

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**ÇOK DEĞİŞKENLİ KONTROL DİYAGRAMINDA KONTROL
DIŞI DURUMA SEBEP OLAN DEĞİŞKENLERİN TOPLULUK
MAKİNE ÖĞRENME ALGORİTMALARI İLE TAHMİNİ**

DOKTORA TEZİ

Deniz DEMİRCİOĞLU DİREN

Enstitü Anabilim Dalı : ENDÜSTRİ MÜHENDİSLİĞİ

Tez Danışmanı : Doç. Dr. Semra BORAN

Ocak 2020

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK DEĞİŞKENLİ KONTROL DİYAGRAMINDA KONTROL
DIŞI DURUMA SEBEP OLAN DEĞİŞKENLERİN TOPLULUK
MAKİNE ÖĞRENME ALGORİTMALARI İLE TAHMİNİ

DOKTORA TEZİ


Deniz DEMİRCİOĞLU DİREN

Enstitü Anabilim Dalı : ENDÜSTRİ MÜHENDİSLİĞİ

Bu tez 31/01/2020 tarihinde aşağıdaki jüri tarafından oybirliği/oyçokluğu ile kabul edilmiştir.



Prof. Dr.
Bayram TOPAL
Jüri Başkanı




Doç. Dr.
Semra BORAN
Üye



Prof. Dr.
İbrahim ÇİL
Üye



Doç. Dr.
Gülşen AYDIN KESKİN
Üye



Dr. Öğr. Uyesi
Tuğba TUNACAN
Üye

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Deniz DEMİRCİOĞLU DİREN

31/01/2020

TEŞEKKÜR

Doktora eğitimim boyunca bilgi ve tecrübeleriyle bana rehberlik eden, çalışmanın planlanmasından yazılmasına kadar olan tüm aşamalarda destekleyen ve her zaman yanımda olduğuna inandığım kıymetli danışman hocam Doç. Dr. Semra BORAN'a en içten saygılarımı ve teşekkürlerimi sunarım.

Tez izleme jürimde olan Prof. Dr. İbrahim ÇİL ve Prof. Dr. Bayram TOPAL hocalarıma bana zaman ayırarak dinledikleri, yapıcı eleştirileri, çalışmanın gelişmesini sağlayan öneri, yorum ve katkılarından dolayı çok teşekkür ederim.

Çalışmamda verdiği manevi destekten dolayı sevgili meslektaşım Arş. Gör. Merve ŞİŞÇİ'ye ve tüm çalışma arkadaşlarıma teşekkürlerimi sunarım.

Tüm hayatım boyunca her alanda beni maddi manevi destekleyen başta sevgili annem Pınar, babam Mustafa ve her an yanımda hissettiğim canım ağabeyim Musa DEMİRCİOĞLU olmak üzere tüm aileme çok teşekkür ederim.

Beni her daim sabır ve anlayışla karşılayan, hiçbir konuda yardımlarını esirgemeyen değerli eşim Selçuk DİREN'e çok teşekkür ederim.

Son olarak hayata geldiğinden beri umut kaynağım olan biricik kızım Nehir DİREN'e çok teşekkür ederim ve bu süreçte ilgimde eksiklik olduysa beni affetmesini dilerim.

İÇİNDEKİLER

TEŞEKKÜR.....	i
İÇİNDEKİLER	ii
SİMGELER VE KISALTMALAR LİSTESİ	vi
ŞEKİLLER LİSTESİ	vii
TABLOLAR LİSTESİ.....	ix
ÖZET.....	xi
SUMMARY	xii
BÖLÜM 1.	
GİRİŞ	1
1.1. Araştırmanın Konusu	2
1.2. Araştırmanın Amacı	3
1.3. Araştırmanın Önemi	4
1.4. Araştırmanın Sınırları.....	5
1.5. Araştırmanın Organizasyonu.....	6
BÖLÜM 2.	
LİTERATÜR ÖZETİ.....	8
2.1. Kontrol Dışı Duruma Neden Olan Değişkenlerin İstatistiksel Yöntemlerle Tespit Edilmesi ile İlgili Çalışmalar.....	8
2.2. Kontrol Dışı Durumlara Neden Olan Değişkenlerin Makine Öğrenme Algoritmalarıyla Tespit Edilmesi ile İlgili Çalışmalar	10
2.2.1. Tekli algoritma ile kontrol dışı durumlara neden olan değişkenlerin tespit edilmesi ile ilgili çalışmalar.....	11
2.2.2. Topluluk algoritmalarıyla kontrol dışı durumlara neden olan değişkenlerin tespit edilmesi ile ilgili çalışmalar.....	13

2.3. Tezin Diğer Çalışmalardan Farkı ve Literatüre Katkısı	15
--	----

BÖLÜM 3.

METODLAR	16
3.1. Hotelling T^2 Kontrol Diyagramı	16
3.1.1. Hotelling T^2 kontrol diyagramı için varsayımlar	18
3.1.1.1. Çok değişkenli normal dağılıma uygunluk varsayımı ...	18
3.1.1.2. Doğrusallık	21
3.1.1.3. Otokorelasyon olmaması	22
3.1.1.4. Varyans kovaryans eşitliği (homojenliği) varsayımı	23
3.2. Mason Young Tracy (MYT) Ayırıştırma Yöntemi	24
3.2.1. Koşullu ve koşulsuz terimlerin tanımlanması ve hesaplanması.	25
3.2.2. Olası MYT ayırıştırmalarının gösterilmesi	26
3.2.3. Koşullu ve koşulsuz terimlerin eşik değerlerinin hesaplanması ve yorumlanması	26
3.3. Makine Öğrenme Algoritmaları	29
3.3.1. Makine öğrenme algoritmalarının tekli kullanılması	32
3.3.1.1. Karar ağaçları	33
3.3.1.2. Naif bayes	35
3.3.1.3. k-En yakın komşu	36
3.3.1.4. Yapay sinir ağları	38
3.3.1.5. Destek vektör makineleri	41
3.3.2. Topluluk makine öğrenme algoritmaları	43
3.3.2.1. Torbalama topluluk yöntemi	45
3.3.2.2. Yükseltme topluluk yöntemi	46
3.3.2.3. Oylama	47
3.3.2.4. Yığılmış genelleme topluluk yöntemi	47
3.3.3. Makine öğrenme performans ölçüm kriterleri	48
3.3.3.1. Hata matrisi	49
3.3.3.2. Kappa istatistiği	50
3.3.4. Dengeli olmayan verinin düzenlenmesi	50

BÖLÜM 4.

KONTROL DIŐI DURUM NEDENLERİNİN TESPİT EDİLMESİ İÇİN

ÖNERİLEN MODEL.....	52
4.1. Veri Toplama ve Düzenleme.....	54
4.2. Kontrol DıŐı Durum Tespiti	54
4.3. Kontrol DıŐı Duruma Neden Olan DeęiŐkenlerin Tespit Edilmesi.....	55
4.4. Veri Seti OluŐturma.....	55
4.5. Önerilen Modelin GeliŐtirilmesi.....	55
4.5.1. Makine öęrenme algoritmalarının tekli olarak kullanılması.....	57
4.5.2. Seçilen makine öęrenme algoritmasının topluluk yöntemleri ile birleŐtirilmesi	57
4.5.3. Topluluk yöntemlerinin yığılmıŐ genelleme topluluk yöntemi ile birleŐtirilmesi	59
4.6. Performans Deęerlendirilmesi.....	60

BÖLÜM 5.

UYGULAMA	61
5.1. Problem Tanımı.....	61
5.2. Çalışmanın Adımlarının Uygulanması.....	64
5.2.1. Veri toplama ve düzenleme.....	64
5.2.2. Kontrol dıŐı duruma neden olan deęiŐkenlerin tespiti.....	65
5.2.2.1. Hotelling T^2 kontrol diyagramının varsayımlarının kontrol edilmesi	65
5.2.2.2. Hotelling T^2 kontrol diyagramının uygulanması	70
5.2.3. Kontrol dıŐı durumlara neden olan deęiŐkenlerin tespit edilmesi.....	73
5.2.4. Veri seti oluŐturma	74
5.2.5. Önerilen modelin geliŐtirilmesi.....	75
5.2.5.1. Makine öęrenme algoritmalarının tekli kullanılması.....	76
5.2.5.2. Seçilen makine öęrenme algoritmasının topluluk yöntemleri ile birleŐtirilmesi	83
5.2.6. Performans deęerlendirmesi.....	85

BÖLÜM 6.

TARTIŞMA VE SONUÇ 87

KAYNAKÇA..... 90

EKLER..... 99

ÖZGEÇMİŞ 106



SİMGELER VE KISALTMALAR LİSTESİ

AKS	: Alt Kontrol Sınırı
ARL	: Ortalama Çalışma Uzunluğu
BK	: Bilgi Kazancı
ÇKA	: Çok Katmanlı Algılayıcı
DVM	: Destek Vektör Makineleri
Ç-DVM	: Çok Sınıflı Destek Vektör Makineleri
ED	: Eşik Değer
GEKK	: Genelleştirilmiş En Küçük Kareler
GI	: Gini İndeksi
GN	: Gerçek Negatif
GP	: Gerçek Pozitif
KA	: Karar Ağaçları
KNN	: K-En Yakın Komşu
KO	: Kazanç Oranı
MCUSUM	: Çok Değişkenli Kümülatif Toplam Kontrol Diyagramı
MYT	: Mason Young Tracy Ayrıştırma Yöntemi
MEWMA	: Çok Değişkenli Üstel Ağırlıklı Hareketli Ortalama Diyagramı
NB	: Naif Bayes
OÇU	: Ortalama Çalışma Uzunluğu
SEKK	: Sıradan En Küçük Kareler
Sig	: Anlamlılık Düzeyi
SMOTE	: Sentetik Azınlık Aşırı Örnekleme Tekniği
TBA	: Temel Bileşen Analizi
ÜKS	: Üst Kontrol Sınırı
YN	: Yanlış Negatif

ŞEKİLLER LİSTESİ

Şekil 1.1. Çalışmanın adımları	6
Şekil 3.1. Normal dağılım dönüşümleri	21
Şekil 3.2. Makine öğrenme ile diğer alanların ilişkisi	29
Şekil 3.3. 3 Katlamalı çapraz doğrulama yöntemi	31
Şekil 3.4. Karar ağacı yapısı	33
Şekil 3.5. K-NN yapısı	37
Şekil 3.6. Sinir ağı yapısı	38
Şekil 3.7. Çok katmalı algılayıcı yapısı	39
Şekil 3.8. 2 sınıflı problem için destek vektör makinesi örneği	41
Şekil 3.9. Topluluk algoritmalarının bileşenleri	44
Şekil 3.10. Topluluk algoritmalarının çeşitleri	45
Şekil 4.1. Çalışmanın akış şeması	53
Şekil 4.2. Önerilen model genel yapısı	56
Şekil 4.3. Temel algoritmaların tekli kullanımı	57
Şekil 4.4. Torbalama yöntemi ile birleştirme	58
Şekil 4.5. Yükseltme yöntemi ile birleştirme	59
Şekil 4.6. Yığılmış genelleme yöntemi ile topluluk algoritmalarının birleştirilmesi	59
Şekil 5.1. Süreç akış şeması	62
Şekil 5.2. Hidrolik pompa kapağı	63
Şekil 5.3. Üstten görünüş	63
Şekil 5.4. Önden görünüş	63
Şekil 5.5. Mahalanobis ve ki-kare değerleri arasındaki serpilme grafiği	69
Şekil 5.6. Hotelling T^2 sonuçları	72
Şekil 5.7. Tekli Algoritmaların performans karşılaştırılması	82
Şekil 5.8. Çalışmanın önerilen modeli	84

Şekil 5.9. Önerilen modelin performansının diğer modeller ile karşılaştırılması... 86



TABLolar LİSTESİ

Tablo 2.1. Makine öğrenme yöntemleri ile yapılmış çok değişkenli kontrol dışı durum tespit çalışmaları.	15
Tablo 3.1. Kernel fonksiyon ve sınıflandırıcı çeşitleri.	42
Tablo 3.2. Oylama birleştirme kuralları.	47
Tablo 3.3. Hata matrisi.	49
Tablo 3.4. Çok sınıflı problemlerde sınıflandırma için hata matrisine bağlı performans kriterleri.	49
Tablo 5.1. Değişkenlerin tanımı.	64
Tablo 5.2. Değişkenler arası korelasyon matrisi ilk durum.	66
Tablo 5.3. Değişkenler arası korelasyon matrisi son durum.	67
Tablo 5.4. Tek değişkenli normal dağılım sonuçları.	68
Tablo 5.5. Çok değişkenli normal dağılım sonuçları.	69
Tablo 5.6. Hotelling T^2 değerleri.	71
Tablo 5.7. MYT koşulsuz parça T^2 değerleri ve kontrol dışı durumlar.	73
Tablo 5.8. Kontrol dışı durumları ve karşılaşılan örnek sayıları.	74
Tablo 5.9. Sınıflandırma için çapraz doğrulama parametreleri.	76
Tablo 5.10. Karar ağacı parametreleri.	77
Tablo 5.11. Karar ağacı performans değerleri.	77
Tablo 5.12. K-NN k parametrelerine göre performans değerleri.	78
Tablo 5.13. K-NN algoritması parametreleri.	78
Tablo 5.14. K-NN performans değerleri.	78
Tablo 5.15. NB performans değerleri.	79
Tablo 5.16. Ç-DVM algoritması.	80
Tablo 5.17. Ç-DVM performans değerleri.	80
Tablo 5.18. YSA algoritması parametreleri.	81
Tablo 5.19. YSA performans değerleri.	81

Tablo 5.20. Karar ağacı-torbalama performans değerleri.....	83
Tablo 5.21. Karar ağacı-adaboost performans değerleri.....	84
Tablo 5.22. Yığılmış genelleme performans değerleri.....	85



ÖZET

Anahtar kelimeler: Çok Değişkenli Kontrol Diyagramları, Makine Öğrenme Algoritmaları, Topluluk Öğrenme Algoritmaları

Çok değişkenli kontrol diyagramları her değişkeni tek tek değerlendirmek yerine çok sayıda değişkeni tek bir diyagram üzerinde değerlendirmektedir. Bu sayede zaman ve iş yükünden kazanım sağlamanın yanında değişkenler arasındaki ilişkiler de değerlendirilmektedir. Bu avantajların aksine oluşan kontrol dışı durumların hangi değişkenlerden kaynaklandığı belirleyememek gibi bir dezavantajı vardır. Ancak sürecin kontrol altına alınabilmesi için hangi değişkenlere düzeltici faaliyetler uygulanması gerektiği bilinmelidir. Bu konuda yardımcı bilimsel yöntemlere ihtiyaç duyulmaktadır. Literatürde kullanılan istatistiksel ve makine öğrenme teknikleri mevcuttur. İstatistiksel yöntemlerin gelecek durumları tahmin edememe eksikliğinden dolayı makine öğrenme yöntemleri kullanılmıştır.

Çalışmada, bu problemi ortadan kaldırmak için kontrol dışı durumların nedenlerini sınıflandıracak makine öğrenme tabanlı bir model geliştirilmiştir. Modelin sınıflandırma doğruluklarının mümkün olan en yüksek oranlarda olması hedeflenmektedir. Tahmin doğruluklarını arttırmak için temel tekli makine öğrenme algoritmalarının optimum parametrelerle çözüm üretmesinin yanında algoritmaları birleştirerek doğrulukları arttırmayı amaçlayan topluluk makine öğrenme algoritmaları kullanılmıştır. Kullanılan 5 temel tekli algoritmanın arasından en başarılı olarak bulunan karar ağacı algoritması, torbalama (bagging) ve yükseltme (boosting) yöntemleriyle ayrı ayrı birleştirilmiş ve doğruluklar artmıştır. Geliştirilen modelde, iyileştirilen bu algoritmalar, yığılmış genelleme (stacking) yöntemi ile birleştirilerek kullanılmıştır. Topluluk algoritmalarının bu şekilde iç içe kullanılmasının tahmin doğruluklarını arttıracığı düşünülmektedir.

Modelin başarısının ispatlanması için gerçek hayat uygulaması yapılmıştır. Temel tekli makine öğrenme algoritmaları ve iki topluluk algoritması ile karşılaştırılarak çalışma başarısı kanıtlanmıştır. Geliştirilen model sayesinde; hem çok değişkenli kontrol diyagramı kullanılarak zaman, maliyet ve değişkenler arası ilişkilerin de dikkate alınması gibi faydalar sağlanmıştır hem de yeni örneklerin kontrol dışı durumların nedenlerinin %98,06 gibi büyük doğruluk oranları ile tespit edilerek hızlıca çözülebileme şansı sunulmuştur.

PREDICTION OF VARIABLES THAT THE CAUSE OF OUT OF CONTROL CONDITION ON MULTIVARIATE CONTROL CHART BY ENSEMBLE MACHINE LEARNING ALGORITHM

SUMMARY

Keywords: Multivariate Control Diagrams, Machine Learning Algorithms, Ensemble Machine Learning Algorithms

Multivariate control charts enable assessment of multi variable on a single chart rather than evaluating them individually. This control chart has the great advantage of not only saving time and workload, but also evaluating the relationships between variables. Contrary to these advantages, there is a disadvantage of not being able to determine which variables arise out of control. However, it should be known which corrective actions should be applied to the variable(s) in order to control the process. Supporting scientific methods are needed in this regard. Statistical and machine learning techniques are available in the literature. Machine learning methods have been used because of the lack of statistical methods to predict future situations.

In this study, a machine learning based model has been developed to prediction of variables that the cause of out of control condition to eliminate this problem. The classification accuracy of the model is aimed to be as high as possible. In order to increase the accuracy of predictions, the basic single machine learning algorithms produce solutions with the most optimum parameters, and ensemble machine learning algorithms aiming to increase the accuracy by combining the algorithms have been used. The decision tree bagging and boosting methods, which were found to be the most successful among the 5 basic single algorithms, were combined separately and the accuracy increased. In the developed model, these two algorithms were combined with the stacking method and the other two machine learning algorithms were used together. Such use of nested ensemble algorithms is thought to improve the prediction accuracy. In order to prove the success of the model, it was applied in real life. The proposed model has been compared with the single machine learning algorithm, and two ensemble algorithms to prove the success of the study. By the help of the developed model, benefits such as consideration of time, cost and the relationships between variables have been obtained by use of multivariate control chart. In addition to fast diagnosis of the cause of out of control condition from new samples with high accuracy up to 98.06%.

BÖLÜM 1. GİRİŞ

Kalitenin önemi her geçen gün artmaktadır. Firmalar rekabet üstünlüğü elde edebilmek için maliyetlerin yanında kalite faktörüne de çok daha fazla önem vermektedirler. Ancak günümüzde süreçlerin giderek daha karmaşık olması nedeniyle kalitenin kontrol edilmesi ve iyileştirilmesi zorlaşmaktadır. Üretimin ayrılmaz bir parçası olarak her noktasında olması gereken kalitenin, değerlendirilmesini kolaylaştırmak için bazı yöntemler geliştirilmiştir. Bu yöntemlerin başında süreçlerden örnek alarak bu örneklerin ölçüm değerlerini bir diyagramda kontrol sınırları ile değerlendirmeyi sağlayan istatistiksel kontrol diyagramları gelmektedir [1]. Uygun kontrol diyagramı kullanımı ile başarılı bir kalite düzeyi elde edilebilecektir.

Süreçler, etki eden değişken sayısı açısından değerlendirildiğinde, tek değişkenli ve çok değişkenli olarak ayrılmaktadır. Değişken sayısı kontrol diyagramı seçimi için kullanıcı açısından en belirgin özellik olsa da, sürecin tek ve çok değişkenli olarak belirlenmesi için değişken sayısından başka özellikler de mevcuttur.

Çok değişkenli bir sürecin değişkenleri çoğu zaman birbirinden bağımsız değildir, aralarında ilişki söz konusudur ve bu nedenle bir grup olarak birlikte değerlendirmek gerekir [2]. Çünkü değişkenleri ayrı ayrı incelemek hem zaman kaybına hem de değişkenlerin aralarındaki ilişkinin kaybedilmesine neden olmaktadır [3]. Bahsedilen bu dezavantajları ortadan kaldırmak için, süreci çok değişkenli olarak değerlendirerek, izlemek ve kontrol dışı durumları tespit etmek gereklidir.

Bu amaçla bütün değişkenleri tek bir istatistiksel noktada değerlendirmeyi sağlayarak aynı diyagramda inceleme şansı sunan çok değişkenli kontrol diyagramları geliştirilmiştir [4-6]. Her bir X adet gözlem biriminde ölçülen p adet değişkene karşılık gelen ölçüm değerler vektörüyle temsil edilen [7] bu diyagramlar, değişkenler

arasındaki ilişkileri yani korelasyonları da incelediği için daha doğru tespitler yapmaktadır [8].

Çok değişkenli kontrol diyagramlarının kontrol dışı durumun belirlenmesindeki başarısına karşılık bu duruma neden olan değişkenler ile ilgili yorum yapamaması gibi çok büyük bir eksikliği mevcuttur [9]. Tek değişkenin etki ettiği süreçlerde kontrol dışı durum ile karşılaşıldığında buna neden olan değişken açıkça bellidir ancak birden fazla değişkenin eş zamanlı olarak incelendiği ilişkili bir süreçte kontrol dışı duruma neden olan değişkenleri tanımlayarak tespit etmek biraz daha karmaşık olmaktadır. Kontrol dışı durumları değişkenler arasındaki ilişkileri de göz önünde bulundurarak tespit etmek için ilave bir yönteme ihtiyaç vardır.

Literatürde kontrol dışı duruma neden olan değişkenlerin tespiti ile ilgili hem istatistiksel yöntemler hem de makine öğrenme algoritmalarını kullanılan çalışmalar mevcuttur. Bunlar çalışmanın ikinci bölümünde yer alan literatür kısmında detaylı olarak ele alınmıştır.

Çok fazla sayıda değişken ve veri içeren üretim süreçlerinde, kontrol dışı duruma neden olan değişkenlerin sınıflandırılması için, kontrol diyagramlarına özel olarak geliştirilen ayrıştırma yöntemi başta olmak üzere, istatistiksel yöntemler her ne kadar çözüm üretse de, verilerden öğrenerek karmaşık işlemler gerçekleştirebilen makine öğrenme algoritmaları gibi kullanışlı olmamaktadır.

Bu çalışmada, kontrol dışı duruma neden olan değişkenlerin tespiti için topluluk makine öğrenme algoritmaları temelli bir model önerilmiştir.

1.1. Araştırmanın Konusu

Bu çalışmanın konusu; çok değişkenli kontrol diyagramları ile karşılaşılan kontrol dışı durumlara neden olan değişkenlerin tespit edilmesini kapsamaktadır. Bu amaçla topluluk makine öğrenme algoritmalarının iç içe kullanıldığı bir sınıflandırma ve tahmin modeli önerilmiştir.

Bu modelle kontrol dışı durumlara neden olan değişkenlerin en doğru şekilde sınıflandırılarak tahmin edilmesi amaçlanmıştır. Sınıflandırma doğruluklarının yüksek olması hem modelde kullanılan algoritmaların başarısına hem de veri setinin doğruluğuna bağlıdır. Bu nedenle modelde kullanılan gerçek ölçüm değerlerine karşılık gelen hedef değerlerini belirlemek için, literatürde en sık kullanılan çok değişkenli kontrol diyagramı olan Hotelling T^2 [2, 3, 10] ve bu diyagram için özel olarak geliştirilmiş Mason Young Tracy (MYT) Ayırıştırma yöntemi [11] kullanılmıştır. Gerçek ölçüm değerleri kullanılarak bir veri seti oluşturulmuş ve kontrol dışı durumlara neden olan değişkenlerin tespit edilmesi amacıyla sınıflandırılmıştır. Sınıflandırma başarılarının artırılması için topluluk makine öğrenme algoritmalarının iç içe kullanıldığı bir model önerilmiştir. Yığılmış genelleme topluluk algoritmasını temel alan bu model, başarısının ispatlanması için, tekli makine öğrenme algoritmaları ve torbalama ve yükseltme olmak üzere iki topluluk algoritması ile karşılaştırılmıştır.

Bu modelin geçerliliğini test etmek için, döküm sürecindeki hata sıklıkları yüksek bir parça üzerinde kalite kontrol problemi incelenmiştir. Kalite değerlendirmesi açısından, çok değişkenli kontrol diyagramlarının varsayımlarına uygun 6 değişkenin eşzamanlı etkili olduğu bir üretim sürecindeki kontrol dışı duruma neden olan değişkenler tespit edilmiştir.

1.2. Araştırmanın Amacı

Çalışmanın amacı çok değişkenli süreçlerde kontrol dışı duruma neden olan değişkenlerin doğru bir şekilde belirlenmek ve tespit edilmiş durumlar için düzeltici faaliyetler ile birlikte süreç iyileştirme çalışmalarını geliştirmektir. Bu amacı gerçekleştirmek için sağlanması gereken alt amaçlar mevcuttur. Bunlar aşağıdaki gibi sıralanmaktadır.

- Çok değişkenin eş zamanlı olarak etki ettiği bir süreçte, değişkenleri kalite açısından tek tek ele almak yerine hepsini birlikte değerlendiren ve bu sayede aralarındaki ilişkiyi göz önünde bulundurarak ciddi bir iş gücü ve maliyet

yükünün azaltılmasını sağlayan çok değişkenli kontrol diyagramlarının kullanmak.

- Geçmiş verilerde çok değişkenli kontrol diyagramlarında meydana gelen kontrol dışı duruma neden olan değişkenlerin uzman bilgisi kullanılarak ya da ortalamalardan kaymalara göre sentetik veri üretmek belirlenmesi yerine gerçek veriler kullanılarak ve bilimsel olarak istatistiksel bir yöntemle tespit etmek
- Değişkenlere ait süreçten elde edilen ölçüm değerlerinin, kontrol dışı duruma neden olan olası değişkenlerle birleştirilmesiyle oluşturulmuş veri seti kullanılarak makine öğrenme algoritmalarıyla sınıflandırma ve tahmin yapmak.
- Kontrol dışı duruma neden olan değişkenlerin en doğru şekilde tespit edilebilmesi için tekli makine öğrenme algoritmalarının en yüksek doğruluk oranlarıyla çalışmasını sağlamak.
- En iyi parametrelere göre sınıflandırmada kullanılan tekli algoritmalar arasında sınıflandırma performansı en yüksek olan algoritmayı seçmek ve performansları yükseltmek için tekli algoritmaları farklı tür topluluk yöntemleri ile birleştirmek. Farklı topluluk algoritmalarının birlikte kullanılması ile doğrulukları daha da arttırmak.

1.3. Araştırmanın Önemi

Çok değişkenin etki ettiği süreçlerde kontrol dışı durum görüldüğünde bunun hangi değişkenlerden dolayı gerçekleştiğini belirlemek, düzeltici faaliyetlerin alınması açısından çok önemlidir. Ancak bu konuda çok değişkenli kontrol diyagramları için özel olarak geliştirilen MYT nin hesaplama zorluğu ve gelecek durumu tahmin edememe gibi eksikleri mevcuttur [12]. Bu nedenle MYT yönteminden veri setini oluştururken hedef değerlerin belirlenmesinde faydalanılmıştır. Tahmin etme probleminin çözümlenebilmesi amacıyla tekli ve topluluk makine öğrenme yöntemleri ile sınıflandırma yapılmıştır.

Algoritmalar kontrol dışı duruma neden olan değişkenleri ne kadar doğru şekilde tahmin ederse önerilecek düzeltme faaliyetleri ve faydalar o kadar başarılı olacaktır. Bu nedenle doğrulukları yükseltmek için topluluk makine öğrenme algoritmalarının iç

içe kullanıldığı bir model geliştirilmiştir. Bunun için literatürde daha önce kontrol dışı durumların sınıflandırılmasında kullanılmamış olan Yığılmış genelleme topluluk algoritması Torbalama ve Yükseltme yöntemleri ile birleştirilerek iç içe topluluk öğrenme algoritmasını kullanan bir model geliştirilmiş ve başarılı sonuçlar elde edilmiştir.

Çalışmada önerilen makine öğrenme algoritma temelli model, süreçlerde değişen durumlara uyum sağlayabilen ve geçmiş dönemlerde kontrol dışı duruma neden olan değişkenleri öğrenerek, yeni dönemlerdeki değişkenleri tahmin edebilen bir model olacaktır.

Ayrıca, Hotelling T^2 kontrol diyagramının 2. Aşamasına alternatif olarak kullanılacak bu model sayesinde, uzman bilgisine ihtiyaç duymadan, kontrol dışı duruma neden olan değişkenlerin doğru şekilde tespit edilmesi iş gücü, zaman ve maliyetler açısından fayda sağlayacaktır.

Çalışmanın bir diğer önemli katkısı da, kontrol dışı duruma neden olan değişkenlerin belirlenmesi sayesinde kalite açısından yapılacak düzeltmelerin ürün de değil süreçte iyileştirmeler sağlayacak olmasıdır.

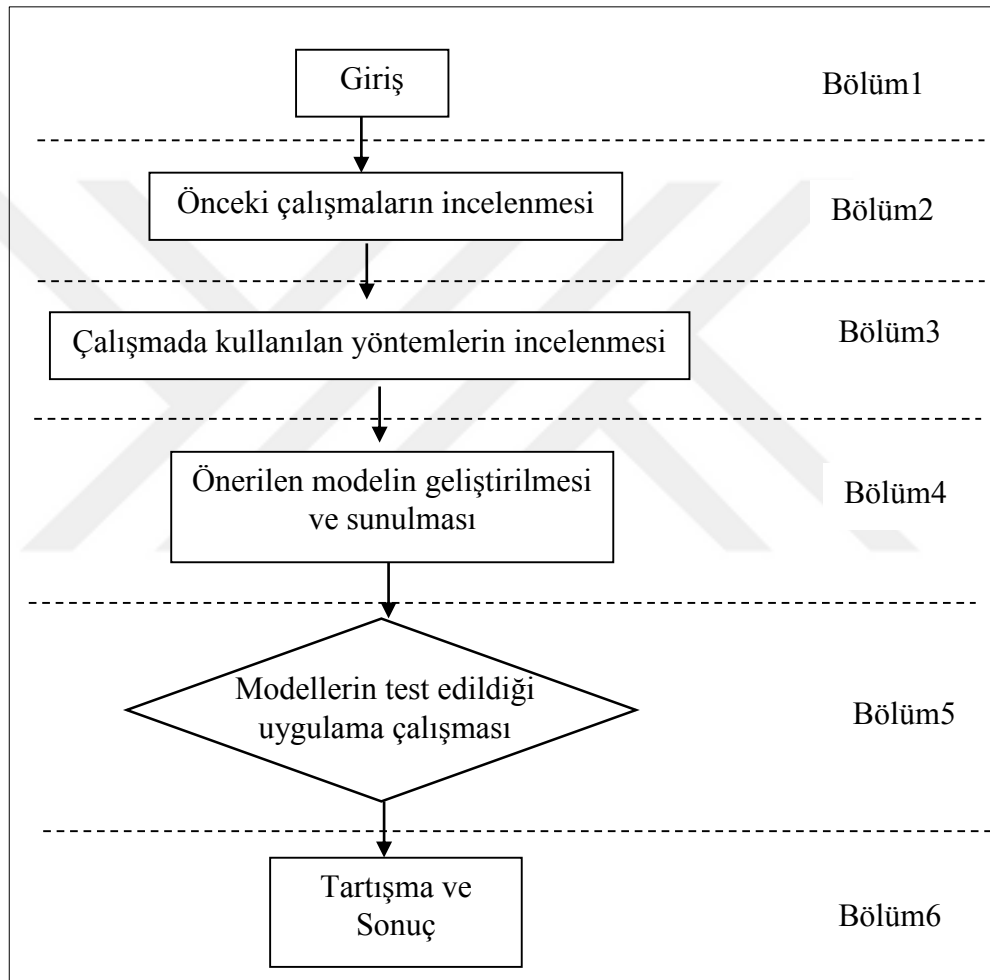
1.4. Araştırmanın Sınırları

Literatürde çok sayıda ve kombinasyonda algoritma ve birleştirme yöntemi mevcuttur ancak çalışmanın sınırlandırılması adına kontrol dışı durumları sınıflandırmak için sadece en temel sınıflandırma algoritmaları (Karar Ağaçları, Naif Bayes (Naive Bayes), Yapay Sinir Ağları, K-En Yakın Komşu, Destek Vektör Makineleri) ve bu algoritmaları birleştirmek için en temel topluluk yöntemleri (torbalama, yükseltme, yığılmış genelleme) kullanılmıştır. Bu model problemlere ve verilere göre özelleştirilerek tüm süreçlerde kullanılabilir.

