T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

BANKA MÜŞTERİLERİNİN MAKİNE ÖĞRENMESİ İLE ANALİZ EDİLEREK SKORLANMASI

Yüksek Lisans Tezi

Tezi Hazırlayan
Fatih KAZOVA

İstanbul, 2025

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

BANKA MÜŞTERİLERİNİN MAKİNE ÖĞRENMESİ İLE ANALİZ EDİLEREK SKORLANMASI

Yüksek Lisans Tezi

Tezi Hazırlayan
Fatih KAZOVA

Öğrenci No
2220003029

ORCID ID
0000-0002-6028-1823

Danışman
Dr. Öğr. Üyesi Ediz ŞAYKOL

İstanbul, 2025

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum ‘‘**Banka Müşterilerinin Makine Öğrenmesi ile Analiz Edilerek Skorlanması**’’ başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, bu tezdeki bütün bilgileri akademik ve etik kurallar içinde elde ettiğimi, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını, patent ve telif haklarını ihlal edici bir davranışımın olmadığını belirtir ve bunu onurumla doğrularım. 08/05/2025

Fatih KAZOVA

T.C.
İSTANBUL BEYKENT ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ MÜDÜRLÜĞÜ
TEZLİ YÜKSEK LİSANS SINAV TUTANAĞI

08/05/2025

Enstitümüz ***Bilgisayar Mühendisliği*** Anabilim Dalı ***Bilgisayar Mühendisliği*** Programı yüksek lisans öğrencilerinden ***2220003029*** numaralı ***Fatih KAZOVA***'nın *"İstanbul Beykent Üniversitesi Lisansüstü Eğitim – Öğretim Yönetmeliği"*nin ilgili maddesine göre hazırlayarak, Enstitümüze teslim ettiği ***"Banka Müşterilerinin Makine Öğrenmesi ile Analiz Edilerek Skorlanması"*** konulu tezini, Yönetim Kurulumuzun 16/04/2025 tarih ve 2025/18 sayılı toplantısında seçilen ve On-Line toplanan biz jüri üyeleri huzurunda, İstanbul Beykent Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 29. maddesinin 3. fıkrası gereğince Zoom programı aracılığıyla on-line olarak aday tarafından savunulmuş ve sonuçta adayın tezi hakkında ***"OYBİRLİĞİ"*** ile ***"KABUL"*** kararı verilmiştir.

İşbu tutanak, 2 nüsha olarak hazırlanmış ve Enstitü Müdürlüğü'ne sunulmak üzere tarafımızdan düzenlenmiştir.

DANIŞMAN
Dr. Öğr. Üyesi Ed*** ŞA***
(İstanbul Beykent Üniversitesi)

ÜYE
Doç. Dr. At*** YI***
(İstanbul Beykent Üniversitesi)

ÜYE
Dr. Öğr. Üyesi So*** SE***
(Kütahya Dumlupınar Üniversitesi)

Adı ve Soyadı : Fatih KAZOVA

Danışmanı : Dr. Öğr. Üyesi Ediz ŞAYKOL

Derecesi ve Tarihi : Yüksek Lisans (Tezli), 2025

Alanı : Bilgisayar Mühendisliği

Anahtar Kelimeler : XGBoost, KMeans, Makine Öğrenmesi, Kredi Skoru

ÖZ

BANKA MÜŞTERİLERİNİN MAKİNE ÖĞRENMESİ İLE ANALİZ EDİLEREK SKORLANMASI

Bankacılık sektörü, firmalara sağladığı kredilerle ekonomik büyümeyi desteklerken, kredi riskinin doğru değerlendirilmesi bu süreçlerin sürdürülebilirliği için kritik bir gerekliliktir. Türkiye'de, firmaların mali yapılarındaki çeşitlilik, geleneksel kredi skorlama yöntemlerinin sınırlarını zorlamaktadır. Bu tez çalışması, banka müşterisi firmaların makine öğrenmesi teknikleriyle analiz edilerek skorlanmasını amaçlamıştır. Çalışmadaki veri setinde, firmaların finansal durumları, risk bilgileri ve çalışan sayısı bilgileri yer almaktadır.

Çalışmada, iki aşamalı bir analiz süreci izlenmiştir. İlk aşamada, K-Means algoritması kullanılarak firmalar risk verilerine göre dört küme oluşturulmuştur. Birinci aşamada elde edilen kümeler bağımsız değişken olarak modele eklenmiştir. İkinci aşamada, denetimli makine öğrenmesi olan XGBoost, Karar Ağacı, Rastgele Orman, KNN ve LightGBM algoritmaları kullanılarak, kredi skoru tahmin edilmiştir. Kredi skoru dört sınıflı (0-1-2-3) bir bağımlı değişkendir. Model performansları; kesinlik, duyarlılık ve F1 skoru gibi kriterlerle değerlendirilerek başarılı sonuçlar elde edilmiştir. Bu çalışmanın firmalara yönelik yapılması, firmaların finansal durumları ve risk verilerine ek olarak denetimsiz makine öğrenmesi yöntemleri ile segmentasyon oluşturulması, oluşturulan bu segmentasyonun kredi skorlamada kullanılması yönleriyle literatürdeki mevcut yaklaşımlardan farklılaşarak özgün bir katkı sunması amaçlanmıştır.

Name and Surname : Fatih KAZOVA

Supervisor : Asst. Prof. Dr. Ediz ŞAYKOL

Degree and Date : Master's (Thesis), 2025

Major : Computer Engineering

Keywords : XGBoost, KMeans, Machine Learning, Credit Score

ABSTRACT

ANALYZING AND SCORING BANK CUSTOMERS WITH MACHINE LEARNING

The banking sector supports economic growth by providing loans to businesses, and the accurate assessment of credit risk is a critical requirement for the sustainability of these processes. In Turkey, the diversity in the financial structures of businesses challenges the limitations of traditional credit scoring methods. This thesis aims to analyze and score bank customer firms using machine learning techniques. The dataset in this study includes information about the financial status, risk data, and the number of firms' employees.

The study follows a two-phase analysis process. In the first phase, firms are grouped into four clusters based on risk data using the K-Means algorithm. The clusters obtained in the first phase are then used as independent variables in the model. In the second phase, supervised machine learning algorithms, including XGBoost, Decision Tree, Random Forest, KNN, and LightGBM, predict the credit score. The credit score is a four-class dependent variable (0-1-2-3). Model performance is evaluated based on criteria such as precision, recall, and F1 score, yielding successful results. This study aims to provide a unique contribution to the existing literature by segmenting firms using unsupervised machine learning methods in addition to considering their financial status and risk data and using the generated segmentation in credit scoring.

İÇİNDEKİLER

Sayfa No.

TABLOLAR LİSTESİ

Sayfa No.

ŞEKİLLER LİSTESİ

Sayfa No.

KISALTMALAR

AR-GE : Araştırma Geliştirme

AI : Artificial Intelligence (Yapay Zekâ)

AUC : Area Under the Curve (Eğri Altındaki Alan)

BIRCH : Balanced Iterative Reducing and Clustering using Hierarchies (Dengeli Yinelemeli Azaltma ve Kümeleme)

DBSCAN : Density-Based Spatial Clustering of Applications with Noise (Gürültüye Dayanıklı Yoğunluk Tabanlı Kümeleme)

GMM : Gaussian Mixture Model (Gauss Karışım Modeli)

KNN : K-Nearest Neighbors (K-En Yakın Komşu)

KOBİ : Küçük ve Orta Büyüklükteki İşletmeler

LightGBM : Light Gradient Boosting Machine (Hafif Gradyan Artırma Modeli)

ML : Machine Learning (Makine Öğrenmesi)

OPTICS : Ordering Points To Identify the Clustering Structure (Noktaları Kümeleme Yapısını Belirlemek İçin Sıralama)

PCA : Principal Component Analysis (Temel Bileşen Analizi)

RF : Random Forest (Rastgele Orman Algoritması)

ROC : Receiver Operating Characteristic (Alıcı İşletim Karakteristik Eğrisi)

SMOTE : Synthetic Minority Over-sampling Technique (Sentetik Azınlık Aşırı Örnekleme Tekniği)

TPR : True Positive Rate (Doğru Pozitif Oranı)

vd. : ve diğerleri

XGBoost : Extreme Gradient Boosting (Aşırı Gradyan Artırma)

SÖZLÜK

Banka: Banka adı altında Türkiye'de kurulan kuruluşlar ile yurtdışında kurulu bankaların Türkiye'deki şubelerini ifade etmektedir.

Denetimli Öğrenme: Etiketlenmiş veri setlerinden öğrenerek tahmin modelleri geliştiren makine öğrenmesi yaklaşımıdır.

Denetimsiz Öğrenme: Etiketlenmemiş verilerden örüntü veya çıkararak analiz yapan makine öğrenmesi yaklaşımıdır.

Gelir Tablosu: Bir firmanın belirli bir dönemde elde ettiği gelirleri, giderleri ve net kârını gösteren finansal rapordur.

Kredi: Bankaların, belirli bir vade ve faiz karşılığında firmalara veya bireylere sağladığı finansal kaynaktır.

Kredi Skoru: Bir firmanın veya bireyin kredi geri ödeme olasılığını temsil eden sayısal bir değerdir.

Makine Öğrenmesi: Verilerden öğrenerek tahmin veya sınıflandırma yapan algoritmalar geliştiren yapay zekâ alt dalıdır.

Öznitelik: Veri setinde bir nesneyi tanımlayan ölçülebilir özelliklerdir

Yapay Zekâ: Makinelerin insan zekâsını taklit ederek öğrenme, analiz ve karar alma yetenekleri kazandığı teknolojidir.

GİRİŞ

Bankacılık sektörü, ekonomilerin işleyişinde vazgeçilmez bir rol oynamaktadır. Bankacılık, finansal kaynakların etkin bir şekilde yönetilmesiyle hem bireylerin hem de firmaların ihtiyaçlarını karşılamaktadır. Türkiye gibi gelişmekte olan bir ekonomide, bankalar özellikle Küçük ve Orta Ölçekli İşletmelerin (KOBİ) büyümesini destekleyen finansal altyapının temel taşlarından birini oluşturmaktadır. Bu sektörün en temel ve önemli faaliyet alanlarından biri, firmalara kredi sağlama süreçleridir. Kredi, işletmelerin sermaye açıklarını kapatmalarına, yeni projelere veya günlük nakit akışlarını düzenlemelerine olanak tanımaktadır. Ancak bu süreç, bankalar için yalnızca bir fırsat değil, aynı zamanda dikkatle yönetilmesi gereken bir risk barındırmaktadır. Bir firmanın krediyi zamanında geri ödeyememesi, bankanın mali dengesini bozabilmektedir. Bu nedenle kredi riskinin doğru bir şekilde analiz edilmesini önem taşımaktadır. Kredi verilen firmaların değerlendirilmesinde; müşteri analizi, sektör analizi ve risk değerlendirmeleri gibi analizler yapılmaktadır. Bu değerlendirmeyi kolaylaştırmak ve daha objektif değerlendirmek amacıyla müşteriler için kredi skorları üretilmektedir.

Kredi skorlama, bir kredi başvurusunun geri ödeme olasılığını tahmin etmeye yarayan bir yöntemi ifade etmektedir. Firmaların mali geçmişleri, mevcut durumlarını ve çeşitli risklerini inceleyerek bankalara yol gösterir. Geleneksel yaklaşımlar, genellikle firmanın bilanço verileri, borç yapısı veya ödeme geçmişi gibi temel göstergelere dayanırken, büyük veri setlerinin karmaşıklığı ve firmalar arasındaki farklılıkların çeşitliliği karşısında yetersiz kalabilmektedir. Türkiye'de bankacılık sektörü, geçmişte yaşanan ekonomik dalgalanmalar ve firmaların heterojen yapıları nedeniyle, risk değerlendirmesinde daha hassas ve esnek araçlara ihtiyaç duymaktadır. Bu ihtiyaç, teknolojinin finansal süreçlere entegrasyonuyla birlikte yeni bir boyut kazanmış ve yapay zekânın, bankacılık sektöründe köklü bir dönüşüme neden olmasının kapılarını aralamıştır.

Yapay zekâ, makinelerin insan zekâsını taklit ederek öğrenme, analiz yapma ve karar alma yeteneklerini geliştirmesiyle, bankacılık sektöründe çığır açan bir yenilik olarak kendini göstermektedir. Yapay zekânın büyük veri setlerini işleme kapasitesi, karmaşık finansal ilişkileri çözme becerisi ve manuel süreçleri

otomatikleştirme avantajlarıyla bankaların operasyonel verimliliğini arttırırken risk yönetiminde daha güvenilir sonuçlar elde edilmesini sağlamıştır.

Yapay zekânın bir alt dalı olan makine öğrenmesi, kredi skorlama süreçlerini daha da ileriye taşıyan bir teknoloji olarak dikkat çekmektedir. Makine öğrenmesi, verilerden öğrenme yeteneğiyle, firmaların finansal verilerindeki gizli örüntüleri ortaya çıkarır ve kredi riskini tahmin etme konusunda geleneksel yöntemlerden çok daha etkili bir performans sergilemektedir. Yapay zekânın gelişimiyle birlikte makine öğrenmesi teknikleri de gelişim göstermiştir. Makine öğrenmesi tekniklerinin büyük veriler karşısında iyi sonuçlar vermesi nedeniyle bankalar tarafından sıklıkla kullanılmaya başlanmıştır. Bankalar müşterilerinin mali tablolarını, gelir raporlarını ve risk durumlarını makine öğrenmesi teknikleri kullanarak ödeme kapasitelerini daha iyi anlamaktadır. Bu nedenlerden ötürü kredi riskinin öngörülmesinde önemli bir sıçrama yaratmıştır. Türkiye'de, özellikle ödeme yapmayan firmaların azınlıkta olduğu dengesiz veri setleriyle karşılaşıldığında, makine öğrenmesi modelleri bu zorlukları aşmak için güçlü bir çözümlerde sunmaktadır.

Bu çalışmada, bankacılık sektöründe makine öğrenmesi ile kredi skorlama süreçlerinin nasıl geliştirilebileceğinin araştırması amaçlamaktadır. Türkiye'deki firmalara yönelik kredi riski analizi özelinde, finansal verilerin (bilanço, gelir tablosu gibi) makine öğrenmesi modellerine entegrasyonu ve bu modellerin performansı üzerinde durulmuştur. Çalışma, firmaların mali durumlarını analiz ederek kredi skorlarını üretmek için hangi algoritmaların daha etkili olduğunu performans kriterlerinin de yardımıyla incelenerek belirlenmiştir. Türkiye'nin bankacılık sektöründeki kendine özgü dinamikleri, özellikle KOBİ'lerin ağırlığı ve ekonomik belirsizlikler, bu araştırmanın temel inceleme alanlarından birini oluşturmaktadır.

Literatür genel olarak incelendiğinde bireysel banka müşterileri için makine öğrenmesi ile kredi skoru çalışmaları olduğu görülmektedir. Ancak bireysel olmayan banka müşterileri (firmalar) için kredi skoru üreten çalışmalar sınırlıdır. Bu çalışmanın makine öğrenmesi teknikleriyle firmalar için kredi skoru üretmesi yönüyle hem sektörel uygulamalara hem de akademik çalışmalara özgün bir katkıda bulunması hedeflenmektedir.

1. BANKACILIK VE KREDİLER

Çalışmanın bu başlığı altında bankacılık, krediler, kredi skorlama ve finansal tablolar kavramları üzerinde durularak makine öğrenmesinin bankacılık sistemi üzerindeki etkisine değinilmiştir.

1.1. Literatür Taraması

Bankacılık sektöründe kredi skorlama süreçlerinin optimizasyonu, finansal risk yönetiminin temel unsurlarından birini oluşturmaktadır. Kredi skorlama hakkında bankalar tarafından çeşitli çalışmalar yapılmaktadır. Aynı zamanda bu konuda akademik çalışmalar bulunmaktadır. Ancak bu konu hakkında genel olarak geleneksel yöntemler kullanılmaktadır. Geleneksel yöntemlerden yapay zekâ ve makine öğrenmesi tabanlı yaklaşımlara geçiş, bu alanda önemli bir dönüşüm yaratmış ve literatürde çeşitli çalışmalarla ele alınmıştır. Bu bölümde, tezin kapsamına uygun olarak literatürde yer alan bankacılıkta kredi skorlama ve makine öğrenmesi uygulamaları incelenmiştir.

Hand ve Henley (1997) tarafından yayımlanan Tüketici kredi puanlamasında İstatistiksel Sınıflandırma Yöntemleri: Bir inceleme, başlıklı çalışma kredi skorlama literatürünün temel taşlarından birini oluşturmaktadır. Bu çalışma, bireysel kredi skorlamasında kullanılan istatistiksel yöntemleri sistematik bir şekilde inceleyerek, bu tekniklerin avantajlarını ve sınırlılıklarını ortaya koymaktadır. Yazarlar, geleneksel yöntemlerin doğrusal varsayımlara dayandığını ve karmaşık veri setlerinde yetersiz kalabileceğini belirtmektedir.

Chawla ve arkadaşları (2002) tarafından geliştirilen SMOTE başlıklı makalede, dengesiz veri setleriyle çalışan kredi skorlama modelleri için büyük bir katkı sağlamıştır. Çalışma, azınlık sınıfı örneklerini sentetik olarak artıran SMOTE algoritmasını tanıtarak, makine öğrenmesi modellerinin performansını iyileştirmeyi amaçlamaktadır. Çalışmada finansal veri setlerinde sıkça görülen sınıf dengesizliği probleminin, standart algoritmaların azınlık sınıfını ihmal etmesine yol açtığını ve SMOTE'un bu sorunu çözmede etkili olduğunu ifade etmektedir.

Lessmann ve arkadaşları (2015) çalışmalarında, kredi skorlamasında modern makine öğrenmesi algoritmalarını geleneksel yöntemlerle karşılaştırmıştır. Araştırmacılar, bu algoritmaların büyük veri setlerinde doğrusal olmayan ilişkileri yakalama kapasitesini test ederek, özellikle finansal risk tahmini için üstün performans sunduğunu ortaya koymaktadır. Bu çalışma, makine öğrenmesi modellerinin doğruluk, kesinlik ve duyarlılık gibi metriklerde geleneksel yöntemleri geride bıraktığını belirmiştir.

Kalaycı (2018) tarafından makine öğrenmesi yöntemleri kullanılarak kredi risk analizi yapılmıştır. KOBİ Şirketlerinin 1 Ocak 2015- 1 Nisan 2016 tarihleri arasında kullandıkları kredi bilgilerini kullanılarak 1 Nisan 2016 - 1 Ekim 2016 tarihleri arasında finansal yaşanma durumu tahmin edilmeye çalışılmıştır. Çalışma kapsamında Lojistik Regresyon, Karar Ağaçları, Rastgele Orman algoritmaları ve Meyilli Hızlandırma yöntemleri kullanılmıştır.

