T.C.
MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY ZEKA ANABİLİM DALI

DENGESİZ VERİ SETLERİNDE HİBRİT YENİDEN ÖRNEKLEME YÖNTEMLERİNİN KARŞILAŞTIRILMASI

HASAN ERSAN YAĞCI

YÜKSEK LİSANS TEZİ

MAYIS 2024
MUĞLA

MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ

Fen Bilimleri Enstitüsü

TEZ ONAYI

HASAN ERSAN YAĞCI tarafından hazırlanan **DENGESİZ VERİ SETLERİNDE HİBRİT YENİDEN ÖRNEKLEME YÖNTEMLERİNİN KARŞILAŞTIRILMASI** başlıklı tezinin, 31/05/2024 tarihinde aşağıdaki jüri tarafından Yapay Zeka Anabilim Dalı'nda yüksek lisans derecesi için gerekli şartları sağladığı oybirliği/oyçokluğu ile kabul edilmiştir.

TEZ SINAV JURİSİ

Doç Dr. Ömer AYDIN **(Jüri Başkanı)**

Bilgisayar Bilimleri ABD.
Celal Bayar Üniversitesi, Manisa — İmza:

Dr. Öğretim Üyesi Nida GÖKÇE NARİN **(Danışman)** — İmza:

İstatistik Bilgi Sistemleri ABD. / Yapay Zeka ABD.
Muğla Sıtkı Koçman Üniversitesi, Muğla

Dr. Öğr. Üyesi Enis KARAARSLAN **(Üye)** — İmza:

Yapay Zeka ABD.
Muğla Sıtkı Koçman Üniversitesi, Muğla

ANA BİLİM DALI BAŞKANLIĞI ONAYI

Dr. Öğretim Üyesi Nida GÖKÇE NARİN — İmza:

Yapay Zeka ABD. Başkanı
İstatistik Bilgi Sistemleri ABD. / Yapay Zeka ABD.
Muğla Sıtkı Koçman Üniversitesi, Muğla

Dr. Öğretim Üyesi Nida GÖKÇE NARİN — İmza:

Danışman
İstatistik Bilgi Sistemleri ABD. / Yapay Zeka ABD.
Muğla Sıtkı Koçman Üniversitesi, Muğla

Savunma Tarihi: 31 / 05 / 2024

Tez çalışmalarım sırasında elde ettiğim ve sunduğum tüm sonuç, doküman, bilgi ve belgelerin tarafımdan bizzat ve bu tez çalışması kapsamında elde edildiğini; akademik ve bilimsel etik kurallarına uygun olduğunu beyan ederim. Ayrıca, akademik ve bilimsel etik kuralları gereği bu tez çalışması sırasında elde edilmemiş başkalarına ait tüm orijinal bilgi ve sonuçlara atıf yapıldığını da beyan ederim.

Hasan Ersan YAĞCI
31/05/24

ÖZET

DENGESİZ VERİ SETLERİNDE HİBRİT YENİDEN ÖRNEKLEME YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Hasan Ersan YAĞCI

Yüksek Lisans Tezi

Fen Bilimleri Enstitüsü

Yapay Zeka Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Nida GÖKÇE NARİN

Mayıs 2024, 34 sayfa

Makine öğrenimi sınıflandırma çalışmalarında, bir sınıfın diğerine göre az temsil edilmesi, verinin dengesiz olması durumu sonuçların azınlık aleyhinde çarpık olmasına neden olmaktadır. Dengesiz veri setleriyle çalışmak için bir takım dengesiz öğrenme yöntemleri bulunmaktadır. Bu tezde, makine öğreniminde dengesiz öğrenme yöntemlerinden yeniden örnekleme yöntemleri ele alındı. Veri seti olarak trafik kazaları sonucunda hukuklaşmaya giden ve gitmeyen sınıfların olduğu, hukuklaşmaya gidenlerin %2 oranında temsil edildiği dengesiz bir veri seti kullanıldı. Veri setindeki hukuklaşmaya giden örneklerin düşük sayısı nedeniyle modelin ön yargılı davranmasını aşmak için, yeniden örnekleme yöntemleri tek başına ve hibrit olarak kullanarak çeşitli denemeler gerçekleştirilmiştir. Bu çalışmada makine öğrenimi algoritması olarak, dengesiz veri setlerinin eğitiminde başarımı yüksek olan rastgele orman kullanılmıştır. Herhangi bir dengesiz öğrenme yöntemi uygulanmadan direkt olarak rastgele orman ile geliştirilen modelde kesinlik skoru %100, duyarlılık %4 ve F1 skor %7 elde edilmiştir. Ele alınan problemde duyarlılık ve kesinlik skorlarının daha önemli olmasından dolayı bu metrikler üzerine çalışılmış ve başarılı olan hibrit modelde skorlar kesinlik skoru için %36, duyarlılık %22 ve F1 skor %28 elde edilmiştir. Bu çalışma sonucunda elde edilen bulgular, yeniden örnekleme yöntemlerinin etkinliği, hibrit yaklaşımların avantajları ve başarım kriterleri stratejilerinin katkıları üzerine odaklanmış ve trafik kazalarında hukuklaşmaya giden olayların tahmin başarısını arttırmayı amaçlamıştır. Çalışmanın amacı yüksek boyutlu ve karmaşık veri setlerinde dengesiz sınıf problemini çözecek hibrit yöntemleri belirlemektir. Bu çalışmanın sonuçları, trafik kazalarının hukuki sonuçlarını ön görmeye yönelik sınıflandırma modelleri tasarlamak isteyen araştırmacılara değerli bir yol haritası sunmaktadır.

Anahtar Kelimeler: Dengesiz Veri Setleri, Makine Öğrenimi, Dengesiz Öğrenme, Sınıflandırma

ABSTRACT

COMPARISON OF HYBRID RESAMPLING METHODS IN IMBALANCED DATASETS

Hasan Ersan YAGCI

Master of Science (M.Sc.)

Graduate School of Natural and Applied Sciences

Department of Artificial Intelligence

Supervisor: Assistant Professor Nida GOKCE NARİN

May 2024, 34 pages

In machine learning classification studies, when one class is underrepresented compared to another, the imbalance in the data can lead to skewed results against the minority class. To address this, several imbalanced learning methods are used to work with imbalanced datasets. In this thesis, resampling methods, which are a part of imbalanced learning methods in machine learning, were examined. The dataset used was an imbalanced dataset consisting of traffic accidents that either resulted in litigation or did not, with cases that went to litigation being represented at a rate of 2%. Due to the low number of litigation cases in the dataset, various experiments were conducted using resampling methods both alone and in hybrid forms to overcome model bias. In this study, Random Forest, which is known for its high performance in training imbalanced datasets, was used as the machine learning algorithm. In the model developed directly with Random Forest without applying any imbalanced learning method, the precision score was 100%, recall was 4%, and F1 score was 7%. Given that recall and precision scores are more significant for the problem at hand, these metrics were focused on. In the successful hybrid model, the scores achieved were 36% for precision, 22% for recall, and 28% for F1 score. The findings of this study focused on the effectiveness of resampling methods, the advantages of hybrid approaches, and the contributions of performance criteria strategies, aiming to improve the prediction success of litigation cases in traffic accidents. The results of this study provide valuable guidelines for researchers designing classification models to predict the legal outcomes of traffic accidents

Keywords: Imbalance Dataset, Machine Learning, Imbalanced Learning, Classification

ÖNSÖZ

Yüksek lisans eğitimim ve tez çalışmam boyunca beni hem yönlendiren hem de motivasyon veren sevgili danışmanım Nida GÖKÇE NARİN'e çok teşekkür ederim.

Hasan Ersan YAĞCI

2024 - Mayıs

İÇİNDEKİLER

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

SEMBOLLER VE KISALTMALAR DİZİNİ

ADABoost Uyarlanabilir Artırma (Adaptive Boosting)

ADASYN Uyarlamalı Sentetik Örnekleme (Adaptive Synthetic Sampling)

AllKNN Tüm K En Yakın Komşular (All KNN)

ANN Yapay Sinir Ağı (Artificial Neural Network)

BorderlineSMOTE Sınır Çizgisi Sentetik Azınlık Örnek Artırma Tekniği (Borderline Synthetic Minority Oversampling Technique)

BT Bilgisayarlı Tomografi (Computed Tomography)

CNN Yoğunlaştırılmış En Yakın Komşu (Condensed Nearest Neighbour)

DT Karar Ağacı (Decision Tree)

ENN Düzenlenmiş En Yakın Komşular (Edited Nearest Neighbours)

FN Yanlış Negatif (False Negative)

FP Yanlış Pozitif (False Positive)

GBM Gradyan Artırma Makineleri (Gradient Boosting Machines)

K-Means K-Ortalama Kümeleme (K-Means Clustering)

KMeansSMOTE K-Means ile Sentetik Azınlık Örnek Artırma Tekniği (KMeans Synthetic Minority Oversampling Technique)

KNN K-En Yakın Komşu (K-Nearest Neighbors)

LR Lojistik Regresyon (Logistic Regression)

NB Naive Bayes (Naive Bayes)

RF Rastgele Orman (Random Forest)

ROS Rastgele Örnek Artırma (Random Over Sampler)

RUS Rastgele Örnek Azaltma (Random Under Sampler)

SMOTE Sentetik Azınlık Örnek Artırma Tekniği (Synthetic Minority Oversampling Technique)

SMOTEN Nominal Özellikler için Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique for Nominal Features)

SMOTENC Nominal ve Sürekli Özellikler için Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique for Nominal and Continuous Features)

SVMSMOTE SVM ile Sentetik Azınlık Örnek Artırma Tekniği (SVM Synthetic Minority Oversampling Technique)

SVM	Destek Vektör Makineleri (Support Vector Machine)
TN	Doğru Negatif (True Negative)
TP	Doğru Pozitif (True Positive)
XGBoost	Aşırı Gradyan Artırma (Extreme Gradient Boosting)

1. GİRİŞ

1.1. Dengesiz Veri Setleri

Sınıflandırma problemlerinde bir ya da bir kaç sınıfın çok az örneğe sahip olması veya çoğunluk sınıflar ile azınlık sınıflar arasındaki örnek sayısında belirgin bir dengesizlik olması, dengesiz veri problemi olarak tanımlanır. Bu durum, özellikle makine öğrenimi modellerinin performansını ve doğruluğunu olumsuz yönde etkilemektedir.

Dengesiz veri setleri üzerinden geliştirilen modellerin çoğunluk sınıfı yönünde yanlı davranması makine öğrenme sürecinde önemli problemlerden biridir. Sınıflardaki gözlem sayılarının eşit olmadığı durumlarda, az temsil edilen sınıf ya da sınıfların model tarafından öğrenilmesi güçleşir ve modelin baskın olan çoğunluk sınıfına ait yanlı bir öğrenme yapması, modelin genelleme yeteneğini olumsuz yönde etkiler.

Dengesiz veya çarpık veri kümelerinden sınıflandırıcıların öğrenilmesi, sınıflandırma problemlerinde pratikte sıklıkla ortaya çıkan önemli bir konudur (Kotsiantis vd., 2006). Özellikle sağlık ve sigorta alanlarında, nadir olayların sıklığı veya azlığı, dengesiz veri setlerinin önde gelen örneklerindendir.

Sağlık verisi üzerinde çalışan Mary ve Claret (2021)'e göre ise veri madenciliği teknolojilerinde son dönemde yaşanan gelişmeler, veri sınıflandırma sürecini büyük ölçüde etkilemiştir. Uygulamaların büyümesi veri hacmini artırmış ve dolayısıyla sınıflandırma görevini oldukça karmaşık hale getirmiştir. Birçok sınıflandırma modeli azınlık sınıfını göz ardı ederek çoğunluk sınıfından yararlanmıştır.

Covid-19 pandemisi sürecinde hastalık şüphesiyle test yaptıran kişilerin verilerinde pozitif sonuç negatif sonuca göre azınlıkta kaldığından ilgili veri setiyle yapılan sınıflandırma çalışmaların çoğunda modelin ön yargılı davrandığı sonuçları negatif tahmin etme eğiliminde olduğu görülmüştür (Javidi vd., 2021).

Sigorta ve bankacılık alanlarında da dengesiz veri setleriyle karşılaşmak mümkündür. Bankacılık işlemlerindeki sahte/dolandırıcı işlemler veya yolculuklarda / tatillerde satılan sigorta veri setleri çoğu zaman dengesizdir. Sahtecilik tahminleme çalışmalarında kullanılan verilerde sahtecilik işlemleri gerçek işlemlere göre çok daha düşük bir

miktardadır. Bu durum, dengesiz veri setleri kullanılarak geliştirilen modellerin sahtekarlık işlemlerini öğrenmesini ve gelecekte olası sahtekarlık durumlarını doğru tahmin etmesini engellemektedir (Kennedy vd., 2023).

Makine öğrenme süreçlerinde, dengesiz veri setleri ile çalışırken, veri setinin özelliklerine ve ele alınan probleme göre kullanılacak sınıflandırma yönteminin seçimi ile doğru veri ön işlem süreçlerinin gerçekleştirilmesi modelin performansı açısından oldukça kritik bir öneme sahiptir. Farklı öğrenme algoritmaları ve farklı ön işlem süreçleri farklı tahmin performanslarına neden olabilir. Burada amaç ilgili ön işlem süreçlerinden sonra en iyi tahmin başarısını verecek öğrenme algoritmasını belirlemektir.

