+.

T.C.

SELÇUK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

AĞ SALDIRI VERİ KÜMELERİNİN SINIFLANDIRILMASINDA DENGELEME İŞLEMİNİN ETKİSİ

Samara Khamees JWAIR JWAIR

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalını

Ekim -2019
KONYA

TEZ KABUL VE ONAYI

Samara Khamees JWAIR JWAIR tarafından hazırlanan "Ağ Saldırı Veri Kümelerinin Sınıflandırılmasında Dengeleme İşleminin Etkisi" adlı tez çalışması 22/07/2019 tarihinde aşağıdaki jüri tarafından oy birliği / ~~oy çokluğu~~ ile Selçuk Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Doç.Dr. Mesut GÜNDÜZ	.....
Danışman Dr.Öğr.Üyesi. Ersin KAYA	.....
Üye Dr.Öğr.Üyesi. Ayşe Merve ACILAR	.....

Yukarıdaki sonucu onaylarım.

Prof. Dr. ....
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

İmza

Samara Khamees Jwair JWAIR

Tarih: 24.10.2019

ÖZET

YÜKSEK LİSANS TEZİ

AĞ SALDIRI VERİ KÜMELERİNİN SINIFLANDIRILMASINDA DENGELEME İŞLEMİNİN ETKİSİ

Samara Khamees Jwair JWAIR

Selçuk Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr.Öğr.Üyesi. Ersin KAYA

2019, 48 Sayfa

Jüri
Dr.Öğr.Üyesi. Ersin KAYA
Doç.Dr. Mesut GÜNDÜZ
Dr.Öğr.Üyesi. Ayşe Merve ACILAR

Sınıflandırma, makine öğrenmesi ve veri madenciliği topluluklarında en önemli görevlerden biridir. Sınıflandırma işleminde sık karşılaşılan sık problemlerden biri veri setindeki sınıf dengesizliği problemidir. Dengesiz veri seti öncelikle iki veya daha fazla sınıfı içeren denetimli makine öğrenmesi bağlamıyla ilgilidir. Çoğu makine öğrenme tekniği için, küçük dengesizlikler problem değildir. İki sınıf varsa, o zaman dengeli veri her sınıf için %50 örnek anlamına gelir. Fakat bir sınıf için %60, diğer sınıf için %40 örnek varsa, herhangi bir önemli performans bozulmasına neden olmamaktadır. Veri setlerinde sınıf dengesizliği yüksek olduğunda sınıflandırma başarısı olumsuz olarak etkilenmektedir. Bu problemi ortadan kaldırmak için ve verilerin dengelenmesini sağlamak için örneklendirme yöntemlerinden biri kullanmaktadır. Örnekleme yöntemi, azınlık ve çoğunluk sınıfı boyutunu değiştirerek eğitim kümesindeki dengesizlik sınıfını ele alan bir yöntemdir. Sınıfları dengelemeye yönelik basit bir veri düzeyi yaklaşımı, bir sınıfı çoğaltma örneklenmesi ya da hemen hemen aynı olan çoğunluk sınıflarının örneklenmesi için orijinal veri kümesinden tekrarlamalı örnekler içerir. Bu stratejilerin her ikisi de herhangi bir öğrenme sisteminde uygulanabilir. Genel olarak, saldırı tespit ve benzeri veri kümelerinde sınıf dengesizliği bulunmaktadır. Bu tez çalışmasında, dengesiz veri kümeleri ele alınarak sentetik azaltma örnekleme tekniği (SMOTE) yöntemi ve diferansiyel evrim algoritması (DE) stratejileri ile bu veri kümelerini dengeli hale getirilip ve sınıflandırma başarıları arttırılmıştır. K-En Yakın Komşuları (K-NN), Destek Vektör Makinesini (SVM) ve C4.5 dengeli veri kümelerini sınıflandırmak için uygulanmıştır. Sonuç olarak, kullanılan dengesiz veri kümeleri dengeli hale geldikten sonra bu veri kümelerinin sınıflandırma başarılarının artması sağlanmıştır.

Anahtar Kelimeler: Diferansiyel Evrim Algoritması, Örnekleme teknikleri, Saldırı Tespit, SMOTE, Sınıflandırma.

ABSTRACT

MS THESIS

THE EFFECT OF BALANCING PROCESS ON CLASSIFYING INTRUSION DETECTION DATASET

Samara Khamees Jwair JWAIR

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF SELÇUK UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE
IN COMPUTER ENGINEERING

Advisor: Asst. Prof. Dr. Ersin KAYA

2019, 48 Pages

Jury
Asst. Prof. Dr. Ersin KAYA
Assoc. Prof. Dr. Mesut GÜNDÜZ
Asst. Prof. Dr. Ayşe Merve ACILAR

Classification is one of the most important tasks in machine learning and data mining communities. One of the common problems encountered in the classification process is the class imbalance problem in the data set. The unbalanced data set is primarily relevant in the context of supervised machine learning involving two or more classes. For most machine learning techniques, small imbalances are not a problem. If there are two classes, then the balanced data means 50% sample for each class. However, if there is a 60% sample for one class and 40% for the other class, it does not cause any significant performance degradation. When class imbalanced is high in datasets, classification success is negatively affected. It uses one of the sampling methods to eliminate this problem and to stabilize the data. The sampling method is a method of addressing the imbalance class in the training set by changing the size of the minority and majority classes. A simple data-level approach to balancing classes includes iterative examples from the original data set for over-sampling of a class or for sampling almost identical majority classes. Both of these strategies can be implemented in any learning system. In general, there are unbalanced class in intrusion detection and similar data sets. In this thesis, unbalanced datasets are handled and synthetic minority sampling (SMOTE) method and differential evolution algorithm (DE) strategies are used to balance these datasets and increase classification accuracy. K-Nearest Neighbors (K-NN), Support Vector Machine (SVM) and C4.5 are applied to classify balanced data sets. As a result, the classification accuracy of the unbalanced data sets increased after the unbalanced data sets became balanced.

Keywords: Classification, Differential Evolution Algorithm, Intrusion Detection Systems, Sampling Technique, SMOTE.

ÖNSÖZ

İlk önce, hem kötü hem de iyi zamanlarımda beni destekledikleri için aileme teşekkür ederim. Ayrıca, danışmanım Dr.Öğr.Üyesi. ERSIN KAYAıya bilgisi, desteği ve sabrından dolayı teşekkür ederim. Son olarak, arkadaşlarıma tez çalışmam süresince bana moral, destek oldukları için mutlu, huzurlu ve yapıcı bir ortamı sağladıkları için teşekkür ederim.

Samara Khamees JWAİR JWAIR
KONYA-2019

İÇİNDEKİLER

Şekil listesi

Çizelge listesi

SİMGELER VE KISALTMALAR

Kısaltmalar

ADASYN	:Dengesiz Veri Kümelerinden Öğrenme İçin Adaptıf Sentetik Örnekleme Yaklaşımı (Adaptive Synthetic Sampling Approach For Imbalanced Learning)
AUC	:Eğri Altındaki Alan (Area under curve)
CMC	:Kontraseptif Yöntem Seçimi (Contraceptive Method Choice)
DE	:Diferansiyel Evrim Algoritmasını (Differential Evolutioun)
Dif	:İki Örnek arasındaki Fark (Difference between two samples)
EA	:Evrimsel Algoritma (Evolutionary Algorithm)
EIT	:Elektriksel empedans tomografi (Electrical Impedance Tomography)
FRST	:Bulanık Kaba Küme Teorisi (Fuzzy Rough Set Theory)
GO	:Geometrik ortalaması (Geometric mean)
GA	:Genetik Algoritmalar (Genetic Algorithms)
HSS	:Hammersley Dizi Örneklemesi (Hammersley Sequence Sampling)
ID3	:Geliştirilmiş Karar Ağacı Algoritması (Improved Decision Tree Algorithm)
IDS	:Saldırı Tespit Sistemleri (Intrusion Detection System)
IPF	:İteratif-Bölme Filtresi (Iterative-Partitioning Filter)
IR	:Dengesizlik Oranı (Imbalanced Rate)
KNN	:En Yakın Komşu (K-nearest neighbour)
MCMC	:Markov Zinciri Monte Carlo (Markov chain Monte Carlo)
ML	:Makine öğreniminde (Machine learning)
MLPUS	:MLP tabanlı azaltma örnekleme tekniği (MLP-based undersampling technique)
MWMOTE	:Çoğunluk ağırlıklı azaltma örnekleme tekniği (Majority weighted minority oversampling technique)
NBOTE	:Komşu Tabanlı Çoğaltma Örnekleme Tekniği (Neighbor Based Oversampling TEchnique)
NCS-R	:Ulusal Komorbidite Anketi Replikasyonu (National Comorbidity Survey Replication)
NLAAS	:Ulusal Latino ve Asya Amerikan Ruh Sağlığı Çalışması (National Latino and Asian American Study of Mental Health)
NRSB	:Komşu Kaba Küme Sınırı (Neighborhood Rough Set Boundary)
NSAL	:Ulusal Amerikan Yaşamı Çalışması (National Study of American Life)
OFS	:Ortogonal İleri Seçim (Orthogonal Forward Selection)
OVA	:Biri-vs-Hepsi (One-vs-All)
OVO	:Biri-vs-Biri (One-vs-One)
PSO	:Parçacık Sürü Optimizasyonu (Particle Swarm Optimization)
PSSP	:Protein İkincil Yapı Tahmini (Protein Secondary Structure Prediction)
Q2T	:Dizi Yapısı Tahmini (Sequence Structure Prediction)
RAMOBoost	:Rastgele Orman (Random Forest)
ROC	:Alıcı Çalışma Eğrisi (Receiver Operating Curve)
RPF	:Radyal Temel Fonksiyon sınıflandırıcısı (Radial Basis Function

	classifier)
RST	:Kaba Küme Teorisi (Rough Set Theory)
SA	:Simüle edilmiş Tavlama (Simulated Annealing)
SIR	:Örnekleme Önemi Yeniden Örnekleme (Sampling Importance Resampling)
SM	:Hassasiyet Ölçüsü (Sensitivity Measure)
SMOTE	:Sentetik Azaltma Çoğaltma Örnekleme Tekniği (Synthetic Minority Oversampling Technique)
SVM	:Destek Vektör Makinesi (Support Vector Machine)
T2T	:Yapı Yapısı (Structure-Structure)
VC	:Vertebral sütun (Vertebral column)

1. GİRİŞ

Gerçek dünyada dengesiz veri kümeleri, veri analizinde yaygın olarak karşılaşılan bir problemdir. Sınıflandırma kategorileri yaklaşık olarak eşit değilse, veri kümesinde dikkate alınan veriler dengesizdir. Son yıllarda, veri dengeleme tekniklerine olan ilgi artmıştır. Çünkü dengeleme teknikleri genellikle gerçek dünya problemlerinden elde edilen dengesiz verilere uygulanabilmektedir. Ayrıca, test verilerinin dağılımı eğitim verilerinden farklı olabilir ve sınıflandırma hatası öğrenme sırasında bilinmeyebilir (Chawla, 2009). Dengesiz veri setleri genellikle tıbbi örüntü tanıma ve veri madenciliği bağlamında birçok pratik uygulamada ortaya çıkabilmektedir. Temel eğitim setinin eşit olarak dağılımı sınıflandırma performansını olumlu olarak etkilemektedir. Bununla birlikte, eğitim setinde dengesiz bir dağılım olduğunda sınıflandırma performansını olumsuz olarak etkilemektedir ve ciddi bir eğilim problemiyle ortaya çıkabilmektedir (Wang ve ark., 2006).

Bire bir telekomünikasyon müşterilerini bulmak, kelime grupları öğrenmek, metin sınıflandırması yapmak, dolandırıcılığı tespit etmek, bilgi taşımak ve filtreleme görevleri dâhil olmak üzere pek çok alanda dengesiz veri setleri mevcuttur (Kotsiantis ve ark., 2006). Dengesiz veri kümeleri sorunu, veri kümesinin yeniden örneklemesi yapılırken dengeli olmasını sağlayarak çözülebilir. Örnekleme yöntemleri bu sorunun üstesinden gelebilmektedir. Örnekleme, azınlık ve çoğunluk sınıfı boyutunu değiştirerek eğitim kümesindeki dengesizlik sınıfını ele alan bir yöntemdir. Basit bir veri seviyesi yaklaşımı dengeleme sınıfları ana veri setinde, azınlık sınıflarına örnekler ekleyerek veya çoğunluk sınıflarını yeniden örnekleyerek sınıfları dengelemeye çalışmaktadır. Bu stratejilerin her ikisi de herhangi bir öğrenme sisteminde uygulanabilir (Ganganwar, 2012). Veri çoğaltma ve veri azaltma yöntemleri, eğitim verilerinin sınıf dağılımını değiştirmek için kullanılabilir ve her iki yöntem de sınıf dengesizliği ile başa çıkmak için kullanılabilir. Veri çoğaltma ve veri azaltma yöntemlerinin, avantajları ve dezavantajları vardır. Veri azaltma yöntemlerini dezavantajı, potansiyel olarak faydalı verileri atmasıdır. Veri çoğaltmanın esas dezavantajı, mevcut örneklerin tam kopyalarını oluşturmasıdır. Veri çoğaltma ikinci bir dezavantajı, eğitim örneklerinin sayısını arttırması ve böylece öğrenme süresini arttırmasıdır. Bu dezavantajların yanı sıra, örneklemenin kullanılmasının en birinci nedeni, tüm öğrenme algoritmalarının maliyete duyarlı uygulamaları olmamasıdır ve bu nedenle örneklemeyi kullanan kapsayıcı tabanlı bir yaklaşım tek seçenektir.

Örneklemeyi kullanmanın ikinci nedeni, veri kümelerinin büyük olması ve öğrenmenin mümkün olabilmesi için eğitim kümesinin boyutunun azaltılması gerektiğidir. Son sebep ise maliyete duyarlı bir öğrenme algoritması yerine örneklemenin kullanılmasına katkıda bulunan, yanlış sınıflandırma maliyetlerinin çoğu zaman bilinmemesidir (Weiss ve ark., 2007). Bazı yayınlarda dengesiz veri kümeleri için azınlık örneklendirme tekniği yerine "sentetik" örnekler alarak SMOTE çoğaltma örneklendirme tekniği kullanılmıştır. SMOTE çoğunluk sınıf değil de azaltma sınıfına çoğaltma örnekleme yapan önemli bir yaklaşımdır (Wang ve ark., 2006). SMOTE tekniği, rastgele örneklemeyi ilerletmek için önerilen, ancak yüksek boyutlu veriler üzerindeki davranışı ayrıntılı bir şekilde araştırılmayan yaygın bir örnekleme yöntemidir. SMOTE'nin çoğu sınıflandırıcı için çoğunluk sınıfındaki sınıflandırmaya yönelik eğilim azalttığı ve rastgele örneklemeden daha az etkili olduğunu gözlemlemişlerdir. SMOTE, değişken sayısı ve bazı değişken seçim türleri azalırsa Öklid mesafesine dayanan KNN sınıflandırıcıları için yararlı olabilir (Blagus ve Lusa, 2012). Bu tez çalışmasında, verileri dengeli haline getirmek için veri kümelerine SMOTE tekniği ve Diferansiyel Evrim (DE) stratejileri uygulanmıştır.

Evrimsel bir hesaplama tekniği olarak, Diferansiyel Evrim, Storn ve Price'ın 1995 yılında algoritmayı tanıttığından beri karmaşık optimizasyon problemlerini çözmesinde çok dikkat çekmiş ve geniş uygulamalar kazanmış bir algoritmadır. Evrimsel bir algoritmanın (EA) yapısını andırır, ancak yeni aday çözümlerin üretilmesinde açgözlü bir seçim şeması kullanılmasıyla geleneksel EA'lardan farklıdır. Diferansiyel Evrim, gerçek dünyadaki birçok küresel optimizasyon problemini çözmede etkilidir. Bununla birlikte, etkinliği kritik olarak uygun popülasyon büyüklüğü ve strateji parametrelerinin ayarlanmasına bağlıdır. Bu nedenle, uygun parametre değerini elde etmek için, zaman alıcı ön parametre ayarlarının yapılması gerekir (Tang ve ark., 2008).

Geleneksel olarak DE algoritmasına ait üç operatör vardır, bu operatörler seçim, mutasyon ve çaprazlama operatörleridir. Bu operatörlerde kullanılan değişkenler vardır, bunlar popülasyon büyüklüğü (NP), ölçek faktörü (F) ve çaprazlama olasılığıdır (CR). Mutasyon, farklı evrim algoritmalarının performansında kilit bir rol oynar ve mutasyonun çeşitli değişkenleri vardır. Mutasyon değişkenlerinin ve parametrelerin seçimi, diferansiyel evrim algoritması araştırmasının en önemli konusudur. Çeşitli mutasyon varyantları ve parametreleri, karşılık gelen birkaç farklı evrim stratejisini

oluşturur. Deneysel parametre çalışmaları ve diferansiyel evrim algoritmasının itibari parametre ayarları yapılmıştır (Ao ve Chi, 2009).

Bu tez çalışmasında, saldırı tespiti sisteminin performansını artırmak için SMOTE tekniği ve DE stratejileri saldırı tespit veri kümesine uygulanmış olup aldığımız sonuçlar arasında bir karşılaştırma yapılmıştır. İnternet, günlük hayatımızın vazgeçilmez bir parçası olup ve birçok önemli işimizi web uygulamaları üzerinden yapmaktayız. Web uygulamalarının sayıları artmakta, bunula birlikte kullanıcı verilerinin risk güvenliği de artmaktadır. Ağ güvenliği için genelde saldırı tespit sistemleri ağlara yapılan saldırıların tespitinde başarılı bir şekilde kullanılmaktadır, bu başarı saldırı tespit sistemindeki kullanılan sınıflandırma algoritmasının öğrenme yeteneğine bağlıdır. Sınıflandırma algoritmalarının öğrenme yeteneğini doğrudan etkileyen veri setlerinin düzenli olmasıdır. Genellikle saldırı tespiti veri kümeleri sınıf dengesizliği problemlerine sahiptir, bu problem saldırı tespit veri kümesinin sınıflandırma başarısını etkilemektedir. Bunun da temel sebebi öğrenme algoritmalarının çok fazla sınıfa sahip olan örnekleri iyi bir şekilde öğrenmesi ve az sınıfa sahip olan örnekleri iyi bir şekilde öğrenememesidir. Daha iyi bir saldırı tespiti sistemi elde etmek için ilk başta, saldırı tespit sisteminin saldırıları ve izinsiz girişleri iyi bir şekilde öğrenmesi gerekmektedir. İzinsiz giriş, güvenlik tekniğinden kaçınarak bilgisayar sistemine yapılan bir tür saldırıdır. Saldırı tespiti, güvenlik sorunlarını ve belirtilerini tespit etmek için bir bilgisayar sisteminde meydana gelen işlemleri kontrol etme ve analiz etme işlemidir. Kullanım şekline göre saldırı tespit sisteminin iki ana stratejisi vardır: kötüye kullanım ve anormallik tespiti. Kötüye kullanma, normal davranışlara benzeyen saldırıların tespiti, anormallik tespiti ise normal davranışa uymayan davranışların belirlenmesidir. Bu mekanizma trafik anormalliklerinin tespitine dayanmaktadır. Anormallik tespit sistemleri doğada uyarlanabilir niteliktedir, yeni saldırı ile başa çıkabilirler. Ancak belirli bir saldırı tipini tanımlayamazlar (Mukherjee ve Sharma, 2012). Metin madenciliği kullanan web uygulamalarında saldırı tespiti sistemi çalışmasında, web uygulamalarına saldırı tespiti veya kötüye kullanımın tespit edilmesine odaklanmışlardır. Ayrıca, metin sınıflandırma kullanarak metin madenciliği mekanizmasını temel alan bir IDS bileşeni sunulmaktadır. Web uygulama sunucusu tarafından oluşturulan normal ve sıra dışı kullanıcı davranışının özelliklerini öğrenebilir, böylece açık kod programlamak veya yazmak zorunda kalmadan kötüye kullanımı tespit edebilir ve sonuç olarak sistem bakımını iyileştirebilir. Telekomünikasyon sistemleri, gizli bilgi ve içerdiği sorumluluklar nedeniyle kritik öneme sahiptir. Bu

nedenle, Telekomünikasyon sistemlerindeki önemli bilgileri korumak amacıyla saldırı tespit sistemleri kullanılmaktadır.

