



**KÜMELEME VE YEREL AYKIRI FAKTÖR TABANLI
AKTİF ÖĞRENME YAKLAŞIMLARI: OTOMOTİV
SEKTÖRÜNDE BİR UYGULAMA**

Fatma Saniye KOYUNCU



T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

**KÜMELEME VE YEREL AYKIRI FAKTÖR TABANLI AKTİF ÖĞRENME
YAKLAŞIMLARI: OTOMOTİV SEKTÖRÜNDE BİR UYGULAMA**

Fatma Saniye KOYUNCU
0000-0003-3661-4154

Prof. Dr. Tülin İNKAYA
(Danışman)

YÜKSEK LİSANS
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

BURSA – 2024
Her Hakkı Saklıdır

ÖZET

Yüksek Lisans Tezi

KÜMELEME VE YEREL AYKIRI FAKTÖR TABANLI AKTİF ÖĞRENME YAKLAŞIMLARI: OTOMOTİV SEKTÖRÜNDE BİR UYGULAMA

Fatma Saniye KOYUNCU

Bursa Uludağ Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Tülin İNKAYA

Dijitalleşme ile birlikte küreselleşme; imalat, finans, enerji, sağlık gibi çeşitli üretim ve hizmet sistemlerinde toplanan verinin artmasını sağlamıştır. Bu sistemlerin verimli ve etkin bir şekilde yönetilebilmesi için yapay zekâ teknikleri kullanılarak toplanan veri ile tahmin ve çıkarımlar yapılmaktadır. Yapay zekâ yöntemleri, öğrenme sürecinde yeterli miktarda etiketli veriye ihtiyaç duymaktadır. Diğer yandan gerçek hayattaki sistemlerde verilerin çoğu etiketsizdir ve etiketleme işlemi maliyetli veya zordur. Bu amaçla aktif öğrenme çözüm yaklaşımı olarak kullanılmaktadır. Aktif öğrenme, modelin çevre ile etkileşime girdiği ve bilgilendirici örneklerin etiketlendiği makine öğrenmesi yöntemlerinden biridir. Bu tezde, zaman serisi sınıflandırma problemi için aktif öğrenme yaklaşımları önerilmiştir. Bu kapsamda, verinin tamamının etiketsiz olduğu durum ele alınmıştır ve iki başlangıç yaklaşımı önerilmiştir. Önerilen başlangıç yaklaşımlarında önce gürültü ya da aykırı veriler başlangıç eğitim kümesinden çıkarılmıştır, sonrasında kalan verilerdeki temsilci örnekler kümeleme ile belirlenmiştir. İlk yaklaşımda gürültüyü tespit etmek amacıyla DBSCAN algoritması; ikinci yaklaşımda ise aykırı verileri tespit etmek amacıyla yerel aykırı faktör (LOF) algoritması kullanılmıştır. Her iki yaklaşımda da veri kümesindeki temsilci örnekleri belirlemek için k-ortalamalar kümeleme yöntemi uygulanmıştır. Belirlenen temsilci örnekler başlangıç aşamasında uzman tarafından etiketlenmiştir. Zaman serilerinin sınıflandırmasında k-en yakın komşu (KNN) algoritması ve aktif öğrenmede sorgu seçimi için belirsizlik örneklemesi kullanılmıştır. Önerilen yaklaşımların performansları, üretim ve sağlık sistemlerinden toplanan dört adet sensör veri kümesi üzerinde test edilmiştir. Deneysel çalışmalarda kümeleme yöntemi, sınıflandırma algoritması ve sorgu sayısının etkileri analiz edilmiştir. Performans kriterleri olarak doğruluk ve ağırlıklı F-ölçütü kullanılmıştır. Ayrıca önerilen yaklaşımlar bir otomobil üreticisinde tedarikçilerin risk tahmininde uygulanmıştır. Deneysel çalışmalar, önerilen yaklaşımların başlangıç eğitim kümesinin belirlenmesinde etkili olduğunu göstermiştir.

Anahtar Kelimeler: Makine Öğrenmesi, Aktif Öğrenme, Kümeleme, Zaman Serisi, Başlangıç, Tedarik Zinciri Risk Yönetimi.

2024, xii + 69 sayfa.

ABSTRACT

MSc Thesis

CLUSTERING AND LOCAL OUTLIER FACTOR-BASED ACTIVE LEARNING APPROACHES: AN APPLICATION TO THE AUTOMOTIVE INDUSTRY

Fatma Saniye KOYUNCU

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

Supervisor: Prof. Dr. Tülin İNKAYA

Globalization along with digitalization has enabled the increase of data collected in various production and service systems such as manufacturing, finance, energy and health. In order to manage these systems efficiently and effectively, predictions and inferences are made with the data collected using artificial intelligence techniques. Artificial intelligence methods require a sufficient amount of labeled data in the learning process. On the other hand, in real-life systems, most of the data is unlabeled and labeling is costly or difficult. For this purpose, active learning is used as a solution approach. Active learning is one of the machine learning methods in which the model interacts with the environment and informative examples are labeled. In this thesis, active learning approaches are proposed for the time series classification problem. In this context, the situation where all the data is unlabeled is considered and two initialization approaches are proposed. In the proposed initialization approaches, noise or outlier data are first removed from the initial training set, and then representative samples in the remaining data are determined by clustering. In the first approach, DBSCAN algorithm to detect noise; in the second approach, the local outlier factor (LOF) algorithm was used to detect outlier data. In both approaches, k-means clustering method was applied to identify representative samples in the dataset. The identified representative samples were labeled by an expert at the initialization stage. The k-nearest neighbor (KNN) algorithm was used for classification of time series and uncertainty sampling was used for query selection in active learning. The performances of the proposed approaches were tested on four sensor datasets collected from production and healthcare systems. In the experimental studies, the effects of clustering method, classification algorithm and number of queries were analyzed. Accuracy and weighted F-measure were used as performance criteria. Additionally, the proposed approaches have been applied to suppliers' risk estimation in an automobile manufacturer. Experimental studies have shown that the proposed approaches are effective in determining the initial training set.

Key words: Machine Learning, Active Learning, Clustering, Time Series, Initialization, Supply Chain Risk Management.

2024, xii + 69 pages.

TEŞEKKÜR

Tez danışmalığımyı yapan, bu süreç boyunca her zaman bilgi ve deneyimiyle yanımda olan ve her koşulda desteğini hissettiren, işini sevgiyle ve sabırla yaparak bana örnek olan danışmanım Prof. Dr. Tülin İNKAYA' ya sonsuz teşekkürlerimi sunarım.

Yüksek lisans eğitimi konusunda beni yönlendiren Prof. Dr. Ali Yurdun ORBAK'a, desteklerini esirgemeyen başta Onur YALÇIN olmak üzere tüm yöneticilerime ve zorlandığım her noktada benimle olan arkadaşlarıma teşekkür ederim.

Hayatım boyunca her zaman ve her koşulda desteğiyle yanımda olan, bana inanan, bugünlere gelmemde büyük emeği olan annem Mevlüde ŞARDAĞ'a ve dedem Mefail ŞARDAĞ'a sonsuz teşekkürlerimi sunarım.

Fatma Saniye KOYUNCU
20/01/2024

İÇİNDEKİLER

	Sayfa
ÖZET.....	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
SİMGELER ve KISALTMALAR DİZİNİ	viii
ŞEKİLLER DİZİNİ.....	xi
ÇİZELGELER DİZİNİ	xii
1. GİRİŞ.....	1
2. KAYNAK ARAŞTIRMASI	4
2.1. Aktif Öğrenme	4
2.2. Aktif Öğrenmede Başlangıç Problemi	8
3. MATERYAL ve YÖNTEM.....	11
3.1. Materyal	11
3.1.1. Zaman Serisi.....	11
3.1.2. Aktif Öğrenme ve Sorgu Stratejisi	12
3.2. Yöntem.....	13
3.2.1. K-En Yakın Komşuluk Yöntemi.....	13
3.2.2. Karar Ağacı Yöntemi	14
3.2.3. Naive Bayes Yöntemi	15
3.2.4. K-Ortalamalar Yöntemi	16
3.2.5. K-Medoid Yöntemi	17
3.2.6. DBSCAN Yöntemi	18
3.2.7. Yerel Aykırı Faktör Algoritması.....	19
3.2.8. Önerilen Yaklaşımlar	21
3.2.9. Performans Değerlendirme Metrikleri	24
3.2.10. Parametrik Olmayan İstatistiksel Testler	26
4. BULGULAR ve TARTIŞMA.....	28
4.1. Veri Kümeleri ve Uygulama Detayları	28
4.2. Deneysel Sonuçlar.....	30
4.3. Otomobil Sektöründe Tedarikçi Risk Durumu için Vaka Analizi.....	44
4.3.1. Tedarik Zinciri Ortamı	46
4.3.2. Veri Kümesi ve Önileme	48
4.3.3. Vaka Analizine İlişkin Bulgular	50
5. SONUÇ.....	54
KAYNAKLAR	56
EKLER.....	60
EK 1 1NN, 3NN ve 5NN Yöntemleri Başlangıç Sorgu Seçimi Doğruluk Sonuçları	61
EK 2 1NN, 3NN ve 5NN Yöntemleri Başlangıç Sorgu Seçimi Ağırlıklı F-ölçütü Sonuçları.....	62
EK 3 Karar Ağacı Yöntemi Başlangıç Sorgu Seçimi Doğruluk Sonuçları.....	63
EK 4 Karar Ağacı Yöntemi Başlangıç Sorgu Seçimi Ağırlıklı F-ölçütü Sonuçları.....	64
EK 5 Gaussian Naive Bayes Yöntemi Başlangıç Sorgu Seçimi Doğruluk Sonuçları	65
EK 6 Gaussian Naive Bayes Yöntemi Başlangıç Sorgu Seçimi Ağırlıklı F-ölçütü Sonuçları.....	66
EK 7 Eğitim süreleri ve yöntemlerin eğitim sürelerine göre sıralamaları	67
ÖZGEÇMİŞ	69

SİMGELER ve KISALTMALAR DİZİNİ

Simgeler	Açıklama
C	Küme
c	Toplam sınıf sayısı
CC	Küme merkezi
D	Veri kümesi
$d(.,.)$	Uzaklık fonksiyonu
ε	Uzaklık eşik değeri
E	Kümeleme hatası
F	F-ölçütü
g	Karşılaştırılan model sayısı
H	Noktalar kümesi
K	En yakın komşuluk sayısı
$K\text{-distance}$	Bir nesnenin K . en yakın komşusu ile arasındaki mesafe
L	Veri kümeleri
Lrd_{minpts}	Yerel erişilebilirlik yoğunluğu
LOF_{minpts}	Yerel aykırı faktör değeri
M	Boyut sayısı
m	Öznitelik sayısı
$minpts$	Minimum nesne sayısı
n	Veri kümesindeki nesne sayısı
nc	Küme sayısı
$N_{K\text{-distance}}$	Bir nesnenin K komşuluk mesafesi
$N\varepsilon(x)$	Bir nesnesinin ε -komşuluğu
o	Nesne
$reachdist_K$	Ulaşılabilirlik mesafesi
p_j	x_i nesnesinin j sınıfına ait olma olasılığı
$P(y x)$	x nesnesinin y sınıf etiketine dahil olma olasılığı
$P(x_i y)$	y sınıfı verildiğinde x_i örneğinin gözlenme olasılığı
q	Veri noktası
q_α	İstatistiksel aralık
R_j	j . yöntemin ortalama sırası
r_i^j	j . karşılaştırma modelinin i . veri kümesindeki sıralaması
S	Değişiklik maliyeti
T	Boyut sayısı
w_i	i . örnek için ağırlık fonksiyonu
x	Nesne
x^*	En az kesin olan nesne
x_i	i . nesne
x_j	j . nesne
x^T	Test nesnesi
x_{ij}	Veri kümesindeki i . nesnenin j . öznitelikle aldığı değer
χ_F^2	Friedman Ki-kare istatistiği
y	Sınıf etiketi
y^T	Test nesnesi x^T 'nin sınıf etiketi
y_i	i . nesnenin sınıf etiketi

$\hat{y}$	En olası sınıf etiketi
μ	Örneklerin ortalama değeri
σ	Örneklerin standart sapması
σ^2	Örneklerin varyansı

Kısaltmalar	Açıklama
AI	Yapay Zeka (Artificial Intelligence)
AHC	Birleştirici Hiyerarşik Kümeleme (Agglomerative Hierarchical Clustering)
APC	Yakınlık Yayılımlı Kümeleme (Affinity Propagation Clustering)
AL	Aktif Öğrenme (Active Learning)
BIRCH	Dengeli İteratif Azaltma ve Hiyerarşik Kümeleme (Balanced Iterative Reducing and Clustering using Hierarchies)
CD	Kritik Mesafe (Critical Distance)
CNN	Evrişimli Sinir Ağı (Convolutional Neural Network)
CALR	Soğuk Başlangıç Aktif Öğrenme (Cold - Start Active Learning)
DBSCAN	Gürültülü Uygulamalarda Yoğunluk Tabanlı Uzamsal Kümeleme (Density - Based Spatial Clustering of Applications with Noise)
DN	Doğru Negatif (True Negative)
DP	Doğru Pozitif (True Positive)
DPEA	Dirichlet Süreci Beklenen Doğruluk (Dirichlet Process Expected Accuracy)
DT	Karar Ağacı (Decision Tree)
Σ -opt	Σ -Optimalite Temelli Grafik Varyansın Minimizasyonu (Σ -optimality Based Graph Variance Minimization)
FFT	En Uzak İlk Gezinme (Furthest - First - Travel)
GDBT	Gradyan Arttırma Karar Ağaçları (Gradient Boosting Decision Trees)
Gen/Disc	Üretici ve Ayrımcı Sınıflayıcılar (Generative and Discriminative Classifiers)
GNB	Gaussian Naive Bayes
IoT	Nesnelerin İnterneti (Internet of Things)
KNN	K-En yakın komşu
k-means+ME	k-ortalamlar + model örnekleri (k-means + model examples)
LOF	Yerel Aykırı Faktör (Local Outlier Factor)
LWAC	Hafif Ağırlıklı Aktif Öğrenme (Light – Weight Active Learning)
NB	Naive Bayes
NETS	Sonsuz Sürekli Zaman Serisi (Never-Ending Time Series)
NN	En Yakın Komşu (Nearest Neighbor)
ODAL	Aykırı Değer Tabanlı Yansız Aktif Öğrenme (Outlier – Based Discriminative Active Learning)
pWrong	En Muhtemel Yanlış Örneklerin Seçimi (The Selection of Most Probable Wrong Samples)
RF	Rastgele Orman (Random Forest)
SCRM	Tedarik Zinciri Risk Yönetimi (Supply Chain Risk Management)
SSE	Karesel Hatalar Toplamı (Sum of Squared Errors)
SVM	Destek Vektör Makinesi (Support Vector Machine)
TED	D-Optimalite Temelli Transdüktif Deneysel Tasarım (D-Optimality Based Transductive Experimental Design)

YN	Yanlış Negatif (False Negative)
YP	Yanlış Pozitif (False Pozitif)



ŞEKİLLER DİZİNİ

	Sayfa
Şekil 3.1. Aktif öğrenme.....	13
Şekil 3.2. KNN algoritması.....	14
Şekil 3.3. k -ortalamalar algoritması.....	17
Şekil 3.4. k -medoid algoritması.....	18
Şekil 3.5. DBSCAN kümeleme algoritması.....	20
Şekil 3.6. LOF algoritması.....	21
Şekil 3.7. Önerilen DBSCAN+ k -ortalamalar akış şeması.....	23
Şekil 3.8. Önerilen LOF+ k -ortalamalar akış şeması.....	24
Şekil 4.1. Friedman ve Nemenyi testleri 1NN, 3NN, 5NN ile başlangıç sorgu örneği seçimi karşılaştırma grafikleri.....	32
Şekil 4.2. Friedman ve Nemenyi testleri DT sınıflandırma ile başlangıç sorgu örneği seçimi karşılaştırma grafikleri.....	33
Şekil 4.3. Friedman ve Nemenyi testleri GNB sınıflandırma ile başlangıç sorgu örneği seçimi karşılaştırma grafikleri.....	35
Şekil 4.4. Friedman ve Nemenyi testleri en iyi altı yöntem için başlangıç sorgu örneği seçimi karşılaştırma grafikleri.....	36
Şekil 4.5. 1NN-LOF+ k -ortalamalar yöntemi için başlangıç sorgu örneği seçimi doğruluk ve ağırlıklı F-ölçütü değerleri.....	38
Şekil 4.6. 1NN-DBSCAN+ k -ortalamalar yöntemi için başlangıç sorgu örneği seçimi doğruluk ve ağırlıklı F-ölçütü değerleri..	39
Şekil 4.7. Önerilen yaklaşımlar ile etiketlenen veriler için doğruluk sonuçları.....	41
Şekil 4.8. Önerilen yaklaşımlar ile etiketlenen veriler için ağırlıklı F-ölçütü sonuçları.....	42
Şekil 4.9. Tedarik zinciri akışı.....	45
Şekil 4.10. Tedarik zinciri risk yönetiminin adımları.....	46
Şekil 4.11. Tedarik zincirinde bilgi ve parça akışı.....	47
Şekil 4.12. Vaka analizinde önerilen yaklaşımlar için başlangıç eğitim örneği seçimi doğruluk ve ağırlıklı F-ölçütü sonucu.....	51
Şekil 4.13. Vaka analizinde önerilen yaklaşımlar ile etiketlenen veriler için doğruluk ve ağırlıklı F-ölçütü sonuçları.....	53

ÇİZELGELER DİZİNİ

	Sayfa
Çizelge 3.1. Karmaşıklık matrisi.....	25
Çizelge 4.1. Veri kümesi özellikleri.....	28
Çizelge 4.2. DBSCAN yönteminde ε değerleri.....	30
Çizelge 4.3. Gürültülü ve aykırı veri sayılar.....	30
Çizelge 4.4. Önerilen yaklaşımların 150 etiketli veri ile eğitim süreleri ve sıralama değerleri.....	44
Çizelge 4.5. Öznitelikler.....	49



1. GİRİŞ

Şirketler üretim ve hizmet sistemlerini daha verimli ve daha esnek hale getirmek, artan küreselleşme ile ortaya çıkan riskleri öngörmek ve önlemek, rekabete ortak olmak için dijital dönüşüm faaliyetlerine destek vermektedirler. Günümüzde riski belirlemek ve yönetmek, rekabete ortak olmak, sistemleri daha verimli kullanmak ve müşteri memnuniyeti açısından önemlidir. Küreselleşme, tedarik zincirini doğrudan etkileyerek firmaları tüm risk ve belirsizliklere karşı karşıya bırakmıştır. Firmalar, belirsizlik ve risklere vaktinde önlem alabilmek için tedarik zincirinde risk yönetimine odaklanmıştır. Yapay zekâ (Artificial Intelligence - AI), nesnelerin interneti (Internet of Things - IoT), akıllı üretim sistemleri gibi yeni teknolojiler; veriyi depolama ve veriden değer yaratma gibi çözümler barındırarak, sistemlerin daha etkili kullanılmasını sağlamaktadır.

Artan dijitalleşme faaliyetleriyle beraber veriyi elde etmek ve depolamak kolaylaşmıştır. Büyük miktarda depolanan veriyi analiz etmek ve veriden değer yaratmak ise zorlu bir süreci içermektedir. Büyük veri kümelerinde, veriden değer elde etmek için makine öğrenmesi yaklaşımları kullanılabilir. Makine öğrenmesi, deneyimleri dikkate alarak öğrenme algoritmalarıyla veriden öğrenen bir tekniktir (Zhou, 2021). Gözetimli, gözetimsiz, pekiştirmeli, yarı gözetimli, topluluk tabanlı ve transfer öğrenme gibi çeşitli öğrenme yöntemleri bulunmaktadır (Das & Behera, 2017). Özellikle gözetimli öğrenme yöntemlerinden sınıflandırma metotları ve gözetimsiz öğrenme yöntemlerinden verileri benzerliklere göre gruplayan kümeleme yöntemleri yaygın olarak kullanılmaktadır.

Makine öğrenmesi yöntemleri, yeni bir veri geldiğinde tahmin yapabilmek için öğrenme aşamasına ihtiyaç duymaktadır. Bir model oluşturmak için öğrenme aşamasında eğitim verisi kullanılmaktadır. Eğitim süreci tamamlandıktan sonra test verisiyle modelin performansı ölçülmektedir (Gupta & Sheng, 2019). Makine öğrenmesi modelinin eğitimi için veride sınıf etiketine ihtiyaç duyulmaktadır. Veri etiketleme sürecinde, zamana ve deneyimli işgücüne ihtiyaç duyulmaktadır. Bazı durumlarda etiketleme işlemi zor ve maliyetlidir.

Aktif öğrenme (Active Learning - AL), yüksek doğruluk ve işgücü ihtiyacını azaltarak düşük etiketleme maliyeti sağlayacak şekilde etiketlenen verinin belirlenmesi için kullanılmaktadır (Aggarwal, 2015). Model eğitimi için az miktarda etiketli veriye ihtiyaç duymaktadır. Aktif öğrenme, etiketleme işlemi için çok sayıda etiketlenmemiş veriden en değerli örnekleri seçmede etkili bir yöntemdir (He vd., 2015). Aktif öğrenme yöntemi kullanmanın avantajlarından biri de, bir uzman ya da veri tabanı tarafından etiketlenmek üzere en bilgilendirici etiketsiz verileri seçmek için çeşitli sorgu örneği seçim yaklaşımları vardır (Kumar & Gupta, 2020). Aktif öğrenme, etiketleme maliyetini azaltması, sorgu seçim stratejilerinin avantajlarından yararlanarak temsil gücü yüksek verilerin seçilmesi için sunduğu çözüm yaklaşımları açısından günümüzde tıp, üretim, görüntü işleme, sensör teknolojisi, stok piyasası gibi alanlarda geliştirilen yapay zekâ yöntemlerinin performansının iyileştirilmesinde kullanılmaktadır.

Aktif öğrenme, makine öğrenmesinin alt sınıflarından biridir. Aktif öğrenme ile az sayıda etiketli veri ile en iyi model performansının elde edilmesi amaçlanmaktadır. Makine öğrenmesinde sıklıkla kullanılan denetimli ve denetimsiz öğrenme ise farklı alt sınıflarda yer almaktadır. Denetimli öğrenmede önceden etiketlenmiş veri ile model eğitimi gerçekleştirilmektedir. Denetimsiz öğrenmede ise veriler etiketlenmemiştir ve amaç verideki gizli örüntüleri keşfetmektir.

Hava durumu, sağlık, enerji, finans gibi birçok sektörde gözlemlenen ve düzenli aralıklarla toplanan veriler mevcuttur. Bu veriler zaman serilerine örnektir. Zaman serisi, her veri noktasının ya gerçek bir değer ya da gerçek değerlerin bir vektörü olduğu, sabit bir zaman aralığında düzenli olarak kaydedilen veridir (Weber vd., 2021).

Bu çalışmada, zaman serisi sınıflandırma problemi üzerinde çalışılmıştır. Elde edilen zaman serilerinde etiket bilgisi bulunmamaktadır. Başlangıç aşamasında etiketli verilerin varlığıyla ilgili çeşitli yaklaşımlar ele alınmaktadır. Barata vd. (2021), Grimova vd. (2018), Jin vd. (2022) hiç etiketli veri olmadığı durumda soğuk başlangıç aktif öğrenme yöntemleri üzerine çalışmıştır. Literatürde yapılan çalışmalar incelendiğinde, zaman serisinde aktif öğrenme yöntemi ile yapılan sınıflandırma ve tahminleme çalışmaları

mevcutken, az ya da hiç etiketli veri olmadığında aktif öğrenmeyi başlatma problemi ile ilgili konularda birtakım eksiklikler görülmüştür.

