ENDÜSTRİYEL SİSTEMLERDE VERİ MADENCİLİĞİ YAKLAŞIMLARININ KULLANIMI ve BİR UYGULAMA

Esengül GÜRBÜZ

T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ENDÜSTRİYEL SİSTEMLERDE VERİ MADENCİLİĞİ YAKLAŞIMLARININ KULLANIMI ve BİR UYGULAMA

Esengül GÜRBÜZ
502006023

Dr. Öğr. Üyesi Mehmet AKANSEL

YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA – 2023
Her Hakkı Saklıdır

ÖZET

Yüksek Lisans Tezi

ENDÜSTRİYEL SİSTEMLERDE VERİ MADENCİLİĞİ YAKLAŞIMLARININ KULLANIMI ve BİR UYGULAMA

Esengül GÜRBÜZ

Bursa Uludağ Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Mehmet AKANSEL

Bu çalışmada üretim sektöründe faaliyet gösteren bir firmanın kaynak makine verileri ele alınmıştır. Gerçekleştirilen çalışma ile kaynak sonucunda oluşan kalite ve kalitesizliği belirleyen değişkenlerin, endüstriyel sistemlerde yaygınlaşmakta olan veri madenciliği yaklaşımlarının sistematik bir biçimde uygulanması ile tespiti hedeflenmektedir.

Veri madenciliği uygulamalarına hazır hale getirmek ve elde edilecek sonuçların kalitesini arttırmak amacıyla veri setine veri önişleme adımları uygulanmıştır. Uygulamaya hazır hale getirilen veriler, kullanılan modellerin performanslarını ölçmek amacıyla eğitim ve test kümelerine ayrılmıştır. Veri madenciliği yöntemlerinden gözetimli öğrenme algoritmaları uygulanmış ve uygulanan modellerin performansları farklı senaryolar için değerlendirilmiştir. Uygulama sonucunda en yüksek performans AdaBoost karar ağacı modeli ile elde edilmiş ve sonuçlar yorumlanmıştır.

Anahtar Kelimeler: Veri madenciliği, üretim süreçleri, kalite kontrol, veri tahminleme, gözetimli öğrenme

ABSTRACT

MSc Thesis

USE OF DATA MINING APPROACHES IN INDUSTRIAL SYSTEMS
and AN APPLICATION

Esengül GÜRBÜZ

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

Supervisor: Asst.Prof. Üyesi Mehmet AKANSEL

In this study, welding machine data of a company operating in the production sector are discussed. The aim of the study is to determine the variables that determine the quality and poor quality resulting from welding, by systematically applying data mining approaches that are becoming widespread in industrial systems.

Data preprocessing steps were applied to the data set in order to make it ready for data mining applications and to increase the quality of the results to be obtained. The data made ready for implementation are divided into training and test sets in order to measure the performance of the models used. Supervised learning algorithms, one of the data mining methods, were applied and the performances of the applied models were evaluated for different scenarios. As a result of the application, the highest performance was obtained with the AdaBoost decision tree model and the results were interpreted.

Keywords: Data mining, production process, quality control, data forecasting, supervised learning

TEŞEKKÜR

Bu tez çalışmasında endüstriyel sistemlerde son zamanlarda artış gösteren veri madenciliği yaklaşımlarına yönelik bir çalışma gerçekleştirilmiştir.

Tez konusu seçiminde isteklerimi göz önünde bulundurup bana yardımcı olan, tez çalışmamın planlanmasında, araştırılmasında ve tamamlanmasında büyük emek veren değerli tez danışman hocam Dr. Öğr. Üyesi Mehmet AKANSEL teşekkürlerimi sunarım. Tez çalışmamda kaynak aramak için yardım talep ettiğim Prof. Dr. Kadir Çavdar hocama ve tüm eğitim hayatım boyunca benden maddi ve manevi desteklerini esirgemeyen her zaman yanımda olan sevgili aileme teşekkürlerimi bir borç bilirim.

Esengül GÜRBÜZ

23/06/2023

İÇİNDEKİLER

Sayfa

ÖZET iv
ABSTRACT v
TEŞEKKÜR vi
SİMGELER ve KISALTMALAR DİZİNİ viii
ŞEKİLLER DİZİNİ ix
ÇİZELGELER DİZİNİ x
1. GİRİŞ 1
2. KAYNAK ARAŞTIRMASI 3
2.1. Veri Madenciliğinin Gelişimi 4
2.2. Veri Madenciliği Modelleri & Metotları 7
2.3. Karar Ağaçları 8
2.4. Lojistik Regresyon 8
2.5. Destek Vektör Makineleri 9
2.6. Veri Madenciliği ile İlgili Çalışmalar 9
3. MATERYAL ve YÖNTEM 12
3.1. Veri Ön İşleme 13
4. BULGULAR 18
5. TARTIŞMA ve SONUÇ 29
KAYNAKLAR 31
EK 1 Aykırı Değer Analizi Sonuçları 33
EK 2 Model Performans Hesaplama 36
ÖZGEÇMİŞ 40

SİMGELER ve KISALTMALAR DİZİNİ

Kısaltmalar	Açıklama
DVM	Destek Vektör Makineleri
EM	Expectation Maximization
KNN	K-Nearest Neighbors
SMO	Sequential Minimal Optimization
VTBK	Veri Tabanından Bilgi Keşfi
PUKÖ	Planla- Uygula- Kontrol Et- Önlem Al

ŞEKİLLER DİZİNİ

Sayfa

Şekil 2.1. Yıl Bazında Üretimde Makine Öğrenmesi ve Veri Madenciliği Kullanımı ……..…. 4
Şekil 2.2. VTBK Sürecinin Aşamaları ……..… 5
Şekil 2.3. DVM Örneği…… ……..…….. 9
Şekil 3.1. Süreç Özeti ……..…….…… 12
Şekil 3.2. Aykırı Değer Analizi Kutu Grafiği …….……..…. 14
Şekil 3.3. Aykırı Değer Analizi İstatistiki Sonuçları ….……...…...…. 15
Şekil 3.4. Normalizasyon İşlemi ……..…….....……. 15
Şekil 4.1. Karar Ağacı Modeli …….…..…… 19
Şekil 4.2. DVM Modeli ……..……...……..……. 21
Şekil 4.3. Korelasyon Matrisi ……..…..…. 23
Şekil 4.4. Özellik Seçimi ile Karar Ağacı Modeli …….....……. 24
Şekil 4.5. Özellik Seçimi ile DVM Modeli ……..…….……. 26

ÇİZELGELER DİZİNİ

		Sayfa
Çizelge 2.1.	Veri Tipleri …….	5
Çizelge 3.1.	Değişkenler …….	13
Çizelge 4.1.	Karar Ağacı Hata Matrisi …….	18
Çizelge 4.2.	AdaBoost Karar Ağacı Hata Matrisi…….	19
Çizelge 4.3.	Lojistik Regresyon Değişken Katsayıları…….	20
Çizelge 4.4.	Lojistik Regresyon Hata Matrisi…….	20
Çizelge 4.5.	DVM Hata Matrisi…….	21
Çizelge 4.6.	Özellik Seçimi…….	22
Çizelge 4.7.	Özellik Seçimi ile Karar Ağacı Hata Matrisi…….	23
Çizelge 4.8.	Özellik Seçimi ile AdaBoost Karar Ağacı Hata Matrisi…….	24
Çizelge 4.9.	Özellik Seçimi ile Lojistik Regresyon Değişken Katsayıları…….	25
Çizelge 4.10.	Özellik Seçimi ile Lojistik Regresyon Hata Matrisi…….	25
Çizelge 4.11.	Özellik Seçimi ile DVM Hata Matrisi…….	26
Çizelge 4.12.	Model Performansları…….	28

1. GİRİŞ

Günümüzde endüstriyel sistemlerde bilimsel yöntemler aktif bir şekilde kullanılmaktadır. Kullanılan bu yöntemlerin çeşitliliği sektörlerdeki rekabet koşulları ile her geçen gün artmaktadır. Teknolojiyle birlikte yeni yöntemlerin oluşturulması, geliştirilmesi ve uygulanmasıyla beraber bilgi ve bilginin yönetimi de aynı derecede önem taşımaktadır. Endüstriyel sistemler temel donanımın yanı sıra bilginin işlendiği ve depolandığı süreçleri de içermektedir. Bu süreçler girdiler, parametreler ve çıktılardan oluşan ve içinde büyük veri yapılarını barındıran sistemlerden meydana gelmektedir. Braha'ya (2013) göre süreçlerden elde edilen büyük ölçekli verilerin doğru bir şekilde modellenmesi için tüm parametrelerin aynı anda dikkate alınması gerekmektedir. Derlenen büyük miktarlı verilerden sistematik olmayan analizler ile yararlı bilgi elde etmek mümkün değildir. Bu nedenle var olan verinin faydalı bilgiye dönüşümü için en uygun teknik ve araçlar kullanılmalıdır.

Verinin sağlıklı bir şekilde kayıt altına alınması ve saklanması için dijital ortamda sistemlerini yöneten üretim tesisleri, süreçlerinden elektronik olarak veri toplamak ve aktarmak için güçlü veri toplama sistemlerini kullanmaktadır. Toplanan ölçüm değerleri, kuruluşların veri tabanlarında saklanmaktadır (Doğan ve Birant, 2021). Geçmiş tarihlerde yapılan bir çalışmaya göre dünya'da bulunan veri miktarının her yirmi ayda bir iki katına çıktığı tahmin edilmiştir (Piatetsky-Shapiro ve ark., 1992). Bu oran günümüzde giderek artmaktadır. Teknolojik yöntemlerin kullanılmasıyla beraber, veriler birleşerek büyük veri yığınlarını oluşturmaktadır. Elde edilen bu verilerin geleneksel yöntemler ile işlenmesi de oldukça güç bir durum haline gelmektedir. Bunun yanı sıra verilerin güvenilirliğinin ölçülmesi karşılaşılan problemlerden biridir. Teknolojinin etkin, ucuz, güncel ve ulaşılabilir hale gelmesi, bilimsel hesap ve modellemelerin geliştirilmesi ile birlikte veri madenciliği yöntemlerinin tercih edilme sıklığı artmıştır.

Veri madenciliği, büyük ölçekli veriler arasından anlamlı ve faydalı bilgilere ulaşmayı sağlayan, kullanılan verilerin uygun algoritmalar ile işlenip veriler arasındaki ilişkilerin tespit edilmesinden, veri eğilimlerinin tahminine kadar birçok yönetimi içerisinde barındıran bilginin açığa çıkarılması işlemidir.

Temelde tahmin edici (denetimli) ve tanımlayıcı (denetimsiz) olmak üzere iki ana başlık altında incelenmektedir. Süreçlerde ölçülen değerlerin çeşitliliğine bağlı olarak veri setlerinin analizi için, sınıflama yöntemleri, tahminleme, birliktelik kuralları, kümeleme ve aykırı değer analizlerinden uygun olan bir veya daha fazla yöntem kullanılabilmektedir.

