

İÇİNDEKİLER

Sıra No	Makale Başlığı	Yazar(lar)	Makale Türü	Sayfa
No	Paper Title	Author(s)	Paper Type	Page
1	Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü <i>Performance analysis of machine learning algorithms for predicting bank loan: Superiority of ensemble learning algorithms</i>	Enes Çelik, Önder Gür	Araştırma Makalesi Research Paper	1-20

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

Enes Çelik ^{*1}, Önder Gür^{2,3}

¹ Kırklareli Üniversitesi, Babaeski Meslek Yüksekokulu, Bilgisayar Bilimleri, Kırklareli, Türkiye, ORCID: 0000-0002-3282-865X

^{2,3} Kırklareli Üniversitesi, Babaeski Meslek Yüksekokulu, Bankacılık ve Sigortacılık, Kırklareli, Türkiye. Ege Üniversitesi, Tire Kutsan Meslek Yüksekokulu, Muhasebe ve Finansman, İzmir, Türkiye, ORCID: 0000-0003-3249-4300

Geliş Tarihi:05.12.2023

Kabul Tarihi:02.04.2024

Özet

Banka kredisi tahmini, finansal risklerin öngörülmesi ve kredi başvurularının etkin bir şekilde değerlendirilmesi bakımından büyük bir öneme sahiptir. Bu amaçla yapılan araştırmada, çeşitli makine öğrenmesi algoritmaları banka kredisi tahmini problemine yönelik performanslarını değerlendirmek amacıyla incelenmiştir. Bu deneylerde genellikle, algoritmaların tahmin doğruluğu, kesinliği, hata oranları gibi performans metrikleri üzerinden ölçülmüştür. Farklı algoritmaların, veri setlerinin büyüklüğü, veri özellikleri ve sınıflandırma performansı gibi faktörler açısından karşılaştırılmıştır. Bu karşılaştırmaların temel amacı, belirli bir veri seti veya problem bağlamında hangi algoritmanın daha iyi performans sergilediğini anlamaktır. Bu çalışmalar, banka kredisi tahmini gibi finansal uygulamalarda daha etkili ve doğru kararlar alınmasına katkıda bulunmuş ve uzman kişilere karar almada yardımcı olur. Algoritmaların avantajları ve dezavantajları, genelleştirilebilirlikleri ve hesaplama karmaşıklıkları gibi faktörler de bu deneylerde dikkate alınmıştır. Sonuç olarak, bu karşılaştırmalı deneyler, finansal alanlarda makine öğrenmesi algoritmalarının performanslarını değerlendirmek için önemli bir rol oynamıştır. Bu çalışmanın sonuçları, Topluluk Öğrenmesi altında bulunan algoritmaların diğerlerine kıyasla banka kredisi tahmini gibi kritik finansal konularda daha başarılı olduğunu ortaya koymuştur.

Anahtar Kelimeler: Banka kredi tahmini, Makine öğrenmesi, Topluluk öğrenmesi

Performance analysis of machine learning algorithms for predicting bank loan: Superiority of ensemble learning algorithms

Abstract

Bank loan prediction is of great importance in predicting financial risks and effectively evaluating loan applications. In the research conducted for this purpose, various machine learning algorithms were examined to evaluate their performance on the bank loan prediction problem. In these experiments, the algorithms were generally measured through performance metrics such as prediction accuracy, precision, and error rates. Different algorithms were compared in terms of factors such as size of datasets, data characteristics and classification performance. The main purpose of these comparisons is to understand which algorithm performs better in the context of a particular data set or problem. These studies have contributed to making more effective and accurate decisions in financial applications such as bank loan prediction and help experts in decision-making. Factors such as advantages and disadvantages of the algorithms, their generalizability and computational complexity were also taken into account in these experiments. Consequently, these comparative experiments have played an important role in evaluating the performance of machine learning algorithms in financial fields. The results of this study revealed that the algorithms under Ensemble Learning are more successful in critical financial issues such as bank loan prediction compared to others.

Keywords: Bank loan prediction, Ensemble learning, Machine learning

*Sorumlu yazar (Corresponding author): Enes Çelik, enes.celik@klu.edu.tr.

Künye Bilgisi: Çelik, E., Gür, Ö. (2024). Banka Kredisi Tahmini İçin Makine Öğrenmesi Algoritmalarının Performans Analizi: Topluluk Öğrenmesi Algoritmalarının Üstünlüğü. Artibilim: Adana Alparslan Türkeş Bilim ve Teknoloji Üniversitesi Fen Bilimleri Dergisi, 7(1), 1-20. Doi: 10.55198/artibilimfen.1400881

1. Giriş

Finansal kuruluşların mali açıdan başarısız olmasında kredilendirme önemli bir rol oynamaktadır. Bu sebeple finansal kuruluşlar, kredi talebinde bulunanların aldıkları borçları zamanında ödeyip ödeyemeyeceğini öngörmeye çalışmaktadır. Özellikle 2020 yılında yaşanan pandemiden kaynaklanan ekonomik durgunluğun önüne geçmek amacıyla hükümetlerin ve merkez bankalarının, yaşamları ve geçim kaynaklarını korumaya yardımcı olmak için özel sektörün daha fazla borçlanmasını desteklemesi nedeniyle küresel borç yaklaşık 226 trilyon USD'ye yükselmiştir. Ancak bu durumun özellikle finansman koşullarının sıkılaşmaya başladığında kırılganlıkları arttırabileceği yönünde tahmin edilmektedir. Kırılganlığın en önemli etkisi, kredi veren finans kuruluşlarının verdiği krediyi zamanında tahsil edememesi veya kredinin ödenmemesi şeklinde sonuçlanabilir. Diğer bir ifadeyle batık kredilerin oluşması ya da ödemelerin aksaması finans kuruluşlarının en büyük gelir kaynağı olan faiz gelirlerinin etkilenmesi anlamı gelmektedir. İşte bu noktada olası kredi geri dönüşlerinde yaşanabilecek problemlerin önüne geçmek amacıyla kredilendirmelerde, bir işletmenin finansal kuruluşlara özellikle bankalara karşı, yükümlülüklerini yerine getirip getirmeme gücünü değerlendirmek için nicel ve nitel teknikler geliştirilmiştir. Finansal kuruluşlar, kredi portföylerinin büyümesiyle kredilendirme uygulamalarının doğruluğunu arttırmak için çeşitli alternatifleri aktif olarak araştırarak bu süreci yönetmeye çalışmaktadır [1]. Kuruluşlar bu süreci yönetmek için çeşitli istatistik temelli araçlar kullanılabilen veya uzman görüşüne dayalı değerlendirmelere başvurulabilmektedir [2]. Ancak insan kararlarına bağlı kalındığında çeşitli hatalara açık olunmasına sebebiyet vermekte ve bu durum maliyetleri daha da arttırmaktadır [3]. Günümüzde veri madenciliğine dayalı istatistik, makine öğrenmesi, yapay zekâ, örüntü tanıma gibi çeşitli disiplinlerle desteklenen farklı türde tahmine dayalı analitik teknikler geliştirilmiştir. Özellikle makine öğrenmesi alanındaki yeni teknikler, veriler arasındaki doğrusal olmayan kalıpları ve zamansal bağımlılıkları tanıyarak büyük veri kümesinden verimli bir şekilde bilgi yakalama yetenekleri sayesinde dünya çapında bir üne kavuşmuştur [4]. Finansal kuruluşlar, makine öğrenmesinin yardımıyla müşteriden müşteriye değişen, finansal ve finansal olmayan bilgi ve belgelere dayanan farklı yapılarıdaki verileri kredi değerlendirme sürecinde kullanarak önemli risk değerlendirmeleri yapılabilir hale gelmiştir [5]. Bununla birlikte kredilerin tahsisine yönelik tahmin için hangi sınıflandırma tekniğinin kullanılacağı konusu çok zor ve uğraştırıcı bir sorun olabilmektedir. Sınıflandırmalarda en önemli faktör başarılı bir sınıflandırıcı model oluşturabilmektir. Ancak model başarısını etkileyen birçok neden bulunmaktadır. Kullanılan test yöntemlerinin yanı sıra veri setine ait özellikler de başarıyı etkileyen unsurlardandır [6]. Bu çalışmada, finansal ve finansal olmayan müşteri bilgilerinin bulunduğu bir veri setiyle kredi kullanılarak kullandırılmayacağına makine öğrenmesiyle çözüm üretilmeye çalışılmıştır. Makine öğrenmesine dayalı kredi değerlendirme yöntemlerinin yanlış sınıflandırmasına bağlı olarak maliyeti oluşturan faktörlerin etkisi de araştırılmalıdır. Makine öğrenmesindeki sınıflandırma hatalarının en büyük etkeni veri setlerinin yapısıdır. Veri setlerinde, bazı değişkenler arasında ilişki olabileceği gibi birbirleriyle alakasız ve gereksiz değişkenler de olabilir [7]. Ayrıca farklı nedenlerden kaynaklı olarak veri setlerinde uygun olmayan (eksik, hatalı, tutarsız, kirli) değerler bulunabilmektedir. Tespit edilen değişkenlerin veya değişken değerlerinin model frekanslarını karıştıracağından ve modelin tahmin gücünü azaltacağından düzeltilmesi, değiştirilmesi ya da silinmesi gerekmektedir [8]. Bu işlemler, veri setindeki geçerliliğini kaybetmiş verileri manipüle etmeden seçilen sınıflandırıcı algoritmalar için gereken

değerlere dönüştürmeye yönelik yapılmalıdır [9]. Çalışmada, sınıflandırmalar arasındaki küçük bir farkın bile önemli tasarruflara dönüşebileceği bilinciyle kredinin kullanılarak kullanılmayacağına farklı tekniklerin metrik sonuçları karşılaştırılarak başarılı olduğu tespit edilen tekniklerle karar verilmeye çalışılmıştır. Çalışmada, kredi veren kuruluşlara yardım edebilmek amacıyla makine öğrenmesi yöntemleri kullanılarak tahminler yapılmıştır. Ayrıca bu çalışmanın özgün tarafı ve literatüre katkıları şunlardır:

- I) Finansal kuruluşlara gelen kredi taleplerine yönelik yeni teknikler kullanılarak tahminleme için kullanılabileceği başarılı bir model önerisi geliştirildi,
- II) Finansal kuruluşların kredi taleplerine yönelik finansal ve finansal olmayan bilgilerin bir arada kullanılması,
- III) Finansal kuruluşlarda yanlış kredi tahsisinden kaynaklı oluşabilecek zaman kaybını ve finansal maliyet zararlarını azaltmaya yardımcı olur,
- IV) Karar alma süreçlerinde kullanılan algoritmaların avantajlarını ve dezavantajlarını ortaya koyarak daha etkili kararlar alınmasına katkıda bulunur,
- V) Finansal alanlarda makine öğrenmesi algoritmalarının etkin bir şekilde kullanılmasına ve daha güvenilir ve açıklanabilir finansal kararların alınmasına katkıda bulunur.

2. Literatür taraması

Bu alanda finans kuruluşlarının en büyük problemi olan kredi riskinin değerlendirilmesine yönelik farklı niteliğe sahip birçok çalışma yapılmıştır. Kredilerin sorunlu hale gelmeden önce tahmin etmeye çalışmak finansal kuruluşlar açısından oldukça önemlidir [39]. Çoğu ulusal ve uluslararası finans kuruluşu temerrüde düşen kredilerden kaynaklı büyük kayıplar yaşamaktadır [15]. Deneyim eksikliği ve borçlunun kredi geçmişi hakkında eksik veya belirsiz bilgiler, kredi veren kuruluşların risklerini artırabilmektedir [10]. Nedeni ne olursa olsun kredilendirmeye yönelik başarısızlığı tahmin etmeye çalışmak, zaman kaybını ve kaynakların yanlış tahsisini azaltmaya yardımcı olmaktadır [11]. Geçmişte finansal kuruluşlarının kredi tahsisinden kaynaklı başarısızlığını önlemeye yönelik birçok çalışma yapılmıştır. Beaver'ın 1966'da tek değişkenli analizi [37], Altman'ın 1968'deki çoklu diskriminant analizi [38] ve Meyer-Pifer'in 1970 yılındaki çoklu regresyon analizi çalışmaları bu alanda yapılan ilk çalışmalar olarak gösterilebilir.

Serengil vd. yaptıkları çalışmada kredi riskini tahmin ederek, işletmelere ve bireylere kredi sağlayan finansal kuruluşlara yardım etmeyi amaçlamışlardır. Özel bir bankaya ait veri setinin sınıf ağırlıkları kullanılarak yapılan çalışmada lojistik regresyon (LR), rastgele orman (RF), destek vektör makineleri (SVM), gradyan arttırma tabanlı (XGBoost ve LightGBM) ve uzun kısa süreli bellek (LSTM) algoritmalarının performanslarını özgüllük, kesinlik, duyarlılık, f1-skor ve dengesizlik doğruluğu gibi metriklerle ölçmüşlerdir. LightGBM algoritması, özgüllük %90, kesinlik %87, duyarlılık %77, f1-skor %60 ve dengesizlik doğruluğunu %82 elde ederek en başarılı algoritma olmuştur [3].

Altan ve Demirci çalışmalarında kredi ürünlerini kullanan müşterilerinin ayırt edilmesi bankaların karlılıklarını ve pazardaki rekabet gücünü arttırabileceğini savunmuştur. Bu amaçla borsaya kotasyon olmuş işletmelerin nakit akış tablolarından faydalanılarak modeli geliştirilmeyi hedeflemişlerdir. Çalışmada XGBoost, Gradient Boosting ve Yapay Sinir Ağları yöntemleri kullanılmış ve karşılaştırılmıştır. Sonuç olarak XGBoost yöntemi %80 doğruluk ile en başarılı model olmuştur [5].

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

Moscato vd. çalışmalarında bir kredinin geri ödenip ödenmeyeceğini tahmin etmek için en iyi performansa sahip kredi riski puanlama modelini tespit etmeyi amaçlamışlardır. Çalışmada Lojistik Regresyon, Rastgele Orman ve Yapay Sinir Ağları kullanılmış ve tahminleri karşılaştırılmıştır. Sonuçlara göre en iyi AUC %71 değerini sağlayan yeniden örnekleme metodolojisiyle kullanılan Rastgele Orman yöntemi olmuştur [10].

Li ve Chen çalışmalarında kredi riskini tahmin etmek için Rastgele Orman, Adaboost, XGBoost, LightGBM toplu öğrenme algoritmaları ile Yapay Sinir Ağları, Lojistik Regresyon, Karar Ağacı, Destek Vektör Makinesi ve Naive Bayes bireysel öğrenme algoritmalarını karşılaştırmıştır. Adaboost dışındaki toplu öğrenme algoritmalarının Doğruluk, AUC, Kolmogorov–Smirnov istatistiği ve Brier Skoru metriklerinde en yüksek performansı gösterdiğini belirtmişlerdir [12].

Sönmez bir çalışmasında bankaların ve diğer finansal kuruluşların müşterileri için bir kredi skorunun belirlenmesine yardımcı olmayı amaçlamıştır. Kredi talebinde bulunan bireysel müşterilerin talebine cevap verilmesi adına Yapay Sinir Ağları ve Karar Ağacı algoritmasını temel alan modeller arasında karşılaştırma yapılmıştır. Yapay Sinir Ağları, Karar Ağacı algoritmasına göre kredi skorlamada daha iyi sınıflandırma başarısı göstermiştir [13].

Bahnsen vd. çalışmalarında bir kredi skoru modeli seçildikten sonra kredinin işleme bağlı finansal maliyetleri azaltmak amacıyla model geliştirmeyi hedeflemişlerdir. Bu amaçla geliştirdikleri maliyete duyarlı Lojistik Regresyon modeli ile Rastgele Orman ve Karar Ağacı modelleri veri setinin %50'si ile eğitilmiş, %25'i ile test edilmiş ve %25'i ile doğrulanmıştır. Elde edilen sonuçlar karşılaştırıldığında geliştirdikleri maliyete duyarlı Lojistik Regresyon modeli ile daha fazla tasarruf edildiğini açıklamışlardır [14].

Son yıllarda, istatistiksel tekniklerin gelişmesiyle finans kuruluşlarının kredi kalitesini değerlendirmek için literatürde görüldüğü üzere çok sayıda model ve yaklaşım önerilmiştir. Birçok çalışmada kredi değerlendirmelerinde model girdisi olarak ana finansal göstergeler kullanılmıştır. Bununla birlikte, bu modeller girdi olarak sadece finansal oranları kullanması, doğrusal olmayan kalıpları yakalamada yetersiz kalmaya başlamıştır [4].

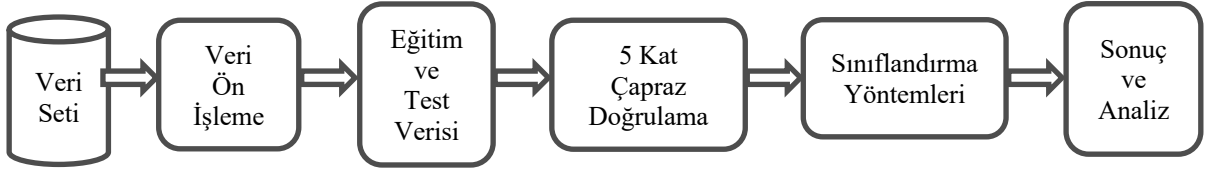
3. Yöntem ve materyal

Makine öğrenmesi yöntemleri mühendislik, sağlık, telekomünikasyon vb. alanlarda olduğu gibi finans alanında da kullanımı her geçen gün artmaktadır. Literatür araştırmasında gösterildiği üzere Karar Ağacı, Destek Vektör Makinesi, Yapay Sinir Ağları, Rastgele Orman, Lojistik Regresyon ve Naive Bayes ağları sınıflandırma tekniklerinin alanda sıklıkla kullanıldığı ve yeni geliştirilen yöntemlerden olan XGBoost, LightGBM ve Catboost gibi gradyan tabanlı torbalayıcı yöntemlerin de kullanılmaya başlandığı görülmüştür. Bu çalışmada daha önce literatürde çok denenmemiş son teknolojiler XGBoost, LightGBM ve Catboost kullanılmış ve diğer sınıflandırma algoritmaları ile kıyaslanmıştır.

