



**TEKSTİL SEKTÖRÜNDE İŞ KAZALARINA NEDEN
OLAN RİSK FAKTÖRLERİNİN VERİ MADENCİLİĞİ
YÖNTEMLERİYLE DEĞERLENDİRİLMESİ**

Büşra TUNÇMAN



T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**TEKSTİL SEKTÖRÜNDE İŞ KAZALARINA NEDEN OLAN RİSK
FAKTÖRLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİYLE
DEĞERLENDİRİLMESİ**

Büşra TUNÇMAN
0009-0005-8361-1708

Prof. Dr. Tülin GÜNDÜZ
(Danışman)

Doç. Dr. Duygu YILMAZ EROĞLU
(Danışman)

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA – 2024
Her Hakkı Saklıdır

ÖZET

Yüksek Lisans Tezi

TEKSTİL SEKTÖRÜNDE İŞ KAZALARINA NEDEN OLAN RİSK FAKTÖRLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİYLE DEĞERLENDİRİLMESİ

Büşra TUNÇMAN

Bursa Uludağ Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Tülin GÜNDÜZ

İkinci Danışman: Doç. Dr. Duygu YILMAZ EROĞLU (Bursa Uludağ Üniversitesi)

Gelişen ve büyüyen iş yerlerinde iş kazaları gün geçtikçe daha kritik bir konu haline gelmektedir. Veri madenciliği, artan iş kazalarına neden olan faktörlerin belirlenmesi ve analiz edilmesinde kullanılabilecek etkili bir yöntemdir. Bu çalışmanın amacı, veri madenciliği yöntemleri kullanılarak tekstil sektöründe iş kazalarının kontrolüne yönelik risk faktörlerinin araştırılması ve analiz edilmesidir. Bu çalışmada 2019-2021 yılları arasında tekstil sektöründe meydana gelen 89.963 iş kazası verisi incelenmiştir. Tahmin sonucuna en çok etki eden özelliklerin sırasıyla kaza sebebi, kaza anında kullanılan araç/metaryel, alt sektör bilgisi ve firma büyüklüğü olduğu görülmüştür. Kaza sonucu yaralanma türlerini tahmin etmek için SVM, Ekstra Ağaçlar, Rastgele Orman, Gradyan Artırma ve XGBoost algoritmaları kullanılmıştır. Veri dengeleme ve parametre optimizasyonu yöntemleri ile algoritmaların tahmin performansı artış göstermiştir. Makro F-skor değeri %70 olarak hesaplanan XGBoost algoritması diğer algoritmalara göre daha iyi performans göstermiştir. SVM algoritması (%69) ve Ekstra Ağaçlar (%68) algoritması, yüksek makro F-skor değerlerine ulaşarak veri setini doğru yorumlayan algoritmalar arasında yer almıştır. Sonuç olarak, veri madenciliği adımlarıyla oluşturulan bir kaza tahmin modelinin, tekstil sektöründe meydana gelebilecek önemli riskleri belirleme konusunda kullanılabileceği anlaşılmıştır.

Anahtar Kelimeler: Tekstil Sektörü, İş Kazaları, İş Güvenliği, Veri Madenciliği, Veri Dengeleme ve Hiperparametre Optimizasyonu

2024, viii + 48 sayfa.

ABSTRACT

MSc Thesis

EVALUATION OF RISK FACTORS CAUSING OCCUPATIONAL ACCIDENTS IN THE TEXTILE SECTOR USING DATA MINING METHODS

Büşra TUNÇMAN

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

Supervisor: Prof. Dr. Tülin GÜNDÜZ

Second Supervisor: Doç. Dr. Duygu YILMAZ EROĞLU (Bursa Uludağ University)

In developing and growing workplaces, occupational accidents are increasingly becoming a more critical issue with each passing day. Data mining is an effective method that can be used to identify and analyze the factors that cause increased occupational accidents. This study aims to investigate and analyze the risk factors for the control of occupational accidents in the textile sector using data mining methods. In this study, 89.963 occupational accident data that occurred in the textile sector between the years 2019-2021 were examined. It has been seen that the features that have the most effect on the forecast result are cause of accident, material agent, sub-sector, and company size, respectively. SVM, Extra Trees, Random Forest, Gradient Boosting, and XGBoost algorithms have been used for the prediction of accidental injury types. The estimation performance of algorithms has increased with data balancing and parameter optimization methods. The XGBoost algorithm, whose macro F-score value was calculated as 70%, performed better than other algorithms. The SVM algorithm (69%) and Extra Trees (68%) algorithm have been among the algorithms that correctly interpreted the data set by reaching high macro F-score values. As a result, it has been understood that an accident prediction model created by data mining steps can be used to identify significant risks that may occur in the textile industry.

Key words: Textile Sector, Occupational Accidents, Work Safety, Data Mining, Data Balancing and Hyperparameter Optimization

2024, viii + 48 pages.

ÖNSÖZ ve/veya TEŞEKKÜR

Bu tez çalışmamın oluşma sürecinde, yardımlarını hiçbir zaman esirgemedi, yönlendirici, anlayışlı ve potansiyelimi ortaya çıkararak beni teşvik eden çok değerli danışmanlarım Prof. Dr. Tülin Gündüz ve Doç. Dr. Duygu Yılmaz Eroğlu' na sonsuz teşekkürlerimi sunarım.

Hem kariyerimde hem de bu tez çalışmasında değerli fikirleriyle bana yol gösteren ve desteklerini esirgemeyen tüm yöneticilerime ve çalışma arkadaşlarıma teşekkürlerimi sunarım.

Hayatımın her alanında desteğini benden esirgemeyen ve her zaman yanımda olduklarını bildiğim aileme gönülden teşekkürlerimi sunarım.

Büşra TUNÇMAN
.../.../.....

İÇİNDEKİLER

	Sayfa
ÖZET.....	i
ABSTRACT.....	ii
ÖNSÖZ ve/veya TEŞEKKÜR.....	iii
SİMGELER ve KISALTMALAR DİZİNİ.....	vi
ŞEKİLLER DİZİNİ.....	vii
ÇİZELGELER DİZİNİ.....	viii
1. GİRİŞ.....	1
2.1. Veri Madenciliği.....	8
2.2. Veri Madenciliğinin Tarihi.....	8
2.3. Veri Madenciliği Süreci.....	9
2.3.1. Problemin tanımlanması.....	9
2.3.2. Verilerin hazırlanması.....	9
2.3.3. Modelin kurulması ve değerlendirilmesi.....	10
2.3.4. Modelin kullanılması.....	10
2.3.5. Modelin izlenmesi.....	10
2.4. Veri Madenciliği Teknikleri.....	10
2.4.1. Sınıflandırma ve regresyon.....	11
2.4.2. Birliktelik kuralları.....	11
2.4.3. Kümeleme.....	11
2.4.4. Veri temizleme.....	12
2.4.5. Veri görselleştirme.....	12
2.4.6. Sinir ağları.....	13
2.4.7. Aykırı değer tespiti.....	13
2.4.8. Tahmin.....	13
2.4.9. Makine öğrenmesi.....	14
2.5. Makine Öğrenmesi Algoritmaları.....	15
2.5.1. Denetimli öğrenme.....	16
2.5.2. Denetimsiz öğrenme.....	18
2.5.3. Pekiştirmeli öğrenme.....	19
2.5.4. Sinir ağları ve derin öğrenme.....	19
2.5.5. Topluluk öğrenme.....	20
2.6. Makine Öğrenmesi Algoritmaları İçin Hiperparametre Optimizasyonu.....	22
2.6.1. Hiperparametre optimizasyonu yöntemleri.....	23
2.7. Sınıf Dengesizliği Problemi İçin Kullanılan Algoritmalar.....	23
2.7.1. RUS (Rastgele az örnekleme).....	24
2.7.2. ROS (Rastgele aşırı örnekleme).....	24
2.7.3. SMOTE (Sentetik azınlık aşırı örnekleme).....	24
2.7.4. NearMiss.....	25
2.8. Tahmin Doğruluğu Yaklaşımı / Performans Ölçütleri.....	25
3. MATERYAL ve YÖNTEM.....	27
3.1. Süreç Özeti.....	27
3.1.1. Veri seti bilgileri / veri toplama.....	28
3.1.2. Veri ön işleme.....	30
4. BULGULAR ve TARTIŞMA.....	34
4.1. Modelin Oluşturulması.....	35
4.1.1. Veri dengeleme yöntemlerinin değerlendirilmesi.....	35

4.1.2. Hiperparametre optimizasyonu yönteminin değerlendirilmesi	37
4.1.3. Model performanslarının değerlendirilmesi.....	37
5. SONUÇ.....	43
KAYNAKLAR	45
ÖZGEÇMİŞ	48



SİMGELER ve KISALTMALAR DİZİNİ

Kısaltmalar	Açıklama
GA	Genetik Algoritma
GSMH	Gayrisafi Millî Hasıla
ILO	Uluslararası Çalışma Örgütü
İSG	İş Sağlığı ve Güvenliği
K-NN	K-En Yakın Komşu/K-Nearest Neighbors
ML	Makine Öğrenmesi/Machine Learning
PSO	Parçacık Sürü Optimizasyonu
ROS	Rastgele Aşırı Örnekleme/Random Over Sampler
RUS	Rastgele Az Örnekleme/Random Under Sampler
SGK	Sosyal Güvenlik Kurumu
SMOTE	Sentetik Azınlık Aşırı Örnekleme Tekniği/Synthetic Minority Oversampling Technique
SVC	Destek Vektör Sınıflandırıcısı/Support Vector Classifier
SVM	Destek Vektör Makinesi/Support Vector Machine
XGBoost	Ekstrem Gradyan Arttırma/eXtreme Gradient Boosting
YSA	Yapay Sinir Ağları

ŞEKİLLER DİZİNİ

	Sayfa
Şekil 1.1. İş kazası istatistikleri.....	1
Şekil 2.1. Veri madenciliği temel alanları.....	8
Şekil 2.2. Yapay zeka ve alt dalları.....	14
Şekil 2.3. Makine öğrenmesi algoritmaları	15
Şekil 2.4. XGBoost algoritmasının yıllar içerisindeki gelişimi	21
Şekil 3.1. Veri madenciliği proses haritası.....	27
Şekil 3.2. Değişken etkinlik grafiği.....	32
Şekil 3.3. Korelasyon Matris (2019-21).....	33
Şekil 3.4. Korelasyon Matris (2019-20 Mart).....	33
Şekil 4.1. Model seçim haritası	35
Şekil 4.2. SVM sınıflandırma raporu	38
Şekil 4.3. SVM karmaşıklık matrisi	38
Şekil 4.4. XGBoost sınıflandırma raporu.....	38
Şekil 4.5. XGBoost karmaşıklık matrisi	38
Şekil 4.6. Gradyan artırma sınıflandırma raporu	39
Şekil 4.7. Gradyan artırma karmaşıklık matrisi	39
Şekil 4.8. Rastgele orman sınıflandırma raporu.....	39
Şekil 4.9. Rastgele orman karmaşıklık matrisi.....	39
Şekil 4.10. Ekstra ağaç sınıflandırma raporu	39
Şekil 4.11. Ekstra ağaç karmaşıklık matrisi	39

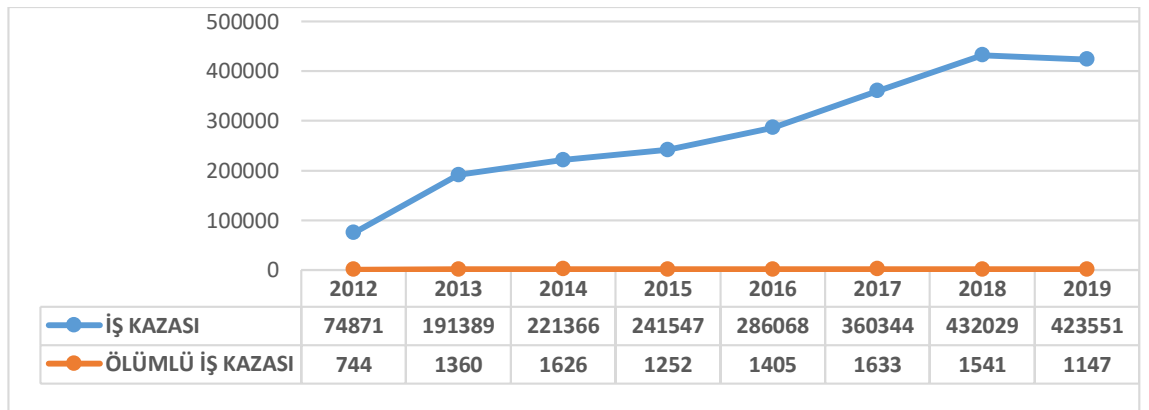
ÇİZELGELER DİZİNİ

	Sayfa
Çizelge 2.1. Karmaşıklık matrisi.....	25
Çizelge 2.2. Sınıflandırma metrikleri.....	26
Çizelge 3.1. Veri seti bilgileri	29
Çizelge 3.1. Veri seti bilgileri (devam).....	30
Çizelge 4.1. Ön işleme sonrası özelliklerin durumu	34
Çizelge 4.2. Veri dengeleme yöntemlerinin başarı metriklerine etkisi.....	36
Çizelge 4.3. Hiperparametre optimizasyonu yönteminin başarı metriklerine etkisi	37
Çizelge 4.4. SVM modelinin yara türlerini tahmin etme başarısı.....	40
Çizelge 4.5. XGBoost modelinin yara türlerini tahmin etme başarısı	41
Çizelge 4.6. Gradyan Artırma modelinin yara türlerini tahmin etme başarısı	41
Çizelge 4.7. Rastgele orman modelinin yara türlerini tahmin etme başarısı	41
Çizelge 4.8. Ekstra ağaç modelinin yara türlerini tahmin etme başarısı.....	42

1. GİRİŞ

İş yerinde, güvenli ve sağlıklı bir çalışma ortamı temel ilke ve haktır. Çalışma hayatında dikkat edilmesi gereken en önemli husus çalışanın sağlığını riske atmayacak güvenli bir çalışma ortamı oluşturmaktır. Bu kapsamda, tehlikeli bir durum veya davranış sonucu meydana gelen iş kazaları ve meslek hataları en büyük problemdir. Güvenli ve sağlıklı bir çalışma ortamı için; hükümetler, işverenler ve işçiler tanımlanmış hak ve sorumluluklarını bilmek ve uygulamakla yükümlüdür. Küresel ölçekte önem arz eden ISG kurallarının temel amacı, iş kazası ve meslek hastalıklarının meydana gelmesini önlemek için en uygun önleyici ve koruyucu tedbirlerin alınmasıdır.

İşin geleceğine insan merkezli yaklaşımı teşvik etme vizyonu ile çalışan Uluslararası Çalışma Örgütü (ILO), iş kazalarını işten kaynaklanan veya iş sırasında ortaya çıkan, ölümcül veya ölümcül olmayan yaralanmayla sonuçlanan olaylar olarak tanımlamaktadır. İş kazaları ve hastalıkları; çalışanlar, işverenler ve ülke ekonomileri üzerinde yıkıcı etkilere yol açmaktadır. Dünya genelinde yılda 270 milyon iş kazası meydana gelmekte ve bu kazalar nedeni ile yaklaşık 360 bin kişi ölmektedir. ILO tarafından paylaşılan analizler, iş kazası ve meslek hastalıklarının küresel GSMH' ın yüzde dördlük kısmının kaybına neden olduğunu göstermektedir (Santos, 2021). Sosyal Güvenlik Kurumu'nun (SGK) kayıtlarına göre ülkemizde 2012 senesinde 74.871 olarak kayda geçen iş kazası verisi 2019 yılında 423.551'e ulaşmıştır (Şekil 1.1.).



Şekil 1.1. İş kazası istatistikleri

Yapılan arařtırmalara g re, iř kazalarına baėlı yaralanmalar nedeniyle acil servislere yapılan bařvurularda ilk sırada %30,1'lik oran ile metal sanayi ve makine kaynaklı sekt rler yer alırken ikinci sırada ise %28,7 oranla tekstil sekt r  yer almaktadır (Serinken ve ark. 2012). Diėer bir arařtırmada ise iř kazalarının nedenleri daha geniř k mede incelenmiř ve tehlikeli davranıř oranı %68,1, tehlikeli durum oranı %11,6 olarak hesaplanmıřtır. Tehlikeli davranıř oranının tehlikeli duruma g re olduk a y ksek olması kazaların k   k iyileřtirmeler ile b y k oranda  nlenebilir olduėunu g stermektedir (G ll oėlu ve Ta gın, 2018).

İř kazaları neticesinde oluřan maddi ve manevi kayıplar, fabrikalardaki  retim artıřıyla doėru orantıda artıř g stermektedir. Hem d nya genelinde hem de kendi  lkemizde sanayileřmenin artması ve teknolojik geliřmeler  zellikle iřyerlerindeki  alıřanların saėlık ve g venlikleri ile ilgili bazı sorunların ortaya  ıkmasına neden olmuřtur. Bu s re te yapılan  alıřmalar neticesinde  alıřma ortam ve kořullarına y nelik birtakım kurallar ve kanunlar d zenlenmiřtir ( akmak, 2019). T rkiye'de ilk b y k iyileřtirme giriřimi 2012 yılında iř saėlıėı ve g venliėi ile ilgili ayrı bir kanun  ıkarılarak yapılmıřtır (Recal ve Demirel, 2021). İřyerlerindeki iř saėlıėı ve g venliėinin saėlanması ve mevcut saėlık ve g venlik řartlarının iyileřtirilmesi i in iřveren ve  alıřanların g rev, yetki, sorumluluk, hak ve y k ml l kleri İř Saėlıėı ve G venliėi Kanunu ile netleřtirilmiřtir. Bu kanun kapsamında iřverenin, mesleki risklerin  nlenmesi ve bu risklerden korunulmasına y nelik  alıřmaları da kapsayacak, iř saėlıėı ve g venliėi hizmetlerinin sunulması i in gerekli d zenlemeleri yapması mecburi hale gelmiřtir.