Sağlık alanında yürütülen çalışmalarda hastalık tespiti kadar önemli olan bir diğer konuda hastalığın erken tanısıdır. Sağlık alanındaki görüntülerde de dengesizlik oranı oldukça yüksektir. Örneğin, akciğer BT görüntüleri ile yapılacak bir çalışmada tümörlü hastaların sayısı sağlıklı bireylere göre çok daha azdır. Bu durumda yapılacak çalışmalarda iki yol izlenebilir: a) eğitim ve test seti eşit sayıda tümörlü ve sağlıklı bireyden oluşacak şekilde dengeli olarak belirlenebilir b) tüm kayıtlar üzerinden dengesiz veri seti problemini aşmak için geliştirilen yöntemler ile model başarımı arttırılabilir. Dengesiz veri setleri üzerinden doğrudan gerçekleştirilen sınıflandırma işlemlerinde negatif sonuçlar yani sağlıklı bireyler fazla olacağından model baskın sınıfları öğrenme eğiliminde olacak ve çoğunlukla negatif tahminler üreterek önyargılı davranacaktır. İki sınıflı sınıflandırma problemlerinde model iki tip hata üretebilir. Hasta olmayan birini hasta olarak etiketlemek (Tip 1 hata), hasta olan bireyi sağlıklı olarak etiketlemek (Tip 2 hata). Akciğer BT örneğinde olduğu gibi, tümörlü sınıfı, tahmin etmenin önemine istinaden Tip 1 hata (Trafimow ve Earp, 2017) ihmal edilebilir.

Literatürde makine öğrenme modellerinin sınıflandırma performanslarını belirlemek için kullanılan çok sayıda metrik bulunmaktadır (Vujovic, 2021). Doğruluk (Accuracy) tahmin edilen doğru sınıflandırmaların bir oranını ifade eder. Dengeli veri setleri için doğruluk değerinin yüksek olması anlamlı iken dengesiz veri setlerinde çoğunluk sınıfını doğru tahmin ederek de yüksek bir doğruluk oranı üreteceğinden model performansı için yanıltıcı olabilir. Kesinlik (Precision), doğru pozitif tahminlerin tüm pozitif tahminlere oranını ifade eder ve yanlış pozitifleri azaltmaya odaklanır. Ancak azınlık sınıfının gerçek performansını yansıtmaz. Duyarlılık (Recall) Doğru pozitif tahminlerin tüm gerçek tahminlere oranını verir. Azınlık sınıfındaki gerçek pozitiflerin ne kadarının doğru tahmin edildiğinin ölçüsüdür ve dengesiz veri setlerinde duyarlılık değerinin yüksek olması daha anlamlıdır. Ancak yanlış negatifleri göz ardı ettiği için performans

değerlendirmesinde tek başına kullanılmaz. Kesinlik ve duyarlılığın harmonik ortalamasından elde edilen F1 skoru dengesiz veri setlerinde sınıflandırma perfomansını belirlemekte kullanılan en güçlü metriktir (Powers vd. (2011). Özellikle dengesiz veri setlerinde modelin azınlık ve çoğunluk sınıflarını tahmin performanslarını değerlendirirken metriklerin ayrı ayrı yorumlanmasına ihtiyaç vardır.
Bu tezde dengesiz veri setleri üzerinden geliştirilen hibrit yeniden öğrenme yöntemlerinin sınıflandırma performansları kesinlik, duyarlılık ve F1 skoru metrikleri üzerinden ayrı ayrı değerlendirilmiştir. Materyal ve yöntem bölümünde söz konusu metrikler detaylı bir şekilde verilmiştir.

1.1.1. İkili sınıflar

En sık karşılaşılan dengesiz veri tipidir. Birçok sınıflandırma çalışmasında pozitif ya da negatif, 1 ya da 0 olma durumu tahmin edilmeye çalışılır. Sigortacılık ve bankacılık sektöründe sahtecilik ve sağlık verileri buna örnek gösterilebilir. Sigorta firmaları açısından müşterilerinin poliçeleri yenileyip yenilemeyeceğinin tahmini, kasko ve trafik sigortası yapıldığında kaza sonrası belirlenen sigorta ücreti nedeniyle ortaya çıkan uyuşmazlıklarda doğan hukuklaşma sürecini öngörebilmek önemlidir. Bankacılıkta veya pek çok sektörde dolandırıcılık tespiti, müşteri terk eğilimi gibi bir çok problem dengesiz veri setleri üzerinden çözülmeye çalışılır. Çünkü bahsedilen durumlar nadiren gerçekleşir ama ciddi para ve prestij kaybına neden olabilir. Bu nedenle azınlık sınıfını doğru tahmin edecek modellere ihtiyaç duyulur. Burez ve Poel (2009) çalışmasında iki sınıflı ve en yaygın dengesiz veri setlerinden olan sadakat verisini kullanarak yeniden örnekleme yöntemlerinden örnek azaltma yöntemi ile modelleme gerçekleştirmişlerdir. Bu veri setinde de gösterildiği gibi sadık olmayanların olduğu sınıf oranı %13'tür.

1.1.2. Çoklu sınıflar

Çoklu sınıflarda dengesizlik problemi ikili sınıflara göre daha karmaşıktır. Tüm sınıflar arasında sadece tek bir sınıf azınlıkta olabilir ya da bir sınıf diğerlerine göre çoğunlukta kalarak dengesizlik durumu oluşturabilir. Bu veri tiplerine örnek olarak memnuniyet anketleri, ürün sınıflandırma ve sağlık verileri verilebilir.

Dengesiz veri durumlarında literatürde çoklu sınıflara uygulanan yöntemler ile ikili sınıfa uygulanan yöntemler aynıdır. Burada amacın ne olduğu önemlidir. Tarekegn (2021), çoklu sınıflarda dengesiz öğrenmeye yönelik yapılan çalışmada, bu konuya dikkat çekerek problemin ikili sınıflardan farklı bir şekilde ele alınması gerektiğini ifade etmiştir.

1.2. Makine Öğrenmesinde Dengesiz Veri Setleri

Kotsiantis vd. (2006)'nin çalışmasında, örneklerin tamamı üzerinde yüksek başarımlı bir sınıflandırma performans arayan geleneksel sınıflandırıcıların dengesiz öğrenme görevleriyle başa çıkmak için uygun olmadığı gösterilmiştir. Çünkü dengesiz veri setlerinde daha az gözleme sahip sınıf elemanları, modelin yanlı olmasından dolayı çoğunluk sınıfına atanma eğilimindedirler. Bu da dengesiz öğrenme kavramını ortaya çıkarmaktadır.

Dengesiz öğrenme problemi, az temsil edilen veri ve sınıf dağılımı çarpıklıkları ile ilgilenir. He ve Garcia (2009)'a göre dengesiz veri kümelerinin doğasında olan karmaşık özelliklerinden dolayı, bu tür verilerden öğrenmek, büyük miktarda ham veriyi verimli bir şekilde bilgi ve bilgi temsiline dönüştürmek için yeni anlayışlar, ilkeler, algoritmalar ve araçlar gerektirir. Yazarlar bu çalışmada; dengesiz veri setlerinin doğasındaki karmaşık özellikler nedeniyle, bu tür verilerden öğrenme gerçekleştirecek yeni bakış açıları, prensipleri, algoritmaları ve araçları araştırmışlardır.

Krawczyk (2016) çalışmasında; dengesiz öğrenmenin yirmi yıldan fazla süren sürekli gelişime rağmen, dengesiz verilerden öğrenmenin hala yoğun araştırmaların odak noktası olduğuna dikkat çekilmektedir. Dengesiz öğrenme alanında veri artırma teknikleri, veri azaltma teknikleri, örnekleri seçerken kullanılacak stratejiler, farklı sınıflandırma algoritmalarının kullanımı mevcuttur.

Bu tezde kullanılan yöntem veri seviyesinde yaklaşımlardan olan yeniden örnekleme (re-sampling) yöntemleridir (Estabrooks vd., 2004). Yeniden örnekleme, veri setindeki örneklerin sayısını değiştirerek, azınlık sınıfındaki örnek sayısını artırma veya çoğunluk sınıfındaki örnek sayısını azaltma amacını taşır. Bu yöntem, modelin azınlık sınıfındaki örnekleri daha iyi öğrenmesine ve bu vesileyle sınıflandırma performansının artmasına yardımcı olmaktadır.

Tezin ikinci bölümünde dengesiz veri setleri ile ilgili literatür çalışmaları, üçüncü bölümünde ise makine öğrenimi sınıflandırma algoritmaları ile yeniden örnekleme yöntemlerinden bahsedilmiştir. Üçüncü bölümde ayrıca çalışmada kullanılan veri setlerinden bahsedilmiştir.

2. LİTERATÜR

Dengesiz veri setlerinin sınıflandırma problemlerinde oluşturduğu sorun ve bunu ele alma yöntemleri ile ilgili literatürde çalışmalar mevcuttur. Bu bölümde literatürdeki çalışmalara ve çözüm yöntemlerine değinilecektir.

Yapılan çalışmalarda dengesiz veri setlerinde makine öğrenimi problemi yöntemler bazında gruplandırılmıştır. Kotsiantis vd. (2006)'nin çalışmasında, yöntemler veri seviyesinde ve algoritma seviyesinde çözümler olarak iki başlık altında incelenmiştir.

Güncel bir çalışma olan Rezvani ve Wang (2023) tarafından yapılan derleme çalışmasında dengesiz öğrenme yöntemleri üç başlığa ayrılmıştır. (1) Veri ön işleme, (2) Algoritmik yapılar ve (3) Hibrit teknikler. Burada veri ön işleme içine giren kısımda verinin modellemeye ve eğitim sürecine hazırlanması ile beraber veri seviyesinde gerçekleştirilen çalışmamızın da konusu olan yeniden örnekleme yöntemleri ele alınmıştır.

Algoritma seviyesindeki yaklaşımlardan olan eşik yöntemi (threshold method) sıklıkla başvurulan yöntemlerden biridir (Zou vs., 2016). İki sınıflı veriler için sınıflandırma eşiği varsayılan değer olarak 0,5 şeklinde belirlenmiştir. Ancak bu eşik dengesiz veriler için ideal değildir. Esposito vd. (2021) karar eşiğini ayarlamak, sınıf dengesizliği sorununu çözmek için iyi bir stratejidir şeklinde açıkladıkları yöntemlerini rastgele orman (random forest) sınıflandırma algoritması ile denemişlerdir. Rastgele orman içindeki her bir karar ağacı sınıflandırma eşiğini otomatik olarak belirleyecek bir algoritma geliştirmişler ve dengesiz veri ile yaptıkları çalışmada başarı sağlamışlardır.

King ve Zeng (2011) tarafından lojistik regresyon (logistic regression) kullanılarak yapılan çalışmada nadir olaylar olarak tanımladıkları dengesiz veri setindeki öğrenme problemini yine sınıflandırma eşiğini düzenleyerek çözmeye çalışmışlardır. Ancak dengesizlik oranı burada çok önemlidir. Dengesizlik oranı arttıkça modelin tahmin başarısı düşmektedir.

Azınlık sınıfının çok düşük oranlarda temsil edilmesi durumunda Esposito vd (2021)'in uyguladığı gibi karmaşık bir çözüm ihtiyacı doğmaktadır. Bu çalışmada iki farklı veri seti için azınlık oranı %7.7 ve %17.5 olarak verilmiştir. Dengesiz veri seti durumlarında kullanılan diğer bir algoritma seviyesinde çözüm yöntemi ise tek sınıf öğrenme (one class learning) yöntemidir. Bu yöntem veri setinde bulunan sınıflardaki dengesizliğin çok belirgin olduğu durumlarda genellikle %10 dan küçük olduğunda ele alınabilir.

Çoğunlukla azınlık sınıfı tanımlamak veya anomali tespiti için kullanılır. Bu yöntemde sadece çoğunluk sınıfına ait örneklerle model oluşturulur ve bu model sayesinde azınlık sınıfı da tespit edilebilir.

Genellikle finansal işlemlerdeki dolandırıcılık işlemlerini tespit etme ya da anomali tespitinde nadir hastalık tanımlamalarda kullanılan tek sınıf öğrenme yöntemi, Seliya vd. (2021) tarafından yapılan bu çalışmada aykırı değerlerin tespiti için kullanılmıştır. Çalışmada tek sınıf öğrenme için veri azaltmanın ve özellik seçiminin öneminden de bahsedilmiştir.

Bir diğer algoritma seviyesindeki yaklaşım ise, maliyetleri karar verme sürecine dahil etmek için sınıflar arasında sabit ve eşit olmayan yanlış sınıflandırma maliyetlerini belirlemektir. Literatürde maliyete duyarlı öğrenme (cost-sensitive learning) olarak geçen bu yöntem Domingos (2002) tarafından önerilmiştir. Maliyete duyarlı öğrenme ile maliyete duyarlı olmayan öğrenme arasındaki temel fark, farklı yanlış sınıflandırmaların ele alınma şeklinden kaynaklanmaktadır.

Maliyete duyarlı öğrenme, verilerdeki dengesizlik nedeniyle farklı maliyetlere neden olan hataları göz önüne alarak modelin eğitimini sağlar. Azınlık ya da çoğunluk sınıflarının yanlış sınıflandırılması durumunda daha yüksek maliyetlere neden olduğu durumlarda kullanılır. Burada çalışmanın amacına göre hareket etmek gerekir.