Bu çalışmada, başlangıçta hiç etiketli veri olmadığı varsayımıyla etiketleme maliyetini azaltmak ve başlangıçta temsil gücü yüksek örneklerin seçimi için kümeleme tabanlı ve yerel aykırı değer tabanlı iki farklı yaklaşım önerilmiştir. Başlangıç örneklerinin seçiminde, aykırı değerleri veri kümesinden çıkarmak için yerel aykırı faktör (Local Outlier Factor-LOF) ve gürültülü verilerin etkisini azaltmak için gürültülü uygulamaların yoğunluk tabanlı uzamsal kümelenmesi (Density-Based Spatial Clustering of applications with noise-DBSCAN) yöntemleri önerilmiştir. Ardından, k -ortalamalar yöntemiyle örnek seçimlerinin gerçekleştirilmesi önerilmiştir. Zaman serisi sınıflandırma yöntemi olarak basit ve etkili olması açısından K -En Yakın Komşuluk (K -Nearest Neighbor-KNN) algoritması kullanılmıştır. Aktif öğrenme aşamasında, en bilgilendirici örnekleri seçmek için belirsizlik örnekleme yöntemi ele alınmıştır. Önerilen yöntemler dört zaman serisinde değerlendirilmiştir. Ayrıca, otomobil üretimi yapan bir firmada, tedarikçi parça teslimatındaki riski tahmin etmede önerilen yaklaşımların kullanıldığı vaka analizi çalışması yapılmıştır.

Önerilen yaklaşımların performansı doğruluk ve ağırlıklı F -ölçütüne (F -weighted) göre değerlendirilmiştir. Önerilen yöntemler, diğer yöntemler ile karşılaştırıldığında aralarında istatistiksel olarak anlamlı farkların olup olmadığını belirlemek amacıyla parametrik olmayan istatistiksel analizler uygulanarak yöntemlerin performansı karşılaştırılmıştır.

Bu çalışmanın organizasyonu aşağıdaki gibidir. Bölüm 2’de Kaynak Araştırması, Bölüm 3’te Materyal ve Yöntem, Bölüm 4’te Bulgular ve Tartışma, Bölüm 5’te Sonuç kısmı yer almaktadır.

2. KAYNAK ARAŞTIRMASI

Bu bölümde tez çalışması ile ilgili çalışmalar iki başlıkta incelenmiştir. Bölüm 2.1’de aktif öğrenme alanında yapılan çalışmalara yer verilmiştir. Bölüm 2.2’de aktif öğrenmede başlangıç problemi ile ilgili çalışmalar yer almaktadır.

2.1. Aktif Öğrenme

Aktif öğrenme (AL), uzmandan veya veri tabanından örneklerin etiketlerini sorgular ve az sayıda etiketli veri ile öğrenme işlemi gerçekleştirebilmektedir (Kumar & Gupta, 2020). Aktif öğrenmede sorgulama stratejilerinin sunduğu avantajlar son zamanlarda birçok araştırmacının dikkatini çekmiş ve bu alanda birçok çalışma önerilmiştir. Aggarwal (2015), Kumar ve Gupta (2020) ve Settles (2009) aktif öğrenme stratejileri ile ilgili çalışmışlardır.

Aggarwal (2015), aktif öğrenmede en bilgilendirici örneklerin seçimi için kullanılan yöntemleri özetlemiştir. Çeşitlilik, performans ve temsiliyet temelli yaklaşımları açıklamıştır.

Kumar ve Gupta (2020), sınıflandırma, kümeleme ve regresyon çalışmalarına yönelik sorgu stratejileri üzerine çalışmışlardır. Aktif öğrenme çalışma senaryolarını, havuz tabanlı (Pool-based), akış tabanlı (Stream-based) ve üyelik sorgusu sentezi tabanlı (Membership query synthesis-based) olmak üzere üçe ayırmışlardır ve her senaryo için sorgu stratejilerini açıklamışlardır.

Settles (2009), aktif öğrenme senaryoları ve sorgu stratejileri üzerine bir çalışma gerçekleştirmiştir. Aktif öğrenmede toplu sorgu örneği seçim yaklaşımı ve diğer öğrenme yaklaşımlarıyla entegrasyonu üzerinde çalışmıştır.

Zaman serisi sınıflandırma problemi yaygın olarak karşılaşılan problemlerdendir. Bir grup çalışmada zaman serilerinde aktif öğrenme yaklaşımı üzerine çalışmıştır. Gweon ve Yu (2021), He vd. (2017), Saeedi vd. (2017) ve Peng vd. (2017) zaman serisi

sınıflandırma problemine yönelik aktif öğrenme yaklaşımı geliştirmişlerdir. Bu çalışmalara ilişkin detaylar şöyledir:

Gweon ve Yu (2021), en yakın komşuya dayalı aktif öğrenme stratejisi önermiştir. Önerilen yaklaşımda, etiketlenmemiş örnekler için Öklid mesafesini hesaplanmış ve sıralanmışlardır. Belirsizliği ölçmek için bilgilendiricilik değeri kullanmışlardır. Karar sınırına yakın olan örneğe yüksek bilgilendiricilik değeri atamışlardır. Bilgilendiricilik değerini hesaplamak için belirsizlik ve fayda ölçütlerini kullanmışlardır. Sınıflandırma algoritması olarak 1-en yakın komşu (1-nearest neighbor – 1NN) yöntemini kullanmışlardır. Öneriler yaklaşımın performansını, doğruluk ve makro ortalama F-ölçütüyle değerlendirmişlerdir. Yöntemlerin ikili karşılaştırmalarını yapmak üzere Wilcoxon signed-rank testini uygulamışlardır. Önerilen yöntemi rassal seçim, Sonsuz Sürekli Zaman Serisi (Never – Ending Time Series – NETS) ve Hafif Ağırlıklı Aktif Öğrenme (Light – Weight Active Learning – LWAC) yöntemleriyle karşılaştırmışlardır. Önerilen yöntemin, rassal seçim yöntemine göre iyi sonuç verdiği sunulmuştur.

He vd. (2017), zaman serisini sınıflandırma problemine aktif öğrenme ve yarı denetimli öğrenme yaklaşımı üzerine çalışmışlardır. Etiketlenmemiş örneklerden en bilgilendirici örnekleri seçmek için belirsizlik ve yerel veri yoğunluğunu dikkate alarak bir örnek seçim stratejisi geliştirmişlerdir. Belirsizliği hesaplamak için kosinüs benzerliğini kullanmışlardır. Ters KNN yöntemiyle yerel veri yoğunluğu ölçütünü kullanmışlardır. En yüksek değere sahip örnek etiketlenmek üzere seçilmiştir. Yakın komşuları belirlemek için ters 1NN denetimli öğrenme yaklaşımı sunmuşlardır. Önerilen yaklaşımı, 1NN algoritması ve etiketli eğitim verisiyle yapmışlardır. Değerlendirme kriteri olarak doğruluk ve F-ölçütü kullanmışlardır.

Saeedi vd. (2017), mobil sağlık verilerini izlemek için çok çalışanlı ve maliyete duyarlı aktif öğrenme yöntemi önermişlerdir. Farklı uzmanlar tarafından yapılan etiketleme maliyetini minimize etmeyi amaçlamışlardır. Mobil sağlık verilerinde başlangıçta etiketli veri bulmak oldukça zordur. Bu nedenle başlangıç öğrenme algoritması, rastgele orman sınıflandırıcısıyla çalışan entropi ve topluluk tabanlı sorgulama stratejisi ve birden fazla uzmanın etiketleme işlemi yaptığı durumda uzman seçim algoritması önermişlerdir.

Peng vd. (2017), zaman serisi sınıflandırma probleminde aktif öğrenme yaklaşımını kullanmışlardır. Etiketsiz verideki en bilgilendirici örnekleri seçmek için belirsizlik ve faydaya dayalı bir değer hesaplamışlardır. Fayda ölçütünü hesaplarken zaman serisi örnekleri arasındaki korelasyonu dikkate almışlardır. Sınıflandırma doğruluk değerini iyileştirmek için en bilgilendirici sorgu örneklerini seçmek amacıyla aç gözlü algoritma önermişlerdir.

Kümeleme yaklaşımlarında, benzer örnekler aynı veri kümesine, benzemeyen örnekler farklı kümelere atanmaktadır (Aggarwal, 2015). Kümeleme, büyük miktarlarda veriyi analiz etme sürecinde makine öğrenmesi yaklaşımlarında önemli bir yere sahiptir. Diğer grup çalışmalarda kümeleme tabanlı aktif öğrenme yaklaşımları üzerine çalışılmıştır.

Zhang ve Dai (2022), zaman serisi tahmin problemi için destek vektör regresyonuna dayalı aktif öğrenme yöntemi önermiştir. Sınıf dengesizliği, modelin performansını etkileyen problemler arasındadır. Bu dengesizliği göz önünde bulundurarak, zaman serileri için tahminlerin farklılaştırılmış maliyetlere sahip olması gerektiği savunulmaktadır. Bu nedenle, dengesiz sınıf dağılımına sahip zaman serisi tahminleme problemini çözmek için verimli ve maliyete duyarlı aktif öğrenme modeli önermişlerdir. Bu yaklaşım çok kriterli aktif ağırlık modeli olarak da adlandırılmıştır. İlk olarak, maliyete duyarlı öğrenme ile örneklerle farklı maliyetler atamak için tahmin edici olarak ağırlıklı destek vektör regresyonu kullanmışlardır. Ardından, destek vektörleri ve destek olmayan vektörler setini elde etmek için ağırlıklı destek vektör regresyonuna küçük bir başlangıç eğitim verisi elde etmişlerdir. İkinci olarak, destek vektör olmayanların eğitim örnekleriyle birlikte etiketlenmemiş örneklerle kümeleme algoritması (kernel k -ortalama algoritması) uygulamışlardır. Kümeleme sonuçlarına dayanarak, birden fazla kriteri birlikte değerlendirerek değerli örnekleri seçmek için iki ardışık aşamalı prosedür önermişlerdir. Her örneği bir kümeye atamışlardır ve sonuçlar kaydedilmiştir. Mesafe hesaplama sonuçlarına göre, kümeler yeniden atama yaparak güncellemişlerdir. Yakınsama sağlanana kadar işlemleri tekrarlamışlardır. Son olarak, seçilen örnekler etiketlenmemiş veri kümesi havuzundan çıkartılmıştır. Durdurma koşulu sağlanana kadar işlemleri tekrarlamışlardır.

Qi ve Ting (2018), yarı denetimli kümeleme algoritması önermişlerdir. Aykırı değerlerin veri üzerindeki etkisini azaltmak için yerel aykırı faktör (Local Outlier Factor- LOF) yöntemini kullanmışlardır. Sonrasında daha iyi bir performans elde etmek için yarı denetimli kümeleme algoritması önermişlerdir. Kümeleme sonuçlarını, benzerlik yayılımı (affinity propagation) kümeleme algoritması aracılığıyla elde etmişlerdir.

Mai vd. (2013), aktif yoğunluk tabanlı kümeleme algoritması önermişlerdir. Yoğunluk tabanlı kümeleme algoritması olarak DBSCAN yöntemini kullanmışlardır. Bir alt sınır benzerliği matrisinin kullanılabilirliğine bağlı olarak, en bilgilendirici ikili alt sınır benzerliklerini dikkate alarak örnek seçim işlemi yapmışlardır. Önerilen yaklaşımın gerçek veri kümelerinde aktif spektral kümeleme gibi diğer yöntemlerden daha iyi performans gösterdiği görülmüştür.

Shim vd. (2021), aktif öğrenme yaklaşımını yarı iletken sektörüne uygulamışlardır. Hatalı tabaka haritalarını kümeler halinde gruplamak için k-ortalamlar kümeleme yöntemini kullanmışlardır ve tabaka haritalarının benzerliğine göre küme düzeyinde etiketleme gerçekleştirmişlerdir. Kümeleme, ilk aşamada etiketlenmemiş tabaka haritalarına uygulanmıştır. Sonraki aşamada, etiketli veri kümesi ile Evrişimli Sinir Ağı (Convolutional Neural Network - CNN) eğitilmiştir ve tabaka haritaları hata türlerine göre sınıflandırılmıştır. Etiketlenmemiş tabaka haritaları için CNN'in küme düzeyindeki belirsizlikleri hesaplanmıştır. Belirsizlikler dikkate alınarak sorgu örnekleri bir mühendis tarafından küme düzeyinde seçilip etiketleme işlemi gerçekleştirilmiştir. Önerilen yaklaşım kullanılarak etiketleme maliyeti düşük ve yüksek performanslı CNN algoritması elde etmişlerdir.

Zhou vd. (2023), kentsel yapılar üzerinde çalışmışlardır. Bina yapısının sınıflandırılması için ön kümeleme yapan aktif öğrenme yaklaşımı önermişlerdir. Ön kümeleme işlemi için DBSCAN algoritmasını kullanmışlardır. Sınıflandırma işlemi için Gradyan Arttırma Karar Ağaçları (Gradient Boosting Decision Trees – GDBT), Rastgele Orman (Random Forest – RF), Destek Vektör Makinesi (Support Vector Machine – SVM), Karar Ağacı (Decision Tree – DT), KNN yöntemlerini kullanmışlardır. Başlangıç eğitim örneği seçimi

için çekirdek örneklere tabakalı örnekleme uygulamışlardır. F-ölçütünü değerlendirme kriteri olarak kullanmışlardır. Sonuçlar, etiketleme işlemi için gereken eforun azaltıldığını göstermiştir.

2.2. Aktif Öğrenmede Başlangıç Problemi

Aktif öğrenme problemlerinde üç farklı başlangıç durumu vardır (Barata vd., 2021). Soğuk başlangıçta (cold-start), etiketlenmiş veri yoktur. Ilık başlangıçta (warm-start), az sayıda etiketli veri vardır. Sıcak başlangıçta (hot-start), tüm veriler etiketlidir. Kang vd. (2004), Hu vd. (2010), Barata vd. (2021), Grimova vd. (2018), Jin vd. (2022), Nguyen ve Smeulders (2004) ve Yang ve Loog (2022) başlangıç problemleri ile ilgili çalışmalar yapmıştır.

Kang vd. (2004), başlangıç eğitim örneklerini seçmek için bir yöntem önermişlerdir. Önerilen yöntemde küme merkezine en yakın örnekler başlangıç örneği olarak seçilmiştir. Sonuçlar rastgele örnek seçim yöntemiyle kıyaslandığında daha yüksek doğruluğa sahip olduğu görülmektedir.

Hu vd. (2010), başlangıç eğitim kümesini elde etmek için deterministik kümeleme yaklaşımları En Uzak İlk Gezinme (Furthest - First - Traversal - FFT), Birleştirici Hiyerarşik Kümeleme (Agglomerative Hierarchical Clustering - AHC), Yakınlık Yayılımlı Kümeleme (Affinity Propagation Clustering - APC) ve deterministik olmayan yaklaşımlardan olan k -ortalamlar (k -means), k -medoid ve başlangıç eğitim seti seçimi için k -ortalamlar yaklaşımının bir varyasyonu olarak , sanal merkezlerden oluşturulan yapay örnekler, model örnekleri olarak adlandırdıkları örnekleri eğitim kümesine ekleyerek karşılaştırmışlardır. Sınıflandırma işlemi için 5NN ve uzaklık ölçüsü olarak kosinüs benzerliği algoritmasını kullanmışlardır. Başlangıçta 10 adet eğitim örneğini belirsizlik örnekleme yöntemiyle seçmişlerdir. Çalışmalarında başlangıç eğitim kümesini oluşturmak için kümeleme yönteminin kullanılmasının aktif öğrenmenin performansını arttırdığı gözlemlenmiştir.

Barata vd. (2021), dengesiz sınıflandırma problemini ele almıştır ve bu problemde soğuk başlangıç aktif öğrenme üzerine çalışmışlardır. Havuz tabanlı aktif öğrenme (pool-based active learning) yöntemiyle toplu örnek seçimi yapmışlardır. Hesaplama açısından verimli olan Aykırı Değer Tabanlı Yansız Aktif Öğrenme (Outlier - Based Discriminative Active Learning - ODAL) yaklaşımını ılık başlangıç metodu gibi kullanarak üç aşamalı aktif öğrenme yöntemi sunmuşlardır. Belirsizlik ölçümü için farklı belirsizlik örnekleme stratejileri kullanmışlardır. Aykırı değer tabanlı yansız aktif öğrenmede ılık başlangıç yönteminin, standart yöntemlere göre daha hızlı şekilde yüksek performans elde edebileceğini göstermişlerdir.

Grimova vd. (2018), cold-start aktif öğrenme probleminin varlığında uyku aşamalarının sınıflandırılması üzerine çalışmışlardır. Başlangıç aşamasında etiketsiz örneklerin sınıflandırılması için Öklid mesafesini kullanarak 1NN algoritmasını kullanmışlardır. Elde edilen eğitim kümesini doğrusal Bayesian sınıflandırıcı ile eğitmişlerdir. Ortalama hata değerini, değerlendirme kriteri olarak kullanmışlardır. Önerilen yöntemin performansı k -ortalamalar yöntemi ile karşılaştırılmıştır ve önerilen yöntemin daha düşük ortalama hata değerine ulaştığı görülmüştür.

Jin vd. (2022), aktif öğrenmede soğuk başlangıç problemi üzerine çalışmışlardır. Soğuk başlangıç aktif öğrenme (Cold-Start Active Learning - CALR) yaklaşımını önermişlerdir. Hiyerarşik kümeleme tabanlı örnek seçim yöntemi kullanmışlardır. Görüntü sınıflandırma alanında çalışmalarını gerçekleştirmişlerdir. Dengeli İteratif Azaltma ve Hiyerarşik Kümeleme (Balanced Iterative Reducing and Clustering using Hierarchies - BIRCH) yöntemini kümeleme için kullanılmıştır. Oluşan kümelerin dağılımını en iyi temsil edecek örnekleri seçmek için bilgi yoğunluğunu hesaplamışlardır. Önerilen yöntemi üç veri kümesine uygulamışlardır. Önerilen yöntem, rassal seçim ve ılık başlangıç aktif öğrenme yöntemleri ile karşılaştırılmıştır. Deneysel sonuçlar, CALR yönteminin başlangıçta etiketli veri kullanmadan daha iyi örnek seçimi gerçekleştirdiğini göstermiştir.

Nguyen ve Smeulders (2004), aktif öğrenme ve kümeleme üzerine çalışmışlardır. Temsil gücü yüksek veri örnekleri elde etmek için başlangıç aşamasında k -medoid kümeleme

yöntemini uygulamışlardır. Sınıf etiketini tahmin etmek için lojistik regresyon yöntemini kullanmışlardır. Mevcut hataya en büyük katkısı olan örneği seçecek şekilde seçim stratejisi kullanmışlardır. Önerilen yaklaşım iki veri kümesinde uygulanmıştır. Performans değerlendirmesini sınıflandırma hatasındaki azalmaya göre yapmışlardır. Önerilen yöntem, rassal seçim, sınıra en yakın örneğin seçimi ve destek vektör makinasındaki medoid merkezlerinin seçimiyle karşılaştırılmıştır. Önerilen yöntemin daha iyi sonuç verdiği görülmüştür.

Yang ve Loog (2022), en yakın komşuluk tabanlı aktif başlangıç yöntemi önermişlerdir. Mesafe ölçütü olarak Öklid mesafesi kullanılmıştır. Etiketlenmemiş veri kümesinden en bilgilendirici örnekleri maksimum entropi yöntemiyle seçmişlerdir. Etiketli veri ve etiketsiz veri arasındaki benzersizliklerin en aza indirilmesine odaklanmışlardır. 16 veri kümesinde çalışma gerçekleştirmişlerdir. Önerilen yöntemle seçilen başlangıç örneklerini, rassal seçim, k-ortalamlar, D-Optimalite Temelli Transdüktif Deneysel Tasarım (D-Optimality Based Transductive Experimental Design – TED), Σ -Optimalite Temelli Grafik Varyansın Minimizasyonu (Σ -Optimality Based Graph Variance Minimization - Σ -opt), En Olası Yanlış Örneklerin Seçimi (The Selection of Most Probable Wrong Samples – pWrong), Üretici ve Ayrımcı Sınıflayıcılar (Generative and Discriminative Classifiers - Gen/Disc), Dirichlet Süreci Beklenen Doğruluk (Dirichlet Process Expected Accuracy – DPEA) seçim yöntemleriyle karşılaştırılmıştır ve önerilen yöntemin daha iyi sonuç verdiği görülmüştür.

Yapılan literatür araştırmasında, zaman serisi sınıflandırma probleminde başlangıçta etiketlenmemiş verilerin kullanılmasını dikkate alarak yapılan sınırlı sayıda çalışma olduğu görülmüştür. Bu boşluğu doldurmak için kümelemeye dayalı yoğunluk ve yerel aykırı değer tabanlı iki farklı başlangıç yöntemi önerilmiştir.

3. MATERİYAL ve YÖNTEM

Bu bölümde tez çalışmasında kullanılan veri kümeleri ve makine öğrenmesi yöntemleri açıklanmıştır. Bölüm 3.1’de tez çalışmasında kullanılan materyal, Bölüm 3.2’de kullanılan yöntemler açıklanmıştır.

3.1. Materyal

Bölüm 3.1.1’de zaman serileri tanımlanmıştır ve Bölüm 3.1.2’de aktif öğrenme yaklaşımı açıklanmıştır.

3.1.1. Zaman Serisi

Bu tez çalışmasında çok değişkenli zaman serisi verileriyle çalışılmıştır. Bir zaman serisi, zamana bağlı ardışık ölçümlerden elde edilen değerleri temsil eder ve zaman serisi sınıflandırması, yeni bir zaman serisinin sınıf etiketini belirlemeyi amaçlar (Esling & Agon, 2012).

Çok değişkenli zaman serisi, birçok değişkenin aynı anda ve aynı zaman aralığında ölçüldüğü ve kaydedildiği veriyi temsil eder (Buza, 2018).

Tek değişkenli ve T -boyutlu zaman serisi $X = [x_1, x_2, \dots, x_T]$ sıralı gerçekteki sayılar kümesidir. Çok değişkenli ve M -boyutlu zaman serisi X ise M farklı tek değişkenli zaman serisinden oluşur ve Denklem 3.1’deki gibi ifade edilmektedir (Fawaz vd., 2019).

$$X = [X^1, X^2, \dots, X^M], X^i \in \mathbb{R}^T \quad (3.1)$$

Veri kümesi $D = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, (X_i, Y_i) çiftlerinden oluşmaktadır. Burada, X_i çok ya da tek değişkenli zaman serisini, Y_i ise bu zaman serisine ilişkin sınıf etiketini göstermektedir (Fawaz vd., 2019).