Artan iř kazalarının  nlenmesi i in mevzuat  ıkarma atılımlarına ek olarak İSG alanında daha geniř arařtırma  alıřmalarına ihtiya  bulunmaktadır. İř kazalarının  nlenmesinde doėru kararlar alınabilmesi i in, iř kazalarının oluřumundaki etkenlerin belirlenmesi ve karar alıcılara sunulması  nemli bir rol oynamaktadır. Bu etkenlerin belirlenmesinde kullanılabilecek en  nemli kaynaklardan birisi ulusal  apta toplanan iř kazaları verisidir. Ancak, kazaları tanımlayan  ok sayıda parametrelerle birlikte toplanan veri boyutunun her ge en g n b y mesi, geleneksel istatistiksel y ntemlerin hesaplama y k n  arttırmıř ve analizlerin uygulanmasını zorlařtırmıřtır. Bu nedenle, b y k veri analizinde veri madenciliėi tekniėi, karmařık bir veri k mesinden deėerli bilgileri keřfetmek i in

istatistiksel yöntemleri yapay zekâ teknikleriyle birleştirerek geleneksel istatistiksel yöntemlere güçlü bir alternatif oluşturmaktadır (Recal, 2022).

Veri boyutu oldukça yüksek olan iş kazalarını analiz etmek veri madenciliği tekniği ile birlikte daha kolay hale gelmiştir. Bu durum, iş kazalarının analiz edildiği çalışmaların literatürdeki payının gün geçtikçe büyümesine neden olmuştur. Bu alandaki çalışmalar da çoğunlukla ulusal ya da uluslararası boyutta iş kazası verileri analiz edilmiştir. Sektörel bazda ise ağır sanayi olarak geçen madencilik, petrokimya ve metal sektöründeki iş kazaları çalışma konusu olmuştur. Sektörel bazdaki çalışmalara örnek olarak Ching-Wu Cheng ve ark. (2022), iş kazalarının analizi için Tayvan'daki petrokimya endüstrisinde 2000 ve 2010 yılları arasında meydana gelen büyük boyuttaki 349 iş kazası vakalarını toplamıştır. Toplanan veriler istatistik paket programı kullanılarak analiz edilmiştir. Kaza faktörlerinin dağılımını ve kurallarını incelemek için veri madenciliği yöntemlerinden sınıflandırma ve regresyon ağacı (CART) kullanılmıştır. Sonrasında, çeşitli değişkenler arasındaki ilişkiyi açıklamak için bir korelasyon katsayısı analizi yapmak üzere bir Ki-kare testi ve Cramer istatistiği (Vc) kullanılmıştır. Literatürdeki bazı çalışmalarda, farklı sektörlerde meydana gelen iş kazaları aynı anda incelenmektedir. Gül ve ark. (2016), Türkiye'deki farklı sektörlerde meydana gelen 234 vakayı içeren iş kazası veri setini iki amaç üzerine analiz etmiştir. İlk aşamada, iş kazası geçiren çalışanların özellikleri ve kaza bilgileri kümeleme algoritması ile analiz edilmiş ve çalışanlara yapılan yatırımın verimliliğini en üst düzeye çıkaracak yönetim şekli araştırılmıştır. İkinci aşamada ise çalışanların özellikleri ile iş kazaları arasında herhangi bir ilişki olup olmadığını varsayan hipotezler ikili karşılaştırma matrislerinde değerlendirilmiştir. Çalışma da açık kaynaklı veri madenciliği programı kullanılmıştır.

Çakır (2019), ticaret gemilerinde meydana gelen iş kazalarını incelediği çalışmasında veri madenciliği yöntemi yardımıyla yaralanma şiddetine etki eden faktörleri ortaya koymuştur. Çalışmada, Tele Sağlık Merkezinde kayıtlı olan kaza verileri içerisinde 2013-2017 yılları arasında gemi adamlarını ilgilendiren 4690 vaka seçilmiştir. Çalışma da üç farklı veri madenciliği yöntemi kullanılmıştır. İlk aşamada birliktelik kuralları yöntemi kullanılarak kaza değişkenler arasındaki ilişkiler incelenmiş ve kaza örüntüleri oluşturulmuştur. İkinci aşama da yaralanma şiddeti ile diğer değişkenler arasındaki

ilişkiler analiz edilmiş ve en önemli değişkenin kaza esnasında yapılan iş olduğu sonucuna varılmıştır. Son aşamada ise lojistik regresyon analizi ile bağımlı ve bağımsız değişkenler arasındaki marjinal etkilerin ortaya çıkarılması hedeflenmiştir. Birliktelik kuralları yöntemi, son zamanlarda iş kazalarının analizi üzerine yapılan çalışmalarda sıklıkla görülmektedir. Bu çalışmalara diğer bir örnek ise, Çakmak (2019) tarafından geliştirilmiştir. Çalışmada, 2016 ve 2017 yıllarında SGK tarafından paylaşılan kaza verilerine ilişkin istatistiki bilgiler kullanılmıştır. Veri madenciliği yöntemleri ile, iş kazalarının meydana gelmesinde birbiriyle ilişkili faktörler belirlenmiştir. Çoğunlukla pazar araştırmalarında kullanılan birliktelik kuralları analizinin iş kazalarının analizi ile ilgili araştırmalarda da kullanılabileceği gösterilmiştir. İş kazalarının oluşumunda sıklıkla kümelenme eğilimi gösteren faktörler için anlamlı kurallar oluşturulmaya çalışılmıştır.

Ayanoğlu ve Kurt (2019), metal sektörü işyerlerine ait kaza verilerini kullanarak, veri madenciliği yöntemi (yapay sinir ağı) ile bir kaza tahmin modeli geliştirmiştir. Bu çalışmada, metal ürünleri imalatında faaliyet gösteren işyerlerinde, 2012-2015 yılları arasında meydana gelen iş kazaları veri seti kullanılmıştır. Çıktı değişkeni kaza şiddeti (hafif yaralanma, yaralanma, uzuv kaybı ve ölümlü kaza) olarak belirlenmiştir. Kaza verileri üzerinde gerekli düzenlemeler yapıldıktan sonra, açık kaynaklı veri madenciliği programı yardımı ile makine öğrenim algoritmalarının performansı izlenerek, en yüksek tahmin doğruluk oranını veren algoritma (Yapay Sinir Ağları) belirlenmiştir. Sonraki aşamada veri setini öğretmek için çift katmanlı ileri beslemeli bir Kaza Tahmin Modeli Yapay Sinir Ağı geliştirilmiştir. Geliştirilen modelin veri kümesi içerisinde yer alan kaza olayları ile denemesi sonucunda, %90 oranında doğrulukla çalıştığı ortaya konulmuştur.

İş kazalarının veri madenciliği yöntemleri ile analiz edildiği çalışmalarda, çıktı değişkeni çalışmanın amacı doğrultusunda farklı şekillerde ifade edilmiştir. Kaza sonucu, kaza şiddeti yaralanma tipi, yaralanan uzuv bu ifadelerden bazılarıdır. Literatürdeki bazı çalışmalarda ise kaza şiddeti, iş günü kaybı süreleri üzerinden değerlendirilmiştir. Bu duruma örnek olarak Recal (2022), 2013-2017 yılları arasında Türkiye’de meydana gelen 187.910 ulusal kaza verisini, iş kazalarını önlemede etkili karar alıcılara sunmak için karar destek sistemine dönüştürmeyi hedefleyen bir çalışma hazırlamıştır. Çalışmada, makine öğrenmesi yöntemi ile kazaların şiddeti üzerinde etkili olan faktörlerin

belirlenmesi ve belirlenen faktörlerin oluşturduğu kaza zincirlerinin birliktelik kural çıkarımı yöntemi ile ortaya çıkarılması olmak üzere iki aşamalı bir veri madenciliği yöntemi geliştirilmiştir.

Choi ve ark. (2020), inşaat sektöründeki endüstriyel kaza verisini lojistik regresyon analizi, karar ağacı analizi, rastgele orman analizi ve AdaBoost yöntemleri ile analiz ederek ciddi kaza risk gruplarını sınıflandırmak için bir tahmin modeli oluşturmuştur. Her bir modelin performansları karşılaştırılmış ve dengesiz iki terimli sınıflandırmaları tahmin etmede en başarılı sonucu rastgele orman analizi (%91,98) sağlamıştır. Benzer şekilde Koç ve ark. (2021), inşaat işçilerinin kaza sonrası sakatlık durumunu tahmin etmek için Rastgele Orman, XGBoost, AdaBoost ve Ekstra Ağaç olmak üzere 4 farklı ağaç tabanlı topluluk makine öğrenimi yöntemlerini kullanmışlardır. Genetik algoritma tabanlı XGBoost en iyi performansı göstermiştir.

Kakhki ve ark. (2019), tarım endüstrisindeki yaralanma şiddeti sonuçlarını sınıflandırmak için SVM, Boosted Trees, and Naive Bayes olmak üzere üç makine öğrenimi yöntemi uygulamıştır. Model değerlendirmesinde F-skor metriği kullanmışlardır. Tüm metrikler hesaplandığında, destek vektör makineleri, diğer tüm modellerden daha iyi performans göstermiştir. Bir sınıflandırıcının performansını değerlendirmek için kullanılan F skor metriği, kesinlik ve hatırlamanın harmonik ortalaması olduğu için en yararlı ölçütlerden biridir (Kakhki, 2019). Khairuddin ve ark. (2022), mesleki yaralanma alanında daha doğru bir tahmin modeli olarak toplu öğrenmenin potansiyelini vurgulamıştır. mesleki yaralanma ciddiyetini tahmin etmek için özelliklere göre optimize edilmiş bir model önermiştir. OSHA'dan alınan 66.405 mesleki yaralanma veri seti, Destek Vektör Makinesi, K-En Yakın Komşular, Naive Bayes, Karar Ağacı ve Rastgele Orman yöntemleri kullanılarak analiz edilmiştir. Model karşılaştırması için Rastgele Orman, daha yüksek doğruluk ve F1 skor değeri ile diğer yöntemlere göre daha başarılı olmuştur.

İş kazaları üzerine yapılan çalışmalar genellikle kaza sonuçlarının tahmini üzerinedir. Sınıflandırma algoritmalarında parametre optimizasyonu modelin tahmin doğrulu açısından oldukça önemlidir. Bir makine öğrenimi modelini farklı problemler üzerinde değerlendirmek için hiper parametrelerinin ayarlanması gerekmektedir.

Hiperparametreler, çeşitli deneme işlemlerinden sonra optimize edilmiş ortalama değerleri elde etmek için eğitim aşamasında değerlendirmeye alınan değişkenlerdir. Makine öğrenimi için hiperparametrelerin optimize edilmesi modellerin performansı üzerinde doğrudan etkilidir (Yang ve Shami, 2020). Sarkar ve ark. (2019), 2010-2013 yılları arasında Hindistan'daki bir entegre çelik fabrikasının iş kazası verilerini kullanarak kaza sonuçlarını tahmin etmek için yapmış olduğu çalışmada parametreleri optimize edilmiş makine öğrenimi tabanlı tahmin modellerini kullanmıştır. Kaza sonuçlarını tahmin etmek için kullanılan SVM ve YSA sınıflandırıcılarının parametrelerini optimize etmek için GA ve PSO olmak üzere iki optimizasyon algoritması kullanılmıştır. Wu ve ark. (2019), manuel parametre seçiminin dezavantajlarının üstesinden gelmek için ızgara araması veya kartezyen hiperparametre araması gibi otomatik arama algoritmaları önermiştir. Kılavuz arama, eğitim kümesindeki olası hiperparametre ifadelerinin tüm kombinasyonu ile bir makine öğrenimi modelini eğitir ve çapraz doğrulama kümesinde önceden tanımlanmış bir metriğe göre performansı değerlendirmektedir. Shekar ve Dagnev (2019), ızgara araması ile aşırı uydurma probleminin çözümü için, örneklemelerin rastgele k adet katmana bölündüğü çapraz doğrulama yöntemini uygulamıştır.

Veri setleri içerisinde bulunan her bir sınıf türüne ait örnek sayıları her zaman dengeli olamayabilir. Çok sınıflı sınıflandırma problemlerinde, bir sınıfın örnekleri diğer bazı sınıflardan önemli ölçüde daha fazla olmaktadır. Bu tip veri kümeleri dengesiz veri kümesi olarak tanımlanır. Bazı durumlarda azınlık sınıfı önemli olabilir. Ancak çoğu sınıflandırıcı, çoğunluk ve azınlık sınıfı örneklem sayılarının aynı miktarda olduğu varsayarak eğitilmektedir (Sarkar ve ark. 2019). Dengesiz sınıf dağılımlarının bulunduğu veri setleri, sınıflandırıcıların performansını kötü yönde etkilemektedir (Bulut, 2016). Yapılan çalışmalar ve istatistikler incelendiğinde kaza sonuç bazında sınıflandırılan iş kazası veri setlerinin dengesiz dağılıma sahip oldukları görülmüştür. Dengesiz veri setlerinin varlığı (yaralanma sayılarının ölümlere kıyasla önemli ölçüde yüksek olması), mesleki ölümlerin tahmin edilmesini zorlaştırmaktadır. Bu tarz veri setlerinde yeniden örnekleme tekniklerinin kullanılması, kaza sonucu tahminlerinde önemli gelişmeler sağlamaktadır (Koç ve Gurgun, 2022). Chawla ve arkadaşları tarafından geliştirilen (2002) SMOTE algoritması, orijinal eğitim setini dengelemek için yüksek hıza sahip bir örnekleme yaklaşımı oluşturmaktadır. Azınlık sınıfı örneklerinin

basit bir kopyasını oluřturmak yerine, sentetik örnekler sunmaktır (Fernández ve ark. 2018).

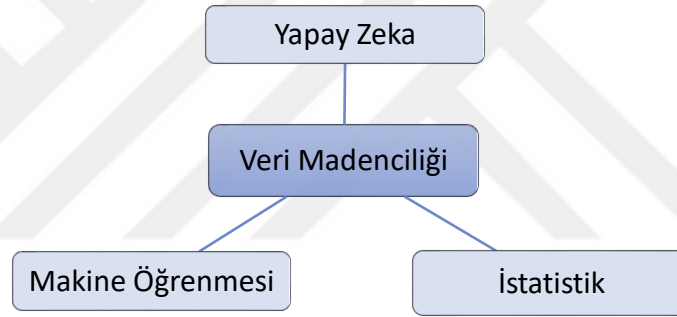
Veri madencilięi yöntemleri kullanılarak analiz edilen iř kazaları üzerine yapılan çalışmalar incelendięinde tekstil sektöründe yařanan iř kazaları sonucu oluřan yaralanma türlerinin tahminine yönelik benzer bir çalışma yapılmadıęı görölmektedir. Bu çalışmanın amacı, veri madencilięi yöntemleri kullanılarak tekstil sektöründe iř kazalarının kontrolüne yönelik risk faktörlerinin araştırılmasıdır. Bu amaç doęrultusunda iř kazaları sonucu oluřan yara türleri, XGBoost (eXtreme Gradient Boosting), SVC (Support Vector Classification), Ekstra Ağaç (Extra Trees), Gradyan Artırma (Gradient Boosting) ve Rastgele Orman (Random Forest) yöntemlerinin sınıflandırma başarıları incelenmiř ve en başarılı yöntem belirlenmiřtir. Çalışmanın geriye kalan kısmı üç ana bölümden oluřmaktadır. İkinci bölümde Kaynak Özetleri, üçüncü bölümde Materyal ve Yöntem, dördüncü bölümde Bulgular ve son bölümde ise Tartışma ve Sonuç'a yer verilmiřtir.

2. KAYNAK ARAŞTIRMASI

Bu bölümde, veri madenciliği yöntemleri ve veri seti kapsamlı olarak anlatılmıştır.

2.1. Veri Madenciliği

Veri madenciliği, gelişmiş analitik teknikleri kullanarak büyük boyutlu veri kümelerinden karar destek sistemlerini yönlendirebilecek önemli bilgileri açığa çıkarma sürecidir. Veri madenciliği yöntemi, mevcut durumun anlamlı veriler ile analiz edilmesine ayrıca olası durumlara karşı tahmin politikalarının geliştirilmesine destek olmaktadır. Veri madenciliğinin temelinde iç içe geçmiş üç bilimsel alan bulunmaktadır. Bunlar; istatistik, yapay zeka ve makine öğrenmesidir (Şekil 2.1).



Şekil 2.1. Veri madenciliği temel alanları

2.2. Veri Madenciliğinin Tarihi

Veri madenciliği, uzun geçmişi olan bir disiplindir. İstatistiksel temelleri, 1763'te Bayes Teoremi'nin ve 1805'te regresyon analizinin keşfiyle başlamıştır. Daha sonra, 1970'lerde veri tabanlarının geliştirilmesi, 1975'te genetik algoritmaların ortaya çıkması ve 1989'da veri tabanlarında bilgi keşfinin yapılmasıyla birlikte veri madenciliği bugünkü anlayışına zemin hazırlanmıştır. 2000'li yıllara gelindiğinde, veri depolama ve teknolojinin hızla gelişmesiyle birlikte veri madenciliği daha güçlü ve üretken hale gelmiştir. Bu dönemde veri madenciliği, birçok alanda uygulanmaya başlanmış ve elde edilen başarılı sonuçlar, veri madenciliği uygulamalarına olan ilgiyi artırmıştır. 2010'lu yıllarda, sosyal medya sitelerinin ve mobil yaşamın yaygınlaşmasıyla birlikte büyük veri kavramı önem

kazanmıştır. Günümüzde nesnelerin interneti, bulut bilişim ve derin öğrenme gibi disiplinlerin de etkisiyle veri çeşitliliği artmış ve bu da veriyi işleme algoritmalarının çeşitlenmesine yol açmıştır (Savaş, 2020).

2.3. Veri Madenciliği Süreci

Veri madenciliği dört adımlı bir süreçten oluşmaktadır. Veri madenciliği sürecinde izlenen adımlar şu şekildedir; problemin tanımlanması, verilerin hazırlanması, modelin kurulması ve değerlendirilmesi, modelin kullanılması ve modelin izlenmesi.

2.3.1. Problemin tanımlanması

Proje amacının tam olarak belirlenmesi ve elde edilecek sonuçların başarı ölçütlerinin tanımlanması, veri madenciliği çalışmalarının en önemli iki unsurudur. Çalışmaya başlamadan önce sorunun kapsamı tam olarak belirlenmelidir. Bu aşamada verilen yanlış kararlar veri setinin oluşturulması ve algoritma seçiminde yanıltıcı olacaktır.