Maliyete duyarlı öğrenmeyi farklı algoritmalarla deneyen Elkan (2001), maliyetlerin mutlaka parasal olması gerekmediğini bunların zaman kaybı veya bir hastalığın ciddiyeti de olabileceğini belirtmiştir. Her algoritmada başarılı sonuçlar alınamayan bu yöntemle ilgili çalışmada detaylar da verilmiştir.

Algoritma seviyesindeki çalışmalar yeterli gelmediğinden yeni yöntemler geliştirilmiştir. 2004 yılında Burez ve Puel (2004) sınıf dengesizliğinin çok ilgi görmediğini belirterek, dengesiz veri setlerinin en sık karşılaştığı alan olan müşteri kaybı, sadakat verisi üzerinde çalışmıştır. Yöntem olarak veri seviyesinde çözüm olan yeniden örneklemeyi kullanmış ve farklı makine öğrenimi algoritmaları ile başarım sağlamıştır.

Algoritma seviyesinde geliştirilen yöntemler güncelliğini yitirmekle beraber son dönemlerde büyük data kavramının oluşmasıyla veri seviyesindeki çözümler daha sık kullanılmaya başlanmıştır. Hasanin vd. (2019) tarafından büyük verideki dengesizlik sorunu veri seviyesindeki yöntemler ile ele alınmış ve çalışmada birçok yöntem kullanılarak tahminleme başarısı artırılmıştır.

Bazı çalışmalarda her iki yöntem beraber tercih edilmiştir. Rezvani ve Wang (2023) tarafından yapılan çalışmada veri seviyesindeki çözüm yöntemleri ve algoritma seviyesindeki çözüm yöntemleri hibrit olarak kullanılmıştır.

Bu çalışmada ele alınan konu veri seviyesindeki çözüm yöntemi olan verilerin yeniden örneklenmesidir. Yeniden örnekleme yöntemleri iki başlık altında ele alınabilir. (1) Yeniden örnekleyerek çoğaltma (2) Yeniden örnekleyerek azaltma. Çalışmada bu iki yöntem bir arada kullanılarak hibrit çözüm yöntemleri de uygulanmış ve yeniden örnekleme yöntemlerinin performansı üzerinde makine öğrenimi algoritmasının seçimi doğrudan etkili olduğu gösterilmiştir.

Kalp krizinin erken tahmini için Wang vd. (2021) tarafından yapılan çalışmada dengesizlik sorununu çözmek üzere veriye ön işlem uygulamıştır. Yeniden örnekleme yöntemlerinden örnek azaltma ve örnek artırma yöntemleri doğru bir ön işlem süreci ile desteklenmiş ve rastgele orman makine öğrenimi algoritması ile dengesizlik sorununa çözüm getirilmeye çalışılmıştır.

Hanafy ve Ming (2021) sigorta verileri ile yaptıkları çalışmada, dengesiz veri seti üzerinde sigorta taleplerinin sıklığını tahmin etmeye çalışmıştır. Dengesizlik oranı %8 olan çalışmada gradyan artırma (Gradient Boosting), Adaboost ve XGBoost makine öğrenimi algoritmaları ile örnek artırma yöntemlerinden sentetik azınlık örnek artırma tekniğini (Synthetic Minority Oversampling Technique - SMOTE) kullanarak başarım sonuçlarını artırmışlardır. Tamamıyla veri seviyesinde çözüm yöntemleri ile yapılan çalışmada dengesiz verilere karşı boosting algoritmaları kullanılmıştır.

Makine öğrenimi algoritmaları ve yeniden örnekleme yöntemleri materyal ve yöntem bölümünde detaylı olarak verilmiştir.

3. MATERYAL VE YÖNTEM

3.1. Makine Öğrenimi Sınıflandırma Algoritmaları

Dengesiz veri setlerinde meydana gelen problemleri aşmak için, uygun sınıflandırma algoritmalarının ve çözüm odağının belirlenmesi çok önemlidir. Bu bölümde, makine öğrenimi sınıflandırma algoritmaları hakkında bilgi verilecektir.

3.1.1. Naive bayes

Temel mantığı Bayes teoremine dayanan Naive Bayes, olasılık temelli bir sınıflandırma algoritmasıdır (Rish, 2001). Sınıflandırma sürecinde, sınıflar arasındaki olasılıkları değerlendirir ve bu olasılıkları kullanarak belirli bir sınıf için bir örnek tahmin yapmaya çalışır. Bu algoritma, çoğunlukla metin sınıflandırma problemlerinde başarılı sonuçlar üretir ve bağımsızlık varsayımını kullanarak hesaplamaları hızlı ve anlaşılır hale getirir.
Naive Bayes algoritması, hesaplama hızı yüksek bir algoritmadır ve bu özelliği nedeni ile büyük veri kümeleri ile iyi çalışır. Basit olmasına rağmen iyi sonuçlar alınabilir. Ancak değişkenler arasındaki gerçek ilişkileri göz ardı etmesi nedeniyle bazı durumlarda yanıltıcı olabilir.
Dengesiz veri setlerinde Naive Bayes algoritması kullanılmak istendiğinde olasılık temelli bir sınıflandırma yöntemi olmasından dolayı, eğitim sonrasında bu olasılık değerlerine müdahale etme yöntemi daha etkili bir yöntemdir. Eşik yöntemi kullanılarak, varsayılan sınıflandırma eşiği olan 0,5 değiştirilerek algoritma seviyesinde yaklaşım daha mantıklıdır.

3.1.2. Karar ağaçları (decision tree)

Karar ağaçları algoritması (CART-Classification and Regression Trees) Breiman vd. (1984), ağaç yapısına benzer bir şekilde veri setini bölerek sınıflandırma yapar. Ağaçta her düğüm bir karar noktasını temsil eder ve aşağıya doğru sınıflandırma yaparak devam eder. Burada bilgi kazancı (entropy) veya jini endeksi (gini) gibi ölçütleri kullanarak en iyi bölünmeyi sağlar.

Karar ağaçları hem sınıflandırma hem de regresyon problemlerinde kullanılabilir. Pratik kullanımı sayesinde sıkça tercih edilir. Ancak büyük ve çeşitli veri kümelerinde aşırı uyumluluk (overfitting) problemi meydana gelir. Özellikle dengesiz veri setlerinde yanlı tahminler üretme ihtimali daha güçlüdür. Karar ağaçları algoritması bağımsız değişken sayısı daha az ve dengesizlik oranı nispeten daha düşük algoritmalar için daha uygundur. Sınıflandırma algoritmaları seçiminde aşırı uyumluluk problemine dikkat edilmelidir.

3.1.3. Rastgele orman (random forest)

Random forest algoritması (Breiman, 2001); yüksek performanslı, esnek ve güçlü bir algoritmadır. Hem sınıflandırma hem de regresyon problemlerinde sıkça kullanılmaktadır. Birçok karar ağacını kullanarak sınıflandırma yapmaya çalışan bir ensemble algoritmasıdır. Her ağaç, rastgele değişkenlerle ve değişken örneklemlerle sınıflandırmaya gider ve her karar ağacının sonuna göre bir karar verilir. Ensemble özelliği sayesinde genelleme yeteneği ile aşırı öğrenmeye karşı dirençli davranır.

Random forest algoritmasında hiper parametre seçimleri çok önemlidir. Karar ağacı sayısı, ağacın derinliği, karar aşamasında seçeceği değişken sayısı gibi parametrelerin optimizasyonu ile doğru ve istikrarlı bir öğrenme süreci geliştirilebilir. Dengesiz veri setlerinde, doğru seçilecek yeniden örnekleme yöntemleri ile başarılı sonuçlar alınabilir. Özellikle yüksek boyutlu ve büyük veri setlerinde başarısı yüksektir.

3.1.4. Gradyan artırma (gradient boosting)

Gradient boosting algoritması (Friedman, 2000), random forest algoritması gibi bir ensemble algoritmasıdır. Aynı şekilde birçok karar ağacı kullanır. Random forest ile temel farkı burada zayıf öğreniciler olan karar ağaçları ardışık olarak ilerler. Bu yöntem sayesinde kararlı bir öğrenim süreci sunar.

Gradient boosting adım adım her önceki öğrenicinin hatalarını düzelterek ilerler. Öğrenme oranı (learning rate) parametresi gradient boosting algoritmasında çok önemlidir. Bu parametrenin optimizasyonu ile başarılı bir sınıflandırma sonucu elde edilebilir.

Random forest algoritmasında olduğu gibi, yüksek boyutlu ve büyük veri setlerinde başarısı yüksektir. Doğru seçilecek yeniden örnekleme yöntemleri ile dengesiz veri setlerinde başarı sağlanabilir.

Gradient boosting modelleri olarak kullanılan popüler algoritmalara AdaBoost, XGBoost ve LightGBM örnek verilebilir (Omer ve Shareef, 2022).

3.1.5. Lojistik regresyon (logistic regression)

Sınıflandırma problemlerinin temel bir algoritması olan logistic regression (Cox, 1958), sigmoid fonksiyonu kullanarak 0 ile 1 arasında olasılık değerleri oluşturur. Logistic regression sadece sınıflandırma problemlerinde kullanılabilir. Hesaplama olarak hızlıdır ve büyük veri setlerinde iyi sonuçlar verir.

Dengesiz veri setlerinde kullanılmak istendiğinde eğitim sonrasında olasılık değerlerine müdahale etme yöntemi daha etkili bir yöntemdir. Eşik yöntemi kullanılarak, varsayılan sınıflandırma eşiği olan 0,5 değiştirilerek algoritma seviyesinde yaklaşım daha mantıklıdır.

3.1.6. Destek vektör makineleri (support vector machines – SVM)

SVM, Vapnik tarafından önerilen iki sınıfı ayıran bir hiperdüzlem oluşturma yöntemi ile sınıflandırma yapan bir algoritmadır (Vapnik, 1995). Hem doğrusal hem de doğrusal olmayan sınıflandırma problemlerinde etkilidir. SVM kullanımında hiper parametre

optimizasyonu çok etkilidir. Margin maksimizasyonu ve kernel seçimi, SVM'nin genelleme yeteneğini artırır. Karmaşık veri setlerinde başarılı sonuçlar verir, fakat büyük veri setlerinde eğitim süreci uzun zaman alması bir dezavantajdır.
SVM ile çekirdek fonksiyonlar ve ağırlıklandırılmış sınıfları kullanarak dengesiz veri setleri problemlerinde başarılı sonuçlar alınabilir.

3.1.7. K-en yakın komşuluk (k-nearest neighbors – KNN)

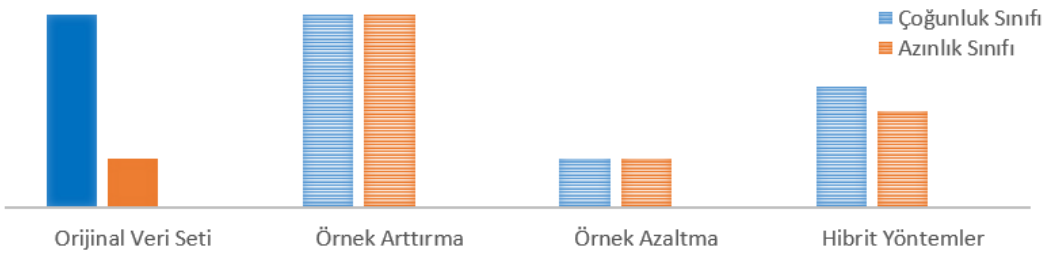
KNN (Fix ve Hoges, 1989), k tane en yakın komşusuna göre bir örneği ilgili sınıfa atayan bir algoritmadır. Sınıflandırma sırasında, örneğin yakındaki evlerin etiketlenmesi, çoğunluk sınıfını belirlemek için kullanılır. Komşu sayısını k değeri, hiperparametredir. Basit ve anlaşılır olmasına rağmen, büyük veri seti ve yüksek boyutlu uzaylarda performans sorunları olabilir.
KNN bazı yeniden örnekleme yöntemlerine temel olmuştur. Bu yöntemler arasında bulunan düzenlenmiş en yakın komşular (Edited Nearest Neighbours), tekrarlayan düzenlenmiş en yakın Komşular (Repeated Edited Nearest Neighbours) ve bu yöntemlerin varyasyonları sayılabilir (Tomek, 1976).
Dengesiz veri setleriyle çalışırken doğru örnekleme yöntemleri ile başarı arttırılabilir. Pagan vd. (2023) tarafından yapılan çalışmada belirtildiği gibi verinin doğru yöntem ile ölçeklendirilmesi de önemli bir husustur.

3.1.8. Yapay sinir ağları (artificial neural networks – ANN)

Biyolojik sinir sisteminden esinlenen, yapay sinir ağları, veri içerisindeki karmaşık örüntüleri öğrenmeye çalışır (McCulloch ve Pitts, 1943). Görüntü, ses ve metin gibi büyük ve karmaşık veri setlerinde etkili sonuçlar verir. Katmanlar arasındaki ağırlıkları öğrenerek karmaşık ilişkileri modelleyebilir. Bununla birlikte, yeterli veri olmadığında aşırı öğrenme eğilimindedir ve eğitim süreçleri hesaplama gücüne ihtiyaç duyar.
ANN ile dengesiz veri setleri ile çalışırken, birden fazla ANN modelini bir araya getirerek ensemble model mimarisi ile başarım artırılabilir. Bunun yanında önceden eğitilmiş bir ANN modelinin parametreleri ile öğrenimi aktararak (transfer learning) dengesiz veri seti problemleri ile başa çıkılabilir.

