getirir. Bunun en önemli sebeplerinden biri gelecekte karşılaşılabilecek olası durumların bilinmemesidir.

Kredinin belirlenen süre içerisinde ilgili bankalara geri ödenmesi, krediyi kullanan borçlunun itibarını olumlu etkileyerek ve bankalar açısından da sürdürülebilirlik ve kârlılık bağlamında önem arz etmektedir. Başka bir ifadeyle, geri ödeme süresindeki gecikmelere bağlı olarak, borçlunun aleyhine bir durum olup maliyetin, krediyi sağlayanların ise lehine bir durum olarak getirinin artmasını ifade etmektedir (Usuç, 2016: 9).

1.1.1.2. Güven Unsuru

Bankalar, kredi talebinde bulunan müşterilerin işlemlerini gerçekleştirebilmek için kısmen veya tam koruma sağlamak amacıyla krediyi kullanan borçludan bir tür teminat vermesini isteyebilir. Bu teminat, kişisel ve kurumsal olabileceği gibi bir varlık taahhüdü ya da yükümlülüklerin yer aldığı bir güvenlik sözleşmesinin imzalanması ile gerçekleştirilebilir. (NACM, 2016).

Bankalarca, güven unsurunun oluşabilmesi için kredi talebinde bulunan kişi ve/veya kurumlarla ilgili kapsamlı incelemeler yapılmaktadır. Bu incelemeler, mali tabloları, diğer bankalarla olan kredi tutarları ve kredilerin kullanım amaçları gibi kredi veren kurum açısından müşterilerin yükümlülüklerini yerine getirip getiremeyeceği hakkında bilgi sahibi olması açısından oldukça önemlidir. Aynı zamanda, kredi alanların da bankalara güvenmesi önemlidir. Bankaların temel olarak beklentileri, kredi talep eden kişilerden geri ödeme süresi içerisinde belirlenen tutarları tahsil etmeleri ve ödeme ile ilgili bir problem oluşmamasıdır. Vade süresinde ilgili borcun ödenmemesi durumunda, bankaların kredi kullanan borçlulara olan itibarını kaybetmesine neden olmaktadır (Aras, Güler, 1996: 5; Usuç, 2016: 9).

1.1.1.3. Risk Unsuru

Risk, bankaların vermiş olduğu kredilerin borçlular tarafından alınan ve vadesinde ödemesi gereken tahsilatın çeşitli nedenlerden dolayı yükümlülüklerin yerine getirilmemesi olarak ifade edilir. Krediler yapısı gereği risk unsurunu barındırmaktadır. Bu riskin artması veya azalması, kredi kullanan borçlular tarafından taahhüdün yerine getirilip getirilmediği ve alınacak teminatın türüne göre değişkenlik göstermektedir. Bankalarca, kredi sağlanan kişi ve/veya kurumların ekonomik gücünün göz önünde bulundurulması gerekmektedir. Kişi ve/veya kurumların olası bir kriz döneminde üzerinde anlaşılan koşulları sağlayamaması durumunda, bankaların kârlılığını olumsuz etkilerken, kredi talep eden borçlular için ise, banka tarafından duyulan güvenin zedelenmesine yol açarak temerrüde düşmesine sebep olur. Bankaların bu tip müşterilerin kredi taleplerine olumsuz yanıt vermesi muhtemeldir. Sonuç olarak, risk faktörü yalnızca içinde bulunulan durumu olumsuz etkilemez. Aynı zamanda, gelecekte oluşabilecek ihtiyaçlara yönelik çeşitli durumları da olumsuz etkilemektedir (Aras, 1996: 5; Kavcıoğlu, 2003: 9; Usuç, 2016:10).

1.1.1.4. Gelir Unsuru

Kredi, bankaların en önemli gelir kaynaklarını oluşturur. Bankalar, hâlihazırda bulunan fonlarını ve farklı şekilde sağlanan kaynaklarını kredi olarak müşterilerine sunmaktadır. Kredi kullanan borçlular, belirli bir faiz veya komisyon ile bankalardan kredi temin etmektedirler. Kredinin gelirini oluşturan faiz ve komisyonlar, kredi türüne

ve mhletine gre deęiřkenlik gstermekte ve banka gelirinin nemli bir blmn oluřturmaktadır. Nakit krediler doęrudan faizin gelirini meydana getirirken bazı nakit kredi eřitlerinde faize ek olarak komisyon gelirleri de saęlanır. Komisyonlar, doęrudan kredi zellięinde bulunmayan nakit olmayan kredilerin gelirini oluřturur. Bankalar, faiz ve komisyonlar ile kr elde etmeyi amalamaktadır. Bylece, paydařlarına kr saęlamanın yanı sıra maddi varlıklarında artıř saęlayarak srdrlebilir olmayı hedeflerler (Usuę, 2016: 10).

1.2. KREDİ FONKSİYONLARI VE AVANTAJLARI

Kredi, ekonomik yapının geliřmesi ve kalkınma amalarının gerekleřtirilmesi iin en nemli finansman kaynakları arasında yer almaktadır. Ekonomik geliřimin ve kalkınmanın srdrlebilirlięi aısından kiři ve/veya kurumlar, ticari faaliyetlerini yerine getirmek iin finansman gereksinimi duyarlar. Bu gereksinimler, finansal kuruluřlar arasında en nemli yere sahip olan bankalar tarafından dn verilen fonlardan karřılanmaktadır. Bu baęlamda, ekonomik faaliyetlerin sreklilięi aısından kiři ve/veya kurumların gereksinimleri olan sermaye, tasarruf sahipleri tarafından kredi olarak saęlanmaktadır. Aynı zamanda kredi, tedavl aracı olma zellięi sayesinde likit paranın tařınmasını azaltırken, ticari iřlemlerin yrtlmesini hızlandırmaktadır. İřletmeler ticari faaliyetlerini gerekleřtirirken sermaye gereksinimlerini krediler ile saęlamaktadırlar. Krediler iřletmelerin, retim ve hizmet faaliyetleri aracılıęıyla lke ekonomisine ve mevcut iř gcne olumlu katkı saęlamasına ve mevcut kořullarda varlıęını srdrmesine olanak saęlamaktadır. Ekonomide tasarruf sahipleri tarafından hazır bulunan atıl durumdaki fonlar, bankalar vasıtasıyla kiři ve/veya kurumlara kredi olarak aktarılır. Bylece, tasarruf sahipleri aısından gelir elde edilirken, fon ihtiyacı olan kiři ve/veya kurumların da finansman ihtiyacı karřılanır. Tasarruf sahipleri tarafından hazır bulunan fon fazlasının bireysel olarak deęerlendirilmesi olduka sınırlı olmaktadır. Bor verilen kiřiye iliřkin borcunu deme gcnn tam olarak doęru deęerlendirilmesi imknlarına sahip deęildirler. Bu deęerlendirme iřleminin uzman kiřiler tarafından fon saęlayan bankalar aracılıęıyla yrtlmesiyle en uygun yatırımcılara ynlendirildięinden, atıl fonların etkin bir řekilde kullanımı saęlanır ve tasarruf sahipleri ile bankalar arasında risk paylařılmış

olur. Böylece, tasarruf sahipleri teşvik edilir ve ekonomik büyüme için gereken en önemli faktörlerden biri olan sermaye stoku sağlanmış olur (Aras, 1996: 6; Katircioğlu, 2006: 7; Kıran, 2013: 4).

1.3. KREDİ KAYNAKLARI

Bankalar varlıklarını ve faaliyetlerini sürdürebilmek için parasal kaynağa ihtiyaç duymaktadırlar. Bu parasal kaynaklar; öz sermaye ile borçlardan (yabancı kaynaklar) oluşur ve büyük bir çoğunluğu yabancı kaynaklardan karşılanmaktadır. Bankalar, borçlanarak veya mevduat ihracı yoluyla mali kaynak elde eder. Bu toplanan mali kaynaklar ile temin edilen varlıklardan, krediler ve menkul kıymetler en önemlileri arasında yer almaktadır (Kılıç, 2019: 46). Finansal kaynaklar; öz kaynak ve yabancı kaynak olmak üzere iki gruba ayrılır. Vergi Usul Kanunu'nun 192. maddesi esasınca öz sermaye; "Aktif toplamı ile borçlar arasındaki fark, müteşebbisin (girişimci) işletmeye mevzu varlığını (öz sermayeyi) teşkil eder." (Vergi Usulü Kanunu, 1961: 3518). Başka bir ifadeyle öz sermaye, işletme ortak veya ortaklarının işletmenin sahip olduğu mevcut varlıkları (öz sermaye + borçları) üzerindeki hakların parasal değeri olarak ifade edilir (Abdioğlu, 2019: 2). Bankalar varlıklarını ve faaliyetlerini yürütebilmek için yapılacak yatırımların öz sermayeler tarafından karşılanamayan sermaye ihtiyacı, yabancı kaynaklardan sağlanmaktadır. Yabancı kaynaklar vadeli bir finansman kaynağı olduğu için geri ödenmek üzere alınan kaynakları ifade etmektedir (Ayboğa, 1991: 4-5). Para piyasasının en temel aracı kuruluşları olarak bilinen bankalar, temel fon kaynaklarını öz sermayesinden değil, yabancı kaynaklarından sağlamaktadırlar (Özdemir, 2019: 68).

Bankaların sahip olduğu kredi kaynaklarını oluşturan öz sermaye ve yabancı kaynak (borç) kalemleri aşağıdaki gibidir (Ayboğa, 1991: 4-5; Kavcıoğlu, 2003: 10; Usta, 1994: 9).

Öz Kaynaklar;

- Ödenmiş Sermaye
- Yedek Akçeler (İhtiyatlar)
- Karşılıklar
- Dağıtılmamış Kârlar
- Fonlar
- Birikmiş Amortismanlar
- Net Kâr

Yabancı Kaynaklar (Borçlar)

• Mevduat (Ulusal Para (Türk lirası) ve Yabancı Para (Döviz))

• T.C. Merkez Bankası tarafından sağlanan krediler,

• Yurt içi Bankalardan alınan krediler,

• Yurt dışı Bankalardan alınan krediler,

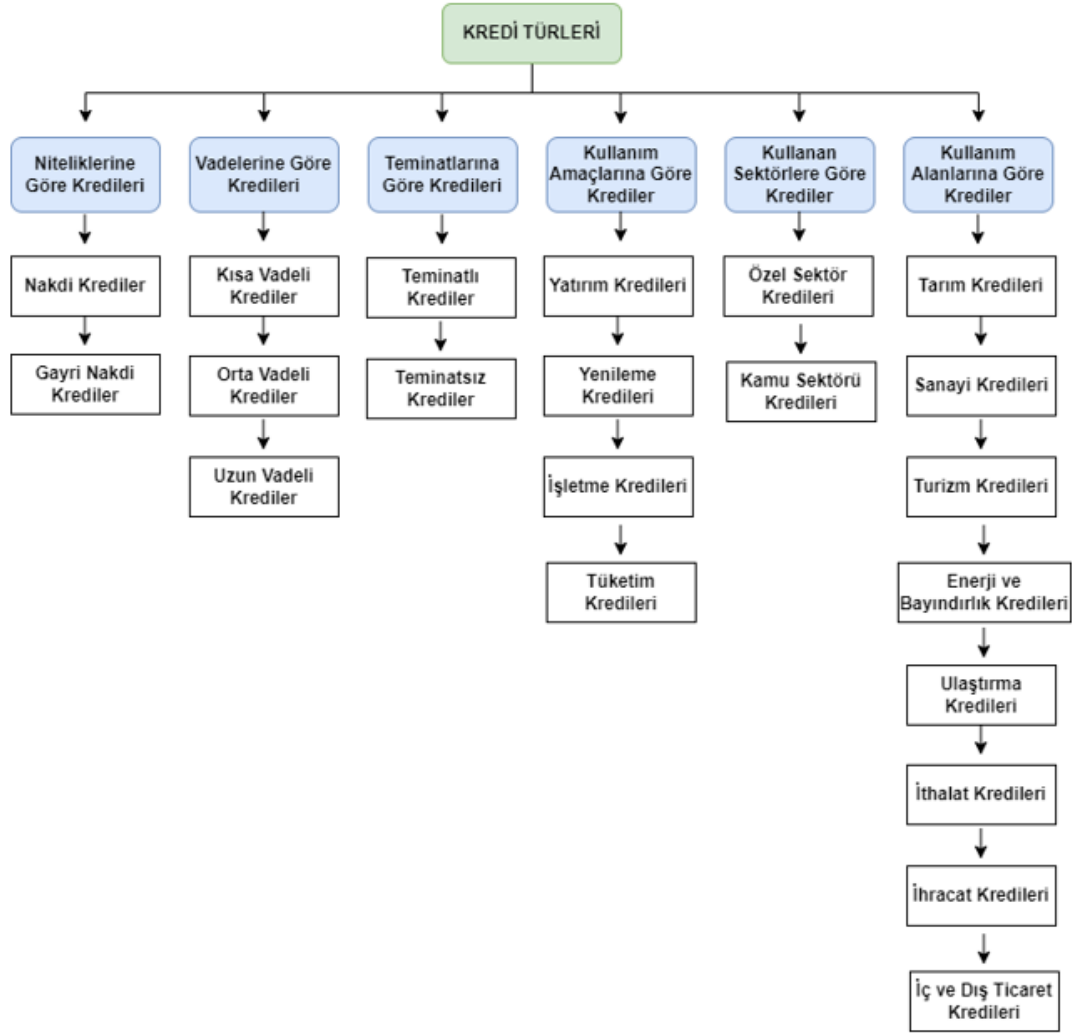
• Muhtelif borçlar (Fonlar, yatırılacak vergiler, ödenmesi gereken havale bedelleri, keşide çeki bedeli ve öz sermaye artırımına araçlık etme gibi bankacılık hizmetleri ile kullandırılmakta olan bazı krediler için alınan nakit blokajlardır.

- İhraç Kredileri
- Tahvil ihracı
- Menkul Kıymetler
- Dağıtılacak Kârlar
- Diğer borçlarından oluşmaktadır.

1.4. KREDİ TÜRLERİ

Kişi ve/veya kurumlar farklı nedenlere bağlı olarak kredi gereksinimi duymaktadırlar. Bu farklı nedenlere göre kredi türlerinde çeşitlilik meydana gelmektedir. Kredi türleri; Nitelikleri, vadeleri, teminatları, kullanım amaçları, kullanan sektörler ve kullanım amaçlarına göre çeşitli türlere ayrılmaktadır. Kredi türleri Şekil 1’de gösterildiği gibidir (Aras, 1996: 9-10).

Şekil 1: Kredi Türleri



Kaynak: Aras, 1996: 9-10; Bakanoğlu, 2005: 16-17.

Bankaların bilançoları incelendiğinde, sağ tarafında fon kaynaklarını oluşturan pasif kısım yer alırken, sol tarafında temin edilen fonların kullanımlarına ilişkin hesaplar aktif kısım yer almaktadır. Pasif bölüme ilişkin hesaplar; öz sermayesi, yurt içi ve yurt dışından alınmış olan krediler, mevduat ve mevduat dışı gibi kaynaklar ile fon ihtiyaçlarını karşılanır ve bu kalemler bilançonun pasif kısmını oluşturur. Pasif kalemlerde gösterilen fon kaynakları ve bankaların finanse ettiği krediler ile menkul kıymet alınması gibi işlemler için kullanılan fonlara ilişkin hesapların yer aldığı bölüm bilançonun aktif kısmını oluşturmaktadır. Ayrıca nakdi krediler olarak bilinen kredi çeşitleri aktif bölümdeki kredileri oluşturur. Bankalar kişi ve/veya kurumların

ihtiyaları doėrultusunda nakdi krediler saėlamaktadır. Bu saėlanan krediler tketim ve yatırım bařta olmak zere, ekonomik istikrarın ve toplam talebin saėlanması iin tzel veya gerek kiřilere saėlanan fon kaynaklarını ifade etmektedir. Bu baėlamda, bireysel ve ticari krediler olmak zere nakdi krediler iki alt gruba ayrılır (Togay, 2013: 99).

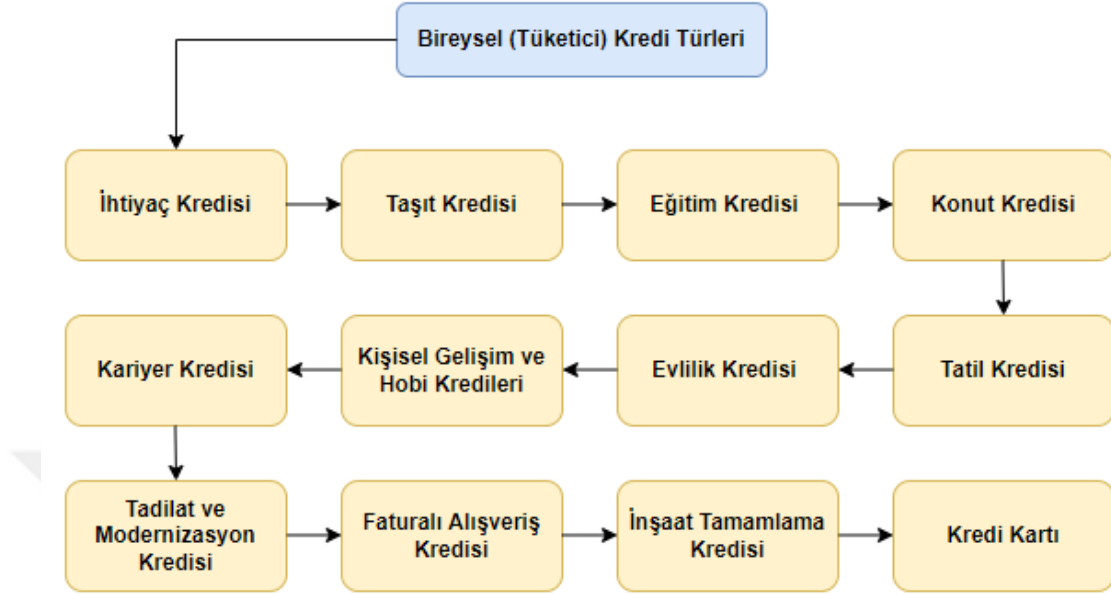
Bu tez kapsamında bahsi geen nakdi kredi trleri detaylandırılmıřtır.

1.4.1. Bireysel Krediler

Nakdi kredileri oluřturan unsurlardan biri olan bireysel krediler, ticari amalı kullanılan kredilerin aksine iřletmeye fon saėlamazlar. Bireysel krediler, kiřilerin ihtiyalarına ynelik saėlanan krediler olarak bilinmektedir. Bu kiřilerin ihtiyalarına gre eřitlilik gstermektedir. Bireysel krediler, gayrimenkul gibi yatırım amaları malların dıřında, bulařık makinesi, eėitim ve saėlık gibi birok alandaki ihtiyaa ynelik olarak kullanılmaktadır. Bankalarca saėlanan bu krediler, bireyler iin ekonomik olması aısından kolaylık saėlar. Genel olarak bankalar, hedeflediėi mřteri profili olan orta ve st gelir dzeyine sahip mřterilerine bireysel kredi kullandırmayı amalar. Bu mřteriler, aktif olarak alıřan veya emeklilik sonucu dzenli gelir elde eden kiřilerdir (Sipahi, 2011: 49). Ayrıca bankalar, belirlenen vade ve faiz oranında eřit taksitlerle geri demek zere kiřilere verdiėi kredilerle kendileri de anaparaya ek olarak faiz elde etmektedirler.

Bireysel krediler genel hatları itibariyle řekil 2'deki gibidir.

Şekil 2: Bireysel Kredilerin (Tüketici Kredi) Türleri



Kaynak: Aksoy, 2005: 74-77.

Bu tez kapsamında bahsi geçen bireysel kredilerin (tüketici kredi) türleri ve bunlara ilişkin açıklamalar kapsamlı bir şekilde ele alınmıştır (Aksoy, 2005: 74-77; Deveci, 2019: 47-53; Sayıcı, 2017: 42-44).

1.4.1.1. İhtiyaç Kredisi

İhtiyaç kredisi, tüzel olmayan kişilere ticari amaçların dışında mal ve hizmet ihtiyaçlarının karşılanabilmesi amacıyla nakit olarak tek seferde kullandırılan, kişilerin belirlenmiş bir faiz oranında ve ödeme aralıkları ile eşit taksitler (anüite) şeklinde geri ödenmek üzere verilen kredi türüdür. Bu kredinin en büyük avantajı nakit olarak tek seferde kullandırılmasından dolayı alışverişlerin taksitli değil peşin olarak yapılmasına ve bunun sağladığı indirimlerden yararlanmaya imkân sunmasıdır. Ayrıca, kredi türleri başlığı altında yer alan teminatlarına göre krediler; teminatlî ve teminatsız olmak üzere iki gruba ayrılan krediler bu kapsam dahilindedir.

1.4.1.2. Taşıt Kredisi

Taşıt kredileri; bireylerin ticari amacı olmaksızın herhangi bir marka aracın tüzel olmayan kişiler tarafından sıfır veya ikinci el taşıtların peşin olarak satın almasına imkân sunan finansman kaynağı sağlayan bir kredidir. Bu krediler, bankalar tarafından gerçek kişilere yönelik olarak belirlenen vade ve faiz oranında eşit taksitlerin ödemesi ile gerçekleştirilir. Bankalar, bireylere kredi imkânı tanırken, aynı zamanda kendilerini de güvence altına almak adına teminat istemektedirler. Bankalar tarafından taşıt kredisi verilirken, bankalar ve bireyler arasında bir anlaşma imzalanır. Bu anlaşma gereğince taşıt kredisi ile alınan araç için bankalar taşıt rehni veya çeşitli teminatların gösterilmesi talep etmektedir ve böylece, banka kendini güvence altına almış olur. Aksi taktirde, bankaların ilgili araç için verilen kredileri geri alması oldukça zordur.

1.4.1.3. Eğitim Kredisi

Eğitim görmekte olan öğrencilerin, eğitim harcamalarına yönelik finanse edilmek üzere reşit olan kişiler için doğrudan kendisine, reşit olmayan kişiler için ise ebeveynleri aracılığı ile banka tarafından sabit faiz ve geri ödenmesi eşit taksitler şeklinde yapılmak üzere finansal kaynak olarak verilen kredidir. Bu kredilerin verilebilmesi için bankanın belirlemiş olduğu koşulların sağlanması gerekmektedir. Kredi talebin için başvuru yapan kişiler öğrenci olduğunu ve eğitim masraflarını belgelemeli ve kredileri geri ödeyebilmeleri için ise hâlihazırda çalışıyor ve taksitleri ödeme gücüne sahip olması gerekmektedir.

1.4.1.4. Konut Kredisi

Bireyler, yaşam standartlarını iyileştirmek ve düzenli bir hayat yaşayabilmek için en önemli unsurlardan biri kişilerin yaşadıkları konutlardır. Bireyler, konut sahibi olmak isterler. Ancak bu konutlar için gerekli finansman kaynaklarını karşılayabilecek ekonomik güçleri olmayabilir veya kısmen karşılayabilecek durumdadır. Böyle bir durumda, kişilerin finansman kaynaklarının sağlanması için başvurdukları ilk yer bankalardır. Bankalar, kişilere belirli değerlendirme kriterlerinden sonra uygun bulmaları hâlinde gerekli finansmanı kısmen veya tamamını sağlamaktadırlar. Bankalarca, bireylere sağlanan bu finansman ile bireyler kira öder gibi ev sahibi olma

imkânı kavuşur. Dahası, satın alınmak istenen konutlar yeni tamamlanmış olabileceği gibi önceden kullanılmış konutlar için de konut kredisi verilmektedir. Bankalarca sağlanan bu kredilerle bankalar kendini güvene almak için satın alınan konutu ipotek etmektedir.

1.4.1.5. Kariyer Kredisi

Dil öğrenme, yüksek lisans ve/veya doktora eğitimi, mesleki açısından kendilerini geliştirebilmeleri için eğitim (programlama, finans, grafik tasarım v.b.) kurslarına yönelik yapılan kariyer harcamalarının finanse edilmesi amacı ile bankalar tarafından gerçek kişilere yönelik olarak belirlenen vade ve faiz oranında eşit taksit veya kredili mevduat şeklinde ödemesi ile gerçekleşen kredilerdir.

1.4.1.6. Kişisel Gelişim ve Hobi Kredisi

Bireyler, ilgi alanlarına göre; yemek, seramik, çinicilik, müzik enstrümanları, sportif faaliyetler gibi hobi ve kişisel gelişime yönelik yapılan harcamaların karşılanması amacıyla bankalar tarafından gerçek kişilere yönelik olarak belirlenen vade ve faiz oranında eşit taksit veya kredili mevduat şeklinde ödemesi ile gerçekleşen kredilerdir. Bu kredi türü doğrudan birey ve banka arasında yapılacağı gibi ilgili faaliyet alanlarındaki kurum ve banka arasında da yapılabilir.

1.4.1.7. Evlilik Kredisi

Evliliklerin en fazla ilk aylarında olan veya evlenme kararı almış ve evlilik sürecindeki çiftlerin, evlilik için gerek çeyiz ve mobilya gerekse düğün ve balayı masrafları için bu süreçte yapmış olduğu her türlü masrafların finanse edilmesi için bankalar tarafından gerçek kişilere yönelik olarak belirlenen vade ve faiz oranında eşit taksit şeklinde ödemesi ile verilen kredilerdir.

1.4.1.8. Tatil Kredisi

Bireylerin yaşamış oldukları ülke sınırları içerisinde ya da kendi ülkesi dışındaki ülkelerde gerçekleştireceği turizm, bilimsel ve mesleki araştırmalar, sağlık, eğlence gibi bireysel veya bir organizasyon dahilinde yapılan seyahat ve her türlü

aktiviteden dolayı oluşabilecek harcamaların karşılanabilmesi için finansman desteğine ihtiyaç duyarlar. Bu finansman desteği, bankalar tarafından belirlenen vade ve faiz oranında eşit taksitler şeklinde geri ödenmek üzere gerçek kişilere verilen kredilerdir.

1.4.1.9. Tadilat ve Modernizasyon Kredisi

Bireylerin yaşamış oldukları konutlarında yapmış oldukları tamirat, bakım ve onarım, yenileme, dekorasyon ve dış cephe kaplama gibi konut içi yaşam alanlarında veya konut dışında yapılacak her türlü masrafların karşılanmasına yönelik olarak bankalar tarafından belirlenen vade ve faiz oranında eşit taksitler şeklinde geri ödenmek üzere gerçek kişilerin kullanımına sunulan kredilerdir.

1.4.1.10. Alışveriş Kredisi

Tüketicilerin, bireysel ve/veya bankalarca anlaşmalı olan kuruluşlarda yapmış olduğu bilgisayar, cep telefonu, televizyon, beyaz eşya ve mobilya gibi yüksek fiyatlı ürünler, kişilerin gelir düzeyinin üstünde olması veya çeşitli nedenlerden dolayı doğrudan nakit olarak karşılanmasının mümkün olmadığı durumlarda alışveriş kredisi kullanılabilir. Bu kredi, bireylerin ihtiyaçlarını karşılamaya olanak sağlarken, aynı zamanda ödeme açısından da kolaylık sağlamaktadır. Bankalar tarafından belirlenen vade ve faiz oranında eşit taksitler şeklinde geri ödenmek üzere gerçek kişilerin kullanımına sunulan kredilerdir.

1.4.1.11. İnşaat Tamamlama Kredisi

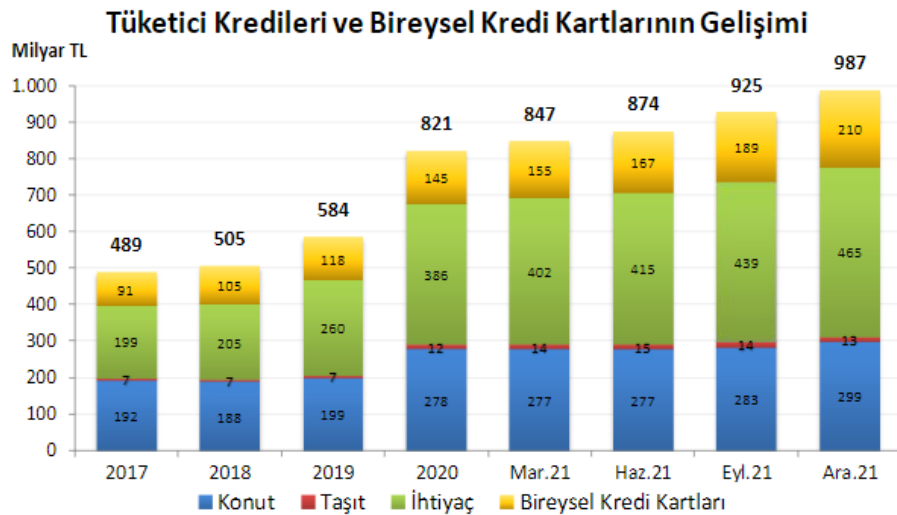
Bireylerin kendi arsaları üzerine yaptıkları konutların veya profesyonel bir ekip tarafından yapılmakta olan ve henüz inşa durumundaki konutların gerçek kişiler tarafından satın alınması ve konutların tamamlanması gereken kısımları için bankalar tarafından belirlenen vade ve faiz oranında eşit taksitler şeklinde geri ödenmek üzere gerçek kişilere verilen kredilerdir.

1.4.1.12. Kredi Kartı

Günlük hayatımızın bir parçası olan kredi kartları, ödeme aracı olarak çeşitli işlemde kullanım imkânı sunmakta ve 5464 sayılı Banka Kartları ve Kredi Kartları Kanunu’nun 3. maddesine göre; “Nakit kullanımı gerekmeksizin mal ve hizmet alımı veya nakit çekme olanağı sağlayan basılı kartı veya fizikî varlığı bulunmayan kart numarasını ifade eder.” biçiminde tanımlanmıştır (BKKK, 2006: 2). Bahsi geçen kanun kapsamından da anlaşılacağı üzere; kişilerin günlük yaşantıları için önemli bir yeri olan kredi kartları, banka ve/veya bankalar gibi yetkinliğe sahip olan diğer kuruluşlar tarafından, insanların satın almak istedikleri mal ve hizmet alımlarını karşılayabilmeleri için doğrudan nakit olarak verilmeyen, belirli kriterler dahilinde limitli olarak verilen ve nakit para çekme özelliği bulunan kısa süreli nakdi olmayan finansal enstrümanlardır.

Kredi kartlarının kullanıcılar için çeşitli avantajları bulunmaktadır. Bu avantajlar; nakit para taşımanın külfeti ve oluşturacağı riski önleme, kişilerin yapmış olduğu alışveriş tarihi ile kredi kartına ait borcun son ödeme tarihine kadar olan süre içerisinde faizsiz kullanım imkânından faydalanma, ilgili karta ait borcun taksitler hâlinde ödeme kolaylığı sağlaması, nakit avans çekimi ve online alışveriş yapılması gibi kullanımı kolaylaştırıcı özellikleri bulunmaktadır (Dülger, 2004: 176).

Şekil 3: Tüketici ve Bireysel Kredi Kartlarının Gelişimlerine İlişkin Örnek



Kaynak: BDDK, 2021: 12.

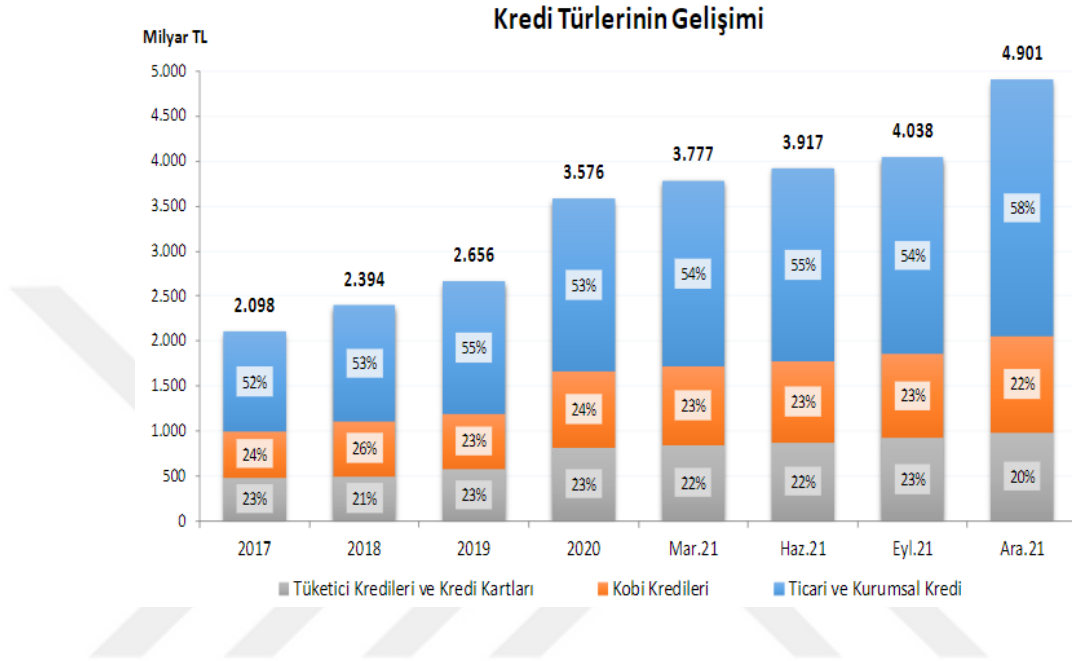
Yukarıdaki şekilde tüketici ve bireysel kredi kartlarının gelişimine ilişkin grafik, BDDK tarafından yayımlanan Aralık 2021 dönemi itibariyle; konut kredilerinin tutarı 299 milyar ₺, bireysel kredi kartlara ait tutar 210 milyar ₺, taşıt kredilerinin tutarı 13 milyon ₺ ve ihtiyaç kredilerindeki tutar ise 465 milyar ₺ seviyesindedir (BDDK, 2021: 12). Bu bağlamda, Şekil 3 incelendiğinde, konut, taşıt ve ihtiyaç kredilerinin birleşimi ile oluşan tüketici kredileri içerisindeki en büyük paya sahip olan kredi türünün ihtiyaç kredisi olduğu görülmektedir.

1.4.2. Ticari Krediler

Ticari krediler, ticaret ve endüstri kollarında faal durumda olan firmalara işletme ve yatırım için gerekli fon kaynaklarının sağlanması amacıyla verilen krediler “ticari krediler” olarak adlandırılmaktadır (Sipahi, 2011: 50). Gerçek ve/veya tüzel kişiler tarafından işletme faaliyetlerinin devam ettirilebilmesi için fon kaynaklarına ihtiyaç duyarlar. Bu fon kaynakları genel olarak bankalar tarafından sağlanan ticari krediler ile elde edilmektedir. İşletmeler çeşitli nedenlerden dolayı ticari krediler almak zorunda kalırlar. Makine ve/veya teçhizat modernizasyonu, ham maddenin tedarik edilmesi, teknolojik inovasyon, işletme kapasitesinin artırılması veya stok finansmanı, çalışan işçilere ödenen maaşlar, elektrik, su gibi işletmenin giderlerini oluşturan kalemler her zaman işletmenin öz sermayesi ile karşılanamamaktadır. İşletmeler konjonktürel yapı içerisinde olduğundan dolayı tahsilat işlemlerinde gecikmeler yaşanmaktadır. Bu nedenle, işletmeler gider kalemlerini karşılayabilmek ve ekonomik faaliyetlerini aksatmamak amacıyla bankalardan fon kaynakları olarak ticari kredi talep ederler. Bu ticari krediler, işletmelerin taleplerine bağlı olarak vade süresi açısından değişkenlik gösterir. İşletmelerin kısa dönemde satıştan elde edilen gelirleri ile giderleri her zaman planlar dahilinde uyum göstermezler. Bu da işletmeler için fon ihtiyacı oluşturur ve bu fonlar kısa vadeli ticari kredilerden sağlanmaktadır. Öte yandan, işletmeler, üretim kapasitelerini genişletmek ve ekonomik istikrarını sürdürmek amacıyla yatırım yapmaları gerekmektedir. Bu yatırımlar, orta ve uzun vadeli finansman gerektirir. Ek olarak, taksitli ticari krediler, ticari kredi kartları, sendikasyon kredileri, döviz endeksli kredi olmak üzere çeşitli ticari krediler bankalar

tarafından sağlanmaktadır (Kaban, 2017: 198; Togay, 2013: 99). BDDK tarafından yayımlanan kredi türleri gelişi örneği Şekil 4’te gösterilmektedir.

Şekil 4: Kredi Türleri Gelişimi Örneği



Kaynak: BDDK, 2021: 11.

Yukarıdaki şekilde kredi türlerinin gelişimine ilişkin grafik, BDDK tarafından yayımlanan Aralık 2021 dönemi itibariyle; ticari ve kurumsal krediler 2.840 milyar ₺, KOBİ kredileri 1.074 milyar ₺ ve tüketici kredileri ve kredi kartları ise 987 milyar ₺ olarak gerçekleşmiştir (BDDK, 2021: 11). Bu bağlamda, Şekil 4 göz önünde bulundurularak, ticari kredilere olan taleplerin sürekli olarak arttığı anlaşılmaktadır. Ayrıca, tüketici kredileri ve kredi kartları; konut, ihtiyaç ve taşıt gibi krediler oluşturmaktadır. Sonuç olarak, kredi türleri gelişimi grafiğine ilişkin ilgili dönem ve yıllar dikkate alındığında, en yüksek payın ticari ve kurumsal kredilere ait olduğu görülmektedir.

1.5. BANKACILIKTA RİSK KAVRAMI VE TÜRLERİ

1.5.1. Bankacılıkta Risk Kavramı

Risk kavramı; kelime kökeni olarak “*tehlike, risk ve şanssızlık*” anlamlarına gelen “*risque*” kelimesinden türetilmiş olan Fransızca bir kelimedir. Bu kelime, İtalyanca “*rischio*” sözcüğünden alıntılanmıştır (Online Etymology Dictionary, 2022). Risk kavramına ilişkin olarak farklı alanlara göre değişik tanımlamalar yer almaktadır.

“*Bankacılıkta risk kavramı; verilen bir kredinin veya girilen bir taahhüdün tahsilinde ya da yerine getirilmesinde başarısızlık ihtimâli olarak tanımlanırken, Sigortacılık alanında risk; yitirme tehlikesinin varlığı, yitirme ihtimâli, belirsizlik, gerçek sonucun beklenen sonuçtan farklı olması ihtimâli, beklenen durumdan başka herhangi bir durumun ortaya çıkması ihtimâli anlamına gelmektedir. Finans alanında ise risk; işletmenin finansal yönüyle ilgili planlanan veya beklenen herhangi bir durumun meydana gelmesinde ortaya çıkan sapma ihtimâlidir.*” (Boyacıoğlu, 2005; Parlakkaya, 1996: 5).

1.5.2. Bankacılıkta Risk Türleri

1.5.2.1 Kredi Riski

Ekonomide tasarruf sahipleri tarafından hazır bulunan atıl durumdaki fonlar, bankalar vasıtasıyla kişi ve/veya kurumlara kredi olarak aktarılır. Bankalar aracılığıyla sağlanan bu krediler, taraflar arasında belirlenen vade ve faiz oranında geri ödenmek üzere bir sözleşme imzalanarak verilir. Kişi ve/veya kurumlar gerek ekonomik koşullar gerekse beklenmeyen durumların ortaya çıkmasından dolayı ödemelerde gecikme yaşamaktadır. Başka bir ifadeyle, kredilerin ifa edilmesi gereken süre içinde ödenmemesi durumu “temerrüt riski” olarak da bilinen durumun ortaya çıkmasına neden olmaktadır. Temerrüt riski, ilgili bankalara ait kredi borçları için sorun oluşturmakla kalmaz, aynı zamanda hem krediyi alan borçlu için çeşitli olumsuz durumların yaşamasına hem de gelecek dönemlerde diğer banka ve/veya banka niteliğinde olan kuruluşlardan kredi alamaması gibi durumlarla karşı karşıya kalmasına neden olur.

BDDK'nın 08 Şubat 2001 tarih 24312 Sayılı Resmi Gazete'de yayımlanan "Bankaların İç Denetim ve Risk Yönetimi Sistemleri Hakkında Yönetmelik'te kredi riski; "Banka müşterisinin yapılan sözleşme gereklerine uymayarak yükümlülüğünü kısmen veya tamamen zamanında yerine getirememesinden dolayı bankanın karşılaştığı durumu ifade eder." şeklinde tanımlanmaktadır (BDDK, 2001). Başka bir deyişle, bankalar tarafından tüzel ve/veya gerçek kişilere vermiş olduğu kredilerin ve satın aldığı menkul kıymetlerin belirlenen vade içerisinde geri ödenmemesi durumunda ortaya çıkan risk olarak ifade edilir (Ertürk, 2010: 66).

1.5.2.2. Piyasa Riski

BDDK'nın 08 Şubat 2001 tarih 24312 Sayılı Resmi Gazete'de yayımlanan "Bankaların İç Denetim ve Risk Yönetimi Sistemleri Hakkında Yönetmelik'te piyasa riski; "Bilanço içi ve bilanço dışı hesaplarda bankalarca tutulan pozisyonlarda finansal piyasadaki dalgalanmalardan kaynaklanan faiz, kur ve hisse senedi fiyat değişmelerine bağlı olarak ortaya çıkan faiz oranı riski, hisse senedi pozisyon riski ve kur riski gibi riskler nedeniyle zarar etme ihtimali " olarak tanımlanmaktadır (BDDK, 2001: 15).

1.5.2.3. Likidite Riski

Bankaların vadeleri geldiğinde mevcut borçlarını ödeme gücü "likidite" olarak ifade edilmektedir. Bankaların vadesi gelen borçlara ilişkin olarak müşterilerine karşı olan yükümlülüklerinin yerine getirilmemesi durumu ise "likidite riski" olarak ifade edilmektedir. Başka bir ifadeyle, likidite riski, gerekli durumlarda nakit toplanamaması riski olarak da ifade edilir. Bankacılık faaliyetlerinin devamlılığının sağlanması için hazırda nakit fon bulundurulması gerekmektedir. Bu fonlar, piyasadaki finansal varlıkların satılması ya da borçlanma yoluyla nakit fon olarak toplanmaktadır. Bankalarca küçük meblağlar hâlinde kısa vadeli olarak toplanır ve bu fonlar, uzun vadeli olarak yüksek tutarlar şeklinde plase edilmektedir. Bankaların her an likidite riski ile karşı karşıya kalmaları olasıdır. Likidite riskini ortaya çıkaran işlemler; kârın azalması, kriz durumları, kısa vadeli varlıklar ve borçlardaki uyumsuzluk sonucu oluşmaktadır (Babuşcu, 2010: 1; Bessis, 2002: 7).

1.5.2.4. Faiz Oranı Riski

BDDK'nın 23 Ekim 2015 tarih 29511 Sayılı Resmi Gazete'de yayımlanan "Bankaların Sermaye Yeterliliğinin Ölçülmesine ve Değerlendirilmesine İlişkin Yönetmelik'te faiz oranı riski; "Bankanın, faiz oranlarındaki değişimlerden etkilenen alım satım hesaplarında yer alan finansal araçlara ilişkin pozisyon durumuna bağlı olarak genel piyasa riski ile spesifik riskten kaynaklı maruz kalabileceği zarar olasılığı " şeklinde tanımlanmaktadır (BDDK, 2015: 2).

Faiz oranı riski, faiz oranlarında oluşan dalgalanmalar sonucunda net faiz geliri ya da faiz gelirlerinde meydana gelen düşüşü ifade etmektedir. Başka bir deyişle, bankalara ait bir ürünün faiz oranlarındaki hareketlenmelerin sonucunda oluşan olumsuz durum veya kârında oluşabilecek muhtemel düşüş olarak ifade edilebilir. Bankaların bilançolarında bulunan kredi ve alacaklar ile tasarruf ya da vadeli mevduatları, faize dayalı olarak maliyetleri ve gelirleri oluşturmaktadır. Bankaların temel görevlerinden biri, olan piyasadaki fon fazlasını toplamak ve bu fonları ihtiyaç sahiplerine ulaştırmada aracılık yapmaktır. Bankalar tarafından toplanan bu fonlar, uzun vadeli olarak plase edilir. Faiz oranı riski, hem borcu alan hem de borcu veren taraf için mevcuttur. Faiz oranları düştüğünde; borç veren alacaklıların vermiş oldukları borca ilişkin olarak gelirlerinde azalma meydana gelirken, borçlular açısından ise faiz oranındaki düşüş, olumlu etki yaparak daha fazla harcama imkânı sunar. Faiz oranları arttığında ise borç veren alacaklıların vermiş oldukları borca ilişkin olarak gelirinde artış meydana gelirken, borçlular açısından ise faiz oranındaki artış, olumsuz etki yaparak daha az harcama yapmalarına neden olmaktadır (Babuşcu, 2010: 1; Bessis, 2002: 4).

1.5.2.5. Operasyonel Risk

Uluslararası Swap ve Türevler Birliği, İngiliz Bankacılar Birliği ve Risk Yönetim Birliği'ne göre Operasyonel risk; "kurum içerisinden kaynaklanan yetersiz veya başarısız iç süreçlerden, sistemler ve insanlardan veya kurum dışından kaynaklanan zarardan oluşan risktir." şeklinde tanımlanmaktadır (Strzelczak, 2007: 12). Potansiyel sistemlerden veya yönetimden kaynaklanan hatalar, kurum için önemli olan personel kaybı, hatalı kontrol ve bilgi kaybı, personellerin görevi kötüye

kullanması gibi çeşitli durumlar kurum içi operasyonel riski oluşturur. Dolandırıcılık ve kabahatler, bilgisayar korsanlığı gibi sistemsel donanımlara yapılan dış saldırılar, terörizm, deprem ve yangın sebebiyle oluşabilecek her türlü doğal afetler sonucu meydana gelebilecek olaylar ise kurum dışı operasyonel riski oluşturmaktadır (Moosa, 2007: 167; Strzelczak, 2007: 12; Van den Brink. 2002:12).