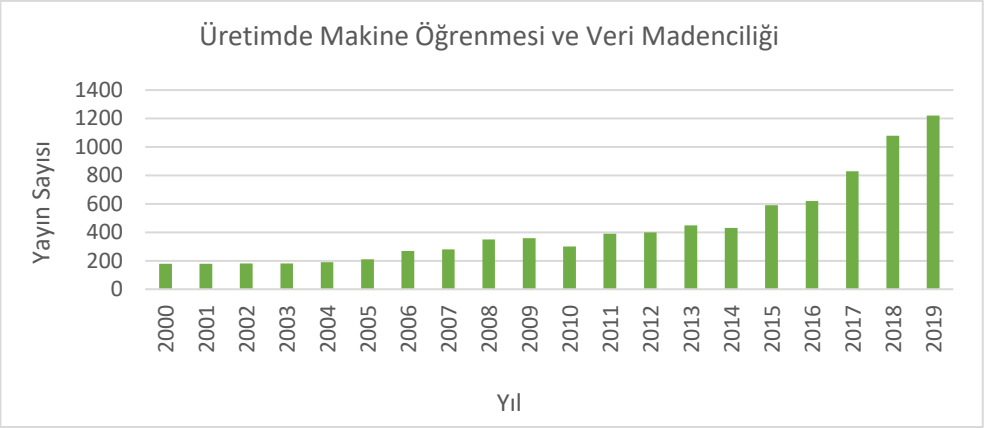
Veri madenciliği alanında yayınlanan makale sayısı her geçen gün artmakla beraber kullanım alanları da farklılık göstermektedir. Hizmet sektörünün bankacılık, pazarlama, telekomünikasyon gibi alanlarında kullanılırken, bilimsel çalışmalarda hastalıkların oluşum risklerinin tahmin edilmesi, ilaç yan etkilerinin tanımlanması vb. farklı alanlarda kullanılmaktadır. Üretim sektöründe ise sağlıklı veri elde edilebilen tüm süreçlerde kullanılabilmektedir. Bu tez çalışmasında da veri madenciliği yöntemleri ile bir üretim sisteminin verileri kullanılarak bir süreç iyileştirme projesi için temel hazırlanması ve literatüre bu alanda katkı sağlanması amaçlanmıştır.

Bu çalışmada, bir otomotiv yan sanayi firmasının kaynak bölümünde, kaynak işleminin yapıldığı makinadan alınan veriler ile kullanılan parametrelerin en uygun düzeylerinin belirlenmesi ve oluşabilecek ıskarta parça sayısının en düşük düzeye indirgenmesi amaçlanmaktadır. Proje kapsamında toplanan veri seti ilk olarak veri önişleme adımlarına tabi tutulmuştur. Çalışmanın kapsamına uygun olarak belirlenen veri madenciliği yöntemlerinden gözetimli öğrenme algoritmaları veri setine uygulanmıştır. En yüksek performansa sahip yöntem ile problem çözümünün elde edilmesi ve mevcut sürecin iyileştirilmesi hedeflenmiştir.

2. KAYNAK ARAŞTIRMASI

Veri madenciliği, büyük ölçekli veriler arasından daha anlamlı veriler elde edilmesini sağlayan, verinin açığa çıkarılması işlemidir. Diğer bir tanımlamaya göre veri madenciliği büyük hacimli veriler arasından örüntüler ve ilişkiler bulma sürecidir (Kerber ve ark., 1995). 1990'lü yıllarda veri madenciliği, büyük hacimli verilerin içerisindeki faydalı bilginin ortaya çıkarılması ile başlamıştır (Fayyad ve Stolorz, 1997). Kantardzic'e göre (2011) veri madenciliği büyük veri setlerinden faydalı veriyi aramanın yanı sıra öncesinde belirlenmemiş kavramların, yinelemeli keşifsel analiz senaryolarından oluşan bir süreçtir. Günümüzde ise veri anlam içeren her türlü sözcük ve sayılardan oluşmakla birlikte, bilginin verilerin bilgisayar ortamlarında işlenebilir hale dönüştürülüp veri madenciliği yöntemlerinin uygulanmasıyla elde edildiği görüşü yaygındır. (Sağın, 2018).

Faydalı bilgiye duyulan ihtiyacın artmasıyla birlikte veri madenciliği yöntemlerinin kullanım oranları da her geçen gün artmaktadır. Günümüzde veri madenciliği farklı metotlarla desteklenmekle beraber veri madenciliği tekniklerinin üretim alanlarında rekabet açısından önemli bir avantaj sağladığı ve yeni bakış açıları getirdiği için yaygın bir şekilde kabul gördüğü belirlenmiştir (Doğan ve Birant, 2021). Üretim sektörlerinde ürün tasarımlarından kaliteye, satın almadan satışa, planlamadan tedarikçi seçimlerine kadar sağlıklı verinin elde edilebildiği tüm alanlarda veri madenciliği uygulamaları kullanılmakta ve yaygınlaşmaktadır. Şekil 1.1'de görüldüğü gibi, yapılan literatür araştırmaları sonucunda 2000'li yılların başından 2019 yılına kadar üretim alanındaki veri madenciliği ve makine öğrenmesi çalışmalarında son yıllardaki artış oranı göze çarpmaktadır. Bu artış nedenlerinin en başında veri madenciliği yöntemlerinin etkin, ucuz ve ulaşılabilir olmasının yanı sıra hizmet ve servis sektöründen üretim sektörüne kadar geniş bir alanda tercih edilmesi de etkili olmuştur.



Şekil 2.1. Yıl Bazında Üretimde Makine Öğrenmesi ve Veri Madenciliği Kullanımı (Doğan ve Birant, 2021)

2.1. Veri Madenciliğinin Gelişimi

Veri kelimesi (verum) Latince'de "gerçek" anlamına gelen temelde varlığı bilinen ham haldeki kayıtların geneline verilen addır. Bununla birlikte, işlenmiş ve farklı boyut kazanmış kayıtlar da farklı çalışmalarda bir girdi olabileceği için veri olarak kabul edilmektedir.

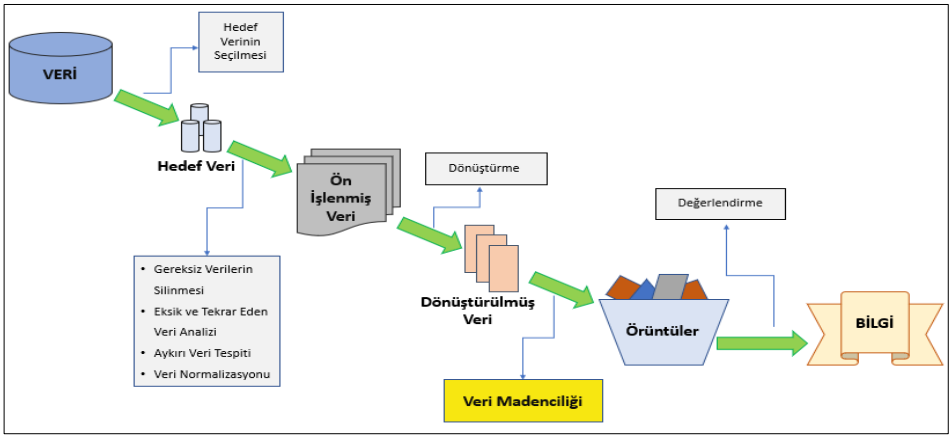
İstatistiksel olarak veriler dört ana başlık altında toplanmaktadır. Bunlar; nümerik, nominal, sıralı ve oransal verilerdir. Nümerik veriler sayısal, sürekli ya da kesikli değerlerden oluşurken, nominal veriler daha çok kategori belirten değerlerden oluşmaktadır. Sıralı veriler nümerik ve kategorik verilerin karışımı olmakla beraber veriler arasında sıralı bir ilişki bulunmaktadır. Son veri tipi olan oransal veriler aynı özniteliğin farklı birimler ile ifade edilmesi sonucu ortaya çıkmaktadır.

Çizelge 2.1'de veri tipleri ve bunlara örnekler verilmiştir.

Çizelge 2.1. Veri Tipleri

Veri Tipleri	Örnekler
Nümerik Veriler	Yaş, sıcaklık, ailedeki çocuk sayısı, akım,vb.
Nominal Veriler	Cinsiyet, medeni durum, göz rengi vb.
Sıralı Veriler	Başarı puanı, bir ürüne karşı tutum vb.
Oransal Veriler	Satış miktarı, yıllık kazanç vb.

Veri madenciliğinin uygulanacağı veriler için öncelikle çalışmanın amacı ve kapsamı belirlendikten sonra faydalı bilgiye ulaşmak amacıyla veri madenciliği yöntemleri sırasıyla takip edilmelidir. Fayyad ve Stolorz (1997) tarafından yayınlanan makalede sadece veri madenciliği adımlarının kullanılması anlamsız kalıpların keşfedilmesine yol açabileceği ifade edilmekte ve verilerden faydalı bilgilerin elde edilebilmesi için verinin hazırlanması, seçilmesi, veri ön işleme, veri madenciliği ve veri madenciliği sonucunda ortaya çıkan örüntülerin değerlendirilmesi ve yorumlanması adımlarının izlenilmesi önerilmektedir. Şekil 2.2'de görüldüğü üzere ham verinin bilgiye dönüşme adımları veri tabanından bilgi keşfi (VTBK) olarak gösterilmiştir.



Şekil 2.2. VTBK Sürecinin Aşamaları (Fayyad ve Stolorz'dan (1996) uyarlanmıştır)

VTBK sürecinin ilk adımı veri seçimi olup burada dikkat edilmesi gereken, çalışmanın hedefine hizmet edebilecek doğru veri setinin seçilmesidir. Veri setinin seçilmesi ile başlayan süreç bir diğer adım olan veri ön işleme süreci ile devam etmektedir. Bu aşamada veri setinin hazırlanması, uygulanacak veri madenciliğinin kalite ve verimliliğinin artmasına yardımcı olacaktır. Veri ön işleme adımları; veri temizleme, birleştirme, dönüştürme ve indirgeme olmak üzere dört ana başlık altında toplanabilir.

Veri ön işleme aşamalarından ilki olan veri temizleme; gereksiz verilerin silinmesi, eksik ve tekrar eden verilerin analizi, aykırı verilerin tespiti gibi gürültü olarak ifade edilen bu olumsuzlukların veri setinden temizlenmesi işlemidir. Piramuthu'ya göre (2004) verilerden örüntülerin çıkarılabilmesi ve veri madenciliği yöntemlerinin uygulanabilmesi için tek olmasa da birincil bilgi kaynağı olan girdilerin yani verilerin temiz olması gerekmektedir. Veri madenciliği süreçlerinde kaynakların yaklaşık % 80'i veri ön işleme ve temizleme adımlarına harcandığı kabul edilmektedir.