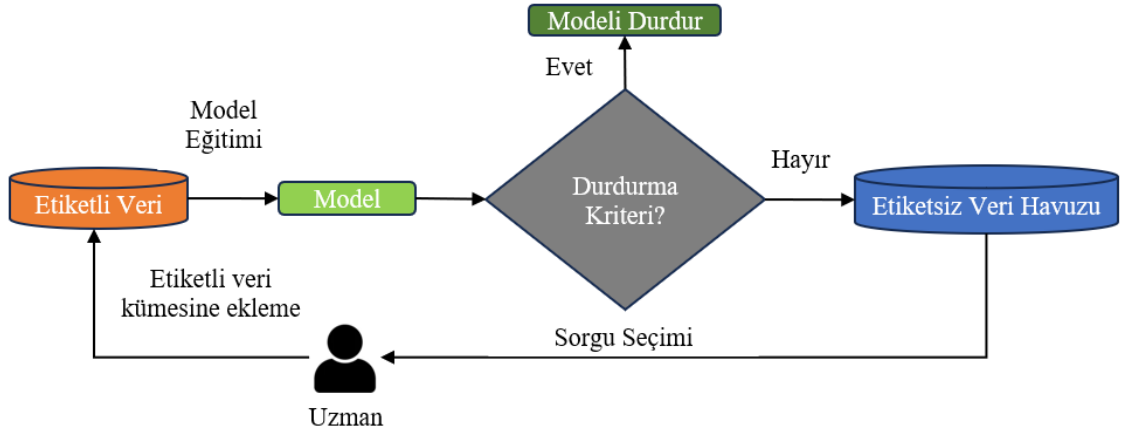
3.1.2. Aktif Öğrenme ve Sorgu Stratejisi

Aktif öğrenme (Active Learning-AL) yapay zekanın bir alt koludur. Aktif öğrenmede ana fikir, algoritma öğreneceği verileri kendi seçmesidir (Şebek vd., 2019). Şekil 3.1’de aktif öğrenme akışı gösterilmektedir (Koyuncu & İnkaya, 2023). Etiketlenmemiş veriler bir veri havuzunda bulunmaktadır ve bu verilerin sınıf etiketi bilinmemektedir. Sınıflandırıcı modelini eğitmek için etiketli verilere ihtiyaç vardır. Bu nedenle etiketlenmemiş veri havuzundan örnekler seçilmektedir. Sorgu stratejisi, etiketleme işlemi için en bilgilendirici örnekleri seçmeyi amaçlar (Aggarwal, 2015). Sorgu örneği seçim aşamasında farklı yaklaşımlar bulunmaktadır. Bu çalışmada belirsizlik örnekleme (uncertainty sampling) sorgu örneği seçim yaklaşımı olarak kullanılmıştır. Belirsizlik örneklemesinde karar sınırına yakın alanlarda bulunan örneklerin öğrenilmesinin daha değerli olduğunu varsayılmaktadır (Aggarwal, 2015). Belirsizlik örnekleme, etiketlenmemiş veri kümesinden en az kesin olan örneği seçer. $P(y|x)$, x örneğinin y sınıf etiketine dahil olma olasılığını göstermektedir. En az kesin olan x^* örneği, Denklem 3.2’ye ve Denklem 3.3’e göre hesaplanmaktadır (Prudencio vd., 2011).

$$x^* = \arg \max_x (1 - P(\hat{y}|x)) \quad (3.2)$$

$$\hat{y} = \arg \max_y (P(y|x)) \quad (3.3)$$

Sorgu örneği seçim işleminden sonra etiketleme işlemi gerçekleştirilmektedir. Verilere etiketleme işlemi yapan bir uzman, makine ya da bir veri tabanı sistemi olabilmektedir. Sorgu örneği, etiketleme işleminin ardından etiketli veri havuzuna eklenir. Bu süreç durdurma kriteri sağlanana veya etiketleme işlemi bitene kadar devam eder (Kumar & Gupta, 2020). Durdurma kriteri, maksimum sorgu örneği sayısı ya da hedeflenen doğruluğun sağlanması olarak düzenlenebilir.



Şekil 3.1. Aktif öğrenme

3.2. Yöntem

Bu bölümde, çalışma kapsamında kullanılan sınıflandırma yöntemleri, kümeleme yöntemleri ve yerel aykırı veri analizi açıklanmıştır. Bölüm 3.2.1, 3.2.2 ve 3.2.3'te sırasıyla K -en yakın komşuluk algoritması, karar ağacı algoritması ve Naive Bayes algoritması açıklanmıştır. Bölüm 3.2.4'te k -ortalamalar algoritması, Bölüm 3.2.5'te k -medoid algoritması, Bölüm 3.2.6'da DBSCAN algoritması ve Bölüm 3.2.7'de LOF algoritması açıklanmıştır. Bölüm 3.2.8'de tez çalışmasında önerilen yaklaşımlar açıklanmıştır. Bölüm 3.2.9, önerilen yaklaşımların performanslarını ölçmek için kullanılan değerlendirme metriklerini açıklamaktadır. Bölüm 3.2.10'da parametrik olmayan istatistiksel testler sunulmuştur.

3.2.1. K-En Yakın Komşuluk Yöntemi

K -en yakın komşu (K -nearest neighbor - KNN) algoritması, ilk olarak 1951 yılında Evelyn ve Hodges tarafından önerilmiştir (Silverman & Jones, 1989). Ardından 1967 yılında Cover ve Hart tarafından değiştirilmiştir (Cover & Hart, 1967). Denetimli öğrenme yöntemlerinden biridir. Mesafe bazlı bir algoritma olup, sınıflandırma işlemi komşu noktaların etiketlerine göre yapılmaktadır. KNN algoritması Şekil 3.2'de verilmiştir (Almomany vd., 2022). $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ veri kümesinde, x_i ve y_i sırasıyla veri örneği i 'yi ve onun sınıf etiketini göstermektedir. Her örneğin m öz niteliği olduğu varsayımıyla x_{ij} , i veri örneğinin j . öz niteliğini göstermektedir. Test örneği x^T 'nin

komşu noktaları bir uzaklık ölçütüne göre belirlenmektedir. Bu çalışmada Öklid uzaklığı kullanılmıştır. Her veri nesnesi x_i ile test örneği x^T arasındaki uzaklık $d(x^T, x_i)$ hesaplanmıştır. Komşuluk parametresi K olmak üzere test örneği x^T 'nin sınıf etiketi y^T , en yakın K komşusunun çoğunluk sınıf etiketine göre belirlenmektedir.

Algoritma 1: K-En Yakın Komşu (KNN) Algoritması

Girdi: D : veri kümesi, K : komşu sayısı x^T : test örneği

Çıktı: Test örneği x^T 'nin sınıf etiketi y^T

1: Öklid uzaklık fonksiyonu hesaplanarak test örneği x^T 'nin K en yakın komşuları bulunur.

$$d(x^T, x_i) = \sqrt{\sum_{j=1}^m (x_j^T - x_{ij})^2}$$

2: K komşu arasında en sık görülen sınıf etiketini dikkate alarak test örneği x^T 'nin sınıfını tahmin eder.

Şekil 3.2. KNN algoritması

3.2.2. Karar Ağacı Yöntemi

Sınıflandırma işlemi için kullanılan bir diğer yöntem Karar Ağacı (Decision Tree-DT) algoritmasıdır. Quinlan tarafından 1993 yılından önerilen bu algoritma basitlik, anlaşılabilirlik ve farklı türlerdeki veriyi kullanabilme özellikleriyle oldukça yaygın olarak kullanılan yöntemlerden biridir (Su & Zhang, 2006). Bu algoritma veri kümesini ağaç gibi alt dallara ayırır. Karar ağacı bir kök düğümle başlar. Ağaçtaki her bir iç düğüm, belirli bir niteliğe ilişkin kararı veya testi, her bir yaprak düğüm ise bir sınıf etiketini temsil eder (Charbuty & Abdulazeez, 2021).

Safsızlık kriteri, bir veri kümesindeki belirsizliğin ölçüsünü gösteren bir metriktir. Belirli bir bölünmenin ya da kararın verileri farklı sınıflara ne kadar iyi ayırdığını belirlemek için kullanılır. Amaç, karar ağacındaki her düğümde daha homojen veri alt kümeleriyle sonuçlanan, safsızlığı en aza indiren bölünmeyi bulmaktır. Bir veri kümesi homojen ise, veri kümesinden herhangi bir safsızlık yoktur (Tangirala, 2020). Entropi ve Gini değeri en bilinen safsızlık kriterlerindedir (Zhang & Gionis, 2023). Bu çalışmada Entropi ve Gini değerleri safsızlık kriteri olarak kullanılmıştır.

$D=\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ veri kümesinde, x_i ve y_i sırasıyla veri örneği i 'yi ve ilgili sınıf etiketini, c sınıf sayısını, p_j ise j sınıfına ait olma olasılığını göstermektedir. Entropi, sınıf etiketlerindeki belirsizliği en aza indirmeyi amaçlar ve 0 ile $\log_2(c)$ arasında bir değer alır. Entropi, Denklem 3.4'teki gibi hesaplanmaktadır (Tangirala, 2020).

$$Entropi = - \sum_{j=1}^c p_j \log_2(p_j) \quad (3.4)$$

Gini değeri, her sınıf için p_j olasılığının karelerin toplamı birden çıkartılarak hesaplanır. Hesaplama sonucunda 0 ile 1 arasında değer alır. Gini değeri, sınıf etiketlerinin homojenliğini arttırmayı hedeflemektedir. Gini değeri, Denklem 3.5'teki gibi hesaplanmaktadır (Tangirala, 2020).

$$GINI = 1 - \sum_{j=1}^c p_j^2 \quad (3.5)$$

3.2.3. Naive Bayes Yöntemi

Gözetimli öğrenme algoritmalarından biri olan Naive Bayes (NB) algoritması olasılığa dayalı sınıflandırma yöntemidir. NB algoritması, Bayes teoremini temel alarak geliştirilmiş bir algoritmadır. Bu algoritmanın büyük veri kümelerine uygulanması ve yorumlanması kolay olduğundan yaygın olarak kullanılan sınıflandırma yöntemlerindendir (Wu vd., 2008).

Bu çalışma kapsamında sayısal verilerle çalışan ve normal dağılım varsayımı yapan Gaussian Naive Bayes (GNB) algoritması kullanılmıştır. Gaussian Naive Bayes yöntemi, sınıfların öncül olasılıklarını ve özelliklerin sınıflara göre ortalamalarını ve standart sapmalarını hesaplar. x_i ve y_i sırasıyla veri örneği ve ilgili sınıf etiketini göstermektedir. $p(x_i|y)$, belirli bir y sınıfı verildiğinde x_i örneğinin gözlenme olasılığını göstermektedir. İlgili y sınıfındaki x_i 'nin ortalama değeri μ_y ve varyansı σ_y^2 olmak üzere GNB olasılığı Denklem 3.6'daki gibi hesaplanmaktadır (Jahromi & Taheri, 2017).

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}} \quad (3.6)$$

3.2.4. K-Ortalamlar Yöntemi

k -ortalamlar (k -means) algoritması 1990 yılında Kaufman ve Rousseauw tarafından önerilmiştir. k -ortalamlar, bir kümeleme algoritmasıdır ve benzer özelliklere sahip veri noktalarından küme oluşturmaya amaçlar. Kümeleme yöntemleri, gözetimsiz öğrenme yöntemlerinden biri olup, örneklerin benzerliğini dikkate alarak benzer örnekleri aynı kümeye, benzerliği olmayan örnekleri farklı kümeye atan bir gruplama yöntemidir. Örnekler arasındaki mesafeyi hesaplayarak ya da örnekler arasındaki benzerliği kullanarak gruplara atama işlemi yapılabilir (Maimon & Rokach, 2005).

k -ortalamlar yöntemi, bir veri kümesi D 'ye ait n örneği nc ayrı kümeye böler ve C kümesine atar. Her küme, kümedeki örneklerin ortalaması μ_{nc} tarafından tanımlanır. Küme atamaları, karesel hatalar toplamı (Sum of Squared Errors - SSE) dikkate alınarak belirlenir. SSE değeri, her noktanın en yakın merkeze olan karesel Öklid uzaklıklarının toplamı olarak tanımlanır. Amaç SSE değerini minimize etmektir (Likas vd., 2003).

SSE değeri, her örneğin küme merkezine olan uzaklığının karesi $\|x_j - CC_i\|^2$, $SSE = \sum_{i=1}^{nc} \sum_{x_j \in C_i} \|x_j - CC_i\|^2$ olarak hesaplanır (Thinsungnoena vd., 2015). Burada, $D = \{x_1, \dots, x_n\}$ veri kümesinde sırasıyla n örnek sayısı, nc küme sayısı, $C = \{C_1, \dots, C_{nc}\}$ kümeler ve CC_i , i . küme merkezini göstermektedir.

k -ortalamlar algoritmasında başlangıç küme merkezleri seçilir. Herhangi bir x_j nesnesi en yakın kümeye atanır. Atama işleminin ardından, her küme için küme merkezleri tekrardan hesaplanır. Herhangi bir küme merkezinde değişiklik olup olmadığını kontrol edilir. Eğer merkezler değişmediyse, algoritma sonlandırılır; aksi takdirde, 2. adıma geri dönlür. k -ortalamlar algoritması Şekil 3.3'te verilmiştir (Thinsungnoena vd., 2015).

Algoritma 2: K-Ortalamlar Algoritması

Girdi: D : veri kümesi, n : örnek sayısı, C : kümeler, CC : küme merkezleri, nc : küme sayısı

Çıktı: küme merkezleri

1: $C = \{C_1, \dots, C_{nc}\}$ kümeleri için başlangıç küme merkezleri

$CC_i = \{CC_1, \dots, CC_{nc}\}$ seçilir.

2: Küme atamaları hesaplanır.

x_j nesnesi, C_i^* kümesine ancak ve ancak

$$i^* = \underset{i=1, \dots, nc}{\operatorname{argmin}} \|x_j - CC_i\|^2$$

sağlanırsa atanır.

3: Küme merkezleri tekrardan hesaplanır.

$$CC_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

5: Eğer küme merkezlerinde değişiklik yoksa algoritma tamamlanır. Aksi takdirde, 2. aşamaya geri gidilir.

Şekil 3.3. k -ortalamlar algoritması

3.2.5. K-Medoid Yöntemi

k -medoid yöntemi, bölümlleme algoritmalarından biri olup Şekil 3.4'te gösterilmektedir (Han vd., 2011). D veri kümesinde n adet örnek bulunmaktadır ve amaç nc küme için nc adet temsilci belirlemektir. Bu sebeple, her veri örneği ile onun temsilcisine olan benzerliklerin toplamı minimize edilmek istenir.

Sırasıyla nc küme sayısı, veri nesnesi x , C_i kümesinin temsili örneği o_i ve benzemezlik fonksiyonu $d(x, o_i)$ için kümeleme hatası $E = \sum_{i=1}^{nc} \sum_{x \in C_i} d(x, o_i)$ hesaplanır. Bir kümenin içindeki elemanlar temsili örnek ile yer değiştirdiğinde (swapping), kümeleme hatası E yeni temsili örnek için yeniden hesaplanır. Küme içindeki elemanların yer değişikliğinden kaynaklanan toplam değiştirme maliyeti S , temsili örneklerden biri temsil olmayan bir örnek ile değiştirildiğinde kümeleme hatası değerindeki fark olarak hesaplanır. Eğer toplam maliyet pozitif ise ($S > 0$), bir değişiklik yapılmaz. Fakat toplam maliyet negatif ise ($S < 0$), temsili örnek o_j , o_{rassal} örneği ile değiştirilir. Toplam maliyette iyileşme olmayıncaya kadar iterasyonlar devam ettirilir.

Algoritma 3: *k*-medoid Algoritması

Girdi: *D*: veri kümesi, *n*: örnek sayısı, *nc*: küme sayısı

Çıktı: *k* kümeler dizini

1: Veri kümesi *D* içinden *nc* örnek veriyi temsili başlangıç değeri olarak seç.

2: **repeat**

3: Her bir örnek, en yakın temsilciye sahip kümeye atanır.

4: Temsili olmayan *o_{rassal}* değeri rassal olarak seçilir.

5: Temsil örneği *o_j* ile temsil olmayan *o_{rassal}* örneğinin toplam değiştirme maliyeti *S* hesaplanır.

6: **if** *S* < 0 **then**; *o_j*, *o_{rassal}* ile değiştirilir.

7: **until** değişiklik olmayana kadar.

Şekil 3.4. *k*-medoid algoritması

3.2.6. DBSCAN Yöntemi

DBSCAN algoritması yoğunluk tabanlı kümeleme yapmak için kullanılan bir algoritma olup 1996 yılında Ester tarafından geliştirilmiştir (Ester vd., 1996). DBSCAN algoritmasının iki önemli değeri bulunmaktadır. Bunlar ϵ ve minimum nesne sayısı (*minpts*) değeridir. ϵ değeri, veri noktaları etrafında oluşturulan kümenin yarıçapını göstermektedir. ϵ -komşuluk, küme içerisinde belirli uzaklıktaki diğer noktaları, *minpts* değeri bir noktanın çekirdek nokta olarak değerlendirilmesi için gereken nokta sayısını göstermektedir. DBSCAN algoritması tamamlandıktan sonra çekirdek nokta, sınır nokta ve gürültü noktası olmak üzere üç nokta oluşur. Bir nokta, ϵ -komşuluk mesafesi içerisinde en az *minpts* kadar noktaya sahipse bu çekirdek nokta olarak adlandırılır. Bir çekirdek noktanın etrafındaki bölgede bulunan ve çekirdek nokta olmayan nokta sınır nokta olarak adlandırılır. Bir nokta çekirdek ya da sınır nokta olarak tanımlanmazsa bu noktaya gürültü noktası denmektedir.

D veri kümesinde *x* ve *q* veri noktaları olmak üzere, $d(x, h)$ uzaklık ölçütünde *x*'in ϵ -komşuluk değeri $N_\epsilon(x)$ formülasyonu Denklem 3.7'de gösterilmektedir (Gan vd., 2020).

$$N_\epsilon(x) = \{q \in D: d(x, q) \leq \epsilon\} \quad (3.7)$$

x veri noktası, $|N_\epsilon(x)| \geq \text{minpts}$ ise çekirdek nokta olarak belirlenir. Eğer $q \in D$ noktası, bir çekirdek nokta olan x 'in ϵ -komşuluğu içinde bulunuyorsa, q noktası x noktasından doğrudan yoğunluk ulaşılabilir durumdadır (Gan vd., 2020).

DBSCAN algoritmasında ilk olarak ϵ ve minpts değerleri belirlenir. D veri kümesindeki tüm noktalar ziyaret edilmemiş olarak dikkate alınmaktadır. Ziyaret edilmemiş her bir x noktası için döngü başlatılmaktadır. x noktası, seçildikten sonra ziyaret edildi olarak işaretlenir. Eğer ϵ yarıçapında x 'in komşu bölgesindeki noktaların sayısı minpts 'den büyük veya eşitse yeni bir C kümesi oluşturulur ve H noktalar kümesi C 'ye atanır. H , x 'in komşuluk yarıçapı ϵ içindeki noktaları kümesini temsil eder. H noktalar kümesindeki her x için, x ziyaret edilmemişse, ziyaret edildi olarak etiketlenir. Eğer x 'in komşu bölgesinde ϵ yarıçapındaki noktaların sayısı minpts değerinden büyükse, bu noktalar x 'e atanır. x herhangi bir kümeye ait değilse, C kümesine atanır. Aksi durumda, x gürültü olarak etiketlenir. Tüm noktalar ziyaret edilene kadar döngü tekrarlanır. Şekil 3.5'te DBSCAN algoritması verilmiştir (Yang vd., 2022).

3.2.7. Yerel Aykırı Faktör Algoritması

Aykırı değer, veri kümesinde bulunan normal veri yapısının dışında kalan değerlere denir (Singh & Upadhyaya, 2012). Aykırı değerler, yerel ve global olmak üzere ikiye ayrılmaktadır. Bir veri noktası, veri kümesindeki tüm noktalardan uzaktaysa global aykırı değer, komşu noktalarını dikkate alarak farklı bir yapı gösteriyorsa yerel aykırı değer olarak tanımlanır ve model performansı performansını doğrudan etkileyen değerler olup, aykırı değerler tespit edilip kaldırılamazsa verilere uygulanan model yanıltıcı olabilir (Alghushairy vd., 2020).

Aykırı değerleri tespit etmek için farklı yaklaşımlar bulunmaktadır. Wang vd. (2019), aykırı değer tespit yöntemlerini istatistik tabanlı, mesafe tabanlı, yoğunluk tabanlı, kümeleme tabanlı, topluluk tabanlı ve öğrenmeye dayalı teknikler olmak üzere altı gruba ayırmıştır. Bu çalışmada, aykırı değerlerin tespitinde yoğunluk tabanlı yöntemlerden olan yerel aykırı faktör (Local outlier factor-LOF) yöntemi kullanılmıştır.

Algoritma 4: DBSCAN Algoritması

Girdi: D : veri kümesi, x : veri noktası, H : noktalar kümesi, ε : komşuluk mesafesi, $minpts$: minimum nesne sayısı

Çıktı: C kümeler

1: Tüm noktalar ziyaret edilmemiş olarak etiketlenir

2: **Do**

3: Rastgele x noktası seçilir

4: x noktası ziyaret edilmiş olarak etiketlenir

5: **if** ε yarıçapında x 'in komşu bölgesindeki noktaların sayısı $minpts$ 'den büyük veya eşitse **then**

6: Yeni bir C kümesi oluştur ve H 'ı C 'ye ata

7: H , x 'in komşuluk yarıçapı ε içindeki noktalar kümesi

8: **for** H 'taki her x noktası için **do**

9: **if** x ziyaret edilmemişse **then**

10: x 'i ziyaret edilmiş olarak etiketle

11: **if** x 'in komşu bölgesinde ε yarıçapındaki noktaların sayısı $> minpts$ **then**

12: Tüm noktaları x 'e ata

13: **if** x herhangi bir kümeye ait değilse **then**

14: x 'i C 'ye ata

15: **end if**

16: **end if**

17: **end if**

18: **end for**

19: çıktı C kümesi

20: **else**

21: x 'i gürültü olarak etiketle

22: **end if**

23: tüm noktalar ziyaret edilene kadar tekrarlanır

Şekil 3.5. DBSCAN algoritması

Yerel aykırı faktör (local outlier factor-LOF) yöntemi ilk olarak 2000 yılında tanıtılmıştır (Breunig vd., 2000). LOF algoritmasında D veri kümesinde p ve o nesne olmak üzere, bu veri nesneleri arasındaki $d(p,o)$ Öklid mesafesi hesaplanır ve küçükten büyüğe sıralanır. K . sıradaki nesnenin p nesnesine olan uzaklığı $K-distance(p)$ hesaplanır. Ardından her bir p nesnesi için K -komşuluk mesafesi $N_{K-distance(p)}(p)$ hesaplanır. K değeri dikkate alınarak, p nesnesinin o nesnesine göre ulaşılabilirlik mesafesi $reachdist_K(p,o)$ hesaplanır. Hesaplanan ulaşılabilirlik mesafesi, p nesnesinin yerel erişilebilirlik yoğunluğu $Lrd_{minpts}(p)$ hesaplanır. Bir önceki aşamada bulunan $Lrd_{minpts}(p)$ değeri kullanılarak $LOF_{minpts}(p)$ değeri hesaplanır ve aykırı değerler elde edilir. Algoritma

sonucunda elde edilen $LOF_{minpts}(p)$ değeri > 1 ise ilgili nesne aykırı değer olarak atanır. Şekil 3.6’da LOF algoritmasının adımları gösterilmektedir (Breunig vd., 2000).

Algoritma 5: Yerel Aykırı Faktör Algoritması

Girdi: D : veri kümesi, $p, o \in D$, $d(.,.)$: uzaklık fonksiyonu, K değeri, $minpts$

Çıktı: Yerel aykırı değerler

1: p ve o nesnesi için $d(p,o)$ Öklid uzaklık fonksiyonu hesaplanır.

2: K . sıradaki nesnenin p nesnesine olan uzaklığı $K-distance(p)$ hesaplanır.

2: p nesnesinin K -komşuluk değerleri hesaplanır.