2.3.2. Verilerin hazırlanması

Veri madenciliği algoritmaları ile analiz edilecek olan verinin öncelikle veri ön işleme aşamasından geçmesi oldukça önem arz etmektedir. Veri ön işleme adımları; eksik değerlerin tespit edilerek veri setinin uygunluğuna göre silinmesi ya da tamamlanması, aykırı değerlere sahip örneklerin çıkarılması ve öznitelik değerlerinin belirli aralıklarda ölçeklendirilmesidir (Aghalarova ve B. Keser, 2022). Bir sonraki aşama ise öznitelik seçimidir. Birçok veri madenciliği algoritması, büyük miktarda özellik veya öznitelik ile iyi performans gösterememektedir. Çalışma kapsamında değer katmayan öznitelikler analiz aşamasındaki hesaplamaları yavaşlatmaktadır. Bu nedenle, veri madenciliği algoritmaları uygulanmadan önce öznitelik seçim tekniklerinin uygulanması gerekmektedir. Öznitelik seçiminin ana hedefi, aşırı öğrenmeyi önlemek ve model performansını iyileştirmektir (Beniwal ve Arora, 2012). Modelin kurulması aşamasında ortaya çıkan sorunlar veri hazırlığı aşamasına geri dönülmesine neden olacaktır. Bu nedenle veri bilimcisinin, verilerin hazırlanması adımıyla oldukça dikkatli olması gerekmektedir.

2.3.3. Modelin kurulması ve değeriendirilmesi

Çalışma amacına en uygun modelin bulunması için birçok sayıda modelin kurulması ve denenmesi gerekmektedir. Bu nedenle verilerin hazırlanması ve modelin kurulması, en iyi sonuç veren modele ulaşana kadar tekrarlanan süreçlerdir. Analizlerin temelini oluşturan veri madenciliği teknikleri iki farklı amaç doğrultusunda ayrılmaktadır. Bunlar, hedef veri setinin tanımlanması ve makine öğrenimi algoritmaları yardımıyla sonuçların tahmin edilmesidir. Veri madenciliği yöntemi, veriler arasındaki anlamlı ilişkileri açığa çıkarırken makine öğrenmesi veriler arasındaki ilişkiyi temsil eden bilgilerden faydalanarak modellere ilişkin tahmini sonuçlar bulur. Veri madenciliği algoritmalarının çıktısı makine öğrenimi algoritmalarının girdisini oluşturmaktadır. (Ersöz ve Çınar, 2021) Modelleme aşamasında, veri ön işleme adımını tamamlayarak analize hazır hale gelen veri makine öğrenmesi yöntemleri ile eğitilerek modeller oluşturulur.

2.3.4. Modelin kullanılması

Tasarlanan ve doğruluğu kabul edilen model doğrudan bir uygulama olabileceği gibi, başka bir uygulamanın girdisi olarak da kullanılabilir. Tasarlanan modeller, amaçlanan problem çözümleri için anlamlı örüntüler ve değeriendirmeler ortaya koymalıdır.

2.3.5. Modelin izlenmesi

Tasarlanan modellerin sürekli izlenmesi ve düzenlenmesi önemlidir. Değişen veriler ve sistem özellikleri, modelin performansını etkileyebilir. Bu nedenle, modelin güncel kalması için düzenli olarak gözden geçirilmesi ve gerekirse yeniden eğitilmesi gerekmektedir. Bu süreç, doğru sonuçlar elde etmek ve güvenilir kararlar vermek için gereklidir.

2.4. Veri Madenciliği Teknikleri

Veri madenciliği yöntemleri, istatistiksel analiz, makine öğrenmesi (ML) ve matematik ile birlikte birbiriyle örtüşen çeşitli öğrenme alanlarından faydalanır. Veri madenciliği

yöntemleri kestirimci ve betimleyici olmak üzere genel başlıkta iki gruba ayrılmaktadır. Kestirimci yöntemler, bilinen bazı değişkenleri kullanarak bilinmeyen bir değişkenin değerini bulmak amacıyla kullanılır. Betimsel yöntemler ise, verileri anlaşılır hale getirmek ve bilgilerin keşfini kolaylaştırmak için verilerin örüntülerini araştırmak amacıyla geliştirilen yöntemlerdir (Çelik, 2019). Veri madenciliğinde sıklıkla kullanılan teknikler aşağıda yer almaktadır.

2.4.1. Sınıflandırma ve regresyon

Sınıflandırma ve regresyon, veri analizi yöntemleri olarak kullanılan iki önemli tekniktir. Bu teknikler, veri analizi sürecinde önemli veri sınıflarını ortaya koyma ve gelecek veri eğilimlerini tahmin etme amacıyla kullanılmaktadır. Sınıflandırma, kategorik değerlerin tahmin edilmesinde kullanılırken, regresyon ise sürekli değerlerin tahmin edilmesinde kullanılır. İş dünyasındaki kullanımına örnek olarak; bir sınıflama modeli banka kredi başvurularını güvenli veya riskli olarak sınıflandırmak için oluşturulabilirken, bir regresyon modeli geliri ve mesleği bilinen potansiyel müşterilerin bilgisayar ürünleri için yapacakları harcamaları tahmin etmek amacıyla kurulabilmektedir (Özekes, 2003).

2.4.2. Birliktelik kuralları

Birliktelik kuralları, bir arada gerçekleşen olaylara ya da özelliklere karar verme işlemidir. Dilek kipli ifadeler, iki veri arasındaki ilişkinin olasılığını göstermektedir. Bu duruma bir müşterinin X ürününü satın alması durumunda Y ürününü de yüksek olasılıkla satın alması durumu gösterilebilir. Birliktelik kuralları ürün katalog tasarımından, satış-pazarlamaya kadar birçok farklı alanda kullanılmaktadır. İş dünyasındaki kullanımına örnek olarak; online alışveriş sitelerinin önerilen bölümleri veri madenciliğinin birliktelik kuralı yöntemi ile oluşturulmaktadır (Taş, 2018).

2.4.3. Kümeleme

Bu veri madenciliği yönteminin amacı, verilerin kendi aralarındaki benzerlik durumuna göre alt gruplara bölünmelerini sağlamaktır. Veri madenciliği kümeleme yöntemi ile oluşturulan her kümenin birbirinden farklı olduğu, ancak küme içerisindeki verilerin

benzer olduđu küme setleri oluşturulmaktadır. Sınıflandırma işleminde, sınıflar önceden belirli iken kümelemede sınıflar önceden belirlenmemiştir. Bu özelliği nedeniyle denetimsiz öğrenmeye örnek teşkil eden kümeleme analizi istatistik, makine öğrenmesi, matematik ve yapay zeka gibi bir çok disiplin tarafından biyoloji, tıp, antropoloji, pazarlama, ekonomi ve telekomünikasyon gibi alanlarda sıkça kullanılmaktadır (Karaibrahimoğlu, 2014).

2.4.4. Veri temizleme

Verilerin düzenlenmesi, yinelenen veya bozuk verilerin ortadan kaldırılması ve boş değerlerin uygun şekilde doldurulması sürecidir. Bu süreç tamamlandığında veriler analiz için hazır hale gelmektedir. 6 farklı veri temizleme yöntemi bulunmaktadır. Bunlar; verilerin doğrulanması, veri türlerinin dönüştürülmesi, gereksiz verilerin kaldırılması, yinelenen verilerin kaldırılması, hataların kaldırılması ve eksik değerlerin tamamlanmasıdır. Verilerin doğrulanması, veri setindeki her bir değerin doğru formatta olup olmadığının kontrol edilmesidir. Veri türlerinin dönüştürülmesi, verilerin veri kümesi genelinde standart olmasını sağlamaktadır. Örneğin, sayısal değişkenler yalnızca sayıları içerirken, dize değişkenleri harfleri, sayıları ve karakterleri içerebilir. Gereksiz verilerin kaldırılması, gereksiz veya uygulanamaz verilerin temizlenmesidir. Bu sayede gerekli veri noktalarına tam vurgu yapılabilir. Yinelenen verilerin kaldırılması, verimliliği artırarak ve hataları azaltarak madencilik sürecinin hızlandırılmasına yardımcı olmaktadır. Hataların kaldırılması, analiz sonuçlarını olumsuz etkileyebilecek yazım hataları, imla hataları ve giriş hatalarının ortadan kaldırılmasıdır. Eksik değerlerin tamamlanması ise, eksik tüm veriler için tahmini bir değerin sağlanmasıdır. Bu sayede çarpık veya yanlış sonuçlara yol açabilecek eksik değerler azaltılmaktadır.

2.4.5. Veri görselleştirme

Verilerin çizelgeler, grafikler, haritalar, diyagramlar ve benzeri araçları kullanarak görsel yollarla sunulmasıdır. Veri bilimcilerin bulgularını sergilemek için kullandıkları en temel yoldur. Verileri görsel olarak temsil etmek için birçok yöntem mevcuttur. Bunlardan

bazıları; karşılaştırma tabloları, ısı haritaları, yoğunluk grafikleri, histogramlar, ağ diyagramları ve dağılım grafikleridir.

2.4.6. Sinir ağları

Yapay sinir ağları, insan beyninin çalışma şeklini dijital olarak taklit etmeye çalışmaktadır. Sinir ağları, birçok bilgisayar işlemcisi verileri işlemek, kararlar almak ve bir insan gibi öğrenmek için birleştirmektedir. Sinir ağları üç ana katmandan oluşmaktadır. Bunlar; girdi, gizli ve çıktı katmanlarıdır. Veriler giriş katmanından girerek, gizli katmanda işlenir ve verilere dayalı olarak işlemin yapıldığı çıkış katmanında çözümlenmektedir. Gizli katman, kullanılan veri miktarına ve gerçekleşen öğrenmeye bağlı olarak birçok işlem katmanından oluşabilmektedir. İş dünyasındaki kullanımına örnek olarak; işletmeler tüketici satın alma modellerini yapay sinir ağları yöntemi ile tahmin ederek pazarlama kampanyalarında belirli demografilere odaklanabilirler.

2.4.7. Aykırı değer tespiti

Genel olarak veri madenciliği teknikleri kalıpları ve eğilimleri bulmayı amaçlarken, aykırı değer tespiti tekniği benzersiz olan verileri aramaktadır. Aykırı değer tespiti, veri setine yanlış girilen veya yanlış örnek sonucu ortaya çıkan hatalı verileri bulmaktadır. Aykırı değer tespiti için en sık kullanılan yöntemler; sayısal aykırı değer, Z-skoru, DBSCAN ve izolasyon ormanıdır.

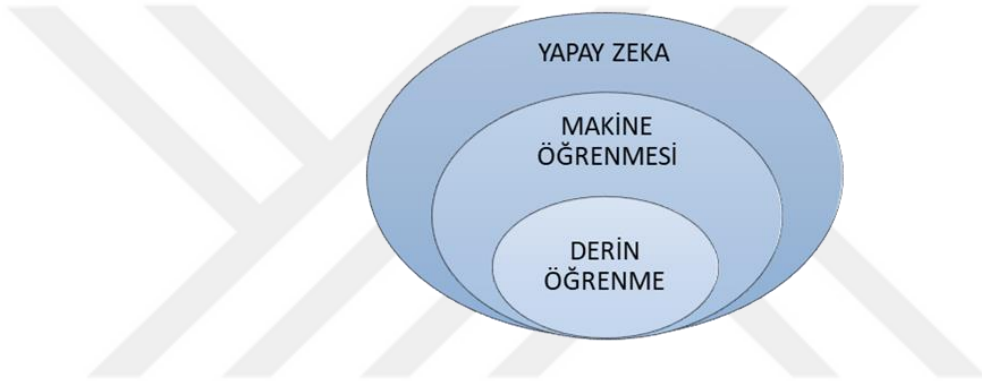
2.4.8. Tahmin

Tahmine dayalı modelleme, verilerin gelecekteki davranışını öngörmeyi amaçlayan bir yöntemdir. Bu modeller, veri kümelerini inceleyerek kalıpları ve eğilimleri tespit ederek gelecekteki olayların olasılıklarını hesaplamaktadır. Tahmine dayalı modelleme, veri madenciliğinin en yaygın kullanımlarından biridir ve özellikle büyük veri kümeleriyle etkili bir şekilde çalışmaktadır. İş dünyasında tahmine dayalı modelleme, birçok alanda kullanılmaktadır. Örneğin, şirketler bu modelleri kullanarak satış projeksiyonlarını belirlemekte ve tüketici satın alma alışkanlıklarını tahmin etmektedir. Aynı zamanda, üreticiler dağıtım ihtiyaçlarını tahmin etmek ve bakım programlarını planlamak için

tahmine dayalı modellemeyi kullanmaktadır. Bu sayede, işletmeler kaynaklarını daha etkin bir şekilde yönetebilir ve operasyonlarını optimize edebilmektedir.

2.4.9. Makine öğrenmesi

Yapay zekanın (AI) bir alt kümesi olarak kabul edilen makine öğrenmesi (ML), bir bilgisayarın doğrudan talimatları kullanmadan öğrenmesine yardımcı olmak için matematiksel modelleri kullanma prosedürüdür (Şekil 2.2). Makine öğrenmesi, veri kalıplarını belirlemek için algoritmaları kullanır. Tasarlanan veri modelinin öğrenilen sistem üzerinde tahmin yürütebilmesi için bu kalıplar kullanılır.



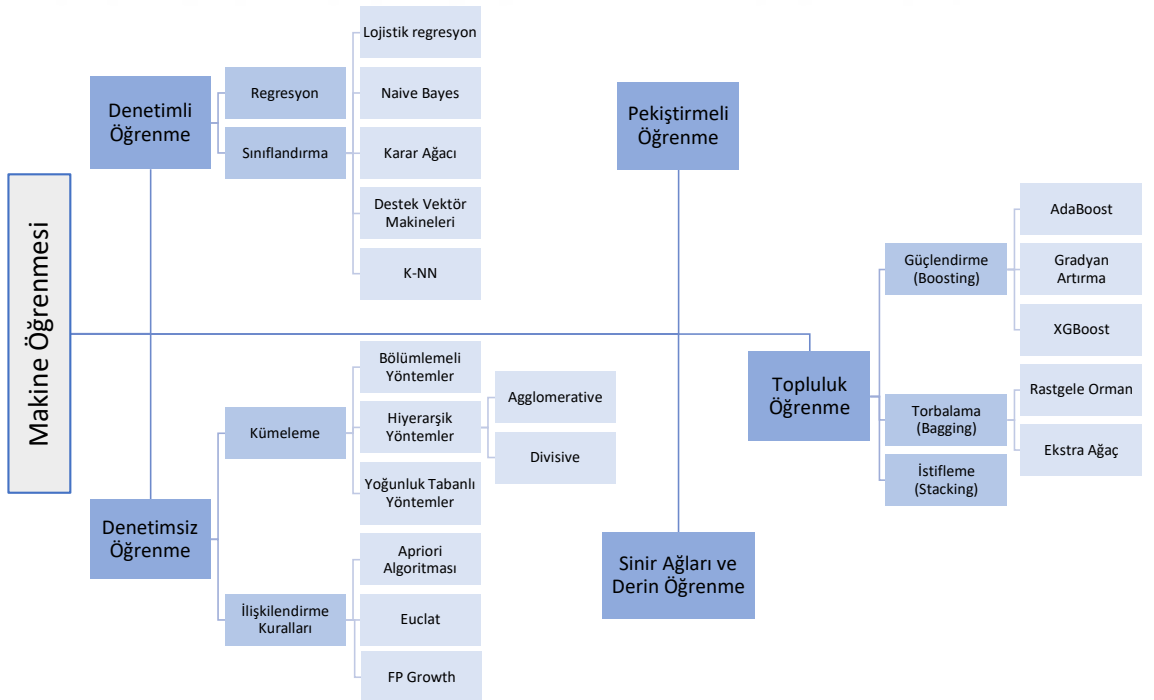
Şekil 2.2. Yapay zeka ve alt dalları

Makine öğrenmesi, tahmine dayalı analizlerin bir türüdür ve veri kümesi büyüdükçe gerçek zamanlı olarak güncellenen algoritmaların uygulanması daha kolay hale gelmektedir. Statik veri kümeleriyle çalışan tahmine dayalı analizlerin ise belirli periyotlarla yenilenmesi gerekmektedir. Makine öğrenmesi, büyük verinin sürekli olarak genişlediği bir ortamda veri mühendisliğiyle önemli bir ilişkiye sahip olmuştur. Veri madenciliği projelerinde, algoritmalar istatistiksel yöntemler kullanılarak sınıflandırma veya tahmin yetenekleri üzerine eğitilmiştir ve temel içgörülerini ortaya çıkarmıştır. Bu içgörüler, karar verme sürecini etkileyerek stratejik büyüme ölçütlerini yönlendirmiştir. Makine öğrenmesi farklı alanlarda kullanılmakta ve bunlara sürekli yenileri eklenmektedir. İşletmeler makine öğrenmesi çalışmaları ile önemli avantajlar elde etmektedir. Örnek olarak; iç görülerini ortaya çıkarma, veri bütünlüğünü geliştirme, kullanıcı deneyimini geliştirme, risk azaltma, müşteri davranışlarını tahmin etme ve maliyetleri azaltma söylenebilir.

2.5. Makine Öğrenmesi Algoritmaları

Makine öğrenmesi algoritmaları, karmaşık veri kümelerini keşfetmek, analiz etmek ve sonuçları anlamlandırmak için kullanılan kod parçacıklarıdır. Her algoritma, belirli bir hedefi gerçekleştirmek için adımlar ve talimatlar kümesinden oluşur. Makine öğrenmesi modelleri, tahmin yapma ve verileri kategorize etme için kullanılabilen desenler oluşturmayı amaçlamaktadır. Bu algoritmalar, geniş veri kümesini yansıtan ve eğitim verilerine dayanan parametreleri kullanmaktadır. Eğitim verileri arttıkça, algoritmalar daha doğru sonuçlar üretebilmektedir.

Makine öğrenmesi alanında geliştirilen algoritmalar sürekli olarak çeşitlenmektedir. Her bir algoritma, farklı amaçlara hizmet etmek üzere kısıtları ve avantajları açısından farklılık göstermektedir. Farklı algoritmalar, verileri farklı şekillerde analiz etmektedir. Makine öğrenmesi algoritmaları, denetimli öğrenme, denetimsiz öğrenme, pekiştirmeli öğrenme, topluluk öğrenme ve sinir ağları ve derin öğrenme olmak üzere beş farklı dalda sınıflandırılmaktadır (Şekil 2.3).