1.5. Araştırmanın Organizasyonu

Bahsedilen amaçlar kapsamında 6 bölümden oluşan çalışmanın adımları Şekil 1.1.'de gösterilmiştir.

Bu başlık altında çalışmadaki bölümler ve bölümlerin kapsamı anlatılacaktır.



Şekil 1.1. Çalışmanın adımları.

Birinci bölümde çalışma konusu ile ilgili genel bir bakış açısı sağlayacak bilgiler verilerek tezin konusu, amacı, kapsamı ve sınırları belirtilmiştir.

İkinci bölümde bu çalışma konusuyla ilgili daha önce yapılmış çalışmaların araştırıldığı literatür özetlerine yer verilmiştir.

Amaçlanan sonuçlara ulaşmak için kullanılması gereken çok değişkenli Hotelling T^2 kontrol diyagramları, MYT ayrıştırma yöntemi, Karar Ağaçları, Naif Bayes, Yapay Sinir Ağları, K-En Yakın Komşu, Destek Vektör Makineleri algoritmalarının hem tekli hem de topluluk şeklinde kullanılması ile oluşturulan algoritma yapıları ve algoritmaların performans kriterleri gibi araç ve yöntemlerin incelenmesi ile ilgili anlatımlar ve temel kavramlar Bölüm 3’de yer almaktadır.

Çalışmanın 4. bölümünde önerilen modelin geliştirilmesi ve farklı modellerle karşılaştırılması sunulmuştur.

Bölüm 5’de önerilen modelin başarısının değerlendirilebilmesi için gerçekleştirilen gerçek hayat probleminin uygulaması gerçekleştirilmiştir. Bu bölümde modelin eğitilebilmesi ve test edilebilmesi için kullanılan algoritmaların parametre değerleri belirlenerek bunlara bağlı sonuçlar hesaplanmıştır. Bu sonuçlar ışığında önerilen modelin başarısı ve sürece katkısı gösterilmiştir.

Son olarak 6. Bölümünde yer alan tartışma ve sonuç kısmında, çalışma değerlendirilerek gelecek önerilerinde bulunulmuştur.

BÖLÜM 2. LİTERATÜR ÖZETİ

Bu bölümde çok değişkenli kontrol diyagramlarında, kontrol dışı duruma neden olan değişkenlerin tespit edilmesi konusunda literatürde yer alan istatistiksel ve makine öğrenme yöntemleri ile ilgili çalışmalar ele alınmıştır.

2.1. Kontrol Dışı Duruma Neden Olan Değişkenlerin İstatistiksel Yöntemlerle Tespit Edilmesi ile İlgili Çalışmalar

Kontrol dışı durumların istatistiksel yöntemlerle belirlendiği çalışmalardan ilki Jackson (1985) tarafından yapılmıştır. Çalışmada geliştirilen temel bileşen analizi (TBA) yöntemi ile boyut azaltma sağlanarak verilerin karmaşıklığı azaltılmış ve arasındaki ilişkiler tanımlanmıştır [13]. Rao ve diğerleri (2013), Jozsef ve Robert (2017) ile Güler ve Bakır (2019) tarafından yapılan çalışmalarda TBA yöntemi kontrol diyagramlarında kullanılmıştır [14, 16].

Temel bileşen analizi yönteminin kalite kontrol diyagramlarıyla birleştirilmesinin ardından Murphy (1987) diskriminant analizini esas alan bir model geliştirmiştir [17]. Kontrol diyagramları için diğer yöntemlere göre daha az tercih edilen bu yöntemin performansı, Pei ve diğerleri (2006) tarafından yapılan uygulama çalışması ile değerlendirilmiştir [18].

Doganaksoy ve arkadaşları (1991) kontrol dışı duruma neden olan değişkenlerin belirlenmesi amacıyla Bonferroni tipi limitleri kullanan yöntem ile bir çalışma sunmuşlardır [19]. Yine aynı amaçla Hawkins (1991) tarafından bireysel değişkenlere yönelik regresyon düzeltmesi kullanılmıştır [20]. Bu yöntem genellikle Bersimis ve arkadaşları (2017) ile Puig ve Ferrer (2014) tarafından yapılan çalışmalarda olduğu gibi diğer yöntemlerle performans karşılaştırma çalışmalarında kullanılmıştır [21, 22].

Bu yöntemlerin yanında, literatürde çok değişkenli kontrol diyagramları için en sık kullanılan yöntem [2], bileşenler analizinden yola çıkarak Hotelling T^2 kontrol diyagramına özel olarak tasarlanmış ve Mason Young Tracy (1995) tarafından geliştirilmiştir [11, 23]. Ardından yine aynı yazarlar tarafından 1997 yılında daha hızlı bir sıralı hesaplama sağlamak için adımlar önerilerek Mason Young Tracy (MYT) Ayırıştırma Yöntemi olarak adlandırılmıştır [24].

MYT yöntemi kullanılarak farklı alanlarda yapılmış çalışmalar mevcuttur. Parra ve Loaiza (2003), Ulen ve Demir (2013) tarafından kimya ve eczacılık alanında uygulama çalışmaları yapılmıştır [25, 26]. Salmona (2005) cam takviyeli plastik boru üretim sürecinde [10], Çetin ve Birgören (2007) pirinç döküm eritme sürecinde [8] ve Boullosa ve arkadaşları (2017) tarafından ise deniz dizel motorun silindir yağlama işlemi için yapılan [27] çalışmalar MYT yöntemiyle uygulamalara örnek olarak gösterilebilir. Üretim süreçleriyle ilgili yapılan uygulamaların yanında, MYT yönteminin hizmet sektöründe de kullanıldığı çalışmalar mevcuttur. Yılmaz (2012) tarafından yapılan çalışmada hastanelerdeki hizmet kalitesini değerlendirebilmek ve bu alanda memnuniyeti ölçmek için MYT yöntemi kullanılmıştır [28].

Ayrıca MYT yönteminin farklı yöntemlerle birleştirildiği, karşılaştırıldığı ya da incelenen değişken sayısının farklılaştırıldığı çalışmalar da mevcuttur. Çalışmalardan birkaçı aşağıda yer almaktadır.

Das ve Prakash (2008) tarafından, ortalamalardan farklı kayma büyüklükleri ile üretilmiş veriler üzerinde üç farklı yöntem ile yapılan karşılaştırma çalışmasında, MYT yönteminin performansının çoğunlukla yüksek olduğunu belirtmiştir [29].

Li ve diğerleri (2008) tarafından yapılan çalışmada, MYT yöntemine Bayesien ağları eklenerek nedensellik esaslı bir T^2 ayırıştırması önerilmiştir. Geliştirilen bu yöntemin hesaplama kolaylığı sağladığı ve sıcak şekillendirme sürecinde yapılan bir uygulama ile sonuçların başarılı olduğu belirtilmiştir [30].

Bir diğ er ç alıřma ise Monarrez (2013) tarafından yapılmıřtır. Bu ç alıřmada Hotelling T^2 diyagramındaki Faz I'in kovaryans matrisinin tersine bakılarak her deęiřken çiftinin iliřkilerine dayanan pratik bir ayırıtırma yöntemi önerilmiř ve MYT yöntemindeki karmařıklıęı azaltarak bařarılı olduęu belirtilmiřtir [31].

Agog ve dięerleri (2014) tarafından yapılan ç alıřmada MYT yöntemi deęiřken sayısı arttırılarak uygulanmıřtır. Yöntemin karmařık iřlem yapısından dolayı genellikle 2 ya da 3 deęiřken kullanıldıęını ancak bu ç alıřmada 4 deęiřken için kullanılarak geniřletildięi belirtilmiřtir [32]. Benzer olarak Akeem ve dięerleri (2015) tarafından yapılan ç alıřma da ise yöntem 5 deęiřken için kullanılmıřtır [33].

Bersimis ve dięerlerinin (2017) de yaptıęı ç alıřmada farklı simülasyon senaryoları geliřtirerek literatürdeki çok sayıdaki yöntemi deęerlendirilmiř ve sonuç olarak YSAve MYT yönteminin dięer yöntemlere göre daha üstün olduęu görölmüřtür [21].

Dięer ç alıřmalardan farklı olarak Huang ve dięerleri (2017) tarafından yapılan ç alıřmada MYT yöntemi, K fonksiyonu ile geliřtirilen T^2 istatistięi üzerine uygulanmıřtır. Önerilen yaklařımın etkinlięi kompozit malzemelerdeki nanopartiküller gibi üretim sürecinde test edilerek bařarılı bulunmuřtur [34].

MYT yöntemi Hotelling T^2 kontrol diyagramına özel olarak geliřtirilmesine raęmen, bir dięer çok deęiřkenli kontrol diyagramı olan, çok deęiřkenli üstel aęırlıklandırılmıř hareketli ortalama, diyagramında da kullanıldıęı ç alıřmalara rastlanmıřtır. Schaffer ve arkadaşları (2001), Muhammed ve arkadaşları (2016), Oktay ve Orçanlı (2018) tarafından yapılan ç alıřmalar bu konuda örnek olarak gösterilebilir [35–37].

2.2. Kontrol Dıřı Durumlara Neden Olan Deęiřkenlerin Makine Öğrenme Algoritmalarıyla Tespit Edilmesi ile İlgili Ç alıřmalar

Literatürdeki kontrol dıřı duruma neden olan deęiřkenlerin belirlenmesi için makine öğrenme algoritmalarının kullanıldıęı ç alıřmalar iki sınıfta incelenmektedir. Temel makine öğrenme sınıflandırma algoritmalarının tek olarak kullanılması ve aynı ya da

farklı türde algoritmaların birlikte kullanımı ile oluşan topluluk algoritmaları ile kontrol dışı duruma neden olan değişkenlerin tespit edilmesi için çalışmalar gerçekleştirilmiştir. Araştırma bu şekilde gruplandırılarak yapılmıştır.

Çalışmada, verilere ve sınıf sayısına uygunluk açısından sınıflandırma algoritmaları arasından, Karar Ağaçları (KA), Naif Bayes (NB), K-En Yakın Komşu (KNN), Çoklu Destek Vektör Makineleri (Ç-DVM) ve Yapay Sinir Ağları (YSA) olmak üzere 5 temel algoritma kullanılmıştır. Literatürde bu algoritmalar ile yapılan çalışmalara ağırlık verilerek incelenmiştir.

2.2.1. Tekli algoritma ile kontrol dışı durumlara neden olan değişkenlerin tespit edilmesi ile ilgili çalışmalar

Chen ve Wang (2004) tarafından kontrol dışı duruma neden olan değişkenlerin ortalamadan kayma büyüklüklerine göre tespit edilmesi amacıyla yapılan çalışmada X^2 diyagramını için yapay sinir ağı tabanlı bir model geliştirilmiş ve performansları değerlendirilerek sunulmuştur [38].

Niaki ve Abbasi (2005) yapay sinir ağlarının bir ağ mimarisi olan çok katmalı algılayıcı (ÇKA) temelli bir model geliştirerek kontrol dışı durumlara neden olan değişkenleri sınıflandırmıştır [12].

Aparisi ve diğerleri (2006) tarafından yapılan çalışmada MYT yönteminin sınıflandırma açısından doğruluk analizi gerçekleştirilmiştir. Çalışmanın ikinci bölümünde ise, doğruluk analizi bir de sinir ağı yaklaşımıyla gerçekleştirilmiştir. Sonuçlara göre; yeni tasarlanan sinir ağının doğruluk performansının MYT yönteminin doğruluk performansından daha iyi olduğu görülmüştür [9].

Çok değişkenli süreçlerdeki varyans kaymalarının sınıflandırılması için Cheng ve Cheng (2008) tarafından yapılan çalışmada, YSA ve DVM uygulanmıştır. Performanslar tahminlerin ortalama sınıflandırma doğruluğuyla değerlendirilmiştir. Buna göre, DVM' nin performansının YSA'ya benzer olduğunun yanında, YSA

algoritmasının kontrol parametre sayısının çok olması ve uygulama adımlarının zor olması gibi zayıf yönlerinin olduğu belirtilmiştir [39].

Aparisi ve Sanz (2010) MEWMA kontrol diyagramlarında yapılmış ilk araştırma özelliği taşıyan çalışmalarında yine yapay sinir ağları kullanılarak kontrol diyagramındaki kontrol dışı durumlar yorumlamışlardır. Tasarlanan bu modelde farklı değişken sayılarına göre doğru sınıflandırma yüzdesi incelenmiş. Ayrıca modelin endüstride kullanımını kolaylaştırmak için bir ara yüz tasarımı geliştirilmiştir [40].

Li ve diğerleri (2013) tarafından yapılan çalışmada proses kayma büyüklüğünün tahmini için optimize DVM yaklaşımı önerilmiştir. Önerilen optimize DVM yaklaşımı ANN ile karşılaştırılmıştır. Bu çalışma sonucunda DVM yaklaşımının en iyi performansı ortaya koyduğu görülmüştür [41].

Huda ve diğerleri (2014) yaptıkları çalışmada uzman bilgisine ihtiyaç duymayan ve az sayısal hesaplama gerektiren YSA tabanlı bir yöntem ile kontrol dışı durumlara neden olan değişkenleri tanımlama modeli geliştirmişlerdir. Sonuçlar, önerilen yaklaşımın başarılı ve kolay uygulanabilir olduğunu göstermiştir [42].

Song ve arkadaşları (2017) kontrol dışı durumları yorumlamak için örnek tabanlı naif Bayes yöntemi önerilmiştir. Yapılan performans karşılaştırmaları sonucunda geliştirilen yöntemin diğer istatistiksel tekniklerden daha iyi performans gösterdiğini belirtilmiştir [43].

Shao ve Lin (2019) tarafından yapılan çalışmada varyans kayması olan çok değişkenli bir süreçteki kontrol dışı durumların nedenlerini belirlemek için zaman gecikmeli yapay sinir ağı temelli bir sınıflandırma modeli geliştirilmiştir. Bu modelin performansı yapay sinir ağı, destek vektör makinesi ve çok değişkenli adaptif regresyon sınıflandırıcısı ile karşılaştırılmıştır. Sonuç olarak geliştirilen modelin daha başarılı olduğu belirtilmiştir [44].

2.2.2. Topluluk algoritmalarıyla kontrol dışı durumlara neden olan değişkenlerin tespit edilmesi ile ilgili çalışmalar

Bu kısımda temel algoritmaların farklı şekillerde birleştirilerek kullanıldığı çalışmalar incelenmiştir.

Karar ağacı sınıflandırma algoritmasının sıralı şekilde birleştirilmesiyle gerçekleştirilen, Guh ve Shiue 2008 tarafından yapılan çalışmada, çok değişkenli kontrol diyagramlarındaki ortalama kaymaları saptamak için basit ve etkili bir model önerilmiştir. Deneysel sonuçlar önerilen modelin öğrenme hızının, YSA temelli bir modelden çok daha hızlı olduğunu göstermektedir [45].

Yu ve Xi (2009) tarafından yapılan çalışmada, YSA için topluluk yaklaşımı kullanılarak kontrol dışı durumların izlenmesi ve teşhisi için öğrenmeye dayalı bir model geliştirilmiştir. Simülasyon ile her değişken için ortalama 5 farklı kayma büyüklüğüne göre veri üretilen çalışmanın sonuçları, önerilen modelin ortalama çalışma uzunluğu (ARL) cinsinden tekli YSA kullanımından daha iyi bir performans gösterdiğini ve kontrol dışı durumların kaynaklarını daha doğru bir şekilde sınıflandırabileceğini göstermektedir [46].

Alfaro ve diğerleri (2009) çalışmalarında, topluluk sınıflandırma ağaçları ile kontrol dışı duruma neden olan değişkenlerin belirlenmesini gerçekleştirmişlerdir. Bu çalışmada topluluk ağaçlarının sınıflandırma doğruluğu açısından çok güçlü bir araç olduğu kanıtlanmıştır [47].

Du ve Xi (2012) tarafından yapılan çalışmada çok değişkenli istatistiksel kontrol diyagramları kullanarak proses ortalama kaymalarının nedenlerini çok sınıflı topluluk DVM algoritması ile sınıflandırmışlardır. Topluluk yöntemi olarak torbalama kullanılmıştır. Önerilen yaklaşımın performansı doğruluk kriterine göre değerlendirilmiştir ve gerçek bir uygulama ile yaklaşımın sınıflandırmada etkili bir performans gösterdiği ispatlanmıştır [48].

Benzer şekilde kontrol dışı duruma nedenleri sınıflandırılması için torbalama topluluk DVM algoritması kullanılarak Cheng ve Lee (2012) tarafından yapılan bir diğer çalışmada geliştirilen yaklaşım geleneksel ayrıştırma yöntemini ile karşılaştırılarak performansının daha başarılı olduğu görülmektedir [49].

Proses ortalamasında meydana gelen kaymaların incelendiği bir diğer çalışmada, Yang (2015) yapay sinir ağları için bir topluluk modeli önermiştir. Ayrıca bu modelin literatürdeki diğer yaklaşımlardan daha iyi bir performans gösterdiği ve kontrol dışı durumların teşhisinde etkili bir yaklaşım olarak gösterilebileceği belirtilmiştir [50].

Karar ağaçlarının paralel şekilde birleştirilmesi ile topluluk modeli oluşturan bir diğer çalışma Jiang and Song (2017) tarafından yapılmıştır. Kontrol dışı durumların teşhisi problemini çözmek için torbalama ve karar ağacına dayalı bir topluluk yöntemi önerilen bu çalışmada topluluk öğrenme yönteminin sınıflandırma performansının daha iyi olduğu kanıtlanmıştır [51].

Asadi ve Farjami (2019) tarafından yapılan çalışmada yine karar ağaçları algoritması topluluk olarak kullanılmıştır. Karar ağaçlarının sıralı şekilde birbirine bağlandığı bu çalışmada dört sınıflandırıcı içeren bir karar ağacı yapısı ve monte carlo simülasyonu kullanılmıştır. Geliştirilen model ARL fonksiyonları, doğruluk, duyarlık ve kesinlik kriterlerine göre karşılaştırılmıştır. Sonuçlar topluluk karar ağacı yapısının performansının daha iyi olduğunu göstermiştir [52].

Makina öğrenme ile ilgili incelenen literatürdeki çalışmalar Tablo 2.1.'de özetlenmektedir. Tablo 2.1. incelendiğinde sıklıkla kullanılan kontrol diyagramı çeşidinin Hotelling T^2 olduğu, kontrol dışı durumlara neden olan değişkenlerin tespiti için son yıllarda daha çok topluluk algoritmalarına başvurulduğu, topluluk algoritmaları içerisinde de en çok algoritmaların sıralı şekilde başlanmasını sağlayan yükseltme yönteminin tercih edildiği görülmektedir

Tablo 2.1. Makine öğrenme yöntemleri ile yapılmış çok değişkenli kontrol dışı durum tespit çalışmaları.

NO	KAYNAK	DİYAGRAM TÜRÜ	ALGORTİMA YAPISI	ALGORİTMA TÜRÜ
1	Cheng ve Wang (2004)	T ²	Tekli	YSA
2	Niaki ve Abbasi (2005)	T ²	Tekli	YSA
3	Aparisi ve ark. (2006)	T ²	Tekli	YSA
4	Guh ve Shiue (2008)	MEWMA MCUSUM	Tekli	KA
5	Cheng ve Cheng (2008)	Varyans T ²	Tekli	DVM
6	Yu ve ark. (2009) [53]	MEWMA MCUSUM	Topluluk	Boosting-YSA
7	Alfaro ve ark. (2009)	T ²	Topluluk	Boosting-KA
8	Aparasi ve Sanz (2010)	Varyans	Tekli	YSA
9	Du ve ark. (2012)		Topluluk	Boosting-DVM
10	Li ve ark. (2013)	T ²	Tekli	DVM
11	Huda ve ark. (2014)	MEWMA T ²	Tekli	YSA
12	Yang (2015)	MEWMA MCUSUM	Topluluk	Boosting-YSA
13	Cheng ve Lee (2016)	T ²	Topluluk	Torbalama- DVM
14	Jiang ve Song (2017)	T ²	Topluluk	Torbalama-KA
15	Song ve arkadaşları (2017)	T ²	Tekli	NB
16	Asadi ve Fariami (2019)	T ²	Topluluk	Boosting-KA
8	Shao ve Lin (2019)	Varyans	Tekli	YSA

2.3. Tezin Diğer Çalışmalardan Farkı ve Literatüre Katkısı

Çalışmada kontrol dışı duruma neden olan değişkenler makine öğrenme algoritmaları ile mevcut örneklerden öğrenerek sınıflandırılmış ve yeni örneklerde rastlanan kontrol dışı durumlara neden olan değişkenlerin tahmin edilmesi hedeflenmiştir.

Sınıflandırma ve tahmin için önerilen model Tablo 2.1.'de görüldüğü gibi diğer çalışmalardan farklı olarak daha önce literatürde kullanılmamış olan Yığılmış Genelleme topluluk yöntemi ile geliştirilmiştir. Bu yöntem kullanılarak diğer iki (Torbalama ve Yükseltme) topluluk yöntemi birleştirilerek kontrol dışı duruma neden olan değişkenler tespit edilmiştir.

BÖLÜM 3. METODLAR

Bu bölümde çalışmada kullanılan yöntemler detaylı olarak ele alınmaktadır.

3.1. Hotelling T^2 Kontrol Diyagramı

Çok değişkenli normal dağılıma dayanan istatistiksel bir uzaklık ölçüsü olan T^2 istatistiğinin çizelgelenmesi ile oluşan [8] ve ilişkili p sayıdaki değişkeni aynı anda izlemeyi sağlayan Hotelling T^2 kontrol diyagramı 1947 yılında Hotelling tarafından geliştirilmiştir [3].

Hotelling T^2 kontrol diyagramı iki aşamadan oluşmaktadır. Birinci aşamada geçmiş veriler hesaplanan üst kontrol sınırına (ÜKS) göre değerlendirilerek sınırı aşan örnekler elenir ve kalan örnekler tekrar diyagramda incelenir. Bu işlem sınırı aşan örnek kalmayana kadar kalan örneklerle tekrarlanır. Kalan örnek sayısı referans veri kümesini oluşturur ve diyagram çiziminin ikinci aşaması olan yeni örneklerin değerlendirildiği adımda ÜKS bu referans sayıya göre hesaplanır. Sınırların hesapları diyagramın aşamalarına göre değiştiği gibi örnekleme alınmasına göre de değişiklik gösterir. Diyagramın hangi amaçla kullanıldığına bağlı olarak Hotelling' in T^2 istatistiği için kullanılacak kontrol sınırlarının seçilmesi önemlidir [3].

Hotelling T^2 , alt grup örnekleme verileri ya da tek gözlem verileri için olmak üzere iki farklı şekilde oluşturulabilir. Çalışmada tek gözlemler yani alt grup sayısı bir olarak kullanılacağından dolayı bu durumda izlenecek adımlar hakkında bilgi verilecektir.

İlk olarak diyagramda işaretlenecek Hotelling T^2 istatistiğini ifade etmek gerekmektedir. Süreçten 1 birimlik örnek ($n=1$) alındığında bu istatistik Eşitlik 3.1'deki gibi oluşturulur.

$$T^2 = (X - \bar{X})'S^{-1}(X - \bar{X}) \quad (3.1)$$

Burada, her biri $n = 1$ olan m örneğin mevcut olduğunu ve p 'nin her örnekte gözlemlenen kalite değişkenlerinin (X) sayısı olduğu varsayılır.

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix} \quad (3.2)$$

$$S = \begin{bmatrix} \bar{s}_1^2 & \bar{s}_{12} & \bar{s}_{13} & \dots & \bar{s}_{1p} \\ & \bar{s}_2^2 & \bar{s}_{23} & \dots & \bar{s}_{2p} \\ & & \bar{s}_3^2 & & \vdots \\ & & & \ddots & \\ & & & & \bar{s}_p^2 \end{bmatrix} \quad (3.3)$$

\bar{X} ve S , bu gözlemlerin sırasıyla örnek ortalama vektörü ve kovaryans matrisini ifade eder ve Eşitlik 3.2 ve Eşitlik 3.3'de gösterildiği gibidir. Örnek büyüklüğü $n = 1$ olduğunda, diyagramın birinci aşamasında kullanılacak sınırların beta dağılımına dayanması gerekmektedir. Bu aynı zamanda, birinci aşamanın sınırları olarak tanımlanır [3].

Üst kontrol sınırı (ÜKS) ve alt kontrol sınırı (AKS) Eşitlik 3.4 ve Eşitlik 3.5'e göre hesaplanmalıdır. Çok değişkenli kontrol diyagramında alt kontrol sınırı sıfır çizgisi olarak alınmaktadır.

$$\text{ÜKS} = \frac{(m-1)^2}{m} \beta_{\alpha, p/2, (m-p-1)/2} \quad (3.4)$$

$$\text{AKS} = 0 \quad (3.5)$$

Burada m örnek sayısı olmak üzere, $\beta_{\alpha, p/2, (m-p-1)/2}$, $p/2$ ve $(m - p - 1)$ parametreleriyle beta dağılımının üst α yüzde noktasını ifade eder [3].

Hotelling T^2 istatistiğinin ikinci aşama için kontrol sınırları ise Eşitlik 3.6 ve Eşitlik 3.7'deki gibidir.

$$\text{ÜKS} = \frac{p(m+1)(m-1)}{m^2 - mp} \quad (3.6)$$

$$\text{AKS} = 0 \quad (3.7)$$

3.1.1. Hotelling T^2 kontrol diyagramı için varsayımlar

T^2 istatistiğini bir kontrol diyagramı istatistiği olarak kullanabilmek için 4 temel varsayım sağlanmalıdır [54] Bu varsayımlar:

1. Çok değişkenli normal dağılıma uygunluk
2. Doğrusallık
3. Otokorelasyon
4. Varyans-Kovaryans Eşitliği'dir.

3.1.1.1. Çok değişkenli normal dağılıma uygunluk varsayımı

Çok değişkenli süreçlerde tek değişkenli süreçlerin uzantısı olan istatistikler kullanılmalıdır. Çünkü değişkenler tek olarak istatistiksel gereklilikleri sağlasa da iki ve daha çok değişkenin arasındaki etkileşimlerden dolayı istatistiksel farklılıklar oluşabilmektedir. Buna benzer şekilde; Hotelling T^2 diyagramının önemli varsayımlarından biri olan normal dağılıma uygunluk konusunda da tüm değişkenler tek değişkenli normal bir dağılıma sahip olsa da, değişkenlere ait vektör çok değişkenli normal dağılıma uygun olmayabilir [55]. Bu nedenle çok değişkenli kontrol diyagramları kullanılırken, çok değişkenli normal dağılım testi yapmak gerekmektedir.

Çok değişkenli normallik varsayımı, tek değişkenli normallik varsayımıyla yakından ilişkili olsa da çok daha karmaşıktır [56]. Her bir boyutun ayrı ayrı analiz edilmesi gerekliliğinin ötesinde bir zorluk teşkil etmektedir. Bu zorluğun bir nedeni, hipotezin

bazı olasılık ifadelerinin eşzamanlı değerlendirilmesinden kaynaklanmaktadır. İkinci bir nedeni ise, p boyutları arasındaki iç korelasyon yapısını hesaba katmaktır ve bu durum genel önem seviyelerini etkilemektedir [7].

Çok değişkenli normal dağılım olasılık yoğunluk fonksiyonu Eşitlik 3.8'de gösterildiği gibidir [2].

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)} \quad (3.8)$$

Burada $-\infty < x_j < \infty, j = 1, 2, \dots, p$

Çok değişkenli normal dağılım için en bilinen testler Mardia, Henze-Zirkler ve Royston testidir. Bu testler ki-kare Q-Q, perspektif ve kontur grafiklerini içeren grafiksel yaklaşımları kapsamaktadır [57].

Eşitlik 3.9'da görülen Henze-Zirkler testi, Mardia'nın çarpıklık ve basıklık testleri gibi diğer test prosedürleriyle tamamlanan, çok değişkenli normalliğin bir ölçüm testidir [58].

$$T_{n,\beta} = \frac{1}{n} \sum_{k=1}^n \sum_{j=1}^n \left[\exp\left(-\frac{\beta^2}{2} \|Y_j - Y_k\|^2\right) \right] - 2(1 + \beta^2)^{-\frac{p}{2}} \quad (3.9)$$

$$\sum_{j=1}^n \left[\exp\left(-\frac{\beta^2}{2(1+\beta^2)} \|Y_j\|^2\right) \right] + n(1 + 2\beta^2)^{-p/2}$$

Burada $T_{n,\beta}$ test istatistiğini göstermektedir. Ayrıca formülün içindeki hesaplamalar Eşitlik 3.10, 3.11, 3.12 ve 3.13'de gösterildiği gibidir.

$$\beta = \frac{1}{\sqrt{2}} \left(\frac{2p+1}{4} \right)^{1/(p+4)} n^{1/(p+4)} \quad (3.10)$$

$$\|Y_j - Y_k\|^2 = (X_j - X_k)' S^{-1} (X_j - X_k) \quad (3.11)$$

$$\|Y_j\|^2 = (X_j - \bar{X}_n)' S^{-1} (X_j - \bar{X}_n) \quad (3.12)$$

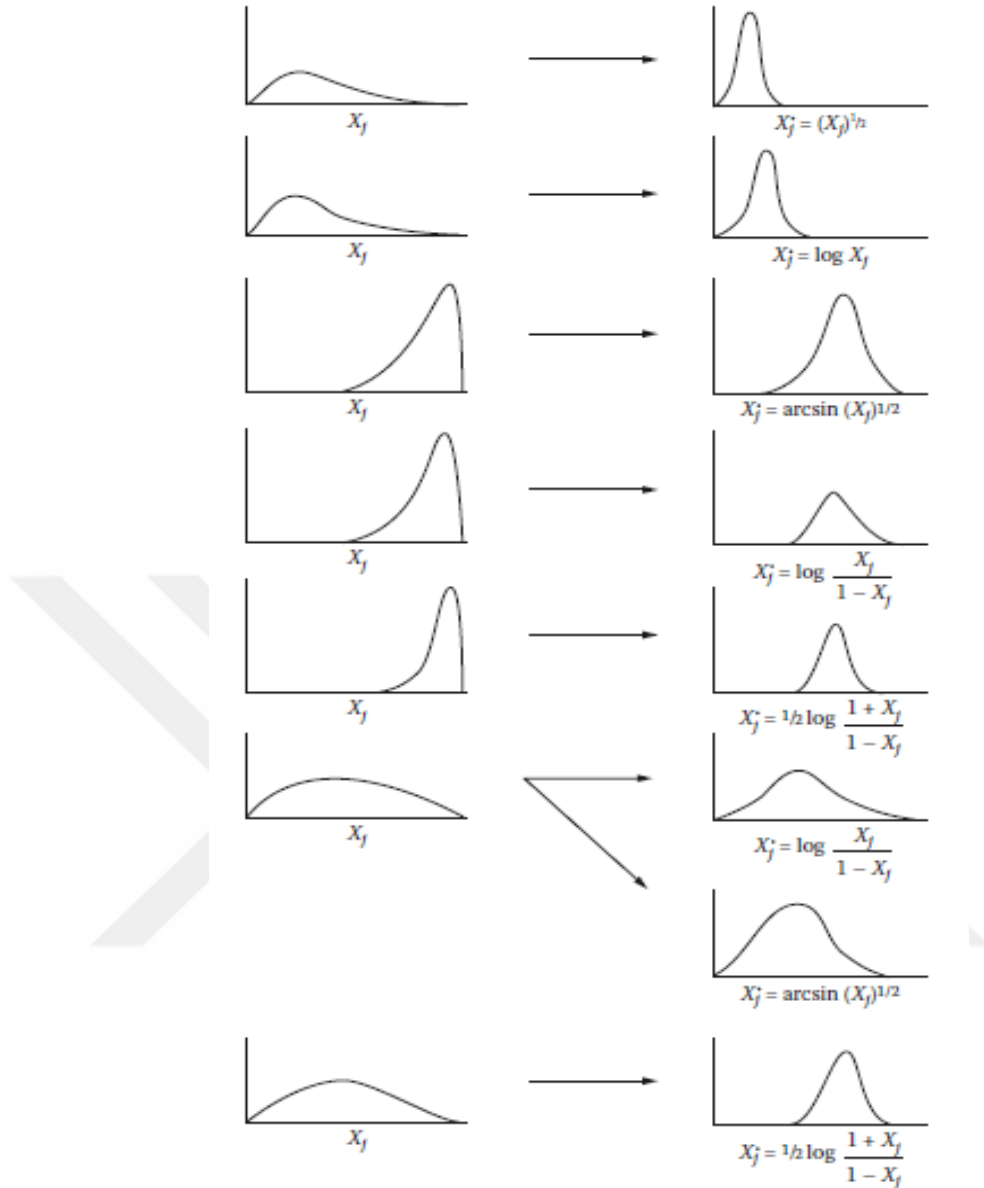
$$S = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)(X_j - \bar{X}_n)' \quad (3.13)$$

Burada \bar{X}_n , örnek ortalama vektörünü, S, kovaryans matrisini ve p değişken sayısını ifade etmektedir.

