

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DENGESİZ KREDİ SKORLAMA
VERİ SETLERİNDE KOLEKTİF ÖĞRENME ALGORİTMALARININ
PERFORMANS DEĞERLENDİRMESİ

Nihan ANKARA

YÜKSEK LİSANS TEZİ

Matematik Mühendisliği Anabilim Dalı

Matematik Mühendisliği Programı

Danışman

Prof. Dr. Hülya ŞAHİNTÜRK

Temmuz, 2019

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
DENGESİZ KREDİ SKORLAMA
VERİ SETLERİNDE KOLEKTİF ÖĞRENME ALGORİTMALARININ
PERFORMANS DEĞERLENDİRMESİ

Nihan ANKARA tarafından hazırlanan tez çalışması 16.07.2019 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Matematik Mühendisliği Anabilim Dalı, Matematik Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Hülya ŞAHİNTÜRK

Yıldız Teknik Üniversitesi

Danışman

Jüri Üyeleri

Prof. Dr. Hülya ŞAHİNTÜRK, Danışman

Yıldız Teknik Üniversitesi

Dr. Öğr. Üyesi Nilgün GÜLER BAYAZIT

Yıldız Teknik Üniversitesi

Dr. Öğr. Üyesi Oğuzhan ÖZTAŞ

İstanbul Üniversitesi Cerrahpaşa

Danışmanım Prof. Dr. Hülya ŞAHİNTÜRK sorumluluğunda tarafımca hazırlanan Dengesiz Kredi Skorlama Veri Setlerinde Kolektif Öğrenme Algoritmalarının Performans Değerlendirmesi başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Nihan ANKARA



*Aileme
ve
emeđi geen tım yakınlarıma*

TEŐEKKÜR

Bu tez alıőmasında finansal kuruluşlar için oldukça önemli bir konu olan kredi skorlama teknikleri incelenmiştir. Kredi skorlama, kredi almak üzere başvuru yapan kiői veya kurumların kredi alma konusundaki yeterliliğini somut veriler üzerinden analiz ederek deęerlendiren ve kredilerin geri ödeme noktasındaki risklilik seviyesini ölçümlendiren bir puanlama sistemi olarak tanımlanmaktadır. Kredi skorlama için, kredi başvurusu yapan borçluların geçmiş verileri incelenmektedir. Bu veriler veri madencilięi teknikleri kullanılarak yorumlanmaktadır.

Tez alıőmasındaki konunun belirlenmesi, araştırılması ve alıőmanın hazırlanma sürecinin her aşamasında bilgilerini, tecrübelerini ve deęerli zamanlarını esirgemeyerek bana her fırsatta yardımcı olan ve başaracağıma dair heveslendirici yorumları olan deęerli hocam Sayın Hülya ŐAHİNTÜRK'e sonsuz teőekkürlerimi sunuyorum.

Yapmış olduęum bu alıőmayı bütün eğitim ve iş hayatımda desteklerini esirgemeyen sevgili aileme ithaf ediyorum.

Nihan ANKARA

İÇİNDEKİLER

KISALTMA LİSTESİ	VII
ŞEKİL LİSTESİ.....	IX
TABLO LİSTESİ	X
ÖZET	XI
ABSTRACT.....	XIII
1 Giriş.....	15
1.1 Literatür Özeti.....	15
1.2 Tezin Amacı	24
1.3 Hipotez.....	25
2 Kredi Skorlama.....	26
2.1 Kredi Skorlama Nedir?.....	26
2.2 Kredi Skorlama Ne Amaçla Yapılır?.....	27
3 Makine Öğrenme Yöntemleri.....	29
3.1 Kolektif Öğrenme Nedir?.....	29
3.2 Kolektif Öğrenme Yöntemleri.....	30
3.2.1 Bagging Yöntemi.....	31
3.2.2 Boosting Yöntemi	32
3.2.2.1 AdaBoost Algoritması	33
3.2.2.2 Gradient Boosting Algoritması	35
3.2.3 Stacking Yöntemi.....	36
3.3 Maliyet Duyarlı Öğrenme Nedir?	36
4 Deneylerin Uygulanması.....	37

4.1 Verilerin Tanıtımı.....	37
4.1.1 German Credit Veri Seti.....	37
4.1.2 Australian Veri Seti.....	38
4.2 Performans Ölçütleri.....	39
4.2.1 Alıcı İşletim Özellikleri Eğrisi Altında Kalan Alan.....	39
4.3 Deneylerin Uygulanması.....	41
4.4 Sonuçlar.....	42
4.4.1 Doğruluk Kriterine Göre	44
4.4.2 Alıcı İşletim Karakteristik Eğri Altında Kalan Alan Kriterine Göre	45
4.4.3 Doğru Sınıflandırma (RİSKLİ Müşteri ve Negatif) Kıyaslaması	45
5 Sonuç ve Öneriler	47
Kaynakça.....	49
Tezden Üretilmiş Yayınlar.....	52

KISALTMA LİSTESİ

AdaBoost	Adaptive Boosting
ANN	Artificial Neural Network
AUC	Area Under the Curve
Bagging	Bootstrap Aggregating
BS	Brier Skoru
Bstacking	Bagging ve Stacking birleşimi
CART	Classification and Regression Tree
CHAID	Chi-square Automatic Interaction Detector
CS	Credit Scoring
DCEID	Dynamic Classifier Ensemble Imbalanced Distribution
DCS	Dynamic Classifier Selection
DES	Dynamic Ensemble Selection
DN	Doğru Negatif
DP	Doğru Pozitif
DPO	Doğru Pozitif Oran
DT	Decision Tree
FS-HB	Feature Selection Base Hybrid Bagging
GA	Genetic Algorithm
GMC	Give Me Some Credit
IBM	International Business Machines
IGR	Information Gain Ratio
kNN	k-Nearest Neighborhood
KS	Kolmogorov–Smirnov İstatistiği

LOLIMOT	Locally Linear Model Tree Algorithm
LR	Logistic Regression
LSSVM	Least Square Support Vector Machines
LSVM	Linear Support Vector Machines
MLP	Multi Layer Perceptron
NB	Naive Bayes
NN	Neural Networks
PCA	Principal Component Analysis
PCC	Percentage Correctly Classified
PG	Partial Gini Index
RBF	Radial Basis Function
RF	Random Forest
RM	Relief Method
ROC	Receiver Operating Characteristic
RS	Random Subspace
SMOTE	Synthetic Minority Oversampling Technique
Subagging	Sub Aggregating
SVM	Support Vector Machines
WSAB	Weight-Selected Attribute Bagging
XGBoost	Extreme Gradient Boosting
YN	Yanlış Negatif
YP	Yanlış Pozitif
YPO	Yanlış Pozitif Oran
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

Şekil 3.1 Bagging Çalışma Adımları.....	33
Şekil 3.2 Boosting Çalışma Adımları.....	34
Şekil 3.3 AdaBoost Çalışma Adımları.....	35
Şekil 3.4 Gradient Boosting Çalışma Adımları.....	37
Şekil 4.1 Karışıklık Matrisi (Confusion Matrix).....	41
Şekil 4.2 DPO ve YPO Eşitlikleri.....	41
Şekil 4.3 ROC Eğrisi.....	42



TABLO LİSTESİ

Tablo 4.1 'German Credit' Veri Seti Tanıtımı.....	38
Tablo 4.2 'Australian' Kredi Veri Seti Tanıtımı.....	39
Tablo 4.3 'German Credit' Maliyet Duyarlı Öğrenme(-) Deney Sonuçları.....	43
Tablo 4.4 'Australian' Maliyet Duyarlı Öğrenme(-) Deney Sonuçları.....	44
Tablo 4.5 'German Credit' Maliyet Duyarlı Öğrenme(+) Deney Sonuçları.....	44
Tablo 4.6 'Australian' Maliyet Duyarlı Öğrenme(+) Deney Sonuçları.....	45
Tablo 4.7 'German Credit' Doğruluk Kriterine Göre Deney Sonuçları.....	45
Tablo 4.8 'Australian' Doğruluk Kriterine Göre Deney Sonuçları.....	45
Tablo 4.9 'German Credit' AUC Kriterine Göre Deney Sonuçları.....	46
Tablo 4.10 'Australian' AUC Kriterine Göre Deney Sonuçları.....	46
Tablo 4.11 'RİSKLİ' Müşteri Tahmin Oranı Deney Sonuçları.....	47
Tablo 4.12 'NEGATİF' Tahmin Oranı Deney Sonuçları.....	47

Dengesiz Kredi Skorlama Veri Setlerinde Kolektif Öğrenme Algoritmalarının Performans Değerlendirmesi

Nihan ANKARA

Matematik Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Hülya ŞAHİNTÜRK

Artan kredi taleplerinin doğru analiz edilmesi ve krediyi geri ödeyen ve geri ödemeyen müşteri tanımlamalarının yapılabilmesi için kredi skorlama (KS) yöntemleri oldukça önemli bir yere sahip olmuştur. Bu nedenle, müşteri verileri analiz edilerek bireysel veya tüzel kişilerin krediyi geri ödeme noktasındaki yeterliliğini ölçmek için KS sistemleri geliştirilmiştir. KS sistemlerinde müşterinin demografik bilgileri, geçmişteki kredi ilişkilerinde sergilediği davranışlar ve finansal bilgileri kullanılarak müşteriler riskli ve risksiz müşteri olarak sınıflandırılmaktadır. Ancak günümüzdeki artan kredi başvuru sayıları bu işlemin sistemsel hesaplanarak otomatik hale getirilmesi ihtiyacını ortaya çıkarmıştır. Bu yüzden, istatistiksel ve makine öğrenme yöntemleri kullanılarak birçok kredi skorlama modeli geliştirilmiştir. Güvenilir kredi skorlama modelleri oluşturulurken karşımıza çıkan en büyük problemlerden biri de sınıflara ait örneklerin dengesiz bir dağılıma sahip olmalarından dolayı, modellerin sınıflama başarılarının düşük olmasıdır. Bu çalışmada kolektif öğrenme yöntemlerinin sınıflara ait örneklerin dengesiz dağıldığı kredi veri kümeleri üzerindeki model performansları karşılaştırılarak hangi kolektif öğrenme yönteminin seçilebileceği tartışılmıştır.

Anahtar Kelimeler: Kredi skollama, kolektif öğrenme



Performance Evaluation of Ensemble Learning Algorithms on Unbalanced Credit Scoring Data Set

Nihan ANKARA

Department of Mathematic Engineering

Doctor of Philosophy Thesis

Advisor: Prof. Dr. Hülya ŞAHİNTÜRK

Credit scoring (CS) methods have been an important role for the analysis of the increasing credit demands correctly and for the identification of the customers who repay and non-repay the loan. Therefore, CS systems have been developed to analyse various customer data and to measure the adequacy of individual or legal persons at the point of repayment of the loan. In CS systems, customers are classified as 'bad' or 'good' classes by using customer demographic information, financial information and behaviours exhibited in past credit dealings. However, nowadays, an increasing number of loan applications have revealed the need to systematically automate this process. Therefore, many credit scoring models were developed using statistical and machine learning methods. One of the challenging problems arising from the creation of reliable credit scoring models is that the models have a low classifier performance because of imbalance class distributions. In this study, the model performances of ensemble learning methods on unbalanced credit datasets are discussed which ensemble learning method can be applicable.

Keywords: Credit scoring, ensemble learning



YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

1.1 Literatür Özeti

Kredi skorlama, kredi talebinde bulunan müşterilerin borcunu geri ödeme konusundaki risklilik seviyesini ölçümlemek için kullanılan modellerdir. Kredi skorlama modelleri günümüzde artan finansal taleplerin daha hızlı ve daha objektif değerlendirilmesini sağlamaktadır. Kredi skorlama modeli oluşturmak için kullanılan veriler geçmiş müşterilerde karşılaşılan tecrübelerle göre hazırlanan veri setleridir.