Veri birleştirme farklı veri tabanları ve kaynaklarındaki verilerin birleştirilmesi işlemidir. Bu aşamada yalnızca verilerin birleştirilmesi değil indirgenmesi işlemi de gerçekleştirilebilir. Veri indirgeme ile fazla olan değişkenlerin veri setinden çıkarılması ve kümeleme yöntemleri kullanılarak uygun veri seti boyutları elde edilmesi amaçlanmaktadır (Oğuzlar, 2004).

Veri dönüştürme veri setindeki değişkenlerin, veri madenciliğine uygun formatlara dönüştürülmesi işlemidir. Normalizasyon işlemleri bu adımda uygulanan yöntemlerden biridir. Veri ön işleme sürecinin son adımı olan veri indirgeme, veri madenciliği yöntemleri ile ilk veri setinden elde edilecek sonuçların, verinin daha küçük hacimlere indirgenmesi ile elde edilmesinin sağlanmasıdır. Veri önişleme sürecinde hedef, veri madenciliği uygulamaları ile elde edilecek sonuçların güvenilirliğini ve kalitesini arttırmaktır.

2.2. Veri Madenciliği Modelleri ve Metotları

Veri madenciliği modelleri tahmin edici ve tanımlayıcı olmak üzere iki başlık altında incelenmektedir (Zhong ve ark., 1999). Tahmin edici modeller, veri setlerinde girdi (parametre) ve çıktıların bir arada bulunduğu temelde öngörülemeyen veriler için sonuçları tahmin edildiği bir veri madenciliği modelidir. Tanımlayıcı modellerde ise, amaç parametrelerin benzer özelliklerine göre bir araya getirilerek sınıflandırılmamış verilerin analiz edilmesidir (Savaş ve ark., 2012). Temelde ikiye ayrılan modeller, işlevlerine göre:

1. Sınıflama ve Regresyon
2. Kümeleme
3. Birliktelik Kuralları

olmak üzere üç başlık altında incelebilir.

Sınıflandırma, bir veri seti için verinin sahip olduğu mevcut sınıfıyla sahip olduğu nitelikler arasındaki ilişkilerin kurulması ve farklı etiketlere ait verilerin tahmin edilmesidir (Işık ve Ulusoy, 2021).

Regresyon değişkenler arasındaki ilişkilerin ölçülmesi ve analiz edilmesinde kullanılan yöntemlerden biridir. Bir girdi ve çıktı arasındaki ilişkinin incelenmesi basit regresyon olarak adlandırılırken, birden fazla değişkenin bulunduğu süreçlerin analizi çoklu regresyon olarak tanımlanmıştır.

Kümeleme yöntemi veri setinde bulunan verilerin sonlu sayıda kümeler ile benzer özelliklerine göre gruplanmasını ve tanımlanmasını sağlamaktadır (Çelik, 2009). Birliktelik kuralları veri setinde bulunan veriler arasındaki ilişkilerin bulunmasına yardımcı olan veri madenciliği yöntemlerinden biridir (Santoso, 2021).

2.3. Karar Ağaçları

Karar ağaçları bir kökten başlayıp her bir yaprak düğümünde mantıksal bir sonuç elde edene kadar devam eden ve bir veri ayırma dizisi olarak tanımlanan ağaç tabanlı bir yöntemdir. Düğüm ve bağlantılar arasındaki ilişkilerin sınıflandırılması sağlamaktadır (Charbuty ve Abdulazeez, 2021).

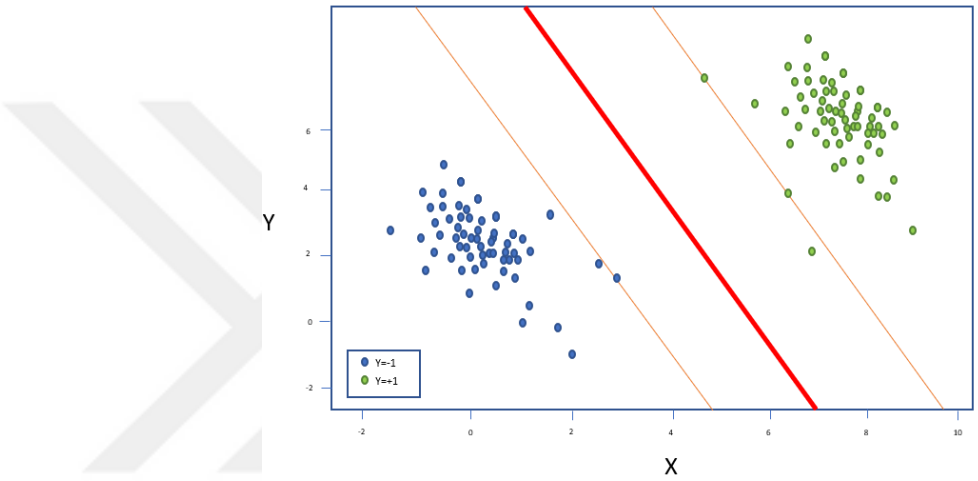
AdaBoost algoritması eğitim kümesinin önce bir zayıf öğrenici ile eğitilip sonrasında çıkan sonuç ile yanlış tahminlenen örneklerin verilerine daha fazla öncelik (ağırlık) verilerek tekrar eğitilmesidir. Diğer bir deyişle daha doğru bir sınıflandırma sonucu elde edene kadar farklı model çıktılarının derlenmesi ve sisteme tekrar verilmesini kapsamaktadır. (Azmi ve Baliga 2020). Genel olarak karar ağacı modelleri oluşturulurken entropi değerleri dikkate alınır. Entropi bir sistemin düzensizliğini ifade ederken, gini indeksi o sistemin safsızlığını (karışık, homojen olmayan) belirtmektedir. Entropi ve gini değerlerinin düşük, verilerden elde edilecek bilgi kazancının yüksek olması karar ağacı modeli ile yapılacak sınıflandırmanın başarısını olumlu yönde etkilemektedir.

2.4. Lojistik Regresyon

Sınıflandırma (classification) yöntemlerinden biri olan lojistik regresyon ikili sonuç veren değişkenleri yani bir sınıfı (kategoriyi) tahmin eden veri analizidir. Amaç bir veya daha fazla değişkenin bağımlı değişken ile olan ilişkisinin ölçülmesidir. Lojistik regresyon sonucunda her bir değişkenin regresyon denklemindeki katsayıları belirlenir.

2.5. Destek Vektör Makineleri

Destek vektör makineleri (DVM) temel olarak düzlem üzerindeki iki sınıftan oluşan noktaların bir doğru veya hiper düzlem ile ayrıştırılması ve sınıflandırılmasıdır. Diğer bir deyişle kategorisi bilinmeyen değişkenlerin öncesinde sınıflandırılmış verileri örnek alarak tahminlenmesi işlemidir. Şekil 3.6 destek vektör makineleri ile sınıflandırmaya bir örnek olarak verilebilir.



Şekil 2.3. DVM Örneği

2.6. Veri Madenciliği ile İlgili Çalışmalar

Kalite kontrol etkinliklerinde iyileştirmeler için altı sigma, Kaizen, PUKÖ, Pareto analizleri gibi metotlar kullanılmaktadır. Üretim sistemlerinde elde edilen kalite verileri çeşitli süreç analizlerinde kullanılmaktadır. Veri madenciliği uygulamalarının, üretim sistemlerinde gerçekleşen kalite kontrol faaliyetlerindeki kullanımları kapsamlı bir şekilde araştırılmıştır. İleriye yönelik yapılabilecek çalışmalar için öneriler ve çalışmanın zorluklarından bahsedilmiştir. (Köksal ve ark., 2011).

Ferreiro ve arkadaşları (2011) yaptıkları çalışmada delme işlemi sırasında oluşan çapakların kalite sorununa neden olup olmayacağının veri madenciliği yöntemleriyle tahmin etmeyi amaçlamıştır. Uygulamada belirlenen özelliklere sahip alüminyum malzeme üzerinde kuru koşullarda yapılan yüksek hızlı delme işlemi verileri kullanılmıştır. Veri kümesi modelin geliştirilmesi için eğitim, doğrulanması içinse test kümesine ayrılmıştır. Sınıf etiketi olarak kabul edilebilir bir çapak miktarı belirlenmiştir. Kullanılan veri madenciliği yöntemleri arasından ID3, Prism ve KNN yöntemlerinin en uygun sonuçları verdiği belirlenmiştir.

Santos ve arkadaşları (2013) lazerle parlatılan bileşenlerin yüzey pürüzlülüğünün tahmin edilmesi ve endüstri standartlarına en uygun olan sınırlandırıcı tabanlı yöntemin belirlenmesini amaçlamıştır. Parlatma işleminde parça pürüzlülüğünü etkileyen önemli noktalar lazer ışınının enerji yoğunluğu, yüzey olarak kullanılan malzeme ve başlangıçtaki parçanın pürüzlülüğü olarak belirtilmiştir. Veri seti 178 örnek içermekte olup her bir veri beş farklı pürüzlülük seviyesinden oluşmaktadır. Verilere veri madenciliği yöntemlerinden C4.5 Karar ağaçları, Torbalama, Boosting, Rotation Forest, Regresyon ve DVM (Destek Vektör Makinesi) metotları uygulanmıştır. Bu yöntemler arasından Rotation Forest metodunun en başarılı sonuç verdiği belirlenmiştir.

Wei ve arkadaşları (2020), veri madenciliğini temel alan veriye dayalı süreç tasarım yöntemlerini incelemiştir. Üretim verilerinin etkili bir rol oynaması için yeni bir veri-bilgi-karar verme modeli önerilmiştir. Çalışmada kullanılan veri seti bir kaynak sürecinin değerlerinden oluşmaktadır. Veri seti, yapılandırılmış ve yapılandırılmamış veriler içermektedir. Yapılandırılmış veriler, zaman serisi verilerinden oluşurken yapılandırılmamış veriler görüntü kayıtlarından oluşmaktadır. Yapılandırılmış veriler için uygun ön işleme adımları gerçekleştirilmiştir. Yapılandırılmamış veriler ise uygun ölçüm, teknikler ve yapay sinir ağı modelleri ile yapılandırılmış verilere dönüştürülmüştür. Uygun hale getirilmiş veriler üzerinde veri madenciliği metotlarından korelasyon, regresyon ve kümeleme algoritmalarından K-Ortalamalar, EM (Expectation Maximization) ve Bayes uygulanmıştır.

