MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİKSEL ÖĞRENME YÖNTEMLERİYLE BAĞIŞÇI DAVRANIŞLARININ İNCELENMESİ

YÜKSEK LİSANS TEZİ

Emre SIRMA

Anabilim Dalı: İstatistik

Programı: İstatistik

Tez Danışmanı: Prof. Dr. Ayça ÇAKMAK PEHLİVANLI

MAYIS 2024

Emre SIRMA tarafından hazırlanan İSTATİSTİKSEL ÖĞRENME YÖNTEMLERİYLE BAĞIŞÇI DAVRANIŞLARININ İNCELENMESİ adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Prof. Dr. Ayça ÇAKMAK PEHLİVANLI
Tez Yöneticisi

Bu çalışma, jürimiz tarafından İSTATİSTİK Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

Başkan : Prof. Dr. Ayça ÇAKMAK PEHLİVANLI

Üye : Prof. Dr. Özge ÇAĞÇAĞ YOLCU

Üye : Dr. Öğr. Üyesi Bilge ÖZLÜER BAŞER

Bu tez, Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygundur.

Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım klavuzuna uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel etik kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- ücret karşılığı başka kişilere yazdırmadığımı (dikte etme dışında), uygulamalarımı yaptırmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

İSTATİSTİKSEL ÖĞRENME YÖNTEMLERİYLE BAĞIŞÇI DAVRANIŞLARININ İNCELENMESİ

ÖZET

Sivil toplum kurumları, ihtiyacı olan bireylere yardım sağlamak ve sosyal sorunlara çözüm bulmak amacıyla faaliyet gösteren, kar amacı gütmeyen kuruluşlardır. Sivil toplum kurumları, faaliyetlerini sürdürebilmek için bağış toplamak zorundadırlar. Bağışçıların desteğini alabilmek için ise çeşitli yaklaşımlar kullanmaktadırlar. Bu yaklaşımlardan bir tanesi de istatistiksel öğrenme yöntemlerini kullanmaktır.

İstatistiksel öğrenme, sivil toplum kurumlarının bağışçılarla etkileşimini artırmak, bağış miktarlarını optimize etmek ve bağışçıların bağışlarına devam etmelerine yönelik stratejiler oluşturmasında kullanılır. İstatistiksel öğrenme yöntemleri, bağışçı profillerini analiz ederek demografik veriler, bağış geçmişi, eğitim düzeyi gibi faktörleri kullanarak bağışçı davranışlarının tahmin edilmesini sağlamaktadır.

Özellikle, düzenli bağış talimatı oluşturan bağışçıların talimatlarına devam edip etmeyecekleri konusunda yapılan tahmin çalışmaları, sivil toplum kuruluşları için kritik öneme sahiptir. Bu çalışma kapsamında, 2010-2024 yılları arasında yapılan bağışları kapsayan düzenli bağış talimatı veren toplam 38.913 bağışçıya ilişkin veri kullanılmıştır. Yapılan tez çalışmasında, farklı istatistiksel öğrenme yöntemleri kullanılarak düzenli bağış talimatı oluşturan bağışçıların talimatlarını iptal edip etmeyeceği üzerine bir tahmin çalışması gerçekleştirilmiştir. Rastgele orman, lojistik regresyon, destek vektör makineleri, naive bayes, XGBoost ve LightGBM modelleri kullanılarak gerçekleştirilen çalışmada, en başarılı modelin belirlenmesi için modellerin tahmin performansları karşılaştırılmıştır. Elde edilen sonuçlara göre XGBoost, LightGBM ve rastgele orman modellerinin özellikle iptal edilen talimatları tahmin etme konusunda diğer modellere göre daha etkili olduğu belirlenmiştir.

Yapılan çalışma ile, sivil toplum kurumlarının bağışçılarıyla etkileşimini artırmak, kaynaklarını daha etkin kullanmak ve sosyal amaçlarına daha etkili bir şekilde hizmet etmek için veri odaklı kaynak geliştirme faaliyetlerinin yürütülmesine katkıda bulunulması hedeflenmektedir.

INVESTIGATION OF DONOR BEHAVIOUR USING STATISTICAL LEARNING METHODS

ABSTRACT

Non-governmental organizations are non-profit organizations that operate to provide assistance to individuals in need and to find solutions to social problems. Non-governmental organizations have to collect donations to continue their activities. They use various approaches to gain the support of donors. One of these approaches is to use statistical learning methods.

Statistical learning is used to create strategies for non-governmental organizations to increase their interaction with donors, optimize donation amounts and enable donors to continue their donations. Statistical learning methods analyze donor profiles and predict donor behavior using factors such as demographic data, donation history and education level.

In particular, prediction studies on whether donors who set regular donation instructions will continue their instructions are of critical importance for non-governmental organizations. The data used in this study belongs to a total of 38,913 donors who gave regular donation instructions, covering donations made between 2010-2024. In the thesis study, a prediction study was carried out on whether donors who created regular donation orders would cancel their orders by using different statistical learning methods. In the study carried out using random forest, logistic regression, support vector machines, naive bayes, XGBoost and LightGBM models, the prediction performances of the models were compared to determine the most successful model. According to the results obtained, it was determined that XGBoost, LightGBM and random forest models were more effective than other models, especially in predicting canceled instructions.

The study aims to contribute to the execution of data-driven resource development activities in order to increase the interaction of non-governmental institutions with their donors, use their resources more effectively and serve their social goals more effectively.

ÖNSÖZ

Bu tez çalışması sürecinde beni her adımda destekleyen, rehberlik eden ve katkılarıyla beni aydınlatan değerli hocam ve tez danışmanım Prof. Dr. Ayça ÇAKMAK PEHLİVANLI'ya, lisansüstü eğitim hayatım boyunca her zaman desteğini esirgemeyen Mimar Sinan Güzel Sanatlar Üniversitesi, İstatistik Bölümü'nün tüm öğretim üyeleri ve elemanlarına, hayatımın her döneminde benimle olan, maddi ve manevi desteğini esirgemeyen aileme teşekkürlerimi sunarım.

Mayıs, 2024 Emre SIRMA

İÇİNDEKİLER

Sayfa

KISALTMALAR

STK **:** Sivil Toplum Kuruluşu
DMEF **:** Direct Marketing Educational Foundation
DVM **:** Destek Vektör Makineleri
CART **:** Classification and Regression Tree
GDM **:** Genelleştirilmiş Doğrusal Modeller

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

1. GİRİŞ

Günümüzde, sivil toplum kuruluşları, toplumun çeşitli kesimlerine yardım etmek ve sosyal fayda sağlamak için önemli bir rol oynamaktadır. Sivil toplum kuruluşları, kurumsal yapılarını sürdürmek ve ihtiyaç duydukları gelirleri elde edebilmek için çeşitli kaynak geliştirme faaliyetleri sürdürmektedir. Bu kuruluşlar genellikle bağışlar yoluyla fon toplamakta ve bu fonları ihtiyaç sahiplerine ulaştırmaktadırlar. Ancak bağış elde etmek için yapılan faaliyetler, STK'lar için yüksek maliyetlere sebep olabilmektedir [1].

Bağışların sürdürülebilirliği ve etkinliği, bağışçı davranışlarının anlaşılması ve tahmin edilmesi ile doğrudan ilişkilidir. Bağışçı davranışlarının analizi ve tahmini, bağış platformları ve yardım kuruluşları için stratejik öneme sahiptir. Bu analizler, bağışçıların ne zaman ve nasıl bağış yapacaklarını belirlemek, bağışçıları motive etmek ve bağış kampanyalarını yönlendirmek için kritik bilgiler sağlar. Bu nedenle, bağışçı davranışlarını etkileyen faktörlerin ve bu davranışların gelecekte nasıl değişebileceğinin anlaşılması büyük önem taşır.

İstatistiksel öğrenme yaklaşımları, büyük miktarda veriyi analiz etme ve desenleri belirleme yeteneğiyle bağışçı davranışlarını anlama ve tahmin etme konusunda güçlü bir araçtır [2]. Bu teknikler, bağışçıların geçmiş bağış alışkanlıklarını, demografik özelliklerini ve diğer ilgili faktörleri kullanarak gelecekteki bağış davranışlarını tahmin etmek için kullanılabilir.

Farklı strateji sonuçlarının tahmin edilmesi ve veri temelli iç görülere dayalı öneriler sunulması STK'ların kaynaklarını optimize etmelerine yardımcı olabilir.

1.1 Tezin Amacı

Bağışçı davranışlarının incelenmesi ve tanımlanması bağış miktarlarında artışa olanak sağlayacaktır. Bu doğrultuda özellikle istatistiksel öğrenme teknikleri yardımıyla bağışçıların davranışlarına ve bağış miktarlarına yönelik tahminler önem arz etmektedir. Bu çalışma kapsamında çeşitli sınıflama algoritmaları aracılığıyla bağışçı

davranışlarının incelenerek elde edilen bulgular üzerinden bağışlar ve bağışçılar bazında hangi değişkenlerin daha fazla etkili olduğu tespit edilmeye çalışılacaktır.

Konuya ilişkin gerek veri gerekse yöntemleri içeren çalışmalar literatürde ayrıntılı olarak araştırılacak ve olumlu/olumsuz yönleri ile irdelenecektir. Bu analizin sonucunda, geliştirilebilecek olan yaklaşımlar belirlenerek, literatürde bulunan eksik ve olumsuz yönlere çözüm olabilecek yaklaşımlar üzerine kapsam belirlenecektir.

Bu tez, düzenli bağışçıların daha önce oluşturdukları düzenli bağış talimatlarını iptal edip etmeyeceklerini tahmin etmeye yönelik olarak hazırlanmıştır. Bu bağlamda, bir sivil toplum kuruluşundan elde edilen veri seti kullanılacaktır. Talimat iptalinin modellenmesinde rastgele orman, lojistik regresyon, destek vektör makineleri, naive bayes, XGBoost, ve LightGBM modelleri kullanılmış ve bu modellerin tahmin performansları karşılaştırılmıştır. Bu modellerin seçilmesinin temel sebepleri, geniş bir makine öğrenimi yelpazesini kapsamaları ve farklı algoritmik yaklaşımları temsil etmeleridir. Rastgele orman, XGBoost ve LightGBM gibi ağaç tabanlı yöntemler, karmaşık ilişkileri modelleyebilme ve ölçeklenebilirlik özelliklerine sahiptirler. Lojistik regresyon ise doğrusal bir model olup, daha basit bir yapıya sahiptir ve katsayıların yorumlanabilirliği açısından avantaj sağlar. Destek vektör makineleri, lineer ve non-lineer ayrım sınıflandırması özelliğine sahiptir. Naive bayes ise olasılık temelli bir modeldir. Modellerin bir arada kullanılması, çeşitli model yaklaşımlarını değerlendirerek genel model performansını artırılmasını sağlamaktadır. Farklı model türlerinin güçlü ve zayıf yönlerini karşılaştırarak, veri setinin karmaşıklığına ve yapısına daha iyi uyum sağlayabilen bir model seçme esnekliği sağlanabilmektedir. Bu yaklaşım, tek bir modelin belirli durumlarda yetersiz kalabileceği durumlarda daha güvenilir tahminler yapılmasını sağlamaktadır.

1.2 Literatür Taraması

Makine öğrenmesi ile bağışçıların geçmiş bağış örüntülerine ve diğer özelliklerine dayanarak gelecekteki bağış davranışlarının tahmin edilmesi sağlanmaktadır. Yüksek bağış olasılığına sahip bağışçılar, bağış toplama kampanyalarında hedef alınabilmektedir. Makine öğrenmesi, düşük bağış olasılığına sahip bağışçılarla iletişim kurma maliyetini azaltır ve bağış kampanyalarının etkinliğini artırır [3].

Worcester Polytechnic Enstitüsü İşletme Okulu'ndan Profesör Andrew C. Trapp, Ulusal Bilim Vakfı (NSF) tarafından 320.000 dolarlık bir hibe alarak, mültecilerin yeni bir ülkeye yerleşme ve entegrasyon şanslarını artırmaya yönelik bir hesaplama aracı geliştirmek üzere bir proje yürütmüştür [4]. Bu proje, mültecileri uygun kaynaklara, özellikle iş olanaklarına, yerleştirmek için makine öğrenimi ve optimizasyon algoritmalarını içeren karmaşık veri hesaplamalarını kullanmaktadır. Trapp ve ark., bu araştırma ile mültecilerin iş bulma olasılıklarını önemli ölçüde artırabilecek yeni yöntemler geliştirmeyi hedeflemişlerdir. Bu kapsamda geliştirdikleri Annie Moore adlı yazılım ile mültecilerin istihdam olasılıklarını tahmin ederek onları uygun yerleşim bölgelerine yönlendirmeye yardımcı olmaktadırlar.

Key, Amerika Birleşik Devletleri'nin Kuzeydoğu bölgesindeki özel bir Katolik lisesinden alınan 10.828 kayıttan oluşan bir veri seti üzerine yaptığı çalışma sonucunda yaşın ve gelirin bağış yapma eğilimi üzerinde önemli etkileri olduğunu belirtmiştir. Yaş ile bağış yapma olasılığı arasındaki ilişki kademeli bir şekilde artıp sonra azalmaktadır. Araştırma, en az bir araba kredisi olan bireylerin ve erkeklerin, bu Katolik lisesine bağış yapma olasılığının daha yüksek olduğunu göstermektedir [5]. Bu bulgular, bağış yapma eğiliminin yaş, gelir ve cinsiyet gibi faktörlere dayandığını göstermektedir.

Key 2001 yılında yaptığı bu çalışmayı büyük bir metropol müzesinden alınan ve 160.484 kayıttan oluşan bir veri seti üzerinde de test etmiştir. Müzeden elde edilen veri seti, üyeliklerin süresi ve seviyesi, ilgi alanları, müze destekli seyahatler, komite katılımı ve çocuk sayısı gibi değişkenler içermektedir. Araştırma, müzenin bağışçılarının genellikle müzeyle güçlü bir ilişkiye sahip olduklarını ve bağış yapma olasılığının kişinin gelir düzeyiyle arttığını gösterirken, yaş bakımından bağış yapanlar ile yapmayanlar arasında fark olmadığını ortaya koymuştur [5].

Malthouse 2001 yılında yayımladığı çalışmada, bir modelin eğitim veri setine optimize edilmiş uyumunun, puanlama modelinin temel amacı olmadığını savundu [6]. Malthouse'e göre potansiyel bir bağışçının puanı ne kadar yüksek olursa, potansiyel bağışçının bağış yapma olasılığı da o kadar yüksek olur. Çalışmada ayrıca, bir modelin kurulması için eğitim verileri yerine, doğrulama verilerinden elde edilen ağırlıkları kullanarak parametreleri tahmin etmek için performansa dayalı bir puanlama modeli önerilmiştir. Bağış veri setlerinin en üst %20 ile %40'ı arasındaki mail derinliğinde

ortalama %3 ile %4'lük bir iyileşme rapor edilmiştir. Promosyonun milyonlarca potansiyel bağışçıya gönderilmesi halinde, iyileşme yüz binlerce dolara ulaşabilir. Malthouse, yöntemindeki ağırlık hesaplamalarının pahalı olması nedeniyle çok sayıda aday değişkenin dahil edilmesine karşı uyarıda bulunmuştur [6].

