
GAZİ ÜNİVERSİTESİ
Gazi
1926

GELİŞMİŞ META MODELLEME İÇİN YİNELEMELİ DENETİMSİZ ÖĞRENME TABANLI ÖZELLİK ÇIKARMA

Bora IŞIKOĞLU

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MART 2025

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Bora IŞIKOĞLU
18/03/2025

GELİŞMİŞ META MODELLEME İÇİN YİNELEMELİ DENETİMSİZ ÖĞRENME TABANLI ÖZELLİK ÇIKARMA

(Yüksek Lisans Tezi)

Bora IŞIKOĞLU

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Mart 2025

ÖZET

Günümüzde denetimli öğrenme yöntemleri büyük başarılar elde etmesine rağmen, yeterli etiketli veriye erişimin sınırlı olduğu durumlarda model performansı ciddi şekilde etkilenmektedir. Bu durum, özellikle karmaşık veri yapılarında gizli kalıpların keşfedilmesini ve modelin genellenebilirliğini zorlaştırmaktadır. Bu tezde, bu probleme çözüm olarak, etiketlenmemiş verilerin bilgi potansiyelinden yararlanarak model doğruluğunu artırmayı amaçlayan yeni bir denetimsiz özellik çıkarım yaklaşımı sunulmaktadır. Önerilen yöntemde, değişken küme parametresine sahip denetimsiz öğrenme algoritmaları kullanılarak veri kümesi üzerinde farklı küme yapılandırmaları keşfedilmekte ve bu süreçten elde edilen çıktıların modele yeni özellikler olarak eklenmesi sağlanmaktadır. Bu yinelemeli entegrasyon, modelin başarımı belirli bir yakınsama noktasına ulaşana kadar sürdürülmekte ve her döngüde model yeniden eğitilmektedir. Yinelemeli özellik çıkarma tabanlı optimizasyon, hem veri içindeki gizli yapıları ortaya çıkarmakta hem de denetimsiz öğrenmenin model iyileştirmede aktif rol almasını sağlamaktadır. Diyabet ve Wisconsin Meme Kanseri Tespit veri setleri üzerinde gerçekleştirilen deneyler, önerilen yöntemin model doğruluğunu anlamlı düzeyde artırdığını ve farklı veri setlerine karşı yüksek uyarlanabilirlik gösterdiğini ortaya koymuştur. Bu tez, önerilen yaklaşımın yalnızca teorik olarak değil, aynı zamanda pratik uygulamalarda da etkili sonuçlar verdiğini ortaya koymakta; denetimsiz öğrenme alanında eğitim sonrası model optimizasyonuna yönelik yeni bir bakış açısı sunmaktadır.

Bilim Kodu	:	92432
Anahtar Kelimeler	:	Denetimsiz öğrenme, özellik çıkarma, k-ortalamalar kümeleme, yinelemeli değerlendirme, yakınsama analizi, gizli desen, yinelemeli iyileştirme
Sayfa Adedi	:	43
Danışman	:	Prof. Dr. Aydın ÇETİN

ITERATIVE UNSUPERVISED LEARNING-BASED FEATURE EXTRACTION FOR ENHANCED META-MODELING

(M. Sc. Thesis)

Bora IŞIKOĞLU

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

March 2025

ABSTRACT

Although supervised learning methods have achieved significant success in recent years, the performance of such models can be severely limited in scenarios where access to labeled data is insufficient. This limitation poses challenges especially in identifying hidden patterns within complex data structures and in enhancing the generalizability of models. In response to this problem, this thesis proposes a novel unsupervised feature extraction approach that leverages the latent information in unlabeled data to improve model accuracy. The proposed method employs unsupervised learning algorithms with a variable clustering parameter to explore diverse cluster structures within the dataset. The resulting cluster outputs are iteratively added to the model as new features. This iterative integration continues until the model's performance reaches a convergence point, with the model being retrained at each cycle. The optimization process based on iterative feature extraction not only reveals latent structures in the data but also enables unsupervised learning to play an active role in model enhancement. The effectiveness of the proposed approach is demonstrated on the Diabetes and Wisconsin Breast Cancer Detection datasets, where empirical results show a significant improvement in model accuracy and high adaptability across different datasets. This thesis highlights that the proposed method provides not only theoretical contributions but also practical value, and offers a new perspective for post-training model optimization within the scope of unsupervised learning.

Science Code	:	92432
Key Words	:	Unsupervised learning, feature extraction, k-means clustering, iterative evaluation, convergence analysis, latent pattern, iterative refinement
Page Number	:	43
Supervisor	:	Prof. Dr. Aydın ÇETİN

TEŞEKKÜR

Bu yüksek lisans tezi, uzun ve zorlu bir sürecin sonucu olarak ortaya çıkmış hem akademik hem de kişisel gelişimime önemli katkılar sağlamıştır. Bu süreçte desteğini ve rehberliğini esirgemeyen danışman hocam Prof. Dr. Aydın ÇETİN'e, değerli bilgisi ve yol göstericiliği için içten teşekkürlerimi sunarım. Onun yönlendirmesi ve teşvik edici yaklaşımı, bu çalışmanın her aşamasında bana ışık tutmuş ve süreci anlamlı kılmıştır.

Ayrıca, tez konumun geliştirilmesine katkı sağlayan, çalışmalarımla ilgili geri bildirimler ve öneriler sunan değerli jüri üyeleri, Prof. Dr. Aysun COŞKUN'a ve Dr. Öğr. Üyesi Sevgi YİĞİT SERT'e teşekkür ederim. Akademik tartışmalar ve eleştiriler, çalışmamın daha sağlam bir temele oturmasını ve araştırmamın kapsamını genişletmemi sağlamıştır.

Süreç boyunca, bilgi ve motivasyonlarıyla yanımda olan, tartışmaları ve paylaşımlarıyla çalışmama katkı sağlayan kıymetli arkadaşım Meerim KUBATOVA'ya da minnettarım.

Son olarak, bu süreçte bana manevi destek sağlayan ve daima yanımda olan aileme, sabırları ve teşvikleri için sonsuz teşekkür ederim.

İÇİNDEKİLER

Sayfa

ÖZET iv
ABSTRACT v
TEŞEKKÜR vi
İÇİNDEKİLER vii
ÇİZELGELERİN LİSTESİ ix
ŞEKİLLERİN LİSTESİ x
SİMGELER VE KISALTMALAR xi
1. GİRİŞ 1
2. LİTERATÜR TARAMASI 3
3. YÖNTEM VE ARAÇLAR 7
3.1. Veri Seti 8
3.2. Geliştirme Ortamı ve Araçlar 11
3.3. Yöntem Seçimi 13
3.3.1. Denetimsiz öğrenmenin kullanımı 15
3.3.2. K-ortalamalar algoritması 16
3.3.3. Random Forest algoritması 17
3.3.4. Optuna 17
3.4. Model ve Değerlendirme Parametreleri 18
3.5. Model Eğitimi ve Yinelemeli Özellik Çıkarımı 19
3.5.1. Önerilen yinelemeli özellik çıkarma tabanlı optimizasyon yöntemi 19
3.6. Sonuçların Kıyaslanması ve Nihai Model Seçimi 21
4. BULGULAR VE DEĞERLENDİRME 23
4.1. Değerlendirme Kriterleri 23
4.2. Önerilen Yöntemin Mevcut Yöntemlerle Kıyaslama Senaryoları 24
4.2.1. Klasik optimizasyon yöntemleriyle elde edilen sonuçlar 24
4.2.2. Optuna ile elde edilen sonuçlar 25
4.2.3. Önerilen yöntem ile elde edilen sonuçlar 26

Sayfa

4.2.4. Önerilen yöntemin optuna ile kullanılmasıyla elde edilen sonuçlar 33
4.2.5. Tüm senaryoların F1-skoru karşılaştırma tablosu..... 34
4.3. Yorum ve Çıkarımlar 34
5. SONUÇ VE ÖNERİLER 37
KAYNAKLAR 39
ÖZGEÇMİŞ 42

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge 3.1. Diyabet veri seti içeriği 8

Çizelge 3.2. Wisconsin meme kanseri veri setinin içeriği 9

Çizelge 3.3. Kullanılacak araç ve kütüphaneler 12

Çizelge 4.1. Diyabet veri kümesi üzerinde önerilen yöntemin elde ettiği F1-Skoru sonuçları 26

Çizelge 4.2. Diyabet veri kümesi üzerinde önerilen yöntemin elde ettiği kesinlik sonuçları 27

Çizelge 4.3. Diyabet veri kümesi üzerinde önerilen yöntemin elde ettiği duyarlılık sonuçları 27

Çizelge 4.4. Diyabet veri seti üzerinde iteratif optimizasyon eğitimi sonuçlarının karmaşıklık matrisleri 28

Çizelge 4.5. Wisconsin veri kümesi üzerinde önerilen yöntemin elde ettiği F1-Skoru sonuçları 30

Çizelge 4.6. Wisconsin veri kümesi üzerinde önerilen yöntemin elde ettiği kesinlik sonuçları 31

Çizelge 4.7. Wisconsin veri kümesi üzerinde önerilen yöntemin elde ettiği duyarlılık sonuçları 31

Çizelge 4.8. Wisconsin meme kanseri veri seti üzerinde iteratif optimizasyon eğitimi sonuçlarının karmaşıklık matrisleri 31

Çizelge 4.9. Tüm senaryoların karşılaştırma tablosu 34

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil 3.1. Geliştirme sürecinin akış şeması 7

Şekil 3.2. Diyabet Veri Setinin Hedef Değişkeni Sınıf Dağılımı 9

Şekil 3.3. Wisconsin Meme Kanseri Veri Seti Hedef Değişkeni Sınıf Dağılımı 11

Şekil 3.4. Denetimli öğrenmenin çalışma prensibi 14

Şekil 3.5. Denetimsiz öğrenmenin çalışma prensibi 15

Şekil 3.6. Tekrarlı eğitimlere dayalı özellik çıkarımının iş akışı 20

Şekil 4.1. Diyabet veri seti üzerinde gerçekleştirilen eğitim sonuçlarının kümelenmiş sütun grafiği 29

Şekil 4.2. Wisconsin meme kanseri veri seti üzerinde gerçekleştirilen eğitim sonuçlarının kümelenmiş sütun grafiği 33

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	**Açıklamalar**
$\mathbf{F_1 - Skoru}$	Duyarlılık ve keskinliğin harmonik ortalaması
Precision	Kesinlik
Recall	Duyarlılık

Kısaltmalar	**Açıklamalar**
BT	Bilgisayar Tomografisi
FILS	İleri Yinelemeli Laplasyen Puanı
FN	False Negative
FP	False Positive
GLCM	Gri Seviye Eşzamanlı Oluşumu
IFILS	Artımlı İleri Yinelemeli Laplasyen Puanı
L1	Lasso
L2	Ridge
PCM	Darbe Kod Modülasyonu
TN	True Negative
TP	True Positive
YZ	Yapay Zekâ

1. GİRİŞ

Günümüzde yapay zekâ (YZ) modelleri, çeşitli disiplinlerde karmaşık sorunları çözmek için yaygın olarak kullanılmaktadır [1]. Bu modellerin başarısı, büyük ölçüde doğru hiperparametrelerin belirlenmesine ve etkili bir özellik çıkarma sürecine bağlıdır. Hiperparametre optimizasyonu ve özellik mühendisliği, YZ modellerinin doğruluk ve genelleştirme kabiliyetini doğrudan etkileyen kritik süreçlerdir. Ancak mevcut yöntemler genellikle yüksek maliyetli, zaman alıcı ve dinamik veri yapıları için yeterince esnek değildir. Bu durum, özellikle karmaşık veri kümelerinde, daha etkili, adaptif ve esnek optimizasyon tekniklerine olan ihtiyacı artırmaktadır.

Hiperparametre optimizasyonu süreçlerinde sıklıkla kullanılan Grid Search, rastgele arama gibi geleneksel yöntemler, yüksek boyutlu veri setlerinde etkisiz hale gelmekte ve yüksek hesaplama maliyetlerine yol açmaktadır. Literatürde yaygın olarak önerilen Bayesian optimizasyon, genetik algoritmalar ve denetimli öğrenme tabanlı yaklaşımlar gibi yöntemler ise statik bir süreç izlemekte ve veri yapısındaki değişikliklere yeterince adapte olamamaktadır. Benzer şekilde, özellik mühendisliği süreçlerinde kullanılan manuel veya denetimli yaklaşımlar, veri setlerindeki gizli kalıpları yeterince iyi yakalayamayabilir. Bu eksiklikler, özellikle dinamik ve karmaşık veri kümeleriyle çalışırken, daha verimli, otomatik ve adaptif çözümlere duyulan gereksinimi artırmaktadır.

Denetimsiz öğrenme (özellikle k-ortalamalar kümelemesi gibi algoritmalar), etiketlenmiş veri gereksinimini ortadan kaldırarak veri yapısının içsel kalıplarını keşfetmek için önemli fırsatlar sunmaktadır. K-ortalamalar kümelemesi gibi yöntemler, denetimsiz öğrenmenin gücünden yararlanarak karmaşık veri kümelerinden anlamlı özellikler çıkarmak için yaygın bir şekilde kullanılmaktadır [2]. Bu yöntemler, verilerin içsel yapısını anlamada güçlü bir araç sunarak, daha az insan müdahalesi gerektiren çözümler sunmaktadır.

Son yıllarda, denetimsiz öğrenme tekniklerinin gelişimi ve analizi, karmaşık veri kümelerinden anlamlı özellikler çıkarmak için yeni fırsatlar sunmuştur. K-ortalamalar kümelemesi gibi yöntemler, etiketli verilere ihtiyaç duymadan verilerdeki gizli kalıpları belirlemek için güçlü araçlar arasında yer almaktadır. Denetimsiz öğrenme, etiketli veri gereksinimini ortadan kaldırarak verilerin içsel yapısını keşfetmeye ve anlamaya olanak

tanımaktadır. Bu bağlamda, özellik çıkarımı için denetimsiz öğrenme yaklaşımları, dinamik veri setleriyle çalışmak için önemli avantajlar sunmaktadır.

