T.C.

BARTIN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MATEMATİK ANABİLİM DALI

YÜKSEK LİSANS TEZİ

DENGESİZ SINIF DURUMUNDA İKİLİ SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

ABDULLAH FAZLI

DANIŞMAN

DOÇ. DR. EMRAH ALTUN

BARTIN-2024

T.C.

BARTIN ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MATEMATİK ANABİLİM DALI

DENGESİZ SINIF DURUMUNDA İKİLİ SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS TEZİ

Abdullah FAZLI

JÜRİ ÜYELERİ

Danışman :

Üye :

Üye :

BARTIN-2024

KABUL VE ONAY

BEYANNAME

Bartın Üniversitesi Lisansüstü Eğitim Enstitüsü tez yazım kılavuzuna göre Doç. Dr. Emrah ALTUN danışmanlığında hazırlamış olduğum "DENGESİZ SINIF DURUMUNDA İKİLİ SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI" başlıklı yüksek lisans tezimin bilimsel etik değerlere ve kurallara uygun, özgün bir çalışma olduğunu, aksinin tespit edilmesi halinde her türlü yasal yaptırımı kabul edeceğimi beyan ederim.

23.10.2024

Abdullah FAZLI

ÖZET

Yüksek Lisans Tezi

DENGESİZ SINIF DURUMUNDA İKİLİ SINIFLANDIRMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Abdullah FAZLI

Bartın Üniversitesi
Lisansüstü Eğitim Enstitüsü
Matematik Anabilim Dalı

Tez Danışmanı: Doç. Dr. Emrah ALTUN

Bartın-2024, sayfa: 37

Bu çalışmada dengesiz sınıf dağılımı durumunda ikili sınıflandırma problemi incelenmiştir. Dengesiz sınıf dağılımı olduğu durumda örnekleme yöntemlerinden yararlanılmaktadır. Böylece, dengeli bir sınıf dağılımı elde edilmektedir. Bu amaçla SMOTE-NC algoritması kullanılmıştır. Lojistik regresyon, destek vektör makineleri ve gradient boosting modellerinin SMOTE-NC algoritması altında elde edilen dengeli veri setlerindeki performansları incelenmiştir. Elde edilen sonuçlara göre SMOTE-NC algoritmasının gradient boosting ile birlikte kullanımı dengesiz sınıf dağılımı durumunda ikili sınıflandırma başarısını artırmaktadır. SMOTE-NC algoritması azınlık sınıfın doğru sınıflandırma oranını yükseltmektedir.

Anahtar Kelimeler: Dengesiz veri setleri, makine öğrenmesi, ikili sınıflandırma, lojistik regresyon

ABSTRACT

M.Sc. Thesis

COMPARISON OF BINARY CLASSIFICATION METHODS IN CASE OF IMBALANCE CLASS

Abdullah FAZLI

Bartın University
Graduate School
Department of Mathematics

Thesis Advisor: Assoc. Prof. Dr. Emrah ALTUN
Bartın-2024, pp: 37

This study investigates the binary classification problem in the case of imbalanced class distribution. In case of imbalanced class distribution, sampling methods are utilized. In this way, a balanced class distribution is obtained. SMOTE-NC algorithm is used for this purpose. The performances of logistic regression, support vector machines and gradient boosting models on balanced data sets obtained under SMOTE-NC algorithm are analyzed. According to the obtained results, the use of SMOTE-NC algorithm together with gradient boosting increases the binary classification performance in case of imbalanced class distribution. SMOTE-NC algorithm increases the correct classification rate of the minority class.

Keywords: Imbalanced data sets, machine learning, binary classification, logistic regression

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

Şekil No **Sayfa No**

TABLOLAR DİZİNİ

Tablo No **Sayfa No**

SİMGELER VE KISALTMALAR DİZİNİ

β	Parametre vektörü
ℓ	Log-olabilirlik fonksiyonu
Ψ	Kayıp fonksiyonu
h	Temel öğrenici fonksiyon

KISALTMALAR

DVM	Destek vektör makineleri
GBM	Gradient Boosting Machines
SMOTE	Synthetic minority over-sampling technique (Sentetik Azınlık Aşırı Örnekleme Yöntemi)
SGD	Steepest gradient descent (En dik iniş algoritması)

1. GİRİŞ

İkili sınıflandırmada bağımlı değişkenin aşırı sıfır değeri içermesi dengesiz sınıf dağılımı olarak nitelendirilmektedir. İkili sınıflandırma algoritmalarının finans, sağlık ve sigortacılık temelinde geniş uygulama alanları bulmaktadır. Kanserli hücrelerin sınıflandırılması, kredi başvurularında ilgili kişiye kredi verilip verilmeyeceği ya da kasko poliçesi yaptırırken risk primlerinin hesaplanmasında kullanılmaktadır.

Dengesiz sınıf dağılımı, destek vektör makineleri, karar ağaçları ve sinir ağları gibi sınıflandırıcıların performansını kötü etkilemektedir (Cateni vd., 2014). Bu sınıflandırıcılar veri seti üzerinde genel performansını maksimize edecek şekilde dizayn edilmişlerdir. Bu nedenle dengesiz sınıf durumundaki başarıları düşüktür. Nadir olayların (rare event) sınıflandırma başarıları düşükken diğer olayların sınıflandırma başarısı yüksektir. Fakat asıl amaç nadir olaylarının doğru sınıflandırılması olduğu için modellerin öngörü performansı gerçek yaşam problemlerinde bu haliyle ile kullanılamaz. (Estabrooks ve Japkowicz, 2004).

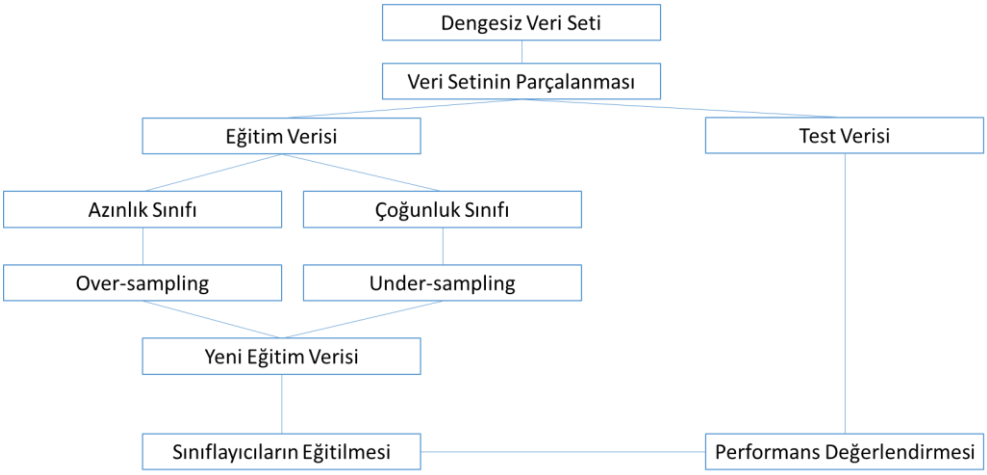
Dengesiz sınıf dağılımının olduğu verilerin sınıflandırılmasında araştırmacılar iki yaklaşım üzerine çalışmaktadır. Birincisi, dengesiz sınıf dağılımını modelleyebilecek yeni algoritma geliştirmek. İkincisi ise veri setini yeniden örnekleme yöntemlerini kullanarak mevcut algoritmaların performansını artıracak şekilde yeniden oluşturmak. Birinci yaklaşımda araştırmacılar yanlış sınıflandırma için bir cezalandırma maliyeti kullanırlar. Bu cezalandırma maliyeti nadir olaylar için daha yüksek iken diğer durum için daha düşüktür. Li vd. (2006) destek vektör makineleri kullanarak dengesiz veri setleri için algoritma geliştirmişlerdir. Soler ve Prim (2007) radyal tabanlı fonksiyonların özel bir durumunu kullanarak yeni bir ikili sınıflandırma yöntemi geliştirmiştir. Bunların yanı sıra, genelleştirilmiş lineer modellere dayanan farklı yaklaşımlar ile de bu tür verilerin modellenmesi için yeni yaklaşımlar önerilmiştir (Chiang vd., 2023).

İkinci yaklaşımda amaç azınlık ve çoğunluk sınıflarının dağılımlarını dengelemektir. Bu aşamada yeniden örnekleme yöntemlerinden yararlanılarak azınlık sınıfın gözlem sayıları artırılır. Azınlık sınıfın gözlem sayısı artırılarak ya da çoğunluk sınıfın gözlem sayıları düşürülerek bir dengeleme sağlanır. Bu iki yöntem sırasıyla düşük örnekleme (under-sampling) ve yüksek örnekleme (over-sampling) olarak bilinmektedir (Seiffert vd., 2009).

Araştırmacılar bu iki yeniden örnekleme yöntemlerini birlikte kullanarak farklı yöntemler geliştirmişlerdir. Chawla (2002) azınlık sınıf için over-sampling yöntemi geliştirmiştir. Bu yöntem SMOTE (Synthetic Minority over-sampling technique) olarak adlandırılmıştır. SMOTE yöntemi diğer makine öğrenme yöntemleri ile birlikte kullanılarak dengesiz veri setlerinin sınıflandırma başarılarının artırılması sağlanmıştır (Wang, 2008). Diğer önemli bir çalışma Cateni vd. (2014) tarafından gerçekleştirilmiştir. Cateni vd. (2014) SUNDO yöntemini geliştirilmişlerdir. Bu yöntem over-sampling ve under-sampling yöntemlerinin birleşiminden oluşmaktadır. Cateni vd. (2014) SUNDO yönteminin etkinliğinin SMOTE yöntemi ile karşılaştırmıştır.

SMOTE yöntemi literatürde oldukça popüler bir yöntem haline gelmiştir. Rahmayanti vd. (2021), sınıflandırma ve regresyon ağacı (classification and regression tree, CART) modelini SMOTE yöntemi ile birlikte kullanarak müşteri kayıp tahmini analizi gerçekleştirmiştir. Gök ve Olgun (2021) rastgele orman (random forest) modeli ile SMOTE algoritmasını birleştirerek COVID-19 verisi üzerine bir çalışma yapmıştır. Joloudari vd. (2023) yapmış oldukları çalışmada, 24 farklı veri setini analiz ederek SMOTE yönteminin evrişimsel sinir ağları ile birlikte kullandığında dengesiz veri setlerinde yüksek sınıflandırma başarısının elde edildiğini göstermişlerdir.

Bu çalışmada amaç, SMOTE yönteminin dengesiz sınıf dağılımındaki etkinliğinin lojistik regresyon (LR), destek vektör makineleri (DVM) ve gradient boosting makineleri (GBM) yöntemlerinde değerlendirilmesidir. Bu amaçla kullanılacak yöntem Şekil 1.1'de özetlenmiştir. Farklı yöntemlerin karşılaştırılması için R programında benzetim çalışması yapılmıştır. Performans ölçütleri olarak doğru sınıflandırma oranı, duyarlılık ve seçicilik değerleri kullanılmıştır.



Şekil 1.1: Dengesiz veri setlerinin modellenmesi

Çalışmanın ikinci bölümünde ikili sınıflandırma modellerine yer verilecektir. Üçüncü bölümde SMOTE algoritmasının çalışma prensibi ele alınacaktır. Dördüncü bölümde ise makine öğrenmesi yöntemleri gerçek bir veri seti üzerinde karşılaştırılacaktır. Beşinci bölümde LR, DVM ve GBM modellerinin etkinlikleri benzetim çalışması ile karşılaştırılmıştır. Çalışma, sonuç ve tartışma bölümü olan altıncı bölüm ile tamamlanacaktır.

2. İKİLİ SINIFLANDIRMA MODELLERİ

2.1. Lojistik Regresyon Modeli

İkili sınıflandırma modellerinde en çok kullanılan model lojistik regresyon modelidir. Bu modelde bağımlı değişken iki değer alır. Örnek olarak, iyileşti/iyileşmedi, pozitif/negatif veya var/yok durumları verilebilir. İlgilenilen olay 1, diğer durum ise 0 olarak kodlanır. Bu durumda Y değişkeni bir indikatör değişkeni görevi görmektedir ve $\Pr(Y=1)=E(Y)$ 'dir.

