T.C
OSTİM TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YAZILIM MÜHENDİSLİĞİ ANA BİLİM DALI
YAZILIM MÜHENDİSLİĞİ
YÜKSEK LİSANS PROGRAMI

ÇALIŞAN KAYBININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİNLEMESİ

YÜKSEK LİSANS

HAZIRLAYAN
ÇETİN KAYA

TEZ DANIŞMANI
Dr. Öğr. Üyesi MURAT ŞİMŞEK

ANKARA-2024

TEZ KABUL VE ONAY

……. tarafından hazırlanan "……." başlıklı bu çalışma, ../../20.. tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Yüksek Lisans/Doktora Tezi olarak kabul edilmiştir.

Kabul Tarihi:...../...../.....

Jüri Üyesi:	**Unvanı Adı SOYADI** … Üniversitesi	_____
Jüri Üyesi:	**Unvanı Adı SOYADI** … Üniversitesi	_____
Jüri Üyesi:	**Unvanı Adı SOYADI** … Üniversitesi	_____
Jüri Üyesi:	**Unvanı Adı SOYADI** … Üniversitesi	_____
Tez Danışmanı:	**Unvanı Adı SOYADI** Ostim Teknik Üniversitesi	_____
İkinci Danışman:	**Unvanı Adı SOYADI** … Üniversitesi	_____

ONAY

Jüri tarafından kabul edilen bu çalışmanın Yüksek Lisans/Doktora Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

…../…../20….

_____

Unvanı Adı SOYADI
Enstitü Müdürü

BİLDİRİM

Enstitü tarafından onaylanan Yüksek Lisans/Doktora tezimin tamamını veya herhangi bir kısmını basılı veya dijital biçimde arşivleme ve aşağıda belirtilen koşullar dâhilinde erişime açma iznini Ostim Teknik Üniversitesine verdiğimi bildiririm. Bu izinle, Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak ve gelecekteki çalışmalar (makale, kitap, lisans, patent vb.) için tezimin tamamının veya bir bölümünün kullanım hakları yalnızca bana ait olacaktır.

Tezimin bütünüyle kendi çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Telif hakkı bulunan ve sahiplerinden yazılı izinle kullanılması zorunlu olan kaynakları, yazılı izin alarak kullandığımı ve istenildiğinde izinlerin suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayımlanan "Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge" kapsamında, tezim, aşağıda belirtilen koşullar haricince, YÖK Ulusal Tez Merkezi ve Ostim Teknik Üniversitesi Açık Erişim Sisteminde erişime açılır.

☐ Enstitü / Fakülte Yönetim Kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.[1]

☐ Enstitü / Fakülte Yönetim Kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren … ay ertelenmiştir.[2]

☐ Tezimle ilgili gizlilik kararı verilmiştir.[3][4]

Tarih

İmza

[1] MADDE 6(1) Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.

[2] MADDE 6(2) Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.

[3] MADDE 7(1) Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.

[4] MADDE 7(2) Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.

OSTİM TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ YÜKSEK LİSANS / DOKTORA TEZİ ORİJİNALLİK RAPORU

Tezin Başlığı:……..

Öğrencinin Adı, Soyadı:……..……..

Tez Danışmanın Unvanı/Adı, Soyadı:……

Anabilim Dalı:…….……..…….

Programı:……..…….

Tarih: ….. / …. / ….

Yukarıda başlığı belirtilen Yüksek Lisans/Doktora tez çalışmama ait Giriş, Ana Bölümler ve Sonuç kısmından oluşan ve toplam..... sayfadan ibaret olan kısmı, …. / …. / ….. tarihinde tez danışmanım/şahsım tarafından …….…… intihal tespit programında incelenmiştir. Orijinallik raporunda aşağıda ifade edilen filtrelemeler uygulanmıştır.

Orijinallik raporuna göre, tezimin benzerlik oranı % 15'dır.

Uygulanan filtrelemeler:

1. Kaynakça (hariç)

2. Alıntılar (hariç)

3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları (hariç)

Tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrencinin İmzası:…….

ONAY

Tarih: …. / …. /…..

Danışmanın Unvanı, Adı, Soyadı, İmzası

ETİK BEYAN

Bu çalışmanın özgün bir çalışma olduğunu, çalışmanın hazırlık, veri toplama, analiz bilgilerin sunumu ve diğer tüm aşamalarında bilimsel etik ve kurallara uygun davrandığımı çalışmada bulunan tüm belge bilgileri akademik etik ve kurallar çerçevesinde elde ettiğimi görsel, işitsel ve yazılı bütün bilgileri ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullanmış olduğum verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterdiğim durumlar dışında tarafımdan kaleme alındığını ve özgün olduğunu, Tez danışmanım Dr. Öğr. Üyesi Murat ŞİMŞEK danışmanlığında ve tarafımdan üretildiğini ve OSTİM Teknik Üniversitesi Tez Yazım Kılavuzuna uygun olarak yazıldığını beyan ederim.

Öğrencinin İmzası

Öğrencinin Adı Soyadı

Tarih

TEŞEKKÜR

Başta, bu süreçte maddi ve manevi desteğini benden esirgemeyen canım aileme sonsuz teşekkür ve minnettarlığımı bildirmekteyim. Aynı zamanda tez sürecinde yardımlarını benden esirgemeyen değerli hocam, Dr. Öğr. Üyesi Murat ŞİMŞEK başta olmakla, destek veren tüm hocalara sonsuz teşekkür ve minnettarlığımı bildirmekteyim.

Tarih

Çetin KAYA

ÖZ

Yazar Adı ve Soyadı : Çetin Kaya

Üniversite : OSTİM Teknik Üniversitesi

Enstitü : Fen Bilimleri Enstitüsü

Program Adı : Bilgisayar Mühendisliği

Tezin Türü : Yüksek Lisans

Sayfa Sayısı : 100

Tarihi : 2024

ÇALIŞAN KAYBININ MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE TAHMİNLEMESİ

Bu tez çalışması, çalışan devrinin kuruluşlar üzerindeki etkilerini anlamak ve bu devri doğru bir şekilde tahmin edebilmek amacıyla gerçekleştirilmiştir. Çalışan devri, bir kuruluşun üretkenliği, kültürü ve karlılığı üzerinde önemli bir etki sahibi olup, bu devrin proaktif bir şekilde yönetilmesi kuruluşların uzun vadeli başarısında kritik bir rol oynar. Bu çalışmada, çalışan devir hızını tahmin etmek için geleneksel makine öğrenimi yöntemleri kullanılarak çeşitli analizler yapılmıştır.

Analiz sürecinde, kullanılan veri setinin dengesiz dağılımı tespit edilmiş ve bu sorunu çözmek için aşağı örnekleme ve yukarı örnekleme yöntemleri uygulanmıştır. Bu yöntemler Rastgele Yukarı Örnekleme (Random Over Sampler: ROS), Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-sampling Technique: SMOTE), Rastgele Az Örnekleme Tekniği(Random Under-Sampling Technique: RUS), Near Miss ve Tomek Links yöntemleridir. Dengelenen veri seti üzerinde, aşırı öğrenmeyi önlemek amacıyla K-Fold Çapraz Doğrulama, Stratified K-Fold Çapraz Doğrulama, Leave-One-Out (LOO) Çapraz Doğrulama ve Group K-Fold Çapraz Doğrulama yöntemleri uygulanmıştır. Makine öğrenmesi modeli olarak yüksek performans metriklerine sahip Rastgele Orman (Random Forest: RF) yöntemi seçilmiş ve modelin performansını artırmak için Manuel(Elle) Ayarlama, Izgara Araması (Grid Search CV), Rastgele Arama, Bayes Optimizasyonu,

Genetik Algoritmalar ve Parçacık Sürüsü Optimizasyonu hiper-parametre ayarlama yöntemleri uygulanmıştır.

Çalışma sonucunda, Rastgele Orman(Random Forest: RF) ve Rastgele Yukarı Örnekleme(Random Over Sampler: ROS) yöntemlerinin birleşik kullanımının, çalışan devri tahmininde yüksek doğruluk sağladığı görülmüştür. Bu bulgu, kuruluşların çalışan devrini daha etkili bir şekilde yönetmelerine ve olası personel kayıplarını önceden tahmin ederek önleyici stratejiler geliştirmelerine olanak tanır. Ayrıca, çalışma sonuçları kuruluşların insan kaynakları yönetimi uygulamalarını geliştirmeleri için değerli içgörüler sunmaktadır.

Bu tez, çalışan devri tahmininde makine öğrenimi yöntemlerinin etkinliğini ortaya koyarak, kuruluşların bu tahminleri stratejik karar alma süreçlerine entegre etmeleri için bir yol haritası sunmaktadır. Gelecek çalışmalar, farklı sektörlerdeki kuruluşlar üzerinde benzer yöntemlerin uygulanabilirliğini ve etkinliğini test ederek, bu alandaki bilgi birikimini genişletebilir.

Anahtar Sözcükler: Çalışan Devri, Makine Öğrenimi, Rastgele Orman, Veri Dengeleme, Çapraz Doğrulama

ABSTRACT

Thesis : Çetin Kaya

University : OSTİM Technical University

Institute : Graduate School of Natural and Applied Sciences

Program's Name : Computer Engineering

Thesis Type: : Master

Pages : 100

Year : 2024

PREDICTING EMPLOYEE ATTRITION WITH MACHINE LEARNING METHODS

This thesis aims to understand the impact of employee turnover on organizations and to accurately predict it. Employee turnover has a significant impact on an organization's productivity, culture and profitability, and proactively managing this turnover plays a critical role in the long-term success of organizations. In this study, several analyses were conducted using traditional machine learning methods to predict employee turnover.

During the analysis process, the unbalanced distribution of the data set used was identified and down sampling and up sampling methods were applied to solve this problem. These methods are Random Over Sampler (ROS), Synthetic Minority Over-sampling Technique (SMOTE), Random Under-Sampling Technique (RUS), Near Miss and Tomek Links. K-Fold Cross-Validation, Stratified K-Fold Cross-Validation, Leave-One-Out (LOO) Cross-Validation and Group K-Fold Cross-Validation methods were applied on the balanced dataset to prevent over-learning. Random Forest (RF) method with high performance metrics was selected as the machine learning model and Manual Tuning, Grid Search CV, Random Search, Bayesian Optimization, Genetic Algorithms and Particle Swarm Optimization hyper-parameter tuning methods were applied to improve the performance of the model.

The results of the study show that the combined use of Random Forest (RF) and Random Over Sampler (ROS) methods provides high accuracy in employee turnover prediction. This finding allows organizations to manage employee turnover more effectively and develop

preventive strategies by anticipating potential staff losses. Furthermore, the study results provide valuable insights for organizations to improve their human resource management practices.

By demonstrating the effectiveness of machine learning methods in employee turnover prediction, this thesis provides a roadmap for organizations to integrate these predictions into their strategic decision-making processes. Future studies can expand the body of knowledge in this area by testing the applicability and effectiveness of similar methods on organizations in different industries.

Keywords: Employee Turnover, Machine Learning, Random Forest, Data Balancing, Cross-Validation

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

SİMGELER VE KISALTMALAR

ROS	Rastgele Yukarı Örnekleme
RUS	Rastgele Aşağı Örnekleme
SMOTE	Sentetik Azınlık Aşırı Örnekleme Tekniği
KNN	K-Nearest Neighbors(K-En Yakın Komşular)
SVM	Support Vector Machine(Destek Vektör Makineleri)
NM	Near Miss
RF	Random Forest(Rastgele Orman)
NB	Naive Bayes
XGB	XGBoost
DT	Decision Three(Karar Ağaçları)
GA	Genetik Arama
GS	Gauss Süreci
PSO	Parçacık Sürü Optimizasyonu
İK	İnsan Kaynakları
ADASYN	Adaptif Sentetik Örnekleme Yöntemi
RFE	Özyinelemeli Öznitelik Eleme
PSO	Parçacık Sürü Optimizasyonu
IQR	Interquartile Range(Çeyrekler Arası Aralık)

1. GİRİŞ

Çalışan devir hızı "çalışan kaybı" şirketler, kurumlar için maliyetli bir problemdir. Bir kuruluşun yaşam döngüsü boyunca karşılaşılabilen en önemli sorunlardan biri olarak görülen çalışan devri, tahmin edilmesi güç olup, sıklıkla kuruluşun vasıflı iş gücünde belirgin boşlukların oluşmasına neden olmaktadır. Özellikle hizmet firmalarında, çalışanların beklenmedik şekilde ayrılması, hizmetlerin zamanında sunulmasını tehlikeye atabilir, genel firma verimliliğini önemli ölçüde düşürebilir ve sonuç olarak müşteri sadakatinin azalmasına yol açabilir. Bu nedenle, kuruluşların etkili işe alma, edinme ve elde tutma stratejileri geliştirmesi, çalışan devrini önlemek ve azaltmak için etkili mekanizmaları uygulaması ve bu durumun altında yatan temel nedenleri anlaması gerekmekte olduğu vurgulanmaktadır.

Genel olarak, çalışan devri istemsiz devir ve gönüllü devir olmak üzere ikiye ayrılabilir. İstem dışı işten ayrılma, genellikle çalışanın sadece kendi isteğiyle işten ayrıldığı bir kuruluştaki üyelik sınırının ötesindeki hareketler olarak tanımlanır. Hafif duygulanımlar sergiler. Diğer taraftan, gönüllü ciro, bir kişinin bir kuruluş arasındaki üyelik sınırının ötesindeki hareketler, üzerinde çalışan ağır duygular besler. Günümüzün rekabetçi ekonomisinde ve artan teknolojik uzmanlaşmasında, verilerin edinilmesi, incelenmesi ve analiz edilmesi, "bilgi ekonomisi" olarak adlandırılan yeni bilgilerin ortaya çıkmasına neden olmaktadır. Bilgi teknolojileri yalnızca bir veri kaynağı değildir, her şeyden önce veri analizini mümkün kılan, büyük veri koleksiyonlarının işlenmesini ve bunlardan bilgi çıkarılmasını mümkün kılan bir faktördür. Veriler, iş süreçleriyle bağlantılı olanlar da dâhil olmak üzere birden fazla sektördeki çoğu kuruluş için stratejik bir varlık haline geldi. Her türden kuruluş yeni teknolojilerin benimsenmesinden yararlanır [1].

Küresel rekabet senaryosu göz önüne alındığında, dünyada vasıflı ve yetenekli insanlar için fırsatlar okyanusu vardır ve iyi bir şans verildiğinde çalışanlar bir kuruluştan diğerine ayrılır. İşyeri verimliliği ve kurumsal hedeflerin zamanında gerçekleştirilmesi üzerindeki olumsuz etkileri nedeniyle, bugünlerde çalışan değişimi tüm kuruluşlar için temel sorun olarak kabul edilmektedir. Bu sorunun üstesinden gelmek için kuruluşlar artık çalışan devir hızını tahmin etmek amacıyla makine öğrenimi tekniklerinden destek alıyor. Tahminlerdeki yüksek hassasiyet sayesinde kuruluşlar, çalışanların şirkette kalması için gerekli önlemleri zamanında alabilir. Verilerin çoğu, tahmin ve modelleme açısından yüksek verimliliğe sahip

olmayan temel İK tabanlı veritabanı sistemlerinden gelmekte ve bu modeller, veri modellerinde çok doğru sonuç vermemekte ve kuruluşların başarılı kararlar almasına yardımcı olamamaktadır [2].

Makine öğrenimi, sistemlerin açık bir programa ihtiyaç duymadan deneyimlerden öğrenmesini ve hareket etmesini sağlayan yapay zekânın kullanıldığı bir uygulama alanıdır. Gelecekteki olaylar, mevcut geçmiş veriler üzerinde makine öğrenmesi algoritmaları çalıştırılarak tahmin edilebilir. Makine öğrenimi, etiketli veri sınıflandırmaları üretir ve aynı zamanda etiketlenmemiş verilerden gizli yapılar oluşturmak için de kullanılabilir. Kuruluşların üst düzey yönetimi, bir çalışanın kuruluştan ayrılma olasılığını önceden tahmin etmek için makine öğrenimi algoritmalarının bu tahmin potansiyelinden yararlanabilir. Bu süreç, yıpranmaya yol açan faktörlerin kontrol altına alınmasına ve bunun gerçekleşmesinin önlenmesine yardımcı olacaktır. Çalışan devri, işverenin karşılaştığı ciddi bir zorluktur. Yetenekleri elde tutmak her organizasyon için çok önemlidir. Dolayısıyla yönetim, çalışanların ayrılma olasılığının yanı sıra ayrılmayı etkileyen faktörleri de tahmin edebilirse, bu, işten ayrılma riskini azaltacak kararların alınmasında etkili olabilir. Makine öğreniminin oynayacağı rol burasıdır. Makine öğrenimi algoritmaları tarafından yapılan tahminler, yönetimin çalışanları elde tutmak için atması gereken proaktif adımları gösterecektir [3].

Bu tez çalışmasında, geleneksel makine öğrenmesi yöntemlerinden, Doğrusal Regresyon, Knn, Svn, Rastgele Orman, Naïve Bayes yöntemleri ile birlikte hibrit yöntemler olan Xgboost ve Adaboost kullanılmıştır. Kullanılan bu yöntemlerin sonucunda dengesiz veri tespit edilmiştir.

Dengesiz veri setlerinin ele alınmasında, Rastgele Aşırı Örnekleme (ROS) ve Rastgele Eksik Örnekleme (RUS) gibi çeşitli örnekleme stratejilerinin yanı sıra Near Miss ve Tomek Links gibi tekniklerin kullanılması, sınıf dengesizliklerinin üstesinden gelinmesinde önemli bir rol oynanmaktadır. Bu yöntemler, azınlık sınıfındaki örneklerin sayısının artırılması veya çoğunluk sınıfındaki örneklerin azaltılması yoluyla veri setlerindeki dengesizliğin azaltılmasına yardımcı olur. Özellikle, Near Miss yöntemi, çoğunluk sınıfından örneklerin seçiminde azınlık sınıfı örneklerine olan "yakınlığı"na göre bir filtreleme yaparken; Tomek Links, çoğunluk ve azınlık sınıfları arasındaki sınırı daha net hale getirerek yanlış sınıflandırılmış örnekleri temizlemek için kullanılır. Bu tekniklerin uygulanması, makine öğrenmesi modellerinin müşteri kaybı tahminlerindeki performansının iyileştirilmesine katkıda bulunurken, aynı zamanda modellerin daha dengeli ve doğru tahminler

yapabilmesine olanak tanımaktadır. Bu şekilde, az temsil edilen sınıflar üzerinde daha iyi bir genelleme kabiliyetine ulaşılması amaçlanmaktadır [4].

Gerçek verilerde, birçok durum vardır ki bir sınıftaki örnek sayısı, bir sınıftaki örnek sayısından çok daha azdır. Diğer sınıflardaki örnekler, dengesiz veri kümesi olarak adlandırılır. Bunun için en popüler çözümlerden biri problemi örneklemedir [5]. İncelenen eserler üzerinde yapılan incelemeye dayalı olarak bazı ilginç eğilimler/sonuçlar gözlemlendi ve bazı önemli bulgular aşağıda özetlenmiştir.

Veri Düzeyi yöntemleri arasında, ilgili çalışmaların deneye dayalı sonuçları genellikle Rastgele Aşırı Örneklemenin (ROS), Rastgele Düşük Örnekleme veya Sentetik Azınlık Aşırı Örnekleme Tekniğinden (SMOTE) daha iyi sınıflandırma performansı verdiğini göstermektedir [6].

Makine öğrenimi algoritmalarının kullanımıyla girişimcilik yetkinliğinin tahmin edilmesine yönelik çeşitli çalışmalar mevcuttur. Çoğu makine öğrenmesi yöntemi dengesiz verilerde daha iyi doğruluk ve F1-skor performansı sergilemektedir. Bu çalışma, tahmin performansını artırmak için dengesiz sınıf işleme yöntemlerinden yararlanmaya odaklanmaktadır. Bu amaçla bu çalışmada dengesiz verileri işlemek için Rastgele Aşırı Örnekleme, Rastgele Az Örnekleme, SMOTE ve NearMiss yöntemleri kullanılmıştır. Dengesiz Veri İşleme yöntemleri ile makine öğrenmesi algoritmalarının performansı makine öğrenmesi algoritmaları ile karşılaştırılmıştır. Karşılaştırmalar performans parametreleri olarak doğruluk, kesinlik, geri çağırma, F1-Skoru kullanılarak yapılmıştır. Karşılaştırma, dengesiz veri kümesi işleme yöntemleri ile makine öğrenmesi algoritmaları kullanılarak gözle görülür bir performans artışı olduğunu göstermektedir [7].

Bu tez, makine öğrenimi algoritmalarının çalışan devir hızını tahmin etme kapasitesini nasıl geliştirebileceğini araştırmayı hedeflemektedir. Araştırmanın odak noktası, dengesiz veri setleriyle başa çıkabilme yöntemlerini geliştirme, çalışan devir hızının tahmininde en verimli makine öğrenimi algoritmalarını saptama ve bu algoritmaların kullanımının etik boyutlarını inceleme olacaktır.

2. ARAŞTIRMA PROBLEMİNİN TANIMI

Çalışanların işten ayrılma eğilimlerinin ve bu durumun organizasyonlar üzerindeki etkilerinin önceden belirlenmesi, günümüz iş dünyasında karşılaşılan kritik sorunlardan biridir. Bu bağlamda, iş gücü devamlılığını sağlamak ve ayrılma oranlarını minimize etmek amacıyla makine öğrenmesi tekniklerinin kullanılması gereksinimi ön plana çıkmaktadır. Bu araştırma, çeşitli makine öğrenmesi modelleri aracılığıyla, çalışanların işten ayrılma olasılığını etkileyen değişkenleri analiz etmekte ve hangi çalışanların ayrılma eğiliminde olduğunu belirleme kapasitesini incelemektedir. İnsan kaynakları yönetimi ve politika yapıcıları için, çalışan memnuniyetini artırma ve ayrılma oranlarını düşürme stratejileri geliştirmede yol gösterici bilgiler sağlama potansiyeline sahiptir. Araştırmanın amacı, yüksek ayrılma olasılığına sahip çalışanları erken bir aşamada tespit ederek, bu bilgiyi stratejik insan kaynakları yönetimi çabalarında kullanmak ve ayrılma oranlarını azaltmaktır. Makine öğrenmesi uygulamalarının, insan kaynakları yönetimi alanında etkin bir şekilde kullanılmasını teşvik ederek iş dünyasına önemli katkılar sunması beklenmektedir.

Makine öğrenmesi modellemelerinde karşılaşılan yaygın bir zorluk olan eksik veri problemi, sınıflar arasındaki eşit olmayan dağılım nedeniyle ortaya çıkar ve modelin tahmin performansına olumsuz etki edebilir. Bankacılık, tıp ve telekomünikasyon gibi çeşitli alanlarda, özellikle kredi kartı dolandırıcılığı, hastalık teşhisi ve kayıp müşteri analizi konularında, sınıf dengesizliğine sahip veri kümelerinin yaygın olduğu literatürde belirtilmektedir. Bu, veri setlerindeki sınıflardan birinin diğerine kıyasla aşırı temsil edildiği veya yetersiz temsil edildiği anlamına gelir.

Literatürde sınıf dengesizliğiyle mücadele yöntemleri, veri seviyeleri, algoritmalar ve topluluk öğrenmesine dayalı çözümler olmak üzere üç ana kategoride incelenir. Bu çözümler, azınlık sınıfındaki örnek sayısının artırılması (örnekleme) ya da çoğunluk sınıfındaki örnek sayısının azaltılması (eksiltme) yoluyla, veri setlerindeki dengesizliğin azaltılmasını amaçlar. Dengesizliklerin giderilmesi, etkili bir sınıflandırma modelinin geliştirilmesinde kritik bir öneme sahiptir. Dengesiz veri setlerinin ele alınması, makine öğrenmesi modellerinin daha dengeli ve doğru tahminler yapabilmesini sağlar, bu da az temsil edilen sınıfların üzerinde daha iyi bir genelleştirme yeteneğine erişilmesine imkân tanımaktadır.

3. LİTERATÜR ARAŞTIRMASI

Bu bölümde çalışan dönüşümü, çalışan devri, dengesiz veri alanlarında yapılan çalışmalardan bahis edilmiştir. Araştırmalar, ağırlık olarak işten ayrılma, çalışma dönüşümü, eksik veri sorunun giderme için kullanılan makine öğrenmesi sınıflandırma yöntemleri ve bunların hibrit çalışmalarından oluşmaktadır.

Makine öğrenmesi tekniklerinin çalışan devir hızı tahmini için kullanımı, son yıllarda birçok araştırmacının ilgisini çekmiştir. Bu bölümde, bu alanda yapılmış önemli çalışmalar detaylı bir şekilde incelenmektedir. Zhao ve arkadaşları [8], lojistik regresyon, karar ağaçları ve rastgele ormanlar gibi makine öğrenmesi yöntemlerini kullanarak bir çalışan devir hızı tahmini modeli geliştirmişlerdir. Araştırmacılar, dengesiz veri setleri için SMOTE (Sentetik Azınlık Örnekleme Tekniği) yöntemini kullanmışlardır. Sonuçlar, rastgele ormanların en iyi performansı gösterdiğini ortaya koymuştur. Ayrıca, araştırmacılar, çalışanların demografik özelliklerinin, iş performanslarının ve maaş memnuniyetlerinin, devir hızı tahmini için önemli faktörler olduğunu belirtmişlerdir. Rgut ve arkadaşları [9], extreme gradient boosting (XGBoost) algoritmasını kullanarak bir çalışan devir hızı tahmini modeli oluşturmuşlardır. Araştırmacılar, özellik seçimi ve hiper-parametre ayarlaması gibi teknikler kullanarak model performansını iyileştirmişlerdir. Sonuçlar, XGBoost modelinin yüksek doğruluk oranları sağladığını göstermiştir. Dengesiz veri setleriyle başa çıkmak için, Ma ve arkadaşları [10] Adaptif Sentetik Örnekleme (ADASYN) tekniğini kullanmışlardır. Araştırmacılar, Karar Ağaçları, Rastgele Ormanlar ve Destek Vektör Makineleri gibi farklı makine öğrenmesi yöntemlerini karşılaştırmışlar ve Rastgele Ormanların en iyi performansı gösterdiğini bulmuşlardır. Chaurasia ve Rosin [11], çalışan devir hızı tahmininde derin öğrenme yöntemlerini kullanmışlardır. Araştırmacılar, çok katmanlı algılayıcı (MLP), uzun kısa süreli bellek (LSTM) ve evrişimli sinir ağları (CNN) gibi farklı derin öğrenme modellerini karşılaştırmışlardır. Sonuçlar, LSTM modelinin en iyi performansı gösterdiğini ortaya koymuştur. Araújo ve arkadaşları [12], çalışan devir hızı tahmininde destek vektör makineleri (SVM) ve rastgele ormanlar gibi makine öğrenmesi yöntemlerini kullanmışlardır. Araştırmacılar, veri setindeki dengesizlik sorununu aşmak için ağırlıklandırma tekniğini uygulamışlardır. Sonuçlar, ağırlıklandırılmış SVM modelinin en iyi performansı gösterdiğini ortaya koymuştur. Song ve arkadaşları [13], çalışan devir hızı tahmini için bir ensemble öğrenme yaklaşımı önermişlerdir. Araştırmacılar, lojistik regresyon, karar ağaçları ve yapay sinir ağları gibi farklı makine öğrenmesi modellerini birleştirerek bir ensemble

modeli oluşturmuşlardır. Sonuçlar, ensemble modelinin tek başına modellere göre daha iyi performans gösterdiğini ortaya koymuştur. Further ve Kızılırmak [14], çalışan devir hızı tahmininde makine öğrenmesi yöntemlerinin kullanımını ve başarısını incelemişlerdir. Araştırmacılar, Lojistik Regresyon, Karar Ağaçları, Rastgele Ormanlar, Destek Vektör Makineleri ve XGBoost gibi farklı yöntemleri karşılaştırmışlardır. Sonuçlar, XGBoost ve Rastgele Ormanların en iyi performansı gösterdiğini ortaya koymuştur. Amin ve arkadaşları [15], çalışan devir hızı tahmini için bir hibrit makine öğrenmesi yaklaşımı önermişlerdir. Araştırmacılar, özellik seçimi için genetik algoritma ve sınıflandırma için rastgele ormanlar kullanmışlardır. Sonuçlar, önerilen hibrit yaklaşımın yüksek doğruluk oranları sağladığını göstermiştir. Chitra ve Uma [16], çalışan devir hızı tahmini için bir naive bayes sınıflandırıcı modeli geliştirmişlerdir. Araştırmacılar, veri setindeki dengesizlik sorununu aşmak için SMOTE yöntemini kullanmışlardır. Sonuçlar, önerilen modelin iyi bir performans gösterdiğini ortaya koymuştur. Jia ve arkadaşları [17], çalışan devir hızı tahmini için bir karar ağacı modeli önermişlerdir. Araştırmacılar, veri setindeki dengesizlik sorununu aşmak için SMOTE ve ADASYN gibi örnekleme tekniklerini kullanmışlardır. Sonuçlar, ADASYN yöntemiyle oluşturulan modelin en iyi performansı gösterdiğini ortaya koymuştur. Kuruzovich ve arkadaşları [18], çalışan devir hızı tahmini için Lojistik Regresyon ve Rassal Orman gibi makine öğrenmesi yöntemlerini kullanmışlardır. Araştırmacılar, dengesiz veri setleriyle başa çıkmak için ağırlıklandırma tekniğini uygulamışlardır. Sonuçlar, ağırlıklandırılmış rassal orman modelinin en iyi performansı gösterdiğini ortaya koymuştur. Zhang ve arkadaşları [19], çalışan devir hızı tahmini için bir gradyan arttırma makine (GBM) modeli önermişlerdir. Araştırmacılar, veri setindeki dengesizlik sorununu aşmak için SMOTE yöntemini kullanmışlardır. Sonuçlar, önerilen GBM modelinin iyi bir performans gösterdiğini ortaya koymuştur. Muazzam ve arkadaşları [20], çalışan devir hızı tahmini için bir derin öğrenme yaklaşımı önermişlerdir. Araştırmacılar, evrişimli sinir ağları (CNN) ve uzun kısa süreli bellek (LSTM) modellerini birleştirerek bir hibrit model oluşturmuşlardır. Sonuçlar, önerilen hibrit modelin yüksek doğruluk oranları sağladığını göstermiştir. Aydın ve arkadaşları [21], çalışan devir hızı tahmini için bir karar ağacı modeli geliştirmişlerdir. Araştırmacılar, veri setindeki dengesizlik sorununu aşmak için ağırlıklandırma tekniğini kullanmışlardır. Sonuçlar, ağırlıklandırılmış karar ağacı modelinin iyi bir performans gösterdiğini ortaya koymuştur. Yao ve arkadaşları [22], çalışan devir hızı tahmini için bir yapay sinir ağı (YSA) modeli önermişlerdir. Araştırmacılar, YSA modelinin performansını iyileştirmek için farklı aktivasyon fonksiyonları ve optimizasyon algoritmaları

denemişlerdir. Sonuçlar, ReLU aktivasyon fonksiyonu ve Adam optimizasyon algoritmasıyla oluşturulan modelin en iyi performansı gösterdiğini ortaya koymuştur. Rong ve arkadaşları [23], çalışan devir hızı tahmini için bir XGBoost modeli önermişlerdir. Araştırmacılar, veri setindeki dengesizlik sorununu aşmak için SMOTE ve ağırlıklandırma tekniklerini kullanmışlardır. Sonuçlar, ağırlıklandırılmış SMOTE ile oluşturulan XGBoost modelinin en iyi performansı gösterdiğini ortaya koymuştur. Goel ve arkadaşları [24], çalışan devir hızı tahmini için bir kural tabanı sınıflandırıcı modeli geliştirmişlerdir. Araştırmacılar, modelin performansını iyileştirmek için bir öznitelik seçim yöntemi kullanmışlardır. Sonuçlar, önerilen modelin iyi bir performans gösterdiğini ortaya koymuştur. Dumitrescu ve arkadaşları [25], çalışan devir hızı tahmini için bir destek vektör makinesi (DVM) modeli önermişlerdir. Araştırmacılar, farklı çekirdek fonksiyonları ve hiper-parametre ayarları denemişlerdir. Sonuçlar, radyal tabanlı çekirdek fonksiyonuyla oluşturulan DVM modelinin en iyi performansı gösterdiğini ortaya koymuştur. Lyu ve arkadaşları [26], çalışan devir hızı tahmini için bir lojistik regresyon modeli geliştirmişlerdir. Araştırmacılar, modelin performansını iyileştirmek için öznitelik seçimi ve hiper-parametre ayarlaması yapmışlardır. Sonuçlar, önerilen modelin iyi bir performans gösterdiğini ortaya koymuştur. Akyol ve arkadaşları [27], çalışan devir hızı tahmini için bir rassal orman modeli önermişlerdir. Araştırmacılar, veri setindeki dengesizlik sorununu aşmak için SMOTE ve ADASYN gibi örnekleme tekniklerini kullanmışlardır. Sonuçlar, ADASYN ile oluşturulan rassal orman modelinin en iyi performansı gösterdiğini ortaya koymuştur. Bu çalışmalar, makine öğrenmesi tekniklerinin çalışan devir hızı tahmini için etkili bir şekilde kullanılabileceğini göstermektedir. Özellikle Rastgele Ormanlar, XGBoost, Destek Vektör Makineleri ve Derin öğrenme modelleri, yüksek doğruluk oranları sağlamaktadır. Ancak, dengesiz veri setleri bu modellerin performansını olumsuz etkileyebilmektedir. Bu nedenle, ROS, SMOTE, ADASYN, ağırlıklandırma gibi teknikler kullanılarak veri dengesizliği sorunu giderilmelidir.

