DENGESİZ VERİ SETLERİNDE SINIFLANDIRMA PROBLEMLERİNİN ÇÖZÜMÜNDE MELEZ YÖNTEM UYGULAMASI

MESTAN ŞAHİN PİR

T.C.

BURSA ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DENGESİZ VERİ SETLERİNDE SINIFLANDIRMA PROBLEMLERİNİN ÇÖZÜMÜNDE MELEZ YÖNTEM UYGULAMASI

Mestan Şahin PİR

0000-0001-8248-0738

Doç. Dr. Duygu YILMAZ EROĞLU
(Danışman)

YÜKSEK LİSANS TEZİ

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA– 2022

Her Hakkı Saklıdır.

TEZ ONAYI

Mestan Şahin PİR tarafından hazırlanan "Dengesiz Veri Setlerimde Sınıflandırma Problemlerinin Çözümünde Melez Yöntem Uygulaması" adlı tez çalışması aşağıdaki jüri tarafından oy birliği/oy çokluğu ile Uludağ Üniversitesi Fen Bilimleri EnstitüsüEndüstri Mühendisliği Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman : Doç. Dr. Duygu YILMAZ EROĞLU

Başkan :	Doç.Dr. Duygu YILMAZ EROĞLU 0000-0002-7730-2707 Bursa Uludağ Üniversitesi Mühendislik Fakültesi, Endüstri Mühendisliği Anabilim Dalı	İmza
Üye :	Prof. Dr. Turgay Tugay BİLGİN 0000-0002-9245-5728 Bursa Teknik Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Anabilim Dalı	İmza
Üye :	Doç. Dr. Tülin İNKAYA 0000-0002-6260-0162 Bursa Uludağ Üniversitesi Mühendislik Fakültesi, Endüstri Mühendisliği Anabilim Dalı	İmza

Yukarıdaki sonucu onaylarım Prof. Dr. Hüseyin Aksel EREN

Enstitü Müdürü

../../….(Tarih)

U.Ü. Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,

ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

10/01/2022

İmza

Mestan Şahin PİR

TEZ YAYINLANMA

FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezin/raporun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma izni Bursa Uludağ Üniversitesi'ne aittir. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet hakları ile tezin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları tarafımıza ait olacaktır. Tezde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığını ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederiz.

Yükseköğretim Kurulu tarafından yayınlanan "**Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge**" kapsamında, yönerge tarafından belirtilen kısıtlamalar olmadığı takdirde tezin YÖK Ulusal Tez Merkezi / B.U.Ü. Kütüphanesi Açık Erişim Sistemi ve üye olunan diğer veri tabanlarının (Proquest veri tabanı gibi) erişimine açılması uygundur.

Doç. Dr. Duygu YILMAZ EROĞLU	Mestan Şahin PİR
Tarih	Tarih
İmza	İmza
Bu bölüme kişinin kendi el yazısı ile okudum anladım yazmalı ve imzalanmalıdır.	Bu bölüme kişinin kendi el yazısı ile okudum anladım yazmalı ve imzalanmalıdır.

ÖZET

Yüksek Lisans Tezi

DENGESİZ VERİ SETLERİNDE SINIFLANDIRMA PROBLEMLERİNİN ÇÖZÜMÜNDE MELEZ YÖNTEM UYGULAMASI

Mestan Şahin PİR

Bursa Uludağ Üniversitesi Fen
Bilimleri Enstitüsü
Endüstri Mühendisliği
Anabilim Dalı

Danışman: Doç. Dr. Duygu YILMAZ EROĞLU

Günümüzde veri toplama teknolojilerinde yaşanan gelişmeler ile veriye bağlı karar destek yöntemlerine olan ilgi ve dolayısıyla veri madenciliğine ilgi arttı. Bu ilgi beraberinde farklı veri türlerinde veri madenciliği çalışmalarının yapılmasını sağladı. Günümüzde numerik ve kategorik verilerin yanında, görüntü tanıma, ses tanıma ve metin madenciliği gibi alanlarda yapılan çalışmalar ile çok sayıda bilimsel ve gerçek hayat çalışması gerçekleşti. Biyomedikal bilişim, örüntü tanıma, dolandırıcılık algılama, doğal dil işleme, tıbbi teşhis, yüz tanıma, metin sınıflandırma, arıza teşhis, anomali tespiti gibi başlıca gerçek hayat uygulama alanlarının yanında, otomom araçlar, Endüstri 4.0, insansız hava araçları gibi yeni teknolijilerde de çalışmaların sayısı arttı. Bu çalışmaların bir kısmında veri setlerinin dengesiz olması, diğer bir deyişle bir sınıf etiketinin belirgin oranda diğer sınıf/sınıflara baskın olması durumu ile karşılaşılmıştır. Bu durumda sınıflandırıcılar çoğunluk sınıfını doğru tahmin ettiği halde, azınlık verilerinde doğru tahminde bulunamamaktadır. Bu da kalite kontrol, tıbbi teşhis gibi çalışma alanlarında ciddi sorunlara yol açmaktadır.

Çalışma kapsamında önerilen melez yöntem ile dengesiz verilerde sınıflandırma problemine çözüm önerilmiştir. Amaç dengesiz verilerde, aşırı örneklemenin yol açtığı aşırı uyum sorunun ve örnek azaltmanın yol açtığı değerli veri kaybının önüne geçilmesi ve başarılı sınıflandırma sonucu almaktır.

İlk olarak dengesiz verinin sınıflandırılması ile ilgili yapılan çalışmalar incelendi. Sonrasında yapılan bu çalışmaların avantaj ve dezavantajlarından yararlanan yeni bir yöntem önerildi. Melez Yöntemin uygulandığı 8 veri seti farklı tip sınıflandırıcılar ile sınıflandırmış, sonuçlar dengesiz veri sınıflandırma problemlerinde sıkça kullanılan SMOTE yöntemi ile dengelenmiş veri setinin sonuçları ile karşılaştırılmıştır. Alınan sonuçlar önerilen yöntemin başarısını doğrulamıştır. Gerçek hayat verilerinde giriş kalite ve proses parametrelerinin iplik kopuşlarının tahmininde kullanılması ile yüksek doğru tahmin oranı ile ipliklerin dokumaya girmesini engelleyebilecek bir karar destek sistemi sunmuştur.

Anahtar Kelimeler: Veri Madenciliği, Dengesiz Veri Setleri, Sınıflandırma, Karar Destek Sistemleri

2022, vi + 47 sayfa.

ABSTRACT

MSc Thesis

HYBRID METHOD APPLICATION TO SOLVE CLASSIFICATION PROBLEMS IN IMBALANCED DATASETS

Mestan Şahin PİR

Bursa Uludağ University
Graduate School of Natural and Applied Sciences Department of Industrial Engineering

Supervisor: Assoc. Prof. Dr. Duygu YILMAZ EROĞLU

Today, the improvements of collecting data technologies and decisions depending on the data-based consequently increased the interest of data mining recently. This interest lead to studies in different data types. These days, besides of numeric and categorical data, visual recognition, voice recognition, text mining etc. has developed many real life and science study. In addition to the main real-life application areas such as biomedical informatics, pattern recognition, fraud detection, natural language processing, medical diagnosis, face recognition, text classification, fault diagnosis, anomaly detection, the number of studies in new technologies such as autonomous vehicles, Industry 4.0, unmanned aerial vehicles it increased. In some of these studies, it was encountered that the data sets were unbalanced, in other words, one class label was significantly dominant over the other class/classes. In this case, although the classifiers predict the majority class correctly but they cannot predict the minority class correctly. This makes serious problem on quality check, medical diagnossis etc.

In this study, hybrid method proposed a solution the classification problem in imbalanced datasets. The aim is to prevent the overfitting problem caused by oversampling and valuable data loss caused by undersampling in imbalanced data, and to obtain successful classification results.

Firstly, the studies on the classification of imbalanced data were examined. Then another method was proposed considering all the studies advantages and disadvantages. Hybrid method was applied to eight datasets, then these datasets were classified with different types of classifiers, and the results were compared with the results of the balanced data set with the SMOTE method, which is frequently used in imbalanced data classification problems. The obtained results confirmed the success of the proposed method. By using the input quality and process parameters in the real yarn data to predict yarn breaks, has presented a decision support system that can prevent yarns from entering the weaving with a high correct prediction rate.

Key words: Data Mining, Imbalanced Datasets, Classification, Decision Support Systems

2022, vi + 47 pages.

İÇİNDEKİLER

Sayfa

KISALTMALAR DİZİNİ

Kısaltmalar	**Açıklama**
ADASYN	Adaptive Synthetic Sampling Method for Imbalanced Data (Dengesiz Veriler için Uyarlanabilir Sentetik Örnekleme Yöntemi)
AUC	Area under the ROC Curve (ROC Eğrisinin Altındaki Alan)
DBSCAN	Density-based spatial clustering of applications with noise (Gürültülü Uygulamaların Yoğunluğa Dayalı Uzamsal Kümelenmesi)
FNR	Pozitif Sınıfa Ait Yanlış Sınıflandırılmış Örnek Oranı
FPR	Negatif Sınıfa Ait Yanlış Sınıflandırılmış Örnek Oranı
FPR	Yanlış Pozitif Oranı
KNN	K-Nearest Neighbors (K-En Yakın Komşu)
MCC	Matthews Correlation Coefficient (Matthews Korelasyon Katsayısı)
RF	Random Forests Algoritması
ROSE	Random Over Sampling (Rastgele Aşırı Örnekleme)
SIMO	A synthetic Informative Minority Over-Sampling (Sentetik Bilgilendirici Azınlık Aşırı Örnekleme)
SLS	Safe Level Smote (Güvenli Bölge Sentetik Azınlık Aşırı Örnekleme Tekniği)
SMOTE	Synthetic Minority Oversampling Technique (Sentetik Azınlık Aşırı Örnekleme Tekniği)
SVM	Support Vector Machines (Destek Vektör Makineleri)
TNR	Negatif Sınıfa Ait Doğru Sınıflandırılmış Örnekler Sayısı
TNR	Gerçek Negatif Oranı
TPR	Pozitif Sınıfa Ait Doğru Sınıflandırılmış Örnekler Sayısı
TPR	Gerçek Negatif Oranı
YSA	Yapay Sinir Ağları

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

1.GİRİŞ

Son yıllarda gelişen veri toplama teknolojileri ve düşen maliyetler ile veri madenciliğine olan ilgi ve kullanım alanı arttı. Biyomedikal bilişim, örüntü tanıma, dolandırıcılık algılama, doğal dil işleme, tıbbi teşhis, yüz tanıma, metin sınıflandırma, arıza teşhis, anomali tespiti gerçek hayat uygulama alanlarına örnek gösterilebilir.

Veri setleri arasında, nadir karşılaşılan etiketlerin tespit edilmesinin hedeflendiği problemlerin artması dengesiz veri setlerine olan eğilimi arttırmıştır. Veri setlerinin dengesiz olarak tanımlanması için bir sınır değer bulunmamaktadır. Dengesiz veri seti, bazı sınıflara ait gözlemlerin diğer sınıflara ait gözlemlere göre fazla olduğu veri setidir. Klasik sınıflandırıcılar bu veri setlerinde kullanılabilir sonuçlar vermemektedirler. Dengesiz veri setlerine odaklanan algoritmaların temel amacı doğruluk oranını artırma ve hata oranını azalmaktır.

İki sınıftan oluşan, bir veri seti üzerinde analizler yapılırken, ilk sınıfa ait 90 örnek ve ikinci sınıfa ait 10 örnek mevcut ise, sınıflandırıcı tüm örnekleri ilk sınıfa atadığında dahi doğruluk oranında %90 değerini yakalayabilir. Oysa ikinci sınıfa ait hiçbir veri sınıfı doğru tahmin edilmedi. Veri setinin bir ürünün üretimine ait ret/onay durumunu verdiğini düşünüldüğünde, müşteriye 10 adet ürün hatalı şekilde gönderilecektir. Bu da dengesiz verilerin klasik sınıflandırıcılar ile çözümlenmeden önce bazı farklı yaklaşımlar ile düzenlemesinin gerekliliğini göstermektedir. Aynı zamanda dengesiz veri setleri için doğruluk oranının yeterli değerlendirme ölçütü olamayacağının diğer değerlendirme kriterlerinin de kullanılması gerektiği de gözlemlenebilir. (G-Ortalama, AUC, F-Ölçütü gibi.)

Bu tez çalışmasında veri eğitim ve test verisi olarak ayrıldıktan sonra, eğitim verisine öncelikle SVM uygulanarak karar sınırına uzak çoğunluk sınıf veriler veri setinden çıkarılacaktır. Sonrasında indirgenmiş veri setinde SLS (Safe Level Smote) yöntemi kullanılarak azınlık verilerin güvenli sınırda çoğaltılması sağlanacaktır. Bu işlem veri seti dengeli hale gelinceye kadar devam edilecektir. Veri setinde sınıf dağılımı %50 ± 5 olduğunda veri setine dengeli denecektir. Yöntemin daha önce yayınlanan SIMO (A Synthetic Informative Minority Over-Sampling) ile benzer olarak bilgi verici azınlık sınıfı verileri çoğaltılırken, farklı olarak bilgi vermeyen çoğunluk sınıfı verileri veri

setinden çıkarmaktadır. Böylece SMOTE yönteminde rastlanan aşırı uyum, örnek azaltmada rastlanan bilgi kaybından kaçınılmaya çalışılmaktadır. Sonrasında sınıflandırma algoritmaları uygulanarak sınıflandırma başarısının arttırılması amaçlanacaktır. Çalışmada gerçek hayat verisi olarak, dokuma fabrikasında dokumaya girecek ipliklerin dokuma aşamasında kopup kopmayacağını önerilen melez yöntem ile tahmin edilmeye çalışılmıştır. Karmaşıklık matrisi üzerinden hem pozitif hem de negative sınıf performansını dikkate alması nedeniyle G-Ortalama ölçütü hesaplanarak karşılaştırmalar yapılacaktır.

Çalışmanın devamında 2. bölümde dengesiz veri setleri tanımlanmış ve bunları ele alma yöntemleri incelenmiş, 3. bölümde performans metrikleri ve hesaplama yöntemleri açıklanmış, 4. bölümde önerilen metodoloji açıklanmış, 5. bölümde uygulama detayları anlatılmış, 6. bölümde ise sonuç ve önerilerde bulunulmuştur.

