MAKİNE ÖĞRENİMİ YÖNTEMLERİNİ KULLANARAK KREDİ ONAY TAHMİNİ

Canan UZMA
Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA

Makale Bilgisi

Düzeltme: 21/04/2024
Teslim: 21/05/2024

Anahtar Kelimeler

Kredi onayı, Makine Öğrenimi, Veri Analizi, finansal kararlar, veri ön işleme, öznitelik seçme, çapraz doğrulama, XGBoost, Random Forest, Decision Tree

Öz

Kredi onay süreci finansal hizmetler sektöründe kritik bir rol oynar ancak doğru bir şekilde değerlendirilmesi zor olabilir. Bu çalışmanın amacı, makine öğrenimi modellerini gözden geçirmek ve kredi onay veri kümelerini analiz etmek için en etkili yöntemleri tartışmaktır. Veri seti, bir kişinin veya kuruluşun kredi itibarını belirlemek için çeşitli finansal özellikler içerir. Eksik verileri temizlemek ve önemli özellikleri belirlemek için veri ön işleme ve özellik seçme aşamaları uygulandı. Farklı makine öğrenimi algoritmalarının performansı, 10 kat çapraz doğrulama yöntemi kullanılarak değerlendirildi.

Sonuçlar XGBoost, Random Forest ve Decision Trees gibi algoritmaların yüksek doğruluğa sahip olduğunu göstermektedir. Üstelik F1 puanı gibi diğer metrikler de XGBoost'un en iyi performansa sahip olduğunu gösteriyor.

Bu araştırma, finansal hizmetler sektöründe daha doğru ve verimli karar almayı mümkün kılmak için yenilikçi makine öğrenimi yaklaşımlarından yararlanmayı amaçlamaktadır.

CREDIT APPROVAL PREDICTION USING MACHINE LEARNING METHODS

Keywords

Credit approval, Machine Learning, Data Analysis, Financial Decisions, Data Preprocessing, Feature Selection, Cross-validation, XGBoost, Random Forest, Decision Tree

Abstract

The loan approval process plays a critical role in the financial services industry but can be difficult to accurately assess. The purpose of this paper is to review machine learning models and discuss the most effective methods for analyzing credit approval datasets. The dataset contains various financial features to determine the creditworthiness of a person or organization. Data preprocessing and feature selection stages were applied to clean missing data and identify important features. The performance of different machine learning algorithms was evaluated using a 10-fold cross-validation method.

The results show that algorithms such as XGBoost, Random Forest and Decision Trees have high accuracy. Moreover, other metrics such as F1 score also show that XGBoost has the best performance.

This research aims to leverage innovative machine learning approaches to enable more accurate and efficient decision making in the financial services industry.

1. GİRİŞ (INTRODUCTION)

Günümüzün finansal ekosistemi, bireylerin ve kuruluşların kredi alımına erişimini kolaylaştıran bir dizi faktöre bağlıdır. Ancak, kredi kurumlarının bu kredilendirme sürecinde karşılaştığı temel zorluklardan biri, potansiyel borçluların krediye uygunluğunu doğru bir şekilde değerlendirmektir. İşte bu noktada, kredi onayı veri setleri, finansal karar alma süreçlerini desteklemek için hayati önem taşıyan bir rol oynamaktadır. Kredi onay kaydı, genellikle bir kişinin veya kuruluşun bir kredi kuruluşundan kredi almaya uygun olup olmadığını belirlemek için kullanılan mali kayıtların ve ilgili bilgilerin bir koleksiyonudur. Bu veri kümeleri çeşitli öğeler içerir.Bu faktörler, bir kişinin veya kuruluşun kredi geçmişi, geliri, çalışma durumu, net değeri ve mevcut kredi durumu gibi bilgileri içerir. Bu zengin veri kaynakları, kredi kuruluşlarının, kredi başvurusu değerlendirme sürecini optimize etmek için makine öğrenimi ve veri analitiği tekniklerini kullanmasına olanak tanır.

Bu makalenin amacı finansal karar alma sürecindeki en etkili makine öğrenimi modellerini gözden geçirmek ve kredi onay veri setlerini analiz etmek için en uygun yöntemleri tartışmaktır. Bu kapsamda, kredi başvurularının değerlendirilmesine yönelik mevcut model ve algoritmaların güçlü ve zayıf yönleri tartışılarak, potansiyel olarak daha etkin ve duyarlı modellerin geliştirilmesine yönelik öneriler sunulmaktadır.

Son olarak, makine öğrenimi ve veri analizi tekniklerinin sürekli gelişimi göz önüne alındığında, finansal hizmetler sektöründe kredi onayı tahmin modellerinin gelecekteki yönü ve eğilimleri değerlendirilmektedir. Bu değerlendirme, kredi kuruluşlarının daha doğru ve verimli kararlar almalarına yardımcı olacak yenilikçi yaklaşımları keşfetmelerine yardımcı olur.

2. YÖNTEMLER (METHODS)

2.1. Veri Seti (Data Set)

Bu veri seti, bireylerin veya kuruluşların bir kredi kurumundan kredi almaya uygunluğunu belirlemek için kullanılan mali kayıtların ve ilgili bilgilerin bir koleksiyonudur. Kredi puanı, gelir, çalışma durumu, kredi vadesi, kredi tutarı, varlık değeri, kredi durumu gibi çeşitli faktörleri içerir. Bu veri kümesi, verilen özelliklere göre kredi onayı olasılığını tahmin eden modeller ve algoritmalar geliştirmek için makine öğrenimi ve veri analizinde yaygın olarak kullanılır.
Veri seti 13 öznitelik ve 4269 örnekten oluşmaktadır. Veri setine ait öznitelikler Tablo 1'de görülmektedir.

Tablo 1-Veri Öznitelik Açıklamaları

Öznitelik	Açıklama
Loan id	Kredi İd
No of dependents	Başvuru sahibinin bakmakla yükümlü olduğu kişi sayısı
Education	Başvuru Sahibinin Eğitimi
Self employed	Başvuru Sahibinin Çalışma Durumu
Income annum	Başvuru Sahibinin Yıllık Geliri
Loan amount	Kredi miktarı
Loan term	Yıllar Olarak Kredi Vadesi
Cibil score	Kredi notu
Residential assets value	Konut varlıkları değeri
Commercial assets value	Ticari varlık değeri
Luxury assets value	Lüx varlık değeri
Bank asset value	Banka varlığı değerleri
Loan status	Kredi onay durumu

2.2. Öznitelik Seçme (Feature Selection)

Veri ön işleme aşamaları, makine öğrenimi modelinin doğruluğunu artırmak için kritik bir rol oynar. Bu aşamalar, genellikle eksik veya aykırı verilerin temizlenmesi, sayısal özelliklerin normalleştirilmesi ve kategorik özelliklerin dönüştürülmesi gibi adımları içerir. Eksik veriler, genellikle ortalama veya medyan değerlerle doldurulur veya eksik olan örnekler veri setinden çıkarılır. Aykırı veriler ise genellikle istatistiksel teknikler veya domain bilgisi kullanılarak ele alınır veya düzeltilir. Sayısal özelliklerin normalleştirilmesi, farklı ölçeklerdeki özellikler arasındaki dengeyi sağlar ve modelin daha iyi performans göstermesini sağlar. Kategorik özellikler ise genellikle one-hot encoding veya label encoding gibi teknikler kullanılarak sayısal değerlere dönüştürülür. Veri setimin ön işlemesinde, label encoding tekniğiyle kategorik veriler sayısal verilere dönüştürüldü.