Yüksek performanslı kredi skorlama modelleri oluşturulması finansal kurumların riski minimize edip maksimum fayda sağlaması için oldukça büyük önem taşımaktadır.

Veri madenciliği teknikleri kullanılarak kredi skorlama modelleri oluşturulması veya oluşturulan modellerin performanslarının artırılması konusu literatürde oldukça önemli ve geniş bir yere sahiptir.

a. Kolektif öğrenme yöntemleri ile ilgili literatürde yapılan çalışmalar,

Lean Yu ve arkadaşları, çalışmalarında çok aşamalı destek vektör makineleri kolektif öğrenme yöntemi önermiştir. Modelde British kredi kartı veri seti kullanılmıştır. Deneysel sonuçlar, önerilen modelin performans iyileştirmesi sağladığını göstermiştir. [1]

Ligang Zhou ve arkadaşları, çalışmalarında kredi skorlaması için en küçük kareler destek vektör makinelerine (LSSVM) dayanan kolektif sınıflandırma modeli ele almışlardır. Deneyler, kolektif modellerin iyi ve sağlam olduğunu ve kredi risk analizi için umut verici çözümler sağlayabileceğini göstermiştir. Modelde German ve İngiltere'deki bir finans şirketi kredi veri setleri kullanılmıştır. [2]

Nan-Chen Hsieh ve arkadaşları, çalışmalarında bir kredi başvuru sahibinin, başlangıçta sağlanan bilgilerden iyi, kötü veya sınırda şeklinde öngörülüp

öngörülmediğine odaklanmışlardır. Bu aslında kredi puanlaması için bir sınıflandırma görevidir. Kredi puanlaması analizi için birçok sınıflayıcı algoritması bulunmaktadır. Daha etkili sınıflandırıcılar oluşturmak için, farklı sınıflandırıcıları bir kolektif sınıflandırıcıya birleştirmek için bazı diğer yaklaşımlar önerilmiştir. Bir kolektif sınıflandırıcıyı oluşturmada önemli bir husus, her bir sınıflandırıcıyı diğer sınıflandırıcılardan mümkün olduğunca farklı kılmaktır. Deneysel sonuçlardan, önerilen kolektif yaklaşımının klasik kolektif sınıflandırıcılarına göre önemli ölçüde daha iyi bir performans verdiği karar verilmiştir. Önerilen sistemin veri temelli doğası onu mevcut hibrit / kolektif kredi puanlama sistemlerinden ayırdığı belirtilmiştir. Deneylerde German kredi veri seti kullanılmıştır. [3]

Jianwu Li ve arkadaşları, çalışmalarında geliştirilmiş bir öznelik Bagging yöntemi, yani ağırlık seçim nitelikli Bagging (Weight-Selected Attribute Bagging(WSAB)) yöntemi önermişlerdir. Özneliklerin ağırlıkları ilk önce doğrusal destek vektör makinesi (Lineer Support Vector Machines(LSVM)) ve ana bileşen analizi (PCA) gibi öznelik değerlendirme yöntemleri kullanılarak hesaplanmıştır. Deneylerde, LSVM kullanılarak WSAB yöntemi, PCA kullanılarak WSAB yöntemi, Random seçilen özelliklerle Bagging yöntemi, Bagging ve AdaBoost yöntemleri karşılaştırılmıştır. Deney sonuçlarında, PCA-WSAB yönteminin doğruluk açısından daha iyi performans gösterdiği belirtilmiştir. Modelde German ve Australian kredi veri setleri kullanılmıştır. [14]

Jie Sun ve arkadaşları, çalışmalarında kredi skorlamasının dinamik modellemesini yapmak için Bagging-Destek Vektör Makineleri yöntemine dayanan artımlı destek vektör makinesi kolektif öğrenme yöntemini ele almışlardır. Önemli özelliklerin tanımlanması için B&B algoritması kullanılmıştır. Deneyde, yeni yöntem artan olmayan destek vektör makineleri toplamı geleneksel modeli ile karşılaştırılmıştır. Deney sonucunda, yeni modelin kredi skorlamasını kurumsal artan bilgiye göre sürekli ve dinamik olarak ayarlayabildiğini, bu da geleneksel modelden daha iyi değerlendirme yeteneğine sahip olduğunu göstermiştir. Modelde Çin kredi veri seti kullanılmıştır. [16]

Wu Hsu-Che ve arkadaşları, çalışmalarında daha doğru sınıflandırıcıların geliştirilmesi için, çoklu veri ön işleme ve denetlenen öğrenme tekniklerinden

oluşan iki aşamalı analiz benimsemiştir. Bu çalışma, kredi derecelendirme tahmini için dokuz sınıflı sınıflandırmayı ele almaktadır. Birinci aşamada, veriyi ön işleme yapmak için özellik seçimi, veri kümeleme ve veri yeniden örnekleme yöntemleri uygulanır ve daha sonra, ikinci aşamada, tahmin modelleri oluşturmak için birkaç sınıflandırma tekniği ve sınıflayıcı topluluğu kullanılır. Elde edilen sonuçlar, Veri Yeniden Örnekleme Yöntemi ile Bagging-karar ağaçlarının yüksek doğruluk oranına (% 82,96) ulaştığını ve önerilen iki aşamalı öngörme modelinin geleneksel tek aşamalı modellerden daha iyi olduğunu göstermiştir. [18]

b. Kolektif öğrenme yöntemlerinin birleşimi ile ilgili literatürde yapılan çalışmalar,

Gang Wang ve arkadaşları, çalışmalarında kurumsal kredi riski tahmini için Boosting ve Random Subspace kolektif öğrenme yöntemlerinin birleştirilmesi yaklaşımını önermişlerdir. Deneysel sonuçlar, Random Subspace-Boosting yönteminin lojistik regresyon, karar ağaçları, yapay sinir ağları, Bagging, Boosting ve Random Subspace gibi yedi yöntem arasında en iyi performansı gösterdiğini ortaya koymaktadır. Modelde China ve Darden kurumsal kredi veri setleri kullanılmıştır. [6]

Gang Wang ve arkadaşları, çalışmalarında Bagging ve Random Subspace olmak üzere iki popüler kolektif yöntemin üzerine kurulu RSB-SVM adlı yeni bir hibrit kolektif öğrenme yaklaşımı önermişlerdir. 239 şirketin finansal kayıtlarını içeren ve Çin Sanayi ve Ticaret Bankası tarafından toplanan kurumsal kredi riski veri seti, önerilen yöntemin etkinliğini ve fizibilitesini göstermek için seçilmiştir. Deneysel sonuçlar, RSB-SVM'nin kurumsal kredi riski değerlendirmesi için alternatif bir yöntem olarak kullanılabileceğini ortaya koymaktadır. [9]

Gang Wang ve arkadaşları, çalışmalarında gürültü verilerinin ve veri gereksiz niteliklerinin etkilerini azaltmak ve nispeten daha yüksek sınıflandırma doğruluğunu elde etmek için Bagging ve Random Subspace kolektif öğrenme stratejisine dayanan çift stratejili kolektif yöntemin birleşimini önermişlerdir. Random Subspace-Bagging karar ağacı ve Bagging-Random Subspace karar ağacı önerilen kolektif ağaçlardır. Modelde German ve Australian kredi veri setleri kullanılmıştır. Performans kriteri olarak doğruluk ve I-II. Tip Hata oranları

kullanılmıştır. Deney sonuçlarında doğruluk açısından Random Subspace-Bagging karar ağaçları ve Bagging-Random Subspace karar ağaçlarının daha iyi performans göstermiştir. I-II Tip Hata oranlarına göre net olarak performansı değerlendirilememiştir. [10]

A.I. Marqués ve arkadaşları, çalışmalarında tek kolektif sınıflandırıcıdan daha iyi performans sonuçları elde etmek amacıyla, farklı sınıflandırıcıları birleştirerek kredi skorlama için yeni bir yöntem önermişlerdir. Birleştirilen kolektif sınıflandırıcı yaklaşımı veri yeniden örnekleme ve öz nitelik alt küme seçim stratejilerini içermektedir. Veri Yeniden Örnekleme için AdaBoost ve Bagging algoritmaları; öz nitelik alt kümesi seçimi için Random Subspace ve Rotation Forest örneklendirilmiştir. Genel olarak, kredi skorlama uygulamalarında önemli ölçüde maliyet tasarrufuna neden olan, iki seviyeli sınıflayıcı kolektif öğrenmeleri Alıcı İşletim Özellikleri Eğrisi Altında Kalan Alan açısından en iyi sonuçları vermiştir. Belirli iki seviyeli kolektif modelin seçimi önemlidir, bu nedenle Bagging ve Rotation Forest modellerinin birlikte kullanılmasının, diğer kombinasyonlardan daha iyi performans gösterdiği belirtilmiştir. Deneylerden elde edilen diğer bir gösterge, AdaBoost algoritmasının bazı öz nitelik alt kümesi seçim yönteminden önce kullanılması, özellikle de Rotation Forest ile birleştirilmesi durumunda, ters sıralamadan dolayı performansı kötü sonucuna ulaşılmıştır. Deneylerde Australian, German, Japanese, UCSD, Iranian and Polish veri setleri kullanılmıştır. [7]

Shashi Dahiya ve arkadaşları, çalışmalarında FS-HB yani özellik seçimi tabanlı hibrit Bagging algoritması modelini önermişlerdir. Veri kümelerindeki önemli özelliklerin seçimi için ki-kare ve ana bileşen analizi özellik seçim yöntemleri kullanılmıştır. Veri bölümlenmeleri 60:40, 70:30, 75:25, 80:20 ve 85:15 olarak denenmiştir. C4.5 karar ağacı ve çoklu katmanlı algılayıcı temel sınıflandırıcılarla sırasıyla bireysel sınıflandırıcı olarak, özellik seçimli bireysel sınıflandırıcı olarak, kolektif sınıflandırıcı temel sınıflandırıcısı olarak ve özellik seçimli kolektif sınıflandırıcı temel sınıflandırıcı olarak kıyaslamalar yapılmıştır. Deney sonuçları, FS-HB modelinin daha az sayıda özellik kullanması ve benzer doğrulukta daha az karmaşık bir model sağlaması sebebi ile en iyi performansı göstermiştir. Model German kredi skorlama veri setinde uygulanmıştır. Veri bölümlenmelerinde ise

diğer bölümlemelere göre 70:30 bölümlemesinin daha iyi sonuçlar verdiği görülmüştür. [28]

Mohammadali Abedini ve arkadaşları, çalışmalarında dört aşamalı hibrit veri madenciliği model önermiştir. Modelde Bagging yönteminden esinlenilmiş olup son aşamadaki oylama yöntemine yeni bir öneride bulunulmuştur. Model German kredi skortlama veri setinde uygulanmıştır. Önerilen yeni model Çoklu Katmanlı Algılayıcı, Lojistik Regresyon, Destek Vektör Makineleri, Bagging(Lojistik Regresyon), Rotation Forest(Çoklu Katmanlı Algılayıcı) sınıflandırıcıları ile kıyaslanmıştır. Bireysel sınıflandırıcılardan hem doğruluk hem I ve II. Tip Hata kriterlerinde göre daha iyi sonuç vermiştir, ancak kolektif sınıflandırıcılardan doğruluk açısından iyi performans sağlamakla beraber I ve II. Tip Hata açısından daha iyi sonuç vermemiştir.[29]