Akpınar (2019) çalışmasında kredi başvuru skor kartının oluşturulmasını makine öğrenmesi yöntemleriyle gerçekleştirmiştir. Çalışmada Lojistik Regresyon, Gradyan Arttırma, Karar Ağaçları ve Rastgele Ormanlar yöntemleri kullanılmıştır. En iyi sonuçlar Gradyan Arttırma modeli ile elde edilmiştir.

Bulut (2019) tarafından tarımsal kredilerin kredi analizinde makine öğrenmesi yöntemleri kullanılmıştır. Çalışmada Karar Ağacı, K-En Yakın Komşu, Naive Bayes, Lojistik Regresyon ve Yapay Sinir Ağları algoritmaları kullanılmıştır.

Türkmen (2021) çalışmasında makine öğrenmesi yöntemleri kullanarak banka pazarlama tahmininde bulunmuştur. Çalışmada Karar Ağacı, K-En Yakın Komşu, Rastgele Orman, Ekstra Ağaçlar, Adaboost, Gradient Boosting gibi yöntemler kullanılmıştır. En iyi sonuçlar Gradyan Artırma Algoritması ve Yapay Sinir Ağları ile elde edilmiştir.

Altın ve Demirci (2022) tarafından çalışmalarında makine öğrenmesi yöntemleri kullanarak Nakit Akış Tablosu Üzerinden Kredi skorlaması yapılmıştır. Çalışmada XGBoost, Gradient Boosting ve Neural Network yöntemleri kullanılmış ve en iyi sonuçlar %80 doğruluk skoru ile XGBoost yöntemi kullanılarak elde edilmiştir.

Sandıkçı ve Şaykol (2024) tarafından çalışmalarında şimdi al sonra öde (BNPL) sistemi üzerine bir davranış skorlama analizi yapılmıştır. Analiz sonucunda

BNPL sistemindeki mevcut durum analiz edilerek, fırsatlar ve zorluklar hakkında skorlama ihtiyacı üzerine bir yaklaşım sunulmuştur.

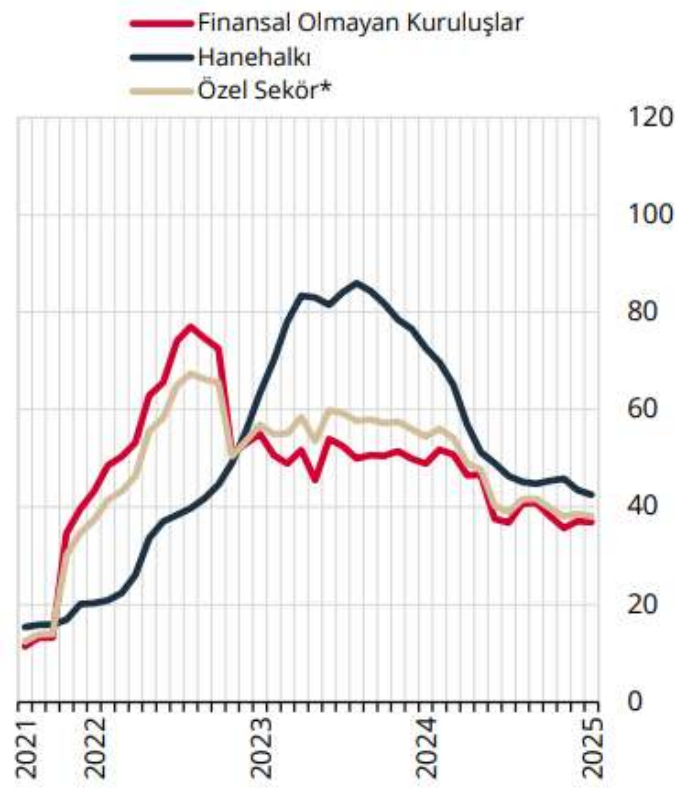
Literatür genel olarak incelendiğinde makine öğrenmesi ile kredi skorlamasın konusunda çeşitli çalışmalar bulunmaktadır. Ancak firmalar için makine öğrenmesi ile kredi skoru üretme konusunda sınırlı sayıda çalışma bulunmaktadır. Bu çalışma ile ticari banka müşterilerinin bilanço, gelir tablosu, kredi hacmi, çalışan sayısı gibi veriler kullanarak önce müşteri segmenti oluşturulmuş daha sonra oluşturulan bu segment modele bağımsız değişken olarak eklenmiş ve denetimli makine öğrenmesi yöntemleri kullanılarak kredi skoru oluşturulmuştur. Segmentasyonun denetimsiz yöntemlerle belirlenmesi ve bu segmentlerin bağımsız değişken olarak denetimli öğrenme algoritmalarına entegre edilmesi, literatürde yeni bir yaklaşımdır. Oluşturulan bu kredi skoru ile bankaların ticari müşterilerine kredi tahsis ederken firmaların risk durumu hakkında bilgi edinmesi ve kredinin tahsis edilmesi konusunda görüşlerine yardımcı olması beklenmektedir. Kullanılan yöntem ile bankaların ticari müşterilere kredi tahsis süreçlerinde karar destek mekanizmalarına katkı sağlanmaktadır.

1.2. Bankacılık

Bankacılık sektörü, modern ekonomilerin temel yapı taşlarından birini oluşturmaktadır. Bankacılık, finansal kaynakların etkin bir şekilde dağıtılmasını sağlayarak bireyler, firmalar ve hükümetler arasında bir finansal köprü oluşturmaktadır (Mishkin, 2019). Bankalar, finansal aracılık, risk yönetimi, ödeme sistemlerinin kolaylaştırılması ve kredi sağlama gibi önemli işlevleriyle ekonomik büyüme ve kalkınmanın sürdürülebilirliğinde kritik bir rol oynamaktadır (Saunders ve Cornett, 2021). Özellikle Türkiye gibi gelişmekte olan ekonomilerdeki bankacılık sektörü Küçük ve Orta Ölçekli İşletmelerin (KOBİ) finansmana erişimini sağlayarak yerel ekonominin canlanmasına katkıda bulunmaktadır. Bu çalışmada, bankacılık sektöründe firmalara yönelik kredi verme ve müşteri kredi skorlama süreçleri ve bu süreçlerin makine öğrenmesi teknikleriyle optimize edilmesi yönleri ele alınmıştır.

Türkiye'de faaliyet gösteren bankalar, türlerine göre üç ana gruba ayrılmaktadır. Bunlar; ticaret (mevduat) bankaları, kalkınma ve yatırım bankaları ile katılım bankalarıdır. Ticaret bankaları kendi içinde kamu, özel, yabancı ve mahalli

bankalar olarak sınıflandırılmaktadır. Kalkınma ve yatırım bankaları ise kamusal sermayeli, özel sermayeli ve yabancı sermayeli olmak üzere üç kategoriye ayrılmaktadır. Bunların yanı sıra, faizsiz bankacılık hizmeti sunan katılım bankaları da Türkiye'de faaliyet göstermektedir (Arabacı, 2018). Ocak 2025 döneminde Türk Bankacılık Sektörünün aktif büyüklüğü 33.366.171 milyon TL olarak gerçekleşmiştir (BDDK, 2025). Şekil 1'de verilen kredilerin 2021-2025 yıllık büyüme oranları görülmektedir.



Şekil 1. Türkiye'deki Kredi Büyümesi

Kaynak: BDDK (2025, 3 Mart). *Aylık bankacılık sektörü verileri*. 3 Mart 2025 tarihinde https://www.bddk.org.tr/Veri/Detay/159 adresinden edinilmiştir.

Bankalar, müşterilerine geniş bir finansal hizmet yelpazesi sunarlar ve bu hizmetlerin temelinde kredi faaliyetleri yer almaktadır. Firmaların sermaye ihtiyaçlarını karşılamak, yeni yatırımları finanse etmek veya nakit akışlarını düzenlemek için bankalardan alınan hizmet, hem mikro hem de makroekonomik dinamikleri etkilemektedir (Berger ve Udell, 2006). Türkiye'de bankacılık sektörü, özellikle 2000'li yıllardan sonra yaşanan yapısal reformlarla güçlendirilmiş ve finansal istikrarın sağlanmasında önemli bir aktör haline getirilmiştir (Özatay, 2011). Bankalar

tarafından tahsis edilen krediler ve bu kredi süreçlerinin etkinliği, firmaların mali verilerinin doğru analiz edilmesi ve risklerin hassas bir şekilde değerlendirilmesi gibi önemli koşullara bağlıdır (Hull, 2018). Türk bankacılık sisteminde, firmalara kredi sağlanırken sektörel riskler, ekonomik dalgalanmalar ve firmaların mali yapıları gibi çok boyutlu faktörler göz önüne alınarak firmalar için çeşitli skorlar üretilmektedir ve kredi tahsis süreçlerinde üretilen bu kredi skorları dikkate alınmaktadır.

Bankacılık sektörünün bir diğer önemli yönü ise risk yönetimi ve finansal sürdürülebilirliktir. Bankalar, kredi portföylerini çeşitlendirerek ve temerrüt risklerini en aza indirerek hem kendilerini hem de müşterilerini koruma altına almaktadır (Saunders ve Cornett, 2021). Türkiye'de, özellikle 2001 krizi sonrasındaki dönemde, bankaların risk yönetimi kapasitelerini artırmak için Basel standartları gibi uluslararası düzenlemeler benimsenerek kredi verme süreçlerinin daha sistematik bir hale gelmesi sağlanmıştır.

Teknolojinin gelişimiyle birlikte bankacılık sektöründe gelişmeler yaşanmıştır. Bu gelişmeler yapay zekâ, veri madenciliği, dijital teknolojiler, nesnelerin interneti, robotik kodlar ve blokzinciri gibi teknolojilerle şekillenmekte olup, bankaların iş süreçlerini ve müşteri hizmetlerini önemli ölçüde dönüştürmektedir. Bu teknolojiler, finansal işlemlerin daha hızlı, güvenli ve verimli bir şekilde gerçekleştirilmesine olanak tanımaktadır (Çelik ve Mangır, 2020). Özellikle son yıllarda banka sayısının da artmasıyla bankalar arasında rekabettin artmış ve bankalar, müşteri deneyiminin iyileştirilesine verilen önem vermiştir. Bu çalışmada, makine öğrenmesi teknikleri, geleneksel yöntemlerin ötesine geçerek bankalara daha doğru ve hızlı risk değerlendirme imkânı sunması hedeflenmektedir.

1.2.1. Kredi

Kredi, bankaların müşterilerine belirli bir vade ve faiz oranı karşılığında sunduğu finansal kaynaktır ve firmalar için işletme sermayesi temini, yatırım finansmanı veya borç yapılandırma gibi çeşitli amaçlarla kullanılmaktadır. Bankalar açısından kredi, temel gelir kaynaklarından biri olmasının yanı sıra önemli bir risk unsuru taşımaktadır. Çünkü kredi geri ödemelerinin zamanında ve tam olarak gerçekleşmesi, bankanın likiditesi, karlılığı ve sermaye yeterliliği için büyük önem

taşımaktadır (Saunders ve Cornett, 2021). Türkiye'de firmalara sağlanan kredilerde özellikle KOBİ'lerin büyümesini desteklemek ve ekonomik kalkınmayı teşvik etmek amacıyla devlet teşvikleriyle de birleştirilmektedir.

Kredi türleri, vade yapısına ve kullanım amacına göre farklılık göstermektedir. Kısa vadeli krediler, genellikle bir yıldan kısa süreli nakit ihtiyaçlarını karşılamak için tasarlanırken uzun vadeli krediler, altyapı projeleri veya makine alımı gibi büyük ölçekli yatırımları finanse etmektedir (Mishkin, 2019). Finansal veri analizinde kredilerin geri ödeme olasılığı, firmanın mali performansı, sektörel riskler ve makroekonomik koşullar gibi çoklu değişkenlere bağlı olmaktadır (Berger ve Udell, 2006).

Kredi verme süreci, bankalar için hem fırsat hem de zorluk barındırmaktadır. Doğru firmalara kredi sağlamak, bankanın karlılığını artırırken, yanlış risk değerlendirmeleri temerrüt riskini ve dolayısıyla mali kayıpları beraberinde getirebilmektedir (Hull, 2018). Türkiye'de bankalar, kredi süreçlerinde bilanço, gelir tablosu ve finansal oran analizi gibi geleneksel yöntemlere başvurmaktadır ancak son yıllarda makine öğrenmesi tabanlı modellerin kullanımıyla bu süreçler daha optimize hale gelmiştir.

1.2.2. Kredi Skoru

Kredi skoru, bir kredi başvurusunun geri ödeme olasılığını değerlendirmek için geliştirilen istatistiksel veya makine öğrenmesi temelli bir yaklaşımı ifade etmektedir. Bankalar, firmaların kredi itibarını belirlemek için bu yöntemi kullanarak risklerini en aza indirmeyi ve kredi portföylerinin kalitesini iyileştirmeyi hedeflemektedir (Thomas vd., 2017). Geleneksel kredi skorlama yöntemleri genellikle lojistik regresyon gibi doğrusal modellere dayanırken, modern yaklaşımlar XGBoost, Rastgele Orman veya LightGBM gibi makine öğrenmesi algoritmalarını benimsemektedir (Lessmann vd., 2015).

Türkiye'de bankacılık sektöründeki kredi skorlama yöntemleri 2000'li yıllardan itibaren teknolojik gelişmelerle birlikte daha yaygın hale gelmiştir. Geleneksel yöntemlerin sınırlılıkları makine öğrenmesi modellerine olan ilgiyi arttırmıştır.

1.3. Finansal Veriler

Finansal veriler, firmaların mali durumlarını operasyonel performanslarını ve risk profillerini analiz etmek için kullanılan verilerdir (Brealey vd., 2014). Bankalar, kredi verme kararlarını desteklemek ve risk yönetimi stratejilerini geliştirmek için bu verilere göre hareket etmektedir. Finansal veriler genellikle mali tablolar, gelir tabloları, nakit akış raporları ve sektörel göstergeler gibi kaynaklardan üretilmektedir. Bankaları kullanan firmalar; Sermaye Piyasası Kurulu (SPK), Türkiye Cumhuriyet Merkez Bankası (TCMB), Bankacılık Düzenleme ve Denetleme Kurumu (BDDK) ve Türk Ticaret Kanunu gibi düzenlemeler gereği bu verileri belirli formatlarda düzenli olarak raporlamaktadır.

Finansal veriler, nicel ve nitel unsurları içerebilmektedir. Bankacılık sektöründe, bu veriler firmaların kredi itibarını değerlendirmek, ödeme kapasitelerini analiz etmek ve olası temerrüt risklerini öngörmek için temel göstergeleri oluşturmaktadır (Altman, 1968). Türkiye'de finansal verilerin kalitesi, firmaların şeffaflık düzeyi ve muhasebe standartlarına uyumu gibi faktörlere bağlı olmaktadır. Makine öğrenmesi modellerinin başarısı, bu verilerin doğruluğuna, eksiksizliğine ve zamanlılığına bağlıdır. Bu nedenle veri ön işleme süreçleri büyük önem taşımaktadır.

Finansal veriler, yalnızca geçmiş performansı değil, aynı zamanda gelecekteki eğilimleri de yansıtabilmektedir. Bu çalışmada, finansal verilerin kredi skorlama modellerine entegrasyonu, bu verilerden türetilen özelliklerin makine öğrenmesi performansına etkisi ve Türkiye'deki uygulamaları üzerinde durulmuştur.

1.3.1. Finansal Tablolar

Finansal tablolar, bir firmanın belirli bir dönemdeki mali durumunu ve performansını özetleyen resmi belgelerdir ve genellikle bilanço, gelir tablosu ve nakit akış tablosu gibi bileşenlerden oluşmaktadır. Bankalar, firmaların finansal tablolarını inceleyerek borç ödeme kapasitelerini, likidite durumlarını ve uzun vadeli finansal sürdürülebilirliklerini değerlendirmektedir (Brealey vd., 2014). Türkiye'de Kamu Gözetimi, Muhasebe ve Denetim Standartları Kurumu (KGK) tarafından belirlenen standartlar, finansal tabloların şeffaflığını ve güvenilirliğini artırmayı amaçlamaktadır.

Bilanço, firmanın varlıklarını yükümlülüklerini ve öz sermayesini detaylandırmaktadır. Bu tablo, firmanın mali yapısını anlamak için temel bir kaynaktır ve borç alacak oranı gibi önemli oranların hesaplanmasında kullanılmaktadır (Altman, 1968).

Finansal tabloların analizi, yalnızca sayısal verilere dayanmaz; aynı zamanda firmanın sektörel konumu ve piyasa koşulları gibi nitel faktörler de dikkate alınmaktadır. Makine öğrenmesi modellerinde, bu tablolardan türetilen özellikler risk tahmini için önemli özellikleri oluşturmaktadır. Türkiye'de finansal tabloların standartlaştırılması, veri kalitesini artırarak bu tür analizlerin doğruluğunu desteklemektedir.

1.3.2. Gelir Tablosu

Gelir tablosu, bir firmanın belirli bir dönemde elde ettiği gelirleri, giderleri, net kâr veya zararını gösteren finansal bir rapordur. Bankalar, firmaların gelir tablolarını analiz ederek operasyonel performanslarını, kârlılıklarını ve nakit akışlarını değerlendirmektedir. Gelir Tablosu Kredi geri ödeme kapasitesini doğrudan etkileyen bir unsurdur (Brealey vd., 2020). Türkiye'de, gelir tabloları genellikle yıllık veya çeyreklik dönemlerde hazırlanır ve firmaların mali durumlarını periyotlar halinde gösterir.

Makine öğrenmesi modellerinde, gelir tablosundan türetilen özellikler kredi riskini tahmin etmek için önemli birer özelliktir (Han vd., 2011). Türkiye'de gelir tablosu verilerinin doğruluğu, firmaların muhasebe politikalarına ve dış denetim süreçlerine bağlıdır. Bankalar tarafından bu tabloların doğruluğu incelenerek analiz edilmektedir.

2. MAKİNE ÖĞRENMESİ

Teknolojinin gelişimiyle birlikte yapay zekânın kullanım alanları artarak yayılmıştır. Makine öğrenmesi de yapay zekânın alt dallarından birini oluşturmaktadır. Bu bölümde yapay zekâ kavramı, denetimli ve denetimsiz makine öğrenmesi algoritmaları ve makine öğrenmesi model seçiminde kullanılan performans kriterlerine değinilmiştir.