Bu tez çalışmasında, dinamik ve yinelemeli bir özellik çıkarma süreci sunan yeni bir hiperparametre optimizasyon yöntemi önerilmektedir. Önerilen yöntem, k-ortalamalar kümelemesiyle elde edilen bilgiyi orijinal veri setine ekleyerek veri setindeki gizli kalıpları keşfetmeyi ve bu özellikler doğrultusunda hiperparametreleri optimize etmeyi amaçlamaktadır. Yöntemin, veri setinin tüm özelliklerini yinelemeli olarak değerlendirip iyileştirerek mevcut modellerin performansını artırma potansiyeline sahip olması amaçlanmıştır. Özellikle, etiketlenmemiş veri kümelerindeki özelliklerin dinamik olarak çıkarılması ve değerlendirilmesi sayesinde, mevcut yöntemlerin sınırlamalarının ötesine geçilmesi hedeflenmiştir.

Önerilen yöntemin hem model doğruluğunu artırmada hem de veri setlerindeki [3,4] gizli kalıpları ortaya çıkarmadaki başarısı çeşitli çalışmalarla değerlendirilmiştir. Bu tez, önerilen yöntemin hem teorik hem de pratik katkılarını detaylandırmış ve gelecekteki çalışmalar için yeni bir bakış açısı sunmayı hedeflemiştir.

Tez, beş bölümden oluşmaktadır. Problemin tanımı ve önerilen yaklaşım Giriş bölümünde detaylandırılmıştır. Mevcut hiperparametre optimizasyonu ve özellik çıkarma yöntemleri ve bu yöntemlerin başarıları ve sınırlamaları Literatür Taraması bölümü altında değerlendirilmiştir. Önerilen yöntem, Yöntem ve Araçlar bölümü altında detaylı olarak açıklanmış, kullanılan algoritmaları, süreçleri ve dinamik yaklaşımları tanıtılmıştır. Tezde, Diyabet ve Wisconsin Meme Kanseri Tespit veri setleri üzerinde gerçekleştirilen deney sonuçlar ve performans değerlendirmeleri Bulgular ve Değerlendirme bölümü altında sunulmuştur. Sonuç bölümünde teze ait genel değerlendirmeler özetlenerek önerilen yöntemin avantajlarını tartışılmış ve gelecekteki çalışmalar için öneriler sunulmuştur.

2. LİTERATÜR TARAMASI

Denetimsiz öğrenme ve özellik çıkarma, verilerdeki gizli kalıpları ortaya çıkarmak ve makine öğrenimi modellerinin performansını artırmak için temel yöntemler arasında yer almaktadır [5]. Bu yöntemler, özellikle karmaşık veri kümeleri üzerinde etkili modeller geliştirilmesinde önemli bir rol oynamaktadır. Hiperparametre optimizasyonu ve dinamik özellik çıkarma teknikleri, model performansını iyileştirmeye yönelik araştırmalarda sıklıkla ele alınmış ve literatürde bu alanda çeşitli yöntemler ve yaklaşımlar önerilmiştir [6,7].

Denetimsiz öğrenme teknikleri arasında, k-ortalamalar kümeleme algoritması, basitliği ve uyarlanabilirliği nedeniyle öne çıkmaktadır. Bu algoritma, darbe kod modülasyonu (PCM) ve en küçük kareler kantizasyonu prensiplerine dayanmaktadır [8]. Literatürde, k-ortalamalar algoritmasının verilerdeki içsel yapıları ortaya çıkarma yeteneği ve farklı veri kümelerindeki etkinliği geniş bir şekilde ele alınmıştır [9]. Ancak, algoritmanın başlangıç merkez noktalarına duyarlılığı ve karmaşık veri kümelerinde yerel optimumlara takılma riski gibi sınırlamaları bulunmaktadır. Bu sorunları aşmak için literatürde k-ortalamaların çeşitli varyantları ve iyileştirilmiş versiyonları geliştirilmiştir.

Özellik çıkarma, model performansını artırmada kritik bir adımdır. Literatürde, özellik seçimi, düzenleme teknikleri ve dönme değişmezliğinin inceliklerini araştırarak etkili özellik çıkarma için kritik olan hususlara ışık tuttu [10]. L1 düzenlemesi, özelliklerin seyrekleştirilmesini sağlarken, L2 düzenlemesi, aşırı öğrenmeyi önlemek için özellik değerlerini küçültmektedir. Çalışmada, L1 ve L2 düzenlemesi arasındaki takaslar ele alınmış ve farklı düzenleme yöntemlerinin özellik önemi üzerindeki etkisine dair değerli içgörüler sağlanmıştır. Bu yöntemlerin kombinasyonu, farklı veri kümeleri ve uygulama senaryoları için esnek çözümler sunmaktadır.

Ayrıca, literatürde İleri Yinelemeli Laplasyen Puanı (FILS) algoritması gibi denetimsiz özellik seçimi teknikleri geliştirilmiştir. FILS, Özellik Önemi (SIG) kriterini kullanarak özellik çıkarma sürecinde önemli bir katkı sunmuştur [11]. Bu ekleme, FILS'deki granülasyon monotonluğu ve belirsizliği ele alarak, daha iyi sınıflandırma performansı gösteren Artımlı İleri Yinelemeli Laplasyen Puanı (IFILS) algoritmasıyla sonuçlanmıştır. Ancak, IFILS algoritmasının artan karmaşıklığı, daha yüksek hesaplama maliyetlerine yol

açabilir ve büyük veri kümeleri veya kaynak kısıtlı ortamlar için pratikliğini sınırlayabilir.

Modern hiperparametre optimizasyon teknikleri arasında OPTUNA öne çıkmaktadır [12]. OPTUNA, dinamik ve adaptif bir arama stratejisi benimseyerek optimizasyon süreçlerini otomatikleştirmekte ve hiperparametre ayarlamalarını daha etkili bir şekilde gerçekleştirmektedir. Literatürde, Bayesian optimizasyon, grid search ve rastgele arama gibi geleneksel yöntemlerin aksine OPTUNA'nın, arama alanını sürekli iyileştirme yeteneği sayesinde daha verimli sonuçlar sunduğu belirtilmiştir. OPTUNA'nın modern yapısı, hiperparametre optimizasyonunda önemli bir katkı sağlamaktadır.

Denetimsiz öğrenme ve özellik çıkarma uygulamalarında, k-ortalamalar algoritmasının farklı veri kümelerindeki etkinliği kanıtlanmıştır. Çalışma [13], k-ortalamalar kümeleme algoritmasının evrimini, varyantlarını ve uygulamalarını tartışarak kapsamlı bir inceleme sunmaktadır. Algoritmanın basitliğini ve verimliliğini vurgular ancak aynı zamanda ilk küme seçimine duyarlılık ve karmaşık veri kümelerini ele alma gibi zorlukları da vurgular. Makale, k-ortalamalar varyantlarının ayrıntılı bir taksonomisini sunar ve gelecekteki araştırma yönlerini belirler. Ancak, bir dezavantajı, yeni çözümlerin veya deneysel doğrulamanın eksikliğidir ve bu da alanın ilerlemesi üzerindeki etkisini sınırlar.

Diğer bir çalışmada bilgisayar tomografisi (BT) tarama görüntülerinden COVID-19 sınıflandırmasında, özellik çıkarma için morfolojik özelliklerin, birinci dereceden dokuların ve gri seviye eşzamanlı oluşumunun (GLCM) kullanımını araştırmaktadır. Özellik çıkarma, doğru sınıflandırma için çok önemlidir ve çalışma, eksantriklik ve kontrast gibi çeşitli özellikleri bu bağlamda değerlendirir. Araştırma, birinci dereceden doku çıkarmanın %50'lik en yüksek doğruluğa ulaştığını ve bu yöntemleri kullanarak COVID-19'u etkili bir şekilde sınıflandırmadaki zorlukları bulmuştur [14]. Ancak, bu çalışmada sınıflandırma doğruluğunun yalnızca %50'lerde kalması, mevcut tekniklerin yetersizliğini ve daha gelişmiş yöntemlerin gerekliliğini göstermektedir. Bununla birlikte, literatürde önerilen k-ortalamalar varyantları, başlangıç küme seçimi duyarlılığı gibi sınırlamaları ele almış ve algoritmanın çeşitli uygulamalardaki başarısını artırmıştır.

Literatürde, k-ortalamalar ve özellik çıkarımı odaklı başka çalışmalar da mevcuttur. İlgili çalışmada, biyometrik bir problem olan kalp sesiyle kimlik tanıma üzerine yaptıkları çalışmada Mel-frekans özniteliklerini k-ortalamalar ile kümeleyerek yeni özellikler

oluşturmuş ve ardından derin sinir ağları ile sınıflandırma yapmışlardır. Sonuç olarak önerilen sistem, %91'in üzerinde bir doğrulukla kişileri ayırt edebilmiştir. Bu çalışma, k-ortalamaların farklı özellik çıkarım teknikleriyle birleştirilerek sağlık gibi alanlarda başarılı uygulanabildiğine güzel bir örnek teşkil etmektedir [15].

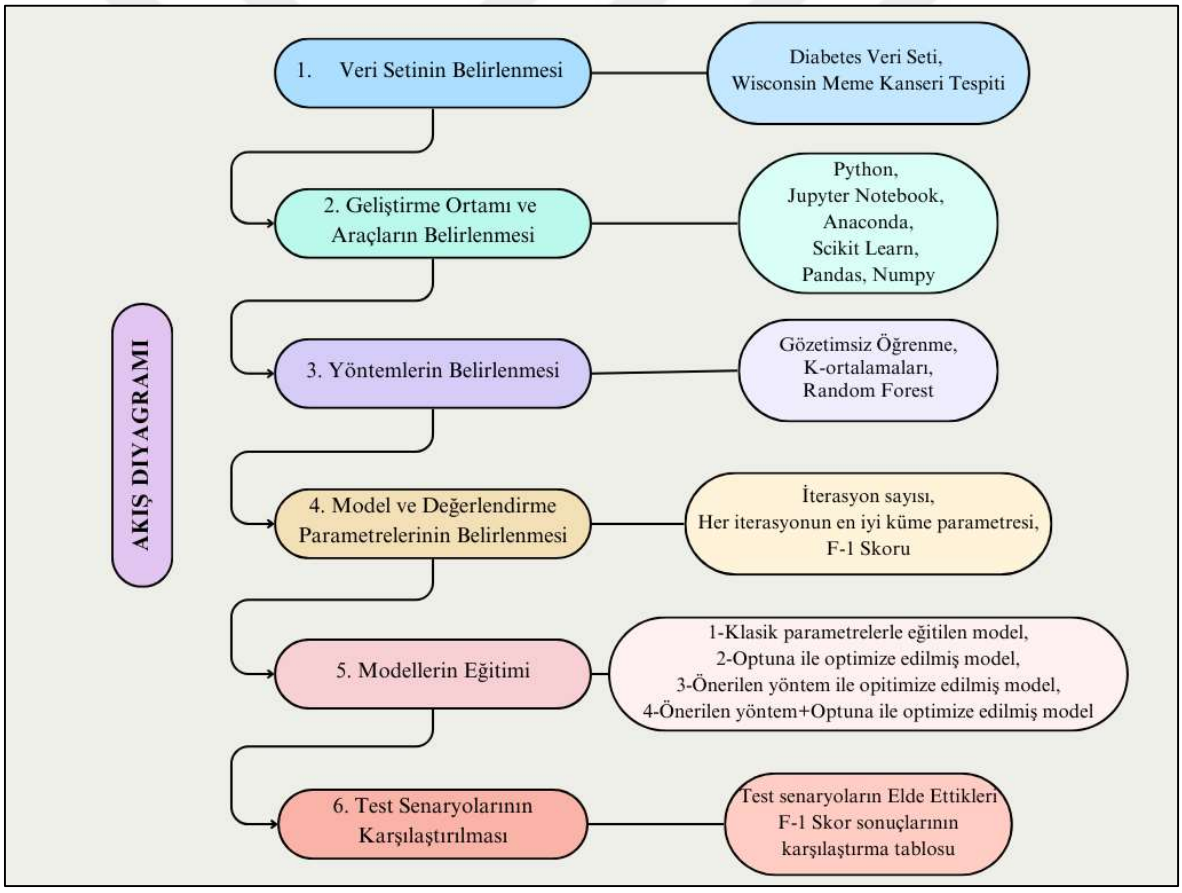
Bu literatürde sunulan yöntemler ve yaklaşımlar, önerilen metodolojinin motivasyonunu ve temelini oluşturmaktadır. Yinelemeli özellik çıkarma ve optimizasyon sürecine odaklanan önerilen yaklaşım, k-ortalamalar algoritmasını dinamik bir şekilde kullanarak orijinal veri setine yeni özellikler eklemekte ve bu özellikleri yinelemeli olarak değerlendirmektedir. Önerilen yöntem, OPTUNA gibi modern hiperparametre optimizasyon teknikleriyle kıyaslandığında, özellik çıkarma ve optimizasyon süreçlerini birleştirerek daha kapsamlı bir çözüm sunmayı hedeflemektedir.

Sonuç olarak, literatür, denetimsiz öğrenme ve özellik çıkarma tekniklerinin önemini vurgularken, bu yöntemlerin daha gelişmiş versiyonlarına duyulan ihtiyacı da ortaya koymaktadır. Mevcut çalışmalar, önerilen yöntemin temelini şekillendirmekte ve literatüre yapılacak katkının çerçevesini belirlemektedir [8,10,11,13,14]. Tezin ilerleyen bölümlerinde, bu tekniklerin detaylı kıyaslamaları ve deneysel sonuçları sunulacaktır.

3. YÖNTEM VE ARAÇLAR

Bu tez çalışmasında veri model optimizasyonunu gerçekleştirmek amacı ile, özellik çıkarımı sürecinde denetimsiz öğrenme temelli meta-model performansını artırmayı hedefleyen yinelemeli bir iş akışı ortaya konmuştur. Genel süreç, ham verilerin işlenmesinden başlayarak özellik çıkarımı, model eğitimi ve performans değerlendirmesi gibi bir dizi aşamayı içermektedir. Bu bölümde, metodolojinin tüm aşamaları ayrıntılı olarak açıklanmış, şekillerle desteklenmiş ve önerilen yöntemin faydaları kapsamlı bir şekilde ele alınmıştır.