$\|Y_j\|^2$; j.örneğin merkeze Mahalanobis uzaklığını, $\|Y_j - Y_k\|^2$ ise; j. ve k.örnekler arasındaki Mahalanobis uzaklığı ifade etmektedir.

Yapılan testler sonunda veri setinin dağılımı çok değişkenli normal dağılıma uygun çıkmazsa, değişkenlerin ayrı ayrı olarak tek değişkenli normal dağılıma uygunluğu kontrol edilmeli ve uygun olmayan değişkenler bulunarak normalleştirilmeli ve testler tekrarlanmalıdır.

Normalleştirme için Şekil 3.1.'deki gibi uygun bir fonksiyon seçilerek dönüşümler yapılmalıdır.



Şekil 3.1. Normal dağılım dönüşümleri [54].

3.1.1.2. Doğrusallık

Çok değişkenli istatistiklerde, analizdeki değişkenlerin doğrusal bir şekilde birbirleriyle ilişkili olması gerekmektedir. Bu varsayımına dayanarak, iki değişken arasında gözlenen doğrusal ilişkinin derecesini değerlendiren Pearson korelasyon katsayısının ya da bir varyantını hesaplanması gerekmektedir. Ayrıca bu varsayımı test etmenin kolay bir yolu da serpilme diyagramını kullanmaktır [56]. Pearson korelasyon katsayısı sadece doğrusal bir ilişki varken hesaplanabilmektedir aksi durumlar göz

önünde bulundurularak serpilme grafiğini de mutlaka kontrol etmek gerekmektedir [59].

X ve Y değişkenler olmak üzere aralarındaki ilişkinin göstergesi olan, Pearson korelasyon Eşitlik 3.14'deki gibi hesaplanır;

$$r = \frac{N \sum XY - \sum X \sum Y}{\sqrt{[N \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}} \quad (3.14)$$

Eğer doğrusal olmayan bir ilişki tespit edilirse, analizleri uygulayabilmek için uygun dönüşüm formülleri kullanılarak doğrusallık varsayımını elde etmeye çalışmak gerekmektedir. Hair (2009)a göre; veri dönüşümüne bir alternatif, ilişkinin doğrusal olmayan kısmını temsil etmek için yeni değişkenler oluşturmak veya değişkenleri analizden çıkartmak olabilir [59].

3.1.1.3. Otokorelasyon olmaması

Otokorelasyon, zaman veya mekanda sıralı gözlemler arasındaki iç korelasyondur [60] ve gözlemlere ait hata paylarının birbirine bağlı olmaması durumudur [61]. Yani bir bağımlılık ölçüsüdür ve durağan süreçleri, çok çeşitli bağımlılık yapıları sergileyen kullanışlı bir modeller ailesi olarak tanımlamaktadır [62]. Otokorelasyon kesit verilerinden ziyade zaman dizilerinde görülmektedir [61].

t ve $t + k$ iki zaman periyodu olmak üzere, t anındaki y_t ile $t + k$ anındaki y_{t+k} değeri arasındaki k değeri kadar sonra olmanın (gecikmenin) kovaryans değeri Eşitlik 3.15'deki gibi hesaplanmaktadır [63].

$$\gamma_k = Cov(y_t, y_{t+k}) = E[(y_t - \mu)(y_{t+k} - \mu)] \quad (3.15)$$

Çalışmada otokorelasyonun olup olmadığını belirlemek için Box Ljung testi kullanılmıştır. Bu istatistik, her bir gecikmeli oto korelasyona bakmak yerine, bir dizi gecikmeli oto korelasyonu incelemektedir [64]. Ljung ve Box (1978) tarafından

geliştirilen ve uygunluk istatistiklerinin iyiliği olarak da adlandırılan prosedürünün değiştirilmiş şekli olan istatistik Eşitlik 3.16'daki gibi hesaplanmaktadır [63].

$$Q_{LB} = T(T + 2) \sum_{k=1}^K \left(\frac{1}{T-k} \right) r_k^2 \quad (3.16)$$

Burada Q_{LB} T gözlemleri için test istatistiğidir ve r de basit oto korelasyon katsayısıdır.

Otokorelasyon mevcutsa tahmin edilen parametrelerin varyansı aşağı doğru bir eğilim gösterecektir demektir böyle olduğunda t ve F testlerinin kullanılması sağlıklı olmamaktadır [61]. Bu durumda yapılabilecek dört seçenek vardır [65]:

1. Modelde bazı önemli değişkenlerin hariç tutulması sonucunda model yanlış tanımlanmış olabilir bu araştırılmalıdır.
2. Eğer saf otokorelasyon ise, uygun dönüşümü kullanılabilir. Bir tür genelleştirilmiş en küçük kareler (GEKK) yöntemi kullanmak gerekebilir.
3. Büyük örneklerde, oto korelasyon için düzeltilmiş sıradan en küçük kareler (SEKK) tahmincilerinin standart hatalarını elde etmek için, heteroseladastisite tutarlı standart hatalar yönteminin bir uzantısı olan Newey-West yöntemini kullanılabilir.
4. Diğer bazı durumlarda da SEKK yöntemi kullanılmaya devam edilmelidir.

3.1.1.4. Varyans kovaryans eşitliği (homojenliği) varsayımı

Çok değişkenli kontrol diyagramlarında değişkenler arası ilişkiler önemli olduğu için varyans kovaryans matrisi incelenmelidir. Değişkenlerdeki değişim varyans, değişkenler arasındaki değişim ise kovaryans ile ölçülmektedir [66]. Varyans kovaryans matrisinin homojen olması varsayımı çok değişkenli tekniklerin doğru uygulanması için çok önemlidir [59].

Çok değişkenli istatistiklerin homojenliği için kullanılan test, aynı zamanda iki değişkenli korelasyonların istatistiksel olarak güçlü bir testi olan Box's M' 'dir. Box's

M testinin sonuçları kabul edilirse, varyansların homojenliği varsayımı reddedilmez bu da varsayımın sağlanması demektir. Box's M, varyans kovaryans eşitliğini test etmenin yanında çok değişkenli normallik varsayımına karşı da duyarlıdır [67].

Box's M istatistiği Eşitlik 3.17'de gösterilmektedir [68].

$$W = v_e \log|S| - \sum_{i=1}^k v_i \log|S_i| \quad (3.17)$$

Burada S_i i. anakütle için kovaryans matrisinin tahminleyicisi ve n_i örnek büyüklüğünü ifade etmektedir.

Ayrıca $v_e = n - k$ ' dir.

Kovaryans, Eşitlik 3.18'deki gibi hesaplanmaktadır.

$$S = \sum_{i=1}^k v_i S_i / (n - k) \quad (3.18)$$

Burada k bağımsız popülasyonu ifade etmektedir. $n = \sum_{i=1}^k n_i$ ve $v_i = n_i - 1$ dir.

Formüllerden görüleceği üzere bu varsayım sadece $n > 1$ olması durumunda yani süreçte alt örnekler olması durumunda geçerlidir, $n=1$ olması durumunda bu varsayım geçersiz olacaktır. Çalışmada $n=1$ olarak kullanıldığından dolayı bu varsayımı kontrol etmeye gerek yoktur [69].

3.2. Mason Young Tracy (MYT) Ayrıştırma Yöntemi

Bu yöntem 1990'lı yıllarda Mason, Young ve Tracy tarafından Hotelling T^2 istatistiğini koşullu ve koşulsuz olarak 2 ortogonal parçaya ayırmak için geliştirilmiştir [11]. Bu sayede kontrol dışı duruma neden olan değişkenlerin tespit edilmesi amaçlanmaktadır.

MYT yöntemi uygulama adımları aşağıdaki gibidir [70]:

1. Koşullu ve koşulsuz terimlerin tanımlanması ve hesaplanması
2. Olası MYT ayrıştırmalarının gösterilmesi
3. Koşullu ve koşulsuz terimlerin eşik değerlerinin hesaplanması ve yorumlanması

3.2.1. Koşullu ve koşulsuz terimlerin tanımlanması ve hesaplanması

Tek değişkenin ele alındığı hesaplamalar koşulsuz terimleri, birden fazla değişkenin etkileşimli olarak hesaplandığı parçalar ise koşullu terimleri ifade etmektedir.

MYT yöntemi içerisindeki ayrıştırmalar Hotelling T^2 kontrol diyagramı başlığı altında Eşitlik 3.1'de $T^2 = (X - \bar{X})'S^{-1}(X - \bar{X})$ ile tanımlanan T^2 istatistiğini esas almaktadır.

Burada $X = (x_1, x_2, \dots, x_p)$ gözlem, \bar{X} ortalama ve S kovaryans matrisi ifade etmektedir.

T^2 istatistiği Eşitlik 3.19'da görüldüğü gibi koşullu ve koşulsuz terimlerin birleşmesiyle oluşmaktadır.

$$T^2 = T_{p-1}^2 + T_{p,1,\dots,p-1}^2 \quad (3.19)$$

Burada Eşitlik 3.20'de gösterilen parça koşulsuz terimleri ifade etmektedir.

$$T_{p-1}^2 = \left(X_i^{(p-1)} - \bar{X}^{(p-1)} \right)' S_{XX}^{-1} \left(X_i^{(p-1)} - \bar{X}^{(p-1)} \right) \quad (3.20)$$

Burada, $\bar{X}^{(p-1)}$; ilk $(p - 1)$ değişkenin n adet çok değişkenli gözlem değerinin ortalama vektörüdür. S_{XX} ise; S 'nin $(p - 1) * (p - 1)$ temel alt matrisidir.

Eşitlik 3.21'de gösterilen parça ise koşullu terimleri ifade etmektedir.

$$T_{p,1,\dots,p-1}^2 = \frac{X_{ip} - \bar{X}_{p,1,\dots,p-1}}{s_{p,1,\dots,p-1}^2} \quad (3.21)$$

Burada; $\bar{X}_{p,1,\dots,p-1} = \bar{X}_p + b'_p (X_i^{(p-1)} - \bar{X}^{(p-1)})$, \bar{X}_p ise p. değişkenin n adet gözlem değerinin örnek ortalamasıdır.

$b_p = S_{XX}^{-1} s_{xX}$; ilk $p - 1$ değişkende p. değişkeninin regresyon katsayılarını tahmin eden boyutlu vektördür.

$$s_{p,1,\dots,p-1}^2 = s_x^2 - s'_{xX} S_{XX}^{-1} s_{xX} \text{ ve } S = \begin{bmatrix} S_{XX} & s_{xX} \\ s'_{xX} & s_x^2 \end{bmatrix} \text{ dir.}$$

Burada, s_{xX} ;değişkenler arasındaki kovaryans vektörüdür, s_x^2 ise; p değişkeninin varyansıdır.

3.2.2. Olası MYT ayrıştırmalarının gösterilmesi

T^2 istatistiğinin koşulsuz terimi Eşitlik 3.21'den de anlaşılacağı gibi tek başına bir T^2 istatistiğini oluşturur. Bu nedenle o terimde kendi içerisinde koşullu ve koşulsuz terimler olacak şekilde ayrıştırılabilir. Ayrıştırma işlemleri gerçekleştirildiğinde Eşitlik 3.22'deki denklem oluşmaktadır.

$$T^2 = T_1^2 + T_{2,1}^2 + T_{3,2,1}^2 + T_{4,1,2,3}^2 + \dots + T_{p,1,\dots,p-1}^2 = T_1^2 + \sum_{j=1}^{p-1} T_{j+1,1,\dots,j}^2 \quad (3.22)$$

3.2.3. Koşullu ve koşulsuz terimlerin eşik değerlerinin hesaplanması ve yorumlanması

Kontrol dışı duruma neden olan değişken (ler) i tespit edebilmek için koşullu ve koşulsuz terimleri tanımlanmış olan eşik değerleriyle karşılaştırarak yorumlamak gerekmektedir. Koşullu ve koşulsuz terimler için tanımlanmış olan eşik değerleri (ED) Eşitlik 3.23 ve 3.24'de gösterilmektedir.

$$ED_{koşullu} = \left(\frac{(n+1)(n-1)}{n(n-p-1)} \right) F_{\alpha,1,n-p-1} \quad (3.23)$$

$$ED_{koşulsuz} = \left(\frac{(n+1)}{n} \right) F_{\alpha,1,n-1} \quad (3.24)$$

Burada k koşullanan değişken sayısı ve n örnek sayısı olmak üzere F dağılımı α düzeyinde incelenmektedir.

Koşullu parça eşik değerini aştığında bu parçada incelenen değişken, koşulsuz parça eşik değerini aştığında ise bu parça ile ilişkili olan değişkenlerin duruma neden olduğu düşünülmektedir.

T^2 ayrıştırmasının $p!$ tane bölünme şekli mevcuttur. Bölünmeye p değişkenlerinden herhangi birini seçerek başlanır. $p - 1$ 'in kalan değişkenleri ilk seçilen karakteristik koşullara göre koşullandırılır. Daha sonra, seçilen ilk iki özellik için şartlandırmak üzere kalan $p - 2$ değişkenlerinden herhangi biri seçilir. Bu şekilde devam etmek, genel istatistiği oluşturan terimlerin farklı sıralarını verir. Bu sıralama toplam değeri değiştirmez, sıralama ve bunların toplam değeri her şekilde birbirine eşittir. Örneğin 3 değişken ($p=3$) için koşullu ve koşulsuz parçalarla oluşturulan 6 farklı ayrıştırma aşağıdaki gibidir:

$$T^2 = T_1^2 + T_{2.1}^2 + T_{3.1,2}^2$$

$$T^2 = T_1^2 + T_{3.1}^2 + T_{2.1,3}^2$$

$$T^2 = T_2^2 + T_{3.2}^2 + T_{1.2,3}^2$$

$$T^2 = T_2^2 + T_{1.2}^2 + T_{3.1,2}^2$$

$$T^2 = T_3^2 + T_{1.3}^2 + T_{2.1,3}^2$$

$$T^2 = T_3^2 + T_{2.3}^2 + T_{1.2,3}^2$$

$p=3$ için eşitlikler ile koşullu ve koşulsuz parçalar yukarıda görülmektedir.

Koşullu ve koşulsuz parçalar ile bunların eşit değerleri kontrol dışı durum oluşturduğunda uygulanması gereken 5 adımdan oluşan sıralı bir hesaplama şeması aşağıdaki gibi özetlenebilir [24].

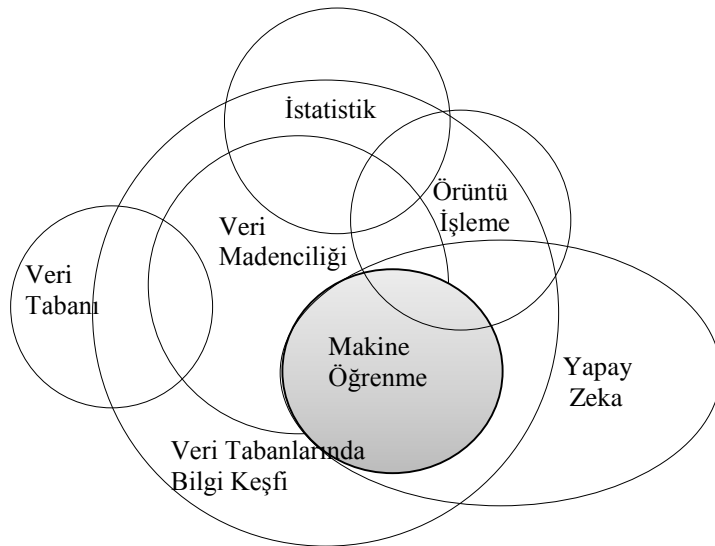
1. X vektörünün her bileşeni için ayrı T^2 istatistikleri hesaplanır. Gözlemleri önemli T_i^2 'yi üreten değişkenleri çıkartılır. Bu değişkenler üzerine yapılan gözlemler bireysel kontrolden çıkmış demektir ve diğer değişkenlerle nasıl ilişkili olduklarını kontrol etmek gerekli değildir. Kayda değer değişkenler çıkarıldığında geriye daha az sayıda değişken kalır ve kalan k adet değişkeninin alt vektörü kontrol edilir. Eğer kontrol dışı durum kalmamışsa, sorunun kaynağını tespit edilmiş demektir.
2. Değişkenler arasındaki korelasyon yapısı incelenerek çok zayıf bir korelasyona sahip değişken çıkartılır ve bu kategoriye giren bir değişkenin katkısı T_i^2 bileşeni tarafından ölçülür. Bu adım isteğe bağlı uygulanmaktadır ancak değişken sayısı çok büyük olduğu durumlarda kullanımı faydalı olacaktır.
3. Eğer silinmemiş k değişkenlerinin alt vektöründe bir kontrol dışı durum kalırsa, tüm $T_{i,j}^2$ terimleri hesaplanır ve tüm değişken çiftleri çalışmadan çıkartılır. (x_i, x_j) , önemli bir $T_{i,j}^2$ terimine sahipse, iki değişkenli ilişkide bir şey yanlış demektir. Bu durumda, dikkate alınan değişken kümesi daha da azaltacaktır. Çıkarılan tüm değişkenleri kontrol dışı durum nedeni için incelemek gereklidir. Kalan alt vektör için T^2 değeri hesaplanır. Kontrol dışı durum yoksa sorunun kaynağının iki değişkenli ilişkilerle ilgili olduğu söylenebilir.
4. Kalan değişkenlerin alt vektörü hala bir kontrol dışı durum içeriyorsa, tüm $T_{i,j,k}^2$ terimleri hesaplanır ve önemli sonuçlar gösteren üç değişkenli terimler çıkartarak kalan alt vektörde kontrol dışı durum olup olmadığını kontrol edilir.
5. Azaltılmış sette hiçbir değişken kalmayana kadar bu yüksek dereceli terimler hesaplamaya devam edilir. Tüm tek terimlerin hesaplanmasının yoğun işlem gerekliliğine neden olması bu yöntemin önemli bir olumsuz özelliğidir.

3.3. Makine Öğrenme Algoritmaları

Gerek üretim gerekse hizmet sektöründe süreçleri kolaylaştırabilmek için insanlarla aynı zekâ seviyesine sahip makineler geliştirmek için çalışmalar yoğun şekilde yapılmaktadır ve bunların günlük hayatımızda da birçok alanda kullanımı gün geçtikçe yaygınlaşmaktadır. Makinelerin zeki olmasını sağlamak için bazı bilgisayar programları ve algoritmalar kullanılmalıdır. Bu amaçla geliştirilen makine öğrenme, sistemlerin doğrudan örnekler, veriler ve deneyimlerden öğrenmelerini sağlayan teknolojidir [71] ve makinelerin tahminler yapmasını, kümelemeyi gerçekleştirmesini, ilişkilendirme kurallarını çıkarmasını veya verilen bir veri setinden kararlar almasını sağlamaktadır [72].

Kullanım alanları öneri sistemleri, kişisel güvenlik, sağlık hizmetleri, akıllı kart ve sistemler, veri güvenliği, kredi skorlama vb gibi özetlenebilir [73].

Makine öğrenme birçok farklı disiplin tarafından kullanılmaktadır ve çalışma prensipleri olarak bunlarla ile iç içe geçmiştir. Bu nedenle zaman zaman bu çalışma konuları birbirine karıştırılmaktadır.



Şekil 3.2. Makine öğrenme ile diğer alanların ilişkisi [72].

Şekil 3.2.'de görüldüğü gibi veri madenciliği, örüntü işleme, veri tabanlarında bilgi keşfi, yapay zeka gibi çalışma alanları ile birlikte kullanılırken analiz yöntemlerinden

en eski ve en temel olan istatistik ile asla kesişmez ve iki bilim dalı tamamen ayrı çalışmaktadır. Ayrıca veri tabanı ve veri işleme de makine öğrenme için kullanılacak veri setinin oluşturulması için yardımcı olsa da aralarında direk bir bağlantı yoktur.

Makine öğrenme uygulamaları temel olarak 5 adımda gerçekleşmektedir.

Bunlar [73];

1. Veri Toplama: Gelecek durumları öğrenmek için geçmişteki verilerin ham olarak toplanması aşamasıdır. Verilerin incelenen hedeflere uygun olarak mümkün oldukça büyük sayılarda toplanması faydalı olacaktır.
2. Verilerin Hazırlanması: Makine öğrenme sisteminde başarılı olmak için verilerin uygun şekilde toplanmasının yanında, gereken unsurlardan biri verilerin düzgün olmasıdır. Bunun için verilere temizleme, ayıklama gibi ön işlemler uygulanmalıdır.
3. Modeli Eğitmek: Bu aşama veriler için bir model kurmak ve bu model içerisinde uygun makine öğrenme algoritmasının seçilerek kullanılması süreçlerini içerir. Yani eldeki verilerle uygun algoritmalar kullanılarak en iyi model oluşturulmalıdır. Model eğitilirken kullanılacak makine öğrenme yöntemlerini sonuç değerinin yapısına göre dört başlık altında toplamak mümkündür.

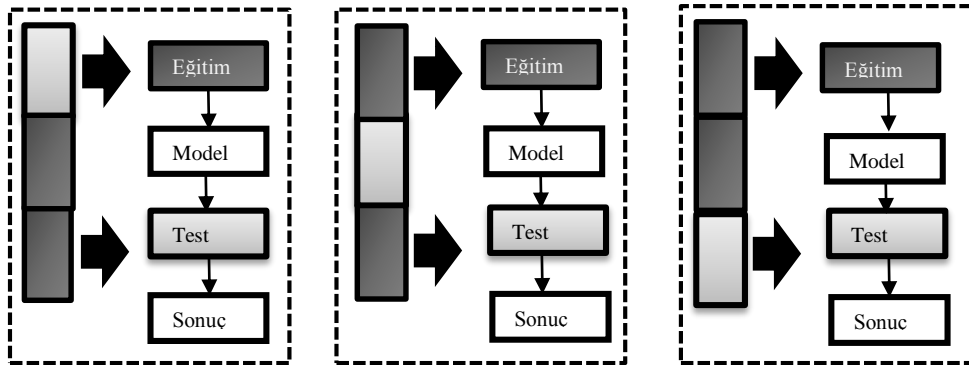
Bunlar; (1) denetimsiz (2) denetimli (3) yarı denetimli ve (4) takviye öğrenme yöntemleri olarak aşağıdaki gibi açıklanabilir [71, 72, 74]:

- a. Denetimsiz Öğrenme: Etiketsiz öğrenme olarak tanımlanabilen bu öğrenme türünde, önceden belirlenmemiş ve bağımlı bir özniteliği olmayan örnekler benzer özelliklerine göre gruplandırılmaktadır. Genellikle kümeleme amaçlı kullanılmaktadır.
- b. Denetimli Öğrenme: Bu öğrenme türü, etiketlenmiş verilerle, her veriyi çıktı sonucuna göre bir veya daha fazla sınıfa ayırarak çalışır. Eğitim verileri olarak bilinen bu etiketli verilerden sistemin nasıl yapılandırıldığını öğrenir ve bu

bilgi yeni verilerin sınıflarını tahmin etmek için kullanılır. Bu eğitim türünde model eğitilirken veri seti genellikle eğitim ve test veri seti olacak şekilde iki bölüme ayrılmaktadır.

- c. Yarı Denetimli Öğrenme: Etiketli birkaç veri noktasının yanında çok sayıda etiketsiz verinin mevcut olduğu analizlerde bunların birleştirilmesinde ve bunların her birinin öğrenilmesinde yarı denetimli öğrenme kullanılabilir.
- d. Takviye Öğrenme: Denetimsiz ve denetimli öğrenme arasında olan bu öğrenme türü, deneyimden öğrenmeye odaklanır. Bir aracının veya bilgisayar programının, çevresiyle etkileşime girmeyi, girdi almayı ve gelecekteki kazanımları en üst düzeye çıkarmak için sıralı kararlar almayı öğrenmesi gereken görevleri ele alır.

Ayrıca çalışmada eğitim yöntemi olarak çapraz doğrulama kullanılmaktadır. Bu yöntem, algoritmaların sınıflandırma performanslarını değerlendirmek için de kullanılan önemli yöntemlerden biridir. Çapraz doğrulama, verileri, biri bir modeli öğrenmek ya da eğitmek, diğeri modeli doğrulamak amacıyla kullanılan iki bölüme ayırarak öğrenme algoritmalarını değerlendirmek ve karşılaştırmak için kullanılan istatistiksel bir yöntemdir [75].



Şekil 3.3. 3 Katlamalı çapraz doğrulama yöntemi [75].

K-katlamalı çapraz doğrulamada, veri ilk önce k eşit boyutlu bölümlere ayrılır. Daha sonra seçilen bir bölüm test kümesi kalan k-1 bölüm eğitim kümesi olarak değerlendirilir. Bir sonraki aşamada farklı bir bölüm test için seçilerek kalanlar eğitim kümesini oluşturur. Seçilecek kümenin bir öncelik sırası ya da önemi

bulunmamaktadır, her bölüm eşit öneme sahiptir. Bu işlem her seferinde farklı bir alt bölüm test kümesi olacak şekilde k kez tekrarlanarak her bölüm hem test hem de eğitim için kullanılmış olur. k tekrara ait sonuçların ortalaması alınarak çapraz doğrulama işlemi tamamlanmış olur. Şekil 3.3.'de, örnek olması adına $k = 3$ alınarak işlemin nasıl gerçekleştirdiği gösterilmiştir. Veri madenciliği ve makine öğrenme uygulamalarında en yaygın olarak $k=10$ olarak kullanılmaktadır [75].

4. Modelin Değerlendirilmesi: Modelin doğruluğunun değerlendirildiği bu aşamada daha önce sisteme hiç tanımlanmamış bir test seti ile geliştirilen modelin başarısının ölçüldüğü adımdır.
5. Performansı Artırmak: Toplanan veri sayısı, verinin ön işlemden geçirilmesi, kullanılan algoritmanın ya da parametrelerinin değiştirilmesi gibi alternatiflerle modelin performansı artırılması gerekebilir.

Sınıflandırma, eğitim verilerindeki etiketli örneklerden öğrenilen denetimli bir öğrenmedir [76].

Algoritmaları, kullanım yapısına göre, tekli ve topluluk olmak üzere iki sınıfta incelemek mümkündür. Tekli algoritmalarda sadece tek bir algoritma çalıştırılmakta ve sonuçlar buna göre elde edilmektedir. Ancak topluluk algoritmalarında aynı ya da farklı türde birden fazla algoritmayı çeşitli yöntemlerle farklı şekillerde birleştirilerek topluluk algoritmaları oluşturularak kullanılmaktadır.

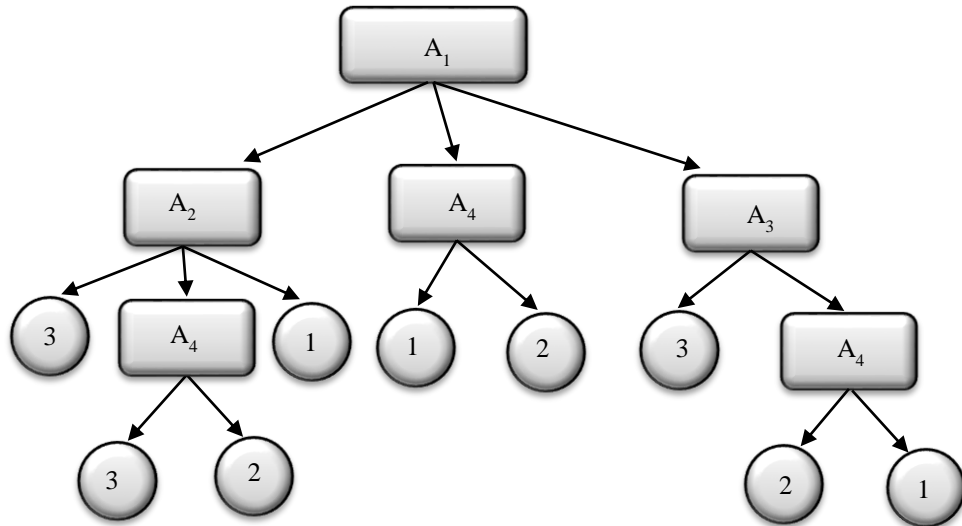
3.3.1. Makine öğrenme algoritmalarının tekli kullanılması

Sınıflandırma algoritmalarından en bilinenleri karar ağaçları, naive bayes sınıflandırıcısı, yapay sinir ağları, destek vektör makineleri, k -en yakın komşu algoritmasıdır [64].

3.3.1.1. Karar ağaçları

Karar Ağacı (KA), örnek uzayının özyinelemeli bir bölünmesi olarak ifade dileyen ve diğer algoritmalara göre en yaygın kullanılan, pratik yöntemlerden biridir ve sınıflandırma algoritması uygun teknikler ile veri setini alt bölümlere ayıran bir yapıdadır [73, 77, 78].

Karar ağacı Şekil 3.4.'de görüldüğü gibi, düğümlerden oluşan bir ağaç yapısına sahiptir. Bu düğümler amaçlarına göre kök düğüm (A_1), ara düğüm (A_2, A_3, A_4), ve yaprak düğüm (1,2,3) olarak adlandırılmaktadır [74]. Algoritmanın çalışma adımları ilk olarak kökten başlar. Ardından ara düğümünden yaprak düğümüne kadar dallanarak devam eder. Ağaçta sınıflar yaprak ile temsil edilir ve her yaprağa sadece tek bir yol gitmektedir [73, 79]. Örnekler, yol boyunca yapılan testlerin sonucuna göre, ağacın köklerinden bir yaprağa doğru sınıflandırılır ve bu sonuçlar birleştirilerek ve yaprağın sınıf tahmininin sınıf değeri olarak alınmasıyla bir kurala dönüştürülebilir [74]. Kullanıcı tarafından kolay anlaşılabilir olması için EĞER-İSE zaman kural kümeleri yeniden temsil edilebilen bu yapı [77] hem nominal hem de sayısal özellikleri içerebilir [74].



Şekil 3.4. Karar ağacı yapısı [73].

Karar ağacı algoritmalarının olumlu yönleri; Anlaşılabilir kurallar çıkartılması, Çok büyük boyuttaki verilerle çalışma yeteneğine sahip olması, Modelin yorumlanması ve anlaşılmasının gayet kolay olması, Sürekli ve kategorik verilerle çalışabilir olması, veri setinde kayıp ve/veya eksik değerler olması durumunda bile çalışabilir olması olarak söylenebilir. Sayılan olumlu yönlerinin yanında; çok sınıflı problemlerde, sürekli değişkenler mevcutken ve eğitim veri sayısı az ya da kısıtlı olduğunda algoritmanın performansını düşmesi gibi olumsuz yönleri de mevcuttur [73].