2. KAYNAK ÖZETLERİ VE KURAMSAL TEMELLER

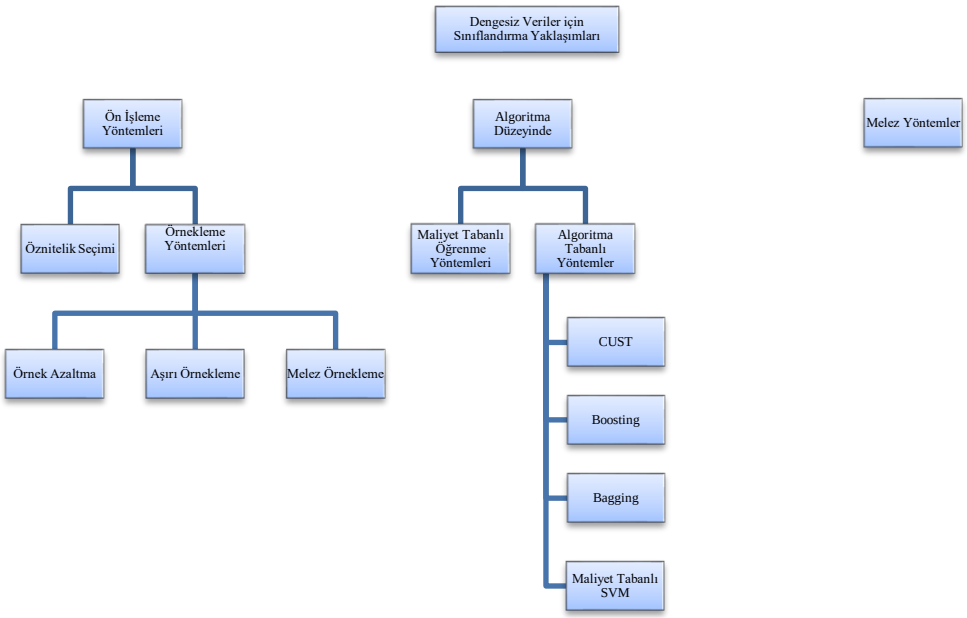
2.1 Dengesiz Veri Setleri

Belirli bir sınıf için bir veri kümesindeki gözlem sayısı diğer sınıftan daha yüksekse, sınıfın çoğunluk sınıfı olduğu söylenir. Diğer bir deyişle, belirli bir sınıf için bir veritabanındaki gözlem sayısı, aynı veritabanındaki diğer sınıftan daha azsa, sınıfın azınlık sınıfı olduğu söylenir. Bu tür veri setlerine dengesiz veri setleri adı verilir (Gong ve Kim, 2017).

Dolandırıcılık tespiti (kredi kartı, telefon görüşmeleri, sigorta), tıbbi teşhis, ağ saldırı tespiti, arıza izleme, kirlilik tespiti, biyomedikal, uzaktan algılama (kara mayını, su altı mayını) ve biyoinformatik gibi birçok gerçek dünya uygulamasında dengesiz veri problem ile karşılaşılabilmektedir.

Şekil 2.1'de de gösterildiği gibi dengesiz verilerde sınıflandırma probleminin çözümü üç ana başlık altında toplanmaktadır.

- Ön işleme yöntemleri (veri düzeyinde)
- Algoritma düzeyinde
- Melez yöntemler

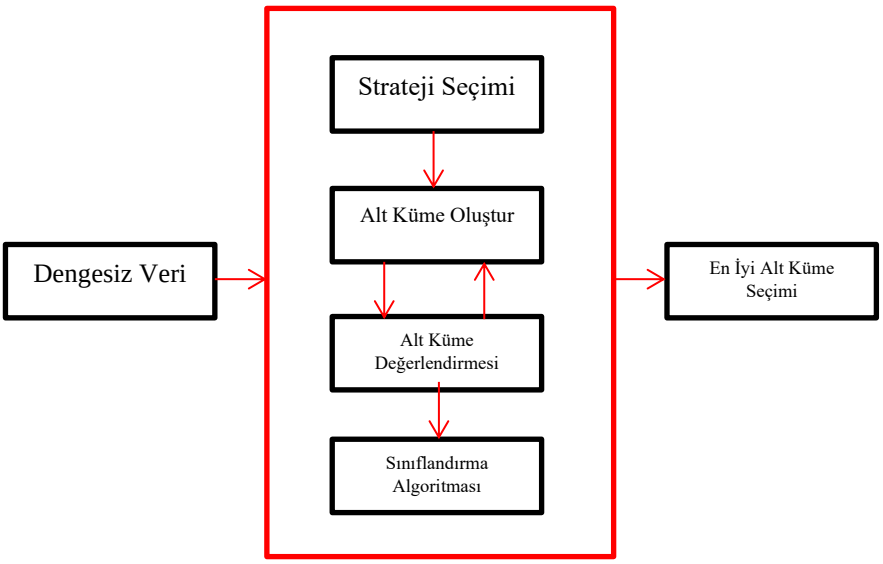


Şekil 2.1. Dengesiz veriler için sınıflandırma yaklaşımları (Kaur ve ark., 2019)

2.1. Ön İşleme Yöntemleri

2.1.1 Öznitelik seçimi

Öznitelik bir veri seti içerisinde bulunan ve hedeflenen model çıktısının oluşturulmasını sağlayacak olan her bir kolon/sütundur. Öznitelik seçimi ise, veri seti içerisinde en yararlı öznitelikleri seçme ve bulma sürecidir. Yüksek boyutlu veri kümelerinden ilgili özelliklerin veya özniteliklerin alt kümesinin seçilmesi, sınıflandırıcının performansını yükseltmeye yardımcı olur (Kaur ve ark., 2019). Şekil 2.2.'de öznitelik seçim süreci gösterilmiştir.



Şekil 2.2. Öznitelik Seçimi (Kaur ve ark., 2019)

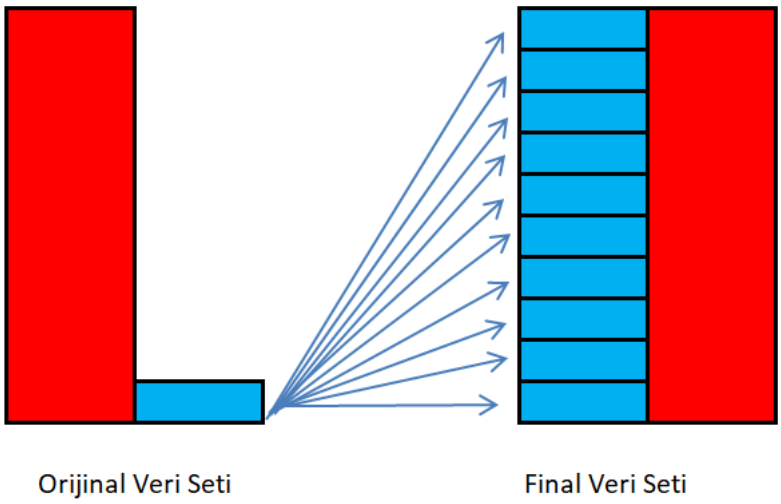
Öznitelik seçimi için Yin ve diğerleri (2013) tarafından yapılan çalışmada iki yeni yöntem önerilmiştir. İlk yöntemde, toplu sınıflar sözde alt sınıflara ayrılmış ve ayrıştırılmış verilerle özellikler değerlendirilmiştir. Diğer yöntem ise, sınıf bilgisi ile hesaplama maliyetini içermeyen hellinger mesafesine dayalı öznitelik seçimidir. Önerilen yaklaşım, gerçek dünya verileri kullanılarak diğer özellik seçim yöntemi ile karşılaştırılmıştır. F-ölçüsü ve AUC gibi değerlendirme ölçütlerinin sonuçlarına dayanarak önerilen yöntemin, performansının oldukça yüksek olduğu kanıtlanmıştır.

Korelasyon katsayısı, ki-kare, Odds oranı, sinyal-gürültü korelasyon katsayısı, bilgi kazanımı, kurtarma, hızlı ve kayan eşik ile özellik değerlendirmesi olmak üzere sekiz öznitelik seçim yöntemini analiz eden bir çalışma yapılmıştır (Bazmara ve Jamali, 2012). Sonucunda, veri kümesindeki özelliklerin sayısına göre veri kümesi için hangi yöntemin uygun olduğunu kanıtlanmış ve dengesiz veri seti için uygun öznitelik seçim modelini seçmek için daha az zaman harcanması sağlanmıştır.

2.1.2 Aşırı Örnekleme

Aşırı örnekleme Şekil 2.3.'te gösterildiği gibi, azınlık sınıfı verilerini çoğaltarak, sınıf dağılımının eşit hale getirilmesini amaçlar.

Rasgele aşırı örnkelme, azınlık sınıfından olan verierin rasgele seçilerek çoğaltılması ve orjinal veri setine eklenmesi ile yapılır. Bu yöntem basittir ancak tam kopyaların aşırı uyuma yol açabileceği öne sürülmüştür (Barista ve ark., 2004).



Şekil 2.3. Aşırı örnekleme gösterimi

En sık kullanılan aşırı örnekleme yöntemi ise SMOTE (Synthetic Minority Oversampling Technique) yaklaşımıdır (Chawla ve ark., 2002). Bu yöntemde rasgele örneklemeden farklı olarak, mevcut azınlık veriler analiz ederek sentetik veriler oluşturur. SMOTE, yeni yapay örneklerde orijinal örneklerin dağılımını gerçekten yansıtamaz. Bu nedenle SMOTE tabanlı yüksek hızda örnekleme yöntemlerini kullanırken, örneklerin dağılımında hata olabilir ve sınıflandırıcının doğruluğunu etkileyebilir. Bu durum, olasılığı artırarak örneklerin yanlış sınıflandırılmasına neden olacaktır (Zheng ve ark., 2015).

Bunkhumpornpat ve ark. (2009), SLS (Safe level SMOTE) adında bir yöntem önermişlerdir. En yakın komşu azınlık örnekleri kullanılarak güvenli seviye belirlenmiş, güvenli seviye bölgesinde aynı ağırlık değerine sahip azınlık verilerini hat boyunca dikkatli bir şekilde örneklemişlerdir. SMOTE ve Borderline SMOTE den daha iyi sonuç aldıklarını çalışmada kanıtlamışlardır.

He ve ark. (2008) ise, sentetik örnekler oluşturmak yerine, azınlık sınıfının içindeki iki sınıf arasındaki örnekleri oluşturan ADASYN adında yeni bir yöntem önermişlerdir. Azınlık sınıfı örnekleri için üretilen sentetik verilerin öğrenilmesi daha zordur. Dolayısıyla ADASYN yöntemi, azınlık örnek sınıfları için ağırlıklı dağılımı kullanır. ADASYN, dengesizlik sınıf öğrenimini iki adımda geliştirir: birincisi, sınıf dengesizliğinin getirdiği önyargıyı azaltmaktır. İkincisi, sınıflandırma kararı sınırını karmaşık örneklere kaydırmaktır. Bu ikisi ise dinamik ağırlık ayarlamaları ve uyarlanabilir öğrenme prosedürü ile gerçekleştirilir (Amin ve ark., 2016).

Rastgele karar ormanları, sınıflandırıcıların karar ağaçlarını üreten ve sınıflandırıcıların sonuçlarını doğrulayan bir topluluk yöntemi olan bir tekniktir (Ho, 1995). Ali ve ark. (2012) tarafından önerilen yöntem, torbalama fikrini ve rastgele özellik seçimini birleştirmektedir. Bölünmeyi belirlemek için sınıflandırma ve regresyon ağaçları, rastgele seçilen giriş değişkenleri arasında bir yeniden örnekleme yapar. Rastgele Ormanın avantajları, aşırı uyum probleminin üstesinden gelmesi, ağaçların budama ihtiyacını ortadan kaldırması, değişkenin önemi ve doğruluğunun üretmesi, aykırı veriler için eğitim verilerinden daha az hassas olmasıdır.

Fernandez-Navarro ve ark. (2011), dengesiz veri sınıflandırmasını iki yöntemle aydınlatmak için dinamik bir aşırı örnekleme yöntem çalışmasında radyal temel işlevleri sinir ağlarını azaltan memetik bir algoritmaya dahil etmişlerdir. Bu yöntemde, sınıf dengesizliği sorununu çözmek için iki aşamada eğitim verileri yeniden örneklenir. İlk aşama, bir aşırı örnekleme tekniği kullanarak azınlık sınıfını arttırmaktan oluşur. Memetik algoritma, farklı aşamalardaki verileri aşırı örneklemekte ve minimum duyarlılık sınıfının yeni modellerini sağlamaktadır.

Saez ve ark. (2016), çok sınıflı dengesizlik sorunu ve sınıf özelliklerinin analizi ile başa çıkmak için bir yüksek hızda örnekleme çalışması yapmışlardır. Bu çalışmada, her sınıftaki önemli örneklerin alt kümelerini bulur ve her biri için bağımsız olarak yüksek hızda örnekleme ile bunları ele alır. Bu metodoloji, çok sınıflı veri kümelerindeki dört farklı türde örneği tespit etmektedir: güvenli, sınırda, nadir ve aykırı değerler.

Diğer bir çalışmada da, dengesiz verileri öğrenmek için rastgele otman tekniğine dayalı iki yöntem önerilmiştir (Chen ve ark., 2004). İlk yöntem olan ağırlıklı random forest,

azınlık sınıfına ek ağırlıklar koymuştur ve böylece azınlık sınıfının yanlış sınıflandırılmasına daha derin bir disiplin uygulanmıştır. İkinci yöntem olan dengeli rastgele orman ile, örnek azaltma çoğunluk sınıf yöntemini ve toplu öğrenme fikrini ilişkilendirir, sınıf dağılımını yapay olarak dağıtır, böylece sınıflar her ağaçta eşit olarak gösterilebilir.

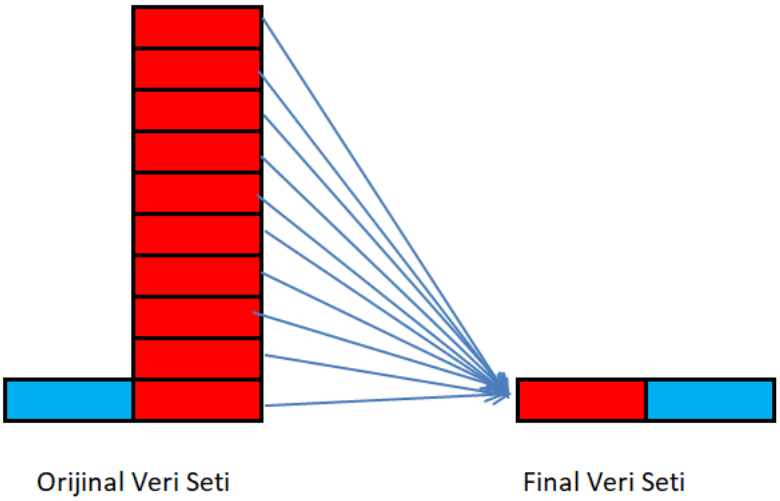
Literatürde, verilerden rastgele ilerlenerek sentetik örnekler oluşturma yoluyla çeşitli sınıf örneklerini dengelemek için kullanılan "Rastgele Yürüme Üstü Örnekleme" çalışması (Zhang ve Li, 2014) da yine dengesiz veri setlerini hedef alarak önerildi. Yöntem, alternatif algoritmalarla önerilen yöntemin, sentetik örnekleri oluşturmak için SMOTE'den aha az zaman harcadığı doğrulanmıştır.

