YZM 4008
Veri Madenciliği Dersi Proje Raporu

Algoritma Adı: FDB-TLABC

Öğrenci Adı ve Soyadı: Özge Gürbüz
Öğrenci No: 385965
Öğrenci İletişim Bilgileri
Cep tel: 0 (545) 572 26 30
***e-mail adresi: ozgegurbuz99@gmail.com**

1.Giriş

Optimizasyon, bir sistemde var olan kaynakların (işgücü,zaman,kapasite) en verimli şekilde kullanılarak belirli amaçlara (maliyet en azaltması veya kar en çoklaması) ulaşmayı sağlayan bir teknoloji olarak tanımlanmaktadır.
Optimizasyonda modelleme ve çözümleme iki önemli bileşen olarak nitelendirilmektedir. Modelleme, gerçek yaşamda karşılaşılan problemin matematiksel olarak ifade edilmesi ; çözümleme ise bu modeli sağlayan en iyi çözümün elde edilmesini kapsamaktadır.Pek çok optimizasyon problemi bulunmaktadır. Ancak bu metotların çoğu belirli bir tür problemin çözümü için geliştirilmiş yöntemlerdir. Bu nedenle optimizasyon probleminin çözümü için gerekli metodun seçimi için optimizasyon probleminin türünün belirlenmesi önemlidir.

Metasezgisel algoritmalar, büyük ölçekli optimizasyon problemlerinin çözümü için optimuma yakın sonuçlar döndüren algoritmalardır. Metasezgisel optimizasyon algoritmaları, genetik, fizik, sürü gibi farklı başlıklar altında incelenmektedir. Örneğin doğada arıların yiyecek bulma davranışlarını inceleyen ve bu hareketler ile gerçek dünya problemlerinin çözümü için uyarlanan Yapay Arı Kolonisi Algoritması veya kuşların doğada yiyecek bulabilmek için sürü olarak izledikleri hareketleri de gerçek dünya problemine uyarlayarak süreci optimize eden Parçacık Sürü Optimizasyon Algoritmaları geliştirilen algoritmalara birer örnektir [1] .

2.Materyal ve Yöntem

Geliştirilen melez algoritmanın anlaşılması için konu üç bölümde detaylandırılmıştır.

2.1. k-En Yakın Komşu Algoritması

K-en yakın komşu (k-nearest neighbors, KNN) algoritması, gözlemlerin birbirlerine olan benzerlikleri üzerinden tahminlerin yapıldığı, gözetimli makine öğrenmesi modellerinde regresyon ve sınıflandırma problemlerinde kullanılan bir algoritmadır.

Bu algoritma kapsamında tahminde bulunmak istediğimiz gözlem birimine en yakın K adet farklı gözlem birimi tespit edilir ve bu K adet gözlem biriminin bağımlı değişkenleri üzerinden ilgili gözlem için tahminde bulunulur [2].

Klasik k-en yakın komşu (k-nearest neighbors, k-NN) algoritmasında, gözlemler arasındaki uzaklık hesaplaması genellikle Öklid bağıntısı kullanılarak yapılır.Öklid bağıntısı , iki nokta arasındaki doğrusal uzaklığı ölçen bir metriktir. İki nokta arasındaki Öklid bağıntısı hesaplamak için aşağıdaki formül kullanılır:

$$d(x, y) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + ... + (xn - yn)^2}$$

Burada, d(x, y) iki nokta x ve y arasındaki Öklid mesafesini temsil eder. k-NN algoritması, bir gözlemi sınıflandırmak veya tahmin etmek için en yakın k komşusunu kullanır. Bu nedenle, bir veri noktasının diğer veri noktalarına olan uzaklıklarını hesaplarken Öklid bağıntısı formülünü kullanarak tüm diğer noktalarla olan uzaklıkları hesaplamak gerekir. Ardından, en yakın k komşuyu belirleyerek sınıflandırma veya tahmin yapılabilir.

