

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

**AĞAÇ TABANLI SINIFLANDIRMA TEKNİKLERİNİN
DENGELİ OLMAYAN VERİ SETLERİNDE
İNCELENMESİ:
KREDİ RİSK ÖRNEĞİ**

Ali ÖZTÜRK

**Danışman
Dr. Öğretim Üyesi Serkan ARAS**

İZMİR – 2021

TEZ ONAY SAYFASI



YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum “*Ağaç Tabanlı Sınıflandırma Tekniklerinin Dengeli Olmayan Veri Setlerinde İncelenmesi: Kredi Risk Örneği*” adlı çalışmanın, tarafımdan, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.



Tarih
22/09/2021

ÖZET

Yüksek Lisans Tezi

Ağaç Tabanlı Sınıflandırma Tekniklerinin Dengeli Olmayan Veri Setlerinde

İncelenmesi: Kredi Risk Örneği

Ali ÖZTÜRK

Dokuz Eylül Üniversitesi

Sosyal Bilimler Enstitüsü

Ekonometri Anabilim Dalı

Ekonometri Programı

Kredi riski, kredi müşterisinin bankadan ödünç aldığı fonu belirlenen vadede ödeyememesi sonucunda, bankanın varlığını tehlikeye sokabilecek bir risk türüdür. Bu nedenle kredi riskine sebep olabilecek müşterilerin doğru tespiti, bankanın faaliyetlerinin devamlılığı açısından büyük önem taşır. Bu noktada müşterilerini kredilendirmek için harekete geçen bankalar, olası kredi riskini bertaraf etmek için kredilendirme süreçlerini büyük bir titizlikle yürütmesi gerekir. Aksi takdirde, süreçte oluşacak bir zafiyet veya ihmal, kredi riskinin doğmasına ve bankanın finansal faaliyetlerinin aksamasına veya durmasına sebep olabilir. Özellikle teknolojinin gelişmesiyle beraber kullanımı yaygınlaşan makine öğrenmesi teknikleri, bankacılık sektöründeki kredilendirme faaliyetlerine entegre edilmiş ve kredi talebinde bulunan müşterilerin risk taşıyıp taşımadığına otomatik olarak karar veren mekanizmaların geliştirilmesine yardımcı olmuştur. Bu kapsamda müşterinin krediyi ödeyip ödeyememesinin doğru tespiti, bankanın finansal ve fiziki varlığını açısından önemli olmasına istinaden, oluşturulan karar mekanizmalarını daha güvenilir ve daha etkin hale getirmek için zamanla farklı istatistiksel ve makine öğrenmesi tekniklerinin arayışına girilmiştir.

Bu çalışmada, bankaların kredilendirme faaliyetlerini otomatik olarak gerçekleştirebileceği ve aynı zamanda kredi riskini azaltarak karlılığını arttıracak bir karar mekanizmasının oluşturulması hedeflenmiştir. Çalışma açık kaynaklı Alman, Avustralya ve HMEQ (Home Equity) kredi veri setleri üzerinde

gerçekleştirilmiştir. Hedeflenen mekanizmayı oluşturmak için on bir farklı makine öğrenmesi sınıflandırma tekniği, dokuz farklı yeniden örnekleme tekniği ve bir adet değişken seçim tekniği kullanılmıştır. Sonuç olarak çalışmada en iyi performans gösteren modeller, ağaç tabanlı modeller olarak tespit edilmiştir. Bu doğrultuda, kredi riski sorununa karşılık olarak geliştirilecek mekanizmaları yukarı yeniden örnekleme tekniği ve ağaç tabanlı model kombinasyonlarıyla hazırlamak ve aynı zamanda bu mekanizmayı değişken seçim tekniği uygulanmayan veri setleriyle çalıştırmak, bankaların kredilendirme süreçlerinde daha doğru kararlar almasına olanak sağlayacağı bulgusuna ulaşılmıştır.

Anahtar Kelimeler: Makine Öğrenmesi; Sınıflandırma Teknikleri, Kredi Risk Analizi, Yeniden Örnekleme, Ağaç Tabanlı Topluluk Öğrenme, Değişken Seçimi

ABSTRACT
Master Thesis
Examination Of Tree-Based Techniques
In Imbalanced Data Sets: The Example Of Credit Risk
Ali ÖZTÜRK

Dokuz Eylul University
Graduate School of Social Sciences
Department of Econometrics
Econometrics Program

Credit risk is a type of risk that may endanger the existence of the bank as a result of the loan customer's failure to pay the funds borrowed from the bank in the specified time. For this reason, the correct identification of customers that may cause credit risk is of great importance for the continuity of the bank's activities. At this point, banks that take action to lend their customers must carry out their lending processes with great care in order to eliminate the possible credit risk. Otherwise, a weakness or negligence that will occur in the process may cause the credit risk to arise and the financial activities of the bank to be interrupted or stopped. Machine learning techniques, which have become widespread especially with the development of technology, have been integrated into lending activities in the banking sector and have helped to develop mechanisms that automatically decide whether customers who request credit carry a risk or not. In this context, different statistical and machine learning techniques have been sought over time to make the decision mechanisms created more reliable and more effective, based on the fact that the correct determination of the customer's ability to pay the loan is important for the financial and physical existence of the bank.

In this study, it is aimed to create a decision mechanism in which banks can perform their lending activities automatically and at the same time increase their profitability by reducing the credit risk. The study was conducted on open source German, Australian and HMEQ (Home Equity) loan datasets. Eleven different machine learning classification techniques, nine different resampling

techniques and one feature selection technique were used to create the targeted mechanism. As a result, the best performing models in the study were determined as tree-based models. In this direction, it has been found that preparing the mechanisms to be developed in response to the credit risk problem with combinations of over-sampling technique and tree-based model and at the same time operating this mechanism with datasets that do not apply the feature selection technique will enable banks to make more accurate decisions in their lending processes.

Keywords: Machine Learning; Classification, Credit Risk Analysis, Resampling, Tree-Based Ensemble Learning, Feature Selection

**AĞAÇ TABANLI SINIFLANDIRMA TEKNİKLERİNİN DENGELİ
OLMAYAN VERİ SETLERİNDE İNCELENMESİ:
KREDİ RİSK ÖRNEĞİ**

İÇİNDEKİLER

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

BİRİNCİ BÖLÜM

KREDİ VE BANKACILIKTA RİSK KAVRAMI

1.1. KREDİ	7
1.1.1 KREDİ UNSURLARI	9
1.1.1.1. Zaman Unsuru	9
1.1.1.2. Güven Unsuru	10
1.1.1.3. Risk Unsuru	10
1.1.1.4. Gelir Unsuru	11
1.2. KREDİNİN FONKSİYONLARI	11
1.3. KREDİ KAYNAKLARI	12
1.4. KREDİ VEREN KURULUŞLAR	13
1.4.1. Kredi ve Kefalet Kooperatifleri	14
1.4.2. Finansman Şirketleri	14

1.4.2.1. Tüketici Finansman Şirketleri	15
1.4.2.2. Satış Finansman Şirketleri	16
1.4.2.3. Ticari Finansman Şirketleri	16
1.4.2.3.1. Faktoring	17
1.4.2.3.2. Finansal Kiralama	17
1.5. KREDİ TÜRLERİ	18
1.5.1. Ticari Krediler	19
1.5.2. Bireysel Krediler	20
1.5.2.1. İhtiyaç Kredisi	22
1.5.2.2. Taşıt Kredisi	23
1.5.2.3. Konut Kredisi	23
1.5.2.4. Eğitim Kredisi	23
1.5.2.5. Kariyer Kredisi	24
1.5.2.6. Evlilik Kredisi	24
1.5.2.7. Tatil Kredisi	24
1.5.2.8. Ev Tadilatı Kredisi	24
1.5.2.9. Kredi Kartı	24
1.6. BANKACILIKTA RİSK KAVRAMI VE RİSK ÇEŞİTLERİ	26
1.6.1. Bankacılıkta Risk Kavramı	27
1.6.2. Bankacılıkta Risk Yönetiminin Amacı	28
1.6.3. Bankacılıkta Risk Çeşitleri	29
1.6.3.1. Kredi Riski	30
1.6.3.2. Piyasa Riski	34
1.6.3.3. Likidite Riski	34
1.6.3.4. Faiz Riski	35
1.6.3.5. Döviz Kuru Riski	36
1.6.3.6. Operasyonel Risk	36

İKİNCİ BÖLÜM

MAKİNE ÖĞRENMESİ TEKNİKLERİNE GENEL BAKIŞ VE YENİDEN ÖRNEKLEME TEKNİKLERİ

2.1. MAKİNE ÖĞRENMESİ	37
2.1.1. Makine Öğrenmesi Avantajları ve Kullanım Alanları	39
2.1.2. Makine Öğrenmesi Türleri	41
2.1.2.1. Denetimli Öğrenme	42
2.1.2.2. Denetimsiz Öğrenme	42
2.1.2.3. Yarı Denetimli Öğrenme	43
2.1.2.4. Pekiştirmeli Öğrenme	43
2.2. TEKİL MAKİNE ÖĞRENMESİ TEKNİKLERİ	43
2.2.1. K-En Yakın Komşu Algoritması	43
2.2.2. Naïve Bayes Yöntemi	47
2.2.3. Lojistik Regresyon Modeli	49
2.2.4. Çok Katmanlı Algılayıcı Sinir Ağları	51
2.2.5. Destek Vektör Makineleri	52
2.3. TOPLULUK ÖĞRENMESİ TEKNİKLERİ	59
2.3.1 Torbalama (Bagging)	60
2.3.1.1. Karar Ağaçları Algoritması	62
2.3.1.2. Rastgele Ormanlar Algoritması	66
2.3.1.3. Ekstra Ağaçlar Sınıflandırıcısı	71
2.3.1.4. Oylama Sınıflandırıcısı	72
2.3.2. Artırma (Boosting)	72
2.3.2.1. AdaBoost	73
2.3.2.2. Gradyan Artırma (Gradient Boosting) Algoritmaları	73
2.4. PERFORMANS ÖLÇÜLERİ	75
2.4.1. Karmaşıklık Matrisi	76
2.4.2. Doğruluk (Accuracy)	77
2.4.3. Kesinlik (Precision)	78
2.4.4. Duyarlılık/Hassasiyet (Recall – Sensitivity)	78
2.4.5. F1- Skoru (F1- Score)	79

2.4.6. Matthews İlişki Katsayısı (Matthews Correlation Coefficient)	79
2.5. YENİDEN ÖRNEKLEME TEKNİKLERİ	80
2.5.1. Aşağı Örneklem (Undersampling)	80
2.5.1.1. Rastgele Aşağı Örneklem (Random Undersampling)	81
2.5.1.2. Near Miss Aşağı Örneklem Tekniği	81
2.5.1.3. Küme Ağırlık Merkezleri (Cluster Centroids)	82
2.5.1.4. Tomek-Links (Tomek Bağlantıları)	83
2.5.2. Yukarı Örneklem (Oversampling)	83
2.5.2.1. Rastgele Yukarı Örneklem (Random Oversampling)	84
2.5.2.2. SMOTE (Sentetik Azınlık Yukarı Örneklem Tekniği)	84
2.5.2.3. Borderline-SMOTE	85
2.5.2.4. SVM-SMOTE	86
2.5.2.5. Uyarlanabilir Sentetik Örneklem (ADASYN)	87

ÜÇÜNCÜ BÖLÜM

UYGULAMA

3.1. ÇALIŞMANIN AMACI VE KAPSAMI	90
3.2. LİTERATÜR	91
3.3. DENEY TASARIMI	102
3.3.1. Veri Setleri	106
3.3.1.1. Avustralya Kredi Veri Seti	106
3.3.1.2. Alman Kredi Veri Seti	106
3.3.1.3. HMEQ Kredi Veri Seti	107
3.3.2. Veri Ön İşleme	108
3.3.2.1. Eksik Değişken Değerlerinin Tamamlanması	109
3.3.2.2. Değişkenlerin Kodlanması	110
3.3.2.3. Veri Dönüşümünün Yapılması	112
3.3.3. Veri Setlerinin Keşfi	112
3.3.3.1. Alman Kredi Veri Setinin Keşfi	113
3.3.3.2. HMEQ Kredi Veri Setinin Keşfi	126
3.3.4. Eğitim ve Test Veri Setlerinin Belirlenmesi	133

3.3.5. Değişken Seçimi Yapılması	134
3.3.5.1. Destek Vektör Makinesi–Yinelemeli Değişken Eleme Yöntemi	135
3.3.6. Yeniden Örnekleme Yöntemlerinin Uygulanması	136
3.3.7. Modelin Kurulumu	138
3.3.8. Modellerin Performans Ölçümü	141
3.4. BULGULAR VE BULGULARIN YORUMLANMASI	141
3.4.1. Alman Kredi Veri Setine ait Performans İstatistikleri	142
3.4.2. Avustralya Kredi Veri Setine ait Performans İstatistikleri	160
3.4.3. HMEQ Kredi Veri Setine ait Performans İstatistikleri	178
3.4.4. Optimal Senaryoların Seçimi	196
SONUÇ	202
KAYNAKÇA	210

KISALTMALAR

ABD	:	Amerika Birleşik Devletleri
ADASYN	:	Adaptive Synthetic Sampling
BDDK	:	Bankacılık Düzenleme ve Denetleme Kurulu
CART	:	Classification and Regression Trees
CHAID	:	Chi-squared Automatic Interaction Detector
DT	:	Decision Trees
DVM	:	Destek Vektör Makineleri
FED	:	Federal Reserve
FN	:	Yanlış Negatif
FP	:	Yanlış Pozitif
GBDT	:	Gradient Boosting Decision Trees
GSYH	:	Gayri Safi Yurt İçi Hasıla
HMEQ	:	Home Equity
KNN	:	K-En Yakın Komşular Algoritması
LR	:	Lojistik Regresyon
MCC	:	Matthews Correlation Coefficient
MLP	:	Multi-Layered Perceptron
NB	:	Naive Bayes
OOB	:	Out of Bag
RF	:	Random Forest
ROS	:	Random Oversampling
RUS	:	Random Undersampling
SMOTE	:	Synthetic Minority Oversampling Technique
SVM	:	Support Vector Machines
TBB	:	Türkiye Bankalar Birliği
TN	:	Doğru Negatif
TP	:	Doğru Pozitif
XTrees	:	Extra Trees

TABLÖLER LİSTESİ

Tablo 1: Karmaşıklık Matrisi	s. 76
Tablo 2: Uygulama Senaryoları	s. 103
Tablo 3: Alman Kredi Veri Setine Ait Değişkenler	s. 107
Tablo 4: HMEQ Kredi Veri Setine Ait Değişkenler	s. 108
Tablo 5: Kategorik Kodlama Öncesi ve Sonrasındaki Toplam Değişken Sayısı	s. 112
Tablo 6: Her Bir Veri Seti için Oluşturulan Eğitim ve Test Veri Setleri	s. 134
Tablo 7: Veri Setinde Bulunan Tüm Değişkenler ile Seçilen Değişkenlerin Sayısı	s. 136
Tablo 8: Toplam Gözlem Sayısı: Alman Kredi Veri Seti	s. 137
Tablo 9: Toplam Gözlem Sayısı: Avustralya Kredi Veri Seti	s. 137
Tablo 10: Toplam Gözlem Sayısı: HMEQ Kredi Veri Seti	s. 138
Tablo 11: Senaryo 1: Alman Kredi Veri Seti Performans İstatistikleri	s. 142
Tablo 12: Senaryo 2: Alman Kredi Veri Seti Performans İstatistikleri	s. 143
Tablo 13: Senaryo 3: Alman Kredi Veri Seti Performans İstatistikleri	s. 144
Tablo 14: Senaryo 4: Alman Kredi Veri Seti Performans İstatistikleri	s. 146
Tablo 15: Senaryo 5: Alman Kredi Veri Seti Performans İstatistikleri	s. 151
Tablo 16: Senaryo 6: Alman Kredi Veri Seti Performans İstatistikleri	s. 152
Tablo 17: Senaryo 7: Alman Kredi Veri Seti Performans İstatistikleri	s. 154
Tablo 18: Senaryo 8: Alman Kredi Veri Seti Performans İstatistikleri	s. 155
Tablo 19: Senaryo 1: Avustralya Kredi Veri Seti Performans İstatistikleri	s. 160
Tablo 20: Senaryo 2: Avustralya Kredi Veri Seti Performans İstatistikleri	s. 161
Tablo 21: Senaryo 3: Avustralya Kredi Veri Seti Performans İstatistikleri	s. 163
Tablo 22: Senaryo 4: Avustralya Kredi Veri Seti Performans İstatistikleri	s. 164
Tablo 23: Senaryo 5: Avustralya Kredi Veri Seti Performans İstatistikleri	s. 169
Tablo 24: Senaryo 6: Avustralya Kredi Veri Seti Performans İstatistikleri	s. 170
Tablo 25: Senaryo 7: Avustralya Kredi Veri Seti Performans İstatistikleri	s. 172
Tablo 26: Senaryo 8: Avustralya Kredi Veri Seti Performans İstatistikleri	s. 173
Tablo 27: Senaryo 1: HMEQ Kredi Veri Seti Performans İstatistikleri	s. 178
Tablo 28: Senaryo 2: HMEQ Kredi Veri Seti Performans İstatistikleri	s. 179
Tablo 29: Senaryo 3: HMEQ Kredi Veri Seti Performans İstatistikleri	s. 181

Tablo 30: Senaryo 4: HMEQ Kredi Veri Seti Performans İstatistikleri	s. 182
Tablo 31: Senaryo 5: HMEQ Kredi Veri Seti Performans İstatistikleri	s. 187
Tablo 32: Senaryo 6: HMEQ Kredi Veri Seti Performans İstatistikleri	s. 188
Tablo 33: Senaryo 7: HMEQ Kredi Veri Seti Performans İstatistikleri	s. 190
Tablo 34: Senaryo 8: HMEQ Kredi Veri Seti Performans İstatistikleri	s. 191
Tablo 35: Senaryolara ait En Yüksek F1-Skoru Değerleri	s. 197
Tablo 36: Senaryolara ait En Yüksek Doğruluk Değerleri	s. 198



ŞEKİLLER LİSTESİ

Şekil 1: Kredi Kaynakları	s. 12
Şekil 2: Kredi Veren Kuruluşların Finansal Sektördeki Payları	s. 13
Şekil 3: Kredi Türleri	s. 18
Şekil 4: Kredi Türlerinin Gelişimi	s. 20
Şekil 5: Bireysel Kredi Türleri	s. 22
Şekil 6: Tüketici Kredileri ve Bireysel Kredi Kartlarının Gelişimi	s. 25
Şekil 7: Takipteki Krediler	s. 33
Şekil 8: Sigmoid Fonksiyonu (“S” Eğrisi)	s. 50
Şekil 9: Çok Katmanlı Algılayıcı Sinir Ağları	s. 52
Şekil 10: Destek Vektör Makineleri Doğrusal Sınıflandırma Örneği	s. 53
Şekil 11: Soft Margin ve Hard Margin Sınıflandırma Örnekleri	s. 57
Şekil 12: C Parametresinin Margin Genişliği Üzerindeki Etkisi	s. 58
Şekil 13: Doğrusal Olmayan DVM Tekniği Ait Sınıflandırma Örneği	s. 59
Şekil 14: Karar Ağacı Yapısı	s. 63
Şekil 15: Rastgele Ormanlar Algoritmasının Yapısı	s. 68
Şekil 16: SMOTE Tekniğine ait Azınlık Sınıf için Sentetik Değerler Üretme	s. 85
Şekil 17: Makine Öğrenmesi Süreçlerine ait Faaliyet Akış Diyagramı	s. 105
Şekil 18: Alman Kredi Veri Seti İçin Korelasyon Matrisi	s. 113
Şekil 19: Başvuran Müşterilerin Cinsiyet ve Medeni Durum ile Yaş Dağılımları	s. 114
Şekil 20: Kredi Miktarı ve Medeni Durum-Cinsiyete Göre Kredi Risk Durumu	s. 115
Şekil 21: Medeni Durum-Cinsiyete Göre Kredi Risk Durumu	s. 115
Şekil 22: Kefil ve Yaşa Göre Kredi Risk Durumu	s. 115
Şekil 23: Müşterilerin Meslekleri ve Krediyeye Başvuru Amaçları	s. 117
Şekil 24: Talep Edilen Kredi Miktarları ve Vadelerinin Dağılımı	s. 118
Şekil 25: Talep Edilen Kredi Miktarı ile Yaş ve Vadenin Dağılımı	s. 118
Şekil 26: Müşterilerin Yaşı ile Krediyeye Başvuru Amaçları	s. 119
Şekil 27: Medeni Durum-Cinsiyet ile Krediyeye Başvuru Amacı	s. 120
Şekil 28: Kabul ve Ret Alan Kredi Tipleri	s. 121
Şekil 29: Talep Edilen Kredi Miktarlarının Kabul ve Ret Durumu	s. 122
Şekil 30: Kredilerin Vadesi ve Kabul-Ret Durumları	s. 123

Şekil 31: Kredi Tipleri ile Vade Sürelerine Göre Kredi Risk Durumu	s. 123
Şekil 32: Müşteriye ait Bazı Bilgilere Göre Kredi Risk Durumu	s. 124
Şekil 33: HMEQ Kredi Veri Setine ait Bazı Özelliklerin Dağılım Grafikleri-1	s. 126
Şekil 34: HMEQ Kredi Veri Setine ait Bazı Özelliklerin Dağılım Grafikleri-2	s. 127
Şekil 35: Müşterinin Kredi Amacı ve Mesleklere Göre Kredi Risk Durumu	s. 128
Şekil 36: Meslek ve Temerrüte Düşme Sayılarına Göre Kredi Risk Durumu	s. 129
Şekil 37: Müşterinin Talep Ettiği Kredi Tutarı ile Kredi Risk Durumu	s. 130
Şekil 38: Müşterinin Mesleği ile Çalıştığı Yıla Göre Kredi Risk Durumu	s. 131
Şekil 39: HMEQ Kredi Veri Seti Korelasyon Matrisi	s. 132
Şekil 40: Mesleklere ve Borç/Gelir Oranına Göre Kredi Risk Durumu	s. 133
Şekil 41: Izgara (Grid) Arama ve Rastgele (Random) Arama	s. 141
Şekil 42: Senaryo 1 ve 2'ye ait F1-Skoru Grafiği: Alman Kredi Veri Seti	s. 147
Şekil 43: Senaryo 3 ve 4'e ait F1-Skoru Grafiği: Alman Kredi Veri Seti	s. 149
Şekil 44: Senaryo 5 ve 6'ya ait F1-Skoru Grafiği: Alman Kredi Veri Seti	s. 157
Şekil 45: Senaryo 7 ve 8'e ait F1-Skoru Grafiği: Alman Kredi Veri Seti	s. 158
Şekil 46: Senaryo 1 ve 2'ye ait F1-Skoru Grafiği: Avustralya Kredi Veri Seti	s. 165
Şekil 47: Senaryo 3 ve 4'e ait F1 Skoru Grafiği: Avustralya Kredi Veri Seti	s. 167
Şekil 48: Senaryo 5 ve 6'ya ait F1 Skoru Grafiği: Avustralya Kredi Veri Seti	s. 175
Şekil 49: Senaryo 7 ve 8'e ait F1 Skoru Grafiği: Avustralya Kredi Veri Seti	s. 176
Şekil 50: Senaryo 1 ve 2'ye ait F1 Skoru Grafiği: HMEQ Kredi Veri Seti	s. 183
Şekil 51: Senaryo 3 ve 4'e ait F1-Skoru Grafiği: HMEQ Kredi Veri Seti	s. 185
Şekil 52: Senaryo 5 ve 6'ya ait F1-Skoru Grafiği: HMEQ Kredi Veri Seti	s. 193
Şekil 53: Senaryo 7 ve 8'e ait F1-Skoru Grafiği: HMEQ Kredi Veri Seti	s. 194
Şekil 54: Senaryoların Veri Setlerine göre Ortalama F1-Skoru Sonuçları	s. 200

GİRİŞ

Risk, kiři ya da kurum fark etmeksizin hayatın her alanında var olan ve kontrol edilemezse eęer, büyük problemlere yol açabilen bir olgudur. Riskin olaya veya duruma göre ortaya çıkma şekli farklı olsa da istisnalar hariç, çoęunlukla tek bir etkiye sebep olabilir; “Zarar”. Oluşacak bazı risklerin kiři veya kurumlarca ne zaman ve ne şekilde oluşacağını bilmesi oldukça zordur, ancak öngörülebilirdir. Öngörülebilir riskler için önceden bir yol haritası belirlemek zararı hafifletebilir. Ancak risklere karşı herhangi bir önlemin alınmaması ya da önlem almada geç kalınmasından dolayı, riske maruz kalan kiři ve kurumların yaşayacağı kaybın bilançosu da ağır olabilir.

Günümüzde finansal kuruluşlar bünyelerinde bulunan müşterilerine günün her saati mali işlem yapma yetkisi verilmiştir. Verilen yetkiye baęlı olarak finansal kuruluşlarda işlem yapan müşteri sayısındaki artış, finansal kuruluşlar açısından her ne kadar karlı gibi gözükse de işin aslı çok farklı olabilir. Özellikle, finansal faaliyetlerde meydana gelen yoğunlukla beraber müşterilerini domine etmede zayıf kalan finansal kuruluşlar, birçok ciddi risk türleriyle karşı karşıya kalabilir. Risk olgusu etkisini özellikle mali işlem yapan sektörlerde yoğun bir şekilde gösterebilir. Bu sektörlerin başında ise bankacılık sektörü gelir. Bankacılık sektörü zaman zaman likidite, piyasa, operasyonel, kredi, faiz ve döviz kuru gibi birçok risk türüyle karşı karşıya kalabilmektedir. Ancak saydığımız bu risk türleri arasında öyle bir risk türü vardır ki; önlem alınmazsa eęer, bankacılık sektörünün varlığını bile sarsabilecek potansiyele sahiptir. Bu risk türü kredi riskidir.

Bankalar günümüzde en popüler finansal kuruluşların başında gelen ve tasarruf ihtiyacı olan gerçek veya tüzel kişilere fon sağlayan finansal aktörlerdir. Bankalar bünyelerinde bulundurdıkları fonları muhafaza etmeyi, farklı yatırım araçlarında değerlendirmeyi, kurumun karlılığını arttırmayı, kiři veya kurumlara bireysel veya ticari amaçla fon kullandırmayı misyon edinmiş devlet baęlı veya özel finansal kurumlardır. Yaklaşık 4000 yıldır varlığını sürdüren bankaların en önemli gelir kaynaklarının başında krediler yer alır. Bankalar varlıklarını sürdürebilmeleri için kredilerden elde ettikleri faiz gelirleriyle ve müşterilerinde topladıkları mevduatlar aracılığıyla hem banka içinde hem de piyasa içerisinde para akışını sağlarlar. Bu sayede, bir taraftan hem karlılığını artırıp hem de banka bünyesindeki mevduat sahibi müşterilerine sabit getiri sağlarken, dięer taraftan ekonominin büyümesine de destek

olabilirler. Bu nedenle bankalar, kurum içindeki karlılığını arttırmak veya piyasalarda istikrarı sağlamak için kredilendirme süreçlerini iyi yönetmek zorundadır. Aksi takdirde bankalar, yok olma riskiyle bile karşı karşıya kalabilecektir.

Bankaların, ticarethaneler gibi, kar etmeyi ve finansal varlığını devam ettirebilmeyi amaçlamaktadır. Bu nedenle ki mevcut aktiflerini akılcıca kullanması karlılığı arttırmak için oldukça önemlidir. Bankaların karlılıkları ise aktif karlılık oranı ve net faiz marjı ile ölçülmektedir. Aktif karlılık oranı aracılığıyla, bankanın mevcut aktiflerini kullandığı yatırımların, karlılık açısından etkin yatırımlar olup olmadığı ve bankanın bu yatırımlardan ne miktarda kar elde ettiği ölçülür. Aktif karlılık oranı, belli bir dönemde elde edilen net kar ile bankanın mevcut toplam aktiflerin birbirine oranlanmasıyla elde edilir. Net faiz marjı ise bankanın yatırımlardan elde ettiği kazanç ile ödeme yükümlülüğünden doğacak kaybın farkı olarak ifade edilir. Özellikle kredi gibi yatırımlardan elde ettiği faiz geliri ile mevduat sahibi müşterilerine ödedikleri faiz tutarı arasında farkın, bankanın mevcut toplam aktiflerine oranlanmasıyla bulunur. Küçükbaş (2017) yaptığı çalışmada, Türk bankacılığında aktif karlılığı etkileyen değişkenlerin sermaye oranı ve bankanın büyüklüğü (toplam aktifleri) olduğu çıkarımında bulunurken; net faiz marjını etkileyen değişkenlerin ise sermaye oranı ve kredi oranı olduğunu tespit etmiştir.

Bankalar için önlenmesi zor riskler, yalnızca bankanın varlığını değil bankanın yabancı kaynaklarını oluşturan mevduat müşterilerinin ve yatırım yapan diğer finansal kurumların varlığını da büyük ölçüde tehdit edebilir. Bankanın maruz kaldığı riskin etkisi, bir müddet sonra bankanın dışına çıkar ve kademeli olarak önce yerel finansal piyasalara daha sonrasında küresel piyasalara sıçrar ve topyekûn olarak finansal zararlara yol açabilir. Bu durumun en iyi örneği; 2007 yılında ABD’de patlak veren Mortgage krizidir. Mortgage kelime anlamı, ipoteğe dayalı ev kredisi demektir. 2000’li yıllarda bankacılık sektöründe artan rekabet ve FED’ in ekonomiyi canlandırmak faiz oranlarını düşürmesi, finansal kurumların ucuza kredi kullandırmasına neden olmuştur. Bu dönemde, ödemeye gücü olan (düşük riskli veya prime) müşterilerin yanında, ödemeye gücü olmayan (yüksek riskli veya subprime) müşterilere de konut kredi verilmiştir. Hatta o dönemde Mortgage kredileri, bankalar tarafından menkul kıymetlere dönüştürülüp yatırımcılara satılmıştır. FED’ in aniden faiz artırımına gitmesi sebebiyle Mortgage kredisi alan subprime müşterilerin kredileri

ödeyememesi, büyük bir finansal krize yol açmıştır. Finansal küreselleşmeyle birlikte müşterilere tanınan finansal piyasalarda faaliyet serbestisi hakkıyla, ülkelerin finansal piyasalarını birbirine bağlı hale getirmiştir. Bu sebeple, Mortgage krizi ülke sınırlarını aşarak, 2008 yılında büyük kredi krizi olarak adlandırılan küresel bir finansal krize dönüşmüş ve tüm ülke piyasalarında çok ciddi mali kayıplara sebep olmuştur.

Kredi derecelendirme kuruluşlarından biri olan Fitch' e göre, ABD'de 2020 yılında, COVID-19'un yarattığı ekonomik sıkıntılara da bağlı olarak, temerrüt (müşterinin kredi borcunu geri ödenmemesi) oranında 2010 yılına bu yana ciddi bir artış yaşandığı ölçülmüştür (Fitch, 2020). BDDK'ye göre Türkiye'de, COVID-19'un patlak vermesinden bu yana maksimum 20 milyar TL'lik tüketici kredisi (kredi kartı dahil), takipteki (temerrütlü) krediye dönüşmüştür. Bu miktar, pandemi sebebiyle ekonomide yaşanan sıkıntılara da bağlı olarak, 2012 yılından bu yana kaydedilen en yüksek takipteki kredi tutarı olmuştur (BDDK, 2021). 2008 yılından beri devam eden ve COVID-19 pandemisi ile de büyük önem kazanan parasal genişleme ve borçlandırma politikaları sayesinde bireysel ve ticari kredilerin kullanımının yanı sıra, ülke borçları da büyük artış yaşanmıştır. Dolayısıyla, olası temerrüt riskinin kontrol altına alınması hususunda, doğru ve etkin risk ölçüm tekniklerinin belirlenmesi de büyük önem kazanmıştır.

Risk ve güven birbiriyle ters ilişkiye sahip iki kavramdır. Müşterilere olan güvenin azalması, riskin artışına neden olarak ve finansal kuruluşların ciddi mali kayıplar yaşamasına sebebiyet verebilir. Bu nedenle, yaşanacak olası bir güven probleminde bankalar, mağdur olmamak adına, müşterilere kredi adı altında sağladığı fonları belli bir güvence bedeliyle vadesinde ödeme şartı koşarlar. Buna bağlı olarak, belirlenen vadelere karşılık olarak faiz oranlarını çok iyi ayarlamak ve ileride yaşanabilecek aksaklıklardan doğacak riskleri de gözüne alarak bankaların, müşterilerinden kendi zararını da karşılayacak miktarda bir güvence bedeli talep etmeleri gerekebilir. Güvence bedeli talebinin yanı sıra bankalar, kredi riskini azaltmak için potansiyel müşterileri hakkında istihbarat faaliyetleri de yürütürler. Bu istihbarat faaliyetleriyle amaçlanan, kredi talebinde bulunan potansiyel müşterileri hakkında güvenilir bilgiler elde ederek, bilgi asimetrisi sebebiyle verilecek yanlış kararlardan doğacak kredi riskinin önüne geçer.

Kredi riskinin ortaya çıkmasına zemin hazırlayan üç ana neden vardır. Bunlardan birincisi müşterinin kredi başvurusu esnasında ibraz ettiği bilgi ve belgelerle bilgi asimetrisi oluşturabilmesidir. Bilgi asimetrisinin oluşması bankaların kredi operasyonlarını, talepte bulunan müşterinin kredi almaya uygun olduğu konusunda yanıltacak ve bu nedenle temerrüt (krediyi geri ödememe) olasılığı yüksek olan müşterinin kredi talebinin onaylanmasına neden olabilir. İkincisi ise kredilendirme işleminde sorumlu personel veya personellerdir. Çünkü personelin gözden kaçıracağı en ufak detay bile riski minimum olan müşterinin kredi talebi reddedilebilirken, riski maksimum olan müşterinin ise ihtiyacı olan krediyi almasına neden olabilir. Bu bağlamda risksiz olan müşterinin talebini reddedip riskli olan müşteriyi kredilendirmesi, bankanın hem karlılığını düşürecek hem de kredi riskiyle karşı karşıya kalmasına neden olabilir. Üçüncü neden ise müşterinin inisiyatif dışında gelişen olaylardır. Müşteri krediyi aldıktan sonraki dönemde işsiz kalması veya iflas etmesinin yanı sıra ekonomik kriz gibi ulusal ya da uluslararası olayların meydana gelmesi, müşteriye krediyi geri ödeme güçlüğü yaşatabilir ve buna bağlı olarak kredi riskinin ortaya çıkmasına neden olabilir. Tüm bunlara ek olarak, yıllar içerisinde kredi talebinde meydana gelen artış, bankanın kredi operasyonlarından sorumlu olan personelin iş yükünü de arttırmıştır. Çok sayıda müşterinin tek tek değerlendirilmesi sebebiyle iş yükünde meydana gelen bu artış, bankaya ve personele hem zaman hem de maliyet açısından bir dezavantaj sağlayabilir. İşte bu tür sorunların gün yüzüne çıkmasıyla beraber bankaların hem etkin hem de hızlı kararlar veren otonom karar mekanizmalarına olan ihtiyacı artar. Özellikle teknolojinin gelişmesiyle beraber yapay zekâ temelli otonom yapıdaki karar mekanizmalarının farklı iş kollarına ve sektörlerle entegre edilmesi hem personelin iş yükünü hafifletmiş hem de ilgili iş kolunda daha sağlıklı ve daha güvenilir kararlar almasına yardımcı olur. Bunun yanı sıra, karar verme faaliyetlerinin personelden alınıp makineye devredilmesi, ilgili sektörün herhangi bir riskle karşı karşıya kalma olasılığını da azaltabilir.

Potansiyel müşterinin kredi almaya uygunluğunu tahmin etmek için kullanılacak veri setleri, dengeli olmayan veri seti olma özelliğine sahip olabilir. Bu nedenle, kredi talebini değerlendirilmesi gibi finansal kökenli gerçek hayat problemlerine ait veri setlerinin büyük bir çoğunluğunda ortaya çıkabilir. Belli sınıflara göre etiketlenen (örneğin; kredi alabilir ya da kredi alamaz şeklinde)

müşterilerin dahil oldukları sınıfın sayıca dengesiz olması, dengeli olmayan veri setlerini oluşturur. Yani 1000 adet müşterinin olduğu bir veri setinde; müşterilerin 600 adedi kredi alabilir, 400 tanesi kredi alamaz olarak etiketlenmiş olsun. İşte bir sınıfın diğer sınıftan sayıca üstün olması dengeli olmayan veri setlerini oluşturur. Kullanılacak bu tip veri setleri, gelecekteki müşterinin kredi alıp almayacağını tahmin ederken çoğunluk sınıfa göre sınıflandırma yapacağından, makine öğrenmesi tekniklerinin tahmin performansını büyük ölçüde etkileyebilir ve ilgili problemin çözümü için yanlış kararlar alınmasına neden olabilir.

Bu çalışmanın amacı, bankacılık sektöründeki kredilendirme faaliyetlerini otomatikleştirebilmek ve kredi riskini minimize ederek bankanın finansal varlığını devam ettirebilmesine katkı sağlayabilmektir. Bunun için, kredi talebinde bulunan riskli müşterileri en doğru şekilde tahmin eden makine öğrenmesi modeli tespit edilecektir. Çalışmada kullanılan makine öğrenmesi modelleri, tekil ve topluluk öğrenmesi modelleri olarak ayrılacak ve kullanılan tekil modeller ile topluluk öğrenmesi modellerinin (özellikle ağaç tabanlı olanların) tahmin performansları birbirleriyle karşılaştırılacaktır. Aynı zamanda, gerçek hayat problemlerine ait veri setlerinde sıkça karşılaşılan ve sınıflandırma modellerinin tahmin performansını etkileyerek yanlış kararlar alınmasına neden olan sınıf dengesizliği sorununu gidermek için bazı yeniden örnekleme teknikleri kullanılacaktır. Çalışmada; Alman, Avustralya ve HMEQ kredi veri setleri kullanılmıştır. Her bir kredi veri seti için sekiz farklı uygulama senaryosu belirlenmiştir. Bu uygulama senaryoların içeriği, değişken seçim tekniğinin kullanılıp kullanılmamasına, veri setini dengeleyen yeniden örnekleme tekniklerine ve tahmin performansı ölçülen tekil ve topluluk öğrenmesi modellerine göre farklılık göstermektedir. Çalışmanın sonucunda ise, belirlenen uygulama senaryolarının arasından optimal sonucun elde edildiği senaryo belirlenmiş ve gelecekte bankacılık sektöründe yaşanabilecek aynı tipteki problemlerin çözümü için önerilmiştir.

Çalışmanın birinci bölümünde, kredinin genel bir tanımı yapılmış ve bankacılık sektörü için kredinin öneminden söz edilmiştir. Aynı zamanda kredinin unsurlarından, kredi kaynaklarından, kredi fonksiyonlarından ve kredi türlerinden bahsedilmiştir. Sonrasında, kredi veren diğer finansal kuruluşların neler olduğundan ve bu finansal kuruluşların işleyişine kısaca değinilmiştir. Tüm bunların yanı sıra,

riskin genel tanımı yapılmıştır. Ayrıca finansal ve bankacılık açısından riskin ne anlam ifade ettiğine değinilmiş ve riskin yönetim amaçları ele alınmıştır. Bölümün sonunda ise bankacılık sektörünün karşılaşılabileceği risk türlerinden ve özellikle bankacılık sektörü için büyük önem taşıyan kredi riskinin ne olduğundan, kapsamından ve bankaların kredi risklerini azaltabilmek için müşterilerinde aradıkları özelliklerden bahsedilmiştir.

Çalışmanın ikinci bölümünde, makine öğrenmesi tekniğinin genel tanımı yapılmış, ayrıca makine öğrenmesi tekniklerinin avantajlarından ve kullanım alanlarından bahsedilmiştir. Sonrasında, çalışmada kullanılan tekil ve topluluk öğrenmesi modelleri tanımlanmıştır. Aynı zamanda, kullanılan bu modellerin tahmin performansını değerlendirmeye yarayan performans ölçüm istatistiklerinden de bahsedilmiştir. Tüm bunların yanı sıra, sınıflandırma modeli olarak kullanılan hem tekil hem de topluluk öğrenmesi modellerinin dengeli olmayan veri setlerindeki tahmin performansını arttırabilmek amacıyla bu veri setlerini dengelemeye yardımcı olan yeniden örnekleme tekniklerinden bahsedilmiş ve buna dayanarak çalışmadaki kredi veri setlerini dengelemek için kullanılan aşağı ve yukarı örnekleme teknikleri hakkında detaylı bilgi verilmiştir.

Çalışmanın üçüncü ve son bölümünde ise, uygulamanın amacından ve kapsamından bahsedilmiştir. Bunun yanında uygulama için kurulan deney tasarımı ana hatlarıyla açıklanmış ve uygulama yapılırken hangi aşamaların takip edileceği hem sözlü hem de görsel olarak ifade edilmiştir. Tüm bunlara ilave olarak, çalışmada kullanılan üç kredi veri seti hakkında detaylı bilgi verilmiş ve bu veri setlerine uygulanan değişken seçim tekniği olan destek vektör makinesi-yinelemeli değişken eleme yönteminden de kısaca bahsedilmiştir. Son olarak, uygulama sonucunda elde edilen sınıflandırma modellerine ait performans ölçüm istatistikleri tablolar haline getirilmiştir. Bu performans ölçüm istatistiklerine dayanarak, değişken seçim tekniği uygulanan ya da uygulanmayan kredi veri setleri arasından, kredi riskine sahip müşterileri tahmin etme performansı yüksek olan sınıflandırma modelleri tespit edilmeye çalışılmıştır.

BİRİNCİ BÖLÜM

KREDİ VE BANKACILIKTA RİSK KAVRAMI

1.1. KREDİ

Kredi, kişi ve kuruluşların, ihtiyacı olan malın ya da paranın gerekli şartları sağlaması halinde, banka veya finansal kuruluşlardan temin edebileceği bir finansman kaynağıdır. Borç veren taraf, borçlunun ihtiyaç duyduğu mal veya paradan belli bir süreliğine vazgeçerek, belirlenen süre içerisinde belli bir risk ve getiri karşılığında borç alan taraftan ödenmesini ister.

Bir diğer tanıma göre ise kredi, satın almaya olanak sağlayan, maddi özellik taşıyan veya taşımayan bir malın, belli bir faiz oranı ile belirlenen süre içerisinde ödenmesi taahhüdü sağlandıktan sonra bir başkasına ihtiyaçlarını karşılaması için verilen bir kaynaktır (Çağırın Tuncer ve Ulusoy, 2019: 1).

Kredi kelimesinin kökeninden bahsetmek gerekirse; kredi kelimesi, eski Latince *“güvenmek, inanmak, emanet etmek”* anlamına gelen *“Credere”* kelimesinden türetilmiştir. Yıllar geçse de kredi kelimesi hem köküne hem de taşıdığı anlama bağlı kalmıştır. Kelimenin kökeninden de anlaşılacağı üzere, borç verenin kredi veya fon sağlaması, ancak borçlunun anaparaya ek olarak belli bir faiz oranı kadar parayı ve söz konusu parayı zamanında ödemesi gibi belirlenen yükümlülükleri yerine getirme yeteneğine ve isteğine güvenmesine bağlıdır (Golin ve Delhaise, 2013: 1).

Bir kredi işlemi, nakdi bir varlığın borç verenden borçluya transferini içerir. Bu nakdi varlık genellikle paradır. Borçlu, ödünç aldığı anaparaya ek olarak belirlenen faiz oranı kadar bir tutarla ileri bir tarihte geri ödeyeceğine söz verir. Kredi veren için bu faiz tutarı, parayı verdikten sonra oluşabilecek gecikmelerin yanı sıra diğer ilgili risk durumları için bir tazminat bedelidir (Benzschawel, 2012: 1).

Bankalar için kredi önemli bir finansal araçtır ve en önemli gelir kaynaklarının başında gelmektedir. Bankalar, hizmet ve görev açısından farklılık gösterse de asıl görevleri; kredi adı altında likidite talebinde bulunan müşterileri ile ödünç kredi vermeye niyetli müşterileri bir araya getirmek ve bu müşterilerin arasındaki likidite alışverişini güvenle yürütebileceği bir ortam sağlamaktır (Kaygısız Ertuğ ve Girginer,

2015: 308). Banka ve diğer kredi veren kurumlar, fon fazlasına sahip şahıs ve kurumlardan elde ettiği mevduat gelirlerini, belli bir faiz oranı ile kredi adı altında, fon yardımına ihtiyaç duyan gerçek ve tüzel kişilere iletilmesinde bir aracılık görevi üstlenmektedir. Bu sebeple hem tasarrufların toplanması hem de bu tasarrufları ihtiyacı olan gerçek ve tüzel kişilere ulaştırılması bakımından bankalar, günümüzde ülke ekonomilerinin istikrarını ve piyasa içerisindeki nakit hareketliliğinin sağlanmasından mükellef, etkili ve önemli kuruluşlar haline gelmiştir (Erdaş ve Eroğlu, 2018: 148).

Bankacılık sektöründe kredi terimi; 5411 sayılı Bankacılık Kanunu'nun 48. Maddesinde de belirtildiği gibi;

Bankalarca verilen nakdî krediler ile teminat mektupları, kefaletler, aval, ciro, kabul gibi gayri nakdî 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, gayri nakdî 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.

Bankalar; likidite sağlayıcıları, vade dönüştürücüleri ve risk yöneticileri gibi temel ekonomik roller üstlenmişlerdir. Bu roller, bankaların fona ihtiyacı olan müşterilerine sağlayacağı sermaye için bilgi, risk analizi ve gözlemleme konusunda avantajlar sağlayarak, doğacak risklere karşı daha hızlı müdahale edebilmesine yardımcı olabilir. Böylelikle öz sermayesini kredi olarak kullandığı müşterilerini takip eder ve beklenmedik risklere karşılık bünyelerinde o riski karşılayacak yeterli miktarda sermayeyi bulundurur. Ayrıca kredilerin faiz oranına dahil edilen risk primiyle beraber kredilerini etkin bir şekilde sigortalamış olurlar (Ayadi, 2019: 6).

Bankacılığın temeli esasen, borç alma talebinde bulunan birey ve kurumların kredi alabilirlik durumunu, yani başka bir deyişle bankanın kredi ve borç verme standartlarını karşılayıp karşılayamadığını belirlenmesidir. Bunu banka kendine has

olan özel borç verme standartları çerçevesinde ve diğer bankalarla rekabet içerisinde belirler. Bu süreçte hiçbir banka diğer bankaların kredi verme standartları konusunda bir bilgiye sahip değildir (Gorton ve He, 2008: 1181).

Diğer yandan bankalar, müşterileri olan kişi ve kurumlar ile ilgili istihbarat faaliyetleri gerçekleştirir. Bunun sonucunda banka, hali hazırda yasalarını ve yükümlülüklerini de hesaba katarak bir değerlendirme yapar. Bu değerlendirme sonucunda, uygunluğu tespit edilmiş kredi talep eden kişi ve kurumlara, güvenceli ya da güvencesiz olarak; para, teminat ya da kefalet vermek gibi sağladığı imkân veya limit olarak tanımlanabilir (Akgüç, 2010: 7).

Literatürde yapılan tanımlarından da anlaşılacağı üzere kredi, güven ve gerekli şartların karşılanması üzerine kuruludur. Borç veren kişinin fon vermeye istekli olması, borç alan kişinin ise alınan fonları geri ödemeye istekli olması ve aldığı fonu ödeme gücüne sahip olması gerekmektedir. Ayrıca borç veren, borçluya sağladığı fondan kar etmek ve aynı zamanda doğabilecek riskleri garanti altına almak istemektedir. Bunun için belli bir faiz oranı belirmesinin yanı sıra borçlunun aldığı fonu belirlenen süre içerisinde ödemesi gibi sorumlulukları da yerine getirmesini ister. Bir sonraki başlıkta ise dört önemli kredi unsurundan bahsedilmiştir.

1.1.1 KREDİ UNSURLARI

Kredi unsurları; zaman, güven risk ve gelir olmak üzere 4 ana unsurdan oluşmaktadır.

1.1.1.1. Zaman Unsuru

Kredinin, bankalar ve kredi veren kurumların potansiyel müşterilerine, ihtiyaç duydukları fonu belli bir süre içerisinde ödenmesi şartıyla verdikleri bir finansman kaynağı olduğunu belirtmiştik. Bu kapsamda kredi açısından zaman unsuru, geri ödenmesi için belirlenen vadenin, kendi içerisinde sürprizler barındırması sebebiyle bankanın gelir ve maliyet dengesini etkileyebilen önemli bir unsurdur. Kredinin zamanında ödenmesi, bankanın varlığı ve karlılığı açısından büyük önem teşkil etmektedir. Borcun geri ödeme süresinin uzaması veya belirlenen süre içerisinde kredinin ödenmemesi, geleceğin belirsizlik taşıması sebebiyle riski de arttırabilir.

Riskin artması, borç verenin karının artmasına, borcu üstlenenin ise maliyetinin artmasına sebep olur (Başar, 2006: 130).

1.1.1.2. Güven Unsuru

Güven, bankalar tarafından kredi verilirken sağlanması gereken unsurların en başında gelmektedir. Finansman kaynağı sağlayan bankalardan kredi alan kişi, kurum ve işletmeler üstlendikleri para, teminat ve kefaletin belirli süre içerisinde geri ödenmesi üzerine bir anlaşma yaparlar. Bankalar bu anlaşmayı yaparken, borcu üstlenene karşı bir güven probleminin oluşmamasını ister (Aras, 1996: 4; Kavcıoğlu, 2003: 6).

Bankalar güveni sağlamak için krediyi kullandırmadan önce potansiyel müşterilerinin mali tablolarını incelemekte ve haklarında bir dizi istihbarat işlemleri yürütmektedir. Güvenilir olmak sadece bankalar için değil, aynı zamanda krediyi talep eden; kişi, kurum ve işletmeler için de çok önemli bir unsurdur. Kredi alan, finansal kaynak elde ederken, aynı zamanda kredi sicili de zarar görmemiş olur. Sicilin zarar görmemesi, müşterilerin ileride oluşacak tasarruf eksikliklerini tamamlamak için ihtiyaç duydukları kredileri temin etmede büyük kolaylık sağlar (Başar, 2006: 130).

1.1.1.3. Risk Unsuru

Risk, kredinin verildiği tarihten kredinin geri ödeneceği tarihe kadarki sürede meydana gelmesi muhtemel olumsuzlukların tamamına denir. Krediler kullandırıldıkları andaki ekonomik şartların iyi olması ve bankalarca kredi talep eden kişinin, talep ettiği fon miktarını ödeyebileceğinin güven ve kanaatine varılması üzerine verilir. Kredi alan kişinin işsiz kalması, kredi alan işletme ve kurumların iflası veya ulusal ve uluslararası ekonomik problemlerin oluşması, kredinin geri ödenmemesi riskini arttıracaktır. Bankalar bu riskleri etkisini minimize etmek için kredi verirken, sağladığı fon ve belirlediği faiz oranıyla zamanında geri ödenmesi şartını koşar. Bu çerçevede bankacılık açısından risk, bankanın alacağı anaparaya ilave faiz geliri ödemesinin zamanında gerçekleşmemesi sonucunda oluşabilecek olası zararların toplamını ifade eder. Kullandırılan kredilerin sorunlu kredilere dönüşmesi ihtimaline karşın bankalar müşterilerinden, hem ödünç verdiği fonu ve faiz gelirini hem de borcun tasfiyesi için açılacak davalardan kaynaklı oluşacak adli masrafları

kapsayacak şekilde bir teminat alması gerekmektedir (Aras, 1996: 4; Başar, 2006: 131).

Bir riskin azalması veya artış göstermesi ise kredi alan müşterinin banka tarafından belirlenen yükümlülükleri yerine getirmemesine ve alınan güvenceye göre değişiklik göstermektedir.

1.1.1.4. Gelir Unsuru

Bankanın amacı, riskini güvence altına alması kadar verdiği krediden de kar etmek istemesidir. Bankanın giderleri, yaptığı yatırımları ve hatta kredilerin müşterilere kullandırılması bile bir maliyet unsurudur. Banka bu maliyetleri karşılamak ve mevduat müşterilerine karşı sorumlulukları yerine getirebilmek için belli faiz ve komisyon miktarlarıyla fon ihtiyacı olan müşterilerine fon sağlar. Bankaların faiz ve komisyon toplamalarının bir diğer sebebi ise finansal piyasalarda varlığını sürdürmek ve özkaynaklarını arttırarak büyüme hedeflerini gerçekleştirmektir (Aras, 1996: 5; Usta, 1996: 3).

1.2. KREDİNİN FONKSİYONLARI

Ülke ekonomilerinin büyüme ve kalkınma hedefleri açısından, o ülkenin, yatırım ve ticari faaliyetlerinin istikrarlı sürdürülebilmesi oldukça önemli bir kısıttır. Ticari faaliyet veya yatırımları gerçekleştirecek girişimciler, bazen faaliyetlerini karşılayacak yeterli fonlara sahip olmayabilir; bir finansal desteğe ihtiyaç duyabilirler. Bu noktada finansal kuruluşlar, her türlü ticari ve yatırım faaliyetlerinin hayata geçirilmesi konusunda fon anlamında en büyük desteği veren kuruluşlar olarak ticari ve sosyal hayatta yer alırlar. Finansal kuruluşların en önemlilerinden olan Bankalar, fon ihtiyacına sahip yatırımcı ve girişimcilere kurum olarak, hem maliyet ve risk unsurlarını aza indirmesi hem de karlarını maksimum düzeye çıkarması menfaati çerçevesinde, elindeki boşta olan mevcut sermaye birikimlerini fon talep edenlerin kullanımına sunar. Bu sundukları boşta olan mevcut sermaye birikimlerine kredi denir. Piyasa içinde para dolaşımı ve ticari faaliyetlerin hızlıca gerçekleşmesi sağlayan krediler, ülkelerin ekonomik ve kalkınma istikrarlarını sürdürebilmesi açısından da önemli fonksiyonlara sahip olan bir finansman kaynağıdır. Kredinin ekonomik olarak fonksiyonlarından bazıları ise şunlardır; Piyasa içerisinde nakdi varlıkların (para vb.)

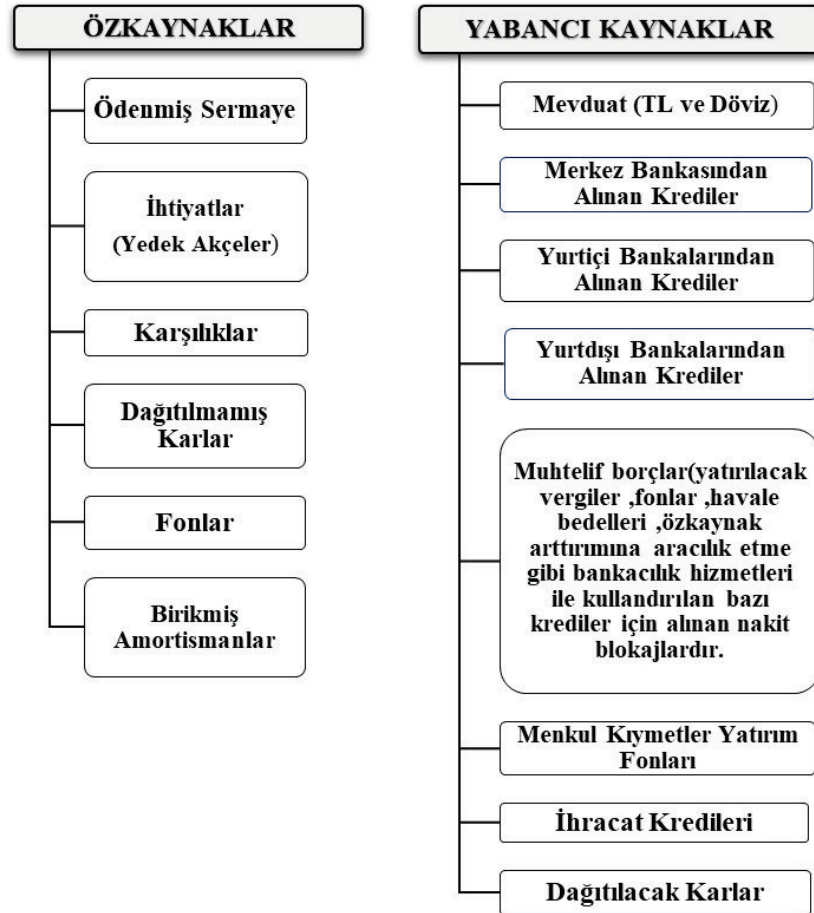
dolaşımına aracı olur (tedavül fonksiyonu). Ticari inovasyonları destekler ve inovasyon için gereken ilgili fon kaynaklarının aktarımını kolaylaştırır. Mevcut tasarrufları fon kaynaklarına dönüştürülmesi sağlayarak, fon ihtiyacı olan girişimcilerin kullanımına sunulmasını sağlar (Şakar, 2011: 5).

1.3. KREDİ KAYNAKLARI

Bankalar varlıklarının devamı açısından fonlara ihtiyaç duyarlar. Bu fonları ihtiyaç duymaları halinde özkaynaklar ve yabancı kaynaklardan temin etmeleri mümkündür. Özkaynaklar kurum içi kaynakları, yabancı kaynaklar ise bir başkasından geri ödeme koşuluyla sağlanan kaynakları temsil etmektedir. Bankalar fonlarının büyük bir kısmını yabancı kaynaklardan sağladığı mevduatlardan elde eder.

Bankaların sahip olduğu özkaynak ve yabancı kaynaklar şekil 1’de gösterildiği gibidir.

Şekil 1: Kredi Kaynakları



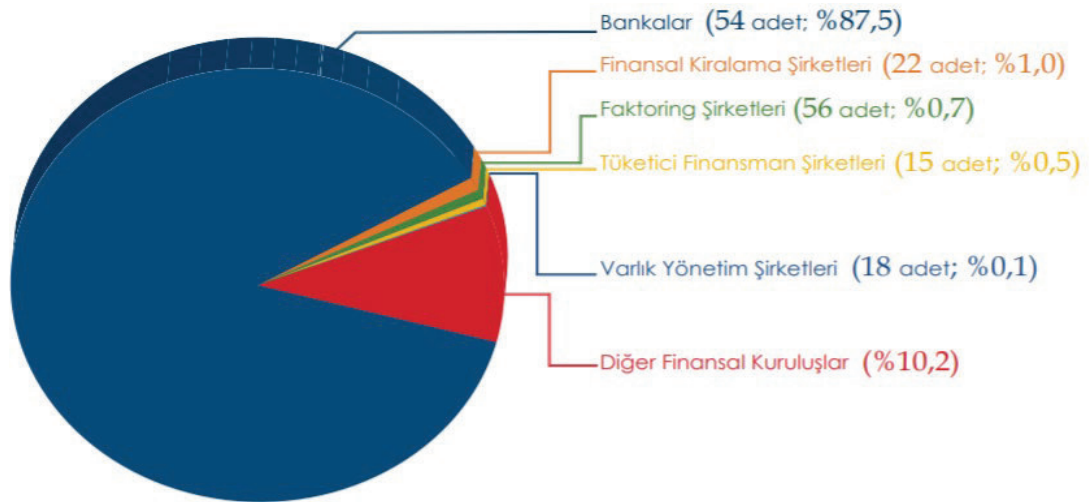
Kaynak: Şakar, 2011: 5-6

1.4. KREDİ VEREN KURULUŞLAR

Her bireyin mutlaka bir finansal desteğe ihtiyacı vardır. Araba alırken, bir iş kurarken ve hatta işletme sahibi ise işletmesi için gereken ekipmanların alınması ve masraflarının karşılanması için bile belli bir miktarda fona sahip olması gerekmektedir. Eğer tüm bu faaliyetleri gerçekleştirmek için tasarrufları ve öz sermayeleri yeterli değilse, ihtiyaç duyulan fonun sağlanmasına yardımcı finansal kuruluşlardan olan bankaların kapısını çalar. Bankalar fon için başvuran kişi ve işletmeler hakkında yaptığı araştırmalar ve incelemeler sonucunda kredi desteği sağlamaya uygun görmediği kişi ve işletmelere finansman kaynağı sağlamaması, bankalara alternatif diğer kredi veren kuruluşlar olan finansman şirketlerinin ortaya çıkmasına neden olmuştur. Bu şirketler, teminat esnekliği ve düşük faiz oranı seçenekleriyle bankalardan finansman sağlayamayan kişi ve işletmeler açısından zamanla cazip kuruluşlar haline gelmiştir. Fakat bugün ülkemizde kredi sağlama açısından en büyük paya sahip finansal kuruluş şekil 2’de de görüldüğü gibi bankalardır. Bankaları takiben sırasıyla finansal kiralama şirketleri (leasing), faktoring şirketleri ve tüketici finansman şirketleri gelir.

BDDK’nın 2020 verilerine göre, son bir yıllık dönemde toplam aktif varlıkları itibariyle bankalar % 36, faktoring şirketleri % 29,7, tüketici finansman şirketleri %30,4 ve finansal kiralama şirketleri % 20 oranında büyümüştür (BDDK, 2021b: 14).

Şekil 2: Kredi Veren Kuruluşların Finansal Sektördeki Payları



Kaynak: BDDK, 2021b: 15

1.4.1. Kredi ve Kefalet Kooperatifleri

Üyesi olan küçük ölçekli işletme ve esnafa ihtiyacı olan krediyi, diğer üyelerden temin ettiği fonlarla sağlayan ve üyelerine düşük faiz oranlarında kredi sağlamak için bankaya kefil olan birliklerdir. Kooperatifler, teminat göstermekte zorlanan küçük işletme ve esnaf adına teminat verirler. Krediyi alan işletme ve esnaf, aldığı finansman kaynağını faiz geliriyle beraber zamanında ödeyememesi halinde borç, Halk Bankasındaki kooperatife ait blokeli hesaplardan temin edilir. Böylece krediyi alan esnaf ve küçük işletmelerin krediyi ödeyememesi veya geciktirmesi halinde itibarlarına gelecek zararlar engellenmiş olur (T.C. Ticaret Bakanlığı, 2019).

1.4.2. Finansman Şirketleri

Finansman şirketleri kişi ve işletmelere bankalardan temin edemediği konut, taşıt, ihtiyaç kredileri ve finansal kiralama faaliyetleri gibi her türlü hizmet ve finansman kaynağını sağlayan, bir veya birden fazla alanda da faaliyet gösterebilen şirketler birliğidir.

1900'lü yılların başında gelişmiş ülkelerde faaliyete başlayıp İkinci Dünya Savaşı'ndan sonra daha da ilgi gören bu şirketler, kişi veya işletmelere finansman kaynağı sağlayarak, ihtiyacı olan mal ve hizmeti satın almasına aracı olan uzmanlaşmış şirketlerdir (Britannica, 2012). Türkiye'de ise finansman şirketleri gelişmiş ülkelerdeki şirketlerin faaliyete başlamasından yaklaşık 80 yıl sonra, kuruluş ve çalışma esasları Hazine Müsteşarlığı tarafından belirlenmiş ve 26 Temmuz 1994 tarihli 22002 sayılı Resmî Gazete' de yayımlanmıştır. Böylelikle ülke çapında bankalara alternatif fon sağlayan kuruluşlar olarak faaliyet göstermeye başlamışlardır (Canbaş ve Doğukanlı, 2012: 297).

Finansman şirketlerinin kuruluşu, tüccarların taksitli kredi vermesine dayanmaktadır. 20. Yüzyılın başlarında otomobil almak isteyen tüketicilerin ellerinde yeterli nakit bulunmaması nedeniyle bankalara kredi talebinde bulunmaktaydı. Talep edilen krediler düşük miktarda ve düşük faiz oranında olduğundan bankalar kredi verme işine pek sıcak bakılmamaktaydı. Bu da hem tüketicinin isteklerinin gerçekleştirilememesine hem de otomobil alımlarının önemli derecede düşmesine sebep olmaktaydı. Bu duruma çözüm olarak otomobil satıcıları taksitli krediler veren

finansman şirketlerini kurmuşlardır. Hem satış hem de karlılık açısından büyük fayda sağlayan bu şirketler, kapsamlarını genişleterek tüketicilerin başka dayanıklı tüketim mallarını satın almasına yardımcı olacak krediler de vermeye başlamışlardır (Mishkin ve Eakins, 2015: W-49).

Finansman şirketler; ticari bankalar, tasarruf kurumları ve kredi birlikleri gibi finansal kuruluşlarının da görevi olan finansal hizmetleri kredilendirmektir. Şirketler, kişi ve kurumlara vereceği uzun ve kısa vadeli finansman kaynaklarını, bankalar gibi mevduat toplama yetkileri olmadığından, kendilerine ait borç senetlerini satarak ya da Ticari bankalar ve yatırımcılardan elde ederler (Saunders ve Cornett, 2018: 70; Mishkin ve Eakins, 2018: W-49).

Bugün de faaliyetlerini sürdürmeye devam eden bu şirketler, faaliyetleri açısından; tüketici finansman şirketleri, satış finansman şirketleri ve ticari finansman şirketleri olmak üzere üçe ayrılır.

1.4.2.1. Tüketici Finansman Şirketleri

Tüketici Finansman şirketleri, araba ve ev eşyaları başta olmak üzere tatil masraflarından hastane hizmetlerine eğitim masraflarından faturalara kadar birçok dayanıklı tüketim mallarını ve hizmetlerini finanse eden finansman şirketleridir. Bu şirketlerin kredileri, nakdi olarak verdikleri krediler ve dayanıklı tüketim mallarının alımı için verdikleri krediler olmak üzere ikiye ayrılır. Dayanıklı tüketim malları için verilen kredilere uygulanan faiz oranları, nakdi olarak verilen kredilere göre daha düşüktür. Bunun sebebi, satın alınan dayanıklı malın krediye teminat olarak gösterilebilmesidir. Nakdi krediler, müşterilerden herhangi bir teminat talep edilmeyen kredilerdir. Bu tip krediler için anlaşmaya atılan bir imza, kredi ihtiyacının karşılanmasında yeterli olabilir. Ancak, bu kredileri talep eden müşterilerin kazanç durumlarında meydana gelebilecek değişimlerin belirsizliğinden ötürü, nakdi krediler, riski yüksek krediler olma özelliğine sahiptir. Finansman şirketleri, kullandığı dayanıklı tüketim malları kredilerinde talep edilen meblağ arttıkça faiz oranlarında da artış olması beklenirken; gösterilen teminat sayesinde, talep edilen kredilerin makul faiz oranlarında alınmasına olanak sağlar. Finansman şirketlerinin kullandığı nakdi kredinin tutarı, dayanıklı malların alımında kullanılan kredilerden daha düşük olmasına rağmen bu kredilerin faiz oranları, dayanıklı tüketim malları kredilerinde

uygulanan faiz oranlarına göre daha yüksektir. Bunun sebebi kredilerin gelecekte ödenmeme ihtimalleri ve bu tür kredi için bir teminatın alınmamasıdır. Bu durum nakdi kredilerin riskinin yükselmesine de neden olur (Rose ve Marquis, 2008: 508; Taş, 1998: 39).

1.4.2.2. Satış Finansman Şirketleri

Satış finansman şirketleri, kişi veya işletmelerin ihtiyacı olan ürün ve hizmetleri, anlaşma yaptığı belli üreticiler, perakendeciler ve toptancılardan satın almaya teşvik etmek için finansman sağlayan şirketlerdir. Satış finansman şirketlerinin en önemli görevi, üretici veya perakende firması tarafından satışı yapılan ürünlerin satın alınması için müşterilere finansman sağlamak ve firmaların satış oranını arttırmasına destek olmaktır. Genellikle bu şirketler, hızlı bir şekilde kredi sağladıkları için bankalarla rekabet etmektedirler. Satış finansman şirketleri satışları arttırmak için genellikle bankalardan ve diğer finansal kuruluşlardan daha düşük faiz oranlarında kredi kullandırır. Satıştan elde edilen karlarla beraber, kredi verildiğinde oluşan zararlar da karşılanmış olur (Mishkin ve Eakins, 2018: W-55).

Tüketicinin ihtiyacı olan dayanıklı tüketim mallarını satan üretici firmalardan, ilgili mallara ait taksitli satış senetlerini satın alarak tüketiciye kredi temin eden satış finansman şirketlerinin çoğunun teftişi ise yine satış senetlerini aldığı ilgili üretici firmalar tarafından yapılmaktadır (Rose ve Marquis, 2008: 508).

1.4.2.3. Ticari Finansman Şirketleri

Ticari finansman şirketleri, ticari firmalara kredi vermeye odaklı olan şirketlerdir. Genellikle orta ve küçük ölçekli üreticiler, toptancılar, perakendeciler ve hizmet kuruluşları dahil olmak üzere çeşitli müşterilere faktoring hizmetleri, taksitli ve vadeli krediler, portföy satın alma işlemleri ve finansal kiralama (leasing) dahil olmak üzere çeşitli hizmetler sağlarlar. Ayrıca müşteri için gerekli olan taşıt, işletme için gerekli büyük makine ve ekipman ihtiyacının yanı sıra finansal kiralama hizmetleri için de finansman vermektedir (AICPA, 2016: 53; Rose ve Marquis, 2008: 509).

Ticari finansman şirketlerince, firmaların ve işletmelerin finansman ihtiyacını karşılayacağı en popüler finansman teknikleri; faktoring ve finansal kiralama. Bu

tekniklerin tanımlarından, sağladığı yararlardan ve kapsamlarından aşağıda kısaca bahsedilmektedir.

1.4.2.3.1. Faktoring

Faktoring, vadeli satış yapan girişimci, işletme ve firmalar için mal ve hizmet satışından sonra alacaklarını tahsil edememesi halinde, satışın belgelenmesine istinaden, oluşacak alacak haklarının faktoring şirketlerine devretmesine dayanır. Faktoringler alacak tahsilatlarını gerçekleştirerek, işini büyütmek isteyen işletmelerin alacaklarını tahsil edilememesi sonucunda oluşacak maddi sıkışıklığın hızlı ve etkili bir şekilde giderilmesine ve işletme için gerekli finansal akış istikrarının sağlanması hususunda önemli bir finansal yöntemdir. Faktoringler; satış sonrası alacaklarının tahsilatı ve vade takiplerinin yapılması ve malların satışının daha kolay yapabileceği piyasaları araştırması gibi faaliyetleri gerçekleştirmekle yükümlüdür.

Faktoring işlemleri genellikle kısa vadeli alacakları kapsadığı gibi nadir de olsa uzun vadeli alacakları da kapsayabilir. Bir kredi tekniği olarak da bilinen faktoringler; finansman sağlama, hizmet verme ve riski üstlenme gibi üç ana fonksiyonu gerçekleştirmeden de sorumludur (Berk, 2014: 213).

1.4.2.3.2. Finansal Kiralama

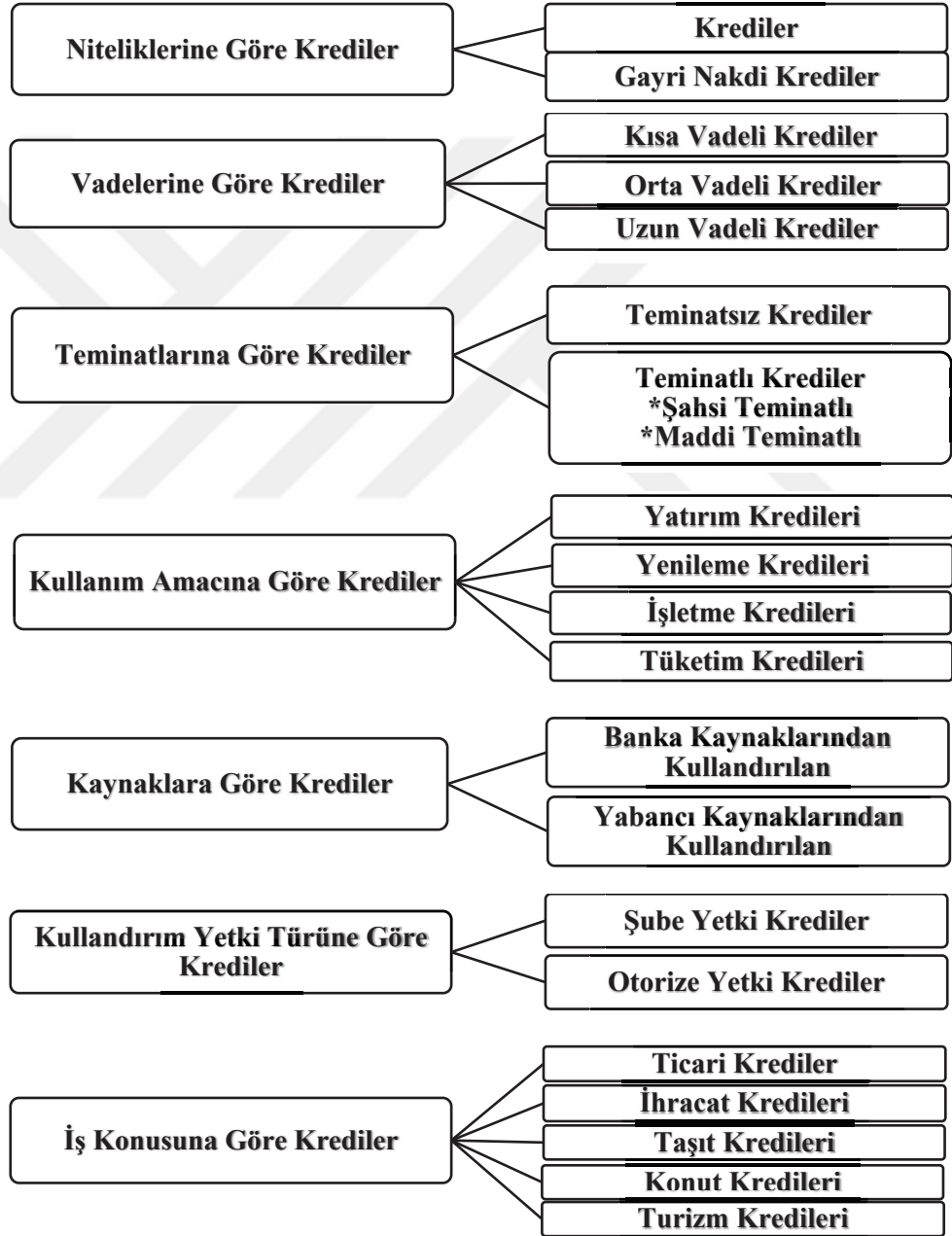
Finansal kiralama, işletme veya firma sahibin ihtiyacı olan iş makineleri, fabrika gibi mülklerin ve ekipmanların ticari finansman şirketleri tarafından alınıp, hiçbir kredi ve özkaynak kullanımını gerektirmeden belirlenen kira bedeli dâhilinde firma ve işletmelere kiraya verilmesine ve sözleşmede belirlenen sürenin sonunda belli bir bedel karşılığında mülkiyetinin kiracıya geçmesine dayalı finansal yöntemdir.

Finansal kiralama avantajları ise şunlardır; Teşvik verme ve işletmelere KDV avantajı sağlar. Yatırımcıların özkaynaklarını daha etkin kullanmasına olanak tanır. Finansal kiralamanın maliyeti banka kredilerinden daha avantajlıdır. İşletmenin iflası durumunda kiralanan makine veya ekipman bu durumdan etkilenmez (Tekin, 2015:230).

1.5. KREDİ TÜRLERİ

Kişi ve işletmeler çeşitli nedenlerden dolayı kredi kullanmak zorunda kalırlar. Bu çeşitli nedenler kredinin farklı şekilde sınıflandırılmasına yol açabilir. Krediler, şekil 3'te de görüldüğü üzere, vade, teminat, nitelik, kullanım amacı, kaynak, kullandırma yetkisi ve iş konusuna göre farklı sınıflara ayrılır (Pamukçuk, 2009: 17).

Şekil 3: Kredi Türleri



Kaynak: Aras, 1996: 8; Ulusoy, 2012a: 267

Bankanın özkaynak, alınan krediler, menkul kıymet ihracı ile sağlanan kaynaklar ve mevduatlar sayesinde oluşturulan fonlar banka bilançosunun pasif bölümünü oluştururken, menkul kıymet alma ve kredileri finanse etme şeklinde fonlarını kullandığı alanların yer aldığı kısım ise bankanın aktif bölümünü oluşturur. Bankaların aktif bölümünü oluşturan kredi türleri nakdi kredilerdir. Nakdi krediler birey ve işletmelere sağlanan kredilerden olup, ekonomik istikrar ve sosyal refahın gelişmesi için toplam talebi dengeleyen ve her türlü ticari, tüketim ve yatırım faaliyetlerini finanse etme görevini üstlenen önemli fon kaynaklarındandır. Bankacılıkta nakdi krediler, bireysel ve ticari krediler olmak üzere iki gruba ayrılır (Togay, 2013: 99).

Bu bölümde ticari kredilerin ve bireysel kredilerin tanım, işlev ve türlerinden detaylıca bahsedilmektedir.

1.5.1. Ticari Krediler

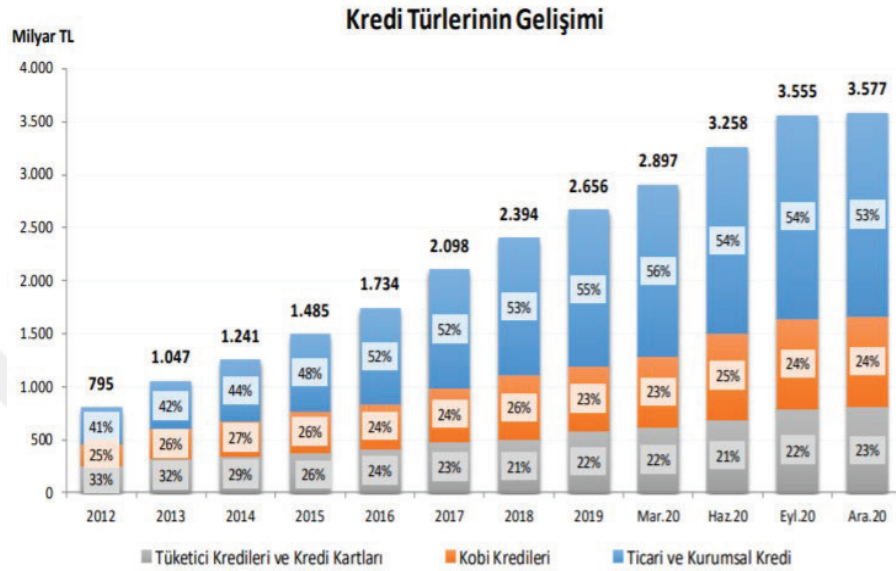
Firmaların, işletme ve sanayi kollarının faaliyet hacimleri yıl boyunca aynı kalmaz. Bu ticari kuruluşlar artan faaliyet hacimleriyle birlikte varlıklarını, hammadde ve malzeme ihtiyaçlarını, işçilerin maaş ödemeleri, vergiler, ticari kuruluşlara ait tüm masrafları bazen kendilerine ait öz sermayesiyle karşılaması mümkün olmaz. Özellikle satışlarından elde ettiği alacaklarının geç tahsil edilmesi gibi durumların meydana gelmesi ticari kuruluşları maddi sıkıntılar içine sokabilir. Bu nedenle ticari kuruluşlar, yaşadıkları maddi sıkıntıları atlatmak için kısa vadeli finansal destek alma ihtiyacı duyarlar. Bunun yanı sıra, ticari kuruluşların orta ve uzun vadeli hedefleri arasında üretim hacimlerini geliştirmek ve yatırımlarını arttırarak ekonomik olarak büyümek vardır. Bu hedefleri için orta ve uzun vadeli finansman kaynaklarına ihtiyaç duyarlar. Bu bağlamda, ticari kuruluşların varlık ihtiyacı ve giderlerinin karşılanmasının yanı sıra, üretim ve büyüme hedeflerini gerçekleştirebilmesi açısından ihtiyaç duyulan ve bankalarca ticari kuruluşlara sağlanan kısa, orta ve uzun vadeli finansman kaynağına ticari krediler denir. Ticari krediler adı altında bankaları sağladığı krediler; taksitli ticari krediler, ticari kredili mevduat hesabı, TL ihracat kredisi, döviz kredisidir (Büker, Aşıkoğlu ve Sevil, 2014: 350; Togay, 2013: 99).