1.5.2.6. Döviz Kuru Riski

Kur riski olarak da bilinen döviz kuru riski, yabancı para kurlarında meydana gelen dalgalanmalardan dolayı ortaya çıkan kâr veya zarar olarak ifade edilmektedir. Başka bir deyişle, döviz piyasalarında yer alan ulusal para birimi karşısında yabancı para birimlerinin değer kazanması ya da değer kaybetmesi sonucu oluşabilecek riskler “döviz kuru riski” olarak adlandırılmaktadır (Ünal ve Altın, 2010: 278).

Döviz kurundan yaşanan dalgalanmalar yalnızca ülke ekonomik faaliyetlerinin işleyişini etkilemekle kalmaz, aynı zamanda piyasadaki firmaların performanslarını da etkiler (Williams, 2018: 2). Döviz kurundaki dalgalanmaların ortaya çıkmasında faiz oranındaki belirsizlik, ekonomik ve siyasal politikalar, yabancı fon akışında oluşan vergi belirsizliği gibi durumlar döviz kurundaki en önemli sebepler arasında yer almaktadır (Sivarajadhanavel ve Chandrakumaramangalam, 2012: 9). Piyasada faal olan birçok firma gibi bankalar da döviz kurundaki değişimlerden etkilenmektedir. Bu değişimler bankaların varlıkları üzerinde etkilidir. Yerli para biriminin yabancı para birimi karşısında değer kaybetmesi durumunda, ilgili bankaların pasifinde döviz fazlasının bulunması ile mevcut borcunu ödemek için daha fazla yerli para kullanması gerekeceğinden zarar, ilgili bankanın aktifinde döviz fazlası olması durumunda ise aynı döviz miktarı için daha fazla yerli para elde edeceğinden dolayı kâr elde eder (Apak ve Atay, 2009: 170-171; Chamberlain ve diğerleri, 1997: 872).

İKİNCİ BÖLÜM

MAKİNE ÖĞRENMESİ TEKNİKLERİ VE YENİDEN ÖRNEKLEME YÖNTEMLERİ

2.1. MAKİNE ÖĞRENMESİ

Yapay Zekâ kapsamında yer alan makine öğrenimi, ses ve konuşma teşhisi, bilgisayar ile algılama, robotik testler, doğal dil işleme gibi farklı uygulamalar için yazılımların geliştirilmesiyle yapay zekânın alt dalı olarak ortaya çıkmıştır. Yapay zekâ ile ilgilenen geliştiriciler, farklı uygulama çalışmalarında, muhtemel girdilere karşılık gelen çıktıların tahmin edilmesinde, doğrudan programa manuel olarak programlanmasının yerine sistemdeki girdi ve çıktı arasındaki ilişkiyi örneklerinin programa tanıtılması ile daha kolay öğreneceğini iddia etmişlerdir (Jordan ve Mitchell, 2015: 255).

Yapay zekâ ile makine öğrenimi arasındaki etkileşim dikkate alındığında, sinir ağları, derin öğrenmenin alt kümesinde yer almaktadır. Makine öğrenmesi; sinir ağları ile derin öğrenmeyi kapsarken, yapay zekâ; makine öğrenimini de kapsamaktadır. Yapay zekâ, belirli kararlar doğrultusunda tahmin elde etmek için verileri işler. Makine öğrenmesinin çeşitli algoritmaları, yapay zekâ ile sadece bu verilerin işlenmesinin sınırlarını aşarak daha ileri aşamalara taşınmasına imkân vermektedir. Başka bir ifadeyle, makine öğrenmesi sayesinde yapay zekânın kapsamı daha da genişlemiştir (SAP, 2021).

Makine öğrenimine, bilgisayar bilimleri başta olmak üzere çeşitli multidisipliner; istatistik, sağlık ve psikoloji gibi bilim dalları ilham kaynağı olmuştur (Muhammad ve Yan, 2015: 946).

Son yıllardaki popülerliği ile çeşitli alanlarda yaygın bir kullanıma sahip olan Makine Öğrenmesi, geçmişteki verilerden elde edilen bilgi ve deneyimlerin kullanılmasıyla verilerden öğrenmenin sağlanması ve gelecekte karşılaşılabilecek problemlerin çözülmesine yardımcı olmak amacıyla kullanılan, yapay zekânın bir çalışma alanı olarak ifade edilmektedir. 1959'da Arthur Samuel tarafından tanıtılmış olan Makine Öğrenmesi kavramı, 1997'de Tom Michael tarafından kapsamı daha da genişletilerek popüleritesi artmıştır.

Makine öğrenmesi, verilerin akıllı bir eyleme dönüştürülmesi sürecinde bilgisayar algoritmalarının geliştirmesi ile ilgilenen bir çalışma alanı olarak adlandırılabilir. Bu alan, güncel verilerin, istatistiksel metotların ve hesaplama gücünün etkililiğinden yararlanması ile eşit zamanlı olarak gelişiminin sağlandığı bir ortamda ortaya çıkmıştır. Verilerdeki hacmin artması, işlem gücü gereksinimini artırdı ve bu da büyük hacimli veri setlerinin analizi için istatistiksel metotların geliştirilmesine katkı sağlamıştır (Lantz, 2013: 7).

Makine öğrenmesi, hem sınıflandırma hem de regresyon yapılması için farklı algoritmaların eğitilmesi ile öğrenme işlemi gerçekleştirir. Giderek artmakta olan veri hacmi, makine öğrenmesinin çalışma sistemi üzerinde olumsuz bir etki oluşturmamaktadır. Aksine, veri sayısının artması ile birlikte tahminleme performansının doğruluğu artar ve daha etkin sonuçlarla genelleştirme yeteneği kazanmış olur (IBM,2020).

Makine öğreniminin temel amacı, mevcut veriler ile algoritmaya verilerin öğretilmesi ve bu öğretilen verilerden genelleme yapmasının beklenmesidir. Aynı zamanda, tahminleme yapılırken hatalarının minimum düzeyde tutulması amaçlanmaktadır.

Makine öğrenimi, veri sayısındaki artış ve gelişen teknoloji ile birlikte birçok sektörde yaygın bir kullanıma sahiptir. Bu sektörlerdeki kullanımının en önemli nedenlerinden biri, sistemdeki davranışları kolayca öğrenme ve uygulama yeteneğine sahip olmasıdır. Sistemdeki bu yetenekler göz önünde bulundurulduğunda, istenilen sonuçların doğru bir şekilde ortaya çıkması için yüksek hacimli eğitim verileri ile makine öğrenmesi bazı durumlarda alanındaki uzman kişilerden daha başarılı performans gerçekleştirilebilir. Aynı zamanda, makine öğrenmesinin kullanımı, zaman, maliyet ve hayatı kolaylaştırıcı olması açısından oldukça avantaj sağlamaktadır. Makine öğrenmesinin işlevselliği sayesinde, günümüzde çeşitli sektörler, iş geliştirme ve strateji yönetimini gerçekleştirmek için makine öğrenmesi algoritmalarını etkin bir biçimde kullanmaktadır. Makine öğreniminin günümüzde birçok örneği bulunmaktadır. Bunlar, Google tarafından gerçekleştirilen ve SEO olarak bilinen arama motoru optimizasyonu başta olmak üzere, günlük olarak sıklıkla kullanılan Instagram'da paylaşılan fotoğraf etiketleme, YouTube, Netflix, e-postalara gelen istenmeyen mesajların filtrelenmesi, kanser teşhisi, kredi kartı dolandırıcılığı,

envanter ve sipariş yöntemi gibi çeşitli alanlarda yaygın olarak kullanılmaktadır (Mitchell ve Brynjolfsson, 2017a: 290) ve (Brynjolfsson ve Mitchell, 2017b: 1531).

Dördüncü Sanayi Devrimi'nin (Endüstri 4.0) mevcut çağında, makine öğrenmesi gerek öğrenme yetenekleri gerekse akıllı kararlar vermesi ile farklı alanlarda popüler hâle gelmiştir. Makine öğreniminin en popüler uygulamaları aşağıdaki gibidir (Sarker, 2021: 15-17).

- Siber güvenlik ve tehdit istihbaratı uygulamaları,
- Nesnelerin İnterneti (IoT) ve akıllı şehirler uygulamaları,
- Trafik ve ulaşım ağı tahmininde,
- Ekonomi ve bankacılık uygulamaları,
- Hastalıkların teşhisi ve COVID-19 salgını,
- E-ticaret ve ürün önerilerinin kategorilendirme,
- Doğal dil işleme (NLP) ve duygu analizi,
- Görüntü, konuşma ve örüntü tanıma tahmininde,
- Robotik ve ileri mühendislik uygulamaları,
- Kullanıcı davranışı analizi ve bağlama duyarlı akıllı telefon gibi birçok alanda kullanılmaktadır.

Makine öğrenmesi genel olarak iki aşamadan oluşmaktadır. Algoritmanın, geçmişte elde edilen bilgi ve deneyimler ile güncel verilerden öğrendiği kümeye “eğitim seti” denir. Eğitim veri kümesi ile eğitilen algoritmanın öğrenme başarısının test edilmesi amacıyla kullanılan kümeye “test seti” denir. Genel olarak, çalışmalarda eğitim veri seti %80 iken, test veri seti %20 olarak kullanılmaktadır.

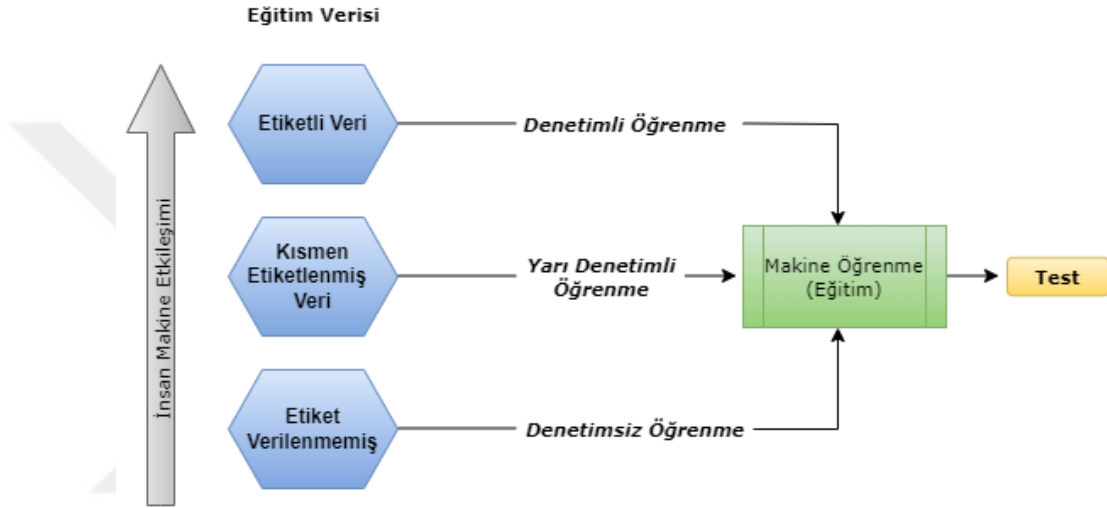
Veri kümesi, eğitim ve test seti olmak üzere iki parçaya bölme işlemi gerçekleştirilirken, eğitim ve test seti üzerinde aşağıdaki adımlar uygulanır (Buczak ve Guven, 2015: 1154).

- Eğitim kümesinde yer alan sınıflar ve bu sınıflara ait öznitelikler tanımlanır.
- Sınıflandırma için gereksiz özellikler çıkarılarak boyutta indirgeme yapılabilir, bu yolla yalnızca gerekli özelliklerin bulunduğu bir alt küme oluşturulur.
- Eğitim veri seti ile modelin öğrenmesi işlemi gerçekleştirilir.

- Bilinmeyen veya modele tanıtılmayan verilerin sınıflandırılmasında, eğitilmiş model kullanılır.

Makine öğrenmesi, eğitim veri kümesindeki verilerin niteliklerine göre kategorilere ayrılabilir. Eğitim verisi ile makine öğrenmesi türlerine göre ilişkilendirilmesi Şekil 5’te gösterildiği gibidir.

Şekil 5: Eğitim Verilerinin Etiket Niteliklerine Göre Kategorize Edilmesi



Kaynak: El Naqa ve diğerleri, 2015: 7.

Makine öğrenmesi algoritmalarının uygulama adımları aşağıdaki gibidir (Kunal, 2015; Lantz, 2013: 17):

Veri Toplama: Veriler metin dosyaları ve elektronik ortamdaki tablolara yazılmış olsun veya yapılandırılmış sorgu dili (SQL) veri tabanında bulunan ham veriler olsun, veri toplama adımı geçmişteki bilgi ve deneyimlere ait veriler ile gelecekte karşılaşılabilecek durumdaki verilerin zeminini oluşturmaktadır. Bu verilerin donanımına bağlı olarak algoritmadan beklenen öğrenme düzeyi de doğru orantılı olacaktır.

Veri Ön İşleme ve Hazırlama: Herhangi bir projenin kalitesi, genel olarak kullanılan verilere bağlıdır. Veri kümesi içerisinde eksik ve aykırı değerler (outlier) ile

ilgili problemlerin tespiti ve çözümünün gerçekleştirilmesinin yapıldığı aşamadır. Makine öğrenimi için önemli olan bu aşamada, dışarıdan sisteme çok fazla müdahale söz konusudur. Aynı zamanda, verilerin keşfi ve geliştirilmesi için en çok emek gerektiren aşamadır.

Modelin Eğitilmesi: Verilerin hazırlanmasının tamamlandığı ve analiz için uygun olduğu aşamadır. Bu aşamada, en doğru algoritma ile verilerin öğrenmesi sağlanır. Analiz için uygun veriler, eğitim ve test verisi olarak iki parçaya ayrılır. Genel olarak, veri kümesinin %80'i eğitim verisi ve %20'lik bölümü test verisi olarak kullanılmaktadır. Algoritmanın eğitim verileri ile eğitildiği ve modelin oluşturulduğu aşamadır.

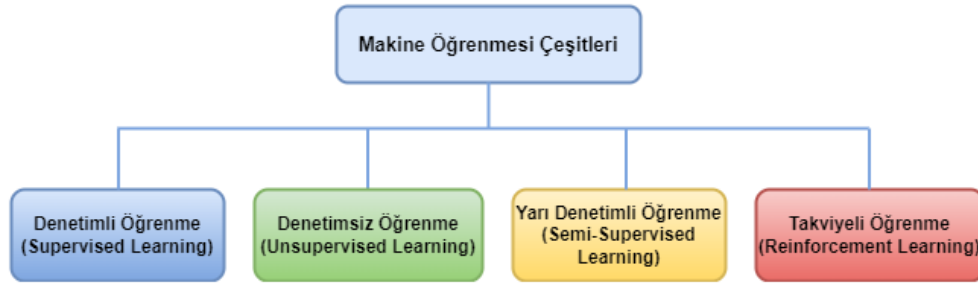
Model Performans Değerlendirmesi: Makine öğrenmesi modeli, algoritmaların deneyimlerine bağlı olarak ne düzeyde öğrendiğinin tespit edilmesi oldukça önemlidir. Bu aşamada, uygun algoritmanın seçimi gerçekleştirilir. Model eğitilirken kullanılan verilerin dışında kalan veriler ile model performansı test edilir.

Performansın İyileştirilmesi: Makine öğrenmesi modelinin performansının değerlendirilmesi aşamasında, istenilen başarının sağlanamadığı durumda model için yeni stratejiler geliştirilebilir veya tamamen başka modele geçiş yapılabilir. Bu nedenle, veri ön işleme ve hazırlama (2.Aşama) süreci model için önem arz etmektedir.

2.1.1. Makine Öğrenmesi Algoritması Türleri

Makine öğrenmesi algoritmaları; denetimli makine öğrenmesi, denetimsiz makine öğrenmesi, yarı denetimli makine öğrenmesi ve takviyeli (pekiştirmeli) makine öğrenmesi olmak üzere dört başlık altında ele alınmıştır.

Şekil 6: Makine Öğrenmesi Çeşitleri



Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

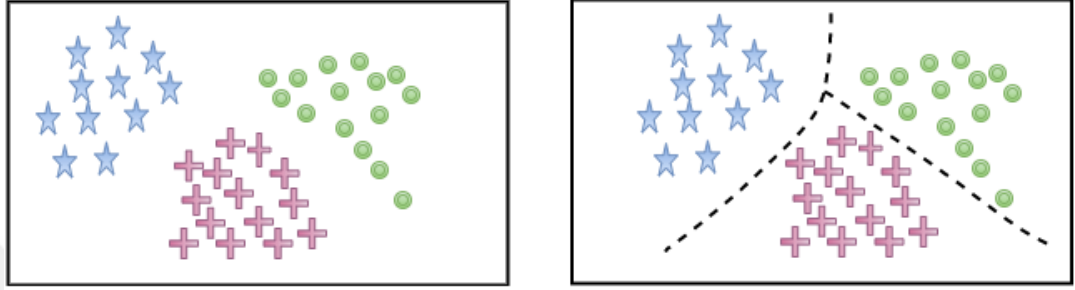
2.1.1.1. Denetimli Makine Öğrenmesi

Veri kümesi içerisinde yer alan sınıfların değerlerinin belli olduğu, etiketlenmiş veya sınıflandırılmış verilerden oluşan öğrenme türü; gözetimli veya denetimli makine öğrenmesi olarak adlandırılmaktadır. Başka bir deyişle, makine öğrenme teknikleri etiketlenmiş eğitim verileri aracılığıyla öğrenme işlemi gerçekleştirilir ve bu verilere bağlı olarak gelecekteki olayların tahmini için algoritmalar tarafından çıktının tahmin edildiği makine öğrenmesi çeşidi olarak da ifade edilmektedir. Burada öğrenme algoritması, modele sunulan ve daha önce modele tanıtılmayan veriler hakkında çıkarım yaparak etkili bir tahmin yapmayı amaçlamaktadır. Ayrıca, girdi (x) değişkeni olarak bilinen bağımsız değişkeni, bir çıktı (y) değişkeni olarak bilinen bağımlı değişkeni ile eşleştirmesi beklenmektedir. Bu eşleştirmeyle, verilen girdiler üzerinden istenilen çıktının üretilmediği durumlarda algoritmaya müdahale edilerek tahminin revize edilmesi mümkündür. Dolayısıyla, girdi değişkenine karşılık gelen bir çıktı değeri olduğu için denetimli makine öğrenmesi, sınıflandırma veya regresyon problemleri için doğrudan uygulanması mümkündür (Bishop ve Nasrabadi, 2006: 3; Han ve diğerleri, 2011: 445; Javatpoint, 2020; James ve diğerleri, 2013: 26; Jordan ve Mitchell, 2015: 257; Loukas, 2020; Mohri ve diğerleri, 2018: 6). Denetimli öğrenme problemleri iki tür olarak ele alınır:

Sınıflandırma: Amaç, kategorik öznitelikteki verilerin tahmin edilmesidir. Örneğin; e-postalara gelen mesajlar; “spam (istenmeyen)” veya “spam değil”, Kişiler; “kredi alabilir” veya “kredi alamaz”, kişi; “hasta” veya “hasta değil”, görüntü sınıflandırma; “beyaz” veya “siyah” gibi kategorilere sahip özellikteki verilerden oluşmaktadır.

Regresyon: Amaç, sayısal özellikteki verilerin tahmin edilmesidir. Örneğin; Konut fiyatları, Emtia (altın, gümüş, petrol v.b.) fiyatları gibi sayısal özellikteki verilerin tahmin edilmesinde kullanılmaktadır.

Şekil 7: Denetimli Makine Öğrenmesi: Sınıflandırma Örneği



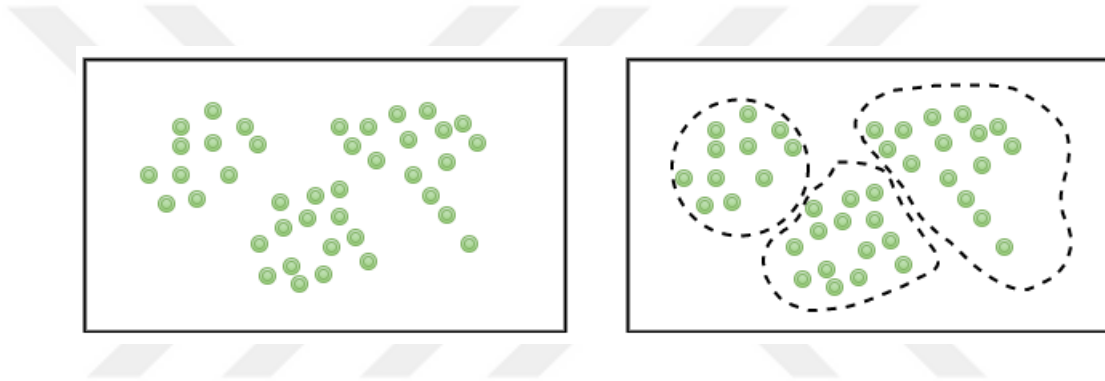
Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

2.1.1.2. Denetimsiz Makine Öğrenmesi

Veri kümesi içerisinde yer alan sınıfların değerlerinin belirsiz olduğu, etiketlenmemiş veya sınıflandırılmamış verilerden oluşan öğrenme türü; gözetimsiz veya denetimsiz makine öğrenmesi olarak adlandırılmaktadır. Denetimsiz makine öğrenmesi, sadece sınıflandırılmamış veya etiketlenmemiş eğitim verilerini girdi olarak alır ve görünürde olmayan bütün noktaların tahmini için kullanılmaktadır. Bu yöntemde, yaygın olarak etiketsiz bir örnek kullanıldığı için performansın nicelik yönünden değerlendirilmesi pek mümkün değildir (Mohri ve diğerleri, 2018: 6). Başka bir ifadeyle, denetimsiz öğrenme, etiketsiz verilerle ilişkili gizli örüntüyü ortaya çıkarmayı hedefler. Burada, veri yapısı tarafından doğru çıktının tahmin edilmesi beklenemez, çünkü veri kümesi bağımlı değişken (y) içermemektedir ve bu nedenle, etiketsiz girdi değerlerine karşı bir çıktı değerleri bilinmediği için sınıflandırma veya regresyon problemleri için direkt olarak uygulanması mümkün değildir. Bunun yerine verilerin keşfine odaklanır ve etiketsiz verilerden gizli örüntülerin belirlenmesi için veri setlerinden çıkarımlar yapılabilir (Loukas, 2020). Etiketlenmemiş eğitim girdileri ile veri yapılarının anlamlandırılmasında; öznitelikler arasındaki ilişkilerin belirlenmesi veya kümeleme yöntemleri kullanılmaktadır. Birbirlerine benzerliği diğerlerine göre daha fazla

olanların bir grupta bulunacağı ve diğer bir gruptaki nesnelerle benzerliğinin az olduğu veya hiç olmayanların bir grupta yer alacak şekilde gruplandırılması kümeleme olarak ifade edilmektedir. Bu bağlamda, gerçekleştirilen farklı gruplara ayırma işlemi kümeleme algoritması ile gerçekleştirilmektedir. Bu algoritmalar, hiyerarşik (hierarchical) ve bölünme (partitional) gibi algoritmalarla kategorize edilebilir. Kümeleme başlıca; iş zekâsı, tıp, web araması ve siber güvenlik gibi çeşitli alanlarda kullanılmaktadır (Béjar Alonso, 2013: 3; Han ve diğerleri, 2011: 444; Javatpoint, 2020; Madhulatha, 2012: 719).

Şekil 8: Denetimsiz Makine Öğrenmesi: Kümeleme Örneği



Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

2.1.1.3. Yarı Denetimli Makine Öğrenmesi

Yarı denetimli makine öğrenmesi, hem denetimli öğrenme hem de denetimsiz öğrenme yöntemlerinin bir araya getirilmesi ile oluşan bir makine öğrenmesi yöntemidir. Bu yöntemdeki eğitim verileri, etiketlenmiş ve etiketlenmemiş verilerden oluşmaktadır. Eğitim veri kümesinde bulunan etiketli veriler, etiketlenmemiş verilere göre görece daha az sayıda bulunmaktadır. Yarı denetimli makine öğrenmesinin işleyişi, etiketlenmemiş verilerle ilgili bir çıkarımda bulunmak amacıyla, girdi değişkeni olarak etiketlenmiş verilerin kullanılması sonucu gerçekleştirilmektedir. Böylece, etiketlenmemiş verilerin etiketlenmesi sağlanmakta ve tahmin performansının başarısının artırılması hedeflenmektedir. Bu yöntem, görüntü sınıflandırma problemleri için yaygın olarak kullanılmaktadır. Ayrıca, denetimli öğrenme modelleri ile kıyaslandığında, az sayıda etiketli verinin kullanılması

algoritmanın kurulum hızını artırmakta, maliyet açısından uygunluk sağlamaktadır (Monkeylearn, 2021).

2.1.1.4. Pekiştirmeli Makine Öğrenmesi

Pekiştirmeli makine öğrenmesi, istenilen davranışların sonucu ödüllendirilmeye veya istenilmeyen davranışların gerçekleştirilmesi sonucunda cezalandırılmaya dayalı olan bir makine öğrenmesi yöntemidir. Bu yöntemde, takviyeli öğrenme ajanı olarak bilinen öğrenen makine, sadece çevreyi algılayabilen ve yorumlayabilen bir özelliğe sahip değil, aynı zamanda harekete geçebilen ve sınama yoluyla öğrenme özelliğine de sahiptir. Bu yöntemde, ajanın güdümlenmesi için gerçekleşmesi istenilen eylemler pozitif ve gerçekleşmesi istenilmeyen eylemlere ise negatif değerler atanmaktadır. Burada, hedeflenen eylemlerin uygulanması ile karşılaşılan durumlardan alınan tepkilere göre geribildirim; ödül veya ceza almaktadır. Ajanın amacı, en iyi çözümün gerçekleştirilmesi amacıyla toplam ödüllerin maksimize edilmesini amaçlamaktadır. Pekiştirmeli makine öğrenmesi başlıca; oyun uygulamaları, robotik alanlar, sinyal kontrolü ve sürücüsüz araçlar gibi alanlarda yaygın olarak kullanılmaktadır (Carew, 2020).

2.2. TEKİL MAKİNE ÖĞRENMESİ TEKNİKLERİ

2.2.1. K-En Yakın Komşu

Gözetimli makine öğrenmesi algoritması olan k -En Yakın Komşu (k -NN) algoritması, en temel ve kolay uygulanabilen sınıflandırma teknikleri arasında yer almaktadır. Verilerin dağılımları hakkında bilgi yetersizliğinin olması durumunda sınıflandırma için yaygın olarak kullanılan, parametrik olmayan bir yöntemdir (Peterson, 2009). Eğitim veri kümelerindeki bağımsız komşu girdilerinin (x) yakınlığının ve bunlara karşılık gelen çıktı (y) değerlerini, veri kümesindeki çıktı değerlerini doğru tahmin etmek için doğrulama veri seti kullanılmaktadır (Fomby, 2008: 1). Diğer bir deyişle, örnek veri kümesindeki sınıfların önceden belli olduğu verileri kullanarak, veri kümelerine sonradan dahil edilecek her örnek için var olan verilere göre mevcut uzaklığı hesaplanır. k -NN algoritmasında mesafe (uzaklık)

ölçümünün doğru seçilmesi sonuçlar açısından önemlidir. Bu doğrultuda Manhattan Mesafe Ölçüsü (2.1), Minkowski Mesafe Ölçüsü (2.2) gibi çeşitli ölçüm yöntemleri kullanılsa da en yaygın olarak Öklid Mesafe Ölçüsü (2.3) kullanılmaktadır (Duda ve diğerleri, 1973; Weinberger ve Lawrence, 2008). Bu bağlamda, iki nokta arasındaki Öklid mesafesi tanımlanması gerekmektedir. Buna göre iki vektör $P (p_1, p_2, \dots, p_n)$ ve $Q (q_1, q_2, \dots, q_n)$ olarak tanımlanmıştır. P ve Q vektörü arasındaki Öklid uzaklığı olarak hesaplanmakta ve gösterimi Eşitlik (2.3) 'teki gibidir (France ve diğerleri, 2012).

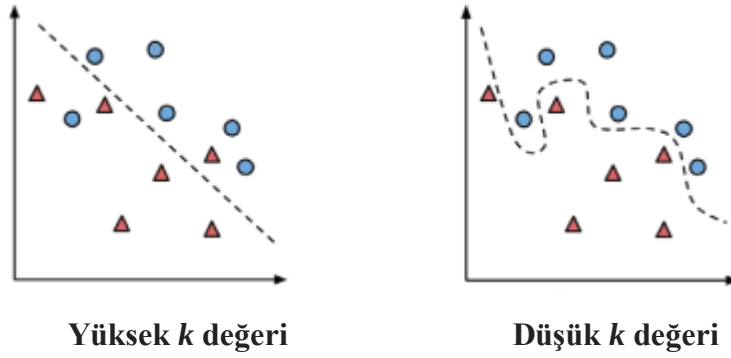
$$d(P, Q)_{\text{Manhattan}} = \sum_{i=1}^n |p_i - q_i| \quad (2.1)$$

$$d(P, Q)_{\text{Minkowski}} = \left(\sum_{i=1}^n |p_i - q_i|^a \right)^{1/a} \quad (2.2)$$

$$d(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (2.3)$$

k -NN'nin uygulanması aşamasında k için en uygun değer seçilmesi gerekmektedir. Seçilen k sayıdaki komşuların yakınlığına göre uygun sınıfa atanmaktadır. k -NN'deki k değerleri kendilerine en yakın komşuların sayılarını ifade etmektedir. k değeri birbirine yakın iki komşu arasındaki karmaşıklık düzeyini minimum hâle getirebilmek için genel olarak tek sayı değeri olarak tercih edilmektedir. Örneğin; bir kredi risk sınıflandırması problemindeki kredinin durumunun tespit edilmesi amacıyla komşu sayısının 4 olarak seçilmesi, seçilen komşuların iki iyi kredi ve iki kötü krediden oluşmasına sebep olabilir. Bu durumda, kredinin hangi komşuya dahil edileceği kesin olarak bilinemediği için belirsiz devam edecektir. Komşu sayısı (k) değerlerinin rassal olarak bir değer verilmesi birçok probleme yol açmaktadır. Seçilen k değerinin sınıflandırma performansının başarısı üzerindeki etkisi bulunmaktadır.

Şekil 9: K -En Yakın Komşu Örneği



Kaynak: Lantz, 2013: 71.

k değerinin çok yüksek bir değer seçilmesi durumunda, gürültülü verilerin tahmin veya varyansı üzerindeki etkisini azaltırken, model yanlılığının (sapmanın) artmasına sebep olmaktadır (Lantz, 2013: 71). Model sapmasının yüksek olması, modelin eğitim verileri üzerindeki uyumsuzluğunun artmasına ve verilere uygun trendi takip edememesine neden olmaktadır. Aynı zamanda k değerinin çok düşük seçilmesi durumunda, model varyansının artmasına buna bağlı olarak da veriler üzerinde aşırı öğrenme durumu gerçekleşir. Bunun sonucu olarak veriler üzerindeki küçük değişikliklerin hata oranı üzerindeki etkisi çok yüksek düzeyde olacaktır.

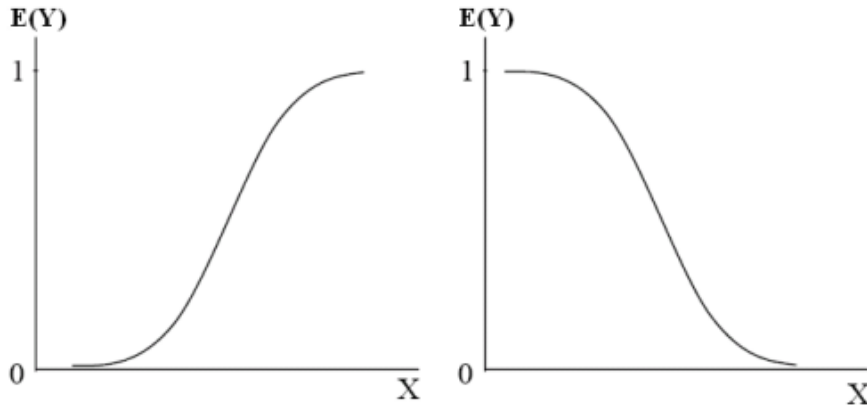
2.2.2. Lojistik Regresyon

Lojistik regresyon veya logit modeller olarak bilinen, bağımsız değişkenlerdeki değerlere bağlı olarak bağımlı (yanıt) değişkenin ikili değerden birine sahip olduğu bir modeldir. Diğer bir deyişle, yanıt değişkeninin sürekli bir değişken olmadığı, yalnızca iki olası sonucu bulunan ve bundan dolayı 0-1 değerlerini alabilen bir değişken olduğu durumlarda kullanılmaktadır (Le ve Eberly, 2016: 351). Lojistik regresyon modelleri bağımlı (yanıt) değişkendeki değerlere bağlı olarak üç gruba ayrılmaktadır. İlki, bağımlı değişkendeki değerler, ikili (binary veya 0-1) kategorik değişkenden oluşmaktadır. Bunlar, İkili Lojistik Regresyon Modeli olarak adlandırılmaktadır. Örneğin; bir kişinin bankaya kredi borcu olabilir veya olmayabilir, bir öğrencinin X dersinden geçmiş olabilir veya olmayabilir, bir kişinin medeni durumu bekar olabilir

veya olmayabilir, bir bireyin beslenme durumu düzenli olabilir veya olmayabilir. İkincisi, bağımlı değişkendeki değerler, ikiden daha fazla değere sahipse Çoklu Lojistik Regresyon Modeli olarak adlandırılmaktadır. Örneğin; 1.Evre, 2.Evre ve 3.Evre gibi kanser evrelerinin sınıflandırılmasında kullanılmaktadır. Son olarak, bağımlı değişkendeki değerler, ikiden fazla değere sahip ve sıralama ölçeği kullanılıyorsa Sıralı Lojistik Regresyon Modeli olarak adlandırılmaktadır. Örneğin; Hiç katılmıyorum, kısmen katılmıyorum, Ne katılıyorum ne de katılmıyorum, Kısmen katılıyorum ve Tamamen katılıyorum gibi anketlerde kullanılan Likert Ölçeği Sıralı Lojistik Regresyon için sıklıkla tercih edilmektedir.

Lojistik regresyon analizi, hem kuramsal hem de ampirik çalışmalar sonucunda yanıt değişkeninin iki sonuçlu olduğunda cevap fonksiyonunun $\left(\frac{p}{1-p}\right)$ şeklinin S veya ters S şeklinde olacağı gösterilmiştir. Yanıt değişkeni, Şekil 10'da gösterildiği gibi bitiş noktaları dışında yaklaşık olarak doğrusaldır. Ayrıca, cevap fonksiyonları 0 ile 1 değerlerinde X ve Y eksenlerine asimptottur (Hosmer -Lemeshow, 1980: 1043-1069).

Şekil 10: S ve Ters S Şeklindeki Olasılık Fonksiyon Grafikleri



Kaynak: Hosmer -Lemeshow, 1980: 1043-1069.

Lojistik fonksiyonun 0 ve 1 sayıları arasında değişen değerlere sahip olması sıklıkla tercih edilmesindeki en önemli faktörlerden biridir. 0' ın altında veya 1' in üstünde bir değer alması söz konusu değildir. Diğer bir deyişle, negatif bir değere sahip

olmaması lojistik regresyonun tercih edilmesindeki ikinci en önemli faktördür (Hosmer – Lemeshow, 1980: 1043-1069).

Matematiksel olarak lojistik regresyon analizi olasılık, odds ve odds'un logirtmasına dayanmaktadır. Lojistik regresyon modelindeki odds 2.4'te, bir olayın meydana gelmesi olasılığının, meydana gelmemesi olasılığına bölünmesiyle elde edilmektedir (Mertler ve Vannatta, 2005).

$$\text{Odds} = \frac{P}{1-P} \quad (2.4)$$

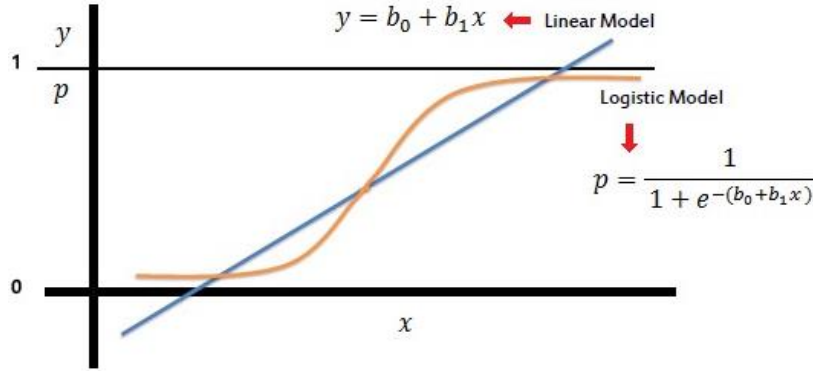
İki Odds değerinin birbirine oranı Odds Ratio (OR) olarak ifade edilmektedir. Odds oranı, Üstel lojistik katsayıları $\text{Exp}(\beta)$ ile sembolize edilir. Ayrıca, Odds oranı asimetrik formdadır. Odds oranının doğal logaritması alınarak asimetrik formdan simetrik hâle dönüştürülerek logit fonksiyonu Eşitlik (2.5)'teki gibi elde edilmektedir (Mertler ve Vannatta, 2005 ; Kleinbaum ve Klein, 2002: 17).

$$\text{logit } P = \ln_e \left[\frac{P}{1-P} \right] \quad (2.5)$$

Odds Ratio değeri 0 ile $+\infty$ değerlerini alırken, logit, parametrelerinde doğrusal, sürekli olabilir ve x aralığına bağlı olarak $-\infty$ ile $+\infty$ arasında değerler alabilir (Hosmer ve diğerleri, 2013: 7). Lojistik regresyon modeli, doğrusal regresyon modeli ile birçok açıdan benzer özellikler taşımaktadır. En basit hâliyle ifade edilirse, sadece bir açıklayıcı X_1 olduğunda, Y'nin olasılığının tahmin edildiği lojistik regresyon modeli Eşitlik (2.6)'da elde edilmekte ve grafiksel gösterimi Şekil 11'de gösterildiği gibidir (Osibanjo ve diğerleri, 2015: 29).

$$P(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}} = \frac{e^{(\beta_0+\beta_1 X_1)}}{1+e^{(\beta_0+\beta_1 X_1)}} \quad (2.6)$$

Şekil 11: Doğrusal Regresyon Modeli ve Lojistik Regresyon Modeli



Kaynak: Sayad, 2019.

Doğrusal olasılık modelinden elde edilen kesikli olmayan (sürekli) değerleri, $-\infty$ 'a yaklaşması durumunda 0 değerine yakınsarken, $+\infty$ 'a yaklaşması durumunda ise 1 değerine yakınsamaktadır (Kleinbaum ve Klein, 2002: 6-7).

Eşitlik (2.6)'da verilen $P(Y)$, Y'nin meydana gelme olasılığını gösterirken, e doğal logaritmayı ifade etmekte ve $\beta_0 + \beta_1 X_1$ (basit doğrusal regresyon) denkleminin doğrusal kombinasyonunu oluşturmaktadır. Doğrusal regresyon denklemine birden fazla açıklayıcı değişken içerek şekilde genişleterek Eşitlik (2.7)'deki gibi çoklu lojistik regresyon modeli elde edilmektedir.

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i)}} = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i)}} \quad (2.7)$$

Doğrusal regresyon modelinde parametre tahmininde en küçük kareler (EKK) yöntemi kullanılırken, lojistik regresyon modelinde ise en yüksek olabilirlik (Maximum Likelihood) yöntemi kullanılmaktadır. Birden fazla bağımsız değişkenin bulunduğu regresyon modeli, bağımsız değişkene ait gerçek değerler ile tahmin değerleri arasındaki farkın kareler toplamını minimize etmek için EKK yöntemi kullanılmaktadır. Lojistik regresyon modeli, doğrusal regresyon modelinin aksine sapmalarını minimize etmek yerine, olayın gerçekleşme olasılığını maksimize yapmakla ilgilenmektedir (Hair ve ark., 1998). Lojistik regresyon analizindeki her bir

açıklayıcı değişkenin anlamlılık düzeylerini test etmek için Wald test istatistiği kullanılmaktadır (Fujimoto, 2003).

2.2.3. Naïve Bayes

Naive Bayes, denetimli makine öğrenimi algoritmalarında yaygın olarak kullanılan bir sınıflandırma algoritmasıdır. Bu algoritma, Bayes teoremine dayalı olasılık tabanlı bir algoritmadır. Başka bir ifadeyle, bir olayın gerçekleştiği biliniyorken, diğer bir olayın meydana gelmesi koşuluna bağlı olarak olayların gerçekleşme olasılıklarının hesaplandığı bir yöntemdir. Naive Bayes sınıflandırıcıları, basit olmasına rağmen, yüksek boyutlu veri setlerinde kolay uygulanabilir ve karmaşık sınıflandırma tekniklerinde diğer alternatif sınıflandırıcılara göre genel olarak daha iyi performans gösterebilmektedir (Domingos ve M. Pazzani, 1997: 121). Uygulanmasının kolay olmasından dolayı birçok alanda sıklıkla tercih edilmektedir. Örneğin; e-postalara gelen spam (istenmeyen) mesajların tespit edilmesi ve e-postaların tasnif edilmesi (ekonomi, reklam v.b.) gibi sağlık alanında ise, gözlenen semptomlar doğrultusunda bireyin hasta olup olmadığının tanısında genellikle kullanılmaktadır (Lantz, 2015: 90). Bayes sınıflandırması, temel bir olasılık modeline dayanır ve modele ilişkin belirsizliklerin, sonuçların koşullu olasılık değerleri hesaplar (Vijaykumar ve Vikramkumar, 2014: 1). Naive Bayes, sınıflandırılmasında, bilinmeyen veri kümelerinin sınıflarının tahmininde Bayes olasılık teorisini kullanmaktadır. Bayes teoremi matematiksel olarak şu şekilde ifade edilir:

X , eğitim veri kümesindeki gözlemleri temsil etmektedir.

$$X = (x_1, x_2, \dots, x_n)$$

C_i , yanıt değişkenine ait i adet tahmin sınıfını temsil etmektedir.

$P(X | C_i)$: i .sınıftan bir örneğin X olma olasılığını temsil eder.

$P(C_i | X)$: X olan bir örneğin, i .sınıftan olma olasılığı (nihai veya ardıl) temsil eder.

$P(C_i)$: yanıt değişkeninin i . Sınıfın ilk olasılığını (önsel) temsil eder.

$P(X)$: Herhangi bir örneğin X olma olasılığını ifade etmektedir.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (2.8)$$

Tüm sınıflar için $P(X)$ sabittir. Sadece $P(X/C_i)P(C_i)$ en yüksek değere sahip olması amaçlanmaktadır. Sınıfın önceki olasılıkları değerleri bilinmiyorsa, bütün sınıfların eşit olasılığa sahip olduğu varsayılmaktadır. Bu varsayım doğrultusunda, $P(C_1) = P(C_2) = \dots = P(C_m)$ olmaktadır. Olasılık değerleri eşit olduğu için $P(X/C_i)$ maksimize edilmiş olacaktır. Aksi halde $P(X/C_i)P(C_i)$ en yüksek değere sahip olacaktır (Han ve Kamber, 2011: 351-352). Naive Bayes, Bayes sınıflandırıcıları tüm tahmin ediciler arasında ilişki olmadığı varsayılır. Bu nedenle tüm sınıftaki yer alan her bir değişken eşit düzeyde katkı sağlamaktadır. Bu varsayımlardan dolayı “Naive” adını almaktadır. Tüm değişkenler için toplu olasılık hesaplamak yerine her bir değişkenin bağımsız olduğu varsayılarak tek tek olasılık değerleri hesaplanır. Naive Bayes algoritmasında kullanılan eğitim verilerine ait her bir gözlem değerleri, kesikli değere sahip olmalıdır. Bu yöntem sürekli verilerle çalışmamaktadır. Yanıt ya da açıklayıcı değişkenlerden kesikli özelliğe sahip olmayan değerler, kategorik veri tipine dönüştürülmelidir. Naive Bayes çalışma adımları: eğitim veri kümesindeki verilen sınıf etiketleri için önsel (denemeler yapılmadan önce gerçekleşen) olasılık değerleri hesaplanır. Her sınıfın her bir özelliği ile olabilirlik olasılığının değerleri bulunur. Bu değerler, koşullu olasılığın bulunmasında kullanılarak nihai (sonsal) olasılık değerleri elde edilir. Sınıf tahmininde ise, girdinin daha yüksek olasılıklı sınıfa ait olduğu göz önünde bulundurulduğunda, sınıflardan hangisinin daha yüksek olasılığa sahip olduğu belirlenir (Navlani, 2018; Wang ve diğerleri, 2020: 143-144).

Naive Bayes matematiksel olarak şu şekilde ifade edilir:

D eğitim veri kümesinde bulunan her bir örnek için n boyutlu uzayda tanımlı olduğunu varsayalım. $X = (x_1, x_2, \dots, x_n)$ ve veri setinde $C_1, C_2, \dots, C_m$ m adet sınıftan oluştuğunu varsayalım. Sınıflandırma Eşitlik (2.8)’de verilen eşitliğin $P(C_i | X)$ nihai olasılığın maksimize edilmesi amaçlanmaktadır. Bu eşitlikte, herhangi bir örneğin X olma olasılığı $P(X)$ sabit olduğuna göre, $P(C_i | X) = P(X / C_i)P(C_i)$ şeklinde basitleştirilir. Birden çok özniteliği bulunan veri setlerinde, hesaplama açısından $P(X / C_i)$, oldukça yüksek maliyetli olmaktadır. Bu hesaplama maliyetlerini düşürebilmek için tüm öznitelikler koşullu olarak birbirinden bağımsız olduğu varsayılmaktadır.

Böylece, $P(X / C_i)$ aşağıdaki Eşitlik (2.9)'da gösterildiği gibi ayrı ayrı çarpım şeklinde yazılabilir (Han ve Kamber, 2011: 351-352).