Önerilen yöntemin sistematik akış şeması ve ilgili adımlar Şekil 3.1'de gösterilmektedir.



Şekil 3.1. Geliştirme sürecinin akış şeması

Süreç, veri setlerinin belirlenmesiyle başlayıp, özellik çıkarımı, model eğitimi ve performans değerlendirmesi ile sonuçlanmaktadır. Bu süreç, Bölüm 3.5'te detaylandırılmıştır.

3.1. Veri Seti

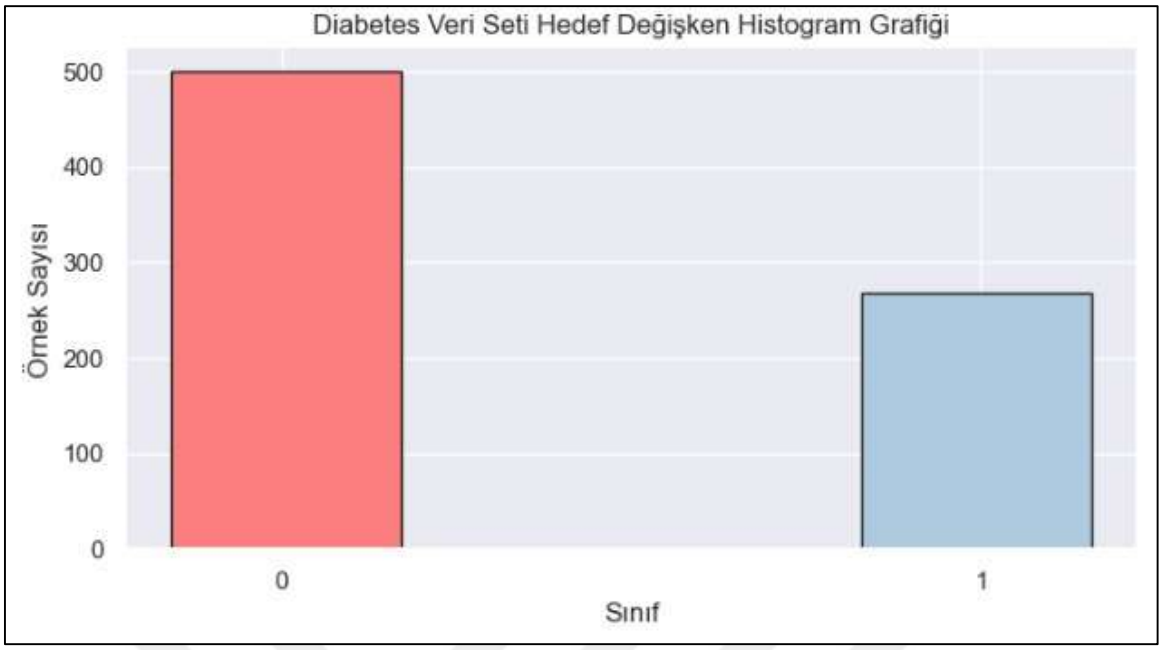
Tezde, Diyabet [3] ve Wisconsin Meme Kanseri [4] veri setleri kullanılmıştır. Bu veri setleri, sınıflandırma problemlerine uygun, yaygın olarak kullanılan ve erişilebilir veri kaynaklarıdır. Veri setleri, UCI Machine Learning Repository'de [16] yer almaktadır ve araştırmacılar tarafından eğitim ve deneysel amaçlar için ücretsiz olarak sağlanmıştır.

Diyabet veri seti, Ulusal Diyabet ve Sindirim ve Böbrek Hastalıkları Enstitüsü (National Institute of Diabetes and Digestive and Kidney Diseases) tarafından oluşturulmuş ve UCI Machine Learning Repository'de kamuya açık hale getirilmiştir [3]. Bu veri seti, Pima Kızılderilileri üzerinde yapılan bir çalışmaya dayanır ve bireylerin çeşitli medikal özellikleri ile diyabet teşhisi arasındaki ilişkiyi incelemek için tasarlanmıştır. Veri seti, tanı ölçümlerine dayanarak bir hastanın diyabet olup olmadığını tahmin etmeyi amaçlamaktadır. Çizelge 3.1, bu veri setlerinin temel özelliklerini özetlemektedir.

Çizelge 3.1. Diyabet veri seti içeriği

No	Özellik Adı	Açıklama
1	Gebelikler	Hamile kalma sayısı
2	Glikoz	Oral glikoz tolerans testinde 2 saatlik plazma glikoz konsantrasyonu
3	Tansiyon	Diyastolik kan basıncı (mm Hg)
4	Cilt Kalınlığı	Triceps deri kıvrım kalınlığı (mm)
5	İnsülin	2-Saatlik serum insülin (mu U/ml)
6	Vücut Kitle İndeksi	Vücut kitle indeksi (kg cinsinden ağırlık/(m cinsinden boy)^2)
7	Diyabet Soyağacı Fonksiyon	Diyabet soyağacı işlevi
8	Yaş	Yaş (yıl)
9	Sonuç	Sınıf değişkeni (0 veya 1)

Çizelge 3.1'de verilen Diyabet Veri Setindeki özniteliklerin tamamı sayısal verilerden oluşmaktadır. Veri setinde toplam 768 örnek mevcuttur. Hedef değişkeninin dağılımı Şekil 3.2'de paylaşılmıştır. Bu şekil incelendiğinde, iki sınıf arasında yaklaşık iki katı kadar bir dengesizlik gözlemlenmektedir.



Şekil 3.2. Diyabet Veri Setinin Hedef Değişkeni Sınıf Dağılımı

Diyabet veri setinde az örnek sayısı bulunmasından kaynaklı olarak test metriklerinden alınan sonuçların yanıltıcı olma ihtimalini engellemek amacı ile eğitim-test veri oranı %75-%25 şeklinde ayarlanmıştır.

Wisconsin Meme Kanseri veri seti, sınıflandırma problemleri için yaygın olarak kullanılan veri kaynağıdır. Bu veri seti, meme kanseri teşhisinde kullanılan örnek bir veri kümesidir ve aslen Wisconsin Üniversitesi Hastanesi'ndeki Dr. William H. Wolberg tarafından oluşturulmuştur [4]. Veri seti, kanserin “malign” (kötü huylu) veya “benign” (iyi huylu) olup olmadığını sınıflandırmak için kullanılmaktadır. Çizelge 3.2, Wisconsin meme kanseri veri setinin içeriği ve bu veri setlerinin temel özelliklerini özetlemektedir.

Çizelge 3.2. Wisconsin meme kanseri veri setinin içeriği

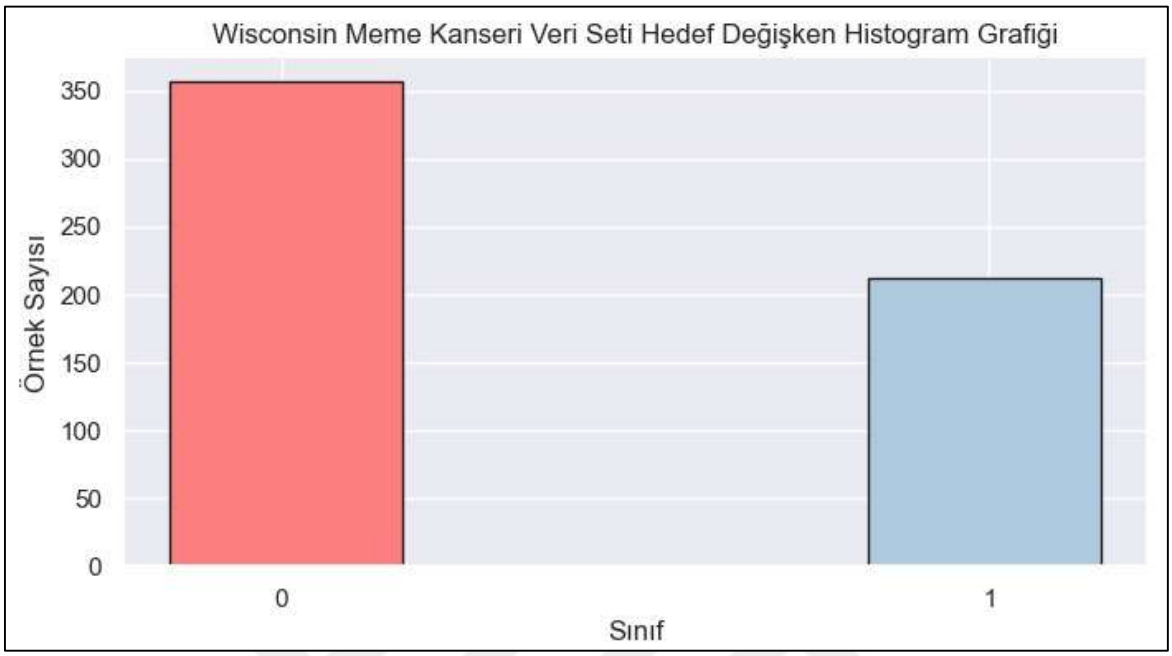
No	Özellik Adı	Açıklama
1	ID Number	Hasta kimliğini temsil eden benzersiz numara.
2	Diagnosis	Teşhis sonucu; 'M' (Malign - Kötü huylu) veya 'B' (Benign - İyi huylu).
3	Radius Mean	Hücre çekirdeğinin yarıçapının ortalaması.
4	Texture Mean	Gri ölçeği yoğunluğunun standart sapması.
5	Perimeter Mean	Hücre çevresinin ortalaması.

Çizelge 3.2. (devam) Wisconsin meme kanseri veri setinin içeriği

6	Area Mean	Hücre alanının ortalaması.
7	Smoothness Mean	Hücre sınırlarının düzgünlüğünün ortalaması.
8	Compactness Mean	Perimeter² / Alan - 1.0 oranı ile ölçülen kompaktlık.
9	Concavity Mean	Hücre çekirdeği sınırlarındaki içbükey kısımların ortalaması.
10	Concave Points Mean	Hücre çekirdeği sınırındaki içbükey noktaların ortalaması.
11	Symmetry Mean	Hücrelerin simetrik yapısının ortalaması.
12	Fractal Dimension Mean	Hücrelerin karmaşıklığını temsil eden fraktal boyutun ortalaması.
13	Radius Worst	Hücre çekirdeği yarıçapının en kötü ölçülen değeri.
14	Texture Worst	Gri ölçeği yoğunluğunun en kötü standart sapması.
15	Perimeter Worst	Hücre çevresinin en kötü ölçülen değeri.
16	Area Worst	Hücre alanının en kötü ölçülen değeri.
17	Smoothness Worst	Hücre sınırlarının en kötü düzgünlük değeri.
18	Compactness Worst	En kötü kompaktlık değeri.
19	Concavity Worst	En kötü içbükey kısımların değeri.
20	Concave Points Worst	En kötü içbükey noktaların sayısı.
21	Symmetry Worst	Hücrelerin en kötü simetrik yapı değeri.
22	Fractal Dimension Worst	Hücrelerin en kötü fraktal boyut değeri.
23	Radius SE	Hücre yarıçapı için standart hata.
24	Texture SE	Doku yoğunluğu için standart hata.
25	Perimeter SE	Hücre çevresi için standart hata.
26	Area SE	Hücre alanı için standart hata.
27	Smoothness SE	Hücre düzgünlüğü için standart hata.
28	Compactness SE	Kompaktlık için standart hata.
29	Concavity SE	İçbükey kısımlar için standart hata.
30	Concave Points SE	İçbükey noktaların standart hatası.
31	Symmetry SE	Hücrelerin simetrik yapısındaki standart hata.
32	Fractal Dimension SE	Hücrelerin karmaşıklığını temsil eden standart hata.

“Diognosis” hedef değişkeninin teşhis sonucunu ifade eden “malign” ve “benign” sınıflarının dağılımı Şekil 3.3’te gösterilmiştir. Dağılıma bakıldığı zaman, bir sınıf diğerinin

yaklaşık iki katını oluşturmaktadır ve bu sınıfın dengesizliğini göstermektedir.



Şekil 3.3. Wisconsin Meme Kanseri Veri Seti Hedef Değişkeni Sınıf Dağılımı

Wisconsin Meme Kanseri veri setindeki örnek sayısı 569'dur. Veri setinde az örnek sayısı bulunması nedeniyle test metriklerinden alınan sonuçların yanıltıcı olma ihtimalini indirgemek için eğitim-test veri oranı %75-%25 şeklinde ayarlanmıştır.

3.2. Geliştirme Ortamı ve Araçlar

Bu tez bünyesinde kullanılan veri hazırlama yöntemleri, model eğitimi kütüphaneleri, optimizasyon algoritmaları ve veriler literatürde yapılan çalışmalar incelenerek ve önerilen problemin çözümüne uygun olarak çeşitli açık kaynaklardan seçilmiştir. Çizelge 3.3'de, bu çalışmada kullanılan araç ve kütüphaneleri özetlemektedir.

Programlama dili seçilirken, literatürde yapılan çalışmaların büyük bir çoğunluğunda kullanılması ve bu sayede sahip olduğu büyük miktardaki kütüphanelerin değerlendirmek amacı ile Python [17] kullanılmıştır. Bu kütüphaneler veri işleme model oluşturma gibi aşamaları daha kolay ve verimli hale getirmektedir.

Geliştirme ortamı olarak Jupyter Notebook'un [18] seçilmesinin sebebi, hücreli çalışma yapısı ile sağladığı avantajlardan kaynaklıdır. Bu hücreli yapı sayesinde her hücrenin çıktısı

kademeli olarak hafızada saklanır ve her bir hücrede yapılan değişikliklerin sonucu tüm programı tekrar çalıştırmadan elde edilebilir.