1. *Veri setinin tanımlanması:* probleme ait n-adet örnek gözlemleri içeren ve problem uzayını temsil etme kabiliyeti yüksek (gözlem uzayını homojen olarak örnekleyen) **X** veri setini oluştur.
2. *Uzaklık bağıntısının belirlenmesi:* gözlemler arasındaki uzaklıkların hesaplanmasında kullanılacak yöntemi belirle.
3. *k-değerinin belirlenmesi:* gözlem sayısına ve veri setinin karakteristiğine bağlı olarak k-komşu sayısı için arama uzayının sınırlarını tanımla.
4. **for each** k_j
   k_j için sınıflandırma performansını $SPk_j = f_{k\text{-}nn}(k_j)$
   if $(SPk_j > SPk_{j-1})$
   $k = k_j$
   end if
   end
5. En iyi sınıflandırma performansı sağlayan k-değerini kaydet
6. Sınıf etiketi belirlenecek olan q sorgu gözlemini tanımla
7. **for i=1:n**
   $D_{[i]}=q$ ile $\mathbf{X_i}$ arasındaki uzaklığı hesapla
   end
8. $Xq_{[k]}=D_{[i]}$ uzaklık dizisinden q sorgu gözlemine en yakın k-adet gözlemi belirle
9. $Xq_{[k]}$ gözlemlerinin sınıflarını dikkate alarak çoğunluk oylaması/ağırlıklı oylama yöntemi kullanarak q-gözleminin sınıfını belirle.

Algoritma 1. k-nn Algoritmasının Sözde Kodu [3].

2.2. FDB-TLABC Algoritması

FDB-TLABC (Fitness-distance balance -The teaching-learning-based artificial bee colony) algoritması, öznitelik seçimi ve sınıflandırma arasındaki dengeyi sağlamak için tasarlanmış bir yöntemdir. Bu algoritma, veri setindeki özniteliklerin frekanslarını analiz ederek, veri setindeki önemli öznitelikleri seçer ve gereksiz olanları elemine eder. Bu şekilde, boyut indirgeme yaparak sınıflandırma performansını iyileştirir.

TLABC (Öğretme-öğrenme temelli yapay arı kolonisi) algoritması, öğretme-öğrenme tabanlı optimizasyon (TLBO) ve yapay arı kolonisi (ABC) algoritmasının birleşiminden oluşan yeni bir hibrit sürü temelli metaheuristik arama algoritmasıdır. Bu hibrit yaklaşım, TLBO'nun açıklayıcı yeteneklerini ve ABC'nin keşif kabiliyetlerini bir araya getirerek küresel optimizasyon problemlerini etkili bir şekilde çözmeyi amaçlamaktadır [4].

TLABC gibi sürü temelli algoritmalarda, seçim sürecini etkili bir şekilde simüle etmek büyük bir zorluktur. Son zamanlarda geliştirilen bir yöntem olan fitness-uzaklık dengelemesi (FDB) doğayı daha etkili bir şekilde taklit etmek için kullanılan güçlü bir tekniktir [4].