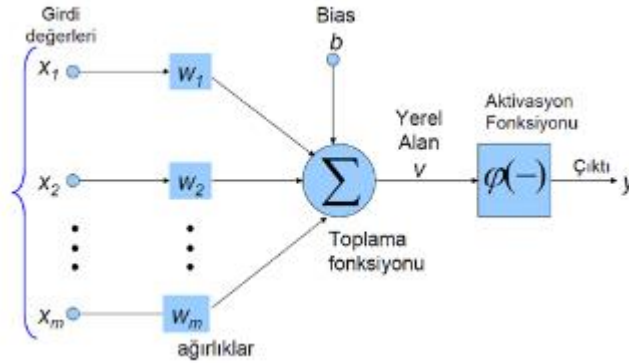
$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^n P(x_k|C_i) \\ &= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i) \end{aligned} \quad (2.9)$$

Naive Bayes, eğitim veri seti için yüksek boyutta veriye gereksinim duymaz. Eğitilen örnek sayısı arttıkça daha iyi çalışmaktadır. Basit olmasına rağmen hızlı ve çok etkilidir. Veri kümesindeki gürültülü veri ve eksik veriler olsa da iyi performans göstermesi gibi avantajları bulunmaktadır. Değişkenlerin öznitelikleri birbirinden bağımsız olduğu varsayıldığı için değişkenler arasındaki ilişkilerin modellenmesi mümkün değildir. Sürekli değişkenlerle çalışmadığı için bu özelliğe sahip değişkenler kategorik hâle dönüştürülerek kullanılmaktadır. Ayrıca, tahmini olasılıklar, tahmin edilen sınıflardan daha az güvenilir olması gibi dezavantajları bulunmaktadır (Lantz, 2013: 95).

2.2.4. Yapay Sinir Ağları

Yapay Sinir Ağları (YSA), doğrusal olmayan (kompleks) işlevleri modelleyebilmek için kullanılan etkin bir yöntem sağlamaktadır (Kohonen,1989). Yapay Sinir Ağları, biyolojik sinir sistemlerinin matematiksel modellerinin genellemeleri olarak geliştirilmiştir (McCulloch ve Pitts, 1943). Başka bir ifadeyle, insan beynindeki bir fonksiyonun istenilen bir görevi yerine getirme metodunu modellemek için kullanılan bir sistemdir. Yapay Sinir Ağları, yapay sinir hücrelerinin (nöron) kendileri arasındaki farklı şekillerdeki etkileşimleri sonucu bağlanmasıyla oluşmaktadır. Bu oluşan yapay sinir hücreleri, katmanlar hâlinindedir. Biyolojik nöronlardan hareketle, yapay bir nöronun kendisi dışındaki nöronlardan gelen sinyalleri bünyesinde bir araya getirerek toplanan bütün sinyallerin belirli bir taban (eşik) seviyesinin üzerine olması durumunda, bu yapay nörona bulunan sinyal diğer nörona gönderilmektedir (Fahey, 2003). Basitleştirilmiş yapay nöronların matematiksel yapısı Şekil 12'de gösterildiği gibidir.

Şekil 12: Yapay Nöron (Sinir Hücrelerinin) Matematiksel Form Örneği



Kaynak: Mijwil, 2017: 1.

Şekil 12'deki yapay nörona, ağırlıkların etkisine göre ($w_1, w_2, \dots, w_n$) hücreye gelen sinyaller ($x_1, x_2, \dots, x_n$) hücre içerisindeki eşik değeri göz önünde bulundurularak doğrusal olmayan aktivasyon fonksiyonu ϕ 'nin yardımıyla çıktı olarak y 'ye dönüştürülmektedir. Bu tür temel yapıdaki yapay sinir ağlarının çıktıları 0 veya 1 değerini almaktadır.

YSA, en basit hâliyle beş bileşenden meydana gelmektedir. Bileşenler ve bunlara ilişkin açıklamalara aşağıda yer verilmiştir.

Girdi: Girdiler (x), yapay nörona dışarıdan alınan verilerden oluşmaktadır. Ağın eğitilmesi sürecinde kullanılan örnekler tarafından oluşacağı gibi diğer nöronlardan veya kendisinden de bilgi alabilir.

Ağırlıklar: Ağırlıklar (w), yapay nörona gelen verilerin önemi ve aksiyonunu göstermektedir. Ağırlıkların büyük veya küçük olmasının, önem düzeyi üzerinde etkisi bulunmamaktadır. Yapay nöronlara ulaşan bilgiler, geçtikleri hatların (bağlantıların) ağırlık değerleriyle çarpılıp toplama fonksiyonuna (çekirdeğe) iletilmektedir.

Toplama Fonksiyonu: Toplama fonksiyonu, ağırlıkların nihai değerlerini toplayarak, o hücrenin ulaşan net girdiyi hesaplayan bir fonksiyondur. Çeşitli toplama fonksiyonları bulunmaktadır. En çok kullanılanı ise ağırlıklı toplamın bulunmasıdır. Ağırlıklı toplamlar, $net = (\sum_{j=1}^n w_j x_j)$ formülüyle elde edilmektedir. Yapay sinir ağlarının yapısına bağlı olarak en uygun toplama fonksiyonu kullanılmaktadır. Bu toplama fonksiyonlarından en yaygın olarak kullanılanları; çarpım, maksimum, minimum, kümülatif toplam fonksiyonlarıdır. Bu fonksiyonun belirlenmesi sınama yoluyla

(Skapura 1996) ve verilere öğretilmek istenilen amaca göre belirlenmektedir. Tüm ağ süreci boyunca aynı fonksiyon kullanılması şart değildir (Öztemel, 2003: 49-50).

Aktivasyon (Transfer) fonksiyonu: Toplama fonksiyonu üzerinden gelen net girdi değerlerini kullanarak hücrenin bu net girdi değerlerine karşılık üretilecek çıktı değeri belirlenmektedir. Çıktıyı belirlemek için çeşitli aktivasyon fonksiyonu kullanılmaktadır. Bunlar; sigmoid, doğrusal, threshold (eşik), hyperbolic tangent (tanjant sigmoid) gaussian, gibi fonksiyonlardır. Aktivasyon fonksiyonlarının en önemli noktası, fonksiyonun yapısına göre (0,1), (-1,1) veya $(-\infty, +\infty)$ çıkış aralık değeri arasındadır. En çok tercih edilen aktivasyon fonksiyonu ise sigmoid fonksiyon (özellikle lojistik sigmoid)'dur. En çok tercih edilen modellerden biri olan Çok Katmanlı Algılayıcı (MLP) gibi modellerin aktivasyon fonksiyonu için en önemli koşul türevlenebilir olmasıdır. Tüm ağ boyunca farklı aktivasyon fonksiyonu kullanılabilir (Lantz, 2013: 209-211).

Çıktı: Aktivasyon fonksiyonundan iletilen değer, yapay nöronun çıktısını oluşturmaktadır. Doğrudan dışarıya aktarılabilceği gibi, diğer bir yapay nöronun girdisi olarak kullanılabilir.

Şekil 12'de gösterilen yapay sinir hücrelerinin girdi değerlerine göre çıktı değerlerine dönüştüren matematiksel fonksiyon Eşitlik (2.10)'daki gibidir (Lantz, 2013: 208).

$$Y(x) = f \left(\sum_{j=1}^n w_j x_i + b \right)$$

(2.10)

Şekil 12'deki nöronların çoğunlukla tabakalı bir yapı oluşturacak şekilde birbirleriyle bağlanması sonucu yapay sinir ağları oluşmaktadır.

Yapay nöronların birleşmesiyle YSA oluşmaktadır. Bu katmanlar oluşurken rassal olarak bir araya gelmemektedirler. Yapay sinir ağlarının yapısı genel olarak üç katmandan oluşmaktadır. Girdi katmanı, gizli (ara) katman ve çıktı katmanı olmak üzere üç katmandan oluşmaktadır. Girdi katmanı, dışarıdan alınan girdileri (verileri) içermekte ve herhangi bir matematiksel işlem gerçekleştirilmemektedir. Bu katmanın temel görevi mevcut girdileri kendinden sonraki katmana iletmektir. Gizli katman,

girdi katmanı tarafından iletilen bilgilerin ilk işleminin gerçekleştiği katmandır. Buradaki işlenen bilgiler çıktı katmanına iletilmektedir. Yapay sinir ağlarının ağ yapısında çeşitli sayıda gizli katman bulunabilir. Çıktı katmanı, gizli katman tarafından alınan bilgileri ağ yapısında kullanılan aktivasyon fonksiyonu ile işlem gerçekleştirilerek çıktıyı oluşturmaktadır.

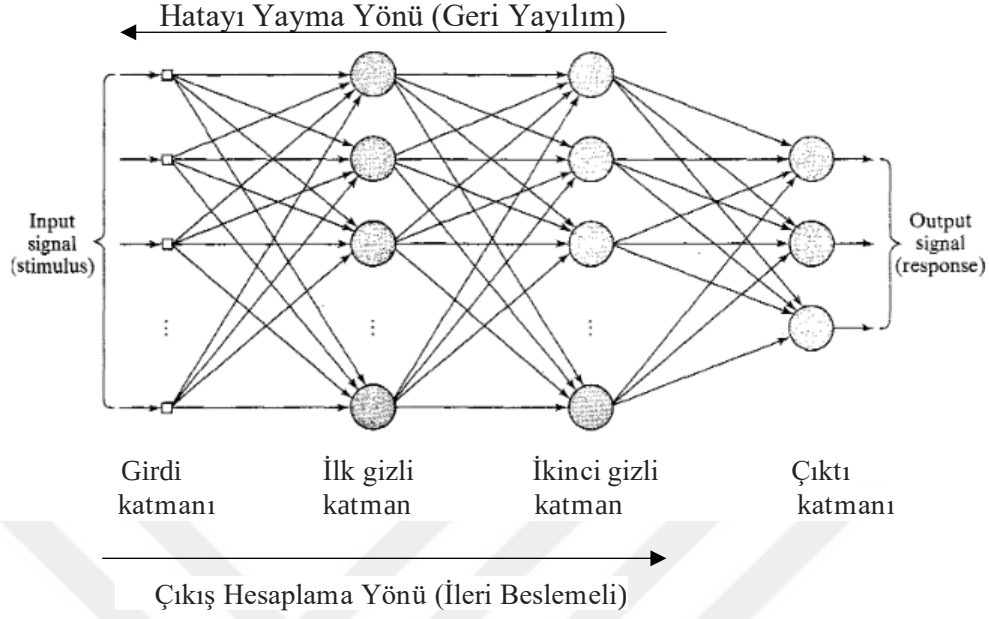
Yapay sinir ağlarının öğrenme stratejileri açısından gözetimli (denetimli) ve gözetimsiz (denetimsiz) yapay sinir ağları, ağ yapılarındaki katmanlar arasındaki akış yönlerine bağlı olarak ise ileri beslemeli (feed-forward) ve geri beslemeli (feed-back veya recurrent) yapay sinir ağları olmak üzere iki türü bulunmaktadır. İleri beslemeli gözetimli yapay sinir ağları, geri beslemeli yapay sinir ağlarına göre daha yaygın bir kullanıma sahiptir (Bataee ve Mohseni, 2011; Svozil ve diğerleri, 1997: 44 - 45).

İleri beslemeli yapay sinir ağları genel olarak yapay sinir hücreleri katmanlara ayrılmış yapıdadırlar. Bu katmanlarda bulunan her bir hücrenin ürettiği çıkıştan elde edilen değer, bir sonraki katmanın giriş değerini oluşturmaktadır. En basit hâliyle; yönler, bir yapay sinir hücresi girdi katmanından başlayarak, rassal sayıdaki gizli (ara) katmandan ve çıktı katmanına kadar devam eden ve nihai bir çıktı elde edilene kadar tek yönlü olarak iletilmesi şeklindedir (Ramchoun ve diğerleri, 2016: 26). Tek yönlü olduğu için hiçbir işlem elemanları arasında bir tekrarlama (döngü) söz konusu olmamaktadır. Çok fazla yapay sinir hücresi, ağ için önemli öznitelikleri veya eğilimlerini benimsemekte ve genelleme yapma özelliğini kaybetmesinden dolayı ezberlemesine sebep olmaktadır. Ağ boyunca tüm işlemler gerçekleştirilerek çıktı değeri oluşmaktadır. Bu çıktı değerleri ile istenilen çıktı değerleri arasında karşılaştırma yapıp bir hata değeri ortaya çıkmaktadır. Hata değerinin sonucuna bağlı olarak ağdaki ağırlıklarla revize edilmektedir. Yapay sinir ağları arasında en yaygın olarak kullanılanı, geri yayılım olarak bilinen ileri beslemeli çok katmanlı ağlardır. İleri besleme, ağ yapısını temsil ederken, geri yayılım ise eğitim metoduna verilen isimdir (Hamid ve Iqbal, 2004: 1118-1120). Bu tez kapsamında ileri beslemeli çok katmanlı algılayıcı (MLP) üzerinde durulacağından dolayı çok katmanlı algılayıcıdan bahsedilmiştir.

Çok Katmanlı Algılayıcı (MLP): Ağ, giriş katmanını oluşturan bir dizi yapay sinir hücrelerinden (nöron veya düğüm), bir veya daha fazla gizli yapay sinir hücreleri katmanından ve bir çıkış yapay sinir hücreleri katmanından meydana gelmektedir.

Girdi veya girdiler, ağ üzerinden katman katman ileriye doğru yayılım eğilimindedirler. Bu sinir ağları, Çok katmanlı algılayıcılar (MLP) olarak adlandırılmaktadır (Maciel ve Ballini, 2008: 7). Tek katmana sahip algılayıcılar, yalnızca doğrusal fonksiyonları tahminleme özelliğine sahiptirler. Katman sayısının arttığı ve tahminlemenin zorlaştığı durumlarla karşılaşmaktadır. Aynı şekilde, doğrusal olmayan regresyon için bir algılayıcının kullanması uygun değildir. Ayrıca, tek katmanlı algılayıcılar, özellikle XOR gibi sorunların doğrusal olmadığı için çözümde yetersiz kalmaktadır (Alpaydın, 2004: 245-246). Bu yetersizliğe çözüm bulunması amacıyla çeşitli araştırmalar sonucu yöntemler önerilmiştir. 1974 yılında, XOR probleminin çözümü için yaklaşım sunulmuştur (Werbos, 1974). 1980'li yıllara gelindiğinde, XOR için çözüm yöntemi bulunmuştur. Bu çözümün temelini oluşturan düşünce hataların geri yayılımının ara katman ile çıktı katmanı arasındaki hataların belirlenmesidir. Basit bir ifadeyle, her girdi değerine karşılık olarak ağ tarafından bir çıktı değeri üretilmektedir. Bu üretilen çıktı değeri ile gerçekte olması beklenen çıktı değerinin minimum yapacak şekilde ağ bağlantılarının ağırlıkları düzenlenmektedir. İstenilen sonucuna ulaşılan kadar bu işlemler devam etmektedir. Böylece, XOR problemi için bir çözüm gerçekleştirilmiş olmaktadır. Çıkış hesaplama yönünün (ileri beslemeli) tersi yönde olduğundan dolayı bu yöntem ise geri yayılım algoritması olarak ifade edilmektedir (McClelland ve diğerleri, 1986: 533-535). Çok katmanlı algılayıcılardaki yapay sinir hücreleri, geri yayımlı algoritma tarafından eğitilmektedir. İleri beslemeli ağda geriye yayımlı çok katmanlı algılayıcı Şekil 13'te gösterildiği gibidir.

Şekil 13: İki Gizli Katmana Sahip Çok Katmanlı Algılayıcının Yapısı



Kaynak: Haykin, 1994: 181.

Çok katmanlı ileri beslemeli ağlarda geri yayılım algoritması kullanılmasının hem avantajları hem de dezavantajları bulunmaktadır. Sayısal öngörü veya sınıflandırma problemlerine kolay uyarlanabilir özelliğe sahiptir. Modelleme yaklaşımları arasında en doğru modelleme gücüne sahip modeller arasında yer almaktadır. Eksik bilgi olsa da çıktı üretmesi gibi avantajları bulunmaktadır. Ağ yapısındaki karmaşıklığın hesaplama güçlüğü oluşturduğu ve bu nedenle yavaş eğitilmesine sebep olmaktadır (Lantz, 2013: 216). Ağ yapısı kurulumu için net bir kural olmadığından dolayı sınama (deneme yanılma) yoluyla en uygun ağ belirlenmesi gibi dezavantajları bulunmaktadır.

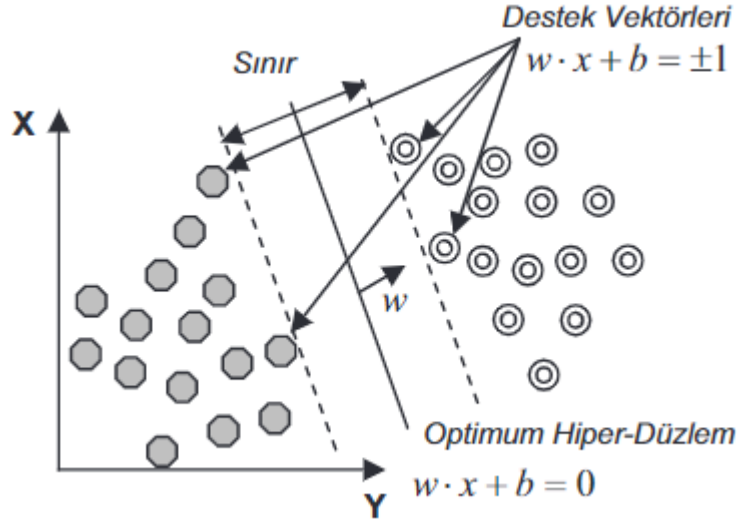
2.2.5. Destek Vektör Makinesi

Destek Vektör Makinesi (DVM), Vapnik tarafından önerilen sağlam teorik temellere dayalı ve yaygın bir kullanıma sahiptir (Vapnik, 1982: 430). Destek vektör makineleri, doğrusal olmayan sınıflandırma, yoğunluk veya fonksiyon tahmininde farklı problemlerin çözümünde kullanılan ve çekirdek (kernel) altyapılı metotlarda çeşitli yeniliklerin ortaya çıkmasını sağlamıştır (Li ve diğerleri, 2006: 11). Destek

vektör makineleri, istatistiksel öğrenme teorisine dayanan bir gözetimli (denetimli) öğrenme algoritmasıdır. Destek vektör makinelerinin temel amacı, sınıflandırma için iki sınıfı maksimum (optimal) olarak bölen hiper düzlem (karar doğrusu) belirlenmesine dayanmaktadır (Vapnik,1995: 290-291). Başka bir ifadeyle, iki farklı sınıfa ait verilerin olduğu bir sınıflandırmada sonradan dahil edilecek verilerin hangi sınıfta olacağına karar verilmesidir. Burada eğitim veri kümeleri yardımıyla optimal bir hiper düzlem seçilir ve genellenebilir olma özelliği için test veri kümeleri ile doğrulama yapılarak belirlenmektedir (Kavzoglu ve Colkesen, 2009: 354). Destek vektör makineleri, veri kümesinin doğrusal ayrılabilir olup olmadığına göre temel olarak üç gruba ayrılmaktadır. Doğrusal Ayrılabilir Durum (Maksimal Marj DVM), Doğrusal Ayrılabilir Olmayan Durum (Yumuşak Marj DVM) ve Doğrusal Olmayan Destek Vektör Makineleri şeklinde üç grupta ele alınmaktadır.

Maksimal Marj Destek Vektör Makineleri: Destek Vektör Makinesi'nin ilk sunulan ve en temel modeli maksimal marj sınıflandırıcısıdır. Maksimal marj (sınır) , sadece öznitelikler uzayında doğrusal olarak bölünebilir (ayrılabilen) veri kümeleri için çalışmaktadır ve bundan dolayı gerçek hayatla ilgili problemlerde kullanılmamaktadır. Destek vektör makinelerinin temelini oluşturmaktadır. Ayrıca, maksimal sınıflandırıcısı, verileri optimal marjlı hiper düzlemle ayırıp, bu ayrılan marjı optimize etmektedir. Marjın, uzayın boyutsallığından bağımsız olduğu, bu ayırım herhangi bir çekirdeğe (kernel) dayalı bir öznitelik uzayında aranabilir. Buradaki amaç, uygun şekilde seçilmiş bir kernel kaynaklı öznitelik uzayında maksimal marj hiper düzleminin belirlenmesidir. Böylece, onu bir dışbükey optimizasyon problemi hâline getirip çözümün daha kolay olması sağlanarak uygulanmaktadır (Cristianini ve Shawe-Taylor, 2000: 94). İlk sunulan destek vektör makinesi, en uygun bölen hiper düzlemi kullanan doğrusal olan bir ayırıcıdır. Maksimal marj hiper düzlemi olarak da bilinen bu hiper düzlem, ikinci dereceden (quadratic) bir optimizasyon ile hesaplanmaktadır (Martin, 2001: 5). Hiper düzlemler sonsuz farklı şekilde çizilebilmektedirler. Destek vektörleri, hiper düzleme en yakın olan vektörlere denir (Sherrod, 2003: 289). Doğrusal olarak ayrılabilen veri kümeleri için hiper-düzlemin belirlenmesi örneği Şekil 14'te gösterilmiştir. Şekil 14 üzerinde destek vektör makinesi ile hiper düzlem arasında kalan alan marj (sınır) olarak adlandırılmaktadır.

Şekil 14: Doğrusal Ayrılabilen Durum İçin Hiper-Düzlem Belirlenmesi



Kaynak: Kavzoglu ve Colkesen, 2010: 76.

N -boyutlu örneklerden oluşan $\{x_i, y_i\}$, $i = 1, \dots, n$ şeklinde iki sınıfa ayrılabilir eğitim veri seti olsun. $y \in \{-1, +1\}$ biçiminde yanıt değişkeni iki sınıfa temsil eden sınıf etiketlerinden oluşmaktadır. Burada $x \in \mathbb{R}^N$, N -boyutlu uzaydır (Huang ve diğerleri, 2002; 727). Bir hiper düzlem $w \cdot x + b = 0$ olarak tanımlanabilir. Burada, x hiper düzlem üzerinde uzanan bir nokta, w bir ağırlık vektörü (hiper düzlemin normali), b ise yanlılık (bias) olarak adlandırılmaktadır. Örneğin; Aylık internet ücreti ve alınan hizmetin süresi olmak üzere iki sınıfa ait veri kümesi olsun. Burada aylık ödenen internet ücreti ve buna bağlı alınan hizmet süresi ile ilgili sonraki dönemlerde kullanıma devam edip etmeyeceğini öngörü yapmak istediğimizi varsayalım. Bu iki sınıf için sınıf etiketleri; devam etmeyenler -1, devam edenler +1 şeklinde ifade edilsin. Etiketlenmemiş yeni gözlem, karar fonksiyonuna bağlı olarak -1 değerine eşit veya daha küçükse, yeni gözlem değeri eğitilen veri kümesinde -1 etiketine ait sınıfa atanır. Doğrusal olarak ayrılabilir iki sınıflı veriler için hiper düzlemler;

$$w \cdot x_i + b \geq +1, \text{ her } y_i = +1 \text{ için}$$

$$w \cdot x_i + b \leq -1, \text{ her } y_i = -1 \text{ için}$$

(2.11)

2.11’de verilen eşitliği birleştirirsek;

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) - 1 \geq 0 \quad \forall_i \quad (2.12)$$

2.11’de verilen eşitlikte tanımlanan ve en iyi hiper düzleme paralel olan bu iki hiper düzlem üzerinde yer alan eğitim veri noktaları destek vektörünü göstermektedir (Burges, 1998: 129; Mathur ve Foody, 2008: 242). Burada düzlemler arasındaki marj 2.13’te yer alan eşitlikte verilmiştir.

$$\frac{2}{\|\mathbf{w}\|} \quad (2.13)$$

Sınıfları en yüksek sınırla ayıran en uygun ayırma hiper düzlemini belirleyebilmek için 2.12’deki eşitliğin kısıtları altında $\|\mathbf{w}\|^2$ ’yi minimum yaparak elde edilmektedir (Song ve diğerleri, 2012: 3307). Böylece kısıtlı kareli optimizasyon problemine dönüştürülmüş ve 2.14’te eşitlikte gösterilmiştir.

$$\min \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\} \quad (2.14)$$

Buna bağlı olarak kısıtlı optimizasyon sınırlamaları ise;

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b}) - 1 \geq 0, \quad y_i \in \{-1, +1\} \quad (2.15)$$

biçiminde ifade edilmektedir (Mathur ve Foody, 2008: 242). Eşitlik 2.14’teki optimizasyon probleminin çözülebilmesi ve sonucunda, sınıflara ait olan destek vektörleri arasındaki mesafenin maksimum yapacak şekilde en uygun ayırma hiper düzlemi belirlenir (Cortes ve Vapnik, 1995: 278). Kısıtlı karesel (kuadratik) optimizasyon probleminin çözümü ile optimal ayırma hiper düzlemi bulunmaktadır. Bu kısıtlı karesel optimizasyon probleminin çözümü ise Lagrange çarpanı yardımıyla yapılmaktadır. Lagrange çarpanı minimizasyon problemlerini dual probleme çevirerek kolay sonuca ulaşmasında kolaylık sağlamaktadır. Lagrange fonksiyonu 2.16’da gösterildiği gibidir (Gunn, 1998: 8).

$$L(w, b, \lambda) = \frac{1}{2} ||w||^2 - \sum_{i=1}^n \lambda_i [y_i(w \cdot x_i + b)] + \sum_{i=1}^n \lambda_i \quad (2.16)$$

2.16’da verilen eşitlikte, $\lambda_i \geq 0$ olmak koşuluyla, eşitlikte yer alan λ_i ‘ler Lagrange çarpanı, w bir ağırlık vektörünü, b ise sabit olarak ifade edilmektedir. Sonuç olarak, doğrusal olarak ayrılabilen problem için karar fonksiyonu aşağıdaki gibidir (Osuna ve diğerleri, 1997: 7-8).

$$f(x) = \text{sign} \left(\sum_{i=1}^n y_i \lambda_i (x \cdot x_i) + b \right) \quad (2.17)$$

Yumuşak (Soft) Marj Destek Vektör Makinesi: Doğrusal olarak ayrılabilen maksimal marj destek vektör makinesinin aksine doğrusal olarak ayrılamayan yumuşak marj destek vektör makinesi, verilerin tam olarak ayrılamadığı durumlarda tercih edilmedir. Eğitim verilerinin hatasız olarak ayrılamadığı ve eğitim setinin en az hata ile ayırmak için kullanılmaktadır (Cortes ve Vapnik, 1995: 280; Cristianini ve Shawe-Taylor, 2000: 103). Bunun nedeni veri kümelerinin gürültülü veri barındırması, boyut sayısının fazla olması buna bağlı olarak kompleks bir yapıda olmasından kaynaklanmaktadır (Li ve diğerleri, 2009: 979). Yanlış sınıflandırma hatasından kaçınmak ve marj vektörünü en uygun yapmak için negatif olmayan bir aylak değişken (slack variable) ξ , 2.15’teki eşitliğe eklenir. Yeni Eşitlik 2.18’de gösterilmiştir (Cristianini ve Shawe-Taylor, 2000: 104).

$$y_i(w \cdot x_i + b) - 1 + \xi_i \geq 0 \quad \text{veya} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad (2.18)$$

$$\xi_i \geq 0, \quad i = 1, \dots, n$$

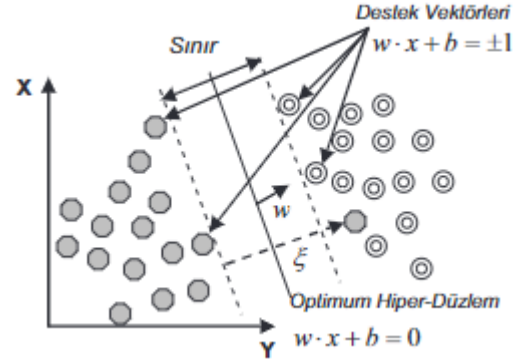
Doğrusal ayrılamayan durum için hiper düzlemin belirlenmesi örneği Şekil 15’te gösterilmiştir.

Şekil 15: Doğrusal Ayrılamayan Durum İçin Hiper Düzlemin Belirlenmesi



(7a)

Doğrusal ayrılamayan
İki sınıf için veri kümesi



(7b)

Doğrusal ayrılamayan veri kümesi için
Hiper düzlemin seçilmesi

Kaynak: Kavzoglu ve Colkesen, 2010: 77.

Yumuşak marj destek vektör makinesi, marj sınırının olmaması gereken taraftaki veri noktaları, bu marj sınırından uzaklaşmasıyla artan bir ceza değeri oluşmaktadır. Bu ceza değeri C ile gösterilmektedir (Fletcher, 2009: 7). 2.18’de verilen eşitlikte, optimal hiper düzlem bulunurken aynı zamanda yanlış sınıflandırmaya bağlı olarak sınıflandırma hataları için bir ağırlık olarak maliyet için C parametresi yer almaktadır. Bu parametre, amaç fonksiyonunu minimum yapacak şekilde düzenlenir (Burges, 1998: 135). Marjı maksimum yapmaya çalışırken, hataların toplamını en az düzeyde tutmak gerekmektedir. 2.18’deki eşitlik kısıtları altında, 2.14’teki eşitlik kullanılarak yeni optimizasyon problemi 2.19’daki gibidir.

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (2.19)$$

2.18’deki denklemde eşitsizlik kısıtlamaları altında, 2.19’daki denklemin ilk kısmı, sınıflar arasında yer alan sınırı maksimum yapabilmek için doğrusal olarak ayrılabilir durumdaki ile aynı şekildedir. 2.19’daki denklemin ikinci kısmı ise, hiper düzlemin olmaması gereken tarafta olan örnekler cezalandırılmaktadır. Eşitlikte

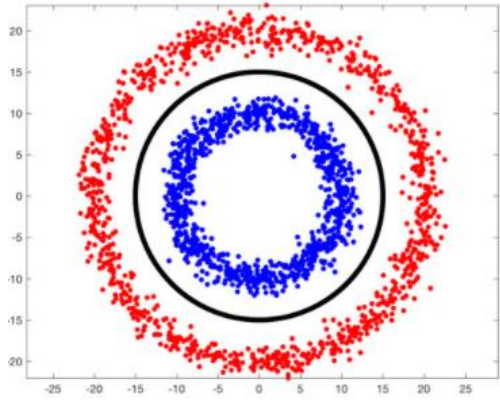
verilen ξ_i 'nin seçilmesi aşamasında çeşitli durumların göz önünde bulundurulması gerekmektedir. 2.19'da elde edilen eşitlikte, C parametresi kullanıcı tarafından belirlenmektedir. Eğitim setinden kaynaklanan olası hataları minimum düzeye indirebilmek ve marjı maksimum düzeye çıkarma arasındaki denge için kilit rol oynamaktadır. C parametresi ne kadar yüksek olursa, veri sınıflandırılmasındaki olası hatalar minimum düzeyde olur (Pal ve Mather, 2004: 1218). 2.19'da verilen eşitliğin çözülebilmesi için w , b ve ξ_i 'ye göre minimize edilirken, λ 'ya göre maksimize edilmesi gereken yeni bir Lagrange formülü 2.20'deki gibidir (Fletcher, 2009: 8).

$$L(w, b, \lambda, r, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \lambda_i [y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^n r_i \xi_i$$

$$\lambda_i \geq 0, r_i \geq 0 \quad \forall i \quad (2.20)$$

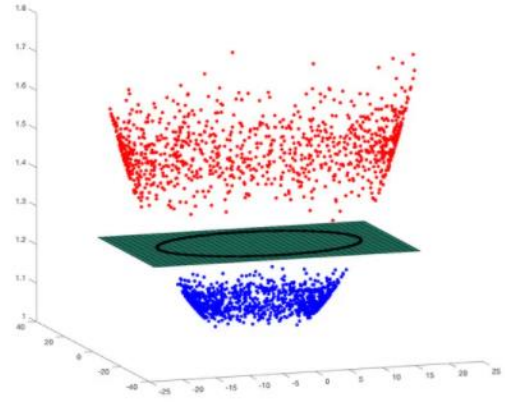
Doğrusal Olmayan Destek Vektör Makinesi: Doğrusal ayrılabilen destek vektör makineleri, tam ayrılabilen veya belirli bir hata ile veri kümeleri ayrılabilirken, doğrusal olmayan vektör makinesi bu iki koşulunun sağlanamadığı durumlarda kullanılmaktadır. Doğrusal olmayan destek vektör makinesi, gerçek hayat problemi için kullanıma elverişlidir. Gerçek hayatta karşılaşılan problemlerin çoğunda veri sınıfları dengesiz bir özelliğe sahiptir ve bu nedenle sınıfları doğru bir şekilde ayırabilmek için uygun hiper düzlemin belirlenmesi kolay olmamaktadır. Bu veri kümelerini ayırabilmek ve doğru hiper düzlemin belirlenmesi amacıyla çekirdek fonksiyonları kullanılmaktadır. En yaygın kullanılan çekirdek fonksiyonları; doğrusal fonksiyon, polinomsal fonksiyon, sigmoid fonksiyon ve radyal tabanlı fonksiyonlar sıklıkla kullanılmaktadır.

Şekil 16: İki Boyutlu Uzaydaki Verilerin Çekirdek Fonksiyonu ile Üç Boyutlu Uzaya Dönüştürülmesi ve Hiper Düzlemin Belirlenmesi Örneği



(16a)

R^2 (İki Boyutlu Uzay)
(Hiper Düzlemle Ayrıl原因amayan)



(16b)

R^3 (Üç Boyutlu Uzay)
(Hiper Düzlemle Ayrıl原因abilen)

Kaynak: Mallick, 2018.

(16a)'da verilen iki boyutlu bir veri dağılımı gösterilmektedir. Burada, karar sınırı siyah renkle gösterilen çemberdir. Bu veri dağılımı doğrusal bir hiper düzlemle ayrılması mümkün değildir. Böyle durumlarda, veri setini ayırabilmek için verilere yeni bir boyut olarak eklenen çekirdek hilesi (Kernel Trick) kullanılmaktadır. Bu çekirdek hilesi tekniği ile veriler mevcut hâlden bir üst boyuta taşınmış olacak ve doğrusal bir ayrılma sağlanmış olacaktır. Böylece, iki boyutlu uzaydaki verilerin çekirdek fonksiyonu ile üç boyutlu uzaya dönüştürülmüştür.

Destek vektör makinelerinin, sınıflandırma veya sayısal bir öngörü problemleri için kullanılması uygundur. Veri kümesi ile ilgili detaylı bilgi olmasa da iyi sonuçlar elde edilebilir. Ek olarak, önceden belirli bir şekilde tanımlanmamış bir veri modeli özelliği taşımayanlar yüksek boyutlu verilerin kullanılmasında iyi çalışmaktadır. Başka bir ifadeyle, yapılandırılmamış veri setleri; e-posta ve web sayfaları gibi yüksek boyutlu veriler için uygun yapıdadır.

Karmaşık yapıdaki veriler uygun bir çekirdek hilesi yardımıyla çözmesi gibi avantajları bulunmaktadır. Optimal çekirdek çözüm işlevinin belirlenmesi oldukça zordur. Yüksek boyutlu veri kümelerinin kullanılması eğitim süresini uzatmaktadır.

Bireysel faktörlerin etkisinden dolayı problemler oluşmakta ve değişkenlerin ağırlıklarına bağlı olarak yorumlanmasını zorlaştırması gibi dezavantajları bulunmaktadır (Auria ve Moro, 2008: 7-8; Fedorovici ve Dragan, 2011: 258).

2.2.6. Karar Ağaçları Algoritması

Karar ağaçları, sınıflandırma ve regresyon problemleri için yaygın bir şekilde kullanılmaktadır. Sınıflandırma ve regresyonların temel amacı, tahminleme yapılması amacıyla yararlanılmak üzere bir model oluşturulmaktadır. Bu model oluşturulurken, eğitim veri kümeleri kullanılmaktadır. Eğitim veri kümesindeki her bir kaydın çeşitli öznitelikleri bulunur. Regresyon için sayısal özellikteki veriler kullanılırken, sınıflandırma için kategorik öznitelikteki veriler kullanılmaktadır (Gehrke, 2003: 3-4). Karar ağaçları amaçlarına bağlı olarak, sınıflandırma ve regresyon için sırasıyla, sınıflandırma ağacı ve regresyon ağacı olarak adlandırılmaktadır (Maimon ve Rokach, 2014: 79).

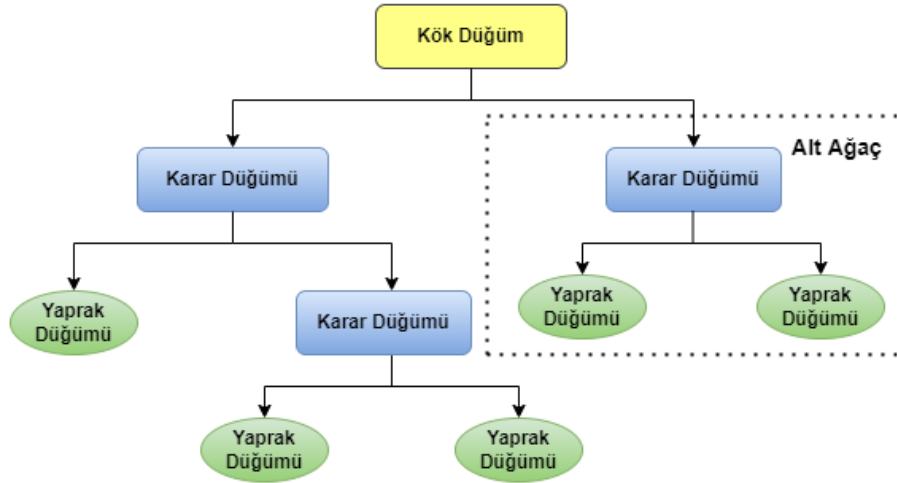
Karar ağaçları, örnek uzayının kendi kendini yineleyen (recursive) bir bölümü olarak adlandırılan sınıflandırıcılardır. Karar ağaçları, kök, dal ve yapraklardan meydana gelen bir ağaç yapısından oluşmaktadır (Sharma ve diğerleri, 2011: 191). Karar ağaçları, köklü bir ağacı meydana getiren düğümlerden oluşmaktadır. Hiyerarşik (aşamalı sırası) yapıda bulunan karar ağaçları, kök düğümden başlayarak dallanmaktadır. Kök düğüm, kendisine doğru gelen hiçbir kenarı bulunmayan bir özelliğe sahiptir. Kök düğüm dışında kalan düğümler, bir kendisine gelen bir kenarı bulunmaktadır. Giden kenarları bulunan bir düğüm, test düğümü olarak adlandırılmaktadır. Diğer tüm düğümler ise, yaprak (karar düğümleri) olarak isimlendirilir. Bir karar ağacında, kök düğümleri iki veya daha çok alt dallara ayrılabilir. Dallara ayrılırken, tek bir özellik dikkate alınmakta ve örnek alan özellik değerlerine göre ayrılır. Sayısal özellikteki durumlarda, bir koşul aralığı olarak ifade edilmektedir. Karar düğümleri (yaprak), optimal amaç değerini temsil eden bir sınıfa tayin edilmektedir. Diğer bir seçenek olarak, karar düğümleri, amaç özelliğinin belirli bir değer bulundurma olasılığı temsil eden bir olasılık vektörüne sahip olabilmektedir (Rokach, 2005: 165-166). Karar ağaçları, diğer kullanılan algoritmaların daha iyi sonuç verdiği modeller olduğunda bile, karar ağaçları çok daha

ön planda olmaktadır. Ön planda olmasının en büyük nedenlerinden biri anlaşılır bir yapıya sahip olmasıdır. Karar ağaçları, belirtilen kararların basit bir şekilde sunulmasını sağlamaktadır (Grabczewski, 2014: 1). ID3, C4.5, C5 algoritmaları Quinlan tarafından, Classification and Regression Tree (CART, Sınıflandırma ve Regresyon Ağacı) Algoritması Breiman ve diğerleri tarafından, Chi-squared Automatic Interaction Detection (CHAID, Ki-Kare Otomatik Etkileşim Algılama) gibi popüler çeşitli karar ağaçları bulunmaktadır (Kumari ve Godara, 2011: 305; Tu ve diğerleri, 2009: 184).

Karar ağacı algoritması kullanarak veri kümesinde sınıflandırma iki aşamalı bir işlemle gerçekleşmektedir. İlk aşama öğrenme aşamadır. Bu aşamada önceden bilgi sahibi olunan bir eğitim veri kümesi, modelin oluşturulması için sınıflandırma algoritması yardımıyla analiz gerçekleştirilir. Eğitilen model, karar ağacı olarak gösterilmektedir. İkinci aşaması ise sınıflandırma aşamasıdır. Bu aşamada test verisi, karar ağacının doğruluğunu test için kullanılmaktadır. Doğruluk oranlarının uygun bulunması hâlinde, mevcut kurallar sonradan dahil edilen verilerin sınıflandırılması için kullanılır (Han ve Kamber, 2011: 18; Özekes, 2003: 69).

Basit bir örnekle açıklamak gerekirse; bir kişinin kredi almak için başvuru yaptığını varsayalım. Bankanın, müşterinin kredi talebini değerlendirmesi için çeşitlik faktörleri gözden geçirmesi gerekmektedir. Geçmişteki kredi notu, geliri, temerrüde düşüp düşmediği gibi çeşitli faktörler olduğunu varsayalım. Kredi için en kritik karar noktası temerrüde düşüp düşmediği olsun. Bu kritik karar, kök düğümde yer almaktadır. Kök düğümdeki soruya ait iki cevap bulunmaktadır; evet ve hayır şeklindedir. Verilen cevaplara göre, kredi verilebilir/kredi verilemez şeklinde kararın verilmesinde takip edilen yolu belirlemektedir. Böylece, banka, müşterinin çeşitli kriterlerini değerlendirdikten sonra nihai kararını vermektedir. Ayrıca, kredi için farklı kriterler bankanın kredi kararını etkilemektedir. Bunlar, kredi talep eden kişinin yaşı, geliri, eğitim durumu, sahip olduğu mal varlığı gibi çeşitlendirilebilir. Burada yaş değişkeni baz alınarak, kredi talebi örneğini biraz daha detaylandıralım. Yaş değişkenine göre, yaşı 25'ten büyük, üniversite mezunu, evi var, arabası var ve kredi durumu iyi olan biri muhtemelen kredi talebi için bankadan olumlu yanıt alacaktır. Şekil 17'de karar ağacının genel yapısı gösterilmiştir.

Şekil 17: Karar Ağacı Yapısı Örneği



Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Karar ağaçlarının en önemli aşamalarından biri kök değişkeninin belirlenmesi aşamasıdır. Belirlenen optimal kök değişkenine göre dallanması sayısı minimum düzeyde olurken, algoritmanın güçlü tahminleme yapmasına olanak sağlayacaktır. Karar ağaçlarının bölünmesinde farklı matematiksel algoritmalar kullanılmaktadır. Bunlardan en çok tercih edileni; Ki-kare testi, Entropi, Gini indeksi ve Bilgi Kazancı (Information Gain)'dır. Ek olarak, ID3'te, C4.5 (J48) ve C5 algoritmaları Entropi temelli bilgi kazancı kullanılmakta; CHAID algoritmasında Ki-kare ve CART algoritmasında ise Gini indeksi kullanılmaktadır (Kumari ve Godara, 2011: 305; Tu ve diğerleri, 2009: 184).

Bilgi kazancının hesaplanması işleminin gerçekleştirilmesinde Entropi ölçüsünde yararlanılmaktadır (Srinivas ve Kumar, 2013: 608). Bilgi kazancı, var olan nitelikler arasındaki en yüksek bilgi kazancı özelliğine sahip olan ağacın kök düğümünde yer almaktadır. Entropi, belirsizlik, düzensizlik ve rassallık ölçüsü olarak ifade edilmektedir. Bir veri setinin alt kümelerinde bulunan gözlemler homojen (purity) ise, veri setinde rassallık veya heterojenlik (impurity) yoktur. Homojenlik oranı ile entropi oranı ters orantılıdır. Homojenlik oranı artarken, Entropi oranı azalmaktadır. Veri yapısındaki heterojenliğin yüksek olması durumunda, entropi de yüksek olmaktadır. Entropi, negatif olmayan bir değere sahiptir. Alt kümelerde bulunan gözlemlerin hepsi bir sınıfa ait olduğunda, o veri setinin entropisi 0 olmaktadır

(Wang, 2017: 3; Fan, 2011: 708). Entropi değeri en küçük olan değerin kök değişkeni olarak seçilmesi uygun olmaktadır. Entropi, her bir sınıf etiketinin olasılığının toplamı ile aynı sınıfa ait etiketin logaritmik olasılık toplamı olarak tanımlanabilir. 2.21'deki formül ile entropi değeri hesaplanmaktadır.

$$\text{Entropi (S)} = p_1 \log_2 p_1 + p_2 \log_2 p_2 + \dots + p_j \log_2 p_j$$

$$\text{Entropi (S)} = - \sum_{i=1}^j p_i \log_2 p_i \quad (2.21)$$

2.21'de verilen eşitlikte, P_i bağımlı değişkende yer alan kategorik değere sahip değişkenlerin meydana gelmesi olasılığını göstermektedir.

Bilgi kazancı, bir sınıfa ait entropi değeri ile sınıfın ve seçilen özelliğin koşullu entropi değeri arasındaki fark olarak ifade edilmektedir. Bilgi kazancı, sınıflandırmadaki bir özelliğinin kullanışlılığını ölçmektedir (Lefkovits ve Lefkovits, 2017: 894). Basit bir ifadeyle, bilgi kazancı, bir f özelliği üzerinde S kümesinin bölünmesindeki önce entropi değeri ile bölündükten sonraki entropi değeri arasındaki fark olarak ifade edilmektedir. 2.22'deki formül ile bilgi kazancı (Information Gain) değeri hesaplanmaktadır.

$$\text{Bilgi Kazancı (S, f)} = - \sum_{i=1}^j p_i \log_2 p_i - \sum_{v=1}^V \frac{|S^v|}{|S|} \text{Entropi (S}^v\text{)}$$

$$\text{Bilgi Kazancı (S, f)} = \text{Entropi (S)} - \text{Entropi (S}^v\text{)} \quad (2.22)$$

2.22'deki verilen eşitlikteki bilgi kazancı değerinin artması, özniteliklerin sınıflandırılması için daha elverişli olduğunu göstermektedir. Optimum bilgi kazancına sahip öznitelik, bölme aşamasında kullanılması en uygun öznitelik olarak seçilmektedir (Wang, 2017: 3).

