Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Yapay Sinir Ağları ve *K*-Ortalamalar Tabanlı Büyük Veri Azaltma Algoritmasının Tasarımı ve Uygulaması[1]

Seyithan TEMEL [a,*], Hamdi Tolga KAHRAMAN [b]

[a] *Yazılım Mühendisliği Bölümü, Of Teknoloji Fakültesi, Karadeniz Teknik Üniversitesi, Ankara, TÜRKİYE*

[b] *Yazılım Mühendisliği Bölümü, Of Teknoloji Fakültesi, Karadeniz Teknik Üniversitesi, Trabzon, TÜRKİYE*

**Sorumlu yazarın e-posta adresi: seyithanntemel@gmail.com*

DOI: 10.29130/dubited.1014161

Öz

Büyük veri azaltma sürecinde karşılaşılan başlıca zorluk, veri setinin homojenliğinin ve problem uzayını temsil yeteneğinin korunmasıdır. Bu durum, büyük veri setleri üzerinde yapılan modelleme çalışmalarında hesaplama karmaşıklığının yeterince azaltılamamasına, geliştirilen modelin orijinal veri setine dayalı olarak geliştirilen modele kıyasla kararlılık ve doğruluk performansının önemli ölçüde azalmasına neden olmaktadır. Bu makale çalışmasının amacı, büyük veri setleri için kararlı ve etkili bir şekilde çalışan veri azaltma algoritması geliştirmektir. Bu amaçla, yapay sinir ağları (YSA) tabanlı problem modelleme modülü ve *K*-ortalamalar tabanlı veri azaltma modülünden oluşan melez bir algoritma geliştirilmiştir. Problem modelleme modülü, büyük veri seti için performans eşik değerlerini tanımlamayı sağlamaktadır. Bu sayede, orijinal veri setinin ve veri azaltma işlemi uygulanmış veri setlerinin problem uzayını temsil yetenekleri ve kararlılıkları analiz edilmektedir. *K*-ortalamalar modülünün görevi ise, veri uzayını *K*-adet kümede gruplamayı ve bu grupların her biri için küme merkezini referans alarak kademeli olarak veri (gözlem) azaltma işlemini gerçekleştirmektir. Böylelikle, *K*-ortalamalar modülü ile veri azaltma işlemi uygulanırken, azaltılmış veri setlerinin performansı ise YSA modülü ile test edilmekte ve performans eşik değerlerini karşılama durumu analiz edilmektedir. Geliştirilen melez veri azaltma algoritmasının performansını test etmek ve doğrulamak amacıyla UCI Machine Learning uluslararası veri havuzunda yer alan üç farklı veri seti kullanılmıştır. Deneysel çalışma sonuçları istatistiksel olarak analiz edilmiştir. Analiz sonuçlarına göre büyük veri setlerinde kararlılık ve performans kaybı yaşanmadan %30-%40 oranları arasında veri azaltma işlemi başarılı bir şekilde gerçekleştirilmiştir.

Anahtar Kelimeler: *Büyük veri, Veri azaltma, K-ortalamalar, Yapay sinir ağları, UCI Machine learning*

Design and Implementation of Artificial Neural Networks and K-Means Based Big Data Reduction Algorithm

Abstract

The main challenge in the big data reduction process is maintaining the homogeneity of the data set and its ability to represent the problem domain. This situation causes the computational complexity to not be sufficiently reduced in modeling studies on large data sets, and the stability and accuracy performance of the developed model decreases significantly compared to the model developed based on the original data set. The purpose of this article study is to develop a stable and effective data reduction algorithm for big data sets. For

this purpose, a hybrid algorithm consisting of artificial neural networks (ANN) based problem modeling module and K-means based data reduction module has been developed. The problem modeling module provides a way to define performance thresholds for a large data set. In this way, the problem space representation capabilities and stability of the original data set and data reduction applied data sets are analyzed. The task of the K-means module is to group the data space into K-numbers clusters and to perform a gradual reduction of data (observation) by referencing the cluster center for each of these groups. Thus, while the data reduction process is applied with the K-means module, the performance of the reduced data sets is tested with the ANN module and the situation of meeting the performance threshold values is analyzed. In order to test and verify the performance of the developed hybrid data reduction algorithm, three different datasets in the UCI Machine Learning international data repository were used. Experimental study results were analyzed statistically. According to the results of the analysis, data reduction between 30% and 40% was successfully performed without any loss of stability and performance in large data sets.

Keywords: *Big data, Data reduction, K-Means, Artificial neural network, UCI Machine learning*

---

[1]ICAIAME 2021 konferansında sunulmuş olup, tam metin olarak basılmıştır.
Geliş: 13/11/2021, Düzeltme: 11/12/2021, Kabul: 16/12/2021

I. GİRİŞ

Günümüzde büyük verinin hızlı ve etkili bir şekilde işlenmesi önemli bir problemdir [1]. Teknolojinin gelişimine bağlı olarak veriye bağlı olarak işleyen uygulamaların ve sistemlerin sayısı da artmıştır. Veri üreten cihaz sayısı ve üretilen verinin parabolik artışı ile büyük veri kavramı hiç olmadığı kadar önemli hale gelmiştir. Büyük veri üzerinde yapılan işlemler büyük veri madenciliği kapsamındadır. Büyük veri madenciliğinde amaç büyük veriyi işleyip değerli veriyi elde etmektir. Bu büyük veriler için önemli bir noktadır çünkü veri büyüdükçe işlenmesi için gereken maliyette parabolik olarak büyür. Veri boyutu azaltma konusunda oldukça güçlü ve güncel melez bir algoritma olarak [2] numaralı çalışma incelenebilir. Benzer şekilde veri madenciliğinde gürültülü verilerin tespit edilmesinde kullanılan melez ve etkili bir algoritma olarak [3] numaralı çalışma incelenebilir.

Literatürde Değerli veriyi elde etmek için kullanılan çeşitli yöntemler bulunmaktadır. Bu yöntemler arasında en çok bilinen algoritmalar Temel Bileşen Analizi (PCA, Princpal Component Analysis) ve Negatifsiz Matris Çarpanlarına Ayırma (NNMF-Non Negative Matrix Factorization) algoritmalarıdır. Bu algoritmalar boyut indirgeme tabanlı algoritmalardır.