1.1. Tezin Amacı

Veri setlerinin hazırlanması veri madenciliği sınıflandırma algoritmalarının performansını doğrudan etkilemektedir, veri seti ne kadar iyi bir şekilde hazırlandıysa o kadar sınıflandırma algoritmasının başarısı artar. Veri setlerinin hazırlanmasında normalizasyon, veri temizleme, özellik seçme ve örneklendirme gibi veri önişleme teknikleri kullanılmaktadır. Günümüzde hemen hemen birçok önemli işlemleri internet üzerinden yapmaktayız, dolayısıyla web tabanlı uygulamalar gittikçe artmaktadır. Bununla birlikte kişilerin bulutta bulunan önemli bilgilerinin riski de artmaktadır, bu riski önlemek ve ağ güvenliği için saldırı tespiti sistemleri tasarlanmaktadır. Tasarlanan saldırı tespit sisteminin performansı sistemde kullanılan sınıflandırma modelinin başarısına bağlıdır. Sınıflandırma algoritmalarının öğrenme yeteneğini doğrudan etkileyen veri setlerinin düzenli olmasıdır ve genellikle saldırı tespiti veri kümelerinde sınıf dengesizliği problemleriyle karşılaşılır. Dolayısıyla saldırı tespit veri kümesindeki sınıf düzensizliği öğrenme modelinin azınlık sınıfı daha iyi bir şekilde öğrenmemesine sebep olur.

Bu tez çalışmamızın amacı saldırı tespit veri setini en yakın komşu (K-NN), karar ağacı (C4.5) ve destek vektör makinesi sınıflandırma algoritmaları ile en iyi bir şekilde sınıflandırmak için her bir sınıf veri boyutuna göre SMOTE ve DE stratejileri veri dengeleme algoritmaları ile dengeli hale getirilmeye amaçlanmıştır. Dengeli hale getirilen veriler sınıflandırma algoritmaları ile sınıflandırılarak sonuçları analiz edilmiştir. Yapılan analiz sonucunda kullanılan dengeleme algoritmasının sınıflandırma üzerindeki etkisi olumlu olarak izlenmiştir.

1.2. Tezin Önemi

Saldırı tespit sistemleri, araştırmacılar tarafından halen üzerinde çalışma yapılan önemli bir alandır ve bu güne kadar araştırmacılar tarafından birçok saldırı tespit sistemi geliştirilmiştir. Bu sistemlerin iyi bir performansa sahip olabilmesi için tasarımlarında veri madenciliği algoritmalarından olan sınıflandırma algoritmaları kullanılmaktadır.

Saldırı tespit veri kümelerinde, veriler ağırlıklı olarak "normal" örneklerden ve küçük bir yüzdeyle "anormal" örneklerden oluşmakta ve bu sınıf dengesizliği problemlerine yol açmaktadır. Sınıf dengesizliği problemlerinde, eğitim verilerinden bir öğrenme modeli oluşturulduğunda genelde daha çok örneğe sahip olan sınıf daha iyi bir şekilde öğrenilebilir. Böyle bir problemi ortadan kaldırmak için az örneklere sahip olan sınıflar için çoğaltma örnekleme (Oversampling) ve/veya çok örneğe sahip olan sınıflar için ise azaltma örnekleme (under-Sampling) veri dengeleme yöntemleri kullanılır.

Tezde dengesiz saldırı tespit verileri ele alınarak ve bu veriler SMOTE ve DE stratejileri veri dengeleme yöntemleri ile dengelenmiştir. Dengeleme işleminden sonra saldırı tespit verileri literatürde yaygın olarak kullanılan K-NN, C4.5 ve SVM sınıflandırma algoritmaları ile sınıflandırılmıştır. Saldırı tespit verisine ek olarak SMOTE ve DE stratejileri veri dengeleme yöntemleri 22 dengesiz veri kümesine daha uygulanarak sınıflandırma başarıları artırmıştır. Dengesiz veri kümeleri dengeli hale geldikten sonra SMOTE ve DE veri dengeleme algoritmalarından elde edilen sonuçlar karşılaştırılmıştır. Bu tezin önemi, düzensiz veri kümeleri SMOTE ve DE stratejileri veri dengeleme yöntemleri ile düzenli hale getirerek sınıflandırma algoritmalarının sınıflandırma başarılarını artırmaktır.

2. KAYNAK ARAŞTIRMASI

Örnekleme teknikleri birçok dengesiz veri setlerinde çok başarılı olarak kullanılmaktadır. Örnekleme teknikleri veri setlerinin sınıflandırma başarılarını arttırmak amacıyla büyük bir önem taşımaktadır. Günümüze kadar araştırmacılar tarafından birçok örneklendirme tekniği geliştirilmiştir. Achlioptas ve ark çalışmalarında temel bileşen analizini hızlandırmak için üç düzeyde rastgele bir mekanizma önermişlerdir. Bu mekanizmalar şunlardır: Eğitimde Gram matrisinin örneklenmesi ve nicelleştirilmesi, kernel yöntemi genişlemelerinin değerlendirilmesinde rastgele yuvarlama ve kernel yönteminin kendisinin değerlendirilmesinde rastgele projeksiyon mekanizmalarıdır. Ana fikir, çekirdek işlevini, beklendiği gibi davranan rastgele bir kernel ile değiştirmektir. Bu süreç, azalan algoritmaların kernel değerlendirmelerinin ağırlıklı toplamlarını hesaplamak için ölçüm konsantrasyonundan yararlanabileceğini göstermektedir (Achlioptas ve ark.; 2002).

Sarah Curtis ve arkadaşları tarafından 2000 yılında nitel araştırmada örnekleme konusuna odaklanılmaktadır. Bu çalışmanın özel bir amacı, Miles ve Huberman (1994) tarafından önerilen kontrol listesinden elde edilen standardın bağlantısını araştırmaktır. Üç çalışma yapmışlar. Bu çalışmaların her biri belirli bir stratejiye sahiptir. Bu üç çalışma örneğinin dikkate alınması, örneklem seçiminde yapılan seçimlerin önemini doğrulamaktadır (Curtis ve ark., 2000).

Hastings tarafından 2018 yılında Markov zincirlerini kullanarak Monte Carlo örnekleme yöntemini detaylı olarak açıklamıştır. Monte Carlo yöntemi, geleneksel sayısal yöntemlere göre daha verimlidir. Ayrıca, Monte Carlo yöntemlerinin uygulanması, yüksek boyutlu olasılık dağılımlarından örneklemeyi gerektirir ve buda bilgisayar zamanı ve analizinde çok zor ve maliyetli olabilmektedir. (Hastings, 2018).

İletişim alanında, örneklemenin etkin bir rolü vardır. Joseph Waksberg tarafından rastgele rakam çevirme için örnekleme yöntemleri çalışmasında, kişisel görüşme anketleri yerine telefon anketlerine artan bir ilgi olduğunu belirtmiştir. Asıl amaç, kişisel görüşmelerin maliyetinden ve daha hızlı bir şekilde uzak durmaktır. Temelde telefon anketleri için kullanılan iki tür örnekleme çerçevesi vardır. İlki, telefon rehberlerinde isim ve numara listesidir. İkincisi, mevcut telefon santrallerinde rastgele rakamlı arama olarak adlandırılan tüm olası dört basamaklı rakamlar kümesidir. Bu yoğun ilgi sonucunda, telefon anketlerine verilen cevapların kalitesi üzerinde bir dizi çalışma yapılmıştır (Waksberg, 1978).

İlginç konjuge olmayan problemler için posterior dağılımın türetilmesindeki zorluk nedeniyle Bayesci istatistiklerinin ilköğretim düzeyinde öğretilmesi zor olabilir. Posterior dağılımını özetlemenin tek bir yöntemi vardır, doğrudan çıkar olasılık dağılımından taklit etmek ve daha sonra simüle edilmiş numuneyi keşfetmektir. Bu makalede, bu amaçla araştırmacı, üç çıkarım problemi için posterior dağılımları simüle etmek için Rubin's Örnekleme Önemi Yeniden Örnekleme algoritması (SIR) kullanılmasını önermiştir. Örnekleme Önemi Yeniden Örnekleme algoritması (SIR), yönteminin çeşitli avantajları vardır. İlk olarak, SIR algoritması çok çeşitli Bayesian çıkarım problemleri için otomatik olarak gerçekleştirilebilir. İkinci olarak, posterior dağılımı özetlemek için, öğrencinin sadece simüle edilmiş bir örneği özetlemesi gerekir. SIR algoritması basit olduğu için standart istatistiksel yazılım programı kullanan öğrenciler tarafından programlanabilir. MINITAB, geniş bir dağılım yelpazesi için rastgele örnekler üretme kabiliyetine sahip olduğundan uygun bir program olsa da, bu örneklerin çeşitli işlevlerini sayabilir ve daha sonra özet istatistikleri ve grafikleri kullanarak simüle edilmiş değerleri özetleyebilir. Bir eğitmen, MINITAB makro tesisini kullanarak SIR algoritmasını çok çeşitli Bayesian sorunları için programlayabilir (Albert, 1993).

SMOTE ilgili makalede, yazarlar Hui Han ve arkadaşları tarafından, veri setlerinde iki tür dengesizlikten bahsedilmiştir. Bunlardan biri sınıf dengesizliği olup bu durumda, sınıfların diğerlerinden daha fazla örneği vardır. Diğeri sınıf içi bir dengesizliktir, bu durumda bir sınıfın bazı altkümelerinin aynı sınıftaki diğer altkümelerden daha az örneği vardır. Daha iyi bir tahmin elde etmek için, sınıflandırma algoritmalarının çoğu, eğitim işleminde her sınıfın sınır çizgisini mümkün olduğu kadar öğrenmeye çalışmıştır. Geleneksel veri madenciliği yöntemleri dengesiz verileri dolayı tatmin edici değildir. Bu çalışmada iki yeni yöntem önermişlerdir, amaç bu sorunu çözmektir. Bu yazıda, iki yeni azaltma örnekleme yöntemi, Borderline-SMOTE1 ve Borderline-SMOTE2'yi sunmuşlardır. Yöntemler, SMOTE sentetik çoğaltma örnekleme tekniğine dayanmaktadır (Han ve ark., 2005).

Azınlık sınıflarının öngörülmesini iyileştirmesi için araştırmacılar Nitesh V. ve arkadaşları Sentetik Azaltma Çoğaltma Örnekleme Tekniğini (SMOTE) ve standart yükseltme prosedürünü birleştiren SMOTEboost algoritması önermişlerdir. Amaç, veri setinde azınlık sınıfından daha iyi bir model edinmek ve grubun genel doğruluğunu arttırmaktır. Deney sonuçları, SMOTEBoost'un, SMOTE'nin kabiliyeti nedeniyle

AdaCost'tan daha yüksek F değerleri elde edebileceğini göstermiştir (Chawla ve ark., 2003).

Diğer bir makalede, yazarlar Enislay Ramentol ve arkadaşları yapılan araştırmalarını genişletmişlerdir. Sentetik Azaltma Örnekleme Tekniği ile birlikte yeni örneklerin yapımı sırasında dengesiz veri setlerinin ön işleme tabi tutulması için yeni bir hibrid yöntem sunmuştur. Kaba Küme Teorisine (RST) ve bir alt kümenin alt yaklaşımına dayanan bir düzenleme tekniğinin uygulanmasıdır. Sunulan yöntem, öğrenme algoritması olarak C4.5 kullanarak iyi sonuçlar gösteren deneysel çalışmayı desteklemiştir. Bu çalışma, SMOTE tarafından çoğaltma örnekleme yapılması ve SMOTERSB olarak adlandırılan yüksek oranda dengesiz veri setlerinin sentetik örneklerinin üzerinde örneklemenin alınması için yeni bir hibrid yaklaşım sunmaktadır. SMOTE kullanarak azınlık sınıfının yeni sentetik örneklerinin oluşturulması ve geliştirilmesi bu yeni örneklerin iyiliği bu arada bir alt kümenin ve Kaba Küme Teorisinin düşük yaklaştırmasına dayanan düzenleme teknikleridir. En büyük katkı, sentetik örnekler üretmek için SMOTE kullanan yeni bir ön işleme yöntemi ve temizleme yöntemi olarak RST sunmaktır. Deneysel analiz sonuçları, SMOTE-RSB tekniğiyle dengelenmiş veri kümeleri çerçevesinde ön işlemler için elde edilen iyi ortalama sonuçları gözlemlemişlerdir (Ramentol ve ark., 2012a).

SMOTE tekniğinin farklı bir alanında, araştırmalar Güvenli Seviye SMOTE adı verilen yeni bir teknik üretmişlerdir. Güvenli Seviye SMOTE adı verilen teknik, aynı seviye boyunca azaltma örneklerini farklı ağırlık dereceleriyle düzgün bir şekilde örneklendirir. Güvenli seviye olarak adlandırılır. Güvenli seviye, en yakın komşu azaltma örneklerini kullanarak hesaplamaktadır. Azaltma örneklerini daha güvenli bir düzeyde daha fazla sentezleyerek, SMOTE ve Borderline-SMOTE'den daha iyi bir doğruluk performansı elde etmiştir (Bunkhumpornpat ve ark., 2009).

Farklı bir çalışmada, David D. ve arkadaşları tarafından denetimli öğrenme için yeni yöntemler geliştirmiştir bu metot Heterojen Belirsizlik Örneklemesi adlı yeni bir metottur. Belirsizlik örnekleme yöntemleri, önceki etiketli örneklere rağmen sınıfları belirsiz olan eğitim örnekleri için sınıf etiketlerini yinelemeli olarak talep eder. Bu yöntemler, bir uzmanın ihtiyaç duyduğu etiket sayısını büyük ölçüde azaltabilir. Bu yöntemle ilgili tek sorun, bir uygulama için en uygun sınıflandırıcının, örneklerin seçimi sırasında eğitilmesi veya kullanılması çok maliyetli olabileceğidir (Lewis ve Catlett, 1994).

Yapılan başka bir araştırmada Sentetik Azaltma Çoğaltma Örnekleme Tekniği (SMOTE), parçacık sürüsü optimizasyonu (PSO) ve radyal temel fonksiyon sınıflandırıcısı (RPF) birleştirilerek güçlü bir mekanizma önerilmiştir. SMOTE, eğitim veri setini dengelemek için pozitif sınıf için sentetik örnekler oluşturmak üzere uygulanır. Örneklenen eğitim verilerine dayanarak önerilen SMOTE + PSO-OFS algoritmasının etkinliği, benzetilmiş bir dengesiz veri seti ve üç gerçek dengesiz veri seti kullanılarak incelenmiştir. Bu üç veri seti, dengesizliği arttırmak amacıyla seçilmiştir. Önerilen algoritmanın etkinliğini açıklamak için benzetilmiş bir dengesiz veri setinde ve üç gerçek dengesiz veri setinden elde edilen deneysel sonuçlar sunulmuştur (Gao ve ark., 2011).

Reshma ve arkadaşları 2015'te önerilen makalede, çok sınıflı bir dengesiz veri setinin sınıflandırılması için bir metodoloji önermişlerdir. Bu metodoloji iki adıma sahiptir: İlk adımda, orijinal veri setini ikili sınıfların alt kümelerine ayırmak için Binarizasyon tekniklerini Biri-vs-Hepsi (OVA) ve Biri-vs-Biri (OVO) kullanmışlardır. Binarizasyon tekniklerini çalışmalarında kullanmalarının sebebi çok sınıflı veri madenciliği algoritmalarının ek komplikasyon gerektirmesi ve veri setlerinin performans seviyesini düşüren sınıf sınırlarını aşmasıdır. Bu durumda, çoklu sınıf problemini, sınıf çiftleştirme teknikleri kullanarak ayırt etmesi kolay olan ikili sınıf probleminin bir alt kümesine dönüştürmüşlerdir. İkinci adımda, SMOTE algoritması dengeli bir veri seti elde etmek için her bir dengesiz ikili sınıf alt kümesine karşı uygulanır. Ayrıca, rastgele orman (RF), iyi performansıyla bilinen bir karar ağacı grubundan biri olan bir yöntem olarak kullanılır. Ayrıca, sınıflandırma hedefine ulaşmak için (RF) kullanılmıştır. Amaç, çok sınıflı dengesiz veri problemini ele almak için SMOTE algoritmasını geliştirmektir. Kullandıkları veriler UCI veri tabanı deposundandır. Veri kümeleri (Landsat, Lenfografi, Hayvanat Bahçesi, Segment, Iris, Araba, Taşıt ve Dalga Biçimi) idir. Sonuçlar SMOTE + OVA algoritmasının dengesiz veri probleminde iyi performans verdiğini göstermiştir (Bhagat ve Patil, 2015).

Bulanık kaba küme teorisini kullanarak, E. Ramentol ve arkadaşları tarafından yapılan araştırmada yeni bir yeniden örnekleme yöntemi tanıtmışlardır. İlk önce, eğitim verilerini yeniden örneklemek için bir SMOTE yöntemi kullanmışlar, ardından dengeli kaba seti teorisine dayanan bir düzenleme tekniği uygulamışlardır. Bu yeni metodoloji SMOTE-FRST olarak dengelenmiş veri kümeleri için yeni bir karma düzenleme metodu olarak adlandırılmıştır. SMOTE-FRST algoritmasının performans testi için KEEL web tabanından veri setleri alınmıştır. Ayrıca, iyi bilinen C4.5 sınıflandırıcısını

öğrenme algoritması olarak kullanmışlar. Deneysel sonuçlarına göre SMOTE-FRST algoritması standart SMOTE algoritmasından birçok veri setinde üstün başarılar sergilemiştir (Ramentol ve ark., 2012b).

