JISTA

Journal of Intelligent Systems: Theory and Applications

SEPTEMBER 2023

ISSN: 2651-3927

VOL 6 NO 2

ARTIFICIAL INTELLIGENT > MACHINE LEARNING > DEEP LEARNING

https://dergipark.org.tr/en/pub/jista

Journal of Intelligent Systems:
Theory and Applications
2023 Volume: 6 Issue: 2

Editorial Boards

Journal of Intelligent Systems:
Theory and Applications
2023 Volume: 6 Issue: 2

Editorial Advisory Board

Editorial Assistants

https://dergipark.org.tr/en/pub/jista

Journal of Intelligent Systems: Theory and Applications

2023 Volume: 6 Issue: 2

Contents

Research Articles

1. **Yapay Sinir Ağları ve Derin Öğrenme Algoritmalarının Kripto Para Fiyat Tahmininde Karşılaştırmalı Analizi** 96-107
Müberra Beyza ODABAŞI, Merve CENGİZ TOKLU

2. **Binary Honey Badger Algorithm for 0-1 Knapsack Problem** 108-118
Gülşen ORUCOVA BÜYÜKÖZ, Hüseyin HAKLI

3. **Makine Öğrenmesi Teknikleri ile Counter-Strike: Global Offensive Raunt Sonuçlarının Tahminlenmesi** 119-129
Vahid SİNAP

4. **Analyzing Big Social Data for Evaluating Environment-Friendly Tourism in Turkey** 130-142
Mahmud ALRAHHAL, Ferhat BOZKURT

5. **Equilibrium Optimizer Algorithm for Optimal Reactive Power Dispatch** 143-151
Erdi DOĞAN

6. **Akıllı Ev Sistemleri için XGBoost Tabanlı Saldırı Tespit Yöntemi** 152-158
Rojbin TEKİN, Orhan YAMAN

7. **Makine Öğrenmesi Yöntemleriyle Anormal İçme Suyu Tüketimlerinin Tespit Edilmesi ve Tahmin Modellerinin Geliştirilmesi** 159-173
İsmail GÜNEY, İhsan Hakan SELVİ

8. **A Study of Ensemble Deep Learning Method Using Transfer Learning for Horticultural Data Classification** 174-180
Gökhan ATALI, Sedanur KIRCI

9. **Kampüs İçi Kapalı Alanlarda Hava Kalitesinin Modellenmesi ve Karar Destek Sistemi Geliştirilmesi** 181-190
Elif CESUR, Cemal EFE

10. **Self Adaptive Methods for Learning Rate Parameter of Q-Learning Algorithm** 191-198
Murat Erhan ÇİMEN, Zeynep GARİP, Yaprak YALÇIN, Mustafa KUTLU, Ali Fuat BOZ

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 6(2) (2023) 96-107
DOI: 10.38016/jista.1228629

Yapay Sinir Ağları ve Derin Öğrenme Algoritmalarının Kripto Para Fiyat Tahmininde Karşılaştırmalı Analizi

Müberra Beyza Odabaşı[1], Merve Cengiz Toklu[2*]

[1] Sakarya Üniversitesi, Endüstri Mühendisliği Bölümü, Sakarya, Türkiye

[2] Sakarya Üniversitesi, Endüstri Mühendisliği Bölümü, Sakarya, Türkiye

muberra.odabasi@ogr.sakarya.edu.tr, mervetoklu@sakarya.edu.tr

Öz

Gelişen teknolojinin sağladığı olanaklar sayesinde internet kullanımıyla gerçekleştirilen işlemlerde artış olmuş ve bu da verilerde artışa neden olmuştur. Bu durum işletmeler için verilerin güvenli bir şekilde saklanması, paylaşılması, kontrolünün sağlaması ve yönetilmesine yönelik yeni teknoloji ihtiyacı doğurmuştur. Bu kapsamda faydalanılabilecek güncel teknolojilerden birisi de blok zinciri (Blockchain) yapısıdır. Blok zinciri yapısı birçok alanda kullanılabilecek bir teknoloji olup günümüzde en popüler kullanım alanı kripto paralar üzerinde olmaktadır. Bu çalışmada önemli alt kripto para birimlerinden biri olan Polkadot kripto para birimi için tahminleme işlemi yapılması amaçlanmıştır. Yapılan çalışmada 20.08.2020 ve 27.02.2023 tarihleri arasındaki veriler kullanılmış olup, bu verilere göre çıktı değer olarak günlük ortalama Polkadot değerinin tahmin edilmesi amaçlanmıştır. Girdi değerleri için kümeler iki farklı şekilde oluşturulmuştur. İlk girdi değerlerinde; Polkadot YouTube arama sayısı, Polkadot Google arama sayısı ve Polkadot hacmi kullanılmıştır. İkinci girdi değerlerinde ise ilk girdi değerlerinden farklı olarak alt kripto paraların lideri Ethereum eklenmiştir. İki farklı girdi yapısından oluşan bu çalışmada Polkadot para birimi günlük ortalama değerlerinin tahminlenebilmesi için yapay sinir ağlarında çok katmanlı algılayıcılar ile derin öğrenme yöntemlerinden olan uzun kısa süreli bellek yapısı kullanılarak tahminleme çalışması yapılmıştır. Sonuçlar incelendiğinde elde edilen yapay sinir ağlarında 4 girdi kümesinden oluşan değerlerin 0,93 korelasyon katsayısı ile daha iyi sonuç verdiği belirlenmiştir.

Anahtar kelimeler: Yapay Sinir Ağları, Derin Öğrenme, Geri Yayılım Algoritması, Uzun Kısa Süreli Bellek, Blok zinciri, Polkadot, Ethereum, Tahminleme

Comparative Analysis of Artificial Neural Networks and Deep Learning Algorithms for Crypto Price Forecast

Abstract

Thanks to the opportunities provided by developing technology, there had an increase in the transactions carried out using the internet. This development also led to an increase in data. This situation created the need for new technology for businesses to store, share, control, and manage data securely. One of the current technologies that can be used in this context is the blockchain structure. The blockchain structure is a technology that can be used in many areas, and the most popular usage area today is cryptocurrencies. In this study, it is aimed to estimate Polkadot cryptocurrency, which is one of the essential sub-cryptocurrencies. In the study, the data between 20.08.2020 and 27.02.2023 are used. According to these data, it aimed to estimate the daily average Polkadot value as the output value. Clusters for input values are created in two different ways. In the first input values; number of Polkadot YouTube search, number of Polkadot Google search, and Polkadot volume are used. Unlike the first input values, Ethereum, the leader of the alt cryptocurrencies, is added in the second input value. In this study, which consists of two different input structures, to estimate the daily average values of the Polkadot currency, an estimation study is carried out using multi-layered sensors in artificial neural networks and a long-short-term memory structure, which is one of the deep learning methods. When the results are examined, it is determined that the values of 4 input sets in the obtained artificial neural networks gave better results with a correlation coefficient of 0.93.

Keywords: Artificial Neural Networks, Deep Learning, Back-propagation Algorithm, Long Short-Term Memory, Blockchain, Polkadot, Ethereum, Prediction

* Sorumlu yazar.
E-posta adresi: mervetoklu@sakarya.edu.tr

Alındı : 4 Ocak 2023
Revizyon : 21 Şubat 2023
Kabul : 1 Nisan 2023

1. Giriş (Introduction)

Blok zinciri teknolojisi, Nakamoto tarafından 2008 yılında "Bitcoin: A Peer-to-Peer Electronic Cash System" başlıklı çalışma ile dünyaya tanıtılan günümüzün önemli teknolojilerinden birisidir (Yavuz vd., 2020). Blok zinciri, zincir şeklinde birbirine bağlı art arda bloklardan oluşan merkeziyetsiz bir yapıya sahiptir. Her blok, işleme ait bilgiler taşımakta olup bir önceki bloğun da şifrelenmiş bilgilerini taşır. Blok zinciri teknolojisi bir veri tabanı olduğu için verilerin saklandığı birçok alanda kullanılabilir.

Blok zinciri teknolojisinde işlem adımları arada bir aracı olmadan yapılabildiğinden daha hızlı ve daha güvenli olmaktadır. Günümüzde özellikle sosyal platformların da etkisiyle kendini kripto paralar üzerinden duyurmuştur. Kripto paralar blok zincirinin bir ürünü olup, blok zincirini sadece kripto paralar üzerinden yorumlamak doğru bir yaklaşım olmayacaktır. Kripto paraların günümüzde finans sektöründe önemli bir hacme sahip olduğu ve talep oranlarının da arttığı görülmektedir. Finansal piyasada yatırım yaparken, uzun vadeli kar elde edebilmek yatırımcılar için önemlidir. Yatırımcılar alternatifler arasında risk ve kar analizi yaparak en az riski içeren seçeneklere yönelmektir. Bunun için yatırımcıların, yatırım yapacakları alternatif hakkında ve o alternatifin yönelimi hakkında bilgi sahibi olmaları alacakları riski en aza indirecektir.

Bu çalışmada alt kripto para birimlerinden olan Polkadot kripto para biriminin günlük ortalama değerini tahminlemek için 20.08.2020 ve 27.02.2023 tarihleri arasındaki veri seti kullanılmıştır. Kullanılan veri setinde iki farklı girdi kümesi oluşturulmuştur. İlk girdi kümesinde Polkadot hacim sayısı, Polkadot Google arama sayısı, Polkadot Youtube arama sayısı bulunmakta olup ikinci girdi kümesinde ise Polkadot hacim sayısı, Polkadot Google arama sayısı, Polkadot Youtube arama sayısı ve Ethereum günlük ortalama değerleri bulunmaktadır. Polkadot kripto para biriminin günlük ortalama değeri üzerinde tahminleme işlemi yapılmıştır. Ethereum önemli alt kripto para birimlerinden olup Polkadot kripto para birimine etkisini ölçmek için iki farklı girdili yapı kullanılmıştır. Çalışmada yapay sinir ağlarında (YSA) çok katmanlı algılayıcılar ve derin öğrenme yöntemlerinden uzun kısa süreli bellek ağları (LSTM) kullanılarak tahminleme yapılmıştır. Yapay sinir ağlarında dört girdili veri seti ile yapılan çalışmanın daha iyi sonuç verdiği belirlenmiştir.

Bu çalışmada önemli alt kripto para birimlerinden olan Polkadot kripto para birimi seçilmiştir. Polkadot kripto para birimi, diğer kripto para birimlerinden farklı olarak arzı sınırlı değildir. Polkadot kripto para birimi, genel ve özel şifreleme sistemleri kullanıp, transfer işlemlerinde bir dijital cüzdandan diğerine göndermeye izin veren yapısı vardır. Ayrıca Polkadot, geliştiricilere kendi projeleri için blok zinciri geliştirmesine imkân tanır. Sahip olduğu paralel zincir (parachain) sistemi sayesinde birden fazla işlemde ağda aksaklık olmadan işlemleri gerçekleşmesine imkân sağlar. Diğer blok zinciri teknolojilerinde olduğu gibi tüm ağ boyunca sıralı değil, her bir parça üzerinden paralel olarak işlenmelerine izin vererek işlemlerin verimini arttırır.