Zhao ve arkadaşlarının 2019 yılında gerçekleştirdiği çalışmada kitle fonlamasında bağış geçmişini analiz etmek için büyük ölçekli gerçek veriler kullanılmıştır [7]. Bağışın tekrar etmesi ve bağışçı sadakati üzerine odaklanılan bu çalışmada iki durumu aynı anda modelleyebilen Joint Deep Survival modeli (JDS) önerilmiştir. Yapılan deneylerde, kitle fonlamasındaki bağışlar analiz edilmiş ve JDS'nin bağış tekrarı ve bağışçı sadakati üzerine tahmin performansı çeşitli yönlerden doğrulanmıştır. Çalışma, kitle fonlamasının uygulama açısından ve derin öğrenme kullanarak sağkalım analizi (survival analysis) için teknik açıdan yeni bir bakış açısı sunmaktadır [7].

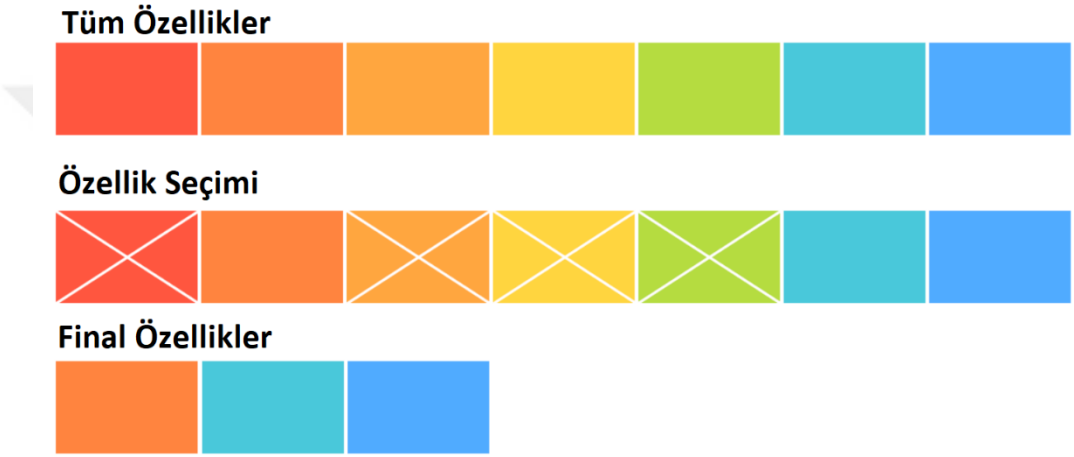
Sivil toplum kuruluşları gibi çevrimiçi bağış platformları da bağışçılarla ilişkiyi sürdürme konusunda zorluklarla karşılaşmaktadır. Althoff ve Leskovec yaptıkları çalışmada, DonorsChoose.org adlı bir bağış platformundaki bağışçı davranışlarını analiz etmişlerdir [8]. Özellikle ilk kez bağış yapanlar üzerinde odaklanarak, bağışçıların tekrar bağış yapma eğilimlerini başarılı bir şekilde tahmin etmek için ilk bağış etkileşimlerinden çok şey öğrenilebileceğini ve bu öngörülerin basit makine öğrenmesi modelleriyle yüksek doğrulukla tahmin edilebileceğini göstermektedir [8].

Doğu Afrika'daki çiftçilere varlık temelli finansman ve tarım eğitimi hizmeti sağlayan One Acre Fund, çiftçilere optimal ekim zamanını tahmin etmek için bir sohbet robotu geliştirmiştir [9]. One Acre Fund tarafından 2023 yılında yapılan bu çalışma, iklim değişikliği nedeniyle zorlaşan yağmur tahminlerini ele almak amacıyla, tahmin modelleri kullanılarak oluşturulan sohbet robotu ile, çiftçilere doğrudan mesajlar gönderilebilmekte ve onlardan geri bildirimler alınarak iletişim kurulabilmektedir. Bu çözüm, çiftçilerin verimliliğini artırmak için ekim zamanını en uygun şekilde belirlemelerine yardımcı olurken gelirlerini artırmıştır. Bu çözüm aynı zamanda One Acre Fund'a geniş kitlelere ulaşma ve destek hizmetlerini ölçeklendirme imkanı sağlamıştır.

2. YÖNTEMLER

2.1. Özellik Seçimi (Feature Selection) ve Yöntemleri

Özellik seçimi, veri önişleme aşamasında gerçekleştirilerek sınıflandırma performansı artırmak bakımından oldukça önemli ve gerekli bir adımdır. Bu sürecin amacı, veri setini en iyi şekilde temsil edebilecek, en ilgili ve bilgilendirici özellikleri içerecek bir özellik alt kümesi seçmektir. Yani, n adet özellik arasından en iyi k özelliği seçme işlemidir (Şekil 2.1).



Şekil 2.1 : Özellik Seçimi [10]

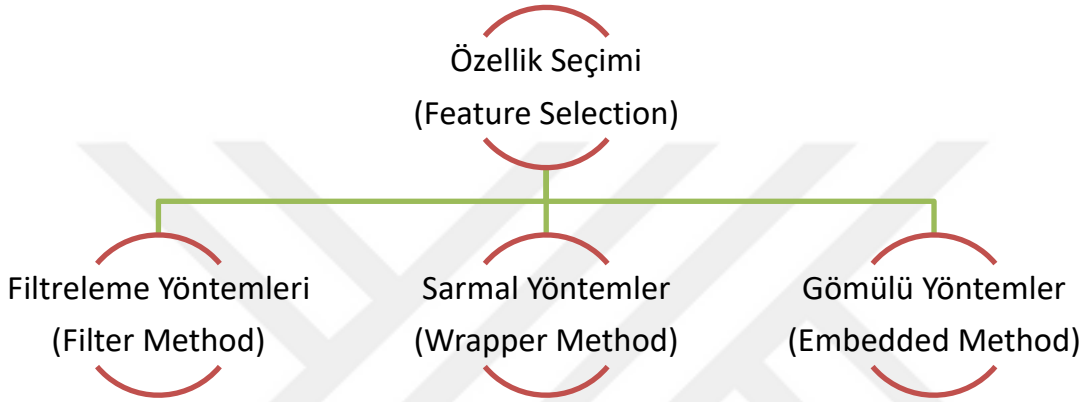
Özellik seçimi, ilgili problem için en yararlı ve önemli özellikleri seçerek veri setindeki özellik sayısını azaltmayı amaçlar. Özellik seçimi işleminin avantajları şunlardır:

- Algoritmanın hızını artırır.
- Gürültülü veriyi azaltır.
- Veri kalitesini artırır.
- Veri kümesini daha basit ve anlaşılır hale getirir.
- Veri depolamak için gereken hafıza miktarını azaltır.
- En iyi özelliklerle çalışmak, elde edilen modelin başarısını artırır.

Özellik seçimi, metin madenciliği, kanser teşhisi, sahtecilik tespiti, müşteri kaybı analizi gibi birçok alanda kullanılmaktadır. Bu amaçla kullanılan özellik seçim yöntemleri arasında, istatistiksel bilgiye dayanan filtreleme yöntemi, özellikler

üzerinde arama yapmayı sağlayan sarmal yöntemler ve en iyi bölme ölçütünü bulmaya dayalı gömülü yöntemler olmak üzere üç temel yaklaşımda incelenebilir [11].

Özellik seçim yöntemleri, Şekil 2.2'de gösterildiği gibi üç kategoriye ayrılır. Filtreleme yöntemleri, ilk olarak özellik seçimi yapar ve ardından öğrenme algoritması çalıştırılır. Sarmal yöntemlerde ise, öğrenme algoritması en iyi özellikleri seçmek için bir araç olarak kullanılır. Gömülü yöntemlerde ise, öğrenme ve özellik seçimi algoritmaları senkron bir şekilde çalışır.



Şekil 2.2 : Özellik Seçim Yöntemleri

2.1.1. Filtreleme Yöntemleri (Filter Methods)

Filtreleme yöntemleri, makine öğrenmesi alanında kullanılan en eski özellik seçim metotlarından biridir. Bu metotlarda, herhangi bir sınıflandırıcı kullanılmadan, belirli (uzaklık, bilgi vb.) ölçümler istatistiksel kriterlere dayalı fonksiyonlar aracılığıyla değerlendirilerek özellik seçimi gerçekleştirilir. Bu yöntemlerde, her bir özellik için veri setinde bir değerlendirme fonksiyonu kullanılarak bir değer elde edilir ve bu değerler arasında en yüksek önem değerine sahip olan özellikler en iyi özellik alt kümesine eklenir [12].

Fisher Skor, T-Skor, Welch T-İstatistiği, Ki-Kare Testi, Bilgi Kazancı, Kazanç Oranı, Korelasyon Tabanlı, Relielf, One-R gibi yöntemler yaygın olarak kullanılan filtreleme yöntemleridir [12].

2.1.2. Sarmal Yöntemler (Wrapper Methods)

Sarmal yöntemlerde özellik seçim süreci bir öğrenme algoritması içerir ve özellikleri öğrenme algoritmasının performansına göre seçer. Her bir özellik için, özellik alt kümesine eklenip eklenmeyeceği veya kaldırılıp kaldırılmayacağı belirlenirken, öğrenme algoritmasının başarı oranını değerlendirir. En iyi tahmin performansını sergileyen özellikler seçilir [13]. Sarmal yöntemler, filtreleme yöntemlerine göre genellikle daha etkilidir, ancak hesaplama maliyeti daha yüksektir. Yaygın olarak kullanılan sarmal yöntemler arasında Ardışık İleri Yönde Seçim, Ardışık Geri Yönde Seçim, L Ekle R Çıkar, Ardışık İleri Yönde Kayan Seçim, Ardışık İleri Yönde Kayan Seçim vb. sayılabilir.

2.1.3. Gömülü Yöntemler (Embedded Methods)

Gömülü yöntemler, makine öğrenimi modellerinin eğitim sürecinde değişken seçimini doğrudan gerçekleştiren tekniklerdir. Bu yöntemler, modelin performansını artırmak için en uygun değişkenleri seçmeye odaklanır ve aynı zamanda modelin karmaşıklığını azaltır [11]. Modelin performansını artırmak için yaygın olarak kullanılan gömülü yöntemler, çeşitli teknikler aracılığıyla değişken seçimini gerçekleştirir ve genellikle doğrusal veya ağaç-tabanlı modellerle ilişkilendirilirler. Bu teknikler arasında veri setinin özelliklerine ve modelin gereksinimlerine bağlı seçilebilen Lasso Regresyon, Ridge Regresyon, Elastic Net Regresyon, rastgele orman ve Gradient Boosting Makine (GBM) vb. bulunmaktadır.

2.2. Sınıflandırma Algoritmaları

2.2.1. Rastgele Orman (Random Forest)

Breiman tarafından bulunan rastgele orman algoritması, topluluk tarafından öğrenilen algoritmalardan bir tanesidir [14]. Komite öğrenme algoritmaları, ortak kararlar elde etmek için birden fazla model oluşturur. Rastgele orman algoritmasının temel öğrenicisi, CART (sınıflandırma ve regresyon ağacı) algoritmasıdır. Bu algoritma, genelleme konusunda oldukça başarılıdır ve bu başarı, varyans-yanlılık ilişkisindeki varyansın etkili bir şekilde azaltılmasına bağlıdır.

Model, ana hedef olarak birbirinden bağımsız yüzlerce karar ağacı oluşturmayı amaçlar. Modelin etkin çalışması için bu tekil karar ağaçlarının istatistiksel olarak

birbirinden bağımsız olması kritiktir. İstatistiksel bağımsızlık yeterince sağlanmazsa, modelin varyansındaki düşüş yeterli olmayabilir ve bu da genel model performansının düşük olmasına neden olabilir. Bu nedenle, rastgele orman algoritması, tekil ağaçların birbirinden bağımsızlığını sağlamak için iki tür rassallık dahil eder. Bu rasyonel unsurlar aynı zamanda modelin adını oluşturur.

Rastgele orman algoritmasının kalibrasyon parametreleri oldukça basit bir şekilde ayarlanabilir. Modelin oluşturacağı tekil ağaçların sayısı (m) ve her bir ağaç için rassal olarak seçilen açıklayıcı değişken sayısı (k), kalibrasyon parametreleri arasında en büyük etkiye sahiptir [15].

Rastgele orman algoritması, temel öğrenici olarak CART algoritmasını ve karar verici olarak modellerin ortalamasını kullandığı için, aşırı uyum (overfitting), yüksek boyutluluk ve uç değerler gibi sorunlara karşı son derece dirençlidir. Bu özellikleriyle, diğer komite modellerinden daha az ön veri işleme gerektiren bir seçenek sunar ve genel tahmin performansı yüksek olabilir. Bu nedenle, bu modeller çekici hale gelir. [16].

2.2.2. Lojistik Regresyon

Lojistik regresyon, genelleştirilmiş doğrusal modeller (GDM) ailesinin bir yöntemidir. Adından da anlaşılacağı gibi parametrik, yapısal ve açıklayıcı değişkenlerle bağımlı değişkenler arasında doğrusal bir ilişki kurmayı amaçlayan doğrusal regresyon benzeri bir modeldir. Bağımlı değişkenin kategorik ölçüm düzeyine sahip olması durumunda, bağımsız değişkenler ile arasındaki ilişki lojistik regresyon kullanılarak ortaya konabilir. Eşitlik 2.1'de lojistik regresyon modeli belirtilmiştir.

$$\boldsymbol{log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k} \tag{2.1}$$

Bu durumda *k*, bağımsız değişkenlerin toplam sayısını temsil eder. Model, katsayıları hesaplayarak *p* değerini elde eder. Model katsayılarıyla hesaplanan *p* değeri, bir olasılık değeridir ve 0 ile 1 arasında olmalıdır.

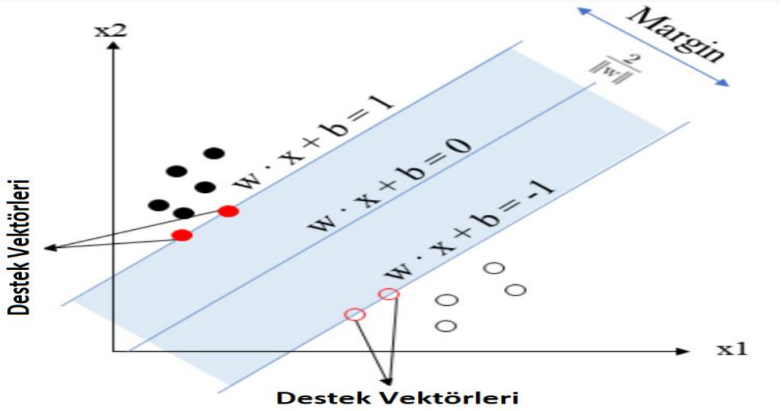
p olasılık değeri Eşitlik 2.2'de verildiği şekilde ifade edilir ve sigmoid fonksiyonu olarak adlandırılır:

$$\boldsymbol{p} = \frac{\mathbf{1}}{\mathbf{1}-\boldsymbol{exp}[-(\boldsymbol{\beta_0}+\boldsymbol{\beta_1 x_1}+\cdots+\boldsymbol{\beta_k x_k})]} \tag{2.2}$$

Lojistik regresyon, açıklayıcı değişkenler doğrusal olmayan ilişkileri modellemek için uygun dönüşümler yapılmadığı sürece doğrusal karar sınırları oluşturur. Bu durum, modelin tahminleme performansının en büyük dezavantajlarından biridir, çünkü gerçek dünyada birçok problem doğası gereği doğrusal olmayan karar sınırlarını gerektirir. Öte yandan, açıklayıcı değişkenlerin eğim parametrelerinin istatistiksel olarak anlamlı olup olmadığını değerlendirebilme olanağı nedeniyle istatistiksel çıkarım problemlerine uygundur. Birçok modelin tahminleme başarısının lojistik regresyondan daha iyi olsa bile lojistik regresyonda istatistiksel çıkarım yapma zorluğu yaşanmaz [16].