Bir başka çalışmada, Profesör Pawlak, Kaba set teorisinin belirsiz ve kararsız bilgiyle başa çıkmak için güçlü bir matematik aracı olduğunu ortaya koymuştur. Bu nedenle yazarlar Feng Hu ve Hang Li, SMOTE ve komşuluk Tabanlı Kaba Set Modeli arasında bağlantı kuran yeni bir yöntem önermişlerdir. Yöntem, Komşuluk Kaba Küme Sınırlı SMOTE (NRSBoundary SMOTE). Önerilen yöntem üç adımdan oluşmaktadır. İlk olarak, sınır bölgesindeki azınlık sınıfı örnekleri ve çoğunluk sınıfı örnekleri daha düşük bir karar yaklaşımıyla hesaplanır. İkinci olarak, her azınlık sınıfı örneği için SMOTE algoritması çağıran sentetik örnekler üretilir. Üçüncüsü, çoğunluk sınıfı örneklerin karar alanlarını etkilemeden rasyonel sentetik örnekleri daha düşük bir karar yaklaşımıyla seçerler. Kullanılan veri setleri UCI'den alınmıştır. Sonuçlar, önerilen yöntemin C4.5, CART ve KNN ile karşılaştırdığınızda SMOTE'den daha iyi performans gösterdiğini göstermektedir. Buna rağmen, SMOTE, SVM kullanırken NRSBoundary-SMOTE'den daha iyidir. Önerilen yöntem çoğaltma örnekleme için aktif bir yöntemdir. Ayrıca, sentetik verileri filtrelemek için daha fazla zaman harcamaktadır. Uzun çalışma süresi nedeniyle büyük bir veri setini işlemek zor olmaktadır (Hu ve Li, 2013).

Alberto ve arkadaşları 2010'da iki adımdan oluşan yeni bir metodoloji önermişlerdir. İlk adımda, biri-vs-biri binarization teknik yöntemini kullanarak veriler iki sınıflı yapıya dönüştürülmüştür. Ancak bu dönüştürülmüş veriler düzenli değildir. İkinci adımda, veriler sınıflandırma algoritmasına aktarılmadan önce iki sınıflı yapıdaki SMOTE örnekleme yöntemi kullanılarak ikili veriler dengelenmiştir. Deneysel sonuçlar kurala dayalı sınıflandırma algoritmasının daha iyi sınıflandırma performansı verdiğini göstermektedir (Fernández ve ark., 2010).

Dengesiz veri kümeleri alanlarında yazarlar Alexander Yun ve arkadaşları, veri çoğaltma ve veri azaltma dengesiz metin veri kümelerinin sınıflandırılmasına etkisi isimli bir çalışma yapmışlardır. Dengesiz bir veri kümesinin sorunu, eğitim kümesinde bir sınıfın çok daha düşük bir olasılığı olduğunda ortaya çıkar ve asıl sorun sınıflandırma sürecidir. Dengesiz bir veri kümesi sorununu çözmenin bir yolu, eğitim setini yeniden örneklendirmektir. İlk olarak, metnin veri kümesinin örneklem metriği, örnekleme tekniği, rastgele örnekleme, örnekleme teknikleri ve değişiklikleri gibi örnekleme yöntemleri kullanılarak uyguladığı örnekleme teknikleri olarak tarif edilmiştir. Rastgele Örnekleme gibi Komşu Tabanlı Rasgele örnekleme, Örnekleme

Teknikleri (NBOTE), Üretken Çoğaltma Örneklemeli Terkedilmiş ve Rastgele çoğaltma Örneklemeli Sentetik Azaltma Örnekleme Tekniği (SMOTE) gibi çoğaltma örnekleme, örneklemeden sonra Naif Bayes, en yakın komşu ve SVM'leri veri tabanında verileri sınıflandırılmıştır. Sonuç olarak, kullanılacak en iyi yeniden örnekleme tekniği çoğu zaman uygun sınıflandırıcıdadır ve veri kümesine bağlıdır (Liu, 2004).

2016'da Varsha Babar ve arkadaşı, temel bir örnek azaltma tekniği (MLPUS) önerdi, veri azaltma örnekleme yaparken bilgi dağıtımını koruyacaktır. Bu önerilen yaklaşım üç adımdan oluşmaktadır. İlk adımda başlangıç MLP yapısı kullanılarak örnekler seçilmişti. İkinci adımda, SM kullanarak önemli örneklerin değerlendirilmesi yapılmıştır. Ve üçüncüsü SM değerlendirmesinde seçilen numuneleri kullanan bir MLP eğitimi vardır. Bu teknik, çoğunluğun yanı sıra azınlık örneklerinden önemli örnekleri tanımlamak için rastgele ölçüm değerlendirmesini kullanmaktadır. Bu aşağı örnekleme tekniği çok sınıflı dengesizlik problemi için genişletilebilir ve bu teknik, makine öğreniminde meydana gelen dengesiz sorunu çözmek için kullanılabilir (Babar ve Ade, 2016).

Sukarna Barua ve arkadaşları yapılan çalışmaya göre, Major Ağırlıklı Azaltma Çoğaltma Örnekleme Tekniği (MWMOTE) adı verilen yeni bir yöntem sunmuştur. MWMOTE ilk önce öğrenmesi zor bilgilendirici azınlık sınıfı örneklerini tanımlar ve onlara en yakın çoğunluk sınıfı örneklerinden Öklid mesafelerine göre ağırlık atar. Sonuçlar, bu yöntemin, genel olarak eğri altındaki alan (AUC) olarak bilinen, geometrik ortalama (G-ortalama) ve alıcı çalışma eğrisi altındaki alan (ROC) gibi çeşitli değerlendirme ölçütleri açısından mevcut diğer yöntemlerden daha iyi olduğunu göstermektedir (Barua ve ark., 2014).

Xu-Ying ve arkadaşı keşifçi sınıf dengesizliği öğrenmesi için düşük örnekleme bu eksikliği gidermek için iki algoritma önermişlerdir. İki yöntem, çoğunluk sınıfından birkaç alt grubu örnekleyen Easy Ensemble (kolay topluluk) ve Balance Cascade (Denge Cascade) öğrencileri sırayla eğitir. Her iki algoritma da çoğunluk sınıfını örneklemeden daha iyi kullanmaktadır ve önerilen yöntemler eğitim süresini azınlık için kullanılabilir. Deneysel sonuçlar, her iki yöntemin de ROC Eğrisi Altındaki Alan, F-ölçüm ve G-ortalama değerlerinin mevcut birçok sınıf dengesizliği öğrenme yönteminden daha yüksek olduğunu göstermektedir. Ayrıca, yaklaşık olarak aynı eğitim süresine sahiptirler ki bu, diğer yöntemlerden daha hızlıdır (Liu ve ark., 2009).

Dengesiz veri setleri için sınıflandırma algoritmalarında, araştırmacı Vaishali Ganganwar yapılan araştırmaya göre, veri ve algoritmik seviyelerde önerilen mevcut

bazı çözümleri hazırlamıştır. Araştırmacıların çalışmasına göre, değiştirilmiş destek vektör makinesinin, kaba küme tabanlı azınlık sınıfına yönelik kural öğrenme yöntemlerinin ve maliyete duyarlı sınıflandırıcının, dengesiz veri kümesinde iyi performans gösterebileceğini kanıtlamışlardır. Araştırmacılar, Hibrit örnekleme tekniklerinin sadece çoğaltma örnekleme yâda örneklem azaltmadan daha iyi olabileceğini kanıtlamışlardır (Ganganwar, 2012).

Haibo He ve arkadaşları 2008 yılında dengesiz veri kümelerinden öğrenme için örnekleme yaklaşımı geliştiren ve dengesiz öğrenme için ADASYN (Uyarlamalı Sentetik Örnekleme Yaklaşımı) adında yeni bir Adaptıf sentetik yöntem önermiştir. ADASYN'nin ana fikri, öğrenmedeki zorluk derecesine göre farklı azınlık sınıfı örnekleri için ağırlıklı bir dağılım kullanmaktır. ADASYN süreci veri dağıtım yoluyla öğrenmeyi iki şekilde geliştirir: sınıf dengesizliğinin yol açtığı eğilim azınlık ve sınıflandırma kararını sınırını uyarlamalı olarak zor örneklere kaydırmaktadır. Bu makalede, ADASYN'in bu alanda güçlü bir yöntem sunduğunu kanıtlamışlar (He ve ark., 2008).

2006'da Zhi-Hua ve arkadaşı Maliyete duyarlı sinir ağlarının eğitiminde örneklemenin ve eşik hareketinin etkisini deneysel olarak araştırdılar. Hem örneklem azaltma hem de çoğaltma örneklem altında kabul edilir. Bu teknikler (örnekleme ve eşik hareketi) eğitim verilerinin dağılımını değiştirebilir. Sonuçlar, çok sınıflı görevlere sahip sinir ağlarında maliyete duyarlı öğrenmenin, iki sınıf görevinden daha zor olduğunu ve sınıf dengesizliği derecesi nedeniyle zorluğun daha yüksek olabileceğini ortaya koymuştur. Ampirik çalışma ayrıca sınıf dengesizliği sorununu çözmede etkili olduğuna inanılan bazı yöntemlerin aslında dengesiz iki sınıf veri kümeleriyle öğrenmede etkili olabileceğini göstermektedir (Zhou ve Liu, 2006).

Literatürde farklı makine öğrenmesi teknikleriyle geliştirilmiş pek çok saldırı tespit sistemi mevcuttur.

Çetin KAYA tarafından saldırı tespit sistemlerinde makine öğrenmesi tekniklerinin kullanılmıştır. Yapılan çalışmada KDD CUP99 ve NSL-KDD veri setleri kullanılarak, makine öğrenmesi tekniklerinden Bayes ağları, destek vektör makinesi, K en yakın komşu algoritması, yapay sinir ağları ve karar ağaçlarının, işlem zamanı, sınıflandırma başarısı, duyarlılık, seçicilik, kesinlik ve F-ölçütü yönünden saldırı tespit sistemlerindeki performansı incelenmiştir. Elde edilen sonuçlara göre sınıflandırıcıları, sınıflandırma başarısına göre değerlendirdiğimizde, normal davranışları ayırt etmede,

karar ağaçları diğer sınıflandırıcılara göre daha başarılıdır. DOS saldırılarının tespitinde KNN, karar ağaçları ve YSA %100'e yakın bir başarıya ulaşmıştır. PROBE saldırılarının doğru tespitinde KNN, YSA ve karar ağaçları daha iyi sonuç vermektedir (Kaya, 2016).

2018 yılında Umut Karacalarlı, KDD99 veri setinde destek vektör makinesinin sınıflandırma performansı üzerine bir araştırma yapmıştır. Bu çalışmada KD 999 veri setinde destek vektör makinesi ile yapılan sınıflandırma performansının Fisher skoru özellik seçim yöntemiyle geliştirilebileceği gösterilmiştir. Elde edilen sonuçlara göre, en iyi isabet oranı (%90.74), Fisher skoruyla birlikte önem puanına göre ilk 5 özelliğinden oluşan bir alt kategoriyi kategorize ederek elde edildi ve doğruluk oranı %87 idi. Fisher puanının %80'idi. KDD99 veri kümesi üzerinde, destek vektör makinesi ile yapılan sınıflandırmanın performansının, Fisher Score özellik seçim yöntemiyle artırılabildiği gösterilmektedir. (Karacalarlı, 2018)

2011 yılında MEHMET CEM yüksek hızlı bilgisayar ağları için daha hızlı bir saldırı tespit metodu önermiştir. Bu tezin sonuçları göre, hızlı izinsiz giriş tespit sistemlerinin, paket yüklerine bağlı olmayan imzalarla tasarlanabileceğini göstermektedir. Yalnızca paket başlığını inceleyerek önemli miktarda izinsiz girişin tespit edilebileceğini varsayarlar (Tarım, 2011).

3. MATERYAL VE YÖNTEM

Son yıllarda, makine öğrenme tekniklerini kullanarak gerçek dünya sorunlarına çözüm girişimlerine ilgi artmıştır. Sınıflandırma kategorileri yaklaşık olarak eşit şekilde temsil edilmezse veri kümesi dengesizdir. Uydu ve radar görüntüleri, petrol sızıntısı tespit etme, kelime telaffuzu öğrenme, metin sınıflandırma vb. dengesiz veri kümeleri farklı alanlarda görülebilir (Chawla, 2009). Dengesiz veri sorunu ile başa çıkmak için yaygın bir uygulama yapay olarak, çoğaltma örnekleme ve / veya azaltma örnekleme yoluyla yeniden dengelemektedir (Ganganwar, 2012). Dengesizliği gidermek için farklı örnekleme yöntemleri kullanılmaktadır. Dengeli veri sağlamak için SMOTE algoritması ve DE stratejileri kullanılmıştır. Bu veri kümeleri dengeli hale geldikten sonra K-NN, C4.5 ve SVM sınıflandırma algoritmaları ile sınıflandırılmıştır.

3.1. Dengesiz Veri

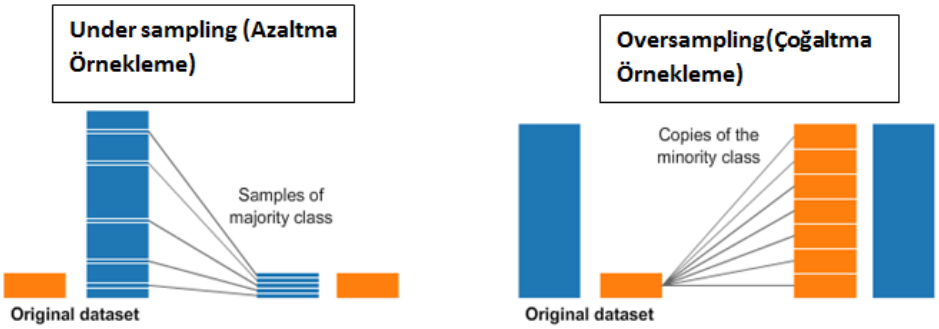
Sınıf dengesizliği sorunu veri madenciliğinde en büyük sorunlardan biridir. Sınıflardaki örnek miktarı yakın değilse, veri kümesi dengesiz olarak adlandırılır. Sınıf dengesizliği problemini üstesinden gelmek için örnekleme yöntemleri geliştirilmiştir. Örnekleme yöntemi, sınıf dengesizliği probleminin üstesinden gelmenin etkin ve popüler bir yoludur. Örnekleme yöntemlerinin amacı, nispeten dengeli bir sınıf dağılımı içeren bir veri kümesi oluşturmaktır. Temel örnekleme yöntemlerinden ikisi, rastgele azaltma örnekleme ve rastgele çoğaltma örneklemedir. Rastgele örneklemede, çoğunluk sınıfı örnekleri daha dengeli bir dağılıma ulaşılıncaya kadar rastgele atılır. Bu önemli bir problem olabilir, çünkü bu tür verilerin kaybı, azınlık ve çoğunluk durumları arasındaki karar sınırlarının bilinmesini zorlaştırabilir, bu da sınıflandırma performansının düşmesine neden olabilmektedir. Rastgele çoğaltma örneklemede, daha dengeli bir dağılıma ulaşılıncaya kadar azınlık sınıfı örnekleri kopyalanır ve veri setinde tekrarlanır. Rastgele çoğaltma örneklemede, örnekler bazen çok yüksek sayıda tekrarlanır. Örnekleri bu şekilde kopyalamak, sınıflandırıcıda çoğaltma öğrenmeye neden olabilir ve sınıflandırıcının performansını son derece zayıf hale getirebilir. Bu sınırlamaların üstesinden gelmek için, daha gelişmiş örnekleme teknikleri geliştirilmiştir. Bu tekniklerden bazıları düşük örnekleme tekniği ve çoğaltma örnekleme tekniğidir (Hoens ve Chawla, 2013).

3.2. Örnekleme

Örnekleme, işlenecek bir veri öğesinin olasılıklı seçim sürecini ifade eder. Örnekleme, büyük veri tabanları için geleneksel veri madenciliği bağlamında uzun süredir kullanılan eski bir istatistiksel tekniktir. Bazı örnekleme stratejileri getirilmiştir, çoğaltma örnekleme, azaltma örnekleme, değiştirme ile rastgele çoğaltma örnekleme, rastgele azaltma örnekleme, bilinen bilgilere dayalı sentetik yeni örnek örneklerle çoğaltma örnekleme ve yukarıdaki tekniklerin kombinasyonları. Bu tezde çoğaltma örnekleme ve azaltma örnekleme kullanılmıştır (Zyt ve ark., 2002).

Çoğaltma örnekleme ve azaltma örnekleme ana işlevi Şekil 3.1'de göstermektedir.

Dengesiz veri kümelerindeki çeşitli çalışmalar, çoğaltma örnekleme ve azaltma örneklemenin farklı çeşitlemelerini kullanmış ve örneklemenin örneklemenin karşılaştırılmasıyla ilgili örneklemelerini sunmuştur (Chawla, 2009). Yeniden örnekleme yöntemlerinin faydası, azınlık ve çoğunluk örnekler arasındaki oran, verilerin diğer özellikleri ve sınıflandırıcının niteliği gibi bir dizi faktöre bağlıdır. Bununla birlikte, yeniden örnekleme yöntemleri önemli avantajlar göstermiştir. Azaltma örnekleme potansiyel olarak yararlı verileri yok edebilirken, çoğaltma örnekleme yapay olarak ayarlanan verinin boyutunu artırır ve sonuç olarak öğrenme algoritmasının hesaplama yükünü artırır (Ganganwar, 2012).



Şekil 3.1. Çoğaltma örnekleme ve Azaltma örnekleme diyagramı.

Şekil 3.1'de görüldüğü gibi, azaltma örnekleme yöntemleri örneği çok fazla olan sınıfların örneklerini azaltmaya çalışmaktadır. Bu azalma işlemi rastgele yapılabilir; bu durumda rastgele düşük örnekleme olarak adlandırılabilir veya düşük örnekleme

bilgilendirilmiş olarak adlandırılan bazı istatistiksel bilgiler kullanılarak yapılabilir Çoğunluk sınıfı örneklerini daha da iyileştirmek için veri tekniklerine bazı azaltma örnekleme yöntemleri ve tekrarlama yöntemleri uygulanır (Shelke ve ark., 2017). Sonuç olarak, eğitim setindeki toplam kayıt sayısı büyük ölçüde azalır. Bu, sınıflandırma sırasında, eğitim süresinin de büyük ölçüde azaldığı anlamına gelir. Çok yüksek veri setleriyle uğraştığı için hafızada önemli bir tasarruf vardır. Bununla birlikte, çoğunluk sınıfından alınan verilerin kaldırılması nedeniyle, doğru bir modelin oluşturulmasında sınıflandırıcı için yararlı olabilecek belgelerin çıkarılması durumunda çok fazla değerli bilginin kaybedilmesi mümkündür. (Liu, 2004). Düşük Örnekleme Teknikleri:

1. Tabanlı azaltma örnekleme tekniği (MLPUS): Örnekleme yaparken bilginin dağıtımını koruyacak MLP tabanlı azaltma örnekleme tekniğidir. MLPUS üç temel mekanizmayı içerir:
   a) Çoğunluk sınıfı örneklerinin derlenmesi
   b) Hassasiyet Ölçüsü (SM) değerlendirmesini kullanarak önemli örneklerin seçimi
   c) SM değerlendirmesinde seçilen örnekleri kullanarak MLP'nin eğitimi (Shelke ve ark., 2017).
2. Easy Ensemble: Easy Ensemble yönteminde, çoğunluk sınıfı birkaç alt gruba ayrılır ve her alt kümenin boyutu bir azınlık sınıfının boyutuna eşittir. Daha sonra her bir alt küme için, tüm azınlık sınıfını ve çoğunluk sınıfı altkümesini kullanan bir sınıflandırıcı kullanılır. Tüm sınıflandırıcılardan elde edilen sonuçlar nihai kararı almak için birleştirilir (Shelke ve ark., 2017).
3. Balanced Cascade: Bu yöntem denetimli öğrenme yaklaşımını izler. Balanced Cascade yöntemi şu şekilde çalışır: Azınlık sınıfı örnek sayısına eşit sayıda örnek içeren çoğunluk sınıfının alt kümesi oluşturulur (Shelke ve ark., 2017).