2. Literatür Araştırması (Literature Review)

Bu kısımda tahminleme üzerine yapılan çalışmalar incelenmiştir. Yatırıma zemin hazırlayan borsalarda doğru yatırımları yapabilmek için hareketli olan bu piyasa hakkında bilgi sahibi olunması gerekmektedir. Bu kapsamda, Kalyoncu (2020) çalışmasında borsa üzerinde tahminleme yapabilmek için k-en yakın komşu algoritması, LSTM, YSA ve ARIMA yöntemlerinden faydalanmış olup en iyi performansı LSTM yönteminin sağladığını belirlemiştir. Son zamanlarda finansal piyasalarda kripto paralar önemli yatırım araçları arasına girmiş olup doğru yatırımı yapabilmek de kritik bir konudur. Demirci (2021) yapmış olduğu çalışma ile LSTM, geçitli tekrarlayan birim ve tekrarlayan sinir ağlarından yararlanarak Ethereum, Bitcoin, Ripple para birimlerinde fiyat tahminlemesi yapmıştır. Kripto paralara olan talebin gün geçtikçe artması ticaret sektöründeki yerinin de merak edilmesine sebep olmuştur. Avşar (2020) kripto paraların uluslararası piyasayalar üzerindeki etkisini belirlemek amacıyla LSTM metodu ile kripto paralar üzerinde tahminleme için analiz bulgularını incelemiştir. Çalışmasında kripto paraların sadece dijital para olarak değerlendirmemesi gerektiği arkasında önemli bir teknolojinin yattığı ve farklı projeleri de desteklediklerini için uluslararası ticarette farklı süreçlerde dijital dönüşüme destek sağlayabileceğini göstermiştir. Baygıner (2022) yapmış olduğu çalışma ile yatırımcıların ne kadar bilgi sahibi olduğunu ve yatırımı etkileyen kriterleri incelemiştir.

Türkiye'de kripto paralar sosyal medyanın da etkisiyle hızla tanınmış olup, haraketli bir yapıda olması nedeniyle hızlı para kazanmak isteyen yatırımcıların ilgisini çekmiştir. Pamuk (2019) çalışmasında sosyal medya ile kripto para birimi fiyat dalgalanması arasında herhangi bir ilişki olup olmadığını makine öğrenme algoritmaları ile belirlemeye çalışmıştır. Çılgın vd. (2020) sosyal medyanın Bitcoin fiyatlarına olan etkisini Bayes, destek vektör makineleri, lojistik regresyon ve YSA yöntemlerinden faydalanılarak araştırmıştır. Köksal vd. (2021) Bitcoin ile ilgili Twitter platformunda yapılan yorumları lojistik regresyon ve Naive Bayes algoritmaları kullanarak duygu analizi çalışması gerçekleştirmiştir. Sonrasında günlük olumlu Bitcoin tweetleri ile Bitcoin açılış değerlerini kullanarak doğrusal regresyon ve rastgele orman regresyon yöntemleri ile kapanış değerleri için tahminleme çalışması gerçekleştirmiştir

Kripto paralarda yatırım yaparken etkili faktörlerin bilincinde yatırım yapmak avantajlı olacaktır. Bu kapsamda Deniz (2020) yaptığı çalışmada Brent petrol ile altının kripto paralar üzerindeki etkisini anlayabilmek için Granger nedensellik analizi ile eş

bütünleşme testi yapmış ayrıca etki-tepki ve varyans ayrıştırması gerçekleştirmiştir. Sel (2020) Bitcoin, Ripple, EOS, Tether ve TRON kripto para birimlerinin pandemi döneminde altın fiyatları üzerindeki etkisini vektör makineleri, çok değişkenli karar ağacı ve rasgele orman regresyon modeli ile belirlemeyi amaçlamıştır. Kartal (2020) çalışmasında k-star algoritması kullanarak makroekonomik değişkenlerin Bitcoin üzerindeki etkilerini incelenmiştir. Bu çalışmalara ek olarak, Evlimoğlu ve Güder (2021) kripto paralardaki balon durumunu incelemiş, geçmişte yaşanan ekonomik balonlar gibi kripto paralarda da balonlar olduğunu, farklılıklar açısından da kimilerinde arzın belirli bir çerçevede olması, kimilerinde ise temel değerlerin bilinmemesinin farklılık sebebi olduğunu ifade etmiştir.

Kripto paralarla ilgili fiyat tahminlemesinde YSA ve farklı tahminleme metotları kullanan çalışmalar literatürde yer almaktadır. Sakız ve Gencer (2017) Bitcoin için tahminleme çalışması yapmıştır. Yapılan çalışmada YSA kullanılarak tahminleme gerçekleştirmiş ve spekülasyona çok açık olan Bitcoin'in tahminleme yaparken klasik tahminleme yöntemleri ile hesaplamasının zor olduğunu belirlemiştir. Şahin (2018) 2012-2018 yılları arasındaki Bitcoin kapanış fiyatlarını baz alarak ARIMA yöntemi ve YSA ile tahminleme yapmıştır. YSA ile yapılan tahminlemenin ARIMA yöntemine göre daha iyi sonuç verdiğini belirlemiştir. Aras (2019) çalışmasında Bitcoin, Ethereum, Ripple ve Litecoin kripto para birimlerinde klasik zaman serileri ve YSA yöntemiyle tahminleme yapmıştır. Atlan vd. (2020) Bitcoin, Ripple ve Ethereum kripto para birimlerinin 24 saatlik, 1 haftalık ve 1 aylık verilerini kullanarak bulanık çıkarım sistemi, YSA ve polinomsal eğri uydurma, LSTM gibi farklı yöntemlerle ileriye yönelik tahminleme çalışması gerçekleştirmiştir. Salman (2020) Bitcoin kripto para birimi ile ilgili olarak teknik ticari göstergeler ve YSA ile fiyat tahminlemesi yapılmıştır. Sel vd. (2020) çalışmasında altın, gümüş, sterlin, euro gibi girdileri kullanarak 2013-2018 yılları arasında Bitcoin'in günlük kapanış değerleri ile tahminleme gerçekleştirmiştir. Akay vd. (2021) Ripple, Binance coin ve Ethereum kripto para birimlerinin 2020-2021 yılları arasındaki verileri ile YSA ve LSTM yöntemlerini kullanarak tahminleme yapmıştır. Günlük açılış değeri, günlük en düşük değer ve en yüksek değerleri girdi değişkenleri, günlük kapanış değeri ise çıktı değişken olarak alınmıştır. Hata ölçütlerinden de faydalanarak gerçeğe en yakın değerlerin YSA ile verildiği belirlenmiştir. Hayradi vd. (2022) destek vektör regresyon algoritması ile günlük Polkadot kapanış değerlerini kullanarak fiyat tahminlemesi yapmıştır.

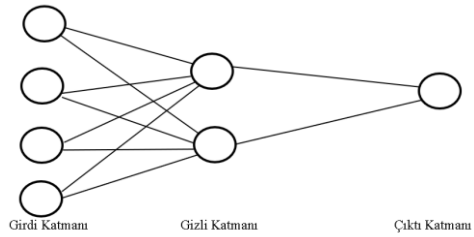
Bu çalışmada hem girdi olarak veri setleri hem de YSA ve derin öğrenme algoritmaları kıyaslanmıştır. Bu sayede verinin önemi vurgulanırken, farkı algoritmalar ile kıyaslama gerçekleştirilmiştir. Sonuç olarak yapılan çalışma ile dört girdili yapay sinir ağında çok katmanlı algılayıcılar için daha iyi sonuç verdiği gözlemlenmiştir.

3. Yöntem (Method)

3.1. Yapay sinir ağları (Artificial neural networks)

YSA öğrenme ve genelleme yapabilme özelliklerinin sağladığı esnek ve güçlü yapısı sayesinde pek çok karar verme sürecinde kullanılan bir yaklaşımdır. YSA, birçok özelliğine bağlı olarak farklı ve karmaşık problemleri çözmede etkili bir sistemdir. Her problemin çözümüne uygun farklı ağ yapıları vardır. Bu problemler için uygun çözüm ağının hangisi olduğuna karar vermek karar verici tarafından gerçekleşir. Bu durum çalışılan probleme göre değişkenlik gösterebilir (Karaatlı vd., 2012).

Teknik açıdan YSA'nın görevi, girdi katmanından gelen bilgileri ara katmanda işleyerek (ağın ağırlık değerleri kullanılarak) çıktıya dönüştürmektir (Yavuz ve Deveci, 2012). YSA'nın girdilerden doğru çıktılar üretebilmesi ağırlıklarının doğru değerler olmasına bağlıdır. YSA genel olarak girdi katman, gizli katman ve çıktı katmandan oluşmaktadır (Şekil 1).



Şekil 1. Yapay sinir ağı model yapısı (Artificial neural network model structure)

Daha geniş kapsamda, YSA yapısı beş temel süreç elemanından oluşmaktadır. Bunlar; girdi değerleri, ağırlıklar, toplama (birleştirme) fonksiyonu, aktivasyon fonksiyonu ve çıktı değerleridir. Girdi değerleri, yapay sinir ağına dışarıdan gelen veriler olup, bu verilerin ağ tarafından öğrenilmesi istenir. Ağırlıklar, girdi olarak gelen verinin hücre üzerindeki etkisini anlamaya yardımcı olmaktadır. Birleştirme fonksiyonu, bir hücredeki net girdiyi hesaplamakta olup, bunun için geliştirilmiş farklı fonksiyonlar bulunmaktadır. Aktivasyon fonksiyonu, net girdi sinyallerini çıkış sinyallerine çevirir. Çıktı katmanına aktarılan değerler üzerinde aktivasyon işlemi yapılmaktadır.

3.2. Derin öğrenme (Deep learning)

Derin öğrenme, en az bir adet yapay sinir ağının kullanıldığı insan beyninin bilgi edinme şeklini taklit eden makine öğrenme türüdür. Nöronlar arası sinyal iletimi ile birbirine bağlı hücreler gibi davranmasını sağlayarak, farklı durumları öğrenip karar sürecini destekleyen nöron bütünü yapay sinir ağını oluşturmaktadır. Tüm derin öğrenme modellerinde yapay sinir ağı olup, her yapay sinir ağının derin öğrenme modeli yoktur. Derin öğrenme algoritmaları en az iki katmana sahiptir, girdi çıktı katmanları da eklendiğinde durumda toplam dört katmandan

oluşmaktadır. Katmanlarda farklı nöron sayıları ve aktivasyon işlemleri olabilmekte olup, her düğümde sonucu optimize edebilmek için ağırlıklar eğitilmektedir.

Phyton pogramlama dilinde yapay sinir ağını oluşturmak için kullanılan iki kütüphane Tensorflow ve Pytorch'tur. Tensorflow kütüphanesine ait Keras sınıfı girdilerin nöronları beslediği ve bu nöronların çıktı sağladığı bir yapı oluşmasını amaçlar. Yapay sinir ağı oluşturmanın farklı yolları vardır ancak Keras sınıfı, katmanlar üzerinde kontrol ve esneklik sağlarken çok girdili ve çıktılı değerleri oluşturmakta kullanılabilmektedir. Eğitim ve test aşamasında optimize edici (optimizer), kayıp değer (loss) ve metrikler tanımlanmakta olup, optimize ediciler arasında adaptive moment estimation (ADAM) algoritması en sık kullanılandır (Turan, 2019). Derin öğrenmede kullanılan farklı algoritmalar aşağıdaki gibi özetlenebilir.