2.1. Yapay Zekâ Kavramı

Yapay zekâ, insan benzeri düşünme süreçlerini taklit eden teknolojiyi ifade etmektedir. Bu teknoloji, büyük miktarda veriyi analiz ederek öğrenme, problem çözme ve karar verme yeteneklerine sahiptir. Yapay zekânın temel amacı, insanların karmaşık problemleri çözmesine yardımcı olmak üzere geliştirilen otomatik sistemler olarak ifade edilebilir. Yapay zekâ makine öğrenmesi, derin öğrenme, doğal dil işleme gibi alt dallara ayrılmaktadır. Bu alt dallar birçok alanda kullanılmaktadır (Russell ve Norvig, 2016). Günümüzde sağlık, finans, lojistik ve üretim gibi alanlarda yapay zekânın kullanımı artarak yaygınlaşmaktadır. Özellikle finans sektöründe kredi skorlama ve risk yönetimi gibi uygulamalarda büyük bir dönüşüm ortaya çıkarmıştır.

2.2. Makine Öğrenmesi Türleri

Makine öğrenmesi, verilerden öğrenme yeteneği kazandırılmış algoritmaların uygulanmasını ve tasarlanmasını kapsayan bir disiplin olarak ifade edilebilir. Makine öğrenmesi denetimsiz öğrenme, denetimli öğrenme ve takviyeli öğrenme olarak üç ana bölüme ayrılmaktadır (Bishop, 2006). Bu bölümde, bankanın müşterisi olan firmaların ve şirketlerin finansal ve kredi verileri analiz edilerek kredi skoru üretmek amacıyla denetimsiz öğrenme teknikleri açıklanmıştır. Denetimsiz öğrenme, etiketlenmemiş veriler üzerinde çalışarak ve veri setindeki gizli yapıları, kümeleri ya da örüntüleri keşfetmeyi amaçlamaktadır (James vd., 2013). Finansal anlamda, bu yöntemler firmaların mali yapılarını anlamak, risk gruplarını belirlemek veya olağandışı davranışları tespit etmek gibi amaçlarla kullanılabilmektedir.

Makine öğrenmesi algoritmalarının kullanım sürecinde, veri setinin yapısına ve çözülmek istenen problemler dikkate alınarak seçimler yapılır. Denetimsiz öğrenme, veri setinde daha önceden tanımlanmış bir değişken bulunmadan çalıştığından dolayı için esneklik sağlar ve finansal verilerdeki belirsizlikleri ve karmaşık durumları ortaya çıkarmak için ideal bir yöntemdir (Hastie vd., 2009). Örnek olarak bir bankada, firmaların ödeme yöntemlerini veya nakit akışlarını analiz ederek hangi firmaların benzer mali özellikler taşıdığını anlamak istenebilir. Bu tür analizler, kredi riskini değerlendirme süreçlerinde önemli bir temel oluşturmaktadır. Aşağıda, denetimsiz makine öğrenmesi ve bu kapsamda kullanılan temel algoritmalar ayrıntılı bir şekilde ele alınarak açıklanmıştır.

2.2.1. Denetimsiz Makine Öğrenmesi

Denetimsiz makine öğrenmesi, etiketlenmemiş verilerle çalışarak veri setindeki temel yapıları ya da örüntüleri açıklamayı hedeflemektedir. Bu yöntem, kümeleme, boyut indirgeme veya aykırı değer tespiti gibi görevler için kullanılmaktadır (James vd., 2013). Finansal verilerle çalışırken, denetimsiz öğrenme, firmaların mali durumların sınıflandırılması, risk profillerinin anlaşılması veya beklenmedik mali davranışların belirlenmesi gibi çalışmalar uygulamalar için önemli rol oynamaktadır. Bir banka, firmaların borç-alacak oranları, nakit akışları ve gelirlerini detaylı incelenerek benzer özelliklere sahip firmaları gruplara ayırarak organize edebilir. Bu gruplar, kredi riskini değerlendirme veya stratejik karar alma süreçlerinde yol gösterme konusunda destek sağlayabilmektedir (Hastie vd., 2009).

Denetimsiz öğrenmenin sağladığı en önemli avantajı, veri setinde önceden belirlenmiş bir çıktı gerektirmemesidir. Bu durum, özellikle finansal veriler gibi karmaşık ve çok boyutlu veri setlerinde keşifsel analizler için büyük bir esneklik sağlayarak avantajlı hale gelmektedir. Fakat bu yöntemlerin sonuçlarının yorumlanabilmesi için uzmanlık bilgisi gerektirmektedir ve algoritmaların performansı veri setinin yapısına göre değişkenlik gösterebilmektedir (Bishop, 2006). Finansal veriler, genellikle doğrusal olmayan ilişkiler ve aykırı değerler içermektedir. Böylece algoritma seçimi ve parametre optimizasyonu önemli hale gelmektedir.

Aşağıda, denetimsiz öğrenme kapsamında sıklıkla kullanılan altı algoritma ayrıntılı olarak açıklanmıştır.

2.2.2. Denetimli Makine Öğrenmesi

Denetimli makine öğrenmesi, etiketlenmiş verilerle çalışır ve verinin özellikleri ile hedef değişken arasındaki ilişkiyi modellemeyi hedeflemektedir. Finansal bağlamda, bu yöntemler firmaların geçmiş performanslarına dayanarak gelecekteki kredi risklerini öngörmek için güçlü bir araç olarak edilebilmektedir. Denetimli makine öğrenmesi, finansal analizlerde, firmaların geçmiş mali performanslarını temel alarak gelecekteki davranışlarını tahmin etmek için yaygın olarak kullanılmaktadır. Bir banka, firmaların borç oranları, nakit akışları ve gelirlerini özellik olarak alıp, kredi geri ödeme olasılıklarını hedef olarak öngörebilmektedir. Denetimli öğrenme, bu tür veri odaklı karar alma süreçlerinde hem doğruluk hem de genelleştirme kapasitesiyle öne çıkmaktadır (Russell ve Norvig, 2016).

Denetimli makine öğrenmesi algoritmaları, genellikle sınıflandırma veya regresyon problemleri için kullanılır ve finansal verilerin hem yapılandırılmış hem de karmaşık doğasına uyum sağlayabilmektedir (Hastie vd., 2009). Algoritma seçimi, veri setinin özellikleri ve problemin gereksinimlerine bağlı olmaktadır (Mitchell, 1997). Bu nedenle, bu çalışmada açıklanan yöntemler finansal risk analizi gibi hassas uygulamalar için dikkatlice değerlendirilmiştir. Denetimli makine öğrenmesi kapsamında beş algoritma; XGBoost, Karar Ağacı, Rastgele Orman, KNN, LightGBM detaylı bir şekilde incelenmiştir.

Denetimli öğrenmenin avantajı, açıkça tanımlanmış bir hedef değişkenle çalışması olarak ifade edilebilmektedir. Finansal risk değerlendirmesi gibi net sonuçların beklendiği alanlarda büyük bir kolaylık sağlamaktadır. Ancak, modelin başarısı veri kalitesine, özellik mühendisliğine ve algoritmanın problemle uyumuna bağlı olmaktadır (Kohavi ve Provost, 1998). Finansal veriler genellikle gürültülü, eksik veya dengesiz olabilmektedir. Bu nedenle, seçilen algoritmaların bu zorluklarla başa çıkabilmesi kritik bir önem taşımaktadır (Domingos, 2012).

2.3. Denetimsiz Makine Öğrenmesi Algoritmaları

Denetimsiz makine öğrenmesi, etiketlenmemiş verilerle çalışarak veri setindeki temel yapıları göre kümeleme oluşturmaktadır. Bu bölümde KMeans, Hiyerarşik, DBSCAN, BIRCH, OPTICS ve GMM algoritmaları üzerinde durulmuştur.

2.3.1. KMeans Algoritması

K-Means, veri noktalarını belirli bir k sayısında kümeye ayırmak için tasarlanmış olan, basit ve etkili bir kümeleme algoritmasıdır. Algoritma şu adımları izler: Öncelikle, k sayıda rastgele merkez seçilir, ardından her veri noktası en yakın merkeze etiketlenir ve bu etiketlendirmelere dayanarak merkezler güncellenir. Bu adımlar, kümeler sabitlenene kadar tekrarlanmaktadır (James vd., 2013). K-Means algoritmasının temel hedefi, kümeler içindeki varyansı en aza indirmektir. Bu hedef maliyet fonksiyonunu minimize ederek gerçekleştirilmektedir.

$$J = \sum_{i=1}^{K} \sum_{\aleph \epsilon C_i} \|\aleph - \mu_i\|^2 \qquad (2.1)$$

Denklem 2.1'deki; K Küme Sayısını, C_i i numaralı kümede bulunan noktaların kümesini, $\aleph$ kümede yer alan bir veri noktasını, μ_i i numaralı kümenin merkezini, $\|\aleph - \mu_i\|^2$ veri noktalarının küme merkezine olan uzaklığın karesini ifade etmektedir (MacQueen 1967).

Finansal verilerde, K-Means algoritması firmaları risk profillerine göre sınıflandırmak için kullanılabilmektedir. Bir banka, firmaların borç-alacak oranları, nakit akışları ve gelirlerini analiz ederek düşük, orta ve yüksek riskli kümeler oluşturabilir. Bu kümeler, kredi verme politikalarını şekillendirmede veya risk yönetimi stratejilerinde rehber olabilmektedir.

K-Means algoritmasının uygulanması sırasında, k değerinin önceden belirlenmesi bir zorluk olarak ortaya çıkarmaktadır. Bu sorunu düzeltmek için, dirsek yöntemi, Silhoutte Skoru, Davies-Bouldin ve Calinski-Harabasz gibi teknikler kullanılır. Dirsek yönteminde, J değerinin k'ye göre grafiği çizilir ve eğimin belirgin şekilde azaldığı nokta optimum k olarak seçilmektedir. (Hastie vd., 2009). Veri incelendiğinde, 3 kümeli bir yapı (düşük, orta, yüksek risk) başlangıçta mantıklı görünebilir, ancak veri setinin yapısına bağlı olarak bu sayı 5 veya daha fazla olabilir.

K-Means'in avantajları arasında hesaplama hızı ve ölçeklenebilirlik bulunmaktadır. Algoritma, büyük veri setlerinde de hızlı sonuçlar üretebilmektedir. Çok fazla firmanın analiz edildiği bir bankacılık senaryosunda önemli bir özelliktir. Ancak, K-Means'in küme şekillerini dairesel varsayması bir sınırlılık olarak karşımıza çıkar. Finansal veriler genellikle doğrusal olmayan ilişkiler içerdiğinden, bu durum algoritmanın performansını etkileyebilmektedir. Örnek olarak, firmaların mali özellikleri düzenli bir şekilde kümelenmeyebilir; bu durumda K-Means yerine daha esnek algoritmalar düşünülebilir. Seçilen denetimsiz makine öğrenme algoritmaları arasından Silhoutte Skoru, Davies-Bouldin ve Calinski-Harabasz gibi metrikler kullanılarak seçim yapılabilmektedir.

2.3.2. Hiyerarşik Algoritması

Hiyerarşik kümeleme, veri noktalarını bir ağaç yapısı şeklinde organize eden bir algoritmadır ve genellikle birleştirici veya bölücü yaklaşımla uygulanmaktadır. Birleştirici yöntemde, her veri noktası başlangıçta tek bir küme olarak kabul edilir ve benzerlik ölçüsüne göre adım adım birleştirilmektedir (Bishop, 2006).

Finansal verilerde, hiyerarşik kümeleme firmaların mali benzerliklerini hiyerarşik bir yapıda analiz etmek için oldukça yararlıdır. Hiyerarşik kümelemenin en önemli avantajlarından biri de küme sayısının önceden belirlenmesi gerekmemesidir. Ağaç yapısı, analizciye esneklik sağlar ve ağaç sayısı istenilen bir seviyede kesilerek farklı küme sayıları elde edilebilir (James vd., 2013). Bir banka, firmaları 5 ana kümeye ayırmayı tercih edebilir, ancak daha detaylı bir analiz için bu sayıyı 10'a çıkarabilmektedir. Bu esneklik, finansal verilerin dinamik yapısına uyum sağlayarak karar vericilere farklı bakış açıları sunmaktadır.

Ancak, hiyerarşik kümelemenin hesaplama maliyeti büyük veri setlerinde önemli bir dezavantaj oluşturmaktadır. Binlerce firmanın analiz edildiği bir senaryoda, algoritmanın çalışma süresi ve bellek kullanımı ciddi bir sorun haline gelebilmektedir. Bu nedenlerden dolayı, hiyerarşik kümeleme genellikle daha küçük veri setleri veya ön analizler için tercih edilmektedir.

2.3.3. DBSCAN Algoritması

Density-Based Spatial Clustering of Applications with Noise (DBSCAN), yoğunluk temelli bir kümeleme algoritmasıdır ve veri setindeki kümeleri yoğunluk farklarına göre tanımlamaktadır. Algoritma, iki temel parametreyle çalışmaktadır. ϵ (mesafe eşiği) ve $MinPt_s$ (minimum nokta sayısı). İşleyişi aşağıdaki şekilde ifade edilmektedir (Ester vd., 1996).

Bir nokta, ϵ mesafesi içindeki komşu sayısı, $MinPt_s$'den büyükse çekirdek nokta olarak sınıflandırılır. Çekirdek noktaların komşuluğundaki diğer noktalar sınır noktalarıdır. Bu kriterleri karşılamayan noktalar gürültü olarak etiketlenir.

Finansal verilerde, DBSCAN aykırı değerleri tespit etmek için son derece etkili bir yöntemdir. DBSCAN'ın geleneksel kümeleme yöntemlerinden farkı, küme şekillerini dairesel varsaymamasıdır. Bu varsayım finansal veriler gibi düzensiz yapıdaki veri setlerinde büyük bir avantaj sağlamaktadır. Anca ϵ ve $MinPt_s$ parametrelerinin doğru seçimi kritik bir öneme sahip olmaktadır. Yanlış değerlerin seçilmesi durumunda, algoritma çok fazla gürültü ya da çok az küme üretebilir. Finansal verilerde, bu parametreler genellikle veri setinin yoğunluk dağılımına göre optimize edilmektedir

2.3.4. BIRCH Algoritması

(Balanced Iterative Reducing and Clustering using Hierarchies) BIRCH, büyük veri setleri için tasarlanmış bir kümeleme algoritması olarak ifade edilebilmektedir (Zhang vd., 1996). Bu algoritma, veriyi özetleyen küçük yapılar oluşturarak bu özetleri bir ağaç yapısında organize eder. Bir ağaç yapısı aşağıdaki bileşenlerden oluşmaktadır.

$$CF = (N,\ \sum_{i=1}^{N} X_i\ , \sum_{i=1}^{N} X_i^2) \qquad (2.2)$$

Denklem 2.2'de ifade edilen; N kümedeki veri sayısını, $\sum_{i=1}^{N} X_i$ kümedeki tüm veri noktalarının toplamını, $\sum_{i=1}^{N} X_i^2)$ kümedeki veri noktalarının kareleri toplamını ifade etmektedir (Zhang vd., 1996.)

Finansal verilerde, BIRCH büyük hacimli firma verilerini hızlıca kümelemek için kullanılabilmektedir. BIRCH algoritmasını kullanarak hem bellek kullanımını optimize edebilirken işlem süresi de kısaltabilmektedir. BIRCH algoritmasının en

büyük avantajı, büyük veri setlerinde etkin bir şekilde çalışabilmesidir. Ancak, algoritma tüm veri setinde kümeleme için uygundur ve yoğunluk farklarına duyarlı değildir. Bu durum, finansal verilerde heterojen yapılarla karşılaşıldığında bir sınırlılık olabilmektedir. Firmaların mali özellikleri farklı yoğunluk bölgelerinde yer alıyorsa, BIRCH bu farklılıkları tam olarak ayrıştıramamaktadır.

2.3.5. OPTICS Algoritması

Ordering Points To Identify the Clustering Structure (OPTICS), DBSCAN'ın genelleştirilmiş bir versiyonudur ve farklı yoğunluk seviyelerine sahip kümeleri tespit edebilmektedir (Ankerst vd., 1999). OPTICS algoritması, her nokta için çekirdek mesafe ve ulaşılabilirlik mesafesini hesaplamaktadır. Bu ölçümler, veri noktalarını bir sıralı yapı içinde organize ederek hiyerarşik kümeler oluşturmaktadır.

Finansal verilerde, OPTICS firmaların mali yapılarının heterojen olduğu durumlarda kullanılabilir. Bazı firmalar yoğun bir küme içinde yer alırken, diğerleri daha seyrek bir yapıda olabilir. OPTICS algoritması, bu farklılıkları yakalayarak firmaları esnek bir şekilde gruplandırabilmektedir. OPTICS algoritmasının avantajı, küme sayısını önceden belirleme gereksiniminin olmaması ve hiyerarşik bir çıktı sunmasıdır. Ancak, algoritmanın karmaşıklığı ve hesaplama maliyeti, büyük veri setlerinde uygulanmasını zorlaştırabilmektedir. Finansal uygulamalarda, OPTICS genellikle daha küçük veya orta ölçekli veri setleriyle sınırlı kalabilmektedir.

2.3.6. GMM Algoritması

Gaussian Mixture Model (GMM), veri noktalarının Gauss dağılımlarının bir karışımıyla modellendiği olasılıksal bir kümeleme yöntemi olarak ifade edilmektedir (Bishop, 2006).

Finansal verilerde, GMM algoritması firmaların risk profillerini olasılıksal bir şekilde sınıflandırmak için kullanılabilmektedir. Bir firmanın düşük riskli kümeye ait olma olasılığı %70, yüksek riskli kümeye ait olma olasılığı %20 olabilir. Bu yumuşak kümeleme özelliği, K-Means gibi sert kümeleme yöntemlerine kıyasla daha esnek bir analiz sunmaktadır.

GMM'nin avantajı ise veri setindeki belirsizlikleri olasılıksal bir çerçevede ele almasıdır. Ancak, modelin başarısı Gauss dağılımı varsayımına bağlıdır ve finansal veriler bu varsayımı karşılamıyorsa performans düşebilmektedir.

2.4. Denetimli Makine Öğrenmesi Algoritmaları

Denetimli makine öğrenmesi, etiketlenmiş verilerle çalışır ve verinin özellikleri ile hedef değişken arasındaki ilişkiyi modellemektedir. Bu bölümde XGBoost, Karar Ağaçları, Rastgele Orman, KNN ve LightGBM algoritmaları üzerinde durulmuştur.