3.2. Yeniden Örneklem Yöntemleri

Dengesiz öğrenme problemini aşmak için bu çalışmada kullanılan yöntemler, veri seviyesinde çözüm yöntemlerinden olan yeniden örneklem yöntemleridir. Yeniden örneklem yöntemleri örnek artırarak yeniden örnekleme ve örnek azaltarak yeniden örnekleme olarak temelde ikiye ayrılırmaktadır. Son dönemlerde hibrit yaklaşımların da kullanılmasıyla Şekil 3.1'de verildiği gibi üç başlık altında incelenmeye başlamıştır (Kimura, 2022).



Şekil 3.1. Yeniden örnekleme

3.2.1. Örnek artırma (oversampling) yöntemleri

Örnek artırma yöntemlerinin amacı, az temsil edilen sınıflardan daha fazla örnek elde etmek yoluyla denge sağlamaktır. Bu yöntemler, modelin az temsil edilen sınıfları daha etkili bir şekilde öğrenmesine yardımcı olmak için mevcut örnekleri kullanarak sentetik örnekler oluşturabilir ya da mevcut örnekleri farklı şekillerde kopyalayabilir (Mohammed vd., 2020).

3.2.1.1. Rastgele örnek artırma (random over sampler – ROS)

Oversampling yöntemlerinin en yalını olan ROS, rastgele olarak azınlık sınıfına ait örnekleri veri içinde çoğaltır (Hayati vd., 2021). Azınlık sınıftan rastgele seçilen örnekler birden fazla kez kullanılarak yeni bir veri oluşturulur. Bu sayede azınlık sınıfı artırılarak dengesizlik azaltılır.

ROS, etkin ve uygulaması basit bir yöntemdir. Verilecek parametrelerle azınlık sınıfının ne kadar artırılacağı veya yeniden örneklenen verinin birbirini tekrarlayıp

tekrarlamayacağı kontrol edilebilir. Menardi ve Torelli (2014) çalışmalarında buna dikkat çekmiştir. ROS'un aşırı uyumluluğa gidebileceğini belirtmiş ve parametrelerin öneminden bahsetmiştir.

3.2.1.2. Sentetik azınlık örnek artırma tekniği (synthetic minority oversampling technique - SMOTE) ve çeşitleri

Oversampling yöntemleri içinde en sık kullanılan yöntem olan SMOTE (Chawla vd., 2002), azınlık sınıfına ait olan örnekler ile sentetik örnekler oluşturarak azınlık sınıfının veri içindeki temsilini güçlendirir. SMOTE bunu yapmak için azınlık sınıfındaki örnekler için en yakın komşuları (KNN) belirler ve komşular arasındaki mesafeyi kullanarak sentetik örnekler oluşturur.

Bu sayede azınlık sınıfındaki yoğunluğu artırarak dengesizlik sorununu azaltır. Ancak, aşırı sentetik örnekler üretme ve aşırı uyum riski gibi bazı zorlukları vardır. Bu nedenle, SMOTE'nin dikkatlice uygulanması ve dengeli bir sentetik örnek oluşturma stratejisinin belirlenmesi önemlidir.

Dengesiz öğrenme üzerine yapılan veri seviyesindeki çözümleri içeren birçok çalışmada SMOTE kullanılmış ve bu çalışmalar neticesinde SMOTE varyantları meydana gelmiştir. Bu varyantlar, Han vd. (2005) tarafından yayınlanan Borderline SMOTE, Last vd. (2017) tarafından çalışılan KMeans SMOTE ve Nguyen vd. (2009) tarafından yayınlanan SVMSMOTE yöntemleridir.

Borderline SMOTE, sınıflar arasındaki sınırları belirlemek için KNN kullanır. Azınlık sınıfında bulunan örneklerden sınıf sınırlarına yakın olanlar seçilir ve bu örneklerin en yakın komşuları kullanılarak sentetik örnekler oluşturulur. Bu yöntem sınıf sınırlarına yakın olan örneklerin çoğaltılmasını sağlar.

KMeans SMOTE, Kmeans (K ortalama) algoritmasını kullanır. Veri setini gruplamak için KMeans kümelemeyi kullanır ve her kümeden bir örnek seçer. Seçilen örnekler arasında en yakın komşu tekniğini kullanarak sentetik örnekler oluşturur. Burada amaç küme merkezlerine yakın sentetik örnekler oluşturulduğu için sınıf sınırlarını korumaktır.

SVMSMOTE, sınıf dengesizliğini gidermek üzere SVM makine öğrenimi algoritmasını kullanır. Azınlık sınıftaki örneklerden sınıf sınırlarına yakın olanlardan en yakın komşu tekniği kullanılarak sentetik örnekler oluşturulur. Burada sınır SVM tarafından belirlenir.

Nominal Özellikler için Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique for Nominal Features - SMOTEN) ve Nominal ve Sürekli

Özellikler için Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique for Nominal and Continuous features - SMOTENC) hakkında yapılan bu çalışmada, SMOTEN ve SMOTENC yöntemlerinin hem nominal hem de sürekli verilerde de başarılı olduğu gösterilmiştir (Mukherje ve Khushi, 2021).

3.2.1.3. Uyarlamalı sentetik örnekleme (adaptive synthetic sampling – ADASYN)

ADASYN, He vd. (2008) tarafından; farklı azınlık sınıfı örnekleri için, öğrenmedeki zorluk düzeylerine göre ağırlıklı bir dağılım kullanmakta olduğu ve öğrenmesi daha zor olan azınlık sınıfı örnekleri için, öğrenmesi daha kolay olan azınlık örneklerine kıyasla daha fazla sentetik veri ürettiğinden bahsederek tanıtılmıştır.
ADASYN yöntemi de SMOTE yönteminin bir varyantı olarak kabul edilir. ADASYN, SMOTE yöntemlerinde olduğu gibi her bir azınlık sınıf için sentetik örnekler üretir ve bu sentetik örnekler azınlık sınıfındaki daha az yoğun bölgelerde üretilir. Bu, azınlık sınıfın yoğunluğu az olan bölgelerde daha fazla dikkate alınmasını sağlar. ADASYN, doğrusal olmayan sınıflar arasındaki dengesizlikleri ele almak için iyi bir seçimdir.
ADASYN yönteminin iki amacı vardır. Birincisi sınıf dengesizliğinin getirdiği önyargıyı azaltmak. Diğeri ise azınlık sınıfa ait örnekleri uyarlanabilir bir şekilde zor örneklere doğru kaydırmaktır.

3.2.2. Örnek azaltma (undersampling) yöntemleri

Örnek azaltma yöntemleri, çoğunluk sınıfına ait örneklerin sayısını azaltarak veri setinde denge oluşturmayı amaçlar. Bu yöntemler modellerin çoğunluk sınıfı hakkında çok fazla bilgi edinmesini önler ve az temsil edilen sınıfların daha iyi öğrenilmesini sağlar.
Örnek azaltma yöntemleri, prototip üretme ve prototip seçme yöntemleri başlıkları altında toplanmıştır. Prototip üretme mevcut örnekleri yeniden örnekleyerek oluşturur. Bu yöntemde, çoğunluk sınıfı ile beraber azınlık sınıfında da kayıplar olabilir. Prototip seçme ise hangi çoğunluk sınıfı örneklerinin çıkarılacağını belirleyerek ilerler. Bu sebeple kullanım amacına uygun yöntem seçilmelidir.

3.2.2.1. Küme merkezleri (cluster cendroids)

Cluster Cendroids yöntemi, prototip üreterek sınıflar arasında denge sağlar Yen ve Len (2006). Seçilecek bir kümeleme algoritması ile sınıfların küme merkezleri belirlenir ve her sınıf için orijinal örnekler yerine küme merkezlerinde yeni örnekler üretilir.
Bu yöntemden en iyi faydayı sağlayabilmek için kullanılacak verinin kümeleye uygun olması gerekmektedir.

3.2.2.2. Rastgele örnek azaltma (random under sampler – RUS)

RUS, kontrollü bir prototip seçerek örnek azaltma yöntemidir (Ali vd., 2019). Örnek azaltma yöntemlerinin uygulaması en basit ve en yalın halidir. RUS, çoğunluk sınıfından rastgele örnekler seçerek verilecek parametreler doğrultusunda azaltıma gider ve sınıflar arasındaki dengesizliği azaltır.
Veri setinde yaptığı rastgele seçimlerle azaltmaya giderken veri kaybına neden olabilir, bu yüzden seçimi ve uygulama yöntemi doğru belirlenmelidir.

3.2.2.3. Yakın örnekler (near miss)

Mani ve Zang (2003) dengesiz verilerdeki dengeleme yöntemlerine KNN kullanarak bir alternatif geliştirmiştir. Yaptıkları çalışmada KNN kullanarak örnekler arasında ilişki kurarak örnek azaltımına gitmişler ve bunu KNN'in uygulama basitliği ile elde etmişlerdir.
Near Miss, azınlık sınıfında olan örnekler ile çoğunluk sınıfı örnekleri arasındaki uzaklıkları dikkate alarak örnek seçimine gider. Near Miss yöntemi de kontrollü bir prototip seçerek örnek azaltma yöntemidir. 3 adet uygulama yöntemi vardır, veriye ve çalışmaya uygun olarak doğru yöntem ve parametrelerle dengesiz veriler üzerine çalışma yürütülebilir.

3.2.2.4. Tomek bağlantıları (tomek links)

Tomek Links, Tomek (1976) çalışmasında tanımlanmıştır. Tomek bağlantısı farklı sınıflara ait iki örnek birbirine en yakın komşu olduğunda ortaya çıkar.

Azınlık ve çoğunluk sınıflarına ait örneklerden en yakın komşu tekniğine göre komşu olanlar çift olarak tomek bağlantısı olarak etiketlenir. Devamında verilecek parametreye göre ya iki örnek ya da çoğunluk sınıfına ait olan örnek veri setinden çıkarılarak dengesizlik sorunu çözülmeye çalışılır. Burada amaç azınlık ve çoğunluk sınıfı arasındaki sınırları belirlemektir.

3.2.2.5. Düzenlenmiş en yakın komşular (edited nearest neighbours)

Edited Nearest Neighbours (Wilson, 1972), en yakın komşular tekniğini kullanarak, verilen parametre doğrultusunda yakın komşularıyla uyuşmayan örnekleri kaldırarak veri kümesini dengelemeye çalışır.
Veri kümesindeki her a örneği için, en yakın komşuları hesaplanır. Eğer a çoğunluk sınıfı örneğiyse ve en yakın komşuları tarafından yanlış sınıflandırılmışsa, a veri kümesinden kaldırılır. Alternatif olarak, eğer a bir azınlık sınıfı örneği ise ve en yakın komşuları tarafından yanlış sınıflandırılmışsa, a'nın komşuları arasındaki çoğunluk sınıfı örnekleri kaldırılır.

3.2.2.6. Tekrarlanmış düzenlenmiş en yakın komşular (repeated edited nearest neighbours)

Repeated Edited Nearest Neighbours, Edited Nearest Neighbours (Tomek, 1976) yöntemini tekrarlayan bir yöntemdir. Bu sayede veri setinde daha fazla örnek azaltımı olmaktadır.

3.2.2.7. Tüm k-en yakın komşular (all KNN)

All KNN yöntemi de Tomek tarafından düzenlenmiş en yakın komşular üstüne yaptığı çalışmada, Repeated Edited Nearest Neighbours yönteminden farklı olarak her tekrar işleminde en yakın komşu sayısını bir arttırarak ilerler (Tomek, 1976).

3.2.2.8. Yoğunlaştırılmış en yakın komşu (condensed nearest neighbour)

Condensed Nearest Neighbour yöntemi Hart (1968) tarafından yapılan çalışmaya dayanmaktadır. Veri setinden seçtiği rastgele örnekleri diğer örneklerle kıyaslayarak doğru sınıfa atamaya çalışır, atayamadığı örnekleri ayırır. Bu yöntem, gereksiz veya fazla örnekleri çıkararak veri setini daha küçük ve daha temsilci hale getirerek sınıf dengesizliğini azaltmayı amaçlar.

3.2.2.9. Komşuluk temizleme kuralı (neighbourhood cleaning rule)

Neighbourhood Cleaning Rule, Condensed Nearest Neighbour yönteminden farklı olarak verileri temizlemek üzerine çalışılmış bir yöntemdir (Laurikkala, 2001).
Azınlık örneklerin doğru bir şekilde sınıflandırılabilmesi için çoğunluk sınıfındaki örneklerle olan ilişkilerini dikkate alarak, en yakın komşuya göre veri setini temizler. Sınıf dengesizliği sorununu çözmek için bu yöntem etkili olabilir ve çoklu sınıflarda başarılı sonuçlar verir.

3.2.2.10. Tek taraflı seçim (one sided selection)

One Sided Selection çoklu sınıflarda yeniden örneklemeyi destekler Kubat ve Matwin (1997)'nin belirttiği üzere, azınlık sınıfına ait örnek, çoğunluk sınıfından en az bir komşusuyla aynı sınıfa aitse, bu örnek korunur ve seçilenler arasına eklenir. Seçilen örnekler ile azınlık sınıfı örnekleriyle birleştirilir. Çoğunluk sınıfı için seçilen örnekler ise karışık örnekleri temizlemek için kullanılır. Çoklu sınıflardaki dengesiz veri seti problemlerinde tercih edilmektedir.