Çizelge 3.3. Kullanılacak araç ve kütüphaneler

No	Araç/Kütüphane	Açıklama
1	Python	Programlama dili
2	Jupyter Notebook	Etkileşimli kodlama ortamı
3	Anaconda	Geliştirme ve bağımlılık yönetimi
4	Scikit-learn	Makine öğrenimi algoritmaları ve araçları
5	K-ortalamalar, Random Forest	Kullanılan algoritmalar
6	Pandas, NumPy	Veri işleme ve manipülasyon
7	Diyabet ve Wisconsin Veri Seti	Çalışmada kullanılan veri setleri

Geliştirme ve bağımlılık yönetimi ortamı olan Anaconda'dan [19], Python ortam yönetimi ve bağımlılık kontrolleri için sağladığı arayüz ve araçlar kullanılmıştır. Bu araçlar sayesinde bağımlılıkların versiyon kontrolü kolayca yapılabilmekte ve Python ortamının anlık kopyası alınabilmektedir.

Scikit-learn [20], bünyesinde bulunan makine öğrenimi algoritmaları ile yöntem geliştirme, karşılaştırma ve test aşamalarında çeşitli makine öğrenimi algoritmalarını hızlı uyarlamak için kullanılmıştır. Aynı zamanda literatürde ulaşılabilecek birçok dokümana sahip olması da yapılacak olan testlerin daha kapsamlı hale gelmesine destek sağlamıştır.

Bu tezde örnek olarak ele alınan Random Forest [21] algoritması, optimizasyonsuz ve optimizasyonlu veri arasındaki gelişme miktarını ölçecek bir temel model olarak seçilmiştir. Deterministic çalışan diğer öğrenme algoritmaları Random Forest yerine yerleştirilerek kullanılabilirdi, bu durum yapılan çalışmanın yüksek uyarlanabilirliğini göstermektedir. Random Forest algoritması iyileşmenin tam sonuçlarını görmek amacı ile literatürden orijinal veri seti üzerinde yüksek sonuçlar vermesi sebebi ile kullanılmıştır. Optimizasyon algoritması olarak seçilebilecek dentimsiz öğrenme yöntmeleri arasından K-ortalamalar algoritmasının [13] optimize edici olarak seçilmesinin birden fazla sebebi bulunmaktadır.
Öncelikle, optimize edici algoritmanın çıktıları veriye geri ekleneceği için hedef değişkeni kullanmayan bir algoritma gereklidir yani denetimsiz bir model eğitimi yapılmalıdır. Aynı

zamanda seçilen algoritma daha sonra meta modelin bir parçası olacağı için deterministik bir şekilde çalışabilmelidir. Seçilen algoritma tekrarlı bir şekilde eğitileceği için hızlı çalışması gerekmektedir. Bu kriterlere uygun olarak K-ortalamalar algoritması seçilmiştir.

Veri işleme ve manipülasyon işlemleri için Pandas [22] ve NumPy [23] kütüphaneleri kullanılmıştır. Bu kütüphanler literatürün yaygın kullanılan yöntemlerini kolay ve hızlı uyarlamayı sağlaması ve yine literatürde bunun gibi örneklerinin bulunmasıdır.

Diyabet ve Wisconsin Meme Kanseri [3,4] veri setlerinin kullanılmasının sebepleri, temelde veri çıkarma işlemi yapıldığı için az miktarda sütun sayısı barındırmaları ve tekrarlı eğitim yapılacağı için örnek sayılarının az olması şeklinde sıralanabilir. Literatürde yapılan diğer çalışmalarda birden çok kez kullanılmış olmaları karşılaştırma ve örnekleme işlemlerini daha kolay hale getirmiştir [24], makine öğrenimi algoritmalarına kolay uyarlaması için sayısal ya da sınıfsal verilere sahip olması da sebepler arasındadır.

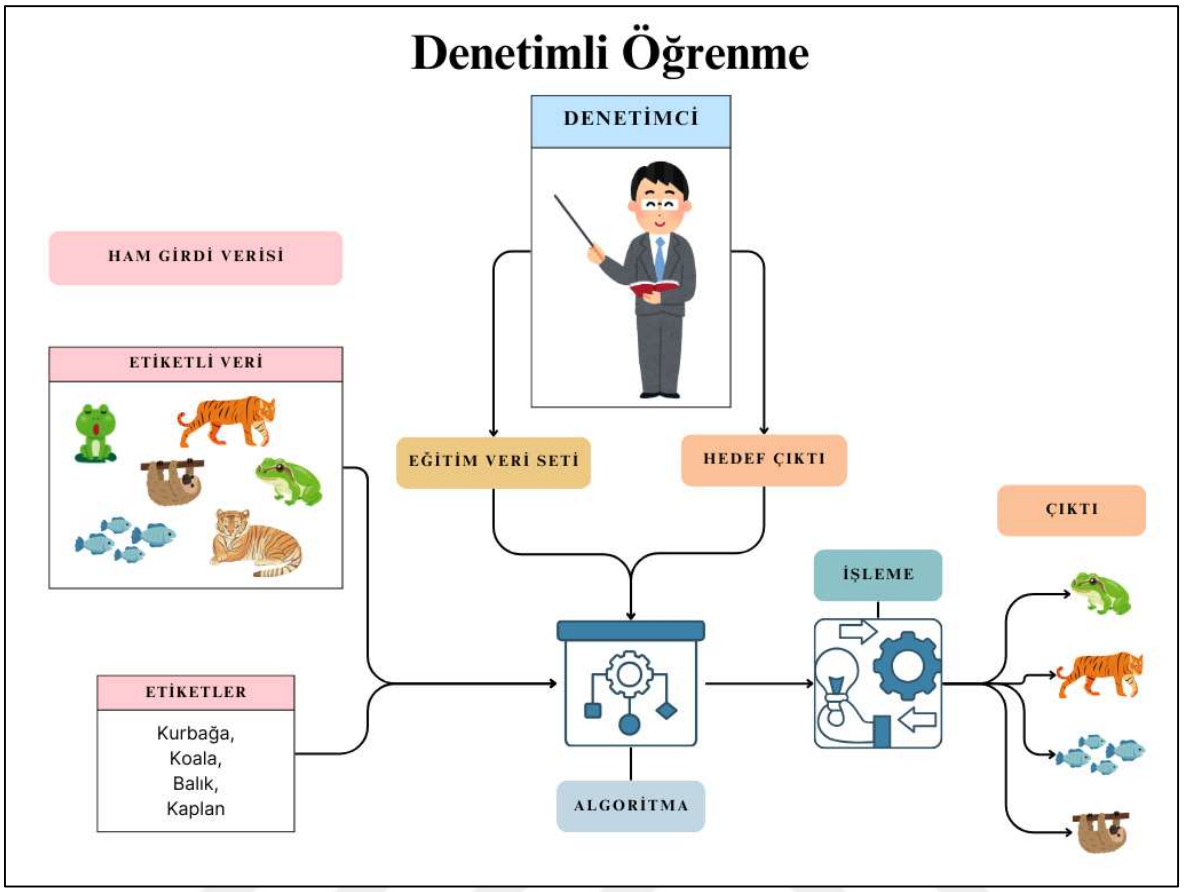
Denetimsiz öğrenme öğrenme algoritmalarının adaptif yapısı, özellik kümesini dinamik bir şekilde geliştirmek için kullanılmış ve hiperparametre optimizasyonuyla desteklenmiştir.

3.3. Yöntem Seçimi

Bu tez çalışmasında, veri işleme ve model performansını iyileştirme süreçlerini optimize etmek amacıyla, kademeli ve iteratif bir yöntem uygulanmıştır. Yöntemlerin seçimi, her bir algoritmanın güçlü yönleri ve kullanım amaçlarına uygunluğu göz önünde bulundurularak yapılmıştır. Aşağıda, kullanılan yöntemler, sıralı uygulamaları ve seçim gerekçeleri detaylandırılmıştır.

Denetimli öğrenme

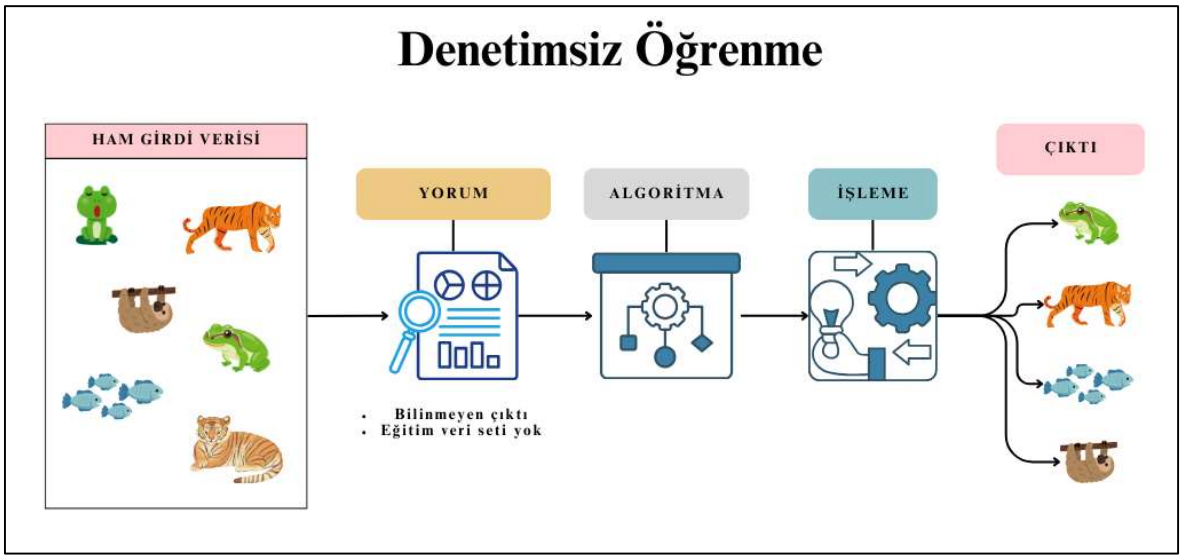
Şekil 3.4, denetimli öğrenme iş akışını görselleştirirken, modelin etiketli veri kullanarak girişleri doğru sonuçlara eşlemeyi öğrenme sürecini detaylandırmaktadır. Etiketli veriler, modelin eğitilmesi sırasında rehberlik sağlamakta ve eğitim sonrası süreçlerde tahmin doğruluğunu artırmaktadır. Ancak, denetimli öğrenmenin aksine, denetimsiz öğrenmede etiket bilgisi kullanılmadan veri kümelerinin analizi gerçekleştirilir.



Şekil 3.4. Denetimli öğrenmenin çalışma prensibi

Denetimsiz öğrenme

Şekil 3.5'te, denetimsiz öğrenmenin çalışma prensibi görülmektedir. Bu süreçte algoritmalar, verilerdeki kalıpları keşfetmek ve girdileri mantıklı gruplara ayırmak için çalışır. Örneğin, hayvanların farklı özelliklerine göre sınıflandırılması gibi karmaşık veri yapılarını analiz etmek, denetimsiz öğrenmenin sağladığı bir avantajdır.



Şekil 3.5. Denetimsiz öğrenmenin çalışma prensibi

3.3.1. Denetimsiz öğrenmenin kullanımı

Denetimsiz öğrenme, hedef değişkenin (etiket) işleme dahil edilmediği bir öğrenme süreci sunar. Bu çalışmada, verilerin iteratif olarak zenginleştirilmesi amacıyla denetimsiz öğrenme yöntemlerinin kullanılmasının temel nedeni, ezberlemeyi önlemek ve modeli yalnızca verilere dayalı olarak eğitmektir. Hedef değişkenin modele dahil edilmesi durumunda, modelin aşırı öğrenme riski artacak ve genelleştirme kabiliyeti zayıflayacaktır.

Bu tezde denetimsiz öğrenme, veri setlerinden özellik çıkarımı amacıyla kullanılmıştır. Denetimli öğrenme ise sınıflandırma aşamasında meta modelin eğitimi için devreye girmiştir. Bu iki öğrenme paradigmasının iş akışları sırasıyla Şekil 3.4 ve Şekil 3.5'te gösterilmiştir.

Şekil 3.5'te gösterilen denetimsiz öğrenme süreci, ham veri setlerinden etiketlere ihtiyaç duymadan özellikler çıkarılmasını sağlar. K-ortalamalar algoritması, veri noktalarını kümelere ayırarak veri setinin gizli yapısını ortaya çıkarır. Denetimsiz öğrenme ayrıca, veri setinde mevcut olan gizli yapıları ve kalıpları ortaya çıkararak, özellikle etiketlenmemiş veri setleri üzerinde etkili sonuçlar sunar. Bu bağlamda, veri setinden çıkarılan ek özelliklerin, model performansını artıracak şekilde işlenmesine olanak sağlar.

3.3.2. K-ortalamalar algoritması

K-ortalamalar kümeleme algoritması [25], denetimsiz öğrenmenin temel yaklaşımlarından biri olarak önerilen metodolojinin merkezinde yer almaktadır. Bu algoritma, veri noktalarını belirli bir kümeleme kriterine göre gruplandırarak veri setinde yer alan içsel kalıpları ve yapıları ortaya çıkarır. Çalışmada, k-ortalamalar kümelemesinin tercih edilmesinin arkasındaki sebepler, algoritmanın sunduğu bir dizi avantaj ile ilişkilidir.

K-ortalamalar kümelemesinin en önemli avantajlarından biri, karmaşık veri kümelerindeki gizli yapıları keşfetme yeteneğidir. Veri noktalarını kümelere ayırarak, özellikler arasındaki benzerlikler ve farklılıklar anlamlı bir şekilde belirlenir. Bu süreç, hem modelin daha iyi bir şekilde eğitilmesine olanak tanır hem de veri setinin anlaşılmasını kolaylaştırır. Özellikle yüksek boyutlu veri setlerinde, k-ortalamalar boyut azaltma açısından kritik bir rol oynar [26]. Kümelerden türetilen yeni özellikler, gereksiz veya düşük önem seviyesine sahip olanların filtrelenmesini sağlayarak daha kompakt ve anlamlı bir veri seti oluşturur.