Azaltma örneklem yöntemlerinin avantajları eğitim verisi çok büyük olduğunda eğitim verisi örneklerinin sayısını azaltarak eğitim süresi ve depolama sorunlarını iyileştirmeye yardımcı olabilir. Dezavantajları ise potansiyel olarak yararlı bilgileri atabilir, bu yararlı bilgiler bazen sınıflandırma modelleri oluşturmak için önemli bilgiler olabilir. Ayrıca, rastgele düşük örnekleme ile seçilen örnek, taraflı bir örnek olabilir.

Nüfusun kesin bir temsili olmayabilir. Bu nedenle gerçek test verileri setiyle yanlış sonuçlara neden olur.

Çoğaltma örnekleme yönteminde, veri setini dengelemek için azınlık sınıfına yeni örnekler eklenir. Bu yöntemler rastgele çoğaltma örnekleme ve sentetik çoğaltma örnekleme şeklinde sınıflandırılabilir. Rastgele çoğaltma örnekleme yönteminde, bir azınlık sınıfının boyutunu artırmak için mevcut azaltma örnekleri çoğaltılmaktadır. Sentetik çoğaltma örnekleme tekniğinde azınlık sınıfı örnekleri için yapay örnekler üretilmiştir. Üretilen bu yeni örnekler azınlık sınıfına gerekli bilgileri ekleyebilir ve sınıflar için çok iyi bir sınıflandırma modeli oluşturur (Shelke ve ark., 2017). Çoğaltma örnekleme, eğitim setinde azınlık sınıfı üyelerinin sayısını artırmaya çalışmaktadır. Çoğaltma örneklemenin avantajı, tüm üyeleri azınlık ve çoğunluk sınıflarından uzak tutmamızdan dolayı orijinal eğitim setinden hiçbir bilginin kaybolmamasıdır. Ancak, dezavantajı, eğitim setinin boyutunu arttırmamızdır. Bu nedenle, eğitim süresini ve eğitim setini tutmak için gereken bellek miktarını da artar. Çok yüksek boyutlu veri kümeleriyle uğraştığımızdan, zaman karmaşıklığını ve bellek karmaşıklığını makul kısıtlar altında tutmak için konusunda dikkatli olmak gerekir. Yeniden örneklemeye harcanan zamanı göz önüne almazsak, düşük örnekleme zamanı ve hafıza kapasitesinden daha iyi performans gösterecektir. Durum böyle olduğu için, uygulanabilir olması için çoğaltma örneklemenin sınıflandırma performansı açısından düşük örneklemeden daha iyi olması gerekir. Düşük örnekleme veya çoğaltma örneklemenin sınıflandırma performansı açısından en iyisi olup olmadığına ilişkin geçmiş araştırmalar kesin sonuçlara ulaşamamıştır. Büyük olasılıkla, çelişkili sonuçlar farklı veri kümelerinin ve sınıflandırma algoritmalarının birleşmesinden kaynaklanmaktadır. Ek olarak, yeniden örnekleme yönteminin seçimi muhtemelen hem alan hem de soruna özgüdür (Liu, 2004).

SMOTE örnekleme tekniği yanında Çoğunluk Ağırlıklı Azaltma Örnekleme Tekniği (MWMOTE), Adaptıf sentetik örnekleme (ADASYN) ve RAMOBoost teknikleri de sıkça kullanılmaktadır. Mevcut sentetik çoğaltma örnekleme yöntemleri, bazı senaryolarda yetersizlikler ve uygunsuzluklara sahip olabilmektedir. Bu sorunların üstesinden gelmek için, MWMOTE yöntem önerilmiştir. MWMOTE'nin amacı iki yönlüdür: örnek seçim sürecini iyileştirmek ve sentetik örnek üretim sürecini iyileştirmek (Shelke ve ark., 2017). Dengesiz veri setini işlemek için, Haibo He ve arkadaşları yeni bir yöntem yaklaşım ADASYN önermiştir. Sentetik örnek üretim

sürecinde, azaltma örneklerinin ağırlıklı dağılımını kullanılmıştır. Azaltma örnekleminin önemine bağlı olarak azaltma örneklemine ağırlık eklenir (Shelke ve ark., 2017). RAMOBoost Arttırılmış Azaltma Çoğaltma Örnekleme, örnekleme ağırlıklarına bağlı olarak sistematik olarak sentetik örnekler üreten bir tekniktir. Bu yöntem iki aşamada çalışır. Birinci aşamada karar hem çoğunluk hem de azınlık sınıflarından öğrenilmesi zor olan örneklere doğru sınır değiştirilir. Sentetik örneklerin üretilmesinde ikinci aşamada, sıralı bir örnekleme olasılık dağılımı kullanılır. RAMOBoost, SMOTE-N yönteminde kullanılan teknikleri seçerse, nominal özelliklere sahip veri kümelerini kullanabilir (Shelke ve ark., 2017). Bu tez çalışmasında kullanılan veri setlerini dengelemek için SMOTE tekniği kullanılmıştır, SMOTE tekniği aşağıda açıklanmıştır.

3.2.1. Sentetik azaltma çoğaltma örnekleme tekniği (SMOTE)

SMOTE azınlık sınıfının, değiştirme ile çoğaltma örnekleme yerine sentetik örnekler oluşturarak çoğaltma örneklendiği bir çoğaltma örnekleme yaklaşımı önermiştir. Bu yönteme sentetik azaltma çoğaltma örnekleme tekniği denir. Bu süreç, el yazısı karakter tanımada başarılı olduğunu kanıtlayan bir teknikten esinlenmiştir. Bu yöntemin sözde kodu vardır. Şekil 3.2'de göstermiştir. SMOTE'nin algoritmasının prosedürü şöyle açıklanabilir: Yeni bir örnek oluşturmak için SMOTE kullanılarak, En yakın komşulardan bir komşu seçilir. İncelenen nitelik ile seçilen komşu arasındaki fark alınır. Fark GAP ile çarpılır (GAP, 0 ile 1 arasında rastgele bir sayı anlamına gelir). Sonuç, seçilin nitelik örneğine eklenir. Bu işlem istenen sonucu elde edene kadar tekrarlanır (Chawla ve ark., 2002). Bu çalışmada, sınıflandırma veri kümesi dengelemiş hale getirmek için SMOTE algoritması kullanmıştır.

SMOTE'nin sözde kodu

SMOTE'nin **Algoritması** (T, N, k)

Giriş:

T: Azınlık sınıfı örneklerinin sayısı

N: % SMOTE Miktarı

K: En yakın Komşular sayısı

Çıktı: $(N / 100) * T$

1.

($Eğer\ N\%\ 100'den\ az\ ise, azınlık\ sınıfı\ örnekleri\ rastgele\ ayarlayın,$ çünkü $bunların\ rastgele\ bir\ y$üzdesi $SMOTEd\ olacaktır.$)

2. **If** $N < 100$
3. **Then** T azınlık sınıfı örneklerini randomize
4. $T = (N/100) * T$
5. $N = 100$
6. **end if**
7. $N\ (int)(N/100)(* SMOTE\ miktarının\ yekpare\ katları\ olarak\ kabul\ edilir\ 100.*)$
8. k = En yakın komşu sayısı
9. $numattrs$ = Niteliklerin sayısı
10. Örnek [ ][ ]: orijinal azınlık sınıfı örnekleri için dizi
11. $Yeni\ endeks$: 0'a ilklendirilen, üretilen sentetik örneklerin sayısını tutar
12. $Sentetik$ [ ][ ]: sentetik örnekleri için dizi

($* Yalnızca\ her\ azınlık\ sınıfı$ örneği için $en\ yakın\ komşuyu\ hesapla.*$)

13. **for** $i \leftarrow 1$ **ile** T
14. İ'ye en yakın komşuları k hesapla ve dizileri nnarray'de kayıtla
15. Doldur $(N, i, nnarray)$
16. **endfor**

$Doldur\ (N, i, nnarray)$ ($* Sentetik$ örnekler üretme $fonksiyonu.$ *)

17. **While** $N \neq 0$
18. 1 ile k arasında bir rasgele sayı seçin, nn olarak adlandırın. Bu adım i'nin en yakın komşularından k birini seçer.
19. **for** $attr \leftarrow 1$ **ile** $numattrs$ için
20. Hesaplamak: $dif = Örnek[nnarray[nn]][attr] - Örnek[i][attr]$
21. Hesaplamak: gap = 0 ile 1 arasında rasgele sayı
22. $Sentetik\ [yeni\ dizin][attr] = Örnek\ [i][attr] + gap * dif$
23. **endfor**
24. $yeni\ indeks$ ++
25. $N = N - 1$
26. **endwhile**
27. **return** ($* Doldurun\ Sonu.*$)

Sözde-Kodun Sonu.

Şekil 3.2. SMOTE sözde kodu

3.3. Sınıflandırma Algoritmaları

Sınıflandırma, geniş bir uygulama yelpazesine sahip, reklam hedeflemesi, spam tespiti, risk değerlendirmesi, tıbbi teşhis ve resim sınıflandırmasıyla birlikte makine öğrenmesinde en yaygın kullanılan tekniklerden biridir. Sınıflandırmanın temel amacı, girdilerden bir kategori tahmin etmektir. Bir makine öğrenimi (ML) algoritmasının performansı büyük ölçüde veri kümesine ve boyuta bağlıdır. Bu nedenle, etkili bir ML algoritması seçmenin makul bir yolu deneme yanılma deneylerine dayanmalıdır (Yucel, 2016). Makine öğreniminde sınıflandırma kavramı genellikle denetimli, denetimsiz ve yarı denetimli öğrenme yöntemleri olarak ele alınmıştır (Pérez-Ortiz ve ark., 2016).

Makine öğrenmesi algoritma türleri: Denetimli, Denetimsiz ve Yarı Denetimli makine öğrenmesidir. Denetimli öğrenmede, operatör makine öğrenme algoritmasına istenen giriş ve çıkışları içeren bilinen bir veri kümesini sağlar ve algoritma, bu giriş ve çıkışlara nasıl ulaşılacağını belirleyen bir model oluşturmaktadır. Denetimli öğrenme yöntemi örnekleri: Sınıflandırma, Regresyon ve Tahmindir. Denetimsiz Makine Öğrenmesi, denetimsiz bir öğrenme sürecinde, büyük veri setlerini yorumlamak ve bu veriler arasındaki ilişkiyi ortaya koyan bir model oluşturmaktadır. Yarı denetimli makine öğrenmesinde, yöntemin çıkışlarına olumlu veya olumsuz olarak geri dönüşler uygulanır ve model bu şekilde oluşturulur. Örnek olarak takviyeli öğrenme yarı denetimli bir öğrenmedir.

En Yaygın ve Popüler Makina Öğrenmesi Algoritmaları K-En Yakın Komşu (Denetimli Öğrenme), Destek Vektör Makinesi Algoritması (Denetimli Öğrenme) ve C4.5 (Karar ağacı) (Denetimli Öğrenme). Bu tezde kullandığımız sınıflandırma algoritmaları. K En Yakın Komşu (K-NN), Destek Vektör Makinesi (SVM) ve (Karar ağacı) C4.5'tir.

3.3.1. K-En Yakın Komşu (K-NN)

K-En Yakın Komşu, yeni bir örnek sorgusunun sonucunun K tane en yakın komşu örneğin kategorisinin çoğunluğuna göre sınıflandırıldığı denetimli bir öğrenme algoritmasıdır. Bu algoritmanın amacı, niteliklere ve eğitim örneklerine dayalı yeni bir nesneyi sınıflandırmaktır. K-NN uygulamalarının basitliği nedeniyle, ağırlıklı K-NN, çekirdek K-NN ve karşılıklı k-NN gibi değiştirilmiş farklı k-NN modelleri önerilmiştir. Bir isteğin komşu bir örnekle ilişkisi temel olarak Öklid mesafesi gibi bir benzerlik ölçüsü ile ölçülür (Ertuğrul ve Tağluk, 2017). K-NN yönteminde önce test verisi

değerleriyle eğitim veri kümesindeki veri değerleri arasındaki Öklid uzaklıkları hesaplanır. Hesaplanan uzaklıklara göre test verisine en yakın mesafedeki k komşu sınıf belirlenir. Şekil 3.3'te K-NN algoritmasının şekli verilmiştir.

K-en yakın komşu algoritması nasıl hesaplanır:

1. K parametresini belirlenir.
2. Sorgu örneği ile tüm eğitim örnekleri arasındaki mesafe hesaplanır.
3. Mesafe sıralanır ve minimum mesafedeki K tane komşu örnek belirlenir.
4. Sınıflandırma için, her bir kategorideki k komşuları arasındaki veri noktalarının sayısını sayın.

KNN'nin uzaklık ölçütleridir (3.1), (3.2), ve (3.3) deki gibi hesaplanır:

Öklid:

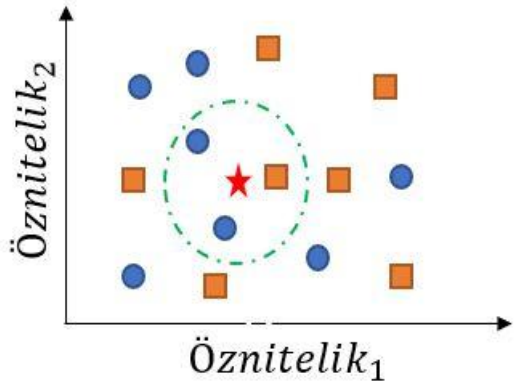
$$D(X, Y) = \sqrt{\sum_{i=1}^{k}(Xi - Yi)^2} \quad (3.1)$$

Manhattan

$$D(X, Y) = \sum_{i=1}^{k}|Xi - Yi| \quad (3.2)$$

Minkowski

$$D(X, Y) = \left(\sum_{i=1}^{k}(|Xi - Yi|)^q\right)^{1/q} \quad (3.3)$$



Şekil 3.3. K En Yakın Komşuluk Algoritmasında K Adet komşuya yakınlık

Veri madenciliği tekniği olarak K-NN, regresyonun yanı sıra sınıflandırmada da çok çeşitli uygulamalara sahiptir. k- NN, birçok alanda basitlik, verimlilik ve sınıflandırma performansı gibi birçok önemli avantaja sahiptir.

Avantajlarına rağmen, K-NN'nin sınıflandırma algoritması bazı dezavantajlara sahiptir. Eğitim seti büyük olduğunda KNN çalışma süresi düşük performansa sahip olabilir. Ayrıca, hangi mesafe ölçütünün kullanılacağı ve en iyi sonuçları elde etmek için hangi özelliğin kullanılacağı net değildir.

3.3.2. Destek Vektör Makinesi (SVM)

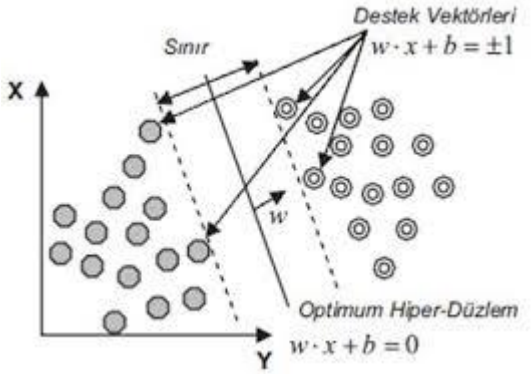
SVM denetimli makine öğrenmesidir ve sınıflandırma için popüler bir stratejidir. Vapnik tarafından 1998 yılında önerilmiş güçlü bir sınıflandırıcıdır. SVM, özellikle alandaki veriler dengesiz ise, sınıflandırma algoritmaları alanında iyi bir seçimdir. Destek vektör sınıflandırma amacı, yüksek boyutlu bir özellik uzayda bir "iyi" ayıran hiper düzlem için etkili bir şekilde arama yapmaktır. 'İyi', genel anlamda bir performans ölçütü anlamına gelir (Mammone ve ark., 2009).

Verilen X örneğini sınıflandırmak için öncelikle en uygun hiperdüzlem bulunur. X örneği SVM yöntemiyle formüle edilir ve f (x) işlevi sıfırdan büyükse pozitif sınıfa atanır, sıfırdan küçükse negatif sınıfa atanır. Destek vektör yöntemi hiperdüzleme en yakın pozitif ve negatif örnekler arasındaki mesafenin (sınır genişliğinin) en yüksek olduğu bir hiperdüzlem bulmaya çalışır. Sınır genişliği (M) Eşitlik 3.4 ve 3.5'teki denklem gibi hesaplanır. Şekil 3.4'te SVM algoritmasının yaklaşımı gösterilmiştir.

Hard-margin : $$y_i \ (\overrightarrow{W} . \ \overrightarrow{X} - b) \geq 1, \ \text{for all } 1 \leq I \leq n . \qquad (3.4)$$

Soft-margin:

$$\left[\frac{1}{n} \sum_{i=1}^{n} \max \ (0,1 \ - \ y_i(\ \omega . x_i - b)) \right] + \ \lambda \ ||w||^2 \qquad (3.5)$$



Şekil 3.4. SVM (Doğrusal olarak ayırma)

Tüm sınıflandırma tekniklerinin, analiz edilen verilere göre aşağı yukarı önemli olan avantaj ve dezavantajları vardır. Yüksek boyutlu uzaylarda etkilidirler ve Karar fonksiyonunda bir takım eğitim noktaları kullanılır. Dezavantajı olasılıksa tahminler üretememe. Örneğin, veriler düzenli bir şekilde dağıtılmadığında veya bilinmeyen bir dağıtıma sahip olduğunda, bu yöntem kullanılabilir. Klasik sınıflandırma teknikleri puanına girmeden önce bilginin, yani dönüştürülmesi gereken finansal oranların değerlendirilmesine yardımcı olabilir (Auria ve Moro, 2008).