Fatemeh Nemati Koutanaei ve arkadaşları, çalışmalarında kredi skortlama alanında daha yüksek performans elde edebilmek için sırasıyla veri toplama, özellik seçimi ve kolektif öğrenme sınıflandırıcılarının hibrit kullanımına dayalı bir model sunmuştur. İlk adımda, veri toplama ve ön işleme yapmışlardır. İkinci adımda sınıflandırma algoritmalarının daha yüksek performansını elde etmek için alt kümede daha iyi özellikler elde etmek için dört özellik seçim algoritması kullanmışlardır. Bu özellik seçim algoritmaları; Genetik Algoritma, Relief Method (RM), Bilgi Kazanç Oranı ve Temel Bileşenler Analizi'dir. Üçüncü adımda sınıflandırma modellemesi yapmışlardır. En iyi sınıflandırma algoritmasını bulmak üzere Naive Bayes, CART karar ağacı, yapay sinir ağları, destek vektör makineleri, sırasıyla bu temel sınıflandırıcıların kullanıldığı Bagging ve AdaBoost modelleri, Random Forest ve Stacking modellerinin karşılaştırmaları yapılmıştır. Deney sonuçları, en iyi algoritmanın Yapay Sinir Ağları-AdaBoost olduğunu ortaya koymuştur. Modelde İran'ın İhracat Geliştirme Bankası veri seti kullanılmıştır.[22]

c. Kolektif öğrenme yöntemlerine ilişkin performans kıyaslaması ile ilgili literatürde yapılan çalışmalar,

Gang Wang ve arkadaşları, çalışmalarında Bagging, Boosting ve Stacking kolektif öğrenme yöntemlerinin temel sınıflandırıcıları lojistik regresyon, karar ağaçları, yapay sinir ağları ve destek vektör makineleri olması durumunda kıyaslamalarını

incelemişlerdir. Modelde German, Australian ve China kredi veri setleri kullanılmıştır. Deneysel sonuçlarda Bagging metodunun Boosting metodundan daha iyi performans gösterdiği, Stacking ve Bagging metodlarının I. Tip Hata ve II. Tip Hata açısından en iyi sonuçları elde ettiği belirtilmiştir.[5]

Chih-Fong Tsai ve arkadaşları, çalışmalarında iki araştırma sorusuna yanıt aramıştır. Birincisi, kolektif öğrenme yöntemlerinden Bagging ve Boosting yöntemlerinden hangi yöntemin daha performanslı olduğunun araştırılması, ikincisi ise kolektif öğrenme sınıflandırıcılar için bir dizi belirli temel sınıflandırıcının bir araya getirilmesinde, en iyi performansı sağlamak için kaç tane sınıflandırıcının kombine edilmesi gerektiğidir. Modelde German, Australian ve Japanese kredi veri setleri kullanılmıştır. Deney sonuçlarında, Boosting yöntemini kullanan 80-100 temel sınıflandırıcıdan oluşan karar ağaçları topluluğunun en iyi performansı gösterdiği belirtilmiştir.[15]

Ning Chen ve arkadaşları, çalışmalarında AdaBoost, Bagging, Rotation Forest, Dagging, Decorate ve Random Subspace olmak üzere altı farklı kolektif öğrenme modellerinin performans kıyaslamasını yapmıştır. Modeller, Australian, French ve German kredi skorlama veri setlerinde uygulanmıştır. Temel sınıflandırıcı olarak Lojistik Regresyon, C4.5 Karar Ağacı, Naive Bayes, Radyal Temel İşlev, Destek Vektör Makineleri, k-En Yakın Komşuluk ve Çoklu Katmanlı Algılayıcı sınıflandırıcıları kullanılmıştır. Bagging ve Random Subspace performanslarını artırabileceği, ancak AdaBoost'un bu konuda Bagging daha kötü performansa sahip olduğu sonucuna varılmıştır.[19]

Stefan Lessmann ve arkadaşları, çalışmalarında bireysel sınıflandırıcıları, homojen ve heterojen kolektif sınıflandırıcıların kıyaslanması yapılmıştır. Modelde Australian, German, Th02, Bene-1, Bene-2, UK, PAK ve GMC(Give Me Some Credit) veri setleri kullanılmıştır. Performans ölçütleri olarak, Percentage Correctly Classified(PCC), AUC, Partial Gini Index (PG), H-testi, Brier Skoru (BS), Kolmogorov-Smirnov İstatistiği(KS) kullanılmıştır.[20]

C.R.Durga devi ve arkadaşları, çalışmalarında AdaBoost, Bagging ve Random Forest olmak üzere üç farklı kolektif öğrenme modellerinin performans kıyaslaması yapmıştır. Modeller German kredi skorlama veri setinde

uygulanmıştır. Tüm modellerin performansı doğruluk, I – II. Tip Hata ve Alıcı İşletim Eğrisi bakımından karşılaştırılmıştır. En iyi sonuncu veren modelin Random Forest modeli olduğu sonucuna ulaşılmıştır.[25]

d. Kolektif öğrenme yöntemlerinin temel sınıflandırıcılarının performans kıyaslaması ile ilgili literatürde yapılan çalışmalar,

A.I. Marqués ve arkadaşları, bu çalışmada beş farklı kolektif öğrenme yönteminin temel sınıflandırıcılarının performanslarını kıyaslamışlardır. Modelde Australian, German, Japanese, Iranian, Polish ve UCSD kredi veri setleri kullanılmıştır. Kolektif öğrenme yöntemleri olarak AdaBoost, Bagging, Random Subspace, DECORATE ve Rotation Forest yöntemleri incelenmiştir. Temel sınıflandırıcı olarak 1-En Yakın Komşuluk, Naive Bayes, Lojistik Regresyon, Çok Katmanlı Algılayıcı, Radyal Temel İşlev, Destek Vektör Makineleri, C4.5 karar ağacı algoritması sınıflandırıcıları uygulanmıştır. Deney sonuçları, doğruluk ve I. Tip Hata açısından C4.5 karar ağacının en iyi performansa sahip olduğunu göstermiştir. Daha sonra en iyi performansa en yakın sırasıyla Çok Katmanlı Algılayıcı, Lojistik Regresyon ve Destek Vektör Makinelerinin performansları olmuştur. 1-En Yakın Komşuluk ve Naive Bayes modellerinin tüm kolektif öğrenme yöntemlerinde en kötü performansı gösterdiği belirtilmiştir.[11]

V. Garcí'a ve arkadaşları, bu çalışmada beş gerçek dünya veritabanı ile Naive Bayes, Lojistik Regresyon, Çok Katmanlı Algılayıcı, Radyal Temel İşlev, Destek Vektör Makineleri, k-En Yakın Komşuluk, RIPPER ve Random Forest modeller gibi kredi skorlama modellerini içeren deneysel bir analiz sunmaktadır. Modelde Australian, German, Japanese, Iranian ve Polish kredi veri setleri kullanılmıştır. Doğruluk, I. ve II. Tip Hata performans kriteri olarak belirlenmiştir. Sonuçların tam kapsamlı analizi Random Forest ve Lojistik Regresyon performansının en yüksek ayrıca en sağlıklı metodlar olduğunu ortaya çıkarmıştır. Bunun tersine Naive Bayes sınıflandırıcıları ve k-En Yakın Komşuluk sınıflandırıcıları doğruluk ve I. Tip Hata performans kriteri açısından en kötü davranışa sahip modeller olarak tanımlamıştır.[8]

Shashi Dahiya ve arkadaşları, çalışmalarında sınıflandırma doğruluğunu artırmak için yedi ayrı modeli bir araya getiren kolektif bir teknik önermiştir.

Bireysel sınıflandırma teknikleri olarak; Lijistik Regresyon, Destek Vektör Makineleri, CART karar ağacı, C5.1 karar ağacı, Chi-square Automatic Interaction Detector (CHAID), En Yakın Komşuluk, QUEST kullanılmıştır. Modelde German kredi veri seti kullanılmıştır. Deney sonuçları kolektif modelin bireysel modellere kıyasla doğruluk açısından daha iyi olduğunu göstermiştir.[21]

Mohammad Siami ve arkadaşları, çalışmalarında müşterinin kredi puanı tahmininin tahmin doğruluğunu ve yorumlanabilirliğini artırmak için yerel doğrusal model ağacı algoritmasının (Locally Linear Model Tree Algorithm (LOLIMOT)) bir uygulaması denemiştir. Veri füzyonu ve özellik seçimi teknikleri ile kredi puanlama alanına göre düzeltme amacıyla algoritma geliştirilmiştir. Deneysel sonuçlar, geliştirilen modelin deneysel sonuçlarda diğer büyük sınıflandırıcılara üstünlüğünden emin olmak için yeterli güven üretmiştir. Deneylerde Australian ve German kredi veri setleri kullanılmıştır.[17]

You Zhu ve arkadaşları, çalışmalarında kolektif öğrenme yöntemlerinin birleşiminin bireysel makine öğrenme yöntemleri ve kolektif öğrenme yöntemlerinden daha iyi performansa sahip olması belirtmiştir. Deney verileri Shenzhen ve Shanghai Menkul Kıymetler Borsası 2012-2013 verilerinden seçilmiştir. Deneylerde Random Subspace ve Boosting birleşiminin en iyi yöntem olduğu sonucuna varılmıştır.[27]

e. Dengesiz kredi skorlama veri seti ile ilgili literatürde yapılan çalışmalar,

Giuseppe Paleologo ve arkadaşları, çalışmalarında verilerin çoğunlukla dengesiz olduğu, bilgilerin eksik olduğu gibi zorluklara sahip IBM İtalyan örneğini temel alan bir vaka çalışması sunmuşlardır. Temel amaç, eksik bilgi, sınıf dengesizliği olan veri noktalarını idare edebilen sağlam modeller oluşturmak ve doğrulamaktır. Kredi skorlaması verileri gibi son derece dengesiz veriler için özellikle uygun olan bir kolektif sınıflandırma tekniğinin kullanımı önerilmektedir. Subbagging, daha genel bir oylama yönteminin belirli bir versiyonuna, çok dengesiz veriler için uygun bagging üzerine odaklanılmaktadır. Karar ağacını onaylayarak, en iyi performansı gösteren sınıflandırıcının elde edildiği ve bir parametrenin ayarlanması yoluyla kabul edilebilir model karmaşıklığını koruduğu gösterilmiştir.[4]

Jin Xiao ve arkadaşları, çalışmalarında dengesiz veri dağılımı problemini çözmek üzere (Dynamic Classifier Ensemble Imbalanced Distribution (DCEID)) dinamik bir kolektif sınıflandırıcı yöntemi önermişlerdir. Her test müşterisi, dinamik sınıflandırıcı seçimi (Dynamic Classifier Selection(DCS)) ve dinamik kolektif seçimi (Dynamic Ensemble Selection (DES)) olmak üzere iki çeşit dinamik kolektif sınıflandırıcı yaklaşımından uygun olanı ile seçilmiştir. Deneyler, gerçek bir telekomünikasyon şirketinin kredi skorlama veri setinde yürütülmüş ve sonuçlar, DCEID ile sınıf dengesizliği sorununun yönetilebileceğini göstermiştir.[12]

Maciej Zięba ve arkadaşları, çalışmalarında sınıf etiketi tekniklerini değiştirilmesi tekniğinin dengesiz veri kümesinin ortaya çıkması ve asimetrik maliyet matrisi sorunu gibi veri madenciliği problemlerinin çözülmesine yardımcı olduğunu incelemişlerdir. Sınıf etiketi tekniklerinin değiştirilmesi ile Bagging, Boosting, Decorate, RIPPER, C4.5 karar ağacı, k-En Yakın Komşuluk, Çok Katmanlı Algılayıcı, Lojistik Regresyon ve Naive Bayes yöntemlerinin performansları karşılaştırılmıştır. Performansların karşılaştırılma kriteri, FN oranı ve ERI değeri olarak belirlenmiştir. FN oranı, çoğunluk sınıfından örnek olarak sınıflandırılan azınlık sınıfından örneklerin sayısının, azınlık sınıfından toplam örnek sayısına bölünmesi olarak tanımlanmıştır. ERI, ağırlıkların yanlış sınıflandırma maliyetlerine eşit olduğu ağırlıklı hata değeri olarak yorumlanmıştır. Deney sonuçlarında, sınıf etiketi tekniklerinin değiştirilmesi ile kolektif öğrenme sınıflandırma teknikleri diğer test edilen sınıflandırıcılardan daha iyi sonuçlar vermiştir. Modelde German kredi veri seti kullanılmıştır.[13]