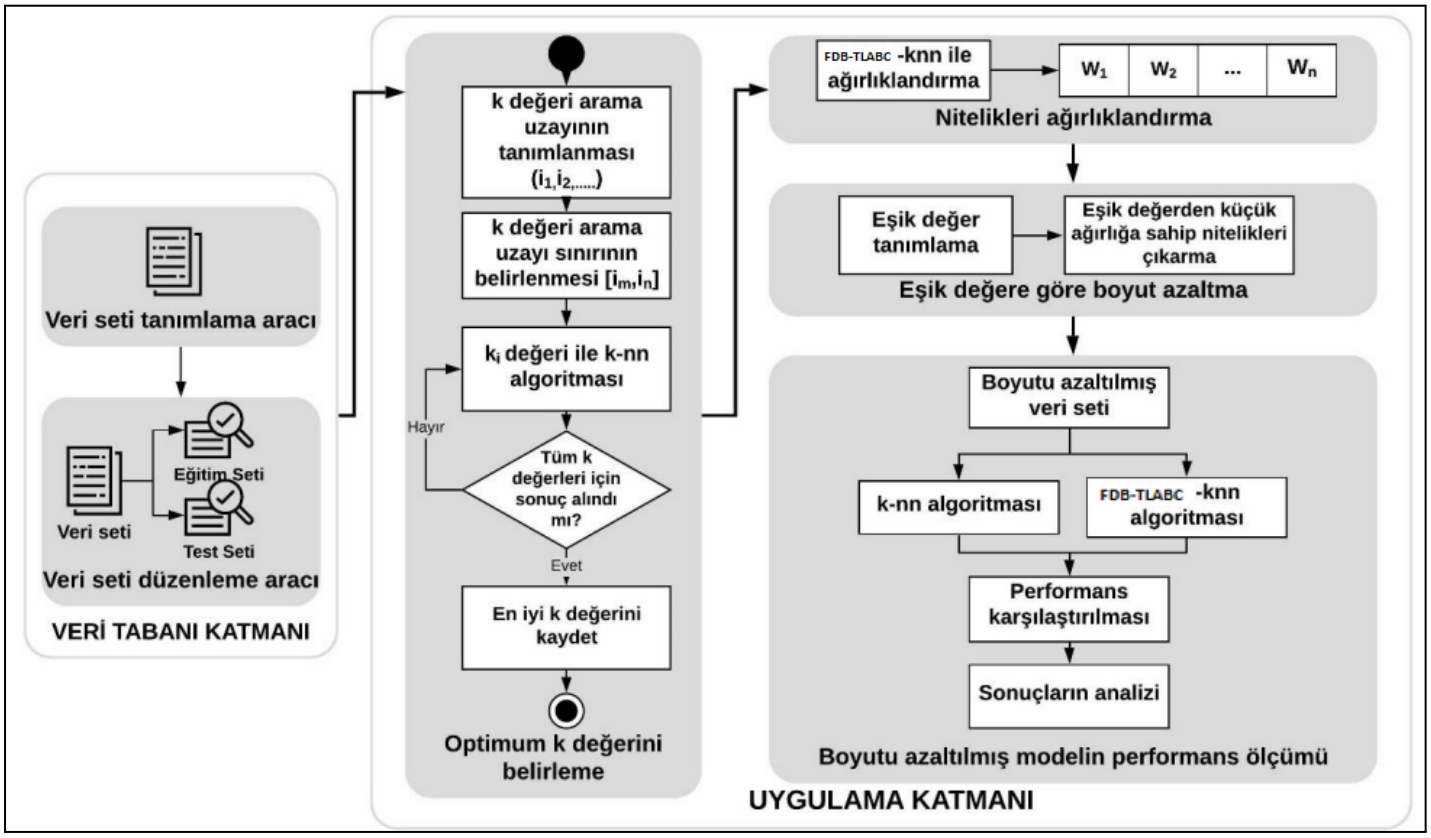
2.3. Önerilen Yöntem:FDB-TLABC ve K-nn ile Sezgisel Boyut İndirgeme

FDB-TLABC Görünümü ve K-nn Konum Yönteminin Birleştirilmesi, doğal zeka ilhamımıyle birlikte düzenlemeleri ve odak tekniklerini kapsamlı bir şekilde kullanmayı amaçlayan bir melez yaklaşımdır. Bu melezleme, global optimizasyon problemlerini yorumlamak için kullanılan FDB-TLABC beklentilerinin problemlerindeki performansı artırmayı hedefler.

FDB-TLABC algoritması, doğayı taklit ederek keşif ve sömürü yeteneklerini birleştirir ve global optimizasyon problemlerinde başarılı sonuçlar verir. K-nn ise bir sınıflandırma algoritması olup, doğru sınıflandırma yapmak için verilerde boyut indirgeme işlemine ihtiyaç duyar. Bu noktada, FDB-TLABC algoritmasının sınıflandırma performansını artırmak için K-nn yöntemiyle birleştirilmesi önemli bir yaklaşımdır.