Sınıflandırma algoritmaları, veri setindeki farklı sınıf etiketleri üzerinden ayrılan farklı oranlardaki kısımlar üzerinde eğitimi gerçekleştirilmekte ve sonrasında veri setinin kalan kısmı üzerinden test edilmektedir. Algoritmanın sınıf etiketleri ile hedef değişkenin sınıf etiketleriyle karşılaştırılarak performans ölçümü için parametreler oluşturulmaktadır [9]. Moscato bir çalışmalarında ifade ettiği üzere büyük hacimli verilere dayanan veri setleri dengesiz olabilmektedir. Ayrıca tüm değişkenlerin kullanılması bir yandan kapsamı artırırken diğer yandan modelin başarısını

azaltabilmektedir. Tahmin yapılmadan önce veri setlerinin analiz edilmesi gerekmektedir, yoksa bu durum yanıltıcı sonuçlara neden olabilmektedir. Bu nedenle veri ön işleme, sürecin ilk adımı olarak gerçekleştirilmelidir. Daha sonra sınıflandırıcının oluşturduğu modelin başarısı parametreler yardımıyla çeşitli metriklerle ölçülerek değerlendirilmelidir [16].

Çalışmada, kredi kullanmış müşterilere ait bilgilerin bulunduğu veri seti analiz edilerek, değişkenler üzerinde önce temizlik daha sonra gerekli dönüştürme işlemleri yapılmıştır. Gerekli hazırlıklar yapıldıktan sonra oluşturulan veri setinin %80'i eğitim, %20'si test için kullanılmıştır (5 kat çapraz doğrulama). Modellerin doğruluk, kesinlik, geri çağırma, f1-skor, ortalama mutlak hata ve ortalama karekök hatası metrikleriyle performansı ölçülmüştür. Modelin akış diyagramı Şekil 1'de verilmiştir.



Şekil 1. Modelin akış diyagramı

3.1. Veri seti

Çalışmada kullanılan veri seti, makine öğrenmesi çalışmalarına katkı sağlayabilmek adına kamuya açık veri bilimi topluluğu olan Kaggle isimli internet sitesinden sağlanmıştır. Kullanılan veri seti, bir bankanın kredi kullanan müşterilerine ait finansal ve finansal olmayan bilgilerini içerir. Veri seti 100.000 adet satır, 18 adet değişken ve 1 adet kredi durumunu gösteren (Loan Status) sınıflandırıcı sütundan oluşmaktadır. Veri setinde yapılan sınıflandırmaya göre 77.361 müşteri kredi borcunu ödemişken, 22.639 müşteri temerrüde düştüğünü göstermektedir [17]. Veri setleri farklı nedenlerden kaynaklı olarak eski, eksik veya uygun olmayan değerleri içerebilmektedir. Amaçlar doğrultusunda modellerin başarılı olabilmesi için verileri temizleme ve dönüştürme ön işlemine tabi tutulmasını gerektirmektedir [18]. Veri hazırlama aşaması, ham veriden yola çıkarak kullanılması elverişli olan veri setine ulaşmaya kadar gereken işlemleri kapsamaktadır [19]. Veri setinin özellikleri, veri yapıları ve açıklamaları ile Tablo 1.'de verilmiştir.

Çalışmada kullanılan veri setinin ham hali olması sebebiyle uygulamada kullanılmadan önce ön analizi yapılmıştır. Yapılan analiz sonucunda değişkenlerin bazılarında hatalı, anlamsız ve eksik olan veri hücreleri bulunduğu anlaşılmıştır. Veri setleri içinde eksik değer içeren alanlar varsa global bir değer verilerek, medyan değeri verilerek veya kategorilere göre ortalama değer verilerek eksik veriler tamamlanabilir ya da bu alanların yok sayılarak silinebilir [20]. Çalışmada kullanılan veri setindeki boş hücreleri az olan “maksimum açık kredi tutarı”, “vergi alacağı” ve “iflas” değişkenlerinin ortalama değerleri alınarak boş veri hücreleri tamamlanmıştır. Ayrıca anlamsız değer taşıyan “kredilendirme numarası”, “müşteri numarası” ve veri hücrelerinin %50'sinin boş olduğu anlaşılan “son temerrütten sonra geçen süre” sütunları ile veri hücrelerinde farklı sayılarda boşluğu olan 206 satır silinmiştir. Sonrasında veri seti, algoritmalar için gereken uygun veri tipine dönüştürme işlemleri yapılmıştır. Bu işlemler

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

sonucunda veri seti 99.794 adet satır, 15 adet değişken giriş olarak modele verilir ve 1 adet sınıflandırıcı sütun (Kredi Durumu) olarak kullanılmak üzere hazır hale getirilmiştir.

Tablo 1. Veri seti öznelikleri

Değişken adı	Değişken açıklaması	Değişkenin veri tipi
Loan ID	Kredilendirme Numarası	Nümerik
Customer ID	Müşteri Numarası	Nümerik
Loan Status	Kredi Durumu (Hedef değişken)	Kategorik
Current Loan Amount	Mevcut Kredi Tutarı	Nümerik
Term	Kredilendirme Dönemi	Kategorik
Credit Score	Kredi Skoru	Nümerik
Annual Income	Yıllık Gelir	Nümerik
Years in current job	Çalışma Süresi	Kategorik
Home Ownership	Ev Sahipliği	Kategorik
Purpose	Kredi Kullanım Amacı	Kategorik
Monthly Debt	Aylık Borç Tutarı	Nümerik
Years of Credit History	Kredi Geçmişi	Nümerik
Months since last delinquent	Son Temerrütten Sonra Geçen Süre	Nümerik
Number of Open Accounts	Açık Hesap Sayısı	Nümerik
Number of Credit Problems	Kredi Problemlerinin Sayısı	Nümerik
Current Credit Balance	Mevcut Kredi Bakiyesi	Nümerik
Maximum Open Credit	Maksimum Açık Kredi Tutarı	Nümerik
Bankruptcies	Başarısızlık / İflas	Kategorik
Tax Liens	Vergi Alacakları / Hacizleri	Kategorik

3.2. Sınıflandırma yöntemleri

3.2.1. Karar Ağaçları (Decision Tree - DT)

Karar ağaçları, tahmine dayalı modelleme ve karar verme için kullanılan temel bir makine öğrenme algoritmasıdır. Karar ağaçları, her düğümün veri kümesindeki belirli bir özelliğe dayalı bir karar noktasını temsil ettiği, düğümlerden ve dallardan oluşan hiyerarşik yapılardır. Kök düğüm tüm veri kümesiyle başlar ve bir dizi yinelemeli bölme yoluyla algoritma, verileri alt kümelere bölmek için en bilgilendirici özellikleri belirler. Bu bölünmeler, sınıflandırma görevleri için Gini Safsızlığı, Entropi, bilgi kazancı ve Regresyon görevleriyle ortalama karesel hata gibi kriterlere göre belirlenir. Süreç, nihai tahminleri veya kararları sağlayan yaprak düğümleri oluşturan bir durdurma kriteri karşılanıncaya kadar

devam eder. Karar ağaçları, yapıları kolay görselleştirmeye ve karar verme sürecinin anlaşılmasına olanak tanıdığından yorumlanabilirlik açısından mükemmeldir [21].

3.2.2. Destek Vektör Makinesi (Support Vector Machine - SVM)

Destek Vektör Makineleri, makine öğreniminde hem sınıflandırma hem de regresyon görevleri için kullanılan güçlü denetimli öğrenme algoritmalarıdır. Destek Vektör Makinesi, hiperdüzlem ile farklı sınıfların en yakın veri noktaları arasındaki mesafe olan marjı maksimuma çıkararak sınıfları en iyi şekilde ayıran veya bir regresyon görevinin sonucunu tahmin eden en uygun hiperdüzlemi bulmayı amaçlar. Algoritma, giriş verilerini, sınıfları etkili bir şekilde ayırmak için bir hiperdüzlemin oluşturulabileceği daha yüksek boyutlu bir alana dönüştürerek çalışır. Hiperdüzlemin seçimi, marjı maksimuma çıkarmanın seçilmesine dayanır, bu da sınıflandırma veya regresyon hatasını en aza indirir. Destek Vektör Makinesi dönüşümü açıkça hesaplamadan girdi verilerini bu yüksek boyutlu alana örtülü olarak eşlemek için bir çekirdek fonksiyonu kullanır, böylece daha yüksek boyutlarda verimli hesaplamaya olanak tanır [22].