Kök düğüm özelliğinin belirlenmesinde yaygın olarak kullanılan kriterler şunlardır [74]:

- Bilgi Kazancı (BK): Bilgi kazancı, entropi ölçüsünü belirsizlik ölçüsü olarak kullanan kararsızlık temelli bir kriterdir. Entropiyi azaltmak kazanç olarak adlandırılır [80]. Entropi aşağıdaki Eşitlik 3.25 ile hesaplanmaktadır.

$$Entropy = (D_1) = - \sum_{i=1}^m p_i \log_2 p_i \quad (3.25)$$

Burada, m durum sayısını, p_i ise durumun olasılığını belirtmektedir.

A niteliğinin D sistemine göre kazancı demek olan Bilgi Kazancı ise Eşitlik 3.26'da gösterildiği gibi sistemin entropi değerinden, özniteliğin entropisinin çıkarılmasıyla hesaplanmaktadır.

$$BK(D, A) = Entropy(D) - \sum P(s) Entropy(D(s)) \quad (3.26)$$

Burada; s, A'nın tüm değerleri ve $P(s) = \frac{|D(s)|}{|D|}$ dir.

- Gini endeksi (GI): Gini endeksi, hedef niteliğin değerlerinin olasılık dağılımları arasındaki farkı ölçen gürültü temelli bir kriterdir. Veri setindeki yanlış sınıflandırma olasılığını ifade etmektedir ve Eşitlik 3.27'de gösterildiği gibi hesaplanmaktadır. Ağaç oluşturulurken en küçük GI değerine sahip olan özellik, kök düğüm olarak seçilir [73].

$$GI(A_{d_i}) = 1 - \sum_j^t p_j^2 \quad (3.27)$$

Burada i seçilen özelliğin durum sayısını, A_{d_i} seçilen özelliğin d_i durumunda iken A veri kümesinin alt kümelerini göstermektedir. j sınıf sayısı, p_j ise; seçilen özellikte d_i durumu için j sınıfının gerçekleşme olasılığıdır.

- Kazanç Oranı (KO): Kazanç oranı, bilgi kazancını Eşitlik 3.28'deki gibi normalleştirmektedir. Payda sıfır olduğunda bu oranın tanımlanamamaktadır [80].

$$KO(a_i, S) = \frac{BK(a_i, S)}{Entropy(a_i, S)} \quad (3.28)$$

3.3.1.2. Naif bayes

Bayes, öncü bir bilgi olduğunda kullanılır ve mantıksal çıkarım için olasılıklı bir yaklaşım sağlar. Örneklemden alınan değer ile öncü bilgiyi birleştirmeyi amaçlamaktadır [81]. Bayesci yaklaşım, olasılıkları doğrudan belirten öğrenme algoritmalarının temelini oluşturmaktadır ve olasılıkları açıkça belirtilemeyen diğer algoritmaları anlamak için yararlı bir bakış açısı sunmaktadır. Öğrenme problemlerine yönelik en pratik yaklaşımlar arasındadır. Bir önceki varsayım olasılığına, çeşitli verileri gözlemeleme olasılıklarına ve gözlemlenen verilere dayanarak bir hipotezin olasılığını hesaplamak için bir yol sağlamaktadır [77].

Bayes Teoremi Eşitlik 3.29'da gösterilmiştir.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \quad (3.29)$$

Burada; $P(X)$: X olayının olasılığı (bağımsız), $P(Y)$: Y olayının olasılığı (bağımsız), $P(X|Y)$: Y 'nin olduğu bilindiğinde X 'in olma olasılığı ve $P(Y|X)$: X 'in olduğu bilindiğinde Y 'nin olma olasılığıdır.

Bayes yaklaşımını temel alan Naif Bayes algoritması mevcut değişkenlerle ulaşmak istenen sonuç arasındaki ilişkiyi tahmin etmeye çalışan bir sınıflandırıcıdır [73]. Naif Bayes algoritması; girdiler arasındaki ilişkileri yok sayarak çok değişkenli bir dağılımı Eşitlik 3.30'da görüldüğü gibi, birden çok tek değişkenli dağılıma indirger [81].

$$p(x|C) = \prod_{j=1}^d p(x_j|C) \quad (3.30)$$

Bu algoritma her çıktının kaç kere görüldüğünü ve bağımsız değişkenin ile bağımlı değişken kombinasyonunun görülme sıklığını hesaplar [73].

3.3.1.3. k-En yakın komşu

En yakın komşu algoritması en benzer etiketli örneklerin sınıfını atayarak, etiketlenmemiş örnekleri sınıflandırma karakteristikleri ile tanımlanır. Bu fikrin sadeliğine rağmen, en yakın komşu yöntemleri son derece güçlüdür. Sınıflandırma için en yakın komşular yaklaşımı KNN algoritması tarafından kullanılır [82]. Komşu sayısı (k) kullanıcı tarafından belirlenmelidir. Sınıflandırmayı başarılı yapabilmek için birden büyük ve eşit oylama olmaması için tek sayı olmalıdır. Ancak, büyük k değerleri de yöntemin etkinliğini azaltır [83]. Eğer k aynı zamanda sonsuzluğa yaklaşırsa, hata oranı Bayes hata oranına yaklaşır.

En yakın komşu sınıflandırıcıları sayısal tahmin için, yani gerçek değerli bir tahmin vermek için de kullanılabilir [76]. Bu algoritma, basit ve etkili olması, kullanılacak veri ile ilgili herhangi bir varsayım gerektirmemesi ve eğitim aşamasının hızlı olması bakımından güçlüdür ancak çok miktarda bellek ihtiyacı olması, nominal özellikler ve eksik veriler mevcutsa ek işlem gerektirmesi ve sınıflandırma aşamasının yavaş olması gibi konularda zayıf noktaları olarak söylenebilir [82].

Bir örneğin en yakın komşularının yerini bulmak, bir mesafe işlevi veya iki örnek arasındaki benzerliği ölçen bir formüle ihtiyaç duyulmaktadır. Mesafeyi hesaplamının birçok farklı yolu vardır. Bunlar Öklid, Manhattan ve Minkowski Uzaklığı olarak farklılık gösterir [73].

Geleneksel olarak kullanılan mesafe metriği (sayısal özellikler için) basit Öklid mesafesi olabilir [82]. Burada p ve q karşılaştırılan iki örnek olmak üzere aralarındaki mesafe Eşitlik 3.31'deki gibi hesaplanabilir [73].

$$\text{Öklid Uzaklığı} = \sqrt{\sum_{i=1}^k (p_i - q_i)^2} \quad (3.31)$$

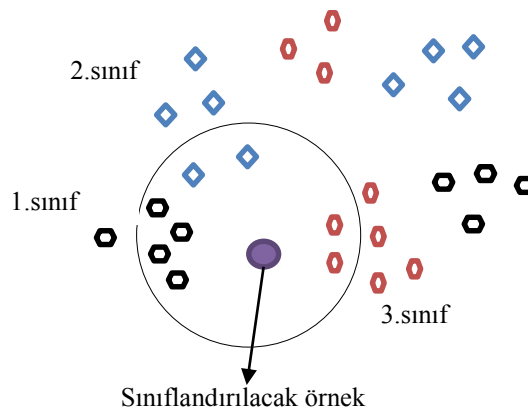
İki örnek arasındaki farkların mutlak değerlerinin toplamı olarak elde edilen Manhattan Uzaklığı, Eşitlik 3.32'deki gibi hesaplanmaktadır [73].

$$\text{Manhattan Uzaklığı} = \sqrt{\sum_{i=1}^k |p_i - q_i|} \quad (3.32)$$

İlk iki mesafenin genelleştirilmiş hali olan Minkowski Mesafesi ise Eşitlik 3.33'deki gibi hesaplanmaktadır [73].

$$\text{Minkowski Uzaklığı} = (\sum_{i=1}^k (|p_i - q_i|)^r)^{1/r} \quad (3.33)$$

Minkowski mesafesi $p=1$ olduğunda Manhattan, $p=2$ olduğunda ise Öklid mesafesini hesaplamaktadır [73].



Şekil 3.5. K-NN yapısı [83].

En yakın komşu karar sorununa bir örnek, Şekil 3.5.'de gösterilmiştir. 3 sınıflı bir problemde yeni örneğin hangi sınıfa ait olduğu aranmaktadır. $K=8$ olarak seçilmiştir.

Bu durumda yeni örneğe en yakın 8 örnek seçilerek içlerinde hangi sınıfa ait örnek sayısı en çoksa yeni örnek o sınıfa atılır denilmektedir. Burada yeni örnek 1. Sınıfa dahil edilecektir.

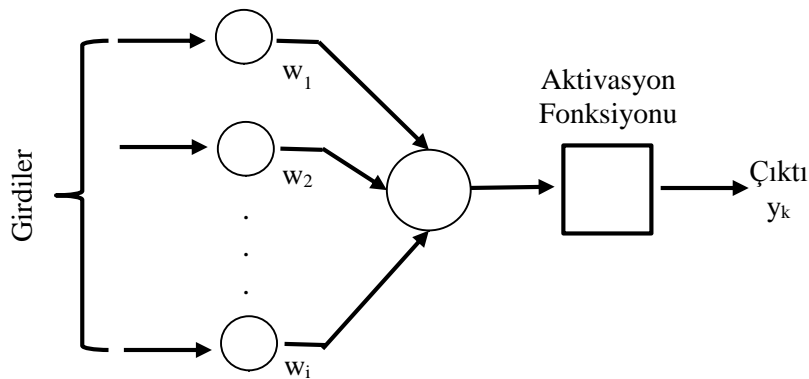
K parametresi seçildikten sonra test örneğinin hangi sınıfta olduğunu seçmek için basit ya da ağırlıklı ortalama kullanılabilir. Basit oylama, seçilen komşu örnekleri içinde sayısı en fazla olan sınıfa atama yapılarak gerçekleşir. Ağırlıklı oylama da ise etkinliklere göre ağırlık verme prensibine dayanmaktadır ve buna göre yakın komşulara uzak komşulara göre daha etkindir ve daha ağırlıklı oylanmaktadır [73]. i ve j verileri arasındaki Öklid mesafesi $d(i,j)$ nin hesaplanması Eşitlik 3.34'de gösterilmektedir.

$$d(i,j) = \frac{1}{d(i,j)'} \quad (3.34)$$

Burada $d(i,j)$ i . ve j . veriler arasındaki Öklid mesafesini göstermektedir.

3.3.1.4. Yapay sinir ağları

Yapay sinir ağları, gerçek bir insan beynine benzer çalışan paralel hesaplama programları içeren önemli bir sınıflandırma aracıdır. Mevcut örnekler ve aralarındaki ilişkilerden yola çıkarak, insanlara özgü olan öğrenme kavramını kullanarak, karar veren bilgisayar sistemleridir. Ağ yapısı, yapay sinir hücrelerinin, sahip olduğu bilgiye göre değişen, bir ağırlık değeriyle birbirine bağlanması ile oluşur [84].



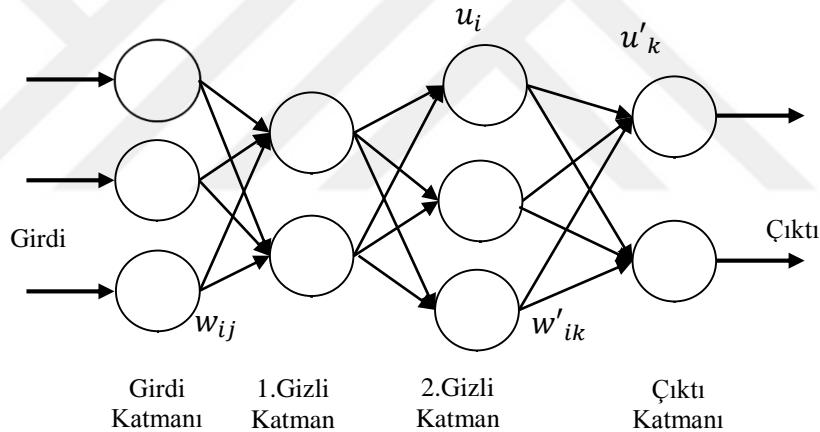
Şekil 3.6. Sinir ağı yapısı [85].

Yapay sinir ağı, Şekil 3.6.'da görülen ve fonksiyonu Eşitlik 3.35'de sunulan yapay nöronlardan oluşmaktadır [86].

$$y = f\left(\sum_i^d w_i x_i\right) \quad (3.35)$$

Burada d toplam girdi sayısını, w ağırlıkları, x ise girdileri göstermektedir.

Tek katman içeren yapay sinir ağı doğrusal olmayan ilişkileri öğrenememektedir. Bu nedenle geliştirilmiş farklı ağ mimarileri mevcuttur. Girdi ile çıktı arasındaki ilişki doğrusal olmadığında Şekil 3.7.'de gösterilen çok katmanlı algılayıcı (ÇKA) kullanılmalıdır [81].



Şekil 3.7. Çok katmanlı algılayıcı yapısı [87].

En sık kullanılan yapay sinir ağı modeli olan ÇKA, girdi, gizli ve çıktı olmak üzere 3 katmandan oluşur. Her katmanda da bilgi iletimini sağlamak için proses elemanları mevcuttur. Katmanlar birbirlerine proses elemanları aracılığıyla bağlanmaktadır. Bilgiler sisteme girdi katmanından eklenir bu katmanda herhangi bir bilgi işleme fonksiyonu gerçekleşmez. Bu katmandan iletilen bilgiler gizli katmanda işlenir. Gizli katman sayısı girdi/çıkta arasındaki ilişkinin durumuna ve karmaşıklığına göre değişir. Son olarak ağı ürettiği çıktının hesaplanarak dış dünyaya sunulduğu çıktı katmanı mevcuttur. Ayrıca girdi ve çıktı katmanındaki proses elemanı sayısı probleme göre belirlenirken, gizli katmanda ki eleman sayısı ise en iyi performansa ulaşmak

amaçlanarak deneme yanılma yolu ile bulunmaktadır. Bilgilerin önem derecelerini gösteren ağırlıklar ise, başlangıçta rastgele belirlenir [84].

En yaygın kullanılan YSA öğrenme tekniği, geri yayımlı sinir ağı algoritmasıdır. Bu teknik, eğitim örneklerinde hata olması durumunda kullanılabilir olması, hedef fonksiyon çıktısı ayrık değerli, gerçek değerli veya birkaç gerçek veya ayrık değerli niteliklere sahip bir vektör olabilmesi ve uzun eğitim süreleri kabul edilebilir olması gibi durumlarda kullanılabilmesi için tercih edilmektedir [77].

Bir YSA'nın en temel işlevi, girdileri alarak bir aktivasyon fonksiyonu uygulayarak çıktıya dönüştürmektir. Aktivasyon fonksiyonları arasında yaygın olarak kullanılanlar; Doğrusal Aktivasyon Fonksiyonu, Parçalı-Doğrusal Aktivasyon Fonksiyonu, İşaret Aktivasyon Fonksiyonu, Log Sigmoid Aktivasyon Fonksiyonu ve Hiperbolik Tanjant Aktivasyon Fonksiyonu olduğu söylenebilir [87].

Transfer fonksiyonları arasında literatürde en sık kullanılan fonksiyon Eşitlik 3.36. da görüldüğü gibi hesaplanan lojistik sigmoid fonksiyonudur [88].

$$y = \frac{1}{1 + e^{-(\sum_i w_i x_i + w_0)}} \quad (3.36)$$

burada, x_i , girdiler; w_i , ağırlık faktörü ve w_0 , biası (hata terimini) ifade etmektedir.

Karar ağacı algoritmasına benzer şekilde, sinir ağları da kararsızdır, çünkü eğitim verilerindeki küçük değişiklikler, hem yapısında hem de parametrelerinde büyük bir değişikliğe yol açabilir [86]. Bu kararsızlığın yanında parametre seçimi ve eğitimin sonlanacağı noktanın bulunması gibi olumsuzluklar olmasına karşın eksik bilgi olduğunda ve belirsiz ortamlarda sorunsuzca çalışarak hesap yapabilme yeteneğine sahip olmak gibi çok önemli olumlu etkileri mevcuttur [84].

3.3.1.5. Destek vektör makineleri

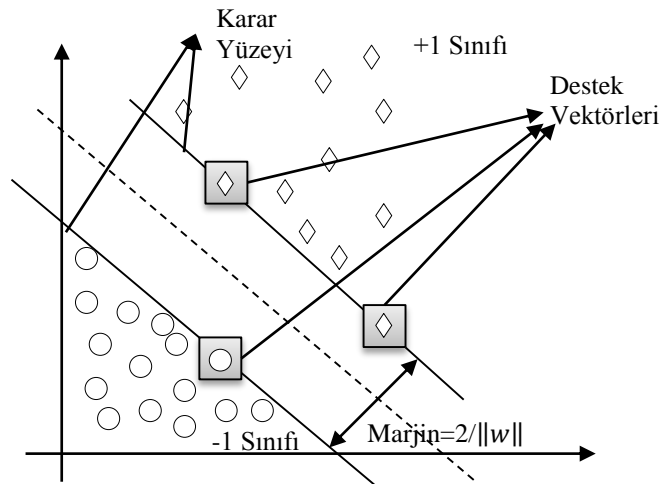
- İki sınıflı destek vektör makineleri

Vapnik ve Cores (1995) tarafından geliştirilen DVM yöntemi hem doğrusal hem de doğrusal olmayan verilerin iki gruplu sınıflandırma ve tahmin problemleri için kullanılmaktadır. Çalışma prensibi verinin boyutunu dönüştürerek karar yüzeyleri belirleyerek en uygun şekilde iki sınıfa ayırmaya dayanmaktadır. Bunu nedenle 2 sınıf birbirinden doğrular ile ayrılmaktadır. İki sınıfı ayırabilmek için birçok farklı karar yüzeyi seçilebilir bu algoritmanın amacı karar yüzeylerini bu iki sınıf arasındaki aralığı en fazla olacak şekilde çizebilmektir. En uygun karar yüzeyi H karar fonksiyon ve sınıf kararları değeri Eşitlik 3.37 ile belirlenmektedir. Burada w ağırlık vektörü, x örnek ve r verinin sınıfını ifade etmektedir. Bu aralığın büyüklüğü sınıflandırmanın doğruluğu için çok önemlidir. $r^t = +1$ ve $r^t = -1$ olmak üzere [81];

$$r^t(w^T x^t + w_0) \geq \pm 1 \quad (3.37)$$

Ayrıca $r^t(w^T x^t + w_0) \geq 0$ olmama hususu çok önemlidir çünkü bu durumda sınıflandırma yapılamaz.

2 sınıflı problem için bir örnek Şekil 3.8.'de gösterilmektedir [89].



Şekil 3.8. 2 sınıflı problem için destek vektör makinesi örneği [89].

Veriyi yüksek boyutlara dönüştürmek için kernel adı verilen fonksiyonlar kullanılır.

x ve y destek vektörleri olmak üzere, kernel fonksiyonları ve karar yüzey çeşitleri Tablo 3.1.'de gösterilmiştir [82, 89].

Tablo 3.1. Kernel fonksiyon ve sınıflandırıcı çeşitleri.

Kernel Fonksiyonu	Sınıflandırıcı türü
$K(x, y) = (x \cdot y)$	Doğrusal Kernel
$K(x, y) = \exp(-\ x - y\ ^2)$	Gaussian RBF Kernel
$K(x, y) = (1 + x \cdot y)^d$	Polinomial Kernel
$K(x, y) = \tanh(x \cdot y - \theta)$	Sigmoid Kernel

– Çok sınıflı destek vektör makineleri

Sınıf sayısı 2'den fazla olduğu durumlarda çok sınıflı destek vektör makineleri kullanılmalıdır. Bu algoritma çözüm için 3 seçenek sunmaktadır [81].

1. seçenek; $K > 2$ iken K tane iki sınıf problemi tanımlanarak her sınıfı öteki sınıflardan ayıran K ayrı ayrıştırıcı olan; $i = 1, \dots, K$ destek vektör makinesi eğitilmektedir. Burada ayrıştırıcısını eğitirken C_i sınıftan örnekler +1, C_k $k \neq 1$ sınıftan örnekler ise -1 olarak sınıflandırılır. Tüm değerler hesaplanarak en büyük olan seçilir.

2. seçenek; problemi birden çok doğrusal alt probleme bölmektir. Bunun için algoritma Eşitlik 3.37'ye benzer olarak $K(K - 1)/2$ tane ayrıştırıcı ikili sınıflandırıcı ile eğitmektir.

İlk iki seçenek de çok sınıflı bir problemi ikili problem olarak yeniden tanımlamaya dayanmaktadır.

3. seçenek ise; Eşitlik 3.38'de görülen, tüm sınıfları içeren çok sınıflı tek bir iyileştirme problemi ele almaktır.

$$\min \frac{1}{2} \sum_{i=1}^K \|w_i\|^2 + C \sum_i \sum_t \xi_i^t \quad (3.38)$$

Burada kısıtlar;

$$w_{z^t}x^t + w_{z^t_0} \geq w_i x^t + w_{i0} + 2 - \xi_i^t, \forall i \neq z^t \text{ ve } \xi_i^t \geq 0$$

şeklindedir.

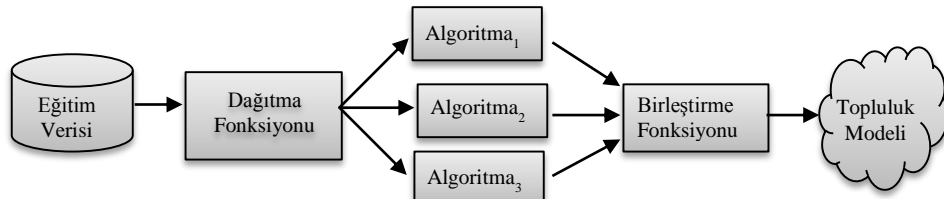
Bu seçenek çok iyi bir yaklaşım olmasına rağmen işlem yükü ve zamanı nedeniyle kullanım açısından diğer seçeneklere göre daha az tercih edilir.

3.3.2. Topluluk makine öğrenme algoritmaları

İstatistik ve yapay zeka gibi çeşitli disiplinlerden araştırmacıların kullandığı topluluk metodolojisi fikri, birden fazla modeli entegre ederek öngörücü bir model oluşturmaktır [90].

Makine öğrenme algoritmalarının tekli kullanılmasının yerine, birleştirilerek topluluk yöntemlerinin kullanılmasının tahmin doğruluğunu arttırmayı sağlayacağı düşünülmektedir. Ancak topluluk algoritmaları oluşturulurken, kullanılan temel algoritma parametrelerinin en iyi şekilde belirlenmesi ve doğru şekilde birleştirilmesi dikkat edilmesi gereken en önemli husustur. Aksi takdirde başarılı sonuçlara ulaşamamanın yanında zaman ve maliyet kayıpları oluşacaktır [81].

Bir topluluğun yani birden fazla algoritmanın genelleme yeteneği genellikle temel algoritmaların tek tek kullanılmasından daha güçlüdür. Çünkü rastlantısal tahminlerden biraz daha iyi sınıflandırma yapmayı sağlayan zayıf algoritmaları, çok kesin tahminler yapabilen güçlü algoritmalarla daha da güçlü hale getirebilirler. Bu nedenle tekli algoritmalar aynı zamanda zayıf algoritmalar olarak da adlandırılır [91].



Şekil 3.9. Topluluk algoritmalarının bileşenleri [82].

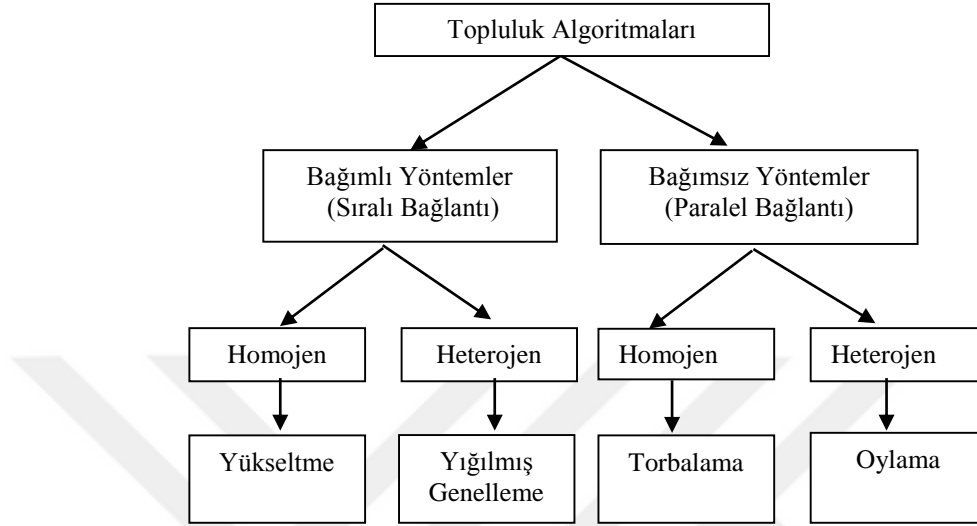
Sınıflandırma görevleri için tipik bir topluluk yöntemi, Şekil 3.9.'daki 3 temel bileşeni içermektedir. Bu bileşenler aşağıda açıklanmaktadır [74].

1. Eğitim seti: Topluluk eğitimi için kullanılan etiketli bir veri setidir. Girdi eğitim verileri bir dizi model oluşturmak için kullanılır.
2. Dağıtma Fonksiyonu: Her modelin tam eğitim veri setini mi yoksa sadece bir örneği mi alacağını belirleyen fonksiyondur. Çeşitli sınıflandırma algoritmalarını eğitmek için girdi verilerini yapay olarak değiştirerek çeşitliliği arttırabilir.
3. Birleştirme Fonksiyonu: Birleştirici, çeşitli sınıflandırıcıların sınıflandırmalarını birleştirmekten sorumludur.

Topluluk algoritmalarını oluşturmak için bağımlı ve bağımsız yöntemler mevcuttur. Bağımlı bir yöntemde, bir sınıflandırıcının çıktısı bir sonraki sınıflandırıcının yapımında kullanılır. Böylece, bir sonraki tekrarlar da öğrenmeyi yönlendirmek için önceki tekrarlar da üretilen bilgiden yararlanmak mümkündür. Bağımsız yöntemlerde ise, her bir sınıflandırıcı bağımsız olarak oluşturulur ve çıktıları birleştirilir [74]. Bağımlı yöntemlerde algoritmalar birbiriyle seri şekilde bağlanırken bağımsız yöntemlerde algoritmalar paralel olarak bağlanmaktadır. Paralel topluluk yöntemlerinin temel esası, tekli algoritmalar arasındaki bağımsızlığı kullanmaktır, çünkü sınıflandırma ve tahmin hatası bağımsız temel öğrencileri birleştirerek önemli ölçüde azaltılabilir [91].

Algoritmaların bağımlı ve bağımsız olarak birleştirilmesinin yanında aynı tekli algoritmanın kullanılması ile homojen olarak, farklı tekli algoritmaların kullanılması ile heterojen olarak oluşturulan topluluk algoritmaları mevcuttur. Topluluk

algoritmalarının birleştirilme prensipleri ile ilgili özet gösterim Şekil 3.10.'da sunulmuştur. Topluluk algoritmalarının çeşitlerinden torbalama ve yükseltme homojen algoritmalar iken, yığılmış genelleme ve oylama heterojen olanlardır [92].



Şekil 3.10. Topluluk algoritmalarının çeşitleri [91, 92].

3.3.2.1. Torbalama topluluk yöntemi

Breiman'ın torbalama algoritması en eski ve en basit, ancak etkili bir topluluk tabanlı algoritmalarıdır [93]. Torbalama, hepsi aynı sentezleme aracı ile türetilmiş birden fazla sınıflandırıcıyı birleştiren bir model üreterek sınıflandırma doğruluğunu artırmayı amaçlayan bir yöntemdir [90] ve kararlar, farklı sınıflandırıcılardan alınan çoğunluk oyu ile verilmektedir [92]. Temel öğrencilerin paralel olarak birleştirildiği torbalama yöntemi çok sınıflı olarak kullanılabilir [91].

Yüksek derecede doğrusal olmayan öğrenciler dengesiz olma eğiliminde olduklarından, yani performansları, veri örneklemindeki düzensizliklerle çok değiştiğinden, Torbalama yönteminin dengesiz temel öğrencilerinde etkili olduğu söylenebilir [91].

3.3.2.2. Yükseltme topluluk yöntemi

Zayıf algoritmaların performansını arttırarak güçlendirmeyi amaçlayan bir diğer topluluk algoritması yükseltme yöntemidir. Bu yöntemin temel amacı, başarısı düşük olarak değerlendirilen öğrenici algoritmalarının istenen başarı oranına yükseltmeye yardımcı olmaktır [82].

Algoritmaların bir önceki öğrenici algoritmanın çıktısını girdi olarak kullanmasını sağlayan ve seri şekilde bağlanan bu yöntemde, torbalamanın aksine, her sınıflandırıcı, bir önceki algoritmanın performansından etkilenir. Daha önce oluşturulmuş sınıflandırıcılar tarafından yapılan sınıflandırma hatalarına daha fazla önem vermektedir [90]. Yani torbalamadaki gibi her bir sınıflandırıcıya eşit oy vermek yerine, her bir sınıflandırıcının performansına dayalı olarak ağırlıklandırılmaktadır [82].

Bu topluluk yöntemi ikili sınıflandırma için tasarlanmıştır ve genellikle verideki gürültüden ciddi şekilde zarar görür. Ancak birçok sınıflandırma probleminde iki sınıf yerine çok sayıda sınıf mevcuttur. Bu durumda Yükseltme yönteminin çok bilinen bir türü olan AdaBoost yönteminin çok sayıda sınıf için genişletilmiş şeklinin kullanılması uygun olacaktır [91].

AdaBoost yöntemi ile, sınıflandırma oranı yüksek olabilecek birçok sınıflandırıcı birleştirilir. Böylece yanlış sınıflandırma oranı azalır. Ayrıca varyansı, zayıf temel algoritmalar tarafından üretilen varyanslardan önemli ölçüde düşük olan bir topluluk sınıflandırıcı üretilir. Bu iki ana nedenden dolayı sınıflandırma performans doğruluğunu arttırdığı düşünülmektedir [90].

Torbalama ve yükseltme yöntemleriyle ilgili yapılan birçok çalışmada yükseltme yönteminin torbalama yönteminden daha başarılı olduğu belirtilmiştir [73].

3.3.2.3. Oylama

Farklı algoritmaları birleştirme konusunda uygulaması en kolay olan yöntem oylamadır. Bu yöntemde çıktılar Eşitlik 3.39. da ki gibi doğrusal bir şekilde birleştirilir [81].

$$y_i = \sum_j w_j d_{ji} \quad (3.39)$$

Burada, d_{ji} değeri, j modelinin i.sınıf için verdiği oy, w_j değeri oyun ağırlığıdır.

$w_j \geq 0$ ve $\sum_j w_j = 1$ dir.

Algoritmalar birleştirilirken Tablo 3.2.'de görüldüğü gibi uygulanması gereken bazı kurallar vardır [81].

Tablo 3.2. Oylama birleştirme kuralları.

Kural	Birleştirme Fonksiyonu
Toplam	$y_i = \sum_{j=1}^L d_{ij}$
Ağırlıklı Toplam	$y_i = \sum_j w_j d_{ji}$ burada; $w_j \geq 0, \sum_j w_j = 1$
Ortanca	$y_i = \text{medyan}_j d_{ji}$
En Küçük	$y_i = \min_j d_{ji}$
En Büyük	$y_i = \max_j d_{ji}$
Çarpım	$y_i = \prod_j d_{ij}$

En sık kullanılan birleştirme yöntemi kullanımı en kolay ve anlaşılır olan toplama kuralıdır. Ortanca kuralı aykırı değerlere karşı kuvvetlidir. En küçük ve en büyük kuralı en kötümser ve en iyimser olanının seçimidir. Çarpım kuralında 0 çıktısı baskındır. Bir model 0 çıktısı verdiğinde diğer modellerin çıktıları ne olursa olsun genel son çıktı 0 olacaktır [81].