BDDK'nın Aralık 2020 dönemi verilerine göre ticari ve kurumsal krediler 1.907 milyar TL iken tüketici kredileri ve kredi kartları 821 milyar TL olarak

gerçekleşmiştir (BDDK, 2021a: 11). Ayrıca şekil 4’e bakarak müşterilerin, ticari kredilere olan talebinin yıllar içerisinde büyük ölçüde arttığı söylenebilir.

Diğer kredi türlerinin gelişimi ise şekil 4’ te detaylı olarak gösterilmektedir. Grafikteki tüketici kredileri; konut, taşıt ve ihtiyaç kredilerinin toplamını içerir.

Şekil 4: Kredi Türlerinin Gelişimi



Kaynak: BDDK, 2021a: 11

1.5.2. Bireysel Krediler

Bireylerin belli ihtiyaçları ve istekleri vardır. İhtiyaçlarını ve isteklerini gerçekleştirebilmek için elde ettiği kazanç ve yapmış oldukları tasarrufları bazen yetersiz kalabilir. Bu durum bireyleri finansal kuruluşlardan fon yardımı almaya iter. Fon sağlama konusunda büyük bir öneme sahip olan bankalar bireylere, olası kredi riskini de bertaraf edecek şekilde, belli vade ve faiz oranıyla hizmet veren ve aynı zamanda bireylerin ihtiyaçlarına ve isteklerine paralel geniş fon yelpazesine sahip finansal kuruluşlardır. Bankaların fon yelpazesinde yer alan hizmetlerinden birisi de bireysel kredilerdir.

Bireysel kredilerin kullanımı 1950’de Avrupa ve Amerika’da, 40 yıl sonra ise Türkiye’de yaygınlaşmaya başlamıştır. Türkiye’de 2000 yılında ekonomik istikrar üzerine alınan kararların ardından düşen enflasyona karşılık olarak faiz oranlarında gerçekleşen düşüş, bireysel kredilere olan talebi artmıştır. Bankaların diğer kredilere oranla daha az risk taşıması ve düşük enflasyon öngörüsü ile ileride daha çok kar

getireceğini düşünen bankalar, kredilendirme konusunda uzun vadeli bireysel kredileri daha çok tercih etmeye başlamıştır (Güney, 2007: 101).

Bireysel krediler, çalışan ya da emekli olup düzenli bir gelire sahip bireylerin ticari amaç için kullanmayacağı gayrimenkul ve taşıtları veya ihtiyacı olan malları ve hizmetleri temin etmek amacıyla, gelire uygun ödeme planı ve belli bir faiz oranı dâhilinde ödenmesi beklenen, bankalarca sağlanan finansman kaynağıdır (Şakar, 2011: 193). Bireysel krediler aynı zamanda menkul veya gayrimenkul nitelikteki malların dışında, çamaşır makinesi, buzdolabı gibi dayanıklı tüketim malları ve ev tadilatı, evlilik, tatil, eğitim gibi birçok hizmetin de satın alınması için kullanılabilen fon kaynakları olarak bankalarca sağlanır (Terzioğlu, 2007: 62).

Bireysel kredileri ticari kredilerden ayıran en temel özelliği, ticari kredileri alan satıcı ve sanayi sektörleri, üretim hacmini arttırarak satışlardan doğacak kazançları yükseltmesiyken, bireysel krediyi alan bireyler ise gelirlerini arttırma amacı gütmeyen bireysel ihtiyaç ve isteklerini temin etmede kullandıkları finansal kaynaktır (İnağ, 1990: 3). Bu çerçevede bireysel kredilerin işlevi, fon talebinde bulunan bireyin alacağı mal ve hizmetin peşin parayla ödenmesi sonucunda doğacak avantajlardan yararlanması için gelecekteki kazancına şimdiden sahip olmasına imkân tanımaktır (Güney, 2007: 115).

Sanayinin gelişimi ve teknolojiye yaşanan devrimler bir tüketim furysı başlatmış ve bireyleri mal ve hizmet konusunda daha talepkar hale getirmiştir. Bireylerin temel amacı kendi sosyal refahlarını yükseltmenin yanında elzem ihtiyaçlarını da karşılamaktır. Bu noktada bireysel krediler, bireylerin taleplerin karşılanması için birtakım mal ve hizmetin alınması amacıyla verilen fonlar olsa da bu fonlar tüketim toplumlarının diğer bir kısmını, ekonomik ve psikolojik açıdan derin buhranlara sürükleyebilir. Özellikle popüler kültürün toplum içerisinde artış yönünde bir trend izlemesi, insanları gereksiz harcama ve tüketim yapmaya zorlayabilir. Bunun sonucunda ise günümüz tüketim toplumlarını aşırı borçlanmaya itebilir. Banka aracılığıyla borçlanan tüketiciler, ileride bankaya borcunu ödeyemez hale gelebilir ve bu da kredi riskinin doğmasına yol açabilir.

Bireysel krediler ülkelerin GSYİH' nin artmasına katkı sağlayarak ülke ekonomilerinin büyümesi açısından olumlu bir etkiye sahiptir. Ancak bireysel kredilere olan taleplerin fazlaşması, ülke ekonomilerinde yaşanan olumlu gelişmeler

yerini zorlu problemlere bırakabilir. Bireysel krediler sayesinde bireylerin alım gücünün artması, satılan mallara olan talebi de arttırdığından fiyatların yükselmesine neden olabilir. Bireysel kredilerin neden olduğu bir diğer problem ise cari işlem açığıdır. Alınan bireysel kredilerin ithal malların alımı üzerinde kullanılması, ülkelerin dış ticaret açığının artmasına ve bunun sonucunda ülkenin cari işlem açığı vermesine yol açar (Yüksel ve Özсарı, 2016: 6).

Bireysel kredilerin bazıları şekil 5’ teki gibidir. Bunun haricinde kişisel gelişim ve hobi kredisi, faturalı alışveriş kredisi, inşaat tamamlama kredisi, doğalgaz dönüşüm kredisi ve kredili mevduat hesabı, bireysel kredilerin arasında yer alan diğer kredilerdir.

Şekil 5: Bireysel Kredi Türleri



Kaynak: Şakar, 2011: 199

1.5.2.1. İhtiyaç Kredisi

İhtiyaç kredisi, tüketicilere ticari amaçla kullanmayacağı mal ve hizmet ihtiyacının karşılanması konusunda yardımcı olan, bankalar tarafından nakit olarak bir kerede ödenen, tüketicinin belirlenen faiz oranında ve belli bir vade farkıyla, eşit veya

değişen taksit imkânları sağlanarak, geri ödemeye yükümlü olduğu finansal kaynaklardır. Tüketicilere taksitle değil peşin alışveriş olanağı tanıyan bu krediler, peşin alışverişten sağlanacak peşinat iskontosundan yararlanmalarını sağlar. İhtiyaç kredileri teminatlı ve teminatsız şekilde verilir. Kredinin ödenmemesinden doğacak riske karşılık, riski karşılayacak kefilin ya da gayrimenkul gibi maddi teminat gösterilmesi sonucu tüketiciye tahsis edilir. (Şakar, 2011: 199).

1.5.2.2. Taşıt Kredisi

Taşıt kredileri; taşıt sahibi olmak isteyen kişilerin, belli marka kriteri olmadan, hususi amaçta kullanacakları sıfır kilometre ya da ikinci el taşıtların satın alınması için finanse edilen kredi türüdür. Bankalar bu krediyi düzenli gelire sahip gerçek kişilere verir ve taksitler halinde geri ödemesini istemektedir (Şakar, 2011: 199; Vurucu ve Arı, 2014: 374-375). Bu kredileri alırken teminat olarak bankalara alınacak taşıt gösterilir. Bu süreçte bankalar alınan taşıtı teminat kabul eder. Taşıtın satıcıdan alıcıya devri gerçekleşmesinden sonra, trafik sigorta ve kasko işlemlerinin yapılması ve taşıtın resmi olarak banka tarafından teminat altına alınmasıyla süreç sonlanır. Bu sürecin ardından taşıt için alınan kredi banka tarafından direkt alıcıya değil satıcıya, isteği dâhilinde, elden ya da banka hesabına yatırılarak ödenir (Deveci, 2019: 47).

1.5.2.3. Konut Kredisi

Bankaların konut sahibi olmak isteyen gerçek kişilere verdikleri kredilerdir. En az bir kefil ve krediyi alımında kullanacağı konutun bankaca ipotek edilmesiyle beraber, uzman incelemesiyle belirlenecek daire bedelinin belli bir kısmı ya da tamamı banka tarafından konut kredisi olarak verilir. Bankaların konuta yapılan deprem, yangın gibi afet sigortalarının, kredi ödeme süresi tamamlana kadar her yıl yenilenmesini ister ve oluşacak masrafların krediyi sağladığı kişiden alınması diğer şartlarından biridir (Şakar, 2011: 201).

1.5.2.4. Eğitim Kredisi

Eğitim masrafları karşılanması için eğitim alan öğrencinin velisine kredili vadeli mevduat şeklinde bankaca kullandırılan ve geri ödemesi eşit ya da değişken taksitlerle yapılan finansal kaynaktır. Kredi alacak kişiden kefil ve öğrencinin kayıt

belgesi ve yapılacak harcamaya dair tutarı gösteren bir belge talep edilir. Verilecek kredinin belgede gösterilen eğitim masrafını geçmemesine ve ayrıca kredi talebinde bulunan velilerin gelirleri taksitleri karşılamaya yeterli olması gerekmektedir (Şakar, 2011: 201).

1.5.2.5. Kariyer Kredisi

Yabancı dil öğrenme, yüksek lisans / doktora programlarına ve mesleki kurslara katılma gibi kariyer açısından kendilerini geliştirmeleri adına yapılan harcamaların karşılanması açısından, taksitli ödeme ya da kredili mevduat olarak fon ihtiyacı olan gerçek kişilere bankalar tarafından finanse edilen kredilerdir (Şakar, 2011: 202).

1.5.2.6. Evlilik Kredisi

Evliliğin en fazla birinci ayında olup ya da yeni evlilik kararı almış çiftlerin evlilikle ilgili oluşacak her türlü masrafın karşılanması için faizin sabit olduğu ve eşit taksitlerle geri ödenmesi yükümlülüğü olan, gerçek kişilere bankalarca sağlanan finansman kaynağıdır (Şakar, 2011: 203).

1.5.2.7. Tatil Kredisi

Yurt içinde veya yurtdışında yapılacak her türlü aktivite ve seyahatten doğacak harcamaları karşılamak amacıyla sabit faizli ve taksitli ödeme gibi yükümlülükle beraber gerçek kişilere bankalarca kullandırılan finansman kaynağıdır (Şakar, 2011: 204).

1.5.2.8. Ev Tadilatı Kredisi

Ev için her türlü tadilat ve dekorasyon masraflarının karşılanması için taksitler halinde geri ödenmesi yükümlülüğüyle bankalarca uygun görülen gerçek kişilere verilen finansman kaynağıdır (Şakar, 2011:205).

1.5.2.9. Kredi Kartı

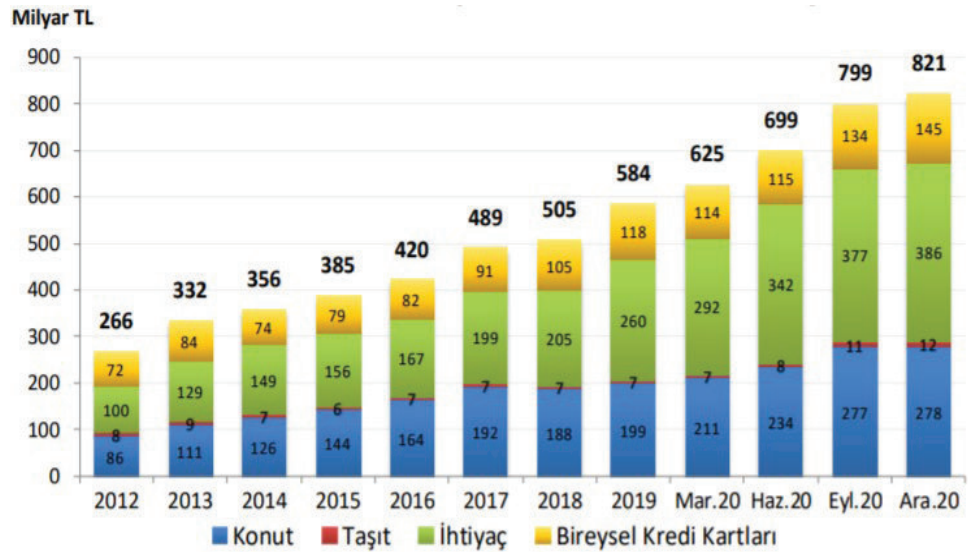
Kredi kartları, bankalar veya diğer kart çıkarmaya yetkili kuruluşlarca sağlanabilen, ihtiyaç duyulan her türlü mal ve hizmeti nakit olarak değil, bankaların

gerçek kişilerin gelirlerine ve ödeme gücüne göre belirlediği limitler arasında satın alınması sağlayan veya nakit para çekmeye olanak tanıyan nakdi olmayan finansal bir araçtır. Ek kart, banka kartı, şirket kartı ve limitli kredi kartı şeklinde temin edilebilen bu kartlar, gerçek kişilerin mal ve hizmetin satın alımında gerekecek kısa vadeli kaynak ihtiyacını gidermek için kullanılan finansal enstrümandır. Ayrıca gerçek kişilere sağlanan finansman, tüketici kredisinde bir defada kullanılırken kredi kartında ise bu finansman, ihtiyaç duyulan herhangi bir zamanda bankalarca belirlenmiş kredi limitleri çerçevesinde kullanılabilir (Kılınç ve Karaoğlu, 2014: 2; Şakar, 2011: 206).

BDDK'nın Aralık 2020'ye ait faaliyet raporunda, bir önceki çeyreğe göre ihtiyaç kredileri artış göstererek 377 Milyar TL olarak belirlenmiştir. İhtiyaç kredilerini, 277 Milyar TL ile konut kredileri, 134 Milyar TL ile Bireysel kredi kartları, 11 Milyar TL ile Taşıt kredileri takip etmektedir. Buna ek olarak, bahsi geçen kredi türlerinin Bireysel krediler içindeki paylarına bakılacak olursa eğer; ihtiyaç kredilerinin payı %45, konut kredilerinin payı %35, bireysel kredi kartlarının payı %17, taşıt kredilerinin payı %1 olarak gerçekleşmiştir.

Diğer yıllara ve dönemlere ait tüketici kredileri ve bireysel kredi kartlarının gelişimi şekil 6'daki gibidir.

Şekil 6: Tüketici Kredileri ve Bireysel Kredi Kartlarının Gelişimi



Kaynak: BDDK, 2021a: 12

1.6. BANKACILIKTA RİSK KAVRAMI VE RİSK ÇEŞİTLERİ

Gelecek, neler yaşanacağını kestiremediğimiz koca bir belirsizlik üzerine kuruludur. Yarınını kestiremeyen bireyler ve işletmeler hedeflerini gerçekleştirme konusunda yapacakları planları ortaya çıkabilecek belirsizlikleri hesaba katarak yapmalılar. Bazen hayata geçirmek istenilen planlar ile gerçekleşenler arasında bir uyumsuzluk söz konusu olabilir. Bu uyumsuzluk, riski temsil etmektedir. Bu uyumsuzluklar özellikle finans piyasalarında yakından takip edilmesi gereken faaliyetlerin başında gelir. Riski aşmak pek mümkün olmadığından riskten etkilenebilecek gerçek ve tüzel kişiler çareyi, riskleri azaltmak ve katlanmak zorunda oldukları risk durumlarını akıllıca yönetmekte bulmuşlardır. Çünkü katlanılan risk ne kadar fazlaysa elde edilen kazanç veya kayıp da o oranda artırır (Büker ve diğerleri, 2014: 65; Özbilgin, 2012: 88).

Belirsizlik ve risk kavramları arasında bir bağ olsa da aslında farklı kavramlardır. Belirsizlik ve riskin arasındaki tek bağ gelecekte gerçekleşmeleridir. Her iki kavram da gelecekte gerçekleşebilecek bir olayların veya durumların meydana gelme olasılığı hakkında aynı bilgi düzeyine sahiptir. Aralarındaki farkı ise, gerçekleşme ihtimalleri ve gelecekte ne şekilde meydana gelecekleri belirler. Belirsizlik, gerçekleşecek olay çeşidi veya olayların gerçekleşme ihtimalleri hakkında bir bilgiye sahip olunamaması durumunda meydana gelir. Risk ise gerçekleşmesi muhtemel olay senaryolarından hangilerinin meydana geleceğinin az çok kestirilebildiği, fakat olaylardan hangisinin gerçekleşeceği konusunda net bir fikre sahip olmanın zor olduğu durumları ifade eder (Cinemre, 2010: 122).

Bu bölümde bankacılık risk kavramı ve risk yönetiminin amaçlarından söz edilmiştir. Ayrıca bankayı etkileyen kredi, piyasa, likidite, döviz kuru, faiz oranı ve operasyonel risk türlerinin tanımlarına, bankacılık için ne anlam ifade ettiğine, bankaların bu risklerle nasıl başa çıktıklarına ve faaliyetleri üzerindeki etkilerine yer verilmiştir.

1.6.1. Bankacılıkta Risk Kavramı

Tüm bankaların tek ortak amacı, banka bünyesinde bulunan sermayeyi kullanarak yüksek getiri sağlamaktır. Bankalar bu sermayenin çoğunu, birçok bireysel mevduat sahibinin tasarruflarını kredi finansmanına dönüştürerek elde etmektedir. Bankalar mevduat ve yatırım fonlarına sahip olanlar ile fona ihtiyacı olanlar arasında aracılık görevini üstlenen kurumlar olup bu aracılık görevi, borç vermek için aldığı mevduatların dönüşümü ile elde ettiği kredi finansmanını fon ihtiyacı olanlara ödünç vermesi faaliyetini kapsar. Bankalar bu faaliyeti yürütürken getirisini maksimize etmesi için kullandığı sermayeyi oluşabilecek risklere karşı da iyi yönetmesi gerekmektedir (Golin ve Delhaise, 2013: 87; Heffernan, 2005: 1).

Bankalar modern çağda meydana gelen yeniliklere de uyum sağlayan finansal kuruluşlardır. Globalleşen finansal düzen ile ulusal sınırları aşır uluslararası alanda da faaliyet gösteren bankalar, yapılan işlemlerin nitelik ve çeşitlerinin değişmesi ile döviz gibi farklı finansal araçları da kendi bünyelerine dâhil ederek faaliyet göstermeye başlamışlardır. Bununla beraber artan fon taleplerinin karşılanması ve karşılanan fon taleplerinden elde edeceği faiz gelirleriyle özkaynaklarını arttırma hedefinin yanı sıra risk kapsamalarını da genişletmişlerdir (Ulusoy, 2012b: 358-359).

Piyasalarda artan rekabet ortamı ve fon temini konusunda bir dar boğazın oluşması, bankaları dövizli işlemler yapmaya itmştir. Bankaların artan dövizli işlemleri sonucunda değişen kur oranları, bankaların büyük zararlara uğramasına neden olmuştur. Ayrıca fon yetersizliğinde kaynaklı dar boğazı gidermek için kullanılan dövizler, beraberinde likidite ve faiz risklerinin artışına sebep olmuştur. Milli ve uluslararası faktörler nedeniyle ortaya çıkan siyasi olaylar ve ekonomik problemler, tasarruflar konusunda yetersiz olan ülkelerde faiz oranlarının artmasına sebep olabilir. Bu duruma hazırlıklı olmayan banka bilançoları yaşanacak problemlere karşı dirençli değilse, büyük risklere maruz kalacaktır. Risk durumlarını kontrol altına almak için bankalarca likidite ve kaynakların kontrollü yönetilmesinin yanı sıra, bilançolarındaki aktif ve pasiflerin vadelerinin uyumu konusunda sağlam politikalar belirlemesinin önemi de artmıştır (Babuşçu, 1997: 60; Erdem, 2008: 193).

Bunun yanı sıra, bankalar karşılaştacağı riskler ve bu risklerden doğacak olumsuzluklara karşı fon kullandırma konusunda büyük bir hassasiyetle hareket etseler de bazen fon sağladığı gerçek ve tüzel kişilerden aynı hassasiyeti göremeyebilir.

Bankaya, aslına uygun olmayan bilgi veya belgelerin ibraz edilmesi ve fonun kullandırıldığı andan itibaren meydana gelebilecek ekonomik dalgalanmalar sonucunda borçlunun başına gelen tehlikeleri kendilerine bildirilmemesi, bankanın kullandığı fonun geri ödenmemesi riskini de artırır (Erdem, 2008: 193).

Risk herkes için farklı anlamlar ifade etse de genel olarak risk, bir dezavantajın ortaya çıkması veya beklenen avantajın belli bir davranış sonucu gerçekleşmeme olasılığını ifade eder. Finansal açıdan risk, gelecekte menfaate paralel olarak değişmesi beklenen faiz, kur oranları gibi ekonomik değerlerin, beklenen değerlerinden negatif sapması olarak tanımlanabilir. Diğer yandan finansal açıdan risk, reel kazanç ile gerçekleşmesi beklenen kazanç arasındaki uyumsuzluk ve gelecekte piyasaların içerisinde bulunan likidite dolaşımında meydana gelebilecek değişkenliklerdir (Büker, Aşıkoğlu ve Sevil, 2015: 66).

Bankacılık açısından risk kavramı ise bankacılık sisteminin, mevcut “*mali, stratejik ve operasyonel hedeflerine*” ulaşmalarını sektöre uğratacak bir olayın gerçekleşmesi olasılığı (IIA, 2009: 7) ve ek kazanç veya beklenmeyen zarar şeklinde ortaya çıkan hedeften sapma değeri olarak ifade edilir. BDDK ise risk kavramını, “*Bir işleme ya da faaliyete ilişkin bir parasal kaybın ortaya çıkması veya bir giderin ya da zararın oluşması halinde ekonomik faydanın azalması ihtimali*” şeklinde ifade eder (BDDK, 2006).

Literatürdeki risk kavramlarını incelediğimizde ise risk, bir yatırımdan para kaybetme şansını ve kazançlardaki değişimi ifade eder (Shim ve Constanas, 2001: 250). Dahası, işlerin yolunda gitmeme olasılığı ve bir olaydan kaynaklanan kaybın büyüklüğü (Apostolik, Went ve Donohuse, 2009: 12) ve gelecekteki olasılıklara göre aktif olarak değerlendirilen tehlikelerin tamamıdır (Giddens, 2002: 22).

1.6.2. Bankacılıkta Risk Yönetiminin Amacı

Değişen dünya düzeni ile iç ve dış faktörlerin ekonomi üzerinde daha etkin olması, aniden meydana gelebilecek siyasi belirsizlikler ve ulusal ya da uluslararası çaptaki kargaşalar, para piyasalarının olumsuz etkilenmesine sebep olmuş ve etkileri bankacılık sektörlerine de yansımıştır. Olumsuzluklardan kaynaklanan zararlar bankalar için risk durumunu temsil ederken, risklerden olabildiğince daha az etkilenmek ve hem ödünç verdikleri mevduatlardan gelecek kârı hem de öz

sermayelerini güvence altına almak istemeleri risk yönetim politikaları belirlemelerine neden olmuştur.

Her alanda risk yönetiminin amacı, riski önceden belirlemenin oluşmasını beklemekten çok daha iyi olacağı düşüncesiyle, gelecekte sorun teşkil edebilecek riskleri belirlemek, incelenen sistemler üzerindeki risk olasılıklarını ölçmek ve olası riskleri kontrol etmek ve izlemektir. Bankacılıkta risk yönetiminin odak noktası hem banka içerisinde hem de para piyasalarında, gelecekte meydana gelebilecek önemli gelişmeleri öngörümleme ve riskli gelişmeleri mümkün olduğunca erken zamanda tanımlama, değerlendirme ve müdahale etmesidir. Bankanın aslında amaçladığı, yatırım yapacağı alanı gözlemleyerek hem kendisini etkileyen ve sürekli değişen para piyasalarında uzun vadeli adaptasyonun sağlamak hem de sahip olduğu sermayenin güvenliğini korumaktır.

1.6.3. Bankacılıkta Risk Çeşitleri

Bankalar günümüzde sadece sermaye açığı ve sermaye fazlası olanlar arasında finansal aracılık yapmakla kalmamaktadır. Aynı zamanda yatırım portföylerini, müşterilerin mevduatlarını ve kendi öz sermayelerini piyasa dalgalanmalarından kaynaklı oluşacak riskli durumlara karşı korumak ve bu riskleri fırsata çevirecek girişimlerde bulunarak hem müşterilerinin sabit getiri sağlamak hem de varlıklarını sürdürebilmek ve kar elde etme açısından önemli bir görev üstlenmektedirler. Günün her saati finansal piyasalarda işlem yapabilme kolaylığı sebebiyle piyasadaki dengelerin çabucak değişebilmesine karşın bankalar, yaşanabilecek gelişmelerin takibini ve değerlendirmesini yaparak oluşabilecek olumlu gelişmeleri fırsata çevirmeye çalışır. Olumsuz gelişmelerde ise yapacağı erken müdahaleler ile yaşanacak parasal ve varlıksal kaybını mümkün mertebe aza indirmeye çalışır. Finansal riskler, yatırım ve ticari gibi finansal faaliyetlerin gelecekte yaşayacağı mali kayıpların olasılığını ifade eder. Birçok finansal hizmet ve işlemin doğasında bir kaybetme riski vardır. Bu bağlamda banka ve diğer finansal kuruluşlar riskleri kredi, piyasa, likidite, faiz oranı, döviz kuru, operasyonel olarak sınıflandırma ihtiyacı duymuşlardır.

1.6.3.1. Kredi Riski

Kredi riski, bir bankanın fon sağladığı gerçek ve tüzel kişilerle yaptığı sözleşmeden doğan yükümlülüklerini yerine getirememe potansiyeli ve bununla ilişkili olarak meydana gelebilecek zararları ifade eder. Bankalar müşterilerine kullandırdıkları fonları, belirlenen süre içerisinde, verilen anaparaya ek faiz getirisiyle geri ödenmesini bekler. Ancak ödemelerde yaşanacak olası gecikmeler ya da borcun tamamının geri ödenmemesi bankaları büyük sıkıntılara sokabilir. Bu durum aynı zamanda temerrüt riski olarak da bilinmektedir. Borçlarını geri ödemeyen ya da ödemelerini geciktiren müşteriler bankalar nezdinde temerrüte düşer. Bunun sonucunda gelecekte doğabilecek finansal ihtiyaçlarını temin etmede büyük zorluklar yaşamasına neden olur. Buna ek olarak, müşterinin kredi itibarının zedelenmesi sebebiyle borç yükümlülüklerinin yeniden yapılandırılması ve müşterinin iflas etmesi durumunda belirlenen yükümlülüklerin yerine getirilememesi, müşteriye temerrüte düşüren diğer olaylar arasında yer alır (Apostolik ve diğerleri, 2009: 14; Bandyopadhyay, 2016: 1; Bessis, 2002: 13; Brown ve Moles, 2016: I/2).

Kredi risklerinin ortaya çıkmasının iki sebebi vardır. Bunlardan ilki, bankaların krediyi ileride borcunu geri ödememe olasılığının yüksek olduğu müşterilerine vermesi (ters seçim) ve müşterinin aldığı krediyi bankaya bildirilen amaç dışında, bankanın onaylamayacağı amaçlar için kullandığının (ahlaki tehlike) tespit edilememesine bağlı içsel sebeplerden kaynaklanabilir (Fidan, 2011: 41). İkincisi ise, finansal ve bankacılık sektörlerinde meydana gelen istikrarsızlık ve artan enflasyon, durgunluk, zayıf döviz kurları gibi ekonomik istikrarsızlıklar ve artan rekabet bağlı olarak iş döngüsünde meydana gelen değişiklikler gibi dışsal sebeplerden kaynaklanabilir (Erdem, 2008: 200; Bandyopadhyay, 2016: 2-3).

Tüm kurumlar veya kişiler, isteyerek ya da istemeyerek riske maruz kalabilir. Özellikle bankacılık sektörünün maruz kaldığı risklerin en başında ise kredi riski gelir. Bugün dünya genelindeki bankaların tamamı, muzdarip oldukları kredi riskine karşın mücadele edebilme konusunda zaman içerisinde büyük tecrübeler edinmiştir. Bankalar, edindikleri bu tecrübeler sayesinde kendi risklerini; tanımlama, ölçme, izleme ve kontrol etmenin yanı sıra olası risklere karşı, riski göğüsleyecek yeterli sermayenin olup olmadığını tespit etme konusunda farkındalığa sahip olma gibi

kazanımlar elde ederek riskin en iyi şekilde yönetilmesine yardımcı olurlar (Bandyopadhyay, 2016: 1).

Kredi riskinin oluşmasındaki en önemli faktörlerden biri de bankaların potansiyel müşteri seçim sürecini doğru yönetememesidir. Kredi kullandırmaya hazırlanan bankalar, kredilerini kullandıracakları potansiyel müşterilerin mali tablolarına, müşterinin talep ettiği krediyi ödemede zorlanıp zorlanmayacağını tespitine, gösterdiği teminatlara, müşterinin ödemeye hevesli olup olmadığının tespiti gibi bir dizi analiz ve incelemelerde bulunur. Bu sorulara aradığı cevapların tamamı bankanın kredilendirme prensiplerini/ilkeleri yani kredilendirmenin 5K'sını oluşturur (Aras, 1996: 49; Berk, 2014: 199; Şenel, 2020: 813). Eğer müşterinin bu prensiplerin/ilkelerin sağlamadığı doğru bir şekilde tespit edilemezse işte o zaman bankaların kredi riskine maruz kalması kaçınılmazdır.

Kredilendirmenin 5K'sı (İngilizcede 5C) karakter, kapasite, kapital, koşullar ve kefalettir. Bu ilkelerden aşağıda kısaca bahsedilmiştir (Torun, 2013: 55).

Karakter (Character): Bankaların, krediyi kullandıracakları potansiyel müşterilerinin geri ödemede gösterebileceği tutumların tespitini yaptığı ilkedir. Bankalar için karakter ilkesi, müşterinin saygınlığı, ödeme takvimine bağlılığı ve koşulsuz şartsız her konuda dürüst ve net olması unsurlarını içermektedir. Potansiyel müşterinin bilgi asimetrisi yaratmayacak şekilde bankalara ilgili belge ve bilgileri ibraz etmesi gerekmektedir. Dahası, müşteriyle yapılacak kredi anlaşması yükümlülüklerinde yer alan, vadesi gelen faiz ödemelerini düzenli yapabileceğinin tespiti konusunda da geçmiş kredi ödemelerindeki sicillerine bakılır. Potansiyel müşterilerinin sicillerinin bozuk olması ve bankaların karakter unsurlarına karşı bir aykırılık tespit edilmesi durumunda, diğer ilkelere uygun olup olmadığı kontrol edilmeden kredi talepleri reddedilir.

Kapasite (Capacity): Bankaların potansiyel müşterilerinin talep ettiği krediyi parasal açıdan ödemeyebilmeye kabiliyetinin tespitini yaptığı ilkedir. Bankalar kredilendirme sürecinde potansiyel müşterinin karakter unsurlarını taşımasının tespitine ek olarak mali açıdan gücünün yeterli olmasına da dikkat eder. Bu açıdan karakter ve kapasite ilkeleri kredilendirme sürecinin ayrılmaz ikilisidir. Verilecek krediler özellikle ticari faaliyetler için kullanılacaksa bankalar, kredilerin bu ticari faaliyetlerden elde edilen kazançlarla ödenmesine önem verir. Bu sebeple kredi

vereceği işletme ve firmaların yürüttüğü ticari faaliyetlerden yüksek kazançlar elde etme potansiyeline sahip olup olmadığına da dikkat ederler.

Kapital (Capital): Bir diğer adı sermaye ilkesi olan bu ilkenin, bankaların potansiyel müşterileri olan firma veya işletmelerin yeterli öz sermaye kaynaklarına sahip olup olmadığının tespitini yaptığı ilkedir. Firma veya işletmelerin ortaklarının sağladığı likidite ve ticari faaliyetlerden elde ettikleri kazançlarını kapsayan öz sermayeleri ne kadar yüksek olursa kredi riski o kadar düşük olur. Bu nedenle öz sermayesi yüksek firma ve işletmelerin bankalar tarafından kredilendirilme olasılığı da artış gösterir.

Koşullar (Conditions): Bankaların potansiyel müşterilerinin ticari olarak faaliyet gösterdiği alanların, ekonomide meydana gelebilecek dalgalanmalardan olumsuz etkilenmeye müsait olup olmadığının tespitini yaptığı ilkedir. Faiz oranı, döviz kuru değişimleri ve hükümet kararları ticari faaliyet gösteren firmaları işletmeleri etkileyen dış olaylardır. Bu olaylar firma ve işletme tekelinde olmadığından meydana gelebilecek finansal zararları engellemek zordur. Bu finansal zararlar sebebiyle firma ve işletmeler kredilerini ödemekte zorlanabilir. Bankalar bu nedenle ticari faaliyetlerinde hızlı gelişim gösteren firma ve işletmelere vereceği kredilerin riskinin düşük olduğunu varsayar.

Kefalet (Collateral): Bankaların potansiyel müşterilerinin talep ettiği krediyi ödeyememesi durumunda krediyi tasfiye edebileceği teminatların olup olmamasının tespitini yaptığı ilkedir. Bankalar kredilerin geri ödemesindeki yaşanacak problemleri azaltmak adına müşterilerinden maddi teminat ve kişisel teminat ister. Maddi güvence araba, ev, iş ekipmanları gibi haciz ve ipotek edilebilecek varlıkları kapsarken, kişisel güvenceler ise müşterinin kefil gösterdiği gerçek kişileri kapsar. Maddi ve kişisel güvenceler bankalar tarafından kabul edilmeden önce varlıkların daha önceden oluşmuş haciz ve ipotek durumlarının olup olmadığının yanı sıra, kefil gösterilen kişinin maddi durumu ve itibarı da detaylıca araştırılır.

Diğer yandan, kredilendirmenin 5K'sı olarak geçen bu ilkelere ilave olarak Yazıcı (2010) çalışmasına “karlılık” ve “kreditinin amacı” ilkelerini, Şakar (2011) ise “yasalara uyum (Compliance)” ve “kredi dokümanları (credit file/documentation)” ilkelerini de ilave etmiştir.

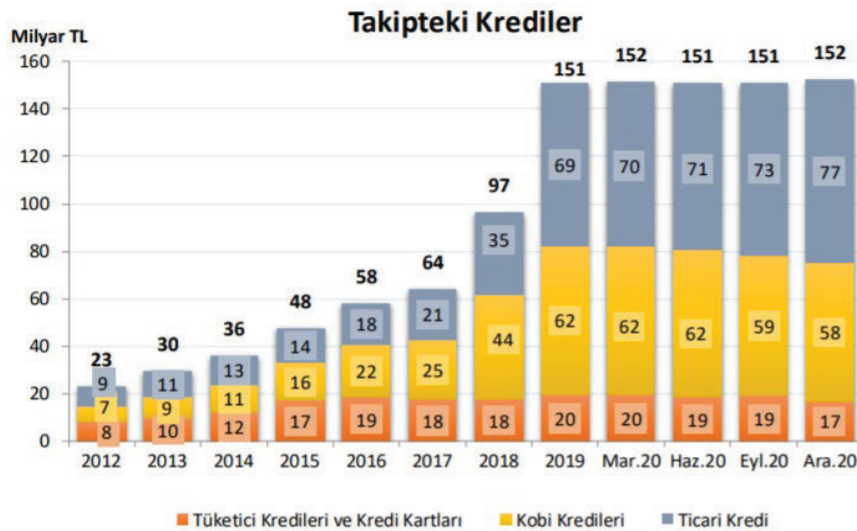
Kredi riski, aynı zamanda sorunlu krediler ve takipteki krediler olarak da ifade edilmektedir. Bir kredinin sorunlu/takipteki krediye dönüşmesi için, gerçek ve tüzel kişiler aldıkları borçları, yapılan kredi anlaşmasında yer alan süre içerisinde bankaya ödenmemesi ya da anlaşmayı aykırı faaliyetlerde bulunması gerekir. Sorunlu/takipteki krediye dönüşme süreleri bireysel ve ticari krediler için 90 gündür (Selimler, 2015: 133).

TBB'nin yayınladığı faaliyet raporuna göre, bankalarca Temmuz-Eylül 2020 dönemi itibariyle idari ve kanuni olarak takibe dönüşen tüketici kredileri ve konut kredilerinin 12,9 milyar TL olarak gerçekleşip, kanuni olarak takibe dönüşen tüketici ve konut kredi miktarı 610 milyon TL olarak gerçekleşmiştir. Söz konusu kredilerin %1'ini taşıt, %7'sini konut, %92'sini ihtiyaç kredileri oluşturmuştur (TBB, 2020: ii-iii).

BDDK'nın yayınladığı faaliyet raporuna göre, bankalarca Aralık 2020'de idari ve kanuni takibe dönüşen kredilerin oranı %4,08 olarak gerçekleşirken, toplam takibe dönüşen kredi tutarı 152 milyar TL olmuştur. Sektörel krediler içerisinde takibe dönüşüm oranları ise inşaat sektöründe %9,29, toptan ticaret ve komisyonculuk sektöründe %5,72 ve enerji ve su kaynakları üretim dağıtım sanayisinde %5,60 olarak gerçeklemiştir (BDDK, 2021a: 14-15).

Diğer yıllara ve dönemlere ait takipteki kredilerdeki değişimler şekil 7'de detaylı olarak gösterilmektedir.

Şekil 7: Takipteki Krediler



Kaynak: BDDK, 2021a: 14

1.6.3.2. Piyasa Riski

Piyasa riski, risk faktörlerinin oluşmasını sağlayan piyasa parametrelerinde meydana gelen olumsuz yöndeki değişimlerden kaynaklı olarak bankanın zarar görme olasılığıdır. Yapılan yatırımlardan beklenen karlar ve borçlardan doğacak alacakların tasfiyesi dâhil olmak üzere piyasada yapılan birçok işlemin faiz, döviz kuru, emtia ve hisse senedi fiyatları ve nakdi değerlerde meydana gelecek oynamalar sebebiyle, bankaların beklentilerinin gerçekleşmemesi ve zarar etme olasılıklarına ilişkin riski ifade eder. Piyasa riski, olumsuz değişimlerden meydana gelen zararların belli bir güven aralığında ölçülmesi temeline dayanan, riske maruz değer (VaR) analizi ile tespit edilir. Piyasa riski, Likidite, faiz, kur ve ülke riskleri olmak üzere dört alt başlık altında incelenir (Apostolik ve diğerleri, 2009: 15; Bessis, 2002: 18; Ulusoy, 2012b: 361; Yazıcı, 2018: 53).

1.6.3.3. Likidite Riski

Likidite riski bankacılıkta en önemli risklerden biridir. Likidite, bir bankanın vadesi gelen yükümlülükleri ciddi zararlara uğramadan karşılayabilmesi olarak tanımlanabilir. Likidite aynı zamanda mevcut müşterilerine ödeme yapabilme yeteneğini temsil ederken, bu ödemeleri zamanında yapabilme gücüne sahip olması veya ödemede yetersiz kalması, tamamen bankanın mevcut nakit rezervleri ve nakde çevrilebilir varlıkları ile ilişkilidir (Apostolik, Donohue ve Went, 2009: 38).

Likidite riski, bir bankanın vadesi geldiğinde, fon giriş ve çıkışı arasında denge kuramadığından, mevduat sahiplerine karşı yükümlülüklerini yerine getirmemesini ifade eder. Bankalar giderlerini karşılayabilmek ve aynı zamanda mevduat sahibi müşterilerinin mevduatlarını sağlamak için belli bir miktar nakit ve kolayca paraya çevirebileceği varlığı ellerinin altında bulundurmaları. Bankalarca ihtiyaç duyulması halinde varlıklarının gereken sürede ve fiyatta nakde çevrilememesi veya satışında problemlerin yaşanması durumunda oluşacak nakit sıkıntıları, likidite riskine ve hatta bankaların iflasına bile sebep olabilir. Dahası, bankadaki paranın tehlikeye girmesine sebep olabilecek olaylar ve bankanın itibarını kötü etkileyecek söylentilerin çıkması nedeniyle tüm müşterilerin aynı anda bankada bulunan mevduatlarını çekmek istemesinin yanı sıra, büyük miktarda kredi kullandırdığı müşterisinin iflası sonucunda, kullandırdığı fonun tasfiyesinin sağlanamaması yine likidite riskine neden

olabilecek önemli unsurlardandır (Apostolik ve diğerleri, 2009: 38; Arman ve Gürman, 1997: 239; Erdem, 2008: 194; Mishkin ve Eakins, 2015: W-50; Özbilgin, 2012: 89).

1.6.3.4. Faiz Riski

Faiz oranı riski, faiz oranında meydana gelecek hareketlenmelerden kaynaklanan potansiyel değer kayıplarını ifade eder. Tüketiciden yatırımcıya borç verenden borç alana kadar herkes faiz oranında meydana gelebilecek değişimlerden etkilenebilir. Aniden yükselen ve düşen faiz oranları piyasa oynaklığını artırır. Piyasada oluşan bu oynaklık, piyasaya endeksli olan gelir, harcama, maliyet ve yatırım dengesinde de büyük değişimlere sebep olur. Faiz oranı yükseldiğinde; borç alan borçtan kaynaklı olarak yüksek maliyetlere maruz kalırken, tüketicinin ise daha fazla tasarruf daha az harcama yapmasına neden olur. Faiz oranının düştüğünde ise borç verenin geliri azalırken, tüketicinin ise daha az tasarruf daha fazla harcama yapmasına neden olur (Apostolik, Donohue ve Went, 2009: 38; Bessis, 2002: 17-18).

BDDK ise faiz oranını riskini “*Faiz oranlarındaki hareketler nedeniyle bankanın alım satım hesapları içerisinde yer alan finansal araçlara ilişkin pozisyon durumuna bağlı olarak maruz kalabileceği zarar olasılığı*” olarak tanımlamıştır. Türk bankacılığında faize endeksli türev araçları pek yaygın olmadığından bilançoları faiz riskine karşı oldukça duyarlıdır (Özün ve Çifter, 2006: 3). Bilançoları açısından faiz oranı riskinden bahsetmek gerekirse, bankanın bilançosundaki kalemleri oluşturan varlıkların faiz oranındaki değişimlere bağlı olarak getirilerinde meydana gelecek olası kayıpları ifade edilebilir. Bankalar kullandığı fonlardan aldığı faiz gelirleri ve müşterilerden topladığı mevduatlar bankanın en önemli finansal kaynaklarıdır. Bankalar kullandıkları fonlardan elde ettikleri gelirlerin bir kısmını giderlerine harcarken, diğer kısmıyla da müşterilerden topladıkları mevduatların faiz gelirlerini öder. Faiz oranında meydana gelecek olumsuz değişiklikler bankaların ödemelerinin aksamasına ve parasal açıdan büyük kayıplar yaşamasına sebep olur (Apostolik, Donohue ve Went, 2009: 38-39; Bessis, 2002: 18). Faiz oranı riski hisse senetleri, emtia ve tahviller dâhil olmak üzere her türlü piyasayı etkileyebilir.

1.6.3.5. Döviz Kuru Riski

Döviz kuru riski, yabancı para kurlarında meydana gelebilecek dalgalanmalara bağlı olarak bankanın dövizde endeksli bilanço kalemlerinde ve yabancı para birimi ile her türlü emtia alış satış işlemleri gibi farklı finansal yatırım enstrümanlarından kar ya da zarar etme olasılığıdır (Apostolik ve diğerleri, 2009: 167; Yazıcı, 2020: 194).

Dövizde endeksli varlıklarının ve yükümlülüklerinin birbiriyle denkleşmemesi durumunda açık pozisyona düşen bankalar, döviz kurlarında meydana gelecek hareketlenmeler sonucu oluşacak fiyat sapmalarını dengelenmekle yükümlüdürler. Bu dengeyi ise Merkez Bankasının döviz alım-satım işlemleri başlamadan önce ya da işlem yapımı durdurulduktan sonra bildirilen kur değerlerini kullanarak sağlarlar. Dahası, müşterilerin karşılaşacağı kur risklerine karşılık, mağduriyet yaşamamaları adına yatırımlarını farklı finansal araçlara taşıyarak, kur değişimlerinden kaynaklı oluşabilecek her türlü zararı iyileştirme sürecine gider. Kur riskinden korunmak için SWAP başta olmak üzere, opsiyon ve forward işlemleri gibi çeşitli türev araçların yardımıyla bankalar, kur dalgalanmalarından kaynaklı oluşabilecek büyük finansal risklerin üstesinden gelebilir. Buna ek olarak, döviz işlemlerinin yapıldığı bankalarda hem yatırımcıları koruma hem de kendilerini garantiye almak adına siyasi, ulusal ve uluslararası gelişmeleri yakından takip etmesi ve yaşanan gelişmeler sebebiyle müşterilerinin yapacağı dövizli işlemler sırasında yaşanabilecek olumsuzluklara karşın tedbirler alarak riski kontrol altına alabilir (Ulusoy, 2012: 362-363).

1.6.3.6. Operasyonel Risk

Operasyonel risk, bankaların hem kurum içinde yaşanan personel hataları, teknik arızalar, sistemin ve denetimin yetersiz olmasından hem de dış olaylara bağlı olarak yaşanan faaliyet aksamalarından kaynaklı meydana gelebilecek potansiyel zararlarıdır. Yazılımsal hatalar, bilgi kaybı, banka personellerinin yetkilerini aşan ve ahlaki değerlere ters düşen işlere kalkışma faaliyetleri, bankanın operasyonel riskini tetikleyen kurum içi olaylardır. Dolandırıcılık, hırsızlık, terörizm, doğal afetler, siyasi ve uluslararası olaylar ise bankanın operasyonel riskini tetikleyen kurum dışı olaylardır (Apostolik ve diğerleri, 2009: 181; Moosa, 2007: 167; Van Grinsven, 2009: 22).

İKİNCİ BÖLÜM

MAKİNE ÖĞRENMESİ TEKNİKLERİNE GENEL BAKIŞ VE YENİDEN ÖRNEKLEME TEKNİKLERİ

2.1. MAKİNE ÖĞRENMESİ

İnternet, çağımızın bilgiye ulaşmak için kullandığı en yaygın araçlarından biri haline gelmiştir. Dünya genelinde 2020 Nisan ayı itibariyle toplam internet kullanıcı sayısı 4.57 milyar kişiye ulaşmıştır. 2019 Ocak ayı ile kıyaslandığında, bir önceki yıla göre kullanıcı sayısı %6 oranında artış göstermiştir (DOMO, 2021). İnternet kullanımının bu denli yaygın hale gelmesindeki en önemli etkenler, teknolojide yaşanan yenilik ve gelişmelerdir. İnternet kullanımına entegre edilmiş teknolojik cihazların icadı ve çevrimiçi olarak erişim sağlanan platformların yaygınlaşması, internet kullanımına olan ihtiyacı büyük oranda artırmıştır.

Hem internetin yaygınlaşması hem de teknolojik cihazların zaman içindeki gelişimi, metin, mesaj, ses, fotoğraf, sunum ve film gibi birçok bilgi ve kaynağa da erişimi kolaylaştırmıştır. Çevrimiçi işlemde tutun da birçok sosyal medya platformuna kadar tüm bu mecraların oluşturduğu, arama günlükleri ve satın alma kayıtları gibi kayıt altında tutulan kişisel faaliyetlerin tamamı veriyi oluşturur. Eğer tüm bu faaliyetlere ait veriler dijital bir ortamda kayıt altına alınıyorsa, bu tür veriler için ‘Büyük Veri (Big Data)’ kavramını kullanmak hiç de yanlış olmaz. Peki, büyük veri ne anlama gelir?

Büyük veri, çeşitli dijital platformlardan hızlı bir şekilde toplanan bilgilerin, işlenmeden, yine dijital bir ortama depolanması sonucu elde edilen hacmi büyük ve düzensiz veri setleridir. Bu tip veri setlerini geleneksel veri işleme araçları ile anlaşılır hale getirmek oldukça zordur. “Her gün toplamda yaklaşık 2.5 Kentyon byte kadar veri elde edilmektedir.” (DOMO, 2018). Bu nedenle, büyük veriyi işleyebilecek ve aynı zamanda bu veriden anlamlı sonuç çıkarabilecek, alan ve araçlara ihtiyaç duyulur. İşte bu ihtiyaç, veri biliminin ortaya çıkmasına neden olmuştur.

Veri bilimi, büyük veriden, çeşitli makine öğrenmesi temellerini ve algoritmaları kullanarak, problem hakkında anlamlı çıkarımlarda bulunmaya yardımcı

olan bir bilim dalıdır. Veri bilimi, yapay zekâ, makine öğrenmesi ve derin öğrenme olmak üzere, birbirlerini kapsayan üç alt daldan oluşur.

Teknolojik cihaz ve çevrimiçi kullanılan tüm platform üreticilerinin sahip oldukları bu araç veya mecraların, değişen dünya düzeninde, yeniliklere kolayca ayak uyduran ve kullanıcının gerçekten ne istediğini bilen sistemlere dönüştürme sürecindeki en önemli kaynağı veridir. Bu süreç için önceki süreçlerde elde edilen verileri anlaşılabilir bilgilere dönüştürerek, sistemlerde iyileştirme yapılmasına önemli ölçüde yardımcı olur. Ayrıca verilerden yararlı bilgilerin keşfedilmesi, yeni iş alanları yaratılmasına ya da sistemi iyileştirecek önemli bir bilgiye erişilmesini sağlar. Bu bilgi keşif ve analiz sürecinde rol oynayan, en kilit tekniklerden biri de makine öğrenmesidir.

Makine öğrenmesi, geçmişte elde ettiği tecrübeleri kullanarak gelecekteki problemlere, bilgisayar, matematik ve istatistik bilimleri yardımıyla çözümler getiren yapay zekâya ait bir çalışma alanıdır. Makine öğrenmesi kavramı, 1959 yılında Arthur Samuel tarafından keşfedilip, 90'ların başında Tom Mitchell ile popülerlik kazanmıştır.

Makine öğrenmesi, bilgisayara, hiçbir açık programlama yapılmadan sadece girdi olarak sunulan verilerle sağladığı öğrenme deneyimini, bilgiye veya uzmanlığa dönüştürme yetkinliğini kazandırır. Bu çerçevede makineye sunulan girdi, tecrübeyi oluşturan eğitim verisine karşılık gelir.

Makine öğrenmesi, iş ve günlük hayat problem veri setleri üzerinden çıkarımda bulunma sürecini otomatik hale getiren bir araçtır. Geçmiş problemlere dayanarak gelecekteki problemler hakkında güçlü çıkarımlarda bulunur. Ayrıca değişen şartlara kolayca uyum sağlayabilmesi nedeniyle, sürece ait modellerin tekrardan eğitilmesine gerek yoktur. Bu durum, modelleme maliyeti ve zaman açısından da tasarruf sağlar. Dahası, makine öğrenmesi geleneksel programlama ile zor çözülen problemleri çözmede oldukça başarılı bir tekniktir. Bugün birçok kurum, iş geliştirme faaliyetleri için farklı makine öğrenmesi tekniklerinden faydalanır.

Makine öğrenmesinin yapay zekâ ile ilişkili olduğu en önemli nokta, bir makineye öğrenme yeteneği kazandırmak ve öğrendiklerini tanıyabilmesidir. Bebeklikten yetişkinliğe kadar her insan hayatının büyük kısmını bir şeyler öğrenmeye adanmıştır. Bir öğrencinin okula başladığı ilk andan itibaren harflerin öğretilmesi ve

öğretilen harfler sayesinde, okuma ve yazmayı sökebilmesi aslında öğrendiği harfleri tanınması ile ilişkilidir. Çünkü öğrencinin harfleri tanınması, ona, önüne çıkacak kelimeleri hem okuma hem de yazma yeteneği kazandıracaktır. İşte bu örnek, öğrenme sürecini ve bir çıktı elde etmeyi anlatan en iyi süreçtir. Bu süreçte makine, öğrenciyi temsil eder. Yapay zekâ sayesinde elde ettiği öğrenme yetisi ile öğrenmesi için sunulan girdiler (harf veya kelime) ile öğrenen ve bunları çıktıya (kelimeleri okuma ve yazma) dönüştürebilen makineler elde edebilmek, makine öğrenmesinin aslında temel çalışma prensibini yansıtır.

Makine öğrenmesinin amacı, tahmin hatasını en aza indirmektir. Ayrıca mevcut model kalıbını başka veri setleri için genellemek veya probleme ait veri setleriyle yeni modelleme kuralları oluşturmaktır (Dangeti, 2017: 8). Makine öğrenmesinin görevleri ise şunlardır;

- İş süreçlerini optimize etme,
- Geçmiş verilere dayanarak, gelecek hakkında tahminlerde bulunma,
- Veri setindeki grup ve küme yapılarını tespit etme,
- Herhangi bir bilgi kaybına sebep olmadan, işlem hızı ve kolaylığı için verilerin boyutlarını indirgeme,
- Probleme ait belirli olayların gerçekleşme olasılıklarını hesaplamadır.

Bir sonraki kısımda makine öğrenmesinin avantajlarından ve kullanım alanlarından bahsedilmiştir.

2.1.1. Makine Öğrenmesi Avantajları ve Kullanım Alanları

Makine öğrenmesi, veriyi öğrendiği ile kalmaz bu veriden önemli bilgileri keşfederek, probleme çözüm üretme sürecinin optimize edilmesine yardımcı olur. Veri setinden keşfettiği farklı desenlerden anlamlı sonuçlar çıkaran makine öğrenmesi, kurum ve firmaların operasyonel süreçlerini geliştiren ve otomatikleştiren sistemler kazanmasını sağlar.

Makine öğrenmesi, sesli yanıt sistemlerinde de etkin bir şekilde kullanılır. Özellikle, kurumların sahip olduğu çağrı merkezi çalışanlarının iş yükünü hafifletmek ve daha hızlı hizmet vermek amacıyla kullandığı bu sistemlerin arka planında farklı

makine öğrenmesi teknikleri kullanılır. Sesli yanıt sistemleri, müşterilerle otomatik olarak konuşan sohbet botlarından oluşur. Bu sohbet botlarının optimizasyonu ise müşterilere ait farklı ses tonlarının eğitilmesiyle sağlanır. Sonuç olarak, makine öğrenmesi, müşterilerden elde edilen sesler sayesinde, sesli yanıt sistemlerine ait sohbet botlarının bilişsel yeteneklerinin geliştirilmesine yardımcı olur.

Makine öğrenmesi, birçok ticari şirketlerin kazançlarını arttırmasına da yardımcı olur. Ticari şirketler elde ettiği veriler ışığında, sadece müşteri memnuniyetini sağlayacak faaliyetlerde bulunmaz, aynı zamanda maliyetleri düşürerek kazançlarını da artırırlar. Dahası, ticari şirketler, makine öğrenmesi sayesinde müşterilerin istek ve ihtiyaçlarını değerlendirebilir. Her bir müşterisi için farklı pazarlama taktikleri geliştirebilir. Bu durum, hem müşterinin istediği şeye hızlıca ulaşmasına yardımcı olacak hem de müşterilere, psikolojik olarak önemsendiğini hissettirecek ve şirkete olan bağlılıklarının artmasını sağlayacaktır. Makine öğrenmesi sayesinde ticari şirketler, bünyelerinden ayrılmaya müsait müşterilerini tespit edebilirler. Bu durumun tespiti ise kurum bünyesinden daha önce ayrılmış müşterilerinin tutulan kayıtlarıyla mümkündür. Ticari şirketler, ayrılacak olan ve ayrılmış müşterilerin ortak özelliklerini analiz ederek, ayrılma riski yüksek müşterilerini tespit eder. Sonrasında, bu müşterileri elinde tutmak için farklı stratejiler geliştirmesine yardımcı olur. Bugün Amazon ve Netflix dahil birçok şirket, makine öğrenmesi tekniklerini, kişiselleştirilmiş kullanıcı önerileri için etkin bir şekilde kullanır.

Makine öğrenmesi, yüz tanıma sistemlerinin optimize edilmesindeki en önemli araçtır. Yüz tanıma sistemleri, günümüzde akıllı telefonların ekran kilitlerinde ve sosyal medya platformlarından olan Instagram ve Snapchat'e ait fotoğraf filtrelerinde yaygın olarak kullanılır. Ayrıca dünyada en popüler sosyal medya platformlarından biri olan Facebook ise yüklenen resimlerdeki kişilerin yüzlerini otomatik olarak tanımlayarak, tanımladığı kişiyi yine otomatik olarak fotoğrafa etiketleme özelliğine sahiptir. İşte tüm bu özelliklerin arkasında makine öğrenmesi teknikleri vardır. Bugün en büyük veri tabanına sahip olan Facebook, yüz tanıma ile ilgili makine öğrenme sistemlerine farklı görsel veriler sağlayarak, bu sistemlerin performans açısından iyileşmesine yardımcı olur.

Makine öğrenmesi, çalışma ortamlarında, çalışanların yükünü hafifletir ve daha verimli çalışmalarını sağlar. Özellikle belgelerin sıralanması ve dosyalanması gibi sorumlulukları bilgisayara devretmek için makine öğrenmesine başvurulur. Bu sayede, başa çıkılması zor ve külfetli iş yükü azaltılmış olur.

Makine öğrenmesi, bugün birçok sanayi iş kolu tarafından da kullanılır. Özellikle imalat sanayisinde kullanılan makinelere, yapılacak işi otomatik olarak gerçekleştirebilmesi için makine öğrenmesi kullanılır. Ayrıca üretim hattında meydana gelebilecek, çözülmesi güç olan sistemsel hataları önceden tespit eder ve onları çözer. Böylelikle, fabrikanın üretimi hiç kesintiye uğramaz. Bu da sürekli olarak ve en az hatayla, üretim yapılmasına olanak tanır. Bu sayede, sanayi sektöründeki firmalara, maliyetlerini azaltma konusunda büyük avantaj sağlar.

Tüm bu bahsedilen alanlar dışında, makine öğrenmesi teknikleri farklı birçok alanda da işleri kolaylaştırmak için etkin olarak kullanılmaya devam edilmektedir. Makine öğrenmesinin yer aldığı diğer belli başlı alanlardan bazıları ise şunlardır (Lantz, 2015: 5);

- E-postada istenmeyen (spam) mesajların filtrelenmesinde,
- Müşterilerin kategorize edilmesinde (Segmentasyon),
- Hava ve uzun vadeli iklim değişikliklerinin tahminlenmesinde,
- Birçok hava olayı ve doğal afetten kaynaklanan mali hasarın aktüeryal tahmininde,
- Hastalıklarla bağlantılı genetik dizilimlerin keşfinde,
- Kredi kartı sahtekarlığı gibi bazı siber suçların engellenmesinde,
- Sürücüsüz araçların algoritmalarının geliştirilmesinde,
- Evde ve ofiste kullanılan enerjinin optimizasyonunda,
- Arama motorlarının geliştirilmesinde,
- Borsa ve kredi risk analizinde kullanılır.

2.1.2. Makine Öğrenmesi Türleri

Makine öğrenmesi türleri; denetimli öğrenme, denetimsiz öğrenme, yarı denetimli öğrenme ve pekiştirmeli öğrenme olmak üzere dört alt başlıkta incelenir.

2.1.2.1. Denetimli Öğrenme

Denetimli öğrenme, denetimi yapacak olan öğrenme algoritmalarından, istenilen çıktıların elde edilmesini sağlayan bir makine öğrenmesi tekniğidir. Denetimli öğrenmeyi diğer öğrenme tekniklerinden ayıran en önemli özellik, kullanılan eğitim verilerinin etiketlenmiş verilerden oluşmasıdır. Bu süreçte hedeflenen, öğrenme algoritmasını geçmişte toplanan verilerle eğiterek, daha önce karşılaşmadığı veri setleri üzerinde de etkin sonuçlar vermesini sağlamaktır. Başka bir deyişle, algoritmayı, daha önce tahmin edilmemiş tüm veri setleri için genellenmiş ve istenilen çıktıları tahmin eden bir otomatik sisteme dönüştürmektir. Denetimli öğrenme tekniği iki temel problem türüne ayrılır; sınıflandırma ve regresyon.

Sınıflandırma problemleri, eğitilmiş denetimli öğrenme algoritmasına verilen yeni girdilerden elde edilen çıktıların, kategorik değer olarak tahmin edildiği problem türleridir. Örneğin, mevcut e-postaların “spam veya spam değil” şeklinde tahmin edilmesi problemleri veya bankaların müşterilerini kredilendirme sürecinde, müşterilerden toplanan bilgiler (yaş, gelir, önceki kredi borçlarını ödeme durumları gibi) doğrultusunda çıktıları “kredi alabilir veya alamaz” şeklinde kategorik olarak tahmin edilebilen problemlerden oluşur. Kredi kartı sahtekârlığında “dolandırıcı veya dolandırıcı değil” probleminin yanı sıra bir tümörün, “iyi huylu veya kötü huylu” teşhisi de yine sınıflandırma problemlerine girer.

Regresyon problemleri, eğitilmiş denetimli öğrenme algoritmasına verilen yeni girdilerden elde edilen çıktıların sayısal değer olarak tahmin edildiği problem türleridir. Hisse senedi fiyatı, gayrimenkul fiyatları ve araba fiyatlarının tahmin edilmesi regresyon problemlerinden bazılarıdır.

2.1.2.2. Denetimsiz Öğrenme

Denetimsiz öğrenme, sadece girdilerden oluşan eğitim veri setinin yapısı keşfedilerek, analiz edilecek test verilerinden anlamlı sonuçlar tahmin etmeye çalışan bir makine öğrenmesi tekniğidir. Veri yapısının keşfi, girdi değerlerine kümeleme (clustering) veya ilişkilendirme (association) tekniklerinden biri uygulanarak yapılır. Daha sonra aralarında benzerlik keşfedilen gözlemler bir araya getirilir ve etiketlenir. Denetimsiz öğrenmede test edilen veri setindeki gözlemlerin tamamı etiketsizdir. Bir başka deyişle, veri seti, bir çıktı (yanıt) değişkenine sahip değildir. Etiketsiz verilere

en iyi örnek, sosyal medya uygulamalarından elde edilen verilerdir. Ayrıca e-ticaret sitelerinin benzer ürünleri beğenen müşterileri kategorize edilmesinde ve yüz tanıma sistemlerinin geliştirilmesinde denetimsiz makine öğrenmesi algoritmaları kullanılır. En temel algoritmaları, K-ortalamlar kümelemesi (k-means clustering) ve temel bileşenler analizi (PCA) 'dır.

2.1.2.3. Yarı Denetimli Öğrenme

Yarı denetimli öğrenme, denetimli ve denetimsiz öğrenme tekniklerinin birleşimi olan bir makine öğrenmesi tekniğidir. Eğitim verisi hem etiketli hem de etiketsiz gözlemlerden oluşur. Ancak eğitim veri setinde yer alan etiketli gözlemler, etiketsiz gözlemlere göre nispeten daha azdır. Algoritma, eğitim verilerindeki etiketli gözlemlerden öğrendiklerini, etiketsiz gözlemleri etiketlemek için kullanır. Bu tekniğin amacı, etiketlenen gözlemleri, tekrarlı olarak eğitim verisine dâhil ederek algoritmanın tahmin performansını iyileştirmektir.

2.1.2.4. Pekiştirmeli Öğrenme

Temsilcinin (ajan), eğitim verisi olmadan, yalnızca ortamda belirlenen politikaları (eylem) uygulayarak aldığı geri bildirimlere (ödül veya ceza) göre hareket eden bir makine öğrenmesi tekniğidir. Temsilcinin amacı, geri bildirimleri değerlendirerek, kendisine yüksek getiri sağlayacak politikanın sırasını öğrenmeye çalışmaktır. Bu makine öğrenmesi tekniği robotik süreçlerde, oyun ve sürücüsüz araç sistemlerinin geliştirilmesinde oldukça etkin bir şekilde kullanılır.

2.2. TEKİL MAKİNE ÖĞRENMESİ TEKNİKLERİ

2.2.1. K-En Yakın Komşu Algoritması

k -En Yakın Komşu (k -NN), denetimli makine öğrenmesi algoritmaları arasında uygulanması en basit olan ve bu sebeple yaygın olarak kullanılan, parametrik olmayan bir algoritmadır (Gutierrez, 2015: 178). Örnek tabanlı öğrenme olarak da bilinen bu algoritma, eğitim aşamasında depolanan örnekleri öğrenmeye değil, ezberlemeye odaklı olduğu için tembel algoritma olarak da bilinir. Bu algoritmanın amacı çıktısı

bilinmeyen gözlemleri, ezberinde tuttuğu eğitim verilerine dayanarak bir sınıfa atar. Algoritmanın etkili sonuç vermesi ancak tahmin yapılacak gözlemin çevresindeki en yakın komşuların seçim parametresine (k değerine) ve gözlemlerle arasındaki mesafenin doğru ölçümüyle ilişkilidir.

Hem sınıflandırma hem de regresyon problemlerinde kullanılan bu algoritmanın, çıktıyı tahmin etmek için izlediği yol aşağıdaki gibidir;

Veri algoritmaya depolanır ve ölçeklendirmesi yapılır: Algoritmanın hafızasına veri depolanır. Depolanan veriye, önce bir veri ön işleme tekniğinden biri olan değişken ölçeklendirme (feature scaling) tekniği uygulanır. Bu değişken ölçeklendirme tekniği, k -en yakın komşu algoritması başta olmak üzere, uzaklık hesaplaması yapan tüm algoritmalar için de oldukça önemli bir veri ön işleme tekniğidir. Girdi değişkenlere (bağımsız değişken) uygulanan bu ölçeklendirme tekniğiyle amaçlanan, veride yer alan gözlem değerlerini birbirine yaklaştırmaktır. Bu sayede gözlem aralığı geniş olan değişkenlerin sonuca olan etkisini azaltılarak, kurulan modellerin daha doğru tahminler yapması sağlanmış olur.

k - en yakın komşu algoritması için özellik ölçeklendirme işlemi, min- max normalleştirme (2.1) ve z-skor standartlaştırma (2.2) olarak iki şekilde yapılabilir (Lantz, 2015: 72-73).

$$X_{(normalleştirme)} = \frac{X - X_{(min)}}{X_{(max)} - X_{(min)}} \quad (2.1)$$

$$X_{(standartlaştırma)} = z = \frac{X - \mu}{\sigma} = \frac{X - Ortalama_{(X)}}{Standart Hata_{(X)}} \quad (2.2)$$

Seçilecek komşu sayısı (k) belirlenir: Komşu sayısı (k)'nın belirlenmesi, seçilen komşuların, bilinmeyen çıktı değerini doğru tahmin etmesi ve modeli gelecek veriler için genelleyebilmesi açısından oldukça önemlidir. k ' ye değer verilirken mutlaka tek sayı olmasına dikkat edilmelidir. Örneğin, bir cinsiyet sınıflandırma probleminde cinsiyetinin tespit edileceği gözlem noktası için komşu sayısı (k)'nın 4

olarak belirlenmesi, seçilen komşuların 2 kadın ve 2 erkekten oluşmasına sebebiyet verebilir. Bu durum algoritmanın, gözlemi hangi sınıfa dâhil edeceğine dair belirsizliğinin devam etmesine neden olur. Bu nedenle k' ye tek sayı vermek, seçilen komşuların çıktı değerleri arasındaki eşitliği bozmada oldukça önemlidir.

k değerini vermek aslında görüldüğü kadar kolay değildir. Çünkü seçilecek komşu sayısının rastgele belirlenmesi, birçok tahminleme ve öğrenme problemine yol açar. k 'ya çok büyük bir değer atanması, gürültülü verilerin tahminlemeye olan etkisini veya varyansı azaltırken, modelin yanlılığının büyük ölçüde artmasına sebebiyet vermektedir (Lantz, 2015: 70). Yanlılık, tahmin edilen değerlerin gerçek değerlerden ne miktarda saptığını ifade eder. Yanlılığın yükselmesi modelin eksik öğrenmesine (underfitting) neden olur. Ayrıca k değerinin aşırı yükselmesi, algoritmayı, bilinmeyen gözlem noktasını en yakın komşuya göre tahmin etmek yerine, hangi çıktı değeri veri içinde çoğunluktaysa bilinmeyen gözlem noktasına o çıktı değerini vermeye iter. Bu sebeple algoritmanın verdiği kararlardan kaynaklı olarak, modelin bilinmeyen gözlem noktalarının hatalı tahmin etme olasılığı yükselir.

k' ya çok küçük bir değer atanması, modelin varyansının artmasına neden olur. Varyansın yükselmesi, algoritmanın eğitim verilerine aşırı bağlı kalmasına ve genelleme özelliğini yitirmesine yol açar. Bu nedenle algoritman bilinmeyen gözlem noktalarının tahmini yaparken tamamen kendini eğitim verilerine teslim etmektedir. Bu da algoritmanın hatalı tahmin yapma oranında artışa sebep olur. Algoritmanın tahminleme yaparken genelleme özelliğini kaybetmesi ve tamamen kendini eğitim verilerine bırakması aşırı öğrenmenin (overfitting) göstergesidir. Örneğin k değeri en küçük değer olan 1 seçildiğinde, algoritma belirlenmemiş gözlem değerine en yakın komşunun sınıf değerini, tahminlenen çıktı değeri olarak alır. Ancak belirlenmemiş gözlem noktası için belki de farklı bir çıktı tahmin edilmesi gerekiyor olabilir. Bu durum algoritmanın yine hatalı tahmin yapmasına neden olur.

Sonuç olarak k parametresi, k -NN için kilit bir parametredir. k parametresi için değer belirlenirken, algoritmanın hatalı karar vermesine sebep olmayacak, dengeli bir değer belirlenmelidir. Genelde $k=5$ değeri, algoritmada tanımlı (default) değerdir. Ancak algoritmanın tahmin performansını iyileştirmek için farklı k değerleri denemekte de fayda vardır.

Bilinmeyen gözlem noktaları ile bütün noktalar arasındaki mesafe ölçülür: k -NN algoritması bilinmeyen gözlem noktaları ile bütün noktalar arasındaki mesafeyi ölçer. Mesafe ölçümündeki amaç, algoritmanın, bilinmeyen gözlem noktalarının mesafe olarak en yakın komşularını tespit etmesidir. k -NN algoritmasının mesafe ölçümü deyince akla ilk gelen dezavantajlarından biri, büyük veri setlerinde başarılı bir performans gösterememesidir. Sebebi, tüm veri noktaları ile bilinmeyen gözlem noktası arasındaki tüm mesafeyi tek tek ölçmek zorundadır. Bu da yavaş bir mesafe hesaplama performansı göstermesine neden olur. Bu işlemi gerçekleştirirken bazı mesafe ölçüm yöntemlerini kullanır. Mesafe ölçüm tekniklerinden en önemlileri; Öklid (2.3), Manhattan (2.4) ve Minkowski'dir (2.5).

$$d_{(ÖKLİD)} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \quad (2.3)$$

$$d_{(MANHATTAN)} = \sum_{i=1}^n |p_i - q_i| \quad (2.4)$$

$$d_{(MINKOWSKI)} = (\sum_{i=1}^n |p_i - q_i|^a)^{\frac{1}{a}} \quad (2.5)$$

Verilerin içerdikleri değişken türlerine göre farklı mesafe ölçüm teknikleri kullanılabilir. Örneğin bir çiçeğin yaprak genişlik ve uzunluk ölçülerini içeren benzer değişkenlere sahip olması durumunda Öklid mesafe ölçüm yöntemi kullanılabilir. Bunun yanı sıra veri seti tür olarak farklı değişkenlere sahipse (boy, kilo, yaş, cinsiyet gibi), Manhattan mesafe ölçüm tekniği oldukça etkili bir yöntemdir (Brownlee, 2016: 100).

Ölçülen mesafe değerlerine göre belirlenen komşu sayısı (k) kadar en yakın gözlemler seçilir: Mesafelerin ölçülmesinden sonra belirlenmemiş gözlem noktası ile diğer tüm noktaların mesafe değerleri karşılaştırılır. Sınıflandırma problemi için, mesafe olarak en yakın komşular arasından tanımlanmış k değeri kadar gözlem noktası seçilir ve sınıflandırılır. Sınıfta çoğunluk olarak hangi çıktı değeri yer alıyorsa, o çıktı, belirlenmemiş gözlem noktasının değeri olarak atanır. Eğer bu regresyon problemi ise seçilen en yakın komşu gözlem noktalarına ait çıktı değerleri toplamalarının ortalaması, belirlenmemiş gözlem noktasının çıktı değeri olarak, algoritma tarafından atanır.

2.2.2. Naïve Bayes Yöntemi

Naive Bayes yöntemi, koşula bağlı olayların gerçekleşme olasılıklarının hesaplanmasına dayalı olarak sınıflandırma yapan bir denetimli makine öğrenmesi yöntemidir. Sınıflandırma aşamasında kullanacağı olasılık değerlerini hesaplamak için Bayes teoremini temel alan basit bir tekniktir. Ancak yöntemin basit olması, etkin çalışmadığı anlamına gelmez. Naive Bayes yöntemi, birçok karmaşık sınıflandırma algoritması ile karşılaştırıldığında, bu algoritmalara yakın ve hatta bazı durumlarda bu algoritmalarından daha iyi sonuçlar verebilir. Günümüzde bu yöntem, e-postaların veya mesajların spam olup olmadığını tespit etmek gibi metin sınıflandırmada işlemlerinde ve gözlenen hastalık belirtilerine göre, kişinin hasta olup olmadığının teşhisinde sıklıkla kullanılabilir (Lantz, 2015: 90).

Naive Bayes yöntemine sunulan eğitim veri setinin gözlemleri ikili (binary) ya da çoklu sınıflardan oluştuğu gibi sayısal değerlerden de oluşabilir. Yöntem, eğitim veri setindeki değişkenlerin her bir gözlem değeri için koşullu olasılık değerlerini hesaplar ve öğrenir. Öğrendiği bu olasılık değerlerini, önceden sınıfı tahmin edilmemiş gözlemleri sınıflandırmak için kullanır. Sınıflandırma aşamasına geçildiğinde etiketlenmemiş gözlem değeri için hesapladığı nihai koşullu olasılık değerlerinden hangisi daha yüksekse o koşulu daha önemli kabul eder. Daha sonra sınıf tahminlemesi yaparken, en yüksek gerçekleşme olasılığına sahip koşullu olayın yanıt değerini, tahmin edilecek olan gözlemlerin yanıt değeri olarak tahmin eder. Yöntemin, öğrendiği koşullara dayanarak sınıflandırma yapması, hızlı ve etkin çalışmasını sağlar. Ayrıca yöntemin öğrendiği veri setinin boyutunun büyük olması, modelin genelleme yeteneğinin artmasına, bu sayede farklı veri setleri için de doğru sınıflandırma yapmasına yardımcı olur.

Naive Bayes sınıflandırma yöntemine ait formül (2.7), Bayes olasılık teoremi formülünden (2.6) türetilmiştir. Formüllere ait ortak bazı parametrelerin açıklamaları ise şunlardır;

- X , eğitim veri setindeki girdilere ait gözlemleri ifade eder.

$$X = (x_1, x_2, x_3, \dots, x_m)$$

- y_i , eğitim veri setindeki yanıt değişkenine (Bağımlı değişken) ait i adet tahmin sınıfını ifade eder.

- $P(y_i|X)$, X olan bir gözlemin y yanıt değişkenin (Bağımlı değişken) i . Sınıf değerine ait olma olasılığını ifade eder (nihai olasılık değeri).
- $P(X|y_i)$, y yanıt değişkenin (Bağımlı değişken) i . Sınıf değerine ait değerin, X gözlemi olma olasılığını ifade eder.
- $P(y_i)$, y yanıt değişkenin (Bağımlı Değişken) i . Sınıf değerine ait ilk olasılık değerini temsil eder.
- $P(X)$, Herhangi bir gözlemin X olma olasılığını temsil eder.

$$P(y_i|X) = \frac{P(X|y_i)P(y_i)}{P(X)} \quad (2.6)$$

Naive Bayes yöntemini “Naive” yapan asıl neden, veri setinde yer alan değişkenlerin birbirinden bağımsız olduğu varsayımdır (Lantz, 2015: 97). Naive Bayes, olasılık hesaplaması yaparken, birbiriyle ilişkili olan değişkenlere bile birbirinden bağımsızlarmış gibi davranır. Yöntem, değişkenlerin etkilerini birbirlerinden ayrı olarak hesaplamaya dâhil ettiği için, herhangi bir değişkende meydana gelebilecek değişimlerden kolay kolay etkilenmez.

Bu çerçevede Naive Bayes yönteminin (2.7) formülündeki, $P(X|y_i)$ koşullu olasılığını hesaplanırken, y yanıt değişkeninin (bağımlı değişken) i . Sınıf değerine ait çıktının gerçekleşmesi durumunda, bu çıktıya karşılık gelecek X 'lere ait gözlem değerlerinin gerçekleşme olasılıklarının ayrı ayrı çarpımlarını hesaplar.

$$\begin{aligned} P(y_i|X) &= P(x_1, x_2, x_3, \dots, x_m|Y_i)P(Y_i) \\ &= \underset{(x_j \in X)}{\operatorname{argmax}} (P(Y_i) \prod_{j=1}^m P(x_j|Y_i)) \end{aligned} \quad (2.7)$$

Naive Bayes yönteminin avantajları ise şunlardır; eğitim ve sınıflandırma süreçleri hızlıdır. Veri setinin içerisinde hatalı veriler veya kayıp veriler olsa bile etkili çalışır. Yöntem, kendini eğitmek için fazla veriye ihtiyaç duymaz. Veri setinde, veri akışından kaynaklı olarak verilerin sürekli değişmesi durumuna algoritma kendini

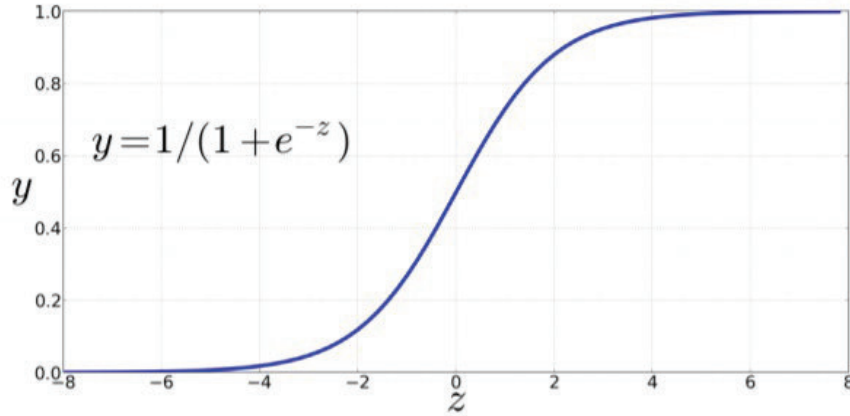
kolayca adapte edebilir. Ayrıca birbiriyle alakasız olan değişkenlere karşı duyarlı değildir. Dezavantajları ise şunlardır; Birbiriyle ilişkili olan değişkenleri bağımsız olarak varsayar. Buna ek olarak kullanılacak veri setindeki girdi değişkenlerine ait gözlem değerleri, kategorik değerlerden çok sayısal değerler içeriyorsa bu tip veri setleri yöntemin çalışma prensibi için pek uygun değildir (Gutierrez, 2015: 173; Lantz, 2015: 97).