2.4.1. XGBoost Algoritması

(Extreme Gradient Boosting) XGBoost, gradyan artırma prensibine dayalı güçlü bir denetimli öğrenme algoritmasıdır ve yüksek doğruluk gerektiren problemlerde sıkça tercih edilmektedir (Chen ve Guestrin, 2016). Zayıf karar ağaçları ardışık bir şekilde birleştirilerek çalışır ve her adımda önceki modellerin hatalarını düzeltmektedir. XGBoost algoritmasının temel hedefi, denklem 2.3'teki kayıp fonksiyonunu minimize etmektir.

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) \sum_{k=1}^{K} \phi(f_k) \qquad (2.3)$$

Denklem 2.3'teki $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ gerçek değer ile tahmin edilen y_i arasındaki kaybı, $\phi(f_k)$ k.ağacın karmaşıklığını cezalandıran düzenleştirme terimini, n veri noktası sayısını, K ise ağaç sayısını ifade etmektedir (Chen ve Guestrin, 2016).

Finansal verilerde, XGBoost algoritması karmaşık ilişkileri modellemek için son derece etkili bir yöntemdir. Bir firmanın kredi riskini değerlendirirken borç-alacak oranı, nakit akışı ve geçmiş ödeme performansı gibi özellikler arasındaki doğrusal olmayan bağlantıları yakalayabilmektedir (Friedman, 2001).

XGBoost algoritmasının gücü, düzenlileştirme mekanizmalarından ve eksik verilerle başa çıkma yeteneğinden gelmektedir. Bu özellikler, aşırı öğrenme riskini azaltırken modelin genelleştirme kapasitesini arttırmaktır (Chen ve Guestrin, 2016). Ayrıca, özellik önem sıralaması sunması, finansal analizlerde hangi mali göstergelerin

daha etkili olduğunu anlamayı sağlamaktadır (Pedregosa vd., 2011). Ancak, algoritmanın parametre optimizasyonu zaman alıcı olabilmektedir. Özellikle büyük veri setlerinde hesaplama maliyeti yüksektir (Witten vd., 2016). Finansal risk analizinde, XGBoost genellikle performans odaklı senaryolarda tercih edilmektedir.

2.4.2. Karar Ağaçları Algoritması

Karar ağaçları, veri setini dallara ayırarak karar kuralları oluşturan temel bir denetimli öğrenme algoritmasıdır. Her dal, bir özellik üzerindeki bir eşik değerine dayanır ve bu bölünmeler bir kayıp ölçütüne (Gini indeksi) göre optimize edilmektedir. Gini indeksi denklem 2.4'teki gibi hesaplanmaktadır.

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \qquad (2.4)$$

Denklem 2.4'teki p_i i. sınıfın olasığılını, C sınıf sayısını ifade etmektedir (Breiman vd., 2017).

Finansal verilerde, karar ağaçları firmaların kredi riskini değerlendirmek için açıklayıcı bir model sunmaktadır. Karar ağaçlarının en büyük avantajı, yorumlanabilir olmasıdır. Bu yorumlanabilirlik, finansal risk değerlendirmelerinde şeffaflık arayanlar için kritik bir özellik oluşturmaktadır (Rokach ve Maimon, 2008). Ancak, tek bir ağaç genellikle aşırı öğrenme durumuna yatkındır ve eğitim verisine aşırı uyum sağlayarak genelleştirme yeteneğini kaybedebilmektedir. Finansal verilerde, bir ağaç yalnızca geçmişteki belirli firmalara özgü kurallar üretebilir ve yeni verilerle başarısız olabilmektedir. Bu sınırlılığı aşmak için, karar ağaçları genellikle toplu yöntemler içinde kullanılmaktadır (Breiman vd., 2017).

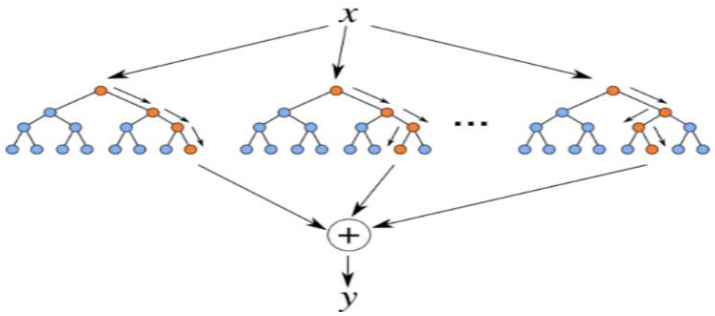
2.4.3. Rastgele Orman Algoritması

Rastgele Orman algoritması, çoklu karar ağaçlarının birleşiminden oluşan bir toplu öğrenme yöntemidir ve varyansı azaltarak genelleştirme performansını arttırmaktadır (Breiman, 2001). Algoritma, veri setinin rastgele alt kümeleri ve özelliklerin rastgele bir seçimiyle çok sayıda ağaç eğitmektedir. Şekil 2'de birden fazla ağacın bir araya gelmesiyle en yüksek puan alan yöntemin seçilmesi görülmektedir. Nihai tahmin, ağaçların sınıflandırma veya ortalaması regresyon ile belirlenir.

Matematiksel olarak, bir rastgele ormanın tahmini denklem 2.5'teki gibi ifade edilmektedir.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(\chi) \qquad (2.5)$$

Denklem 2.5'te; $\hat{y}$ nihai tahmini, $h_t(\chi)$ t. Ağacın tahminini, T ağaç sayısını ifade etmektedir (Breiman, 2001).



Şekil 2. Rastgele Orman Algoritmasının Yapısı

Kaynak: Zilyas, D., ve Yılmaz, A. (2023). Makine öğrenmesi yöntemleri ile eğitim başarısının tahmini modeli. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 14*(3), 437-447.

Finansal verilerde, rastgele orman aşırı öğrenme riskini azaltarak güvenilir kredi skoru tahminleri üretebilmektedir. Bir banka, firmaların kredi riskini değerlendirirken yüzlerce ağacın ortak kararına dayanarak daha dengeli bir skor elde edebilmektedir (Ho, 1995).

Rastgele orman algoritmasının avantajı, hem doğruluk hem de stabilite sunması olarak ifade edilebilmektedir. Ayrıca, özellik önem sıralaması sağlayarak hangi mali göstergelerin kredi riski üzerinde daha etkili olduğunu belirtebilmektedir (Breiman, 2001). Ancak, çok sayıda ağaç kullanılması hesaplama maliyetini arttırarak modelin yorumlanabilirliği tek bir karar ağacına kıyasla azalabilmektedir. Finansal analizlerde, rastgele orman genellikle büyük veri setleriyle çalışırken ve yüksek doğruluk arandığında tercih edilmektedir (Liaw ve Wiener, 2002).

2.4.4. KNN Algoritması

K-En Yakın Komşu (KNN), bir veri noktasını en yakın k komşusuna göre sınıflandıran bir denetimli öğrenme algoritması olarak ifade edilmektedir (Cover ve Hart, 1967). Mesafe genellikle denklem 2.6'daki gibi Öklid formülüyle hesaplanmaktadır.

$$d\,(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2.6)$$

Denklem 2.6'daki; x, y iki veri noktasını, x_i ve y_i i. özellik değeri, n ise özellik sayısını göstermektedir (Cover ve Hart, 1967).

Finansal verilerde, KNN algoritması benzer mali özelliklere sahip firmaları karşılaştırmak için kullanılabilmektedir. Bir banka, bir firmanın borç-alacak oranı ve nakit akışını geçmişteki benzer firmalarla eşleştirerek kredi riskini tahmin edebilmektedir (Altman, 1968). KNN algoritmasının avantajı, basitliği ve herhangi bir eğitim aşaması gerektirmemesidir. Tahminler doğrudan veri setine dayanmaktadır. Ancak, büyük veri setlerinde hesaplama maliyeti yüksektir çünkü her tahmin için tüm veri setiyle mesafe hesaplaması yapılmaktadır. Ayrıca, k değerinin seçimi performansı doğrudan etkilemektedir. Çok küçük bir k gürültüye duyarlı hale getirirken, çok büyük bir k genelleştirmeyi zorlaştırabilmektedir (Duda vd., 2006).

2.4.5. LightGBM Algoritması

LightGBM (Light Gradient Boosting Machine), büyük veri setlerinde hızlı ve verimli bir gradyan artırma algoritmasıdır (Ke vd., 2017). XGBoost'a benzer şekilde çalışmaktadır. Ancak histogram tabanlı öğrenme ve yaprak bazlı büyüme gibi yenilikçi yaklaşımlar kullanmaktadır. Histogram tabanlı öğrenme, sürekli özellik değerlerini ayrık kutulara bölerek hesaplama maliyetini azaltmaktadır. Kayıp fonksiyonu XGBoost ile aynıdır.

Finansal verilerde, LightGBM yüksek hacimli veri setleriyle çalışmak için kullanılabilmektedir. Bir banka, binlerce firmanın mali verilerini analiz ederek kredi skoru üretmek istediğinde, LightGBM hızlı eğitim süreleri ve yüksek doğruluk sunmaktadır.

LightGBM algoritmasının avantajı, büyük veri setlerinde ve yüksek boyutlu özellik uzaylarında etkin bir şekilde çalışabilmesi olarak ifade edilebilmektedir. Ayrıca, eksik verilerle başa çıkma yeteneği ve paralel işlem desteği finansal analizlerde pratik bir kolaylık sağlamaktadır. Ancak, yaprak bazlı büyüme stratejisi aşırı öğrenme riskini artırabilmektedir. Bu durum dikkatli parametre optimizasyonu gerektirmektedir (Friedman vd., 2000).

2.5. Performans Kriterleri

Makine öğrenmesi modellerinin performansını değerlendirmek, finansal veri analizi gibi yüksek hassasiyet gerektiren alanlarda temel bir gereklilik olarak ifade edilebilir. Banka müşterisi firmaların finansal ve kredi verilerini analiz ederek kredi skoru üretmek için geliştirilen modellerin etkinliği, yalnızca genel tahmin başarısına değil, aynı zamanda farklı sınıfların doğru bir şekilde ayırt edilmesine bağlıdır (Provost ve Fawcett, 2013). Performans kriterleri, modelin güçlü yönlerini ve zayıflıklarını sistematik bir şekilde analiz etmektedir. Bu analizler finansal risk yönetiminde stratejik karar alma süreçlerini desteklemektedir (Han vd., 2011). Çalışmanın bu bölümde, denetimli makine öğrenmesi modellerinin performansını ölçmek için kullanılan temel metrikler ve araçlar anlatılmıştır.

2.5.1. Kesinlik

Kesinlik, bir modelin pozitif olarak sınıflandırdığı tahminler arasında gerçekte pozitif olanların oranını ölçmektedir. Matematiksel olarak denklem 2.7'deki gibi ifade edilmektedir.

$$\text{Kesinlik} = TP / (TP + FP) \qquad (2.7)$$

Denklem 2.7'deki; TP (Doğru Pozitifler) modelin doğru bir şekilde pozitif olarak sınıflandırdığı örnekleri, FP (Yanlış Pozitifler) modelin yanlışlıkla pozitif olarak sınıflandırdığı örnekleri göstermektedir.

Finansal analizde kesinlik, modelin ödeme yapmayan firmaları tespit ederken ne kadar güvenilir olduğunu gösterir. Yüksek bir kesinlik değeri, yanlış pozitif

oranının düşük olduğunu ifade eder; bu, bir bankanın gereksiz yere kredi reddetme riskini azaltır (Sokolova ve Lapalme, 2009).

2.5.2. Duyarlılık

Duyarlılık, gerçekten pozitif olan örnekler arasında modelin doğru tahmin ettiği pozitiflerin oranını ölçmektedir. Hesaplama yöntemi denklem 2.8'deki gibidir.

$$\text{Duyarlılık} = TP / (TP + FN) \quad (2.8)$$

Denklemde 2.8'de; TP (Doğru Pozitifler) modelin doğru bir şekilde pozitif olarak sınıflandırdığı örnekleri, FN (Yanlış Negatifler) modelin yanlışlıkla negatif olarak sınıflandırdığı pozitif örnekleri ifade etmektedir.

Finansal risk analizinde duyarlılık, modelin ödeme yapmayan firmaları ne kadar iyi yakaladığını belirtmektedir. Yüksek duyarlılık, yanlış negatif oranını düşürerek bir bankanın potansiyel temerrüt risklerini gözden kaçırma olasılığını azaltmaktadır (Fernández vd., 2018).

2.5.3. F1 Skoru

F1 skoru, kesinlik ve duyarlılık arasındaki harmonik ortalamayı temsil etmektedir. Dengesiz veri setlerinde model performansını değerlendirmek için sıkça kullanılmaktadır. F1 skorunun hesaplanma yöntemi denklem 2.9'daki gibidir.

$$F1 = 2 * (\text{Kesinlik} * \text{Duyarlılık}) / (\text{Kesinlik} + \text{Duyarlılık}) \quad (2.9)$$

F1 skoru, kesinlik ve duyarlılık arasında bir denge sağlamaktadır. Bu denge, finansal veri setlerinde azınlık sınıfının doğru tahmin edilmesi gereken durumlarda çok önemlidir (Chawla vd., 2002). Finansal risk modellemesinde F1 skoru, yanlış pozitiflerden ve yanlış negatiflerden kaçınmaya yardımcı olmaktadır (Provost ve Fawcett, 2013).

2.5.4. Doğruluk

Doğruluk, modelin tüm tahminler arasında doğru yaptığı tahminlerin oranını ölçmektedir. İlgili işlemlerin denklemi 2.10'daki gibi ifade edilmektedir.

$$\text{Doğruluk} = (TP + TN) / (TP + TN + FP + FN) \quad (2.10)$$

Doğruluk, modelin genel performansını yansıtır ve dengeli veri setlerinde anlamlıdır. Ancak, finansal veri setlerinde sınıf dengesizliği yaygın olduğundan doğruluk yanıltıcı olabilmektedir (He ve Garcia, 2009). Bu nedenle, model performans analizinde doğruluk diğer metriklerle birlikte kullanılmalıdır (James vd., 2013).

2.5.5. Makro Ortalama

Makro ortalama, her sınıfın performans metriklerinin aritmetik ortalamasını alarak sınıflara eşit ağırlık vermektedir. Makro kesinlik denklem 2.11'deki gibi hesaplanmaktadır.

$$\text{Makro Kesinlik} = (\text{Kesinlik\_pozitif} + \text{Kesinlik\_negatif}) / 2 \quad (2.11)$$

Makro ortalama, azınlık ve çoğunluk sınıflarının performansını dengeli bir şekilde değerlendirmektedir (Sokolova ve Lapalme, 2009). Bu metrik, sınıf dengesizliği durumunda azınlık sınıfının performansını göz ardı etmeyi önlemektedir. Ancak sınıf büyüklüklerini dikkate almamaktadır (Fernández vd., 2018).

2.5.6. Ağırlıklı Ortalama

Ağırlıklı ortalama, her sınıfın metriklerini sınıf büyüklüklerine göre ağırlıklandırarak hesaplamaktadır. Finansal veri setlerinde, ağırlıklı ortalama, çoğunluk sınıfının performansına daha fazla ağırlık vermektedir (Brown ve Mues, 2012). Ağırlıklı ortalama, genel performansı yansıtmaktadır. Ancak azınlık sınıfının önemini gölgede bırakabilmektedir (Provost ve Fawcett, 2013).

2.5.7. Karmaşıklık Matrisi

Karmaşıklık matrisi, modelin tahminlerini gerçek sınıflarla karşılaştıran bir tablodur ve aşağıdaki bilgileri içermektedir.

TP: Doğru pozitifler,

TN: Doğru negatifler,

FP: Yanlış pozitifler,

FN: Yanlış negatifler.

Karmaşıklık matrisi, modelin sonucunda sonuçlarını nasıl sınıflandırdığını görselleştirmektedir (Fawcett, 2006). Karmaşıklık matrisi, karar vericilere modelin performansını ayrıntılı bir şekilde analiz etme imkânı sunmaktadır (Kohavi ve Provost, 1998).

2.5.8. ROC Eğrisi

Receiver Operating Characteristic (ROC) eğrisi, duyarlılık ile 1-özgüllük arasındaki ilişkiyi grafiksel olarak göstermektedir. Özgüllük denklem 2.12'deki gibi ifade edilmektedir.

$$\text{Özgüllük} = TN / (TN + FP) \qquad (2.12)$$

ROC eğrisinin altında kalan alan (AUC) modelin sınıfları ayırt etme yeteneğini ölçmektedir. AUC 1'e yakınsa model mükemmel, 0.5'e yakınsa mükemmellikten uzaktır (Bradley, 1997).

2.5.9. Öğrenme Eğrisi

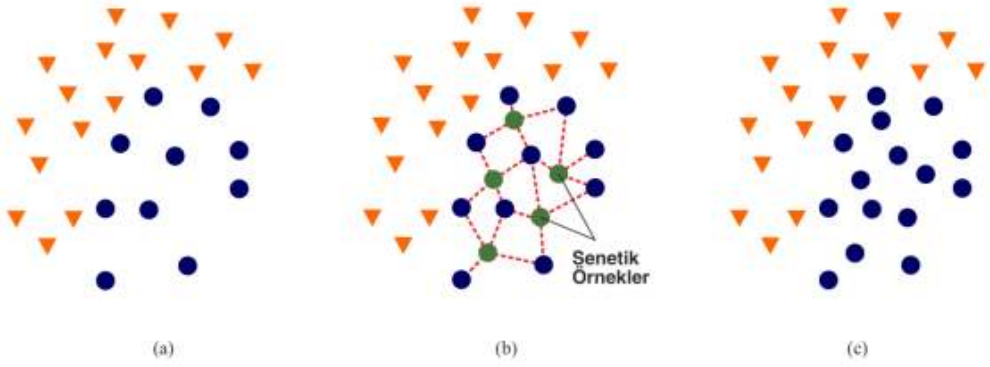
Öğrenme eğrisi, eğitim ve doğrulama veri setlerindeki performansın veri miktarıyla nasıl değiştiğini göstermektedir. Öğrenme eğrisi modelin aşırı öğrenme veya eksik öğrenme durumunu ortaya koymaktadır (Mitchell, 1997).

2.6. SMOTE

(Synthetic Minority Over-sampling Technique) SMOTE, dengesiz veri setlerindeki azınlık sınıfını temsil eden örnekleri artırarak sınıflardaki dağılımı dengelemeyi amaçlayan yöntemdir (Chawla vd., 2002). SMOTE geleneksel örnekleme yöntemlerinden farklı çalışmaktadır. SMOTE mevcut azınlık sınıfı örneklerini kopyalamak yerine, örnekler arasında sentetik veri noktaları oluşturmaktadır. Bu yaklaşım ile modelin azınlık sınıfını öğrenme kapasitesi artarak aşırı öğrenme riskini azalmaktadır (Batista vd., 2004). Denklem 2.13'de SMOTE ile sentetik veri üretme görülmektedir.

$$x_{yeni} = x_i + \lambda.(x_{nn} - x_i) \qquad (2.13)$$

Denklem 2.13'te; x_{yeni} üretilen sentetik veriyi, x_i orijinal azınlık sınıf örneği, x_{nn} rastgele seçilen k-komşusunu ifade etmektedir (Chawla vd., 2002).