Modelin karmaşıklığını azaltarak, veri setindeki en önemli özellikleri seçmeyi amaçlar. Bu adım, gereksiz veya yetersiz bilgi içeren özellikleri çıkartarak modelin daha genelleştirilmiş ve daha hızlı çalışmasını sağlar. Örneğin, eğitim seviyesi, kendi işinde çalışma durumu gibi özellikler, özellik seçimi aşamasında önemlidir.

Bu bağlamda, veri setimdeki "education", "self employed" ve "loan status" gibi özelliklerin int değerlere dönüştürülmesi ve sonrasında "loan status" özelliğinin tahmin edileceği modelden çıkarılması, feature selection aşamasının bir örneğidir. Bu adım, modelin daha kesin ve etkili bir şekilde çalışmasını sağlar ve gereksiz karmaşıklığı ortadan kaldırarak performansı arttırır.

2. Sınıflandırma Yöntemleri (Regression andClassification Methods)

2.3.1. KNN

K-En Yakın Komşular (kNN), belirli sınıflara sahip bir veri kümesine yeni bir veri eklendiğinde, bu yeni verinin hangi sınıfa ait olduğunu belirlemek için kullanılan bir yöntemdir. Yeni eklenen verinin sınıfını tespit etmek için, bu verinin sınıfları bilinen (eğitim seti) tüm verilere olan uzaklığı hesaplanır. Ardından, bu uzaklıklar yakından uzağa doğru sıralanır ve komşuluklar belirlenir. Parametre olarak belirlenen k değerine göre, k tane en yakın komşunun sınıfları incelenir ve en çok gözlemlenen sınıf, eklenen verinin sınıfı olarak belirlenir. Bu sayede, yeni verinin sınıfı diğer verilere göre benzerliklerine dayanarak belirlenmiş olur.

2.3.2 Naive Bayes

Naive Bayes, özellikler arasında bağımsızlık varsayımı yaparak en yüksek olasılığa sahip sınıfı tahmin eden bir sınıflandırma algoritmasıdır. Bu algoritma, tahmine dayalı modelleme için basit ancak güçlü bir yöntemdir. Özellikle sinyal ve görüntü işleme gibi alanlarda sıkça kullanılan bir sınıflandırma ve tahmin algoritmasıdır. Naive Bayes'te, bir sınıftaki belirli özelliklerin diğer özelliklerle ilişkisi olmadığı varsayılır. Bu sayede, verilerin sınıflandırılması ve tahmin edilmesi süreçlerinde etkili sonuçlar elde edilir.

2.3.3. LSVM

En Küçük Kareler Vektör Makineleri (LSVM), verileri optimal bir şekilde iki kategoriye ayırmak için n-boyutlu bir hiperdüzlem oluşturan bir destek vektör makinesi (DVM) türüdür. DVM modelleri, yapay sinir ağlarıyla yakından ilişkilidir. Sigmoid kernel fonksiyonu kullanan bir DVM ise, iki katmanlı, ileri beslemeli bir yapay sinir ağına benzeyen bir yapıya sahiptir.

DVM'nin dikkate değer özelliklerinden biri, veri seti üzerinde ortalama hata karesini minimize eden ampirik risk minimizasyonu prensibi yerine, istatistiksel öğrenme teorisindeki yapısal risk minimizasyonu prensibini benimsemesidir. Bu prensip, modelin genelleme yeteneğini artırmak için modelin karmaşıklığını kontrol eder.

DVM'nin temel varsayımlarından biri, eğitim kümesindeki tüm örneklerin bağımsız ve benzer şekilde dağılmış olmasıdır. Bu varsayım, modelin doğruluğunu ve güvenilirliğini artırmaya yöneliktir.

2.3.4. Random Forest

Rastgele Orman (Random Forest), verinin ve özelliklerin rastgele alt kümeleri üzerinde eğitilen bir dizi karar ağacından oluşan bir topluluk oluşturur. RF, yaygın ve güçlü bir ensemble denetimli sınıflandırma yöntemidir. Üstün doğruluğu ve dayanıklılığı nedeniyle, biyoinformatik ve tıbbi görüntüleme gibi çeşitli makine öğrenimi uygulamalarında etkili bir şekilde kullanılmaktadır.

RF, bir "orman" oluşturan bir dizi karar ağacından meydana gelir. Her karar ağacı, bagging algoritması kullanılarak oluşturulur. RF'yi eğitmek için, belirli parametreler ve eğitim veritabanı gereklidir. RF, hatırlama (recall), doğruluk (precision) ve f-puanını (f-score) hesaplamak için kullanılan bir kesme noktasını ayarlamaya da imkan tanır.

2.3.5. MLP

Çok Katmanlı Algılayıcı (MLP), üç farklı katmandan oluşur: girdi katmanı, gizli katman ve çıktı katmanı. Girdi katmanı, verilerin alındığı katmandır ve her nöron bir özelliği temsil eder, bu nedenle özellik sayısı kadar nöron içerir. Çıktı katmanı, sınıfların belirlendiği katmandır ve modele bağlı olarak tek bir nöron olabileceği gibi, sınıf çeşitliliği kadar da olabilir. Gizli katman, girdi ve çıktı katmanları arasında yer alır ve verilerin ara işlemlerden geçtiği katmandır. Gizli katmanların sayısı ve nöron sayısı, eğitimin kalitesini önemli ölçüde etkileyen faktörlerdir, ancak kesin değildir.

Kullanılan eğitim algoritması, ağırlıkları güncelleyerek hatanın karesini en aza indirmeyi amaçlar. Bu

süreçte, gerçek çıktı ile tahmin edilen çıktı arasındaki farkı minimize etmek için ağırlıkların iteratif olarak ayarlanması gerçekleştirilir. Bu şekilde, modelin doğruluğu ve performansı artırılmaya çalışılır.

2.3.6. XGBoost

XGBoost, yüksek performansı, esnekliği ve hızlı çalışmasıyla öne çıkar. Temelde, bir dizi zayıf öğreniciyi (genellikle karar ağaçları) ardışık olarak ekleyerek hataları minimize etmeye odaklanır. Her yeni ağaç, önceki ağaçların hatalarını düzeltmeye yönelik olarak eğitilir. Bu iteratif süreç, hataların aşamalı olarak azaltılmasını sağlayarak modelin tahmin performansını artırır.