3.2.3. Gauss Naif Bayes (Gaussian Naive Bayes - GNB)

Gauss Naif Bayes, Bayes teoremi prensiplerine göre çalışan ve veri kümesindeki özelliklerin birbirinden bağımsız olduğunu varsayan olasılıksal bir sınıflandırma algoritmasıdır. Özellikle Gauss dağılımını takip eden sürekli verilerle uğraşırken, sınıflandırma görevleri için özellikle yararlıdır. "Naif" yön, özellik bağımsızlığı varsayımını ifade eder; bu her özelliğin sonuca bağımsız ve eşit bir katkı sağladığı anlamına gelir. Gauss Naif Bayes algoritması, verilerde mevcut özellikler göz önüne alındığında sınıfın olasılığının belirlenmesini içeren Bayes teoremini kullanarak belirli bir örneğin belirli bir sınıfa ait olma olasılığını hesaplar. Bunu, özelliklerin normal bir dağılım izlediğini varsayarak, Gauss dağılımını kullanarak her sınıf için her bir özelliğin olasılık yoğunluklarını hesaplayarak yapar [23].

3.2.4. K-En Yakın Komşular (K-Nearest Neighbors - KNN)

K-En Yakın Komşular, hem sınıflandırma hem de regresyon görevleri için kullanılan parametrik olmayan ve örnek tabanlı bir denetimli öğrenme algoritmasıdır. Özellik uzayında benzer veri noktalarının birbirine yakın olduğu varsayılarak benzerlik ilkesine göre çalışır. K-En Yakın Komşular açık bir eğitim aşaması yer almaz; bunun yerine algoritma tüm eğitim veri kümesini ezberler. K-En Yakın Komşular, yeni görünmeyen bir veri noktasıyla sunulduğunda, özellik uzayındaki en yakın K adet komşusunu, genellikle Öklid mesafesi olan seçilmiş bir mesafe ölçüsüne göre tanımlar. Bu komşular, yeni veri noktası ile eğitim setindeki diğer tüm noktalar arasındaki mesafe ölçülerek belirlenir. Algoritma daha sonra yeni veri noktasını, sınıflandırma görevleri için en yakın K adet komşusu arasında çoğunluk oyu ile sınıflandırır ve bu komşular arasında en yaygın olan sınıf etiketini atar [24].

3.2.5. Yapay Sinir Ağları (Artificial Neural Network - ANN)

Yapay Sinir Ağı, biyolojik nöronların iletişim şeklini taklit ederek karmaşık sorunları çözmek için tasarlanmış, insan beyninin sinir yapısından ilham alan hesaplamalı bir modeldir. Giriş gizli ve çıkış katmanları halinde düzenlenmiş birbirine bağlı düğümlerden oluşan Yapay Sinir Ağı, bilgiyi bir dizi matematiksel işlem aracılığıyla işler. Her düğüm veya yapay nöron, giriş sinyallerini alır, bu girişlere ağırlıklar uygular, bunları toplar ve ardından sonucu bir çıktı üretmek için bir aktivasyon fonksiyonundan geçirir. Eğitim sırasında ağ, tahmin edilen ve gerçek çıktılar arasındaki farkı en aza indirmek için geri yayılım gibi yinelemeli öğrenme algoritmaları yoluyla ağırlıklarını ayarlar ve yeni,

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

görünmeyen veriler üzerinde genelleme yapma ve doğru tahminler veya sınıflandırmalar yapma yeteneğini geliştirir. Yapay Sinir Ağının uyarlanabilirliği, ölçeklenebilirliği ve doğrusal olmayan ilişkileri modelleme yeteneği, bunları görüntü ve konuşma tanıma, doğal dil işleme, finansal tahmin ve sağlık hizmetleri teşhisleri dahil olmak üzere çeşitli alanlarda uygulanabilir [25].

3.2.6. Rastgele Ormanlar (Random Forest - RF)

Rastgele Orman, eğitim sırasında çok sayıda karar ağacı oluşturarak ve sınıfların modunu (sınıflandırma) veya bireysel ağaçların ortalama tahminini (regresyon) çıkararak çalışan, makine öğreniminde bir topluluk öğrenme tekniğidir. Ormandaki her karar ağacı, eğitim verilerinin rastgele bir alt kümesi kullanılarak oluşturulur ve ağacın her düğümünde, çeşitli ağaç kümelerine yol açacak şekilde, bölme için rastgele bir özellik alt kümesi dikkate alınır. Tahminlerde bulunurken, her ağaçtan yeni veriler geçer ve nihai tahmin, genellikle sınıflandırma için bir oylama mekanizması veya regresyon için ortalama alma aracılığıyla, tüm ağaçlardan elde edilen sonuçların toplanmasıyla belirlenir. Rastgele Ormanlar, birden fazla zayıf öğreniciyi (karar ağaçları) birleştirerek, gürültüye karşı dayanıklılık gösteren ve görünmeyen verilere iyi bir şekilde genelleme yapma eğiliminde olan sağlam, daha doğru bir model oluşturmak için aşırı uyumu azaltır. Yüksek boyutlu verileri işleme, önemli özellikleri belirleme ve sağlam tahminler sağlama yeteneği onu finans uygulamalarında popüler bir seçim haline getirdi [26].

3.2.7. Lojistik Regresyon (Logistic Regression - LR)

Lojistik Regresyon, belirli bir sınıfa ait bir gözlemin olasılığını tahmin etmeyi amaçlayan, ikili sınıflandırma görevleri için kullanılan istatistiksel bir yöntemdir. Öncelikle regresyondan ziyade sınıflandırma için kullanılır. Bu algoritma, belirli bir olayın meydana gelme olasılığını tahmin ederek bağımsız değişkenler ile kategorik bağımlı değişken arasındaki ilişkiyi modeller. Giriş özelliklerini 0 ile 1 arasındaki bir olasılık değerine eşlemek için lojistik işlevi (sigmoid işlevi) kullanarak çalışır. Lojistik regresyon modeli, giriş özelliklerinin doğrusal bir kombinasyonunu hesaplar ve bu daha sonra aşağıdaki denklemi kullanarak bir olasılığa dönüştürülür. Eğitim sırasında model, tipik olarak gradyan inişi gibi optimizasyon tekniklerini kullanarak, girdi verilerini, gerçek sonuçları gözlemeleme olasılığını en üst düzeye çıkararak parametreleri ayarlar. Model eğitildikten sonra, bir sınıfa veya diğerine ait olma olasılığını belirleyerek yeni verileri sınıflandırabilir ve genellikle nihai sınıflandırmaya karar vermek için bir eşik değeri uygular [27].

3.2.8. Stokastik Gradyan Azalma (Stochastic Gradient Descent - SGD)

Stokastik Gradyan Azalması, makine öğrenmesi ve matematiksel optimizasyonda kullanılan temel bir optimizasyon algoritmasıdır. Fonksiyonun değerini minimuma indiren yönde yinelemeli olarak hareket ederek bir fonksiyonun minimumunu bulmayı amaçlayan yinelemeli bir yöntemdir. Tüm veri kümesinin eğimini hesaplayan geleneksel Gradient Descent'ten farklı olarak Stokastik Gradyan Azalması, her yinelemede daha küçük bir alt küme veya veri kümesinden tek bir rastgele gözlem üzerinde çalışır, bu da onu özellikle büyük veri kümeleri için hesaplama açısından verimli hale getirir. Bu rastgelelik, optimizasyon sürecine gürültü katar ancak daha hızlı yakınsamaya izin verir ve Stokastik Gradyan Azalmayı dışbükey olmayan ve yüksek boyutlu optimizasyon sorunları için oldukça uygun hale getirir. Stokastik Gradyan Azalması, parametre uzayında atılan adımların boyutunu belirleyen öğrenme oranını ayarlayarak, verilen amaç fonksiyonunu en aza indiren optimal parametre grubuna

kademeli olarak yaklaşmayı amaçlamaktadır. Stokastik doğasına rağmen, hiper parametrelerin ve öğrenme oranlarının uygun şekilde ayarlanmasıyla Stokastik Gradyan Azalması genellikle tatmin edici bir çözüme yakınsar ve bu da onu makine öğreniminde tercih sebebi olur [28].