Bu melez yaklaşım, ilk olarak FDB-TLABC algoritmasının boyut indirgeme yeteneklerini kullanarak özniteliklerin etkin bir şekilde ağırlıklandırılmasını sağlar. Daha sonra, elde edilen ağırlıklandırılmış öznitelikler K-nn sınıflandırma algoritmasına giriş olarak verilir. Bu şekilde, FDB-TLABC algoritmasının optimize ettiği öznitelikler, K-nn sınıflandırma algoritmasıyla birlikte kullanılarak daha kesin ve etkili bir sınıflandırma süreci elde edilir.

FDB-TLABC ve K-nn'in melezlenmesi, global optimizasyon problemlerini çözmek için kullanılan FDB-TLABC algoritmasının sınıflandırma performansını artırırken, aynı zamanda boyut indirgeme yeteneklerinden yararlanmayı sağlar. Bu melez yaklaşım, farklı alanlarda kullanılan ve başarılı sonuçlar veren sezgisel optimizasyon ve sınıflandırma tekniklerinin güçlerini birleştirerek daha güçlü bir çözüm sağlar.



Şekil 1. FDB-TLABC ve knn ile Sezgisel Boyut İndirgeme Sürecinin Öğeleri [3].

Veri Tabanı Katmanı

Bu çalışmada ilk olarak veri seti incelenip, gerekli düzenlemeler yapılmıştır. Yüklenen veri seti iki ayrı alt veri setine bölünmüştür: Eğitim seti ve test seti. Modelin eğitim setiyle eğitilmesi sağlanmakta ve test setiyle başarısı ölçülmektedir. Veri setlerinde toplam örnek sayısının yaklaşık olarak %70'i eğitim, %30'u test için kullanılmıştır.

Uygulama Katmanı

Optimum k değerini belirleme: incelenen makalede yer alan çalışmadaki veri seti kullanmış ve bu çalışmada belirlenen k değeri kullanılmıştır [3].

FDB-TLABC ve knn algoritması, niteliklerin ağırlıklandırılması için kullanılan bir yöntemdir. Bu yöntemde, sezgisel k-nn algoritması ile niteliklerin probleme olan etkisi incelenir ve buna göre ağırlıklandırma işlemi gerçekleştirilir. FDB-TLABC algoritması, bu ağırlıkların çözüm adayları olduğu bir meta-sezgisel arama algoritmasıdır. Çözüm adayları, 0 ile 1 arasında sınırlanmıştır, yani en uygun ağırlıklar 0 ile 1 arasında aranır. Amaç fonksiyonu ise sezgisel k-nn'dir. Ağırlıkların uygunluk değerleri, hedef (amaç) fonksiyonundan dönen sınıflandırma hata değeri ile ölçülür. Bu sayede, FDB-TLABC ve knn algoritması ile niteliklerin ağırlıklandırılması, niteliklerin sınıflandırma performansına olan etkisini değerlendirerek en uygun ağırlıkların bulunmasını sağlar.

FDB-TLABC algoritması ile nitelik seçimi/boyut azaltma aşamasında, en uygun ağırlıkları arama işlemi sonlandırıldıktan sonra, problemin boyutunu azaltma/nitelik çıkarımı aşamasına geçilir. Bu aşamada, bir eşik değeri kullanılır ve 0 ile 1 arasındaki ağırlıklar arasında, eşik değerden daha düşük ağırlığa sahip nitelikler çıkarılır [3].