Bu makale çalışmasında amaç, büyük veri içerisindeki değerli veriye ulaşmak ve en az sayıda veriyle büyük veri setlerinin problemi temsil etme yeteneğini sağlayabilmektir. Böylelikle problem uzayını homojen bir şekilde özetleyen bir veri setinin oluşturulması hedeflenmektedir. Bu doğrultuda geliştirilen yöntem büyük veri içerisindeki her bir veri noktasını işleyip değerli veriyi arayacaktır. Literatürde büyük veride değerli veriye ulaşma konusunda verinin kendisini azaltarak azaltma işlemi yapan herhangi bir yönteme rastlanmamıştır. Geliştirilen yöntem yapay sinir ağları kullanılarak tahmin problemlerinde tahmin performansına göre değerlendirilmiştir. Bu değerlendirmeye göre yapay sinir ağlarının tahmin performans hatası değerli veri setinde daha düşük, orijinal veri setinden daha yüksek olmalıdır. Bu amaçla boyut sayısı ve veri sayısı bakımından çeşitli veri setleri ile yapay sinir ağı eğitilip tahmin performans hatası ölçülmüştür. Birden çok tekrar sonucu elde edilen sonuçlar ikili karşılaştırma yöntemi ile karşılaştırılmış olup geliştirilen yöntemin tahmin performansının daha iyi olduğu görülmüştür.

Makalede sırasıyla yöntem, deneysel çalışma ve sonuçlar kısımlarından bahsedilecektir. Yöntem kısmında geliştirilen yöntemin öğeleri olan K-Ortalamalar algoritması ve Yapay Sinir Ağları'ndan bahsedilecek ve daha sonra geliştirilen yöntemin işleyişi ve çalışması ele alınacaktır. Bir sonraki kısım

olan Deneysel Çalışma kısmında kullanılan veri setleri hakkında bilgi verilip, orijinal veri setinin tahmin performansı ve geliştirilen yöntemin tahmin performansı sunulacaktır. Son kısım olan Sonuçlar kısmında deneysel çalışmalarda elde edilen sonuçlar ele alınacaktır.

II. YÖNTEM

Bu bölümde veri azaltma algoritmasının öğeleri olan K-Ortalamalar ve Yapay Sinir Ağlarını ve işleyişi açıklanmaktadır. Bir sonraki kısımda Küme Merkezine Uzaklık Tabanlı Orantılı Veri Azaltma (KMUTOVA) yönteminin detaylarından bahsedilecektir.

A. K-ORTALAMALAR

K-Ortalamalar (KO) yöntemi uzun yıllardan beri en çok kullanılan bölümleyici yöntemlerden biridir. Nesne sınıflandırma, görüntü bölümleme, veri madenciliği, makine öğrenmesi gibi bilişim uygulamaları yanında iktisat, müşteri yönetimi, pazarlama, biyoinformatik ve mühendislik araştırmaları gibi hemen her alanda en çok kullanılan yöntemler arasında en çok karşılaşılan yöntemler arasındadır. K-Ortalamaklar yönteminin en önemli avantajı uygulanmasının basit olmasıdır [4, 5-10].

K-Ortalamalar yöntemi verilerin benzer özelliklerine göre ayrılması için verilerin küme merkezlerine olan uzaklığını kullanmaktadır. K-Ortalamalar yöntemi Denklem (1)' de yer alan J amaç fonksiyonunu minimize etmeyi hedefler.

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2 \quad (1)$$

Denklem (1)'de yer alan J amaç fonksiyonunda *k* küme sayısını, *n* nesne(veri) sayısını belirtir. Denklemde yer alan $x_i^{(j)}$ *j*. kümeye ait *i*. nesne(veri) ve c_j *j* kümesinin küme merkezidir.

$$d_{ij} = \|x_i^{(j)} - c_j\|^2 \quad (2)$$

Denklem (2)'de yer alan d_{ij} *i*. nesne'nin *j*. küme merkezine uzaklığı belirten denklemdir.

$$c_j = \sum_{i=1}^{n_j} \frac{x_{ij}}{n_j}, \; 1 \leq j \leq k \quad (3)$$

K-Ortalamalar yöntemi aşağıdaki adımları takip etmektedir [11]:

1) Veri setinden rastgele k adet küme merkezi seçilir.
2) Veri noktaları ile küme merkezleri arasındaki uzaklık değerleri hesaplanır.
3) Veri noktaları kendilerine en küçük uzaklığa sahip olan merkezlerin ait olduğu kümelere atanır.
4) Küme merkezleri Denklem (3)'e göre güncellenir.
5) Küme değiştiren veri noktaları yoksa ya da birbirini izleyen iki adımda hata karelerindeki artış tanımlanmış bir yaklaşma değerine eşit veya küçükse kümeleme sona erdirilir, değilse 2. adıma geçilerek işlemler tekrarlanır.

Algoritma 1. K-Ortalamalar Algoritması [12]
Girişler : D = {d1, d2,.....,dn} //*D* veri setinden *n* veri.

k // Belirlenen küme sayısı
Çıkışlar : *k* kümeye ayrılmış veri seti
Adımlar : **1.** Başlangıç merkezleri random olarak *k* tane seçilir. **2.** Tekrar et: Her veriyi(*di)* en yakın ağırlık merkezine sahip kümeye atayın; Her küme için yeni ortalama hesaplayın; Yakınsama kriterleri sağlanınca bitir.

B. YAPAY SİNİR AĞLARI

Yapay Sinir Ağları (YSA) insan beyninin öğrenme yapısını modellemek için geliştirilmiştir. İnsanlarda olduğu gibi makinelere de olaylardan öğrenme, sonuç çıkarma ve karar verme yetenekleri kazandırılmak istenmiştir.

YSA'nın temel olarak 2 bileşenden oluşmaktadır. Bu bileşenler Öğrenme Algoritması ve Aktivasyon Fonksiyonudur. YSA'da girdiler ağırlıklar ile çarpılarak net girdi elde edilir. Elde edilen net girdi YSA'nın temel bileşenlerden olan Öğrenme Algoritması'na girdi olarak sağlanır. Ağdaki veri biyolojik ağlarda sinaps olarak bilinen, YSA'da ise ağırlık diye adlandırdığımız uçlar da tutulur. Öğrenme Algoritması en uygun ağırlık değerlerini bulmayı hedefler. En uygun ağırlık değerlerini girdi değerlerini ve mevcut ağırlık değerlerini çeşitli işlemlerden geçirerek bulmayı amaçlar.