3.2.9. İkinci Dereceden Diskriminant Analizi (Quadratic Discriminant Analysis - QDA)

İkinci Dereceden Diskriminant Analizi, belirli bir gözlemin bir veri kümesi içindeki belirli bir sınıfa ait olma olasılığını modellemek ve tahmin etmek için kullanılan istatistiksel bir sınıflandırma tekniğidir. Doğrusal Diskriminant Analizinin (LDA) bir çeşididir ancak her sınıfın kendi kovaryans matrisine sahip olduğunu varsayarak sınıflar arasındaki karar sınırının modellenmesinde daha fazla esnekliğe izin vermesi açısından farklılık gösterir. İkinci Dereceden Diskriminant Analizi, her sınıfa ait verilerin normal şekilde dağıldığını varsayar ve bu varsayımı, her sınıf için ortalama ve kovaryans matrisi de dahil olmak üzere sınıf yoğunluklarının parametrelerini tahmin etmek için kullanır. Tüm sınıfların aynı kovaryans matrisini paylaştığını varsayan Doğrusal Diskriminant Analizinden farklı olarak İkinci Dereceden Diskriminant Analizi, bu kısıtlamayı gevşeterek özellikler ve sınıflar arasındaki daha karmaşık ilişkileri yakalamasına olanak tanır. İkinci Dereceden Diskriminant Analizi, bu tahmin edilen parametrelerden yararlanarak, Bayes teoremini kullanarak her sınıfa ait bir gözlemin sonsal olasılığını hesaplar ve gözlemi en yüksek olasılığa sahip sınıfa atar [29].

3.2.10. AdaBoost (Adaptive Boosting)

AdaBoost, genellikle karar ağaçlarının performansını iyileştirmek ve sağlam bir tahmine dayalı model oluşturmak için tasarlanmış, topluluk öğrenme yöntemidir. AdaBoost'un arkasındaki temel konsept, bir dizi zayıf öğrenicinin aynı veri seti üzerinde sırayla eğitilmesini ve sonraki her öğrenicinin öncekilerin yanlış sınıflandırdığı örnekler daha fazla odaklanmasını içerir. Eğitim sırasında AdaBoost, yanlış sınıflandırılan örnekler daha yüksek ağırlıklar atar ve esas olarak sınıflandırılması zor örnekleri vurgular. Model yinelenirken, sonraki zayıf öğrenenler bu zorlu örnekler öncelik verecek şekilde ayarlanıyor ve bu da genel tahmin doğruluğunu toplu olarak iyileştiriyor. Her zayıf öğrenici, bireysel önemlerinin verileri sınıflandırmadaki performanslarına göre belirlendiği, ağırlıklı bir kombinasyon aracılığıyla nihai tahmine katkıda bulunur. Bu yinelenmeli süreç aracılığıyla AdaBoost, birden fazla zayıf öğrenicinin güçlü yönlerinden etkili bir şekilde yararlanarak güçlü bir sınıflandırıcı oluşturur [30].

3.2.11. Gradyan Artırma (Gradient Boosting - GB)

Gradyan Artırma, özellikle regresyon ve sınıflandırma problemlerinde tahmine dayalı modeller oluşturmak için kullanılan güçlü bir makine öğrenme tekniğidir. Güçlü bir tahmin modeli oluşturmak için birden fazla zayıf öğreniciyi (karar ağaçları) sırayla birleştirerek çalışır. Geleneksel güçlendirme algoritmalarından farklı olarak Gradyan Artırma, yinelenmeli bir süreç yoluyla önceden tanımlanmış bir kayıp fonksiyonunu en aza indirerek modeli optimize etmeye odaklanır. Başlangıçta algoritma, tahminlerde bulunmak için genellikle bir karar ağacı olan basit bir model oluşturur. Daha sonra bu tahminler ile gerçek değerler arasındaki artıkları veya hataları analiz eder. Sonraki modeller, önceki yinelenmelerdeki yanlış sınıflandırılmış veya kötü tahmin edilen örnekler vurgulayarak bu artıkları tahmin etmek için eğitilir. Her yeni zayıf öğreniciye, şu ana kadar oluşturulan birleştirilmiş model setinin yaptığı hataları azaltmakla görev verilir. Gradyan Artırma, birden fazla zayıf öğrenicinin tahminlerini eklemeli bir şekilde bir araya getirerek, veriler içindeki karmaşık kalıpları etkili bir şekilde yakalayan sağlam bir tahmin modeli oluşturur [31].

3.2.12. XGBoost (Extreme Gradient Boosting)

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

XGBoost, özellikle regresyon ve sınıflandırma problemlerinde olağanüstü performansı ile güçlü ve yaygın olarak kullanılan bir makine öğrenmesi algoritmasıdır. XGBoost, genellikle karar ağaçlarının tahminlerini birleştiren bir topluluk yöntemi olarak çalışır. Sıralı olarak çok sayıda ağaç oluşturur ve her ağaç bir öncekinin yaptığı hataları düzeltir. XGBoost, mevcut topluluğun eksikliklerini tamamlayan yeni ağaçlar ekleyerek belirli bir amaç fonksiyonunu en aza indirir. Gradyanları hesaplayarak ve modeli yinelemeli olarak güncelleyerek kayıp fonksiyonunu optimize etmeyi içeren bir gradyan artırma çerçevesi kullanır. XGBoost, aşırı uyumu kontrol etmek için düzenleme teknikleri, eksik verilerin karmaşık bir şekilde işlenmesi ve paralel işleme yeteneklerini güçlendirerek onu büyük ölçekli veri kümeleri için oldukça verimli hale getiren ölçeklenebilir ve verimli bir algoritma tasarımı sunar [32].

3.2.13. LightGBM (Light Gradient Boosting Machine)

LightGBM, Microsoft tarafından geliştirilen, büyük veri kümelerini yönetme ve gradyan artırma karar ağacı modelleri için hızlı ve etkili eğitim gerçekleştirme konusunda üstün olan bir makine öğrenmesi çerçevesidir. LightGBM, Gradyan Tabanlı Tek Taraflı Örnekleme (GOSS) ve Özel Özellik Paketleme (EFB) ile birleştirilmiş benzersiz bir teknik kullanır. GOSS, örnekleri gradyanlarına göre örnekleyerek, daha büyük gradyanlara sahip örnekleri vurgulayarak ve daha küçük gradyanlara sahip olanları atarak eğitim sürecini geliştirmeye odaklanır. Bu yaklaşım, doğruluktan ödün vermeden hesaplama maliyetini önemli ölçüde azaltır. EFB, ayrıcalıklı özellikleri birleştirmek, bellek kullanımını azaltmak ve eğitim sürecinin verimliliğini artırmak için kullanılır. LightGBM, geleneksel seviye bazlı yaklaşım yerine yaprak bazlı ağaç büyüme stratejisini takip ederek ağacın daha derinlik odaklı bir şekilde büyümesine olanak tanır ve bu da genellikle daha yüksek doğruluk sağlar. Ayrıca sürekli özellikler için en uygun bölünme noktalarına yaklaşmak amacıyla histogram tabanlı algoritmalar uygulayarak eğitim sürecini daha da hızlandırır [33].

3.2.14. CatBoost (Categorical Boosting)

CatBoost, öncelikle sınıflandırma ve regresyon problemleri için kullanılan, denetimli öğrenme görevleri için tasarlanmış bir makine öğrenmesi algoritmasıdır. Yandex tarafından geliştirilen CatBoost, kategorik özellikleri kapsamlı ön işleme gerek kalmadan verimli ve sağlam bir şekilde işleme yeteneğiyle öne çıkar. Kategorik değişkenlerin yarattığı zorlukları ele alan yenilikçi bir uygulamayla karar ağaçlarında gradyan artırmayı kullanır. CatBoost simetrik bir ağaç yapısı kullanır ve sıralı artırma olarak bilinen gradyan hesaplaması için yeni bir algoritma kullanır. Aşırı uyumu azaltan ve genellemeyi geliştiren, kategorik verileri işlemek için rastgele permütasyonların kullanılması gibi teknikleri içerir. CatBoost, ağaçlardaki eksik verilerle başa çıkmak için özel bir yöntem kullanarak eksik veri kümelerinin etkili bir şekilde işlenmesini sağlar. Optimizasyon mekanizmaları, aşırı uyumu önlemek için gelişmiş düzenleme tekniklerini içerir ve bu da onu yüksek tahmin doğruluğunu korurken karmaşık veri kümelerini işleme konusunda özellikle usta kılar. Genel olarak CatBoost'un temel güçlü yönleri, kategorik değişkenleri etkili bir şekilde yönetme, eksik verileri ele alma ve güçlü tahmin performansı sunma yeteneğinde yatmaktadır [34].