3.2.2.11. Örnek zorluğu eşiği (instance hardness threshold)

Instance Hardness çoklu sınıflardaki yeniden örneklemeyi destekler ve yeniden örneklenecek olan örneklerden, doğru sınıflandırılma olasılığı en yüksek olanları korur (Threshold Smith vd., 2014).
Instance Hardness Threshold yöntemi, bir eşik değer belirler ve model bu eşik değerinden daha zor örnekleri kabul etmez veya dikkate almaz. Karmaşık veya belirsiz

örneklerin etkisi azaltılır ve model daha güvenilir ve istikrarlı örnekler üzerinde odaklanır.

3.2.3. Hibrit yaklaşımlar

Hibrit yaklaşımlar hem örnek artırma hem de örnek azaltma yeniden örnekleme yöntemlerinin beraber kullanılması durumudur. Hibrit yaklaşımlar, dengesiz veri setleri üzerinde daha etkili sınıflandırma sağlamayı amaçlar.
Batista vd. (2003) çalışmalarında hibrit bir yöntem deneyerek, SMOTE ve tomek links yöntemlerini bir arada kullanmışlardır. Chris vd. (2009) tarafından hibrit olarak denemeler yapılmış ve başarılı sonuçlar alındığı görülmüştür.
Bu çalışmada yeniden örnekleme yöntemleri tek başına uygulandığı gibi hibrit olarak da kullanılmıştır. Bulgular kısmında detayları verilmiştir.

3.2.4. Veri seti

Çalışmada yeniden örnekleme yöntemlerinin tek başına ve hibrit olarak uygulandığı 4 adet veri seti kullanılmıştır ve söz konusu veri setlerinin özellikleri Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Veri setleri ve özellikleri

Veri Seti Adı	Satır Sayısı	Sütun Sayısı	Dengesizlik Oranı
Kredi Kartı Verisi	1319	11	%22
Covid-19 Verisi	158962	11	%12
Sigorta Verisi	47725	45	%1.5
Hukuklaştırma Verisi	817738	155	%3

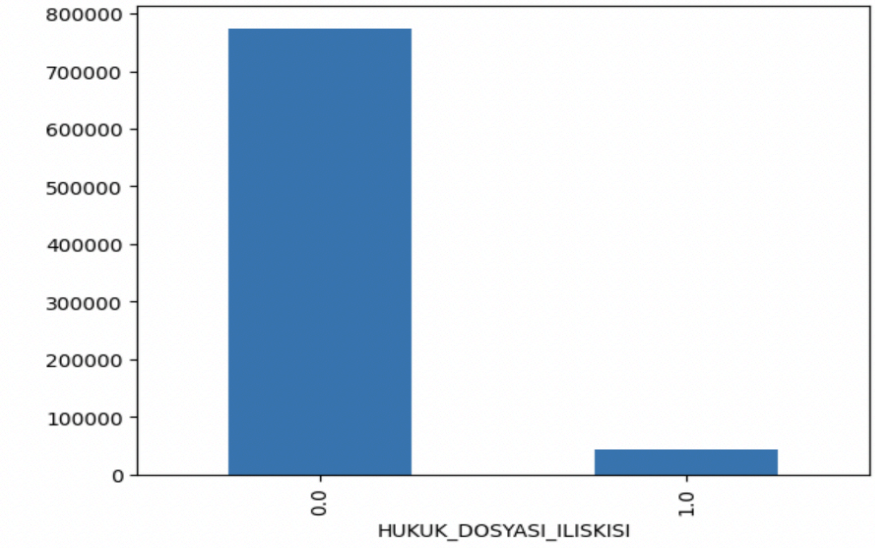
Kredi kartı verisi 1319 satır ve 11 sütundan oluşmaktadır. Diğer veri setlerine göre daha küçük boyutludur ve %22 dengesizlik oranı ile dengesizlik oranı diğer veri setlerine göre düşüktür. Kredi kartı verisi uygulanacak makine öğrenimi algoritmasının seçimi için kullanılmıştır.

Covid-19 verisi 158962 sütun ve 11 sütundan oluşmaktadır. Örnek sayısı yüksek olsa da sütun sayısı bakımından küçük boyutlu kabul edilebilir. Dengesizlik oranı ise %12'dir. Bu veri seti üzerinde yeniden örnekleme yöntemleri denenmiştir.

Sigorta verisi ise 47725 satır ve 45 sütundan oluşmaktadır. Dengesizlik oranı % 1.5'tur. Hukulaştırma verisine hacim ve dengesizlik olarak en yakın veri setidir. Bu veri seti üzerinde de yeniden örnekleme yöntemleri denenmiştir.

Tez çalışmasının temel amacı bir sigorta şirketine ait trafik kazalarındaki hukuklaşmaya (dava sürecine) giden örnekleri tahminleyebilmektir. Çalışmada kullanılan temel veri seti özel bir sigorta firmasına ait trafik sigortası poliçe, onarım, hukuk ilişkisini gösteren anonimleştirilmiş hukuklaşma verisidir. KVKK ve gizlilik sözleşmeleri nedeniyle Firma ismi çalışmada verilmemiştir. Veri setinin bağımlı değişkeni olan "hukuk_dosyası_ilişkisi", ilgili kayıtlardaki kaza/hasarların hukuklaşma, yargıya intikal, sürecine gidip gitmediğini göstermektedir. Hukuklaştırma veri setinde; her kayıtta araç ile ilgili hasar durumu, onarım aşamaları, değişen parçalar, maliyet, plaka, kaza bilgileri gibi veriler bulunmaktadır.

Çalışma kapsamında kullanılan data 817738 satır ve 155 sütundan oluşmaktadır. Hukuklaşmaya giden ve gitmeyen olmak üzere 1 ya da 0 olarak 2 sınıf mevcuttur. Bağımlı değişken olarak ele alınan "hukuk_dosyası_ilişkisi" değişkeninde dengesizlik durumu mevcuttur (Şekil 3.2.).

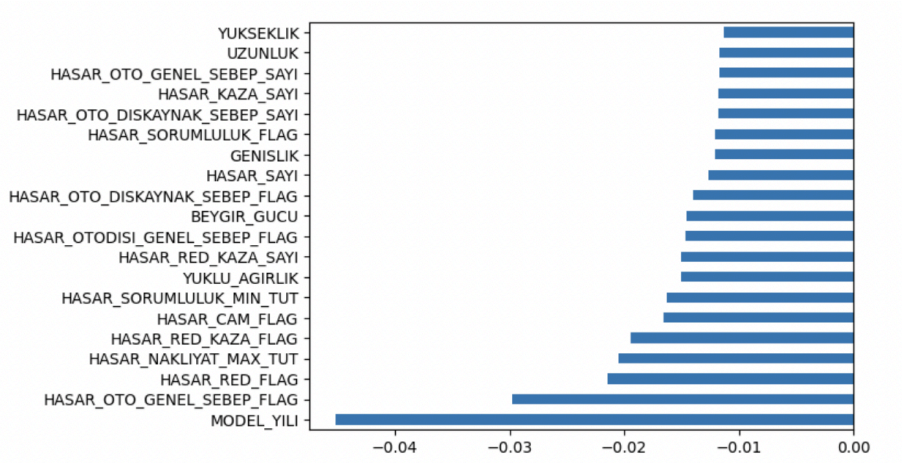


Şekil 3.2. Sınıf dağılımı

Hukuklaşmaya giden 1 sınıfı için veri sayısı 43324, hukuklaşmaya gitmeyen 0 sınıfı için ise veri sayısı 774414 adettir. Verinin sadece yüzde 5'i çalışma için önemli olan ve

tahminlemede zorlanılan hukuklaşmaya giden verilere aittir. Bu durum büyük bir dengesizlik sorunu yaratmakta ve yalın bir şekilde eğitilen modeller tüm sonuçları 0 sınıfına atayacak şekilde yanlı davranmaktadır. Kesinlik, duyarlılık ve f1 skor performansları başarısız çıkmaktadır.

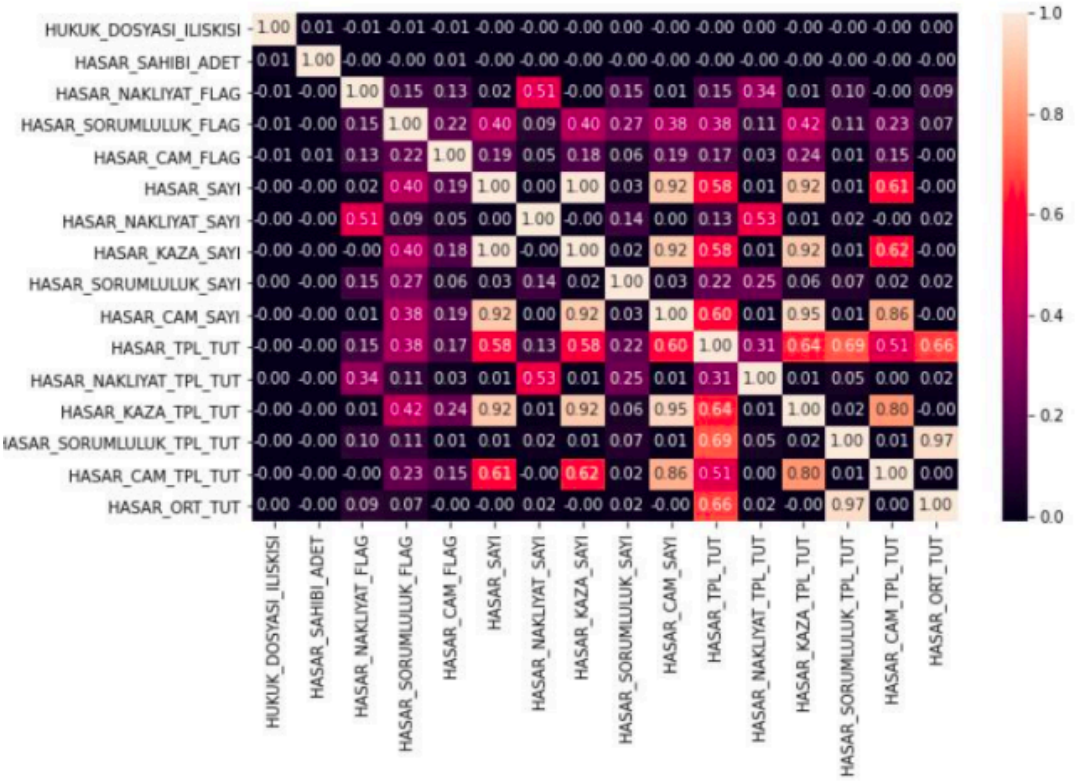
Veri setinde bulunan 154 bağımsız değişkenin 104 tanesi sayısal, 51 tanesi ise kategorik verilerden oluşmaktadır. Veri setini makine öğrenimi sürecine hazırlamak üzere yapılan keşifsel veri analizi ile sayısal olmayan veriler sayısala dönüştürülmüş, bağımlı değişken ile bağımsız değişkenler arasında korelasyon kontrolü yapılarak düzenlemeye gidilmiştir (Şekil 3.2.). Grafikte bağımlı değişken ile bağımsız değişkenler arasındaki ters korelasyon durumu gösterilmiştir. Yükseklik ve uzunluk değişkenleri, grafikte görüldüğü gibi 0 değerine çok yakındır. Korelasyonu düşük olan değişkenler modele dahil edilmemiştir.



Şekil 3.3. Bağımlı değişken ile Bağımsız değişkenler arasındaki korelasyon durumu

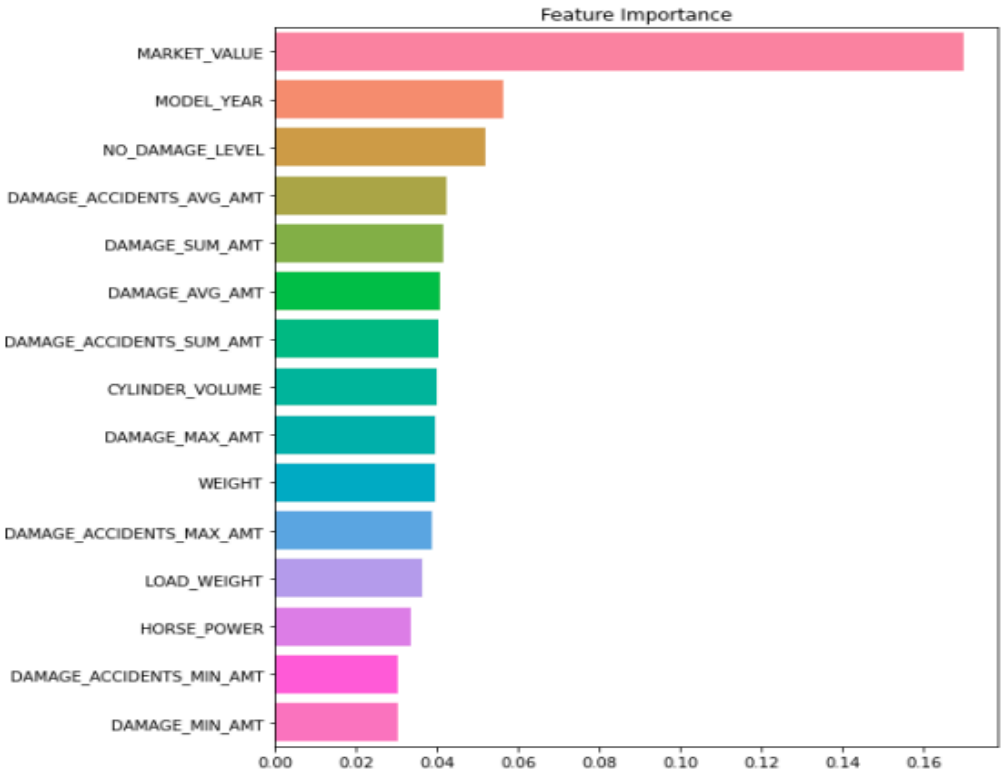
Makine öğrenmesinde özellik seçiminde bağımsız değişkenlerin kendi aralarındaki korelasyonlar incelenir. Yüksek korelasyonlu değişkenlerin birlikte kullanımı modelin benzer bilgiyi birden fazla kez kullanmasına neden olacağından hem model karmaşıklığı ve hesaplama gücünü arttırır hemde aşırı uyum problemine neden olabilir. Bu çoklu bağlantı (multicollinearity) sorunu olarak adlandırılır. Yüksek korelasyonlu bağımsız değişkenler modele birlikte dahil edilirse modelin güvenilirliği azalır ve model kararsız hale gelir. Bu çalışmada, özellik seçinimde bağımsız değişkenler arasındaki korelasyonu incelemek için Şekil 3.3'de bir parçası verilen bir ısı haritası (heatmap) kullanılmıştır. Isı haritası üzerinde yüksek korelasyona sahip değişkenler belirlenmiş ve bu değişkenler

çıkarılmıştır. Bu işlem, modelin doğruluğunu ve genel performansını artırmak için yapılmıştır.