Şekil 3. SMOTE Sentetik Veri Üretme Yöntemi

Kaynak: Özgün, S.(2022). *Dengesiz bal peteği veri setinde sınıflandırma performansının analizi* [Yüksek lisans tezi]. Selçuk Üniversitesi.

Şekil 3'te SMOTE ile sentetik veri üretme yöntemi görülmektedir. (a) dengesiz veri setini, (b) sentetik veri oluşturma sürecini, (c) ise dengeli hale getirilen veri setini göstermektedir.

SMOTE özellikle finansal verilerde kredi skoru üretiminde dengesiz sınıf problemini çözmek için etkili bir yöntemdir. SMOTE'un en büyük avantajı, azınlık sınıfını temsil eden verileri arttırarak modelin genelleştirme yeteneğini arttırıyor olmasıdır. Finansal risk analizlerinde, bu yöntem ödeme yapmayan firmaların

özelliklerini daha doğru tespit ederek yanlış negatif oranını azaltabilmektedir (He ve Garcia, 2009). Ancak, SMOTE'un sentetik veri üretmesi, veri setinin orijinal dağılımını değiştirebilmektedir. Bu durum bazı durumlarda modelin gerçek dünyadaki performansını etkileyebilmektedir (Blagus ve Lusa, 2013). Aynı zamanda, k değerinin seçimi ve özellik uzayının boyutu, yöntemin etkinliğini belirleyen kritik faktörlerdendir.

SMOTE, finansal uygulamalarda genellikle denetimli makine öğrenmesi algoritmalarıyla birlikte kullanılmaktadır Bir banka SMOTE ile dengelenmiş bir veri setini XGBoost ile eğitebilir ve kredi geri ödeme olasılıklarını daha hassas bir şekilde tahmin edebilmektedir (Burez ve Van den Poel, 2009). SMOTE, veri dengesizliğini gidermeye çalışan diğer yöntemlere kıyasla veri çeşitliliğini koruyarak model performansını arttırmaktadır(Fernández vd., 2018). Bu kombinasyon ile özellikle azınlık sınıfının kritik olduğu iflas durumu gibi senaryolarda büyük bir avantaj sağlamaktadır.

3. BULGULAR

Bu çalışmada banka müşterisi firmaların yıl sonu verileri anonimleştirilerek kullanılmıştır. Veriler ile model kurulmadan önce daha anlamlı sonuçlar elde etmek için Z-Skor yöntemi ile uç değerler temizlenmiş ve oran olmayan değişkenler standartlaştırılmıştır. Denetimsiz öğrenme aşamasında ayrıca boyut indirgeme amacıyla Temel Bileşen Analizi(PCA) uygulanmıştır. Denetimli makine öğrenmesi algoritmalarında bağımlı değişken olarak belirlenen Kredi Skorunun dengesiz dağılım göstermesi nedeniyle SMOTE kullanılmış ve veri dengesizliği giderilerek daha anlamlı sonuçlar elde edilmiştir. Çalışmadaki analizlerde Jupyter Notebook kullanılmıştır. Çalışmada kullanılan verilerle ilgili Tablo 1'de açıklayıcı bilgiler bulunmaktadır.

Tablo 1. Öznitelik Bilgileri

Öznitelik	Açıklama
Aktif	Firmanın tüm varlıklarının toplamıdır. Bilanço aktifini oluşturur ve işletmenin sahip olduğu kaynakları ifade eder.
Aktif Devir Hızı	Firmanın varlıklarını ne kadar verimli kullandığını gösteren bir orandır. (Net Satışlar / Toplam Aktifler).
Atanan Küme	Veri kümesindeki her verinin ait olduğu kümedir. (Denetimsiz öğrenme sonuçlarıyla elde edilen kümeler).
Cari Oran	Kısa vadeli borçların kısa vadeli varlıklarla karşılanabilirliğini gösteren orandır. (Kısa Vadeli Varlıklar / Kısa Vadeli Borçlar).
Çalışan Sayısı	Firmanın toplam çalışan sayısıdır.
Diğer Kısa Vadeli Borçlar	Kısa vadeli borçlar kategorisinde yer almayan ancak 1 yıl içinde ödenmesi gereken borçlar.
Duran/Aktif	Firmanın duran varlıklarının aktiflerine oranı. Duran varlıklar genellikle uzun vadeli yatırımları ve maddi duran varlıkları içermektedir.
FAVÖK Marjı	Firmanın faaliyet kârının net satışlara oranı. Faaliyet kârlılığını göstermektedir. (FAVÖK / Net Satışlar).
Kasa Hazır Değer	Firmanın nakit ve nakit benzeri varlıklarının toplamını göstermektedir.
Kısa Vadeli Borç/Aktif	Firmanın kısa vadeli borçlarının toplam aktiflerine oranını göstermektedir.
Kısa Vadeli Ticari Borçlar	Firmanın, tedarikçilere olan kısa vadeli ticari borçlarıdır.
Kredi Skoru	Bu çalışmadaki bağımlı değişkendir. Firmanın kredi derecesini belirleyen puandır. Bu çalışmada 0,1,2,3 olmak üzere dört sınıf bulunmaktadır.

Nakdi Limit Doluluk Oranı	Firmanın nakdi kredi limitlerinin kullanım oranıdır.
Net Dönem Kârı	Firmanın belirli bir dönemde elde ettiği net kârı göstermektedir.
Net İşletme	Firmanın temel faaliyetlerinden elde ettiği net kârdır.
Net Kâr/Aktif	Firmanın net kârının toplam aktiflerine oranıdır. Firmanın aktiflerini ne kadar verimli kullandığını göstermektedir.
Net Satışlar	Firmanın mal ve hizmet satışlarından elde ettiği gelirdir.
Ödenmiş Sermaye	Firmanın hissedarlarına karşı ödenmiş olan sermaye tutarıdır.
Stok Devir	Firmanın stoklarını ne kadar hızlı sattığını gösteren orandır. (Satılan Malın Maliyeti / Ortalama Stoklar).
Ticari Borç Devir Hızı	Firmanın ticari borçlarını ödeme hızını gösteren orandır. (Ticari Borçlar / Satın Alma Harcamaları).
Toplam Borç/FAVÖK	Firmanın toplam borçlarının FAVÖK'e oranıdır. Firmanın borç yükünün faaliyet kârlılığına göre ne kadar sürdürülebilir olduğunu göstermektedir.
Toplam Nakdi Riski Artışı	Firmanın nakit akışındaki potansiyel riski ve likidite sıkıntılarının artışını ifade eden bir orandır.
Toplam Nakit Açığı	İşletmenin belirli bir dönem içinde nakit girişleri ile nakit çıkışları arasındaki farkı ifade etmektedir.

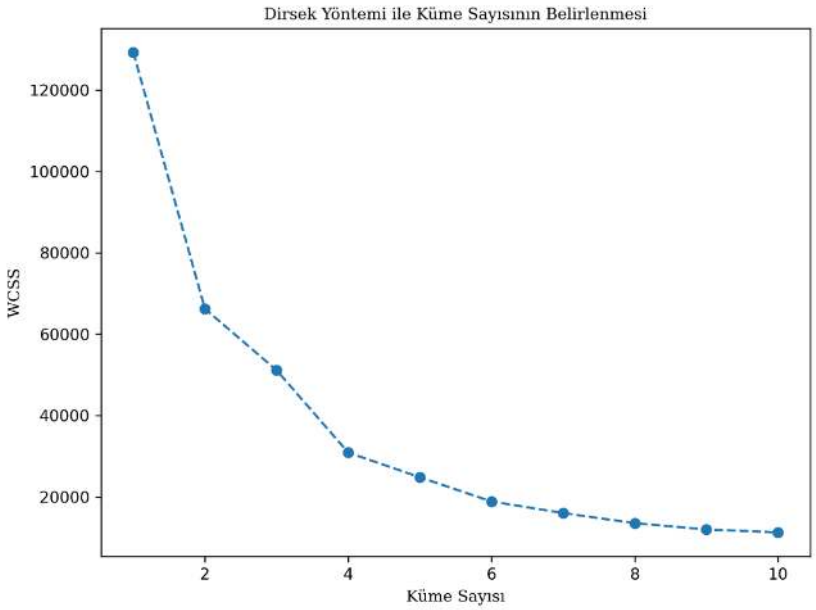
Kaynak: Yazar tarafından oluşturulmuştur.

3.1. Denetimsiz Makine Öğrenmesi Sonuçları

Bu bölümde denetimsiz makine öğrenmesi modellerinin sonuçları incelenmiştir. İncelemeler sonucunda en iyi sonuçların elde model seçilerek kümeleme yapılmıştır.

3.1.1. KMeans Algoritması

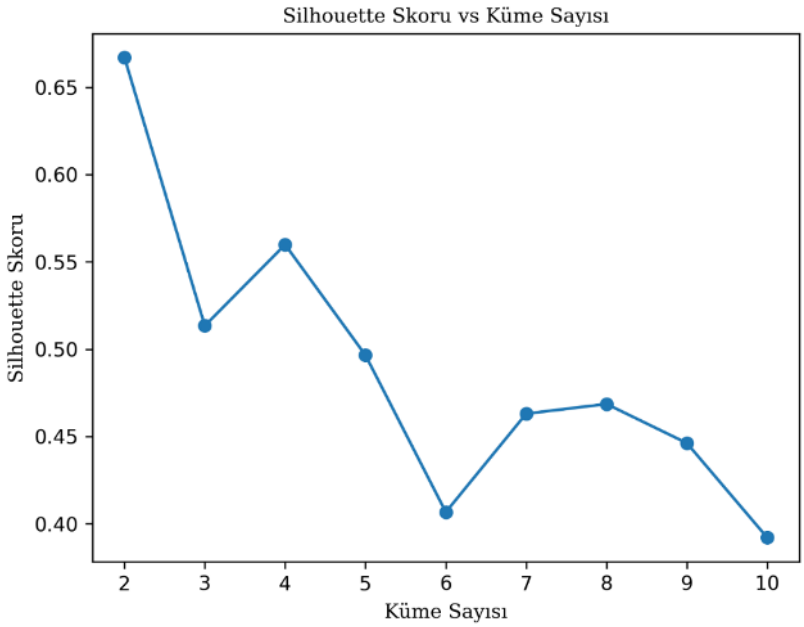
Çalışmanın bu bölümünde denetimsiz makine öğrenmesi algoritmaları kullanılmıştır. Küme sayısı ve kullanılacak algoritmayı belirlemek amacıyla Dirsek yöntemi, Silhoutte Skoru, Davies-Bouldin ve Calinski-Harabasz İndeksi kullanılmıştır. Şekil 4'te Dirsek Yönteminin grafiği görülmektedir. Şekilde 4'te görüldüğü üzere kırılım 3 ve 4 küme sayısında yoğunlaşmaktadır.



Şekil 4. Dirsek Yöntemi ile Küme Sayısının Belirlenmesi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 5'te Silhoutte Skorunun grafiği görülmektedir. Grafiğe göre 2 ve 4 küme sayısında en iyi skorlar gözlemlenmiştir. Dirsek yöntemi ve Silhoutte skorları birlikte değerlendirildiğinde ve çalışmanın amacına da uygun olması sebebiyle küme sayısı 4 olarak belirlenmiştir.

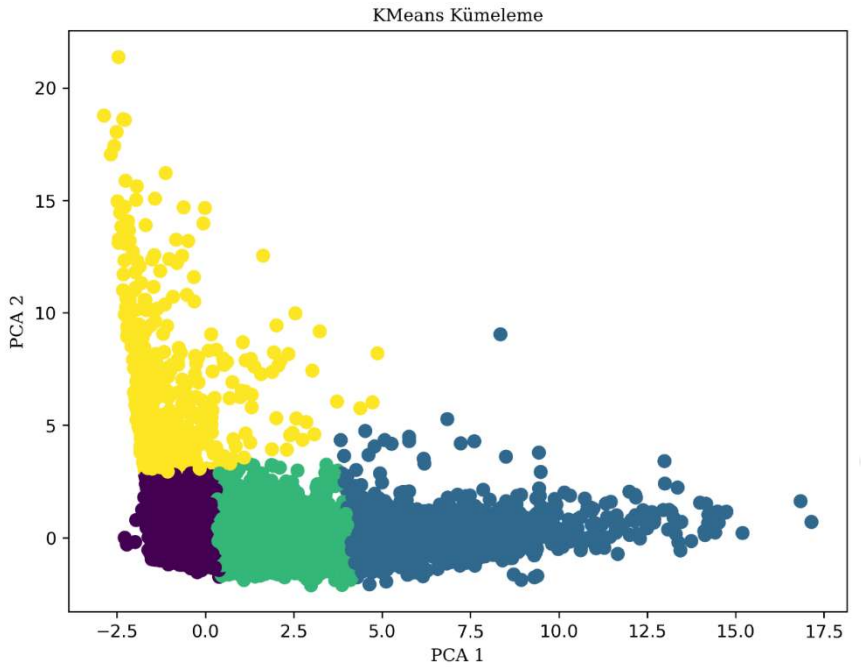


Şekil 5. Silhouette Skoru ve Küme Sayısı

Kaynak: Yazar tarafından oluşturulmuştur.

Belirlenen küme sayısı denetimsiz makine öğrenmesi algoritmalarında kullanılmıştır. Çalışmada; KMeans Kümeleme, Hiyerarşik Kümeleme, DBSCAN Kümeleme, BIRCH Kümeleme, OPTICS Kümeleme ve GMM Kümeleme algoritmaları kullanılmıştır. En iyi sonuçların elde edildiği algoritma belirlenerek elde edilen kümeler çalışmada kullanılmıştır.

Şekil 6'da KMeans algoritması kullanılarak çizilmiş olan grafik görülmektedir. Grafikte görüldüğü üzere küme sınırlarının belirli olduğu ve KMeans algoritması ile başarılı kümeleme oluşturulmuştur.

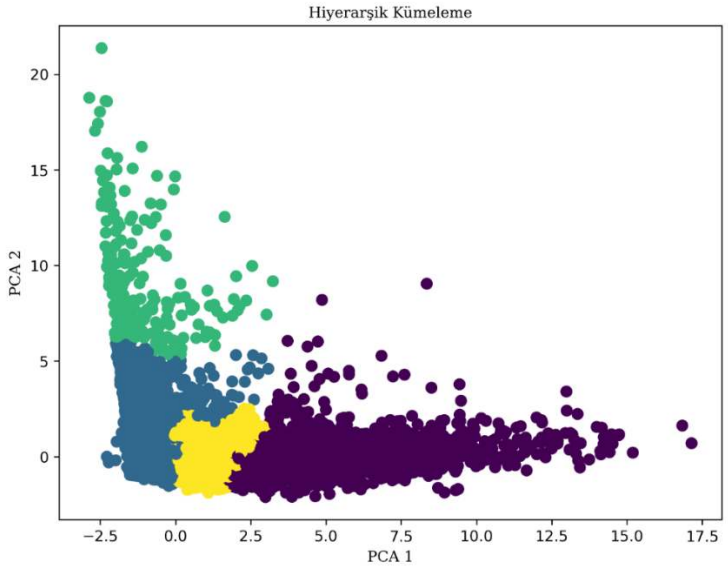


Şekil 6. KMeans Kümeleme

Kaynak: Yazar tarafından oluşturulmuştur.

3.1.2. Hiyerarşik Algoritması

Şekil 7'de Hiyerarşik kümeleme sonuçları grafiği görülmektedir. Hiyerarşik kümeleme sonuçları da KMeans algoritmasına yakın görünmektedir. Ancak bazı noktalarda sınırlar net değildir.

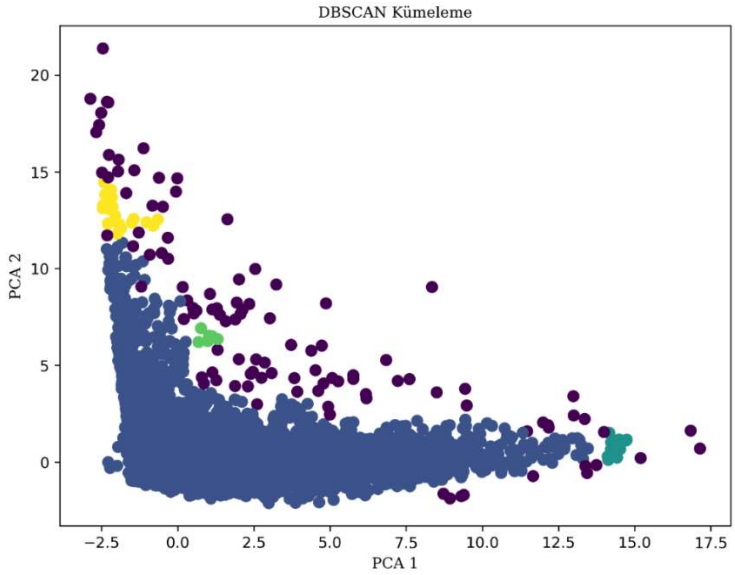


Şekil 7. Hiyerarşik Kümeleme

Kaynak: Yazar tarafından oluşturulmuştur.

3.1.3. DBSCAN Algoritması

Şekil 8'de DBSCAN algoritması kümeleme grafiği görülmektedir. DBSCAN algoritmasından elde edilen kümeleme sonuçları başarılı görünmemektedir. Ayrıca küme sınırları da belirgin değildir.

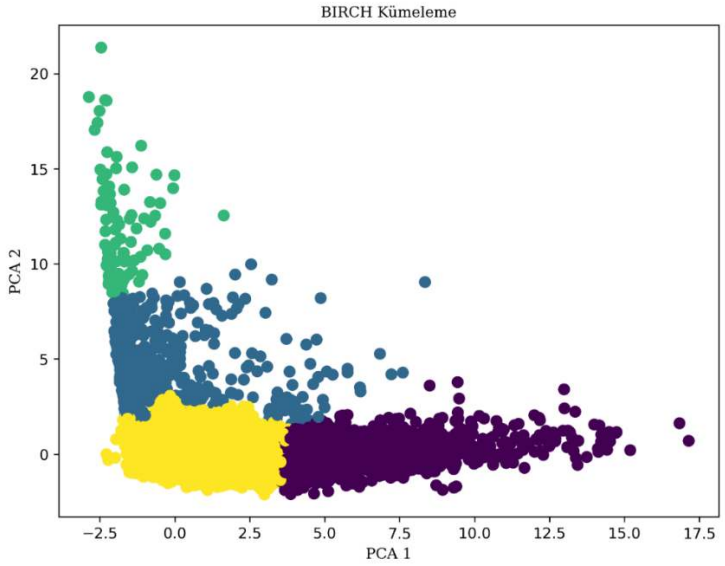


Şekil 8.DBSCAN Kümeleme

Kaynak: Yazar tarafından oluşturulmuştur.

