T.C.
KONYA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

YAPAY SİNİR AĞI KULLANILARAK DENGESİZ VERİ KÜMELERİNDE SINIFLANDIRMA BAŞARISININ ARTIRILMASI

Fatih DİKDERE

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Ocak-2021
KONYA

TEZ KABUL VE ONAYI

Fatih DİKDERE tarafından hazırlanan "Yapay Sinir Ağı Kullanılarak Dengesiz Veri Kümelerinde Sınıflandırma Başarısının Artırılması" adlı tez çalışması …/…/… tarihinde aşağıdaki jüri tarafından oy birliği / oy çokluğu ile Konya Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri **İmza**

Başkan
Dr.Öğr.Üyesi Burak YILMAZ

Danışman
Dr.Öğr.Üyesi Ersin KAYA

Üye
Dr.Öğr.Üyesi Sedat KORKMAZ

Yukarıdaki sonucu onaylarım.

Prof. Dr. Saadettin Erhan KESEN
Enstitü 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.

Fatih DİKDERE

26.02.2021

ÖZET

YÜKSEK LİSANS TEZİ

YAPAY SİNİR AĞI KULLANILARAK DENGESİZ VERİ KÜMELERİNDE SINIFLANDIRMA BAŞARISININ ARTIRILMASI

Fatih DİKDERE

Konya Teknik Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr.Öğr.Üyesi Ersin KAYA

2021, … Sayfa

Jüri
Dr.Öğr.Üyesi Ersin KAYA
Dr.Öğr.Üyesi Burak YILMAZ
Dr.Öğr.Üyesi Sedat KORKMAZ

Veri kümelerinde, sınıflar arasında dengeli bir dağılım bulunmaması sonucunda dengesiz veri kümeleri ortaya çıkmaktadır. Bu dengesiz veri kümelerinde karşılaşılan en büyük problemlerden biri ise sınıflandırma başarısıdır. Sınıflandırma başarısı çoğunluk sınıfında yüksek değerlere yakın iken, azınlık sınıfında sınıflandırma başarısında yanlışlıklar ve hatalar görülmektedir. Bu tez çalışmasında dengesiz dağılım gösteren veri kümelerinde sınıflandırma başarısının artırılmasına yönelik çalışmalar yapılmıştır. Sınıflandırma başarısının artırılması için yapay sinir ağlarından yararlanılmıştır. Bu çalışmada yapay sinir ağları kullanılarak yedi yöntem önerilmiş olup sınıflandırma sonuçları için geometrik ortalama ve f ölçüsü metriklerinden yararlanıp, bu metriklerin değerlendirilmesi için de Friedman testi istatistik ölçüsünden faydalanılmıştır. Bu yöntemlerde en başarılı sonuç elde edilen yöntemde yapay sinir ağları ile rastgele örnekler üretilmiş olup bu örnekler bir eşik değeriyle sınırlandırılmıştır. Tez çalışmasında yapılan yöntemlerden alınan sonuçlar, orijinal veri kümesinin sonuçları ve temel SMOTE yöntemi sonuçları ile karşılaştırılırmıştır. Karşılaştırılma sonucunda başarılı sonuçlar elde edilmiş olup dengesiz veri kümelerinde sınıflandırma başarısı artırılmıştır.

Anahtar Kelimeler: Dengesiz veri kümeleri, ikili sınıf etiketi, SMOTE, yeniden örnekleme

ABSTRACT

MS THESIS

INCREASING THE CLASSIFICATION SUCCESS IN IMBALANCED DATA SETS USING ARTIFICIAL NEURAL NETWORK

Fatih DİKDERE

Konya Technical University
Institute of Graduate Studies
Department of Computer Engineering

Advisor: Asst.Prof.Dr. Ersin KAYA

2021, … Pages

Jury
Asst.Prof.Dr. Ersin KAYA
Asst.Prof.Dr. Burak YILMAZ
Asst.Prof.Dr. Sedat KORKMAZ

In data sets, imbalanced data sets emerge as a result of not having a balanced distribution among classes. One of the biggest problems encountered in these unbalanced data sets is classification success. While the classification success is close to high values in the majority class, inaccuracies and errors are observed in the classification success in the minority class. In this thesis, studies have been conducted to increase the success of classification in data sets with uneven distribution. Artificial neural networks have been used to increase the classification success. In this study, seven methods using artificial neural networks are proposed, and geometric mean and f measure metrics are used for classification results, and Friedman's means evaluation statistics measure is used to evaluate these metrics. In these methods, random samples were produced with artificial neural networks in the method with the most successful results, and these samples were limited to a threshold value. The results obtained from the methods in the thesis study were compared with the results of the original data set and the results of the basic SMOTE method. Successful results were obtained as a result of the comparison, and classification success was increased in unbalanced data sets.

Keywords: Binary class tag, imbalanced data sets, SMOTE, resampling

ÖNSÖZ

Bu tez çalışmasına değerli bilgileri ile katkıda bulunan danışman hocam Sayın Dr.Öğr. Üyesi Ersin KAYA'ya, jüri üyelerim Sayın Dr.Öğr.Üyesi Burak YILMAZ'a ve Sayın Dr.Öğr.Üyesi Sedat KORKMAZ'a ve diğer hocalarıma teşekkür eder, saygılarımı sunarım.

Fatih DİKDERE
KONYA-2021

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

ANOVA:	Analysis Of Variance
ARFF:	Attribute Relation File Format
AUC:	Area Under The Curve
CPU:	Central Proccessing Unit
CSV:	CommaSeperated Values
CURE:	Clustering Using Representatives
DAT:	Data File
DMLP:	Discrimination-based Multilayer Perceptron
FN:	False Negative
FP:	False Positive
GPU:	Graphics Processing Unit
IDE:	Integrated Development Environment
IPF:	Iterative-Partitioning Filter
LVQ:	Learning Vector Quantization
ÖYRÖ:	Önerilen Yöntem Rastgele Örnekleme
ÖYOK:	Önerilen Yöntem Otomatik Kodlayıcı
ÖYOKED:	Önerilen Yöntem Otomatik Kodlayıcı Eşik Değeriyle
ÖYT:	Önerilen Yöntem Tekrarlı
ÖYSA:	Önerilen Yöntem SMOTE Azınlık
ÖYSÇ:	Önerilen Yöntem SMOTE Çoğunluk
ÖYED:	Önerilen Yöntem Eşik Değeriyle
RMLP:	Recognition-based Multilayer Perceptron
ROC:	Receiver Operating Characteristic
RST:	Rough Set Theory
SELU:	Scaled Exponential Linear Unit
SMOTE:	Synthetic Minority Oversampling Technique
SVM:	Support Vector Machine
TN:	True Negative
TP:	True Positive
TXT:	Text
WEKA:	Waikato Environment for Knowledge Analysis

1. GİRİŞ

1.1. Dengesiz Veri Kümeleri

Son yıllarda artan teknolojik gelişmelerle, kullanılan teknolojik cihazların sayısı ve bu teknolojik cihazlarla beraber veri miktarlarında önemli oranda artış gözlemlenmiştir. Artan veri miktarı, verilerin işlenmesi, verilerin sınıflandırılması gibi konularda da problemleri beraberinde getirmiştir. Burada verilerin dağılım hakkında en önemli kavramlardan biri, veri setindeki dengesiz dağılımı kavramı olmuştur. Yapılan tez çalışmasında iki sınıflı dengesiz veri kümeleri üzerinde çalışılmıştır. İki sınıflı veri kümeleri bir veya sıfır, hasta veya sağlıklı, hata veya başarı gibi iki tane sonucun yer aldığı veri kümeleridir. İkili veri kümelerinin dağılımında bir sınıfın diğer sınıftan fazla olduğu durumlarda dengesizlik ortaya çıkmaktadır.

Dengesiz veri kümelerinde sınıflar sadece hasta ve hasta değil, erkek ve kadın gibi iki çeşitten değil ikiden fazla da olabilmektedir. Yani veri kümeleri çoklu sınıflardan oluşabilmektedir. Örnek olarak renkleri ele aldığımızda; renkler temel olarak kırmızı, yeşil ve maviden oluşmaktadır. Renkler ile ilgili bir veri kümesinde kullanıcıdan girdi olarak alınan bir rengin hangi renge daha yakın olduğu tutulabilir. Burada görüldüğü üzere veri kümesinde ikiden fazla sınıf yani kırmızı, yeşil ve mavi bulunmaktadır.

Dengesizlik teknik anlamda incelendiğinde ise sınıflar arasında eşit olmayan bir dağılım gösteren herhangi bir veri kümesi dengesiz veri kümesi olarak kabul edilmektedir. Bu oranlar 100:1, 100:40, 1000:2, 100000:1 gibi oranlarda olabilmekte iken dengesizlik oranları için kesin bir tanım yoktur. Tez çalışmasında dengesizlik oranları 1.5 ve 9 arasında değişen veri kümeleri kullanılmıştır.

Dengesiz veri kümelerine sağlık, araç, biyoloji, yazılım, veri tabanları, iletişim gibi alanlarda sıkça rastlanmaktadır. Örnek olarak tıp alanındaki veri kümesini ele alıp kanser veri kümesi üzerinden ilerlensin: Bu veri kümesinde 10000 negatif yani hasta olmayanlar ve 10 pozitif yani hasta olanlar olmak üzere 10010 kayıt var olduğu kabul edilsin. Burada negatif sınıfına çoğunluk sınıfı ve pozitif sınıfına ise azınlık sınıfı adı verilmektedir. Hem çoğunluk hem de azınlık sınıfı için ideal ve dengeli bir sınıflandırma yöntemine ihtiyacımız vardır. Ne yazık ki gerçek hayatta sınıflandırıcılar çoğunluk sınıfını yüksek bir doğruluk ile sınıflandırırken azınlık sınıfını ise neredeyse sınıflandıramayacak kadar düşük bir başarı oranına sahiptir. Burada görülen en önemli

problem ise pozitif hastalarda sınıflandırma başarısızlığından dolayı hasta değilmiş gibi sınıflandırılmasıdır.

Bu nedenle çoğunluk sınıfının sınıflandırma başarısını tehlikeye atmadan azınlık sınıfının sınıflandırma başarısını artıracak yöntemlere ihtiyaç duyulmaktadır. Bu yöntemlerin kategorileri ise örnekleme yöntemleri, maliyet duyarlı yöntemler, çekirdek tabanlı yöntemler, aktif öğrenme yöntemleri ve diğer yöntemlerdir (He ve Garcia, 2009). Örnekleme yöntemleri:

- Rastgele aşırı örnekleme ve alt örnekleme
- Bilgilendirilmiş alt örnekleme
- Veri üretimi ile sentetik örnekleme
- Uyarlanabilir sentetik örnekleme
- Veri temizleme ile örnekleme
- Kümeleme tabanlı örnekleme
- Örnekleme ve arttırma entegrasyonu

Maliyet duyarlı yöntemler:

- Uyarlamalı arttırma ile maliyete duyarlı veri alanı ağırlığı
- Maliyet duyarlı karar ağaçları
- Maliyet duyarlı sinir ağları

Çekirdek tabanlı yöntemler:

- Örnekleme yöntemiyle entegrasyon
- Çekirdek modifikasyon metotları

Ve aktif öğrenme yöntemleri olarak kategorize edilmiştir.

Rastgele aşırı örnekleme tekniğinde azınlık sınıfından rastgele olarak seçilen örnekler orijinal veri kümesine eklenerek azınlık ve çoğunluk örnekleri arasında denge sağlanmaya çalışılmaktadır. Bu yöntemin dezavantajı ise birden çok örnek aynı veri kümesinde bulunacağından aşırı öğrenmeye sebep olmaktadır.

Rastgele alt örnekleme yönteminde ise çoğunluk sınıfından rastgele olarak seçilen örnekler orijinal veri kümesinden çıkarılarak azınlık ve çoğunluk örnekleri arasında denge sağlanmaya çalışılmaktadır. Bu yöntemin dezavantajı ise karar vermede önemli olabilecek verilerin veri kümesinden silinmesine yol açabilmektedir.

SMOTE yöntemi ise dengesiz dağılım gösteren veri kümelerinin dengeli hale getirilip sınıflandırma sonuçlarının iyileştirilmesi için en çok kullanılan ve başarılı yöntemlerdendir (Chawla ve ark., 2002). Veri kümesindeki azınlık örnek sayısını istenilen oranda arttırmak için kullanılır.

1.2. Tez Çalışması Kapsamı

Yapılan tez çalışmasının amacı, dengesiz veri kümelerinin sınıflandırma başarısı üzerinedir. Bu çalışmadaki önerilen yöntem ile dengesiz veri kümelerinin sınıflandırma başarısını artırma konusunda başarılı sonuçlar elde edilmiştir. Bu önerilen yöntemde ise dengesiz veri kümelerinde azınlık sınıfı ile çoğunluk sınıfı veri sayıları eşitlenmeye çalışılmıştır. Veriler eşitlendikten sonra sınıflandırma işlemi uygulanıp bu işlemin sonucunda geometrik ortalama ve f ölçüsü değerleri karşılaştırılmıştır. Geometrik ortalama ve f ölçüsü sonuçlarının hangi yöntemde başarılı olduğuna karar vermede Friedman testi istatistik ölçüsünden yararlanılmıştır. Veri kümelerinde sınıflar arasındaki veri sayılarının eşitlenebilmesi için yedi yöntem önerilmiştir. Önerilen yöntemler orijinal veri kümesi ve SMOTE yöntemi sonuçlarıyla karşılaştırılmıştır. Sınıflandırma başarısının artırılması için önerilen yedi yöntem ise:

1. Yapay sinir ağları kullanılarak rastgele örnekleme yöntemi. Bu Önerilen Yöntem Rastgele Örnekleme (ÖYRÖ) olarak adlandırılmıştır.
2. Yapay sinir ağları üzerinde otomatik kodlayıcı tekniği ile rastgele örnekleme yöntemi. Bu Önerilen Yöntem Otomatik Kodlayıcı (ÖYOK) olarak adlandırılmıştır.
3. Yapay sinir ağları üzerinde otomatik kodlayıcı tekniği ile rastgele örnekleme yöntemine çıkış ve giriş arasında eşik değeriyle rastgele örnekleme yöntemi. Bu önerilen yöntem Otomatik Kodlayıcı Eşik Değeriyle (ÖYOKED) olarak adlandırılmıştır.
4. Yapay sinir ağları kullanılarak azınlık sınıfının tekrarlı olarak veri kümesine eklenmesiyle örnekleme yöntemi. Bu Önerilen Yöntem Tekrarlı (ÖYT) olarak adlandırılmıştır.
5. SMOTE yönteminin ürettiği örnekler ile azınlık sınıfıyla eğitilmiş yapay sinir ağları üzerinde örnekleme yöntemi. Bu Önerilen Yöntem SMOTE Azınlık (ÖYSA) olarak adlandırılmıştır.
6. SMOTE yönteminin ürettiği örnekler ile çoğunluk sınıfıyla eğitilmiş yapay sinir ağları üzerinde örnekleme yöntemi. Bu Önerilen Yöntem SMOTE Çoğunluk (ÖYSÇ) olarak adlandırılmıştır.
7. Yapay sinir ağları üzerinde rastgele örnekleme yönteminde çıkış ve giriş arasında eşik değeriyle örnekleme yöntemi. Bu Önerilen Yöntem Eşik Değeriyle (ÖYED) olarak adlandırılmıştır.

Önerilen yöntemlerin sınıflandırma sonuçlarının değerlendirilebilmesi için beş sınıflandırma algoritması kullanılmıştır. Bu sınıflandırma algoritmaları:

1. AdaBoost.M1
2. K en yakın komşu (KNN)
3. K star
4. Çok katmanlı yapay sinir ağları (MLP)
5. Sıralı asgari optimizasyon (SMO)

Bu yöntem ve sınıflandırma algoritmalarının sonuçları için dengesizlik oranları 1.5 ile 9 arasında değişen on veri kümesi ve bu veri kümelerinin beş farklı dağılımı kullanılmıştır. Veri kümelerinin detayları bölüm 3.5'te açıklanmıştır.

1.3. Tez Dokümanı Kapsamı

Yapılan tez çalışmasına ilişkin bu dokümanda ilk olarak birinci bölümde dengesiz veri kümeleri ve tez çalışmasına dair bir giriş yapılmıştır. İkinci bölümde yine dengesiz veri kümeleri, dengesiz veri kümeleri için yaygın olarak kullanılan yöntem olan SMOTE ve dengesiz veri kümeleri hakkında yapılan çalışmalar için kaynak araştırması yer almıştır. Üçüncü kısımda materyal ve yöntem hakkında detaylı bilgiler verilmiştir. Dördüncü kısım ise bu yedi yöntem ve beş sınıflandırma algoritmasının sonuçlarına yer verilen araştırma sonuçları ve tartışma bölümüdür. Son olarak beşinci bölümde ise sonuçlar ve öneriler yer almıştır.

2. KAYNAK ARAŞTIRMASI

Verilerin dağılımı, verilerin dağılımının yeniden yapılması ve sınıflandırma sonuçlarının iyileştirilmesi alanlarında çalışmalar yapılmıştır. Kaynak araştırmaları çizelge 2.1'de gösterilmiştir.