Evrişimli sinir ağları (CNN), ileri beslemeli sinir ağı modeli olup nesne algılama ve görüntü işlemede kullanılmaktadır. Tekrarlayan sinir ağları (RRN), LSTM ağlarından gelen çıktı değerlerinin, girdi olarak girilmesine izin verir ve dahili belleği sayesinde bir önceki girdileri ezberler böylece döngü oluşmasını sağlayan yapılar oluşturur. Uzun kısa süreli bellek ağlarına (LSTM), bu çalışmada kullanıldığından dolayı daha geniş yer verilmiştir. Uzun kısa süreli bellek, tekrarlayan sinir ağı türüdür. Diğer tekrarlayan sinir ağlarından farklı olarak uzun süre hatırlayabilen bellek hücreleri vardır. RNN'nin uzun vadede bellekteki bilgiyi tahmin etmekte zorlandığı, yakın tarihli işlemlerde daha iyi sonuç verdiği gözükmektedir. Bunlardan farklı olarak LSTM yapısı bilgiyi uzun süre saklayabilmektedir. Geleneksel sinir ağlarında girdi alınır ve bir çıktı üretir. LSTM yapısında ise veriler tekrarlayan şekilde işlenir, ilk adımda girdi alınır ve bu sonraki adımlarda çıktıyı etkilemek için kullanılabilir. Bu yineleme işlem adımları sayesinde LSTM yapılarında veri dizileri öğrenilebilir. Sıralı olan verilerde uzun vadeli verileri LSTM yapısı öğrenip bellek hücrelerinde tutmaktadır. Bellek hücrelerinde üç adet kapı bulunmakta olup bunlar giriş, çıkış ve unutma kapısıdır. Bellek hücresi ve hücre durumu LSTM yapısını oluşturmaktadır. Giriş kapısında belirli zaman aralığında hücre durumuna ne kadar yeni bilgi eklendiği kontrol edilir. Bu sigmoid işlevi kullanılarak düzenlenir, belirli zamandaki giriş ve önceki hücre çıkışı kullanılarak hatırlanacak veriler filtrelenir. Tahn işlevi ile (-1) ve (+1) arası çıktı veren olası tüm değerleri içeren bir vektör oluşturulur. Daha sonraki işlemde gerekli bilgileri elde etmek için vektör değerleri ve düzenlenmiş değerler çarpılır. Çıkış kapısında ise bu zaman aralığında çıktıyı üretirken hücre durumunda ne kadar bilgi kullanıldığını kontrol edilir. Unutma kapısında, geçmiş zaman adımlarında ne kadar bilginin mevcut zaman adımlarında tutulacağı kontrol edilir. Hücre durumunda artık işe yaramayan bilgiler unutma kapısına kaldırılır. Çok Katmanlı Algılayıcılar (MLP), ileri beslemeli sinir ağı olup birçok algılayıcı katmana sahiptir. Giriş katmanını besleyen çok katmanlı algılayıcılar, nöron katmanları sayesinde sinyali tek yönde geçecek şekilde grafiklere bağlar. Çok katmanlı algılayıcılar, girdi katmanla ile gizli katman arasındaki ağırlığın hesaplanmasında kullanılır. Radyal Temelli Fonksiyon Ağları (RBFN), ileri beslemeli yapay sinir ağı olup aktivasyon işlemlerinde radyal tabanları işlevleri kullanmaktadır. Giriş katmanını besleyen giriş vektörüne sahiptir. Sınıflandırma, zaman serisi, regresyon analizi gibi birçok kısımda kullanılmaktadır. Kendi Kendini Düzenleyen Haritalar (SOM), veri boyutlarında küçülmeye giderek veri görselleştirmeyi sağlayarak, yüksek boyutlu görselleştirilemeyen verilere anlam kazandırmaktadır. En olası girdi vektöründe hangi ağırlık olduğunu bulmak için tüm düğümler incelenir ve hak eden düğüm en iyi eşleşen birim olarak isimlendirilir. En iyi eşleşen düğüm çevresini inceler böylece sayı git gide azalır. Bir düğüm en iyi eşleşen birime ne kadar yakın ise ağırlık o kadar fazla değişmektedir. Üretken Düşman Ağları (GAN) sahte veri üretmeyi öğrenen oluşturucu ve yanlış bilgiyi öğrenen ayrımcı olmak üzere iki kısımdan oluşur. Eğitim verilerine benzeyen veri setleri oluşturmaktadır. Sahte veri ile gerçek veri arasında ayrım yapmayı öğrenir. Eğitim sırasında sahte veriler üretilir ayrımcı da bunların yanlış olduğunu söylemeyi öğrenir. Kısıtlanmış Boltzmann Makineleri (RBM), girdi değerleri üzerinden olasılık dağılımını öğrenmekte olup regresyon, boyut azaltma, konu modelleme, iş birlikçi filtreleme gibi alanlarda kullanılmaktadır. Derin İnanç Ağları (DBN), katmanlar arası bağlantıları olup, Boltzmann makinesi yığınıdır. Her katman önceki ve sonraki ile iletişim kurar. Derin inanç ağlarını eğiten aç gözlü algoritması olup, üretken ağırlıkları öğrenmek için katman katman yaklaşım uygular. Otomatik kodlayıcılar, ileri beslemeli sinir ağı olup denetimsiz öğrenme problemlerini çözmek için geliştirilmiştir.

4.Uygulama (Application)

Bu çalışmada Polkadot kripto para birimi için YSA'da fiyat modellemesi yapılmıştır. Polkadot para biriminde gün bazlı tahminleme yapabilmek için YSA'da çok katmanlı algılayıcılar ve derin öğrenme yöntemlerinden LSTM yapısı kullanılarak geliştirme yapılmıştır. Bu çalışmada 20.08.2020 ve 27.02.2023 tarihleri arasındaki veriler kullanılmış olup, nöron sayısı, test ve eğitim verilerindeki sayılar seçilirken çapraz doğrulama yönteminden yararlanılmıştır. Çıktı değeri olarak Polkadot'un günlük ortalama değeri alınmıştır. Ortalama değeri hesabında ilgili güne ait en yüksek ve en düşük değerlerin ortalaması alınmıştır. Girdi parametreleri iki farklı şekilde test edilmiştir. Bunun sebebi Polkadot kripto para biriminin Ethereum ağından türemiş olmasından dolayı Ethereum kripto para biriminin Polkadot üzerindeki etkisini görebilmektir. Bu nedenle ilk girdi parametreleri; Google arama sayısı, YouTube arama sayısı ve Polkadot

hacim değeri olarak belirlenmiştir. İkinci girdi parametreleri ise; Google arama sayısı, YouTube arama sayısı, Polkadot hacim değeri ve Ethereum ortalama değeri olarak belirlenmiştir.

Yapılan çalışma ile Polkadot ortalama değerinin, Ethereum ortalama değeri dikkate alınarak ve alınmayarak iki farklı şekilde tahminlenmesi amaçlanmıştır. Bunun için yapay sinir ağlarında çok katmanlı algılayıcılar ile derin öğrenmede LSTM yapısı kullanılarak başarı oranı korelasyon katsayısı üzerinden değerlendirme yapılarak karşılaştırılmıştır.

4.1. Yapay sinir ağlarında ağ tasarımı (Network design in artificial neural networks)

Bu çalışmada belirlenen parametreler doğrultusunda, Polkadot fiyatlarının yer aldığı veri seti kullanılmıştır. Polkadot fiyatlarını tahmin etmede kullanılan bir takım girdi parametreleri bulunmaktadır. Parametrelerde günlük değerler kullanılmış olup Ethereum ve Polkadot fiyatların da günlük en yüksek ve en düşük fiyatın ortalaması alınarak Polkadot fiyatında tahminleme işlemi yapılmıştır. Tahminleme işlemi yapılmadan önce veri seti üzerinde normalizasyon işlemi yapılmıştır.

Yapay sinir ağını modellerken aktivasyon fonksiyonu olarak sigmoid (Logsig) fonksiyonu kullanılmıştır. Daha sonra normalize edilen verilerin transpozesi alınmıştır. Hem doğrusal hem doğrusal olmayan fonksiyonlarda çıktı üretilebiliyor olmasından dolayı sigmoid fonksiyonu tercih edilmiştir. Denklem (1)'de sigmoid fonksiyonun formülü verilmiştir. Bu denkleme bağlı olarak elde edilen normalize edilmiş veriler Tablo 1'de gözükmektedir.

$$X_{norm} = (X - X_{min})/X_{max} - X_{min} \quad (1)$$

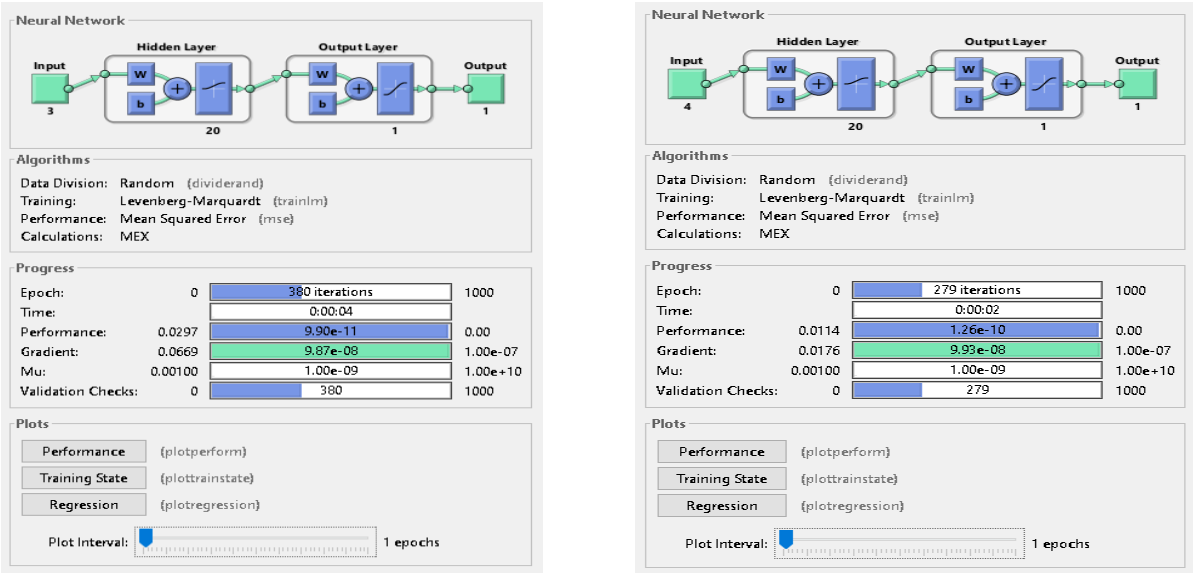
Tablo 1. Normalize edilmiş veriler (Normalized data)

Ethereum NM Data	Polkadot NM Data	NM Hacim	NM Youtube	NM WEB
0,63828023	0,82115869	0,1476941	0,7012987	0,20588235
0,73648822	0,73047859	0,26444834	0,06493506	0,29411765
0,64567961	0,60957179	0,17104495	0,28571429	0,30882353
0,72273337	0,72292191	0,18661218	0,46753247	0,22058824
0,81341916	0,91939547	0,32282545	0,07792208	0,13235294
0,944561	0,96473552	0,13485114	0,44155844	0,20588235
1	0,95717884	0,12239735	0,4025974	0,16176471

*Tüm veriler kısıtlı alan dolayısıyla gösterilemediği için verilerin bir kısmı örnek olarak verilmiştir.