Boyutu azaltılmış modelin performansı ölçülür. Niteliklerin çıkarılması işleminden sonra, modelin sınıflandırma performansına bakılır. Klasik k-nn algoritması ve sezgisel k-nn algoritmalarının sınıflandırma hata değerleri incelenir. Eğer bu performans hata değerleri, nitelik çıkarılmadan önceki modele göre daha düşükse, yani sınıflandırma performansı düşmemiş hatta iyileşmişse, başarı elde edilmiş olur. Bu durumda, eşik değere göre nitelik seçimi/boyut azaltma işlemi etkili bir şekilde gerçekleştirilmiş demektir [3].

3.Deneysel Sonuçlar

Bu bölümde uygulamada kullanılan veri setleri, önerilen yöntem kullanılarak veri setlerinde boyut azaltma ve nitelik çıkarımı, çıkarılan bu niteliklerden sonra veri setinin sınıflandırma başarısına bakılmaktadır.

3.1. Ayarlar

Algoritma ayarları için Tablo 1'de gösterilen değerler aynı şekilde kullanılmıştır.

Algoritma	Parametre Değerleri
FDB-TLABC	n=20, maxFE = 9500
k-nn	Uzaklık Bağıntısı = Öklid

Tablo 1. Algoritmaların Parametre Değerleri

3.2. Veri Seti

Bu çalışmada geliştirilen algoritma UCI Machine Learning veri havuzundan temin edilen veri seti üzerinde tatbik edilmiştir. Veri seti seçilirken sınıflandırma problemine uygun bir veri seti olmasına, eksik/hatalı veri olmamasına dikkat edilmiştir.

	Diabetic Retinopathy Debrecen
Boyut	20
Eğitim Örnek	805
Test Örnek	345

Tablo 2. Veri Seti Bilgisi

Kullanılan veri setindeki bağımlı değişken sayısı 1 olmakla birlikte bağımsız değişken sayısı 19 ‘dur.

3.3. Deneysel Sonuçların Analizi

İncelenen makalede yer alan çalışmadaki veri seti kullanmış ve veri seti için k komşu sayısı 7 olarak alınmıştır.

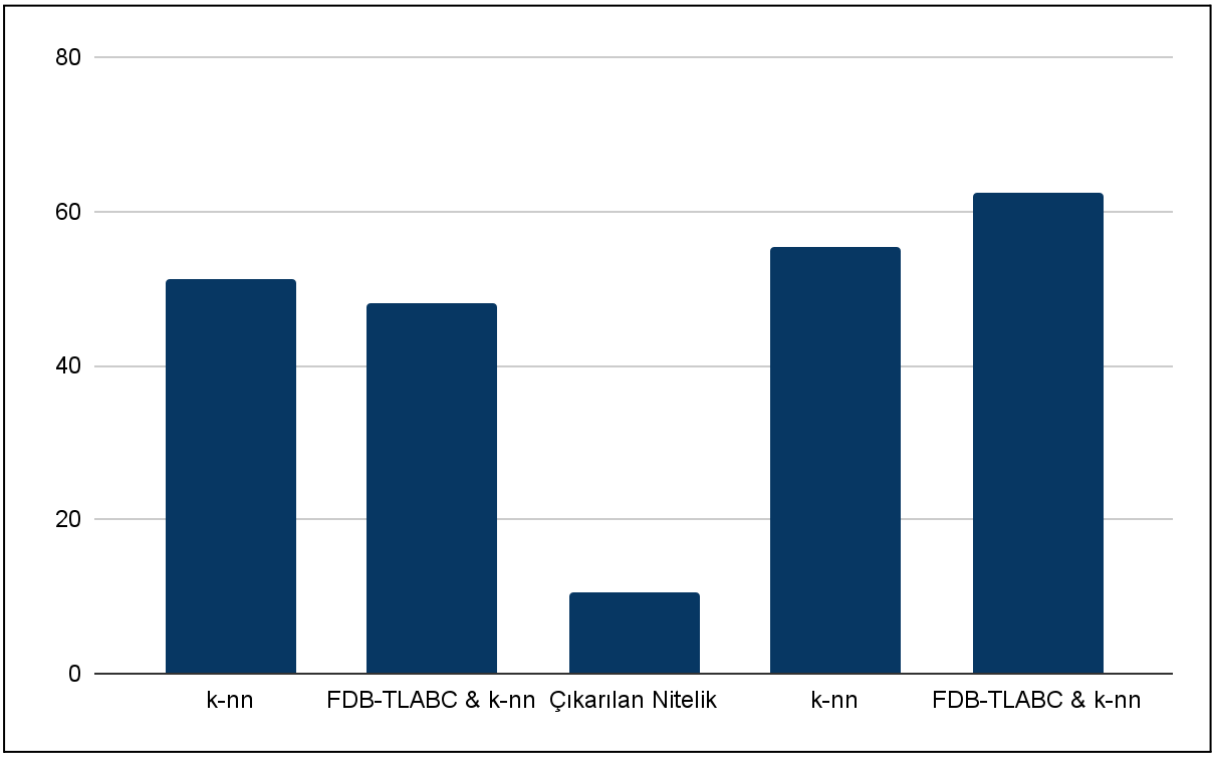
Algoritma		**Diabetic Retinopathy Debrecen**
FDB-TLABC	En İyi	37,68
	En Kötü	45,79
	Ortalama	49,85
	Std. Sapma	5,84
k-nn	Sonuç	44,63