2.2.3. Destek Vektör Makineleri

Vladimir Vapnik ve Alexey Chervonenkis tarafından 1963 yılında geliştirilen Destek Vektör Makinesi (Support Vector Machines-SVM), etiketli bir eğitim verisi kümesinden girdi-çıktı haritalama fonksiyonları üreten, regresyon veya sınıflandırma amaçlı kullanılabilen denetimli bir öğrenme yöntemidir [17]. Başka bir deyişle, eğitim verilerindeki herhangi bir noktadan en uzak iki sınıf arasında bir karar sınırı belirleyen vektör uzayı tabanlı bir makine öğrenme yöntemidir. Bu yöntem, N boyutlu bir uzayda optimal bir hiper düzlemi tahmin etmek için eğitim veri setini kullanarak ikili bir sınıflandırma tekniğini ifade eder [18].



Şekil 2.3 : Destek Vektör Makinaları Grafiksel İşleyişi [19]

Model, farklı sınıflardaki gözlemler arasındaki uzaklığı maksimize edecek şekilde bir

karar düzlemi bulmayı hedefler [20]. Bu yöntem, tahmine dayalı kontrol, ses-yüz tanıma, yazı tanıma, metin-görsel sınıflandırma gibi birçok alanda uygulanabilmektedir [18].

2.2.4. Naive Bayes

Elde edilen verilere dayanarak, önceden belirlenmiş sınıflara ait olma olasılıklarını baz alan bir algoritmadır. Bu algoritma, istatistiksel Bayes teorisine dayanarak bir rassal değişken için olasılık dağılımı içindeki koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir. Başka bir ifadeyle, hangi verinin hangi sınıfa ve hangi olasılıkla ait olduğunu tahmin etmeye yönelik bir algoritmadır. Bayes teoremine dayanan bu algoritma, değişkenler arasında koşullu bağımsızlık varsayımı altında, bir sonucun gözlenmesinde, birden fazla bağımsız değişkenin etkileşimini inceleyerek, sonucun ortaya çıkmasında hangi değişkenin daha etkin bir rol oynadığına dair ilişkiyi araştırmaktadır [21].

2.2.5. XGBoost

Chen ve Guestrin tarafından 2016 yılında geliştirilen XGBoost, bir gradyan artırma (gradient boosting) algoritmasıdır [22]. Bu algoritma, regresyon, sınıflandırma, sıralama ve kullanıcı tanımlı hata fonksiyonları gibi çeşitli görevleri destekler. XGBoost, Gradient boosting algoritmasının daha hızlı, daha esnek ve dağıtık sistemlerde çalışması mümkün olan bir uyarlaması olarak öne çıkar. Gradient boosting'e ek olarak, düzenlileştirme (regularization) ve paralel işleme gibi avantajlara sahiptir. Ancak, parametre sayısının fazla olması nedeniyle iyi bir parametre seçimi yapılması önemlidir; aksi halde aşırı öğrenme ya da az öğrenme sorunları ortaya çıkabilir [22].

XGBoost, sistem kullanımı ve algoritma için optimize edilmiş bir yapıya sahiptir. Artırma (Boosting), sıralı ilerleyen bir yapıya sahip olmasına rağmen, XGBoost paralelleştirme kullanarak ağaç yapısının yaprak ve düğümlerinin hesaplanmasını hızlandırır. XGBoost, ağaçların oluşturulmasında açgözlü (greedy) fonksiyonları kullanarak bölünme için durdurma kriterlerini belirler ve bölünme kısmında negatif kayıp (negative loss) kriteri ile değerlendirme yapar. Ağaçlar budanırken, maksimum derinlik parametresini kullanarak geriye doğru budama işlemini gerçekleştirir, bu da

hesaplama performansını artırır. Aşırı uyumu kontrol altına almak ve önlemek amacıyla, daha karmaşık modelleri cezalandırmak için L1 (Lasso) ve L2 (Ridge) hatalarını kullanır [22]. XGBoost, seyrek (sparse) verilerle uyumlu bir şekilde çalışabilir. Eğitim sırasında oluşan bilgi kaybını kullanarak otomatik olarak öğrenilen değerlerle eksik gözlemleri doldurur. Bu sayede veride az bulunan özellikleri doğal olarak kabul eder ve farklı seyreklik düzeylerini daha etkili bir şekilde işler.

2.2.6. LightGBM

Microsoft DMTK (Distributed Machine Learning Toolkit) projesi kapsamında 2017 yılında geliştirilen LightGBM, Ke ve arkadaşları tarafından ortaya çıkarılmış bir gradyan artırma algoritmasıdır [23]. Bu algoritma, regresyon, sınıflama ve sıralama problemlerini çözmek için kullanılır. Dağıtık sistemler ile çalışabilen, özel tanımlanmış hata fonksiyonunu destekleyen bir algoritmadır.

LightGBM, histogram tabanlı bir çalışma prensibine sahiptir. Bu prensip, sürekli değere sahip değişkenleri kategorik hale getirerek eğitim süresini kısaltmak ve kaynak kullanımını düşürmek amacını taşır. Ayrıca, LightGBM diğer çoğu ağaç tabanlı algoritmadan farklı olarak yaprak tabanlı bir yaklaşım benimser. Bu sayede ağaçlar yatayda büyüyerek aşırı öğrenmeyi engeller ve ağaç derinliği fazla artmadığı için bu alanda etkin bir performans sergiler.

2.3. Çapraz Doğrulama

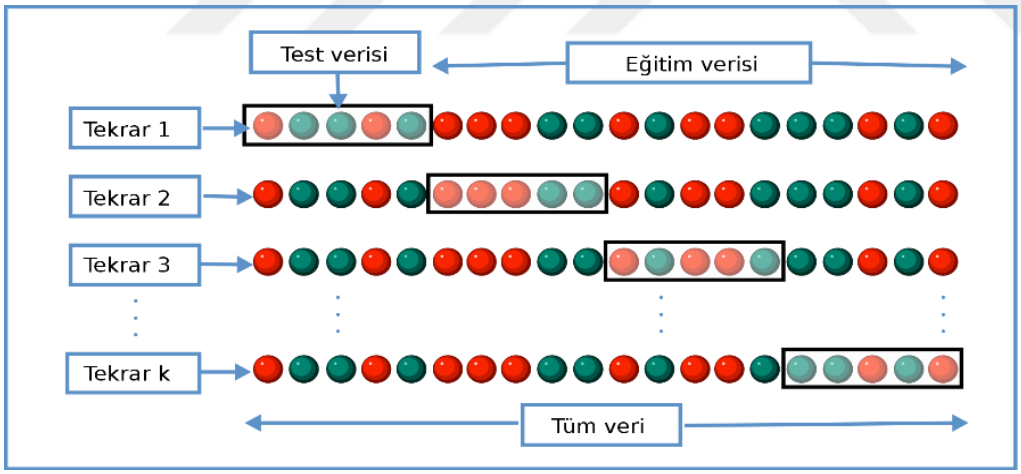
K katlı çapraz doğrulama yöntemi, model değerlendirmesi için son derece önemlidir ve modelin geliştirilme aşamasında aşırı öğrenme ve eksik öğrenmeyi tespit etme amacını taşır [24]. Aşırı öğrenme, modelin eğitim kümesindeki örüntüleri gerçek gözlemlerin yerine öğrenmesini ifade eder. Bu durumda, model eğitim aşamasında kullanılan veri kümesini öğrenir, ancak yeni gözlemler için başarılı tahminler yapamaz [25]. Genellikle aşırı öğrenme modelleri, eğitim aşamasını düşük hata oranlarıyla tamamlarken, test aşamasında yüksek hata oranlarıyla tahmin yapar [25]. Eksik öğrenme durumunda ise model, gözlemlerdeki örüntüyü eksik bir şekilde öğrenir. Bu durumda, model eğitim aşamasında kullanılan veri kümesini tam anlamıyla öğrenemez. Bu nedenle eksik öğrenme modelleri, eğitim ve test aşamalarını düşük hata

oranlarıyla tamamlar. Bu tür sorunları önlemek için sınıflandırma modellerinde özellikle aşırı öğrenmeyi ve eksik öğrenmeyi önlemek için k-katlı çapraz doğrulama, erken durdurma ve budama gibi birçok yöntem geliştirilmiştir. Bu yöntemler arasında *k* katlı çapraz doğrulama yöntemi diğerlerinden daha yaygın olarak kullanılmaktadır. *K* katlı çapraz doğrulama yöntemine göre, ilk olarak eğitim sürecinde kullanılacak eğitim kümesi karıştırılır ve ardından eşit büyüklükteki *k* alt kümelere bölünür. Bu işlemler *k* kez tekrarlanarak her iterasyonda sıradaki alt küme eğitim veri kümesinden çıkarılır ve test kümesi olarak kullanılır. Eğer *k* sayısı örnek büyüklüğüne eşitse, buna "birini dışarıda bırak" (leave-one-out) denir [26].

K katlı çapraz doğrulamada, modele gelen veriler kısımlara bölünür; her bir kısma parça adı verilir. Test veri kümesi için sınıflandırma başarısı (SB) ölçütü hesaplanmaktadır. Buna ait denklem, Eşitlik 2.3'te verilmiştir.

$$SB_i = \frac{Doğru\ sınıflandırılmış\ örnek\ sayısı}{Test\ veri\ kümesindeki\ örnek\ sayısı} * 100\ ve\ i = 1, 2, \dots, k \qquad (2.3)$$

k katlı çapraz doğrulama yöntemi Şekil 2.4'te ayrıntılı olarak gösterilmiştir.



Şekil 2.4 : *k* Katlı Çapraz Doğrulama [27]

Bu model, *k* değerinin 10 olduğu bir durumu ele alarak, test edilmek üzere 1. parçadaki verileri bir kenara ayırır ve kalan 9 parçayı eğitim amacıyla kullanır. Kısacası, bu 9 parça, modelin eğitimi için kullanılarak kalan 1 parça üzerinde eğitilen modeli test eder. Bu işlem, 10 kat çapraz geçerleme örneği için 10 defa tekrarlanır.

Her parça için modelin testi sırasında doğruluk istatistikleri değerlendirilir ve kullanılan ölçütler, değerlendirilen modelin türüne bağlı olarak belirlenir.

Sınıflandırma modellerinde genellikle sınıflandırma başarıları, karmaşıklık matrisi ve hata oranları gibi ölçütler kullanılır [28].

2.4. Değerlendirme Ölçütleri

2.4.1. Karmaşıklık Matrisi (Confusion Matrix)

Sınıflama problemlerinde model tahmin performansını değerlendirmek için kullanılan karmaşıklık matrisi, tahmin edilen hedef değişken ile gerçek hedef değişken değerlerini karşılaştırarak oluşturulur. İki sınıflı bir sorun için karmaşıklık matrisi, Çizelge 2.1'de gösterildiği şekildedir.

Çizelge 2.1 : Karmaşıklık Matrisi

		Gerçek Sınıf	
		Pozitif	Negatif
Tahmin Edilen Sınıf	Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
	Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Çizelge 2.1'de verilen;

DP, gerçekte pozitif olan ve pozitif olarak doğru sınıflandırılan,

DN, gerçekte negatif olan ve negatif olarak doğru sınıflandırılan,

YN gerçekte pozitif olan ve negatif olarak yanlış sınıflandırılan,

YP ise gerçekte negatif olan ve pozitif olarak yanlış sınıflandırılan gözlemleri temsil etmektedir.

Doğruluk, kesinlik, ROC eğrisi, duyarlılık, özgüllük ve F Ölçütü değerleri, karmaşıklık matrisi ile hesaplanabilir.

2.4.2. Doğruluk (Accuracy)

Doğruluk, pozitif ve negatif sonuçların toplamının toplam gözlem sayısına oranıdır. Dengesiz veri setlerinde, bu değerler yanıltıcı olabilir. Yüksek sayıda olan baskın sınıfa ait gözlemlerin yüksek oranda doğru sınıflandırılması, doğruluk oranını

artırabilir. Ancak bu durum, az olan sınıfın doğru sınıflandırıldığı bilgisini sağlamaz. Bu nedenle, doğruluk, dengesiz veri setlerinde yanıltıcı bir sonuç verebilir.

Doğruluk, Eşitlik 2.4'teki formülle doğru sınıflandırılan sonuçların toplam sonuçlara oranı olarak hesaplanır.

$$\boldsymbol{Doğruluk} = \frac{\boldsymbol{DP + DN}}{\boldsymbol{DP + DN + YP + YN}} \tag{2.4}$$

2.4.3. Kesinlik (Precision)

Kesinlik, gerçek değeri pozitif olan ve pozitif olarak sınıflandırılan gözlem sayısının, tahmin değeri pozitif olan tüm gözlemlerin toplamına oranı olarak Eşitlik 2.5'teki gibi hesaplanmaktadır.

$$\boldsymbol{Kesinlik} = \frac{\boldsymbol{DP}}{\boldsymbol{DP + YP}} \tag{2.5}$$

2.4.4. Duyarlılık (Sensitivity, Recall, True Positive Rate)

Doğru-pozitif (DP) oranı olarak da bilinen bu ölçüt pozitif olarak tahmin edilmesi gereken gözlemlerin ne kadarının pozitif yani doğru tahmin edildiğinin oranıdır ve Eşitlik 2.6 ile hesaplanır.

$$\boldsymbol{Hassaslık} = \frac{\boldsymbol{DP}}{\boldsymbol{DP + YN}} \tag{2.6}$$

2.4.5. Özgüllük (Specifity)

Özgüllük, gerçekte negatif sınıfa ait olan gözlemlerin sınıflarının model tarafından doğru tahmin edilme oranıdır ve Eşitlik 2.7'deki formül ile hesaplanır.

$$\boldsymbol{Özgüllük} = \frac{\boldsymbol{DN}}{\boldsymbol{DN + YP}} \tag{2.7}$$

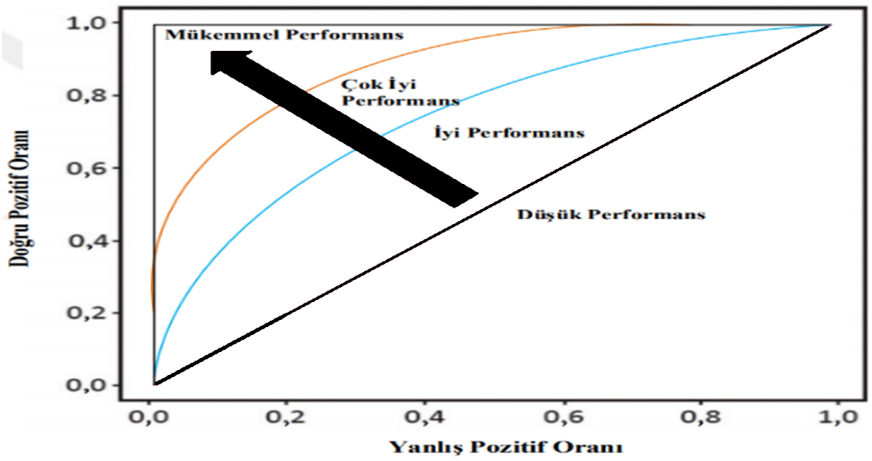
2.4.6. F Ölçütü

Kesinlik ve Hassaslık değerlerinin harmonik ortalamasıdır. Bazı durumlarda duyarlılığın artması kesinliğin düşmesine neden olabilir. Bu iki ölçütün dengede olması önemlidir ve denge F skoru ile ölçülür ve 1.0'a yakın olması beklenir. F ölçütü, Eşitlik 2.8'de belirtildiği şekilde hesaplanır.