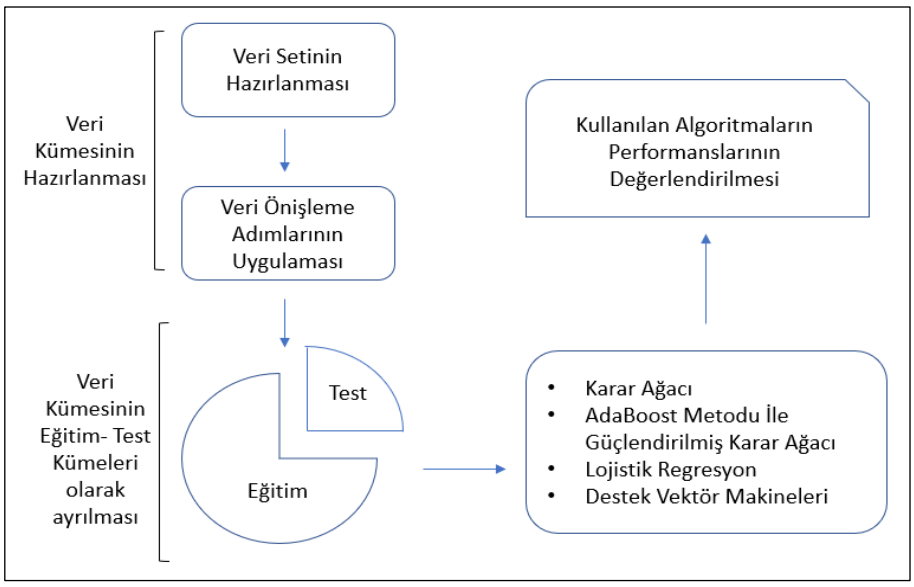
Canlı ve Toklu (2019) yaptıkları çalışmada bir üretim firmasında montaj hattında bulunan tezgâhtan alınan kamyon ön kapı menteşe verilerini kullanmıştır. Veri seti 17046 örnek ve 19 değişken içermektedir. Bu veriler arasından işlemin kritikliği ile ilişkili değişkenler belirlenerek bir veri seti oluşturulmuştur. Bu değerlere veri ön işleme ve veri görselleştirme adımları uygulanmıştır. Veriler farklı oranlarda test ve eğitim kümelerine bölündükten sonra sınıflandırma algoritmaları kullanılarak kalite kontrol sürecine yönelik bir karar destek sistemi geliştirilmiştir. Toplanan veriler üzerinde C4.5, Naive Bayes, SMO ve Random Forest algoritmalarının performansları karşılaştırılmıştır. Bu algoritmalar ile üretim sırasında parçanın kalite uygunluk düzeyinin tahmin edilmesi ve işlemler tamamlanmadan önce parçanın arızalı olarak sınıflanması halinde oluşabilecek ek maliyetlerin önlenmesi amaçlanmıştır. Uygulanan yöntemler arasında en uygun sınıflandırma metodu C4.5 karar ağaçları olarak belirlenmiştir.

Doğan ve Birant (2021), yaptıkları çalışma ile üretim süreçlerinde veri madenciliği ve makine öğrenimi tekniklerinin uygulanabilirliğine yönelik güncel ve kapsamlı bir literatür taraması sunmuştur. Bu tekniklerin literatürdeki kullanım oranlarının yirmi yıllık dönem içindeki artışı sayısal olarak ifade edilmiştir. Literatür taraması dört ana başlıkta ele alınmıştır. Bunlar çizelgeleme, kalite, izleme ve arızadır. Ana başlıkların içerisinde kümeleme, sınıflandırma, regresyon vb. algoritmalar, destek vektör makineleri, sinir ağları, öğrenme türleri, derin öğrenme, toplu öğrenme, performans dahil olmak üzere mevcut çözümleri tartışmaktadır. Ele alınan yöntemlerin üretim süreçlerine sağlayacağı avantajlar, yaşanabilecek zorluklara karşı çözüm yöntemleri ve ileri araştırmalar için izlenebilecek yollar belirtilmiştir.

3. MATERYAL VE YÖNTEM

Bu çalışmada, üretim sektöründe hizmet veren bir firmadan alınan gaz altı kaynak makinesi verileri kullanılmıştır. Veri seti kaynak sonucunda oluşan parçanın kalitesini içeren bir çıktıdan ve kaynak sürecini etkileyen 15 adet değişkenden oluşmaktadır. Toplamda 600 adet veri içermektedir. Kaynak işlemi sonucunda oluşan parçanın kalite uygunluğu veri setinde Uygun/Uygun Değil (OK/NOK) olarak belirtilmiştir. Çalışmada makine öğrenmesi, veri analizi ve eğitim için uygun olan ve bulut tabanlı Python kodlarının çalıştırılabildiği bir uygulama kullanılmıştır. Firma gizliliği sebebiyle çalışmada kullanılan veriler paylaşılmamıştır.

İlk olarak veri setinin hazırlığı ile başlayan süreç veri önişleme adımlarının uygulanması ve veri setinin eğitim/test kümesi olarak ayrılmasıyla, veri madenciliği modellerinin uygulanabilirliği için hazır hale getirilmiştir. Belirlenen modellerin uygulanması ve sonuçların değerlendirilmesi ile çalışmanın tamamlanması amaçlanmıştır. Şekil 3.1'de yapılan çalışmanın akışı kısaca özetlenmiştir.



Şekil 3.1. Süreç Özeti

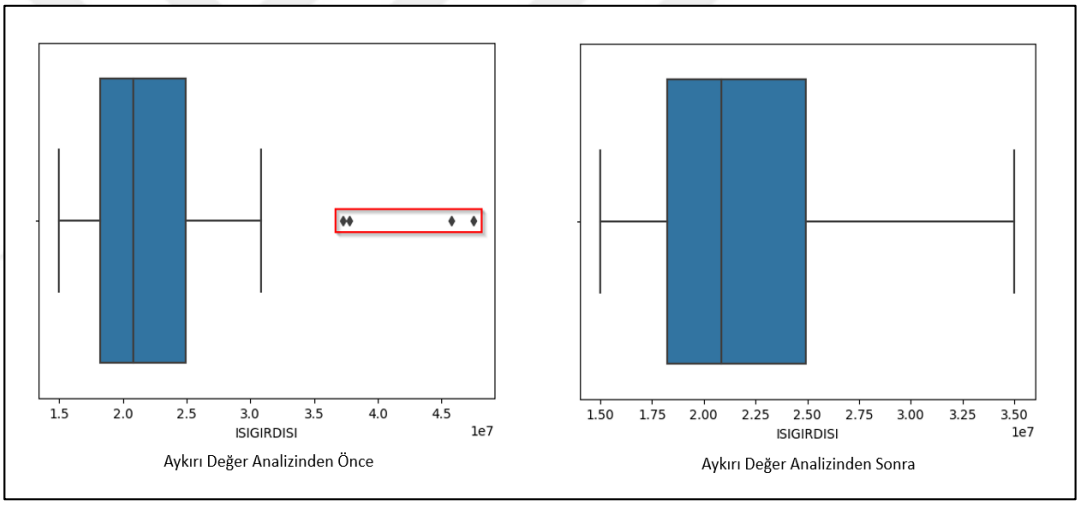
3.1 Veri Ön İşleme

Veri setine veri önişleme adımlarından; gereksiz verilerinin silinmesi, tekrar eden verilerin analizi, eksik verilerin tespiti, normalizasyon, aykırı değer analizi ve veri tiplerinin değiştirilmesi işlemleri uygulanmıştır. Veri madenciliği model performanslarını etkilemeyecek, tüm değerleri aynı ya da farklı olan kategorik değişkenler belirlenmiş ve veri kümesindeki 14 değişken, gereksiz verilerin silinmesi ile 10 adet niteliğe indirgenmiştir. Bu işlem sonrasında iş yeri, üretim ID, işlem sıra numarası, ekipman numarası ve kaynak numarası değişkenleri veri setinden çıkarılmıştır. Veri setinde eksik veri bulunmadığı tespit edilmiştir. İlk aşamadaki temizlikten sonra veri madenciliği modellerinin uygulanacağı veri kümeleri hazırlanmıştır. Çalışmanın ikinci adımında uygulanacak modeller için veri önişleme adımlarına ek olarak veri setine özellik seçimi uygulanmıştır. Özellik seçimi ile model performanslarındaki değişimin belirlenmesi hedeflenirken değişkenlerin hedef değişken ile arasındaki ilişkisinden faydalanılmıştır. Bu yönteme ek olarak değişkenlerin, çıktı (öznitelik) ile arasındaki korelasyon değerleri de göz önünde bulundurularak veri setinde kullanılacak nitelikler belirlenmiştir. İlk adımda kullanılacak değişkenler Çizelge 3.1'de açıklamalarıyla verilmiştir.

Çizelge 3.1. Değişkenler

Değişkenler	Açıklama
Akım	Amper cinsinden elektriksel yük
Gerilim	Voltaj, iki nokta arasındaki potansiyel fark
Tel Hızı	Zaman cinsinden tüketilen tel uzunluğu (m/min)
Gaz Kullanımı	Kaynak sürecinde kullanılan gaz miktarı (l/min)
Robot Hızı	Lineer ve dönel hareket hızları
Isı Girdisi	Kaynak ile parça arasında mm'e başına veridiği ısı (joule)
S1/S2	Kaynak parametreleri, boğaz kalınlığı
P1/P2	Penetrasyon, kaynak derinliği

Aykırı gözlemler veri setlerindeki genel dağılımın dışında yer alıp, veri setinden elde edilebilecek sonuçların yanlılığına ve istatistiksel analizlerin yanlışlığına sebep olabilmektedir. Belirlenen her bir değişken için aykırı değer analizi işlemi gerçekleştirilmiş olup yapılan analizler sonucunda veriler içerisinde aykırı değerlerin bulunduğu tespit edilmiştir. Aykırı değer analizinde kutu grafiğinden faydalanılmıştır. Şekil 3.2'de 'Isı Girdisi' niteliği için aykırı değer analizi öncesi ve sonrası görselleştirilmiştir. Aykırı veriler bulundukları çeyrekler dikkate alınarak minimum veya maksimum değer ataması metodu kullanılarak değiştirilmiştir. Aykırı değer analizinden sonra normal dağılımın dışında çıkan değerlerin giderildiği görülmüştür. Aykırı değer analiz veri madenciliği modellerinin uygulandığı programda gerçekleştirilmiştir.



Şekil 3.2. Aykırı Değer Analizi Kutu Grafiği

Aykırı değer analizi uygulanırken her bir niteliğin minimum, maksimum ve ortalama gibi istatistiki değerlerinin korunması dikkate alınmıştır.