Şekil 2.3. Makine öğrenmesi algoritmaları

2.5.1. Denetimli öğrenme

Denetimli öğrenme algoritmaları, etiketli eğitim verilerinden öğrenerek öngörülemeyen verilerin sonuçlarını tahmin etmektedir. Etiketli veri kümesi hem girdi hem de çıktı değerlerine sahiptir. Model eğitilirken, veriler genellikle 80:20 oranında bölünmektedir. Bu durumda verilerin %80'i eğitim verisi geri kalanı ise test verisi olarak kullanılmaktadır. Eğitim verileri hem girdiyi hem de çıktıyı beslemektedir. Eğitim verileri kullanılarak farklı algoritmalar yardımı ile model oluşturulmaktadır. Model başarısı ise, test verilerinin tahmin ettiği değerler ile gerçek değerler karşılaştırılarak hesaplanmaktadır.

Denetimli öğrenme, sınıflandırma ve regresyon algoritma modellerini içermektedir. Regresyon modelleri, çıktı değişkeni süreklilik gösteren nümerik değerlerden oluşan veri setlerinde kullanılmaktadır. Regresyon modelleri, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi anlamak için kullanılmaktadır. Bu tarz veri setlerinde, iki veya daha fazla değişken arasında bir ilişki bulunmaktadır. Bir değişkende meydana gelen değişiklik diğer değişkendeki değişiklik ile ilişkilendirilmektedir. Örnek olarak, iş deneyimi- maaş veya boy-kilo değişkenleri arasındaki ilişki gösterilebilir. Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi anlamak için regresyon kullanılır.

Sınıflandırma modelleri, çıktı değişkeni kategorik ifadelerden oluşan, yani 2 veya daha fazla sınıftan oluşan veri setlerinde kullanılmaktadır. Sınıflandırma, test verilerinin belli kategorilere doğru şekilde atanması için algoritmalarla yararlanmaktadır. Sınıflandırma algoritmaları ile veri kümesindeki etiketi bilinmeyen örneklerin nasıl etiketlenmesi veya tanımlanması gerektiğine dair bazı sonuçlar çıkarılmaktadır. Sınıflandırma problemlerinde çıktı değişkeni, kırmızı-yeşil-mor veya çalışan-emekli gibi kategorik ifadelerden oluşmaktadır. Yaygın kullanılan sınıflandırma algoritmaları; Lojistik Regresyon, Naive Bayes, Karar Ağacı, SVM (destek vektör makineleri) ve K-En Yakın Komşu (K-NN)'dur.

David Cox (1972) tarafından geliştirilen Lojistik Regresyon Algoritması, veri setindeki gözlemlerin gruplara atanmasında kullanılan bir yöntemdir. Var olan verilerle bir

sınıflandırma modeli elde edilir ve yeni gözlemlerin sınıflara atanması sağlanır. Bu analiz, bağımlı değişkenin birden fazla kategorili olduğu durumlarda bağımsız değişkenlerle neden-sonuç ilişkisini belirlemek için kullanılmaktadır. İlgilenilen değişkenin modelde yer aldığı durumda, daha iyi tahminler elde edilmekte ve bu değişkenin model için önemli olduğu yorumlanmaktadır. Araştırmacılar, birden çok etkenin olduğu durumlarda, etkenlerin bağımlı değişken üzerindeki etkisini tek tek öğrenmenin yanı sıra birlikte etkilerini incelemek için lojistik regresyonu tercih etmektedirler (Başer ve arkadaşları, 2021).

18. yüzyılda Thomas Bayes tarafından ortaya konulan yaklaşımlara dayanan Naive Bayes Algoritması, Bayes teoremine dayanan ve sınıflandırma problemlerinin çözümünde kullanılan denetimli bir öğrenme algoritmasıdır. Sınıf koşullu bağımsızlık ilkesini benimseyen sınıflandırma yaklaşımıdır. Bu ifade, belirli bir sonucun olasılığında bir özelliğin varlığının diğerinin varlığını etkilemediği ve her tahmin edicinin o sonuç üzerinde eşit etkiye sahip olduğu anlamına gelmektedir. Genel olarak, yüksek boyutlu eğitim veri seti içeren metin sınıflandırmasında kullanılmaktadır.

J.R. Quinlan tarafından 1986 yılında geliştirilen Karar Ağacı Algoritması hem sınıflandırma hem de regresyon problemlerinde kullanılabilen denetimli bir öğrenme tekniğidir ancak daha çok sınıflandırma problemlerinin çözümünde tercih edilmektedir. İç düğümlerin veri kümesinin özelliklerini, dalların karar kurallarını ve her yaprak düğümün sonucu temsil ettiği ağaç yapılı bir sınıflandırıcıdır. Karar Ağaçlarının, karar aşamasında insanın düşünme yeteneğini taklit etmesi ve yapı itibarıyla ağaç sekline benzemesi algoritmayı daha anlaşılır hale getirmektedir. Karar ağacının çalışma prensibinde, veri kümesinin sınıfını tahmin etmek için algoritma ağacın kök düğümünden başlamaktadır. Devamında, ikili karşılaştırmalar ile uygun dalı takip edilerek bir sonraki düğüme atlanmaktadır. Bir sonraki düğüm için, algoritma tekrardan öznelik değerini diğer alt düğümlerle karşılaştırır ve ağacın yaprak düğümüne ulaşana kadar çalışmaya devam etmektedir (Çelik ve ark, 2022). Karar ağaçları, tüm olası sonuçların düşünülmesine yardımcı olmaktadır. Ancak karar ağaçlarını karmaşık hale getiren birçok katman bulunmaktadır. Rastgele Orman algoritması kullanılarak aşırı uyum sorunu çözülebilmektedir.

1963 yılında Vladimir Vapnik ve Alexey Chervonenkis tarafından temelleri atılan Destek Vektör Makineleri (SVM) Algoritması, hem sınıflandırma hem de regresyon problemlerinde kullanılabilen denetimli bir öğrenme tekniğidir. SVM algoritmasının amacı, veri noktalarını belirgin bir şekilde sınıflandıran N boyutlu bir uzayda bir hiper düzlem bulmaktır. Sınıflandırma problemlerinde, iki veri noktası sınıfı arasındaki mesafenin maksimumunda olduğu bir hiper düzlem çizilmektedir (Kaynar ve ark, 2018). Bu hiper düzlem karar sınırı olarak bilinir ve düzlemin her iki tarafındaki veri noktası sınıflarını ayırmaktadır. Hiper düzlemin boyutu, özelliklerin sayısına bağlıdır. SVM algoritması, “kernel trick” olarak adlandırılan özel bir tekniğe sahiptir. SVM çekirdeği, düşük boyutlu girdi uzayını alarak onu yüksek boyutlu bir uzaya dönüştürmektedir. Bu sayede, ayrılabilir olarak tanımlanan problem ayrılabilir probleme dönüştürülebilir. Doğrusal olmayan ayırma probleminde daha kullanışlıdır. SVM algoritması çalışma prensibi olarak öncelikle karmaşık veri dönüşümleri yapılmaktadır. Sonrasında ise tanımlı etiketlere veya çıktılarına göre verileri ayırma prosesleri bulunmaktadır (Tuncer ve ark, 2022).

1967 yılında T. M. Cover ve P. E. Hart tarafından önerilen K-En Yakın Komşu (K-NN) Algoritmasının çalışma prensibi, öncelikle veri kümesini eğitim ve test verisi olarak ikiye bölmektir. Bir gözlemi sınıflandırmak için, nitelik uzayında ele alınan gözlemin eğitim kümesindeki her bir gözleme olan uzaklığı ayrı ayrı hesaplanmaktadır. Önceden belirlenen k sayısı kadar en yakın komşunun hangi sınıflara ait olduğuna bakılarak, en çok gözlem sayısı hangi sınıftaysa, ele alınan örnek veri o sınıfa dahil edilmektedir. Algoritmanın performansı, k değeri ile birlikte uzaklık hesaplama yöntemi tarafından da etkilenmektedir. Bu nedenle, farklı uzaklık hesaplama yöntemleri kullanılarak elde edilen sonuçların karşılaştırmalı analizi yapılabilmektedir (Başer ve ark. 2021).

2.5.2. Denetimsiz öğrenme

Denetimsiz öğrenme, etiketlere sahip olmayan veri kümelerini analiz etmek ve kümelemek amacıyla makine öğrenmesi algoritmalarını kullanmaktadır. Bu algoritmalar, gizli kalıpları veya veri gruplarını keşfetmektedir. Bilgilerdeki benzerlikleri ve farklılıkları keşfetme özelliği sayesinde, veri analizi, çapraz satış stratejileri, müşteri

segmentasyonu ve görüntü tanıma için ideal çözüm yöntemi haline gelmiştir. Denetimsiz öğrenme modelleri, kümeleme ve ilişkilendirme analizi olmak üzere iki temel görev için kullanılmaktadır.

Kümeleme, verilerin, kendi kümesinde benzerlik, farklı kümelerde farklılık göstereceği şekilde gruplandırılması yöntemidir. Kümeleme analizinde, veri nesneleri arasındaki ortak değerler bulunur. Bu ortak noktaların varlığına ve yokluğuna göre veriler gruplandırılır. Kümeleme; hiyerarşik, bölümlenmeli ve yoğunluk tabanlı olmak üzere 3 farklı yöntemden oluşmaktadır.

İlişkilendirme analizi, veri setindeki değişkenler arasındaki ilişkileri bulmayı amaçlayan kural tabanlı bir yöntemdir. Bu yöntemler, işletmelerin farklı ürünler arasındaki ilişkileri daha iyi anlamalarını sağlayan pazar sepeti analizi için sıklıkla kullanılmaktadır. Apriori, Eclat ve FP-Growth olmak üzere 3 farklı yöntemden oluşmaktadır. En sık kullanılan yöntem Apriori algoritmasıdır.

2.5.3. Pekiştirmeli öğrenme

Pekiştirmeli öğrenme ile, sistemin kendi etkileşimlerinin sonuçlarını gözlemleyerek önceki tecrübelerden öğrenmesi sağlanır. Bu algoritma denetimli öğrenmenin aksine deneme yanılma yöntemi ile öğrenmeyi gerçekleştirmektedir. Verinin analiz edilmesi sonucunda aldığı bildirimler ile kullanıcıları en doğru sonuca yönlendiren davranışsal bir öğrenme yöntemidir (Gürsoy, 2022) Alınan geri bildirimler ile sistem ödül ve cezalar açısından değerlendirilerek denge durumu oluşana kadar parametre ayarlama süreci devam etmektedir.

2.5.4. Sinir ağları ve derin öğrenme

Derin öğrenme, yapay sinir ağlarının büyük boyutlu verilerden öğrenme yeteneği olan bir makine öğrenme alt alanıdır. Sinir ağlarının katmanları tarafından desteklenen derin öğrenme modelleri, birden fazla veri kaynağından bilgi alabilir ve insan müdahalesine gerek duymadan gerçek zamanlı analiz yapabilmektedir. Derin öğrenme için özel olarak optimize edilen grafik işleme birimleri, eğitim modellerinde aynı anda birden fazla

hesaplama yapabilme yeteneğine sahiptir. Sinir ağı, etiketlenmemiş ve yapılandırılmamış verilerdeki gizli yapıları keşfedebilme yeteneğine sahiptir.

2.5.5. Topluluk öğrenme

Topluluk öğrenme yöntemleri, optimum tahmin modeli üretmek için birkaç temel modeli birleştiren bir makine öğrenme yöntemidir. Veri kümesine uygulanan modellerin her biri tarafından yapılan farklı tahminler, topluluk öğrenme modelinin temelini oluşturmaktadır. Topluluk öğrenme modeli ile nihai sonucu elde etmek için farklı modellerin tahminlerinin sonuçlarını birleştirilmektedir.

Bazı problemlerin karmaşık bir karar sınırı olabilir. Bu durumda tek bir sınıflandırıcının uygun sınırı oluşturması oldukça zordur. Çoklu doğrusal sınıflandırıcılardan oluşan bir topluluk, herhangi bir polinom karar sınırı oluşturabilir. Bilgi birleşimi ile üretilen model, tek başına alınan temel öğrencilerden daha iyi performansa sahiptir. Bu durumun örneği aşağıdaki şekilde gösterilmiştir. Topluluk öğrenme modelleri Güçlendirme (Boosting), Torbalama (Bagging) ve İstifleme (Stacking) olmak üzere üç temel görev için kullanılmaktadır. En popüler topluluk algoritmaları Rastgele Orman, XGBoost ve LightGBM' dir.

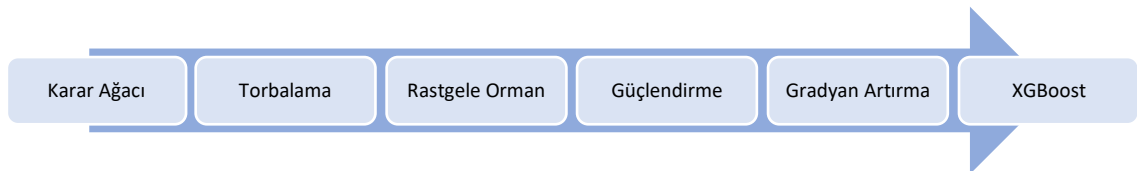
Güçlendirme (Boosting), eğitim hatalarını en aza indirmek amacıyla bir dizi zayıf öğrenciyi güçlü öğrencilerle birleştiren bir topluluk öğrenme yöntemidir. Güçlendirmede, rastgele bir veri örneği seçilir, bir modelle oturtulur ve sonrasında sırayla eğitilir. Sonuç olarak, her model bir diğerinin zayıflıklarını telafi etmeye çalışmaktadır. Tek bir tahmin kuralı oluşturmak için her sınıflandırıcının zayıf kuralları yineleme ile birleştirilmektedir. Genel anlamda, makine öğrenmesi çalışmalarında kapsamlı kullanımı bulunan üç temel güçlendirme algoritması türü vardır: AdaBoost, Gradyan Artırma ve XGBoost.

AdaBoost, Freund ve Schapire (1995) tarafından formüle edilen istatistiki bir meta sınıflandırma algoritmasıdır. AdaBoost tekniği yinelemeli olarak çalışmaktadır. Yanlış sınıflandırılan veri noktalarını tanımlayarak eğitim hatasını en aza indirecek şekilde

ağırlıklandırma optimizasyonu yapmaktadır. Model, en güçlü tahmin ediciyi oluşturana kadar tekrarlı bir şekilde optimize etmeye devam etmektedir.

Tianqi Chen ve Carlos Guestrin tarafından 2016 yılında oluşturulan Gradyan Artırma Algoritması, prosedür olarak önceki aşamalarda yapılan hataları düzenleyen tahmin edicilerin sırasıyla bir gruba eklendiği güçlendirme tekniğidir. Ayrıca, AdaBoost gibi veri noktalarının ağırlıklarını değiştirmek yerine önceki tahmin edicilerin artı kalan hatalarını kullanmaktadır. Gradient Boosting ismi, gradient descent algoritması ile boosting tekniğini birleşiminden gelmektedir (Khorsheed, 2021).

XGBoost, gradyan artırma çerçevesi kullanan karar ağacı tabanlı bir makine öğrenimi algoritmasıdır. Veri bilimcileri ve araştırmacılar tarafından makine öğrenimi modellerini optimize etmek için kullanılmaktadır. XGBoost algoritması, gradyan artırma algoritmasının doğaçlama bir versiyonudur. Chen ve Guesterin' in 2016 yılında yayınlanan çalışmaları ile popüler hale gelmiştir. Her iki algoritmanın da çalışma prosedürü benzerlik göstermektedir. XGBoost, büyük veri kümelerinde hız, kullanım kolaylığı ve performans için tasarlanmıştır (Daldır, 2021). Parametre optimizasyonu veya ayarlama gerektirmez, bu da kurulumdan sonra konfigürasyona ihtiyaç duymadan hemen kullanılabilmesi anlamına gelmektedir. XGBoost' un eğitim süreci hızlıdır ve birden çok küme arasında paralelleştirilebilir. Bu nedenle Gradyan Artırma Algoritmasına kıyasla yüksek performans sunmaktadır. Üç ana gradyan artırma biçimini (Gradyan Artırma, Stokastik GB ve Normalleştirilmiş GB) gerçekleştirme yeteneğine sahiptir. Karar ağacı tabanlı XGBoost algoritmasının yıllar içerisindeki gelişimi Şekil 2.4' de gösterilmiştir.



Şekil 2.4. XGBoost algoritmasının yıllar içerisindeki gelişimi

Torbalama (Bagging) yönteminde kullanılan popüler algoritmalar Rastgele Orman ve Ekstra Ağaçlardır. Leo Breiman tarafından ortaya koyulan Rastgele Orman Algoritması hem sınıflandırma hem de regresyon problemleri için kullanılabilir. Karmaşık bir sorunu çözmek ve modelin performansını iyileştirmek için birden çok sınıflandırıcıyı

birleştirme süreci olan topluluk öğrenme kavramına dayanmaktadır. Rastgele Orman Algoritması, belirli bir veri kümesinin çeşitli alt kümelerinde bir dizi karar ağacı içeren ve bu veri kümesinin tahmin doğruluğunu iyileştirmek için ortalamayı alan bir sınıflandırıcıdır. Böylece, modelin doğruluğunu artırarak fazla uydurma sorununun önüne geçmektedir (Khorsheed, 2021).

Ekstra Ağaçlar Algoritması, Rastgele Ormanlara benzer şekilde çok sayıda karar ağacını eğiten ve bir tahmin çıktısı almak için karar ağaçları grubundan sonuçları toplayan bir toplu makine öğrenimi yaklaşımıdır. Rastgele Orman Algoritması, karar ağaçlarının yeterince farklı olmasını sağlamak için eğitim verilerinin farklı varyasyonlarını seçmek üzere torbalamayı kullanmaktadır. Ancak Ekstra Ağaç Algoritması (Geurts ve ark, 2006), karar ağaçlarını eğitmek için tüm veri setini kullanmaktadır. Bu nedenle, tekil karar ağaçları arasında yeterli farklılıkları sağlamak için, bir özelliğin bölüneceği ve alt düğümlerin oluşturulacağı değerler rastgele seçilmektedir. Tüm veri kümesinin kullanılması modelin yanlılığını azaltmasına olanak tanımaktadır. Hesaplama maliyeti açısından Ekstra Ağaçlar, Rastgele Orman'dan çok daha hızlıdır. Bunun nedeni, Rastgele Orman' da kullanılan açgözlü algoritma yerine Ekstra Ağaçlar' ın özelliklerin bölüneceği değeri rastgele seçmesidir (Ahmad ve ark, 2018).