3.3.2.4. Yığılmış genelleme topluluk yöntemi

Wolpert tarafından geliştirilen yığılmış genelleme, bir algoritmanın farklı türdeki bireysel algoritmaları birleştirmek için eğitildiği genel bir prosedürdür [91]. Çoklu

sınıflandırma modelini birleştirmek için meta-öğrenme tabanlı bir topluluk algoritması olan bu yöntemde eğitim seti iki ayrı sete (yani bir eğitim seti ve bir test seti) bölünür. Temel öğrenme algoritmalarının tahminlerine ve doğru cevaplarına dayanarak, bir meta-öğrenici eğitilir [94].

Burada, temel tekli algoritmalar birinci seviye öğrenen olarak adlandırılırken, birleştirici ikinci seviye öğrenen veya meta öğrenen olarak adlandırılır. Temel fikir, birinci seviye öğrencileri orijinal eğitim veri setini kullanarak eğitmek ve daha sonra, birinci seviye öğrencilerin çıktılarının orijinal durumdayken girdi özellikleri olarak kabul edildiği ikinci seviye öğrenciyi eğitmek için yeni bir veri seti oluşturmaktır. Birinci seviye öğrenenler, genellikle farklı öğrenme algoritmaları uygulanarak üretilir ve bu nedenle, yığılmış topluluklar genellikle heterojendir [91]. İkinci seviyedeki meta veri seti ise tüm algoritmaların tahminlerinden oluşur [94]. Bu algortmada, tahmin edilen hataları birinci öğrencinin çıktılarına eklemek, hataları düzelterek gelişmiş bir son sınıflandırma kararı sağlayabilir [95].

3.3.3. Makine öğrenme performans ölçüm kriterleri

Makine öğrenme algoritmalarıyla kurulan modellerin performansları, gelecekteki veriler üzerindeki öngörücü performansı tahmin etmek, belirli bir hipotez uzayından en iyi performans gösteren modeli seçerek öngörü performansını artırmak ve/veya problem için en uygun makine öğrenme algoritmasını belirlemek gibi nedenlerden dolayı değerlendirilmelidir [96]. Bu değerlendirme aşaması makine öğrenme modelleri için kritik önem taşımaktadır.

Algoritmaların performanslarını ölçmek için birçok kriter vardır. Ancak kullanım amacına göre kullanılacak kriterler farklılık gösterir. Regresyon uygulaması yapılıyorsa, hatalarla ilgili ölçüm yapılır, sınıflandırma yapılıyorsa, doğruluklarla ilgili ölçümler yapılır [97].

Çalışmada sınıflandırma ve tahmin yapıldığı için bu alanda kullanılan, hata matrisi, doğruluk, duyarlılık, kesinlik, F ölçütü, kappa istatistiği gibi kriterler açıklanacaktır.

3.3.3.1. Hata matrisi

Makine öğrenme algoritmalarının performanslarını değerlendirmek için kullanılan hata matrisi, sonuçları görsel ve anlaşılır şekilde ifade ettiğinden dolayı sıklıkla kullanılmaktadır. Tablo 3.3.'de görüldüğü gibi sütunlar gerçek değerleri, satırlarda tahmin değerlerini ifade etmektedir [73].

Tablo 3.3. Hata matrisi [73].

	Gerçek Pozitif Sınıf	Gerçek Negatif Sınıf
Tahmin Pozitif Sınıf	GP	YN
Tahmin Negatif Sınıf	YP	GN

Burada;

GP: test kümesindeki gerçek pozitif örnek sayısı,

GN: test kümesindeki gerçek negatif örnek sayısı,

YP: test kümesindeki yanlış pozitif örnek sayısı,

YN: test kümesindeki yanlış negatif örnek sayısı

olarak ifade edilmektedir.

Hata matrisinden yola çıkılarak hesaplanabilecek performans ölçütleri içerisinde çoklu sınıf problemlerini değerlendirmek için kullanılacak kriterler Tablo 3.4.'de gösterilmektedir [98].

Tablo 3.4. Çok sınıflı problemlerde sınıflandırma için hata matrisine bağlı performans kriterleri [98].

Kriter	Formül	Açıklama
Ortalama Doğruluk	$\frac{\sum_{i=1}^l \frac{g_{pi} + g_{ni}}{g_{pi} + y_{ni} + y_{pi} + g_{ni}}}{l}$	Sınıfların ortalama etkinliği
Ortalama Hata Oranı	$\frac{\sum_{i=1}^l \frac{y_{pi} + y_{ni}}{g_{pi} + y_{ni} + y_{pi} + g_{ni}}}{l}$	Sınıfların ortalama hata oranı
Ortalama Hassasiyet	$\frac{\sum_{i=1}^l \frac{g_{pi}}{g_{pi} + y_{pi}}}{l}$	Sınıf başına hassasiyetin ortalaması
Ortalama Kesinlik	$\frac{\sum_{i=1}^l \frac{g_{pi}}{g_{pi} + y_{ni}}}{l}$	Sınıf başına kesinlik ortalaması
Ortalama F ölçümü	$\frac{2 * h_M * k_M}{h_M + k_M}$	Sınıf başına F ölçümü

Burada, g_{pi} ; i.sınıftaki gerçek pozitif sayısını, g_{ni} ;i.sınıftaki gerçek negatif sayısını, y_{pi} ;i.sınıftaki yanlış pozitif sayısını, y_{ni} ;i.sınıftaki yanlış negatif sayısını h_M ; hassasiyetin makro ortalamasını, k_M ;kesinliğin makro ortalamasını ifade etmektedir.

3.3.3.2. Kappa istatistiği

Kappa istatistiği, sınıflandırma doğruluğunu, doğru tahmin olasılığındaki şans faktörünü hesaba katarak değerlendirmektedir. Eşitlik 3.40. da görüldüğü gibi hesaplanmaktadır [82].

$$k = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)} \quad (3.40)$$

Burada $\text{Pr}(a)$ ve $\text{Pr}(e)$ sırasıyla gerçek ve beklenen değerlerin, sınıflandırıcı ve gerçek değerler arasındaki uyum oranını ifade etmektedir.

Kappa değerleri, modelin tahminleri ile gerçek değerler arasında mükemmel bir uyum olduğunu gösteren maksimum 1 değerine kadar değişmektedir. Birden az değerler hatalı tahminleri göstermektedir. Değerler yaygın olarak şu şekilde yorumlanmaktadır [82]:

- Kötü tahmin = 0,20'den az
- Makul düzey tahmin = 0,20 – 0,40
- Orta düzey tahmin = 0,40 ila 0,60
- İyi tahmin = 0,60 ila 0,80
- Çok iyi tahmin = 0,80 – 1,00

3.3.4. Dengeli olmayan verinin düzenlenmesi

Veri kümesindeki sınıflar yaklaşık olarak eşit olarak temsil edilmiyorsa veri kümesi dengesiz denilebilir. Makine öğrenimi algoritmalarının performansı genellikle öngörücü doğruluğu temel almaktadır. Ancak, veriler dengesiz olduğunda genellikle

çoğunluk sınıfı çok az hata ile tahmin edilirken, azınlık olan sınıf (lar) tahmin edilemez. Bu durumda öngörücü doğruluk kullanılmasının yanıltıcı olacağı söylenilebilir [99].

Verideki sınıf dengesizliği iki şekilde ele alınmaktadır. Birincisi, eğitim örneklerine farklı ağırlıklar atamasıdır. Diğeri ise, ya azınlık sınıfını fazla örnekleyerek ve / veya çoğunluk sınıfını az örnekleyerek orijinal veri kümesini yeniden örneklemektir [99].

Yeniden örnekleme için yaygın olarak kullanılan Sentetik Azınlık Aşırı Örnekleme (SMOTE) yöntemi, azınlık sınıftan sentetik numuneler üreten bir örnekleme tekniğidir. Sentetik olarak azınlık sınıfındaki veri sayısını çoğunluk sınıfındaki veri sayısına eşitleyen bu yöntem, dengeli veya neredeyse dengeli sınıflı bir eğitim seti elde etmek için kullanılır. SMOTE örnekleri, azınlık sınıftan iki benzer örneğin doğrusal kombinasyonlarıdır ve Eşitlik.3.41. ile elde edilmektedir [100].

$$s = x + u (x^R - x) \quad (3.41)$$

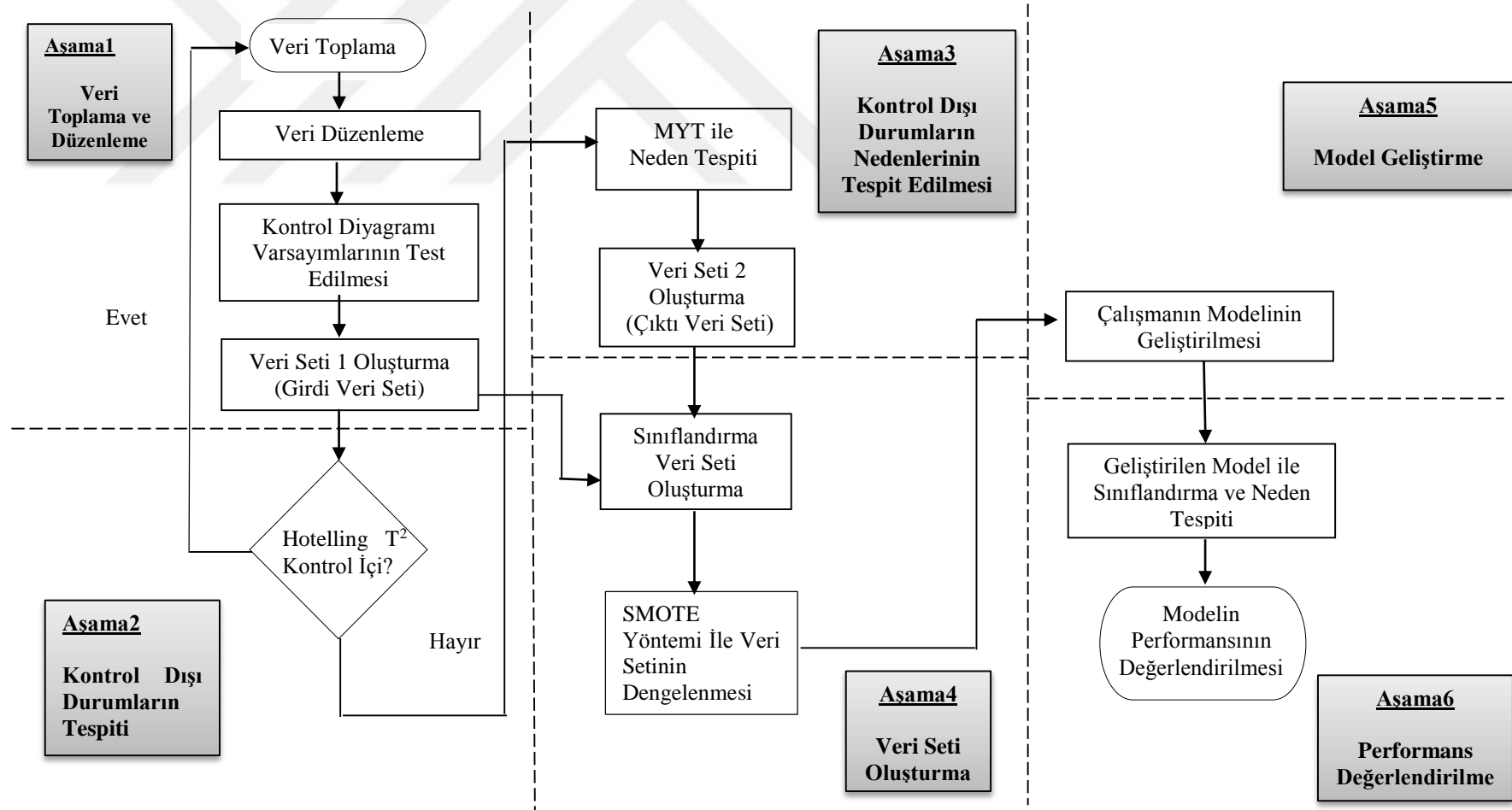
Burada x^R ve x iki benzer sınıf olmak üzere x^R , x 'in en yakın 5 azınlık sınıfı arasından rastgele seçilir. u ise 0 ile 1 arasında rastgele bir sayıdır.

BÖLÜM 4. KONTROL DIŐI DURUM NEDENLERİNİN TESPİT EDİLMESİ İÇİN ÖNERİLEN MODEL

Çalışmanın temel amacı; üretim sürecinde karşılaşılabacak kontrol dışı durumlara neden olan değişkenlerin makine öğrenme algoritmaları ile tespit etmektir. Bu amaç doğrultusunda, kontrol dışı duruma neden olan değişkenlerin mevcut durumlardan öğrenerek, yeni örnek kontrol dışı durum oluşturuyor ise, bunun hangi değişkenden dolayı olduğunun tespit edilmesi hedeflenmektedir. Bunu sağlamak için, mümkün olan en yüksek doğruluklarla kontrol dışı durum nedenlerinin tespit edilmesi gerekmektedir. Bu doğrultuda ilk olarak kontrol dışı durumlar, Hotelling T^2 kontrol diyagramı ile belirlenecektir. Ardından kontrol dışı durumlara neden olan değişken ya da değişken grubu MYT yöntemi ile tespit edilecektir. Elde edilen bu sonuçlarla oluşturulan veri seti ile gelecek dönemler için tahminde bulunan makine öğrenme topluluk algoritmaları kullanılarak sınıflandırma yapılacaktır.

İyileştirmenin hangi değişken üzerinde yapılacağına doğru tahmin edilerek belirlenmesi gereksiz zaman ve kaynak tüketilmesini önleyecektir. Tahmin doğruluklarının artırılması amacıyla geliştirilen bu modelde ilk olarak, belirlenen optimum parametrelere göre, temel tekli algoritmaların performansları karşılaştırılarak en başarılı olan seçilecektir. Algoritmaların topluluk olarak kullanılmasının başarı oranının artırarak tekli kullanılmasından daha başarılı olduğu bilinmektedir [101–104]. Bu nedenle seçilen bu algoritma torbalama ve yükseltme algoritmaları ile paralel ve sıralı olarak birleştirilerek performansları değerlendirilecektir. Son olarak bu iki topluluk algoritması yığılmış genelleme topluluk algoritması ile birleştirilerek geliştirilen bir model önerilmiştir.

Çalışma Şekil 4.1.'de sunulan altı aşamadan oluşmaktadır.



Şekil 4.1. Çalışmanın akış şeması.

Aşamalar şu şekildedir;

1. Aşama: Veri toplama ve düzenleme
2. Aşama: Hotelling T^2 kontrol diyagramı ile kontrol dışı durumların belirlenmesi
3. Aşama: Kontrol dışı durumların nedenlerinin tespit edilmesi
4. Aşama: Veri seti oluşturma
5. Aşama: Önerilen modelin geliştirilmesi
6. Aşama: Modelin performansının değerlendirilmesi

Aşamaların içerdiği adımlar ve önerilen model aşağıda detaylı olarak anlatılmaktadır.

4.1. Veri Toplama ve Düzenleme

Bu aşama, çalışmada kullanılacak parçaya ait incelenen kalite karakteristiklerinin yani değişkenlerin ölçüm değerlerini içeren veri setinin toplandığı aşamadır. Veri setinin toplanmasının ardından ölçüm değerlerinin aykırı, eksik ya da tutarsız değer içerip içermediği kontrol edildikten sonra varsa bu değerler veri setinden temizlenmelidir. Ardından çok değişkenli kontrol diyagramının 4 temel varsayımı kontrol edilerek, verilerin kontrol diyagramında değerlendirme uygunluğu ve değişkenler test edilmelidir. Bu işlemlerin sonucunda sınıflandırmada kullanılacak veri setinin girdi değerleri elde edilecektir.

4.2. Kontrol Dışı Durum Tespiti

Kontrol diyagramlarında istatistiksel olarak hesaplanarak bulunan sınırların dışına çıkan örnekler kontrol dışı durumları oluşturmaktadır. Çalışmanın bu aşamasında önceki adımda belirlenmiş olan ilişkili değişkenlerin ölçüm değerlerini içeren veri seti kullanılarak, Hotelling T^2 kontrol diyagramı ile süreçteki üst kontrol sınırını aşan örnekler belirlenmektedir.

4.3. Kontrol Dışı Duruma Neden Olan Değişkenlerin Tespit Edilmesi

Kontrol dışı durumlar Hotelling T^2 ile belirlenmiş olsa da bu durumların nedenleri bilinmemektedir. Çok değişkenli kontrol diyagramlarının bu eksikliğinden dolayı ilave teknik ve yöntemlere ihtiyaç duyulmaktadır.

Bu aşama da Hotelling T^2 için özel olarak geliştirilen MYT ayrıştırma yöntemi ile örneklerin hangi değişkenden dolayı üst kontrol sınırı dışına çıktığı tespit edilmektedir. Bu şekilde tespit edilen kontrol dışı durumlara neden olan değişkenler, bir sonraki aşama için, sınıflandırmada kullanılacak veri setinin çıktılarını oluşturmaktadır.

4.4. Veri Seti Oluşturma

Girdiler, 1.aşamada belirlenmiş olan değişken ve bunların ölçüm değerleri ile elde edilmiştir. Çıktılar ise 2.aşamada belirlenen kontrol dışı durumların nedenlerinin tespit edildiği 3.aşamada elde edilmiştir. Son olarak, bu iki veri setinin örnekler temel alınarak karşılıklı olarak birleştirilmesi neticesinde uygulamada modeli eğitmek ve test etmek için kullanılan veri seti oluşturulmuştur. Veri setindeki sınıflar arasındaki dengesizliği gidermek için smote sentetik veri üretme yönteminden yararlanılmıştır.

Önerilen model literatürdeki diğer çalışmalardaki [12, 47, 51] gibi sadece hatalı veriler üzerinde sınıflandırma yapmaktadır.

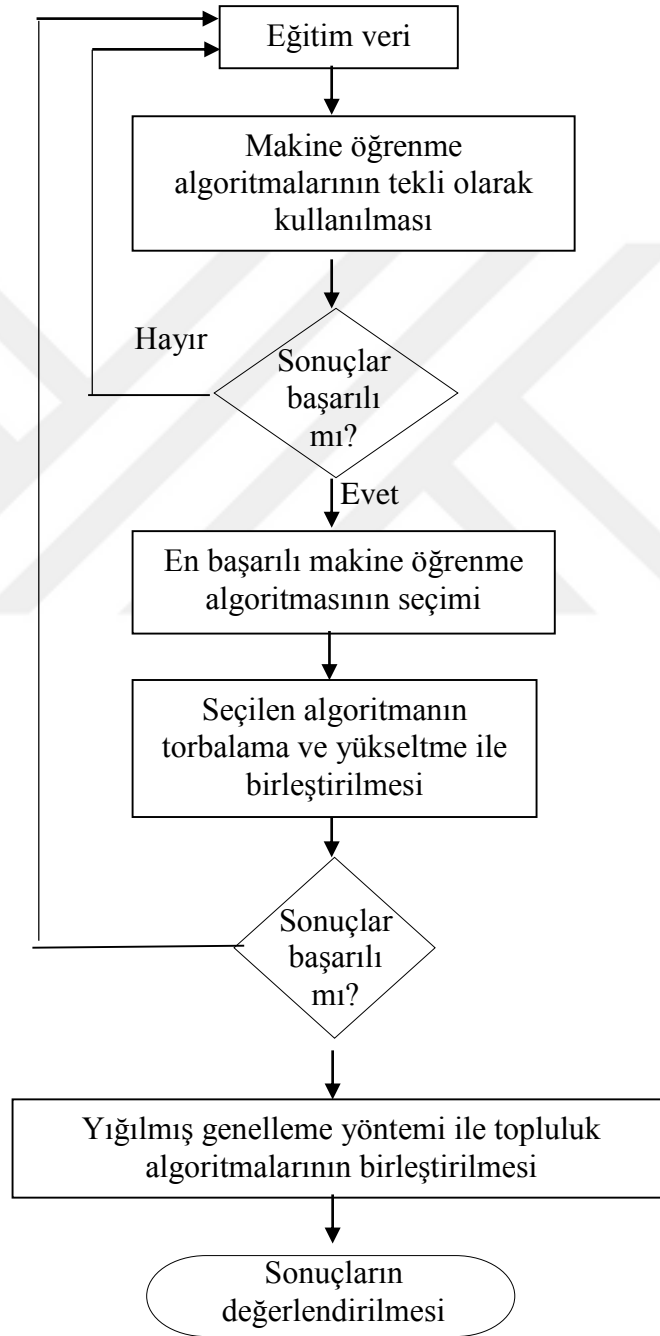
4.5. Önerilen Modelin Geliştirilmesi

Üretim sürecinde karşılaşılan kontrol dışı duruma neden olan değişkenlerin tespit edilmesi için önerilen bu modelde yöntem olarak daha önce kullanılmamış olan topluluk algoritmalarının birleştirilmesi hedeflenmektedir.

Önerilen model üç adımdan oluşmaktadır. İlk olarak makine öğrenme algoritmaları tekli olarak kullanılarak, belirlenen kriterlere göre en başarılı olan algoritma seçilecektir. İkinci adımda bu algoritma sıralı ve paralel olarak birleştirilerek

performansları değerlendirilecektir. Üçüncü adımda ise, topluluk algoritmaları başarılı görülürse, bu iki topluluk algoritması bir diğer topluluk algoritması olan yığılmış genelleme yöntemi ile birleştirilecektir.

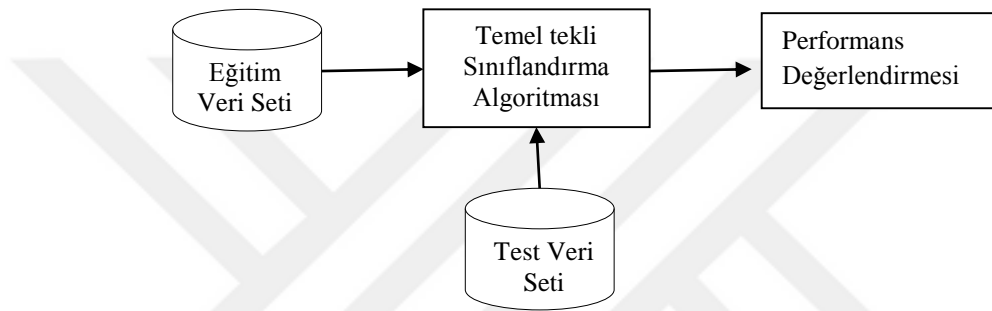
Önerilen modelin yapısı Şekil 4.2.'de gösterilmiştir.



Şekil 4.2. Önerilen model genel yapısı.

4.5.1. Makine öğrenme algoritmalarının tekli olarak kullanılması

Modelde torbalama ve yükseltme algoritmaları ile birleştirilmek için kullanılacak algoritmanın belirlenmesini içeren bu adımda temel 5 makine öğrenme algoritması kullanılacaktır. Şekil 4.3.'de görüldüğü gibi algoritmaların her biri için tekrarlanarak veri seti ile eğitilmiş ve test edilerek performansları belirlenerek birbirleriyle karşılaştırılarak değerlendirilecektir. Çalışmada kullanılacak tekli algoritmalar; KA, K-NN, YSA, NB, Ç-DVM'dir.



Şekil 4.3. Temel algoritmaların tekli kullanımı.

Çalışmada algoritmalar için en uygun parametreler belirlenecektir. Parametre değerlerinden kaynaklanacak doğruluk oranları ile ilgili farklılıkların ortadan kaldırılması için çalışmanın ilerleyen adımlarında da bu adımda belirlenen parametre değerleri kullanılacaktır.

4.5.2. Seçilen makine öğrenme algoritmasının topluluk yöntemleri ile birleştirilmesi

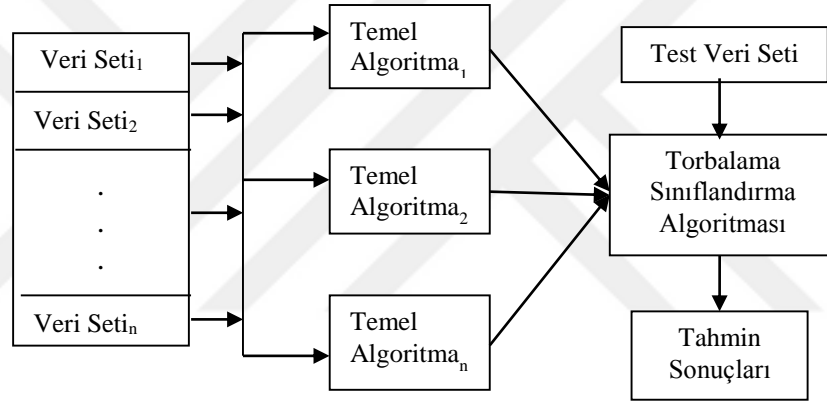
Bu adımda, ilk adımda seçilmiş olan makine öğrenme algoritması torbalama ve yükseltme topluluk algoritmaları ile paralel ve sıralı şekillerde birleştirilerek kullanılacaktır.

Bu yöntemler sayesinde, algoritmaların sonuçları birleştirilerek, sınıflandırma doğruluk oranlarının tek algoritma kullanımına göre daha yüksek olması hedeflenmektedir.

- Torbalama topluluk yöntemi ile birleştirilmesi

Torbalama topluluk yönteminde, tekli algoritmalar, eğitim veri setinin eşit örnek sayılarıyla ayrılmasıyla oluşturulan farklı veri seti kümeleri (veri seti₁, veri seti₂, ..., veri seti_n) ile eğitilerek, paralel şekilde bağlanmaktadır. Bu yöntemde kararlar basit ortalama ile oylama sonucunda verilir. Oylama, en yüksek oranda rastlanan sınıfın seçilmesi ile gerçekleştirilir.

Bir önceki adımda en başarılı olarak seçilen algoritma Şekil 4.4.'de görüldüğü gibi birleştirilerek tahmin doğruluk performansları değerlendirilir.

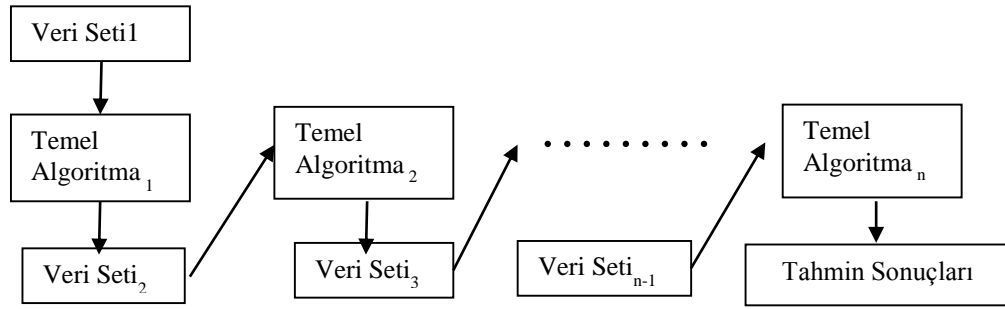


Şekil 4.4. Torbalama yöntemi ile birleştirme.

- Yükseltme topluluk yöntemi ile birleştirilmesi

Yükseltme yönteminde, torbalama yöntemine göre en temel fark, her veri seti kümesinde eğitilen temel algoritma, eğitim verileri üzerinde sınıflandırma sonucu iyi olan bir algoritmaya, kötü olandan daha yüksek bir ağırlık verilecek şekilde, ağırlıklandırılarak yeni algoritma için girdi şeklinde kullanılmasıdır.

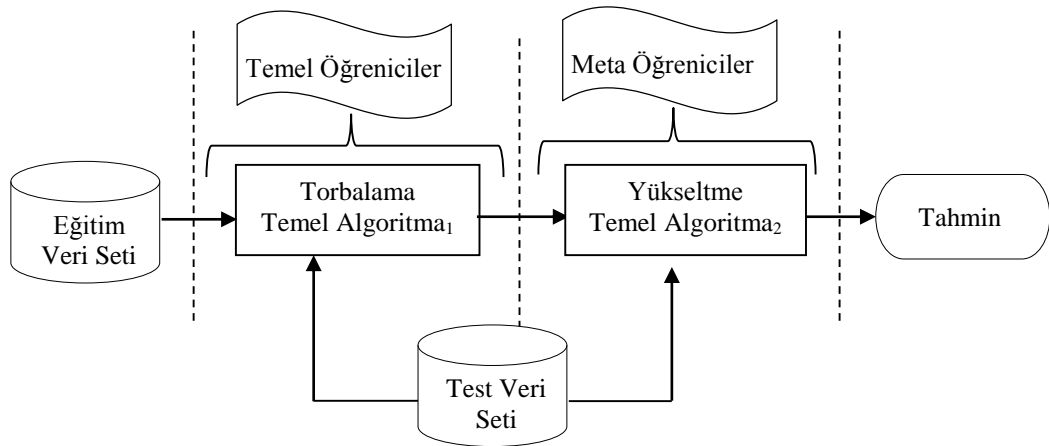
Yükseltme topluluk yöntemi ile birleştirilme işlemini Şekil 4.5.'de görüldüğü gibidir.



Şekil 4.5. Yükseltme yöntemi ile birleştirme.

4.5.3. Topluluk yöntemlerinin yığılmış genelleme topluluk yöntemi ile birleştirilmesi

Çalışmada önerilen modelin oluşturulduğu bu adım, literatürdeki çalışmalardan farklı olarak, topluluk algoritmalarının diğer bir topluluk algoritması ile birleştirilmesi temeline dayanmaktadır. Topluluk algoritmalarının kullanımının başarıyı artırdığı birçok çalışmada görülmüştür. Başarı oranları yükseltilmiş olan yöntemlerin birleştirilmesiyle oluşturulan modelin başarı oranlarını daha da yükselteceği düşünülmektedir. Topluluk algoritmalarının birleştirilmesine dayayan model Şekil 4.6.'da görüldüğü gibi kurulmaktadır.



Şekil 4.6. Yığılmış genelleme yöntemi ile topluluk algoritmalarının birleştirilmesi.

Seçilen bu iki topluluk yöntemi yığılmış genelleme algoritması ile birleştirilerek eğitilir ve performans sonuçları elde edilir. Sonuçlar bu modelin diğerlerine göre

sınıflandırma performansları açısından daha başarılı olduğunu gösterirse bu modelin uygulanabilirliği ispatlanacaktır.

4.6. Performans Değerlendirilmesi

Önerilen model veri seti ile eğitilerek performansı çeşitli kriterlere göre değerlendirilerek eğer değerlendirme sonuçları başarılı görülürse modelin uygunluğuna karar verilecektir. Sınıflandırma algoritmalarının performans karşılaştırmaları için, doğruluk, duyarlılık, kesinlik ve kappa istatistiği gibi kriterleri kullanılmıştır.



