Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 27, Sayı 2, 181-190, 2023

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 27, Issue 2, 181-190, 2023

DOI: 10.19113/sdufenbed.1090596

Dengesiz Sınıf Dağılımında Kayıp Gözlem Sorunu için Topluluk Öğrenmesi Sonuçlarının İstatistiksel Değerlendirmesi

Enis GÜMÜŞTAŞ[1], Ayça ÇAKMAK PEHLİVANLI[*2]

[1,2]Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, 34380, İstanbul, Türkiye

(Alınış / Received: 20.03.2022, Kabul / Accepted: 07.12.2022, Online Yayınlanma / Published Online: 25.08.2023)

Anahtar Kelimeler
Kayıp Veri Atama,
Dengesiz Sınıf Öğrenmesi,
Topluluk Öğrenmesi,
İstatistiksel Değerlendirme,
Wilcoxon Sıra Testi,
XGBoost

Öz: Son yıllarda gelişen teknoloji sürekli akan, farklı yapılarda ve yüksek boyutlarda verileri de beraberinde getirmiştir. Bu hızlı değişim ve veri setlerinde rastlanan problemler özellikle geleneksel yöntemleri bir noktadan sonra yetersiz bırakmaktadır. Bu çalışma kapsamında iki önemli veri problemi ele alınmıştır: i) kayıp gözlem içeren veri setleri ve ii) dengesiz sınıf dağılımı içeren veri setleri. Bu çalışmanın amacı aynı anda hem kayıp gözlem hem de dengesiz sınıf dağılımı sorununa sahip veri setlerini çeşitli kayıp gözlem atama yöntemleri kullanarak doldurmak ve elde edilen veri üzerinde topluluk öğrenme algoritmalarının başarı düzeylerini değerlendirmektir. Uygulama için sensörler aracılığıyla toplanan veri setinde eğitim için 59000 gözlemden oluşan negatif sınıfa karşılık 1000 adet pozitif sınıfa ait gözlem bulunmaktadır. Elde edilen modeller %2.4 oranında dengesiz sınıf dağılımına sahip sınama verisi ile sınanmıştır. Ayrıca veri setinde bulunan değişkenlerin yaklaşık %99'unda %82'ye varan kayıp veri söz konusudur. Bu kayıp gözlemler sıcak deste ataması, ortalama, ortanca, tepe değeri, çoklu atama, beklenti en büyükleme ve k en yakın komşu yöntemleri ile giderilmeye çalışılmıştır. Atama metodu ile eksik veri tamamlaması yapılan veri setleri Extra Trees, Random Forest, Gradient Boosting, LightGBM ve XGBoost gibi algoritmalar ile karşılaştırmalı sınanmış, en iyi sonuç XGBoost algoritması ile elde edilmiştir.

Statistical Evaluation of Ensemble Learning Outcomes for Missing Value Problem in Imbalanced Class Distribution

Keywords
Missing Data Imputation,
Class Imbalanced Learning,
Ensemble Learning,
Statistical Evaluation,
Wilcoxon Rank Test
XGBoost

Abstract: Rapid developments in technology have brought data in different structures and high dimensions in recent decades. Due to this rapid changes and problems encounters in data sets, it has been inevitable that traditional methods replaced with machine learning methods. Within the range of this study, two important data problems are discussed: data sets with i) missing observations and ii) imbalanced class distribution. This study aims to fill the datasets that have both missing observation and imbalanced class distribution problems at the same time by using various missing observation imputation methods and to assess the success levels of ensemble learning algorithms on the obtained data. In the data set collected through sensors for the application, there are 59000 observations for training versus 1000 positive observations for the negative class. The models obtained were tested with the data with an imbalanced class distribution of 2.4%. In addition, approximately 99% of the features in the data set have missing data up to 82%. These missing observations have been tried to be eliminated by hot deck imputation, mean, median, mode, multiple imputation, expectation maximization, and *k* nearest neighbour methods. Datasets completed with the imputation methods were comparatively tested with algorithms such as Extra Trees, Random Forest, Gradient Boosting, LightGBM, and XGBoost, and the most promising result was obtained with the XGBoost algorithm.

*İlgili yazar: ayca.pehlivanli@msgsu.edu.tr

1. Giriş

Dengesiz sınıf dağılımı ve kayıp gözlemlerin veri setinde bulunması özellikle sınıflama problemlerinde çok önemli sorunlara ve zorluklara neden olmaktadır. Özellikle teknolojik alandaki hızlı gelişime bağlı olarak üretilen verinin miktarındaki artış ile birlikte veri yapıları da değişime uğramıştır. Bu değişimin etkisi veri kaynaklarındaki çeşitliliği ön plana çıkarmaya başlamıştır. Dolayısı ile büyük çaplı akıllı uygulamalar, birbirleri ile etkileşime girebilen elektronik cihazlar, sosyal medya gibi kaynaklardan elde edilen veri setlerini analiz etmek için bilinen yaklaşımlar tek başına yetersiz kalmış, istatistik tabanlı makine öğrenmesi yaklaşımları son derece yaygın hale gelmiştir.

Veri toplama sırasında karşılaşılaşılabilen sistematik veya yazılım kaynaklı sorunlar, verilerin bozulmasına ya da kaydedilememesine neden olarak kayıp/eksik veri sorunu ortaya çıkarır. Veri analizi sürecinin en gerekli ve etkili aşaması veri ön işlemedir. Kayıp/eksik veri sorunu özellikle bu aşamada belirlenip çözülmelidir. Kayıp verileri oluşum nedenlerine göre gruplayan kayıp veri mekanizmalarına ilişkin ilk çalışma 1976'da Rubin tarafından gerçekleştirilmiştir [1]. Bu çalışmayı 1977 yılında Dempster, Laird ve Rubin'in en çok olabilirlik yöntemlerini adımsal olarak kullanan beklenti en büyükleme (expectation maximization) yaklaşımını sundukları çalışma takip etmiştir [2]. 1988'de Little, tüm veri setini kullanan çok değişkenli bir yaklaşım önermiştir [3]. Buna göre, veri setindeki her bir değişken için ortalama farkları eşzamanlı hesaplayan *t*-testine benzer yaklaşım ile kayıp gözlem mekanizması belirlenip kayıp veri ataması için kullanılacak yöntemler seçilebilmektedir [3].