SIMO olarak anılan yeni bir yöntemde ise, eğitim ve test verisi olarak ayrıldıktan sonra eğirim verisi üzerine SVM uygulanarak sınıflar arası karar sınırları belirlenir (Piri ve ark., 2018). Çoğunluk sınıfına yakın azınlık değerleri Safe Level Smote yöntemi ile çoğaltılır. Amaç azınlık verisini çoğaltırken aşırı öğrenmeden kaçınmak için sadece bilgi verici azınlık verilerinin çoğaltılmasıdır. Bu işleme veri dengeli hale gelene kadar devam edilir. Sonrasında SVM yöntemi ile sınıflandırma yapılır. Çalışma sonucu G-Ortalama ölçütüne göre değerlendirilmiş, sık kullanılan SMOTE ve rastgele aşırı örnekleme gibi yöntemlere karşı başarılı sonuç göstermiştir.

Krawczyk ve ark. (2019) tarafından önerilen çalışmada, çok sınıflı problemlere adanmış yeni bir veri örnekleme algoritması olan çok sınıflı radyal tabanlı yöntem ile yüksek hızda örnekleme çalışması yapıldı. Yalnızca azınlık sınıfı özelliklerini kullanan mevcut çok sınıflı yüksek hızda örnekleme yaklaşımlarının aksine, tüm sınıflardan gelen bilgileri dikkate alındı. Yapay örnek oluşturma süreci, karşılıklı sınıf dağılımının değerinin çok küçük olduğu alanları keşfederek yönlendirilir. Bu şekilde, zor veri dağıtımlarıyla başa çıkabilen ve mevcut yöntemlerin eksikliklerini hafifletebilen akıllı bir yüksek hızda örnekleme prosedürü sağlandı. Önerilen algoritmasının kullanışlılığı, kapsamlı deneysel çalışma temelinde değerlendirilmiş olup, sonuçlar kapsamlı bir istatistiksel analiz ile desteklenmiştir.

2.1.3 Örnek Azaltma

Örnek azaltma çoğunluk sınıfına ait verilerin azaltılarak sınıf dengesinin sağlanması yöntemidir. Şekil 2.4.'de örnek azaltma yöntemi görselleştirilmiştir.



Şekil 2.4. Örnek azaltma gösterimi

Rasgele alt-örnekleme yöntemi, azınlık ve çoğunluk sınıfları makul bir boyuta erişene kadar çoğunluk sınıfından verileri rasgele kaldıran sezgisel olmayan bir yöntemdir. Bu veri azaltma işlemi belki de sınıflama ile ilgili yaralı bilginin de atılmasına neden olabilir (He ve ark., 2009).

Örnek azaltma için önerilen diğer bir yöntemde ise, düzenlenen en yakın komşu kuralı ve komşuluk temizleme kuralı kullanılmıştır, üç yakın komşusunun en az ikisinden farklı olan herhangi bir örnek kaldırılmıştır. Bu fikre dayalı olarak, yanlış sınıflandırılan çoğunluk sınıf örneklerini kaldırılır ve gürültü olarak kabul edilir. Bu arada, bir azınlık sınıfı örneği yanlış sınıflandırılsa, onun çoğunluk sınıfına ait komşuları kaldırılır (Laurikkala, 2001).

KNN algoritmasına dayalı olarak alt örnekleme yöntemi olarak önerilen bir çalışmada ise, verileri dengelemek için, her bir sınıfın temel komşu sayısına göre örnekler kaldırılmıştır (Beckmann ve ark., 2015). Önerilen algoritma 33 veri seti üzerinde test edilmiş ve 6 metot ile karşılaştırılmıştır. Diğer yöntemlerle karşılaştırılan sonuçlar, KNN alt örnekleme yönteminin geçerliliği doğrulanmıştır. Yöntem, aynı zamanda, sınıfın üst

üste binmesini engelleyerek karar yüzeyini temizleme işlevi görür ve gürültülü verileri temizler. Sonuçlara göre, KNN az örnekleme yönteminin dengesiz verileri dengelemek için iyi bir makine öğrenimi yaklaşımı olduğunu kanıtlamıştır.

Tomek bağlantıları, sınır çizgisini ve gürültülü verileri tanımlamak için kullanılan alt örnekleme yöntemidir (Chawla, 2009). Tomek bağlantıları, örnekleme yöntemlerinin ürettiği örtüşmeyi ortadan kaldırmak için kullanılan veri temizliği için de kullanılır. Başka bir deyişle, tomek bağlantıları, karşıt sınıfların minimum mesafeli en yakın komşularının bir kombinasyonu olarak tanımlanır. Alt örnekleme yönteminde çoğunluk sınıfı örnekleri kaldırılır (He ve ark., 2009).

Rao ve ark. (2012), sınıf dengesizliği problemini çözmek için görselleştirme kümeleme tekniklerinden biri olan OPTICS'i kullanan alt-örnekleme yaklaşımı ile çalışma yapmışlardır. Çoğunluk sınıfı OPTICS kümeleme tekniğini kullanarak alt örneklenir. Çoğunluk veri seti üzerinde farklı kümeleri tanımlamak için kümeleme algoritması uygulanır. OPTICS sonucu çoğunluk veri setindeki kümelerin sayısını tanımlamak için kullanılır. Zayıf veya aykırı kümeleri tanımlamak ve onları çoğunluk alt kümesinden silmek gerekir. Silme işleminin miktarı veri setinin benzersiz özelliklerine bağlıdır. Zayıf veya aykırı kümeleri çıkardıktan sonra yeni bir çoğunluk alt kümesi oluşur. Yeni çoğunluk alt kümesi ve azınlık alt kümesi yeni ve büyük bir olasılıkla dengeli veri seti oluşturmak için birleştirilir. Bu yeni oluşan dengeli veri seti temel algoritmaya uygulanır. Önerilen yöntemin ROC, F-ölçüsü, hassasiyet, gerçek pozitif oranı ve gerçek negatif oranı değerlendirme ölçülerinde klasik ve yeni yöntemlerden iyi sonuç verdiği doğrulanmıştır.

Diğer bir çalışmada ise, sınıf dengesizliği sorununu ele almak için karınca kolonisi optimizasyonu fikrine dayanan sezgisel bir alt örnekleme yöntemi önerilmiştir (Yu ve ark., 2013). Algoritma, verilerdeki gürültülü verileri çıkarmak için özellik seçme yöntemiyle başlar. Seçim sıklığı temelinde, önemli ve bilgilendirici çoğunluk sınıf örnekleri projelendirilir. Önerilen yöntem, optimal olan çoğunluk denge setini sağlar. Önerilen yöntemin dezavantajı, basit örnekleme yaklaşımlarıyla karşılaştırıldığında daha fazla zaman almasıdır.

DBSCAN algoritması çoğunluk sınıfının uygun örneklerini seçmek için etkili bir alt örnekleme tekniği olarak önerildi (Mirzaei ve ark., 2020). Yapılan çalışmaya göre, çoğunluk sınıfından en uygun örnekler seçilir ve eğitim setini dengelemek için diğer çoğunluk sınıfı örnekler çıkarılır. On beş dengesiz veri setinin üzerindeki deneysel sonuçlar, önerilen yöntemin diğer altı ön işleme yöntemine kıyasla üstünlüğünü göstermektedir.

Rastgele bir ilk seçime dayalı olarak mevcut eğitim setinden en bilgilendirici örnekleri seçmek için bir Naive Bayes sınıflandırıcısının kullanımından yararlanan bir örnek azaltma yaklaşımı Aridas ve diğerleri (2019) tarafından önerildi. Yöntem, küçük tabakalı bir başlangıç eğitim setinde bir Naive Bayes sınıflandırma modelini öğrenerek başlar. Daha sonra, modelin en belirsiz olduğu örneklerle temel modelini yinelemeli olarak öğretir ve bazı kriterler karşılanana kadar onu yeniden dener. Çalışma sonuçları, önerilen örnekleme yönteminin, birkaç uygun metrikle ilgili olarak ve uygun bir istatistiksel test prosedürü gerçekleştirerek, diğer yeniden örnekleme tekniklerinin aksine karşılaştırılabilir sınıflandırma performansına ulaştığını göstermiştir.

2.2 Algoritma Düzeyinde Yöntemler

Algoritmik düzeyde yöntemler genellikle iç yaklaşım olarak adlandırılır, çünkü yeni sınıflandırma algoritma tasarımını kullanır veya dengesiz veriler tarafından üretilen önyargıların üstesinden gelmek için mevcut algoritmaların geliştirilmesi ile ortaya çıkmışlardır (Spelmen ve Porkodi, 2018).

Algoritmik yöntemler ve maliyet tabanlı yöntemler olarak ikiye ayrılırlar.

2.2.1 Algoritmik yöntemler

Dengesiz veri sınıflandırma problemlerini çözmek için yeni algoritmalar oluşturmak ya da mevcut algoritmaları dengesiz veri setleri için geliştirme yaklaşımları algoritmik merkezli yaklaşımlar olarak adlandırılırlar.

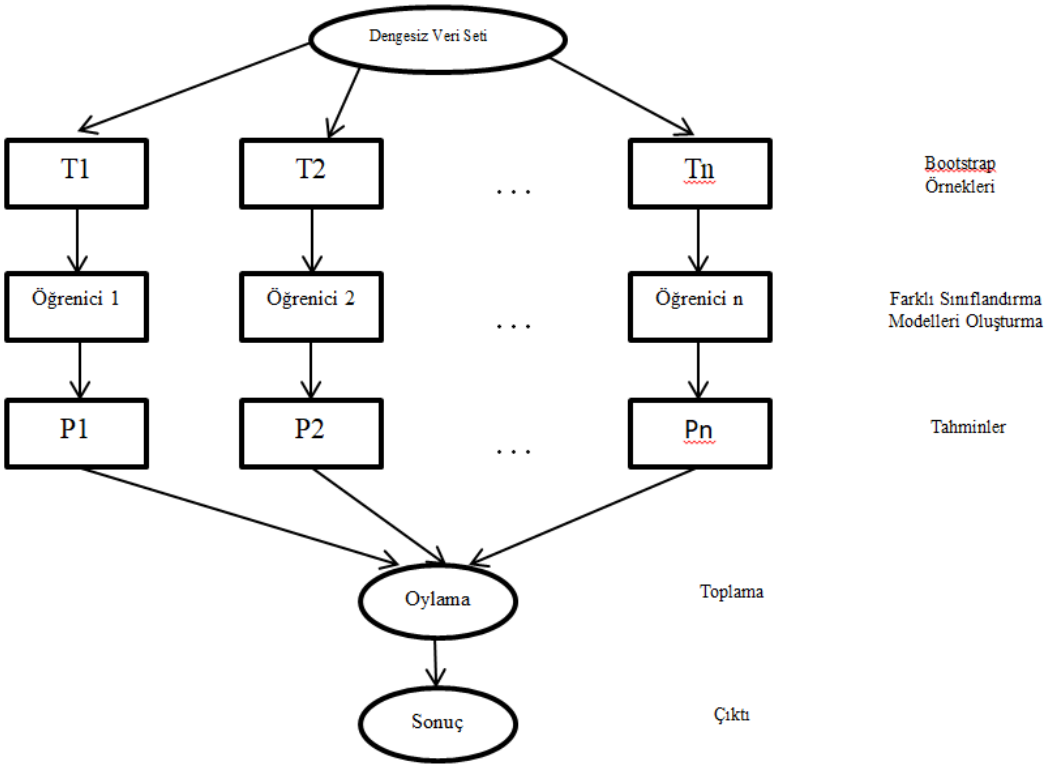
Küme tabanlı örnek azaltma yöntemi, rasgele örnek azaltmaya alternatif olarak küme tabanlı örnek azaltma yöntemi önerilmiştir (Zhang ve ark., 2010). Küme tabanlı az örnekleme yönteminde eğitim veri seti n sayıda kümeye bölünmüştür. Ardından, ayrılan

kümelerden doğru örnekler seçilmiştir. Bu yöntemin arkasındaki temel fikir, eğitim veri alanının n farklı kümeye ayrılması ve her kümenin ayırt edilebilir özellikler ortaya koymasıdır.

Kümelenmeye dayalı farklı bir örnek azaltma tekniğinde ise ilk olarak, eğitim veri setinden çoğunluk sınıfında bulunan gürültülü ve güvenilmez örnekler silinerek temizlenmiştir (Sowah ve ark., 2016). Daha sonra, çoğunluk sınıftaki örneklerin geri kalanı n kümeye bölünmüştür. Deney, diğer mevcut algoritmalardan daha iyi sonuçlar veren iki makine öğrenme algoritması C4.5 Karar Ağacı ve OneR kullanılarak, sınıf dengesizliği problemi olan 16 kıyaslama veri seti üzerinde doğrulanmıştır.

Torbalama metodu, var olan bir eğitim setinden yeni eğitim setleri türeterek temel öğreniciyi yeniden eğiten bir yöntemdir. Bagging'de amaç, yeni veri setleri türeterek farklılıkları oluşturmak ve bu sayede toplam sınıflandırma başarısını artırmaktır (Kaur ve ark., 2019).

Şekil 2.5, dengesiz veri sınıflandırmasının üstesinden gelmek için torbalama algoritmasının genel fikrini temsil etmektedir.



Şekil 2.5. Torbalama algoritması **(**Kaur ve ark., 2019)

Varyansı ve önyargıyı ortadan kaldırmak için çalışan makine öğrenimi topluluğu meta algoritması Boosting olarak adlandırılır. Etkili ve doğru tahmin kuralları, çeşitli zayıf ve yanlış verileri modele entegre ederek oluşturulur. Yanlış ve zayıf sınıflandırma sonuçlarına sahip verilere ait kuralları aramanın, doğru tahmin kuralıyla karşılaştırıldığında çok daha basitleştirilmiş olduğu bu yöntemde ileri sürülmektedir.

2.2.2 Maliyet tabanlı yöntemler

Maliyete duyarlı öğrenme teknikleri, yanlış sınıflandırılmış örneklerle ilişkili maliyeti bulan maliyete özgü tekniklerdir. Genellikle, bunlar yanlış sınıflandırma maliyetini araştırmaya götürür. Örnekleme yöntemlerine kıyasla, maliyete duyarlı öğrenme yöntemleri, yanlış sınıflandırma maliyetinin verilerden belirlenememesi ve maliyetleri belirlerken zorluk çıkması nedeniyle daha az popülerdir. Örnekleme yöntemlerinin uygulanması kolaydır. Ancak maliyete duyarlı öğrenme, hesaplama açısından daha etkili bir tekniktir.