$$F = 2 * \frac{Kesinlik * Hassaslık}{Kesinlik + Hassaslık} \tag{2.8}$$

2.4.7. ROC (Receiver Operating Characteristics) Eğrisi

ROC olarak adlandırılan Alıcı İşlem Karakteristiği eğrisi, farklı sınıflardaki tek bir gözlemin orta değerlerine dayanarak istatistiksel bilgi sunar. Ancak, değişkenler arasındaki orta değerler birbirine yakın olduğunda, ROC eğrisi testi geçen gözlemlerin özelliklerini belirlemede yeterince iyi bir ayırıcı olmayabilir. ROC eğrisi, çalışma şekli olarak karmaşıklık matrisinden faydalanır. Bu bağlamda, iki sınıflı bir problemde ROC eğrisi için X ekseni yanlış pozitif oranını (Yanlış Pozitif / (Yanlış Pozitif + Doğru Negatif)), Y ekseni ise doğru pozitif oranını (Doğru Pozitif / (Doğru Pozitif + Yanlış Negatif)) ifade eder. ROC eğrisinin altındaki alanı ifade eden AUC (Eğri Altındaki Alan) değeri, sınıflandırma yöntemlerinin başarısını değerlendirmek için kullanılır. AUC değerinin yüksek olması, istatistiksel olarak daha anlamlı bir sonuç elde edildiğine işaret eder. Şekil 2.5'te görülen üç farklı eğri arasında, mavi renkteki eğri en yüksek AUC değerine sahiptir, yeşil renkteki eğri ise en düşük AUC değerine sahiptir. Son yıllarda, AUC değeri, sınıflandırmanın başarısını kanıtlamak için birçok çalışmada kullanılmaktadır [29].



Şekil 2.5 : ROC Eğrisi [30]

2.5. Yazılım ve Kütüphaneler

Bu araştırmada, veri analizi ve modelleme işlemleri Python dilinde yazılmıştır. Veri içe aktarma ve ön işleme aşamalarında *pandas* kütüphanesi, sayısal hesaplamalar için *numpy*, veri görselleştirmede *matplotlib* kütüphanesi, model uydurma, seçme ve değerlendirmede ise *scikit-learn* kütüphanesi kullanılmıştır.

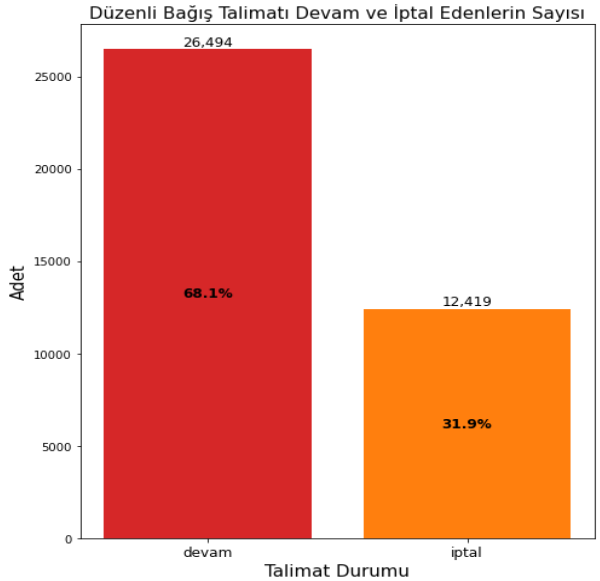
3. VERİ SETİ VE UYGULAMA

3.1. Veri Seti Açıklaması

Çalışmada kullanılan veri seti, Türkiye'de faaliyet gösteren bir STK'ya ait olup talepleri doğrultusunda ismi tez kapsamında verilmeyecektir. Veri seti içerisinde STK'ya ait düzenli bağış talimatı oluşturmuş bağışçılara ait bağış bilgileri ve demografik bilgiler yer almaktadır. Veri seti içerisindeki bağış bilgileri genel olarak bağış tarihleri, bağış kategorilerine ilişkin bağış adedi ve bağış tutarı şeklindedir. Düzenli bağışlara ait bulunan bilgilerle birlikte, bağışçıların düzenli bağış talimatları dışında tek seferlik bağışları da yer almaktadır. Veri ön işleme sürecinin ardından modellemede kullanılan veri seti içerisinde yer alan değişken listesi EK-1'de belirtilmiştir. Veri seti doğrulama, eğitim ve test verisi olarak üçe ayrılmıştır. Tezin amacı doğrultusunda tahmin edilmesi istenen hedef değişken talimat_durumu isimli değişken olup talimat_durumu bağışçıların STK için daha önce oluşturdukları düzenli bağış talimatlarına devam edip etmediklerini göstermektedir.

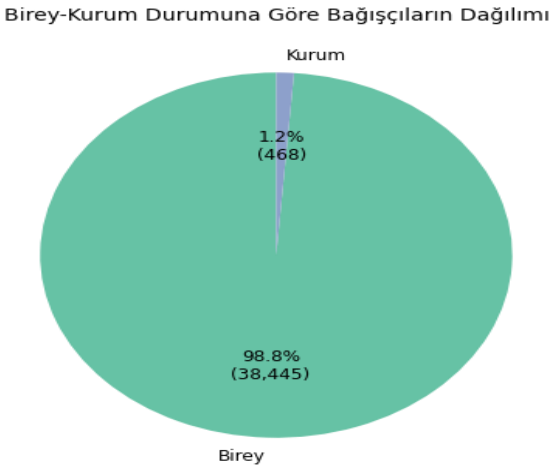
3.2. Açıklayıcı Veri Analizi

Veri seti 2010-2024 yılları arasında yapılan bağışları kapsamakta olup düzenli bağış talimatı veren toplam 38.913 bağışçı bulunmaktadır. Düzenli bağış talimat durumunun dağılımını gösteren Şekil 3.1'de de görüldüğü üzere talimatı devam edenlerin sayısı 26.494 (%68,1), talimatını iptal edenlerin sayısı ise 12.419 (%31,9)'dur.



Şekil 3.1 : Düzenli Bağış Talimat Durumlarının Dağılımı

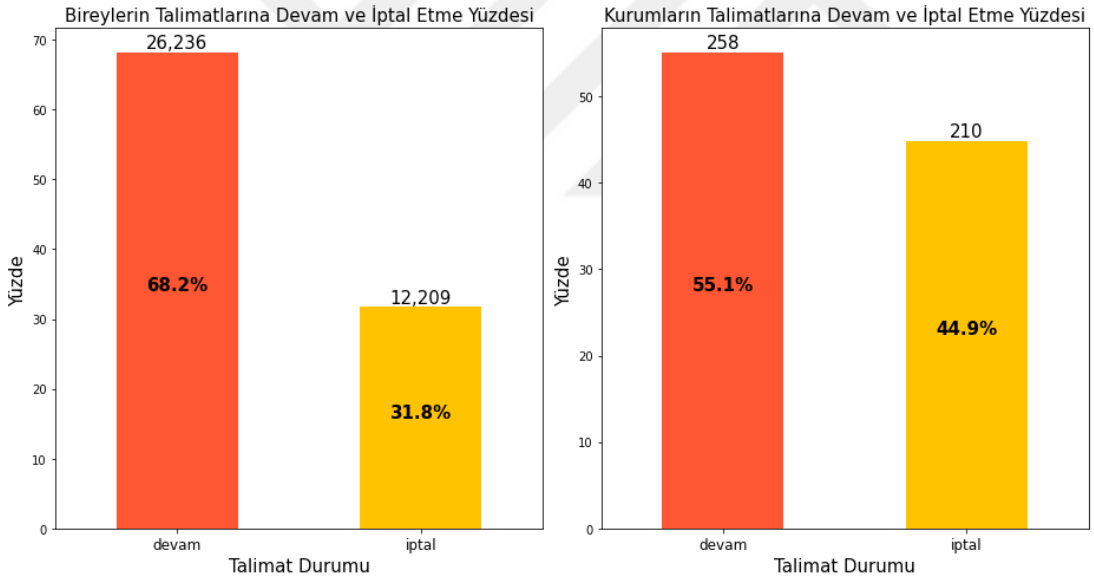
Bağışçıların birey - kurum olma durumlarına göre bağışçı sayılarının dağılımı;



38.913 bağışçının 38.445'i birey, 468'i kurumdur. Yüzdesel olarak bakıldığında düzenli bağış talimatı oluşturan bağışçıların %98,8'i birey, %1,2'si kurumdur.

Şekil 3.2 : Bağışçı Sayısının Birey-Kurum Durumuna Göre Dağılımı

Bağışçıların birey - kurum olmalarına göre düzenli bağış talimatına devam ve iptal etme yüzdesi;

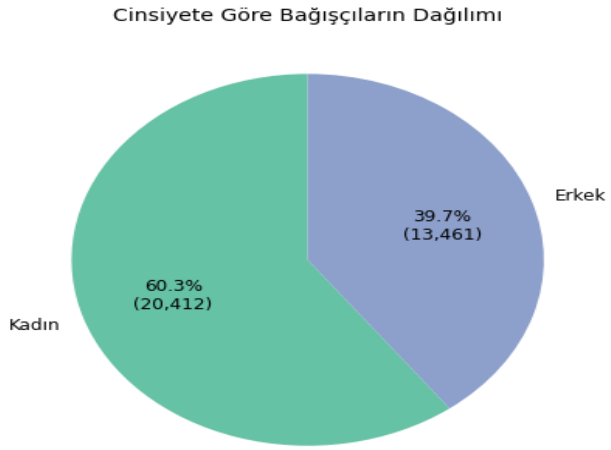


Şekil 3.3 : Düzenli Bağış Talimat Durumlarının Birey - Kurum Durumuna Göre Dağılımı

Bağışçıların birey ve kurum olma durumlarına göre düzenli bağış talimatlarına devam edip etmeme oranlarına bakıldığında; Bireylerde iptal oranı %31,8 iken kurumlarda bu oran %44,9'dur.

Cinsiyete göre bağışçı sayılarının dağılımı;

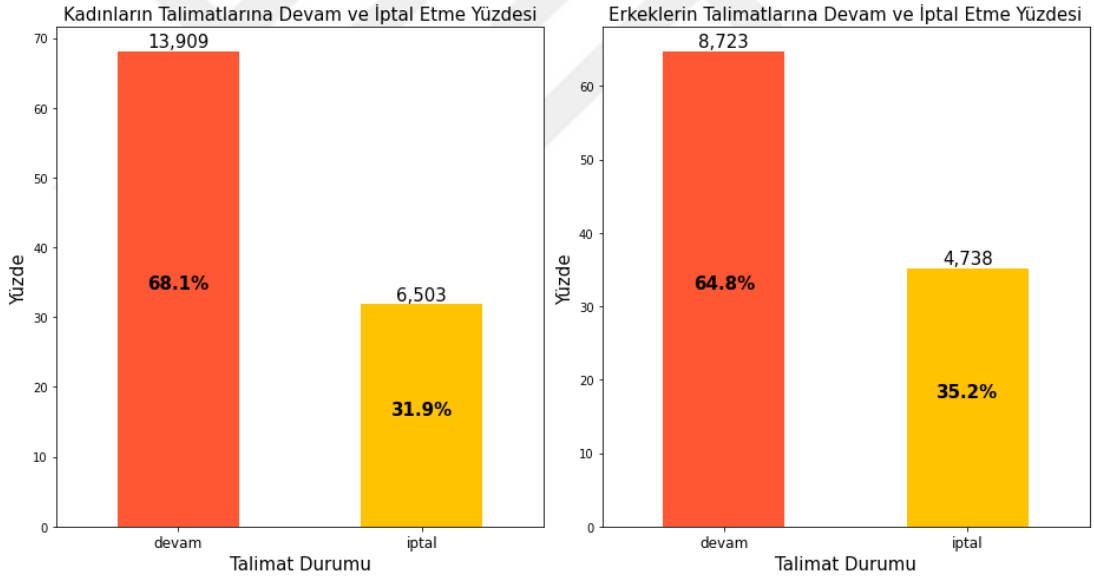
Veri seti içinde 38.913 bağışçı arasında cinsiyet bilgisi bulunanların sayısı 33.873'tür.



Şekil 3.4'te de görüldüğü üzere 33.873 bağışçının 20.412 (%60,3)'si kadın, 13.461 (%39,7)'i erkektir.

Şekil 3.4 : Bağışçı Sayısının Cinsiyete Göre Dağılımı

Cinsiyete göre düzenli bağış talimatına devam ve iptal etme yüzdesi;



Şekil 3.5 : Düzenli Bağış Talimat Durumlarının Cinsiyete Göre Dağılımı

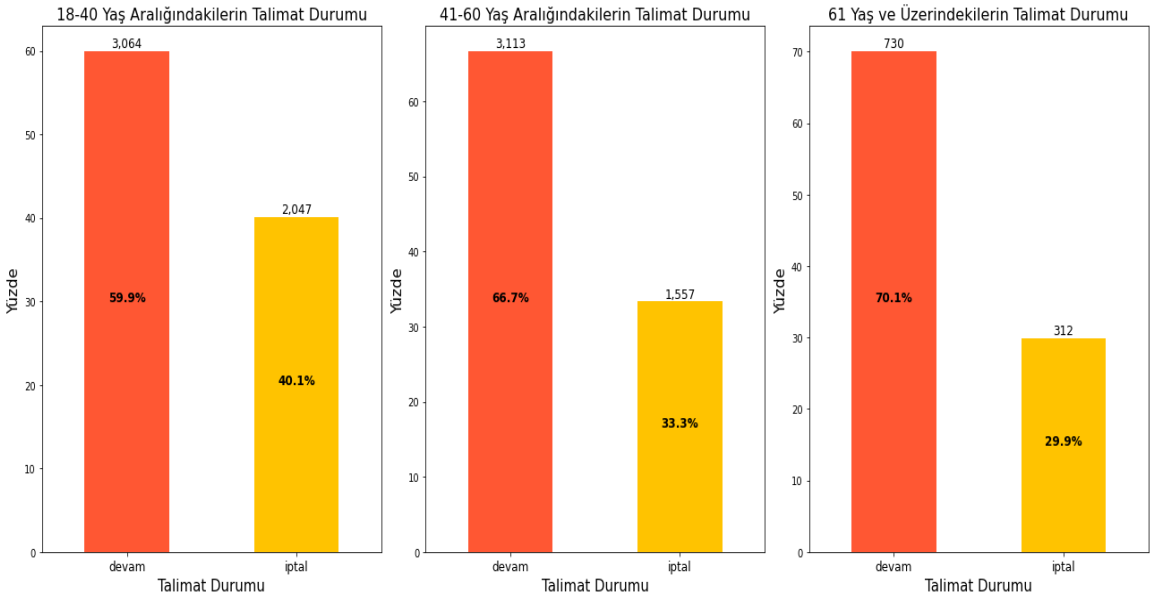
Şekil 3.5'te verilen cinsiyete göre düzenli bağış talimatlarına devam edip etmeme oranlarına bakıldığında ise; kadınlarda iptal oranı %31,9 iken erkeklerde bu oran %35,2'dir.