Veri setinde bulunan sınıfların dağılımları arasında önemli ölçüde farklılık olması dengesiz veri problemini ortaya çıkarır. Sınıf dengesizliği sorunu, kredi kartı sahtekarlığı [4, 5], banka müşterilerinin ödeme alışkanlıkları [6] gibi finans alanında yapılan çalışmalarda, kanser [7], diyabet teşhisi [8] gibi sağlık alanında yapılan çalışmalarda, doğal dil işleme [9], uydu görüntülerinden nesne tanıma [10] gibi çeşitli alanlarda karşılaşılan yaygın bir sorundur. Dengesiz sınıf dağılımı içeren veri setleri üzerinde yapılan çalışmalar genel olarak üç grupta değerlendirilebilir; i) veri düzeyinde (data-level), ii) algoritma düzeyinde (algorithm-level), iii) maliyete duyarlı (cost-sensetive). Dengesiz sınıf dağılımını ortadan kaldırmak için veri düzeyinde değerlendirilen yöntemlerden rassal az örnekleme (random under sampling), rassal aşırı örnekleme (random over sampling) ve sentetik veri üretme en yaygın olanlarıdır. Rassal az örneklemede çoğunluk olan sınıflardan azınlık olan sınıftaki gözlem sayısı kadar örnek rasgele çekilerek sınıflara ait gözlem sayıları eşitlenir. Rassal aşırı örnekleme yaklaşımında ise bu işlem tersine çevrilir. Sınıflara ait gözlem sayılarını çoğunluk sınıf sayısına eşitlemek adına, azınlık sınıftan örnek seçimi rastgele ve iadeli olarak gerçekleştirilir. Ancak kullanılan örnekleme yöntemlerinin, çoğunluk sınıfta olup eğitime alınmayan bazı etkili gözlemlerin kaybedilmesi, azınlık sınıfının sayısının artırılması ile elde edilen veri setindeki artış ile eğitim süresinin artması ve beraberinde aşırı öğrenmeye neden olması gibi olumsuz yan etkilerine de dikkat etmek gerekir [11, 12]. Veri düzeyinde değerlendirilen son yaklaşım olan sentetik veri üretme, veri dağılımını temel alarak benzer gözlemler üretir. Bu yaklaşımı kullanan en bilinen yöntem sentetik azınlık aşırı örnekleme yöntemi (SMOTE-Synthetic Minority Oversampling Technique) Chawla ve diğ. tarafından önerilmiştir [11]. Dengesiz sınıf dağılım problemlerinde azınlık olan sınıftaki gözlem sayısının sentetik olarak çoğaltılması temeline dayanan bu yöntemin farklı versiyonları önerilmiştir. Bunlardan biri olan Borderline-SMOTE, Han ve diğ. tarafından sadece karar sınırında (decision boundary) olduğu düşünülen gözlemlere uygulanarak sunulmuştur [13].

Van Hulse ve diğ. 2007 yılında, yüksek dengesizlik oranına sahip otuzun üzerinde gerçek yaşam verisini on bir sınıflama ve yedi örnekleme yöntemini karşılaştırmalı olarak kullanmış ve rassal az örneklemenin daha iyi sonuç verdiğini belirtmiştir [14]. He ve diğ. yaptıkları araştırma ve kapsamlı derleme çalışmalarında dengesiz veri sorununun doğası gereği ortaya çıkışını ve bu sorunun giderilmesine ilişkin yöntemleri incelemiş, kullanılması uygun değerlendirme ölçütlerine ilişkin öneri ve görüşlerini paylaşmışlardır [15]. Ayrıca dengesiz sınıf dağılımı problemi içeren verilerde öğrenme için azınlıkta olan sınıfın dağılımını temel alarak dengesizlik nedeni ile ortaya çıkan yanlılığı azaltmak üzere uyarlanabilir sentetik örnekleme (ADASYN - Adaptive Synthetic Sampling Approach) yaklaşımı önermiş, böylece dengesizlik nedeni ile sınıflaması güç olan gözlemler için karar sınırlarını uyarlanabilir hale getirmişlerdir [15, 16]. Batista ve diğ. rassal aşırı örnekleme ile SMOTE yöntemlerini yirmi iki gerçek yaşam verisi üzerine uygulamış ve yedi sınıflama algoritması ile örnekleme yöntemlerinin başarımı genel olarak %30'dan fazla arttıramayacağını ifade etmişlerdir [17].

Haixiang ve diğ. tarafından yapılan derlemede dengesiz sınıf dağılımı problemini içeren ve son on yılda yayımlanan beş yüzün üzerinde çalışma incelenmiştir [18]. Buna göre, yeniden örneklemeye dayalı topluluk öğrenme yöntemlerinin özellikle kimya, biyomedikal gibi alanlarda derlenen klinik verilerin sabit yapıda olmaları nedeni ile daha yaygın olarak kullanıldığı belirtilmiştir. Öte yandan, finans, işletme gibi alanlarda çoklukla tercih edilen maliyete duyarlı modellerin başarımına ek olarak firmaların karlarını artırmak önemli hedef olduğundan maliyete duyarlı öğrenme yaklaşımları tercih edilmiştir. Aynı çalışmada verilerin karmaşık olması nedeni ile bilgi

teknolojileri alanında dengesiz sınıf dağılımına sahip verilerde öğrenmenin zor olduğu değerlendirilmiştir. Ağ akışları, görüntü, metin gibi yapılandırılmamış veriler için değişken çıkarımına ek olarak dinamik veri içeren çalışmalarda geleneksel yaklaşımlar yerine çevrimiçi öğrenme kullanılabileceği ifade edilmiştir [18].

Dengesiz sınıf dağılımını inceleyen çalışmalarda veri düzeyinde de yapılan çalışmalara ek olarak algoritma düzeyinde de çeşitli yaklaşımlar bulunmaktadır. Veri setindeki tüm gözlemlere başlangıçta eşit bir ağırlık vererek her bir iterasyon sonunda aşamalı olarak doğru sınıflanmayan gözlemlere ait ağırlıkları güncelleme temeline dayanan boosting (arttırma) Schapire ve Freund tarafından önerilmiş bir topluluk öğrenme algoritmasıdır [19, 20]. Chawla ve diğ., benzer bir ağırlıklandırma yaklaşımını temel alan, boosting ve SMOTE yöntemini melezleyen bir algoritma önermişlerdir. SMOTEBoost adını verdikleri bu algoritma, yanlış sınıflanan gözlemlere eşit ağırlık vermek yerine, SMOTE ile elde edilen ve azınlık sınıfa ait gözlemlerin ağrılıklarını güncelleyerek dolaylı olarak dengesizliği gidermeye çalışmaktadır [21]. Dengesiz sınıf dağılımına sahip veri setlerinde çoğunluk/baskın sınıfın alt kümelerini kullanan yaklaşımlarda en önemli sorun çoğunluk sınıftaki çoğu gözlemin katkısının göz ardı edilmesidir. Bu sorunu gidermek amacı ile önerilen EasyEnsemble yaklaşımı baskın sınıftan tek bir alt küme yerine birkaç alt kümeyi örnekleyerek her birini bir öğrenme algoritması ile eğiterek modeller oluşturur ve bu modellerin sonuçlarını birleştirir [22]. Aynı çalışmada sunulan BalanceCascade yaklaşımında ise eğitim sürecini sıralı olarak yaparak baskın sınıfta bulunan ve model tarafından doğru sınıflanan gözlemler çıkarılır. Her iki yaklaşımda da sonuçlar benzer metotlara göre oldukça iyi bir performans ile daha kısa sürede elde edilmiştir [22]. Seiffert ve diğ. tarafından çarpık dağılıma sahip veriler için önerilen RUSBoost algoritması, SMOTEBoost algoritmasına daha anlaşılır ve hızlı bir alternatif sunmaktadır [23]. Birçok veri seti, öğrenme algoritması ve çeşitli değerlendirme ölçütleri kullanılarak elde edilen uygulama sonuçlarına göre önerilen algoritmanın dengesiz veriler için benzerlerine göre performans ve hız bakımından daha etkili olduğu ifade edilmiştir [23].