2.2.3. Lojistik Regresyon Modeli

Lojistik regresyon, açıklayıcı değişkenlerdeki gözlem değerlerine dayalı olarak yanıt değişkenlerini ikili(binary) sınıf olarak tahmin eden bir modeldir. Üç çeşit lojistik regresyon modeli vardır. Bunlardan ilki, ikili lojistik regresyon modelidir. Bu modelin yanıt değişkenleri ikili (binary) sınıf değerlerinden oluşur. Örneğin: evet/hayır, hasta/hasta değil, var/yok şeklinde sınıf değerleri alırlar. Bazen bu sınıf değerleri 1-0 değerleri şeklinde etiketlenerek yanıt değişken sınıfları oluşturulabilir. Genelde 1 değeri gerçekleşen olayın sınıfı olarak etiketlenir (evet veya hasta gibi). İkincisi, üç veya üçten fazla yanıt değişken sınıf değerlerinden oluşur. Örneğin: Doktor, Öğretmen, Eczacı, Polis gibi mesleklerin tahmin edilmesini içeren sınıflandırma problemlerinde çoklu lojistik regresyon modeli kullanılır. Üçüncüsü ise üç veya üçten fazla sınıfa sahip, özellikle anketlerde kullanılan Likert ölçeği gibi yanıt değişken sınıfı derecelendirme değerlerinden oluşan, sınıflandırma problemlerinde kullanılan sıralı lojistik regresyon modelidir.

Lojistik regresyon, lineer olasılık modeliyle hesaplanan tahmini yanıt değerlerini, ikili olayların gerçekleşme olasılığına dönüştürür. Bu dönüşümü yaparken, lojistik fonksiyonu bir diğer adıyla Sigmoid fonksiyonunu kullanır. Sigmoid fonksiyonu, hesaplanan sürekli değerlerin 0 ile 1 sayıları arasında olasılık değerleri almasını sağlayan bir fonksiyondur. Bu fonksiyon, her bir sürekli değişkenin karşılığına gelecek olasılık değerlerinin birleştiği noktalardan, şekil 8’ de gösterildiği gibi, “S” şeklinde bir eğri oluşturur. Ayrıca “S” eğrisinde, lineer olasılık modelinden elde edilen sürekli değerler $+\infty$ ’a doğru gittiğinde, hesaplanan olasılık değerleri asimptotik olarak 1 değerine yakınsarken, sürekli değerleri $-\infty$ ’a doğru gittiğinde, hesaplanan olasılık değerleri ise asimptotik olarak 0 değerine yakınsar.

Şekil 8: Sigmoid Fonksiyonu (“S” Eğrisi)



Kaynak: Jurafsky ve Martin, 2020: 3

Sigmoid fonksiyonu $-\infty$ ile $+\infty$ aralığında yer alan sürekli tahmin değerlerini olasılık değerlerine dönüştürdüğünden, lineer olasılık modeli aracılığıyla hesaplanan sürekli değerlerin bu aralıklarda yer alması gerekir. Sigmoid fonksiyonu ile elde edilen olasılık değerlerinin kategorik olarak sınıflandırılması için belli bir eşik (treshold) değeri belirlenir. Belirlenen eşik değeri genellikle 0.5'tir. 0.5 değerine eşit veya büyük olan olasılık değerlerinin yanıt değişken sınıfı 1 olarak belirlenirken, 0.5 değerinden küçük olan olasılık değerlerinin yanıt değişken sınıfı 0 olarak belirlenir.

2.7'deki eşitlik, sürekli değerlerin hesaplandığı lineer olasılık modelidir. Sonrasında lineer olasılık modelinden hesaplanan sürekli değerlerin(z), 2.8'deki Sigmoid Fonksiyonu ile olasılık değerlerine(y) dönüşümü sağlanır. En sonunda ise belirlenen eşik değeri altında kalan olasılık değerleri 0, eşik değerinin üstünde kalan olasılık değerleri ise 1 olacak şekilde kategorik yanıt değişkenlerine dönüştürülür.

$$\ln \left[\frac{y}{1-y} \right] = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + u = z \quad (2.7)$$

$$y = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\dots+\beta_n X_n+u)}} = \frac{1}{1+e^{-z}} \quad (2.8)$$

Lojistik regresyonu, Lineer regresyondan ayıran en önemi iki farktan birincisi, sürekli değerlerle değil kategorik değerlerle ilgilenmesidir. İkinci en önemli fark ise

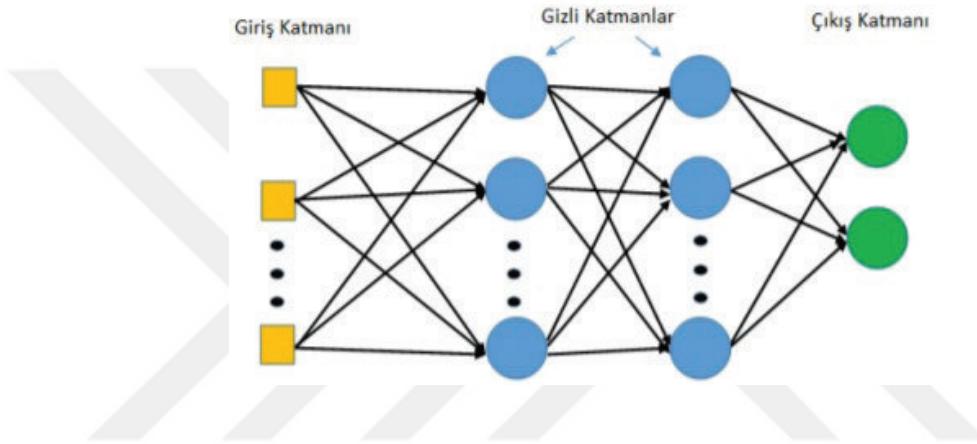
lineer regresyonda açıklayıcı değişkenlerin parametreleri en küçük kareler (EKK) yöntemi ile tahmin edilirken, lojistik regresyonda ise parametreler En Yüksek Olabilirlik(Maximum Likelihood) tahmin yöntemiyle hesaplanır. En Yüksek Olabilirlik tahmin yönteminin amacı, açıklayıcı değişkenlere ait parametre değerlerinin içerisinde olayın gerçekleşme olasılığını en yüksek yapacak, en iyi model parametre değerlerini belirleyebilmektir.

2.2.4. Çok Katmanlı Algılayıcı Sinir Ağları

Tek katmanlı algılayıcılar yalnızca doğrusal sınıflandırma yapabilir. Ancak bazen daha karmaşık sınıflandırma problemleriyle ilgilenmek de gerekebilir. Özellikle XOR problem türünün çözümünde tek katmanlı algılayıcıları kullanarak başarılı sonuçlar elde etmek pek mümkün değildir. Bu nedenle XOR problemlerini çözmek için araştırmalara girilmiştir. XOR probleminin çözümü için araştırmalar yürütüldüğü sırada, 1980'lerin sonlarına doğru Rumelhart ve diğerleri (1986) tarafından geri yayılım (back-propagation) algoritmasının geliştirilmesi, çok katmanlı algılayıcı sinir ağının ortaya çıkmasına olanak tanımıştır. Çok katmanlı algılayıcı (MLP), ileri beslemeli sinir ağının bir tamamlayıcısıdır. Şekil 9'da gösterildiği gibi çok katmanlı algılayıcılar sinir ağı giriş katmanı, çıkış katmanı ve gizli katman olmak üzere üç tür katmandan oluşur. Bir katmanın algılayıcıları, bir sonraki katmanın algılayıcılarıyla ileri yönde birbirlerine bağlanır. Katmanlar arasında döngüsel bir ara bağlantı yoktur ve katmanlardaki algılayıcılar kendi içlerinde birbirlerine bağlı değildir. Ayrıca katmanlar arasında geri yayılım ve ileri yayılım olmak üzere iki tür geçiş yöntemi bulunur. Geri yayılım geçiş türünün uygulanması, hata değerini optimize etmek adına katmanlar arasındaki bağlantı ağırlık değerlerini ayarlamaya yardımcı olur. İleri yayılım geçiş türünün uygulanması ise sinir ağlarının çıktılarının elde edilmesine ve hataların hesaplanmasına yardımcı olur. Giriş katmanı, işlenecek girdileri kabul eden ve bir sonraki katmana ileten kısımdır. Giriş katmanında herhangi bir işlem yerine getirilmemesi nedeniyle, toplam katman sayısı hesaplanırken giriş katmanı dikkate alınmaz. Giriş katmanını bir veya birden fazla gizli katman izler. Giriş ve çıkış katmanı arasına yerleştirilen rastgele sayıda gizli katman ise MLP' nin hesaplama merkezidir. Asıl sınıflandırma işlemi gizli katmanda başlar. Nihai tahmin işlemleri, çıkış katmanı tarafından gerçekleştirilir. MLP'deki nöronlar, geri yayılım (back-propagation)

öğrenme algoritması ile eğitilir. MLP'ler, çeşitli aktivasyon fonksiyonlarının kullanılabileceği şekilde tasarlanmıştır. En çok kullanılan aktivasyon fonksiyonları; doğrusal, sigmoid, tang, hard limiter ve treshold' dur. Ayrıca MLP sinir ağı, doğrusal olarak ayıramayan problemleri çözüme ulaştırmada oldukça başarılı bir performans gösterir. MLP'nin başlıca kullanıldığı durumlar ise; örüntü sınıflandırması, tanıma ve tahminlemedir.

Şekil 9: Çok Katmanlı Algılayıcı Sinir Ağları



Kaynak: Zaccane, 2016: 108

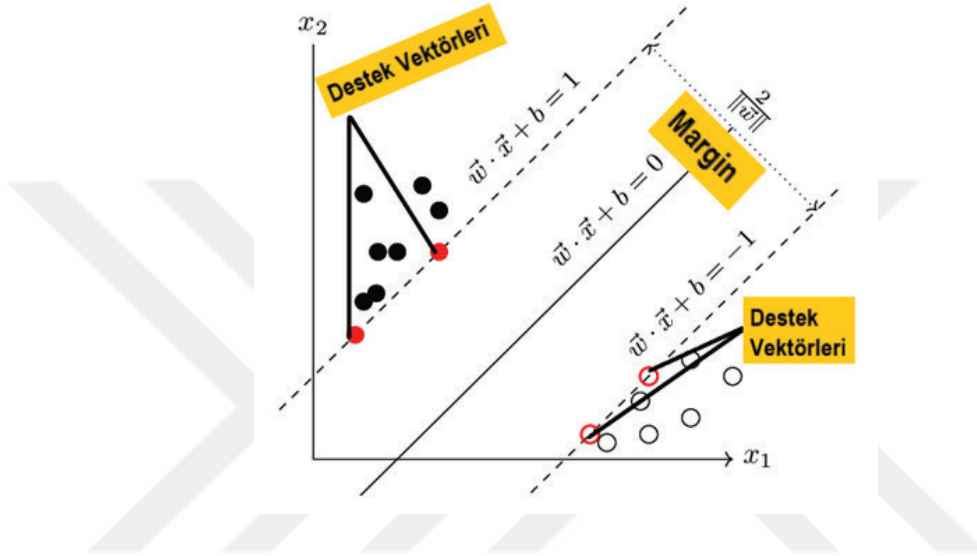
2.2.5. Destek Vektör Makineleri

Destek vektör makineleri veri setinin içerdiği yanıt değişkenine ait sınıfları, belirlenen bir sınır yardımıyla birbirinden ayırabilen denetimli makine öğrenmesi algoritmasıdır. Belirlenen bu sınıra hiper düzlem adı verilir. Hiper düzlemlerin gösterimi veri setinin boyutuna göre değişiklik gösterir. Örneğin; hiper düzlem iki boyutlu veri setinde doğru şeklindeki, üç boyutlu veri setinde ise düz bir yüzey şeklindedir. Algoritmanın belirleyeceği hiper düzlem sayısı birden fazla olabilir. Ancak algoritmanın asıl amacı, bu hiper düzlemler arasından sınıfları doğrusal ve en doğru şekilde ayırtıran optimal hiper düzlemi belirleyebilmektir.

Optimal hiper düzlemin belirlenmesinde ise destek vektörleri ve margin genişliği büyük rol oynar. Destek vektörleri, hiper düzleme en yakın noktalara denir. Her iki sınıfa ait destek vektörleri ile hiper düzlem arasındaki alana ise margin adı

verilir. Destek vektörlerinde meydana gelecek deęişim margin genişliğinin deęişmesine neden olurken, belirlenen destek vektörleri dışındaki dięer noktalardaki deęişimler margin genişliğini etkilemez. Doğrusal sınıflandırmaya ait örnek şekil 10'daki gibidir.

Şekil 10: Destek Vektör Makineleri Doğrusal Sınıflandırma Örneęi



Kaynak: Yang ve dięerleri, 2019: 2

Algoritma, eğitim veri setindeki yanıt deęişkenlerine ait deęerleri, $\vec{w} \cdot \vec{x} + b$ karar fonksiyonuyla -1 veya +1 şeklinde yeniden etiketleyerek, doğrusal olarak ayırabilen bir hiper düzlem belirlemeye çalışır. Ancak bu hiper düzlemi belirlerken, 2.9'daki margin formül deęerinin maksimum olduęu durumda, yani destek vektörleri baz alınarak, sağ ve sol tarafta belirlenen iki sınır çizgisinin (şekil 10'daki kesik çizgili doğrular) arasındaki mesafe maksimum olduęunda, bu çizgilere eşit uzaklıkta bir hiper düzlem belirlenir. Böylece algoritma, gelecekte sınıf etiketi bilinmeyen farklı gözlem noktalarının hangi sınıfa dahil olacağını tahmin etmesi için eğitilmiş olur.

$$= \max \frac{2}{\|\vec{w}\|} \quad (2.9)$$

Etiketlenmemiş gözlem noktasının, algoritma tarafından etiketi tahmin edilmeye çalışırken, 2.10'daki kısıtlardan birini sağlaması gerekir. Etiketlenmemiş

yeni gözlem noktasının karar fonksiyonu değeri +1'e eşit ya da +1'den büyük olduğu durumda algoritma, etiketsiz yeni gözlem noktasını, eğitim sürecinde +1 olarak etiketlediği değerlerin olduğu sınıfa dahil eder. Bu sınıf, şekil 10'daki $\vec{w} \cdot \vec{x} + b = 1$ olan sınır çizgisinin üst tarafında yer alan bölgedir. Karar fonksiyonun -1'e eşit ya da -1'den küçük olması durumunda ise algoritma, etiketsiz yeni gözlem noktasını, eğitim sürecinde -1 olarak etiketlediği değerlerin olduğu sınıfa dahil eder. Bu sınıf ise $\vec{w} \cdot \vec{x} + b = -1$ olan sınır çizgisinin alt tarafında yer alan bölgedir.

$$f(x_i) = \begin{cases} \vec{w} \cdot \vec{x}_i + b \geq +1, & y_i = 1 \\ \vec{w} \cdot \vec{x}_i + b \leq -1, & y_i = -1 \end{cases}$$

ya da

(2.10)

$$y_i * (\vec{w} \cdot \vec{x}_i + b) \geq +1$$

Destek vektör makineleri için önemli olan, margin genişliği yüksek olan ve aynı zamanda doğrusal sınıflandırma yapabilecek bir hiper düzlem belirleyebilmektir. Bu nedenle 2.9'daki margin formül değerinin maksimum olması için paydasında bulunan Öklid normunun ($||\vec{w}||$) minimize edilmesi gerekir. Öklid normu, orijin ile $\vec{w}$ vektörü arasındaki uzaklığı ifade eder. Bu problem, 2.11'deki modelde de ifade edildiği gibi, amaç fonksiyonu minimize edilen ve aynı zamanda sahip olduğu kısıtları doğrusal olan bir kısıtlı optimizasyon problemidir. Bu problemin kısıtında yer alan $\forall$ ifadesi "her $\vec{x}_i$ vektör değerleri için" anlamına gelir (Lantz, 2015: 244).

$$= \min \frac{1}{2} ||\vec{w}||^2$$

(2.11)

$$y_i * (\vec{w} \cdot \vec{x}_i + b) \geq +1, \forall \vec{x}_i$$

$||\vec{w}||$ minimize etme problemi aynı zamanda karesel (kuadratik) programlama problemi olarak da bilinir. Karesel programlama problemi ismini, 2.11'deki amaç fonksiyonunda yer alan değişkenin karesel ifadesinden alır. Bu kısıtlı optimizasyon probleminden optimal bir çözüm elde etmek için lagrange çarpanı tekniğini kullanılır.

Lagrange çarpanı, kısıtlamalara tabi tutulan bir fonksiyonun yerel maksimum ve yerel minimum noktalarını bulan bir optimizasyon tekniğidir. Bu tip kısıtlı optimizasyon problemlerinin çözümü lagrange çarpanı yardımıyla sağlanır.

Destek vektör makineleri, verilerin ayrıştırılabilirliği açısından doğrusal ve doğrusal olmayan destek vektörleri olmak üzere ikiye ayrılır. Bu aşamaya kadar hep veri setindeki etiketli noktaların doğrusal olarak ayrıştırılmasının, yani gözlem değerlerine karşılık gelen etiketli noktaların doğru bir şekilde sınıflandırılmasının öneminden bahsettik. Veri seti, eğer doğrusal olarak birbirinden ayrıştırılıyorsa buna doğrusal destek vektör makineleri denir. Doğrusal DVM'nin en nihai amacı, veri setindeki yanıt değişkene ait değerleri ilgili kısıta uyacak şekilde sınıflandırarak, margin genişliğinin optimize etmek olduğundan bahsetmiştik.

Bu çerçevede, 2.11'deki kısıtlı optimizasyon problemi modeli aslında hard margin DVM sınıflandırıcısının problem modeline de eşittir. Hard (sert) margin, etiketlenmiş noktaların doğrusal olarak sınıflandırılması prensibiyle çalışan ve sınıflandırma konusunda katı kurallara sahip, doğrusal DVM'ye ait bir sınıflandırma türüdür. Veri setindeki etiketli değerlerin hatalı sınıflandırılmasına kesinlikle tolerans göstermez. $y_i * (\vec{w} \cdot \vec{x}_i + b) \geq +1$ kısıtı yardımıyla +1 ya da -1 olarak etiketlenmiş noktaların, kendi sınıfı dışında başka bir sınıfta veya margin bölgesi içerisinde yer alması, hard margin sınıflandırma için kesinlikle kabul edilemez durumlardır. Örneğin, +1 olarak sınıflandırılan bir noktanın, -1 sınıfında yer alması veya bu noktanın margin bölgesinde yer alması durumunda, hard margin sınıflandırma koşulu karşılanmamış olur. Böylece algoritma, optimal margin genişliğini bulma konusunda başarısız olur.

Hard margin sınıflandırması ile alakalı iki temel özellik söz konusudur. Bunlardan ilki, sadece verilerin birbirinden doğrusal olarak ayrışabildiği durumlarda faaliyet göstermesidir. İkincisi ise aykırı değerlere karşı oldukça duyarlı olmasıdır (Géron, 2017: 146).

Ancak gerçek hayat problemlerinde, aykırı veya gürültülü gözlem değerleri içeren ve doğrusal olarak ayrılamayan bir veri seti ile karşılaşma olasılığımız daha yüksektir. Problem veri seti, içerisinde gürültülü ve aykırı gözlem değerlerini barındırması, hard margin sınıflandırma modelinin eğitim veri setini aşırı öğrenme (overfitting) riskini yükseltir. Modelin veri setini aşırı öğrenmesi, margin genişliği

büyük, optimal bir hiper düzlem belirlenmesini engeller. Bunun yanı sıra, eğitim veri setinde bulunan aykırı değerler nedeniyle margin genişliğinin küçülmesi ise hiper düzlemin mevcut sınıflardan birine çok yakın bir konumda çizilmesine sebep olur. Bu da sınıflandırıcı modelin, etiketsiz değerlerin sınıfını hatalı tahmin etmesine yol açar. Yani bir başka deyişle, model gelecekte tahmin edilecek değerler için genelleme yeteneğini kaybeder. Ayrıca istenilen optimal margin genişliği ve hiper düzlemin belirlenememesi, eğitim veri setinde içerisinde sınıflandırma problemini temsil eden önemli gözlem değerlerinin göz ardı edilmesine de neden olabilir.

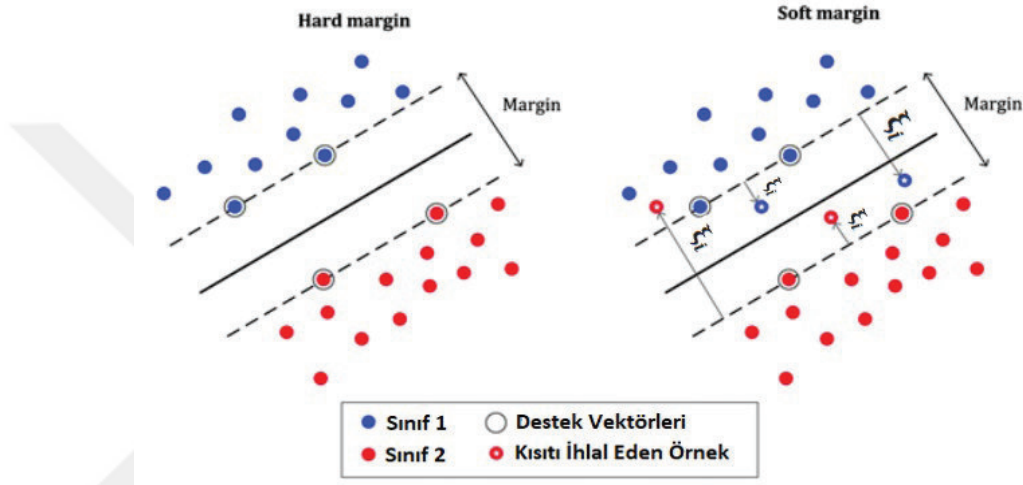
Bu tip sorunların üstesinden gelebilmek aslında gerçek hayat problemlerine uygun ve aynı zamanda sınıflandırma konusunda esnek olan bir DVM modelinden geçer. Sınıflandırma için gereken esneklik ise ancak soft (yumuşak) margin sınıflandırıcısı ile mümkündür. Soft margin sınıflandırıcısı, adından da anlaşılacağı üzere, hard margine göre oldukça esnek sınıflandırma kurallarına sahip bir sınıflandırma türüdür. Veri setindeki etiketli değerlerin hatalı sınıflandırılmasına tolerans gösterir. +1 ile -1 olarak etiketlenmiş noktaların, kendi sınıfı dışında başka bir sınıfta veya margin bölgesi içerisinde yer alması, soft margin sınıflandırıcısı için kabul edilebilir durumlardır.

Soft margin sınıflandırıcısının amacı, eğitim veri seti doğrusal olarak ayrıştırılamasa bile, margin genişliğini eniyilemektir. Bu nedenle soft margin sınıflandırıcısı amacına ulaşmak için eğitim veri setindeki etiketli değerlerin, hatalı sınıflandırılmasına veya sınır çizgilerinin dışına çıkmasına izin verir. Aslında bu durum, hatalı sınıflandırılmasına veya margin alanında yer almasına izin verilen etiketli değerlerin sayısı ile yakından ilişkilidir. Yanlış sınıflandırılan değerlerin mümkün olduğunca en az sayıda olması veya az sayıda noktanın sınır ihlali gerçekleştirmesi, optimal margin genişliğinin elde edilmesi açısından oldukça önemli kriterlerdendir. Ayrıca margin genişliği ve noktaların yaptığı sınır ihlalleri arasındaki dengenin sağlanması, modelin gelecekteki veriler için genelleme yeteneğini koruması açısından da önemlidir.

Soft marginin sınır ihlallerine karşı gösterdiği esnekliğin etkisi, 2.12'deki kısıtlı optimizasyon probleminin amaç fonksiyonuna ve kısıtlarına esnek değişken (slack variable) olarak dahil edilir. $\xi_i(X_i)$ ile ifade edilen bu değişken, hatalı sınıflandırılan noktaların bulundukları konum ile gerçekte ait oldukları sınıflar

arasındaki mesafeye karşılık gelir. Bu duruma ait sınıflandırma örneği şekil 11'deki görselde (soft margin kısmında) gösterildiği gibidir. $\xi_i = 0$ olması durumu, noktanın kendi sınıfında yer aldığını ifade eder. $0 < \xi_i < 1$ aralığında olması durumu, noktanın margin bölgesinde yer aldığını ifade eder. $\xi_i \geq 1$ olması durumu ise noktanın başka sınıf içerisinde yer aldığını ifade eder. Ayrıca ξ_i değişkeni negatif olmama kısıtına uyacak bir değer alması gerekir.

Şekil 11: Soft Margin ve Hard Margin Sınıflandırma Örnekleri



Kaynak: Medium, 2019

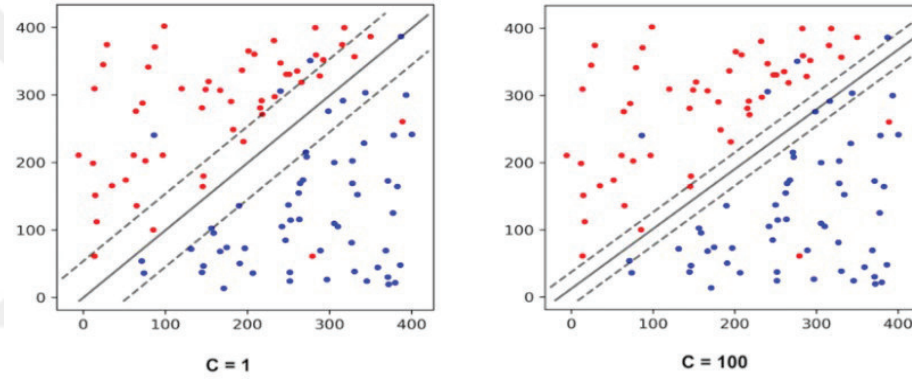
2.12'deki kısıtlı optimizasyon problemine ait modelin amaç fonksiyonunda yer alan bir diğer ifade ise C parametresidir. C parametresinin görevi, margin genişliği ile hatalı sınıflandırma arasındaki dengeyi sağlamaktır. Bunun yanı sıra C parametresi, hatalı sınıflandırmaya karşı gösterilen toleransın aşırı olması durumunu kontrol altına almaya yardımcı, önemli bir parametredir. En uygun C parametresi değeri ise çapraz doğrulama (cross validation) ile belirlenir. Ayrıca modelin eğitim veri setini aşırı öğrenme (overfitting) riskinin söz konusu olduğu durumda, C parametresinin değerini azaltmak sorunu çözmeye yardımcı olacaktır.

$$\min \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^n \xi_i \quad (2.12)$$

$$y_i * (\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i, \quad \forall \vec{x}_i, \xi_i \geq 0, i = 1, \dots, n$$

C parametresi azaltıldığında, model, yüksek bir margin genişliği elde etmekle beraber hatalı sınıflandırma yapmaya daha fazla tolerans gösterecektir. C parametresi artırıldığında ise model, dar bir margin genişliğine sahip olmakla beraber hatalı sınıflandırma yapmaya daha az tolerans gösterecektir. Bu nedenle çok yüksek bir C değeri, modelinin yanlılığını artırırken, çok düşük bir C değeri ise modelin varyansını arttıracaktır. Modelde eksik öğrenme veya aşırı öğrenme problemi olma durumunu engelleyebilmek açısından C parametresi seçimine dikkat edilmelidir. Ayrıca yüksek bir C değeri seçmek, ξ_i değişkeninin etkisini azaltırken düşük bir C değeri seçmek ise ξ_i değişkeninin etkisini artırır.

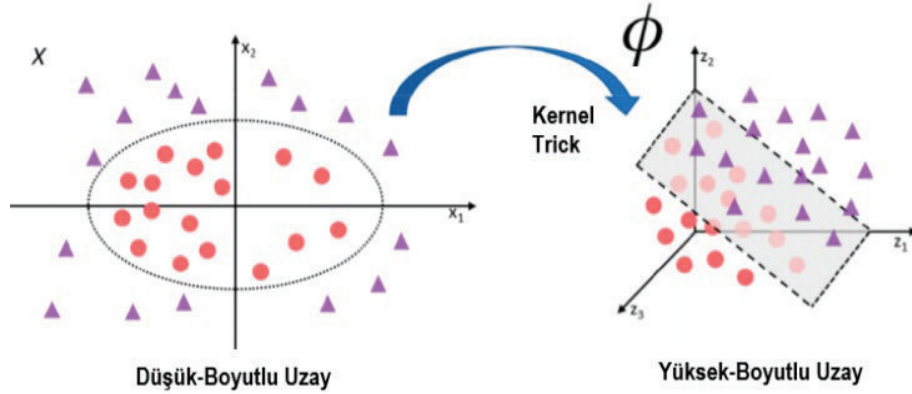
Şekil 12: C Parametresinin Margin Genişliği Üzerindeki Etkisi



Kaynak: Medium, 2020

Düşük boyuta sahip ve doğrusal olarak ayrılamayan veri setine gerçek hayat problemlerinde sık rastlamak mümkündür. Bu tip veri setleri için kullanılacak en etkili sınıflandırma tekniği doğrusal olmayan DVM'dir. Bu teknik, doğrusal olarak ayrılması mümkün olmayan eğitim veri setindeki etiketli değerlerin, çekirdek hilesi (kernel trick) teknikleri yardımıyla yüksek bir boyuta taşıyarak, doğrusal olarak birbirinden ayrılmasını sağlar. Bazı çekirdek hilesi teknikleri ise şunlardır; doğrusal çekirdek, polinomik çekirdek, sigmoid çekirdek ve gaussian çekirdek' tir.

Şekil 13: Doğrusal Olmayan DVM Tekniği Ait Sınıflandırma Örneği



Kaynak: Yang ve diğerleri, 2019: 10

Destek vektör makineleri günümüzde yüz tanıma, metin tanıma, görüntülerin sınıflandırılması ve biyoinformatik alanında oldukça yaygın kullanılır. Ayrıca Destek vektör makineleri algoritmalarının avantaj ve dezavantajları mevcuttur. Avantajları şöyledir; Çok boyutlu veri setine ait problemleri çözmede oldukça etkilidir. Buna ek olarak, küçük miktarda eğitim veri seti ile çok iyi sonuçlar verir. Dezavantajları ise şöyledir; uygun çekirdek (kernel) ve parametre değerleri seçmeye çalışmak, hesaplama yoğunluğuna neden olabilir. Eğitim veri setinde miktar olarak meydana gelecek artış, algoritma performansının düşmesine neden olabilir.

2.3. TOPLULUK ÖĞRENMESİ TEKNİKLERİ

“Birlikten Kuvvet Doğar.” Türkçe’ de yer alan bu atasözü, aslında bu tekniği anlatan en iyi atasözüdür. Topluluk öğrenmesi, birden fazla makine öğrenmesi modelini birlik haline getirerek, kuvvetli tahminler yapmayı amaçlayan bir tekniktir. Bu teknik, bir sınıflandırma veya regresyon problemini bireysel makine öğrenmesi modellerine uygulayarak tahmin performansını iyileştirmeye çalışır. Bireysel olarak modele dahil edilen makine öğrenmesi tekniklerine temel veya zayıf öğreniciler de denir. Birden fazla temel öğrenicinin bir araya getirilmesiyle yapılan tahminlemenin sonuçları, tek bir temel öğreniciyle yapılan tahminlemenin sonuçlarına göre çok daha iyidir. Ayrıca temel öğreniciler bir araya getirildiklerinde, problemin farklı yönlerini

modellemede oldukça yüksek başarı gösterirler. Bu sayede temel öğrencilerin birinde yaşanacak modelleme hatası, bir başka temel öğrenci tarafından telafi edilebilir.

Topluluk öğrenmesi teknikleri k -NN, destek vektör makineleri, lojistik regresyon gibi farklı temel öğrencilerin bir araya getirilmesiyle oluşturulabilirken, sadece karar ağaçlarının bir araya getirilmesiyle de topluluk öğrenmesi tekniklerini oluşturmak mümkündür. Bu çerçevede, topluluk öğrenmesi tekniklerini oluşturmak için kullanılacak modellerin, güçlü ve çeşitli temel öğrenci modellerden seçilmesi, daha doğru ve daha sağlam bir model elde edilmesine yardımcı olacaktır. Bir topluluk öğrenmesi modelinin yüksek doğruluk (accuracy) oranı ve sağlamlık (robustness) gibi iki önemli faktöre sahip olması, tahmin performansını artırır.

Topluluk öğrenmesi teknikleri aynı zamanda temel öğrencilerin aşırı öğrenme (overfitting) ya da eksik öğrenme (underfitting) olma riskine karşın, bu modellerin sahip oldukları varyansın ve yanlılığın dengelenmesini sağlar.

Topluluk öğrenmesi teknikleri çalışma prensibi olarak ardışık ve paralel olmak üzere ikiye ayrılır. Ardışık çalışan modeller daha çok temel öğrencilerin birbirlerine bağıllığından faydalanır. Adaboost, gradient boosting gibi boosting tabanlı teknikler, ardışık çalışma prensibine sahiptir. Paralel olarak çalışan modeller ise temel öğrencilerin arasındaki bağımsızlık durumundan faydalanır. Rastgele ağaçlar algoritması gibi bagging tabanlı teknikler, paralel çalışma prensibine sahiptir. Bahsi geçen topluluk öğrenmesi tekniklerinin işlevlerinden ve çalışma prensiplerinden ise detaylı olarak bahsedilecektir.

2.3.1 Torbalama (Bagging)

Bagging yöntemi, birden fazla temel öğrenciden elde ettiği tahmin sonuçlarını birleştiren ve bu sonuçlardan nihai bir tahmin yapan topluluk öğrenmesi tekniğidir. 1994 yılında Leo Breiman tarafından keşfedilen bu teknik, önyükleme (bootstrapping) ve birleştirme (aggregation) işlemlerinin temeline dayanır. Bu nedenle Bagging yöntemleri literatürde aynı zamanda bootstrapping-aggregation olarak da geçer.

Önyükleme işlemi ile probleme ait orijinal eğitim veri setinden rastgele ve tekrarlı olarak örnekler seçilir ve bunlardan altküme veri setleri oluşturur. Daha sonra elde edilen bu altküme veri setleri, temel öğrencileri eğitmek için kullanır.

Oluşturulan altküme veri setlerine dahil edilecek örnek sayısı (n), orijinal eğitim veri setinin sahip olduğu örnek sayısından (m) küçük olmalıdır ($n < m$). Ayrıca altkümeleri oluşturmak için örneklerin tekrarlı olarak seçilmesi, seçilen örneklerin ilgili altkümelerde birden fazla kez yer alabileceği anlamına gelir. Bu bakımdan özyineleme yöntemi, seçilen örneği tekrardan orijinal eğitim veri setine iade etmesi nedeniyle, oluşturulan ve oluşturulacak olan eğitim verisine ait altkümelerde aynı örneklerin yeniden yer alma şansı oldukça yüksektir.

Altküme veri setlerinin elde edilmesinden sonra, temel öğrenciler, kendileri için ayrı ayrı oluşturulan bu altküme veri setleriyle eğitilir ve eğitilen model test edilir. Test sonucunda, her bir temel öğrenci için bir tahmin değeri elde edilir. Bu tahmin değerlerinin bir araya getirilmesiyle genel bir tahmin değeri belirlenir. Belirlenen tahmin değeri, bagging yönteminin elde ettiği nihai sonuçtur. Nihai sonucun elde edilmesi sınıflandırma veya regresyon problemine göre değişiklik gösterir. Eğer problem, sınıflandırma problemi ise çoğunluk oyu yöntemi kullanılır. Yani, temel öğrencilerden elde edilen kategorik tahmin değerlerinden hangisi çoğunluktaysa, nihai sonuç olarak o kategorik tahmin değeri belirlenir. Eğer problem, regresyon problemi ise elde edilen sayısal tahmin değerlerinin ortalaması, nihai tahmin değeri olarak belirlenir.

Bagging yöntemlerinin en önemli görevi varyansı düşürmektir. Bu sayede, yöntemin içerisinde yer alan temel öğrencilerin overfitting (aşırı öğrenme) olma riskini ortadan kaldırır ve temel öğrencilerin gelecekteki veri setleri için genelleme yeteneği iyileştirilir. Böylece daha doğru tahmin yapan temel öğrenciler elde edilmiş olur. Bu da bagging yönteminin daha güçlü tahminler yapmasını sağlar. Bagging yöntemi prensibiyle çalışan en popüler algoritma ise rastgele ormanlar (random forest) algoritmasıdır. Bu bölümde ekstra karar ağaçları, oylama sınıflandırıcısı ve rastgele ormanlar algoritmasından bahsedilmiştir. Ayrıca bu bölümde hem Ekstra Ağaçlar sınıflandırıcısında hem de Rastgele Ormanların oluşmasında büyük rol oynayan karar ağaçları algoritması hakkında bilgi sahibi olmak adına, ne olduğundan ve işleyişinden detaylıca bahsedilmiştir.

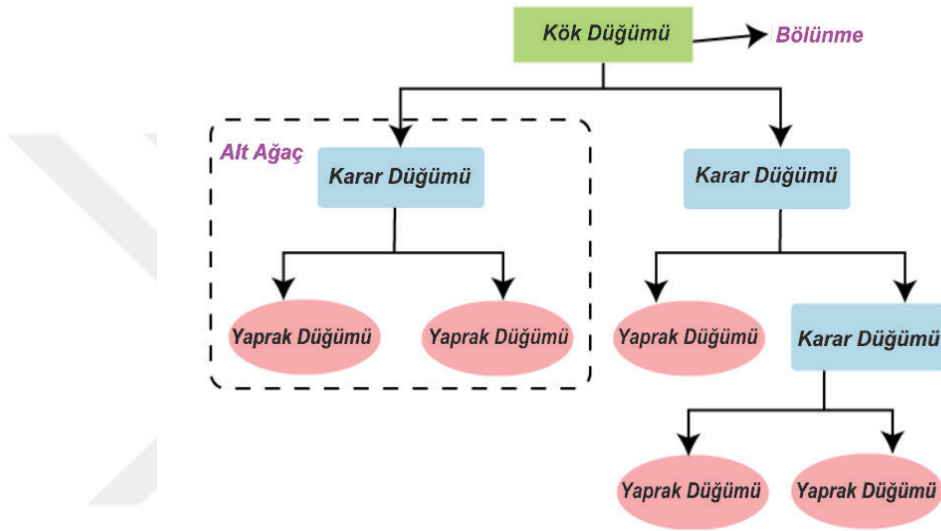
2.3.1.1. Karar Ağaçları Algoritması

Karar Ağaçları hem regresyon hem de sınıflandırma problemleri için kullanılan algoritmalar. Başta rastgele ormanlar algoritması olmak üzere tüm ağaç tabanlı algoritmaların temelinde karar ağaçları algoritması vardır. Karar ağaçları ayrıca heterojen/karmaşık yapıdaki veri setlerini, homojen/anlaşılır hale getirerek karar verme sürecini kolaylaştıran popüler algoritmaların başında gelir. Hiyerarşik yapıya sahip olan bu algoritma, dallanmaya kök düğümden başlar. Kök düğüm, problemi en iyi temsil eden ve heterojen yapıya sahip, eğitim veri setindeki girdi değişkenlerden biridir. Kök düğümler iki veya daha fazla dala ayrılabilir. Dalların ucundaki alt düğümlere ise karar düğümleri denir. Bu karar düğümleri, problem açısından önemli diğer girdi değişkenlerine karşılık gelir. Dallanmanın devam etmediği düğümlere ise yaprak düğümler denir. Yaprak düğümleri nihai sonuçları taşıyan düğümler olup, ayrıca karar aşamasında referans alınacak sonuç düğümleri olma özelliğine sahiptir. Yaprak düğümlerinin alacağı nihai çıktı değeri aslında problemin türüne göre değişiklik gösterir. Sınıflandırma problemlerinde yaprak düğüme ulaşıldığında, hangi kategorik değer çoğunluktaysa düğüme o değer atanır. Regresyon problemlerinde ise, probleme ait veri setinde yer alan sayısal çıktı değerlerinin ortalaması alınır ve yaprak düğüme nihai sonuç olarak atanır. ID3, C4.5, CART ve CHAID başta olmak üzere çeşitli karar ağacı algoritmaları bulunur.

Her bir düğüm aslında probleme ait bir soruya veya bir yanıtla tekabül eder. Kök düğümden yaprak düğüme ulaşmaya kadar yinelemeli olarak devam eden bu süreçte, ağacın sahip olduğu dallar, her bir düğümden sorulan sorulara (karar düğüm/kök düğüm) ait yanıtların (girdi değişkenine ait gözlem değerleri) temsilidir. Bu karar sürecinde ilgili sorulara ait yanıtları takip ederek, nihai yanıt (yaprak düğüme) ulaşmaya çalışılır. Örneğin, bir banka müşteriye kredi verme sürecinde olsun. Bu süreçte banka ilgili müşterilere ait bilgilere ve belgelere bakarak müşterinin kredi alıp almayacağına karar verecektir. İlgili banka için önemli olan soru, geçmiş kredilerini düzenli ödeyip ödemediği olduğunu varsayalım. Bu aslında problemin girdi değişkenidir ve kök düğümü temsil eder. Bu soruya ait iki yanıt bulunur; evet ve hayır. Bu evet ve hayır cevapları aslında algoritmanın nihai cevaba (kredi alabilir/kredi alamaz) ulaşmada takip edeceği yolu (dalları) belirler. Banka bu süreçte farklı birçok soru sorup, bu soruya nasıl cevap buluyorsa o şekilde yoluna devam eder. Bazı

müşteriler için sorduğu tek bir soru bankalar için kredi alıp almamasına yeterli olurken, bazı müşteriler için bu soruya verilen cevabın akabinde, aylık kazancı ve mal varlığı gibi farklı sorular da sorulur. Bu soruların sonunda amaçlanan ise bankanın, müşterinin verdiği cevaplara göre kredi alıp almayacağına karar vereceği nihai sonuca (yaprak düğüme) ulaşabilmesidir.

Şekil 14: Karar Ağacı Yapısı



Kaynak: Arain, 2021

Probleme ait modelleri oluştururken, problemi en iyi temsil eden girdi değişken ve/veya değişkenlerini belirleyebilmek oldukça önemlidir. Bu sayede doğru değişken seçimi, probleme çözüm bulma yolunda hızlı ve uygun adımlar atarak doğru tahminlerde bulunmamıza yardımcı olur. Özellikle, karar ağaçları algoritmasında problemi açıklayan girdi değişkenlerinin seçilmesi hem dallanma sayısını azaltacak hem de algoritmanın daha etkin tahminler vermesine fayda sağlayacaktır. Bu nedenle, karar ağaçları algoritmasında kök düğüm için problemi en iyi açıklayan bir girdi değişkenin seçimi oldukça önemlidir. Bu çerçevede, oluşturulacak karar ağacında kök düğümde yer alacak olan değişkeni belirlemede bazı seçim teknikleri kullanılır. Bu teknikler arasında en yaygın olanı bilgi kazancıdır (information gain). Bu teknik dışında, gini indeksi, ki-kare ve varyans azaltma gibi algoritmanın türüne göre ağacın

hangi girdi değişkeninden bölünmeye başlayacağını belirlemeye yardımcı farklı dallanma kriterleri de mevcuttur.

Bilgi kazancının hesaplanmasında entropi ölçüsünden faydalanılır. Entropi, rastgelelik ölçüsü olarak ifade edilir. İlgili ağacı oluşturmaya başlanmadan önce hesaplanması gereken ilk ölçümdür. Veri setindeki gözlem değerlerinin ne oranda heterojenliğe sahip olduğunu ölçmek için kullanılır. Daha çok sınıflandırma problemlerine ait karar ağaçlarının bölünmesinde kullanılan bu metrik, kök düğümden yaprak düğümlere kadar ağacın inşasında kullanılan değişkenlerin rastgelelik oranını ölçmektedir.

Veri setindeki heterojenlik (impurity) yüksek olduğunda, entropi değeri yüksektir. Homojenlik (purity) oranı yüksek olduğunda ise entropi oranı düşüktür. Buna bağlı olarak bilgi kazancı, veri setinde problemi en iyi temsil eden değişkene göre bölümlendirerek Entropi oranında meydana gelecek değişimlerin ölçülmesini sağlar. Bölünmenin başlayacağı girdi değişkenin, sınıf hakkında yeterli bilgi oranına sahip olup olmadığını ölçer. Bu sayede değişkenin, bilgi kazancı değerine bakarak bölünmenin yapılacağı değişkeni seçmemize ve karar ağacını oluşturmamıza yardımcı olur. Karar ağacı algoritması, bilgi kazancını maksimize edecek değişkenleri önce bölmeye çalışır. Böylelikle, Entropi değerini düşürerek veri setinin homojen olarak ayrılmasına yardımcı olur.

2.13'teki formül ile veri setine ait genel Entropi hesaplanır. P_i değeri yanıt değişkenindeki kategorik değerlerin olasılığını temsil eder. Örneğin, kredi alabilir veya kredi alamaz olarak etiketlenen 30 gözlemin 14'ü 'kredi alabilir' olarak, 16'sının ise 'kredi alamaz' olarak etiketlendiğini varsayalım. Kredi alabilir için P_i değeri $14/30$ 'a eşit olur. Kredi alamaz için P_i değeri ise $16/30$ 'a eşit olur. Bu iki olasılığın logaritmik değerleri ile kendi değerlerinin negatif haliyle çarpımlarının toplamı Entropi oranını verir. S_1 ise toplam gözlem değerini ifade eder. İlgili girdi değişkenine ait gözlemlere karşılık gelen etiketli (kategorik) değerlerin sayısı oransal olarak birbirine eşit olduğunda yani, $P_i = 0.5$ ise entropi değeri 1'e eşit olurken, bütün gözlemlerin tamamı aynı etiketli(kategorik) değerlerden oluşuyorsa yani, $P_i = 1$ ise entropi değeri 0 (sıfır)'a eşit olur.

$$Entropi(S_1) = \sum_{i=1}^n -P_i \log P_i \quad (2.13)$$

2.14' teki formülde w_i ağırlığı, ilgili değişkenin ilgili gözlem değerine karşılık gelen etiketlerin sayısı ile toplam etiket sayısının birbirine oranlanmasıyla hesaplanır. Entropi oranı ise yalnızca ilgili girdi değişkeni için hesaplanır. Her bir girdi değişkeni için hesaplanan Entropi değeri, 2.13'teki Entropi formülüyle aynı hesaplama mantığına sahiptir. Fakat bu kez, ilgili girdi değişkenine ait her bir gözlem için ayrı ayrı hesaplama yapılır. Hesaplanan w_i ağırlığı ile ilgili değişkenin gözlem değerleri için hesaplanan Entropi 'nin çarpımı bize ilgili tüm girdi değişkenleri için hesaplanan Entropi' yi verir.

$$Entropi(S_2) = \sum_{i=1}^n w_i Entropi(Her Bir Değişken için) \quad (2.14)$$

2.15'teki formül ise veri seti için hesaplanan genel entropi ile ilgili tüm girdi değişkenleri için hesaplanan Entropi değerinin farkı, bilgi kazancını verir. Bilgi kazancı yüksek olan girdi değişkenini kök düğüm için seçmek, homojen alt ağaçlar elde etme olasılığını daha da yükseltecektir.

$$Bilgikazancı(F) = Entropi(S_1) - Entropi(S_2) \quad (2.15)$$

Diğer algoritmalarda olduğu gibi bu algoritmada da aşırı öğrenme (overfitting) sorunuyla karşılaşmak mümkündür. Karar ağaçlarında aşırı uyum engelleyecek iki yol vardır; durdurma ve budama parametrelerini ayarlamak.

Karar ağacındaki dallanmanın fazla oluşu, algoritmanın eğitim setini ne kadar iyi öğrendiğini gösterir. Ancak ağaçtaki aşırı dallanma her zaman olumlu sonuçlar doğurmayabilir. Bir karar ağacında durdurma koşulunun olmaması, algoritmanın sürekli düğüm oluşturmaya neden olur. Bu da modelin doğruluk oranını artırır. Bu durum, aslında modelin aşırı öğrenme riskini gösterdiğinin işaretidir. Büyüyen ağaç hacmi aslında çözülmeye çalışılan problemi çok iyi temsil etse de daha önce test

edilmemiş problem veri setleri üzerinde genelleme yeteneğini kaybeder. Tüm bunlara ek olarak, durdurma kriterinin olmaması, karar ağaçlarını anlaşılması zor ve karmaşık algoritmalara dönüştürür. Bu nedenle, algoritma için uygun bir durdurma kriteri belirlemek, tüm bu olayların gerçekleşmesini engellemek için oldukça önemlidir.

Aşırı öğrenmeyle başa çıkmanın bir diğer yolu ise ağacı budamaktır. Budama işlemi, alışılmışın aksine, ağacın alt dallarından üst dallarına doğru uygulanan bir işlemidir. Bu işlemle beraber amaçlanan gereksiz düğümleri çıkararak modelin veri setini aşırı öğrenmesini engellemektir. Ayrıca ağacın budanması, daha önce test edilmemiş veri setlerine karşı modelin genelleme yeteneğini artırır. Ağaç budaması için bir karmaşıklık parametresi belirlenir. Bu karmaşıklık parametresi, hacmi büyük ağaçları daha okunaklı ve anlaşılması kolay hale getirir. Ayrıca ağacın hacminin büyümesini de engeller. En uygun karmaşıklık parametresini, çapraz doğrulama (cross-validation) yöntemi yardımıyla belirlemek mümkündür.

2.3.1.2. Rastgele Ormanlar Algoritması

Rastgele Ormanlar Algoritması, birden fazla tekil karar ağacının bir araya getirilmesiyle meydana gelen bir denetimli makine öğrenmesi tekniğidir. Breiman tarafından geliştirilen bu algoritma, hem sınıflandırma hem de regresyon problemlerinde etkin olarak kullanılır (Breiman, 2001). Ancak sınıflandırma problemlerinde daha başarılı bir performans gösterir. Breiman aslında bu algorithmada kendi geliştirmiş olduğu iki temel yöntemi de birleştirmiştir. Bunlardan biri CART algoritması, bir diğeri ise Bagging yöntemidir. CART algoritması ile kurulurken, her bir ağaç için üretilecek alt veri kümeleri ise Bagging yöntemindeki gibi eğitim veri setinden tekrarlı olarak rastgele seçilen örneklerden oluşur. Ayrıca algorithmadaki tekil ağaçlar birbirinden bağımsızdır ve birbirlerine paralel çalışırlar.

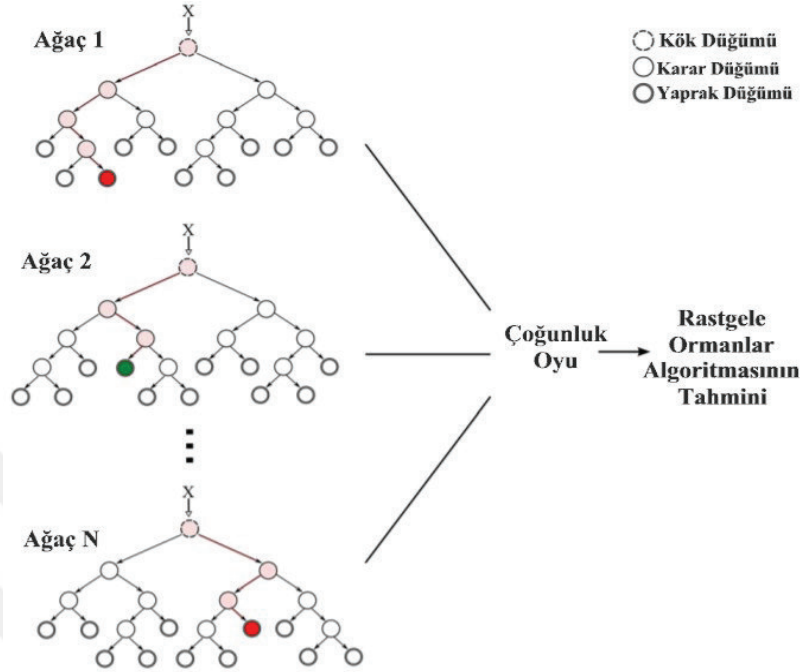
Rastgele Ormanlar algoritmasını bagging yönteminden ayıran en önemli özellik rastgele değişken seçimi yapmasıdır. Rastgele ormanlar algoritması, rastgele alt uzay (random subspace) yöntemi (Ho, 1998) yardımıyla, ağaçların dallanma sırasında kullanacağı değişkenleri rastgele olarak seçer. Bu sayede hem örnek seçiminde hem de değişken seçiminde rastgeleliğin etkisini kullanarak modeldeki ağaç çeşitliliği sağlanır. Buna ek olarak, problemin farklı yönlerini yansıtan ağaçlar oluşturularak performans açısından kuvvetli bir tahmin modelleri elde edilmiş olur.

Rastgele alt uzay yöntemi, seçtiği değişkenleri ağaçların eğitilmesinde kullanılacak olan alt küme eğitim veri setlerinden seçer. Seçimden sonra elde edilen veri setleri ile tahmin modelleri eğitilir. Gereksiz değişkenlerin alt küme eğitim veri setlerinden elimine edilmesiyle oluşturulan yeni veri setleriyle eğitilen tüm zayıf öğrenciler, orijinal eğitim setiyle eğitilen tekil modellere göre çok daha iyi sınıflandırma performansı gösterebilirler. Ayrıca eğitilen zayıf öğrencilerden elde edilecek sonuçların birleşimi, orijinal veri setinin kullanıldığı tek bir sınıflandırma modelinden elde edilecek sonuca göre çok daha üstün olabilir.

Rastgele ormanlar algoritmasının kurulmasında ve performansında önem arz eden parametreler bulunur. Bu parametreler; oluşturulacak ağaç sayısı (N), rastgele seçilen örneklerden oluşan alt kümeler (x). Algoritma için belirlenen ağaç sayısı olan N değeri, alınacak optimal sınıflandırma performansına göre değişir. Bu nedenle, optimal sonucu elde edene kadar farklı ağaç sayılarında algoritmayı çalıştırmakta fayda vardır. Rastgele ormanlar algoritması, aşırı öğrenmeye karşı oldukça dirençlidir. Buna bağlı olarak, algoritmayı oluşturan ağaç topluluklarının sayısının fazla olması, overfitting (aşırı öğrenme) olması için herhangi bir tehlike teşkil etmemektedir. Buna ilave olarak, probleme ait X hacimli eğitim veri setinden tekrarlı olarak rastgele seçilen örneklerden her bir karar ağacını eğitiminde kullanılacak x hacimli alt kümeler oluşturulur. Bu x hacimli alt küme eğitim veri setinin $2/3$ 'ü modeli eğitmek için kullanılırken (inbag), geriye kalan $1/3$ 'lük kısmı (out-of-bag) ise modeli test etmek ve modelin sahip olduğu iç hata oranını (OOB hata oranı) belirlemek için kullanılır.

OOB hata oranı, algoritmanın bünyesinde problemin çözümü için yeterli sayıda karar ağacını barındırıp barındırmadığını tespit etmede kullanılır. Algoritma, problem çözümü için kullanılacak yeterli karar ağacı sayısına ulaştığı andan itibaren algoritmaya eklenecek her bir ağaçtan elde edilecek OOB hata oranı sabit bir değere yaklaşır. OOB hata oranı, aynı zamanda algoritmanın en uygun nihai tahmin değerini elde etmesine de yardımcı olur. İlgili karar ağaçlarının hata oranı düşükse, o ağaçların elde ettiği tahmin değeri, algoritmanın vereceği nihai kararda daha ağır basar. Bu nedenle algoritma hata oranları düşük ve yüksek olan ağaçları karar aşamasında birbirinden ayırmak için her bir ağaca ağırlık değeri verir. OOB hata oranı yüksek olan ağaçlar düşük bir ağırlık değeri alırken, OOB hata oranı yüksek olan ağaçlar daha yüksek ağırlık değeri alırlar.

Şekil 15: Rastgele Ormanlar Algoritmasının Yapısı



Kaynak: Machado ve diğerleri, 2015: 5

Rastgele ormanlar algoritmasının bir diğer önemli parametresi ise rastgele seçilen değişken (tahminci) sayısıdır (m). Sınıflandırma probleminde, dallanma için kullanılacak olan rastgele seçilen değişkenlerin sayısı (m), alt küme eğitim veri setinde yer alan M adet değişken (tahminci) sayısının karekökü kadardır ($m = \sqrt{M}$). Regresyon probleminde ise dallanma için kullanılacak değişken sayısı (m) ise $m = M/3$ 'tür.

Rastgele Ormanlar algoritmasında bulunan karar ağaçlarının dallanmaya başlaması için problemi en iyi temsil eden değişkeni bulmak oldukça önemlidir. Bunun yanı sıra, seçilecek değişkenin sahip olduğu örneklerin diğerlerine göre daha homojen bir yapıya sahip olması gerekmektedir. Belirlenecek en uygun değişkenler, süreci en iyi şekilde yansıtacak ve algoritmanın kısa bir sürede doğru tahminler elde etmesini sağlayacaktır. Ayrıca doğru değişkenlerin seçilmesi, problemin temelinde yatan önemli noktaları tespit edilmesi yardımcı olarak, algoritmanın tahmin performansını arttıracaktır. Bu nedenle, CART karar ağaçlarının dallanmaya başladığı kök düğüme yerleştirilecek rastgele seçilen değişkenler arasından en uygun değişkeni bulmak için

gini indeksi (2.16) tekniği kullanılır. Seçilen değişkenden dallanma işlemine başlanır. Bu değişken için en iyi dallanma kriteri yine gini indeksi ile belirlenir. Gini indeksi haricinde; standart yöntemler, örnekler arasında yakınlık (proximity) rastgele ormanlar algoritması için kullanılan diğer değişken (tahminci) önem değerini hesaplama tekniklerinden bazılarıdır.

Ana düğümün sahip olduğu her alt düğüm için gini safsızlığı değeri 2.16'daki formülle hesaplanır. p , tüm veri kümesini ifade eder. N , sınıf sayısını ifade eder. p_i ise alt düğümde yer alan i . sınıfa ait örneğin, yine alt düğümde toplam örnek sayısına oranıdır. Gini safsızlık değeri aslında değişkene alt düğümlerin heterojenliğini ölçer.

$$Gini\ Safsızlığı(p) = 1 - \sum_{i=1}^N p_i^2 \quad (2.16)$$

Her bir alt düğüm için hesaplanan gini safsızlık değerleri, ilgili değişkenin sahip olduğu her bir alt düğüm için belirlenen ağırlık değerleri ile çarpılır. Her bir düğüm için belirlenen ağırlık değeri, ana düğüme atanacak potansiyel değişkenin dallanmadan sonra oluşan ilgili alt düğümlerinde yer alan toplam örnek ile yine ana düğüme atanacak potansiyel değişkenin sahip olduğu toplam örneğe oranlanmasıyla elde edilir. Böylelikle ana düğüme gelmesi muhtemel her bir değişken için gini indeksi değeri belirlenmiş olur. Gini indeksine ait hesaplama formülü 2.17'de gösterildiği gibidir.

$$Gini\ İndeksi = \sum_{j=1}^c w_j * Gini\ safsızlığı(p_j) \quad (2.17)$$

Gini indeksi değişkene ait alt düğümlerin safsızlık değerleri yardımıyla, ana düğüm yer alacak değişkenler için heterojenlik hesabı yapar. Gini değerinin 0 veya 0'a yakın olması, değişkenin içerisinde yer alan sınıfların yapısının homojen, yani neredeyse tamamının benzer sınıflardan oluştuğu anlamına gelir. Eğer gini değeri, 0.5 veya 1 değerleri arasındaysa değişkende yer alan sınıfların yapısının heterojen yani, birbirinden farklı sınıflardan oluştuğu anlamına gelir. Her bir değişken için hesaplanan

gini değerleri arasından en düşük gini değerine sahip değişken, kök düğüme atanır ve dallanmaya o düğümden başlanır. Dallanmanın devamında oluşan her bir alt düğüme ise sırasıyla düşük gini değerine sahip diğer değişkenler yerleştirilir. Bu süreç yaprak düğüme ulaşılan kadar devam eder.

Rastgele ormanlar algoritmasının çalışma prensibi aşağıdaki gibidir;

Aşama1: Orijinal eğitim veri setinden (X) tekrarlı olarak rastgele seçilen örneklerle alt küme eğitim veri setleri (x) oluşturulur. Alt küme eğitim veri setinin 2/3'ü karar ağaçları için eğitim veri seti (inbag) olarak kullanılırken, 1/3'ü test veri seti (OOB) olarak kullanılır.

Aşama2: Oluşturulan alt küme veri setindeki değişkenler (M) içerisinde önemli olan değişkenler (m) rastgele olarak seçilir. Seçilecek önemli değişken sayısı, sınıflandırma probleminde $\sqrt{M}$ iken regresyon probleminde değişken sayısı $M/3$ 'tür. Bu değişkenler arasından en çok bilgi kazancı sağlayan değişken belirlenir ve dallanmaya, bilgi kazancı yüksek olarak tespit edilen değişkenden başlanır. Dallanma kriterini belirlemek için gini indeksini hesaplanır. Böylece ağaç, dallanma için belirlenen değişkenin yer aldığı düğümün iki alt dala ayrılmasıyla beraber büyümeye başlar. Bu süreç, veri seti için yaprak düğüme ulaşılan kadar devam eder ve bu işlem, algoritmanın içinde yer alan tüm karar ağaçları için uygulanır.

Aşama3: Elde edilen her bir karar ağacı OOB veri setiyle test edilir ve tahmin değerleri elde edilir. Bunun yanında modelin sahip olduğu iç hata oranı (OOB hata oranı) da belirlenir. Hesaplanan OOB hata oranlarına göre her bir karar ağacına bir ağırlık verilir. Böylece, algoritmanın karar verme süreci kolaylaştırılmış olur.

Aşama4: Her bir ağaçtan elde edilen tahmin değerleri bir araya toplanır. Ağırlığı yüksek olan ağaçların tahminlerine karar aşamasında öncelik tanınarak, problem sınıflandırma problemiye çoğunlukta hangi değer tahmin edilmişse algoritmanın nihai tahmin değeri olarak o değer alınır. Eğer regresyon problemiye ağaçlardan elde edilen tahminlerin ortalama değeri, algoritmanın nihai tahmin değeri olarak belirlenir.

Tüm bu aşamalar belirlenen maksimum ağaç sayısına (N) ulaşılan kadar ve en iyi sonuç elde edilene kadar tekrarlı olarak devam eder.

Rastgele ormanlar algoritmasının bazı avantajları ve dezavantajları mevcuttur. Avantajları şöyledir; Üretilen ağaçlar budanmaz. Veri seti kategorik veya sürekli ya da hem sürekli hem de kategorik değerlerden oluşan değişkenler içeren veri setlerine uygulanabilir. Sınıf sayısı ikiden fazla olan yanıt değişkenine sahip veri setlerine ve hacmi büyük olan veri setlerine de kolayca uygulanabilir. Model overfitting (aşırı öğrenme) olmaya karşı dirençlidir. Veri setinde yer alan kayıp ve gürültülü örnek sayısının fazla olduğu durumda da sınıflandırma başarısı oldukça yüksektir. Yanıt değişkeni, dengeli olmayan veri setlerini sınıflamada oldukça iyi bir performans göstermektedir. Orijinal veri seti eğitim ve test veri seti olarak ayrılmadan oluşturulan karar ağaçlarını eğitmek ve test etmek de mümkündür. Dezavantajları ise şöyledir; Sınıflandırma problemlerinde iyi performans bir performans gösterse de regresyon problemlerinde pek de iyi bir performans gösterdiği söylenemez. Modelin oluşturduğu karar ağaçlarını, tekil ağaçlardaki gibi görsel olarak görmek pek mümkün değildir. Bunun sebebi, modelin karmaşık bir yapıya sahip olmasıdır. İşte bu yüzden bazı ağaç modellerinin elde ettiği sonuçları nasıl elde edildiğine dair bilgiye ulaşmak zordur (Akman, Genç ve Ankaralı, 2011: 39).

2.3.1.3. Ekstra Ağaçlar Sınıflandırıcısı

Ekstra ağaçlar sınıflandırıcısı, rastgele ormanlar algoritmasındaki gibi eğitim veri setine ait alt kümeler yardımıyla inşa edilen ağaç topluluklarından oluşan bir sınıflandırma tekniğidir. Ekstra ağaçlar sınıflandırıcılarını rastgele ormanlar algoritmasından ayıran en önemli özelliği, oluşturulan ağaçlara budama işlemi yapılmamasıdır. Ayrıca oluşturulacak ağaçların kullanacağı alt kümeler, özyinelemeli (bootstrapping) şekilde oluşturulan alt kümelerden değil, tüm veri setinin kullanıldığı alt kümelerden oluşur. Özyinelemeli (bootstrapping) yöntemle alt kümeleri oluşturmak yerine tüm eğitim veri setinden oluşturulan alt kümeleri kullanmak, modelin yanlılığını azaltacaktır. Buna ek olarak, bir ağacı düğümlere ayırmak için gerekli olan değişkenlerin seçimi rastgele yapılır. Ekstra ağaçlar sınıflandırıcısı arkasındaki temel fikir, değişken seçiminin rassal hale getirilmesidir. Böylece, ekstra ağaçlar sınıflandırıcı için varyansı düşürmede kullanılacak rassallaştırma stratejisi, diğer modellerde kullanılacak rassallaştırma stratejilerine göre çok daha başarılı olacaktır. Ekstra ağaçlar sınıflandırıcısı için belirlenecek nihai tahmin, çoğunluk oyu

yardımla oluşturulan ağaçlardan elde edilen tüm sonuçların bir araya getirilmesiyle elde edilir. Model içerisinde yer alan tüm karar ağaçlarının yukarıdan aşağıya doğru belirlenen rassallaştırma stratejisine bağlı olarak bölündüğünden bu sınıflandırma modeline aynı zamanda aşırı rassallaştırılmış ağaçlar (extremely randomized trees) da denir.

2.3.1.4. Oylama Sınıflandırıcısı

Oylama sınıflandırıcısı, farklı makine öğrenmesi sınıflandırma modellerini birleştirir ve kullanılan her bir modelden elde ettiği tahmin değerlerini toplar. Topladığı bu tahmin değerlerini, oylamaya dayalı olarak yeni bir gözlemin sınıf etiketini tahmin etme amacıyla kullanır. Kullanılan oylama, sert veya yumuşak olmak üzere iki şekilde olabilir. Sert oylama durumunda temel olarak çoğunluk oyu kullanılır. Bu durumda, en fazla oyu alan sınıf, sınıflandırıcı tarafından kullanılacak nihai tahmin değeri olur. Yumuşak oylama için, her sınıflandırıcının sınıf olasılıklarının ortalaması alınarak bir tahmin yapılır. En iyi olasılık değerine sahip sınıf etiketi nihai tahmin değeri olarak kullanılır. Ayrıca ağaç tabanlı topluluk teknikleri, oylama sınıflandırıcısı için temel tahmin modelleri olarak kullanılır.

2.3.2. Artırma (Boosting)

Boosting (arttırma) tekniği, bagging tekniğinin küçük bir varyasyonudur. Bu teknik birden fazla zayıf öğreniciyi bir araya getirerek, tahmin gücü kuvvetli bir model oluşturmaya çalışır. Boosting tekniğinin görevi, zayıf öğreniciler tarafından yanlış tahmin edilen örneklerin hatalarını gidermektir.

Boosting tekniğinin özelliği, zayıf öğrenicileri ardışık olarak çalıştırma prensibine dayanmasıdır. Ayrıca zayıf öğreniciyi eğitmek için altküme veri setini oluştururken bagging' deki önyükleme işlemini kullanır. Boosting yöntemini bagging yönteminden ayıran en önemli özellik, eğitim veri setindeki her bir örneğe ağırlık değeri vermesidir. Örneklerin ağırlık değerleri, zayıf öğrenicinin örnekleri yanlış veya doğru tahmin etmesine göre her iterasyonda güncellenir. Tüm örnekler doğru tahmin edilene kadar, örneklere yapılan bu ağırlık güncellemesi devam eder.

Boosting tekniği prensibiyle çalışan en önemli algoritmalar; adaptive boosting (Adaboost), gradient boosting algoritmalarıdır.

2.3.2.1. AdaBoost

Adaboost algoritması, boosting yönteminin mantığını en iyi anlatan algoritmadır. Adaboost algoritması, zayıf öğrenici toplulukları oluşturarak güçlü tahminler elde etmeyi amaçlayan algoritmadır. İkili (binary) sınıflandırma problemlerinde daha etkin performans gösteren bu algoritma, zayıf öğrenici olarak genellikle tek seviyeli karar ağaçlarını kullanır. Adaboost algoritmasının çalışma prensibi aşağıdaki gibidir;

Aşama 1: Orijinal eğitim veri setinden rastgele seçilen örneklerden bir altküme veri seti oluşturulur. (Tüm örnekler eşit ağırlığa sahiptir.)

Aşama 2: Oluşturulan altküme veri setiyle zayıf öğrenici eğitilir ve “*orijinal eğitim veri seti*” ile test edilir. (Zayıf öğrenici genellikle tek seviyeli karar ağacıdır.)

Aşama 3: Test aşamasından sonra yanlış tahmin edilen örnekler tespit edilir ve ağırlık değerleri güncellenir. Doğru tahmin edilen örneklerin ağırlık değerleri düşürülürken, yanlış tahmin edilen örneklerin ağırlık değerleri yükseltilir.

Aşama 4: Örnekler üzerinde ağırlık güncellemesi yapıldıktan sonra, örnekleri güncel ağırlığa sahip altküme veri setiyle zayıf öğrenici tekrar eğitilir ve test edilir.

Bu süreç, tüm örnekler doğru bir şekilde tahmin edilene kadar aynı sırada devam eder.

2.3.2.2. Gradyan Artırma (Gradient Boosting) Algoritmaları

Hem sınıflandırma hem de regresyon problemlerinde kullanılan bir diğer topluluk öğrenmesi tekniği de gradyan artırma algoritmalarıdır. Friedman (2001) tarafından geliştirilen bu algoritma, birden fazla zayıf öğreniciyi birleştirerek güçlü tahmin yeteneği kazanır. Boosting tekniklerde olduğu gibi algoritmada yer alan zayıf öğreniciler, ardışık çalışma prensibine sahiptir. Gradyan artırma algoritmasını boosting tekniklerden ayıran en önemli özelliği ise bir önceki iterasyonda yanlış sınıflandırılan örneklere ağırlıklandırma işlemi yapılmamasıdır. Gradyan artırma algoritması, hedef çıktılar ile topluluktan elde edilen nihai tahminlerin farkını alarak kalıntılar elde eder. Bu algoritma aslında, önceki iterasyondaki zayıf öğreniciden elde ettiği kalıntılarla bünyesindeki zayıf öğrenicileri eğiten bir topluluk öğrenmesine dayanır. Gradyan artırma algoritmaları, güçlü tahminleme yapabilmesi için bileşenler üzerine kurulmuştur. Bu bileşenlerden en önemlileri; problem için uygun olan

türevlenebilir kayıp fonksiyonunun belirlenmesi, algoritmanın daha güçlü tahminler verebilmesi için kullanılacak zayıf öğrencilerin seçilmesi ve belirlenen kayıp fonksiyonunu optimize edilebilmesi için algoritmaya yeni zayıf öğrencilerin ilave edilmesidir. Gradyan artırma algoritmalar için bu bileşenlerin en iyi şekilde belirlenmesi, algoritmanın tahmin gücünü üst düzeye çıkaracak, bununla beraber yanlılığı en aza indirecektir.

Gradyan artırma algoritmasında her yeni bir zayıf öğrenci, bir önceki zayıf öğrenciden elde edilen kalıntıları kullanarak oluşturulur. Buradaki amaç, zayıf öğrencilere ait kayıp fonksiyonlarını optimize ederek hedef çıktı ile nihai tahmin değerleri arasındaki farkı düşürmektir. Başka bir deyişle, ilgili zayıf öğrenciye ait kayıp fonksiyonuna ait değeri 0 (sıfır)'a yaklaştırmaktır. Kayıp fonksiyonu, modelin güvenilirliğini yansıtan ve modeldeki hem kayıpları hem de kalıntıları ölçmeye yardımcı bir fonksiyondur. Regresyon ve sınıflandırma problemleri için farklı kayıp fonksiyonları bulunur. Bu kayıp fonksiyonlarının en önemli özelliği, türevlenebilir fonksiyonlar olmasıdır. En sık kullanılan kayıp fonksiyonları şunlardır; Regresyon problemleri için kare kayıp fonksiyonu, mutlak kayıp fonksiyonu ve Huber-M kayıp fonksiyonudur. Sınıflandırma problemleri için kullanılan fonksiyon ise üstel kayıp fonksiyonudur. Kayıp fonksiyonu değeri ne kadar küçükse, modelin doğruluğu o kadar yüksek olacaktır.

Gradyan artırma algoritması, zayıf öğrenci olarak karar ağaçları kullandığında gradyan artırma karar ağaçları (gradient boosting decision trees veya GBDT) adını alır. GBDT algoritmasında zayıf öğrenci olarak kullanılan karar ağaçları makine öğrenmesi teknikleri, sınıflandırma konusunda oldukça başarılı geçmişe sahiptir. Başarısının yanında kısıtlamaları söz konusudur; karar ağaçları için yeterli veri sağlandığında karar ağacının derinliğinde artış olacaktır. Ancak, ağacın üstel olarak büyümesi ağacın derinliğini sınırlandıracaktır. GBDT algoritması, bu sorunun üstesinden gelmek için kullanılabilir. Tek bir karar ağacı gereğinden fazla aşırı uyum gösterebilirken, GBDT algoritması overfitting (aşırı öğrenme) riskini azaltabilir. Zayıf öğrenci olarak dahil edilen karar ağaçları modelleri, gradyan kaybını azaltmak için GBDT algoritmasında gerçekleşen her bir iterasyonda güncellenir. Bu durum, her bir iterasyon zayıf öğrencilerin eğitim başarısını artırır. Gerçekleşen iterasyon sayısının fazla olması eğitim hatasını azaltabilir. GBDT algoritmasının eğitim veri setine çok

fazla uyum göstermesi aşırı öğrenme riskini artırır. Bu da eğitim hatasında artışa yol açar. Aşırı öğrenme riskini azaltmak için ridge, lasso ve elasticnet gibi farklı düzenlileştirme (regularization) işlemleri kullanılır. Düzenlileştirme işlemlerine aynı zamanda küçültme işlemleri olarak da bilinir. Ek olarak, bir öğrenme oranı parametresi de belirlenir. Öğrenme oranını azaltmak, aşırı öğrenme durumunu kontrol altına almaya yardımcı olabilir. Öğrenme oranında küçültmenin fazla olması, eğitim hatasını artırır. Bu nedenle öğrenme oranı üzerinde yapılacak küçültme, hem eğitim hatası hem de aşırı öğrenme riskini dengeleyecek miktarda olması gerekir (Alam ve diğerleri, 2020: 201183).

2.4. PERFORMANS ÖLÇÜLERİ

Bir problemi çözmek için sadece modellemek yetmez. Bunun yanında, probleme en iyi çözümü getirecek olan modeli belirleyebilmek de çok önemlidir. Bu nedenledir ki ilgili problem için en iyi sonucu veren modelleri tespit edebilmek hem zaman hem de maliyet açısından bizlere büyük avantajlar sağlayacaktır. En iyi sonucu veren bu modelin tespiti ise ancak performansını ölçüm istatistiklerinin hesaplanmasıyla mümkündür.

Performans ölçüm istatistikleri, makine öğrenmesi modelleri arasından, en iyi performansa sahip modeli bulmamıza yardımcı olan değerlendirme enstrümanlarıdır. Ölçüm istatistikleri, ilgili problem üzerinde uygulanan modellerin performansları hakkında çıkarsama yapılmasına yardımcı olur. Eğer ilgili problem sınıflandırma problemiyse, modelin sınıflandırmayı doğru şekilde gerçekleştirilmesi gerekir. Bu durum aslında model için bir başarı kriteridir. Modelin, sınıflandırma yaparken bu kritere bağlı kalması, iyi bir performans göstereceğinin işaretidir. Özetle, sınıflandırmada başarı oranı yüksek olan optimal bir modelin belirlenmesi, test veri setindeki kategorik yanıt değişkenlerinin ne kadarının doğru sınıflandırıldığıyla ilişkilidir.

Bu bölümün devamında, sınıflandırma modellerinin değerlendirmek için kullanılacak çeşitli performans ölçüm metriklerinden detaylıca bahsedilmiştir.

2.4.1. Karmaşıklık Matrisi

Karmaşıklık matrisi, sınıflandırma modeli tarafından hem doğru hem de yanlış sınıflandırılan kategorik değerlerin sayısını veren bir matristir. Karmaşıklık matrisi, bir performans ölçüm metriği olarak kullanılmaz. Ancak matris içerisinde yer alan değerler, diğer performans ölçüm metriklerinin hesaplanmasında kullanılır. Bu açıdan karmaşıklık matrisi, bir modelin performansının hesaplanmasında büyük rol oynar. Karmaşıklık matrisinin yapısı tablo 1’de gösterildiği gibidir.

Genelde elde edilen matris 2x2 boyutundadır. Ancak yanıt değişkende yer alan kategorik değerlerin sayısına göre matris boyutu büyüyebilir ve $n \times n$ boyutunda bir matrise de dönüşebilir.

Tablo 1: Karmaşıklık Matrisi

GERÇEK		TAHMİN	
		POZİTİF	NEGATİF
	POZİTİF	Doğru Pozitif (TP)	Yanlış Pozitif (FP)
	NEGATİF	Yanlış Negatif (FN)	Doğru Negatif (TN)

Sınıflandırma yapan modelin performansını tespit edebilmek için model tarafından gerçek yanıt değerlerinden kaçır tanesinin doğru ve yanlış tahmin edildiğinin belirlenmesi gerekir. Bu durumu bir sınıflandırma problemi üzerinde açıklamaya çalışalım.

Bir bankanın, potansiyel müşterilerini kredilendirme sürecinde olduğunu varsayalım. Bu banka, müşterilerinden ve istihbarat işlemlerinden elde ettiği bilgilere dayanarak, müşterinin kredi alıp almamasına karar vermek zorundadır. Veri setinin sahip olduğu girdi değişkenlerine ait gözlem değerleri; alabilir (1), alamaz (0) şeklinde etiketlenmiş olsun. Bu etiketler problemin yanıt değişkeni oluşturur. Ayrıca matriste yer alan pozitif ve negatif kategorik değerleri, örnek sınıflandırma problemimizdeki karşılığı ise sırasıyla, alabilir (1) ve alamaz (0) kategorik değerleridir.