Lojistik regresyon modeli hala araştırmacılar tarafından ikili sınıflandırma problemleri için kullanılmaktadır. Vatansever (2014), Ural vd. (2015), Ayan ve Değirmenci (2018) ve Aktümsek ve Göker (2018), lojistik regresyon modeli ile firmalar için finansal başarısızlık modellemesi gerçekleştirmişlerdir. Bircan (2004) ise lojistik regresyon modelinin uygulamasını sağlık verilerine üzerine gerçekleştirmiştir. Oğuzlar (2005) ise suçlu profilinin belirlenmesi ve sınıflandırılmasında lojistik regresyon modelini kullanmıştır.

Lojistik regresyonda koşullu olabilirlik fonksiyonu Eşitlik (1)'de verildiği gibi tanımlanır.

$$\Pr\left(Y=y_i \middle| X=x_i\right)=\prod_{i=1}^{n} p\left(x_i\right)^{y_i}\left(1-p\left(x_i\right)^{1-y_i}\right) \tag{1}$$

Eşitlik (1)'de $p(x)$ fonksiyonu lineer bir fonksiyondur. Bağımsız değişkenlerdeki her bir birimlik artış veya azalış $p(x)$ fonksiyonu üzerinde büyük etkiye sahiptir. Ayrıca $p(x)$ fonksiyonu 0-1 aralığında tanımlı olmalıdır. Bu yüzden, $p(x)$ fonksiyonu üzerinde logit dönüşümü uygulanır. Logit dönüşümü Eşitlik (2)'de verildiği gibidir.

$$\log\left(\frac{p_i}{1-p_i}\right)=\operatorname{logit}\left(p_i\right)=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik} \tag{2}$$

Eşitlik (2) p için çözülürse, Eşitlik (3) elde edilir.

$$p_i=\frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)} \tag{3}$$

Eşitlik (3)'de $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ik})'$ $(k+1)\times 1$ boyutlu bağımsız değişken vektörü, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ ise $(k+1)\times 1$ boyutlu regresyon parametre vektörüdür. Lojistik regresyon için olabilirlik fonksiyonu ise Eşitlik (4)'de verilmiştir.

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \tag{4}$$

Log-olabilirlik fonksiyonu ise Eşitlik (5)'de tanımlandığı gibidir.

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \\ &= \sum_{i=1}^{n} \log(1 - p_i) + \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1 - p_i}\right) \end{aligned} \tag{5}$$

Eşitlik (3), Eşitlik (5)'de yerine yazılırsa, Eşitlik (6) elde edilir.

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \log\left(1 + \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)\right) \tag{6}$$

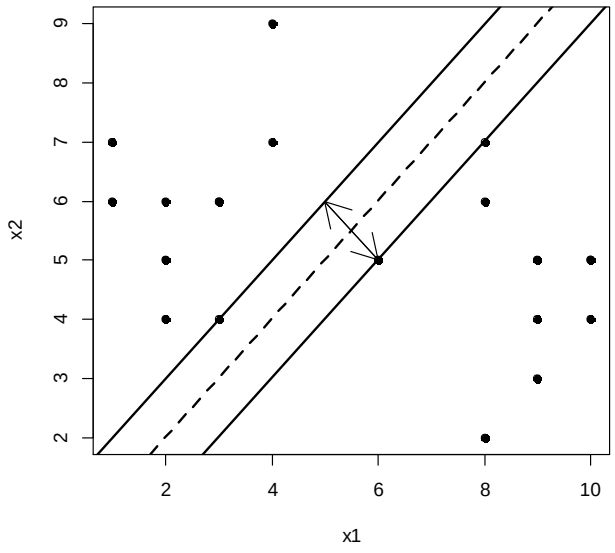
Eşitlik (6)'da β_j parametrelerine göre kısmi türevler alınarak elde edilen skor vektörlerinin 0 için eşanlı çözümleri $\boldsymbol{\beta}$ vektörünün en çok olabilirlik (EÇO) tahmin edicisini verir. Fakat Eşitlik (6) Nelder-Mead algoritması yardımıyla doğrudan maksimize edilerek parametre tahminleri elde edilebilir. Bu amaçla R programında optim fonksiyonu kullanılabilir.

2.2. Destek Vektör Makineleri

DVM, finans, mühendislik ve sağlık bilimleri gibi konularda geniş uygulama alanlarına sahiptir. Kavzoğlu ve Çölkesen (2010), uydu görüntülerinin sınıflandırılmasında kernel tabanlı DVM modellerini kullanmıştır. Yavut vd. (2014) DVM ve yapay sinir ağları modellerini kullanarak BİST-100 endeks tahmini gerçekleştirmiştir. Ayhan ve Erdoğmuş (2014) DVM için çekirdek fonksiyonu seçim sürecini rastgele bloklar deney tasarımı modeli ile gerçekleştirmiştir. Küçüksille ve Ateş (2013), istenmeyen e-postaların tespiti ve sınıflandırılması için DVM modelini kullanmışlardır.

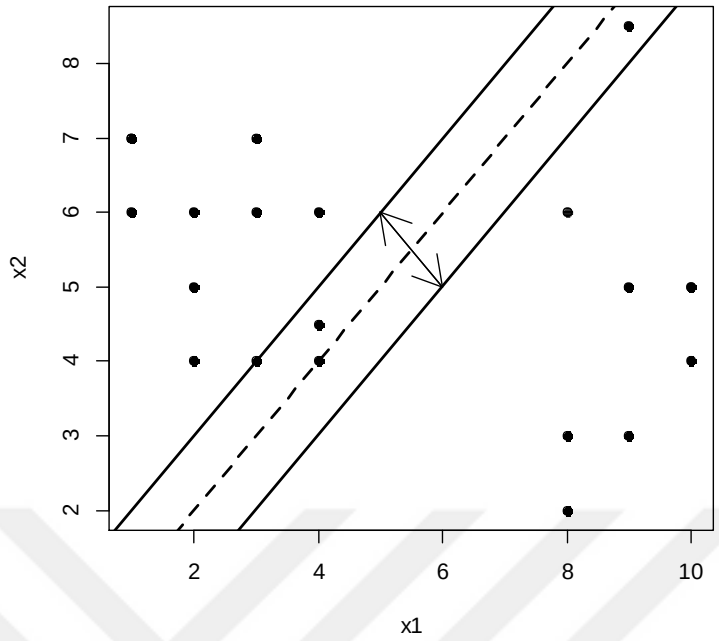
DVM, $g(x) = w^T x + b$ şeklinde tanımlanan hiper-düzlemde maksimum marjin değerine göre iki sınıfı ayırır. Doğrusal olarak ayrılabilen iki sınıflı yapıda doğrusal DVM modeli kullanılır. Belirlenen destek vektörler içerisinde herhangi bir gözlem kalmıyorsa hard-marjin, eğer kalıyorsa soft-marjin durumu vardır. Şekil 2.1'de hard-marjin durumu ele alınmıştır. Şekil 2.1'de görüldüğü gibi belirlenen destek vektörler içerisinde kalan herhangi bir gözlem bulunmamaktadır. x_i noktalar kümesi ve w_1, w_2 doğrusal olarak ayrılabilen iki sınıf olmak üzere, hiper-düzlemdeki noktalar arası uzaklık $|g(x)|/\|w\|$ olarak ifade edilir. DVM, $x \in w_1$ için $w^T x + b = 1$ ve $x \in w_2$ için $w^T x + b = -1$ koşullarını sağlayan w ve b parametrelerinin bulunmasını amaçlar. DVM Eşitlik (7)'de verilen optimizasyon probleminin çözümüne dayanmaktadır.

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 \\ &s.t. \\ &y_i\left(w_i^T x + b\right) \geq 1, i = 1, 2, ..., N \end{aligned} \tag{7}$$



Şekil 2.1: Hard-marjin DVM

Eğer veri doğrusal olarak tam ayrılabilir yapıda değilse, bu durumda DVM gevşek değişkenler yardımıyla hatalı sınıflandırmaya izin verir ve durum soft-marjin olarak isimlendirilir. Şekil 2.2'de soft-marjin durumu gösterilmiştir.



Şekil 2.2: Soft-marjin DVM

Soft-marjin durumunda DVM aşağıda verilen optimizasyon probleminin çözümüne dayanmaktadır.

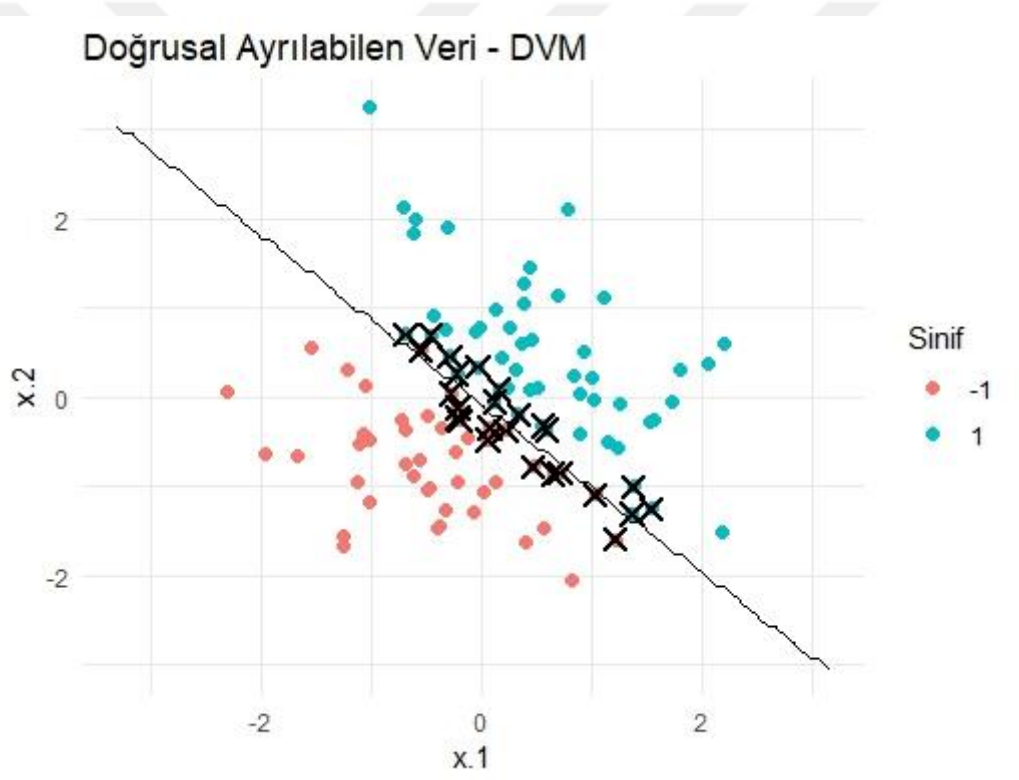
$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\delta_i \\ &s.t. \\ &y_i\left(w_i^T x + b\right) \geq 1-\delta, i = 1,2,...,N \\ &\delta_i \geq 0, i = 1,2,...,N \end{aligned} \tag{8}$$

Eşitlik (8)'de δ_i gevşek değişkenleri göstermektedir. C değeri ise düzenleme terimi olarak ifade edilmektedir. C artırıldıkça daha sıkı bir marj elde edilir ve yanlış sınıflandırma sayısının en aza indirilmesine daha fazla önem verilir (Awad, ve Khanna, 2015).