İnsan kaynakları disiplininde, çalışan devri üzerine yapılan geniş çaplı araştırmalar bulunmaktadır. İzleyen bölüm, fen bilimleri, sosyal bilimler, iş dünyası ve veri analizi disiplinlerinde gerçekleştirilen literatür taraması vasıtasıyla, iş probleminin tanımını, çalışan devrinin nedenlerini ve bu devrin sonuçlarını ele alarak konuya dair kapsamlı bir değerlendirme yapmayı amaçlamaktadır.

Çalışan devrinin incelenmesi, işgücü yönetimi ve organizasyonel davranış konularında kritik bir öneme sahiptir; ancak, bu alanda yapılan literatür taramalarında bazı eksikliklerin olduğu görülmektedir. İşte bu eksikliklere ilişkin bazı gözlemler:

Sektörel ve Coğrafi Çeşitlilik: Araştırmaların çoğu, belirli sektörlerle veya coğrafi alanlarla sınırlı kalmakta, bu da farklı sektör ve coğrafyalardaki çalışan devri dinamiklerinin karşılaştırmalı olarak yeterince analiz edilmemesine yol açmaktadır.

KOBİ'lere Yönelik Araştırmaların Eksikliği: Büyük şirketlere odaklanılması, küçük ve orta ölçekli işletmelerin karşılaştığı benzersiz zorluklar ve ihtiyaçlar hakkında yapılan araştırmaların göz ardı edilmesine neden olmaktadır.

Nesiller Arası Analizlerin Yetersizliği: Farklı nesillerin işgücü piyasasındaki davranış ve beklentileri üzerine yapılan çalışmalar, nesiller arasındaki farklılıkların devir oranlarına olan etkisini yeterince ele almamaktadır.

Teknolojik Değişiklikler: İş yerlerinde teknolojik değişikliklerin hızla evrimi, bu değişikliklerin çalışan devri üzerindeki etkilerinin eksiksiz bir şekilde incelenmemesine yol açmaktadır.

Psikolojik ve Duygusal Boyutlar: Çalışanların ayrılma kararlarını etkileyen psikolojik ve duygusal boyutlar üzerine derinlemesine yapılan araştırmaların sayısının yetersiz olduğu görülmektedir.

Uzaktan ve Hibrit Çalışma Modelleri: Pandemi sonrasında yaygınlaşan uzaktan ve hibrit çalışma modellerinin, çalışan devri üzerindeki etkilerini inceleyen çalışmaların yetersiz kaldığı belirlenmiştir.

Çalışan Devrinin Maliyetleri: Organizasyonlara olan maliyetleri açısından çalışan devrinin incelenmesi, genellikle dolaylı maliyetleri (bilgi kaybı, ekip dinamiklerindeki bozulmalar gibi) yeterince kapsamamaktadır.

Ayrıca bu tez çalışması kapsamında kullanılan, teknoloji sektöründeki IBM HR Analytics Employee Attrition & Performance veri setinin kullanılması ile ilgili bazı eksiklikler aşağıdaki gibi listelenmiştir.

Yeterince veri ön işleme ve veri son işleme teknikleri kullanılmadığı gözlemlenmiştir.

Veri Ön İşleme Yöntemleri:

Eksik değerlerin tamamlanması, verideki gürültünün giderilmesi, verilerin standartlaştırılması, normalleştirilmesi, özelliklerin ölçeklendirilmesi, kategorik verinin sayısallaştırılması, özelliklerden yeni bilgilerin elde edilmesi, önemli özelliklerin belirlenmesi, verinin temizlenmesi, aykırı değerlerin belirlenip işlenmesi, verinin dönüştürülmesi, boyutunun azaltılması, farklı veri kaynaklarının birleştirilmesi, verinin bölümlere ayrılması işlemleri, veri işleme sürecinde sıkça başvurulan yöntemler arasında yer alır.

Veri Son İşleme Yöntemleri:

Elde edilen sonuçların analiz edilmesi, tahminlerin gerçek değerlerle ölçeklendirilmesi, model performansının değerlendirilmesi için metriklerin kullanılması, modelin optimizasyonunun yapılması, hataların incelenmesi, özelliklerin öneminin değerlendirilmesi, sonuçların grafiklerle gösterilmesi, modelin kalibrasyonunun gerçekleştirilmesi, tahmin sonuçlarının son kullanıcı için anlaşılır hale getirilmesi, modelin periyodik olarak güncellenmesi, elde edilen bilgilerin karar destek sistemlerine entegre edilmesi, bulguların raporlanması, bu sürecin son aşamalarında uygulanan yöntemlerdendir.

Yukarıda belirtiğim veri ön işleme, veri son işleme adımlarını olabileceğinden bol alternatifli olacak şekilde bu çalışmada kullanılmıştır. Amaç daha yüksek doğruluğa sahip sonuçları veren yüksek performanslı modeller oluşturmaktır. Oluşan bu sonuçları sektörel bazda kurum ve kuruluşların kullanımına sunarak faydalanmalarını sağlamaktır.

Bu tez araştırması, makine öğrenimi algoritmalarının çalışan devir hızının tahmin edilmesindeki kullanımı üzerine önemli bir katkı olarak öne çıkmaktadır. Mevcut literatüre derinlemesine bir bakış atan tez, literatürdeki eksiklikleri belirlemekte ve bu eksikliklerin nasıl giderileceğine dair yol göstermektedir. Çalışan devir hızını düşürme ve çalışanların kurumda tutulmasını sağlama konusunda, makine öğrenimi algoritmaları önemli bir araç olma potansiyeline sahiptir.

Araştırmanın ilerlemesi için öneriler şunlardır:

Dengesiz verilerle başa çıkmak amacıyla yeni yöntemlerin geliştirilmesi,

- Çalışan devir hızını tahmin etmede üstün performans gösteren makine öğrenimi algoritmalarının belirlenmesi,
- Makine öğrenimi algoritmalarının etik kullanımı üzerine araştırmalar yapılması,

- Makine öğrenimi algoritmalarının farklı sektörlerde ve iş güçlerinde çalışan devir hızını tahmin etme kapasitesinin nasıl değişebileceğine dair araştırmaların genişletilmesi gerekmektedir.

Bu çalışma, kurumlar için önemli bir maliyet kaynağı olan çalışan devir oranlarının makine öğrenmesi teknikleri aracılığıyla nasıl etkin bir biçimde öngörülebileceğine dair bir inceleme sunmaktadır. Bu bağlamda, çalışmanın, dengesiz veri setlerinin ele alınması, çalışan devir hızının tahmininde üstün sonuçlar sunan makine öğrenmesi algoritmalarının tespiti ve bu algoritmaların kullanımının etik boyutlarının incelenmesi gibi alanlarda literatüre katkıda bulunduğu gözlemlenmiştir.

Literatüre sunulan katkılar şu şekilde özetlenebilir:

Dengesiz Veri Setlerinin İşlenmesi: Bu tez, dengesiz veri setlerini işleme konusunda ileri düzey yöntemler geliştirilmesiyle alakalı değerli bilgiler sağlar. Makine öğrenmesi modellerinin, özellikle az temsil edilen sınıfların doğru bir şekilde tanımlanabilmesi için gerekli olan daha doğru ve güvenilir tahminler yapabilmesine olanak tanıyan bu yöntemler, modelin genel doğruluğunu artırmaktadır.

Çeşitli Makine Öğrenimi Algoritmalarının Değerlendirilmesi: Çalışma, çeşitli makine öğrenimi algoritmalarının (örneğin, Random Forest, XGBoost, AdaBoost) çalışan devir tahminindeki etkinliğinin derinlemesine analizini sunar. Bu analiz, hangi algoritmaların bu özel tahmin görevinde en etkili olduğunu belirlemekte ve bu alanda daha bilinçli kararlar alınmasını sağlamaktadır.

Hiperparametre Optimizasyon Yöntemlerinin Karşılaştırılması: Çalışma, modelin performansını maksimize etmek için çeşitli hiperparametre optimizasyon tekniklerini karşılaştırır. Bu tekniklerin etkili bir şekilde nasıl kullanılacağına dair sağlanan bilgiler, modelin tahmin kabiliyetini iyileştirmekte önemli bir role sahiptir.

Etik Kullanımın Vurgulanması: Algoritmaların etik kullanımı üzerine yapılan tartışmalar, teknolojinin insan kaynakları yönetimi gibi hassas alanlarda nasıl sorumlu bir şekilde kullanılması gerektiğine dair önemli perspektifler sunmaktadır.

Sonuç olarak, bu tez araştırması, çalışan devir hızının tahmin edilmesi konusunda makine öğrenmesi algoritmalarının potansiyelinden nasıl yararlanılabileceğini göstermekle kalmayıp, dengesiz veri setleriyle başa çıkma, makine öğrenmesi algoritmalarının performans değerlendirmesi ve etik kullanım gibi konularda literatürdeki mevcut boşlukları

doldurmayı hedeflemektedir. Bu çalışma, iş gücü devamlılığının sağlanması ve kurumların çalışan devir oranlarını azaltma yönündeki çabalarına destek olacak bilgiler sunarak, insan kaynakları yönetimi alanında önemli bir referans kaynağı olma potansiyeline sahiptir.

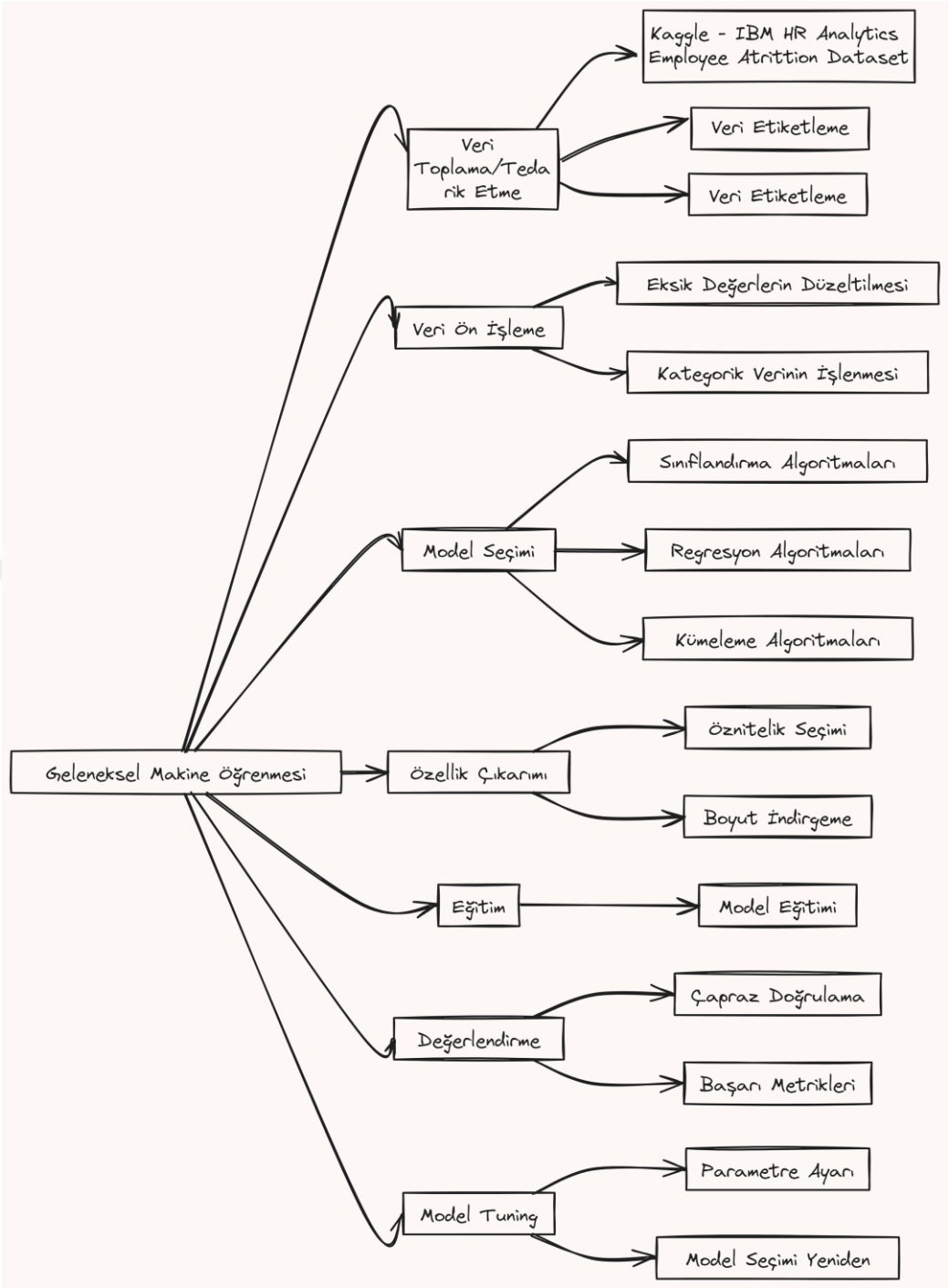
4. UYGULAMANIN KAPSAMI, VERİ ÖN İŞLEME FAALİYETLERİ

Bu tez çalışması, özellikle çalışan devri dinamiklerini derinlemesine incelemektedir. Araştırmanın odak noktasını oluşturan problem, çalışanların işten ayrılma eğilimlerinin altında yatan sebepleri ve bu eğilimlerin zamanlamasını kapsamlı bir şekilde anlamak üzerine kurulmuştur. Bu derinlemesine anlayış, organizasyonların çalışanlarını daha etkin bir biçimde işte tutmalarına, geliştirilmesi gereken alanları tespit etmelerine ve potansiyel yeni işe alımları önceden planlama imkânı sağlama potansiyeline sahiptir.

Araştırma, eski çalışanlara ait verileri barındıran bir veri seti üzerinden yürütülmüştür.

4.1. Tez Çalışması Akış Adımları

Tez çalışmasında bir makine öğrenmesi yöntemlerini uyguladığımız çalışmanın tüm adımları Visio programı kullanarak süreç adımları çizilerek Şekil 1'deki gibi gösterilmiştir.



Şekil 1. Süreç akış diyagramı

Yukarıdaki Şekil 1'de tüm adımlarda yapılan işlemlerin özet bilgileri sırasıyla aşağıda açıklanmıştır. Veri toplama, veri ön işleme, veri keşfi ve analizi, özelik mühendisliği, veri bölme, model seçimi ve eğitimi, model değerlendirme, Hiperparametre ayarlama, son modelin eğitilmesi, modelin dağıtılması, modelin izlenmesi ve güncellenmesi işlemlerinden oluşmaktadır.

4.2. Veri Setinin Temin Edilmesi ve Tanıtımı

Bu tez çalışmasında, IBM HR Analytics Employee Attrition & Performance veri kaynağından[28] elde edilen, 1.470 çalışanın verilerini ve bu çalışanlara dair çeşitli bilgileri içeren bir İK veri seti temin edilmiştir. Bu veri seti, çalışan kaybının temel nedenlerini anlamak ve çalışanların işten ayrılma zamanlarını tahmin etmek amacıyla kullanılacaktır [29].

Yukarıda tanımlanan veri seti, toplamda 35 değişken içermekte olup, bu değişkenler 26 tanesi sayısal (int64) ve geriye kalan 9 tanesi kategorik (object) türlerde sütunlar olarak yer almaktadır. Veri setinin değişken isimleri ve tanımları Tablo 1'de detaylandırılmıştır.

Tablo 1. Değişken Listesi ve Tanımları

Değişken İsmi	Açıklaması
Yaş	Çalışanın yaşı
İşten Ayrılma	Ayrılma ya da ayrılmama
İş Seyahati	İş gereksinimine göre seyahat
Günlük Ücret	Çalışanların günlük maaş oranı
Departman	Çalışanların ait olduğu bölüm
Eve Uzaklık	KM cinsinden evden uzaklık
Eğitim	Eğitim Türü
Eğitim Alanı	Eğitim alanı
Çalışan Sayısı	Raporlanan kişi sayısı
Çalışan Numarası	Çalışan Kodu/ID
Çalışma Ortamı Memnuniyeti	Çevre Memnuniyeti
Cinsiyet	Çalışanın cinsiyeti
Saatlik Ücret	Çalışanların saatlik maaş oranı
İşe Bağlılık	Anket 360'a göre
İş Seviyesi	Çalışan Notu
İş Rolü	Satış Yöneticisi, Araştırma Bilimcisi, Laboratuvar Teknisyeni, Üretim Direktörü, Sağlık Hizmetleri Temsilcisi, Müdür, Satış Temsilcisi, Araştırma Direktörü, İnsan Kaynakları
İş Memnuniyeti	İş Memnuniyeti Anketine dayanmaktadır (Daha çok yapılan görevle ilgilidir)
Medeni Durum	Medeni Durum
Aylık Gelir	Net Gelir
Aylık Ücret	Aylık CTC
Çalışılan Şirket Sayısı	Çalışılan Şirket Sayısı
18 Yaş Üstü	Evet/Hayır
Fazla Mesai	Uygunluk
Maaş Artış Yüzdesi	% Yürüyüş

Performans Değerlendirmesi	Değerlendirme
İlişki Memnuniyeti	360 geri bildirim
Standart Çalışma Saatleri	Çalışma Saatleri
Hisse Senedi Seçeneği Seviyesi	Hisse Senedi Opsiyonu Seviyesi
Toplam Çalışma Yılları	Çalışma yılı

4.3. Veri Ön İşleme

4.3.1. Veriyi Sisteme Aktarma

Veri setinin seçimi tamamlandıktan sonra, daha önceden belirlenen araç ve tekniklerin yardımıyla, makine öğrenimi çalışması için gereken ortamın hazırlığına yönelik adımlar atılmıştır. Geliştirme sürecinde, PyCharm geliştirme ortamının kullanımı tercih edilerek, bu ortam üzerinde yapılandırma işlemleri gerçekleştirilmiştir. Kod sürüm yönetimi açısından, çalışmanın kodlama aşamaları ve dokümantasyonlarının etkin bir şekilde yönetilebilmesi amacıyla GitHub sürüm oluşturma aracı kullanılmıştır. Bu sayede, tez çalışmasının temel bileşenlerini oluşturan kodlamaların, GitHub üzerinden sistematik bir şekilde yönetimi sağlanmıştır. Ayrıca, Kaggle platformu üzerinden edinilen ve csv formatında bulunan veri seti, pandas kütüphanesinin imkanlarından faydalanılarak çalışma ortamına entegre edilmiştir. Bu süreç, veri setinin analiz ve işlenmesi aşamalarının verimli bir şekilde gerçekleştirilmesine olanak tanımıştır.

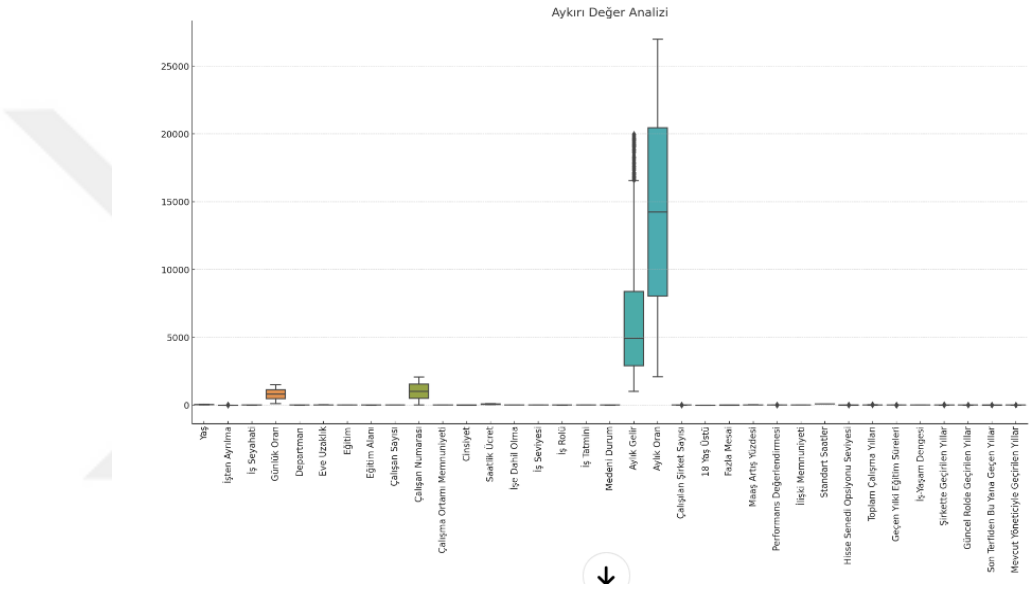
4.3.2. Verinin İşlenmeye Hazır Hale Getirilmesi için Temizlenmesi

Veri üzerinde sırasıyla aşağıdaki faaliyetler yapılmıştır.

- Değişken yapıları incelenmiştir.
- Değişken tipleri incelenmiştir.
- Tip dönüşümleri yapılmıştır.(Kategorik değişken tipleri object iken float tipe dönüştürülmüştür.)
    - Makine öğrenimi algoritmalarının tahmin yapısında genellikle sayısal değerler tercih edilir. Bu bağlamda, kategorik değişkenlerin sayısal değerlere dönüştürülmesi gereklilik arz eder. Bu dönüşüm için, özellikle çok sayıda

benzersiz değer içeren kategorik özelliklerde, özelliğin önemini doğru bir şekilde yansıtmak adına hem Etiket Kodlama hem de Tek Sıcak Kodlama teknikleri kullanılacaktır, bu teknikler aşağıda detaylandırılmıştır.

- Kategorik olan değişkenler dikey sütün halinde ve yatay sütün haline getirilerek encoding uygulanmıştır.
- Veriye max min scaler uygulanarak outler olan değerlere 0.5 Aralığında bir değerler atanarak aralık daraltılmıştır.
- Attration(yıpranma) değerlerindeki dengesizlik için unbalasing(dengesiz) metodlar uygulanarak "Evet" sayıları yükseltilmiştir.



Şekil 2. Aykırı değer analizi

Şekil 2'de aykırı değerlerin tespitinde boxplot yöntemi kullanılarak gösterilmiştir. Kutuların dışında kalan noktalar, bu grafikte aykırı değerleri işaret etmektedir.

Söz konusu veri setinde, bazı sütunlarda dikkat çekici miktarda aykırı değerlerin bulunduğu belirlenmiştir. Özellikle "Günlük Oran", "Aylık Gelir", "Aylık Oran" ve "Toplam Çalışma Yılları" sütunları bu aykırılıkların yoğun olduğu alanlar olarak öne çıkmaktadır. Bu tür aykırı değerlerin nasıl işleneceği konusunun, analiz sürecinde önemli bir yer tuttuğu anlaşılmaktadır.

4.3.3. Keşifsel Veri Görselleştirilmeleri

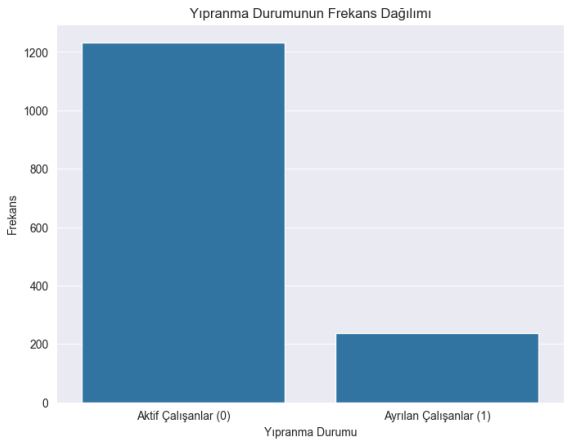
Şekil 3'teki histogramlar incelendiğinde, sayısal özelliklere ilişkin çeşitli gözlemler yapılabilir. Bu histogramların çoğu, kuyruk ağırlıklı yapıdadır ve bazı dağılımların, örneğin

Aylık Gelir, Mesafeden Ev Uzaklık ve Şirketteki Yıllar, sağa çarpık olduğu görülmektedir. Normal dağılıma yakınlaşmak için, model uygulanmadan önce verilerin dönüştürülmesi gerekmektedir. Yaş dağılımı ise, hafif sağa çarpık bir normal dağılım göstermekte ve personelin büyük bir bölümünün 25 ile 45 yaş aralığında olduğu belirlenmektedir. Çalışan Sayısı ve Standart Saatler gibi, tüm çalışanlar için sabit değerler gösteren özelliklerin gereksiz olabileceği düşünülmektedir. Ayrıca, özelliğin yarı-üniform dağılımı göz önünde bulundurularak, Çalışan Numarası'nın çalışanlar için benzersiz bir tanımlayıcı olabileceği öngörülmektedir.