Polkadot kripto para biriminin günlük ortalama değeri üzerinde tahminleme yapabilmek için, simülasyon ve eğitim verilerinden oluşan girdi ve çıktı kümeleri oluşturulmuştur. Yapay sinir ağlarında çok katmanlı algılayıcılarda yapılan geliştirmede MATLAB programından yararlanılmıştır. Yapay sinir ağlarında birçok model bulunmakta olup, bu çalışmada geri yayılım algoritması kullanılarak bir tahminde bulunulduğu için "feed-forward backdrop" tipi seçilmiştir. Eğitim fonksiyonu (training function) olarak, tahminlemede en çok kullanılan "traingdx" tercih edilmiştir. "Adaptation learning function" olarak "learngdx" seçilmiştir. Tahmin verilerinde daha iyi sonuç verdiği için katman sayısı iki olarak belirlenmiştir. Tasarlanan ağda bulunan yapay sinir ağı hücreleri için birleştirme fonksiyonu olarak toplama fonksiyonu kullanılmıştır. Aktarma fonksiyonu olarak ara katmanda bulunan yapay sinir hücreleri için sigmoid fonksiyonu, çıktı katmanında bulunan yapay sinir hücreleri içinse, doğrusal fonksiyon kullanılmıştır. Ağın eğitiminde eğitmenli öğrenme dizisi ile hatayı geriye doğru yayan "Levenberg-Marquardt" geri yayılım algoritması kullanılmıştır.

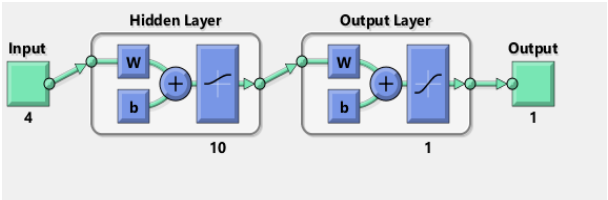
Polkadot para birimini tahminlemek için yapay sinir ağları çok katmanlı algılayıcı algoritması üzerinde yapılan geliştirmede, iki ayrı kümede girdi parametreleri oluşturularak buna göre ağ yapısı oluşturulmuştur. Oluşturulan ilk ağ yapısında üç parametreden oluşan Polkadot Google arama sayısı, Polkadot YouTube arama sayısı, Polkadot hacmi girdi değerleri, Polkadot günlük ortalama değeri ise çıktı değerlerini vermektedir. Yapay sinir ağları çok katmanlı algılayıcı algoritması üzerinde oluşturulan ikinci ağ yapısında, dört girdi parametreden oluşan Polkadot Google arama sayısı, Polkadot YouTube arama sayısı, Polkadot hacmi, Ethereum günlük ortalama değeri girdi değerlerini, Polkadot günlük ortalama değeri ise çıktı değerini vermektedir. Geliştirme test edilirken çapraz doğrulama yapılmış olup 10 ve 20 olmak üzere sırayla farklı nöron değerleri ve %70, %80 ve %90 eğitim verisi olacak şekilde karşılaştırma testi yapılmıştır. Üç girdili ve dört girdili tasarlanan ağın yapısı şekil 2'de gösterilmiştir.



(a) (b)

Şekil 2. (a) Üç girdili tasarlanan ağ yapısı (Network structure designed with three inputs). (b) Dört girdili tasarlanan ağ yapısı (Network structure designed with four inputs)

Şekil 3'te dört girdili 10 nöronlu tasarlanmış ağ yapısı bulunmaktadır.



Şekil Hata! Belgede belirtilen stilde metne rastlanmadı.**.** Dört girdili 10 nöronlu tasarlanan ağ yapısı (Network structure designed with 10 neurons and four inputs)

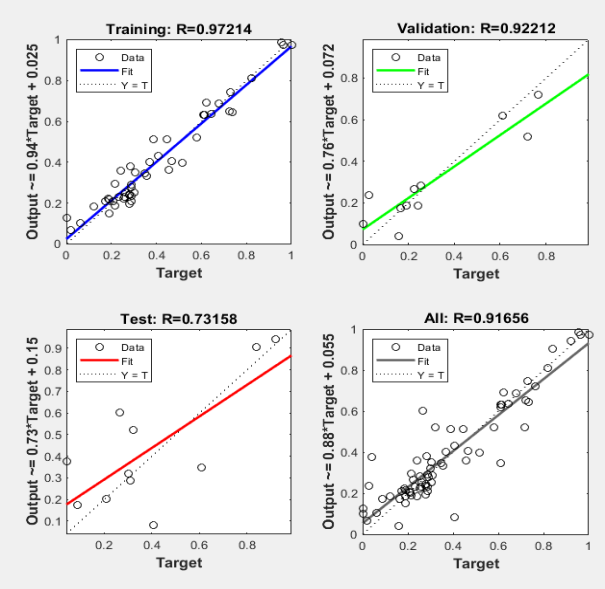
4.2 Yapay sinir ağlarının eğitimi ve test edilmesi (Training and testing of artificial neural networks)

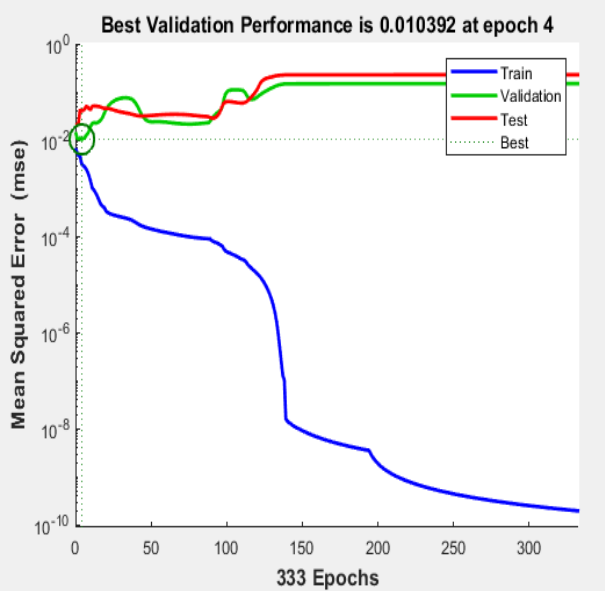
Ağın eğitilmesi aşamasında dört girdili ve üç girdili olmak üzere iki farklı girdi kümesi oluşturulmuştur.

Çıktı değerlerinde hangi değerlerde daha iyi sonuç verdiğini gözlemlemek için çapraz doğrulama yöntemi kullanılmıştır. Bunun için 10 ve 20 nöron değerleri seçilerek test edilmiştir.

Ayrıca 20.08.2020 ve 27.02.2023 tarihleri arasındaki veriler için %70, %80 ve %90 oranlarında eğitim verisi olacak şekilde test edilmiştir. Test sonuçlarına göre, 10 nöronlu %70 eğitim seti oranına göre "Traning R" değeri 0,97214, "Validation R" değeri 0,92212, "Test R" değeri 0,73158 ve "All R" değeri 0,91656 olarak bulunmuştur (Şekil 4).

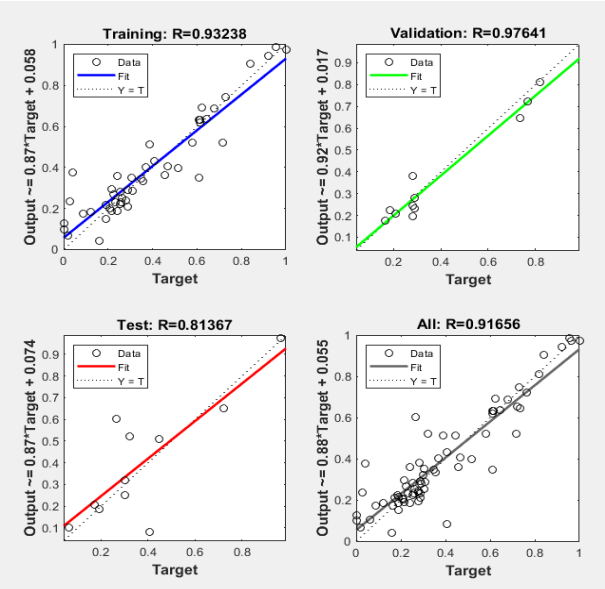
Şekil 5'te, 3 girdi parametresi, 10 nöron ve %80 oranındaki eğitim seti için ilgili veriler ve grafikler verilmiştir. Benzer şekilde Şekil 6'da 3 girdi parametresinden oluşan veri setinde, 10 nöronlu %90 eğitim seti kullanıldığında elde edilen veriler ve grafikler verilmiştir.

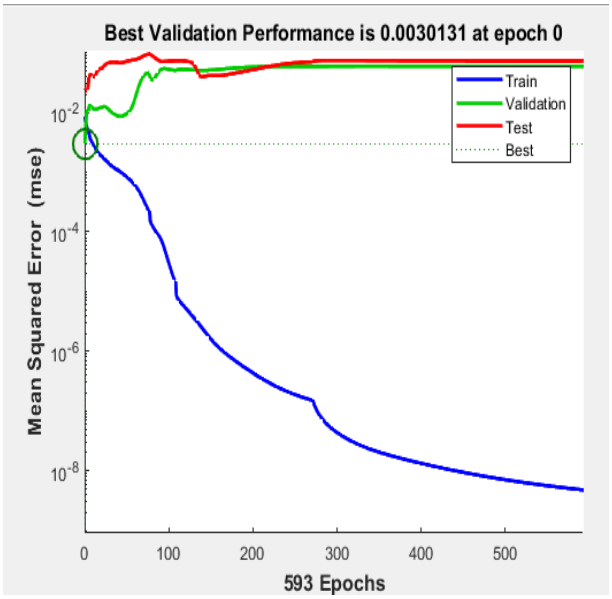




(a) (b)