Zongyuan Zhao ve arkadaşları, çalışmalarında verilerin dengesiz dağılımını düzenlemek için 'Average Random Choosing' yöntemi kullanmıştır. Sınıflandırma için üç katman içeren ileri beslemeli sinir ağı modeli (Multi-Layer Perceptron) kullanılmıştır. Model German kredi skorlama veri setinde uygulanmıştır. Veri setinde farklı eğitim-doğrulama-test dağılım oranlarına göre uygun olan dağılım tespit edilmeye çalışılmıştır. Data dağılımları 800:100:100, 900:50:50, 600:200:200 olarak bölümlenmiştir. Dengesiz veri dağılımı ile en kabul edilebilir oranın 800:100:100 olduğu belirtilmiştir. Kullanılan modelin doğruluk açısından performans iyileştirmesini sağladığı gösterilmiştir.[23]

Hong Wang, Qingsong Xu ve arkadaşları, çalışmalarında büyük dengesiz veriler göz önüne alındığında, kredi skorlama problemleriyle başa çıkabilmek için kolektif öğrenme yönteminde temel sınıflandırıcı olarak lojistik regresyon sınıflandırıcısının kullanılmasını incelemiştir. Veriler ilk önce Kümeleme ve Bagging algoritmaları ile dengelenmiş ve çeşitlendirilmiştir. Daha sonra, kredi risklerini değerlendirmek için bir Lasso-lojistik regresyon kolektif öğrenme yöntemi uygulanmıştır. Önerilen yöntemin Alıcı İşletim Özellikleri Eğrisi Altında Kalan Alan ve F-ölçüsü açısından, karar ağacı, Lasso-lojistik regresyon ve Random Forest gibi popüler kredi puanlama modellerinden daha iyi performans gösterdiği belirtilmiştir. Çalışmada kullanılan geniş kredi veri seti Kaggle.com veri madenciliği yarışma sitesinden alınmıştır.[24]

Uma R. Salunkhe ve Suresh N. Mali, çalışmalarında dengesiz veriler göz önüne alındığında kolektif öğrenme yönteminin performans değerlendirmesini yapmıştır. 8 farklı veri seti kullanılmıştır. Kolektif öğrenme yöntemi olarak Bagging ve Stacking yöntemleri ile çalışılmıştır. Dengesiz verilerin dengeli hale getirilmesi için SMOTE algoritması kullanılmıştır.[26]

1.2 Tezin Amacı

Bankacılık sisteminde, kredi verme konusunda müşterilerin belirlenmesinde kredi skorlama modelleri oldukça etkin rol oynamaktadır. Bu amaçla literatürde istatistik ve makine öğrenmesi yöntemleri kullanılarak bir çok kredi skorlama modeli geliştirilmiştir. Kredi skorlama modeli geliştirilirken kullanılan veri kümelerinde sınıflara ait örneklerin dengesiz bir dağılıma sahip olmalarından dolayı, modellerin doğruluk oranı düşük olmaktadır. Bu durum kredi skorlama sistemlerinde hala üstesinden gelinmesi gereken bir problem olarak görülmektedir. Bu çalışmada kolektif öğrenme algoritmalarını maliyete duyarlı öğrenme yöntemiyle birlikte kullanarak elde edilen modellerin performansını karşılaştırıp özellikle 'RİSKLİ' sınıf etiketine sahip sınıfın tespit edilmesi için en etkin kolektif öğrenme yönteminin belirlenebilmesine odaklanılmıştır.

1.3 Hipotez

Bu tezde Bagging ve AdaBoost kolektif öğrenme yöntemleri Karar Ağaçları, Destek Vektör Makineleri ve k-En Yakın Komşuluk temel sınıflandırıcıları ile iki farklı kredi veri seti üzerinde çalıştırılmıştır. Ayrıca Bagging ve AdaBoost için maliyet duyarlı öğrenme yöntemi kullanılarak azınlık sınıflandırma grubunun ceza puanı artırılmıştır. Bu yapılan çalışmada farklılık olarak dengesiz veri setini dengeli hale getirmeden kredi skorlama modelinde 'RİSKLİ' müşteri sınıfın tahminine odaklanılmıştır. Ayrıca bu tez, kolektif öğrenmeler konusunda daha sonra yapılacak çalışmalar için Türkçe kaynak niteliği taşımaktadır.



Kredi Skorlama, kredi almak üzere başvuru yapan kişi veya kurumların kredi alma konusundaki yeterliliğini somut veriler üzerinden analiz ederek değerlendiren ve kredileri geri ödeme noktasındaki risklilik seviyesini ölçümlendiren bir puanlama sistemidir.[34]

2.1 Kredi Skorlama Nedir?

Bireysel veya tüzel kişilerden oluşan tüketiciler, ihtiyaçlarını karşılamak üzere finansal kuruluşlardan satın alma gücü talep ederler. Talep edilen bu güç belli koşullarla finansal kuruluşlar tarafından tüketicilere ödünç olarak verilir. Bu ödünç verilen satın alma gücü 'kredi' kavramı ile ifade edilir. Alınan kredilerin belirli koşullara bağlı olarak finansal kuruluşlara geri ödemelerinin yapılması finansal kuruluşlar açısından oldukça önemli bir husustur. Bu sebeple doğru tüketiciye doğru krediyi vermek tüm finansal kuruluşlar tarafından hedeflenmektedir.

Tüketicilerin almış olduğu kredileri geri ödeyip ödememeleri konusu, finansal kuruluşların tüketicilerin risklilik açısından sıralanması gerekliliğini ortaya koymuştur. Kredi talebinde bulunan tüketicilerin krediye erişim ihtimalleri risklilik seviyesi büyük olandan küçük olana doğru artmaktadır. Tüketicilerin sıralanması durumu kredi kullandırımında daha önce öngörülmesi gereken kritik bir konudur.

Kredi analizi, tüketicilerin hangi şartlar altında kredilendirilmesi ve ne kadar riske sahip olması gerekliliği gibi durumları içeren geniş bir konudur. Kredi analizi, finansal olan ve finansal olmayan çalışmalar şeklinde iki bölümden oluşur. Finansal analiz, tüketicilerin finansal bilgilerinden yola çıkarak kredi alma konusunda yeterli olup olmadığı ile ilgilenir. Finansal olmayan analiz ise, tüketicilerin kredi alma performanslarının finansal veriler dışında kalan konularını inceler. Kredi analizi, büyük ölçekli ve küçük ölçekli tüketicilerin kredi alma

performanslarını incelemek için derecelendirme ve skorelama yöntemlerini kullanır. Büyük ölçekli tüketiciler için derecelendirme yöntemi kullanılırken küçük ölçekli tüketicilerde skorelama yöntemleri kullanılır. Derecelendirme, sübjektif ve objektif veriler üzerinden yapılır. Ancak skorelama sübjektif veriler ile ilgilenmez, yalnızca objektif verileri kullanır.

Kredi skorelama, krediye erişim sağlamak isteyen tüketicilerin kredi alma konusundaki yeterliliğinin bir puanlama ile ölçülmesini sağlayan bir sistemdir. Kredi skorelama sisteminde asıl amaç kredi almak isteyen tüketicinin nitel ve nicel verilerinin kredi skorelama modellerinde anlamlandırılması ile tüketici için genel bir risk profili belirlemektir. Modellerden çıkan sonuçlar, tüketiciye kredi verilip verilmemesi, verilecekse ne kadar limitli bir kredi verileceği gibi finansal kuruluşların oluşabilecek zararlarını önleyen kararlar sunmaktadır.

Kredi skorelama yalnızca tüketicinin kredi alma konusundaki yeterliliğini ölçümleyen bir sistem değildir. Aynı zamanda, finansal kuruluşların kaliteli kredi başvuruları elde etmesini, kredi kullanan tüketicilerin daha sonra da elde tutulmasını ve bir portföy oluşturarak tüketicilerin davranışlarının kontrol edilmesini de sağlamaktadır. Kredi skorelama amaçlarına göre çok çeşitli olabilmektedir. Örneğin, Başvuru Sahibini Skorelama (Applicant Scoring) ve Davranış Skorelaması (Behavioral Scoring).

Başvuru Sahibini Skorelama, ilk kez kredi başvurusunda bulunan bir tüketicinin beklenen şekilde davranıp davranmayacağını konu alır.

Davranış Skorelaması, hali hazırda kredi kullanmış olan borçluların risklilik seviyesini ölçümlemeyi konu alır.

Kredi skorelama modellerinde tüketicinin ödeme performansını etkileyen tüm özellikler skorelama modeline katılır.

2.2 Kredi Skorelama Ne Amaçla Yapılır?

Günümüzde bireysel veya tüzel kişiler, ihtiyaçlarını karşılamak için sıklıkla finansal kuruluşlardan kredi talebinde bulunmaktadır. Kredi riski, finans piyasasında borç alan tüketicinin aldığı borcu geri ödememesi veya zamanında

ödememesi gibi durumlarda finansal kuruluşun uğrayacağı zararı ifade etmektedir. Artan kredi talepleri ile finansal kuruluşlar için kredi risklerinin minimize edilmesi ihtiyacı oldukça önemli bir konu olmaya başlamıştır.

Kredi riskinin minimize edilmesi, talep edilen kredi tutarının borçlu tarafından ödemeye uygun olup olmaması veya finansal kuruluşların kredi dağıtımında azalan mevduat miktarlarını en uygun borçlular ile değerlendirmesi gibi konular kredi skorlama ihtiyacının başlıca sebepleri arasında yer almaktadır.

Kredi skorlama için müşterilerin geçmiş finansal, demografik, iş/meslek ve finansal hesaplarının davranışları gibi göstergelerden oluşan veri setlerine gerek duyulmaktadır. Örneğin, kredi talebinde bulunan bir borçlunun daha önceden almış olduğu bir krediyi ödememesi veya geciktirmesi halinde yeni bir kredi başvuru talebinin olumlu sonuçlanma ihtimali düşecektir. Kredi riski kavramında tüketicinin almış olduğu krediyi geri ödemedi en az 3 ay gecikmesi temerrüt diye adlandırılmaktadır. Temerrüt olasılığı ise gecikmenin gerçekleşme ihtimalinin 0 ile 1 arasındaki bir değerini ifade etmektedir. Kredi skorlama sisteminde temerrüt olasılığı tüketicinin yeni kredi talebi için oldukça önemli bir veri girdisi olarak kabul edilmektedir.

Finansal kuruluşların vermiş olduğu kredilerden kaynaklı zararlarının oluşmaması veya daha az etkileyici şekilde zararlarının oluşması için, kredi talepleri sırasında kredi skorlama sistemi ile tüketicilerin risklilik seviyelerinin tespit edilmesi ihtiyacı bulunmaktadır.

3.1 Kolektif Öğrenme Nedir?

Kolektif öğrenmeler, birden fazla makine öğrenme tekniklerinin aynı problemi çözmek üzere eğitildiği bir makine öğrenme yöntemidir.[5] Doğru sınıflandırmayı bulmak için çoklu sınıflandırıcılar olarak da adlandırılan kolektif öğrenmeleri, yüz veya binlerce modeli ortak bir amaçla birleştirilmesi ile daha büyük bir grup olarak sürece katılmaktadır. Kolektif öğrenmelerinin çalışma fikri, genel sınıflandırma sonuçlarını iyileştirmek için birden fazla kaynaktan bilgi birleştirme fikrini kullanmaktadır. Başka bir deyişle, varyans ve önyargı hatalarını azaltmak ya da tahminleri geliştirmek için çeşitli makine öğrenme tekniklerini tek bir tahmin modeline birleştiren meta algoritmalarıdır.