XGBoost, özellikle büyük veri kümeleri ve karmaşık modeller için uygundur ve çeşitli hiperparametre optimizasyon teknikleriyle daha da güçlendirilebilir. Ayrıca, paralel hesaplama ve düzenlilik gibi özellikleri sayesinde hız ve doğruluk açısından üstün performans sergiler.
bırakma riskini tahmin etmede başarılı olduğunu göstermiştir.

a. 10 Kat Çarpraz Doğrulama (10-Fold Cross-Validation)

Bir modelin genelleme performansının doğru bir şekilde değerlendirilmesi makine öğrenmesi algoritmaları açısından önemlidir. Bunu yapmanın en etkili yollarından biri 10 kat çapraz doğrulamadır. Bu yöntemde veri seti eşit büyüklükte 10 alt gruba bölünür, her bir alt grup doğrulama seti olarak kullanılır ve geri kalan 9 grup da eğitim seti olarak kullanılır. Bu işlem 10 kez tekrarlanır ve her seferinde doğrulama kümesi olarak farklı bir alt küme seçilir. Sonuç olarak model 10 farklı veri seti üzerinde eğitilmiş ve test edilmiştir.
Bu çalışmada her grup için doğruluk oranı 10 kat çapraz doğrulama kullanılarak hesaplandı. Ayrıca algoritmanın F1 puanı, kesinlik ve geri çağırma değerleri 10 kat çapraz doğrulama kullanılarak hesaplandı.

Tablo 2-k-fold cross validation

3. BULGULAR VE TARTIŞMA (FINDINGS AND DISCUSSION)

3.1 Performans Metrikleri (Performance Metrics)

Doğruluk (Accuracy), doğru olarak sınıflandırılan örneklerin yüzdesini ifade eder. Yani, tüm örnekler içinde doğru tahmin edilenlerin oranını belirtir.

Duyarlılık (Recall), pozitif olarak sınıflandırılması gereken örneklerin ne kadarının doğru bir şekilde pozitif olarak tahmin edildiğini gösterir. Başka bir deyişle, gerçek pozitiflerin ne kadarının doğru bir şekilde tespit edildiğini ifade eder.

Kesinlik (Precision), pozitif olarak tahmin edilen değerlerin gerçekte kaçının pozitif olduğunu gösterir. Yani, pozitif olarak tahmin edilenler içinde gerçekten pozitif olanların oranını ifade eder.

$$Doğruluk = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Kesinlik = \frac{TP}{TP+FP}$$

$$Duyarlılık = \frac{TP}{TP+FN}$$

Tablo 3- Doğruluk,Kesinlik,Duyarlılık Formülleri

F1 Skoru (F1 Score), bir testin doğruluğunu ölçen bir metriktir ve kesinlik ile duyarlılığın harmonik ortalamasını temsil eder. Bu, bir modelin kesinlik ve duyarlılığının dengesini sağlar. F1 Skoru, mükemmel kesinlik ve duyarlılık durumunda 1'e, en kötü durumda ise 0'a yaklaşır.

$$F1 - Skor = \frac{2 \times Duyarlılık \times Kesinlik}{Duyarlılık + Kesinlik}$$

Tablo 4-F1 Score Hesaplanması

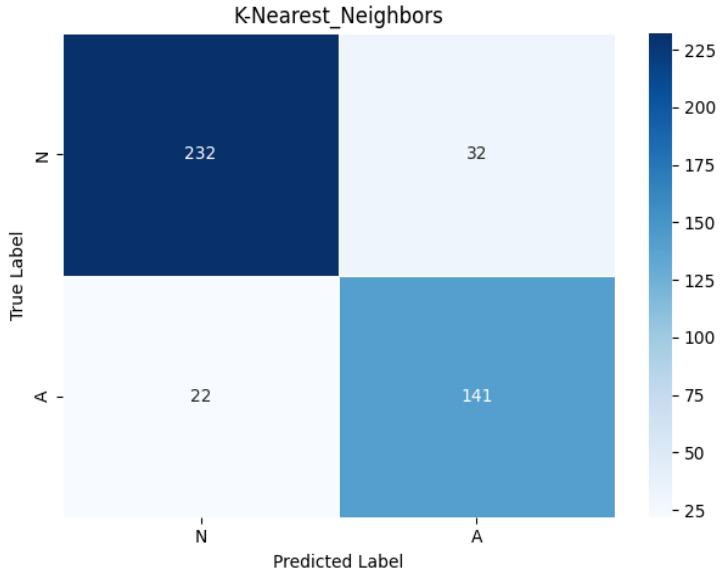
AUC-ROC Eğrisi (AUC-ROC Curve), sınıflandırma problemlerinde performansı değerlendirmek için kullanılan bir ölçümdür. ROC eğrisi, çeşitli eşik değerlerinde duyarlılık ve özgünlük arasındaki ilişkiyi gösterir. AUC ise ROC eğrisinin altında kalan alanı ifade eder ve modelin sınıflar arasındaki ayrım gücünü temsil eder. Yani, AUC ne kadar yüksekse, modelin sınıfları ayırt etme yeteneği o kadar iyidir.

3.2. Deneysel Sonuçlar (Experimental Results)

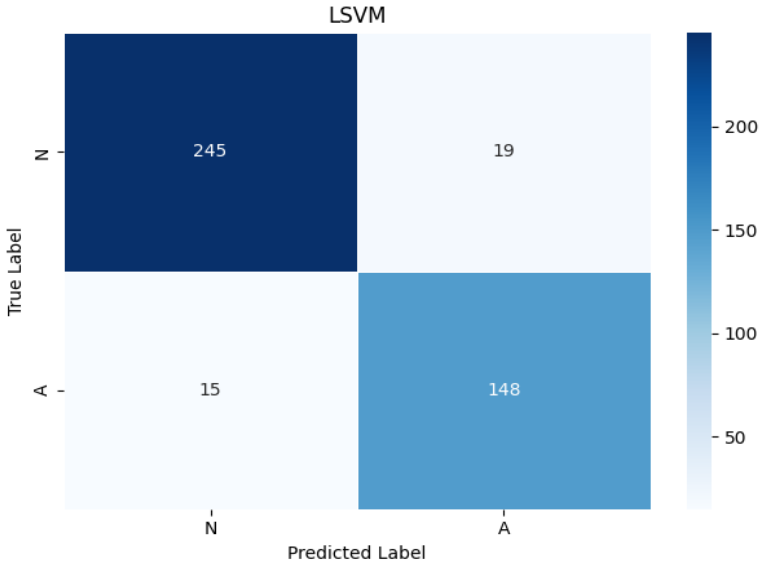
13 öz niteliğe sahip veri setiyle yapılan çalışma sonucunda KNN %90 , Naïve Bayes %94 LSVM %93, RBF SVM %94,MLP %94, Random Forest %98 Decision Tree %98 ve en yüksek doğruluk oranıyla XGBoost %98 doğruluk oranına sahiptir. Doğru ve yanlış sınıfları ayırt etmedeki başarıyı gösterir diğer metriklerin dengesini ifade eden F1 Score incelendiğinde de en yüksek oranı XGBoost algoritması sahip olmuştur.Diğer en yüksek doğruluk oranları Decision Tree ve Random Forest algoritmalarına aittir.Devamında sıralama RBF SVM,MLP,Naive Bayes,LSVM ve KNN şeklindedir. Modellerin performans metrikleri Tablo 2'de gösterilmektedir.