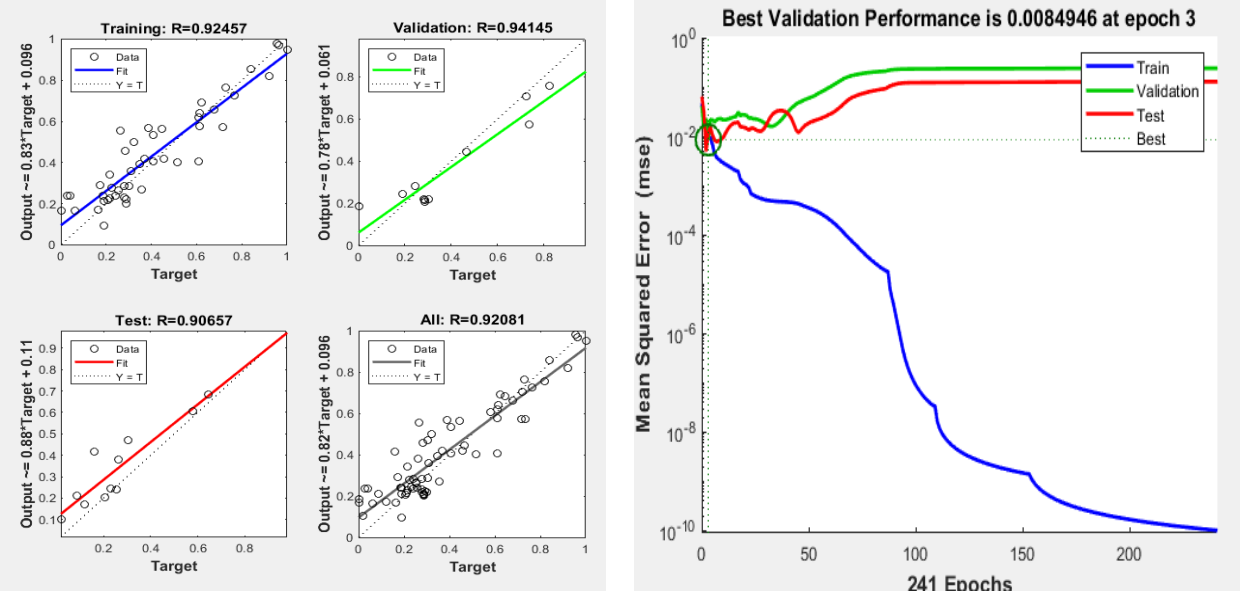
Şekil 1. (a) 3 girdili 10 nöronlu %70 eğitim verili gerçek-tahmin değeri grafiği (Predicted versus actual values scatter plot with 3 input, 10 neuron and 70% training data) (b) 3 girdili 10 nöronlu %70 eğitim verili Epoch-MSE grafiği (Number of epochs vs. MSE of 3 input, 10 neuron and 70% training data)





(a) (b)

Şekil 5. (a) 3 girdili 10 nöronlu %80 eğitim verili gerçek-tahmin değeri grafiği (Predicted versus actual values scatter plot with 3 input, 10 neuron and 80% training data) (b) 3 girdili 10 nöronlu %80 eğitim verili Epoch-MSE grafiği (Number of epochs vs. MSE of 3 input, 10 neuron and 80% training data)



(a) (b)

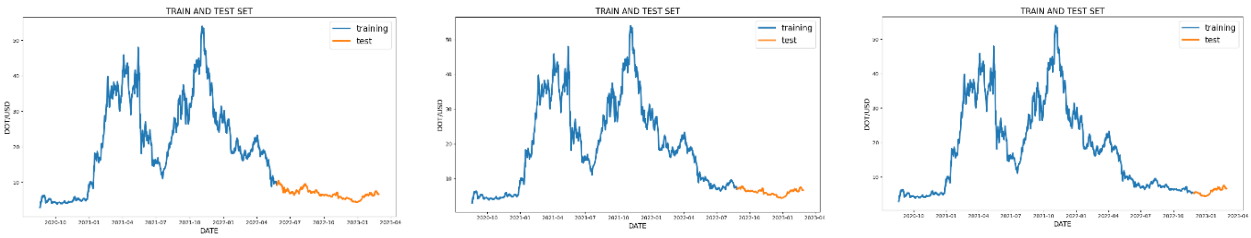
Şekil 6. (a) 3 girdili 10 nöronlu %90 eğitim verili gerçek-tahmin değeri grafiği (Predicted versus actual values scatter plot with 3 input, 10 neuron and 90% training data) (b) 3 girdili 10 nöronlu %90 eğitim verili Epoch-MSE grafiği (Number of epochs vs. MSE of 3 input, 10 neuron and 90% training data)

10 nöronlu işlem adımlarında uygulanan adımlar 20 nöron için de gerçekleştirilmiştir. 3 girdili 20 nöronlu %70 eğitim verisi kullanılarak "Training R" değeri 0,94048, "Validation R" değeri 0,78926, "Test R" değeri 0,9475, "All R" değeri ise 0,91059 olarak belirlenmiştir. 3 girdili 20 nöronlu, %80 eğitim verisi kullanılarak "Training R" değeri 0,95548, "Validation R" değeri 0,97673, "Test R" değeri 0,3744, "All R" değeri ise 0,91059 olarak tespit edilmiştir. 3 girdili 20 nöronlu, %90 eğitim verisi kullanılarak "Training R" değeri 0,88966, "Validation R" değeri 0,91949, "Test R" değeri 0,97199 ve "All R" değeri 0,91059 olarak belirlenmiştir. Eğitim veri seti oranı arttıkça başarı oranı da artmaktadır. Aynı adımlar 4 girdili veri seti için gerçekleştirildiğinde 10 nöron %90 eğitim seti için daha iyi sonuç verdiği gözlemlenmiştir. Sonuçlar incelendiğinde yapay sinir ağlarında 4 girdili 10 nöronlu %90 eğitim setinin olduğu 0,93944 değeri en iyi sonucu verdiği gözlemlenmiştir.

4.3 Derin öğrenme ile tahminleme (Prediction with deep learning)

Derin öğrenme, insan beyni işlevinden ilham almış yapay sinir ağlarını kullanan algoritmalarla alakalı makine öğreniminin bir alt dalıdır (Tuncer, 2022). Derin öğrenme yapay zekânın eğitilmesine olanak sağlayarak ilgili girdi değerlerine göre çıktı değerlerinin tahmin edilmesini sağlamaktadır.

Bu çalışmada derin öğrenme yöntemlerinden LSTM kullanılarak, Anaconda Navigator uygulaması üzerinden Jupyter Notebook kullanılarak Python 3 programlama dilinde geliştirme yapılmıştır. Polkadot fiyat tahmini için yapılan geliştirmede farklı kütüphaneler kullanılmıştır. Bunlar; numpy, pandas, matplotlib, sklearn, tensorflow kütüphaneleridir. Şekil 7'de, %70 eğitim seti, %80 eğitim seti ve %90 eğitim setine göre tarih bazlı test ve eğitim verilerinin dağılımı verilmiştir.

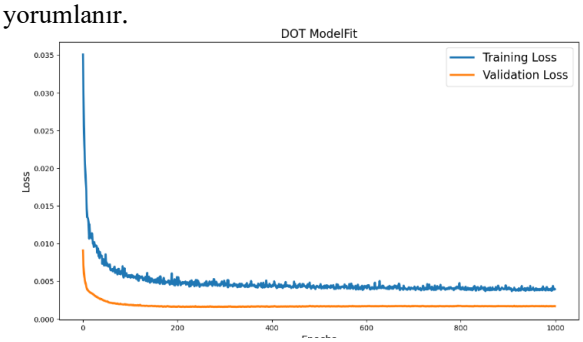


(a) (b) (c)

Şekil 7. (a) %70 eğitim verili test ve eğitim verilerinin tarih bazlı dağılımı (Date-based distribution of test and training data with 70% training data) (b) %80 eğitim verili test ve eğitim verilerinin tarih bazlı dağılımı (Date-based distribution of test and training data with 80% training data) (c) %90 eğitim verili test ve eğitim verilerinin tarih bazlı dağılımı (Date-based distribution of test and training data with 90% training data)

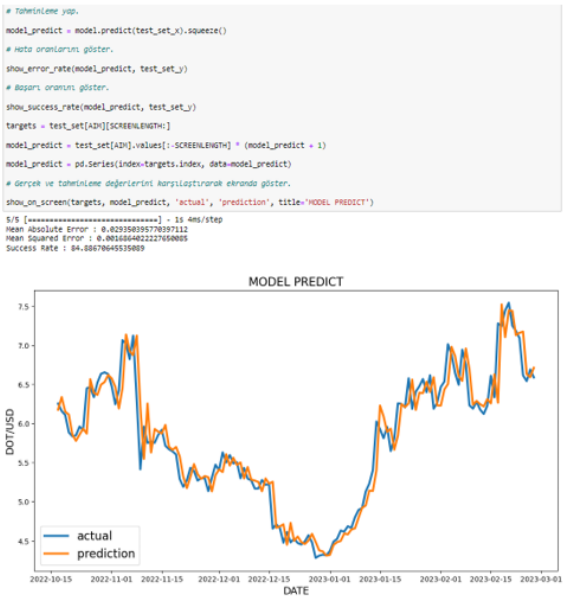
RNN, tekrarlayan sinir ağı olup, işlemlerin daha kalıcı olması için kendi içinde bir döngü yapısında çalışmaktadır. Veri setinde bulunan değer nörona verildiğinde tahmin üretilmektedir. Bu tahmin daha sonra gelen veri ile tekrar giriş yapılıp, tekrar kullanılmaktadır. Kısaca çıkan sonuç bir sonrakini beslemektedir. LSTM ise tekrarlayan sinir ağlarının farklı bir sürümü olup, hafızadaki geçmiş verileri hatırlamayı sağlar. LSTM modeli geri yayılım algoritması olarak eğitmekte olup, zaman serileri ile sınıflandırmada tahmin edilmesinde uygundur. Bu nedenle bu çalışmada LSTM tercih edilmiştir.

Şekil 8'de 10 nöronlu %70 eğitim verisi ile epochs değerlerine (döngü sayısı) göre ortalama karesel hata (Mean Squared Error-MSE) değeri gösterilmektedir. MSE değeri sıfıra ne kadar yakınsa o kadar başarılı diye yorumlanır.



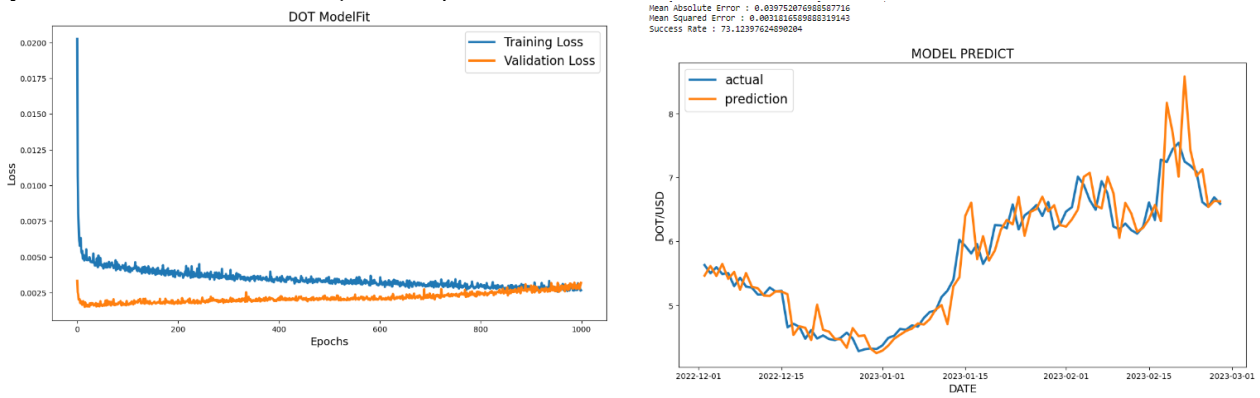
Şekil 8. Epochs-MSE grafiği (Epochs-MSE chart)

Şekil 9'da 10 nöronlu %80 eğitim verisi ile gerçek-tahmin grafiği ve hata/başarı oranları bulunmaktadır. Ayrıca ortalama karesel hata (Mean Squared Error-MSE), ortalama mutlak hata (Mean Absolute Error-MAE) ve başarı oranlarının yer aldığı değerler mevcuttur.