Karar ağaçlarında ortaya çıkması muhtemel problemlerden biri aşırı öğrenme (overfitting)'dir. Algoritma seçilen bir özelliği her bölmesinde ortaya çıkan her kurala yeni bir terim eklenmektedir. İlgisiz veya gürültü içeren verilerin olduğu bir ağaç oluşturulması hâlinde, algoritmaya sonradan dahil edilen her bir örneği sınıflandırırken hatalarla karşılaşılacak ve kötü bir performans göstermesine neden olacaktır. Böylece, sınıfın genelleme yeteneği etkisini kaybedecektir. Aşırı öğrenmenin engellenebilmesi için budama yöntemi kullanılmaktadır. Budama yöntemi;

Ön Budama (Pre-pruning, İleri Budama): Ön budama aşamasında gereksiz dalların oluşmasının engellenmesidir.

Budama Sonrası (Post-pruning, Geriye Doğru Budama): Budama sonrası aşamasında ise karar ağacının oluşturulması ve akabinde önemli olmayan dalların karar ağacından çıkarılması işlemidir.

Budama öncesi ve budama sonrası yapılan işlemlerle algoritmanın genellebilirliğini artırmaktadır (Bramer, 2007: 125-126). Tüm algoritmalarda olduğu gibi karar ağaçlarında da avantajlar ve dezavantajlar bulunmaktadır. Avantajları; Karar ağacı, hem regresyon hem de sınıflandırma problemleri için uygulanabilir. Yorumlanabilir olması, anlaşılması ve görselleştirilmesi bakımından kolay olduğu söylenebilir. Kullanılan diğer algoritmalarla kıyaslandığında, veri hazırlık sürecinde daha az emek harcanmaktadır. Karar ağaçları, eksik veya gürültülü verilerden genel olarak etkilenmemektedir. Sayısal ve kategorik değişken kullanımı için uygundur. Dezavantajları ise; Karar ağaçları için en önemli problemlerden biri aşırı öğrenmedir. Sürekli sayısal değişkenler için kullanıma pek uygun değildir. Veri yapısındaki küçük değişimler, ağaç yapısında önemli ölçüde değişikliğe sebep olmaktadır. Son olarak, karmaşıklık durumuna göre, modelin eğitilmesi zorlaşabilir ve maliyeti artırabilir.

2.3. TOPLULUK ÖĞRENME TEKNİKLERİ

Topluluk öğrenmesi modelleri, tekil modellerde olduğu gibi modelleri birbirinden bağımsız olarak değil, bir bütün olarak bir araya getirerek güçlü bir öngörü yapmak için kullanılan tekniktir. Başka bir ifadeyle, topluluk öğrenmesi, birden fazla sınıflandırıcıların tahminlerini birleştirerek sınıflandırmanın doğruluğunu

iyileştirmeyi amaçlamaktadır (Steinbach ve diğerleri, 2016: 159). Topluluk öğrenmesinde, farklı sınıflandırıcılar bulunmaktadır. Bu sınıflandırıcılar, birbirinden bağımsız şekilde zayıf veya temel öğrenciler olarak isimlendirilmektedir. Topluluk öğrenme tekniğın temel mantığı, çeşitli temel öğrencilerden alınan kararların, tek temel öğrenciden alınan kararlara göre daha iyi sonuçlar vereceğini öngörmektedir. Bu bağlamda, çeşitli temel öğrenciden alınan kararların, yalnızca bir temel öğrenciden alınan karardan daha yüksek güvenilirliğe sahip olması ve böylece daha yüksek doğruluğa sahiptir. Örneğin; Rassal 1000 kişiye sorulan karmaşık yapıdaki bir sorudan alınan cevapların, çoğu durumlarda uzman kişinin sorulara verdiği yanıttan daha iyi sonuçlar verdiği görölmektedir. Doğruluğu artırıp, varyansını azaltmak için geliştirilmiş olan topluluk öğrenme teknikleri, güven tahmini, öznelik seçme, hata düzeltimi, dengesiz veri kümelerinde ve durağan özellikte bulunmayan dağılımlarda sapmayı öğrenmesi gibi birçok makine öğrenmesi problemlerinin çözölmesinde başarılı bir şekilde kullanılmaktadır (Polikar, 2012: 1-2).

Topluluk öğrenme teknikleri iki yöntem kullanılarak sınıflandırılabilir:

Ortalama Yöntemi: Birden çok tahmin edici kullanarak tahminde bulunmak ve bu tahminlerin ortalamasının alınması işlemidir. Ortalama yöntemi ile bir araya getirilen tahmin edicinin varyansı azaltmakta ve tek tahmin edicilere göre daha iyi performans göstermektedir. Ortalama kullanılan topluluk öğrenme yöntemleri; bagging (özyinelemeli birleştirme) ve rassal ormanlar gibi yöntemlerdir.

Artırma (Boosting) Yöntemi: Temel tahmin ediciler ardışık olarak oluşturulur ve bir araya getirilen tahmincilerin sapmasının azaltılması amaçlanmaktadır. Birçok zayıf modellerin birleştirilmesiyle güçlü bir topluluk oluşmaktadır (Korniiichuk ve Boryczka, 2021: 3678). Adaboost, gradient boosting gibi yöntemler kullanılmaktadır.

Özyinelemeli Birleştirme ve Artırma yöntemi detaylıca anlatılmıştır.

Topluluk öğrenmesi tekniklerinin sınıflandırılmasında kullanılan ilk teknik olan Bagging (Bootstrap Aggregating – Özyinelemeli Birleştirme) sınıflandırma yöntemi; Sınıflandırma yöntemlerinin sonuçlarını iyileştirmek amacıyla geliştirilen bir tekniktir. Bu teknik, 1994 yılında Leo Breiman tarafından önerilmiştir. Özyinelemeli (Bootstrap) ve Birleştirme (Aggregating) ifadelerinin kısaltması ile “**Bagging**” kavramı türetilmiştir (Breiman, 1996: 2). Bagging, tahminle ilişkili olan varyansı

azaltmak ve tahmin sürecini daha da iyileştirmek için çeşitli sınıflandırma ve regresyon yöntemleri ile kullanılan bir yöntemdir. Orijinal eğitim veri kümelerinin önyüklenmiş kopyaları ve çeşitli sınıflandırıcıların eğitimini içermektedir. Burada, orijinal veri setlerinden rassal örnek seçilerek her bir sınıflandırıcının eğitilmesi için yeni alt veri setleri oluşturulmaktadır. Böylece, çeşitli alt kümelerin kullanılmasından dolayı yeniden örnekleme yöntemiyle çeşitlilik sağlanmaktadır. Yeniden örneklemeyle ilgili olarak aynı örnek, birden fazla alt kümede yer alabilir. Son olarak, sınıflandırıcıların her birine test aşamasında bilinmeyen bir örneğin hangi sınıfa ait olduğunu tespit etmek için çoğunluk oy veya ağırlık kullanılarak belirlenmektedir (Galar ve diğerleri, 2011: 468). Basit bir ifadeyle, veri kümelerinden çeşitli sayıda önyükleme örneği alınır, alınan bu örneklerle bazı tahminleme yöntemleri uygulanır ve sonrasında sonuçlar ortaya çıkmaktadır. Elde edilen sonuçlar, regresyon problemleri ve sınıflandırma problemleri için farklı şekilde uygulanmaktadır. Regresyon problemleri, sayısal değerlerden oluştuğu için ortalama alınarak tahmin değeri elde edilmektedir. Sınıflandırma problemleri, çoğunluk oylaması yöntemi ile tahmin edilmektedir. Sınıflandırmada çoğunluk oyunun seçildiği sınıf tahmin değeri olarak kullanılmaktadır (Sutton, 2005: 305). Kullanılan bu yöntemlerle hem varyans azaltılır hem de aşırı öğrenme (overfitting) sorunun önüne geçilmiş olacaktır ve böylece genelleme yeteneği artırılmış olmaktadır. Bagging yöntemine dayalı olarak kullanılan en yaygın makine öğrenmesi yöntemi Rassal Ormanlar (Random Forest)'dır. Bu yöntem dışında, oylama sınıflandırıcısı, karar ağaçları ve ekstra ağaçlar gibi yöntemler sıklıkta tercih edilmektedir (Zhou, 2012: 48). Bahsi geçen her bir yöntem detaylandırılarak anlatılmıştır.

2.3.1. Rassal Ormanlar Algoritması

Rassal Ormanlar Algoritması, 2001 yılında Leo Breiman tarafından geliştirilmiştir. Rassal Ormanlar (RF) veya Karar Ağacı Ormanları olarak bilinen yöntem, sadece karar ağaçlarının topluluklarına odaklanan denetimli bir makine öğrenmesi yöntemidir. Bu yöntem, tek bir karar ağacının değil, birden çok karar ağacının bir araya gelmesiyle oluşmaktadır (Breiman, 2001: 5). Rassal ormanlar, bir yanıt hesaplayabilmek için birden fazla farklı modellerden elde edilen sonuçların

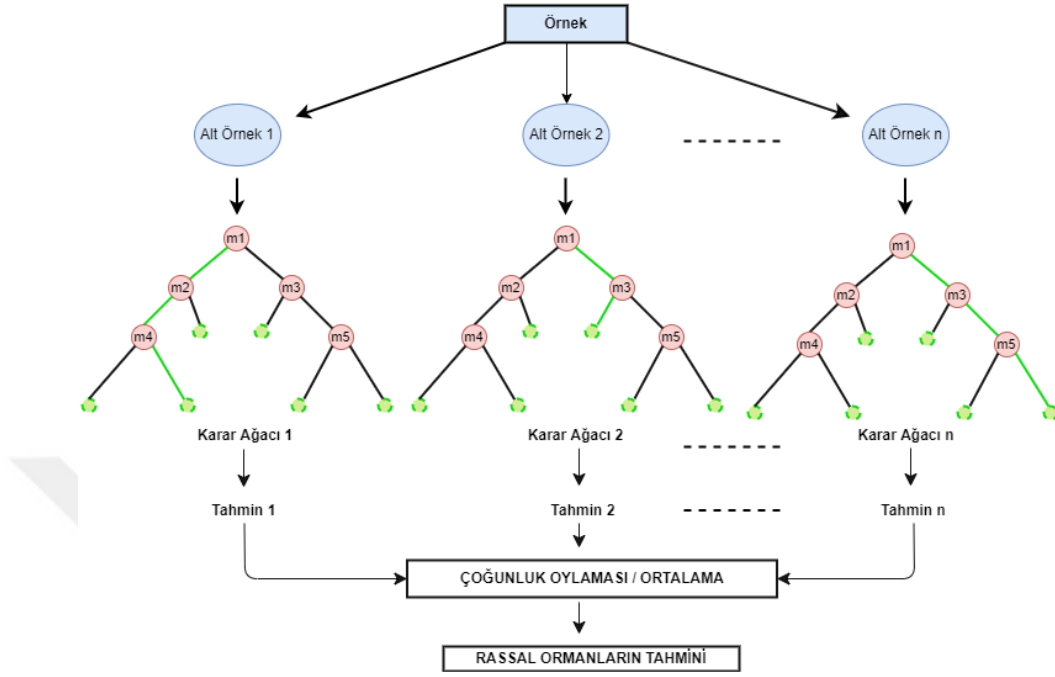
kullanıldığı bir topluluk öğrenme modelidir. Genel olarak, çeşitli ağaçların kullanıldığı tahminlerden elde edilen sonuçlar, yalnız bir ağacın tahmininden daha iyi sonuç vermektedir (Horning, 2010: 2). Rassal ormanlar algoritması, sadece sınıflandırma problemlerinde değil, aynı zamanda regresyon problemlerinde de etkili bir şekilde kullanılmaktadır. Sınıflandırma problemlerindeki performansı göz önünde bulundurulduğunda, regresyon problemlerine göre daha iyi olduğu söylenebilmektedir. Rassal orman yöntemi, özyinelemeli birleştirme (Bagging) ile Sınıflandırma ve Regresyon Ağacı (CART) yöntemlerinin birleştirilmesiyle oluşmaktadır. Rassal ormanlarda bulunan her bir karar ağacı, özyinelemeli yönteminin yardımıyla orijinal veri kümesinden örnekler seçilmektedir. Her bir karar düğümünde yer alan bütün değişkenler arasından istenilene uygun olarak rassal değişkenlerin seçilmesi ile meydana gelmektedir. Sınıflandırma ve regresyon ağacı yöntemleri ile ağaçlar oluşturulmaktadır ve bu ağaçlarda budama işlemi yapılmamaktadır. (Breiman, 2011; Akman ve diğerleri, 2011: 37).

Rassal ormanlar algoritmasında, özyinelemeli birleştirme yönteminin kullanılmasının iki önemli nedeni bulunmaktadır. İlki, rassal öznitelikler seçimi kullanıldığında özyinelemeli birleştirme kullanımının doğruluğunu artırdığı görülmektedir. İkincisi, özyinelemeli birleştirmenin modelin iç hatasının Out Of Bag (OOB) hesaplanmasıdır (Breiman, 2011: 11). Ayrıca, rassal ormanlar algoritmasının Bagging yönteminden ayıran en önemli nokta rassal değişkenin seçilmesidir. Sözgelimi, m adet örneğin bulunduğu açıklayıcı değişkenlerden rassal olarak p adet örnek seçildiğinde; açıklayıcı değişkenlerin sayısı, seçilen rassal örnek sayısından fazla ise rassal ormanlar algoritmasının, sınıflandırma hatası ve modelin iç hatasının minimize edilmesi sağlanabilir (Gareth ve diğerleri, 2013: 320). Rassal ormanlar sınıflandırıcısı ile bir ağacın oluşturulması için kullanıcı tarafından iki parametrenin tanımlanmış olması gerekmektedir. Bunlar, en iyi bölünmenin gerçekleştirilmesi için her bir düğümde yer alan rassal değişkenlerin sayısı (m) ve geliştirilecek olan ağaç sayısı (N) olarak ifade edilmektedir (Pal, 2005: 218). Bir ağaçta bulunan karar düğümleri başına rassal seçilen özniteliklerin sayısı orman sınıflandırmasının hata oranını belirlemektedir. Rassal ormanlar sınıflandırıcısının hata oranı, herhangi iki araç arasındaki bağlantıya ve her bir ağaçtaki sınıflandırma gücüne dayanmaktadır. Rassal olarak seçilen özniteliklerin azaltılması, sınıflandırma ağaçlarındaki

korelasyonu ve her bir ağaçtaki sınıflandırma gücünün azalmasına sebep olmaktadır. Benzer şekilde, rassal olarak seçilen özniteliklerin artırılmasında ise, sınıflandırma ağaçlarındaki korelasyon ve ağaçtaki sınıflandırma gücünün artmasına sebep olmaktadır. Sınıflandırıcının veri kümesinin ne derecede iyi performans gösterip göstermediği Out-Of-Bag (OBB) hata oranı ile gösterilmektedir. Alt küme eğitim veri kümesi, modelin eğitilmesi için $2/3$ 'lük (inBag) bölümü, modelin test edilmesi için ise $1/3$ 'lük bölümü (Out Of Bag) ve model içerisinde bulunan iç hata oranının (OOB Hata Oranı) belirlenmesi amacıyla kullanılmaktadır. Diğer bir ifadeyle, out-of-bag model, her bir ağaç için önyükleme örneğinden k . Ağacın oluşturulması için veri kümesi içerisinde $1/3$ 'ü (Out Of Bag) dışarıda bırakılır. Aynı zamanda bu $1/3$ 'lük örnek, tüm ağacın ortalamasının alınmasıyla yanlış sınıflandırma sonuçlarını ve k . Ağacın test edilmesi amacıyla kullanılmaktadır (Alam ve Vuong, 2013: 664).

Modelin iç hata oranı, algoritmadaki problemin çözümlenebilmesi için yeterli kadar karar ağaçlarının bulunup bulunmadığının belirlenmesi amacıyla kullanılmaktadır. Oluşturulan her bir ağaca başlangıçta hesaplanan OOB hata oranını göz önünde bulundurularak her bir ağaç için ağırlık değeri verilmektedir. Ağaçtaki her bir ağırlık değerleri ile OOB hata oranları arasında ters yönlü bir ilişki vardır. Hata oranı OOB en yüksek olan ağaçlar düşük bir ağırlık değerine sahipken, OOB hata oranı en düşük olan ağaçlar ise yüksek bir ağırlık değerine sahiplerdir. Tanımlanan ağırlık değerlerine göre bütün ağaçların sınıf tahmini işleminin gerçekleştirilmesi için bir oylamaya yönteminden yararlanılmaktadır. Oylama işlemleri arasından en yüksek oy değerini sahip ağaç, sınıf tahmini olarak seçilmektedir (Akman ve diğerleri, 2011: 37).

Şekil 18: Rassal Ormanlar Algoritmasının Genel Yapısı



Kaynak: Dmitrievski, 2018.

Tüm algoritmalarda olduğu gibi rassal ormanlar algoritmasında da avantajlar ve dezavantajlar bulunmaktadır. Avantajları; eğitim verileri ile test verileri arasındaki ilişkiden dolayı performansı zayıf bir sınıflandırıcı ile karşılaşma ihtimalini azaltmaktadır. Ayrıca, son derece esnek ve yüksek bir doğruluğa sahiptir. Eksik verilerin olması durumunda da verilerin büyük bir çoğunluğu doğruluğunu korumaktadır. Son olarak, rassal ormanlar tek bir karar ağacıyla karşılaştırıldığında daha az varyansa sahiptirler. Dezavantajları ise şu şekildedir; rassal ormanların oluşturulması, karar ağaçlarına göre çok daha uğraştırıcı ve zaman alıcıdır. Hesaplama açısından çok fazla kaynağa ihtiyaç duymaktadır (NewTechDojo, 2017).

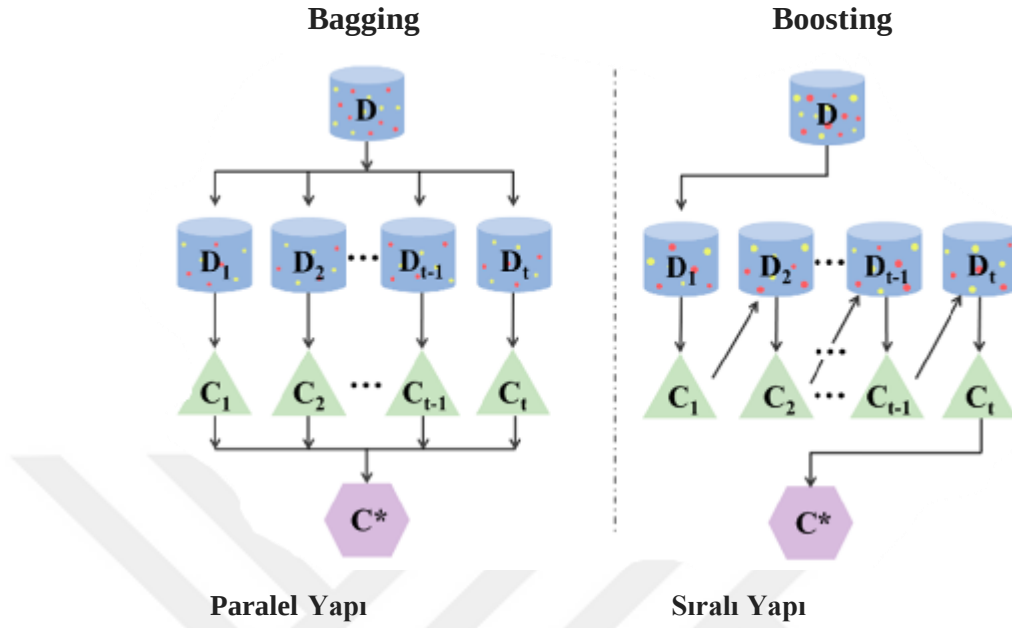
Topluluk öğrenmesi tekniklerinin sınıflandırılmasında kullanılan ikinci teknik ise Artırma (Boosting) sınıflandırma yöntemi; Artırma yöntemi, 1989'da Schapire tarafından geliştirilmiş bir yöntemdir. Bu yöntem, 1997'de Freund ve Schapire tarafından daha da geliştirilmiş ve 2001 yılında Friedman'ın katkılarıyla yapılan son çalışmalar orijinal gelişmeyi bir adım daha ileri taşımıştır. Artırma algoritması, zayıf

veya temel olarak adlandırılan tekil sınıflandırıcıların doğruluğunu artırmak amacıyla kullanılan bir topluluk öğrenme yöntemidir. Bu algoritma, zayıf (temel) öğrenciler tarafından oluşturulan yanlış öngörü yapılan örneklerin hatalarının giderilmesi amacıyla kullanılmaktadır. Artırma yöntemi, özyinelemeli birleştirme (Bagging) yöntemi ile benzerlik göstermektedir. özyinelemeli birleştirme yönteminde, her bir yeni örneğin oluşturulmasında bulunan her gözlemin seçilme olasılığı eşittir. Ancak Artırma yöntemi, örneğe sonradan dahil edilen her bir gözlemin seçilmesinde farklı olasılıkları kullanmaktadır (Džeroski ve Ženko, 2004: 255; De Ville, 2006: 167).

Boosting yöntemi, her artırma iterasyonunda, temel bir sınıflandırıcı oluşturmakta ve yanlış sınıflandırılan vektörler bağlı olarak çeşitli örnek dağılımı yardımıyla eğitilmektedir. Yanlış sınıflandırılan sınıflandırıcılar daha sonra ağırlık oylama kullanılarak yeni bir son (nihai) sınıflandırıcıda birleştirilmektedir. Böylece, sınıflandırıcının performansının artmasına yardımcı olmaktadır. Ayrıca, yanlış sınıflandırıcının minimum düzeyde kalmasını sağlamaktadır (Ramalho ve Medeiros, 2006: 2). Diğer bir ifadeyle, sınıflandırıcıların birbiri ardından eğitilmesiyle, zayıf öğrencilerin (weak learner) güçlü bir öğrenci (strong learner) hâline gelmesi amacıyla iteratif olarak gerçekleştirilmektedir. Bu işlemlerin uygulanmasında, örnekler bir ağırlık değeri verilmektedir. Bu ağırlık değeri, temel öğrencilerin tahminleme başarısına göre revize edilmektedir. Sonuç olarak, başarı istenilen düzeye ulaştığında nihai tahmin elde edilir ve işlem sonlandırılmaktadır. Boosting yöntemine dayalı algoritmalar; Adaboost, Gradient Boosting, XGboost gibi algoritmalar sıklıkla tercih edilmektedir. Bu tez kapsamında, Gradient Boosting algoritmasına değinilmiştir.

Bagging ve Boosting algoritmalarının oluşturulması: Bagging algoritmasında her bir öğrenci birbirinden bağımsız ve paralel bir yapıya sahipken, Boosting algoritmasında ise yeni öğrenciler birbirini izleyen bir yapıya sahiptir. Bagging ve Boosting algoritmasının işleyişi Şekil 20’de gösterildiği gibidir.

Şekil 19: Bagging ve Boosting Algoritma Yapısı Örneği



Kaynak: Byrne ve diğerleri, 2019: 10530.

2.3.2. Gradyan Artırma Karar Ağaçları

Gradient Boosting Algoritması, 2001’de Friedman tarafından geliştirilmiş olan makine öğrenmesi yöntemidir. Gradient Boosting algoritması, karar ağaçları başta olmak üzere, zayıf tahmin modellerinin birleştirilmesiyle güçlü bir tahmin modeli üretilmektedir. Bu algoritma, sınıflandırma ve regresyon problemlerinin çözümünde kullanılan bir makine öğrenme yöntemidir. Gradient Boosting algoritması, Boosting yöntemine dayalı ve çalışma prensipleri açısından benzerlik göstermektedir. Farklı temel öğreniciler sıralı olarak oluşturulur ve kompleks bir modelde birleştirilmektedir. Genel olarak, Gradient Boosting yöntemi, karar ağaçları ile birlikte kullanılır ve bunlar “Gradyan Artırma Karar Ağaçları (Gradient Boosted Decision Trees / GBDT)” olarak adlandırılmaktadır. Gradyan Artırma Karar Ağaçları, belirli bir veri setindeki optimal parametrelerin tahmini için kayıp fonksiyon (loss function) ve düzenleme (regularization) kullanılmaktadır. Model yapısındaki karmaşıklığın ve aşırı öğrenme (overfitting) sorununun minimum düzeye indirilmesi amacıyla düzenleme terimi kullanılmaktadır. Kayıp fonksiyonu ise, gerçek çıktı değeri ile tahminden elde edilen

çıktı değeri arasındaki farkın hesaplanmasına dayanmaktadır. Ayrıca, kayıp fonksiyonları türevlenebilir özelliktedir (Marksfeld, 2018: 9). Yanıt değişkenlerine bağlı olarak sınıflandırılmaktadır. Regresyon problemlerinde; Gaussian L_2 (en küçük kareler hatası), Laplace L_1 (ortalama mutlak sapmalar) ve Huber gibi kayıp fonksiyonları sürekli değişkenler için sıklıkla tercih edilirken, sınıflandırma problemlerinde; Binom, Üstel ve Adaboost gibi kayıp fonksiyonları kategorik değişkenler için yaygın olarak tercih edilmektedir (Natekin ve Knoll, 2013: 4). Kayıp fonksiyon değerinin yüksek olması, model doğruluğunun başarısını azaltırken, kayıp fonksiyon değerinin düşük olması, model doğruluğunun başarısını artırmaktadır.

Makine öğrenmesi yöntemlerinin oluşturulmasında verilerden öğrenme modeline ilişkin en mühim problem, nihai modelin genelleme becerisidir. Burada, GBDT algoritması, çeşitli sayıda ağacın bir araya gelmesiyle oluşmasından dolayı veri kümesinin eğitim için ayrılan verilerin fazla uyum göstermesi durumunda, aşırı öğrenme (overfitting) riski ile karşılaşmasını artırmaktadır. Aşırı öğrenmeden kaynaklı eğitimde karşılaşılan hataların daha da artmasına sebep olmaktadır. Aşırı öğrenme etkilerini azaltmak için alt örnekleme (subsampling), daralma / küçülme (shrinkage) ve erken durdurma (early stopping) gibi çeşitli düzenleme (regularization) yöntemleri kullanılmaktadır (Natekin ve Knoll, 2013: 8).

Tüm algoritmalarda olduğu gibi GBDT algoritmasının avantajları ve dezavantajları bulunmaktadır. Avantajları; veri ön işleme süreci gerektirmez ve eksik verilerin olması algoritmanın çalışmasını engellememektedir. Bu algoritma, çeşitli kayıp fonksiyonları optimize etme özelliğine sahip ve farklı hiper-parametre (Hyperparameter) yardımıyla kullanılan fonksiyonlar daha esnek bir hâle getirilmektedir. GBDT algoritması, sınıflandırma ve regresyon problemleri için etkili bir şekilde çalışmaktadır. Böylece, tahmin doğruluğunun daha yüksek olması sağlanmaktadır. Dezavantajları ise GBDT Modelleri, çeşitli sayıda ağaç (> 1000) kullanımını gerektirmektedir. Ağaç sayısının yüksek olması farklı sorunların ortaya çıkmasına neden olmaktadır. Bunlar; zaman, hafıza (bellek) kullanımı ve hesaplama maliyetlerinin artmasına sebebiyet vermektedir. Ayrıca, çok sayıda ağaçların ardışık bir şekilde oluşturulduğu göz önünde bulundurulduğunda eğitim süresi uzamaktadır. Verilerin gürültülü olması, tahminleme aşamasında hataların oluşmasına sebep olacaktır (Kumar, 2020).

2.3.3. Ekstra Ağaçlar Sınıflandırıcısı

Aşırı Rassallaştırılmış Ağaçlar (Extremely Randomized Trees) olarak bilinen Ekstra Ağaçlar Algoritması, alışılmış olan yukarıdan başlayan ve aşağıya belirlenen yöntemle göre, budanmamış karar ağaçlarından veya regresyon ağaçlarından meydana gelen bir topluluk oluşturulmaktadır. Diğer ağaç tabanlı yöntemlerle kıyaslandığında iki önemli farkı bulunmaktadır. Bunlardan ilki, kesim noktaları tamamen rassal seçilerek düğümlere ayrılmasıdır. İkincisi, ağacın büyütülmesi için tüm öğrenme örneğinin kullanılmasıdır. Diğer bir deyişle, önyükleme kopyası yapılmamaktadır. Ekstra ağaçların tahminlerinde, sınıflandırma problemleri için çoğunluk oyu kullanılırken, regresyon problemlerinde ise aritmetik ortalama ile nihai tahmin elde edilecek şekilde bir araya getirilmektedir. Ayrıca, ekstra ağaçlar varyans ve yanlılık açısından değerlendirildiğinde, özyinelemeli kopyalarını kullanmak yerine tüm öğrenme örneğinin kullanılması, yanlılığın minimum düzeyde olmasını sağlayacaktır. Buna ek olarak, rassal değişken seçimi yöntemine dayanan ekstra ağaçlar yöntemi, varyansın azaltılmasında kullanılan rassallaştırma tekniği, başka modellerde yer alan rassallaştırma tekniklerden daha etkin sonuçlar vermektedir (Geurts ve diğerleri, 2006: 5-6).

2.3.4. Oylama Sınıflandırıcısı

Topluluk oylama sınıflandırıcısı, çoğunluk oylama yoluyla sınıflandırma benzeri olan ve farklı makine öğrenimi sınıflandırıcılarından yararlanılan bir meta sınıflandırıcıdır. Bu farklı makine öğrenimindeki sınıflandırma modellerinin oluşturulmasında eğitim veri setleri ile modeller eğitilmektedir. Eğitilen farklı her bir model için test setleri yardımıyla test edilmektedir. Kullanılan her bir modelin sonucundan elde edilen tahmin değerleri bir araya getirilmektedir. Bu toplanan tahmin değerleri, sonradan dahil edilen gözlemlerin sınıf etiketlerinin tahmini için oylamaya dayalı olarak gerçekleştirilmektedir. Kısaca burada yapılan işlemler:

Eğitim veri kümesi girdi olarak alınır ve sınıflandırma modelleri ($C_1, C_2, \dots, C_m$) şeklinde oluşturulmaktadır. Ardından, her bir sınıflandırıcılardan elde edilen tahmin değerleri ($P_1, P_2, \dots, P_m$) alınır ve oylama işlemi gerçekleştirilir. Son olarak, oylama işleminin tamamlanmasıyla nihai olasılık (P_f) değeri elde edilmektedir (Mahabub ve Habib, 2019: 3).

2.3.4.1. Sert Oylama (Hard Voting)

Sert Oylama, çoğunluk oylamasının en basit hâlidir. Burada, en yüksek oy değerine sahip olan sınıf, sınıflandırıcının kullanacağı son tahmin değeri olarak alınır. Sınıf etiketi $\hat{y}$, her sınıflandırıcı ise C_j 'dir. Sert oylama, 2.23'te verilen eşitlikteki gibidir (Mahabub ve diğerleri, 2019: 808).

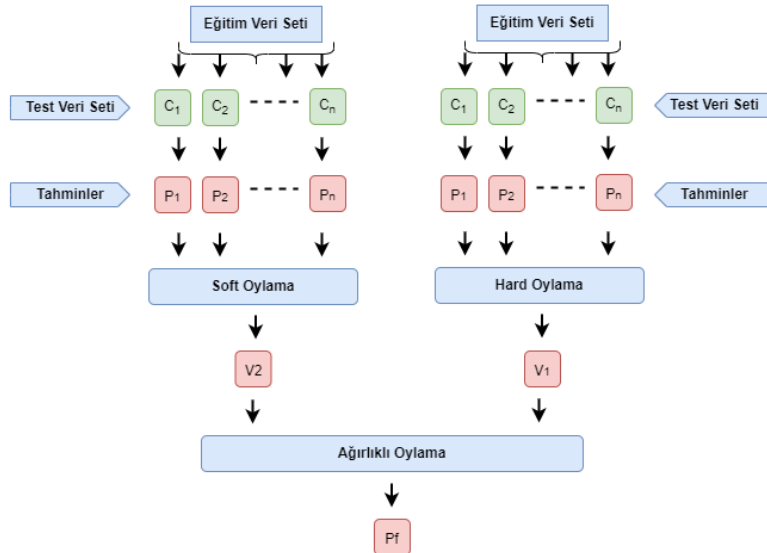
$$\hat{y} = \text{mod} \{ C_1(x), C_2(x), \dots, C_m(x) \} \quad (2.23)$$

2.3.4.2. Yumuşak Oylama (Soft Voting)

Yumuşak oylamada, her sınıf etiketleri, sınıflandırıcının tahmin edilen olasılık değerleri P 'ye bağlı olarak tahmin edilmektedir. Tahminler, her bir modelden elden edilen tahminlerin olasılık değerleri toplanmakta ve bu toplanan değer arasından maksimum olasılık değerine sahip olan sınıf etiketi nihai tahmin olarak seçilmektedir. Burada, sınıflandırıcıların doğru bir şekilde ölçülmesi durumunda tercih edilmektedir. Yumuşak oylama, 2.24'te verilen eşitlikle hesaplanmaktadır (Mahabub, 2020: 4).

$$\hat{y} = \arg \max_i \sum_{j=1}^m W_j P_{ij} \quad (2.24)$$

Şekil 20: Oylama Sınıflandırıcısı Soft ve Hard Yapısı Örneği



Kaynak: Rustam ve diğerleri, 2020: 7.

2.4. PERFORMANS DEĞERLENDİRME ÖLÇÜTLERİ

Sınıflandırma problemlerine ait en optimal çözümü gerçekleştiren modelin belirlenebilmesi için farklı performans değerlendirme ölçütleri kullanılmaktadır. Bu doğrultuda, makine öğrenmesi algoritmalarının tahmin performanslarının değerlendirilmesinde doğruluk, hassasiyet, kesinlik, F1-Skor ve MCC değerleri yaygın olarak kullanılmaktadır. Dengeli yapıya sahip veri kümeleri için yalnızca doğruluk performans ölçüm istatistiği sonucuna bakmak en iyi model seçimini belirlemede yeterli olmaktadır. Ancak bu durum dengeli olmayan veri kümeleri için yeterli olmamaktadır. Bu nedenle, dengesiz veri seti söz konusu iken en iyi model seçiminde literatürde sıklıkla F1-skoru ve MCC performans ölçütleri kullanılmaktadır. Bu tez kapsamında kullanılan makine öğrenmesi performans ölçütleri alt bölümde detaylandırılmıştır.

2.4.1. Karmaşıklık Matrisi (Confusion Matrix)

Makine öğrenmesinde, karışıklık veya hata matrisi olarak adlandırılan karmaşıklık matrisi, ikili (binary) sınıflandırma problemleri performanslarının değerlendirilmesi amacıyla yaygın bir şekilde kullanılmaktadır. Karmaşıklık matrisi, modelin sınıflandırma sonuçlarını gösteren gerçek sonuç değerleri ile tahminlenen sonuç değerlerini içermektedir. Burada, sınıflandırma modellerine ait doğru ve yanlış tahmin edilen değerlerini gösteren bir matris söz konusudur. Karmaşıklık matrisinde, satır bazında gerçek sonuçların değerleri yer alırken, sütun bazında ise tahmin sonuçları yer almaktadır. Bahsi geçen matrise ilişkin yapı Tablo 1’de gösterildiği gibidir. Burada örneklendirilen yapı 2x2 boyutundadır. Genel olarak 2x2 boyutundaki matris kullanılmakla birlikte, hedef değişkenine göre matris boyutu değişkenlik göstermektedir (Feng ve diğerleri, 2020: 732).

Tablo 1: Karmaşıklık Matrisi (Confusion Matrix) Yapısı

KARMAŞIKLIK MATRİSİ		TAHMİN SONUÇLAR	
		POZİTİF	NEGATİF
GERÇEK SONUÇLAR	POZİTİF	Doğru Pozitif (TP)	Yanlış Negatif (FN) (Tip II Hata)
	NEGATİF	Yanlış Pozitif (FP) (Tip I Hata)	Doğru Negatif (TN)

Kaynak: Feng ve diğerleri, 2020: 733.

Doğru Pozitif: Bir örneğin pozitif olduğuna ilişkin doğru tahminlerin sayısını ifade etmektedir. Gerçekte pozitif olan bir sınıfın, doğru olarak pozitif tahmin edildiğini göstermektedir.

Örneğin; gerçek sınıf kredi alabilir ve sınıflandırıcı doğru bir şekilde kredi alabilir olarak tahmin etmesi durumudur.

Doğru Negatif: Bir örneğin negatif olduğuna ilişkin doğru tahminlerin sayısını ifade etmektedir. Gerçekte negatif olan bir sınıfın, doğru olarak negatif tahmin edildiğini göstermektedir.

Örneğin; gerçek sınıf kredi alamaz ve sınıflandırıcı doğru bir şekilde negatif olarak tahmin etmektedir.

Yanlış Pozitif (False Positive veya FP): Bir örneğin pozitif olduğuna ilişkin yanlış tahminlerin sayısını ifade etmektedir. Gerçekte negatif olan bir sınıfın, yanlış olarak pozitif tahmin edildiğini göstermektedir.

Örneğin; gerçek sınıf kredi alamaz. Ancak sınıflandırıcı yanlış bir şekilde kredi alabilir olarak tahmin etmektedir.

Yanlış Negatif (False Negative veya FN): Bir örneğin negatif olduğuna ilişkin yanlış tahminlerin sayısını ifade etmektedir. Gerçekte pozitif olan bir sınıfın, yanlış olarak negatif tahmin edildiğini göstermektedir.

Örneğin; verilen sınıf kredi alabilir. Ancak sınıflandırıcı yanlış bir şekilde kredi alamaz olarak tahmin etmektedir (Data Science, 2019; Chicco ve Jurman, 2020: 4; Zeng, 2020: 2083).

Tip I Hata: Model tahmini pozitif, kişi kredi alabilir ve kişinin gerçek durumuna bakılırsa negatif ve kredi alamaz şeklindedir. Burada model pozitif olarak tahmin edilmiş fakat bu yanlıştır. Gerçekleşen bu durum, hipotez testlerinde Tip I Hata olarak adlandırılmaktadır.

Tip II Hata: Model tahmini negatif, kişi kredi alamaz ve kişinin gerçek durumuna bakılırsa pozitif ve kredi alabilir şeklindedir. Burada model negatif olarak tahmin edilmiş fakat bu yanlıştır. Gerçekleşen bu durum, hipotez testlerinde Tip II Hata olarak adlandırılmaktadır (Amanet, 2020).

2.4.2. Doğruluk

Doğruluk, doğru bir biçimde sınıflandırılmış olan veri ögelerinin, tüm veri ögelerine oranı olarak ifade edilen bir performans değerlendirme ölçütüdür. Bu metrik, sıklıkla tercih edilmektedir ancak modelin performansı ile ilgili detaylı bilgi vermemektedir. Model başarısının değerlendirilmesinde tek kriter olarak doğruluk oranı yetersiz olabilir. Genel olarak tercih edilmesine rağmen, bazı durumlarda, veri setindeki yanıt değişkene ait sınıflarının dengeli olmadığı durumlarda, sınıflandırma performans değerlendirilmesi için uygun değildir (Vakili ve diğerleri, 2020: 5). Örneğin; kredi alabilen ve kredi alamayan 1000 müşteri olduğunu varsayalım. 1000 kişi içerisinde yalnızca 100 kişinin kredi alabilen olduğunu varsayımı altında, kredi alabilen ancak kredi almasının uygun bulunmadığı (Yanlış Negatif) müşterilerin olması istenmemektedir. Burada, kredi alamayan kişilerin sayısı, kredi alabilen kişilerden daha fazla olduğu için model, kredi alabilen müşterilere göre daha çok doğru sınıflandırma yapacak ve modeldeki doğruluğu artacaktır. Sonuç olarak, bir sınıfın sayısının fazla olması doğruluk oranını artırmakta ancak tek ölçüt olarak değerlendirilmesi modelle ilgili yanılgıya neden olmaktadır.

Doğruluk ölçütü, matematiksel olarak Eşitlik (2.25)'te gösterildiği gibidir (Vakili ve diğerleri, 2020: 5).

$$\text{Doğruluk (Accuracy)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.25)$$

2.4.3. Kesinlik (Precision / Positive Predictive Value)

Model kesinlik ölçütü, model tarafından tahmin edilen bütün pozitif tahminler içerisinde, pozitiflerin gerçekte doğru olan kısmını tahmin etme yeteneğini göstermektedir. Kesinlik, pozitif ve negatif örneklerden elde edilen sonuçları birleştirir ve bu nedenle yanlış pozitif değerleri kesinlik puanı üzerinde etkili olacaktır. Bu etkiyi azaltmak ve modelin daha yüksek kesinlik değerine sahip olması için yanlış pozitif değerlerin minimum tutulması gerekmektedir. Böylece, doğru pozitiflerin değeri artmakta ve kesinlik değeri daha yüksek sonuç vermektedir. Kesinlik ölçütü, dengeli olmayan veri setlerinin değerlendirilmesinde anlamlı bir ölçüt olarak kullanılmaktadır (Bradley ve diğerleri, 2006: 2; Kumar, 2022). Kesinlik ölçütü, Eşitlik (2.26)'da gösterildiği gibidir.

$$\text{Kesinlik (Precision)} = \frac{TP \text{ (Doğru Pozitif)}}{TP \text{ (Doğru Pozitif)} + FP \text{ (Yanlış Pozitif)}}$$

(2.26)

2.4.4. Hassasiyet (Recall / Sensitivity)

Model hassasiyet ölçütü, sınıflandırıcının gerçek pozitifleri doğru bir biçimde tahmin etme yeteneğini göstermektedir. Burada, var olan veri kümesi içerisinde yer alan bütün pozitif değerlerden, gerçekte pozitif olan tüm değerlerin tespit edilmesinde sınıflandırıcının ne ölçüde iyi performans gösterdiğini ölçmeye yardımcı olmaktadır. Hassasiyet değerinin yüksek olması, yanlış negatif değerlerinin düşük olmasını gerektirir. Hassasiyet ölçütü, dengeli olmayan veri kümelerinin tahminleme başarısının değerlendirilmesinde anlamlı bir ölçüt olarak kullanılmaktadır (Kumar, 2022). Hassasiyet ölçütü, Eşitlik (2.27)'de gösterildiği gibidir.

$$\text{Hassasiyet (Recall)} = \frac{TP \text{ (Doğru Pozitif)}}{TP \text{ (Doğru Pozitif)} + FN \text{ (Yanlış Negatif)}}$$

(2.27)

2.4.5. F1 Skoru (F1 Score)

F1 skoru, veri kümesindeki model doğruluğunun belirlenmesi amacıyla kullanılan bir ölçüttür. Pozitif veya negatif olarak sınıflandırılmış olan ikili (binary) sınıflandırma problemlerinin yorumlanması için kullanılmaktadır. F1 skoru, hassasiyet ve kesinlik istatistiklerinin birleşmesiyle oluşur ve aynı zamanda, harmonik ortalaması olarak da tanımlanmaktadır. F1 skoru, yanlış pozitifler ile yanlış negatifleri göz önünde bulundurmaktadır. Bu ölçüt, sınıf dengesizliği olduğu durumlarda, doğruluk ile karşılaştırıldığında daha kullanışlıdır. Böylece, F1 skoru modelin doğru seçilmesi açısından etkilidir (Wood ve Joshi, 2016).

F1 skoru [0,1] aralığındadır. F1 skoru, Eşitlik (2.28)'de gösterildiği gibidir (Chicco ve Jurman, 2020: 5).

$$F_1 \text{ Skoru} = \frac{2.TP}{2.TP + FP + FN} = 2 * \frac{\text{kesinlik} * \text{hassasiyet}}{\text{kesinlik} + \text{hassasiyet}} \quad (2.28)$$

2.4.6. Matthews Korelasyon Katsayısı

Veri setlerinin dengeli olmadığı durumlarda tercih edilen ölçütlerden biri de Matthews Korelasyon Katsayısı (MCC)'dir. Bu ölçüt, gerçek değerler ile tahmin edilen değerler arasındaki korelasyonun bir ölçüsü olarak ifade edilmektedir. MCC ölçütü, [-1, +1] arasında değerler almaktadır. Burada, MCC = -1 değeri, gerçek değerler ile tahmin edilen değer arasındaki negatif yönde bir ilişki olduğunu gösterir. Bu ilişki, sınıflandırma modelinin tamamen yanlış tahmin edildiğini ifade eder. Aynı zamanda, MCC'nin minimum değeri -1'dir. MCC = 0 değeri, yapılan tahminlerin rassal olarak oluşturulduğu anlamı taşımaktadır. Son olarak, MCC = 1 değeri, gerçek değerler ile tahmin edilen değer arasındaki pozitif yönde bir ilişki olduğunu gösterir. Bu ilişki, sınıflandırma modelinin tamamen doğru tahminler ürettiği anlamını taşır. (Boughorbel ve diğerleri, 2017: 6; Chicco ve diğerleri, 2021: 78371; Chicco ve Jurman, 2020: 5).

MCC ölçütü, Eşitlik (2.29)'da gösterildiği gibidir (Al-Abbasi ve diğerleri, 2020: 3126).