3.3.3. C4.5 (Karar ağacı)

Makine öğrenmesi ve veri madenciliği topluluklarında en sık kullanılan algoritmalardan biridir. C4.5 ile birleştirilmiş düşük örnekleme, diğer algoritmaları değerlendirmek için yararlı bir başlangıç noktasıdır (Drummond ve Holte, 2003)

3.6'da entropi denklemi gösterilmektedir:

Entropi

$$H(X) = - \sum_{i=1}^{n} P(Xi) \log_p P(Xi) \qquad (3.6)$$

C4.5, hem kategorik hem de sayısal değeri ele alır. C4.5, ID3'ün bir evrimidir. C4.5 algoritması, bu verileri tekrar tekrar bölerek verilen veriler için bir karar ağacı oluşturur. C4.5 algoritması, verileri bölen olası tüm testleri göz önünde bulundurur ve en iyi bilgi kazanımını sağlayan bir test seçilecektir. Bu algoritma, ID3'ün geniş karar ağacının lehine eğilmesine ortadan kaldırmaktadır (Mohankumar ve ark., 2016). C4.5 algoritmasının avantajları ve dezavantajları vardır, avantajı: Kolayca yorumlanabilecek modeller oluşturur, uygulaması kolay, hem kategorik hem de sürekli değerleri

kullanabilir. Dezavantajı: Veriyi iyi bir şekilde açıklamayan aşırı karmaşık ağaçlar üretilebilir. Bu durumda ağaç dallanması takip edilemeyebilir.

3.4. Veri Kümeleri

Gerçek dünya problemlerinde birçok veri kümesi dengesiz olmaktadır, bu veri kümelerinin dengesiz olmaları sınıflandırma başarılarını olumsuz olarak etkilemektedir. Bu problemi ortadan kaldırmak için bu dengesiz veri kümelerini dengeli hale getirmek sınıflandırma başarısı açısından bize büyük avantaj sağlamaktadır. Bu tez çalışmasında, KEEL web tabanından 22 dengesiz veri seti kullanılmıştır. Kullanılan veri setlerinin düşük dengesizlik oranları aşağıdaki denklem 3.7'e göre hesaplanmıştır.

$$Düşük\ dengesizlik\ oranı = \frac{Çoğunluk\ sınıftaki\ örnek\ sayısı}{Azınlık\ sınıftaki\ örnek\ sayısı} \tag{3.7}$$

Denklem 3.7'den elde edilen sonuçlar 1.5 ile 9 değerleri arasında olmaktadır. Kullanılan veri setlerinin özelikleri Çizelge 3.1'de verilmiştir.

Çizelge 3.1. Veri setlerinin özelikleri

Veri setleri	Nitelik sayısı	Dengesizlik oranı	Örnek sayısı	Sınıf sayısı
Ecoli 1	7	3.36	336	2
Ecoli 2	7	5.46	336	2
Ecoli 3	7	8.6	336	2
Ecoli-0_vs_1	7	1.86	220	2
Glass 0	9	2.06	214	2
Glass 1	9	1.82	214	2
Glass 6	9	6.38	214	2
Glass-0-1-2 3_vs_4-5-6	9	3.2	214	2
Haberman	3	2.78	306	2
Iris0	4	2	150	2
New-thyroid 1	5	5.14	215	2
New-thyroid 2	5	5.14	215	2
Page-blocks 0	10	8.79	5472	2
Pima	8	1.87	768	2
Segment 0	19	6.02	2308	2
Vehicle 0	18	3.25	846	2

Vehicle 1	18	2.9	846	2
Vehicle 2	18	2.88	846	2
Vehicle 3	18	2.99	846	2
Wisconsin	9	1.86	683	2
Yeast 1	8	2.46	1484	2
Yeast 3	8	8.1	1484	2

Bu veri setlerine ek olarak Kyoto Üniversitesi internet trafik verilerinden alınan veriler de kullanılmıştır (Kyoto Dataset, 2006). Bu veriler saldırı tespit verileri olarak adlandırılır. Saldırı tespit sistemleri, araştırmacılar tarafından halen üzerinde çalışma yapılan önemli bir çalışma alanıdır ve bu güne kadar araştırmacılar tarafından birçok saldırı tespit sistemleri geliştirilmiştir. Bu sistemler veri madenciliği algoritmalarından olan sınıflandırma algoritmalarını kullanılmaktadırlar.

Saldırı tespit veri kümelerinde, veriler ağırlıklı olarak "normal" örneklerden ve küçük bir yüzdeyle" anormal" örneklerden oluşmakta ve bu sınıf dengesizliği problemlerine yol açmaktadır. Sınıf dengesizliği problemlerinde, eğitim verilerinden bir öğrenme modeli oluşturulduğunda genelde daha çok örneğe sahip olan sınıf daha iyi bir şekilde öğrenilebilir. Böyle bir problemi ortadan kaldırmak için az örneklere sahip olan sınıflar için örnek artırma (Oversampling) ve/veya çok örneğe sahip olan sınıflar için ise azaltma örnekleme (Under Sampling) veri dengeleme yöntemleri kullanılır. Bu tez çalışmasında, yukarıdaki verilere ek olarak dengesiz saldırı tespit verileri de ele alınarak bu veriler veri dengeleme yöntemleri ile dengelenmiştir. Saldırı tespit veri setinin orijinal halinde 233428 örnek bulunmaktadır. Veri kümesinin büyük olması nedeniyle örnekleme yöntemine başvurulmuştur. Bu tez çalışmasında orijinal veriden seçilen 6000 örnek kullanılmıştır. Bu örneklerden 4000 tanesi çoğunluk, 2000 tanesi azınlık olacak şekilde rastgele seçilmiştir. Saldırı tespit veri setinin özelikleri Çizelge 3.2'de verilmiştir.

Çizelge 3.2. Saldırı tespit veri setinin özelikleri

Veri Seti	Nitelik	Dengesiz oranı (IR)	Örneği
Saldırı tespit	18	2.06	6000 örnek

Saldırı tespit veri setinin nitelik isimleri ve açıklamaları Çizelge 3.3'te verilmiştir.

Çizelge 3.3. Saldırı tespit veri setinin nitelik isimleri ve açıklamaları

Nitelik isimleri	Açıklamaları
Süre	Bağlantının uzunluğu (saniye sayısı)
Servis	Bağlantının servis türü, örneğin, http, telnet, vb
Kaynak bayt	Kaynak IP adresi tarafından gönderilen veri baytlarının sayısı
Hedef bayt	Hedef IP adresi tarafından gönderilen veri baytlarının sayısı
Say	Kaynak IP adresi ve hedef IP adresi son iki saniye içinde geçerli bağlantıdakilerle aynı olan bağlantıların sayısı.
Aynı srv hızı	Count özelliğinde aynı hizmete bağlantıların% si
Serror hızı	Count özelliğinde "SYN" hatası olan bağlantıların% si
Srv terör oranı	Srv sayımında "SYN" hatası olan bağlantıların yüzdesi (hizmet tipi, son iki saniyedeki mevcut bağlantı ile aynı olan bağlantıların sayısı) özelliği
Bayrak	Özetin yazıldığı sırada bağlantının durumu (genellikle bağlantı sonlandırıldığında). Farklı durumlar aşağıdaki bölümde özetlenmiştir
IDS tespiti	IDS'nin (Intrusion Detection System) bağlantı için bir uyarı tetikleyip tetiklemediğini yansıtır; "0", herhangi bir uyarının tetiklenmediği ve bir Arap rakamı ('0' hariç), farklı uyarı türlerini ifade eder. Parantez, bağlantı sırasında gözlenen aynı uyarının numarasını gösterir.
Kötü amaçlı yazılım algılama	Kötü amaçlı yazılım olarak da bilinen kötü amaçlı yazılımın bağlantıda gözlenip gözlemlenmediğini belirtir; '0', kötü amaçlı yazılım gözlemlenmediği anlamına gelir ve bir dize, bağlantıda gözlenen ilgili kötü amaçlı yazılımları belirtir. Kötü amaçlı yazılımları tespit etmek için 'clamav' yazılımını kullandık. Parantez, bağlantı sırasında gözlenen aynı kötü amaçlı yazılımın sayısını gösterir.
Ashula tespiti	Özel yazılım kullanılarak bağlantıda kabuk kodlarının ve kullanım kodlarının kullanılıp kullanılmadığı anlamına gelir; "0", hiçbir barkod ve istismar kodunun gözlemlenmediği anlamına gelir ve bir Arap rakamı ('0' hariç) farklı türde barkodların veya kullanım kodlarının kullanılması anlamına gelir. Parantez, bağlantı sırasında gözlenen aynı kabuk kodunun veya istismar kodunun numarasını gösterir.
Etiket	Oturumun saldırı olup olmadığını gösterir; "1", oturumun normal olduğu, "-1", oturumda bilinen bir saldırı görüldüğü ve "-2", oturumda bilinmeyen bir saldırı görüldüğü anlamına gelir
Kaynak Port Numarası	Oturumda kullanılan kaynak port numarasını gösterir.
Hedef Port Numarası	Oturumda kullanılan hedef port numarasını gösterir
Start Time	Oturumun ne zaman başladığını gösterir
Süre	Oturumun ne kadar süre kurulduğunu gösterir
En son sınıfı temsil edin	

3.5. Karışıklık Matrisi

Bir Karışıklık matrisi, çoğu zaman gerçek değerlerin bilindiği bir dizi test verisi üzerinde bir sınıflandırma modelinin performansını tanımlamak için kullanılan bir tablodur. Bir algoritmanın performansının görselleştirilmesine izin verir. Bir sınıflandırma problemine ilişkin tahmin sonuçlarının bir özetidir (Data School, 2014)

Şekil 3.5'te Karışıklık matrisi göstermiştir.

		Tahmin Edilen		Toplam
		C^+	C^-	
Gerçek	C^+	DP Doğru Pozitif	YN Yanlış Negatif	Gerçek Pozitif Sayısı
	C^-	YP Yanlış Pozitif	DN Doğru Negatif	Gerçek Negatif Sayısı
Toplam		Tahmin Pozitif Sayısı	Tahmin Negatif Sayısı	Toplam örnek Sayısı

Şekil 3.5. Karışıklık Matrisi

Doğruluk, doğru sınıflandırılan pozitif örneklerin sayısına doğru negatif örneklerin sayısı ekilerek toplam örneklerin sayısına bölünerek bulunmuş. Doğruluk Denklem 3.8e göre hesaplanmaktadır.

$$Doğruluk = \frac{DoğruPozitif + Doğru\ Negatif}{Toplam\ örnek\ sayısı} \quad (3.8)$$

Duyarlılık, doğru sınıflandırılan pozitif örneklerin yani doğru pozitif örneklerin sayısının doğru pozitif ile yanlış negatif örneklerin toplamına oranıdır, duyarlılık aynı zamanda hassasiyeti ifade etmektedir. Duyarlılık Denklem 3.9'e göre hesaplanmaktadır.

$$Duyarlılık = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + \text{Yanlış Negatif}} \quad (3.9)$$

Kesinlik, doğru sınıflandırılan yani doğru pozitif örneklerin sayısının doğru pozitif ile yanlış pozitif örneklerin sayıları toplamına oranıdır. Kesinlik Denklem 3.10'e göre hesaplanmaktadır.

$$Kesinlik\ (Gerçek\ Olumlu) = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Pozitif} \quad (3.10)$$

Özgünlük, doğru sınıflandırılan negatif örneklerin sayısının doğru negatif ile yanlış pozitif örneklerin sayıları toplamına oranıdır. Özgünlük denklem 3.11'e göre hesaplanmaktadır.

$$\text{Özgünlük} = \frac{Doğru\ \text{Negatif}}{Doğru\ \text{Negatif} + Yanlış\ Pozitif} \quad (3.11)$$

F-Ölçüsü: F-Ölçüt hem yanlış pozitifleri hem de yanlış negatifleri hesaba katar. Ancak F1, özellikle düzensiz bir sınıf dağılımınız varsa, genellikle doğruluktan daha kullanışlıdır. F-Ölçüsü Denklem 3.12'e göre hesaplanmaktadır.

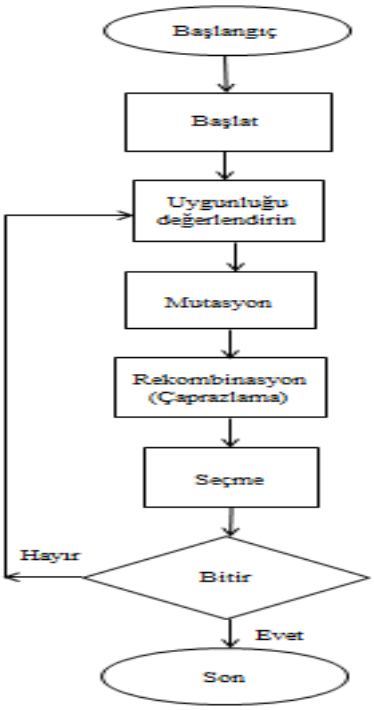
$$F - \text{Ölçüsü} = \frac{2 * \text{Kesinlik} * Duyarlılık}{\text{Kesinlik} + Duyarlılık} \quad (3.12)$$

3.6. Diferansiyel Evrim Algoritması

Diferansiyel Evrim (DE) Ken Price ve Rainer Storn tarafından önerilmiştir. DE'nin diferansiyel evrim olarak adlandırılmasının nedeni, doğal olarak matematiksel (geometrik) medyadan kaynaklanmaktadır. Büyük ve doğal olan karmaşık bir evrim sürecinin küçük ve basit matematiksel bir modeli olarak tanımlanır. Yani, kolay ve verimlidir. DE, sürekli problemlerde global optimizasyon amaçlı en başarılı evrimsel algoritmalardan biridir. DE diğer yöntemlerle karşılaştırıldığında basit yapı ve uygulama, düşük hesaplama karmaşıklığı, yüksek performans gibi birçok avantaja sahiptir. Diferansiyel mutasyon operatörünün yapısı ve geçişi, sürekli problemler için DE'yi daha uygun hale getirmiştir (Noghabi ve ark., 2015). Evrimsel algoritmalarda (EA) bulunan arama parametreleri iki sınıfa ayrılabilir: parametre ayarlama ve parametre kontrolü. Parametre ayarlama DE'nin parametre ayarı, parametre seçimine göre kullanıcı yükünü en aza indirir. Parametre kontrolü, çalışma sırasında değiştirilen başlangıç parametre değerleriyle bir çalışmaya başlama miktarını belirlediği için bir alternatiftir (Liu ve Lampinen, 2005).

DE'nin avantajı: popülasyon çokluğunu koruması ve yerel arama kapasitesini arttırmasıdır. DE'nin dezavantaj: yakınsama kararsızlığıdır (Wu ve ark., 2011).

Diferansiyel Evrimin akış diyagramı şekil 3.6'de gösterilmiştir.



Şekil 3.6. Diferansiyel Evrim Akış diyagramı

Şekil 3.6'de görüldüğü gibi diferansiyel evrim algoritması mutasyon, rekombinasyon ve seçme işlemlerinden oluşmaktadır. Bu işlemleri aşağıda detaylı olarak açıklanmıştır.

3.6.1. Mutasyon

Mutasyon DE'nin operatörlerinden biridir, bu operatörü geliştirmek için pek çok çaba yapılmaktadır. Aslında mutasyon, DE değişkenlerini ayırt etmenin anahtarıdır. DE'nin diferansiyel evrimi ifade ettiği DE / X / Y / Z gösterimi uygulandığında, X hedefi belirtir, Y, rastgele seçilen fark vektörlerinin sayısını belirtir ve Z, çaprazlama operatörünün tipini belirtir. Çaprazlama operatörünün ve Mutasyon operatörünün amacı, algoritmaya daha fazla çeşitlilik, keşif ve güvenilirlik sağlamaktır. Mutasyon operatöründe üç vektör rastgele seçilmiştir (Noghabi ve ark., 2015).

3.6.2. Seçme

Bir deneme vektörü üretmek için mutasyon işleminde hangi bireylerin yer alacağını belirlemek ve ebeveynlerden veya yavrulardan hangisinin bir sonraki nesle aktarılacağını belirlemek için seçim kullanılmıştır. Rastgele seçim genellikle fark

vektörlerinin hesaplandığı bireyleri seçmek için kullanılmıştır. DE uygulamalarının çoğu için, hedef vektör rastgele seçilir veya en iyi birey seçilir (Engelbrecht, 2007).

3.6.3. Çaprazlama (Rekombinasyon)

Çaprazlama operatörü DE algoritmasının dinamiklerinden biridir, mutasyon operatörü zaten bireylerin rekombinasyonuna dayandığından, çaprazlama rolü biraz farklıdır. Sadece mevcut elemanı ve mutasyon tarafından üretilenleri birleştirerek yeni bireylerin yapımına izin verir (Varun Kumar ve Panneerselvam, 2017).

Diferansiyel Evrimin beş farklı strateji ile kullanmıştır:

$Y = örneği\ orijinal\ hali$

$X = elde\ ettiğimiz\ crossover\ yapılacak\ hali$

$F_{weight} =$ ölçek $faktör$ü

$CR = geçme\ olasılğı$

$FM_{mui} = rand(1, D) < F_{CR}$

$FM_{mpo} = FM_{mui} < 0.5$

$X = cat(1, X, X_{temp})$

$F_{weight} = 0.5$

$F_{CR} = 0.7$

Strateji 1

$x_{syn} = x_{nearest(3,:)} + F_{weight} * (x_{nearest(1,:)} - x_{nearest(2,:)})$

$x_{syn} = y.* FM_{mpo} + x_{syn}.* FM_{mui}$

Strateji 2

$x_{syn} = y + F_{weight} * (MinorCenter - y) + F_{weight} * (x_{nearest(1,:)} - x_{nearest(2,:)})$

$x_{syn} = y.* FM_{mpo} + x_{syn}.* FM_{mui}$

Strateji 3

$x_{syn} = MinorCenter + (x_{nearest(1,:)} - x_{nearest(2,:)}) * ((1 - 0.9999) * rand(1, D) + F_{weight})$

$x_{syn} = y.* FM_{mpo} + x_{syn}.* FM_{mui}$

Strateji 4

$f1 = ((1 - F_{weight}) * rand + F_{weight})$

$x_{syn} = x_{nearest(3,:)} + (x_{nearest(1,:)} - x_{nearest(2,:)} * f1)$

$x_{syn} = y.* FM_{mpo} + x_{syn}.* FM_{mui}$

Strateji 5

$x_{syn} = x_{nearest(3,:)} + F_{weight} * (x_{nearest(1,:)} - x_{nearest(2,:)})$

$x_{syn} = x_{nearest(3,:)} + 0.5 * (F_{weight} + 1.0) * (x_{nearest(1,:)} + x_{nearest(2,:)}\ 2 * x_{nearest(3,:)})$

$x_{syn} = y.* FM_{mpo} + x_{syn}.* FM_{mui}$

3.7. Saldırı Tespit Sistemleri (IDS)

Günümüzde İnternet ağlarının artması ile birlikte, veri değişiminin güvenliği temel bir görev olarak görülmektedir. Bu nedenle güvenlik araçlarının kullanımı gün geçtikçe artmaktadır. Saldırı tespit sistemleri (STS) bu araçlar arasındadır. Yalnızca bir ağdan gelen bir mesajı "uyarı" olarak etiketleyebilirler, ancak sistem durumunu tanımlayamazlar.