Yaş aralıklarına göre bağışçı sayılarının dağılımı;

Veri setinde yer alan 38.913 bağışçı içinde yaş bilgisi bulunanların olanların sayısı oranı yaklaşık %27,8 (10.823)'dir. 10.823'tür. Yaş bilgisi olan bağışçıların yaş aralıkları, 18-40, 41-60, 60 yaş ve üzeri şeklinde gruplandırılmıştır. Bu aralıklara ilişkin dağılım Şekil 3.6'da belirtilmiştir.

Şekil 3.6 : Bağışçıların Yaş Aralıklarına Göre Dağılımı

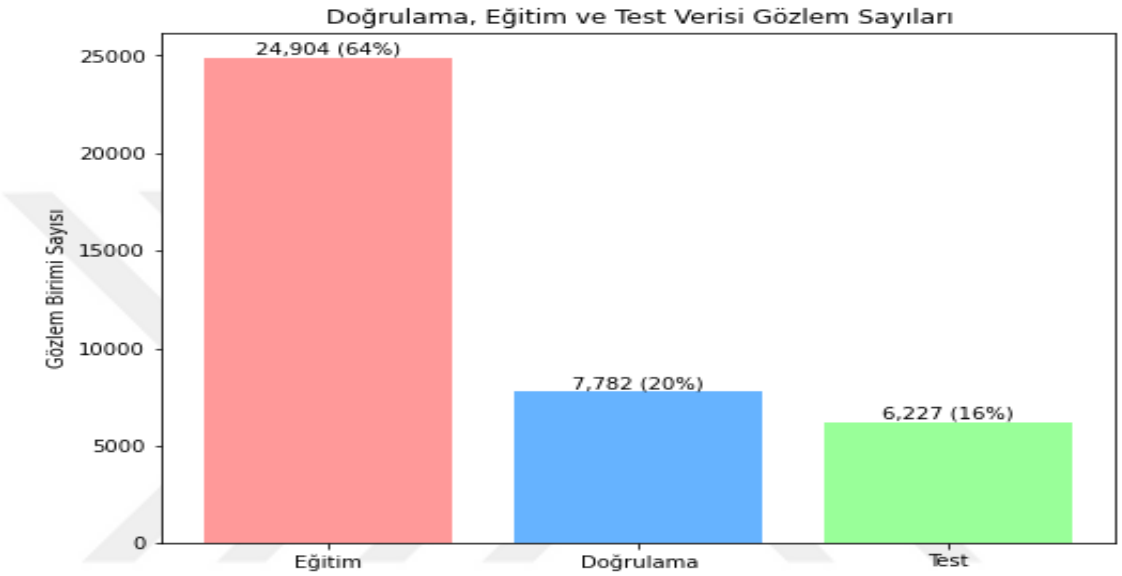
Yaş aralıklarına göre düzenli bağış talimatlarına devam edip etmeme incelendiğinde ise 18-40 yaş aralığındaki bağışçıların düzenli bağış talimatlarını iptal etme oranlarının daha fazla olduğu görülmektedir (Şekil 3.7).



Şekil 3.7 : Düzenli Bağış Talimat Durumlarının Yaş Aralıklarına Göre Dağılımı

3.3. Veri Ön İşleme ve Modelleme

Tez çalışmasında kullanılan veri seti 38.913 gözlem biriminden ve hedef değişkeni de dahil olmak üzere 60 özellikten oluşmaktadır. Veri seti doğrulama, eğitim ve test verisi olmak üzere 3'e bölünmüştür. %80-20 kuralı baz alınarak veri setinin %80'i eğitim verisi, %20'si ise parametre optimizasyonu için doğrulama verisi olarak ayrılmıştır. %80 olarak ayrılmış olan eğitim verisinin %20'si ise test verisi olarak ayrılmış, dağılım Şekil 3.8'de grafikte gösterilmiştir.



Şekil 3.8 : Doğrulama, Eğitim ve Test Verisi Gözlem Sayıları

3.3.1. Özellik Dönüşümü ve Veri Normalizasyonu

Hazırlanan çalışmada var olan değişkenler üzerinden yeni değişkenler oluşturulmuştur. Oluşturulan değişkenler;

- Bağış tutarı üzerinden bağışçıların yıllık bazda yaptıkları toplam bağışları gösteren değişkenler eklenmiştir. Bu değişkenler; 2024_yili_bagis_tutari, 2023_yili_bagis_tutari, 2022_yili_bagis_tutari, 2021_yili_bagis_tutari, 2020_yili_bagis_tutari şeklindedir.
- Bağış türleri üzerinden de bağışçıların bağış türü bazında yaptıkları toplam bağış tutarları ve adetlerini gösteren değişkenler oluşturulmuştur.
- Bağış tarihleri üzerinden de değişkenler oluşturulmuştur. Oluşturulan değişkenler, ilk bağış tarihi ve son bağış tarihinden bugüne kadar geçen gün, ay, yıl sayısını gösteren değişkenlerdir.

- Bağışçılar bireysel ve kurumsal bağışçılar olmak üzere ikiye ayrılmaktadır. bagisci_kayit_tipi özelliği üzerinden bagisci_kayit_tipi_Birey ve bagisci_kayit_tipi_Kurum şeklinde iki farklı kukla (dummy) değişken oluşturulmuştur.
- Hedef özellik durumundaki talimat_durumu özelliği içindeki "devam" ve "iptal" ifadeleri ikili olarak kodlanmıştır. Bunun sebebi lojistik regresyon, DVM ve naive bayes algoritmalarının doğrudan kategorik verilerle çalışamaması ve sayısal girdi gerektirmesidir. 0, devam 1 ise iptal'i temsil etmektedir.

Veri seti için Min-Max Ölçeklendirme Yöntemi ile modelde kullanılmadan önce normalizasyon işlemi yapılmıştır.

3.3.2. Hiperparametre Ayarlaması

Modellerin performansı genellikle model parametrelerinin (hiperparametrelerin) uygun bir şekilde ayarlanmasına bağlıdır. Hiperparametreler, bir makine öğrenimi modelinin yapılandırılmasını ve eğitilmesini etkileyen ayarlanabilir parametrelerdir. Bu parametreler, modelin karmaşıklığını, genelleme yeteneğini ve eğitim süresini etkilemektedir.

Yapılan çalışmada modelleme işleminden önce en iyi performansı sağlayacak parametreleri belirlemek için hiperparametre ayarlaması gerçekleştirilmiştir. Bu ayarlama sürecinde Grid Search yöntemi kullanılmış ve 5 kat çapraz doğrulama ile birleştirilmiştir. Grid Search ile farklı hiperparametre kombinasyonlarının denenmesi sağlanarak en iyi performansı sağlayanları belirlenmiştir.

Doğrulama verisi üzerinden gerçekleştirilen 5 katlı çapraz geçerleme performansını ölçmek için doğruluk skoru kullanılmıştır. Tez kapsamında kullanılan her bir model için seçilen parametreler Çizelge 3.1'de listelenmiştir.

Çizelge 3.1 : Algoritmaların Hiperparametre Değerleri

Algoritma	Hiperparametre	Hiperparametre Açıklaması	Seçilen Hiperparametre
Rastgele Orman	**n_estimators**	Karar ağaçlarının sayısı	1000
	max_depth	Karar ağaçlarının maksimum derinliği	10
	min_samples_split	Bir düğümün ikiye bölünmeden önce en az kaç örnek içermesi gerektiği	2
	max_features	Her bölünmenin rastgele seçilecek olan özelliklerin maksimum sayısı	30
Lojistik Regresyon	**penalty**	Ceza türü	l1
	C	Düzenlileştirme parametresi	10000
	solver	Optimizasyon algoritması	liblinear
	max_iter	Maksimum iterasyon sayısı	100
Destek Vektör Makineleri	**C**	Düzenlileştirme parametresi	100
	gamma	Rbf', 'poli' ve 'sigmoid' için çekirdek katsayısı	1
	kernel	Çekirdek türü	poly
Naive Bayes	**var_smoothing**	Varyansı sıfıra yaklaşan özniteliklerin hesaplanması sırasında düzeltilmiş bir süreklilik düzeltmesidir.	0.035
XGBoost	**max_depth**	Bir ağacın maksimum derinliği	5
	learning_rate	Öğrenme oranı	0.1
	n_estimators	Ağaç sayısı	200
	min_child_weight	Çocuk düğümde gereken minimum örnek ağırlığının toplamı (hessian)	1
	subsample	Eğitim örneklerinin alt örnek oranı	1.0
	colsample_bytree	Sütunların alt örneklemesine yönelik bir parametre ailesidir	0.5
	gamma	Ağacın yaprak düğümünde daha fazla bölüm oluşturmak için gereken minimum kayıp azaltımı	0
LightGBM	**num_leaves**	Bir ağaçtaki maksimum yaprak sayısı	30
	max_depth	Bir ağacın maksimum derinliği	10
	learning_rate	Öğrenme oranı	0.1
	n_estimators	Ağaç sayısı	200
	boosting_type	Boosting algoritması	dart

3.3.3. Özellik Seçimi ve Model Performanslarının Değerlendirilmesi

Tez çalışmasında kullanılan makine öğrenimi modelleri rastgele orman, lojistik regresyon, destek vektör makineleri, naive bayes, XGBoost ve LightGBM'dir. Modellerin performansının değerlendirilmesi için doğruluk, kesinlik, duyarlılık, özgüllük, F ölçütü ve ROCAUC ölçütleri kullanılmıştır.

Belirtilen algoritmalara ait özellik seçimi yapılmadan önce değerlendirme ölçütleri hesaplandığında Çizelge 3.2'deki değerler elde edilmiştir.

Çizelge 3.2 : Modelleme Sonuçları - Özellik Seçimi Yapılmadan Önce

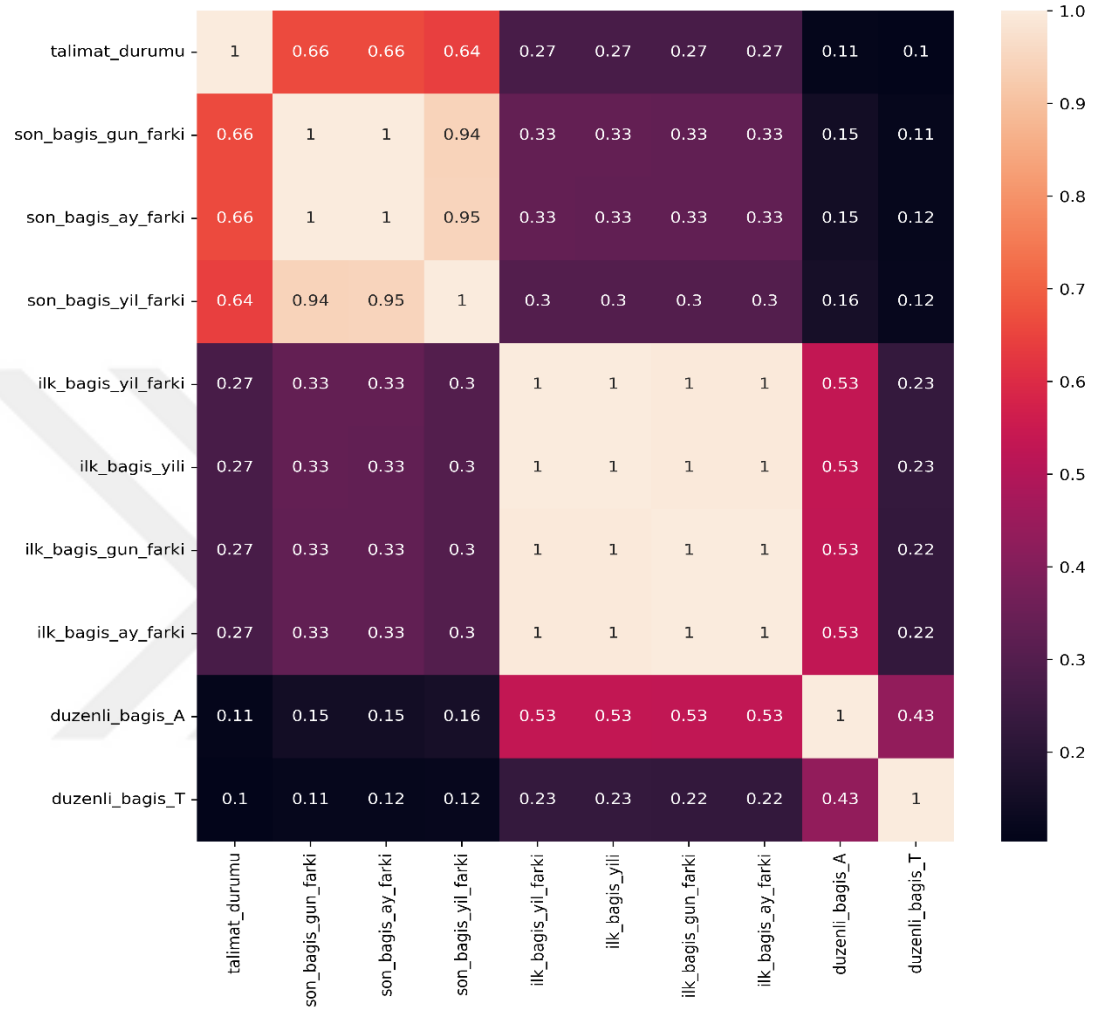
	Doğruluk	Kesinlik	Duyarlılık	Özgüllük	F Ölçütü	ROCAUC
Rastgele Orman	0,93	0,85	0,94	0,92	0,89	0,98
Lojistik Regresyon	0,89	0,90	0,74	0,96	0,81	0,96
Destek Vektör Makineleri	0,89	0,89	0,74	0,96	0,81	0,96
Naive Bayes	0,85	0,89	0,62	0,96	0,73	0,90
XGBoost	0,93	0,86	0,94	0,92	0,90	0,98
LightGBM	0,93	0,85	0,95	0,92	0,90	0,98

Çizelge 3.2'de belirtildiği gibi modeller için hesaplanan ölçüt değerleri yüksek değerlere sahiptir. Bu durum, modellerin eğitim verilerine aşırı öğrenme sağlamış olabilir. Modeller için elde edilen yüksek ölçüt değerlerinin aşırı öğrenme sonucunda olup olmadığının değerlendirilmesi gerekmektedir.

Bu sebeple, aşırı öğrenme durumunu değerlendirmek için ilk olarak korelasyon analizi yapılmıştır. Korelasyon analizi, bağımsız değişkenler ile hedef değişken arasındaki ilişkileri inceleyerek yüksek derecede ilişkili özelliklerin belirlenmesini sağlamıştır. Ardından, rastgele orman modeli kullanılarak değişkenlerin önem düzeyleri ile özellik seçimi gerçekleştirilmiş ve modellerin performansı yeniden değerlendirilmiştir. Bu adım, modellerin karmaşıklığını azaltmayı ve yalnızca önemli özelliklere odaklanarak daha güvenilir tahminler elde edilmesini amaçlamaktadır.