Çizelge 2.1. Dengesiz veri kümeleri kaynak araştırmaları

Yazar	Yıl	Çalışma
Anand ve ark.	1993	An improved algorithm for neural network classification on imbalanced training sets
Caruna	2000	Learning From Imbalanced Data: Rank Metrics and Extra Tasks
Japkowicz	2000	Learning From Imbalanced Data: A Comparison of Various Strategies
Chawla ve ark	2002	SMOTE: Synthetic Minority Over-sampling Technique
Estabrooks ve ark.	2004	A Multiple Resampling Method for Learning from Imbalanced Data Sets
Han ve ark.	2005	Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning
Ayhan	2009	Multi-class classification methods utilizing Mahalanobis Taguchi system and a re-sampling approach for imbalanced data sets
Bunkhumpornpat	2009	Safe-Level SMOTE: Safe-Level Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem
He ve Garcia	2009	Learning from Imbalanced Data
Öztürk	2009	SVM classification for imbalanced datasets with multi objective optimization framework
Sun ve ark.	2009	Classification of Imbalanced Data: A Review, Classification of Imbalanced Data: A Review
Jeatrakul ve ark.	2010	Classification of Imbalanced Data by Combining the Compleentary Neural Network and SMOTE algorithm
Maciejewski ve Stefanowski	2011	Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data
Ramentol ve ark.	2011	SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory
Nakamura ve ark.	2013	LVQ-SMOTE – Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data
Sarmanova	2013	Veri madenciliğinde sınıf dengesizliği sorununun giderilmesi
Sáez ve ark.	2015	SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, Information Sciences
Aydın	2016	Dengesiz veri setlerinde trafik işaretlerini tanıma
Bulut	2016	Sınıflandırıcı Topluluklarının Dengesiz Veri Kümeleri Üzerindeki Performans Analizleri
Krawczyk	2016	Learning from imbalanced data: open challenges and future directions
Wang ve ark.	2016	Training deep neural networks on imbalanced data sets
Douzas ve Bacao	2017	Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning
Ma ve Fan	2017	CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests
Ankara	2019	Dengesiz kredi skorlama veri setlerinde kolektif öğrenme algoritmalarının performans değerlendirmesi

Douzas ve Bacao	2019	Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE
Fotouhi	2019	A comprehensive data level analysis for cancer diagnosis on imbalanced data
Gümüştaş	2019	Kayıp gözlem içeren dengesiz veri setlerinin topluluk öğrenme algoritmaları le sınıflandırılması
Haklı	2019	Sınıf dengesizliği sorununu çözmek için kullanılan algoritmaların farklı sınıflandırma yöntemlerinde performanslarının karşılaştırılması
Turhan	2019	Kolektif öğrenmede sınıf dengesizliği problemi
Sağlam	2020	Gürültülü gözlemler durumunda dengesiz veride öğrenme için yeni bir yaklaşım

Anand ve ark. (1993) tarafından yapılan çalışmada dengesiz veri kümelerinin sinir ağı sınıflandırması üzerine çalışılmıştır. İkili sınıflı dengesiz veri kümelerinde her bir sınıf için hatayı azaltma yönünde bir çalışma yapılmıştır. Sonuç olarak sinir ağlarında öğrenme oranı hızlandırılmıştır.

Caruna (2000) tarafından yapılan çalışmada dengesiz veri kümeleri ile ilgili iki esas problem olan küçük veri kümeleri için öğrenme zorlukları ve değerlendirme süreçlerinden bahsedilmiştir. Değerlendirme yöntemlerinde standart yöntemler ile diğer geleneksel değerlendirme ve yumuşak değerlendirme yöntemlerinden ve bunların dönüşümlerinden de bahsedilmiştir. Öğrenme sürecindeki ekstra yöntemlerden biri olan çok görevli öğrenme yöntemi ele alınmıştır. Bu çok görevli öğrenme süreci açıklanıp sonuçlar verilmiştir. Sonuç olarak bu yumuşak değerlendirme ve çok görevli öğrenme yönteminin birleşiminin performansı artırdığı gözlemlenmiştir.

Japkowicz (2000) tarafından yapılan çalışmada dengesiz veri kümeleri üzerinde standart sınıflandırıcıların performansı üzerine çalışılmıştır ve bu problemle başa çıkmak için kullanılan yöntemlerin performans karşılaştırılması yapılmıştır. Bu çalışmada da iki sorun üzerinde tartışılmıştır. Ayrıca burada iki önerilmiştir. Bunlar ayrımcılık temelli çok katmanlı algılayıcılar (DLMP) ve tanıma temelli çok katmanlı algılayıcılardır (RLMP). Bu yöntemleri kullanarak elde edilenler ve var olan stratejiler ile karşılaştırmalar yapılmıştır. Sonuç olarak ise bu çalışmada dengesiz veri kümesi üzerine yapılan yöntemler üzerinde durulmuştur ve gelecekte yapılabilecek çalışmalar için öneriler sunulmuştur.

Chawla ve ark. (2002) tarafından yapılan çalışmada SMOTE yöntemi açıklanmıştır. Bu yöntemde azınlık sınıfı aşırı örnekleme ve çoğunluk sınıfı örnek azaltma yönteminin ROC alanında daha iyi performans gösterdiği belirtilmiştir. Ayrıca bu yöntem ile daha iyi sınıflandırma başarısı elde edilmiştir. Bu yöntemin daha önce dengesiz veri alanında yapılmış çalışmalarla karşılaştırması yapılmış ve bu yeni

yöntemin başarısı da sınıflandırma sonuçları ile gösterilmiştir. Ayrıca bu yeni algoritmaya ait sözde kod ise bölüm 3.1.1.2'de verilmiştir. Bu çalışma ABD enerji bakanlığı tarafından Sandia Ulusal Laboratuvarları ASCI VIEWS Veri Keşfi Programı aracılığıyla kısmen desteklenmiştir. Sonuç olarak aşırı örnekleme alanında yeni bir yöntem bulunmuş ve bu çalışma ileriki çalışmalara da destek olmuştur.

Estabrooks ve ark. (2004) çalışmasında dengesiz veri kümelerinden çıkarım yapmak için yeniden örnekleme yöntemleri üzerinde durmuştur. Burada aşırı örnekleme ve çoğunluk sınıfı örnek azaltma yöntemlerinin sorunları üzerine deneysel bir çalışma yapılmıştır.

Han ve ark. (2005) yaptıkları çalışmada SMOTE yöntemini kullanarak dengesiz verilerden madencilik için yeni sınır çizgisi SMOTE yöntemini önermişlerdir. Bu yöntemin ise 1 ve 2 olarak iki metodu bulunmaktadır. Burada da örnek üretirken tehlikeli bir alan vardır ve burada örnek üretilmemeye çalışılmıştır. Burada ilk olarak yöntemin aşamalarından ve algoritmasından bahsedilmiştir. Toplam dört adet veri kümesi kullanılmış olup dengesiz dağılım oranları birbirinden farklıdır. Bu önerilen iki yöntem orijinal veri kümesi, rastgele örnek üretme ve orijinal SMOTE yöntemi ile sonuçları karşılaştırılmıştır. Sonuçları karşılaştırmak için C4.5 yönteminden ve bunların başarılarını hesaplayabilmek için de başarı ölçüm yöntemlerinden kullanılmıştır. Sonuçlar ise yöntemin işe yaradığını göstermiştir. İleriki çalışmalarda ise bu tehlikeli alanla ilgili çalışmalar yapılabilecektir ve yöntemin diğer yöntemler ile denenmesi de önerilmiştir.

Ayhan (2009) tarafından yapılan bir yüksek lisans çalışmasında Mahalanobis Taguchi Sistem'i ile ilk olarak iki sınıflı dengesiz veri kümeleri için yeniden örnekleme yöntemi geliştirilmiştir. Mahalanobis Taguchi Sistem'i ile ikinci olarak ise çok sınıflı dengesiz veri kümeleri için de yeni sınıflandırma yöntemleri geliştirilmiştir.

Bunkhumpornpat ve ark. (2009) yaptıkları çalışmasında SMOTE yöntemini kullanarak dengesiz verilerden, veri madenciliği için güvenli seviye SMOTE yöntemini önermişlerdir. Güvenli seviye, aynı çizgide farklı ağırlık derecesi olarak adlandırılmaktadır. Bu çalışmada güvenli seviye belirlenerek sentetik örnek üretilip üretilemeyeceğine karar verilmiştir. Önerilen yöntemin testi için burada benzerlik oranları birbirinden farklı olan sadece iki adet veri kümesi kullanılmıştır. Sonuçların karşılaştırılması orijinal veri kümesi, orijinal SMOTE yöntemi ve önceden önerilen sınır güvenli SMOTE yöntemi ile karşılaştırmalar yapılmıştır. Sonuçları karşılaştırmak için çeşitli sınıflandırma algoritmaları ve bunların başarılarını hesaplayabilmek için yine

başarı ölçüm yöntemleri kullanılmıştır. Sonuç olarak ise daha iyi sonuçlar alındığını ve gelecekteki çalışmalarda yapılabilecek yöntemler üzerinde durulmuştur.

He ve Garcia (2009) tarafından yapılan çalışmada ilk olarak dengesiz verinin tanımlaması yapılmış, dengesiz verinin ne olduğu detaylıca incelenmiş ve yıllara göre bu konu ile ilgili yapılan çalışmaların artışına değinilmiştir. Dengesiz öğrenme probleminin giderilmesi için yöntemler kategorize edilmiştir. Bu kategoriler örneklendirme yöntemleri, maliyet tabanlı öğrenmeler, çekirdek tabanlı yöntemler, aktif öğrenme yöntemleri ve diğer öğrenme yöntemleri olarak beşe ayırmışlardır. Ayrıca her kategori kendi içerisinde farklı öğrenme yöntemlerine de ayrılmaktadır. Bu yöntemler sonucunda başarının ölçülebilmesi için de çeşitli değerlendirme yöntemleri kullanılmıştır. Bunlar tekil değerlendirme yöntemleri, alıcı işletim karakteristiği eğrisi gibi yöntemlerdir. Sonuç olarak bu açıklanan yöntemlerin ve tekniklerin ileriki çalışmalara yön göstermesi ve yardımcı olması ümit edilmiştir.

Öztürk (2009) tarafından yapılan bir yüksek lisans çalışmasında dengesiz veri kümelerindeki sınıflandırma başarısını artırmak için algoritma seviyesinden bir çalışma yapılmıştır. Burada kullanılan algoritma SVM yani destek vektör makineleri ile sınıflandırma başarısını artırmak hedeflenmiştir.

Sun ve ark. (2009) tarafından yazılan derleme makalesinde dengesiz veriler ile sınıflandırma konusu ele alınmıştır. Burada dengesiz verilerin ne olduğu konusundan ve sınıflandırma sorunlarından bahsedilmiştir. Dengesiz verilerin dolandırıcılık tespiti, medikal tanımlamalar, ağ saldırı tespiti gibi alanlarda kullanıldığına dair örnekler verilmiştir. Bu dengesiz veri kümelerinde dengesizlik sorunu üzerinde durulmuştur ve nasıl olduğu yönünde açıklamalar yapılmıştır. Karar ağaçları, destek vektör makineleri, k en yakın komşu gibi standart olarak kullanılan sınıflandırma algoritmalarının dengesiz veri kümeleri üzerindeki öğrenme zorluklarından bahsedilmiştir. Sınıflandırma sonucu ölçüm metriklerinden söz edilmiştir. Burada da ölçüm yöntemi olarak f ölçüsü, geometrik ortalama ve ROC eğrisinden yararlanılmıştır. Yine dengesiz öğrenme seviyeleri üç kategoride ele alınmıştır. Bu çalışmanın sonucunda ise dengesiz veriler üzerine detaylı bir çalışma yapıldığından, gelecekte bu konuya olan araştırmalara fikir sunması beklenmekte ve daha çok ikili sınıflar üzerinde çalışılıp ikiden fazla sınıflı dengesiz veri kümeleri üzerine de daha çok çalışma yapılması gerektiğinden bahsedilmiştir.

Jeatrakul ve ark. (2010) tarafında yapılan çalışmada tamamlayıcı sinir ağı ve SMOTE algoritması birleştirilerek dengesiz verilerin sınıflandırılması üzerine

çalışılmıştır. Önerilen yöntemin sınıflandırma sonuçları için üç sınıflandırma algoritması kullanılmıştır. Sonuç olarak çoğunlukla diğer yöntemlerden daha iyi performans gösterdiği görülmüştür.

Maciejewski ve Stefanowski (2011) yaptıkları çalışmada SMOTE yöntemini kullanarak dengesiz verilerden madencilik için SMOTE yönteminin genişletilmiş hali olarak yerel komşu SMOTE yöntemini ortaya koymuşlar ve bu yöntemin orijinal SMOTE algoritması, diğer genelleştirilmiş sınır çizgisi ve güvenli seviye SMOTE algoritması ile karşılaştırılması yapılmıştır. Burada ilk olarak yöntemin aşamalarından ve algoritmasından bahsedilmiştir. Rastgele örnekler oluşturulurken ayrıca örnek oluşturulup oluşturulamayacağına dair bir kontrol yapılmıştır. Toplam on dört adet veri kümesi kullanılmış olup dengesizlik oranları birbirinden farklıdır. Sonuçları karşılaştırmak için çeşitli sınıflandırma algoritmaları ve bunların başarılarını hesaplayabilmek için de yine başarı ölçüm yöntemleri kullanılmıştır. Sonuç olarak önerilen yöntemin başarılı sonuçlar verdiği gözlemlenmiştir.

Ramentol ve ark. (2011) yaptıkları çalışmada SMOTE yöntemini ve kaba küme teorisini kullanarak karma önişleme tabanlı yeni bir yöntem önermişlerdir. Bu yöntem iki aşamadan oluşmaktadır. Birinci aşamada SMOTE yöntemi kullanılarak yeni örnekler oluşturulmuştur. İkinci aşamada ise kaba küme teorisi (RST) dayalı bir temizleme işlemi uygulanır. Bu yöntem de önceden yapılmış olan SMOTE yöntemleri ile karşılaştırılmıştır. Burada bu tez çalışmasında da kullanılan veri kümelerinden de yararlanılmıştır. Bu yöntem de yeni bir yaklaşım olduğundan algoritması da verilmiştir. Sonuç olarak ise yüksek seviyede dengesizlik gösteren veri kümeleri için güzel sonuçlar alındığı belirtilmiştir.

Nakamura ve ark. (2013) yaptıkları çalışmada biyomedikal veriler için vektör niceleme tabanlı sentetik azınlık aşırı örnekleme tekniğini (LVQ-SMOTE) önermişlerdir. Mevcut SMOTE yönteminin biyomedikal verilere uygulanıp sınıflandırma algoritmalarında sorunlar yaşandığı bilindiğinden sentetik örnek üretme tekniği önerilmiştir. Bu yöntemde sekiz veri seti ile beş sınıflandırma algoritmasıyla testler yapılmıştır ve bu sınıflandırma algoritmalarının dördünde SMOTE algoritmasından daha iyi sonuçlar alındığı gösterilmiştir.

Sarmanova (2013) tarafından yapılan bir yüksek lisans çalışmasında iki sınıflı dengesiz veri kümeleri için sınıflandırma performanslarının karşılaştırmaları yapılmıştır. Ayrıca bu çalışmada dengesiz veri kümelerini daha performanslı

sınıflandırmak için RusAda yöntemi önerilip diğer sınıflandırma algoritmaları ile performansları karşılaştırılmıştır.

Sáez ve ark. (2015) tarafından yapılan çalışmada dengesiz sınıflandırmada gürültülü ve sınır çizgisi örnekleri probleminin filtreleme ile yeniden örnekleme yöntemi ele alınmıştır. Burada SMOTE yöntemindeki sorunlardan dolayı yinelemeli bölünme filtresi (IPF) adı verilen yinelemeli topluluk tabanlı bir yöntem önerilmiştir. Ayrıca bu filtreleme yaklaşımının da tanımlaması yapılmıştır.

Aydın (2016) tarafından yapılan bir yüksek lisans çalışmasında ise dengesiz veri kümelerinden trafik işaretlerinin tanınmasına odaklanılmıştır. Bu çalışma ile konforlu ve güvenli sürüş için bir yöntem sağlanmıştır.

Bulut (2016) çalışmasında kolektif öğrenme yöntemlerinin dengesiz dağılım gösteren veri kümeleri üzerindeki sınıflandırma başarısı analiz edilmiştir ve bu yöntemlere dair bir değerlendirme yapılmıştır.

Krawczyk (2016) derleme makalesinde ise dengesiz veri üzerinde çalışılmıştır. Dengesiz veri kümeleri üzerine uzun yıllardır yapılan çalışmalardan, bu konu ile ilgili olarak şu an da bilenen sorunlardan ve çözüm yöntemlerinden ve gelecekteki dengesiz veri üzerine yapılacak çalışmaların yönelimlerinden bahsedilmiştir. İlk olarak dengesiz verilerden öğrenme yapabilmek için üç ana yöntemden bahsedilmiştir. Bunlar veri seviyesinde yöntemler, algoritma seviyesinde yöntemler ve bu iki yöntemin avantajlarını ele alan karma metotlardır. Daha sonrasında gerçek hayat üzerindeki dengesiz veri problemlerine örnek verilmiştir: Nadir veya az sıklıkta yapılan aktivitelerin tespiti, kanser şiddetinin analizi, endüstriyel makinelerde arıza tespiti, video işlemi gibi farklı konularda birçok örnek bulunmaktadır. Dengesiz veri kümelerinde ikili sınıflar ile ikiden fazla sınıflı veri kümeleri ve bunlar ile ilgili yapılan çalışmalardan bahsedilmiştir. İkili sınıflar olarak hasta veya hasta değil, zararlı veya zararsız gibi, ikiden fazla sınıflar için ise yeterince gelişmiş bir çalışma yapılmamıştır. Sonuç olarak ise dengesiz verilerin hala gelişmekte olan bir konu olduğundan ve ileriki çalışmalarda dikkat edilmesi gereken noktalardan bahsedilmiştir.