Şekil 9. Dört girdili gerçek-tahmin grafiği ve hata/başarı oranları (Actual-prediction graph and error-success rates)

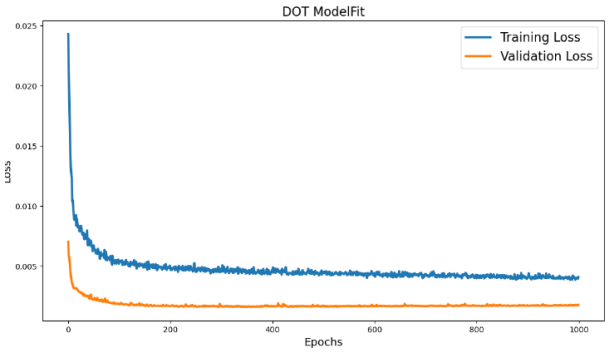
Şekil 10'da üç girdili %70 eğitim seti ile oluşturulmuş grafikler gözükmektedir. Şekil 10(a) Epoch-MSE grafiğini verirken, Şekil 10(b)'de gerçek ve tahmin değeri grafiği gözükmektedir. Ayrıca Şekil 10(b)'de yer alan başarı oranına bakıldığında 73,123 çıktığı gözükmektedir.

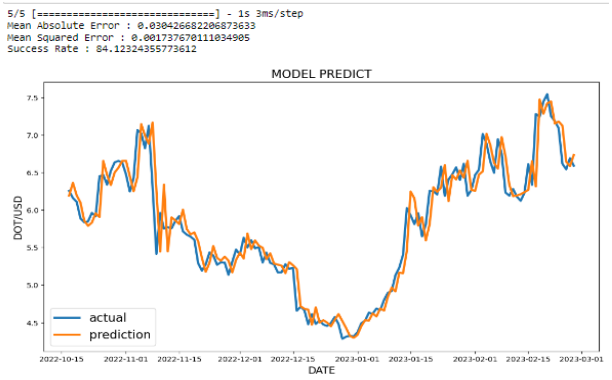


(a) (b)

Şekil 10. (a) Epochs-MSE grafiği (Epochs-MSE chart) (b) 3 girdili gerçek-tahmin grafiği ve hata/başarı oranları (Actual-prediction graph and error-success rates)

Şekil 11'de 3 girdili %80 eğitim seti ile oluşturulmuş grafikler gözükmektedir. Şekil 11(a) Epoch-MSE grafiğini verirken, Şekil 11(b)'de gerçek ve tahmin değeri grafiği gözükmektedir. Ayrıca Şekil 11(b)'de yer alan başarı oranı 84,123 olarak okunmaktadır.

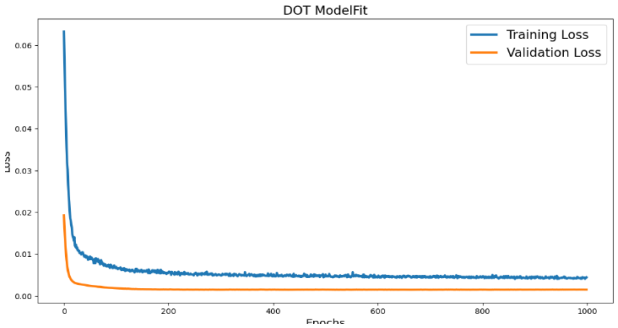


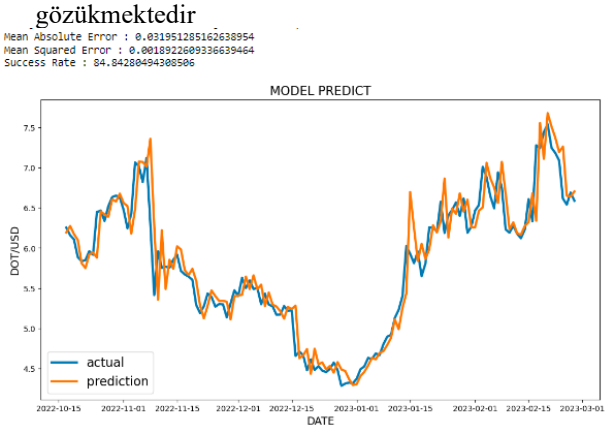


(a) (b)

Şekil 11. (a) Epochs-MSE grafiği (Epochs-MSE chart) (b) 3 girdili gerçek-tahmin grafiği ve hata/başarı oranları (Actual-prediction graph and error-success rates)

Şekil 12'de 3 girdili %90 eğitim seti ile oluşturulmuş grafikler gözükmektedir. Şekil 12(a) Epoch-MSE grafiğini verirken, Şekil 12(b) gerçek ve tahmin değeri grafiği gözükmektedir. Ayrıca Şekil 12(b)'de yer alan başarı oranına bakıldığında 84,842 çıktığı gözükmektedir

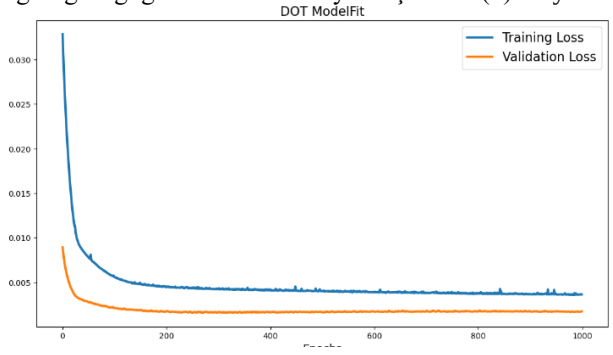


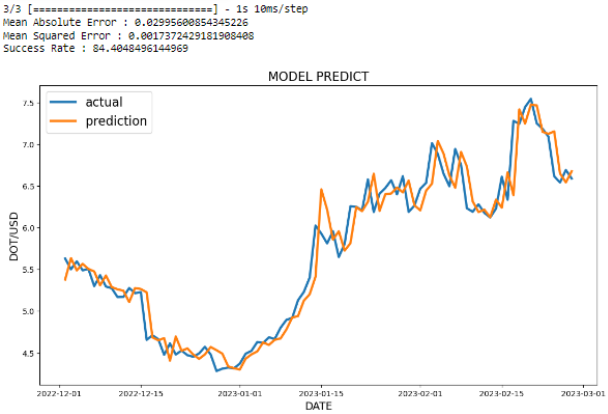


(a) (b)

Şekil 12. (a) Epochs-MSE grafiği (Epochs-MSE chart) (b) 3 girdili gerçek-tahmin grafiği ve hata/başarı oranları (Actual-prediction graph and error-success rates)

Şekil 13'te 4 girdili %70 eğitim seti ile oluşturulmuş grafikler gözükmektedir. Şekil 13(a) Epoch-MSE grafiğini verirken, Şekil 13(b)'de gerçek ve tahmin değeri grafiği gözükmektedir. Ayrıca Şekil 13(b)'de yer alan başarı oranına bakıldığında 84,4048 çıktığı gözükmektedir.

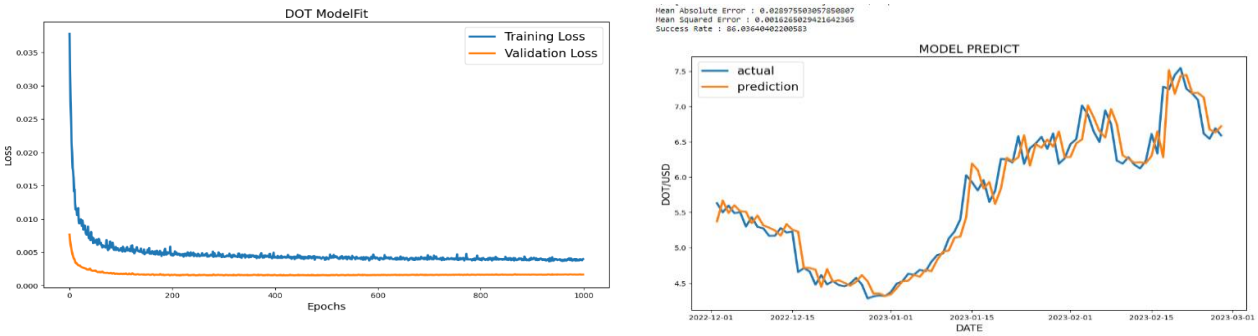




(a) (b)

Şekil 13. (a) Epochs-MSE grafiği (Epochs-MSE chart) (b) 4 girdili gerçek-tahmin grafiği ve hata/başarı oranları (Actual-prediction graph and error-success rates)

Şekil 14'de 4 girdili %90 eğitim seti ile oluşturulmuş grafikler gözükmektedir. Şekil 14(a) Epoch-MSE grafiğini verirken, Şekil 14(b)'de gerçek ve tahmin değeri grafiği gözükmektedir. Ayrıca Şekil 14(b)'de yer alan başarı oranı 86,036 olarak gözükmektedir.



(a) (b)

Şekil 14. (a) Epochs-MSE grafiği (Epochs-MSE chart) (b) 4 girdili gerçek-tahmin grafiği ve hata/başarı oranları (Actual-prediction graph and error-success rates)

Sonuçlara göre, eğitim veri sayısı arttığında başarı oranın arttığı görülmüştür. Ayrıca dört girdili olarak Ethereum kripto para biriminin de değerlerini eklemenin çalışmaya olumlu yansıdığı tespit edilmiştir. Bu çalışmadan elde edilen sonuçlar genel bir tablo halinde Tablo 2 ve Tablo 3'te verilmiştir.

Tablo 2. 4 girdili derin öğrenme ve yapay sinir ağı sonuç değerleri (Deep learning and neural network result values for 4 input)

Yöntem	R	MSE	RMSE
Derin Öğrenme	0,86036	0,01626	0,12751
Yapay Sinir Ağları	0,93944	0,01165	0,10793

Tablo 2'de korelasyon katsayısı (R) ile kıyaslandığında yapay sinir ağlarının daha iyi sonuç verdiği gözlemlenmiştir. Tablo 2'de dört girdi parametresi, Tablo 3'te ise 3 girdi parametresi için derin öğrenme ve yapay sinir ağları üzerindeki en yüksek değerler verilmiştir. İki algoritmada da 10 nöron ve %90 test verisi daha iyi sonuç verdiği gözlemlenmiştir.