BÖLÜM 5. UYGULAMA

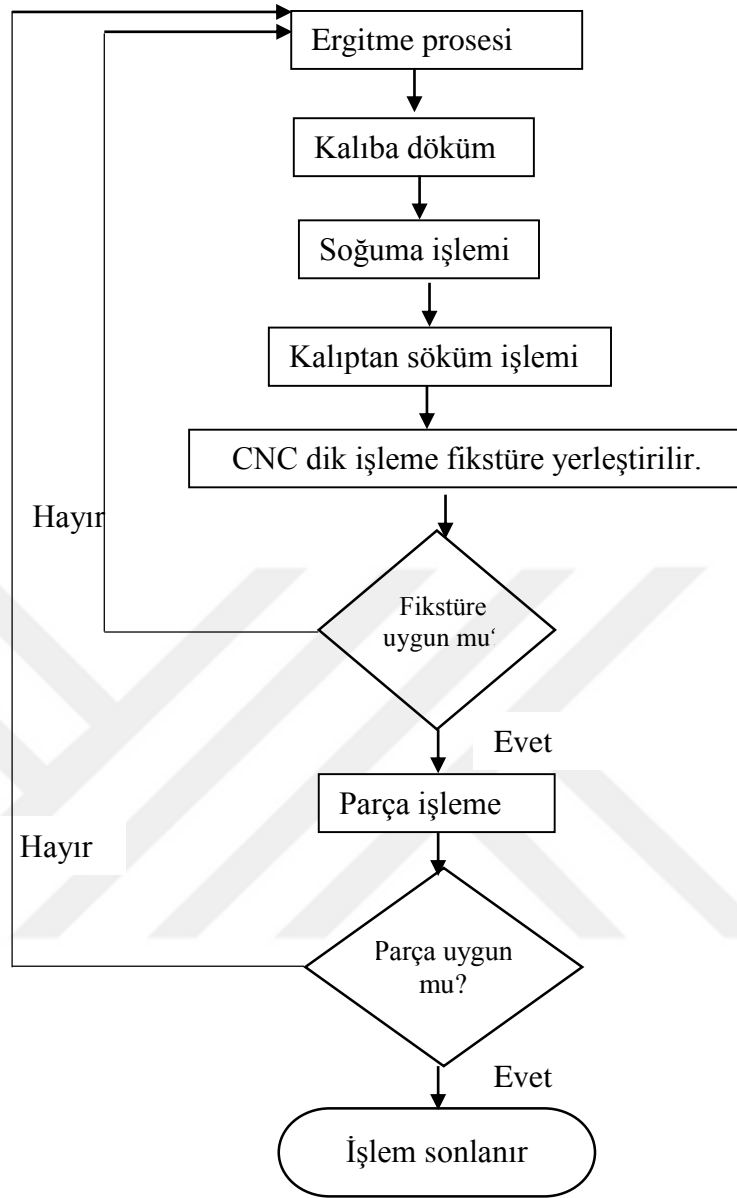
5.1. Problem Tanımı

Çalışma Türkiye’de faaliyet gösteren otomotiv yan sanayi tedarikçisi olan bir işletmede gerçekleştirilmiştir. Müşterilerden gelen montaj problemleri yoğunluğu nedeniyle işletme açısından incelenmesi uygun görülen parça olarak çelik hidrolik pompa kapağı seçilmiştir.

Hidrolik pompa kapağı parçasının döküm üretimindeki işlemleri içeren süreç akış şeması Şekil 5.1.’de gösterilmektedir.

Çalışmada incelenen çelik pompa kapağı parçasının üretimi için ilk olarak çelik, ark ocağında ergitilmektedir daha sonra ergimiş çelik kalıba dökülerek soğutma işlemi uygulanmaktadır. Soğuma işlemi tamamlanan parça kalıptan çıkartılarak CNC dik işleme merkezinde fikstüre yerleştirilmektedir. Parçanın kalite değerlendirmesi bu aşamada başlamalıdır. Değerlendirme sırasında aşağıdaki iki temel kalite problemi ile karşılaşılabilir.

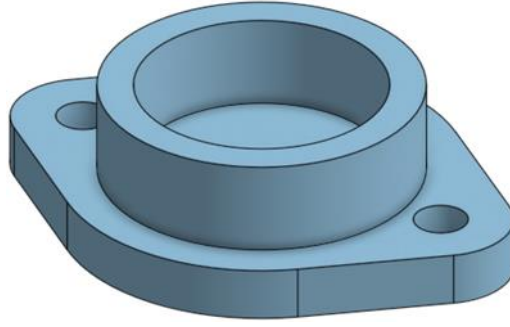
1. Parçanın üzerinde etkili olan değişkenlerdeki ölçü hatalarından dolayı CNC tezgâha yerleşememesi.
2. Yerleşme sorunu olmamasına rağmen fikstüre bağlanan parçanın dökümdeki ölçüm hatalarından dolayı, CNC de gerektiği gibi işlenememesi.



Şekil 5.1. Süreç akış şeması.

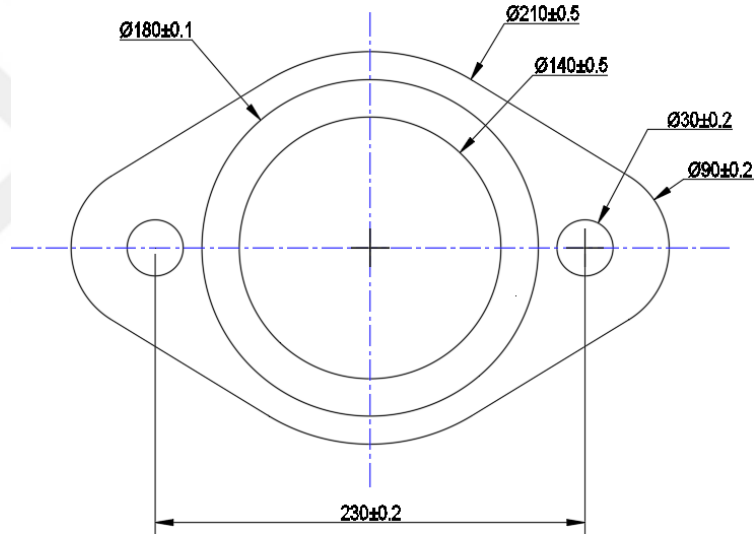
Hidrolik pompa kapağı parçasının 3 boyutlu görünüşü Şekil 5.2.'de görülmektedir. Kalite uzmanları tarafından parçanın üzerinde kaliteye etki eden 8 tane kalite karakteristiği yani değişken belirlenmiştir. Değerlendirmeler bu değişkenlere göre gerçekleştirilecektir.

Parça üzerine etki eden değişkenler; 1.delik çapı (x_1), 2.delik çapı (x_2), büyük dış çap (x_3), delikler arası mesafe (x_4), yanak yüksekliği (x_5), yanak dış çapı (x_6), yanak iç çapı (x_7), kapak et kalınlığı (x_8) olarak belirlenmiştir.



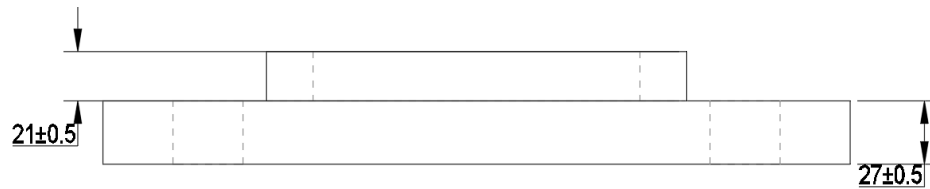
Şekil 5.2. Hidrolik pompa kapağı.

Çalışmada değerlendirilen parçanın kalite açısından incelenen değişkenlerine ait ölçülerin bulunduğu, üstten görünüşünü içeren teknik resim Şekil 5.3.'de görülmektedir.



Şekil 5.3. Üstten görünüş.

Parçanın önden görünüşünü içeren teknik resimleri Şekil 5.4.'de görülmektedir.



Şekil 5.4. Önden görünüş.

Değişkenlerin tanımı, ortalama ve spesifikasyon değerleri Tablo 5.1.'deki gibidir.

Tablo 5.1. Değişkenlerin tanımı.

Değişken	Tanımlama	Spesifikasyon (mm)	Değ	Tolerans(mm)
x ₁	1.Delik Çapı	30		±0,2
x ₂	2.Delik Çapı	30		±0,2
x ₃	Büyük Dış Çap	210		±0,5
x ₄	Delikler Arası Mesafe	230		±0,2
x ₅	Yanak Yüksekliği	21		±0,5
x ₆	Yanak Dış Çapı	180		±0,1
x ₇	Yanak İç Çapı	140		±0,5
x ₈	Kapak Et Kalınlığı	27		±0,5

Otomotiv sektöründe kalite çok önemli ve kritik bir faktördür bu nedenle hataların oluşmadan ortadan kaldırılması için kalitesizliğe sebep olan değişkenlerin mümkün oldukça hızlı ve doğru şekilde tespit edilerek süreçlerde sürekli iyileştirme yapılması gerekmektedir. Çalışmada incelenen parçanın üretim miktarları yüksek olması ve değişkenlerdeki toleransların çok düşük olmasından dolayı daha seri, dikkatli ve hassas hareket edilmesi gerekmektedir. Yukarıda bahsedilen 8 değişkenin değerlerinin ayrı ayrı izlenmesi, değişkenler arası ilişki, zaman ve iş yükü kaybına neden olacaktır. Bu nedenle problem çözümü için çok değişkenli kontrol diyagramlarının kullanılması fayda sağlayacaktır.

5.2. Çalışmanın Adımlarının Uygulanması

Bir önceki bölümde tanımlanmış olan çalışmanın uygulama adımları uygulanarak sonuçları ele alınmıştır.

5.2.1. Veri toplama ve düzenleme

Günde 8 saat 3 vardiya çalışan firmada kalite değerlendirmesi, süreçten saat başı tek örnek alınarak yapılmaktadır. Çalışmada firmanın uzun süreli toplamış olduğu geçmiş veriler kullanılmıştır. Veri seti, 8 değişkene ait 26.700 tane ölçüm değerinden oluşmaktadır. Veri setindeki aykırı, uç ya da eksik değerler veri ön işleme adımlarıyla ilk olarak değişken ve örnek uygunluğu açısından incelenmiştir. Sonuç olarak veri setinde aykırı ya da uç değerlere rastlanmamış. Daha sonra eksik veriler incelenerek 3 tane örnekte ölçüm değerleri açısından eksik değişkenler olduğu görülmüştür. Bu 3 örnek elenerek kalan 26697 örnek üzerinde kalite değerlendirmesi esas alınmıştır.

5.2.2. Kontrol dışı duruma neden olan değişkenlerin tespiti

Çalışmada gerçekleştirilen uygulama süreci çok değişkenli olduğu için süreçlerde rastlanan kontrol dışı örnekleri tespit etmek için çok değişkenli kontrol diyagramlarından biri olan Hotelling T^2 kullanılmıştır. Bu kontrol diyagramı T^2 istatistiğini temel aldığından dolayı çok değişkenli istatistik varsayımlarını sağlayıp sağlamadığı kontrol edilmelidir.

5.2.2.1. Hotelling T^2 kontrol diyagramının varsayımlarının kontrol edilmesi

Çok değişkenli kontrol diyagramı varsayımlarının sağlanıp sağlanmadığı kontrol edilerek uygun görülmeyen varsayımlar için gerekli değişim ve dönüşümlerle uygunluk sağlanmalıdır.

Kontrol edilmesi gereken varsayımlar; doğrusallık, normal dağılıma uygunluk, otokorelasyon olmaması ve varyans-kovaryans eşitliği varsayımları olmasına rağmen çalışmada örnek büyüklüğü 1 olduğundan varyans-kovaryans varsayımı geçersiz olmaktadır. Çünkü bu varsayımın geçerli olması için örnek büyüklüğünün 1'den büyük olması gerekmektedir. Ancak çalışmada örneklem büyüklüğü 1 olduğundan dolayı varyans kovaryans varsayımının kontrolüne gerek kalmamaktadır. Bu nedenle 3 varsayımın sağlanıp sağlanmadığı incelenmiştir.

– Doğrusallık

Doğrusallık varsayımında değişkenler arasında doğrusal bir ilişki olup olmadığı araştırılmaktadır. Bu varsayımın geçerliliğini belirleyebilmek için değişkenler arasındaki korelasyon katsayıları hesaplanıp anlamlılıklarına bakılmıştır.

Doğrusal ilişkinin varlığının incelenmesinde iki değişken arasında lineer ilişki olup olmadığını belirlemek için Pearson Korelasyon (PK) katsayıları ve anlamlılık düzeyi (sig.) ile aşağıdaki hipotezlerle test edilmiştir.

H₀: Değişkenler arasında korelasyon katsayısı 0'dır.

H₁: Değişkenler arasında korelasyon katsayısı 0'dan farklıdır.

Hipotezlerin geçerliliği için α değeri 0,05 olarak belirlenmiştir. Anlamlılık düzeyi $>0,05$ olduğunda H₀ hipotezi kabul edilmektedir yani değişkenler arasında doğrusal bir ilişki yoktur. Anlamlılık düzeyi $<0,05$ olduğunda H₀ hipotezi reddedilmektedir yani değişkenler arasında doğrusal bir ilişki olduğu şeklinde değerlendirilmiştir.

Bu durumda Tablo 5.2.'deki değerler incelendiğinde bazı değişkenlerin bir ya da daha fazla değişkenle arasında çok zayıf ilişki bulunmasına rağmen diğerleriyle çok yüksek ilişkiler mevcuttur. Örneğin 1.delik çapı (x₁) değişkeninin 2.delik çapı (x₂) değişkeni ile aralarında (-) 0,75 katsayı değeri ile ters yönlü kuvvetli bir doğrusal ilişkisi mevcuttur ancak diğer değişkenlerle daha zayıf ilişkiler mevcuttur. Bunun yanında değişkenlerden 2 tanesinin (büyük dış çap (x₃) ve kapak et kalınlığı (x₈)) diğer hiçbir değişkenle ilişkisi bulunmadığı anlaşılmaktadır.

Tablo 5.2. Değişkenler arası korelasyon matrisi ilk durum.

		X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈
x ₁ :1.delik çapı	P.K.	1	-0,75**	0,01	0,63**	0,048**	-0,004	0,062**	0,007
	Sig.		0,000	0,881	0,000	0,000	0,520	0,000	0,264
x ₂ :2.delik çapı	P.K.		1	-0,002	0,18**	-0,008	-0,001	-0,048**	0,004
	Sig.			0,725	0,003	0,195	0,852	0,000	0,562
x ₃ :büyük dış çap	P.K.			1	-0,010	-0,008	0,001	-0,001	0,008
	Sig.				0,104	0,184	0,874	0,881	0,183
x ₄ :delikler arası mesafe	P.K.				1	-0,019**	-0,009	-0,038**	0,006
	Sig.					0,002	0,162	0,000	0,344
x ₅ :yanak yüksekliği	P.K.					1	-0,21**	0,037**	-0,004
	Sig.						0,001	0,000	0,508
x ₆ :yanak dış çapı	P.K.						1	-0,011	0,001
	Sig.							0,065	0,907
x ₇ :yanak iç çapı	P.K.							1	-0,003
	Sig.								0,606
x ₈ :kapak et kalınlığı	P.K.								1
	Sig.								

**0,01 düzeyinde önemli olan korelasyon

Tablo 5.2.'ye göre korelasyon katsayıları incelendiğinde en kuvvetli ve ters yönlü ilişki, 0,75 değeriyle, 1. ve 2. delik çapı arasında olduğu gözlenmiştir. Yani bir delik çapı büyürken diğerinin küçülmesi beklenmektedir.

Bunun yanında bazı değişkenler arasında zayıf pozitif ya da negatif ilişkiler bulunmaktadır. Ayrıca, büyük dış çap (x_3) ve kapak et kalınlığı (x_8) değişkenlerinin, diğer değişkenlerle arasında doğrusal bir ilişkisi bulunmamaktadır. Bu nedenle çok değişkenli kontrol diyagramı ile değerlendirmesine gerek görülmemiştir. Bu değişkenler ilişkisiz olduğundan dolayı tek değişkenli kontrol diyagramları ile ayrı ayrı ele alınabilir. Büyük dış çap (x_3) ve kapak et kalınlığı (x_8) değişkenleri veri setinden çıkartılarak doğrusallık varsayımı 6 değişken için tekrarlanmıştır.

Sonuçlar incelenerek uygun görülürse diğer varsayımlarla devam edilmelidir.

6 değişken arasındaki ilişkilerin değerlendirilmesi için hesaplanan korelasyon değerleri Tablo 5.3.'de gösterilmektedir. Kalan 6 değişkenin anlamlılık düzeyleri incelendiğinde genel olarak anlamlı görülmektedir yani doğrusal ilişki bulunmaktadır.

Kalan değişkenler 1. delik çapı (x_1), 2. delik çapı (x_2), delikler arası mesafe (x_3), yanak yüksekliği (x_4), yanak dış çapı (x_5) ve yanak iç çapı (x_6) olarak yeniden tanımlanmıştır.

Tablo 5.3. Değişkenler arası korelasyon matrisi son durum.

		X1	X2	X3	X4	X5	X6
x1:1.delik çapı	P.K.	1	-0,75**	0,63**	0,048**	-0,004	0,062**
	Sig.		0,000	0,000	0,000	0,520	0,000
x2:2.delik çapı	P.K.		1	0,18**	-0,008	-0,001	-0,048**
	Sig.			0,003	0,195	0,852	0,000
x3:delikler arası mesafe	P.K.			1	-0,019**	-0,009	-0,038**
	Sig.				0,002	0,162	0,000
x4:yanak yüksekliği	P.K.				1	-0,21**	0,037**
	Sig.					0,001	0,000
x5:yanak dış çapı	P.K.					1	-0,011
	Sig.						0,065
x6:yanak iç çapı	P.K.						1
	Sig.						

**0,01 düzeyinde önemli olan korelasyon

Yeni korelasyon sonuçlarına göre; x_5 dışındaki tüm değişkenlerde diğer değişkenlerin kuvvetli ilişkili olduğu görülmüştür. x_5 değişkeninde ise sadece x_4 değişkeni ile kuvvetli bir ilişki olduğu ve diğer değişkenlerle zayıf ilişkiler olduğu belirlenmiştir. Kuvvetli ilişkili değişken bulunduğundan dolayı, x_5 değişkeninin de doğrusallık varsayımına uygun olduğuna ve diğer varsayımlar için değerlendirilecek değişkenler

arasında ele alınmasına karar verilmiştir. Korelasyon sonuçlarına göre doğrusallık varsayımını sağlayan 6 değişkene göre diğer varsayımlara devam edilmiştir.

- Çok değişkenli normal dağılıma uygunluk varsayımı

Çok değişkenli normal dağılıma uygunluk için değişkenlerin tek tek normal dağılıma uyması gereklidir. Bunun için öncelikli olarak normal dağılıma uygunluğu test etmek için hipotezler oluşturulmuştur.

H_0 : Değişkenin değerleri normal dağılıma uygundur.

H_1 : Değişkenin değerleri normal dağılıma uygun değildir.

Hipotezlerin geçerliliği için α değeri 0,05 olarak belirlenmiştir.

6 kalite değişkenine ait elde edilen normal dağılım uygunluk test sonuçları Tablo 5.4.'de yer almaktadır. Tüm değişkenler için p değerinin 0,05'den büyük olduğu ve bu nedenle tüm değişkenlerin tek tek normal dağılıma uygun olduğu söylenebilir.

Tablo 5.4. Tek değişkenli normal dağılım sonuçları.

Test	Değişken	KS değeri	P değeri
Kolmogorov-Smirnov	x_1	0,004	>0,150
Kolmogorov-Smirnov	x_2	0,004	>0,150
Kolmogorov-Smirnov	x_3	0,004	>0,150
Kolmogorov-Smirnov	x_4	0,005	>0,150
Kolmogorov-Smirnov	x_5	0,003	>0,150
Kolmogorov-Smirnov	x_6	0,002	>0,150

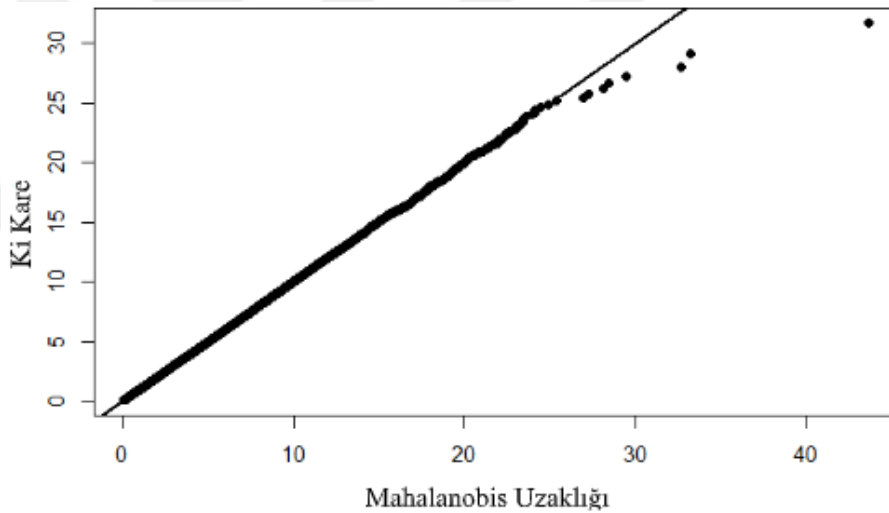
Değişkenlerin tümünün tek değişkenli normal dağılıma uygunluğu kanıtlandıktan sonra tüm değişkenlerin birlikte değerlendirildiği çok değişkenli normal dağılım incelenmiştir çünkü tek değişkenli normal dağılım sonuçları çok değişkenli normal dağılım ile ilgili yorum yapmak için yeterli değildir. Bu test için de tek değişkenli normal dağılımda olduğu gibi hipotez oluşturulmuştur. R programı aracılığıyla Henze-Zirkler's testi ile değerlendirilen çok değişkenli normal dağılım sonuçları ve Q-Q diyagramı Tablo 5.5. ve Şekil 5.5.'deki gibidir.

Tablo 5.5. Çok deęişkenli normal daęılım sonuçları.

Test	Deęişken	P deęeri	Normallik
Henze-Zirkler	x_1, \dots, x_6	0,4539	Evet

Tablo 5.5.'de görüldüğü gibi 6 deęişken birlikte ele alınarak deęerlendirilen çok deęişkenli normal daęılıma uygunluk test sonuçları $p > 0,05$ olduğundan varsayıma uygun olarak bulunmuştur.

Henze- Zirkler testini desteklemek için bir de serpilme diyagramı incelenmiştir. Bunun için 6 deęişkene ait 26697 örneğin Mahalanobis deęerleri küçükten büyüğe doğru sıralanmış ve her bir deęere karşılık χ^2 deęerleri bulunmuştur. Elde edilen Mahalanobis ve Ki- kare deęerleri arasındaki serpilme diyagramı Şekil 5.5.'de gösterilmiştir.



Şekil 5.5. Mahalanobis ve ki-kare deęerleri arasındaki serpilme grafięi.

Diyagram incelendiğinde çok deęişkenli normal daęılımı sağlayacak şekilde doęrusal bir ilişki elde edildiği görülmektedir

- Otokorelasyon Olmaması Varsayımı

Ürün üzerinde incelenen 6 deęişken için Ek 1-6 da yer alan Box-Ljung tablolarına göre, otokorelasyon varlığı Box-Ljung istatistięi ile aőađıdaki hipoteze göre test edilmiştir.

Değişkenlerin zamana bağımsızlığının incelendiği bu varsayımda SPSS Paket programı kullanılmıştır. Her değişken için aşağıdaki hipotezin geçerliliği kontrol edilmiştir.

H_0 : Değişkenin ardışık değerleri arasında otokorelasyon yoktur.

H_1 : Değişkenin ardışık değerleri arasında otokorelasyon vardır.

Hipotezlerin geçerliliği için α değeri 0,05 olarak belirlenmiştir. Ek 1-6 da sunulan değerler incelendiğinde 6 değişken için de α değeri $>0,05$ olduğu için H_0 hipotezi kabul edilerek problem için otokorelasyon sorunun olmadığı görülmüştür.

Sonuç olarak, sürecin kalite açısından izlemek ve kontrol için Hotelling T^2 kontrol diyagramının uygunluğunu belirleyen bütün varsayımların sağlandığı görülmüştür.

5.2.2.2. Hotelling T^2 kontrol diyagramının uygulanması

Çalışmada her seferinde tek birimin alındığı ve altı değişkenin değerlendirildiği 26697 örnek için Hotelling T^2 kontrol diyagramı oluşturulmuştur.

26697 adet geçmiş veri için uygulanan Hotelling T^2 diyagramının 1.aşamasının sonuçları Şekil 5.6.'da görülmektedir. Diyagram üzerinde görülen her nokta o örneğe ait altı tane parametrenin ölçüm değerinin birleşik değeridir.

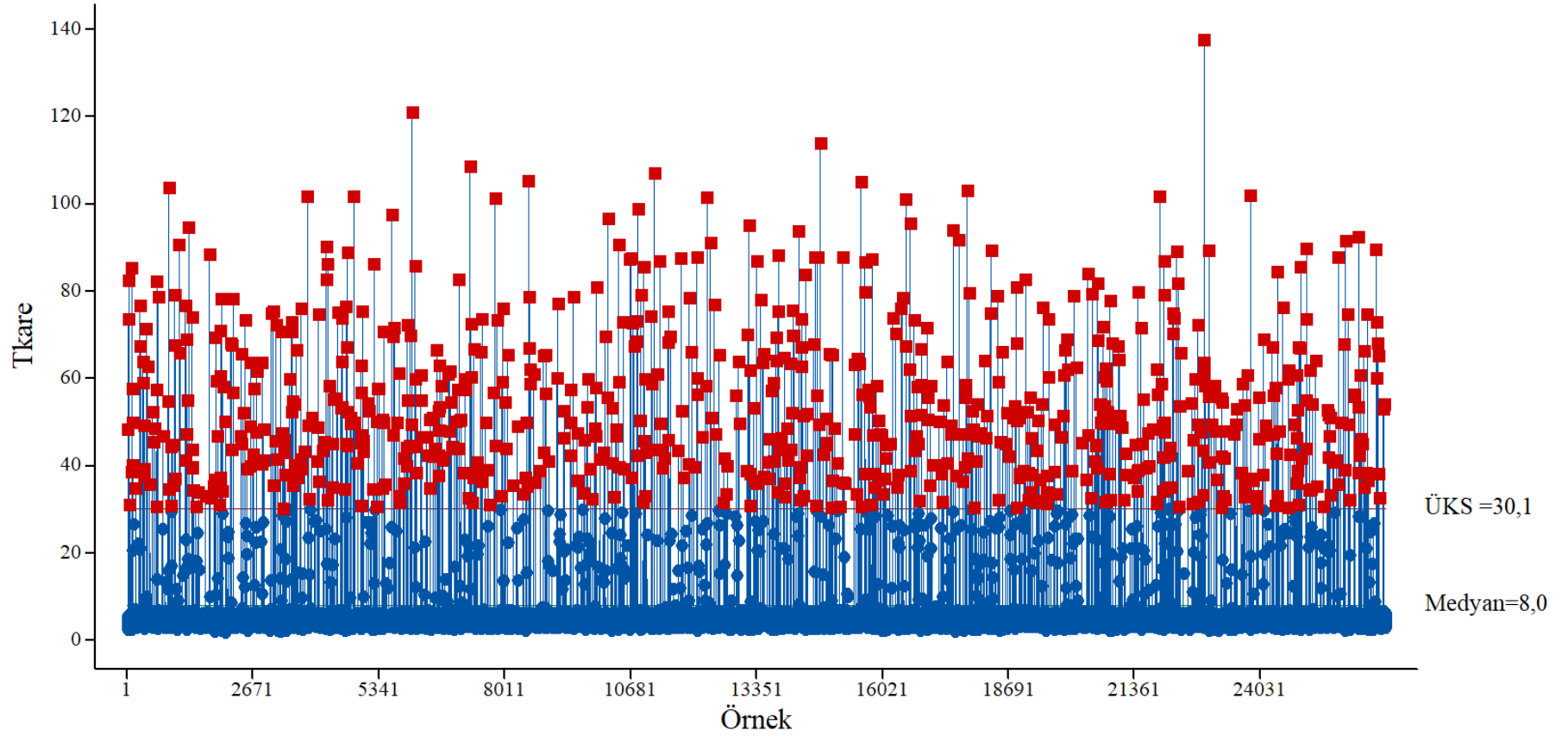
Çok değişkenli kontrol diyagramlarının alt kontrol sınırı yoktur. Üst kontrol sınırı ise teorik bölümde gösterilen Eşitlik 3.4 ile hesaplanmış ve 30,1 olarak bulunmuştur. Çalışmada kullanılan veri seti için ÜKS değeri 30,1 olarak elde edilmiştir.

Sıfır çizgisi ile 30,1 arasında kalan T^2 değerleri kontrol sınırları altında kabul edilen örneklerdir. Buna karşılık ÜKS değerini aşan örnekler kontrol dışı durumları oluşturmaktadır. Mevcut durumda 25893 adet örneğin kontrol sınırları içerisinde 804 adet örneğin ise kontrol sınırları dışarısında olduğu görülmüştür.

İlk 50 örneğe ait ölçüm değerlerinden oluşan T^2 değerleri Tablo 5.6.'da örnek olarak sunulmuştur. Tablodaki değerler incelendiğinde 50 örnek arasında 18, 27 ve 38 numaralı üç örneğin ÜKS değerini geçtiği bundan dolayı kontrol sınırı dışı durum oluşturduğu görülmüştür. Geriye kalan 47 örnek ÜKS altında kalmıştır bu nedenle incelenmesi gerekmemektedir.

Tablo 5.6. Hotelling T^2 değerleri.

Örnek No	T^2	Örnek No	T^2	Örnek No	T^2	Örnek No	T^2	Örnek No	T^2
1	3,247	11	3,962	21	4,492	31	4,313	41	3,639
2	3,949	12	4,115	22	3,250	32	5,312	42	3,989
3	4,592	13	2,957	23	4,760	33	4,922	43	3,789
4	3,452	14	3,434	24	4,130	34	5,484	44	2,971
5	3,721	15	2,202	25	3,721	35	4,684	45	5,910
6	4,778	16	5,403	26	4,457	36	4,472	46	4,310
7	4,565	17	3,463	27	73,473	37	5,001	47	3,747
8	5,112	18	48,205	28	4,893	38	82,395	48	4,471
9	5,116	19	5,141	29	4,941	39	3,186	49	4,021
10	4,520	20	5,180	30	4,251	40	4,499	50	5,044



Şekil 5.6. Hotelling T^2 sonuçları.

5.2.3. Kontrol dışı durumlara neden olan değişkenlerin tespit edilmesi

Çok değişkenli kontrol diyagramında görülen kontrol dışı durumların düzeltilebilmesi için bunların hangi değişkenden kaynaklandığının belirlenmesi gerekmektedir. Ancak Hotelling T^2 kontrol diyagramı ile bu mümkün olmadığı için MYT ayrıştırma yöntemi kullanılmıştır. Veri setindeki kontrol dışı durumlara neden olan değişkenler bir önceki bölümde anlatılan Eşitlik 3.19 -3.24'e göre belirlenmiştir.

Kontrol dışı 804 örnekten seçilen 10 örneğe ait 6 değişken için MYT ayrıştırma yöntemi uygulama sonuçları aşağıdaki Tablo 5.7.'deki gibidir. Tabloda kontrol dışı duruma etki eden koşulsuz parçalar (T_j^2) ve sonuç değerleri ile kontrol dışı duruma neden olan değişkenler görülmektedir.

Tablo 5.7. MYT koşulsuz parça T^2 değerleri ve kontrol dışı durumlar.

Örnek No	T_1^2	T_2^2	T_3^2	T_4^2	T_5^2	T_6^2	Sonuç
19	9,818	0,298	0,560	0,272	2,978	35,053	100001
27	1,141	36,305	9,005	25,855	1,093	0,099	011100
38	24,641	6,589	11,685	0,704	8,655	31,180	111011
52	17,048	0,883	6,350	0,333	0,787	5,257	101001
92	1,094	6,014	27,093	0,760	2,775	0,421	011000
95	0,252	1,093	0,486	12,159	33,232	38,175	000111
109	0,575	0,509	15,589	15,671	11,380	14,084	001111
110	6,452	0,848	10,892	0,354	13,798	16,987	101011
115	4,532	0,640	29,088	0,894	3,802	1,148	101000
175	0,686	0,735	10,161	21,790	0,879	1,049	001100

Tablo 5.7.'de sonuçlarda gösterilen ifade de; 0 kontrol altı olan değişkenleri 1 ise kontrol dışı duruma neden olan değişkenleri temsil etmektedir. Örneğin; 100001 durumu; kontrol dışı durumun x_1 ve x_6 değişkenlerinden dolayı oluştuğu, diğer değişkenlerin ise hesaplanan eşik değerine göre kontrol sınırı içerisinde kaldığı ve kontrol dışı duruma etkili olmadığı anlamına gelmektedir.