Tablo 5-Modellerin Performans Metrikleri

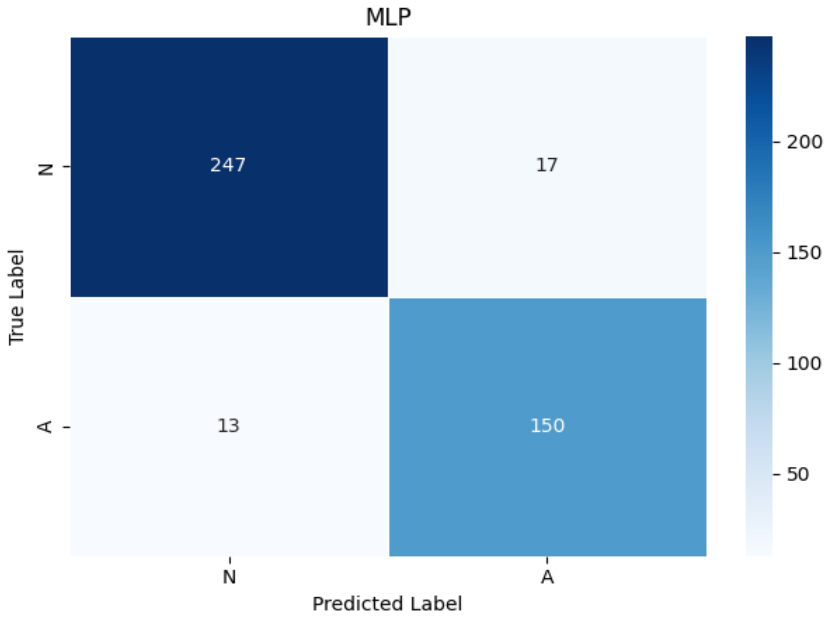
Model	F1 Score	Precision	Recall	Accuracy
K-Nearest_Neighbors	0.904903	0.904220	0.905876	0.881944
Naive Bayes	0.946125	0.940003	0.952620	0.933240
LSVM	0.933108	0.932547	0.933929	0.916842
RBF SVM	0.949845	0.943208	0.956779	0.938161
Random Forest	0.985236	0.988442	0.982096	0.981497
XGBoost	0.988196	0.991781	0.984663	0.985244
Decision Tree	0.981974	0.981137	0.982884	0.977749
MLP	0.947733	0.949250	0.946405	0.934882



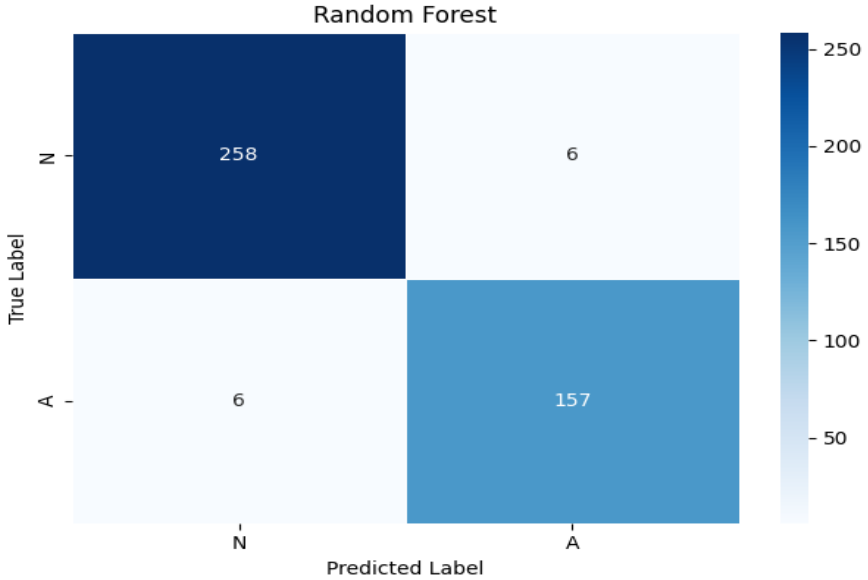
Tablo 6-K-Nearest-Neighbors Karmaşıklık Matrisi