Yinelemeli özellik çıkarımı bağlamında, k-ortalamalar algoritmasının verimliliği dikkat çekicidir. Algoritmanın hızlı çalışması, özellik kümesinin kısa sürede değerlendirilmesini ve iyileştirilmesini sağlar. Bu hız, önerilen metodolojideki iteratif süreçte algoritmanın birçok kez uygulanmasını mümkün kılar. Aynı zamanda algoritmanın esnek ve uyarlanabilir yapısı, her iterasyonda veri setindeki desenlerin değişmesine göre kendini güncelleyerek, farklı kümeleme senaryolarını keşfetme imkânı sunar.

Denetlenmeyen bir teknik olan k-ortalamaların bir diğer önemli avantajı, hedef değişkenlerle doğrudan bir bağımlılık oluşturmamasıdır. Denetimli öğrenme tekniklerinde, hedef değişkenlerin eğitim sürecine istemeden karışma riski mevcuttur. Ancak k-ortalamalar gibi denetlenmeyen yöntemler, yalnızca giriş verileri üzerinde çalışarak bu türden bir "hedef sızıntısı" riskini ortadan kaldırır. Bu durum, modelin genelleme kabiliyetini artırır ve daha güvenilir bir performans değerlendirmesi sağlar.

K-ortalamalar algoritmasının bir başka katkısı da yorumlanabilirliğe yöneliktir. Kümeleme sürecinden elde edilen sonuçlar, veri setindeki örüntülerin anlaşılmasını kolaylaştırır. Örneğin, kümelerden türetilen yeni özellikler, veri noktalarının daha iyi tanımlanmasına ve modelin doğruluğunun artırılmasına katkıda bulunur. Bu bağlamda, k-ortalamalar

algoritması yalnızca meta-model performansını artırmakla kalmaz, aynı zamanda veri setindeki karmaşık örüntüleri yakalayarak çalışmanın genel etkinliğini güçlendirir.

Sonuç olarak, önerilen metodolojide k-ortalamaların etkin bir şekilde kullanımı, özellik çıkarımı sürecinin hem zaman açısından verimli hem de sonuç açısından etkili olmasını sağlamaktadır. Bu yaklaşım, algoritmanın temel avantajlarını yinelemeli iş akışına entegre ederek modelin doğruluğunu ve genelleme yeteneğini artırmayı hedefler.

3.3.3. Random Forest algoritması

Random Forest [21], güçlü bir tahmin performansı sunması ve heterojen veri setlerinde etkili çalışması nedeniyle modelin son değerlendirici olarak seçilmiştir. Random Forest'ın tercih edilmesinin başlıca nedenleri şunlardır:

- Doğruluk ve Genelleştirme: Random Forest, birden fazla karar ağacının çıktılarını birleştirerek daha yüksek doğruluk ve genelleştirme kabiliyeti sunar. Veri setinin iteratif olarak zenginleştirilmesiyle elde edilen yeni özelliklerin değerlendirilmesinde bu algoritma oldukça etkilidir.
- Kararlılık ve Esneklik: Random Forest, farklı veri türlerini iyi bir şekilde işleyebilir ve modelin stabilizasyonuna katkıda bulunur. Özellikle, K-ortalamalar ile çıkarılan kümeler gibi ek özelliklerin anlamlı bir şekilde değerlendirilmesi için uygundur.
- Aşırı Öğrenmeye Dayanıklılık: Karar ağaçlarının bir araya gelmesiyle oluşan bu yöntem, aşırı öğrenme riskini azaltır ve veri setindeki gürültüye karşı dayanıklıdır [27].

Bu çalışmada Random Forest, K-ortalamalar ile zenginleştirilmiş veri seti üzerinde eğitilmiş ve modelin tahmin performansını değerlendirmek için kullanılmıştır. Ayrıca, iterasyonlar arasında sonuçların karşılaştırılması için bir referans modeli olarak görev yapmıştır.

3.3.4. Optuna

Optuna tarafından optimize edilmiş bir Random Forest modeli, önerilen yöntemin etkinliğini artırmak ve kıyaslama yapabilmek için kullanılmıştır [28]. Optuna, hiperparametre optimizasyonunda kullanılan adaptif bir araçtır.

Optuna;

- Hızlı ve Verimli Arama: Optuna, Bayesian optimizasyon teknikleriyle hiperparametre arama sürecini hızlandırır ve gereksiz kombinasyonları elimine eder.
- Dinamik Yapı: Optuna, iterasyonlar arasında dinamik olarak öğrenme sağlayarak, modelin farklı parametre kombinasyonlarında nasıl performans gösterdiğini anlayabilir.
- En İyi Performansın Sağlanması: Optuna'nın optimize ettiği Random Forest modeli, diğer yöntemlerle elde edilen sonuçlarla karşılaştırılarak, önerilen yöntemin etkisini daha net bir şekilde ortaya konmasına yardımcı olabilir.

Optuna, bu tezde Random Forest'ın hiperparametrelerini optimize ederek, optimize edilmiş veri seti üzerinde en iyi sonuçları elde etmeye odaklanmıştır.

3.4. Model ve Değerlendirme Parametreleri

Bu çalışmada, modelin ve değerlendirme süreçlerinin parametrelerinin belirlenmesi, önerilen yöntemin başarısını optimize etmek ve karşılaştırmalı analizler yapmak amacıyla dikkatle planlanmıştır. K-ortalamalar algoritması ve meta modelin iteratif değerlendirme süreçleri hem özellik çıkarımı hem de model doğruluğunun iyileştirilmesi için kritik bir rol oynamıştır.

K-ortalamalar algoritmasının başarısı, küme sayısı (k) gibi hiperparametrelerin doğru seçimine bağlıdır. Bu çalışmada, veri setindeki gizli kalıpları en iyi şekilde temsil eden özellikleri çıkarmak için her iterasyonda farklı küme sayıları test edilmiştir. Test edilen küme sayıları geniş bir aralıkta belirlenmiş, her iterasyonda en iyi sonucu veren küme sayısı seçilmiştir.

Her iterasyon için $F_1 - \text{Skoru}$, modelin başarısını ölçmek için bir temel kriter olarak kullanılmıştır. $F_1 - \text{Skorları}$, küme sayılarının model üzerindeki etkilerini değerlendirmede kritik bir metrik sunmuş ve en yüksek skoru sağlayan parametreler sonraki iterasyonlarda kullanılmıştır. Bu süreç, her iterasyonun etkili bir şekilde optimize edilmesini ve modelin iteratif öğrenme sürecinde tutarlı bir şekilde iyileşmesini sağlamıştır.

Bununla birlikte, kaç iterasyon yapılacağı önceden belirlenmiştir. Bu sınır, işlem

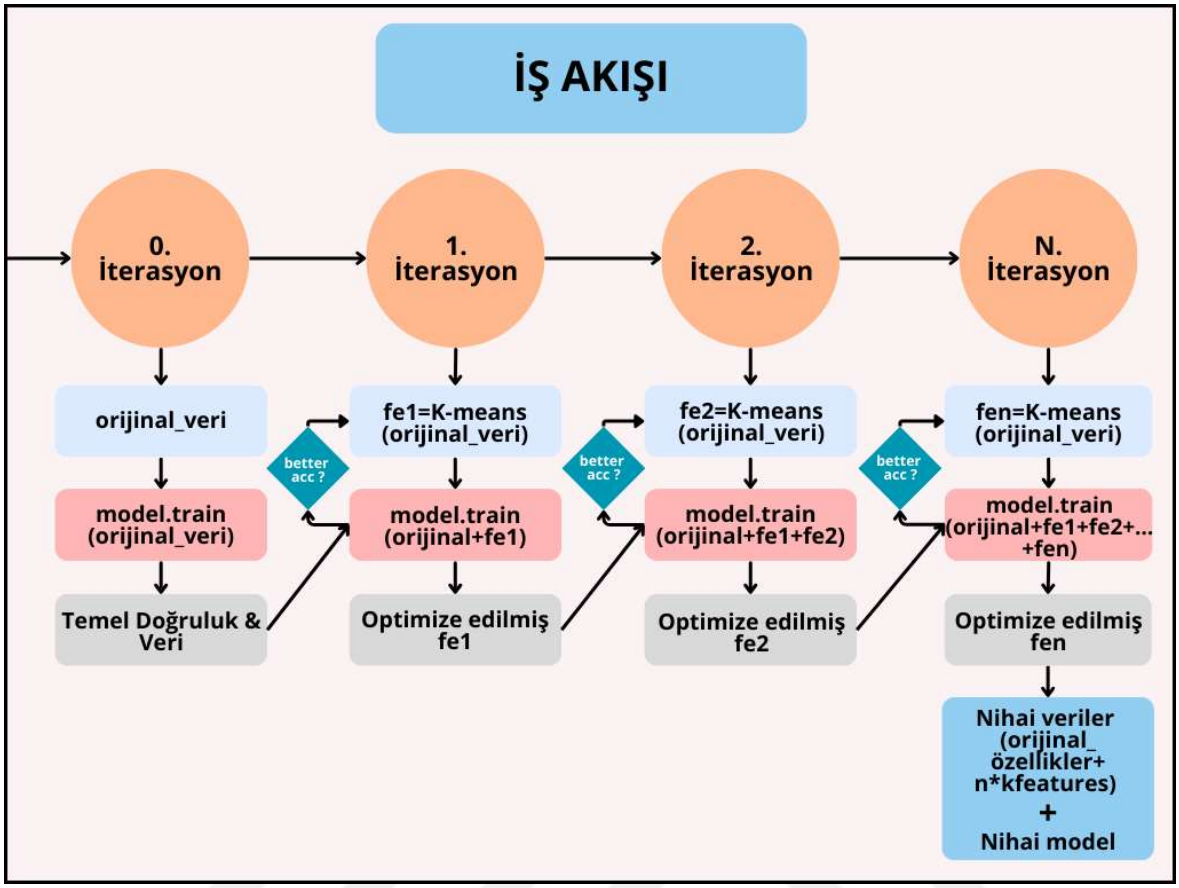
maliyetlerini kontrol altında tutmak ve aşırı hesaplama süresi gerektiren durumları önlemek için belirlenmiştir. İterasyonların sınırlandırılması, yöntemin pratik uygulanabilirliğini artırmış ve yöntem, belirlenen iterasyon sayısına ulaşıldığında veya model performansında daha fazla bir iyileşme gözlemlenmediğinde durdurulmuştur.

3.5. Model Eğitimi ve Yinelemeli Özellik Çıkarımı

Orijinal veri setiyle eğitilen temel modelin ardından, yinelemeli bir süreçle k-ortalamalar kümelemesi kullanılarak yeni özellikler çıkarılmış ve meta model sürekli olarak iyileştirilmiştir.

3.5.1. Önerilen yinelemeli özellik çıkarma tabanlı optimizasyon yöntemi

Önerilen metodoloji, k-ortalamalar kümelemesinin yinelemeli bir iş akışı ile uygulanmasını temel alır. K-ortalamalar kümelemesi, yinelemeli bir iş akışıyla uygulanmış ve her iterasyonda veri setine yeni özellikler eklenmiştir. Bu süreçte, her iterasyonda veri setinden yeni özellikler çıkarılarak, meta modelin sürekli olarak iyileştirilmesi hedeflenmiştir. Şekil 3.6, bu iş akışını görsel olarak özetlemektedir ve sürecin dinamik yapısını ortaya koymaktadır.



Şekil 3.6. Tekrarlı eğitimlere dayalı özellik çıkarımının iş akışı

Veri hazırlığı ve başlangıç

Yöntemin ilk aşamasında, veri setlerinin seçimi ve ön işlenmesi gerçekleştirilmektedir. Veri setlerinde olası eksik değerlerin doldurulması, normalizasyon ve standartlaştırma işlemleri, veri setinin k-ortalamalar algoritması için uygun hale getirilmesi sağlanmıştır. Dolayısıyla, veri analizi ve veri işlenmesi için eksik değerler düşürülmüş, normalizasyon yöntemi olarak MinMaxScaler kullanılmıştır [29].

Temel model eğitimi

Başlangıçta, orijinal veri setiyle bir temel model eğitilmiş ve bu modelin performans metrikleri hesaplanmıştır. Bu metrikler kesinlik, duyarlılık ve f1-skoru olmak üzere probleme göre belirlenir. Bu aşama, önerilen metodolojinin etkinliğini karşılaştırmak için bir referans noktası sağlar.

İlk özellik çıkarma

İlk iterasyonda, k-ortalamalar algoritması veri setine uygulanmıştır. Minimumdan maksimum (1-200 manuel olarak kullanıcı tarafından belirlenmiştir) değere kadar değişen küme sayıları kullanılarak veri setindeki potansiyel desenler analiz edilmiş ve her kümeye karşılık gelen Şekil 3.1'de "fe1" şeklinde ifade edilmiş özellikler veri setine eklenmiştir.

Yeni özelliklerin entegrasyonu

Kümeleme sürecinden elde edilen yeni özellikler, orijinal veri setine entegre edilmiş ve model tekrar eğitilmiştir. Bu aşamada modelin performansı tekrar değerlendirilmiştir. Eğer model doğruluğu artırılmışsa, süreç bir sonraki iterasyonla planlı sonlanana kadar devam etmektedir.

Hiperparametre optimizasyonu

GridSearch yöntemi ile her iterasyon için en uygun kümeleme parametresi belirlenir. Bu, k-ortalamaların her iterasyonda dinamik bir şekilde ayarlanmasını sağlar ve algoritmanın veri setindeki değişimlere uyum sağlamasına olanak tanır. Bu süreç, özellik kümesinin dinamik olarak geliştirilmesini ve nihai modelin doğruluğunun optimize edilmesini sağlamıştır.

Yinelemeli iterasyon ve sürecin sonlandırılması

Süreç, her iterasyonda yeni özellikler eklenerek devam eder. Ancak model performansında bir iyileşme gözlenmediğinde, süreç sonlandırılır ve elde edilen son model nihai çıktı olarak kaydedilmiştir.