Wang ve ark. (2016) tarafından yapılmış çalışmada dengesiz veri kümeleri üzerinde derin sinir ağları eğitimi üzerine çalışılmıştır. Burada önerilen yeni yöntem ise ortalama hata karesi yönteminin geliştirilmiş sürümü olan ortalama yanlış hata karesi yöntemidir. Önerilen bu yöntem ile hem çoğunluk hem de azınlık sınıfı için sınıflandırma hatalarının eşit olarak yakalanabilmesi amaçlanmıştır.

Douzas ve Bacao (2017) tarafından yapılmış çalışmada ise yeni bir aşırı örnekleme yöntemi olan kendi kendini düzenleyen harita aşırı örneklemesi yöntemi önerilmiştir. Bu yöntem üç aşamadan oluşmaktadır. Kendi kendini düzenleyen harita iki boyutlu olarak oluşturulur, küme içi sentetik örnekler üretilir ve kümeler arası sentetik örnekler üretilmektedir. Yirmi altı veri kümesi ve iki tane sınıflandırma algoritması ile test edilip diğer yöntemlerden başarılı sonuçlar alındığı gözlemlenmiştir.

Ma ve Fan (2017) tarafından yapılan çalışmada CURE-SMOTE algoritmasını ve özellik seçimli ve parametre tabanlı rastgele orman algoritması üzerinde karma algoritması önerilmiştir. Bu önerilen yöntem önceden yapılan SMOTE algoritmalarıyla karşılaştırılmıştır. Önerilen CURE-SMOTE algoritması orijinal veri kümesi dağılımına daha yakın olduğundan sınıflandırma sonuçlarında da başarılı olmuştur ve karma algoritma ise orijinal rastgele orman algoritmasından daha performanslı olduğu gösterilmiştir.

Haklı (2018) tarafından yapılan bir biyoistatistik alanı bütünleşik doktora tezinde dengesiz veri kümelerinin performansı üzerine çalışılmıştır. Burada farklı algoritmalar ile senaryolar çok sayıda tekrarlanmıştır.

Ankara (2019) tarafından yapılan bir yüksek lisans çalışmasında kredi talepleri için alınan kredinin analizi ve müşterinin bu krediyi geri ödemesiyle ilgili olarak kredi değerlendirme alanında bir çalışma yapılmıştır. Burada da dengesiz veri kümeleri üzerinde kolektif öğrenme yöntemlerinin başarıları karşılaştırılmıştır.

Douzas ve Bacao (2019) yapmış oldukları çalışmada SMOTE yönteminin geliştirilmiş sürümü olan geometrik SMOTE (G-SMOTE) önerilmiştir. Seçilen azınlık örnek çevresinde ve giriş alanının geometrik bölgesinde sentetik örnekler üretilmektedir. Bu bölge ise hiperküre olarak tanımlanmıştır. Bu çalışmanın sonunda ise G-SMOTE algoritmasının uygulaması Python programlama dilinde sunulmuştur.

Fotouhi ve ark. (2019) tarafından yapılan çalışmada dengesiz verilerde kanser teşhisi üzerine kapsamlı bir analiz yapılmıştır. Bu çalışmada on bir aşırı örnekleme ve yedi tane de çoğunluk örneği azaltma tekniği kullanılmıştır. Sonuç olarak her veri kümesi için her tekniğin sonuçları listelenmiştir.

Gümüştaş (2019) tarafından yapılan bir yüksek lisans çalışmasında veri kümelerindeki kayıp veriler üzerine çalışılmıştır ve kayıp verileri dolduran yöntemlerin sınıflandırma performansında da dengeli veri kümeleri üzerinde daha iyi sonuçlar verdiği gözlemlenmiştir. Bundan dolayı dengesiz veri kümeleri üzerinde de kayıp

gözlem içeren dengesiz veri kümelerine odaklanılmış ve sınıflandırma başarıları da ele alınmıştır.

Turhan (2019) tarafından yapılan bir yüksek lisans çalışmasında ise dengesiz veri kümeleri üzerinde yeniden örnekleme yöntemi kullanılarak hastalık tanılarının sınıflandırılması üzerine çalışılmıştır. Burada sınıflandırma için kolektif öğrenme yöntemlerinden yararlanılmıştır.

Sağlam (2020) tarafından yapılan bir yüksek lisans çalışmasında aşırı örnekleme ve SMOTE yönteminin sorunları ele alınıp bu sorun için yeniden örnekleme yöntemi önerilmiştir. Önerilen bu yönteme ise boosting ile SMOTE adı verilmiştir.

.

3. MATERYAL VE YÖNTEM

Bu bölümde dengesiz veri kümelerinden, dengesiz veri kümelerindeki sorunların çözümünde yer alan yöntemlerden, yapay sinir ağı sistemlerinden, sınıflandırma konusundan, veri kümelerinden, yazılım geliştirme ortamından ve tez çalışmasında kullanılan yöntemlerden bahsedilmiştir.

3.1. Dengesiz Veri Kümeleri

Teknik olarak bakılırsa sınıfları arasında eşit olmayan bir dağılım gösteren herhangi bir veri kümesi dengesiz veri kümesi olarak kabul edilmektedir. Veri kümesinin sınıfları arasındaki dengesizlik oranı birbirine yakın olmakla beraber aralarında çok fazla bir fark da bulunabilmektedir. Örnek olarak tıp alanında çok nadir görülen bir hastalık için, hastalığın görülme oranı 1000:1, 10000:1, 100000:1 gibi oranlarında olabilirken bir ülkenin, şehrin vb. yerleşim yerlerinde ise kadın ve erkek nüfus oranı birbirine yakın olmaktadır.

Dengesiz veri kümelerinde sınıflar sadece hasta ve hasta değil, erkek ve kadın gibi iki çeşitten değil ikiden fazla da olabilmektedir. Yani veri kümeleri çoklu sınıflardan oluşabilmektedir. Örnek olarak kırmızı, yeşil ve maviden oluşan bir veri kümesinde kullanıcıdan girdi olarak alınan bir rengin hangi renge daha yakın olduğu tutulabilmektedir. Burada görüldüğü üzere veri kümesinde ikiden fazla sınıf yani kırmızı, yeşil ve mavi bulunmaktadır.

Tıp alanındaki veri kümesini tekrar incelendiğinde: Bu veri kümesinde 10000 negatif yani hasta olmayanlar ve 10 pozitif yani hasta olanlar olmak üzere 10010 kayıt var olduğu kabul edilsin. Burada negatif sınıfına çoğunluk sınıfı ve pozitif sınıfına ise azınlık sınıfı adı verilmektedir. Hem çoğunluk hem de azınlık sınıfı için ideal ve dengeli bir sınıflandırma yöntemine ihtiyacımız vardır. Ne yazık ki gerçek hayatta sınıflandırıcılar çoğunluk sınıfını yüksek bir doğruluk ile sınıflandırırken azınlık sınıfını ise neredeyse sınıflandıramayacak kadar düşük bir başarı oranına sahiptir. Burada görülen en önemli problem ise pozitif hastalarda sınıflandırma başarısızlığından dolayı hasta değilmiş gibi sınıflandırılmasıdır. Çizelge 3.1'de bu duruma örnek olarak bir sınıflandırıcıdan alınan sonuçların karmaşıklık matrisi verilmiştir. Toplam pozitif hasta sayısı 80 olmasına rağmen sadece 4 hasta pozitif olarak sınıflandırılabilmiştir. Görüldüğü üzere sınıflandırma oranı %5'tir ve çok düşük bir oranı elde edilmiştir.

Çizelge 3.1. Örnek bir sınıflandırıcıdan alınmış sonuçların karmaşıklık matrisi

		Tahmini Veriler	
Gerçek Veriler		Negatif	Pozitif
	Negatif	9642	6
	Pozitif	76	4

Dengesiz veri kümeleri iki gruba ayrılmaktadır. İlk grup içsel dengesiz veri kümeleri olarak adlandırılır ve ikinci gruptan daha yaygındır. Dengesizlik veri kümesinde bulunmaktadır yani herhangi bir nedenden dolayı dengeli olan bir veri kümesi dengesiz veri kümesine dönüşmemiştir.

İkinci grup ise dışsal dengesiz veri kümeleri olarak adlandırılır. Zaman ve saklama gibi faktörler etki eder.

Dengesiz veri kümeleri ile ilgili en büyük sorunlardan birisi de sınıflandırma problemidir. Sınıflandırma algoritmaları dengesiz veri kümelerinde çoğunluk sınıfı yönünde bir başarı gösterirken azınlık sınıfını yönünde ise bir başarısızlık durumu bulunmaktadır. Bunun sebebi ise azınlık sınıfının verilerinin sayısının az olmasından dolayıdır. Kaynak araştırması bölümünde de belirtildiği gibi dengesiz veri kümelerinin sınıflandırılması üzerine çalışmalar yapılmaktadır. Bu yöntemler örnekleme yöntemleri, maliyet duyarlı yöntemler, çekirdek tabanlı yöntemler, aktif öğrenme yöntemleri ve diğer yöntemler olarak kategorilendirilmiştir (He ve Garcia, 2009). Örnekleme yöntemleri:

- Rastgele aşırı örnekleme ve alt örnekleme
- Bilgilendirilmiş alt örnekleme
- Veri üretimi ile sentetik örnekleme
- Uyarlanabilir sentetik örnekleme
- Veri temizleme ile örnekleme
- Kümeleme tabanlı örnekleme
- Örnekleme ve arttırma entegrasyonu

Maliyet duyarlı yöntemler:

- Uyarlamalı arttırma ile maliyete duyarlı veri alanı ağırlığı
- Maliyet duyarlı karar ağaçları
- Maliyet duyarlı sinir ağları

Çekirdek tabanlı yöntemler:

- Örnekleme yöntemiyle entegrasyon
- Çekirdek modifikasyon metotları

Ve aktif öğrenme yöntemleri olarak kategorize edilmiştir.

3.1.1. Dengesiz veri kümelerinde örnekleme yöntemleri

Dengesiz öğrenme uygulamalarında örnekleme yöntemlerinin kullanımı, dengeli dağılım sağlamak için dengesiz bir veri kümesinin değiştirilmesinden oluşmaktadır. Yapılan çalışmalar, birkaç temel sınıflandırıcı için dengeli bir veri setinin, dengesiz bir veri setine kıyasla gelişmiş genel sınıflandırma performansı sağladığını göstermiştir. Bu sonuçlar, dengesiz öğrenme için örnekleme yöntemlerinin kullanımını haklı çıkarmıştır (He ve Garcia).

3.1.1.1. Rastgele aşırı örnekleme ve rastgele alt örnekleme

Rastgele aşırı örnekleme tekniğinde azınlık sınıfından rastgele olarak seçilen örnekler orijinal veri kümesine eklenerek azınlık ve çoğunluk örnekleri arasında denge sağlanmaya çalışılmaktadır. Bu yöntemin dezavantajı birden çok örnek aynı veri kümesinde bulunacağından aşırı öğrenmeye sebep olmaktadır.

Rastgele alt örnekleme yönteminde ise çoğunluk sınıfından rastgele olarak seçilen örnekler orijinal veri kümesinden çıkarılarak azınlık ve çoğunluk örnekleri arasında denge sağlanmaya çalışılmaktadır. Bu yöntemin dezavantajı ise karar vermede önemli olabilecek verilerin veri kümesinden silinmesine yol açabilmektedir.

3.1.1.2. Sentetik Azınlık Veri Üretme Tekniği (SMOTE)

Dengesiz dağılım gösteren veri kümelerinin dengeli hale getirilip sınıflandırma sonuçlarının iyileştirilmesi için en çok kullanılan ve başarılı yöntemlerdendir (Chawla ve ark., 2002). Veri kümesindeki azınlık örnek sayısını istenilen oranda arttırmak için kullanılır.

1. Girdi: X_{nxp}, Y_{nx1} ve K komşu kayısı
2. $i = 1,2, \ldots, p$ için X için

   $minler \leftarrow \min(x_{*i})$ leri içeren vektör

   $makslar \leftarrow \max(x_{*i})$ leri içeren vektör

   $$X_{*i} = \frac{x_{*i} - minler_i}{makslar_i - minler_i}$$

   hesaplanır ve tüm değişkenler için maksimum ve minimum değerler tutulur. NOT: Bu dönüşümü yapmanın amacı daha gerçekçi yakın komşular elde etmektedir.
3. $N_{sentetik} \leftarrow N_{neg} \leftarrow N_{poz}$
4. $C \leftarrow |\frac{N_{sentetik}}{N_{poz}}|$ değerlerinden oluşan N_{poz} uzunluğunda vektör.
5. C içerisinde rastgele $N_{sentetik} - toplam(C)$ gözlemi 1 artır.

   NOT: Tam denge sağlandı.
6. $i = 1$'den N_{poz}'a kadar

   $k_i \leftarrow X$ içerisinde X_i'ye kendisi dışındaki en yakın K komşunun indeksleri

   NOT: k, $N_{poz} \times K$'lık bir matris.
7. $X_{smote} \leftarrow N_{sentetik} \times p$'lik boş matris.
8. $i = 1$'den N_{poz}'a kadar
   a. $eyklar.id \leftarrow 1$ ile K arasında C_i adet sayı çek
   b. $eyklar \leftarrow k_i$ içerisinden $eyklar.id$ indeksli komşulara ait gözlemleri yerine koymalı şekilde çek
   c. $\delta \leftarrow 0$ ile 1 arasında C_i adet sayı türet
   d. $x_{yeni} = x_i + (eyklar - x_i) \times \delta$
   e. x_{yeni}'yi x_{smote}'a yerleştir.
9. $X_{yeni} \leftarrow X_{smote}$ ile X'i birleştir.
10. $N_{yeni} = N + N_{sentetik}$
11. $i = 1$'den p'ye kadar X_{yeni} için

    $$X_{*i} = X_{*i} \times (makslar_i - minler_i) + minler_i$$

    şeklinde 2.adımda uygulanan dönüşüm geri alınır.
12. Çıktı: X_{yeni}

Şekil 3.2. SMOTE yöntemi sözde kodu (Sağlam, 2020)

Şekil 3.2'de çalışma adımları sözde kodu gösterilmiş bu tekniğin kısaca çalışma adımları aşağıdaki gibidir:

- Özellik vektörü ve bir örneğe en yakın komşu belirlenir.

- İkisi arasındaki fark alınır.
- Aradaki fark 0 ve 1 arasında rastgele bir sayı ile çarpılır.
- Özellik vektörüne bir önceki adımda bulunan sayı eklenir.
- İstenilen örnek sayısı kadar bu aşamalar tekrarlanır.

Bu tekniğin avantajları:

- Tekniğin anlaşılması ve programlanması kolaydır.
- Aşırı öğrenme yani aynı veriyi sürekli öğrenip ezberleme sorununu azaltır.

3.1.1.3. Uyarlanabilir sentetik örnekleme

Bu yöntemde SMOTE algoritmasının komşularını dikkate almadan, her azınlık örneği için aynı sayıda sentetik veri üretmesinden kaynaklanan sorunu ortadan kaldırmak için kullanılan bir yöntemdir. Bu yöntem sınır güvenli SMOTE yöntemidir. Burada örnek üretilmemesi gereken bir tehlikeli alan bulunmaktadır (Han ve ark).

Dengesiz veri kümeleriyle ilgili yapılan çalışmaların değerlendirilmesi için bazı yöntemlere ihtiyaç bulunmaktadır. Yöntemler ve sınıflandırma hakkında bilgiler ise bölüm 3.3'te açıklanmıştır.

3.2. Yapay Sinir Ağları

İnsanda bulunan sinir sisteminden esinlenerek geliştirilmiş makine öğrenmesi tekniklerinden bir tanesidir. Bu teknikte kurallar ve bağlantılar ile yapay sinir ağı bölümleri arasında iletişim sağlanmaktadır. Örneğin girişinden bir veriyi alarak eğitilmiş sistemi uyarladığımızda bu sistem kuralları ile bu veriyi ne yapacağına karar verip işlemektedir. Yapay sinir ağı sistemlerinin de avantaj ve dezavantajları bulunmaktadır. Avantajlarından bazıları:

- Bilgiler ağın tamamında ve dağıtık hafıza yapısında tutulur.
- Karmaşık ilişkileri öğrenme ve modelleme yeteneğine sahiptir.
- İlk öğrenme işleminden sonra bir örüntü oluşturup ağ üzerinde genelleştirme yapılabilir.
- Kendi kendine öğrenme sürecine sahiptir.
- Belirli bir hata oranına sahiptir.