3.3. Değerlendirme metrikleri

Makine öğrenmesi yöntemlerinde herhangi bir sınıflandırıcının kalitesi, karışıklık matrisi için gerekli olan doğru pozitif tahminlerin sayısı 'TP', yanlış pozitif tahminlerin sayısı 'FP', doğru negatif tahminlerin sayısı 'TN' ve yanlış negatif tahminlerin sayısı 'FN' elde

edilerek çeşitli metriklerle ölçülür. Bu ölçütler, sınıflandırıcının ne kadar iyi performans gösterdiğine dair net bilgi vermektedir [16]. Çalışmada kullanılan yöntemlerin sınıflandırma performansı doğruluk (accuracy), kesinlik (precision), geri çağırma (recall), f1-skoru (f1-score), ortalama mutlak hata (MAE), ortalama karekök hata (RMSE) metrikleri ile ölçülmüştür [35].

3.3.1. Karışıklık Matrisi (Confusion Matrix)

Karışıklık matrisi, ikili sınıflandırmanın yorumlanmasında kullanılan ve performans ölçümlerinin güçlü yanları ve sınırlamaları nispeten kolay anlaşılır kılmaktadır [36]. Karışıklık matrisi, sınıflandırma modelinin tahminlerinin doğruluğunu özetleyen bir ($n \times n$) tablodur. n sınıf sayısını temsil etmek için kullanılmaktadır. Bir ikili sınıflandırma probleminde $n = 2$ 'dir. Diğer bir ifadeyle, gerçek sınıf etiketleri ile modelin tahmin ettiği sınıf etiketleri arasındaki korelasyonu ölçmeye yarayan bir tablodur [16].

Tablo 2. Karışıklık matrisi

	Tahmin Edilen Pozitif	Tahmin Edilen Negatif
Gerçek Pozitif	TP	FN
Gerçek Negatif	FP	TN

3.3.2. Doğruluk

Hata oranı ölçümünün tamamlayıcısı olan doğruluk ölçütü, doğal olarak test setindeki doğru sınıflandırılmış örneklerin, toplam örnek sayısına bölünmesiyle hesaplanmaktadır [36]. Doğruluk formül 1'de gösterilmiştir:

$$\text{Doğruluk} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

3.3.3. Kesinlik

Kesinlik ölçütü, sınıflandırma sonucunda pozitif olarak tahmin edilenlerin ne ölçüde doğru olarak tahmin edildiğini vermektedir [36]. Kesinlik formül 2'de gösterilmiştir:

$$\text{Kesinlik} = \frac{TP}{TP+FP} \quad (2)$$

3.3.4. Geri Çağırma

Geri çağırma ölçütü, sınıflandırıcının gerçekte pozitif sınıfa ait olan verileri doğru olarak tahmin etme oranını vermektedir [36]. Geri çağırma formül 3'de gösterilmiştir:

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

$$\text{Geri Çağırma} = \frac{TP}{TP+FN} \quad (3)$$

3.3.5. F1-Skoru

F1-skoru, kesinlik ve duyarlılık değerlerinin harmonik ortalaması alınarak hesaplanmaktadır. Dengeli F1-Skoru, sınıflandırıcının duyarlılık ve kesinliği eşit olarak ağırlıklandırması ile oluşturulmaktadır [36]. F1-Skoru formül 4’de gösterilmiştir:

$$\text{F1-Skoru} = \frac{2 (\text{Kesinlik} * \text{Duyarlılık})}{\text{Kesinlik} + \text{Duyarlılık}} \quad (4)$$

3.3.6. Ortalama Mutlak Hata (Mean Absolute Error - MAE)

Ortalama Mutlak Hata, tahmine dayalı bir modelin tahminlerinin veya tahminlerinin doğruluğunu değerlendirmek için kullanılan istatistiksel bir ölçümdür. Bir veri kümesindeki tahmin edilen değerler ile gerçek değerler arasındaki hataların ortalama büyüklüğünü ölçer. Matematiksel olarak MAE, her tahmin edilen değer ile ona karşılık gelen gerçek değer arasındaki mutlak farklar alınarak, bu farklar toplanarak ve ardından bunların tüm veri kümesi genelinde ortalaması alınarak hesaplanır. Orijinal verilerle aynı birimlerdeki hataların ortalama büyüklüğünü temsil ettiğinden MAE'nin yorumlanmasını kolaylaştırır. Düşük MAE değerleri, tahmin edilen ve gerçek değerler arasındaki daha küçük ortalama hataları yansıtan, daha iyi model performansını gösterir [35]. MAE formül 5’de gösterilmiştir:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |x_i - x| \quad (5)$$

3.3.7. Ortalama Karekök Hata (Root Mean Square Error - RMSE)

Ortalama Karekök Hata, bir veri kümesi içindeki tahmin edilen değerler ile bunlara karşılık gelen gerçek değerler arasındaki farkları değerlendirerek tahmine dayalı bir modelin tahminlerinin doğruluğunu değerlendirmek için kullanılan istatistiksel bir ölçüdür. RMSE, öncelikle tahmin edilen ve gerçek değerler arasındaki farkların karelerinin hesaplanması, bu farkların karelerinin toplanması, bu toplamın ortalamasının alınması ve son olarak bu ortalamanın karekökünün elde edilmesi yoluyla hesaplanır. RMSE’de hataların karesinin alınmasının iki temel etkisi vardır; tüm hataların pozitif olmasını sağlar ve daha büyük hataları küçük olanlardan daha önemli ölçüde cezalandırır, böylece tahmin edilen ve gerçek değerler arasındaki daha büyük sapmalara karşı duyarlı hale gelir. Daha düşük RMSE değerleri daha iyi model performansını gösterir; bu da tahmin edilen ve gerçek değerler arasındaki ortalama kare farklarının daha küçük olduğunu gösterir [35]. RMSE formül 6’da gösterilmiştir:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

4. Bulgular

4.1. Veri analizi

Çalışmada, makine öğrenmesi yöntemi kullanılarak oluşturulan modellerden önce verilere ön analiz yapılmıştır. Analizde öne çıkan temel bulgularda “Başarısızlık/İflas” ile “Problemlili Kredilerin Sayısı” arasında, “Vergi Alacağı/Haczi” ile “Problemlili Kredilerin Sayısı” arasında doğru orantılı çok yüksek ilişkiye rastlanılmıştır. Ayrıca “Aylık Borç” ile “Yıllık Kazanç”, “Açık Hesap Sayısı”, “Mevcut Kredi Bakiyesi” arasında yüksek, “Yıllık Kazanç” ile “Mevcut Kredi Bakiyesi” arasında doğru orantılı ilişkiye rastlanılmıştır. Diğer yandan, “Kredi Durumu” ile “Kredi Skoru” arasında ters orantılı ilişki görülmüştür. Değişkenler arasındaki korelasyon ilişkisine dair daha detaylı bilgi için Resim 3'te gösterilen verilere bakılabilir.



Şekil 2. Özniteliklerin ilişki dereceleri

4.2. Modellerin metrik değerleri

Çalışmanın bu aşamasında tekniklerin oluşturduğu modellerin performansının ölçülmesinde doğruluk, kesinlik, geri çağırma, f1-skoru metrikleri ile hata değerlerinin ölçülmesinde ortalama mutlak hata ve ortalama karekök hata metrikleri kullanılmıştır. Modeller oluşturulurken birkaç kez deneme yapılmış, en başarılı metriklerin elde edilmesinde veri setinin %80'i eğitim ve %20'si test seti (5 kat çapraz doğrulama) ayrımı ile elde edildiği görülmüştür. Modellerin oluşturduğu doğruluk metriği baz alınarak yapılan genel değerlendirmede 12 model başarılı performans gösterirken GNB ile QDA iyi bir performans gösterememiştir. CatBoost doğruluk seviyesinin diğer yöntemlerden daha başarılı olmuştur ama çalışma süresi olarak diğer yöntemlere göre daha yavaştır. Çalışmanın en başarılı performansına sahip modelleri sırasıyla CatBoost, XGBoost ve LightGBM yöntemleri hem doğruluk hem de hata kriterlerine

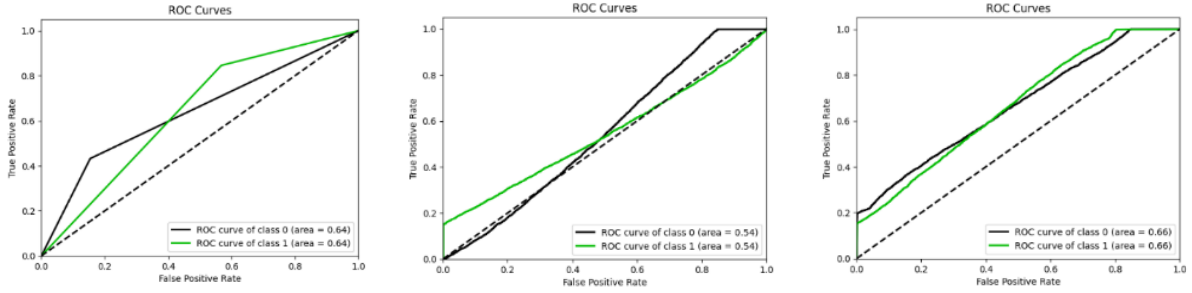
Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

göre daha başarılıdır. Modellerin çalışma süreleri Google Colab'da geliştirilen uygulama ile ölçülmüştür. GNB ve QDA çalışma süresi açısından başarılı olmuş ama doğruluk açısından başarılı olamayınca, sürenin en kısa olması bir anlam ifade etmiyor.