İstifleme (Stacking), çoklu sınıflandırmaları veya regresyon modellerini bir meta sınıflandırıcı veya meta regresör aracılığıyla birleştiren bir toplu öğrenme tekniğidir. Temel düzey modeller, eksiksiz bir eğitim setine dayalı olarak eğitilir. Daha sonra tüm modellerin çıktıları, nihai olarak örnekleri tahmin eden meta sınıflandırıcı adı verilen başka bir sınıflandırıcıya girdi olarak kullanılır. İki katmanlı sınıflandırıcı kullanmanın ardındaki sezgi, eğitim verilerinin uygun şekilde öğrenilip öğrenilmediğini belirlemektir.

2.6. Makine Öğrenmesi Algoritmaları İçin Hiperparametre Optimizasyonu

Hiperparametre optimizasyonu, bir makine öğrenmesi algoritması için belirlenen başarı metriğine göre en uygun hiperparametre kombinasyonunu bulma işlemidir. Hiperparametre optimizasyonu ile model karmaşıklığı dengelenerek aşırı öğrenme (overfitting) ve eksik öğrenme (underfitting) dengesi sağlanabilir. Hiperparametre optimizasyonu için çalışma ve veri setine uygun seçilmiş ve iyi tasarlanmış temel bir

modele ihtiyaç vardır. Başarısız bir model üzerinde hiperparametre optimizasyonu yapmak sonucu değiştirmeyecektir. Bu nedenle veri ön işleme ve özellik mühendisliği kısımları oldukça önemlidir. İyi bir model kurulduktan sonra hiperparametre optimizasyonu yapmak daha iyi sonuçlar elde edilmesini sağlayacaktır.

2.6.1. Hiperparametre optimizasyonu yöntemleri

Her bir makine öğrenmesi algoritması, onlarca hiperparametre ve bu hiperparametrelerin alabileceği onlarca farklı değer seçeneklerine sahiptir. Modelden başarılı sonuç almak amacıyla hiperparametre kombinasyonları tek tek deneyerek en iyi kombinasyonu seçmek oldukça zordur. Bu nedenle hiperparametre optimizasyonu için farklı yöntemler geliştirilmiştir. GridSearchCV ve RandomizedSearchCV bu yöntemlerdendir.

GridSearchCV, belirli bir model için en uygun hiperparametre değerlerini belirlemek amacıyla hiperparametre ayarlaması yapmaktadır. GridSearch işlemi sırasında, model eğitilirken çapraz doğrulama yöntemi kullanılmaktadır. Standart bir makine öğrenimi sürecinde, veriler eğitim ve test kümeleri olarak ikiye ayrılmaktadır. Eğitim veri seti, modelin eğitimi için kullanılırken, test veri seti ise modelin performansını değerlendirmek için kullanılmaktadır. Ancak bu yaklaşım, farklı test veri setlerinden kaynaklanan varyans problemlerine neden olabilmektedir. Bu problemin üstesinden gelmek için K-Katlı Çapraz Doğrulama yöntemi kullanılmaktadır. K-Katlı Çapraz Doğrulama sürecinde veriler K adet gruba ayrılır. K-1 seti eğitim için kullanılırken, geriye kalan bir set test için kullanılır. Algoritma, K kez eğitilir ve test edilir. Her yinelemede, modelin performansı kaydedilir ve sonuç aşamasında tüm performansların ortalaması alınır. Bu sayede modelin genel performansı daha güvenilir bir şekilde değerlendirilir.

2.7. Sınıf Dengesizliği Problemi İçin Kullanılan Algoritmalar

Sınıflandırma algoritmaları genellikle eğitim setlerinin iyi dengelendiğini varsaymaktadır. Algoritmaların amacı, doğru tahmin oranının maksimize edilmesidir. Ancak gerçek hayattaki veri kümelerinde sınıflardan bazıları çok az örnekle temsil edilebilirken, diğer sınıflar çok sayıda örnekle temsil edilmektedir. Bu durumda, etiket

bilgisi az olan örnekler için model yeterince eğitilmediğinden modelin bu grup için hatalı tahminler yapması muhtemeldir.

Dengesiz veri seti, bazı sınıflara ait gözlemlerin diğer sınıflara ait gözlemlere göre fazla olduğu veri setidir. Klasik sınıflandırıcılar bu veri setlerinde kullanılabilir sonuçlar vermemektedirler. Dengesiz veri setlerine odaklanan algoritmaların temel amacı doğruluk oranını artırma ve hata oranını azalmaktır. (Pir, 2022)

2.7.1. RUS (Rastgele az örnekleme)

Çoğunluk sınıfı gözlemlerinin azınlık sınıfı gözlem sayısına eşit olacak şekilde çoğunluk sınıfından rastgele gözlemlerin çıkarılmasına rastgele az örnekleme denmektedir. İlk durumda çoğunluk sınıfı azınlık sınıfından çok fazla gözleme sahip iken son durumda azınlık ile çoğunluk sınıfı gözlem sayıları birbirine eşittir. Bu işlemin yapılması, sınıflama yapılırken hem süreyi azaltmakta hem de sınıflama performanslarının doğruluğunu arttırmaktadır.

2.7.2. ROS (Rastgele aşırı örnekleme)

Bu yöntem, azınlık sınıfında bulunan gözlemler arasından rastgele gözlemleri tekrar azınlık sınıfına ekleyerek gözlem sayısının artmasını sağlayan örnekleme yöntemidir. İlk durumda çoğunluk sınıfı azınlık sınıfından çok fazla gözleme sahip iken son durumda azınlık sınıfı gözlem sayısı artmaktadır. Bu yöntemin dezavantajlarından birisi sınıflama yapılırken süreyi arttırmasıdır.

2.7.3. SMOTE (Sentetik azınlık aşırı örnekleme)

SMOTE algoritması (Chawla ve ark. 2002), sentetik azınlık örnekler üretilmesi sağlayan bir aşırı örnekleme yöntemidir. Veri bilimi projelerinden en sık kullanılan yöntemlerden bir tanesidir. Yöntemin ana fikri, gereken aşırı örnekleme miktarına bağlı olarak azınlık sınıfının örnekleri arasındaki en yakın k komşudan rastgele komşular seçilir ve aşağıdaki adımlar takip edilerek yeni azınlık sınıfı örnekleri oluşturulur. Böylece, aşırı öğrenme sorununun önüne geçilir ve iyi bir sınıflandırma performansı sunulur.

Sentetik örnekler şu şekilde üretilir:

- İncelenen öznelik vektörü (E_i) ile en yakın komşusu (E_j) arasındaki fark alınır,
- Bu fark 0 ile 1 arasında rastgele bir sayı(δ) ile çarpılır,
- Çıkan sonuç incelenen özellik vektörüne eklenir ve yeni örnek oluşur.

$$E_{yeni} = E_i + (E_i - E_j) \delta \quad (2.1)$$

2.7.4. NearMiss

NearMiss, dengesizlik içeren veri kümelerinde kullanılan bir örnekleme yöntemidir. NearMiss, azınlık sınıfına ait örneklerin çoğunluk sınıfına ait örneklerle olan uzaklığını hesaplayarak bilgi kaybını önlemeyi hedeflemektedir. Bu yöntem, yapısal olarak K-En Yakın Komşu (K-NN) algoritmasına dayanmaktadır.

2.8. Tahmin Doğruluğu Yaklaşımı / Performans Ölçütleri

Sınıflandırma metrikleri, makine öğrenmesi modellerinin performansını değerlendirmek için kullanılmaktadır. Bir sınıflandırma modelinin performansını yansıtan çeşitli değerlendirme kriterleri bulunmaktadır. Her bir makine öğrenmesi modeli farklı kısıtlara ve çalışma prosedürlerine sahiptir. Bu nedenle, problem için uygun bir değerlendirme metriği seçmek, model kurulumu ve optimizasyonu için önemlidir. Karmaşıklık matrisi, makine öğrenmesi modelinin tahmin performansı hakkında bilgi sağlamaktadır (Çelik ve arkadaşları, 2022). Sınıflandırma problemleri için performans metriklerinin hesaplanmasında kullanılan gerçek ve tahmin edilen değerler, karmaşıklık matrisine dayanmaktadır. Karmaşıklık matrisindeki alanlara karşılık gelen ifadeler Çizelge 2.1.'de gösterilmektedir.

Çizelge 2.1. Karmaşıklık matrisi

Gerçek Değerler	Negatif	Yanlış Pozitif	Doğru Negatif
	Pozitif	Doğru Pozitif	Yanlış Negatif
		Pozitif	Negatif
Tahmin Değerleri			

DN (Doğru Negatif), gerçek sınıf etiketi negatif olan değerlere negatif sınıf tahmini yapılarak sonucun doğru tahmin edildiğini göstermektedir. DP (Doğru Pozitif), gerçek sınıf etiketi pozitif olan değerlere pozitif sınıf tahmini yapılarak sonucun doğru tahmin edildiğini göstermektedir. YP (Yanlış Pozitif), gerçek sınıf etiketi negatif olan değerlere pozitif sınıf tahmini yapılarak sonucun yanlış tahmin edildiğini göstermektedir. YN (Yanlış Negatif), gerçek sınıf etiketi pozitif olan değerlere negatif sınıf tahmini yapılarak sonucun yanlış tahmin edildiğini göstermektedir.

Çizelge 2.2.'de tüm performans ölçütlerinin hesaplanma yöntemleri gösterilmiştir. Doğruluk, veri setinde tahmin edilen değerlerin doğruluk oranını göstermektedir (Denklem 2.2). Kesinlik, test bölümünde pozitif olarak tahmin edilen örneklerin ne kadarının doğru sınıflandırıldığını gösterir (Denklem 2.3). Duyarlılık, pozitif olarak tahmin etmemiz gereken işlemlerin ne kadarını pozitif olarak tahmin ettiğimizi gösteren bir metriktir. Duyarlılık oranının düşük çıkması yanlışların fazla olduğunu doğruların az olduğunu göstermektedir (Denklem 2.4). F-Skor, kesinlik ve duyarlılığın harmonik ortalamasını vermektedir (Denklem 2.5). Makro F-skor metriği, F-skor değerlerinin ortalamasını temsil etmektedir.

Çizelge 2.2. Sınıflandırma metrikleri

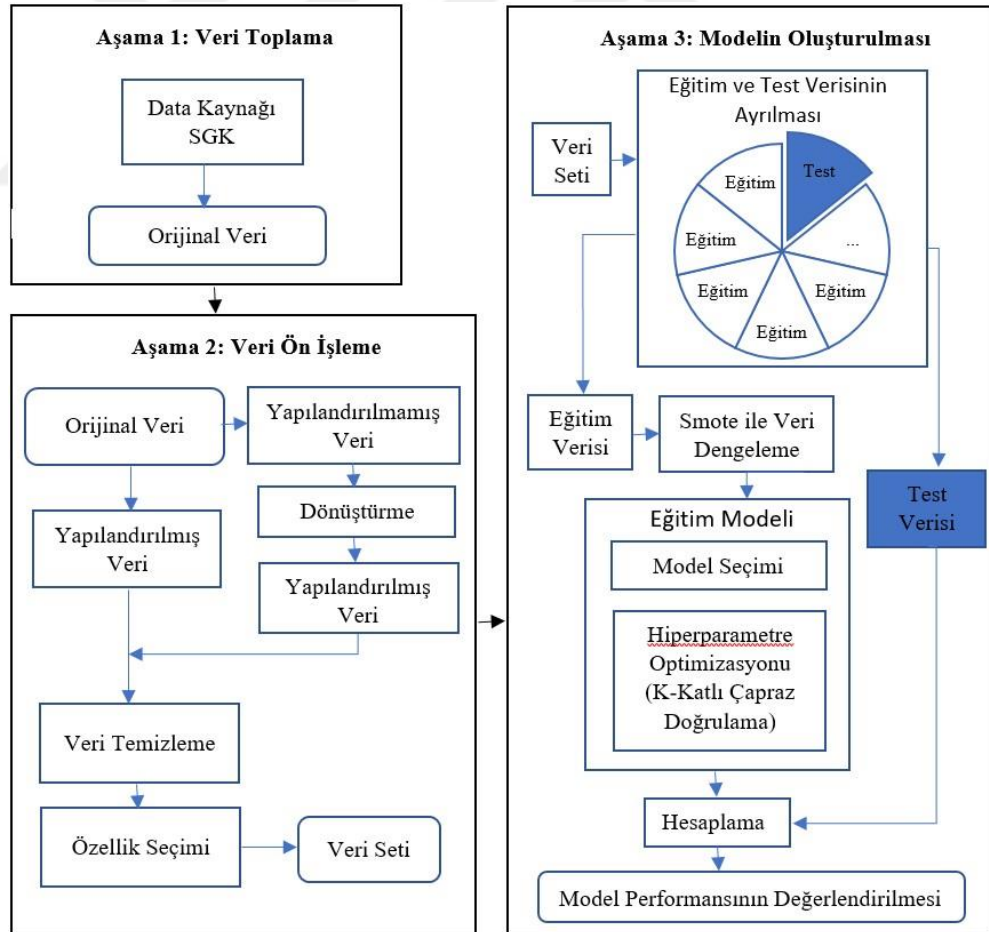
Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F- Skor
$\frac{TP+TN}{TP+TN+FP+FN}$ (2.2)	$\frac{TP}{TP + FP}$ (2.3)	$\frac{TP}{TP + FN}$ (2.4)	$2 * \frac{\text{kesinlik} * \text{duyarlılık}}{\text{kesinlik} + \text{duyarlılık}}$ (2.5)

3. MATERYAL ve YÖNTEM

Çalışma kapsamında, tekstil sektöründeki iş kazaları 3 aşamada analiz edilmiştir. İlk aşamada analiz için gerekli veriler toplanmış, ikinci aşamada toplanan veriler ön işlemden geçirilmiş ve son aşamada modeller oluşturulmuş ve performansları değerlendirilmiştir. Bu bölümde analiz için kullanılan yöntemler süreç haritasına uygun şekilde anlatılmıştır (Şekil 3.1).

3.1. Süreç Özeti

Bu çalışmada, Şekil 3.1’de gösterilen süreç haritası izlenmiştir. Bu yol haritasında veri madenciliği sürecinde takip edilen tüm adımlar ayrıntılı olarak gösterilmiştir. İlk aşamada, son 3 yılda tekstil sektöründe meydana gelen iş kazası verisi çalışma için gerekli görülen öznitelikler bağlamında SGK’ dan talep edilmiştir.



Şekil 3.1. Veri madenciliği proses haritası

İkinci aşamada, SGK tarafından gönderilen veri setindeki bilgiler incelenmiş ve bazı verilerin dönüşümü ile çalışma için anlamlı olduğu düşünülen yeni öznitelikler türetilmiştir. Yeni özniteliklerin eklenmesi ile elde edilen veri seti, veri ön işleme adımlarına tabi tutulmuştur. Veri setindeki tüm öznitelikler istatistiksel olarak incelenmiştir. Nümerik değişkenler içerisindeki aykırı değerler tespit edilip temizlenmiştir. Kategorik değişkenler içerisinde ise yeterli örnekleme sahip olmayan ifadeleri temizlemek amacıyla öznitelik bazında yüzdesel olarak sifıra denk gelen ifadeler silinmiştir. Veri temizleme işlemi sonrasında istatistik tabanlı özellik seçme yöntemleri aracılığıyla her bir girdi değişkeni ile hedef değişken arasındaki ilişki değerlendirilmiş ve hedef değişkenle en güçlü ilişkiye sahip olan girdi değişkenleri seçilmiştir.

Veri ön işleme süreci tamamlandıktan sonra son aşama olan model oluşturma aşamasına geçilmiştir. Bu aşamada veri seti, eğitim ve test verisi olmak üzere iki parçaya ayrılmıştır. Dengesiz dağılıma sahip olan sınıf etiketlerinin hem tahmin oranını yükseltmek hem de hata oranını düşürmek amacıyla farklı algoritmalar denenmiş ve en başarılı sonucu Sentetik Azınlık Yüksek Örnekleme Tekniği (SMOTE) algoritması vermiştir. Veri dengeleme yöntemleri yalnızca eğitim setine uygulanmalıdır. Test setine uygulanması durumunda değerlendirme sonucu yanıltıcı olacaktır. Bu doğrultuda dengeli hale getirilen eğitim veri seti için uygun modeller tespit edilmiştir. Seçilen makine öğrenmesi modelleri için belirlenen başarı metriğine göre en uygun hiperparametre kombinasyonları grid arama fonksiyonu ile oluşturulmuştur. Bu aşamada 5 katlı çapraz doğrulama işlemi uygulanmış ve farklı test veri setlerinden elde edilen doğruluk değerlerinin farklılık gösterme ihtimaline karşı önlem alınmıştır. Hazır hale gelen makine öğrenmesi modelleri eğitim veri seti ile eğitilmiştir. Eğitim süreci tamamlandıktan sonra test veri seti ile her model için doğruluk, kesinlik, duyarlılık ve F1 Skor değerleri hesaplanmış ve model performansları karşılaştırılmıştır.

3.1.1. Veri seti bilgileri / veri toplama

Bu çalışmada kullanılan veri seti, tekstil sektöründeki şirketler tarafından Sosyal Güvenlik Kurumu'na (SGK) bildirilen iş kazası verilerinden oluşmaktadır. SGK tarafından gönderilen veri seti 20 farklı özelliği içermektedir. Özelliklerin ismi, tanımı, tipi (kategorik/nümerik) ve kategorik olan özelliklerin sahip olduğu farklı ifadelerin

miktarı Çizelge 3.1'de gösterilmektedir. Nümerik olan özellikler yalnızca gerçek sayılar içerdiğinden farklı ifade miktarı '-' olarak gösterilmiştir. Çalışma kapsamında, 2019-2021 yılları arasında tekstil sektöründe gerçekleşen 89.963 iş kazası verisi analiz edilmiştir. Yaralanma türü, etkilenen vücut bölgeleri ve kaza sonucu gibi veriler, kazadan sonra kayda geçilebilecek verilerdir. Bu nedenle, bu çıktı verileri etiket seçenekleri olarak kullanılmıştır. Yapılan çalışmalar incelendiğinde, kaza sonucu verisi etiket olarak değerlendirildiğinde dengesiz veri seti sorununun modelin başarısını önemli ölçüde olumsuz etkilediği görülmektedir. Bu nedenle, yaralanma türü verisi etiket olarak seçilmiştir.