Genel saldırganların amacı genel olarak maddi menfaat sağlama, politik, ticari ve ekonomik yönden rakiplerine karşı avantaj sağlama, sahip olamadığı ek kaynaklara sahip olma isteği, kurumsal veya ulusal çıkar elde etme isteği gibi, çok çeşitli konuları içerilmektedir.

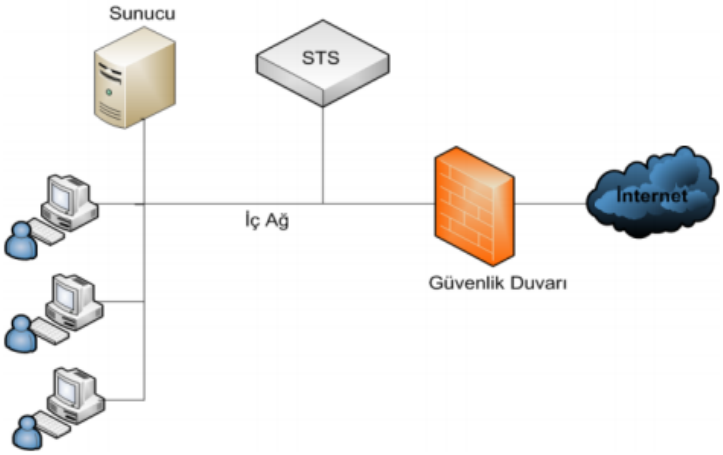
Güvenlik duvarları, saldırı tespit sistemleri vb. gibi farklı güvenlik araçları vardır. Bir ağdaki güvenlik düzeyini ilerletmek için kullanılırlar. Saldırı tespit sistemlerinin temel sorunlarından biri, düşük seviyeli soyutlama ile birçok uyarının verilmesidir. Dolayısıyla, izinsiz giriş tespit sisteminden alınan bağlı uyarılar için bir yöntem vardır, bu yöntem, uyarı sayısını azaltmaya ve ayrıca daha yüksek düzeyde uyarılar üretmeye çalışır. Tipik sistemlerde, güvenlik duvarları, virüsten koruma yazılımı ve Saldırı Tespit Sistemleri gibi araçlar, bir ağı korumaya ve olası saldırılara karşı korumaya çalışmaktadır. Bu araçlar, bilgisayar saldırılarının etkilerini azaltmak için uygun çözümlerdir, ancak kapsamlı bir koruma yöntemi olarak görülemezler. Son zamanlarda dikkat çeken araçlardan biri saldırı tespit sistemleridir (STS). Ağın ve güvenlik aygıtlarının genişletilmesi her zaman ağ öğelerinden gelen birçok uyarının ağ yöneticisine gönderilmesine neden olur. Bununla birlikte, bu alarmların sayısı çok yüksektir ve soyutlama seviyeleri o kadar düşüktür ki analizleri ağ yöneticileri için pratikte mümkün değildir. Ek olarak, bu uyarıların çoğu yanlış pozitiftir. Saldırı tespit sistemleri, ağın güvenlik durumunu ve analizlerini izlemek için mevcut yöntemlerden biridir. Saldırı tespit sistemleri, bilgisayar saldırılarının etkisini azaltmak için uygun bir çözümdür. Ancak, ağ için olası zararları korumak ve önlemek için kapsamlı bir çözüm olarak kabul edilemezler. Bunun birçok nedeni vardır, bunlar aşağıdaki örneklerde belirtilmiştir: Alınan uyarılar yüksek, yanlış uyarılar ve karmaşık saldırıları keşfetme yetersizliği bu nedenlerdir (Pourabbas, 2014).

İzinsiz giriş tespit sistemi işlemleri aşağıdaki gibidir:

- Hem kullanıcı hem de sistem işlemlerini izle ve analiz yapma
- Sistem yapılandırmasını ve güvenlik hassasiyetini analiz yapma

- Genel saldırı modellerini tanımak
- Anormal olayların etkilerini analiz edebilme
- Kullanıcı güvenlik politikası ihlallerini izleyin

Şekil 3.7'de Saldır Tespit Sistemi yerleşimi gösterilmiştir.



Şekil 3.7. Saldırı tespit sistemi

Bu çalışmada kullanılan veriler, İzinsiz Giriş Tespiti verileri, Kyoto Üniversitesi'nden Honeypots'tan alınmıştır.

4. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

Bu tez çalışmasında SMOTE ve DE stratejileri veri dengeleme algoritmaları kullanılarak KEEL web sitesinden 22 veri kümesi ve Kyoto Üniversitesi'nin Honeypots'larından veri saldırı tespiti veri kümeleri dengesiz durumundan dengeli bir hale gelmiştir. Bu veri kümeleri dengeli hale geldikten sonra K-NN, C4.5 ve SVM sınıflandırma algoritmaları ile sınıflandırılmıştır. Bu bölümde SMOTE ve DE stratejileri veri dengeleme algoritmalarından elde edilen sonuçlar analiz ederek tartışılmıştır.

SMOTE ve DE stratejileri veri dengeleme algoritmaların kodları Matlab programlama dili kullanılarak kodlanmıştır. SMOTE ve DE stratejileri veri dengeleme algoritmaları için veri kümeleri üzerinde uygulandıktan sonra 10 deneme sonucunda sınıflandırma algoritmalarında sınıflandırma başarıların geometrik ortalaması (GO) ve standart sapması hesaplanmıştır. Geometrik ortalaması aşağıdaki denklem 4.1'e göre elde edilmektedir.

$$GO = \left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}} \tag{4.1}$$

Denklem 4.1'de N sınıflandırma sonuçlarının sayılarını ve x_i sınıflandırma sonucunu temsil etmektedir. Sınıflandırma sonuçlarının en iyi sonucu almaktan ise tüm sonuçların ortalaması alınmasında değerlendirme açısından fayda vardır. Sınıflandırma geometrisinin yanı sıra sınıflandırma başarılarının standart sapma değerleri aşağıdaki denklem 4.2'e göre hesaplanmıştır.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4.2}$$

Denklem 4.2'de σ standart sapma değerini ifade etmektedir, N sonuçların sayısı x_i sonuçların içinde x. sonuç, $\bar{x}$ sonuçların aritmetik ortalaması. Standart sapma değeri elde edilen sonuçların ortalamaya ne kadar yakın olduğunu gösterir.

22 dengesiz veri kümelerine SOMTE ve DE stratejileri veri dengeleme algoritmaları uygulandıktan sonra her veri kümesinin K-NN, C4.5 ve SVM sınıflandırma algoritmaları ile sınıflandırma başarıları elde edilmiştir, K-NN sınıflandırma algoritması için elde edilen sınıflandırma başarılarının geometrik ortalama ve standart sapma değerleri aşağıda Çizelge 4.1'de verilmiştir.

Çizelge 4.1. K-NN sınıflandırma algoritmasının sınıflandırma başarıları

	Orjinal	SMOTE		Strateji1		Strateji2		Strateji3		Strateji4		Strateji5	
Veriler	GO	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma
Ecoli 1	0.7920	**0.8554**	0.0122	0.8505	0.0129	0.8408	0.0094	0.8124	0.0126	0.8553	**0.0051**	0.8248	0.0064
Ecoli 2	0.9143	**0.9166**	0.0091	0.9122	0.0107	0.9142	0.0049	0.9099	0.0063	0.9135	0.0086	0.9041	**0.0020**
Ecoli 3	0.7171	0.8547	0.0178	0.8225	0.0183	0.8372	0.0212	0.8072	0.0184	**0.8614**	0.0151	0.8029	**0.0134**
Ecoli 0_vs_1	0.9615	0.9631	0.0038	0.9613	0.0065	0.9597	**0.0019**	0.9634	0.0037	0.9596	0.0063	**0.9650**	0.0033
Glass 0	0.7770	0.7883	0.0100	**0.7983**	0.0112	0.7866	0.0078	0.7749	**0.0049**	0.7859	0.0086	0.7960	0.0102
Glass 1	0.7984	0.8053	0.0117	0.7951	0.0124	0.7896	0.0095	0.7948	**0.0066**	**0.8125**	0.0202	0.8122	0.0110
Glass 6	0.9207	**0.9325**	0.0146	0.9228	0.0089	0.9323	0.0138	0.9245	**0.0055**	0.9298	0.0111	0.9240	0.0084
Glass 0-1-2-3_vs_4-5-6	0.9372	**0.9471**	0.0074	0.9426	0.0069	0.9457	0.0077	0.9418	0.0041	0.9441	0.0045	0.9402	**0.005**
Haberman	0.5225	0.5494	0.0210	0.5528	0.0173	0.5345	0.0196	0.5463	0.0196	**0.5600**	**0.0117**	0.5361	0.0204
Iris	1	1	0	1	0	1	0	1	0	1	0	1	0
New thyroid 1	0.9826	0.9815	0.0077	**0.9958**	**0.0014**	0.9899	0.0046	0.9858	0.0073	0.9802	0.0078	0.9944	0.0019
New thyroid 2	0.9852	0.9902	0.0038	0.9972	0.0020	0.9972	0.0020	0.9859	0.0041	0.9916	0.0040	**0.9986**	**0.0014**
Page block 0	0.8680	0.8967	0.0047	0.8691	0.0018	0.8649	**0.0014**	**0.8995**	0.0030	0.8987	0.0016	0.8754	0.0023
Pima	0.6314	0.6602	0.0062	0.6551	0.0066	0.6472	0.0075	**0.6747**	0.0082	0.6621	0.0109	0.6527	**0.0053**
Segment 0	0.9898	0.9911	0.0012	**0.9913**	0.0008	0.9892	0.0007	0.9911	0.0017	0.9906	0.0011	0.9897	**0.0001**
Vehicle 0	0.9111	0.9270	0.0065	0.9288	0.0071	0.9128	0.0041	0.9123	**0.0017**	**0.9297**	0.0049	0.9236	0.0050
Vehicle 1	0.5438	0.5925	0.0134	0.5928	0.0097	0.5545	0.0047	0.5466	**0.0039**	**0.6096**	0.0110	0.5895	0.0120
Vehicle 2	0.9165	**0.9318**	0.0057	0.9156	0.0045	0.9201	0.0038	0.9153	**0.0017**	0.9210	0.0049	0.9235	0.0028
Vehicle 3	0.5691	**0.6248**	0.0150	0.6040	0.0063	0.5828	0.0046	0.5734	**0.0041**	0.6188	0.0099	0.5971	0.0100
Wisconsin	0.9549	0.9578	0.0043	0.9595	0.0038	0.9611	**0.0015**	0.9595	0.0025	**0.9620**	0.0027	0.9606	0.0042
Yeast 1	0.6137	0.6547	0.0079	**0.6725**	**0.0046**	0.6598	0.0082	0.6315	0.0048	0.6528	0.0055	0.6310	0.0048
Yeast 3	0.7795	**0.8695**	0.0086	0.8644	0.0082	0.8556	0.0081	0.8203	0.0086	0.8643	0.0088	0.8370	**0.0043**

Çizelge 4.1'de sınıflandırma sonuçlarının geometrik ortalama ve standart sapma değerleri analiz edildiğinde, geometrik ortalama sonuçlarında Ecoli 1, Ecoli 2 ve Glass 6 veri kümelerinde sırasıyla 0.8554, 0.9166 ve 0.9325 değerleri ile SMOTE veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir. DE Strateji1 veri dengeleme yönteminde Glass 0, New thyroid 1, Segment 0 ve Yeast 1 veri kümeleri sırasıyla 0.7983, 0.9958, 0.9913 ve 0.6725 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. DE Strateji3 veri dengeleme yönteminde Page block 0 ve Pima veri kümeleri sırasıyla 0.8995 ve 0.6747 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. DE Strateji4 veri dengeleme yönteminde Ecoli 3, Glass 1, Haberman, Vehicle 0, Vehicle 1 ve Wisconsin veri kümeleri sırasıyla 0.8614, 0.8125, 0.5600, 0.9297, 0.6096 ve 0.9620 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. DE Strateji5 veri dengeleme yönteminde Ecoli 0_vs_1 ve New thyroid 2 veri kümeleri sırasıyla 0.9650 ve 0.9986 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir.

Kullanılan veri kümelerinin sınıflandırma başarıların standart sapmalarına bakıldığında New thyroid 1 ve Yeast 1 veri kümelerinde sırasıyla 0.0014 ve 0.0046 değerleri ile Strateji1 veri dengeleme yönteminde en iyi standart sapma başarısı kaydetmiştir. DE Strateji2 veri dengeleme yönteminde Ecoli 0_vs_1, Page block 0 ve Wisconsin veri

kümeleri sırasıyla 0.0019, 0.0014 ve 0.0015 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi standart sapma değerleri elde etmiştir. DE Strateji3 veri dengeleme yönteminde Glass 0, Glass 1, Glass 6, Vehicle 0, Vehicle 1, Vehicle 2 ve Vehicle 3 veri kümeleri sırasıyla 0.0049, 0.0066, 0.0055, 0.0017, 0.0039, 0.0017 ve 0.0041 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. DE Strateji4 veri dengeleme yönteminde Ecoli 1ve Haberman veri kümeleri sırasıyla 0.0051 ve 0.0117 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. DE Strateji5 veri dengeleme yönteminde Ecoli 2, Ecoli 3, Glass 0-1-2-3_vs_4-5-6, New thyroid 2, Pima, Segment ve Yeast 3 veri kümeleri sırasıyla 0.0020, 0.0134, 0.005, 0.0014, 0.0053, 0.0001 ve 0.0043 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir.

Sonuç olarak, K-NN sınıflandırma algoritmasında SMOTE veri dengeleme yöntemi toplam 22 veri kümesinden 7 veri kümesinde daha iyi geometrik ortalama değerleri elde etmiştir. Strateji3 ve Strateji5 veri dengeleme yöntemi toplam 22 veri kümesinden 7 veri kümesinde daha iyi standart sapma değerleri elde etmiştir, bu da sınıflandırma sonuçlarının birbirlerine yakın sonucu elde ettiğini göstermektedir.

22 veri kümesi SMOTE ve DE stratejileri ile dengelendikten sonra C4.5 sınıflandırma algoritmasından elde edilen sınıflandırma sonuçlarının geometrik ortalama ve standart sapma değerleri Çizelge 4.2'de sunulmuştur.

Çizelge 4.2. C4.5 sınıflandırma algoritmasının sınıflandırma başarıları

	Orjinal	SMOTE		Strateji1		Strateji2		Strateji3		Strateji4		Strateji5	
Veriler	GO	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma
Ecoli 1	0.8381	**0.8567**	0.0227	0.8435	0.0143	0.8315	0.0151	0.8407	**0.0106**	0.8501	0.0131	0.8566	0.0122
Ecoli 2	0.7882	0.8631	0.0198	**0.8758**	**0.0154**	0.8418	0.0167	0.8477	0.0113	0.8648	0.0200	0.8337	0.0181
Ecoli 3	0.6713	0.7787	0.0229	0.7438	0.0262	0.7394	0.0304	0.7396	0.0308	**0.7791**	**0.0176**	0.7053	0.0437
Ecoli 0_vs_1	0.9724	0.9722	0.0067	0.9723	0.0047	**0.9741**	**0.0042**	0.9738	0.0055	0.9684	0.0089	0.9648	0.0046
Glass 0	0.7731	0.7696	**0.0131**	0.7801	0.0234	0.7849	0.0199	**0.7974**	0.0191	0.7714	0.0299	0.7506	0.0254
Glass 1	0.6402	0.7380	**0.0240**	0.7179	0.0249	0.6859	0.0262	0.6990	0.0293	**0.7516**	0.0255	0.7432	0.0277
Glass 6	0.8342	0.8760	0.0292	0.8813	0.0307	**0.9119**	0.0247	0.8455	0.0271	0.8657	0.0185	0.8757	**0.0181**
Glass 0-1-2-3_vs_4-5-6	0.8835	0.8886	0.0172	0.8971	0.0175	0.9078	0.0219	0.8932	0.0187	**0.9079**	**0.0131**	0.8771	0.0171
Haberman	0.4539	0.5377	0.0223	0.5239	0.0241	0.5219	0.0199	**0.5704**	0.0395	0.5515	**0.0193**	0.5316	0.0295
Iris	1	1	0	1	0	1	0	1	0	1	0	1	0
New thyroid 1	0.9042	**0.9511**	**0.0101**	0.9295	0.0120	0.9141	0.0109	0.9353	0.0142	0.9510	0.0108	0.9465	0.0172
New thyroid 2	0.9303	**0.9461**	0.0104	0.9233	0.0147	0.9292	0.0127	0.9393	**0.0091**	0.9448	0.0129	0.9443	0.0184
Page block 0	0.9085	**0.9415**	0.0038	0.9174	0.0052	0.9165	0.0066	0.9377	0.0031	0.9400	0.0055	0.9197	**0.0030**
Pima	0.6741	0.6929	0.0195	0.6864	0.0140	0.6877	**0.0139**	**0.6972**	0.0159	0.6968	0.0140	0.6488	0.0212
Segment 0	0.9774	0.9930	0.0012	0.9874	0.0015	0.9921	**0.0010**	0.9913	0.0022	**0.9934**	0.0014	0.9922	0.0011
Vehicle 0	0.9304	**0.9324**	**0.0076**	0.9278	0.0077	0.9080	0.0102	0.9211	0.0101	0.9295	0.0093	0.9062	0.0156
Vehicle 1	0.6334	0.6833	0.0173	0.6721	0.0193	0.6565	**0.0130**	0.6731	0.0210	**0.6855**	0.0146	0.6718	0.0229
Vehicle 2	0.9405	0.9517	**0.0045**	0.9426	0.0077	0.9531	0.0070	0.9339	0.0068	**0.9550**	0.0071	0.9383	0.0077

Vehicle 3	0.6220	0.6906	0.0152	0.6828	0.0195	0.6843	**0.0144**	0.6788	0.0219	**0.7178**	0.0216	0.6854	0.0250
Wisconsin	0.9416	**0.9559**	0.0056	0.9506	0.0061	0.9534	0.0051	0.9517	0.0076	0.9557	0.0064	0.9485	**0.0040**
Yeast 1	0.6461	0.6600	0.0134	0.6513	0.0135	0.6398	**0.0090**	0.6317	0.0147	**0.6629**	0.0114	0.6467	0.0130
Yeast 3	0.8244	0.8916	0.0094	0.8673	**0.0067**	0.8663	0.0101	0.8503	0.0074	**0.8977**	0.0094	0.8708	0.0158

Çizelge 4.2'de sınıflandırma sonuçlarının geometrik ortalama ve standart sapma değerleri analiz edildiğinde, geometrik ortalama sonuçlarında Ecoli 1, New thyroid 1, New thyroid 2, Page block 0 ve Wisconsin veri kümelerinde sırasıyla 0.8567, 0.9511, 0.9461, 0.9415, 0.9324 ve 0.9559 değerleri ile SMOTE veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir. DE Strateji1 veri dengeleme yönteminde Ecoli 2 veri kümelerinde 0.8758 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. DE Strateji2 veri dengeleme yönteminde Ecoli 0_vs_1 ve Glass 6 veri kümelerinde sırasıyla 0.9741 ve 0.9119 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. Glass 0, Haberman ve Pima veri kümelerinde sırasıyla 0.7974, 0.5704, ve 0.6972 değerleri ile DE Strateji3 veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir. DE Strateji4 veri dengeleme yönteminde Ecoli 3, Glass 1, Glass 0-1-2-3_vs_4-5-6, Segment 0, Vehicle 1, Vehicle 2, Vehicle 3, Yeast 1 ve Yeast 3 veri kümelerinde sırasıyla 0.7791, 0.7516, 0.9079, 0.9934, 0.6855, 0.9550, 0.7178, 0.6629 ve 0.8977 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir.