Varyans kavramı, bir veri seti üzerinde yapılan tahminlerde elde edilen tahminlerin aritmetik ortalamasının etrafındaki dağılımları yorumlamaya yarayan bir hata türüdür. Bir modelde yüksek varyans olması eğitim verilerinin, modelin yeni veriler üzerinde doğru tahminler yapamayacak kadar yakından eşleşen bir model oluşturmasına (over-fitting) ve dolayısı ile düşük performanslı modeller oluşturulmasına sebep olur.

Önyargı kavramı ise, bir veri seti üzerinde yapılan tahminlerde gerçek değerden uzak değerler olduğunu ölçülemeye yarayan bir hata türüdür. Bir modelde önyargı hatasının yüksek olması doğru tahminlerin kaçırılması ve dolayısı ile düşük performanslı modeller üretilmesine sebep olur.

Modellerin komplekslik seviyesinin artışı varyans ve önyargı için farklı değişimlere; varyans değerinin artmasına, bunun aksine önyargının düşmesine neden olur. Optimal bir model kompleksliği ile çok daha performanslı modeller elde edilebilir.

3.2 Kolektif Öğrenme Yöntemleri

Kolektif öğrenme yöntemleri iki grupta toplanmaktadır;

Paralel Kolektif Öğrenme Yöntemi, her biri paralel olarak bağımsız bir model oluşturan farklı öğrenme algoritmalarının birleşimidir. Bu gruptaki yöntemler daha basit bir yaklaşım olması sebebi ile anlaşılır bir karar verme sürecine sahiptir. Bunun yanı sıra verilerde gürültü seviyesinin yüksek olması durumunda sıralı kolektif öğrenme yöntemlerine göre daha iyi performans sergilemektedir. Bu gruptaki kolektif öğrenme yönteminin varyansı düşürme özelliği sebebi ile topluluk olarak tek bir temel sınıflandırıcının elde ettiği tahminden daha yüksek doğruluğa sahip olduğu söylenebilmektedir. Bu yöntemlere Bagging kolektif öğrenme yöntemi örnek verilebilir.

Sıralı Kolektif Öğrenme Yöntemi, ilk algoritmanın bir model oluşturmayı öğrenip daha sonraki algoritmanın bir önceki modelde yapılan yanlış düzeltilmesi şeklinde ilerleyen bir yaklaşımı olan modeldir. Bu gruptaki kolektif öğrenme yönteminin önyargıyı düşürme özelliği sebebi ile topluluk olarak tek bir temel sınıflandırıcının elde ettiği tahminden daha yüksek doğruluğa sahip olduğu söylenebilmektedir. Bu yöntemde kararsız temel sınıflandırıcı kullanılması ile daha güçlü tahmin modeli elde edilmektedir.[35] Bu yöntemlere Boosting kolektif öğrenme yöntemi örnek verilebilir.

Kolektif öğrenmeler için farklı bir grup da Hibrit Yöntemlerdir. Bu yöntemde de bir dizi temel sınıflandırıcı kullanılabilir, ancak farklı öğrenme teknikleri kullanılarak eğitilebilmektedir. Stacking, hibrit yöntemlerde en çok bilinen yöntemdir.

Bagging, Boosting ve Stacking yöntemleri çok iyi bilinen kolektif öğrenme yöntemleridir. Bu kolektif yöntemlerin kıyaslanmasına yönelik geçmişte akademik çalışmalar bulunmaktadır.[5]

3.2.1 Bagging Yöntemi

Her bir modelin bireysel tahminlerinden daha doğru tahminler yapmak üzere çoklu makine algoritmalarının paralel çalışması ile bir araya gelmesini sağlayan, basit olmasına karşın güçlü performans sergileyen bir makine öğrenme yöntemidir. Bu yöntem Leo Breiman tarafından formüle edilmiş ve adı "Bootstrap Aggregating" ifadesinin kısaltmasından çıkarılmıştır.[30]

Bagging yöntemini kullanabilmek için öncelikle temel sınıflandırıcı algoritmasının belirlenmesi gerekmektedir. Bir eğitim veri setinden değişim ile rastgele örnekleme (random sampling with replacement) yöntemi kullanılarak eğitim altküme veri setleri üretilmektedir. Bagging yönteminde herhangi bir verinin yeni üretilen veri setinde görülme olasılığı bulunmaktadır.

Bu üretilen alt küme veri setlerine aynı temel sınıflandırıcı uygulanarak çeşitli tahminler elde edilmektedir. Her bir tahminleme birbirinden bağımsız çalışmaktadır. Elde edilen tahminlerin birleşiminden sonuç elde edilirken regresyon için ortalama; sınıflandırma için ise çoğunluk oylama işlemi yapılmaktadır.

Bagging yöntemini kullanmanın avantajları;

- Birçok zayıf sınıflandırıcının bireysel performansından daha iyi performans sergilemektedir.
- Bir modelin yeni veriler üzerinde doğru tahminler yapamayacak kadar yakından eşleşen bir model oluşturma eğiliminde olması durumundan (over-fitting) kaçınmayı sağlamaktadır.
- Varyansı azaltmaktadır.

Bagging yöntemini kullanmanın dezavantajları;

- Modelin yorumlanabilirlik kaybı oluşmaktadır.
- Doğru modellenmemeye durumlarında yüksek önyargı hataları bulunmaktadır.
- Yüksek doğruluk sağlmasına rağmen hesaplama açısından maliyetli bir yöntem olarak görülmektedir.

Bagging yönteminin çalışma adımları Şekil 3.1 de gösterilmiştir.

GİRDİ	Veri seti $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ Temel öğrenme algoritması \mathcal{L} Öğrenme tur sayısı T
İŞLEM	<ol style="list-style-type: none"> 1. for $t = 1, 2, \dots, T$ 2. $h_t = \mathcal{L}(D, \mathcal{D}_{bs})$ \mathcal{D}_{bs} bootstrap dağılımı 3. end
ÇIKTI	$H(x) = \operatorname{argmax}_{y \in E} \sum_{t=1}^T 1(y = h_t(x))$

Şekil 3.1 Bagging Çalışma Adımları

3.2.2 Boosting Yöntemi

Bir dizi zayıf sınıflandırıcıyı daha güçlü bir sınıflandırıcıya dönüştüren, her bir iterasyonun sıralı bir şekilde işletilerek daha doğru tahminleme yapılmasını amaçlayan bir makine öğrenme yöntemidir. Bu yöntemde amaç, her bir iterasyonda yanlış sınıflandırılan verilerin ağırlıklarının arttırılması ile bir sonraki iterasyonda yanlış sınıflandırılmış verilere odaklanarak modelin doğruluğunu arttırmaktır. Boosting algoritmalarının doğuşu, Kearns ve Valiant'ın [1989] oluşturduğu ilginç bir kuramsal soruya verilen cevaptan kaynaklanmıştır. Bu kuramsal soru, zayıf bir öğrencinin potansiyel olarak güçlü bir öğrenciye yükseltilebildiği çıkarımının yapılmasını sağlamıştır.[31]

Boosting yöntemini kullanmanın avantajları;

- Birçok zayıf sınıflandırıcının bireysel performansından daha iyi performans sergilemektedir.
- Varyansı azaltmaktadır.
- Hataları optimize etmeye odaklanması sebebi ile önyargı hatasının azaltılması konusunda daha iyi performans gösterir.

Boosting yöntemini kullanmanın dezavantajları;

- Bir modelin yeni veriler üzerinde doğru tahminler yapamayacak kadar yakından eşleşen bir model oluşturma eğiliminde olması durumuna (over-fitting) yatkınlığı bulunmaktadır.
- Yüksek doğruluk sağlamasına rağmen hesaplama ve zaman açısından maliyetli bir yöntem olarak görülmektedir.

Boosting yönteminin çalışma adımları Şekil 3.2 de gösterilmiştir.

GİRİDİ	<p>Örnek dağılım \mathcal{D}</p> <p>Temel öğrenme algoritması \mathcal{L}</p> <p>Öğrenme tur sayısı T</p>
İŞLEM	<ol style="list-style-type: none"> 1. $\mathcal{D}_1 = \mathcal{D}$ Dağılımı başlat 2. for $t = 1, \dots, T$ 3. $h_t = \mathcal{L}(\mathcal{D}_t)$ \mathcal{D}_t dağılımından zayıf öğreniciyi eğit. 4. $\varepsilon_t = P_{x \sim \mathcal{D}_t}(h_t(x) \neq f(x))$ h_t nin hatalarını değerlendir. 5. $\mathcal{D}_{t+1} = Adjust_Distribution(\mathcal{D}_t, \varepsilon_t)$ 6. end
ÇIKTI	$H(x) = Combine_Outputs(\{h_1(x), \dots, h_t(x)\})$

Şekil 3.2 Boosting Çalışma Adımları

Boosting yöntemlerinin en iyi bilinen algortimaları arasında AdaBoost ve Gradient Boosting algoritmaları yer almaktadır.

3.2.2.1 AdaBoost Algoritması

1988 yılında Kearns tarafından sorulan bir sorudan ortaya çıkan Boosting yöntemini gerçeklemek üzere 1996 yılında Freund ve Schapire tarafından ilk pratik algoritma olarak AdaBoost algoritması sunulmuştur.[32]

Boosting yönteminin çalışma adımlarında bahsi geçen *Adjust_Distribution* ve *Combine_Outputs* ifadeleri gerçek bir algoritmayı ifade etmemektedir.

Bu ifadeleri gerçeklemek üzere etkili bir destek algoritması olarak AdaBoost ifade edilmektedir. AdaBoost algoritması zayıf bir sınıflandırıcıyı verilerin ağırlıklı

sürümleri ile yinelemeli olarak sürdürmesi ve her bir yinelemede verilerin yanlış sınıflandırılmış verileri daha büyük ağırlıklandırarak yeniden değerlendirilmesi şeklinde çalışmaktadır.

AdaBoost yöntemi diğer yöntemlerle kıyaslandığında tahmin hızının yüksek olması, daha az hafıza kullanması, uygulanabilir olması gibi özelliklerinden dolayı tercih edilmektedir.[33]

AdaBoost algoritmasının çalışma adımları Şekil 3.3 te gösterilmiştir.

GİRİDİ	Veri seti $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ Temel öğrenme algoritması \mathcal{L} Öğrenme tur sayısı T
İŞLEM	<ol style="list-style-type: none"> 1. $\mathcal{D}_1 = 1/m$ Ağırlık Dağılımı başlat 2. for $t = 1, \dots, T$ 3. $h_t = \mathcal{L}(D, \mathcal{D}_t)$ \mathcal{D}_t dağılımı ile D den bir öğrenciyi eğit. 4. $\epsilon_t = P_{x \sim \mathcal{D}_t}(h_t(x) \neq f(x))$ h_t nin hatalarını değerlendir. 5. $\epsilon_t > 0.5$ olana kadar devam et. 6. $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$ h_t nin ağırlığını belirle. 7. $\mathcal{D}_{t+1} = \frac{\mathcal{D}_t(x)}{Z_t} \times \begin{cases} \exp(-\alpha_t) & \text{if } h_t(x) = f(x) \\ \exp(\alpha_t) & \text{if } h_t(x) \neq f(x) \end{cases}$ Z_t'nin \mathcal{D}_{t+1}'in bir dağılım olmasını sağlayan normalizasyon faktörü olduğu dağılımı güncelleyin. 8. end
ÇIKTI	$H(x) = \text{sign} \left(\sum_{t=1}^T (\alpha_t h_t(x)) \right)$

Şekil 3.3 AdaBoost Çalışma Adımları

3.2.2.2 Gradient Boosting Algoritması

Gradient Boosting algoritması zayıf öğrenenlerden oluşan bir gruptan güçlü bir öğrenci oluşturmaya çalışan farklı bir boosting algoritma türüdür. Gradient Boosting algoritması da AdaBoost algoritmasına benzerdir, ancak bazı farklı yönleri bulunmaktadır. Bu yöntemde boosting problemi bir optimizasyon problemi olarak ele alınır. Yöntemin amacı, eğitim setindeki kayıp fonksiyonunun ortalama değerini minimize eden bir fonksiyon bulmaktır. Bu fikir ilk olarak Leo Breiman tarafından geliştirilmiştir.