- **Doğru Pozitif (TP):** Sınıflandırma modelinin, veri setindeki pozitif (1) etiketli gözlemleri, tekrardan pozitif (1) etiketli olarak tahmin etmesi anlamına gelir. Aynı zamanda, “Sınıflandırma modeli, kredi alabilen müşterilerden kaçını ‘kredi alabilir’ şeklinde doğru tahmin etmiştir?” sorusunun cevabıdır.
- **Doğru Negatif (TN):** Sınıflandırma modelinin, veri setindeki negatif (0) etiketli gözlemleri, tekrardan negatif (0) etiketli olarak tahmin etmesi anlamına gelir. Aynı zamanda, “Sınıflandırma modeli, kredi alamayan müşterilerden kaçını ‘kredi alamaz’ şeklinde doğru tahmin etmiştir?” sorusunun cevabıdır.
- **Yanlış Pozitif (FP):** Sınıflandırma modelinin, veri setindeki negatif (0) etiketli gözlemleri, pozitif (1) etiketli olarak yanlış tahmin etmesi anlamına gelir. Aynı zamanda, “Sınıflandırma modeli, kredi alamayan müşterilerden kaçını ‘kredi alabilir’ şeklinde yanlış tahmin etmiştir?” sorusunun cevabıdır.
- **Yanlış Negatif (FN):** Sınıflandırma modelinin, veri setindeki pozitif (1) etiketli gözlemleri, negatif (0) etiketli olarak yanlış tahmin etmesi anlamına gelir. Aynı zamanda, “Sınıflandırma modeli, kredi alabilen müşterilerden kaçını ‘kredi alamaz’ şeklinde yanlış tahmin etmiştir?” sorusunun cevabıdır.

2.4.2. Doğruluk (Accuracy)

Doğruluk, sınıflandırma modellerinin, sınıflandırmada gösterdiği performansın ölçümü için kullanılan temel bir ölçüm istatistiğidir. Yaygın olarak kullanılan bu istatistik, sınıflandırma modeli aracılığıyla, veri setinde yer alan kategorik değerlerin hangi oranda doğru sınıflandırıldığının tespitine yardımcı olur. Doğruluk oranının olabildiğince 1’e yakın olması istenir. Doğruluk oranının yüksek olması, sınıflandırma modelinin doğru sınıflandırmada etkin bir performans gösterdiği anlamına gelir. Ancak sınıflandırma modelinin performansını ölçerken sadece doğruluk ölçüm istatistiğine bakmak yanıltabilir. Özellikle, dengeli olmayan veri setlerini test eden sınıflandırma modeli için uygun bir performans ölçüm istatistiği değildir. Çünkü herhangi bir kategorik değer sayıca diğerinden üstün olması, modelin sınıflandırma ilgisini sayıca fazla olan kategorik değere kaydıracaktır. Başka bir deyişle, model, kategorik değerlerden birini diğerine göre ‘sayıca’ daha doğru sınıflandırdığı için modelin doğruluk oranını artacaktır. Bu da modelin performansı

hakkında yanlış çıkarım yapılmasına neden olur. Doğruluk oranı, formül 2.14'teki gibi hesaplanır.

$$\text{Doğruluk} = \frac{\text{Doğru Pozitif}(TP) + \text{Doğru Negatif}(TN)}{\text{Doğru Pozitif}(TP) + \text{Doğru Negatif}(TN) + \text{Yanlış Pozitif}(FP) + \text{Yanlış Negatif}(FN)} \quad (2.14)$$

2.4.3. Kesinlik (Precision)

Sınıflandırma modelinin, pozitif etiketli olarak sınıflandırdığı tüm gözlemlerin içerisinde, asıl etiketi pozitif olan gözlemleri hangi oranda doğru sınıflandırdığını ölçmeye yardımcı bir istatistiktir. Örneğin, modelin, kredi alabilir olarak etiketlediği müşteriler ile kredi almaya uygun olmayan fakat, kredi alabilir olarak 'yanlış' etiketlediği müşterilerin içerisinde, aslen 'kredi alabilir' olarak etiketlenen müşterileri hangi oranda doğru sınıflandırıldığını ölçmeye yarar. Performans ölçüm istatistiğine ait formül, 2.15'teki gibidir.

$$\text{Kesinlik} = \frac{\text{Doğru Pozitif}(TP)}{\text{Doğru Pozitif}(TP) + \text{Yanlış Pozitif}(FP)} \quad (2.15)$$

2.4.4. Duyarlılık/Hassasiyet (Recall – Sensitivity)

Sınıflandırma modelinin, pozitif olarak tahmin ettiği etiketli gözlemler ile problem veri setinde yer alan pozitif etiketli gözlemlerin, hangi oranda birbiriyle örtüşüğünü ölçmeye yardımcı istatistiktir. Örneğin, modelin, kredi alabilir olarak etiketlediği müşteriler ile kredi almaya uygun olan fakat, kredi alamaz olarak 'yanlış' etiketlediği müşterilerin içerisinde, aslen 'kredi alabilir' olarak sınıflandırdığı gözlemlerin oranını ölçmeye yardımcı istatistiktir. Duyarlılık oranının düşük olması, modelin pozitif etiketli gözlemleri yanlış sınıflandırma eğilimini artırır. Performans ölçüm istatistiğine ait formül, 2.16'daki gibidir.

$$\text{Duyarlılık} = \frac{\text{Doğru Pozitif}(TP)}{\text{Doğru Pozitif}(TP) + \text{Yanlış Negatif}(FN)} \quad (2.16)$$

2.4.5. F1- Skoru (F1- Score)

Duyarlılık ve kesinlik oranlarından birinin yüksek diğerinin ise düşük olması, modelin performansı hakkında çıkarım yapmamızı zorlaştırabilir. F1- skoru, bu tip durumlarda kullanılan en etkin performans ölçüm istatistiğidir. Bu istatistik aslında duyarlılık ve kesinlik ölçüm istatistiklerinin harmonik ortalamasıdır. Ayrıca F1-skor performans ölçüm istatistiği, dengeli olmayan veri setlerini sınıflandıran modellerin performans ölçümünde oldukça başarılı sonuçlar verir. Performans ölçüm istatistiğine ait formül, 2.17'deki gibidir.

$$F1 - Skoru = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \quad (2.17)$$

2.4.6. Matthews İlişki Katsayısı (Matthews Correlation Coefficient)

Yanıt değişkeni, ikili (binary) değerlerden oluşan problemleri sınıflandıran modellerin performansının ölçülmesinde kullanılan, popüler ölçüm istatistiklerinden biridir. Dengeli olmayan veri setlerini sınıflandıran modellerin performansını ölçmede de oldukça başarılı sonuçlar verir. MCC metriği, -1 ile +1 arasında değer alır. -1 eşit veya -1'e yakın bir değer alması, gerçek ve tahmin değerleri arasında ters bir ilişkinin olduğu anlamına gelir. Bu nedenle, ilgili sınıflandırma modelinin kötü bir performans gösterdiği sonucu çıkarılabilir. +1 eşit veya +1'e yakın bir değer alması ise, gerçek ve tahmin değerleri arasında ilişkinin doğrusal olduğu anlamına gelir (Chicco and Jurman, 2020). Bu nedenle, ilgili sınıflandırma modelinin iyi bir performans gösterdiği sonucu çıkarılabilir. 0'a eşit olması ise sınıflandırma modelinin, etiketli örnekleri rastgele sınıflandırdığı anlamına gelir.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (2.18)$$

2.5. YENİDEN ÖRNEKLEME TEKNİKLERİ

Bazı veri setlerinde etiketli değerler sayıca birbirine eşit olmayabilir. Örneğin, bir veri setinin sahip olduğu 1000 adet etiketli gözlemin 600'ü “kredi alabilir”, 400'ü ise “kredi alamaz” olarak etiketlenmiş olduğunu varsayalım. İşte bu tür veri setlerine “dengeli olmayan veri setleri” denir. Dengeli olmayan veri setleri, bazı modellerin sınıflandırma performansını kötü etkileyebilir. Bunun sebebi, ikili kategorik değere sahip bir veri setindeki herhangi bir kategorik değer, sayıca diğerinden fazla olmasıdır. Sınıflandırıcı model, sınıflandırma yaparken çoğunluk sınıfına odaklandığı için azınlık sınıfı görmezden gelir. Bu da sınıflandırma modelinin sayıca fazla olan etiketli örneklerin sınıfını, diğer etiketli örneklere göre daha doğru tahmin etmiş gözükmesine yol açar. Buna dayanarak, ilgili model için “sözde” yüksek bir tahmin performans oranı elde edilmesine neden olur. Aslında bu durum, etiketli değerleri sınıflandırmada en iyi performansa sahip modeli seçme konusunda bizi yanıltır. Bunun önüne geçebilmek için bazı yeniden örnekleme teknikleri kullanılır. Yeniden örnekleme tekniklerinin temel görevi, etiketli değerlerin sayısını suni etiketli değerler üreterek veya bazı etiketli değerleri veri setinden çıkararak, sınıfların sayıca birbirine yaklaşmasını sağlamaktır; yani veri setini dengelemektir.

Bu çerçevede, yeniden örnekleme türleri yukarı örnekleme (oversampling) ve aşağı örnekleme (undersampling) adı altında iki ana başlıkta incelenir.

2.5.1. Aşağı Örnekleme (Undersampling)

Aşağı örnekleme, problem veri setinde bulunan etiketli örneklerin sınıfları arasındaki eşitsizliği gidermek için kullanılan bir tekniktir. Veri setinin sahip olduğu azınlık ve çoğunluk örnek sınıfları arasındaki sayısal eşitsizliği gidererek veri setini dengeli hale getirmeye çalışır. Aşağı örnekleme tekniği, sınıfları dengelemek için veriden çıkarılması gereken örnek sayısını belirlemede, azınlık örnek sınıfında bulunan toplam etiketli örneklerin sayısını referans alırken sınıfları dengeleme sürecinde uygulayacağı işlemlerin tamamını, çoğunluk örnek sınıfı üzerinde uygular. Yani, çoğunluk sınıfının sahip olduğu bazı etiketli örnekler sınıftan çıkarılarak, azınlık sınıfına sayıca eşitlenmeye çalışılır. Rastgele aşağı örnekleme, near miss, küme

ağırlıkları merkezi (cluster centroids) ve Tomek-links olmak üzere çeşitli aşağı örnekleme teknikleri bulunur.

2.5.1.1. Rastgele Aşağı Örnekleme (Random Undersampling)

Veri setini dengeli hale getirmek için, çoğunluk sınıftaki etiketli örneklerin bir kısmını eksilterek, azınlık sınıfta bulunan toplam etiketli örnek sayısına eşitlemeye çalışan bir yöntemdir. Bu yöntem, veri setini dengelemek amacıyla çoğunluk sınıfta bulunan etiketli örnekleri eksiltme işlemini rastgele yapar. Rastgele olarak yapılan örnek eksiltme işlemi, azınlık sınıf ile çoğunluk sınıf sayıca birbirine eşitleninceye kadar devam eder. Ancak bu veri eksiltme işlemi, problem için önemli olan bilgilerin kaybolmasına neden olabilir. Bu da sınıflandırma modelinin, eğitim veri setine karşı az uyum (underfitting) göstermesine yol açar. Rastgele aşağı örnekleme yöntemi, tüm bu olumsuzluklara rağmen, uygulaması basit ama etkili bir yöntemdir.

2.5.1.2. Near Miss Aşağı Örnekleme Tekniği

Near miss aşağı örnekleme tekniği, Zhang ve Mani (2003) tarafından geliştirilmiştir. Azınlık ve çoğunluk sınıflarına ait örnekleri sayıca dengelemek için çoğunluk sınıfına ait örnekleri eksiltmeye dayalı bir tekniktir. Ancak örnekleri eksiltirken yaşanabilecek bilgi kayıplarını azaltmayı amaçlanabilir. Near miss aşağı örnekleme tekniği, eksiltilmesi gereken örnekleri belirlerken belirli bir nokta ile karşı sınıfa ait en yakın veya en uzak noktalar arasındaki ortalama mesafeleri kullanır. Bu sayede, problem için önemli bilgilerin kaybolmasını engelleyerek problemi daha iyi temsil eden dengeli bir veri seti elde edilmiş olur. Azınlık ve çoğunluk sınıflara ait tüm örneklerin arasındaki mesafe en yakın komşu yöntemi kullanılarak belirlenir. Ayrıca near miss tekniğinin iki sınıf arasında fark oluşturmayı amaçlayan üç farklı türü vardır; Nearmiss-1'e göre bazı azınlık sınıfa ait olan örneklere yakın olan çoğunluk sınıfa ait örnekler seçilir. Bu yöntemde sınıfları dengeli hale getirmek için çoğunluk sınıfa ait örnekler seçilirken, en yakın üç azınlık sınıfı örneğine olan ortalama uzaklığı en küçük olan örnekler dikkate alınır. Nearmiss-2'ye göre azınlık sınıfa ait olan bazı örneklere uzak olan çoğunluk sınıfa ait örnekler seçilir. Bu yöntemde sınıfları dengeli hale getirmek için çoğunluk sınıfa ait örnekler seçilirken, en uzak üç azınlık sınıfı örneğine

olan ortalama uzaklığı en küçük olan örnekler dikkate alınır. Nearmiss-3'e göre, her bir azınlık sınıf örneği için, belirli bir sayıda en yakın çoğunluk sınıf örneği seçilir.

2.5.1.3. Küme Ağırlık Merkezleri (Cluster Centroids)

Aşağı örnekleme tekniklerinin kullanmasındaki temel mesele aslında sınıfları dengeleme olduğundan bahsetmiştik. Ancak, sınıfları dengelemek için yapılacak veri eksiltme işlemi aslında problem için çok önemli bilgilerin kaybıyla karşı karşıya bırakabilir. Özellikle kredi skorlama ve risk analizi üzerine yapılan çalışmalarda, kullanılacak veri setleri gerçek hayat problemlerine ait veri setleri olduğundan dengeli olması çok olası değildir. İşte bu sebeple bu tür veri setleri üzerine uygulanacak veri dengeleme işlemleri özellikle kredilendirme süreci gibi bankacılık faaliyetleri için önem arz eden bilgilerin kaybına yol açabilir. Ayrıca veri setini dengeli hale getirilememesi, kullanılacak sınıflandırma yöntemlerinin hatalı sınıflandırma yapmasına yol açarak sınıflandırma yöntemlerinden elde edilecek verimin azalmasına sebep olabilir. Bu nedenle hem bilgi kaybını önleyecek hem de sınıflandırma tekniklerinin verimini artırmak için sınıfları dengeli hale getirecek bir Aşağı Örnekleme tekniği geliştirilmiştir.

Küme ağırlık merkezleri aşağı örnekleme tekniği, Yen ve Lee (2009) tarafından geliştirilmiştir. Çoğunluk sınıfına ait örnek sayısını azınlık sınıfa ait örnek sınıfına yaklaştırmak için kullanılan bu teknik, adından da anlaşılacağı üzere aslında kümeleme yaklaşımına dayanır. Bu tekniğin görevi, örnek sayısını azaltmak olup çoğunluk sınıfına ait örneklerin kümelenmesi gerektiği unutulmamalıdır. Buna ek olarak, bilgi kaybı yaşanmaması adına verilerin benzerliklerine göre kümelenmesi gerekmektedir. Veriler etiketli olduğundan, kümeleme yöntemleri, yalnızca kümelere ait ağırlık merkezini bulmada kullanılır. Bundan hareketle k-ortalımalı kümeleme yöntemlerine göre ağırlık merkezi oluşturarak çoğunluk sınıfına ait örnekleri eksiltir. Küme ağırlık merkezi, değişken uzayında çoğunluk sınıfa ait örnekler üzerinden tüm değişkenler için ortalama değişken vektörü elde edilerek tespit edilir. Çoğunluk sınıfına ait örneklerden oluşan kümeler için ağırlık merkezleri bulunduktan sonra bulunan küme ağırlık merkezlerine en uzak noktada yer alan çoğunluk sınıfına ait örnekler, en az öneme sahip örnekler olarak kabul edilir. En az öneme sahip olduğu

tespit edilen örnekler, çoğunluk sınıfa ait kümelerden çıkarılır. Böylece, çoğunluk sınıfa aşağı örnekleme yapılarak azınlık sınıfa denk bir sınıf haline getirilmiş olur.

2.5.1.4. Tomek-Links (Tomek Bağlantıları)

Tomek bağlantıları, Tomek (1976) tarafından geliştirilen bir aşağı örnekleme tekniğidir. İki farklı sınıfa ait olan ve birbirlerine mesafe olarak en yakın iki örnek, tomek bağlantısı oluşturur. İki örnek arasında bir tomek bağlantısı olması, örneklerden birinin gürültülü örnek olduğuna veya iki örneğin karar sınırına yakın olduğu anlamına gelir. Tomek bağlantıları gürültülü veya karar sınırına yakın olup yanlış sınıflandırmaya eğilimli çoğunluk örneklerini eğitim veri setinden kaldırarak, azınlık ve çoğunluk sınıflarındaki örnek sayılarını birbirine eşitlemeye çalışır.

Tomek bağlantılarını başka bir şekilde tanımlamak gerekirse, O_i ve O_j olmak üzere farklı sınıflara ait iki örneğin seçilmiş olduğunu varsayalım. O_i ve O_j örneklerinin arasındaki mesafe hesaplanır ve iki örnek arasındaki mesafe $d(O_i, O_j)$ şeklinde gösterilir. Eğer $d(O_i, O_k) < d(O_i, O_j)$ veya $d(O_j, O_k) < d(O_i, O_j)$ gibi bir O_k örneği bulunmuyorsa (O_i, O_j) örnekleri arasında bir tomek bağlantısı vardır (Kubat ve Matwin, 1997: 182). Tomek bağlantıları aynı zamanda veri temizleme işlemi için de kullanılır. Veri setinde yer alan gürültülü örnekleri tespit edip veri setinden çıkarılır. Ancak aşağı örnekleme yerine verileri temizleme için kullanıldığında azınlık sınıfına ait örnekleri de eğitim veri setinden çıkartabilir.

2.5.2. Yukarı Örnekleme (Oversampling)

Yukarı örnekleme yöntemleri de aşağı örnekleme yöntemleri gibi problem veri setinde bulunan etiketli örneklerin sınıfları arasındaki eşitsizliği gidermek için kullanılan bir tekniktir. Veri setinin sahip olduğu azınlık ve çoğunluk örnek sınıfları arasındaki sayısal eşitsizliği gidererek veri setini dengeli hale getirmeye çalışır. Yukarı örnekleme tekniği, sınıfları dengelemek için üretilmesi gereken örnek sayısını belirlemede, çoğunluk örnek sınıfında bulunan toplam etiketli örneklerin sayısını referans alır. Dengeleme sürecinde uygulayacağı işlemlerin tamamını ise azınlık örnek sınıfı üzerinde uygular. Yani, azınlık sınıfın sahip olduğu etiketli örneklerin üzerine suni örnekler ekleyerek, çoğunluk sınıfına sayıca eşitlenmeye çalışılır. Rastgele yukarı

örnekleme, SMOTE, Borderline-SMOTE, SVM-SMOTE ve uyarlanmış sentetik örnekleme (ADASYN) gibi çeşitli yukarı örnekleme teknikleri bulunmaktadır.

2.5.2.1. Rastgele Yukarı Örnekleme (Random Oversampling)

Veri setini dengeli hale getirmek için, azınlık sınıfta bulunan etiketli örnekleri kopyalayarak, azınlık ve çoğunluk sınıfta bulunan etiketli örnek sayısını birbirine eşitlemeye çalışan bir yöntemidir. Bu yöntem, veri setini dengelemek amacıyla azınlık sınıfta bulunan mevcut eksikliği, yine azınlık sınıfın sahip olduğu etiketli örneklerin kopyasıyla tamamlar. Ancak bu durum, eğitim veri setinde aynı değerlerin tekrarlanmasına sebep olur. Bu da sınıflandırma modelinin, eğitim veri setine karşı overfitting (aşırı öğrenme) göstermesine yol açar.

2.5.2.2. SMOTE (Sentetik Azınlık Yukarı Örnekleme Tekniği)

SMOTE, yukarı örnekleme yöntemleri arasında günümüzde yaygın olarak kullanılan ve çoğu zaman sınıfları dengeleme konusunda başarılı performans gösteren tekniklerden biridir. Chawla ve diğerleri (2002) tarafından geliştirilen bu teknik, rastgele yukarı örnekleme tekniğine göre farklı bir çalışma prensibine sahiptir. SMOTE tekniği, sınıfları dengelemek için rastgele yukarı örnekleme tekniğinde olduğu gibi azınlık sınıftaki örnekleri kopyalamaz. Onun yerine, azınlık sınıfta bulunan örneklerin k-en yakın komşularını belirleyerek, yeni sentetik örnekler üretmeye çalışır. Böylece, sınıfları dengelemek için rastgele yukarı örnekleme tekniği yerine SMOTE tekniğinin kullanılması, modeli, eğitim veri setine karşı yaşayacağı aşırı öğrenme (overfitting) tehlikesinden de kurtarmış olur.

SMOTE tekniğinin yeni sentetik örnekler üretirken izlediği aşamalar aşağıdaki gibidir (Yavaş, Güran ve Uysal, 2020: 261);

Aşama 1: Azınlık sınıfta yer alan örneklerin bir kısmı ya da tamamı için belirlenecek k değeri kadar en yakın komşuları belirlenir. Ancak, belirlenecek komşular yine azınlık sınıfında yer alan örneklerdir. Azınlık sınıfındaki örnekler için komşu örnekler belirlenirken, çoğunluk sınıfta yer alan örnekler hesaba katılmaz. Ayrıca komşuların belirlenmesi için referans alınacak azınlık sınıftaki örnek x_i olarak tanımlanır.

Aşama 2: Belirlenen en yakın komşu örnek (x_j) ile komşusu seçilen azınlık sınıftaki örnek (x_i) arasındaki fark bulunur.

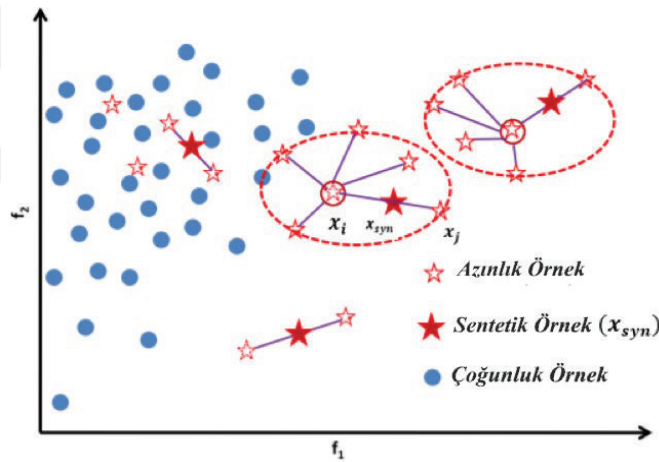
Aşama 3: (0,1) aralığından rastgele bir sayı seçilir ve seçilen bu sayı α parametresine atanır. Bu parametre daha sonra, Aşama 2’deki örneklerin farkından elde edilen değerle çarpılır.

Aşama 4: 2.20’deki formül yardımıyla azınlık sınıf için yeni bir sentetik örnek üretilir.

$$x_{syn} = x_i * (x_j - x_i) * \alpha \quad (2.19)$$

Yukarıdaki tüm aşamalar, azınlık ve çoğunluk sınıfları sayıca birbirine eşitleninceye kadar tekrar tekrar devam eder.

Şekil 16: SMOTE Tekniğine ait Azınlık Sınıf için Sentetik Değerler Üretme



Kaynak: Raghuwanshi ve Shukla, 2020: 2

2.5.2.3. Borderline-SMOTE

Borderline-SMOTE tekniği, He ve diğerleri (2005) tarafından geliştirilen yukarı örnekleme tekniğidir. SMOTE tekniğinin bir çeşidi olan Borderline-SMOTE’nin asıl amacı; sınıfları birbirinden ayıran sınıra yakın olan azınlık sınıfa ait örnekleri tespit etmek ve bu örnekleri kullanarak sentetik veri üretmektir. Borderline-SMOTE tekniğini SMOTE tekniğinden ayıran en önemli özellik, yukarı örnekleme yapılacak azınlık örneğin seçim şeklidir. SMOTE tekniğinde örnekler

rastgele seçilirken, Borderline-SMOTE tekniğinde ise iki sınıfı ayıran sınıra en yakın olan azınlık örnekleri, sentetik veri üretilmesi için seçilir. Borderline-SMOTE tekniği eğitim aşamasında sınıfların sınırını olabildiğince öğrenmeye çalışır. Karar sınırına yakın örnekler ve gürültülü örnekler, yanlış sınıflandırmaya eğilimli olan tehlikeli örnekler olarak adlandırılır. Tehlikeli örnekler sınıflandırma için oldukça önemlidir. Bu tehlikeli örneklerin tespiti yine azınlık sınıfın sahip olduğu komşuların tespiti ile yapılır. Bunun tespiti için SMOTE' de olduğu gibi k -en yakın komşu algoritması kullanır. Sınıra yakın olan azınlık örnekleri için belirlenen komşulara bakılarak azınlık örnekler etiketlenir ve daha sonra tehlikeli olarak etiketlenen azınlık örnekler baz alınarak sentetik örnekler üretilir. Komşulara göre azınlık sınıfa ait örneklerin etiketleri şu şekilde belirlenir (Alam ve diğerleri, 2020: 201181);

- k -en yakın komşuların tamamı çoğunluk sınıfına ait örneklerse azınlık örneği gürültülü olarak kabul edilir.
- k -en yakın komşuların yarısından azı çoğunluk sınıfına aitse verinin güvenli olduğu kabul edilir.
- k -en yakın komşuların en az yarısı çoğunluk sınıfına aitse tehlikeli örnek olarak kabul edilir.

Borderline-SMOTE tekniği, yukarı örnekleme yaparken güvenli olarak etiketlediği örnekleri yok sayar. Daha sonra tehlikeli olarak etiketlenen azınlık örnekleri ile k -en yakın komşuları arasında sentetik örnekler üretilir. Bu örnekler SMOTE tekniğine dayalı olarak üretilir. Borderline-SMOTE yukarı örnekleme yaparak sınıflandırma modelinin iki sınıfı ayıran sınırı yanlış öğrenmesini engelleyebilir. Bu durum aslında, sınıflar arasındaki sınırın güçlenmesini sağlar. Tüm bunlara ek olarak, Borderline-SMOTE1 ve Borderline-SMOTE2 olmak üzere iki türü bulunmaktadır.

2.5.2.4. SVM-SMOTE

SVM-SMOTE, Nguyen ve diğerleri (2011) tarafından geliştirilen bir yukarı örnekleme tekniğidir. Borderline-SMOTE' ye alternatif olarak geliştirilen bu teknik, sınıf sınırına en yakın azınlık örnekleri arasından yanlış sınıflandırılmış ya da tehlikeli olarak etiketlenen azınlık örnekler yardımıyla yukarı örnekleme yapar. Örneklerin tespiti için Borderline-SMOTE'de kullanılan k -en yakın komşu algoritması yerine,

destek vektör makinesi (DVM) algoritmasını kullanır. Çevresindeki çoğunluk sınıfa ait örneklerin yoğunluğuna bağlı olarak, her bir azınlık sınıfı destek vektörü ile belirlenen en yakın komşuları bir çizgi ile birleştirir. Bu çizgiler boyunca ya birleştirilen iki örnek arasında bir yerde ya da çizgi üzerinde bu iki örneğin dışında, fakat bu iki örneğe mesafe olarak çok yakın bir konumda, rastgele olarak sentetik veri üretir.

2.5.2.5. Uyarlanabilir Sentetik Örneklem (ADASYN)

Uyarlanabilir sentetik yukarı örneklem tekniği, He ve diğerleri (2008) tarafından geliştirilmiştir. Temel amacı, SMOTE örneklem tekniğinde olduğu gibi azınlık sınıfını kopyalamadan sentetik veri üreterek sınıfları dengelemektir. ADASYN algoritması, her bir azınlık sınıf örneği için oluşturulması gereken sentetik örneklerin sayısına yoğunluk dağılımı $\hat{r}_i$ ile otomatik olarak karar verebilir. ADASYN algoritması, dengeli olmayan veri setlerindeki azınlık ve çoğunluk sınıflarındaki örnekleri sayıca dengelemek için belli bir miktar sentetik örnek üretir. Sınıfları dengeleme işlemi için uygulanacak işlemlerin tamamı azınlık sınıfa ait örnekler üzerine uygulanır. Ancak bazen sentetik örneklerin üretileceği azınlık sınıfına ait orijinal gözlem noktasının öğrenilmesi zordur. Bunun sebebi, belirlenen komşuların çoğunun, çoğunluk sınıfına ait örneklerden oluşması ve bu örneklerin orijinal azınlık sınıfa ait örneklerle oldukça benzemesidir.

ADASYN algoritmasının temel amacı, yoğunluk dağılımına dayanarak düşük yoğunluğa sahip alanlardaki azınlık sınıfları için örnek üretmektir. Örnekleri üretmek için çoğunluk sınıfının daha yoğun olduğu alanlardaki azınlık sınıfına ait örnekleri tespit etmeye çalışır. Azınlık örneklerinin bulunduğu alanın düşük yoğunluğa sahip olması, bu tip örneklerin öğrenilmesini zorlaştırır. Buna bağlı olarak, bu tip örnekleri sınıflandırmak oldukça güçleşir. Bu nedendir ki ADASYN yukarı örneklem tekniği tüm dikkatini sınıflandırılması güç olan örnekler üzerine verir. Bundan hareketle azınlık sınıfa ait örnekler, öğrenilme zorluklarına göre farklı ağırlıklar verilir. Böylece, öğrenmesi zor olan azınlık sınıfa ait örnek için üretilecek sentetik örnek sayısı, öğrenmesi kolay azınlık sınıf ait örnek için üretilecek örnek sayısından daha fazla olacaktır ve örneklerin daha kolay öğrenilmesine yardımcı olacaktır.

Azınlık sınıfı için sentetik üretimi aslında iki ana aşama üzerine kuruludur; Düşük yoğunluğa sahip olan alandaki azınlık örneklerini tespit etmek ve bu örnekleri baz alarak SMOTE tekniği yardımıyla azınlık sınıfı için sentetik veri üretmektir. ADASYN tekniği ile sentetik örnek üretim aşamalarıyla ilgili detaylı bilgi aşağıda verilmiştir. Ancak, sentetik örnek üretme aşamalarından bahsetmeden önce tekniğin girdilerinden söz etmekte fayda vardır. Öncelikle D_{tr} , m tane örneğin bulunduğu bir eğitim veri setidir. x_i ise 1’den m ’ e kadar ($i = 1, \dots, m$) olan n boyutlu X uzayında yer alan bir örneği temsil etmektedir. $y_i \in Y = \{-1, 1\}$ ise x_i ‘ye karşılık gelen sınıf etiketlerini temsil eder. Bunlara ek olarak, m_k ve m_l ise sırasıyla azınlık ve çoğunluk sınıflarının sahip olduğu örnek sayısını temsil ettiğini varsayalım. Azınlık ve çoğunluk sınıflarında yer alan örneklerin toplamı, eğitim veri seti içerisinde yer alan toplam örnek sayısı olan m ’ i verir ($m_k + m_l = m$). Ek olarak, $m_k \leq m_l$ durumu ise azınlık sınıf için sentetik örnek üretilmesi gerektiğini gösterir.

ADASYN tekniğine ait sentetik örnek üretme aşamaları sırasıyla aşağıdaki gibidir.

Aşama 1: Öncelikle sınıflar arasındaki dengesizliğin oranı (2.20)’deki formül yardımıyla hesaplanır.

$$d = \frac{m_k}{m_l}, \quad d \in (0, 1] \quad (2.20)$$

Aşama 2: Eğer $d \leq d_t$ ise azınlık sınıf için üretilecek sentetik örneklerin sayısı (2.21)’de yer alan formülle hesaplanır. (d_t , veri setindeki dengesizliği tolere edebilmek için belirlenen maksimum eşik değerine karşılık gelir.)

$$G = (m_k - m_l) * \beta, \quad \beta \in [0, 1] \quad (2.21)$$

Formüldeki G , azınlık sınıfı için üretilecek toplam sentetik örneklerin sayısını ifade eder. Denklemden yer alan β parametresi ise uyarlanabilir sentetik algoritmasının (ADASYN) veri setine uygulandıktan sonra m_k ve m_l ‘nin istenilen denge seviyesini belirtmek için kullanılan parametredir. $\beta = 1$ durumunda, genelleme sürecinden sonra dengeli bir veri setinin elde edileceği sonucu çıkartılabilir.

Aşama 3: Azınlık sınıfına ait olan her x_i gözlem noktası için öklid mesafesine göre k -en yakın komşuları belirlenir ve (2.22)' deki formül yardımıyla r_i oranı hesaplanır.

$$r_i = \frac{\Delta_i}{K}, \quad r_i \in [0, 1], \quad i = 1, \dots, m_k \quad (2.22)$$

Formüldeki r_i , belirlenen komşular arasında çoğunluk sınıfın baskınlık oranının gösterir. r_i 'nin 1'e çok yakın bir değer çıkması, azınlık sınıfına ait bir gözlem noktası için belirlenen komşuların çoğunluk sınıfa ait gözlem noktalarından oluştuğu ve azınlık sınıfı örneklerini öğrenmenin zorlaştığı anlamına gelir. Ayrıca Δ_i , x_i 'nin k -en yakın komşuları arasında yer alan çoğunluk sınıfa ait örnek noktalarının sayısıdır. K ise seçilecek komşu sayısıdır.

Aşama 4: Bir önceki aşamada bulunan r_i oranı (2.23)'teki formül yardımıyla normalize edilir. Formüldeki $\hat{r}_i$ değeri yoğunluk dağılımını temsil eder. ($\sum_i \hat{r}_i = 1$)

$$\hat{r}_i = r_i / \sum_{i=1}^{m_k} r_i \quad (2.23)$$

Aşama 5: Her bir azınlık örneği x_i için oluşturulması gereken sentetik örneklerin sayısı (2.24)' teki formül yardımıyla hesaplanır. Formüldeki G , azınlık sınıfı için üretilecek toplam sentetik örneklerin sayısını ifade etmekteydi (Aşama2'de G değeri hesaplanmıştır).

$$g_i = \hat{r}_i * G \quad (2.24)$$

Aşama 6: Azınlık sınıfındaki her bir x_i için g_i sentetik örnekler üretilir. Daha sonra, x_i örneği için belirlenen komşuların arasından azınlık sınıfa ait bir x_{zi} gözlem noktası rastgele seçilir. Sentetik örnekler (2.25)' teki formül yardımıyla hesaplanır.

$$S_i = x_i + (x_{zi} - x_i) * \lambda \quad (2.25)$$

Formüldeki $(x_{zi} - x_i)$, n boyutlu uzaydaki fark vektörüdür. λ ise rastgele bir sayıdır ve 0 ile 1 arasında değer alır ($\lambda \in [0,1]$).

ÜÇÜNCÜ BÖLÜM

UYGULAMA

3.1. ÇALIŞMANIN AMACI VE KAPSAMI

Bankalar, kredi adı altında gerçek ve tüzel kişilere fon sağlayan en önemli finansal kuruluşlardır. Bankacılık sektörünün yüklendiği bu misyona paralel olarak karşılaşılabileceği en büyük risklerin başında ise kredi riski gelmektedir. Bu nedenle bankalar potansiyel müşterilerini kredilendirirken, temerrüt (kreditin geri ödenmemesi) potansiyeli yüksek olan müşterileri de doğru tespit etmesi gerekmektedir.

Günümüzde teknolojinin gelişmesiyle beraber, gelişen teknolojiyi en iyi kullanan sektörlerden biri de bankacılık sektörüdür. Teknolojinin gelişmesi, bankacılık sektöründe bir çığır açmış ve dünyanın dört bir yanında 7 gün 24 saat, toplumun her kesimine hizmet veren finansal kuruluşlar haline gelmiştir. Teknoloji sayesinde, bankacılık faaliyetlerine de getirilen kolaylıklarla beraber müşteriler, nerede olursa olsun tek bir tıkla bile kredi talep edebilme yetkinliği kazanmıştır. Bankalara farklı kanallardan erişilebilme kolaylığı, her geçen gün kredi talep eden müşterilerin sayısında da artışa neden olmuştur. Müşteri sayısındaki bu artış, bankaların kredilendirme faaliyetlerini elle yürütemeyecek hale getirmiştir. Bu durum, bankaların karar verme mekanizmalarını olumsuz etkilemiş ve kredi riskinin artmasına sebep olmuştur. Bankalar çareyi hem iş yükünü azaltacak hem de bir yığın müşterinin arasından kredi riskine sebep olacak müşterileri doğru bir şekilde ayıklayacak karar mekanizmaları geliştirmekte bulmuşlardır. Karar verme sürecini daha güvenilir ve daha pratik hale getirecek bu mekanizmaların temelinde ise, çağın en popüler bilim dalı olan yapay zekânın alt çalışma alanında yer alan makine öğrenmesi bulunmaktadır.

Çalışmanın amacı; kredi ihtiyacı olan müşteriler tarafından ibraz edilen bilgi ve belgeler ışığında, kredi veren bankaların hem karar verme mekanizmasını otomatikleştiren hem de kredi riskini azaltabilecek en iyi denetimli makine öğrenmesi tekniğini tespit etmektir. Tespit edilecek optimal tahmin performansına sahip sınıflandırma modeli sayesinde, kredi riski yüksek olan müşterilerin tespiti daha hızlı ve kolay yapılacaktır, aynı zamanda bankalara zaman ve maliyet açısından avantaj

sağlayacaktır. Bunun yanı sıra, kullanılacak makine öğrenmesi tekniklerinin tahmin performansını etkileyen dengeli olmayan veri setleri sorununa, yeniden örnekleme teknikleri uygulanarak bu çalışmada çözüm getirilmeye çalışılacaktır. Bu çalışmada ayrıca, farklı örnekleme teknikleriyle dengelenmiş kredi veri setleri üzerinde test edilen tekil ve topluluk öğrenmesi tekniklerine ait tahmin performansları karşılaştırılacak ve en yüksek performansa sahip olan makine öğrenmesi modeli, gelecekte meydana gelebilecek benzer problemlerin çözümü için önerilecektir. Bunun yanı sıra, topluluk öğrenmesi tekniklerinin içerisinde yer alan ağaç tabanlı sınıflandırma tekniklerinin, kredi riski bulunan müşterileri sınıflandırmada tekil sınıflandırma tekniklerine göre daha başarılı bir performans gösterip göstermediği de incelenecektir.

3.2. LİTERATÜR

Kredi riski veya skortlama problemleri ile Makine öğrenmesi tekniklerini birlikte kullanan literatürde sayısız çalışma bulunmaktadır. Ancak kredi riski ve makine öğrenmesi tekniklerini bir arada kullanan en eski çalışma, tüketici kredi veri setini kullanarak müşterilerin kredi almaya uygun olup olmadığını değerlendirmek için lineer diskriminant analizini kullanan Durand (1941)'a aittir. Bu çalışmanın ardından Altman (1968) ise, Z-skor modeli geliştirmek için değişken seçimi için çoklu diskriminant analizi kullanmıştır. İki farklı veri seti üzerinde çoklu diskriminant analizi sınıflandırma tekniğini kullanan Altman, kullandığı model, her iki veri seti için sırasıyla %96 ve %79 doğruluk oranları elde etmiştir. Altman'ın bu çalışmasından sonra Çoklu Diskriminant Analiz tekniği kullanılarak farklı sınıflandırma problemlerine çözüm için kullanılmaya devam edilmiştir.

90'ların sonlarına kadar tekil makine öğrenmesi tekniklerinde yaşanan devrimler, karar ağaçları (CART) (Grajski ve diğerleri, 1986), destek vektör makineleri (Cortes ve Vapnik, 1995) gibi tekniklerin yanında bir dizi yeni tekniğin de keşfiyle beraber başka bir boyuta taşınmış; Artık sınıflandırma problemleri için daha hızlı ve daha etkin çözümler elde edilmeye başlanmıştır. Değişim bu modellerle sınırlı kalmamış, tekil makine öğrenmesi tekniklerini etkilerinin bir araya getirilmesiyle daha güçlü ve daha hızlı tahmincilerin elde edilmesine de öncü olmuşlardır. Önce boosting

(Freund, 1995), bagging (Breiman, 1996) gibi teknikler ve sonrasında rastgele ormanlar algoritması (Breiman, 2001) olmak üzere çeşitli topluluk ve ağaç tabanlı teknikler ortaya çıkmıştır.

Bu bölümün devamında, 2000-2020 yılları arası kredi riski analizi üzerine yapılan çalışmalardan oluşan bir literatür araştırması yapılmıştır. Son 20 yıl içerisinde, gerek tez konusuna mümkün olduğunca yakın çalışmalar gerekse tez için kullanılacak teknikleri ve veri setlerini içeren çalışmalardan bahsedilmiştir. Literatür araştırması ile amaçlanan, yıllar geçtikçe kredi risk analizi yapan makine öğrenmesi tekniklerinde tahmin performansı açısından iyileşme olduğunu daha net gözler önüne sermektir. Ayrıca tez konusuna yakın çalışmaların deney sonuçları ile tezde yapılacak deneylerin sonuçlarını birbiriyle karşılaştırma fırsatı sunar. Kronolojik bir sırada, son 20 yıla ait çalışmalardan aşağıda bahsedilmiştir.

Fan ve Palaniswami (2000) çalışmasında, sınıflandırma problemini çözmek için destek vektör makineleri modelini kullanmıştır. DVM çekirdeğinde (kernel) öklid mesafesine dayalı bir kriter kullanılarak iflas durumunu tahmin eden modeli seçmenin yeni bir yolu sunulmuştur. Problem için üç geleneksel sınıflandırma modeli ve DVM yaklaşımına dayalı bir alternatif bir model kullanılarak karşılaştırmalı bir çalışma yapılmıştır. Kullanılan yöntemler, UCI *Machine Learning Repository Data Set*'e ait Avustralya kredi veri setine uygulanarak karşılaştırılmıştır. Çalışmanın sonucunda, DVM modeli diğer kullanılan üç yönteme göre daha başarılı bir performans göstermiştir.

Giudici (2001) çalışmasında, Markov Zinciri Monte Carlo hesaplama teknikleriyle birleştirilen bayes yöntemlerinin, kredi skorlama ve modelleri kıyaslama gibi faaliyetleri yüksek boyutlu karmaşık veri setinin analizinde de başarıyla gerçekleştirilebileceğini göstermiştir. Ayrıca kullanılan modeller Güney Almanya veri seti üzerinde uygulanmış ve model belirlemek ve çıkarımları yerelleştirmek için koşullu bağımsızlık grafikleri kullanmıştır. Böylece modelleme esnekliğinde ve hesaplamaların verimliliğinde önemli bir kazanç sağladığı sonucu ortaya çıkmıştır.

Lee ve diğerleri (2002) çalışmasında, geri yayılım sinir ağlarını geleneksel diskriminant analizi yaklaşımıyla birleştirerek modellerin kredi skorlarını tahmin performansı ölçülmüştür. Öngörülen hibrit yaklaşım, geleneksel sinir ağları modelinden çok daha hızlı bir performans göstermiştir. Önerilen hibrit yaklaşım ve

diğer geleneksel modeller UCI *Machine Learning Repository Data Set*'e ait Tayvan kredi veri setine uygulanmıştır. Ayrıca önerilen hibrit model sayesinde tahmin edilen kredi skorlarının doğruluğu artmıştır ve hibrit yaklaşım, geleneksel diskriminant analizinden ve lojistik regresyondan daha iyi performans göstermiştir.

Baesens ve diğerleri (2003) çalışmasında, 8 adet gerçek hayat kredi veri seti üzerinden sınıflandırma tekniklerinin performansı ölçülür. Kullanılan veri setlerinin bazıları, Benelüks (Belçika, Hollanda ve Lüksemburg İngilizce kısaltmaları) ve Birleşik Krallık (UK)'taki finansal kuruluşlardan temin edilmiştir. Diğer veri setleri ise UCI *Machine Learning Repository Data Set* ait Almanya ve Avustralya veri setleridir. Bu çalışmanın temel amacı, lojistik regresyon, karar ağaçları ve sinir ağları gibi temel sınıflandırma modelleri değerlendirmek ve birbiriyle karşılaştırmaktır. Aynı zamanda, destek vektörleri makineleri gibi çekirdek tabanlı sınıflandırma modeller ve en küçük kareler destek vektör makineleri (LS-SVM) modellerinin, sınıflandırma performansları ve uygunluğunu tespit etmektir. Çalışmanın sonucunda, en küçük kareler DVM ve sinir ağları sınıflandırma teknikleri performans olarak en iyi iki teknik olmuştur. Ayrıca lineer diskriminant analizi ve lojistik regresyon gibi temel sınıflandırma teknikleri de kredi skorlamada iyi bir performans göstermiştir.

Huang ve diğerleri (2004) çalışmasında, problemi açıklama gücüne sahip bir model sağlamak amacıyla, çalışmanın yapıldığı zamanda daha yeni keşfedilmiş bir makine öğrenmesi tekniği olan destek vektör makineleri (DVM) kullanmıştır. Daha sonra bu model, geri yayılım sinir ağını (GSA) modeli ile karşılaştırılmıştır. Amerika Birleşik Devletleri ve Tayvan pazarı için hem GSA hem de DVM yöntemleri için %80 civarında tahmin doğruluğu elde edilmiştir. Bununla birlikte, DVM' nin performansında az da olsa bir iyileşme görülmüştür.

West ve diğerleri (2005) çalışmaları için kredi skorlama ve kurumların iflas tahmini dahil olmak üzere finansal karar alanlarında karar destek sistemlerine daha doğru tahminleme yapan modeller kazandırmak için önemli araştırmalar yapılmıştır. Önceki çalışmaların odak noktası, basit parametrik modeller, veri yoğunluklarını doğrudan tahmin eden parametrik olmayan modeller ve sinir ağları gibi doğrusal olmayan model tanıma modellerini içeren çeşitli yöntemler arasından en iyi olan temel tahmin modelini belirlemek olduğu tespit edilmiştir. Ancak, yalnızca tek bir temel sınıflandırıcıya göre finansal kararlar vermek yanıltıcı olabileceği kanısına varılmıştır.

Bu nedenle birden fazla temel modeli bir arada kullanarak tahminleme yapmanın, tek bir modeli kullanarak tahminleme yapmaya göre çok daha doğru bir genelleme yeteneğine sahip olduğu öne sürülmüştür. Çalışmada, çapraz doğrulama (CV), torbalama (bagging) ve güçlendirme (boosting) gibi topluluk öğrenmesi tekniklerini kullanılırken, temel sınıflandırıcı olarak da sadece Çok katmanlı algılayıcı sinir ağı kullanılmıştır. İlgili modeller Avustralya, Almanya kredi ve iflas veri seti olmak üzere üç farklı gerçek hayat finansal veri seti üzerinde test edilmiştir. Sonuç olarak, sinir ağı topluluğunun genelleme yeteneğinin en iyi temel modelin genelleme yeteneğinden daha üstün olduğu tespit edilmiştir.

Huang ve diğerleri (2006) çalışmasında, Çin'deki Tanian isimli ticaret bankasından elde edilen veri setiyle, bankaya ait müşterilerin kredi sicillerini yine müşterilerin hesap bilgilerine göre tahmin etmeye çalışmaktadır. Bunun için birden farklı sınıflandırma algoritmaları kullanılmış ve problem için en etkili algoritmanın tespiti için bir dizi deney yürütülmüştür. Buna bağlı olarak, deneylerden çıkacak sonuçla beraber ileride banka gibi finansal kurum ve kuruluşların riski kontrol altına almak için aktif kullanabileceği ve doğru kararlar almasına yardımcı olacak algoritmanın keşfedilmesi amaçlanmıştır. Veri seti üzerinden daha hızlı ve doğru tahminler yapılabilmesi için temel bileşenler analizi (PCA) ile değişken indirgemesi yapılmıştır. Algoritma olarak, ağaç tabanlı ve temel algoritmaların yanı sıra, yapay sinir ağlarını çoklu algılayıcısı da kullanılmıştır. Çalışmada amaçlanan bir diğer durum, gerçek hayat problemlerine ait veri setlerinin sahip olduğu iki önemli sorun olan ayrık gözlem ve veri setinin içerdiği sınıf dengesizliği üzerine bir çözüm getirmektir. Bu sayede hem dengeli olmayan sınıfların hem de ayrık gözlemlerin etkisini azaltarak daha etkin tahminler elde etmeyi hedeflemektedirler. Çalışmada bu iki büyük sorunun çözümü için iki temel yöntem önermiştir. Yöntemlerden birincisi, sınıf özniteliklerinin sahip olduğu değerlere göre eğitim veri kümesindeki örneklerin rastgele seçilmesidir. Böylece, veri dağılımı her sınıf için eğitim veri setinde dengeli hale gelmesi sağlanmıştır. Yöntemlerden ikincisi ise ID3 sınıflandırıcı ve PRISM filtresinin kombinasyonunu kullanarak hibrit bir sınıflandırıcı oluşturmaktır. Bu sayede, her sınıf için tutarlı bir şekilde yüksek doğru pozitif (TP) tahmin oranları elde edilebilir olmasıdır. Sonuç olarak, sınıfların sayı olarak eşitlenmesiyle birlikte problem veri seti içerisinde gizli olan bilgi kolayca tespit edilebilir olduğu ortaya çıkmıştır.

Ayrıca çalışmada, gerçek hayat problemlerine ait dengeli olmayan veri setlerinde bir veri ön işleme yöntemi önerilmiştir.

Huang ve diğerleri (2007) çalışmasında, potansiyel müşterilerini kredilendirme sürecinde olan bankaların, sahip olduğu bilgilerden hareketle müşterilerinin kredi skorlarını belirlediği ve bu işi sezgisel olarak yapan bankalara bu süreçte iş yükünü hafifletecek bir yaklaşım sunmuşlardır. Çalışmanın yapıldığı dönemde, destek vektör makineleri (DVM) tekniği aktif olarak kullanılmaktadır. Buna istinaden, başvuru sahibinin girdi değişkenlerinden kredi skorlarını doğru değerlendirmek için hibrit DVM tabanlı kredi skorlama modellerini oluşturacak üç farklı strateji kullanmışlardır. UCI veri tabanındaki Avustralya ve Almanya olmak üzere iki kredi veri seti, DVM sınıflandırıcısının doğruluğunu göstermek için deney verisi olarak seçilmiştir. Sinir ağları, genetik algoritma ve karar ağacı sınıflandırıcılarıyla karşılaştırıldığında, DVM sınıflandırıcı nispeten az girdi değişken ile aynı sınıflandırma doğruluğuna ulaşmıştır. Tüm bunlara ek olarak, genetik algoritmaları DVM sınıflandırıcıyla birleştirerek önerilen hibrit GA-DVM stratejisi aynı anda özellik seçim görevini ve model parametrelerinin optimizasyonunu gerçekleştirmiştir. Deneylerden elde edilen sonuçlara dayanarak, DVM' nin mevcut makine öğrenmesi yöntemlerine önemli bir katkısının olduğunu göstermektedir.

Yeh ve Lien (2009) çalışmasında, Tayvan'daki temerrüte düşmüş müşterilerin altı farklı makine öğrenmesi modeliyle temerrüt olasılıklarının tahmin doğruluğunu karşılaştırmaktadır. Risk yönetimi açısından, tahmini temerrüt olasılığının doğruluğunun sonucu, güvenilir veya güvenilir olmayan şeklinde ikili değerlerin doğru bir şekilde sınıflandırılmasından elde edilen sonuca göre çok daha önemli olduğunu savunmuşlardır. Çalışmada, temerrüdün gerçek olasılığı bilinmediğinden bu çalışma aynı zamanda temerrüdün gerçek olasılığını tahmin etmek için yeni bir "sıralama düzeltme yöntemi" önermişlerdir. Veri seti olarak UCI *Machine Learning Repository Data Set*'e ait Tayvan kredi veri seti kullanılmıştır. k - en yakın komşu yöntemi, lojistik regresyon, diskriminant analizi, Naive Bayes, sinir ağları ve sınıflandırma ağaçları modelleri kullanılmış ve AUC performans ölçüm istatistiği ile test edilmiştir. Ayrıca altı makine öğrenmesi tekniği arasından gerçek temerrüt olasılığını doğru tahmin edebilen tek teknik, yapay sinir ağları olarak tespit edilmiştir.

Yu ve diğerleri (2010) çalışmasında, bireysel kredi risklerini değerlendirebilmek için, German Credit Data ve Australian Credit Approval olmak üzere iki adet açık kaynaklı veri kullanılmıştır. Lojistik regresyon, karar ağacı (C4.5), destek vektör makinesi ve sinir ağları algoritmalarını ilgili kredi verilerine uygulayarak karşılaştırmışlardır. Karşılaştırma sonucunda, Lojistik regresyon ve destek vektör makineleri algoritmaları diğer algoritmalara göre sınıflandırma konusunda yüksek başarı göstermişlerdir. Ancak, destek vektör makineleri diğer algoritmalara kıyasla daha yüksek sağlamlık ve genelleme yeteneği göstermiştir. Beklenenin aksine, sinir ağları algoritması sınıflandırmada zayıf performans göstermiştir.

Wang ve diğerleri (2011) çalışmasında hem istatistiksel hem de yapay zekâ tekniklerinin kredi skorlama üzerindeki performansları tespit edilmeye çalışılmıştır. Kredi skorlama için bir dizi literatür araştırması yapılmış bunun sonucunda kullanılan teknikler arasında topluluk öğrenmesi teknikleri temel tekniklere göre daha iyi bir performans sahip olabileceğini önermişlerdir. Bundan hareketle çalışmada, bagging (torbalama), boosting (artırma) ve stacking (yığınlama) gibi üç farklı topluluk öğrenmesi tekniği ile lojistik regresyon, karar ağaçları, yapay sinir ağları ve destek vektör makineleri gibi dört farklı tekil makine öğrenmesi tekniği kullanılmıştır. Çalışmanın amacı, tekil modeller ile topluluk öğrenmesi tekniklerini karşılaştırmaktır. Bu teknikler, Almanya, Avustralya ve Çin kredi veri setine uygulanmıştır. Almanya ve Avustralya veri setleri UCI *Machine Learning Repository Data Set*’e ait veri setleri olup, Çin kredi veri set ise Çin’ e ait temel bir ticari ve endüstriyel bankadan elde edilmiştir. Çalışmanın sonucunda, topluluk öğrenmesi modelleri tekil modellerden çok daha iyi performans göstermiştir. Topluluk öğrenmesi teknikleri kendi arasında kıyaslandığında, bagging yöntemi boosting yöntemine göre tüm veri setlerinde çok daha iyi performans göstermiştir.

Marqués ve diğerleri (2012) çalışmada, topluluk sınıflandırma modellerine çeşitlilik kazandırmak için farklı modelleri birlikte kullanan iki seviyeli sınıflandırıcı toplulukları oluşturmak ve bu oluşturulan modellerden iyi bir tahmin performansı elde etmeyi amaçlamıştır. Bu toplulukların oluşturulması için adaboost ve bagging gibi yeniden örnekleme teknikleri ile rastgele alt uzay ve rotasyon ormanları gibi değişken seçim yöntemlerinin farklı kombinasyonlarının tahmin performansını açısından iyi bir

sonuç verip vermediği araştırılmaktadır. Çalışmada, iki seviyeli sınıflandırıcı topluluklarının yanı sıra, k - en yakın komşu yöntemi, lojistik regresyon, destek vektör makineleri, C4.5 karar ağaçları, adaboost, bagging, rastgele alt uzay yöntemi, çok katmanlı algılayıcı sinir ağı ve rotasyon ormanları gibi temel sınıflandırma modelleri de kullanılmıştır. Oluşturulan tüm iki seviyeli sınıflandırma toplulukları ile temel sınıflandırma modelleri, altı farklı kredi veri seti kullanılarak sınıflandırma performansları test edilmiştir. Kullanılan veri setleri, UCI *Machine Learning Repository Data Set*'e ait Almanya, Avustralya, İran, Polonya, Japonya kredi veri setleridir. Ayrıca Kaliforniya Üniversitesi tarafından 2007 yılında düzenlenen veri madenciliği yarışmasında kullanılan UCSD veri setidir. Çalışmanın sonucunda, bu yeni iki seviyeli sınıflandırıcı topluluklarının kredi problemleri için uygun bir çözüm oluşturduğunu ve aynı zamanda hem geleneksel topluluk öğrenmesi tekniklerinde hem de temel sınıflandırma tekniklerinden çok daha iyi performans gösterdiği tespit edilmiştir.

Waad ve diğerleri (2013) çalışmasında, popüler olan kredi risk analizi ve skorlama problemleri hakkında birçok çalışmanın yapıldığını ve yapılan çalışmaların çoğunda teknikleri ve veri setini bilinçsiz şekilde kullanıldığını tespit etmişlerdir. Özellikle boyutu büyük olan veri setlerine herhangi bir ön işleme yapılmadan çeşitli sınıflandırma modellerinde test edilmeye çalışılması, tatmin edici olmayan sonuçların elde edilmesine sebep olabilir. Bundan hareketle, hem veri setindeki gereksiz değişkenleri ayıklamak için hem de sınıflandırıcı modellerin performanslarını artırabilmek adına temel filtreleme (filtre) ve sarmalıyıcı (wrapped) yöntemlere ek olarak, yeni bir füzyon yaklaşımı da sunmuşlardır. Veri seti olarak UCI *Machine Learning Repository Data Set*'e ait Almanya ve Avustralya'ya ait kredi veri setlerinin dışında ev kredisi (HMEQ) veri seti kullanılmıştır. On adet değişken seçim yöntemi bu veri setleri üzerinde kullanılmış ve hem Topluluk öğrenmesi modellerine hem de tekil modellere performans ölçümü yapılmıştır. Çalışmada kullanılan modeller, rastgele ormanlar algoritması, Naive Bayes, lojistik regresyon ve çok katmanlı algılayıcılar modelleri kullanılmıştır. Sonuç olarak, Almanya, Avustralya ve HMEQ veri setleri için genel olarak önerilen bir füzyon yaklaşımıyla en iyi sonuç elde edilmiştir.

Zurada ve diğerleri (2014) çalışmasında, kredi sağlayan kurum ve kuruluşların, potansiyel müşterilere kredi verme sürecindeki risklerini en aza indirmeyi amaçlamıştır. Bu nedenle, tespit edilecek makine öğrenmesi modelinin, potansiyel müşterinin temerrüte düşme durumunu doğru bir şekilde değerlendirmesi hedeflenmektedir. Hedefe ulaşmak için tekil makine öğrenmesi tekniklerini içeren beş farklı makine öğrenmesi tekniği kullanılmıştır. Bu teknikler; lojistik regresyon, sinir ağları, radyal temel fonksiyonlu sinir ağları, destek vektör makineleri ve karar ağaçları' dır. Çalışmada ayrıca geçek hayat problemine ait beş farklı veri seti kullanılmıştır. Kullanılan veri setleri gözlem sayısına, değişken sayısına ve türüne, eksik değerlerin sayısına ilave olarak yanıt değişkenindeki iyi kredi/kötü kredi sayısı açısından farklı özelliklere sahiptir. Çalışmada kullanılan veri setler; Almanya, Avustralya, SAS1, SAS2 ve Farmer kredi veri setleri kullanılmıştır. SAS1 ve SAS2 veri setleri, ev kredisi (HMEQ) veri setinden elde edilmiştir. Sınıflandırma doğruluğunu test etmek için ROC Eğrisi kullanılmıştır. Çalışmanın sonucunda, hiçbir yöntem diğer yöntemlere göre performans açısından üstünlük göstermemiştir. Ancak, gelecek dönemlerde elde edilecek aynı tip veri setlerine uygun modeller olabileceği kanısına varılmıştır.

Vinod Kumar ve diğerleri (2016) çalışmasında, Lendingclub adlı kişiler arası (P2P) çevrimiçi kredi alışverişine imkân tanıyan sistemin kredi riskini analiz etmişlerdir. Lendingclub.com adlı siteden temin edilen 2013-2015 yılları arasındaki veri seti üzerinde analizlerini yapılmıştır. Bu çalışmada veri setine eğitim modelleri olarak karar ağaçları, rastgele ormanlar ve bagging gibi topluluk tabanlı makine öğrenmesi algoritmaları ve değişken seçimi teknikleri kullanılmıştır. Algoritmalar, temerrütlü veya riskli kredileri belirlerken potansiyel iyi kredileri tercih edecek şekilde optimize edilmiştir. Rastgele ormanlar algoritmasının doğruluk oranı ve karar ağaçlarının kesinlik oranı yüksek çıkmıştır. Performans bakımından rastgele ormanlar, riskli kredileri belirlemede güçlü bir performans gösterirken, karar ağaçları ise iyi kredileri bulmada daha güçlü performans göstermektedir. Sonuç olarak, önerilen algoritmalar hem riskli hem de iyi kredileri belirlediğinden yatırım getirisi yüksek olacağına varılmıştır.

Abellán ve Castellano (2017) çalışmasında, Marqués ve Diğerleri (2012)'nin temel ve topluluk öğrenmesi tekniklerinin performansını belirlemek için yaptığı

çalışmayı baz alarak, yeni topluluk sınıflandırma teknikleri eklemiş ve bir önceki çalışmayı genişletmişlerdir. Çalışmalarında topluluk öğrenmesi modellerinin her birini farklı tekil modeller ile kurarak, en iyi performansı sağlayan topluluk öğrenmesi tekniği araştırılmıştır. Temel sınıflandırma modeli olarak kullanılan lojistik regresyon, destek vektör makineleri, C4.5 karar ağaçları ve çok katmanlı algılayıcı sinir ağı (MLP)'na ilave olarak güvenilir karar ağaçları modeli kullanılmıştır. Sayılan bu tekil modeller ile kurulan topluluk öğrenmesi teknikleri ise adaboost, bagging, rastgele alt uzay yöntemi ve rotasyon ormanlarına ek olarak DECORATE tekniği de kullanılmıştır. Önceki çalışmada kullanılan, UCI *Machine Learning Repository Data Set'e* ait Almanya, Avustralya, İran, Polonya, Japonya ve Kaliforniya Üniversitesi tarafından 2007 yılında düzenlenen Veri madenciliği yarışmasında kullanılan UCSD veri setleri bu çalışmada tekrar kullanılmıştır. Kullanılan her bir veri seti için 5 farklı topluluk öğrenmesi tekniği ile 5 farklı temel modelin birleşiminden toplamda 25 farklı sınıflandırma modeli oluşturulmuştur. Elde edilen sonuçlara göre temel sınıflandırma modellerini kredi skorlama ve iflas problemleri için topluluk öğrenmesi teknikleriyle birlikte kullanmanın farklı bir yaklaşım olarak sunulmuştur. Ayrıca topluluk öğrenmesi tekniklerinin tek bir temel modelle oluşturularak performansının ölçülmesi aslında seçilecek modelin kilit model olmadığını göstermiştir. Bu nedenle, kredi veri setleri üzerinde farklı topluluk öğrenmesi ve temel sınıflandırma modelleri kombinasyonlarını sınıflandırma test etmenin daha iyi sonuçlar sunabileceği bu çalışma sayesinde ispatlanmıştır.

Xia ve diğerleri (2018) çalışmasında, torbalama (bagging) ile yığınlama (stacking) yöntemlerinden oluşan heterojen bir kredi skorlama modeli oluşturulmuştur. Bu önerilen modele, bstacking adı verilmiştir. Bstacking tahmin performansını ölçmek için, Alman kredi veri seti, Avustralya kredi veri seti, eşler arası (P2P) çevrimiçi kredi platformları olan We.com ve Lendingclub.com sitelerinden temin edilen veri setleri üzerinde tahmin performansı ölçülmüştür. Daha sonra hem temel sınıflandırma hem de topluluk öğrenmesi yöntemleri ile karşılaştırılmıştır. Performans ölçüm metrikleri olarak; AUC, AUC-H metriği ve Brier skoru kullanılmıştır. Sonuç olarak, oluşturulan bstacking modeli diğer temel topluluk öğrenmesi yöntemlerinden daha iyi sonuç vermiştir.

Namvar ve diğerleri (2018) çalışmasında, Lendingclub adlı kişiler arası(P2P) çevrimiçi kredi sitesinden temin edilen veri setiyle potansiyel borçlunun kredilerini ödeyip ödemeyeceğini konusunda kredi risk tahmini yapılmıştır. Kullanılan veri seti dengeli olmayan veri setidir. Ayrıca 2016-2017 yıllarına ait 636 bin gözlem ve 145 değişken içerir. Çalışmanın amacı, kredi risk analizinde sıklıkla karşılaşılan dengeli olmayan veri setleri dengeli hale getirerek, en iyi performansa sahip sınıflandırma algoritmalarını tespit etmektir. Veri setini dengelemek için aşağı örnekleme, yukarı örnekleme ve bu iki yöntemi içeren hibrit teknikler kullanılmıştır. Dahası, sınıflandırma algoritması olarak, lineer diskriminant analizi, lojistik regresyon ve rastgele ormanlar algoritması kullanılmıştır. Sonuç olarak, rastgele ormanlar algoritması rastgele aşağı örnekleme yöntemi kullanılarak dengelenen veri setinde, diğer teknik ve modellere göre daha başarılı bir tahmin performansı göstermiştir.

Shen ve diğerleri (2019) çalışmasında, literatürdeki kredi risk değerlendirmesi çalışmalarını detaylı şekilde incelenmiş ve araştırmalar sonucunda geri yayılma (back propagation veya BP) sinir ağları tekniğinin kredi risk problemlerinde etkin sonuçlar verdiğini tespit etmiştir. Bunun yanı sıra, geleneksel makine öğrenmesi tekniklerinin topluluk sınıflandırıcı tekniklere göre daha zayıf kaldığını da gözlemlemiştir. Bundan yola çıkarak çalışmalarında bireysel kredi riskini değerlendirmek için SMOTE yeniden örnekleme tekniğine dayalı topluluk öğrenmesi modeli ve bir sınıflandırıcı optimizasyon tekniği önermişlerdir. Dengesiz veri setlerinin kredi değerlendirme modelinin performansı üzerindeki olumsuz etkiyi azaltmak için SMOTE tekniğini yanıt değişkenini dengelemek için kullanmıştır. Ayrıca parçacık sürüsü optimizasyon (PSO) algoritması, BP sinir ağlarındaki en iyi ağırlıkları ve sapmaları aramak için kullanılmıştır. Daha sonra optimize edilen BP sinir ağları sınıflandırma modeline bağlı olarak, AdaBoost topluluk öğrenme tekniğini geleneksel sınıflandırma teknikleriyle birleştirerek kendilerine has bir topluluk öğrenmesi modeli geliştirmişlerdir. Çalışmanın amacı, önerilen modelin diğer geleneksel modellere göre daha iyi performans gösterip göstermediğini araştırmaktır. Bu amaca bağlı olarak önerilen model performansını iyileştirmek için uygun parametre değerleri tespit edilmeye çalışılmıştır. Önerilen modeli sınıflandırma performansı açısından karşılaştırmak için sınıflandırma ağaçları, Naive Bayes, lojistik regresyon, en yakın komşu algoritması, destek vektör makineleri ve BP sinir ağları modeli kullanılmıştır. Tüm bu yöntemler

UCI *Machine Learning Repository Data Set* ait gerçek hayat problemlerine ait veri setleri olan Almanya ve Avusturalya veri setlerine uygulanmıştır. Çalışmanın sonucunda ise önerdikleri topluluk öğrenmesi tekniği diğer tüm tekniklerden genel olarak daha iyi bir performans gösterdiği tespit edilmiştir.

Alam ve diğerleri (2020) çalışmasında, modellerin yorumlanmasından ziyade sınıflandırma açısından en iyi performans gösteren modeli bulunması amaçlanmaktadır. Özellikle bankacılık sektörünün artan kredi kartı temerrüt riskini kolay tespit etmek ve tehlikeyi azaltmak için sınıflandırma performansı iyi bir modeli sunarak işleri kolaylaştıracağını savunmuşlardır. Çalışmada hem dengeli olmayan veri setlerini hem de dengeli hale getirdiği veri setleri için ayrı ayrı model tahmin performansları ölçülmüştür. Çalışmada birden fazla model kullanılmıştır. Aynı zamanda çalışmada, Güney Almanya, Tayvan ve Belçika olmak üzere üç farklı veri seti kullanılmıştır. İlk önce dengeli olmayan veri setleri için sınıflandırma yapılmıştır. Doğruluk oranları sırasıyla, Güney Almanya için %70, Tayvan için %66,9 ve Belçika içinse %65 olarak bulunmuştur. Daha sonra önerilen veri setini dengeli hale getirme teknikleri kullanılarak model tahmin performansı hesaplanmıştır. Modellerin doğru sınıflandırma oranlarında büyük ölçüde değişim olmuştur. Elde edilen doğruluk oranları, Güney Almanya için %87, Tayvan için %89 ve Belçika içinse %87,1 olarak bulunmuştur. Dengeli olmayan veri setleri ile karşılaştırıldığında sınıflandırıcılar dengeli veri setlerini sınıflandırmada daha yüksek başarı elde etmişlerdir. Ayrıca yukarı örnekleme tekniklerinin uygulandığı veri setleri, aşağı örnekleme tekniklerine göre sınıflandırma aşamasında daha fazla verim alınmıştır. Sonuç olarak gradient boosted karar ağaçları (GBDT) metodu diğer geleneksel sınıflandırma modellerine göre en iyi sonucu vermiştir. Dahası, veri setini dengelemek için k-means SMOTE tekniği kullanıldığında, gradient boosted karar ağaçları modelinin en iyi sonucu verdiği gözlenmiştir.

Li ve Chen (2020) çalışmasında, Amerika merkezli P2P olarak kredi sağlayan Lendingclub.com adresli siteden 2018 yılına ait kredi seti üzerinden kredi risk analizi yapmıştır. Çalışmanın amacı; kredi risk oranını düşürmek için kredi veri setini bazı topluluk ve geleneksel tekniklere uygulayarak en iyi performansı gösteren tekniği tespit etmektir. Böylece sınıflandırma performansı yüksek olan modeli belirleyerek risk taşıyan müşterileri kolayca belirlenmesine yardımcı olur. Çalışmada, 5 topluluk

öğrenmesi yöntemi ve 5 geleneksel yöntem kullanılmıştır. Topluluk öğrenmesi yöntemlerinden yığınlama (stacking), rastgele ormanlar, XGBoost, Extreme XGBoost ve AdaBoost yöntemleri kullanılmıştır. Geleneksel yöntemlerden ise Naive Bayes, karar ağaçları, destek vektör makineleri ve yapay sinir ağları kullanılmıştır. Çalışmada performans metriği olarak Doğruluk, AUC, Kolmogorov–Smirnov istatistiği ve brier skoru kullanılmıştır. Çalışmada kullanılan veri seti toplamda 36394 örnek ve 150 değişkenden oluşmaktadır. Ön işleme teknikleri ile veri setinde kayıp gözlemin bulunduğu satırlar ve önemsiz değişkenler veri setinden çıkarılmıştır. Topluluk öğrenmesi yöntemleri arasından en iyi performans rastgele ormanlar algoritması gösterirken, geleneksel yöntemler arasından en iyi performans gösteren lojistik regresyon yöntemi olmuştur. Ancak genel olarak bakıldığında hem geleneksel algoritmalar hem de diğer topluluk öğrenmesi yöntemleri arasından rastgele ormanlar algoritması en iyi performans gösteren yöntem olmuştur.

3.3. DENEY TASARIMI

Makine öğrenmesi teknikleri için en temel görev bilindiği üzere probleme en optimal çözümü bulmaktır. Bunun yolu ise problem çözümü için oluşturulacak sistematik bir deney tasarımıdır. Sistematik olarak oluşturulan bir deney tasarımının, problemin tanımlanmasından en optimal yöntemin tespitine kadarki süreci en iyi şekilde temsil etmesi oldukça önemlidir. Sürecin sistematik olarak planlanması, araştırmacının problemin çözümüne en doğru ve en hızlı şekilde ulaşmasına fayda sağlayacaktır. Böylelikle araştırmacının çözüm aradığı problem üzerine harcadığı efor ve maliyet açısından tasarruf etmesine yardımcı olacaktır. Deney tasarımının avantajı yalnızca zamandan veya maliyetten tasarruf etmek değildir. Deney tasarımı, probleme ne şekilde yaklaşılacağına ve problem için optimal çözüme kavuşturacak araç ve modellerin kullanılacağına karar vermede araştırmacıya kılavuzluk da edebilir. Başka bir deyişle, problemi anlamak ve probleme uygun araç ve modellerle süreci iyileştirmeye çalışmak ve devamında gelecek diğer tüm süreçlerde de benzer problemlerin çözümünde rehber olarak kullanılmasına olanak sağlamaktadır. Dahası, çalışmalara ait deney tasarımının şeffaf bir şekilde sunulması, araştırmacının probleme çözüm bulma sürecinde farklı fikirler ortaya atmasına katkı sağlayacaktır. Bu durum,

araştırmacının ele aldığı benzer veya farklı problem türleri için oluşturacağı deney tasarımına dahil edeceği araç ve model yelpazesini çeşitlendirmesine imkân tanıyacaktır.

Bu çalışma için sekiz farklı uygulama senaryo oluşturulmuştur. Uygulama senaryolarının içerikleri; ilgili veri setine değişken seçiminin uygulanıp uygulanmasına ve veri setinin dengelendiği yeniden örnekleme tekniklerine göre farklılık göstermektedir. Oluşturulan senaryoların içeriği Tablo 2’deki gibidir.

Tablo 2: Uygulama Senaryoları

Uygulama Senaryoları	Çalışmada Kullanılan Her Bir Kredi Veri Seti için Oluşturulan Uygulama Senaryolarının İçeriği
<i>Senaryo 1</i>	Değişken Seçimi ve Aşağı Örnekleme Teknikleriyle Topluluk Öğrenmesi Modellerinin Tahmin Performansı
<i>Senaryo 2</i>	Değişken Seçimi ve Aşağı Örnekleme Teknikleriyle Tekil Modellerin Tahmin Performansı
<i>Senaryo 3</i>	Tüm Değişkenler ve Aşağı Örnekleme Teknikleriyle Topluluk Öğrenmesi Modellerinin Tahmin Performansı
<i>Senaryo 4</i>	Tüm Değişkenler ve Aşağı Örnekleme Teknikleriyle Tekil Modellerin Tahmin Performansı
<i>Senaryo 5</i>	Değişken Seçimi ve Yukarı Örnekleme Teknikleriyle Topluluk Öğrenmesi Modellerinin Tahmin Performansı
<i>Senaryo 6</i>	Değişken Seçimi ve Yukarı Örnekleme Teknikleriyle Tekil Modellerin Tahmin Performansı
<i>Senaryo 7</i>	Tüm Değişkenler ve Yukarı Örnekleme Teknikleriyle Topluluk Öğrenmesi Modellerinin Tahmin Performansı
<i>Senaryo 8</i>	Tüm Değişkenler ve Yukarı Örnekleme Teknikleriyle Tekil Modellerin Tahmin Performansı

Oluşturulan bu senaryolar Alman, Avustralya ve HMEQ kredi veri seti olmak üzere dengeli olmayan üç kredi veri seti üzerinde uygulanmış ve ilgili uygulama senaryolarının makine öğrenmesi modellerinin riskli müşterileri tahmin etme performansına etkileri ölçülmüştür. Bunun yanı sıra, senaryoları test etmek için kullanılan denetimli makine öğrenmesi modelleri, tekil ve topluluk öğrenmesi modelleri olarak ayrılacak ve kullanılan tekil modeller ile topluluk öğrenmesi modellerinin (özellikle ağaç tabanlı olan modellerin) tahmin performansları birbirleriyle karşılaştırılacaktır. Çalışmanın sonucunda ise, belirlenen uygulama senaryoları arasında optimal sonucun elde edildiği senaryo belirlenecek ve gelecekte bankacılık sektöründe yaşanabilecek aynı tipteki problemlerin çözümü için genelleştirilecek ve önerilecektir.

Çalışmada veri setini dengelemeye yardımcı RUS, tomet-links, cluster centroids, near miss olmak üzere dört aşağı örnekleme (undersampling) ve ADASYN, ROS, SMOTE, SVMSMOTE, Borderline-SMOTE olmak üzere beş yukarı örnekleme (oversampling) tekniği kullanılmıştır. Bunun yanı sıra, dengeli hale getirilen veri setleri üzerinde tahmin performansı ölçülen 11 adet denetimli makine öğrenmesi

sınıflandırma tekniği kullanılmıştır. Kullanılan sınıflandırma tekniklerinin 5'i topluluk öğrenmesi tekniklerine, geriye kalan 6'sı ise tekil sınıflandırma tekniklerine aittir. Kullanılan 5 topluluk öğrenmesi tekniği arasından 3'ü ise ağaç tabanlıdır. Kullanılan topluluk öğrenmesi teknikleri; rastgele ormanlar (RF), gradyan artırma karar ağaçları (GBDT), ekstra ağaçlar (XTrees), yumuşak ve sert oylamadır. RF, GBDT ve XTrees sınıflandırma teknikleri, çalışmada kullanılan ağaç tabanlı topluluk öğrenmesi teknikleridir. Kullanılan tekil sınıflandırma teknikleri ise lojistik regresyon, k- en yakın komşu (KNN), Naive Bayes (NB), karar ağaçları (DT), çok katmanlı algılayıcı sinir ağı (MLP) ve destek vektör makineleri (SVM)' dir. Ayrıca, modellerin tahmin performanslarını ölçmek için 5 farklı performans ölçüm istatistiği kullanılmıştır.