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP). (TP + FN). (TN + FP). (TN + FN)}}$$

(2.29)

2.5. YENİDEN ÖRNEKLEME TEKNİKLERİ

Veri kümelerindeki sınıfların, çeşitli özniteliklerine göre farklı sayıda gözlem içermesi durumu “dengesiz veri kümeleri” olarak adlandırılır. Dengesiz veri kümelerini dengeli hâle getirebilmek için yeniden örnekleme teknikleri kullanılmaktadır. Bu örnekleme teknikleri ile analizler daha doğru ve güvenilir bir sonuç vermektedir. Basit bir örnekle açıklamak gerekirse; 15000 örneğin bulunduğu bir veri seti olduğunu varsayalım. Bu veri seti içerisinde 10000 adet gözlem kredi alabilir, 5000 adet gözlem kredi alamaz olduğu varsayımı altında, sayıca fazla olan sınıf, çoğunluk sınıfı olarak adlandırılırken, görece daha az olan sınıf, azınlık sınıfı olarak adlandırılmaktadır. Veri setlerinin dengeli olmama durumu çeşitlik probleme yol açmaktadır. Sözgelimi, ikili sınıflandırma problemlerinde, sayıca daha fazla olan sınıf göz önünde bulundurularak tahmin yapıldığı için sayıca az olan sınıf göz ardı edilir ve böylece, çoğunluk sınıfı, daha çok doğru sınıf olarak tahmin edilir. Aslında, bu gerçek durumu yansıtmamaktadır. Ayrıca, optimal model seçiminde yanılgıya neden olmaktadır. Bu tür problemlerin oluşmaması için yeniden örnekleri teknikleri ile veri kümelerinin dengeli hâle getirilmesi amaçlanır. Bu doğrultuda, Aşağı Örnekleme (Undersampling), Yukarı Örnekleme (Oversampling), Aşağı ve Yukarı Örnekleme Kombinasyonları ve Örnekleyiciler Topluluğu gibi teknikler kullanılarak, dengesiz veri kümelerinin dengeli veri kümelerine dönüştürülmesi bu çalışma kapsamında incelenmiştir.

2.5.1. Aşağı Örnekleme

Aşağı Örnekleme, dengeli olmayan veri kümelerinde bulunan sınıf dengesizliği probleminin önüne geçebilmek ve sınıfların dengeli bir hâle getirilmesi için kullanılan bir yöntemdir. Bu yöntemde, azınlık sınıfının sayısı ile çoğunluk sınıfının sayısı eşitlenmesi amaçlanmaktadır. Aşağı Örnekleme yöntemi, sınıfların dengelenmesi için veri kümesindeki çoğunluk sınıfında yer alan örneklerin sayısının azaltılmasında,

azınlık sınıfında yer alan örneklerin sayısını göz önünde bulundurur ve bu sayı eşitleninceye kadar örnek eksiltme işlemleri gerçekleştirilir. Böylece, veri kümeleri dengeli hâle getirilmiş olacaktır. Bu işlemler farklı Aşağı Örneklemeye yöntemleri ile gerçekleştirilse de en yaygın olarak Rassal Aşağı Örnekleme tekniği kullanılmaktadır. Aşağı örneklemeye yöntemi Şekil (21a)'daki gibi görselleştirilebilir.

2.5.1.1. Rassal Aşağı Örneklemeye (Random Undersampling - RUS)

Aşağı Örneklemeye yöntemine dayalı olan Rassal Aşağı Örneklemeye (RUS) yöntemi, dengeli olmayan veri kümelerinde, çoğunluk sınıfında yer alan örneklerin bir bölümü azaltılarak, azınlık sınıfındaki örneklerin sayısı ile dengelemeyi amaçlayan bir yöntemdir. Eğitim veri kümesindeki çoğunluk sınıfı örnekleri, azınlık ve çoğunluk sınıfı arasında belirlenen düzeye ulaşmaya kadar devam etmektedir. Rassal olarak yapılan eksiltme işleminin yol açtığı en önemli problemlerden biri, eksiltelen çoğunluk sınıfına ilişkin gerekli bilgilerin atılması söz konusu olabilir ve bu durum iki sınıf arasında, karar için önemli olan detaylara ulaşılmasının önüne geçebilir. Böylece, sınıflandırmadaki model, eğitim verileri ile uyum sağlayamayarak var olan ilişkiyi ortaya çıkaramayabilir. Bu durum, eğitim verileri ve test verileri arasında hatanın meydana gelmesine neden olmaktadır. Bahsi geçen bu olumsuzluklara rağmen, RUS yöntemi, basit ve kolay uygulanabilirliği açısından sıklıkla kullanılmaktadır (Alam ve diğerleri, 2020: 201179; Ali ve diğerleri, 2019: 392).

2.5.2. Yukarı Örneklemeye

Yukarı Örneklemeye yöntemleri, Aşağı Örneklemeye yöntemlerinde olduğu gibi dengeli olmayan veri kümelerinde bulunan sınıf dengesizliği probleminin önüne geçebilmek ve sınıfların dengeli bir hâle getirilmek için kullanılan bir yöntemdir. Bu yöntemde, azınlık sınıfının sayısı ile çoğunluk sınıfının sayısının eşitlenmesi amaçlanmaktadır. Yukarı Örneklemeye yöntemi, sınıfların dengelenmesi için veri kümesindeki azınlık sınıfında yer alan örneklerin sayısının artırılmasında, çoğunluk sınıfındaki yer alan örneklerin sayısı göz önünde bulundurulur ve iki sınıf eşitleninceye kadar örnek ekleme işlemleri gerçekleştirilir. Böylece, veri kümeleri dengeli hâle getirilmiş olacaktır. Bu işlemler farklı Yukarı Örneklemeye farklı yöntemler

ile gerçekleştirilse de en yaygın olarak Rassal Yukarı Örneklem tekniği kullanılmaktadır. Yukarı Örneklem yöntemi Şekil (21b)'deki gibi görselleştirilebilir.

2.5.2.1. Rassal Yukarı Örneklem (Random Oversampling - ROS)

Yukarı Örneklem yöntemine dayalı olan Rassal Yukarı Örneklem (ROS) yöntemi, dengeli olmayan veri kümelerinde, azınlık sınıfında yer alan örneklerin rassal olarak çoğaltılması ile çoğunluk sınıfı ve azınlık sınıfındaki örneklerin sayısını dengelemeyi amaçlayan bir yöntemdir. Bu yöntemde, yalnızca azınlık sınıfındaki örnekler kopyalanır ve eğitim verisinde kullanılmaktadır. ROS yöntemi, azınlık sınıfındaki örnekler, orijinal veri kümesindeki örneklerin çoğaltılarak kullanılmasının yol açtığı en önemli problemlerden biri, aşırı öğrenme (overfitting)'dir. Bu problem, modelin eğitilmesinde verilerin ezberlemesine ve test aşamasında farklı verilerle karşılaşması durumunda sınıflandırıcı doğru bir şekilde sınıflandırma yapamamasına ve sınıflandırıcılar için eğitim süresinin artmasına neden olmaktadır (Alam ve diğerleri, 2020: 201179; Kotsiantis ve diğerleri, 2006: 4).

Şekil 21: Aşağı Örneklem ve Yukarı Örneklem Yapısı Örneği

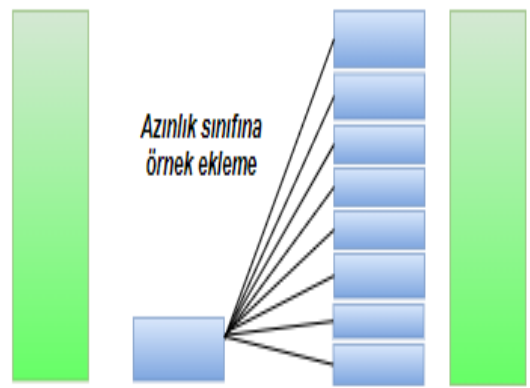
a) Aşağı Örneklem



Orijinal veri seti

Dengelenen veri seti

b) Yukarı Örneklem



Orijinal veri seti

Dengelenen veri seti

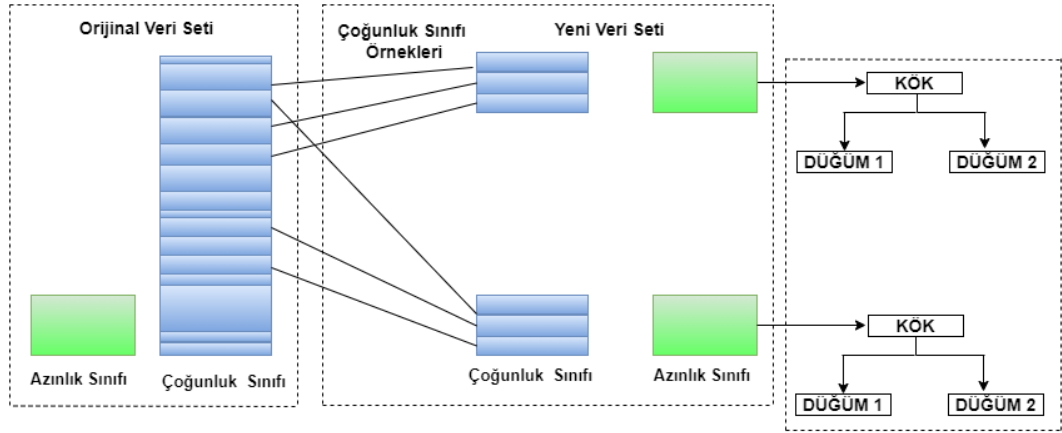
Kaynak: Mohammed ve diğerleri, 2020: 244.

2.6. ÖRNEKLEYİCİLER TOPLULUĞU

2.6.1. Dengeli Özyinelemeli Birleştirme Sınıflandırıcısı (BBC)

Dengeli olmayan veri kümelerinde, sınıf dengesizliği sorununun çözümü için kullanılan diğer bir yöntem, Dengeli Özyinelemeli Birleştirme Sınıflandırıcısı (Balanced Bagging Classifier) yöntemidir. Bu yöntemde, azınlık sınıfındaki örnekler ile çoğunluk sınıfındaki örneklerin sayısının dengelenmesi amacıyla alt örnekleme kullanılarak, orijinal veri seti örneklemesinden yeni alt kümeler oluşturulmaktadır (Barros ve diğerleri, 2019: 3). Daha sonra, oluşturulan bu alt kümeler, çeşitli sayıda özyinelemeli eğitim örneklerine birbirinden farklı ağaç tabanlı sınıflandırıcıların yerleştirilmesiyle, ikili sınıflandırma modellerinden oluşan toplulukları içermektedir. Son olarak, elde edilen tahmin değerleri, çoğunluk oyu yardımıyla her bir topluluktan bağımsız olarak elde edilmektedir. Bu durum, varyansın azalmasına ve güçlü bir tahmin yapılmasına olanak sağlamaktadır (Schlögl, 2020: 1). Dengeli özyinelemeli birleştirme sınıflandırma örneği Şekil 22’de gösterildiği gibidir.

Şekil 22: Dengeli Özyinelemeli Birleştirme Sınıflandırıcısı Örneği



Kaynak: Barros ve diğerleri, 2019: 4.

2.7. AŞAĞI VE YUKARI ÖRNEKLEME KOMBİNASYONLARI

2.7.1. SMOTE-Tomek Links

Dengeli olmayan veri kümelerinde, sınıf dengesizliği sorununun çözümü için kullanılan diğer bir yöntem, Sentetik Azınlık Yukarı Örneklem Tekniği (SMOTE) ile Tomek Bağlantıları (Tomek-Links) yöntemlerinin birleşimi olarak bilinen SMOTE-Tomek Links yöntemi, 2003 yılında Batista ve diğerleri tarafından tanıtılmıştır. Bu yöntem, sınıfların eşitlenmesi amacıyla çoğunluk sınıfında bulunan örneklerin sayısının azaltılması ve azınlık sınıfında bulunan örneklerin sayılarının artırılması için kullanılan bir hibrit yöntemdir. Azınlık sınıfı örnekleri, yeni yapay sentetik azınlık sınıfı örneklerinin eklenmesi için SMOTE yöntemi ve çoğunluk sınıfı örneklerinin azaltılması için de Tomek Bağlantıları yöntemi kullanılmaktadır. Tomek Bağlantıları yöntemi, çoğunluk sınıfındaki örneklerin azaltılmasında, azınlık ve çoğunluk sınıfında bulunan en yakın iki örnek arasında Tomek Bağlantıları oluşturulmaktadır. Tomek Bağlantılarının oluşması, sınıflardaki örneklerden yalnızca birinin gürültülü olduğu veya sınıflardaki iki örneğin de sınıfların sınırına yakın olduğunu ifade etmektedir. Ayrıca, bu bağlantı, doğru sınıflandırma özelliği bulunmayan çoğunluk sınıfındaki örneklerin atılması ile farklı iki sınıftaki örneklerin sayılarının dengelenmesini amaçlamaktadır (Batista ve diğerleri, 2003: 5; Elhassan ve diğerleri, 2016: 2; Zheng ve diğerleri, 2021: 113695-113696).

SMOTE-Tomek Links yönteminin aşamaları aşağıdaki gösterildiği gibidir (Viadinugroho, 2021).

I. Aşama: SMOTE yönteminin başlangıç aşamasıdır. Veriler azınlık sınıfından rassal olarak seçilir.

II. Aşama: Rassal veriler ile k en yakın komşu arasında bulunan mesafe hesaplanır.

III. Aşama: 2. Aşamada çıkan fark, 0 ile 1 arasında rassal bir sayı ile çarpılır ve çıkan sonuç azınlık sınıfına sentetik örnek olarak eklenmektedir.

IV. Aşama: Azınlık sınıfındaki örnekler olması gereken düzeye gelene kadar II. ve III. aşamalar tekrarlanır ve gereken düzeye ulaşıldıktan sonra SMOTE sonlandırılır.

V. Aşama: Tomek Bağlantılarının başlangıç aşamasıdır. Veriler çoğunluk sınıfından rassal olarak seçilir.

VI. Aşama: Rassal olarak seçilen verilerin, en yakın komşusu azınlık sınıfındaki verilerden gelirse (Tomek Bağlantısı oluşturulması), o zaman Tomek Bağlantılarının kaldırılması gerekmektedir.

2.7.2. SMOTE-ENN

Dengeli olmayan veri kümelerinde, sınıf dengesizliği sorununun çözümü için kullanılan diğer bir yöntem, Sentetik Azınlık Yukarı Örnekleme Tekniği (SMOTE) ile Düzenlenmiş En Yakın Komşu (Edited Nearest Neighbor veya ENN) yöntemlerinin kombinasyonu olarak bilinen SMOTE-ENN yöntemi, 2004 yılında Batista ve diğerleri tarafından önerilmiştir. Bu yöntemde, Azınlık sınıfı örnekleri, yeni yapay sentetik azınlık sınıfı örneklerinin eklenmesi için SMOTE yöntemi ve gözlem sınıfı ile K -en yakın komşu çoğunluğu arasındaki farklı sınıfa ait olduğu tespit edilen gözlemlerin iki sınıftan çıkarılması işlemi ENN yöntemi ile gerçekleştirilmektedir. ENN yöntemi, veri kümesindeki gürültülü verilerin çıkarılması için etkin bir şekilde kullanılan bir tekniktir (Muaz ve diğerleri, 2020: 481; Viadinugroho, 2021).

SMOTE-ENN yönteminin aşamaları aşağıda verildiği gibidir (Viadinugroho, 2021).

I. Aşama: SMOTE yönteminin başlangıç aşamasıdır. Veriler azınlık sınıfından rassal olarak seçilir.

II. Aşama: Rassal veriler ile k en yakın komşu arasında bulunan mesafe hesaplanır.

III. Aşama: 2. Aşamada çıkan fark, 0 ile 1 arasında rassal bir sayı ile çarpılır ve çıkan sonuç azınlık sınıfına sentetik örnek olarak eklenir.

IV. Aşama: Azınlık sınıfındaki örnekler olması gereken düzeye gelene kadar II. ve III. aşamalar tekrarlanır ve gereken düzeye ulaşıldıktan sonra SMOTE yöntemi sonlandırılır.

V. Aşama: ENN yönteminin başlangıç aşamasıdır. k en yakın komşu sayısı belirlenir. Eğer herhangi bir k sayısı belirlenmediyse, $k = 3$ olarak belirlenir.

VI. Aşama: Veri setinde yer alan diğer gözlemler arasında bulunan gözlemin k en yakın komşusu belirlenir ve sonrasında çoğunluk sınıfın k en yakın komşudan alınır.

VII. Aşama: Gözlemin bulunduğu sınıf ve buna en yakın komşusundan, çoğunluk sınıfının farklı olması durumunda, gözlem ve bu gözleme en yakın komşu veri setinden atılmaktadır.

VIII. Aşama: Azınlık ve çoğunluk sınıfındaki örnekler olması gereken düzeye ulaşana kadar II. ve III. aşamalar tekrarlanır ve gereken düzeye ulaşıldıktan sonra ENN sonlandırılır.



ÜÇÜNCÜ BÖLÜM

UYGULAMA

3.1. ÇALIŞMANIN AMACI

Bankalar, fon arz eden ve fon talebinde bulunan müşteriler arasında bir aracılık görevi gören finansal kuruluşlardır. Bu sayede, gelir elde etmektedirler. Bankalar, müşterilerin ihtiyaçları doğrultusunda fon olarak kredi vermektedirler. Müşterilere verilen bu krediler, bankalar açısından yüksek bir kredi riskini de beraberinde getirmektedir. Bankaların iyi bir değerlendirme sonrası en uygun müşterilerine kredi vermeleri gerekmektedir. Aksi takdirde, müşteriler tarafından alınan kredilerin geri ödenmemesi durumunda bankalar kredi riski başta olmak üzere çeşitli risklerle karşılaşacaktır. Bu durumla baş etmek için daha objektif ve daha doğru yöntemlerin kullanılması bankalar açısından olumlu etki oluşturacaktır. Gelişen teknoloji ile birlikte bankalar daha sistematik hâle gelmiştir. Müşterilerin kredi talebinde bulunmak için doğrudan bankaya gitmeden her an kredi taleplerini gerçekleştirmeleri mümkündür. Bu avantaj sayesinde krediye olan talep artırmıştır. Bu durum beraberinde kredi tahsis eden personellerin iş yükünü de artırmaktadır. Bankalar kredilendirme faaliyetlerini pratik bir şekilde gerçekleştirmek için farklı yöntemlere ihtiyaç duymaktadırlar. Gelişen teknoloji ile birlikte farklı alanlarda kullanılan makine öğrenmesi yöntemleri, bankaların kredilendirme sürecine uyarlanan yöntemler ile riskli müşterilerin tespit edilmesini kolaylaştırmıştır.

Bu tez kapsamında yürütülen çalışmanın amacı, müşterilerden alınan bilgiler doğrultusunda, kredilendirme sürecinin gerçekleştirilmesinde dengeli olmayan veri kümelerinde karşılaşılan sınıf dengesizliği probleminin çözümü sayesinde hem riskli müşterilerin tespit edilmesi ile kredi riskinin azaltılması ve dolayısıyla bankaların kârının artmasına olumlu katkı sunması amaçlanmıştır. Bu doğrultuda, makine öğrenmesine ait en iyi performans gösteren sınıflandırma modeli de belirlenmiştir. Bu sınıflandırma modeli ile bankaların kredilendirme sürecinde risk teşkil eden müşterilerin saptanması hız kazanacaktır. Ayrıca bu çalışmada kullanılan gerçek hayat problemine ait dengeli olmayan veri kümelerinin makine öğrenmesi yöntemlerinin performansını olumsuz etkilememesi için yeniden örnekleme teknikleri ve bu tekniklerin hibrit yöntemleri ele alınmıştır. Ek olarak, dengelenen bu veri kümeleri

üzerinde tekil modeller ve topluluk öğrenmesi modelleri performansları birbiri ile kıyaslanmış ve en iyi performansı gösteren örnekleme yöntemi ve model birleşimi belirlenmiştir. Aynı zamanda gelecekte karşılaşılabilecek benzer nitelikteki problemlerin çözümüne katkı sağlaması amaçlanmıştır.

3.2. LİTERATÜR TARAMASI

Malekipirbazari ve Aksakalli (2015) çalışmasında, dünyanın bilinen en büyük eşler arası kredi ağı olan *LendingClub.com* sitesinden 2012-2014 yılları arasında halka açık işlem kayıtları kullanılarak kredibilite durumundaki iyi borçluların tespit edilmesi amacıyla Rassal Orman (RF) tekniği kullanılmıştır. Fair Isaac Corporation (FICO) olarak bilinen ve Amerika Birleşik Devletleri tarafından kredi değerlendirilmesinde de kullanılan FICO puanlama kriteri, yaklaşık 350.000 kişinin kredi ve finansal özellikleri gibi bilgileri içeren veri setleri kullanılmıştır. Çalışmada kredi risk puanlanmasının güvenilirliğinin iyileştirilmesi için standart olmayan finansal özelliklere yer verilmekte ve kredibilite açısından iyi durumdaki borçluların belirlenebilmesi için rassal orman (RF), destek vektör makine (SVM), *k*-en yakın komşu (*k*-NN) ve lojistik regresyon (LR) gibi makine öğrenmesi teknikleri kullanılmıştır. Sonuç olarak, FICO puan kriteri, P2P kredi notları ve diğer tüm sınıflandırıcılar göz önünde bulundurulduğunda rassal ormanların en iyi sonucu verdiği görülmüştür.

Dahiya ve diğerleri (2016) çalışmasında, kredi risk değerlendirilmesinin doğru bir şekilde yapılmasının finansal kurum ve bankalar için büyük önem arz ettiğini ifade etmiştir. Bankalardaki olası kayıpları önlerken, kârlılığa olumlu katkı sağlamaktadır. Çalışmada üç model geliştirilmiştir. Bu modellerden ilki, tek MLP modelidir. İkinci olarak, bu tek MLP modelin öngörü doğruluğunun değerlendirilebilmesi için öznelik seçme tekniği kullanılmıştır. Son olarak, doğruluk değerlerini daha da iyileştirmek için Bagging-hibrit MLP yöntemi uygulanmıştır. Geliştirilen modellerin doğrulukları Avustralya ve Alman veri setleri üzerinde karşılaştırılmıştır. Bu veri setleri, “*UCI Machine Learning Repository*” üzerinden alınmış gerçek hayat kredi veri setleridir. Eğitim ve test veri setleri %70:30’luk bölünerek ikili karşılaştırma yapılabilmesi için üç deney gerçekleştirilmiştir. MLP kullanılarak bagging ile öznelik seçme tekniğinin birleşimiyle oluşturulan hibrit yöntemler, Avustralya ve Alman veri kümeleri sırasıyla;

%90,50 ve %80 doğruluk değerlerine sahiptir. Ek olarak, eğitim ve test verilerinin en iyi sonuçları %70:30'a bölünmesiyle daha iyi katkı sağladığı görülmüştür.

Khemakhem ve diğerleri (2018) çalışmasında, Tunus'taki çeşitli sektörde faaliyet gösteren firmalar yöntem kredisi almak için Tunus bankasına kredi başvurusu yapmıştır. Verilen krediler çeşitli istatistiksel yöntemler ve yapay zekâ kombinasyonları kullanılarak yeniden örnekleme yöntemleri ile analiz edilmiştir. Çalışmada 2011-2012 yılları arasında talep edilen çeşitli kredi verileri kullanılmıştır. Tunus'taki belirlenen 408 şirket arasından 300'ü istenilen tüm mükellefiyetleri yerine getirmiştir. 108'i kredi güvenilirliği olmayan borçlulardan oluşmaktadır. 3 nicel ve 22 nitel veri olmak üzere 25 değişken ile kredi risk analizi yapılmıştır. Çalışmanın amacına uygun olarak dengesiz verileri dengeli hâle getirebilmek için Rassal yukarı örnekleme (ROS) ve Sentetik azınlık aşırı öğrenme tekniği (SMOTE) örnekleme teknikleri kullanılarak sınıflandırma performansları üzerindeki etkileri araştırılmıştır. Lojistik regresyon, yapay sinir ağları ve destek vektör makineleri gibi sınıflandırıcılar analizde kullanılmıştır. Doğruluk performans ölçüsü açısından incelendiğinde, destek vektör makineleri ve rastgele yukarı örnekleme kombinasyonu en yüksek doğruluk değerine sahiptir. Dengeli hâle getirilen verilerin performansları değerlendirildiğinde yapay zekâ teknikleri, lojistik regresyona göre daha iyi sonuç göstermiştir. Sonuç olarak, rassal yukarı örnekleme teknikleri ile yapay zekâ kombinasyonlarının birleşiminin performans üzerindeki etkilerine bakılarak alınan kredilerin geri ödenmesinin tahmin edilebilir olması açısından önemli bir katkı sağlanmıştır.

Shen ve diğerleri (2019) çalışmasında, Kredi risk analizi ile ilgili yapılan önceki çalışmaları incelenmiş ve araştırma sonucunda geri yayılım sinir ağları teknikleriyle kredi risk değerlendirmesinde anlamlı sonuçlara ulaşıldığını gözlemlemiştir. Çalışmada SMOTE tabanlı topluluk modeli ile bireysel kredi risk değerlendirilmesi için sınıflandırıcıya dayalı bir optimizasyon yöntemi önerilmiştir. Dengesiz veri setlerinin kredi risk değerlendirme sürecinde model performansı üzerindeki istenmeyen etkilerini azaltmak için SMOTE tekniği kullanılmıştır. Ek olarak, Parçacık sürüsü optimizasyonu (PSO) algoritmaları, BP sinir ağlarıyla en iyi ağırlıkları belirlemek ve sapmaları aramak için kullanılmıştır. BP sinir ağları sınıflandırıcısının optimize edilmiş değerlerini ve AdaBoost topluluk öğrenme tekniklerini klasik tekniklerle birleştirerek yeni bir topluluk öğrenme modeli

geliştirmişlerdir. Geliştirilen modelin topluluk öğrenme üzerindeki performanslarını değerlendirebilmek için herkese açık gerçek veriler olan *UCI Machine Learning Repository* üzerinden alınmış Alman ve Avustralya kredi veri setlerinden faydalanılmıştır. Elde edilen sonuçlara göre çalışmada geliştirilen modelin klasik sınıflandırma modelleri ile karşılaştırıldığında daha anlamlı sonuçlar verdiği görülmektedir.

Hou ve diğerleri (2020) çalışmasında, Kredi risk değerlendirilmesi sürecinde kullanılan dengesiz veri setleri için META-DESKNN-MI adı verilen yeni birleşik Dinamik topluluk seçimi (DES) modeli önermiştir. Verileri dengeli hâle getirebilmek için SMOTE yöntemi kullanılmıştır. Sınıflandırıcıları değerlendirirken kullanılan azınlık örneklerin önemini göstermek için çoklu sınıf dengesizliği (DES-MI) kullanılmıştır. Topluluk öğrenme sınıflandırıcılarının performanslarını daha etkili hâle getirmek için META-DES ve DES-KNN kullanılarak çerçeveleri birleştirilmiştir. Tüm bunlar kredi risk değerlendirilmesi amacıyla <https://www.kaggle.com/wendykan/lending-club-loan-data> sitesinden alınan P2P kredi veri kümesine META-DESKNN-MI uygulanmıştır. Önerilen model, DES'in performansını iyileştirdiğini tespit etmişlerdir. Deneysel çalışmalar, KEEL üzerinden alınan 15 tane dengesiz veri seti üzerinde test edilmiştir. Sonuç olarak, çalışmada önerilen yeni DES modeli, bilinen yedi DES modelinden daha iyi sonuçlar vermiştir. P2P kredi veri kümelerinden elde edilen sonuçlar, yeni DES modelinin kredi puanlamasında kullanılan topluluk öğrenme modellerinden daha düşük bir seviyede tip 1 hata değerine ulaştığını göstermiştir.

Niu ve diğerleri (2020) çalışmasında, P2P kredi verilerinde kullanılan dengeli olmayan veri dağılımı (REMDD) yöntemi ile yeniden örnekleme dayanan topluluk öğrenme yöntemini önermiştir. REMMD yöntemi, çoğunluk sınıfı veri dağılımına (UMCDD) dayanan alt örnekleme yöntemini kullanarak sınıfların dengelenmesi için kullanılmıştır. Eğitim verilerine ait alt kümeleri dengeli duruma getirebilmek için UMCDD tekniği kullanılmıştır. REMMD yönteminin sınıflandırma performansının daha da etkili bir hâle getirebilmek için, doğrulama setinde yüksek performansa sahip sınıflandırıcılar, sınıf tahmininin belirlenmesi amacıyla kullanılmıştır. Daha sonra, bu önerilen modelin performansını değerlendirmek ve daha avantajlı olduğunu göstermek için üç veri seti kullanılmıştır. Bu veri setleri; PaiPaiDai (PPD), Prosper ve Lending Club üzerinden alınan üç P2P'ye ait kredi verileridir. Sonuç olarak önerilen model

olan REMMD, P2P'nin dengeli olmayan kredi verilerinin değerlendirilmesinde daha iyi sonuç vermiştir.

Jin ve diğerleri (2021) çalışmasında, kredi tahminini doğru ve istikrarlı bir şekilde öngörülebilir olması için hibrit genetik algoritmaya dayalı birçok aşamadan oluşan yeni bir topluluk modelini önermiştir. Kredi puanlama model verilerinin dengesiz olması sebebiyle model performansı üzerinde olumsuz etkisi olduğundan bahsetmişlerdir. Bu etkiyi azaltmak için klasik örnek sertlik eşiği (Instance Hardness Threshold) kullanılmış ve veri dengesizliği problemi için çoğunluk oylama stratejileri kullanılmıştır. Daha sonra veri kümesindeki özellik seçilmesi aşamasında önerilen yeni hibrit genetik algoritma, temel sınıflandırıcıların her birinin optimal alt küme değerlerini seçmek için kullanılmıştır. Temel sınıflandırıcıların tahmin güçlerinin sonuçlarını bir araya getirmek için, topluluk öğrenme modeli ile temel sınıflandırıcıların optimal değerlerini birleştiren yığınlama (stacking) yaklaşımını benimsemişlerdir. Geliştirilen modelin performansını incelemek için BACC, F-score, G-mean ve Recall ölçüleri kullanılmıştır. Çalışmada önerilen modellerin etkinliğini test etmek için *UC Irvine Machine Learning Repository* üzerinden alınan Alman, Polonya-1 ve Polonya-2 gerçek veri setlerinden faydalanılmıştır. Sonuç olarak, önerilen modelin kredi puanlama modelleri üzerindeki etkisi diğer karşılaştırma yöntemlerinden daha iyi performans gösterdiği tespit edilmiştir.

Moscato ve diğerleri (2021) çalışmasında, P2P kredi platformundaki kullanıcıların temerrüde düşüp düşmeyeceği hususundaki kredi risk puanlamasında en sık kullanılan tekniklere dayalı bir karşılaştırma çalışması gerçekleştirmiştir. Ayrıca çalışma kapsamında literatürde en yaygın olarak kullanılan örnekleme tekniklerinden yararlanılmıştır. Çalışmanın amacına uygun olarak dengesiz veri setleri P2P kredi platformundan farklı değerlendirme kriterleri ile incelenmiş ve elde edilen sonuçlar makine öğrenmesi teknikleriyle karşılaştırmıştır.

Xiong ve Huang (2021) çalışmasında, Kredi risk performanslarının değerlendirilmesi sürecinde temel sınıflandırıcıyı kapsayan topluluk öğrenmesi modelleri kullanılmıştır. Çalışmanın amacı, Maksimum bilgi katsayısından (MIC-CCS) faydalanarak topluluk öğrenmesi modellerinin sınıflandırma gücünü artırmak için korelasyona dayalı bir sınıflandırıcı geliştirmektir. Çalışmada iki önemli husus üzerinde durulmuştur. İlki, temel sınıflandırıcıların doğru bir şekilde seçilerek topluluk

öğrenmesi modellerinin sınıflandırma performanslarının etkinliğini artırmak için farklı çözüm önerileri sunulmuştur. İkincisi, önceki çalışmalardan farklı olarak kompleks veri kümeleri içerisindeki doğrusal olmayan ve fonksiyonelliğini kaybeden ilişkileri tespit edebilmek için MIC-CCS tekniği önerilmiştir. Bu doğrultuda önerilen MIC-CCS tekniği, *UCI Machine Learning Repository* üzerinden alınan Avustralya, Tayvan ve Alman veri setleri, yaygın olarak kullanılan 8 sınıflandırma yöntemi ve 4 topluluk öğrenmesi modellerini kıyaslayan bir kredi risk analizi yapılmıştır. Bu yöntemler; lojistik regresyon (LR), doğrusal diskriminant analizi (LDA), naive Bayes (NB), k-en yakın komşular (k-NN), destek vektör makinesi (SVM), sinir ağları (NN), C5.0 karar ağaçları (DT), rassal orman (RF) ve hafif gradyan artırma makineleri (LGB) gibi temel sınıflandırma modelleridir. Ayrıca Pearson-CCS, geriye doğru seçim (BW-CCS), ileriye doğru seçim (FW-CCS) ve sekiz temel sınıflandırıcının tümü (ALL) ve topluluk öğrenme modelleri de kullanılarak MIC-CCS yöntemi ile karşılaştırılmıştır. Sonuç olarak, sınıflandırmada kullanılan model performansları AUC performans ölçütüne göre değerlendirilmiştir. Önerilen MIC-CCS tekniği, AUC'ye göre topluluk öğrenmesi modellerinde en etkin sonucu vermiştir. Sınıflandırma yöntemleri arasında en iyi performans LDA göstermiştir. Meta sınıflandırıcısı olarak en iyi sonuçlar Avustralya verisi için destek vektör makinesi, Alman ve Tayvan verisi için sırasıyla rassal orman ve naive Bayes modellerinde ölçülmüştür.

Wang (2021) çalışmasında, dengesiz veri kümeleri kullanılarak kredi riskinin tahminlemesi için topluluk öğrenme modelleri oluşturmuş ve finansal veri kümelerinin öznitelikleri ile ilgili algoritmalar geliştirmiştir. Topluluk öğrenme modelleri, Parçacık Sürü Optimizasyonu (PSO) ve SMOTE ile optimal hâle getirilerek Çok Çekirdekli Bulanık C-Ortalamları (MK-FCM) birleştirme tekniğinden oluşmuştur. Veri setlerini dengeli hâle getirebilmek için SMOTE tekniği kullanılmıştır. Daha sonra, Bulanık C-Ortalamlar (FCM), sınıflandırıcıların performansını artırmak için çeşitli çekirdek işlevselliğinin özelliklerini sentezleme gücüne sahip bir aday sınıflandırıcı olarak MK-FCM, çok çekirdekli işlev tarafından oluşturulmuştur. PSO ise, MK-FCM'de kullanılan parametrelerinin en optimal değerlerini elde etmek için kullanılmıştır. Çalışmada kullanılan Çin misafirperverliği ve turizm firmalarının yer aldığı listeden seçilen, kredi riski olan konaklama ve turizm firmaları ile kredi riski olmayan konaklama ve turizm firmaları normal örnekler olmak üzere mevcut listelerden

seçilmiştir. 2007-2016 yıllarında Çin Şanghay Borsası ile Shenzhen Borsası'ndan konaklama ve turizm verilerine ait 251 örnek arasından 24 tanesi riskli şirket, geriye kalan 227 şirket kredi riski olmayan normal örneklerden oluşmuştur. Önerilen model, kredi riski performansı açısından 4 temel sınıflandırıcı ve 3 ön işleme yöntemiyle karşılaştırılabilir duruma getirilmiştir. Sonuç olarak, 4 önemli avantajla karşılaştırılmıştır. İlki, finansal verilerdeki örnek karmaşıklığını azaltmaya yardımcı olan azalan boyutun çokluğu yöntemidir. İkincisi, verileri dengeli hâle getirmek için kullanılan SMOTE'dir. Üçüncü olarak, FCM'nin geliştirilmesi için kullanılan çok çekirdekli işlevler, finansal veri kümelerindeki kompleks yapıyı ve dengesiz dağılım üzerinde verilerin işlenmesine olanak sağlamıştır. Dördüncüsü ise, MK-FCM'nin optimal parametre değerleri, geliştirilmiş PSO tarafından daha iyi ayarlanmıştır.

Dumitrescu ve diğerleri (2022) çalışmasında, kredi puanlaması durumunda, rassal ormanların klasik lojistik regresyondan daha etkili sınıflandırma performansı göstermiştir. Çalışmada lojistik regresyonun etkililiğini artırmak amacıyla karar ağaçlarından elde edilen bilgilere dayalı, cezalandırılmış lojistik ağaç regresyonu (PLTR) tekniği önerilmiştir. Önerilen model olan PLTR, lojistik regresyonu içsellik açısından değerlendirilebilirken, kredi puanlamasındaki veri kümelerinde oluşabilecek doğrusal olmayan durumların tespit edilmesinde de yardımcı olmaktadır. Çalışma kapsamında, “*Give me some credit*” isimli yaygın olarak kullanılan dört veri kümeleri kullanılmıştır. İstatistiki performansların değerlendirilmesi aşamasında literatürde en sık kullanılan beş doğruluk kriteri ve altı yöntem kullanılmıştır.

Bu kriterler sırasıyla; AUC, Brier Skoru (BS), Kısmi Gibi indeksi (PGI), Kolmogorov Smirnov (KS) ve doğru sınıflandırılmış olayların yüzdesi (PCC) 'dir. Kullanılan yöntemler ise; doğrusal lojistik yöntemi, destek vektör makine, rassal orman, PLTR, sinir ağı, doğrusal olmayan lojistik regresyon ve doğrusal olmayan lojistik regresyon ve Alasso yöntemlerinin birleşimi kullanılmıştır. Yöntemler ve değerlendirilme kriterleri göz önünde bulundurulduğunda PLTR yöntemi, PCC ve BS dışındaki diğer yöntemlerden daha iyi performansa ulaşmıştır. Bu sonuçlara bakılarak PLTR, tüm lojistik regresyon yöntemlerinden daha iyi sonuç göstermiştir. Kredi puanlaması açısından ele alındığında, lojistik regresyon ve karar ağaçlarını birleştiren bir tekniğin kullanılması, istatistiksel modelleme açısından da bir çözüm sunmuştur. Önerilen PLTR'nin performansını ve uygulanabilirliğinin değerlendirilmesi açısından

kredi temerrüt veri kümeleri, Monte Carlo Simülasyonları ile karşılaştırmaya dayanan deneysel bir uygulama yapılmıştır. Ek olarak, “*Housing*”, “*Australian dataset*” ve “*Taiwan dataset*” isimli üç gerçek hayat veri setleri de kullanılmıştır. Sonuç olarak, PLTR, rassal ormanlarla kıyaslandığında daha rekabetçi, örneklem dışı performansı açısından doğrusal olmayan ve doğrusal lojistik regresyondan daha etkili performans gösterdiği tespit edilmiştir.

Liu ve diğerleri (2022) çalışmasında, önceki çalışmalar incelenmiş ve belirli perspektifler doğrultusunda kredi puanlaması performansının daha da güçlendirmek için iki gradyan artırılmış karar ağaçları (AugBoost-RFS ve AugBoost-RFU) yöntemi önerilmiştir. Önerilen AugBoost-RFS ve AugBoost-RFU yöntemlerinde, temel sınıflandırıcıların, öznelilikleri artırma yoluyla çeşitlendirilmesi sağlanmıştır. Ağaç tabanlı gömme teknolojileri, hem gradyan artırılmış karar ağaçlarının yorumlanabilirliğini hem de öznelilik artırma sürecini kolaylaştırmıştır. Çalışma kapsamında, “*Fannie Mae (Fannie)*”, “*Bankfear*”, “*Lending Club*” ve “*Give me some credit (Give)*” isimli dört yüksek boyutlu kredi puanlama veri setiyle analiz yapılmıştır. Ayrıca, önerilen yöntemlere bir alternatif olarak AugBoost-NN yöntemi sunulmuştur. Daha sonra, denetimli gradyan artırılmış karar ağaçları (AugBoost-RFS ve AugBoost-NN) ile denetimsiz gradyan artırılmış karar ağaçları (AugBoost-RP, AugBoost-PCA ve AugBoost-RFS) analiz sonuçları incelendiğinde, kredi puanlaması değerlendirilmesi aşamasında gözetimli GBDT daha iyi performansa yol açtığı tespit edilmiştir.

Yuan ve diğerleri (2022) çalışmasında, bankaların kredi verme kararlarının belirlenmesi aşamasında, temerrüde düşecek bir firmada yanlış bir tahmin yapılması ile kredi verilmesi kararı ve temerrüde düşmeyen bir firmanın da yanlış tahmin yapılmasıyla kredi verilmemesi kararının potansiyel müşteri ve ciddi bir mali kayba yol açacağını belirtmiştir. Çalışmada, kredi veren bankaların ve banka dışı finans kurum ve kuruluşlarının kredi puanlaması için destek vektör makine ve k-ortalama kümelemenin birleştirilmesiyle oluşan melez iki aşamalı (SVDD) bir yöntem önerilmiştir. 2000-2019 yılları arasında Çin borsasındaki 3425 şirket verileri kullanılarak analiz yapılmıştır. Sonuç olarak, önerilen SVDD hibrit yöntem, model doğruluğunun tahmininde tek aşamalı modellerle kıyaslandığında daha iyi performans

göstermiş ve iki aşamalı hibrit yöntemin 5 senelik temerrüt tahmininde başarılı olduğu tespit edilmiştir.

3.3. DENEY TASARIMI AŞAMASI

Makine öğrenmesine teknikleriyle bir sınıflandırma problemine ilişkin en iyi çözümü gerçekleştiren yöntemin belirlenmesi oldukça önemlidir. Bu yöntemin hem doğru hem de güvenilir olması yapılan analiz için önem arz eder. Buna yönelik olarak, ilgilenilen problemin çözümü için sistemli bir şekilde oluşturulan ve çeşitlendirilmiş bir deney tasarımı oluşturmak fayda sağlayacaktır. Bu deney tasarımı amaç, problem için en optimal çözümün bulmada minimum seviyede bir deney tasarlanması ile tüm süreç içerisinde optimal çıktı değerini veren yöntemin tespit edilmesidir. Deney tasarımının en önemli avantajı nihai çıktının optimize edilmesi için girdilerin hangi düzeyde tutulması gerektiğini belirlemesidir. Ayrıca deney tasarımı yapılan analizin optimal çözüme hızlı bir şekilde ulaşması açısından kolaylık sağlayacaktır. Sonuç olarak deney tasarımı sadece mevcut problemin çözümü için değil, aynı zamanda gelecekte karşılaşılabilecek benzer problemlerin çözümü için de önem arz etmektedir.

Bu çalışmanın uygulamasında kullanılmak üzere dört farklı senaryo hazırlanmıştır. Bu senaryolar ve içerik bilgileri Tablo 2’de gösterilmiştir.

Tablo 2: Uygulama Aşamasında Yer Alan Senaryolar

Uygulama Senaryoları	Uygulamada Kullanılan Senaryolara İlişkin İçerik Bilgileri
Senaryo 1	Rassal Aşağı Örneklem ve Rassal Yukarı Örneklem Tekniği ile Tekil Modellerine ait Tahmin
Senaryo 2	Rassal Aşağı Örneklem ve Rassal Yukarı Örneklem Tekniği ile Topluluk Öğrenmesi Modellerine ait Tahmin
Senaryo 3	Aşağı ve Yukarı Örneklem Yöntemlerinin Kombinasyonu ile Örnekleyiciler Topluluğu ve Tekil Modellerine ait Tahmin
Senaryo 4	Aşağı ve Yukarı Örneklem Yöntemlerinin Kombinasyonu ile Örnekleyiciler Topluluğu ve Topluluk Öğrenmesi Modellerine ait Tahmin

Bu tez kapsamında kullanılmış olan dengeli olmayan üç farklı veri seti; Alman kredi veri seti, Avustralya kredi veri seti ve HMEQ kredi verisinin dengeli hâle getirilmesinde “Rassal yukarı örneklem, Rassal aşağı örneklem, Dengeli

Özyinelemeli Birleştirme Sınıflandırıcısı , SMOTE-ENN Combine, SMOTE-Tomek Links gibi hem aşağı örnekleme yöntemleri hem de yukarı örnekleme yöntemleri ve bunların birleştiren teknikler kullanılmıştır. Dengeli hâle getirilmiş olan veri kümelerinin tahmin performanslarının değerlendirmesi için 11 farklı makine öğrenmesi tekniğinden faydalanılmıştır. Bu teknikler kendi içerisinde farklı türlere ayrılmaktadır. K-en yakın komşu, Naive Bayes, Lojistik Regresyon, Destek Vektör Makineleri, Çok Katmanlı Algılayıcı ve Karar Ağaçları gibi teknikler tekil sınıflandırma modelleri olarak adlandırılırken, Rassal Ormanlar, Gradyan Artırma Karar Ağaçları, Extra Ağaçlar, Sert ve Yumuşak oylama gibi teknikler topluluk öğrenmesi modelleri olarak adlandırılmaktadır. Ek olarak, bu tekniklerin performanslarının değerlendirilmesi için 5 farklı ölçüm istatistiği kullanılmıştır. Bu performans ölçütleri; Doğruluk, Kesinlik, Hassasiyet, F1-Skor ve Matthews Korelasyon Katsayısı'dır.

Uygulama aşamasında kullanılmış olan makine öğrenmesi teknikleri ile ilgili süreç 6 adımdan oluşmaktadır. Bu adımlar;

Problemin Tanımlanması

Veri kümelerinin Hazırlanması

Veri Ön İşleme Süreci

Modellerin Belirlenmesi ve Oluşturulması

Modellere ait Performans Ölçümlerinin Elde Edilmesi

Sonuçların Değerlendirilmesi ve Raporlanması

3.3.1. Veri Setleri ve Açıklamaları

Çalışmada, açık kaynaklı site üzerinden alınan üç farklı kredi verisi kullanılmıştır. Bu veri setleri; “*UCI Machine Learning Repository*” isimli açık kaynaklı internet sitesi üzerinden Almanya ve Avustralya adlı iki veri seti alınmıştır. Açık kaynaklı erişime sahip olan diğer bir internet sitesi “*Kaggle*” üzerinden ise HMEQ (Mortgage) adlı veri seti alınmıştır. Tez çalışması kapsamında yararlanılan tüm veri setlerine ilişkin açıklamalar bu bölüm altında detaylandırılmıştır.