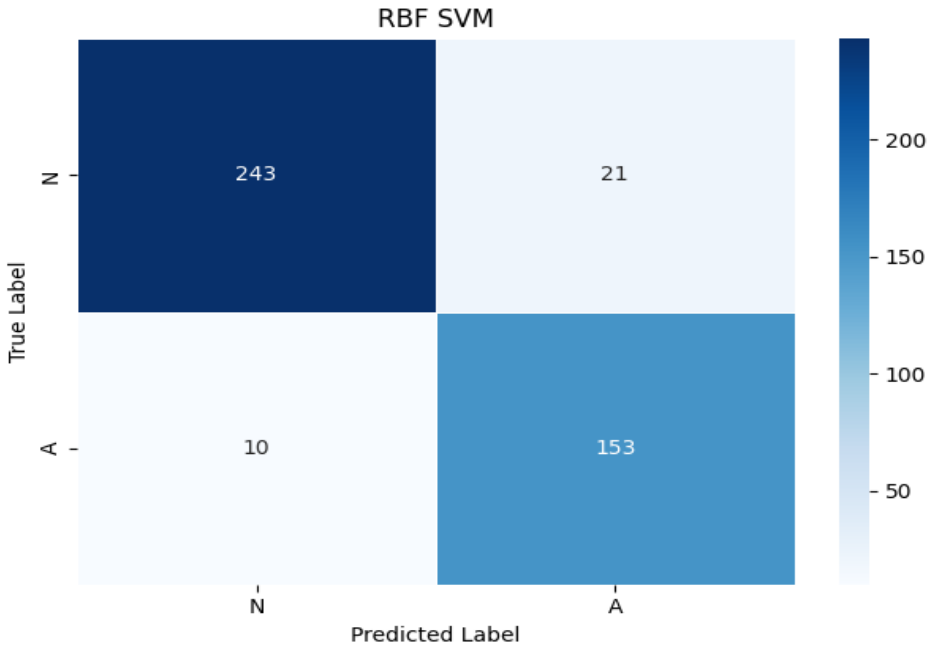
Tablo 7-LSVM Karmaşıklık Matrisi



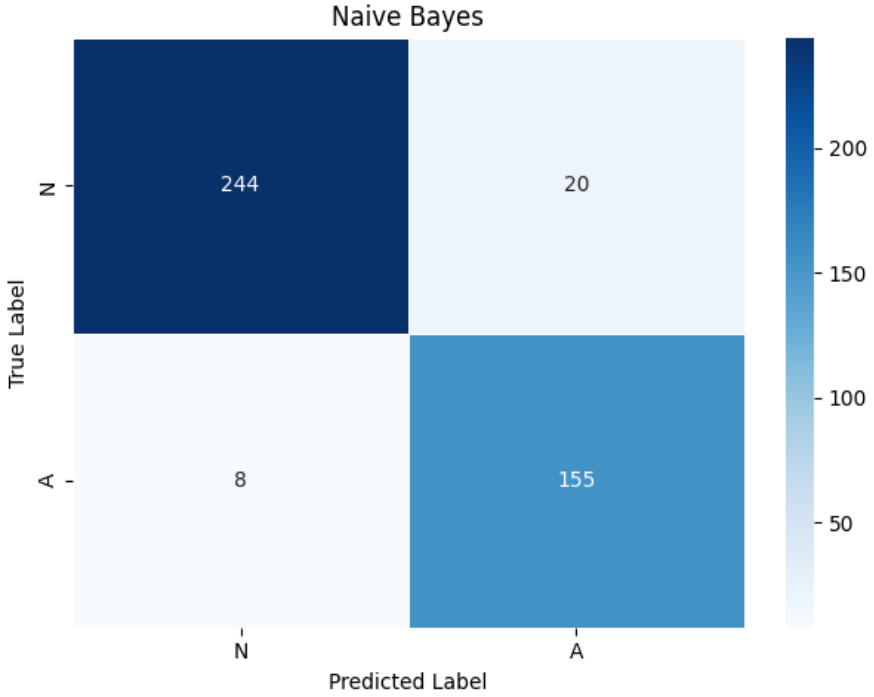
Tablo 8-MLP Karmaşıklık Matrisi



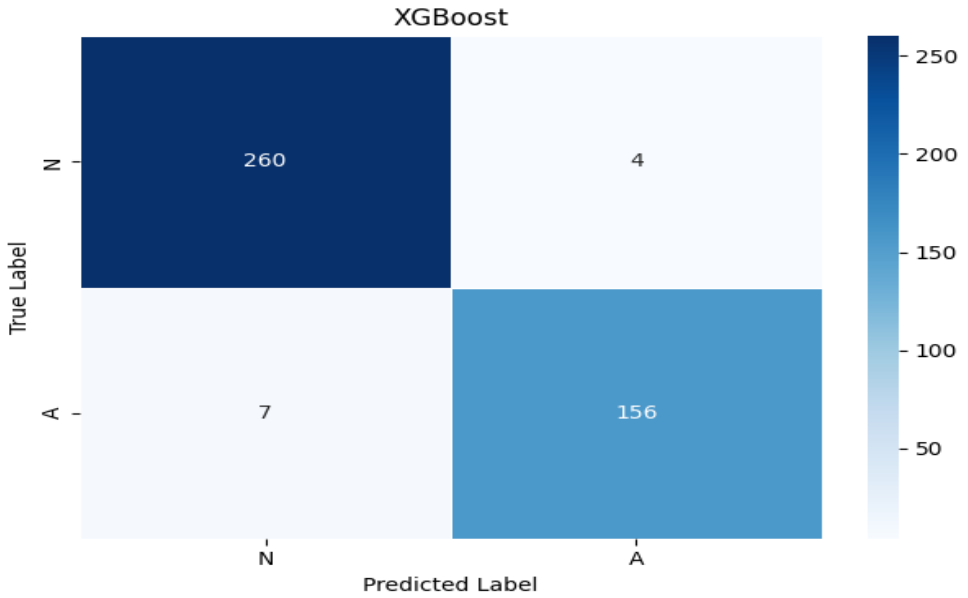
Tablo 9- Random Forest Karmaşıklık Matrisi