Kullanılan veri kümelerinin sınıflandırma başarıların standart sapmalarına bakıldığında Glass 0, Glass 1, New thyroid 1, Vehicle 0 ve Vehicle 2 veri kümelerinde sırasıyla 0.0131, 0.0240, 0.0101, 0.0076 ve 0.0045 değerleri ile SMOTE veri dengeleme yönteminde en iyi standart sapma başarısı kaydetmiştir. Ecoli 2 ve Yeast 3 veri kümelerinde sırasıyla 0.0154 ve 0.0067 değerleri ile Strateji1 veri dengeleme yönteminde en iyi standart sapma başarısı kaydetmiştir. DE Strateji2 veri dengeleme yönteminde Ecoli 0_vs_1, Pima, Segment 0, Vehicle 1, Vehicle 3 ve Yeast 1 veri kümelerinde sırasıyla 0.0042, 0.0139, 0.0010, 0.0130, 0.0144 ve 0.0090 veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir. Ecoli 1 ve New thyroid 2 veri kümelerinde sırasıyla 0.0106 ve 0.0091 değerleri ile DE Strateji3 veri dengeleme yönteminde en iyi standart sapma başarısı kaydetmiştir. . DE Strateji4 veri dengeleme yönteminde Ecoli 3, Glass 0-1-2-3_vs_4-5-6 ve Haberman veri kümelerinde sırasıyla 0.0176, 0.0131 ve 0.0193 veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir. DE Strateji5 veri dengeleme yönteminde Glass 6 ve Page block 0 veri

kümelerinde sırasıyla 0.0181 ve 0.0030 veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir.

Sonuç olarak, C4.5 sınıflandırma algoritmasında Strateji4 veri dengeleme yöntemi toplam 22 veri kümesinden 9 veri kümesinde daha iyi geometrik ortalama değerleri elde etmiştir. Strateji2 veri dengeleme yöntemi toplam 22 veri kümesinden 6 veri kümesinde daha iyi standart sapma değerleri elde etmiştir, bu da sınıflandırma sonuçlarının birbirlerine yakın sonucu elde edilmesini göstermektedir.

22 veri kümesi SMOTE ve DE stratejileri ile dengelendikten sonra SVM sınıflandırma algoritmasından elde edilen sınıflandırma sonuçlarının geometrik ortalama ve standart sapma değerleri Çizelge 4.3'te sunulmuştur.

Çizelge 4.3. SVM sınıflandırma algoritmasının sınıflandırma başarıları

	Orjinal	SMOTE		Strateji1		Strateji2		Strateji3		Strateji4		Strateji5	
Veriler	GO	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma	GO	St Sapma
Ecoli 1	0.8381	0.8990	0.0065	0.8988	0.0066	0.8949	0.0052	0.8727	0.0068	0.8988	0.0046	**0.8991**	**0.0027**
Ecoli 2	0.7882	0.9063	0.0056	0.9059	0.0026	0.9019	0.0055	0.8993	0.0062	0.9045	0.0062	**0.9076**	**0.0022**
Ecoli 3	0.6713	0.8892	0.0046	0.8907	0.0059	**0.8935**	0.0073	0.8853	0.0120	0.8919	0.0032	0.8811	**0.0019**
Ecoli 0_vs_1	0.9724	0.9742	0.0029	0.9725	0.0045	0.9696	0.0033	**0.9762**	0.0072	0.9743	0.0015	0.9760	**0.0020**
Glass 0	**0.7731**	0.7310	0.0112	0.7410	0.0120	0.7294	0.0105	0.6994	0.0174	0.7307	0.0093	0.7390	**0.0079**
Glass 1	**0.6402**	0.4934	**0.0089**	0.4802	0.0108	0.4829	0.0095	0.1687	0.0499	0.4881	0.0104	0.4861	0.0103
Glass 6	0.8342	0.9192	0.0095	0.9254	**0.0025**	**0.9295**	0.0037	0.9261	0.0085	0.9237	0.0050	**0.9213**	0.0066
Glass 0-1-2-3_vs_4-5-6	0.8941	0.9048	0.0089	0.9109	0.0084	0.9109	0.0079	0.8915	0.0096	0.9088	0.0072	0.9123	**0.0057**
Haberman	0.4539	0.5776	0.0096	0.6218	0.0040	**0.6461**	**0.0032**	0.5407	0.0128	0.5734	0.0059	0.5713	0.0056
Iris	1	1	0	1	0	1	0	1	0	1	0	1	0
New thyroid 1	0.9042	**0.9944**	0.0020	0.9883	0.0084	0.9898	0.0072	0.9855	0.0058	0.9930	0.0029	0.9824	**0**
New thyroid 2	0.9303	**0.9944**	0.0023	0.9898	0.0072	0.9898	0.0078	0.9826	0.0083	**0.9944**	0.0021	0.9824	**0**
Page block 0	**0.9085**	0.1836	**0.0415**	0.1901	0.0622	0.1755	0.0720	0.1066	0.0341	0.1340	0.0927	0.3713	0.0576
Pima	0.6741	0.7379	0.0110	0.7412	0.0090	0.7396	0.0062	0.7364	0.0079	0.7362	**0.0049**	**0.7438**	0.0099
Segment 0	0.9774	0.9935	0.0011	0.9930	0.0007	0.9923	**0.0003**	0.9929	0.0006	0.9933	0.0008	**0.9964**	0.0006
Vehicle 0	0.9304	0.9630	**0.0019**	0.9611	0.0034	0.9604	0.0033	0.9470	0.0070	**0.9635**	0.0038	0.7259	0.0037
Vehicle 1	0.6334	0.7697	0.0143	0.6907	0.0289	0.6460	0.0252	0.5283	0.0504	0.7749	0.0133	**0.8014**	**0.0079**
Vehicle 2	0.9405	0.9489	0.0034	0.9460	0.0050	0.8849	0.0244	0.9394	0.0071	0.9493	0.0046	**0.9559**	**0.0011**
Vehicle 3	0.6220	0.7170	0.0339	0.6485	0.0433	0.6505	0.0414	0.4494	0.0524	0.7198	0.0248	**0.7614**	**0.0113**
Wisconsin	0.9416	0.9731	0.0032	0.9696	0.0022	0.9633	0.0030	0.9715	**0.0018**	**0.9732**	0.0023	0.9688	0.0021
Yeast 1	0.6461	0.7064	0.0047	0.7164	**0.0019**	**0.7178**	0.0025	0.6207	0.0043	0.7045	0.0028	0.7054	0.0022
Yeast 3	0.8244	0.8953	0.0048	0.8886	0.0023	0.8925	0.0040	0.8699	0.0039	0.8969	0.0030	**0.8973**	**0.0013**

Çizelge 4.3'de sınıflandırma sonuçlarının geometrik ortalama ve standart sapma değerleri analiz edildiğinde, geometrik ortalama sonuçlarında Glass 0, Glass 1 ve Page block 0 veri kümelerinde sırasıyla 0.7731, 0.6402 ve 0.9085 değerleri ile orijinal veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir. SMOTE veri dengeleme yönteminde New thyroid 1 ve New thyroid 2 veri kümelerinde sırasıyla

0.9944 ve 0.9944 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. DE Strateji2 veri dengeleme yönteminde Ecoli 3, Glass 6, Haberman ve Yeast1 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. DE Strateji4 veri dengeleme yönteminde New thyroid 2, Vehicle 0 ve Wisconsin veri kümelerinde sırasıyla 0.9944, 0.9635 ve 0.9732 9119 değerleri ile diğer veri dengeleme yöntemlerinden daha iyi başarı elde etmiştir. Ecoli 1, Ecoli 2, Glass 6, Pima, Segment 0, Vehicle 1, Vehicle 2, Vehicle 3 ve Yeast 3 veri kümelerinde sırasıyla 0.8991, 0.9076, 0.9213, 0.7438, 0.9964, 0.8014, 0.9559, 0.7614 ve 0.8973 değerleri ile DE Strateji5 veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir.

Kullanılan veri kümelerinin sınıflandırma başarıların standart sapmalarına bakıldığında Glass 1, Page block 0 ve Vehicle 0 veri kümelerinde sırasıyla 0.0089, 0.0415 ve 0.0019 değerleri ile SMOTE veri dengeleme yönteminde en iyi standart sapma başarısı kaydetmiştir. DE Strateji4 veri dengeleme yönteminde Glass 6 ve Yeast 1 veri kümelerinde sırasıyla 0.0025 ve 0.0019 veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir. Haberman ve Segment 0 veri kümelerinde sırasıyla 0.0032 ve 0.0003 değerleri ile DE Strateji2 veri dengeleme yönteminde en iyi standart sapma başarısı kaydetmiştir. DE Strateji3 veri dengeleme yönteminde Wisconsin veri kümelerinde 0.0018 veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir. DE Strateji5 veri dengeleme yönteminde Wisconsin veri kümelerinde Ecoli 1, Ecoli 2, Ecoli 3, Ecoli 0_vs_1, Glass 0, Glass 0-1-2-3_vs_4-5-6, New thyroid 1, New thyroid 2, Vehicle 0 veri kümelerinde sırasıyla 0.0027, 0.0022, 0.0019, 0.0020, 0.0079, 0.0057, 0, 0, 0.0079, 0.0011, 0.0113 ve 0.0013 0019 veri dengeleme yönteminde en iyi sınıflandırma başarısı elde etmiştir.

Sonuç olarak, SVM sınıflandırma algoritmasında Strateji5 veri dengeleme yöntemi toplam 22 veri kümesinden 9 veri kümesinde daha iyi geometrik ortalama değerleri elde etmiştir. Strateji5 veri dengeleme yöntemi toplam 22 veri kümesinden 12 veri kümesinde daha iyi standart sapma değerleri elde etmiştir, bu da sınıflandırma sonuçlarının birbirlerine yakın sonucu elde edilmesini göstermektedir.

Sınıflandırma başarılarının geometrik ve standart sapma testlerinin yanı sıra sınıflandırma başarılarına göre algoritmalarının sıralama ortalaması hesaplanmıştır, Çizelge 4.4'de SMOTE ve DE stratejilerinin sıralama ortalaması sonuçları verilmiştir.

Çizelge 4.4. Sıralama ortalaması

	Orijinal	SMOTE	Strateji1	Strateji2	Strateji3	Strateji4	Strateji5
k-NN	1.9130	**5.5217**	4.7826	3.6087	3.1739	5.0000	4.0000
C4.5	1.8696	5.3043	4.1304	3.8696	3.7826	**5.8696**	3.1739
SVM	2.4348	**5.0870**	4.6522	3.5652	2.3043	4.9130	5.0435

Çizelge 4.4'de sıralama ortalamasına bakıldığında K-NN sınıflandırma algoritması 5.5217 değeri ile SMOTE veri dengeleme yöntemi iyi başarı elde etmiştir. C4.5 sınıflandırma algoritması 5.8696 değeri ile stratej4 veri dengeleme yöntemi iyi başarı elde etmiştir. SVM sınıflandırma algoritması 5.0870 değeri ile SMOTE veri dengeleme yöntemi iyi başarı elde etmiştir.

Yukarıdaki verilere ek olarak Kyoto veri kümesi de SMOTE ve DE stratejileri ile dengelenip daha sonra K-NN, C4.5 ve SVM sınıflandırma algoritmaları ile sınıflandırma başarı ölçütleri karışıklık matrisi üzerinden hesaplanmıştır. K-NN, C4.5 ve SVM sınıflandırma algoritmalarının Karışıklık matrisleri sırasıyla Çizelge 4.5, 4.6 ve 4.7'de verilmiştir.

Çizelge 4.5. K-NN sınıflandırma algoritmasının karışıklık matrisi

Sınıflandırma algoritması (K-NN)	Karışıklık Matrisi
Orijinal	571 229 154 246
SMOTE	509 291 140 260
Strateji 1	540 260 148 252
Strateji 2	568 232 156 244
Strateji 3	511 289 146 254
Strateji 4	509 291 140 260
Strateji 5	**595 205** **165 235**

Çizelge 4.5'te Karışıklık matrisi incelendiğinde K-NN sınıflandırma algoritması doğru pozitif ve doğru negatif veri kümelerinde 595 ve 235 değeri ile strateji5 veri dengeleme yöntemi iyi başarı elde etmiştir.

Sonraki Çizelge 4.6'de C4.5 sınıflandırma algoritmaları ile sınıflandırma başarı ölçütleri Karışıklık matrisleri gösterilmiştir.

Çizelge 4.6. C4.5 sınıflandırma algoritmasının karışıklık matrisi

Sınıflandırma algoritması (C4.5)	Karışıklık Matrisi
Orijinal	554 246 224 156
SMOTE	485 315 224 176
Strateji 1	**575 225** **213 187**
Strateji 2	551 249 244 156
Strateji 3	486 314 221 179
Strateji 4	498 302 226 174
Strateji 5	509 290 243 157

Çizelge 4.6'de Karışıklık matrisi incelendiğinde C4.5 sınıflandırma algoritması doğru pozitif ve doğru negatif veri kümelerinde 575 ve 187 değeri ile strateji 1 veri dengeleme yöntemi iyi başarı elde etmiştir.

Çizelge 4.7'de SVM sınıflandırma algoritmaları ile sınıflandırma başarı ölçütleri Karışıklık matrisleri gösterilmiştir.

Çizelge 4.7. SVM sınıflandırma algoritmasının karışıklık matrisi

Sınıflandırma algoritması (SVM)	Karışıklık Matrisi
Orijinal	**575 225** **213 187**
SMOTE	420 380 205 195
Strateji 1	354 446 178 222
Strateji 2	467 333 245 155
Strateji 3	388 412 207 193
Strateji 4	379 421 196 204
Strateji 5	450 350 210 190

Çizelge 4.7'de Karışıklık matrisi incelendiğinde SVM sınıflandırma algoritması doğru pozitif ve doğru negatif veri kümelerinde 575 ve 187 değeri ile orijinal veri dengeleme yöntemi iyi başarı elde etmiştir.

Karışıklık matrisleri üzerinden doğruluk, duyarlılık, özgünlük, kesinlik, hassasiyet ve F-ölçüsünün sonuçlarını hesaplanmıştır. K-NN sınıflandırma algoritmasının doğruluk, duyarlılık, özgünlük, kesinlik, ve F-ölçüsünün sonuçları Çizelge 4.8'de verilmiştir.

Çizelge 4.8. K-NN sınıflandırma algoritmasının sınıflandırma başarı ölçütleri

K-NN	Doğruluk	Duyarlılık	Özgünlük	Kesinlik	Hassasiyet (Recall)	F-ölçüsü
Orijinal	0.6808	0.7141	0.6146	**0.7875**	0.7141	0.7489
SMOTE	0.6403	0.6361	0.6492	0.7839	0.6361	0.7022
Strateji 1	0.6595	0.6746	0.6295	0.7846	0.6746	0.7254
Strateji 2	0.6767	0.7104	0.6095	0.7845	0.7104	0.7455
Strateji 3	0.6375	0.6384	0.6361	0.7782	0.6384	0.7013
Strateji 4	0.6408	0.6366	**0.6497**	0.7842	0.6366	0.7026
Strateji 5	**0.6918**	**0.7441**	0.5875	0.7830	**0.7441**	**0.7630**

Çizelge 4.8'de sonuçlara göre K-NN sınıflandırma algoritmasını doğruluk başarı ölçüsü 0.6918 değeri ile strateji 5 veri dengeleme yöntemi iyi başarı elde etmiştir. Duyarlılık başarı ölçüsü 0.744 değeri ile strateji 5 veri dengeleme yöntemi iyi başarı elde etmiştir. Bu tezde, Özgünlük başarı bizim için daha önemledir çünkü saldırıyı bulamalı önemli ve bu çizelgede özgünlük başarı ölçüsü 0.6497 değeri ile strateji 4 veri dengeleme yöntemi iyi başarı elde etmiştir. Kesinlik başarı ölçüsü 0.7875 değeri ile orijinal veri dengeleme yöntemi iyi başarı elde etmiştir. Hassasiyet başarı ölçüsü 0.7441 değeri ile strateji 5 veri dengeleme yöntemi iyi başarı elde etmiştir. F-ölçüsü başarı ölçüsü 0.7630 değeri ile strateji 5 veri dengeleme yöntemi iyi başarı elde etmiştir.

C4.5 sınıflandırma algoritmasının doğruluk, duyarlılık özgüllük, hassasiyet, kesinlik ve F-ölçüsünün sonuçları Çizelge 4.9'de verilmiştir.

Çizelge 4.9. C4.5 sınıflandırma algoritmasının sınıflandırma başarı ölçütleri

C4.5	Doğruluk	Duyarlılık	Özgünlük	Kesinlik	Hassasiyet (Recall)	F-ölçüsü
Orijinal	0.5912	0.6920	0.3892	0.6940	0.6920	0.6929
SMOTE	0.5510	0.6058	0.4410	0.6844	0.6058	0.6427
Strateji 1	**0.6352**	**0.7184**	0.4680	**0.7297**	**0.7235**	**0.7241**
Strateji 2	0.5892	0.6894	0.3890	0.6930	0.6894	0.6910
Strateji 3	0.5542	0.6070	**0.4480**	0.6878	0.6070	0.6448
Strateji 4	0.5598	0.6220	0.4355	0.6879	0.6220	0.6532
Strateji 5	0.5553	0.6372	0.3911	0.6768	0.5918	0.6564

Çizelge 4.9'de sonuçlara göre C4.5 sınıflandırma algoritmasını doğruluk başarı ölçüsü 0.6352 değeri ile strateji 1 veri dengeleme yöntemi iyi başarı elde etmiştir. Duyarlılık başarı ölçüsü 0.7184 değeri ile strateji 1 veri dengeleme yöntemi iyi başarı elde etmiştir. Bu tezde, Özgünlük başarı bizim için daha önemledir çünkü saldırıyı bulamalı önemli ve bu çizelgede özgünlük başarı ölçüsü 0.4480 değeri ile strateji 3 veri dengeleme yöntemi iyi başarı elde etmiştir. Kesinlik başarı ölçüsü 0.7297 değeri ile strateji 1 veri dengeleme yöntemi iyi başarı elde etmiştir. Hassasiyet başarı ölçüsü 0.7235 değeri ile strateji 1 veri dengeleme yöntemi iyi başarı elde etmiştir. F-ölçüsü başarı ölçüsü 0.7241 değeri ile strateji 1 veri dengeleme yöntemi iyi başarı elde etmiştir.