Şekil 3.3'te parametrelerden biri olan ısı girdisi değişkeninin diğer istatistiki değerleri değişmezken, standart sapma değerinin azaldığı görülmektedir. Aykırı değer bulunan değişkenlerin istatistiki sonuçları EK 1'de verilmiştir.

```
Önceki İstatistikler: count    5.990000e+02
mean      2.166706e+07
std       4.107061e+06
min       1.501871e+07
25%       1.824346e+07
50%       2.083644e+07
75%       2.494845e+07
max       4.752634e+07
Name: ISIGIRDISI, dtype: float64
```

```
Sonraki İstatistikler: count    5.990000e+02
mean      2.161981e+07
std       3.889214e+06
min       1.501871e+07
25%       1.824346e+07
50%       2.083644e+07
75%       2.494845e+07
max       3.500594e+07
Name: ISIGIRDISI, dtype: float64
```

Şekil 3.3. Aykırı Değer Analizi İstatistiki Sonuçları

Veri setinde tekrar eden verilerin bulunmadığı belirlenmiştir. Uygulanan veri önişleme adımları ile belirlenen veri setindeki değişkenlerin farklı birimlerde değer alması sebebiyle veri kümesine Min-Max normalizasyon işlemi uygulanmıştır. Burada amaç veride bulunan en büyük değerin 1, en küçük değerin 0 olacak şekilde normalize işleminin gerçekleştirilmesidir. Şekil 3.4'te değişkenlerden birkaçının normalizasyondan önce ve sonra aldığı değerlerin bir kısmı gösterilmiştir.

	TELHIZI	GAZKULLANIM	ROBOTHIZI	ISIGIRDISI
0	5.039	15.440	0.007914	17921178.64
1	5.296	16.807	0.007000	20561597.20
2	4.626	17.788	0.006926	17974069.25
3	4.943	17.986	0.006822	21976053.65
4	6.011	17.918	0.007000	22185163.53

→

	TELHIZI	GAZKULLANIM	ROBOTHIZI	ISIGIRDISI
0	0.608079	0.488491	0.111457	0.089286
1	0.647256	0.738217	0.056746	0.170510
2	0.545122	0.917428	0.052317	0.090913
3	0.593445	0.953599	0.046091	0.214022
4	0.756250	0.941176	0.056746	0.220454

Şekil 3.4. Normalizasyon İşlemi

Veriler içerdikleri değerlere göre sınıflandırılmakta olup bu çalışmada ele alınan veri kümesindeki değişkenler farklı veri tiplerine sahiptir. Ele alınan veri setindeki değişkenler sayısal ve kategorik değerler içermektedir. Sayısal değer içeren veriler uygulamada kullanılacak programa reel sayı anlamına gelen 'Float' tipinde verilmiştir. Çıktı değeri ise kategorik veri tipine sahip olup diğer verilerle arasındaki örüntü ve ilişkilerin analizini gerçekleştirebilmek adına 0 ve 1 değerlerini alacak şekilde değiştirilmiştir. Veri setinde 'Uygunluk' niteliğindeki Uygun Değil (NOK) değerine sahip veriler '0' iken Uygun (OK) değerine sahip veriler '1' olacak şekilde düzenlenmiştir.

Veri ön işleme adımlarının tamamlanmasının ardından veriler kullanılacak algoritmalar için eğitim ve test olmak üzere iki gruba ayrılmıştır. Eğitim için ayrılan veri seti ile uygulanan modeller eğitilirken, test veri seti ile modellerin performanslarının ölçülmesi amaçlanmaktadır. Eğitim ve test veri setleri % 70 eğitim-% 30 test kümesi olacak şekilde ayrılmıştır. Eğitim ve test verilerine uygulanan modellerin performansları değerlendirilmiş ve başarı oranları karşılaştırılmıştır.

Veri seti uygulanacak modeller için eğitim ve test kümelerine ayrıldıktan sonra belirlenen sınıflandırma yöntemleri ile farklı sınıflara ait verilerin tahmin edilmesi amaçlanmıştır. Bu çalışmada kullanılan sınıflandırma algoritmaları Karar Ağaçları, AdaBoost ile güçlendirilmiş Karar Ağaçları, Lojistik Regresyon ve Destek Vektör Makineleridir.

Çalışmada veri madenciliği modelleri iki aşamada tamamlanmıştır. Sürecin ilk adımında özellik seçimi yöntemleri uygulanmadan 12 değişken kullanılarak karar ağacı tabanlı modeller, lojistik regresyon ve destek vektör makineleri yöntemleri uygulanmıştır.

Çalışmanın ikinci adımda ise özellik seçiminin ardından sırasıyla modeller uygulanmış ve performansları her iki veri seti içinde tablolar halinde verilmiştir.

Veri madenciliği algoritmalarından veri setine ilk olarak karar ağacı modeli uygulanmıştır. Karar ağacı oluşturulurken bilgi kazancı ve entropi değerleri dikkate alınmıştır. Her bir model, iki durum için de ayrıca AdaBoost algoritmasıyla desteklenmiştir.

Diğer bir sınıflandırma modeli olan lojistik regresyon modeli veri setine uygulanmıştır. Tahmin edilen sınıf etiketinin 0-1 değer alması sebebiyle bu regresyon yöntemi tercih edilmiştir. Her bir bağımlı değişken ve bağımsız değişken için regresyon katsayıları belirlenmiş ve regresyon denklemi oluşturulmuştur. Oluşturulan modelin tahmin sonuçları hata matrislerinde verilmiştir.

Son veri madenciliği modeli DVM veri setine uygulanmıştır. Model kaliteli ve kalitesiz parça sınıflarını bir düzlem ile birbirinden ayırarak, tahmin edilecek test verilerinin bu düzlemler arasındaki dağılımını göstermektedir. Her bir veri için gerçekte aldığı değer ile tahmin edildiği değer model üzerinde gösterilmekte olup hata matrisleri oluşturulmuştur. Tüm model sonuçları çalışmanın birinci ve ikinci adımında performans değerleri birlikte verilmiştir.

4. BULGULAR

Veri ön işleme adımları ile hazır hale getirilmiş veriye ilk olarak temelde entropi değerini baz alan karar ağacı fonksiyonu uygulanmıştır. Entropi ile veri setindeki belirsizliklerin seviyesi tespit edilmiştir. Entropi değerlerinden faydalanan bilgi kazancı yöntemi ile veri kümesinin en iyi şekilde bölünmesi veya sınıflandırması amaçlanmıştır. Karar ağacı oluşturulurken bilgi kazancı yüksek olan değişkenin seçilmesi veri kümesinin en iyi şekilde bölünmesine veya sınıflandırılmasına yardımcı olmuştur. Bilgi kazancı ile belirlenen değişken P1 olup bu değişken üzerinden karar ağacı modeli veri setine uygulanmıştır.

Model sonucunda hata matrisi sonuçları Çizelge 4.1'de verilmiştir. Hata matrisi verilerin gerçekte aldığı değerleri ve model ile tahmin ettiği sonuçları içermektedir. Modelin performansı EK 2'de hesaplanmıştır.

Çizelge 4.1. Karar Ağacı Hata Matrisi

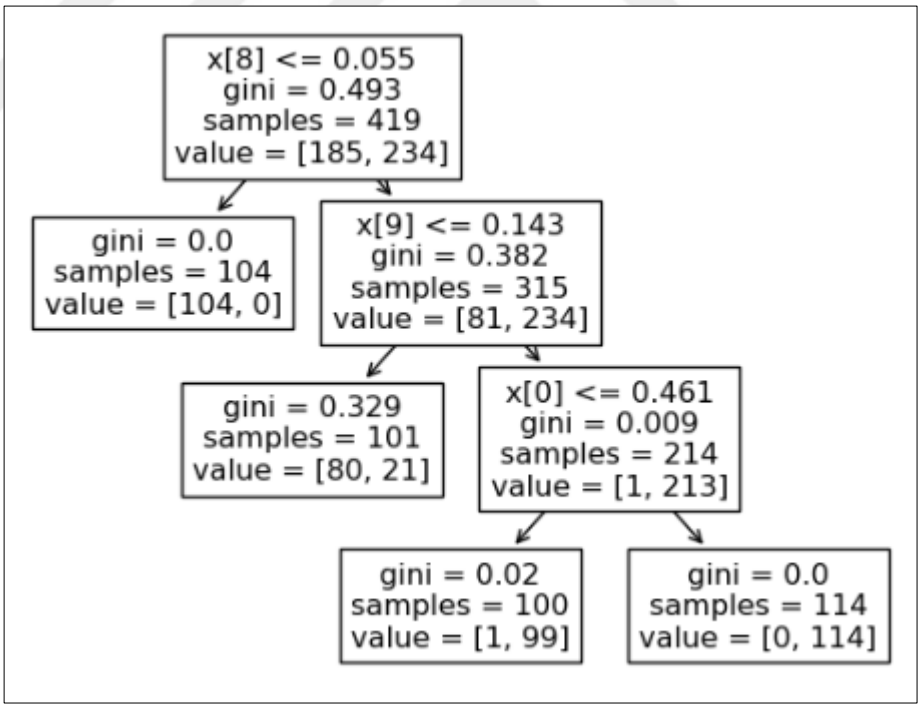
Tahmin / Gerçek	0	1
0	60	25
1	19	76

Uygulanan karar ağacı modeli AdaBoost algoritması ile güçlendirilip veri setine tekrar uygulanmıştır. Bu metot ile yanlış tahminlenen örneklere daha fazla ağırlık verilerek modelin performansının arttırılması amaçlanmıştır. AdaBoost algoritması kullanılarak uygulamadaki veri sayısı, modelin başarısı dikkate alınarak belirlenmiştir.

Model sonucunda hata matrisi sonuçları Çizelge 4.2'de verilmiş olup modelin başarısı % 95,5 olarak hesaplanmıştır. Model sonucunda oluşturulan karar ağacı Şekil 4.1'de gösterilmiştir. Uygulanan karar ağacı modelinde ilk düğümde P1 değişkeninde 0,055 değerinden küçük veya eşit olan veriler için test verileri iki kısma ayrılmıştır. Karar ağacı P2 değişkenini ve sonrasında akım değerlerini temel alarak ağaç dallanmasını gerçekleştirmiştir.

Çizelge 4.2. AdaBoost Karar Ağacı Hata Matrisi

Tahmin / Gerçek	0	1
0	47	0
1	4	39



Şekil 4.1. Karar Ağacı Modeli

Uygulanan diğer bir model Lojistik Regresyon'dur. Lojistik regresyon ile veri setinde bulunan değişkenlerin sınıf etiketine olan etkileri hesaplanmıştır. Her bir değişkenin regresyon denkleminde aldığı değerler Çizelge 4.3'te verilmiştir.

Çizelge 4.3. Lojistik Regresyon Değişken Katsayıları

Lojistik Regresyon Değerleri	**AKIM**	**GERİLİM**	**TELHIZI**	**GAZKULLANIM**	**ROBOTHIZI**
	0.82643057	0.69161595	-0.85322701	0.50457462	-0.06758785
UYGUNLUK OK	**ISIGİRDİSİ**	**S1**	**S2**	**P1**	**P2**
0.27065789	-0.39042416	-0.28839511	1.08567304	1.43073745	1.57529663

Model sonucunda hata matrisi sonuçları Çizelge 4.4'te verilmiştir. Modelin performansı EK 2'de hesaplanmıştır.