$N_{K-distance(p)}(p) = \{q \in D \setminus \{p\} \mid d(p, q) \leq K - distance(p)\}$

3: p nesnesinin o nesnesine olan ulaşılabilirlik mesafesi hesaplanır.

$$reachdist_K(p, o) = \max \{K - distance(o), d(p, o)\}$$

4: p nesnesinin yerel erişilebilirlik yoğunluğu $Lrd_{minpts}(p)$ hesaplanır.

$$Lrd_{minpts}(p) = 1 / \frac{\sum_{o \in N_{minpts}(p)} reachdist_{minpts}(p, o)}{|N_{minpts}(p)|}$$

5: p nesnesi için LOF değeri hesaplanır.

$$LOF_{minpts}(p) = \frac{\sum_{o \in N_{minpts}(p)} \frac{Lrd_{minpts}(o)}{Lrd_{minpts}(p)}}{|N_{minpts}(p)|}$$

Şekil 3.6. LOF algoritması

3.2.8. Önerilen Yaklaşımlar

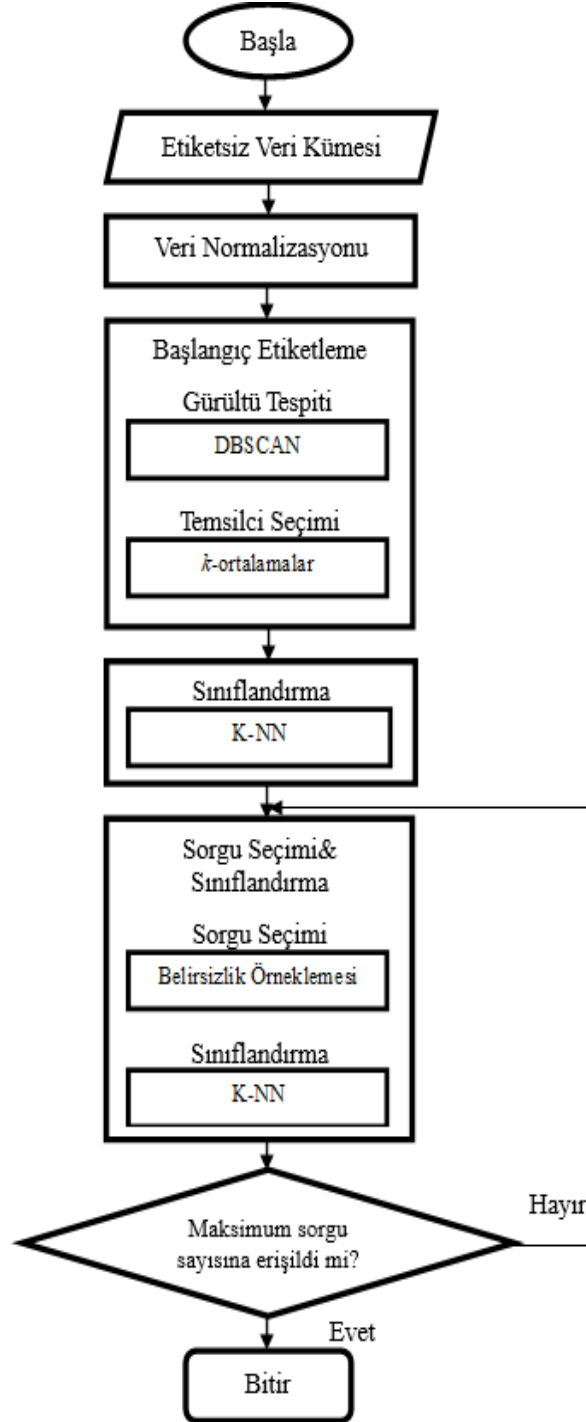
Bu tez çalışmasında zaman serisi sınıflandırma probleminin çözümü için iki yaklaşım önerilmektedir. Başlangıçta tüm veriler etiketsizdir. Etiketsiz veri kümesine öznitelikler arasındaki ölçüm farklılıklarının ortadan kaldırılması için normalizasyon uygulanmıştır. Soğuk başlangıçtaki etiketlemeyi doğru yapabilmek, model performansını artırmak ve daha az veri kullanarak daha iyi öğrenme stratejisi elde etmek amacıyla üzere iki farklı başlangıç örnek seçim algoritması önerilmiştir. Sınıflandırma işlemi KNN algoritmasıyla yapılmıştır. Ardından temsil gücü yüksek sorgu örneklerin seçimi ve etiketlenmesi için sorgulama stratejisi kullanılmıştır.

İlk yaklaşımda, öncelikle etiketsiz veri kümesine Z-skor normalizasyon uygulanmıştır. Z-skor normalizasyon, x nesne, μ ortalama ve σ standart sapma olmak üzere Denklem 3.8'deki gibi hesaplanmaktadır.

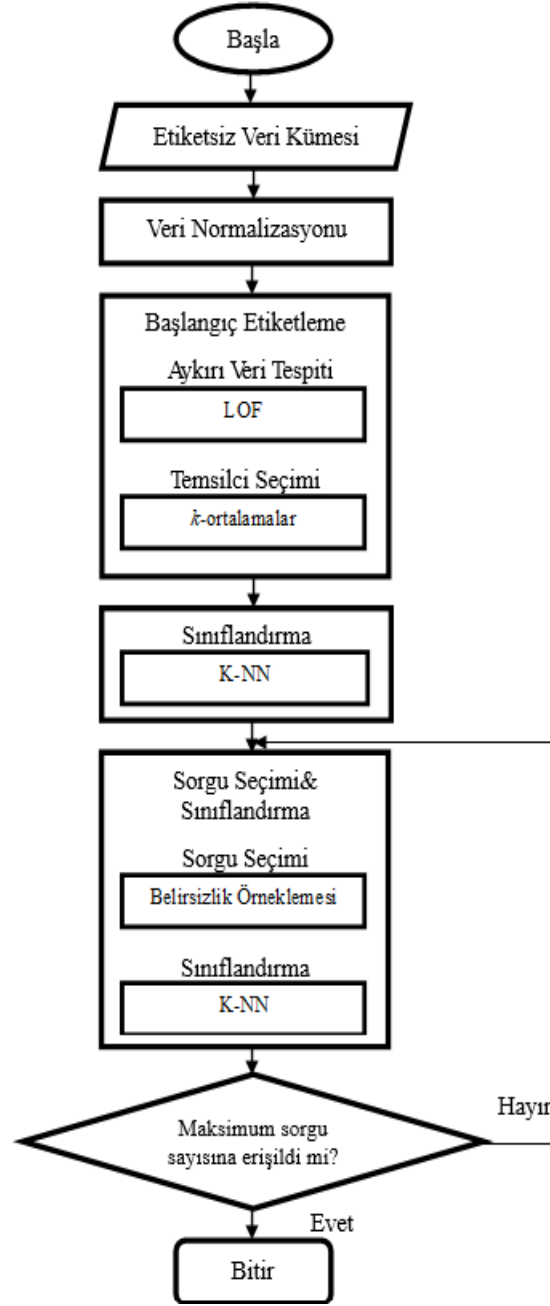
$$Z = \frac{x - \mu}{\sigma} \quad (3.8)$$

Başlangıçta etiketlenecek örneklerin seçiminde, öncelikle gürültülü verilerin tespiti ve etiketsiz veri havuzundan çıkarılması için DBSCAN algoritması uygulanmıştır. Etiketsiz veri havuzunda çekirdek ve sınır noktalara k -ortalamlar yöntemi uygulanarak nc adet eğitim örneği seçilir. Seçilen eğitim örnekleri, KNN algoritması ile sınıflandırılır. Ardından sorgu seçimi belirsizlik örnekleme yöntemiyle gerçekleştirilir. Seçilen sorgu örnekleri KNN algoritması ile sınıflandırılır. Belirlenen maksimum sorgu sayısına ulaşırsa algoritma sonlandırılır. Aksi durumda; sorgu örneği seçim ve etiketleme işlemine devam edilir. Önerilen DBSCAN+ k -ortalamlar yöntemi Şekil 3.7'de gösterilmiştir.

İkinci yaklaşımda, etiketsiz veri kümesine Z-skor normalizasyon uygulanmıştır. Başlangıçta etiketlenecek örneklerin seçiminde aykırı verilerin tespiti ve etiketsiz veri havuzundan çıkarılması için LOF algoritması uygulanmıştır. Etiketsiz veri havuzunda kalan noktalara k -ortalamlar yöntemi uygulanarak nc adet eğitim örneği seçilir. Seçilen eğitim örnekleri, KNN algoritması ile sınıflandırılır. Ardından sorgu seçimi belirsizlik örnekleme yöntemiyle gerçekleştirilir. Seçilen sorgu örnekleri KNN algoritması ile sınıflandırılır. Belirlenen maksimum sorgu sayısına ulaşırsa algoritma sonlandırılır. Aksi durumda; sorgu örneği seçim ve etiketleme işlemine devam edilir. Önerilen LOF+ k -ortalamlar yöntemi Şekil 3.8'de gösterilmiştir.



Şekil 3.7. Önerilen DBSCAN+k-ortalamalar akış şeması



Şekil 3.8. Önerilen LOF+k-ortalamlar akış şeması

3.2.9. Performans Değerlendirme Metrikleri

Önerilen yaklaşımların performansı doğruluk ve ağırlıklı F-ölçütü ile değerlendirilmiştir. Doğruluk değeri, Çizelge 3.1’de yer alan karmaşıklık matrisi dikkate alınarak hesaplanmıştır.

Çizelge 3.1. Karmaşıklık matrisi

		Tahmini Sınıf	
		Pozitif	Negatif
Gerçek Sınıf	Pozitif	Doğru Pozitif	Yanlış Negatif
	Negatif	Yanlış Pozitif	Doğru Negatif

Doğru pozitif (DP) değeri, modelin doğru bir şekilde pozitif olarak sınıflandırdığı pozitif örneklerin sayısını, doğru negatif (DN) değeri, modelin doğru bir şekilde negatif olarak sınıflandırdığı negatif örneklerin sayısını, yanlış pozitif (YP), modelin hatalı bir şekilde pozitif olarak sınıflandırdığı negatif örneklerin sayısını ve yanlış negatif (YN), modelin negatif olarak sınıflandırdığı pozitif örneklerin sayısını göstermektedir. Doğruluk, doğru pozitif ve doğru negatif değerlerin tüm değerlere bölünmesiyle hesaplanmaktadır ve Denklem 3.9’da verilmiştir.

$$\text{Doğruluk} = \frac{DP+DN}{DP+DN+YP+YN} \quad (3.9)$$

Kullanılan bir diğer değerlendirme yöntemi ağırlıklı F-ölçütüdür. F-ölçütü, kesinlik ve duyarlılık metriklerinin harmonik ortalaması olarak tanımlanır (Flach, 2019). F-ölçütü (F), Denklem 3.10’deki gibi hesaplanmaktadır.

$$F = \frac{2DP}{2DP+YP+YN} \quad (3.10)$$

Ağırlıklı F-ölçütü, F-ölçütü değerinin her sınıf için ağırlıklı ortalamasını ifade etmektedir. i . sınıfın ağırlığını ifade eden w_i değeri, Denklem 3.11’deki gibi hesaplanmaktadır. i . sınıf için hesaplanan F_i değeriyle w_i değeri çarpılarak ağırlıklı F-ölçütü hesaplanmaktadır. Ağırlıklı F-ölçütü, Denklem 3.12’deki gibi hesaplanmaktadır.

$$w_i = \frac{i. \text{ sınıftaki örnek sayısı}}{\text{toplam örnek sayısı}} \quad (3.11)$$

$$\text{Ağırlıklı F-ölçütü} = \sum_{i=1}^C w_i F_i \quad (3.12)$$

3.2.10. Parametrik Olmayan İstatistiksel Testler

Modelleri karşılaştırmak için farklı performans değerlendirme yöntemleri kullanılabilmektedir. Yöntemler arasında istatistiksel olarak anlamlı farkların olup olmadığını belirlemek amacıyla parametrik olmayan istatistiksel testler kullanılmaktadır. Bu çalışmada, parametrik olmayan Friedman testi ve post-hoc Nemenyi testi kullanılmıştır.

Friedman testi 1937 yılında geliştirilmiş bir test olup, karşılaştırılan yöntemlerin ortalama sıralamalarını hesaplar ve sonuçları karşılaştırır (Friedman, 1937). r_i^j , L adet veri kümesi ve g adet karşılaştırma modeli için j . karşılaştırma modelinin i . veri kümesindeki sıralamasını göstermektedir. Her bir yöntemin ortalama sıralaması $R_j = \frac{1}{L} \sum_{i=1}^L r_i^j$ olarak hesaplanmaktadır (Demšar, 2006).

Friedman testinde, gruplar arasında fark olmadığını gösteren H_0 hipotezi kurulur. Eğer Friedman testi sonucunda elde edilen p-değeri $< 0,05$ ise H_0 hipotezi red edilir ve gruplar arasında anlamlı fark var demektir. p-değeri $\geq 0,05$ ise H_0 hipotezi red edilemez ve gruplar arasında anlamlı fark yok demektir. Friedman testi, $k-1$ serbestlik derecesine göre dağılım sonucu vermekte olup Denklem 3.13'teki gibi hesaplanmaktadır (Ma vd., 2022).

$$\chi_F^2 = \frac{12L}{g(g+1)} \left[\sum_j R_j^2 - \frac{g(g+1)^2}{4} \right] \quad (3.13)$$

H_0 hipotezi reddedildiğinde yöntemler arasındaki farkı belirlemek için Nemenyi testi gibi post-hoc ikili karşılaştırmalar testi yapılmaktadır (Elamir, 2022). Bu test her yöntem için dengeli bir dağılım varsayımı yapar (Pohlert, 2014).

Nemenyi testi, 1963 yılında geliştirilmiştir (Nemenyi, 1963). Tüm popülasyonlardan alınan verilerin bir arada sıralandığı iki yönlü çoklu karşılaştırma testidir (Liu & Chen, 2012). Nemenyi testi, kritik mesafeyi dikkate alır. Student t-testi kullanılarak elde edilen aralık istatistiği q_α olup kritik mesafe formülü Denklem 3.14’te gösterilmektedir (Ma vd., 2022).

$$CD = q_\alpha \sqrt{\frac{g(g+1)}{6L}} \quad (3.14)$$

4. BULGULAR ve TARTIŞMA

Bu bölümde önerilen yaklaşımların dört farklı veri kümesine uygulanması ile otomotiv sektöründeki vaka analizi bulguları sunulmuş ve tartışılmıştır. Bölüm 4.1’de uygulama yapılan veri kümeleri ve uygulama detayları hakkında bilgi verilmektedir. Bölüm 4.2’de veri kümelerinden elde edilen bulgular ve tartışma sunulmuştur. Bölüm 4.3’te ise otomotiv sektöründeki vaka analizi sunulmuştur.

4.1. Veri Kümeleri ve Uygulama Detayları

Bu çalışmada önerilen yaklaşımların performansını ölçmek amacıyla dört zaman serisi sınıflandırma verisi ele alınmıştır. Bu veri kümeleri ECG5000, Wafer, FordA ve FordB veri kümeleridir (Bagnall, 2024). Veri kümelerinin özellikleri Çizelge 4.1’de gösterilmektedir.

Çizelge 4.1. Veri kümesi özellikleri

Veri Kümesi	Eğitim Örneği Sayısı	Test Örneği Sayısı	Öznitelik Sayısı	Sınıf Sayısı
ECG5000	500	4500	140	5
Wafer	1000	6164	152	2
FordA	3601	1320	500	2
FordB	3636	810	500	2

ECG5000 veri kümesi, Physionet’ten indirilen 20 saatlik EKG verisini içerir. Veri setinde iki aşamalı veri ön işleme işlemi yapılmıştır. İlki kalp atışlarını veri kümesinden ayırmak, ikincisi ise kalp atışı verilerini interpolasyon yöntemiyle eşit uzunluklara dönüştürmektir. Bu veri kümesinde 500 eğitim örneği ve 4500 test verisi olmak üzere 5000 örnek bulunmaktadır. Veri kümesinde beş farklı sınıf bulunmaktadır. Dengesiz sınıf problemi vardır.

Wafer veri kümesi, yarı iletken üretimine ilişkin verileri içermektedir. Veri kümesi, tabaka silikonlarının üretimi sırasında sensörlerden kaydedilen verileri içermektedir. Veri kümesinde normal ve anormal olmak üzere iki sınıf vardır. Normal ve anormal sınıf

etiketleri arasında dengesiz sınıf etiketi dağılımı mevcuttur. Veri kümesinde 1000 eğitim örneği ve 6164 test örneği olmak üzere toplam 7164 veri bulunmaktadır.

FordA ve FordB veri kümeleri, otomotiv üretiminde kaydedilen motor gürültü ölçümlerini içermektedir. FordA eğitim ve test verileri normal çalışma koşullarında, minimum gürültü kirliliği altında oluşturulmuştur. Toplam 4921 örnek ve iki sınıf bulunmaktadır. FordB eğitim kümesi normal çalışma koşullarında, test kümesi ise gürültülü koşullar altında oluşturulmuştur. FordB veri kümesinde toplam 4446 örnek ve iki sınıf bulunmaktadır. Her iki veri kümesinde düşük oranda da olsa dengesiz sınıf etiketi dağılımı problemi vardır.

Deneysel çalışmada etiketsiz eğitim kümeleri, başlangıç örneklerinin seçilmesinde kullanılmıştır. Test veri kümeleri, modelin performansını değerlendirmek için kullanılmıştır.

Başlangıç aşamasında algoritmayı eğitmek için sırasıyla beş ile on arasında eğitim örneği seçilmiştir. Tüm deneysel çalışmalar Python platformunda gerçekleştirilmiştir. Başlangıç eğitim örneklerinin seçiminde kullanılan her yöntem için beş tekrar yapılmıştır ve bu çalışmada yapılan tekrarların ortalamaları sunulmuştur.

Önerilen yaklaşımların performansını karşılaştırmak için farklı sınıflandırma ve kümeleme yöntemleri kullanılmıştır. Başlangıç aşamasında temsilci örneklerin belirlenmesinde rassal seçim ile k -ortalamlar, k -medoid ve DBSCAN yöntemleri karşılaştırılmıştır. Karşılaştırılan sınıflandırma algoritmaları ise şunlardır: KNN, DT, GNB.

KNN yönteminde en yakın komşuların sayısı belirlemek için $K = \{1, 3, 5\}$ değerleri dikkate alınmıştır ve bu çalışmada 1NN, 3NN ve 5NN olarak bahsedilecektir. KNN uzaklık fonksiyonu olarak Öklid mesafesi kullanılmıştır.

DT yönteminde, safsızlık kriteri olarak Gini ve Entropi, ağacın maksimum derinliği için $\{1, \dots, 10\}$, iç düğümü bölmek için gereken minimum örnek sayısı $\{2, \dots, 10\}$ ve bir yaprak

düğümde olması gereken minimum örnek sayısı $\{1, \dots, 5\}$ ile deneysel çalışma yapılmıştır. En iyi doğruluk sonucu veren parametre değerleri için sınıflandırıcı eğitilmiştir.

DBSCAN yönteminde, $minpts = 4$ olarak alınmıştır (Gaonkar & Sawant, 2013). ϵ -komşuluk değeri, k -mesafe grafiği çizilerek belirlenmiştir (Gaonkar & Sawant, 2013). Burada, $k = 4$ değeri ele alınmıştır. Her veri kümesi için kullanılan ϵ değerleri Çizelge 4.2’de verilmiştir. LOF yöntemi uygulanırken $minpts = 4$ değeri kullanılmıştır.

Çizelge 4.2. DBSCAN yönteminde ϵ değerleri

Veri Kümesi	ECG5000	Wafer	FordA	FordB
ϵ değeri	10	11	23	24

Başlangıç aşamasında DBSCAN ve LOF yöntemiyle belirlenen gürültülü ve aykırı veriler etiketsiz veri kümesinden çıkartılmıştır. Çizelge 4.3’te önerilen yaklaşımlar için DBSCAN ve LOF yöntemleriyle veri kümesinden çıkartılan gürültülü ve aykırı veri sayıları verilmiştir.

Çizelge 4.3. Gürültülü ve aykırı veri sayıları

	Tedarik Zinciri Verisi	ECG5000	Wafer	FordA	FordB
Gürültülü veri sayısı	6	73	6	664	652
Aykırı veri sayısı	39	30	76	11	0

Doğruluk ve ağırlıklı F-ölçütü değerlerinin elde edilmesi için Python, parametrik olmayan Friedman ve Nemenyi istatistiksel analizi için R yazılımı kullanılmıştır.

4.2. Deneysel Sonuçlar

Bu bölümde önerilen yaklaşımların veri kümelerine uygulanması ile elde edile sonuçlar sunulmuş ve tartışılmıştır.

Tüm yöntemler, başlangıçta seçilen örnek sayısı $nc = (5, \dots, 10)$, 150 sorgu örneği ve etiketleme işlemi için beş kez tekrarlanmıştır. Bu analizin sonucunda doğruluk ve ağırlıklı F-ölçütü değerlerinin ortalamaları alınmıştır. 1NN, 3NN ve 5NN yöntemleri için ortalama

doğruluk sonuçları EK 1’de ve ağırlıklı F-ölçütü sonuçları EK 2’de, DT için ortalama doğruluk sonuçları EK 3’te ve ağırlıklı F-ölçütü sonuçları EK 4’te ve GNB için ortalama doğruluk sonuçları EK 5’te ve ağırlıklı F-ölçütü sonuçları EK 6’da verilmiştir.

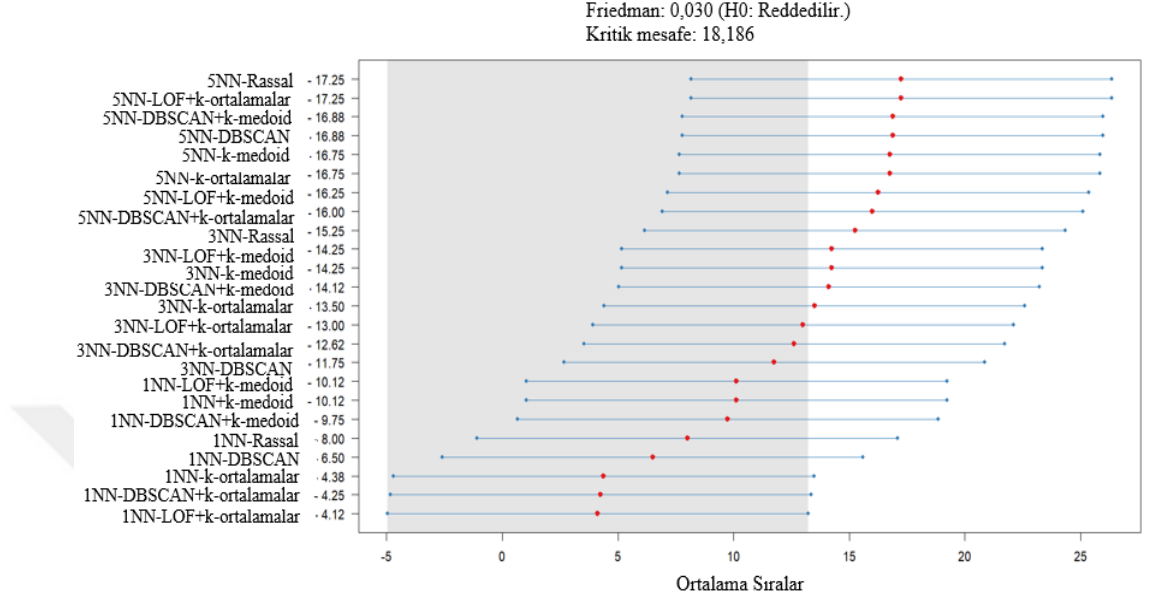
Önerilen yaklaşımlar ve karşılaştırılan yöntemler için Friedman ve Nemenyi testi uygulanmıştır ve analiz sonucunda elde edilen grafikler sunulmuştur. Grafikte y eksen, ortalama sıralamaları göstermektedir. Gri bölge, güven aralıklarını temsil etmektedir. Güven aralıkları, gruplar arasındaki sıralama farklılıklarının istatistiksel olarak anlamlı olup olmadığını değerlendirmek için kullanılır. %95 güven aralığında, her bir yöntem için kritik uzaklıklar gösterilmektedir. H_0 hipotezinin reddedilmesi yöntemler arasında istatistiksel olarak anlamlı fark olduğunu göstermektedir.

KNN sınıflandırıcısı içeren 24 yöntem karşılaştırılmıştır. 1NN, 3NN ve 5NN yöntemlerinde yapılan Friedman ve Nemenyi testleri için doğruluk (a) ve ağırlıklı F-ölçütü (b) grafikleri Şekil 4.1’de verilmiştir.