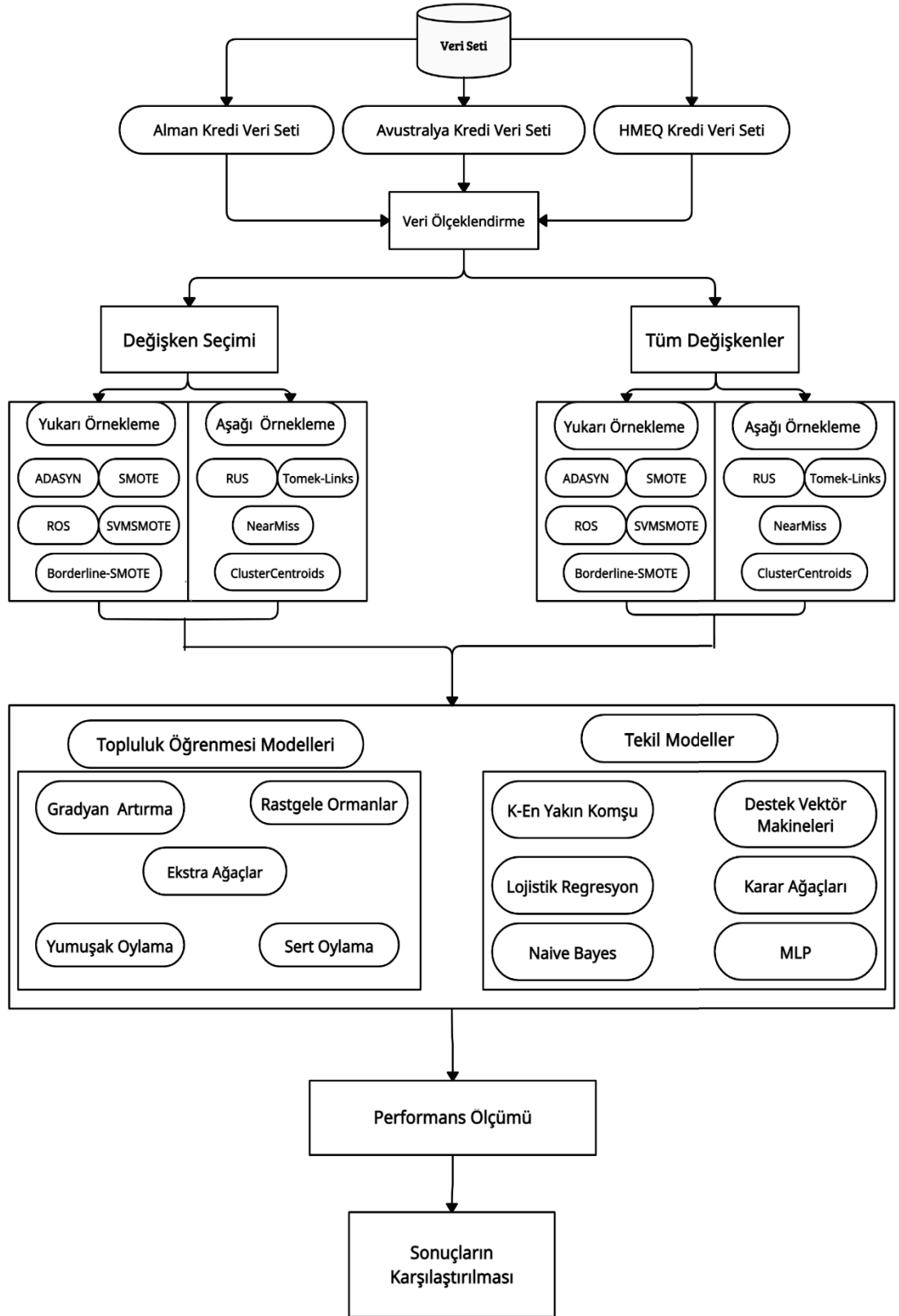
Uygulamaya ait makine öğrenmesi süreçleri 7 adımda özetlenmiştir. Bu 7 adım sırasıyla şöyledir;

- Problemin Tanımlanması
- Veri Setlerini Belirlenmesi
- Veri Ön İşleme Tekniklerinin Uygulanması
- Değişken Seçimi Yapılması
- Modellerin Belirlenmesi ve Kurulması
- Modellerin Performansının Ölçülmesi
- Bulguların Raporlanması ve Yorumlanması

Yukarıda yer alan makine öğrenmesi süreçlerine ait adımlar, çalışmanın ilerleyen kısımlarında ayrı başlıklar haline getirilmiş ve bu adımların, yürütülen uygulama için öneminden kısaca bahsedilmiştir. Bunların yanı sıra, ilgili adımlarda, yürütülen uygulama kapsamında hangi faaliyetlerin gerçekleştirildiğinden ve gerçekleştirilen bu faaliyetler için hangi tekniklerin kullanıldığının yanı sıra, bu kullanılan tekniklerin ne amaçla kullanıldığına da detaylıca değinilmiştir.

Bu çalışma için yürütülen uygulama kapsamında gerçekleştirilen faaliyetler ve bu faaliyetlerin gerçekleştirilmesi için kullanılan teknikleri adım adım gösteren faaliyet akış diyagramı Şekil 17'de gösterildiği gibidir. Bu faaliyet akış diyagramı, aynı zamanda yukarıda bahsedilen makine öğrenmesi süreçlerine ait her bir adımda gerçekleştirilen faaliyetleri özetler niteliktedir.

Şekil 17: Makine Öğrenmesi Süreçlerine ait Faaliyet Akış Diyagramı



3.3.1. Veri Setleri

Çalışmada, kredi veri seti olarak üç adet açık kaynaklı veri kullanılmıştır. Veri setinin ikisi; Almanya ve Avustralya'ya ait kredi veri setleridir. Bu iki veri seti, *UCI Machine Learning Repository* adlı açık kaynaklı veri setlerinin bulunduğu bir internet sitesinden temin edilmiştir. Diğer veri seti ise HMEQ ev kredisi (Mortgage) veri setidir. HMEQ kredi veri seti ise “Kaggle” adında bir veri bilimi platformu olan internet sitesinden elde edilmiştir. Bir sonraki kısımda bu veri setlerinden taşıdığı değişken sayısı, değişkenlerin hangi örnek tiplerine (kategorik ve sayısal) sahip olduğundan detaylıca bahsedilmiştir.

3.3.1.1. Avustralya Kredi Veri Seti

Avustralya kredi veri seti, *Statlog Australian Credit Approval Data* adıyla *UCI Machine Learning Repository* adlı veri tabanında yer alan bir gerçek hayat problemine ait bir veri setidir. 690 adet kredi başvurusu içeren bu veri seti, 15 değişkenden oluşur. Değişkenlerin 7'si kategorik, 8'i ise sayısal örnek tiplerine sahiptir. Yanıt değişkeni ikili kategorik değerlerden oluşur (kredi alabilir/kredi alamaz). Avustralya kredi veri seti dengeli bir veri seti değildir. Kredi için başvuran müşterilerin 307'si “kredi alabilir” olarak etiketlidir. Geriye kalan 383'ü ise “kredi alamaz” olarak etiketlidir. Veri setinde bulunan 14 girdi değişkenin içerdiği örnek tipleri haricinde hakkında herhangi bir bilgi verilmemiştir. Bu sebeple, değişkenler; A1, A2, ..., A15 şeklinde isimlendirilmiştir.

3.3.1.2. Alman Kredi Veri Seti

Alman kredi veri seti, *Statlog German Credit Data* adıyla *UCI Machine Learning Repository* adlı veri tabanında yer alan bir gerçek hayat problemine ait veri setidir. 1000 adet kredi başvurusu içeren bu veri seti 21 değişkenden oluşur. Değişkenlerin 13'ü kategorik, 7'si ise sayısal örnek tiplerine sahiptir. Yanıt değişkeni ikili kategorik değerlerden oluşur (kredi alabilir/kredi alamaz). Almanya kredi veri seti dengeli bir veri seti değildir. Kredi için başvuran müşterilerin 700'ü “kredi alabilir” olarak etiketlidir. Geriye kalan 300'ü ise “kredi alamaz” olarak etiketlidir. Veri setinde bulunan 20 girdi değişkeni, banka müşterilerinin kredi geçmişini, hesap bakiyelerini, kredi amacını, kredi miktarını, istihdam durumunu, kişisel bilgileri, yaşı, barınma ve

iş unvanını içermektedir. Veri setinde yer alan değişkenler hakkında detaylı bilgi Tablo 3'teki gibidir.

Tablo 3: Alman Kredi Veri Setine Ait Değişkenler

Değişken İsimleri	Açıklamaları
<i>Status</i>	Borçlunun Bir Bankadaki Hesap Durumu
<i>Duration</i>	Ay Cinsinden Kredi Vadesi
<i>Credit History</i>	Önceki veya Mevcut Anlaşmaları
<i>Purpose</i>	Kredi Başvuru Amacı
<i>Amount</i>	Aldığı Toplam Kredi Miktarı
<i>Savings</i>	Borçlunun Sahip Olduğu Toplam Tasarruflar
<i>Employment Duration</i>	Borçlunun Mevcut İşindeki Çalışma Süresi
<i>Installment Rate</i>	Kredi İçin Belirlenen Taksit Oranı
<i>Personal Status- Sex</i>	Borçlunun Medeni Durumu- Cinsiyeti
<i>Other Debtors</i>	Kredi İçin Kefil Durumu
<i>Present Residence</i>	Mevcut İkamet Adresinde Yaşadığı Süre
<i>Property</i>	Borçlunun Mülkünün Artan Düzende Sıralaması
<i>Age</i>	Borçlunun Yaşı
<i>Other Installment Plans</i>	Kredi Alınan Banka Dışında Borçlunun Taksit/Kredi Yükü
<i>Housing</i>	Oturduğu Evin Durumu (Kira, Kendi Evi, Kira Ödemiyor)
<i>Number Credits</i>	Aldığı Tüm Kredilerin Geçmişi
<i>Job</i>	Borçlunun İş Düzeyi
<i>People Liable</i>	Finansal Olarak Bakmakla Yükümlü Olduğu Kişi Sayısı
<i>Telephone</i>	Borçlunun Adına Kayıtlı Bir Telefon Var mı?
<i>Foreign Worker</i>	Borçlu Yabancı Bir İşçi Mi?
<i>Credit Risk</i>	Kredi Riski Var/Yok

Kaynak: Alam ve diğerleri, 2020: 201178

3.3.1.3. HMEQ Kredi Veri Seti

Ev sermayesi kredi veri seti (HMEQ), ev kredisi (Mortgage) almak için 5960 adet başvuruyu içeren bir gerçek hayat problemine ait bir veri setidir. Veri seti, toplamda 12 değişkene sahiptir. Değişkenlerin 3'ü kategorik, 9'u ise sayısal örnek tiplerine sahiptir. Yanıt değişkeni ikili kategorik değerlerden oluşur. Kredi için başvuran müşterilerin 1189 'u "temerrüte düşen (kredi alamaz)" olarak etiketlenmiştir. Geriye kalan 4771 müşteri ise "kredilerini düzenli olarak ödemiş (kredi alabilir)" olarak etiketlenmiştir. Veri setindeki 11 girdi değişkeni, müşterilerin temerrüt durumlarıyla beraber azı mali ve kişisel bilgileri de içermektedir. Veri setine ait tüm değişkenler ve detaylı açıklamaları Tablo 4'te gösterildiği gibidir.

Tablo 4: HMEQ Kredi Veri Setine Ait Değişkenler

Değişken İsimleri	Açıklamaları
<i>Loan</i>	Müşterinin Talep Ettiği Kredi Tutarı
<i>Mortdue</i>	Müşterinin Ödemesi Gereken Ev Kredisinin (Mortgage) Tutarı
<i>Value</i>	Müşterinin Mevcut Mülkünün Değeri
<i>Yoj</i>	Müşterinin Mevcut İşinde Çalıştığı Toplam Yıl Sayısı
<i>Derog</i>	Müşterinin Temerrüte Düşme Sayısı
<i>Delinq</i>	Müşterinin Ödemediği Kredi Taksiti Sayısı
<i>Clage</i>	Müşterinin En Son Aldığı Kredinin Taksit Süresi (Ay Cinsinden)
<i>Ninq</i>	Müşteriyle İlgili En Son Yapılan Kredi Sorgulama Sayısı
<i>Cln</i>	Kredi Taksit Sayısı
<i>Debtinc</i>	Borç/Gelir Oranı
<i>Reason</i>	Müşterinin Krediyi Alma Sebebi (Mali Destek, Ev Yenileme)
<i>Job</i>	Müşterinin Mesleği
<i>Bad</i>	Risk Yok / Riskli

3.3.2. Veri Ön İşleme

Gerçek hayat problemlerine ait olan veri setleri hem araştırmacı hem de uygulanacak model için karmaşık yapılara sahiptir. Bu karmaşıklık, probleme ait veri setinin yapısında bulunan önemli desenlerin tanımlanabilmesine engel olabilir. Buna bağlı olarak, modelin veri setini öğrenmesi güçleşebilir ve problem için optimal çözümün bulunması noktasında büyük sıkıntılar yaşanabilir. Problem için elde edilen veriler, genelde işlenmemiş (ham) verilerdir. Modelin, bu tip verilerden analiz yapması oldukça güçtür. Bu nedenle, karmaşıklığı gidermek ve yorumlanabilir yapıları öne çıkarmak adına, veri setine ön işleme yöntemleri uygulanır. Bu yöntemler; veri bütünleştirme, temizleme, dönüştürme ve indirgeme yöntemleridir.

Çalışmada kullanılan veri setlerine uygulanan bazı ön işleme yöntemleri aşağıdaki gibidir;

- Eksik Değişken Değerlerinin Tamamlanması
- Değişkenlerin Kodlanması
- Eğitim ve Test Veri Setlerinin Belirlenmesi

- Veri Dönüşümünün Yapılması
- Değişken Seçiminin Yapılması
- Yeniden Örnekleme Yöntemlerinin Uygulanması

Kullanılan ön işleme yöntemleri hakkında detaylı bilgi, bir sonraki bölümde yer almaktadır.

3.3.2.1. Eksik Değişken Değerlerinin Tamamlanması

Gerçek hayat problemlerine ait veri setlerinde eksik değerlere sahip değişkenlerle karşılaşmak olasıdır. Verinin gizliliğinden veya örneğin kaybindan dolayı oluşan bu sorun modellenen problem için optimal sonuçlar elde etmeyi zorlaştırır. Bu nedenle, eksik gözlemle başa çıkabilmek için bazı çözüm yöntemleri geliştirilmiştir. Bu çözüm yöntemlerinden en popüler iki yöntem, eksik gözlemlerin veri setinden silinmesi ve değişkenlere ait mod, medyan ve ortalama gibi tanımlayıcı istatistikleri ile eksik değişken değerlerinin tamamlanmasıdır. Eksik değişken değerlerine sahip gözlemleri veri setinden çıkarmak, bazı önemli bilgilerin kaybolmasına sebebiyet verebilir. Bu da kullanılacak tahmin modelinin, problem için iyi bir tahminleme performansı göstermesine engel olacaktır. Bu sebeple mod, medyan veya ortalama ile eksik değişken değerlerini tamamlamak hem bilgi kaybına sebebiyet vermeyecek hem de veri setine eklenen yeni değişken değerleri, veri setinde yer alan diğer değerlerle kıyaslandığında bir aykırılık göstermeyecektir.

Kullanılan veri setleri içerisinde, sadece HMEQ kredi veri seti eksik değişken değerlerine sahiptir. HMEQ kredi veri setinde toplamda 5271 tane eksik değişken değeri bulunur. Eksik olan bu değerler, yer aldığı değişkenin sahip olduğu örnek tiplerine göre mod ve ortalama değerleri ile doldurulmuştur. Değişken sayısal örnek tiplerinden oluşuyorsa, değişkenlerin ortalama değerleri kullanılmıştır. Eğer değişken kategorik örnek tiplerinden oluşuyorsa, mod değerleri kullanılmıştır. Veri setinde yer alan; LOAN, MORTDUE, VALUE, YOJ, DEROG, DELINQ, CLAGE, NINQ, CLNO, DEBTINC değişkenleri, sayısal örnek tiplerine sahiptir. Bu değişkenlerin sahip olduğu eksik değerlerden her biri, değişkenin sahip olduğu ortalama değerleriyle tamamlanmıştır. REASON ve JOB değişkenleri kategorik örnek tiplerine sahip olduğundan, değişken içerisinde en çok tekrar eden değişken değeriyle yani mod değerleri ile tamamlanmıştır.

3.3.2.2. Değişkenlerin Kodlanması

Gerçek hayat problemlerine ait veri setleri; sayısal (numeric), kategorik (nominal) ve sıralı (ordinal) olmak üzere birçok veri tipini bünyesinde barındırabilir. Ancak, kategorik ve sıralı veri tipleri genellikle metin (string) türünde oluşturulmaktadır. Bu nedenle, kullanılacak makine öğrenmesi modellerine, metin türünde olan veri tiplerini entegre etmek oldukça zordur. Bunun sebebi, makine öğrenmesi modellerinin sayısal tabanlı işlemler yapmasından dolayı metin türündeki veri tiplerini benimseyememesidir. Buna dayanarak, metin türünde kategorik ve sıralı veri tipine sahip değişkenlere, sayısal değerler atanması gerekir. Bu sayısal değer atama işlemine kodlama (encoding) denir. Kodlama işlemi, değişken içerisinde yer alan metin türündeki verilerin, hiçbir bilgi kaybı yaşanmadan, sayısal değerlere atanmasıdır (Örneğin; Öğretmen için 1, Avukat için 2, Doktor için 3 gibi).

Python'a ait Scikit-Learn kütüphanesinde, kategorik ve sıralı veri tiplerine sahip değişkenler için kodlama işlemleri farklı şekilde gerçekleştirilir. Değişkende yer alan veriler sıralanabiliyorsa bu tip veriler sıralıdır (Örneğin, üniversitedeki harf notları). Bu veri tipleri için Scikit-Learn kütüphanesindeki "*OrdinalEncoder()*" fonksiyonu kullanılmaktadır. Değişkende yer alan veriler arasında bir ilişki yoksa veya sıralanamıyorsa bu tip veriler kategorik verilerdir. Bu veri tipleri için Scikit-Learn'deki "*OneHotEncoder()*" fonksiyonu kullanılır. One Hot Encoder fonksiyonu, değişken içerisinde yer alan sınıfları yeni bir değişkene dönüştürür. Örneğin, meslek türlerinin yer aldığı değişken olsun. Bu değişken içerisinde, Öğretmen, Doktor ve Avukat olmak üzere üç meslek grubu olsun. Fonksiyon, Meslek gruplarına ait metin türündeki verilerin her biri için meslek_öğretmen, meslek_doktor, meslek_avukat gibi sayısal veri tiplerine sahip farklı değişkenler oluşturur. Bu sayede kategorik veriler sayısallaştırılarak, makine öğrenmesi modellerine daha kolay entegre edilebilir. Ek olarak, Python'da kategorik veri tiplerini sayısal hale getiren diğer bir fonksiyon "*LabelEncoder()*" fonksiyonudur. LabelEncoder fonksiyonu, genellikle Evet-Hayır gibi ikili(binary) kategorik yapıda bulunan değişkenleri, 0-1 şeklinde sayısal hale getirir.

Çalışmada kullanılan veri setleri için değişken türlerine göre farklı kodlama fonksiyonları kullanılmıştır. Her bir veri seti için değişkenin sahip olduğu veri tiplerine göre hangi kodlama fonksiyonu kullanıldığına dair bilgiye aşağıda yer verilmiştir;

- Alman Kredi Veri Seti: Veri setinde 10 adet kategorik (nominal), 3 adet sıralı(Ordinal) veri tipine sahip değişkenler bulunmaktadır. Kategorik (nominal) olarak belirlenen değişkenler; Status, Purpose, Savings, Personal Status-Sex, Other Debtors, Property, Other Installment Plans, Housing, Foreign Worker, Telephone. Bu değişkenlere, *OneHotEncoder()* fonksiyonu uygulanmıştır. Sıralı (Ordinal) olarak belirlenen değişkenler; Employment Duration, Credit History, Job. Uygulama yapılırken, tüm bu değişkenlere, *OrdinalEncoder()* fonksiyonu uygulanmıştır. Veri setinin yanıt değişkeni, Credit Risk değişkenidir. Bu değişken, ikili kategorik veri tipine sahiptir. Bu nedenle, *LabelEncoder()* fonksiyonu uygulanmıştır.
- Avustralya Kredi Veri Seti: Veri setinde toplamda 8 kategorik değişken vardır. Bu değişkenler hakkında detaylı bilgi bulunmamaktadır. Sadece değişkenlerin veri tipi bilinmektedir. Uygulama yapılırken, kategorik olan tüm değişkenlere *OneHotEncoder()* fonksiyonu uygulanmıştır. Veri setinin yanıt değişkeni ikili kategorik veri tipine sahiptir. Bu nedenle, veri setinin yanıt değişkenine *LabelEncoder()* fonksiyonu uygulanmıştır.
- HMEQ Kredi Veri Seti: Veri setinde toplam 2 adet kategorik değişken vardır. Kategorik olarak belirlenen değişkenler; REASON ve JOB değişkenleridir. Uygulama yapılırken, tüm bu değişkenlere *OneHotEncoder()* fonksiyonu uygulanmıştır. Veri setinin yanıt değişkeni, BAD değişkenidir. BAD değişkeni ikili kategorik veri tipine sahiptir. Bu nedenle, veri setinin yanıt değişkenine *LabelEncoder()* fonksiyonu uygulanmıştır.

Çalışmada kullanılan veri setlerinin sahip olduğu değişken sayıları, Tablo 5’te verilmiştir. Bu tabloda, orijinal veri setindeki toplam değişken sayısı ile değişken kategorik kodlama (encoding) fonksiyonu uygulandıktan sonraki toplam değişken sayıları verilmiştir.

Tablo 5:Kategorik Kodlama Öncesi ve Sonrasındaki Toplam Değişken Sayısı

<i>Veri Setleri</i>	<i>Orijinal Veri Setindeki Toplam Değişken Sayısı</i>	<i>Kodlama Sonrası Toplam Değişken Sayısı</i>
<i>Alman Kredi Veri Seti</i>	20	50
<i>Avustralya Kredi Veri Seti</i>	14	42
<i>HMEQ Kredi Veri Seti</i>	12	18

3.3.2.3. Veri Dönüşümünün Yapılması

Veri dönüşümü, veri setindeki değerleri, belli bir aralıkta yer alan değerlere çeviren bir ön işleme yöntemidir. Veri dönüşümü, aykırı ve yüksek değerlere sahip değişkenlerin etkisi azaltmaya yardımcı olur. Bu sayede, yüksek gözlem değerine sahip değişkenlerin tahmin üzerindeki baskınlığını azaltarak, modelin tahmin performansını iyileştirir. Çalışmada kullanılan veri dönüştürme yöntemi Standartlaştırma işlemidir. Standartlaştırma işlemi, değişkenlerin ortalaması 0 varyansı 1 olması kriterine dayanarak, değişken değerlerini normal dağılıma yaklaştıran bir ön işleme yöntemidir. Çalışma için kullanılan üç veri setine de standartlaştırma işlemi uygulanmıştır. Ek olarak, standartlaştırma işlemi scikit-learn’de “*preprocessing.StandardScaler()*” fonksiyonu kullanılarak yapılmaktadır.

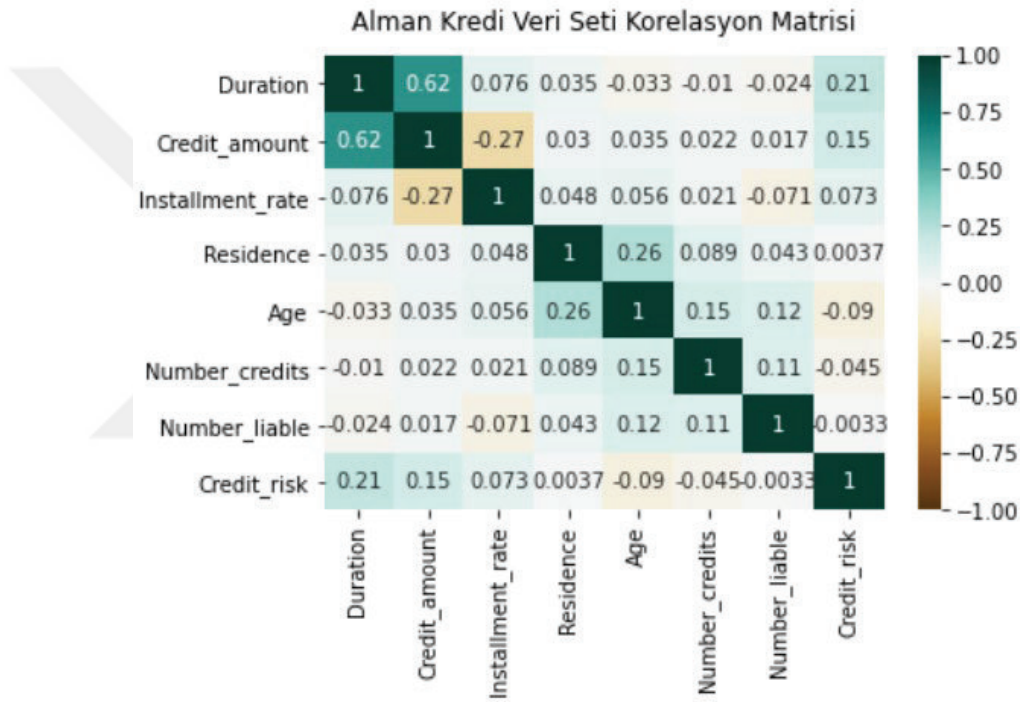
3.3.3. Veri Setlerinin Keşfi

Modelin, probleme ait veri setini anlaması ve anladıklarından çözüme yardımcı yapılar çıkarmasının öneminden sıklıkla bahsedilmektedir. Ancak, iş sadece bununla sınırlı değildir. Kullanılacak model kadar araştırmacının da problemi anlaması ve bazı çıkarımlar yapması gerekir. Çünkü model, sadece sayısal veya kategorik bir değer verir, fakat değer ne anlama geldiği konusunda bir fikir vermez. Bu değeri problemle ilişkilendirmek ve yorumlamak ise araştırmacıya düşer. İşte bu yüzden, araştırmacının da tüm bu süreçleri başarıyla yürütmesi için probleme hâkim olması gerekir. Araştırmacı için problemin özünü anlamak ve probleme ne şekilde yaklaşacağını belirlemek ise veri setini keşfetmesinden geçer. Araştırmacı, veri setini keşfetmek için

keşifçi veri analizini kullanır. Keşifçi veri analizi, temel istatistiklerle yakalayamayacağımız bilgileri daha gözle görülür hale getirir. Bunun için, histogram ve daire grafiği gibi farklı türde veri görselleştirme araçları kullanılır. Bu çalışmada kullanılan Alman ve HMEQ kredi veri setlerini anlayabilmek için çeşitli grafikler kullanılmış ve yorumlanmıştır. Avustralya veri setine ait değişkenlerin açıklamaları olmadığından keşfetmek zordur. Bu nedenle, görselleştirme araçları kullanılmamıştır.

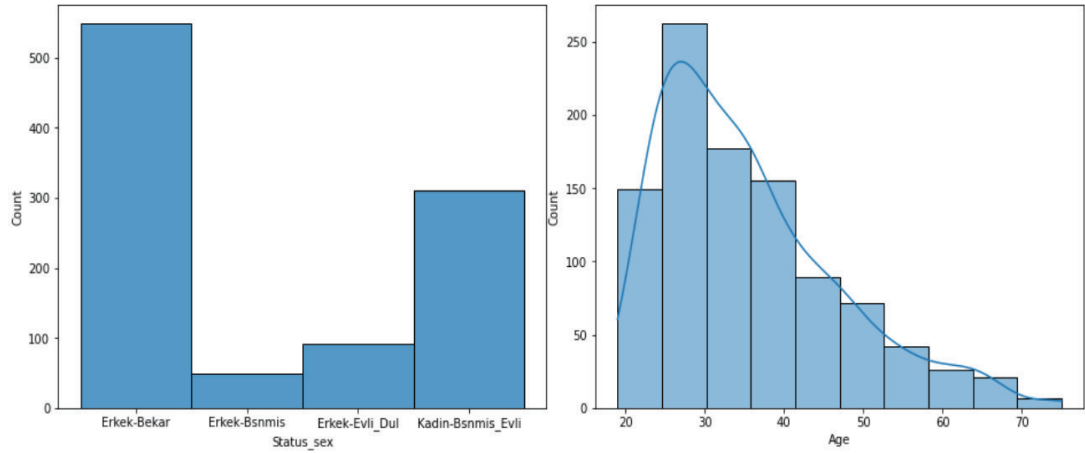
3.3.3.1. Alman Kredi Veri Setinin Keşfi

Şekil 18: Alman Kredi Veri Seti İçin Korelasyon Matrisi



Alman kredi veri seti için elde edilen Şekil 18'deki korelasyon matrisine göre kredi miktarı ve kredi vadesi arasındaki korelasyon değeri %62 oranındadır. Bu değer bize, iki değişken arasında orta derecede pozitif yönde bir ilişki olduğunu gösterir. Ayrıca talep edilen kredi miktarı ile ödenecek taksit oranı arasındaki korelasyon değeri %0.27 oranındadır. Bu değer bize, iki değişken arasında negatif yönde zayıf bir ilişki olduğunu gösterir. Bahsedilen değişkenler dışında, Korelasyon matrisi için genel bir değerlendirme yapıldığında ise aralarında pozitif yönde çok zayıf bir ilişki bulunan değişkenlerin sayısı çoğunluktadır. Bu değişkenlerin korelasyon değeri %0.03-0.26 arasında dağılmaktadır. Ek olarak, aralarında negatif yönde çok zayıf bir ilişki bulunan değişkenlerin korelasyon değeri ise %0.03-%9 aralığında dağılmaktadır.

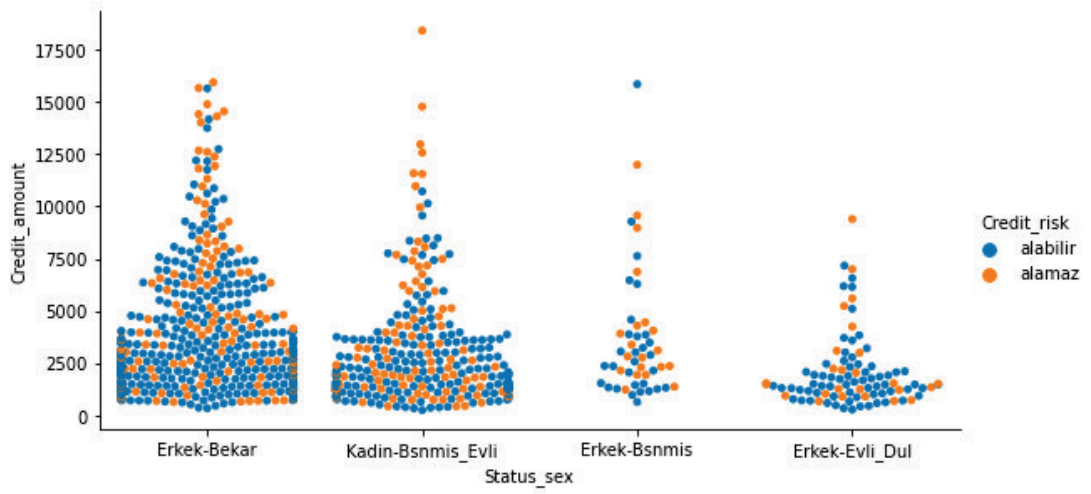
Şekil 19: Başvuran Müşterilerin Cinsiyet ve Medeni Durum ile Yaş Dağılımları



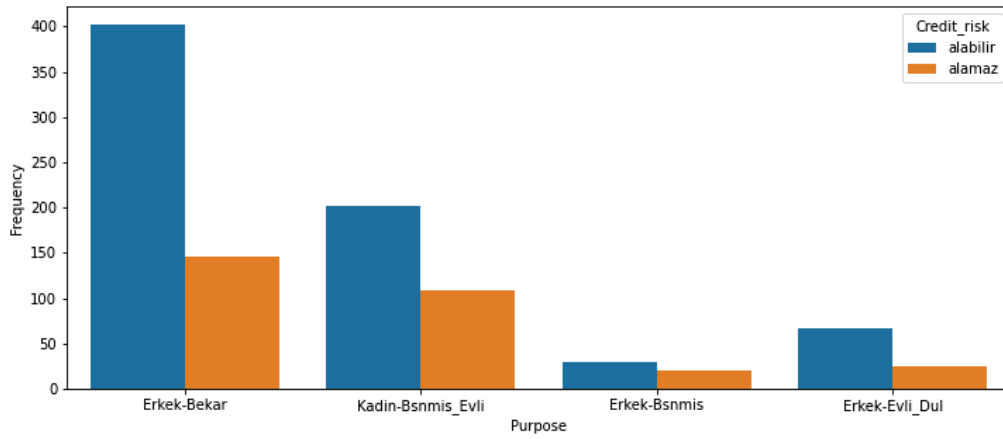
Sol tarafta yer alan sütun grafiğine göre kredi için başvuran müşterilerin çoğunluğu erkek olup, medeni durumları bekarıdır. Hem erkek olup hem de medeni durumu bekar olan müşterilerin sayısı ise 550'nin üzerindedir. Başvuran müşterilerin yaklaşık 300'ü kadın olup, evli veya boşanmıştır. Ayrıca başvuru yapan erkeklerin 100'ü evli veya dul iken yaklaşık 50 kişi ise boşanmıştır.

Sağ tarafta ise histogram ve poligon grafiği verilmiştir. Kredi için başvuru yapan müşterilerin yaş aralığı genel olarak yaklaşık 20 ile 75 yaş arasında değişmektedir. Aynı zamanda grafiği detaylı incelediğimizde, krediye başvuru yapan müşterilerin çoğunluğunun yaşı, 20 ile 40 yaş aralığındadır. Özellikle, 25-30 yaş aralığında kredi için başvuru yapan müşterilerin sayısı, diğer yaş aralıklarında olup, kredi için başvuran müşteri sayısından daha fazladır. Genel bir çıkarım yapmak gerekirse, genç ve orta yaşlı müşterilerin finansal tasarruf ihtiyacı, yaşlı dediğimiz yaş grubunda yer alan müşterilerden daha fazladır. Bunun yanı sıra, histogram ve poligon grafiklerine bakıldığında yaş değişkeni için sağa çarpık bir dağılıma sahip olduğu çıkarımında bulunabiliriz. Sağa çarpık olduğunun sağlamasını yapmak istersek ya da poligon grafiğini çizdirilmese bile, ilgili değişkenin; ortalama, medyan ve mod değerlerine bakılarak da anlaşılabilir. Yaş değişkeni için medyan değeri 33, mod değeri 27, ortalama değeri 35,54 olarak hesaplanmıştır. Yaş değişkeninin, $mod < medyan < ortalama$ durumuna sahip olması nedeniyle, sağa çarpık bir dağılım gösterdiği böylece doğrulanmış olur.

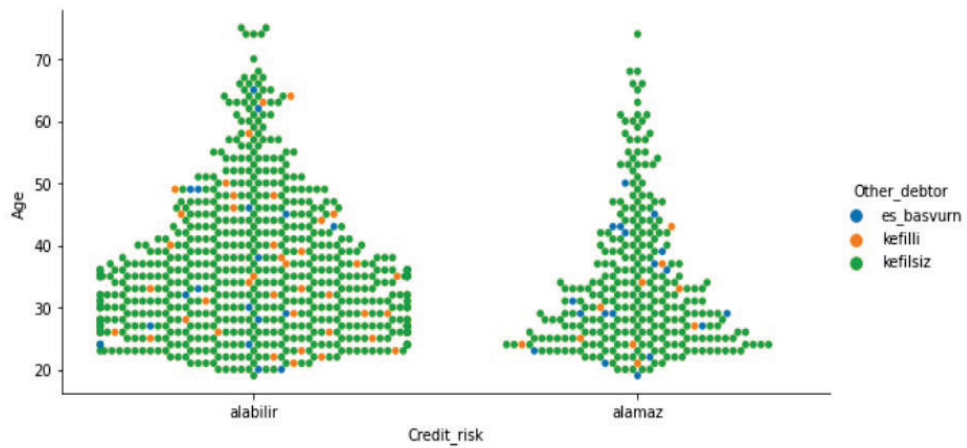
Şekil 20: Kredi Miktarı ve Medeni Durum-Cinsiyete Göre Kredi Risk Durumu



Şekil 21: Medeni Durum-Cinsiyete Göre Kredi Risk Durumu



Şekil 22: Kefil ve Yaşa Göre Kredi Risk Durumu



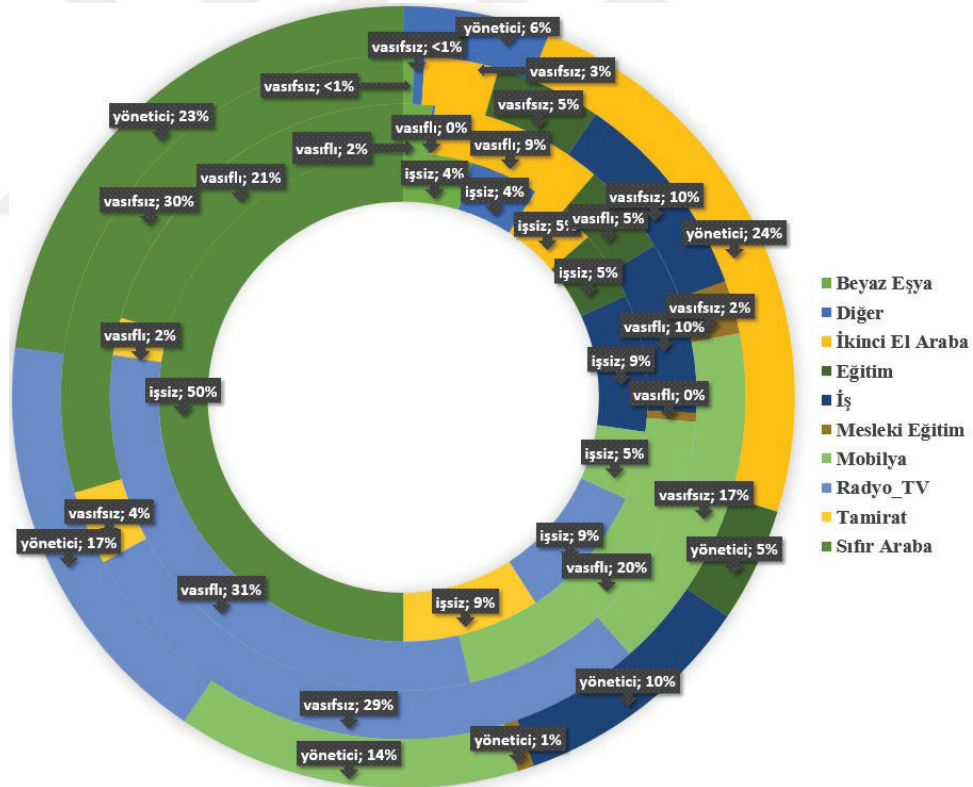
Şekil 20’de, kredi başvurusunda bulunan müşterilerin, talepte bulundukları kredi miktarı ve medeni durumu- cinsiyetine göre kredi risk durumu yer almaktadır. Hem erkek olup hem de bekar olan müşterilerin büyük bir çoğunluğu yaklaşık 1000 ila 5000 Alman markı aralığında kredi talebinde bulunmuşlardır. Kadın olup , aynı zamanda evli veya boşanmış olan müşterilerin çoğunluğu yaklaşık 1000 ila 4000 Alman markı aralığında kredi talebinde bulunmuşlardır. Boşanmış olup, aynı zamanda erkek olan müşterilerin çoğunluğu yaklaşık olarak 1000 ila 5000 Alman markı aralığında kredi talebinde bulunurken, evli veya dul olan diğer erkek müşterilerin çoğunluğu ise yaklaşık olarak 1000 ila 2500 Alman markı aralığında kredi talep etmişlerdir. Ayrıca Şekil 20 hakkında genel bir yorum yapmak gerekirse, talep edilen kredi miktarındaki artış, cinsiyet ve medeni durum farketmeksizin, çoğunlukla banka tarafından reddedilmiştir. Başka bir deyişle, düşük miktarda kredi talebinde bulunan müşterilerin, yüksek miktarda kredi talebinde bulunan müşterilere göre kredi alabilen sayısı daha fazladır.

Şekil 21’deki sütun grafiğinde de görüldüğü üzere, medeni duruma ve cinsiyete göre bakıldığında, müşterilerin çoğunluğu kredi alabilmiştir. Özellikle, hem erkek olup hem de medeni durumu bekar olan yaklaşık 550 müşteriden 400’ü kredi alabilirken, geriye kalan yaklaşık 150 müşteri ise kredi alamamıştır. Kadın olup evli veya boşanmış olan yaklaşık 300 müşteriden 200’ü kredi alabilirken, geriye kalan yaklaşık 100 müşteri ise kredi alamamıştır. Erkek olup, evli veya dul olan yaklaşık 100 müşteriden 60’ı kredi alabilirken, geriye kalan yaklaşık 40 müşteri ise kredi alamamıştır. Son olarak, boşanmış 50 erkek müşterinin yaklaşık olarak 30’u kredi alabilirken, geriye kalan 20 müşteri kredi alamamıştır. Kredi alamayan müşterilerin kredi alamaması, kredi başvurusu yapan müşterinin kredilendirilme açısından, finansal kuruluş tarafından riskli görülmesiyle ilişkilidir.

Finansal kuruluşlar kredilendirme sürecinde, kendilerini garantiye almak için potansiyel müşterilerinden teminatlar ister. Bu teminatlardan biri de müşterinin ilgili finansal kuruluşa bir kefil göstermesidir. Müşterinin kefil olarak göstereceği birinin olması, bankaların müşteri açısından yaşayacağı problemlerden doğacak riskleri biraz olsun azaltmış olacaktır. Şekil 22’ye bakıldığında krediye başvuran müşterilerin büyük bir çoğunluğu bir kefile sahip değildir. Buradan iki sonuç çıkabilir; Birincisi, finansal kuruluş için kefileden daha önemli teminat unsurlarını veya müşteri bankanın

kredilendirme süreci için belirlediği diğer tüm şartları karşılıyor olabilir. İkincisi ise veri seti oluşturulurken, belirlenen örneklemin içerisinde yer alan gözlemlerin büyük çoğunun kefil olmayan müşterilerden oluşması olabilir. Bu grafiği oluşturmamın bir diğer amacı ise yaş arttıkça bankanın kefil isteme zorunluluğunun artıp artmadığını öğrenmektir. Çünkü, yaş ilerledikçe müşterinin vefatı veya hastalıklarının artması ve çalışamayacak hale gelmesi durumu söz konusudur. Gerçekleşmesi olası bu durumlar, banka için büyük bir risk unsurudur. Bu nedenle banka müşterisinden teminat olarak kefil göstermesini şart koşabilir. Ancak bu veri setinde öyle bir durum söz konusu değildir. Müşterilerin çoğu kefilsiz olarak kredi alabilmiştir. Ayrıca müşterinin yaşının büyük olması ve kefilinin olmaması kredi almasına engel teşkil etmediği sonucu ortaya çıkar.

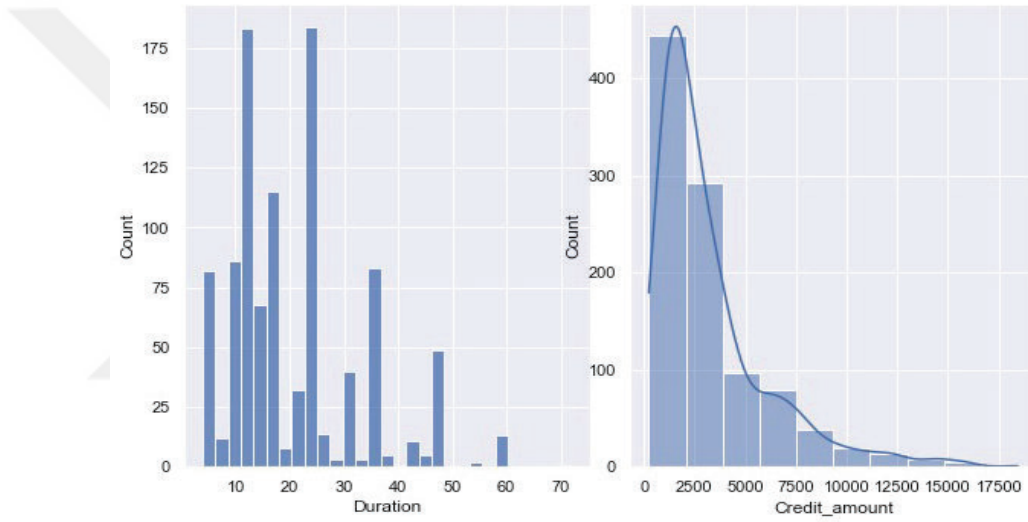
Şekil 23: Müşterilerin Meslekleri ve Krediye Başvuru Amaçları



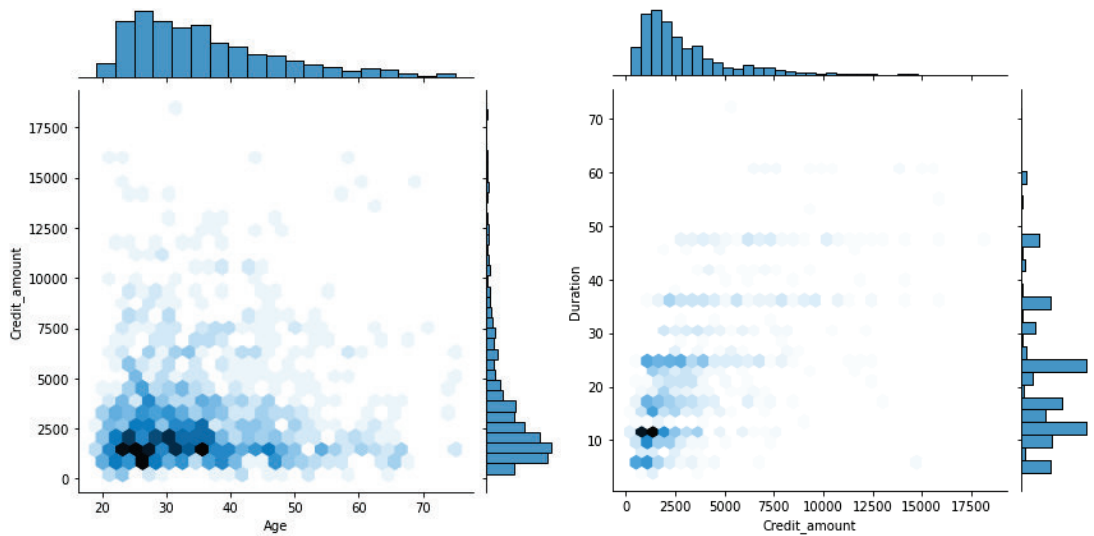
Şekil 23'te ise, müşterilerin meslekleri ve krediye başvuru amaçlarının yer aldığı bir halka grafik oluşturulmuştur. Vasıflı, yönetici, vasıfsız ve işsiz olmak üzere dört farklı meslek grubu vardır. İşsiz olan müşterilerin %50'si ile vasıfsız olan müşterilerin %30'unun krediye başvuru sebepleri, çoğunlukla sıfır araba alımı içindir.

Yönetici müşterilerin %24'ünün ise ikinci el araba alımı içindir. Bunun yanı sıra, vasıflı olan müşterilerin %31'i ise radyo veya televizyon alımı için kredi başvurusunda bulunmuşlardır. Kredi talebinde bulunan müşteriler için önemli olan diğer başvuru amaçları ise şöyledir; Vasıflı müşterilerin %21'i sıfır araba alımı için kredi başvurusunda bulunmuştur. Vasıfsız müşterilerin %29'u radyo veya televizyon alımı için krediye başvurmuştur. Tamirat, iş ve radyo/televizyon alma amaçlarının her biri için %9 oranında işsiz müşteri kredi talebinde bulunmuştur. Müşterilerin diğer başvuru amaçlarının mesleklerine göre dağılımı yukarıdaki halka grafiğinde yer almaktadır.

Şekil 24: Talep Edilen Kredi Miktarları ve Vadelerinin Dağılımı



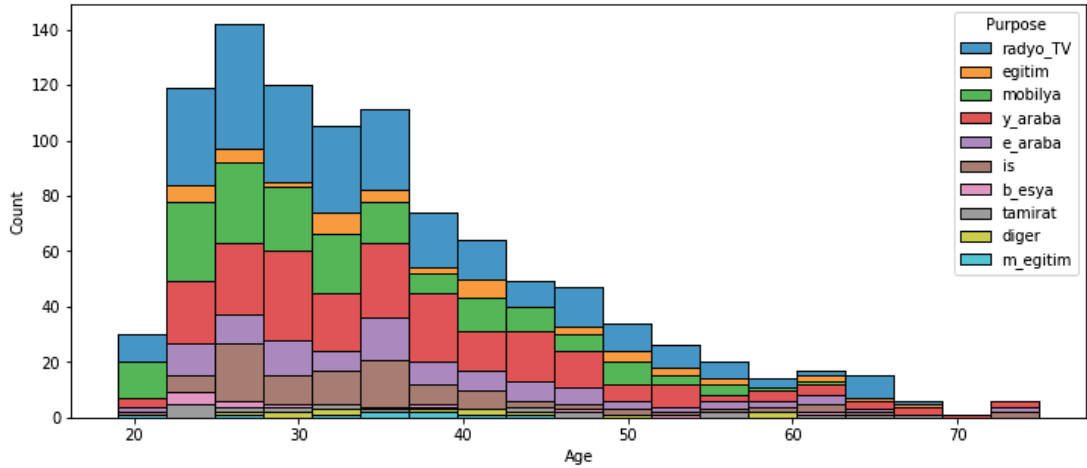
Şekil 25: Talep Edilen Kredi Miktarı ile Yaş ve Vadenin Dağılımı



Şekil 25'te sol tarafta yer alan, müşterilerin yaşı ile talep ettikleri kredi miktarlarının histogram ve “Hexbin” grafiğinin yer aldığı ortak bir grafik bulunmaktadır. Hexbin grafiğinde ilk çarpan koyu bölgedir. Bu koyu bölge, veri setindeki gözlem değerlerinin yığıldığı aralığı kolayca anlamamıza yardımcı olur. Yaşı 20 ile 40 aralığında olan müşterilerin talep ettikleri kredi miktarı genellikle 1000 ile 2500 Alman markı aralığında değişmektedir. Ayrıca bu grafiğe bakarak değişkenlerin aralarında ilişki de yorumlanabilir. Grafiğe göre yaştaki artış talep edilen kredi miktarını çok etkilediği söylenemez. Bu nedenle iki değişken arasında zayıf bir ilişkinin olduğu görülmektedir.

Şekil 25'te sağ tarafta ise müşterilerin talep ettikleri kredi miktarı ve vade süreleri karşılaştırılmıştır. Bu grafikte, talep edilen kredi miktarındaki artışa orantılı olarak vade süresinde de bir artış olduğu görülmektedir. Bu da değişkenlerin arasında pozitif bir ilişki olduğunu gösterir, fakat aralarında ilişkinin kuvveti çok güçlü değildir. Vade süreleri çoğunlukla 5 ile 25 ay arasındadır. Ancak, banka tarafından müşterilere verilen kredinin vade süresi en yoğun olduğu vade aralığı 10 ile 15 aydır. Müşterilerin çoğu düşük miktarda kredi talebinde bulundukları için kredi talebine paralel olarak vade sürelerinin de kısaldığını söylemek mümkündür.

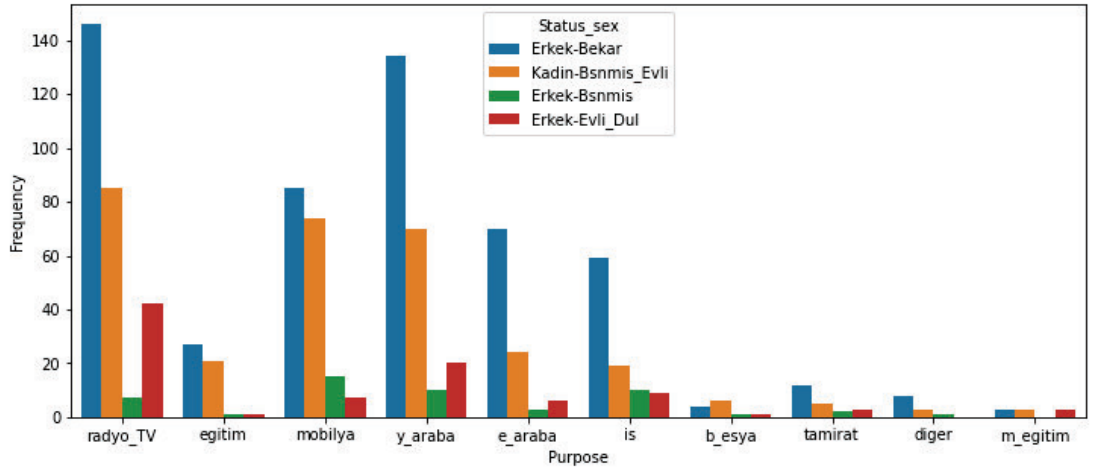
Şekil 26: Müşterilerin Yaşı ile Krediye Başvuru Amaçları



Şekil 26'ya göre tüm yaş aralıkları için popüler olan başvuru amacı, radyo ve televizyon alımı içindir. Radyo ve televizyon alımından sonra genel olarak en popüler kredi amaçları; mobilya, sıfır otomobil (y_araba) ve ikinci el otomobil (e_araba) alımıdır. Aynı zamanda, yaş aralığı 25 ile 45 aralığında değişen müşteriler yeni bir

kurma amacıyla kredi başvurusunda bulunmuşlardır. 19 ila 23 yaş aralığındaki genç grubun eğitim kredisi için başvuramamaları oldukça şaşırtıcıdır. Şaşırtıcı olmasının sebebi, o yaş aralığındaki kişilerin, üniversite çağında olması sebebiyle öğrenci olma olasılığından kaynaklıdır. Bunun yanı sıra, hemen hemen tüm yaş gruplarındaki müşteriler, eğitim amacıyla kredi başvurusunda bulunmuşlardır. Özellikle orta yaş diyebileceğimiz 30 ila 50 yaş aralığında bulunan müşterilerin eğitim kredisi başvurusunda bulunmalarının sebebi, çocuklarının eğitimi olabilir. Şekil 27' yi inceleyerek bu düşüncüyü ispatlamak daha mümkün olacaktır.

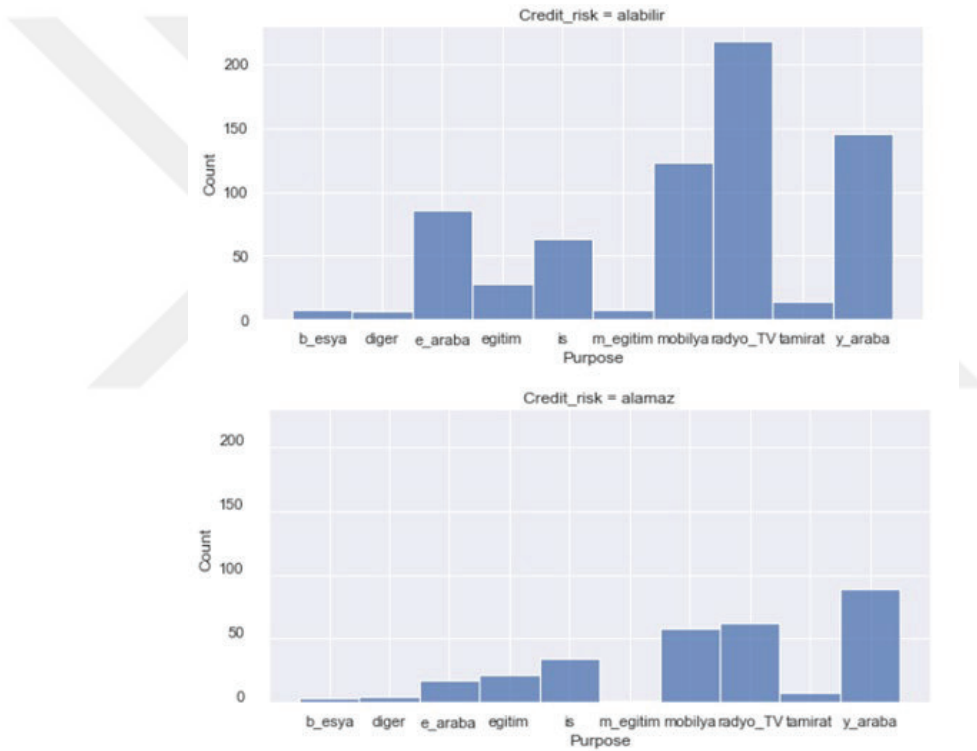
Şekil 27: Medeni Durum-Cinsiyet ile Krediye Başvuru Amacı



Şekil 27'de müşterilerinin cinsiyet-medeni durumları ile krediye başvuru amaçları yer almaktadır. Cinsiyeti erkek olup, aynı zamanda bekar olan müşterilerin yaklaşık 145'i, televizyon veya radyo almak için kredi talebinde bulunmuşlardır. Ayrıca Erkek olup, medeni durumu bekar olan müşterilerden yaklaşık 130'u, sıfır otomobil (y_araba), 90'ı mobilya, 70'i ikinci el otomobil (e_araba) alımı için ve 60'ı ise yeni bir iş kurmak için kredi başvurusunda bulunmuştur. Başvurdukları diğer krediler ise sırasıyla eğitim, tamirat, beyaz eşya, mesleki eğitim ve diğer kredilerdir. Cinsiyeti kadın olup, evli veya boşanmış olan müşteriler çoğunlukla radyo ve televizyon almak için kredi başvurusunda bulunmuşlardır. Bu müşterilerden yaklaşık 90'ı radyo ve televizyon için krediye başvururken; 75'i mobilya, 70'i sıfır otomobil, 30'u ikinci el otomobil alımı için ve 25'i ise yeni bir iş kurmak için kredi başvurusunda bulunmuştur. Başvurdukları diğer krediler ise sırasıyla beyaz eşya, tamirat, mesleki eğitim ve diğer kredilerdir. Şekil 26'da bahsedildiği üzere eğitim kredisine başvuran,

30 ila 50 yaş aralığında bulunan müşterilerin çoğunluğunun medeni durumu boşanmış veya evli olan kadın müşterilerden oluştuğunu ve başvuru amaçlarının çocuklarının eğitimi olduğu yorumunda bulunmak hiç de yanlış olmaz. Hem erkek olup hem de evli veya dul olan müşterilerin yaklaşık 40'ı radyo ve televizyon almak için kredi talebinde bulunmuşlardır. Bunun dışında, yaklaşık olarak 15'i mobilya, 25'i sıfır otomobil, 10' u ikinci el otomobil alımı için ve 15'i ise yeni bir iş kurmak için kredi talebinde bulunmuştur. Başvurdukları diğer krediler ise sırasıyla eğitim, tamirat, beyaz eşya, mesleki eğitim kredileridir.

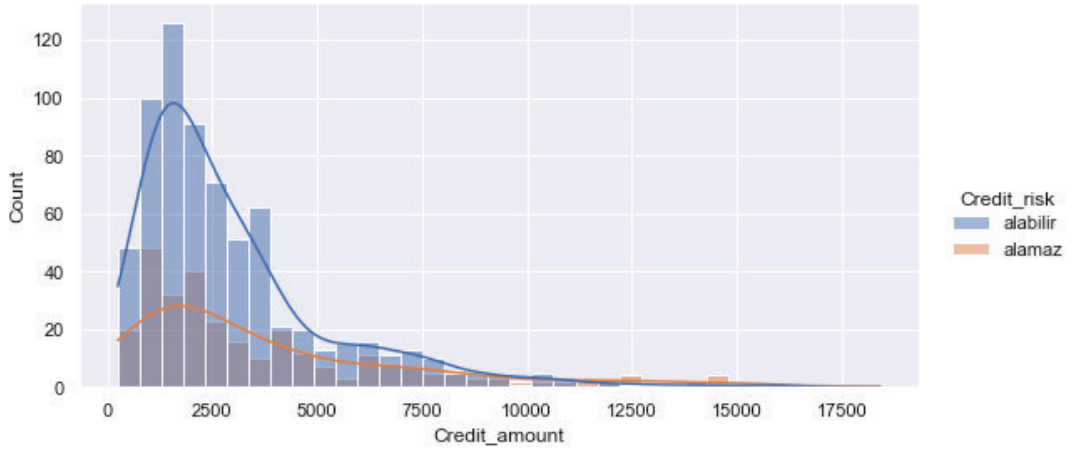
Şekil 28: Kabul ve Ret Alan Kredi Tipleri



Şekil 28'de ise kabul ve ret alan kredilere ait sütun grafikleri yer almaktadır. Üstte yer alan sütun grafiği, başvurduğu kredi tipine göre müşterinin alabildiği kredilerin sayısını temsil etmektedir. Altta yer alan sütun grafiği ise başvurduğu kredi tipine göre müşterinin alamadığı kredilerin sayısını temsil etmektedir. Üst tarafta yer alan sütun grafiğine baktığımızda, radyo ve televizyon alımı için başvuran 200'den fazla müşterinin kredi talebi, banka tarafından kabul edilmiştir. Bu kredi tipi aynı zamanda, diğer kredi tiplerine göre en fazla kabul alan kredi tipidir. Mobilya alımı için krediye başvuran 100'den fazla müşterinin kredi talebi, banka tarafından kabul

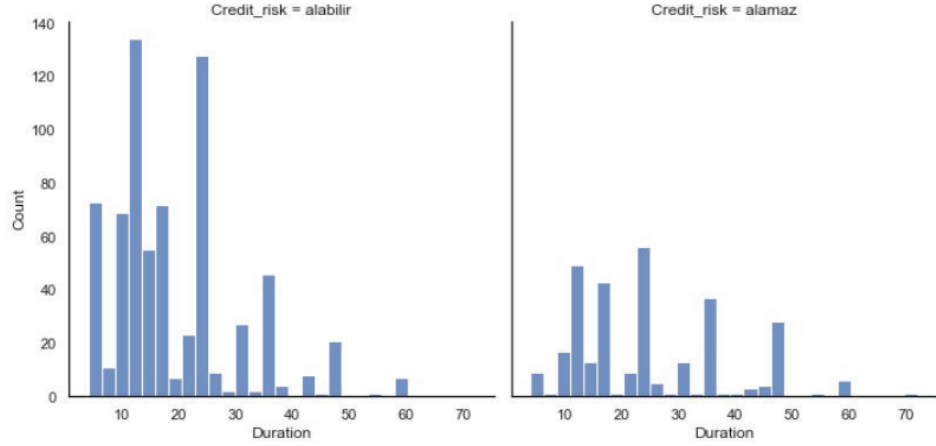
edilmiştir. Sıfır otomobil alımı için krediye başvuran yaklaşık 150 müşterinin kredi talebi, banka tarafından kabul edilmiştir. Ek olarak, ikinci el otomobil almak isteyen müşterilerin 80'inin, iş kurmak isteyen müşterilerin yaklaşık 60'ının ve eğitim için krediye başvuran müşterilerin yaklaşık 50'sinin kredi talebi, banka tarafından kabul edilmiştir. Diğer kredi tiplerinin kabul durumları, üst tarafta yer alan sütun grafiğinde gösterildiği gibidir. Alt tarafta bulunan sütun grafiğini incelediğimizde ise sıfır otomobil alımı için krediye başvuran yaklaşık 90 müşterinin, kredi talebi banka tarafından reddedilmiştir. Bu kredi tipi aynı zamanda, diğer kredi tiplerine göre en fazla ret alan kredi tipidir. Ek olarak, radyo veya televizyon alımı için krediye başvuran yaklaşık 70 müşterinin, mobilya alımı için krediye başvuran yaklaşık 60 müşterinin ve iş kurmak için kredi talebinde bulunan yaklaşık 40 müşterinin kredi talebi reddedilmiştir. Diğer kredi tiplerinin ret durumları, alt tarafta yer alan sütun grafiğinde gösterildiği gibidir.

Şekil 29: Talep Edilen Kredi Miktarlarının Kabul ve Ret Durumu



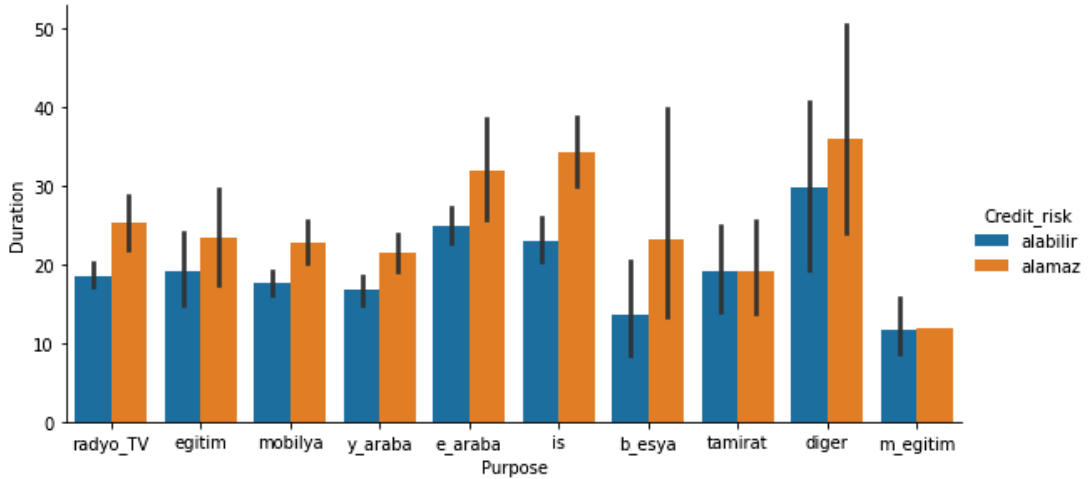
Şekil 29'da müşterilerin talep ettiği kredi miktarlarının kabul ve ret durumları gösterilmiştir. Müşterilerin çoğunun, talep ettikleri kredi miktarını aldığı aşıkardır. Özellikle, 1000-4000 Alman markı arasında kredi talebinde bulunan çoğu müşterinin, kredi talepleri kabul edilmiştir. Ayrıca 4000-4500 Alman markı aralığında ise kabul edilen kredi miktarı sayısı ile reddedilen kredi miktarı sayısı nerdeyse birbirine eşittir. Dağılımları hakkında genel bir yorum yapmak gerekirse eğer, kabul edilen kredi miktarları ile reddedilen kredi miktarlarının sağa çarpık bir dağılım gösterdiklerini söylemek mümkündür.

Şekil 30: Kredilerin Vadesi ve Kabul-Ret Durumları



Şekil 30’da müşterilerin aldıkları kredilerin vade sürelerine kabul ve ret durumları verilmiştir. Her iki grafik için genel bir değerlendirme yapmak gerekirse, vade süresinin artmasıyla ret alan kredi talebinin az da olsa arttığıdır. Özellikle, vadesi süresi 40-50 ay arasında olan kredilerde ret sayısının arttığını, sağdaki grafikte görmek mümkündür.

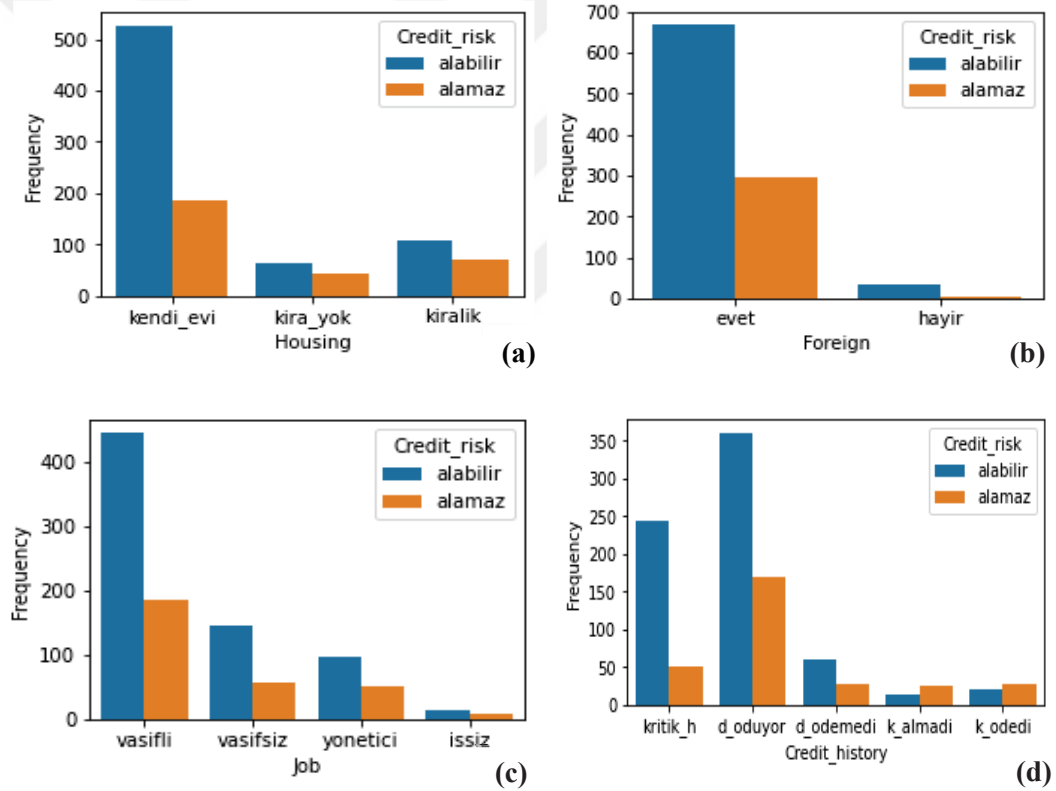
Şekil 31: Kredi Tipleri ile Vade Sürelerine Göre Kredi Risk Durumu



Şekil 31’de görüldüğü üzere vade süresindeki artış, müşterilerin başvurduğu tüm kredi tiplerinde ret sayısının artmasına neden olmuştur. Çubuk grafiklerini daha detaylı incelediğimizde; vade süresi 20 ayın üzerinde olan radyo/televizyon, mobilya, iş ve diğer farklı kredi tiplerine başvuran müşterilerin, kredi talepleri reddedilmiştir. Bu kredi tiplerinin yanı sıra, vade süresinin 10 ayın üzerinde çıkması ise buzdolabı, tamirat, sıfır otomobil ve ikinci el otomobil alımı için başvuran müşterilerin ise kredi

taleplerinin çoğunlukla reddedildiği görülmektedir. Eğitim ve mesleki eğitim kredi tipleri için durum oldukça farklıdır. Eğitim kredileri için bankanın talebi reddedeceği vade eşiği oldukça düşük olduğu görülmektedir. Yani, eğitim kredileri için belirlenen vade düşük olsa da müşterinin kredi talebinin reddedilme olasılığı oldukça yüksektir. Mesleki eğitim için kredi talebinde bulunan müşteriler için vade süresi ne olursa olsun, kredi taleplerinin kabul edilme olasılığı oldukça yüksektir. Sonuç olarak, banka için vade süresinin artması, talep edilen krediyi ödemesini riske sokacağından, belli bir vade sınırından sonra müşterilerin başvurduğu kredilerin reddedilme olasılığı yükselmektedir.

Şekil 32: Müşteriye ait Bazı Bilgilere Göre Kredi Risk Durumu



Şekil 32’de; müşterin ikamet ettiği evin durumu, müşterinin yabancı uyruklu olup olmaması, müşterinin mesleği ve müşterinin kredi geçmişinin başvurduğu kredinin kabul veya ret alması üzerine etkisi incelenmiştir.

(a)’daki sütun grafiğine bakıldığında kendi evine sahip olan 500’den fazla müşterinin kredi talebi kabul edilmişken, yaklaşık 180 müşterinin ise kredi talebi reddedilmiştir. İkamet ettiği ev kendisine ait olmayan ve aynı zamanda kira da

ödemeyen (kira_yok) yaklaşık 50 müşterinin kredi talebi kabul edilirken, yaklaşık 30 müşterinin ise kredi talebi reddedilmiştir. İkamet ettikleri eve kira ödeyen 100'den fazla müşterinin kredisi talebi kabul edilirken, yaklaşık 80 müşterinin ise kredi talebi reddedilmiştir. Genel bir yorum yapmak gerekirse, kredi talebi kabul olan müşterilerin büyük çoğunluğunun kendisine ait evi bulunmaktadır. Müşterinin ev sahibi olması, aslında banka için güvence oluşturuyor olsa da kredinin kabul veya reddedilmesi üzerinde pek etkisi yoktur. Çünkü, kirada oturan ya da ev sahibi olmayan müşterilerin de kabul edilen kredi talepleri bulunmaktadır.

(b)'deki sütun grafiğine bakıldığında, yabancı olan 600'ün üzerinde müşterinin kredi talebi kabul edilirken, 250'nin üzerinde müşterinin kredi talebi reddedilmiştir. Ayrıca yabancı olmayan müşterilerin yaklaşık olarak 50'sinin kredi talebi kabul edilmiştir. Kredi talebi reddedilen müşteri sayısı ise oldukça azdır. Genel bir yorum yapmak gerekirse kredi başvurusunda bulunan müşterinin yabancı uyruklu olması, kredi alması için bir problem oluşturmaz.

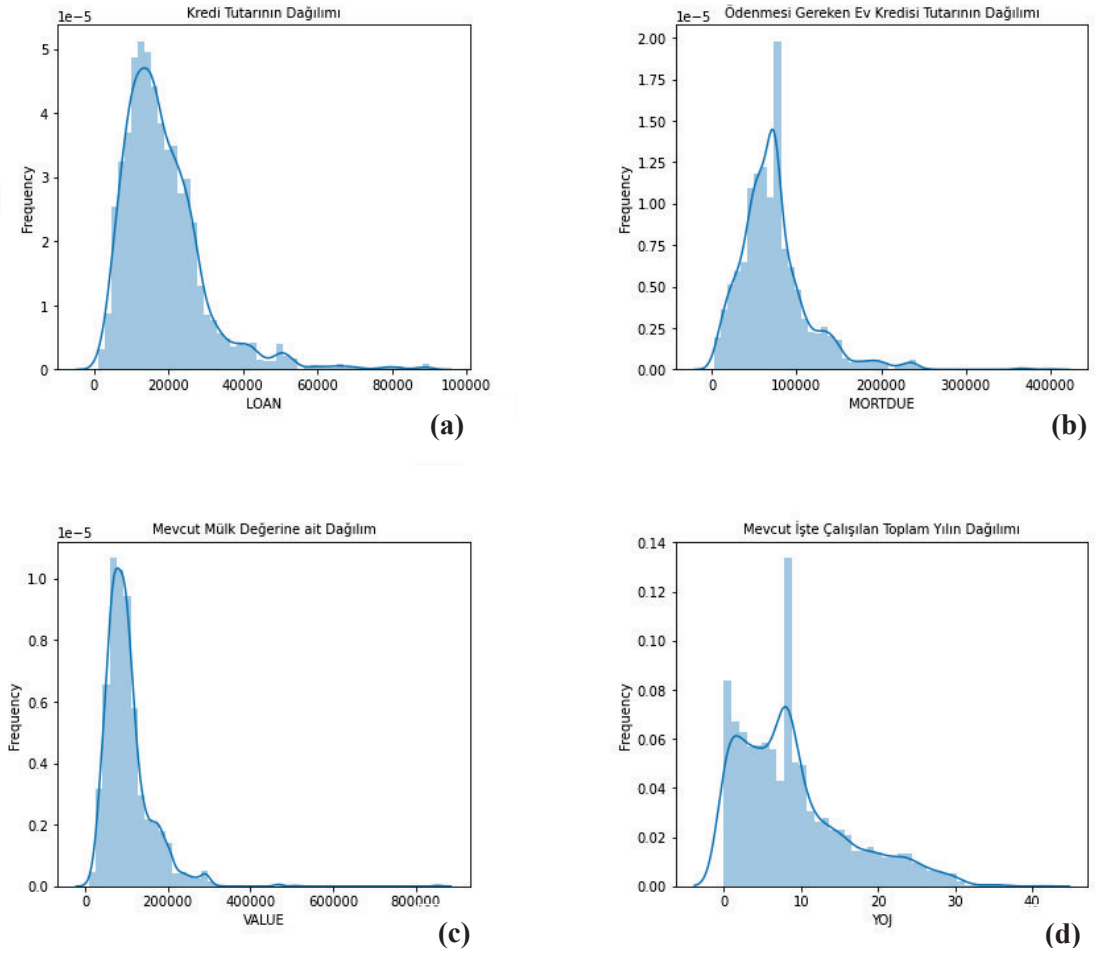
(c)'deki sütun grafiğine bakıldığında, vasıflı olan 400'ün üzerinde müşterinin kredi talebi kabul edilirken, 150'den fazla müşterinin ise kredi talebi reddedilmiştir. Vasfi olmayan 150'den fazla müşterinin kredi talebi kabul edilirken, 50'den fazla müşterinin ise kredi talebi reddedilmiştir. Yönetici olan 100'ün üzerinde müşterinin kredi talepleri kabul edilirken, 50'nin üzerinde müşterinin kredi talebi reddedilmiştir. Ayrıca işsiz olan müşterilerin kabul edilen kredi taleplerinin sayısı, reddedilen kredi talebinden fazladır. Genel bir yorum yapmak gerekirse, müşterinin mesleği kredi talebinin kabul edilmesi veya reddedilmesi üzerinde etkisinin olmadığını söylemek mümkündür. Çünkü her meslek grubunda, kabul edilen kredi talebi reddedilen kredi talebinden daha fazladır.

(d)'deki sütun grafiğine bakıldığında, geçmiş kredilerini düzenli ödeyen 350'den fazla müşterinin kredi talebi kabul edilirken, 150'den fazla müşterinin kredi talebi reddedilmiştir. Kritik hesaba sahip veya başka bankadan (kritik_h) kredi alan 200'den fazla müşterinin kredi talebi kabul edilirken, yaklaşık 50 müşterinin kredi talebi reddedilmiştir. Kredisini düzenli ödemeyen 50'den fazla müşterinin kredi talebi kabul edilirken, yaklaşık 40 müşterinin kredi talebi reddedilmiştir. Ayrıca daha önce hiç kredi almayan (k_almadı) ve gecikmeli olarak kredilerini ödeyen (k_ödedi) müşterilerin kabul edilen kredi talebi, reddedilen kredi talebine göre daha düşüktür.

Genel bir yorum yapmak gerekirse eğer, müşterinin kredi geçmişi bazı noktalarda önemli gibi görünse de genel olarak bakıldığında, kredi alımını etkileyecek kadar önemli bir işlevi olduğu söylenemez.

3.3.3.2. HMEQ Kredi Veri Setinin Keşfi

Şekil 33: HMEQ Kredi Veri Setine ait Bazı Özelliklerin Dağılım Grafikleri-1



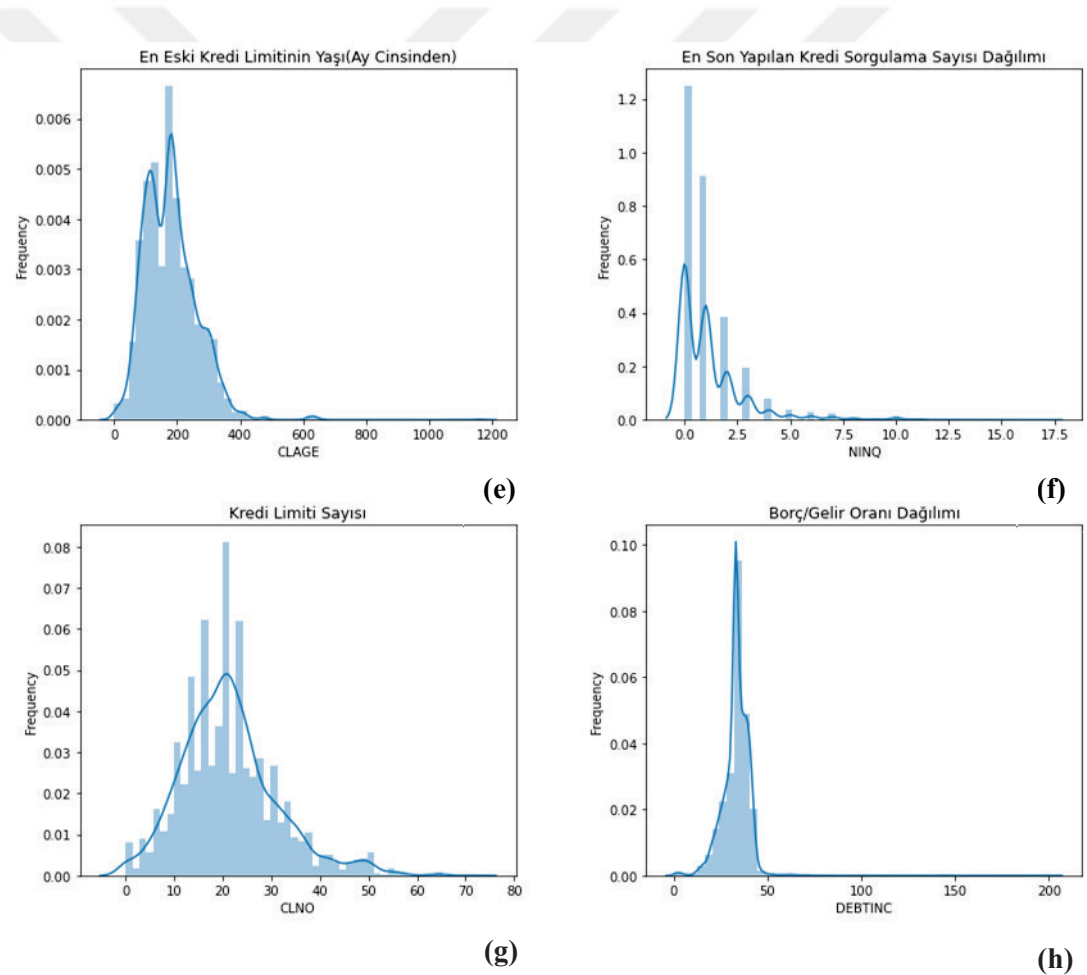
(a)'daki grafik, kredi tutarının dağılımını göstermektedir. Müşterilerin talepte bulundukları kredi tutarı genel olarak 1.100 ila 90.000 Amerikan doları aralığında dağılmaktadır. Ancak grafiğe baktığımızda, müşteriler tarafından en çok talep edilen kredi tutarı, 1.100 ila 30.000 Amerikan doları aralığında olduğu görülmektedir.