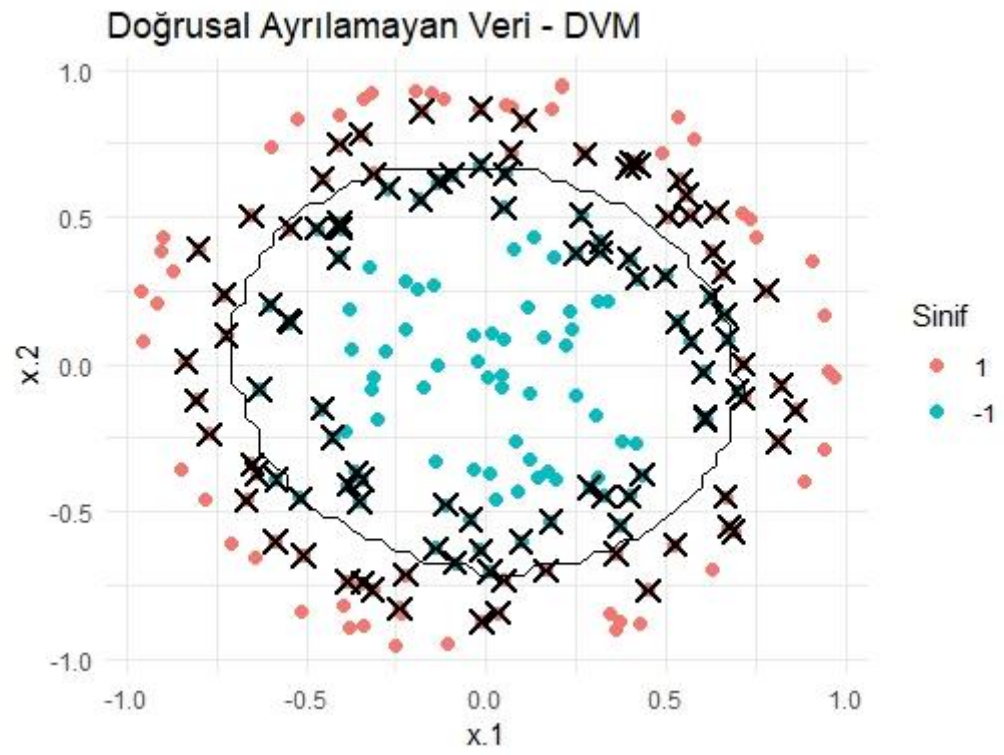
Eğer veri doğrusal olarak ayrılamıyorsa, soft-marjin DVM hatalı sınıflandırma noktalarının sayısını minimize edebilecek bir hiper-düzlem bulamaz. Bu durumda kernel fonksiyonları yardımıyla veri, daha yüksek boyutlu bir uzaya dönüştürülür ve bu uzaya kernel uzayı denir. Kernel uzayında veri, doğrusal olarak ayrılabilen bir yapıya dönüşür. Kernel, Hermisyen pozitif yarı-tanımlı bir matris olup Mercer teoremini sağlamaktadır. (Awad, ve Khanna, 2015). En çok kullanılan kernel fonksiyonları, radyal tabanlı fonksiyon, polinom fonksiyonu

hiperbolik tanjant fonksiyonudur. DVM, e1071 paketinde tanımlı olan svm fonksiyonu kullanılarak R yazılımında uygulanabilir.

DVM modeli altında doğrusal ayrılabilen ve ayrılamayan durumları göstermek için R programında benzetim çalışması yapılmıştır. Doğrusal ayrılabilen durum için iki bağımsız değişken normal dağılımdan üretilmiştir. Doğrusal ayrılamayan durum için ise dairesel formda veriler üretilmiştir. $x_1 = r\cos(\theta)$ ve $x_2 = r\sin(\theta)$ değişkenleri tanımlanmıştır. Doğrusal ayrılabilen veri seti doğrusal kernel fonksiyon ile ayrılamayan veri ise radyal tabanlı kernel fonksiyonu ile modellenmiştir. Sonuçlar Şekil 2.3 ve 2.4'te verilmiştir.



Şekil 2.3: Doğrusal ayrılabilen veri altında DVM sonuçları



Şekil 2.4: Doğrusal ayrılamayan veri altında DVM sonuçları

2.3. Gradient Boosting

Gradient Boosting Machines (Gradyan Arttırma Makineleri, GBM) yöntemi en çok kullanılan makine öğrenmesi algoritmalarından biridir. Birçok uygulama alanında kendisine yer bulmuştur. Üstüner vd. (2020) GBM yöntemi ile tarımsal ürünlerin sınıflandırılması problemini ele almışladır. Atasoy ve Demiröz (2021), GBM modelinin yanı sıra birçok makine öğrenmesi yöntemini kullanarak prostat kanserinde tümör oluşum durumunu incelemişlerdir. Tütüncü ve Gürsakal (2023) bankacılık sektöründe önemli bir konu olan temerrüt risk durumunu modellemiş ve en iyi sonucu veren yöntemin GBM modeli olduğu sonucuna varmıştır.

$\mathbf{x} = \left(x_1, x_2, ..., x_k \right)$ açıklayıcı değişkenler vektörü, y bağımlı değişken vektörü olmak üzere modellenecek veri kümesi $\left(\mathbf{x}, y \right)_{i=1}^{N}$ şeklinde ifade edilsin. Amaç bu iki vektör arasında bilinmeyen fonksiyonel yapının elde edilmesidir, $\hat{f}\left(x \right) : x \to y$. Bu durum Eşitlik (9)'da belirtilen kayıp fonksiyonunun minimize edilmesi problemini oluşturur.

$$\hat{f}(x) = y,$$
$$\hat{f}(x) = \arg\min_{f(x)} \Psi\left(y, f(x)\right) \tag{9}$$

Eşitlik (9), beklenen değerler üzerinden ifade edilirse, Eşitlik (10) elde edilir.

$$\hat{f}(x) = \hat{f}(x,\theta)$$
$$\hat{\theta} = \arg\min_{\theta} E_x\left[E_y\left(\Psi\left[y, f(x,\theta)\right]\right)\middle| x\right] \tag{10}$$

Eşitlik (10)'da verilen optimizasyon probleminin kapalı formda bir çözümü bulunmamaktadır. Bu nedenle sayısal yöntemler kullanılarak çözülmesi gerekmektedir.

M itiresyon sayısı olmak üzere parametre tahmini $\hat{\theta} = \sum_{i=1}^{M} \hat{\theta}_i$ şeklinde ifade edilir. Parametre tahminlerinin elde edilmesinde steepest gradient descent (SGD) algoritması kullanılmıştır. Gözlemlenen veri seti için, deneysel kayıp fonksiyonunun azaltılmasına dayanmaktadır.

$$J(\theta) = \sum_{i=1}^{N} \Psi\left(y_i f\left(x_i, \hat{\theta}\right)\right) \tag{11}$$

SGD algoritması, kayıp fonksiyonunun gradyanı yönünde ardışık iyileştirmelerine dayanır. Kayıp fonksiyonun gradyanı $\nabla J(\theta)$ şeklinde ifade edilmektedir. $\hat{\theta}$ parametresi artan şekilde ifade edilmiştir. $\hat{\theta}_t$, $\hat{\theta}$ için t. artış adımını, $\hat{\theta}^t$ ise 1. adımdan t. adıma kadar alan artışlar toplamını ifade etmektedir. SGD algoritmasının adımları aşağıda verilmiştir (Natekin ve Knoll, 2013).

1. $\hat{\theta}_0$ için bir başlangıç değeri belirle.
2. Her bir iterasyon için 3-6 adımlarını tekrar et.
3. Tüm önceki iterasyon değerlerini kullanarak $\hat{\theta}^t$ değerini hesapla

$$\hat{\theta}^t = \sum_{i=0}^{t-1} \hat{\theta}_i$$

4. $\hat{\theta}^t$ noktasındaki kayıp fonksiyonun gradyan değerini hesapla

$$\nabla J(\theta) = \left[\frac{\partial J(\theta)}{J(\theta_i)} \right]_{\theta=\hat{\theta}^t}$$

5. Yeni artan parametre değerini hesapla

$$\hat{\theta}_t \leftarrow -\nabla J(\theta)$$

6. $\hat{\theta}_t$ değerini adım 3'e ekle.

Boosting yöntemleri ile makine öğrenmesi yöntemleri arasındaki temel fark, optimizasyon sürecinin fonksiyon uzayı üzerinde gerçekleşmesidir. Bu nedenle fonksiyon tahminleri toplanabilir formda Eşitlik (12)'de gibi tekrar tanımlanır.

$$\hat{f}(x) = \hat{f}^M(x) \sum_{i=0}^{M} \hat{f}_i(x) \tag{12}$$

Eşitlik (12)'de M iterasyon sayısını göstermektedir. Friedman (2001) temel öğrenici (base-learner) fonksiyonlara dayanan bir algoritma geliştirmiştir. Temel öğrenici fonksiyonlar $h(x,\theta)$ olarak tanımlanmaktadır. ρ ise optimal adım büyüklüğünü göstermektedir. Fonksiyon tahmini için optimizayon problemi Eşitlik (13)'de verilmiştir.

$$\begin{aligned} \hat{f}_t &\leftarrow \hat{f}_{t-1} + \rho_t h(x,\theta_t) \\ (\rho_t,\theta_t) &= \arg\min \sum_{i=1}^{N} \Psi\left(y_i, \hat{f}_{t-1}\right) + \rho h(x_i,\theta) \end{aligned} \tag{13}$$

Eşitlik (13)'de $\Psi(\cdot)$ kayıp fonksiyonu, $h(\cdot)$ ise temel öğrenici fonksiyonu göstermektedir. Sınıflandırma algoritmasında, GBM modelinde kayıp fonksiyon için adaboost kayıp fonksiyonu (Schapire, 2003) ya da binom kayıp fonksiyonu kullanılmaktadır. Temel öğrenici fonksiyonlar olarak ise radyal tabanlı fonksiyonlar veya p-splines fonksiyonlar tercih edilmektedir.

3. DENGESİZ VERİ SETLERİ İÇİN SMOTE ALGORİTMASI

Veri madenciliğinde, ilgilenilen örneklerin nadir olduğu dengesiz sınıflandırma sorunuyla karşılaşılmaktadır. Eğitim verisi içerisinde ilgilenilen olayın sıklığının çok az olması, sınıflandırıcıyı çoğunluk sınıfı örneklerinden çok fazla öğrenmesine neden olur. Model, azınlık sınıfı örnekleri için tanıma gücünden yoksun ve yetersiz uyuma sahip olacaktır (Li vd., 2017).

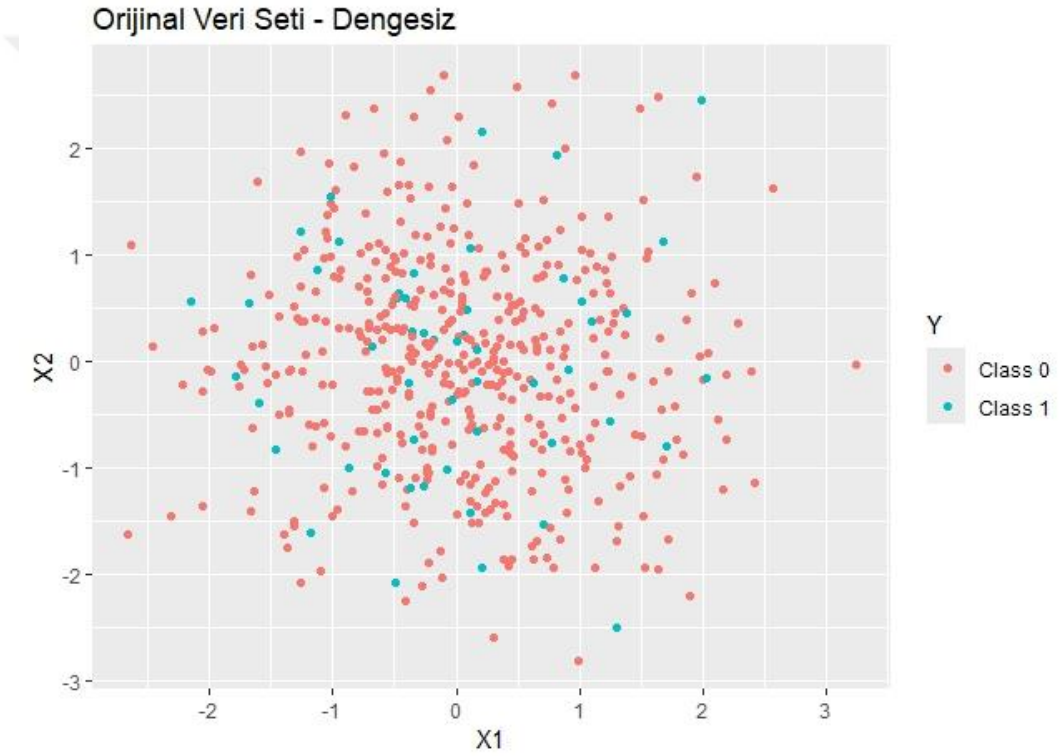
Sınıflayıcı üzerindeki çoğunluk sınıfının öğrenme sürecinde etkisini azaltmak için örnekleme yöntemleri önerilmiştir. Bu yöntemlerin temel amacı, dengesiz veri setini dengeli hale getirmek ve homojen bir dağılım elde etmektir. Azınlık sınıfı üzerinden yapılan örnekleme yöntemlerine over-sampling, çoğunluk sınıfı üzerinden yapılan örnekleme yöntemlerine ise under-sampling denir. Under-sampling yönteminde, çoğunluk sınıfındaki gözlem sayısı rastgele olarak azaltılır. Over-sampling yönteminde ise, azınlık sınıfındaki gözlemler çoğaltılarak dengeli bir sınıf dağılımı oluşturulmak amaçlanır.