Bu çalışmada sürece etki eden 6 parametre olmasından dolayı $(2^6 - 1) = 63$ tane olası kontrol dışı durum mevcuttur [12]. İncelenen veri setinde rastlanan bu olası kontrol dışı durumlarının görüldüğü örnek adetleri Tablo 5.8.'de gösterilmiştir.

Tablo 5.8. Kontrol dışı durumları ve karşılaşılan örnek sayıları.

Durum	Örnek sayısı (adet)	Durum	Örnek sayısı (adet)	Durum	Örnek sayısı (adet)
000001	14	010110	5	101011	5
000010	4	010111	12	101100	10
000011	7	011000	24	101101	10
000100	17	011001	20	101110	5
000101	21	011010	5	101111	7
000110	10	011011	5	110000	12
000111	13	011100	11	110001	16
001000	18	011101	10	110010	7
001001	30	011110	11	110011	2
001010	8	011111	3	110100	16
001011	14	100000	8	110101	13
001100	30	100001	15	110110	5
001101	23	100010	12	110111	3
001110	11	100011	11	111000	12
001111	10	100100	100	111001	13
010000	8	100101	25	111010	3
010001	23	100110	5	111011	7
010010	10	100111	4	111100	13
010011	5	101000	13	111101	9
010100	16	101001	15	111110	8
010101	14	101010	8	111111	0

Tablo incelendiğinde olası durumlar içerisinde 111111 yani tüm değişkenlerin kontrol dışı durum üzerinde etkili olduğu bir duruma rastlanmamıştır. Bu nedenle veri setinde 62 tane farklı kontrol dışı durumun mevcut olduğu görülmektedir. Toplamda 804 tane kontrol dışı durum bulunmaktadır bunlar içerisinde 100 adet ile en çok 100100, yani x_1 ve x_4 değişkenlerinden dolayı oluşan kontrol dışı duruma rastlanmıştır. En az karşılaşılan kontrol dışı durumları ise 3'er adet örnek ile 011111, 110111 ve 111010 durumlarıdır.

Kontrol dışı durumlar çalışmanın bundan sonraki bölümlerinde sınıf olarak adlandırılacaktır.

5.2.4. Veri seti oluşturma

Bu aşama da girdi ve çıktı veri seti karşılıklı şekilde birleştirilmektedir. Girdi veri seti süreçten toplanan örneklere ait ölçüm değerlerinden oluşmaktadır. Çıktı veri seti ise önceki adımlarda yapılan hesaplamalar sonucunda elde edilen kontrol dışı durumlara neden olan değişkenleri ifade eden sınıflardan oluşmaktadır.

Tablo 5.8.'deki deęerler incelendięinde 62 tane sınıfın mevcut olduęu ve sınıflara ait örnek sayılarının da 3 ila 100 arasında deęişiklik gösterdięi görülmüştür. Bu durum sınıflar arasında örnek sayıları açısından dengesiz bir veri seti oluşturmaktadır. Dengesiz veri setlerinin sınıflandırma doğruluklarında başarıları olumsuz etkileyeceęinden dolayı sınıf sayılarının dengelenmesinin daha doğru olacaęı düşünülmektedir. Çalışmada gerçek veri kullanıldığı için en yüksek örnek sayısına sahip sınıftaki örnek sayısı olan 100 sayısı referans alınarak smote yöntemi yardımıyla tüm sınıflar için 100'er tane örnek tamamlanacak şekilde sentetik veri üretilmiştir. Çalışmaya 62 kontrol dışı sınıfa ait 6200 adet veri ile devam edilmiştir.

5.2.5. Önerilen modelin geliştirilmesi

Çalışmanın bu aşamasında sürece yeni gelen bir örneğin sınıfının tahmin edilmesini amaçlayan makine öğrenme algoritmalarını temel alan bir model geliştirilmiştir. Makine öğrenme algoritmaları kullanılarak geliştirilen bu model; mevcut durumları sınıflandırarak öğrenmekte ve bu sayede yeni durumlardaki kontrol dışı duruma neden olan deęişkenlerin sınıfını tahmin etmektedir. Yöntemin sınıflandırma doğruluęu ne kadar yüksekse sınıf tahmini o kadar doğru yapılabilmektedir. Bu nedenle algoritmaları kullanırken bakılması gereken temel konu, en uygun parametreler kullanılarak en yüksek doğruluęu yakalayabilmektir. Tüm algoritmalara ait parametrelerin uygun deęerleri sezgisel olarak denenerek bulunmuştur. Ayrıca geliştirilen modelde algoritmaların başarı oranlarını arttırmayı saęlayan topluluk makine öğrenme algoritmaları temel alınmıştır.

Topluluk algoritmalarına geçilmeden önce temel algoritmalar tekli olarak denenerek performansı en yüksek olan algoritma seçilerek bu algoritmanın topluluk yöntemleriyle birleştirilmesi hedeflenmektedir. Bu nedenle ilk olarak temel 5 tekli makine öğrenme algoritması ile veri seti sınıflandırılmış ve performansları ölçülmüştür. Daha sonra topluluk algoritmalarına geçiş yapılmıştır.

5.2.5.1. Makine öğrenme algoritmalarının tekli kullanılması

Bu bölümde temel makine öğrenme algoritmalarının tek olarak kullanılması durumunda, sınıflandırma performanslarını en iyi yapan parametreler, veri setine ait ön bilgiler göz önünde bulundurularak, sezgisel olarak tahmin edilmiştir. Modeller her parametreye göre tekrar tasarlanarak eğitilmiş ve sonuçları elde edilmiştir. Uygun parametre değerlerinin belirlenmesi ile algoritmalarından elde edilen başarı oranlarının karşılaştırması yapılmıştır.

Sınıflandırma algoritmaları ile kurulan modellerin eğitim aşamaları için çapraz doğrulama yöntemi kullanılmıştır. Çapraz doğrulama parametreleri, temel tekli algoritmalara göre denenerek, ortak olarak performansları yükselttiğine karar verilen değerlerle belirlenmiştir. Çalışmanın tutarlı olması adına tüm modeller aynı parametreler kullanılarak eğitilmiştir.

Çapraz doğrulama için kullanılan parametreler Tablo 5.9.'da gösterilmektedir.

Tablo 5.9. Sınıflandırma için çapraz doğrulama parametreleri.

Parametre	Değer
Katlama Sayısı	10
Örnekleme Türü	Otomatik

Katlama sayısı yapılan çalışmalarda sıklıkla 10 olarak alınmıştır [75, 101, 105–108]. Çalışmada da benzer şekilde katlama sayısı 10 değeri ile başarılı sonuçlar elde edilmiştir. Ayrıca örnekleme türü otomatik olarak seçilmiştir ve sonuç değerleri nominal olduğu için katlamalı örnekleme kullanılmıştır.

Tüm algoritmalar için belirlenen parametre değerleri kullanılarak elde edilen sınıflandırma başarı oranlarını değerlendirmek için çok sınıflı performans kriterleri kullanılmıştır. Bunlar; doğruluk, sınıflandırma hatası, kappa, ağırlıklı ortalama duyarlılık ve ağırlıklı ortalama kesinliktir.

Kullanılan 5 temel makine öğrenme algoritmasının performanslarını en iyi yapan parametreler ve sonuç değerleri aşağıda sunulmuştur.

– Karar ağacı algoritması

Karar ağacı algoritması için kullanılan parametrelerden bazıları Tablo 5.10.'da gösterilmektedir. Ağaçtaki bölme işlemi için daha önce yapılmış çalışmalara benzer olarak [109, 110] niteliklerin seçileceği kriter, entropiyi hesaplayarak en az değerli olanı bölünme kriteri olarak seçen, bilgi kazancı olarak belirlenmiştir. Maksimum derinlik ağacın derinliğini kısıtlamak için kullanılır ve örnek veri setinin büyüklüğüne göre değişmektedir. 0-30 arası 31 değer denenerek 20 olarak seçilmiştir. Güven seviyesi ağacı budamanın kötümser hata hesaplamasında kullanılan bir parametredir. Minimum kazanç, bir düğümün kazancını ifade eder ve düğümlerin bölünmesi kazanç değerlerine bağlıdır. 0-1 arası 0,1 adım artışla 11 deneme yapılarak seçilmiştir. Bu değerlerin asgari kazançtan büyük olması istenmektedir. Düşük değerler az bölünmeye çok yüksek değerler de bölünmenin durmasına neden olmaktadır. Birkaç kriter ise programın varsayılan değerleri ile çalıştırılmıştır.

Tablo 5.10. Karar ağacı parametreleri.

Parametre	Değer
Kriter	Bilgi kazancı
Maksimum derinlik	20
Güven seviyesi	0,1
Min.kazanç	0,1

Bu parametrelere göre uygulanan karar ağacı algoritması için performans kriterlerine ait sonuçlar Tablo 5.11.'de gösterilmiştir.

Tablo 5.11. Karar ağacı performans değerleri.

Kriter	Değer
Doğruluk	% 93,74
Sınıflandırma Hatası	% 6,26
Kappa	% 0,936
Ağırlıklı Ortalama Duyarlılık	% 93,74
Ağırlıklı Ortalama Kesinlik	% 94,48

Tablo genel olarak incelendiğinde algoritmanın %93,74 doğruluk, %94,48 kesinlik ve %6,26 hata oranı ile performans gösterdiği görülmektedir. Ayrıca kappa 0,936 değeri ile 0,8-1.00 arasında olduğundan dolayı çok iyi tahmin olarak değerlendirilmektedir.

– K-NN algoritması

Sınıflandırma için kullanılan en yakın komşu sayısı (k) denenerek belirlenmiştir. Daha önce bahsedildiği gibi sınıflara sağlıklı şekilde atama yapabilmek için k değeri tek sayı olarak seçilmelidir. Çalışmada k değeri ilk olarak 1 seçilerek denemeye başlanmıştır. Tablo 5.12.'de görüldüğü gibi doğruluk oranlarına göre, en yüksek performans k değeri 3 iken elde edilmiştir. Doğruluk oranı k=9 olana kadar sabit kalarak sonra azalmaya başladığından dolayı 7 deneme arasından ilk yüksek değere rastlanmasından dolayı k=3 olarak alınmıştır.

Tablo 5.12. K-NN k parametrelerine göre performans değerleri.

k	Doğruluk oranı
1	%86,53
3	%88,85
5	%88,85
7	%88,85
9	%88,85
11	%88,74
13	%86,53

En yakın komşuları tespit etmek için kullanılan ölçü türü parametresi, veri seti sayısal değerler içerdiğinden dolayı sayısal ölçümler ve mesafe uzaklığı da en sık kullanılan uzaklık çeşidi olması sebebiyle [111] Öklid uzaklığı olarak seçilmiştir. K-NN için kullanılan parametreler Tablo 5.13.'de gösterilmektedir.

Tablo 5.13. K-NN algoritması parametreleri.

Parametre	Değer
K	3
Ölçü türü	Sayısal
Mesafe ölçümü	Öklid uzaklığı

Bu parametrelere göre uygulanan K-NN algoritması için performans kriterlerine ait sonuçlar Tablo 5.14.'de gösterilmiştir.

Tablo 5.14. K-NN performans değerleri.

Kriter	Değer
Doğruluk	% 88,85
Sınıflandırma Hatası	% 11,15
Kappa	% 0,887
Ağırlıklı Ortalama Duyarlılık	% 88,85
Ağırlıklı Ortalama Kesinlik	% 89,95

Tablo genel olarak incelendiğinde algoritmanın %88,85 doğruluk, %89,95 kesinlik ve %11,15 hata oranı ile performans gösterdiği görülmektedir. Ayrıca kappa 0,887 değeri ile 0,8-1.00 aralığında olduğundan dolayı çok iyi tahmin olarak değerlendirilmektedir

– Naif bayes algoritması

Naif bayes algoritmasında uygulamasında sınıflandırma, sadece tek parametre, Laplace korelasyonlara göre yapılmıştır bunun dışında herhangi bir parametre bulunmamaktadır [110].

Bu parametrelere göre uygulanan NB algoritması için performans kriterlerine ait sonuçlar Tablo 5.15.'de gösterilmiştir.

Tablo 5.15. NB performans değerleri.

Kriter	Değer
Doğruluk	%75.73
Sınıflandırma Hatası	%24.27
Kappa	%0.753
Ağırlıklı Ortalama Duyarlılık	%75.73
Ağırlıklı Ortalama Kesinlik	%78.70

Tablo genel olarak incelendiğinde algoritmanın %75,73 doğruluk, %78,70 kesinlik ve %24,27 hata oranı ile performans gösterdiği görülmektedir. Ayrıca kappa 0,753 değeri ile 0,6-0,8 aralığında olduğundan dolayı iyi tahmin olarak değerlendirilmektedir.

– Çok sınıflı destek vektör makinesi algoritması

Çalışmada ele alınan süreç çok sınıflı olduğu için Ç-DVM algoritmasının türü, bu duruma uygun olarak çok sınıflı olarak belirlendi. Sınıflandırma için Du ve arkadaşlarının çalışmasında başarısı kanıtlanmış olan bire karşı bir yaklaşımı kullanılmıştır [48]. Kernel yani çekirdek işlevinin türü de, geçmiş çalışmalar [48, 102, 112, 113] ve veri seti yapısı göz önünde bulundurularak belirlenmiştir. Sezgisel olarak denenmiş ve en yüksek doğrulukları sağlayan çekirdek işlev türüne radyal tabanlı fonksiyon olarak karar verilmiştir. Diğer parametreler programın varsayılan değerleri

ile çalıştırılmıştır. Ç-DVM algoritmasının parametreleri Tablo 5.16.'da görüldüğü gibidir.

Tablo 5.16. Ç-DVM algoritması.

Parametre	Değer
SVM yaklaşımı	Bire karşı bir
Kernel Türü	Radyal tabanlı fonksiyon

Bu parametrelere göre uygulanan Ç-DVM algoritması için performans kriterlerine ait sonuçlar Tablo 5.17.'de gösterilmiştir.

Tablo 5.17. Ç-DVM performans değerleri.

Kriter	Değer
Doğruluk	%90,58
Sınıflandırma Hatası	%9,42
Kappa	%0,781
Ağırlıklı Ortalama Duyarlılık	%90,58
Ağırlıklı Ortalama Kesinlik	%90,51

Tablo genel olarak incelendiğinde algoritmanın %90,58 doğruluk, %90,51 kesinlik ve %9,42 hata oranı ile performans gösterdiği görülmektedir. Ayrıca kappa 0,781 değeri ile 0,8-1.00 aralığında olduğundan dolayı iyi tahmin olarak değerlendirilmektedir

– Yapay Sinir Ağları Algoritması

Çalışmada, benzer çalışmalarda olduğu gibi [9, 12, 114], ileri beslemeli geri yayımlı çok katmanlı algılayıcı sinir ağı kullanılmıştır. Ağ yapısında, değişkenlerden oluşan 6 girdi, 100 er nöron içeren 2 tane gizli katman ve sınıflardan oluşan 62 tane çıktı mevcuttur. Sinir ağına ait kullanılan parametreler Tablo 5.18.'de gösterilmiştir.

Sınıflandırma ve tahmin çalışmalarında olduğu gibi aktivasyon fonksiyonu sigmoid olarak kullanılmıştır [38, 53, 115]. Sinir ağlarında hata fonksiyonun değerini azaltmak için her bağlantının ağırlığı güncellenir. Bu işlemin kaç kez tekrarlanacağı eğitim döngüsü parametresi kullanılarak 200-500 aralığında 50 birimde bir olacak şekilde 7 kez denenmiş ve 500 olarak belirlenmiştir. Öğrenme oranı ve diğer parametreler ise varsayılan değerler kabul edilerek kullanılmıştır [44].

Tablo 5.18. YSA algoritması parametreleri.

Parametre	Değer
Aktivasyon fonksiyonu	Sigmoid
Eğitim Döngüsü	500
Öğrenme oranı	0,01

Bu parametrelere göre uygulanan YSA algoritması için performans kriterlerine ait sonuçlar Tablo 5.19.'da gösterilmiştir.

Tablo 5.19. YSA performans değerleri.

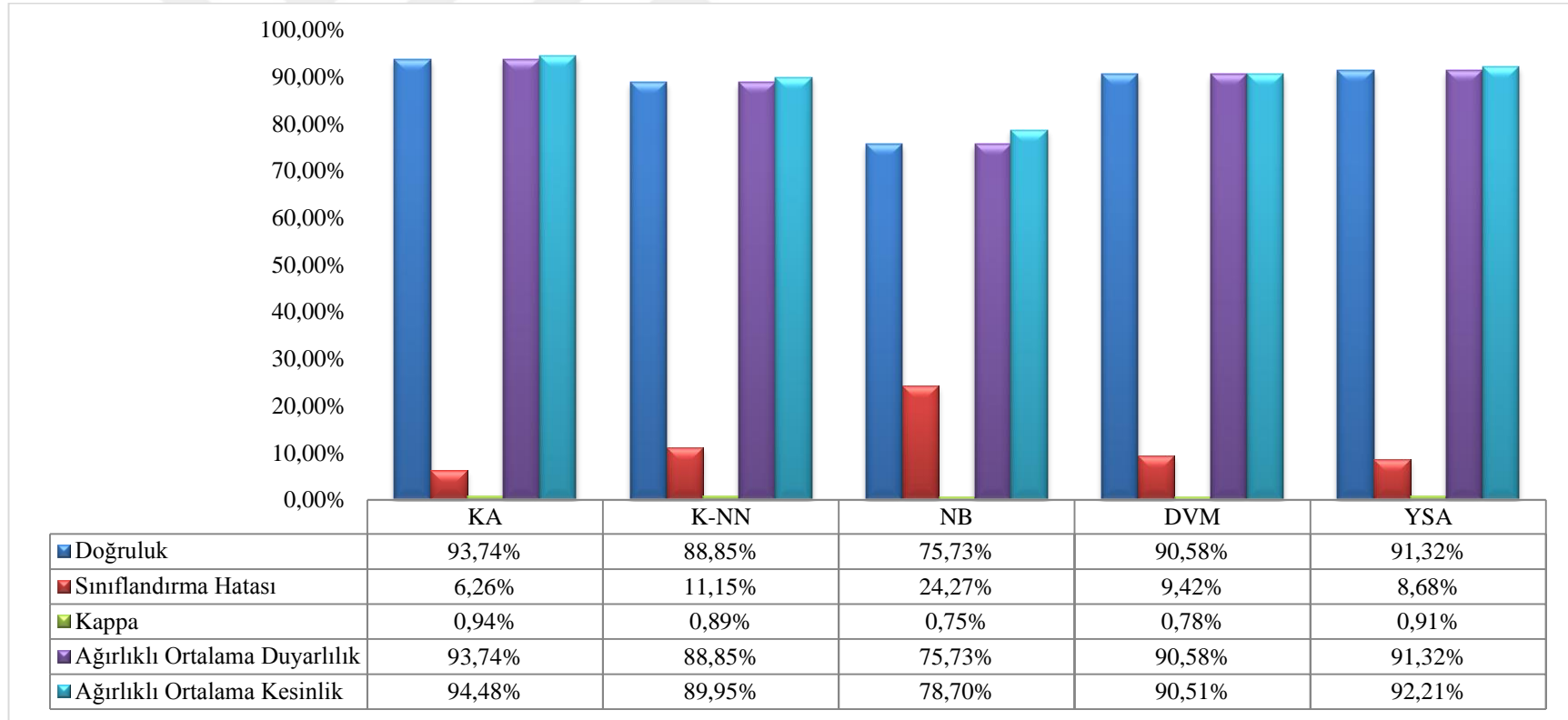
Kriter	Değer
Doğruluk	%91.32
Sınıflandırma Hatası	%8.68
Kappa	%0.912
Ağırlıklı Ortalama Duyarlılık	%91.32
Ağırlıklı Ortalama Kesinlik	%92.21

Tablo genel olarak incelendiğinde algoritmanın %91,32 doğruluk, %92,21 kesinlik ve %8,68 hata oranı ile performans gösterdiği görülmektedir. Ayrıca kappa 0,912 değeri ile 0,8-1.00 aralığında olduğundan dolayı çok iyi tahmin olarak değerlendirilmektedir

Temel 5 makine öğrenme algoritmasının performansları karşılaştırılırsa; karar ağaçlarının %93,74 doğruluk ile sınıflandırma için en başarılı algoritma olduğu görülmektedir. Ardından sırasıyla, %91,32 doğruluk ile YSA algoritması, 90,58% doğruluk ile Ç-DVM algoritması, %88,85 doğruluk ile K-NN algoritması son olarak da %75,73 doğruluk ile NB algoritması gelmektedir.

Algoritmaların performans kriter değerleri Şekil 5.7.'de özetlenmektedir. Grafik incelendiğinde; hata oranı, kappa, ağırlıklı ortalama duyarlılık ve ağırlıklı ortalama kesinlik kriterlerine göre de karar ağaçları algoritmasının, literatürdeki diğer bazı çalışmalara benzer olarak [45, 116–119], en başarılı sınıflandırma algoritması olduğu ve diğer 4 algoritmaya göre üstün olduğu görülmektedir.

Sonuç olarak çalışmada en başarılı olarak görülen karar ağaçları 5 temel algoritma arasından çalışmanın temel sınıflandırma algoritması olarak seçilmiştir.



Şekil 5.7. Tekli Algoritmaların performans karşılaştırılması.

5.2.5.2. Seçilen makine öğrenme algoritmasının topluluk yöntemleri ile birleştirilmesi

Tekli algoritmalar arasından en yüksek sınıflandırma başarısı ile seçilen karar ağaçları algoritmasının daha önce belirlenmiş parametreler kullanılarak torbalama ve yükseltme yöntemleri ile birleştirilerek performansları belirlenmiştir. Torbalama ve yükseltme algoritmalarının tekrar sayıları ise varsayılan değer olan 10 olarak alınmıştır.

– Torbalama topluluk yöntemi ile birleştirilmesi

Algoritmaların paralel şekilde birleştirildiği yöntem olan Torbalama algoritmasının karar ağacını 10 tekrarlı şekilde birleştirilmesi ile elde edilen sonuçlar Tablo 5.20.'de gösterilmektedir.

Tablo 5.20. Karar ağacı-torbalama performans değerleri.

Kriter	Değer
Doğruluk	% 94,97
Sınıflandırma Hatası	% 5,03
Kappa	% 0,949
Ağırlıklı Ortalama Duyarlılık	% 94,97
Ağırlıklı Ortalama Kesinlik	% 95,46

Sonuçlar incelendiğinde karar ağaçları torbalama yöntemi ile birleştirildiğinde sınıflandırma doğruluğu %93,74 den %94.97 ye yükselmiştir. Benzer şekilde hata oranının da %6,26 dan %5,03 e azaldığı görülmektedir.

– Adaboost topluluk yöntemi ile birleştirilmesi

Algoritmaların sıralı şekilde birleştirildiği yöntem olan Adaboost algoritması ile karar ağacını 10 tekrarlı olarak birleştirilmesi ile elde edilen sonuçlar Tablo 5.21.'de gösterilmektedir.

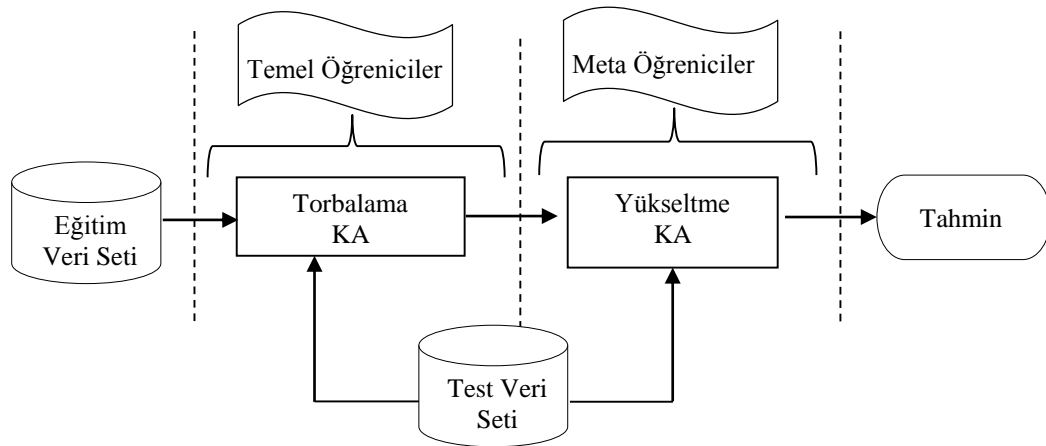
Tablo 5.21. Karar ağacı-adaboost performans değerleri.

Kriter	Değer
Doğruluk	% 95,08
Sınıflandırma Hatası	% 4,92
Kappa	% 0,950
Ağırlıklı Ortalama Duyarlılık	% 95,08
Ağırlıklı Ortalama Kesinlik	% 95,56

Sonuçlar incelendiğinde karar ağaçlarının adaboost yöntemi ile sıralı şekilde birleştirilmesinin doğrulukları arttırdığı görülmektedir. Karar ağaçları tekli kullanılması durumunda doğruluk oranları %93,74 ve paralel şekilde birleştirilerek oluşturulan torbalama yöntemi ile %94,97 iken adaboost yöntemi ile %95.08 doğruluk oranına ulaşılmıştır.

- Topluluk yöntemlerinin yığılmış genelleme topluluk yöntemi ile birleştirilmesi

İki yöntemden farklı çalışma prensibi olan yığılmış genelleme yönteminde farklı türdeki sınıflandırma algoritmaları sıralı olarak birleştirilmektedir. Çalışmanın modeli Karar ağacı-Torbalama ve Karar ağacı-Adaboost topluluk algoritmalarının birleştirilmesi ile oluşmaktadır. Bu bilgiler ışığında önerilen model yapısı Şekil 5.8.'de sunulmuştur.



Şekil 5.8. Çalışmanın önerilen modeli.

5.2.6. Performans deęerlendirmesi

Tekli karar aęacı algoritmasının topluluk algoritmalarıyla birleřtirilmesi sonucunda sınıflandırma performanslarının arttığı görülmüřtür. Bu iyileřtirmelerin olduęu topluluk algoritmalarının yığılmıř genelleme topluluk yöntemi ile bir arada kullanılmasının performansları daha da iyileřtireceęi düşünülerek bir model geliřtirilmiřtir.

Bu model sonucunda elde edilen sınıflandırma performansları Tablo 5.22.'de gösterilmektedir.

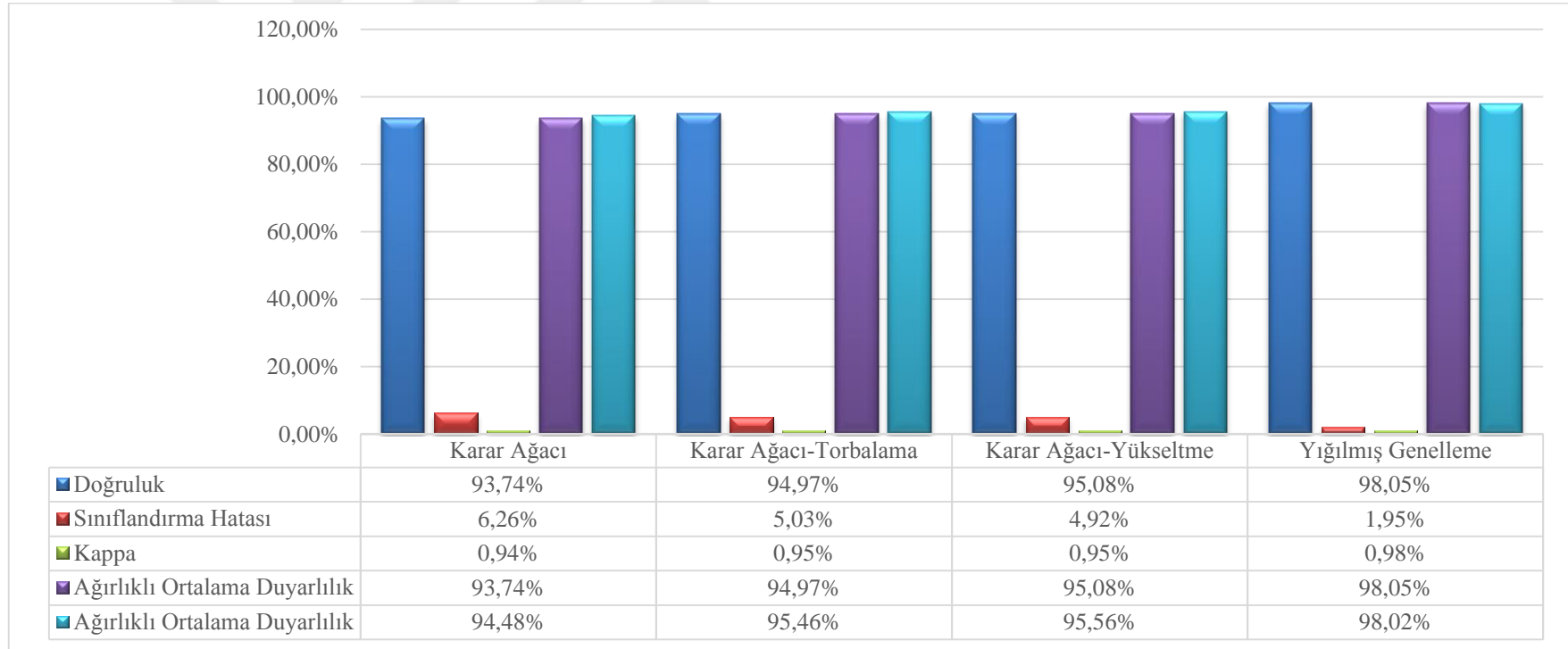
Tablo 5.22. Yığılmıř genelleme performans deęerleri.

Kriter	Deęer
Doęruluk	% 98,06
Sınıflandırma Hatası	% 1,94
Kappa	% 0,980
Aęırlıklı Ortalama Duyarlılık	% 98,06
Aęırlıklı Ortalama Kesinlik	% 98,27

Karar aęacı algoritmasının tekli olarak, topluluk algoritmalarının sıralı ve paralel olarak birleřtirilmesi ile ve topluluk algoritmalarını birleřtirilmesi sonucunda elde edilen sınıflandırma performansları Őekil 5.9.'da gösterilmektedir.

Sonuçları deęerlendirerek performansların topluluk algoritmalarıyla ayrı ayrı ve birlikte kullanılmasının performansları git gide arttırdığı görülmüřtür. Karar aęacı algoritması tekli kullanılırken sınıflandırma doęruluęu 93,74% iken, karar aęacı-torbalama 94,97%, karar aęacı-yükseltme 95,08% ve en bařarı olarak görülen yığılmıř genelleme yöntemi ile oluřturulan modelin doęruluk performansı ise; 98,06% olarak elde edilmiřtir.

Böylece geliřtirilen modelin daha yüksek doęrulukla sınıflandırma yapma yeteneęine sahip olması fikri gerçekleřtirilmiř olmuřtur.



Şekil 5.9. Önerilen modelin performansının diğer modeller ile karşılaştırılması.

BÖLÜM 6. TARTIŞMA VE SONUÇ

Tek deęişkenli kontrol diyagramlarında kolaylıkla tespit edilebilen kontrol dıőı durumlara neden olan deęişkenler süreç çok deęişkenli olduęunda, sadece kontrol diyagramları kullanılarak tespit edilememektedir. Bu durumda kalitenin izlenmesi ve deęerlendirilmesi için kullanılan çok deęişkenli kontrol diyagramları karşılaőılan kontrol dıőı duruma neden olan deęişkenler ile ilgili hiçbir bilgi vermemektedir. Bu noktada kontrol dıőı durumları analiz etmek için farklı yöntemler kullanılmalıdır.