Şekil 3.4. Bağımsız değişkenler korelasyon durumu

Hukuklaşma veri seti üzerinde veri ön işleme kapsamında değişken seçimine gidilmiştir (Gunes vd., 2023). Bunun yanında rastgele orman algoritmasının öznitelik seçimi yöntemi ile modelin başarımına katkıda bulunacak değişkenler belirlenmiştir (Şekil 3.4.). Rastgele orman algoritması ile her bir değişkenin modeldeki önem derecesi görselleştirilmiştir. Bu grafik, hangi değişkenlerin model performansı için daha kritik olduğunu ve hangi değişkenlerin daha az etkili olduğunu açıkça göstermektedir. Daha yüksek önem derecesine sahip değişkenler, modelin tahmin gücüne daha fazla katkıda bulunurken, daha düşük önem derecesine sahip değişkenler, modele olan etkisinin sınırlı olduğunu göstermektedir. Bu sayede, modelin performansını artırmak ve gereksiz değişkenlerden arındırmak mümkün olmuştur.



Şekil 3.5. Rastgele orman öznitelik seçimi

Veri setinde bulunan boş (null) değerlere yönelik özellik mühendisliği (feature engineering) yapılarak mevcut boş değerler doldurulmuştur. Doldurulamayan, eksik olduğu değerlendirilen, veri toplanırken hata oluştuğu gözlemlenen kayıtlar veri setinden çıkarılmıştır..

Tüm ön işlem süreçlerinin sonunda modelde ve yeniden örnekleme yöntemlerinde kullanılan veri seti 156638 satır ve 81 sütundur. Azınlık sınıfı 5116, çoğunluk sınıfı ise 151522 satırdan oluşmaktadır. Dengesizlik oranı %3'tür.

3.3. Başarım Ölçütleri

Sınıflandırma problemlerinin başarım ölçütleri, modelin performansını ölçmek ve sonuçları değerlendirmek için kullanılır (Goutte ve Gaussier, 2005). Makine öğrenimi süreci sonunda tahminler ile gerçek değerler karşılaştırılarak modelin; doğruluğu (accuracy), kesinliği (precision), duyarlılık (recall) ve F1 skoru (F1 score) ölçütleri oluşturulur.

Kesinlik : Doğru pozitiflerin (True Positives - TP) toplam pozitif tahminlere (True Positives + False Positives - TP + FP) oranıyla hesaplanır. Kesinlik, bir sınıfa tahmin edilen örneklerin ne kadarının gerçekten o sınıfa ait olduğunu gösterir. Kesinlik değerinin yüksek olması, modelin bir sınıfı tahmin ettiğinde genellikle doğru tahminleme yaptığını gösterir.

Duyarlılık : Doğru pozitiflerin (True Positives - TP) toplam gerçek pozitiflere (True Positives + False Negatives - TP + FN) oranıyla hesaplanır. Duyarlılık, doğru bir şekilde tahmin edilen gerçek pozitiflerin miktarını gösterir. Duyarlılık değerinin yüksek olması, modelin o sınıfa ait örnekleri genellikle kaçırmadığını gösterir.

F1 Skor: Kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. Dengesiz veri setleri için uygun bir performans ölçütüdür. F1 skor, bir sınıflandırma modelinin hem kesinlik hem de duyarlılık performansını ölçer.

Doğruluk: Doğru sınıflandırılan tüm örneklerin toplam örnek sayısına oranı ile tanımlanır. Bir sınıflandırma modelinin toplam doğruluk oranını hesaplar. Başka bir deyişle, modelin tüm sınıfları doğru bir şekilde sınıflandırma yeteneğini göstermektedir. Doğruluk değeri 0 ve1 arasında değer alır ve 1'e yakın değer modelin performansının daha iyi olduğunu ifade eder. Dengesiz veri setlerinde doğruluk değeri yüksek olsada azınlık sınıfını doğru tahmin edilememiş olma ihtimali olduğu için bu metriği kullanmak uygun değildir. Ancak sınıf dağılımı eşit olduğunda kullanılabilir.

Duyarlılık ve kesinlik arasında bir denge vardır. Yüksek kesinlik, düşük duyarlılık ile sonuçlanabilir ve tam tersi de geçerlidir. Bir modelin kesinlik değeri yüksekse, o sınıfı tahmin ederken nadiren yanlış alarm verir, ancak sınıfın bazı örneklerini kaçırabilir. Bununla birlikte, bir modelin duyarlılık değeri yüksekse, çoğu sınıfı doğru bir şekilde tahmin edebilir, ancak diğer sınıflardan örnekleri yanlışlıkla bu sınıfa dahil edebilir. Bu nedenle, özellikle dengesiz verilerle çalışırken kesinlik ve duyarlılık değerlerinin birlikte değerlendirilmesi, bir sınıflandırma modelinin performansının daha kapsamlı bir şekilde anlaşılmasına yardımcı olur. Bu iki değerin harmonik ortalaması olan F1 skor da kullanılmalıdır.

Başarım ölçütleri üzerinden değerlendirme yaparken çalışmanın amacına göre ilerlenmeli, kullanılan veri seti iyi analiz edilerek sonuçlar buna göre ele alınmalıdır. Çalışmanın amacına göre bazı metrikler ihmal edilebilir, performans sadece bir metrik üzerinde toplanmak istenebilir. Bu çalışmada bizim kullandığımız başarım ölçütleri, dengesiz öğrenmede kullanılması gereken, kesinlik, duyarlılık ve F1 skor'dur.

4. BULGULAR

Çalışmada python programlama dili kullanılarak hukuklaştırma veri seti üzerinde, veri seti başlığında belirtilen ön işlemler ve data temizliği yapılmıştır. Makine öğrenimi algoritmaları uygulanmadan önce veri setleri hold out metodu gereği yüzde 20 test, yüzde 80 eğitim verisi olarak ikiye ayrılmıştır (Kohavi, 2001). Yapılan tüm farklı uygulamalarda veri setleri aynı oranda bölünmüştür.

Bu çalışmanın amacı dengesiz öğrenmede veri seviyesinde çözüm yöntemlerinden olan yeniden örneklemeyi değerlendirmektir. Bu nedenle seçilecek makine öğrenimi algoritmasının aşırı uyumluluğa karşı başarılı sonuçlar vermesi ve yöntemlerin testi için sabit tutulması gerekmektedir.

Makine öğrenimi algoritması seçiminde bağımsız değişken sayısının fazlalığı önemli olmuştur. Yüksek boyutlu verilerle sınıflandırma, makine öğreniminde önemli bir zorluktur çünkü yüksek boyutlu verilerdeki özelliklerin çokluğu, anlamlı modellerin tanımlanmasını zorlaştırır, bu da aşırı uyum ve sınıflandırma performansının düşmesine yol açar (Tran, 2023).

Çizelge 4.1. Kredi kartı veri seti makine öğrenimi sınıflandırma algoritmaları sonuçları

Makine Öğrenimi Algoritması	**Sınıf**	**Kesinlik**	**Duyarlılık**	**F1 Skor**
KNN	Azınlık	0.85	0.92	0.89
	Çoğunluk	0.98	0.95	0.96
SVM	Azınlık	0.74	1.00	0.85
	Çoğunluk	1.00	0.89	0.94
Karar Ağaçları	Azınlık	0.92	0.96	0.94
	Çoğunluk	0.99	0.98	0.98
Rastgele Orman	***Azınlık***	***0.92***	***1.00***	***0.96***
	Çoğunluk	***1.00***	***0.97***	***0.99***

Makine öğrenimi algoritmasının seçimi için 1319 örnekten oluşan ve daha düşük boyutlu, 11 bağımsız değişkenli, kredi kartı harcama verisi ile makine öğrenimi sınıflandırma algoritmaları denenmiştir. Makine öğrenimi sınıflandırma algoritmalarından literatürde en fazla kullanılan KNN, karar ağaçları, SVM ve rastgele orman denenmiş ve yapılan denemelerde daha düşük boyutlu bir veri olmasına rağmen azınlık sınıfı olan evet örneklerinin tahmin başarım ölçütlerinde rastgele orman yüksek başarı vermiştir (Çizelge 4.1).

Bu kapsamda hem dengesiz verilerde hem de yüksek boyutlu verilerde başarılı sonuçlar veren ensemble model olan rastgele orman seçilmiştir.

Makine öğrenimi algoritmasının seçimi sonrasında Covid-19 veri seti üzerinde yeniden örnekleme yöntemleri denenmiştir. Veri seti üzerinde veri temizliği ve mühendisliği kapsamında; boş değerlerin düşürülmesi ve aykırı değerlerin çıkarılması işlemleri yapılmıştır. Yeniden örnekleme yöntemleri tek başına uygulandığı gibi hibrit olarak da uygulanmıştır.

Covid-19 veri seti üzerinde başarılı sonuçlar dengesizlik oranının azaldığı hibrit yöntemler ve tek başına RUS yöntemi vermiştir. Burada Covid-19 veri setinin düşük boyutlu (az sayıda sütun) olması bu nedenle bir çok örneğin kendini tekrarlaması etkin olmuştur. Bu sebeple RUS metodu rastgele örnek azaltarak model başarısını yükseltmiştir (Çizelge 4.2).

Çizelge 4.2. Covid-19 veri seti yeniden örnekleme yöntemleri sonuçları

Yöntem	Dengesizlik Oranı	Sınıf	Kesinlik	Duyarlılık	F1 Skor
Yeniden Örnekleme Yöntemi Kullanılmadan	%12	Azınlık	0.61	0.57	0.59
		Çoğunluk	0.96	0.97	0.96
SMOTE	%33	Azınlık	0.55	0.63	0.59
		Çoğunluk	0.97	0.96	0.96
Random Under Sampler	***%33***	***Azınlık***	***0.53***	***0.64***	***0.58***
		Çoğunluk	***0.97***	***0.95***	***0.96***
SMOTE + Random Under Sampler	***%41***	***Azınlık***	***0.53***	***0.64***	***0.58***
		Çoğunluk	***0.97***	***0.95***	***0.96***
ENN	%15	Azınlık	0.53	0.63	0.58
		Çoğunluk	0.97	0.95	0.96
SMOTE + ENN	%17	Azınlık	0.12	0.76	0.21
		Çoğunluk	0.96	0.50	0.66
ENN + SMOTE + Random Under Sampler	***%33***	***Azınlık***	***0.53***	***0.64***	***0.58***
		Çoğunluk	***0.97***	***0.95***	***0.96***

Çalışmada kullanılan anonimleştirilmş sigorta verisi olan hukuklaşma veri setine en yakın veri seti olan sigorta verisi üzerinde yine aynı yöntemler sırayla denenmiştir. Veri ön işleme kapsamında; boş değerlerin düşürülmesi ve aykırı değerlerin çıkarılması işlemleri yapılmıştır.

Dengesizlik oranının %1.5 olduğu sigorta verisinde başarılı sonuçlar hibrit yöntemler ile alınmıştır. F1 Skor'un yanında başarım olarak duyarlılık ölçütüne de bakılmıştır. %3 olan duyarlılık başarılı yöntemlerde %18'e çıkarılmıştır ve çoğunluk sınıfının başarısı da göz önüne alınmıştır (Çizelge 4.3).