YSA'nın diğer bir temel öğesi olan Aktivasyon Fonksiyonu bir değişkeni farklı bir boyuta taşıyan doğrusal veya doğrusal olmayan bir fonksiyondur. YSA'da Aktivasyon Fonksiyonu yapay sinir hücresi girdi verileri üzerinde işlem yaparak buna karşılık gelen net çıktı sonuçları elde eder [13-20].

Algoritma 2. YSA Temel Adımları [21]	
i.	Algoritmayı başlatmak için tüm nöronlar arasındaki bağlantılara rastgele ağırlık atanır.
ii.	Girişler ve bağlantılar kullanarak Gizli Düğümlerin aktivasyon oranı bulunur.
iii.	Gizli düğümlerin ve Çıkış bağlantılarının aktivasyon oranını kullanarak, Çıkış Düğümlerinin aktivasyon oranı bulunur.
iv.	Çıkış düğümündeki hata oranı bulunur ve Gizli Düğümler ile çıkış Düğümleri arasındaki tüm bağlantılar yeniden kalibre edilir.
v.	Çıktı düğümünde bulunan Ağırlıkları ve hatayı kullanarak, Gizli Düğümlerdeki hata indirilir.
vi.	Gizli düğüm ve giriş düğümleri arasındaki ağırlıkları yeniden ayarlanır.
vii.	Bitirme kriteri sağlanana kadar süreç tekrarlanır.
viii.	Son bağlantı ağırlıklarının kullanılması, çıktı düğümlerinin etkinleştirme oranını belirler.

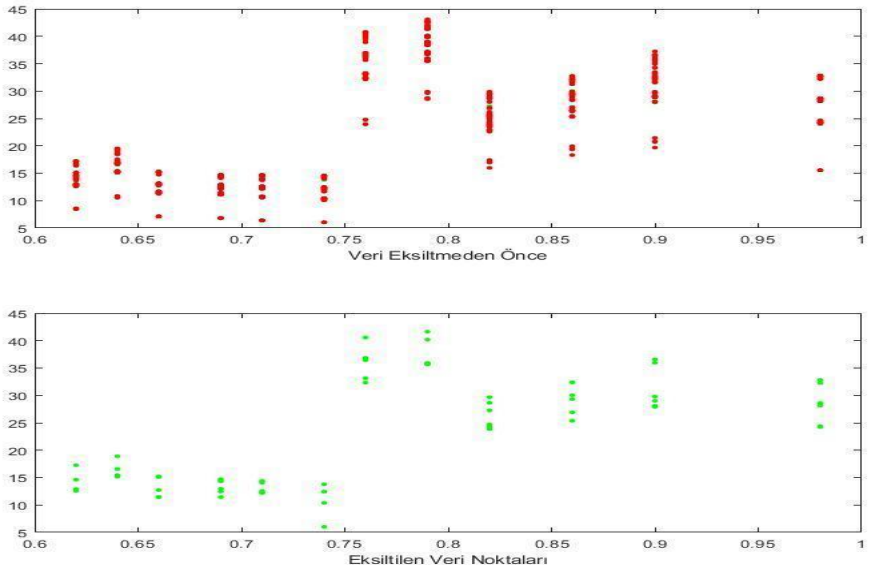
C. GELİŞTİRİLEN YÖNTEM: K-ORTALAMALAR TABANLI BÜYÜK VERİ AZALTMA ARACI

Geliştirilen yöntem olan Küme Merkezine Uzaklık Tabanlı Orantılı Veri Azaltma (KMUTOVA) algoritması girdi setini kümelere ayırarak kümeler üzerinden işlem yapan bir algoritmadır. Algoritma adımlarına göre veri *n* adet kümeye ayrılır. Bu işlemin amacı girdi setini bir bütün olarak değil birer parça olarak ele almak ve ele alınan kümeler (parçalar) içindeki baskın özellikleri öne çıkarmaktır. Girdi setinin kümelere(parçalara) ayrıştırılmaması durumunda tüm girdi seti içerisindeki baskın özellikler ortaya çıkacaktır [24-28]. Girdi setinin kümelere ayrıldıktan itibaren ele alınacak küme içerisinde azaltılacak veri belirleme işlemlerine geçilebilir. Bu işlem küme içerisindeki verilerin küme merkezlerine olan uzaklığı hesaplanarak belirlenir. Küme merkezine uzaklık hesaplamasında Denklem (4)'te verilen Öklid uzaklık bağıntısı kullanılmıştır.

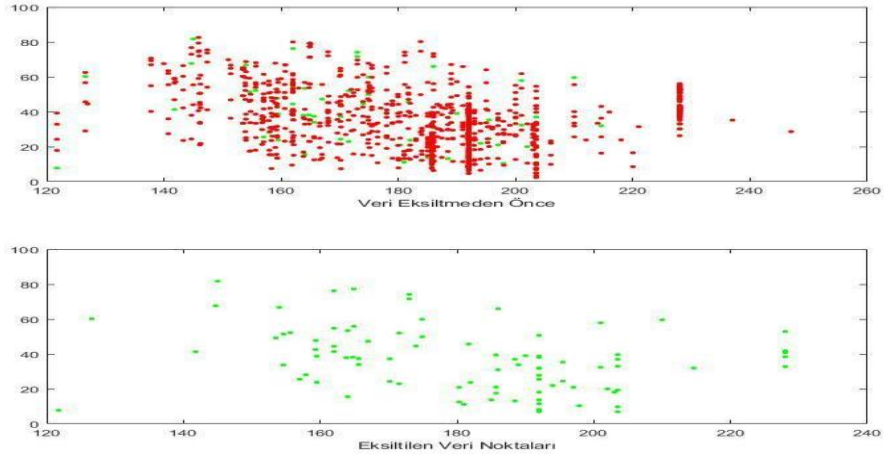
$$S_i = \sqrt{\sum_{j=1}^{n} \left(v_{ij}^{k} - v_{j}^{k}\right)^2} \tag{4}$$

Denklem (4)'te verilen j verilerin boyutunu temsil etmektedir. v^k_{ij} k kümesinin i. elemanını ve v^k_j k. kümenin küme merkezini temsil etmektedir. Küme içerisindeki tüm verilerin küme merkezlerine uzaklıkları hesaplandıktan sonra veri noktalarının küme merkezine en uzak olandan küme merkezine en yakın olan doğru sıralanması gerçekleştirir.