Salem ve diğ. eksik/kayıp veri ve dengesiz sınıf dağılımı ile birlikte gürültülü özelliklere sahip SECOM veri seti ile yaptıkları çalışmada veri budama, değişken seçimi, eksik gözlem tamamlama, sınıflama gibi yaklaşımları içeren yaklaşık üç yüz kombinasyon incelemişlerdir [24]. Yarı iletken madde üretiminde sensörler yardımı ile toplanan bu veri setindeki sınıf dengesizliği SMOTE ile giderilmiş, en yakın komşu algoritmasını temel alan kayıp gözlem ataması ile düzenlenen veride en iyi sonuç lojistik regresyon ile elde edilmiştir [24]. Liu ve ark. 2020 yılında yayımladıkları çalışmalarında sınıf dengesizliği, verilerin düşük kalitede olması, gürültü ve sınıf çakışmaları gibi problemleri göz önünde bulundurarak bir çerçeve önermişlerdir. Çalışma kapsamında elde edilen bulgular göstermiştir ki, önerilen çerçeve hesaplama bakımından verimli, çakışan sınıflar ve çarpık dağılım durumunda sağlam performans gösteren ve birçok sınıflama yöntemi ile kolaylıkla entegre olabilen bir yapıya sahiptir [25].

Razavi-Far ve diğ. temel olarak torbalama (bagging) ve artırma (boosting) tabanlı topluluk algoritmalarını örnekleme yöntemleri ile entegre ettikleri yeni bir dengesiz sınıf öğrenme algoritması önermişlerdir. Çalışmada geliştirilen bu yaklaşım birçok veri setine uygulanmış ve sonuç olarak torbalama tabanlı yöntemlerin arttırma tabanlı yöntemlere göre daha iyi sonuçlar verdiğini değerlendirmişlerdir [26].

Veri düzeyi ve algoritma düzeyinde değerlendirilen dengesiz sınıf dağılımı çalışmalarına ek olarak maliyete duyarlı (cost-sensitive) çalışmalar da literatürde yer bulmuştur [27-29]. Maliyete duyarlı yaklaşımların temel mantığı, baskın ve azınlık sınıflar için eşit olamayan yanlış sınıflama maliyetleri atayarak değerlendirme yapmaktır. Zong ve diğ. önerdiği ağırlıklı aşırı öğrenme makineleri (WELM-weighted extreme learning machine) ile azınlık sınıfın ağırlığını artırarak sınıf dengesizliği sorunu çözmeye çalışmışlardır [28].

Gerçek yaşam verileri çoğunlukla hem kayıp gözlem hem de dengesiz sınıf dağılımı sorununa sahiptir. Çalışmada literatürde hala güncelliğini koruyan ve çözüm arayışı devam bu iki sorun bir arada ele alınmaya çalışılmıştır. Bu amaçla çeşitli eksik/kayıp gözlem tamamlama yaklaşımları kullanarak topluluk (ensemble) öğrenme algoritmalarının performansları karşılaştırmalı olarak incelenmiştir. Çalışma kapsamında kayıp/eksik veri tamamlama amacı ile sıcak deste ataması, ortalama, ortanca ve tepe değeri ile atama, çoklu atama, beklenti en büyükleme ve k en yakın komşu gibi çeşitli atama ve doldurma yöntemleri kullanılarak tamamlanmış veri setleri oluşturulmuştur. Elde edilen bu veri setlerindeki dengesiz sınıf dağılımını gidermek amacı ile topluluk algoritmaları karşılaştırmalı olarak uygulanmıştır. Bu amaç ile kullanılan extra trees, rasgele orman, gradient boosting, LightGBM ve XGBoost gibi topluluk öğrenme algoritmalarının performansları sensörler aracılığıyla toplanan dengesiz sınıf dağılımına ve kayıp veriye sahip bir veride karşılaştırılarak yorumlanmıştır [30].

2. Materyal ve Metot

2.1. Dengesiz veri kavramı

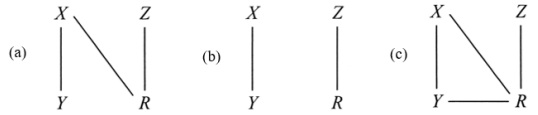
Veri setlerinde her sınıfa ait örnek sayısı olarak tanımlanan sınıf dağılımının dengesiz yapıda olması, bir başka deyişle sınıflardan birinin diğer sınıfa oranla baskın olması modelleme performansını

etkileyen ve dikkatle incelenmesi gereken önemli bir problemdir. Veri bilimi ve yapay zekâ adı altında anılan makine öğrenmesi, veri madenciliği gibi alanlarda öğrenme sürecinde kullanılacak verinin yapısı ve kalitesi model başarısını doğrudan etkileyen önemli bir faktördür. Çoğu standart makine öğrenimi algoritması, veri setinde dengeli sınıf dağılımının olduğunu varsayar. Buna karşın kredi kartı hareketlerine ilişkin sahteciliğin belirlenmesi, endüstriyel sistemlerin ya da bir uçağın operasyonel durumundaki arıza ve hata tespiti, nadir görülen hastalıkların teşhisi gibi gerçek dünya uygulamalarında problemlerin doğası gereği sınıf dağılımı dengesizlikleri ile sıkça karşılaşılmaktadır [15, 26].

Veri setlerinde bulunan sınıflara ait örnek sayılarının eşit dağılmaması yani sınıf dengesizliği nedeni ile çarpıklık durumunda ortaya çıkan en önemli sorun eğitim sırasında sınıflandırıcıların çoğunluk sınıf örneklerini doğru, azınlık sınıfa ait örnekleri yanlış etiketleme eğiliminde olmasıdır [26]. Bu durumun sonunda model özellikle az örneğe sahip sınıfı yeterince öğrenemediği için sınama sonucunda hatalı ve yanlı tahminler ortaya çıkabilir.