Şekil 3. Değişken dağılımları

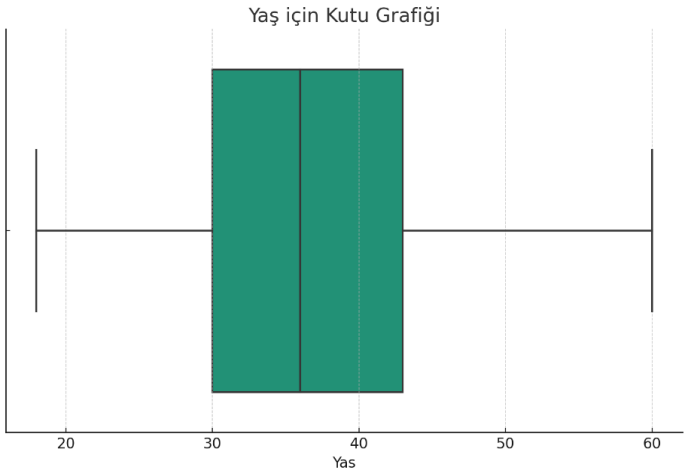
Sayısal özelliklere ilişkin bilgilere ve histogramlara dayanarak birkaç gözlem yapılabilir: Çoğu histogram kuyruk ağırlıklıdır; aslında bazı dağılımlar sağa çarpıktır (örneğin, Aylık Gelir, Evden Uzaklık Mesafesi, Şirketteki Yıllar). Verilere bir model uydurmadan önce normal dağılıma yaklaşmak için veri dönüştürme yöntemleri gerekebilir. Yaş dağılımı hafif sağa çarpık normal bir dağılım olup personelin büyük bir kısmı 25 ile 45 yaş arasındadır. EmployeeCount ve StandardHours tüm çalışanlar için sabit değerlerdir. Bunların gereksiz özellikler olması muhtemeldir. Özelliğin yarı-üniform dağılımı göz önüne alındığında, Çalışan Numarasının çalışanlar için benzersiz bir tanımlayıcı olması muhtemeldir.

Bağımlı değişken olarak seçtiğimiz Yıpranma değişkenin veri seti içerisindeki dağılımı aşağıda Şekil 4'teki gibidir. Aktif çalışan sayısının 1200 civarında olduğu, işten ayrılan personelin 200 küsure yakın olduğu görülmektedir.



Şekil 4. Yıpranma değişkenin veri seti içindeki sayısal dağılımı

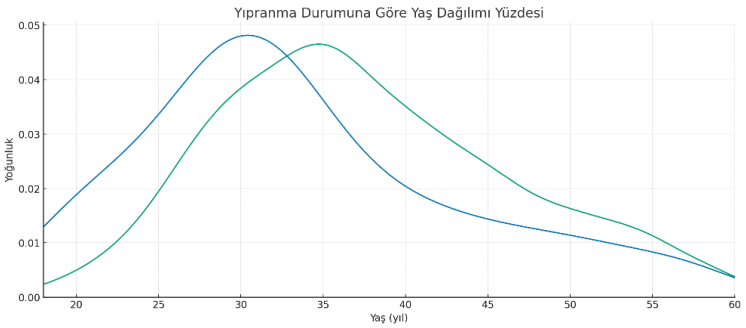
Grafikten elde edilen verilere dayanarak, veri setimizdeki sınıf dağılımının dengesiz bir yapıda olduğu net bir şekilde gözlemlenmektedir. Belirtilen verilere göre, hâlihazırda şirkette çalışmakta olan personelin oranı %83,9 iken, şirketten ayrılmış olan personelin oranı %16,1 olarak tespit edilmiştir. Makine öğrenmesi algoritmalarının, her bir sınıfın benzer sayıda örneğe sahip olduğu veri setlerinde daha başarılı sonuçlar ürettiği bilinmektedir. Bu bağlamda, makine öğrenmesi algoritmalarının verimli bir şekilde uygulanabilmesi adına, söz konusu hedef özelliğe ait dengesizliğin öncelikle giderilmesi önem arz etmektedir.



Şekil 5. Yaş değişkenine ait kutu grafiği

"Yaş" değişkenine ait kutu grafiği Şekil 5'te gibi, veri setindeki yaş dağılımının merkezi eğilimini, yayılımını ve olası aykırı değerleri göstermektedir. Grafiğin incelenmesi

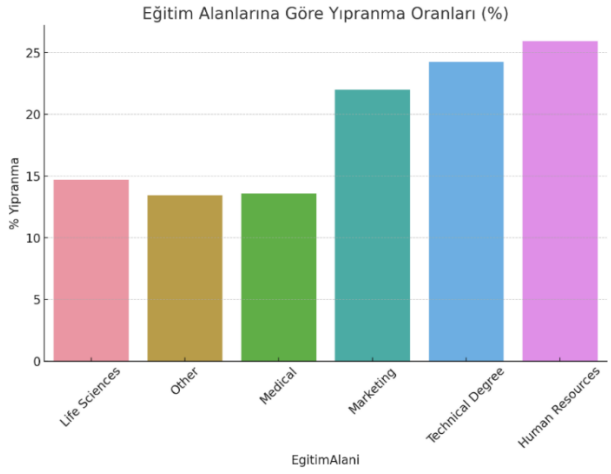
sonucunda, yaş dağılımının genel olarak dengeli bir yapıda olduğu ve belirgin aykırı değerlerin bulunmadığı tespit edilmiştir.



Şekil 6. Yıpranma durumuna göre yaş dağılımı yüzdesi

"Yıpranma" durumuna göre "Yaş" dağılımının yoğunluk grafiği çizilmiştir. Şekil 6'daki grafik, işten ayrılma eğiliminde olan (Yıpranmış Çalışanlar) ve işten ayrılmayan (Yıpranmamış Çalışanlar) çalışanların yaş dağılımlarını karşılaştırmalı olarak sunmaktadır.

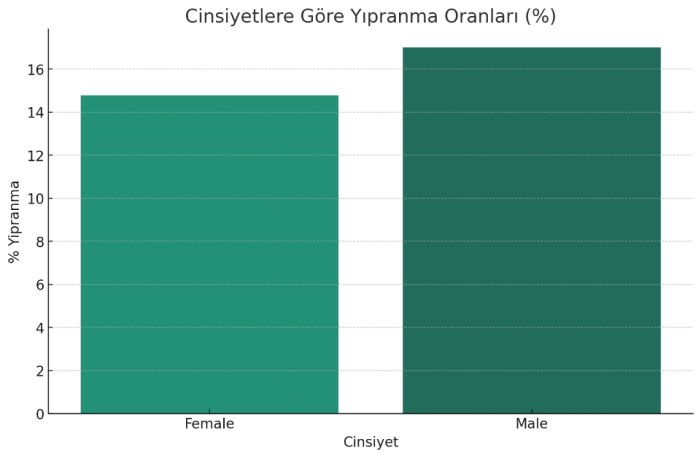
Yıpranma durumu evet olan kayıtlardaki çalışanlar kategorisi, işten ayrılma eğilimi gösteren çalışanları ifade etmektedir. Yıpranma durumu hayır olan kayıtlardaki çalışanlar kategorisi ise işten ayrılmayan çalışanları temsil etmektedir. Grafikten elde edilen veriler, bu iki grup arasındaki yaş dağılımlarındaki farklılıkların gözlemlenmesi açısından değerlendirilebilir. Bu grafik üzerinden daha detaylı bir analiz gerçekleştirilmesi talep edildiğinde veya başka bir analiz ihtiyacı ortaya çıktığında, ilgili değerlendirmeler yapılabilir.



Şekil 7. Eğitim alanına göre yıpranma oranları

Veri setindeki "Eğitim Alanı"na göre "Yıpranma" oranları, Şekil 7'deki bar grafiği ile görselleştirilmiştir. Bu grafik, farklı eğitim alanlarında yer alan çalışanların işten ayrılma oranlarını yüzdesel olarak sergilemektedir. Her bir eğitim alanı, ilgili alandaki çalışanların

yıpranma oranlarını yansıtmaktadır. Grafik incelendiğinde, eğitim alanlarına göre yıpranma oranlarının nasıl farklılık gösterdiği gözlemlenmektedir.



Şekil 8. Cinsiyetlere göre yıpranma oranları

Veri setinde "Cinsiyet" kategorisine göre "Yıpranma" oranları Şekil 8'deki bar grafiği ile görselleştirilmiştir. Bu grafik, farklı cinsiyet kategorilerindeki çalışanların işten ayrılma oranlarını yüzdesel olarak sergilemektedir. Her bir cinsiyet, ilgili cinsiyetteki çalışanların yıpranma oranlarını yansıtmaktadır.

Grafik incelendiğinde, cinsiyetler arasındaki yıpranma oranlarının nasıl farklılık gösterdiği gözlemlenmektedir.

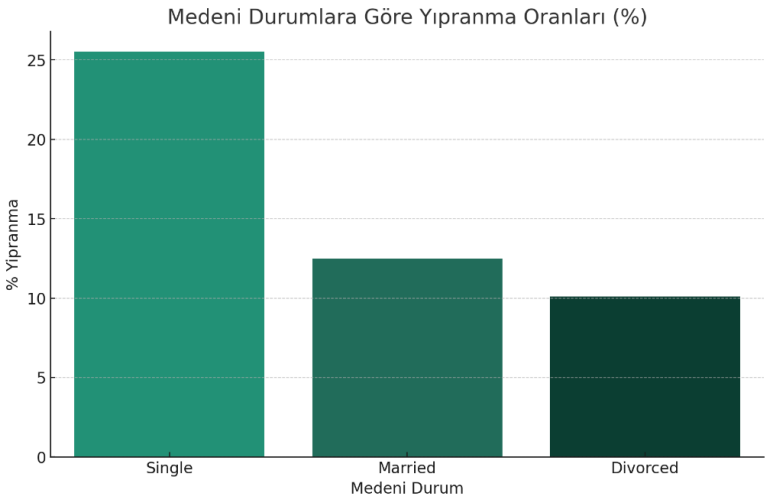
Veri setinde yer alan çalışanların medeni durumlarına ilişkin dağılım şu şekildedir:

Evli (Married) olarak belirtilen çalışan sayısı: 673

Bekâr (Single) olarak belirtilen çalışan sayısı: 470

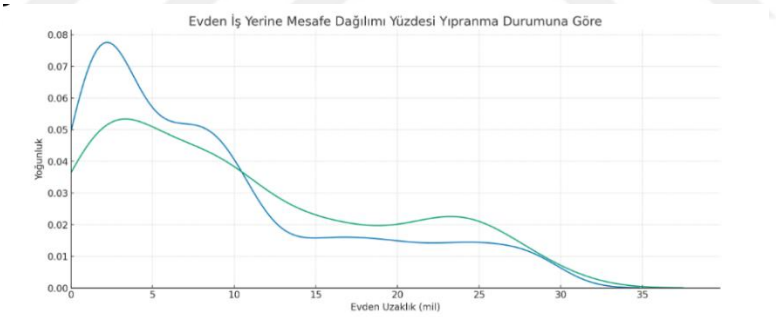
Boşanmış (Divorced) olarak belirtilen çalışan sayısı: 327

Bu dağılım, veri setinde bulunan çalışanların medeni durumlarının sayısal bir özetini sunmaktadır.



Şekil 9. Medeni Duruma göre Yıpranma oranları

Veri setinde "Medeni Durum" kategorisine göre "Yıpranma" oranları Şekil 9'daki bar grafiği ile görselleştirilmiştir. Bu grafik, farklı medeni durum kategorilerindeki çalışanların işten ayrılma oranlarını yüzdesel olarak sergilemektedir. Her bir medeni durum, ilgili durumdaki çalışanların yıpranma oranlarını yansıtmaktadır. Grafik incelendiğinde, medeni durumlar arasındaki yıpranma oranlarının nasıl farklılık gösterdiği gözlemlenmektedir.



Şekil 10. Evden İş Yerine Mesafe dağılımı yüzdesi Yıpranma durumuna göre dağılımı

Veri setinde bulunan "Evden Uzaklık" değişkenine dayanarak, aktif çalışanlar ve eski çalışanlar arasındaki mesafe dağılımının yoğunluk grafiği Şekil 10'da çizilmiştir. Bu grafik, şirkette hala çalışmakta olan kişilerle (Yıpranma 'Hayır') ve şirketten ayrılmış kişilerle (Yıpranma 'Evet') arasındaki evden iş yerine mesafe dağılımını göstermektedir.

Grafiğin analizi, aktif çalışanlar ve eski çalışanlar arasında evden iş yerine mesafe dağılımındaki farklılıkları belirlemek için kullanılabilir.

4.4. Model Seçimi

Bu tez araştırmasında 3.2. İşlem adımında ve 3.3. İşlem adımında detaylandırılan veri seti için en uygun olacak modeller 3.Adımdaki LİTERATÜR ARAŞTIRMASI işlem adımındaki incelenen diğer ilişkili araştırmalar göz önünde bulundurularak en uygun yöntemlerin Geleneksel makine öğrenmesi sınıflandırma ve hibrit algoritmalarının olduğu gözlemlenmiştir.

4.5. Özellik Çıkarımı(Değişken Seçimi)

Bu tezde, kullanılan veri setinden yapılan analiz faaliyetleri sonunda, "Yıpranma" değişkenini bağlı değişken (hedef) olarak belirlendi. Özellik mühendisliği yapmak, bu değişkeni etkileyebilecek faktörleri belirlemek ve modelleme için veri setini hazırlamak anlamına gelmektedir. "Yıpranma" değişkeni, çalışanların işten ayrılma eğilimi ya da memnuniyetsizlik durumunu gösteren bir ikili (binary) değişken görünümdedir. Bu tür bir değişken için sınıflandırma modeli kullanılabilir. Özellik mühendisliği yaparken, "Yıpranma" ile potansiyel olarak ilişkili olabilecek özellikler üzerinde durulacaktır. Bu süreç şunları içerecektir:

Korelasyon Analizi: "Yıpranma" ile diğer sayısal özellikler arasındaki korelasyonları inceleyerek hangi özelliklerin önemli olabileceği tespit edilecektir.

Kategorik Değişken Analizi: Kategorik değişkenlerin "Yıpranma" üzerindeki etkisini incelenecektir.

Yeni Özellikler Oluşturma: Mevcut verilerden, "Yıpranma" ile ilişkili olabilecek yeni özellikler türetilecektir.

İlk adım olarak "Yıpranma" ile diğer sayısal değişkenler arasındaki korelasyonları incelenerek ve bu bilgiye dayanarak ilerlenecektir. "Yıpranma" değişkenini sayısal bir formata daha önceki veri ön işleme adımları basamağında gerçekleştirmiştik.

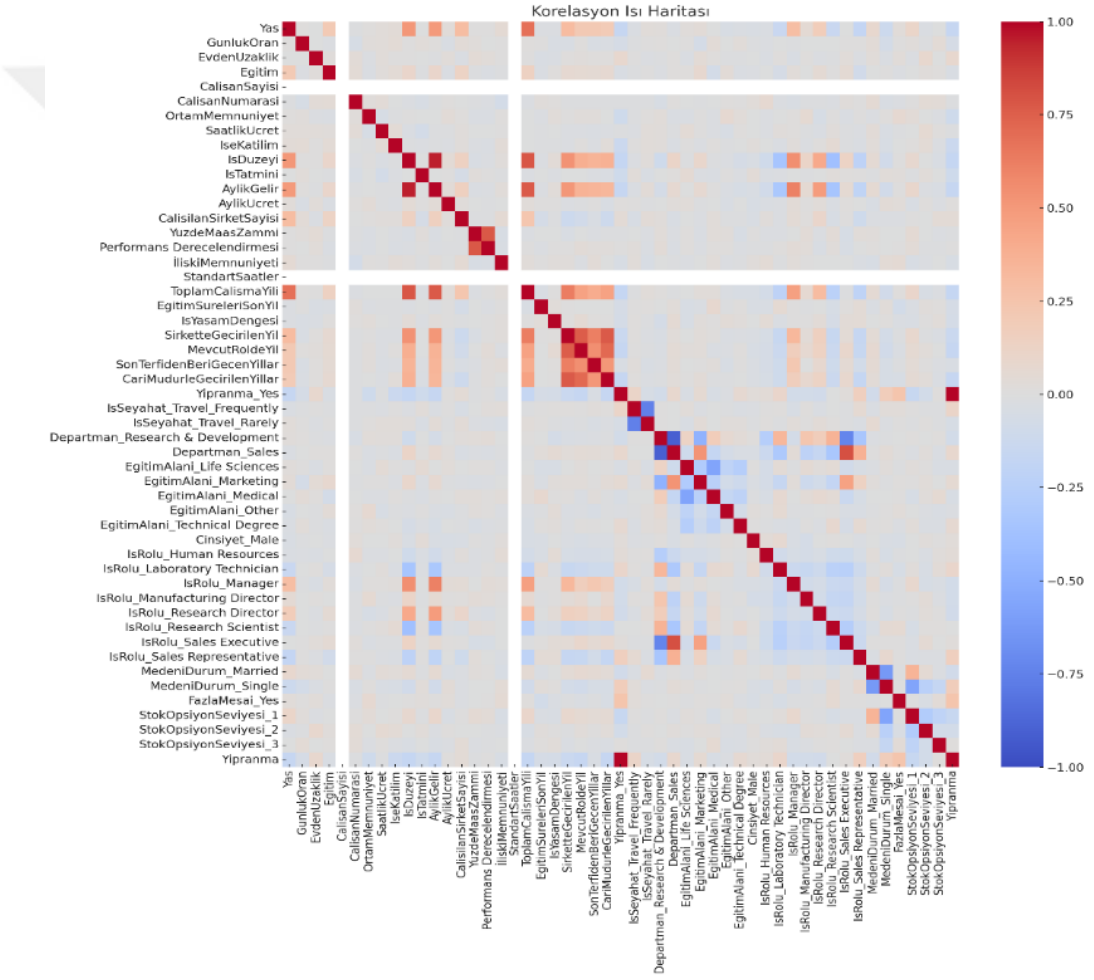
"Yıpranma" değişkeni ile diğer değişkenler arasındaki korelasyonlar Şekil 11'de gösterildiği gibi hesaplanmıştır. Yapılan hesaplanma sonrası önemli bulgular elde edilmiştir. Bulgular aşağıdaki gibi açıklanmıştır.

Pozitif Korelasyonlar için, 'FazlaMesai_Yes', 'MedeniDurum_Single', ve 'IsSeyahat_Travel_Frequently' gibi değişkenler, "Yıpranma" ile pozitif korelasyona sahiptir. Bu, bu özelliklerin yüksek değerlerinin, yıpranmanın olasılığını artırdığını göstermektedir.

Negatif Korelasyonlar için, 'Yas', 'AylikGelir', 'IsDuzeyi' gibi değişkenler, "Yıpranma" ile negatif korelasyona sahiptir. Bu, bu özelliklerin yüksek değerlerinin, yıpranmanın olasılığını azalttığını göstermektedir.

Bazı değişkenler ('CalisanSayisi', 'StandartSaatler') herhangi bir korelasyon göstermemekle beraber, bu değişkenlerin "Yıpranma" üzerinde önemli bir etkisi olmadığını göstermektedir.

Bu korelasyonlar, hangi özelliklerin "Yıpranma" tahmin modelinde daha önemli olabileceğine dair bir fikir vermektedir. Ancak, korelasyon tek başına nedensellik anlamına gelmez ve modelin başarısını değerlendirmek için bu özelliklerin bir model içinde test edilmesi gerekmektedir.

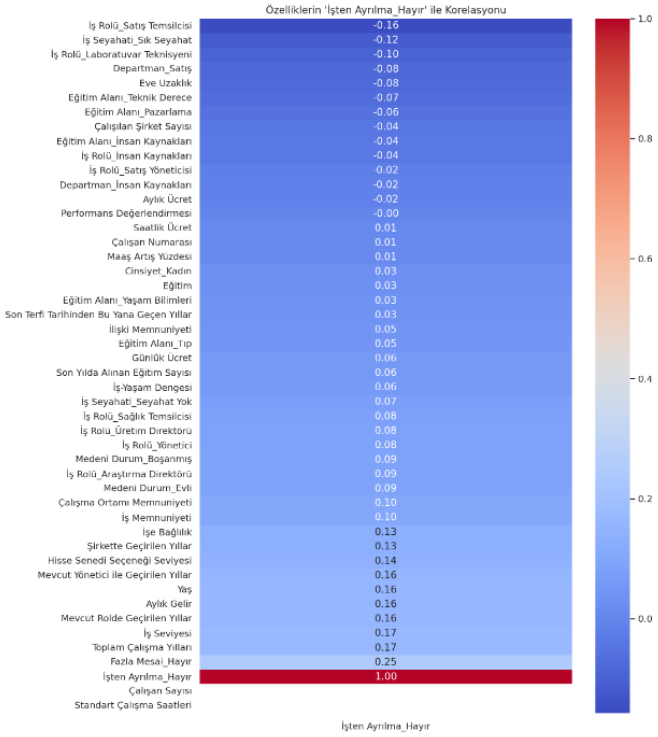


Şekil 11. Korelasyon Isı Haritası

Korelasyon ısı haritası, veri setimizdeki değişkenler arasındaki ilişkileri görsel bir şekilde göstermektedir. Bu ısı haritasında, her bir değişken çifti arasındaki korelasyon katsayılarını renklerle ifade edilmiştir. Kırmızı renkler, pozitif korelasyonu göstermektedir. Bu, bir değişkenin değeri arttığında diğer değişkenin değerinin de artma eğiliminde olduğunu

göstermektedir. Mavi renkler, negatif korelasyonu göstermektedir. Bu, bir değişkenin değeri arttığında diğer değişkenin değerinin azalma eğiliminde olduğunu göstermektedir. Renk yoğunluğu, korelasyonun gücünü göstermektedir. Daha yoğun renkler, daha güçlü bir korelasyonu ifade etmektedir.

Bu harita, hangi değişkenlerin birbiriyle daha güçlü bir ilişkiye sahip olduğunu ve bunların "Yıpranma" değişkeniyle olan ilişkilerini anlamada yardımcı olabilmektedir. Bu bilgi, modelleme aşamasında hangi özelliklerin dâhil edilmesi gerektiği konusunda karar vermek için kullanılabilmektedir.



Şekil 12. İşten Ayrılma Hayır durumunun diğer değişkenlerle korelasyonu

Şekil 12'deki görselde sunulan ısı haritası, 'İşten Ayrılma_Hayır' durumu ile diğer özellikler arasındaki korelasyonları sayısal değerlerle ifade etmektedir. Pozitif değerlerin, bir özelliğin artması halinde işten ayrılmama ihtimalinin yükseldiğini göstermesi, negatif değerlerin ise bu durumun tam tersini, yani özelliğin artması ile işten ayrılma ihtimalinin yükseldiğini ifade etmesi dikkate alınmıştır. Bu korelasyon değerleri, hangi özelliklerin işten ayrılmayı daha fazla etkilediğine dair önemli ipuçları sunmakta ve modelleme sürecinde kritik bir rol oynamaktadır.

Veri setinin incelenmesi sonucunda, eksik veya hatalı veri değerlerinin bulunmadığı ve tüm özelliklerin uygun veri tipine sahip olduğu anlaşılmıştır. Performans Derecelendirmesi,

Aylık Ücret, Çalışılan Şirket Sayısı ve Eve Uzaklık gibi faktörlerin hedef özelliklerle pozitif bir korelasyon içinde olduğu görülmüştür. Öte yandan, Toplam Çalışma Yılları, İş Seviyesi, Güncel Görevde Geçirilen Yıllar ve Aylık Gelir gibi özelliklerin hedef özelliklerle negatif bir ilişkisi tespit edilmiştir.

Veri setindeki gözlemlerin çoğunluğu Halen Aktif Çalışanlarla ilgili olduğundan, dengesiz bir yapıya sahip olduğu belirlenmiştir. Çalışan Sayısı, Çalışan Numarası, Standart Saatler ve 18 Yaş Üstü gibi analiz için gereksiz bazı özelliklerin de olduğu saptanmıştır.

Diğer dikkat çekici bulgular şöyle sıralanabilir:

- Bekâr çalışanların, evli ve boşanmış olanlara kıyasla ayrılanlar arasında daha büyük bir oranı temsil ettiği gözlemlenmiştir.
- Çalışanların yaklaşık %10'unun şirketteki ikinci yıllarında işten ayrıldığı kaydedilmiştir.
- Daha yüksek maaş alan ve daha fazla sorumluluk üstlenen sadık çalışanların, diğerlerine göre daha az ayrıldığı tespit edilmiştir.
- İş yerlerine uzak mesafede yaşayan kişilerin, diğerlerine göre daha fazla ayrıldığı belirlenmiştir.
- Sık seyahat eden bireylerin, diğerlerine kıyasla daha yüksek oranda işten ayrıldığı gözlemlenmiştir.
- Fazla mesai yapan bireylerin, diğerlerine göre daha fazla işten ayrıldığı tespit edilmiştir.
- Satış Temsilcisi pozisyonundaki çalışanların, veri setinde önemli bir yüzdeyle ayrılanlar arasında yer aldığı belirlenmiştir.

Daha önceden birkaç şirkette çalışmış olan bireylerin, diğerlerine kıyasla daha fazla işten ayrıldığı saptanmıştır. Yukarıdaki bulgular benzer araştırma ve veri seti ile çalışacak araştırmaların faydalanabilmesi için eklenmiştir.

Bu dikkat çekici bulguların makine öğrenmesiyle olan ilişkisi, bu tür verilerin işten ayrılma eğilimlerini tahmin etmek ve anlamak için kullanılabilecek değerli içgörüler sağlamasıdır. Makine öğrenmesi algoritmaları, veri setlerindeki desenleri ve ilişkileri tespit etme yeteneğine sahiptir. Dolayısıyla, yukarıda belirtilen bulgular, özellikle işten ayrılma olasılığını artıran faktörlerin belirlenmesinde kritik öneme sahiptir.

Makine öğrenmesi modelleri, bu tür faktörlere dayanarak, hangi çalışanların işten ayrılma olasılığının daha yüksek olduğunu tahmin edebilir. Örneğin, bekar çalışanlar, uzak mesafede yaşayan kişiler, sık seyahat edenler, fazla mesai yapanlar veya belli pozisyonlardaki (örneğin, Satış Temsilcisi) çalışanların ayrılma eğilimlerinin daha yüksek olduğunu belirleyebilir.

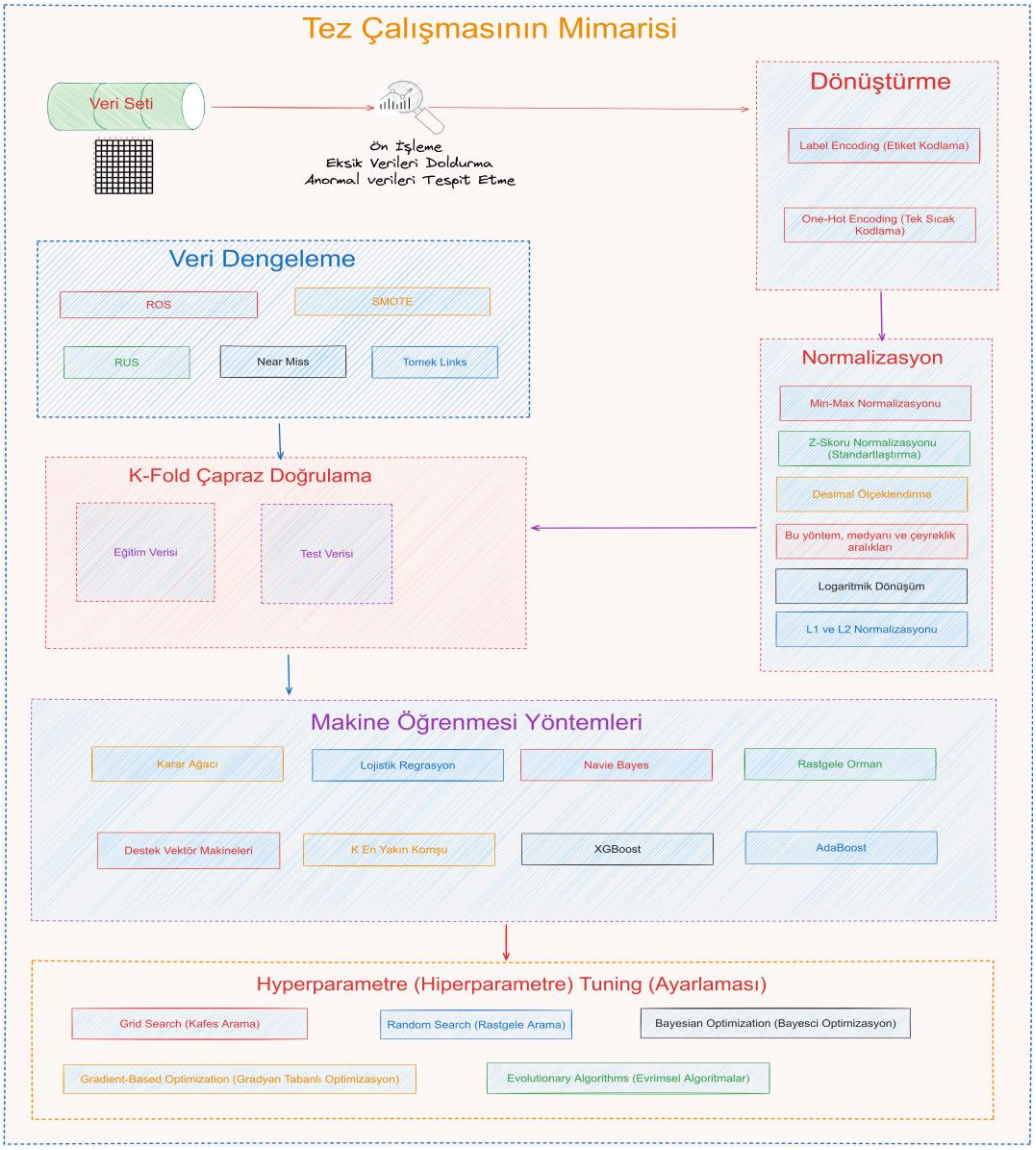
Bu bulguları analiz ederek, özellikle işten ayrılma riski yüksek olan grupları hedefleyen özelleştirilmiş müdahale stratejileri geliştirilebilir. Örneğin, daha yüksek maaş ve sorumluluk sunarak sadık çalışanları elde tutma veya iş yerine yakın konut imkanları sunarak uzakta yaşayanların ayrılma oranını düşürme gibi stratejiler uygulanabilir.