Çizelge 4.3. Sigorta veri seti yeniden örnekleme yöntemleri sonuçları

Yöntem	Dengesizlik Oranı	Sınıf	Kesinlik	Duyarlılık	F1 Skor
Yeniden Örnekleme Yöntemi Kullanılmadan	%1.5	Azınlık	0.16	0.03	0.05
		Çoğunluk	0.99	1.00	0.99
SMOTE	%33	Azınlık	0.05	0.11	0.07
		Çoğunluk	0.99	0.97	0.98
Random Under Sampler	%33	Azınlık	0.05	0.54	0.08
		Çoğunluk	0.99	0.83	0.90
SMOTE + Random Under Sampler	***%41***	***Azınlık***	***0.05***	***0.18***	***0.08***
		Çoğunluk	***0.99***	***0.95***	***0.97***
ENN	%1.5	Azınlık	0.12	0.08	0.10
		Çoğunluk	0.99	0.99	0.99
SMOTE + ENN	***%18***	***Azınlık***	***0.05***	***0.18***	***0.08***
		Çoğunluk	***0.99***	***0.95***	***0.97***
ENN + SMOTE + Random Under Sampler	%33	Azınlık	0.05	0.19	0.07
		Çoğunluk	0.99	0.94	0.96

Hukuklaşma veri setinde gerçekleştirilen ön işlem sürecinden sonra 817738 satır ve 155 sütun olan veri seti 156638 satır ve 81 sütuna düşmüştür. Makine öğrenimi algoritması seçiminde bağımsız değişken sayısının fazlalığı önemli olmuştur. Yüksek boyutlu verilerle sınıflandırma, makine öğreniminde önemli bir zorluktur çünkü yüksek boyutlu verilerdeki özelliklerin çokluğu, anlamlı modellerin tanımlanmasını zorlaştırır, bu da aşırı uyum ve sınıflandırma performansının düşmesine yol açar (Tran, 2023).

Hukuklaşmaya gidecek dosyaların tespiti, hukuki süreçlerin hızlı ve doğru bir şekilde başlatılabilmesi için çok önemlidir. Mağdurların haklarını arayabilmeleri için gerekli olan hukuki sürecin gecikmesine veya tamamen atlanmasına neden olan bu tür dosyaların kaçırılması mümkündür. Sonuç olarak, bu tür bir durumda duyarlılığın yüksek olması ve gerçek hukuklaşma vakalarının kaçırılmaması çok önemlidir. Duyarlılık, modelin azınlık sınıfını iyi tanımladığını gösterir. Hukuklaşmaya giden dosyaların az sayıda ve önemli olduğu bir veri setinde yüksek duyarlılık, modelin bu önemli vakaları kaçırmamasını sağlar.

Yapılan denemelerde düşük duyarlılığın artırılması amaçlanmıştır. Bunun yanında kesinlik skoru da modelin pozitif tahminlerinin ne kadarının doğru olduğunu gösterdiğinden mümkün olduğunca artırılmaya çalışılmış ve dolaylı olarak iki metriğin harmonik ortalaması olan f1 skor da kontrol edilmiştir.

Toplam 6 adet farklı yeniden örnekleme yöntemi ve hibrit kullanımı denenmiştir. Elde edilen sonuçlar karşılaştırılmıştır (Çizelge 4.4). Hiç bir çalışmada parametre ayarlaması yapılmamış, algoritma ve yöntemler en yalın haliyle uygulanıp deneme yapılmış ve sonuçlar buna göre değerlendirilmiştir.

Çizelge 4.4. Hukuklaşma veri seti yeniden örnekleme yöntemleri sonuçları

Yöntem	**Dengesizlik Oranı**	**Sınıf**	**Kesinlik**	**Duyarlılık**	**F1 Skor**
Yeniden Örnekleme Yöntemi Kullanılmadan	%3	Azınlık	1.00	0.04	0.07
		Çoğunluk	0.97	1.00	0.98
SMOTE	%33	Azınlık	0.33	0.19	0.24
		Çoğunluk	0.97	0.99	0.98
Random Under Sampler	%30	Azınlık	0.09	0.34	0.14
		Çoğunluk	0.98	0.91	0.94
SMOTE + Random Under Sampler	%42	Azınlık	0.25	0.22	0.24
		Çoğunluk	0.97	0.98	0.98
ENN	%4	Azınlık	0.31	0.23	0.27
		Çoğunluk	0.97	1.00	0.98
SMOTE + ENN	***%16***	***Azınlık***	***0.36***	***0.22***	***0.28***
		Çoğunluk	***0.97***	***0.99***	***0.98***
ENN + SMOTE + Random Under Sampler	%33	Azınlık	0.26	0.23	0.24
		Çoğunluk	0.97	0.99	0.98

İlk olarak rastgele orman makine öğrenimi algoritması direkt olarak uygulanmış ve 1 sınıfı için kesinlik 1.00 - duyarlılık 0.04 - f1 skor ore 0.07 elde edilmiştir. 0 sınıfı için ise kesinlik 0.98 - duyarlılık 1.00 - f1 skor 0.99 elde edilmiştir. Görüldüğü üzere model 1 sınıfını tahmin edememekte ve bu sınıfı öğrenememektedir.

Daha sonra oversampling yöntemlerinden olan SMOTE tek başına uygulanmış ve örnek artırım oranı olarak yüzde 50 kullanılmıştır. 1 sınıfı için kesinlik 0.33 - duyarlılık 0.19 - f1 skor 0.24, 0 sınıfı için kesinlik 0.97 - duyarlılık 0.99 - f1 skor 0.98 elde edilmiştir. 1 sınıfı için tüm başarım ölçütlerinde artış görülmüştür.

Undersampling yöntemlerinden RUS tek başına uygulanmış ve örnek azaltım oranı olarak yüzde 50 kullanılmıştır. 1 sınıfı için kesinlik 0.09 - duyarlılık 0.34 - f1 skor 0.14, 0 sınıfı için kesinlik 0.98 - duyarlılık 0.89 - f1 skor 0.93 elde edilmiştir. Burada bir önceki denemeye göre 1 sınıfı için duyarlılık değeri artsa da kesinlik ve harmonik ortalamaları olan f1 skor değeri düşmüştür. 0 sınıfı için ise duyarlılık ve f1 score değerlerdi düşmüştür.

Bu düşüşü dengelemek üzere hibrit bir yöntem kullanılmış, önce oversampling yöntemi olan SMOTE ve devamında undersampling yöntemi olan RUS kullanılmıştır. SMOTE için örnek artırım oranı yüzde 30, RUS için örnek azaltım oranı yüzde 70 seçilmiştir. 1 sınıfı için kesinlik 0.25 - duyarlılık 0.22 - f1 skor 0.24, 0 sınıfı için kesinlik 0.97 - duyarlılık 0.98 - f1 skor 0.98 elde edilmiştir. Sonuçlara bakıldığında SMOTE'un undersampling yönteminden kaynaklı çoğunluk sınıfını öğrenememe sorununa çözüm getirdiği görülmüştür. 0 sınıfı için değerler 0.97-1.00 arasına geri gelmiştir. Bunun yanında 1 sınıfı için düşük olan kesinlik ve f1 score değerleri artmış, duyarlılık değeri 0.22'ye gelmiştir.

SMOTE'un başarım ölçütlerine katkısının yanında daha etkili bir undersampling yöntemi kullanmak üzere ENN ile denemelere devam edilmiştir. ENN tek başına uygulanmıştır ve 1 sınıfı için kesinlik 0.31 - duyarlılık 0.23 - f1 skor 0.27, 0 sınıfı için kesinlik 0.97 - duyarlılık 0.98 - f1 skor 0.98 elde edilmiştir. ENN precision için güzel sonuçlar vermiş ve 1 sınıfı için öğrenememe sorunu oluşturmamıştır.

Hibrit yaklaşım olarak önce ENN sonrasında SMOTE olmak üzere beraber kullanılmıştır. SMOTE için örnek artırım oranı yüzde 20 alınmıştır. 1 sınıfı için kesinlik 0.36 - duyarlılık 0.22 - f1 skor 0.28, 0 sınıfı için kesinlik 0.97 - duyarlılık 0.99 - f1 skor 0.98 elde edilmiştir. Daha önce denenen hibrit yöntemden ve tek kullanılan ENN'den daha başarılı sonuçlar elde edilmiştir. 1 sınıfı için kesinlik değeri 0.36 alınan en yüksek değerdir. 0 sınıfı için öğrenememe sorunu oluşmamıştır.

Son bir hibrit yaklaşım denemesi olarak önce ENN, sonrasında yüzde 30 örnek artırım oranıyla SMOTE, devamında yüzde 50 örnek azaltım oranıyla RUS uygulanmıştır. 1 sınıfı için kesinlik 0.26 - duyarlılık 0.23 - f1 skor 0.24, 0 sınıfı için kesinlik 0.97 -

duyarlılık 0.98 - f1 skor 0.98 elde edilmiştir. Bir önceki hibrit modelin başarısını RUS, çoğunluk sınıfından rastgele örneklem azaltarak verinin yapısını bozarak, düşürmüştür.

5. TARTIŞMA VE SONUÇLAR

Bu tezde dengesiz öğrenme problemlerinde veri seviyesinde çözüm olan yeniden örnekleme yöntemleri üzerine çalışılmıştır. Veri seti olarak kullanılan hukuklaşma verisinde sınıflar arasında yüzde 3 oranında çok yüksek bir dengesizlik durumu bulunmaktadır. Literatürdeki çalışmalarda daha düşük dengesizlik durumları ile çalışılarak performans en yüksek seviyelere çıkarılmaya çalışılmıştır.

Tezdeki amacımız yüksek boyutlu verilerde çok yüksek dengesizlik durumunda dahi uygun yöntem ve hibrit yöntemlerle bu dengesizlik sorununun aşılabileceğini göstermektir. Bu nedenle yapılan denemelerde parametre ayarlaması yapılmamış, yalın bir halde yeniden örnekleme yöntemleri ile başarım artırılmaya çalışılmıştır.

Çalışmamıza yeniden örnekleme yöntemlerinden önce makine öğrenimi seçimi için yapılan çalışmayla başlanmıştır. Bu kapsamda dengesizlik oranı daha düşük %22 olan ve düşük boyutlu bir veri seti olan kredi kartı verileri ile makine öğrenimi sınıflandırma algoritmaları test edilmiştir. Yapılan denemelerde rastgele orman yönteminin karar ağaçları, SVM ve KNN makine öğrenimi algoritmalarından daha başarılı sonuçlar verdiği görülmüştür. F1 skor karşılaştırmasının yanında azınlık sınıfının kesinlik ve duyarlılık skorları da dikkate alınarak yapılan denemelerde rastgele orman yöntemi seçilmiştir.

Sonrasında elimizde olan 3 adet dengesiz veri seti üzerinde yeniden örnekleme yöntemleri denenmiştir.

İlk olarak dengesizlik oranı %12 olan düşük boyutlu covid-19 veri seti üzerinde yeniden örnekleme yöntemleri uygulanmıştır. Yeniden örnekleme yöntemlerinden önce veri setine veri ön işleme adımı olarak sadece aykırı değerlerin ve boş değerlerin veriden çıkarılması uygulaması yapılmıştır. En başarılı sonuçlar RUS yeniden örnekleyerek azaltma yönteminin olduğu hibrit yöntemler ve tek başına uygulandığı denemelerde alınmıştır. Covid-19 veri seti düşük boyutlu, az değişken içeren ve örnek sayısı fazla olan bir veri setidir. Bu sebeple RUS yöntemi çoğunluk sınıfından rastgele örnek çıkarırken tekrarlayan örnekleri çıkartmış ve modelin başarısına katkı sağlamıştır. Ancak bu veri seti üzerinde hibrit yöntemlerin başarılı sonuçlar verdiği de görülmüştür.

İkinci deneme dengesizlik oranın %1.5 olduğu ve fazla değişkene sahip yüksek boyutlu sigorta veri seti ile yapılmıştır. Bu veri setinde de yeniden örnekleme yöntemlerinden önce veri setine veri ön işleme adımı olarak sadece aykırı değerlerin ve boş değerlerin veriden çıkarılması uygulaması yapılmıştır. Bu çalışmada hukuklaşma veri setine en

yakın veri seti olan sigorta verisi üzerinde yapılan denemelerde iki yöntemin kullanıldığı hibrit yöntemler en başarılı sonuçları vermiştir. SMOTE ve ENN ile SMOTE ve RUS yöntemleri en iyi f1 skor ve azınlık sınıfı için en iyi kesinlik, duyarlılık skorlarını üretmiştir.

En son olarak bu 3 veri seti üzerindeki çalışmalardan hareketle anonimleştirilmiş hukulaştırma verisi üzerinde denemeler yapılmıştır. Hukuklaşma veri seti üzerinde yapılan ön işlem ve değişken seçimi sonucunda 81 bağımsız değişken ile %3 dengesizlik oranına sahip bir veri seti elde edilmiştir. Denemelerde parametre ayarlaması yapılmamıştır. Hem bu durum nedeniyle hem de dengesiz bir veri seti olması nedeniyle çalışma amacımızın da yeniden örnekleme yöntemlerini incelemek ve etkilerini gözlemlemek olduğundan kullandığımız makine öğrenimi sınıflandırma algoritmasını, kredi kartı veri seti üzerinde yaptığımız denemelerden hareketle ensemble metod olan ve bagging ile çalışarak aşırı uyumluluğa dayanıklı olan rastgele orman olarak seçilmiştir.

Yeniden örnekleme yöntemlerinden ise diğer oversampling yöntemlerine temel olan SMOTE ile en yalın undersampling yöntemi olan RUS ve diğer undersampling yöntemlere temel olan ENN kullanılmıştır. Sigorta ve Covid-19 veri setlerinde yapılan denemeler bu kararı desteklemiştir.