2.2. Kayıp veri kavramı

Veri bilimi ve yapay zekâ kapsamında yapılan çalışmalarda kullanılan gerçek yaşam verilerinde ortaya çıkan en temel ve yaygın sorun bir veya birden fazla özellik için eksik/kayıp değerlere sahip olma durumudur. Bu sorun, veri toplama sırasında teknik ya da pratik bir nedenden dolayı veri kaydedilememesi durumunda ortaya çıkar. Longford tarafından en temel ifade ile gözlenmesi beklenen veri seti ile gözlenen veri seti arasındaki fark olarak tanımlanan eksik veri, Rubin tarafından önerilen üç sınıfta incelenmiştir [31, 32]. Bu sınıflar için uygun tanım Little ve Rubin tarafından aşağıdaki gibi verilmiştir [32]. Şekil 1'de verilen **X**, eksik veriye sahip *(Nxp)* boyutunda bir matris ve $\mathbf{X}_g$ bu matriste gözlemlenen girdiler, **Y** çıktı vektörü ve **Z** = (**Y, X**), $\mathbf{Z}_g$=(**Y,** $\mathbf{X}_g$) olsun. **R** ise eksik gözleme karşılık gelen x_{ij} =1 ve diğer durumlarda x_{ij} =0 olan gösterge matrisi olarak verilsin. Rubin tarafından kayıp veri mekanizmasını tanımlamak için önerilen sınıflardan ilki olan ve Şekil 1a'da verilen rastgele kayıp (missing at random - RK) mekanizması X'te eksik veri olması olasılığının **Y** değişkeni ile ilişkili olması yani $P(\mathbf{R} \mid \mathbf{Z}) = P(\mathbf{R} \mid \mathbf{Z}_g)$ şeklinde bağımlı olması durumu olarak tanımlanır. Benzer biçimde Şekil 1b'de sunulan tümüyle rastgele kayıp (missing completely random - TROK) mekanizması $P(\mathbf{R} \mid \mathbf{Z}) = P(\mathbf{R})$ biçiminde bağımsız olarak yani **X** ve **Y** değişkenlerine yanıt alınamama durumunun birbirinden etkilenmediği durum, bir başka ifade ile **Y**'deki eksik değer olasılığının **X**'e bağlı olmaması durumu olarak tanımlanır [31, 33]. Son kayıp veri mekanizması olan rastgele olmayan kayıp (missing not at random - ROK) ise Şekil 1c verildiği üzere **X**'te ortaya çıkan eksik veri durumunun hem kendisiyle hem de çıktı değişkeni ile ilişkili olmasıdır [31, 33].



Şekil 1. (a) Rastgele Kayıp Mekanizması (RK) (b) Tamamıyla Rastgele Kayıp Mekanizması (TROK) (c) Rastgele Olmayan Kayıp Mekanizması (ROK).

Kayıp veri sorunun giderilmesi için literatürde bulunan yöntemler temel olarak silme ve atama yöntemleri olarak ayrılmıştır. Geleneksel yöntemlerden olan silme yöntemleri oldukça basit ve yaygın kullanıma sahiptir. Ancak basit olmaları beraberinde çeşitli dezavantajlar ortaya koymuş, bu durum da alternatif yöntem arayışını gerektirmiştir. En çok olabilirlik ve çoklu atama yaklaşımları silme yöntemlerine karşı alternatif olarak sunulan yöntemler arasında öne çıkmıştır [34]. Liste bazında (list-wise) ve çiftler bazında (pair-wise) olmak üzere iki şekilde uygulanabilen silme yöntemlerinin her iki durumunda da eksik değer içeren gözlemlerin veri setinden çıkarılması gerekmektedir. Atama yöntemleri ise tekli atama (sıcak deste, soğuk deste, ortalama ile atama, regresyon ataması) ve çoklu atama olmak üzere iki gruba ayrılmaktadır. Çoklu atama yöntemleri, tekli atama yöntemlerinin bir arada kullanıldığı yöntemler olarak tanımlanabilir.

Eksik değer içeren ve içermeyen gözlemler sırası ile bağımlı ve bağımsız olarak tanımlandığında, tekli atama yöntemlerinde temel amaç istatistiksel olarak anlamlı ve açıklama düzeyi yüksek modeller oluşturmaktır. Regresyon modellerinin yanı sıra regresyon ataması için çeşitli makine öğrenmesi (karar ağaçları, destek vektör makineleri, *k* en yakın komşu vb.) yöntemleri de kullanılabilmektedir. Tahmine dayalı algoritmalara ek olarak yerine koyma yöntemlerinden de yararlanılır. Veri setinde yer alan eksik değerler ortalama, ortanca ya da tepe değeri gibi veri içerisinde bulunan ve hesaplanan değerler ile değiştirilir [30].

Literatürde bağışçı (donör) olarak bilinen sıcak ve soğuk deste atamaları da tekli atama yöntemleri arasında değerlendirilir. Sıcak ve soğuk deste atamasında temel fark atama için kullanılacak değerin seçildiği veri setidir. Sıcak deste atamasında eksik değer ataması verinin kendisinin içinden rasgele seçilen değerler ile yapılırken, soğuk deste atamasında bu işlem başka bir veri setinden rasgele seçim ile yapılmaktadır [30].

Beklenti en büyükleme (BEB) ve çoklu atama (en çok olabilirlik çoklu atama) yöntemleri en çok olabilirlik tabanlı yöntemlerdendir. Bu yöntemler, en çok olabilirlik kullanarak atama modelinin parametrelerini tahmin eder. Her bir gözlemin bağımsız olması durumunda, ortaya çıkma

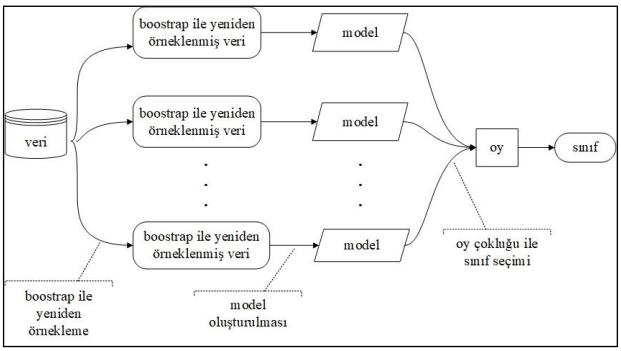
olasılıklarının çarpımını en büyük yapan parametre tahminine en çok olabilirlik tahmini denmektedir [30]. Bir parametre için bulunacak en çok olabilirlik tahmini, gözlenme olasılığının en yüksek olduğu gözlemleri veren parametrenin değeridir. BEB algoritması yinelemeli ve iki aşamalı bir yöntem olarak ilk defa Dempster ve diğ. tarafından önerilmiştir [4]. Eksik değerler için olası en iyi kestirimler beklenti (B) aşamasında hesaplanırken, en büyükleme (EB) aşamasında eksik değer ataması yapıldığında dağılımı tanımlamaya ilişkin kestirimler (ortalama, standart sapma, korelasyon gibi) elde edilmektedir. BEB algoritması, kestirilen değerlerdeki değişimler önemsenmeyecek düzeye gelene kadar yinelenir [35].