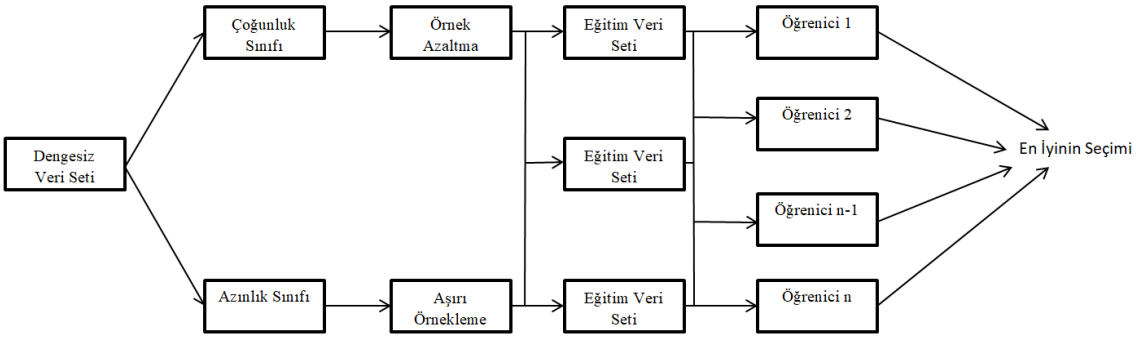
Yapılan bir çalışmada en iyi özellik alt kümesini, içsel parametreleri ve yanlış sınıflandırma maliyet parametrelerini eşzamanlı olarak optimize ederek sınıflandırma performansını iyileştirmek için doğrudan maliyete duyarlı SVM'nin hedef işlevine değerlendirme ölçüsünü (AUC ve G-ortalama) dahil eden bir algoritma önerildi (Cao ve ark., 2013).

Dhar ve Cherkassky (2014), SVM tabanlı çalışmalarında, U-SVM formülasyonunu farklı yanlış sınıflandırma maliyetleri olan sorunlara genişletmiş ve maliyete duyarlı U-SVM'yi dengesiz veri setleri için önermişlerdir.

Çok hedefli bir uyarlanabilir özellik seçim ölçüsü ve karar ağaçlarını oluşturmak ve test etmek için basit ama etkili bir çalışmada, diğer birçok karar ağacı öğrenme algoritması gibi açgözlü bir öznitelik seçim ölçüsü kullanmak yerine, algoritma ağaçtaki her düğümde test etmek üzere uygun bir öznitelik bulmak için rastgele bir öznitelik seçim ölçüsü kullanıldı (Qiu ve ark., 2017). Çalışmada, spesifik olarak, ağaç oluşturmadaki tüm nitelikler alanında rastgele bir arama yapılır ve ortaya çıkan model rastgele seçilmiş karar ağacı olarak adlandırılır. Bu şekilde, yöntem toplam test maliyetini önemli ölçüde düşürürken aynı zamanda rakiplerine kıyasla daha yüksek sınıflandırma doğruluğunu korur.

2.3 Melez Yöntemler

Melez yöntemler ön işleme ve algoritma düzeyinde yöntemlerinin birleştirişmesi ile oluşmuş yöntemlerdir. Amaç sınıf degensizliği probleminin üstesinden gelmek ve daha iyi doğruluk sonuçlarına ulaşmaktır. Melez yöntemlerin olası çalışma kombinasyonları Şekil 2.6’ da gösterilmiştir.



Şekil 2.6. Dengesiz veri sınıflandırması için olası melez kombinasyonlar **(**Kaur ve ark., 2019)

Hastane enfeksiyonunun tespiti için yapılan bir çalışmada, pozitif veya enfekte (% 11) ve negatif (% 89) vakalar arasındaki önemli dengesizlik bulunmaktadır. Sınıf dengesizliğini gidermek için;

- Hem nadir pozitiflerin yüksek hızda örneklemesinin hem de enfekte olmayan çoğunluğun örnek azaltması yoluyla oluşturulan sentetik vakalara (prototipler) dayandığı yeni bir yeniden örnekleme yaklaşımı,
- Pozitif vakaların tanınmasını arttırmak için asimetrik kenar boşluklarının ayarlandığı destek vektör makinelerinin kullanıldığı algoritma önerildi (Cohen ve ark., 2006). Sonuçlar IB1, Naive Bayes, C4.5 ve AdaBoost algoritmaları ile kıyaslanmış ve daha başarılı olduğu doğrulanmıştır.

Diğer bir melez yöntemde, dengesiz veri setinde azınlık türünün sınıflandırılmış performansını etkili bir şekilde geliştirmek için, K-ortalamaları kümesine ve genetik algoritmaya dayalı bir tür azınlık örnekleme yöntemi önerildi (Yong, 2012). Azınlık örnek türünü kümelemek ve gruplamak için K-ortalamaları algoritması kullanıldı ve her kümede yeni örneği elde etmek ve geçerli onayı devam ettirmek için genetik algoritma kullanıldı. Metodun geçerliliği, KNN ve SVM sıralayıcı kullanılarak simülasyon deneyi ile kanıtlanmıştır.

Dengesizlik sınıflandırma sorununu ele almak için RHSBoost geliştirilen toplu sınıflandırma yönteminde bir güçlendirme şeması altında rastgele alt örnekleme ve rastgele aşırı örnekleme yöntemi kullanır (Gong ve Kim, 2017). Deneysel sonuçlara göre,

RHSBoost, dengesizlik verileri için başarılı bir sınıflandırma modeli olarak görünmektedir.

Yapılan başka bir çalışmada ise kompakt ve doğru bir model elde etmek için bilgi tabanının yapısının hiyerarşik bir şekilde genişletilmesi ve bir genetik kural seçim sürecinin kullanılması yoluyla basit bir dilbilimsel bulanık modelin iyileştirilmesine dayanan hiyerarşik bulanık kurala dayalı bir sınıflandırma sisteminin (HFRBCS) kullanılması önerilmiştir (Fernandez ve ark., 2009).

Rastgele düşük örneklemeyi torbalama algoritması ile birleştiren, en basit ve en doğru topluluklardan biri olan RUSBoost'u temel alan yeni bir topluluk oluşturma algoritması olan EUSBoost Galar ve ark. (2013) tarafından geliştirilmiştir. Bu çalışma evrimsel düşük örnekleme yaklaşımının kullanılmasıyla temel sınıflandırıcıların performansını artıran mevcut önerileri iyileştirmeyi amaçlamıştır. Ayrıca, her bir temel sınıflandırıcıyı eğitmek için çoğunluk sınıfı örneklerinin farklı alt kümelerinin kullanımını destekleyen çeşitliliği teşvik etmiştir. Algoritma iki sınıflı dengesiz problemlere odaklanmıştır. Sonuçlar RusBoost yönteminin sonuçları ile karşılaştırılmış ve daha başarılı sonuçlar verdiği doğrulanmıştır.

Ramentol ve ark. (2016), SMOTE-FRST adında yeni bir dengesiz öğrenme ön işleme algoritması önermişlerdir. Çalışmada SMOTE tarafından sunulan sentetik azınlık örneklerini ve gerçek çoğunluk örneklerini temizlemek için iki farklı eşiği kullanarak bulanık kaba küme teorisine (FRST) dayalı bir örnek seçimi stratejisi ile birleştirilmiştir.

Diğer bir çalışmada SMOTE ve parçacık sürüsü optimizasyonu destekli radyal temel işlevi sınıflandırıcısını birleştirerek iki sınıflı dengesiz sınıflandırma problemleri için güçlü bir yöntem önerilmiştir (Gao ve ark., 2011). Sonuçlar dört farklı veri seti üzerinde, üç yöntem ile karşılaştırılmış ve başarılı olduğu doğrulanmıştır.

SVM modellemesine, maliyete duyarlı öğrenme, aşırı ve az örnekleme dahil olmak üzere farklı "yeniden dengeleme" yöntemlerini dahil edilen bir çalışma yapıldı. Bu çalışmada önerilen dört SVM tabanlı algoritmadan GSVM-RU algoritması hem etkinlik hem de verimlilik açısından en etkili olanıdır (Tang ve ark., 2008). GSVM-RU, örnek azaltma

sürecinde veri temizlemenin olumlu katkısını en üst düzeye çıkarırken bilgi kaybının olumsuz etkisini en aza indirebildiği için etkilidir.

İki sınıflı dengesiz veri setlerinin sınıflandırma problemini çözmek için yapılan başka bir çalışmada, ilk olarak SMOTE ile azınlık örneklerinin sayısı arttırılmış, OSS (One Side Selection) yöntemi ile de çoğunluk sınıfı örnekleri azaltılmıştır (Cao ve Zhai, 2015). Çalışmada, sınıflandırıcı olarak ise SVM kullanılmıştır.

2.4 Performans Metrikleri

Model seçimi ve model değerlendirme, makine öğreniminde iki önemli süreçtir. Bu nedenle performans ölçümleri, bir sınıflandırıcının hem etkililiğini değerlendirmek hem de öğrenmesine rehberlik etmek için temel göstergelerdir (Haixiang ve ark., 2017). Sınıflandırma problemlerinde çoğunlukla doğruluk değerlendirme ölçütü olarak kullanılır ancak bu ölçüt dengesiz veri setlerinde tek ölçüt olarak kullanılmamalıdır. %10 azınlık ve %90 çoğunluğa sahip bir veri setinde tüm azınlık sınıfları yanlış da tahmin edilse doğruluk oranı %90 olacaktır. Ancak bu oran modelin başarısının yanlış yorumlanmasına neden olabilir.

Sınıflandırma problemlerinde değerlendirme için genellikle karmaşıklık matrisi kullanılmaktadır. Çizelge 2.1'de karmaşıklık matrisi gösterilmiştir (Sarmanova, 2013).

Çizelge 2.1. Karmaşıklık Matrisi

	Tahmin Pozitif sınıf	**Tahmin Negatif sınıf**
Gerçek Pozitif sınıf	Gerçek Pozitif TP	Yanlış Negatif FN
Gerçek Negatif sınıf	Yanlış Pozitif FP	Gerçek Negatif TN

TP: pozitif sınıfa ait doğru sınıflandırılmış örneklerin sayısı.
TN: negatif sınıfa ait doğru sınıflandırılmış örneklerin sayısı.
FP: negatif sınıfa ait yanlış sınıflandırılmış örneklerin sayısı.

FN: pozitif sınıfa ait yanlış sınıflandırılmış örneklerin sayısı

- Doğruluk; herhangi bir sınıflandırma modeli için, olası tüm toplam örnekler arasında doğru tahmin edilen örneklerin sayısını ölçer. 2.1 numaralı formülde doğruluk oranı gösterilmiştir.

$$Doğruluk\ Oranı = \frac{\mathrm{TP+TN}}{\mathrm{TP+FN+FP+TN}} \qquad (2.1)$$

- Duyarlılık; bir model tarafından doğru bir şekilde tahmin edilen olumlu örneklerin ölçüsüdür. 2.2 numaralı formül ile hesaplanmaktadır. Bazen gerçek pozitif oranı (TPR) olarak da adlandırılır ve başka bir değerlendirme metriğine, yani anma'ya eşdeğerdir.

$$\mathrm{Duyarlılık\ (Gerçek\ Pozitif\ Oranı)} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (2.2)$$

- Özgüllük; bir model tarafından doğru bir şekilde tahmin edilen olumsuz örneklerin ölçüsüdür. Bazen gerçek negatif oranı (TNR) olarak da adlandırılır. 2.3 numaralı formül ile hesaplanmaktadır.

$$\mathrm{Özgüllük}\ (Gerçek\ Negatif\ Oranı) = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (2.3)$$

- Kesinlik; gerçek pozitiflerin (TP) tahmin edilen toplam pozitif örnek sayısına oranı olarak tanımlanır. 2.4 numaralı formül ile hesaplanmaktadır.

$$Kesinlik = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (2.4)$$

- F-ölçüsü; hem gerçek pozitif oranı (True Positive Rate TPR) hem de kesinliği değerlendirir, özellike öğrenme doğruluğu pozitif sınıf üzerinde odaklanır, 2.5 numaralı formül ile hesaplanmaktadır. Diğer bir deyişle, F-ölçütü, anma ve kesinlik arasındaki bir harmonik ortalamadır.

$$F - \mathrm{Ölçütü} = \frac{2\times Kesinlik\times Anma}{\mathrm{Kesinlik+Anma}} \qquad (2.5)$$

- G-Ortalama; hem pozitif sınıf hem de negatif sınıf performansını dikkate alır ve onları birleştirmek için geometrik ortalamayı kullanır. 2.6 numaralı formül ile hesaplanmaktadır. Yüksek G-Ortalama değeri, hem pozitif sınıfı hem de negatif sınıf için yüksek tahmin doğruluğuna sahip olduğu zaman elde edilebilir.

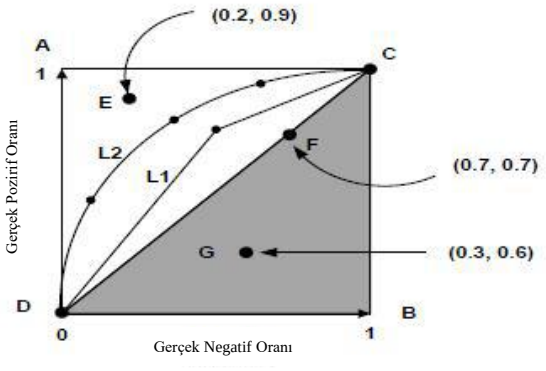
$$G - Ortalama = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (2.6)$$

Çizelge 2.2'de performans metriklerinin hesaplanması çizelge olarak verilmiştir.

Çizelge 2.2. Performans metrikleri hesaplamaları

Metrik	Hesaplama
F-Ölçütü	(2×Kesinlik×Anma)/(Kesinlik+Anma)
G-Ortalama	√ TPR×TNR
Gerçek Pozitif Oranı	TP/(TP+FN)
Yanlış Pozitif Oranı	FP/(FP+TN)
Gerçek Negatif Oranı	TN/(TN+FP)
Kesinlik	TP/(TP+FP)
Anma	TPR

- Alıcı çalışma karakteristiğinin (ROC) eğri altındaki alan (AUC), özellikle ikili sınıflandırıcılar için dengesiz sınıf varlığında genel bir değerlendirme tekniği olarak sıkça kullanılır. AUC grafiği Şekil 2.7'de gösterilmiştir. ROC eğrisi, çeşitli karar eşikleri boyunca gerçek pozitif oranı (TPR) ile yanlış pozitif oranı (FPR) arasındaki olası tüm çatışmaları gösterir ve AUC değerlendirme metriği, bu eğriyi [0,5, 1] aralığında bir değere dönüştürür, burada 1 değeri, mükemmel bir sınıflandırıcıyı gösterir. 0,5 değeri ya da daha düşük değerler, sınıflandırıcının rastgele tahminden daha iyi çalışmadığı anlamına gelir.