Azınlık sınıfındaki gözlem sayısı oldukça düşük olduğunda aynı gözlem değerlerini kullanarak dengeli bir sınıf oluşturmak, sınıflandırma yöntemlerinin etkinliğini istenilen düzeyde arttırmaz. Çünkü azınlık sınıfı bu yöntemlerle istenilen düzeyde temsil edilemez (Ling ve Li, 1998; Japkowicz, 2000). Bu durum için sentetik veri üretme yöntemleri geliştirilmiştir. Chawla vd. (2002), SMOTE yöntemini önermişlerdir. Azınlık sınıfını over-sampling yaparak çoğaltmak yerine, azınlık sınıfına sentetik gözlemler eklemeye dayanmaktadır. Bu işlem yapılırken azınlık sınıfındaki gözleme ilişkin k-en yakın komşular belirlenir ve rastgele seçilir. Eğer azınlık sınıfı %200 oranında artırılacaksa, belirlenen k-en yakın komşulardan 2 tanesi seçilir. Bu yöntem sadece sürekli veriler üzerinde çalışmaktadır. SMOTE algoritması aşağıda verilmiştir (Chawla vd., 2002).

Adım 1: Azınlık sınıf kümesi A olsun. Her $x \in A$ için
x'in k-en yakın komşuları, x ile A kümesindeki her bir örnek arasındaki Öklid uzaklığı hesaplanarak belirlenir.
Adım 2: Örnekleme oranı N değerini belirle. Her $x \in A$ için
N örnek ($x_1, x_2, ..., x_N$) k-en yakın komşularından rastgele seçilir ve A_1 kümesini oluştururlar.
Adım 3: Her $x_k \in A_1, k = 1, 2, 3, ..., N$ için aşağıda eşitlik ile yeni örneklem üretilir.

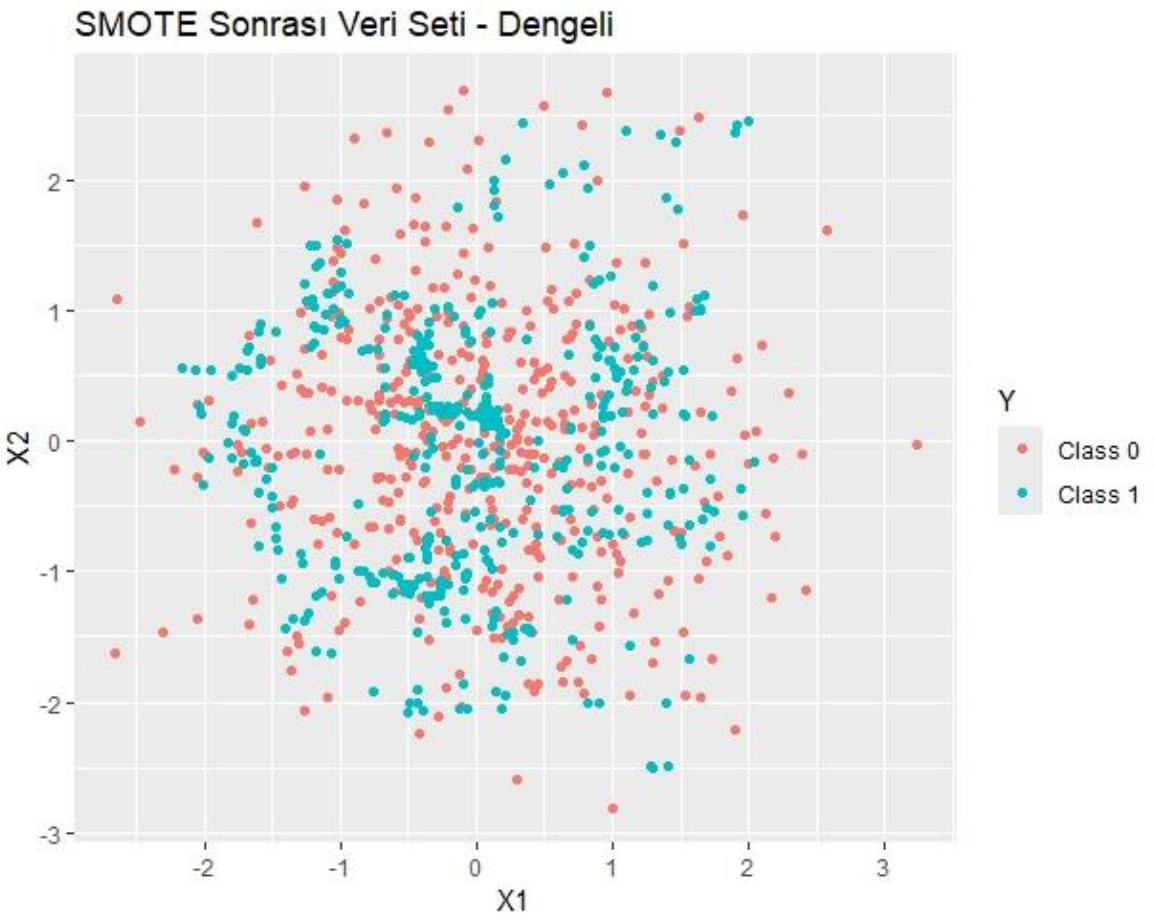
$x' = x + \text{rand}(0,1)|x - x_k|$. Burada $\text{rand}(0,1)$, 0 ile 1 arasında üretilen rastgele sayıyı ifade eder.

R programında SMOTE algoritmasına ilişkin bir benzetim gerçekleştirilmiştir. Bu amaçla, 2 bağımsız değişkenden 1 bağımlı değişkenden oluşan sentetik bir veri seti üretilmiştir. Bağımsız değişkenler normal dağılımdan, bağımlı değişken ise bernoulli dağılımdan üretilmiştir. Örneklem büyüklüğü n=500 olarak belirlenmiştir. Şekil 3.1'de benzetimle üretilen dengesiz veri setinin dağılımı verilmiştir. Burada dengesizlik oranı %10.4 olarak hesaplanmıştır.



Şekil 3.1: Simülasyonla üretilen dengesiz veri seti

Şekil 3.2'de SMOTE algoritması ile elde edilen dengeli veri setinin dağılımı verilmiştir. Azınlık sınıfının oranı %10.4'den %50'e çıkarılarak dengeli bir sınıf dağılımı elde edilmiştir. Burada, k değeri 5, örnekleme oranı N=100 olarak alınmıştır. Örnekleme oranının %100 alınması, azınlık sınıfı ile çoğunluk sınıfının gözlem sayılarının eşitlenmesi demektir.



Şekil 3.2: SMOTE algoritması sonrası elde edilen dengeli veri seti

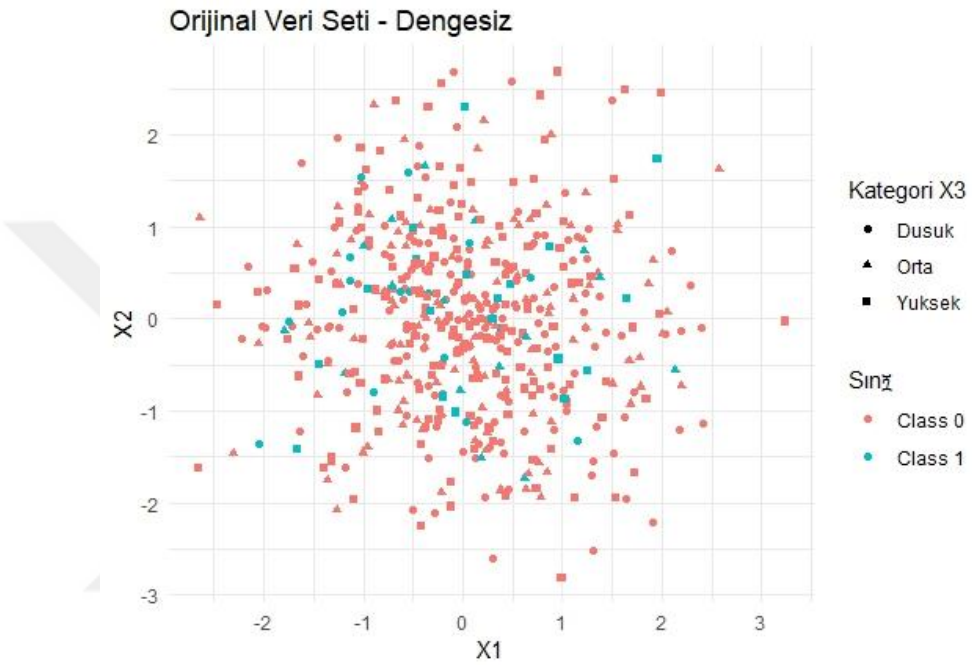
Hem sürekli hem de kategorik veriler üzerinde çalışan SMOTE-NC algoritması geliştirilmiştir (Wijaya vd., 2018). Bu algoritma modifiye edilmiş Öklid uzaklığına dayanmaktadır.

$$\Delta(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2+\sum_{j=1}^{q}\text{med}^2} \tag{14}$$

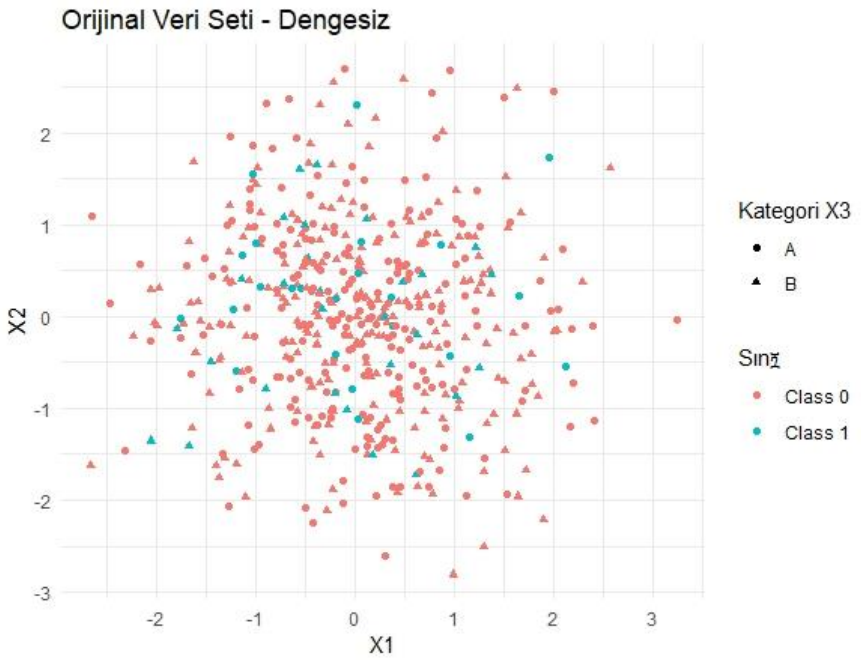
Eşitlik (14)'de $\Delta(x,y)$ gözlemler arasındaki uzaklığı, p sürekli değişkenlerin sayısını, q ise kategorik değişkenlerin sayısını göstermektedir. Burada med değeri ise azınlık sınıfındaki sürekli değişkenlerin standart sapmalarının medyan değeridir.