(b)'deki grafik, ödenmesi gereken ev kredisi tutarının dağılımını gösterir. Başvuruda bulunan müşterilerin çoğunun, ödemesi gereken ev kredisi tutarı genel olarak 2.063 ila 400.000 Amerikan doları aralığında yoğunlaştığı görülmektedir.

(c)' deki grafik, müşterinin sahip olduğu mevcut mülk fiyatının dağılımını göstermektedir. Müşterilerin sahip olduğu mevcut mülklerinin değeri genel olarak 8.000 ila 860.000 Amerikan doları aralığında dağılmaktadır. Ancak müşterilerin mülklerinin değeri, çoğunlukla 8.000 ila 200.000 Amerikan doları aralığında dağılmaktadır.

(d)' deki grafik, müşterilerin mevcut işindeki çalıştığı toplam yılın dağılımını göstermektedir. Müşterilerin çalıştıkları toplam yıl sayısı, genel olarak 0 ila 41 yıl aralığında dağılmaktadır. Ancak grafik detaylı incelendiğinde, 0 ila 18 yıl aralığında çalışan müşteri sayısının oldukça fazla olduğu göze çarpmaktadır.

Şekil 34: HMEQ Kredi Veri Setine ait Bazı Özelliklerin Dağılım Grafikleri-2



(e)'deki grafik, müşterilerin sahip olduğu en eski kredi limiti yaşına ait dağılımı göstermektedir. Kredi limitinden bahsetmek gerekirse, finansal kuruluş tarafından müşteri için uygun görülen, sınırlı kredi tutarı olarak bilinmektedir. Belirlenen tutarın

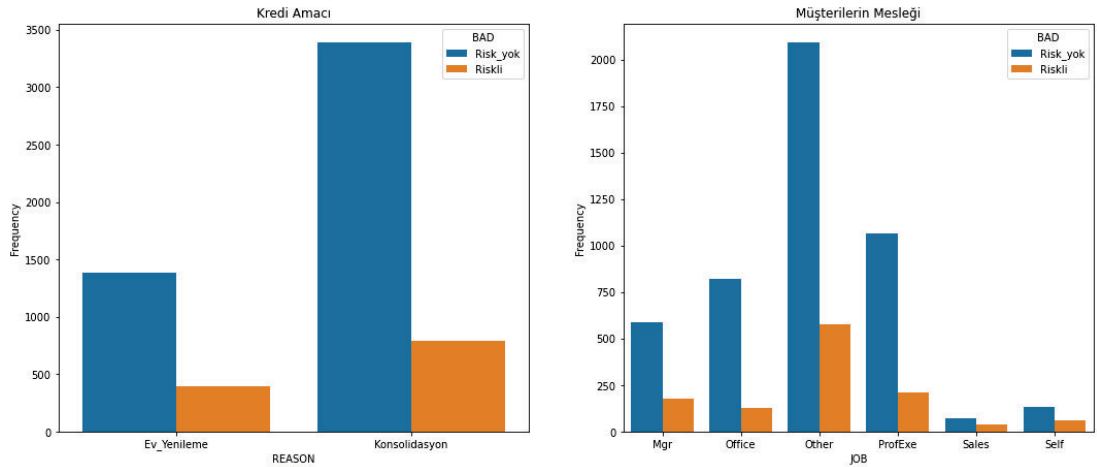
tamamı kullanılabilir ya da kullandığın miktarın faiziyle beraber ödenebilir kredilerdir. Müşteri ancak, finansal kuruluşun belirlediği limit çerçevesinde parayı kullanabilir. Belirlenen limitin üzerinde miktara erişimi yoktur. Üç farklı türde kredi limitleri bulunmaktadır. Bunlar; kişisel, ev sermayesi ve iş kredi limiti türleridir. Grafiğe dönecek olursak, müşterinin sahip olduğu kredi limitlerinin yaşı çoğunlukla 200 ile 400 ay aralığında dağılmaktadır.

(f)'deki grafik, müşteri için en son yapılan kredi sorgulama sayısının dağılımını gösterir. Grafiğe göre kredi veren finansal kurumun, kredi sorgulaması yapmadığı müşteri sayısının daha fazla olduğu söylemek mümkündür. Müşteriler için yapılan kredi sorgulama sayısı, genel olarak 0 ile 10 arasındadır.

(g)'deki grafik, kredi limiti sayısını göstermektedir. Müşteriler için belirlenen kredi limiti sayısı çoğunlukla 0 ile 40 arasında dağılmaktadır. Grafiğe göre kredi limiti sayısı değişkeninin normal bir dağılıma sahip olduğu görülmektedir.

(h)'deki grafik, müşterilerin sahip olduğu borç/gelir oranını göstermektedir. Müşterilerin borç/gelir oranı çoğunlukla %25-50 oranında dağıldığı görülmektedir.

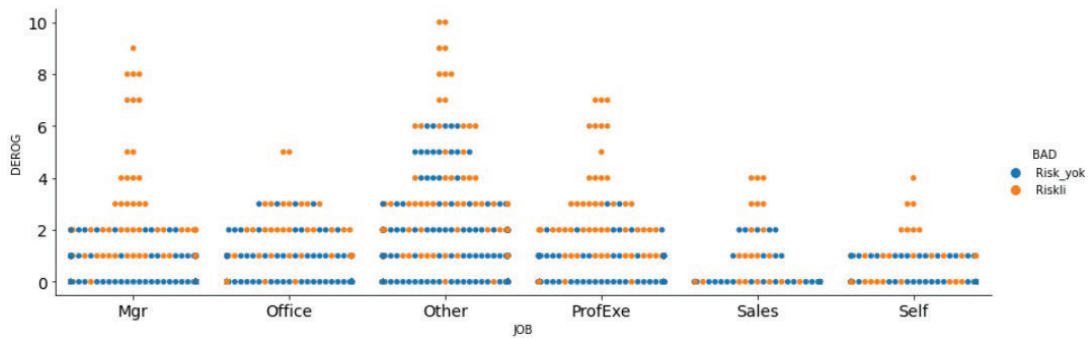
Şekil 35:Müşterinin Kredi Amacı ve Mesleklere Göre Kredi Risk Durumu



Şekil 35'in sol tarafında, müşterilerin kredi kullanım amaçlarına ait bir sütun grafiği bulunmaktadır. Müşteriler kredi talebinde bulunmalarının iki tip amacı vardır. Bunlar; borç konsolidasyonu ve ev yenilemedir. Borç konsolidasyonu, borca ait vadenin değişimidir. Başka bir deyişle, kısa vadeli olarak alınan borçların vadesini uzatmak ya da vadesini süresiz hale getirmektir.

Soldaki sütun grafiğine göre; borç konsolidasyonu için kredi talebinde bulunan 3000’ nin üzerinde müşterinin, herhangi bir risk teşkil etmediği ve bu yüzden kredi talebinin kabul edildiği görülmektedir. Bunun yanı sıra, 500’ ün üzerinde müşteri ise riskli olarak tespit edildiği ve kredi talebi reddedildiği görülmektedir. Ayrıca evlerini yenilemek için krediye talebinde bulunan 1000’ in üzerinde müşterinin risk teşkil etmediği ve bu yüzden kredi talebinin kabul edildiği görülmektedir. Bununla beraber, yaklaşık 500 müşterinin riskli olarak tespit edildiği ve bu yüzden kredi talebinin reddedildiği görülmektedir. Sağ tarafta yer alan sütun grafiğinde ise müşterilerin kredi risk durumu, meslekleri üzerinden incelenmiştir. Verilen beş meslek dışında başka bir iş yapan (Other) 2000’inin üzerinde müşterinin, herhangi bir risk teşkil etmediği ve bu yüzden kredi talebinin kabul edildiği görülmektedir. Ayrıca 500’den fazla müşterinin ise risk teşkil ettiği ve bu yüzden kredi talebinin reddedildiği görülmektedir. Akabinde ise kredi talebinde bulunan profesyonel yönetici (Profexe) olan müşteriler için yaklaşık olarak 1000’in üzerine kabul; 200’ün üzerinde ret, memur (Office) olan müşteriler için 750’nin üzerinde kabul; yaklaşık 200 ret, yönetici (Mgr) müşteriler için 500’üzerinde kabul; 200’ün üzerinde ret cevabı çıkmıştır. Satıcı (Sales) ve kendi işine sahip (Self) müşteri için ise kredi risk durumu ise grafikte gösterildiği gibidir. Genel olarak yorum yapmak gerekirse müşterin kalifiye olması, banka nezdinde riskini azalttığını ve kredi başvurusunda bulunan müşterilerin, başvurularının kabul edilme şansını artırdığını söylemek mümkündür.

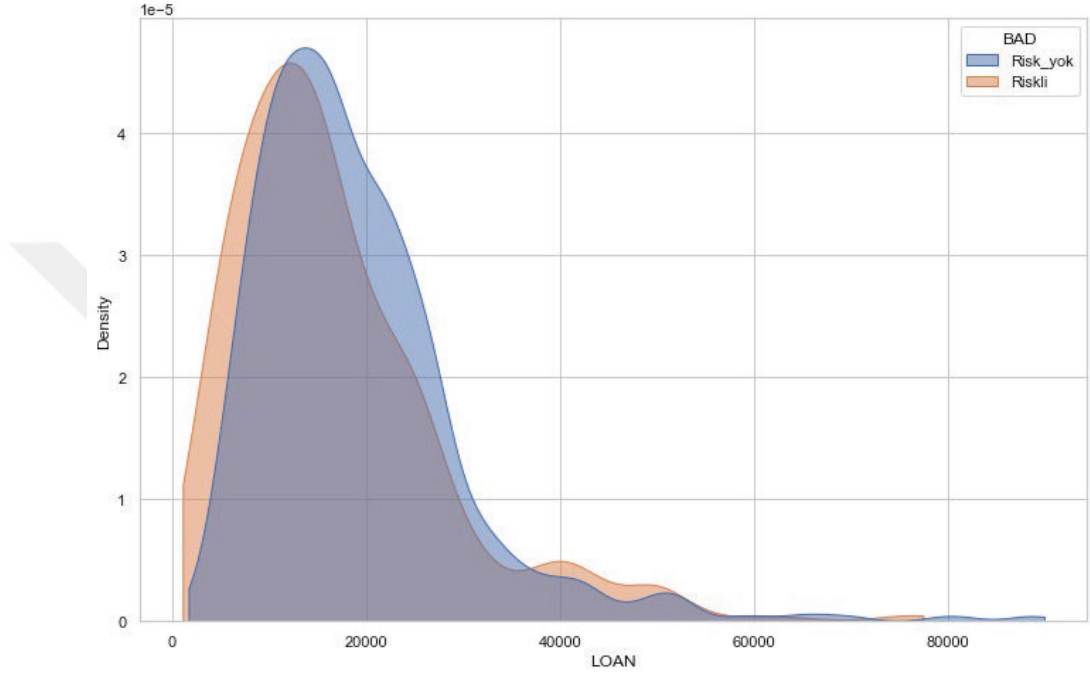
Şekil 36: Meslek ve Temerrüte Düşme Sayılarına Göre Kredi Risk Durumu



Şekil 36’da müşterilerin meslekleri ile temerrüte düşme sayılarına göre kredi risk durumları gösterilmiştir. Grafikten genel bir çıkarım yapmak gerekirse, meslek grubu ne olursa olsun banka tarafından temerrüte düşürülen müşteriler bulunmaktadır.

Başka bir deyişle, müşterin kalifiye olması risk taşımadığı anlamına gelmemektedir. Ek olarak, ödenmeyen ya da geç ödenen kredileri yüzünden banka tarafından temerrüte düşürülen müşterilerin, temerrüte düşme sayıları arttıkça riskin arttığını ve dolayısıyla kredi taleplerinin ret alma olasılığının oldukça yüksek olduğunu görmek mümkündür.

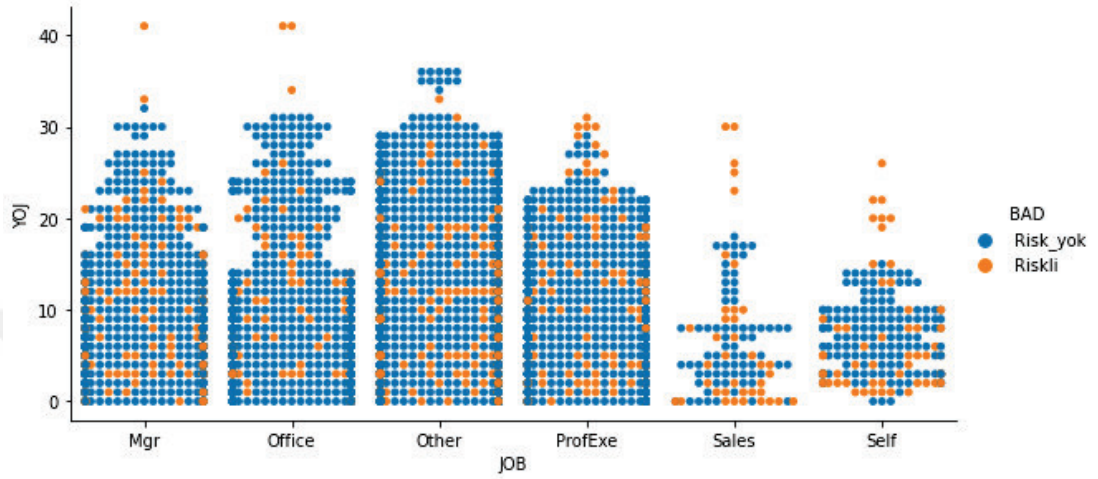
Şekil 37: Müşterinin Talep Ettiği Kredi Tutarı ile Kredi Risk Durumu



Şekil 37’de, müşterilerin talep ettiği kredi tutarları üzerinden kredi risk dağılımları da gösterilmektedir. Banka tarafından kabul edilen ve reddedilen çoğu kredi tutarı, genel olarak aynı aralıkta dağılmıştır. Ayrıca 40.000 ile 50.000 Amerikan doları arasında kredi talebinde müşterilerin riski yüksektir. Bu sebeple, reddedilen kredi tutarı, kabul alan kredi tutarından daha yüksek olduğu görülmektedir. Aynı durum, 70.000 ile 80.000 Amerikan doları arasındaki kredi taleplerinde de gerçekleştiği görülmektedir. Tüm bunlara ek olarak, kredi talepleri kabul edilen müşterilerin sayısı, kredi talebi reddedilen müşterilerden daha fazla olduğunu bu grafikte de görmek mümkündür. Genel bir yorum yapmak gerekirse eğer, talep edilen kredi tutarındaki artış, riski az da olsa arttırdığı görülmektedir. Sadece buna dayanarak, talep edilen kredi tutarındaki artışın, riski de genel olarak arttırdığını söylemek yanlış olabilir. Çünkü, yaklaşık 1000 ila 10.000 Amerikan doları gibi düşük bir aralıkta da, kredi talebinde bulunan müşterilerin oldukça yüksek oranda risk taşıdıklarını

söylemek mümkündür. Bu yüzden, risk durumunu tek bir noktaya bağlamak; bankaların, müşterilerin risk durumları hakkında yanlış çıkarımlarda bulunmasına sebep olacaktır.

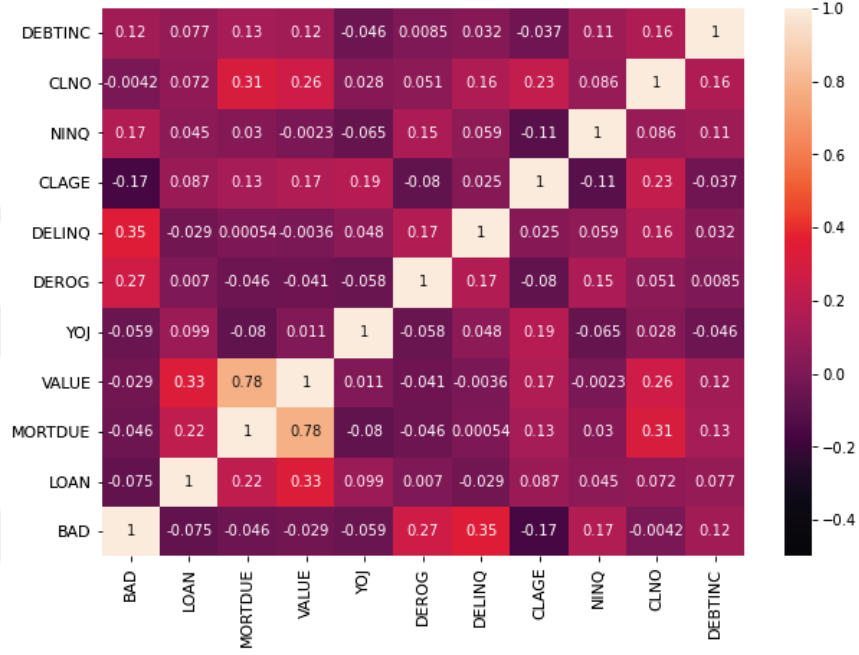
Şekil 38: Müşterinin Mesleği ile Çalıştığı Yıla Göre Kredi Risk Durumu



Şekil 38’de, kredi talebinde bulunan müşterin meslekleri ile çalıştıkları toplam yıl sayısına, yani tecrübelerine göre kredi risk durumları incelenmiştir. Mesleği yönetici (Mgr) olan müşterilerin tecrübelerindeki artış, genel olarak riskli müşteri sayısını düşürdüğünü söylenebilir. Ancak, tecrübe yılı 35’in üzerinde olan yönetici müşterilerin ise risk teşkil etmeleri sebebiyle kredi taleplerinin reddedildiği görülmektedir. Mesleği memur (Office) olan müşterilere baktığımızda ise diğer meslek gruplarına kıyasla, daha az riskli müşteri barındırdığı görülmektedir. Ayrıca müşterinin tecrübesindeki artış, müşterinin bulunduğu meslek grubundaki riskli müşteri sıklığını azalttığı söylenebilir. Ek olarak yönetici meslek grubunda olduğu gibi tecrübeleri 35 yılın üzerinde olan müşterilerin, risk teşkil etmesi sebebiyle kredi taleplerinin çoğunlukla reddedildiğini de söylemek mümkündür. Diğer (Other) meslek gruplarında çalışan müşterilerin tecrübelerindeki artışın, genel olarak riskli müşteri sıklığını azalttığı söylenebilir. Yönetici ve memur meslek gruplarına kıyasla tecrübesi 35 yılın üzerinde olan ve diğer meslek gruplarında riskli müşterilerin olmadığını görmek mümkündür. Profesyonel yönetici (ProExe), satıcı (Sales) ve kendi işinin sahibi (Self) müşteriler için genel bir yorum yapmak gerekirse eğer, bu meslek gruplarında yer alan müşterilerin tecrübelerindeki artışa orantılı olarak taşıdıkları riskin, banka nezdinde düştüğünü görmek mümkündür. Özellikle, 0-5 yıl arasında

tecrübeye sahip satıcı ve kendi işine sahip müşterilerin risk yoğunluğunun fazla olduğunu görmek mümkündür. Bu durumdan şöyle bir sonuç çıkarılabilir; müşterinin bulunduğu meslek grubu kredi talebinin kabulünde veyahut reddinde etken olabilir. Başka bir deyişle, girişimci diyebileceğimiz, satıcı ve kendi işine sahip müşterilerin, bir kurum ve şirket altında çalışan müşterilerden daha riskli olduğu söylenebilir.

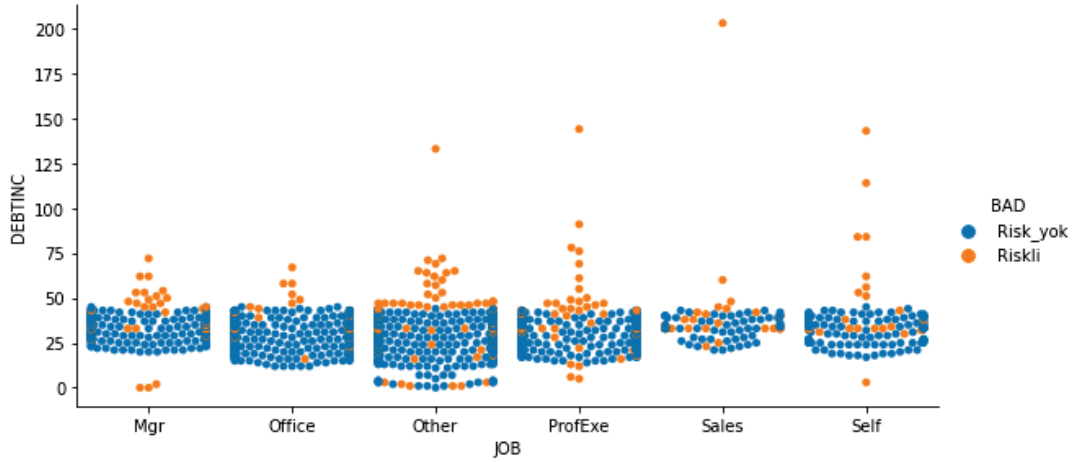
Şekil 39: HMEQ Kredi Veri Seti Korelasyon Matrisi



Şekil 39'daki HMEQ kredi veri seti için elde edilen korelasyon matrisine göre, Müşterinin mevcut mülkünün değeri (Value) ve ödemesi gereken ev kredisi tutarı (Mortdue) arasındaki korelasyon değeri %78'dir. Bu değer bize, iki değişken arasında yüksek derecede pozitif yönde bir ilişki olduğunu gösterir. Müşterinin ödemesi gereken ev kredisi tutarı (Mortdue) ve kredi limiti sayısı (Cln) arasındaki korelasyon değeri %31'dir. Bu değer bize, iki değişken arasında pozitif yönde zayıf bir ilişki olduğunu gösterir. Müşterinin mevcut mülkünün değeri (Value) ve talep ettiği kredi tutarı (Loan) arasındaki korelasyon değeri %33'tür. Bu değer bize, iki değer arasında pozitif yönde zayıf bir ilişki olduğunu gösterir. Müşterinin mevcut mülkünün değeri (Value) ve kredi limiti sayısı (Cln) arasındaki korelasyon değeri %26'dır. Bu değer bize, iki değişken arasında pozitif yönde zayıf bir ilişki olduğunu gösterir. Müşterinin ödemesi gereken ev kredisi tutarı (Mortdue) talep ettiği kredi tutarı (Loan) arasındaki korelasyon değeri %22'dir. Bu değer bize, iki değişken arasında pozitif yönde zayıf

bir ilişki olduğunu gösterir. Müşterinin en eski kredi limitinin yaşı (Clage) ve kredi limiti sayısı (Cln0) arasındaki korelasyon değeri %23'tür. Bu değer bize, iki değişken arasında pozitif yönde zayıf bir ilişki olduğunu gösterir.

Şekil 40: Mesleklere ve Borç/Gelir Oranına Göre Kredi Risk Durumu



Şekil 40'ta, mesleklere ve borç/gelir oranlarına göre müşterilerin risk durumları incelenmiştir. Borç/gelir oranı, 50 veya 50'nin üzerinde olan tüm meslek gruplarındaki müşterilerin tamamının riskli olduğu görülmektedir. Bu nedenle, kredi taleplerinin reddedilme olasılığı oldukça yüksektir. Grafik hakkında genel bir yorum yapmak gerekirse; mesleklerine bakılmaksızın, borç/gelir oranı 25-50 arasında dağılan tüm müşterilerin kredi riski oldukça düşüktür. Başka bir deyişle, meslek borç/gelir oranı 25-50 dağılan müşterilerin kredi talebinin kabul edilme oranı oldukça yüksektir.

3.3.4. Eğitim ve Test Veri Setlerinin Belirlenmesi

Kullanılacak tahmin modelleri eğitmek ve test etmek için orijinal veri seti kullanılır. Bu orijinal veri setinin bir kısmı modeli eğitmek için ayrılırken, diğer kısmı ise modeli test etmek için kullanılır. Modelin eğitimi için ayrılan veri setine, eğitim veri seti denir. Modelin tahmin performansını test etmek için kullanılan veri setine ise test veri seti denir. Belirlenen test veri seti ile model eğitilemeyeceği unutulmamalıdır. Bunun yanı sıra, kullanılacak olan eğitim ve test veri setlerinin hacimleri birbirinden farklıdır. Eğitim veri setinde yer alan gözlem sayısı, test veri setinde bulunan gözlem sayısından fazladır. Bunun sebebi, kurulacak olan modelin, problemi her yönüyle

öğrenmesini istemektir. Eğitim veri setindeki gözlem değerlerinin fazla olması, modelin problem hakkında daha fazla bilgi sahibi olmasına imkân sağlayacaktır. Veri seti içerisindeki bilinmeyen yapıları keşfederek daha güçlü tahminler yapmasına da neden olacaktır. Çünkü, bir modelin güçlü tahminler yapmasının yolu ancak modelin veri setini çok iyi tanımasında geçer.

Çalışmada kullanılan her bir veri seti için oluşturulan eğitim ve test veri setlerinin sayısı Tablo 6’da gösterildiği gibidir. Ayrıca Tablo 6’da veri setindeki toplam gözlem sayısı ve veri setinde yer alan sınıfların dengesi hakkında da bilgi sahibi olmak mümkündür.

Tablo 6: Her Bir Veri Seti için Oluşturulan Eğitim ve Test Veri Setleri

<i>Veri Setleri</i>	<i>Toplam Gözlem Sayısı</i>	<i>Sınıf Dengesi (Çoğunluk: Azınlık)</i>	<i>Eğitim Veri Seti</i>	<i>Test Veri Seti</i>
<i>Alman Kredi Veri Seti</i>	1000	700:300	800	200
<i>Avustralya Kredi Veri Seti</i>	690	383:307	565	125
<i>HMEQ Kredi Veri Seti</i>	5960	4771:1189	4887	1073

3.3.5. Değişken Seçimi Yapılması

Gerçek hayat problemlerine ait veri setleri, gereğinden fazla değişken bünyesinde barındırabilir. Çok fazla girdi değişkenine sahip veri setleri ile modeli eğitmeye çalışmak, tahmin performansını düşürür ve en optimal çözüm değerini bulmak zorlaşır. Bu durum, araştırma için maliyet ve zaman kaybı demektir. Bu durumun üstesinden gelebilmek için değişken seçim yöntemleri geliştirilmiştir. Değişken seçim yöntemleri, farklı yöntem ve araçların yardımıyla problem ile değişkenler arasındaki ilişkiyi daha anlaşılır hale getirmeye yardımcı olur. Veri setinde yer alan tüm değişkenler arasından, problemle ilişkisi en zayıf olan girdi değişkenler tespit edilir. Tespit edilen bu değişkenler, veri setinden elimine edilir. Böylece veri

setinde yer alan alakasız değişkenlerin sayısı azaltılarak, daha hızlı çalışan ve daha doğru sonuçlar tahmin eden modeller elde edilmiş olur. Ayrıca değişken sayısının fazla oluşu, süreci karmaşık bir hale getireceğinden modelin ezber yeteneğini arttırır. Başka bir deyişle, değişken sayısının artışı modelin veri setini aşırı öğrenmesine (overfitting) neden olur. İşte bu nedenle, veri setindeki değişkenlere boyut indirgemesi yapılması gerekebilir. Böylece, modellerin veri setine göstereceği aşırı uyum riski kontrol altına alınmış olur. Veri setinden anlamlı bilgileri çekmek için etkili bir araç olan değişken seçim teknikleri; filtreleme (filter), sarmalayıcı (wrapped) yöntemler ve gömülü (embedded) yöntemler olmak üzere üç alt başlıkta incelenir.

Çalışma için yürütülen uygulamada, gömülü (embedded) değişken seçim yöntemlerinden biri olan destek vektör makineleri-yinelemeli değişken eleme yöntemi kullanılmıştır. Kullanılan değişken seçimi yöntemine ait detaylı bilgi, bir sonraki alt başlıkta yer almaktadır.

3.3.5.1. Destek Vektör Makinesi–Yinelemeli Değişken Eleme Yöntemi

Destek vektör makinesi–yinelemeli değişken eleme yöntemi (*support vector machine-recursive feature elimination* veya *SVM-RFE*), Guyon ve diğerleri (2002) tarafından geliştirilen bir değişken seçim yöntemidir. Başlangıçta gen seçimi için kullanılan bu yöntem, zamanla farklı problemlere ait veri setleri içerisinde yer alan önemli girdi değişkenlerin belirlenmesinde kullanılmaya başlanmıştır. SVM-RFE yöntemi amacı, veri setinde alakasız değişkenleri veri setinden eleyerek modellerin performansını optimize edecek değişkenleri belirlemeye çalışmaktır. Alakasız değişkenlerin tespitini ise değişkenler için hesapladığı sıralama değeri ile yapar. Sıralama değerlerini hesaplamak için Destek vektör makinesi modelinden elde edilen ağırlık vektörüne ihtiyaç duyulur. Destek vektör makinesinden elde edilen bu ağırlık vektörü yardımıyla, her bir değişken için bir sıralama değeri hesaplar. Hesaplanan sıralama değerlerine dayanarak, değişkenlerin yüksek değerden düşük değere doğru sıralandığı bir alt küme oluşturulur. Daha sonra, bu alt kümeden, en düşük sıralama değerine sahip değişkenler geriye doğru yinelemeli olarak alt kümeden çıkarılır. Değişken seçimine ait bu süreç, önemli değişkenler belirlenene kadar, alt kümede kalan değişkenler üzerinden tekrarlanır.

Tablo 7: Veri Setinde Bulunan Tüm Değişkenler ile Seçilen Değişkenlerin Sayısı

<i>Veri Setleri</i>	<i>Orijinal Veri Setindeki Toplam Girdi Değişken Sayısı</i>	<i>Kodlamadan Sonraki Toplam Değişken Sayısı</i>	<i>Değişken Seçim Yöntemiyle Seçilen Değişken Sayısı</i>
<i>Alman Kredi Veri Seti</i>	20	50	37
<i>Avustralya Kredi Veri Seti</i>	14	42	27
<i>HMEQ Kredi Veri Seti</i>	12	18	13

3.3.6. Yeniden Örneklem Yöntemlerinin Uygulanması

Gerçek hayat problemlerine ait veri setlerinin en büyük problemlerin biri de, veri seti olarak dengeli bir yapıya sahip olmamasıdır. Veri setleri üzerinde hâkim olan bu dengesizlik durumu, makine öğrenmesi modellerini performans açısından zayıflatacaktır. Bir modelin performans olarak zayıflaması demek, o modelin tahmin başarısının düşük olması demektir. Dengeli olmayan veri setleri ile eğitilen modeller veri setinde sayıca hangi sınıf üstünse, o sınıf daha baskın gelecektir ve baskın olan sınıfa göre, yanlı bir tahminleme yapacaktır. Bu durum, modelin sunacağı nihai tahmin değerine olan güveni de azaltacaktır. Bahsedilen sıkıntıların üstesinden gelmek için, dengeli yapıya sahip olmayan veri setlerini dengeli hale getiren birçok farklı teknik geliştirilmiştir.

Yürütülen uygulamadaki amaçlardan biri de farklı örneklem tekniklerini uygulayarak sınıflandırma modellerinin tahmin performansını hangi oranda iyileştirdiğini karşılaştırmaktır. Bu nedenle, çalışmada veri setlerini dengelemek adına toplamda 9 tane yeniden örneklem tekniği kullanıldı. Bunlardan 5'i yukarı örneklem (oversampling) tekniği, geriye kalan 4'ü ise aşağı örneklem (undersampling) tekniğidir. Kullanılan yukarı örneklem teknikleri; Rastgele yukarı örneklem (ROS), ADASYN, SMOTE, Borderline-SMOTE ve SVM-SMOTE teknikleridir. Kullanılan aşağı örneklem teknikleri ise Rastgele aşağı örneklem (RUS), near miss, cluster centroids ve tomed-links teknikleridir. Kullanılan tekniklere ait detaylı bilgi, çalışmanın ikinci bölümünde bulunan “Yeniden Örneklem Teknikleri” adlı başlık

altında yer almaktadır. Ek olarak, bu çalışmanın uygulaması için belirlenen deney tasarımı, veri setini dengelemek için uygulanacak yeniden örnekleme teknikleri, değişken seçimi yapıldıktan sonra uygulanmıştır.

Tablo 8: Toplam Gözlem Sayısı: Alman Kredi Veri Seti

<i>Alman Kredi Veri Seti</i>			
	<i>Örnekleme Yöntemleri</i>	<i>Örnekleme Yapılmadan Önce Eğitim Veri Seti</i>	<i>Örnekleme Yapıldıktan Sonra Eğitim Veri Seti</i>
<i>Yukarı Örnekleme</i>	ROS/ SMOTE/ SVM SMOTE/ Borderline SMOTE	800	1122
	ADASYN	800	1086
<i>Aşağı Örnekleme</i>	Cluster Centroids/ NearMiss/ RUS	800	476
	Tomek-Links	800	749

Tablo 9: Toplam Gözlem Sayısı: Avustralya Kredi Veri Seti

<i>Avustralya Kredi Veri Seti</i>			
	<i>Örnekleme Yöntemleri</i>	<i>Örnekleme Yapılmadan Önce Eğitim Veri Seti</i>	<i>Örnekleme Yapıldıktan Sonra Eğitim Veri Seti</i>
<i>Yukarı Örnekleme</i>	SMOTE/ SVM SMOTE/ Borderline-SMOTE	565	610
	ROS	565	616
	ADASYN	565	599
<i>Aşağı Örnekleme</i>	Cluster Centroids	565	519
	NearMiss	565	454
	RUS	565	513
	Tomek-Links	565	539

Tablo 10: Toplam Gözlem Sayısı: HMEQ Kredi Veri Seti

<i>HMEQ Kredi Veri Seti</i>			
<i>Örnekleme Yöntemleri</i>		<i>Örnekleme Yapılmadan Önce Eğitim Veri Seti</i>	<i>Örnekleme Yapıldıktan Sonra Eğitim Veri Seti</i>
<i>Yukarı Örnekleme</i>	SMOTE/ SVM SMOTE/ Borderline-SMOTE	4887	7840
	ROS	4887	7846
	ADASYN	4887	7828
<i>Aşağı Örnekleme</i>	Cluster Centroids/ NearMiss/ RUS	4887	1946
	Tomek-Links	4887	4871

3.3.7. Modelin Kurulumu

İlgili problem için nihai sonuca ulaşmaya yardımcı olan en kilit aşamalardan biri de modelin doğru kurulmasıdır. Modelin doğru araç ve parametre değerleri ile kurulması, amaçlanan optimal nihai tahmin değerlerinin elde edilmesine olanak tanır. Ayrıca parametre değerlerinin doğru belirlenmesi, modelin yaşayabileceği aşırı öğrenme ve eksik öğrenme maruziyetini en aza indirerek, bu iki durum arasındaki dengenin korunmasını sağlayacaktır. Ancak, optimal sonuca ulaşmak sanıldığı kadar da kolay değildir. Optimum model yapısına ulaşmak için farklı parametre kombinasyonlarını tek tek denemek gerekebilir. Bu iş, araştırmacı için oldukça zahmetli olacaktır. Problem yapısına ve veri setine göre değer aralığı değişen parametreler, hiper parametre (hyperparameter) olarak adlandırılır. Hiper parametreler model parametreleri ile karıştırılmamalıdır. Model parametrelerini hiper parametrelerden ayıran en önemli özellik, model tarafından öğrenilme durumudur. Model parametreleri eğitim aşamasında tekrarlı olarak model tarafından öğrenilirken, hiper parametreler ise model tarafından öğrenilmezler. Aynı zamanda hiper parametreler, modelden haricidir ve eğitim veri setleriyle belirlenemezler. Ayrıca hiper parametreler, modelin eğitilmesine başlanmadan önce belirlenir. Modelin

eđitimi boyunca belirlenen hiper parametreler aynı kalmaya devam eder. Bu parametre türü için deęer aralıđını belirlemek, yalnızca arařtırmacının tekelindedir. Arařtırmacı elinde olan veri seti ve ele aldıđı problem için çeřitli deęerleri içeren kombinasyonlar oluşturur. Bu kombinasyonlar önceden ele alınmıř, benzer veri setleri ve problemler ıřıđında oluşturulabilir. Dahası, arařtırmacı kendi sezgilerine ve edindiđi tecrübelerine dayanarak da hiper parametre için çeřitli deęer kombinasyonları oluşturabilir. Bu nedenle hem arařtırmacının iř yükünü azaltan hem de modelin optimizasyonu için uygun parametrelerin kolayca belirlenmesine yardımcı olan farklı hiper parametre seçim yöntemleri geliştirilmiřtir. Bařka bir deyiřle, arařtırmacının belirlediđi farklı parametre kombinasyonları arasından en uygun kombinasyonu otomatik olarak belirleme gereksinimi, rastgele arama (random search) ve ızgara arama (grid search) gibi farklı hiper parametre seçim yöntemlerinin ortaya çıkmasına yol açmıřtır. Hiper parametre seçim yöntemlerinden bahsetmeden önce, hiper parametre seçimi hakkında bilgi vermekte fayda vardır.

Hiper parametre seçimi (hyperparameter tuning), modelin optimizasyonu için en uygun hiper parametre deęerlerini seçme sürecine denir. Hiper parametre seçim süreci, üç ana ařamadan oluřmaktadır; Hiper parametre deęerlerini belirleme, belirlenen deęerlerin farklı kombinasyonlarıyla modeli eđitme ve modelin performansını ölçme. Hiper parametre seçiminin, modelin optimizasyonunu sađlama konusunda avantaj sađlasa da dezavantajı da mevcuttur. Belirlenen hiper parametre deęerleri, modelin karmařıklıđı ya da eđitim veri setinin boyutuna bađlı olarak iřlem süresinde meydana gelecek uzama, hesaplama maliyetinde artışa neden olabilir.

Izgara arama (grid search) ve rastgele arama (random search) teknikleri, günümüzde yapılan çalıřmalarda yaygın olarak kullanılan hiperparametre seçim tekniklerindendir. Birbirleriyle kıyaslandıđında farklı avantajlara ve dezavantajlara sahip olan bu seçim teknikleri, özellikle model için seçilecek optimal hiperparametre deęerlerinin otomatik olarak tespit edilmesi konusunda farklı stratejiler izlerler.

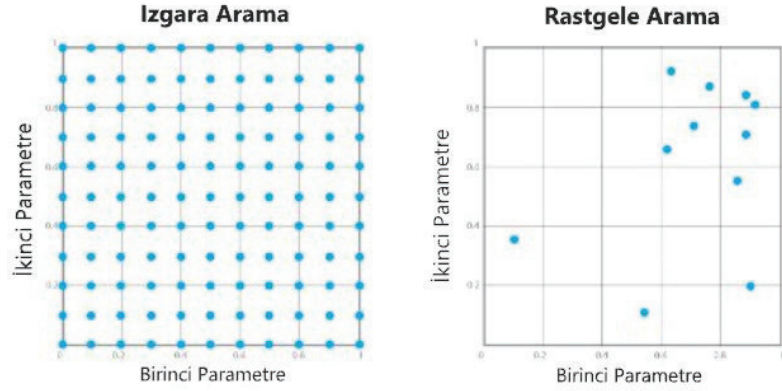
Rastgele arama, belirlenen hiper parametre deęerleri arasından rastgele seçilen deęerlere dayanarak model optimizasyonunu sađlamaya çalıřan bir arama yöntemidir. Aynı zamanda rastgele arama, belirlenen tüm hiper parametre deęerlerinin tamamını kullanmaz. Bu nedenle hesaplama maliyetleri oldukça düşüktür. Onun yerine rastgele olarak seçilen hiper parametre deęerlerinden farklı kombinasyonlar oluşturur.

Belirlenen kombinasyonlarla model eğitilir ve model performansını optimize eden değer kombinasyonları belirlenmeye çalışılır.

Izgara arama, araştırmacı tarafından belirlenen hiper parametre değerlerinin tamamını kullanarak, modeli optimize etmeye çalışan bir arama yöntemidir. Bu hiper parametre seçim yöntemi, belirlenen hiper parametre değerlerinin kartezyen çarpımıyla değerlendirme yapar. Bu yöntemin arama alanı, çok boyutlu eş hücrelerdir. Bu çok boyutlu eş hücrelerin tamamı, ızgara (grid)'yı oluşturur. Oluşturulan eş hücrelerin köşegenlerindeki noktalar, aslında hiper parametre kombinasyonlarını oluşturur. Izgara aramada, aramanın yapılacağı her bir hücrede yer alan hiper parametre değer kombinasyonlarıyla model eğitilir ve değerlendirilir. Rastgele aramanın aksine, ızgara aramada hiper parametre değerleri rastgele seçilmez. Izgara arama, hiper parametre değer kombinasyonlarının tamamını denemek zorundadır. Bu özellik, ızgara aramanın yüksek bir hesaplama maliyetine sahip olmasına neden olur.

Tüm bunlara ek olarak, ızgara arama, bilgisiz arama (uninformed search) grubunda yer alır. Izgara aramanın, arama yaptığı bir hücredeki kombinasyon değerlerinden elde ettiği sonuç, yine yöneme, yapması gereken bir sonraki hamle hakkında ipucu vermez. Bunun sebebi, aramanın yapıldığı hücredeki kombinasyonu kullanması sonucu elde edilen bilgileri, arama yapacağı diğer hücrelere taşımamasıdır. Böylece, bir modeli oluşturduktan sonra akabinde oluşturulacak bir diğer modelin, gösterdiği performans açısından optimal bir model olup olmayacağını kestiremez. Izgara arama hiper parametre seçim yönteminin, dezavantajı olduğu kadar avantajı da söz konusudur. Izgara arama, yaptığı kapsamlı arama yapması sebebiyle en iyi sonucu sağlayan bir hiper parametre seçim yöntemi olma özelliğine sahiptir. Izgara arama ve rastgele arama yöntemlerinin sahip olduğu arama uzayı, şekil 40' daki gibidir. Çalışma için kullanılan her bir sınıflandırma modelinin optimize edilebilmesi için hiper parametre seçimi yönteminden biri olan ızgara arama (grid search) kullanılmıştır.

Şekil 41: Izgara (Grid) Arama ve Rastgele (Random) Arama



Kaynak: Sharma, 2016

3.3.8. Modellerin Performans Ölçümü

Kullanılan modellerin tahmin performansını ölçmek için çeşitli performans ölçüm metrikleri kullanılır. Bu ölçüm metrikleri bize kullanılan tahmin modellerinin performansını karşılaştırma olanağı tanır. Bu çalışma için yürütülen uygulamada, kullanılan sınıflandırma modellerinin tahmin performanslarını belirlemek için 5 farklı performans ölçüm istatistiği kullanılmıştır. Kullanılan istatistikler; doğruluk, precision, recall, F1-skoru, MCC değerleridir. Kullanılan metriklere ait detaylı bilgi, çalışmanın ikinci bölümde bulunan “*Performans Ölçüleri*” adlı başlık altında yer almaktadır.

3.4. BULGULAR VE BULGULARIN YORUMLANMASI

Çalışmanın bu bölümünde, farklı yeniden örnekleme teknikleri kullanılarak dengeli hale getirilen kredi veri setleriyle birlikte, sınıflandırma başarısı ölçülen topluluk ve tekil modellere ait performans istatistikleri yer almaktadır. Ölçülen performans istatistikleri, çalışma kapsamında oluşturulan senaryolara göre ayrı ayrı tablolar haline getirilmiştir. Sonrasında, ilgili kredi veri setini sınıflandırma konusunda optimal performansın elde edildiği yeniden örnekleme tekniği-model kombinasyonlarına ait en yüksek değerler, tablo üzerinde koyulaştırılmış ve bu değerler yorumlanmıştır.

3.4.1. Alman Kredi Veri Setine ait Performans İstatistikleri

Tablo 11: Senaryo 1: Alman Kredi Veri Seti Performans İstatistikleri

Aşağı Örneklemeye	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	RF	0.6950	0.7742	0.5053	0.6115	0.4016
	GBDT	0.6950	0.7419	0.5055	0.6013	0.3862
	XTrees	0.6850	0.7581	0.4947	0.5987	0.3799
	Hard	0.7000	0.7903	0.5104	0.6203	0.4163
	Soft	0.7100	0.7742	0.5217	0.6234	0.4225
Near Miss	RF	0.6500	0.7419	0.4600	0.5679	0.3243
	GBDT	0.6450	0.7581	0.4563	0.5697	0.3260
	XTrees	0.6600	0.7258	0.4688	0.5696	0.3298
	Hard	0.6800	0.7097	0.4889	0.5789	0.3499
	Soft	0.6950	0.7903	0.5052	0.6164	0.4095
Cluster Centroids	RF	0.6100	0.7903	0.4298	0.5568	0.2983
	GBDT	0.6300	0.8387	0.4483	0.5843	0.3513
	XTrees	0.6750	0.7581	0.4845	0.5912	0.3662
	Hard	0.6900	0.7903	0.5000	0.6125	0.4027
	Soft	0.6100	0.8226	0.4322	0.5667	0.3170
Tomek-Links	RF	0.7550	0.5323	0.6226	0.5739	0.4059
	GBDT	0.7550	0.6129	0.6032	0.6080	0.4299
	XTrees	0.7700	0.5806	0.6429	0.6102	0.4488
	Hard	0.7400	0.6129	0.5758	0.5938	0.4033
	Soft	0.7300	0.6290	0.5571	0.5909	0.3921

Tablo 11’de farklı aşağı örneklemeye teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait bazı performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi uygulanan ve aşağı örneklemeye teknikleriyle dengelenen Alman kredi veri seti aracılığıyla sınıflandırma başarısı test edilen topluluk modellerinin tahmin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %77 ile Tomek links-XTrees aşağı örneklemeye tekniği-model kombinasyonunda ölçülmüştür. Ancak, sadece doğruluk değerleri göz önünde bulundurularak, makine öğrenmesi tekniklerinin performansı hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örneklemeye tekniği-topluluk öğrenmesi modeli kombinasyonu belirlerken, kullanılan diğer tüm ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda, doğruluk ölçüm istatistiğine ek olarak kesinlik (precision) ve MCC performans ölçüm istatistiklerine bakıldığında da en iyi performans yine Tomek links-XTrees kombinasyonunda ölçülmüştür. Duyarlılık (recall) açısından en yüksek tahmin performansı cluster centroids-XTrees kombinasyonunda ölçülürken, F1-skor

açısından en iyi performans RUS-soft oylama kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 1’de, kredi riski bulunan müşterileri tahmin etme konusunda yüksek tahmin performansı gösteren aşağı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, Tomek links-XTrees’e aittir; yani değişken seçim tekniği uygulanan ve Tomek links aşağı örnekleme tekniğiyle dengelenen Alman kredi veri setinde yer alan riskli müşterileri, senaryo 1’de en doğru tahmin eden topluluk öğrenmesi modeli ekstra ağaçlar (XTrees)’dir.

Tablo 12: Senaryo 2: Alman Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	KNN	0.6850	0.7258	0.4945	0.5882	0.3645
	NB	0.7250	0.8226	0.5368	0.6497	0.4665
	LR	0.6900	0.8226	0.5000	0.6220	0.4191
	SVM	0.6700	0.8226	0.4811	0.6071	0.3929
	MLP	0.7300	0.6935	0.5513	0.6143	0.4171
	DT	0.6000	0.8065	0.4237	0.5556	0.2950
Near Miss	KNN	0.5750	0.1452	0.2195	0.1748	-0.0994
	NB	0.4450	0.8710	0.3439	0.4932	0.1403
	LR	0.6600	0.5484	0.4595	0.5000	0.2477
	SVM	0.6100	0.7581	0.4273	0.5465	0.2803
	MLP	0.6050	0.7742	0.4248	0.5486	0.2828
	DT	0.6550	0.7097	0.4632	0.5605	0.3150
Cluster Centroids	KNN	0.6200	0.8387	0.4407	0.5778	0.3389
	NB	0.3400	0.9677	0.3158	0.4762	0.0546
	LR	0.7400	0.7581	0.5595	0.6438	0.4591
	SVM	0.7300	0.7258	0.5488	0.6250	0.4304
	MLP	0.5600	0.3710	0.3194	0.3433	0.0153
	DT	0.6650	0.7419	0.4742	0.5786	0.3446
Tomek-Links	KNN	0.6800	0.5323	0.4853	0.5077	0.2720
	NB	0.3700	0.9677	0.3261	0.4878	0.1180
	LR	0.7650	0.6935	0.6056	0.6466	0.4742
	SVM	0.7450	0.5645	0.5932	0.5785	0.3961
	MLP	0.7450	0.6290	0.5821	0.6047	0.4176
	DT	0.7000	0.5000	0.5167	0.5082	0.2925

Tablo 12’de farklı aşağı örnekleme teknikleri ve tekil model kombinasyonlarına ait bazı performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve aşağı örnekleme teknikleriyle dengelenen Alman kredi veri seti aracılığıyla sınıflandırma başarısı test edilen tekil modellerin tahmin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %76,50 ile Tomek links-LR

aşağı örnekleme tekniği-model kombinasyonunda ölçülmüştür. Ancak sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği ve tekil model kombinasyonlarını belirlerken, kullanılan diğer tüm performans ölçüm istatistiğine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiğine ek olarak, kesinlik (precision) ve MCC performans ölçüm istatistiklerine bakıldığında da en iyi performans yine Tomek links-LR kombinasyonunda ölçülmüştür. Duyarlılık (recall) açısından en yüksek tahmin performansı, hem Tomek links-NB hem de cluster centroids-NB kombinasyonlarına ölçülürken, F1-skor açısından en iyi performans ise, RUS-NB kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin tamamı göz önünde bulundurulduğunda Senaryo 2’de, kredi riski bulunan müşterileri tahmin etme konusunda yüksek tahmin performansı gösteren aşağı örnekleme tekniği-tekil model kombinasyonu, Tomek links-LR’ ye aittir; yani değişken seçim tekniği uygulanan ve Tomek links aşağı örnekleme tekniğiyle dengelenen Alman kredi veri setinde yer alan riskli müşterileri, senaryo 2’de en doğru tahmin eden tekil model lojistik regresyon (LR)’dur.

Tablo 13: Senaryo 3: Alman Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	RF	0.7050	0.7742	0.5161	0.6194	0.4155
	GBDT	0.6600	0.7742	0.4706	0.5854	0.3542
	XTrees	0.6950	0.7742	0.5053	0.6115	0.4016
	Hard	0.7100	0.7903	0.5213	0.6282	0.4302
	Soft	0.7150	0.7903	0.5269	0.6323	0.4372
Near Miss	RF	0.6950	0.8065	0.5051	0.6211	0.4175
	GBDT	0.6700	0.7419	0.4792	0.5823	0.3514
	XTrees	0.6750	0.7903	0.4851	0.6012	0.3825
	Hard	0.6800	0.7419	0.4894	0.5897	0.3652
	Soft	0.6950	0.6935	0.5059	0.5850	0.3641
Cluster Centroids	RF	0.6400	0.8387	0.4561	0.5909	0.3638
	GBDT	0.5900	0.8065	0.4167	0.5495	0.2825
	XTrees	0.6950	0.7742	0.5053	0.6115	0.4016
	Hard	0.6500	0.8710	0.4655	0.6067	0.3952
	Soft	0.5850	0.8871	0.4198	0.5699	0.3273
Tomek-Links	RF	0.7750	0.5806	0.6545	0.6154	0.4588
	GBDT	0.7750	0.6290	0.6393	0.6341	0.4717
	XTrees	0.7350	0.5161	0.5818	0.5470	0.3620
	Hard	0.7450	0.6129	0.5846	0.5984	0.4120
	Soft	0.7700	0.6613	0.6212	0.6406	0.4722

Tablo 13'te farklı aşağı örnekleme teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve aşağı örnekleme teknikleriyle dengelenen Alman kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %77,50 ile hem Tomek links-RF hem de Tomek links-GBDT aşağı örnekleme tekniği-model kombinasyonlarında ölçülmüştür. Ancak, sadece doğruluk değerleri göz önünde bulundurularak, makine öğrenmesi tekniklerinin performansı hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu belirlerken, kullanılan diğer performans ölçüm istatistiklerine de bakmakta fayda vardır. Duyarlılık (recall) açısından en iyi tahmin performansı, cluster centroids-soft oylama kombinasyonunda ölçülmüştür. Kesinlik (precision) açısından en iyi performans, Tomek links-RF kombinasyonunda ölçülmüştür. F1-skor ve MCC açısından en iyi performans, Tomek links-soft oylama kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 3'de kredi riski bulunan müşterileri tahmin etme konusunda yüksek performans gösteren aşağı örnekleme tekniği-topluluk modeli kombinasyonu, Tomek links-soft oylamaya aittir; yani değişken seçim tekniği uygulanmayan ve Tomek links aşağı örnekleme tekniğiyle dengelenen Alman kredi veri setinde yer alan riskli müşterileri, senaryo 3'te en doğru tahmin eden topluluk öğrenmesi modeli yumuşak oylamadır.

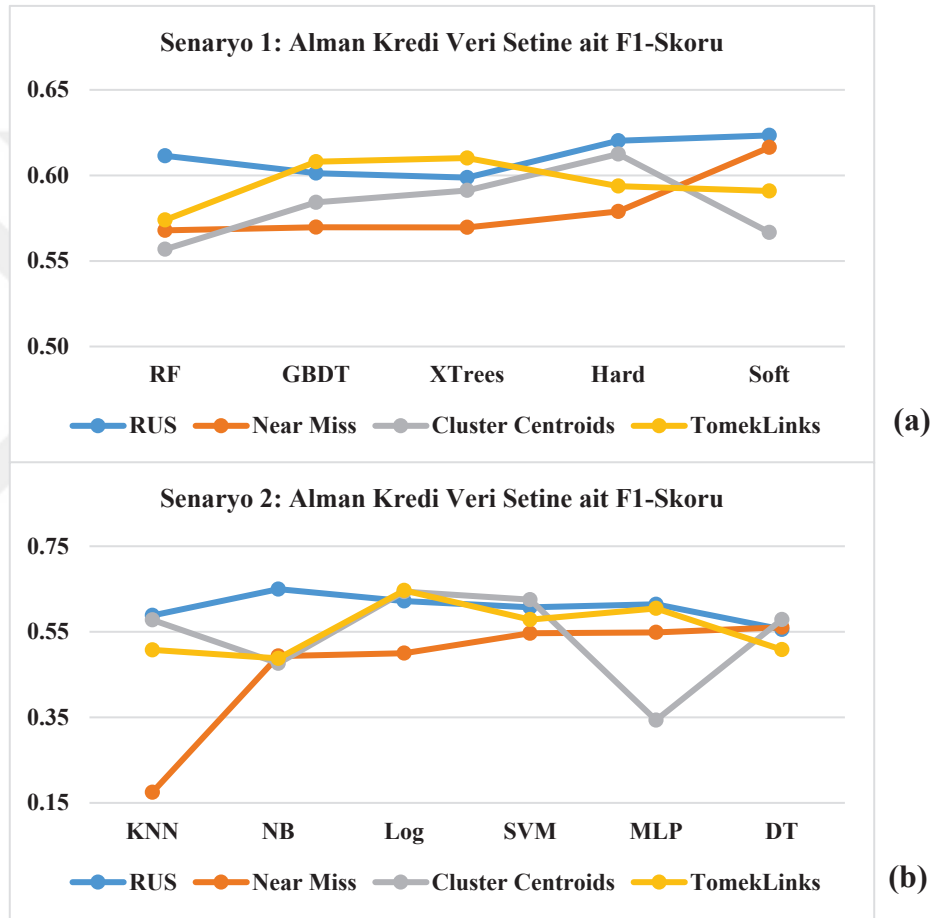
Tablo 14: Senaryo 4: Alman Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	KNN	0.6850	0.6935	0.4943	0.5772	0.3496
	NB	0.7250	0.8548	0.5354	0.6584	0.4824
	LR	0.7100	0.8065	0.5208	0.6329	0.4380
	SVM	0.7150	0.7903	0.5269	0.6323	0.4372
	MLP	0.6550	0.7258	0.4639	0.5660	0.3230
	DT	0.6000	0.8065	0.4237	0.5556	0.2950
Near Miss	KNN	0.6800	0.5806	0.4865	0.5294	0.2924
	NB	0.6200	0.1290	0.2667	0.1739	-0.0394
	LR	0.6850	0.7419	0.4946	0.5935	0.3722
	SVM	0.6500	0.7258	0.4592	0.5625	0.3162
	MLP	0.6400	0.6774	0.4468	0.5385	0.2786
	DT	0.6000	0.8065	0.4237	0.5556	0.2950
Cluster Centroids	KNN	0.5900	0.9032	0.4242	0.5773	0.3442
	NB	0.3650	0.9194	0.3184	0.4730	0.0533
	LR	0.6800	0.7581	0.4896	0.5949	0.3731
	SVM	0.6650	0.8226	0.4766	0.6036	0.3865
	MLP	0.7000	0.6935	0.5119	0.5890	0.3715
	DT	0.6050	0.7903	0.4261	0.5537	0.2920
Tomek-Links	KNN	0.6750	0.5323	0.4783	0.5038	0.2640
	NB	0.4250	0.9516	0.3450	0.5064	0.1839
	LR	0.7500	0.6774	0.5833	0.6269	0.4432
	SVM	0.7300	0.5323	0.5690	0.5500	0.3579
	MLP	0.6800	0.3871	0.4800	0.4286	0.2122
	DT	0.7550	0.6129	0.6032	0.6080	0.4299

Tablo 14’te farklı aşağı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve aşağı örnekleme tekniklerinin uygulandığı Alman kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen tekil modellerin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %75,50 ile Tomek links-DT aşağı örnekleme tekniği-model kombinasyonunda ölçülmüştür. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği ve tekil model kombinasyonu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda, doğruluk ölçüm istatistiğine ek olarak, kesinlik (precision) performans ölçüm istatistiğine göre en iyi performans, Tomek links-DT kombinasyonunda ölçülmüştür. F1-skor ve MCC açısından en iyi performans, RUS-NB kombinasyonunda ölçülmüştür. Duyarlılık (recall) açısından en yüksek tahmin performansı, tomek links-NB kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda

yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 4'te kredi riski bulunan müşterileri tahmin etme konusunda yüksek tahmin performansı gösteren aşağı örnekleme tekniği-tekil model kombinasyonu RUS-NB'ye aittir; yani değişken seçim tekniği uygulanmayan ve RUS aşağı örnekleme tekniğiyle dengelenen Alman kredi veri setinde yer alan riskli müşterileri, senaryo 4'te en doğru tahmin eden tekil model Naive Bayes (NB) dir.

Şekil 42: Senaryo 1 ve 2'ye ait F1-Skoru Grafiği: Alman Kredi Veri Seti



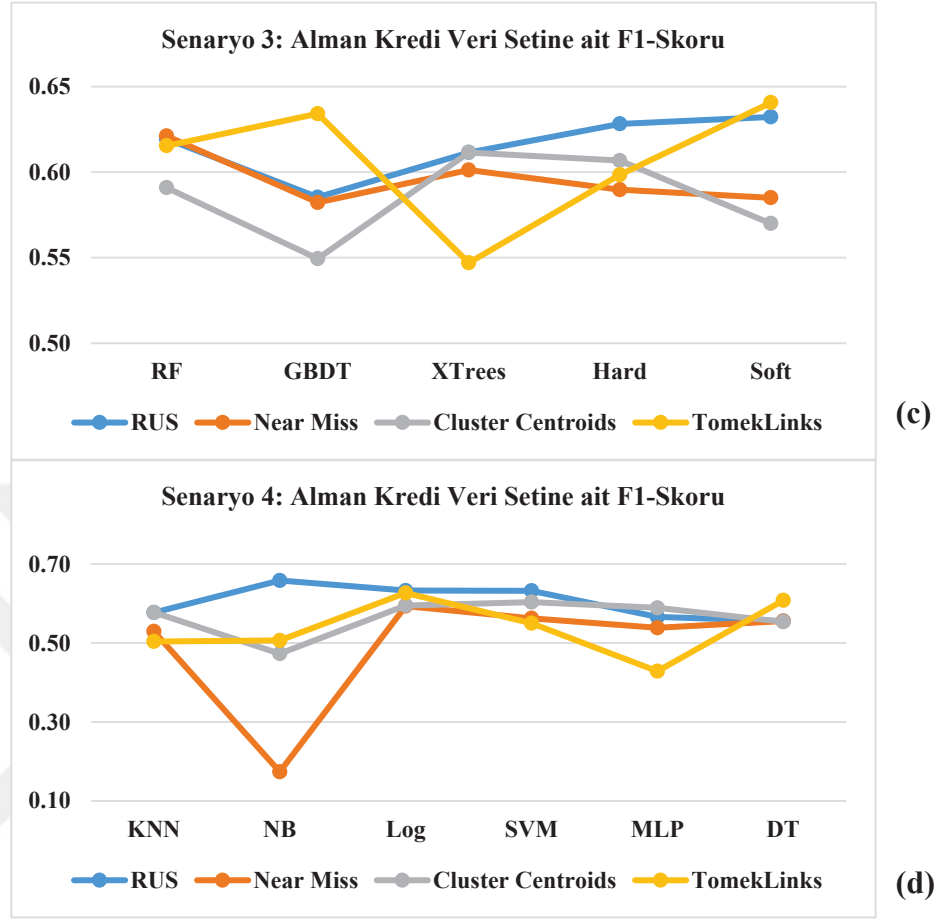
Yukarıdaki çizgi grafikleri, aşağı örnekleme teknikleriyle dengelenen Alman kredi veri seti üzerinde sınıflandırma performansı ölçülen tekil ve topluluk modellerine ait F1-skoru değerlerinin dağılımını göstermektedir. Bu grafiklere bakılarak model ve aşağı örnekleme teknikleri için elde edilen çıkarımlar ise şöyledir;

(a) grafiğine göre; kullanılan tüm topluluk öğrenmesi modellerinin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model hard oylamadır. Aynı zamanda, tüm topluluk modelleri içerisinde genel olarak en düşük performansa

sahip model rastgele ormanlar (RF)'dır. Topluluk modellerinin nerdeyse tamamı, en yüksek tahmin performansını RUS aşağı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir. Kullanılan topluluk öğrenmesi tekniklerinde yer alan; rastgele ormanlar (RF), gradyan artırma karar ağaçları (GBDT) ve ekstra ağaçlar (XTrees) modelleri ağaç tabanlı modellerdir. (b) grafiğine göre; kullanılan tüm tekil modellerin arasından ortalama olarak en düşük tahmin performansını gösteren model KNN 'dir. KNN, diğer sınıflandırma modelleri ile karşılaştırıldığında ise özellikle, nearmiss aşağı örnekleme teknikleriyle dengelenen Alman kredi veri setinde en düşük tahmin performansını elde etmiştir. Bunun yanı sıra, kullanılan tüm tekil modellerin arasından ortalama olarak en yüksek tahmin performansına sahip model ise lojistik regresyon (LR)'dur. Tekil modellerin tamamı en yüksek tahmin performansını, RUS aşağı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir.

(a) ve (b) grafiklerine göre; kullanılan tüm sınıflandırma modellerinin arasından ortalama olarak en düşük performansa sahip model, k-en yakın komşu (KNN) modelidir. Bunun yanı sıra senaryo 1 ve 2'de kullanılan tekil ve topluluk öğrenmesi modellerinin performansları kıyaslandığında, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu söylenebilir.

Şekil 43: Senaryo 3 ve 4’e ait F1-Skoru Grafiği: Alman Kredi Veri Seti



(c) grafiğine göre; kullanılan tüm topluluk modellerinin arasından ortalama en yüksek tahmin performansının ölçüldüğü model rastgele ormanlar (RF)’dır. Rastgele ormanlar modeli, senaryo 1’de en düşük performansa sahip topluluk öğrenmesi tekniği olarak belirlense de senaryo 3’te en yüksek tahmin performansını gösteren topluluk öğrenmesi tekniği olarak belirlenmiştir. Aynı zamanda, kullanılan tüm topluluk modellerinin arasından ortalama olarak en düşük performansa sahip model GBDT modelidir. Topluluk öğrenmesi tekniklerine ait en yüksek performans değerleri, RUS örnekleme tekniğiyle dengelenen veri setinde ölçülmüştür.

(d) grafiğine göre; kullanılan tüm tekil modellerin arasından, ortalama olarak en düşük tahmin performansına sahip yöntem Naive Bayes (NB) yöntemidir. Bunun yanı sıra, kullanılan tüm tekil modellerin arasından ortalama olarak en yüksek tahmin performansının elde edildiği model lojistik regresyon (LR) modelidir. Tekil modeller en yüksek tahmin performansını, RUS örnekleme tekniğiyle dengelenen veri seti

üzerinde göstermiştir. Buna ek olarak, senaryo 2’de en düşük tahmin performansa sahip KNN tekil modeli, senaryo 4’te daha yüksek tahmin performansı göstermiştir.

(c) ve (d) grafiklerini kıyaslırsak eğer; kullanılan tüm sınıflandırma modellerinin arasından genel olarak en düşük performansın ölçüldüğü yöntem Naive Bayes (NB)’tir. Bunun yanı sıra senaryo 3 ve 4’te kullanılan tekil ve topluluk öğrenmesi modellerinin performansları kıyaslandığında, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu da söylenebilir.

(a) ve (c) grafikleri kıyaslırsak eğer; Senaryo 2’nin, Alman kredi veri setinde bulunan riskli müşterileri sınıflandırmada kullanılacak olan tüm tekil modellerin tahmin performansını arttırdığını söylemek mümkündür; yani senaryo 2, çalışmada kullanılan tüm tekil modellerin tahmin performansını arttırmada senaryo 4’e göre daha başarılı olmuştur.

(b) ve (d) grafiklerini kıyaslırsak eğer; Senaryo 1’in, Alman kredi veri setinde bulunan riskli müşterileri sınıflandırmak için kullanılacak olan topluluk öğrenmesi modellerinin tahmin performansını arttırdığını söylemek mümkündür; yani senaryo 1, çalışmada kullanılan tüm topluluk öğrenmesi modellerinin tahmin performansını arttırmada senaryo 3’e göre daha başarılı olmuştur.

Tablo 15: Senaryo 5: Alman Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	RF	0.7950	0.6129	0.6909	0.6496	0.5072
	GBDT	0.7550	0.5645	0.6140	0.5882	0.4150
	XTrees	0.7600	0.5161	0.6400	0.5714	0.4120
	Hard	0.7550	0.6129	0.6032	0.6080	0.4299
	Soft	0.7600	0.5968	0.6167	0.6066	0.4341
ADASYN	RF	0.7700	0.5806	0.6429	0.6102	0.4488
	GBDT	0.7600	0.5806	0.6207	0.6000	0.4293
	XTrees	0.7750	0.5968	0.6491	0.6218	0.4629
	Hard	0.7750	0.6613	0.6308	0.6457	0.4813
	Soft	0.7500	0.6129	0.5938	0.6032	0.4209
SMOTE	RF	0.7350	0.5000	0.5849	0.5391	0.3569
	GBDT	0.7500	0.6290	0.5909	0.6094	0.4263
	XTrees	0.7500	0.5161	0.6154	0.5614	0.3914
	Hard	0.7450	0.6129	0.5846	0.5984	0.4120
	Soft	0.7350	0.6129	0.5672	0.5891	0.3947
Borderline-SMOTE	RF	0.7800	0.6290	0.6500	0.6393	0.4813
	GBDT	0.7750	0.6452	0.6349	0.6400	0.4764
	XTrees	0.7600	0.5806	0.6207	0.6000	0.4293
	Hard	0.7600	0.6129	0.6129	0.6129	0.4390
	Soft	0.7450	0.6290	0.5821	0.6047	0.4176
SVM-SMOTE	RF	0.7700	0.5968	0.6379	0.6167	0.4532
	GBDT	0.7600	0.6129	0.6129	0.6129	0.4390
	XTrees	0.7550	0.5323	0.6226	0.5739	0.4059
	Hard	0.7350	0.5968	0.5692	0.5827	0.3889
	Soft	0.7150	0.5806	0.5373	0.5581	0.3488

Tablo 15’te farklı yukarı örnekleme teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve yukarı örnekleme tekniklerinin uygulandığı Alman kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %79,50 ile ROS-RF örnekleme tekniği-model kombinasyonuna aittir. Ancak sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm metriklerine de bakmakta fayda vardır. Bu bağlamda, doğruluk ölçüm istatistiğine ek olarak kesinlik (precision), F1-Skoru ve MCC performans ölçüm metriklerine bakıldığında da en iyi performans yine ROS-RF kombinasyonunda ölçülmüştür. Duyarlılık (recall) açısından en yüksek tahmin performansı, ADASYN-hard oylama kombinasyonunda ölçülürken, Sonuç

olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 5'te kredi riski bulunan müşterileri tahmin etme konusunda yüksek tahmin performansı gösteren yukarı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, ROS-RF' dir; yani, değişken seçimi uygulanan ve ROS yukarı örnekleme tekniğiyle dengelenen Alman kredi veri setinde yer alan riskli müşterileri, senaryo 5'te en doğru tahmin eden topluluk öğrenmesi modeli rastgele ormanlar (RF)'dir.

Tablo 16: Senaryo 6: Alman Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	KNN	0.7050	0.5000	0.5254	0.5124	0.3013
	NB	0.3600	0.9677	0.3226	0.4839	0.0991
	LR	0.7450	0.7903	0.5632	0.6577	0.4804
	SVM	0.7600	0.4032	0.6944	0.5102	0.3895
	MLP	0.7500	0.6613	0.5857	0.6212	0.4375
	DT	0.7250	0.4839	0.5660	0.5217	0.3324
ADASYN	KNN	0.7000	0.5000	0.5167	0.5082	0.2925
	NB	0.3650	0.9516	0.3224	0.4816	0.0880
	LR	0.7150	0.7903	0.5269	0.6323	0.4372
	SVM	0.7450	0.3871	0.6486	0.4848	0.3489
	MLP	0.7400	0.6613	0.5694	0.6119	0.4207
	DT	0.6900	0.5000	0.5000	0.5000	0.2754
SMOTE	KNN	0.6950	0.5000	0.5082	0.5041	0.2839
	NB	0.3750	0.9516	0.3260	0.4856	0.1066
	LR	0.6950	0.7903	0.5052	0.6164	0.4095
	SVM	0.7550	0.3871	0.6857	0.4948	0.3741
	MLP	0.7600	0.7097	0.5946	0.6471	0.4716
	DT	0.6850	0.4839	0.4918	0.4878	0.2604
Borderline-SMOTE	KNN	0.7000	0.5000	0.5167	0.5082	0.2925
	NB	0.3850	0.9516	0.3296	0.4896	0.1238
	LR	0.7400	0.7581	0.5595	0.6438	0.4591
	SVM	0.7500	0.3871	0.6667	0.4898	0.3613
	MLP	0.7150	0.6290	0.5342	0.5778	0.3676
	DT	0.6900	0.5323	0.5000	0.5156	0.2883
SVM-SMOTE	KNN	0.7050	0.5000	0.5254	0.5124	0.3013
	NB	0.3900	0.9516	0.3315	0.4917	0.1320
	LR	0.7050	0.8065	0.5155	0.6289	0.4311
	SVM	0.7650	0.4355	0.6923	0.5347	0.4068
	MLP	0.7500	0.6935	0.5811	0.6324	0.4492
	DT	0.6600	0.5323	0.4583	0.4925	0.2405

Tablo 16'da farklı yukarı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler,

değişken seçimi ve yukarı örnekleme tekniklerinin uygulandığı Alman kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen tekil modellerin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %76,50 ile SVMSMOTE-SVM kombinasyonuna aittir. Ancak sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-tekil model kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda, kesinlik (precision) performans ölçüm metriğine göre en iyi performans ROS-SVM kombinasyonunda ölçülmüştür. F1-skoru ve MCC performans ölçüm istatistiklerinde ise en iyi performans, ROS-LR yukarı örnekleme tekniği-tekil model kombinasyonunda ölçülmüştür. Duyarlılık (recall) açısından en iyi performans, ROS-NB örnekleme tekniği-tekil model kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 6'da kredi riski bulunan müşterileri tahmin etme konusunda yüksek tahmin performansı gösteren yukarı örnekleme tekniği-tekil model kombinasyonu, ROS-LR'dir; yani değişken seçimi tekniği uygulanan ve ROS yukarı örnekleme tekniğiyle dengelenen Alman kredi veri setinde yer alan riskli müşterileri, senaryo 6'da en doğru tahmin eden tekil model lojistik regresyon (LR) dur.

Tablo 17: Senaryo 7: Alman Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	RF	0.7800	0.5968	0.6607	0.6271	0.4729
	GBDT	0.8000	0.5806	0.7200	0.6429	0.5118
	XTrees	0.7500	0.4677	0.6304	0.5370	0.3787
	Hard	0.7650	0.5806	0.6316	0.6050	0.4390
	Soft	0.7550	0.5806	0.6102	0.5950	0.4198
ADASYN	RF	0.7700	0.5806	0.6429	0.6102	0.4488
	GBDT	0.7600	0.5806	0.6207	0.6000	0.4293
	XTrees	0.7750	0.5968	0.6491	0.6218	0.4629
	Hard	0.7750	0.6613	0.6308	0.6457	0.4813
	Soft	0.7500	0.6129	0.5938	0.6032	0.4209
SMOTE	RF	0.7500	0.5161	0.6154	0.5614	0.3914
	GBDT	0.7650	0.5645	0.6364	0.5983	0.4346
	XTrees	0.7750	0.5323	0.6735	0.5946	0.4477
	Hard	0.7550	0.5484	0.6182	0.5812	0.4104
	Soft	0.7450	0.5484	0.5965	0.5714	0.3911
Borderline-SMOTE	RF	0.7450	0.5323	0.6000	0.5641	0.3862
	GBDT	0.7600	0.5484	0.6296	0.5862	0.4203
	XTrees	0.7600	0.5161	0.6400	0.5714	0.4120
	Hard	0.7400	0.5323	0.5893	0.5593	0.3766
	Soft	0.7450	0.5645	0.5932	0.5785	0.3961
SVM-SMOTE	RF	0.7550	0.5806	0.6102	0.5950	0.4198
	GBDT	0.7600	0.6129	0.6129	0.6129	0.4390
	XTrees	0.7650	0.5323	0.6471	0.5841	0.4264
	Hard	0.7600	0.6129	0.6129	0.6129	0.4390
	Soft	0.7500	0.6129	0.5938	0.6032	0.4209

Tablo 17’de farklı yukarı örnekleme teknikleri ve topluluk öğrenmesi modelleri kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve yukarı örnekleme tekniklerinin uygulandığı Alman kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk öğrenmesi modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %80 ile ROS-GBDT örnekleme tekniği-model kombinasyonuna aittir. Ancak sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiğine ek olarak, kesinlik (precision) ve MCC performans ölçüm istatistiklerine bakıldığında da en iyi performans yine ROS-GBDT kombinasyonunda ölçülmüştür. Duyarlılık (recall) ve F1-skoru açısından en yüksek tahmin performansı,

ADASYN-hard oylama kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 7’de kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren yukarı örnekleme tekniği-tekil model kombinasyonu, ROS-GBDT’dir; yani değişken seçim tekniği uygulanmayan ve ROS yukarı örnekleme tekniğiyle dengelenen Alman kredi veri setinde yer alan riskli müşterileri, senaryo 7’de en doğru tahmin eden topluluk öğrenmesi modeli gradyan artırma karar ağaçları (GBDT) dır.