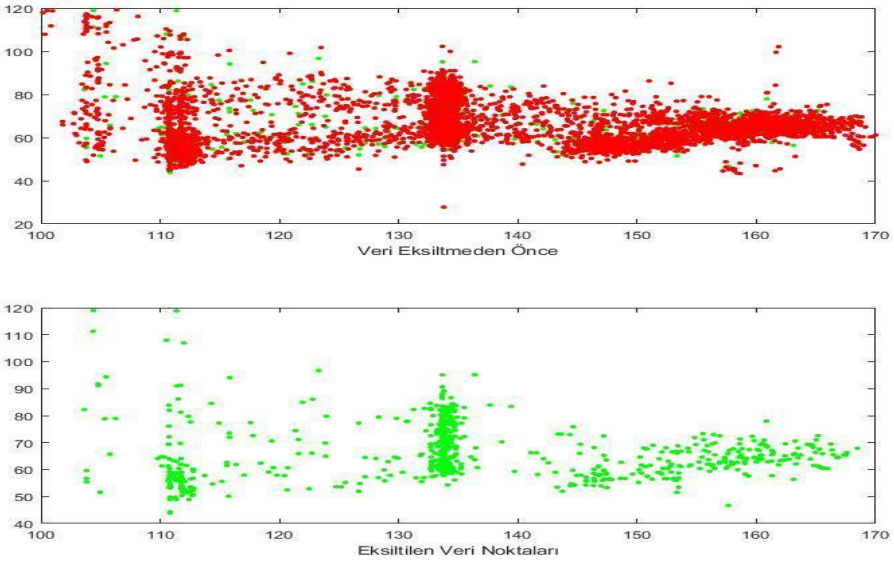
Bu işlemden sonra çıkarttırılacak (eksiltilecek, azaltılacak) veri noktaları seçilir. Bu seçim rastgele bir şekilde veya küme merkezine en uzak olan veriler olacak şeklinde veya küme merkezine en yakın veriler olacak şekilde yapılmak yerine küme merkezine uzaklıklarına göre sıralanmış veriler üzerinden bir atlama (belirli sayıda veri geçilerek sonraki verilere ulaşma) işlemi gerçekleştirilerek yapılır. Bu işlemin amacı küme merkezine en uzak veya küme merkezine en yakın verileri eksiltmeyerek veri düzleminin demografik yapısını (kümenin temsil yeteneği) bozmamaktır. Bu açıklama her bir veri setinde eksiltmeden önceki düzlemdeki verilerin gösterimi ve eksiltilen verilerin düzlemdeki durumu grafikler ile gösterilmiştir. Şekil 1'de gösterilmiştir.



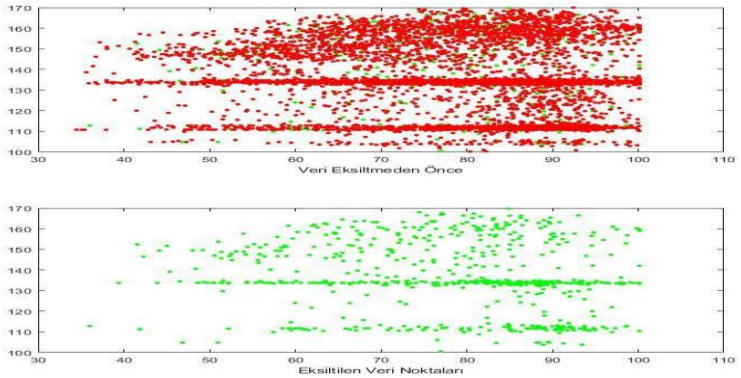
Şekil 1. *Enerji verimliliği veri seti veri eksiltmeden önceki ve sonraki dağılım grafiği.*



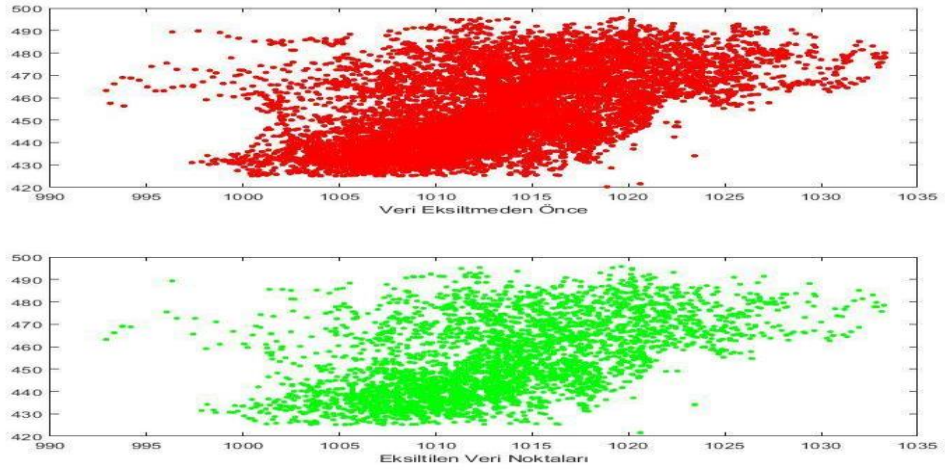
Şekil 2. *Beton başınç dayanımı veri seti veri eksiltmeden önceki ve sonraki dağılım grafiği.*



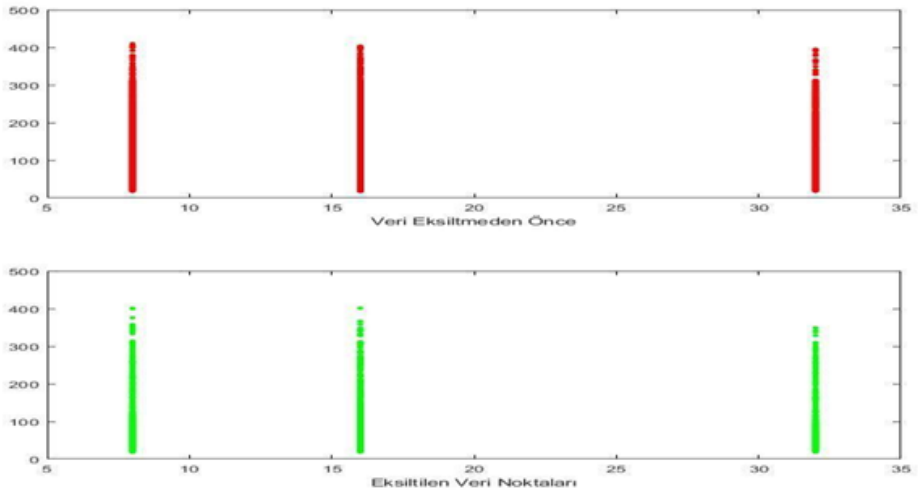
Şekil 3. *Parkinson hastalığı veri seti veri eksiltmeden önceki ve sonraki dağılım grafiği.*