3.6. Sonuçların Kıyaslanması ve Nihai Model Seçimi

Nihai model, klasik yöntemler ve OPTUNA gibi modern optimizasyon teknikleriyle karşılaştırılmıştır. Bölüm 4.2'de sunulan sonuçlar, önerilen yöntemin etkinliğini ve avantajlarını ortaya koymuştur, hiperparametre optimizasyonu ve özellik çıkarımı açısından avantajlar gözlemlenmiş ve değerlendirilmiştir.

Bu yinelemeli iş akışı, yalnızca modelin doğruluğunu artırmayı değil, aynı zamanda veri setindeki gizli desenleri daha derinlemesine anlamayı hedeflemektedir. Önerilen metodoloji, hem veri biliminde hem de makine öğreniminde özellik çıkarımı sürecine yeni bir bakış açısı sunmaktadır.

Bu metodoloji, k-ortalamalar algoritmasının dinamik kullanımı ve hiperparametre optimizasyonunu birleştirerek özellik çıkarımı sürecini sistematik hale getirmiştir. Yinelemeli iş akışı, veri setindeki gizli desenleri keşfetmiş ve nihai modelin doğruluğunu optimize etmiştir. Metodoloji, Bölüm 4.3'te sunulduğu üzere klasik ve modern yöntemlerle kıyaslandığında üstünlük sağlamış ve model performansını artırmada etkili bir yaklaşım paylaşmıştır.

4. BULGULAR VE DEĞERLENDİRME

Önerilen metodolojinin etkinliğini değerlendirmek için çeşitli kriterler ve karşılaştırmalı analizler kullanılmıştır. Çalışmanın temel amacı, model performansının farklı veri kümelerinde iteratif özellik çıkarımı yoluyla nasıl iyileştirilebileceğini göstermek olmuştur. Değerlendirme süreci, hem $\mathrm{F}_1-\mathrm{Skoru}$ gibi ölçütlere hem de yöntemimizin her iterasyondaki katkısının detaylı incelenmesine odaklanmıştır. Aşağıda, değerlendirme kriterleri, analiz edilen veri kümeleri ve sonuçlara dair ayrıntılar sunulmuştur.

4.1. Değerlendirme Kriterleri

Önerilen metodolojinin başarısını ölçmek için aşağıdaki kriterler dikkate alınmıştır.
$\mathrm{F}_1-\mathrm{Skoru}$, Modelin tahmin doğruluğunu ve dengeli performansını değerlendiren bir ölçüttür. Hem kesinlik (precision) hem de duyarlılık (recall) değerlerini dikkate alarak genel tahmin başarısını yansıtır. $\mathrm{F}_1-\mathrm{Skoru}$, her iterasyon sonrası modeldeki iyileştirmeleri net bir şekilde ortaya koymak için kullanılmıştır.

$$Doğruluk = \frac{TN + TP}{TN + TP + FN + FP} \tag{4.1}$$

$$Kesinlik = \frac{TP}{TP + FP} \tag{4.2}$$

$$Duyarlılık = \frac{TP}{TP + FN} \tag{4.3}$$

$$\mathrm{F}_1 - Skoru = 2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{4.4}$$

İterasyonların etkisi, yinelemeli özellik çıkarımının her aşamada model performansına yaptığı katkılar detaylı olarak analiz edilmiştir. Bu analiz, iterasyon bazında model performansında meydana gelen artışları gözlemlemek ve yöntemimizin etkinliğini değerlendirmek için önemlidir.

En iyi kümeleme parametresi, her iterasyonda kullanılan k-ortalamalar kümeleme algoritmasının en iyi hiperparametresi optimize edilmiştir. Bu parametre, her aşamada veri setindeki yeni özelliklerin çıkarılmasını etkili hale getirmiştir. Devamlı iterasyonlar süresince arka arkaya gerçekleştirilen üç eğitim boyunca (“patience” değeri kullanıcı

tarafından değiştirilebilir) iyileşme gerçekleşmezse (tercih edilen metrik bazında, bu çalışmada F-1 Skoru kullanılmıştır) eğitim sonlanır. Tüm iterasyonlar süresince en iyi metrik değerle ardışık iterasyonlar kayıt edilir. Ayrıca bu üç eğitim süresince iyileşmeme kriteri tetiklenmezse eğitimi sonlu hale getirmek için bir sert limit belirlenir (bu çalışmada 20 kullanılmıştır).

Veri kümesi çeşitliliği, Wisconsin Meme Kanseri Tespiti ve Diyabet veri setleri gibi farklı özelliklere sahip veri kümeleri üzerinde yapılan analizlerle metodolojinin genelleştirilebilirliği değerlendirilmiştir. Bu, algoritmanın farklı veri yapılarına uyarlanabilirliğini ve genel uygulanabilirliğini test etmek için kritik bir unsurdur.

4.2. Önerilen Yöntemin Mevcut Yöntemlerle Kıyaslama Senaryoları

Önerilen yöntemin etkinliğini değerlendirmek için klasik optimizasyon yöntemleri, yalnızca Optuna, önerilen yöntem ve bu yöntemin Optuna ile birleştirilmiş hali üzerinde çeşitli kıyaslama çalışmaları yapılmıştır. Bu süreçte Diyabet Veri Seti ve Wisconsin Meme Kanseri Tespiti Veri Seti kullanılarak performans değerlendirilmiş, elde edilen sonuçlar her bir veri senaryosunun güçlü ve zayıf yönlerini ortaya koymaktadır.

4.2.1. Klasik optimizasyon yöntemleriyle elde edilen sonuçlar

Klasik optimizasyon yöntemleri, hiperparametre optimizasyonunda yaygın olarak kullanılan geleneksel yaklaşımlardır. Bu yöntemler arasında Grid Search ve Random Search öne çıkmaktadır. Grid Search, hiperparametre arama alanındaki tüm kombinasyonları değerlendirirken, Random Search yalnızca rastgele bir alt küme üzerinde çalışır.

Hiç optimizasyon yapılmamış modeller $F_1 - \text{Skoru}$ açısından Diyabet veri setinde %67,96, Wisconsin Meme Kanseri Tespiti veri setinde %91,14 sonuçlarını almıştır.

Diyabet veri seti

Diyabet veri setinde Grid Search, $F_1 - \text{Skorunu}$ %68,5'e kadar çıkarabilmiştir. Ancak, tüm kombinasyonları deneme gerekliliği, işlem süresini önemli ölçüde artırmıştır. Random Search, işlem süresini düşürmüş olsa da, F-1 Skorunu maksimum %68'e çıkarabilmiştir.

Wisconsin meme kanseri tespiti veri seti

Bu veri setinde Grid Search, $F_1 - Skoru$'nu %92,3 ile aynı seviyede kalmıştır. Random Search de daha düşük maliyetle çalışmasına rağmen benzer doğruluk değerlerini korumuştur, $F_1 - Skoru$ 0,92,6'te kalmıştır.

Sonuç olarak, klasik yöntemler küçük veri setlerinde kabul edilebilir sonuçlar verse de büyük veri setleri için yüksek maliyetli ve verimsiz hale gelmektedir. Ayrıca, veri setindeki gizli kalıpları keşfetme konusunda sınırlı bir yetenek sergilerler.

4.2.2. Optuna ile elde edilen sonuçlar

Optuna, adaptif ve dinamik optimizasyon yetenekleriyle geleneksel yöntemlere göre üstün bir performans sergiler. Bayesian optimizasyon tabanlı bir strateji kullanarak, gereksiz kombinasyonları elemeyi ve hiperparametre arama sürecini hızlandırmayı başarmaktadır.

Diyabet veri seti

Optuna ile yapılan çalışmalar $F_1 - Skoru$'nu başlangıç değerinden %68,6'ya yükseltmiştir. Optuna'nın dinamik yapısı, hiperparametre aramasında Grid Search'e göre %30 daha düşük bir işlem maliyeti sağlamıştır. Özellikle, optimal kümeleme parametrelerinin belirlenmesinde büyük bir esneklik sağlamış ve modeli hem doğruluk hem de genelleştirme açısından optimize etmiştir.

Wisconsin meme kanseri tespiti veri seti

Optuna, bu veri setinde $F_1 - Skorunu$ %90'una ulaşmıştır. Bunun sebebi, yaptığı budama işlemleri sırasında lokal minimuma takılmıştır.

Yalnızca Optuna kullanıldığında elde edilen sonuçlar, modern optimizasyon tekniklerinin etkinliğini kanıtlamakla birlikte, veri setlerindeki gizli özellikleri otomatik olarak çıkaramaması, Optuna'nın eksiklikleri olarak değerlendirilmiştir.

4.2.3. Önerilen yöntem ile elde edilen sonuçlar

Önerilen yöntem, denetimsiz öğrenme tabanlı dinamik özellik çıkarma süreci ile hiperparametre optimizasyonunu birleştirerek geleneksel yöntemlerin ötesine geçmektedir. Özellikle veri setlerindeki gizli kalıpları ortaya çıkararak özellik kümesini iyileştirme yeteneği, yöntemin en büyük avantajıdır.

Diyabet veri seti

Diyabet veri kümesi üzerinde yapılan analizler, önerilen metodolojinin iteratif özellik çıkarımı ile model performansını önemli ölçüde artırabildiğini göstermektedir. Çizelge 4.1'te, iterasyonlar boyunca $F_1 - \text{Skoru}$ ve en iyi kümeleme parametreleri yer almaktadır. Ayrıca, Çizelge 4.2 ve Çizelge 4.3'te 8 iterasyon boyunca elde edilen kesinlik ve duyarlılık skoru sonuçları ve en iyi kümeleme parametresi yer almaktadır.

Çizelge 4.1. Diyabet veri kümesi üzerinde önerilen yöntemin elde ettiği F1-Skoru sonuçları

İterasyon	$F_1 - \text{Skoru}$	En İyi Küme Parametresi (1-200)
0	0,6796	-
1	0,7636	181
2	0,7850	3
3	0,7778	18
4	0,7706	49
5	0,7850	42
6	0,7778	194
7	0,8000	184
8	0,7963	49

Başlangıçta %67,96 olan $F_1 - \text{Skoru}$, sekiz iterasyon sonucunda %79,63'e ulaşmıştır. Bu, %11,67'luk bir iyileşmeye karşılık gelmektedir. Özellikle 7. iterasyonda büyük bir sıçrama gözlemlenmiştir. Bu durum, metodolojinin iterasyon bazlı özellik çıkarımında başarılı bir şekilde iyileştirme sağladığını göstermektedir. Çizelge 4.1'de gösterilen özelliklerin iteratif bir şekilde eklenmesi, modelin tahmin doğruluğunu artırırken, aynı zamanda daha sağlam

ve genelleştirilebilir bir yapıya ulaşmasına olanak tanımıştır.

Çizelge 4.2. Diyabet veri kümesi üzerinde önerilen yöntemin elde ettiği kesinlik sonuçları

İterasyon	Kesinlik	En İyi Küme Parametresi (1-200)
0	0,7142	-
1	0,75	181
2	0,7925	3
3	0,7778	18
4	0,7636	49
5	0,7925	42
6	0,7778	194
7	0,7857	184
8	0,7963	49

Çizelge 4.2'de Diyabet veri kümesi üzerinde önerilen yöntemin elde ettiği kesinlik sonuçları sunulmuştur. Başlangıçta %71,42 olan kesinlik skoru, sekiz iterasyon sonucunda %79,63'e ulaşmıştır. Bu, %8,21'lik bir iyileşmeye karşılık gelmektedir. Bu durum, metodolojinin iterasyon bazlı özellik çıkarımında başarılı bir şekilde iyileştirme sağladığını göstermektedir. Çizelge 4.2'de, $F_1 - \text{Skoru}$ sonuçlarında olduğu gibi kesinlik skoru sonuçlarında da özelliklerin iteratif eklenmesi, modelin tahmin doğruluğunu artırırken, aynı zamanda daha sağlam ve genelleştirilebilir bir yapıya ulaşmasına olanak tanımıştır.

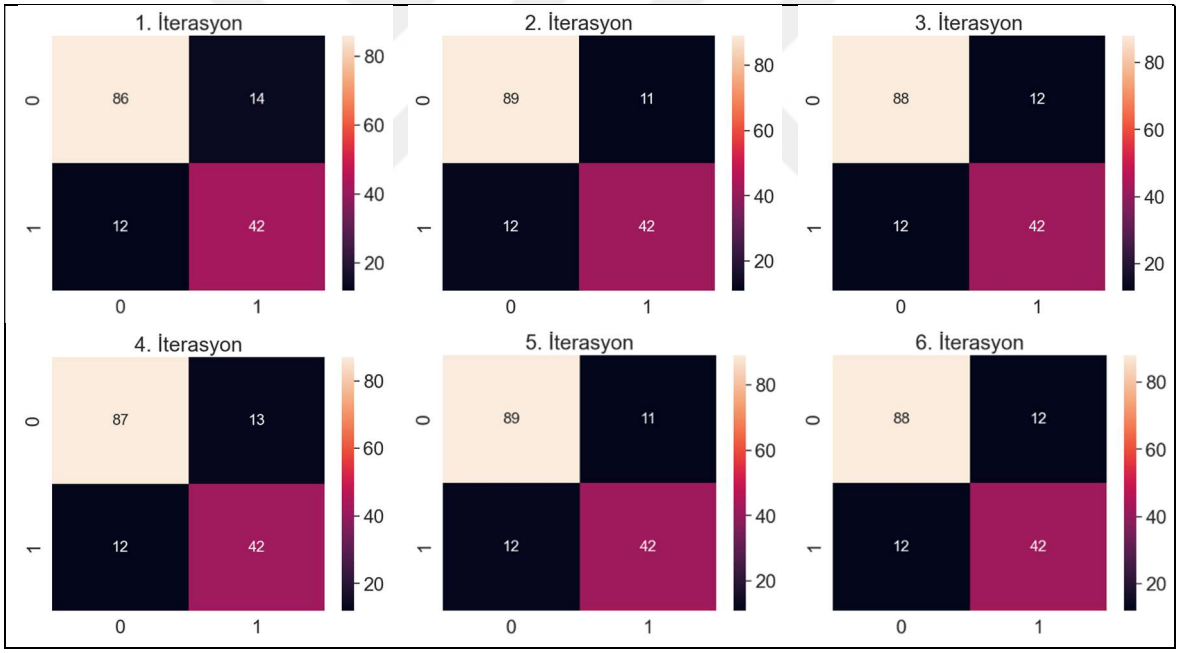
Çizelge 4.3. Diyabet veri kümesi üzerinde önerilen yöntemin elde ettiği duyarlılık sonuçları