Bu bölümde, aşırı öğrenme durumunu azaltmak için yapılacak özellik seçimi işlemlerinin, model performansını nasıl etkilediği daha ayrıntılı bir şekilde değerlendirilmiştir. Yapılan analiz, bağışçı davranışlarını daha doğru bir şekilde tahmin etme yeteneğimizi artırarak, daha etkili stratejiler geliştirmemize olanak sağlayacaktır.

Özelliklerin hedef özellikle olan ilişkilerini daha detaylı incelemek amacıyla korelasyon analizi gerçekleştirilecektir. Bu analiz, seçilen özelliklerin hedef özellikle olan ilişkilerini görsel olarak göstererek, modelin açıklanabilirliğini artırmaya ve sonuçların daha kapsamlı bir şekilde değerlendirilmesine katkı sağlayacaktır.



Şekil 3.9 : En Yüksek İlişkiye Sahip İlk 10 Değişkenin Heatmap İle Gösterimi

Hedef özellik olan talimat_durumu ile en yüksek ilişkiye sahip durumdaki 10 sütun Şekil 3.9'da belirtilmiştir.

Yüksek Korelasyonlu Özelliklerin Kaldırılması;

Şekil 3.9'daki Heatmap grafiğinde, hedef özellik ile korelasyon değeri 0.7'den büyük ve -0.7'den küçük olan özellikler tespit edilmiştir. Bu durum, yüksek korelasyonlu özelliklerin veri setinde fazladan ve gereksiz bilgi taşıyabileceğini işaret etmektedir. Yüksek korelasyonlu özelliklerin kaldırılmasının sebebi, hem modelin karmaşıklığının azaltılmak istenmesi hem de gereksiz özelliklerin modele olan etkisinin azaltılmasıyla daha iyi bir genelleme yeteneğinin sağlanmasının hedeflenmesidir.

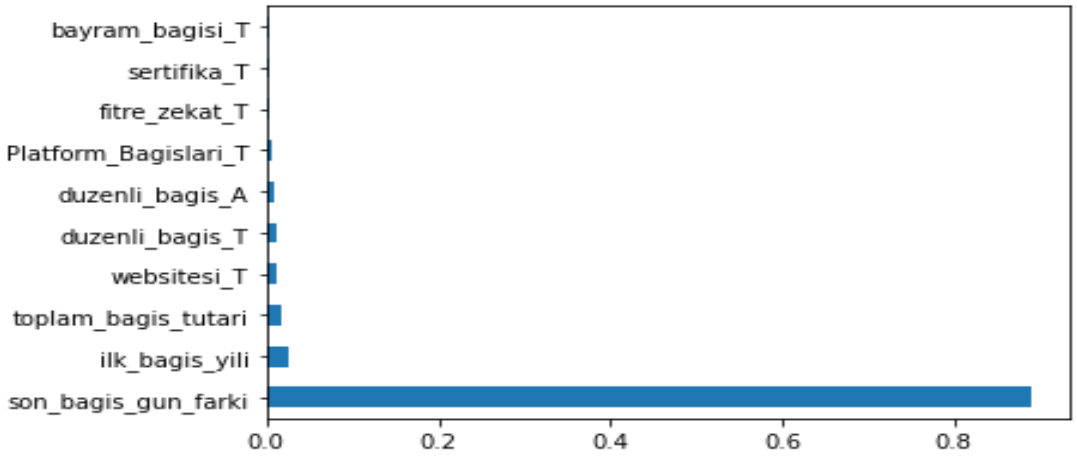
Bu nedenle, yapılan uygulamada yüksek korelasyonlu özellikler veri setinden kaldırılmıştır. Yüksek korelasyona sahip durumda 25 özellik tespit edilmiştir ve bu özellikler veri setinden çıkarılmıştır.

Veri setinden çıkarılan özellikler; 2020_yili_bagis_tutari, 2021_yili_bagis_tutari, 2022_yili_bagis_tutari, 2023_yili_bagis_tutari, 2024_yili_bagis_tutari, Platform_Bagislari_A', ayni_bagis_T, bagis_adedi, bagis_hediyeleri_T, bagisci_kayit_tipi_Kurum', bayram_bagisi_tutar_A, etkinlik_T, genel_bagis_A, genel_bagis_T', ilk_bagis_ay_farki, ilk_bagis_gun_farki, ilk_bagis_yil_farki, ilk_son_bagis_ay_farki, internet_bankaciligi_A, internet_bankaciligi_T, maraton_T, son_bagis_ay_farki, son_bagis_yil_farki, veli_bagisi_toplam_miktari, websitesi_A

Korelasyon analizinin sonrasında rastgele orman algoritması ile özellik önem düzeylerinin hesaplanmasıyla özellik seçimi gerçekleştirilmiştir. Model için uygun parametreler, hiperparametre optimizasyon yöntemlerinden birisi olan çapraz doğrulama yöntemi ile 5 kat uygulanarak belirlenmiştir. Çapraz doğrulama sonuçlarına göre en iyi performansı sağlayan hiperparametre kombinasyonu belirlenmiş ve seçilen en iyi hiperparametrelerle nihai model oluşturulmuştur.

Özellik Önem Sıralamasının Belirlenmesi;

Eğitilmiş rastgele orman modeli ile özelliklerin önem düzeyleri Şekil 3.10'da gösterildiği gibi elde edilmiştir. Böylece, en yüksek önem düzeyine sahip ilk 10 özelliğin modelin performansına olan katkısı görülebilmektedir.



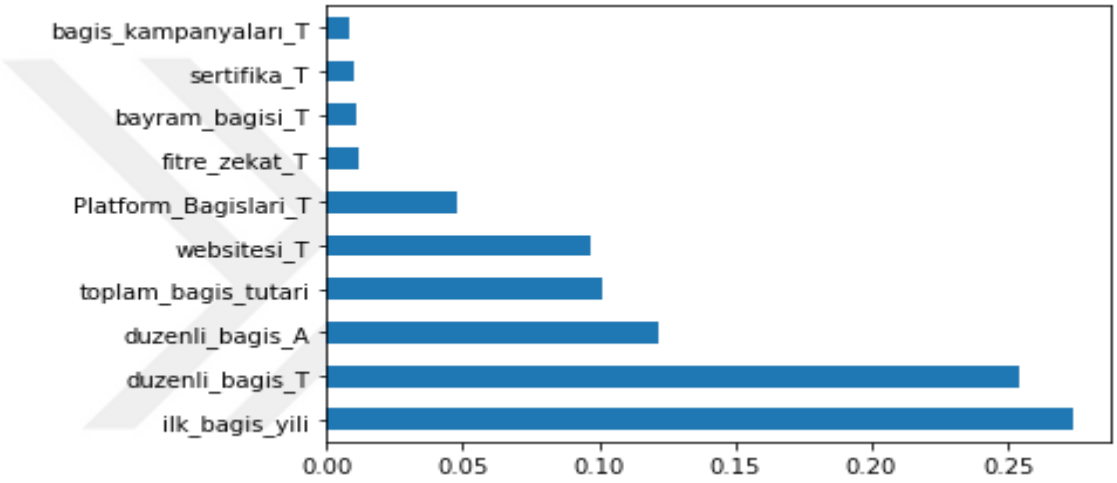
Şekil 3.10 : Rastgele Orman Modeline Göre Özellik Önem Düzeyleri

Özellik önem sıralaması yapılırken, özelliklerden bazıları belirgin bir biçimde daha yüksek önem düzeyine sahip olabilir. Bu durum, önem düzeyi yüksek olan

değişkenlerden kaynaklı olarak modelin eğitim verilerine aşırı uyum sağladığı (aşırı öğrenme/overfitting) anlamına gelebilir.

Bu bağlamda, Şekil 3.10'daki özellik önem sıralamasında ilk özellik olan son_bagis_gun_farki özelliğinin diğerlerinden belirgin şekilde daha yüksek bir değer aldığı gözlemlenmiştir. Bu aşamada, modelin genelleme yeteneğini artırmak ve olası aşırı öğrenmeyi araştırmak ve önlemek amacıyla, son_bagis_gun_farki değişkeni çıkarılarak önem sıralaması yeniden hesaplanmıştır.

Geri kalan özellikler için özellik önem sıralamasına bakıldığında Şekil 3.11 elde edilmiştir;



Şekil 3.11 : Rastgele Orman Modeline Göre Değişken Önem Düzeylerinin Son Durumu

Yapılan özellik seçimi işlemleri kapsamında korelasyon analizi sonucu 25 özellik, sonrasında rastgele orman ile özellik çıkarımı işlemi yapılarak 1 özellik daha veri setinden çıkarılmıştır. Veri setinde kalan özellik sayısı, hedef değişken de dahil olmak üzere 60 iken 26 özellik çıkarılması ile 34 olmuştur. 34 özellik EK-1'de belirtilmiştir.

Özellik çıkarımı sonrasında veri seti üzerinde modellerin performansları incelendiğinde Çizelge 3.3'te belirtilen sonuçlar elde edilmiştir.

Çizelge 3.3 : Modelleme Sonuçları - Özellik Seçimi Yapıldıktan Sonra

	Doğruluk	Kesinlik	Duyarlılık	Özgüllük	F Ölçütü	ROCAUC
Rastgele Orman	0,82	0,78	0,61	0,92	0,69	0,87
Lojistik Regresyon	0,78	0,74	0,46	0,92	0,57	0,81
Destek Vektör Makineleri	0,79	0,72	0,55	0,90	0,63	0,83
Naive Bayes	0,78	0,66	0,64	0,84	0,65	0,77
XGBoost	0,82	0,79	0,61	0,92	0,69	0,87
LightGBM	0,82	0,80	0,60	0,93	0,69	0,87

Özellik seçimi yapıldıktan sonra elde edilen sonuçlarda, duyarlılık ve özgüllük arasında belirgin bir dengesizlik olduğu gözlemlenmiştir. Bu dengesizlik, modelin duyarlılığının istenen seviyede olmaması durumunda, düzenli bağış talimatlarına devam eden bağışçıların iptal etme eğilimlerini belirlemede yetersiz kalınabileceğini göstermektedir. Bu nedenle, farklı eşik değerleri kullanarak model değerlendirilmesi yeniden yapılacaktır.

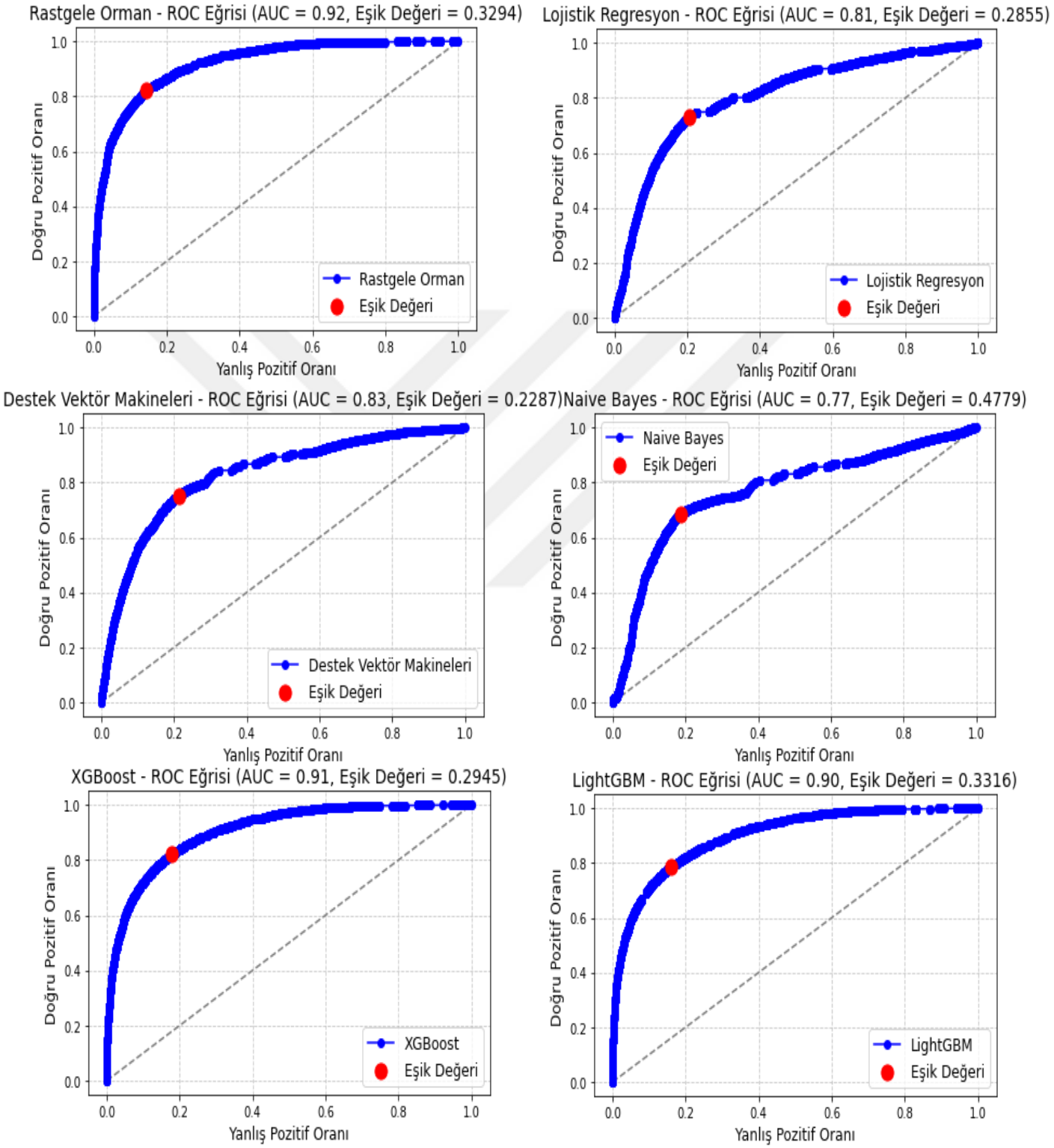
Yeni eşik değerlerinin kullanarak yapılan bu değerlendirme, modellerin duyarlılık ve özgüllük ölçütleri arasındaki dengeyi optimize etme amacını taşımaktadır. Bu işlem, bağışçıların düzenli bağış talimatlarını iptal etme eğilimlerinin daha hassas bir şekilde belirlenmesini sağlayacaktır. Bu dengenin sağlanmasıyla birlikte, çalışmanın amacına daha uygun bir model elde edilmesi amaçlanmaktadır.

3.3.4. Eşik Değerlerin Tespit Edilmesi ve Model Başarılarına Etkisi

Uygulama kapsamında kullanılan sınıflandırma modellerinin çıktılarını optimize etmek için en iyi eşik değerleri ROC analizi ile belirlenmiştir. Bu işlem, modellerin doğruluğunu, duyarlılığını ve özgüllüğünü iyileştirmeye yöneliktir.

Eğitim verisi üzerinden elde edilen sınıflama tahmin değerleri üzerinden elde edilen sınıf olasılıkları ve gerçek sınıf değerleri kullanılarak ROC eğrisinin farklı eşikler için modelin doğruluğunu temsil eden TPR ve FPR değerleri hesaplanmıştır. Modelin denge ve performansını değerlendirmek üzere, TPR ve FPR değerleri için hesaplanan en yüksek geometrik ortalamaya ilişkin değer en iyi eşik değeri olarak belirlenmiştir.