Çalıőmada kontrol dıőı durumların nedenlerinin araőtırılması için makine öğrenme algoritmaları kullanılmıőtır. Algoritmaları eęitmek için kullanılan veri setinin çıktı deęerleri ya da örneklere karşılık gelen durumların nedenleri, MYT ayrıőtırma yöntemi ile belirlenmiőtir. Ancak MYT ayrıőtırma yöntemi hesaplama ačasından yoęundur ve kontrol dıőı sinyale neden olan deęişkenin tanısında deęişkenlerin büyük bir boyutu için sınırlı bir yeterlilięe sahiptir [30]. Ayrıca bu yöntem öğrenen bir yapıda deęildir sadece eldeki verilerden sonuç üretir. Çalıőmanın ileriye dönük önemli bir faydasının kurulan modelle kalite süreçlerine iyileőtirme önerileri için yol göstermek olduęu düşünöldüęünden makine öğrenme algoritmaları ile devam edilmiőtir. Bu sayede model kalite süreçlerinde kullanılacak akıllı öneri sistemlerine entegre olabilecektir.

Bu nedenlerle çalıőmada kullanılan makine öğrenme algoritmalarının doęru tanı koyabilmeleri için performanslarının en yüksek düzeyde olması istenmektedir. Bunu saęlayabilmek için topluluk makine öğrenme yöntemlerine başvurulmuőtur. Farklı algoritmaları birleőtirmeyi saęlayan bir topluluk yöntemi olan Yıęılmış genelleme algoritması ile topluluk algoritmaları birleőtirilmiőtir. Torbalama ve Yükseltme topluluk yöntemleri ile tekli algoritmanın birleőtirilerek iyileőtirilmesi saęlanmış, ardından iyileőtirilmiő iki yöntem birleőtirilmiőtir. Bu sayede daha da başarılı sonuçlar elde edilmiőtir.

Geliştirilen modelde kullanılan yığılmış genelleme birleştirme yöntemine daha önce kalite kontrol alanında ya da kontrol dışı durumların nedenlerinin belirlenmesi konusunda bir çalışmada rastlanmamıştır. Bu açıdan yeniliği olan bu modelin temel amacı, başarılı olan topluluk yöntemlerini birleştirmek ve başarıyı arttırmaktır.

Hidrolik pompa kapağının döküm sürecindeki kontrol dışı durumların nedenlerinin tespit edilmesi ile ilgili yapılan uygulama çalışmasının sonuçlarında elde edilen yüksek başarı oranları ile amaçlanan hedefe ulaşılmıştır. Bu da çalışmanın faydası olarak düşünülmektedir. Benzer bir sürece uygulanabilir bir modeldir daha farklı süreçlerde kullanılacağı zaman parametreler ile ilgili değişiklikler gerekebilir.

Ayrıca bu katkılarının yanında geliştirilen model sayesinde çok değişkenli kontrol diyagramlarına gerek kalmadan yeni alınan örneklerin kontrol dışı olması durumunda hangi değişkenden kaynaklandığı tahmin edilebilecektir.

Döküm parçalarının istenilen özelliklere uymaması, CNC tezgâhına oturmaması gibi durumlarda parça hurdaya gönderileceğinden dolayı bu sorunlar yaşanmadan engellenmesi büyük önem taşımaktadır. Bu anlamda modelin uygulama yapılan firmaya katkıları değerlendirilirse; değişkenlerin önceden tespit edilerek ortadan kaldırılmasını sağlamak hem zaman hem maliyet hem de iş gücü açısından büyük faydalar sağlayacaktır denilebilir.

Gelecek çalışması olarak,

- Daha çok çeşitte temel algoritma ve bunların farklı birleştirilme çeşitleri ile kullanılması ile performansların yükseltilmesi sağlanabilir.
- Algoritmaların parametrelerinin sezgisel olarak denenerek bulunması yerine bir optimizasyon tekniği kullanarak iş gücü ve zamandan tasarruf edilmesi planlanmaktadır.
- Optimizasyon tekniği kullanarak sadece tekli değil topluluk algoritmalarının da parametrelerinin belirlenmesi sağlanabilir.

- Bu çalışmada da olduğu gibi literatürdeki diğer çalışmalarda da kontrol dışı durumlara neden olan değişkenler belirlenerek çalışmalar sonlandırılmış ancak devamında iyileştirme önerileri sunan bir model bulunmamaktadır. Önerilen modelin çıktılarını kullanarak devamına uzman sistemle bütünleşik bir kalite tespit uyarı ve öneri sistemi eklenebilir.
- Değişkenlere ait kesin olan değerleri değiştirmek için bulanık çıkarım sistemi ile entegre bir şekilde kullanılabilir.



KAYNAKÇA

- [1] Birgören, B., İstatistiksel Kalite Kontrolü, 1. baskı. Nobel Akademik, 16-17, 2015.
- [2] Robert J. C. Y., Mason L., Multivariate Statistical Process Control With Industrial Applications, Society for Industrial and Applied Mathematics, 5-17, 2002.
- [3] Montgomery D. C., Introduction to Statistical Quality Control, 6. Baskı. John Wiley & Sons, 499-507, 2009.
- [4] Hotelling H., Multivariable quality control—illustrated by the air testing of sample bombsight, McGraw Hill, 111-184, 1947.
- [5] Woodall W. H., Ncube M. M., Multivariate CUSUM quality-control procedures, *technometrics*, 27(3), 285–292, 1985.
- [6] Lowry, C. A., Woodall, W. H., Champ, C. W., Rigdon, S. E., A multivariate exponentially weighted moving average control chart, *Technometrics*, 34(1), 46–53, 1992.
- [7] Fuchs, C., Kenett, R., Multivariate Quality Control Theory Applications, 1. baskı. Marcel Dekker, 11, 1998.
- [8] Çetin, S., Birgören B., Çok değişkenli kalite kontrol çizelgelerinin döküm sanayiinde uygulanması, *Gazi Üniv. Müh. Mim. Fak. Der.*, 22(4), 809–818, 2007.
- [9] Aparisi, F., Avendaño, G., ve Sanz, J., Techniques to interpret T^2 control chart signals, *IIE Trans., Institute Ind. Eng.*, 38(8), 647–657, 2006.
- [10] Salmona, A. M. Ö., Fırat, O.S.Ü., Bir endüstri işletmesinde çok değişkenli istatistiksel süreç kontrolü, 25–27, 2005.
- [11] Mason, R. L., Tracy, N. D., Young, J. C., Decomposition of T^2 for multivariate control chart interpretation, *J. Qual. Technol.*, 27(2), 99–108, 1995.
- [12] Niaki, S. T. A., Abbasi. B., Fault diagnosis in multivariate control charts using artificial neural networks, *Qual. Reliab. Eng. Int.*, 21(8), 825–840, 2005.

- [13] Jackson, J. E., Multivariate quality control, *Commun. Stat. Theory Methods*, 14(11), 2657–2688, 1985.
- [14] Rao, O. R. M., Subbaiah, K.V., Rao, K. N., Rao T. S., Application of multivariate control chart for improvement in quality of hotmeal-a case study, *Int. J. Qual. Res.*, 7(4), 623–640, 2013.
- [15] Mihalkó, J. ve Rajkó, R., Investigation of multivariate statistical process control in r enviroment, *Analecta Tech. Szeged.*, 11(2), 36–40, 2017.
- [16] Güler, Z. Ö., Bakır, M. A., Signal detection based on independent components in mmultivariate statistical quality control, *Econometrics Letters*, 5(2), 14-24, 2019.
- [17] Murphy, B. J., Selecting out of control variables with the T^2 multivariate quality control procedure, *Stat.*, 36(5), 571-581, 1987.
- [18] Pei, X., Yamashita, Y., Yoshida, M., ve S. Matsumoto, Discriminant analysis and control chart for the fault detection and identification, *Comput. Aided Chem. Eng.*, 21, 1281-1286, 2006.
- [19] Doganaksoy, N., Faltin, F. W., Tucker, W. T., Identification of out of control quality characteristics in a multivariate manufacturing environment, *Commun. Stat. - Theory Methods*, 20(9), 2775–2790, 1991.
- [20] Hawkins, D. M., Multivariate Quality Control Based on Regression-Adiusted Variables”, *Technometrics*, 33(1), 61–75, Şub. 1991.
- [21] Bersimis, S., Sgora, A., Psarakis, S., Methods for interpreting the out-of-control signal of multivariate control charts: A comparison study, *Qual. Reliab. Eng. Int.*, 33(8), 2295–2326, 2017.
- [22] Vidal-Puig, S., Ferrer, A., A comparative study of different methodologies for fault diagnosis in multivariate quality control, *Commun. Stat. Simul. Comput.*, 43(5), 2014
- [23] Özel, S. Çok değişkenli kalite kontrolün döküm sanayiinde uygulanması, Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü, Endüstri Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi, 34, 2005.
- [24] Mason R. L., Tracy N. D., Young, J. C., A practical approach for interpreting multivariate T^2 control chart signals, *J. Qual. Technol.*, 29(4), 369–501, 1997
- [25] Parra, M. G., P. Loaiza, R., Application of the multivariate T^2 control chart and the Mason Tracy Young decomposition procedure to the study of the consistency of impurity profiles of drug substances, *Qual. Eng.*, 16(1), 127–142, 2003.

- [26] Ulen, M., Demir, I., Application of multivariate statistical quality control in pharmaceutical industry, *Balk. J. Math.*,1, 93–105, 2013.
- [27] Boullosa, D., Larrabe, J. L., Lopez, A., Gomez M. A., Monitoring through T^2 Hotelling of cylinder lubrication process of marine diesel engine, *Appl. Therm. Eng.*, 110, 32–38, 2017.
- [28] Yılmaz, H., Çok değişkenli istatistiksel süreç kontrolü: Bir hastane uygulaması, İstanbul Teknik Üniversitesi, 2012.
- [29] Das, N., Prakash, V., Interpreting the out-of-control signal in multivariate control chart — a comparative study, *Int. J. Adv. Manuf. Technol.*, 37, 966–979, 2008.
- [30] Li, J., Jin, J., Shi, J., Causation-based T^2 decomposition for multivariate process monitoring and diagnosis, *J. Qual. Technol.*, 40 (1), 46–58, Oca. 2008.
- [31] Monarrez M. R. P., “Practical decomposition method for T^2 hotelling chart”, *Int. J. Ind. Eng. Theory Appl. Pract.*, 2056(5), 401–411, 2013.
- [32] Agog, N. S., Dikko, H. G., Asiribo, O. E., Determining out-of-control variable(s) in a multivariate quality control chart, *Sci. Africana*, 13(2), 266–280, 2014.
- [33] Akeem, A. A., Yahaya, A., Asiribo, O., Hotelling’s T^2 decomposition: approach for five process characteristics in a multivariate statistical process control, *Am. J. Theor. Appl. Stat.*, 4(6), 432–437, 2015.
- [34] Huang, X., Xu, J., Zhou, Q., Multi-scale diagnosis of spatial point interaction via decomposition of the k function-based T^2 statistic, *J. Qual. Technol.*, 49(3), 213–227, 2017.
- [35] Schaffer, J. R., Vandenhul, S., The decomposition of the mewma statistic, *Proceedings of the Annual Meeting of the American Statistical Association*, 1–6, 2001.
- [36] Muhammad, N., Nagwai, A. P., Asiribo, E. O., Yahaya, A., Detecting assignable signals via decomposition of mewma statistic, 4(1), 25–30, 2016.
- [37] Oktay, E., Orcanli, K., Döküm Sanayisinde Metal Alaşım Oranlarına Hotelling T^2 ve MEWMA Kontrol Grafikleri Uygulamaları”, *Social Sciences Research Journal*, 7 (1), 114-135, 2018.
- [38] Chen. L. H., Wang T. Y., Artificial neural networks to classify mean shifts from multivariate χ^2 chart signals, *Comput. Ind. Eng.*, 47(2–3), 195–205, 2004.

- [39] Cheng, C. S., Cheng, H. P., Identifying the source of variance shifts in the multivariate process using neural networks and support vector machines, *Expert Syst. Appl.*, 35(1–2), 198–206, 2008.
- [40] Aparisi, F., Sanz J., Interpreting the out-of-control signals of multivariate control charts employing neural networks, *World Acad. Sci. Eng. Technol.*, 61, 226–230, 2010.
- [41] Li, T., Hu, S., Wei, Z., Liao, Z., A Framework for Diagnosing the Out-of-Control Signals in Multivariate Process Using Optimized Support Vector Machines, *Math. Probl. Eng.*, 2013(2), 1–9, 2013.
- [42] Huda, S., Abdollahian, M., Mammadov, M., Yearwood, J., Ahmed S., Sultan I., A hybrid wrapper-filter approach to detect the source(s) of out-of-control signals in multivariate manufacturing process, *Eur. J. Oper. Res.*, 237(3), 857–870, 2014.
- [43] Song, H., Xu, Q., Yang, H., Fang, J., Interpreting out-of-control signals using instance-based bayesian classifier in multivariate statistical process control, *Commun. Stat. Simul. Comput.*, 46(1), 2017.
- [44] Shao, Y. E., Lin, S. C., Using a time delay neural network approach to diagnose the out-of-control signals for a multivariate normal process with variance shifts, *Mathematics*, 7(10), 2019.
- [45] Guh, R. S., Shiue Y. R., An effective application of decision tree learning for on-line detection of mean shifts in multivariate control charts, *Comput. Ind. Eng.*, 55(2), 475–493, 2008.
- [46] Yu J., Xi L., Zhou X., Identifying source(s) of out-of-control signals in multivariate manufacturing processes using selective neural network ensemble, *Eng. Appl. Artif. Intell.*, 22(1), 141–152, 2009.
- [47] Alfaro, E., Alfaro, J.L., Gamez M., Garcia N., A boosting approach for understanding out-of-control signals in multivariate control charts, *Int. J. Prod. Res.*, 47(24), 6821–6834, 2009.
- [48] Du, S., Lv, J., Xi, L., On-line classifying process mean shifts in multivariate control charts based on multiclass support vector machines, *Int. J. Prod. Res.*, 50(22), 6288–6310, 2012
- [49] Cheng, C.S., Lee H.T., Identifying the out-of-control variables of multivariate control chart using ensemble SVM classifiers, *J. Chinese Inst. Ind. Eng.*, 29(5), 314–323, 2012
- [50] Yang, W. A., Monitoring and diagnosing of mean shifts in multivariate manufacturing processes using two-level selective ensemble of learning vector quantization neural networks, *J. Intell. Manuf.*, 26(4), 769–783, 2015.

- [51] Jiang, J., Song, H.-M., Diagnosis of out-of-control signals in multivariate statistical process control based on bagging and decision tree, *Asian Bus. Res.*, 2(2), 2017.
- [52] Asadi, A., Farjami Y., Online mean shift detection in multivariate quality control using boosted decision tree learning, *J. Syst. Manag.*, sayı 2, 081–106, 2019.
- [53] Yu, J. Bo., Xi, L. Feng., A neural network ensemble-based model for on-line monitoring and diagnosis of out-of-control signals in multivariate manufacturing processes, *Expert Syst. Appl.*, 36(1), 909–921, 2009.
- [54] Pituch, K. A., Stevens, J., *Applied multivariate statistics for the social sciences: analyses with SAS and IBM's SPSS*, 6. baskı, 144,227, Routledge, 2016.
- [55] Brenner D., Bilodeau M., *Theory of Multivariate Statistics*, 3. baskı. Springer, 1999.
- [56] Meyers, L.S., Gamst, G., Guarino, A.J., *Applied multivariate research: design and interpretation*. SAGE Publications, 2005.
- [57] Korkmaz, S., Goksuluk, D., Zararsiz G., MVN: An R Package for Assessing Multivariate Normality, *Contributed research articles, The R Journal*, 151-162, 6(2), 2014.
- [58] Jayaram, N., Baker, J.W., Statistical tests of the joint distribution of spectral acceleration values, *Bull. Seismol. Soc. Am.*, 98(5), 2231–2243, 2008.
- [59] Hair, F. J., Black, W. C., Babin, B. J., Anderson, R. E., *Multivariate Data Analysis*, Pearson Education Limited, 7.baskı, 2014.
- [60] Kendall, M. G., Buckland, W. R., *A dictionary of statistical terms-*. Longman for the International Statistical Institute, 3. Baskı, 1975.
- [61] Kılıçbay, A., *Ekonometrinin Temelleri*, İstanbul Üniversitesi Yayın İktisat Fakültesi, 1986.
- [62] Brockwell, P. J., Davis, R. A., *Introduction to Time Series and Forecasting*, 2. baskı. Springer, 2002.
- [63] Montgomery, D. C., Jennings, C. L., Kulahcı, M., *Introduction to Time Series Analysis and forecating*, A John Wiley & Sons, 2008.
- [64] Warner, R., *Spectral Analysis of Time Series Data*, Guilford Press, 1998.
- [65] Gujarati, D. N., *Basic Econometrics*, McGraw-Hill Companies, 4. baskı, 754, 2004.

- [66] Rencher, A. C., *Methods of Multivariate Analysis*: 2.baskı. John Wiley & Sons, 2002.
- [67] Burdinski, T. K., Evaluating univariate, bivariate and multivariate normality using graphical procedures, evaluating normality, *Multiple Linear Regression Viewpoints*, 26 (2), 15-2.,2000.
- [68] Timm, N. H., *Applied Multivariate Analysis*, Springer, 133, 2002.
- [69] Doğruel, H. M., Tek ve çok değişkenli proses kontrol diyagramları ve beyaz eşya sektöründe uygulanması, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, Sayısal Yöntemler Bilim Dalı, Yüksek Lisans Tezi, 75, 2010.
- [70] Mason, R. L., Champ, C. W., Tracy, N. D., Wierda, S. J., Young, J. C. Assessment of multivariate process control techniques, *J. Qual. Technol.*, 29(2), 140–143, 1997.
- [71] The Royal Society, *Machine learning: the power and promise of computers that learn by example*, 5-6, 2017.
- [72] Mohammed, M., Khan, M. B., Bashier, E. B. M., *Machine learning: Algorithms and applications*, 1. baskı. CRC Press, 5-11, 2016.
- [73] Bilgin, M., *Veri Biliminde Makine Öğrenmesi Makine Öğrenmesi Teorisi ve Algoritmaları*, 2. baskı. Papatya Bilim, 31-138, 2018.
- [74] Maimon, L., Rokach, O., *Data Mining and Knowledge Discovery Handbook*, 2.baskı. Springer London, 165-174, 2010.
- [75] Refaeilzadeh, P., Tang, L., Liu, H., *C Cross-Validation*, Springer, Boston, 1-3, 2009.
- [76] Han, J., Kamber, M., Pei, J., *Data Mining. Concepts and Techniques*, The Morgan Kaufmann Series in Data Management Systems, 3. Baskı, 2012.
- [77] Mitchell, T. M., *Machine Learning*, McGraw-Hill Science, 52-155, 2014.
- [78] Agrawal, R., Imielinski, T., ve Swami, A., Mining Association Rules between Sets of Items in Large Databases, *ACM SIGMOD*, 1–10, 1993.
- [79] Utgoff, P. E. Berkman, N. C., Clouse, J. A., *Decision Tree Induction Based on Efficient Tree Restructuring*, Kluwer Academic Publishers, 29, 5-44, 1997.
- [80] Quinlan, J. R. *Induction of Decision Trees*, Kluwer Academic Publishers, *Machine Learning*, 81-106, 1986.
- [81] Alpaydın, E., *Yapay Öğrenme*, 3. baskı. Boğaziçi Üniversitesi, 207-341, 2012.

- [82] Lantz, B., *Machine learning with R : learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Birmingham: Packt Publishing Ltd; 66-343, 2013.
- [83] Nilsson NJ. *Introduction to Machine Learning*. An early draft of a proposed textbook. Robotics Laboratory, Department of Computer Science, Stanford University: Stanford, 70-74, 1996.
- [84] Öztemel E., *Yapay Sinir Ağları*, İstanbul, Papatya Yayınları, 7, 2003.
- [85] Haykin S., *Neural Networks: A Comprehensive Foundation*. Pearson Prentice Hall, 2.baskı, 1998.
- [86] Kuncheva, L., *Combining Pattern Classifiers Methods and Algorithms*, Wiley-Interscience, 1.baskı, 68, 2004.
- [87] Yadav, M., Yadav, A., Kumar N., *An Introduction to Neural Network Methods for Differential Equations*, Springer, 2015.
- [88] Priddy, K. L., Keller, P. E., *Artificial Neural Networks: An Introduction*, SPIE Publications, 2005.
- [89] Cortes, C., Vapnik, V., *Support-vector networks*, *Mach. Learn.*, 20(3), 273–297, 1995.
- [90] Rokach, L., *Ensemble-based classifiers*, *Artif. Intell. Rev.*, 33(1–2), 1–39, 2010.
- [91] Zhou, Z. H., *Ensemble methods: foundations and algorithms* Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Taylor & Francis, 2012.
- [92] Gowda, S., Kumar, H., Imran, M., *Ensemble Based Learning with Stacking, Boosting and Bagging for Unimodal Biometric Identification System*, 30-36, 2018.
- [93] Zhang, Y., Ma, C., *Ensemble Machine Learning*. Springer US, 2012.
- [94] Onan, A., *Particle Swarm Optimization Based Stacking Method with an Application to Text Classification*, *Acad. Platf. J. Eng. Sci.*, 6(2), 134–141, 2018.
- [95] Oza, N. C., Tumer, K., *Classifier ensembles: Select real-world applications*, *Inf. Fusion*, 9(1), 4–20, 2008
- [96] Raschka, S., *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning*, 1-49, 2018.

- [97] Zheng, A., *Evaluating Machine Learning Models*, O'Reilly Medis, Inc, 2015.
- [98] Hossin, M, Sulaiman, M., N, A Review on Evaluation Metrics for Data Classification Evaluations, *Int. J. Data Min. Knowl. Manag. Process*, 5(2), 01–11, 2015.
- [99] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., SMOTE: Synthetic minority over-sampling technique”, *J. Artif. Intell. Res.*, 16, 321-357, 2002.
- [100] Blagus, R., Lusa, L., SMOTE for high-dimensional class-imbalanced data, *BMC Bioinformatics*, 14(16), 1471–2103, 2013.
- [101] Zhang, Y., Li, M., Han, S., Ren, Q., Shi, J., Intelligent identification for rock-mineral microscopic images using ensemble machine learning algorithms, *Sensors*, 19(9), 1-14, 2019.
- [102] Farhan, S., Fahiem, M. A., Tauseef, H., An ensemble-of-classifiers based approach for early diagnosis of alzheimer’s disease: Classification using structural features of brain images, *Comput. Math., Methods Med.*, 2014, 2014.
- [103] Zhu, Y., Xie, C., Wang, G. J., Yan X. G., Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China’s SME credit risk in supply chain finance, *Neural Comput. Appl.*, 28, 41-50, 2017.
- [104] He, S., Wang, G. A., Zhang, M., Cook, D. F., Multivariate process monitoring and fault identification using multiple decision tree classifiers, *Int. J. Prod. Res.*, 51(11),3355-3371, 2013.
- [105] Jonathan, O., Omoregbe, N., Misra, S., Empirical Comparison of Cross-Validation and Test Data on Internet Traffic Classification Methods, *Journal of Physics: Conference Series*, 1299(1), 1-9, 2019.
- [106] Karimi, S., Yin, J., Baum, J., Evaluation methods for statistically dependent text”, *Comput. Linguist.*, 41(3), 539–548, 2015.
- [107] Ramezan, C. A., Warner, T. A., Maxwell, A. E., Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification”, *Remote Sens.*, 11(185), 1-22, 2019.
- [108] Yu, Y., Feng, Y., Modified Cross-Validation for Penalized High-Dimensional Linear Regression Models, *J. Comput. Graph. Stat.*, 23(4), 1009-1027, 2014.
- [109] Dreiseitl, S., Machado, O, L., Kittler, H., Vinterbo, S., Billhardt, H., Binder, M., A comparison of machine learning methods for the diagnosis of pigmented skin lesions, *J. Biomed. Inform.*, 34(1), 28-36, 2001.

- [110] Anwar, H., Qamar, U., ve Qureshi, A. W. M., Global optimization ensemble model for classification methods, *Sci. World J.*, 2014, 1-9, 2014.
- [111] Hu, L. Y., Huang, M. W., Ke, S. W., Tsai, C. F., The distance function effect on k-nearest neighbor classification for medical datasets, *Springerplus*, 5(1), 2016.
- [112] Lu, C. J., Shao, Y. E., Li, P. H., Mixture control chart patterns recognition using independent component analysis and support vector machine, *Neurocomputing*, 74(11), 1908-1914, 2011.
- [113] Onel, M., Kieslich, C. A., Pistikopoulos, E. N., A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: Application to the Tennessee Eastman process”, *AIChE J.*, 65(3), 992–1005, 2019.
- [114] Salehi, M., Kazemzadeh, R. B., Salmasnia, A., On line detection of mean and variance shift using neural networks and support vector machine in multivariate processes, *Appl. Soft Comput. J.*, 12(9), 2973–2984, 2012.
- [115] Maleki, M. R., Amiri, A., Simultaneous Monitoring of Multivariate-attribute Process Mean and Variability Using Artificial Neural Networks, *J. Qual. Eng. Prod. Optim.*, 1(1), 43–54, 2015.
- [116] Ali, M., Eyduran. E., Tariq. M. M., Jan, S., Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in harnai sheep, *Pak. J. Zool.*, 47(6), 1579–1585, 2015.
- [117] Manesh, S. G., Fathzadeh, A., Mehrjardi, R. T., Comparison of artificial neural network and decision tree models in estimating spatial distribution of snow depth in a semi-arid region of Iran, *Cold Reg. Sci. Technol.*, 122, 26-35, 2016.
- [118] Ekiz, B., Baygul, O., Yalcintan, H., Ozcan, M., Comparison of the decision tree, artificial neural network and multiple regression methods for prediction of carcass tissues composition of goat kids, *Meat Sci.*, 161, 2020.
- [119] Ahmad, A., Mustapha, A., Zahadi, E. D., Masah, N., Yahaya, N. Y., Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus, *Communications in Computer and Information Science*, 188(2011), 537-545, 2011.

EKLER

EK 1: x_1 deęişkeni otokorelasyon sonuçları

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	-,003	,006	,251	1	,616
2	-,003	,006	,487	2	,784
3	-,003	,006	,719	3	,869
4	-,005	,006	1,298	4	,862
5	-,004	,006	1,797	5	,876
6	,000	,006	1,798	6	,937
7	,010	,006	4,560	7	,713
8	-,006	,006	5,454	8	,708
9	-,001	,006	5,488	9	,790
10	,003	,006	5,705	10	,839
11	,021	,006	17,723	11	,088
12	-,005	,006	18,405	12	,104
13	,001	,006	18,442	13	,141
14	-,007	,006	19,721	14	,139
15	,001	,006	19,736	15	,182
16	,004	,006	20,169	16	,213

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

EK 2: x_2 deęişkeni otokorelasyon sonuçları

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	,003	,006	,202	1	,653
2	-,013	,006	4,390	2	,111
3	,002	,006	4,492	3	,213
4	,007	,006	5,846	4	,211
5	,002	,006	6,012	5	,305
6	-,006	,006	7,034	6	,318
7	,002	,006	7,151	7	,413
8	-,001	,006	7,179	8	,517
9	-,002	,006	7,288	9	,607
10	-,004	,006	7,679	10	,660
11	-,006	,006	8,765	11	,644
12	,007	,006	10,038	12	,613
13	,003	,006	10,265	13	,672
14	,001	,006	10,300	14	,740
15	-,004	,006	10,761	15	,769
16	-,002	,006	10,837	16	,819

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

EK 3: x_3 deęişkeni otokorelasyon sonuçları

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	,001	,006	,021	1	,885
2	,003	,006	,236	2	,889
3	-,007	,006	1,391	3	,708
4	-,003	,006	1,645	4	,801
5	-,016	,006	8,103	5	,151
6	,005	,006	8,840	6	,183
7	,012	,006	12,407	7	,088
8	-,004	,006	12,924	8	,114
9	-,013	,006	17,350	9	,044
10	-,004	,006	17,698	10	,060
11	,002	,006	17,854	11	,085
12	,009	,006	20,225	12	,063
13	,000	,006	20,228	13	,090
14	-,001	,006	20,285	14	,121
15	-,003	,006	20,605	15	,150
16	-,004	,006	20,944	16	,181

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

EK 4: x_4 deęişkeni otokorelasyon sonuçları

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	-,004	,006	,486	1	,486
2	-,007	,006	1,815	2	,403
3	-,008	,006	3,658	3	,301
4	,001	,006	3,673	4	,452
5	-,009	,006	5,697	5	,337
6	-,012	,006	9,716	6	,137
7	-,008	,006	11,446	7	,120
8	-,001	,006	11,453	8	,177
9	-,003	,006	11,779	9	,226
10	-,004	,006	12,293	10	,266
11	-,005	,006	12,924	11	,298
12	,002	,006	13,037	12	,366
13	,002	,006	13,119	13	,439
14	-,003	,006	13,372	14	,497
15	,003	,006	13,654	15	,552
16	-,002	,006	13,769	16	,616

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

EK 5: x_4 deęişkeni otokorelasyon sonuçları

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	-,004	,006	,486	1	,486
2	-,007	,006	1,815	2	,403
3	-,008	,006	3,658	3	,301
4	,001	,006	3,673	4	,452
5	-,009	,006	5,697	5	,337
6	-,012	,006	9,716	6	,137
7	-,008	,006	11,446	7	,120
8	-,001	,006	11,453	8	,177
9	-,003	,006	11,779	9	,226
10	-,004	,006	12,293	10	,266
11	-,005	,006	12,924	11	,298
12	,002	,006	13,037	12	,366
13	,002	,006	13,119	13	,439
14	-,003	,006	13,372	14	,497
15	,003	,006	13,654	15	,552
16	-,002	,006	13,769	16	,616

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

EK 6: x_5 değişkeni otokorelasyon sonuçları

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	-,007	,006	1,190	1	,275
2	-,003	,006	1,464	2	,481
3	,001	,006	1,492	3	,684
4	-,013	,006	5,921	4	,205
5	-,003	,006	6,186	5	,289
6	-,005	,006	6,879	6	,332
7	,012	,006	10,421	7	,166
8	-,005	,006	10,983	8	,203
9	,000	,006	10,983	9	,277
10	-,007	,006	12,470	10	,255
11	-,008	,006	14,236	11	,220
12	-,007	,006	15,410	12	,220
13	-,002	,006	15,477	13	,279
14	,006	,006	16,535	14	,282
15	-,003	,006	16,765	15	,333
16	,000	,006	16,769	16	,401

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

EK 7: x_6 deęişkeni otokorelasyon sonuçları

Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	,000	,006	,006	1	,939
2	-,002	,006	,095	2	,954
3	-,006	,006	1,048	3	,790
4	-,002	,006	1,125	4	,890
5	,002	,006	1,248	5	,940
6	-,005	,006	1,869	6	,931
7	,004	,006	2,402	7	,934
8	,003	,006	2,633	8	,955
9	-,005	,006	3,297	9	,951
10	-,003	,006	3,509	10	,967
11	-,004	,006	3,928	11	,972
12	-,002	,006	4,078	12	,982
13	-,001	,006	4,110	13	,990
14	,002	,006	4,277	14	,994
15	,005	,006	4,977	15	,992
16	-,009	,006	7,301	16	,967

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

ÖZGEÇMİŞ

Deniz DEMİRCİOĞLU DİREN, 13.05.1984'de Sakarya'da doğdu. İlk, orta ve lise eğitimini Adapazarı'nda tamamladı. 2003 yılında Sakarya Üniversitesi Endüstri Mühendisliği bölümünü kazanarak 2007 yılında mezun oldu ve endüstri mühendisi unvanı aldı. 2008 yılında Sakarya Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim dalında başladığı yüksek lisans eğitimini 2011 yılında tamamladı ve aynı yıl Sakarya Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim dalında doktora eğitimine başladı. 2009 yılında Sakarya Üniversitesi Uzaktan Eğitim Araştırma ve Uygulama Merkezinde Öğretim Görevlisi olarak çalışmaya başladı. Evli ve bir çocuk annesidir.