Şekil 4. *Gaz tribünü emisyonu veri seti veri eksiltmeden önceki ve sonraki dağılım grafiği.*

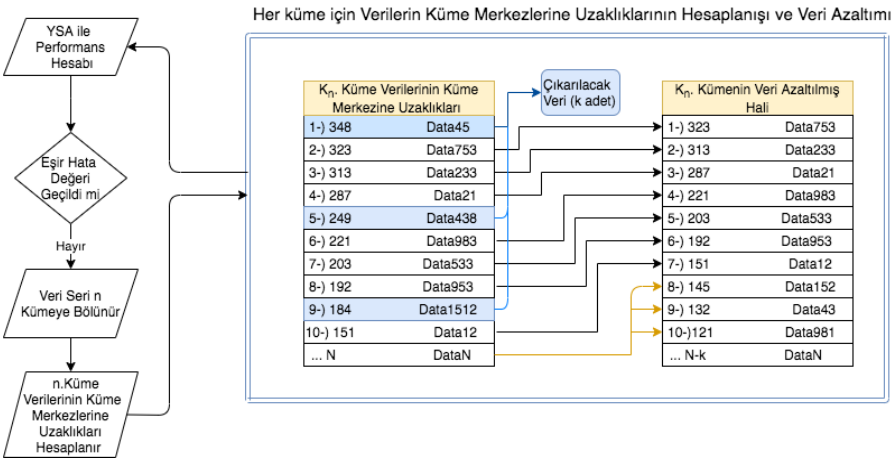


Şekil 5. *Kombine çevrim santrali veri seti veri eksiltmeden önceki ve sonraki dağılım grafiği.*



Şekil 6. *CPU çekirdek performansı veri seti veri eksiltmeden önceki ve sonraki dağılım grafiği.*

Şekil [1-6]'da gösterildiği gibi eksiltilecek(veri setinden çıkarılacak) veriler küme içerisinde belirli bir bölgeden seçilmeyip küme genelinden eksiltmeler yapılmıştır ve böylece küme demografisi bozulmadan korunmuştur.



Şekil 7. *Veri eksiltme döngüsü.*

Şekil 7'de KMUTOVA yönteminin veri eksiltme döngüsü görselleştirilmiştir. Yöntem içerisinde kümelere ayrılan veri seti üzerinden işlem yapılmaktadır. Ele alınan kümede veri azaltılması gerçekleştirilmiştir.

Algoritma 3. Küme Merkezine Uzaklık Tabanlı Orantılı Veri Azaltma Yöntemi Sözde Kodu	
1.	**Girişler:** Veri Seti
2.	**Çıkışlar:** Azaltılmış Veri Seti *D* veri seti tanımlanır
	Adımlar: 1. Veri seti n kümeye bölünür. 2. Seçilen küme verilerinin küme merkezine uzaklığı hesaplanır. 3. Küme verileri küme merkezlerine olan uzaklığa göre en uzak olandan en yakın olana doğru sıralanır. 4. Sıralamada bulunan noktada veri eksiltilir. 5. Küme içerisindeki veri sayısının eksiltme sayısına oranı kadar veri atlanır. **6.** Eksiltme oranı tamamlanmışsa eksiltilmiş veri seti döndürülür, tamamlanmamışsa 4. Adımdan işlemlere devam edilir.

III. DENEYSEL ÇALIŞMA

Bu bölümde uygulamada kullanılan veri setleri, geliştirilen yöntem kullanılarak veri setlerinde veri azaltma ve veri azaltıldıktan sonra altı farklı veri setinin tahmin başarısına bakılmaktadır.

A. AYARLAR

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

***Tablo 1**. Algoritmaların parametre değerleri.*

Algoritma	Parametre Değerleri
KMUTOVA	Atlama Oranı=(Küme Veri Sayısı) * %10
K-Ortalamalar	K(Küme Sayısı)=10
YSA	Gizli Katman Sayısı=1; Gizli Katman Hücre Sayısı=10, Hata Metriği= Ortalama Mutlak Yüzde Hata (MAPE),

B. VERİ SETLERİ

Çalışmada geliştirilen algoritma UCI Machine Learning [28] veri havuzundan temin edilen altı farklı veri seti üzerinde tatbik edilmiştir. Veri seti seçilirken tahmin problemine uygun bir veri seti olmasına, veri setlerinin farklı boyutlarda olmasına ve eksik/hatalı veri olmamasına dikkat edilmiştir.

Enerji Verimliliği veri setinde [29] taklit edilen 12 farklı bina şeklini kullanarak enerji analizi yapılmaktadır. Binalar, diğer parametrelerin yanı sıra cam alanı, cam alanı dağılımı ve yönelim

açısından farklılık gösterir. Veri seti, iki bağımlı nitelik değerli yanıtı tahmin etmeyi amaçlayan 768 örnek ve 8 bağımsız nitelik içerir. Bu çalışmada bağımlı niteliklerden sadece biri tahmin edilecek veri olarak kullanılacaktır.

Beton Basınç Dayanımı veri seti [30]: İnşaat mühendisliğinde en önemli malzeme betondur. Beton basınç dayanımı, beton yaşı ve bileşenlerin oldukça doğrusal olmayan bir fonksiyonudur. Beton Basınç Dayanımı veri seti 8 bağımsız nitelik ve 1 bağımlı nitelikten oluşur. Bağımsız nitelikler; çimento miktarı, yüksek fırın cürufu miktarı, uçan kül miktarı, su miktarı, süperakınlaştırıcı miktarı, kaba agrega miktarı, ince agrega miktarı ve yaşıdır. Yaş değişkeni gün ile ifade edilir, miktar bildiren nitelikler ise kilogram ile ifade edilir. Tahmin edilecek bağımlı nitelik ise beton basınç dayanımıdır.