SMOTE-NC algoritmasının uygulaması R programında gerçekleştirilebilir. Bu amaçla, RSBID paketindeki SMOTE_NC fonksiyonu kullanılmıştır. SMOTE-NC algoritmasının test edilmesi için benzetim ile 2 sürekli, 2 kategorik değişkenden oluşan toplam 4 bağımsız

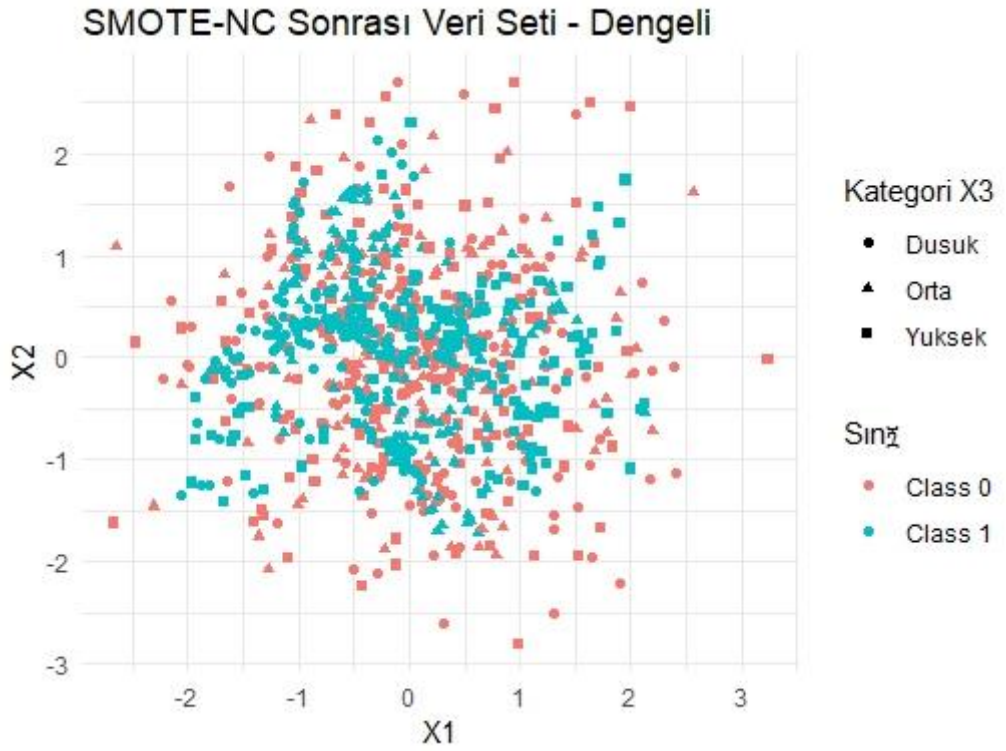
değişkenli bir veri seti üretilmiştir. Kategorik değişkenlerden biri 2 düzeyli, diğeri ise 3 düzeyli olarak üretilmiştir. Gözlem sayısı 500 olarak belirlenmiştir. Şekil 3.3 ve 3.4'de iki ve üç düzeyli kategorik değişkenler altında üretilen veri setinin görselleştirilmesi verilmiştir. Dengesizlik oranı yine %10.4 olarak belirlenmiştir. Dengeli veri setinin grafikleri ise Şekil 3.5 ve 3.6'da verilmiştir.



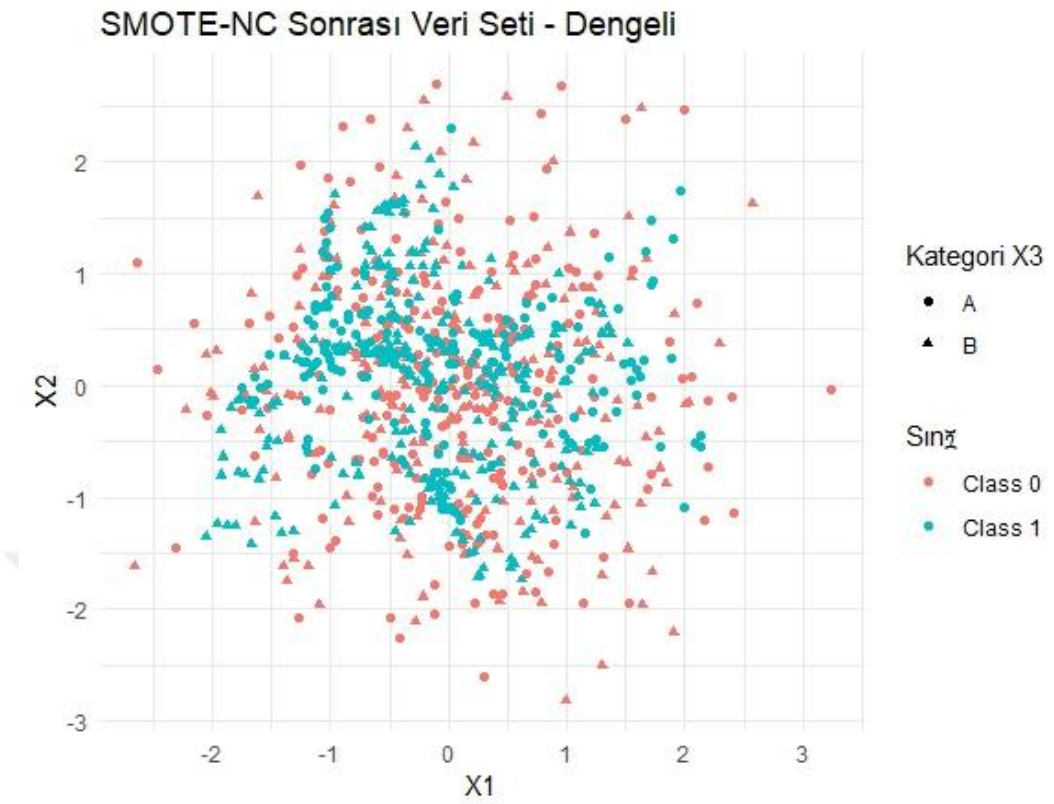
Şekil 3.3: Üç düzeyli kategorik değişken altında üretilen dengesiz veri setinin görselleştirmesi



Şekil 3.4: İki düzeyli kategorik değişken altında üretilen dengesiz veri setinin görselleştirmesi



Şekil 3.5: SMOTE-NC sonrası elde edilen dengeli veri setinin üç düzeyli kategorik değişken altında görselleştirmesi



Şekil 3.6: SMOTE-NC sonrası elde edilen dengeli veri setinin iki düzeyli kategorik değişken altında görselleştirmesi

4. UYGULAMA

Bu bölümde SMOTE-NC yönteminin dengesiz veri setlerindeki başarısı incelenecektir. SMOTE-NC algoritması ile elde edilen dengeli veri setleri kullanılarak, lojistik regresyon, DVM ve GBM modellerinin etkinlikleri karşılaştırılacaktır. Ayrıca modellerin SMOTE-NC ile üretilmiş dengeli veri seti ile orijinal veri seti üzerindeki başarıları da değerlendirilecektir.

4.1. Performans Ölçütleri

İkili sınıflandırma modellerinde, model başarılarının değerlendirilmesinde hata matrisi (confusion matrix) kullanılır. Hata matrisinin gösterimi Tablo 4.1'de verilmiştir.

Tablo 4.1: Hata matrisinin gösterimi

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

Tablo 4.1'de satırlarda gerçek durumlar, sütunlarda ise tahmin edilen durumlar yer almaktadır. TN, doğru tahmin edilen negatif gözlemleri, FN ise gerçekte pozitif olup negatif sınıflandırılan gözlemleri göstermektedir. TP doğru tahmin edilen pozitif gözlemleri FP ise gerçekte negatif olup pozitif olarak sınıflandırılan gözlemleri göstermektedir.

Doğru sınıflandırma oranı ise doğru tahmin edilen pozitif ve negatif gözlemlerinin toplam gözlem sayısına oranı olarak ifade edilir.

$$\text{Doğruluk} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

Hata oranı ise 1-Doğruluk olarak ifade edilmektedir. Hata oranı dengesiz veri setleri için iyi bir performans ölçütü değildir. Bu nedenle kesinlik (precision) ve duyarlılık (recall) değerleri kullanılır.

$$\begin{aligned}
&\text{Pozitif Tahmin Değeri}\left(\text{Kesinlik, Precision}\right) = \frac{TP}{TP+FP} \\
&\text{Duyarlılık}\left(\text{Recall,Sensitivity}\right) = \frac{TP}{TP+FN} \\
&\text{Seçicilik}\left(\text{Specificity}\right) = \frac{TN}{TN+FP} \\
&\text{Negatif Tahmin Değeri} = \frac{TN}{TN+FN}
\end{aligned} \tag{16}$$

Kesinlik, pozitif olarak sınıflandırılan durumların gerçekte kaçının pozitif olduğunu ölçer. Diğer bir ifadeyle pozitif tahmin değeridir. Duyarlılık ise gerçekte pozitif olan durumların yüzde ne kadarının pozitif olarak sınıflandırıldığını ölçer. Seçicilik ise gerçekte negatif olan durumların yüzde ne kadarının negatif olarak sınıflandırıldığını ölçer. Negatif tahmin değeri ise, negatif olarak tahmin edilen durumların yüzde kaçının gerçek negatif olduğunu gösterir.

Diğer bir ölçüt ise F1 skor değeridir. F1 skor değeri Eşitlik (17) ile hesaplanır.

$$F1 = 2\times\left(\frac{\text{Kesinlik}\times\text{Duyarlılık}}{\text{Kesinlik}+\text{Duyarlılık}}\right) \tag{17}$$

F1 skor değeri duyarlılık ve kesinlik değerlerinin harmonik ortalamasını göstermektedir.

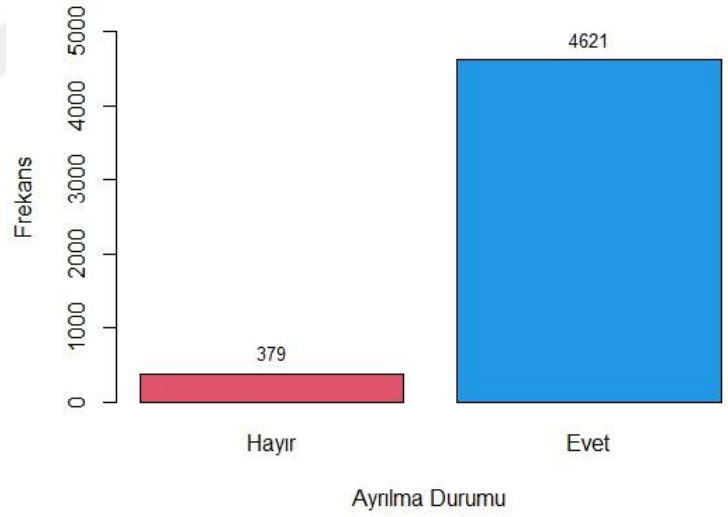
4.2. Veri

Kullanılan veri seti Kaggle platformundan alınmıştır. Veriye buradan ulaşabilirsiniz. Kullanılan değişkenler Tablo 4.2'de verilmiştir. Ayrılma değişkeni bağımlı değişken olup diğer 9 değişken ise bağımsız değişkendir.