İterasyon	Duyarlılık	En İyi Küme Parametresi (1-200)
0	0,6481	-
1	0,7778	181
2	0,7778	3
3	0,7778	18
4	0,7778	49
5	0,7778	42
6	0,7778	194
7	0,8148	184
8	0,7963	49

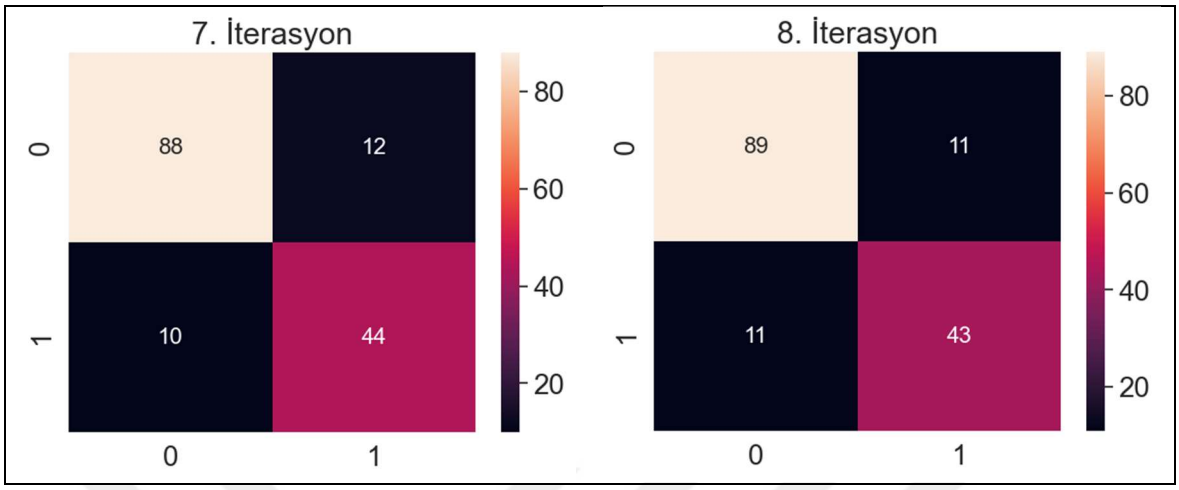
Başlangıçta %64,81 olan duyarlılık skoru, sekiz iterasyon sonucunda %79,63'e ulaşmıştır. Bu, %14,82'lik bir iyileşmeye karşılık gelmektedir. Bu durum, metodolojinin iterasyon bazlı özellik çıkarımında başarılı bir şekilde iyileştirme sağladığını göstermektedir. Çizelge 4.3'te, $F_1 -$ Skoru ve kesinlik sonuçlarında olduğu gibi duyarlılık skoru sonuçlarında da özelliklerin iteratif eklenmesi, modelin tahmin doğruluğunu artırırken, aynı zamanda daha sağlam ve genelleştirilebilir bir yapıya ulaşmasına olanak tanımıştır.

Çizelge 4.4'te önerilen yöntemin Diyabet veri seti kullanılarak her iterasyonda elde ettiği sonuçlarının karmaşıklık matrisleri verilmiştir. Karmaşıklık matrislere bakıldığında, TP ve TN değerlerinin son iterasyonlara doğru yükseldiği gözlemlenmiştir.

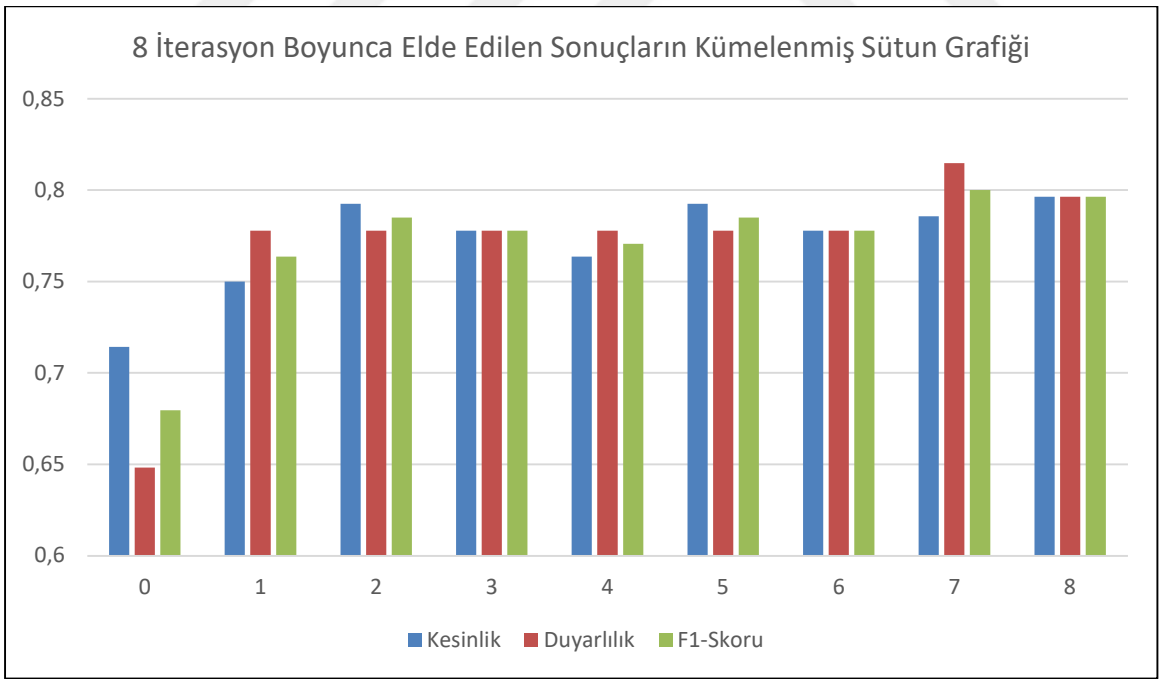
Çizelge 4.4. Diyabet veri seti üzerinde iteratif optimizasyon eğitimi sonuçlarının karmaşıklık matrisleri



Çizelge 4.4. (devam) Diyabet veri seti üzerinde iteratif optimizasyon eğitimi sonuçlarının karmaşıklık matrisleri



Şekil 4.1'te Diyabet veri seti üzerinde yapılan eğitim sonuçlarının kümelenmiş grafiği sunulmuştur. Bu değerlendirme testleri incelendiğinde, sonuçlar ilk 0-2 iterasyon boyunca yükselme göstermiştir.



Şekil 4.1. Diyabet veri seti üzerinde gerçekleştirilen eğitim sonuçlarının kümelenmiş sütun grafiği

Daha sonrasında, küçük bir düşüş gözlemlenmiştir. Ancak bir önceki en yüksek değeri lokal maksimum olabileceği ihtimali ile iterasyon devam ettirilmiştir, sonuçlarda da görüldüğü

üzere ilerleyen iterasyonlarda daha yüksek ve dolayısı ile global maksimum kabul edilebilecek sonuçlar alınmıştır.

Wisconsin meme kanseri tespiti veri seti

Wisconsin Meme Kanseri veri kümesinde yapılan analizlerde başlangıç $F_1 - \text{Skoru}$ %91,14 olarak belirlenmiştir. Beşinci iterasyonda bu değer en yüksek seviyesine ulaşarak %97,73 olmuştur. Çizelge 4.5, iterasyonlar boyunca, $F_1 - \text{Skoru}$ ve en iyi küme parametrelerini göstermektedir.

Başlangıçtaki yüksek $F_1 - \text{Skoru}$ nedeniyle iyileştirme aralığı sınırlı olmuştur. Buna rağmen, model performansında %6,59 oranında bir artış sağlanmıştır. Bu durum, veri setinin başlangıç özelliklerinin oldukça güçlü olduğunu ve ileri iterasyonların katkısının sınırlı kaldığını göstermektedir. Çizelge 4.5'teki performans düşüşler, algoritmanın stabilizasyon süreçlerinin optimize edilmesi gerektiğini ortaya koymaktadır.

Çizelge 4.5. Wisconsin veri kümesi üzerinde önerilen yöntemin elde ettiği F1-Skoru sonuçları

İterasyon	$F_1 - \text{Skoru}$	En İyi Küme Parametresi
0	0,9114	-
1	0,9425	87
2	0,9535	26
3	0,9655	11
4	0,9655	151
5	0,9773	21

Çizelge 4.6'da zaten yüksek başlayan kesinlik skoru ilerleyen iterasyonlarda 1.0'a ulaşmıştır. Bunun sebebi veri setinde buluna örnek sayısının az olmasıdır. Bununla birlikte Çizelge 4.5'te bulunan F1-Skorları karşılaştırıldığında ikinci iterasyondan itibaren kesinlik çerçevesinde bir katkı olmamıştır.

Çizelge 4.6. Wisconsin veri kümesi üzerinde önerilen yöntemin elde ettiği kesinlik sonuçları

İterasyon	Kesinlik	En İyi Küme Parametresi
0	0,9230	-
1	0,9762	87
2	1,0	26
3	1,0	11
4	1,0	151
5	1,0	21

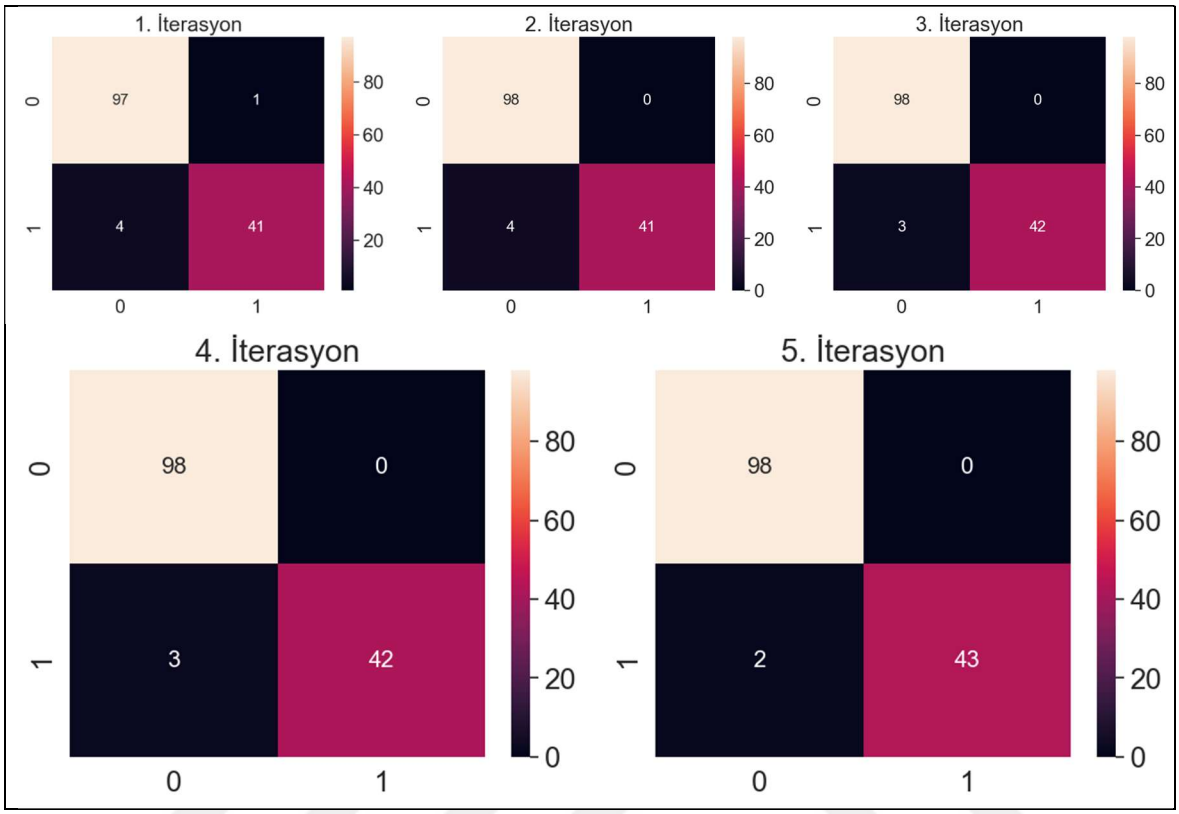
Çizelge 4.7'te verilen duyarlılık skorları kesinlik 1.0'a ulaştıktan sonra eğitimin F1-Skorunun gelişmesine katkı sağlayan metriktir. Duyarlılıkta gerçekleşen iyileşmeler, eğitimin bir sonraki iterasyona geçmesine olanak tanımıştır.