Parkinson hastalığı veri seti [31] uzaktan semptom ilerleme izlemesi için bir tele-izleme cihazının altı aylık denemesine alınan erken evre Parkinson hastalığı olan 42 kişiden alınan çeşitli biyomedikal ses ölçümlerinden oluşur. Kayıtlar otomatik olarak hastanın evinde çekilmiştir. Veri setinin nitelikleri denek numarası, denek yaşı, denek cinsiyeti motor UPDRS, toplam UPDRS ve 16 biyomedikal ses ölçümü içerir. Verilerin temel amacı, 16 ses ölçüsünden motor ve toplam UPDRS puanlarını ('motor_UPDRS' ve 'toplam_UPDRS') tahmin etmektir. Bu çalışmada motor UPDRS tahmin edilecektir.

Gaz Tribünü Emisyonu veri seti [32] baca gazı emisyonlarını, yani CO ve NOx'i (NO NO2) incelemek amacıyla Türkiye'nin kuzey batı bölgesinde bulunan bir gaz türbininden bir saat boyunca (ortalama veya toplam yoluyla) toplanan 36733 adet sensör ölçümü örneğini içermektedir. Ortam değişkenlerine ek olarak gaz türbini parametrelerini (Türbin Giriş Sıcaklığı ve Kompresör Çıkış basıncı gibi) içerir. Veri seti, özellik olarak ortam değişkenlerini kullanarak türbin enerji verimini (TEY) tahmin etmek için iyi kullanılır.

Kombine Çevrim Santrali veri seti [33] santral tam yükte çalışmaya ayarlandığında 6 yıl boyunca (2006-2011) bir Kombine Çevrim Santralinden toplanan verilerden oluşan veri setidir. 4 bağımsız parametre saatlik ortalama, Sıcaklık (T), Ortam Basıncı (AP), Bağıl Nem (RH) ve Egzoz Vakumundan (V) oluşan ortam değişkenleridir. 1 bağımlı parametre ise tesisin net saatlik elektrik enerjisi (EP) çıkışıdır.

GPU çekirdeği performansı veri seti [34] bir SGEMM GPU çekirdeği kullanarak, A*B = C'nin çalışma süresini ölçer. Tüm zamanlar milisaniye cinsinden ölçülür. 14 bağımsız parametre (bağımsız nitelik) 1 bağımlı parametre (bağımlı nitelik) vardır. Bağımlı nitelik (Run) parametre A*B=C işleminin bağımsız parametreler doğrultusunda ölçüm süresidir.

***Tablo 2**. Veri setleri özellikleri.*

Veri Seti	Boyut	Eğitim Örnek	Test Örnek	Toplam Örnek
Enerji Verimliliği Veri Seti	9	691	77	768
Beton Basınç Dayanımı Veri Seti	9	927	103	1030
Parkinson Hastalığı Veri Seti	22	5299	576	5875
Gaz Tribünü Emisyonu Veri Seti	11	6670	741	7411
Kombine Çevrim Santrali Veri Seti	5	43056	4784	47840
GPU Çekirdek Performansı Veri Seti	15	58982	6553	65535

Tablo 2'de bu makalede geliştirilen yöntem dahilinde kullanılan veri setleri ile ilgili verinin boyutu, YSA eğitim aşamasında veri setinin kullanılan Eğitim Örnek Sayısı ve Test Örnek Sayısı, ayrıca verinin Toplam Örnek Sayısı bilgileri sunulmuştur.

A. ANALİZ SONUÇLARI

Deneysel sonuçlar geliştirilen yöntem ile belirtilen yüzdelerle veri azaltması yapıldıktan sonra 25 performans hesaplama tekrarı işleminin ardından ortaya çıkan sonuçlar arasında en iyi, en kötü ve ortalama değerleri, ayrıca sonuçların standart sapma değerlerini içermektedir. Karşılaştırma yapmak amacıyla Orijinal Veri (Eksiltme Yapılmadan Önceki Veri Seti) setinin performans değerleri 25 tekrar sonucunda ortaya çıkan en iyi, en kötü ve ortalama değerleri olarak Tablo 3'te sunulmuştur. Orijinal veriye %5 hata miktarı eklenerek algoritmanın orijinal veri setinin performans değerine göre %5 daha kötü tahminler üretmesi tolore edilebilir olarak sunulmuştur. Karşılaştırma hata oranı eklenmiş performans değeri ve eksiltme sonrası 25 tekrar sonucunda çıkan ortalama değer arasında yapılmıştır. Buna göre veri setinin orijinal performans değerine eklenen %5 hata oranı ekledikten sonra çıkan değer eksiltme sonucu 25 performans hesaplama tekrarının ortalamasından yüksek ise (daha iyi tahminler üretmişse) algoritma eksiltme oranında başaralı sayılarak koyu renkte işaretlenmiştir. Eksiltme işlemi orijinal veri seti üzerinden %10 eksiltme sonrası değerler hesaplanıp, %10 eksiltilmiş veri seti üzerinden %10 daha eksiltme yapılarak %20 eksiltme sağlanmış olur. Bu işlem sonucunda Orijinal veri setinin yaklaşık olarak %19'u eksiltilmiş olur. Bu işlem bir sonraki %10 eksiltme için de aynı şekilde işlemektedir. Deneysel sonuçlarda belirtilen yüzdeler %10 + %10 şeklinde oluşan yüzdelerdir.

Performans metriği olarak veri setinin yapay sinir ağlarının tahmin başarısını ölçmek amacıyla Ortalama Mutlak Yüzde Hata (MAPE) kullanılmıştır. Deneysel sonuçlar Tablo 3'te verildiği gibidir.

Tablo 3. *Eksiltilmemiş orijinal verinin performans değerleri.*

Orijinal Veri Seti	En İyi	En Kötü	Ortalama	Standart Sapma	Ortlama Değere %5 Hata Değeri Eklendikten Sonra Ort Değer
Enerji Verimliliği Veri Seti	2.258	5.398	3.232	0.822	3.393
Beton Basınç Dayanımı Veri Seti	15.547	23.220	18.639	2.331	19.570
Parkinson Hastalığı Veri Seti	7.598	39.463	16.689	8.385	17.523
Gaz Tribünü Emisyonu Veri Seti	0.344	0.401	0.379	0.014	0.397
Kombine Çevrim Santrali Veri Seti	0.670	0.702	0.683	0.007	0.717
GPU Çekirdek Performansı Veri Seti	11.193	16.486	13.713	0.822	14.398

Tablo 3'te verilen değerler orijinal(eksiltilmemiş-azaltılmamış) veri setinin 25 tekrar sonucu ortaya çıkan Ortalama Mutlak Yüzde Hata (MAPE) değerlerinin En İyisi, En Kötüsü, 25 MAPE değerinin Ortalaması ve 25 MAPE değerinin Standart Sapmaları verilmiştir. Ayrıca 25 tekrar sonucu elde edilen MAPE değerlerinin ortalamasına %5 Hata toleransı eklenerek kıyaslama yapılacak olan değeri(Hata toleransı eklenmiş Ortalama MAPE) sunulmuştur.