Şekil 4.1’den aşağıdaki çıkarımlar yapılmaktadır:

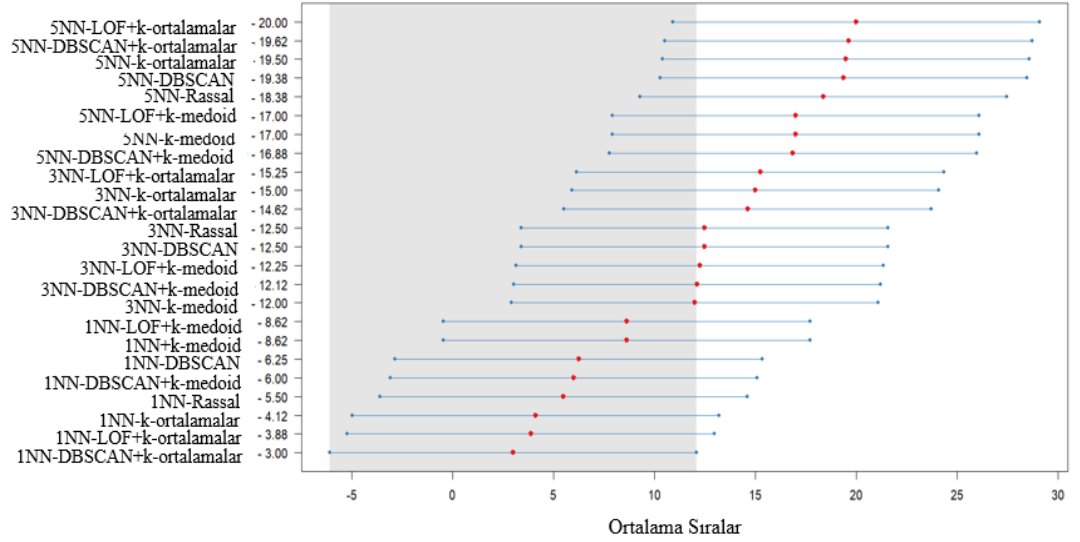
- Doğruluk grafiğinde, %95 güven aralığında bazı yöntemler için kritik uzaklıkların çakıştığı görülmektedir. Bu yöntemler arasında istatistiksel olarak anlamlı fark olmadığı görülmektedir. Friedman testi sonucunda elde edilen p-değerine bakıldığında, $p = 0,030 < 0,05$ olduğu için H_0 hipotezi reddedilir ve yöntemler arasında istatistiksel olarak anlamlı fark olduğu sonucuna ulaşılmaktadır. Ortalama sıralamalara bakıldığında en iyi iki yöntem sırasıyla 1NN-LOF+k-ortalamlar ve 1NN-DBSCAN+k-ortalamlar olarak belirlenmiştir.
- Ağırlıklı F-ölçütü grafiğinde, %95 güven aralığında bazı yöntemler için kritik uzaklıkların çakıştığı görülmektedir. Bu yöntemler arasında istatistiksel olarak anlamlı fark olmadığı görülmektedir. Friedman testi sonucunda elde edilen p-değerine bakıldığında, $p = 0,0 < 0,05$ olduğu için H_0 hipotezi reddedilir ve yöntemler arasında istatistiksel olarak anlamlı fark olduğu sonucuna

ulaşılmaktadır. Ortalama sıralamalara bakıldığında en iyi iki yöntem sırasıyla 1NN-DBSCAN+k-ortalamalar ve 1NN-LOF+k-ortalamalar olarak belirlenmiştir.



(a)

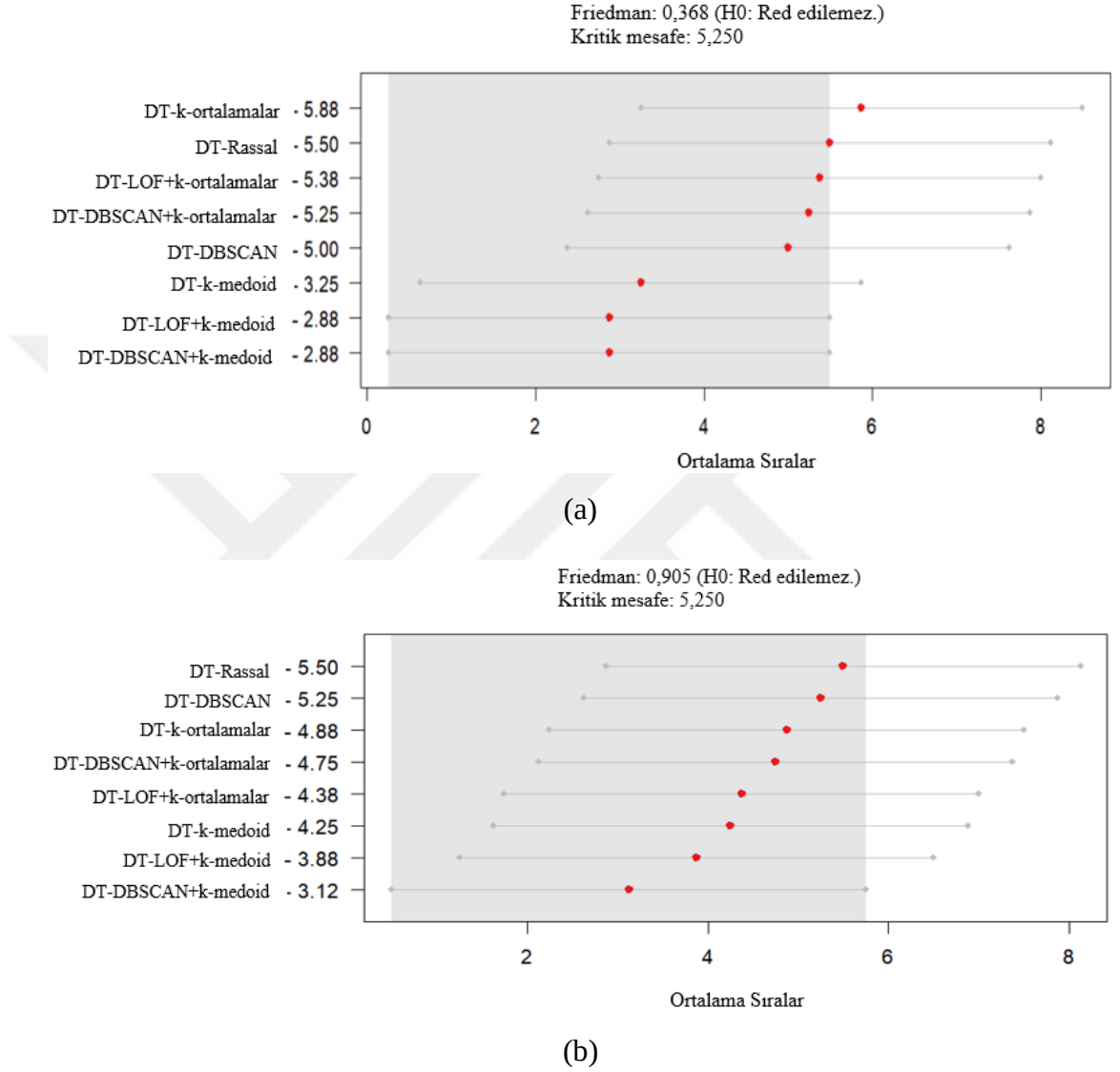
Friedman: 0,00 (H0: Reddedilir.)
Kritik mesafe: 18,186



(b)

Şekil 4.1. Friedman ve Nemenyi testleri 1NN, 3NN, 5NN ile başlangıç sorgu örneği seçimi karşılaştırma grafikleri

DT sınıflandırıcılarında sekiz yöntem karşılaştırılmıştır. Bu karşılaştırmada, yapılan Friedman ve Nemenyi testleri için doğruluk (a) ve ağırlıklı F-ölçütü (b) grafikleri Şekil 4.2’de verilmiştir.



Şekil 4.2. Friedman ve Nemenyi testleri DT sınıflandırma ile başlangıç sorgu örneği seçimi karşılaştırma grafikleri

Şekil 4.2’den aşağıdaki çıkarımlar yapılmaktadır:

- Doğruluk grafiğinde, %95 güven aralığında yöntemlerin kritik uzaklıklarının çakıştığı ve yöntemler arasında istatistiksel olarak anlamlı fark olmadığı görülmektedir. Friedman testi sonucunda elde edilen p-değerine bakıldığında,

$p = 0,368 \geq 0,05$ olduğu için H_0 hipotezi red edilemez ve yöntemler arasında istatistiksel olarak anlamlı fark olmadığı sonucuna ulaşılmaktadır. Ortalama sıralamalara bakıldığında en iyi iki yöntem sırasıyla DT-DBSCAN+k-medoid ve DT-LOF+k-medoid olarak belirlenmiştir.

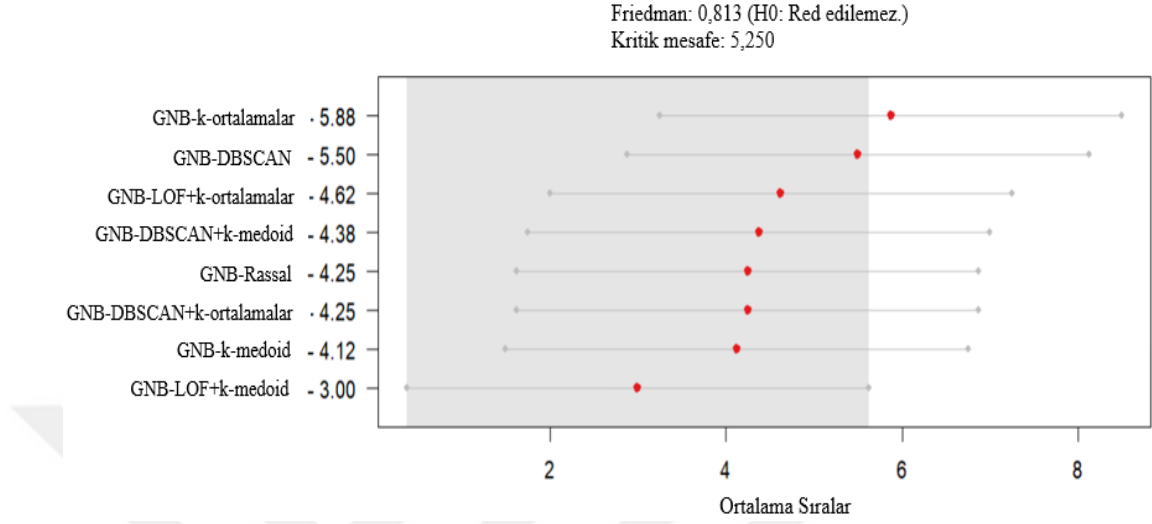
- Ağırlıklı F-ölçütü grafiğinde, %95 güven aralığında yöntemlerin kritik uzaklıklarının çakıştığı ve yöntemler arasında istatistiksel olarak anlamlı fark olmadığı görülmektedir. Friedman testi sonucunda elde edilen p-değerine bakıldığında, $p = 0,905 \geq 0,05$ olduğu için H_0 hipotezi red edilemez ve yöntemler arasında istatistiksel olarak anlamlı fark olmadığı sonucuna ulaşılmaktadır. Ortalama sıralamalara bakıldığında en iyi iki yöntem sırasıyla DT-DBSCAN+k-medoid ve DT-LOF+k-medoid olarak belirlenmiştir.

GNB sınıflandırıcılarında sekiz yöntem karşılaştırılmıştır. Bu karşılaştırmada, yapılan Friedman ve Nemenyi testleri için doğruluk (a) ve ağırlıklı F-ölçütü (b) grafikleri Şekil 4.3'te verilmiştir.

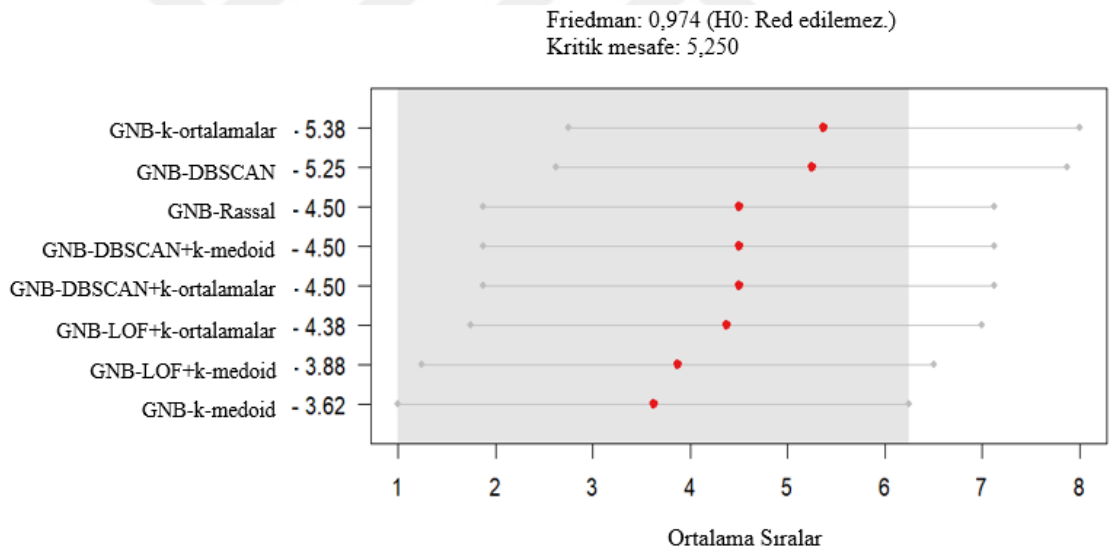
Şekil 4.3'ten aşağıdaki çıkarımlar yapılmaktadır:

- Doğruluk grafiğinde, %95 güven aralığında yöntemlerin kritik uzaklıklarının çakışması nedeniyle yöntemler arasında istatistiksel olarak anlamlı fark olmadığı görülmektedir. Friedman testi sonucunda elde edilen p-değerine bakıldığında, $p = 0,813 \geq 0,05$ olduğu için H_0 hipotezi red edilemez ve yöntemler arasında istatistiksel olarak anlamlı fark olmadığı sonucuna ulaşılmaktadır. Ortalama sıralamalara bakıldığında en iyi iki yöntem sırasıyla GNB-LOF+k-medoid ve GNB-k-medoid olarak belirlenmiştir.
- Ağırlıklı F-ölçütü grafiğinde, %95 güven aralığında yöntemlerin kritik uzaklıklarının çakışması nedeniyle yöntemler arasında istatistiksel olarak anlamlı fark olmadığı görülmektedir. Friedman testi sonucunda elde edilen p-değerine bakıldığında, $p = 0,974 \geq 0,05$ olduğu için H_0 hipotezi red edilemez ve yöntemler arasında istatistiksel olarak anlamlı fark olmadığı sonucuna ulaşılmaktadır.

Ortalama sıralamalara bakıldığında en iyi iki yöntem sırasıyla GNB-*k*-medoid ve GNB-LOF+*k*-medoid olarak belirlenmiştir.



(a)

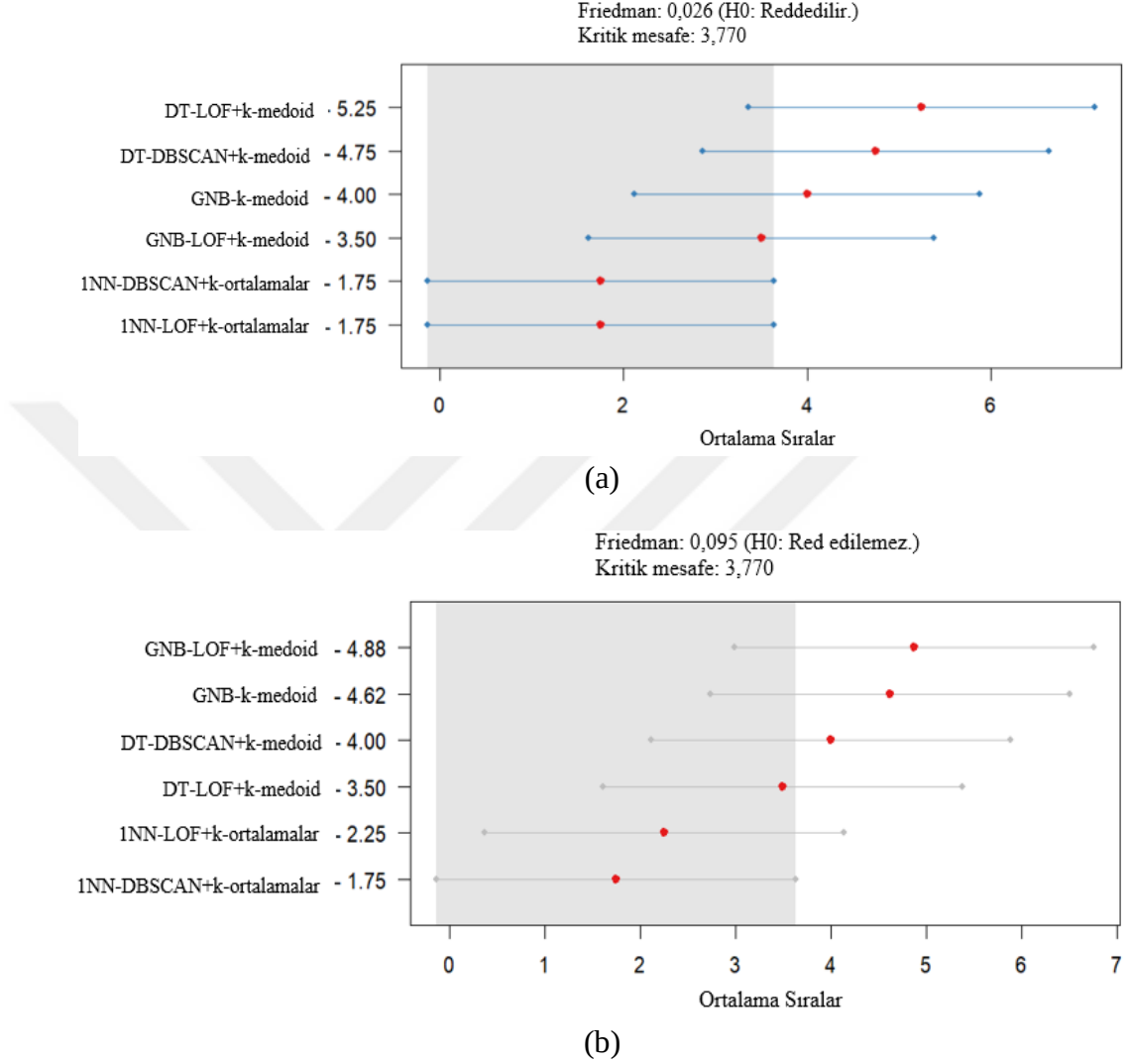


(b)

Şekil 4.3. Friedman ve Nemenyi testleri GNB sınıflandırma ile başlangıç sorgu örneği seçimi karşılaştırma grafikleri

Her bir sınıflandırma yöntemi için en iyi iki yöntem seçilmiştir ve toplam altı yöntem için Friedman ve Nemenyi testleri tekrarlanmıştır. Buna göre, KNN sınıflandırıcılarda 1NN-LOF+*k*-ortalamalar ve 1NN-DBSCAN+*k*-ortalamalar, DT’de DT-DBSCAN+*k*-medoid ve DT-LOF+*k*-medoid, GNB’de GNB-LOF+*k*-medoid ve GNB-*k*-medoid karşılaştırma

için belirlenmiştir. En iyi altı yöntem için yapılan Friedman ve Nemenyi testleri doğruluk (a) ve ağırlıklı F-ölçütü (b) grafikleri Şekil 4.4'te verilmiştir.



Şekil 4.4. Friedman ve Nemenyi testleri en iyi altı yöntem için başlangıç sorgu örneği seçimi karşılaştırma grafikleri

Şekil 4.4'ten aşağıdaki çıkarımlar yapılmaktadır:

- Doğruluk grafiğinde, %95 güven aralığında bazı yöntemler için kritik uzaklıkların çakıştığı görülmektedir. Buna göre yöntemler arasında istatistiksel olarak anlamlı fark olmadığı görülmektedir. Friedman testi sonucunda elde edilen p-değerine bakıldığında, $p = 0,026 < 0,05$ olduğu için H_0 hipotezi reddedilir ve yöntemler arasında istatistiksel olarak anlamlı fark olduğu sonucuna ulaşılmaktadır.

Ortalama sıralamalara bakıldığında en iyi iki yöntem sırasıyla 1NN-LOF+k-ortalamar ve 1NN-DBSCAN+k-ortalamar olarak belirlenmiştir.

- Ağırlıklı F-ölçütü grafiğinde, %95 güven aralığında bazı yöntemler için kritik uzaklıkların çakıştığı görülmektedir. Güven aralıklarına göre yöntemler arasında istatistiksel olarak anlamlı fark olmadığı görülmektedir. Friedman testi sonucunda elde edilen p-değerine bakıldığında, $p = 0,095 > 0,05$ olduğu için H_0 hipotezi red edilemez ve yöntemler arasında istatistiksel olarak anlamlı fark olmadığı sonucuna ulaşılmaktadır. Ortalama sıralamalara bakıldığında en iyi iki yöntem sırasıyla 1NN-LOF+k-ortalamar ve 1NN-DBSCAN+k-ortalamar olarak belirlenmiştir.

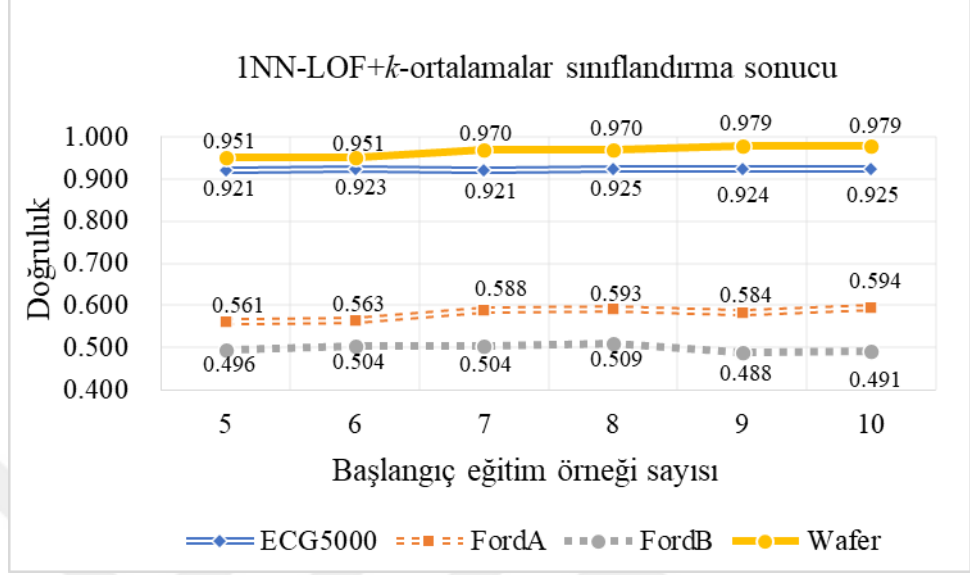
Analizler sonucunda 1NN-LOF+k-ortalamar ve 1NN-DBSCAN+k-ortalamar yaklaşımlarının diğer yöntemlere göre iyi sonuç verdiği gözlemlenmiştir.

Şekil 4.5'te en iyi yöntemlerden biri olan 1NN-LOF+k-ortalamar yöntemi için başlangıçta seçilen eğitim örneği sayısına göre doğruluk (a) ve ağırlıklı F-ölçütü (b) değerleri sunulmuştur.