3.1.4. BIRCH Algoritması

Şekil 9'da BIRCH algoritması kullanılarak çizilmiş olan grafik görülmektedir. Kümelemenin genel olarak iyi olduğu ama bazı noktalarda sınırların belirgin olmadığı görülmektedir.

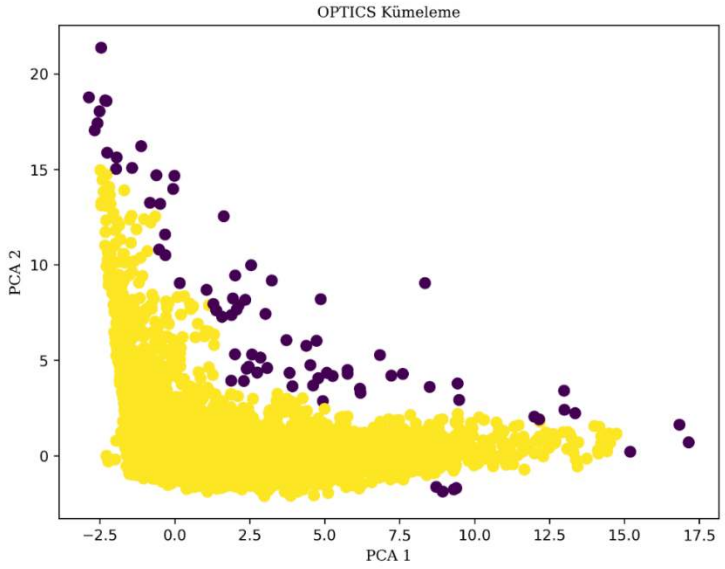


Şekil 9.BIRCH Kümeleme

Kaynak: Yazar tarafından oluşturulmuştur.

3.1.5. OPTICS Algoritması

Şekil 10'da OPTICS algoritması kullanılarak çizilmiş olan kümeleme grafiği görülmektedir. Grafikte görüldüğü tek bir kümede yığılma olduğu görülmektedir.

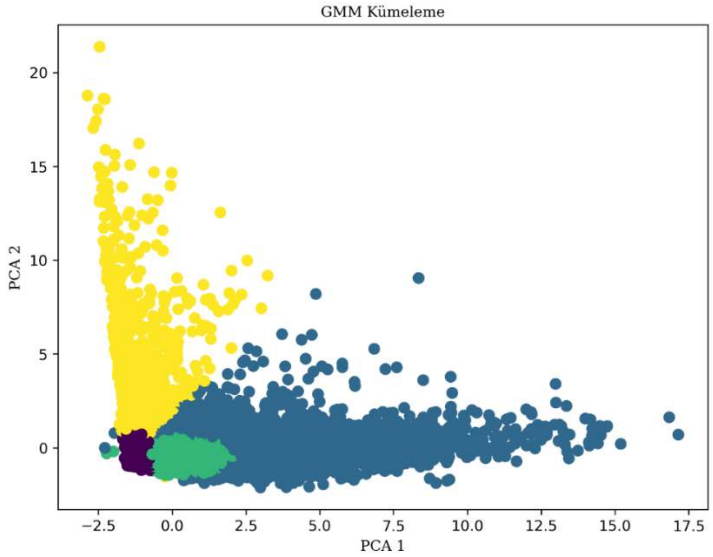


Şekil 10. OPTICS Kümeleme

Kaynak: Yazar tarafından oluşturulmuştur.

3.1.6. GMM Algoritması

Şekil 11'de GMM algoritması kullanılarak çizilmiş olan kümeleme grafiği görülmektedir. Grafikte görüldüğü üzere kümeleme sonuçları istenilen düzeyde dağılmamıştır.



Şekil 11. GMM Kümeleme

Kaynak: Yazar tarafından oluşturulmuştur.

3.1.7. En İyi Modelin Belirlenmesi

Grafiklerin oluşturulmasından sonra kümeleme başarısını gösteren Silhouette Skoru, Davies-Bouldin İndeksi ve Calinski-Harabasz İndeksi incelenmiştir. Ayrıca bu kriterlerle birlikte kümelerde oluşan dağılımlar da gözlemlenmiştir.

Tablo 2. Başarı Kriterleri

Algoritma	Silhouette Skoru	Davies-Bouldin İndeksi	Calinski-Harabasz İndeksi	Küme Dağılımı
KMeans	0.5604	0.6421	23459.7668	{0: 15974, 1: 1233, 2: 4336, 3: 562}
Agglomerative	0.4448	0.7067	17371.6426	{0: 2482, 1: 15564, 2: 204, 3: 3855}
DBSCAN	0.6482	1.3982	611.6756	{-1: 107, 0: 21962, 1: 10, 2: 6, 3: 20}
BIRCH	0.6366	0.6292	13091.4346	{0: 1417, 1: 757, 2: 87, 3: 19844}
OPTICS	0.7699	0.9213	1015.9796	{-1: 79, 0: 22026}
GMM	0.3687	0.8143	12847.6349	{0: 9765, 1: 3360, 2: 7312, 3: 1668}

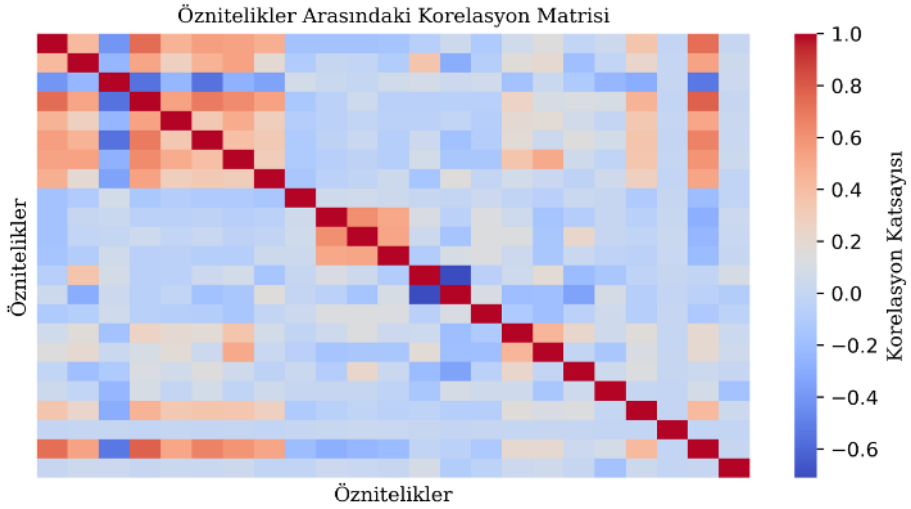
Kaynak: Yazar tarafından oluşturulmuştur.

Bu indekslere ek olarak kümeleme algoritmaları verinin doğal yapısına ve dağılımına da uygun seçilmelidir (Jain 2010). Tablo 2'de yer alan kriterlerden Silhouette yüksek, Davies-Bouldin düşük, Calinski Harabasz indeksinin ise yüksek olması istenilen durumdur. Ancak bu kriterlerle birlikte küme dağılımları, grafikteki görünümleri ve kümeleme yapılan konunun içeriğiyle olan uyumlulukları da önemlidir. Tablodaki kriterler, küme dağılımı ve grafiği incelendiğinde en iyi kümelemenin KMeans ile yapıldığı belirlenmiştir. KMeans algoritması kullanılarak belirlenen kümeler veri setine eklenerek denetimli öğrenme algoritmaları kullanılmıştır. Belirlenen bu kümeler KMeans algoritması ile oluşturulan firmaların segmentini oluşturmaktadır.

KMeans algoritması ile kurulan kümeleme sonuçları incelendiğinde küme dağılımında sıfır kümesinde ağırlık olduğu görülmektedir. Ancak bu durum çalışma için sorun teşkil etmemektedir çünkü çalışmada kullanılan veri setinde Kobi firmaları çoğunluğu oluşturmaktadır.

3.2. Denetimli Makine Öğrenmesi Sonuçları

Çalışmada denetimsiz öğrenme algoritması olan KMeans kullanılarak elde edilen kümeler veri setine eklenmiştir. Eklenen bu kümeler firmaların segmentlerini göstermektedir. Çalışmada ilk olarak değişkenler arasındaki korelasyon matrisi incelenmiştir. Aşağıdaki grafikte korelasyon matrisi görülmektedir. Şekil 12'deki korelasyon matrisinde görüldüğü üzere değişkenler arasındaki ilişki 0.8'den küçüktür.



Şekil 12. Korelasyon Matrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Kurulan modellerde daha anlamlı sonuçlar elde etmek için uç değerler Z-skor yöntemiyle temizlenmiş ve veri standartlaştırılmıştır. Çalışmada kullanılan veri setindeki Kredi Skoru bağımlı değişken olarak belirlenmiştir. Kredi Skoru; 0,1,2,3 değerini alan bir değişkendir. Kredi skoru değişkeninin dağılımı incelendiğinde bir dengesizlik olduğu görülmektedir. Bu dengesizliği gidermek amacıyla SMOTE kullanılmıştır. Aşağıdaki tabloda SMOTE öncesi ve sonrası dağılım görülmektedir. Tabloda 3'te görüldüğü üzere SMOTE kullanılmadan önce belirgin bir dengesizlik bulunmaktaydı ve bu dengesizlik kurulan modellerin anlamlılığını düşürebilmektedir. Daha anlamlı ve başarı oranı yüksek modeller oluşturmak için dengesizliğin giderilmesi önemlidir. Bu çalışmada da denetimli makine öğrenmesi algoritmaları, SMOTE ile dengesizliğin giderilmesi sonrası elde edilen veri setiyle kullanılmıştır.

Tablo 3. SMOTE Sonuçları

Sınıf	SMOTE Öncesi	SMOTE Sonrası
0	263	12881
1	1041	12881
2	12881	12881
3	2716	12881

Kaynak: Yazar tarafından oluşturulmuştur.

Veri seti eğitim ve testlerine ayırdıktan sonra sırasıyla XGBoost, Karar Ağacı, Rastgele Orman, KNN ve LightGBM algoritmaları kullanılmıştır.

3.2.1. XGBoost Algoritması

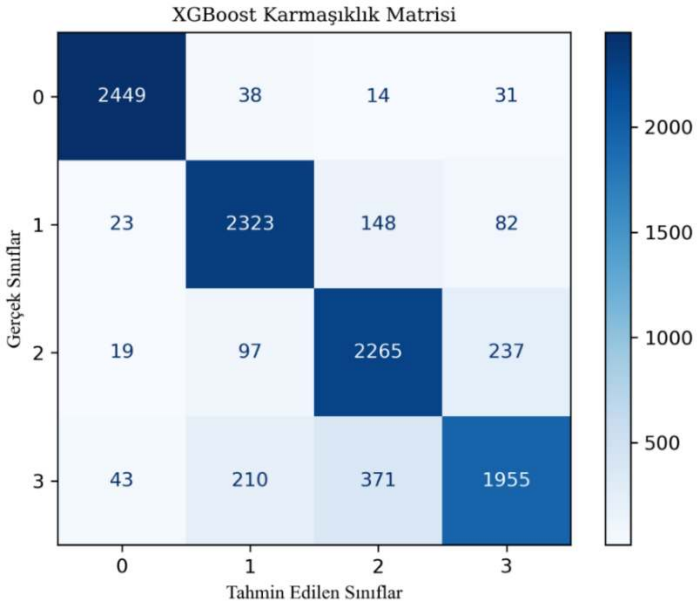
Denetimli makine öğrenmesi algoritmalarından olan XGBoost algoritması kullanılarak oluşturulan modelin sınıflandırma raporu Tablo 4'teki gibidir.

Tablo 4. XGBoost Sınıflandırma Raporu

Sınıf	Kesinlik	Duyarlılık	F1 Skoru
0	0.97	0.97	0.97
1	0.87	0.9	0.89
2	0.81	0.87	0.84
3	0.85	0.76	0.80
Doğruluk	0.87	-	-
Makro Ortalama	0.87	0.87	0.87
Ağırlıklı Ortalama	0.87	0.87	0.87

Kaynak: Yazar tarafından oluşturulmuştur.

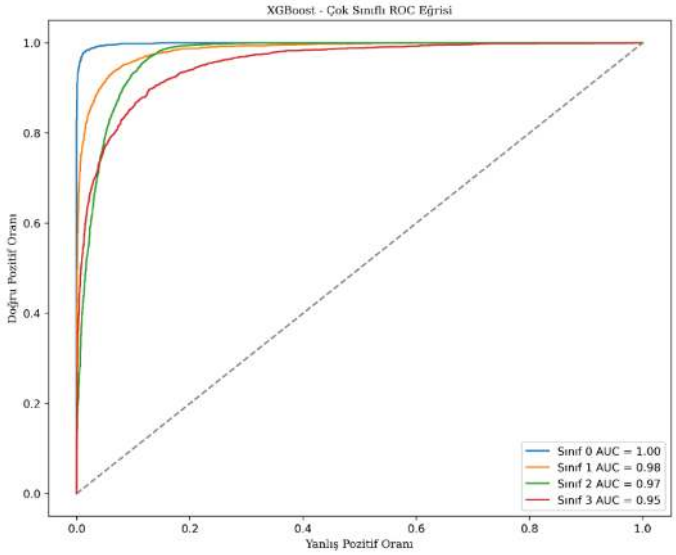
Tabloda 4'te görüldüğü gibi 0,1,2 sınıfları için sonuçlar genel olarak başarılıdır. Ancak 3. Sınıflandırma için başarı oranları geliştirilebilir düzeydedir.



Şekil 13. XGBoost Karmaşıklık Matrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 13'te karmaşıklık matrisi sonuçları görülmektedir. 1,2,3 sınıflandırma sonuçları başarılıdır anca 3. sınıflandırma tahminlerinde geliştirilebilir sonuçlar görülmektedir.



Şekil 14. XGBoost ROC Eğrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 14'te XGBoost modelinin ROC eğrisi görülmektedir. ROC eğrisinde elde edilen sonuçlarda sınıflandırma raporundaki sonuçlarla paraleldir. Genel olarak model performansının iyi olduğu görülmektedir.

3.2.2. Karar Ağaçları Algoritması

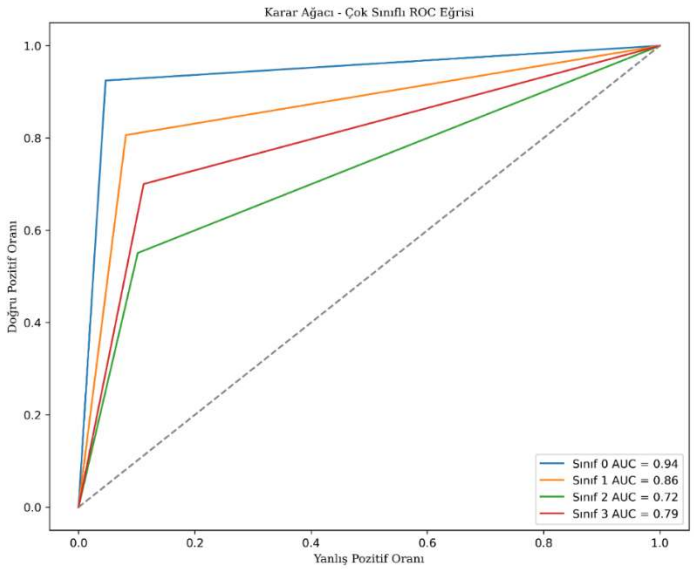
Karar Ağacı sınıflandırma raporu Tablo 5'teki gibidir. Sınıflandırma raporuna göre 0 ve 1 sınıflarındaki başarı oranları yüksektir. Ancak 2 ve 3 sınıflarına bakıldığında başarı oranlarının düşük olduğu görülmektedir. Genel olarak model geliştirilebilir durumdadır.

Tablo 5. Karar Ağacı Sınıflandırma Raporu

Sınıf	Kesinlik	Duyarlılık	F1 Skoru
0	0.87	0.92	0.89
1	0.77	0.81	0.79
2	0.65	0.55	0.6
3	0.68	0.7	0.69
Doğruluk	0.74	-	-
Makro Ortalama	0.74	0.75	0.74
Ağırlıklı Ortalama	0.74	0.74	0.74

Kaynak: Yazar tarafından oluşturulmuştur.

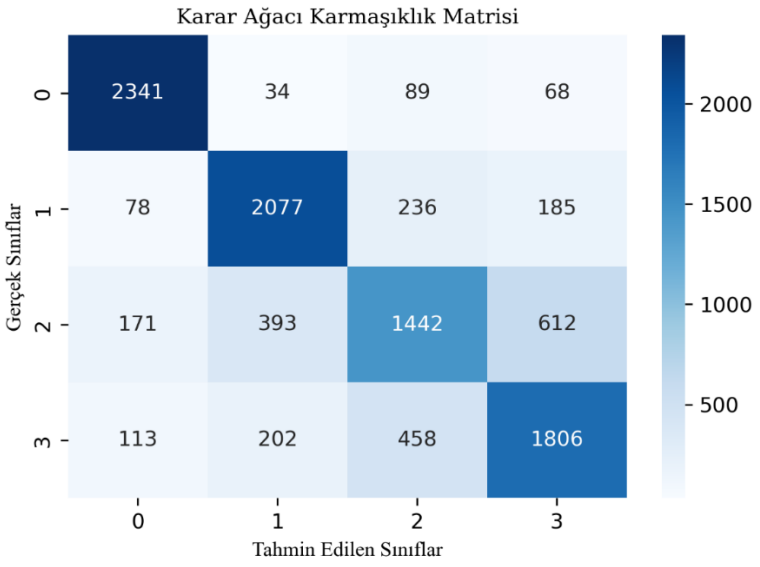
Şekil 15'te ROC eğrisi görülmektedir. ROC eğrisinde görüldüğü üzere sonuçlar yine geliştirilebilir durumdadır.



Şekil 15. Karar Ağacı ROC Eğrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 16'da karmaşıklık matrisi görülmektedir. Karmaşıklık matrisinde de görüldüğü üzere 2. Sınıf hatalı tahminleri yüksek durumdadır.