3.3.1.1. Alman Kredi Veri Seti

“UCI Machine Learning Repository” adlı internet sitesinde bulunan “Statlog German Credit Data Set” adı ile tanımlanmış olan Alman kredi veri kümesi, gerçek hayat problemine ait bir veri setidir. Veri kümesi içerisinde 1000 adet kredi örneği bulunmakta ve bu veri seti 1 bağımlı (yanıt) değişken ve 20 bağımsız değişken olmak üzere toplam 21 değişkenden oluşmaktadır. Bu değişkenlerin 7 tanesi sayısal, 13 tanesi ise kategorik veri türündedir. Bağımlı değişkene ilişkin olarak, “kredi alabilir veya kredi alamaz” şeklinde ikili kategorik veri şeklindedir. Kredi başvuru talebinde bulunan müşteriler arasından 300 kişi “kredi alamaz” şeklinde etiketlenmişken, geriye kalan 700 kişi “kredi alabilir” şeklinde etiketlenmiştir. Alman kredi veri kümesi dengesiz veri özelliğine sahip bir veri kümesidir. Alman veri setinde yer alan değişkenler ve bunlara ilişkin detaylı açıklamalar Tablo 3’te gösterildiği gibidir.

Tablo 3: Alman Kredi Veri Kümesindeki Değişkenlere İlişkin Bilgiler

Değişken Adları	Değişken Türü	Öznitelik Tanımları (Attribute Description)
Status	categorical	Borçlunun bankadaki mevcut hesap durumu
Duration	quantitative	Ay cinsinden kredi süresi
Credit History	categorical	Sözleşmenin önceki veya mevcut kredi geçmişi
Purpose	categorical	Kredi başvuru amacı
Amount	quantitative	Toplam kredi tutarı
Savings	categorical	Borçlunun toplam tasarrufu
Employment Duration	ordinal	Bir borçlunun mevcut kuruluştaki çalışma süresi
Installment rate	ordinal	Kredi için belirlenmiş olan taksit oranı
Personal status sex	categorical	Borçlunun cinsiyeti ve medeni durumu
Other Debtors	categorical	Borçlunun kredisine kefil olan borçlu
Present Residence	ordinal	Mevcut ikametgahta yaşama süresi
Property	ordinal	Borçlunun mülkünün artan bir şekilde sıralaması
Age	quantitative	Borçlunun yaşı
Other Installment Plans	categorical	Kredi veren banka dışındaki her türlü kredi/taksit yükü
Housing	categorical	Borçlunun mevcut ikametindeki ev durumu
Number Credits	ordinal	Alınan kredinin tüm geçmişi
Job	ordinal	Borçlunun iş statüsü
People Liable	quantitative	Finansal olarak bakmakla yükümlü olduğu toplam kişi
Telephone	binary	Borçlunun adına kayıtlı bir sabit hattın durumu
Foreign Worker	binary	Borçlu yabancı bir işçi mi?
Credit Risk	binary	Kredi riski yok / kredi riski var

Kaynak: Alam ve diğerleri, 2020: 201178.

3.3.1.2. Avustralya Kredi Veri Seti

“UCI Machine Learning Repository” adlı internet sitesinde bulunan “Statlog Australian Credit Approval Data Set” adı ile tanımlanmış olan Avustralya kredi veri kümesi, gerçek hayat problemine ait bir veri setidir. Veri kümesi içerisinde 690 adet kredi örneği bulunmakta ve bu veri seti 1 bağımlı değişken ve 14 bağımsız değişken olmak üzere toplam 15 değişkenden oluşmaktadır. Bu değişkenlerin 8 tanesi sayısal, 7 tanesi ise kategorik veri türündedir. Bağımlı değişken, “kredi alabilir veya kredi alamaz” şeklinde ikili kategorik veri türünden oluşmaktadır. Kredi başvuru talebinde bulunan müşteriler arasından 383 kişi “kredi alamaz” şeklinde etiketlenmişken, geriye kalan 307 kişi “kredi alabilir” şeklinde etiketlenmiştir. Avustralya kredi veri kümesi dengesiz veri özelliğine sahip bir veri kümesidir. Avustralya veri setinde yer alan 14 değişkenin içeriği hakkında herhangi bir bilgi bulunmazken, değişkenlere ait örnek tiplerine ilişkin bilgiler mevcuttur. Değişken isimleri bilinmediği için bu değişkenler; X1, X2,....., X14, X15 değişken isimleriyle temsil edilmiştir.

3.3.1.3. HMEQ Kredi Veri Seti

“Kaggle” adlı internet sitesinde bulunan “Ev sermayesi kredi veri kümesi” adı ile tanımlanmış olan HMEQ veri kümesi, gerçek hayat problemine ait bir veri setidir. Veri kümesi içerisinde 5960 adet kredi örneği bulunmakta ve bu veri seti 1 bağımlı değişken ve 11 bağımsız değişken olmak üzere toplam 12 değişkenden oluşmaktadır. Bu değişkenlerin 9 tanesi sayısal, 3 tanesi ise kategorik veri türündedir. Bağımlı değişken, “kredi alabilir (Temerrüde düşmeyen) veya kredi alamaz (Temerrüde düşen)” şeklinde ikili kategorik veri türündedir. Kredi başvuru talebinde bulunan müşteriler arasından 1189 kişi “kredi alamaz” şeklinde etiketlenmişken, geriye kalan 4771 kişi “kredi alabilir” şeklinde etiketlenmiştir. HMEQ kredi veri kümesi dengesiz veri özelliğine sahip bir veri kümesidir. HMEQ veri setinde yer alan değişkenler ve bunlara ilişkin detaylı açıklamalar Tablo 4’te verilmiştir.

Tablo 4: HMEQ Kredi Veri Kümesindeki Değişkenlere İlişkin Bilgiler

Değişken Adları	Öznitelik Tanımları (Attribute Description)
<i>Loan</i>	Talep edilen kredi miktarı
<i>Mortdue</i>	Ödenmesi gereken ev kredisi (Mortgage) tutarı
<i>Value</i>	Mevcut mülkün değeri
<i>Yoj</i>	Mevcut işinde toplam çalışılan yıl
<i>Derog</i>	Temerrüt (Kredi borcunu ödeyememesi) sayısı
<i>Delinq</i>	Ödenmemiş olan kredi taksitlerinin sayısı
<i>Clage</i>	Aylık olarak ödenmesi gereken en son alınan kredi taksiti
<i>Ninq</i>	Son kredi sorgulama sayısı
<i>Clno</i>	Kredi taksit sayısı
<i>Debtinc</i>	Borç-gelir oranı
<i>Reason</i>	Krediyi alma gerekçesi
<i>Job</i>	Meslek
<i>Bad</i>	Kredi riski yok/ kredi riski var

Kaynak: Snart, 2021.

3.3.2. Veri Ön İşleme Süreci

Veri ön işleme süreci, analizin en önemli bölümünü oluşturan süreçlerden biridir. Bu süreçte, özellikle karmaşık ve yapılandırılmamış gerçek hayat problemlerinin kullanıldığı veri kümeleri ile doğrudan analiz yapılması, hem analizin güvenilirliği açısından hem de tahmin performansı açısından sakınca oluşturmaktadır. Bu nedenle, makine öğrenmesi yöntemleri ile analiz yapmadan önce çalışmada kullanılacak olan her veri kümesinin analiz için uygun hâle getirilmesi gerekmektedir. Böylece, uygun formdaki veriler ile yapılan analiz sonucu karmaşık yapı ortadan kalkmaktadır. Bu çalışma kapsamında yer alan üç farklı veri setine ait veri ön işleme süreci ve buna ilişkin bilgiler alt bölümlerde kapsamlı bir şekilde ifade edilmiştir.

3.3.2.1. Eksik Verilerin Tamamlanması

Tez kapsamında kullanılan üç farklı veri seti arasından yalnızca HMEQ kredi veri kümesi içerisinde eksik değerler bulunmaktadır. Bu veri kümesinde 5271 adet kayıp değer yer almaktadır. Bu değerler örnek tiplerinin türüne göre tamamlanmıştır. Eksik değerlere sahip değişken veri tipine göre; aritmetik ortalama veya mod değerleri kullanılarak doldurulmuştur. Kayıp gözlemler, kategorik özellikte bulunan nitel

verilerden oluşuyorsa, bu değerler mod değerleri yardımıyla tamamlanmıştır. Eğer eksik değişken değerleri, sayısal özellikte bulunan nicel verilerden oluşuyorsa, bu değerler aritmetik ortalama değerleri yardımıyla tamamlanmıştır. HMEQ veri seti içerisinde bulunan; “Loan”, “Mortdue”, “Value”, “Yoj”, “Derog”, “Delinq”, “Clage”, “Ninq”, “Cln”, “Debtinc” sayısal özellikteki nicel verilere ait eksik değişken değerlerinin her biri ilgili değişkene ait ortalama değerleri kullanılarak tamamlanmıştır. Ayrıca, HMEQ veri seti içerisinde bulunan; “Reason”, “Job” değişkenleri ise kategorik özelliğe sahip olan nitel verilerden oluştuğu için bu değişkenlere ait her bir eksik değişken değeri ilgili değişken içerisinde en çok yinelenen değer olarak bilenen mod değerlerinin kullanımı ile tamamlanmıştır.

3.3.2.2. Değişkenlerin Özniteliklerine Göre Kodlanması

Python programlama dili temeline dayalı olan ve machine öğrenmesi modellerin oluşturulması aşamasında yararlanılan “Scikit-Learn” kütüphanesi içerisinde, nominal ve ordinal veri özelliğindeki değişkenlere ait kodlama işlemlerinin gerçekleştirilmesi için farklı yöntemler kullanılmaktadır. Nominal özelliğe sahip veriler, sıralanması gerekmeyen ya da sıralanabilir özelliği bulunmayan, nitel özellikteki verilerden oluşur. Bu veri tipleri için “scikit-learn” kütüphanesindeki “*OneHotEncoder()*” fonksiyonu ile kategorik özellikteki nitel verilerin, sayısal özellikteki yeni bir değişkene dönüştürme işlemi gerçekleştirilir. Söz gelimi; “Lise”, “Lisans”, “Lisansüstü” gibi üç farklı kategori eğitim durumlarını gösterebilir. Bu sınıfların metin özelliğindeki verileri “*OneHotEncoder()*” komutunun kullanılması ile birlikte; eğitim_lise, eğitim_lisans, eğitim_lisansüstü gibi isimlerle kategorik değişkenler sayısallaştırılmış ve makine öğrenmesi modellerinin oluşturulması için uygun hâle getirilmiştir. Dahası, Python programlama dilinde kategorik özellikteki verilerin sayısal hâle getirilmesi için “*LabelEncoder()*” fonksiyonu yaygın bir kullanıma sahip diğer bir yöntem olarak öne çıkmaktadır. Bu fonksiyon, “*OneHotEncoder()*” fonksiyonundan farklı olarak, ikili kategorik değişken özelliğindeki verileri 0-1 şeklinde “kredi alabilir” veya “kredi alamaz” gibi iki yanıtlı hedef değişkenlerinin sayısallaştırılması için kullanılmaktadır. Son olarak, veri kümesi içerisindeki değişken değerleri sıralanabilir özellikte ise bu ordinal değişken değerleri için “*OrdinalEncoder()*” fonksiyonu kullanılabilir.

Bu tez kapsamında, üç farklı veri kümesinde yer alan değişkenlerin türlerine bağlı olarak çeşitli kodlama fonksiyonları kullanılmış ve ilgili değişkenler sayısal hâle getirilmiştir. Bu veri kümeleri içerisindeki her bir değişkene özgü uygulanan fonksiyon aşağıda detaylandırılmıştır.

HMEQ Kredi Veri Kümesi: Mortgage ev kredisi veri seti içerisinde 3 kategorik değişken bulunmaktadır. Bu değişkenler iki farklı kodlama ile sayısallaştırılmıştır. Hedef değişkeni olarak bilinen bağımlı değişken “*BAD*” ikili veri tipine sahip olduğu için “*LabelEncoder()*” uygulanırken, kategorik özelliği taşıyan “*Reason*” ve “*Job*” için ise “*OneHotEncoder()*” fonksiyonu uygulanmıştır.

Avustralya Kredi Veri Kümesi: Avustralya kredi veri seti hakkında detaylı bilgi bulunmamakla birlikte, değişkenlere dair de herhangi bir açıklamalara yer verilmediğinden bu değişkenlerin yalnızca veri tipi hakkında bilgi bulunmaktadır. Bu veri tiplerine bakıldığında, 7 kategorik değişken olduğu bilgisine ulaşılmıştır. Bu veri seti kümesinde yer alan tüm kategorik değişkenler için “*OneHotEncoder()*” fonksiyonu uygulanmış olup, hedef değişken için ise “*LabelEncoder()*” fonksiyonu uygulanarak sayısallaştırma işlemleri gerçekleştirilmiştir.

Alman Kredi Veri Kümesi: Veri setleri arasında en detaylı bilgi sahibi olduğumuz veri kümesi olan Alman kredi veri seti içerisinde yer alan “*Credit_history*”, “*Duration*”, “*Employment*”, “*Job*” gibi ordinal veri tipindeki değişkenler için “*OrdinalEncoder()*” fonksiyonu uygulanırken, diğer kategorik özellikteki değişkenler için “*OneHotEncoder()*” fonksiyonu ve hedef değişkeni olan “*Credit Risk*” için ise “*LabelEncoder()*” fonksiyonu uygulanmıştır.

Çalışma kapsamında kullanılmış olan ham veri setlerindeki değişkenlere uygulanan kategorik kodlama sonrasında, değişken sayısında artış meydana gelmiştir. Söz gelimi; Ham veri kümesinde bulunan değişkenler: Alman kredi verisi, HMEQ kredi verisi ve Avustralya kredi verisi olmak üzere üç veri setinde sırasıyla bulunan değişken sayıları; **20, 12 ve 14** iken, sayısallaştırma işlemleri sonucunda değişken sayıları sırasıyla; **50, 18 ve 42** hâlini almıştır.

3.3.2.3. Verilerin Dönüştürülmesi Aşaması

Verilerin dönüştürülmesi aşaması çeşitli fonksiyonlardan faydalanılmaktadır. Bunlar içerisinde en sık kullanılanları minimum-maximum (min-max) ve z-skor' dur. Veri kümesi içerisinde yer alan tüm değerler, belirlenen bir veri aralığına indirgenerek bir ölçeklendirme (dönüşüm) gerçekleştirilir. Bu yolla değişkenler arasındaki ölçek farklılıkları ortadan kaldırarak, küçük ve büyük ölçekli değişkenlerin oluşturulan modelin parametrelerini eşit düzeyde etkilemeleri sağlanmaya çalışılır. Aynı zamanda veri setindeki yüksek gözlem değerlerin veri seti içerisindeki etkisinin azaltılması da bu dönüşümlerle sağlanabilir. Bu tez kapsamında kullanılan üç farklı veri kümesinin her biri için verilerin dönüştürülmesi aşamasında standartlaşma işlemi yapılarak veriler ölçeklendirilmiştir. Bu normalleştirme işlemi olarak da bilinen standartlaşma işleminde; $\mu=0$ ve $\sigma^2=1$ parametrelili normal dağılıma dönüşüm gerçekleştirilir.

Çalışma kapsamında standartlaştırma işlemi, Python programına ait “Scikit-Learn” kütüphanesi içerisinde “*sklearn.preprocessing.StandardScaler()*” fonksiyonu ile gerçekleştirilmiştir.

3.3.3. Veri Kümelerinin Keşfi Aşaması

Veri kümelerinin keşfi aşaması, araştırmanın en önemli noktalarından biridir. Bu aşamanın doğru bir şekilde gerçekleştirilmesi hem araştırmacı tarafından problemin doğru olarak belirlenmesi hem de belirlenen probleme ilişkin olarak yöntem ve tekniklerin doğru seçilmesi için önemlidir. Araştırmacı ilgili probleme ait çözümü gerçekleştirmek için veri setini iyi analiz etmeli ve bu verilerden bir çıkarım yapmalıdır. Temel istatistiksel bilgilerin yeterli olmadığı durumlarda verilerin keşfedilmesi için “*keşfedici veri analizi*” yöntemi kullanılmaktadır. Keşfedici veri analizinin en büyük avantajı, verilerin görselleştirilebilir hâle getirilmesini sağlamasıdır. İlgili veriler, problemin çözümü için araştırmacı tarafından belirlenen yöntemlerle kullanılacak olan model ve/veya modellere karar vermede incelenir. Ancak, sadece bu modellerin elde edilmesi tek başına yeterli değildir. Araştırmacı, veri kümeleri ve analiz süreci hakkında detaylı bilgi sahibi olmalıdır. Aksi takdirde, faydalanılan modeller sonucunda elde edilen çıktı değerlerinin ne anlama geldiği hakkında araştırmacı bilgi veremez. Dolayısıyla, araştırmacının sürece iyi hâkim

olması beklenmektedir. Bu çalışma kapsamında kullanılan üç veri kümesi arasından yer kazanmak amacıyla yalnızca HMEQ kredi veri seti görselleştirilerek bu veri seti hakkında detaylı bilgi sunulmuştur.

3.3.3.1. HMEQ Ev Kredi Veri Kümesinin Keşif Aşaması

Şekil 23: HMEQ Ev Kredisi Korelasyon Matrisi



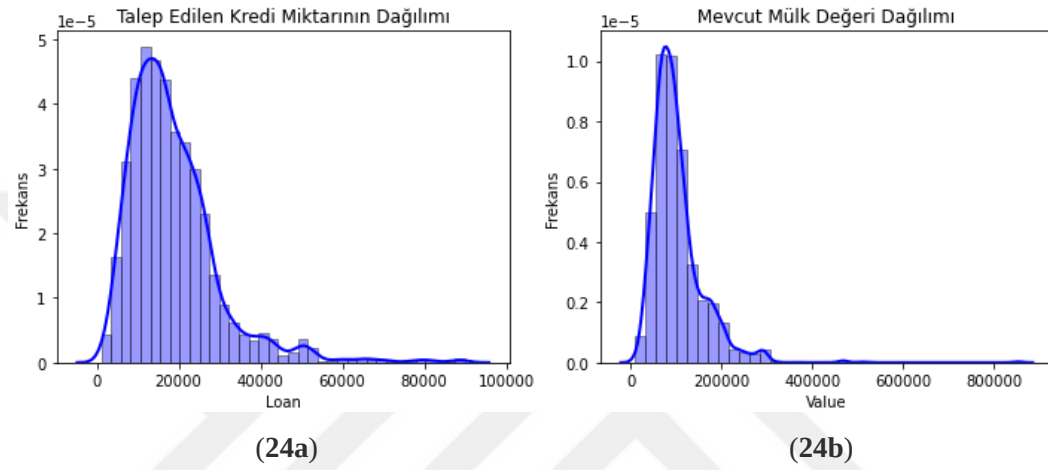
Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Şekil 23'te HMEQ ev kredisi veri kümesinden elde edilmiş olan korelasyon matrisi incelendiğinde, kredi talebinden bulunan borçlunun (müşteri) talep ettiği kredi tutarı (**Loan**) ile borçlunun mülkünün değeri (**Value**) arasındaki korelasyon değeri 0.33'tür. Bu değerın anlamı, talep edilen kredi tutarı ile borçlunun mülk değeri arasında **pozitif yönde düşük kuvvetli doğrusal bir ilişki** olduğudur. Borçlunun ödemesi gereken ev kredi tutarı (**Mortdue**) ile borçlunun mülkünün değeri (**Value**) arasındaki korelasyon değeri 0.79'dur. Bu değerın anlamı, borçlunun ödemesi gereken ev kredisi ile borçlunun mülk değeri arasında **pozitif yönde kuvvetli doğrusal bir ilişki** olduğudur. Borçlunun ödemesi gereken ev kredi tutarı (**Mortdue**) ile kredi taksit

sayısı (**Cln**) arasındaki korelasyon 0.32’dir. Benzer yorumlamaları diğer değişkenler arasında da yapmak mümkündür.

HMEQ Mortgage kredisine ait değişkenlerden bazıları görselleştirilerek detaylandırılmıştır. Bu değişkenlere ilişkin bilgiler aşağıda verilmiştir.

Şekil 24: Talep Edilen Kredi Miktarı ve Mevcut Mülk Değerine ait Grafikler

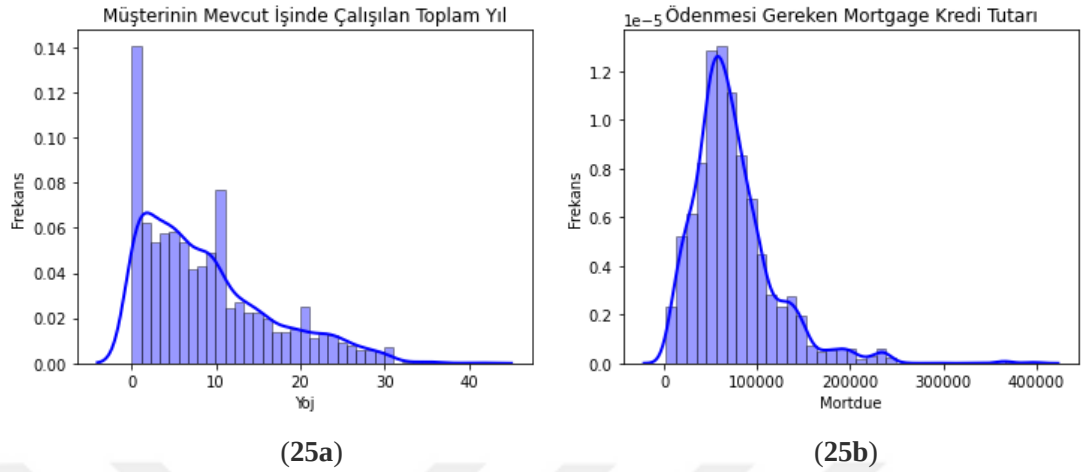


Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Grafik 24a’da görüldüğü üzere, kişilerin talep ettiği kredi tutarlarına ait dağılımın grafiği gösterilmektedir. Kredi talebinde bulunan müşterilerin, talep ettikleri kredi tutarları genel olarak, 1.000 ile 95.000 ABD doları (\$) aralığında olduğu görülmektedir. Grafik detaylı incelendiğinde ise en çok talep edilen kredi tutarının ortalama olarak, 1.000 ile 35.000 ABD doları (\$) aralığında olduğu anlaşılmaktadır.

Grafik 24b’de, borçlunun sahip olduğu mevcut mülkün değerine ait dağılımın grafiği gösterilmektedir. Müşterilere ait olan mevcut mülkün genel olarak, 8.500 ile 88.000 ABD doları (\$) aralığında olduğu görülmektedir. Grafik detaylı incelendiğinde ise müşterilere ait olan mevcut mülkün değeri ortalama olarak, 8.500 ile 300.000 ABD doları (\$) aralığında olduğu anlaşılmaktadır.

Şekil 25: Müşterinin Çalıştığı Toplam Yıl ve Ödenmesi Gereken Kredi Tutarı

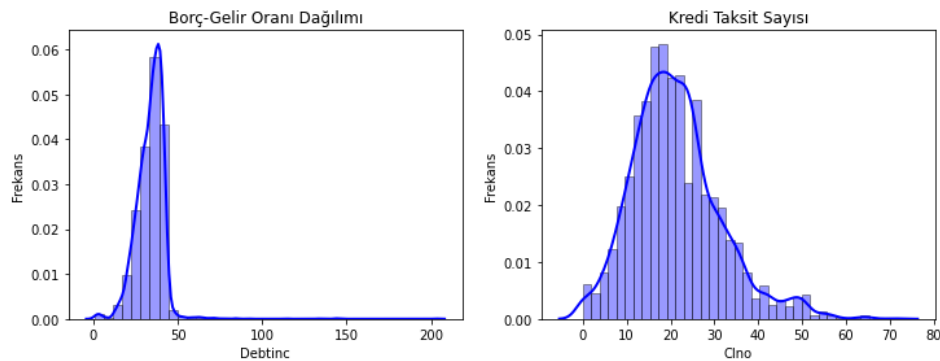


Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Grafik 25a’da, kişilerin mevcut işlerinde toplam çalışılan yıla ait dağılımı göstermektedir. Müşterilerin mevcut işlerinde toplam çalıştıkları yılların 0 ile 45 yıl arasında değiştiği görülmektedir. Grafik detaylı incelendiğinde ise müşterilerin mevcut işlerinde toplam çalıştıkları yılların ortalama olarak 0 ile 25 yıl arasında olduğu anlaşılmaktadır.

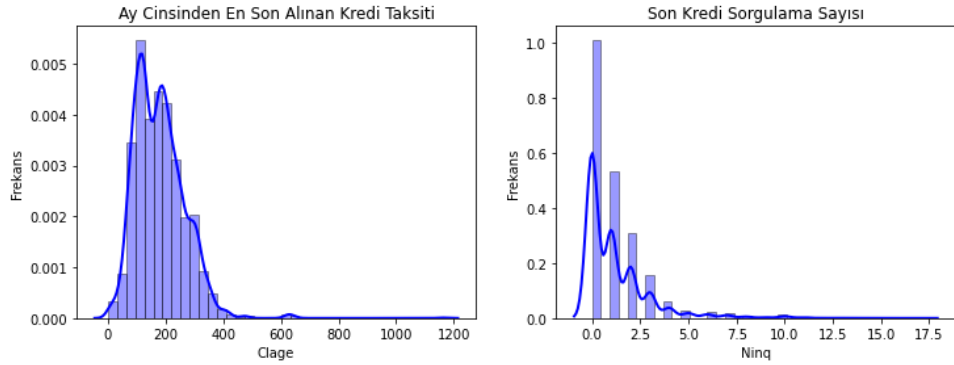
Grafik 25b’de, müşterilerin ödemesi gereken ev kredisi tutarına ait dağılımı göstermektedir. Müşterilerin ödemesi gereken ev kredisi tutarı 2.100 ile 450.000 ABD doları (\$) aralığında değişmektedir. Grafik detaylı incelendiğinde ise ödenmesi gereken mevcut kredi tutarının ortalama olarak 2.100 ile 220.000 ABD doları (\$) aralığında dağıldığı görülmektedir.

Şekil 26: Borç- Gelir Oranı ve Kredi Taksit Sayısına ait Dağılım Grafikleri



Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

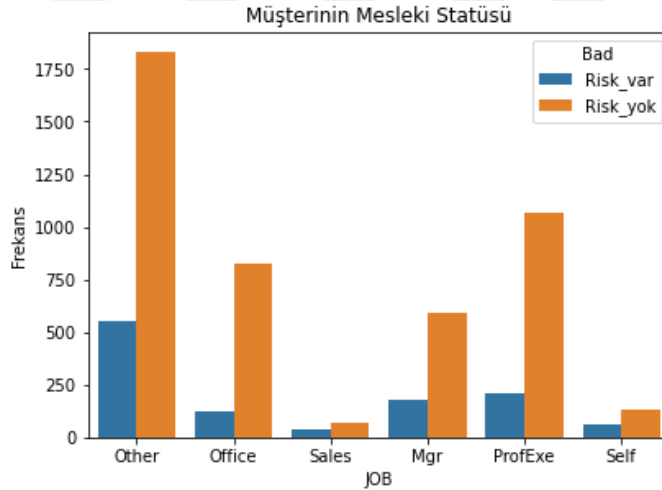
Şekil 27: En Son Alınan Kredi Taksiti ve Son Kredi Sorgulama Sayısı



Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Bu veri setine ait diğer değişkenler için oluşturulmuş Şekil 26 ve Şekil 27 ‘de verilmiştir. Benzer yorumlamalar bu grafiklerden yapılabilir.

Şekil 28: Mesleki Statülerine Göre Kredi Riskinin İncelenmesi

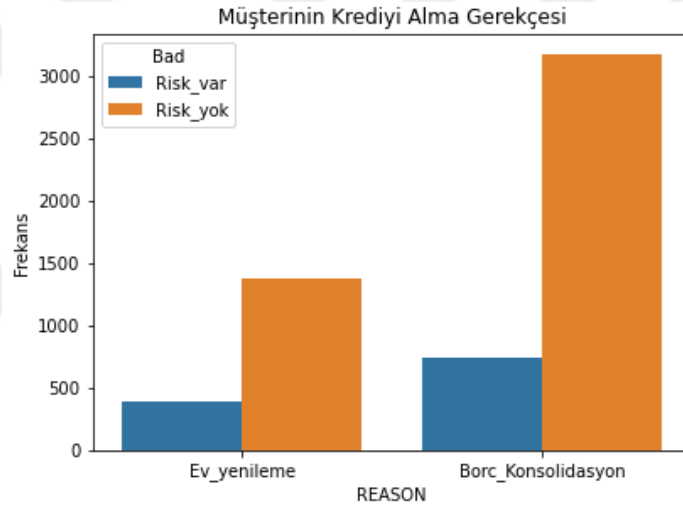


Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Şekil 28’de müşterilerin mevcut mesleki pozisyonlarına göre kredi risk durumuna ait grafiği göstermektedir. Müşterilerin mesleki pozisyonları ve kredi risk durumu incelendiğinde, mesleği profesyonel yönetici (Professional Executive) olan müşteriler arasından yaklaşık 1100 kişinin riskli olmayıp kredisi kabul edilirken, yaklaşık 200 kişinin ise riskli olduğu için kredisi kabul edilmemiştir. Mesleği memur (Office) olan müşteriler arasından yaklaşık 800 kişinin kredi talebinin riskli olmadığı

için kredi talebi kabul edilirken, yaklaşık 150 kişinin ise riskli olduğu için kredi talebi olumsuz sonuçlanmıştır. Ek olarak, mesleği yönetici (Mgr) olan müşteriler arasında yaklaşık 600 kişinin kredi talebinin riskli olmadığı için olumlu sonuçlanırken, yaklaşık 200 kişinin ise riskli olduğu için kredi talebi olumsuz sonuçlanmıştır. Ayrıca, belirtilen meslek grubu dışında kalan diğer (Other) mesleklerle ilgilenen müşteriler arasında kredi talebinden bulunan yaklaşık 2200 kişinin kredisinin riskli olmadığı için kabul edilirken, yaklaşık 600 kişi riskli olduğu için kredisi kabul edilmemiştir. Bu yorumlardan anlaşılabacağı üzere müşterilerin mesleki statüleri ve itibarları arttıkça, bankalara yapılan kredi talepleri olumlu sonuçlandığı anlaşılmaktadır.

Şekil 29: Müşterinin Krediyi Alma Gerekçesine Göre Kredi Risk Durumu

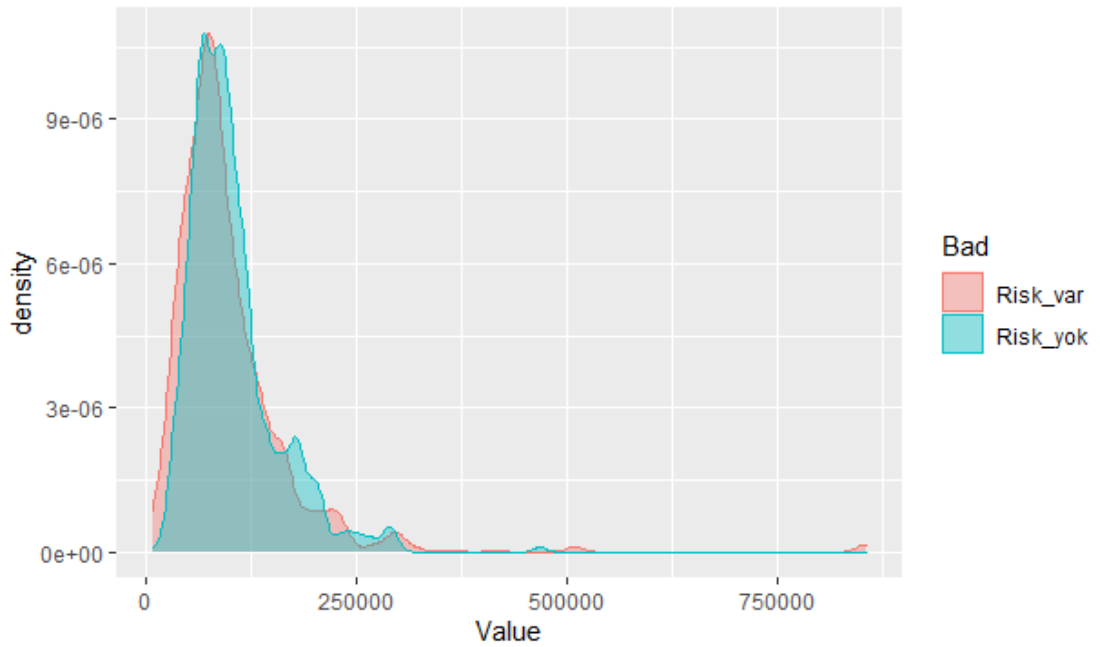


Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Kredi talebinde bulunan müşterilerin, kredi alma gerekçesini gösteren grafik yukarıda Şekil 29'da verilmiştir. Müşteriler farklı amaçlar doğrultusunda kredi talebinde bulunmuşlardır. Burada, müşterilerin kredi talebinden bulunma nedenlerinden ilki, ev yenileme amacıyla alınan krediler iken, ikincisi ise borç konsolidasyonudur. Kredi talebinde bulunan kişiler, tadilat, yenileme ve dekorasyon gibi çeşitli nedenlerden dolayı ev yenileme kredisi almaktadır. Borçlu tarafından alınmış olan borcun ödenmeyip, belirli yükümlülüklerle birlikte yeniden yapılandırılarak ileri bir tarihte ödenmesi anlamına gelen borç konsolidasyonu, müşteriler tarafından başvuru olan diğer bir kredi talebini oluşturmaktadır. Şekil 29'dan anlaşıldığı üzere, ev yenileme amacıyla kredi talebinde bulunan yaklaşık 1400

müşterinin, kredi talebinin gerçekleştirilmesinde herhangi bir riskli durumun olmadığı ve dolayısıyla kredi talebinin olumlu sonuçlandığı görülmektedir. Ek olarak, ev yenileme amacıyla kredi talebinde bulunan yaklaşık 450 müşterinin ise kredi talebinin gerçekleştirilmesinin riskli olduğu ve böylece, kredi talebinin olumsuz sonuçlandığı görülmektedir. Ayrıca, aynı grafik üzerinde gösterilen diğer bir kredi talebini oluşturan borç konsolidasyonu amacıyla kredi talebinde bulunan yaklaşık olarak 3400 müşteriye ait kredi talebinin gerçekleştirilmesinde herhangi bir riskin olmadığı ve dolayısıyla kredi talebinin olumlu sonuçlandığı görülmektedir. Buna ilaveten, kredi talebinde bulunan yaklaşık 750 müşterinin ise kredi talebinin gerçekleştirilmesinin riskli olduğu ve dolayısıyla kredi talebinin olumsuz sonuçlandığı da anlaşılmaktadır.

Şekil 30: Mevcut Mülkün Değerine Göre Kredi Risk Durumunun İncelenmesi

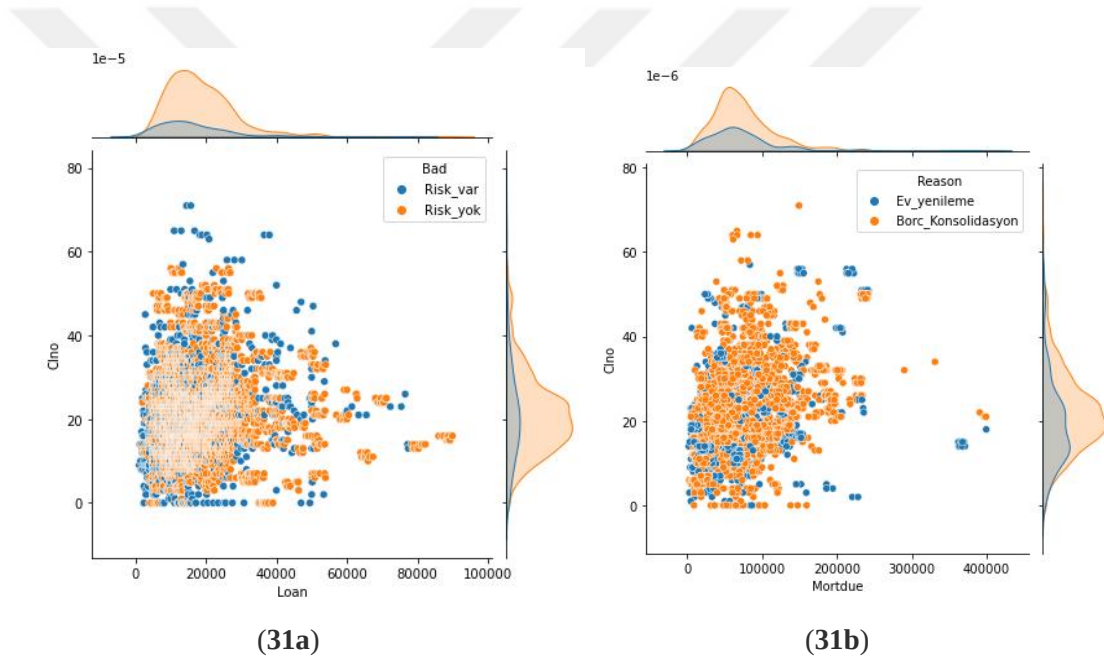


Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Mevcut mülkünün değeri ile kredi riski durumunu gösteren grafik yukarıda Şekil 30’da gösterilmiştir. Şekil 30 genel olarak incelendiğinde, kredi risk durumunun birbirine çok yakın olduğu görülmektedir. Şekil 30’da görüldüğü üzere, müşterilerin mevcut mülkünün değeri ile kredi risk durumu incelendiğinde, kredi riskinin olumsuz sonuçlandığı durumlar, olumlu sonuçlandığı durumdan daha fazla olduğu anlaşılmaktadır. Yaklaşık 1.000 ile 6.500 ABD doları (\$) gibi oldukça düşük bir aralık

incelendiğinde, kredi riskinin çok yüksek olduğu görülmektedir. Öte yandan, müşterinin mevcut mülk değeri 6.500 ile 10.000 ABD doları (\$) aralığında iken riskli kredi durumunun riskli olmayan kredi durumundan daha az olduğu görülmektedir. Ek olarak, müşteri mülkü değerinin yaklaşık 12.000 ile 15.000 ABD doları (\$) aralığı, 20.000 ile 23.000 ABD doları (\$) aralığı, 24.000 ile 36.000 ABD doları (\$) aralığında da benzer durumların olduğu açıktır. Başka bir deyişle, bu söz konusu aralıklarda müşterilerin kredilerin daha düşük riskli olduğu ve bu yüzden bankalarca kredi talepleri kabul edildiği sonucuna varılmıştır.

Şekil 31: Talep Edilen Kredi ile Ödenmesi Gereken Kredi Tutarı ve Sayısı



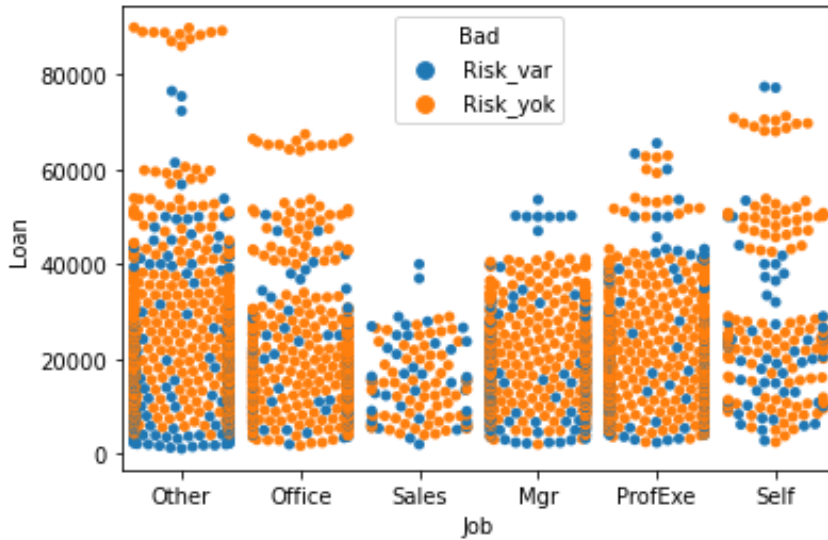
Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Şekil (31a) 'da gösterilen, kredi talebinde bulunan müşterilerin, talep etmiş olduğu kredi tutarı ile kredi taksit sayısına göre kredi risk durumu yukarıda **jointplot** grafiğinde gösterilmiştir. Şekil (31a) incelendiğinde, ilk göze çarpan nokta, flu bölgedir. Bu bölge, HMEQ veri kümesinde bulunan gözlem değerlerinin yoğun olduğu değer aralığını göstermektedir. Bu grafik, iki farklı değişkene ait ilişkisel yorumlamanın yapılmasına olanak sağlar. Talep edilen kredi tutarı ve taksit sayısı arasında pozitif yönde ancak orta düzeyde bir ilişki olduğu görülmektedir. Genel olarak, talep edilen kredi tutarının 1.000 ile 35.000 ABD doları ve kredi taksit sayısının

10 ile 35 olduğu aralıkta yoğunluk olduğu görülmektedir. Kredi talebinde bulunan müşterilerin, düşük miktarlarda talep ettiği kredilerin taksit sayısındaki artışa bağlı olarak, kredi riskini artırdığı söylenebilir. Ek olarak, 60.000 ABD doları ve üstü kredi talepleri incelendiğinde, taksit sayısının maksimum 25 olduğu ve bu aralıkta genel olarak kredilerin risk teşkil etmediği görülmektedir.

Şekil (31b) 'de gösterilen, ödenmesi gereken ev kredisi tutarı ile kredi taksit sayısına göre kredi alma gerekçesi yukarıda **jointplot** grafiğinde gösterilmiştir. Şekil (31b) genel olarak incelendiğinde, 1.000 ile 200.000 ABD doları ve 10 ile 40 kredi taksit sayısı aralığındaki bölgenin yoğun olduğu ve borç konsolidasyonu amacıyla kredi alındığı görülmektedir. Bu bölge, HMEQ veri kümesinde bulunan gözlem değerlerinin yoğun olduğu değer aralığını göstermektedir. Bu grafik, iki farklı değişkene ait ilişkisel yorumlamanın yapılmasına olanak sağlar. Müşterilerin ödemesi gereken ev kredisi tutarı ile kredi taksit sayısı arasında pozitif yönde ancak orta düzeyde bir ilişki olduğu anlaşılmaktadır.

Şekil 32: Müşterinin Mesleği ve Kredi Talep Tutarına Göre Kredi Riski



Kaynak: Tez Yazarı Tarafından Hazırlanmıştır.

Şekil 32, kredi talebinde bulunan müşterilerin, talep etmiş olduğu kredi tutarı ile mesleki statülerine göre kredi riski durumunu temsil eden **swarmplot** grafiği verilmiştir. Şekil 32'de yer alan meslek grupları dışında kalan diğer (Other) meslek grubu incelendiğinde, yaklaşık 1.000 ile 3.000 ABD doları (\$) aralığı ve yaklaşık 40.000 ile

42.000 ABD doları (\$) aralığında müşteri talebinde bulunan müşterilerin kredilerinin kabul edilmediği görülmektedir. Ek olarak, benzer bir duruma yaklaşık 75.000 ABD doları seviyelerinde rastlandığını söylemek mümkündür. Öte yandan, diğer meslek grubuna ait genel bir değerlendirme yapılacak olursa, kredi talebinin kabul edildiği durumun daha fazla olduğu görülmektedir. Mesleği memur (Office) olan ve kredi talebinde bulunan müşteriler incelendiğinde, kredi için talep edilen tutarın artması riski artırmamış ve bankalarca kredi talepleri olumlu sonuçlanmıştır. Mesleği satıcı (Sales) olan ve kredi talebinde bulunan müşteriler incelendiğinde ise yaklaşık 1.000 ile 30.000 ABD doları (\$) aralığında riskli olan ve riskli olmayan kredilerin sayılarının eşit olduğu ancak yaklaşık 35.000 ABD doları (\$) ve üstü için kredilerin riskli olduğu ve kredi taleplerinin kabul edilmediği görülmektedir. Mesleği yönetici (Mgr) olan müşterilerin, yaklaşık 1.000 ile 45.000 ABD doları (\$) aralığındaki kredi taleplerinin genel olarak riskli bir durum olmadığı için kabul edildiği ancak 45.000 ABD doları ve üstü kredi taleplerinin kabul edilmediği anlaşılmaktadır. Ek olarak mesleği profesyonel yönetici (Professional Executive) olan müşteriler ile mesleği yönetici olanlar kıyaslandığında, 45.000 ABD doları (\$) seviyesine kadar risk durumlarının benzerlik gösterdiği ancak 45.000 ABD doları ve üstü kredi talepleri için profesyonel yöneticiler açısından daha risksiz bir durumun olduğu ve kredi taleplerinin kabul edildiği söylenebilir. Kendi işi (Self) olan müşterilerin, yaklaşık 1.000 ile 25.000 ABD doları (\$) aralığında kredi risk durumlarının benzer olduğu, yaklaşık 30.000 ile 40.000 ABD doları (\$) aralığı ile 78.000 ABD doları ve üstü kredi taleplerinin ise riskli olduğu için kabul edilmediği söylenebilir.

3.3.4. Eğitim ve Test Veri Kümelerinin Hazırlanması Aşaması

Makine öğrenmesi için oluşturulacak tahmin modellerinin eğitilebilmesi için ham veri kümesi eğitim ve test veri kümeleri olmak üzere iki bölüme ayrılmaktadır. Ham veri kümelerinin büyük bir çoğunluğu modellerin eğitilmesi için kullanılırken, geriye kalan kısım ise modelin test edilmesi amacıyla kullanılmaktadır. Ham verilerden öğrenme için ayrılan veriler “eğitim veri seti” olarak adlandırılırken, modellerin test edilmesi için kullanılan veriler “test veri seti” olarak adlandırılır. Model eğitilirken gözlem sayısının büyük olması tahmin edilen model parametrelerinin daha kararlı olmasına ve buradan elde edilen sonuçların genelleme

yeteneğinin güçlü olmasına yol açar. Basit bir ifadeyle, veri büyüklüğünün artması makine öğrenme tekniklerinin en büyük sorunu olan aşırı uyum sağlama probleminden kaçınmada avantaj sağlayacaktır. Böylece genelleme yeteneği artmış modeller oluşturulacağından daha önce hiç görmediği verilerde de iyi performans gösterme şansı artacaktır. Bu tez kapsamında kullanılmış olan Alman kredi veri kümesi, Avustralya kredi veri kümesi ve HMEQ kredi veri kümesi olmak üzere üç farklı veri setinin sırasıyla; **800, 565, 4887**'si eğitim verisi ve **200, 125, 1073**'ü ise test verisi olarak ayrılmıştır.