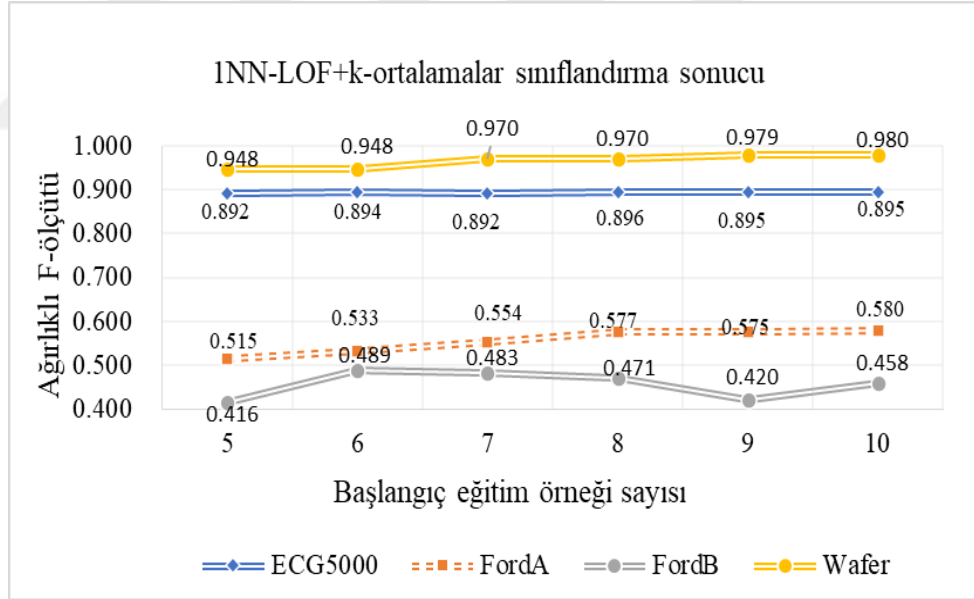
Şekil 4.5'ten aşağıdaki çıkarımlar yapılmaktadır:

- Doğruluk değeri için; Wafer veri kümesi için başlangıçta seçilen eğitim örneği sayısı arttıkça başlangıç sorgu seçim doğruluğu artmaktadır. ECG5000 veri kümesinde değerlerin, eğitim örneği sayısı arttıkça iyileştiği görülmüştür. FordA ve FordB veri kümesine bakıldığında diğer veri kümelerine göre düşük doğruluk değerleri elde edilmiştir. Eğitim örneği sayısı arttıkça doğruluk FordA için iyileşmiş olsa da, FordB için düşüş göstermiştir.
- Ağırlıklı F-ölçütü değeri için; Wafer veri kümesinde eğitim örneği sayısı arttıkça ağırlıklı F-ölçütü değerinde artış görülmüştür. ECG5000 veri kümesinde eğitim örneği sayısının artmasıyla iyileşme görülse de yüksek artış elde edilmemiştir. FordA veri kümesinde eğitim örneği sayısı arttıkça, ağırlıklı F-ölçütü değeri

artmaktadır. FordB’de en iyi deęer $nc = 6$ ’da elde edilmiřken, eęitim rneęi sayısı arttıķa $nc = 9$ ’a kadar dřř gzlemlenmiřtir.



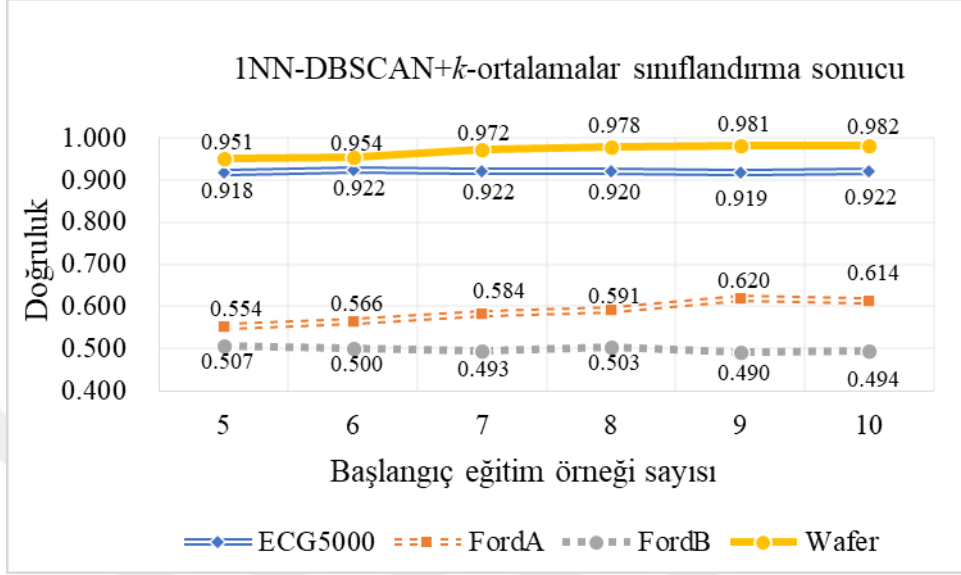
(a)



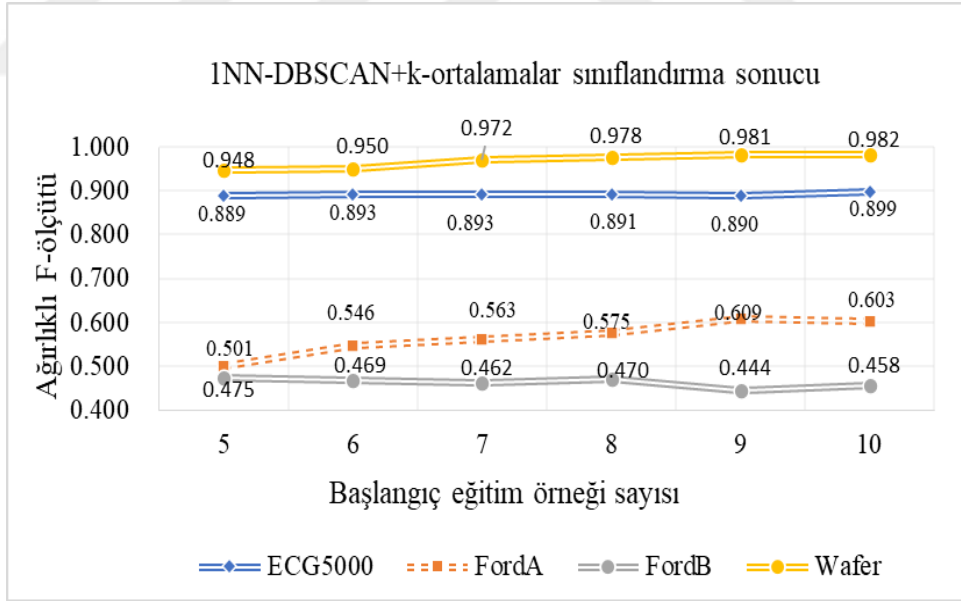
(b)

řekil 4.5. 1NN-LOF+k-ortalamlar yöntemi için başlangıç sorgu örneęi seçimi doęruluk ve ağırlıklı F-ölçütü deęerleri

Şekil 4.6’da en iyi yöntemlerden 1NN-DBSCAN+k-ortalamlar yöntemi için başlangıçta seçilen eğitim örneği sayısına göre ortalama doğruluk (a) ve ağırlıklı F-ölçütü (b) değerleri sunulmuştur.



(a)



(b)

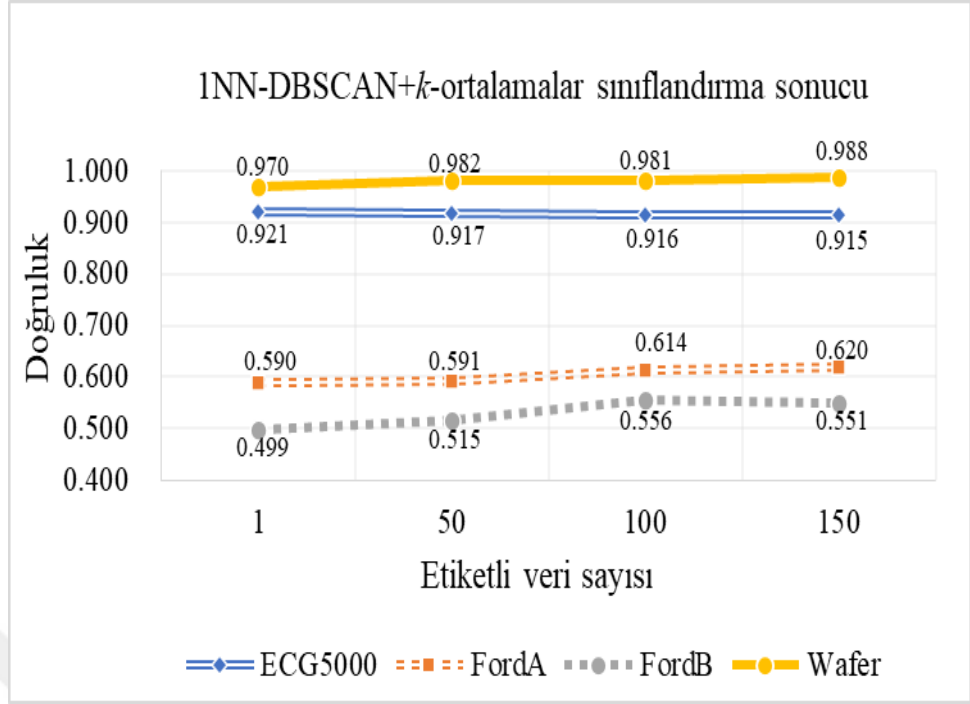
Şekil 4.6. 1NN-DBSCAN+k-ortalamlar yöntemi için başlangıç sorgu örneği seçimi doğruluk ve ağırlıklı F-ölçütü değerleri

Şekil 4.6’dan aşağıdaki çıkarımlar yapılmaktadır:

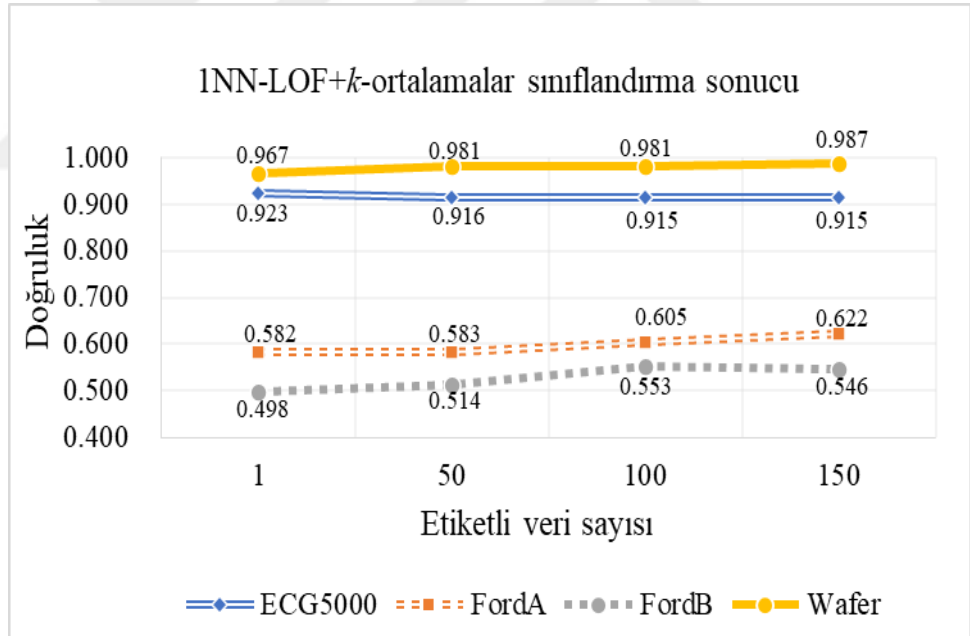
- Doğruluk değeri için; Wafer veri kümesi için başlangıçta seçilen eğitim örneği sayısı arttıkça başlangıç sorgu seçim doğruluğu artmaktadır. ECG5000 veri kümesi içinde değerlerin eğitim örneği sayısı arttıkça, düşük oranda da olsa performansın iyileştiği görülmüştür. Eğitim örneği sayısı arttıkça FordA için iyileşme olmuştur ancak, FordB için düşüş görülmektedir.
- Ağırlıklı F-ölçütü değeri için; Wafer veri kümesinde eğitim örneği sayısı arttıkça ağırlıklı F-ölçütü değerinde artış görülmüştür. ECG5000 veri kümesinde performans stabil kalmıştır. FordA veri kümesinde eğitim örnek sayısı arttıkça, ağırlıklı F-ölçütü değeri artarken, FordB’de azaldığı görülmüştür.

Şekil 4.7’de sorgu örneği etiketleme sürecinde 1NN-DBSCAN+k-ortalamalar ve 1NN-LOF+k-ortalamalar yöntemleri için doğruluk değerleri gösterilmektedir. Wafer veri kümesinde, etiketlenen sorgu örneği sayısı arttıkça her iki yöntem için performansta artış gözlemlenmiştir. ECG5000 veri kümesinde, etiketlenen veri sayısı arttıkça doğruluk oranlarında düşüş görülmüştür. FordA veri kümesinde etiketlenen veri sayısı arttıkça artış gözlemlenmiştir. FordB veri kümesinde her iki yöntem için etiketli veri sayısı arttıkça doğruluk değerlerinde 100. eğitim örneğinin etiketlenmesine kadar artış görülmüştür.

Şekil 4.8’de sorgu örneği etiketleme sürecinde 1NN-DBSCAN+k-ortalamalar ve 1NN-LOF+k-ortalamalar yöntemleri için ağırlıklı F-ölçütü değerleri görülmektedir. Wafer veri kümesinde, sorgu örneği sayısı arttıkça her iki yöntem içinde artış gözlemlenmiştir. ECG5000 veri kümesinde, etiketlenen veri sayısı arttıkça ağırlıklı F-ölçütünde düşüş görülmüştür. FordA veri kümesinde etiketlenen sorgu örneği sayısı arttıkça artış gözlemlenmiştir. FordB veri kümesinde her iki yöntem içinde etiketli veri sayısı arttıkça doğruluk değerlerinde 100. eğitim örneğinin etiketlenmesine kadar artış görülmüştür.

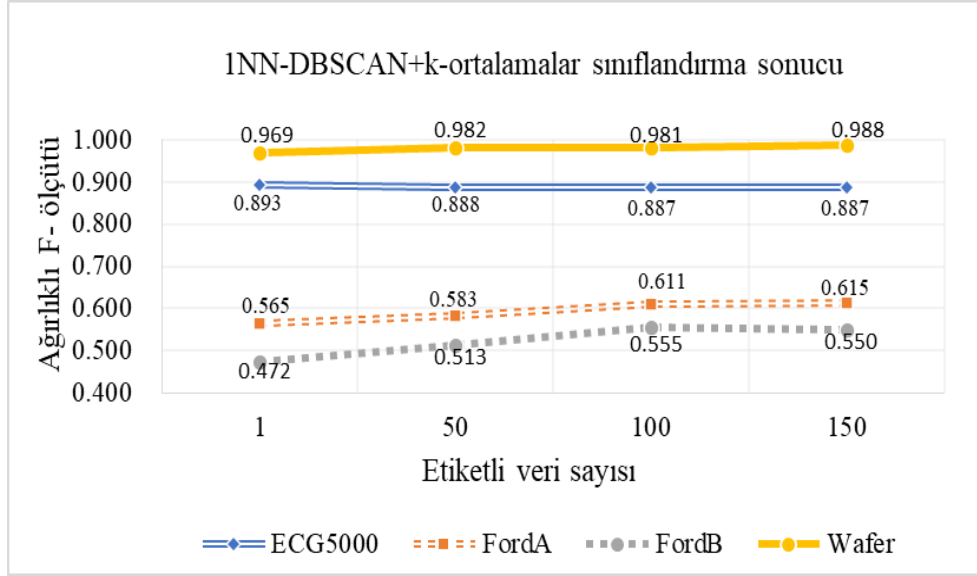


(a)

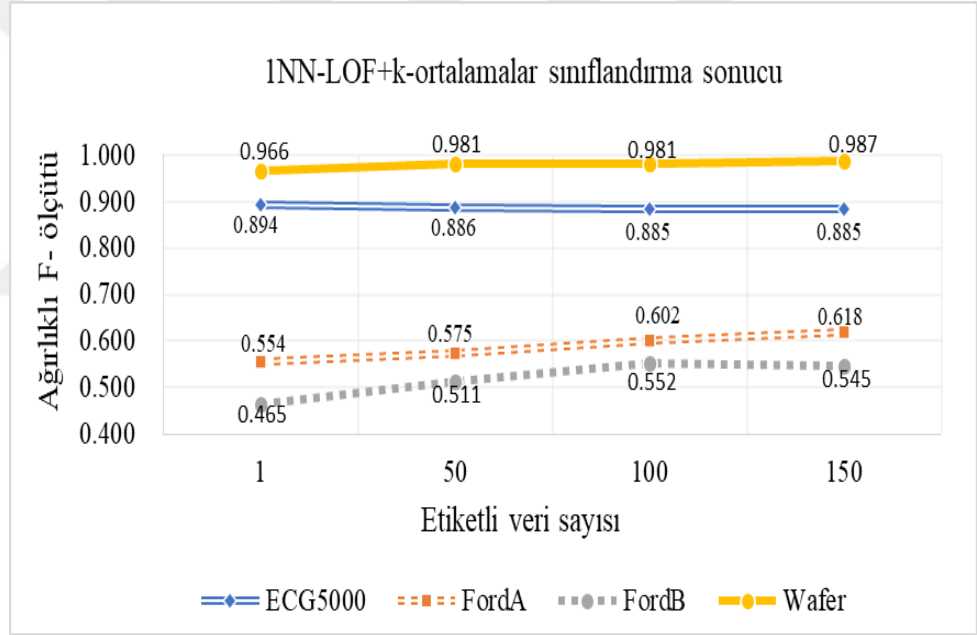


(b)

Şekil 4.7. Önerilen yaklaşımlar ile etiketlenen veriler için doğruluk sonuçları



(a)



(b)

Şekil 4.8. Önerilen yaklaşımlar ile etiketlenen veriler için ağırlıklı F-ölçütü sonuçları

Aktif öğrenmenin performansını değerlendirmek için tüm etiketli veri ile eğitilen sınıflandırma modeli ve 150 adet etiketli veri ile eğitilen model karşılaştırılmıştır. Bu kapsamda, ele alınan dört veri kümesi 1NN sınıflandırma algoritması ile eğitilmiştir. Buna göre, sonuçlar şöyledir:

- Wafer veri kümesinde tüm etiketli veri ile eğitilen modelde hem doğruluk hem de ağırlıklı F-ölçütü değerleri 0,996 olarak elde edilmiştir. 150 adet etiketli veri ile eğitilen 1NN-DBSCAN+k-ortalamalar modelinde ise doğruluk ve ağırlıklı F-ölçütü değerleri sırasıyla 0,988 ve 0,987'dir. Buna göre, aktif öğrenme ve denetimli öğrenme performansları arasında doğruluk ölçütüne göre %0,8; ağırlıklı F-ölçütüne göre ise %0,9 fark olduğu görülmektedir.
- ECG5000 veri kümesinde tüm etiketli veri ile eğitilen modelde doğruluk ve ağırlıklı F-ölçütü değerleri sırasıyla 0,927 ve 0,924'tür. 150 adet etiketli veri ile eğitilen 1NN-DBSCAN+k-ortalamalar modelinde ise doğruluk ve ağırlıklı F-ölçütü değerleri sırasıyla 0,915 ve 0,887'dir. Bu durumda, denetimli öğrenme modeline ait performansın aktif öğrenme modeline ait performanstan doğruluk ölçütüne göre %1,3; ağırlıklı F-ölçütüne göre %4,0 oranında daha iyi olduğu gözlenmiştir.
- FordA veri kümesinde tüm etiketli veri ile eğitilen modelde doğruluk ve ağırlıklı F-ölçütü değerleri sırasıyla 0,670 ve 0,669'dur. 150 adet etiketli veri ile eğitilen 1NN-LOF+k-ortalamalar modelinde ise doğruluk ve ağırlıklı F-ölçütü değerleri sırasıyla 0,622 ve 0,618'dir. Aktif öğrenme ile denetimli öğrenme modellerinin performansları arasında doğruluk ölçütüne göre %7,2; ağırlıklı F-ölçütüne göre %7,7 fark vardır.
- FordB veri kümesinde tüm etiketli veri ile eğitilen modelde hem doğruluk hem de ağırlıklı F-ölçütü değerleri 0,610'dur. 150 adet etiketli veri ile eğitilen 1NN-DBSCAN+k-ortalamalar modelinde ise doğruluk ve ağırlıklı F-ölçütü değerleri sırasıyla 0,551 ve 0,550'dir. Buna göre, denetimli öğrenme modeli aktif öğrenme modeline kıyasla doğruluk ölçütünde %9,7; ağırlıklı F-ölçütünde ise %9,8 oranında daha iyi performans göstermektedir.

Sonuç olarak, aktif öğrenme modelleri, denetimli öğrenme modelleri ile karşılaştırıldığında Wafer ve ECG5000 veri kümelerinde başarılı sonuçlar vermiştir.

Diğer yandan, FordA ve FordB veri kümelerinde aktif öğrenme modellerinin performansında iyileştirme potansiyeli bulunmaktadır.

150 etiketli veri ile eğitilen 40 yöntem, her biri veri kümesi için eğitim sürelerine göre küçükten büyüğe sıralanmıştır ve en düşük eğitim süresi olan yöntemin sıralaması 1, en büyük eğitim süresine sahip yöntemin sıralaması 40'tır. Tüm yöntemlerin eğitim süreleri ve sıralama değerleri EK 7'de verilmiştir. Önerilen yaklaşımlar için eğitim süreleri ve sıralama değerleri ise Çizelge 4.4'te özetlenmiştir. Buna göre, 1NN-LOF+k-ortalamlar yönteminin eğitim süresi, 1NN-DBSCAN+k-ortalamlar yönteminin eğitim süresinden daha uzundur.

Çizelge 4.4. Önerilen yaklaşımların 150 etiketli veri ile eğitim süreleri ve sıralama değerleri

Veri kümesi	1NN-DBSCAN+k-ortalamlar		1NN-LOF+k-ortalamlar	
	Eğitim Süresi (sn.)	Sıralama	Eğitim Süresi (sn.)	Sıralama
Wafer	70	17	108	25
ECG5000	60	12	92	22
FordA	79	16	125	25
FordB	90	19	112	24

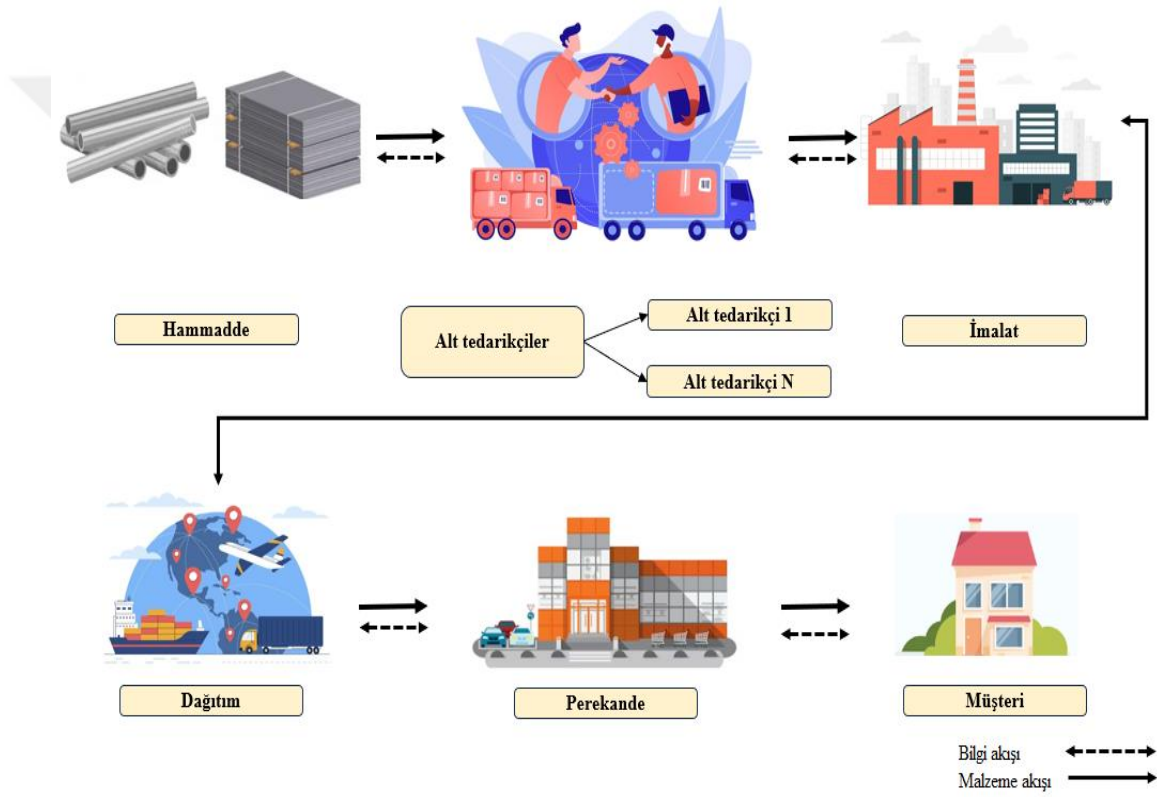
4.3. Otomobil Sektöründe Tedarikçi Risk Durumu için Vaka Analizi

Hammadde ya da yarımamüllerin tedarikçilerden alınıp, üreticiye oradan da son tüketici olan müşteriye sevkine kadar sıralı şekilde birbirine bağlı bu sürece tedarik zinciri denir (Zhang vd., 2023). Günümüz tedarik zinciri ağı sayesinde üretici firmalar aynı ülkedeki tedarikçilerin yanı sıra denizaşırı ülkeler, Avrupa ve Asya ülkeleri gibi dünyanın her yerindeki tedarikçilerden hammadde ya da yarı mamül temin edebilmektedir. Üretici firma, temin ettiği yarımamulleri son ürüne dönüştürür. Ürünler bayilere sevk edilir ve bayiler aracılığıyla son müşteriye teslim edilir. Tedarik zinciri akışı Şekil 4.9'da gösterilmektedir.