Sonuç olarak, makine öğrenmesi modelleri, işten ayrılma eğilimlerini anlamak ve proaktif çözümler geliştirmek için bu bulguları kullanarak şirketlerin insan kaynakları stratejilerini iyileştirmelerine yardımcı olabilir. Bu yaklaşım, çalışan memnuniyetini artırarak iş gücü devir hızını azaltmaya yönelik etkili bir adımdır.

Şekil 2'deki Süreç ve akış diyagramı çizelgesindeki Eğitim, Değerlendirme ve Model Tuning başlıklarının faaliyetleri 4. Adımdaki BULGULAR VE YORUM başlığı altında gerçekleştirilmiştir.

5. UYGULAMADA KULLANILAN YÖNTEMLER VE METOTLAR

Tez araştırmasının bu bölümünde, üzerinde çalışma yapılacak veri setinin araştırma süresince aşağıda Şekil 13'te gösterilen yöntemler araştırılarak tanımlanmıştır. Sonrasında tez araştırması içeresinde 5.BULGULAR VE YORUMLAR bölümünde kullanılmıştır.



Şekil 13. Tez Çalışması Mimari Diyagramı

Bu tez çalışmasında kullanmış olunan Şekil 13'teki mimari yapının kullanımının benzer çalışmalarda faydalı ve pratik olacağından dolayı önerilmektedir.

5.1. Veri Normalizasyonu

Veri normalizasyonu, bilgisayar bilimlerinde ve özellikle makine öğrenimi uygulamalarında sıklıkla başvurulan bir ön işleme tekniğidir. Bu yöntemin temel amacı, veriler arasında önemli ölçüde farklılıklar bulunduğunda bu verileri standart bir formata getirerek işlemektir. Ayrıca, farklı ölçeklendirme sistemlerine sahip verilerin birbirleriyle uyumlu bir şekilde kıyaslanabilmesini sağlar. Veri normalizasyonu, çeşitli matematiksel fonksiyonlar aracılığıyla farklı ölçeklerdeki verileri aynı ölçeğe indirgeyerek, bu verilerin karşılaştırılabilir olmasını mümkün kılar [30]. Normalleştirme işlemi, amaç fonksiyonunun değerlendirilmesi sırasında her bir öznitelik boyutunun etkisinin eşitlenmesine yardımcı olur ve iteratif çözümleme süreçlerinde yakınsamanın daha hızlı gerçekleşmesine olanak tanır. [31]. Beş farklı normalizasyon yöntemi mevcuttur. Bunlar, min-maks, z-skor, ondalık(desimal) ölçekleme, logaritmik dönüşüm ve kütle çekirdekli normalizasyonu yöntemleridir.

5.1.1. Min-Maks Normalizasyonu

Min-maks normalizasyon yöntemi, veri setindeki değerlerin 0 ile 1 arasında yeniden ölçeklendirilmesi için kullanılır. Bu yöntemde, herhangi bir veri noktası, veri setindeki minimum ve maksimum değerler temel alınarak düzenlenir. Bu süreç, en düşük değerleri sıfıra, en yüksek değerleri ise bir değerine dönüştürmeyi amaçlar ve böylece tüm verileri 0 ile 1 arasındaki bir aralığa yerleştirir. İşlem, her bir giriş değerinin, veri setindeki en küçük değerden çıkarılması ve sonucun, maksimum ve minimum değerler arasındaki farka bölünmesi ile gerçekleştirilir. Yöntemin formülü eşitlik(3.1) ile gösterilmektedir. Bu yöntem, normalize edilen veriyi belirlemek için z, giriş değerini ifade etmek için x, veri setindeki en küçük değeri belirtmek için min(x) ve en büyük değeri belirtmek için maks(x) terimlerini kullanır. Bu işlem, veriler arası ilişkileri korurken, onları belirli bir standart aralığa sığdırmayı sağlar [32].

$$z = \frac{x - \min(x)}{maks(x) - \min(x)} \tag{3.1}$$

5.1.2. Z-Skor Normalizasyonu

Z-Skor Normalizasyonu veya Standartlaştırma, veriler üzerinde uygulandığında, bu verilerin ortalamasını sıfıra ve standart sapmasını bir değerine dönüştürme amacı güder. Bu işlem sırasında, her bir veri noktası, ilgili veri setinin ortalamasından çıkarılır ve elde edilen fark, veri setinin standart sapmasıyla bölünür. Bu yöntemin temel formülü, bir veri noktasının yeniden hesaplanmasını, bu veri noktasının orijinal değerinin, veri setinin ortalamasından çıkarılması ve ardından bu sonucun standart sapmaya bölünmesi şeklinde tanımlar.

5.1.3. Ondalık(Decimal) Ölçekleme Normalizasyonu

Ondalık(Desimal) Ölçekleme ise, verileri önceden belirlenmiş bir desimal noktasına göre yeniden boyutlandırma işlemidir. Bu süreçte, her bir veri noktası belirli bir sayıya (örneğin 10 veya 100) bölünerek ölçeklendirilir.

5.1.4. Logaritmik Dönüşüm Normalizasyonu

Logaritmik Dönüşüm, veri noktalarını logaritmik bir ölçeğe dönüştürür. Bu yöntem özellikle, verilerin geniş bir aralıkta dağıldığı veya dağılımın belirgin bir şekilde sağa veya sola çarpık olduğu durumlarda tercih edilir [33].

5.1.5. Kütle Çekirdekli Normalizasyonu((Robust Scaling))

Kütle Çekirdekli Normalizasyon ya da Robust Ölçekleme teknikleri, veri setlerinin medyanını ve çeyrekler arası mesafeyi temel alarak ölçeklendirme işlemi gerçekleştirir. Bu yöntem, özellikle aykırı değerlerin varlığında onların oluşturabileceği olumsuz etkileri minimize etmek adına tercih edilir, böylece veriler aykırı değerlere karşı daha dayanıklı bir şekilde ölçeklendirilmiş olur. Eşitlik (3.2) içerisinde tanımlanan bu metot, veri içerisinde aykırı değerlerin bulunması halinde özellikle avantajlıdır; zira bu yöntem, ortalama yerine medyanı ve genişlik yerine kartil aralığını kullanmaktadır. Kartil aralığı, birinci ve üçüncü kartiller arasındaki fark olarak ifade edilen IQR (Interquartile Range) ile hesaplanır ve verilerin ölçeklendirilmesi bu aralık dikkate alınarak yapılır [34].

$$x = \frac{xi - medyan}{p_{75} - p_{25}} \tag{3.2}$$

X’: ölçeklenen edilmiş öznitelik

$medyan$:Özniteliğin medyanı

$p75$: Özniteliğin 3. Kartil değeri

$p25$: Özniteliğin 1. Kartil değeri

Tablo 2. Normalizasyon yöntemlerinin Günlük Kazanç(DailyRate) sütunu üzerindeki uygulama sonuçlarının karşılaştırılması

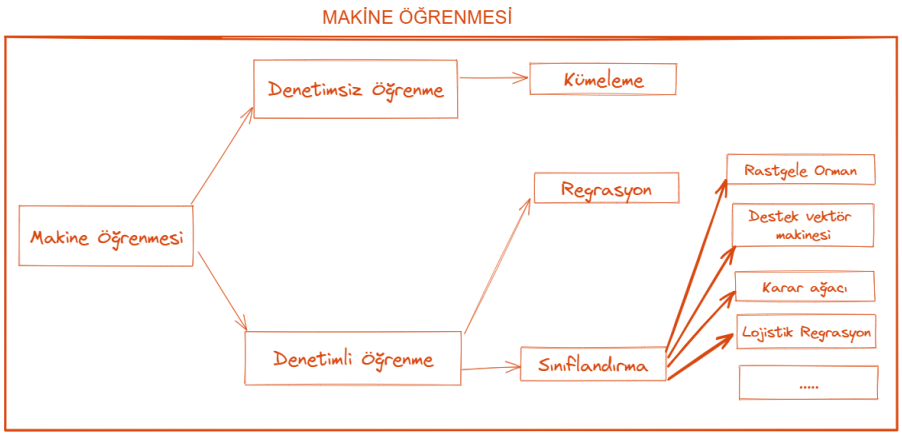
Sıra	Orjinal	Min-Maks	Z-Skoru	Ondalık Ölçekleme	Logaritmik Dönüşüm	Kütle Çekirdekli Ölçekleme
0	1102	0.716	0.743	0.735	7.006	0.434
1	279	0.127	-1.298	0.186	5.635	-0.756
2	1373	0.910	1.414	0.916	7.225	0.825
3	1392	0.923	1.461	0.929	7.239	0.853
4	591	0.350	-0.524	0.394	6.384	-0.305

Tablo 2’de gösterildiği gibi, yapılan karşılaştırma, farklı normalizasyon yöntemlerinin GünlükKazanç(DailyRate) değişkeni üzerindeki etkilerini açıkça ortaya koymaktadır. Her bir yöntemin, veri setinin özelliklerine ve makine öğrenimi modelinin gereksinimlerine bağlı olarak sunduğu farklı avantajlar bulunmaktadır. Bu bağlamda, model eğitimine başlamadan önce hangi normalizasyon yönteminin en uygun olduğunun belirlenmesi önem arz etmektedir. Tablo 2’deki değerler göz önünde bulundurularak, veri setimizdeki değerleri “0” ile “1” aralığında normalize formunu oluşturabilen model olan min-maks yöntemi seçilerek bu tez araştırmasında uygulanmıştır.

5.2. Geleneksel Makine Öğrenimi ve Yöntemleri

Geleneksel makine öğrenimi yöntemi, veri setlerinde yer alan örüntüleri ve ilişkileri saptamak amacıyla istatistiksel analiz tekniklerinden faydalanır. Bu yaklaşım, veriler üzerinde derinlemesine incelemeler yaparak, bunlar arasındaki bağlantıları ve düzenleri ortaya çıkarmayı hedefler. Elde edilen bu örüntüler, ileriye dönük tahminlerde bulunmak veya belirlenen bir işlevi gerçekleştirmek üzere kullanıma sunulur.

Geleneksel makine öğrenimi, Şekil 14’te gösterildiği gibi denetimli öğrenme ve denetimsiz öğrenme olmak üzere iki ana kategoriye ayrılır.

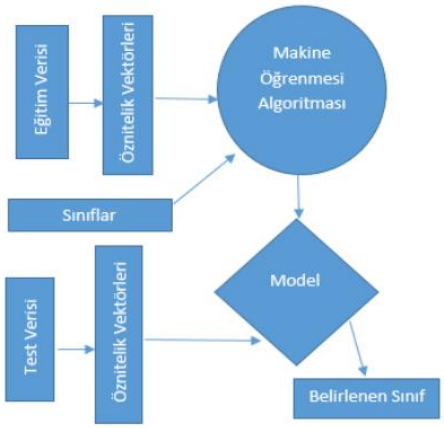


Şekil 14. Makine öğrenimi sınıflandırması

5.2.1. Denetimli öğrenme

Makine öğrenimi alanındaki araştırmaların önemli bir bölümü, denetimli öğrenme üzerine yoğunlaşmaktadır ve bu öğrenme yöntemi, multimedya içeriklerinin işlenmesi gibi çeşitli uygulama sahalarında kendine yer bulmaktadır. Denetimli öğrenmenin temelinde, önceden etiketlenmiş eğitim verilerinin bulunması yatar. Bu durum, eğitim süreci sırasında, öğrenme sistemine hangi örneklerin hangi etiketlerle ilişkilendirileceğine dair yönlendirme yapacak bir 'gözetmen' varlığına işaret eder. Çoğunlukla, bu etiketler sınıflandırma sorunlarında karşılaşılan sınıf etiketleridir. Denetimli öğrenme algoritmaları, bu eğitim verilerini kullanarak modeller geliştirir ve geliştirilen bu modeller, etiketlenmemiş yeni verilerin sınıflandırılmasında kullanılır [35].

Denetimli Makine Öğrenmesi, etiketlenmiş verilerin kullanılmasıyla sistemin eğitilmesi sürecidir. Bu süreçte, veri setindeki her bir örneğin giriş ve çıkış bilgileri sistem tarafından öğrenilir. Örneğin, metin sınıflandırma görevlerinde giriş, metnin kendisi olup, çıkış ise bu metnin ait olduğu kategori olarak tanımlanır. Sistemin performansının doğrulanması için test veri seti kullanılır ve bu aşamada, kategorisi önceden belirlenmemiş test verilerine, eğitim sürecinde öğrenilen çıkış değerlerinden biri atanır. Denetimli öğrenme süreci, Şekil 15'te sunulan şekilde yürütülür.



Şekil 15. Denetimli öğrenme modeli süreci [36]

Denetimli öğrenme yaklaşımında, sınıflandırma sorunları ele alınmakta ve eğitilen model, test verileri üzerinden tahminlerde bulunmak için kullanılmaktadır. Sınıflandırma ve tahmin işlemlerinde Destek Vektör Makinesi, Yapay Sinir Ağları, Lojistik Regresyon, Naive Bayes, Multinomial Naive Bayes, k-En Yakın Komşu, Rastgele Orman ve Karar Ağaçları gibi algoritmalar denetimli öğrenme modellerinin oluşturulmasında sıklıkla tercih edilen yöntemler arasında yer alır [36].

5.2.1.1. Lojistik regresyon

Lojistik regresyon, bağımlı değişkenin iki muhtemel sonuca (0 ve 1) sahip olduğu ikili sınıflama sorunlarında tercih edilen bir model olarak karşımıza çıkar. Makine öğrenmesinin ötesinde, uygulamalı bilimlerin çeşitli alanlarında ve gerçek dünya sorunlarının çözümünde sıklıkla başvurulan bir yöntemdir. İkili bir bağımlı değişken ile bir veya birden fazla bağımsız değişken arasındaki ilişkiyi modellemek için kullanılan lojistik regresyon, bir olayın meydana gelme ihtimalini tahmin etmeyi amaçlar. Bu model, belirli bir denklem çerçevesinde, bir olayın gerçekleşme şansını hesaplamak için eşitlik(3.3) kullanılır.

$$p = \frac{e^{a+bx}}{1+e^{a+bx}} \tag{3.3}$$

Olayın gerçekleşmeme olasılığı da ise $1 - p$ olmak üzere logit fonksiyonu eşitlik(3.4)'te verilmiştir:

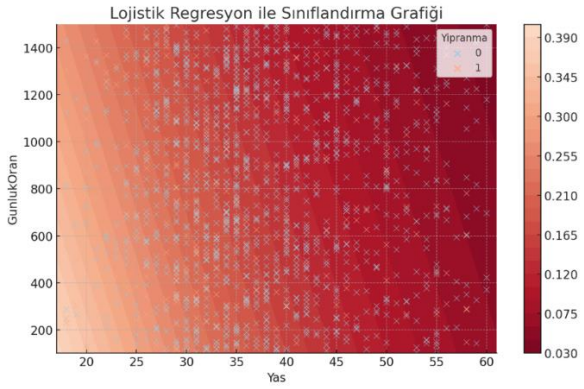
$$logit(p) = ln(\frac{p}{1-p})$$

(3.4)

Lojistik regresyon, logit dönüşümünün tahmin edilmesi sürecinde, bir denklemin katsayılarını belirlemek için kullanılan bir yöntemdir [37].

Lojistik regresyon modeli kullanılarak bir sınıflandırma grafiği oluşturulmuş ve bu grafik Şekil 15'te sunulmuştur. Söz konusu modelin görselleştirilmesi amacıyla, genellikle, karar sınırının (decision boundary) görselleştirilmesini kolaylaştırmak için iki boyutlu bir özellik uzayı tercih edilmektedir. Mevcut veri seti, çok sayıda özellik içermekte olduğundan, bu özellikler arasından yalnızca iki tanesi seçilerek bu özellikler temel alınarak bir sınıflandırma grafiği oluşturulmuştur.

İki özellik seçip, bu özellikler üzerinden lojistik regresyon modeli eğitip ve sonuçları Şekil 16'daki gibi görselleştirildi. Özellik seçimi, genellikle veriye özgüdür ve hangi özelliklerin en anlamlı olduğuna bağlıdır. Ancak burada, örnek olarak rastgele iki özellik seçilmiştir. Daha sonra bu özelliklerle bir karar sınırı grafiği çizilmiştir.



Şekil 16. Lojistik regresyon sınıflandırma grafiği

Bu grafikte: İki sayısal özellik (grafikteki x ve y eksenleri) seçildi. Mavi ve kırmızı renkler, modelin bu iki özellik temelinde 'Yipranma' değişkeninin 'Evet' veya 'Hayır' sınıflarını tahmin etme olasılığını göstermektedir. Mavi alanlar 'Hayır', kırmızı alanlar ise 'Evet' sınıfına daha yüksek olasılık göstermektedir. Noktalar, gerçek veri noktalarını temsil etmektedir. Farklı renkler farklı sınıflara ait gözlemleri göstermektedir. Bu görselleştirme, seçilen iki özellik ile 'Yipranma' arasındaki ilişkiyi ve modelin nasıl kararlar verdiğini göstermektedir.

5.2.1.2. Destek vektör makineleri (SVM)

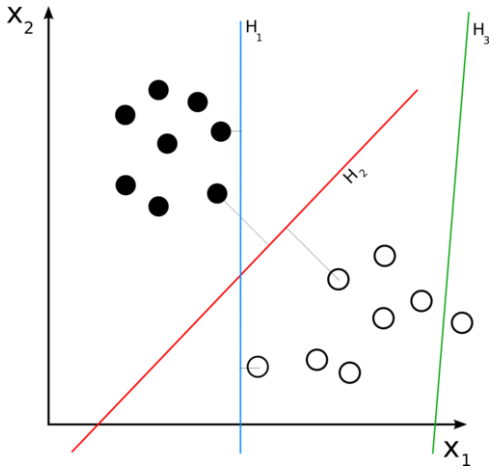
Destek Vektör Makineleri (SVM), 1990'larda Vapnik ve çalışma arkadaşları tarafından ortaya konulan, iki farklı sınıfı birbirinden ayırma amacı güden, istatistiksel öğrenme teorisine dayanan bir gözetimli makine öğrenmesi modelidir. Bu algoritma, sınıflandırma ve regresyon görevlerinde etkin bir şekilde kullanılabilmektedir. Eğitim verilerinin bulunduğu düzlem üzerinde, iki sınıf arasındaki en geniş marjı sağlayacak bir karar sınırı belirlenmesine imkan tanır.

Verinin her bir noktası eşitlik(3.5)'te verilen şekilde tanımlanır.

$$\{(x_i, y_i)|xi \in R^d, y_i \ \in \{-1,1\}\}_{i-1}^{n} \tag{3.5}$$

Formülde x bir girdiyi, y ise -1 ve 1 ile temsil edilen bir sınıfı belirtir. Düzlemde her bir nokta wx-b şeklinde ifade edilir. Burada w düzleme dik olan normal vektörü, b ise kayma miktarıdır. Destek vektör makinesi, karesel optimizasyon yöntemi ile ayırma sınırının bulunmasını sağlar.

Şekil 17'deki görselde, iki farklı sınıfa ait veriler ve bu verileri ayıran karar sınırı görülmektedir.

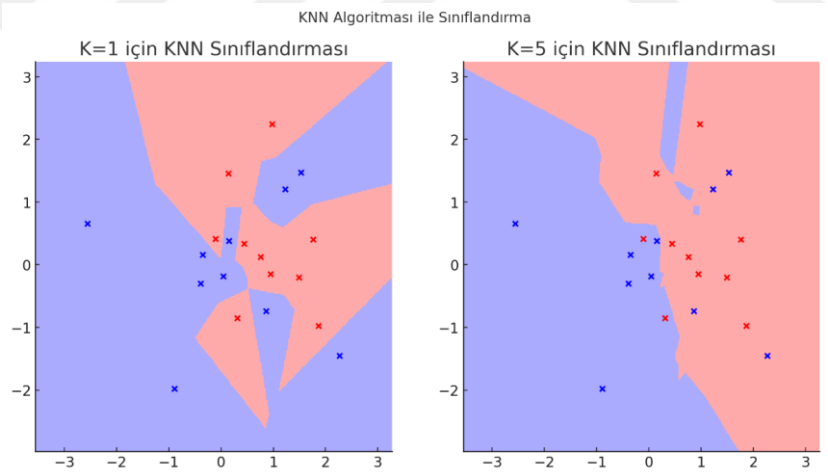


Şekil 17. Destek vektör makinesi [37]

Ayrıca, destek vektörler de belirtilmiştir. SVM, bu destek vektörleri üzerinden en iyi karar sınırını ve marjı bulur, böylece yeni verileri sınıflandırmak için kullanılır. Bu algoritma, eğitim verilerindeki sınıfları ayıran bir sınır (hyperplane) bulmaya çalışır. SVM, karmaşık veri kümelerinde iyi çalışan bir algoritmadır. Örneğin, bir görüntüdeki nesneleri sınıflandırmak için kullanılabilir [37].

5.2.1.3. K-En yakın komşu (K-Nearest Neighbors: KNN)

K-En yakın komşu(K-Nearest Neighbors: KNN) algoritması, sınıflandırma ve regresyon problemleri için kullanılan basit ve etkili bir makine öğrenmesi yöntemidir. Bir model, sınıflandırılacak olan bir noktanın, önceden belirlenmiş sınıflara ait noktalar içinden, seçilen bir K miktarına kadar en yakın olanlarına dayalı olarak hangi sınıfa ait olduğunun tespit edilmesine imkân tanır. Bu yaklaşımda, en yakın noktaların belirlenmesi sürecinde çoğunlukla Öklid uzaklığı kullanılır. Seçilecek olan ideal K değeri, üzerinde çalışılan veri setinin özelliklerine göre değişkenlik gösterebilir. Yüksek K değerlerinin kullanılması, sınıflandırma sürecindeki gürültünün azaltılmasına yardımcı olurken, diğer yandan sınıflar arası sınırların daha az keskin olmasına yol açar. Bu algoritmanın temel uygulama adımı, veri setinin öncelikle eğitim ve test olmak üzere iki ayrı kümeye ayrılması üzerine kuruludur. Test veri kümesinden seçilen bir örneğin sınıflandırılması sürecinde, bu örneğin nitelik uzayındaki konumu dikkate alınarak, eğitim veri kümesinde yer alan diğer örneklerle arasındaki mesafeler tek tek hesaplanır. Bu süreçte, yeni sınıflandırılacak gözlem ile eğitim setindeki mevcut gözlemler arasındaki yakınlık, belirlenen bir k sayısına kadar en yakın gözlemlere göre değerlendirilir [38]. Şekil 20'de, K-en yakın komşu (KNN) algoritmasının iki farklı 'K' değeri için uygulanmış hali gösterilmiştir.



Şekil 18. KNN Algoritması ile sınıflandırma

Bu grafikler, algoritmanın nasıl çalıştığını ve K'nın değerinin sonuçlara nasıl etki ettiğini gösteriyor. Grafikteki iki alt grafik, K'nın farklı değerleri (1 ve 5) için KNN algoritmasının sonuçlarını göstermektedir. Her iki durumda da, iki farklı sınıf mavi ve kırmızı renklerle temsil edilmiştir. K = 1 (Sol grafik): Bu durumda, her veri noktası yalnızca en yakın komşusunun sınıfına göre sınıflandırılır. Bu, çok detaylı ve gürültülü bir sınırlara yol açar ve genellikle aşırı uyuma (overfitting) örnek olabilir. K = 5 (Sağ grafik): K'nın değeri

arttığında, sınıflandırma için daha fazla komşu dikkate alınır. Bu, karar sınırlarını daha düzgün ve genelleştirilmiş hale getirir. Ancak, K çok büyük olursa, modelin basitleşmesi ve alt uyuma (underfitting) riski ortaya çıkabilir [39].

5.2.1.4. Karar Ağaçları

Karar ağacı, her bir düğümün bir nitelik (özellik) ifade ettiği, her bir bağlantının (dal) ise bir karara işaret ettiği ve her bir yaprak düğümünün bir sonucu temsil ettiği hiyerarşik bir yapıdır. Bu yapı içerisinde, ağacın her bir düğümü, sınıflandırılmak üzere olan bir grubun özelliklerini yansıtırken, dallar ise bu düğümlerin alabileceği değerlerin çeşitliliğini gösterir. [40].

Literatürde çeşitli karar ağacı algoritmaları mevcuttur ve bu algoritmalar, temel aldıkları ağaç oluşturma prensiplerine bağlı olarak farklı veri setlerinde değişken sınıflandırma performansları sergileyebilirler. ID3, J48, C4.5 ve C5, karar ağacı algoritması alanında sıkça rastlanan ve geniş çapta tanınan yöntemler arasındadır. Özellikle C4.5 algoritması, sınıflandırma sürecinde en belirgin özelliğin belirlenmesi aşamasında entropi kavramını temel alır. Entropi, veri setindeki belirsizlikleri ve rasgelelik düzeylerini değerlendirmek için kullanılan bir ölçümdür. $p1$,$p2$,…,pn gibi olasılık durumları için, bu olasılıkların toplamının 1'e eşit olması gerektiği varsayılır. Bu prensip altında, entropi hesaplaması belirli bir formülasyon ile eşitlik(3.6) gibi gerçekleştirilir.

$$I(P) = -\sum_{i=1}^{k} pi * \log(pi) \tag{3.6}$$

Veri tabanında bulunan tüm özniteliklerin entropileri hesaplanmasının ardından, her bir özniteliğin Bilgi(D,S) değeri eşitlik(3.7)'de verildiği gibi hesaplanır.

$$\text{Bilgi}(\text{D},\text{S}) = \sum_{i}^{n} \frac{Ti}{T} l(T_i) \tag{3.7}$$

Elemanların bilgi değerleri hesaplandıktan sonra, elde edilen kazanç değerleri, eşitlik(3.8)'deki gibi belirli bir formülasyona göre hesaplanır. Bu hesaplama sonucunda en yüksek kazanç değerine sahip eleman, karar ağacının en üstünde yer alan kök düğüm olarak belirlenir.

Bilgi değerleri hesaplanan elemanların kazanç değerleri eşitlik(3.8)'de verildiği üzere hesaplanır ve kazanç değeri maksimum olan eleman en üstteki kök düğümüne yerleştirilir.

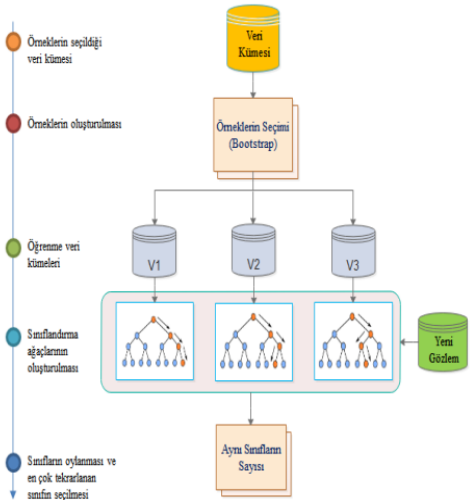
$$\text{Kazanç}(\text{D},\text{S}) = \text{Bilgi}(\text{S}) - \text{Bilgi}(\text{D},\text{S})$$

(3.8)

Literatürde yer alan diğer karar ağacı algoritmaları da, muhtelif veri setleri üzerinde yüksek derecede sınıflandırma başarısı sergilemektedir. Örneğin, J48 algoritması, ağaç yapısını oluştururken, sınıflandırma performansını yükseltmeyi hedefleyerek, performansı düşük dalları budama esasına dayanır. Aynı zamanda, karar ağaçları temel alınarak geliştirilen hibrit yöntemlerin, sınıflandırma başarısını iyileştirdiğine dair bulgular da literatürde mevcuttur. Karar ağacı algoritmaları, analiz ve yorumlama kolaylığı sunması ve birden fazla sonucu olan problemleri çözmede etkili olması nedeniyle tercih edilmektedir. Ancak, bu algoritmaların zaman zaman karmaşık dallanmalara yol açabileceği ve aşırı uyum (overfitting) gibi sorunlara sebebiyet verebileceği göz önünde bulundurulmalıdır [41].

5.2.1.5. Rastgele Orman Sınıflandırması Algoritması

Rastgele orman regresyon modelinde, ağaçların sayısının artırılmasıyla genellikle modelin performansının yükselmesi beklenir; ancak ağaç sayısının artışının veri setinin niteliği ve büyüklüğüne bağlı olarak her durumda performansı artıracağının garantisi yoktur. Bu sebepten ötürü, çeşitli ağaç sayılarıyla modellerin performanslarının karşılaştırılması önerilir. Modelin dezavantajlarına gelince, tek bir karar ağacının sunduğu gibi, sonuçların ağaç yapısı üzerinden görsel bir sunum sağlanamaması ve modelin karmaşıklığı sebebiyle çoklu karar ağaçlarının değerlendirme süreçlerinin gözlemlenememesi sayılabilir. Şekil 19'da Rastgele Orman Algoritmasının işlem adımları sunulmaktadır [42].