2.3. Topluluk öğrenme algoritmaları

Çalışma kapsamında kullanılan sınıflama yöntemleri ortak dil kullanmak açısından Gümüştaş ve diğ. tarafından 2021 yılında yapılan çalışmada belirtilen isim ve kısaltmalar ile ifade edilmiştir [36]. Buna göre, bu çalışmada kullanılan yöntem ve kısaltmalar şöyledir; *"RF: rasgele orman (random forest), ExtraTrees: aşırı rasgeleleştirilmiş ağaçlar (extremly randomized trees), XGBoost: aşırı gradyan arttırma (extreme gradient boosting), LightGBM: hafif gradyan arttırma"* [36].

Birbirinden bağımsız olarak rastgele tahminden biraz daha başarılı (en az %51) tahminde bulunan öğreniciler zayıf öğrenici olarak tanımlanmaktadır. Topluluk öğrenmesi birden fazla zayıf öğreniciden elde edilen sonuçları bir arada değerlendirerek daha kararlı ve başarılı sonuçlar elde etmek amacı ile tasarlanmış esnek bir yapıdır. Her bir zayıf öğreniciden elde edilen bağımsız sonuçların oluşturduğu ortak karar ile daha başarılı ve düşük değişkenliğe sahip bir yaklaşım hedeflenir [37]. Gerek sınıflama gerekse regresyon ile yapılan tahminlerin doğruluğu gürültü, değişkenlik, yanlık gibi sorunlar ile doğrudan ilişkilidir. Tahmin ile gerçek değer arasındaki farklılık dolayısı ile değişkenlik ve yanlılık, birden fazla öğreniciden oluşan topluluk öğrenme yöntemleri ile en aza indirilmeyi amaçlar.

Yaygın kullanılan topluluk öğrenme algoritmaları genel olarak üç yaklaşım ile gerçekleştirilir; torbalama (bagging: boostrap aggregating), arttırma (boosting) ve yığma (stacking). Özellikle yüksek korelasyon durumunda rastgele seçilmiş daha küçük alt eğitim setleri ile birden fazla kurulan modellerin birleştirilmiş sonucu olarak tanımlanabilen bagging yaklaşımı 1996 yılında Breiman tarafından ortaya konmuştur [38]. Bagging temel olarak ilk defa 1979 tarihinde Jackknife yöntemine alternatif olarak sunulan, 1994 tarihinde ise Efron ve Tibshirani tarafından geliştirilerek sunulan yeniden örnekleme yöntemi Bootstrap'i kullanmaktadır [39, 40]. Veri setinden olabildiğince yüksek düzeyde bilgi alabilmek adına bootstrap işlemi defalarca tekrarlanmakta, böylece yeniden örneklenen alt eğitim veri setleri ile eğitilen modellerin sonuçları literatürde topluluk öğrenmesinin temelini oluşturan "çokluğun bilgeliği" (wisdom of crowd) temelinde birleştirilerek çok daha tutarlı ve başarılı sonuçlar elde edilebilmektedir [38, 41]. Şekil 2'de genel aşamaları sunulan bagging yaklaşımının uygulaması ***m*** büyüklüğünde orijinal veri setinden bootstrap yaklaşımı ile ***p*** büyüklüğünde gözlem içeren alt örneklemler seçilir. ***p*** büyüklüğünde her bir alt örneklem belirlenen bir öğrenici (sınıflayıcı/regresör) ile eğitilerek *k* adet model kurulur. Her bir modelden elde edilen sonuçlar problemin yapısına göre ortalama ya da çoğunluk oylaması (majority voting) ile birleştirilerek sonuç elde edilir.



Şekil 2. Bagging yönteminin aşamaları.

Topuluk öğrenmesi için bir diğer yaklaşım olan arttırma (boosting) yaklaşımı birden fazla zayıf öğrenicinin etkilerini güçlü bir öğrenicide birleştirme temeline dayanır. Çoğu boosting yaklaşımında eğitimler zayıf öğreniciler ile ardışık olarak yapılır. Elde edilen her sonuç bir sonraki zayıf öğrenici (sınıflayıcı) için yeni bir bilgi olarak aktarılır ve her aşamada iyileşme beklenir. AdaBoost, Gradient Boosting gibi yöntemler boosting yaklaşımını kullanmaktadır [42].

Kısaca yığılma (stacking) olarak bilinen üçüncü yaygın topluluk öğrenme yaklaşımı da yığılmış genellemedir (stacked generalization) [43]. Bu yaklaşımın bagging yaklaşımından farkı farklı alt eğitim setleri ile elde edilmiş modelleme sonuçlarının oylama gibi fonksiyonlar yerine başka bir öğrenici ile birleştirilmesidir. Her bir modelden gelen sonuçlar birleştirici olarak belirlenen öğreniciye girdi olarak verilir ve sonuçlar elde edilir.

2.4. Uygulama tasarımı ve veri seti

Dengesiz sınıf dağılımına sahip veri setlerinde gözlenen kayıp/eksik gözlem sorununun topluluk öğrenmesi algoritmaları ile incelenmesine ilişkin bu çalışmada uygulamalar Python ve R açık kaynak programlama dilleri ile gerçekleştirilmiştir. Modelleme aşamasında xgboost, scikit-learn ve lightgbm kütüphaneleri kullanılmıştır.

Çalışmada kullanılan veri seti İsveç menşeili çekici ve ağır kamyon üreticisi SCANIA tarafından sağlanmış olup sensörler aracılığı ile toplanmıştır. Genel olarak kamyon ve çekicilerdeki frenleme ve vites değiştirme gibi çeşitli işlevlerde kullanılan basınçlı havayı üreten hava basıncı sistemine (APS) ilişkin bilgileri içermektedir. Veri 2016 yılında düzenlenmiş olan 15th International Symposium on Intelligent Data Analysis kapsamında gerçekleşen endüstriyel bir yarışma için anonim olarak paylaşılmıştır. SCANIA veri seti eğitim ve sınama olmak üzere iki dosya halinde paylaşılmış olup, anonim olması nedeni ile değişkenlerine ilişkin ayrıntılar bilinmemektedir. Veride toplam 171 değişken olup, tahmin edilmesi istenen hedef değişken hava basıncı sisteminin belirli bir bileşeni için arızanın olması (pozitif) ve olmaması (negatif) olmak üzere iki kategoriye sahiptir. Eğitim setinde 59000 adet negatif sınıfa ait, 1000 adet pozitif sınıfa ait 60000 gözlem bulunmaktadır. 16 bin gözlemden oluşan sınama seti ise 15625 negatif sınıf ve 375 pozitif sınıf dağılımı ile %2.4 dengesizlik oranı içermektedir. Eğitim verisindeki bu oran ise yaklaşık %1.7 olarak hesaplanmaktadır. Veri seti kayıp/eksik gözlemler bakımından incelendiğinde ise 169 değişkende kayıp veri sorunu gözlenmiştir. Hem eğitim hem sınama verisinde kayıp veri oranı %82'ye kadar çıkmıştır. Veri setlerinde azınlık sınıfta (pozitif etiketli) bulunan gözlem sayısının baskın sınıfta (negatif etiketli) bulunan gözlem sayısına oranı ile ölçülen dengesizlik oranının (imbalanced ratio) ayrıntılı dağılımı ve sınıflara göre gözlem sayıları Tablo 1'de sunulmuştur.