Üretici firmalar, gelişen küresel tedarik zinciri ağıyla dünyanın her yerinden hammadde ya da bitmiş ürün temin edip, dünyanın her yerine hammadde ve bitmiş ürünleri gönderebilir

hale gelmiştir. Bu noktada firmalar hem dış hem de kendi iç süreçlerindeki tedarik zinciri akışını yönetmektedir.

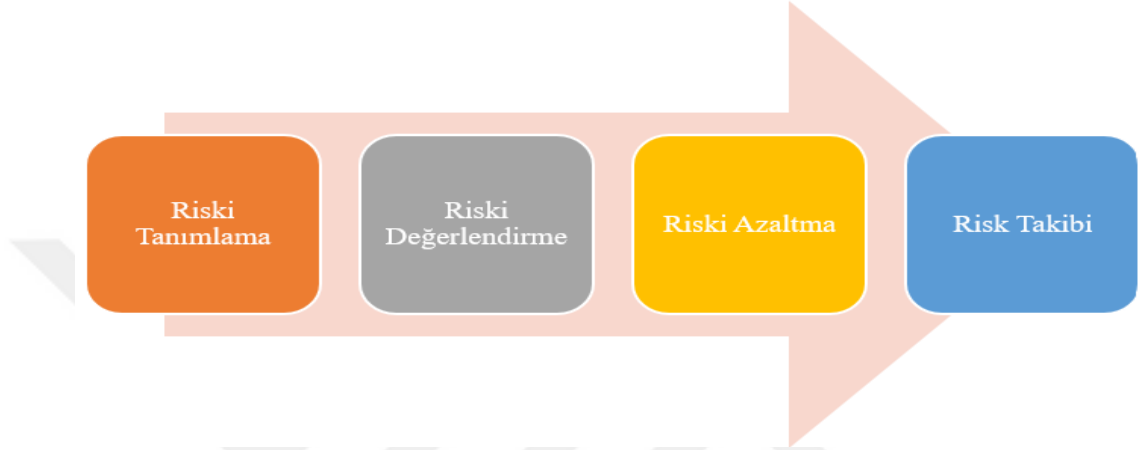
Tedarik zinciri taşıma/dağıtım, teslim süresi, üretim süreci, sipariş döngüsü ve doğruluğu, depolama gibi faktörlerden direkt olarak etkilenmektedir ve her bir süreç potansiyel risk barındırmaktadır (Gaudenzi & Borghesi, 2006). Tedarik zincirinin her aşamasında karmaşık ve sürekli değişen bir yapı mevcuttur. Bu değişen talep ve süreçlere ayak uydurabilmek için firmalar çeşitli çözüm yolları geliştirmektedir.



Şekil 4.9. Tedarik zinciri akışı

Firmalar öngörülebilir riskleri (örneğin kapasite problemi, grev vb.) tahmin edip onlara yönelik proaktif aksiyon alabilmektedir, ancak öngörülemez riskler birçok problemi de beraberinde getirmektedir. Depremler, savaşlar, ekonomik krizler, Covid 19 salgını gibi bir çok öngörülemez risk tedarik zinciri akışını sektöre uğratmaktadır. Bu ve bunun gibi belirsizlikler tedarik zinciri yönetimini olumsuz etkilemektedir. Bir firmanın amacı doğru yerde, doğru zamanda müşteri taleplerini karşılamaktır. Bu sebeple firmalar

tedarik zincirindeki risklere zamanında aksiyon alabilmek için tedarik zinciri risk yönetimine büyük önem vermektedir. Tedarik Zinciri Risk Yönetimi (Supply Chain Risk Management – SCRM) aşamaları Şekil 4.10’da gösterilmektedir. SCRM’nin aşamaları sırasıyla riski tanımlamak, riski değerlendirmek, riski azaltmak ve riskin takibidir (Yang vd., 2023).



Şekil 4.10. Tedarik zinciri risk yönetiminin adımları

Tedarik zincirinde, belirsiz talepler ve tedarik süreci risk oluşturabileceği gibi doğal afetler, ekonomik krizlerde risk oluşturmaktadır. Birçok belirsizliğin, risklerin ve krizlerin yaşandığı bu süreçte üretici firmalar üretimlerini sürdürebilmek için parça temininde risk olup olmadığını önceden öngörebilmeyi hedeflemektedir. Bu vaka analizi, risk tanımlama ve risk değerlendirme aşamalarına odaklanmaktadır.

4.3.1. Tedarik Zinciri Ortamı

Bu bölümde, önerilen yaklaşımlar bir otomobil üreticisinde tedarik zinciri risk yönetimi için uygulanmıştır. Firma, son müşteriler için bitmiş otomobiller üretmektedir. Otomobil üretmek için gerekli hammadde farklı ülkelerde ve kendi ülkesinde bulunan çeşitli tedarikçilerden temin edilmektedir. Üretim kapasitesi, teslim süresi ve nakliye şekli tedarikçiye ve lokasyona göre değişmektedir. Tedarik zincirinin her noktasında belirsizlikler ve risklerle karşılaşmaktadır. Bu nedenle tedarik zincirinde riskin değerlendirilmesi önemlidir.

değiştirilmemesi için parça temin süreçlerindeki riskleri önceden tahmin etmeyi amaçlamaktadır.

Vaka analizinde üretici firmadan elde edilen zaman serileri kullanılmıştır. Mevcut sistemde temin uzmanları tedarikçinin üretim kapasitesini ve bakiye durumunu üreticinin üretim planına göre analiz etmektedir. Ayrıca tedarikçinin geçmiş siparişlerini, stok, bakiyelerini kontrol etmek için ek süre gerekir. Uzmanlar bu girdileri göz önünde bulundurarak her parça siparişindeki riski tahmin ederek etiketleme işlemini gerçekleştirmektedir. Bu süreçleri yönetmek için 20 uzman çalışmaktadır ve her uzman en az 100 parçadan sorumludur. Bu prosedür her hafta her parça siparişi için tekrarlanmaktadır. Bu nedenle, etiketleme görevi üretici için maliyetlidir. Bu vaka çalışması, her bir parça siparişinin riskini en az etiketleme maliyeti ve yüksek doğrulukla belirlemeyi amaçlamaktadır.

4.3.2. Veri Kümesi ve Önışleme

Vaka analizinde amaç, üç haftalık planlama dönemi içerisinde bir parça siparişinin tesliminde herhangi bir risk olup olmadığını tahmin etmektir. Bu nedenle problem sınıflandırma problemi olarak formüle edilmiş ve otomobil üreticisinden sekiz haftalık süre boyunca veriler toplanmıştır.

Çıktı (sınıf etiketi), belirli bir zaman diliminde bir parça siparişinin risk durumu olarak tanımlanmaktadır ve ikili bir değişkendir. Temin sürecinde parça siparişinin 1 değeri yüksek risk taşıdığını, 0 değeri ise riskin düşük olduğunu göstermektedir. Uzman görüşleriyle bir parça siparişinin risk durumunu etkileyen öznitelikler belirlenmiştir. Çizelge 4.5'te bu çalışmada kullanılan öznitelikler ve veri türleri gösterilmektedir. Bu öznitelikler zamanla ilgili, tedarikçiyle ilgili ve üreticiyle ilgili öznitelikler olmak üzere üç kategoriye ayrılmaktadır.

Zamanla ilgili öznitelikler, ilgili yılın hafta ve ay bilgisini göstermektedir.

Tedarikçiyle ilgili öznitelikler, tedarikçinin faaliyet gösterdiği ülkeyi, haftalık üretim kapasitesini, teslimat süresinde sorun olup olmadığını, siparişleri son dört haftadaki karşılama performansını ve ekstra nakliye maliyetinin varlığını göstermektedir. Her ne kadar her tedarikçi için her parçanın nakliye şekli belirlenmiş olsa da siparişin aciliyetine göre farklı bir nakliye şekli kullanılabilir. Bu durumda ekstra nakliye maliyeti ortaya çıkmakta olup, ekstra nakliye maliyetinin varlığı siparişin teslim riskini etkileyebilmektedir. Örneğin bir parça için önceden tanımlanmış nakliye aracı tırdır. Kritiklik durumuna göre parça uçak kargo temin edilirse, ekstra nakliye maliyetleri ortaya çıkmaktadır.

Üreticiyle ilgili öznitelikler, parça kategorisi, son üç haftadaki fiili üretim miktarları ve son üç hafta ile bulunan hafta dahil olmak üzere gelecek üç haftadaki öngörülen üretim miktarları yer almaktadır.

Çizelge 4.5. Öznitelikler

Öznitelik Grubu	Öznitelik Adı	Açıklama	Veri Tipi
Zaman	Hafta	Yılın haftası	Sayısal
	Ay	Yılın ayı	Sayısal
Tedarikçi	Ülke	Tedarikçi konumu	Nominal
	Kapasite	Haftalık üretim kapasitesi	Sayısal
	Sevk süresi	Teslimat süresinde sorunun varlığı	İkili
	Bakiye durumu	Sipariş tamamlama durumu	Sayısal
	Sevk durumu	Akış dışı sevkiyat durumu	İkili
Firma	Parça sınıfı	Ürün grubu	Nominal
	Üretim adedi	Haftalık gerçekleşen üretim adedi	Sayısal
	Tahmini üretim	Üç haftalık tahmini üretim adedi	Sayısal

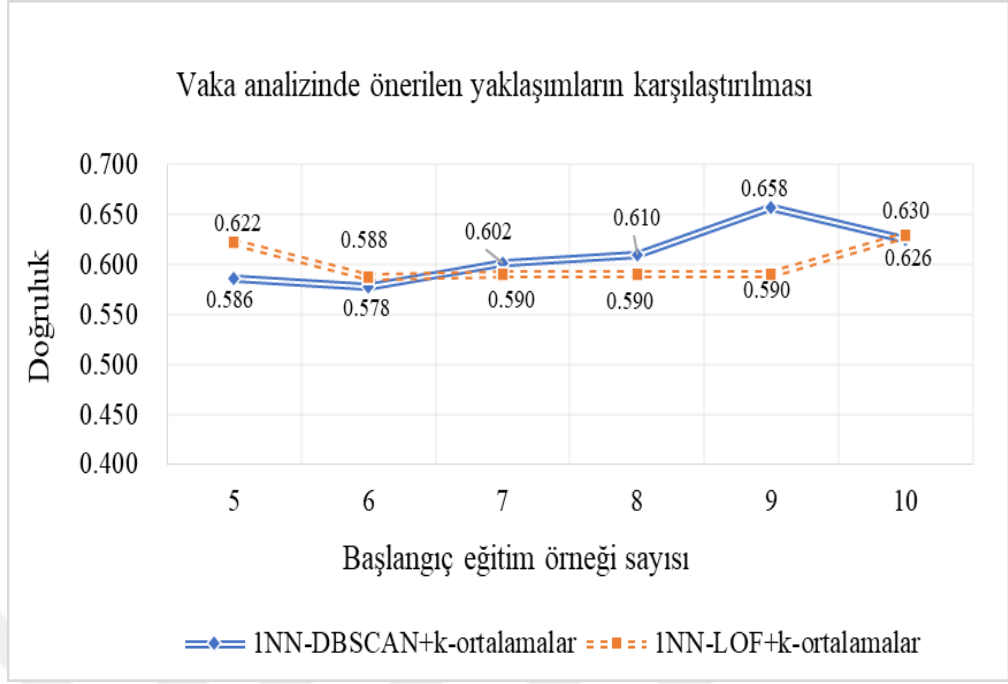
Firmadan elde edilen veri kümesi, haftalık olarak düzenli periyotlar ile hazırlanmıştır ve bu veri kümesi zaman serisi özelliklerini barındırmaktadır. Veri kümesinde 21 nitelik ve 400 örnek bulunmaktadır. Veri ön işleme aşamasında, nominal niteliklerin sayısal niteliklere dönüştürülmesi için 1-0 ikili kodlama uygulanmıştır. Ayrıca öznitelikleri standartlaştırmak için z-skor normalizasyon uygulanmıştır. Veri kümesinde eksik veri bulunmamaktadır. Veri kümesi dengesiz sınıf dağılımına sahiptir.

4.3.3. Vaka Analizine İlişkin Bulgular

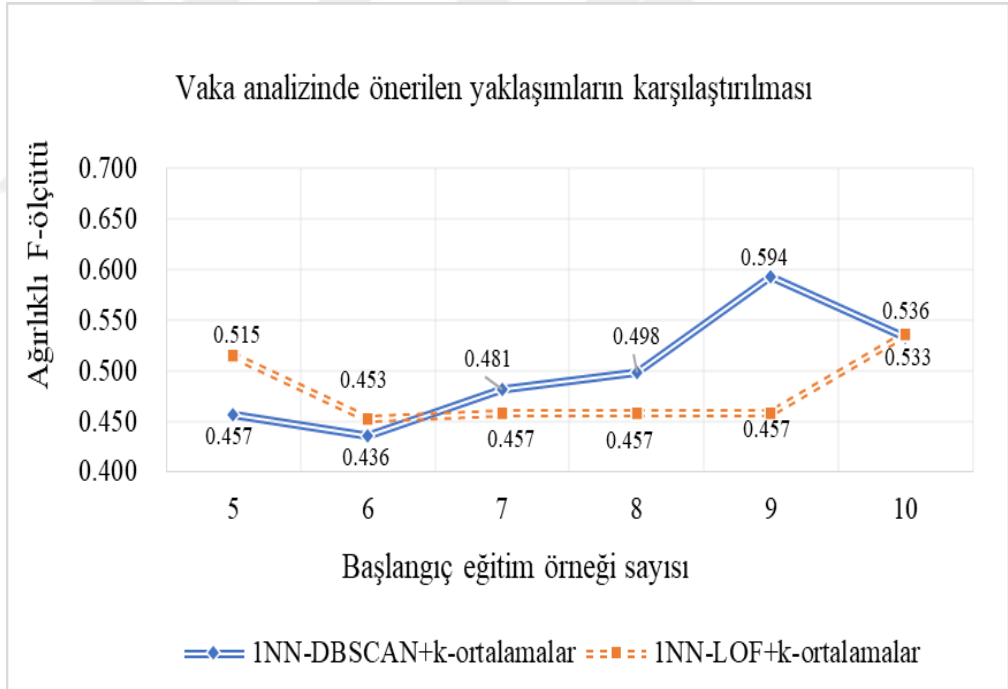
Firmadan elde edilen veri kümesine önerilen yaklaşımlar uygulanmıştır. Başlangıç eğitim örneği seçimine göre önerilen yöntemler içinde elde edilen doğruluk (a) ve ağırlıklı F-ölçütü (b) sonuçları Şekil 4.12’de sunulmuştur.

Şekil 4.12’den aşağıdaki çıkarımlar yapılmaktadır:

- Doğruluk grafiğinde en iyi sonuç $nc = 9$ için 0,658 değeriyle 1NN-DBSCAN+k-ortalamlar yönteminde elde edilmiştir. Eğitim örneği sayısı arttıkça doğruluk değerinin arttığı gözlemlenmiştir. 1NN-LOF+k-ortalamlar yönteminde en yüksek değer 0,626 ile $nc = 10$ ’da elde edilmiştir. 1NN-LOF+k-ortalamlar yönteminde başlangıç eğitim örneği sayısı arttıkça doğruluk değeri artmıştır. 1NN-DBSCAN+k-ortalamlar yönteminde $nc = (5,...,9)$ aralığında artış gözlemlenmiştir.
- Ağırlıklı F-ölçütü grafiğinde; en iyi sonuç $nc = 9$ için 0,594 değeriyle 1NN-DBSCAN+k-ortalamlar yönteminde elde edilmiştir. $nc = (5,...,9)$ aralığında eğitim sayısı arttıkça performans artmıştır ancak $nc = 10$ için ağırlıklı F-ölçütü değerinin 0,533’e düştüğü gözlemlenmiştir. 1NN-LOF+k-ortalamlar yönteminde en yüksek değer 0,536 ile $nc = 10$ ’da elde edilmiştir ve seçilen eğitim örneği sayısında artış oldukça, ağırlıklı F-ölçütü değerinin de arttığı gözlemlenmiştir.



(a)



(b)

Şekil 4.12. Vaka analizinde önerilen yaklaşımlar için başlangıç eğitim örneği seçimi doğruluk ve ağırlıklı F-ölçütü sonucu

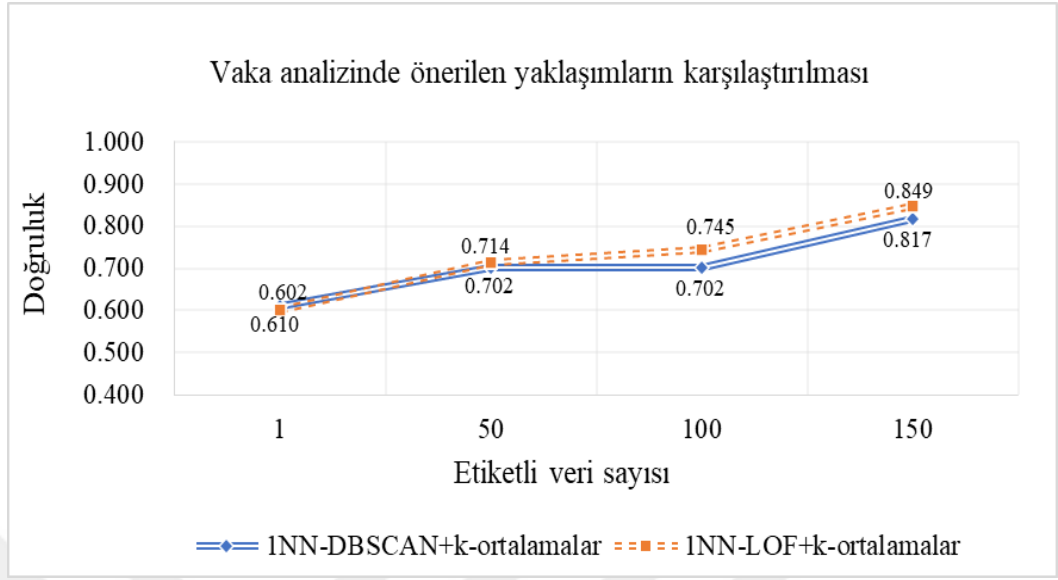
Sorgu örneği etiketlenme durumuna göre her iki yöntem içinde elde edilen doğruluk (a) ve ağırlıklı F-ölçütü (b) sonuçları Şekil 4.13'te sunulmuştur.

Şekil 4.13'ten aşağıdaki çıkarımlar yapılmaktadır:

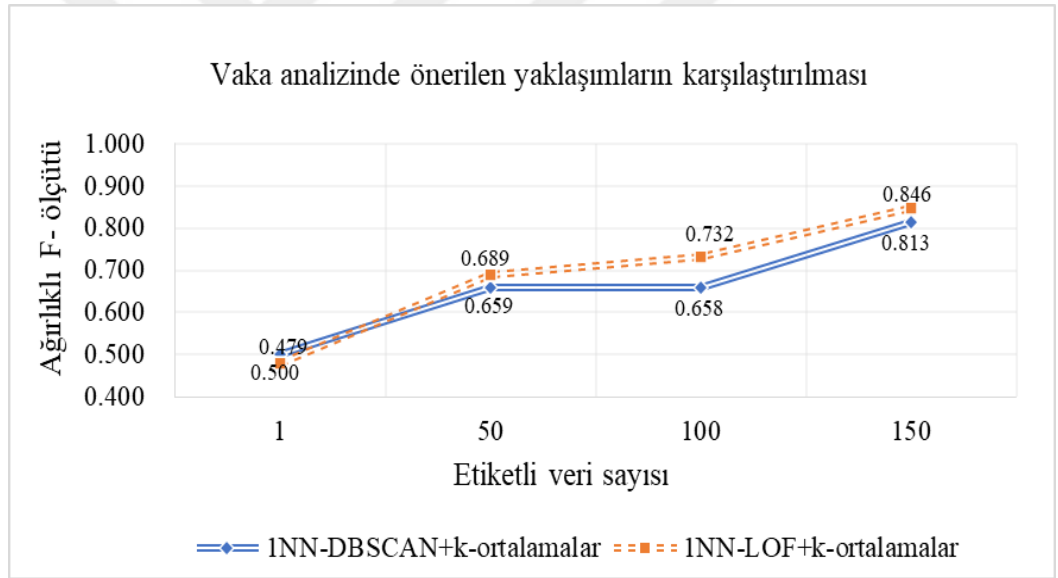
- Doğruluk grafiğinde; 1NN-DBSCAN+k-ortalamalar yönteminde, 1. ve 50. sorgu örnekleri arasında artış gözlemlenmiştir. 50. ve 100. sorgu örnekleri arasında değişim gözlemlenmemiştir. 150. sorguya erişildiğinde doğruluk oranında artış gözlemlenmiş ve iyi sonuç elde edilmiştir. 1NN-LOF+k-ortalamalar yönteminde sorgu örneği sayısı arttıkça, doğruluk değerinde artış gözlemlenmiştir.
- Ağırlıklı F-ölçütü grafiğinde; 1NN-DBSCAN+k-ortalamalar yönteminde, 1. ve 50. sorgu örnekleri arasında artış gözlemlenmiştir. 50. ve 100. sorgu örnekleri arasında az da olsa düşüş gözlemlenmemiştir. 150. sorguya erişildiğinde ağırlıklı F-ölçütünde artış elde edilmiştir. 1NN-LOF+k-ortalamalar yönteminde sorgu örneği sayısı arttıkça, değerinde artış gözlemlenmiştir.

Vaka analizinde aktif öğrenmenin performansını değerlendirmek için tüm etiketli veri ile eğitilen sınıflandırma modeli ve 150 adet etiketli veri ile eğitilen model karşılaştırılmıştır. Buna göre, tüm etiketli veri ile eğitilen modelde hem doğruluk hem de ağırlıklı F-ölçütü değerleri 0,880 olarak elde edilmiştir. 150 adet etiketli veri ile eğitilen 1NN-LOF+k-ortalamalar modelinde ise doğruluk ve ağırlıklı F-ölçütü değerleri sırasıyla 0,849 ve 0,846'dır. Bu durum, aktif öğrenme ve denetimli öğrenme performansları arasında doğruluk ölçütüne göre %3,5; ağırlıklı F-ölçütüne göre %3,8 oranında fark olduğu görülmektedir. Bu nedenle, vaka analizinde aktif öğrenme modeli ile başarılı sonuç elde edilmiştir.

Vaka analizinde 150 etiketli veri ile eğitilen 40 yöntem, eğitim sürelerine göre küçükten büyüğe sıralanmıştır ve en düşük eğitim süresi olan yöntemin sıralaması 1, en büyük eğitim süresine sahip olan yöntemin sıralaması 40'tır. Buna göre, 1NN-DBSCAN+k-ortalamalar yönteminin eğitim süresi 64 saniye ile 18. sırada, 1NN-LOF+k-ortalamalar yönteminin eğitim süresi 86 saniye ile 24. sırada yer almıştır. LOF yönteminin DBSCAN yöntemine göre daha uzun çalıştığı gözlemlenmiştir.