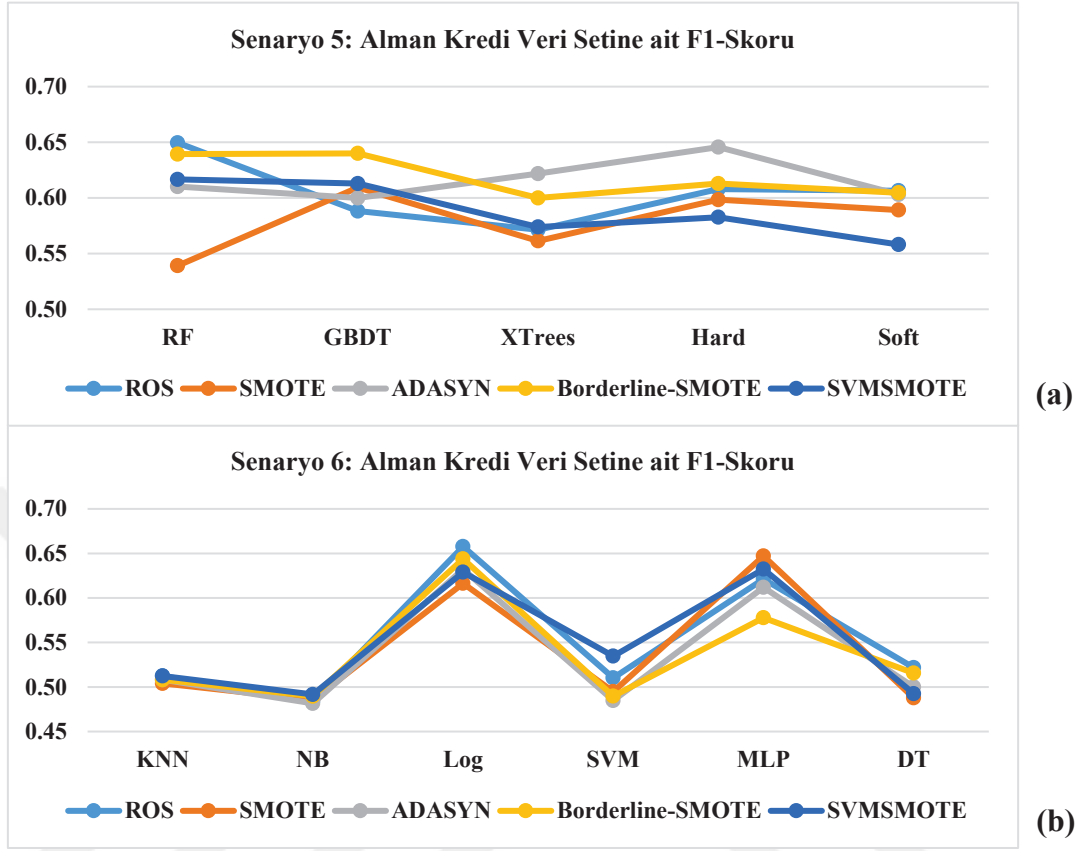
Tablo 18: Senaryo 8: Alman Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	KNN	0.6900	0.4516	0.5000	0.4746	0.2562
	NB	0.4150	0.9355	0.3392	0.4979	0.1532
	LR	0.7350	0.7581	0.5529	0.6395	0.4516
	SVM	0.7250	0.1774	0.7333	0.2857	0.2606
	MLP	0.7050	0.5000	0.5254	0.5124	0.3013
	DT	0.7000	0.5645	0.5147	0.5385	0.3177
ADASYN	KNN	0.6850	0.4516	0.4912	0.4706	0.2474
	NB	0.4100	0.8871	0.3313	0.4825	0.1019
	LR	0.7250	0.7903	0.5385	0.6405	0.4514
	SVM	0.7150	0.2097	0.6190	0.3133	0.2289
	MLP	0.7100	0.5968	0.5286	0.5606	0.3468
	DT	0.7100	0.5323	0.5323	0.5323	0.3221
SMOTE	KNN	0.6900	0.4516	0.5000	0.4746	0.2562
	NB	0.4400	0.8871	0.3438	0.4955	0.1459
	LR	0.7150	0.7742	0.5275	0.6275	0.4296
	SVM	0.7250	0.2097	0.6842	0.3210	0.2621
	MLP	0.6700	0.4194	0.4643	0.4407	0.2080
	DT	0.7150	0.6129	0.5352	0.5714	0.3613
Borderline-SMOTE	KNN	0.6900	0.4516	0.5000	0.4746	0.2562
	NB	0.4300	0.9194	0.3434	0.5000	0.1594
	LR	0.7300	0.7581	0.5465	0.6351	0.4442
	SVM	0.7200	0.1935	0.6667	0.3000	0.2425
	MLP	0.6850	0.4516	0.4912	0.4706	0.2474
	DT	0.6600	0.5645	0.4605	0.5072	0.2548
SVM-SMOTE	KNN	0.6900	0.4516	0.5000	0.4746	0.2562
	NB	0.4550	0.8871	0.3503	0.5023	0.1666
	LR	0.7350	0.7903	0.5506	0.6490	0.4658
	SVM	0.7200	0.2097	0.6500	0.3171	0.2450
	MLP	0.7050	0.5161	0.5246	0.5203	0.3074
	DT	0.7050	0.5645	0.5224	0.5426	0.3259

Tablo 18’de farklı yukarı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve yukarı örnekleme tekniklerinin uygulandığı Alman kredi

veri seti aracılığıyla, sınıflandırma başarısı test edilen tekil modellerin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %73,50 ile hem SVMSMOTE-LR hem de ROS-LR örnekleme tekniği-model kombinasyonlarına aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-tekil model kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiğine ek olarak, MCC ve F1-skoru performans ölçüm istatistiklerine bakıldığında da en iyi performans, SVMSMOTE-LR kombinasyonuna aittir. Kesinlik (precision) açısından en iyi performans, ROS-SVM kombinasyonunda ölçülürken, duyarlılık (recall) açısından en iyi performans ise ROS-NB kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda; senaryo 8’de, kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren yukarı örnekleme tekniği-tekil model kombinasyonu, SVMSMOTE-LR’ dir; yani değişken seçim tekniği uygulanmayan ve SVMSMOTE yukarı örnekleme tekniğiyle dengelenen Alman kredi veri setinde yer alan riskli müşterileri, senaryo 8’de en doğru tahmin eden tekil model lojistik regresyon (LR) dur.

Şekil 44: Senaryo 5 ve 6'ya ait F1-Skoru Grafiği: Alman Kredi Veri Seti



Yukarıdaki çizgi grafikleri, yukarı örnekleme teknikleriyle dengelenen Alman kredi veri seti üzerinde sınıflandırma performansı ölçülen tekil ve topluluk modellerine ait F1-skoru değerlerinin dağılımını gösterilmektedir. Bu grafiklere bakılarak model ve yukarı örnekleme teknikleri için elde edilen çıkarımlar ise şöyledir;

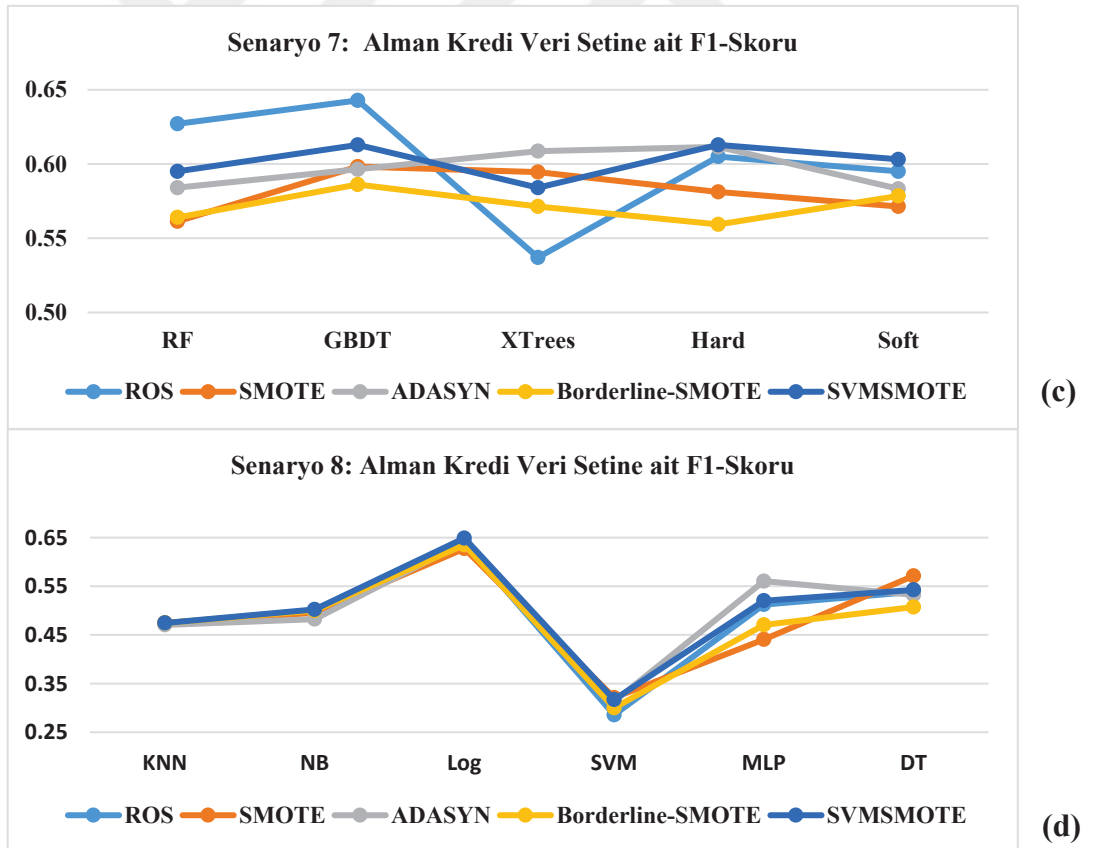
(a) grafiğine göre; kullanılan tüm topluluk öğrenmesi modellerinin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model rastgele ormanlar (RF) modelidir. Aynı zamanda, tüm topluluk modellerinin arasından ortalama olarak en düşük tahmin performansın ölçüldüğü model ekstra ağaçlar (XTrees) modelidir. Topluluk modellerinin nerdeyse tamamı, en yüksek tahmin performansını Borderline-SMOTE yukarı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir.

(b) grafiğine göre; kullanılan tüm tekil modellerin arasından ortalama olarak en düşük tahmin performansının ölçüldüğü yöntem Naive Bayes (NB)'tir. Aynı zamanda Naive Bayes, kullanılan tüm yukarı örnekleme teknikleriyle dengelenen veri

setindeki riskli müşterileri sınıflandırmada en düşük tahmin performansını göstermiştir. Naive bayes dışında, senaryo 2’de tahmin performansı düşük olan diğer model ise destek vektör makineleri (SVM)’dir. Bunun yanı sıra, kullanılan tüm tekil modellerin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model lojistik regresyon (LR) modelidir. Buna ek olarak, tekil modellerin tamamı en yüksek tahmin performansını, ROS yukarı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir.

(a) ve (b) grafiklerine göre; kullanılan tüm sınıflandırma modelleri arasından ortalama olarak en düşük tahmin performansının ölçüldüğü yöntem Naive Bayes (NB)’tir. Bunun yanı sıra senaryo 5 ve 6’da kullanılan tekil ve topluluk öğrenmesi modellerinin performanslarını kıyaslırsak eğer, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu söylenebilir.

Şekil 45: Senaryo 7 ve 8’e ait F1-Skoru Grafiği: Alman Kredi Veri Seti



(c) grafiğine göre; kullanılan tüm topluluk modellerinin arasından en yüksek tahmin performansının ölçüldüğü model GBDT modelidir. Aynı zamanda, kullanılan

tüm topluluk modellerinin arasından ortalama olarak en düşük tahmin performansının ölçüldüğü model ekstra ağaçlar (XTrees) modelidir. Bunun yanı sıra, çalışmada kullanılan topluluk öğrenmesi tekniklerinin nerdeyse tamamı ortalama olarak en yüksek tahmin performansını Borderline-SMOTE örnekleme tekniğiyle dengelenen veri setinde göstermiştir.

(d) grafiğine göre; kullanılan tüm tekil modellerin arsından ortalama olarak en düşük tahmin performansının ölçüldüğü yöntem Naive Bayes (NB)'tir. Bunun yanı sıra, kullanılan tüm tekil modellerin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model lojistik regresyon (LR) modelidir. Buna ek olarak, çalışmada kullanılan tekil modellerin nerdeyse tamamı, en yüksek tahmin performansını SVM-SMOTE örnekleme tekniğiyle dengelenen veri setinde göstermiştir.

(c) ve (d) grafiklerini kıyaslırsak eğer; incelenen dört senaryoda kullanılan tüm sınıflandırma modelleri arasından genel olarak en düşük tahmin performansının ölçüldüğü model Naive Bayes (NB)'tir. Bunun yanı sıra senaryo 7 ve 8'de kullanılan tekil ve topluluk öğrenmesi modellerinin performanslarını kıyaslırsak eğer, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu söylenebilir.

(a) ve (c) grafiklerini kıyaslırsak eğer; Senaryo 6'nın, Alman kredi veri setinde bulunan riskli müşterileri sınıflandırmada kullanılacak olan tüm tekil modellerin tahmin performansını arttırdığını söylemek mümkündür. Yani, senaryo 6, çalışmada kullanılan tüm tekil modellerin tahmin performansını arttırmada, senaryo 8'e göre daha başarılı olmuştur.

(b) ve (d) grafiklerini kıyaslırsak eğer; Senaryo 5'in, Alman kredi veri setinde bulunan riskli müşterileri sınıflandırmak için kullanılacak olan topluluk öğrenmesi modellerinin tahmin performansını arttırdığını söylemek mümkündür. Yani, senaryo 5, çalışmada kullanılan tüm topluluk öğrenmesi modellerinin tahmin performansını artırma konusunda senaryo 7'ye göre daha başarılı olmuştur.

Çalışmada, Alman kredi veri setini sınıflandırmak için kullanılan sekiz farklı senaryodan genel olarak elde edilen çıkarımlar ise şöyledir;

- SMOTE tabanlı yukarı örnekleme teknikleri, çalışmada kullanılan sınıflandırma tekniklerinin tahmin performansını artırma konusunda ADASYN yukarı örnekleme tekniğine göre daha başarılı olmuştur.

- Aşağı örnekleme tekniklerinin yer aldığı senaryoların tümünde, kullanılan sınıflandırma modellerinin tamamı ortalama olarak en yüksek tahmin performansını rastgele aşağı örnekleme (RUS) tekniğiyle dengelenen Alman kredi veri setinde elde etmiştir.
- Alman kredi veri setine değişken seçim tekniğinin uygulanması, senaryo 6 hariç, değişken seçiminin uygulandığı diğer tüm senaryolarda yer alan makine öğrenmesi tekniklerinin tahmin performansına önemli bir etkisinin olmadığı görülmüştür.

3.4.2. Avustralya Kredi Veri Setine ait Performans İstatistikleri

Tablo 19: Senaryo 1: Avustralya Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	RF	0.8768	0.8868	0.8103	0.8468	0.7463
	GBDT	0.8696	0.8868	0.7966	0.8393	0.7330
	XTrees	0.8696	0.9245	0.7778	0.8448	0.7419
	Hard	0.8623	0.9057	0.7742	0.8348	0.7245
	Soft	0.8696	0.8868	0.7966	0.8393	0.7330
Near Miss	RF	0.8406	0.8113	0.7818	0.7963	0.6657
	GBDT	0.6594	0.1698	0.7500	0.2769	0.2322
	XTrees	0.8623	0.9057	0.7742	0.8348	0.7245
	Hard	0.7971	0.7925	0.7119	0.7500	0.5825
	Soft	0.8406	0.8302	0.7719	0.8000	0.6690
Cluster Centroids	RF	0.8560	0.7708	0.8409	0.8043	0.6924
	GBDT	0.8400	0.7500	0.8182	0.7826	0.6580
	XTrees	0.8400	0.7708	0.8043	0.7872	0.6595
	Hard	0.8400	0.7708	0.8043	0.7872	0.6595
	Soft	0.8560	0.7500	0.8571	0.8000	0.6920
Tomek-Links	RF	0.8478	0.8868	0.7581	0.8174	0.6945
	GBDT	0.8623	0.9245	0.7656	0.8376	0.7296
	XTrees	0.8478	0.8679	0.7667	0.8142	0.6899
	Hard	0.8696	0.9057	0.7869	0.8421	0.7372
	Soft	0.8696	0.9057	0.7869	0.8421	0.7372

Tablo 19’da farklı aşağı örnekleme teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait bazı performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve aşağı örnekleme teknikleriyle dengelenen Avustralya kredi veri seti aracılığıyla sınıflandırma başarısı test edilen topluluk modellerinin tahmin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %87,68 ile RUS-RF aşağı örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk

değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu belirlerken, kullanılan diğer tüm ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda, doğruluk ölçüm istatistiğine ek olarak, F1-Skoru ve MCC performans ölçüm istatistiğine bakıldığında da en iyi performans yine RUS-RF kombinasyonunda ölçülmüştür. Duyarlılık (recall) açısından en yüksek tahmin performansı, RUS-XTrees kombinasyonunda ölçülürken, kesinlik (precision) açısından ise en iyi performans, cluster centroids-soft oylama kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 1’de kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren aşağı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, RUS-RF’ dir; yani değişken seçim tekniği uygulanan ve RUS aşağı örnekleme tekniğiyle dengelenen Avustralya kredi veri setinde bulunan riskli müşterileri, senaryo 1’de en doğru tahmin eden topluluk öğrenmesi tekniği rastgele ormanlar (RF)’dır.

Tablo 20: Senaryo 2: Avustralya Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	KNN	0.8768	0.9057	0.8000	0.8496	0.7501
	NB	0.7609	0.4151	0.9167	0.5714	0.5025
	LR	0.8696	0.9245	0.7778	0.8448	0.7419
	SVM	0.8696	0.9245	0.7778	0.8448	0.7419
	MLP	0.8768	0.9245	0.7903	0.8522	0.7544
	DT	0.8478	0.8491	0.7759	0.8108	0.6859
Near Miss	KNN	0.8623	0.8868	0.7833	0.8319	0.7200
	NB	0.6957	0.2264	0.9231	0.3636	0.3574
	LR	0.8623	0.9245	0.7656	0.8376	0.7296
	SVM	0.8696	0.9434	0.7692	0.8475	0.7473
	MLP	0.8551	0.8868	0.7705	0.8246	0.7072
	DT	0.8188	0.7925	0.7500	0.7706	0.6218
Cluster Centroids	KNN	0.8240	0.6458	0.8611	0.7381	0.6239
	NB	0.7600	0.3750	1.0000	0.5455	0.5195
	LR	0.8400	0.7917	0.7917	0.7917	0.6618
	SVM	0.8720	0.8750	0.8077	0.8400	0.7352
	MLP	0.8320	0.7292	0.8140	0.7692	0.6402
	DT	0.8800	0.8750	0.8235	0.8485	0.7502
Tomek-Links	KNN	0.8551	0.8302	0.8000	0.8148	0.6962
	NB	0.7754	0.4528	0.9231	0.6076	0.5340
	LR	0.8841	0.9434	0.7937	0.8621	0.7718
	SVM	0.8841	0.9434	0.7937	0.8621	0.7718
	MLP	0.8188	0.8113	0.7414	0.7748	0.6255
	DT	0.8768	0.9057	0.8000	0.8496	0.7501

Tablo 20’de farklı aşağı örnekleme teknikleri ve tekil model kombinasyonlarına ait bazı performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve aşağı örnekleme teknikleriyle dengelenen Avustralya kredi veri seti aracılığıyla sınıflandırma başarısı test edilen tekil modellerin tahmin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %88,41 ile hem Tomek links-LR hem de Tomek links-SVM aşağı örnekleme tekniği-tekil model kombinasyonlarına aittir. Ancak sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği ve tekil model kombinasyonlarını belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda, doğruluk ölçüm istatistiğine ek olarak F1-skor, duyarlılık (recall) ve MCC performans ölçüm istatistiklerine bakıldığında en iyi performans yine hem Tomek links-LR hem de Tomek links-SVM kombinasyonlarında ölçülmüştür. Duyarlılık (recall) açısından en yüksek tahmin performansının ölçüldüğü diğer kombinasyon ise near miss-SVM kombinasyonudur. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önüne alındığında, Senaryo 2’de kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren aşağı örnekleme tekniği-tekil model kombinasyonu hem Tomek links-LR hem de Tomek links-SVM’ ye aittir; yani değişken seçim tekniği uygulanan ve Tomek links aşağı örnekleme tekniğiyle dengelenen Avustralya kredi veri setinde yer alan riskli müşterileri, senaryo 2’de en doğru tahmin eden tekil modeller hem destek vektör makineleri hem de lojistik regresyondur.

Tablo 21: Senaryo 3: Avustralya Kredi Veri Seti Performans İstatistikleri

Aşağı Örneklemeye	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	RF	0.8960	0.9216	0.8393	0.8785	0.7906
	GBDT	0.8400	0.8627	0.7719	0.8148	0.6780
	XTrees	0.8480	0.8235	0.8077	0.8155	0.6864
	Hard	0.8720	0.8824	0.8182	0.8491	0.7398
	Soft	0.8800	0.9020	0.8214	0.8598	0.7578
Near Miss	RF	0.8560	0.7917	0.8261	0.8085	0.6936
	GBDT	0.8400	0.7917	0.7917	0.7917	0.6618
	XTrees	0.8240	0.7500	0.7826	0.7660	0.6254
	Hard	0.8400	0.7500	0.8182	0.7826	0.6580
	Soft	0.8320	0.7500	0.8000	0.7742	0.6415
Cluster Centroids	RF	0.8560	0.7708	0.8409	0.8043	0.6924
	GBDT	0.8000	0.7708	0.7255	0.7475	0.5829
	XTrees	0.8480	0.7917	0.8085	0.8000	0.6775
	Hard	0.8560	0.7708	0.8409	0.8043	0.6924
	Soft	0.8640	0.7708	0.8605	0.8132	0.7094
Tomek-Links	RF	0.8720	0.8333	0.8333	0.8333	0.7294
	GBDT	0.8400	0.8125	0.7800	0.7959	0.6648
	XTrees	0.8480	0.8125	0.7959	0.8041	0.6801
	Hard	0.8560	0.7708	0.8409	0.8043	0.6924
	Soft	0.8320	0.7292	0.8140	0.7692	0.6402

Tablo 21’de farklı aşağı örneklemeye teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve aşağı örneklemeye teknikleriyle dengelenen Avustralya kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %89,60 ile RUS-RF aşağı örneklemeye tekniği- topluluk öğrenmesi model kombinasyonuna aittir. Ancak, sadece doğruluk değerini göz önünde bulundurularak, performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örneklemeye tekniği-topluluk öğrenmesi modeli kombinasyonu belirlerken, kullanılan diğer performans ölçüm istatistiklerine de bakmakta fayda vardır. Doğruluk ölçüm istatistiğine ek olarak, duyarlılık (recall), F1-skor ve MCC açısından da en iyi performans RUS-RF kombinasyonunda ölçülmüştür. Kesinlik (precision) açısından en iyi performans, cluster centroids-soft oylama kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önüne alındığında, Senaryo 3 için, kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren

aşağı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, RUS-RF' ye aittir; yani değişken seçim tekniği uygulanmayan ve RUS aşağı örnekleme tekniğiyle dengelenen Avustralya kredi veri setinde yer alan riskli müşterileri, senaryo 3'te en doğru tahmin eden topluluk öğrenmesi modeli rastgele ormanlar (RF)'dir.

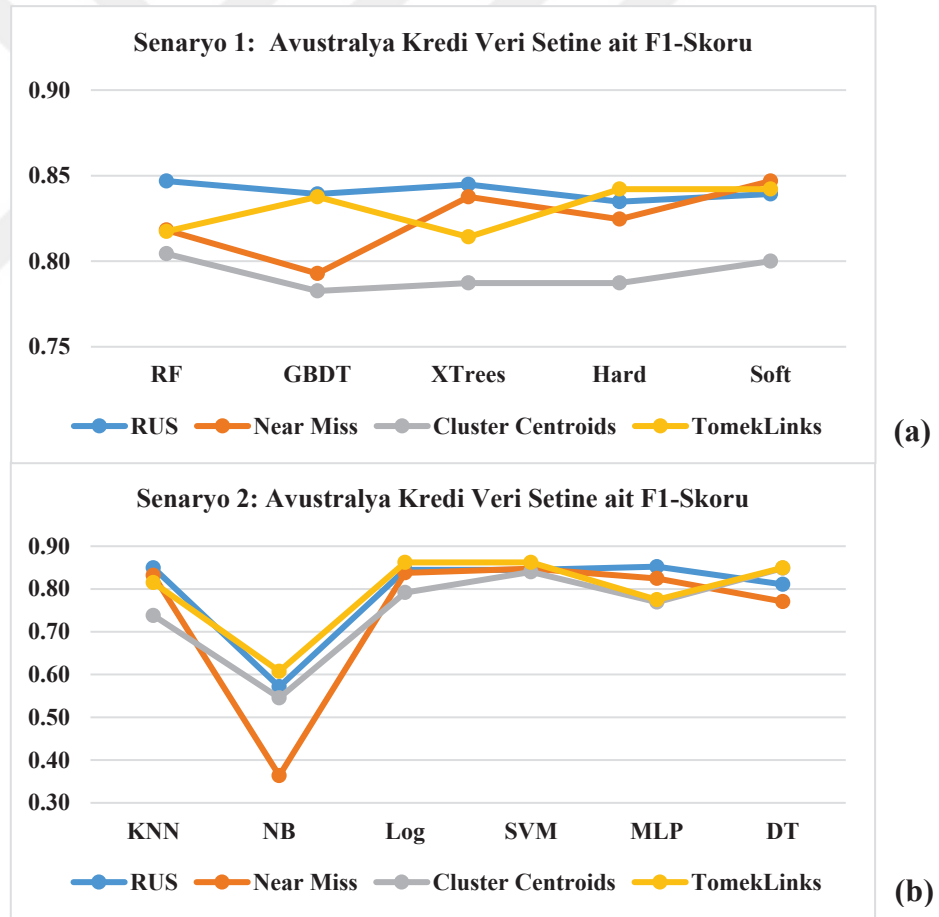
Tablo 22: Senaryo 4: Avustralya Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	KNN	0.8480	0.8039	0.8200	0.8119	0.6845
	NB	0.8000	0.5686	0.9063	0.6988	0.5947
	LR	0.8880	0.9412	0.8136	0.8727	0.7802
	SVM	0.8960	0.9608	0.8167	0.8829	0.7989
	MLP	0.8960	0.9020	0.8519	0.8762	0.7876
	DT	0.8800	0.8627	0.8462	0.8544	0.7524
Near Miss	KNN	0.8480	0.7292	0.8537	0.7865	0.6747
	NB	0.6480	0.1458	0.7000	0.2414	0.1916
	LR	0.8800	0.9375	0.7895	0.8571	0.7633
	SVM	0.7840	0.7292	0.7143	0.7216	0.5453
	MLP	0.8160	0.7708	0.7551	0.7629	0.6127
	DT	0.8240	0.8125	0.7500	0.7800	0.6351
Cluster Centroids	KNN	0.8560	0.7917	0.8261	0.8085	0.6936
	NB	0.7040	0.2500	0.9231	0.3934	0.3776
	LR	0.8880	0.9167	0.8148	0.8627	0.7725
	SVM	0.8080	0.7083	0.7727	0.7391	0.5891
	MLP	0.8400	0.7917	0.7917	0.7917	0.6618
	DT	0.8240	0.7500	0.7826	0.7660	0.6254
Tomek-Links	KNN	0.8640	0.7500	0.8780	0.8090	0.7097
	NB	0.7360	0.3333	0.9412	0.4923	0.4545
	LR	0.8400	0.7917	0.7917	0.7917	0.6618
	SVM	0.8240	0.7500	0.7826	0.7660	0.6254
	MLP	0.8400	0.7917	0.7917	0.7917	0.6618
	DT	0.8480	0.8333	0.7843	0.8081	0.6833

Tablo 22'de farklı aşağı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve aşağı örnekleme tekniklerinin uygulandığı Avustralya kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen tekil modellerin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %89,60 ile hem RUS-SVM hem de RUS-MLP kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği ve tekil model kombinasyonu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda

vardır. Bu bağlamda doğruluk ölçüm istatistiğine ek olarak, duyarlılık (recall), F1-skor ve MCC performans ölçüm istatistiklerine göre en iyi performans RUS-SVM kombinasyonuna aittir. Kesinlik (precision) açısından en yüksek tahmin performansı, RUS-NB kombinasyonuna aittir. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önüne alındığında, Senaryo 4'te kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren aşağı örnekleme tekniği-tekil model kombinasyonu, RUS-SVM' ye aittir; yani değişken seçim tekniği uygulanmayan ve RUS aşağı örnekleme tekniğiyle dengelenen Avustralya kredi veri setinde yer alan riskli müşterileri, senaryo 4'te en doğru tahmin eden tekil model destek vektör makineleridir.

Şekil 46: Senaryo 1 ve 2'ye ait F1-Skoru Grafiği: Avustralya Kredi Veri Seti



Yukarıdaki çizgi grafikleri, aşağı örnekleme teknikleriyle dengelenen Avustralya kredi veri seti üzerinde sınıflandırma performansı ölçülen tekil ve topluluk

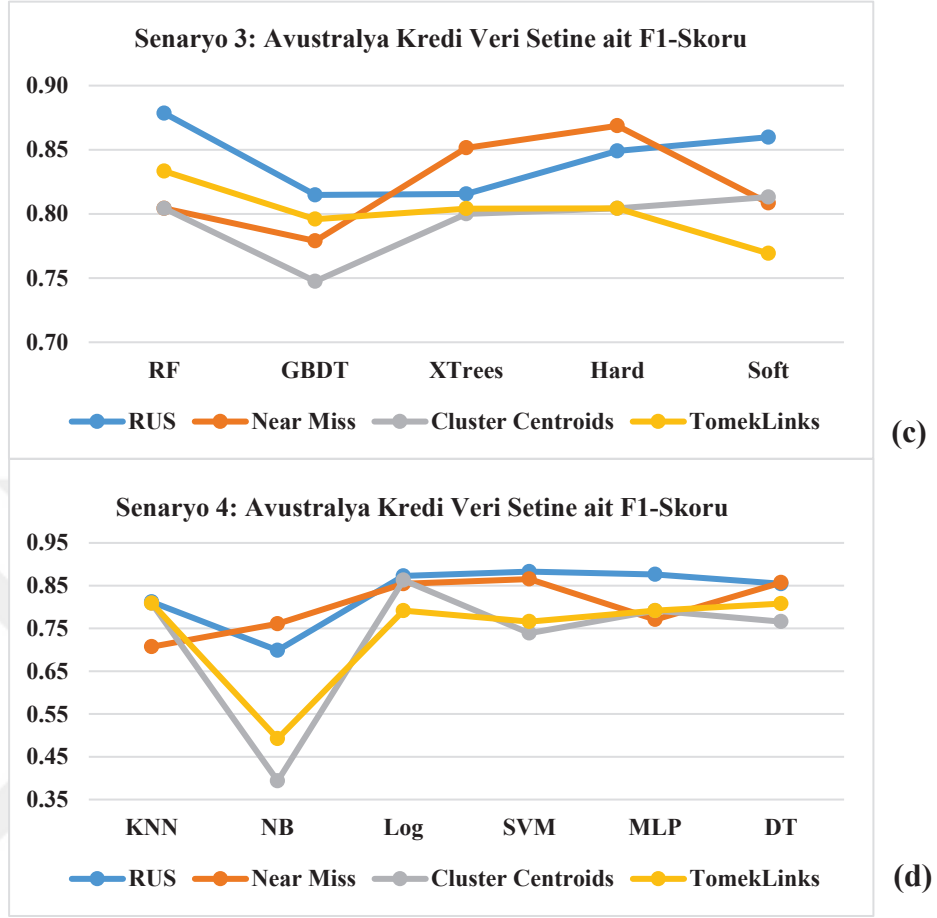
modellerine ait F1-skoru deęerlerinin daęılımını gsterilmektedir. Bu grafiklere bakılarak model ve aŗaęı rnekleme teknikleri iin elde edilen ıkarımlar ise řyledir;

(a) grafięine gre; kullanılan tm topluluk ęrenmesi modellerinin arasından ortalama olarak en yksek tahmin performansının lldę model soft oylamadır. Aynı zamanda, kullanılan tm topluluk modellerinin arasından ortalama olarak en dřk tahmin performansının lldę model ekstra aęalar (XTrees) modelidir. Topluluk modellerinin nerdeyse tamamı, en yksek tahmin performansını RUS aŗaęı rnekleme teknięiyle dengelenen veri setini sınıflandırırken gstermiŗtir

(b) grafięine gre; kullanılan tm tekil modellerinin arasından ortalama olarak en dřk tahmin performansının lldę yntem Naive Bayes (NB)'tir. Aynı zamanda Naive Bayes, tm aŗaęı rnekleme teknikleriyle dengelenen Avustralya kredi veri setinde en dřk tahmin performansını gsteren yntem olmuŗtur. Bunun yanı sıra, kullanılan tm tekil modellerinin arasından ortalama olarak en yksek tahmin performansının lldę model destek vektr makineleri (SVM) dir. Buna ek olarak, tekil modellerin tamamı ortalama olarak en yksek tahmin performansını, RUS aŗaęı rnekleme teknięiyle dengelenen veri setinde bulunan riskli mŗterileri sınıflandırırken gstermiŗtir.

(a) ve (b) grafiklerini kıyaslırsak eęer; kullanılan tm sınıflandırma modellerinin arasından ortalama olarak en dřk tahmin performansının lldę yntem Naive Bayes (NB)'tir. Bunun yanı sıra senaryo 1 ve 2'de kullanılan tekil ve topluluk ęrenmesi modellerinin performanslarını kıyaslırsak eęer, topluluk ęrenmesi modellerinin tahmin performansının daha yksek olduęu sylenebilir.

Şekil 47: Senaryo 3 ve 4'e ait F1 Skoru Grafiği: Avustralya Kredi Veri Seti



(c) grafiğine göre; kullanılan tüm tekil modellerin arasında ortalama olarak en düşük tahmin performansına sahip yöntem Naive Bayes (NB)'tir. Bunun yanı sıra, kullanılan tüm tekil modellerin arasında ortalama olarak en yüksek tahmin performansının ölçüldüğü model lojistik regresyon (LR) modelidir. Ek olarak, çalışmada kullanılan tekil modellerin neredeyse tamamı, ortalama olarak en yüksek tahmin performansını RUS aşağı örnekleme tekniğiyle dengelenen veri setinde elde etmiştir.

(d) grafiğine göre; kullanılan tüm topluluk modellerinin arasında en yüksek tahmin performansının ölçüldüğü model hard oylama modelidir. Aynı zamanda, kullanılan tüm topluluk modellerinin arasında ortalama olarak en düşük tahmin performansının ölçüldüğü model GBDT modelidir. Bunun yanı sıra, çalışmada kullanılan topluluk öğrenmesi tekniklerinin neredeyse tamamı en yüksek tahmin performansını RUS örnekleme tekniğiyle dengelenen veri setinde elde etmiştir.

(c) ve (d) grafiklerini kıyaslarsak eğer; kullanılan tüm sınıflandırma modellerinin arasından genel olarak en düşük tahmin performansının ölçüldüğü yöntem Naive Bayes (NB)'tir. Bunun yanı sıra senaryo 3 ve 4'te kullanılan tekil ve topluluk öğrenmesi modellerinin performanslarını kıyaslarsak eğer, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu söylenebilir.

(a) ve (c) grafiklerini kıyaslarsak eğer; senaryo 2'nin, Avustralya kredi veri setinde bulunan riskli müşterileri sınıflandırmada kullanılacak olan tüm tekil modellerin tahmin performansını arttırdığını söylemek mümkündür. Yani, senaryo 2, çalışmada kullanılan tüm tekil modellerin tahmin performansını artırma konusunda senaryo 4'e göre daha başarılı olmuştur.

(b) ve (d) grafiklerini kıyaslarsak eğer; senaryo 1'in, Avustralya kredi veri setinde bulunan riskli müşterileri sınıflandırmak için kullanılacak olan topluluk öğrenmesi modellerinin tahmin performansını arttırdığını söylemek mümkündür. Yani, senaryo 1, çalışmada kullanılan tüm topluluk öğrenmesi modellerinin tahmin performansını artırma konusunda senaryo 3'e göre daha başarılı olmuştur.

Tablo 23: Senaryo 5: Avustralya Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	RF	0.8640	0.8627	0.8148	0.8381	0.7219
	GBDT	0.8960	0.9020	0.8519	0.8762	0.7876
	XTrees	0.8160	0.7647	0.7800	0.7723	0.6180
	Hard	0.8720	0.8627	0.8302	0.8462	0.7370
	Soft	0.8800	0.8824	0.8333	0.8571	0.7548
ADASYN	RF	0.8640	0.7917	0.8444	0.8172	0.7100
	GBDT	0.8400	0.7708	0.8043	0.7872	0.6595
	XTrees	0.8240	0.7708	0.7708	0.7708	0.6280
	Hard	0.8480	0.7708	0.8222	0.7957	0.6758
	Soft	0.8400	0.7500	0.8182	0.7826	0.6580
SMOTE	RF	0.8560	0.7708	0.8409	0.8043	0.6924
	GBDT	0.8480	0.7917	0.8085	0.8000	0.6775
	XTrees	0.8320	0.7708	0.7872	0.7789	0.6436
	Hard	0.8400	0.7708	0.8043	0.7872	0.6595
	Soft	0.8400	0.7500	0.8182	0.7826	0.6580
Borderline-SMOTE	RF	0.8560	0.7708	0.8409	0.8043	0.6924
	GBDT	0.8560	0.8542	0.7885	0.8200	0.7019
	XTrees	0.8320	0.7500	0.8000	0.7742	0.6415
	Hard	0.8320	0.7500	0.8000	0.7742	0.6415
	Soft	0.8400	0.7500	0.8182	0.7826	0.6580
SVMS-MOTE	RF	0.8560	0.7708	0.8409	0.8043	0.6924
	GBDT	0.8320	0.7917	0.7755	0.7835	0.6464
	XTrees	0.8320	0.7500	0.8000	0.7742	0.6415
	Hard	0.8320	0.7500	0.8000	0.7742	0.6415
	Soft	0.8400	0.7500	0.8182	0.7826	0.6580

Tablo 23'te farklı yukarı örnekleme teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve yukarı örnekleme tekniklerinin uygulandığı Avustralya kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %89,60 ile ROS-GBDT örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiğine ek olarak, kesinlik (precision) ve F1-skoru, MCC ve duyarlılık (recall) performans ölçüm istatistikleri açısından en iyi performans yine, ROS-GBDT kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz

önünde bulundurulduğunda, senaryo 5’te kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren yukarı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, ROS-GBDT’ ye aittir; yani değişken seçim tekniği uygulanan ve ROS yukarı örnekleme tekniğiyle dengelenen Avustralya kredi veri setinde yer alan riskli müşterileri, senaryo 5’te en doğru tahmin eden topluluk öğrenmesi modeli gradyan artırma karar ağaçlarıdır.

Tablo 24: Senaryo 6: Avustralya Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	KNN	0.8480	0.7843	0.8333	0.8081	0.6833
	NB	0.7760	0.4902	0.9259	0.6410	0.5532
	LR	0.8800	0.9020	0.8214	0.8598	0.7578
	SVM	0.9040	0.9608	0.8305	0.8909	0.8128
	MLP	0.8800	0.8824	0.8333	0.8571	0.7548
	DT	0.8800	0.8824	0.8333	0.8571	0.7548
ADASYN	KNN	0.8560	0.7500	0.8571	0.8000	0.6920
	NB	0.7520	0.3750	0.9474	0.5373	0.4904
	LR	0.8320	0.7708	0.7872	0.7789	0.6436
	SVM	0.8240	0.7500	0.7826	0.7660	0.6254
	MLP	0.8320	0.8125	0.7647	0.7879	0.6498
	DT	0.8880	0.8542	0.8542	0.8542	0.7633
SMOTE	KNN	0.8400	0.7083	0.8500	0.7727	0.6573
	NB	0.7440	0.3750	0.9000	0.5294	0.4630
	LR	0.8560	0.8125	0.8125	0.8125	0.6956
	SVM	0.8240	0.7500	0.7826	0.7660	0.6254
	MLP	0.8240	0.7292	0.7955	0.7609	0.6235
	DT	0.8560	0.8125	0.8125	0.8125	0.6956
Borderline-SMOTE	KNN	0.8320	0.7292	0.8140	0.7692	0.6402
	NB	0.7520	0.3750	0.9474	0.5373	0.4904
	LR	0.8400	0.7917	0.7917	0.7917	0.6618
	SVM	0.8240	0.7500	0.7826	0.7660	0.6254
	MLP	0.8240	0.7292	0.7955	0.7609	0.6235
	DT	0.8800	0.8542	0.8367	0.8454	0.7474
SVM-SMOTE	KNN	0.8400	0.7500	0.8182	0.7826	0.6580
	NB	0.7520	0.3750	0.9474	0.5373	0.4904
	LR	0.8400	0.7917	0.7917	0.7917	0.6618
	SVM	0.8240	0.7500	0.7826	0.7660	0.6254
	MLP	0.8000	0.8333	0.7018	0.7619	0.5982
	DT	0.8400	0.7708	0.8043	0.7872	0.6595

Tablo 24’te farklı yukarı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve yukarı örnekleme tekniklerinin uygulandığı Avustralya kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk modellerinin gücünü

gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %90,40 ile ROS-SVM örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-tekil model kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiğine ek olarak, kesinlik (precision) ve F1-skoru, MCC performans ölçüm istatistikleri açısından da en iyi performans yine ROS-SVM yukarı örnekleme tekniği-tekil model kombinasyonunda ölçülmüştür. Duyarlılık (recall) açısından en iyi performans ROS-NB kombinasyonunda elde edilmiştir. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önüne alındığında, senaryo 6'da kredi riski bulunan müşterileri doğru tahmin etmede yüksek performans gösteren yukarı örnekleme tekniği-tekil model kombinasyonu, ROS-SVM'dir; yani değişken seçim tekniği uygulanan ve ROS yukarı örnekleme tekniğiyle dengelenen Avustralya kredi veri setinde yer alan riskli müşterileri, senaryo 6'da en doğru tahmin eden tekil model destek vektör makineleri (SVM)'dir.

Tablo 25: Senaryo 7: Avustralya Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	RF	0.8640	0.8627	0.8148	0.8381	0.7219
	GBDT	0.8480	0.8431	0.7963	0.8190	0.6890
	XTrees	0.8560	0.8431	0.8113	0.8269	0.7041
	Hard	0.8640	0.8627	0.8148	0.8381	0.7219
	Soft	0.8640	0.8627	0.8148	0.8381	0.7219
ADASYN	RF	0.8640	0.8627	0.8148	0.8381	0.7219
	GBDT	0.8720	0.8824	0.8182	0.8491	0.7398
	XTrees	0.8560	0.8235	0.8235	0.8235	0.7019
	Hard	0.8560	0.8431	0.8113	0.8269	0.7041
	Soft	0.8640	0.8627	0.8148	0.8381	0.7219
SMOTE	RF	0.8640	0.8627	0.8148	0.8381	0.7219
	GBDT	0.8880	0.9020	0.8364	0.8679	0.7726
	XTrees	0.8640	0.8431	0.8269	0.8350	0.7194
	Hard	0.8720	0.8824	0.8182	0.8491	0.7398
	Soft	0.8720	0.8824	0.8182	0.8491	0.7398
Borderline-SMOTE	RF	0.8640	0.8627	0.8148	0.8381	0.7219
	GBDT	0.8560	0.8627	0.8000	0.8302	0.7070
	XTrees	0.8720	0.8627	0.8302	0.8462	0.7370
	Hard	0.8560	0.8431	0.8113	0.8269	0.7041
	Soft	0.8560	0.8431	0.8113	0.8269	0.7041
SVM-SMOTE	RF	0.8720	0.8824	0.8182	0.8491	0.7398
	GBDT	0.8640	0.8627	0.8148	0.8381	0.7219
	XTrees	0.8640	0.8627	0.8148	0.8381	0.7219
	Hard	0.8560	0.8627	0.8000	0.8302	0.7070
	Soft	0.8560	0.8431	0.8113	0.8269	0.7041

Tablo 25’te farklı yukarı örnekleme teknikleri ve topluluk öğrenmesi modelleri kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve yukarı örnekleme tekniklerinin uygulandığı Avustralya kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk öğrenmesi modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %88,80 ile SMOTE-GBDT örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiklerine ek olarak, kesinlik (precision), duyarlılık (recall), F1-skoru ve MCC performans ölçüm metriklerine bakıldığında da en iyi performans yine SMOTE-GBDT kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans

istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 7’de kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren yukarı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, SMOTE-GBDT’ dir; yani değişken seçim tekniği uygulanmayan ve SMOTE yukarı örnekleme tekniğiyle dengelenen Avustralya kredi veri setinde yer alan riskli müşterileri, senaryo 7’de en doğru tahmin eden topluluk öğrenmesi modeli gradyan artırma karar ağaçları (GBDT)’dir.

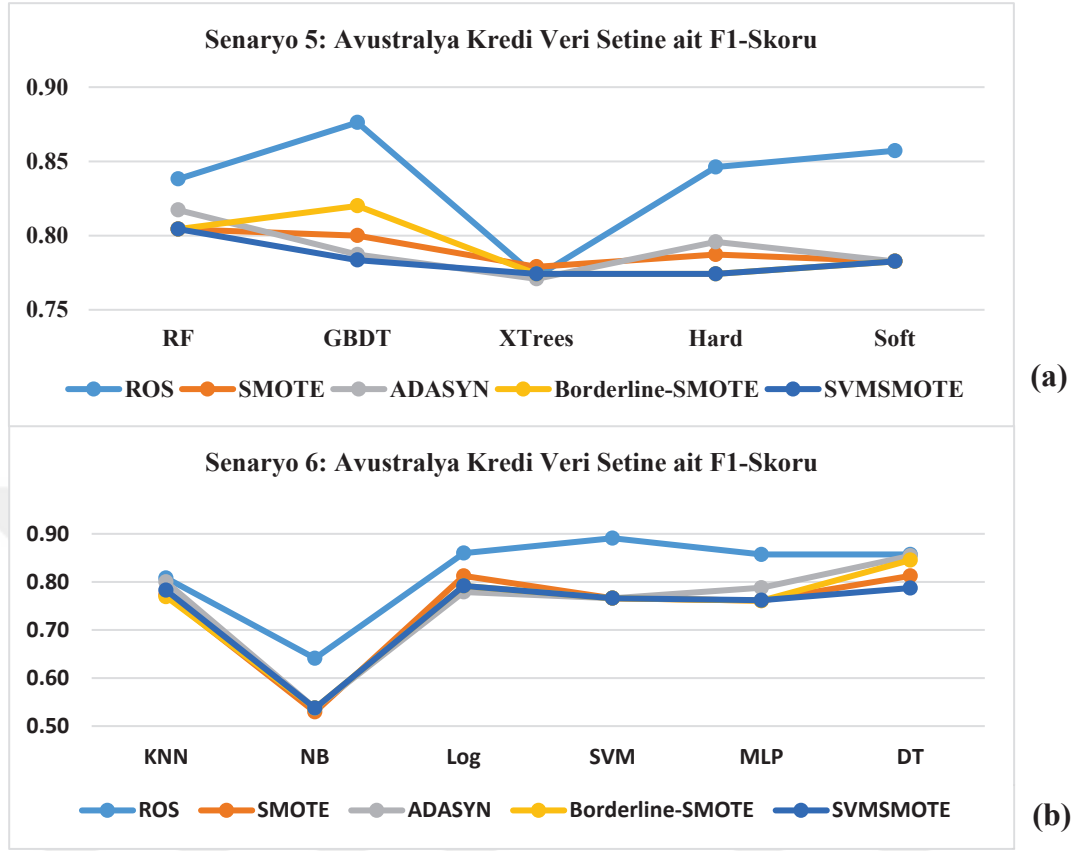
Tablo 26: Senaryo 8: Avustralya Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	KNN	0.8560	0.8431	0.8113	0.8269	0.7041
	NB	0.7920	0.5490	0.9032	0.6829	0.5787
	LR	0.8960	0.9412	0.8276	0.8807	0.7943
	SVM	0.8560	0.8627	0.8000	0.8302	0.7070
	MLP	0.8640	0.9020	0.7931	0.8440	0.7291
	DT	0.8800	0.9020	0.8214	0.8598	0.7578
ADASYN	KNN	0.8480	0.8039	0.8200	0.8119	0.6845
	NB	0.8000	0.5490	0.9333	0.6914	0.6007
	LR	0.8800	0.9216	0.8103	0.8624	0.7617
	SVM	0.8640	0.8431	0.8269	0.8350	0.7194
	MLP	0.8480	0.8627	0.7857	0.8224	0.6924
	DT	0.8880	0.9412	0.8136	0.8727	0.7802
SMOTE	KNN	0.8480	0.8235	0.8077	0.8155	0.6864
	NB	0.8080	0.5686	0.9355	0.7073	0.6164
	LR	0.8800	0.9216	0.8103	0.8624	0.7617
	SVM	0.8720	0.9020	0.8070	0.8519	0.7433
	MLP	0.8480	0.8824	0.7759	0.8257	0.6964
	DT	0.8800	0.9020	0.8214	0.8598	0.7578
Borderline-SMOTE	KNN	0.8560	0.8235	0.8235	0.8235	0.7019
	NB	0.7920	0.5294	0.9310	0.6750	0.5849
	LR	0.8880	0.9412	0.8136	0.8727	0.7802
	SVM	0.8480	0.8235	0.8077	0.8155	0.6864
	MLP	0.8640	0.8824	0.8036	0.8411	0.7251
	DT	0.8800	0.9020	0.8214	0.8598	0.7578
SVM-SMOTE	KNN	0.8560	0.8235	0.8235	0.8235	0.7019
	NB	0.7840	0.5294	0.9000	0.6667	0.5626
	LR	0.8720	0.9020	0.8070	0.8519	0.7433
	SVM	0.8640	0.8627	0.8148	0.8381	0.7219
	MLP	0.8560	0.9020	0.7797	0.8364	0.7150
	DT	0.8640	0.9020	0.7931	0.8440	0.7291

Tablo 26’da farklı yukarı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve yukarı örnekleme tekniklerinin uygulandığı Avustralya

kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen tekil modellerin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %89,90 ile ROS-LR örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-tekil model kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiğine ek olarak, MCC, duyarlılık (recall) ve F1-skoru performans ölçüm istatistiklerine bakıldığında da en iyi performans, ROS-LR kombinasyonunda ölçülmüştür. Kesinlik (precision) açısından en iyi performans, ROS-NB kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda, senaryo 8'de kredi riski bulunan müşterileri doğru tahmin etmede yüksek performans gösteren yukarı örnekleme tekniği-tekil model kombinasyonu, ROS-LR' dir; yani, değişken seçim tekniği uygulanmayan ve ROS yukarı örnekleme tekniğiyle dengelenen Avustralya kredi veri setinde yer alan riskli müşterileri, senaryo 8'de en doğru tahmin eden tekil model lojistik regresyon (LR)'dur.

Şekil 48: Senaryo 5 ve 6'ya ait F1 Skoru Grafiği: Avustralya Kredi Veri Seti



Yukarıdaki çizgi grafikleri, yukarı örnekleme teknikleriyle dengelenen Avustralya kredi veri seti üzerinde sınıflandırma performansı ölçülen tekil ve topluluk modellerine ait F1-skoru değerlerinin dağılımını gösterilmektedir. Bu grafiklere bakılarak model ve yukarı örnekleme teknikleri için elde edilen çıkarımlar ise şöyledir;

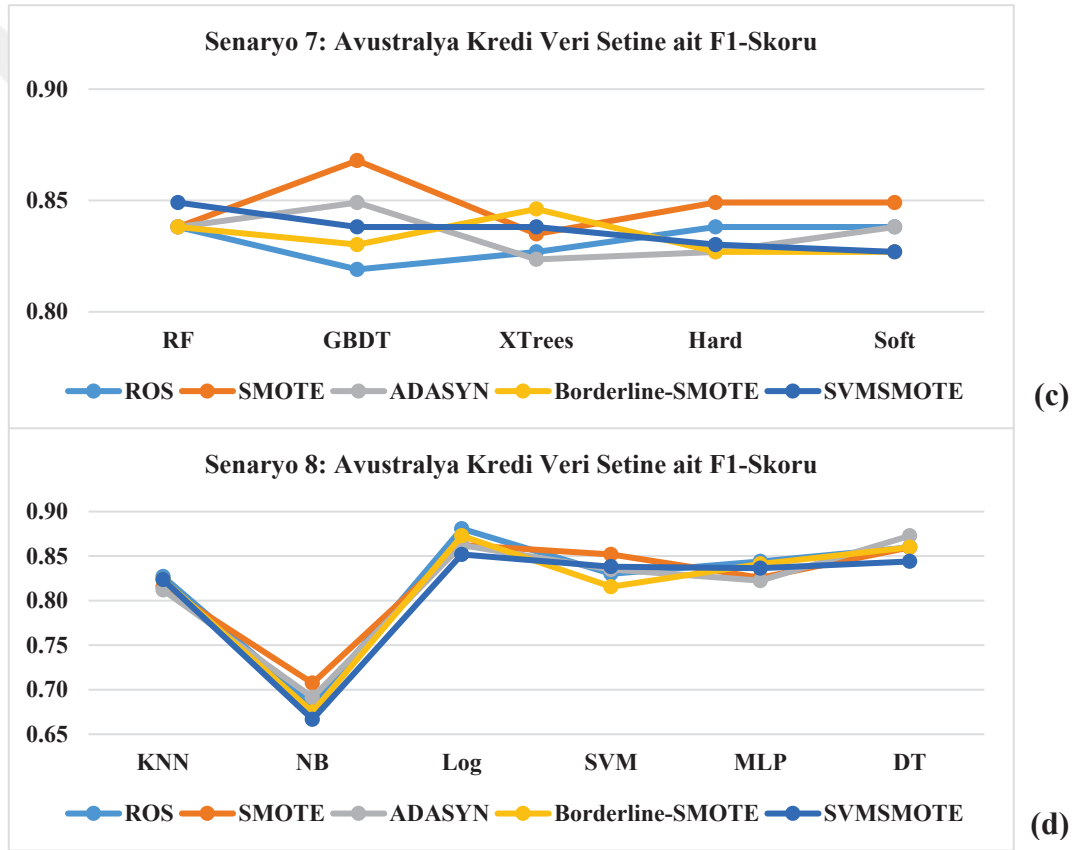
(a) grafiğine göre; kullanılan tüm topluluk öğrenmesi modellerinin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model rastgele ormanlar (RF) modelidir. Aynı zamanda, kullanılan tüm topluluk modelleri arasından ortalama olarak en düşük performansın ölçüldüğü model ise ekstra ağaçlar (XTrees) modelidir. Topluluk modellerinin nerdeyse tamamı, en yüksek tahmin performansını ROS yukarı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir.

(b) grafiğine göre; kullanılan tüm tekil modellerin arasından ortalama olarak en düşük tahmin performansının ölçüldüğü yöntem Naive Bayes (NB)'tir. Aynı zamanda Naive Bayes, çalışmada kullanılan tüm yukarı örnekleme teknikleriyle dengelenen Avustralya kredi veri setinde en düşük tahmin performansını elde etmiştir. Bunun yanı sıra, kullanılan tüm tekil modeller arasında ortalama olarak en yüksek

performans ise karar ağaçları (DT) modeline aittir. Buna ek olarak; tekil modellerin tamamı en yüksek tahmin performansını, ROS yukarı örnekleme tekniğiyle dengelenen veri setinde bulunan riskli müşterileri sınıflandırırken göstermiştir.

(a) ve (b) grafiklerini kıyaslırsak eğer; kullanılan tüm sınıflandırma modelleri arasında ortalama olarak en düşük tahmin performansına sahip yöntem Naive Bayes (NB)'tir. Bunun yanı sıra senaryo 5 ve 6'da kullanılan tekil ve topluluk öğrenmesi modellerinin performanslarını kıyaslırsak eğer, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu söylenebilir.

Şekil 49: Senaryo 7 ve 8'e ait F1 Skoru Grafiği: Avustralya Kredi Veri Seti



(c) grafiğine göre; kullanılan tüm topluluk modellerinin arasından en yüksek tahmin performansının ölçüldüğü model GBDT modelidir. Bunun yanı sıra, çalışmada kullanılan topluluk öğrenmesi tekniklerinin neredeyse tamamı en iyi tahmin performansını SMOTE örnekleme tekniğiyle dengelenen veri setinde elde etmiştir.

(d) grafiğine göre; kullanılan tüm tekil modellerin arasından ortalama olarak en düşük tahmin performansının ölçüldüğü yöntem Naive Bayes (NB)'e aittir. Bunun

yanı sıra, kullanılan tüm tekil modellerin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model lojistik regresyon (LR) modelidir. Buna ek olarak, çalışmada kullanılan tekil modellerin nerdeyse tamamı, en yüksek tahmin performansını ROS örnekleme tekniğiyle dengelenen veri setinde elde etmiştir.

(c) ve (d) grafiklerini kıyaslırsak eğer; kullanılan tüm sınıflandırma modellerinin arasından en düşük tahmin performansın ölçüldüğü sınıflandırma tekniği, Naive Bayes (NB)'tir. Bunun yanı sıra senaryo 7 ve 8'de kullanılan tekil ve topluluk öğrenmesi modellerinin performanslarını kıyaslırsak eğer, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu söylenebilir.

(a) ve (c) grafiklerini kıyaslırsak eğer; senaryo 8'in, Avustralya kredi veri setinde bulunan riskli müşterileri sınıflandırmada kullanılacak olan tüm tekil modellerin tahmin performansını arttırdığını söylemek mümkündür. Yani, senaryo 8, çalışmada kullanılan tüm tekil modellerin tahmin performansını arttırmada, senaryo 6'ya göre daha başarılı olmuştur.

(b) ve (d) grafiklerini kıyaslırsak eğer; senaryo 5'in, Avustralya kredi veri setinde bulunan riskli müşterileri sınıflandırmak için kullanılacak olan topluluk öğrenmesi modellerinin tahmin performansını arttırdığını söylemek mümkündür. Yani, senaryo 7, çalışmada kullanılan tüm topluluk öğrenmesi modellerinin tahmin performansını artırma konusunda senaryo 5'e göre daha başarılı olmuştur.

Çalışmada Avustralya kredi veri seti için kullanılan sekiz farklı senaryodan genel olarak elde edilen çıkarımlar ise şöyledir;

- Aşağı örnekleme tekniklerinin yer aldığı senaryoların tümünde, kullanılan sınıflandırma modellerinin tamamı ortalama olarak en yüksek tahmin performansını rastgele aşağı örnekleme (RUS) tekniğiyle dengelenen Avustralya kredi veri setinde elde etmiştir.
- Yukarı örnekleme tekniklerinin yer aldığı senaryoların tümünde, kullanılan sınıflandırma modellerinin tamamı ortalama olarak en yüksek tahmin performansını rastgele yukarı örnekleme (ROS) tekniğiyle dengelenen Avustralya kredi veri setinde elde etmiştir.
- Avustralya kredi veri setine değişken seçim tekniğinin uygulanması, senaryo 1 dışında, değişken seçiminin uygulandığı diğer tüm senaryolarda yer alan makine

öğrenmesi tekniklerinin tahmin performansını arttırmada önemli bir etkisinin olmadığı görülmüştür.

- Çalışmada kullanılan tüm makine öğrenmesi tekniklerinin arasından düşük tahmin performansının ölçüldüğü yöntem, Naive Bayes (NB) yöntemidir.

3.4.3. HMEQ Kredi Veri Setine ait Performans İstatistikleri

Tablo 27: Senaryo 1: HMEQ Kredi Veri Seti Performans İstatistikleri

Aşağı Örneklemeye	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	RF	0.8695	0.8333	0.6338	0.7200	0.6471
	GBDT	0.8667	0.8194	0.6299	0.7123	0.6367
	XTrees	0.9143	0.9028	0.7331	0.8091	0.7614
	Hard	0.8947	0.8241	0.7036	0.7591	0.6958
	Soft	0.9059	0.8009	0.7489	0.7740	0.7153
Near Miss	RF	0.8456	0.8145	0.5941	0.6871	0.6008
	GBDT	0.8221	0.8266	0.5481	0.6592	0.5665
	XTrees	0.8289	0.9274	0.5529	0.6928	0.6220
	Hard	0.8414	0.9194	0.5743	0.7070	0.6376
	Soft	0.8540	0.8750	0.6028	0.7138	0.6397
Cluster Centroids	RF	0.8733	0.8400	0.6540	0.7354	0.6626
	GBDT	0.8406	0.8533	0.5818	0.6919	0.6092
	XTrees	0.9543	0.8622	0.9151	0.8879	0.8598
	Hard	0.9189	0.8133	0.8026	0.8079	0.7566
	Soft	0.9329	0.8222	0.8525	0.8371	0.7951
Tomek-Links	RF	0.9236	0.6713	0.9295	0.7796	0.7490
	GBDT	0.9366	0.7454	0.9253	0.8256	0.7943
	XTrees	0.9366	0.6991	0.9805	0.8162	0.7955
	Hard	0.9310	0.6713	0.9797	0.7967	0.7765
	Soft	0.9245	0.6343	0.9856	0.7718	0.7546

Tablo 27’de farklı aşağı örneklemeye teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait bazı performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve aşağı örneklemeye teknikleriyle dengelenen HMEQ kredi veri seti aracılığıyla sınıflandırma başarısı test edilen topluluk modellerinin tahmin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %95,43 ile cluster centroids-XTrees aşağı örneklemeye tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örneklemeye tekniği-topluluk öğrenmesi modeli kombinasyonu belirlerken, kullanılan diğer tüm ölçüm

istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiğine ek olarak, kesinlik(precision), F1-skor ve MCC performans ölçüm istatistikleri açısından da en iyi performans yine cluster centroids-XTrees kombinasyonunda ölçülmüştür. Duyarlılık (recall) açısından en yüksek tahmin performansı, near miss-XTrees kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda senaryo 1’de, kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren aşağı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, cluster centroids-XTrees’ dir; yani değişken seçim tekniğinin uygulanan ve cluster centroids aşağı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde yer alan riskli müşterileri, senaryo 1’de en doğru tahmin eden topluluk öğrenmesi modeli ekstra ağaçlar (XTrees) modelidir.

Tablo 28: Senaryo 2: HMEQ Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	KNN	0.9171	0.8565	0.7613	0.8061	0.7557
	NB	0.7782	0.4444	0.4486	0.4465	0.3078
	LR	0.6897	0.7176	0.3630	0.4821	0.3278
	SVM	0.8621	0.8241	0.6181	0.7063	0.6295
	MLP	0.7614	0.6806	0.4401	0.5345	0.4004
	DT	0.7763	0.7778	0.4667	0.5833	0.4702
Near Miss	KNN	0.7886	0.8831	0.4955	0.6348	0.5436
	NB	0.7777	0.3105	0.4503	0.3675	0.2442
	LR	0.2081	1.0000	0.2081	0.3444	0.0000
	SVM	0.7626	0.8105	0.4600	0.5869	0.4721
	MLP	0.7836	0.8185	0.4880	0.6114	0.5049
	DT	0.8238	0.6492	0.5669	0.6053	0.4944
Cluster Centroids	KNN	0.9627	0.8622	0.9557	0.9065	0.8852
	NB	0.7642	0.3822	0.4300	0.4047	0.2590
	LR	0.6971	0.7111	0.3810	0.4961	0.3374
	SVM	0.8891	0.7378	0.7345	0.7361	0.6659
	MLP	0.7390	0.6756	0.4234	0.5205	0.3722
	DT	0.8155	0.7156	0.5458	0.6192	0.5083
Tomek-Links	KNN	0.9543	0.7778	0.9941	0.8727	0.8549
	NB	0.7856	0.3935	0.4620	0.4250	0.2957
	LR	0.8313	0.3102	0.6768	0.4254	0.3780
	SVM	0.9245	0.7315	0.8729	0.7960	0.7545
	MLP	0.9040	0.6667	0.8229	0.7366	0.6843
	DT	0.8788	0.6019	0.7471	0.6667	0.5989

Tablo 28’de farklı aşağı örnekleme teknikleri ve tekil model kombinasyonlarına ait bazı performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve aşağı örnekleme teknikleriyle dengelenen HMEQ kredi veri seti aracılığıyla sınıflandırma başarısı test edilen tekil modellerin tahmin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %96,27 ile cluster centroids-KNN aşağı örnekleme tekniği-tekil model kombinasyonlarına aittir. Ancak sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği ve tekil model kombinasyonlarını belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiklerine ek olarak, duyarlılık (recall), F1-skoru ve MCC performans ölçüm istatistiklerine bakıldığında en iyi performans yine, cluster centroids-KNN aşağı örnekleme tekniği-tekil model kombinasyonlarında ölçülmüştür. Kesinlik (precision) açısından en yüksek tahmin performansının ölçüldüğü diğer kombinasyon ise near miss-LR kombinasyonudur. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda senaryo 2’de, kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren aşağı örnekleme tekniği-tekil model kombinasyonu, cluster centroids-KNN’ dir; yani değişken seçim tekniğinin uygulanan ve cluster centroids aşağı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde yer alan riskli müşterileri, senaryo 2’de en doğru tahmin eden tekil model k-en yakın komşu (KNN) modelidir.

Tablo 29: Senaryo 3: HMEQ Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	RF	0.8938	0.8889	0.6920	0.7782	0.7194
	GBDT	0.8872	0.8933	0.6745	0.7686	0.7080
	XTrees	0.9413	0.9333	0.8140	0.8696	0.8352
	Hard	0.9273	0.8933	0.7882	0.8375	0.7935
	Soft	0.9301	0.8889	0.8000	0.8421	0.7992
Near Miss	RF	0.8444	0.8711	0.5868	0.7013	0.6228
	GBDT	0.8444	0.8756	0.5863	0.7023	0.6247
	XTrees	0.8304	0.9378	0.5567	0.6987	0.6300
	Hard	0.8649	0.8844	0.6258	0.7330	0.6633
	Soft	0.8742	0.9022	0.6424	0.7505	0.6868
Cluster Centroids	RF	0.8788	0.8622	0.6621	0.7490	0.6811
	GBDT	0.8416	0.8800	0.5806	0.6996	0.6219
	XTrees	0.9674	0.8933	0.9481	0.9199	0.9001
	Hard	0.9394	0.8844	0.8361	0.8596	0.8216
	Soft	0.9515	0.8756	0.8914	0.8834	0.8529
Tomek-Links	RF	0.9236	0.6844	0.9333	0.7897	0.7578
	GBDT	0.9273	0.7022	0.9349	0.8020	0.7703
	XTrees	0.9422	0.7289	0.9939	0.8410	0.8212
	Hard	0.9282	0.6711	0.9805	0.7968	0.7751
	Soft	0.9245	0.6578	0.9737	0.7851	0.7624

Tablo 29’da farklı aşağı örnekleme teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve aşağı örnekleme teknikleriyle dengelenen HMEQ kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %96,74 ile cluster centroids-XTrees aşağı örnekleme tekniği-topluluk öğrenmesi model kombinasyonuna aittir. Ancak, sadece doğruluk değerini göz önünde bulundurularak, performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu belirlerken, kullanılan diğer performans ölçüm istatistiklerine de bakmakta fayda vardır. Doğruluk ölçüm istatistiğine ek olarak, F1-skor ve MCC açısından da en iyi performans yine, cluster centroids-XTrees aşağı örnekleme tekniği-topluluk öğrenmesi model kombinasyonunda ölçülmüştür. Kesinlik (precision) açısından en iyi performans, Tomek links-XTrees kombinasyonunda ölçülürken, duyarlılık açısından en iyi performans ise near miss-XTrees kombinasyonunda ölçülmüştür. Sonuç olarak;

tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda senaryo 3’te, kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren aşağı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, cluster centroids-XTrees’ dir; yani değişken seçim tekniği uygulanmayan ve cluster centroids aşağı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde yer alan riskli müşterileri, senaryo 3’te en doğru tahmin eden topluluk öğrenmesi modeli ekstra ağaçlar (XTrees)’ dir.

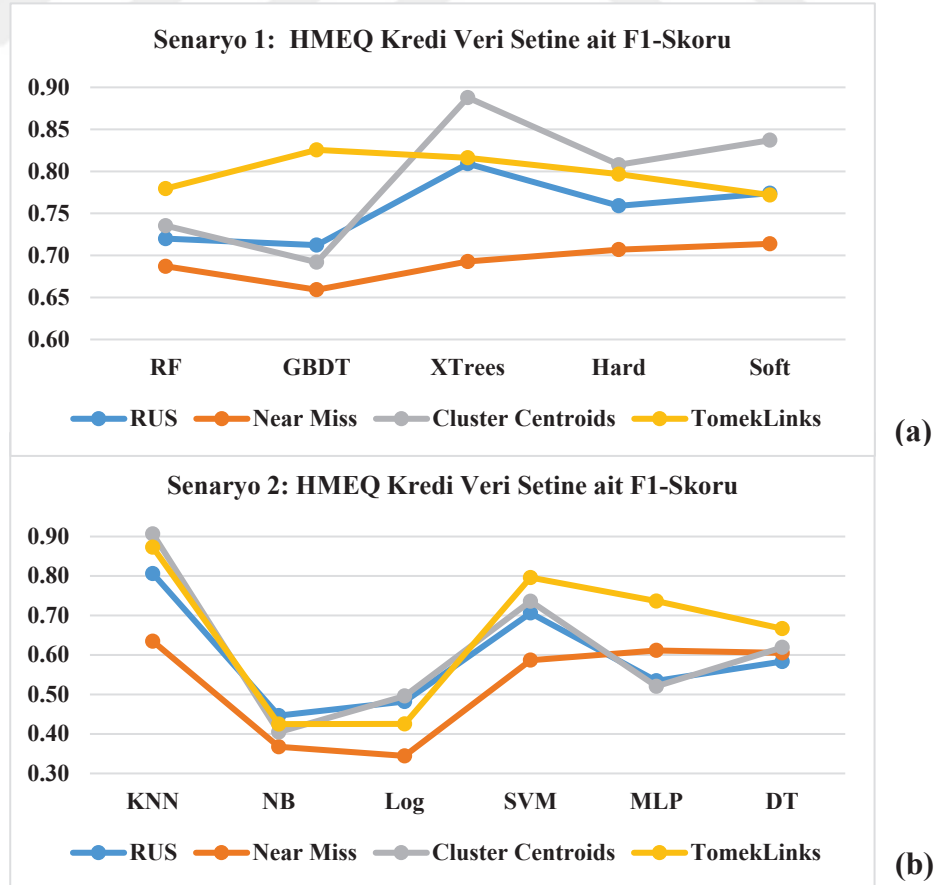
Tablo 30: Senaryo 4: HMEQ Kredi Veri Seti Performans İstatistikleri

Aşağı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
RUS	KNN	0.9329	0.8933	0.8072	0.8481	0.8069
	NB	0.7679	0.4400	0.4459	0.4430	0.2964
	LR	0.7120	0.7067	0.3955	0.5072	0.3533
	SVM	0.8826	0.8622	0.6713	0.7549	0.6884
	MLP	0.7717	0.7111	0.4706	0.5664	0.4365
	DT	0.8024	0.7689	0.5195	0.6201	0.5105
Near Miss	KNN	0.8006	0.9067	0.5139	0.6559	0.5726
	NB	0.7801	0.3778	0.4696	0.4187	0.2876
	LR	0.6664	0.7289	0.3557	0.4781	0.3114
	SVM	0.7819	0.8578	0.4886	0.6226	0.5230
	MLP	0.7400	0.6578	0.4229	0.5148	0.3643
	DT	0.8313	0.6844	0.5833	0.6299	0.5243
Cluster Centroids	KNN	0.9702	0.8756	0.9801	0.9249	0.9086
	NB	0.7698	0.4133	0.4471	0.4296	0.2860
	LR	0.6980	0.7067	0.3813	0.4953	0.3361
	SVM	0.9394	0.8356	0.8704	0.8526	0.8148
	MLP	0.7577	0.7644	0.4538	0.5695	0.4432
	DT	0.8015	0.7778	0.5178	0.6217	0.5132
Tomek-Links	KNN	0.9590	0.8133	0.9892	0.8927	0.8740
	NB	0.7754	0.4000	0.4592	0.4276	0.2897
	LR	0.8145	0.2533	0.6477	0.3642	0.3216
	SVM	0.9413	0.7556	0.9551	0.8437	0.8165
	MLP	0.8900	0.5422	0.8905	0.6740	0.6398
	DT	0.8975	0.7156	0.7778	0.7454	0.6823

Tablo 30’da farklı aşağı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve aşağı örnekleme tekniklerinin uygulandığı HMEQ kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen tekil modellerin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %97,02 ile cluster centroids-KNN aşağı örnekleme tekniği-tekil model kombinasyonuna aittir. Ancak, sadece

doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği ve tekil model kombinasyonu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda, Doğruluk ölçüm istatistiğine ek olarak kesinlik (precision), F1-skor ve MCC performans ölçüm istatistiklerine göre en iyi performans yine cluster centroids-KNN kombinasyonuna aittir. Duyarlılık (recall) açısından en yüksek tahmin performansı, near miss-KNN kombinasyonuna aittir. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda senaryo 4’te, kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren aşağı örnekleme tekniği-tekil model kombinasyonu cluster centroids-KNN’ dir; yani değişken seçim tekniği uygulanmayan ve cluster centroids aşağı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde yer alan riskli müşterileri, senaryo 4’ te en doğru tahmin eden tekil model k-en yakın komşu (KNN) modelidir.

Şekil 50: Senaryo 1 ve 2’ye ait F1 Skoru Grafiği: HMEQ Kredi Veri Seti



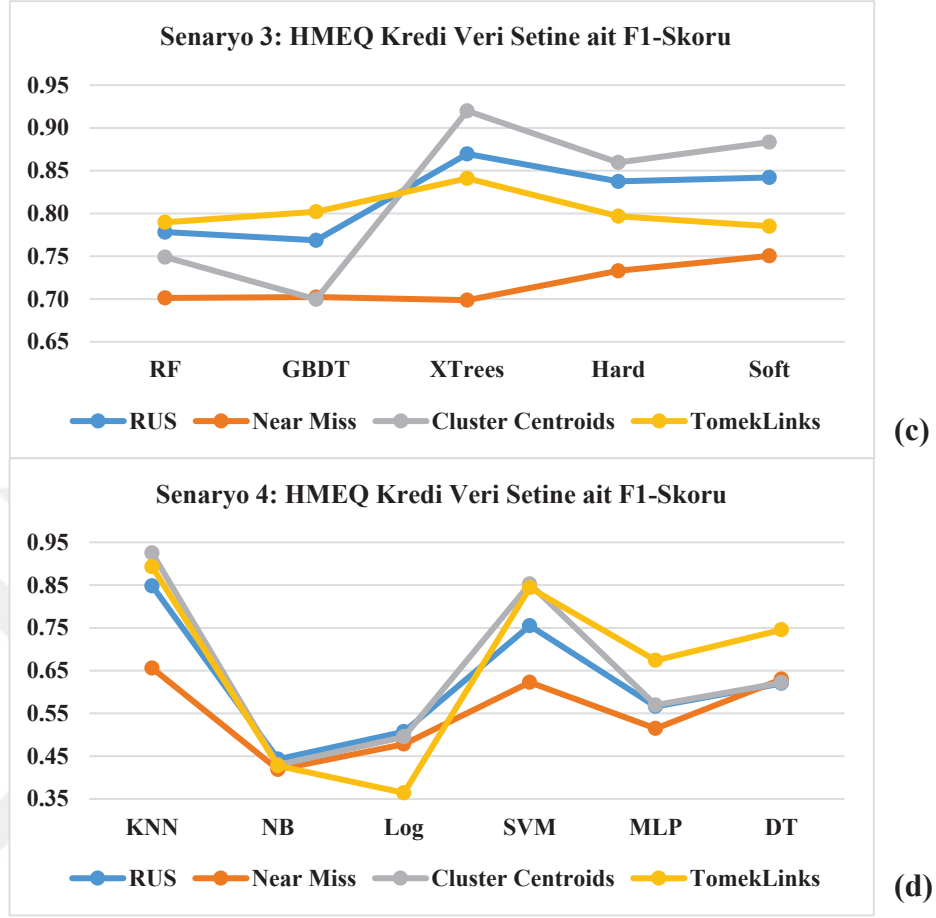
Yukarıdaki çizgi grafikleri, aşağı örnekleme teknikleriyle dengelenen HMEQ kredi veri seti üzerinde sınıflandırma performansı ölçülen tekil ve topluluk modellerine ait F1-skoru değerlerinin dağılımını gösterilmektedir. Bu grafiklere bakılarak model ve aşağı örnekleme teknikleri için elde edilen çıkarımlar ise şöyledir;

(a) grafiğine göre; kullanılan tüm topluluk öğrenmesi modellerinin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model ekstra ağaçlar (XTrees)'dir. Aynı zamanda, kullanılan tüm topluluk modellerinin arasından ortalama olarak en düşük tahmin performansının ölçüldüğü model ise GBDT modelidir. Topluluk modellerinin nerdeyse tamamı, en yüksek tahmin performansını Tomek links aşağı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir

(b) grafiğine göre; kullanılan tüm tekil modelleri arasından ortalama olarak en düşük tahmin performansının ölçüldüğü teknikler Naive Bayes (NB) ve lojistik regresyon (LR) teknikleridir. Bunun yanı sıra, kullanılan tüm tekil modeller arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model ise k-en yakın komşu (KNN) modelidir. Ek olarak, tekil modellerin nerdeyse tamamı en yüksek tahmin performansını, Tomek links aşağı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir.

(a) ve (b) grafiklerini kıyaslırsak eğer; kullanılan tüm sınıflandırma modellerinin arasından ortalama olarak en düşük tahmin performansının ölçüldüğü teknikler Naive Bayes (NB) ve lojistik regresyon (LR) teknikleridir. Bunun yanı sıra senaryo 1 ve 2'de kullanılan makine öğrenmesi yöntemleri tahmin performansına göre kıyaslandığında, topluluk öğrenmesi modellerinin tekil modellere karşın performans açısından bir üstünlüğünün bulunduğunu söylemek mümkündür.

Şekil 51: Senaryo 3 ve 4'e ait F1-Skoru Grafiği: HMEQ Kredi Veri Seti



(c) grafiğine göre; kullanılan tüm topluluk modellerinin arasında en yüksek tahmin performansının ölçüldüğü model ekstra ağaçlar (XTrees) modelidir. Aynı zamanda, kullanılan tüm topluluk modellerinin arasında ortalama olarak en düşük tahmin performansının ölçüldüğü model GBDT modelidir. Bunun yanı sıra, çalışmada kullanılan topluluk öğrenmesi tekniklerinin neredeyse tamamı en yüksek tahmin performansını cluster centroids aşağı örnekleme tekniğiyle dengelenen veri setinde elde etmiştir. Ağaç tabanlı ve oylama yöntemine ait modelleri performans açısından karşılaştırıldığında ise oylama yöntemi modelleri, ağaç tabanlı modellere göre tahmin performansı daha iyi olduğunu söylemek mümkündür.

(d) grafiğine göre; kullanılan tüm tekil modellerin arasında ortalama olarak en düşük tahmin performansının ölçüldüğü sınıflandırma teknikleri Naive Bayes (NB) ve lojistik regresyon (LR) teknikleridir. Bunun yanı sıra, kullanılan tüm tekil modellerin arasında ortalama olarak en yüksek tahmin performansının ölçüldüğü model k-en yakın komşu (KNN) modelidir. Bunun yanı sıra, çalışmada kullanılan tekil

modellerin nerdeyse tamamı, en yüksek tahmin performansını Tomek links aşağı örnekleme tekniğiyle dengelenen veri setinde elde etmiştir.

(c) ve (d) grafiklerini kıyaslarsak eğer; senaryo 3 ve 4'te kullanılan tüm makine öğrenmesi sınıflandırma tekniklerinin arasından genel olarak en düşük tahmin performansının ölçüldüğü yöntem Naive Bayes (NB)'tir. Aynı zamanda senaryo 3 ve 4'te yer alan makine öğrenmesi yöntemleri tahmin performanslarına göre kıyaslandığında, topluluk öğrenmesi modellerinin tekil modellere karşın performans açısından bir üstünlüğünün bulunduğunu söylemek mümkündür.

(a) ve (c) grafiklerini kıyaslarsak eğer; senaryo 4'ün, HMEQ kredi veri setinde bulunan riskli müşterileri tahmin etme konusunda kullanılacak olan tüm tekil modellerin tahmin performansını arttırdığını söylemek mümkündür; yani senaryo 4, çalışmada kullanılan tüm tekil modellerin tahmin performansını arttırmada, senaryo 2'ye göre daha başarılı olmuştur.

(b) ve (d) grafiklerini kıyaslarsak eğer; senaryo 3'ün, HMEQ kredi veri setinde bulunan riskli müşterileri sınıflandırmak için kullanılacak olan topluluk öğrenmesi modellerinin tahmin performansını arttırdığını söylemek mümkündür; yani senaryo 3, çalışmada kullanılan tüm topluluk öğrenmesi modellerinin tahmin performansını arttırmada senaryo 1'e göre daha başarılı olmuştur.