3.3.5. Yeniden Örneklemme Tekniklerinin Uygulanması Aşaması

Gerçek yaşam problemlerine ait veri kümelerinde karşılaşılan en temel problemlerden biri, sınıflandırma problemleri için verilerin eşit dağılmaması sonucunda ortaya çıkan “dengesiz veri seti” durumudur. Verilerin dengeli bir yapıya sahip olmaması model performansı üzerinde olumsuz bir etki oluşturmaktadır. Bu durum kullanılan sınıflandırıcılara ait parametrelerin daha çok çoğunluk sınıfını öğrenecek şekilde tahmin edilmesine yol açar. Dengesiz bir veri yapısına sahip olan veri kümelerinin kullanılması ile oluşturulan modellerden veri kümesi içerisinde yer alan azınlık sınıfına karşın çoğunluk sınıfının fazla olması nedeniyle çoğunluk sınıfı, model üzerinde daha hâkim olacak ve böylece, modelin yanıltıcı bir tahmin yapmasına sebep olacaktır. Basit bir ifadeyle, dengesiz veri kümelerinde çoğunluk sınıfının azınlık sınıfına göre daha baskın olması modelin doğruluğu üzerinde yanıltıcı bir etki oluşturmaktadır. Bahsi geçen problemlerin önüne geçebilmek için verilerin analiz öncesi dengeli hâle getirilmesine yönelik önerilmiş çeşitli yöntemler bulunmaktadır. Bu tez kapsamında kullanılmış olan dengeli olmayan üç farklı veri setinin dengeli hâle getirilmesinde; Rassal Yukarı Örneklemme, Rassal Aşağı Örneklemme, Dengeli Özyinelemeli Birleştirme Sınıflandırıcısı, SMOTE-ENN Combine, SMOTE-Tomek Links gibi hem aşağı örneklemme yöntemleri hem de yukarı örneklemme yöntemleri ile birlikte bunların birleştirilmesi ile oluşan teknikler kullanılmıştır.

3.4. SONUÇLARIN DEĞERLENDİRİLMESİ VE YORUMLANMASI

3.4.1. Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları

Tablo 5: Dengelenmemiş Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları

Ham Veri Kümesi	Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Tekil Modeller	KNN	0,7250	0,4137	0,5333	0,4660	0,2889
	NB	0,6500	0,6551	0,4318	0,5205	0,2770
	Log	0,7100	0,6551	0,5000	0,5671	0,3623
	SVM	0,7500	0,5517	0,5714	0,5614	0,3867
	MLP	0,7250	0,6551	0,5205	0,5801	0,3852
	DT	0,7000	0,3793	0,4782	0,4230	0,2267
Topluluk Öğrenmesi Modelleri	RF	0,7850	0,4827	0,6829	0,5656	0,4397
	GBDT	0,7400	0,5344	0,5535	0,5438	0,3622
	XTrees	0,7750	0,4482	0,6666	0,5360	0,4085
	Hard	0,7450	0,5344	0,5636	0,5486	0,3714
	Soft	0,7400	0,5344	0,5535	0,5438	0,3622

Tablo 5’te, Dengelenmemiş Alman kredi veri kümesi tekil ve topluluk öğrenmesi modellerine ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Alman kredi veri setinin dengelenmeden önceki sonuçları hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise tekil ve topluluk öğrenmesi modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 5’te gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %78 ile Rassal ormanlar algoritmasında ölçülmüştür. Fakat, sadece doğruluk kriteri üzerinden değerlendirme yapılması, kullanılmış olan makine öğrenmesi yöntemlerinin performansları açısından yanlış bir çıkarımda bulunma sebebiyet verebilir. Bu nedenle, tüm değerlendirme kriterleri göz önüne alınarak doğru bir değerlendirme yapılmalıdır. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %68’dir. Hassasiyet açısından %65,5 ile Naive Bayes, Lojistik regresyon ve Çok Katmanlı Algılayıcı’da ölçülmüştür. F1-skoru açısından ise en iyi performans Çok Katmanlı Algılayıcı’da ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak

değerlendirildiğinde, müşterilerin kredi riskini tahmin etme hususunda en etkili performansı Çok Katmanlı Algılayıcı göstermiştir.

Tablo 6: Senaryo 1: Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları

RUS ve ROS Tekniği	Tekil Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Random Under Sampler (RUS)	KNN	0,6550	0,7413	0,4432	0,5548	0,3278
	NB	0,6250	0,6896	0,4123	0,5161	0,2617
	Log	0,6550	0,7241	0,4421	0,5490	0,3188
	SVM	0,6700	0,7241	0,4565	0,5599	0,3387
	MLP	0,6500	0,6896	0,4347	0,5333	0,2944
	DT	0,6250	0,8793	0,4285	0,5762	0,3701
Random Over Sampler (ROS)	KNN	0,6800	0,4655	0,4500	0,4576	0,2308
	NB	0,6300	0,7413	0,4215	0,5375	0,2958
	Log	0,6500	0,7241	0,4375	0,5454	0,3123
	SVM	0,7200	0,1206	0,5833	0,2000	0,1633
	MLP	0,7050	0,4137	0,4897	0,4485	0,2508
	DT	0,6600	0,5517	0,4324	0,4848	0,2405

Tablo 6’da, rassal aşağı örnekleme ve rassal yukarı örnekleme tekniği ile tekil modellere ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Alman kredi veri setinin dengelenmesi amacıyla hem rassal aşağı örnekleme hem de rassal yukarı örnekleme teknikleri üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise tekil modellere ait tahmin sonuçlarını rapor etmiştir.

Tablo 6’da gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %72 ile ROS-SVM tekil modeller kombinasyonunda ölçülmüştür. Fakat, sadece doğruluk kriteri üzerinden değerlendirme yapılması, kullanılmış olan makine öğrenmesi yöntemlerinin performansları açısından yanlış bir çıkarımda bulunma sebebiyet verebilir. Bu nedenle, tüm değerlendirme kriterleri göz önüne alınarak doğru bir değerlendirme yapılmalıdır. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %52 ile ROS-SVM kombinasyonunda ölçülürken, hassasiyet, F1-skor ve MCC açısından ise en iyi performans RUS-DT kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak

değerlendirildiğinde, senaryo 1’de, müşterilerin kredi riskini tahmin etme hususunda en etkili performansı RUS-DT göstermiştir. Başka bir ifadeyle, Alman kredi veri kümesi ile müşterilerin kredi riskinin tahmin edilmesi hususunda mevcut veri setinde örnekler yeni örneklerin eklenmesine gerek olmadığı ve tekil modellerden karar ağaçları ile tahmin yapılabileceği sonucuna ulaşılır.

Tablo 7: Senaryo 2: Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları

RUS ve ROS Tekniği	Topluluk Öğrenme Modelleri	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Random Under Sampler (RUS)	RF	0,7050	0,6896	0,4938	0,5755	0,3705
	GBDT	0,6650	0,7413	0,4526	0,5620	0,3409
	XTrees	0,6900	0,7413	0,4777	0,5810	0,3743
	Hard	0,6700	0,7413	0,4574	0,5657	0,3475
	Soft	0,6950	0,7068	0,4823	0,5734	0,3644
Random Over Sampler (ROS)	RF	0,7650	0,5172	0,6122	0,5607	0,4045
	GBDT	0,7650	0,4827	0,6222	0,5436	0,3944
	XTrees	0,7500	0,3965	0,6052	0,4791	0,3364
	Hard	0,7650	0,5172	0,6122	0,5607	0,4045
	Soft	0,7600	0,5172	0,6000	0,5555	0,3944

Tablo 7’de, rassal aşağı örnekleme ve rassal yukarı örnekleme tekniği ile topluluk öğrenme modellerine ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Alman kredi veri setinin dengelenmesi amacıyla hem rassal aşağı örnekleme hem de rassal yukarı örnekleme teknikleri üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise topluluk öğrenme modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 7’de gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %76,5 ile ROS-RF, ROS-GBDT ve ROS-Hard topluluk öğrenme modelleri kombinasyonunda ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, hassasiyet açısından performans ölçüm değerine ait en yüksek sonuç %74 ile RUS-GBDT, RUS-XTrees ve RUS-Hard kombinasyonunda ölçülürken, F1-skoru açısından ise en iyi performans RUS-XTrees kombinasyonunda ölçülmüştür. Öte yandan, kesinlik açısından performans ölçüm değerine ait en yüksek

sonuç ROS-GBDT kombinasyonunda ölçülürken, MCC açısından ise en iyi performans %40 ile hem ROS-RF hem de ROS-Hard kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 2’de, müşterilerin kredi riskini tahmin etme hususunda en etkili performansı RUS-XTrees göstermiştir. Başka bir ifadeyle, Alman kredi veri kümesi ile müşterilerin kredi riskinin tahmin edilmesi hususunda mevcut veri setindeki örneklere ek olarak, yeni örneklerin eklenmesine gerek olmadığı ve topluluk öğrenme modellerinden ise Extra ağaçlar (XTrees) tekniği ile tahmin yapılabileceği sonucuna ulaşıldı.

Tablo 8: Senaryo 3: Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları

Aşağı ve Yukarı Örneklemeye Yöntemleri Kombinasyonu ile Örnekleyiciler Topluluğu	Tekil Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
SMOTE-ENN Combine	KNN	0,6300	0,6379	0,4111	0,5000	0,2414
	NB	0,6800	0,4655	0,4500	0,4576	0,2308
	Log	0,5800	0,7758	0,3879	0,5172	0,2536
	SVM	0,7200	0,4482	0,5200	0,4814	0,2926
	MLP	0,5950	0,8103	0,4017	0,5371	0,2922
	DT	0,6250	0,6551	0,4086	0,5033	0,2436
SMOTE-Tomek Links	KNN	0,6800	0,4482	0,4482	0,4482	0,2229
	NB	0,5600	0,7241	0,3684	0,4883	0,1989
	Log	0,6700	0,6724	0,4534	0,5416	0,3129
	SVM	0,7200	0,1551	0,6000	0,2465	0,1945
	MLP	0,7050	0,5517	0,4923	0,5203	0,3093
	DT	0,6800	0,6034	0,4605	0,5223	0,2942
Balanced Bagging Classifier (BBC)	KNN	0,7250	0,4137	0,5333	0,4660	0,2889
	NB	0,6500	0,6551	0,4318	0,5205	0,2770
	Log	0,7100	0,6551	0,5000	0,5671	0,3623
	SVM	0,7500	0,5517	0,5714	0,5614	0,3867
	MLP	0,7250	0,6551	0,5205	0,5801	0,3852
	DT	0,7000	0,3793	0,4782	0,4230	0,2267

Tablo 8’de, aşağı ve yukarı örneklemeye yöntemleri kombinasyonu ile örnekleyiciler topluluğu ve tekil modellere ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Alman kredi veri setinin dengelenmesi amacıyla

aşağı ve yukarı örnekleme yöntemlerinin kombinasyonu ile örnekleyiciler topluluğunu kullanılmış ve sınıflandırma başarılarının test edilmesi için ise tekil modellere ait tahmin sonuçlarını rapor etmiştir.

Tablo 8’de gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %75 ile Balanced Bagging Classifier-SVM çiftinden oluşan model kombinasyonu vermiştir. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %57,14 ile Balanced Bagging Classifier-SVM tekil modeller kombinasyonunda ölçülürken, hassasiyet performans ölçüm değerine ait en yüksek sonuç %81,03 ile SMOTE-ENN Combine- MLP kombinasyonunda ölçülmüştür. Öte yandan, F1-skoru açısından en iyi ölçüm değeri %58,01 ile Balanced Bagging Classifier -MLP kombinasyonunda ölçülürken, MCC açısından en iyi performans %38,67 ile Balanced Bagging Classifier-SVM kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 3’te Balanced Bagging Classifier-MLP müşterilerin kredi riskini tahmin etme hususunda etkili performans göstermiştir.

Tablo 9: Senaryo 4: Alman Kredi Veri Setine ait Performans Ölçüm Sonuçları

Aşağı ve Yukarı Örnekleme Yöntemleri Kombinasyonu ile Örnekleyiciler Topluluğu	Topluluk Öğrenmesi Modelleri	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
SMOTE-ENN Combine	RF	0,6800	0,7068	0,4659	0,5616	0,3436
	GBDT	0,6800	0,7931	0,4693	0,5897	0,3875
	XTrees	0,6600	0,6896	0,4444	0,5405	0,3078
	Hard	0,6800	0,7241	0,4666	0,5675	0,3521
	Soft	0,6650	0,6551	0,4470	0,5314	0,2975
SMOTE- Tomek Links	RF	0,7550	0,5000	0,5900	0,5420	0,3789
	GBDT	0,7400	0,4800	0,5600	0,5185	0,3435
	XTrees	0,7650	0,5000	0,6100	0,5523	0,3994
	Hard	0,7650	0,5100	0,6100	0,5607	0,4045
	Soft	0,7650	0,5300	0,6000	0,5688	0,4098
Balanced Bagging Classifier (BBC)	RF	0,7850	0,4827	0,6829	0,5656	0,4397
	GBDT	0,7400	0,5344	0,5535	0,5438	0,3622
	XTrees	0,7750	0,4482	0,6666	0,5360	0,4085
	Hard	0,7550	0,4655	0,6000	0,5242	0,3681
	Soft	0,7500	0,4482	0,5909	0,5098	0,3521

Tablo 9’da, aşağı ve yukarı örnekleme yöntemleri kombinasyonu ile örnekleyciler topluluğu ve topluluk öğrenmesi modellerine ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Alman kredi veri setinin dengelenmesi amacıyla aşağı ve yukarı örnekleme yöntemlerinin kombinasyonu ile örnekleyciler topluluğunu kullanılmış ve sınıflandırma başarılarının test edilmesi için ise topluluk öğrenmesi modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 9’da gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %78,5 ile Balanced Bagging Classifier-RF çiftinden oluşan model kombinasyonu vermiştir. Fakat, sadece doğruluk kriteri üzerinden değerlendirme yapılması, kullanılmış olan makine öğrenmesi yöntemlerinin performansları açısından yanlış bir çıkarımda bulunmaya sebebiyet verebilir. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %68,29 ile Balanced Bagging Classifier-RF kombinasyonunda ölçülürken, hassasiyet performans ölçüm değerine ait en yüksek sonuç %79,31 ile SMOTE-ENN Combine-GBDT kombinasyonunda ölçülmüştür. Öte yandan, MCC açısından en iyi performans %43,97 ile Balanced Bagging Classifier-RF kombinasyonunda ölçülürken, F1-skoru açısından ise en iyi ölçüm değeri %58,97 ile SMOTE-ENN Combine-GBDT kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 4’te SMOTE-ENN Combine-GBDT müşterilerin kredi riskini tahmin etme hususunda etkili performans göstermiştir.

3.4.2. Avustralya Kredi Veri Setine ait Performans Ölçüm Sonuçları

Tablo 10: Dengelenmemiş Avustralya Kredi Veri Setine ait Performans Ölçüm Sonuçları

Ham Veri Kümesi	Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Tekil Modeller	KNN	0,8550	0,7692	0,9090	0,8333	0,7144
	NB	0,8405	0,8000	0,8524	0,8253	0,6801
	Log	0,8840	0,8769	0,8769	0,8769	0,7673
	SVM	0,8695	0,8615	0,8615	0,8615	0,7382
	MLP	0,8913	0,8461	0,9166	0,8799	0,7830
	DT	0,8695	0,8307	0,8852	0,8571	0,7386
Topluluk Öğrenmesi Modelleri	RF	0,8550	0,7846	0,8947	0,8360	0,7120
	GBDT	0,8695	0,8615	0,8615	0,8615	0,7382
	XTrees	0,8768	0,8307	0,9000	0,8640	0,7537
	Hard	0,8768	0,8153	0,9137	0,8617	0,7552
	Soft	0,8768	0,8153	0,9137	0,8617	0,7552

Tablo 10’da, Dengelenmemiş Avustralya kredi veri kümesi tekil ve topluluk öğrenmesi modellerine ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Avustralya kredi veri setinin dengelenmeden önceki sonuçları hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise tekil ve topluluk öğrenmesi modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 10’da gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %89 ile Çok Katmanlı Algılayıcı algoritmasında ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, hassasiyet performans ölçüm değerine ait en yüksek sonuç %87,69 ile lojistik regresyonda ölçülmüştür. Kesinlik, F1-skoru ve MCC açısından ise en iyi performans Çok Katmanlı Algılayıcı’da ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, müşterilerin kredi riskini tahmin etme hususunda en etkili performansı Çok Katmanlı Algılayıcı göstermiştir.

Tablo 11: Senaryo 1: Avustralya Kredi Veri Seti Performans Ölçüm Sonuçları

RUS ve ROS Tekniği	Tekil Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Random Under Sampler (RUS)	KNN	0,8478	0,7692	0,8928	0,8264	0,6983
	NB	0,8260	0,8000	0,8253	0,8125	0,6506
	Log	0,8768	0,8769	0,8636	0,8702	0,7530
	SVM	0,8768	0,8769	0,8636	0,8700	0,7500
	MLP	0,8550	0,8150	0,8688	0,8412	0,7093
	DT	0,8913	0,9076	0,8676	0,8872	0,7831
Random Over Sampler (ROS)	KNN	0,8550	0,7846	0,8947	0,8360	0,7120
	NB	0,8405	0,8000	0,8524	0,8253	0,6801
	Log	0,8840	0,8615	0,8888	0,8750	0,7672
	SVM	0,8550	0,8307	0,8571	0,8437	0,7089
	MLP	0,8695	0,8307	0,8852	0,8571	0,7386
	DT	0,8333	0,7846	0,8500	0,8160	0,6659

Tablo 11’de, rassal aşağı örnekleme ve rassal yukarı örnekleme tekniği ile tekil modellere ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Avustralya kredi veri setinin dengelenmesi amacıyla hem rassal aşağı örnekleme hem de rassal yukarı örnekleme teknikleri üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise tekil modellere ait tahmin sonuçlarını rapor etmiştir.

Tablo 11’de gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %89,1 ile RUS-DT tekil modeller kombinasyonunda ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %89,4 ile ROS-KNN kombinasyonunda ölçülürken, hassasiyet, F1-skoru ve MCC açısından ise en iyi performans RUS-DT kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 1’de, müşterilerin kredi riskini tahmin etme hususunda en etkili performansı RUS-DT göstermiştir. Başka bir ifadeyle, Avustralya kredi veri kümesi ile müşterilerin kredi riskinin tahmin edilmesi hususunda mevcut veri setinde örneklere yeni örneklerin eklenmesine gerek olmadığı ve tekil modellerden karar ağaçları ile tahmin yapılabileceği sonucuna ulaşılır.

Tablo 12: Senaryo 2: Avustralya Kredi Veri Seti Performans Ölçüm Sonuçları

RUS ve ROS Tekniği	Topluluk Öğrenmesi Modelleri	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Random Under Sampler (RUS)	RF	0,8623	0,8461	0,8593	0,8527	0,7235
	GBDT	0,8623	0,8461	0,8500	0,8527	0,7235
	XTrees	0,8840	0,8400	0,9000	0,8730	0,7678
	Hard	0,8623	0,8153	0,8833	0,8480	0,7244
	Soft	0,8768	0,8615	0,8750	0,8682	0,7526
Random Over Sampler (ROS)	RF	0,8695	0,8461	0,8730	0,8593	0,7381
	GBDT	0,8478	0,7692	0,8928	0,8264	0,6983
	XTrees	0,8768	0,8307	0,9000	0,8640	0,7537
	Hard	0,8695	0,8307	0,8852	0,8571	0,7386
	Soft	0,8695	0,8461	0,8730	0,8593	0,7381

Tablo 12’de, rassal aşağı örnekleme ve rassal yukarı örnekleme tekniği ile topluluk öğrenmesi modellerine ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Avustralya kredi veri setinin dengelenmesi amacıyla hem rassal aşağı örnekleme hem de rassal yukarı örnekleme teknikleri üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise topluluk öğrenme modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 12’de gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %88,4 ile RUS-XTrees kombinasyonunda ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, hassasiyet açısından performans ölçüm değerine ait en yüksek sonuç %86,1 ile RUS-Soft kombinasyonunda ölçülürken, F1-skoru açısından ise en iyi performans RUS-XTrees kombinasyonunda ölçülmüştür. Öte yandan, kesinlik açısından performans ölçüm değerine ait en yüksek sonuç %90 ile hem RUS-XTrees hem de ROS-XTrees kombinasyonunda ölçülürken, MCC açısından ise en iyi performans %76 ile RUS-XTrees kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 2’de, müşterilerin kredi riskini tahmin etme hususunda en etkili performansı RUS-XTrees göstermiştir. Başka bir ifadeyle, Avustralya kredi veri kümesi ile müşterilerin kredi riskinin tahmin edilmesi hususunda mevcut veri

setindeki örneklere yeni örneklerin eklenmesine gerek olmadığı ve topluluk öğrenmesi modellerinden XTrees ile tahmin yapılabileceği sonucuna ulaşılır.

Tablo 13: Senaryo 3: Avustralya Kredi Veri Seti Performans Ölçüm Sonuçları

Aşağı ve Yukarı Örnekleme Yöntemleri Kombinasyonu ile Örnekleyiciler Topluluğu	Tekil Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
SMOTE-ENN Combine	KNN	0,7753	0,7384	0,7741	0,7559	0,5485
	NB	0,8405	0,8000	0,8524	0,8253	0,6801
	Log	0,8695	0,8461	0,8730	0,8593	0,7381
	SVM	0,8405	0,7692	0,8771	0,8196	0,6826
	MLP	0,8695	0,8153	0,8983	0,8548	0,7397
	DT	0,8695	0,8307	0,8852	0,8571	0,7386
SMOTE-Tomek Links	KNN	0,8550	0,7846	0,8947	0,8360	0,7120
	NB	0,7608	0,6153	0,8333	0,7079	0,5300
	Log	0,8768	0,8615	0,8750	0,8682	0,7526
	SVM	0,8695	0,8000	0,9122	0,8524	0,7415
	MLP	0,8623	0,8000	0,8965	0,8455	0,7258
	DT	0,8333	0,8000	0,8387	0,8188	0,6653
Balanced Bagging Classifier (BBC)	KNN	0,8550	0,7692	0,9090	0,8333	0,7144
	NB	0,8405	0,8000	0,8524	0,8253	0,6801
	Log	0,8840	0,8769	0,8769	0,8769	0,7673
	SVM	0,8695	0,8615	0,8615	0,8615	0,7382
	MLP	0,8913	0,8461	0,9166	0,8800	0,7830
	DT	0,8695	0,8307	0,8852	0,8571	0,7386

Tablo 13'te, aşağı ve yukarı örnekleme yöntemleri kombinasyonu ile örnekleyiciler topluluğu ve tekil modellere ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Avustralya kredi veri setinin dengelenmesi amacıyla aşağı ve yukarı örnekleme yöntemlerinin kombinasyonu ile örnekleyiciler topluluğu üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise tekil modellere ait tahmin sonuçlarını rapor etmiştir.

Tablo 13'te gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %89,13 ile Balanced Bagging Classifier-MLP tekil modeller kombinasyonunda ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %91,66 ile Balanced Bagging Classifier-MLP tekil modeller kombinasyonunda ölçülürken,

hassasiyet performans ölçüm değerine ait en yüksek sonuç %87.69 ile Balanced Bagging Classifier-Log kombinasyonunda ölçülmüştür. Öte yandan, F1-skoru açısından en iyi ölçüm değeri %88 ile Balanced Bagging Classifier-MLP kombinasyonunda ölçülürken, MCC açısından en iyi performans %78,30 ile Balanced Bagging Classifier-MLP kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 3'te Balanced Bagging Classifier-MLP müşterilerin kredi riskini tahmin etme hususunda etkili performans göstermiştir.

Tablo 14: Senaryo 4: Avustralya Kredi Veri Seti Performans Ölçüm Sonuçları

Aşağı ve Yukarı Örneklemeye Yöntemleri Kombinasyonu ile Örnekleyiciler Topluluğu	Topluluk Öğrenmesi Modelleri	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
SMOTE-ENN Combine	RF	0,8768	0,8461	0,8870	0,8661	0,7528
	GBDT	0,8695	0,8307	0,8852	0,8571	0,7386
	XTrees	0,8840	0,8461	0,9016	0,8730	0,7678
	Hard	0,8768	0,8461	0,8870	0,8661	0,7528
	Soft	0,8913	0,8461	0,9166	0,8799	0,7830
SMOTE-Tomek Links	RF	0,8913	0,8615	0,9032	0,8818	0,7820
	GBDT	0,8768	0,8615	0,8750	0,8682	0,7526
	XTrees	0,8695	0,8307	0,8852	0,8571	0,7386
	Hard	0,8768	0,8307	0,9000	0,8640	0,7537
	Soft	0,8768	0,8307	0,9000	0,8600	0,7500
Balanced Bagging Classifier (BBC)	RF	0,8550	0,7846	0,8947	0,8360	0,7120
	GBDT	0,8695	0,8615	0,8615	0,8615	0,7382
	XTrees	0,8768	0,8307	0,9000	0,8640	0,7537
	Hard	0,8768	0,8153	0,9137	0,8617	0,7552
	Soft	0,8768	0,8153	0,9137	0,8617	0,7552

Tablo 14'te, aşağı ve yukarı örneklemeye yöntemleri kombinasyonu ile örnekleyiciler topluluğu ve topluluk öğrenmesi modellerine ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, Avustralya kredi veri setinin dengelenmesi amacıyla aşağı ve yukarı örneklemeye yöntemlerinin kombinasyonu ile örnekleyiciler topluluğu üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise topluluk öğrenmesi modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 14’te gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %89,13 hem SMOTE-ENN Combine-Soft hem de SMOTE-Tomek Links-RF topluluk öğrenmesi modelleri kombinasyonunda ölçülmüştür.. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %91,66 ile SMOTE-ENN Combine-Soft kombinasyonunda ölçülürken, hassasiyet performans ölçüm değerine ait en yüksek sonuç %86,15 ile SMOTE-Tomek Links-RF, SMOTE-Tomek Links-GBDT ve Balanced Bagging Classifier-GBDT kombinasyonlarında ölçülmüştür. Öte yandan, MCC açısından en iyi performans %78,30 ile SMOTE-ENN Combine-Soft kombinasyonunda ölçülürken, F1-skoru açısından da en iyi ölçüm değeri %87,99 ile SMOTE-ENN Combine-Soft kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 4’te SMOTE-ENN Combine-Soft müşterilerin kredi riskini tahmin etme hususunda etkili performans göstermiştir.

3.4.3. HMEQ Mortgage Veri Setine ait Performans Ölçüm Sonuçları

Tablo 15: Dengelenmemiş HMEQ Kredi Veri Setine ait Performans Ölçüm Sonuçları

Ham Veri Kümesi	Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Tekil Modeller	KNN	0,9471	0,9989	0,9389	0,9680	0,8301
	NB	0,7852	0,8794	0,8561	0,8676	0,3000
	Log	0,8229	0,9612	0,8405	0,8968	0,3303
	SVM	0,9270	0,9769	0,9348	0,9554	0,7606
	MLP	0,9144	0,9664	0,9294	0,9475	0,7192
	DT	0,8775	0,9402	0,9097	0,9247	0,5987
Topluluk Öğrenmesi Modelleri	RF	0,9169	0,9758	0,9245	0,9495	0,7248
	GBDT	0,9228	0,9727	0,9336	0,9527	0,7469
	XTrees	0,9370	0,9937	0,9321	0,9619	0,7949
	Hard	0,9236	0,9895	0,9209	0,9540	0,7476
	Soft	0,9244	0,9916	0,9202	0,9545	0,7509

Tablo 15’te, Dengelenmemiş HMEQ kredi veri kümesi tekil ve topluluk öğrenmesi modellerine ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, HMEQ kredi veri setinin dengelenmeden önceki sonuçları hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise tekil ve topluluk öğrenmesi modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 15’te gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %94,71 ile K-en yakın komşu algoritmasında ölçülmüştür. Fakat, sadece doğruluk kriteri üzerinden değerlendirme yapılması, kullanılmış olan makine öğrenmesi yöntemlerinin performansları açısından yanlış bir çıkarımda bulunma sebebiyet verebilir. Bu nedenle, tüm değerlendirme kriterleri göz önüne alınarak doğru bir değerlendirme yapılmalıdır. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, hassasiyet, kesinlik, F1-skoru ve MCC açısından da en iyi performans K-en yakın komşu algoritmasında ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, müşterilerin kredi riskini tahmin etme hususunda en etkili performansı K-en yakın komşu algoritması göstermiştir.

Tablo 16: Senaryo 1: HMEQ Veri Setine ait Performans Ölçüm Sonuçları

RUS ve ROS Tekniği	Tekil Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Random Under Sampler (RUS)	KNN	0,9169	0,9308	0,9641	0,9472	0,7555
	NB	0,7734	0,8364	0,8750	0,8553	0,3370
	Log	0,7500	0,7672	0,9059	0,8308	0,3832
	SVM	0,8548	0,8637	0,9504	0,9049	0,6131
	MLP	0,8305	0,8616	0,9215	0,8905	0,5227
	DT	0,8020	0,8123	0,9314	0,8678	0,4987
Random Over Sampler (ROS)	KNN	0,9471	0,9989	0,9389	0,9680	0,8301
	NB	0,7667	0,8343	0,8689	0,8513	0,3129
	Log	0,7390	0,7589	0,8993	0,8231	0,3573
	SVM	0,9186	0,9696	0,9315	0,9501	0,7330
	MLP	0,8825	0,9224	0,9302	0,9263	0,6372
	DT	0,8901	0,9423	0,9220	0,9320	0,6454

Tablo 16’da, rassal aşağı örnekleme ve rassal yukarı örnekleme tekniği ile tekil modellere ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, HMEQ Ev Kredisi veri setinin dengelenmesi amacıyla hem rassal aşağı örnekleme hem de rassal yukarı örnekleme teknikleri üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise tekil modellere ait tahmin sonuçlarını rapor etmiştir.

Tablo 16’da gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %94,71 ile ROS-KNN tekil modeller kombinasyonunda ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %96,41 ile RUS-KNN kombinasyonunda ölçülürken, hassasiyet, F1-skoru ve MCC açısından ise en iyi performans ROS-KNN kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 1’de, müşterilerin kredi riskini tahmin etme hususunda en etkili performansı ROS-KNN göstermiştir. Başka bir ifadeyle, HMEQ ev kredisi veri kümesi ile müşterilerin kredi riskinin tahmin edilmesi hususunda mevcut veri setinde örneklere ek olarak, yeni örnekler eklenerek örnek sayısı artırılmalı ve tekil modellerden K-en yakın komşu ile tahmin yapılabileceği sonucuna ulaşılır.

Tablo 17: Senaryo 2: HMEQ Veri Setine ait Performans Ölçüm Sonuçları

RUS ve ROS Tekniği	Topluluk Öğrenmesi Modelleri	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
Random Under Sampler (RUS)	RF	0,8741	0,8815	0,9578	0,9181	0,6589
	GBDT	0,8758	0,8857	0,9558	0,9194	0,6592
	XTrees	0,9018	0,9035	0,9718	0,9364	0,7315
	Hard	0,8909	0,9056	0,9557	0,9300	0,6888
	Soft	0,8968	0,9129	0,9560	0,9340	0,7015
Random Over Sampler (ROS)	RF	0,9286	0,9664	0,9456	0,9456	0,7704
	GBDT	0,9261	0,9716	0,9382	0,9382	0,7593
	XTrees	0,9303	0,9916	0,9265	0,9265	0,7714
	Hard	0,9261	0,9790	0,9321	0,9321	0,7571
	Soft	0,9303	0,9853	0,9316	0,9316	0,7711

Tablo 17’de, rassal aşağı örnekleme ve rassal yukarı örnekleme tekniği ile topluluk öğrenmesi modellerine ait performans ölçüm sonuçları yer almaktadır. Bu

performans ölçümleri, HMEQ Ev Kredisi veri setinin dengelenmesi amacıyla hem rassal aşağı örnekleme hem de rassal yukarı örnekleme teknikleri üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise topluluk öğrenme modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 17’de gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %93,03 ile hem ROS-XTrees hem de ROS-Soft kombinasyonunda ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, hassasiyet açısından performans ölçüm değerine ait en yüksek sonuç %99,16 ile ROS-XTrees kombinasyonunda ölçülürken, F1-skoru açısından en iyi performans ROS-RF kombinasyonunda ölçülmüştür. Öte yandan, kesinlik açısından performans ölçüm değerine ait en yüksek sonuç %97,18 ile RUS-XTrees kombinasyonunda ölçülürken, MCC açısından ise en iyi performans %77,14 ile ROS-XTrees kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 2’de müşterilerin kredi riskini tahmin etme hususunda en etkili performansı ROS-RF göstermiştir. Başka bir ifadeyle, HMEQ ev kredisi veri kümesi ile müşterilerin kredi riskinin tahmin edilmesi hususunda mevcut veri setinde örneklere ek olarak, yeni örnekler eklenerek örnek sayısı artırılmalı ve topluluk öğrenmesi modellerinden rassal ormanlar ile tahmin yapılabileceği sonucuna ulaşılır.

Tablo 18: Senaryo 3: HMEQ Veri Setine ait Performans Ölçüm Sonuçları

Aşağı ve Yukarı Örneklemeye Yöntemleri Kombinasyonu ile Örnekleyiciler Topluluğu	Tekil Modeller	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
SMOTE-ENN Combine	KNN	0,9211	0,9643	0,9387	0,9513	0,7445
	NB	0,7692	0,8291	0,8759	0,8519	0,3344
	Log	0,7416	0,7599	0,9017	0,8248	0,3651
	SVM	0,9068	0,9486	0,9358	0,9422	0,7029
	MLP	0,8783	0,9203	0,9271	0,9237	0,6236
	DT	0,8842	0,9224	0,9322	0,9272	0,6435
SMOTE-Tomek Links	KNN	0,9505	0,9979	0,9435	0,9699	0,8409
	NB	0,7701	0,8322	0,8744	0,8528	0,3315
	Log	0,7399	0,7515	0,9075	0,8222	0,3761
	SVM	0,9177	0,9643	0,9349	0,9494	0,7324
	MLP	0,9052	0,9570	0,9269	0,9417	0,6906
	DT	0,8481	0,8962	0,9124	0,9042	0,5377
Balanced Bagging Classifier (BBC)	KNN	0,9471	0,9989	0,9389	0,9680	0,8301
	NB	0,7852	0,8794	0,8561	0,8676	0,3000
	Log	0,8229	0,9612	0,8405	0,8968	0,3303
	SVM	0,9270	0,9769	0,9348	0,9554	0,7606
	MLP	0,9144	0,9664	0,9294	0,9475	0,7192
	DT	0,8775	0,9402	0,9097	0,9247	0,5987

Tablo 18’de, aşağı ve yukarı örneklemeye yöntemlerinin kombinasyonu ile örnekleyiciler topluluğu ve tekil modellere ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, HMEQ ev kredisi veri setinin dengelenmesi amacıyla aşağı ve yukarı örneklemeye yöntemlerinin kombinasyonu ile örnekleyiciler topluluğu üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise tekil modellere ait tahmin sonuçlarını rapor etmiştir.

Tablo 18’de gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %95,05 SMOTE-Tomek Links-KNN tekil model kombinasyonunda ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %94,35 ile SMOTE-Tomek Links-KNN kombinasyonunda ölçülürken, hassasiyet performans ölçüm değerine ait en yüksek sonuç %99,89 ile Balanced Bagging Classifier -KNN kombinasyonlarında ölçülmüştür. Öte yandan, MCC açısından en iyi performans %84,09 ile SMOTE-Tomek Links-KNN kombinasyonunda ölçülürken, F1-skoru açısından ise en iyi ölçüm

değeri %96,99 ile SMOTE-Tomek Links-KNN kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 3'te SMOTE-Tomek Links-KNN müşterilerin kredi riskini tahmin etme hususunda etkili performans göstermiştir.

Tablo 19: Senaryo 4: HMEQ Veri Setine ait Performans Ölçüm Sonuçları

Aşağı ve Yukarı Örneklemeye Yöntemleri Kombinasyonu ile Örnekleyiciler Topluluğu	Topluluk Öğrenmesi Modelleri	Doğruluk Accuracy	Hassasiyet Recall	Kesinlik Precision	F1-Skor	MCC
SMOTE-ENN Combine	RF	0,9186	0,9507	0,9477	0,9492	0,7441
	GBDT	0,9270	0,9612	0,9482	0,9547	0,7673
	XTrees	0,9454	0,9748	0,9577	0,9662	0,8255
	Hard	0,9278	0,9612	0,9492	0,9552	0,7702
	Soft	0,9345	0,9696	0,9496	0,9595	0,7897
SMOTE-Tomek Links	RF	0,9295	0,9643	0,9484	0,956	0,7745
	GBDT	0,9437	0,9700	0,9530	0,9653	0,8188
	XTrees	0,9479	0,9779	0,9579	0,9678	0,8331
	Hard	0,9437	0,9779	0,9530	0,9653	0,8188
	Soft	0,9412	0,9853	0,9437	0,9641	0,8088
Balanced Bagging Classifier (BBC)	RF	0,9169	0,9758	0,9245	0,9495	0,7248
	GBDT	0,9228	0,9727	0,9336	0,9527	0,7469
	XTrees	0,9370	0,9937	0,9321	0,9619	0,7949
	Hard	0,9236	0,9895	0,9209	0,9540	0,7476
	Soft	0,9244	0,9916	0,9202	0,9545	0,7509

Tablo 19'da, aşağı ve yukarı örneklemeye yöntemlerinin kombinasyonu ile örnekleyiciler topluluğu ve topluluk öğrenmesi modellerine ait performans ölçüm sonuçları yer almaktadır. Bu performans ölçümleri, HMEQ ev kredisi veri setinin dengelenmesi amacıyla aşağı ve yukarı örneklemeye yöntemlerinin kombinasyonu ile örnekleyiciler topluluğu teknikleri üzerinden hesaplanmış ve sınıflandırma başarılarının test edilmesi için ise topluluk öğrenmesi modellerine ait tahmin sonuçlarını rapor etmiştir.

Tablo 19'da gösterilen performans ölçüm sonuçlarına göre; doğruluk performans ölçüm istatistik değeri incelendiğinde, en yüksek sonucu %94,79 ile SMOTE-Tomek Links-XTrees topluluk öğrenmesi modelleri kombinasyonunda

ölçülmüştür. Doğruluk performans ölçüm istatistiği dışında kalan diğer ölçüm istatistikleri incelendiğinde, kesinlik performans ölçüm değerine ait en yüksek sonuç %95,79 ile SMOTE-Tomek Links-XTrees kombinasyonunda ölçülürken, hassasiyet performans ölçüm değerine ait en yüksek sonuç %99,37 ile Balanced Bagging Classifier-XTrees, kombinasyonunda ölçülmüştür. Öte yandan, MCC açısından en iyi performans %83,31 ile SMOTE-Tomek Links-XTrees kombinasyonunda ölçülürken, F1-skoru açısından da en iyi ölçüm değeri %96,78 ile SMOTE-Tomek Links-XTrees kombinasyonunda ölçülmüştür. Özetle, tabloda gösterilen performans sonuçlarına ait çıktılar genel olarak değerlendirildiğinde, senaryo 4'te SMOTE-Tomek Links-XTrees müşterilerin kredi riskini tahmin etme hususunda etkili performans göstermiştir.

3.4.4. Senaryolar Arasından Optimal Seçim İşlemi

Bu kısımda gerçekleştirilmiş olan uygulama senaryoları için belirlenen örnekleme yöntemleri ve model kombinasyonları arasından bir seçim işlemi gerçekleştirilmiştir. Bu seçim işlemi, hem doğruluk hem de etkililik açısından en iyi performansı gerçekleştiren yöntem ve senaryonun seçilmesine dayanmaktadır. Bu en iyi yöntemin belirlenmesi aşamasında, uygulanan senaryolardaki yöntemlerden elde edilen performans değerleri arasında karşılaştırma yapılarak, maksimum olan değere ait yöntem, en iyi yöntem olarak belirlenir. Ancak optimal yöntem seçimi, ilgilenilen probleme göre performans ölçütünde değişkenlik göstermektedir. Örneğin, dengeli bir veri kümesinde doğruluk performans ölçüm istatistiğine bakılması yeterlidir. Ancak kredi risk analizi gibi dengeli olmayan veri kümelerinin performansı ile ilgili bir çıkarımda bulunmak için salt doğruluk ölçütü yeterli olmamaktadır. Böyle dengesiz bir yapıdaki veri kümeleri için yaygın bir kullanıma sahip olan F1-skoru ve MCC performans ölçüm istatistiği açısından değerlendirme yapılması daha doğru olacaktır. Dengesiz veri kümeleri ile ilgili yapılan çalışmalarda, yanlış pozitifler ve yanlış negatifler genel olarak eşit düzeyde öneme sahiptirler. Bu bağlamda, yürütülen kredi risk analizi çalışmasında, risk teşkil etmeyen müşterilerin riskli olarak etiketlenmesinden kaynaklanan kaybedilen maliyet ile risk teşkil eden müşterilerin risksiz olarak etiketlenmesinden kaynaklanan kaybedilen maliyetin bankalar nezdinde oluşturduğu etki aynı öneme sahiptir. Bu nedenle, hassasiyet ölçütü yanlış negatifler ile hesaplanırken, kesinlik ölçütü ise yanlış pozitifler ile hesaplanmaktadır. Ayrıca

hassasiyet ve kesinlik ölçütünün harmonik ortalaması olarak da bilinen F1-skoru, dengeli olmayan veri kümelerinin değerlendirilmesi açısından önem arz etmektedir. Uygulamanın değerlendirilmesi aşamasında, F1-skoruna ek olarak, MCC kriteri de göz önünde bulundurularak bir değerlendirme yapılacaktır. Bu tez kapsamında, üç farklı veri kümesi üzerinde gerçekleştirilen dört farklı senaryodan elde edilen en iyi sonuçlara ait yeniden örnekleme yöntemi ve model kombinasyonuna ilişkin performans ölçüm değerleri hem F1-skoru hem de MCC kriterlerine göre Tablo 20 ve Tablo 21’de gösterilmiştir.. Daha sonra, F1-skoru ve MCC açısından seçilen optimal değerler arasından nihai örnekleme yöntemi ve model birleşimi seçilerek tablolarda koyu renk ile temsil edilmiştir.

Tablo 20: F1-Skoru Açısından En İyi Senaryolar

UYGULAMA SENARYOLARI				
Veri Kümesi	Senaryo 1	Senaryo 2	Senaryo 3	Senaryo 4
Alman Kredi Veri Kümesine ait Örnekleme Yöntemi ve Seçilen Model Kombinasyonu	0,5762	0,5810	0,5801	0,5897
	RUS-DT	RUS-XTrees	Balanced Bagging Classifier-MLP	SMOTE-ENN Combine-GBDT
Avustralya Kredi Veri Kümesine ait Örnekleme Yöntemi ve Seçilen Model Kombinasyonu	0,8872	0,8730	0,88	0,8799
	RUS-DT	RUS-XTrees	Balanced Bagging Classifier-MLP	SMOTE-ENN Combine-Soft
HMEQ Ev Kredisi Veri Kümesine ait Örnekleme Yöntemi ve Seçilen Model Kombinasyonu	0,9680	0,9456	0,9699	0,9678
	ROS-KNN	ROS-RF	SMOTE-Tomek Links-KNN	SMOTE-Tomek Links-XTrees

Tablo 21: MCC Açısından En İyi Senaryolar

UYGULAMA SENARYOLARI				
Veri Kümesi	Senaryo 1	Senaryo 2	Senaryo 3	Senaryo 4
Alman Kredi Veri Kümesine ait Örnekleme Yöntemi ve Seçilen Model Kombinasyonu	0,3701	0,4045	0,3867	0,4397
	RUS-DT	ROS-RF / ROS- Hard	Balanced Bagging Classifier – SVM	Balanced Bagging Classifier - RF
Avustralya Kredi Veri Kümesine ait Örnekleme Yöntemi ve Seçilen Model Kombinasyonu	0,7831	0,7678	0,7830	0,7830
	RUS-DT	RUS-XTrees	Balanced Bagging Classifier - MLP	SMOTE-ENN Combine-Soft
HMEQ Ev Kredisi Veri Kümesine ait Örnekleme Yöntemi ve Seçilen Model Kombinasyonu	0,8301	0,7714	0,8409	0,8331
	ROS-KNN	ROS-XTrees	SMOTE-Tomek Links-KNN	SMOTE-Tomek Links-XTrees

Uygulama senaryolarına ait en optimal sonuçlara ait çıktı değerleri Tablo 20 ve Tablo 21’de gösterilmiştir. Alman kredi veri kümesine uygulanan dört senaryo arasından F1-skoruna ait en iyi senaryo değerleri incelendiğinde, sınıflandırma performansı açısından en optimal sonuç %58,97 ile senaryo 4’te yer alan SMOTE-ENN Combine-Gradyan Artırma Karar Ağaçları kombinasyonunda ölçülmüştür. Ek olarak, Alman kredi veri kümesine uygulanan dört senaryo arasından MCC ait en iyi senaryo değerleri incelendiğinde, sınıflandırma performansı açısından %43,97 ile senaryo 4’te yer alan Balanced Bagging Classifier-Rassal Ormanlar kombinasyonunda ölçülmüştür. Başka bir deyişle, model başarısının orta düzeyde olduğu söylenebilir.