Tablo 4.2: Kullanılan değişkenler ve tanımları

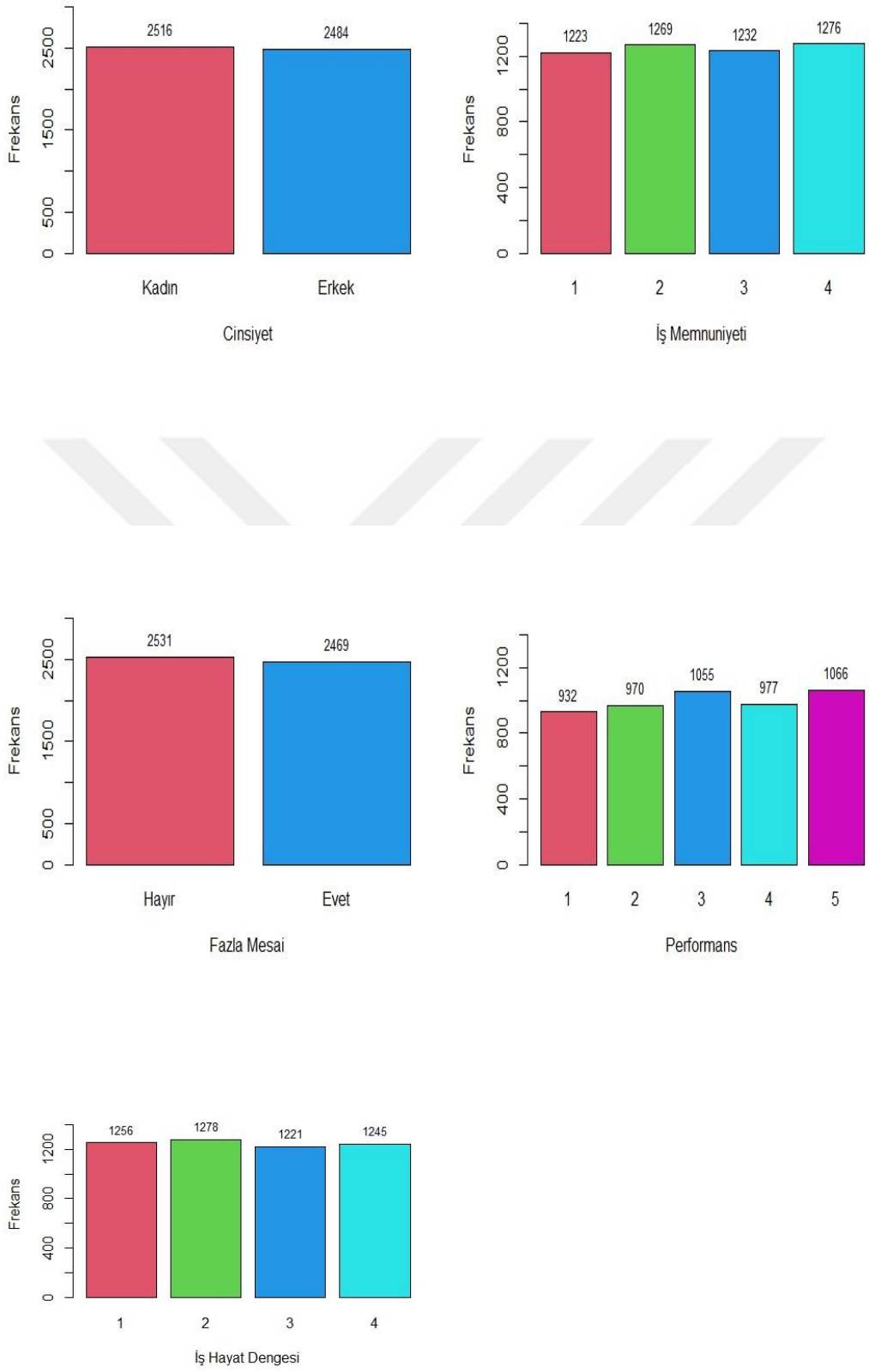
Değişkenler	Tanım Kümesi	Açıklama
Ayrılma	İkili (0/1)	Çalışanın şirketten ayrılıp ayrılmadığını gösterir
Yaş	R+	Çalışanın yaşı
Cinsiyet	Faktör (2 düzeyli)	Çalışanın cinsiyeti
Eve Uzaklık	R+	Ev ve iş yeri arası mesafe

İş Memnuniyeti	Faktör (5 düzeyli)	İş yerinden memnuniyeti gösterir
Aylık Gelir	R+	Çalışanın aylık geliri
Şirkette Çalışma Süresi	Z+	Çalışanın şirkette geçirdiği yıl sayısı
Fazla Mesai	İkili (0/1)	Çalışanın fazla mesai yapıp yapmadığını gösterir
Performans	Faktör (5 düzeyli)	Çalışanın ölçülen performans değeri
İş Hayat Dengesi	Faktör (5 düzeyli)	Çalışanın iş sorumluluklarını kişisel yaşamıyla ne kadar iyi dengelediğini gösterir



Şekil 4.1: Ayrılma değişkeninin dağılımı

Şekil 4.1'de ayrılma durumu değişkeninin dağılımı verilmiştir. Burada azınlık sınıfı 0 (hayır) kategorisidir. Toplam gözlem sayısı 5000 olup işten ayrılanların sayısı sadece 379'dur. Bu nedenle dengesiz bir sınıf dağılımı mevcuttur. Ayrılanların oranı sadece %7.58'dir. Şekil 4.2'de ise cinsiyet, iş memnuniyeti, fazla mesai, performans ve iş hayat dengesi değişkenlerinin grafikleri verilmiştir. İlgili değişkenlerin düzeyleri arasında dengeli bir dağılım olduğu dikkat çekmektedir.



Şekil 4.2: Cinsiyet, iş memnuniyeti, fazla mesai, performans ve iş hayat dengesi değişkenlerinin grafikleri

Tablo 4.3: Nicel değişkenlere ait betimsel istatistik değerleri

Betimsel İstatistikler	Yaş	Eve Uzaklık	Aylık Gelir	Şirkette Çalışma Süresi
Minimum	22	1	3000	0
Maksimum	59	29	7999	19
Değişim Aralığı	37	28	4999	19
Medyan	41	15	5444.5	10
Ortalama	40.700	15.239	5478.177	9.554
Standart Hata	0.155	0.118	20.455	0.082
Varyans	119.405	69.945	2091933	33.243
Standart Sapma	10.927	8.363	1446.351	5.766

Tablo 4.3'de nicel değişkenlere ilişkin betimsel istatistik değerleri verilmiştir. Ortalama yaş 40.7, eve uzaklık 15.239, gelir 5478.177 ve şirkette çalışma süresi 9.554 yıl olarak hesaplanmıştır.

4.3. Model Sonuçları

Bu bölümde LR, DVM ve GBM modellerinin performanslarını SMOTE-NC kullanılarak test edilecektir. Öncelikle veri eğitim ve test verisi olarak ikiye ayrılmıştır. Verinin %80'i eğitim verisi, %20'si ise test verisi olarak bölümlenmiştir. SMOTE-NC algoritması sadece eğitim verisi üzerinde uygulanmıştır. Test verisi üzerinde uygulanmamıştır.

Azınlık sınıfı 0 kategorisi olduğu için performans ölçütlerinden seçicilik ve negatif tahmin değeri bizim için önemlidir.

4.3.1. Lojistik Regresyon Sonuçları

Model sonuçları hem SMOTE-NC kullanılarak hem de kullanılmadan ayrı ayrı elde edilmiştir. Tablo 4.4'de SMOTE-NC kullanılmadan elde edilen sonuçlar verilmiştir. Tablo 4.4'de verilen sonuçlara göre 4 değişken anlamlı bulunmuştur. Bunlar yaş, eve uzaklık, aylık gelir ve şirkette çalışma süresidir. Parametre tahminleri yorumlanırken, odds oranları yorumlanır. Bu nedenle elde edilen parametre tahminlere üstel fonksiyon dönüşümü yapılır. Örneğin, diğer değişkenlerin etkisi sabit tutulduğunda, yaş 1 birim arttığında, kişinin işte ayrılması, ayrılmamasına göre $\exp(0.1071) = 1.113$ kat artar. Benzer şekilde diğer

parametre tahminleri de aşağıdaki gibi yorumlanabilir.

Eve uzaklık 1 birim arttığında, kişinin işten ayrılması, ayrılmamasına göre $\exp(0.2367) = 1.267$ kat artar.

Aylık gelir 1 birim arttığında, kişinin işten ayrılması, ayrılmamasına göre $\exp(-0.0002) = 0.9998$ kat artar. Burada odds değeri 1 değerinin altında olduğu için odds değerinin tersi yorumlanabilir. Kişinin işten ayrılmaması, ayrılmasına göre 1.0002 kat artar.

Şirkette çalışma süresi 1 birim arttığında, kişinin işten ayrılması, ayrılmamasına göre $\exp(-0.1424) = 0.867$ kat artar. Benzer şekilde, kişinin işten ayrılmaması ayrılmasına göre 1.153 kat artar.

Tablo 4.4: Lojistik regresyon model sonuçları

Değişkenler	Tahmin	Std. Hata	z-değeri	p-değeri
Sabit	42.24	2642.000	0.016	0.987
Yaş	0.1071	0.012	8.83	**<0.001**
Cinsiyet (Erkek)	0.4263	0.231	1.847	0.065
Eve Uzaklık	0.2367	0.019	12.408	**<0.001**
İş Memnuniyeti (2)	0.3069	2440.000	<0.001	0.9999
İş Memnuniyeti (3)	-24.08	1747.000	-0.014	0.9890
İş Memnuniyeti (4)	-24.01	1747.000	-0.014	0.9890
Aylık Gelir	-0.0002	<0.001	-2.806	**0.0050**
Şirkette Çalışma Süresi	-0.1424	0.022	-6.379	**<0.001**
Fazla Mesai	24.18	1223.000	0.02	0.9842
Performans (2)	0.1599	2768.000	<0.001	1.0000
Performans (3)	-23.86	1982.000	-0.012	0.9904
Performans (4)	-23.44	1982.000	-0.012	0.9906
Performans (5)	-23.71	1982.000	-0.012	0.9905
İş Hayat Dengesi (2)	-0.0955	0.329	-0.29	0.7714
İş Hayat Dengesi (3)	0.1145	0.324	0.354	0.7236
İş Hayat Dengesi (4)	-0.1763	0.319	-0.553	0.5806

Modelin test verisi üzerindeki sınıflandırma başarısı ise Tablo 4.5'de verilmiştir. Genel sınıflandırma başarısı %97.2 olarak hesaplanmıştır. Azınlık sınıfını doğru sınıflandırma başarısı ise %80'dir. Pozitif tahmin başarı oranı %82.1, negatif tahmin başarı oranı ise %98.3'dür.

Tablo 4.5: Lojistik regresyon sınıflandırma başarısı

Doğruluk	0.972
Duyarlılık	0.986
Seçicilik	0.800
Pozitif Tahmin Değeri	0.984
Negatif Tahmin Değeri	0.822

Lojistik regresyon modeli bu sefer, SMOTE-NC algoritması ile birlikte kullanılmıştır. Elde edilen sonuçlar Tablo 4.6'da verilmiştir. Elde edilen sonuçlara göre, yaş, eve uzaklık, aylık gelir, şirkette çalışma süresi ve iş hayat dengesi değişkenleri istatistiksel olarak anlamlı bulunmuştur.

Tablo 4.6: SMOTE-NC ile lojistik regresyon model sonuçları

Değişkenler	Tahmin	Std. Hata	z-değeri	p değeri
Sabit	41.900	1388.000	0.030	0.976
Yaş	0.143	0.011	13.172	**<0.001**
Cinsiyet (Erkek)	0.205	0.162	1.261	0.207
Eve Uzaklık	0.361	0.018	20.103	**<0.001**
İş Memnuniyeti (2)	0.259	1281.000	0.000	1.000
İş Memnuniyeti (3)	-27.000	913.600	-0.030	0.976
İş Memnuniyeti (4)	-26.850	913.600	-0.029	0.977
Aylık Gelir	0.000	**<0.001**	-3.384	**0.001**
Şirkette Çalışma Süresi	-0.139	0.018	-7.932	**<0.001**
Fazla Mesai	27.040	645.000	0.042	0.967
Performans (2)	0.358	1452.000	**<0.001**	1.000
Performans (3)	-26.670	1044.000	-0.026	0.980
Performans (4)	-26.030	1044.000	-0.025	0.980
Performans (5)	-26.380	1044.000	-0.025	0.980

İş Hayat Dengesi (2)	0.079	0.226	0.352	0.725
İş Hayat Dengesi (3)	0.458	0.222	2.063	**0.039**
İş Hayat Dengesi (4)	-0.187	0.219	-0.854	0.393

SMOTE-NC ile elde edilen lojistik regresyon modelinin sınıflandırma başarısı ise Tablo 4.7'de verilmiştir. Seçicilik değeri yükselmiştir. Fakat, negatif tahmin değeri düşmüştür.