Yöntemde ilk olarak zayıf bir öğrenci ele alınmaktadır. Her bir yinelemede performansı artırmak ve güçlü bir öğrenci oluşturmak için başka bir zayıf öğrenci modele eklenmektedir. Yeni zayıf öğrenciler, mevcut öğrencilerin yetersiz performans gösterdikleri alanlara yoğunlaşmaktadır. Bu ekleme işlemi kayıp fonksiyonunun kaybını azaltmaktadır. Yinelemeli bir şekilde modele eklenen her yeni zayıf öğrenciden sonra kayıp fonksiyonu tekrar hesaplanmaktadır. Kayıp, hata kalıntılarını, yani gerçek değer ile tahmin edilen değer arasındaki farkı temsil etmektedir. Bu kayıp değeri kullanılarak, hata kalıntılarını en aza indirmek için tahminler güncellenmektedir. Yineleme sayısı arttıkça hata kalıntıları sıfır değerine yaklaşmaktadır.

Gradient Boosting algoritmasının çalışma adımları Şekil 3.4 te gösterilmiştir.

GİRĐİ	Veri seti D Kayıp fonksiyonu L Temel öğrenme algoritması \mathcal{L}_{\square} Yineleme sayısı M Öğrenme oranı η
İŞLEM	<ol style="list-style-type: none"> 1. $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta}_0 = \arg \min_{\theta} \sum_{i=1}^n L(y_i, \theta);$ 2. for $m = 1, \dots, M$ do, 3. $\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f(x)=\hat{f}^{m-1}(x)};$ 4. $\hat{\phi}_m = \arg \min_{\phi \in \mathbb{R}, \beta} \sum_{i=1}^n [(-\hat{g}_m(x_i)) - \beta \phi(x_i)]^2;$ 5. $\hat{\rho}_m = \arg \min_{\rho} \sum_{i=1}^n L(y_i, \hat{f}^{m-1}(x_i) + \rho \hat{\phi}_m(x_i));$ 6. $\hat{f}_m(x) = \eta \hat{\rho}_m \hat{\phi}_m(x);$ 7. $\hat{f}^m(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x);$ 8. end
ÇIKTI	$\hat{f}(x) = \hat{f}^M(x) = (x + a)^n = \sum_{m=0}^M \hat{f}_m(x)$

Şekil 3.4 Gradient Boosting Çalışma Adımları

3.2.3 Stacking Yöntemi

Stacking, çoklu sınıflandırma modellerini bir ara sınıflandırıcı ile birleştirmek için kullanılan bir topluluk öğrenme tekniğidir. Stacking yöntemi farklı tipteki sınıflandırıcıları birleştirmek için kullanılması yönüyle oldukça cazip bir yöntem olmasına rağmen Bagging ve Boosting yöntemleri kadar yaygın kullanıma sahip değildir.

3.3 Maliyet Duyarlı Öğrenme Nedir?

Maliyete Duyarlı Öğrenme, veri sınıflandırmalarında yanlış sınıflandırılmaların maliyetlerinin belirlenmesi ile yanlış sınıflandırmaları dikkate alan bir öğrenme türüdür.[36] Bu öğrenme yöntemi, odaklanılan sınıfın yanlış sınıflandırılması durumundaki ceza puanının artırılarak riski ortadan kaldırmayı hedeflemektedir.

4.1 Verilerin Tanıtımı

Bu çalışmada, kredi skorlama konusunda oldukça bilinen 'German Credit' ve 'Australian' gerçek dünya kredi veri setleri kullanılmıştır.

'German Credit' ve 'Australian' veri setleri UCI Repository makine öğrenme veri ambarı web sitesinden elde edilmiştir.

4.1.1 German Credit Veri Seti

Veri seti, 20 özellik, 1 sınıf etiketi ve 1000 veri içeren orta büyüklükteki eğitim veri seti olarak paylaşılmıştır. Sınıf etiketi değerleri, 'RİSKSİZ' ve 'RİSKLİ' olarak belirlenmiştir. 'RİSKSİZ' sınıf etiketi, kredi borcunu geri ödeyen, 'RİSKLİ' sınıf etiketi ise kredi borcunu geri ödemeyen gerçek kişi müşteri kitlesini ifade etmektedir. Eğitim veri setinde sınıf etiketlerine göre veri dağılımları şu şekildedir;

- 'RİSKSİZ' olarak sınıflandırılan veri sayısı 700,
- 'RİSKLİ' olarak sınıflandırılan veri sayısı ise 300

Veri seti özellikleri arasında müşterilere ait birikim hesaplarının durumu, bankadaki varolan kredileri, aylık geliri gibi finansal ve yaş, medeni durum gibi demografik bilgiler bulunmaktadır.

Değişkenlerin adı, tipi Tablo 4.1. de belirtilmiştir;

Tablo 4.1 'German' Kredi Veri Seti Tanıtımı

DEĞİŞKEN ADI	DEĞİŞKEN TİPİ	DEĞİŞKEN ADI	DEĞİŞKEN TİPİ
Sınıf Etiketi	NOMINAL	İkametgah Süresi	NÜMERİK
Çek Hesap Durumu	NOMINAL	Mülkiyet Büyüklüğü	NOMINAL
Vade	NÜMERİK	Yaş	NÜMERİK
Kredi Geçmişi	NOMINAL	Diğer Ödeme Planları	NOMINAL

Amaç	NOMINAL	Ev Sahibi Durumu	NOMINAL
Kredi Miktarı	NÜMERİK	Varolan Kredileri	NÜMERİK
Birikim Durumu	NOMINAL	İş	NOMINAL
Çalışma Durumu	NOMINAL	Bağımlıların Sayısı	INTEGER
Ödeme Planı	NÜMERİK	Telefon Var/Yok	NOMINAL
Medeni Hal	NOMINAL	Yabancı Çalışan Mı	NOMINAL
Diğer Borç Durumu	NOMINAL		

4.1.2 Australian Veri Seti

Veri seti, 14 özellik, 1 sınıf etiketi ve 690 veri içeren kredi kart başvuruları eğitim veri seti olarak paylaşılmıştır. Özelliklerden 6 tanesi nümerik, 8 tanesi ise nominal veri içermektedir. Veri özelliklerinin açıklamaları belirtilmemiştir. Eğitim veri setinde sınıf etiketlerine göre veri dağılımları şu şekildedir;

- Pozitif sınıf veri sayısı 307,
- Negatif sınıf veri sayısı ise 383

Bu veri setinde değişkenlerin açıklamaları bulunmamaktadır. Değişkenlerin adı ve değişken tipi Tablo 4.2. de belirtilmiştir;

Tablo 4.2 'Australian' Kredi Veri Seti Tanıtımı

DEĞİŞKEN ADI	DEĞİŞKEN TİPİ	DEĞİŞKEN ADI	DEĞİŞKEN TİPİ
A1	NOMINAL	A9	NOMINAL
A2	NÜMERİK	A10	NÜMERİK
A3	NÜMERİK	A11	NOMINAL
A4	NOMINAL	A12	NOMINAL

A5	NOMINAL	A13	NÜMERİK
A6	NOMINAL	A14	NÜMERİK
A7	NÜMERİK	A15 (Class)	NOMINAL
A8	NOMINAL		

4.2 Performans Ölçütleri

Makine Makine öğrenme teknikleri model karşılaştırması için Doğruluk Oranı (Accuracy), Duyarlılık (Recall) ve Kesinlik (Precision) ölçülmektedir. Ama son yıllarda kullanılan model başarılarını karşılaştırmak için Alıcı İşletim Özellikleri Eğrisi Altında Kalan Alan (AUC-Area Under the Curve) kullanılmaktadır. [37]

4.2.1 Alıcı İşletim Özellikleri Eğrisi Altında Kalan Alan

Alıcı İşletim Özellikleri Eğrisi (Receiver Operating Characteristics(ROC) Curve), ikili sınıflandırma sistemlerinde, sistemin başarısına karar vermek için kullanılan bir performans ölçütüdür. ROC eğrilerinde, x ekseninde YPO (Yanlış Pozitif Oran) yer alırken, y ekseninde DPO (Doğru Pozitif Oran) bulunmaktadır. DPO ve YPO ikilileri ROC eğrisini oluşturur. [38]

DPO ve YPO değerlerinin anlaşılabilmesi için Karışıklık Matrisi (Confusion Matrix) Şekil 4.2.1 de gösterilmiştir.

Karışıklık Matrisi		Tahmin Değerleri	
		Negatif	Pozitif
Gerçek Değerler	Negatif	Doğru Negatif (DN)	Yanlış Pozitif (YP)
	Pozitif	Yanlış Negatif (YN)	Doğru Pozitif (DP)

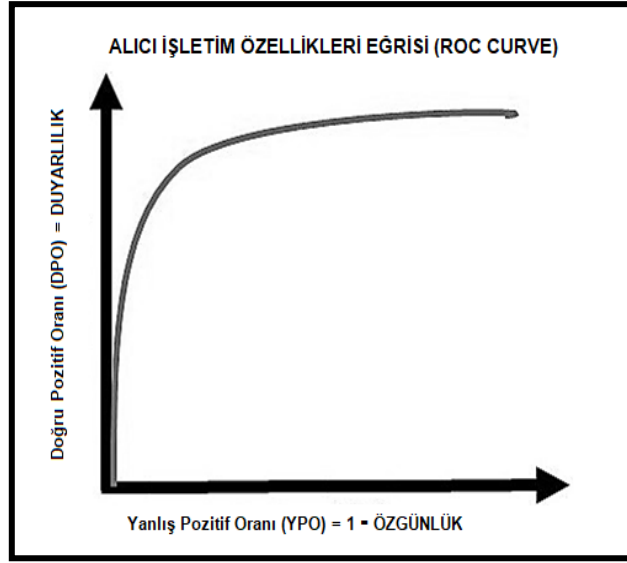
Şekil 4.1 Karışıklık Matrisi (Confusion Matrix)

$$\text{DUYARLILIK} = \frac{DP}{DP+YN}$$

$$1\text{-ÖZGÜNLÜK} = \frac{YP}{DN+YP}$$

Şekil 4.2 DPO ve YPO Eşitlikleri

Sistemin başarısının tek bir değer ile ifade edilmesi ROC eğrisinin altında kalan alan (AUC-Area Under the Curve) ile ifade edilmektedir. Bu alan değeri ne kadar büyük ise sistemin güvenilirlik değeri de o kadar yüksek olur. ROC Eğrisi ve altında kalan alan (AUC) Şekil 4.2.3 te gösterilmiştir.



Şekil 4.3 ROC Eğrisi

4.3 Deneylerin Uygulanması

Bu çalışmada, kredi skorlama modelleri için 'RİSKLİ' olarak sınıflandırılan müşteri kitlesinin önceden tahmin edilme oranının kolektif öğrenme yöntemlerinden Bagging veya Boosting açısından farklılığı incelenmiştir. Boosting yöntemi olarak iyi bilinen AdaBoost yöntemi ile çalışılmıştır.