Tablo 31: Senaryo 5: HMEQ Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	RF	0.9226	0.7361	0.8595	0.7930	0.7492
	GBDT	0.9338	0.7361	0.9191	0.8175	0.7848
	XTrees	0.9292	0.6620	0.9795	0.7901	0.7702
	Hard	0.9394	0.7222	0.9689	0.8276	0.8044
	Soft	0.9413	0.7222	0.9811	0.8320	0.8112
ADASYN	RF	0.9256	0.7661	0.8610	0.8108	0.7667
	GBDT	0.9306	0.7500	0.9000	0.8182	0.7807
	XTrees	0.9541	0.8091	0.9647	0.8801	0.8571
	Hard	0.9430	0.7903	0.9245	0.8522	0.8210
	Soft	0.9407	0.7823	0.9209	0.8459	0.8137
SMOTE	RF	0.9301	0.7733	0.8788	0.8227	0.7819
	GBDT	0.9245	0.7378	0.8830	0.8039	0.7623
	XTrees	0.9534	0.8089	0.9630	0.8792	0.8556
	Hard	0.9385	0.7733	0.9206	0.8406	0.8075
	Soft	0.9310	0.7422	0.9126	0.8186	0.7829
Borderline-SMOTE	RF	0.9292	0.7778	0.8706	0.8216	0.7795
	GBDT	0.9282	0.7378	0.9022	0.8117	0.7739
	XTrees	0.9562	0.8311	0.9541	0.8884	0.8644
	Hard	0.9432	0.7956	0.9227	0.8544	0.8228
	Soft	0.9404	0.7867	0.9171	0.8469	0.8138
SVM-SMOTE	RF	0.9292	0.7689	0.8782	0.8199	0.7787
	GBDT	0.9310	0.7600	0.8953	0.8221	0.7837
	XTrees	0.9553	0.8089	0.9733	0.8835	0.8617
	Hard	0.9404	0.7822	0.9215	0.8462	0.8136
	Soft	0.9366	0.7644	0.9198	0.8350	0.8014

Tablo 31’de farklı yukarı örnekleme teknikleri ve topluluk öğrenmesi modeli kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve yukarı örnekleme tekniklerinin uygulandığı HMEQ kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %95,62 ile BorderlineSMOTE-XTrees örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm istatistiklerine ek olarak, duyarlılık (recall), F1-skoru ve MCC performans ölçüm metrikleri açısından en iyi performans, yine BorderlineSMOTE-XTrees yukarı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonunda ölçülmüştür.

Kesinlik (precision) açısından en iyi performans SVMSMOTE-XTrees kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önünde bulundurulduğunda senaryo 5’te, kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren yukarı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, BorderlineSMOTE-XTrees’ dir; yani değişken seçim tekniği uygulanan ve BorderlineSMOTE yukarı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde yer alan riskli müşterileri, senaryo 5’te en doğru tahmin eden topluluk öğrenmesi modeli ekstra ağaçlar (XTrees)’ dır.

Tablo 32: Senaryo 6: HMEQ Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	KNN	0.9525	0.7685	0.9940	0.8668	0.8488
	NB	0.7782	0.4491	0.4491	0.4491	0.3102
	LR	0.7055	0.6898	0.3744	0.4853	0.3314
	SVM	0.9236	0.7454	0.8564	0.7970	0.7530
	MLP	0.9217	0.7500	0.8438	0.7941	0.7480
	DT	0.8760	0.5926	0.7399	0.6581	0.5889
ADASYN	KNN	0.9457	0.7715	0.9599	0.8554	0.8300
	NB	0.7780	0.4892	0.4679	0.4783	0.3375
	LR	0.6784	0.7419	0.3656	0.4898	0.3317
	SVM	0.9021	0.7097	0.7976	0.7511	0.6922
	MLP	0.8949	0.7339	0.7541	0.7439	0.6778
	DT	0.8736	0.6828	0.7017	0.6921	0.6127
SMOTE	KNN	0.9534	0.8044	0.9679	0.8786	0.8557
	NB	0.7698	0.4311	0.4491	0.4399	0.2952
	LR	0.7176	0.6844	0.3990	0.5041	0.3485
	SVM	0.9199	0.7733	0.8325	0.8018	0.7525
	MLP	0.8928	0.7200	0.7570	0.7380	0.6711
	DT	0.8677	0.6267	0.7085	0.6651	0.5847
Borderline-SMOTE	KNN	0.9543	0.8089	0.9681	0.8814	0.8586
	NB	0.7754	0.4711	0.4649	0.4680	0.3257
	LR	0.6775	0.7333	0.3659	0.4882	0.3266
	SVM	0.9208	0.7511	0.8535	0.7991	0.7523
	MLP	0.8928	0.7200	0.7570	0.7380	0.6711
	DT	0.8844	0.6933	0.7393	0.7156	0.6437
SVM-SMOTE	KNN	0.9525	0.7956	0.9728	0.8753	0.8528
	NB	0.7670	0.4178	0.4413	0.4292	0.2832
	LR	0.7018	0.7111	0.3855	0.5000	0.3431
	SVM	0.9152	0.7244	0.8490	0.7818	0.7331
	MLP	0.9096	0.7556	0.8019	0.7780	0.7218
	DT	0.8910	0.7022	0.7596	0.7298	0.6624

Tablo 32’de farklı yukarı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, değişken seçimi ve yukarı örnekleme tekniklerinin uygulandığı HMEQ kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %95,43 ile BorderlineSMOTE-KNN örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-tekil model kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm metriğine ek olarak, duyarlılık (recall), F1-skoru ve MCC performans ölçüm metrikleri açısından da en iyi performans yine, BorderlineSMOTE-KNN kombinasyonunda ölçülmüştür. Kesinlik (precision) açısından en iyi performans ise ROS-KNN kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önüne alındığında, senaryo 6’da kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren yukarı örnekleme tekniği-tekil model kombinasyonu, BorderlineSMOTE-KNN’ dir. Yani, değişken seçim tekniği uygulanan ve BorderlineSMOTE yukarı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde yer alan riskli müşterileri, senaryo 6’da en doğru tahmin eden tekil model k-en yakın komşu (KNN) modelidir.

Tablo 33: Senaryo 7: HMEQ Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Topluluk Öğrenmesi Modelleri	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	RF	0.9320	0.7600	0.9000	0.8241	0.7866
	GBDT	0.9301	0.7422	0.9076	0.8166	0.7800
	XTrees	0.9413	0.7200	1.0000	0.8372	0.8187
	Hard	0.9432	0.7689	0.9505	0.8501	0.8225
	Soft	0.9404	0.7511	0.9548	0.8408	0.8135
ADASYN	RF	0.9404	0.8267	0.8815	0.8532	0.8165
	GBDT	0.9338	0.7733	0.8969	0.8305	0.7931
	XTrees	0.9664	0.8711	0.9655	0.9159	0.8969
	Hard	0.9543	0.8444	0.9314	0.8858	0.8589
	Soft	0.9525	0.8222	0.9439	0.8789	0.8526
SMOTE	RF	0.9366	0.8089	0.8792	0.8426	0.8041
	GBDT	0.9376	0.7911	0.8990	0.8416	0.8055
	XTrees	0.9618	0.8444	0.9694	0.9026	0.8822
	Hard	0.9534	0.8222	0.9487	0.8810	0.8555
	Soft	0.9469	0.7867	0.9516	0.8613	0.8346
Borderline-SMOTE	RF	0.9413	0.8178	0.8932	0.8538	0.8184
	GBDT	0.9441	0.8089	0.9146	0.8585	0.8262
	XTrees	0.9637	0.8444	0.9794	0.9069	0.8882
	Hard	0.9543	0.8222	0.9536	0.8831	0.8585
	Soft	0.9562	0.8178	0.9684	0.8867	0.8646
SVM-SMOTE	RF	0.9413	0.8178	0.8932	0.8538	0.8184
	GBDT	0.9441	0.8089	0.9146	0.8585	0.8262
	XTrees	0.9637	0.8444	0.9794	0.9069	0.8882
	Hard	0.9543	0.8222	0.9536	0.8831	0.8585
	Soft	0.9562	0.8178	0.9684	0.8867	0.8646

Tablo 33’te farklı yukarı örnekleme teknikleri ve topluluk öğrenmesi modelleri kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm değişkenlerin var olduğu ve yukarı örnekleme tekniklerinin uygulandığı HMEQ kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen topluluk öğrenmesi modellerinin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %96,64 ile ADASYN-XTrees örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm metriğine ek olarak duyarlılık (recall), F1-skoru ve MCC performans ölçüm istatistiklerine bakıldığında da en iyi performans yine, ADASYN-XTrees kombinasyonunda ölçülmüştür. Kesinlik (precision) açısından ise en iyi performans

ROS-XTrees kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önüne alındığında, senaryo 7 için kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren yukarı örnekleme tekniği-topluluk öğrenmesi modeli kombinasyonu, ADASYN-XTrees’ dir. Yani, değişken seçim tekniği uygulanmayan ve ADASYN yukarı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde yer alan riskli müşterileri, senaryo 7’de en doğru tahmin eden topluluk öğrenmesi modeli ekstra ağaçlar (XTrees)’ dir.

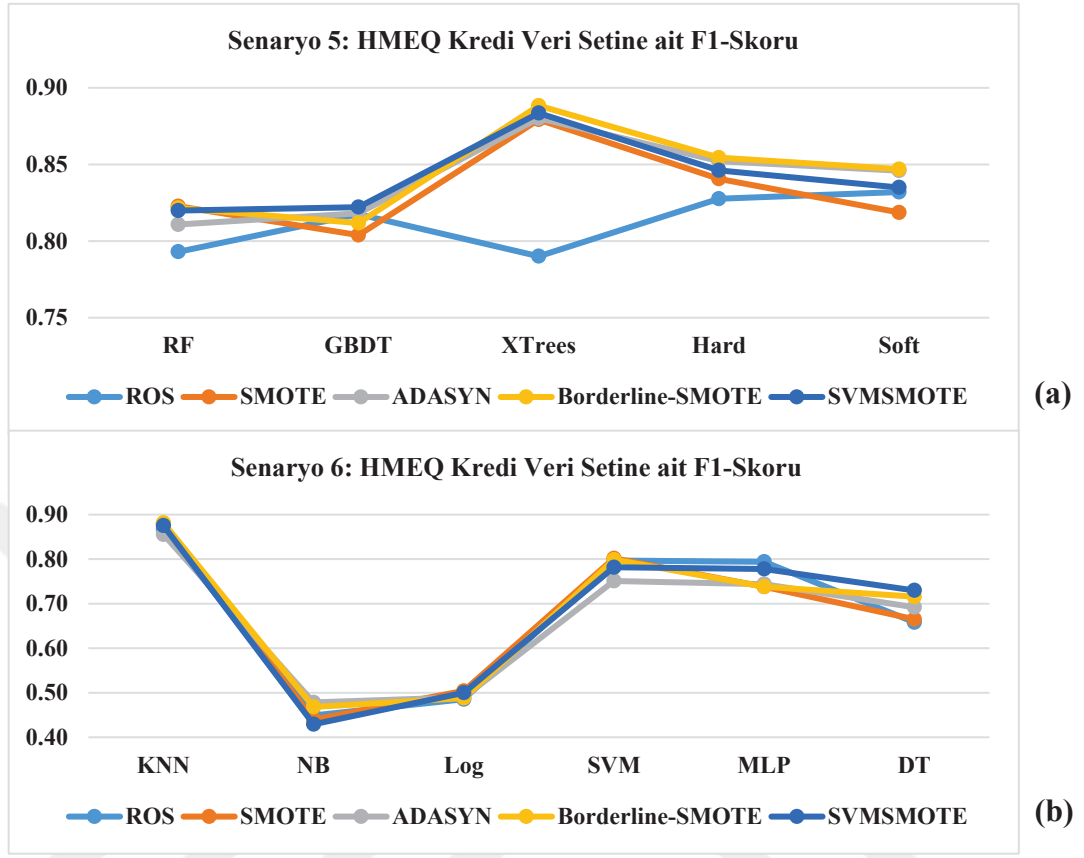
Tablo 34: Senaryo 8: HMEQ Kredi Veri Seti Performans İstatistikleri

Yukarı Örnekleme	Tekil Modeller	Doğruluk (%)	Recall (%)	Precision (%)	F1-Skor (%)	MCC
ROS	KNN	0.9581	0.8089	0.9891	0.8900	0.8711
	NB	0.7689	0.4311	0.4470	0.4389	0.2935
	LR	0.7120	0.6889	0.3934	0.5008	0.3438
	SVM	0.9338	0.7600	0.9096	0.8281	0.7924
	MLP	0.9059	0.7822	0.7719	0.7770	0.7174
	DT	0.8816	0.6178	0.7722	0.6864	0.6204
ADASYN	KNN	0.9637	0.8400	0.9844	0.9065	0.8884
	NB	0.7745	0.4933	0.4644	0.4784	0.3350
	LR	0.6766	0.7378	0.3656	0.4890	0.3281
	SVM	0.9245	0.7600	0.8636	0.8085	0.7642
	MLP	0.8826	0.7644	0.7020	0.7319	0.6579
	DT	0.9003	0.7333	0.7783	0.7551	0.6931
SMOTE	KNN	0.9627	0.8311	0.9894	0.9034	0.8856
	NB	0.7651	0.4267	0.4384	0.4324	0.2844
	LR	0.7018	0.6978	0.3839	0.4953	0.3358
	SVM	0.9282	0.7689	0.8737	0.8180	0.7760
	MLP	0.9115	0.7511	0.8125	0.7806	0.7261
	DT	0.8844	0.6800	0.7463	0.7116	0.6406
Borderline-SMOTE	KNN	0.9609	0.8222	0.9893	0.8981	0.8798
	NB	0.7670	0.4133	0.4408	0.4266	0.2808
	LR	0.6980	0.7111	0.3819	0.4969	0.3385
	SVM	0.9459	0.7911	0.9418	0.8599	0.8316
	MLP	0.9152	0.7822	0.8073	0.7946	0.7413
	DT	0.8947	0.7111	0.7692	0.7390	0.6740
SVM-SMOTE	KNN	0.9609	0.8222	0.9893	0.8981	0.8798
	NB	0.7670	0.4133	0.4408	0.4266	0.2808
	LR	0.6980	0.7111	0.3819	0.4969	0.3385
	SVM	0.9459	0.7911	0.9418	0.8599	0.8316
	MLP	0.9152	0.7822	0.8073	0.7946	0.7413
	DT	0.8947	0.7111	0.7692	0.7390	0.6740

Tablo 34’te farklı yukarı örnekleme teknikleri ve tekil model kombinasyonlarına ait performans istatistikleri yer almaktadır. Bu istatistikler, tüm

değişkenlerin var olduğu ve yukarı örnekleme tekniklerinin uygulandığı HMEQ kredi veri seti aracılığıyla, sınıflandırma başarısı test edilen tekil modellerin gücünü gösterir. Tabloya göre; doğruluk açısından en yüksek performans, %96,37 ile ADASYN-KNN örnekleme tekniği-model kombinasyonuna aittir. Ancak, sadece doğruluk değerleri göz önünde bulundurularak performans hakkında bir sonuca varmak yanıltıcı olabilir. Bu nedenle, optimal bir örnekleme tekniği-tekil model kombinasyonunu belirlerken, kullanılan diğer tüm performans ölçüm istatistiklerine de bakmakta fayda vardır. Bu bağlamda doğruluk ölçüm metriğine ek olarak, MCC, duyarlılık (recall) ve F1-skoru performans ölçüm istatistiklerine bakıldığında da en iyi performans yine, ADASYN-KNN kombinasyonunda ölçülmüştür. Kesinlik (precision) açısından en iyi performans, SMOTE-XTrees kombinasyonunda ölçülmüştür. Sonuç olarak; tabloda yer alan performans istatistiklerinin geneli göz önüne alındığında, senaryo 8’de kredi riski bulunan müşterileri doğru tahmin etme konusunda yüksek performans gösteren yukarı örnekleme tekniği-tekil model kombinasyonu, ADASYN-KNN’ dir. Yani, değişken seçim tekniği uygulanmayan ve ADASYN yukarı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde yer alan riskli müşterileri, senaryo 8’de en doğru tahmin eden tekil model k-en yakın komşu (KNN) modelidir.

Şekil 52: Senaryo 5 ve 6'ya ait F1-Skoru Grafiği: HMEQ Kredi Veri Seti



Çizgi grafikleri, yukarı örnekleme teknikleriyle dengelenen HMEQ kredi veri seti üzerinde sınıflandırma performansı ölçülen tekil ve topluluk modellerine ait F1-skoru değerlerinin dağılımını gösterilmektedir. Bu grafiklere bakılarak model ve yukarı örnekleme teknikleri için elde edilen çıkarımlar ise şöyledir;

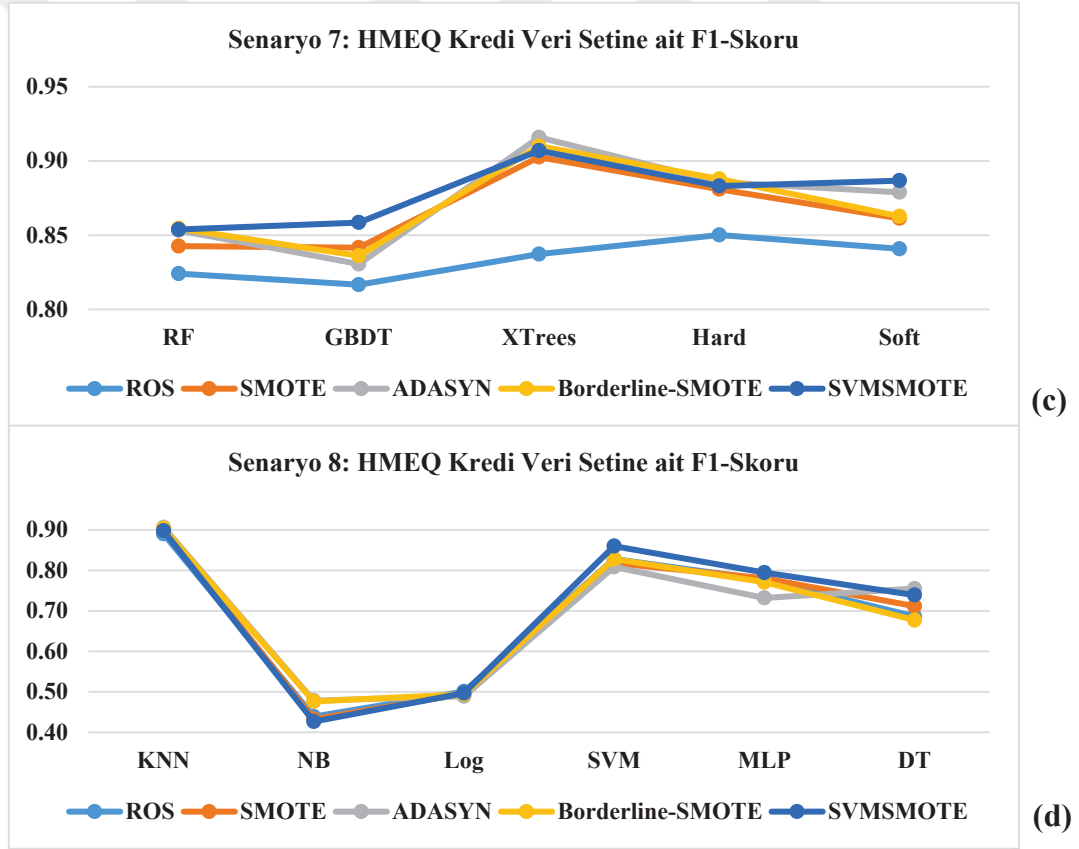
(a) grafiğine göre; kullanılan tüm topluluk öğrenmesi modellerinin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model ekstra ağaçlar (XTrees) modelidir. Topluluk modellerinin neredeyse tamamı, en yüksek tahmin performansını Borderline-SMOTE yukarı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir.

(b) grafiğine göre; kullanılan tüm tekil modelleri arasında ortalama olarak en düşük tahmin performansı Naive Bayes (NB) ve lojistik regresyon (LR)'a aittir. Aynı zamanda Naive Bayes ve lojistik regresyon, tüm yukarı örnekleme teknikleriyle dengelenen HMEQ kredi veri setindeki riskli müşterileri sınıflandırmada en düşük tahmin performansını göstermiştir. Bunun yanı sıra, kullanılan tüm tekil modellerin

arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model ise k-en yakın komşu (KNN) modelidir. Ek olarak, tekil modellerin tamamı en yüksek tahmin performansını, SVMSMOTE yukarı örnekleme tekniğiyle dengelenen veri setini sınıflandırırken göstermiştir.

(a) ve (b) grafiklerine göre; kullanılan tüm sınıflandırma modellerinin arasından ortalama olarak en düşük tahmin performansının ölçüldüğü yöntem Naive Bayes (NB)'tir. Bunun yanı sıra senaryo 5 ve 6'da kullanılan tekil ve topluluk öğrenmesi modellerinin tahmin performanslarını kıyaslırsak eğer, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu söylenebilir.

Şekil 53: Senaryo 7 ve 8'e ait F1-Skoru Grafiği: HMEQ Kredi Veri Seti



(c) grafiğine göre; kullanılan tüm topluluk modellerinin arasından en yüksek tahmin performansının ölçüldüğü model ekstra ağaçlar (XTrees) modelidir. Aynı zamanda, kullanılan topluluk öğrenmesi tekniklerinin arasından en düşük tahmin performansının ölçüldüğü model GBDT'dir. Bunun yanı sıra, çalışmada kullanılan

topluluk öğrenmesi tekniklerinin nerdeyse tamamı en yüksek tahmin performansını SVM-SMOTE yukarı örnekleme tekniğiyle dengelenen veri setinde elde etmiştir.

(d) grafiğine göre; kullanılan tüm tekil modellerin arasından ortalama olarak en düşük tahmin performansının ölçüldüğü teknikler Naive Bayes (NB) ve lojistik regresyon (LR) teknikleridir. Bunun yanı sıra, kullanılan tüm tekil modellerin arasından ortalama olarak en yüksek tahmin performansının ölçüldüğü model ise k-en yakın komşu (KNN) modelidir. Bunun yanı sıra, çalışmada kullanılan tekil modellerin nerdeyse tamamı, en yüksek tahmin performansını SVM-SMOTE örnekleme tekniğiyle dengelenen veri setinde elde etmiştir.

(c) ve (d) grafiklerini kıyaslırsak eğer; kullanılan tekil sınıflandırma tekniklerinin arasından en düşük performansın ölçüldüğü teknikler Naive Bayes (NB) ve lojistik regresyon (LR) teknikleridir. Bunun yanı sıra senaryo 7 ve 8’de kullanılan tekil ve topluluk öğrenmesi modellerinin tahmin performanslarını kıyaslırsak eğer, topluluk öğrenmesi modellerinin tahmin performansının daha yüksek olduğu söylenebilir.

(a) ve (c) grafiklerini kıyaslırsak eğer; senaryo 8’in, HMEQ kredi veri setinde bulunan riskli müşterileri sınıflandırmada kullanılacak olan tüm tekil modellerin tahmin performansını arttırdığını söylemek mümkündür. Sonuç olarak; senaryo 8, çalışmada kullanılan tüm tekil modellerin tahmin performansını arttırmada, senaryo 6’ya göre daha başarılı olmuştur.

(b) ve (d) grafiklerini kıyaslırsak eğer; senaryo 7’nin, HMEQ kredi veri setinde bulunan riskli müşterileri sınıflandırmak için kullanılacak olan topluluk öğrenmesi modellerinin tahmin performansını arttırdığını söylemek mümkündür. Sonuç olarak; senaryo 7, çalışmada kullanılan tüm topluluk öğrenmesi modellerinin tahmin performansını artırma konusunda senaryo 5’e göre daha başarılı olmuştur.

Çalışmada HMEQ kredi veri seti için kullanılan sekiz farklı senaryodan genel olarak elde edilen çıkarımlar ise şöyledir;

- Çalışmada kullanılan tüm makine öğrenmesi teknikleri arasından genel olarak en düşük tahmin performansının görüldüğü yöntem, Naive Bayes (NB) yöntemidir.
- Aşağı örnekleme tekniklerinin yer aldığı senaryoların tümünde, kullanılan sınıflandırma modellerinin tamamı ortalama olarak en yüksek tahmin

performansını Tomek links aşağı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde elde etmiştir.

- Yukarı örnekleme tekniklerinin yer aldığı senaryoların tümünde, kullanılan sınıflandırma modellerinin nerdeyse tamamı ortalama olarak en yüksek tahmin performansını SVM-SMOTE yukarı örnekleme tekniğiyle dengelenen HMEQ kredi veri setinde elde etmiştir.
- HMEQ kredi veri setine değişken seçim tekniğinin uygulamanın, makine öğrenmesi tekniklerinin tahmin performansına önemli bir etkisinin olmadığı görülmüştür.

3.4.4. Optimal Senaryoların Seçimi

Bu çalışmada, kredi riskini azaltacak ve aynı zamanda karı maksimize edecek bir senaryonun tespit edilip bankacılık sektöründe benzer problemlerin çözümü için önerileceğinden söz etmiştik. Bu bağlamda oluşturulan senaryoların içerdiği tekniklerin arasından, kredi riski yüksek olan müşterilerin en doğru ve en etkin şekilde tespit edecek optimal bir teknik belirlenmeye çalışılacaktır. Bu optimal tekniğin belirlenmesi ise ancak, çalışma için oluşturulan senaryolarda yer alan tekniklerin tahmin performanslarının ölçülmesiyle mümkün olacaktır. Fakat problemin yapısına ve önemine bağlı olarak, performansı test edilecek teknikler için temel alınacak performans ölçüm istatistiği de değişkenlik göstermektedir. Örneğin; problem dengeli bir veri setine sahipse eğer, senaryoda bulunan tekniklerin performansı hakkında bir çıkarımda bulunmak için doğruluk ölçüm istatistiğine bakmak yeterli olacaktır. Ancak, kredi risk analizinin yapıldığı çalışmalarda, kredi riski bulunan müşteriler ile kredi riski bulunmayan müşterilerin sayısındaki dengesizlik sebebiyle bu tip problemlerde sadece doğruluk ölçüm istatistiğine göre en iyi modele karar vermek sağlıklı olmayacaktır.

Kredi risk analizi üzerine yürütülen çalışmalarda, riskli olduğu halde risksiz olarak sınıflandırılan müşteriler (yanlış pozitif) ile risksiz olduğu halde riskli olarak sınıflandırılan müşteriler (yanlış negatif) genellikle aynı öneme sahiptir. Çünkü yanlış pozitif olarak etiketlenen müşterilerden kaynaklanan kayıp ile yanlış negatif olarak etiketlenen müşterilerden kaynaklanan kaybın etkisi bankacılık sektörü açısından aynı olacaktır. Bu nedenle, yanlış pozitifte sınıflandırılan müşterilerin sayısı ile hesaplanan

duyarlılık ölçüsü ile yanlış negatifte sınıflandırılan müşterilerin sayısı ile hesaplanan kesinlik ölçüm istatistiği arasındaki dengeyi harmonik ortalamayla sağlayan F1-skoru ölçüm istatistiği, kredi risk analizi üzerinde yürütülen çalışmaların neredeyse tümünde en iyi sınıflandırma modelini belirlemede temel alınan bir performans ölçüm istatistiklerinden biri olmuştur. Doğruluk ölçüm istatistiğine ek olarak, F1-skoru ölçüm istatistiği de bu çalışma kapsamında kullanılan makine öğrenmesi modelleri arasından en yüksek tahmin performansına sahip sınıflandırma modelinin belirlenmesinde de temel alınacak ölçüm istatistiğinden olacaktır.

Çalışmanın bu kısmında; Alman, Avustralya ve HMEQ kredi veri setiyle çalıştırılan sekiz farklı senaryodan elde edilen optimal örnekleme tekniği-model kombinasyonlarına ait performans istatistikleri tablo 35 ve tablo 36’da bir araya getirilmiş ve karşılaştırılmıştır. Uygulama kapsamında elde edilen en optimal örnekleme tekniği-model kombinasyonları ise tablolar üzerinde belirginleştirilmiştir.

Tablo 35: Senaryolara ait En Yüksek F1-Skoru Değerleri

Veri Setleri	SENARYOLAR							
	Senaryo 1	Senaryo 2	Senaryo 3	Senaryo 4	Senaryo 5	Senaryo 6	Senaryo 7	Senaryo 8
<i>Alman</i>	0.6234	0.6497	0.6406	0.6575	0.6496	0.6577	0.6457	0.6490
Örnekleme T.	RUS	RUS	Tomek	RUS	ROS	ROS	ADASYN	SVMSMOTE
Model	Soft	NB	Soft	NB	RF	LR	Hard	LR
<i>Avustralya</i>	0.8468	0.8621	0.8785	0.8829	0.8762	0.8909	0.8679	0.8807
Örnekleme T.	RUS	Tomek	RUS	RUS	ROS	ROS	SMOTE	ROS
Model	RF	LR/SVM	RF	SVM	GBDT	SVM	GBDT	LR
<i>HMEQ</i>	0.8879	0.9065	0.9199	0.9249	0.8884	0.8814	0.9159	0.9065
Örnekleme T.	Cluster	Cluster	Cluster	Cluster	B. SMOTE	B. SMOTE	ADASYN	ADASYN
Model	XTrees	KNN	XTrees	KNN	XTrees	KNN	XTrees	KNN

Tablo 36: Senaryolara ait En Yüksek Doğruluk Değerleri

	SENARYOLAR							
Veri Setleri	Senaryo 1	Senaryo 2	Senaryo 3	Senaryo 4	Senaryo 5	Senaryo 6	Senaryo 7	Senaryo 8
<i>Alman</i>	0.7700	0.7400	0.7750	0.7550	0.7950	0.7650	0.8000	0.7350
Örnekleme T.	Tomek	Tomek	Tomek	Tomek	ROS	SVMSMOTE	ROS	SVMSMOTE ROS
Model	XTrees	LR	RF GBDT	DT	RF	SVM	GBDT	LR
<i>Avustralya</i>	0.8768	0.8841	0.8960	0.8960	0.8960	0.9040	0.8880	0.8960
Örnekleme T.	RUS	Tomek	RUS	RUS	ROS	ROS	SMOTE	ROS
Model	RF	LR/SVM	RF	SVM MLP	GBDT	SVM	GBDT	LR
<i>HMEQ</i>	0.9543	0.9627	0.9674	0.9702	0.9562	0.9543	0.9664	0.9637
Örnekleme T.	Cluster	Cluster	Cluster	Cluster	B. SMOTE	B. SMOTE	ADASYN	ADASYN
Model	XTrees	KNN	XTrees	KNN	XTrees	KNN	XTrees	KNN

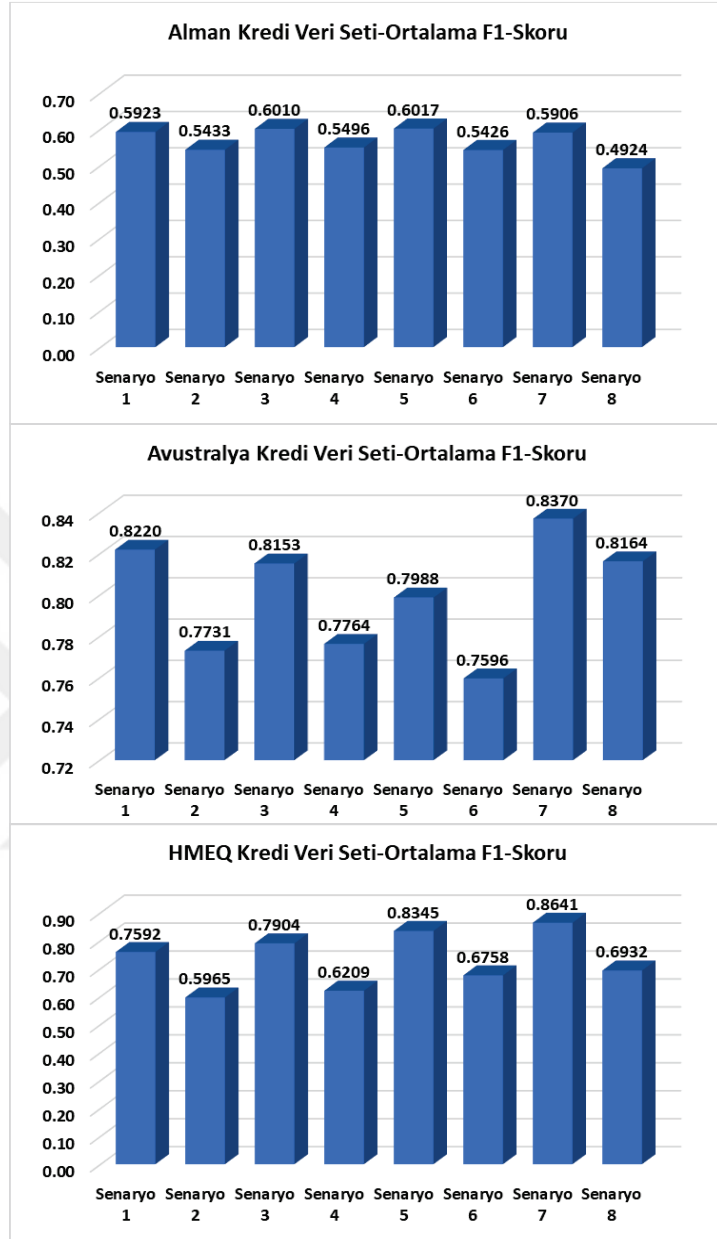
Alman kredi veri setiyle çalıştırılan senaryolar arasından, doğruluk açısından, en optimal sınıflandırma performansı ölçüldüğü örnekleme tekniği-model kombinasyonu senaryo 7'deki rastgele yukarı örnekleme (ROS)-gradyan artırma karar ağaçları (GBDT) kombinasyonudur; yani değişken sayısı 50 olan ve ROS yukarı örnekleme tekniğiyle dengelenmiş Alman kredi veri setiyle çalıştırılan topluluk öğrenmesi teknikleri arasından en optimal sınıflandırma performansını gösteren teknik, gradyan artırma karar ağaçları (GBDT)'dir. Alman kredi veri setiyle çalıştırılan senaryolar arasından, F1-skoru açısından, en optimal sınıflandırma performansı ölçüldüğü örnekleme tekniği-model kombinasyonu, senaryo 6'daki rastgele yukarı örnekleme (ROS)-lojistik regresyon (LR) kombinasyonudur; yani değişken sayısı 50'den 37'ye düşürülmüş ve ROS yukarı örnekleme tekniğiyle dengelenmiş Alman kredi veri setiyle çalıştırılan tekil modeller arasından en optimal sınıflandırma performansını gösteren model, lojistik regresyon (LR)'dur. Hem doğruluk hem de F1-skoru ölçüm istatistikleri açısından belirlenen optimal kombinasyonlar farklılık göstermektedir. Bu belirsizliği ortadan kaldırmak için çalışmada kullanılan diğer performans ölçüm tekniklerine bakmakta vardır. Diğer performans ölçüm

istatistiklerinden elde edilen sonuçlar göz önünde bulundurulduğunda, en optimal sınıflandırma performansına sahip kombinasyon, ROS-GBDT olarak belirlenmiştir.

Avustralya kredi veri setiyle çalıştırılan senaryolar arasından hem doğruluk hem de F1-skoru açısından en optimal sınıflandırma performansının ölçüldüğü örnekleme tekniği-model kombinasyonu, senaryo 6'daki rastgele yukarı örnekleme (ROS)-lojistik regresyon (LR) kombinasyonudur; yani değişken sayısı 42'den 27'ye düşürülmüş ve ROS yukarı örnekleme tekniğiyle dengelenmiş Avustralya kredi veri setiyle çalıştırılan temel modeller arasından en optimal sınıflandırma performansını gösteren model, lojistik regresyon (LR)'dur.

HMEQ kredi veri setiyle çalıştırılan senaryolar arasından en optimal sınıflandırma performansı ölçüldüğü örnekleme tekniği-model kombinasyonu, senaryo 4'teki cluster centroids-k en yakın komşular modeli (KNN) kombinasyonudur; yani değişken sayısı 18 olan ve cluster centroids aşağı örnekleme tekniğiyle dengelenmiş HMEQ kredi veri setiyle çalıştırılan tekil modeller arasından en optimal sınıflandırma performansını gösteren model, k-en yakın komşular modeli (KNN)'dir.

Şekil 54: Senaryoların Veri Setlerine göre Ortalama F1-Skoru Sonuçları



Şekil 41’de yer alan sütun grafikleri; Alman, Avustralya ve HMEQ kredi veri setlerine uygulanan her bir senaryoya ait F1-skoru değerlerinin genel ortalamalarını göstermektedir. Sütun grafiklerinin oluşturulmasındaki amaç; çalışmada kullanılan makine öğrenmesi tekniklerinin tamamından elde edilen F1-skoru değerlerinin genel ortalamasına bakarak, sekiz farklı senaryonun performansı hakkında çıkarım yapmaktır.

Sütun grafiklerini detaylı incelediğimizde ise, Alman kredi veri setini sınıflandırmada en iyi performansı senaryo 5 göstermiştir; yani değişken seçim tekniği uygulanan ve yukarı örnekleme tekniğiyle dengelenen Alman kredi veri setinde, ortalama olarak en iyi sınıflandırma performansı gösteren model, topluluk öğrenmesi modelleri olduğu sonucuna ulaşılabilir.

Avustralya ve HMEQ kredi veri setlerini sınıflandırmada ise en iyi performansı, senaryo 7 göstermiştir; yani değişken seçim tekniği uygulanmayan ve yukarı örnekleme tekniğiyle dengelenen veri setlerinde, en yüksek sınıflandırma performansı topluluk öğrenmesi modellerinde elde edildiği sonucuna ulaşılabilir.

Genel bir çıkarımda bulunmak gerekirse, her ne kadar Alman kredi veri setini sınıflandırmada en iyi performansı senaryo 7 göstermemiş olsa da seçilen senaryo 5 ve senaryo 7 birbirlerine benzer özellikleri barındırmaktadır. Bu iki senaryo, hem veri setlerini dengelerken yukarı örnekleme tekniklerini hem de ilgili veri setlerini sınıflandırırken topluluk öğrenmesi modellerini kullanır. Ayrıca Alman kredi veri setine değişken seçim tekniğinin uygulanması, topluluk öğrenmesi modellerinin sınıflandırma performansı üzerinde az da olsa olumlu bir etki yaratırken; Avustralya ve HMEQ kredi veri setlerine değişken seçim tekniği uygulanmaması ise, topluluk öğrenmesi modellerinin sınıflandırma performansı üzerinde olumlu bir etki yarattığını söylemek mümkündür. Sonuç olarak; yukarı örnekleme teknikleriyle dengelenen kredi veri setlerinin sınıflandırılmasında topluluk öğrenmesi teknikleri genel olarak daha iyi performans gösterdiği söylenebilir.

SONUÇ

Bu çalışma kapsamında, bankalara kredi talebinde bulunan potansiyel müşterilerden istenen demografik (yaş, meslek vb.) ve ekonomik (maaş vb.) bilgiler ışığında, kredilendirme faaliyetlerini otomatik olarak gerçekleştirebilecek ve aynı zamanda kredi riskini azaltarak karlılığı arttıracak bir karar mekanizmasının oluşturulması hedeflenmiştir. Hedeflenen karar mekanizmasını oluşturmak için, çağımızın en popüler bilimlerinden biri olan makine öğrenmesine ait sınıflandırma tekniklerinden faydalanılmıştır. Bankalar, oluşturulan karar mekanizmalarında kullandığı makine öğrenmesi tekniklerini geçmişte müşterileri için aldığı kararlarla eğiterek, gelecekte kredi talebinde bulunan müşterilerin kredi riski taşıyıp taşımadığına; yani kredi almaya uygun olup olmadığına otomatik olarak karar verecektir.

Bir karar mekanizması geliştirmek kadar, bu mekanizmayı performans açısından iyileştirmek de oldukça önemlidir. Çünkü, mekanizmanın vereceği kararlar ne kadar etkin ve doğru olursa, bankacılık sektöründe elde edilecek kâr o kadar fazla olacaktır. Bunun yanı sıra, oluşturulacak mekanizmanın performans olarak iyileştirilmesi, bankaların kredi operasyonlarına hem zaman hem de maliyet açısından da büyük avantaj sağlayacaktır. Ancak bazı problemler için optimal bir çözüm elde etmek güçtür. Çözümü güç olan problemler ise genellikle gerçek hayat problemleridir. Gerçek hayat problemlerine ait verilerin nerdeyse tamamı, makine öğrenmesi tekniğinin kolayca öğrenebileceği veya anlamlı yapılar çıkarabileceği bir yapıya sahip değildir. Bu nedenle ilgili veri setlerini, farklı teknik veya araçlarla beraber makine öğrenmesi tekniklerinin kolaylıkla öğrenebileceği bir forma dönüştürmek gerekir. Bu noktada, çalışmanın bankacılık sektörü için önemini yanı sıra, geliştirilecek olan karar mekanizmasının sahip olduğu teknik yapının işleyişinden de detaylıca bahsedilmiştir.

Yapılan çalışmada, bankacılık sektörü için geliştirilecek olan karar verme mekanizmasının etkinliğini ölçmek için gerçek hayat problemi olan 3 farklı kredi veri seti kullanılmıştır. Kullanılan kredi veri setleri; Alman, Avustralya ve HMEQ kredi verileri olup, özellikle literatürdeki kredi risk analizi üzerine yapılan çalışmalarının çoğunda kullanılan açık kaynaklı verilerdir.

Uygulama kapsamında kullanılacak olan kredi veri setlerine ait değişkenler hem sayısal hem de kategorik ve sıralı veri tiplerine sahiptir. Ancak kategorik ve sıralı veri tipine sahip değişkenlerin, özellikle değişkenin sahip olduğu veriler metin türündeysen, bu metin türündeki verileri sayısal değerlerle kodlamak gerekir. Sayısal değerlere kodlamamızın sebebi ise çalışma kapsamında kullanacağımız çoğu makine öğrenmesi tekniğinin metin türündeki verileri temelinde bulunan matematiksel fonksiyonlar sebebiyle işleyememesidir. Sıralı veri tipine sahip değişkende yer alan veriler kodlanırken, kodlanan her veri için yeni bir değişken oluşturulmuş ve kodlanan veri ilgili değişkene atanmıştır. Bu durum değişken sayısının artışına neden olmuştur. Orijinal halinde girdi değişken sayısı değişken sayısı 20 olan Alman kredi veri setinin kodlamadan sonraki değişken sayısı 50 olmuştur. Orijinal halinde girdi değişken sayısı 14 olan Avustralya kredi veri setinin kodlamadan sonraki değişken sayısı 42 olmuştur. Son olarak orijinal halinde girdi değişken sayısı 12 olan HMEQ kredi veri setinin kodlamadan sonraki değişken sayısı ise 18 olmuştur.

Karar mekanizması genel olarak üç ana aşamadan oluşmaktadır. Bu aşamaların iş akış sırası, oluşturulan sekiz senaryo için farklılık gösterebilmektedir. İlk aşama, değişken seçimidir. Bu aşama özellikle değişken seçimi yapılan senaryolar için önem arz ederken, değişken seçim yapılmayan senaryolarda ise direkt ikinci aşamaya geçilir. Bu bağlamda, gereksiz veya problemin çözümüyle ilişkisi daha az olan değişkenlerin, değişken seçim tekniği ile tespit edilmesinin akabinde bu değişkenlerin göz ardı edilmesi veya değişkenlerin tamamının uygulamaya dahil etmenin, karar mekanizmasının performansı üzerinde olumlu bir etki yaratıp yaratmadığı incelenmiştir. Değişken seçimi uygulanan senaryolarda değişken seçim tekniği olarak destek vektör makinesi–yinelemeli değişken eleme tekniği kullanılmıştır.

İkinci aşamada, çalışma kapsamında kullanılacak olan dengeli olmayan kredi verileri dengeli hale getirilecektir. Bankacılık ve finans sektörlerine ait problemler de gerçek hayat problemleridir. Gerçek hayat problemlerine ait veri setinde karşılaşılabilecek en büyük problem biri de sınıf dengesizliğidir. Sınıf dengesizliğine sahip veri setinin makine öğrenmesi teknikleri tarafından öğrenilmesi oldukça zordur. Bu durum, makine öğrenmesi tekniklerinin tahmin performansında düşüşe sebep olur. Bu nedenle bankacılık sektörüne ait veri setlerinde yaygın olan sınıf dengesizliği problemi müşterinin riskli olup olmadığının tespitinin sağlıklı yapılabilmesi için giderilmesi

gerekmektedir. Bu aşamada; Alman, Avustralya ve HMEQ kredi veri setlerindeki sınıf dengesizliğinin giderilmesi için yukarı ve aşağı olmak üzere iki tip yeniden örnekleme tekniği kullanılmıştır. Yukarı örnekleme tekniklerinin kullanıldığı senaryolarda rastgele yukarı örnekleme (ROS), SMOTE, SVMSMOTE, ADASYN, Borderline-SMOTE olmak üzere 5 farklı örnekleme tekniği kullanılırken, aşağı örnekleme tekniklerinin kullanıldığı senaryolarda ise; near miss, Tomek-links, rastgele aşağı örnekleme (RUS), cluster-centroids olmak üzere 4 farklı örnekleme tekniği kullanılmıştır.

Üçüncü aşama ve son aşama ise, ilk iki aşamada kredi verileri üzerinde gerçekleştirilen faaliyetlerin, çalışma kapsamında kullanılacak makine öğrenmesi tekniklerinin performansına olan etkisinin incelendiği aşamadır. Bu aşamada, ilk iki aşamada revize edilen Alman, Avustralya ve HMEQ kredi veri setlerinin her biriyle ayrı ayrı eğitilen ve test edilen tekil ve topluluk öğrenmesi modellerinin sınıflandırma performansı ölçülmüştür. Bu aşamada, toplamda 11 adet makine öğrenmesi tekniği kullanılmıştır. Bunlardan 6'sı tekil model iken 5'i ise topluluk öğrenmesi modelleridir. Çalışmada kullanılan tekil modeller; Naive Bayes, k-en yakın komşu, çok katmanlı algılayıcı sinir ağları, karar ağaçları, destek vektör makineleri ve lojistik regresyondur. Çalışmada kullanılan topluluk öğrenmesi modelleri ise; rastgele ormanlar, gradyan artırma karar ağaçları, ekstra ağaçlar, yumuşak oylama ve sert oylamadır.

Çalışma kapsamında kredi riski sınıflandırma problemlerinin çözümüne en uygun karar verme mekanizmasının tespiti için 8 farklı uygulama senaryosu oluşturulmuştur. Revize edilen; Alman, Avustralya ve HMEQ kredi verileriyle çalıştırılan her bir senaryodan elde edilen optimal sonuçlar aşağıda verilmiştir. Uygulamadan elde edilen sonuçların daha anlaşılabilir olması açısından uygulanacak olan her bir yeniden örnekleme tekniği ile makine öğrenmesi tekniği, kombinasyon haline getirilmiş ve yorumlanırken de kombinasyon kalıbı kullanılmıştır.

İlk senaryoda, değişken seçim tekniği uygulanmış ve aşağı örnekleme teknikleriyle dengelenmiş Alman, Avustralya ve HMEQ kredi veri setleriyle eğitilen ve test edilen topluluk öğrenmesi tekniklerinin sınıflandırma performansı ölçülmüştür. Değişken sayısı 50'den 37'ye düşürülen Alman kredi veri setiyle çalıştırılan kombinasyonlar arasından en iyi sınıflandırma performansına sahip kombinasyon Tomek links-ekstra ağaçlar kombinasyonudur. Değişken sayısı 42'den 27'ye

düşürülen Avustralya kredi veri setiyle çalıştırılan kombinasyonlar arasından en iyi sınıflandırma performansına sahip kombinasyon ise rastgele aşağı örnekleme tekniği-rastgele ormanlar kombinasyonudur. Değişken sayısı 18'den 13'e düşürülen HMEQ kredi veri setiyle çalıştırılan kombinasyonların arasından en iyi sınıflandırma performansına sahip kombinasyon ise cluster centroids-ekstra ağaçlar kombinasyonudur.

İkinci senaryoda, değişken seçim tekniği uygulanmış ve aşağı örnekleme teknikleriyle dengelenmiş Alman, Avustralya ve HMEQ kredi veri setleriyle eğitilen ve test edilen tekil modellerin sınıflandırma performansı ölçülmüştür. Değişken sayısı 50'den 37'ye düşürülen Alman kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından, en iyi sınıflandırma performansına sahip kombinasyon Tomek links-lojistik regresyon kombinasyonudur. Değişken sayısı 42'den 27'ye düşürülen Avustralya kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına sahip kombinasyon ise hem Tomek links-lojistik regresyon hem de Tomek links-destek vektör makineleri kombinasyonlarıdır. Değişken sayısı 18'den 13'e düşürülen HMEQ kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına sahip kombinasyon ise cluster centroids-k en yakın komşular kombinasyonudur.

Üçüncü senaryoda, değişken seçim tekniği uygulanmamış ve aşağı örnekleme teknikleriyle dengelenmiş Alman, Avustralya ve HMEQ kredi veri setleriyle eğitilen ve test edilen topluluk öğrenmesi tekniklerinin sınıflandırma performansı ölçülmüştür. Değişken sayısı 50 olan Alman kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına sahip kombinasyon Tomek links-yumuşak (soft) oylamadır. Değişken sayısı 42 olan Avustralya kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına sahip kombinasyon ise rastgele aşağı örnekleme tekniği-rastgele ormanlar kombinasyonudur. Değişken sayısı 18 olan HMEQ kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansı sahip kombinasyon ise cluster centroids-ekstra ağaçlar kombinasyonudur.

Dördüncü senaryoda, değişken seçim tekniği uygulanmamış ve aşağı örnekleme teknikleriyle dengelenmiş Alman, Avustralya ve HMEQ kredi veri

setleriyle eğitilen ve test edilen tekil modellerin sınıflandırma performansı ölçülmüştür. Değişken sayısı 50 olan Alman kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasında en iyi sınıflandırma performansına sahip kombinasyon rastgele aşağı örnekleme-Naive Bayes kombinasyonudur. Değişken sayısı 42 olan Avustralya kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasında en iyi sınıflandırma performansına sahip kombinasyon rastgele aşağı örnekleme destek vektör makineleri kombinasyonudur. Değişken sayısı 18 olan HMEQ kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasında en iyi sınıflandırma performansına sahip kombinasyon cluster centroids-k en yakın komşular algoritması kombinasyonudur.

Beşinci senaryoda, değişken seçim tekniği uygulanmış ve yukarı örnekleme teknikleriyle dengelenmiş Alman, Avustralya ve HMEQ kredi veri setleriyle eğitilen ve test edilen topluluk öğrenmesi tekniklerinin sınıflandırma performansı ölçülmüştür. Değişken sayısı 50'den 37'ye düşürülen Alman kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasında en iyi sınıflandırma performansına sahip kombinasyon rastgele yukarı örnekleme-rastgele ormanlar kombinasyonudur. Değişken sayısı 42'den 27'ye düşürülen Avustralya kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasında en iyi performansa sahip kombinasyon ise rastgele yukarı örnekleme-gradyan artırma karar ağaçları kombinasyonudur. Değişken sayısı 18'den 13'e düşen HMEQ kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasında en iyi sınıflandırma performansına sahip kombinasyon ise BorderlineSMOTE-ekstra ağaçlar kombinasyonudur.

Altıncı senaryoda, değişken seçim tekniği uygulanmış ve yukarı örnekleme teknikleriyle dengelenmiş Alman, Avustralya ve HMEQ kredi veri setleriyle eğitilen ve test edilen tekil modellerin sınıflandırma performansı ölçülmüştür. Değişken sayısı 50'den 37'ye düşürülen Alman kredi veri setiyle çalıştırılan kombinasyonlar arasında örnekleme tekniği-model kombinasyonları arasında en iyi sınıflandırma performansa sahip kombinasyon rastgele yukarı örnekleme tekniği-lojistik regresyon kombinasyonu olarak belirlenmiştir. Değişken sayısı 42'den 27'ye düşürülen Avustralya kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasında en iyi sınıflandırma performansına sahip kombinasyon ise rastgele yukarı

örnekleme tekniği-destek vektör makineleri kombinasyonudur. Değişken sayısı 18'den 13'e düşürülen HMEQ kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına sahip kombinasyon ise BorderlineSMOTE-k en yakın komşular algoritması kombinasyonudur.

Yedinci senaryoda, değişken seçim tekniği uygulanmamış ve yukarı örnekleme teknikleriyle dengelenmiş Alman, Avustralya ve HMEQ kredi veri setleriyle eğitilen ve test edilen topluluk öğrenmesi tekniklerinin sınıflandırma performansı ölçülmüştür. Değişken sayısı 50 olan Alman kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına sahip kombinasyon rastgele yukarı örnekleme-gradyan artırma karar ağaçları kombinasyonudur. Değişken sayısı 42 olan Avustralya kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına sahip kombinasyon SMOTE-GBDT kombinasyonudur. Değişken sayısı 18 olan HMEQ kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonlar arasından en iyi sınıflandırma performansına sahip kombinasyon ise ADASYN-ekstra ağaçlar kombinasyonudur.

Sekizinci ve son senaryoda, değişken seçim tekniği uygulanmamış ve yukarı örnekleme teknikleriyle dengelenmiş Alman, Avustralya ve HMEQ kredi veri setleriyle eğitilen ve test edilen topluluk öğrenmesi tekniklerinin sınıflandırma performansı ölçülmüştür. Değişken sayısı 50 olan Alman kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına sahip kombinasyon SVMSMOTE-lojistik regresyon kombinasyonudur. Değişken sayısı 42 olan Avustralya kredi veri setiyle çalıştırılan örnekleme tekniği-model kombinasyonları arasından en iyi sınıflandırma performansına kombinasyon ise rastgele yukarı örnekleme tekniği-lojistik regresyon kombinasyonudur. Değişken sayısı 18 olan HMEQ kredi veri setiyle çalıştırılan kombinasyonlar arasından en iyi sınıflandırma performansına sahip kombinasyon ise ADASYN-k en yakın komşular kombinasyonudur.

Çalışma kapsamında kullanılan tüm makine öğrenmesi teknikleri arasından en düşük sınıflandırma performansına sahip yöntem Naive Bayes olarak tespit edilmiştir. Li ve Chen (2020) kredi skorlama üzerine yaptıkları çalışmada Naive Bayes yönteminin sınıflandırma problemlerinde zayıf kalmasının sebebini, tahmin edilen

sonucun bir önceki olasılığa dayanıyor olabilme ihtimaline bağlamışlardır. Bu nedenle, kurulmuş olan önceki modelin nihai sonucun üzerinde etkisinin olabileceğini düşünmüşlerdir (Li, Chen; 2020: 13). Ayrıca HMEQ kredi veri seti aracılığıyla elde edilen sonuçlar, diğer kredi veri setlerinden elde edilen sonuçlara göre önemli ölçüde daha iyi çıkmıştır (HMEQ>Avustralya>Alman). Bunun yanı sıra, Alman kredi veri setiyle çalıştırılan senaryoların içerdiği kombinasyonlar arasından en optimal sonucun elde edildiği kombinasyon, değişken seçim tekniğinin uygulanmadığı senaryo 7’de yer alan ROS-GBDT kombinasyonudur. Avustralya kredi veri setiyle çalıştırılan senaryoların içerdiği kombinasyonlar arasından en optimal sonucun elde edildiği kombinasyon, değişken seçim tekniğinin uygulandığı senaryo 6’da yer alan ROS-SVM kombinasyonudur. HMEQ kredi veri setiyle çalıştırılan senaryoların içerdiği kombinasyonlar arasından en optimal sonucun elde edildiği kombinasyon, değişken seçim tekniğinin uygulanmadığı senaryo 4’te yer alan cluster centroids-KNN kombinasyonudur.

Sonuç olarak çalışmada kullanılan tüm makine öğrenmesi teknikleri arasından en iyi sınıflandırma performansına sahip modeller, ağaç tabanlı topluluk öğrenmesi modelleri olarak tespit edilmiştir. Bu doğrultuda, özellikle bankacılık sektörünün müzdarip olduğu kredi riski sorununa karşılık olarak geliştirilecek otomatik karar verme mekanizmalarını yukarı örnekleme tekniği-ağaç tabanlı modellerle oluşturmak, bankaların kredilendirme süreçlerinde hem doğru hem de etkin karar almalarına olanak tanıyacaktır. Bu sayede oluşabilecek kredi riskini minimize edilebilecek ve buna paralel olarak bankanın karı da artacaktır. Buna ek olarak probleme ait veri setlerini değişken seçim tekniği uygulamadan geliştirilen karar mekanizmasında çalıştırmak, mekanizmanın sınıflandırma performansını artırdığını söylemek mümkündür.

Çalışmanın geliştirilmesi amacıyla, uygulama kısmında iyileştirmeler yapılabilir. Özellikle oluşturulan karar mekanizmalarının işlevselliğini arttırmak adına farklı teknikler bu çalışmanın kapsamına dahil edilebilir ve çalışmadaki sınıflandırma modellerinin performansı üzerindeki etkileri de detaylıca araştırılabilir. Bu bağlamda, önerilecek teknikler ve faaliyetler, bankacılık sektöründe meydana gelen aynı tip problemlerin çözümlerinin iyileştirilmesinde de kullanılabilir. Çalışmada oluşturulan mekanizmanın geliştirilmesi için önerilen teknikler ise aşağıdaki gibidir.

- Çalışma kapsamında kullanılan Destek Vektör Makinesi–Yinelemeli Değişken Eleme tekniğini kullanmak yerine; lasso, elasticnet gibi farklı türde değişken seçim teknikleri kullanılabilir. Farklı değişken seçim tekniklerinin kullanıldığı veri setleriyle çalışma kapsamında oluşturulmuş senaryolarla tekrar çalıştırılabilir ve kullanılan yeni değişken seçim tekniklerinin sınıflandırma performansına etkisinin olup olmadığı incelenebilir.
- Çalışmada kapsamında sınıf dengesizliğini gidermek için kullanılan aşağı ve yukarı yeniden örnekleme tekniklerine ilave olarak SMOTE-Tomek links, SMOTE-ENN gibi yukarı ve aşağı örnekleme tekniklerinin birleşimi olan hibrit yaklaşımlar kullanılabilir. Bu sayede kullanılan hibrit yaklaşımlarla dengelenen veri setleriyle çalıştırılan senaryoların sınıflandırma performansının, diğer yukarı ve aşağı örnekleme teknikleriyle dengelenmiş veri setleriyle çalıştırılan senaryoların performansına göre daha iyi olup olmadığı araştırılabilir.
- Çalışmada kullanılan topluluk öğrenmesi modellerine ilave olarak; XGBoost, LightGBM, CatBoost gibi farklı türdeki topluluk öğrenmesi modelleri de çalışma kapsamına dahil edilebilir. Çalışmada oluşturulan örnekleme tekniği-model kombinasyonları, bu topluluk öğrenmesi modelleri içinde oluşturulabilir ve ilgili kredi verileriyle çalıştırılabilir. Böylece sonradan eklenen bu üç model ile geriye kalan tüm modellerin sınıflandırma performansları karşılaştırılabilir. Çalışmaya dahil edilen bu üç topluluk öğrenmesi tekniği, çalışmada uygulanan değişken seçim tekniği dışında farklı bir değişken seçim tekniği uygulanmış ve hibrit yaklaşımlarla dengelenmiş veri setleriyle çalıştırıldığında sınıflandırma performansının diğer topluluk ve tekil modellere göre daha iyi bir sınıflandırma performansı gösterip göstermediği de bu çalışma kapsamında araştırılabilir.

KAYNAKÇA

Abellán, J. ve Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*. 73: 1–10.

AICPA. (2016). *Audit and Accounting Guide Depository and Lending Institutions: Banks and Savings Institutions, Credit Unions, Finance Companies, and Mortgage Companies*. New York, USA: Wiley.

Akgüç, Ö. (2010). *Kredi Taleplerinin Değerlendirilmesi* (7. Baskı.). İstanbul: Arayış Basım ve Yayıncılık.

Akman, M., Genç, Y. ve Ankaralı, H. (2011). Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama. *Türkiye Klinikleri Journal of Biostatistics*. 3(1): 36–48.

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, Khushi, M. (2020). An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. *IEEE Access*. 8(November, 2020): 201173–201198.

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*. 23(4): 589–609.

Apostolik, R., Went, P. ve Donohuse, C. (2009). *Foundations of Banking Risk: An Overview of Banking, Banking Risks, and Risk-Based Banking Regulation* (1th ed.). Hoboken, New Jersey.: JohnWiley & Sons, Inc.

Arain, F. N.(04.04.2021). *Decision Tree Classification Algorithm*. <https://www.devops.ae/decision-tree-classification-algorithm/> (Erişim Tarihi: 25.04.2021)

Aras, G. (1996). *Ticari Bankalarda Kredi Portföyünün Yönetimi* (1. Baskı.). Ankara: SPK Yayın No:30.

Arman, T. ve Gürman, T. (1997). *Bankalarda Finansal Yönetime Giriş*. İstanbul: Türkiye Bankalar Birliği Yayını.

Ayadi, R. (2019). *Banking Business Models*. Palgrave Macmillan Studies in Banking and Financial Institutions (1th Ed.). Cham, Switzerland: Springer International Publishing.

Babuşçu, Ş. (1997). *Bankacılıkta Risk Derecelendirmesi(Rating) ve Türk Bankacılık Sektörüne Uygulanması*. Ankara: SPK Yayın No:94.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. ve Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*. 54(6): 627–635.

Bandyopadhyay, A. (2016). *Managing Portfolio Credit Risk in Banks*. (1st Ed.). Delhi, India: Cambridge University Press.

Başar, M. (2006). Ticari Bankaların Fon Kullanımı. *Bankacılık Uygulamaları* (1. Baskı.). (ss. 121–156). Eskişehir: T.C. Anadolu Üniversitesi Yayınları.

BDDK. (2006). Bankaların İç Sistemleri ve İçsel Sermaye Yeterliliği Değerlendirme Süreci Hakkında Yönetmelik. <https://www.mevzuat.gov.tr/File/GeneratePdf?mevzuatNo=19864&mevzuatTur=KurumVeKurulusYonetmeligi&mevzuatTertip=5>. (Erişim Tarihi: 10.01.2021).

BDDK. (08.02.2021a). Türk Bankacılık Sektörü Temel Göstergeleri Aralık 2020. <https://www.bddk.org.tr/Duyuru/EkGetir/733?ekId=757>.(Erişim Tarihi: 15.03.2021).

BDDK. (10.06.2021b). Bankacılık Düzenleme ve Denetleme Kurumu Yıllık Faaliyet Raporu 2020. <https://www.bddk.org.tr/Duyuru/EkGetir/839?ekId=771>.(Erişim Tarihi: 29.06.2021).

Benzschawel, T. (2012). *Credit Risk Modelling: Facts, Theory and Applications*. London, UK: Risk Books, a Division of Incisive Media Investments Ltd.

Berk, N. (2014). Kısa Vadeli Finansman. *Finansal Yönetim-I* (3. Baskı.), (ss. 192–218). Eskişehir: T.C. Anadolu Üniversitesi Yayınları.

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

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

Breiman, L. (2001). Random Forests. *Kluwer Academic Publishers*. 45: 5–32.

Britannica, T. E. of E. (14.06.2012). Finance company. *Encyclopedia Britannica*. <https://www.britannica.com/topic/finance-company>.(Eriřim Tarihi: 15.12.2020).

Brown, K. ve Moles, P. (2016). *Credit Risk Management* (4th Ed.). Edinburgh, UK: Edinburgh Business School.

Brownlee, J. (2016). *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End* (1st Ed.).

Büker, S., Ařıkoęlu, R. ve Sevil, G. (2014). *Finansal Yönetim* (8. Baskı.). Ankara: Sözkese Matbaacılık.

Çaęıran Tuncer, S. ve Ulusoy, Y. (2019). Kredi ve Kredi İşlemlerinin Hukuki Açıdan Deęerlendirilmesi. *Pamukkale Journal of Eurasian Socioeconomic Studies*. 6(2): 43–69.

Canbař, S. ve Doęukanlı, H. (2012). *Finansal Pazarlar* (5. Baskı.). Adana: Karahan Kitabevi.

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

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.

Cinemre, H. A. (2010). *Tarım İşletmecilięi ve Planlama* (2. Baskı.). Samsun: OMÜ Ziraat Fakóltesi Yayınları. <https://avys.omu.edu.tr/storage/app/public/esin.hazneci/126051/11>. Risk ve Belirsizlik Yönetimi.pdf.(Eriřim Tarihi: 03.01.2021).

Cortes, C. ve Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*. 20(3): 273–297.

Dangeti, P. (2017). *Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R* (1th Ed.). Birmingham, UK.: Packt Publishing.

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

DOMO. (2018). Data Never Sleeps 6.0. <https://web-assets.domo.com/blog/wp-content/uploads/2018/06/18-domo-data-never-sleeps-6.png>. (Erişim Tarihi: 22.02.2021).

DOMO. (2021). Data Never Sleeps 8.0. <https://web-assets.domo.com/blog/wp-content/uploads/2020/08/20-data-never-sleeps-8-final-01-Resize.jpg>. (Erişim Tarihi: 22.02.2021).

Durand, D. (1941). *Risk Elements in Consumer Instalment Financing*. (Tech. Ed.). New York, USA: National Bureau of Economic Research.

Erdaş, M. L. ve Eroğlu, A. (2018). Bazı Geri Ödemelerin Geometrik Değişimli Olduğu Yeni Bir Borç Ödeme Model Önerisi: Örnek Uygulamalar. *Sosyal, Beşeri Ve İdari Bilimler'de Akademik Araştırmalar- V* (ss. 147–160). Ankara: Gece Kitaplığı.

Erdem, E. (2008). *Para Banka ve Finansal Sistem* (2. Baskı.). Ankara: Detay Yayıncılık.

Fan, A. ve Palaniswami, M. (2000). Selecting Bankruptcy Predictors Using a Support Vector Machine Approach. *Proceedings of the International Joint Conference on Neural Networks*. 6: 354–359.

Fidan, M. M. (2011). Kobi Kredi Piyasasında Asimetrik Bilgi ve Ahlaki Tehlike: Laleli Örneği. *Maliye ve Finans Yazıları*. 1(90): 41–57.

Fitch. (13.07.2020). Non-Rating Action Commentary: US Default Rate to Hit 5.5%, Highest in 10 Years. <https://www.fitchratings.com/research/corporate-finance/us-default-rate-to-hit-5-5-highest-in-10-years-13-07-2020>. (Erişim Tarihi: 17.05.2021).

Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*. 121(2): 256–285.

- Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems* (1st Ed.). USA: O'Reilly Media, Inc.
- Giddens, A. (2002). *Runaway World: How Globalisation Is Reshaping Our Lives*. New York, USA: Routledge.
- Giudici, P. (2001). Bayesian data mining, with application to benchmarking and credit scoring. *Applied Stochastic Models in Business and Industry*. 17(1): 69–81.
- Golin, J. ve Delhaise, P. (2013). *The Bank Credit Analysis Handbook: A Guide for Analysts, Bankers and Investors* (2. Edition.). Singapur: John Wiley & Sons Singapore Pte. Ltd.
- Gorton, G. B. ve He, P. (2008). Bank credit cycles. *Review of Economic Studies*. 75(4): 1181–1214.
- Grajski, K. A., Breiman, L., Di Prisco, G. V. ve Freeman, W. J. (1986). Classification of EEG Spatial Patterns with a Tree-Structured Methodology: CART. *IEEE Transactions on Biomedical Engineering*. BME-33(12):1076–1086.
- Guyon, I., Weston, J., Barnhill, S. ve Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*. (46): 389–422
- Güney, A. (2007). *Banka İşlemleri* (1. Baskı.). İstanbul: Beta Basım Yayım.
- Gutierrez, D. D. (2015). *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R* (1st Ed.). NJ, USA: Technics Publications.
- Han, H., Wang, W. Y. ve Mao, B. H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *International Conference on Intelligent Computing 2005*(ss. 878–887). Hefei, China.
- He, H., Bai, Y. A., Garcia, E. ve Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)* (ss. 1322–1328).
- Heffernan, S. (2005). *Modern Banking* (1st Ed.). London, UK: John Wiley & Sons Ltd.

- Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.20(8): 832–844.
- Huang, C., Chen, M. ve Wang, C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*. 33(4): 847–856.
- Huang, Y., Hung, C. ve Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*. 7(4): 720–747.
- Huang, Z., Chen, H., Hsu, C., Chen, W. ve Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*. 37(4): 543–558.
- IIA. (2009). IIA position paper: The role of internal auditing in enterprise-wide risk management. *The Institute of Internal Auditors*. (January, 2009): 1–8.
- İnağ, N. (1990). *Tüketici Kredileri*. Ankara: T.C. Merkez Bankası Araştırma ve Eğitim Genel Müdürlüğü Tartışma Tebliği No:9011.
- Jurafsky, D. ve Martin, J. H. (30.12.2020). Speech and Language Processing. *Stanford University*.<https://web.stanford.edu/~jurafsky/slp3/5.pdf>. (Erişim Tarihi: 29.01.2021)
- Kavcıoğlu, Ş. (2003). *Ticari Bankalarda Sorunlu Kredilerin Yönetimi, Çözüm Yolları ve Takibi*. İstanbul: Türkmen Kitabevi.
- Kaygısız Ertuğ, Z. ve Girginer, N. (2015). Evaluation of banks' commercial credit applications using the analytic hierarchy process and grey relational analysis: A comparison between public and private banks. *South African Journal of Economic and Management Sciences*. 18(3): 308–324.
- Kılınç, N. ve Karaoğlu, N. (2014). Tüketim Toplumu ve Bireysel Krediler; Türkiye Üzerine Bir Değerlendirme. *EconWorld2014@Prague, International Conference in Economics*: 1–16. 3-5 Eylül 2014.
- Kumar, A. ve Jain, M. (2020). *Ensemble Learning for AI Developers* (1st Ed.). India: Apress Media LLC.

Kubat, M. ve Matwin, S. (1997). Addressing the curse of imbalanced training sets: one sided selection. *14th International Conference on Machine Learning* (ss. 179–186). Nashville, USA.

Küçükbay, F. (2017), Banka karlılığını etkileyen faktörler: Avrupa Birliği bankaları ve Türk bankaları arasında bir karşılaştırma. *Yönetim ve Ekonomi*, 24(1), 137-149.

Lantz, B. (2015). *Machine Learning with R* (2nd Ed.). Birmingham, UK.: Packt Publishing.

Lee, T., Chiu, C., Lu, C. ve Chen, I. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*. 23(3): 245–254.

Li, Y. ve Chen, W. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics*. 8(10): 1756.

Machado, G., Mendoza, M. R. ve Corbellini, L. G. (2015). What Variables are Important in Predicting Bovine Viral Diarrhea Virus? A Random Forest Approach. *Veterinary Research*. 46(1): 85.

Marqués, A. I., García, V. ve Sánchez, J. S. (2012). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*, 39(12): 10916–10922.

Medium. (13.02.2019). *Math behind SVM(Support Vector Machine)*.
<https://medium.com/@ankitnitjsr13/math-behind-svm-support-vector-machine-864e58977fdb> (Erişim Tarihi:07.02.2021)

Medium. (25.08.2020). *Nedir Bu Destek Vektör Makineleri? (Makine Öğrenmesi Serisi-2)*.
<https://medium.com/deep-learning-turkiye/nedir-bu-destek-vekt%C3%B6r-makineleri-makine-%C3%B6%C4%9Frenmesi-serisi-2-94e576e4223e> (Erişim Tarihi: 09.02.2021)

Mishkin, F. S. ve Eakins, S. G. (2015). Finance Companies(Chapter on The Web). *Financial Markets and Institutions*(8th Ed.), (ss. W48–W59). Courier/Kendallville, USA: Pearson Education Limited.
https://wps.pearsoned.co.uk/wps/media/objects/15998/16382113/web_chap/M27_MI_SH3624_08_SE_Ch27_W-48-W59.pdf .(Erişim Tarihi: 20 Aralık 2020).

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

Namvar, A., Siami, M., Rabhi, F. ve Naderpour, M. (2018). Credit risk prediction in an imbalanced social lending environment. *International Journal of Computational Intelligence Systems*. 11(1): 925.

Nguyen, H. M., Cooper, E. W. ve Kamei K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*. 3(1): 4–21.

Özbilgin, İ. G. (Temmuz 2012). Risk ve Risk Çeşitleri. *Bilişim Kültür Dergisi*, 87–93. <http://www.bilisimdergisi.org/s145>.(Erişim Tarihi: 02.01.2021).

Özün, A. ve Çifter, A. (2006). Bankaların Hisse Senedi Getirilerinde Faiz Oranı Riski: Dalgacıklar Analizi ile Türk Bankacılık Sektörü Üzerine Bir Uygulama. *Bankacılık Dergisi*. (59): 3–15.

Pamukçuk, E. B. (2009). *Kredi ve İhracat Kredi Sigortaları*. İstanbul: Avcıol Basım Yayım.

Raghuwanshi, B. S. ve Shukla, S. (2020). SMOTE Based Class-Specific Extreme Learning Machine for Imbalanced Learning. *Knowledge-Based Systems*. 187: 1–17.

Rose, P. S. ve Marquis, M. H. (2008). *Money and Capital Markets: Financial Institutions Instruments in the Global Marketplace* (10th Ed.). New York, USA: The McGraw-Hill Companies, Inc.

Rumelhart, D. E., Hinton, G. E. ve Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*. 323: 533-536.

Selimler, H. (2015). Sorunlu Kredilerin Analizi, Banka Finansal Tablo Ve Oranları Etkisini Değerlendirilmesi. *Finansal Araştırmalar ve Çalışmalar Dergisi*. 7(12): 131–172.

Sharma, A. (01.08.2016). *Contest 2nd Place: Automating Data Science*. <https://www.kdnuggets.com/2016/08/automating-data-science.html> (Erişim Tarihi : 05.05.2021).

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: 1–17.

Shim, J. K. ve Constat, M. (2001). *Encyclopedic Dictionary of Finance and Banking* (1. Baskı.). Boca Raton, Florida: CRC Press.

Şakar, B. (2011). *Banka Kredileri ve Yönetimi* (3. Baskı.). İstanbul: Beta Basım Yayım.

Şenel, C. (2020). Banka Kredileri ve Türk Bankacılık Sektöründe Kredilerin Uluslararası Karşılaştırmalı Analizi (Bank Loans and International Comparative Analysis of Loans in the Turkish Banking Sector). *Journal of Business Research - Turk*. 12(1): 810–825.

T.C. Ticaret Bakanlığı. (01.08.2019). Esnaf ve Sanatkarlar Kredi Kefalet Kooperatifleri.<https://ticaret.gov.tr/kooperatifcilik/sikca-sorulan-sorular/esnaf-ve-sanatkarlar-kredi-kefalet-kooperatifleri>.(Erişim Tarihi: 13 Aralık 2020).

Taş, C. (1998). *Dünya’da ve Türkiye’de Tüketici Finansman Şirketleri ve Uygulama Örnekleri*. (Yüksek Lisans Tezi). Marmara Üniversitesi Bankacılık ve Sigortacılık Enstitüsü.

TBB. (Mayıs 2020). Tüketici Kredileri ve Konut Kredileri. https://www.tbb.org.tr/Content/Upload/istatistikraporlar/ekler/1461/Tuketici_Kredileri_Raporu-Eylul_2020.pdf.(Erişim Tarihi: 20.01.2021)

Tekin, M. (2015). *Girişimcilik ve Küçük İşletme Yöneticiliği* (12. Baskı.). Konya: Günay Ofset.

Terzioğlu, K. (2007). *AB’ye Uyum Sürecinde Türkiye’de Bireysel Kredilendirme Politikasında Yapılan Değişiklikler ve Bu Değişikliklerin Hizmetiçi Eğitimi*.(Yayınlanmamış Yüksek Lisans Tezi). Ankara: Gazi Üniversitesi Eğitim Bilimleri Enstitüsü.

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

- Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems Man and Communications*. 6: 769-772.
- Torun, T. (2013). Kredi Değerlendirme Süreci, Mali Tahlil ve Kredi Hesaplamaları. *Bankalarda Kredi Yönetimi* (2. Baskı.), (ss. 50–97). Eskişehir: T.C. Anadolu Üniversitesi Yayınları.
- Ulusoy, T. (2012a). Kredi Türleri. *Bankacılık: Giriş ve İlkeler* içinde (1. Baskı.). (ss. 261–269). İstanbul: Beta Basım Yayım.
- Ulusoy, T. (2012b). Bankacılıkta Risk Yönetimi. *Bankacılık: Giriş ve İlkeler* (1. Baskı.). (ss. 357–369). İstanbul: Beta Basım Yayım.
- Usta, M. (1996). *Temel Kredi Bilgileri*. İstanbul: Pamukbank T.A.S. Eğitim Yayınları.
- Van Grinsven, J. (2009). *Improving operational risk management* (2nd Ed.). Amsterdam, The Netherlands: IOS Press BV.
- Vinod Kumar, L., Natarajan, S., Keerthana, S., Chinmayi, K. M. ve Lakshmi, N. (2016). Credit Risk Analysis in Peer-to-Peer Lending System. *2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA)*: 193–196.
- Vurucu, M. ve Arı, M. U. (2014). *A' dan Z' ye Bankacılık Yasal Mevzuat - Ürün ve Hizmetler Uygulamalar* (1. Baskı.). Ankara: Seçkin Yayıncılık.
- Waad, B., Ghazi, B. M. ve Mohamed, L. (2013). A Three-Stage Feature Selection Using Quadratic Programming for Credit Scoring. *Applied Artificial Intelligence*. 27(8): 721–742.
- Wang, G., Hao, J., Ma, J. ve Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*. 38(1): 223–230.
- West, D., Dellana, S. ve Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & Operations Research*. 32(10): 2543–2559.

- Xia, Y., Liu, C., Da, B. ve Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*. 93: 182–199.
- Yang, J., Awan, A. J. ve Vall-Llosera, G. (2019). Support Vector Machines on Noisy Intermediate Scale Quantum Computers. *Department of Cloud Systems and Platforms, Ericsson Research*, (September, 2019).
- 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.
- Yavaş, M., Güran, A. ve Uysal, M. (2020). Covid-19 Veri Kümesinin SMOTE Tabanlı Örneklem Yöntemi Uygulanarak Sınıflandırılması. *European Journal of Science and Technology*. (Special Issue): 258–264.
- Yazıcı, M. (2010). *Bankacılıkta Kredi Tahsisi*. Ankara: Seçkin Yayıncılık.
- Yazıcı, M. (2018). *İşletmelerde Risk Yönetimi ve Karar Verme*. İstanbul: Beta Basım Yayım.
- Yazıcı, M. (2020). Finansal Piyasalarda Risk ve Risk Yönetimi. *Finansal Piyasalar ve Kurumlar* (ss. 189–212). Bursa: Ekin Basın Yayın Dağıtım.
- Yeh, I. ve Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*. 36(2): 2473–2480.
- Yen, S. J. ve Lee, Y. S. (2009). Cluster-Based Under-Sampling Approaches for Imbalanced Data Distributions. *Expert Systems with Applications*. 36(3): 5718–5727.
- Yu, L., Yue, W., Wang, S. ve Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications*. 37(2): 1351–1360.
- Yüksel, S. ve Özşarı, M. (2016). Türkiye’deki Bireysel Krediler ile Enflasyon ve Cari İşlemler Açığı Arasındaki Nedensellik İlişkisinin İncelenmesi. *EconWorld2016@ImperialCollege Proceedings* (ss. 1–13). London, UK.

Zaccone, G. (2016). *Getting Started with TensorFlow* (1st Ed.). Birmingham, UK.: Packt Publishing.

Zhang, J. P. ve Mani, I. (2003). KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. *Proceeding of International Conference on Machine Learning (ICML 2003)*. (August, 2003):1-7.

Zurada, J., Kunene, N. ve Guan, J. (2014). The Classification Performance of Multiple Methods and Datasets : Cases from the Loan Credit Scoring Domain. *Journal of International Technology and Information Management*. 23(1): 57-82.