Çizelge 3.1. Veri seti bilgileri

Özellik İsmi	Tanım	Özellik Tipi	Miktar
Kimlik Numarası	İş kazası geçiren çalışanın kimlik numarası (kapalı ifade).	kategorik	89.963
Cinsiyet	İş kazası geçiren çalışanın cinsiyeti.	kategorik	2
Yaş	İş kazası geçiren çalışanın yaşı.	nümerik	-
Medeni Durum	İş kazası geçiren çalışanın medeni durumu.	kategorik	7
Alt Sektör	İş kazası geçiren çalışanın çalıştığı alt sektör.	kategorik	20
Şehir	İş kazası geçiren çalışanın yaşadığı şehir.	kategorik	77
İş Deneyimi	Çalışanın iş kazası geçirdiği şirketteki iş deneyimi (gün).	nümerik	-
Sicil Numarası	Çalışanın iş kazası geçirdiği şirketteki sicil numarası (kapalı ifade).	kategorik	89.963
Eğitim Durumu	İş kazası geçiren çalışanın eğitim durumu.	kategorik	11
İş Ünvanı	İş kazası geçiren çalışanın iş ünvanı.	kategorik	1242
Kaza Saati	Kazanın gerçekleştiği saat.	kategorik	24
Kaza Yılı	Kazanın gerçekleştiği yıl.	kategorik	3
Kaza Ayı	Kazanın gerçekleştiği ay.	kategorik	12
İş Başı Saati	Çalışanın kaza gününde işe başladığı saat.	kategorik	24

Çizelge 3.1. Veri seti bilgileri (devam)

Özellik İsmi	Tanım	Özellik Tipi	Miktar
Şirket Büyüklüğü	Kazanın meydana geldiği işyerindeki çalışan sayısı.	nümerik	-
Kaza Nedeni	İş kazasına neden olan durum / aktivite.	kategorik	9
Kullanılan Araç	Kaza meydana gelmeden hemen önce çalışan tarafından kullanılan araç, nesne veya alet.	kategorik	22
Vücudun Etkilendiği Bölge	İş kazası sonucunda çalışanın vücudunun etkilendiği bölgeler.	kategorik	9
Yaralanma Türü	İş kazası sonucunda meydana gelen yaralanma türü.	kategorik	14
Kaza Sonucu	İş kazası sonucunda çalışanın durumu.	kategorik	5

3.1.2. Veri ön işleme

Veri seti, tekstil sektöründeki şirketler tarafından SGK'ya bildirilen iş kazası verilerinden oluşmaktadır. SGK tarafından gönderilen veri seti 20 farklı özelliği içermektedir. Kimlik Numarası ve İşletme Sicil Numarası gibi özelliklerde her bir iş kazasına özgü belirli bilgiler bulunmaktadır (Kişisel Verilerin Korunması Kanunu gereği tüm kişisel bilgiler SGK tarafından bize kapalı ifadelerle sağlanmıştır). Veri ön işleme aşamasında ilk adım olarak bu ifadeler temizlenmiştir. Ayrıca, kazanın gerçekleştiği ay ve yıl verileri kullanılarak "pandemi-öncesi" ve "pandemi-sonrası" olarak tanımlanan yeni bir özellik türetilmiştir. Mart 2020'den önceki dönem "pandemi-öncesi" olarak tanımlanırken, mart 2020'den sonra olan dönem "pandemi-sonrası" olarak tanımlanmıştır. Bu özelliğin kazanın sonucu üzerindeki etkisi modelleme aşamasında test edilmiştir. Veri setine özellik ekleme ve çıkarma işlemleri yapıldıktan sonra, veri ön işleme adımları gerçekleştirilmiştir. Veri setindeki tüm özellikler istatistiksel olarak incelenmiştir. Kategorik özelliklerde, yeterli sayıda örnekleme sahip olmayan (örneklem oranı sıfır olan) ifadeler temizlenmiştir. 20 farklı özellik bilgisi ile tahmin edilmesi beklenen

yaralanma türü çıktı etiketi toplamda 14 farklı ifade içermektedir. Çizelge 3.2’de, veri setindeki her bir etiketin miktarı ve oranı belirtilmektedir.

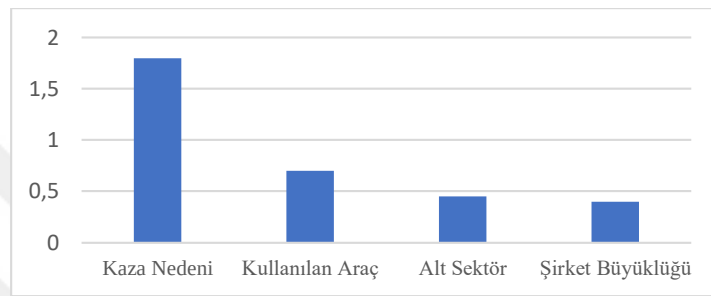
Çizelge 3.2. Yaralanma türü etiketleri

Etiket	Miktar	%
Yaralar ve yüzeysel yaralanmalar	44.707	49,7%
Diğer belirtilen yaralanmalar	19.126	21,3%
Çıkık, burkulma ve gerilmeler	13.503	15,0%
Bilinmeyen veya belirsiz yara türü	5.649	6,3%
Kemik kırıkları	3.129	3,5%
Yanıklar ve donma	1.696	1,9%
Zehirlenme ve enfeksiyonlar	1.430	1,6%
Şok	178	0,2%
Çoklu yaralanmalar	130	0,1%
Sarsıntı ve iç yaralanmalar	125	0,1%
Vücudun bir bölümünün kaybı	118	0,1%
Aşırı ısı, ışık ve radyasyon etkileri	53	0,1%
Ses, titreşim ve basınç etkileri	39	0,0%
Boğulma ve nefes darlığı	12	0,0%

'Diğer belirtilen yaralanmalar' ve 'Bilinmeyen veya belirsiz yara türü' açık bir tanımı olmadığı için analiz dışı bırakılmıştır. İş kazalarından kaynaklanan nadir yaralanma tiplerinin makine öğrenimi modelinin tahmin performansını etkilememesi için, örneklem büyüklüğü binde birin altında olan yaralanma türü etiketleri analize dahil edilmemiştir. Bu etiketler 'çoklu yaralanmalar', 'sarsıntı ve iç yaralanmalar', 'vücudun bir bölümünün kaybı', 'aşırı ısı, ışık ve radyasyon etkileri', 'ses, titreşim ve basınç etkileri', 'boğulma ve nefes darlığı' ve 'beyin sarsıntısı ve iç yaralanmalar' dır. Yapılan araştırma ve analizler sonucunda, 'çıkık, burkulma ve gerilmeler' ve 'kemik kırıkları' olarak tanımlanan yaralanma türlerinin benzer olaylar sonucunda meydana geldiği gözlemlenmiştir. Bu bilgilere dayanarak, iki etikete ait örnekler birleştirilerek 'çıkık, burkulma, gerilmeler ve kemik kırıkları' başlığı altında yeni bir etiket tanımlanmıştır. Veri Yaralanma türü etiketinin sayısal karşılığı şu şekildedir: yanıklar ve donma 0, yaralar ve yüzeysel

yaralanmalar 1, zehirlenme ve enfeksiyonlar 2, çıkık, burkulma, gerilmeler ve kemik kırıkları 3 ve şok 4.0 Değerlendirmeye alınan yaralanma türü etiketleri toplam verinin % 72'sine denk gelmektedir.

Veri temizleme işleminden sonra Filtreleme Tabanlı Özellik Seçme Yöntemi kullanılarak her bir girdi değişkeninin çıktı etiketiyle olan ilişkisi değerlendirilmiştir. Çıktı etiketiyle en güçlü ilişkiye sahip girdi değişkenlerinin sırasıyla kaza nedeni, kullanılan araç, alt sektör ve şirket büyüklüğü olduğu gözlemlenmiştir (Şekil 3.2).

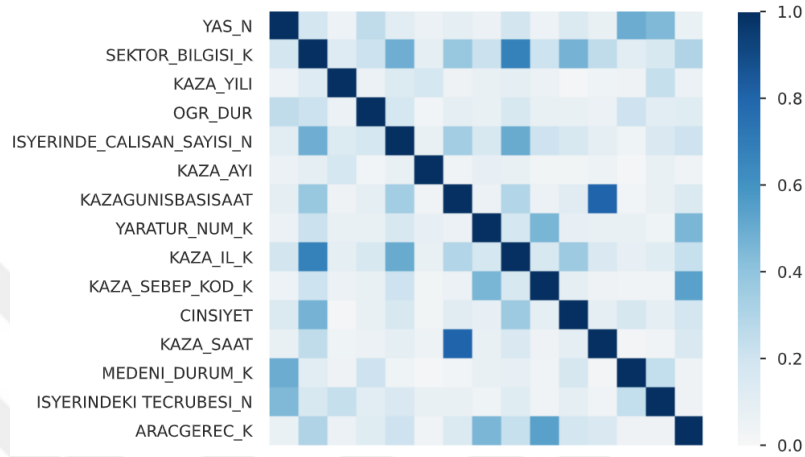


Şekil 3.2. Değişken etkinlik grafiği

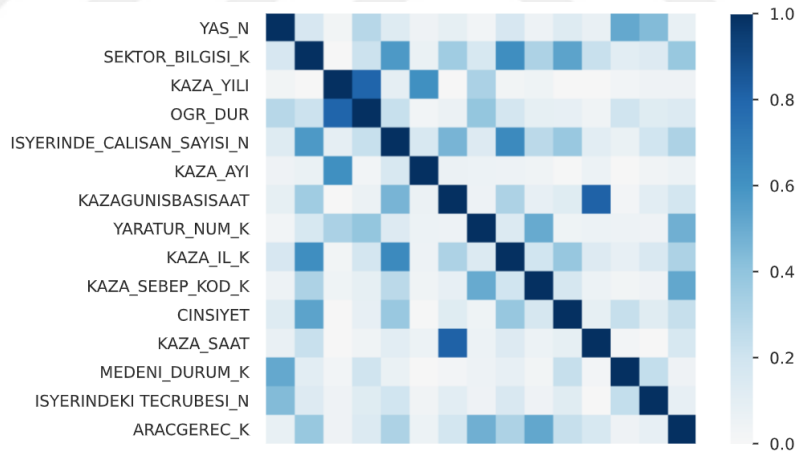
2020 yılında başlayan Covid-19 pandemisi, çalışanların mümkün olduğunca evden çalışmalarına ve işyerlerinden uzak kalmalarına neden olmuştur. Dolayısıyla, işyerinde çalışılan gün ve saatlerde azalma olmuştur. Pandeminin iş kazaları üzerindeki etkisi, kategorik değişkenler arasında tutarlı bir şekilde çalışan Phi_K korelasyon analizi kullanılarak incelenmiştir. Pandemi sürecinin dahil olduğu ve olmadığı iş kazaları ayrı korelasyon matrisleri üzerinde incelenmiştir.

Şekil 3.3, pandemi dönemini içeren üç yıllık (2019-2021) iş kazası verilerinin analiz sonuçlarını gösterirken, Şekil 3.4 ise sadece pandemi öncesi (2019 yılı ocak ayı başından 2020 yılı mart ayının sonuna kadar geçen süre) iş kazası verilerinin analiz sonuçlarını göstermektedir. Korelasyon matrisi üzerindeki dolgu renkleri, açık mavi ile koyu mavi arasındaki renk skalasını içermektedir. Korelasyon matrisindeki ikili değerlendirmelerde; korelasyon değeri 0 olan açık mavi dolgu rengi iki özellik arasında ilişki olmadığını, korelasyon değeri 1 olan koyu mavi dolgu rengi ise iki özellik arasında kuvvetli ilişki olduğunu göstermektedir. İki matrisinde dolgu renklerine bakıldığında pandemi sürecinin dahil olduğu korelasyon matrisindeki ikili ilişkilerin birçoğunun, pandemi sürecinin dahil

olmadığı korelasyon matrisindeki karşılıklıklarına göre korelasyon değerlerinin daha düşük olduğu gözlemlenmiştir. Bu durum pandemi sürecinin dahil edilmediği iş kazası veri setinde, özellik ve etiket ilişkilerinin daha güçlü olduğunu göstermektedir (Şekil 3.4). Bugün büyük ölçüde etkisini yitiren pandemi sürecinin analizi yanılmaması adına sadece pandemi öncesi meydana gelen iş kazaları analiz edilmiştir.



Şekil 3.3. Korelasyon Matris (2019-21)



Şekil 3.4. Korelasyon Matris (2019-20 Mart)

4. BULGULAR ve TARTIŞMA

Bu bölümde, ikinci bölümde açıklanan yöntemler adım adım iş kazası veri setine uygulanmış ve her adımda model performans değerleri kaydedilerek karşılaştırılmıştır. Veri ön işleme adımları (özellik seçimi, kategorik verilerin sayısal verilere dönüştürülmesi, normalizasyon), model oluşturma (veri dengesizliği, hiperparametre optimizasyonu) ve model performansını hesaplamak için yapılan tüm testler Python 3'te (Van Rossum, 2008) gerçekleştirilmiştir.

Veri temizleme sürecinde ilk olarak değişkenler incelenmiştir. Kategorik değişkenlerde, yeterli örnekleme sahip olmayan ifadeler temizlenmiştir. Ardından, çıktı etiketleri detaylı bir şekilde incelenmiştir. Belirsiz tanımlara sahip iki çıktı etiketi veri setinden çıkarılmıştır. Ayrıca, nadir görülen yaralanma tiplerinin makine öğrenme modelinin performansını etkilememesi için binde bir oranından az örneklem sayısına sahip olan yaralanma türü etiketleri analizden çıkarılmıştır. Son olarak, korelasyon analiziyle yanıltıcı etkisi kanıtlanan pandemi sürecinde meydana gelen iş kazaları veri setinden çıkarılmıştır. Veri ön işleme adımı yapılan analizler ve değerlendirmeler sonucunda, makine öğrenme algoritmalarının tahmin performansını optimize etmek için veri setindeki örnekleme sayısı 89.963'ten 11.710'a düşürülmüştür. Veri ön işleme adımı sonrasında veri setindeki değişkenlerin son durumu Çizelge 4.1'de gösterilmektedir. Nümerik olan özellikler için kategori sayısı '-' olarak gösterilmiştir.

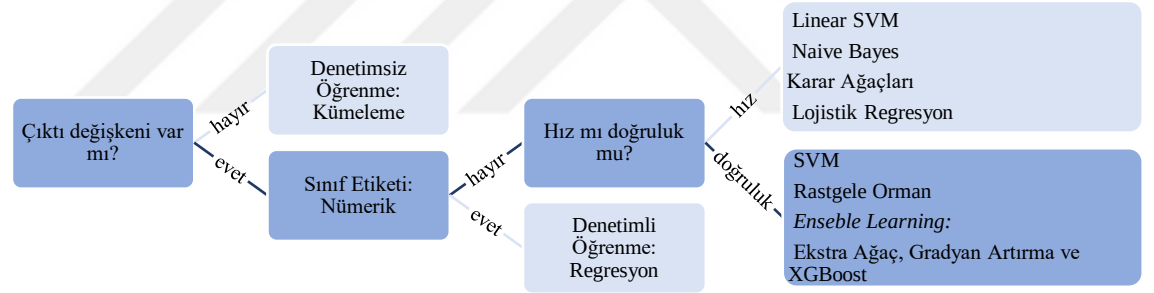
Çizelge 4.1. Ön işleme sonrası özelliklerin durumu

Özellik	Kategori Sayısı	Feature	Kategori Sayısı
Cinsiyet	2	Kaza Saati	24
Yaş	-	Kaza Yılı	2
Medeni Durum	6	Kaza Ayı	12
Alt Sektör	16	İş Başı Saati	24
Şehir	71	Şirket Büyüklüğü	-
İş Deneyimi	-	Kaza Nedeni	8
Eğitim Durumu	11	Kullanılan Araç	19
İş Ünvanı	567	Yara Türü	5

4.1. Modelin Oluşturulması

Kaza sonucu oluşan yara türlerinin doğru tahmin edilebilmesi için öncelikle tekstil sektörüne ait iş kazaları veri setini makine öğrenmesi için uygun hale getirecek yöntemler uygulanmış ve veri rassal olarak %80 oranında eğitim ve %20 oranında test setine ayrılmıştır. Daha sonra, veri setimizin girdi değişkenlerini normalleştirmek amacıyla özellik ölçeklendirme yöntemlerinden MaxAbs Ölçekleyici kullanılmıştır. Böylece tüm özellikler maksimum değeri kullanılarak ortak bir ölçeğe getirilmiştir.

Farklı veri tipleri ve farklı problemler için farklı tahmin modelleri kullanılmaktadır. Çalışmada kullanılan tahmin modellerinin seçiminde, Şekil 4.1'deki model seçimi haritası model seçimi konusunda yol gösterici olmuştur. Veri seti ve çalışmanın amacı göz önünde bulundurularak, SVM (Destek Vektör Makinesi), Rastgele Orman, Ekstra Ağaç, Gradyan Artırma ve XGBoost (Aşırı Gradyan Artırma) algoritmalarının kullanılmasına karar verilmiştir.