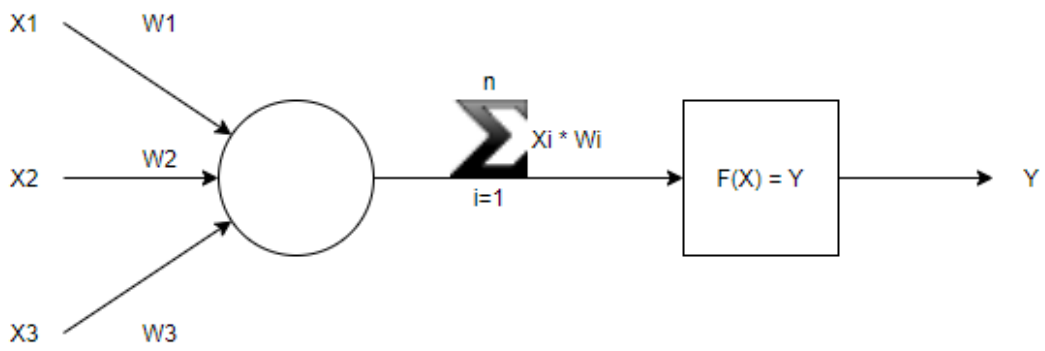
Bazı dezavantajları ise:

- Sistem kaynaklarına bağımlıdır.
- Eğitim sürecinde zorluklar yaşanmaktadır.
- Ağa bilgiler verilmeden önce sayısal değerlere çevrilmelidir.

Yapay sinir ağları beş bölümden oluşmaktadır. Bu bölümler:

- Girdi
- Ağırlık
- Toplama fonksiyonu
- Aktivasyon fonksiyonu
- Çıktı

3.2.1. Yapay sinir ağı bölümleri



Şekil 3.3. Toplam fonksiyonu ile yapay sinir hücresi yapısı

Şekil 3.3'te bir yapay sinir ağı hücresinin yapısı verilmiştir. Burada bir hücre, gelen bilgiyi alıp ağırlıklar ile girişler arasında toplam fonksiyonuna göre net girdi elde eder. Bu net girdiyi ise bir aktivasyon fonksiyonuna verip çıktısını üretir. Örnekte aktivasyon fonksiyonu olarak toplama kullanılmıştır. Burada ağırlıklar giriş verilerinin hücreye olan etkisini tanımlamaktadır.

Çizelge 3.2. Toplama fonksiyonları

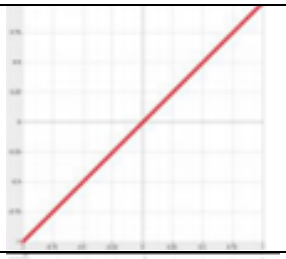
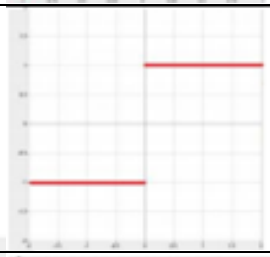
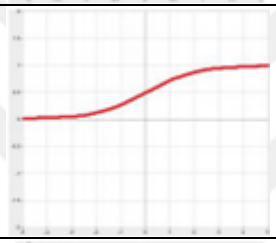
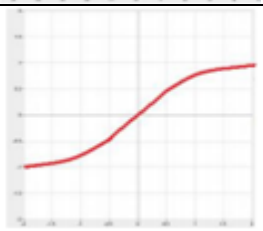
Fonksiyon	**Formül**	**Açıklama**
Toplam	$Net = \sum_{i=1}^{N} Xi * Wi$	Ağırlık ve giriş verisi çarpılıp sonuçlar toplanır.
Çarpım	$Net = \prod_{i=1}^{N} Xi * Wi$	Ağırlık ve giriş verisi çarpılıp sonuçlar çarpılır.
Maksimum	$Net = Max(Xi * Wi)$	Ağırlık ve giriş verileri ayrı ayrı çarpılıp en yüksek olan alınır.
Minimum	$Net = Min(Xi * Wi)$	Ağırlık ve giriş verileri ayrı ayrı çarpılıp en düşük olan alınır.
Kümülatif Toplam	$Net = Net(önceki) + \sum_{i=1}^{N} Xi * Wi$	Eski net verisi ile yeni bulunan değer toplanır.

Toplama fonksiyonları ise üretilen bu net girdinin işlenip bir çıktı elde edilmesini sağlamaktadır. Toplama fonksiyonları çeşitleri ve detayları çizelge 3.2'de açıklanmıştır.

- Toplam fonksiyonu: Ağırlık ve giriş verisi çarpılıp sonuçlar toplanır.
- Çarpım fonksiyonu: Ağırlık ve giriş verisi çarpılıp sonuçlar çarpılır.
- Maksimum fonksiyonu: Ağırlık ve giriş verileri ayrı ayrı çarpılıp en yüksek olan alınır.
- Minimum fonksiyonu: Ağırlık ve giriş verileri ayrı ayrı çarpılıp en düşük olan alınır.
- Kümülatif toplam fonksiyonu: Eski net verisi ile yeni bulunan değer toplanır.

Aktivasyon fonksiyonları ise net girdinin hesaplanması için kullanılan çeşitli fonksiyonlardır. Doğrusal aktivasyon fonksiyonu, adım aktivasyon fonksiyonu, sigmoid aktivasyon fonksiyonu, tanjant hiperbolik aktivasyon fonksiyonu, eşik değer aktivasyon fonksiyonu ve sinüs aktivasyon fonksiyonları yer almaktadır. Bu aktivasyon fonksiyonlarına ait detaylar çizelge 3.3'te verilmiştir.

Çizelge 3.3. Aktivasyon fonksiyonları (Çayıroğlu, 2020)

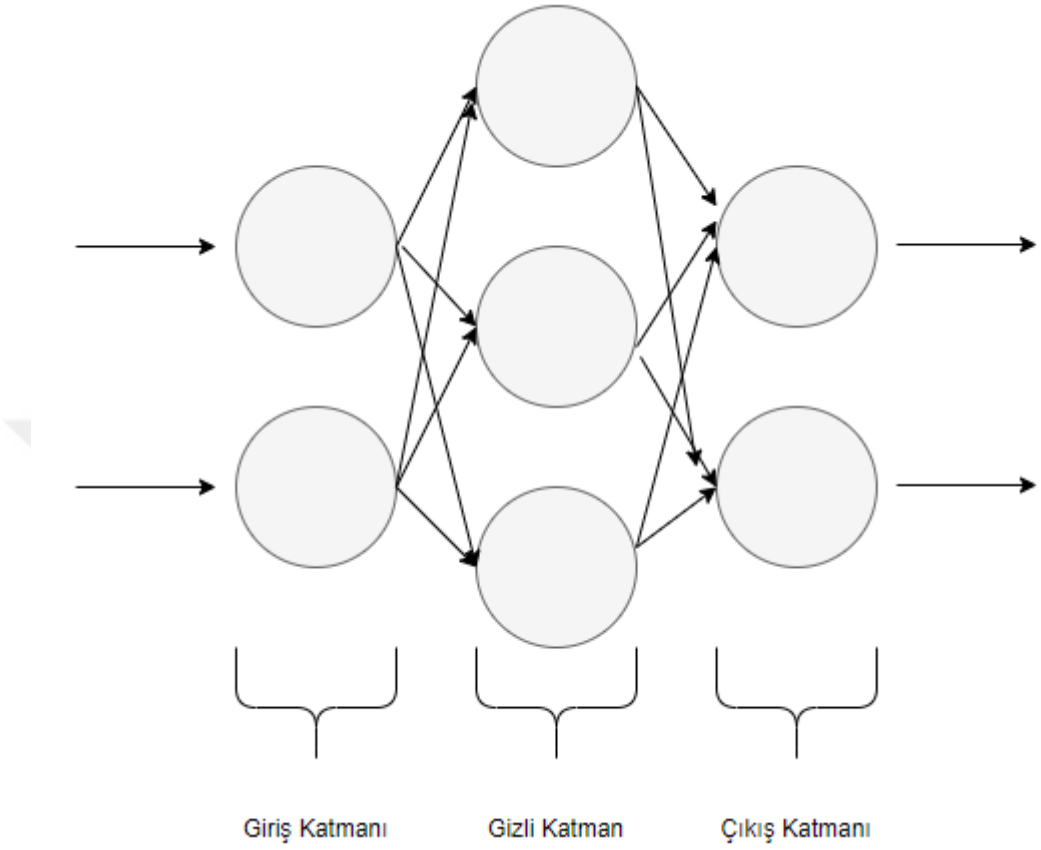
Aktivasyon Fonksiyonu	**Grafik**	**Formül**	**Açıklama**
Doğrusal		$F(Net) = A * Net$ (A sabit bir sayı)	Doğrusal problemler çözmek amacıyla aktivasyon fonksiyonu doğrusal bir fonksiyon olarak seçilebilir.
Adım		$F(Net) = \begin{cases} 1 \; if \; Net > Eşik \; Değer \\ 0 \; if \; Net \leq Eşit \; Değer \end{cases}$	Gelen net girdinin belirlenen bir eşik değerin altında veya üstünde olmasına göre hücrenin çıktısı 1 veya 0 değerini alır.
Sigmoid		$F(Net) = \frac{1}{1 + e^{-Net}}$	Sürekli ve türevi alınabilir bir fonksiyondur. Yapay sinir ağı uygulamalarında en sık kullanılan fonksiyondur.
Tanjant Hiperbolik		$F(Net) = \frac{e^{Net} + e^{-Net}}{e^{Net} - e^{-Net}}$	Sigmoid fonksiyonuna benzerdir. Değer aralığı sigmoid ek olarak 0 ile -1 arasında da değer alabilir.
Eşik Değer		$F(Net) = \begin{cases} 0 \; if \; Net \leq 0 \\ Net \; if \; 0 < Net < 1 \\ 1 \; if \; Net \geq 1 \end{cases}$	Gelen bilgilere göre 0,1 ve kendisi olabilen bir fonksiyondur.
Sinüs		$F(Net) = Sin(Net)$	Öğrenilmesi düşünülen olayların sinüs fonksiyonuna uygun dağılım gösterdiği durumlarda kullanılır.

3.2.2. Yapay sinir ağı çeşitleri

Yapay sinir ağları yapılarına, öğrenme algoritmalarına ve öğrenme zamanına göre olmak üzere üç kategoride çeşitlenmektedir.

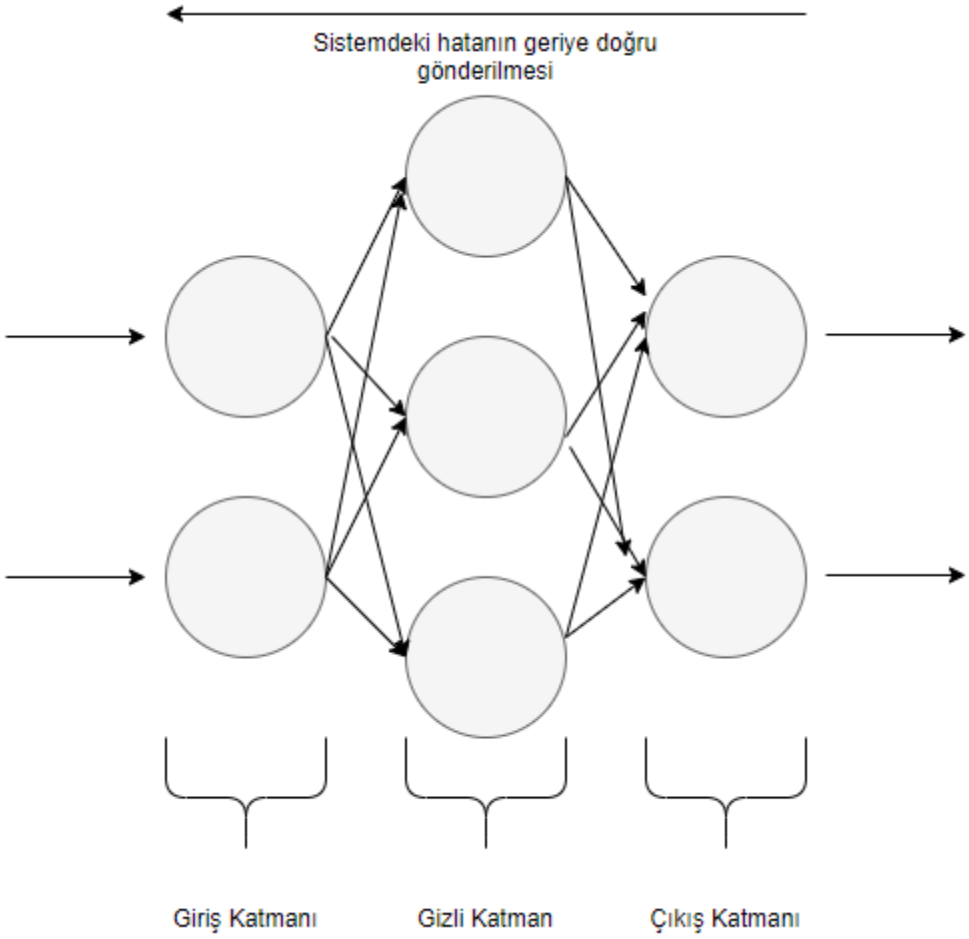
Yapılarına göre yapay sinir ağları ileri beslemeli ve geri beslemeli olarak ikiye ayrılmaktadır. İleri beslemeli yapay sinir ağlarında bir hücre bir bilgiyi ileri doğru

iletirken, geri beslemeli yapay sinir ağlarında ise sadece ileri yönlü değil geri yönlü de bir bilgi iletimi olmaktadır.



Şekil 3.4. Örnek ileri beslemeli yapay sinir ağı modeli

Şekil 3.4'te 1 giriş katmanı, 1 gizli katman ve 1 çıkış katmanı yer alan ileri beslemeli yapay sinir ağı gösterilmiştir. Bu yapay sinir ağı modelinde giriş katmanında alınan bilgi tek yönlü olarak çıkış katmanına doğru iletilmektedir.



Şekil 3.5. Örnek geri beslemeli yapay sinir ağı modeli

Şekil 3.5.'te ise 2 giriş katmanı, 3 gizli katman ve 2 çıkış katmanı yer alan geri beslemeli yapay sinir ağı gösterilmiştir. Burada sistemde bir hata bulunmaktadır ve bu hata geri yönlü olarak sisteme iletilmekte ve tekrar sistem ileri yönlü olarak çalışmaktadır.

Yapay sinir ağları öğrenme algoritmalarına göre ise danışmanlı, danışmansız ve destekleyici öğrenme olarak üçe ayrılmaktadır. Danışmanlı yapay sinir ağı modelinde ağa verilen her girişin çıktısı da sisteme verilmektedir. Böylece ağın kendi ağırlıklarını ayarlaması sağlanmaktadır. Danışmansız yapay sinir ağı modelinde girişlere karşılık bir çıktı verilmez. Son kategori destekleyici yapay sinir ağı modelinde ise her çıkış değerlerine karşılık bir puanlama yapılmaktadır.

Yapay sinir ağları öğrenme zamanına göre ise statik ve dinamik olarak ikiye ayrılmaktadır. Statik yapay sinir ağı modelinde sistem bir kez eğitilir ve sürekli o sistem

kullanılırken dinamik yapay sinir ağı modelinde eğitilmiş olan sistem kullanılırken tekrar tekrar kendini eğitmesi ve güncellenmesi sağlanmaktadır.

3.3. Sınıflandırıcı Değerlendirme Metrikleri

Makine öğrenmesi, yapay zekâ gibi bir yöntem ile eğitilmiş veri kümelerinin dışarıdan test örneği verilerek hangi sınıfa ait olduğunun tahmin edilmesini sağlayan algoritmalara sınıflandırıcılar denir. Sınıflandırıcılar kategorilere ayrılır. Bu kategoriler çok sayıdadır. Ağaçlar, kurallar, fonksiyonlar gibi kategorilere örnekler verilebilmektedir. Bu çalışmada kullanılan sınıflandırma algoritmaları da açıklanmıştır.

Sınıflandırma sonuçlarının yorumlanabilmesi için eğitilmiş sistemin test edilmesi gerekmektedir. Eğitim veri kümesinin test seti olarak kullanılması, test veri kümesi sağlanması, çapraz doğrulama ve yüzdelik bölme yöntemleri kullanılmaktadır.

Elde edilmiş bu başarı oranlarının değerlendirilebilmesi için de bazı kavramlar yer almaktadır.