Şekil 16. Karar Ağacı Karmaşıklık Matrisi

Kaynak: Yazar tarafından oluşturulmuştur.

3.2.3. Rastgele Orman Algoritması

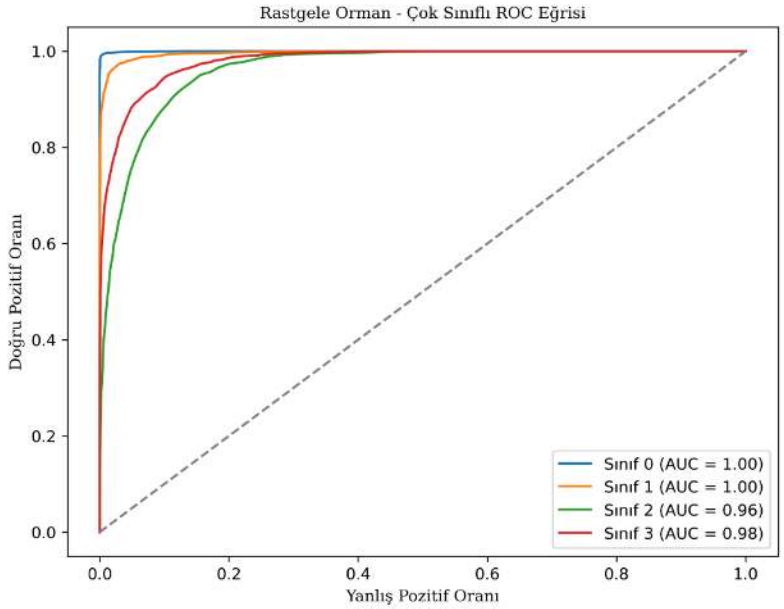
Rastgele Orman algoritması sınıflandırma bilgileri Tablo 6'daki tablodaki gibidir. Sınıflandırma raporuna göre başarı oranlarının iyi olduğu görülmektedir. Ancak 2.sınıf içim başarı geliştirilebilir durumdadır.

Tablo 6. Rastgele Orman Sınıflandırma Raporu

Sınıf	Kesinlik	Duyarlılık	F1 Skoru
0	0.99	0.99	0.99
1	0.91	0.97	0.94
2	0.88	0.75	0.81
3	0.83	0.9	0.86
Doğruluk			0.9
Makro Ortalama	0.9	0.9	0.9
Ağırlıklı Ortalama	0.9	0.9	0.9

Kaynak: Yazar tarafından oluşturulmuştur.

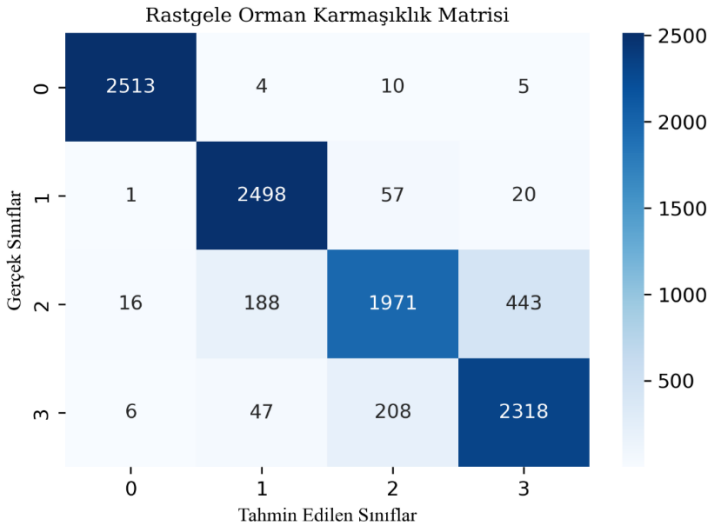
Şekil 17'de Rastgele Orman ROC eğrisi görünmektedir. ROC eğrisine göre sınıflandırmalar başarılı görünmektedir.



Şekil 17. Rastgele Orman ROC Eğrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 18'de karmaşıklık matrisi görülmektedir. Karmaşıklık matrisinde de görüldüğü üzere 1 ve 2. Sınıflar için oldukça iyi sonuçlar elde edilmiştir. Ancak 2 ve 3. Sınıf tahminleri geliştirilebilir durumdadır.



Şekil 18. Rastgele Orman Karmaşıklık Matrisi

Kaynak: Yazar tarafından oluşturulmuştur.

3.2.4. KNN Algoritması

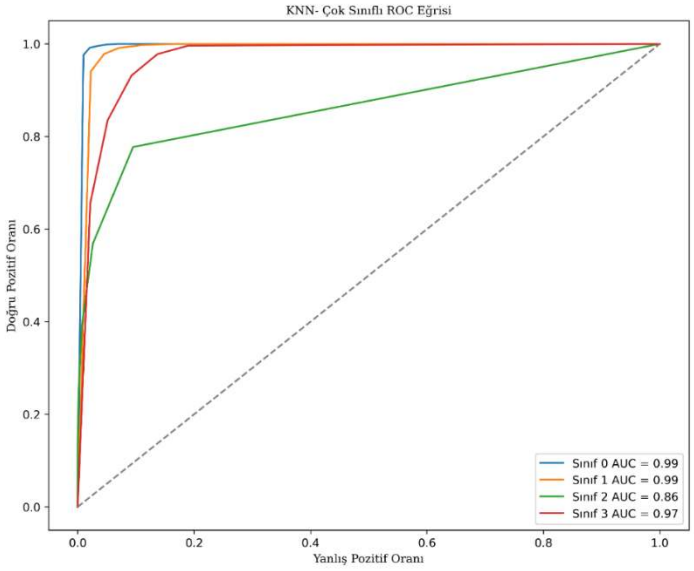
Tablo 7'de KNN sınıflandırma sonuçları görülmektedir. Sınıflandırma raporuna göre 0, 1 ve 3 sınıflarındaki başarı oranları yüksektir. Ancak 2.sınıfa bakıldığında başarı oranlarının düşük olduğu görülmektedir. Genel olarak model geliştirilebilir durumdadır.

Tablo 7. KNN Sınıflandırma Raporu

Sınıf	Kesinlik	Duyarlılık	F1 Skoru
0	0.89	0.99	0.94
1	0.8	0.99	0.89
2	0.93	0.41	0.57
3	0.77	0.93	0.84
Doğruluk	0.83		
Makro Ortalama	0.85	0.83	0.81
Ağırlıklı Ortalama	0.85	0.83	0.81

Kaynak: Yazar tarafından oluşturulmuştur.

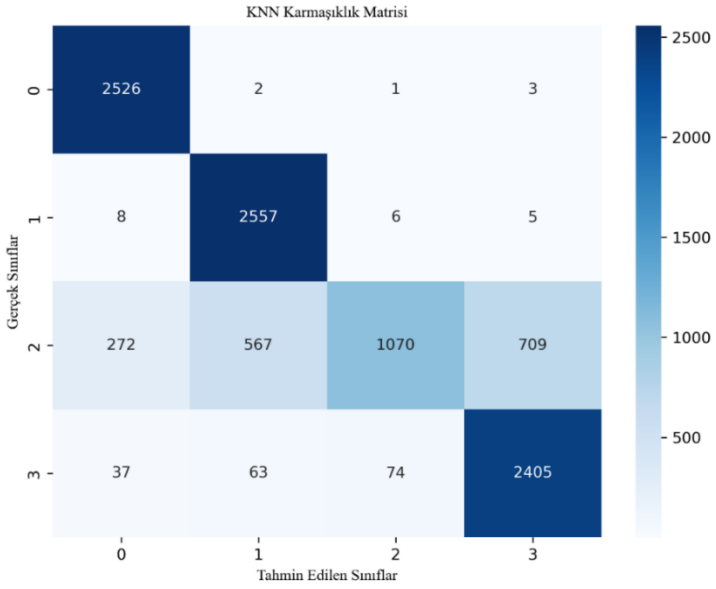
KNN ROC grafiği Şekil 19'daki gibidir. ROC grafiğine göre 2.sınıf başarısı diğer sınıflara göre daha düşüktür.



Şekil 19. KNN ROC Eğrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 20'de karmaşık matrisi görülmektedir. Karmaşıklık matrisine göre 0 ve 1. sınıflar için çok başarılı tahminler elde edilmiştir. Ancak özellikle 2.sınıf başarı oranının düşük olduğu görülmektedir.



Şekil 20. KNN Karmaşıklık Matrisi

Kaynak: Yazar tarafından oluşturulmuştur.

3.2.5. LightGBM Algoritması

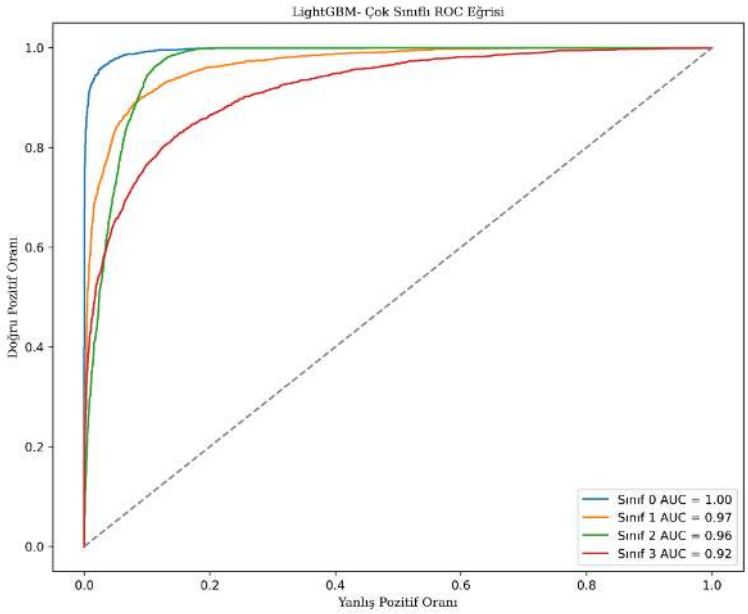
LightGBM sınıflandırma sonuçları Tablo 8'de yer almaktadır. Sınıflandırma raporuna göre 0,1 ve 2 sınıflarındaki başarı oranları yüksektir. Ancak 3.sınıfa bakıldığında başarı oranının düşük olduğu görülmektedir. Genel olarak model geliştirilebilir durumdadır.

Tablo 8. LightGBM Sınıflandırma Raporu

Sınıf	Kesinlik	Duyarlılık	F1 Skoru
0	0.94	0.94	0.94
1	0.84	0.85	0.84
2	0.78	0.9	0.84
3	0.81	0.68	0.74
Doğruluk			0.84
Makro Ortalama	0.84	0.84	0.84
Ağırlıklı Ortalama	0.84	0.84	0.84

Kaynak: Yazar tarafından oluşturulmuştur.

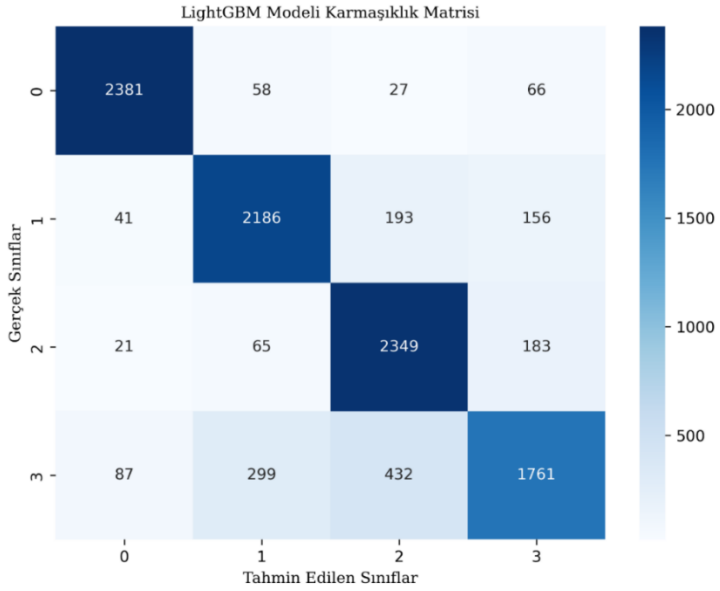
ROC eğrisi grafiği Şekil 21'de yer almaktadır. ROC eğrisi grafiğine göre 3.sınıf başarısının daha düşük olduğu görülmektedir.



Şekil 21. LightGBM ROC Eğrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Karmaşıklık matrisi Şekil 22'de yer almaktadır. Karmaşıklık matrisinde de 3. sınıf için hatalı tahminlerin yüksek olduğu görülmektedir.



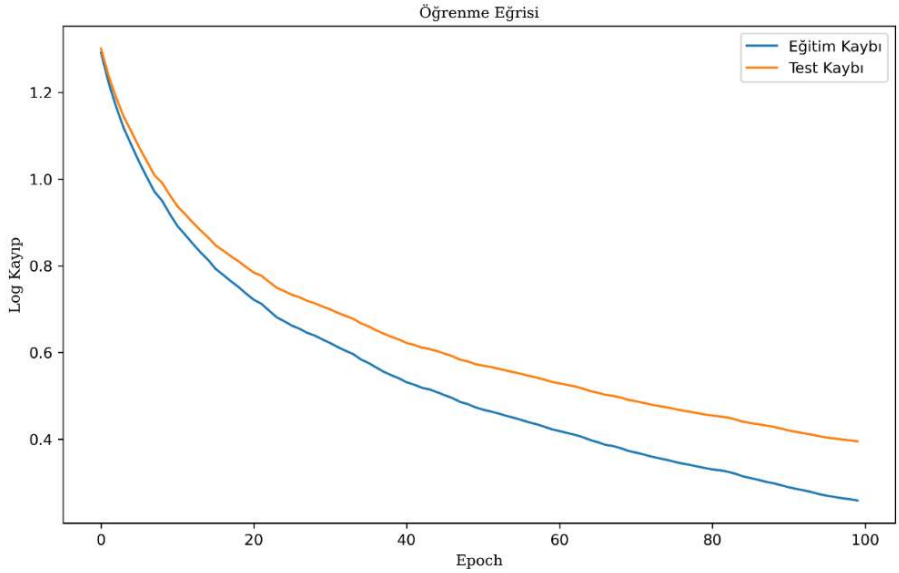
Şekil 22. LightGBM Karmaşıklık Matrisi

Kaynak: Yazar tarafından oluşturulmuştur.

KMeans kümeleme algoritması ile elde edilen kümeler veriye eklendikten sonra denetimli makine öğrenmesi algoritmaları kullanılmıştır. Kullanılan algoritmalar arasından en iyi sonuçları veren algoritmanın seçilmesi başarı oranlarına göre yapılmıştır.

3.3. En İyi Modelin Belirlenmesi

Denetimli makine öğrenmesi algoritmalarından; XGBoost, Karar Ağacı, Rastgele Orman, KNN ve LightGBM kullanılarak kurulan modellerden en iyi sonuçların XGBoost modeli ile elde edildiği belirlenmiştir. XGBoost modelin anlamlılığını ve başarısını incelemek amacıyla çeşitli incelemeler yapılmıştır.



Şekil 23. XGBoost Öğrenme Eğrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 23'te görüldüğü üzere eğitim ve test kayıpları birbirine yakınsamaktadır. Kurulan modelin başarılı olduğunu göstermektedir.

Tablo 9. Çapraz Doğrulama

Başarı Oranı Türü	Başarı Oranı
Eğitim Seti	0.9583
Çapraz Doğrulama (1)	0.8701
Çapraz Doğrulama (2)	0.8738
Çapraz Doğrulama (3)	0.8712
Çapraz Doğrulama (4)	0.8745
Çapraz Doğrulama (5)	0.8702
Ortalama Başarı Oranı	0.872

Kaynak: Yazar tarafından oluşturulmuştur.

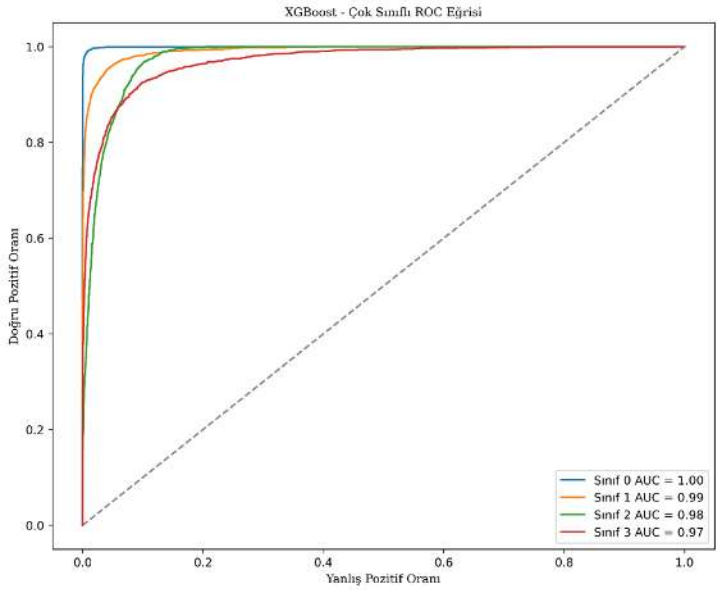
Tablo 9'da görüldüğü üzere XGBoost modelin çapraz doğrulama sonuçları görülmektedir. Çapraz doğrulama sonuçları da aşırı öğrenme sorunu olmadığını destekler niteliktedir.

Tablo 10. En İyi Parametrelerle XGBoost Sınıflandırma Raporu

Sınıf	Kesinlik	Duyarlılık	F1 Skoru
0	0.98	0.98	0.98
1	0.91	0.93	0.92
2	0.84	0.88	0.86
3	0.88	0.82	0.85
Doğruluk	0.9	-	-
Makro Ortalama	0.9	0.9	0.9
Ağırlıklı Ortalama	0.9	0.9	0.9

Kaynak: Yazar tarafından oluşturulmuştur.