Kolektif öğrenme yöntemlerinin temel sınıflandırıcıları olarak k-En Yakın Komşuluk, Destek Vektör Makineleri ve Karar Ağaçları ailesinden REPTree sınıflandırma teknikleri kullanılmıştır. Bu şekilde, kolektif öğrenme yöntemlerinin farklı prensiplerde çalışan temel sınıflandırıcılar üzerindeki etkileri incelenmiştir.

Dengesiz veri dağılımına sahip veri setlerinde uygulanan Maliyet Duyarlı Sınıflandırma yöntemi uygulanmıştır. Maliyet Duyarlı Sınıflandırma yönteminin uygulanması ile AdaBoost ve Bagging algoritmalarında sonuçların Doğruluk ve Alıcı İşletim Özellikleri Eğrisi Altında Kalan Alan (AUC-Area Under the Curve) açısından değişimleri gözlemlenmiştir.

4.4 Sonuçlar

WEKA programında, 'German Credit' ve 'Australian' veri setleri kullanılarak AdaBoost ve Bagging kolektif öğrenme yöntemi Karar Ağaçları, k-En Yakın Komşuluk ve Destek Vektör Makineleri temel sınıflandırıcıları ile çalıştırılmıştır. Elde edilen sonuçlar German Credit veri seti için Tablo 4.3, Australian veri seti için Tablo 4.4 te belirtilmiştir.

German Credit veri seti için hem doğruluk hem de Alıcı İşletim Karakteristik Eğri Altında Kalan Alan kriterine göre Bagging kolektif öğrenme algoritması AdaBoost algoritmasından ve temel sınıflandırıcılardan daha yüksek başarı sağlamıştır. Destek Vektör Makineleri temel sınıflandırıcısının tek başına kullanılması sonucu dışında tüm sonuçlarda temel sınıflandırıcıların kolektif öğrenme ile kullanımının hem doğruluk hem de Alıcı İşletim Karakteristik Eğri Altında Kalan Alan kriterine göre daha yüksek başarı elde edildiği görülmüştür.

Tablo 4.3 'German Credit' Maliyet Duyarlı Öğrenme (-) Deney Sonuçları

Bagging/AdaBoost	Baz Algoritma	RİSKSİZ Müşteri Örneklem Sayısı	Doğru Sınıflandırma (RİSKSİZ Müşteri)	Oran	RİSKLİ Müşteri Örneklem Sayısı	Doğru Sınıflandırma (RİSKLİ Müşteri)	Oran	Doğruluk	AUC
AdaBoost	Karar Ağaçları	700	595	0,85	300	146*	0,487*	74,1	0,731
Bagging	Karar Ağaçları	700	622	0,889*	300	132	0,44	75,4*	0,781*
-	Karar Ağaçları	700	601	0,859	300	117	0,39	71,8	0,72
AdaBoost	Destek Vektör Makineleri	700	610	0,871*	300	140	0,467	75	0,745
Bagging	Destek Vektör Makineleri	700	606	0,866	300	145	0,483*	75,1*	0,786*
-	Destek Vektör Makineleri	700	610	0,871*	300	141	0,47	75,1*	0,671
AdaBoost	k-En Yakın Komşuluk	700	624	0,891	300	118	0,393	74,2	0,656
Bagging	k-En Yakın Komşuluk	700	634*	0,906*	300	118	0,393	75,2*	0,744*
-	k-En Yakın Komşuluk	700	624	0,891	300	118	0,393	74,2	0,727

Tablo 4.4 ‘Australian’ Maliyet Duyarlı Öğrenme (-) Deney Sonuçları

Bagging/ AdaBoost	Baz Algoritma	POZİTİF Örneklem Sayısı	Doğru Sınıflandırma (POZİTİF)	Oran	NEGATİF Örneklem Sayısı	Doğru Sınıflandırma (NEGATİF)	Oran	Doğruluk	AUC
AdaBoost	Karar Ağaçları	307	242	0,788	383	335	0,875	83,6232	0,905
Bagging	Karar Ağaçları	307	262	0,853	383	336*	0,877*	86,6667*	0,932*
-	Karar Ağaçları	307	267*	0,87*	383	321	0,838	85,2174	0,902
AdaBoost	Destek Vektör Makineleri	307	250	0,851	383	326	0,814*	83,4783	0,912*
Bagging	Destek Vektör Makineleri	307	281	0,915	383	308	0,804	85,3623	0,876
-	Destek Vektör Makineleri	307	284*	0,925*	383	306	0,799	85,5072*	0,862
AdaBoost	k-En Yakın Komşuluk	307	250	0,814	383	320	0,836	82,6087	0,858
Bagging	k-En Yakın Komşuluk	307	257*	0,837*	383	326*	0,851*	84,4928*	0,904*
-	k-En Yakın Komşuluk	307	253	0,824	383	324	0,846	83,6232	0,896

WEKA programında, ‘German Credit’ ve ‘Australian’ veri setleri kullanılarak Karar Ağaçları, k-En Yakın Komşuluk ve Destek Vektör Makineleri temel sınıflandırıcıları ile maliyet duyarlı öğrenme yöntemi uygulanarak çalıştırılmıştır. Maliyet duyarlı öğrenme yönteminde, ‘RİSKLİ’ veya ‘NEGATİF’ sınıfın yanlış sınıflandırılma ceza puanı diğer sınıfın 2 katı olacak şekilde ayarlanmıştır. Elde edilen sonuçlar German Credit veri seti için Tablo 4.5, Australian veri seti için Tablo 4.6 de belirtilmiştir.

Yapılan deneylerde kolektif öğrenme algoritmalarına maliyet duyarlı öğrenme kullanımının anlamlı sonuçlar elde etmediği tespit edilmiştir. Her iki veri seti için de doğruluk kriterine göre en yüksek başarı Destek Vektör Makineleri temel sınıflandırıcısı ile, Alıcı İşletim Karakteristik Eğri Altında Kalan Alan kriterine göre ise en yüksek başarı k-En Yakın Komşuluk temel sınıflandırıcısı ile elde edilmiştir.

Tablo 4.5 ‘German Credit’ Maliyet Duyarlı Öğrenme (+) Deney Sonuçları

Baz Algoritma	RİSKSİZ Müşteri Örneklem Sayısı	Doğru Sınıflandırma (RİSKSİZ Müşteri)	Oran	RİSKLİ Müşteri Örneklem Sayısı	Doğru Sınıflandırma (RİSKLİ Müşteri)	Oran	Doğruluk	AUC
Karar Ağaçları	700	515	0,736	300	191	0,637	70,6	<u>0,709</u>
Destek Vektör Makineleri	700	525	0,75	300	202	0,673*	72,7*	0,712
k-En Yakın Komşuluk	700	477	0,681	300	191	0,637	<u>66,8</u>	0,727*

Tablo 4.6 ‘Australian’ Maliyet Duyarlı Öğrenme (+) Deney Sonuçları

Baz Algoritma	POZİTİF Örneklem Sayısı	Doğru Sınıflandırma (POZİTİF)	Oran	NEGATİF Örneklem Sayısı	Doğru Sınıflandırma (NEGATİF)	Oran	Doğruluk	AUC
Karar Ağaçları	307	218	0,71	383	349	0,911	<u>82,1739</u>	0,897
Destek Vektör Makineleri	307	239	0,779	383	351	0,916	85,5072*	<u>0,847</u>
k-En Yakın Komşuluk	307	231	0,752	383	352	0,919*	84,4928	0,903*

4.4.1 Doğruluk Kriterine Göre

‘German Credit’ ve ‘Australian’ veri setleri için elde edilen deney sonuçlarının maliyet duyarlı öğrenme uygulanıp uygulanmamasına göre doğruluk değerleri karşılaştırılmıştır. Karşılaştırmalar German Credit veri seti için Tablo 4.7, Australian veri seti için Tablo 4.8 de belirtilmiştir.

German Credit veri setinde, tüm temel sınıflandırıcılar için Bagging kolektif öğrenme algoritmasının, AdaBoost, Maliyet Duyarlı öğrenme ve yalnızca temel sınıflandırıcı kullanımına göre daha üstün olduğu görülmüştür. Australian veri setinde bu durum Destek Vektör Makineleri temel sınıflandırıcısı dışında diğer temel sınıflandırıcılar için de sağlanmıştır.

Tablo 4.7 ‘German Credit’ Doğruluk Kriterine Göre Deney Sonuçları

DOĞRULUK	MALİYET DUYARLI ÖĞRENME (-)	AdaBoost	Bagging	MALİYET DUYARLI ÖĞRENME (+) (RİKSİZ: 1.0 / RİSKLİ: 2.0)
k-En Yakın Komşuluk	74,2	74,2	75,2*	66,8
Karar Ağaçları	71,8	74,1	75,4*	70,6
Destek Vektör Makineleri	75,1*	75	75,1*	72,7

Tablo 4.8 ‘Australian’ Doğruluk Kriterine Göre Deney Sonuçları

DOĞRULUK	MALİYET DUYARLI ÖĞRENME (-)	AdaBoost	Bagging	MALİYET DUYARLI ÖĞRENME (+) (POZİTİF: 1.0 / NEGATİF: 2.0)
k-En Yakın Komşuluk	83,6232	82,6087	84,4928	84,4928*
Karar Ağaçları	85,2174	83,6232	86,6667	82,1739
Destek Vektör Makineleri	85,5072*	83,4783	85,3623	85,5072*

4.4.2 Alıcı İşletim Karakteristik Eğri Altında Kalan Alan Kriterine Göre

'German Credit' ve 'Australian' veri setleri için elde edilen deney sonuçlarının maliyet duyarlı öğrenme uygulanıp uygulanmamasına göre Alıcı İşletim Karakteristik Eğri Altında Kalan Alan değerleri karşılaştırılmıştır. Karşılaştırmalar German Credit veri seti için Tablo 4.9, Australian veri seti için Tablo 4.10 da belirtilmiştir.

German Credit veri setinde, tüm temel sınıflandırıcılar için Bagging kolektif öğrenme algoritmasının, AdaBoost, Maliyet Duyarlı öğrenme ve yalnızca temel sınıflandırıcı kullanımına göre daha üstün olduğu görülmüştür. Australian veri setinde bu durum Destek Vektör Makineleri temel sınıflandırıcısı dışında diğer temel sınıflandırıcılar için de sağlanmıştır. Australian veri setinde Destek Vektör Makineleri temel sınıflandırıcısı için en iyi sonuç AdaBoost için elde edilmiştir.

Tablo 4.9 'German Credit' AUC Kriterine Göre Deney Sonuçları

AUC	MALİYET DUYARLI ÖĞRENME (-)	AdaBoost	Bagging	MALİYET DUYARLI ÖĞRENME (+) (RİSKSİZ: 1.0 / RİSKLİ: 2.0)
k-En Yakın Komşuluk	0,727	0,656	0,744*	0,727
Karar Ağaçları	0,72	0,731	0,781*	0,709
Destek Vektör Makineleri	0,671	0,745	0,786*	0,712

Tablo 4.10 'Australian' AUC Kriterine Göre Deney Sonuçları

AUC	MALİYET DUYARLI ÖĞRENME (-)	AdaBoost	Bagging	MALİYET DUYARLI ÖĞRENME (+) (POZİTİF: 1.0 / NEGATİF: 2.0)
k-En Yakın Komşuluk	0,896	0,858	0,904*	0,903
Karar Ağaçları	0,902	0,905	0,932*	0,897
Destek Vektör Makineleri	0,862	0,912*	0,876	0,847

4.4.3 Doğru Sınıflandırma (RİSKLİ Müşteri ve Negatif) Kıyaslaması

'German Credit' veri seti için elde edilen deney sonuçlarının maliyet duyarlı öğrenme uygulanıp uygulanmamasına göre 'RİSKLİ' müşteri tahmin oranı değerleri karşılaştırılmıştır. Karşılaştırmalar Tablo 4.11 de belirtilmiştir.