Şekil 2.7. AUC grafiği (ROC altında kalan alan) (Sarmanova, 2013)

- MAUC; AUC'yi çok sınıflı problemlere genişletmek açık bir araştırma konusu olsa da, tüm sınıf çiftlerinin AUC değerinin ortalamasını alan MAUC ölçüsü (Haixiang ve ark., 2017), çoğunlukla araştırmalarda çok sınıflı dengesiz veri öğrenmede kullanılır ve 2.7 numaralı formül ile gösterilir;

$$MAUC = \frac{2}{C(C-1)} \sum_{i<j} A(i,j) \qquad (2.7)$$

C = sınıf sayısı

A (i, j) = sınıf i ve sınıf j arasındaki AUC'dir.

- MCC (*Matthews Correlation Coefficient*), Matthews korelasyon katsayısında çıktı değeri -1 ile +1 arasında değişmektedir. 0 değeri rastgele sınıflandırma durumunu, -1 değeri sınıflandırmanın başarısız olduğunu, +1 ise sınıflandırma başarısının tam doğru olduğunu göstermektedir (Matthews, 1975). MCC 2.8 numaralı formül ile hesaplanmaktadır.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \qquad (2.8)$$

Bu çalışmada hem pozitif sınıf hem de negatif sınıf performansını dikkate alan G-Ortalama değerlendirme ölçütü olarak kullanılacaktır.

3. MATERYAL ve YÖNTEM

Piri S. ve ark. (2018) tarafından yapılan çalışmada önerilen SIMO yönteminde veri eğitim ve test verisi olarak ayrıldıktan sonra eğitim verisi üzerine SVM uygulanarak sınıflar arası karar sınırları belirlenmiştir. Çoğunluk sınıfına yakın azınlık değerleri SLS yöntemi ile çoğaltılmıştır. Amaç azınlık verisini çoğaltırken aşırı öğrenmeden kaçınmak için sadece bilgi verici azınlık verilerinin çoğaltılmasıdır. Bu işleme veri dengeli hale gelinceye kadar devam edilmiştir. Sınıflandırıcı olarak SVM yöntemi kullanılmıştır. Sonuçlar G-Ortalama ölçütüne göre değerlendirilmiş, klasik örnekleme yöntemlerinden başarılı sonuç verdiği doğrulanmıştır.

Bu çalışmada, eğitim verisine önce SVM uygulanarak karar sınırına uzak çoğunluk sınıf verileri veri setinden çıkarılmıştır. Sonrasında örnek azaltma işlemine uğrayan veri, hala dengeli değil ise SLS yöntemi ile azınlık verilerinin güvenli sınırda çoğaltılması sağlanmıştır. Bu işlem veri seti dengeli hale gelinceye kadar devam edilmiştir. Tez kapsamındaki çalışmanın SIMO dan farkı, SIMO da sadece bilgi verici azınlık sınıfı verileri çoğaltılırken, önerdiğimiz yöntemde azınlık sınıfının çoğaltılmasının yanında, bilgi vermeyen çoğunluk sınıfı verileri çıkarılarak örnek azaltmada problem olan değerli veri kaybı önlenmeye çalışılmıştır. Böylece SMOTE de karşılaşılan aşırı uyum sorununun önüne geçilmeye çalışılmış, rastgele örnek azaltma yönteminde rastlanan değerli veri kaybının önüne geçilmeye çalışılmıştır. Sonrasında sınıflandırma algoritmaları uygulanarak sınıflandırma başarısının arttırılması amaçlanmıştır.

3.1 Kullanılan Veri Setleri

Çalışma kapsamında kullanılan veri setleri Kaggle ve UCI veri havuzundan alınmıştır, 1 adette gerçek hayat verisine yer verilmiştir. Veri setleri, karşılaştırma yapabilmek için literatür taramasında benzer çalışmalarda kullanılan veri setleri arasından seçilmiştir. Verilerde min-max normalizasyon işlemi uygulanmıştır. Veri setlerine ait özet bilgi Çizelge 3.1'de gösterilmiştir.

Çizelge 3.1. Veri setleri

Veri Setleri	Öznitelik Sayısı	Gözlem Sayısı	Pozitif Sınıf	Negatif Sınıf	Dengesizlik Oranı (Np/Nn)
Climate	21	594	494	46	10,74
Diabetes	9	768	500	268	1,87
Liver	11	583	416	167	2,49
Haberman	4	305	224	81	2,77
Transfusion	5	748	570	178	3,20
Ionosphore	32	351	225	126	1,79
Column_2c	7	310	210	100	2,10
TexYarn	8	979	937	42	22,31

Gerçek Veri Seti: TexYarn: Bir dokuma firmasının gerçek verilerinin bulunduğu veri seti, dokumaya gidecek olan ve tedarikçiden temin edilen ipliğin kalite kontrol aşamasında ve kontrol sonrası onay aldıktan sonraki süreçlerdeki proses parametreleri ile işlem gördükten sonra dokuma aşamasında kopup kopmayacağını tahmin etmeye çalışacaktır. Buna göre, eğer işlem görmüş iplik kopacak olarak etiketlenirse, dokuma aşamasına girmeyecek ve böylece verimlilik kaybı engellenmiş olacaktır. Veri setindeki öznitelikler ve açıklamaları aşağıdaki gibidir:

1. Giriş Kontrol Parametreleri: Aşağıdaki öznitelikler, tedarikçiden gelen iplik lotlarına, kalite kontrol aşamasında uygulanan testlerdir. Sadece belirlenen aralıklardaki değerleri gerçekleyebilen iplikler üretim alanına alınabilir.
   a. Kaynama Çekme (Değerler;1,4 – 67 aralığında değişmektedir.)
   b. Kopma Yükü (Değerler;124,19 – 2329,2 aralığında değişmektedir.)
   c. Mukavemet (Değerler;1,4 – 4,85 aralığında değişmektedir.)
   d. Numara Denye (Değerler;30 – 673 aralığında değişmektedir.)
   e. Uzama (Değerler;14 – 221,84 aralığında değişmektedir.)
2. Üretim Parametreleri: Aşağıdaki öznitelikler, iplik lotlarının üretim alanına alındıktan sonra, dokuma aşamasına geçmeden önceki süreçlerde gördüğü işlemlerdeki proses parametrelerini göstermektedir.

a. Bekleme Süresi (Değerler; 30 – 50 aralığında değişmektedir.)
b. İplik Fikse Sıcaklık (Değerler; 80 – 122 aralığında değişmektedir.)

Bu veri setinde amaç, bahsi geçen 7 özniteliği kullanarak dokumaya girmeden önce ipliğin kopup kopmayacağını tahmin etmek olacaktır. Önerilecek karar destek sistemi ile kopuş yaşayacak ipliklerin üretime girmesi engellenerek verimlilik artışı sağlanabilecektir.

3.2 Kullanılan Yazılım ve Paketler

Uygulamada istatiksel yazılım geliştirme ve veri analizi programı olan R programının "R i386 4.0.3" versiyonu kullanılmıştır (Anonim, 2020). Kullanılan paketler ve parametre değerleri Çizelge 3.2'de özetlenmiştir.

Çizelge 3.2. Kullanılan algoritmalar ve parametreler

Algoritma	R paketi	Parametreler
SVM	e1071	kernel= Radial/Linear; sigma=0.01,0.015 ; C=0.75,1,1.25
KNN	class	k=1:20;preProc= "center","scale"
RF	randomForest	mtry= 1:10; method='rf'; metric= 'Accuracy'
YSA	nnet	decay = 0.001,0.01, 0.1; size = 1:10
SLS	smotefamily	K=5; C=5

3.3 Karşılaştırmada Kullanılan Algoritmalar

3.3.1 K-en yakın komşu algoritması (KNN)

1951 yılında Evelyn Fix ve Joseph Hodges tarafından önerilen K-En yakın komşu algoritmasının çalışma mantığını en basit şekilde özetleyen cümle "Bana arkadaşını söyle sana kim olduğunu söyleyeyim." şeklindedir. Algoritma test verisine en yakın K adet komşuyu bulup, bu komşulukta baskın olan sınıfı etiketini test verisine sınıf etiketi olarak atar.

K sayısı komşuluk sayısını ifade eder, uzaklık için farklı uzaklık ölçüm kriterleri mevcuttur. K sayısının küçük bir değer olması aşırı öğrenmeye, yüksek bir değer olması ise genellemeye yol açar.

Uzaklık ölçüm kriterleri (Balaban ve Kartal, 2015);

n adet nitelik, $a_r(x_{i,j})$ x'in r. niteliğindeki değerini ve x_i, x_j gibi iki örnek arasındaki uzaklık olmak üzere;

- Öklid uzaklığı 3.1 numaralı formül ile gösterilmiştir.

$$d_{öklid}(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2} \quad (3.1)$$

- Manhattan uzaklığı 3.2 numaralı formül ile gösrerilmiştir.

$$d_{manhattan}(x_i, x_j) = \sum_{r=1}^{n}\left|a_r(x_i) - a_r(x_j)\right| \quad (3.2)$$

- Hamming uzaklığı 3.3 numaralı formül ile gösterilmiştir.

$$d_{hamming}(x_i, x_j) = \sum_{r=1}^{n} \mathrm{I}(a_r(x_i), a_r(x_j)); \mathrm{I}(\mathrm{x}, \mathrm{y}) \quad (3.3)$$
$$= \begin{cases} 0, eğer\ x = y \\ 1\ eğer\ x \neq y \end{cases}$$

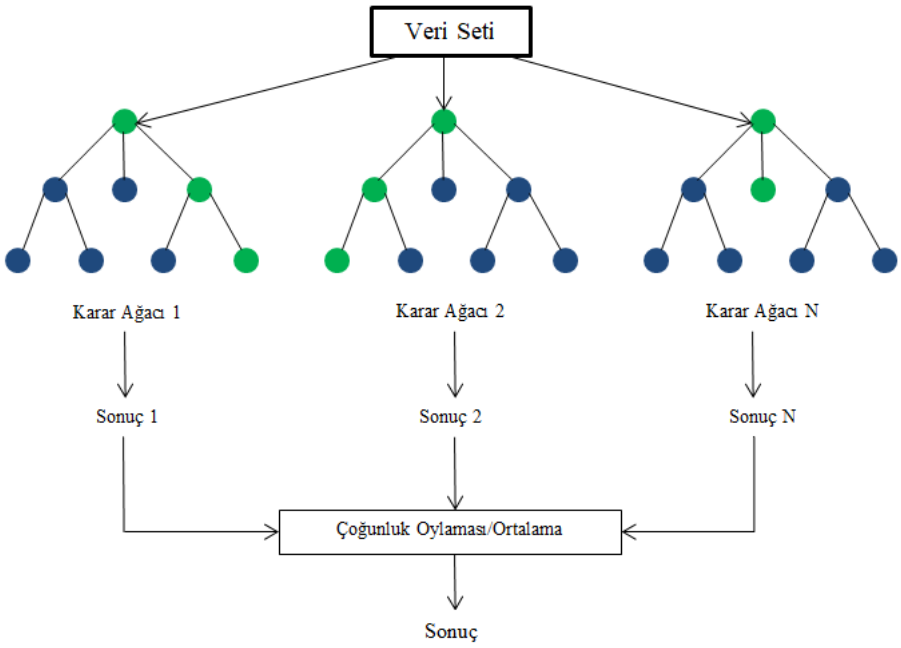
- Kosinüs uzaklığı 3.4 numaralı formül ile gösterilmiştir.

$$d_{kosinüs}(x_i, x_j) = \frac{x_i . x_j}{\|x_i\|\|x_j\|} = \cos\theta \quad (3.4)$$

3.3.2 Rastgele orman algoritması (RF)

Random Forests (rastgele ormanlar) algoritması, her ağacın bağımsız olarak örneklenen rastgele bir vektörün değerlerine bağlı ve ormandaki tüm ağaçlar için aynı dağılıma sahip olduğu, ağaç tahmin edicilerinin bir kombinasyonudur. Şekil 3.1 de RF algoritmasının gösterimi bulunmaktadır. Denetimli bir sınıflandırma algoritmasıdır. Sınıflandırma sonucu karar ağaçlarının çoğunluk oylaması ile belirlenir. RF için genelleme hatası,

ormandaki ağaç sayısı arttıkça bir sınıra yakınsamaktadır. Bir ağaç sınıflandırıcı ormanının genelleme hatası, ormandaki tek tek ağaçların gücüne ve aralarındaki korelasyona bağlıdır. Dahili tahminler hatayı, gücü ve korelasyonu izler ve bunlar, bölmede kullanılan özelliklerin sayısını artırmaya verilen yanıtı göstermek için kullanılır. Dahili tahminler değişken önemini ölçmek için de kullanılır (Breiman, 2001).



Şekil 3.1. Rastgele orman algoritması

3.3.3 Destek vektör makineleri algoritması (SVM)

Destek vektör makineleri iki veri sınıfını birbirinden ayırabilecek en iyi karar sınırı/hiper düzlem bulmayı amaçlayan bir sınıflandırma algoritmasıdır. Hiper düzlemin geniş olması iki sınıfın daha kolay şekilde birbirinden ayrılmasına olanak sağlar (Boser ve ark., 1992).

SVM doğrusal olarak ayrılabilen ve doğrusal olarak ayrılamayan olaran iki grupta incelenmektedir.

- Doğrusal olarak ayrılabilen SVM;

Bu durumda verileri iki sınıfa ayırabilmek için kullanılabilecek doğrusal bir hiper düzlem vardır.