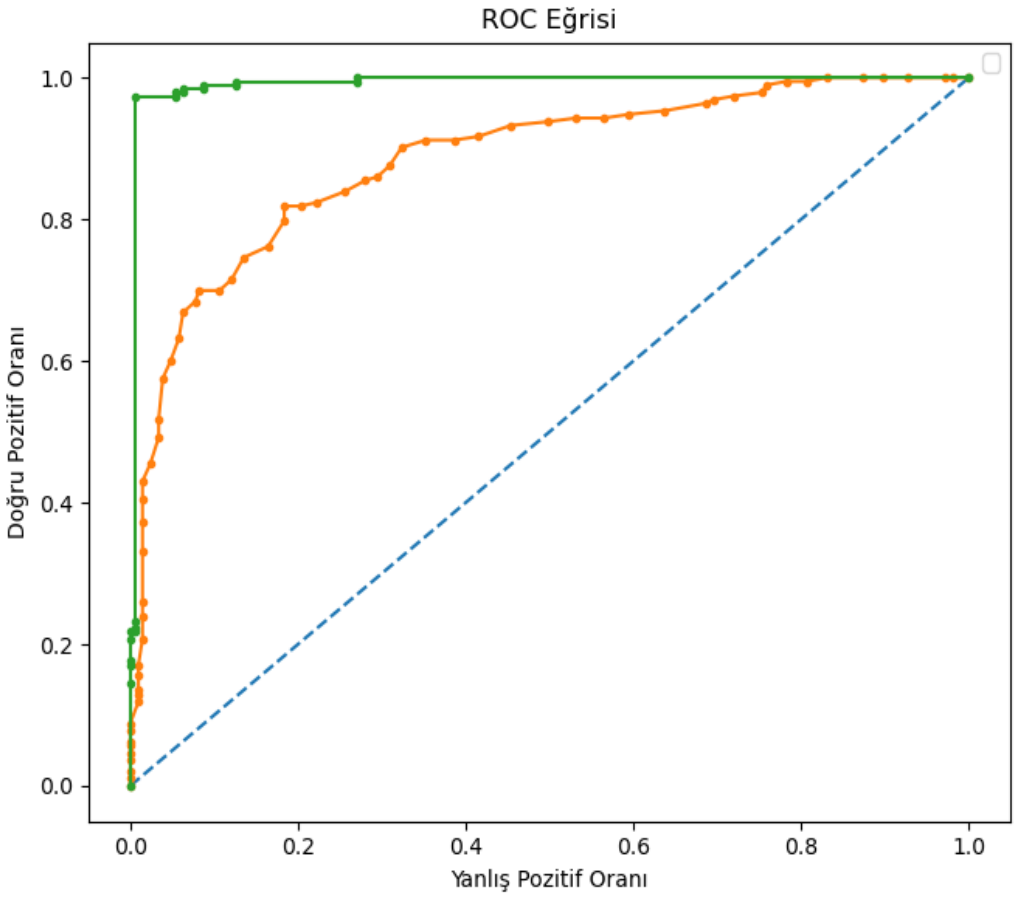
Çizelge 3.4. Karmaşıklık matrisi

		Tahmini Veriler	
Gerçek Veriler		Pozitif	Negatif
	Pozitif	TP	FN
	Negatif	FP	TN

Çizelge 3.4'te karmaşıklık matrisi gösterilmiştir. Buradaki kavramları adım adım açıklanmıştır. Gerçek veriler veri kümesinde bulunan gerçek sınıf verilerini temsil etmekte iken tahmini değerler eğitilmiş sistemin tahmin ettiği sınıf verilerini temsil etmektedir. TP kavramı doğru pozitif yani gerçekte pozitif olup sınıflandırma sonucunda da pozitif olarak bulunanları ifade eder. FP kavramı yanlış pozitif yani gerçekte negatif olup sınıflandırma sonucunda pozitif olarak bulunanları ifade eder. FN kavramı yanlış negatif yani gerçekte pozitif olup sınıflandırma sonucunda negatif olarak bulunanları ifade eder. TN kavramı doğru negatif yani gerçekte negatif olup sınıflandırma sonucunda da negatif olarak bulunanları ifade eder. Örnek vermek gerekirse bir kişi hasta ve hasta olarak tahmin edildi ise TP, bir kişi hasta değil ama

hasta olarak tahmin edildi FP, bir kişi hasta ama hasta değil olarak tahmin edildi FN ve bir kişi hasta değil ve hasta değil olarak tahmin edildi TN hücresinde yer alır.

Diğer değerlendirme ölçütü ise ROC eğrisidir. Bu 2 boyutlu olarak çizilen bir grafiktir. X ekseninde yanlış pozitifler (FP) ve Y ekseninde gerçek pozitifler (TP) yer alır. Burada değerlendirme için eğri altında kalan alan (AUC) kullanılmaktadır. Eğri altında kalan 1'e yaklaştıkça sınıflandırma başarısı da artmaktadır.



Şekil 3.6. Örnek ROC Eğrisi

Doğruluk değeri, doğru olarak tahmin ettiğimiz hücrelerin toplamının toplam veri setindeki örnek sayısına bölümüdür.

$$\text{Doğruluk} = (TP + TN) / (TP + TN + FP + FN) \quad (3.1)$$

Hata oranı, değeri ve doğruluk değerinin toplamı birdir.

Hata Oranı = 1-Doğruluk (3.2)

Kesinlik, pozitif olarak atanmış verilerin başarı oranını belirtir.

Kesinlik = TP / (TP + FP) (3.3)

Hassasiyet gerçek pozitif verilerin sınıflandırma başarısını belirtir.

Hassasiyet = TP / (TP + FN) (3.4)

Seçicilik gerçek negatif verilerin sınıflandırma başarısını belirtir.

Seçicilik = TN / (TN + FP) (3.5)

F1 skoru kesinlik ve hassasiyetin harmonik ortalamasıdır.

F1 Skoru = 2 * Kesinlik * Hassasiyet / (Kesinlik + Hassasiyet) (3.6)

3.3.1. AdaBoost.M1 sınıflandırma algoritması

Adaboost sınıflandırma algoritması zayıf olarak adlandırılan farklı sınıflandırma algoritmalarının hata oranları dikkate alınıp ağırlıklandırma yöntemi ile daha iyi sonuçlar alınmasını sağlayan bir uyarlamalı boosting algoritmasıdır. Freund ve Robert tarafından 1996 yılında önerilmiştir.

3.3.2. K en yakın komşu sınıflandırma algoritması

Bir test veri kümesindeki verilerin hangi sınıfa ait olduğunu bulmak için k-en yakın komşu kadar uzaklık hesaplaması yapılır. Bu en yakın uzaklık hesaplaması için Öklid, Manhattan, Minkowski gibi uzaklık ölçüsü algoritmaları kullanılır. Hesaplanan bu uzaklıklar içerisinde hangi sınıfa daha fazla uzaklık olarak yakınsa bu test verisi o sınıfa dâhil edilmektedir. Bu algoritmanın sözde kodu şekil 3.7’de verilmiştir.

- Eğitim ve test verilerini al
- En yakın komşu sayısını (k) seç
- Veri kümesindeki her test verisi için
    - Eğitim verisindeki veriler ile arasındaki uzaklıkları hesapla
    - Uzaklık hesaplamalarını küçükten büyüğe doğru sırala
    - En yakın komşu sayısı (k) kadarını al
    - Seçilen k komşu içerisinde en fazla hangi sınıfa sahipse test verisinin o sınıfa dahil et

Şekil 3.7. K en yakın komşu sınıflandırma algoritmasının sözde kodu

3.3.3. K star algoritması

Entropinin mesafe ölçüsü olarak kullanılmasının birçok faydası vardır. Sembolik nitelikleri, gerçek değerli nitelikleri ve eksik değerleri ele almak için tutarlı bir yaklaşım sağlar. K star sınıflandırma algoritması da bu tür bir entropi mesafe ölçüsü kullanan örnek tabanlı bir algoritmadır (Cleary ve Trigg, 1995).

Bu algoritma iki avantaj sunmaktadır:

1. Sadece arama ile araştırılması gereken köşeler bellekte saklanmaktadır.
2. Hedef temelli bir arama işlemi için sezgisel bir yöntem kullanmaktadır.

Bu algoritma iki arama algoritmasına dayanmaktadır:

1. En kısa yolu bulmak için A star algoritması
2. Oluşturulan bir grafik yapısında k çözüm yolları arayan Dijkstra algoritması

3.3.4. Sıralı asgari optimizasyon

Sıralı asgari optimizasyon sınıflandırma algoritması, destek vektör makinelerinin eğitimi sırasında ortaya çıkan karesel programlamanın çözümü için kullanılmaktadır. Bu karesel programlama probleminin analitik olarak çözülmektedir ve uzun süren optimizasyonu önlemektedir. Matris hesaplamalarından kaçınıldığından dolayı bu algoritma için gereken bellek miktarı lineer olup çok büyük veri kümeleri üzerinde de

çalışabilmektedir. Bu algoritma John tarafından Microsoft araştırma laboratuvarlarında geliştirilmiştir. (Platt, 1998)

Ayrıca bu algoritma, yer ve zaman verimliliğinden ödün vermeden uygulanması çok daha kolaydır ve yakınsamayı sağlama altında ideal bir regresyon doğruluğu elde edilebilmektedir. Bu nedenle belirli bir teorik ve pratik önemi bulunmaktadır.

3.4. Friedman Testi

Friedman testi, tekrarlanan ölçümlerle beraber parametrik olamayıp tek yönlü ANOVA'ya bir alternatiftir. Ölçülen bağımlı değişken sıralı olduğunda gruplar arasındaki farkları test etmek için kullanılır. Friedman testi ise karşılaştırılmak istenen sonuçlar etiketleri ve verileri ile bir veri kümesi gibi tabloya aktarılır. Sonrasında ise alınan sonuçlar satırlar karşılaştırılıp 1'den veri kümesi kolon sayısına kadar değerlendirme yapılmaktadır.

Çizelge 3.5. Friedman testi örnek atama

İndeks	Etiket1	Etiket2	Etiket3	Atama1	Atama2	Atama3
0	50	60	70	1	2	3
1	45	35	85	2	1	3
2	25	55	35	1	3	2
3	85	75	55	3	2	1
4	65	95	70	1	3	2

Çizelge.3.6 Friedman testi sonuçları

Atama	Sonuç
1	(1+2+1+3+1) / 5 = 1.6
2	(2+1+3+2+3) / 5 = 2.2
3	(3+3+2+1+2) / 5 = 2.2

Çizelge 3.5'te örnek bir atama tablosu verilmiş olup ortalama değerlendirme sonuçları ise çizelge 3.6'da gösterilmiştir.

Bu tez çalışmasında ortalama değerlendirme sonuçlarının alınabilmesi için SPSS programı kullanılmıştır.

3.5. Veri Kümeleri

Bu tez çalışmasında 6 farklı konuda toplam 10 adet veri kümesi kullanılmıştır. Bu veri kümelerinin her biri 5 kez çapraz doğrulama olarak ve her parçada özellik

sayısının bir kısmı test verisi olarak bir kısmı da eğitim verisi olarak ayrılmıştır. Bu veri kümesinin de arff formatına benzer şekilde kendine özgü dat formatı bulunmaktadır. Bu veri kümesi çalışmada ele alırken dat formatı ile csv, txt ve arff formatlarını da üretecek kod geliştirmeleri de yapılmıştır. Ayrıca veri kümelerinin pozitif sonucu 1 ile negatif sonucu 0 ile değiştirilmiştir.

```
@relation yeast1
@attribute Mcg real [0.11, 1.0]
@attribute Gvh real [0.13, 1.0]
@attribute Alm real [0.21, 1.0]
@attribute Mit real [0.0, 1.0]
@attribute Erl real [0.5, 1.0]
@attribute Pox real [0.0, 0.83]
@attribute Vac real [0.0, 0.73]
@attribute Nuc real [0.0, 1.0]
@attribute Class {positive,negative}
@inputs Mcg, Gvh, Alm, Mit, Erl, Pox, Vac, Nuc
@outputs Class
@data
0.58, 0.61, 0.47, 0.13, 0.50, 0.00, 0.48, 0.22, negative
0.43, 0.67, 0.48, 0.27, 0.50, 0.00, 0.53, 0.22, negative
0.64, 0.62, 0.49, 0.15, 0.50, 0.00, 0.53, 0.22, negative
0.58, 0.44, 0.57, 0.13, 0.50, 0.00, 0.54, 0.22, positive
0.42, 0.44, 0.48, 0.54, 0.50, 0.00, 0.48, 0.22, negative
```

Şekil 3.8. Örnek dat dosyası içeriği

3.5.1. Ecoli veri kümesi

Bu veri kümesi, içme sularından e.coli bakterisi yolu ile bulaşan bir hastalığın verilerini içeren medikal bir veri kümesidir. Burada ecoli ile ilgili sadece “ecoli-0_vs_1-5” veri kümesi kullanılmıştır. Dengesizlik oranı 1.86’dır. Bu veri kümesinde yedi öznitelik yer almaktadır. Eğitim veri kümelerinde 176, test veri kümelerinde 44 veri yer almaktadır.

3.5.2. Glass veri kümesi

Bu veri kümesinde bir maddenin cam olup olmadığı ile ilgili bilgiler yer almaktadır. Burada kullanılan “glass-0-1-2-3_vs_4-5-6” veri kümesi 3.2, “glass0” veri kümesi 2.06, “glass1” veri kümesi 1.82 ve “glass6” veri kümesi ise 6.38 dengesizlik oranına sahiptir. Bu veri kümesinde dokuz öznitelik yer almaktadır. Eğitim veri kümelerinde 171, test veri kümelerinde 43 veri yer almaktadır.

3.5.3. Haberman veri kümesi

Bu veri kümesi 1958 ve 1970 yılları arasında Chicago Üniversitesinde Billings hastanesinde meme kanseri ameliyatı geçiren hastaların hayatta kalmasıyla ilgili bir veri kümesidir. Dengesizlik oranı ise 2.78'dir. Bu veri kümesinde üç öznitelik yer almaktadır. Eğitim veri kümelerinde 245, test veri kümelerinde 61 veri yer almaktadır.

3.5.4. New thyroid veri kümesi

Bu veri kümesinde tiroit hastalığı ile ilgili bilgilerin yer aldığı medikal veri kümesidir. Bu kısımda kullanılan "new-thyroid1" ve "new-thyroid2" veri kümeleri 5.14 dengesizlik oranına sahiptir. Bu veri kümesinde ise beş öznitelik yer almaktadır. Eğitim veri kümelerinde 172, test veri kümelerinde 43 veri yer almaktadır.

3.5.5. Pima veri kümesi

Pima veri kümesi bir hastanın diyabet olup olmadığına ilişkin medikal bir veri kümesidir. Bu veri kümesinin dengesizlik oranı 1.87'dir. Ayrıca bu veri kümesinde sekiz öznitelik yer almaktadır. Eğitim veri kümelerinde 614, test veri kümelerinde 154 veri yer almaktadır.

3.5.6. Wisconsin veri kümesi

Bu veri kümesi de göğüs kanseriyle ilgili medikal bir veri setidir. Bu veri kümesinin dengesizlik oranı 1.86'dır. Burada dokuz öznitelik yer almaktadır. Eğitim veri kümelerinde 546, test veri kümelerinde 137 veri yer almaktadır.

3.6. Geliştirme Ortamı

Yapılan tez çalışmasında programlama dili, programlama dili geliştirme ortamı (IDE), yapay sinir ağının geliştirilmesi, yapay sinir ağı için kullanılan kütüphane, sınıflandırma sonuçlarının karşılaştırılması ve veri setlerinin hazırlanması için çeşitli yöntemler kullanılmıştır.

Yapılan çalışmada Python programlama dili kullanılmıştır. Python programlama dilinin ise v3 sürümü tercih edilmiştir. Bu programlama dili ile geliştirme yapabilmek için ise PyCharm IDE'si kullanılmıştır. Bu programlama dilinin kullanılması sebepleri ise sade, kolay anlaşılabilir ve geliştirme yapmak için gerekli olan kütüphanelerin sunulmuş olmasıdır.

Yapay sinir ağı geliştirmesi için Keras kütüphanesinden yararlanılmıştır. Keras Python'da yazılmış açık kaynak kodlu bir kütüphanedir. Farklı makine öğrenmesi kütüphaneleri ile de beraber çalışabilmektedir. TensorFlow, Theano gibi kütüphaneleri desteklemektedir. Bu çalışmada Google'ın sunduğu TensorFlow tercih edilmiştir. Keras tercih edilmesi sebepleri ise farklı makine öğrenme kütüphanelerini desteklemesi, modellerin kolayca oluşturulup test edilmesi ve hem CPU hem de GPU ile öğrenme seçeneklerinin olmasıdır.

Yapay sinir ağından yararlanılarak üretilen veri kümelerinin sınıflandırma başarılarının test edilmesinde WEKA kütüphanesinden yararlanılmıştır. WEKA kütüphanesi Java ile yazılmış ve ücretsiz olarak dağıtılmaktadır. Sınıflandırmaların yanı sıra veri ön işleme, görselleştirme gibi özellikleri de desteklemektedir ve birçok sınıflandırma algoritması içermektedir. Ayrıca kendine özgü arff dosya yapısını içermektedir. Bu arff dosya yapısına ait bir örnek şekil 3.9'da verilmiştir.