Şekil 19. Rastgele orman algoritmasının akış şeması [42]

5.2.1.6. Naive Bayes sınıflandırma algoritması

Naive Bayes sınıflandırıcısı, olasılık tabanlı sınıflandırma metotlarının en sınırlı spektrumunda yer alır ve belirlenmiş sınıflar ile örneklerin bu sınıflara nasıl dağıtıldığı önceden bilinir durumdadır. Özellikle metin sınıflandırması alanında etkinliği ispatlanmış bir yöntemdir. Bayes teoremine dayanan bu yaklaşımda, n boyutlu bir uzayda tanımlanmış X vektörü için, m sayıda sınıfın bulunduğu bir veri setinde, son olasılığı en yüksek sınıfa ait etiketi bulmayı amaçlar. Bu yöntem, eşitlik(3.9)'de gibi bir formülasyon ile ifade edilir. [43].

$$P(C_i \mid \mathbf{X}) = P(\mathbf{X} \mid C_i) P(C_i) \tag{3.9}$$

5.2.1.7. XGBoost sınıflandırma algoritması

2015 yılında Chen ve He tarafından geliştirilen XGBoost (eXtreme Gradient Boosting), bir dizi karar ağacını temel alan ve derinlikleri farklı ağaçların oluşturulması ile çeşitli algoritmaların optimizasyonunu gerçekleştirebilen bir topluluk öğrenme yöntemidir. Bu algoritma, regresyon ve sınıflandırma gibi problemlerin çözümünde yüksek düzeyde verimlilik ve etkinlik sunmaktadır.

Aşırı Gradyan Artırma olarak da bilinen XGBoost algoritması, özellikle karar ağaçları ve makine öğrenmesi alanlarında sıkça kullanılan bir yöntem olup, sınıflandırma, regresyon ve sıralama gibi farklı görevlerde üstün performans sergileyen bir denetimli öğrenme aracı olarak öne çıkmaktadır. Denetimli öğrenme, etiketlenmiş eğitim verileri kullanılarak tahminleyici bir model oluşturma sürecini ifade eder ve bu modelle, üretim hatalarının tespiti veya belirli bir gün için sıcaklık ve nem tahminleri gibi problemlere çözümler üretilebilir. XGBoost, artırma yöntemini kullanarak etkili modeller geliştirir ve bu modeller, hem regresyon hem de sınıflandırma görevlerinde kullanılabilir. Bu algoritmanın temelinde, amaç fonksiyonunun optimizasyonu yatar ve ölçeklenebilirlik açısından önemli bir avantaj sunar. Tek bir makinede, mevcut popüler yöntemlerden on kat daha hızlı çalışma kapasitesine sahip olan sistem, aynı zamanda milyarlarca örneğe kadar ölçeklenebilir şekilde dağıtılmış veya hafıza sınırlı ortamlarda da etkin bir performans sergileme yeteneğine sahiptir [44].

5.2.1.8. AdaBoost Sınıflandırma Algoritması

Kolektif öğrenme, tekil modellerin ve bu modellerin kararlarının bir araya getirilmesiyle meydana gelen öğrenci topluluğu kavramını ifade eder. Bu yaklaşım, genellikle, bireysel öğrenme yöntemlerine kıyasla sınıflandırma performansında bir artış sağlar. AdaBoost, bu alanda sıklıkla tercih edilen boosting algoritmalarından biridir ve ilk kez Freund ve Schapire tarafından sunulmuştur [45]. Diğer kolektif öğrenme yöntemlerine kıyasla, tahmin hızının yüksek olması, daha az hafıza ihtiyacı duyması ve uygulanabilirliği gibi avantajlar sebebiyle AdaBoost algoritması tercih edilmektedir. Bu çalışmada ele alınan veri seti, örnek tabanlı bireysel öğreniciler için uygun olduğundan, ardışık topluluk öğrenme metodu olan Boosting kullanılmıştır. Boosting metodu, öğrenicilerin eğitiminde, önceki temel öğrenicilerin hatalı bulduğu örnekleri önceliklendirmek çalışır. Bagging yöntemiyle karşılaştırıldığında, Bagging'de her iterasyonda tüm örnekler eşit şansa sahipken, Boosting'de her iterasyonda örneklerin seçilme olasılıkları güncellenir. Bu, yöntemin yanlış kararlar üzerine yoğunlaşmasını sağlayarak özellikle hatalı kararlara odaklanılmasını teşvik eder.

AdaBoost algoritması, her bir özelliğe dayanarak oluşturulan zayıf sınıflandırıcıları birleştirerek işlev görür. Bu zayıf sınıflandırıcılar, her özelliğin negatif ve pozitif örnekler için ağırlıklı ortalamalarının alınması ve belirlenen karar sınırlarına göre oluşturulmasıyla tanımlanır. İlerleyen aşamalarda, en düşük hata oranlarına sahip zayıf sınıflandırıcılar seçilerek daha kuvvetli bir sınıflandırıcı oluşturulur. Bu güçlü sınıflandırıcının parçası olmayan zayıf sınıflandırıcıların özellikleri bu aşamada elenir. Algoritmanın temel yapısını açıklayan pseudo kod ise belirli adımlar çerçevesinde geliştirilmiştir [46].

Verilen: Eğitim veri seti (X, y), burada X eğitim örnekleri, y etiketler

Parametreler: Zayıf öğrenici türü, öğrenici sayısı T

Başlangıç:

1. Her eğitim örneğine eşit ağırlık ata,

$$w_i = \frac{1}{N} \;\; i = 1,2, \dots, N \tag{3.10}$$

2. T kadar iterasyon için:

Her Iterasyon için:

a) Zayıf öğreniciyi eğitim veri seti üzerinde eğit (ağırlıkları dikkate alarak)

b) Hata değerini hesapla:

$$err = \frac{\text{err } = \text{ Sum(w\_i } * \text{ I(y\_i ! = y\_pred\_i))}}{\text{Sum(w\_i)}} \tag{3.11}$$

Burada I koşulun doğruluğuna göre 1 veya 0 değerini almaktadır.

c) Öğrenici ağırlığını hesapla:

$$\text{alpha} = 0.5 * \log((1 - \text{err}) / \text{err}) \tag{3.12}$$

d) Güncelle ağırlıkları:

$$w_i = w_i * \exp(\text{alpha } * \text{ I(y\_i ! = y\_pred\_i))} \tag{3.13}$$

Doğru sınıflandırılanlar için ağırlık azalır, yanlışlar için artar.

e) Ağırlıkları normalize et.

Sonuç:

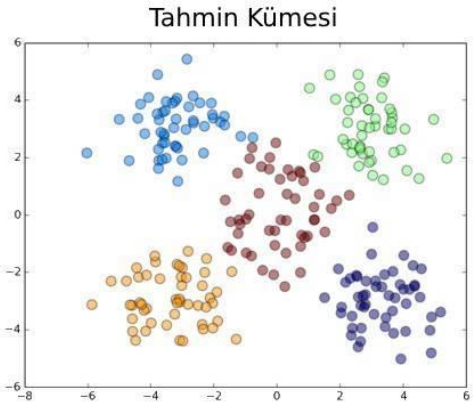
- Tüm öğrenicilerin ve ağırlıkların bir kombinasyonu ile bir son model oluşturur.

- Son model, öğrenici ağırlıklarıyla ağırlıklı oylama yaparak sınıflandırma gerçekleştirir.

5.2.2. Denetimsiz Öğrenme

Denetimsiz makine öğrenmesi, veri setlerinin içerisindeki önceden tanımlanmamış yapıları veya desenleri otomatik olarak keşfetmeye yönelik bir yaklaşımdır. Bu süreçte, makine öğrenmesi algoritmaları herhangi bir insan müdahalesi olmaksızın, veriler arasındaki ilişkileri, benzerlikleri veya farklılıkları belirleyebilir. Özellikle, veri setlerindeki gizli grupları (kümeleri) bulmak, verilerin boyutunu azaltmak ve veri setinin daha derinlemesine analizini sağlamak için kullanılır. Yönetimsiz öğrenmenin en bilinen uygulamalarından biri kümeleme algoritmalarıdır. Bu algoritmalar, benzer özelliklere sahip veri noktalarını tespit ederek, bu noktaları ortak gruplara yerleştirir. Örneğin, müşteri davranış verileri üzerinden yapılan bir analizde, benzer alışveriş eğilimlerine sahip müşteri segmentleri tespit edilebilir. Bu segmentasyon, pazarlama stratejilerinin daha hedefli ve etkili bir şekilde planlanmasına olanak tanır. Bir diğer önemli uygulama ise boyut indirgemedir. Yüksek boyutlu verilerin,

daha az boyutlu bir temsile indirgenmesi işlemi, verilerin daha kolay işlenmesini ve anlaşılmasını sağlar. Bu sayede, makine öğrenmesi modellerinin eğitim süreçleri hızlanır ve veri görselleştirme teknikleri daha etkili bir şekilde kullanılabilir [47].

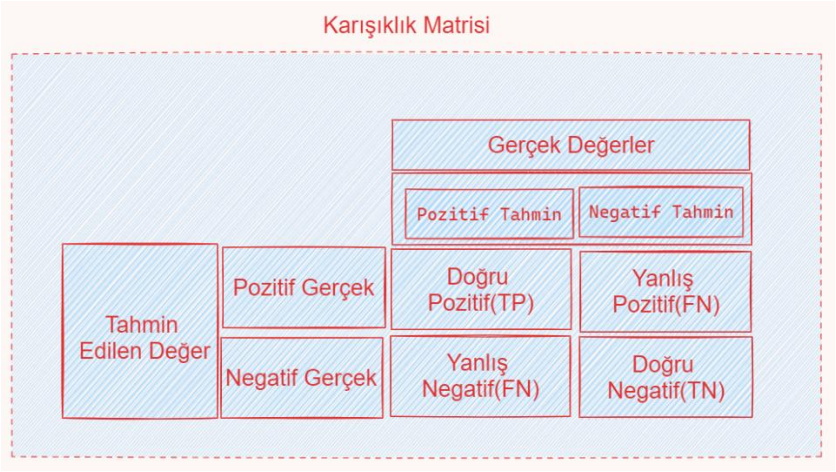


Şekil 20. Denetimsiz öğrenme kümeleme modeli veri dağılımı [47]

Denetimsiz makine öğrenmesi yöntemlerinden en bilinenlerden biri olan k-ortalama kümeleme algoritması, verilerin yakınlıkları, uzaklıkları ve benzerlikleri gibi önemli ölçütlere göre analiz edilerek sınıflara Şekil 20'deki gibi ayrılmasıdır.

5.2.3. Model Performans Değerlendirme Ölçütleri

Makine öğrenmesinde, sınıflandırma modellerin performansını ölçmede ve değerlendirmede hedeflenen tahmin değeler ile gerçek değerlerin karşılaştırıldığı, Şekil 21'deki Karışıklık(Hata) Matrisi kullanılır.



Şekil 21. Karışıklık matrisi

Doğruluk (Accuracy), tüm sınıflardan (olumlu ve olumsuz), kaç tanesini doğru tahmin ettik [48].

Doğruluk(Accuracy), sınıflandırma görevlerinde yaygın olarak başvurulan bir değerlendirme ölçütüdür ve doğru bir şekilde sınıflandırılan örneklerin, toplam örnek sayısına bölünmesiyle elde edilir.

$$Doğruluk = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3.14}$$

Kesinlik (Precision), bir sınıflandırma modelinin sonuçlarının ne derece güvenilir olduğunu ifade eden bir metriktir. Bu, pozitif olarak işaretlenmiş olan örneklerin, pozitif olarak sınıflandırılan tüm örnekler içindeki oranıyla hesaplanır.

$$Kesinlik = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \tag{3.15}$$

Duyarlılık(Recall), pozitif olarak işaretlenmiş örneklerin, aslında pozitif olan tüm örnekler içerisindeki payını gösterir.

$$Duyarlılık = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{3.16}$$

F-Skor(F-Measure), kesinlik ve duyarlılık değerlerinin birleştirilmesiyle elde edilen bir ölçümdür. Bu metrik, bir sistemin kesinlik veya duyarlılık açısından ne kadar iyi optimize edildiğini değerlendirmede kullanılır.

$$F = \frac{2 * Duyarlılık * Kesinlik}{Duyarlılık * Kesinlik} \tag{3.17}$$

Eşitliklerde kullanılan TP, FP, TN, FN değerleri sırasıyla;

TP (True Positive Rate): Pozitif olan aynı zamanda sınıflandırıcı tarafından da pozitif olarak sınıflandırılmış yorumların sayısını gösterir.

FP (False Positive Rate): Pozitif olan ancak sınıflandırıcı tarafından pozitif olarak sınıflandırılmamış yorumların sayısını gösterir.

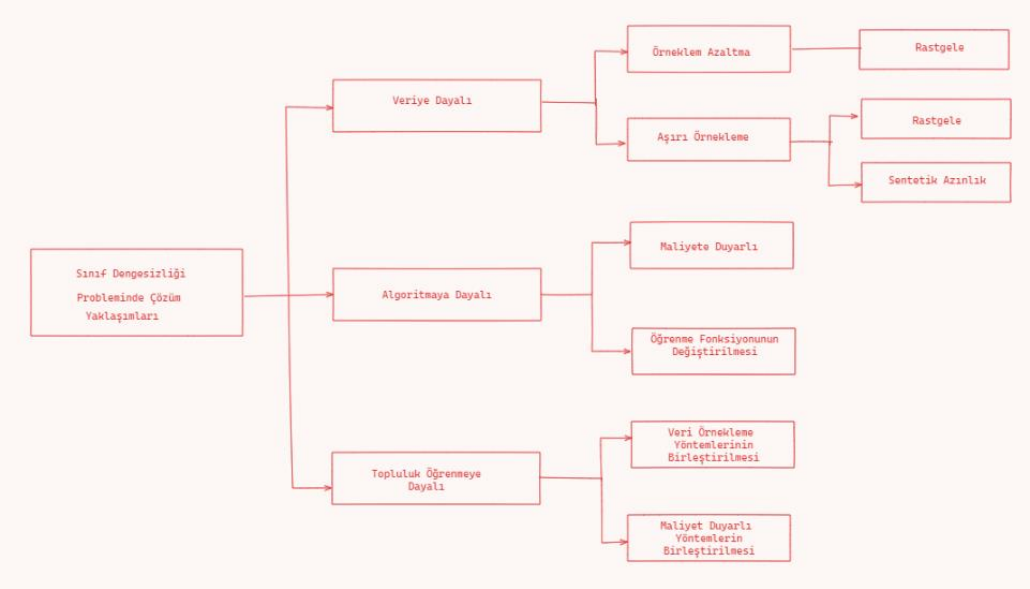
TN (True Nagative Rate): Negatif olan aynı zamanda sınıflandırıcı tarafından da negatif olarak sınıflandırılmış yorumların sayısını gösterir.

FN (False Nagative Rate): Negatif olan ancak sınıflandırıcı tarafından negatif olarak sınıflandırılmamış yorumların sayısını gösterir [49].

5.2.4. Dengesiz Veri Yöntemleri

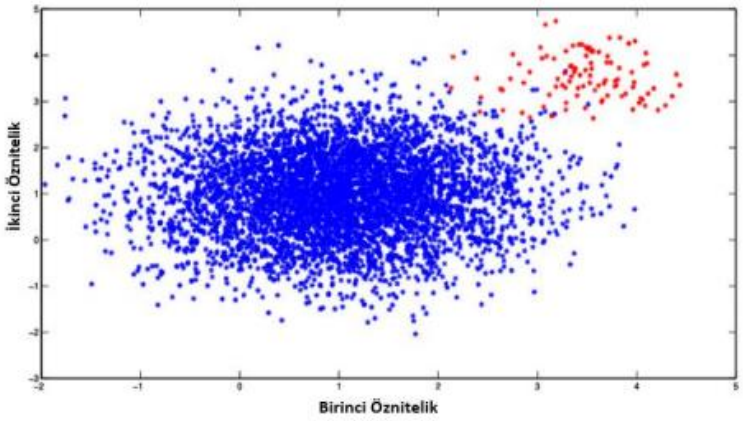
Herhangi bir veri kümesi, teknik açıdan incelendiğinde, sınıflar arasında eşit olmayan bir dağılım sergiliyorsa, bu durum dengesiz kabul edilir [50]. Literatürde yapılan çalışmalar incelediğinde; bankacılık (kredi kartı dolandırıcılığı), tıp (hastalık teşhisi), telekomünikasyon (kayıp müşteri analizi) gibi birçok alanda sınıf dengesizliğine sahip veri kümeleri ile karşılaşılmaktadır.

Sınıf dengesizliğine bir örnek vermek gerekirse, elimizde ikili sınıf değişkenine (hasta / hasta değil) sahip; 980 kişi hasta, 20 kişi hasta değil olmak üzere 1000 adetlik bir sağlık veri seti olsun. Oran açısından bakıldığında hasta kişilerin sayısı, hasta olmayan kişilerin sayısına göre bariz bir üstünlük kurduğu görülmektedir. Veri setleri içerisinde sınıflardan birinin çok sayıda, diğer sınıf az sayıda veri ile temsil ediliyor ise sınıf dengesizliğinden bahsetmek mümkündür. Sınıf dengesizliği ise, sınıflandırma modellerinin doğru sınıflandırma performansını olumsuz yönde etkileyen bir durumdur [51]. Sınıf dengesizliğinin çözümü için literatürde kullanılan yöntemleri; veri seviyelerine dayalı, algoritmalara dayalı ve topluluk öğrenmesine dayalı olarak üç başlıkta incelemek mümkündür. Şekil 22'de sınıf dengesizliğinin çözümü için kullanılan yöntemlerin bir görseli verilmiştir [52].



Şekil 22. Sınıf dengesizliği problemi çözüm yaklaşımları

Şekil 23'de sınıf dengesizliği probleminin bir örneği gösterilmiştir [53]. Mavi renkli sınıf değerinin kırmızı renkli sınıf değerinden belirgin şekilde fazla olduğu gözükmektedir. Bu da veri setinde bir sınıf dengesizliği olduğunun bir göstergesidir.



Şekil 23. Sınıf dengesizliği problemi örneği [53]

Literatürde sınıf dengesizliğini giderebilmek adına yeniden örnekleme yöntemleri sıklıkla kullanılan teknikler arasındadır. Bu yöntemler içinde dengesiz veri durumunun olumsuz etkisini azaltmak için örneklem azaltma (undersampling), aşırı örnekleme (oversampling) ve sentetik azınlık aşırı örnekleme (SMOTE) yöntemleri ön plana çıkmaktadır.

Aşağıdaki gibi gruplama yapabiliriz.

- (Yeniden Örnekleme)Resampling
  - Yukarı Örnekleme(Oversampling)
    - Rastgele Aşırı Örnekleme Tekniği (ROS)
    - Sentetik azınlık aşırı örnekleme (SMOTE)
  - Aşağı Örnekleme(Undersampling)
    - Rastgele Örneklem Azaltma Tekniği (RUS)
    - NearMiss aşağı örnekleme
    - Tomek links aşağı örnekleme

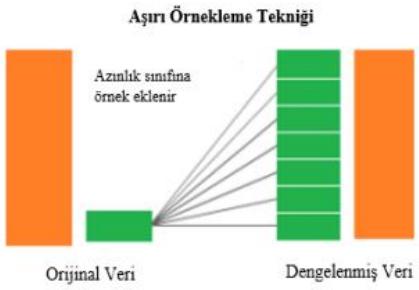
Her yöntemin kendine özgü avantajları ve dezavantajları bulunmaktadır. Örneğin, örneklem azaltma yönteminde, çoğunluk sınıfına ait verilerin sayısının azaltılmasıyla birlikte, bilgi kaybının meydana gelmesi olası bir sonuçtur. Aşırı örnekleme yönteminde ise, azınlık sınıfına ait verilerin sayısının artırılması, modelin aşırı uyum problemiyle karşı karşıya kalmasına neden olabilmektedir. Eğitim süreleri ve bellek kullanımı açısından

değerlendirildiğinde, örneklem azaltma yöntemiyle veri setinin boyutunun küçültülmesi, modelin eğitim süresinin önemli ölçüde kısalmasına ve daha az bellek kullanımına imkân tanır. Aşırı örnekleme yöntemindeyse, veri sayısının artırılmasıyla eğitim sürelerinin uzaması ve daha yüksek bellek kullanımı kaçınılmazdır [54].

Bu tezde, rastgele aşırı örnekleme (Random Over-Sampling Technique: ROS), rastgele örneklem azaltma (Random Under-Sampling: RUS), sentetik azınlık aşırı örnekleme tekniği(Synthetic Minority Over-sampling Technique: SMOTE), Near Miss ve Tomek links yöntemleri kullanılmıştır.

5.2.4.1. Rastgele Aşırı Örnekleme Tekniği (Random Over-Sampling Technique: ROS)

Rastgele aşırı örnekleme yöntemi, basit fakat etkili bir yeniden örnekleme stratejisi olarak kabul edilmektedir. Bu yaklaşımda, azınlık sınıfına ait öğeler, rastgele bir şekilde seçilmekte ve seçilen bu öğelerin çoğaltılması sonucunda yeni bir eğitim setine dâhil edilmektedir. Rastgele aşırı örnekleme işlemi gerçekleştirilirken, bir hususun göz önünde bulundurulması gerekmektedir: Yeniden örnekleme işleminde, öğelerin yeni eğitim setinden değil, orijinal veri setinden rastgele seçilmesi esastır. Bu, seçimin rastgeleliğinin korunması için önem taşımaktadır; zira aksi takdirde, seçim sürecinde rastgeleliğin bozulması riski ortaya çıkacaktır [55]. Rastgele aşırı örneklem tekniği(ROS) Şekil 24'teki gibi bir dağılım göstermektedir.



Şekil 24. Rastgele aşırı örnekleme örneği [56]

5.2.4.2. Rastgele Örneklem Azaltma Tekniği (Random Under-Sampling: RUS)

Rastgele örneklem azaltma, eğitim setlerindeki dengesizlikleri gidermek için basit bir yöntem olarak ele alınır. Bu süreçte, çoğunluk sınıfına ait veriler, azınlık sınıfı ile

aralarındaki oran kabul edilebilir bir seviyeye ulaşana kadar rastgele şekilde çıkarılır. Teorik açıdan, rastgele örneklem azaltmanın karşılaştığı zorluklardan biri, çıkarılan çoğunluk sınıfı verilerinin hangi bilgileri içerdiğinin kontrol edilememesidir. Bu durum, özellikle azınlık ve çoğunluk sınıfları arasındaki karar sınırlarına dair kritik bilgilerin kaybedilmesine yol açabilir. Basit bir yaklaşım olmasına karşın, rastgele örneklem azaltmanın deneysel çalışmalarda en etkili yeniden örnekleme yöntemlerinden biri olarak ortaya çıktığı belirlenmiştir. Şekil 25'te örneklem azaltma tekniğinin görsel olarak verilmiştir [56].

Bu bağlamda, iki farklı sınıf bulunmaktadır: yeşil ve turuncu renklerde temsil edilenler. Örneklem azaltma tekniği uygulandığında, turuncu renkli sınıfa ait verilerin (çoğunluk sınıfı), yeşil renkli sınıfa ait verilerle (azınlık sınıfı) sayısal olarak eşitlenmesi amacıyla azaltılması gerekmektedir. Bu işlem, sınıflar arasındaki dengenin sağlanması için gerçekleştirilir.

Şekil 25. Rastgele örneklem azaltma örneği [56]

5.2.4.3. Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE)

Sentetik Azınlık Örneklem Arttırma Tekniği (SMOTE), 2002 yılında Chawla ve arkadaşları tarafından önerilen bir veri önişleme yöntemidir. Bu teknik, azınlık sınıfı için yerine koyma yöntemi veya sadece mevcut örneklerin tekrarlanması yerine, 'sentetik' örnekler oluşturarak aşırı örnekleme yapar. Böylece, azınlık sınıfının örnek sayısı, sentetik verilerle artırılarak denge sağlanmış olur [57].

SMOTE (Sentetik Azınlık Örneklem Arttırma Tekniği), dengesiz sınıflandırmada önemli bir rol oynayan öncü bir veri ön işleme tekniğidir. Bu teknik literatürde tanıtıldığından beri,

çeşitli senaryolarda performansını iyileştirmek amacıyla birçok uzantısı ve alternatifi önerilmiştir. Yüksek popülaritesi ve büyük etkisi nedeniyle, SMOTE makine öğrenimi ve veri madenciliği alanlarında en etkili veri ön işleme ve örnekleme algoritmalarından biri olarak kabul edilmektedir [58].

Bu yöntemin temelinde, k-En Yakın Komşu (k-EYK) algoritması kullanılır. Yöntem, rastgele seçilen bir azınlık sınıfı örneği için kullanıcının belirlediği sayıda en yakın komşuyu inceler. Algoritma, seçilen örnek ile diğer azınlık sınıfı örnekleri arasında en yakın noktayı belirler ve ardından bu iki nokta arasında interpolasyon yaparak yeni bir sentetik veri üretir. Eğer üretilecek sentetik veri sayısı orijinal veri kümesi sayısından azsa, orijinal veriler kullanılır. Ancak üretilecek sentetik veri sayısı orijinal veri kümesi sayısını aşıyorsa, algoritma belirlenen aşırı örnekleme oranına uygun olarak yinelemeli bir şekilde sentetik veriler üretir [59].

SMOTE algoritmasının adımları şu şekilde ifade edilebilir [60] :

Adım 1: Azınlık sınıfındaki her bir verinin (xi) k en yakın komşusuna bakılır,

Adım 2: Azınlık sınıfındaki veri (xi) ile k en yakın komşusundaki gözlemin (xj) farkı alınır,

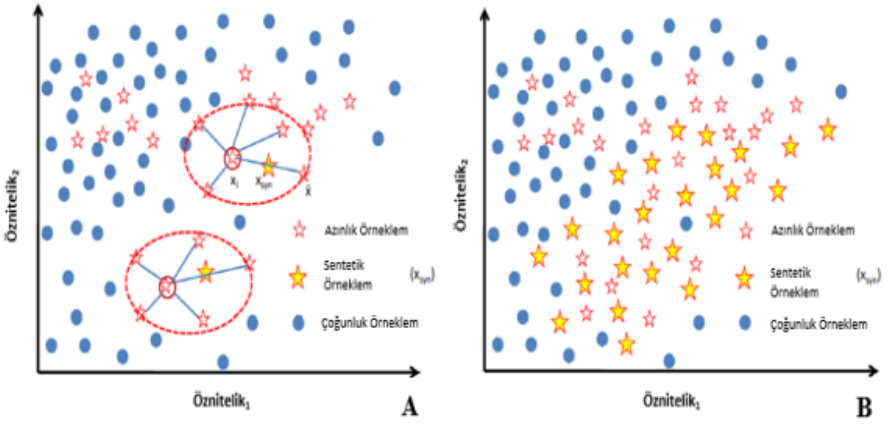
Adım 3: 0 ile 1 arasında rastgele belirlenen bir sayı ile Adım 2'deki değer çarpılır,

Adım 4: Eşitlik(3.14)'deki eşitlik sayesinde yeni sentetik veriler ($xsyn$) üretilir.

$$x_{syn=x_i+(x_j-x_i)*rand(0,1)} \tag{3.18}$$

Adım 5: Belirlenen sayıda veri elde edilebilmesi için Adım 1-4 tekrarlanır.

SMOTE yönteminde k=5 için sentetik veri üretiminin bir örneği Şekil 26 (A) ve Şekil 26 (B)'de gösterilmiştir [61].



Şekil 26. SMOTE, Rastgele seçilen bir azınlık örneğinin dağılımı [61]

5.2.4.4. Near Miss Aşağı Örnekleme Yöntemi

Near Miss örnekleme yöntemi, çoğunluk sınıfı örneklerini azaltarak sınıf dengesizliğini gidermeyi amaçlar. Bu yöntemde, çoğunluk ve azınlık sınıfları arasındaki Öklid uzaklığına göre örnekler seçilir. Üç farklı Near Miss yöntemi vardır: Birinci yöntemde, azınlık sınıfına en yakın çoğunluk sınıfı örnekleri seçilir. İkinci yöntem, azınlık sınıfının en uzak örneklerine en yakın olan çoğunluk sınıfı örneklerini seçer. Üçüncü yöntem iki adımdan oluşur; ilk olarak azınlık sınıfının her örneği için en yakın komşular belirlenir, sonra çoğunluk sınıfından bu komşulara en uzak olan örnekler seçilir. Bu yöntemler, sınıf dengesizliğini etkili bir şekilde azaltmaya yardımcı olur.

5.2.4.5. Tomek Links Örnekleme Dengeleme Yöntemi

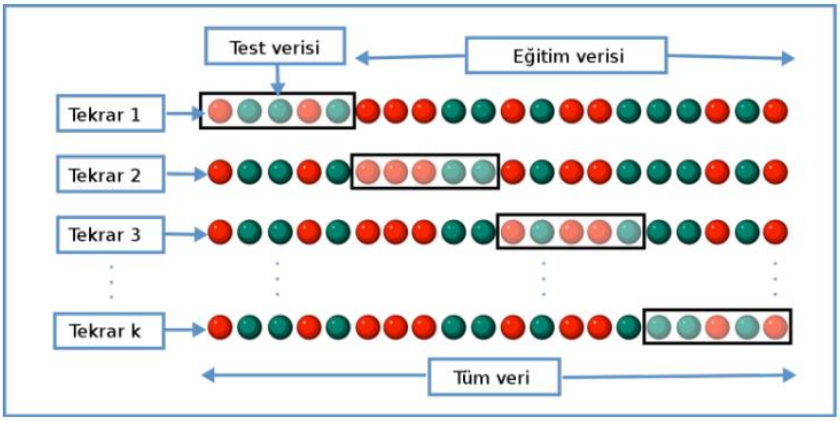
Tomek Links veri dengeleme yöntemi, sınıf dengesizliği olan veri setlerini düzenlemek için kullanılan bir tekniktir. Bu yöntem, azınlık ve çoğunluk sınıfları arasındaki sınırı temizleyerek sınıflandırma performansını iyileştirmeyi amaçlar. İsmi, Ivan Tomek tarafından 1976'da yapılan bir çalışmaya dayanmaktadır [62].