Hiper düzlem 3.5 numaları formül ile bulunabilmektedir, x girdi vektörü, w hiper düzlemde ağırlık vektörü ve b sapma olmak üzere karar sınırları;

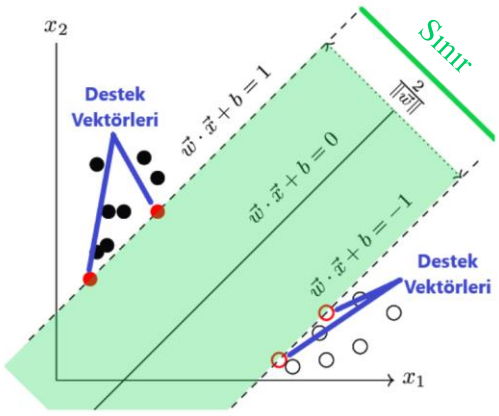
$$wx + b = 0 \tag{3.5}$$

Marjini en büyükleme 3.6 numaralı formülde gösterildiği gibi bulunmaktadır.

$$marjin = \frac{2}{\|\vec{w}\|^2} \tag{3.6}$$

Kısıtlar 3.7 numaralı formülde gösterilmiştir.

$$f(\vec{x_i}) = \begin{cases} 1\,, eğer\ \vec{w}.\vec{x_i} + b \geq 1 \\ -1, eğer\ \vec{w}.\vec{x_i} + b \leq -1 \end{cases} \tag{3.7}$$



Şekil 3.2. Doğrusal olarak ayrılabilen SVM karar düzlemi

- Doğrusal olarak ayrılamayan SVM;

Gerçek dünyada karşılaşılan sorunların büyük bir kısmı, veri setini doğrusal bir şekilde ayırabilecek tek bir hiper düzlemin olmadığı verileri içermektedir. Bu sorunu çözmek için, veriler nispeten daha yüksek boyutlu bir uzaya eşlenir ve sonra orada bir hiper düzlem tanımlanır (Maglogiannis, 2007). 3.8 numaralı formülde eşleme çözümünde SVM formülü verilmiştir.

y_i = çıktı etiketi olmak üzere 3.8 numaralı denklem Lagrange çarpanları yöntemi ile çözüldüğünde,

$$\min \frac{(\| w \|)^2}{2}$$
$$y_i(wx + b) \geq 1 y_i \in (-1, +1) \quad (3.8)$$
$$L_p(w, b, a) = \frac{1}{2} \| w \| - \sum_{i=1}^{l} a_i [y_i(wx + b) - 1]$$

Bu durumda sınıflandırma fonksiyonu formül 3.9'daki gibi olacaktır.

$$f(x) = sgn\big((wx_i) + \mathrm{b}\big) = \mathrm{sgn}\big(\sum_{i=1}^{i} y_i \, a_i \varphi(x_i) \varphi\big(x_j\big) + b\big) \quad (3.9)$$

Doğrusal olarak ayrılamayan SVM'de, hesaplanması gereken miktarlar $(x_i)\varphi\big(x_j\big)$, temel özellikleri olan skaler çarpımlardır. Buna kernel fonksiyonu (K) denmektedir. Kernel fonksiyonu kullanıldığında SVM 3.10 numaralı formülde gösterildiği gibi formüle edilecektir.

$$f(x) = \mathrm{sgn}\big(\sum_{i=1}^{i} y_i \, a_i K\big(x_i, x_j\big) + b\big) \quad (3.10)$$

Çalışmalarda Radial Basis ve Sigmoid kernel sıkça kullanılmaktadır. Bu kerneller 3.11 ve 3.12 nolu formüller ile gösterilmiştir (Akın ve Terzi, 2021).

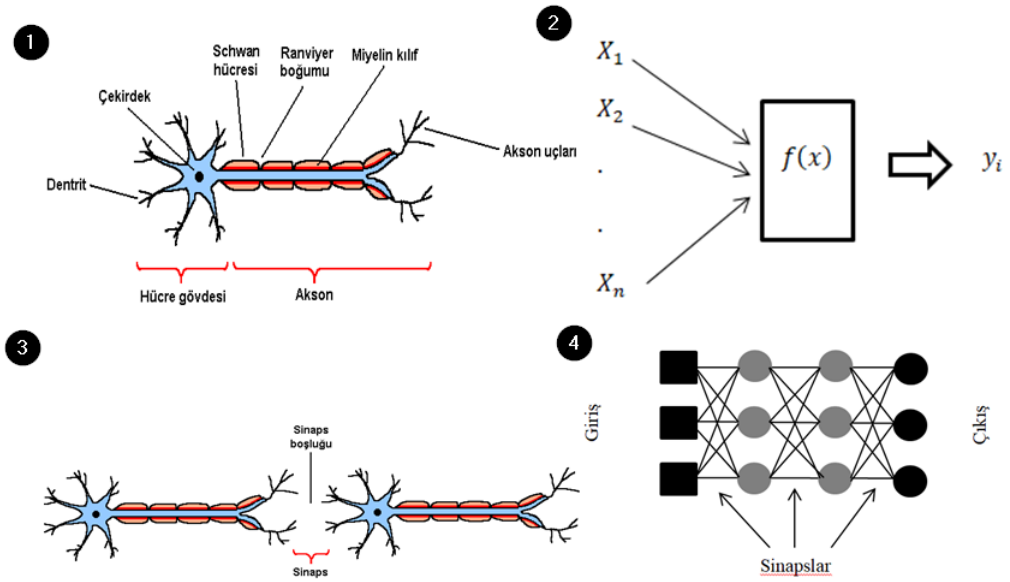
$$Radial\ Basis : K\big(X_i, X_j\big) = e \quad (3.11)$$
$$Sigmoid\ Kerneli: K\big(X_i, X_j\big) = \tanh(kX_i, Xj - \delta) \quad (3.12)$$

Bu çalışmada SVM doğrusal ve Radial Basis kernel için ayrı ayrı denenmiştir.

3.3.4 Yapay sinir ağları algoritması (YSA)

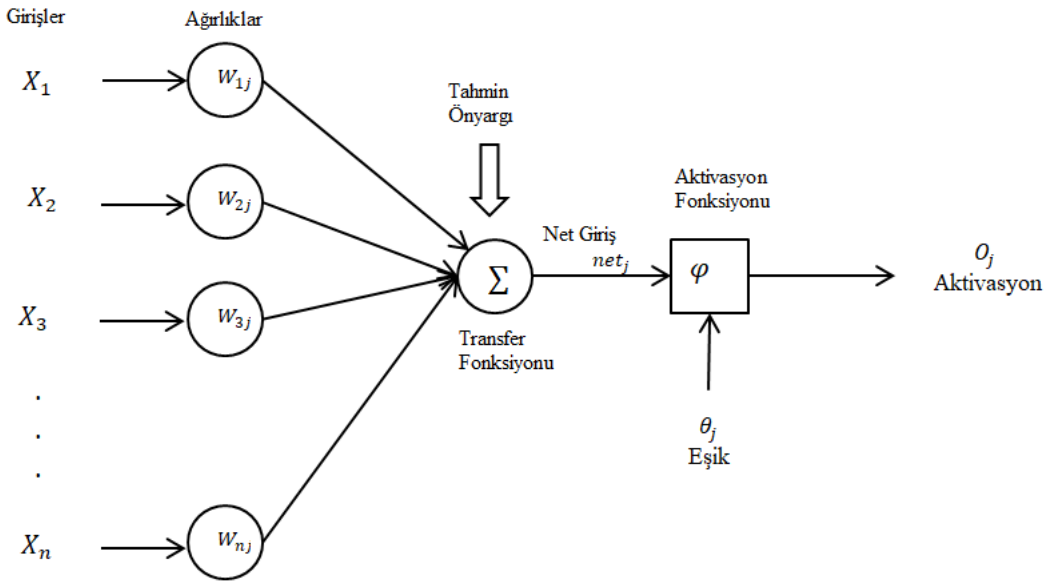
Yapay sinir ağları ilk olarak Warren McCulloch ve Walter Pitts tarafından önerilmiştir. Yapay sinir ağları insan beyninin çalışma prensibini taklit ederek öğrenme sürecinin matematiksel olarak modellenmesidir. Bu yol ile öğrenme, hatırlama, genelleme yapma yolu ile topladığı verilerden yeni veri üretebilme gibi temel işlevleri yapabilmektedir (McCullock ve Pitts, 1956).

Şekil 3.3'te biyolojik sinir ağları ile yapay sinir ağları arasındaki benzerlik gösterilmiştir.



Şekil 3.3. Biyolojik sinir hücresi ve yapay sinir ağı (Maltarollo ve ark., 2013)

Şekil 3.4'de gösterildiği gibi, Girilen n adet veri ağırlıklarla çarpılır ve tüm veriler toplanır, sonrasında önyargı eklenir bunun sonucunda net yargı elde edilir. Net girdi aktivasyon fonksiyonundan geçirilir ve bir veri çıktısı elde edilmiş olur.



Şekil 3.4. Yapay sinir hücresi

Aktivasyon fonksiyonları bir nöronun aktive edilip edilmeyeceğine ağa girişinin önemli olup olmadığına daha basit matematiksel işlemler kullanarak karar veren fonksiyonlardır.

Başlıca aktivasyon fonksiyonları (Baheti, 2021);

- İkili adım fonksiyonu, nöronun aktivasyonunun belirli bir eşik değerine bağlı olduğu fonksiyondur.
- Doğrusal aktivasyonu fonksiyonu, nöron aktivasyonunun giriş ile basit bir regresyon modeli ile bağlı olduğu fonksiyondur.
- Doğrusal olmayan aktivasyon fonksiyonları ise nöron aktivasyonunun basit bir regresyon modeli ile yapılmadığı durumda kullanılan fonksiyon türüdür. Başlıca doğrusal olmayan fonksiyonlar;
    a) Sigmoid / Lojistik fonksiyonu 0 ile 1 arasındaki değerleri verir.
    b) Tanh fonksiyonu
    c) ReLU fonksiyonu
    d) Sızdıran ReLU fonksiyonu
    e) Parametrik ReLU fonksiyonu
    f) Üstel doğrusal birimler (ELU's) fonksiyonu
    g) Softmax fonksiyonu
    h) Swish fonksiyonu
    i) Gauss hatası doğrusal birimi (GELU) fonksiyonu
    j) Ölçekli üstel doğrusal birim (SELU) fonksiyonu

Bu çalışmada lojistik fonksiyon kullanılmıştır.

3.4 Veri Hazırlama

Verileri R programına “Utils” paketi içindeki “read.csv” komutu kullanarak tanıtılmaktadır.

Veriler içeri alındıktan sonra min-max normalizasyon işlemini 3.13 numaralı formül kullanarak yapılmaktadır (Balaban ve Kartal, 2015. Bu işlem sırasında tahmin edilecek sınıf etiketi bu işleme dahil edilmemektedir.
Min-max normalizasyon yöntemi;

v: A niteliğime ait normalize edilmek istenen değer

v': v'nin normalize edilen değeri

min_A: A niteliğinin en küçük değeri

max_A : A niteliğinin en büyük değeri

$yeni\_min_A$: Normalizasyon sonucunda elde edilmek istenen en küçük değer

$yeni\_max_A$: Normalizasyon sonucunda elde edilmek istenen en büyük değer

$$v' = \frac{v - min_A}{max_A - min_A}(yeni\_max_A - yeni\_min_A) + yeni\_min_A \qquad (3.13)$$

Tahmin edilecek sınıf numerik veya kategorik olduğu durumlarda sınıf etiketi "0", "1" etiketi ile değiştirilmektedir. R içerisindeki SMOTE paketleri azınlık sınıf etiketini "1" olarak tanımaktadır. Bu nedenle ham veride çıktı etiketi "0", "1" olduğu durumlarda azınlık verisi "1" değilse, etiketlerde değişim yapılmaktadır. Sonrasında veriler faktör olarak tanımlanmaktadır.

Veride, eksik veri ya da kategorik veri olduğu durumlarda bu veriler veriden çıkarılarak devam edilir.

3.5 Eğitim – Test Verisi Ayırma

Veriler R içerisindeki "caret" paketinin "createDataPartition" fonksiyonu ile ayrılmaktadır. Ayrılan test ve eğitim verileri sonraki işlemlerde de kullanıldığı için ayırma işlemi öncesi "set.seed()" fonksiyonu ile sabitlenir.

Veriler %80 eğitim ve %20 test verisi şeklinde ayrılmaktadır. Önce "createDataPartition" ile %80'lik dilime ait verilerin indekslerine ulaşılmaktadır. Bu veri "train" verisine atanır, daha sonra %20'lik veri test verisine atanmaktadır. Sonrasında sınıflandırma algoritmalarında kolaylık sağladığı için sınıf etiket değerleri çıkarılarak "train_x, train_y, test_x, test_y" veri setleri oluşturulmaktadır. Test verisi içerisindeki dengesizlik oranına bakılmamaktadır. Örnekleme işlemi sırasında, test verisinde aşırı örnekleme ya da örnek azaltma kullanılmamaktadır. Orjinal veri setinden alınan test seti ile işlemlere devam edilmektedir.

3.6 Melez Aşırı Örnekleme ve Alt Örnekleme Yöntemi

Örnek azaltma ve çoğaltma işlemleri sadece eğitim verisi üzerine uygulanmaktadır. Test verisi üzerinde işlem yapılmamıştır.

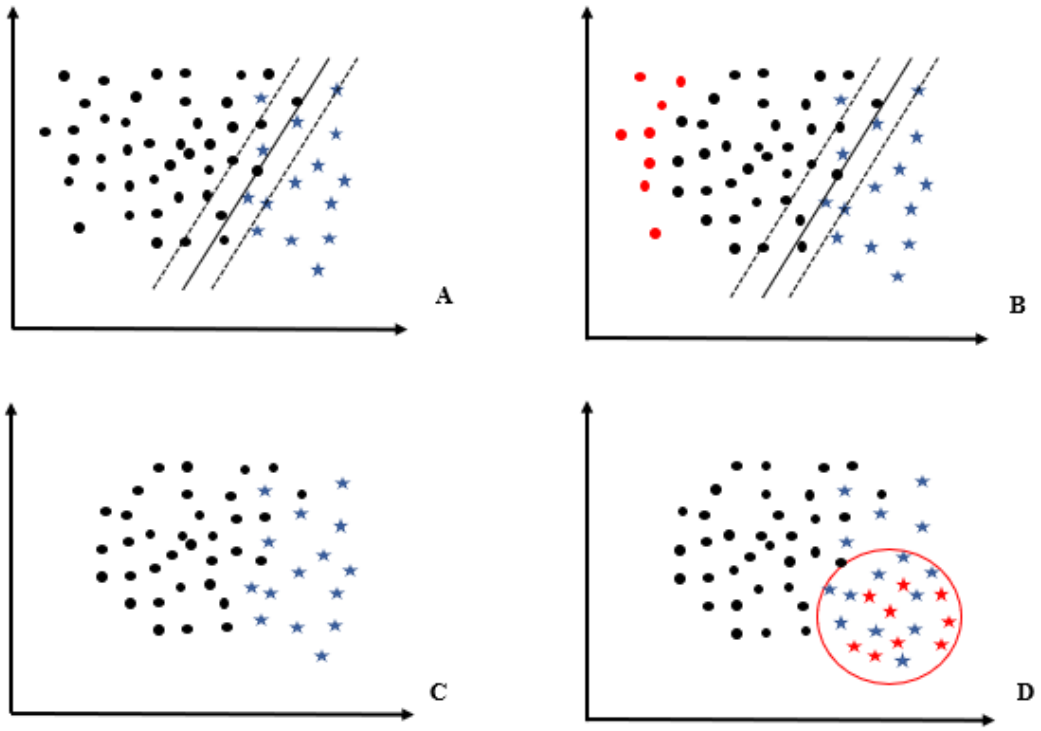
Melez yöntemin uygulanmasında, örnek azaltma adımında kullanılan SVM fonksiyonu için "e1071" paketi kullanılmış olup, hem radial hem de lineer kernelleri için veri türetilmiştir.