Tablo 3. Algoritmaların Sınıflandırma Hata Değerleri

Uygulamanın bir sonraki aşamasında ise boyut azaltma işlemi gerçekleşmiştir. Eşik değerden düşük ağırlığa sahip nitelikler çıkarılmıştır. Eşik değere bağlı olarak çıkarılan nitelik sayısı ve çıkarılan niteliklerden sonra algoritmaların sınıflandırma hata değerleri Tablo 4’te verilmiştir.

Eşik Değeri		0.02	0.04
Çıkarılan Nitelik Sayısı	Ortalama	0,005	0,018
	Ortanca	0,005	0,018
	Std.Sapma	0,007	0,017
k-nn	En İyi	44,63	46,95
	En Kötü	52,75	57,39
	Ortalama	48,88	52,94
	Std.Sapma	4,07	5,38
FDB-TLABC	En İyi	40	37.68
	En Kötü	42,31	45.79
	Ortalama	41,15	41,93
	Std.Sapma	1,15	4,07

Tablo 4. Boyut Azaltıldıktan Sonra Algoritmaların Sınıflandırma Hata Değerleri



Şekil 3. Veri Setinde Çıkarılan Niteliklerin Sayısı (%) ve Sınıflandırma Doğruluk Değerleri.(Boyut azaltılmadan önceki ve boyut azaltıldıktan sonraki k-nn & FDB-TLABC algoritmalarının sınıflandırma doğruluk değerleri gösterilmektedir.)

Şekil 3'te her veri seti için nitelik çıkarma/boyut azaltma yapılmadan önce k-nn ve FDB-TLABC ve knn uygulanarak elde edilen ortalama sınıflandırma doğruluk değerleri, o veri setinin ortalama niteliklerinin yüzde kaçının çıkarıldığı ve bu çıkarılan niteliklerden sonra k-nn ve FDB-TLABC ve knn algoritmaları uygulanarak elde edilen ortalama sınıflandırma doğruluk değerleri verilmiştir.

Şekil 3’e bakılarak veri setinin niteliklerinin %10,5’i çıkarıldığında performansının düşmediği hatta k-nn ve FDB-TLABC algoritmalarında iyileşme olduğu görülmüştür.

4. Sonuç ve Tartışma

Bu çalışmada, sınıflandırma problemleri için k-nn ve FDB-TLABC algoritmalarının melezlenmesiyle bir meta-sezgisel boyut indirgeme algoritması önerilmiştir. Önerilen algoritma, öznitelik sayısını azaltarak sınıflandırma performansını iyileştirmeyi hedeflemektedir.