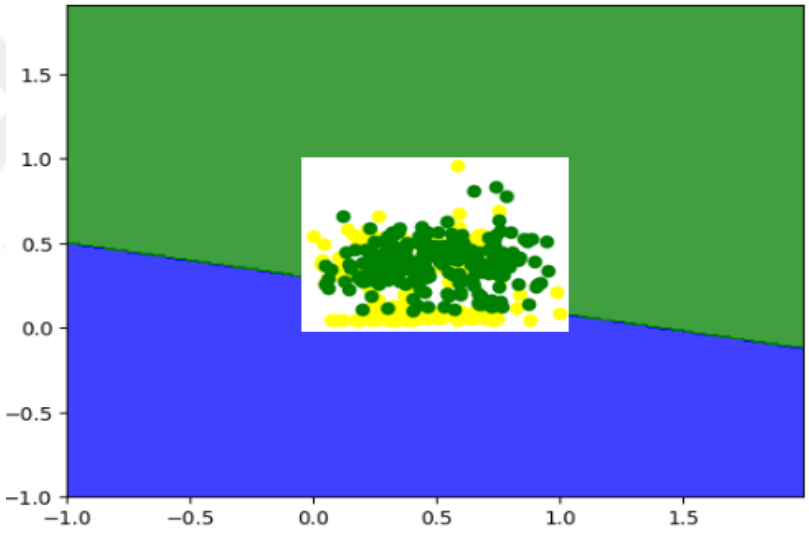
Çizelge 4.4. Lojistik Regresyon Hata Matrisi

Tahmin / **Gerçek**	0	1
0	66	16
1	12	86

Çalışmanın ilk adımında son olarak DVM modeli uygulanmış ve değişkenlerin sınıf etiketinin tahminlenmesi hedeflenmiştir. Model sonucunda hata matrisi sonuçları Çizelge 4.5'te verilmiştir. Modelin performansı EK 2'de hesaplanmıştır. DVM modeli ile tahminlenen test verilerinin grafik üzerinde dağılımı Şekil 4.2' de gösterilmiştir.

Çizelge 4.5. DVM Hata Matrisi

Tahmin / Gerçek	0	1
0	48	37
1	12	83



Şekil 4.2 DVM Modeli

Düzlem yeşil ve mavi renk olmak üzere ikiye ayrılmıştır. Yeşil renkli bölge içinde kalan alan parçaların kaliteli yani değerinin bire eşit, mavi renkli bölgenin içinde kalan alan ise parçaların kalitesiz yani sıfıra eşit olduğunu göstermektedir.

Modelin uygulanması ile incelenen veriler noktalar halinde gösterilmiş olup sarı noktaların kalitesiz parça, yeşil noktaların kaliteli parça olduğu tahmin edilmektedir.

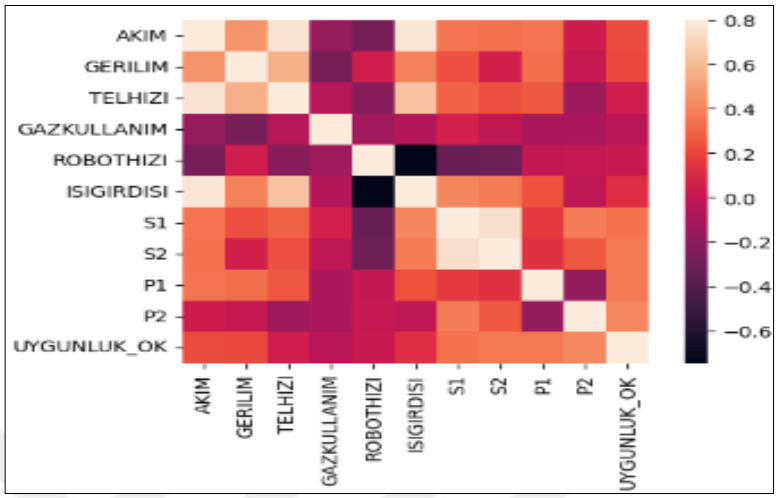
Grafik üzerinde de hata matrisi ile elde edilen sonuçlar ve verilerin dağılımları gösterilmektedir.

Çalışmanın ikinci adımında özellik seçimi yöntemleri kullanılarak değişkenler ile sınıf etiketi arasındaki ilişkiler belirlenmiş olup süreç üzerinde yüksek etki düzeyine sahip özellikler ortaya çıkarılmıştır. Çizelge 4.6'da özellik seçimi yöntemleri kullanılarak seçilen değişkenler görülmektedir. Özellik seçimi yapılırken SelectKBest fonksiyonundan faydalanılmıştır. Fonksiyona ek olarak değişkenlerin sınıf etiketiyle arasındaki ilişkiye göre sıralanması veri önişleme adımında uygulanan korelasyon yöntemi ile sağlanmıştır.

Çizelge 4.6. Özellik Seçimi

	KBestSelect	**Korelasyon**
AKIM	8	9
GERILIM	7	10
TEL HIZI	5	8
GAZ KULLANIM	10	6
ROBOT HIZI	9	7
ISI GIRDISI	6	5
S1	3	3
S2	4	4
P1	1	2
P2	2	1

Şekil 4.3'te korelasyon ile her bir değişkenin diğer değişkenlerle olan ilişkisi gösterilmiştir.



Şekil 4.3. Korelasyon Matrisi

Özellik seçimi ve korelasyon matrisi yöntemi ile Tel hızı, S1, S2, P1 ve P2 değişkenleri belirlenmiştir. Çalışmanın ilk adımında uygulanan modeller belirlenen parametreler için de sırasıyla uygulanmıştır. İlk olarak karar ağacı modeli uygulanmıştır. Model sonucunda hata matrisi sonuçları Çizelge 4.7'de verilmiştir. Modelin performansı EK 2'de hesaplanmıştır.

Çizelge 4.7. Özellik Seçimi ile Karar Ağacı Hata Matrisi

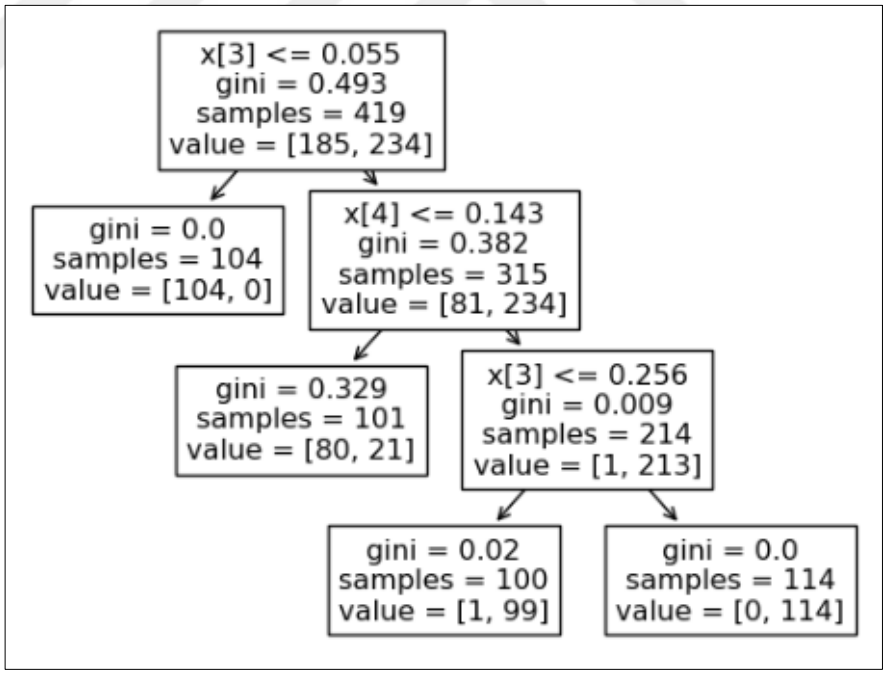
Tahmin / Gerçek	0	1
0	64	21
1	12	83

Uygulanan karar ağacı modeli AdaBoost algoritması ile güçlendirilip veri setine tekrar uygulanmıştır. Model sonucu Çizelge 4.8'de verilmiştir. Modelin performansı EK 2'de hesaplanmıştır.

Çizelge 4.8. Özellik Seçimi ile AdaBoost Karar Ağacı Hata Matrisi

Tahmin / Gerçek	0	1
0	47	0
1	4	39

Oluşturulan karar ağacı modeli Şekil 4.4'te verilmiştir. Kök düğüm olarak modelin P1 değişkenini belirlediği görülmektedir



Şekil 4.4. Özellik Seçimi ile Karar Ağacı Modeli

P1 değişkeninin 0,055'ten küçük veya eşit olması durumuna göre ağaç yapısı dallanmaktadır. Dallardan ilki 104 örneklem içermekte olup, tümü kalitesiz parça olarak tahminlenmiştir. Diğer dalda P2 değişkeninin 0,143'ten küçük veya eşit olduğu durumlar için ağaç yapısı dallanmaya devam etmiştir. P2 değişkenine göre ilk dallanmada 101 örneklem içerisinden 80 adet verinin kalitesiz parça olarak doğru tahmin edildiği tespit edilmiştir. Diğer dalda ise P1 değişkeninin 0,256'dan küçük veya eşit olduğu durumlar için veriler kaliteli ve kalitesiz parça tahminleri olmak üzere iki dala ayrılmıştır.

Karar ağacı modellerinden sonra veri setine lojistik regresyon modeli uygulanmıştır. Lojistik regresyon ile seçilen her bir değişkenin regresyon katsayısı hesaplanmış Çizelge 4.9'da belirtilmiştir. Model sonucunda hata matrisi ise Çizelge 4.10'da verilmiş olup lojistik regresyon modeli ile tahminlenmesi yapılan verilerin aldıkları değerler gösterilmiştir. Modelin performansı EK 2'de hesaplanmıştır.

Çizelge 4.9. Özellik Seçimi ile Lojistik Regresyon Değişken Katsayıları

UYGUNLUK_OK	TELHIZI	S1	S2	P1	P2
0.11483743	-0.30888678	0.62189208	0.5201934	1.55936622	1.5600132

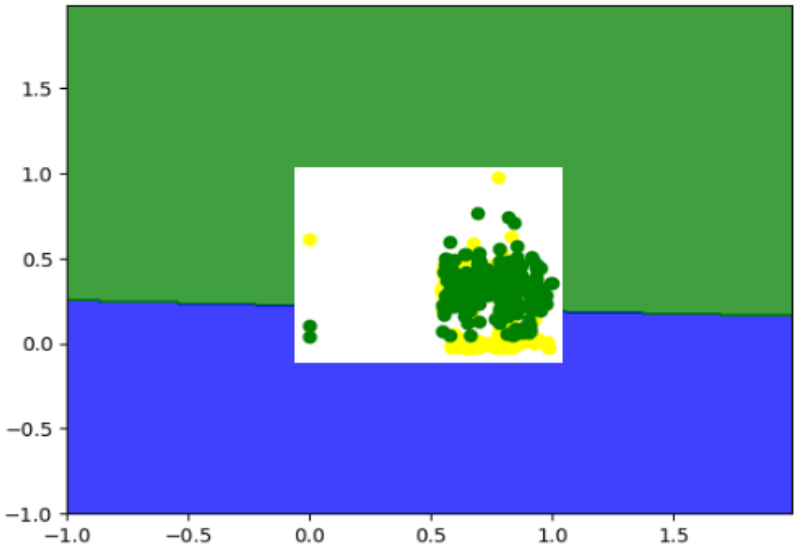
Çizelge 4.10. Özellik Seçimi ile Lojistik Regresyon Hata Matrisi

Tahmin / Gerçek	0	1
0	59	23
1	15	83

Çalışmanın son adımında DVM modeli uygulanmış ve değişkenlerin sınıf etiketinin tahmin edilmesi hedeflenmiştir. Model sonucunda hata matrisi sonuçları Çizelge 4.11'de verilmiştir. Modelin performansı EK 2'de hesaplanmıştır. DVM modeli ile tahmin edilen test verilerinin grafik üzerinde dağılımı Şekil 4.5'te gösterilmiştir.