Karar değişken değerleri (decision.values) bulunarak çoğunluk veri sınıfına ait karar değerlerinin ilk çeyrek öncesinde kalan değerler veriden çıkarılmıştır.

Sonrasında indirgenmiş eğitim verisine "smotefamily" paketindeki SLS fonksiyonu ile azınlık sınıf verileri safe level smote yöntemi ile çoğaltılmıştır.

Safe level smote yöntemi ile üretilen veriler "syn_data" içinde bulunmaktadır. Bu veriler indirgenmiş eğitim verisine eklenerek azınlık sınıf veri sayısı arttırılmıştır. Bu işlem sırasında üretilen yeni çıktı değerleri faktör olmadığı için tekrar faktöre çevrilmiştir.

Bu işlemlere azınlık ve çoğunluk sınıflarının oranı %50 ± 5 oluncaya kadar devem edilmiştir. Ardından sınıflandırıcı algoritmalar uygulanarak sonuçlar dengesiz dağılama sahip ham veri ve SMOTE uygulanan veri ile karşılaştırılmıştır.



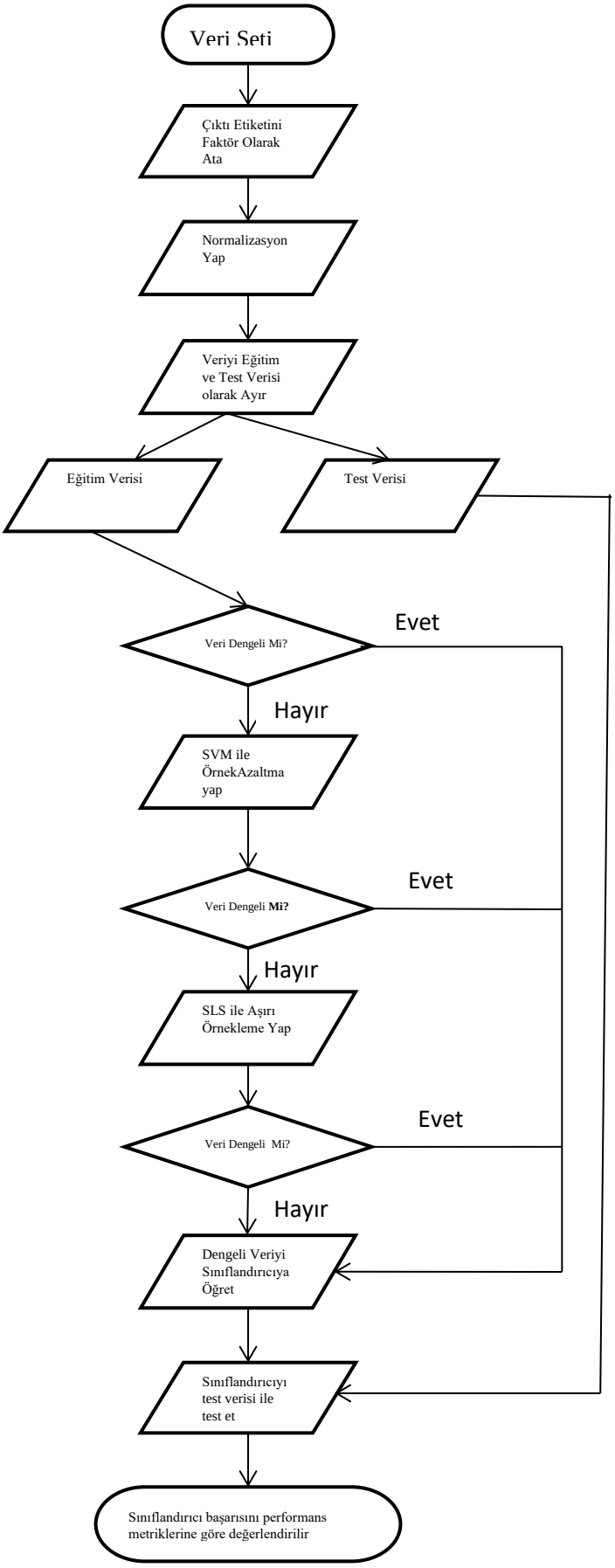
Şekil 3.5. Melez yöntem adımları **A)** SVM uygulanması **B)** Karar sınırına uzak çoğunluk verilerinin seçilmesi **C)** Örnek azaltma işleminin yapılması **D)** SLS ile aşırı örnekleme yapılması

Şekil 3.5'te akış şeması verilen melez yöntemin uygulama adımları;

1. Adım: Sınıf etiketleri "0", "1" olarak atanır.
2. Adım: Min-max normalizasyonu yapılır.
3. Adım: Veri eğitim ve test seti olarak ayrılır.
4. Adım: Eğitim verisinin dengeli olup olmadığı kontrol edilir. Denge koşulu sınıflar değılımının %50 ± 5 olmasıdır. Dengeli ise 9. adıma gidilir.
5. Adım: Şekil 3.5 a'da gösterildiği gibi dengesiz olan eğitim verisine SVM uygulanır.
6. SVM ile karar sınırına uzak çoğunluk verilerine örnek azaltma işlemi uygulanır (Şekil 3.5 b, c).
7. Adım: Eğitim verisinin dengeli olup olmadığı kontrol edilir. Dengeli ise 9. adıma gidilir.

8. Adım: Şekil 3.5 d'de gösterimi yapılan, SLS ile aşırı örnekleme yapılır. 4. Adıma dönülür.
9. Adım: Dengeli hale gelen veri sınıflandırma algoritmaları ile sınıflandırılır.
10. Adım: Sınıflandırıcılar test verisi ile test edilerek karmaşıklık matrisi üzerinden sonuçlar yorumlanır.

Şekil 3.6.'da melez yöntem akış şeması gösterilmiştir.



Şekil 3.6. Melez yöntem akış şeması

4. BULGULAR

Melez yöntem algoritması incelenen 8 veri setinden 6'sında diğer yöntemlere gore daha başarılı sonuç vermiştir. Veri setleri bazında elde edilen sonuçlar aşağıda paragraflar halinde derlenmiştir. İncelenen veri setlerine göre, önerilen melez yöntemin daha başarılı sonuç verdiği sınıflandırmalarda, dengesizlik oranı 2,49'a kadar olan verilerde melez yöntem - Radial SVM, 2,49'dan yüksek dengesizlik oranlarında ise melez yöntem – Linear SVM başarılı olmuştur. Veri setleri, sınıflandırıcılar ve algoritmalar ile elde edilen sonuçlar Çizelge 4.1'deki gibidir.

- Climate veri setinde en iyi sonucun, SMOTE uygulanmış veri seti ile yapay sinir ağları algoritmasında % 96 doğruluk ve 0,87 G-Ortalama değeriyle alındığı gözlemlenmiştir. Radial SVM ile melez yöntem uygulanan veri seti ve SVM algoritması ise % 95 doğruluk ve 0,808 G-Ortalama değeri ile ikinci en iyi sonucu vermiştir. Tüm algoritmalarda en kötü sonucu dengesiz veri seti vermiştir.
- Diabetes veri setinde en iyi sonucun radial SVM ile melez yöntem uygulanan veri seti ile yapay sinir ağları algoritmasında % 75 doğruluk ve 0,727 G-Ortalama değeriyle alındığı gözlemlenmiştir. SMOTE uygulanan veri seti ile SVM algoritması ise % 71 doğruluk ve 0,713 G-Ortalama değeri ile ikinci en iyi sonucu vermiştir. Tüm algoritmalarda en kötü sonucu ise dengesiz veri seti vermiştir.
- Liver veri setinde en iyi sonucun radial SVM ile melez yöntem uygulanan veri seti ile yapay sinir ağları algoritmasında % 69 doğruluk ve 0,736 G-Ortalama değeriyle alındığı gözlemlenmiştir. SMOTE uygulanan veri seti ile yapay sinir ağları algoritması ise % 67 doğruluk ve 0,703 G-Ortalama değeri ile ikinci en iyi sonucu vermiştir. Tüm algoritmalarda en kötü sonucu ise dengesiz veri seti vermiştir.
- Haberman veri setinde en iyi sonucun linear SVM ile melez yöntem uygulanan veri seti ile yapay sinir ağları algoritmasında % 75 doğruluk ve 0,705 G-Ortalama değeriyle alındığı gözlemlenmiştir. SMOTE uygulanan veri seti ile yapay sinir ağları algoritması ise % 68 doğruluk ve 0,703 G-Ortalama değeri ile ikinci en iyi sonucu vermiştir. En kötü sonuç ise dengesiz veri setine SVM algoritmasının uygulanması ile alınmıştır.

- Transfusion veri setinde en iyi sonucun linear SVM ile melez yöntem uygulanan veri seti ile yapay sinir ağları algoritmasında % 77 doğruluk ve 0,732 G-Ortalama değeriyle alındığı gözlemlenmiştir. Radial SVM ile melez yöntem uygulanan veri seti ile yapay sinir ağları algoritması % 79 doğruluk ve 0,728 G-Ortalama değeri ile ikinci en iyi sonucu vermiştir. Tüm algoritmalarda en kötü sonucu ise dengesiz veri seti vermiştir.
- Ionosphore veri setinde en iyi sonucun radial SVM ile melez yöntem uygulanan veri seti ile random forest algoritmasında % 97 doğruluk ve 0,959 G-Ortalama değeriyle alındığı gözlemlenmiştir. Dengesiz veri ve SMOTE uygulanan veri seti %95 doğruluk ve 0,948 G-Ortalama değeri ile ikinci en iyi sonucu vermiştir. En kötü sonuç % 87 doğruluk ve 0,800 G-Ortalama değeri ile dengesiz veri setine K-en yakın komşuluk algoritması uygulanması ile alınmıştır.
- Column_2c veri setinde en iyi sonucun dengesiz veriye ait veri seti ile random forest algoritmasında % 83 doğruluk ve 0,842 G-Ortalama değeriyle alındığı gözlemlenmiştir. % 82 doğruluk ve 0,830 G-Ortalama değerleriyle linear SVM ile melez yöntem ve radial SVM ile melez yöntem uygulanan veri setleri random forest algoritmasında en iyi ikinci sonucu vermişlerdir. En kötü sonuç dengesiz veri setine yapay sinir ağları algoritması uygulanması ile % 77 doğruluk ve 0,717 G-Ortalama değerinde alınmıştır.
- TexYarn veri setinde en iyi sonucun YSA algoritmasında linear SVM ile melez yöntemin uygulandığı veri setinde % 98 doğruluk ve 0,991 G-Ortalama değeriyle alındığı gözlemlenmiştir. Linear SVM ile melez yöntem uygulanan veri seti SVM algoritmasında % 98 doğruluk ve 0,970 G-Ortalama değeri ile ikinci en iyi sonucu vermiştir. En kötü sonuç ise dengesiz veri setine KNN ve YSA algoritmaları uygulanmasında % 95 doğruluk ve 0 G-Ortalama değerleri ile alınmıştır.

Çizelge 4.1. Karşılaştırmalı sınıflandırma sonuçları

Veri Seti	Sınıflandırıcı	Önerilen-Radial SVM		Dengesiz Veri		SMOTE		Önerilen-Linear SVM	
		Doğruluk	G-Ortalama	Doğruluk	G-Ortalama	Doğruluk	G-Ortalama	Doğruluk	G-Ortalama
Climate	KNN	% 82	0,502	% 92	0,333	% 76	**0,656**	% 81	0,535
	RF	% 91	0,332	% 90	0	% 93	**0,471**	% 91	0
	SVM	% 95	**0,808**	% 91	0	% 91	0,726	% 94	0,738
	YSA	% 95	0,742	% 95	0,667	% 96	**0,873**	% 94	0,738
Diabetes	KNN	% 70	**0,694**	% 73	0,611	% 70	0,679	% 67	0,653
	RF	% 72	**0,703**	% 71	0,654	% 71	0,699	% 70	0,664
	SVM	% 74	0,710	% 75	0,673	% 71	**0,713**	% 71	0,707
	YSA	% 75	**0,727**	% 74	0,464	% 69	0,699	% 72	0,684
Liver	KNN	% 60	0,574	% 69	0,509	% 65	**0,605**	% 68	0,576
	RF	% 75	**0,649**	% 72	0,174	% 73	0,602	% 75	0,628
	SVM	% 64	0,682	% 71	0,463	% 66	0,696	% 71	**0,700**
	YSA	% 69	**0,736**	% 64	0,573	% 67	0,703	% 71	0,678
Haberman	KNN	% 63	0,555	% 73	0,477	% 63	**0,631**	% 66	0,538
	RF	% 65	0,530	% 75	**0,577**	% 66	0,538	% 71	0,562
	SVM	% 76	0,584	% 73	0,000	% 78	**0,631**	% 73	0,477
	YSA	% 73	0,648	% 75	0,533	% 68	0,703	%75	**0,705**
Transfusion	KNN	% 68	**0,616**	% 79	0,567	% 65	0,597	% 67	0,580
	RF	% 75	0,640	% 80	0,548	% 73	0,658	% 72	**0,677**
	SVM	% 79	0,718	% 77	0,169	% 67	**0,718**	% 73	0,704
	YSA	% 79	0,728	% 78	0,375	% 73	0,668	% 77	**0,732**
Ionosphore	KNN	% 92	**0,894**	% 87	0,800	% 92	**0,894**	% 92	**0,894**
	RF	% 97	**0,959**	% 95	0,948	% 95	0,948	% 94	0,947
	SVM	% 95	**0,938**	% 95	**0,938**	% 95	**0,938**	% 90	0,904
	YSA	% 87	0,815	% 88	0,825	% 91	**0,872**	% 87	0,815
Column_2c	KNN	% 79	0,764	% 80	**0,805**	% 77	0,736	% 72	0,719
	RF	% 82	0,830	% 83	**0,842**	% 80	0,805	% 82	0,830
	SVM	% 72	0,743	% 77	0,736	% 70	0,746	% 77	**0,792**
	YSA	% 75	0,756	% 77	0,717	% 74	0,744	% 77	**0,792**
TexYarn	KNN	% 96	0,497	% 95	0	% 96	0,497	% 96	**0,609**
	RF	% 96	0,782	% 96	0,784	% 95	0,605	% 97	**0,856**
	SVM	% 96	0,920	% 95	0	% 82	0,901	% 94	**0,970**
	YSA	% 96	0,779	% 96	0	% 86	0,924	% 98	**0,991**

Elde edilen sonuçların Piri ve ark. (2018) SIMO algoritmasında elde ettiği ve makalede yer alan sonuç değerleri ile karşılaştırılması Çizelge 4.2'de özetlenmiştir. Eğitim ve test verisi makale ile aynı şekilde ayrılmamıştır.