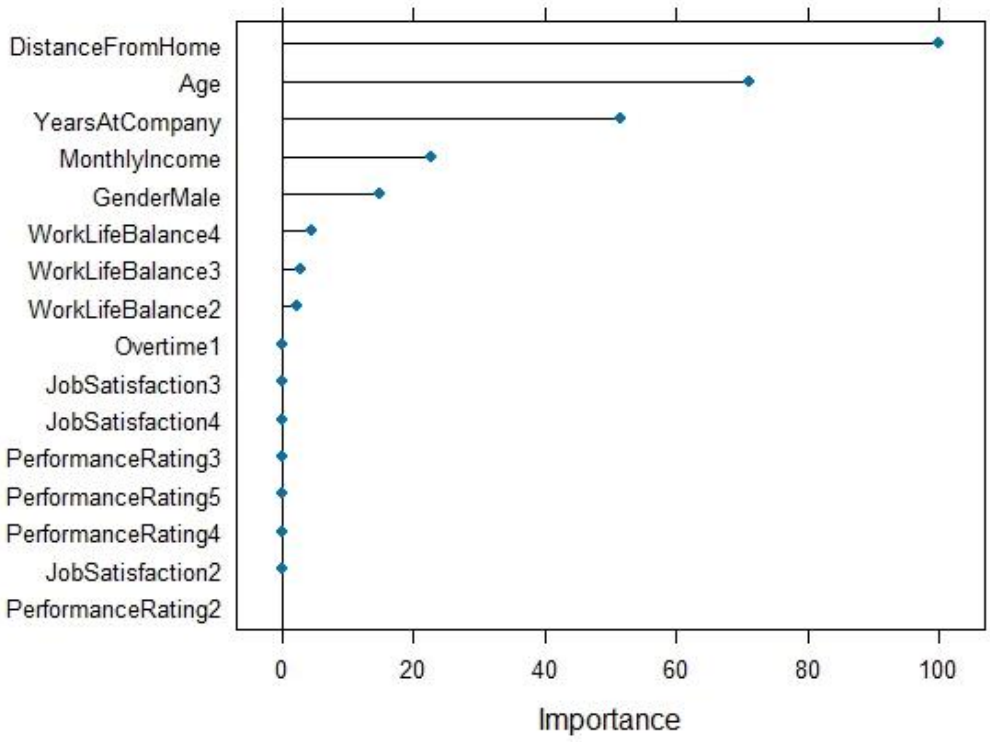
Tablo 4.7: SMOTE-NC ile lojistik regresyon sınıflandırma başarısı

Doğruluk	0.959
Duyarlılık	0.957
Seçicilik	0.973
Pozitif Tahmin Değeri	0.997
Negatif Tahmin Değeri	0.651

İki modelden elde edilen hata matrisleri ise Tablo 4.8'de verilmiştir. Tablo 4.8'de görüldüğü gibi SMOTE-NC modeli, azınlık sınıfının doğru sınıflandırma başarısını artırmıştır.

Tablo 4.8: Lojistik regresyon modellerinin hata matrisleri

SMOTE-NC LOJİSTİK REGRESYON			LOJİSTİK REGRESYON		
Tahmin	Gerçek Durum		Tahmin	Gerçek Durum	
	Hayır	Evet		Hayır	Evet
Hayır	73	39	Hayır	60	13
Evet	2	885	Evet	15	911



Şekil 4.3: Lojistik regresyon modeli için değişken önem grafiği

Şekil 4.3'de lojistik regresyon modeli için en önemli değişkenler verilmiştir. Buna göre, evin işyerine uzaklığı, yaş ve şirketteki çalışma süresi en önemli ilk 3 değişken olarak belirlenmiştir.

4.3.2. DVM Sonuçları

DVM modeli için radyal tabanlı kernel fonksiyonu kullanılmıştır. Elde edilen modelin sınıflandırma başarısı Tablo 4.9'da verilmiştir.

Tablo 4.9: DVM model sonuçları

Doğruluk	0.980
Duyarlılık	0.990
Seçicilik	0.853
Pozitif Tahmin Değeri	0.988
Negatif Tahmin Değeri	0.876

Tablo 4.10'da ise SMOTE-NC kullanılarak elde edilen DVM model sonuçları verilmiştir.

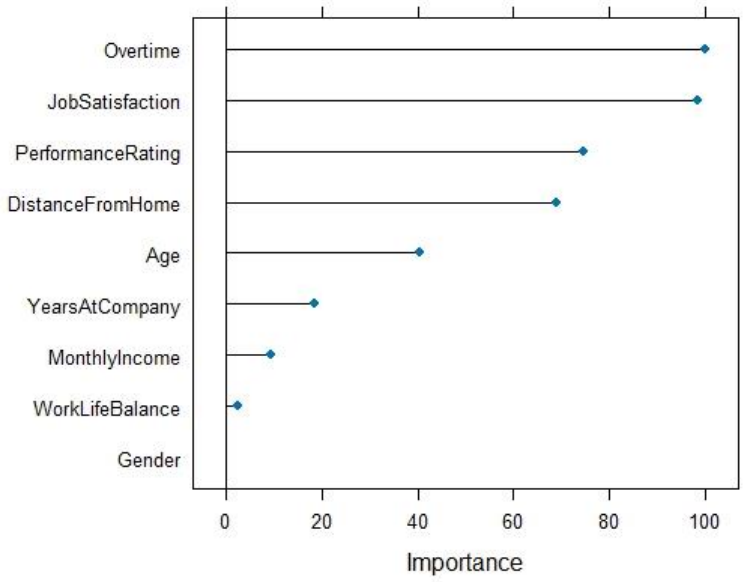
Tablo 4.10:. SMOTE-NC ile DVM model sonuçları

Doğruluk	0.968
Duyarlılık	0.965
Seçicilik	1
Pozitif Tahmin Değeri	1
Negatif Tahmin Değeri	0.701

SMOTE-NC kullanılarak eğitilen DVM modelinin seçicilik değerinin oldukça yüksek olduğu görülmektedir. Ancak negatif tahmin değeri düşmüştür. Her iki durum için de elde edilen hata matrisleri ise Tablo 4.11'de verilmiştir.

Tablo 4.11: SVM modellerinin hata matrisleri

SMOTE-NC SVM			SVM		
Tahmin	Gerçek Durum		Tahmin	Gerçek Durum	
	Hayır	Evet		Hayır	Evet
Hayır	75	32	Hayır	64	9
Evet	0	892	Evet	11	915



Şekil 4.4: DVM modeli için değişken önem grafiği

Şekil 4.4'de DVM modeli için en önemli değişkenler verilmiştir. Buna göre, fazla mesai, iş tatmini ve performans en önemli ilk 3 değişken olarak belirlenmiştir.

4.3.3. GBM Sonuçları

GBM modelinin uygulaması R programında caret paketi ile yapılmıştır. Tablo 4.12'de SMOTE-NC yöntemi olmadan GBM sonuçları verilmiştir. Modelin doğruluk, duyarlılık ve seçicilik değeri %100 olarak elde edilmiştir. Tablo 4.13'de ise SMOTE-NC kullanılarak elde edilen GBM sonuçları verilmiştir. Tüm değerlendirme kriterlerine göre model performansı %100 olarak elde edilmiştir.

Tablo 4.12: GBM model sonuçları

Doğruluk	1
Duyarlılık	1
Seçicilik	1
Pozitif Tahmin Değeri	1
Negatif Tahmin Değeri	1

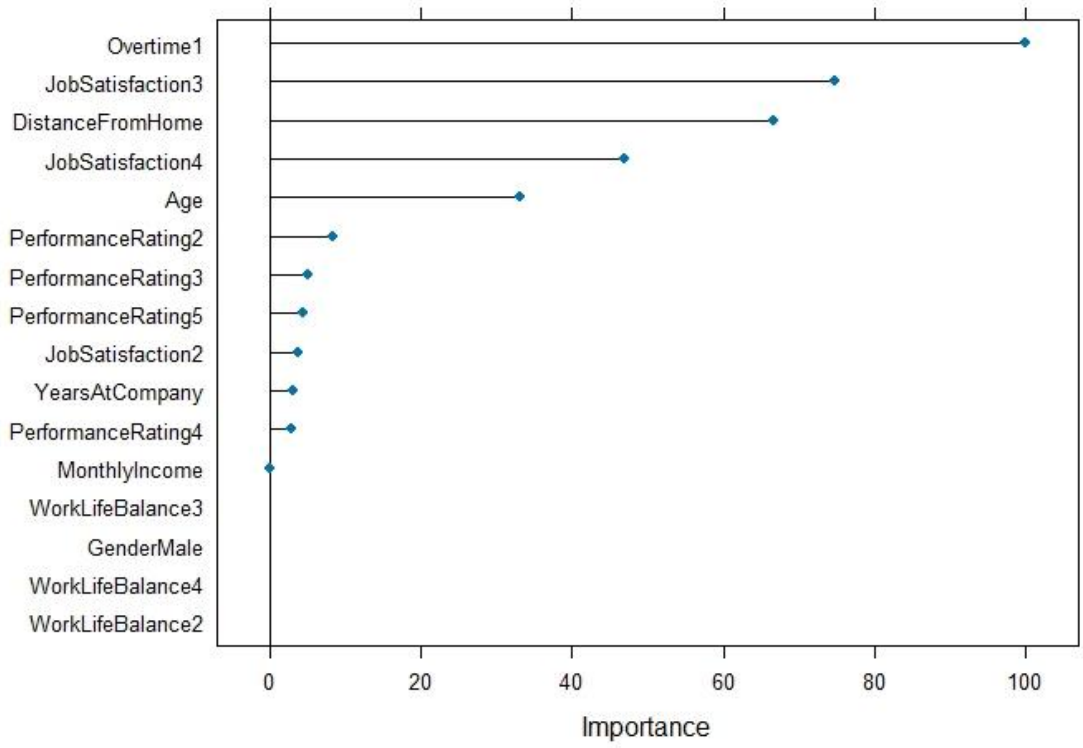
Tablo 4.13: SMOTE-NC ile GBM model sonuçları

Doğruluk	1
Duyarlılık	1
Seçicilik	1
Pozitif Tahmin Değeri	1
Negatif Tahmin Değeri	1

Tablo 4.14'de ise hata matrisleri verilmiştir. Tablo 4.14'de görüldüğü gibi hatalı sınıflandırma oranı GBM modeli için her iki durumda da %0 olarak elde edilmiştir.

Tablo 4.14: GBM modellerinin hata matrisleri

SMOTE-NC GBM			GBM		
Tahmin	Gerçek Durum		Tahmin	Gerçek Durum	
	Hayır	Evet		Hayır	Evet
Hayır	75	0	Hayır	75	0
Evet	0	924	Evet	0	924



Şekil 4.5: GBM modeli için değişken önem grafiği

Şekil 4.5'de GBM modeli için en önemli değişkenler verilmiştir. Buna göre, fazla mesai, iş tatmini ve iş yerinin eve uzaklığı en önemli ilk 3 değişken olarak belirlenmiştir.

4.3.4. En İyi Modelin Belirlenmesi

En iyi modelin belirlenmesinde F1 skor değeri kullanılmıştır. SMOTE-NC algoritmasından elde edilen eğitim verisi üzerinde yapılan model tahminlerine göre hesaplanan F1 değerleri Tablo 4.15'de verilmiştir. Tablo 4.15'de verilen sonuçlara göre en iyi model GBM modeli olarak belirlenmiştir. DVM modeli LR modelinden daha iyi sonuçlar vermiştir. F1 değerleri SMOTE-NC ile eğitilen modellerde daha düşük olmuştur. Bunun nedeni, seçicilik değeri

artmasına rağmen negatif tahmin değerinin azalmasıdır. Bu durumu daha iyi anlamak için benzetim çalışması yapılmıştır.