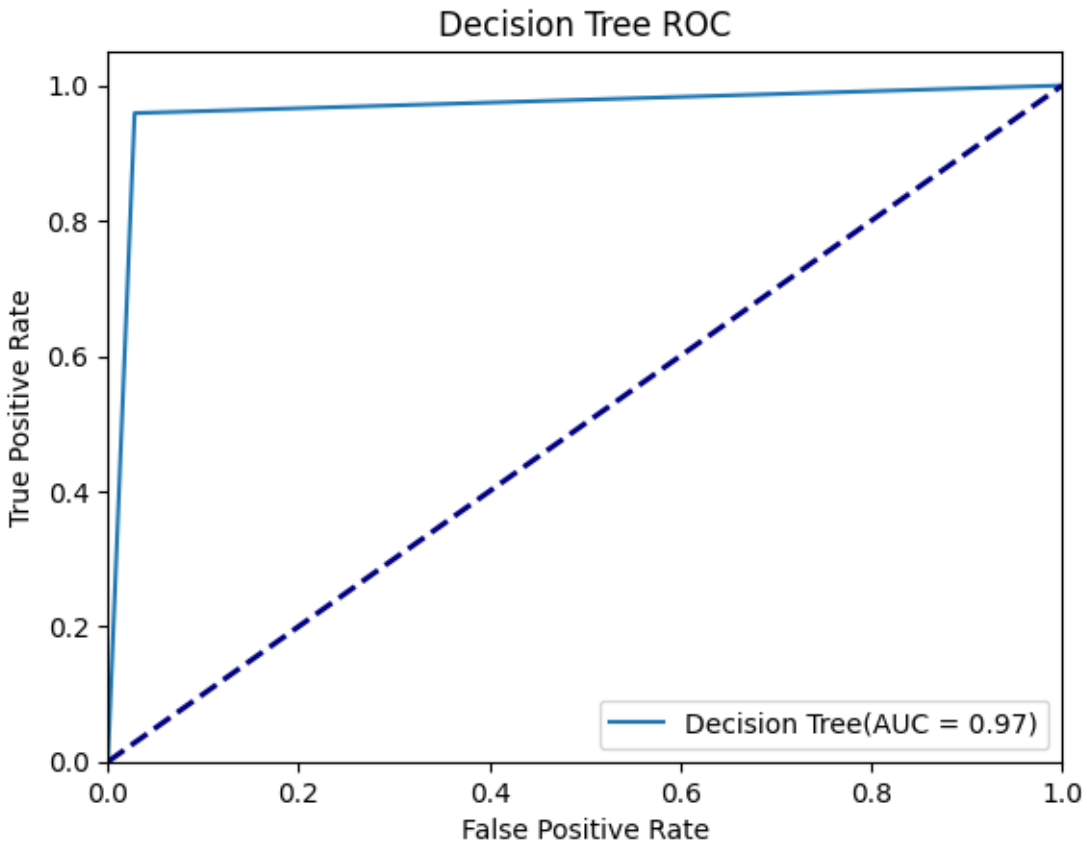
Tablo 10-RBF SVM Karmaşıklık Matrisi

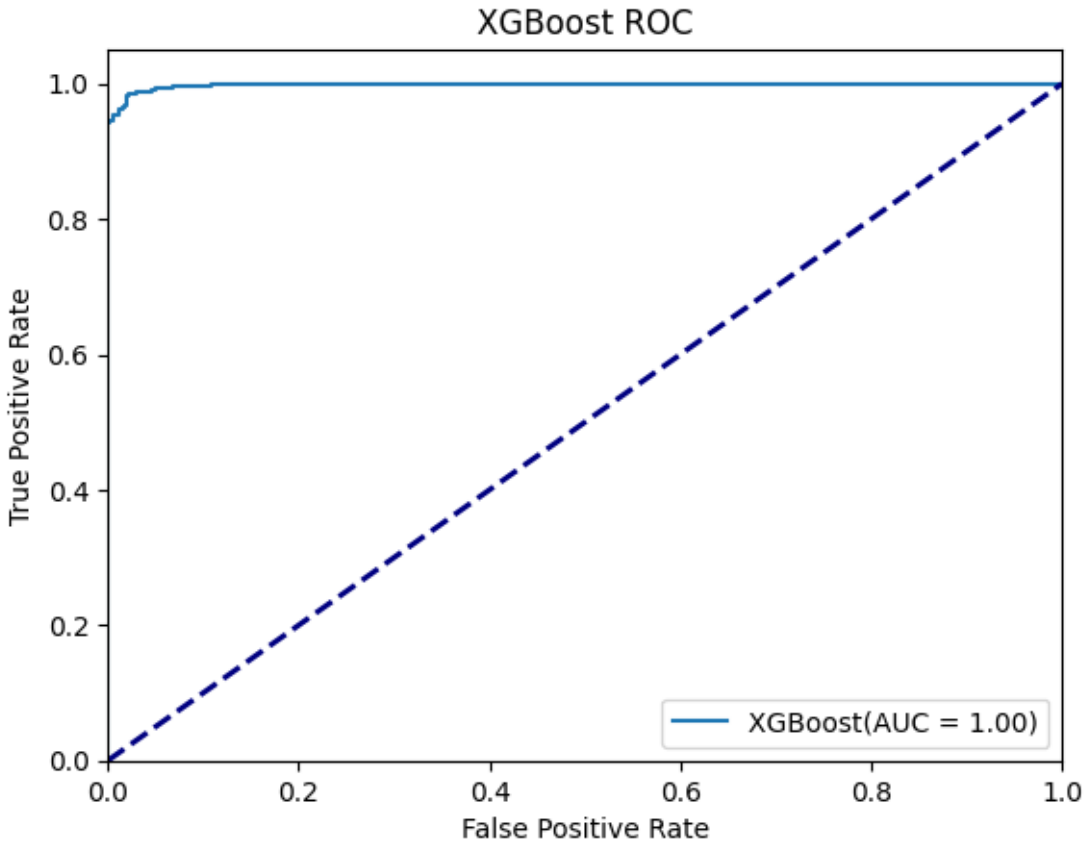


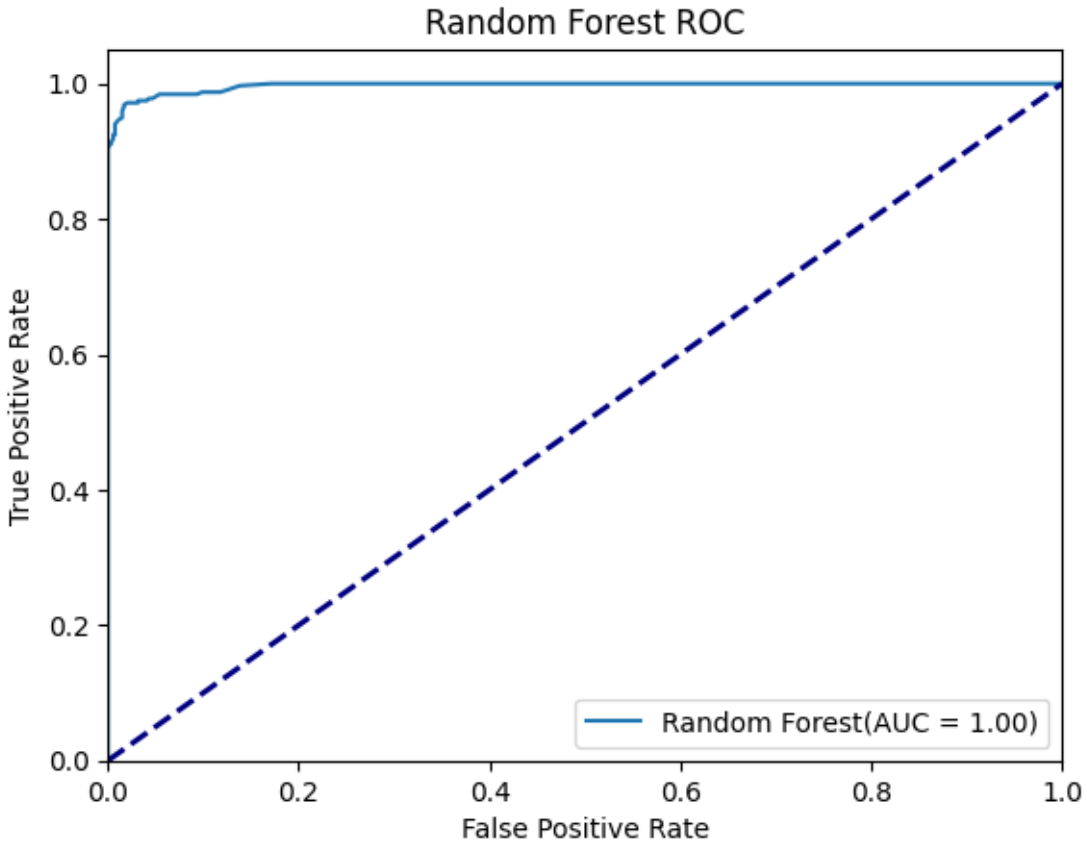
Tablo 11-Naive Bayes Karmaşıklık Matrisi