Bu değer, modelin hem duyarlılığını hem de özgüllüğünü optimize eder. Her bir model için ayrı yapılan ROC eğrisi analizi sonucunda hesaplanan AUC ve en iyi eşik değerleri, Şekil 3.12'deki grafiklerde gösterilmiştir.



Şekil 3.12 : Algoritmalara Ait Eşik Değerlerinin ROC Eğrisi ile Tespiti

Eşik değerlerine göre ölçütlere tekrar bakıldığında;

Ölçütler, veri setine herhangi bir özellik seçimi uygulanmamış hali, korelasyon analizi ve rastgele orman modeline göre çıkarılan özellikler sonrası olmak üzere 2 farklı durum özelinde incelenmiştir. Yapılan inceleme test verisi üzerinde gerçekleştirilmiştir.

Belirtilen algoritmalara ait özellik seçimi yapılmadan değerlendirme ölçütleri hesaplandığında Çizelge 3.3'teki değerler elde edilmiştir.

Çizelge 3.4 : Eşik Değerleri Uyguladıktan Sonraki Modelleme Sonuçları - Özellik Seçimi Yapılmadan

	Doğruluk	Kesinlik	Duyarlılık	Özgüllük	F Ölçütü	ROCAUC
Rastgele Orman	0,92	0,83	0,97	0,90	0,89	0,98
Lojistik Regresyon	0,91	0,85	0,86	0,93	0,86	0,96
Destek Vektör Makineleri	0,91	0,88	0,83	0,95	0,86	0,96
Naive Bayes	0,85	0,88	0,62	0,96	0,73	0,90
XGBoost	0,93	0,83	0,96	0,91	0,89	0,98
LightGBM	0,93	0,83	0,97	0,91	0,89	0,98

Çizelge 3.4'te eşik değerleri belirlemeden önceki duruma benzer şekilde, genellikle tüm modeller için hesaplanan ölçütlere bakıldığında yüksek performanslar gözlemlenmiştir. Bu sebeple aşırı öğrenme riskini gidermek adına korelasyon analizi ve rastgele orman modeline göre çıkarılan özellikler sonrası değerlendirme ölçütleri hesaplanmış ve Çizelge 3.5'teki değerler elde edilmiştir.

Çizelge 3.5 : Eşik Değerleri Uyguladıktan Sonraki Modelleme Sonuçları - Özellik Seçimi Yapılarak

	Doğruluk	Kesinlik	Duyarlılık	Özgüllük	F Ölçütü	ROCAUC
Rastgele Orman	0,79	0,66	0,75	0,81	0,70	0,87
Lojistik Regresyon	0,77	0,62	0,73	0,79	0,67	0,81
Destek Vektör Makineleri	0,77	0,62	0,75	0,78	0,68	0,83
Naive Bayes	0,77	0,63	0,68	0,81	0,66	0,77
XGBoost	0,80	0,67	0,75	0,83	0,71	0,87
LightGBM	0,80	0,67	0,74	0,83	0,71	0,87

Bağışçıların düzenli bağış talimatlarını iptal etme durumunu tahmin etmek için yaptığımız analizde, duyarlılık ölçütü üzerinden modellerin performansı değerlendirilmiştir. Duyarlılık, gerçekte pozitif olan durumların ne kadarının doğru bir şekilde tanımlandığını gösterir ve bağışçıların düzenli bağış talimatlarını iptal etme durumlarını doğru bir şekilde tespit edebilme başarısıyla ilgilidir.

Çizelge 3.5'teki sonuçlara göre, performans açısından en iyi modeller XGBoost, rastgele orman ve destek vektör makineleri olarak belirlenmiştir. Bu modeller, pozitif sınıfı önemli bir ölçüde doğru bir şekilde sınıflandırabilmektedir. Ancak, diğer ölçütler de dikkate alındığında, XGBoost modelinin genel performansının diğer modellere göre daha yüksek olduğu gözlemlenmektedir.

Değerlendirme ölçütlerine göre model performanslarına bakıldığında, duyarlılık ölçütü öncelik alınarak değerlendirme yapılmış olup XGBoost, rastgele orman ve destek vektör makineleri en başarılı üç model olarak belirlenmiştir. Belirlenen bu 3 modelin performanslarını daha ayrıntılı bir şekilde incelemek amacıyla karmaşıklık matrisleri de incelenmiştir. Karmaşıklık matrisleri, bir modelin gerçek ve tahmin edilen sınıfları ne kadar doğru bir şekilde sınıflandırdığını gösterir. Bu matrisler, modelin ne kadar doğru sınıflandırdığını, hangi hataları yaptığını ve hangi sınıfların daha iyi veya daha kötü tahmin edildiğinin belirlenmesini sağlamaktadır.

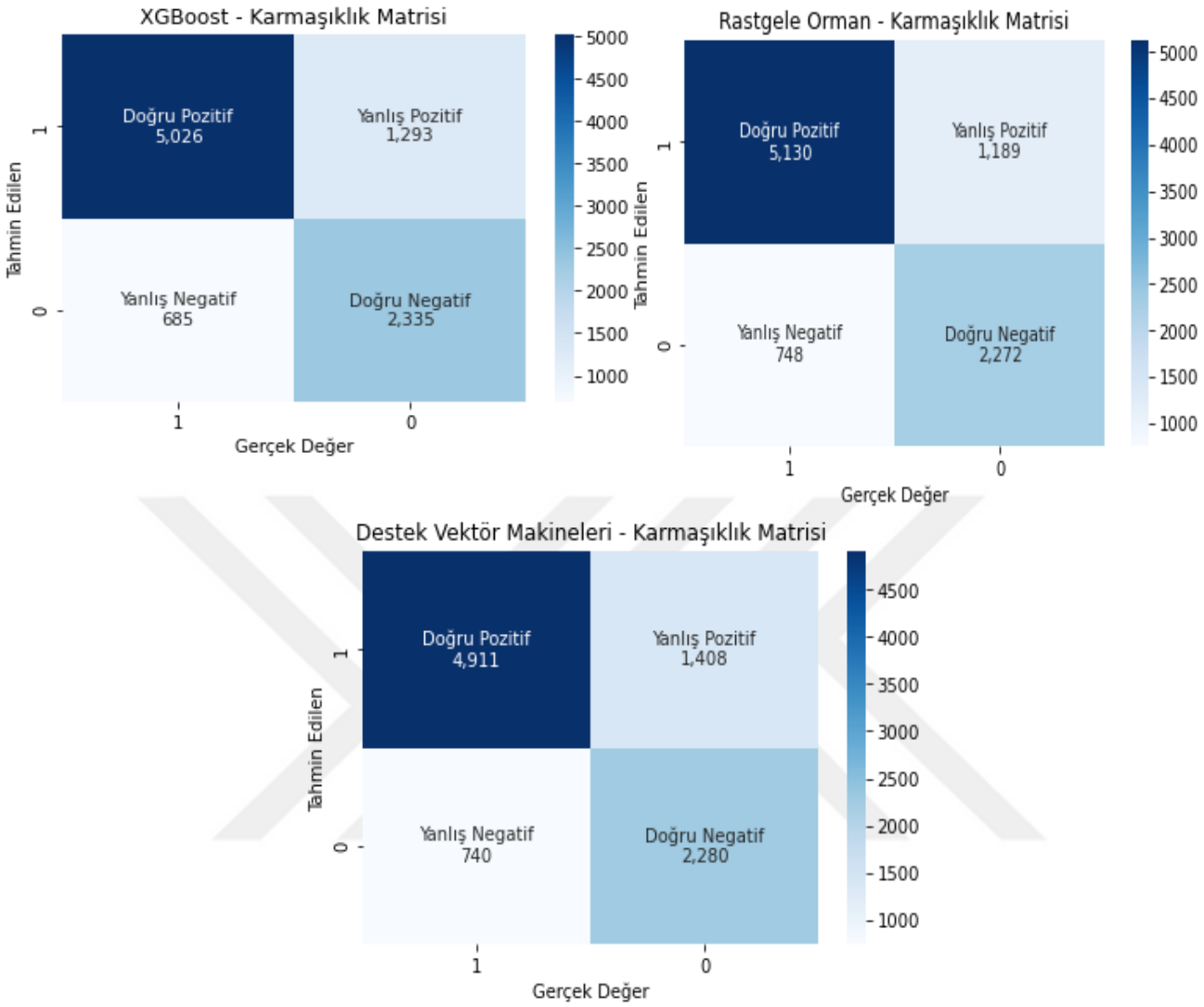
Yukarıdaki değerlendirme test verisi üzerinden gerçekleştirilmiştir. Çizelge 3.6'da doğrulama verisi üzerinden yapılan değerlendirme sonuçları paylaşılmıştır.

Çizelge 3.6 : Doğrulama Verisi Üzerinden Modellere Ait Ölçütlerin Hesaplanması

	Doğruluk	Kesinlik	Duyarlılık	Özgüllük	F Ölçütü	ROCAUC
Rastgele Orman	0,78	0,65	0,74	0,81	0,69	0,86
Lojistik Regresyon	0,77	0,62	0,73	0,79	0,67	0,80
Destek Vektör Makineleri	0,77	0,62	0,73	0,78	0,67	0,82
Naive Bayes	0,77	0,64	0,66	0,82	0,65	0,76
XGBoost	0,79	0,67	0,73	0,82	0,70	0,87
LightGBM	0,79	0,67	0,73	0,83	0,70	0,87

Doğrulama verisi ve test verisi üzerinden model performansları karşılaştırıldığında, modellerin her iki farklı veri seti üzerinde benzer performans gösterdiği görülmektedir.

Karmaşıklık Matrisleri;



Şekil 3.13 : Algoritmalara Ait Karmaşıklık Matrisleri

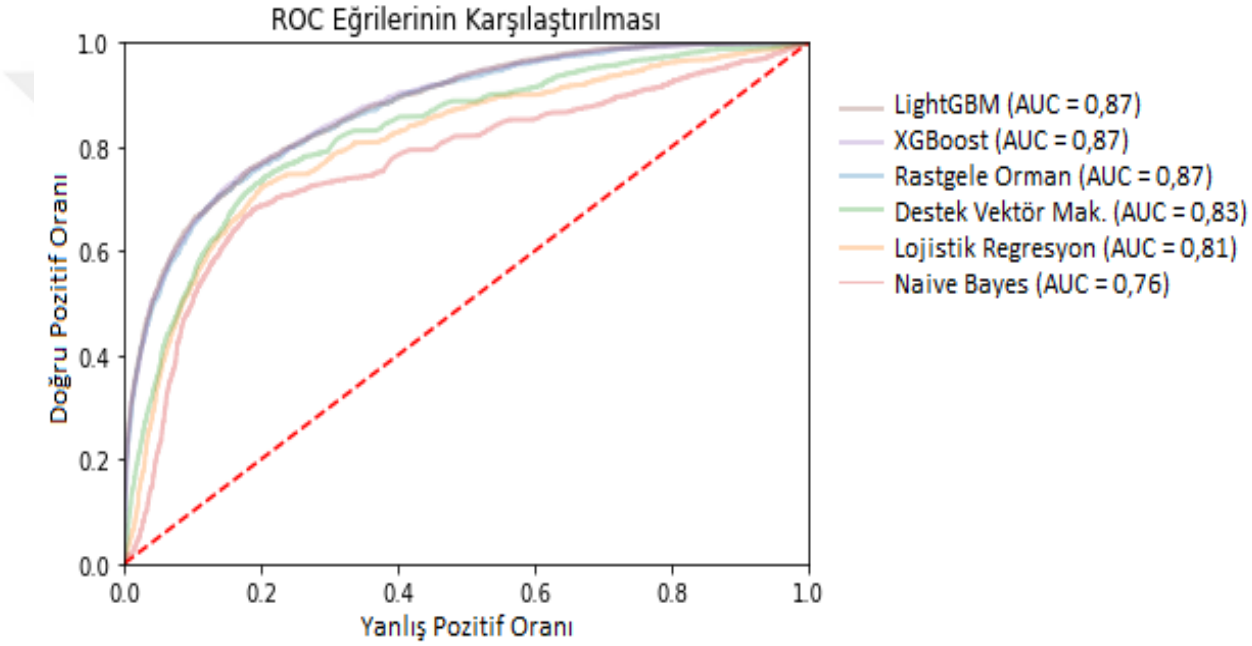
XGBoost, rastgele orman ve destek vektör makineleri modelleri için hesaplanan karmaşıklık matrisleri Şekil 3.13'te belirtilmiştir.

Karmaşıklık matrisleri üzerinden modellerin karşılaştırılması;

Karmaşıklık matrisleri, model performansının derinlemesine analiz edilmesinde önemli bir araçtır. Çalışmada yer alan XGBoost, rastgele orman ve destek vektör makineleri modellerinin karmaşıklık matrisleri incelenerek, bağışçıların düzenli bağış talimatlarını iptal etme eğilimini tahmin etmedeki başarıları ayrıntılı bir şekilde değerlendirilmiştir. Diğer modellere ait karmaşıklık matrisleri ise Ek-2'de belirtilmiştir.

Rastgele orman, destek vektör makineleri (SVM) ve XGBoost modellerinin karmaşıklık matrislerini incelediğimizde, her üç modelin de benzer performans gösterdiği tespit edilmiştir, ancak rastgele ormanın diğerlerine göre daha yüksek doğru sınıflandırmalar elde ettiği görülmektedir.

Karmaşıklık matrislerinin sunduğu bilgilerin yanı sıra, modellerin ROC eğrilerini de incelenmiştir. ROC eğrileri, duyarlılık ve özgüllük arasındaki ilişkiyi göstererek modellerin performansını değerlendirmemize yardımcı olur. Bu analiz, modellerin farklı eşik değerlerindeki performansını karşılaştırmamıza ve en uygun modeli seçmemize yardımcı olacaktır.



Şekil 3.14 : Modellere Ait ROC Eğrilerinin Karşılaştırılması

Şekil 3.14, çalışmada kullanılan 6 modele ait ROC eğrilerini göstermektedir.

Test verisi üzerinden hesaplanan ROC eğrilerine göre modellerin karşılaştırılması;

Bağışçıların düzenli bağış talimatlarını iptal etme eğilimini tahmin etmek için 6 farklı model değerlendirilmiştir. En yüksek AUC değeri (0,8722) LightGBM modeline aittir. Bu sonuç, LightGBM modelinin diğer modellere göre en yüksek tahmin yeteneğine sahip olduğunu gösterir. XGBoost modeli de benzer bir performans sergilerken (AUC: 0,8721), rastgele orman biraz daha düşük bir AUC değeriyle etkili bir sınıflandırma sağlamıştır (AUC: 0,869). Destek vektör makineleri ve lojistik regresyon modelleri, diğerlerine göre daha düşük AUC değerlerine sahiptir (sırasıyla 0.827 ve 0.807). En düşük AUC değeri (0,764) ise naive bayes modeline aittir, bu durum da naive bayes

modelinin diğerlerine göre daha zayıf bir sınıflandırma performansı olduğunu göstermektedir.