Bu çalışmanın temel katkıları şunlardır:

1. Melezlenen k-nn ve FDB-TLABC algoritmaları: K-nn algoritması, yakın komşuluk temelli bir sınıflandırma yöntemidir, ve FDB-TLABC algoritması öznitelik seçimi ve sınıflandırma arasındaki dengeyi sağlamak için tasarlanmıştır. Bu çalışmada, bu iki algoritma birleştirilerek daha iyi bir sınıflandırma performansı elde edilmiştir.
2. Meta-sezgisel boyut indirgeme algoritması: Önerilen algoritma, meta-sezgisel bir yaklaşım kullanarak boyut indirgeme yapmaktadır. Bu yaklaşım, veri setine bağımlı olmadan genel bir boyut indirgeme algoritması sunar. Bu şekilde, farklı veri setleri ve sınıflandırma problemleri üzerinde uygulanabilirlik sağlar.
3. Performans değerlendirmesi: Deneylerde, önerilen algoritmanın sınıflandırma performansını değerlendirmek için çeşitli performans ölçütleri kullanılmıştır. Elde edilen sonuçlar, önerilen algoritmanın daha az öznitelik kullanarak daha yüksek doğruluk, hassasiyet, duyarlılık elde ettiğini göstermektedir.

Bu çalışmanın sınırlamaları da bulunmaktadır. Öncelikle, önerilen algoritmanın genel performansı, veri setinin özelliklerine bağlı olabilir. Farklı veri setleri üzerinde daha fazla deney yapılması ve sonuçların karşılaştırılması gerekmektedir. Ayrıca, bu çalışmada sadece k-nn ve FDB-TLABC algoritmalarının melezlenmesi üzerinde durulmuştur. Başka sınıflandırma algoritmaları veya boyut indirgeme yöntemleri ile de benzer çalışmalar yapılabilir.

Önerilen algoritmanın gerçek dünya uygulamalarında kullanılabilirliği ve performansı da değerlendirilmelidir. Büyük veri setleri, gerçek zamanlı uygulamalar ve farklı sınıflandırma problemleri üzerinde algoritmanın etkinliği daha detaylı olarak incelenebilir.

Sonuç olarak, bu çalışmada önerilen k-nn ve FDB-TLABC algoritmalarının melezlenmesiyle oluşturulan meta-sezgisel boyut indirgeme algoritması, sınıflandırma problemlerinde öznitelik sayısını azaltarak daha iyi bir sınıflandırma performansı sunmaktadır. Bu çalışma, sınıflandırma alanında öznitelik seçimi ve boyut indirgeme konularında yeni bir yaklaşım sunmaktadır. Ayrıca, algoritmanın gerçek dünya uygulamalarında kullanılabilirliği ve performansı daha fazla çalışma gerektiren bir konudur. Bu çalışma, sınıflandırma problemlerine yönelik daha etkili boyut indirgeme yöntemlerinin geliştirilmesi için bir adım olarak önemli bir katkı sunmaktadır.

Kaynakça

1. https://www.mshowto.org/metasezgisel-algoritmalar.html#close (Ziyaret tarihi:22 Mayıs 2023)
2. https://www.miuul.com/not-defteri/k-en-yakin-komsu-algoritmasi-nasil-calisir (Ziyaret tarihi: 22 Mayıs 2023)
3. KAHRAMAN, H., Büşra, A. R. A. S., & YILDIZ, O. (2020). SINIFLANDIRMA PROBLEMLERİ İÇİN AGDE-TABANLI META-SEZGİSEL BOYUT İNDİRGEME ALGORİTMASININ GELİŞTİRİLMESİ. Mühendislik Bilimleri ve Tasarım Dergisi, 8(5), 206-217.
4. Duman, S., Kahraman, H. T., Sonmez, Y., Guvenc, U., Kati, M., & Aras, S. (2022). A powerful meta-heuristic search algorithm for solving global optimization and real-world solar photovoltaic parameter estimation problems. Engineering Applications of Artificial Intelligence, 111, 104763.