Şekil 4.1. Model seçim haritası

4.1.1. Veri dengeleme yöntemlerinin değerlendirilmesi

Sınıflandırma ve tahmin modellemelerinde, her bir örneklemin çıktısı (sınıf veya kategori), öğrenmenin gerçekleştiği girdi değişkenlerindeki bilgilere dayanarak tahmin edilmektedir. İş kazaları kapsamındaki çalışmalarda yaygın yaklaşım, çıktı değişkeni olarak kaza sonucu parametrelerinin ve girdi değişkeni olarak çalışan, iş, iş yeri ve kaza bilgilerinin kullanılmasıdır (Recal, 2022). Yapılan çalışmalar ve istatistikler incelendiğinde kaza sonuç bazında sınıflandırılan iş kazası veri setlerinin dengesiz dağılıma sahip oldukları görülmüştür. Yüksek oranda dengesiz veri kümelerinin sınıflandırılmasında, çoğunluk sınıfının lehine olan doğruluk değerleri modelin genel

başarısında yanıltıcı olacaktır. Dengesiz veri setlerinde model performansı ölçülürken düşük kesinlik (precision) ve yüksek duyarlılık (recall) veya tam tersi başarı metrikleri elde edilebilir. Bu durumda modelleri doğru bir şekilde karşılaştırmak zorlaşmaktadır. Bu durumda modellerin performansını uygun bir şekilde karşılaştırmak için F-skor metriği kullanılmalıdır. F1 skor, precision ve recall değerlerini aynı anda ölçerek değerlerin harmonik ortalamasını göstermektedir. Bu çalışmada, modelin oluşturulması ve değerlendirilmesi sürecinde F-skor metriği temel alınmıştır. Ayrıca, yüksek örneklem büyüklüğüne sahip olan etiketin yüksek F-skor değeri, ağırlıklı F-skor değerini yükseltmektedir. Bu nedenle model performansları, F-skor değerlerinin ortalamasını temsil eden ‘makro F-skor’ değeri kullanılarak karşılaştırılmıştır.

Dengesiz veri setlerinde tahmin gücünü arttırmak amacıyla kullanılan farklı veri dengeleme yöntemleri bulunmaktadır. Bu yöntemlerden RUS (Rastgele az örnekleme), ROS (Rastgele aşırı örnekleme), SMOTE (Sentetik Azınlık Aşırı Örnekleme), Near Miss algoritmaları eğitim veri setine uygulanmış ve en başarılı sonucu veren SMOTE (Sentetik Azınlık Aşırı Örnekleme) algoritması seçilmiştir. Çizelge 4.2’de tüm senaryoların başarı değerleri gösterilmektedir.

Çizelge 4.2. Veri dengeleme yöntemlerinin başarı metriklerine etkisi

Yöntemler	Metrikler	SVM	Ekstra Ağaç	Rastgele Orman	Gradyan Artırma	XGBoost
Orjinal	accuracy	0.76	0.76	0.77	0.76	0.76
	macro F-score	0.58	0.62	0.62	0.61	0.63
Near Miss	accuracy	0.25	0.34	0.34	0.42	0.37
	macro F-score	0.25	0.24	0.24	0.28	0.31
RUS	accuracy	0.61	0.61	0.61	0.59	0.63
	macro F-score	0.50	0.46	0.45	0.42	0.46
ROS	accuracy	0.76	0.76	0.76	0.75	0.76
	macro F-score	0.57	0.52	0.50	0.53	0.57
SMOTE	accuracy	0.76	0.76	0.76	0.75	0.75
	macro F-score	0.67	0.64	0.64	0.64	0.64

4.1.2. Hiperparametre optimizasyonu yönteminin değerlendirilmesi

Parametreler, makine öğrenimi algoritmalarının veri setindeki çıktı etiketlerini tahmin etmek için kullandığı yapılandırma değişkenleridir. Seçilen modellerden başarılı sonuçlar elde etmek için ideal hiperparametre değerlerinin belirlenmesi gerekmektedir. Bu amaçla, GridSearch yöntemi kullanılarak 5-kat 2-tekrar konfigürasyonu ile tüm olası kombinasyonlar denenmiş ve en iyi kombinasyonu sağlayan parametreler modellere uygulanmıştır. Hiperparametre seçimi modülleri, F-skor değerini optimize edecek şekilde tasarlanmıştır. Çizelge 4.3’ de veri dengeleme yöntemi ve hiperparametre optimizasyonu yönteminin başarı metriklerine etkisi görülmektedir.

Çizelge 4.3. Hiperparametre optimizasyonu yönteminin başarı metriklerine etkisi

Performans Geliştirme Adımları	Metrikler	SVM	Ekstra Ağaç	Rastgele Orman	Gradyan Artırma	XGBoost
Original	accuracy	0.76	0.76	0.77	0.76	0.76
	macro F-score	0.58	0.62	0.62	0.61	0.63
SMOTE	accuracy	0.76	0.76	0.76	0.75	0.75
	macro F-score	0.67	0.64	0.64	0.64	0.64
SMOTE +GRID SEARCH	accuracy	0.76	0.78	0.75	0.76	0.78
	macro F-score	0.69	0.68	0.64	0.65	0.70

4.1.3. Model performanslarının değerlendirilmesi

Veri ön işleme ve model oluşturma süreçleri veri madenciliği adımlarına uygun şekilde tasarlanmıştır. Smote ile veri dengeleme ve hiperparametre optimizasyonu adımları tamamlanan modeller test edilmiştir. Son aşama olan model performansının değerlendirilmesi sürecinde ise makine öğrenim algoritmalarının tahmin başarıları sınıflandırma raporu ve karmaşıklık matrisleri aracılığıyla karşılaştırılmıştır. Şekil 4.2 ve Şekil 4.3’ de SVM modelinin %69 makro F-skor ve %76 accuracy değerine ulaştığını ve ‘yanıklar, kaynar su ile kavrulma ve donmalar’ yara türü dışında tüm yaralanma türlerinin büyük oranda tahmin edildiği, Şekil 4.4 ve Şekil 4.5’ de XGBoost modelinin %70 makro F-skor ve %78 accuracy değerine ulaştığını ve ‘yanıklar, kaynar su ile kavrulma ve donmalar’ yara türü dışında tüm yaralanma türlerinin büyük oranda tahmin edildiği, Şekil

4.6 ve Şekil 4.7’de Gradyan Artırma modelinin %65 makro F-skor ve %76 accuracy değerine ulaştığını ve ‘yanıklar, kaynar su ile kavrulma ve donmalar’ ve ‘şok’ yara türü dışında tüm yaralanma türlerinin büyük oranda tahmin edildiği, Şekil 4.8 ve Şekil 4.9’da Rastgele Orman modelinin %64 makro F-skor ve %75 accuracy değerine ulaştığını ve ‘yanıklar, kaynar su ile kavrulma ve donmalar’ ve ‘şok’ yara türü dışında tüm yaralanma türlerinin büyük oranda tahmin edildiği, Şekil 4.10 ve Şekil 4.11’ de Ekstra Ağaç modelinin %68 makro F-skor ve %78 accuracy değerine ulaştığını ve ‘yanıklar, kaynar su ile kavrulma ve donmalar’ yara türü dışında tüm yaralanma türlerinin büyük oranda tahmin edildiği görülmektedir. Sınıf etiketlerinin sayısal eşdeğerleri aşağıdaki gibidir: yanıklar, kaynar su ile kavrulma ve donmalar: 0, yaralar ve yüzeysel yaralanmalar: 1, zehirlenme ve enfeksiyonlar: 2, çıkıklar, burkulmalar, incinmeler ve kemik kırıkları: 3, ve şok: 4.

	precision	recall	f1-score	support
Yanıklar, kaynar su ile kavrulma ve donmalar	0,48	0,37	0,42	87
Yaralar ve yüzeysel yaralanmalar	0,8	0,83	0,82	1505
Zehirlenme ve enfeksiyonlar	0,95	0,76	0,84	25
Çıkıklar, burkulmalar, incinmeler ve kemik kırıkları	0,71	0,68	0,7	868
Şok	0,82	0,6	0,69	15
accuracy			0,76	2500
macro avg	0,75	0,62	0,69	2500
weighted avg	0,76	0,69	0,76	2500

Şekil 4.2. SVM sınıflandırma raporu

	precision	recall	f1-score	support
Yanıklar, kaynar su ile kavrulma ve donmalar	0,54	0,33	0,41	87
Yaralar ve yüzeysel yaralanmalar	0,79	0,9	0,84	1505
Zehirlenme ve enfeksiyonlar	0,95	0,84	0,89	25
Çıkıklar, burkulmalar, incinmeler ve kemik kırıkları	0,78	0,63	0,69	868
Şok	0,89	0,53	0,67	15
accuracy			0,78	2500
macro avg	0,79	0,65	0,70	2500
weighted avg	0,78	0,78	0,77	2500

Şekil 4.4. XGBoost sınıflandırma raporu

predicted label \ true label	0	1	2	3	4
0	32	37	1	16	1
1	22	1256	0	226	1
2	2	4	19	0	0
3	11	264	0	593	0
4	0	4	0	2	9

Şekil 4.3. SVM karmaşıklık matrisi

predicted label \ true label	0	1	2	3	4
0	29	39	1	18	0
1	18	1351	0	135	1
2	1	2	21	1	0
3	6	317	0	545	0
4	0	5	0	2	8

Şekil 4.5. XGBoost karmaşıklık matrisi

	precision	recall	f1-score	support
Yanıklar, kaynar su ile kavrulma ve donmalar	0,57	0,33	0,41	89
Yaralar ve yüzeysel yaralanmalar	0,78	0,87	0,82	1473
Zehirlenme ve enfeksiyonlar	0,9	0,68	0,78	28
Çıkıklar, burkulmalar, incinmeler ve kemik kırıkları	0,73	0,63	0,68	903
Şok	0,75	0,43	0,55	7
accuracy			0,76	2500
macro avg	0,75	0,59	0,65	2500
weighted avg	0,75	0,76	0,75	2500

Şekil 4.6. Gradyan artırma sınıflandırma raporu

predicted label	0	1	2	3	4
0	29	37	0	23	0
1	9	1279	2	182	1
2	3	3	19	3	0
3	10	325	0	568	0
4	0	4	0	0	3
true label	0	1	2	3	4

Şekil 4.7. Gradyan artırma karmaşıklık matrisi

	precision	recall	f1-score	support
Yanıklar, kaynar su ile kavrulma ve donmalar	0,46	0,33	0,39	87
Yaralar ve yüzeysel yaralanmalar	0,76	0,87	0,81	1505
Zehirlenme ve enfeksiyonlar	0,86	0,76	0,81	25
Çıkıklar, burkulmalar, incinmeler ve kemik kırıkları	0,72	0,57	0,64	868
Şok	1	0,4	0,57	15
accuracy			0,75	2500
macro avg	0,76	0,59	0,64	2500
weighted avg	0,74	0,75	0,74	2500

Şekil 4.8. Rastgele orman sınıflandırma raporu

predicted label	0	1	2	3	4
0	29	38	1	19	0
1	21	1314	1	169	0
2	1	5	19	0	0
3	11	361	1	495	0
4	1	6	0	2	6
true label	0	1	2	3	4

Şekil 4.9. Rastgele orman karmaşıklık matrisi

	precision	recall	f1-score	support
Yanıklar, kaynar su ile kavrulma ve donmalar	0,64	0,32	0,43	87
Yaralar ve yüzeysel yaralanmalar	0,78	0,9	0,84	1505
Zehirlenme ve enfeksiyonlar	0,9	0,76	0,83	25
Çıkıklar, burkulmalar, incinmeler ve kemik kırıkları	0,78	0,63	0,69	868
Şok	1	0,47	0,64	15
accuracy			0,78	2500
macro avg	0,82	0,61	0,68	2500
weighted avg	0,78	0,78	0,77	2500

Şekil 4.10. Ekstra ağaç sınıflandırma raporu

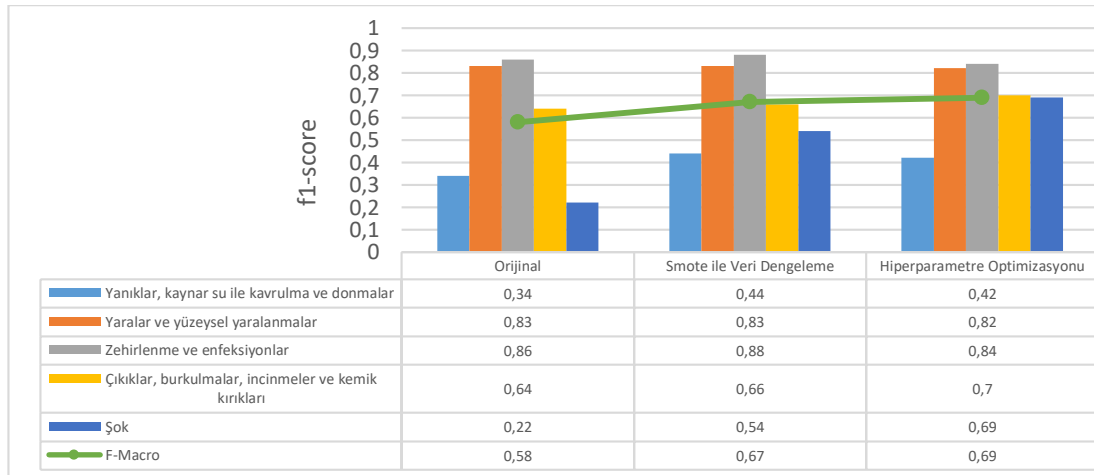
predicted label	0	1	2	3	4
0	28	42	1	16	0
1	12	1354	0	139	0
2	1	5	19	0	0
3	3	322	0	543	0
4	0	5	1	2	7
true label	0	1	2	3	4

Şekil 4.11. Ekstra ağaç karmaşıklık matrisi

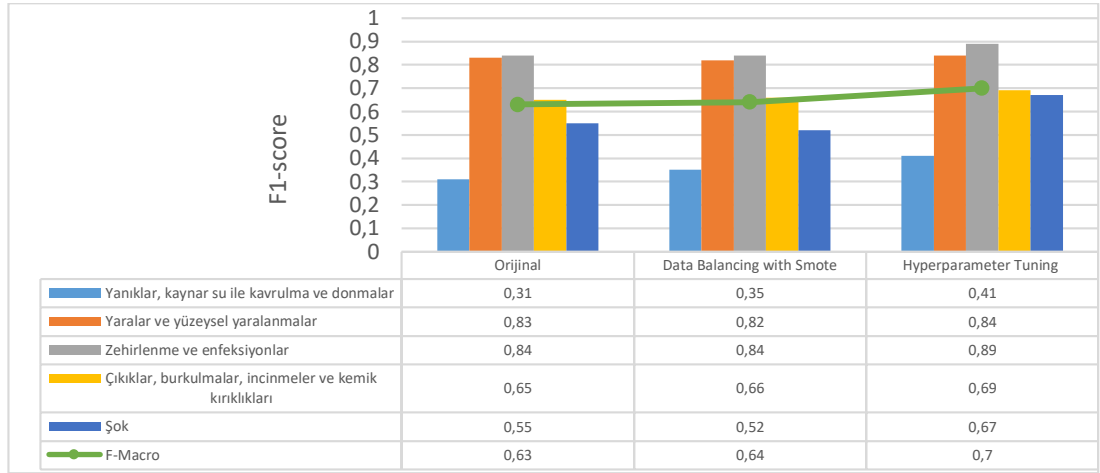
Yapılan analizler sonucunda tahmin başarısı en yüksek olan algoritmaların sırasıyla XGBoost, SVM, Ekstra Ağaç, Gradyan Artırma ve Rastgele Orman olduğu görülmüştür. Modelin oluşturulması sürecinde model performansını büyük ölçüde etkileyen veri dengeleme ve hiperparametre optimizasyonu adımlarının modellerin yara türleri üzerindeki tahmin başarıları üzerindeki etkileri Çizelge 4.4, 4.5, 4.6, 4.7 ve 4.8’de özetlenmiştir.

SVM modelinin %58 olan makro F-skor değerinin veri dengeleme yöntemi ile %67, ardından hiperparametre optimizasyonu yöntemi ile %69’a yükseldiği görülmüştür. XGBoost modelinin %63 olan makro F-skor değerinin veri dengeleme yöntemi ile %64, ardından hiperparametre optimizasyonu yöntemi ile %70’e yükseldiği görülmüştür. Gradyan Artırma modelinin %61 olan makro F-skor değerinin veri dengeleme yöntemi ile %64, ardından hiperparametre optimizasyonu yöntemi ile %65’e yükseldiği görülmüştür. Rastgele Orman modelinin %62 olan makro F-skor değerinin veri dengeleme yöntemi ile %64’e yükseldiği, ardından hiperparametre optimizasyonu yöntemi ile %64’te sabit kaldığı görülmüştür. Ekstra Ağaç modelinin %62 olan makro F-skor değerinin veri dengeleme yöntemi ile %64, ardından hiperparametre optimizasyonu yöntemi ile %68’e yükseldiği görülmüştür. Elde edilen sonuçlara göre, XGBoost algoritmasının en yüksek tahmin doğruluk yüzdesini vermesi nedeniyle, kaza tahmin modellerinde daha başarılı bir temel oluşturacağı öngörülmektedir.