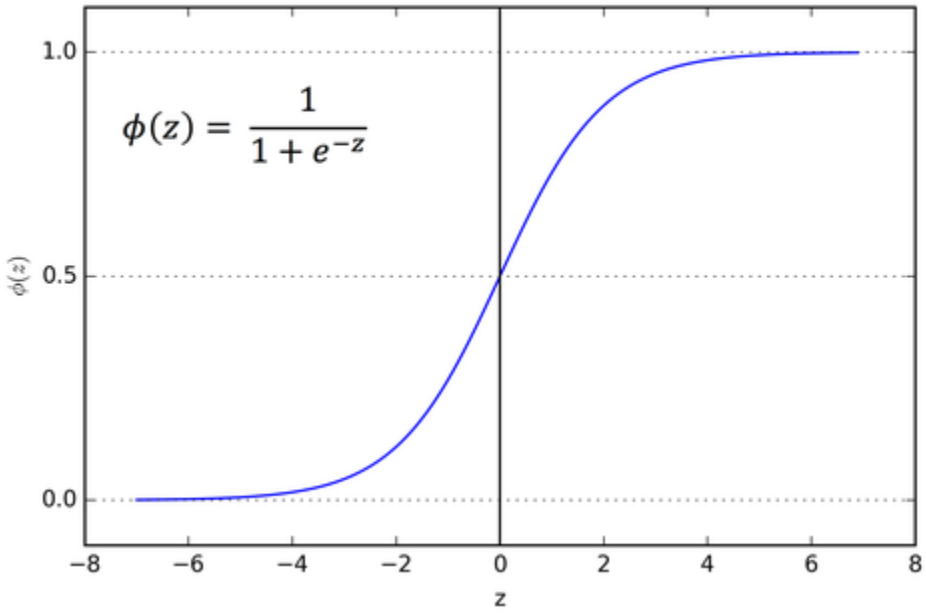
```
@relation açıklayıcı_isim

@attribute attribute1 veri_tipi
@attribute attribute2 veri_tipi
@attribute attribute3 veri_tipi
@attribute class_attribute veri_tipi

@data
data1, data2, data3, class
data1, data2, data3, class
data1, data2, data3, class
```

Şekil 3.9. Örnek arff dosya içeriği

3.7. Önerilen Yöntemler

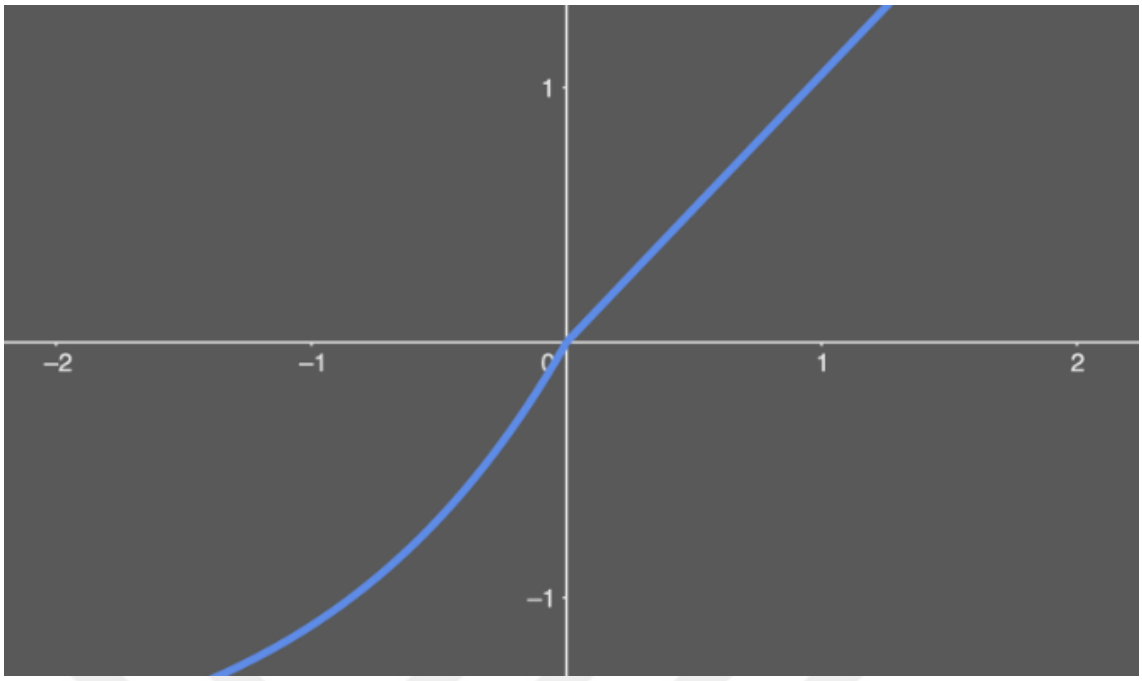
Bu tez çalışmasında ilk olarak yapay sinir ağı tasarımı gerçekleştirilmiştir. Burada tasarlanan yapay sinir ağında giriş ve çıkış katmanları ile bir ve iki adet gizli katman da kullanılmıştır. Bir adet gizli katman ile iki adet gizli katmanda alınan sonuçlarının yakınlığından ve çalışma süresinden dolayı bir adet gizli katman tercih

edilmiştir. Gizli katmandaki girdi sayısı, nitelik sayının bir fazlası ve nitelik sayısını 2 katının bir fazlasının sonucu ile karşılaştırılıp nitelik sayının bir fazlası olarak kabul edilmiştir. Yapay sinir ağı modelinde aktivasyon fonksiyonları kullanılmıştır. Aktivasyon fonksiyonları bir girdiyi işleyerek çıktı üretilmesini sağlar. Giriş ve gizli katmanda sigmoid aktivasyon fonksiyonu kullanılmıştır. Bu sigmoid fonksiyonuna ait grafik şekil 3.10'da gösterilmiştir.



Şekil 3.10. Sigmoid fonksiyon

Çıkış katmanında ise SELU aktivasyon fonksiyonu kullanılmıştır. Bu SELU fonksiyonuna ait grafik ise şekil 3.11'de gösterilmiştir.

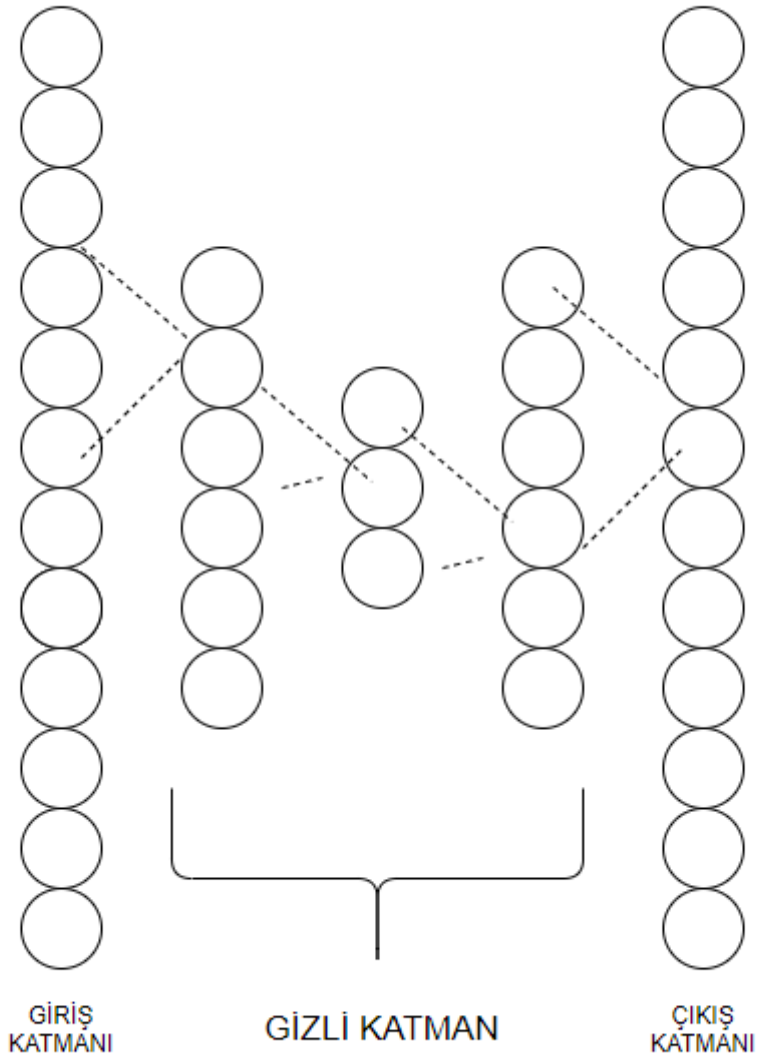


Şekil 3.11. SELU fonksiyonu

Yapay sinir ağının eğitimi aşamasında ise 1000, 2000 ve 10000 tekrar sayıları ile denemeler yapılmış ve 10000 tekrar kullanılmıştır.

ÖYRÖ'de bir veri kümesinin beş farklı dağılımında da eğitim verileri azınlık ve çoğunluk verileri olarak ikiye ayrılmıştır. Bu azınlık ve çoğunluk verilerinin her biri ile iki adet yapay sinir ağı modelleri oluşturulmuştur. Azınlık veri kümesindekilerin sayısı ile çoğunluk veri kümesindekilerin sayısı eşitlenmeye çalışılmıştır. Bu aşamada veri kümesindeki örneklerin tamamı sıfır ile bir arasına normalize edilmiştir. Bu veri kümesi dengesiz olmayacak şekilde azınlık veri kümesi eşitlenene kadar rastgele bir veriler üretilmiş ve her iki yapay sinir ağı ile de test edilip hangi yapay sinir ağı modeline yakın olduğu belirlenmiştir. Eğer azınlık sınıfına daha yakın ise azınlık veri kümesine dâhil edilmiştir. Çoğunluk ve azınlık örneklerinin sayısı eşitlendikten sonra bölüm 3.3'te belirtilen sınıflandırıcılar ile test edilmiştir.

ÖYOK'de yapay sinir ağı tasarımı değiştirilmiştir. Bu tasarıma otomatik kodlayıcı adı verilmektedir. Bu yöntem ile verinin belirli bir miktara kadar boyutunun küçültülüp tekrar açılması işlemidir. Tez çalışmasında ise gizli katmanlar üzerine uygulanmış ve burada gizli katmandaki nöron sayısı en az üç olacak şekilde kodlama yapılmıştır.



Şekil 3.12. 12 girişli veri kümesi üzerinde otomatik kodlayıcı uygulanması

Şekil 3.12'de 12 özellikli veri kümesine ait bir otomatik kodlayıcı yapısına sahip yapay sinir ağı modeli gösterilmiştir. Burada 12 girişten sonra 2 ile bölünerek en az 3 girdiye sahip olacak şekilde veri sıkıştırması yapılmaktadır. 12'den sonra 6, 6'dan sonra 3 olup en az rakama ulaşılmış ve buradan sonra açma işlemi yapılmaktadır. Tam tersi işlem ile 2 ile çarparak ilerlenmektedir. 3'ten sonra 6 ve 6'dan sonra 12 olup veri kümesindeki özellik sayısına ulaşılmıştır. Bu yöntem ile tez çalışmasında gerçekleştirilen ilk yönteme benzer olarak sadece azınlık örnekleri üzerinde yapay sinir ağı modeli oluşturulup yeni veriler oluşturulup sınıflandırma başarısı test edilmiştir.

ÖYOKED'de ise bu otomatik kodlayıcı yapısına bir fark miktarı eklenip üretilen örnekler arasında bir eşik değer belirlenmiştir. Bu fark miktarı ile üretilen veri ve üretilen verinin azınlık ve çoğunluk yapay sinir ağlarına uzaklığı karşılaştırılıp eğer uzaklıklar fark miktarından büyükse bu örnek dikkate alınmamıştır. Eğer uzaklık fark miktarından küçük ise yapay sinir ağlarında karşılaştırma yapılıp eğer azınlık sınıfına daha yakın ise veri kümesine dâhil edilmiştir. Bu fark miktarının belirlenmesinde ise veri kümesindeki örnek sayısı ve nitelik sayısı etkili olmuştur. Fark miktarı küçüldükçe benzer örnekleri üretip veri kümesinde önceden üretilen örnekler sisteme dâhil edilmeyeceğinden sistemin çalışma süresi ile ters orantılı olarak artmıştır.

ÖYT'de azınlık verileri tekrarlı olarak artırılmıştır. Örneğin 150 çoğunluk verisi ve 50 azınlık verisi var ise burada 50 azınlık verisinin üzerine 50 azınlık verisi eklenerek çoğunluk verisine ulaşılıncaya kadar bu işleme devam edilmiştir. Bu önerilen yöntem ile örnekler birbirine yakın hale getirilerek daha başarılı sonuçlar elde edilmeye çalışılmıştır. Ayrıca üretilen örnekler ile veri kümesinde olan örnekler arasında daha önce bulunup bulunmadığı kontrolü de eklenmiştir.

ÖYSA'da SMOTE yöntemi ile üretilen örnekler azınlık yapay sinir ağı modeline verilip elde edilen veriler, veri kümesine dâhil edilip veri kümesindeki azınlık ve çoğunluk örnekleri eşitlenmiştir.

ÖYSÇ'da SMOTE yöntemi ile üretilen örnekler çoğunluk yapay sinir ağı modeline verilip elde edilen veriler, veri kümesine dâhil edilip veri kümesindeki azınlık ve çoğunluk örnekleri eşitlenmiştir.

ÖYED'de ise ÖYRÖ'de yapılan senaryoya bir fark miktarı eklenmiştir. Bu fark miktarı ile üretilen veri ve üretilen verinin her iki yapay sinir ağı modeline uzaklığı karşılaştırılıp fark miktarında büyük bir sonuç varsa bu üretilen veri sisteme dâhil edilmemiştir. Bu fark miktarının belirlenmesinde ise veri kümesindeki örnek sayısı ve nitelik sayısı etkili olmuştur. Fark miktarı küçüldükçe sistemin çalışma süresi ters orantılı olarak artmıştır. Bu yöntem ise tez çalışmasında hedeflenen "dengesiz veri kümelerinde sınıflandırma başarısının" artırılması hedefine ulaştırmıştır.

4. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

Bu bölümde yapılan tez çalışmasında geliştirilen yöntemlere ait sonuçlar paylaşılmıştır. Yapılmış olan yedi yöntemin sonuçları açıklanmış ve yorumlanmıştır. Çizelge 4.1 ve çizelge 4.2'de AdaBoost.M1, çizelge 4.3 ve çizelge 4.4'te k en yakın komşu(k=3), çizelge 4.5 ve çizelge 4.6'da k star, çizelge 4.7 ve çizelge 4.8'de çok katmanlı yapay sinir ağı ve çizelge 4.9 ve çizelge 4.10'da sıralı asgari optimizasyon algoritması sonuçları açıklanmıştır. Çizelge 4.11'de ÖYRÖ, çizelge 4.12'de ÖYOK, çizelge 4.13'te ÖYOKED, çizelge 4.14'te ÖYT, çizelge 4.15'te ÖYSA, çizelge 4.16'da ÖYSÇ ve çizelge 4.17'de ise ÖYED'nin geometrik ortalama ve f ölçüsü sonuçlarının Friedman testi sonuçları açıklanmıştır.

Çizelge 4.1. AdaBoost.M1 yöntemi ile geometrik ortalama sonuçları

Veri Kümesi	**Orijinal**	**SMOTE**	**ÖYRÖ**	**ÖYOK**	**ÖYOKED**	**ÖYT**	**ÖYSA**	**ÖYSÇ**	**ÖYED**
Ecoli0vs1_5	0.97	0.97	0.97	**0.98**	0.97	0.97	0.96	0.97	0.97
Glass0123vs456	0.9	0.91	0.9	**0.92**	**0.92**	0.91	**0.92**	0.9	**0.93**
Glass0	0.8	0.78	0.76	0.68	0.8	0.77	0.79	0.78	**0.81**
Glass1	0.64	0.69	**0.7**	0.64	0.68	0.58	0.57	0.64	0.68
Glass6	0.91	0.92	0.92	0.92	0.92	**0.93**	0.91	0.92	0.89
Haberman	0.48	0.6	0.58	0	0.59	0	0.57	0.52	**0.63**
New Thyroid 1	1	0.97	0.97	0.96	0.97	0.94	0.99	1	1
New Thyroid 2	0.96	0.96	**0.99**	0.95	0.95	0.92	0.96	**0.98**	0.94
Pima	0.67	0.7	0.66	0.63	0.7	0.63	0.7	0.69	**0.71**
Wisconsin	0.93	0.94	0.94	0.92	0.94	0.92	0.92	0.93	**0.95**
Ortalama	0.826	0.844	0.839	0.76	0.844	0.757	0.829	0.833	**0.851**

Çizelge 4.1'deki AdaBoost.M1 yöntemi geometrik ortalama sonuçları incelendiğinde; ÖYRÖ glass1 ve new thyroid 2 veri kümelerinde, ÖYOK ecoli0vs1_5 ve glass0123vs456 veri kümelerinde, ÖYOKED glass0123vs456 veri kümesinde, ÖYT glass6 veri kümesinde, ÖYSA glass01234vs456 veri kümesinde, ÖYSÇ new thyroid 2 veri kümesinde ve ÖYED glass01234vs456, glass0, haberman, pima ve wisconsin veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.2. Adaboost.M1 yöntemi ile f ölçüsü sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.97	0.97	0.97	**0.98**	0.96	0.97	0.96	0.96	0.96
Glass0123vs456	0.89	0.89	**0.9**	**0.9**	**0.9**	**0.9**	**0.9**	**0.91**	**0.9**
Glass0	0.78	0.74	0.72	0.65	**0.88**	0.74	**0.88**	0.75	0.77
Glass1	0.66	0.67	**0.7**	0.64	**0.68**	0.58	0.64	0.64	0.67
Glass6	0.9	0.88	0.9	**0.91**	0.9	**0.93**	0.9	0.9	0.88
Haberman	0.5	0.56	**0.61**	0	**0.6**	0	0.52	0.52	**0.62**
New Thyroid 1	1	0.97	0.95	0.96	0.96	0.94	1	1	1
New Thyroid 2	0.97	0.97	**0.98**	0.95	0.94	0.92	0.92	**0.98**	0.95
Pima	0.69	0.69	0.67	0.63	**0.7**	0.63	0.69	0.69	**0.7**
Wisconsin	0.93	0.94	0.94	0.93	0.94	0.92	0.93	0.93	**0.95**
Ortalama	0.829	0.828	**0.834**	0.755	**0.835**	0.753	0.828	0.828	**0.84**