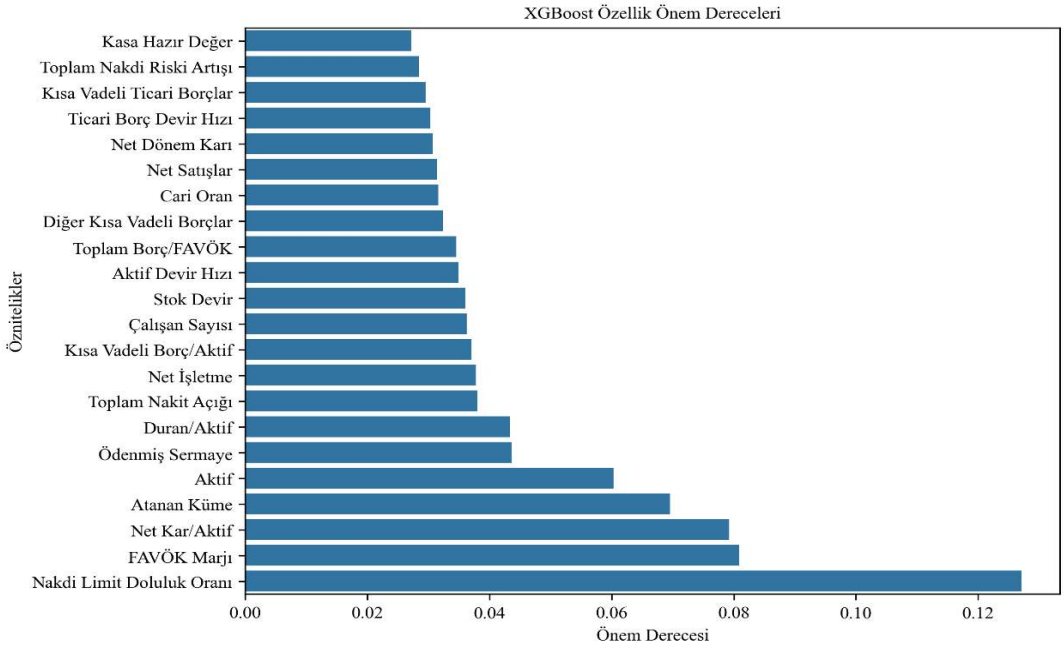
Tablo 10'da sınıflandırma sonuçları görülmektedir. GridSearchCV kullanılarak en iyi parametreler aranmış ve bulunan en iyi parametreler kullanarak yeniden XGBoost modeli kurulmuş ve tahmin sonuçları yukarıdaki tablodaki gibi görülmektedir. Tabloda da görüldüğü üzere başarı oranları yükselmiştir.



Şekil 24. En İyi Parametrelerle XGBoost ROC Eğrisi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 24'te ROC eğrisi sonuçları görülmektedir. Modelin başarı oranının yüksekliğini destekler niteliktedir.

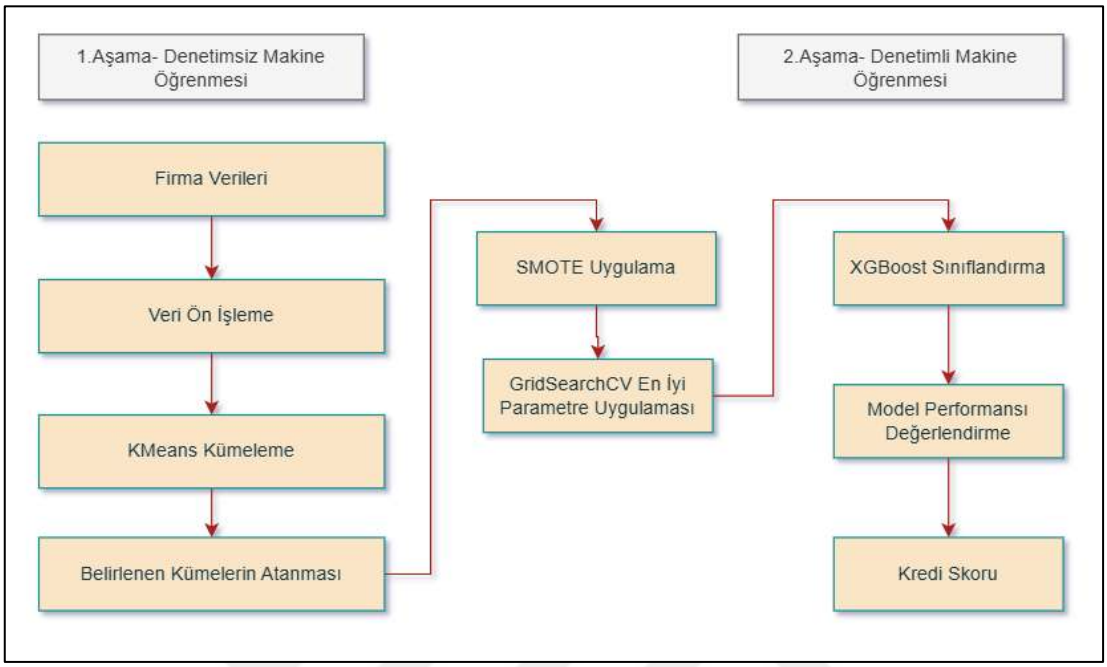


Şekil 25. XGBoost Özellik Önem Dereceleri

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 25'te özniteliklerin önem dereceleri görülmektedir. En etkili olan özniteliğin Nakdi limit doluluk oranı olduğu görülmektedir. Denetimli öğrenme yapılmadan önce KMeans algoritması ile oluşturulan kümelerin çalışmaya etkisinin yüksek olduğu görülmektedir. Atanan küme değişkeni en önemli dördüncü özniteliktir. Bu durum kümelemenin başarılı yapıldığı ve XGBoost algoritması kullanılarak kurulan modelde katkısının yüksek olduğu sonucunu destekler niteliktedir. Çalışmanın birinci aşamasında elde edilen segmentasyon modele dahil edilmeden denetimli makine öğrenme modelleri kullanıldığında başarı oranı daha düşük modeller elde edilmiştir. Bu durum birinci aşamada elde edilen segmantasyonun kullanımının model başarısını yükselttiğini göstermektedir. Şekil 25'te görüldüğü

üzere Atanan kümenin en önemli dördüncü öznitelik olması da bu durumu destekler niteliktedir.



Şekil 26.Sistemin Genel Mimarisi

Kaynak: Yazar tarafından oluşturulmuştur.

Şekil 26'da çalışmada kullanılan sistemin genel mimarisi yer almaktadır. Bu çalışmada önce denetimsiz makine öğrenmesi ile kümeleme yapılmış, atanan kümelerde veriye eklenerek denetimli makine öğrenmesi algoritmaları kullanılmıştır. Kullanılan algoritmalardan en iyi sonuçlar XGBoost ile elde edilmiştir. Bu çalışma ile firmalar için denetimsiz makine öğrenmesi algoritması ile yeni bir segment oluşturulmuş ve firmaların kredi skorları XGBoost model ile tahmin edilmiştir. Çalışmada firmaların kullanılması ve firmalar için yeni bir segment oluşturulması ile farklı bir bakış açısı kazandırılmıştır. Elde edilen kümelemenin bağımsız değişken olarak alınması ve bağımlı değişken olan kredi skorlarının denetimli makine öğrenmesi algoritmalarıyla tahmin edilmesi yönüyle de literatüre katkı sağlanmıştır.

SONUÇ

Bu tez çalışması, banka müşterisi firmaların makine öğrenmesi teknikleriyle analiz edilerek kredi skorlarının oluşturulmasını amaçlamış ve Türkiye bankacılık sektöründe kredi riski değerlendirme süreçlerinin geliştirilmesine yönelik yenilikçi bir yaklaşım sunmuştur. Çalışmada, firmalara ait 19 öznitelik içeren bir veri setinden yararlanarak iki aşamalı bir analiz süreci izlemiştir. İlk aşamada, denetimsiz makine öğrenmesi yöntemiyle K-Means algoritması kullanılarak firmalar dört kümeye ayrılmıştır. Bu kümeler, firmaların finansal ve risk özelliklerine (net satışlar, net kâr, borç oranları, çalışan sayısı gibi) dayalı segmentler oluşturmuştur. Oluşturulan model ve Silhouette Skoru ile Davies-Bouldin İndeksi gibi metriklerle seçilmiştir. İkinci aşamada, denetimli öğrenme teknikleriyle kredi skorlaması gerçekleştirilmiştir. İlk aşamada elde edilen kümeler bağımsız değişken olarak modele eklenmiştir. XGBoost, Karar Ağacı, Rastgele Orman, KNN ve LightGBM algoritmalarıyla kurulan modeller, kredi skorunu dört sınıflı (0-1-2-3) bir çıktı olarak tahmin etmiş ve kesinlik, duyarlılık, F1 skoru gibi performans kriterleriyle değerlendirilmiştir. Kurulan modelde, ROC ve Öğrenme eğrileri ile yapılan incelemeler sonucunda aşırı öğrenme olmadığı görülmüştür.

Çalışmanın temel sonuçları, segmentasyonun kredi skorlama sürecine entegrasyonunun tahmin doğruluğunu artırdığını ortaya koymaktadır. Denetimsiz öğrenme ile oluşturulan dört küme, firmaların heterojen yapısını anlamlı bir şekilde sınıflandırmış ve bu kümeler, denetimli öğrenme modellerinde ek bir bilgi katmanı olarak modeli geliştirmiştir. Özellikle XGBoost algoritması, büyük veri setindeki karmaşık ilişkileri yakalama kapasitesiyle öne çıkmış ve diğer algoritmalara göre daha iyi performans sergilemiştir. Türkiye'de bankacılık sektöründe, firmaların finansal tabloları, risk durumları ve ek göstergeleri bir arada değerlendirildiğinde, makine öğrenmesi modellerinin risk tahmini için güçlü bir araç sunduğu gözlemlenmiştir. Bu sonuçlar, firmaların yalnızca mali tablolarına değil, aynı zamanda operasyonel dinamiklerine dayalı bir analizle daha iyi anlaşılabileceğini göstermiştir. Çalışma sonucunda, makine öğrenmesi yöntemlerinin bankacılık sektöründe kredi tahsis süreçlerini dönüştürme potansiyeline sahip olduğu görülmüştür. Segmentasyon ve skorlama kombinasyonu, firmaların risk profillerini daha ayrıntılı bir şekilde ortaya koyarak bankaların karar alma süreçlerini desteklemektedir.

Bu çalışmanın firmalara yönelik yapılması, firmaların finansal durumları ve risk verilerine ek olarak denetimsiz makine öğrenmesi yöntemleri ile segmentasyon oluşturulması, oluşturulan bu segmentasyonun kredi skorlamada kullanılması yönleriyle literatürdeki mevcut yaklaşımlardan farklılaşarak özgün bir katkı sunması amaçlanmıştır.

Çalışma sonucunda elde edilen skor, bankacılık sektörü ve ilgili paydaşlar için önemli öneriler sunmaktadır. Bankalar, bu çalışmada geliştirilen segmentasyon ve skorlama yöntemini kullanarak, firmalara kredi tahsis süreçlerinde daha bilinçli kararlar alabilir. Özellikle KOBİ'lere yönelik risk analizi, bu firmaların finansmana erişimini kolaylaştırırken bankaların temerrüt riskini azaltabilir. Bir banka, düşük riskli olarak skorlanan firmalara daha uygun faiz oranları sunarak hem müşteri memnuniyetini arttırırken portföyünü de çeşitlendirebilmektedir. Ayrıca, bu modeller, bankaların iç denetim süreçlerinde kullanılarak kredi portföylerinin düzenli bir şekilde izlenmesi ve optimize edilmesi sağlanabilir. Çalışma sonucunda, makine öğrenmesinin finansal veri analizi alanındaki uygulamalarına dair yeni bir perspektif sunmakta ve segmentasyonla zenginleştirilmiş skorlama yaklaşımını ileriye taşıyacak araştırmalara zemin hazırlamaktadır. Türkiye'deki bankacılık sektörüne özgü dinamikler dikkate alındığında, bu çalışma hem teorik hem de uygulamalı katkılarıyla, kredi riski yönetiminde yenilikçi bir yaklaşım oluşturarak literatüre katkı sağlamıştır.

KAYNAKÇA

Akpınar, N. (2019). *Makine öğrenmesi teknikleriyle kredi başvuru skor kartının oluşturulması* [Yüksek lisans tezi]. Yıldız Teknik Üniversitesi.

Altan, G., ve Demirci, S. (2022). Makine öğrenmesi ile nakit akış tablosu üzerinden kredi skorlaması: XGBoost yaklaşımı. *Journal of Economic Policy Researches, 9*(2), 397-424.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609.

Ankerst, M., Breunig, M. M., Kriegel, H. P., ve Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record, 28*(2), 49-60.

Arabacı, H. (2018). Türkiye'de bankacılık sektörünün gelişimi. *Meriç Uluslararası Sosyal ve Stratejik Araştırmalar Dergisi, 2*(3), 25-42.

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

Berger, A. N., ve Udell, G. F. (2006). A more complete conceptual framework for SME finance. *Journal of Banking and Finance, 30*(11), 2945-2966.

Bishop, C. M., ve Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. (Vol. 4, No. 4, p. 738). New York: Springer.

Blagus, R., ve Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics, 14*, 1-16.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145-1159.

Brealey, R. A., Myers, S. C., ve Allen, F. (2014). *Principles of corporate finance*. McGraw-hill.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32.

Breiman, L., Friedman, J., Olshen, R. A., ve Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Brown, I., ve Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033

Bulut, M. A. (2019). *Kredi analizinde makine öğrenmesi kullanımı: Tarımsal kredilerde uygulama örneği.* [Yüksek lisans tezi]. Eskişehir Osmangazi Üniversitesi.

Burez, J., ve Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626-4636.

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.

Chen, T., ve Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Cover, T., ve Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21-27.

Çelik, S. B., ve Mangır, F. (2020). Bankacılık sektörünün dijitalleşmesi: Dünyada ve Türkiye'de durum analizi. *Cyberpolitik Journal, 5*(10), 260-282.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Duda, R. O., ve Hart, P. E. (2006). *Pattern classification.* John Wiley and Sons.

Ester, M., Kriegel, H. P., Sander, J., ve Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. *In KDD* (Vol. 96, No. 34, pp. 226-231).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., ve Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 10, No. 2018). Cham: Springer.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Friedman, J., Hastie, T., ve Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics, 28*(2), 337-407.

Han, J., Kamber, M., ve Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Han, J., Kamber, M., ve Pei, J. (2012). *Data mining: Concepts and techniques*, Waltham: Morgan Kaufmann Publishers.

Hand, D. J., ve Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series a (Statistics in Society), 160*(3), 523-541.

Hastie, T., Tibshirani, R., Friedman, J. H., ve Friedman, J. H. (2009). The elements of statistical learning: data mining*, Inference, and Prediction* (Vol. 2, pp. 1-758). New York: Springer.

He, H., ve Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284.

Ho, T. K. (1995, August). Random decision forests. *In Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

Hull, J. C. (2018). *Risk management and financial institutions* (5th ed.). Wiley.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.

James, G., Witten, D., Hastie, T., ve Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, No. 1). New York: Springer.

Kalaycı, S. (2018). *Makine öğrenmesi yöntemleri ile kredi risk analizi* [Yüksek lisans tezi]. İstanbul Teknik Üniversitesi.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... ve Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*.

Kohavi, R. (1998). Glossary of terms. *Machine Learning, 30*, 271-274.

Lessmann, S., Baesens, B., Seow, H. V., ve Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124-136.

Liaw, A., ve Wiener, M. (2002). Classification and regression by random forest. *R News, 2*(3), 18-22.

MacQueen, J. (1967, January). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* Volume 1: Statistics (Vol. 5, pp. 281-298). University of California press.

Mishkin, F. (2007). Money, banking and financial markets. *New Horizons,* Paris, France.

Mitchell, T. M., ve Mitchell, T. M. (1997). *Machine learning* (Vol. 1, No. 9). New York: McGraw-hill.

Özatay, F. (2011).Türkiye'de bankacılık sektörü: 2001 krizi sonrası dönüşüm. *İktisat İşletme ve Finans, 26*(305), 9-34.

Özgün, S.(2022). *Dengesiz bal peteği veri setinde sınıflandırma performansının analizi* [Yüksek lisans tezi]. Selçuk Üniversitesi.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... ve Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research, 12*, 2825-2830.

Provost, F., ve Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc.

Rokach, L., ve Maimon, O. (2010). Classification trees. *Data Mining and Knowledge Discovery Handbook,* 149-174.

Russell, S. J., ve Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Sandıkcı, Z. N., ve Şaykol, E. (2024). Şimdi al sonra öde kredi sistemi üzerine bir davranış ve skorlama analizi. *Beykent Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, 16*(2), 41-54.

Saunders, A., ve Cornett, M. M. (2021). *Financial institutions management: A risk management approach* (10th ed.). McGraw-Hill Education.

Sokolova, M., ve Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*(4), 427-437.

Thomas, L., Crook, J., ve Edelman, D. (2017). Credit scoring and its applications. *Society for industrial and Applied Mathematics.*

Türkmen, E. (2021). *Makine öğrenmesi yöntemleri ile banka pazarlama tahmini* [Yüksek lisans tezi]. İstanbul Kültür Üniversitesi.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., ve Data, M. (2005). Practical machine learning tools and techniques. *In Data Mining* (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier.

Zhang, T., Ramakrishnan, R., ve Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM Sigmod Record, 25*(2), 103-114.

Zilyas, D., ve Yılmaz, A. (2023). *Makine öğrenmesi yöntemleri ile eğitim başarısının tahmini modeli.* Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 14(3), 437-447.

İnternet Kaynağı

BDDK (2025, 3 Mart). *Aylık bankacılık sektörü verileri.* 3 Mart 2025 tarihinde https://www.bddk.org.tr/Veri/Detay/159 adresinden edinilmiştir.

TOPLUMSAL KATKI

Bu tez çalışmasında, Türkiye'deki bankacılık sektöründe firmaların kredi riski değerlendirmesindeki yetersizlikleri ele alarak, firmaların finansmana erişim zorluğu gibi önemli bir toplumsal problemi incelemiştir. Geleneksel yöntemlerin hassasiyet eksikliği, ödeme kapasitesi yüksek firmaların reddedilmesine veya riskli firmalara kredi verilmesine yol açarak bankaların kayıplarını, firmaların iflasını ve dolaylı olarak işsizliği arttırmaktadır. Bu durum, ekonomik istikrarı ve toplumsal refahı tehdit ederken, bölgesel kalkınma farklarını derinleştirmektedir. Çalışma, makine öğrenmesi ile bu soruna çözüm sunmayı amaçlamıştır. Çalışmadaki veri setiyle, K-Means algoritması ile segmentasyon, XGBoost algoritması ile dört sınıflı kredi skoru üretilmiştir. Sonuçlar, modellerin risk tahmini doğruluğunu artırdığını göstermiştir. Bu durum firmaların finansmana erişimini kolaylaştırarak istihdamı ve ekonomik büyümeyi destekler. Ayrıca, politika yapıcılara veri temelli öneriler sunarak risk yönetimini modernize etme fırsatı vermektedir. Sonuçlar, bankaların kredi süreçlerini iyileştirerek daha doğru tahsis sağlarken temerrüt riskini azaltmaktadır. Bankalar bu skorları otomatik sistemlere entegre ederek hızlı ve objektif kararlar alabilirler.