***Tablo 4**. Deneysel çalışma sonuçları.*

Veri Seti		Eksiltme Oranı->	%10	%20	%30	%40	%50	%60
Enerji Verimliliği Veri Seti	KMUTOVA	En İyi	2.282	2.331	2.319	2.302	2.291	2.593
		En Kötü	15.651	11.993	8.558	4.422	9.821	6.677
		Ortalama	4.294	**3.344**	3.670	**3.246**	3.809	**3.255**
		Std.Sapm	3.481	2.021	1.442	1.132	1.902	0.952
Beton Basınç Dayanımı Veri Seti	KMU-TOVA	En İyi	14.327	15.540	16.201	14.863	13.894	14.982
		En Kötü	27.983	23.747	24.998	21.849	26.816	24.729
		Ortalama	**19.377**	**19.061**	20.300	**19.271**	20.117	**19.492**
		Std.Sapm	3.961	1.915	2.507	1.586	2.946	2.618
Parkinson Hastalığı Veri Seti	KMU-TOVA	En İyi	9.115	9.316	7.929	8.998	7.843	6.156
		En Kötü	39.813	39.095	28.717	30.164	37.275	59.388
		Ortalama	20.829	20.175	**16.124**	**17.863**	18.736	20.449
		Std.Sapm	8.133	8.779	6.057	5.778	8.640	12.022
Gaz Tribünü Emisyonu Veri Seti	KMU-TOVA	En İyi	0.347	0.360	0.341	0.341	0.367	0.354
		En Kötü	0.445	0.428	0.421	0.428	0.613	0.466
		Ortalama	**0.385**	**0.386**	**0.384**	**0.387**	0.404	**0.394**
		Std.Sapm	0.022	0.017	0.019	0.020	0.052	0.030
Kombine Çevrim Santrali Veri Seti	KMUTOVA	En İyi	0.669	0.671	0.669	0.665	0.665	0.671
		En Kötü	0.695	0.693	0.693	0.691	0.694	0.700
		Ortalama	**0.681**	**0.681**	**0.680**	**0.681**	**0.682**	**0.684**
		Std.Sapm	0.007	0.005	0.004	0.006	0.006	0.008
GPU Çekirdek Performansı Veri Seti	KMUTOVA	En İyi	12.033	11.066	11.937	11.745	11.387	11.946
		En Kötü	17.744	22.316	16.361	38.766	17.810	19.761
		Ortalama	**13.951**	**14.028**	**13.624**	14.862	**13.766**	**14.235**
		Std.Sapm	1.499	2.367	1.084	5.706	1.572	1.800

Deneysel çalışma sonuçlarında koyu ile ifade edilen değerler KMUTOVA yöntemi uygulanması sonucu azaltılmış verinin ortalama MAPE değeri orijinal(azaltılmamış) veri setinin ortalama MAPE değerinden daha iyi sonuçları ifade etmektedir. Elde edilen sonuçlara göre KMUTOVA yöntemi ile azılmış veri setleri farklı oranlardaki veri azaltma yüzdelerinde orijinal veri setinin hata toleransı eklenmiş MAPE değerine göre daha düşük bir hata değeri sunmaktadır. Bu sonuçlar KMUTOVA yönteminin veri setlerini azaltırken orijinal veri setine göre değerli veriyi elde tuttuğunun önemli bir göstergesidir.

VI. SONUÇ

Bu makale çalışmasında veri azaltma amaçlı melez bir algoritma başarılı bir şekilde geliştirilmiştir. Geliştirilen algoritma büyük veri setlerinde veri sayısını azaltmak için yapay sinir ağlarından ve k-ortalamalar yönteminden faydalanmaktadır. Yapay sinir ağları veri setlerinin modellenmesinde performans değerlendirmesi için kullanılırken, k-ortalamalar algoritması ise veri setini benzerliklerine göre ayrılmış gözlemlerden oluşturma ve küme içindeki dağılımı homojen bir şekilde azaltma işlemlerinde kullanılmıştır. Geliştirilen Küme Merkezine Uzaklık Tabanlı Orantılı Veri Azaltma algoritmasının tahmin problemlerinde tahmin performansı üzerindeki etkisi incelenmiştir. Performans hata sınırları içerisinde veri azaltılması amaçlanmıştır. Bu amaca yönelik veri azaltılması için Küme Merkezine Uzaklık Tabanlı Orantılı Veri Azaltma algoritması kullanılmıştır. Sonuç olarak, altı farklı özellikteki veri seti üzerinde yürütülen deneysel çalışmalarda %10 - %50 arasında veri çıkarıldıktan sonra veri setlerinin temsil yeteneklerinde bir bozulma olmadığı, hatta gürültülü verilerin tespit edilmesi ve çıkarılması sayesinde az veriden oluşan veri setlerinin problem uzayını daha da başarılı bir şekilde temsil ettikleri görülmüştür. Bu durum, önerilen algoritmanın gürbüz bir veri azaltma yöntemi olduğunun açık bir işaretidir.

TEŞEKKÜR: Bu çalışma, TÜBİTAK 2209-A Üniversite Öğrencileri Araştırma Projeleri Destekleme Programı tarafından 2021 yılında desteklenmiştir. (Proje No:1919B012004718)

V. KAYNAKLAR

[1] HT. Kahraman, "A novel and powerful hybrid classifier method: Development and testing of heuristic k-nn algorithm with fuzzy distance metric," *Data & Knowledge Engineering*, c. 103, ss. 44-59, 2016.

[2] HT. Kahraman, B. Aras, & O. Yıldız. "Sınıflandırma Problemleri İçin Agde-Tabanlı Meta-Sezgisel Boyut İndirgeme Algoritmasının Geliştirilmesi," *Mühendislik Bilimleri ve Tasarım Dergisi*, c. 8, s. 5, ss. 206-217, 2020.