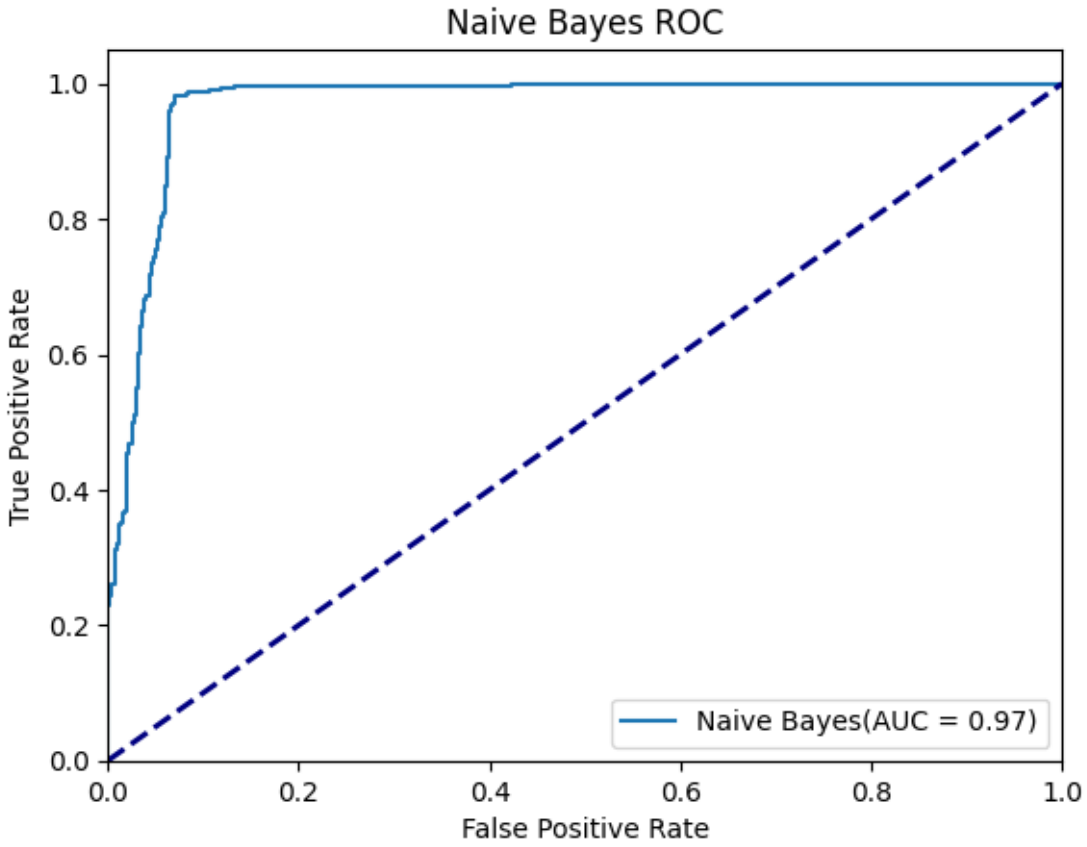


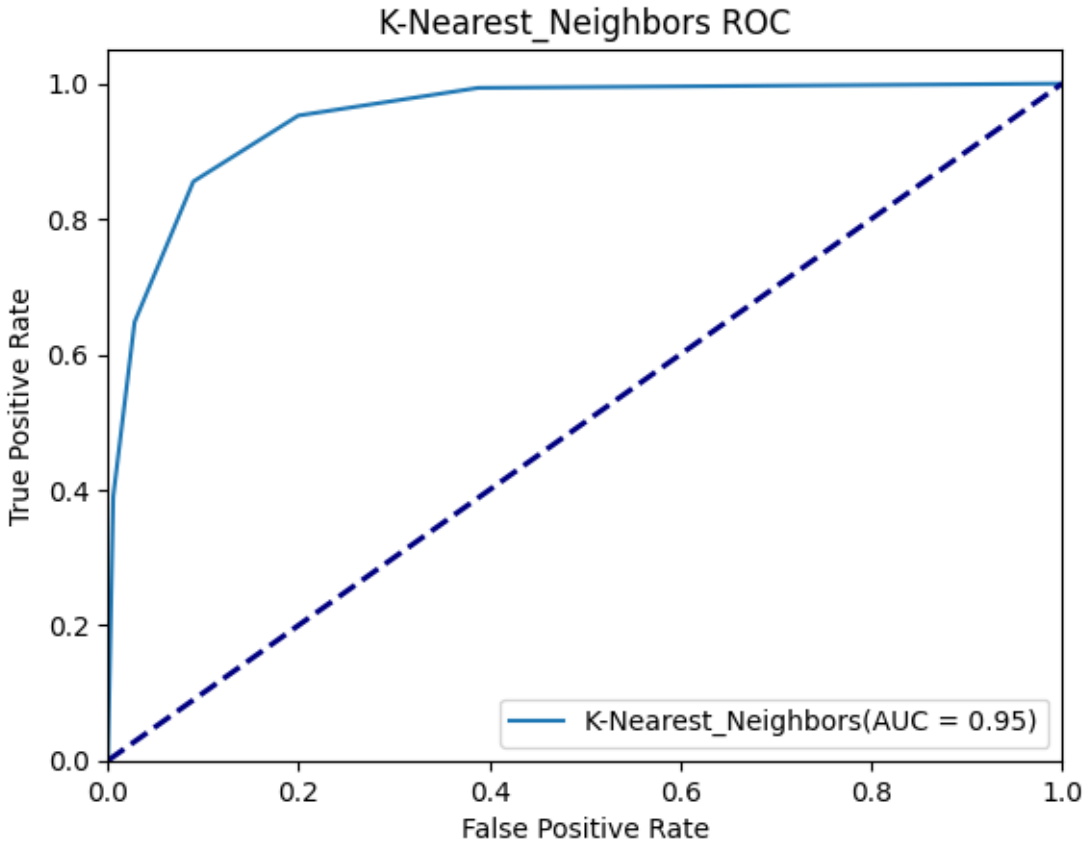
Tablo 12-XGBoost Karmaşıklık Matrisi

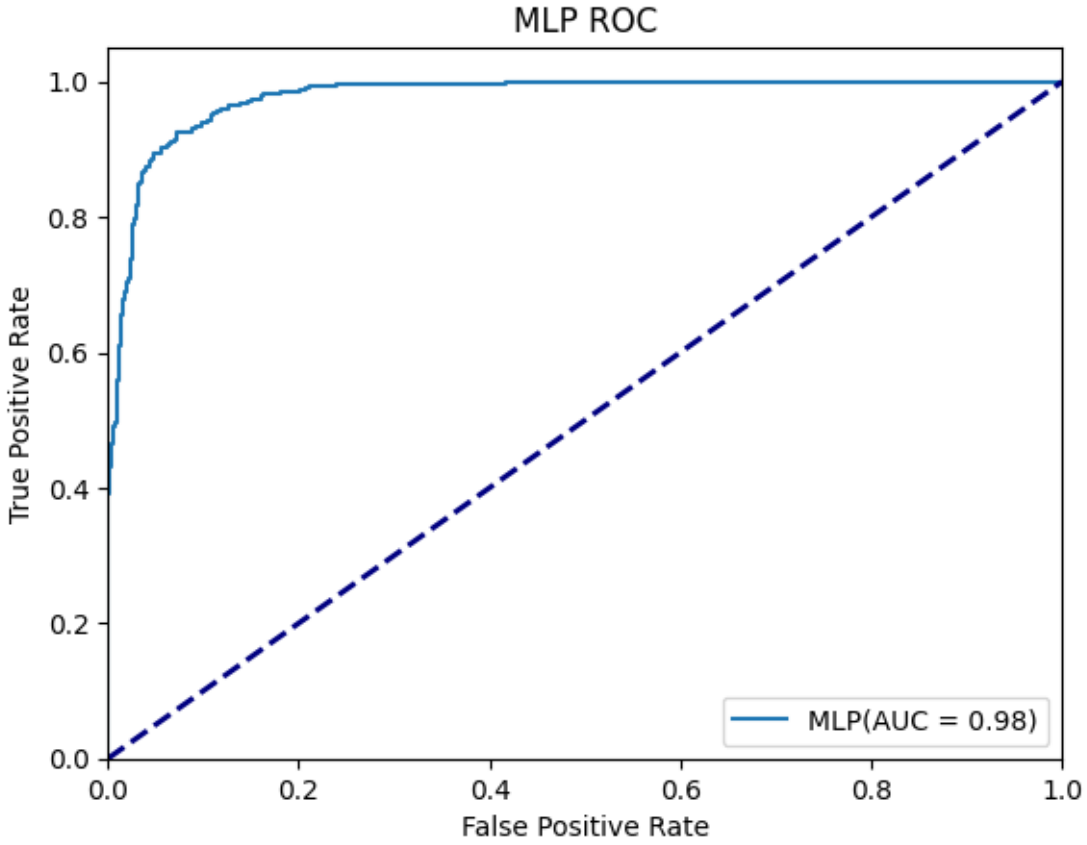


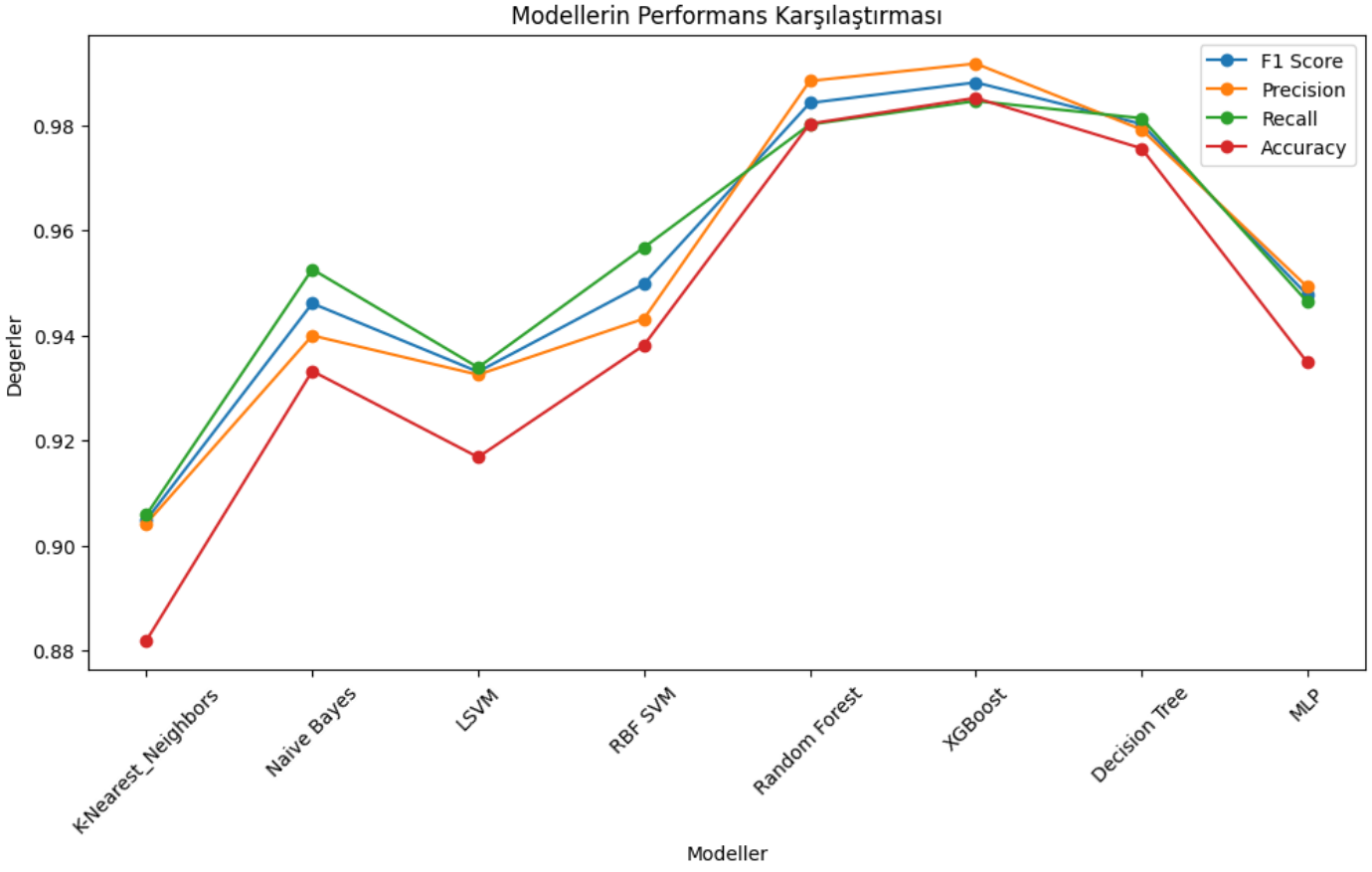

XGBoost ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
XGBoost(AUC = 1.00)


Random Forest ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest(AUC = 1.00)


Naive Bayes ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Naive Bayes(AUC = 0.97)


K-Nearest_Neighbors ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
K-Nearest_Neighbors(AUC = 0.95)


MLP ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
MLP(AUC = 0.98)


Modellerin Performans Karşılaştırması
F1 Score
Precision
Recall
Accuracy
Degerler
0.88
0.90
0.92
0.94
0.96
0.98
K-Nearest_Neighbors
Naive Bayes
LSVM
RBF SVM
Random Forest
XGBoost
Decision Tree
MLP
Modeller

4.SONUÇLAR (CONCLUSIONS)

Sonuç olarak, kredi onayı süreçlerinin makine öğrenimi teknikleri ve veri analizi ile değerlendirilmesi, finansal hizmetler sektöründe hayati bir rol oynamaktadır. Veri ön işleme ve öznitelik seçme adımları sayesinde, kredi değerlendirme modellerinin doğruluğu ve verimliliği artırılmıştır. 10 kat çapraz doğrulama gibi yöntemlerin uygulanması, makine öğrenimi algoritmalarının sağlam bir şekilde değerlendirilmesini sağlar. Bulgularımız, XGBoost, Random Forest ve Decision Tree gibi algoritmaların kredi onayı tahmin görevlerinde yüksek doğruluk oranları (%98) elde etmedeki etkinliğini vurgulamaktadır.