Tablo 1. Eğitim ve sınama setlerine ilişkin gözlem sayıları ve dengesizlik oranı.

	Negatif	Pozitif	Dengesizlik Oranı (%)
Eğitim	59000	1000	1.7
Sınama	15625	375	2.4

Veri setindeki eksik veri mekanizmasının kayıp/eksik gözlem ataması yapmadan belirlenmesi gerekmektedir. Bu amaç doğrultusunda kurulan hipotez Little'ın TROK testi ile değerlendirilmiştir. Testin sonucunda *ki*-kare 3756.395 ve $\boldsymbol{p}$>0.05 olarak bulunduğu için $\boldsymbol{H_0}$ hipotezi reddedilmemiştir.

H_0: Veri setindeki kayıp gözlemler tamamen rastgele olarak kayıptır.
H_1: Veri setindeki kayıp gözlemler tamamen rastgele olarak kayıp değildir.

Bu bulgu doğrultusunda eksik veri mekanizmasının TROK olduğu 0.05 yanılma düzeyinde bulunmuş, dolayısı ile sorununun giderilmesi için uygun yöntemlerin silme ve atama olduğu belirlenmiştir. Çalışma kapsamında kayıp gözlem ataması için aritmetik ortalama, ortanca, tepe değeri, *k*-nn, çoklu atama, sıcak deste ve beklenti en büyükleme yöntemleri ilgili Python ve R kütüphaneleri ile kullanılmıştır.

2.5. Değerlendirme ölçütleri

Çalışmalarda uygulanan yöntemlerin etkinliğini değerlendirmek için çeşitli performans ölçütleri kullanılabilir. Özellikle doğruluk (accuracy) ölçütü standart sınıflandırıcıları değerlendirmek için kullanılan en yaygın ölçüttür. Ancak doğruluk ölçütü sınıf dağılımında çoğunluk sınıfa olan eğilimin yarattığı yanlılık nedeni ile çarpık sınıf verileri için iyi bir ölçüt değildir ve bu durum çeşitli çalışmalarda da gösterilmiştir [44, 45]. Bu nedenle, çalışma kapsamında model değerlendirme ölçütleri olarak doğruluk (accuracy) yerine, sınıf dengesizliğini ayrıştırmada etkili olabilen kesinlik (precision), duyarlılık (recall, sensitivity), eğri altında kalan alan (area under curve) ve F1 skoru ölçütleri yansız bir değerlendirme için tercih edilmiştir. Bu istatistiklere ilişkin hesaplamalar karışıklık matrisi ile gerçekleştirilmektedir.

Özellikle dengesiz sınıf dağılımına sahip veri setlerinin değerlendirilmesinde doğruluk yerine duyarlılık ve kesinlik değerlerinin harmonik ortalaması alınarak hesaplanan F1 skoru tercih edilmektedir [46]. Eşitlik (1)'e göre hesaplanan ve değişim aralığı [0,1] olan bu değerin 1'e yakın olması beklenir.

$$F1\ Skoru = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık} \qquad (1)$$

3. Bulgular

Bu bölümde, önceki bölümlerde ifade edilen uygun yöntemler aracılığı ile kayıp gözlem ataması yapılarak yeni veri setleri elde edilmiştir. Bu veri setleri Extra Trees, Gradient Boosting, LightGBM, Random Forest ve XGBoost algoritmaları ile sınıflandırılmış ve model başarımlarını karşılaştırmalı olarak değerlendirebilmek için F1 skor hesaplanmıştır.

Doğruluk (accuracy) doğru tahmin edilen gözlemlerin oranını vermektedir. Bilindiği ve bölüm 2.5'te de ifade edildiği gibi özellikle dengesiz verilerin sınıflandırılma sonuçlarının değerlendirmesinde yanlıdır ve tercih edilmez. Diğer taraftan duyarlılık (recall) ve kesinlik (precision) oldukça önemli ölçütlerdir. F1, her iki değerlendirme ölçütünü (duyarlılık ve kesinlik) dengeli olarak ağırlıklandıran ve yüksek F1 değeri için her ikisinin de yüksek değere sahip olmasını gerektiren bir istatistiktir. Ayrıca F1 skoru çok düşük ya da çok yüksek değerlerden etkilenmediğinden F1 skorunu en büyüklemek, duyarlılık, kesinlik, doğruluk gibi ölçütlerden daha yansız olacaktır. Bu nedenle, çalışmada tüm değerlendirmeler F1 skor temel alınarak yapılacaktır.

Tablo 2'de çeşitli kayıp gözlem ataması ile elde edilmiş veri setleri ve topluluk sınıflandırma

modellerine ait F1 skorları verilmiştir. Tabloya göre tüm atama yöntemleri ile elde edilen veri setleri için XGBoost ile elde edilen sonuçların özellikle F1 skoru düzeyinde en iyi sonuçları verdiği görülmektedir. Sınıflandırma algoritmalarının başarısını yorumlamak ve net bir bakış açısı getirmek adına Tablo 3'te verilen Wilcoxon sıra testi için sıralama F1 skoru kullanılarak yapılmıştır. Her bir atama yöntemi için elde edilen F1 skorları kendi içlerinde sıralanarak birer sıra değeri verilmiştir. Eşit değere sahip olanlar için ise ortalama sıra değeri atanmıştır.

Tablo 2. Modellere göre F1 skoru.

Atama Yöntemleri	Extra Trees	Gradient Boosting	LightGBM	Random Forest	XGBoost	Ortalama
Sıcak Deste	0.7976	0.8012	0.8179	0.7951	0.8516	0.8127
Aritmetik Ort.	0.8024	0.8076	0.8247	0.8193	0.8588	0.8226
Mod	0.8102	0.7954	0.8314	0.8103	0.8576	0.8210
Medyan	0.8109	0.8035	0.8333	0.8091	0.8622	0.8238
Çoklu Atama	0.7988	0.7847	0.8232	0.8116	0.8571	0.8151
BEB	0.7896	0.8024	0.8024	0.7951	0.8408	0.8061
KNN-5	0.8085	0.7959	0.8172	0.8073	0.8614	0.8181
Ortalama	0.8026	0.7987	0.8214	0.8068	0.8556	

Tablo 3. İstatistiksel değerlendirme için sıra özeti**.**

Atama Yöntemleri	Extra Trees	Gradient Boosting	LightGBM	Random Forest	XGBoost
Sıcak Deste	4	3	2	5	1
Aritmetik Ort.	5	4	2	3	1
Mod	3.5	5	2	3.5	1
Medyan	3	5	2	4	1
Çoklu Atama	4	5	2	3	1
BEB	5	2.5	2.5	4	1
KNN-5	3	5	2	4	1
Toplam	27.5	29.5	14.5	26.5	7