5.2.5. K katlamalı çapraz doğrulama

Makine Öğrenimi algoritmalarının uygulanmasında, modelin eğitime başlanabilmesi için hiper parametrelerin belirlenmesi zorunludur. Çapraz doğrulama yöntemi, tahmin edici modellerin genelleştirme kapasitesini ölçmek ve aşırı uyum veya yetersiz uyum sorunlarını engellemek amacıyla kullanılan bir veri yeniden örnekleme stratejisidir [63]. Bu çalışma

kapsamında, K-n katlamalı çapraz doğrulama yöntemi kullanılmıştır ve k değeri olarak 10 belirlenmiştir. Doğrulama süreci, verilen n değerine uygun olarak 10 kez tekrar edilmiş ve her tekrarda elde edilen doğruluk oranları ağırlıklı ortalama yöntemiyle birleştirilerek modelin tahmin başarısının artırılması amaçlanmıştır [64].

K-Katlamalı çapraz doğrulama, çapraz doğrulama teknikleri arasında yaygın olarak tercih edilen bir yöntemdir. Bu yöntemde, K parametresi, bir veri setinin kaç adet bölüme ayrılacağını ifade eder. Bu bölümlerden biri doğrulama seti olarak ayrılırken, makine öğrenimi modeli, geriye kalan K-1 bölüm üzerinde eğitim gerçekleştirir. K-Katlamalı çapraz doğrulamanın her bir adımında, farklı bir bölüm doğrulama seti olarak işlev görür ve bu süreç, K adet değerlendirme puanı (örneğin, doğruluk) elde edilene kadar devam eder. Son olarak, Şekil 27'deki açıklandığı üzere, modelin genel performansını değerlendirmek için bu puanların ortalaması alınarak nihai bir model puanı hesaplanır [65].



Şekil 27. Çapraz Doğrulama [65]

5.3. Hiper Parmetre Ayarlama Yöntemleri

Makine öğrenmesinde "Hyperparameter Tuning" (Hiperparametre Ayarlama), bir makine öğrenmesi modelinin performansını en üst düzeye çıkarmak için kullanılan parametrelerin (hiperparametrelerin) ayarlanması sürecidir. Bu parametreler genellikle algoritmanın kendisi tarafından öğrenilemez, bu yüzden doğru değerlerin manuel olarak veya otomatik yöntemlerle belirlenmesi gerekir.

Hiperparametre ayarlaması için aşağıdaki yöntemler önerilmektedir:

- Manuel(Elle) Ayarlama
- Izgara Araması (Grid Search CV)
- Rastgele Arama

- Bayes Optimizasyonu
- Genetik Algoritmalar
- Parçacık Sürüsü Optimizasyonu

5.3.1. Manuel Arama Modeli

Manuel Arama yönteminde, deneyimlerden yola çıkarak başlangıçta seçilen hiperparametre değerleri üzerinden modelin eğitim süreci başlatılır. Modelin performansı, tahminleme doğruluğu üzerinden değerlendirildikten sonra, hiperparametre değerlerinde yapılan ayarlamalarla süreç tekrarlanır. Bu işlem, memnun edici bir doğruluk seviyesine ulaşılana dek sürdürülür. Rastgele Orman algoritmasında kullanılan temel parametreler aşağıdaki gibidir:

- criterion fonksiyonu, bir bölünmenin kalitesini ölçmek için kullanılır.
- max_depth, ağaçların ulaşabileceği maksimum derinliği belirler.
- max_features, bir düğümün bölünmesi sırasında göz önünde bulundurulan maksimum özellik sayısını ifade eder.
- min_samples_leaf, bir yaprak düğümünde bulunması gereken minimum örnek sayısını tanımlar.
- min_samples_split, bir düğümün bölünmesi için gereken minimum örnek sayısını belirtir.
- n_estimators, ormandaki ağaçların toplam sayısını gösterir.

Rastgele Orman hiperparametreleriyle ilgili daha detaylı bilgi, Scikit-Learn'in resmi dokümantasyonunda yer almaktadır.

5.3.2. Izgara Araması (Grid Search) Modeli

Izgara Arama, hiperparametre ayarlamada kullanılan geleneksel yöntemlerden biri olarak tanımlanabilir. Bu yaklaşım, önceden belirlenmiş hiperparametre değerlerinin her bir mümkün kombinasyonunu test etmeyi içerir. İlk adımda, seçilen hiperparametreler için makul değerlerin bir Kartezyen çarpımı oluşturulur. Ardından, belirli bir makine öğrenimi algoritması, bu hiperparametre kombinasyonlarının her biri kullanılarak eğitilir. Eğitim sürecinin performansı, çapraz doğrulama tekniği ile değerlendirilir, böylece oluşturulan ızgara üzerindeki her bir kombinasyon için bir performans puanı elde edilir. En yüksek performansı sağlayan hiperparametre kombinasyonu tespit edilir ve bu kombinasyon, seçilen

model için kullanılır. Deneyim gerektirmeyen bu yöntem, en uygun hiperparametre değerlerinin belirlenmesini sağlamakla birlikte, hiperparametrelerin sayısının fazla olduğu durumlarda yüksek hesaplama gücü gerektirebilir ve büyük hiperparametre uzayını tarama süreci uzun sürebilir. N adet farklı değere sahip k hiperparametre için, Izgara Aramasının hesaplama karmaşıklığı O(N^k) şeklinde artmaktadır [66].

5.3.3. Rastgele Arama (Random Search) Modeli

Rastgele Arama yöntemi, hiperparametrelerin rastgele kombinasyonlarını kullanarak en etkili modeli bulmayı amaçlayan bir tekniktir. Bu yöntem, arama alanını örnekleme ve belirlenen bir olasılık dağılımına göre değerlendirmeler yapma esasına dayanır. Örnek olarak, 100.000 olası kombinasyonun tamamını değerlendirmek yerine, bu yöntemle 1000 rastgele hiperparametre seti incelenir. Rastgele Arama sürecinde değerlendirilecek kombinasyonların sayısı, optimizasyon süreci başlamadan önce belirlenir. Bu bağlamda, n adet değerlendirme gerçekleştiren Rastgele Arama'nın hesaplama karmaşıklığı O(n) olarak ifade edilir [67]. Rastgele Arama yönteminin önemli bir eksikliği, seçimlerini yaparken önceki deneyimlerden elde edilen bilgileri dikkate almaması ve sonraki adım için bir strateji belirlememesidir. Hiperparametrelerin belirli bir aralıkta rastgele seçilen değerlerini dener, genellikle ızgara aramasından daha hızlıdır.

5.3.4. Bayes Tabanlı Optimizasyon Modeli

Bayes Optimizasyonu, bu bağlamda bilgiye dayalı bir arama yöntemi olarak öne çıkar. Her iterasyonda, bu algoritma önceki adımdan elde edilen bilgileri kullanarak öğrenir ve sonraki öğrenme adımını şekillendirir. Bayes Optimizasyonu, hiperparametre kombinasyonlarının bir alt kümesini örnekleme yöntemiyle Rastgele Aramaya benzerlik gösterse de, kombinasyonların seçilme yöntemi açısından Rastgele Aramadan ayrılır. Bayes Araması, hiper parametre ayarlama sürecini, bir kara kutu fonksiyonunun optimizasyonu olarak ele alır. Bu süreçte, modelin son tahmin doğruluğu puanı göz önünde bulundurularak, algoritmanın en uygun yapıya indirgenmesi için herhangi bir küresel optimizasyon yöntemi uygulanabilir. Bu optimizasyon sürecinde, hedeflenen nihai modele yakınsamak adına ara aşamada kullanılan modele "vekil model" adı verilir. Bayes Araması, birçok türev modeli barındırmakla birlikte, en sık kullanılanı Gauss Süreci'ni (GS) temel alır. Bu süreçte, ilk adım olarak birkaç hiper parametre kombinasyonu rastgele seçilip test edilir. Bu rastgele seçilen

hiper parametre değerleri, amaç fonksiyonunun ilk modelinin oluşturulmasında kullanılır [68].

5.3.5. Genetik Arama Modeli

Metasezgisel algoritmalar, biyolojik kuramlardan esinlenerek geliştirilen yaklaşımlardır ve özellikle sürekli olmayan ve dışbükey olmayan optimizasyon sorunlarını çözme kapasiteleri ile öne çıkarlar. Bu algoritmalar, her iterasyonda, bir nesli ifade eden bir popülasyon ile işe koyulur. Bir nesil, çeşitli özelliklere sahip bireylerden, yani "kromozomlar" ile temsil edilen potansiyel çözüm setlerinden meydana gelir. Her bir nesil içindeki potansiyel çözümler, genel bir optimum çözüm bulunana kadar değerlendirilir [69]. Metasezgisel algoritmalar, popülasyon oluşturma ve seçim metodolojileri bakımından birbirlerinden ayrılırlar. Bu algoritmaların bilinen bir temsilcisi Genetik Algoritma'dır (GA). GA, çevresel faktörlere en uygun adaptasyonu sergileyen bireylerin hayatta kalma şansının daha fazla olduğu ve bu yolla yeteneklerini sonraki nesillere aktarabilecekleri evrimsel teoriye dayanır [70]. Sonraki nesiller, ebeveynlerinden miras aldıkları özelliklerle, ebeveynlerinin daha iyi veya daha kötü versiyonlarına sahip bireyler üretebilmektedir. Bu bireyler arasında hayatta kalma ve üreme başarısı bakımından farklılıklar gözlemlenmektedir. Daha iyi özelliklere sahip olan bireylerin hayatta kalma ve yavru bırakma olasılığı daha yüksektir. Bu sayede daha iyi genlerin bir sonraki nesillere aktarılması sağlanmaktadır. Daha az uyumlu özelliklere sahip olan bireyler ise zamanla elenmektedir. Birkaç nesil sonra, en iyi kromozomları (özellikleri) taşıyan bireyler popülasyonda baskın hale gelerek küresel optimum olarak tanımlanan bir duruma ulaşılmaktadır [71]. Hiperparametre optimizasyon probleminde, her bir hiperparametre bir kromozom tarafından temsil edilmektedir. Hiperparametre değerleri, temsili kromozomun ondalık değerine göre ayarlanmaktadır. Her kromozom, ikili basamak sisteminde kodlanmış bir dizi genden oluşmaktadır. Optimum parametreleri belirlemek için bu kromozomların genleri üzerinde kromozom seçimi, çaprazlama ve mutasyon işlemleri uygulanmaktadır. Yüksek uygunluk fonksiyonu değerlerine sahip kromozomlar seçilerek bir sonraki nesile aktarılma şansları artırılmaktadır. Bir sonraki nesilde bu kromozomlar, ebeveynlerinin en iyi özelliklerini taşıyan yeni kromozomlar üretmektedir. Çaprazlama işlemi, arama uzayında çözümlerin karışımını elde etmek için kullanılır. Farklı kromozomların genleri oranlanarak yeni kromozomlar üretilmektedir. Bu sayede çözüm uzayında yeni ve daha iyi çözümler keşfedilmesi sağlanmaktadır [72]. Mutasyon, genetik algoritmaların bir parçası olarak, kromozomlardaki bir ya da birden fazla genin rastgele

modifiye edilmesi sürecidir. Bu süreç, popülasyondaki çeşitliliği artırarak ve yeni çözüm yollarını ortaya çıkararak genetik algoritmaların etkinliğini artırmak amacıyla kullanılır [73]. Çaprazlama ve mutasyon süreçleri, gelecek nesillerdeki çeşitliliği artırır ve farklı özelliklerin bulunmasını sağlar, böylece olumlu özelliklerin gözden kaçırılma riskini minimize eder [74]. Başlangıç adımı, uygun bir tahmin ile ayarlanarak algoritmanın yakınsama hızını artırabilmektedir. Bu nedenle, her optimizasyon probleminde önemli bir adım olarak kabul edilir. Yakınsama işlemi sırasında başlangıç değerleri yinelemeli olarak iyileştirilse bile, arama uzayı içinde ümit vermeyen yerel bölgelere hapsolmadan yakınsamayı optimum noktaya yaklaştıran değerleri seçmek her zaman daha yararlı olacaktır. Genetik Algoritma'nın (GA) avantajlarından biri, iyi başlangıç değerlerinin seçiminde aşırı çaba gerektirmeden rastgele seçilmesine de izin vermesidir. Mutasyon, çaprazlama ve seçim işlemleri, küresel optimumun kaçırılmamasını sağlayarak algoritmanın sağlamlığını artırmaktadır. Ancak, GA sıralı bir yürütme algoritması olduğundan paralelleştirme olasılığı daha düşüktür. Bu durum, GA'nın zaman karmaşıklığını O(n^2) mertebesine yükseltmekte ve düşük yakınsama hızı olarak kabul edilmesine neden olmaktadır [75].

5.3.6. Parçacık Sürü Optimizasyonu Modeli

Parçacık Sürü Optimizasyonu (PSO), optimizasyon problemlerini çözmek için kullanılan güçlü bir meta-sezgisel algoritmadır. PSO, kuş sürüsü veya balık sürüsü gibi doğal popülasyonların davranışlarından ilham alarak çalışır. [76]. Parçacık Sürü Optimizasyonu (PSO), bir arama uzayını yarı rastgele biçimde taramak için bir grup parçacığın (sürü) kullanıldığı, iteratif bir optimizasyon sürecini ifade eder. Bu süreçte, optimizasyon, sürünün arama uzayındaki hareketleri vasıtasıyla gerçekleştirilir [77]. Parçacık Sürü Optimizasyonu (PSO) sürecinde, bilgi paylaşımı ve işbirliği, bir grup parçacığın bilgileri arasında gerçekleştirilir. Bu süreçte, sürüyü oluşturan parçacıklar, kendi en iyi bilinen pozisyonlarını, şu anki pozisyonlarını ve hızlarını içeren bir vektörle temsil edilirler. Parçacıkların konumları ve hızları belirlendikten sonra, her bir parçacığın mevcut konumu ve tahmini performansı değerlendirilir. İlerleyen yinelemelerde, her parçacığın hızı, önceki yinelemelerden elde edilen bilgilere ve küresel en uygun konuma dayanarak güncellenir. Parçacıklar, bu yeni hız vektörlerine göre hareket etmeye devam ederler [78]. Bu süreçler, belirlenen yakınsama ya da sonlandırma kriterleri karşılanana dek yinelenir. PSO'nun, çaprazlama ve mutasyon işlemlerini barındırmaması, GA'ya kıyasla uygulamasını daha basit

kılar. GA içerisinde, bilgi tüm kromozomlar arasında paylaşılır ve bu sayede popülasyon, optimal bölgeye doğru düzenli bir ilerleme kaydeder. PSO'da ise, sadece bireysel en iyi ve genel en iyi parçacıkların bilgisi diğerlerine aktarılır. PSO algoritmasının hesaplama karmaşıklığı O(nlogn) olarak ifade edilir [79]. PSO, GA'ya göre genellikle daha hızlı bir yakınsama sürecine sahiptir. Bunun yanı sıra, PSO'nun paralelleştirilmesi daha basittir çünkü parçacıklar bağımsız hareket eder ve yalnızca her iterasyon sonunda bilgi alışverişi yaparlar. Fakat, GA ile karşılaştırıldığında, PSO'da popülasyonun etkin bir şekilde başlatılması gerekmektedir. Eğer uygun bir başlangıç yapılmazsa, PSO algoritması, özellikle ayrık hiperparametreler söz konusu olduğunda, yalnızca yerel optimumları bulabilir ve global optimumu gözden kaçırabilir. Bu durumu gidermek amacıyla, PSO'nun yakınsama hızını ve performansını artırmaya yönelik çeşitli popülasyon başlatma teknikleri geliştirilmiştir [80]. Ancak, popülasyon başlatma yaklaşımlarından birinin kullanılması, PSO'nun karmaşıklığını ve yürütme süresini artıracaktır.

Tablo 3. Hiper parametre ayarlama modellerinin karşılaştırılması

Özellik/ Algoritma	**Manuel Aramam**	**Izgara Araması**	**Rastgele Araması**	**Bayes Optimizasyonu**	**Genetik Algoritmalar**	**Parçacık Sürü Optimizasyonu**
Açıklama	Hiperparametrelerin elle ayarlanması.	Tüm olası parametre kombinasyonlarını deneme.	Parametre değerlerinin rastgele seçilmesi.	Olasılık modeline dayalı optimizasyon.	Evrimsel süreçlerle parametre optimizasyonu.	Sürü zekasına dayalı optimizasyon.
Zaman Verimliliği	Değişken	Düşük	Orta	Yüksek	Orta-Yüksek	Orta-Yüksek
Uzay Verimliliği	Düşük	Düşük	Düşük	Yüksek	Yüksek	Yüksek
Kullanım Kolaylığı	Yüksek	Orta	Yüksek	Orta	Orta	Orta
Ölçeklenebilirlik	Düşük	Düşük	Yüksek	Yüksek	Yüksek	Yüksek
En İyi Kullanım	Basit modeller veya deneyimli kullanıcılar için.	Küçük parametre uzayları için.	Büyük parametre uzayları için hızlı çözüm.	Karmaşık parametre uzayları için.	Çok boyutlu ve karmaşık uzaylar için.	Karmaşık problemler ve geniş arama uzayları için.

Model Bağımsızlığı	Evet	Evet	Evet	Çoğunlukla	Evet	Evet
Optimizasyon Süreci	Kullanıcı deneyimine bağlı.	Sistematik ve kapsamlı.	Rastgele ve esnek.	Veriye dayalı ve iteratif.	Evrimsel ve adaptif.	Sürü davranışı ve adaptif.

Optimizasyon yöntemlerinin özelikleri Tablo 3'te listelenmiştir.

Hiperparametre ayarlaması, özellikle geniş veri kümeleri ve karmaşık modeller söz konusu olduğunda, zaman ve kaynak bakımından yoğun bir süreçtir. Bu durum, bazen yaklaşık metotların kullanılmasını veya deneme sayısının sınırlanmasını gerektirir. Süreci kolaylaştırmak amacıyla otomatik hiperparametre ayarlama araçları (AutoML) tercih edilmektedir.

6. BULGULAR VE YORUM

Bu çalışmanın dördüncü bölümünde yer alan Şekil 13'de görselleştirilen model, Python programlama dili ve onun güçlü kütüphaneleri olan "pandas", "numpy" ve "sklearn" yardımıyla hayata geçirilmiştir. Bu kütüphaneler, çeşitli istatistiksel yöntemler ve klasik makine öğrenimi sınıflandırma tekniklerinin kullanımını mümkün kılarak, modelin oluşturulmasında temel taşları oluşturmuştur. PyCharm geliştirme aracının imkanlarından yararlanılarak gerçekleştirilen bu süreç, 11. nesil Intel(R) Core(TM) i7-1165G7 @ 2.80GHz hızında çalışan işlemci ve 16 GB RAM belleğe sahip bir diz üstü bilgisayar kullanılarak, Windows 10 Enterprise LTSC işletim sistemi yüklü ortamda test edilmiştir. Bu donanım ve yazılım özellikleri, model geliştirme ve test süreçlerinin etkin bir şekilde yürütülmesine olanak sağlamıştır.

6.1. Hazırlanan Veri Seti Modellerin Çalıştırılması

Bu çalışmanın ilk aşamasında, veri seti eğitim ve test setleri olarak ayrılmıştır. Bu ayrımın başarıyla tamamlandığı görülmekte olup, eğitim seti 1176 örnek, test seti ise 294 örnek içermektedir. Hem eğitim hem de test setlerinde 35 bağımsız değişken bulunmaktadır.

Bu çalışmanın bir sonraki aşamasında, sınıflandırma algoritmalarının uygulanması ve bu algoritmalar için karışıklık matrisi ile performans metriklerinin hesaplanması gerçekleştirilecektir. İlk olarak, çeşitli modeller tanımlanıp eğitilecek, ardından bu modellerle test veri seti üzerinde tahminler yapılacak ve son olarak her model için doğruluk, geri çağırma, kesinlik ve F1 skoru gibi performans metrikleri hesaplanacaktır.

Bu sürecin uygulanmasında aşağıdaki adımlar takip edilecektir:

- Modellerin tanımlanması.
- Eğitim seti üzerinde her modelin eğitilmesi.
- Test seti üzerinde tahminlerin yapılması.
- Karışıklık matrisi ve performans metriklerinin hesaplanması.
- Elde edilen sonuçların karşılaştırılabilir bir tablo formatında sunulması.

Test ve eğitim veri setimize sırasıyla, Lojistik Regresyon(LR), Destek Vektör Makinesi(SVM), En Yakın Komşu (KNN)), Karar Ağaçları(DT), Rastgele Orman

Sınıflandırması(RF), Naïve Bayes(NB) ve hibrit modeller olan XGBoost(XGB) ile AdaBoost sınıflandırma algoritmaları uygulanmıştır. Oluşan sonuçlar aşağıdaki Tablo 4'de gösterilmiştir.

Tablo 4. Geleneksel makine öğrenimi algoritmaları için performans değerlendirme ölçütleri, Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

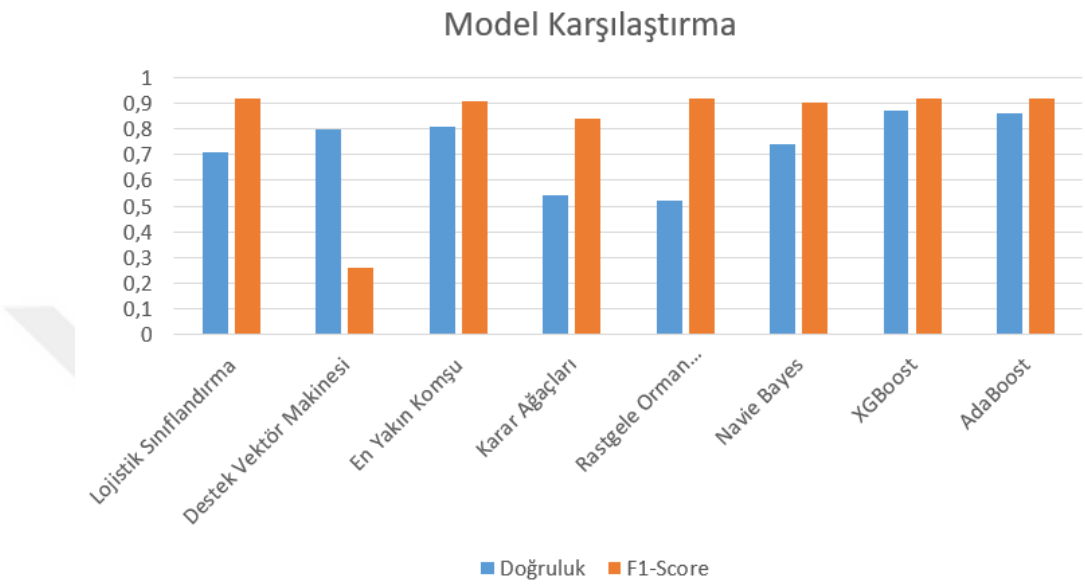
ML Algoritmalar	A	P 0	P 1	R 0	R 1	F1-0	F1-1
LR	0.87	0.88	0.68	0.97	0.32	0.92	0.43
SVM	0.86	0.86	0.92	1.00	0.15	0.92	0.26
KNN	0.84	0.85	0.45	0.99	0.06	0.91	0.11
DT	0.76	0.85	0.27	0.85	0.28	0.85	0.32
RF	0.85	0.85	0.53	0.98	0.10	0.92	0.21
NB	0.83	0.87	0.45	0.93	0.30	0.90	0.36
XGB	0.85	0.87	0.62	0.97	0.23	0.92	0.33
AdaBoost	0.86	0.88	0.59	0.95	0.34	0.92	0.43

Tablo 4, geleneksel makine öğrenimi algoritmalarının performans değerlendirme ölçütlerini kapsamaktadır. Bu ölçütler, Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru (F1-score) olarak belirlenmiştir. Makine öğrenimi algoritmaları; Lojistik Regresyon (LR), Destek Vektör Makineleri (SVM), K-En Yakın Komşu (KNN), Karar Ağaçları (DT), Rastgele Ormanlar (RF), Naive Bayes (NB), XGBoost (XGB) ve AdaBoost olarak sıralanmıştır.

Her algoritmanın performansı, sınıflandırma problemlerinde sıkça kullanılan iki sınıf (0 ve 1) için ayrı ayrı incelenmiştir. Burada '0' ve '1', sınıflandırma yapılan kategorileri temsil etmektedir. Kesinlik (Precision), modelin pozitif olarak tahmin ettiği örneklerin gerçekten ne kadarının pozitif olduğunu; Duyarlılık (Recall), gerçek pozitif örneklerin ne kadarının doğru olarak pozitif olarak tahmin edildiğini gösterir. F1 Skoru ise, Kesinlik ve Duyarlılık ölçütlerinin harmonik ortalamasıdır ve her iki ölçütü de dengeli bir şekilde değerlendirmeyi amaçlar.

Analize göre, LR, AdaBoost ve SVM algoritmaları, diğerlerine göre daha yüksek Doğruluk ve F1 Skorları ile öne çıkmaktadır. Özellikle, LR ve AdaBoost algoritmaları '1' sınıfı için Kesinlik ve Duyarlılık dengesini sağlayarak yüksek F1 Skorlarına ulaşmıştır. Diğer yandan, KNN ve RF algoritmaları, özellikle '1' sınıfı için düşük Duyarlılık değerleri ile dikkat çekmektedir.

Bu değerlendirmeler, belirli bir sınıflandırma problemi için en uygun makine öğrenimi algoritmasının seçiminde önemli bir rehber oluşturur. Özellikle, farklı sınıflar arasındaki dengesizliklerin ve performans ölçütlerinin önemi göz önünde bulundurulmalıdır. Bu analiz, makine öğrenimi modellerinin performansının kapsamlı bir şekilde değerlendirilmesi ve optimizasyonu için temel bir adımı temsil eder.



Şekil 28. Sınıflandırma algoritmalarının performanslarının karşılaştırılması

Hem Tablo 4'te hem de Şekil 28'de gösterildiği gibi "0" kategorisinde olan yani işe devam eden çalışanları ifade eden değerlerin, "1" kategorisinde işten ayrılanlara göre daha yüksek olduğunu göstermektedir. Bu durum veri setimizde dengesiz(imbalanced) veri dağılımını problemini ifade etmektedir. Yıpranma hedef değişkenin "Evet" durumunun veri seti içerisinde azınlık sayıda dağılım gösterdiği gözlemlenmiştir.

Eksik veri(İmbalanced Dataset) çözmek için veri dengeleme yöntemlerini bölüm ikideki açıklandığı gibi araştırılmıştır. Sonrasında aşağı ve yukarı yönde veri seti dengelemesi için ROS, SMOTE, RUS, Near Miss(NM) ve Tomek Links(TL) yöntemleri uygulanmıştır. Öncelikle dengesiz veri kümesi işleme algoritmaları uygulanarak veri kümesi dengelenmiştir. Bu aşamada Rastgele Aşırı Örnekleme (ROS), Rastgele Eksik Örnekleme (RUS), SMOTE, NearMiss(NM) ve Tomek Links(TL) yöntemleri ayrı ayrı veri setine uygulanmıştır. Daha sonra, makine öğrenmesi yöntemleri dengeli veri kümesine uygulanarak her model için sonuçlar oluşturulmuştur. Devamında, dengesiz veri kümesine uygulanan makine öğrenmesi yöntemlerinin sonuçları ile dengesiz veri kümesi işleme

teknikleri kullanılarak dengelenmiş veri kümesine uygulanan makine öğrenmesi yöntemleriyle karşılaştırılmıştır.

Lojistik Regresyon(LR) ile veri dengeleme yöntemlerinin uygulanması sonrası oluşan sonuçlar aşağıdaki Tablo 5'te gösterilmiştir.

Tablo 5. Dengesiz veri yöntemleri ile lojistik regresyon(LR) modeli performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

ML Algoritm	A	P 0	P 1	R 0	R 1	F1-0	F1-1
ROS-LR	0.72	0.88	0.81	0.79	0.89	0.83	0.85
SMOTE-LR	0.73	0.86	0.86	0.86	0.86	0.86	0.86
RUS-LR	0.70	0.71	0.73	0.73	0.70	0.72	0.72
NM-DT	0.70	0.69	0.77	0.81	0.63	0.74	0.69
TL-DT	0.87	0.89	0.71	0,97	0,38	0,93	0,50

Tablo 5, dengesiz veri setleri üzerinde uygulanan çeşitli yöntemlerle geliştirilen lojistik regresyon (LR) modellerinin ve karar ağaçları (DT) kullanılarak yapılan denemelerin performans ölçütlerini içermektedir. Performans ölçütleri; Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru (F1-score) olarak detaylandırılmıştır. Bu değerlendirme, Random Over Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), Random Under Sampling (RUS), NearMiss (NM) ve Tomek Links (TL) gibi yöntemlerin uygulanması sonucunda elde edilen sonuçları karşılaştırmaktadır.