Tablo 4.15: Modellerin F1 skor değerleri

Modeller	SMOTE-NC	
	Yok	Var
	F1-Skor	F1-Skor
LR	0.811	0.781
DVM	0.864	0.824
GBM	1	1

5. BENZETİM ÇALIŞMASI

Bu bölümde, LR, DVM ve GBM modellerinin dengesiz sınıf dağılımında modelleme başarısı benzetim çalışması ile karşılaştırılmıştır. Dengesizlik oranı %10 olarak belirlenmiştir. Sonuçlar, duyarlılık ve pozitif tahmin değerleri üzerinden yorumlanmıştır. Bunun nedeni, dengesiz sınıf dağılımında doğruluk ve seçicilik değerleri azınlık sınıfın sınıflandırması başarısı çok düşük ya da 0 bile olsa yüksek çıkmasıdır. Amaç, azınlık sınıfının yüksek düzeyde doğru sınıflandırılmasının sağlanmasıdır.

Benzetim çalışmasında iki bağımsız değişken standart normal dağılımdan üretilmiştir. Bağımlı değişken ise bernoulli dağılımından üretilmiştir. Örneklem büyüklüğü n=1000 olarak belirlenmiştir. Her bir örneklem %80 eğitim ve %20 test verisi olmak üzere iki parçaya bölünmüştür. LR, DVM ve GBM modelleri, hem SMOTE-NC hem de SMOTE-NC olmadan eğitilmiştir. Elde edilen sonuçlar Tablo 5.1'de verilmiştir.

SMOTE-NC olmadan eğitilen modellerin hepsi, azınlık sınıfının sınıflandırılmasında başarısız olmuştur. Tüm pozitif ve negatif değerleri, negatif olarak sınıflandırmışlardır. Bu nedenle pozitif tahmin değerleri hesaplanamamıştır. SMOTE-NC algoritması ile eğitilen modellerde ise doğruluk değeri düşmüş, ancak duyarlık ve pozitif tahmin değerleri artmıştır. Diğer bir ifadeyle, azınlık sınıfının, sınıflandırma başarısı yükselmiştir. Burada, SMOTE-NC modeli ile en iyi sonucu GBM modeli vermiştir.

Tablo 5.1: LR, DVM ve GBM modellerinin benzetim sonuçları

		Ölçütler	Modeller		
			LR	GBM	DVM
SMOTE-NC	Hayır	Doğruluk	0.900	0.900	0.900
		Duyarlılık	0	0	0
		Seçicilik	1	1	1
		Pozitif Tahmin Değeri	-	-	-
		Negatif Tahmin Değeri	0.900	0.900	0.900
	Evet	Doğruluk	0.535	0.645	0.52
		Duyarlılık	0.300	0.350	0.30
		Seçicilik	0.561	0.677	0.544

		Pozitif Tahmin Değeri	0.070	0.108	0.068
		Negatif Tahmin Değeri	0.878	0.9037	0.875

İkinci benzetim çalışmasında ise bağımlı değişken için veri üretme süreci değiştirilmiştir. Toplam gözlemlerin yarısı bernoulli dağılımından diğer yarısı lojistik fonksiyon kullanılarak, lojistik regresyondan üretilmiştir. Lojistik regresyon modeli için parametre değerleri $\beta_0 = 2,\ \beta_1 = 0.5$ ve $\beta_2 = -0.5$ değerleri kullanılmıştır. Dengesizlik oranı %10 korunmuştur. Sonuçlar Tablo 5.2'de verilmiştir.

Tablo 5.2'de verilen değerlere göre LR ve GBM modellerinin duyarlılık ve pozitif tahmin değerleri birbirine yakınken DVM modelinin azınlık sınıfını doğru sınıflandırma başarısı %0 olmuştur. SMOTE-NC algoritması kullanılarak elde edilen sonuçlar incelendiğinde, duyarlılık değeri en yüksek olan model DVM, pozitif tahmin değeri en yüksek olan model ise LR modeli olmuştur.

Tablo 5.2: LR, DVM ve GBM modellerinin farklı veri üretme süreci için benzetim sonuçları

		Ölçütler	Modeller		
			LR	GBM	DVM
SMOTE-NC	Hayır	Doğruluk	0.805	0.795	0.830
		Duyarlılık	0.088	0.117	0
		Seçicilik	0.951	0.933	1
		Pozitif Tahmin Değeri	0.272	0.266	-
		Negatif Tahmin Değeri	0.835	0.837	0.830
	Evet	Doğruluk	0.600	0.585	0.575
		Duyarlılık	0.588	0.588	0.617
		Seçicilik	0.600	0.584	0.566
		Pozitif Tahmin Değeri	0.232	0.224	0.225
		Negatif Tahmin Değeri	0.8772	0.873	0.878

6. SONUÇ VE TARTIŞMA

Çalışmada LR, DVM ve GBM modelleri ele alınmıştır. Bu modellerin dengesiz veri setlerindeki başarıları incelenmiştir. Uygulama çalışmasında Kaggle platformunda yer alan veri seti kullanılmıştır. Dengesiz sınıf dağılımında örnekleme yöntemleri kullanılarak dengeli bir sınıf dağılımı elde edilmesi amaçlanmıştır. Hem sürekli ve hem de kategorik veriler üzerinde çalışabilen SMOTE-NC algoritması tercih edilmiştir. Modeller hem orijinal eğitim verisi ile hem de SMOTE-NC ile elde edilen veri ile eğitilmiştir.

GBM modelinin gerçek veri seti üzerindeki başarısı hem LR hem de DVM modellerinden oldukça yüksektir. Lojistik regresyon modelinin test verisi üzerindeki başarısı DVM modelinden daha düşük olarak elde edilmiştir. Her iki model içinde SMOTE-NC yöntemiyle elde edilen sonuçlar incelendiğinde, SMOTE-NC yöntemi, DVM modelinin sınıflandırma başarısını, lojistik regresyon modeline göre, daha çok iyileştirmiştir. Her üç modelin başarısı F1 skor değerlerine göre incelendiğinde, en iyi model GBM modelidir.

SMOTE-NC yöntemi, dengesiz veri setlerindeki sentetik gözlem üretme başarısı, makine öğrenmesi yöntemlerinin aşırı öğrenme ya da eksik öğrenme problemlerini de ortadan kaldırmaktadır. Elde edilen sonuçların genelleştirilmesi açısından farklı dengesiz sınıf dağılımları kullanılarak üretilen veri setlerinde gerçekleştirilecek benzetim çalışması önem taşımaktadır. Yapılan benzetim çalışması sonucunda, tamamen rastgele üretilen sınıf dağılımı altında en başarılı sonucu SMOTE-NC GBM modeli vermiştir.

KAYNAKLAR

Aktümsek, E ve Göker, İ. E. K. (2018). Mali başarısızlık tahminlemesinde sektör bazlı bir karşılaştırma. *İşletme Araştırmaları Dergisi*, 10(4): 401-421.

Atasoy, N. A., & Demiröz, A. (2021). Makine öğrenmesi algoritmaları kullanılarak prostat kanseri tümör oluşumunun incelenmesi. *Avrupa Bilim ve Teknoloji Dergisi*, 29: 87-92.

Awad, M. ve Khanna, R. (2015). *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Germany: Springer Nature.

Ayan, T. Y. Ve Değirmenci, N. (2018). Firma Finansal Başarısızlık Öngörüsü için Bir Lojistik Regresyon Modeli. *Uluslararası İktisadi ve İdari İncelemeler Dergisi*, 77-88.

Ayhan, S. ve Erdoğmuş, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 9(1): 175-201.

Bircan, H. (2004). Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. *Kocaeli Üniversitesi Sosyal Bilimler Dergisi*, 8:185-208.

Cateni, S., Colla, V. ve Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135: 32-41.

Chawla, N. V., Bowyer, K. W., Hall, L. O. ve Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321-357.

Chiang, J. Y., Lio, Y., Hsu, C. Y., Ho, C. L. ve Tsai, T. R. (2023). Binary Classification with Imbalanced Data. *Entropy*, 26(1): 1-15.

Estabrooks, A., Jo, T. ve Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1): 18-36.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5): 1189-1232.

Gök, E. C. ve Olgun, M. O. (2021). SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. *Neural Computing and Applications*, 33(22): 15693-15707.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. *In Proceeding of the International Conference. on Artificial Intelligence*, 56: 111-117.

Joloudari, J. H., Marefat, A., Nematollahi, M. A., Oyelere, S. S ve Hussain, S. (2023). Effective class-imbalance learning based on SMOTE and convolutional neural

networks. *Applied Sciences*, 13(6): 1:34

Kavzoğlu, T. ve Çölkesen, İ. (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. *Harita Dergisi*, 144(7): 73-82.

Küçüksille, E. U. Ve Ateş, N. (2013). Destek vektör makineleri ile yaramaz elektronik postaların filtrelenmesi. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 6(1): 1-7.

Li, J., Fong, S., Hu, S., Chu, V. W., Wong, R. K., Mohammed, S. ve Dey, N. Rare event prediction using similarity majority under-sampling technique. In Soft Computing in Data Science: Third International Conference, SCDS 2017, Yogyakarta, Indonesia, November 27–28, 2017, 23-29

Li, P., Chan, K. L ve Fang, W. Hybrid kernel machine ensemble for imbalanced data sets. In 18th International Conference on Pattern Recognition, Hong Kong, 20-24 August 2006, 1-4.

Ling, C. X. ve Li, C. (1998, August). Data mining for direct marketing: Problems and solutions. In Proceedings of the fourth international conference on knowledge discovery and data mining. NY, USA, August 1998, 73-79.

Natekin, A. ve Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7(21):1-21

Oğuzlar, A. (2005). Lojistik regresyon analizi yardımıyla suçlu profilinin belirlenmesi. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 19(1): 21-35.

Rahmayanti, I. A., Saifudin, T. ve Ana, E. (2021). Applyıng Smote-Nc on Cart Algorıthm To Handle Imbalanced Data In Customer Churn Predıctıon: A Case Study of Telecommunıcatıons Industry. *Journal of Syntax Literate*, 6: 1321-1337.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*, New York: Springer.

Seiffert, C., Khoshgoftaar, T. M. ve Van Hulse, J. (2009). Hybrid sampling for imbalanced data. *Integrated Computer-Aided Engineering*, 16(3): 193-210.

Soler, V. ve Prim, M. (2007). Rectangular basis functions applied to imbalanced datasets. In Artificial Neural Networks–ICANN 2007: 17th International Conference, Porto, Portugal, September 9-13, 2007, Proceedings, Part I 17 (pp. 511-519). Springer Berlin Heidelberg.

Tütüncü, T. E. ve Gürsakal, S. (2023). Kredi Temerrüt Riskini Tahmin Etmede Makine Öğrenme Algoritmalarının Karşılaştırılması. Avrupa *Bilim ve Teknoloji Dergisi*, 50: 14-22.

Ural, K., Gürarda, Ş. ve Önemli, M. B. (2015). Lojistik regresyon modeli ile finansal

başarısızlık tahminlemesi: Borsa İstanbul'da faaliyet gösteren gıda, içki ve tütün şirketlerinde uygulama. *Muhasebe ve Finansman Dergisi*, 67: 85-100.

Üstüner, M., Abdikan, S., Bilgin, G. ve Şanlı, F. B. (2020). Hafif gradyan artırma makineleri ile tarımsal ürünlerin sınıflandırılması. *Türk Uzaktan Algılama ve CBS Dergisi*, 1(2): 97-105.

Vatansever, K. (2014). Finansal başarısızlığın öngörülmesinde çok kriterli karar verme analizine dayalı bir araştırma. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi*, 41: 163-176

Wang, H. Y. Combination approach of SMOTE and biased-SVM for imbalanced datasets. In 2008 IEEE international joint conference on neural networks, China, 1-8 June 2008, 228-231.

Wijaya, J., Soleh, A. M. ve Rizki, A. (2018). Penanganan data tidak seimbang pada pemodelan Rotation Forest keberhasilan studi mahasiswa Program Magister IPB. *Xplore: Journal of Statistics*, 2(2): 32-40.

Yakut, E., Elmas, B. & Yavuz, S. (2014). Yapay Sinir Ağları Ve Destek Vektör Makineleri Yöntemleriyle Borsa Endeksi Tahmini. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 19(1): 139-157.

ÖZGEÇMİŞ