(a)



(b)

Şekil 4.13. Vaka analizinde önerilen yaklaşımlar ile etiketlenen veriler için doğruluk ve ağırlıklı F-ölçütü sonuçları

5. SONUÇ

Bu çalışmada üretim ve hizmet sektörlerinde zaman serisi sınıflandırma problemi çalışılmıştır. Artan teknolojik faaliyetler sonucunda, günümüzde verileri depolamak kolaylaşmıştır. Rekabete ortak olmak ve verilerden elde edilen bilgiyi yönetmek önemlidir. Depolanan verilerde çoğunlukla etiket bilgisi olmamaktadır. Etiketleme süreci, zamana ve işgücüne ihtiyaç duyması açısından zor ve maliyetlidir. Bu kapsamda, etiketleme maliyetinin azaltmak ve model performansını geliştirmek için aktif öğrenme yaklaşımı önerilmiştir.

Önerilen yaklaşımlarda, tüm veriler etiketsizdir. Verilere Z-skor normalizasyonu işlemi uygulanmıştır. Başlangıçta etiketlenecek verilerin seçimi, modelin iyi genelleme yeteneği sunması için önemlidir. Bu çalışmada, gürültülü ve aykırı değer tabanlı iki farklı yaklaşım önerilmiştir. Başlangıç örnek seçimi sırasında, ilk yaklaşımda DBSCAN yöntemiyle etiketsiz veri kümesindeki gürültülü veriler belirlenip, veri setinden çıkartılmıştır. Ardından k -ortalamlar yöntemiyle başlangıç örnek seçim işlemi gerçekleştirilmiştir. Diğer yaklaşımda, LOF yöntemiyle etiketsiz veri kümesindeki aykırı veriler belirlenip, veri setinden çıkartılmıştır. Ardından k -ortalamlar yöntemiyle başlangıç örnek seçim işlemi gerçekleştirilmiştir. Seçilen örnekler, 1NN algoritması ile eğitilmiştir. Belirlenen sayıda sorgu örneğinin seçimi belirsizlik örnekleme yöntemiyle yapılmıştır.

Üretim ve sağlık sistemleri ile yapılan deneysel çalışmalar, önerilen yöntemin, küçük komşuluk boyutlarında doğruluğu arttırdığını göstermiştir. Küçük komşuluk değerinde eğitim örneği sayısı arttıkça elde edilen sonuçlarda iyileşme görülmüştür. Önerilen yaklaşımların, literatürdeki diğer yöntemler karşılaştırıldığında etiketleme maliyetini azaltmada ve model performansını arttırmada etkili sonuç verdiği görülmüştür.

Öte yandan, gürültülü ve aykırı veriler etiketsiz veri kümesinden çıkarılmasına rağmen önerilen yaklaşımlar gürültü varlığında sınırlamalara sahiptir.

Gelecek çalışmalarda, aktif öğrenmede bilgi kazanımı tabanlı ve çeşitlilik tabanlı sorgu örneği seçim stratejileri değerlendirilebilir. Ayrıca aktif öğrenmenin performansını

artırmak için kümeleme tabanlı sorgu seçim stratejileri geliştirilebilir. Bir diğer gelecek çalışma ise önerilen yaklaşımların yarı denetimli öğrenme yöntemleriyle entegre edilmesidir.



KAYNAKLAR

- Aggarwal, C. C. (2015). *Data mining: the textbook* (C. 1). Springer.
- Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2020). *A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams*. <https://doi.org/10.3390/bdcc>
- Almomany, A., Ayyad, W. R., & Jarrah, A. (2022). Optimized implementation of an improved KNN classification algorithm using Intel FPGA platform: Covid-19 case study. *Journal of King Saud University - Computer and Information Sciences*, 34(6), 3815-3827. <https://doi.org/10.1016/j.jksuci.2022.04.006>
- Bagnall, T. *The UEA & UCR Time Series Classification Repository*. Geliş tarihi 18 Ocak 2024, gönderen <http://www.timeseriesclassification.com>
- Barata, R., Leite, M., Pacheco, R., Sampaio, M. O. P., Ascensão, J. T., & Bizarro, P. (2021). Active learning for imbalanced data under cold start. *Proceedings of the Second ACM International Conference on AI in Finance*, 1-9.
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93-104.
- Buza, K. (2018). Time series classification and its applications. *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, 1-4.
- Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20-28.
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21-27.
- Das, K., & Behera, R. N. (2017). A survey on machine learning: concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(2), 1301-1309.
- Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. İçinde *Journal of Machine Learning Research* (C. 7).
- Elamir, E. A. H. (2022). A Graphical Approach for Friedman Test: Moments Approach. *arXiv preprint arXiv:2202.09131*.
- Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1), 1-34.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & others. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, 96(34), 226-231.
- Flach, P. (2019). Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. *Proceedings of the AAAI conference on artificial intelligence*, 33(01), 9808-9814.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200), 675-701.
- Gan, G., Ma, C., & Wu, J. (2020). *Data clustering: theory, algorithms, and applications*. SIAM.
- Gaonkar, M. N., & Sawant, K. (2013). AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset. *International Journal on Advanced Computer Theory and Engineering*, 2(2), 11-16.
- Gaudenzi, B., & Borghesi, A. (2006). Managing risks in the supply chain using the AHP method. *The International Journal of Logistics Management*, 17(1), 114-136.

- Grimova, N., Macas, M., & Gerla, V. (2018). Addressing the Cold Start Problem in Active Learning Approach Used For Semi-automated Sleep Stages Classification. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2249-2253.
- Gupta, B. B., & Sheng, Q. Z. (2019). *Machine learning for computer and cyber security: principle, algorithms, and practices*. CRC Press.
- Gweon, H., & Yu, H. (2021). A nearest neighbor-based active learning method and its application to time series classification. *Pattern Recognition Letters*, 146, 230-236.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.
- He, G., Duan, Y., Li, Y., Qian, T., He, J., & Jia, X. (2015). Active learning for multivariate time series classification with positive unlabeled data. *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, 178-185.
- He, G., Li, Y., & Zhao, W. (2017). An uncertainty and density based active semi-supervised learning scheme for positive unlabeled multivariate time series classification. *Knowledge-Based Systems*, 124, 80-92.
- Hu, R., Mac Namee, B., & Delany, S. J. (2010). *Off to a good start: Using clustering to select the initial training set in active learning*.
- Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4), 917-963.
- Jahromi, A. H., & Taheri, M. (2017). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. *In 2017 Artificial intelligence and signal processing conference (AISP)*, 209-212.
- Jin, Q., Yuan, M., Li, S., Wang, H., Wang, M., & Song, Z. (2022). Cold-start active learning for image classification. *Information Sciences*, 616, 16-36.
- Kang, J., Ryu, K. R., & Kwon, H.-C. (2004). Using cluster-based sampling to select initial training set for active learning in text classification. *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings 8*, 384-388.
- Koyuncu, F. S., & İnkaya, T. (2023). An Active Learning Approach Using Clustering-Based Initialization for Time Series Classification. *International Symposium on Intelligent Manufacturing and Service Systems*, 224-235.
- Kumar, P., & Gupta, A. (2020). Active learning query strategies for classification, regression, and clustering: a survey. *Journal of Computer Science and Technology*, 35, 913-945.
- Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2), 451-461.
- Liu, Y., & Chen, W. (2012). A SAS macro for testing differences among three or more independent groups using Kruskal-Wallis and Nemenyi tests. *Journal of Huazhong University of Science and Technology [Medical Sciences]*, 32(1), 130-134.
- Ma, J., Xia, D., Wang, Y., Niu, X., Jiang, S., Liu, Z., & Guo, H. (2022). A comprehensive comparison among metaheuristics (MHs) for geohazard modeling using machine learning: Insights from a case study of landslide displacement prediction. *Engineering Applications of Artificial Intelligence*, 114. <https://doi.org/10.1016/j.engappai.2022.105150>
- Mai, S. T., He, X., Hubig, N., Plant, C., & Böhm, C. (2013). Active density-based clustering. *2013 IEEE 13th International Conference on Data Mining*, 508-517.

- Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook* (C. 2, Sayı 2005). Springer.
- Nemenyi, P. B. (1963). *Distribution-free multiple comparisons*. Princeton University.
- Nguyen, H. T., & Smeulders, A. (2004). Active learning using pre-clustering. *Proceedings of the twenty-first international conference on Machine learning*, 79.
- Peng, F., Luo, Q., & Ni, L. M. (2017). ACTS: an active learning method for time series classification. *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, 175-178.
- Pohlert, T. (2014). The pairwise multiple comparison of mean ranks package (PMCMR). *R package*, 27(2019), 9.
- Prudencio, R. B. C., Soares, C., & Ludermir, T. B. (2011). Uncertainty sampling methods for selecting datasets in active meta-learning. *The 2011 International Joint Conference on Neural Networks*, 1082-1089.
- Qi, L., & Ting, L. (2018). Active semi-supervised affinity propagation clustering algorithm based on local outlier factor. *2018 37th Chinese Control Conference (CCC)*, 9368-9373.
- Saeedi, R., Sasani, K., & Gebremedhin, A. H. (2017). Co-MEAL: Cost-optimal multi-expert active learning architecture for mobile health monitoring. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 432-441.
- Šebek, J., Schaabova, H., & Krajca, V. (2019). Active learning approach for EEG classification using neural networks: a review. *2019 E-Health and Bioengineering Conference (EHB)*, 1-4.
- Settles, B. (2009). *Active learning literature survey*.
- Shim, J., Kang, S., & Cho, S. (2021). Active cluster annotation for wafer map pattern classification in semiconductor manufacturing. *Expert Systems with Applications*, 183, 115429.
- Silverman, B. W., & Jones, M. C. (1989). E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *International Statistical Review / Revue Internationale de Statistique*, 57(3), 233-238. <https://doi.org/10.2307/1403796>
- Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, 9(1), 307.
- Su, J., & Zhang, H. (2006). A fast decision tree learning algorithm. *Aai*, 6, 500-505.
- Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, 11(2), 612-619.
- Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K., Kerdprasopb, N., & others. (2015). The clustering validity with silhouette and sum of squared errors. *learning*, 3(7).
- Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *Ieee Access*, 7, 107964-108000.
- Weber, M., Auch, M., Doblander, C., Mandl, P., & Jacobsen, H. A. (2021). Transfer Learning with Time Series Data: A Systematic Mapping Study. *IEEE Access*, 9, 165409-165432. <https://doi.org/10.1109/ACCESS.2021.3134628>
- Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., & others. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14, 1-37.

- Yang, M., Lim, M. K., Qu, Y., Ni, D., & Xiao, Z. (2023). Supply chain risk management with machine learning technology: A literature review and future research directions. *Computers & Industrial Engineering*, 175, 108859.
- Yang, Y., & Loog, M. (2022). To Actively Initialize Active Learning. *Pattern Recognition*, 131, 108836.
- Yang, Y., Qian, C., Li, H., Gao, Y., Wu, J., Liu, C.-J., & Zhao, S. (2022). An efficient DBSCAN optimized by arithmetic optimization algorithm with opposition-based learning. *The Journal of Supercomputing*, 78(18), 19566-19604.
- Zhang, D., Tang, Y., & Yan, X. (2023). Supply chain risk management of badminton supplies company using decision tree model assisted by fuzzy comprehensive evaluation. *Expert Systems*, e13275.
- Zhang, G., & Gionis, A. (2023). Regularized impurity reduction: accurate decision trees with complexity guarantees. *Data mining and knowledge discovery*, 37(1), 434-475.
- Zhang, J., & Dai, Q. (2022). A cost-sensitive active learning algorithm: toward imbalanced time series forecasting. *Neural Computing and Applications*, 34(9), 6953-6972.
- Zhou, P., Zhang, T., Zhao, L., Qi, Y., Chang, Y., & Bai, L. (2023). Pre-clustering active learning method for automatic classification of building structures in urban areas. *Engineering Applications of Artificial Intelligence*, 123, 106382.
- Zhou, Z.-H. (2021). *Machine learning*. Springer Nature.

EKLER

- EK 1** 1NN, 3NN ve 5NN yöntemleri başlangıç sorgu seçimi doğruluk sonuçları
- EK 2** 1NN, 3NN ve 5NN yöntemleri başlangıç sorgu seçimi ağırlıklı F-ölçütü sonuçları
- EK 3** Karar ağacı yöntemi başlangıç sorgu seçimi doğruluk sonuçları
- EK 4** Karar ağacı yöntemi başlangıç sorgu seçimi ağırlıklı F-ölçütü sonuçları
- EK 5** Gaussian Naive Bayes yöntemi başlangıç sorgu seçimi doğruluk sonuçları
- EK 6** Gaussian Naive Bayes yöntemi başlangıç sorgu seçimi ağırlıklı F-ölçütü sonuçları
- EK 7** Eğitim süreleri ve yöntemlerin eğitim sürelerine göre sıralamaları

EK 1 1NN, 3NN ve 5NN Yöntemleri Başlangıç Sorgu Seçimi Doğruluk Sonuçları

	Veri kümeleri			
Yöntemler	ECG5000	FORDA	FORDB	WAFER
1NN-DBSCAN	0,86	0,53	0,51	0,93
1NN-DBSCAN+k-ortalamalar	0,92	0,59	0,50	0,97
1NN-DBSCAN+k-medoid	0,90	0,57	0,47	0,90
1NN-k-ortalamalar	0,92	0,58	0,50	0,97
1NN-k-medoid	0,91	0,56	0,46	0,90
1NN-LOF+k-ortalamalar	0,92	0,58	0,50	0,97
1NN-LOF+k-medoid	0,91	0,56	0,46	0,90
1NN-Rassal	0,85	0,53	0,51	0,90
3NN-DBSCAN	0,76	0,50	0,51	0,90
3NN-DBSCAN+k-ortalamalar	0,86	0,53	0,49	0,89
3NN-DBSCAN+k-medoid	0,90	0,52	0,50	0,89
3NN-k-ortalamalar	0,84	0,53	0,50	0,89
3NN-k-medoid	0,90	0,52	0,48	0,89
3NN-LOF+k-ortalamalar	0,86	0,53	0,50	0,89
3NN-LOF+k-medoid	0,90	0,52	0,48	0,89
3NN-Rassal	0,83	0,51	0,50	0,89
5NN-DBSCAN	0,66	0,49	0,50	0,89
5NN-DBSCAN+k-ortalamalar	0,70	0,50	0,50	0,89
5NN-DBSCAN+k-medoid	0,78	0,52	0,50	0,89
5NN-k-ortalamalar	0,63	0,50	0,50	0,89
5NN-k-medoid	0,83	0,52	0,48	0,89
5NN-LOF+k-ortalamalar	0,64	0,50	0,50	0,89
5NN-LOF+k-medoid	0,83	0,52	0,48	0,89
5NN-Rassal	0,70	0,50	0,50	0,89

EK 2 1NN, 3NN ve 5NN Yöntemleri Başlangıç Sorgu Seçimi Ağırlıklı F-ölçütü Sonuçları

Yöntemler	Veri kümeleri			
	ECG5000	FORDA	FORDB	WAFER
1NN+DBSCAN	0,83	0,52	0,48	0,91
1NN+DBSCAN+k-ortalamalar	0,89	0,57	0,46	0,97
1NN+DBSCAN+k-medoid	0,88	0,53	0,46	0,86
1NN+k-ortalamalar	0,89	0,56	0,46	0,97
1NN+k-medoid	0,88	0,50	0,45	0,86
1NN+LOF+k-ortalamalar	0,89	0,56	0,46	0,97
1NN+LOF+k-medoid	0,88	0,49	0,45	0,86
1NN+Rassal	0,83	0,51	0,49	0,87
3NN+DBSCAN	0,69	0,47	0,44	0,86
3NN+DBSCAN+k-ortalamalar	0,82	0,46	0,41	0,84
3NN+DBSCAN+k-medoid	0,87	0,39	0,49	0,84
3NN+k-ortalamalar	0,81	0,45	0,42	0,84
3NN+k-medoid	0,87	0,39	0,48	0,84
3NN+LOF+k-ortalamalar	0,82	0,45	0,42	0,84
3NN+LOF+k-medoid	0,87	0,38	0,48	0,84
3NN+Rassal	0,78	0,45	0,45	0,84
5NN+DBSCAN	0,55	0,40	0,40	0,84
5NN+DBSCAN+k-ortalamalar	0,61	0,40	0,37	0,84
5NN+DBSCAN+k-medoid	0,71	0,35	0,46	0,84
5NN+k-ortalamalar	0,53	0,41	0,39	0,84
5NN+k-medoid	0,78	0,36	0,45	0,84
5NN+LOF+k-ortalamalar	0,56	0,39	0,39	0,84
5NN+LOF+k-medoid	0,78	0,36	0,45	0,84
5NN+Rassal	0,60	0,40	0,40	0,84

EK 3 Karar Ağacı Yöntemi Başlangıç Sorgu Seçimi Doğruluk Sonuçları

Yöntemler	Veri kümeleri			
	ECG5000	FORDA	FORDB	WAFER
DT-DBSCAN	0,75	0,50	0,51	0,86
DT-DBSCAN+k-ortalamalar	0,79	0,51	0,50	0,85
DT-DBSCAN+k-medoid	0,81	0,51	0,50	0,88
DT-k-ortalamalar	0,76	0,51	0,50	0,85
DT-k-medoid	0,82	0,51	0,49	0,88
DT-LOF+k-ortalamalar	0,81	0,51	0,50	0,83
DT-LOF+k-medoid	0,81	0,51	0,49	0,89
DT-Rassal	0,74	0,50	0,50	0,86

EK 4 Karar Ağacı Yöntemi Başlangıç Sorgu Seçimi Ağırlıklı F-ölçütü Sonuçları

Yöntemler	Veri kümeleri			
	ECG5000	FORDA	FORDB	WAFER
DT+DBSCAN	0,71	0,45	0,49	0,84
DT+DBSCAN+k-ortalamalar	0,77	0,46	0,47	0,85
DT+DBSCAN+k-medoid	0,79	0,48	0,48	0,84
DT+k-ortalamalar	0,74	0,47	0,45	0,85
DT+k-medoid	0,80	0,46	0,48	0,84
DT+LOF+k-ortalamalar	0,78	0,47	0,45	0,84
DT+LOF+k-medoid	0,79	0,46	0,48	0,85
DT+Rassal	0,71	0,47	0,48	0,83

EK 5 Gaussian Naive Bayes Yöntemi Başlangıç Sorgu Seçimi Doğruluk Sonuçları

Yöntemler	Veri kümeleri			
	ECG5000	FORDA	FORDB	WAHER
GNB-DBSCAN	0,65	0,49	0,50	0,87
GNB-DBSCAN+k-ortalamalar	0,85	0,51	0,50	0,82
GNB-DBSCAN+k-medoid	0,79	0,52	0,50	0,89
GNB-k-ortalamalar	0,70	0,52	0,50	0,75
GNB-k-medoid	0,84	0,51	0,50	0,89
GNB-LOF+k-ortalamalar	0,86	0,52	0,50	0,80
GNB-LOF+k-medoid	0,84	0,52	0,50	0,89
GNB-Rassal	0,68	0,50	0,50	0,89

EK 6 Gaussian Naive Bayes Yöntemi Başlangıç Sorgu Seçimi Ağırlıklı F-ölçütü Sonuçları

Yöntemler	Veri kümeleri			
	ECG5000	FORDA	FORDB	WAFER
GNB-DBSCAN	0,55	0,37	0,37	0,83
GNB-DBSCAN+k-ortalamalar	0,82	0,39	0,34	0,81
GNB-DBSCAN+k-medoid	0,72	0,35	0,40	0,84
GNB-k-ortalamalar	0,65	0,39	0,36	0,77
GNB-k-medoid	0,81	0,35	0,42	0,84
GNB-LOF+k-ortalamalar	0,83	0,39	0,36	0,79
GNB-LOF+k-medoid	0,80	0,35	0,42	0,84
GNB-Rassal	0,59	0,36	0,36	0,84

EK 7 Eğitim Süreleri ve Yöntemlerin Eğitim Sürelerine Göre Sıralamaları (Eğitim süreleri saniye olarak verilmiştir ve parantez içindeki değerler eğitim sürelerine göre yöntemin sıralamasını göstermektedir.)

Yöntemler	Veri kümeleri			
	ECG5000	FORDA	FORDB	WAFER
1NN-Rassal	1101 (39)	365 (37)	885 (40)	1247 (39)
1NN-DBSCAN	468 (36)	510 (40)	297 (32)	351 (36)
1NN-DBSCAN+k-medoid	261 (31)	342 (36)	304 (34)	275 (33)
1NN-LOF+k-medoid	221 (27)	249 (29)	230 (27)	229 (26)
1NN-k-medoid	213 (26)	246 (28)	259 (29)	256 (30)
1NN-LOF+k-ortalamalar	92 (22)	125 (25)	112 (24)	108 (25)
1NN-k-ortalamalar	84 (21)	108 (21)	114 (25)	85 (19)
1NN-DBSCAN+k-ortalamalar	60 (12)	79 (16)	90 (19)	70 (17)
3NN-Rassal	1121 (40)	316 (35)	868 (39)	1284 (40)
3NN-DBSCAN	407 (35)	245 (27)	204 (26)	297 (35)
3NN-DBSCAN+k-medoid	283 (34)	313 (34)	312 (35)	265 (31)
3NN-LOF+k-medoid	259 (30)	289 (33)	248 (28)	235 (28)
3NN-k-medoid	221 (28)	250 (30)	354 (37)	250 (29)
3NN-LOF+k-ortalamalar	104 (24)	113 (23)	101 (22)	100 (23)
3NN-k-ortalamalar	83 (20)	101 (19)	89 (18)	90 (21)
3NN-DBSCAN+k-ortalamalar	61 (14)	76 (14)	81 (16)	67 (15)
5NN-Rassal	1091 (38)	381 (39)	633 (38)	1183 (38)
5NN-DBSCAN	468 (37)	375 (38)	284 (31)	404 (37)
5NN-LOF+k-medoid	275 (33)	279 (31)	271 (30)	274 (32)
5NN-k-medoid	272 (32)	243 (26)	298 (33)	234 (27)
5NN-DBSCAN+k-medoid	224 (29)	285 (32)	351 (36)	283 (34)
5NN-k-ortalamalar	82 (19)	104 (20)	89 (17)	86 (20)
5NN-LOF+k-ortalamalar	98 (23)	120 (24)	112 (23)	106 (24)
5NN-DBSCAN+k-ortalamalar	62 (15)	81 (17)	81 (15)	70 (16)
DT-k-medoid	64 (16)	33 (7)	34 (7)	20 (7)
DT-Rassal	51 (10)	55 (10)	45 (9)	40 (10)
DT-DBSCAN	41 (9)	67 (12)	62 (12)	50 (12)
DT-LOF+k-medoid	39 (8)	63 (11)	61 (11)	44 (11)
DT-DBSCAN+k-medoid	28 (7)	44 (8)	36 (8)	30 (8)
DT-LOF+k-ortalamalar	9 (3)	24 (6)	22 (4)	12 (4)
DT-DBSCAN+k-ortalamalar	9 (2)	22 (4)	21 (3)	12 (5)
DT-k-ortalamalar	8 (1)	19 (1)	18 (1)	10 (1)
GNB-DBSCAN	110 (25)	51 (9)	50 (10)	35 (9)
GNB-DBSCAN+k-medoid	79 (18)	109 (22)	94 (21)	77 (18)
GNB-k-medoid	64 (17)	71 (13)	68 (13)	56 (14)

EK 7 Eğitim Süreleri ve Yöntemlerin Eğitim Sürelerine Göre Sıralamaları (Eğitim süreleri saniye olarak verilmiştir ve parantez içindeki değerler eğitim sürelerine göre yöntemin sıralamasını göstermektedir.) (devamı)

Yöntemler	Veri kümeleri			
	ECG5000	FORDA	FORDB	WAFER
GNB-Rassal	60 (13)	95 (18)	92 (20)	96 (22)
GNB-LOF+k-ortalamalar	58 (11)	78 (15)	72 (14)	54(13)
GNB-k-ortalamalar	15 (6)	22 (3)	20 (2)	12 (3)
GNB-LOF+k-ortalamalar	12 (5)	24 (5)	26 (6)	12 (6)
GNB-DBSCAN+k-ortalamalar	11 (4)	21 (2)	22 (5)	11 (2)