Çizelge 4.2'daki AdaBoost.M1 yöntemi f ölçüsü sonuçları incelendiğinde; ÖYRÖ glass0123vs456, glass1, haberman ve new thyroid 2 veri kümelerinde, ÖYOK ecoli0vs1_5, glass0123vs456 ve glass6 veri kümelerinde, ÖYOKED glass0123vs456, glass0, glass1, haberman ve pima veri kümelerinde, ÖYT glass0123vs456 ve glass6 veri kümelerinde, ÖYSA glass0123vs456 ve glass0 veri kümelerinde, ÖYSÇ glass0123vs456 ve new thyroid 2 veri kümelerinde ve ÖYED glass0123vs456, haberman, pima ve wisconsin veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.3. K en yakın komşu (k=3) yöntemi ile geometrik ortalama sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.96	0.95	0.96	0.96	0.95	0.96	0.96	0.96	**0.97**
Glass0123vs456	0.9	**0.92**	0.9	0.89	0.92	0.91	0.92	0.91	**0.93**
Glass0	0.78	0.78	0.77	0.73	0.7	0.73	**0.81**	0.78	**0.79**
Glass1	0.73	**0.78**	0.74	0.76	0.76	0.76	0.73	0.72	0.75
Glass6	0.81	0.86	0.86	**0.88**	**0.89**	0.84	0.85	0.83	**0.88**
Haberman	0.46	**0.59**	0.56	0.48	0.56	0.48	0.51	0.49	0.56
New Thyroid 1	0.95	**0.98**	0.97	0.94	0.97	0.94	0.95	0.95	0.95
New Thyroid 2	0.91	**0.98**	0.95	0.91	0.94	0.91	0.91	0.91	0.94
Pima	0.67	0.7	0.67	0.67	0.7	0.67	0.68	0.67	0.7
Wisconsin	0.95	0.96	0.96	0.96	0.96	0.96	0.95	0.96	0.96
Ortalama	0.812	**0.85**	0.834	0.818	0.842	0.82	0.827	0.818	0.843

Çizelge 4.3'teki K en yakın komşu (k=3) yöntemi geometrik ortalama sonuçları incelendiğinde; ÖYOK glass6 veri kümesinde, ÖYOKED glass6 veri kümesinde, ÖYSA glass0 veri kümesinde, ÖYED ecoli0vs1_5, glass0123vs456, glass0 ve glass6 veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.4. K en yakın komşu (k=3) yöntemi ile f ölçüsü sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.96	0.95	0.96	0.96	0.95	0.96	0.96	0.96	**0.97**
Glass0123vs456	0.9	0.9	0.89	0.89	**0.91**	**0.92**	**0.91**	0.9	**0.92**
Glass0	0.77	0.75	0.76	0.72	0.75	0.73	**0.78**	0.77	0.77
Glass1	0.74	0.77	0.75	**0.78**	0.77	0.77	0.74	0.74	0.76
Glass6	0.86	0.88	0.86	**0.89**	0.87	0.86	**0.89**	0.85	0.85
Haberman	0.48	0.54	**0.57**	0.51	0.54	0.51	0.53	0.51	**0.55**
New Thyroid 1	0.95	**0.97**	0.96	0.94	0.94	0.94	0.95	0.95	0.94
New Thyroid 2	0.93	**0.97**	0.94	0.93	0.92	0.93	0.93	0.93	0.94
Pima	0.68	0.69	0.69	0.68	**0.7**	0.68	0.69	0.68	**0.7**
Wisconsin	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Ortalama	0.822	**0.837**	0.833	0.825	0.83	0.83	0.833	0.824	0.835

Çizelge 4.4'teki K en yakın komşu (k=3) yöntemi f ölçüsü sonuçları incelendiğinde; ÖYRÖ haberman veri kümesinde, ÖYOK glass1 ve glass6 veri kümelerinde, ÖYOKED glass0123vs456 ve pima veri kümelerinde, ÖYT glass0123vs456 veri kümesinde, ÖYSA glass0123vs456, glass0 ve glass6 veri kümelerinde ve ÖYED ecoli0vs1_5, glass01234vs456, haberman ve pima veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.5. K star yöntemi ile geometrik ortalama sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.94	0.94	**0.95**	0.94	**0.95**	**0.95**	**0.95**	0.93	**0.95**
Glass0123vs456	0.82	**0.89**	0.87	0.88	0.87	0.86	0.86	0.87	0.86
Glass0	0.81	**0.82**	0.81	0.82	**0.84**	0.79	0.79	0.8	0.81
Glass1	0.79	0.81	**0.82**	0.8	**0.82**	0.75	0.75	0.8	0.79
Glass6	0.88	0.88	**0.92**	0.88	**0.92**	**0.9**	**0.9**	**0.9**	**0.91**
Haberman	0.31	0.52	**0.54**	0.27	0.52	0.52	0.52	0.51	**0.6**
New Thyroid 1	0.93	0.95	**0.97**	0.93	0.92	0.95	0.95	0.95	**0.97**
New Thyroid 2	0.93	0.97	0.97	0.93	0.94	0.95	0.95	0.95	0.97
Pima	0.61	**0.68**	0.63	0.61	0.64	0.64	0.64	0.65	0.65
Wisconsin	0.92	0.93	0.93	0.91	**0.96**	0.93	0.93	**0.94**	**0.94**
Ortalama	0.794	0.839	0.841	0.797	0.838	0.824	0.824	0.83	**0.845**

Çizelge 4.5'teki K star yöntemi geometrik ortalama sonuçları incelendiğinde; ÖYRÖ ecoli0vs1_5, glass1, glass6, haberman ve new thyroid 1 veri kümelerinde, ÖYOKED ecoli0vs1_5, glass0, glass1, glass6 ve wisconsin veri kümelerinde, ÖYT ecoli0vs1_5 ve glass6 veri kümelerinde, ÖYSA ecoli0vs1_5 ve glass6 veri kümelerinde, ÖYSÇ glass6 ve wisconsin veri kümelerinde ve ÖYED ecoli0vs1_5, glass6, haberman, new thyroid 1 ve winconsin veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.6. K star yöntemi ile f ölçüsü sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.95	0.95	0.95	0.94	0.95	0.95	0.95	0.93	0.95
Glass0123vs456	0.86	0.89	0.89	0.89	0.88	**0.9**	0.87	0.89	0.89
Glass0	0.79	0.78	0.78	**0.8**	**0.8**	0.79	0.76	0.78	0.79
Glass1	0.81	0.82	0.82	0.81	0.82	0.81	0.76	0.81	0.8
Glass6	0.91	0.89	**0.94**	0.91	0.9	0.9	0.87	0.88	**0.92**
Haberman	0.35	0.51	**0.56**	0.32	**0.52**	0.32	**0.54**	**0.52**	**0.59**
New Thyroid 1	0.94	0.95	0.95	0.94	0.91	0.91	0.93	0.93	0.95
New Thyroid 2	0.95	0.95	0.95	0.94	0.92	0.93	0.93	0.93	0.95
Pima	0.63	**0.67**	0.64	0.63	0.65	0.63	0.65	0.65	0.66
Wisconsin	0.93	0.94	0.94	0.92	**0.95**	0.92	0.94	0.94	**0.95**
Ortalama	0.812	0.835	0.842	0.81	0.83	0.81	0.82	0.826	**0.845**

Çizelge 4.6'daki K star yöntemi f ölçüsü sonuçları incelendiğinde; ÖYRÖ glass6 ve haberman veri kümelerinde, ÖYOK glass0 veri kümesinde, ÖYOKED glass0 ve haberman veri kümelerinde, ÖYOKED glass0, haberman ve wisconsin veri kümelerinde, ÖYT glass0123vs456 veri kümesinde, ÖYSA haberman veri kümesinde, ÖYSÇ haberman veri kümesinde ve ÖYED glass6, haberman ve wisconsin veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.7. Çok katmanlı yapay sinir ağı yöntemiyle geometrik ortalama sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.97	0.97	0.97	0.96	0.96	0.97	0.97	0.96	0.97
Glass0123vs456	0.89	0.9	0.89	0.9	0.89	0.9	**0.91**	0.88	0.89
Glass0	0.77	0.78	0.75	0.78	0.77	0.74	**0.81**	**0.82**	0.75
Glass1	0.62	0.7	0.7	**0.71**	0.67	**0.71**	0.68	**0.71**	**0.71**
Glass6	0.83	0.86	**0.9**	**0.89**	0.85	**0.89**	0.86	0.86	0.85
Haberman	0.47	0.59	0.59	0.5	0.57	0.5	0.58	0.53	0.58
New Thyroid 1	0.94	0.95	**0.97**	**0.97**	**0.98**	**0.97**	**0.97**	**0.97**	**0.98**
New Thyroid 2	0.94	0.94	**0.97**	**0.95**	**0.97**	**0.95**	0.94	0.94	0.94
Pima	0.72	0.71	0.69	0.7	0.72	0.7	0.68	0.7	0.68
Wisconsin	0.95	0.96	0.95	0.93	0.95	0.93	0.96	0.94	0.95
Ortalama	0.81	0.836	**0.838**	0.829	0.833	0.83	0.836	0.831	0.83

Çizelge 4.7'deki çok katmanlı yapay sinir ağı yöntemi geometrik ortalama sonuçları incelendiğinde; ÖYRÖ glass6, new thyroid 1 ve new thyroid 2 veri kümelerinde, ÖYOK glass1, glass6, new thyroid 1 ve new thyroid 2 veri kümelerinde, ÖYOKED new thyroid 1 ve new thyroid 2 veri kümelerinde, ÖYT glass1, glass6, new thyroid 1 ve new thyroid 2 veri kümelerinde, ÖYSA glass0123vs456, glass0, ve new thyroid 1 veri kümelerinde, ÖYSÇ glass0, glass1 ve new thyroid 1 veri kümelerinde ve ÖYED glass1 ve new thyroid 1 veri kümesinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.8. Çok katmanlı yapay sinir ağı yöntemiyle f ölçüsü sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.97	0.97	0.97	0.96	0.96	0.97	0.97	0.96	0.97
Glass0123vs456	0.89	0.89	0.88	**0.9**	**0.9**	**0.9**	**0.91**	0.89	0.88
Glass0	0.76	0.76	0.73	0.76	0.75	0.73	**0.79**	**0.81**	0.73
Glass1	0.63	0.69	**0.7**	**0.73**	0.67	**0.7**	0.68	**0.7**	**0.7**
Glass6	0.85	0.86	**0.88**	**0.9**	0.83	**0.9**	**0.88**	**0.87**	0.83
Haberman	0.51	0.56	**0.6**	0.54	**0.58**	0.54	**0.62**	0.55	**0.6**
New Thyroid 1	0.95	0.95	0.95	**0.97**	**0.96**	**0.97**	**0.97**	**0.97**	**0.97**
New Thyroid 2	0.95	0.94	0.95	0.95	0.94	0.95	0.94	0.94	0.94
Pima	**0.73**	0.7	0.7	0.7	0.7	0.71	0.7	0.68	0.68
Wisconsin	0.95	0.95	0.94	0.93	0.94	0.93	0.95	0.94	0.94
Ortalama	0.819	0.827	0.83	0.834	0.823	0.83	**0.841**	0.831	0.824

Çizelge 4.8'deki çok katmanlı yapay sinir ağı yöntemi f ölçüsü sonuçları incelendiğinde; ÖYRÖ glass1, glass6 ve haberman veri kümelerinde, ÖYOK glass0123vs456, glass1, glass6 ve new thyroid 1 veri kümelerinde, ÖYOKED glass0123vs456, haberman ve new thyroid 1 veri kümelerinde, ÖYT glass0123vs456, glass1, glass6 ve new thyroid 1 veri kümelerinde, ÖYSA glass0123vs456, glass0, glass6, haberman ve new thyroid 1 veri kümelerinde, ÖYSÇ glass0, glass1, glass6 ve new thyroid 1 veri kümelerinde ve ÖYED glass1, haberman ve new thyroid 1 veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.9. Sıralı asgari optimizasyon yöntemiyle geometrik ortalama sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.95	0.96	0.96	0.96	0.96	0.96	**0.97**	0.96	**0.97**
Glass0123vs456	0.85	0.88	0.87	0.83	**0.9**	0.84	0.87	**0.89**	0.88
Glass0	0.32	0.64	**0.7**	0.26	0.6	0.37	**0.66**	**0.72**	**0.71**
Glass1	0	0.45	0.32	0	0.36	0	**0.46**	0.35	0.38
Glass6	0.86	0.89	**0.92**	**0.9**	**0.9**	0.85	**0.91**	0.89	**0.91**
Haberman	0	0.54	0.5	0.24	0.53	0.24	**0.61**	0.5	**0.61**
New Thyroid 1	0.73	**0.98**	0.91	0.69	0.93	0.69	0.96	0.82	0.96
New Thyroid 2	0.71	**0.97**	0.94	0.69	0.94	0.69	0.94	0.82	0.96
Pima	0.68	0.74	0.68	0.37	0.73	0.37	0.74	0.7	0.72
Wisconsin	0.96	0.96	**0.97**	0.86	0.96	0.84	0.96	0.96	**0.97**
Ortalama	0.606	0.801	0.777	0.58	0.781	0.59	**0.808**	0.761	0.807

Çizelge 4.9'da ki sıralı asgari optimizasyon yöntemi geometrik ortalama ölçüsü sonuçları incelendiğinde; ÖYRÖ glass0, glass6 ve wisconsin veri kümelerinde, ÖYOK glass6 veri kümesinde, ÖYOKED glass0123vs456 ve glass6 veri kümelerinde, ÖYSA ecoli0vs1_5, glass0, glass1, glass6 ve haberman veri kümelerinde, ÖYSÇ glass0123vs456 ve glass0 veri kümelerinde ve ÖYED ecoli0vs1_5, glass0, glass6, haberman ve wisconsin veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.10. Sıralı asgari optimizasyon yöntemiyle f ölçüsü sonuçları

Veri Kümesi	Orijinal	SMOTE	ÖYRÖ	ÖYOK	ÖYOKED	ÖYT	ÖYSA	ÖYSÇ	ÖYED
Ecoli0vs1_5	0.95	0.96	0.96	0.96	0.95	0.96	**0.97**	**0.97**	**0.97**
Glass0123vs456	0.88	0.86	0.86	0.85	**0.89**	0.84	0.86	0.86	0.87
Glass0	0.38	0.6	**0.67**	0.29	0.59	0.41	**0.61**	**0.61**	**0.67**
Glass1	0	0.43	0.38	0	0.42	0	**0.44**	**0.44**	0.41
Glass6	0.88	0.89	**0.9**	**0.92**	0.83	0.89	**0.9**	**0.9**	0.87
Haberman	0	0.57	0.55	0.3	**0.59**	0.3	**0.6**	**0.6**	**0.6**
New Thyroid 1	0.8	0.95	0.93	0.77	0.95	0.77	**0.96**	**0.96**	**0.96**
New Thyroid 2	0.78	0.95	0.95	0.76	0.95	0.76	0.95	0.95	**0.96**
Pima	0.71	0.73	0.7	0.41	0.73	0.41	0.73	0.73	0.71
Wisconsin	0.95	0.95	**0.96**	0.84	**0.96**	0.83	0.95	0.95	**0.96**
Ortalama	0.633	0.789	0.786	0.61	0.786	0.62	**0.797**	**0.797**	**0.798**

Çizelge 4.10'da ki sıralı asgari optimizasyon yöntemi geometrik ortalama ölçüsü sonuçları incelendiğinde; ÖYRÖ glass0, glass6 ve wisconsin veri kümelerinde, ÖYOK glass6 veri kümesinde, ÖYOKED glass0123vs456, haberman ve wisconsin veri kümelerinde, ÖYSA ecoli0vs1_5, glass0, glass1, glass6, haberman ve new thyroid 1 veri kümelerinde, ÖYSÇ ecoli0vs1_5, glass0, glass1, glass6, haberman ve new thyroid 1 veri kümelerinde ve ÖYED ecoli0vs1_5, glass0, haberman, new thyroid 1, new thyroid 2 ve wisconsin veri kümelerinde başarılı sonuçlar elde edilmiştir.

Çizelge 4.11. ÖYRÖ geometrik ortalama ve f ölçüsü Friedman testi sonuçları

Sınıflandırma Algoritması	Geometrik Ortalama			F Ölçüsü		
	Orijinal	SMOTE	ÖYRÖ	Orijinal	SMOTE	ÖYRÖ
AdaBoost.M1	1.7	**2.3**	2	1.9	1.9	**2.2**
K en yakın komşu (k=3)	1.4	**2.65**	1.95	1.65	**2.3**	2.05
K star	1.15	2.3	**2.55**	1.5	2.15	**2.35**
Çok katmanlı yapay sinir ağı	1.55	**2.35**	2.1	**2**	**2**	**2**
Sıralı asgari optimizasyon	1.1	**2.6**	2.3	1.35	**2.4**	2.25

Çizelge 4.11'de ÖYRÖ'ye ait geometrik ortalama ve f ölçüsü sonuçları verilmiştir. Geometrik ortalama sonuçlarında k star algoritmasında, F ölçüsü sonuçlarında AdaBoost.M1, k star ve çok katmanlı yapay sinir ağı algoritmalarında daha başarılı sonuçlar elde edilmiştir.