Tablo 3. 3 girdili derin öğrenme ve yapay sinir ağı sonuç değerleri (Deep learning and neural network result values for 3 input)

Yöntem	R	MSE	RMSE
Derin Öğrenme	0,84842	0,01800	0,13416
Yapay Sinir Ağları	0,91059	0,01975	0,14053

5. Sonuçlar (Conclusions)

Teknolojinin gelişmesi ile bazı yatırımcılar fiziki yatırım araçlarından ziyade sanal yatırım araçlarına yönelmektedir. Güncel teknolojiler arasında yer alan blok zinciri teknolojisinin bir ürünü olan kripto paralar da son zamanların popüler yatırım araçları arasına girmeyi başarmıştır. Bir alım satım işleminin bazen saatler, günler sürdüğü geleneksel para birimleri yerine kripto paralarda saniyeler içinde işlemler tamamlanabilmektedir. Ayrıca blok zinciri teknolojisi sayesinde güvenilir olduğu düşünülmektedir. Kripto paralarda merkezi kontrolün olmaması, şeffaf olması, değiştirilemez olması kötü kullanımın minimum olmasını sağlayacaktır.

Bu çalışma kapsamında Polkadot kripto para biriminin 20.08.2020 ve 27.02.2023 tarihleri arasındaki veriler kullanılarak tahminleme yapılmıştır. Tahminleme işlemi için yapay sinir ağlarında çok katmanlı algılayıcılar kullanılırken derin öğrenme metotlarında ise LSTM kullanılmıştır. Polkadot kripto birimini etkileyen girdi değerleri iki farklı şekilde ele alınmıştır. Bunlardan ilkinde Polkadot Youtube arama sayısı, Polkadot web arama sayısı, Polkadot hacim değerini oluşturmaktadır, ikinci girdi kümesinde Polkadot Youtube arama sayısı, Polkadot web arama sayısı, Polkadot hacim değerine ek olarak önemli alt kripto para birimlerinden olan Ethereum da katılarak üzerindeki etkisi anlaşılmaya çalışılmıştır. Çalışma yapılırken çapraz doğrulama yöntemi kullanılarak eğitim verilerinin oranı belirlenmiştir. Ayrıca nöron sayısı belirlenirken 10 ve 20 şeklinde test ederek belirlenmiştir. Yapılan çalışma neticesinde yapay sinir ağlarında çok katmanlı geri yayılım algoritmasında Ethereum değerinin de etkilendiği dört girdili 10 nöron sayılı, %90 eğitim verisinin olduğu veri setinde tahminleme yapıldığında 0,93 korelasyon katsayısı ile en yüksek başarıya ulaştığı görülmüştür. Aynı çalışma derin öğrenme yöntemlerinden LSTM kullanılarak gerçekleştirildiğinde en iyi sonuç 10 nöron ve dört girdili, %90 eğitim verisinin olduğu 0,86 korelasyon katsayısı ile elde edilmiştir. Yapılan çalışmada dört girdili, 10 nöronlu, %90 eğitim verisi ile yapılan çalışmanın hem yapay sinir ağlarında hem de derin öğrenmede daha iyi sonuç verdiği belirlenmiştir.

Yapılan çalışma ile farklı girdi değerlerinin, nöron ve eğitim verisini oranlarının başarıyı nasıl etkilediği belirlenmiştir. Ayrıca farklı algoritmalar üzerinde geliştirme yapmanın başarı oranlarında farklı değerler ortaya çıkarabileceği gösterilmiştir.

İleriki çalışmalarda farklı girdi parametreleri, farklı model yapıları, farklı algoritmalar kullanılarak kıyaslamalar yapılması doğru sonuca ulaşma önem arz etmektedir.

Teşekkür (Acknowledgment)

Bu çalışma, Sakarya Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim dalında gerçekleştirilen, 2. yazarın danışmanlığında 1. yazara ait yüksek lisans tezinden üretilmiştir.

Kaynaklar (References)

Aras, S., 2019. Kripto para fiyatlarının klasik ve yapay sinir ağı modelleri ile tahmini. Kafkas Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 10(20), 608-640.

Atlan, F., Pençe, İ., Çeşmeci, M., 2020. Kripto paralardan Bitcoin, Ethereum ve Ripple için yapay zekâ ile online fiyat tahmin modeli. 28th Signal Processing and Communications Applications Conference (SIU), 2165-0608.

Avşar, İ.İ., 2020. Kripto paralar ve uluslararası ticaret üzerine bir araştırma: Bibliyometrik, LSTM ve kümeleme analizi. Doktora Tezi, Hasan Kalyoncu Üniversite ve Gaziantep Üniversitesi.

Akay, M.K., Canik, F., Yeşilyurt, C., ve Günkut, M.Ş., 2021. Yapay zeka teknikleri ile kripto para değeri tahmini, Ekonomi Bilimleri Dergisi, 14(1), 72-101.

Baygıner, O., 2022. Kripto para piyasaları ve Türkiye'de insanların piyasalara yaklaşımı. Yüksek Lisans Tezi, Üsküdar Üniversitesi.

Çılgın, C., Ünal, C., Alıcı, S., Akkol E., ve Gökşen, Y., 2020. Metin sınıflandırmada yapay sinir ağları ile Bitcoin kodları ve sosyal medyadaki beklentilerin analizi. Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi, 4(1), 106-126.

Deniz, E.A., 2020. Finansal piyasalarda kripto para uygulamaları: Kripto para fiyatlarını etkileyen faktörler. Yüksek Lisans Tezi, Işık Üniversitesi.

Demirci, E., 2021. Kripto Para Fiyatlarının LSTM ve GRU modelleri ile tahmini. Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi.

Evlimoğlu, U., Güder, M., 2021. Tarihteki ekonomik balonlar ışığında kripto paralara genel bir bakış. Abant Sosyal Bilimler Dergisi, 21(3), 469-496.

Hayradi, D., Hâkim, A., Atmaja, D., Yutia, S., 2022. Implementation of support vector regression for Polkadot cryptocurrency price prediction. Int. J. Inform. Visualization, 6(1-2), 201-207

Kalyoncu, Ş., 2020. Borsa analizi ve tahmini için derin öğrenme ağları. Yüksek Lisans Tezi, İstanbul Sabahattin Zaim Üniversitesi.

Karaatlı, M., Helvacıoğlu, Ö., Ömürberk, N., ve Tokgöz, G., 2012. Yapay sinir ağları yöntemi ile otomobil satış tahmini. Uluslararası Yönetim İktisat ve İşletme Dergisi, 8(17), 87-100.

Kartal, C., 2020. K-Star algoritması ile Bitcoin fiyatları modelleme. Business & Management Studies: An International Journal, 8 (1), 213–231.

Köksal, B., Erdem, G., Türkeli, C., Öztürk, Z.K., 2021. Twitter'da duygu analizi yöntemi kullanılarak Bitcoin değer tahminlemesi. Düzce Üniversitesi Bilim ve Teknoloji, 9(3), 280-297.

Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system. Decentralized Business Review, 21260.

Pamuk, Ö.G., 2019. Cryptocurrency price prediction by using social media data. Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi.

Sakız, B., Gencer, A. H., 2017. Yapay sinir ağları ile bitcoin fiyatını tahminleme. In International Conference on Eurasian Economies, 438-444.

Salman, M., 2020. Price prediction of different cryptocurrencies using technical trade indicators and machine learning. Yüksek Lisans Tezi, Altınbas Üniversitesi.

Sel, A., 2020. Pandemi sürecinde altın fiyatları ile Kripto para ilişkisinin makine öğrenme metotları ile incelenmesi. Journal of Statistics & Applied Science, 1(2), 85-98

Sel, A., Zengin, N., Yıldız, Z., 2020. Alternatif yatırım araçları ile Bitcoin fiyatları arasındaki ilişkinin yapay sinir ağları ile tahmini. Sivas Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 21(2), 157-169.

Şahin, E.E., 2018. Kripto para Bitcoin: ARIMA ve yapay sinir ağları ile fiyat tahmini. Fiscaoeconomia, 2(2), 74-92.

Tuncer, A., 2022. LSTM metodu kullanılarak rüzgar hızının tahmin edilmesi. Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi.

Turan, S., 2019. Uzun kısa süreli hafıza ve geçitli yinelenen birim ile Borsa İstanbul 100 endeks değeri tahmini üzerine bir uygulama. Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi.

Yavuz, S., Deveci, M., 2012. İstatiksel normalizasyon tekniklerinin yapay sinir ağın performansına etkisi. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 40, 167-187.

Yavuz, U., Özen, Ü., Taş, K., Çağlar, B., 2020. Yapay sinir ağları ile Blockchain verilerine dayı Bitcoin fiyat tahmini, Bilişim Sistemleri ve Yönetim Araştırmaları Dergisi, 2(1), 1-9.

Research Article
Journal of Intelligent Systems: Theory and Applications 6(2) (2023) 108-118
DOI: 10.38016/jista.1200225

Binary Honey Badger Algorithm for 0-1 Knapsack Problem

Gülşen Orucova Büyüköz[1*], Hüseyin Haklı[2]

[1] Department of Mathematics And Computer Science, Necmettin Erbakan University, Konya, Türkiye

[2] Department of Computer Engıneerıng, Necmettin Erbakan University, Konya, Türkiye

gorucova@erbakan.edu.tr, hhakli@erbakan.edu.tr

Abstract
Honey Badger Algorithm (HBA) is one of the recently proposed optimization techniques inspired by the foraging behavior of honey badger. Although it has been successfully applied in solving continuous problems, the algorithm cannot be implemented directly in binary problems. A binary version of HBA is proposed in this study for the 0-1 Knapsack Problem (0-1 KP). To adapt the binary version of HBA, V- Shaped, S-Shaped, U-Shaped, T-Shaped, Tangent Sigmoid, O-Shaped, and Z-Shaped transfer functions are used. Each transfer function was tested by computational experiments over 25 instances of 0-1 KP and compared results. According to the results obtained, it was observed that O1 was the best TF among 25 TFs. In addition, the proposed algorithm was compared with three different binary variants, such as BPSO, MBPSO, and NGHS. Experimental results and comparison show that the proposed method is a promising and alternative algorithm for 0-1 KP problems.

Keywords: Binary Honey Badger Algorithm, 0-1 Knapsack Problems, Transfer Functions, Binary Optimization.

0-1 Sırt Çantası Problemi İçin İkili Bal Porsuğu Algoritması

Öz
Bal Porsuğu Algoritması (HBA), son zamanlarda önerilen optimizasyon tekniklerinden biridir ve bal porsuğunun yiyecek arama davranışından esinlenmiştir. Sürekli problemlerin çözümünde başarılı bir şekilde uygulanmasına rağmen, algoritma doğrudan ikili problemlerde uygulanamaz. Bu çalışmada 0-1 Sırt Çantası Problemi (0-1 KP) için HBA'nın ikili versiyonu önerilmiştir. HBA'nın ikili versiyonunu uyarlamak için V-Şekilli, S-Şekilli, U-Şekilli, T-Şekilli, Tanjant Sigmoid, O-Şekilli, Z-Şekilli transfer fonksiyonları (TF) kullanılmaktadır. Her transfer fonksiyonu 25 0-1 KP problemi için test edilmiş ve sonuçlar karşılaştırılmıştır. Elde edilen sonuçlara göre 25 TF arasından en iyi TF'nin O1 olduğu görülmüştür. Ayrıca bu algoritma BPSO, MBPSO, NGHS gibi üç farklı ikili varyant ile karşılaştırılmıştır. Deneysel sonuçlar ve karşılaştırmalar önerilen yöntemin 0-1 KP problemleri için umut verici ve alternatif bir araç olduğunu göstermektedir.