Çizelge 4.7. Wisconsin veri kümesi üzerinde önerilen yöntemin elde ettiği duyarlılık sonuçları

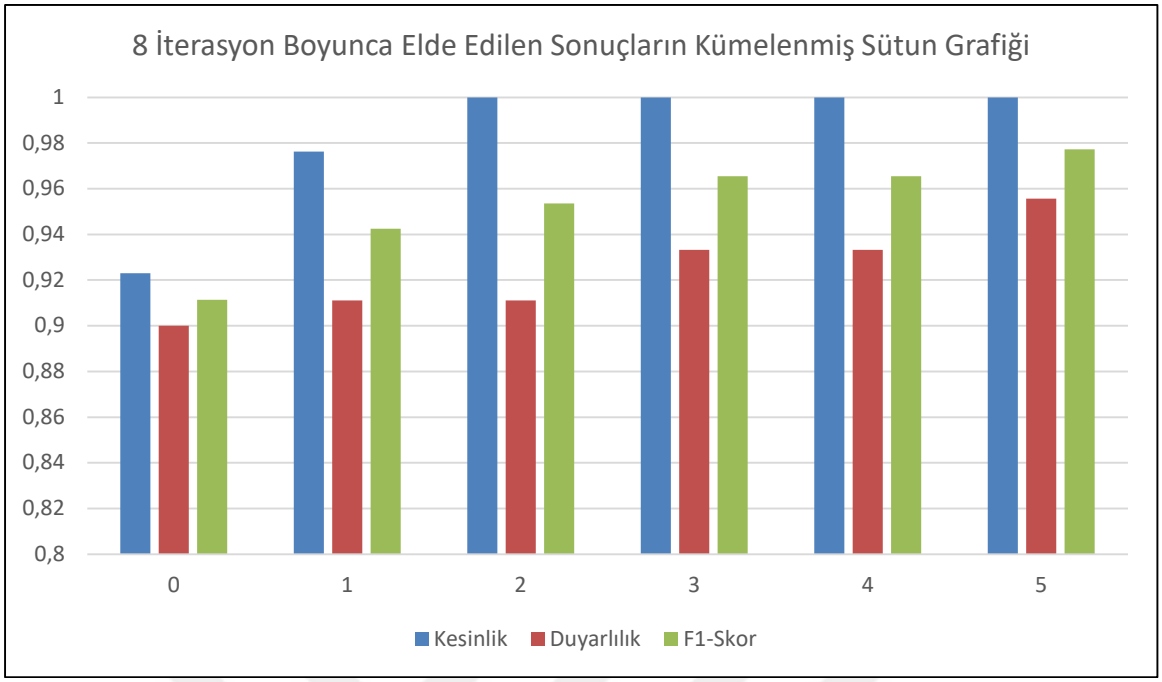
İterasyon	Duyarlılık	En İyi Küme Parametresi
0	0,9	-
1	0,9111	87
2	0,9111	26
3	0,9333	11
4	0,9333	151
5	0,9556	21

Çizelge 4.8'de önerilen yöntemin Wisconsin Meme Kanseri veri seti kullanılarak her iterasyonda elde ettiği sonuçlarının karmaşıklık matrisleri verilmiştir. Karmaşıklık matrislere bakıldığında, TP ve TN değerlerinin son iterasyonlara doğru yükseldiği gözlemlenmiştir.

Çizelge 4.8. Wisconsin meme kanseri veri seti üzerinde iteratif optimizasyon eğitimi sonuçlarının karmaşıklık matrisleri



Şekil 4.2'de Wisconsin meme kanseri veri seti üzerinde yapılan eğitim ve değerlendirme testleri özetlemek gerekirse, sonuçlar ilk 0-5 iterasyon boyunca yükselme göstermiştir. Daha sonrasında, eğitimde bir iyileşme gözlemlenmemiştir. Dolayısıyla, iteratif eğitim beşinci iterasyonda kesilmiştir.



Şekil 4.2. Wisconsin meme kanseri veri seti üzerinde gerçekleştirilen eğitim sonuçlarının kümelenmiş sütun grafiği

4.2.4. Önerilen yöntemin optuna ile kullanılmasıyla elde edilen sonuçlar

Önerilen yöntemin Optuna ile birleştirilmesi, hiperparametre optimizasyonu ile özellik çıkarım süreçlerini sinerjik biçimde entegre ederek en yüksek performans iyileştirmesini hedefler.

Diyabet veri seti

Bu kombinasyon, $F_1 -$ Skorunu %69,3'e yükseltmiş ve model performansını tek başına kullanıldığı senaryoya göre daha da iyileştirmiştir. Optuna'nın adaptif yapısı, çıkarılan yeni özelliklerle dinamik bir şekilde çalışarak daha verimli bir optimizasyon süreci sağlamıştır.

Wisconsin meme kanseri tespiti veri seti

Kombine yöntem, başlangıç $F_1 -$ Skorunu %94,33'e yükselterek iyileşme sağlamıştır. Bu sonuçlar hem özellik çıkarma sürecinin hem de hiperparametre arama stratejisinin güçlü bir şekilde entegre edildiğini göstermektedir.

Her ne kadar Optuna'nın tek başına kullanıldığı senaryoya göre daha iyi sonuçlar almış olsa da sadece önerilen yöntemin kullanıldığı senaryodan daha düşük sonuçlar elde edilmiştir.

4.2.5. Tüm senaryoların F1-skoru karşılaştırma tablosu

Çizelge 4.9. Tüm senaryoların karşılaştırma tablosu

No	Modeller	$F_1 - Skoru$ (%)	
		Diyabet veri seti	Wisconsin veri Seti
1	Temel model	67,96	91,1
2	Klasik optimizasyon	68,5	92,6
3	Optuna	68,6	90
4	Önerilen yöntem+Optuna	69,3	94,33
5	Önerilen yöntem	80	97,73

Alınan tüm eğitim ve değerlendirme sonuçlarının gösterildiği Çizelge 4.9'da yapılan her işlem çoğunlukla alınan sonuçlarla iyileştirilmiştir. Yapılan analizlere göre, Optuna örnek sayısı az olan veri setlerinde etkisini gösterememektedir [30]. Aynı zamanda, Optuna ile önerilen yöntem birlikte kullanıldığında, önerilen yöntemin özellik çıkarma yeteneği tek başına olduğu senaryoya göre daha düşük olmaktadır. Dolayısıyla, daha önceden optimize edilmiş veri setleri yerine direkt önerilen yöntemin uygulanması hem zaman hem de başarı metrikleri çerçevesinde daha makul kalmaktadır.

4.3. Yorum ve Çıkarımlar

Bu tezde önerilen yaklaşım, seçilen iki farklı veri seti üzerinde test edilerek model performansını artırma konusunda etkili sonuçlar ortaya koymuştur. Diyabet veri setinde %9,69; Wisconsin Meme Kanseri veri setinde ise %2,59 oranında iyileşme sağlanmıştır. Başlangıç performans düzeylerinin bu iyileşme oranları üzerindeki etkisi gözlemlenmiş; özellikle daha düşük başlangıç performansına sahip veri setlerinde yöntemin daha yüksek fayda sağladığı görülmüştür. Her yinelemede hiperparametrelerin optimize edilmesi, model başarımına anlamlı katkı sağlamış, ancak bu sürecin daha büyük veri kümeleri için verimlilik açısından yeniden ele alınması gerektiği ortaya çıkmıştır. Ayrıca yöntemin mevcut hali, bazı manuel adımlar içermekle birlikte, bu adımların otomatikleştirilmesi geniş veri setlerine

ölçeklenebilirliği ve uygulama maliyetini olumlu yönde etkileyebilir. Farklı veri yapıları üzerinde elde edilen olumlu sonuçlar ise önerilen yaklaşımın genelleştirilebilir yapısını desteklemekte ve çeşitli uygulama alanları için potansiyel sunduğunu göstermektedir.

Önerilen yöntem, iteratif özellik çıkarımı yoluyla model performansını artırmak için etkili bir yaklaşım sunmaktadır. Bununla birlikte, daha büyük veri kümeleri için işlem maliyetlerini optimize etmek ve hiperparametre aramalarını hızlandırmak adına ek iyileştirmelere ihtiyaç duyulmaktadır. Bu iyileştirme en iyi küme sayılarını bulmak ve son model ile test etmek amacı ile oluşturulacak bir yatay alt küme ile sağlanabilir. Bu yatay alt küme üzerinden çıkarılan küme sayıları daha sonra son eğitim için kullanılabilir. Bununla birlikte küme sayıları belirlenirken eş zamanlı işlem (Multi-Threading [31]) yapılması yine toplam eğitim süresini azaltacaktır. Küme sayıları bulunurken hedef metrik bazında aynı sonucu veren kümeler arama ağacının bir parçası olarak kaydedilip kullanıcıya sunulan bir parametre ile daha sonra denenebilir. Bu bulgular, yöntemin genişletilebilirliğini artırmak için değerli içgörüler sunmaktadır.

5. SONUÇ VE ÖNERİLER

Bu tezde, meta-model doğruluğunu artırmaya yönelik yenilikçi bir özellik çıkarma metodolojisi önerilmiş ve iteratif denetimsiz öğrenme yaklaşımı detaylı bir şekilde incelenmiştir. Denetimsiz öğrenme yoluyla, veri kümesinden elde edilen içgörüler doğrultusunda dinamik olarak geliştirilen bir özellik setiyle meta-modelin performansı artırılmıştır.

Çalışmadan elde edilen bulgular, iteratif özellik çıkarma yaklaşımının meta-model doğruluğunu artırmadaki etkinliğini ortaya koymaktadır. Yinelemeli optimizasyon sırasında kullanılan algoritmanın parametrelerinin dinamik olarak ayarlanması, algoritmanın veri içindeki farklı desenleri yakalamasını sağlamış ve adaptasyon kabiliyetini artırmıştır. Ayrıca, değerlendirme sürecinde elde edilen verilerin orijinal özellik kümesine entegrasyonu, meta-modelin sürekli olarak iyileştirilmesine katkıda bulunmuştur.

Bu tezde sunulan iteratif denetimsiz öğrenme metodolojisi, hem özellik çıkarma için k-ortalamalar kümelemenin potansiyelini hem de veri kümelerinin dinamik doğasına uyum sağlamanın önemini vurgulamaktadır. Deneysel bulgularda gözlemlenen olumlu sonuçlar, önerilen yaklaşımın makine öğrenimi modellerinin yorumlanabilirliğini ve doğruluğunu artırmadaki uygulanabilirliğini ortaya koymaktadır.

Bu yaklaşımın farklı veri kümeleri üzerinde uygulanması ve diğer denetimsiz öğrenme algoritmalarıyla performansının değerlendirilmesi, yöntemle ilgili daha fazla içgörü sağlayabilir.

Ek özelliklerin entegrasyonu ya da iteratif sürecin iyileştirilmesi, bu yaklaşımın sürekli olarak gelişmesine katkı sağlayabilir.

Özellikle büyük ölçekli veri kümelerinde hesaplama maliyetlerini optimize edecek stratejiler geliştirilebilir.

Bu tezde, özellik çıkarma ve model iyileştirme için denetimsiz öğrenme yöntemlerinin gelişen alanına bir katkı sunmaktadır. Elde edilen sonuçlar, önerilen metodolojinin yalnızca

deneysel doğruluk açısından değil, aynı zamanda esneklik, genellenebilirlik ve model geliştirme süreçlerine katkı bakımından da güçlü bir potansiyel taşıdığını ortaya koymaktadır. Özellik çıkarma sürecinde denetimsiz öğrenme algoritmalarının etkin şekilde kullanılması, etiketlenmemiş verilerin değerlendirilmesi açısından önemli bir avantaj sunmakta; bu da önerilen yöntemi, etiketli veri erişiminin kısıtlı olduğu senaryolarda uygulanabilir bir çözüm haline getirmektedir. Bu yönüyle çalışma, gelecekte yapılacak araştırmalar için hem metodolojik hem de uygulamalı açıdan sağlam bir temel oluşturmaktadır.

KAYNAKLAR

1. Sarker, I. H. (2022). AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, 3(2), 158.

2. Piernik, M., Morzy, T. (2021). A study on using data clustering for feature extraction to improve the quality of classification. *Knowledge and Information Systems*, 63(7), 1771-1805.

3. İnternet: Kaggle. Diabetes dataset. URL: https://www.kaggle.com/datasets/mathchi/diabetes-data-set. Son Erişim Tarihi: 21 Nisan 2025.

4. Wolberg, W. H., Street, W. N., Heisey, D. M., Mangasarian, O. L. (1995). Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*, 26(7), 792-796.

5. Al-Yaseen, W. L., Othman, Z. A., Nazri, M. Z. A. (2017). Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. *Expert Systems with Applications*, 67, 296-303.

6. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

7. Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

8. Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129-137.

9. Ahmed, M., Seraj, R., Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8), 1295.

10. Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Kanada, 78-86.

11. Jiang, X., Yu, X., Wang, B., Ma, L., Guan, Y. (2023). DECAF: An interpretable deep cascading framework for ICU mortality prediction. *Artificial Intelligence in Medicine*, 135, 102521.

12. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, 4-8 August 2019, 2623-2631.

13. Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., Jia, H. (2023). K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178-210.

14. Auliasari, K. (2021). Feature extraction and K-means clustering approach to classify the COVID-19 lung CT-scan image. *International Journal of Information System and Computer Science*, 5(3), 171-183.

15. Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 13-17 August 2016, 785-794.

16. Dua, D., Graff, C. (2017). UCI Machine Learning Repository. *University of California, Irvine, School of Information and Computer Science* (İnternet). URL: http://archive.ics.uci.edu/ml (Son Erişim Tarihi: 21 Nisan 2025).

17. İnternet: Python Software Foundation. Python Language Reference (Sürüm 3.9). URL: https://docs.python.org/3/reference. Son Erişim Tarihi: 21 Nisan 2025.

18. İnternet: Project Jupyter. Jupyter Notebook (açık kaynak kodlu defter uygulaması). URL: https://jupyter.org. Son Erişim Tarihi: 21 Nisan 2025.

19. İnternet: Anaconda. Anaconda Distribution (Sürüm 2023.03). URL: https://www.anaconda.com. Son Erişim Tarihi: 21 Nisan 2025.

20. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

21. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

22. McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, Austin, TX, 51-56.

23. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362.

24. Strelcenia, E., Prakoonwit, S. (2023). Improving classification performance in credit card fraud detection by using new data augmentation. *AI*, 4(1), 172-198.

25. MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA: University of California Press, 1, 281-297.

26. Chan, S. H. (2021). *Introduction to Probability for Data Science*. Ann Arbor, MI: Michigan Publishing.

27. LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

28. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, 2623-2631.

29. Singh, D., Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97, 105524.

30. Tetko, I. V., Karpov, P., Van Deursen, R., Godin, G. (2024). State-of-the-art augmented neural networks in chemistry: what can we learn from them? *Journal of Cheminformatics*, 16(1), 8.

31. Aslam, M., Sherwani, R. A. K., Saleem, M. (2021). Vague data analysis using neutrosophic Jarque–Bera test. *PLOS ONE*, 16(12), e0260689.



Gazili olmak ayrıcalıktır