Avustralya kredi veri kümesine uygulanan dört senaryo arasından F1-skoruna ait en iyi senaryo değerleri incelendiğinde, sınıflandırma performansı açısından en optimal sonuç %88,72 ile senaryo 1’de yer alan rassal aşağı örnekleme-karar ağaçları kombinasyonunda ölçülmüştür. Ek olarak, Avustralya kredi veri kümesine uygulanan dört senaryo arasından MCC ait en iyi senaryo değerleri incelendiğinde, Balanced Bagging Classifier-MLP ve SMOTE-ENN Combine-Soft’un aynı sonucu verdiği görülmekte ve az bir farkla sınıflandırma performansı açısından %78,31 ile senaryo

1’de yer alan rassal ařađı rneklemekarar ađađları kombinasyonu seilmiřtir. Bařka bir deyiřle, model bařarisının yksek dzeyde olduđunu gstermektedir.

HMEQ Ev kredisi veri kmesine uygulanan drt senaryo arasından F1-skoruna ait en iyi senaryo deđerleri incelendiđinde, sınıflandırma performansı aısından en optimal sonu %96,99 ile senaryo 3’te yer alan SMOTE-Tomek Links-K-en yakın komřu kombinasyonunda llmřtr. Ek olarak, HMEQ Ev kredisi veri kmesine uygulanan drt senaryo arasından MCC kriterine gre en iyi senaryo deđerleri incelendiđinde, sınıflandırma performansı aısından %84,09 ile senaryo 3’ te yer alan SMOTE-Tomek Links-K-en yakın komřu kombinasyonunda tespit edilmiřtir. Bařka bir deyiřle, model bařarisının yksek dzeyde olduđu sylenebilir.



SONUÇ ve ÖNERİLER

Bu tez kapsamında, kredi talebinde bulunan müşterilerin, bankalar tarafından kredi taleplerinin gerçekleştirilebilmesi için müşterilerden istenen çeşitli bilgilere dayanarak, müşterilerin riskli olup olmama durumlarının tespiti için makine öğrenmesi tekniklerinden faydalanılmıştır. Ayrıca sınıf dengesizliği problemi eli alınarak, son yıllarda önerilmiş yeni dengeleme yöntemlerinin incelenen makine öğrenme teknikleri ile etkileşimi incelenmiştir. Bu kapsamda sınıflandırma modelleri için en iyi performans sergileyen makine öğrenme ve dengeleme tekniği kombinasyonunun belirlenmesi amaçlanmıştır. Bu yöntem ile bankalar, riskli müşterileri belirleyerek kredi riskini azaltabilecek kârın artmasını sağlayabilecektir. Bu doğrultuda, bankalara entegre edilen karar verme yöntemi ile makine öğrenmesi algoritmaları ilgili bankaların geçmiş dönemlerde kredi talebinde müşteriler için belirlenen kriterlere göre aldığı kararlardan oluşan verilerle eğitilmektedir. Bu eğitilen veriler ve kullanılan yöntemler yardımıyla, kredi talep eden yeni müşteriler otomatik olarak değerlendirilerek kredi alabilir veya kredi alamaz şeklinde nihai karara ulaşılmaktadır. Bankalarca, bu yöntemin kullanılması hem iş yükünü hem de maliyetin azaltılması gibi çeşitli avantajları beraber getirecektir. Ancak tüm bu avantajların yanı sıra, uygulama aşamasında kullanılan gerçek hayat problemlerinin doğrudan makine öğrenme algoritmalarına entegre edilmesiyle iyi sonuçlar elde edilmesi söz konusu olmayabilir. Bu verilerin dengesiz bir yapıda olması algoritmaların performansı etkilemektedir. Bu nedenle, veriler uygun hâle dönüştürülerek ilgili makine öğrenmesi algoritmaları ile kullanılması gerekmektedir.

Çalışmanın uygulama aşamasında, dengesiz bir yapıda olan gerçek hayat problemlerine ait üç veri seti; Alman kredi veri kümesi, Avustralya kredi veri kümesi ve HMEQ ev kredisi veri kümesi içerisindeki sayısal nitelik taşımayan değişkenler öncelikle sayısal forma dönüştürülmüştür. Bu dönüştürme işlemi, değişkenler arasında ilişki bulunmayan ve nitel özellikte bulunan kategorik değişkenler için Python diline ait Scikit-learn kütüphanesi içinden “*OneHotEncoder()*”, ikili kategorik veri özelliğinde bulunan değişkenler için “*LabelEncoder()*” ve sıralanabilir özellikteki değişkenler için ise “*OrdinalEncoder()*” fonksiyonları kullanılmıştır. Bu nedenle, ilgili veriler sayısallaştırılmış ve mevcut veri sayısında yer alan değişkenlerde artış

olmuştur. Sonuç olarak, makine öğrenmesi algoritmaları için değişken uygun bir yapıya dönüştürülmüştür.

Riskli müşterilerin tespit edilebilmesi için temel olarak iki aşamadan oluşmaktadır:

İlk aşama, veri setinin ön işleme süreci aşamadır. Bu aşamada, dengesiz bir yapıda bulunan veri kümesi dengeli bir yapıya getirilmektedir. Uygulama aşamasında kullanılan gerçek hayat problemlerinde en sık karşılaşılan problemlerin başında sınıf dengesizliği problemi gelmektedir. Bu problem, uygulanan makine öğrenmesi algoritmalarının performansını olumsuz etkilemekte ve tahmin gücünü azaltmaktadır. Bu tez kapsamında kullanılan, bankacılık ve finans sektörüne ait gerçek hayat problemlerindeki sınıf dengesizliği problemini çözmek ve risk teşkil eden müşterilerin doğru bir şekilde tespit edilmesi için aşağıdaki yöntemlerle azınlık ve çoğunluk sınıfı dengelenmiştir. Bu doğrultuda, rassal aşağı örnekleme, rassal yukarı örnekleme teknikleri, aşağı ve yukarı örnekleme yöntemlerinin kombinasyonu; SMOOTE-ENN ve SMOTE-Tomek Links ve örnekleyiciler topluluğu; Dengeli Özyinelemeli Birleştirme Sınıflandırıcısı teknikleri ile sınıflar dengelenmiştir.

İkinci aşama ise üç farklı veri kümesinin her biri eğitim ve test verisi seti olarak ayrıldıktan sonra makine öğrenmesi algoritmalarına ait performanslar tekil ve topluluk öğrenmesi modelleri açısından ele alınmıştır. Bu doğrultuda, KNN, NB, LR, SVM, MLP ve DT olmak üzere 6 farklı tekil model kullanılırken, RF, GBDT, XTrees, Hard ve Soft olmak üzere 5 farklı topluluk öğrenmesi modelinde faydalanılmıştır.

Çalışmanın uygulama aşamasında, ilgilenilen kredi riski problemine ilişkin veriler için en iyi performansı sergileyen sınıf dengeleme yönteminin belirlenmesi amacıyla dört farklı senaryo hazırlanmıştır. Bu senaryolar, üç farklı veri kümesine uygulanmıştır.

Çalışmanın uygulamasında kullanılan üç farklı veri kümesi için dengelenmemiş sonuçlarına ait performans ölçüm istatistiklerine yer verilmiştir. Dengelenmemiş veri kümeleri de bazı dengelenmiş veri kümelerine göre etkin sonuçlar göstermiştir. Bu bulgu, veri kümesinin hacmi ve özelliklerine göre farklılık gösterebilir. Dolayısıyla, buradan elde edilen sonuç sınıf dengeleme tekniklerinin her zaman ve her veri setinde daha iyi performanslara yol açmayacağını göstermiştir.

İlk senaryoda, dengesiz bir yapıya sahip olan Alman, Avustralya ve HMEQ kredi veri setinin dengelenmesi amacıyla hem rassal aşağı örnekleme hem de rassal yukarı örnekleme teknikleri kullanılmış ve sınıflandırma başarılarının test edilmesi için ise tekil modellere ait performans tahmin sonuçları kullanılarak değerlendirme yapılmıştır. Alman kredi veri kümesi için uygulanan senaryolar incelendiğinde, örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını rassal aşağı örnekleme-karar ağaçları göstermiştir. Avustralya kredi veri kümesi için uygulanan senaryolar incelendiğinde, örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansına rassal aşağı örnekleme-karar ağaçları ile ulaşılmıştır. HMEQ ev kredisi veri kümesi için uygulanan senaryolar incelendiğinde ise örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını rassal yukarı örnekleme-K-en yakın komşu göstermiştir.

İkinci senaryoda, Alman, Avustralya ve HMEQ kredi veri setlerinin dengelenmesi amacıyla hem rassal aşağı örnekleme hem de rassal yukarı örnekleme teknikleri kullanılmış ve sınıflandırma başarılarının test edilmesi için ise topluluk öğrenmesi modellerine ait performans tahmin sonuçları raporlanmıştır. Alman kredi veri kümesi için uygulanan senaryolar incelendiğinde, örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını rassal aşağı örnekleme-extra ağaçlar göstermiştir. Avustralya kredi veri kümesi için uygulanan senaryolar incelendiğinde, örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını extra ağaçlar sergilemiştir. HMEQ ev kredisi için uygulanan senaryolar incelendiğinde ise örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını rassal yukarı örnekleme-rassal ormanlar göstermiştir.

Üçüncü senaryoda, Alman, Avustralya ve HMEQ ev kredisi kredi veri setinin dengelenmesi amacıyla aşağı ve yukarı örnekleme yöntemlerinin kombinasyonu ile örnekleyciler topluluğu kullanılmış ve sınıflandırma başarılarının test edilmesi için ise tekil modellere ait performans tahmin sonuçları kullanılarak değerlendirme yapılmıştır. Alman kredi veri kümesi için uygulanan senaryolar incelendiğinde, örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını Balanced Bagging Classifier-Çok Katmanlı Algılayıcı göstermiştir. Avustralya kredi veri kümesi için uygulanan senaryolar incelendiğinde, örnekleme yöntemi ve model

birleşimi açısından en etkili sınıflandırma performansını yine Balanced Bagging Classifier-Çok Katmanlı Algılayıcı sergilemiştir. HMEQ Ev kredisi veri kümesi için uygulanan senaryolar incelendiğinde ise örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansı SMOTE-Tomek Links-KNN tarafından sağlanmıştır.

Dördüncü senaryoda, dengesiz bir yapıya sahip olan Alman, Avustralya ve HMEQ ev kredisi kredi veri setinin dengelenmesi amacıyla aşağı ve yukarı örnekleme yöntemlerinin kombinasyonu ile örnekleyiciler topluluğu kullanılmış ve sınıflandırma başarılarının test edilmesi için ise topluluk öğrenmesi modellerine ait performans tahmin sonuçları kullanılarak değerlendirme yapılmıştır. Alman kredi veri kümesi için uygulanan senaryolar incelendiğinde, örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını SMOTE-ENN Combine-GBDT göstermiştir. Avustralya kredi veri kümesi için uygulanan senaryolar incelendiğinde, örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını SMOTE-ENN Combine-Soft tarafından kaydedilmiştir. HMEQ ev kredisi veri kümesi için uygulanan senaryolar incelendiğinde ise örnekleme yöntemi ve model birleşimi açısından en etkili sınıflandırma performansını SMOTE-Tomek Links-Extra Ağaçlar ile elde edilmiştir.

Çalışmanın uygulama aşamasında kullanılan üç farklı veri kümesinin göstermiş olduğu performans sonuçlarına göre, HMEQ ev kredisi hem Alman hem de Avustralya kredi verisine göre kayda değer bir performans sergilemiştir. Buna ilaveten, diğer veri kümelerinin göstermiş oldukları performans başarıları sırasıyla, Avustralya ve Alman veri kümesi şeklindedir. Ayrıca, başarılarına göre, örnekleme yöntemi ve model birleşimleri sırasıyla; HMEQ ev kredisi için uygulanan senaryolar arasından en iyi performansı senaryo 3'te yer alan SMOTE-Tomek Links-KNN, Avustralya kredi veri kümesine uygulanan senaryolar arasından en iyi performansı senaryo 1'de yer alan RUS-DT ve Alman kredi veri kümesine uygulanan senaryolar arasından en iyi performansı ise senaryo 4'te yer alan SMOTE-ENN Combine-GBDT birleşimi olduğu tespit edilmiştir.

Çalışma kapsamında uygulanan tüm senaryolar ve makine öğrenmesi yöntemlerine ait modeller açısından genel bir değerlendirme yapılacak olursa, topluluk öğrenmesi modelleri, tekil modellere göre daha iyi sınıflandırma performansı

göstermiştir. Bu bağlamda, aşağı ve yukarı örnekleme yöntemlerinin bir arada kullanıldığı hibrit yöntem ve topluluk öğrenmesi modelleri yardımıyla riskli müşterilerin tespit edilmesi bankacılık ve finans sektöründeki kredi veren diğer kuruluşlar için iş yükü, zaman, maliyet ve kaynak kullanım masraflarını azaltarak olumlu etki sağlarken, aynı zamanda kredi riskini de azaltarak kârın artmasını sağlayacaktır.

Gelecek çalışmalar kapsamında, uygulama aşamasında kullanılan makine öğrenmesi algoritmaları ve sınıflandırma performans ölçütlerine ek olarak, farklı yöntem ve AUC-ROC eğrisi gibi performans ölçütleri değerlendirme için kullanılabilir. Aynı zamanda, bu çalışmada kullanılan yöntemlerin kapsamı genişletilebilir ve mevcut bulgularla kıyaslanabilir.

Gelecek çalışmalar kapsamında, mevcut çalışmadaki yöntemler için değişken seçim yöntemleri uygulanabilir. Bu yöntemler; Lasso, Ridge ve Elastic Net gibi değişken seçimi yöntemleri olabilir. Böylece, değişken seçimi yöntemlerini ilgili veri setleri ve kullanılan senaryolar üzerinde uygulayarak, değişken seçiminin performans üzerindeki etkisi araştırılabilir. Ayrıca, mevcut çalışmada sınıf dengesizliğinin giderilmesi için kullanılan yöntemlere ek olarak, Bagging Classifier, Balanced Random Forest Classifier, RUS Boost, Easy Ensemble Classifier gibi örnekleyiciler topluluğu yöntemleri kullanılarak, söz konusu veri setlerine mevcut senaryolar üzerinden yeniden bir performans değerlendirmesi gerçekleştirilebilir.

KAYNAKÇA

Abdioğlu, H. (2019). İşletmelerde Öz Sermaye Karşılaştırılması Yoluyla Ticari Karın Hesaplanması.

Akman, M., Genç, Y. ve Ankaralı, H. (2011). Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama/Random Forests Methods and an Application in Health Science. *Türkiye Klinikleri Biyoistatistik*. 3(1): 36.

Aksoy, E. (2005). *Türk bankacılık sisteminde bireysel krediler ve bankacılık altyapısının paylaşımı (Ortak ATM ve POS kullanımı)*. (Yayımlanmamış yüksek lisans tezi). İstanbul: Beykent Üniversitesi Sosyal Bilimler Enstitüsü.

Al-Abassi, A., Sakhnini, J. ve Karimipour, H. (2020). Unsupervised Stacked Autoencoders for Anomaly Detection on Smart Cyber-physical Grids. *In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (ss. 3123-3129).

Alam, M. S. ve Vuong, S. T. (2013). Random forest classification for detecting android malware. *In 2013 IEEE international conference on green computing and communications and IEEE Internet of Things and IEEE cyber, physical and social computing* (pp. 663-669).

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S. ve Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access*. 8: 201173-201198.

Ali, H., Salleh, M. N. M., Hussain, K., Ahmad, A., Ullah, A., Muhammad, A. ve Khan, M. (2019). A review on data preprocessing methods for class imbalance problem. *International Journal of Engineering & Technology*. 8: 390-397.

Ali, N., Neagu, D. ve Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences*. 1(12): 1-15.

Alpaydın, E. (2004). *Introduction To Machine Learning*. England: The MIT Press.

Amanet, H. (26.05.2020). *Model Değerlendirme Metrikleri-1 : Karmaşıklık Matrisi (Confusion Matrix)*. [https://textdecipher.com/model-degerlendirme metrikleri %E2 %80%8A-%E2%80%8A1-karmasiklik-matrisi-confusion-matrix/](https://textdecipher.com/model-degerlendirme-metrikleri-%E2%80%8A-%E2%80%8A1-karmasiklik-matrisi-confusion-matrix/), (25.02.2022).

Apak, S. ve Atay, E. (2009). Bankacılık Sektöründe Operasyonel Riskin Minimize Edilmesi ve Altı Sigma Uygulamaları. *Trakya Üniversitesi Sosyal Bilimler Dergisi*. 11(1): 165-180.

Aras, G. (1996). *Ticari Bankalarda Kredi Portföyünün Yönetimi*. Ankara: SPK Yayınları.

Auria, L. ve Moro, R. A. (2008). *Support Vector Machines (SVM) as a Technique for Solvency Analysis* (No. 811). DIW Berlin, German Institute for Economic Research.

Ayboğa, H. (1991). *Anonim şirketlerin iştirak ilişkilerinde Yeniden Değerleme Değer artış Fonu ile Gayrimenkul Satış karının Esas Sermayeye Eklenmesinde Muhasebe ilişkileri*. (Doktora Tezi). İstanbul: Marmara Üniversitesi Sosyal Bilimler Enstitüsü.

Babuşcu, Ş. (2010). “Bankalarda Risk Yönetimi (1)”. *Ekonomist Dergisi*. Sayı: 12: 1.

Bakanoğlu, D. (2005). *Bankalarda Kredi ve Kredi Riski Yönetimi*. (Yayınlanmamış Doktora Tezi). İstanbul: Kültür Üniversitesi Sosyal Bilimler Enstitüsü.

BKKK. (01.03.2006). Banka Kartları ve Kredi Kartları Kanunu. [https://www.mevzuat.gov.tr/ MevzuatMetin/1.5.5464.pdf](https://www.mevzuat.gov.tr/MevzuatMetin/1.5.5464.pdf), (10.04.2022).

Barros, T. M., Souza Neto, P. A., Silva, I. ve Guedes, L. A. (2019). Predictive models for imbalanced data: A school dropout perspective. *Education Sciences*. 9(4): 275.

Bataee, M. ve Mohseni, S. (2011). Application of Artificial Intelligent Systems in ROP Optimization: A Case Study in Shadegan Oil Field. *In SPE middle east unconventional gas conference and exhibition*. OnePetro.

Batista, G. E., Bazzan, A. L. ve Monard, M. C. (2003, December). Balancing Training Data for Automated Annotation of Keywords: a Case Study. *In WOB* (ss. 10-18).

Batista, G. E., Prati, R. C. ve Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*. 6(1): 20-29.

BDDK. (08.02.2001). *Bankaların İç Denetim ve Risk Yönetimi Sistemleri Hakkında Yönetmelik*. <https://www.resmigazete.gov.tr/eskiler/2001/02/20010208.htm#7>, (15.04.2022).

BDDK.(23.10.2015). *Bankaların Sermaye Yeterliliğinin Ölçülmesine ve Değerlendirilmesine İlişkin Yönetmelik*. <https://www.resmigazete.gov.tr/eskiler/2015/10/20151023-9.pdf>, (14.04.2022).

Béjar, J. (2013). Strategies and Algorithms for Clustering Large Datasets: A Review. *Computer Science*, 1-20.

Bessis, J. (2002). *Risk Management in Banking* (2nd Ed.). West Sussex, England: John Wiley & Sons Ltd.

Bishop, C. M. ve Nasrabadi, N. M. (2006). Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*. 16(4): 049901-049901.

Boughorbel, S., Jarray, F. ve El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS One*. 12(6): 0177678.

Boyacıoğlu, M. (2005). *Bankalarda Derecelendirme (Rating) ve Türk Bankacılık Sektörü Üzerine Ampirik Bir Çalışma*. (Doktora Tezi). Konya: Selçuk Üniversitesi Sosyal Bilimler Enstitüsü.

Bradley, A. P., Duin, R. P. W., Paclik, P. ve Landgrebe, T. C. W. (2006). Precision-Recall Operating Characteristic (P-ROC) Curves in Imprecise Environments. *In 18th International Conference on Pattern Recognition (ICPR'06)* (pp.123-127). Cambridge , United Kingdom.

Bramer, M. (2007). Avoiding Overfitting of Decision Trees. *Principles of Data Mining*. 8: 119-134.

Breiman, L. (1996). Bagging Predictors. *Machine learning*. 24(2): 123-140.

Breiman, L. (2001). Random forests. *Machine learning*. 45(1): 5-32.

Breiman, L. , Friedman J., Olshen R. ve Stone C. (1984). Classification and Regression Trees. Wadsworth Int. Group. 37(15): 237-251.

Brynjolfsson, E. ve Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*. 358(6370): 1530-1534.

Buczak, A. L. ve Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys and tutorials*. 18(2): 1153-1176.

Burges, C.J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*. 2(2): 121–167.

Carew, J. (10.09.2020). Reinforcement Learning. <https://www.techtarget.com/search-enterprise/ai/definition/reinforcement-learning>, (15.03.2022).

Chamberlain, S., Howe, J. S. ve Popper, H. (1997). The exchange rate exposure of US and Japanese banking institutions. *Journal of banking & finance*. 21(6): 871-892.

Chicco, D. ve Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*. 21(1): 1-13.

Cortes, C. ve Vapnik, V. (1995). Support-Vector Networks. *Mach Learn*. 20: 273–297.

Cristianini, N. ve Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. UK: Cambridge University Press.

Chicco, D., Warrens, M. J. ve Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. *IEEE*. 9: 78368-78381.

Dahiya, S., Handa, S. S. ve Singh, N. P. (2016). Impact of Bagging on MLP Classifier for Credit Evaluation. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE (pp. 3794-3800).

Data Science. (29.04.2019). *What is Confusion Matrix and Advanced Classification Metrics?*. <https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html>, (24.02.2022).

De Ville, B. (2006). *Decision trees for business intelligence and data mining: Using SAS enterprise miner*. Tang Kinh Cac.

Deveci, M. (2019). *Türk Bankacılık Sektöründe Mobil Bankacılığın Bireysel Krediler Açısından İncelenmesi*. (Yayınlanmamış Doktora Tezi). İstanbul: Marmara Üniversitesi Sosyal Bilimler Enstitüsü.

Dmitrievski. (06.07.2018). Random Decision Forest in Reinforcement Learning. <https://www.mql5.com/en/articles/3856>, (13.02.2022).

Duda, R. O. ve Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: Wiley.

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

Dölger, M. V. (2004). *Türk Ceza Hukuku'nda Bilişim Suçları*. (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.

Džeroski, S. ve Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one?. *Machine learning*. 54(3): 255-273.

Edwards, B. (Ed.). (2004). *Credit management handbook*. Gower Publishing, Ltd..

El Naqa, I. ve Murphy, M. J. (2015). What is machine learning?. *In machine learning in radiation oncology* (pp. 3-11). Cham, Switzerland.

Elhassan, T. ve Aljurf, M. (2016). Classification of Imbalance Data Using Tomek Link (t-link) Combined With Random Under-Sampling (RUS) as a Data Reduction Method. *Global J Technol Optim*. 2(11): 1-12.

Ertürk, H. (2010). Bankacılık Sektörünün Karşılaştığı Riskler Ve Risk Yönetimi. *Denetim*. 4: 62-70.

Fahey, C. (2003). *Colin Faheys Guide to Artificial Neural Networks*. http://www.colinfahey.com/2003apr20_neuron/2003apr20_neuron.htm, (02.02.2022).

Fan, R., Zhong, M., Wang, S., Zhang, Y., Andrew, A., Karagas, M. ve Moore, J. H. (2011). Entropy-Based Information Gain Approaches to Detect and to Characterize Gene-Gene and Gene-Environment Interactions/Correlations of Complex Diseases. *Genetic epidemiology*. 35(7): 706-721.

Fedorovici, L.O. ve Dragan, F. (2011). A Comparison Between a Neural Network and a SVM and Zernike Moments Based Blob Recognition Modules. *6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)* (ss.253-258). Timisoara, Romania.

Feng, Y., Wang, D., Yin, Y., Li, Z. ve Hu, Z. (2020). An XGBoost-based casualty prediction method for terrorist attacks. *Complex & Intelligent Systems*, 6(3): 721-740.

Fletcher, T. (2009). *Support Vector Machines Explained*. Tutorial Paper.

Fomby, T. B. ve Barber, T. (2008). *K-Nearest Neighbors Algorithm: Prediction and Classification*. Dallas,USA: Southern Methodist University.

France, S. L., Carroll, J. D. ve Xiong, H. (2012). Distance metrics for high dimensional nearest neighborhood recovery: Compression and normalization. *Information Sciences*. 184(1): 92-110.

Fujimoto, K. (2003). *Application of Multinomial and Ordinal Regressions to the Data of Japanese Female Labor Market*. (Doctoral Dissertation). USA: University of Pittsburgh.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H. ve Herrera, F. (2011). An Overview of Ensemble Methods for Binary Classifiers in Multi-Class Problems: Experimental Study on One-vs-One and One-vs-All schemes. *Pattern Recognition*. 44(8):1761-1776.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. ve Herrera, F. (2011). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 42(4): 463-484.

Gareth, J., Daniela, W., Trevor, H. ve Robert, T. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

Geurts, P., Ernst, D. ve Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*. 63(1): 3-42.

Gehrke, J. (2003). Decision Trees. *The Handbook of Data Mining* (pp. 3-24). Editors Nong Ye. New Jersey: Lawrence Erlbaum Associates Inc.

Golin, J. ve Delhaise, P. (2013). *The Bank Credit Analysis Handbook: A Guide for Analysts, Bankers and Investors* (2nd Ed.). Singapur: John Wiley and Sons Singapore Pte. Ltd.

Grąbczewski, K. (2014). *Meta-learning in Decision Tree Induction*. Cham, Switzerland: Springer International Publishing.

Gunn, S. R. (1998). Support vector machines for classification and regression. *ISIS technical report*. 14(1): 5-16.

Gupta, S., Kumar, D. ve Sharma, A. 2011. Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. *Indian Journal of Computer Science and Engineering*. 2(2): 188-195.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. ve Tatham, R. L. (1998). Multivariate data analysis . Uppersaddle River. *Multivariate Data Analysis (5th ed)* Upper Saddle River. 5(3): 207-219.

Hamid, S. A. ve Iqbal, Z. (2004). Using neural networks for forecasting volatility of S&P 500 Index futures prices. *Journal of Business Research*. 57(10): 1116-1125.

Han, J., Pei, J. ve Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. New York: MacMillan College Publishing Co.

Horning, N. (2010). Random Forests: An Algorithm for Image Classification and Generation of Continuous Fields Data Sets. *International Conference On Bioinformatics for Spatial Infrastructure Development in Earth and Allied Sciences* 2010 (pp.1–6). Osaka, Japan.

Hosmer Jr, D. W., Lemeshow, S. ve Sturdivant, R. X. (2013). *Applied Logistic Regression* (398). John Wiley & Sons.

Hosmer, D. W. ve Lemeshow, S. (1980). Goodness of Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics-Theory and Methods*. 9(10): 1043–1069.

Hou, W. H., Wang, X. K., Zhang, H. Y., Wang, J. Q. ve Li, L. (2020). A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. *Knowledge-Based Systems*. 208: 106462.

Huang, C., Davis, L.S. ve Townshend, J.R.G., 2002, An assessment of support vectormachines for land cover classification. *International Journal of Remote Sensing*. 23: 725–749.

James, G., Witten, D., Hastie, T. ve Tibshirani, R. (2013). *An Introduction to Statistical Learning* (112). New York: Springer.

Javapoint.(2020). *Unsupervised Machine Learning*. <https://www.javatpoint.com/unsupervised-machine-learning>. (18.03.2022).

Jin, Y., Zhang, W., Wu, X., Liu, Y. ve Hu, Z. (2021). A Novel Multi-Stage Ensemble Model With a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data. *IEEE Access*. 9: 143593-143607.

Jordan, M. I. ve Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*. 349(6245): 255-260.

Joshi, R. (09.09.2016). *Accuracy, Precision, Recall and F1 Score: Interpretation of Performance Measures*. <https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/>, (27.02.2022).

Kaban, İ. (2017). Kullanımı Yaygın Olan Teminat Türleri Çerçevesinde Bankalarda Nakdi Krediler İç in Ayrılacak Özel Karşılıkların Muhasebeleştirilmesi. *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*. 39(1): 195-211.

Katircioğlu, H. (2006). *Basel-II Standartlarına Göre Ticari Bankalarda Kredi Taleplerinin Değerlendirilmesi ve Bir Uygulama*. (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: Yıldız Teknik Üniversitesi Sosyal Bilimler Enstitüsü.

Kavcıoğlu, Ş. (2003). *Ticari Bankalarda Sorunlu Kredilerin Yönetimi, Çözüm Yolları ve Takibi*. İstanbul: Türkmen Kitabevi.

Kavzoglu, T. ve Çölkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*. 11(5): 352-359.

Kavzoğlu, T. ve Çölkesen, I. (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. *Harita Dergisi*. 144(7): 73-82.

Khemakhem, S., Ben Said, F. ve Boujelbene, Y. (2018). Credit Risk Assessment for Unbalanced Datasets Based on Data Mining, Artificial Neural Network and Support Vector Machines. *Journal of Modelling in Management*. 13(4): 932-951.

Kılıç, M. (2019). Bankaların Finansal Yapısının Karlılık Üzerindeki Etkisi: Özel Sermayeli Mevduat Bankaları Üzerine Bir İnceleme. *Uluslararası İktisadi ve İdari İncelemeler Dergisi*. (24): 45-58.

Kıran, S. (2013). *Finansal Oran Analizi ile Kredi risk Yönetimi: Bir Uygulama*. (Yayınlanmamış Yüksek Lisans Tezi). İstanbul: İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M. ve Klein, M. (2002). *Logistic Regression*. New York: Springer International Publishing.

Kohonen, T. (1989). *Self-Organizing and Associative Memory* (3rd ed.). Berlin: Springer-Verlag.

Korniichuk, R. ve Boryczka, M. (2021). Averaging and Boosting Methods in Ensemble-Based Classifiers for Text Readability. *Procedia Computer Science*. 192: 3677-3685.

Kotsiantis, S., Kanellopoulos, D. ve Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*. 30(1): 25-36.

Kumar, A. (20.01.2022). *Accuracy, Precision, Recall and F1-Score-Python Examples*. https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/#What_is_Recall_Score, (26.02.2022).

Kumar, A. (20.05.2020). *Introduction to the Gradient Boosting Algorithm*. <https://medium.com/analytics-vidhya/introduction-to-the-gradient-boosting-algorithm-c25c653f826b>, (22.02.2022).

Kumari, M. ve Godara, S. (2011). Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. *I. International Journal of Computer Science and Technology*. 2: 304-308.

Kunal (11.06.2015). *Machine Learning Basics for a Newbie*. <https://www.analyticsvidhya.com/blog/2015/06/machine-learning-basics/>, (24.03.2022).

Lantz, B. (2013). *Machine learning with R*. Birmingham, UK: Packt Publishing.

Le, C. T. ve Eberly, L. E. (2016). *Introductory Biostatistics* (2nd ed.). New Jersey: John Wiley & Sons.

Lefkovits, S. ve Lefkovits, L. (2017). Gabor feature selection based on information gain. *Procedia Engineering*. 181: 892-898.

Li, S., Li, H., Li, M., Shyr, Y., Xie, L. ve Li, Y. (2009). Improved prediction of lysine acetylation by support vector machines. *Protein and peptide letters*. 16(8): 977-983.

Li, Y., Zhang, W. ve Lin, C. (2006). Simplify support vector machines by iterative learning. *Neural Information Processing: Letters and Reviews*. 10(1): 11-17.

Liu, W., Fan, H. ve Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*. 189: 116034.

Loukas, S. (10.06.2020). *What is Machine Learning: Supervised, Unsupervised, Semi-Supervised and Reinforcement learning methods*. <https://towardsdatascience.com/what-is-machine-learning-a-short-note-on-supervised-unsupervised-semi-supervised-and-aed1573ae9bb>, (17.03.2022).

Maciel, L. S. ve Ballini, R. (2008). Design a Neural Network for Time Series Financial Forecasting: Accuracy and Robustness Analysis. *Anales do 9º Encontro Brasileiro de Finanças*. Sao Pablo, Brazil.

Madhulatha, T. S. (2012). An Overview on Clustering Methods. arXiv preprint arXiv:1205.1117.

Mahabub, A. (2020). A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. *SN Applied Sciences*. 2(4): 1-9.

Mahabub, A. ve Habib, A. Z. S. B. (2019, December). A voting approach of modulation classification for wireless network. In *Proceedings of the 6th international conference on networking, systems and security* (pp. 133-138).

Mahabub, A., Mahmud, M. I. ve Hossain, M. F. (2019). A robust system for message filtering using an ensemble machine learning supervised approach. *ICIC Express Letters, Part B: Applications*. 10(9): 805-812.

Maimon, O. Z. ve Rokach, L. (2014). *Data Mining with Decision Trees: Theory and Applications* (2nd Ed.). USA: World scientific.

Malekipirbazari, M. ve Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*. 42(10): 4621-4631.

Mallick, S. (11.07.2018). *Support Vector Machine*. <https://learnopencv.com/support-vector-machines-svm/>, (03.02.2022).

Marksfeld, A. E. (2018). *Strojové učení v sociodemografické segmentaci zákazníků telekomunikační společnosti*. (Unpublished Bachelor's Thesis,). Czech Republic: Czech Technical University Computer and Information Center.

Martin, S. B. (2001). *Techniques in Support Vector Classification*. (Yayınlanmış Doktora Tezi). USA: Colorado State University.

Mathur, A. ve Foody, G. M. (2008). Multiclass and binary SVM classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*. 5(2): 241-245.

McClelland, J. L., Rumelhart, D. E. ve Hinton, G. E. (1986). The appeal of parallel distributed processing. MIT Press, Cambridge MA, 3-44.

McCulloch, W. S. ve Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*. 5(4): 115-133.

Mertler, C. A. ve ÖC Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Glendale, CA: Pyrczak Publishing.

Mijwil, M. (2017). Yapay Sinir Ağlar Yapısı ve Fonksiyonu. *Bilgisayar Bilimleri*, 1-4.

Mitchell, T. (1997). *Machine Learning*. New York, USA: The McGraw-Hill Companies, Inc.

Mitchell, T. ve Brynjolfsson, E. (2017). Track how technology is transforming work. *Nature*. 544(7650): 290-292.

Mohammed, R., Rawashdeh, J. ve Abdullah, M. (2020, April). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*: 243-248.

Mohri, M., Rostamizadeh, A. ve Talwalkar, A. (2018). Adaptive Computation and Machine Learning. *Foundations of Machine Learning* (2nd ed.). Cambridge, USA: The MIT Press.

MonketLearn, (2021). *An Introduction to Machine Learning*. <https://monkeylearn.com/machine-learning/>, (14.03.2022).

Moosa, I. A. (2007). Operational risk: a survey. *Financial Markets, Institutions & Instruments*. 16(4): 167-200.

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

Muaz, A., Jayabalan, M. ve Thiruchelvam, V. (2020). A Comparison of Data Sampling Techniques for Credit Card Fraud Detection. *International Journal of Advanced Computer Science and Applications (IJACSA)*. 11(6).

Muhammad, I. ve Yan, Z. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing*. 5(3).

NACM (2016). *How Credit Works*. http://web.nacm.org/pdfs/educ_presentations/Principles_Ch1_v3.pdf, (01.04.2022).

Natekin, A. ve Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*. 7:21.

Navlani, A. (04.12.2018). *Naive Bayes Classification Tutorial using Scikit-learn*. <https://www.datacamp.com/tutorial/naive-bayes-scikit-learn>, (01.02.2022).

NewTechDojo. (27.12.2017). *Learn Random Forest Using Excel*. <https://www.newtechdojo.com/learn-random-forest-using-excel/>, (14.02.2022).

Niu, K., Zhang, Z., Liu, Y. ve Li, R. (2020). Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. *Information Sciences*. 536: 120-134.

Online Etymology Dictionary. (2022). *Online Etymology Dictionary*. <https://www.etymonline.com/>, (12/04/2022).

Osibanjo, F. S., Olalude, G. A., Akintunde, M. O. ve Ajala, A. G. (2015). Application of Logistic Regression Model to Admission Decision of Foundation Programme at University of Lagos. *International Journal of Mathematics and Statistics Studies*. 3(4): 27-41.

Osuna, E.E., Freund, R. ve Girosi, F. (1997). *Support Vector Machines: Training and Applications*. (Doctoral dissertation). Massachusetts: Institute of Technology and Artificial Intelligence Laboratory.

Özdemir, B. Türk Bankacılık Sektöründe Sermaye Yapısının Performans Üzerine Etkisi. *Adnan Menderes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*. 6(1): 65-76.

Özekes, S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*. 2(3): 65-82.

Öztemel, E. (2003). *Yapay Sinir Ağları*. İstanbul: PapatyaYayıncılık.

Domingos, P. ve Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Machine learning*. 29(2): 103-130.

Pal, M. (2005). Random Forest Classifier For Remote Sensing Classification. *International Journal of Remote Sensing*. 26(1): 217-222.

Pal, M. ve Mather, P. M. (2004). Assessment of the effectiveness of support vector machines for hyperspectral data. *Future generation computer systems*. 20(7): 1215-1225.

Parlakkaya, R.(1996). *İşletmelerde Mali Risk ve Riskin Muhasebeleştirilmesi*. (Yayınlanmamış Doktora Tezi). Konya: Selçuk Üniversitesi Sosyal Bilimler Enstitüsü.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*. 4(2):1883.

Polikar R. Ensemble learning. In: Zhang C, Ma Y (eds). Ensemble machine learning. *Springer Science, Business Media*. 2012:1-34.

Quinlan, J.R. (1993). *Programs for Machine Learning*. Los Altos: Morgan Kaufmann.

Ramalho, G. L. ve Medeiros, F. N. (2006). Using Boosting to improve Oil Spill Detection in SAR Images. *In 18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, (pp. 1066-1069).

Ramchoun, H., Idrissi, M. A. J., Ghanou, Y. ve Ettaouil, M. (2016). *Multilayer Perceptron: Architecture Optimization and Training*. *Int. J. Interact. Multim. Artif. Intell.* 4(1): 26-30.

Rokach, L. (2005). *Data Mining and Knowledge Discovery Handbook*. New York: Springer Science+ Business Media.

Rustam, F., Mehmood, A., Ullah, S., Ahmad, M., Khan, D. M., Choi, G. S. ve On, B. W. (2020). Predicting pulsar stars using a random tree boosting voting classifier (RTB-VC). *Astronomy and Computing*. 32: 100404.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*. 3(3): 210–229.

SAP. (2021). *Was ist maschinelles Lernen?*. <https://www.sap.com/germany/insights/what-is-machine-learning.html>, (20.03.2022).

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*. 2(3): 1-21.

Sayad, S. (2019). *Logistic Regression*. https://www.saedsayad.com/logistic_regression.htm (23.01.2022).

Sayıcı, S.C. (2017). *İhtiyaç Kredilerinde Yapay Sinir Ağları Uygulaması*. (Yayımlanmamış Doktora Tezi).İstanbul: Kadir Has Üniversitesi Sosyal Bilimler Enstitüsü.

Schlögl, M. (2020). A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach. *Accident Analysis & Prevention*. 136: 105398.

Shen, F., Zhao, X., Li, Z., Li, K. ve Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications*. 526: 121073.

Sherrod, P. H. (2003). Classification and Regression Trees and Support Vector Machines for Predictive Modeling and Forecasting. *DTREG program manual*, 281-295.

Sipahi, B. (2011). *Bankacılık İşlemleri ve Muhasebesi*. İstanbul: Türkmen Kitabevi.

Sivarajadhanavel, P. ve Chandrakumaramangalam, S.(2012). Exchange rate risk in the foreign exchange market: a challenge on corporate profitability. *Bonfring International Journal of Industrial Engineering and Management Science*. 2(3), 08-11.

Skapura, D. M. (1996). *Building Neural Networks*. Addison-Wesley Professional.

Snart, H. (19.07.2021). *Understanding your data: A series on the importance of Exploratory Data Analysis*. <https://blogs.sas.com/content/hiddeninsights/2021/07/19/understanding-your-data-a-series-on-the-importance-of-exploratory-data-analysis/>, (18.04.2022).

Song, X., Duan, Z. ve Jiang, X.(2012). Comparison of artificial neural networks and support vector machine classifiers for land cover classification in Northern China using a SPOT-5 HRG image. *International Journal of Remote Sensing*. 33(10): 3301-3320.

Srinivas, D. S. ve Kumar, M. A. (2013). Attribute and Information Gain Based Feature Selection Technique For Cluster Ensemble: Hybrid Majority Voting Based Variable Importance Measure. *IJITR*. 1(6): 607-610.

Strzelczak, S. (2007). *Operational Risk Management*. Stanislaw Strzelczak: Warsaw University of Technolog.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics*. 24: 303-329.

Svozil, D., Kvasnicka, V. ve Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*. 39(1): 43-62.

T.C. Resmî Gazete (19.10.2005). *Bankacılık Kanunu*. <https://www.resmigazete.gov.tr/eskiler/2005/11/20051101M1-1.htm>, (27.03.2022).

Tan, P. N., Steinbach, M. ve Kumar, V. (2016). *Introduction to Data Mining*. India: Pearson Addison-Wesley.

Togay, S. (2013). Nakdi Krediler: Ticari Krediler. *Bankalarda Kredi Yönetimi* (2. Baskı.). (ss. 98–120). Eskişehir: T.C. Anadolu Üniversitesi Yayınları.

Tu, M. C., Shin, D. ve Shin, D. (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. *In 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing* . (pp.183-187).

Usta, M. (1994). *Kredi Taleplerinin Değerlendirilmesi ve Tahsisi*. İstanbul: Pamuk bank T.A.Ş. Eğitim Bölümü Yayınları.

Usuğ, C. (2016). *Temel Kredi Bilgileri*. İstanbul: T.C. Anadolu Üniversitesi Yayınları.

Ünal, O. ve Altın, H. (2010). Döviz kur riski ile şirket değeri arasındaki ilişkinin İMKB otomotiv sektöründe analizi. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*. 1(26): 277-287.

Vadapalli. (24.12.2020). *Pros and Cons of Decision Tree Regression in Machine Learning*. <https://www.upgrad.com/blog/pros-and-cons-of-decision-tree-regression-in-machine-learning/>, (12.02.2022).

Vakili, M., Ghamsari, M. ve Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for Iot Data Classification. arXiv preprint arXiv:2001.09636.

Van den Brink, G. J. (2002). *Operational risk: the new challenge for banks*. Macmillan.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. New York: Springer Science+ Business Media, Inc.

Vergi Usul Kanunu.(04.01.1961). *Bilanço Esasına Göre Defter Tutma*. [https:// www.mevzuat.gov.tr/mevzuatmetin/ 1.4.213.pdf](https://www.mevzuat.gov.tr/mevzuatmetin/1.4.213.pdf), (05.03.2022).

Viadinugroho, R. (18.04.2021). *Combining SMOTE with Tomek Links for imbalanced classification in Python*. <https://towardsdatascience.com/imbalanced-classification-in-python-smote-tomek-links-method-6e48dfe69bbc>, (04.03.2022).

Viadinugroho, R.(31.05.2021). *Combine SMOTE with Edited Nearest Neighbor (ENN) using Python to balance your dataset*. <https://towardsdatascience.com/imbalanced-classification-in-python-smote-enn-method-db5db06b8d50>, (05.03.2022)

Vijaykumar, B. ve Vikramkumar, T. (2014). Bayes and Naive Bayes Classifier. *ArXiv*. 1404.0933.

Wang, L. (2021). Imbalanced Credit Risk Prediction Based on SMOTE and Multi-Kernel FCM Improved by Particle Swarm Optimization. *Applied Soft Computing*. 114: 108153.

Wang, Y., Li, Y., Song, Y., Rong, X. ve Zhang, S. (2017). Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree. *Algorithms*, 10(4): 124.

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning a case study of bank loan data. *Procedia Computer Science*. 174: 141-149.

Weinberger, K. Q., & Saul, L. K. (2008). Fast solvers and efficient implementations for distance metric learning. In *Proceedings of the 25th international conference on Machine learning* (ss. 1160-1167).

Werbos, P. (1974). *Beyond regression: New Tools for Prediction and Analysis in the Behavior Science*. (Unpublished Doctoral Dissertation). Cambridge: Harvard University.

Williams, H. (2018). An Empirical Investigation Of The Impact Of Exchange Rate Fluctuations On The Performance Of Selected Listed Firms In Nigeria. *Journal of Business Management and Economic Research*. 2: 1-10.

Wood, T. (2020). *What is the F-score?*. <https://deepai.org/machine-learning-glossary-and-terms/f-score#:~:text=The%20F%2Dscore%2C%20also%20called,positive%20or%20'negative',> (27.02.2022).

Xiong, Z. ve Huang, J. (2021). Prediction of Credit Risk with An Ensemble Model: A Correlation-Based Classifier Selection Approach. *Journal of Modelling in Management*. 16(4): 1-20.

Yang, X., Wang, Y., Byrne, R., Schneider, G. ve Yang, S. (2019). Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical reviews*. 119(18): 10520-10594.

Yuan, K., Chi, G., Zhou, Y. ve Yin, H. (2022). A novel two-stage hybrid default prediction model with k-means clustering and support vector domain description. *Research in International Business and Finance*. 59: 101536.

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics-Theory and Methods*. 49(9): 2080-2093.

Zheng, H., Sherazi, S. W. A. ve Lee, J. Y. (2021). A Stacking Ensemble Prediction Model for the Occurrences of Major Adverse Cardiovascular Events in Patients With Acute Coronary Syndrome on Imbalanced Data. *IEEE Access*. 9: 113692-113704.

Zhou, Z. H. (2012). *Ensemble Methods: Foundations and Algorithms*. Boca Raton, Florida: CRC Press.