Çizelge 4.11. Özellik Seçimi ile DVM Hata Matrisi

Gerçek \ Tahmin	0	1
0	48	37
1	9	86



Şekil 4.5. Özellik Seçimi ile DVM Modeli

Özellik seçiminden önce uygulanan modelin sonucundan farklı olarak, kaliteli parça tahmini performansındaki artış yeşil noktaların mavi düzlem üzerindeki azalışından da gözlemlenmektedir.

Veri setinde özelik seçimi yapılmadan önce uygulanan karar ağacı modellinde, kalitesiz olan parçaların tahmininin kaliteli parçaların tahmininden daha başarılı olduğu gözlemlenmiştir. Veri seçimi yapıldıktan sonra uygulanan karar ağacı modelinde ise kaliteli ve kalitesiz parçaların tahmininde iyileşmeler gözlemlenmiştir. Modelin genelinde ise özellik seçimi ile % 6,2 düzeyinde iyileşme sağlanmıştır.

Uygulanan karar ağacı modelinin performansının arttırılması amacıyla model AdaBoost algoritması ile desteklenmiştir. Karar ağacı modelinin AdaBoost algoritmasıyla desteklenmesi sonucunda model performansında artış gözlemlenmiştir. Özellik seçimi olmadan ve seçim yapıldıktan sonraki performanslar değerlendirildiğinde ise sonuçların değişmediği belirlenmiştir. Karar ağacının en üstünde bulunan başlangıç düğümü her iki ağaç modelinde de P1 değişkenidir. Bu durumun sebebi karar ağacı oluşturulurken entropi ve bilgi kazancının temel alınmasıdır. Her iki ağaç modeli için de sınıflandırma sonuçları çalışmada verilmiştir.

Öznitelik seçiminden önce uygulanan lojistik regresyon modeli ile her bir değişken için regresyon katsayıları belirlenmiştir. Model sonucunda tel hızı, robot hızı, ısı girdisi ve S1 değişkeninin parçanın kalitesi yani sınıf etiketi ile negatif bir ilişki içerisinde olduğunu göstermiştir. Kalan değişkenler ise parça kalitesi ile pozitif bir ilişki içerisinde olduğu tespit edilmiştir. Lojistik regresyon modelinde kalite açısından uygun parçanın tahminlenmesinin uygun olmayan parçanın tahminlenmesinden % 0,31 daha başarılı olduğu tespit edilmiştir. Özellik seçiminin ardından modelin genel başarısında % 5,6 oranında düşüş görülmüştür. Değişkenlerden tel hızı sınıf etiketi ile negatif yöndeki ilişkiyi korurken, S1 değişkeni pozitif yönde bir eğilim sağlamıştır.

Veri setine son olarak DVM modeli uygulanmıştır. İlk aşamada diğer modellerin genelinde de gözlemlendiği gibi kalite açısından uygun olmayan parçaların tahmini, kaliteli parçaların tahmin edilmesinden daha başarılı olduğu belirlenmiştir. Uygulanan tüm modeller içerisinde kaliteli parçanın tahminlenme başarısı en düşük olan DVM modeli olduğu tespit edilmiştir.

Veri setinde özellik seçimi yapıldıktan sonra DVM modeli uygulandığında kalitesiz parça tahmininde herhangi bir değişiklik yaşanmazken modelin genel performansında % 1,7 oranında artış gözlenmiştir.

Özellik seçiminin uygulandığı ve uygulanmadığı veri kümeleri için her bir modelin başarısı Çizelge 4.12'de gösterilmiştir.

Çizelge 4.12. Model Performansları

Model / Performans	Karar Ağacı	Özellik Seçimi ile Karar Ağacı	AdaBoost Karar Ağacı	Özellik Seçimi ile AdaBoost Karar Ağacı	Lojistik Regresyon	Özellik Seçimi ile Lojistik Regresyon	DVM	Özellik Seçimi ile DVM
Kaliteli Parça Tahmini %	75,24	79,80	100	100	84,31	78,30	69,16	70
Kalitesiz Parça Tahmini %	75,94	84,21	92,15	92,15	84,61	79,72	80	84,21
Modelin Genel Performansı %	75,5	81,7	95,5	95,5	84,5	78,9	72,8	74,4

En başarılı sonucu veren AdaBoost algoritması ile desteklenmiş karar ağacı modelidir. Uygulanan model ile ağaç yapısı çıkarılarak ürünlerin kaliteli ya da kalitesiz olmalarını etkileyen değişkenler ve değişkenlerin koşulları belirlenmiştir.

5. TARTIŞMA VE SONUÇ

Yapılan çalışma ile üretim sektöründeki bir işletmenin kaynak verileri kullanılarak parçaların kalitesine etki eden faktörlerin veri madenciliği yöntemleri ile belirlenebilirliği gösterilmiştir. Çalışmanın ilk adımında özellik seçimi yapılmadan veri madenciliği modelleri uygulanmıştır. İkinci adımda ise özellik seçiminin ardından ilk adımda uygulanan veri madenciliği modelleri tekrarlanmış ve iki adımdaki model performansları karşılaştırılmıştır. Bu işlemler makine öğrenimi, veri analizi ve eğitim için uygun olan tamamen bulut tabanlı Python kodlarının yazılıp yürütebildiği bir uygulama kullanılarak gerçekleştirilmiştir.

AdaBoost algoritması ile desteklenmiş karar ağacı en başarılı sonucu veren model olarak belirlenmiştir. Uygulanan model ile ağaç yapısından faydalanarak parçaların kaliteli ya da kalitesi olmasını etkileyen değişkenler ve değişkenlerin koşulları belirlenmiştir.

Yapılan bu çalışma ile üretim sektöründe faaliyet gösteren firmalarda üretim sonucunda elde edilen parçaların kalitesinin iyileştirilmesi ve sürece etki eden değişkenlerin tespit edilmesinde veri madenciliği yöntemlerinin uygulanabilirliği ortaya konmuştur. Uygulamalar sonucunda elde edilen bilgiler işletmelerin benzer süreçlerinde bir kaynak ve farklı bir bakış açısı olarak kullanılabilir.

Elde edilen sonuçlar ile parça kalitesini etkileyen değişkenler için önleyici bakım yöntemleri geliştirilerek üretim süreçlerinde iyileştirme gerçekleştirilirken maliyete olan etkisi azaltılabilir. Bunlara örnek olarak kaynak kabin sıcaklığı ölçer, enerji analizörleri ve sensörler verilebilir.

Uygulanan modellerin genelinde kalitesiz parça tahmin performanslarının kaliteli parça tahminlerinden daha başarılı olduğu tespit edilmiştir. Kullanılan veri setindeki örneklem sayılarının arttırılması ve veri dağılımlarının dengelenmesi ile iki tahmin arasındaki eşitsizliğin giderilebileceği öngörülmüştür.

Gözetimsiz öğrenme veri madenciliği alanında veriye dayalı bir yaklaşım olup etiketlenmemiş veri kümelerini kullanarak veriler arasındaki ilişkileri anlamak ve veri noktalarını benzer özelliklere göre gruplandırmak için kullanılır. Veri madenciliği, veri analitiği ve desen tanıma gibi alanlarda kullanılmaktadır. Gözetimsiz öğrenme algoritmalarına örnek olarak kümeleme (K-Means, hiyerarşik), anomalileri algılama ve Apriori algoritmaları verilebilir. Çalışmaya ilerleyen süreçlerde gözetimsiz öğrenme algoritmaları da uygulanarak, ele alınan verilerin benzer özelliklerine göre gruplandırılması ve etiketlenmemiş verilerin analizlerinin yapılması ile çalışmanın eksik kalabileceği noktalarda desteklemesi sağlanabilir.

KAYNAKLAR

Azmi, S. S., & Baliga, S. (2020). An Overview of Boosting Decision Tree Algorithms utilizing AdaBoost and XGBoost Boosting strategies. International Research Journal of Engineering and Technology, May.

Braha, D. (Ed.). (2013). Data mining for design and manufacturing: methods and applications (Vol. 3). Springer Science & Business Media.

Canlı, H., & Toklu, S. (2019). Implementation of Decision Support System with Data Mining Methods in the Quality Control Process of the Automotive Sector. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 7(1). https://doi.org/10.29130/dubited.427900

Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01). https://doi.org/10.38094/jastt20165

Çelik, M. (2009). Bir otomotiv yan sanayi kuruluşunda veri madenciliği uygulaması (Doctoral dissertation, Bursa Uludag University (Turkey)).

Chen, L. Da, Sakaguchi, T., & Frolick, M. N. (2000). Data mining methods, applications, and tools. Information Systems Management, 17(1). https://doi.org/10.1201/1078/43190.17.1.20000101/31216.9

Doğan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. In Expert Systems with Applications (Vol. 166). Elsevier Ltd. https://doi.org/10.1016/j.eswa.2020.114060

Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. Future Generation Computer Systems, 13(2–3). https://doi.org/10.1016/s0167-739x(97)00015-0

Ferreiro, S., Sierra, B., Irigoien, I., & Gorritxategi, E. (2011b). Data mining for quality control: Burr detection in the drilling process. Computers and Industrial Engineering, 60(4). https://doi.org/10.1016/j.cie.2011.01.018

Işık, K., & Ulusoy, S. K. (2021). Determining the factors that affect the production time in metal industry utilizing data mining methods. Journal of the Faculty of Engineering and Architecture of Gazi University, 36(4). https://doi.org/10.17341/gazimmfd.736659

Kantardzic, M. (2011). Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition. In Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition. https://doi.org/10.1002/9781118029145

Kerber, R., Livezy, B., Simoudis, E., (1995). In: Goonatilake, S., Khebbal, S. (Eds.), A Hybrid System for Data Mining. Intelligent Hybrid Systems. John Wiley, New York.