Her bir yöntemin etkinliği, '0' ve '1' olarak ifade edilen iki sınıf üzerindeki performansları ile değerlendirilmektedir. Bu iki sınıf, sınıflandırma problemlerinde hedeflenen kategorileri temsil etmektedir. Kesinlik ölçütü, bir modelin pozitif olarak tahmin ettiği durumların gerçekte ne kadarının pozitif olduğunu; Duyarlılık ise, gerçek pozitif durumların ne oranda doğru olarak pozitif tahmin edildiğini gösterir. F1 Skoru, Kesinlik ve Duyarlılık ölçütlerinin harmonik ortalaması alınarak hesaplanır ve her iki ölçüt arasında dengeli bir performans değerlendirmesi sunar.

Analize göre, TL yöntemi ile zenginleştirilmiş DT modeli, diğer yöntemlere kıyasla '0' sınıfı için yüksek Doğruluk ve F1 Skoru elde etmiş, bu da onu dengesiz veri setleri için özellikle etkili bir yöntem olarak öne çıkarır. ROS ve SMOTE yöntemleri ile geliştirilen LR modelleri de, özellikle '1' sınıfı için dengeli Kesinlik, Duyarlılık ve F1 Skorları sağlayarak dikkat çekmiştir.

Bu analiz, dengesiz veri setlerinin işlenmesinde kullanılan farklı yöntemlerin etkinliğinin kapsamlı bir şekilde değerlendirilmesi gerektiğini göstermektedir. Her bir yöntemin, veri setinin özelliklerine ve hedeflenen model performans kriterlerine bağlı olarak değişken sonuçlar üretebileceği belirlenmiştir. Bu değerlendirme, makine öğrenimi modellerinin geliştirilmesi ve iyileştirilmesi sürecinde önemli bir referans noktası sağlar.

Destek Vektör Makinesi(SVM) ile veri dengeleme yöntemlerinin uygulanması sonrası oluşan sonuçlar aşağıdaki Tablo 6'da gösterilmiştir.

Tablo 6. Dengesiz veri yöntemleri ile Destek Vektör Makinesi (SVM) modeli performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

ML Algoritm	A	P 0	P 1	R 0	R 1	F1- 0	F1-1
ROS-SVM	0.84	0.88	0.81	0.79	0.89	0.83	0.85
SMOTE-SVM	0.86	0.86	0.86	0.86	0.86	0.86	0.86
RUS-SVM	0.72	0.73	0.73	0.70	0.72	0.72	0.73
NM-SVM	0.72	0.77	0.81	0.63	0.74	0.69	0.77
TL-SVM	0,87	0,87	0,88	0,99	0,25	0,93	0,38

Tablo 6, dengesiz veri setleri üzerinde farklı dengeleme yöntemleri uygulandıktan sonra Destek Vektör Makinesi (SVM) modelinin performans değerlerini içermektedir.

Analiz sonuçlarına göre, SMOTE ve TL yöntemleri ile iyileştirilmiş SVM modelleri, diğer dengeleme tekniklerine göre daha yüksek Doğruluk ve F1 Skorları elde etmiştir. Bu, özellikle SMOTE yönteminin '1' sınıfı için ulaştığı dengeli Kesinlik, Duyarlılık ve F1 Skorları ile kendini göstermektedir. RUS ve NM yöntemleri, özellikle '1' sınıfı için daha

düşük performans değerleri sergileyerek, bu yöntemlerin bazı durumlarda veri kaybına yol açabileceğini ve modelin genel performansını olumsuz etkileyebileceğini ortaya koymaktadır.

Bu değerlendirme, dengesiz veri setlerinin işlenmesi sürecinde kullanılabilecek farklı yöntemlerin etkinliğini ve bu yöntemlerin model performansı üzerindeki etkilerini gözler önüne sermektedir. Yöntem seçimi, veri setinin özellikleri ve modelin maksimize etmeye çalıştığı performans ölçütleri göz önünde bulundurularak yapılmalıdır. Bu analiz, makine öğrenimi alanında model geliştirme ve optimizasyon çalışmaları için önemli bir referans teşkil eder.

En Yakın Komşu(KNN) ile veri dengeleme yöntemlerinin uygulanması sonrası oluşan sonuçlar aşağıdaki Tablo 7'de gösterilmiştir.

Tablo 7. Dengesiz veri yöntemleri ile En Yakın Komşu (KNN) modeli performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

ML Algoritm	A	P 0	P 1	R 0	R 1	F1-0	F1-1
ROS-KNN	0.82	0.90	0.77	0.72	0.92	0.80	0.84
SMOTE-KNN	0.79	0.99	0.71	0.59	0.99	0.74	0.83
RUS-KNN	0.65	0.63	0.69	0.74	0.57	0.69	0.62
NM-KNN	0.72	0.56	0.66	0.81	0.38	0.67	0.48
TL-KNN	0,84	0,86	0,61	0,97	0,20	0,91	0,30

Tablo 7, dengesiz veri setlerine uygulanan farklı yöntemler sonrasında En Yakın Komşu (KNN) modelinin performans ölçütleri üzerindeki etkilerini göstermektedir. Bu ölçütler; Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru (F1-score) olarak tanımlanmıştır. Değerlendirilen yöntemler arasında Random Over Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), Random Under Sampling (RUS), NearMiss (NM) ve Tomek Links (TL) bulunmaktadır.

Performans değerlendirmesinde '0' ve '1' sınıfları, sınıflandırma işleminde ayrım yapılan iki ana kategori olarak karşımıza çıkar. Kesinlik, modelin pozitif olarak işaretlediği durumların, gerçekten ne kadarının pozitif olduğunu; Duyarlılık ise, gerçekte pozitif olan örneklerin, model tarafından ne oranda doğru tahmin edildiğini ifade eder. F1 Skoru, Kesinlik ve Duyarlılık değerlerinin harmonik ortalaması ile hesaplanır ve bu iki ölçüt arasında bir denge kurulmasını sağlar.

Analiz sonuçları, ROS ve TL yöntemleri ile iyileştirilmiş KNN modellerinin, diğer yöntemlere göre daha yüksek Doğruluk ve F1 Skorlarına ulaştığını ortaya koymaktadır. SMOTE yöntemi, özellikle '0' sınıfı için son derece yüksek bir Kesinlik oranı elde etmiş, ancak bu, '1' sınıfı için Duyarlılık değerinin düşük kalmasına neden olmuştur. RUS ve NM yöntemleriyle elde edilen modeller, genel olarak daha düşük performans değerleri sergileyerek, bu yöntemlerin veri setinin yapısına bağlı olarak bazı sınırlamalara sahip olabileceğini göstermektedir.

Bu değerlendirme, dengesiz veri setlerinin analizinde ve işlenmesinde kullanılan yöntemlerin model performansına olan etkisini gözler önüne serer. Yöntem seçimi, sınıflandırma probleminin özellikleri ve hedeflenen performans ölçütleri dikkate alınarak yapılmalıdır. Analiz, modelin performansını artırmak ve en uygun yöntemi seçmek için kritik bir rehberlik sunar.

Karar Ağaçları ile veri dengeleme yöntemlerinin uygulanması sonrası oluşan sonuçlar aşağıdaki Tablo 8'de gösterilmiştir.

Tablo 8. Dengesiz veri yöntemleri ile Karar Ağaçları(DT) modeli performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

ML Algoritm	A	P 0	P 1	R 0	R 1	F1-0	F1- 1
ROS- DT	0.74	1.00	0.65	0.47	1.00	0.64	0.79
SMOTE- DT	0.67	0.83	0.63	0.49	0.87	0.61	0.73
RUS- DT	0.50	0.52	0.54	0.70	0.35	0.59	0.43
NM- DT	0.51	0.50	0.50	0.59	0.41	0.54	0.45

TL-DT	0.79	0,87	0,35	0,85	0,39	0,86	0,37

Tablo 8, dengesiz veri setleri üzerine uygulanan çeşitli dengeleme yöntemlerinin Karar Ağaçları (Decision Trees, DT) modeli üzerindeki etkisini gösteren performans ölçütlerini sunmaktadır.

Analiz sonuçlarına göre, ROS yöntemiyle geliştirilen DT modeli, '0' sınıfı için Kesinlikte mükemmel bir performans sergileyip, F1 Skorunda yüksek bir değer elde etmiştir. Bununla birlikte, bu yüksek Kesinlik, '1' sınıfı için Duyarlılığın düşük kalmasıyla dengelenmiştir. SMOTE ve TL yöntemleri, her iki sınıf için dengeli bir performans sunarken, RUS ve NM yöntemleri genel olarak daha düşük performans değerleri ile dikkat çekmiştir. Özellikle RUS yöntemi, '0' sınıfı için yüksek bir Duyarlılık sergilese de, bu, Kesinlik ve F1 Skorlarında düşük değerlere yol açmıştır.

Bu değerlendirme, dengesiz veri setlerinin analizinde ve işlenmesinde kullanılan farklı yöntemlerin Karar Ağaçları modelinin performansına etkilerini ortaya koymaktadır. Yöntem seçimi, sınıflandırma probleminin özellikleri ve hedeflenen performans ölçütleri dikkate alınarak yapılmalıdır. Bu çalışma, modelin performansını iyileştirmek ve en uygun yöntemi belirlemek için önemli bir referans sunar.

Rastgele Orman Algoritması ile veri dengeleme yöntemlerinin uygulanması sonrası oluşan sonuçlar aşağıdaki Tablo 9'da gösterilmiştir.

Tablo 9. Dengesiz veri yöntemleri ile Rastgele Orman(RF) modeli performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

ML Algoritm	**A**	**P 0**	**P 1**	**R 0**	**R 1**	**F1- 0**	**F1- 1**
ROS- RF	0.96	0.97	0.95	0.95	0.98	0.96	0.96
SMOTE- RF	0.88	0.90	0.85	0.84	0.91	0.87	0.88
RUS- RF	0.59	0.57	0.66	0.80	0.40	0.67	0.50
NM- RF	0.60	0.60	0.68	0.77	0.49	0.67	0.57

TL-RF	0.85	0,86	0,81	0,99	0,19	0,92	0,30

Tablo 9, dengesiz veri setlerinin işlenmesi için uygulanan yöntemler sonrasında Rastgele Orman (RF) modelinin performans ölçütleri ile ilgili ayrıntılı bilgiler sunmaktadır.

Analize göre, ROS yöntemi ile iyileştirilmiş RF modeli, Doğruluk, Kesinlik, Duyarlılık ve F1 Skoru açısından yüksek performans sergileyerek dikkat çekmektedir. Bu yöntem, modelin her iki sınıf üzerinde dengeli ve yüksek bir başarı elde etmesini sağlamıştır. SMOTE yöntemi de benzer şekilde, özellikle '1' sınıfı için yüksek Duyarlılık ve F1 Skorları ile olumlu sonuçlar sunmuştur. Öte yandan, RUS ve NM yöntemleri, genellikle daha düşük performans değerleri göstererek, bu tekniklerin bazı durumlarda modelin genel başarısını olumsuz yönde etkileyebileceğini ortaya koymuştur.

Bu değerlendirme, dengesiz veri setlerinin işlenmesinde kullanılan farklı yöntemlerin Rastgele Orman modelinin performansına etkisini kapsamlı bir şekilde göstermektedir. Yöntem seçimi, veri setinin yapısına ve modelin hedeflenen performans kriterlerine uygun olarak yapılmalıdır. Bu analiz, model geliştirme sürecinde önemli bir kılavuz olarak kullanılabilir.

Naive Bayes(NB) Algoritması ile veri dengeleme yöntemlerinin uygulanması sonrası oluşan sonuçlar aşağıdaki Tablo 10'da gösterilmiştir.

Tablo 10. Dengesiz veri yöntemleri ile Naive Bayes(NB) modeli performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

ML Algoritm	**A**	**P 0**	**P 1**	**R 0**	**R 1**	**F1- 0**	**F1- 1**
ROS-NB	0.75	0.73	0.76	0.78	0.72	0.76	0.74
SMOTE-NB	0.73	0.74	0.73	0.72	0.75	0.73	0.75
RUS-NB	0.71	0.71	0.72	0.72	0.71	0.72	0.71
NM-NB	0.72	0.71	0.75	0.77	0.68	0.74	0.71

TL-NB	0,85	0,88	0,56	0,94	0,37	0,91	0,44

Tablo 10, dengesiz veri setlerine yönelik olarak uygulanan farklı yöntemlerin Naive Bayes (NB) modeli üzerindeki etkilerini ve bu yöntemlerin modelin performans ölçütlerindeki sonuçlarını içermektedir.

Analiz sonuçlarına göre, TL yöntemi ile iyileştirilmiş NB modeli, diğer yöntemlere göre Doğruluk ve F1 Skoru açısından dikkat çekici bir performans sergilemektedir. Bu yöntem, özellikle '0' sınıfı için yüksek Kesinlik ve Duyarlılık değerleri ile öne çıkmaktadır. Diğer taraftan, ROS, SMOTE, RUS ve NM yöntemleri ile iyileştirilmiş modeller, birbirine yakın ve tutarlı performans değerleri göstermektedir. Bu sonuçlar, kullanılan yönteme bağlı olarak NB modelinin dengesiz veri setleri üzerindeki tahmin başarısının farklılık gösterebileceğini ortaya koymaktadır.

Bu değerlendirme, dengesiz veri setleriyle çalışırken Naive Bayes modelinin performansını artırmak amacıyla kullanılabilecek yöntemlerin etkilerini gözler önüne sermektedir. Her bir yöntemin veri setinin özelliklerine ve modelin hedeflenen performans kriterlerine uygun şekilde seçilmesi gerektiği vurgulanmaktadır. Bu analiz, modelin geliştirilmesi ve iyileştirilmesi sürecinde önemli bir kaynak olarak değerlendirilebilir.

XGBoost(XGB) Algoritması ile veri dengeleme yöntemlerinin uygulanması sonrası oluşan sonuçlar aşağıdaki Tablo 11'de gösterilmiştir.

Tablo 11. Dengesiz veri yöntemleri XGBoost(XGB) modeli performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

ML Algoritm	**A**	**P 0**	**P 1**	**R 0**	**R 1**	**F1- 0**	**F1- 1**
ROS-XGB	0.95	0.96	0.94	0.94	0.97	0.95	0.95
SMOTE-XGB	0.94	0.93	0.95	0.94	0.94	0.94	0.94
RUS- XGB	0.54	0.52	0.56	0.76	0.31	0.62	0.40

NM-XGB	0.56	0.54	0.60	0.63	0.44	0.63	0.44
TL-XGB	0.87	0.88	0.71	0.97	0.35	0.92	0.47

Tablo 11, dengesiz veri setlerinin işlenmesi için uygulanan farklı yöntemlerin XGBoost (XGB) modelinin performans ölçütlerine olan etkilerini ortaya koyar.

Analiz sonuçları, ROS ve SMOTE yöntemleri ile iyileştirilmiş XGB modellerinin yüksek Doğruluk, Kesinlik, Duyarlılık ve F1 Skorları ile öne çıktığını göstermektedir. Bu iki yöntem, modelin her iki sınıf üzerinde dengeli ve yüksek bir başarı elde etmesini sağlamıştır. RUS ve NM yöntemleri, genel olarak daha düşük performans değerleri göstererek, bu tekniklerin modelin genel başarısını sınırlayabileceğini ortaya koymaktadır. TL yöntemiyle iyileştirilmiş model, özellikle '0' sınıfı için yüksek bir Kesinlik değeri sergilese de, '1' sınıfı için Duyarlılık ve F1 Skoru daha düşük kalmıştır.

Bu değerlendirme, dengesiz veri setleri üzerinde çalışırken XGBoost modelinin performansını artırmaya yönelik farklı yöntemlerin etkilerini kapsamlı bir şekilde gösterir. Yöntem seçimi, veri setinin yapısına ve modelin hedeflenen performans kriterlerine göre dikkatlice yapılmalıdır. Bu analiz, model geliştirme sürecinde önemli bir kaynak olarak kullanılabilir ve en uygun yöntemin seçilmesinde yardımcı olabilir.

AdaBoost hibrit modeli birlikte veri dengeleme yöntemleri uygulanması sonrası oluşan sonuçlar aşağıdaki Tablo 12'de gösterilmiştir.

Tablo 12. Dengesiz veri yöntemleri ile AdaBoost hibrit modelinin performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

ML Algoritm	A	P 0	P 1	R 0	R 1	F1- 0	F1- 1
ROS-AdaBoost	0.56	0.64	0.54	0.27	0.85	0.38	0.66
SMOTE- AdaBoost	0.62	0.74	0.58	0.37	0.87	0.49	0.70
RUS- AdaBoost	0.52	0.51	0.53	0.71	0.32	0.60	0.40
NM- AdaBoost	0.60	0.57	0.68	0.81	0.39	0.67	0.49

TL - AdaBoost	0.85	0.87	0.67	0.98	0.25	0.92	0.36

Tablo 12, dengesiz veri setlerine yönelik çeşitli yöntemler uygulanarak oluşturulan AdaBoost hibrit modelinin performans metriklerini içermektedir.

Analiz sonuçları göstermiştir ki, TL yöntemi ile geliştirilen AdaBoost modeli, Doğruluk, Kesinlik ve F1 Skoru açısından dikkate değer performans sergilemiştir, özellikle '0' sınıfı için elde edilen yüksek Kesinlik ve F1 Skoru bu durumu desteklemektedir. SMOTE ve NM yöntemleri de, '1' sınıfı için Duyarlılık değerlerinde iyileşmeler sağlamış ve modelin bu sınıfı doğru bir şekilde tanıma yeteneğini artırmıştır. Ancak, RUS yöntemi ile geliştirilen model, genel olarak daha düşük performans değerleri göstererek, bu teknikle yapılan iyileştirmelerin sınırlı kaldığını göstermektedir.

Bu analiz, AdaBoost modelinin dengesiz veri setleri üzerindeki performansını iyileştirmek için kullanılabilecek yöntemlerin etkilerini göstermektedir. Yöntem seçimi, veri setinin karakteristikleri ve modelin hedeflenen performans kriterlerine göre yapılmalıdır. Bu çalışma, model geliştirme ve iyileştirme süreçlerinde yol gösterici bir kaynak olarak değerlendirilebilir.

Tablo 5,6,7,8,9,10,11 ve Tablo 12'de gösterildiği gibi yapılan performans karşılaştırma ölçümleri sonucunda dengesiz veri yöntemi olarak ROS ile sınıflandırma modeli olarak ta Rastgele Orman Modeli seçilmiştir. Aşırı öğrenmenin olmaması için çapraz doğrulama yöntemi olarak k-Katmanlı(k-Fold) yöntemi de uygulanmıştır.

Dengesiz çalışan kaybı verisi ile geleneksel makine öğrenmesi sınıflandırma algoritmaları yöntemleri kullanılarak performans ölçütleri oluşturulmuştur. Sonuçlar detaylıca değerlendirilmiştir. Oluşan bu sonuçları iyileştirmek için veri ön işleme(pre-processing) ve sonrası işleme(post-processing) faaliyetleri uygulanmıştır.

Bir modelin Hiper parametreleri için sonuçları değerlendirmek için farklı yöntemler uygulanmıştır. Bu yöntemler sırasıyla, Manuel Arama, Izgara Araması (Grid Search CV), Rastgele Arama, Bayes Optimizasyonu, Genetik Algoritmalar ve Parçacık Sürü Optimizasyon oluşmaktadır.

Tablo 13. Dengesiz Veri Yöntemleri İle Rastgele Orman(Random Forest: RF), Ros ve Hiperparametre Yöntemlerinin Uygulanması

Model	Manuel Arama	Izgara Araması (Grid Search CV)	Rastgele Arama	Bayes Optimizasyonu	Genetik Algoritmalar	Parçacık Sürü Optimizasyon
ROS- RF	0.91	0.97	0.96	0.86	0.95	0.94

Tablo 13, Random Over Sampling (ROS) yöntemi ile zenginleştirilmiş Rastgele Orman (Random Forest: RF) modelinin, farklı hiperparametre optimizasyon tekniklerinin uygulanması sonucu elde edilen performans metriklerini sunmaktadır. Değerlendirilen optimizasyon teknikleri arasında Manuel Arama, Izgara Araması (Grid Search CV), Rastgele Arama, Bayes Optimizasyonu, Genetik Algoritmalar ve Parçacık Sürü Optimizasyonu bulunmaktadır. Performans, her bir teknik için modelin doğruluğu (accuracy) ile ölçülmüştür.

Tablodaki değerlendirme, her bir hiperparametre ayarlama yönteminin ROS yöntemi ile iyileştirilmiş RF modeli üzerindeki etkisini göstermektedir. Izgara Araması, modelin en yüksek doğruluk değerine ulaştığı yöntem olarak öne çıkmaktadır. Bu, Izgara Aramasının model için en uygun hiperparametre kombinasyonunu bulma konusunda etkili olduğunu göstermektedir. Rastgele Arama ve Genetik Algoritmalar da yüksek doğruluk değerleri ile dikkat çekmektedir, bu da bu yöntemlerin hiperparametre seçimi konusunda güçlü alternatifler sunabileceğini ortaya koymaktadır. Bayes Optimizasyonu ve Parçacık Sürü Optimizasyonu yöntemleri de rekabetçi doğruluk skorlarına ulaşmış, ancak Manuel Arama yöntemi, görece daha düşük bir performans sergilemiştir.

Bu analiz, Rastgele Orman modelinin performansını maksimize etmek için kullanılabilecek çeşitli hiperparametre optimizasyon tekniklerinin etkinliğini göstermektedir. Her bir yöntemin, modelin veri seti ve özel gereksinimleriyle uyumlu şekilde seçilmesi gerektiği vurgulanmaktadır. Bu çalışma, model optimizasyonu süreçlerinde önemli bir kaynak olarak kullanılabilir ve en uygun hiperparametre ayar yönteminin belirlenmesinde yardımcı olabilir.

Hiperparametre optimizasyonu sürecinde, Izgara Araması (Grid Search CV) gibi yöntemler aracılığıyla modellerin çeşitli parametre kombinasyonları üzerinde kapsamlı bir araştırma gerçekleştirilmekte ve bu süreç, en iyi model performansını sağlayacak parametre kombinasyonlarının belirlenmesine yardımcı olmaktadır. Bu çalışmada, sınıflandırma modelinin performansının artırılması amacıyla Izgara Araması (Grid Search CV) yöntemiyle

hiperparametre optimizasyonu gerçekleştirilmiş ve modelin başarımı, doğruluk, kesinlik, geri çağırma ve F1 skoru gibi metrikler aracılığıyla değerlendirilmiştir. Bu süreç, modelin farklı parametre kombinasyonlarının sistematik bir şekilde test edilmesini ve en uygun hiperparametre setinin belirlenmesini sağlamaktadır. Sonuçların analizi, belirlenen metrikler kullanılarak yapılmış ve modelin genel performansı bu ölçütler çerçevesinde incelenmiştir. Bu yaklaşım, modelin her bir yönünün performansının kapsamlı bir şekilde anlaşılmasını sağlamakta ve modelin farklı senaryolarda nasıl davranabileceği hakkında değerli bilgiler sunmaktadır.

ROS, Izgara Araması (Grid Search CV) ile Rastgele Orman Modeli ve Hiperparametre ayarlama yöntemleri kullanarak Tablo 14'teki gibi iyileştirme yapılmıştır.

Tablo 14. Dengesiz Veri Yöntemleri İle Rastgele Orman(RF), Ros ve Izgara Araması (Grid Search CV) modeli performans ölçütleri Doğruluk (Accuracy: A), Kesinlik (Precision: P), Duyarlılık (Recall: R) ve F1 Skoru (F1-score:F1)

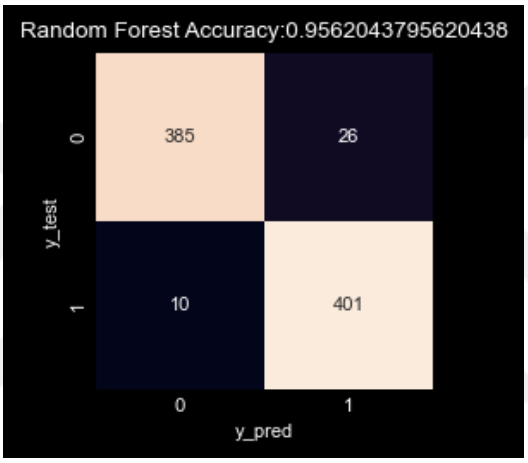
ML Algorithm	A	P 0	P 1	R 0	R 1	F1 0	F1 1
ROS- RF	0.97	0.97	0.94	0.94	0.98	0.96	0.96

Tablo 14, dengesiz veri setleri üzerinde Random Over Sampling (ROS) yöntemi kullanılarak geliştirilen ve Izgara Araması (Grid Search CV) ile hiperparametreleri optimize edilen Rastgele Orman (Random Forest, RF) modelinin performans metriklerini sunmaktadır. Performans ölçütleri arasında Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru (F1-score) bulunmaktadır. '0' ve '1', sınıflandırma modeli tarafından tahmin edilen iki farklı sınıfı ifade etmektedir. Kesinlik ve Duyarlılık, sırasıyla her bir sınıf için pozitif olarak tahmin edilen örneklerin gerçekte ne kadarının pozitif olduğunu ve gerçekte pozitif olan örneklerin ne kadarının doğru olarak pozitif olarak tahmin edildiğini gösterir. F1 Skoru ise, Kesinlik ve Duyarlılık değerlerinin harmonik ortalamasıdır ve dengeli bir performans ölçümü sağlar.

Modelin performansı yüksek bir Doğruluk skoru ile öne çıkmaktadır, bu da genel olarak modelin tahminlerinin yüksek bir doğrulukta olduğunu gösterir. Kesinlik ve Duyarlılık değerleri '0' ve '1' sınıfları için sırasıyla oldukça yüksektir, bu da modelin her iki sınıfı da başarılı bir şekilde tahmin edebildiğini işaret eder. Özellikle F1 skorları, her iki sınıf için de yüksek değerler almıştır, bu da modelin dengesiz veri setleri üzerindeki performansının etkili olduğunu göstermektedir.

Bu sonuçlar, dengesiz veri setleri üzerinde çalışırken Rastgele Orman algoritmasının ve ROS yöntemi ile birlikte hiperparametre optimizasyonu kullanımının etkili olduğunu göstermektedir. Özellikle, bu yöntemlerin ve algoritmanın pratik uygulamalarda, çalışan devri gibi karmaşık sınıflandırma problemlerinde başarılı sonuçlar verebileceği anlaşılmaktadır. Bu nedenle, makine öğrenimi modellerinin geliştirilmesi ve iyileştirilmesi süreçlerinde bu bulguların dikkate alınması önerilir.

Model sonuçlarının karışıklık(hata) matrisinde gösterimi ise Şekil 29'da gösterilmiştir. Gerçekte durumu çalışmaya devam eden olup modelin doğru tahmin ettiği çalışan sayısı 385, yanlış tahmin ettiği çalışan sayısı ise 26'dır. Aynı şekilde gerçekten işten ayrılan ve modelin doğru tahmin ettiği çalışan sayısı 401, yanlış tahmin ettiği çalışan sayısı 10'dur.

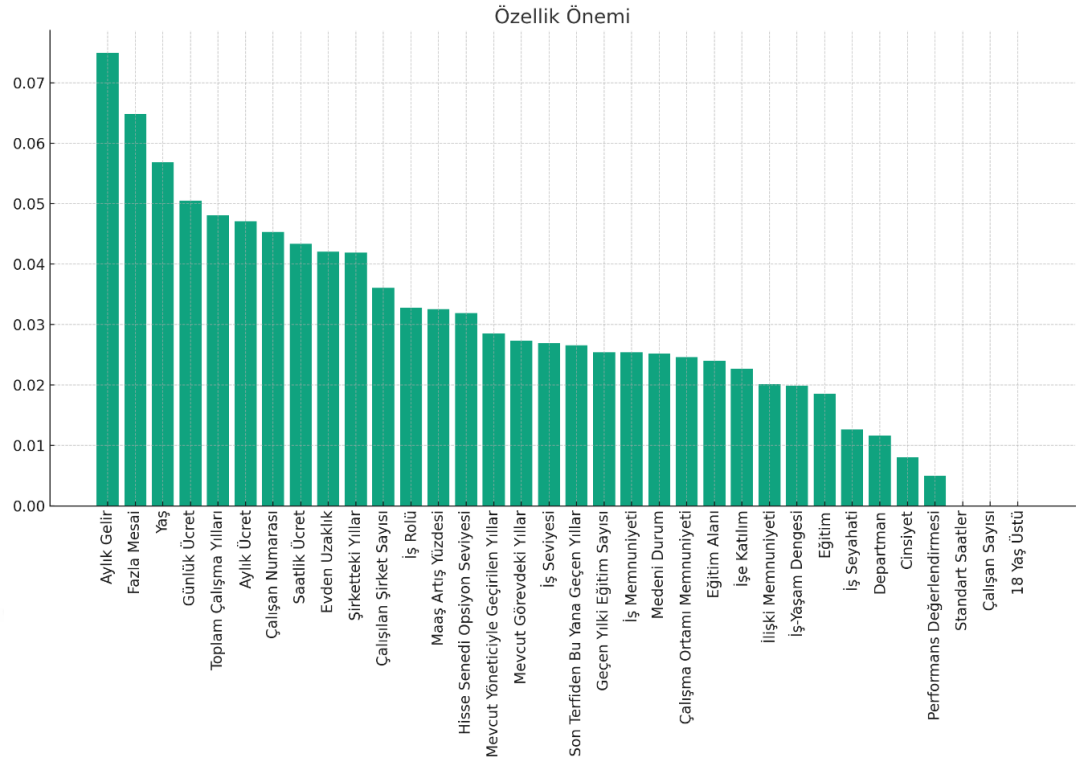


Şekil 29. Rastgele Orman Karışıklık(Hata) Matrisi

Tablo 15. En İyi Model Parametreleri

Model En İyi Parametreler	max_depth	min_samples_leaf	min_samples_split	n_estimators
ROS-RF	3	2	2	100

Şekil 30'da görselleştirilmiş özellik önemleri analizi, Rastgele Orman(Random Forest)modelinin tahmin sürecinde her bir özelliğin etkisini göstermektedir.