SVM sınıflandırma algoritmasının doğruluk, duyarlılık, özgüllük, hassasiyet, kesinlik ve F-ölçüsünün sonuçları Çizelge 4.10'da verilmiştir.

Çizelge 4.10. SVM sınıflandırma algoritmasının sınıflandırma başarı ölçütleri

SVM	Doğruluk	Duyarlılık	Özgünlük	Kesinlik	Hassasiyet (Recall)	F-ölçüsü
Orijinal	**0.6352**	**0.7187**	0.4680	**0.7297**	**0.7187**	**0.7241**
SMOTE	0.5126	0.5252	0.4869	0.6781	0.5252	0.5720
Strateji 1	0.4793	0.4432	**0.6654**	0.6662	0.4432	0.5004
Strateji 2	0.5185	0.5849	0.3886	0.6592	0.5849	0.6157
Strateji 3	0.4835	0.4854	0.4827	0.6478	0.4854	0.5236
Strateji 4	0.4860	0.4751	0.5123	0.6661	0.4751	0.5413
Strateji 5	0.5333	0.5636	0.4759	0.6877	0.5636	0.5978

Çizelge 4.10'de sonuçlara göre SVM sınıflandırma algoritmasını doğruluk başarı ölçüsü 0.6352 değeri ile orijinal veri dengeleme yöntemi iyi başarı elde etmiştir. Duyarlılık başarı ölçüsü 0.7187 değeri ile orijinal veri dengeleme yöntemi iyi başarı elde etmiştir. Bu tezde, Özgünlük başarı bizim için daha önemledir çünkü saldırıyı bulamalı önemli ve bu çizelgede özgünlük başarı ölçüsü 0.6654 değeri ile strateji 1 veri dengeleme yöntemi iyi başarı elde etmiştir. Kesinlik başarı ölçüsü 0.7297 değeri ile orijinal veri dengeleme yöntemi iyi başarı elde etmiştir. Hassasiyet başarı ölçüsü 0.7187 değeri ile orijinal veri dengeleme yöntemi iyi başarı elde etmiştir. F-ölçüsü başarı ölçüsü 0.7241 değeri ile orijinal veri dengeleme yöntemi iyi başarı elde etmiştir.

Çizelge 4.11, Çizelge 4.12, ve Çizelge 4.13 saldır tespit veri kümesinin sırasıyla K-NN, C4.5 ve SVM SMOTE ve DE stratejilerinin çoğunluk ve azınlık değerleri verilmiştir.

Çizelge.4.11. K-NN saldır tespit veri setinin çoğunluğu ve azınlığı

Saldır Tespit	Azınlık başarı oranı	Çoğunluk başarı oranı
Orijinal	0.6556	0.6624
SMOTE	0.6395	0.6539
Strateji 1	0.6467	0.6550
Strateji 2	0.6579	0.6623
Strateji 3	0.6292	0.6383
Strateji 4	0.6369	0.6467
Strateji 5	0.6611	0.6611

Çizelge 4.12. C4.5 saldır tespit veri setinin azınlığı ve çoğunluğu

Saldır Tespit	Azınlık başarı oranı	Çoğunluk başarı oranı
Orijinal	0.4863	0.5183
SMOTE	0.5165	0.5429
Strateji 1	0.5797	0.5797
Strateji 2	0.5000	0.5229
Strateji 3	0.5175	0.5383
Strateji 4	0.5050	0.5374
Strateji 5	0.4795	0.5160

Çizelge 4.13. SVM'nın saldır tespit veri setinin azınlığı ve çoğunluğu

Saldır Tespit	Azınlık başarı oranı	Çoğunluk başarı oranı
Orijinal	0.5797	0.5797
SMOTE	0.3932	0.4729
Strateji 1	0.3814	0.4644
Strateji 2	0.4111	0.5176
Strateji 3	0.3168	0.4781
Strateji 4	0.3905	0.4872
Strateji 5	0.3187	0.5070

Yukarıdaki çizelge 4.11, Çizelge 4.12, ve Çizelge 4.13 de görüldüğü gibi her veri kümesi için sınıfların örnek sayılarına göre azınlık ve çoğunluk değerleri hesaplanmıştır**.**

Azınlık az veriler demektir ve çoğunluk çok fazla veriler demektir. Nasıl hesaplandı Karışıklık Matristen hesapladık:

Azınlık başarısı = $\frac{\mathrm{TN}}{\mathrm{TN+FN}}$

Çoğunluk başarısı = $\frac{TP}{TP+\mathrm{FP}}$

5. SONUÇLAR VE ÖNERİLER

5.1. Sonuçlar

Dengesiz veriler birçok gerçek dünya problemi verisinde bulunmaktadır. Bu problemi çözümü için farklı yöntemler vardır. SMOTE, özellik alanındaki veri noktaları arasındaki Öklid Uzaklığı tarafından değerlendirilen en yakın komşulara dayanan bir tekniktir. Bu tez çalışmasında verileri dengelemek için SMOTE ve DE stratejileri kullanmıştır. DE stratejileri yüksek doğruluk elde etmek için kullanılır. DE büyük ve doğal olarak karmaşık bir evrim sürecinin küçük ve basit bir matematiksel modelidir. Ayrıca DE, basit yapı ve uygulama, düşük hesaplama karmaşıklığı ve yüksek performans gibi diğer yöntemlerle karşılaştırıldığında birçok avantaja sahiptir.

Saldırı tespit sistemleri, halen üzerinde araştırma yapılması gereken önemli bir çalışma alanıdır. Daha etkin saldırı tespit sistemi tasarlamak için makine öğrenme teknikleri sıklıkla kullanılmaktadır. Yapılan çalışmada saldırı tespit veri setleri kullanılarak, makine öğrenmesi tekniklerinden karar ağaçlarının C4.5, SVM ve K-NN sınıflandırma algoritması ve sınıflandırma başarısı, doğruluk, duyarlılık, özgüllük, hassasiyet, kesinlik ve F-ölçüsünün yönünden saldırı tespit sistemlerindeki performansı incelenmiştir.

Sonuçlara göre, 22 veri kümesinde yapılan analize K-NN sınıflandırma algoritmasında geometrik ortalama değerleri K-NN sınıflandırma algoritmasında SMOTE veri dengeleme yöntemi daha iyi sınıflandırma göstermiştir. Standart sapma değerleri strateji3 ve strateji5 daha iyi sınıflandırma başarılarının elde etmiştir.

C4.5 sınıflandırma algoritmasında geometrik ortalama değerleri Strateji4 veri dengeleme yöntemi daha iyi sınıflandırma başarılarının elde etmiştir. Standart sapma değerleri Strateji2 veri dengeleme yöntemi daha iyi sınıflandırma başarılarının elde etmiştir.

SVM sınıflandırma algoritmasında geometrik ortalama değerleri Strateji5 veri dengeleme yöntemi daha iyi sınıflandırma başarılarının elde etmiştir. Standart sapma değerleri strateji5 veri dengeleme yöntemi daha iyi sınıflandırma başarılarının elde göstermektedir.

Saldırı tespit veri kümesine yapılan analize ve sonuçları baktığında, Karışıklık matrisi değerlerine baktığımızda K-NN sınıflandırma algoritması strateji5 veri dengeleme yöntemi iyi başarı elde etmiştir.

C4.5 ağları sınıflandırma algoritması strateji 1'de daha iyi sınıflandırma işlemi yapmaktadır.

Bunun dışında SVM sınıflandırma algoritması orijinal halinde daha iyi sınıflandırma işlemi göstermiştir.

Başarı ölçütleri analizinde K-NN sınıflandırma algoritmasının doğruluk başarı ölçüsü strateji 5'te veri dengeleme yöntemi iyi başarı elde etmiştir. Özgünlük strateji 4'te, kesinlik ile orijinal 'da, hassasiyet strateji 5'te ve F-ölçüsü strateji 5'te veri dengeleme yöntemi iyi başarı elde etmiştir.

C4.5 sınıflandırma algoritmasının doğruluk başarı ölçüsü strateji 1'de, özgünlük strateji 3'te, duyarlılık strateji 1'de, kesinlik ile strateji 1'de, hassasiyet strateji 1'de ve F-ölçüsü strateji 1'de veri dengeleme yöntemi iyi başarı elde etmiştir.

SVM sınıflandırma algoritmasının doğruluk başarı ölçüsü orijinal halinde, duyarlılık orijinal halinde, özgünlük strateji 1'de, duyarlılık strateji 1'de, kesinlik ile orijinal halinde, hassasiyet strateji 1'de ve F-ölçüsü strateji 1'de veri dengeleme yöntemi iyi başarı elde etmiştir.

Vurgulanan bulgular, gelecekte makine öğrenme teknikleri ile daha etkin saldırı tespit kümesine ve DE stratejileri tasarlamak isteyen araştırmacılara faydalı olabilir.

5.2. Öneriler

Dengesiz veri kümeleri, bilimsel alanda artan pratik uygulamalarla son zamanlarda artan gerçek bir problemdir. Literatürde geliştirilen diğer dengeleme yöntemleri kullanılarak sınıflandırma başarıları incelenebilin.

Bu çalışmada kullanılan veri setleri dışında KDD, NLS-KDD gibi farklı veri setleri araştırılarak seçilen dengelem algoritmaları bu veri setlerinin uygun olanlarına uygulanabilir. Bu çalışmada kullanılan sınıflandırma algoritmaları dışındaki algoritmalar dengelenmiş verilerle kullanılarak anlarında başarıları ölçülebilir.

KAYNAKLAR

Achlioptas, D., McSherry, F. ve Schölkopf, B., Sampling Techniques for Kernel Methods.

Achlioptas, D., McSherry, F. ve Schölkopf, B., 2002, Sampling techniques for kernel methods, *Advances in neural information processing systems*, 335-342.

Albert, J. H., 1993, Teaching Bayesian statistics using sampling methods and MINITAB, *The American Statistician*, 47 (3), 182-191.

Ao, Y. ve Chi, H., 2009, Experimental Study on Differential Evolution Strategies, *IEEE*, 19-24.

Auria, L. ve Moro, R. A., 2008, Support vector machines (SVM) as a technique for solvency analysis.

Babar, V. ve Ade, R., 2016, A Novel Approach for Handling Imbalanced Data in Medical Diagnosis using Undersampling Technique, *Communications on Applied Electronics (CAE), Foundation of Computer Science FCS, New York*, 5.

Barua, S., Islam, M. M., Yao, X. ve Murase, K., 2014, MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning, *IEEE Transactions on Knowledge and Data Engineering*, 26 (2), 405-425.

Bhagat, R. C. ve Patil, S. S., 2015, Enhanced SMOTE algorithm for classification of imbalanced big-data using random forest, *2015 IEEE International Advance Computing Conference (IACC)*, 403-408.

Blagus, R. ve Lusa, L., 2012, Evaluation of SMOTE for high-dimensional class-imbalanced microarray data, *IEEE*, 89-94.

Bunkhumpornpat, C., Sinapiromsaran, K. ve Lursinsap, C., 2009, Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, *Pacific-Asia conference on knowledge discovery and data mining*, 475-482.

Chawla, N. V., Bowyer, K. W., Hall, L. O. ve Kegelmeyer, W. P., 2002, SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research 16*.

Chawla, N. V., Lazarevic, A., Hall, L. O. ve Bowyer, K. W., 2003, SMOTEBoost: Improving prediction of the minority class in boosting, *European conference on principles of data mining and knowledge discovery*, 107-119.

Chawla, N. V., 2009, Data mining for imbalanced datasets: An overview, In: Data mining and knowledge discovery handbook, Eds: Springer, p. 875-886.

Curtis, S., Gesler, W., Smith, G. ve Washburn, S., 2000, Approaches to sampling and case selection in qualitative research: examples in the geography of health, *Elsevier*, 50, 1001-1014.

Data School, 2014, Simple guide to confusion matrix terminology, https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/: [15/09/2019].

Drummond, C. ve Holte, R. C., 2003, C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, *Workshop on learning from imbalanced datasets II*, 1-8.

Engelbrecht, A. P., 2007, Computational intelligence: an introduction, John Wiley & Sons, p.

Ertuğrul, Ö. F. ve Tağluk, M. E., 2017, A novel version of k nearest neighbor: Dependent nearest neighbor, *Applied Soft Computing*, 55, 480-490.

Fernández, A., Del Jesus, M. J. ve Herrera, F., 2010, Multi-class imbalanced data-sets with linguistic fuzzy rule based classification systems based on pairwise learning, *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 89-98.

Ganganwar, V., 2012, An overview of classification algorithms for imbalanced datasets, *International Journal of Emerging Technology and Advanced Engineering*, 2 (4), 42-47.

Gao, M., Hong, X., Chen, S. ve Harris, C. J., 2011, A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems, *Neurocomputing*, 74 (17), 3456-3466.

Han, H., Wang, W.-Y. ve Mao, B.-H., 2005, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, *International conference on intelligent computing*, 878-887.

Hastings, W. K., 2018, Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *OXFORD*, 57, 97-109.

He, H., Bai, Y., Garcia, E. A. ve Li, S., 2008, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, 1322-1328.

Hoens, T. R. ve Chawla, N. V., 2013, Imbalanced datasets: from sampling to classifiers, *Imbalanced Learning: Foundations, Algorithms, and Applications*, 43-59.

Hu, F. ve Li, H., 2013, A novel boundary oversampling algorithm based on neighborhood rough set model: NRSBoundary-SMOTE, *Mathematical Problems in Engineering*, 2013.

Karacalarlı, U., 2018, Performance Increase of Intrusion Detection Systems Utilizing Support Vector Machine (Svm) By Feature Selection, *EGE University*.

Kaya, Ç., 2016, Use of Machine Learning Techniques in Intrusion Detection Systems: Comparative Analysis of Performance.

Kotsiantis, S., Kanellopoulos, D. ve Pintelas, P., 2006, Handling imbalanced datasets: A review, *GESTS International Transactions on Computer Science and Engineering*, 30 (1), 25-36.

Kyoto Dataset, 2006, Traffic Data from Kyoto University's Honeypots, http://www.takakura.com/Kyoto_data/:

Lewis, D. D. ve Catlett, J., 1994, Heterogeneous uncertainty sampling for supervised learning, In: Machine learning proceedings 1994, Eds: Elsevier, p. 148-156.

Liu, A. C., 2004, The effect of oversampling and undersampling on classifying imbalanced text datasets, *The University of Texas at Austin*.

Liu, J. ve Lampinen, J., 2005, A fuzzy adaptive differential evolution algorithm, *Soft Computing*, 9 (6), 448-462.

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

Mammone, A., Turchi, M. ve Cristianini, N., 2009, Support vector machines, *Wiley Interdisciplinary Reviews: Computational Statistics*, 1 (3), 283-289.

Mohankumar, M., Amuthakkani, S. ve Jeyamala, G., 2016, Comparative analysis of decision tree algorithms for the prediction of eligibility of a man for availing bank loan, *Age*, 19, 60.

Mukherjee, S. ve Sharma, N., 2012, Intrusion detection using naive Bayes classifier with feature reduction, *Procedia Technology*, 4, 119-128.

Noghabi, H. S., Mashhadi, H. R. ve Shojaei, K., 2015, Differential Evolution with Generalized Mutation Operator for Parameters Optimization in Gene Selection for Cancer Classification, *arXiv preprint arXiv:1510.02516*.

Pérez-Ortiz, M., Jiménez-Fernández, S., Gutiérrez, P., Alexandre, E., Hervás-Martínez, C. ve Salcedo-Sanz, S., 2016, A review of classification problems and algorithms in renewable energy applications, *Energies*, 9 (8), 607.

Pourabbas, F., 2014, An approach for classifying alerts of intrusion detection systems, *The institute of sciences*.

Ramentol, E., Caballero, Y., Bello, R. ve Herrera, F., 2012a, SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, *Knowledge and information systems*, 33 (2), 245-265.

Ramentol, E., Verbiest, N., Bello, R., Caballero, Y., Cornelis, C. ve Herrera, F., 2012b, SMOTE-FRST: a new resampling method using fuzzy rough set theory, In: Uncertainty Modeling in Knowledge Engineering and Decision Making, Eds: World Scientific, p. 800-805.

Shelke, M. M. S., Deshmukh, P. R. ve Shandilya, V. K., 2017, A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique.

Tang, H., Xue, S. ve Fan, C., 2008, Differential evolution strategy for structural system identification, *Elsevier*.

Tarım, M. C., 2011, A Faster Intrusion Detection Method for High-Speed Computer Networks, *Middle East Technical University*.

Varun Kumar, S. ve Panneerselvam, R., 2017, A study of crossover operators for genetic algorithms to solve VRP and its variants and new sinusoidal motion crossover operator, *Int. J. Comput. Intell. Res*, 13 (7), 1717-1733.

Waksberg, J., 1978, Sampling methods for random digit dialing, *Journal of the American Statistical Association*, 73 (361), 40-46.

Wang, J., Xu, M., Wang, H. ve Zhang, J., 2006, Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding, *IEEE*.

Weiss, G. M., McCarthy, K. ve Zabar, B., 2007, Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?, *DMIN*, 7, 35-41.

Wu, Y.-C., Lee, W.-P. ve Chien, C.-W., 2011, Modified the performance of differential evolution algorithm with dual evolution strategy, *International conference on machine learning and computing*, 57-63.

Yucel, A., 2016, Predictive Text Analytics and Text Classification Algorithms.

Zhou, Z.-H. ve Liu, X.-Y., 2006, Training cost-sensitive neural networks with methods addressing the class imbalance problem, *IEEE Transactions on Knowledge and Data Engineering*, 18 (1), 63-77.

Zyt, J., Klosgen, W. ve Zytkow, J., 2002, Handbook of data mining and knowledge discovery, Oxford university press, p.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Samara Khamees Jwair JWAIR
Uyruğu : Iraklı
Doğum Yeri ve Tarihi : IRAK- BAĞDAT 19 Haziran 1991
Telefon : 00905382114563
Faks :
E-Posta : Samarajassim91@gmail.com

EĞİTİM

Derece		**Adı, İlçe, İl**		**Bitirme Yılı**
Lise	:	**9 Nissan kız lisesi,**	**KERKÜK**	**2009**
Üniversite	:	**AL- Qalam Üniversitesi,**	**KERKÜK**	**2015**
Yüksek Lisans	:			
Doktora	:			

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**

UZMANLIK ALANI

YABANCI DİLLER (Arapça, Türkçe, İnglizce)

BELİRTMEK İSTEĞİNİZ DİĞER ÖZELLİKLER

YAYINLAR Jwair, S VE KAYA. E., 2018. The Effect Of Balancing Process On classifying Unbalancing Data Set. (ICENTE'18)''