- Diabetes veri seti için SIMO algoritmasında % 75,48 G-Ortalama değeri elde edilmiştir. Melez algoritmada ise bu değer % 72,7 olmuştur.
- Liver veri setin için SIMO algoritmasında % 68,62 G- Ortalama değeri elde edilmiştir. Melez algoritmada ise bu değer % 73,6 olmuştur.
- Ionosphore veri seti için SIMO algoritmasında % 84,69 G-Ortalama değeri elde edilmiştir. Melez algoritmada ise bu değer % 95,9 olmuştur.

Çizelge 4.2. SIMO ve önerilen melez yöntem karşılaştırması

Yöntem\ Veri Seti	Diabetes	Liver	Ionosphore
SIMO	**% 75,48**	% 68,62 %	% 84,69
Önerilen Melez Yöntem	% 72,7	**% 73,6**	**% 95,9**

İki çalışmada da ortak kullanılan 3 veri seti incelendiğinde 2 veri setinde melez algoritmanın daha başarılı olduğu gözlemlenmiştir.

Sarmonova'nın (Sarmanova, 2013) önerdiği RusAda algoritmasının tez çalışmasında yer alan sonuç değerleri ile önerilen melez yöntemin karşılaştırma sonuçları Çizelge 4.3'te gösterilmiştir. Eğitim ve test verisi tez ile aynı şekilde ayrılmamıştır.

- Diabetes veri seti için RusAda algoritmasında % 76,15 G-Ortalama değeri elde edilmiştir. Melez algoritmada ise bu değer % 72,7 olmuştur.
- Haberman veri seti için RusAda algoritmasında % 64,79 G-Ortalama değeri elde edilmiştir. Melez algoritmada ise bu değer % 70,5 olmuştur.
- Transfusion veri seti için RusAda algoritmasında % 69,21 G-Ortalama değeri elde edilmiştir. Melez algoritmada ise bu değer % 73,2 olmuştur.
- Ionosphore veri seti için RusAda algoritmasında % 90,14 G-Ortalama değeri elde edilmiştir. Melez algoritmada ise bu değer % 95,9 olmuştur.

Çizelge 4.3. RusAda ve önerilen melez yöntem karşılaştırması

Yöntem\ Veri Seti	Diabetes	Haberman	Transfusion	Ionosphore
RusAda	**% 76,15**	% 64,79	% 69,21	% 90,14
Önerilen Melez Yöntem	% 72,7	**% 70,5**	**% 73,2**	**% 95,9**

İki çalışmada da ortak kullanılan 4 veri seti incelendiğinde 3 veri setinde melez algoritmanın daha başarılı olduğu gözlemlenmiştir.

5. SONUÇ

Gerçekleştirilen çalışma kapsamında öncelikli olarak dengesiz veri setleri üzerine yapılan çalışmalar incelenmiştir. Literatürde dengesiz veriyi hedef alan yöntemlerin çıkış noktaları ve geçmişte yapılan yöntemler ile kıyaslanmaları incelenerek yapılacak çalışmanın mevcut çalışmaların avantaj ve dezavantajlarından yararlanması amaçlanmıştır. Sonuçların değerlendirmesinde literatür ile de karşılaştırılabilmesi için bu çalışmalarda sıklıkla kullanılan G-Ortalama ölçütü değerlendirme ölçütü olarak seçilmiştir.

Veri setlerinin seçiminde daha önce yapılan çalışmalar ile kıyaslanabilmesi için sıkça kullanılan veri setleri çalışma kapsamına dahil edilmiştir. Gerçek hayatta uygulanabilirliğinin kanıtlanması için de gerçek hayat verisi ile de çalışma yapılmıştır. Seçilen veriler melez yöntemin uygulanmasının ardından sonuçlar, aynı verilelerin kullanıldığı diğer algoritmalar ile karşılaştırılması yapılarak başarılı sonuçlar aldığı gözlemlenmiştir.

Gerçek hayat verisi olan TexYarn verisinde yapılan çalışmada klasik yöntemlerden daha başarılı sonuçlar vererek, dokuma firmasında kopma riski bulunan ipliklerin dokuma işlemine girmeden tespit edilmesinin mümkün olduğunu kanıtlamıştır. Bu yönüyle karar destek sistemi olarak kullanılmasının önü açılmıştır.

SIMO yöntemi bilgi verici azınlık verilerine odaklanırken bu çalışma bilgi verici olmayan çoğunluk verileri ile de ilgilenmiştir. Aşırı örneklemeden kaynaklanan aşırı uyumu ortadan kaldırmak için sadece bilgi verici azınlık verilerini çoğaltırken, bilgi kaybının önüne geçebilmek için değerli bilgi taşımayan çoğunluk verileri veri setinden çıkarılmıştır. Örnek azaltmada SVM kullanılarak, karar sınırına uzak, diğer bir deyişle bilgi vericiliği az olan çoğunluk verilerini çıkartılması ile, literature örnek azaltma işlemlerinde yeni bir yaklaşım getirmiştir.

Yapılan çalışma ile örnek azaltma ya da aşırı örnekleme yöntemlerinin tekil kullanımı yerine birlikte kullanılarak bu yöntemlerin dezavantajlarının minimize edilmesine olanak sağlayacağını göstermiştir. Yine dengesiz veriyi dengeli hale getirirken verilerin toplu halde işleme tutulması yerine, verilerin bilgi verici olup olmaması ya da verilerin sınıf

etiketinin tahmininde değerli bir etki yaratıp yaratmamasına göre veri bazında incelenelerek veri ön işlemesinin yapılmasının sınıflandırma sonucuna olumlu etki yarattığı göstermiştir. Gelecek çalışmalarda klasik yöntemlerden farklı olarak bu noktalara odaklanılması, dengesiz veri setlerinin sınıflandırma probleminin çözümünde olumlu etki yaratacağını göstermektedir.

Önerilen yöntemin literatürde incelenen yöntemlerle benzer olarak iki sınıflı dengesiz veriler üzerinde çalışması, çok sınıflı verilerde çalışmamış olması sınırlı bir kullanım alanı sunmaktadır. Çalışma farklı sınıflandırma algoritmalarını da içine alacak şekilde genişletilebilir. Aynı zamanda parameter optimizasyonu üzerine çalışılmasının sınıflandırma başarısına etkisi incelenebilir. Gelecek çalışmaların parameter optimizasyonu ve çok sınıflı veri setleri üzerine odaklanması dengesiz verilerin ele alınmasında gelişme sağlayacaktır.

KAYNAKLAR

Akin, P., & Terzi, Y. 2021. Comparison of Unbalanced Data Methods for Support Vector Machines. *Turkiye Klinikleri Journal of Biostatistics*, *13*(2).

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. 2012. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 272.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... & Hussain, A. 2016. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, 4, 7940-7957.

Anonim, 2020. Previous Releases of R for Windows
https://cran.r-project.org/bin/windows/base/old/4.0.3/ - (Erişim Tarihi :02.02.2021)

Aridas, C. K., Karlos, S., Kanas, V. G., Fazakis, N., & Kotsiantis, S. B. 2019. Uncertainty based under-sampling for learning naive Bayes classifiers under imbalanced data sets. *IEEE Access*, 8, 2122-2133.

Balaban, M. E., & Kartal, E. 2015. Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları. *Çağlayan Kitabevi, İstanbul*.

Batista, G. E., Prati, R. C., & Monard, M. C. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1), 20-29.

Beckmann, M., Ebecken, N. F., & de Lima, B. S. P. 2015. A KNN undersampling approach for data balancing. *Journal of Intelligent Learning Systems and Applications*, 7(04), 104.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1), 5-32.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. 2009. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In Pacific-Asia conference on knowledge discovery and data mining (pp. 475-482). Springer, Berlin, Heidelberg.

Cao, L., & Zhai, Y. 2015. Imbalanced data classification based on a hybrid resampling SVM method. In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, 1533-1536. IEEE.

Cao, P., Zhao, D., & Zaiane, O. 2013. An optimized cost-sensitive SVM for imbalanced data learning. In *Pacific-Asia conference on knowledge discovery and data mining* ,280-292. Springer, Berlin, Heidelberg.

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chen, C., Liaw, A., & Breiman, L. 2004. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110(1-12), 24.

Cohen, G., Hilario, M., Sax, H., Hugonnet, S., & Geissbuhler, A. 2006. Learning from imbalanced data in surveillance of nosocomial infection. *Artificial intelligence in medicine*, 37(1), 7-18.

Dhar, S., & Cherkassky, V. 2014. Development and evaluation of cost-sensitive universum-SVM. *IEEE transactions on cybernetics*, *45*(4), 806-818.

Fernández, A., del Jesus, M. J., & Herrera, F. 2009. Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *International Journal of Approximate Reasoning*, 50(3), 561-577.

Fernández-Navarro, F., Hervás-Martínez, C., & Gutiérrez, P. A. 2011. A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognition*, 44(8), 1821-1833.

Galar, M., Fernández, A., Barrenechea, E., & Herrera, F. 2013. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern recognition*, 46(12), 3460-3471.

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

Gong, J., & Kim, H. 2017. RHSBoost: Improving classification performance in imbalance data. *Computational Statistics & Data Analysis*, 111, 1-13.

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

He, H., & Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284.

He, H., Bai, Y., Garcia, E. A., & Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* ,1322-1328. IEEE.

Ho, T. K. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* Vol. 1, 278-282. IEEE.

Jamali, I., Bazmara, M., & Jafari, S. 2012. Feature Selection in Imbalance data sets. *International Journal of Computer Science Issues (IJCSI)*, 9(3), 42.

Kaur, H., Pannu, H. S., & Malhi, A. K. 2019. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, 52(4), 1-36.

Krawczyk, B., Koziarski, M., & Woźniak, M. 2019. Radial-based oversampling for multiclass imbalanced data classification. *IEEE transactions on neural networks and learning systems*, 31(8), 2818-2831.

Laurikkala, J. 2001. Improving identification of difficult small classes by balancing class distribution. In *Conference on Artificial Intelligence in Medicine in Europe* ,63-66. Springer, Berlin, Heidelberg.

Maglogiannis, I. G. (Ed.). 2007. *Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies* (Vol. 160). Ios Press.

Maltarollo, V. G., Honório, K. M., & da Silva, A. B. F. 2013. Applications of artificial neural networks in chemical problems. *Artificial neural networks-architectures and applications*, 203-223.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2), 442-451.

McCullock, W. S., & Pitts, W. 1956. A Logical Calculus of Ideas Immanent in Nervous Activity. Archive copy of 27 November 2007 on Wayback Machine. *Avtomaty [Automated Devices] Moscow, Inostr. Lit. publ*, 363-384.

Mirzaei, B., Nikpour, B., & Nezamabadi-Pour, H. 2020. An under-sampling technique for imbalanced data classification based on DBSCAN algorithm. In *2020 8th Iranian Joint Congress on Fuzzy and intelligent Systems (CFIS)*,21-26. IEEE.

Piri, S., Delen, D., & Liu, T. 2018. A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems*, 106, 15-29.

Pragati Baheti, 2021. 12 Types of Neural Network Activation Functions: How to Choose? https://www.v7labs.com/blog/neural-networks-activation-functions - (Erişim Tarihi :12.12.2021)

Qiu, C., Jiang, L., & Li, C. 2017. Randomly selected decision tree for test-cost sensitive learning. *Applied Soft Computing*, 53, 27-33.

Ramentol, E., Gondres, I., Lajes, S., Bello, R., Caballero, Y., Cornelis, C., & Herrera, F. 2016. Fuzzy-rough imbalanced learning for the diagnosis of High Voltage Circuit Breaker maintenance: The SMOTE-FRST-2T algorithm. *Engineering Applications of Artificial Intelligence*, 48, 134-139.

Rao, K. N., Rao, T. V., & Lakshmi, D. R. 2012. A Novel Class Imbalance Learning using Ordering Points Clustering. *International Journal of Computer Applications*, 51(16).

Sáez, J. A., Krawczyk, B., & Woźniak, M. 2016. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 57, 164-178.

Sarmanova, A. 2013. Veri madenciliğindeki sınıf dengesizliği sorununun giderilmesi.

Sowah, R. A., Agebure, M. A., Mills, G. A., Koumadi, K. M., & Fiawoo, S. Y. 2016. New cluster undersampling technique for class imbalance learning. *International Journal of Machine Learning and Computing*, 6(3), 205-214.

Spelmen, V. S., & Porkodi, R. 2018. A review on handling imbalanced data. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*,1-11. IEEE.

Tang, Y., Zhang, Y. Q., Chawla, N. V., & Krasser, S. 2008. SVMs modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1), 281-288.

Yin, L., Ge, Y., Xiao, K., Wang, X., & Quan, X. 2013. Feature selection for high-dimensional imbalanced data. *Neurocomputing*, 105, 3-11

Yong, Y. 2012. The research of imbalanced data set of sample sampling method based on K-means cluster and genetic algorithm. *Energy Procedia*, 17, 164-170.

Yu, H., Ni, J., & Zhao, J. 2013. ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing*, 101, 309-318.

Zhang, H., & Li, M. 2014. RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, 20, 99-116.

Zhang, Y. P., Zhang, L. N., & Wang, Y. C. 2010. Cluster-based majority under-sampling approaches for class imbalance learning. In *2010 2nd IEEE International Conference on Information and Financial Engineering* ,400-404. IEEE.

Zheng, Z., Cai, Y., & Li, Y. 2015. Oversampling method for imbalanced classification. *Computing and Informatics*, 34

ÖZGEÇMİŞ

Adı Soyadı	: Mestan Şahin Pir
Doğum Yeri ve Tarihi	: Tire/İZMİR – 27.04.1992
Yabancı Dil	: İngilizce
Eğitim Durumu	
Lise	: Ortaklar Anadolu Öğretmen Lisesi
Lisans	: Balıkesir Üniversitesi – Endüstri Mühendisliği
Yüksek Lisans	: Bursa Uludağ Üniversitesi – Endüstri Mühendisliği
Çalıştığı Kurum/Kurumlar	: Beyçelik Gestamp – Üretim Planlama Mühendisi Ak Pres Otomotiv – Lojistij Uzmanı Nobel Otomotiv – Malzeme Planlama Sorumlusu
İletişim (e-posta)	: mestansahin@hotmail.com