Tablo 3'te verilen test sonuçlarına göre XGBoost veri seti üzerinde en başarılı sonucu vermiştir. LightGBM ikinci başarılı sınıflandırma algoritması olurken, az farkla sırası ile random forest, Extra Trees ve gradyan boosting gelmektedir. Öte yandan Tablo 2 incelendiğinde XGBoost algoritması, medyan ve *k* en yakın komşu atama yöntemleri ile elde edilen veri setleri ile en yüksek F1 skorunu vermiştir. Atama ile oluşturulan veri setleri üzerine yapılan genel değerlendirmede ortanca ataması yapılmış veri seti ile elde edilen ortalama F1 skorunun 0.8238 ile en yüksek değere sahip olduğu gözlenmiştir. Aritmetik ortalama ataması ile elde edilen veri seti 0.821 ile en yüksek F1 skoruna sahip ikinci veri seti olurken, tepe değeri atanarak tamamlanan veri seti en yüksek üçüncü F1 skoruna sahiptir. Sonuçlar göstermiştir ki, yerine koyma yöntemleri ile yapılan eksik gözlem doldurma yaklaşımları sıcak deste, çoklu atama, beklenti en büyükleme ve *k* en yakın komşu yöntemleri ile yapılan eksik gözlem atamalarına göre daha tercih edilebilir durumdadır.

4. Tartışma ve Sonuç

Gelişen teknoloji ile artan veri üretimi, farklılaşan veri kaynakları ve dolayısı ile değişen veri türleri bu çalışmanın da odak noktalarından olan önemli veri problemlerini de beraberinde getirmektedir. Tüm bu sorunlar ve hızlı değişimler geleneksel yöntemleri yetersiz kılmış, özellikle makine öğrenmesi yöntemlerini veri bilimi alanında oldukça önemli konuma gelmiştir. Bilindiği üzere makine öğrenmesi yöntemlerinin başarılı sonuçlar verebilmesinde veri setinin analize uygun ve sorunsuz olması önemli bir rol oynamaktadır. Karşılaşılan en önemli sorunlardan olan eksik/kayıp gözlem ve dengesiz sınıf dağılımının bir arada olması durumu bu çalışma kapsamında çeşitli yaklaşımlar ile incelenmiştir.

Çalışmanın temel amacı, gerçek yaşam uygulamalarında sıklıkla karşılaşılan sözü edilen iki sorunun belirlenmesi ve giderilmesine yönelik bilgi vermek, ayrıca her ikisinin aynı anda görülmesi durumunda farklı sınıflandırma algoritmaları ile karşılaştırmalı bir değerlendirme yapmaktır. Uygulama için seçilen veri setinde eksik değerleri doldurulmadan önce kayıp gözlem mekanizmasının belirlenmesi için Little'ın TROK testi kullanılmıştır. Eksik değer doldurulması TROK testi ile belirlenen mekanizmaya göre atama ve yerine koyma yöntemleri uygulanarak gerçekleştirilmiştir. Aynı zamanda dengesiz sınıf dağılımı problemine de sahip olan veri seti için Extra Trees, Gradient Boosting, LightGBM, Random Forest ve XGBoost topluluk öğrenme yöntemleri modelleme amacı ile kullanılmıştır. Tamamlanmış veri setleri üzerinden alınan sonuçlar göstermiştir ki; eksik/kayıp gözlem sorununu gidermek için kullanılan yerine koyma yöntemi ile elde edilen veri setlerinde modellerin ortalama başarısı daha yüksektir. Bu durum veri setinin taşıdığı bilginin kullanılmasının daha başarılı ve anlamlı sonuçlar verebileceğini ortaya koymuştur.

Sunulan bulgulardan çıkarılan bir diğer sonuç ise topluluk öğrenme yaklaşımlarından arttırmaya dayanan yöntemlerin, torbalama yaklaşımı kullanan yöntemlere kıyasla daha iyi sonuçlar verdiğidir. Bu değerlendirmelerin tamamı dengesiz sınıf dağılımı probleminden dolayı ağırlıklı olarak F1 skoru ile yapılmıştır. F1 skoru temel alındığında en başarılı sonuç XGBoost, en başarılı atama yaklaşımı medyan ataması olarak elde edilmiştir. Öte yandan, çalışma kapsamında kullanılan veri setine ait değişken bilgilerinin kısıtlı olması (anonim) olması özellikle kayıp gözlem atama uygulamalarını güçleştirmiştir. Unutulmamalıdır ki, gerçek yaşam verilerinde veri setlerine ilişkin düzenlemeler özellikle değişkenlere ait bilgiler dikkate alınarak yapıldığında daha anlamlı olmaktadırlar.

Teknolojik ilerlemeler ile veri setlerindeki artış ve karmaşıklık dengesiz sınıf dağılımı ve eksik gözlem sorununu güncel tutmayı gerektirmekte olup bu alanda çalışmalar devam etmektedir. Yapılabilecek çalışmalar arasında dengesiz veri problemlerinde bir diğer yaklaşım olan örnekleme tabanlı yaklaşımlar ya da bu yöntemleri kullanan melez yaklaşımların kullanılması düşünülebilir. Buna ek olarak değişken seçim yöntemlerinden yararlanarak dengesiz sınıf dağılımından dolayı ortaya çıkan ve gürültü yarattığı düşünülen değişkenler belirlenerek veri setinden çıkarılabilir ve/veya dağılımının çarpık olduğu görülen değişkenler üzerinde çeşitli düzeltme yöntemleri kullanılarak veri düzeyinde yaklaşımlar denenebilir. Ayrıca modelleme aşamasından sonra elde edilen sınıf olasılıkları kullanılarak eşik optimizasyonu (threshold optimization) yapılarak F1 skoru iyileştirilebilir.

Etik Beyanı

Bu çalışmada, "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması gerekli tüm kurallara uyulduğunu, bahsi geçen yönergenin "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbirinin gerçekleştirilmediğini taahhüt ederiz.

Teşekkür / Belirtme

Bu çalışma, Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı Yüksek Lisans Programı'nda, Enis Gümüştaş tarafından, Doç. Dr. Ayça Çakmak Pehlivanlı danışmanlığında tamamlanan "Kayıp Gözlem İçeren Dengesiz Veri Setlerinin Topluluk Öğrenme Algoritmaları ile Sınıflandırılması" başlıklı Yüksek Lisans tezinden üretilmiştir. Tezin inceleme ve değerlendirme aşamasında yapmış oldukları katkılardan dolayı jüri üyelerine teşekkür ederiz.

Kaynakça

[1] Rubin, D. B. 1976. Inference and missing data. Biometrika, 63(3), pp. 581-592.

[2] Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B: Methodological, 39(1), pp. 1-22.

[3] Little, R. J. 1988. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), pp. 1198-1202.

[4] Chan, P., and Stolfo, S. 1998. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In Proc. of Knowledge Discovery and Data Mining, pp:164–168.