Hukuklaşma dosyalarını kaçırmamak adına duyarlılık metriği önceliklendirilmiştir. Ancak kesinlik skoru da yanlış pozitiflerin (hukuklaşmaya gitmeyen ancak hukuklaşmaya giden olarak sınıflandırılan dosyaların) sayısını azaltmaktadır. Duyarlılık skorunun daha önemli olması yanında kesinlik skoru da düşük kalmamalıdır. Bu nedenle duyarlılık başta olmak üzere kesinlik skoruna da odaklanılmıştır.

Random forest sabit tutularak yapılan çalışmalarda görülmüştür ki; hibrit bir yöntemin kullanılması başarıyı arttırmıştır. Tek başına oversampling ya da undersampling yöntemi kullanmanın üstünde bir başarı vermiştir.

En iyi sonuç SMOTE ve ENN yeniden örnekleme yöntemlerinin beraber kullanıldığı denemede alınmıştır. 1 sınıfı için kesinlik 0.36 - duyarlılık 0.22 - f1 skor 0.28, 0 sınıfı için kesinlik 0.97 - duyarlılık 0.99 - f1 skor 0.98 elde edilmiştir.

Bu başarının devamında hibrit modellere üçüncü bir yöntem eklenmiş ancak bunun datayı manipüle ederek başarısını düşürdüğü görülmüştür. Çünkü sentetik olarak artırılan ya da azaltılan bir sınıfa tekrar aynı işlemi uygulamak verinin gerçekliğini bozmaktadır.

Bulgular kısmındaki çizelgelerde her veri setinin uygulanan yeniden örnekleme yöntemi sonrasındaki dengesizlik oranı gösterilmiştir. Dengesizlik oranının azalmasının modelin başarısına etkisi olmamıştır. Hukuklaşma veri setinde de görüldüğü üzere en iyi başarıyı

veren modelde dengesizlik oranı %16'dır. Diğer veri setlerinde de dengesizlik oranının etkili olmadığı gözükmektedir.

Hanafy ve Ming (2021)'in sigorta verileriyle yaptıkları benzer çalışmada kullanılan veri 16 bağımsız değişkenden oluşmakta ve dengesizlik oranı % 8'dir. Elde ettikleri başarı modellerin parametre ayarlamasıyla %90'ın üzerine çıkmıştır (Hanafy ve Ming, 2021).

Ancak bu tez çalışmasında kullanılan veri setindeki bağımsız değişken sayısı çok daha yüksektir ve sadece yeniden örnekleme yöntemlerinin model başarısına etkisi incelendiği için model eğitiminde parametre ayarlaması da yapılmamıştır.

Sonuç olarak, bu çalışmada literatüdeki çalışmalara oranla yüksek dengesizlik durumlarında ve yüksek boyutlu veri setlerinde yeniden örnekleme yöntemlerinin daha başarılı olduğu gösterilmiştir. Buna ek olarak örnek azaltma ve arttırma yöntemlerinin birlikte kullanılarak hibrit yeniden örnekleme ile başarım daha da arttırılmıştır. İlerideki çalışmalarda hibrit yeniden örnekleme yöntemleri ile oluştulan veri seti ile geliştirilecek modeller üzerinden hiper parametre optimizasyonu da gerçekleştirilerek tahmin başarısı daha da arttırılabilir.

KAYNAKLAR

Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas. "Handling imbalanced datasets: A review." GESTS international transactions on computer science and engineering 30.1 (2006): 25-36.

Kennedy, R.K.L., Salekshahrezaee, Z., Villanustre, F. et al. Iterative cleaning and learning of big highly-imbalanced fraud data using unsupervised learning. J Big Data 10, 106 (2023).

H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009.

J. Burez, D. Van den Poel,Handling class imbalance in customer churn prediction,Expert Systems with Applications,Volume 36, Issue 3, Part 1,2009,Pages 4626-4636,ISSN 0957-4174.

Zou, Quan & Xie, Sifa & Lin, Ziyu & Wu, Meihong & Ju, Ying. (2016). Finding the Best Classification Threshold in Imbalanced Classification. Big Data Research. 5. 10.1016/j.bdr.2015.12.001.

Trafimow, David & Earp, Brian. (2017). Null Hypothesis Significance Testing and Type I Error: The Domain Problem. New Ideas in Psychology. 45. 19-27. 10.1016/j.newideapsych.2017.01.002.

Vujovic, Zeljko. (2021). Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications. Volume 12. 599-606. 10.14569/IJACSA.2021.0120670.

Javidi, M., Abbaasi, S., Naybandi Atashi, S. et al. COVID-19 early detection for imbalanced or low number of data using a regularized cost-sensitive CapsNet. Sci Rep 11, 18478 (2021).

Estabrooks, Andrew & Jo, Duke Taeho & Japkowicz, Nathalie. (2004). A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence. 20. 18-36.

Adane Nega Tarekegn, Mario Giacobini, Krzysztof Michalak, A review of methods for imbalanced multi-label classification, Pattern Recognition, Volume 118, 2021, 107965, ISSN 0031-3203.

Powers, David & Ailab,. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2. 2229-3981. 10.9735/2229-3981.

A. J. Mary and S. P. A. Claret, "Imbalanced Classification Problems: Systematic Study and Challenges in Healthcare Insurance Fraud Detection," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1049-1055, doi: 10.1109/ICOEI51242.2021.9452828.

Ali, Haseeb & Salleh, Mohd & Hussain, Kashif & Ullah, Ayaz & Ahmad, Arshad & Naseem, Rashid. (2019). A review on data preprocessing methods for class imbalance problem. 390-397. 10.14419/ijet.v8i3.29508.

Rish, Irina. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell. 3.

Zahi M. Omer, Hussain Shareef, Comparison of decision tree based ensemble methods for prediction of photovoltaic maximum current, Energy Conversion and Management: X, Volume 16, 2022, 100333, ISSN 2590-1745.

Salim Rezvani, Xizhao Wang, A broad review on class imbalance learning techniques, Applied Soft Computing, Volume 143, 2023, 110415, ISSN 1568-4946.

Esposito, Carmen & Landrum, Gregory & Schneider, Nadine & Stiefl, Nikolaus & Riniker, Sereina. (2021). GHOST: Adjusting the Decision Threshold to Handle

Imbalanced Data in Machine Learning. Journal of Chemical Information and Modeling. 10.1021/acs.jcim.1c00160.

King, G.; Zeng, L. Logistic Regression in Rare Events Data. Polit. Anal. 2001, 9, 137– 163, DOI: 10.1093/oxfordjournals.pan.a004868 There is no corresponding record for this reference.

Seliya, N., Abdollah Zadeh, A. & Khoshgoftaar, T.M. A literature review on one-class classification and its potential applications in big data. J Big Data 8, 122 (2021).

Domingos, Pedro. (2002). MetaCost: A General Method for Making Classifiers Cost-Sensitive. Proceedings of the Fifth ACM SIGKDD Int'l. Conf. on Knowledge Discovery & Data Mining. 10.1145/312129.312220.

Elkan, Charles. (2001). The Foundations of Cost-Sensitive Learning. Proceedings of the Seventeenth International Conference on Artificial Intelligence: 4-10 August 2001; Seattle. 1.

Hasanin, T., Khoshgoftaar, T.M., Leevy, J.L. et al. Severely imbalanced Big Data challenges: investigating data sampling approaches. J Big Data 6, 107 (2019).

Wang, Meng & Yao, Xinghua & Chen, Yixiang. (2021). An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3057693.

Hanafy, Mohamed & Ming, Ruixing. (2021). Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches. International Journal of Advanced Computer Science and Applications. 12. 493-499. 10.14569/IJACSA.2021.0120656.

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Friedman, Jerome. (2000). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 29. 10.1214/aos/1013203451.

D. R. Cox, The Regression Analysis of Binary Sequences, Journal of the Royal Statistical Society: Series B (Methodological), Volume 20, Issue 2, July 1958, Pages 215–232.

Evelyn Fix and J. L. Hodges, Jr. International Statistical Review / Revue Internationale de Statistique Vol. 57, No. 3 (Dec., 1989), pp. 238-247 (10 pages)

Pagan, Muasir & Zarlis, Muhammad & Candra, Ade. (2023). Investigating the impact of data scaling on the k-nearest neighbor algorithm. Computer Science and Information Technologies. 4. 135-142. 10.11591/csit.v4i2.p135-142.

McCulloch, W.S., Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133 (1943).

Goutte, Cyril & Gaussier, Eric. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Lecture Notes in Computer Science. 3408. 345-359. 10.1007/978-3-540-31865-1_25.

Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery, 28:92–122, 2014.

Kimura, Takuma. (2022). Customer Churn Prediction with Hybrid Resampling and Ensemble Learning.. 1-23.

R. Mohammed, J. Rawashdeh and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 243-248.

Hayati, Mardhiya & Mutmainah, Siti & Ghufran, Syed. (2021). Random and Synthetic Over-Sampling Approach to Resolve Data Imbalance in Classification.

International Journal of Artificial Intelligence Research. 4. 86. 10.29099/ijair.v4i2.152.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

Mukherjee, Mimi & Khushi, Matloob. (2021). SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing, 878–887. Springer, 2005.

H. M. Nguyen, E. W. Cooper, K. Kamei, "Borderline over-sampling for imbalanced data classification," International Journal of Knowledge Engineering and Soft Data Paradigms, 3(1), pp.4-21, 2009.

Felix Last, Georgios Douzas, and Fernando Bacao. Oversampling for imbalanced learning based on k-means and smote.

He, Haibo, Yang Bai, Edwardo A. Garcia, and Shutao Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," In IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328, 2008

Yen, Show-Jane & Lee, Yue-Shi. (2006). Cluster-based Under-sampling Approaches for Imbalanced Data Distributions. Expert Systems with Applications. 36. 5718-5727. 10.1016/j.eswa.2008.06.108.

P. Hart, "The condensed nearest neighbor rule," In Information Theory, IEEE Transactions on, vol. 14(3), pp. 515-516, 1968

I. Mani, I. Zhang. “kNN approach to unbalanced data distributions: a case study involving information extraction,” In Proceedings of workshop on learning from imbalanced datasets, 2003

Ivan Tomek. Two modifications of cnn. IEEE Trans. Systems, Man and Cybernetics, 6:769–772, 1976.

Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, pages 408–421, 1972.

Ivan Tomek. An experiment with the edited nearest-neighbor rule. IEEE Transactions on systems, Man, and Cybernetics, 6(6):448–452, 1976.

Peter Hart. The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory, 14(3):515–516, 1968.

Jorma Laurikkala. Improving identification of difficult small classes by balancing class distribution. In Conference on Artificial Intelligence in Medicine in Europe, 63–66. Springer, 2001.

M. Kubat, S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection,” In ICML, vol. 97, pp. 179-186, 1997.

D. Smith, Michael R., Tony Martinez, and Christophe Giraud-Carrier. “An instance level analysis of data complexity.” Machine learning 95.2 (2014): 225-256.

Gustavo EAPA Batista, Ana LC Bazzan, and Maria Carolina Monard. Balancing training data for automated annotation of keywords: a case study. In WOB, 10–18. 2003.

Seiffert, Chris & Khoshgoftaar, Taghi & Van Hulse, Jason. (2009). Hybrid sampling for imbalanced data. Integrated Computer-Aided Engineering. 16. 193-210. 10.3233/ICA-2009-0314.

Tran, Cao Truong. (2023). Ensemble Learning Approaches for Classification with High-Dimensional Data. Journal of Science and Technique. 12. 10.56651/lqdtu.jst.v12.n1.659.ict.

Kohavi, Ron. (2001). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 14.

Güneş, V., Kırca, S., Yagcı, H. E., Narın, N. G. (2023). Variable Selection with Machine Learning in the Legalization Process for Traffic Insurance. The Eurasia Proceedings of Science Technology Engineering and Mathematics, 22, 237-246.

ÖZGEÇMİŞ

Kişisel Bilgiler
Ad Soyad :H**********
Uyruk : *******
Doğum Yeri ve Tarihi: **/**/***
Medeni Hali :*****
Telefon : -
E-posta : *******@**********

Eğitim

Alınan Derece	**Aldığı Kurum/Üniversite**	**Mezuniyet Yılı**
Lise	Deniz Lisesi	2008
Lisans	Deniz Harp Okulu	2012
Yüksek Lisans	Muğla Sıtkı Koçman Üniversitesi	-

İş Tecrübesi

Yıl	**Yer**	**Pozisyon**
2012	Muğla	Deniz Subayı
2019	Singapur (Uzaktan)	Veri Bilimci
2021	İstanbul	Veri Bilimci
2022	İstanbul	Veri Analitiği Müdürü
2024	İstanbul	Veri Bilimci

Yabancı Dil

İngilizce	**Seviye**
Yazma	C2
Konuşma	C2
Anlama	C2
Okuma	C2

Bilimsel Faaliyetler

1. H. E. Yağcı, A. Atçılı, ve S. Sezer, “A Review on Deep Learning Models for Satellite Imagery”, Adv. Artif. Intell. Res., c. 1, sy. 2, ss. 73–79, 2021.
2. Güneş, V., Kırca, S., Yagcı, H. E., Narın, N. G. (2023). Variable Selection with Machine Learning in the Legalization Process for Traffic Insurance. The Eurasia Proceedings of Science Technology Engineering and Mathematics, 22, 237-246.