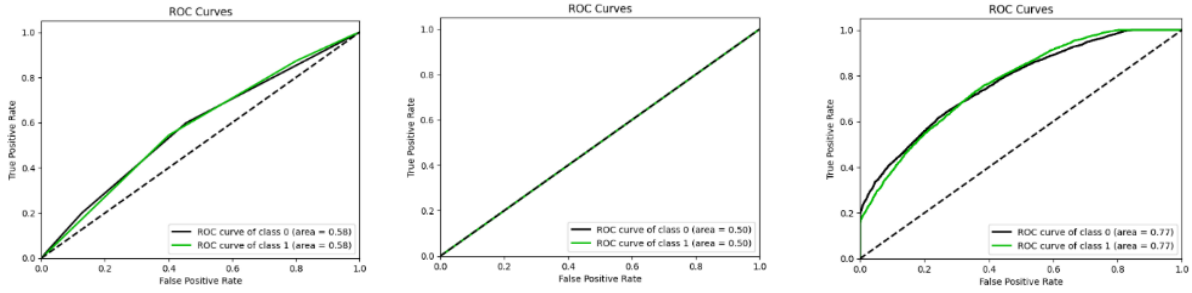
Tablo 3. Modellere ait metrik değerleri

Modeller	Doğruluk	Kesinlik	Geri Çağırma	F1-Skor	Ortalama Mutlak Hata	Ortalama Karekök Hatası	Çalışma Süresi
DT	0.75169	0.8348	0.8458	0.8403	0.2483	0.4983	33.77
SVM	0.77238	0.7723	1.0	0.8715	0.2276	0.4770	184.90
GNB	0.34625	0.9941	0.1545	0.2674	0.6537	0.8085	2.09
KNN	0.71987	0.7872	0.8733	0.8280	0.2801	0.5292	19.13
ANN	0.77238	0.7723	1.0	0.8715	0.2276	0.4770	160.71
XGBoost	0.81847	0.8159	0.9878	0.8936	0.1815	0.4260	394.72
LightGBM	0.81802	0.8107	0.9970	0.8943	0.1819	0.4265	32.10
CatBoost	0.81958	0.8147	0.9920	0.8946	0.1804	0.4247	819.92
SGD	0.79227	0.7917	0.9919	0.8806	0.2077	0.4557	6.72
AdaBoost	0.81672	0.8091	0.9981	0.8937	0.1832	0.4281	118.69
RF	0.81702	0.8084	1.0	0.8940	0.1829	0.4277	90.66
LR	0.81301	0.8080	0.9939	0.8914	0.1869	0.4324	21.43
GB	0.81737	0.8096	0.9982	0.8941	0.1826	0.4273	192.17
QDA	0.35728	0.9671	0.1737	0.2946	0.6427	0.8016	2.18

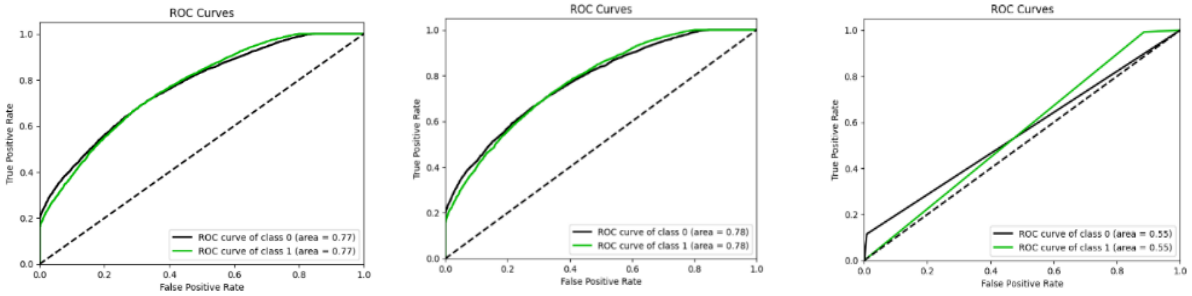
ROC eğrisi, sınıflandırma modellerinin performansını görsel olarak değerlendirmek için faydalı bir araçtır. Ancak, tek başına AUC veya eğri şekli üzerinden yorum yapmak yerine, geniş bir perspektiften değerlendirme yapmak daha önemlidir. ROC eğrisi, sol üst köşeden geçen bir çizgiyi temsil eden rastgele sınıflandırıcıdan ne kadar uzak olursa, modelin performansının o kadar iyi olduğunu gösterir. ROC eğrilerine baktığımızda CatBoost diğer modellerden daha başarılıdır. İlave olarak XGBoost, LightGBM, AdaBoost ve Gradient Boosting'de model performansı oldukça başarılıdır.



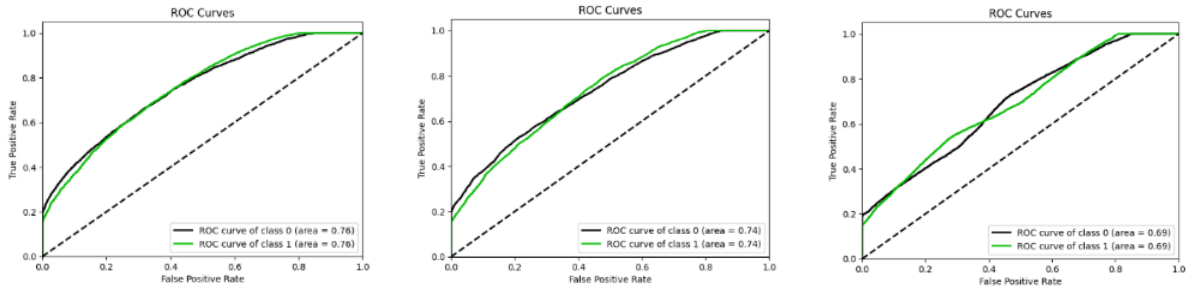
Şekil 3. ROC eğrileri DT - SVM - GNB



Şekil 4. ROC eğrileri KNN - ANN - XGBoost

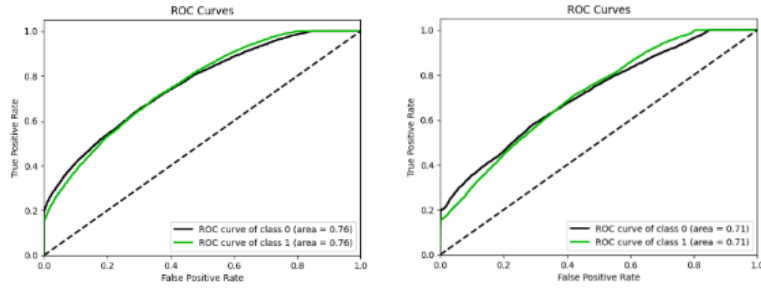


Şekil 5. ROC eğrileri LightGBM - CatBoost - SGD



Şekil 6. ROC eğrileri AdaBoost - RF - LR

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü



Şekil 7. ROC eğrileri GB – QDA

ROC eğrisinin altında kalan alan (AUC - Area Under the Curve), sınıflandırma modelinin performansını ölçen bir metriktir. AUC değeri, 0 ile 1 arasında bir değer alır. AUC'nin değeri 1'e ne kadar yakınsa, modelin performansı o kadar iyidir. Modellerin AUC değerlerine baktığımızda en yüksek performans 0,78 değeri ile CatBoost modelinde Şekil 5. gösterilmiştir. AUC değerlerine göre CatBoost modelini 0,77 değeri ile XGBoost, LightGBM modelleri takip etmiştir. ROC eğrisi gerçek pozitif oranı ve yanlış pozitif oranı grafiğini çizer. Bu çalışmada kredi verilmesi gerekenler ve kredi verilmemesi gerekenleri yüksek oranda doğru tahmin etmiştir. AUC performans değerlerine göre CatBoost algoritması diğer algoritmalara göre daha başarılı olduğu Tablo 3. de gösterilmiştir.