Köksal, G., Batmaz, I., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. In Expert Systems with Applications (Vol. 38, Issue 10, pp. 13448–13467). https://doi.org/10.1016/j.eswa.2011.04.063

Oğuzlar, A. (2004). Veri Ön İşleme. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21.

Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery workbench for exploring business databases. International Journal of Intelligent Systems, 7(7). https://doi.org/10.1002/int.4550070708

Piramuthu, S. (2004). Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research, 156(2). https://doi.org/10.1016/S0377-2217(02)00911-6

Rokach, L., & Maimon, O. (2006). Data mining for improving the quality of manufacturing: A feature set decomposition approach. Journal of Intelligent Manufacturing, 17(3). https://doi.org/10.1007/s10845-005-0005-x

Sagin, A. N., & Ayvaz, B. (2018). Determination of Association Rules with Market Basket Analysis: Application in the Retail Sector. Southeast Europe Journal of Soft Computing, 7(1). https://doi.org/10.21533/scjournal.v7i1.149

Santos, P., Maudes, J., Bustillo, A., & Rodríguez, J. J. (2013). Improvements in modelling of complex manufacturing processes using classification techniques. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7906 LNAI. https://doi.org/10.1007/978-3-642-38577-3_69

Santoso, M. H. (2021). Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom. Brilliance: Research of Artificial Intelligence, '

Savaş, S., Topaloğlu, N. & Yılmaz M. (2012). Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 1

Wei, W., Yuan, J., & Liu, A. (2020). Manufacturing data-driven process adaptive design method. Procedia CIRP, 91, 728–734. https://doi.org/10.1016/j.procir.2020.02.230

Zhang, Z., Zhang, C., & Zhang, S. (2003). An agent-based hybrid framework for database mining. Applied Artificial Intelligence, 17(5–6). https://doi.org/10.1080/713827179

EK 1 Aykırı Değer Analizi Sonuçları

```
Önceki İstatistikler: count    599.000000
mean      18.237997
std        0.571629
min       17.120000
25%       17.920000
50%       18.200000
75%       18.455000
max       24.490000
Name: X2       dtype: float64

Sonraki İstatistikler: count    599.000000
mean      18.206699
std        0.411335
min       17.120000
25%       17.920000
50%       18.200000
75%       18.455000
max       19.257500
Name: X2      , dtype: float64

Önceki İstatistikler: count    599.000000
mean       6.010618
std        0.845775
min        1.050000
25%        5.366000
50%        6.164000
75%        6.592000
max        7.610000
Name: X3       dtype: float64

Sonraki İstatistikler: count    599.000000
mean       6.022990
std        0.789309
min        3.527000
25%        5.366000
50%        6.164000
75%        6.592000
max        7.610000
Name: X3      , dtype: float64

Önceki İstatistikler: count    599.000000
mean       6.010618
std        0.845775
min        1.050000
25%        5.366000
50%        6.164000
75%        6.592000
max        7.610000
Name: X4      , dtype: float64

Sonraki İstatistikler: count    599.000000
mean       6.022990
std        0.789309
min        3.527000
25%        5.366000
50%        6.164000
75%        6.592000
max        7.610000
Name: X4      , dtype: float64
```

```
Önceki İstatistikler: count    599.000000
mean      16.903915
std        0.903255
min       12.766000
25%       16.566500
50%       17.097000
75%       17.459500
max       18.240000
Name: X5         , dtype: float64

Sonraki İstatistikler: count    599.000000
mean      16.966993
std        0.710243
min       15.227000
25%       16.566500
50%       17.097000
75%       17.459500
max       18.240000
Name: X5         , dtype: float64
```

```
Önceki İstatistikler: count    599.000000
mean       0.007710
std        0.001067
min        0.006052
25%        0.007000
50%        0.007000
75%        0.008000
max        0.022758
Name: X6        , dtype: float64

Sonraki İstatistikler: count    599.000000
mean       0.007667
std        0.000821
min        0.006052
25%        0.007000
50%        0.007000
75%        0.008000
max        0.009500
Name: X6        , dtype: float64
```

```
Önceki İstatistikler: count    599.000000
mean       1.658698
std        0.449896
min        0.000000
25%        1.520000
50%        1.740000
75%        1.935000
max        3.020000
Name: X7  dtype: float64

Sonraki İstatistikler: count    599.000000
mean       1.704374
std        0.293224
min        0.897500
25%        1.520000
50%        1.740000
75%        1.935000
max        2.557500
Name: X7  dtype: float64
```

```
Önceki İstatistikler: count    599.000000
mean       2.464157
std        2.084475
min        0.000000
25%        1.960000
50%        2.050000
75%        2.480000
max       15.320000
Name: X8   dtype: float64

Sonraki İstatistikler: count    599.000000
mean       2.196244
std        0.504820
min        1.180000
25%        1.960000
50%        2.050000
75%        2.480000
max        3.260000
Name: X8   dtype: float64
```

```
Önceki İstatistikler: count    599.000000
mean       0.971402
std        0.789544
min        0.000000
25%        0.150000
50%        0.870000
75%        1.650000
max        4.340000
Name: X9   dtype: float64

Sonraki İstatistikler: count    599.000000
mean       0.970668
std        0.786604
min        0.000000
25%        0.150000
50%        0.870000
75%        1.650000
max        3.900000
Name: X9   dtype: float64
```

```
Önceki İstatistikler: count    599.000000
mean       0.971402
std        0.789544
min        0.000000
25%        0.150000
50%        0.870000
75%        1.650000
max        4.340000
Name: X10 dtype: float64

Sonraki İstatistikler: count    599.000000
mean       0.970668
std        0.786604
min        0.000000
25%        0.150000
50%        0.870000
75%        1.650000
max        3.900000
Name: X10 dtype: float64
```

EK 2 Model Performans Hesaplama

1. Karar ağacı modeli performans hesaplama

Kaliteli Parça Tahmini Performansı %: 100* 76 (doğru tahmin edilmiş kaliteli parça sayısı) / 101 (tahmin edilen tüm kaliteli parça sayısı) = 75,24

Kalitesiz Parça Tahmini Performansı %: 100* 60 (doğru tahmin edilmiş kalitesiz parça sayısı) / 79 (tahmin edilen tüm kalitesiz parça sayısı) = 75,94

Model Genel Performansı %: 100* 136 (doğru tahmin edilmiş tüm parça sayısı) / 180 (tahmin edilen tüm parça sayısı) = 75,5

2. Özellik seçimi ile karar ağacı modeli performans hesaplama

Kaliteli Parça Tahmini Performansı %: 100* 83 (doğru tahmin edilmiş kaliteli parça sayısı) / 104 (tahmin edilen tüm kaliteli parça sayısı) = 79,80

Kalitesiz Parça Tahmini Performansı %: 100* 64 (doğru tahmin edilmiş kalitesiz parça sayısı) / 76 (tahmin edilen tüm kalitesiz parça sayısı) = 84,21

Model Genel Performansı %: 100* 147 (doğru tahmin edilmiş tüm parça sayısı) / 180 (tahmin edilen tüm parça sayısı) = 81,7

3. AdaBoost karar ağacı modeli performans hesaplama

Kaliteli Parça Tahmini Performansı %: 100* 39 (doğru tahmin edilmiş kaliteli parça sayısı) / 39 (tahmin edilen tüm kaliteli parça sayısı) = 100

Kalitesiz Parça Tahmini Performansı %: 100* 47 (doğru tahmin edilmiş kalitesiz parça sayısı) / 41 (tahmin edilen tüm kalitesiz parça sayısı) = 92,15

Model Genel Performansı %: 100* 86 (doğru tahmin edilmiş tüm parça sayısı) / 100 (tahmin edilen tüm parça sayısı) = 95,5

4. Özellik seçimi ile AdaBoost karar ağacı modeli performans hesaplama

Kaliteli Parça Tahmini Performansı %: 100* 39 (doğru tahmin edilmiş kaliteli parça sayısı) / 39 (tahmin edilen tüm kaliteli parça sayısı) = 100

Kalitesiz Parça Tahmini Performansı %: 100* 47 (doğru tahmin edilmiş kalitesiz parça sayısı) / 41 (tahmin edilen tüm kalitesiz parça sayısı) = 92,15

Model Genel Performansı %: 100* 86 (doğru tahmin edilmiş tüm parça sayısı) / 100 (tahmin edilen tüm parça sayısı) = 95,5

5. Lojistik regresyon modeli performans hesaplama

Kaliteli Parça Tahmini Performansı %: 100* 86 (doğru tahmin edilmiş kaliteli parça sayısı) / 102 (tahmin edilen tüm kaliteli parça sayısı) = 84,31

Kalitesiz Parça Tahmini Performansı %: 100* 66 (doğru tahmin edilmiş kalitesiz parça sayısı) / 78 (tahmin edilen tüm kalitesiz parça sayısı) = 84,61

Model Genel Performansı %: 100* 152 (doğru tahmin edilmiş tüm parça sayısı) / 180 (tahmin edilen tüm parça sayısı) = 84,5

6. Özellik seçimi ile lojistik regresyon modeli performans hesaplama

Kaliteli Parça Tahmini Performansı %: 100* 83 (doğru tahmin edilmiş kaliteli parça sayısı) / 106 (tahmin edilen tüm kaliteli parça sayısı) = 78,30

Kalitesiz Parça Tahmini Performansı %: 100* 59 (doğru tahmin edilmiş kalitesiz parça sayısı) / 74 (tahmin edilen tüm kalitesiz parça sayısı) = 79,72

Model Genel Performansı %: 100* 142 (doğru tahmin edilmiş tüm parça sayısı) / 180 (tahmin edilen tüm parça sayısı) = 78,9

7. DVM modeli performans hesaplama

Kaliteli Parça Tahmini Performansı %: 100* 83 (doğru tahmin edilmiş kaliteli parça sayısı) / 120 (tahmin edilen tüm kaliteli parça sayısı) = 69,16

Kalitesiz Parça Tahmini Performansı %: 100* 48 (doğru tahmin edilmiş kalitesiz parça sayısı) / 60 (tahmin edilen tüm kalitesiz parça sayısı) = 80

Model Genel Performansı %: 100* 131 (doğru tahmin edilmiş tüm parça sayısı) / 180 (tahmin edilen tüm parça sayısı) = 75,

8. Özellik Seçimi ile DVM modeli performans hesaplama

Kaliteli Parça Tahmini Performansı %: 100* 86 (doğru tahmin edilmiş kaliteli parça sayısı) / 123 (tahmin edilen tüm kaliteli parça sayısı) = 70

Kalitesiz Parça Tahmini Performansı %: 100* 48 (doğru tahmin edilmiş kalitesiz parça sayısı) / 57 (tahmin edilen tüm kalitesiz parça sayısı) = 84,21

Model Genel Performansı %: 100* 134 (doğru tahmin edilmiş tüm parça sayısı) / 180 (tahmin edilen tüm parça sayısı) = 74,4