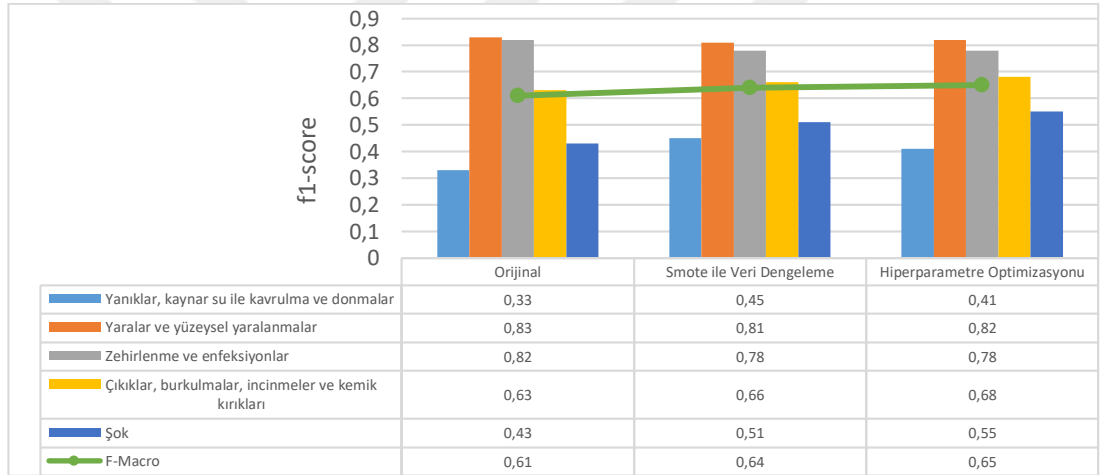
Çizelge 4.4. SVM modelinin yara türlerini tahmin etme başarısı



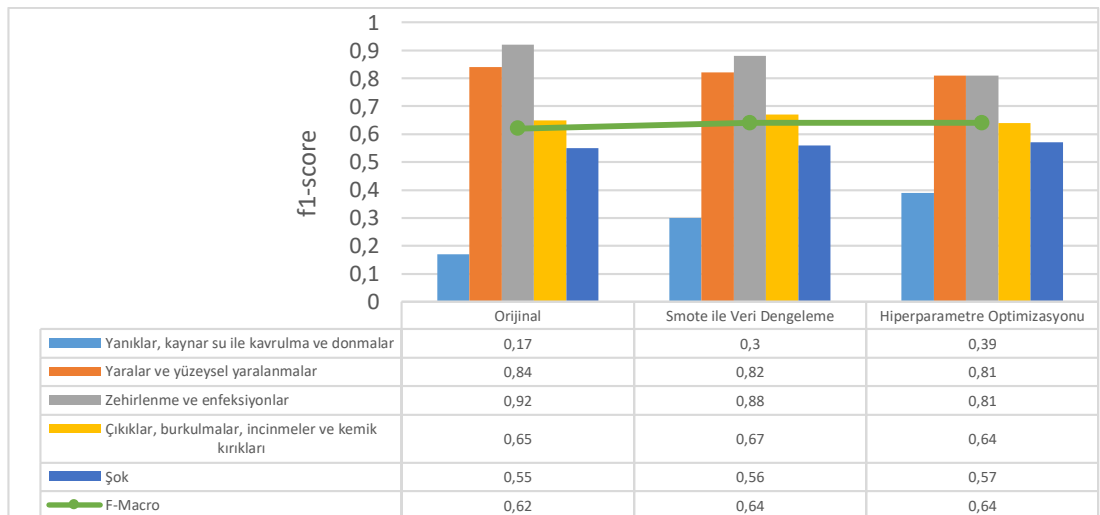
Çizelge 4.5. XGBoost modelinin yara türlerini tahmin etme başarısı



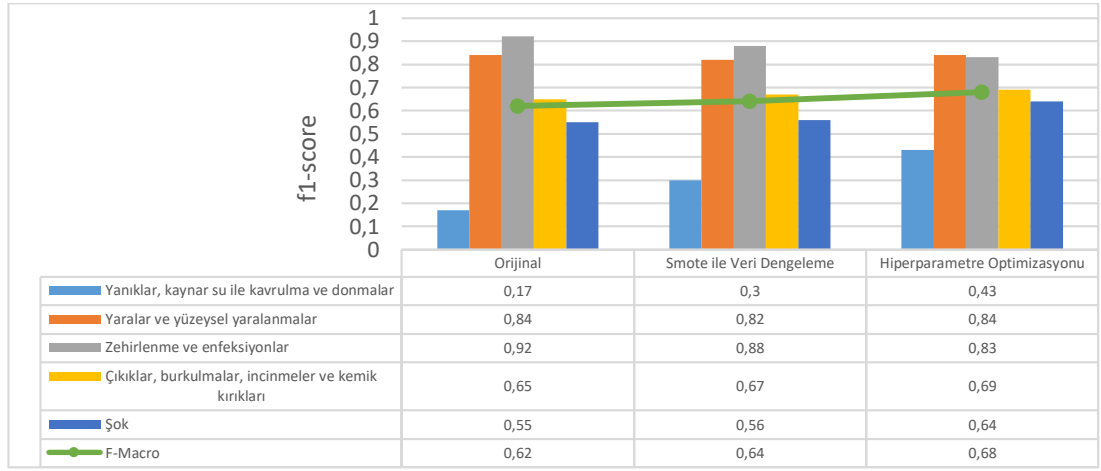
Çizelge 4.6. Gradyan Artırma modelinin yara türlerini tahmin etme başarısı



Çizelge 4.7. Rastgele orman modelinin yara türlerini tahmin etme başarısı



Çizelge 4.8. Ekstra ağaç modelinin yara türlerini tahmin etme başarısı



5. SONUÇ

Bu çalışma kapsamında, tekstil sektöründeki işyerlerine ait 2019-2021 yıllarında gerçekleşen kaza verileri kullanılarak, veri madenciliği yöntemi ile iş kazalarının kontrolüne yönelik risk faktörleri analiz edilmiş ve kaza tahmin modeline temel oluşturacak model performansları değerlendirilmiştir. Bu kapsamda, SGK tarafından gönderilen tekstil sektörü işyerlerinde gerçekleşen iş kazalarından bir kaza veri kümesi oluşturulmuştur. Kaza veri kümesi, veri analiz yöntemleri uygulanarak makine öğrenmesi için daha anlamlı bir veri kümesine dönüştürülmüştür. Sonrasında veri kümesine ve çalışmanın amacına en uygun algoritmalar seçilmiştir. Veri dengeleme ve hiperparametre optimizasyonu ile makine öğrenmesi modellerinin veri kümesini daha iyi performans göstererek yorumlaması sağlanmıştır. Böylece; SVM algoritması makro F-skör değeri %58'den %69'a, Ekstra Ağaç algoritması makro F-skör değeri %62'den %68'e, Rastgele Orman algoritması makro F-skör değerini %62'den %64'e, Gradyan Artırma algoritması makro F-skör değeri %61'den %65'e ve XGBoost algoritması makro F-skör değeri %63'ten %70'e yükselmiştir. Yapılan çalışmalar neticesinde veri kümesini en iyi yorumlayan makine öğrenmesi algoritmasının XGBoost olduğu görülmüştür. Geliştirilen modelin veri kümesi içerisinde yer alan kaza olaylarını 5 kat 2 tekrar konfigürasyonu ile denemesi sonucunda, %70 oranında doğrulukla çalıştığı ortaya konulmuştur. SVM algoritması %69 ve Ekstra Ağaç algoritması %68 makro F-skör tahmin başarıları oranıyla veri kümesini iyi derecede yorumlayan algoritmalar arasına girmiştir.

Veri kümesinde tahmin değişkeni olarak seçilen yara türü etiketlerinin örneklem bazında dengesiz dağılımı, başarı metriği olarak seçilen makro F-skör değerinin düşük çıkmasına neden olmaktadır. Hem veri kümesinde hem de tahmin modellerinde de yapılan revizyonlar sonucu makro F-skör değeri daha iyi seviyelere getirilmiş oldu. En iyi performansı veren XGBoost algoritmasının yara türü etiketleri bazında başarı değerleri sırasıyla; zehirlenme ve enfeksiyonlar %89, yaralar ve yüzeysel yaralanmalar %84, çıkıklar, burkulmalar, incinmeler ve kemik kırıklar %69, şok %67 ve yanıklar, kaynar su ile kavrulma ve donmalar %41 olarak tespit edilmiştir. Örneklem sayıları ve tahmin başarıları değerleri birlikte değerlendirildiğinde; zehirlenme ve enfeksiyonlar yara türünün

az örneklem ile test edilmesine rağmen yüksek oranda tahmin başarısına sahip olduğu görülmüştür.

Bu çalışma kapsamında incelenen kaynaklara göre, dengesiz dağılıma sahip veri setleri için XGBoost algoritmasının diğer algoritmalara göre daha iyi performans sergilediği gözlemlenmiştir. Ayrıca, veri dengeleme ve hiperparametre optimizasyonu yöntemlerinin, dengesiz dağılıma sahip veri setleri için tahmin başarısı açısından önemli olduğu belirlenmiştir. Sonuç olarak, veri madenciliği adımlarıyla oluşturulan bir kaza tahmin modelinin, tekstil sektöründe meydana gelebilecek önemli riskleri belirleme konusunda kullanılabileceği anlaşılmıştır. Kaza tahmin modeli ile tekstil sektöründe faaliyet gösteren işyerlerinde meydana gelen kazalar sonucunda oluşabilecek yaralanma türleri, çalışanlar ve çalışma süreçleriyle ilgili temel kriterlere dayanarak tahmin edilebileceği görülmüştür. Yaralanma türlerinin tahmin edilmesi, tehlikeli durumların belirlenmesine ve tekstil sektörü işyerleri için risk haritaları oluşturulmasına katkı sağlayacaktır. Bu nedenle, kaza tahmin modellerinin yaygın kullanımıyla işyerlerinde iş sağlığı ve güvenliği (OSH) personelinin potansiyel kazalara karşı proaktif önlemler alması ve iş kazalarının sayı ve şiddetinde azalma görülmesi beklenmektedir. Gelecekteki çalışmalarda araştırmacılar, daha kapsamlı bir kaza tahmin modeli geliştirmek için evrensel verilerle veri setini genişleterek bu çalışmanın bulgularını geliştirebilirler. Ayrıca, tehlikeli durumları tespit edebilen ve tekstil sektöründeki işçilere ve yöneticilere uyarı verebilen gerçek zamanlı izleme sistemlerinin geliştirilmesine odaklanabilir. Bu sistemler, işyerinden gelen verileri analiz etmek ve kazalara yol açmadan önce potansiyel riskleri tespit etmek için makine öğrenmesi algoritmalarını kullanabilir.

KAYNAKLAR

- Ayanoğlu C. & Kurt M. (2019). Proposal of an occupational accident forecasting model with data mining methods in metal sector. *Ergonomi*, 2(2), 78–87. doi: <https://doi.org/10.33439/ERGONOMI.481861>
- Aghalarova S. & Bozkurt Keser S. (2022). Öğrencilerin Akademik Performanslarının Tahmin Edilmesi için AutoML Tekniğinin Uygulanması. *El-Cezeri*, 9(2), 394 – 412. <https://doi.org/10.31202/ecjse.946505>
- Ahmad M. W., Reynolds J. & Rezgui Y. (2018). Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. *Journal of cleaner production*, 203, 810-821. doi: <https://doi.org/10.1016/j.jclepro.2018.08.207>
- Bulut F. (2016). Performance Analysis of Ensemble Methods on Imbalanced Datasets. *Bilişim Teknolojileri Dergisi*, 9(2), 153-159. doi: 10.17671/btd.81137
- Beniwal S. & Arora J. (2012). Classification and Feature Selection Techniques in Data Mining. *International Journal of Engineering Research & Technology (IJERT)*, 1(6).
- Cheng C. W., Yao H. Q. & Wu T. C. (2013). Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, 26(6), 1269-1278. doi: <https://doi.org/10.1016/j.jlp.2013.07.002>
- Choi J., Gu B., Chin S. & Lee J.S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. *Automation in Construction*, 110, 102974. doi: <https://doi.org/10.1016/j.autcon.2019.102974>
- Çakmak H. (2019). Analysis of current occupational accidents raw data by data mining process (Publication No. 614088) [Master's dissertation, Gazi University]. Ulusal Tez Merkezi.
- Çakır E. (2019). Workplace hazards and occupational risks: A research on occupational accidents aboard merchant ships (Publication No. 564634) [Doctoral dissertation, Dokuz Eylül University]
- Çelik E. (2019). Customer Churn Analysis Based On Machine Learning By Using Data Mining Techniques In Telecommunication Sector. [Master's dissertation, Yeditepe University]
- Çelik S., Çetinkaya Bozkurt Ö. & Ekşili N. (2022). Çalışan Performansı Ölçeğindeki İfadelerin Karar Ağacı Algoritması ile Belirlenmesi. *Mehmet Akif Ersoy Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 9(1), 561-584. doi: 10.30798/makuiibf.913478
- Çelik H.İ. & Dülger L.C. (2022). A Novel Industrial Application of CNN Approach: Real Time Fabric Inspection and Defect Classification on Circular Knitting Machine. *Textile and Apparel*, 32(4), 344-352. doi: <https://doi.org/10.32710/tekstilvekonfeksiyon.1017016>
- Daldır I. (2021). Havacılık Sektöründe Uçuş Gecikmelerinin Makine Öğrenmesi Temelli Analizi Ve Tahmini. [Doctoral dissertation, Akdeniz University]
- Ersöz F. & Çınar Y. (2021). Veri Madenciliği ve Makine Öğrenimi Yaklaşımlarının Karşılaştırılması: Tekstil Sektöründe bir Uygulama. *Avrupa Bilim ve Teknoloji Dergisi*, 29, 397 – 414. <https://doi.org/10.31590/ejosat.1035124>
- Fernández A., García S., Herrera F. & Chawla N.V. (2018). SMOTE for Learning from

- Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61, 863-905. doi: <https://doi.org/10.1613/jair.1.11192>
- Güllüoğlu, E. N. & Taçgın, E. (2018). Türkiye Tekstil Sektöründe İstihdam ve İş Kazalarının Analizi. *Tekstil ve Mühendis*, 25(112), 344-354. <https://doi.org/10.7216/1300759920182511208>
- Gül M., Guneri A.F., Yılmaz F. & Celebi O. (2016). Analysis of the relation between the characteristics of workers and occupational accidents using data mining. *The Turkish Journal of Occupational / Environmental Medicine and Safety*, 1(4), 102-118.
- Gürsoy G. (2022). Makine öğrenmesi algoritmaları ile kalp hastalığı tahmini. [Master's dissertation, Maltepe University]
- Koc K., Ekmekcioğlu Ö. & Gurgun A.P. (2021). Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. *Automation in Construction*, 131, 103896. doi: <https://doi.org/10.1016/j.autcon.2021.103896>
- Kakhki FD., Freeman SA. & Mosher G.A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science*, 117, 257-262. doi: <https://doi.org/10.1016/j.ssci.2019.04.026>
- Khairuddin M.Z.F., Hui P.L., Hasikin K., Razak N.A.A., Lai K.W., Saudi A.S.M. & Ibrahim S.S. (2022). Occupational Injury Risk Mitigation: Machine Learning Approach and Feature Optimization for Smart Workplace Surveillance. *International Journal of Environmental Research and Public Health*, 19(21), 13962. doi: <https://doi.org/10.3390/ijerph192113962>
- Karaibrahimoğlu A. (2014). Veri Madenciliğinden Birliktelik Kuralı ile Onkoloji Verilerinin Analiz Edilmesi: Meram Tıp Fakültesi Onkoloji Örneği. [Doctoral dissertation, Selçuk University]
- Kaynar O., Arslan H., Görmez Y. & Işık Y. E. (2018). Makine Öğrenmesi ve Öznitelik Seçim Yöntemleriyle Saldırı Tespiti. *Bilişim Teknolojileri Dergisi*, 11(2), 175-185. doi: 10.17671/gazibtd.368583
- Khorsheed, R. M. (2021). Building sensor-based real-time predictive maintenance system by utilizing artificial intelligent techniques. [Doctoral dissertation, Istanbul Technical University]
- Koc K. & Gurgun A.P. (2022). Scenario-based automated data preprocessing to predict severity of construction accidents. *Automation in Construction*, 140, 104351. doi: <https://doi.org/10.1016/J.AUTCON.2022.104351>
- Mathews D.G. (2016). Data Mining and Machine Learning Algorithms for Workers' Compensation Early Severity Prediction. [Master's dissertation, Middle Tennessee State University]
- Özekes S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Dergisi*, 65-82.
- Özlüer Başer B., Yangın M. & Sarıdaş S. (2021). Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 25(1), 112-120. doi: 10.19113/sdufenbed.842460
- Pir M. Ş. (2022). Dengesiz Veri Setlerinde Sınıflandırma Problemlerinin Çözümünde Melez Yöntem Uygulaması. (Publication No. 720846) [Doctoral dissertation, Bursa Uludağ University]
- Recal, F. & Demirel, T. (2021). Predicting accident severity with machine learning. *Journal of Intelligent & Fuzzy Systems*, 40, 10981–10998. doi:10.3233/JIFS-202099

- Recal F. (2022). Creating a decision support framework from national occupational accidents data with machine learning approach (Publication No. 743913) [Doctoral dissertation, Yıldız Teknik University] Ulusal Tez Merkezi.
- Santos A.J.R. (2021, February 11). Learning from work-related accidents [Conference presentation]. Towards safe, healthy and declared work in Ukraine – ILO, Ukraine. https://www.ilo.org/wcmsp5/groups/public/---europe/---ro-geneva/---sro-budapest/documents/genericdocument/wcms_769666.pdf
- Sarkar S., Vinay S., Raj R., Maiti J. & Mitra P. (2019). Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, 106, 210-224. doi: <https://doi.org/10.1016/j.cor.2018.02.021>
- Serinken M., Türkçüer İ., Dağlı B., Karcioğlu Ö., Zencir M., Uyanık E. (2012). Work-related injuries in textile industry workers in Turkey. *Ulus Travma Acil Cerrahi Dergisi*, 18(1), 31-36. doi: 10.5505/tjtes.2012.54376
- Shekar B.H. & Dagnev G. (2019). Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data [Conference presentation]. Second International Conference on Advanced Computational and Communication Paradigms (1-8), Gangtok, Hindistan. doi:10.1109/ICACCP.2019.8882943
- Sarkar S., Khatedi N., Pramanik A. & Maiti J. (2019). *An Ensemble Learning-Based Undersampling Technique for Handling Class-Imbalance Problem*. Lecture Notes in Electrical Engineering, Springer.
- Savaş S. (2020). Veri Madenciliğinin Tarihi ve Kullanıldığı Alanlar. <https://medium.com/veri-madencili%C4%9Fi/veri-madencili%C4%9Finin-tarihi-ve-kullan%C4%B1d%C4%B1%C4%9F%C4%B1-alanlar-19a1896c1505>
- Taş Y. (2018). Birliktelik Kuralları Madenciliği ve Bir Uygulama. (Publication No. 494114) [Master's dissertation, Cumhuriyet University]
- Tuncer E. & Bolat E. D. (2021). Destek Vektör Makinaları ile EEG Sinyallerinden Epileptik Nöbet Sınıflandırması. *Politeknik Dergisi*, 25(1), 239-249. doi: 10.2339/politeknik.672077
- Wu J., Chen X.Y., Zhang H., Xiong L.D., Lei H. & Deng S.H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, 17(1), 26-40. <https://doi.org/10.11989/JEST.1674-862X.80904120>
- Yang L. & Shami A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415, 295-316. doi: <https://doi.org/10.1016/j.neucom.2020.07.061>