Tablo 4. Modellere ait AUC değerleri

Modeller	AUC
DT	0.64
SVM	0.54
GNB	0.66
KNN	0.58
ANN	0.50
XGBoost	0.77
LightGBM	0.77
CatBoost	0.78
SGD	0.55
AdaBoost	0.76
RF	0.74
LR	0.69
GB	0.76
QDA	0.71

5. Tartışma

Bu çalışmada, veri ön işleme aşamasında 206 adet kayıt silinmiştir. Bu oran veri kaybı oranında oldukça azdır. Bu durum sınıflandırma algoritmalarının başarı ve performansını artırmıştır. Veri setindeki sınıfların dağılımı dengesizse, yani bir sınıf diğerlerine göre çok daha fazla temsil ediliyorsa, doğruluk oranı yanıltıcı olabilir. Bu çalışmadaki veri seti çok dengesiz veya aşırı dengesiz değildir. Bir diğeri bu dengesizliği ortadan kaldırmak için çapraz doğrulama ölçümleri yapılmıştır. Bu çalışmada azınlık sınıfları ihmal edilmemiştir. Çapraz doğrulama olarak 5 kat uygulanmıştır. 10 kat çapraz doğrulamada işlem süreci uzadığı için tercih edilmedi. Isı haritasına göre “Başarısızlık/İflas” ile “Problemlili Kredilerin Sayısı” arasında ilişki sayesinde gelecekteki çalışmalara bile ışık tutabilir. Özniteliklerin azaltılması veya indirgenmesi noktasında deneyler yapılmamıştır, zaten öznitelik sayısı oldukça azdır. Değerlendirme metriklerine göre, hata değerlerine göre ve ROC eğrilerine göre CatBoost algoritması oldukça başarılı olmuştur. Son teknoloji algoritmalarından olan; CatBoost, XGBoost ve LightGBM bu çalışmadaki performansları diğeri algoritmalara oldukça fark atmıştır.

6. Sonuç

Bu çalışmada, banka kredisi tahmini alanında çeşitli makine öğrenmesi algoritmalarının performansları değerlendirilmiş ve bu algoritmaların finansal risk öngörüsü ve kredi değerlendirmesi gibi kritik uygulamalardaki etkinliği incelenmiştir. Yapılan deneylerde algoritmaların tahmin doğruluğu, kesinliği ve hata oranları gibi performans metriklerine göre CatBoost, XGBoost ve LightGBM algoritmaları diğerlerine göre daha başarılı olduğu gözlemlenmiştir. Algoritmaların başarı oranının artmasıyla beraber çalışma süreleri diğeri algoritmalara göre uzun sürmüştür. Buradaki öncelikli kriter doğruluk ve hata kriterleridir. Bu çalışma ile finansal kurum kredisi veya banka kredisi tahmini gibi finansal uygulamalarda daha etkili kararlar alınmasına yardımcı olmuş ve uzmanlara rehberlik sağlar. Sonuçlar, bu karşılaştırmalı deneylerin finansal alanda makine öğrenmesi algoritmalarının performanslarını anlamak ve değerlendirmek için önemli bir araç olduğunu ortaya koymuştur. Özellikle, Topluluk Öğrenmesi altında bulunan algoritmaların banka kredisi tahmini gibi kritik finansal konularda daha yüksek başarı elde ettiği gözlemlenmiştir. Gelecek çalışmalarda kişiye özel kredi önerisi problemi araştırılarak, son teknoloji transformer (dönüştürücüler) ile ilgili deneyler yapılması planlanıyor.

Araştırma ve yayın etiği beyanı: Bu çalışmada bilimsel araştırma ve yayın etiğine uygun davranılmış ve yazarlar tarafından kabul edilmiştir.

Etik kurul onayı: Bu çalışmada etik kurul onayına gerek yoktur.

Yazarların katkısı: Bu çalışmaya yazarlar eşit oranda katkı sağlamıştır.

Çıkar çatışması beyanı: Herhangi bir kurum ya da kişiler ile çıkar çatışması bulunmamaktadır.

Kaynakça

- [1] Baesens, B., Van Gestel, B.T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J. (2003). Benchmarking State-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54, 627-635.
- [2] Crone, S. F., Finlay, S. (2012). Instance sampling in credit scoring: an empirical study of sample size and balancing. International Journal of Forecasting, Elsevier, 28 (1), 224-238.

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

- [3] Serengil, S.I., Imece, S., Tosun, U.G., Buyukbas, E.B., Koroglu, B. (2022). A comparative study of machine learning approaches for non performing loan prediction with explainability. *International Journal of Machine Learning and Computing*, 12 (5), 208-214.
- [4] Provenzano, A.R., Trifirò, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., Pera, G.L., Spadaccino, M., Massaron, L., Nordio, C. (2020). Machine learning approach for credit scoring. *arXiv: Statistical Finance*.
- [5] Altan, G., Demirci, S. (2022). Makine öğrenmesi ile nakit akış tablosu üzerinden kredi skorlaması: Xgboost yaklaşımı. *Journal of Economic Policy Researches*, 9 (2), 397-424.
- [6] Alan, A., Karabatak, M. (2020). Veri seti - sınıflandırma ilişkisinde performansa etki eden faktörlerin değerlendirilmesi. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 32 (2), 531-540.
- [7] Xia, Y., Liu, C., Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research Applications* 24, 30-49.
- [8] Nisbet, R., Miner, G., Yale, K. (2018). *Handbook of statistical analysis and data mining applications*. Second edition. Academic Press – Elsevier, London.
- [9] Gür, Ö., Tarhan Mengi, B. (2022). Hile tespitinde makine öğrenmesi yöntemlerinin kullanılması ve model performanslarının değerlendirilmesi. *İşletme Araştırmaları Dergisi*, 14 (4), 3053-3065.
- [10] Moscato, V., Picariello, A., Sperli, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Application*, 165, 113986.
- [11] Meyer, P.A., Pifer, H.W. (1970). Prediction of bank failures. *Journal of Finance*, 25(4), 853-868.
- [12] Li, Y., Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, 8 (10), 1756.
- [13] Sönmez, F. (2015). Kredi skorunun belirlenmesinde yapay sinir ağları ve karar ağaçlarının kullanımı bir model önerisi. *Anadolu Bil Meslek Yüksekokulu Dergisi*, 37, 1-22.
- [14] Bahnsen, A.C., Aouada, D., Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit scoring. *13th International Conference on Machine Learning and Applications*, Detroit, MI, USA, 263-269.
- [15] Kim, J., Choi, K., Kim, G., Suh, Y. (2012). Classification cost: An empirical comparison among traditional classifier, cost-sensitive classifier, and metacost. *Expert Systems with Application*, 39 (4), 4013-4019.
- [16] Bhatia, P. (2019). *Data mining and data warehousing principles and practical techniques*, First Published. Cambridge University Press, United Kingdom.
- [17] Bank Loan Status Dataset - Future Loan Status Prediction via Classification Models. (2018). <https://www.kaggle.com/datasets/zaurbegiev/my-dataset> (11.11.2023).
- [18] Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining*. John Wiley & Sons, New Jersey.

- [19] Akpınar, H. (2017). Veri madenciliği veri analizi. 2. Baskı. Papatya Yayıncılık Eğitim, İstanbul.
- [20] Han, J.W., Kamber, M., Pei, J. (2012). Data mining concepts and techniques. Third Edition, Morgan Kaufmann, Waltham.
- [21] Kingsford, C., Salzberg, S. L. (2008). What are decision trees?. *Nature Biotechnology*, 26 (9), 1011-1013.
- [22] Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology*, 24 (12), 1565-1567.
- [23] Bustamante, C., Garrido, L., Soto, R. (2006). Comparing fuzzy naive bayes and gaussian naive bayes for decision making in robocup 3d. In *Mexican International Conference on Artificial Intelligence* (pp. 237-247). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [24] Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4 (2), 1883.
- [25] Zou, J., Han, Y., So, S. S. (2009). Overview of artificial neural networks. *Artificial neural networks: methods and applications*, 14-22.
- [26] Qi, Y. (2012). Random forest for bioinformatics. *Ensemble machine learning: Methods and applications*, 307-323.
- [27] LaValley, M. P. (2008). Logistic regression. *Circulation*, 117 (18), 2395-2399.
- [28] Bottou, L. (2012). Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition* (pp. 421-436). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [29] Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, 3 (2), 145-180.
- [30] Margineantu, D. D., Dietterich, T. G. (1997). Pruning adaptive boosting. In *ICML*, Vol. 97, pp. 211-218.
- [31] Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurobotics*, 7, 21.
- [32] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.
- [33] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
- [34] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.
- [35] Singh, P. K., Pramanik, P. K. D., Debnath, N. C., Choudhury, P. (2019). A novel neighborhood calculation method by assessing users' varying preferences in collaborative filtering. In *CATA* (pp. 345-355).
- [36] Japkowicz N. (2011). Performance evaluation for learning algorithms. Cambridge University Press, Cambridge.

Banka kredisi tahmini için makine öğrenmesi algoritmalarının performans analizi: Topluluk öğrenmesi algoritmalarının üstünlüğü

- [37] Beaver, W.H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4, 71-111.
- [38] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23 (4), 589-609.
- [39] Ampountolas, A., Nde, T. N., Date, P. Constantinescu, C. (2021). A machine learning approach for micro-credit scoring. *Risks*, MDPI AG, 9 (3), 50:1-20.