[3] F. Arslan, & HT. Kahraman. "Yapay zekâ tabanlı büyük veri yönetim aracı," *Journal of Investigations on Engineering and Technology*, c. 2, s. 1, ss. 8-21, 2019.

[4] Ö. Köroğlu, & HT. Kahraman. "K-Ortalamalar Tabanlı En Etkili Meta-Sezgisel Kümeleme Algoritmasının Araştırılması," *Mühendislik Bilimleri ve Tasarım Dergisi*, c. 8, s. 5, ss. 173-184, 2020.

[5] N. Gokilavani and B. Bharathi, "Test case prioritization to examine software for fault detection using PCA extraction and K-means clustering with ranking," *Soft Computing*, vol. 25, no. 7, pp. 5163-5172, 2021.

[6] M. Sivaguru and M. Punniyamoorthy, "Performance-enhanced rough *k* k-means clustering algorithm," *Soft Computing*, vol. 25, no. 2, pp. 1595-1616, 2021.

[7] Z. Wang, Y. Zhou, and G. Li, "Anomaly Detection by Using Streaming K-Means and Batch K-Means," *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*. IEEE, vol. 5, pp. 11-17, 2020.

[8] Y. Li, and H. Wu, "A clustering method based on K-means algorithm," *Physics Procedia* vol. 25, pp. 1104-1109, 2012.

[9] CU. Kumari, SJ. Prasad, and G. Mounika, "Leaf disease detection: feature extraction with K-means clustering and classification with ANN," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, pp. 1095-1098, 2019.

[10] VP. Murugesan, and P. Murugesan, "A new initialization and performance measure for the rough k-means clustering," *Soft Computing*, vol. 24, no. 15, pp. 11605-11619, 2020.

[11] OJ. Oyelade, OO. Oladipupo, and IC. Obagbuwa, "Application of k Means Clustering algorithm for prediction of Students Academic Performanc,." *International Journal of Computer Science and Information Security, IJCSIS*, Vol. 7, No. 1, pp. 292-295, 2010.

[12] M. Yedla, SR. Pathakota, and TM. Srinivasa, "Enhancing K-means clustering algorithm with improved initial center," *International Journal of computer science and information technologies* vol. 1, no. 2, pp. 121-125, 2010.

[13] BP. Koustubh, VV. Nair, and S. Kumaravel, "Anomaly Detection in Hybrid Electric Vehicles Using ANN Based Support Vector Data Description," *2018 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*. IEEE, pp. 14-24, 2018.

[14] A. Pannu, "Artificial intelligence and its application in different areas," *Artificial Intelligence*, vol. 4, no. 10, pp. 79-84, 2015.

[15] N. Kayarvizhy, S. Kanmani, and RV. Uthariaraj, "ANN models optimized using swarm intelligence algorithms," *WSEAS Transactions on Computers* vol. 13, no. 45, pp. 501-519, 2014.

[16] L. Cavallaro, "Artificial neural networks training acceleration through network science strategies," *Soft Computing* vol. 24, no. 23, pp. 17787-17795, 2020.

[17] H. Yaşar, "A novel approach for estimation of coronary artery calcium score class using ANN and body mass index, age and gender data," *2018 4th International Conference on Computer and Technology Applications (ICCTA)*. IEEE, pp. 184-187, 2018.

[18] J. Xu, "ANN based on IncCond algorithm for MPP tracker," *2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*. IEEE, pp. 129-134, 2011.

[19] S. Akhmedova, and E. Semenkin, "Co-operation of biology related algorithms meta-heuristic in ANN-based classifiers design," *2014 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, pp. 2207-2214,2014.

[20] S. Anitha, and M. Vanitha, "Optimal Artificial Neural Network based Data Mining Technique for Stress Prediction in Working Employees." *Soft Computing*, vol. 25, no. 17, pp. 11421-11428, 2021.

[21] T. Srivastaya, (October 20, 2014).How does Artificial Neural Network (ANN) algorithm work? [Online]. Avaiable: https://www.analyticsvidhya.com/blog/2014/10/ann-work-simplified/

[22] C. Yilmaz, HT. Kahraman and S. Söyler, "Passive mine detection and classification method based on hybrid model," *IEEE Access*, c. 6, ss. 47870-47888, 2018.

[23] R. Bayindir, I. Colak, S. Sagiroglu and HT. Kahraman, "Application of adaptive artificial neural network method to model the excitation currents of synchronous motors," *IEEE*, vol. 2, pp. 498-502, 2012.

[24] A. Radhika, and MS. Masood, "Effective dimensionality reduction by using soft computing method in data mining techniques," *Soft Computing* vol. 25, no. 6, pp. 4643-4651, 2021.

[25] T. Karin and D. Mondial, "Data Reduction and Deep-Learning Based Recovery for Geospatial Visualization and Satellite Imagery," *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, vol.16, no. 3, pp. 439-454, 2020.

[26] SL. Wong, BY. Ooi and SY Liew, "Data Reduction with Real-Time Critical Data Forwarding for Internet-of-Things," *2019 International Conference on Green and Human Information Technology (ICGHIT)*. IEEE, pp. 1-6, 2019.

[27] A. Moitra, NO. Malott and PA. Wilsey, "Cluster-based data reduction for persistent homology," *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 327-334, 2018.

[28] D. Dua and C. Graff , (2019) UCI Machine Learning Repository [Online]. Avaiable: http://archive.ics.uci.edu/ml

[29] T. Athanasios and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings* vol. 49, pp. 560-567, 2012.

[30] IC. Yeh, "Modeling of strength of high-performance concrete using artificial neural Networks," *Cement and Concrete research*. pp. 1797-1808, 1998.

[31] T. Athanasios."Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests," *Nature Precedings*. pp. 1-1, 2009.

[32] H. Kaya, P. Tüfekcin and E. Uzun, "Predicting co and no x emissions from gas turbines: novel data and a benchmark pems," *Turkish Journal of Electrical Engineering & Computer Sciences* vol. 27, no. 6, pp. 4783-4796, 2019.

[33] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *International Journal of Electrical Power & Energy Systems* vol. 60, pp. 126-140, 2014.

[34] B. Rafael, EG. Paredes and R. Pajarola, "Sobol tensor trains for global sensitivity analysis," *Reliability Engineering & System Safety* vol. 183, pp. 311-322, 2019.