[5] Fu K., Cheng D., Tu Y., Zhang L. 2016. Credit Card Fraud Detection Using Convolutional Neural Networks. Neural Information Processing. ICONIP 2016. Lecture Notes in Computer Science, vol 9949. Springer, Cham.

[6] Sanz, J. A., Bernardo, D., Herrera, F., Bustince, H., and Hagras, H. 2015. A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. Fuzzy Systems, IEEE Transactions on, 23(4), pp. 973–990

[7] Mitchell P.S., Parkin R.K., Kroh E.M., et al. 2008. Circulating microRNAs as stable blood-based markers for cancer detection. Proc. of the National Academy of Sciences, 105(30) pp. 10513-8.

[8] Oh, S., Lee, M. S. And Zhang, B.T. 2011. Ensemble learning with active example selection for imbalanced biomedical data classification. IEEE-ACM Trans. on Computational Biology and Bioinformatics (TCBB), 8(2), pp. 316–325

[9] Li, Y., Sun, G., & Zhu, Y. 2010. Data imbalance problem in text classification. IEEE 2010 3rd Int. Symposium on Information Processing, pp. 301-305.

[10] Kubat, M., Holte, R.C. and Matwin, S. 1998. Machine learning for the detection of oil splis in radar images Machine Learning, 30, pp.195–215.

[11] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16, pp. 321–357.

[12] Drummond, C. and Holte, R. C. 2003. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In Workshop on Learning from Imbalanced Datasets II, vol. 11, pp. 1-8.

[13] Han, H., Wang, W. Y. and Mao, B. H. 2005. Borderline-SMOTE: A new oversampling method in imbalanced data sets learning. In International Conference on Intelligent Computing (ICIC'05). Lecture Notes in Computer Science 3644, pp. 878-887, Springer-Verlag.

[14] Van Hulse, J., Khoshgoftaar, T.M. and Napolitano, A. 2007. Experimental perspectives on learning from imbalanced data. In Proc. of the 24th Int. Conf. on ML (ICML), pp. 17–23.

[15] He, H., Bai, Y., Garcia, E. A. and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks, pp. 1322-1328.

[16] He, H., Garcia, E. A. 2009. Learning from Imbalanced Data, IEEE Trans. Knowledge and Data Eng., 21(9), pp. 1263-1284.

[17] Batista, G. E. D. A. P. A., Silva, D. F. and Prati, R. C. 2012. An Experimental Design to Evaluate Class Imbalance Treatment Methods, 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, pp. 95-101.

[18] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, pp. 220-239.

[19] Schapire, R. E. 1990. The strength of weak learnability. Machine learning, 5(2), 197-227.

[20] Freund, Y., and Schapire, R. E. 1996. Experiments with a new boosting algorithm. Proc. of the 13th International Conference on International Conference on Machine Learning, ICML' 96, pp. 148-156.

[21] Chawla N.V., Lazarevic A., Hall L.O. and Bowyer K.W. 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. Knowledge Discovery in Databases: PKDD 2003. Lecture Notes in Computer Science, vol 2838. Springer, Berlin, Heidelberg.

[22] Liu, X. Y., Wu, J. and Zhou, Z. H. 2009. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, 39(2), pp. 539-550.

[23] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. 2009. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40(1), pp. 185-197.

[24] Salem, M., Taheri, S., and Yuan, J. S. 2018. An Experimental Evaluation of Fault Diagnosis from Imbalanced and Incomplete Data for Smart Semiconductor Manufacturing. Big Data and Cognitive Computing, 2(4), 30.

[25] Liu, Z., Cao, W., Gao, Z., Bian, J., Chen, H., Chang, Y., and Liu, TY. 2020. Self-paced Ensemble for Highly Imbalanced Massive Data Classification, IEEE 36th International Conference on Data Engineering (ICDE), pp. 841-852.

[26] Razavi-Far, R., Farajzadeh-Zanajni, M., Wang, B., Saif, M. and Chakrabarti, S. 2021. Imputation-Based Ensemble Techniques for Class Imbalance Learning, IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 5, pp. 1988-2001.

[27] Zhou, Z.-H. and Liu, X.-Y. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Trans. Knowledge. Data Eng., vol. 18, pp. 63–77.

[28] Zong, W., Huang, G.-B. and Chen, Y. 2013. Weighted extreme learning machine for imbalance learning, Neurocomputing, vol. 101, pp. 229–242.

[29] Wang, J., Zhao, P. and Hoi, S. C. H. 2014. Cost-sensitive online classification, IEEE Trans. Knowledge. Data Eng., vol. 26(10), pp. 2425–2438.

[30] Gümüştaş, E. 2019. Kayıp gözlem içeren dengesiz veri setlerinin topluluk öğrenme algoritmaları ile sınıflandırılması. Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 48s, İstanbul.

[31] Longford, N. T. 2004. Missing data and small area estimation in the UK Labour Force Survey. Journal of the Royal Statistical Society: Series A: Statistics in Society, 167(2), pp. 341-373.

[32] Little, R.J.A. and Rubin, D.B. 1987. Statistical Analysis with Missing Data. John Wiley & Sons, New York.

[33] Oğuzlar, A. 2001. Alan araştırmalarında kayıp değer problemi ve çözüm önerileri. Ulusal Ekonometri ve İstatistik Sempozyumu, Çukurova Üniversitesi Adana, 20(22), pp. 1-28.

[34] Allison, Paul. 2001. Missing data. Sage University Papers Series on Quantitative Applications in the Social Sciences. 07-136.

[35] Alpar, R. 2003. Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş 1, Nobel Akademik Yayıncılık, 404s.

[36] Gümüştaş, E. ve Çakmak Pehlivanlı, A. 2021. In-Silico Mutajenisite Tahmininde İstatistiksel Öğrenme Modeli. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi 25, pp. 365-370.

[37] Dietterich, T. G. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, 40(2), pp. 139-157.

[38] Breiman, L. 1996. Bagging predictors. Machine learning, 24(2), pp. 123-140

[39] Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7(1), pp. 1-26.

[40] Efron, B. and Tibshirani, R. 1994. An introduction to the bootstrap. Chapman & Hall/CRC.

[41] Surowiecki, J. 2004. The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economics, Societies and Nations., Little, Brown.

[42] Freund, Y. and Schapire, R.E. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences, 55(1), pp. 119-139.

[43] Wolpert, D. H., 1992. Stacked generalization, Neural Networks, 5(2), pp. 241-259.

[44] Maloof, M. A. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. Workshop on Learning from Imbalanced Datasets II vol. 2.

[45] Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 40(12), pp. 3358-3378.

[46] Lipton, Z. C., Elkan, C., and Naryanaswamy, B. 2014. Optimal thresholding of classifiers to maximize F1 measure. Joint European Conf. on Machine Learning and Knowledge Discovery in Databases pp. 225-239. Springer, Berlin, Heidelberg.