Tablo 4.11 'RİSKLİ' Müşteri Tahmin Oranı Deney Sonuçları

BAZ ALGORİTMA	MALİYET DUYARLI ÖĞRENME (-)	AdaBoost	Bagging	MALİYET DUYARLI ÖĞRENME (+) (RİKSİZ: 1.0 / RİSKLİ: 2.0)
k-En Yakın Komşuluk	0,393	0,393	0,393	0,637*
Karar Ağaçları	0,39	0,487	0,44	0,637*
Destek Vektör Makineleri	0,47	0,467	0,483	0,673*

'Australian' kredi veri seti için elde edilen deney sonuçlarının maliyet duyarlı öğrenme uygulanıp uygulanmamasına göre 'NEGATİF' tahmin oranı değerleri karşılaştırılmıştır. Karşılaştırmalar Tablo 4.12 de belirtilmiştir.

Tablo 4.12 'NEGATİF' Tahmin Oranı Deney Sonuçları

BAZ ALGORİTMA	MALİYET DUYARLI ÖĞRENME (-)	AdaBoost	Bagging	MALİYET DUYARLI ÖĞRENME (+) (POZİTİF: 1.0 / NEGATİF: 2.0)
k-En Yakın Komşuluk	0,846	0,836	0,851	0,919*
Karar Ağaçları	0,838	0,875	0,877	0,911*
Destek Vektör Makineleri	0,799	0,814	0,804	0,916*

'RİSKLİ' müşteri veya 'NEGATİF' veri sınıfının tahmin oranının en yüksek olduğu yöntemin maliyet duyarlı öğrenme yöntemi olduğu görülmüştür.

Maliyet duyarlı öğrenme yöntemi, dengesizlik seviyesinin daha az olduğu veri setinde negatif sınıfın tahmin edilme oranını daha çok iyileştirebilmiştir.

Bankacılık sisteminde, kredi verme konusunda müşterilerin belirlenmesinde kredi skorlama modelleri oldukça etkin rol oynamaktadır. Bu amaçla literatürde istatistik ve makine öğrenmesi yöntemleri kullanılarak bir çok kredi skorlama modeli geliştirilmiştir. Kredi skorlama modeli geliştirilirken kullanılan veri kümelerinde sınıflara ait örneklerin dengesiz bir dağılıma sahip olmalarından dolayı, modellerin doğruluk oranı düşük olmaktadır. Bu durum kredi skorlama sistemlerinde hala üstesinden gelinmesi gereken bir problem olarak görülmektedir. Bu çalışmada kolektif öğrenme algoritmalarını maliyete duyarlı öğrenme yöntemiyle birlikte kullanarak elde edilen kolektif modellerin performansını karşılaştıran en etkin modellerin belirlenmesine çalışılmıştır. 'RİSKLİ' müşteri kitlesinin önceden tahminlenmesindeki başarı oranına odaklanılmıştır.

Bu amaçla Bagging ve AdaBoost kolektif öğrenme yöntemleri Karar Ağaçları, Destek Vektör makineleri ve k-En Yakın Komşuluk temel sınıflandırıcıları ile iki farklı kredi veri seti üzerinde çalıştırılmıştır. Ayrıca temel sınıflandırıcılar için maliyet duyarlı öğrenme yöntemi kullanılarak azınlık sınıflandırma grubunun ceza puanı artırılmıştır.

German Credit veri seti için hem doğruluk hem de Alıcı İşletim Karakteristik Eğri Altında Kalan Alan kriterine göre Bagging kolektif öğrenme algoritması AdaBoost algoritmasından ve temel sınıflandırıcılardan daha yüksek başarı sağlamıştır. Destek Vektör Makineleri temel sınıflandırıcısının tek başına kullanılması sonucu dışında tüm sonuçlarda temel sınıflandırıcıların kolektif öğrenme ile kullanımının hem doğruluk hem de Alıcı İşletim Karakteristik Eğri Altında Kalan Alan kriterine göre daha yüksek başarı elde edildiği görülmüştür.

Her iki veri seti için de doğruluk kriterine göre en yüksek başarı Destek Vektör Makineleri temel sınıflandırıcısı ile, Alıcı İşletim Karakteristik Eğri Altında Kalan

Alan kriterine göre ise en yüksek başarı k-En Yakın Komşuluk temel sınıflandırıcısı ile elde edilmiştir.

German Credit veri setinde, tüm temel sınıflandırıcılar için Bagging kolektif öğrenme algoritmasının, AdaBoost, Maliyet Duyarlı öğrenme ve yalnızca temel sınıflandırıcı kullanımına göre daha üstün olduğu görülmüştür. Australian veri setinde bu durum Destek Vektör Makineleri temel sınıflandırıcısı dışında diğer temel sınıflandırıcılar için de sağlanmıştır. Australian veri setinde Destek Vektör Makineleri temel sınıflandırıcısı için en iyi sonuç AdaBoost için elde edilmiştir.

'RİSKLİ' müşteri veya 'NEGATİF' veri sınıfının tahmin oranının en yüksek olduğu yöntemin maliyet duyarlı öğrenme yöntemi olduğu görülmüştür.

Maliyet duyarlı öğrenme yöntemi, dengesizlik seviyesinin daha az olduğu veri setinde negatif sınıfın tahmin edilme oranını daha çok iyileştirebilmiştir.

- [1] L. Yu ve ark, "Support vector machine based multiagent ensemble learning for credit risk evaluation", *Expert Systems with Applications* 37, pp.1351–1360, 2010.
- [2] L. Zhou ve ark, "Least squares support vector machines ensemble models for credit scoring", *Expert Systems with Applications* 37, pp.127–133, 2010.
- [3] N. Hsieh ve ark, "A data driven ensemble classifier for credit scoring analysis", *Expert Systems with Applications* 37, pp. 534–545, 2010.
- [4] G. Paleologo ve ark, "Subagging for credit scoring models", *European Journal of Operational Research* 201, pp.490–499, 2010.
- [5] G. Wang ve ark, "A comparative assessment of ensemble learning for credit scoring", *Expert Systems with Applications* 38, pp.223–230, 2010.
- [6] G. Wang ve ark, "Study of corporate credit risk prediction based on integrating boosting and random subspace", *Expert Systems with Applications* 38, pp.13871–13878, 2011.
- [7] A.I. Marqués ve ark, "Exploring the behaviour of base classifiers in credit scoring ensembles", *Expert Systems with Applications* 39, pp.10244–10250, 2012.
- [8] V. García ve ark, "Non-parametric Statistical Analysis of Machine Learning Methods for Credit Scoring", *Management Intelligent Systems*, pp.263–272, 2012.
- [9] G. Wang ve ark, "A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine", *Expert Systems with Applications* 39, pp.5325–5331, 2012.
- [10] G. Wang ve ark, "Two credit scoring models based on dual strategy ensemble trees", *Knowledge-Based Systems* 26, pp.61–68, 2012.
- [11] A.I. Marqués ve ark, "Two-level classifier ensembles for credit risk assessment", *Expert Systems with Applications* 39, pp.10916–10922, 2012.
- [12] J. Xiao ve ark, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution", *Expert Systems with Applications* 39, pp.3668–3675, 2012.
- [13] M. Zięba ve ark, "Ensemble Classifier for Solving Credit Scoring Problems", *IFIP International Federation for Information Processing*, pp.59–66, 2012.
- [14] J. Li ve ark, "Weight-Selected Attribute Bagging for Credit Scoring", *Hindawi Publishing Corporation Mathematical Problems in Engineering*, vol 2013, no 379690, pp. 13, 2013.
- [15] C. Tsai ve ark, "A comparative study of classifier ensembles for bankruptcy prediction", *Applied Soft Computing* 24, pp.977–984, 2014.
- [16] J. Sun ve ark, "Dynamic credit scoring using B&B with incremental-SVM-ensemble", *Kybernetes*, vol 44, no 4, pp. 518-535, 2014.

- [17] M. Siami ve ark, "An application of locally linear model tree algorithm with combination of feature selection in credit scoring", *International Journal of Systems Science*, vol 45, no 10, pp.2213–2222, 2014.
- [18] W. Hsu-Che ve ark, "Two-stage credit rating prediction using machine learning techniques", *Kybernetes*, vol 43, no 7, pp. 1098-1113, 2014.
- [19] N. Chen ve ark, "Comparative study of classifier ensembles for cost-sensitive credit risk assessment", *Intelligent Data Analysis* 19, pp.127–144, 2015.
- [20] S. Lessmann ve ark, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research", *European Journal of Operational Research* 247, pp.124–136, 2015.
- [21] S. Dahiya ve ark, "Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set", *Industrija*, vol 43, no 4, 2015.
- [22] F. N. Koutanaei ve ark, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring", *Journal of Retailing and Consumer Services* 27, pp.11–23, 2015.
- [23] Z. Zhao ve ark, "Investigation and improvement of multi-layer perceptron neural networks for credit scoring", *Expert Systems with Applications* 42, pp.3508–3516, 2015.
- [24] H. Wang ve ark, "Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble", *PLOS ONE*, 2015.
- [25] C.R.Durga devi ve ark, "A Relative Evaluation of the Performance of Ensemble Learning in Credit Scoring", *IEEE International Conference on Advances in Computer Applications (ICACA)*, 2016.
- [26] U. R. Salunkhe, Suresh N. Mali, "Classifier Ensemble Design for Imbalanced Data Classification A Hybrid Approach", *International Conference on Computational Modeling and Security (CMS 2016) Procedia Computer Science* 85, pp.725 – 732, 2016.
- [27] Y. Zhu ve ark, "Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance", *Neural Comput & Applic* 28, pp.41 – 50, 2017.
- [28] S. Dahiya ve ark, "A feature selection enabled hybrid-bagging algorithm for credit risk evaluation", *WILEY Expert Systems*, 2017.
- [29] M. Abedini ve ark, "Customer credit scoring using a hybrid data mining approach", *Kybernetes*, vol 45, no 10, pp. 1576-1588, 2018.
- [30] K. Machová ve ark, "A Comparison of the Bagging and the Boosting Methods Using the Decision Trees Classifiers", *ComSIS*, vol 3, no 2, 2006.
- [31] Z. Zhou, *Ensemble Methods Foundations and Algorithms*, Cambridge UK: CRC Press, 2012.
- [32] D. Nielsen, "Tree Boosting With XGBoost", *Norwegian University of Science and Technology*, 2016.

- [33] F. Bulut, "Sınıflandırıcı Topluluklarının Dengesiz Veri Kümeleri Üzerindeki Performans Analizleri", *Bilişim Teknolojileri Dergisi*, Cilt: 9, Sayı: 2, Syf. 153-158, 2016.
- [34] M. Akça, Kredi Skorlama Nedir?, <http://mustafaakca.com/kredi-skorlama-nedir/>, 12 Mayıs 2015.
- [35] Xia, Liu ve Li, "A boosted decision tree approach using Bayesian hyperparameter optimization for credit scoring", *Expert Systems with Applications* 78, pp.225-241, 2017.
- [36] C. X. Ling, V. S. Sheng, "Cost-Sensitive Learning and the Class Imbalance Problem", *Encyclopedia of Machine Learning*, 2008.
- [37] T. Fawcett, "An introduction to ROC analysis", *Pattern Recognition Letters*, vol 27, no 8, 2006.
- [38] S. Kılıç, "Klinik Karar Vermede ROC Analizi", *Journal of Mood Disorders*, vol 3, no 3, 2013.



Konferans Bildirileri

- [1] N.Ankara, H.Şahintürk, “Dengesiz Kredi Skorlama Veri Setlerinde Kolektif Öğrenme Algoritmalarının Performans Değerlendirmesi”, *Global İşletme Araştırmaları Kongresi*, 30-31 Mayıs 2019.