Sonuç olarak, karmaşıklık matrisi sonuçlarına göre XGBoost, rastgele orman ve destek vektör makineleri modelleri, AUC değerlerine göre de LightGBM, XGBoost ve rastgele orman modelleri bağışçı davranışlarını tahmin etme konusunda en etkili modeller olarak belirlenmiştir. Diğer modeller ise performans açısından geride kalmıştır.

4. SONUÇ

Bu tez çalışması, bağışçı davranışlarını tahmin etmek amacıyla makine öğrenimi modellerinin kullanılabilirliğini araştırmıştır. Çalışma kapsamında, farklı makine öğrenimi algoritmaları (rastgele orman, lojistik regresyon, destek vektör makineleri, naive bayes, XGBoost ve LightGBM) kullanılarak bağışçı davranışlarının tahmini üzerine bir model oluşturulmuştur.

Yapılan analizler sonucunda, XGBoost, LightGBM ve rastgele orman modellerinin bağışçı davranışlarını tahmin etme konusunda diğer modellere göre daha etkili olduğu belirlenmiştir. Bu modeller, doğruluk, kesinlik, özgüllük, F ölçütü ve ROCAUC değerlerinde diğer modellere kıyasla daha yüksek performans göstermiştir. Ayrıca, Karmaşıklık matrisi analizi ve ROC eğrisi analiziyle elde edilen AUC değerleri de XGBoost, LightGBM ve rastgele orman modellerinin diğerlerine göre daha başarılı olduğunu desteklemiştir. Diğer taraftan, lojistik regresyon ve naive bayes modelleri, diğer modellere kıyasla daha düşük performans sergilemiştir. Özellikle, naive bayes modelinin AUC değerinin diğer modellere göre en düşük olduğu belirlenmiştir.

Bu çalışma, bağışçı davranışlarının tahmini için kullanılabilecek makine öğrenimi modellerinin performansını karşılaştırmak ve en etkili modeli belirlemek amacıyla önemli bir adım olmuştur. XGBoost, LightGBM ve rastgele orman gibi modellerin bağışçı davranışlarını daha doğru bir şekilde tahmin etme yeteneği, kuruluşların bağış toplama stratejilerini optimize etmelerine ve kaynaklarını daha etkin bir şekilde kullanmalarına yardımcı olabilir. Gelecekteki çalışmalar, daha fazla veri ve farklı özelliklerin kullanılmasıyla bu modellerin performansını daha da artırabilir.

Bu çalışmanın sonuçları, bağış toplama ve sivil toplum kuruluşlarının stratejik karar alma süreçlerinde faydalı olabilir ve makine öğrenimi tekniklerinin bağışçı davranışlarını anlama ve tahmin etmede önemli bir araç olarak kullanılmasını teşvik edebilir. Bağışçıların bağış davranışlarına göre kategorize edilmesi, aramaların, e-posta, sms gönderimlerinin kişiselleştirilmesine imkan sağlayacaktır. Yapılan çalışma üzerinden ileride, bağışçıların düzenli bağış talimatlarına devam ve iptal durumları göz önünde bulundurularak özelleştirilmiş kaynak geliştirme faaliyetleri yürütülerek bağışçı bağlılığı artırılabilir. Bu nedenle, bu çalışmanın sonuçları, sivil toplum kuruluşlarının kaynak geliştirme faaliyetlerini iyileştirmek ve daha geniş kitlelere ulaşabilmeleri için önemli bir kılavuz olabilir.

KAYNAKLAR

[1] **Andreoni, J. & Payne, A.** (2013). Charitable Giving, Handbook of Public.

[2] **Sathyamurthi, A. V.** (2022). Predicting Major Donor Prospects using Machine Learning. Master's thesis, Acadia University.

[3] **Tiernan, K., Singh G.** (2019). How data analysis can help your charity to enhance its fundraising results, Erişim: 26.04.2024, https://thenonprofittimes.com/column_database/predictive-analytics/

[4] **Duffy, A.** (2018). Using Machine Learning and Optimization to Improve Refugee Integration, Erişim: 26.04.2024, https://www.wpi.edu/news/using-machine-learning-and-optimization-improve-refugee-integration

[5] **Key, J.** (2001). Enhancing fundraising success with custom data modeling, International Journal of Nonprofit and Voluntary Sector Marketing, 6(4), 335–346.

[6] **Malthouse, E. C.** (2001). Assessing the performance of direct marketing scoring models. Journal of Interactive Marketing, 15(1), 49–62.

[7] **Zhao, H., B. Jin, Q. Liu, Y. Ge, E. Chen, X. Zhang, T. Xu.** (2019). Voice of charity: Prospecting the donation recurrence & donor retention in crowdfunding. IEEE Transactions on Knowledge and Data Engineering.

[8] **Althoff, T., J. Leskovec.** (2015). Donor retention in online crowdfunding communities: A case study of donorschoose.org. Proceedings of the 24th International Conference on World Wide Web (pp. 34-44).

[9] **Veersma, E., Sananka A.** (2023). One Acre Fund - Creating a Chatbot With Limited Resources to Forecast Optimal Seeding Time in Sub-Saharan Africa, Erişim: 26.04.2024, https://community.dataiku.com/t5/Dataiku-Frontrunner-Awards/One-Acre-Fund-Creating-a-Chatbot-With-Limited-Resources-to/ta-p/27985

[10] **Barinov, A.** (2022). The Use of Machine Learning for Customer Churn Prediction, Erişim: 03.04.2024, https://intelliarts.com/blog/machine-learning-for-customer-churn-prediction-in-insurance/

[11] **Saeys, Y., Inza, I. ve Larranaga, P.** (2007). A review of feature selection techniques in bioinformatics, Bioinformatics 23, pp. 2507–2517.

[12] **Budak, H.** (2018). Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22, 21-31.

[13] **Das, S.** (2001). Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. Proceedings of the Eighteenth International Conference on Machine Learning, (pp. 74-81). Williamstown, June 28- July 1.

[14] **Breiman, L.** (1996). Bagging Predictors, Machine Learning, 24(2), 123-140.

[15] **Breiman, L.** (2001). Random Forests, Machine Learning, 45(1), 5-32.

[16] **Kuhn, M. ve Johnson K.** (2013). Applied Predictive Modeling, Springer

[17] **Uncuoğlu, E.** (2018). Destek Vektör Makinaları Kullanarak Kişisel Termal Konfor Modellemesi. (Yüksek Lisans Tezi). Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli

[18] **Schölkopf B. ve Smola A.** (2002), Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, Adaptive computation and machine learning, MIT Press

[19] **Meric, E. ve Ozer C.** (2023). Symptom based health status prediction via Decision tree XGBoost, LDA SVM and random forest. Computational Intelligence, Data Analytics and Applications, Selected papers from the International Conference on Computing, Intelligence and Data Analytics (ICCIDA), (pp.193-207). March 2023.

[20] **S, Haykin** (2009). Neural Networks and Learning Machines, Pearson, Upper Saddle River, NJ

[21] **Tontuş, H.** (2020). Sarf Malzeme Kullanımından Veri Madenciliği Birliktelik Kurallarının Elde Edilmesi, Kuralların Analizi Ve Sınıflandırılması. (Yüksek Lisans Tezi). Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara

[22] **Chen, T.Q. ve Guestrin, C.** (2016). Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (pp. 785-794). San Francisco, August 13-17.

[23] **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: Annual Conference on Neural Information Processing Systems, (pp. 3146-3154). Long Beach, December 4-9.,

[24] **Singh, A., Tiwari, V. ve Tentu, A. N.** (2018). A Machine Vision Attack Model on Image Based CAPTCHAs Challenge: Large Scale Evaluation. In International Conference on Security, Privacy, and Applied Cryptography Engineering, Springer, Cham, December 15-19, Kanpur, 52-64.

[25] **Arlot, S. ve Celisse, A.** (2010), A survey of cross-validation procedures for model selection, Statistics surveys, 4, 40-79.

[26] **Wong, T.T.** (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recognition, 48(9), 2839- 2846.

[27] **Doppala, B. P., Janarthanan M., Bhattacharyya D.** (2021). A Novel Approach to Predict Cardiovascular Diseases Using Machine Learning, Springer, Berlin, January, 71–80.

[28] **Wiens, T. S., Dale, B. C., Boyce, M. S. ve Kershaw, G. P.** (2008). Three way k-fold cross validation of resource selection functions, Ecological Modelling, 212(3-4), 244-255.

[29] **Sokolova, M., Japkowicz, N. ve Szpakowicz, S.** (2006). F-score and ROC: a family of discriminant measures for performance evaluation. In

Australasian Joint Conference on Artificial Intelligence, Springer, Cham, January, Berlin, 1015-1021.

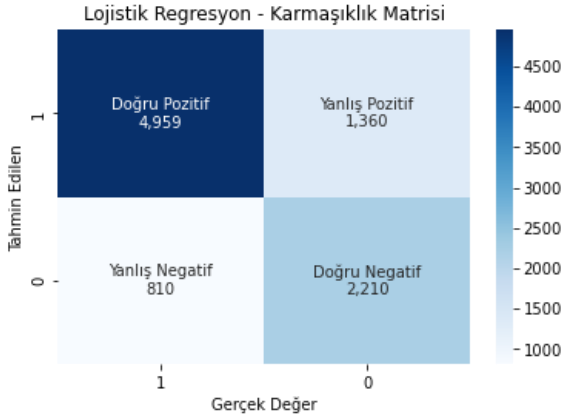
[30] **Gantley, M., Whitehouse, H. ve Bogaard, A.** (2018). Material Correlates Analysis (MCA): An Innovative Way of Examining Questions in Archaeology Using Ethnographic Data. Advances in Archaeological Practice, 6(4): 328-341.

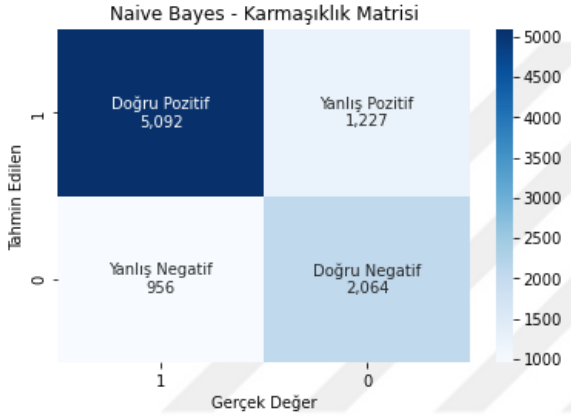
EKLER

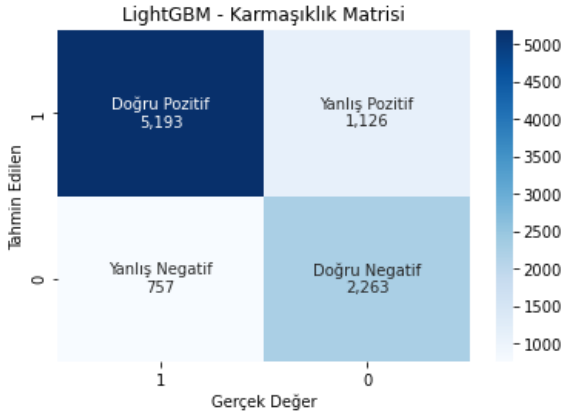
EK-1

Değişken	Değişken Açıklaması
toplam_bagis_tutari	Tüm bağışların toplamı
x2010_oncesi_toplam_bagis_tutari	Bağışçının 2009 yılı ve öncesinde yaptığı tüm bağışların toplamı
ilk_bagis_yili	Bağışçının ilk bağışının gerçekleştiği yıl
platform_bagislari_T	İnternette yer alan platformlar üzerinden yapılan toplam bağış tutarı
telemarketing_A	Bağışçılarla yapılan telefon görüşmeleri ile elde edilen toplam bağış adedi
telemarketing_T	Bağışçılarla yapılan telefon görüşmeleri ile elde edilen toplam bağış tutarı
adak_bagisi_A	Adak olarak yapılan toplam bağış adedi
adak_bagisi_T	Adak olarak yapılan toplam bağış tutarı
aylik_tek_seferlik_destek_paketleri_A	Destek paketleri halinde yapılan toplam bağış adedi
aylik_tek_seferlik_destek_paketleri_T	Destek paketleri halinde yapılan toplam bağış tutarı
ayni_bagis_tutar_A	Eşya, hizmet veya bedelsiz kullandırma ile yapılan toplam bağış adedi
Bayram Bağışı_T	Kurban Bayramı ve Ramazan Bayramı dönemlerinde yapılan toplam bağış tutarı
bagis_hediyeleri_A	Satın alınan hediyeler ile yapılan toplam bağış adedi
bagis_kampanyalari_A	Oluşturulan kampanyalar üzerinden yapılan toplam bağış adedi
Bağış Kampanyaları_T	Oluşturulan kampanyalar üzerinden yapılan toplam bağış tutarı
diger_telif_hakki_gelirler_A	Telif gelirleri ile yapılan toplam bağış adedi
diger_telif_hakki_gelirler_T	Telif gelirleri ile yapılan toplam bağış tutarı
duzenli_bagis_A	Toplam düzenli bağış adedi
duzenli_bagis_T	Toplam düzenli bağış tutarı
etkinlik_A	Etkinlikler ile yapılan toplam bağış adedi
fidye_A	Fidye olarak yapılan bağışların toplam adedi
fidye_T	Fidye olarak yapılan bağışların toplam tutarı
fitre_zekat_A	Fitre ve zekat olarak yapılan bağışların toplam adedi
fitre_zekat_T	Fitre ve zekat olarak yapılan bağışların toplam tutarı
maraton_A	Maraton kampanyaları özelinde yapılan toplam bağış adedi
sertifika_A	Sertifika olarak yapılan toplam bağış adedi
Sertifika_T	Sertifika olarak yapılan toplam bağış tutarı
ozel_gun_bagislari_A	Özel günler (doğum günü, nikah vb.) için yapılan toplam bağış adedi
ozel_gun_bagislari_T	Özel günler (doğum günü, nikah vb.) için yapılan toplam bağış tutarı
sartli_bagis_A	Bağışçıların belirli bir şart doğrultusunda yaptıkları toplam bağış adedi
sartli_bagis_T	Bağışçıların belirli bir şart doğrultusunda yaptıkları toplam bağış tutarı
websitesi_T	Kurum web sitesi üzerinden yapılan bağış toplamı
bagisci_kayit_tipi_Birey	Bağışçının birey ise 1 kurum ise 0 olarak gösteren özellik
talimat_durumu	Düzenli bağış talimatının devam ettiğini veya iptal olma durumunu gösteren özellik

EK-2


Lojistik Regresyon - Karmaşıklık Matrisi
Doğru Pozitif
4,959
Yanlış Pozitif
1,360
Yanlış Negatif
810
Doğru Negatif
2,210
Tahmin Edilen
Gerçek Değer
1
0
4500
4000
3500
3000
2500
2000
1500
1000


Naive Bayes - Karmaşıklık Matrisi
Doğru Pozitif
5,092
Yanlış Pozitif
1,227
Yanlış Negatif
956
Doğru Negatif
2,064
Tahmin Edilen
Gerçek Değer
1
0
5000
4500
4000
3500
3000
2500
2000
1500
1000


LightGBM - Karmaşıklık Matrisi
Doğru Pozitif
5,193
Yanlış Pozitif
1,126
Yanlış Negatif
757
Doğru Negatif
2,263
Tahmin Edilen
Gerçek Değer
1
0
5000
4500
4000
3500
3000
2500
2000
1500
1000