Geleceğe yönelik olarak, makine öğrenimi ve veri analizi tekniklerinin sürekli gelişimi, kredi onayı sistemlerinde daha fazla ilerleme vaat etmektedir. Gelecek araştırmalar, bu modellerin performansını ve güvenilirliğini artırmak için yenilikçi yaklaşımları keşfetmeye odaklanmalıdır. Sonuç olarak, finansal kuruluşların kredi onay süreçlerinde daha doğru ve verimli kararlar almalarına olanak tanıyacaktır.

5.LİTERATÜR KARŞILAŞTIRMASI

	precision	recall	f1-score	support
Approved	0.99	0.96	0.98	633
Rejected	0.96	0.99	0.98	632
accuracy			0.98	1265
macro avg	0.98	0.98	0.98	1265
weighted avg	0.98	0.98	0.98	1265

KAYNAK DOSYALAR (SUPPLEMENTARY FILES)

Bu çalışmada kullanılan veri setlerine aşağıdaki web adresinden ulaşılabilir:

https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset

TEŞEKKÜR (ACKNOWLEDGMENTS)

Bu çalışma, Sayın Murat GÖK tarafından desteklenmiştir. Ayrıca yapmış oldukları katkılardan dolayı Yalova Üniversitesine teşekkür ederiz.

KAYNAKÇA(ACKNOWLEDGMENTS)

[1] Murat GÖK, Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi, FenBilimleri Dergisi, 2007

[2] Wang, Y. et al. (2021). Impact of Data Preprocessing Techniques on Credit Approval Models. International Conference on Data Analysis and Machine Learning

[3] https://medium.com/deep-learning-turkiye/model-performans%C4%B1n%C4%B1-de%C4%9Ferlendirmek-metrikler-cb6568705b1
[4] https://www.kaggle.com/code/hasansezertaan/machine-learning-dersleri-s-n-fland-rma