Çizelge 4.12. ÖYOK geometrik ortalama ve f ölçüsü Friedman testi sonuçları

Sınıflandırma Algoritması	Geometrik Ortalama			F Ölçüsü		
	Orijinal	SMOTE	ÖYOK	Orijinal	SMOTE	ÖYOK
AdaBoost.M1	1.95	**2.45**	1.6	2.15	**2.2**	1.65
K en yakın komşu (k=3)	1.6	**2.6**	1.8	1.8	**2.35**	1.85
K star	1.55	**2.75**	1.7	1.9	**2.45**	1.65
Çok katmanlı yapay sinir ağı	1.5	**2.3**	2.2	1.85	1.95	**2.2**
Sıralı asgari optimizasyon	1.7	**2.8**	1.5	1.8	**2.7**	1.5

Çizelge 4.12'de ÖYOK'ye ait geometrik ortalama ve f ölçüsü sonuçları verilmiştir. Geometrik ortalama sonuçlarında herhangi bir algoritmada başarılı sonuçlar elde edilemezken F ölçüsü sonuçlarında çok katmanlı yapay sinir ağı algoritmasında daha başarılı sonuç elde edilmiştir.

Çizelge 4.13. ÖYOKED geometrik ortalama ve f ölçüsü Friedman testi sonuçları

Sınıflandırma Algoritması	Geometrik Ortalama			F Ölçüsü		
	Orijinal	SMOTE	ÖYOKED	Orijinal	SMOTE	ÖYOKED
AdaBoost.M1	1.6	**2.25**	2.15	1.95	1.85	**2.2**
K en yakın komşu (k=3)	1.35	**2.55**	2.1	1.75	**2.25**	2
K star	1.2	2.35	**2.45**	1.65	**2.2**	2.15
Çok katmanlı yapay sinir ağı	1.5	**2.5**	2	2.05	**2.15**	1.8
Sıralı asgari optimizasyon	1.1	**2.65**	2.25	1.3	**2.4**	2.3

Çizelge 4.13'te ÖYOKED'ye ait geometrik ortalama ve f ölçüsü sonuçları verilmiştir. Geometrik ortalama sonuçlarında k star algoritmasında, F ölçüsü sonuçlarında AdaBosst.M1 algoritmasında daha başarılı sonuçlar elde edilmiştir.

Çizelge 4.14. ÖYT geometrik ortalama ve f ölçüsü Friedman testi sonuçları

Sınıflandırma Algoritması	Geometrik Ortalama			F Ölçüsü		
	Orijinal	SMOTE	ÖYT	Orijinal	SMOTE	ÖYT
AdaBoost.M1	2.05	**2.5**	1.45	**2.25**	2.2	1.55
K en yakın komşu (k=3)	1.5	**2.7**	1.8	1.75	**2.4**	1.85
K star	1.65	**2.7**	1.65	2	**2.35**	1.65
Çok katmanlı yapay sinir ağı	1.55	**2.3**	2.15	1.85	1.9	**2.25**
Sıralı asgari optimizasyon	1.7	**2.9**	1.4	1.7	**2.75**	1.55

Çizelge 4.14'te ÖYT'ye ait geometrik ortalama ve f ölçüsü sonuçları verilmiştir. Geometrik ortalama sonuçlarında herhangi bir algoritmada başarılı sonuçlar elde edilemezken F ölçüsü sonuçlarında çok katmanlı yapay sinir ağı algoritmasında daha başarılı sonuçlar elde edilmiştir.

Çizelge 4.15. ÖYSA geometrik ortalama ve f ölçüsü Friedman testi sonuçları

Sınıflandırma Algoritması	**Geometrik Ortalama**			**F Ölçüsü**		
	Orijinal	**SMOTE**	**ÖYSA**	**Orijinal**	**SMOTE**	**ÖYSA**
AdaBoost.M1	1.9	**2.3**	1.8	2.05	1.85	**2.1**
K en yakın komşu (k=3)	1.4	**2.6**	2	1.55	2.2	**2.25**
K star	1.3	**2.55**	2.15	1.85	**2.5**	1.65
Çok katmanlı yapay sinir ağı	1.4	**2.3**	**2.3**	1.75	1.85	**2.4**
Sıralı asgari optimizasyon	1.1	2.35	**2.55**	1.3	2.05	**2.65**

Çizelge 4.15'te ÖYSA'ya ait geometrik ortalama ve f ölçüsü sonuçları verilmiştir. Geometrik ortalama sonuçlarında çok katmanlı yapay sinir ağı yönteminde SMOTE yöntemi ile aynı sonuç elde edilmiştir. Ayrıca geometrik ortalama sonuçlarında sıralı asgari optimizasyon algoritmasında, F ölçüsü sonuçlarında AdaBoost.M1, k en yakın komşu (k=3), çok katmanlı yapay sinir ağı ve sıralı asgari optimizasyon algoritmalarında daha başarılı sonuçlar elde edilmiştir.

Çizelge 4.16. ÖYSÇ geometrik ortalama ve f ölçüsü Friedman testi sonuçları

Sınıflandırma Algoritması	**Geometrik Ortalama**			**F Ölçüsü**		
	Orijinal	**SMOTE**	**ÖYSÇ**	**Orijinal**	**SMOTE**	**ÖYSÇ**
AdaBoost.M1	1.65	**2.35**	2	2	1.95	**2.05**
K en yakın komşu (k=3)	1.5	**2.65**	1.85	1.8	**2.4**	1.8
K star	1.3	**2.65**	2.05	1.85	**2.45**	1.7
Çok katmanlı yapay sinir ağı	1.65	**2.4**	1.95	1.9	**2.05**	**2.05**
Sıralı asgari optimizasyon	1.1	**2.6**	2.3	1.3	2.05	**2.65**

Çizelge 4.16'da ÖYSÇ'ye ait geometrik ortalama ve f ölçüsü sonuçları verilmiştir. Geometrik ortalama sonuçlarında herhangi bir algoritmada başarılı sonuçlar elde edilemezken F ölçüsü sonuçlarında AdaBoost.M1, çok katmanlı yapay sinir ağı ve sıralı asgari optimizasyon algoritmalarında daha başarılı sonuçlar elde edilmiştir.

Çizelge 4.17. ÖYED geometrik ortalama ve f ölçüsü Friedman testi sonuçları

Sınıflandırma Algoritması	Geometrik Ortalama			F Ölçüsü		
	Orijinal	SMOTE	ÖYED	Orijinal	SMOTE	ÖYED
AdaBoost.M1	1.6	2.05	**2.35**	1.95	1.8	**2.25**
K en yakın komşu (k=3)	1.2	2.35	**2.45**	1.6	2.15	**2.25**
K star	1.2	2.35	**2.45**	1.55	2.1	**2.35**
Çok katmanlı yapay sinir ağı	1.6	**2.5**	1.9	2.1	**2.15**	1.75
Sıralı asgari optimizasyon	1.05	2.4	**2.55**	1.4	2.15	**2.45**

Çizelge 4.17'de ÖYED'ye ait geometrik ortalama ve f ölçüsü sonuçları verilmiştir. Geometrik ortalama sonuçlarında AdaBoost.M1, k en yakın komşu (k=3), k start ve sıralı optimizasyon algoritmalarında, F ölçüsü sonuçlarında ise geometrik ortalama sonuçları ile aynı algoritmalarda daha başarılı sonuçlar elde edilmiştir.

Yapılan bu tez çalışmasında belirli bir eşik değeri vererek önerilen yöntemin 4 sınıflandırma algoritmasında başarılı ve 1 sınıflandırma algoritmasında ise başarısız sonuç aldığı gözlemlenmiştir. Çizelge 4.18'de ise başarılı yöntemin sonuçları toplu olarak gösterilmiştir.

5. SONUÇLAR VE ÖNERİLER

5.1. Sonuçlar

Yapılan tez çalışmasında genel olarak dengesiz veri kümelerinde sınıflandırma başarısının artırılmasına yönelik çalışmalar yapılmıştır. Bu kapsamda yapay sinir ağları ile rastgele örnek üretimi, yapay sinir ağının gizli katmanlarının sayısı arttırılarak otomatik kodlayıcı yapay sinir ağı ile rastgele örnek üretimi, bu otomatik kodlayıcı yapay sinir ağına bir eşik değeri vererek üretilen örneklerin kısıtlanması, azınlık sınıfındaki örneklerin tekrarlı olarak veri kümesine eklenmesi, SMOTE yöntemi örneklerinin azınlık ve çoğunluk yapay sinir ağlarından geçirilip rastgele örnek elde edilmesi ve ilk yapılan rastgele örnekleme yöntemine bir eşik değeri koyarak üretilen örneklerin kısıtlanması olmak üzere yedi adet yöntem önerilmiştir. Bu yöntemlerin test edilmesi için de on veri kümesi ve her veri kümesinin beş çapraz doğrulaması kullanılmıştır. Test edildikten sonra sonuçların yorumlanabilmesi için geometrik ortalama ve f ölçüsü kıstasları kullanılmıştır. Ayrıca geometrik ortalama ve f ölçüsü sonuçlarının alınması için "AdaBoost.M1", "K en yakın komşu", "K star", "Çok katmanlı yapay sinir ağı" ve "Sıralı optimizasyon algoritması" olmak üzere beş sınıflandırma algoritması kullanılmıştır. Burada önerilen yöntemler orijinal sınıflandırma sonuçları ve SMOTE yöntemi ile karşılaştırılmıştır. Hangi yöntemin daha başarılı olduğunun belirlenmesi için ise Friedman testi yöntemi kullanılmıştır.

Araştırma sonuçları ve tartışma bölümündeki Friedman testi sıralama sonuçları incelendiğinde son yöntem olan yapay sinir ağlarını kullanarak eşik değeriyle rastgele örnek üretimi yönteminin geometrik ortalama ve f ölçüsü sonuçlarının Friedman istatistiklerinin "Çok katmanlı yapay sinir ağı" sınıflandırma algoritmasında başarısız, diğer dört sınıflandırma algoritmasında ise başarılı sonuçlar elde ettiği gözlemlenmiştir. Bu tez kapsamında dengesiz veri kümelerinin sınıflandırma başarısı önerilen yöntemler kullanılarak artırılmıştır.

5.2. Öneriler

Bu tez çalışmasında yapılan çalışmalar sonucunda ilerideki çalışmalar için öneriler bu bölümde verilmiştir.

Bu tez çalışmasında on veri kümesi kullanılmış olup dengesizlik oranı 1.5 ve 9 arasında olan diğer veri kümeleri ile dengesizlik oranı 9 üzerinde olan veri kümeleri ile de çalışmalar yapılabilir

Bu tez çalışmasında kullanılan beş sınıflandırma algoritması haricinde diğer sınıflandırma algoritmaları ile de test yapılabilir.

Sınıflandırma sonuçları için geometrik ortalama ve f ölçüsü metrikleri kullanılmış olup bu metriklerin değerlendirilmesi için de Friedman ortalama sıralama yöntemi kullanılmıştır. Yine bu yöntemlerin haricinde ROC eğrisi gibi metrikler ile diğer istatistiksel yöntemler ile değerlendirmeler yapılabilir.

KAYNAKLAR

Anand, R., Mehrotra, K.G., Mohan, C.K. and Ranka, S., 1993, An improved algorithm for neural network classification on imbalanced training sets, *IEEE Transactions on Neural Networks*, 4 (6), 962-969

Ankara, N., 2019, Dengesiz kredi skorlama veri setlerinde kolektif öğrenme algoritmalarının performans değerlendirmesi, yüksek lisans, fen bilimleri enstitüsü, Yıldız Teknik Üniversitesi, 1-53

Aydın, Y., 2016, Dengesiz veri setlerinde trafik işaretlerini tanıma, yüksek lisans, *fen bilimleri enstitüsü*, Atatürk Üniversitesi, 1-142

Ayhan, D., 2009, Multi-class classification methods utilizing Mahalanobis Taguchi system and a re-sampling approach for imbalanced data sets, yüksek lisans, *fen bilimleri enstitüsü,* Orta Doğu Teknik Üniversitesi, 1-99

Bulut, F., 2016, *Bilişim Teknolojileri Dergisi*, 9 (2), 153-159

*Bunkhumpornpat, C., Sinapiromsaran, K., and 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*, 1, 475-482

*Caruna, R., 2000, Learning From Imbalanced Data: Rank Metrics and Extra Tasks, Prof.Am.Assoc. for Artifical Intelligence (AAII) conf, WS-00-05, 51-57

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P., 2002, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artifical Intelligence Research, 16, 321-357

Cleary, J.G. and Trigg, L.E., 1995, K: An Instance-based Learner Using an Entropic Distance Measure, Machine Learning Proceedings, , 108-114

Çayıroğlu, İ., İleri Algoritma Analizi-5 Yapay Sinir Ağları [online], http://www.ibrahimcayiroglu.com/Dokumanlar/IleriAlgoritmaAnalizi/IleriAlgoritmaAnalizi-5.Hafta-YapaySinirAglari.pdf, [Ziyaret Tarihi:07.05.2020]

Douzas, G., and Bacao, F., 2017, Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning, *Expert Systems with Applications*, 82, 40-52

Douzas, G., and Bacao, F., 2019, Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE, *Information Sciences*, 501, 118-135

Estabrooks, A., Jo, T., and Japkowicz, N., 2004, A Multiple Resampling Method for Learning from Imbalanced Data Sets, *Computational Intelligence*, 20 (1), 18-36

Fotouhi, S., Asadi, S., and Kattan, M.W., 2019, A comprehensive data level analysis for cancer diagnosis on imbalanced data, *Journal of Biomedical Informatics*, 90, 103089

Gümüştaş, E., 2019, Kayıp gözlem içeren dengesiz veri setlerinin topluluk öğrenme algoritmaları le sınıflandırılması, yüksek lisans, *fen bilimleri enstitüsü*, Mimar Sinan Güzel Sanatlar Üniversitesi, 1-60

Haklı, D.A., 2018, Sınıf dengesizliği sorununu çözmek için kullanılan algoritmaların farklı sınıflandırma yöntemlerinde performanslarının karşılaştırılması, doktora, *sağlık bilimleri enstitüsü*, Hacettepe Üniversitesi, 1-102

*Han, H., Wang, W.Y. and Mao, B.H., 2005, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning International conference on intelligent computing, 1, 878-887

He, H. and Garcia, E.A., 2009, Learning from Imbalanced Data, *IEEE Transactions On Knowledge And Data Engineering*, 21 (9), 1263-1284

*Japkowicz, N., 2000, Learning From Imbalanced Data: A Comparison of Various Strategies, *AAAI Workshop on learning from imbalanced data sets*, WS-00-05, 10-16

*Jeatrakul, P., Wong, K.W. and Fung, C.C., 2010, Classification of Imbalanced Data by Combining the Compleentary Neural Network and SMOTE algorithm, *Internation conference on Neural Information Processing*, 2, 152-159

Krawezyk, B., 2016, Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence*, 5(4), 221-232

Ma, L., and Fan, S., 2017, CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests, BMC Bioinformatics, 18 (1), 169

*Maciejewski, T., Stefanowski, J., 2011, Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data, IEEE Symposium on Computational Intelligence and Data Mining (CIDM), , 104-111

Nakamura, M., Kajiwara, Y., Otsuka, A. and Kimura, H., 2013, LVQ-SMOTE – Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data, *BioData Mining*, 6 (1), 16

Öztürk, A., 2009, SVM classification for imbalanced datasets with multi objective optimization framework, yüksek lisans, *fen bilimleri enstitüsü*, Koç University, 1-80

*Platt, J.C., 1998, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Microsoft Researh, 21

Ramentol, E., Caballero, Y., Bello, R., and Herrera, F., 2011, 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

Sáez, J.A., Luengo, J., Stefanoski, J. and Herrera, F., 2015, SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, *Information Sciences,* 291, 184-203

Sağlam, F., 2020, Gürültülü gözlemler durumunda dengesiz veride öğrenme için yeni bir yaklaşım, yüksek lisans, *fen bilimleri enstitüsü*, Ondokuz Mayıs Üniversitesi, 1-91

Sarmanova, A., 2013, Veri madenciliğinde sınıf dengesizliği sorununun giderilmesi, yüksek lisans, fen bilimleri enstitüsü, Yıldız Teknik Üniversitesi, 1-81

Sun, Y., Wong, A.K.C., and Kamel, M.S., 2009, *Classification of Imbalanced Data: A Review*, *International Journal of Pattern Recognition and Artificial Intellience*, 23 (04), 687-719

Turhan, S., 2019, Kolektif öğrenmede sınıf dengesizliği problemi, yüksek lisans, fen bilimleri enstitüsü, Muğla Sıtkı Koçman Üniversitesi, 1-100

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q. and Kennedy, P.J.,2016, Training deep neural networks on imbalanced data sets, *International Joint Conference on Neural Networks (IJCNN),* , 4368-4374