Şekil 30. Rastgele Orman(Random Forest) ile Özeliklerin Önemini Gösteren Grafik

Modelin karar verme mekanizmasında, bazı özelliklerin diğerlerine göre daha belirgin bir etkiye sahip olduğu görülmektedir. Bu, özelliklerin öneminin uzunlukları ile temsil edildiği çubuk grafiklerle ifade edilir. En uzun çubuklar, tahmin sürecinde en fazla bilgi sağlayan özelliklere işaret ederken, daha kısa çubuklar daha az etkili özellikleri temsil etmektedir.

Bu tür bir analiz, modelin hangi özellikler üzerine inşa edildiğini ve bu özelliklerin ayrılma tahminlerine olan etkilerinin ne derecede olduğunu anlamada yardımcı olur. Özellikle, modelin performansını iyileştirmek veya belirli özelliklere bağlı aşırı uyumu azaltmak amacıyla, hangi özelliklerin dikkate alınması gerektiği veya modifiye edilmesi gerektiği hakkında önemli bilgiler sunar.

Ayrıca, bu analiz, modelin karar verme sürecinin anlaşıla bilirliğini ve şeffaflığını artırmada kritik bir rol oynamaktadır. Karar verme mekanizmalarının açık bir şekilde sunulması, özellikle kararların sonuçlarının önemli olduğu durumlarda, modelin güvenilirliğini ve kabul edilebilirliğini güçlendirir.

Sonuç olarak, bu görselleştirme, makine öğrenimi modelinin nasıl çalıştığını ve hangi faktörlere dayandığını daha iyi kavramak için önemli bir araçtır. Özellikle karmaşık veri setlerinde, modelin hangi özelliklere odaklandığını ve bu özelliklerin öneminin neden

olduğunu anlamak, modelin etkinliğini ve doğruluğunu artırmak açısından önemlidir. Bu, modelin daha da geliştirilmesi ve uygulama alanına özgü ihtiyaçların daha iyi karşılanması için temel bir adımdır.

Bu tez araştırması, çalışmanın çeşitli makine öğrenimi modelleri kullanılarak çalışanların işten ayrılma olasılıklarının nasıl tahmin edilebileceğine dair kapsamlı bir analiz sunduğunu göstermektedir. Araştırma, özellikle dengesiz veri setleriyle başa çıkma yöntemleri, çalışan ayrılıklarını tahmin ederken yüksek performans gösteren modellerin seçimi ve etik kullanım üzerine odaklanmıştır. Bu çerçevede, Random Over Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE) ve çeşitli hiper-parametre ayarlama yöntemleri gibi dengesiz veri setlerini işleme yöntemleri incelenmiştir. Ayrıca, Rastgele Orman (Random Forest) gibi algoritmaların çalışan devri tahmininde yüksek doğruluk sağladığı belirlenmiştir.

Bu çalışma, organizasyonların çalışan devrini daha etkin bir şekilde yönetmelerine ve potansiyel personel kayıplarını önceden tahmin ederek önleyici stratejiler geliştirmelerine olanak tanıyan bulgular sunmaktadır. Makine öğrenimi algoritmalarının, çalışan devri tahminindeki etkinliğini göstermek ve kuruluşların bu tahminleri stratejik karar alma süreçlerine entegre etmeleri için bir yol haritası sunmak, tezin temel katkıları arasında yer alır. Gelecek çalışmalar, farklı sektörlerdeki organizasyonlar üzerinde benzer yöntemlerin uygulanabilirliğini ve etkinliğini test ederek, bu alandaki bilgi birikimini daha da genişletebilir.

Bu değerlendirme, makine öğrenimi tekniklerinin, çalışan ayrılıklarını önceden tahmin etme ve bu bilgileri kullanarak insan kaynakları stratejilerini geliştirme potansiyeline sahip olduğunu vurgulamaktadır. Dolayısıyla, bu tez, iş gücü yönetimi alanında önemli bir referans noktası olarak kabul edilebilir ve insan kaynakları yönetimi pratiklerinde makine öğrenimi teknolojilerinin entegrasyonunu teşvik edebilir.

7. SONUÇ

Bu tez çalışmasında birçok geleneksel makine öğrenmesi sınıflandırma yöntemi tahmin metodu incelenmiş ve her birinin kullanıldığı veri üzerinde göreceli olarak başarılı performans sergilediği gözlemlenmiştir. İncelenen çalışmalarda geleneksel sınıflandırma tahmin metotları ile bunların hibrit kullanılmasının performans sonuçları karşılaştırılmıştır. Bazı çalışmalar kurdukları modellerin eksik verisi üzerinde performans iyileştirme çalışmaları yapılmıştır. Bazı çalışmalar ise eksik veri sorununu giderip hibrit model oluşturarak performans iyileştirme yapmışlardır. Bu çalışmada, veri tiplerinin kendine has özelliklerine ve aralarındaki bağımlılıklara odaklanılmıştır. Bu bağlamda, çeşitli veri tiplerinin tahminleme sürecine ne derecede uygun olduğu detaylı bir şekilde incelenmiştir.

Bu araştırmada, IBM HR Analytics Employee Attrition & Performance tarafından sağlanan ve 1.470 çalışanı kapsayan bir İK veri seti üzerinde çalışılmıştır. Çalışan kaybı tahmin problemine yönelik olarak, sekiz farklı sınıflandırma modelinin uygulanması ve bu modellerin hesaplanan metrikler üzerinden değerlendirilmesi gerçekleştirilmiştir. Analiz kapsamında, her bir model için doğruluk, çapraz doğrulama skoru, kesinlik, duyarlılık ve F1-skor gibi metrik değerler hesaplanmıştır.

Bu çalışmada, geleneksel makine öğrenimi algoritmalarının eğitimi için veri setinin %75'i kullanılmış ve test için ise %25'i ayrılmıştır. Dengesiz veriler üzerine yapılan bu çalışmada, yıpranma değişkenine ilişkin evet ve hayır oranları dikkate alınarak, işten ayrılma durumlarının tahmini yapılmıştır. Hayır oranları yetersiz olan veri setinin dengelenmesi için, hem veri aşağı hem de yukarı dengeleme teknikleri uygulanmıştır. İlk aşamada, Rastgele Orman sınıflandırma algoritması yüksek F1-skore değerleriyle dikkat çekmiş olup, bu değerler 0.92 ve 0.19 olarak belirlenmiştir. Veri dengeleme tekniklerinin uygulanmasının ardından, Rastgele Orman Algoritması, ROS ve Izgara Araması (Grid Search) Modeli yöntemlerinin kullanılmasıyla elde edilen en yüksek performans değerleri, F1-skore olarak “0.96” ve “0.96” olarak tespit edilmiştir.

Bu tez çalışması kapsamında, veri ön işleme (pre-processing) ve son işleme (post-processing) süreçleri titizlikle yürütülmüştür. Son işleme aşamasında, aşırı öğrenmeyi önlemek amacıyla k-katmanlı çapraz doğrulama yöntemi uygulanmıştır. Ayrıca, performansın iyileştirilmesi için hiperparametre ayarlama süreçlerinden biri olan Izgara

Araması (Grid Search) Modeli kullanılarak en uygun model parametreleri belirlenmiş ve bu parametrelerle çalışmalar yürütülmüştür.

Bu çalışmanın sonuçlarına göre, çözülen iki problemin ve bu problemlere yönelik çözüm yöntemlerinin değerlendirilmesi neticesinde, gelecekteki sınıflandırma çalışmaları için bu tezde uygulanan modeller arasından en iyi performans metrik değerlerine sahip olan Rastgele Orman Sınıflandırması ile ROS modelinin kullanılması tavsiye edilmektedir.

Bu çalışma, sınıf dengesizliği sorununun üstesinden gelmek için güçlü ve yenilikçi bir yaklaşım sunmaktadır ve makine öğrenimi topluluğu için önemli bir katkı sağlamaktadır. Ayrıca daha büyük ve gerçek sektörel insan kaynağı eksik veri setleri kullanarak, yüksek CPU sahip bilgisayarlarda model parametre ayarlamaları ile daha iyi sonuçların elde edilebileceği önerilmektedir.

Bu çalışma kapsamında, kullanılan veri setinin özellikleri ve modelleme sürecinde benimsenen yaklaşımlar göz önünde bulundurulduğunda bazı kısıtların varlığından bahsedilebilir. İlk olarak, analiz sürecinde kullanılan veri seti, belirli bir sektöre ve coğrafi bölgeye özgüdür, bu da elde edilen bulguların genel geçerliği konusunda sınırlamalar getirmektedir. İkincil olarak, modelin eğitiminde kullanılan makine öğrenmesi tekniklerinin seçimi ve parametre ayarlamaları, çeşitli denemeler sonucunda belirlenmiş olup, farklı tekniklerin veya parametrelerin kullanılması durumunda farklı sonuçlar elde edilebileceği akılda tutulmalıdır. Son olarak, çalışmada dengesiz veri setlerinin işlenmesi için uygulanan yöntemler, veri dağılımındaki dengesizliği azaltmaya yönelik olmuş ve bu yöntemlerin seçimi, çalışmanın sonuçları üzerinde etkili olmuştur.

Elde edilen bulgular, çalışan devrinin tahmin edilmesinde makine öğrenmesi tekniklerinin potansiyelini ortaya koymaktadır. Rastgele Orman ve XGBoost gibi yöntemler, özellikle dengesiz veri setleri üzerinde etkili sonuçlar sunmuş, çalışan devrinin öngörülmesi konusunda önemli ipuçları sağlamıştır. Bu tekniklerin kullanımı, organizasyonların çalışan memnuniyeti ve sadakati konularında stratejik kararlar almasına yardımcı olabilir. Bununla birlikte, makine öğrenmesi modellerinin etik kullanımı, özellikle insan kaynakları yönetimi gibi hassas alanlarda dikkatle ele alınmalıdır.

Gelecek çalışmalarda, farklı sektörlerden ve coğrafi bölgelerden elde edilecek daha geniş ve çeşitli veri setleri kullanılarak modelin genel geçerliliğinin artırılması önerilmektedir. Ayrıca, dengesiz veri setlerini daha etkin bir şekilde işlemek için yeni yöntemlerin geliştirilmesi ve test edilmesi, modelin tahmin başarısını daha da iyileştirebilir. Son olarak,

makine öğrenmesi tekniklerinin etik ve sorumlu kullanımına yönelik daha kapsamlı araştırmalar yapılması, bu teknolojilerin insan kaynakları yönetimi alanındaki uygulamalarının sürdürülebilirliğini destekleyecektir.

Gelecekteki çalışmalarda, modelleme süreçlerinin iyileştirilmesine yönelik olarak daha gelişmiş modellerin kullanılmasına, farklı veri kaynaklarının entegre edilerek analizlerin zenginleştirilmesine ve analizlerin gerçek zamanlı olarak yapılabilmesi kapasitesinin artırılmasına odaklanılabilir. Bu yönelimler, elde edilen sonuçların daha geniş bir perspektiften değerlendirilmesini sağlayarak, modelin genel geçerliliğini ve uygulanabilirliğini artırma potansiyeline sahiptir. Aynı zamanda, çoklu veri kaynaklarının bir araya getirilmesi, farklı perspektiflerden elde edilen bilgilerin birleştirilmesini mümkün kılarken, gerçek zamanlı analiz yeteneği, organizasyonların dinamik iş dünyasında karşılaştıkları değişikliklere daha hızlı ve etkili bir şekilde yanıt vermelerini destekleyebilir. Bu genişletilmiş yaklaşım, çalışma alanına yeni ve değerli katkılar sunarak, ilgili literatürdeki mevcut boşlukların doldurulmasına yardımcı olabilir.

KAYNAKLAR

[1] Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inform. Allied Res. J. 4 (2013)

[2] Cockburn, I.; Henderson, R.; Stern, S. Yapay Zekanın İnovasyona Etkisi. Yapay Zeka Ekonomisinde: Bir Gündem ; University of Chicago Press: Chicago, IL, ABD, 2019; s. 115–146. [Google Akademik]

[3] Hong, W.-C., Pai, P.-F., Huang, Y.-Y., & Yang, S.-L. (2005). Application of Support Vector Machines in Predicting Employee Turnover Based on Job Performance. Advances in Natural Computation, 668–674. doi:10.1007/11539087_85

[4] Aylak, B. L., Oral, O., Yazıcı, K., Yapay Zeka ve Makine Öğrenmesi Tekniklerinin Lojistik Sektöründe Kullanımı, El-Cezerî Journal of Science and Engineering Vol: 8, No: 1, 2021 (74-93) DOI :10.31202/ecjse.776314

[5] Begum, M., Suneetha, T.V., Shreshta, C., Shivani, B., Nargees, , 2020, An Analysis of the Use of Machine Learning for Employee Attrition Prediction – A Literature Review, Usha,Published 2020, Computer Science, Business

[6] Diala, P., Wang, H., 2019, Prediction of Customer Churn in Telecommunication using Machine Learning Algorithms, School of Computer Science and Applied Mathematics University of the Witwatersrand, Johannesburg

[7] Danquah, R. A., 2020, Handlıng Imbalanced Data: A Case Study For Bınary Class Problems, Department of Mathematics Southern Illinois University Edwardsville, IL 62026 raddoda@siue.edu

[8] Zhao, Y., Hrnjic, E., & Murthy, S. (2021). Employee turnover prediction using machine learning. International Journal of Business Analytics, 8(1), 100-118.

[9] Rut, G., Singh, R., & Dixit, A. (2021). Employee turnover prediction using extreme gradient boosting. In 2021 International Conference on Computational Performance Evaluation (ComPE) (pp. 53-58). IEEE.

[10] Ma, Y., Zhao, Y., Liu, L., He, X., & Liu, J. (2020). A novel employee turnover prediction model by machine learning. In 2020 IEEE International Conference on Knowledge Graph (ICKG) (pp. 149-156). IEEE.

[11] Chaurasia, V., & Rosin, A. (2022). Deep learning for employee turnover prediction. IEEE Access, 10, 32447-32459.

[12] Araújo, M. C., Resende, P. A., Petri, G., Galdino, A. L., & Pereira, V. S. (2021). Machine learning for employee turnover prediction in small software companies. IEEE Access, 9, 19387-19398.

[13] Song, H., Zhu, Y., & Li, J. (2020). Ensemble machine learning approach for employee turnover prediction. IEEE Access, 8, 145561-145573.

[14] Further, J., & Kızılırmak, M. (2021). Employee turnover prediction with machine learning techniques: A review. IEEE Access, 9, 121439-121460.

[15] Amin, A., Shehzad, K., Khan, M., Ali, I., & Anwar, S. (2018). Density-based traffic signal control prediction using supervised machine learning techniques. In 2018

International Conference on Frontiers of Information Technology (FIT) (pp. 229-234). IEEE.

[16] Chitra, C., & Uma, V. (2020). Employee turnover prediction using naive bayes classifier. In 2020 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-5). IEEE.

[17] Jia, L., Huang, Z., & Xie, N. (2021). Employee turnover prediction based on decision tree models. In 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 693-698). IEEE.

[18] Kuruzovich, J., Krivenko, P., Zakharov, R., & Ershov, E. (2019). Human capital prediction using machine learning techniques. In 2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1-7). IEEE.

[19] Zhang, X., Zhao, Y., Huang, J., Liu, J., & Liu, L. (2021). Employee turnover prediction using gradient boosting machine. In 2021 3rd International Conference on Human-Computer Interaction (IHCI) (pp. 141-145). IEEE.

[20] Muazzam, A., Butt, M. A., Shah, S. A., & Mahmood, K. (2021). Deep learning for employee turnover prediction. In 2021 6th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (ISITIA) (pp. 1-5). IEEE.

[21] Aydın, D., Yıldız, O., & Aktaş, N. (2022). Employee turnover prediction with machine learning methods and cost-sensitive learning. Computational Intelligence and Neuroscience, 2022.

[22] Yao, S., Yang, F., Gao, W., & Xi, H. (2021). Employee turnover prediction based on neural network. In 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) (Vol. 5, pp. 2364-2369). IEEE

[23] Rong, G., Mendez, G., Assi, E. B., Bajic, B., & Chawla, N. V. (2020). Understanding SBML's Flat" Line: An Analysis of Loss Functions for Oversampling-based Classifiers in Imbalanced Datasets. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 5372-5379.

[24] Goel, S., Butun, E., Mogali, S., & Kar, U. (2022). Employee Turnover Prediction with Machine Learning and Rule-based Classifiers. International Journal of Data Science and Analytics, 13(1), 1-15.

[25] Dumitrescu, R., Calin, O., & Kalantarzadeh, S. (2019). Employee Turnover Prediction Using Support Vector Machines with Bayesian Hyper-parameter Optimization. In 2019 10th International Conference on Information and Communication Systems (ICICS) (pp. 56-61). IEEE.

[26] Lyu, Y., Zhang, C., Rios, K., Vuong, V., & Hoang, D. T. (2020). Employee turnover prediction with machine learning: A reliable approach. In Proceedings of the 4th SPRing-8 Symposium (pp. 1-7).

[27] Akyol, M., Fazlıoğlu, U., Vayvay, Ö., & Demirel, Ö. C. (2021). Automating employee turnover prediction using machine learning models. Journal of Enterprise Information Management, 35(1), 67-89.

[28] Şimşek, M., Daş, A. S. ,The Effect of Handling Imbalanced Datasets Methods on Prediction of Entrepreneurial Competency in University Students, 2022,

www.iceans.org

[29] Hamza, B., "Employee Churn Model w/ Strategic Retention Plan", Kaggle, https://www.kaggle.com/code/hamzaben/employee-churn-model-w-strategic-retention-plan/ (January, 20, 2023)

[30] Balcioglu, Y., Artar, M., 2022, Çalışanların İşten Ayrılma Olasılığının Makine Öğrenmesi İle Tahmini : K-En Yakın Komşu Algoritması İle, acedemia.com, 74145031

[31] Mohammed, A. J., Hassan, M. M., & Kadir, D. H. (2020). Improving classification performance for a novel imbalanced medical dataset using SMOTE method. International Journal, 9(3), 3161-3172.

[32] Benhar, H., Idri, A., & Fernández-Alemán, J. L. (2020). Data preprocessing for heart disease classification: A systematic literature review. Computer Methods and Programs in Biomedicine, 195, 105635.

[33] Altunbey Özbay, F., Özbay, E., Makine Öğrenmesi Algoritmalarının Kalp Yetmezliği Hastalarının Hayatta Kalma Tahmini Üzerindeki Performans Karşılaştırılması, 2. Medıterranean Scıentıfıc Research And Innovatıon Congress, 2022, Firat University, Faculty of Engineering, Department of Software Engineering, ORCID: 0000-0003-0629-6888 , ORCID: 0000-0002-9004-4802

[34] Oğuzlar, A., Veri Ön İşleme, Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Sayı: 21, Temmuz-Aralık 2003, ss. 67-76.

[35] POLATGİL, M., Veri Ölçekleme ve Eksik Veri Tamamlama Yöntemlerinin Makine, Öğrenmesi Yöntemlerinin Başarısına Etkisinin İncelenmesi, Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 11 (2023) 78-88, DOI:10.29130/dubited.948564

[36] Cunningham, P. , Kordon, M., Delany, S. J. , 2008, Denetimli Öğrenme. İçinde: Cord, M., Cunningham, P. (eds) Multimedya için Makine Öğrenimi Teknikleri. Bilişsel Teknolojiler. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2

[37] Sevli, O., Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması, Avrupa Bilim ve Teknoloji Dergisi, Sayı 16, S. 176-185, Ağustos 2019, www.ejosat.com ISSN:2148-2683

[38] ÖZLÜER BAŞER, B. , YANGIN, M., SARIDAŞ, E. S., Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü Dergisi, Cilt 25, Sayı 1, 112-120, 2021

[39] GÖK, M., Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi, Dergipark, GU J Sci, Part C, 5(3): 139-148 (2017)

[40] Gültepe, Y., Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine Karşılaştırmalı Bir Değerlendirme, Avrupa Bilim ve Teknoloji Dergisi Sayı 16, S. 8-15, Ağustos 2019

[41] PALA, M. A., ÇİMEN, M. E., BOYRAZ, Ö. F., YILDIZ, M. Z., BOZ, A. F., Meme Kanserinin Teşhis Edilmesinde Karar Ağacı Ve KNN Algoritmalarının Karşılaştırmalı Başarım Analizi, International Symposium on Innovative Technologies in Engineering and Science 22-24 November 2019 (ISITES2019 Şanliurfa - Turkey)

[42] PEKER, M., ÖZKARACA, O., KESİMAL, B., Enerji Tasarruflu Bina Tasarımı için

Isıtma ve Soğutma Yüklerini Regresyon Tabanlı Makine Öğrenmesi Algoritmaları ile Modelleme, BİLİŞİM TEKNOLOJİLERİ DERGİSİ, CİLT: 10, SAYI: 4, EKİM 2017

[43] Nizam, H., Akın, S. S., Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması, Türkiye'de İnternet Konferansı, 2019

[44] Zhou J., Li E., Wang M., Chen X., Shi X., Jiang L., Feasibility of Stochastic Gradient Boosting Approach for Evaluating Seismic Liquefaction Potential Based on SPT and CPT Case Histories, Journal of Performance of Constructed Facilities, 2019, 33: 04019024.

[45] Şimşek, H.; Demiral. Y., Aslan. Ö, Toğrul. B. Ü. Bir Üniversite Hastanesinde Koroner Kalp Hastalarına Uygulanan Tedavi Oranları, DEÜ Tıp Fakültesi Dergisi. 2012; 2, 111-117.

[46] Estevez, P.A.; Tesmer, M., Perez, C.A., Zurada, J.M. Normalized Mutual Information Feature Selection, IEEE Neural Networks. 2009; 20, 189-201.

[47] Bilgin, M. (2017). Gerçek veri setlerinde klasik makine öğrenmesi yöntemlerinin performans analizi. Breast, 2(9), 683.

[48] Kaya, Ç., Şimşek, M., 2023, Prediction of Employee Turnover with Imbalance Dataset Using Machine Learning Methods, International Journal Of Advanced Natural Sciences and Engineering Researches E-ISSN:2980-0811

[49] Kyanar, O., Yıldız, M., Görmez, Y., Albayrak, A., Makine Öğrenmesi Yöntemleri ile Duygu Analizi Sentiment Analysis with Machine Learning Techniques, International Artificial Intelligence and Data Processing Symposium (IDAP'16), September 17-18, 2016

[50] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

[51] Aydın, M. A. (2020). Müşteri Kaybı Tahmininde Sınıf Dengesizliği Problemi. Journal of Polytechnic. https://doi.org/10.2339/politeknik.734916

[52] Devi, D., Biswas, S. K., & Purkayastha, B. (2020). A Review on Solution to Class Imbalance Problem: Undersampling Approaches. 2020 International Conference on Computational Performance Evaluation (ComPE), 626-631. Shillong, India: IEEE https://doi.org/10.1109/ComPE49325.2020.9200087

[53] 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, C. 52. Association for Computing Machinery. https://doi.org/10.1145/3343440

[54] Turhan Sultan, Özkan, Y., Yürekli, B. S., Suner, A., & Doğu, E. (2020). Comparison of Ensemble Learning Methods for Disease Diagnosis in Presence of Class Unbalanced: Case of Diabetes. Turkiye Klinikleri Journal of Biostatistics, 12(1), 16-26. https://doi.org/10.5336/biostatic.2019-66816

[55] Liu, H., Yu, L., Efficient Feature Selection via Analysis of Relevance and

Redundancy,Journal of Machine Learning Research 5 (2004) 1205–1224

[56] Mohammed, A. J. (2020). Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method. International Journal of Advanced Trends in Computer Science and Engineering, 9(3), 3161-3172. https://doi.org/10.30534/ijatcse/2020/104932020

[57] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Içinde Journal of Artificial Intelligence Research (C. 16).

[58] Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Içinde Journal of Artificial Intelligence Research (C. 61).

[59] Harman, G. (2021). Destek Vektör Makineleri ve Naive Bayes Sınıflandırma

Algoritmalarını Kullanarak Diabetes Mellitus Tahmini. European Journal of Science and Technology. https://doi.org/10.31590/ejosat.1041186

[60] Yavaş, M., Güran, A., & Uysal, M. (2020). Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması. European Journal of Science and Technology, 258-264. https://doi.org/10.31590/ejosat.779952

[61] Raghuwanshi, B. S., & Shukla, S. (2021). Classifying imbalanced data using

SMOTE based class-specific kernelized ELM. International Journal of Machine Learning and Cybernetics, 12(5), 1255-1280. https://doi.org/10.1007/s13042-020-01232-1

[62] Durahim, A. O., Dengesiz Öğrenme İçin Örnekleme Tekniklerinin Karşılaştırılması, Yönetim Bilişim Sistemleri Dergisi , Cilt:1, Sayı:3, Yıl:2016, Sayfa 181-191, ISSN: 2148-3752, 2016

[63] Berrar, D., “Cross-validation”, Encyclopedia of Bioinformatics and Computational Biology 1 (542-545), 2019

[64] Göçgün, Ö. F., Onan, A., Amazon Ürün Değerlendirmeleri Üzerinde Derin Öğrenme/Makine Öğrenmesi Tabanlı Duygu Analizi Yapılması, Avrupa Bilim ve Teknoloji Dergisi Özel Sayı 24, S. 445-448, 2021

[65] EMEÇ1, M., ÖZCANHAN, M. H., Makine Öğrenmesi Algoritmalarında Hiper Parametre Belirleme, Mühendislik Öncü ve Çağdaş Çalışmalar, ORCID No: 0000-0002-9407-1728, 2023

[66] Koehrsen, W. (2018). Hyperparameter Tuning the Random Forest in Python. towardsdatascience.com. https://towardsdatascience.com/hyperparameter-tuningthe-random-forest-in-python-using-scikit-learn-28d2aa77dd74 adresinden 17.04.2023 tarihinde alınmıştır.

[67] Shekar, B.H.; Dagnew, G. Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. InProceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), Gangtok, India, 25–28 February 2019; pp. 1–8

[68] Lorenzo, P.R.; Nalepa, J.; Kawulok, M.; Ramos, L.; Ranilla, J. Particle swarm optimization for hyper-parameter selection in deep neural networks. In Proceedings of the Genetic and Evolutionary Computation Conference, Berlin, Germany, 15–19

July 2017.

[69] Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. J. Mach. Learn. Res. 2012, 13, 281–305

[70] Nguyen, V. Bayesian Optimization for Accelerating Hyper-Parameter Tuning. In Proceedings of the 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Sardinia, Italy, 3–5 June 2019; pp. 302–305.

[71] Hensman, J.; Fusi, N.; Lawrence, N. Gaussian processes for big data. In Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence, Bellevue, WA, USA, 11–13 July 2013; pp. 282–290.

[72] Man, K. F., Tang, K. S., & Kwong, S. (1996). Genetic algorithms: concepts and applications [in engineering design]. IEEE transactions on Industrial Electronics, 43(5), 519-534.

[73] Friedrich, T.; Kötzing, T.; Krejca, M.S.; Sutton, A.M. The Compact Genetic Algorithm is Efficient Under Extreme Gaussian Noise. IEEE Trans. Evol. Comput. 2017, 21, 477–490.

[74] Itano, F.; de Abreu de Sousa, M.A.; Del-Moral-Hernandez, E. Extending MLP ANN hyper-parameters Optimization by using Genetic Algorithm. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.

[75] Srinivas, M.; Patnaik, L. Adaptive probabilities of crossover and mutation in genetic algorithms. IEEE Trans. Syst. Man Cybern. 1994, 24, 656– 667.

[76] Smullen, D.; Gillett, J.; Heron, J.; Rahnamayan, S. Genetic algorithm with self-adaptive mutation controlled by chromosome similarity. In Proceedings of the 2014 IEEE Congress on Evolutionary Computation (CEC), Beijing, China, 6–11 July 2014; pp. 504–511.

[77] Yang, L.; Shami, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 2020, 415, 295–316.

[78] Lobo, F.; Goldberg, D.; Pelikan, M. Time Complexity of genetic algorithms on exponentially scaled problems. In Proceedings of the GECCO Genetic and Evolutionary Computation Conference, Las Vegas, NV, USA, 10– 12 July 2000.

[79] Shi, Y.; Eberhart, R.C. Parameter selection in particle swarm optimization. In Evolutionary Programming VII; Porto, V.W., Saravanan, N., Waagen, D., Eiben, A.E., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 591–600.

[80] Chuan, L.; Quanyuan, F. The Standard Particle Swarm Optimization Algorithm Convergence Analysis and Parameter Selection. In Proceedings of the Third International Conference on Natural Computation (ICNC 2007), Haikou, China, 24–27 August 2007; Volume 3, pp. 823–826.