Anahtar kelimeler: İkili Bal Porsuğu Algoritması, 0-1 Sırt Çantası Problemleri, Transfer Fonksiyonları, İkili Optimizasyon.

1. Introduction

Many real-world problems, such as scheduling problems (Kaya et al., 2020), (Deng, Xu and Zhao, 2019) placement of wind turbines (Deng, Xu and Zhao, 2019; Hakli, 2019), vehicle routing (Halat and Ozkan, 2021), optimization of seismic isolation parameters (Çerçevik and Avşar, 2020), etc. use meta-heuristic optimization methods due to traditional solution methods that are insufficient. One of these problems is the knapsack problem.

0-1 KP has a prominent part in many real-world applications such as decision-making processes, exploiting resources optimally, database storage, investment strategies, and network formation. This problem is one of the fundamental NP-hard problems that achieves the maximum profit and the minimum cost in combinatorial optimization (Bansal and Deep, 2012), (Rooderkerk and van Heerde, 2016).

Recently, 0-1 KP has been applied by many swarm-intelligence and population-based optimization algorithms. Meta-heuristic optimization algorithms presented for continuous search space must be adapted to binary structure for tackling discrete optimization problems. Transfer functions are widely preferred approaches to discretization of a continuous algorithm. The binary Particle Swarm Optimization algorithm was modified (MBPSO) and used to solve some 0-1 KPs and

* Corresponding Author
E-mail: gorucova@erbakan.edu.tr

Recieved : 6 Nov 2022
Revision : 9 Mar 2023
Accepted : 1 Jun 2023

multidimensional KPs, results were compared with Binary PSO algorithm (Bansal and Deep, 2012). Cuckoo Search (CS) algorithm was transformed to a binary version using the sigmoid function (Gherboudj, Layeb and Chikhi, 2012). The binary Monkey Algorithm (BMA) was developed by (Zhou, Chen and Zhou, 2016). BMA was employed with the greedy algorithm to strengthen the local search ability to overcome fall into local optimal solutions. Also, 0-1 KP was considered by Binary Monarch Butterfly Optimization (BMBO) using S-shaped transfer functions and repair operator (Feng *et al.*, 2017). Social Spider Algorithm was adapted to binary search space with sigmoid function and repair algorithm to overcome 0-1 KPs (Nguyen, Wang and Truong, 2017). In another study (Rizk-Allah and Hassanien, 2018), Binary Bat Algorithm (BBA) was established based on the V-shaped and S-shaped transfer functions and used to cope with 0-1 KP. The Differential Evolution Algorithm was designed to apply to binary problems and the binary version was tested on the 0-1 KPs (Ismail M. Ali, 2018). A binary variant of Flower Pollination Algorithm (BFPA) with sigmoid transfer function was introduced, and repair operator and penalty function were employed to improve the solution quality (Abdel-Basset, El-Shahat and El-Henawy, 2019). Using V-Shaped and S-Shaped transfer functions, Marine Predators Algorithm (MPA) was moved from continuous to discrete space (Abdel-Basset *et al.*, 2021). The binary version of the Equilibrium Algorithm (BEA) was proposed for tackling 0–1 KP. Because the standard Equilibrium Optimizer (EO) was presented to overcome continuous optimization problems, EO was transformed to BEO with V-Shaped and S-Shaped TFs. Results showed that among those transfer functions, V3 was the best one (Abdel-Basset, Mohamed and Mirjalili, 2021). Ali et al. proposed a new binary technique that makes a simple differential evolution algorithm adequate for solving binary optimization issues (Ali, Essam and Kasmarik, 2021). The Giza Pyramids Construction (GPC) algorithm was proposed with accumulative and multiplicative penalty functions to determine infeasible solutions in the binary version of GPC (Harifi, 2022). On the other hand binary version of Slime Mould Algorithm (SMA) was presented to convert a continuous variable to a binary by employing eight different transfer functions (Abdollahzadeh *et al.*, 2021). A Quantum Inspired Social Evolution Algorithm (QSE) (Pavithr and Gursaran, 2016) was obtained by hybridizing Social Evolution Algorithm with the QSE, and the method was compared with different algorithms. Cohort Intelligence (CI) (Kulkarni and Shabir, 2016) was inspired by individuals' social, natural and social learning to learn from each other. Several cases of 0–1 KP were applied using CI, and the various parameters influencing the solution quality were discussed. One of the global optimization strategies was the complex-valued encoding method. Zhou, Li, et al. applied the method to the bat algorithm, and the sigmoid function was used for obtaining the discrete value (Zhou, Li and Ma, 2016). In order to achieve the good solutions that increases the total value without overcapacity of knapsack by Grey Wolf Optimization (GWO) and K-means algorithm was merged and dealt with the complexity of the algorithm (Yassien *et al.*, 2017). Genetic Algorithm, Branch and Bound, Simulated Annealing, Dynamic Programming, Greedy Search algorithms were compared for obtained 0-1 KPs results and discussed in (Ezugwu *et al.*, 2019). Improved Whale Optimization Algorithm (IWOA) was performed by the sigmoid transfer function to convert the real-valued solutions into binary and combining the penalty function with the fitness function to evaluate performance single and multidimensional 0-1 KPs are solved (Abdel-Basset, El-Shahat and Sangaiah, 2019). Due to fact that Dragonfly Algorithm (DA) performs on continuous search space, angle modulation mechanism was used for DA to adapt the algorithm works in the binary space (Wang, Shi and Dong, 2021). Moreover, Hybrid Harmony Search Algorithm with distribution estimation was introduced (Liu *et al.*, 2022), Hybrid Rice Optimization (HRO) was merged with Binary Ant Colony Optimization (BACO) algorithm to increase the convergence speed and search efficiency (Shu *et al.*, 2022).

Although many meta-heuristic optimization methods have been applied to overcome the 0-1 KPs, HBA has not been used to this problem. However, HBA presents a successful performance for continuous optimization problems, but a remarkable binary version of HBA is not seen in literature (Hashim *et al.*, 2022). This paper proposes binary versions of HBA with transfer functions and applies to several 0-1 KPs. The rest of this paper are formed as follows: Section 2 explains 0-1 KPs and original HBA. Section 3 presents binary version of HBA and implementation of transfer functions. In Section 4, the experiment results and comparison of transfer functions are conducted. The last section covers the conclusion of the study and provides some possible future directions.

2. Materials and Method

2.1. 0-1 Knapsack Problem

The 0-1 KP problem, proposed by Dantzig (Dantzig, 1957), is based on the knapsack, which has a capacity $C > 0$ and contains a set of n items $(x_1, x_2, \ldots, x_n)$. For each x_i item has $p_i > 0$ profit and $w_i > 0$ weight $(i = 1,2,\ldots,n)$. If x_i item is selected x_i=1 and $x_i = 0$ if x_i is not selected into the knapsack. The goal of this issue is to achieve a maximum profit from the items selected for the knapsack and the weights of all chosen items must be less or equal to the capacity of the knapsack. Mathematically formulation is given below (Rooderkerk and van Heerde, 2016);

fitness function: $\max_i \sum_{i=1}^{n} x_i p_i$ (1)

subject to

$$\sum_{i=1}^{n} x_i w_i \leq C, \quad x_i \in \{0,1\}, i = 1,2,\ldots,n \tag{2}$$

2.2. Honey Badger Algorithm

Honey Badger Algorithm (HBA) is a search strategy used to solve mathematical optimization problems inspired by the honey badger's foraging behavior. This algorithm is proposed by Hashim et al. (Hashim *et al.*, 2022). The honey badger's digging and dynamic foraging behavior are formulated in the exploration and exploitation phases. In the case of digging, it uses its sense of smell to predict the location of prey; Once reached, it moves around the prey to catch the prey. In the case of honey, the honey badger takes the guide of the honey guide bird to find the beehive directly. The steps of the Honey Badger algorithm are given below.

The algorithm starts by generating a randomly population of candidate solutions with the help of the following equation.

$$x_i = lb_i + r_1(ub_i - lb_i) \tag{3}$$

where x_i is honey badger's $i$ts position, lb_i and ub_i are the lower and upper bounds of the search space, respectively. r_1 is a random number between 0 and 1. The other important notation is intensity (I) which is related to the concentration power of the prey and the distance between it and the prey. I_i is the odor intensity of the prey; if the odor is high, the honey badger will move quickly and vice versa. This odor intensity is inversely proportional to the distance of the honey badger from the prey. There S is the source power or concentration power, d_i is the distance between prey and honey badger, and r_2 is a random number between 0 and 1.

$$I_i = r_2 \frac{S}{4\pi d_i^2} \tag{4}$$

$$S = (x_i - x_{i+1})^2 \tag{5}$$

$$d_i = x_{prey} - x_i \tag{6}$$

The intensity factor (α) controls the time change randomness to provide a soft transition from exploration to exploitation. The decreasing factor α is updated with decreasing iterations in a random time. Where $C \geq 1$ is constant (default value is 2), t_{max} is maximum number of iterations.

$$\alpha = C.exp\left(\frac{-t}{t_{max}}\right) \tag{7}$$

For escaping from the local optimum algorithm is used an F flag that changes the search direction of the algorithm. In equation (9), the property of flag F is given.

The main phases of HBA are the digging phase and the honey phase. In first phase, honey badger draws the path in the form of Cardioid. This motion is simulated by equation (8)

$$x_{new} = x_{prey} + F.\beta.I.x_{prey} + F.r_3.\alpha.d_i.|cos(2\pi r_4).[1 - cos(2\pi r_5)]| \tag{8}$$

where, x_{prey} is the prey position, x_{new} honey badger's new position, $\beta \geq 1$ (default 6) honey badger's ability to reach food, d_i is the distance between prey and the ith honey badger and r_3, r_4, r_5 are random numbers different from each other between 0 and 1. The F flag changes the search direction and is defined as follows.

$$F = \begin{cases} 1, & r_6 \leq 0.5 \\ -1, & else \end{cases} \tag{9}$$

where r_6 is a random value in range [0,1].

In the second phase honey badger's pursuit of the honey guide bird is shown by the equation below.

$$x_{new} = x_{prey} + F.r_7.\alpha.d_i \tag{10}$$

where r_7 is a random like r_6. For more information on HBA, see (Hashim *et al.*, 2022).

The flowchart of the HBA method is demonstrated in Figure 1.

2.3. Transfer Functions (TFs)

Selecting an appropriate transfer function is an important decision to increase efficiency, as transfer functions play a significant role in converting the continuous search space to binary space. PSO algorithm is adapted to binary space with the help of the sigmoid function, which is defined as follows (J. Kennedy, 1997):

$$sigm\left(v_i^d(t)\right) = \frac{1}{1 + e^{-v_i^d(t)}} \tag{11}$$

where $v_i^d(t)$ is the following velocity of the i^{th} particle in the d^{th} dimension. The position, $x_i^d(t+1)$ is updated by the following equation:

$$x_i^d(t+1) = \begin{cases} 1, & if\ r \geq sigm\left(v_i^d(t)\right) \\ 0, & else \end{cases} \tag{12}$$

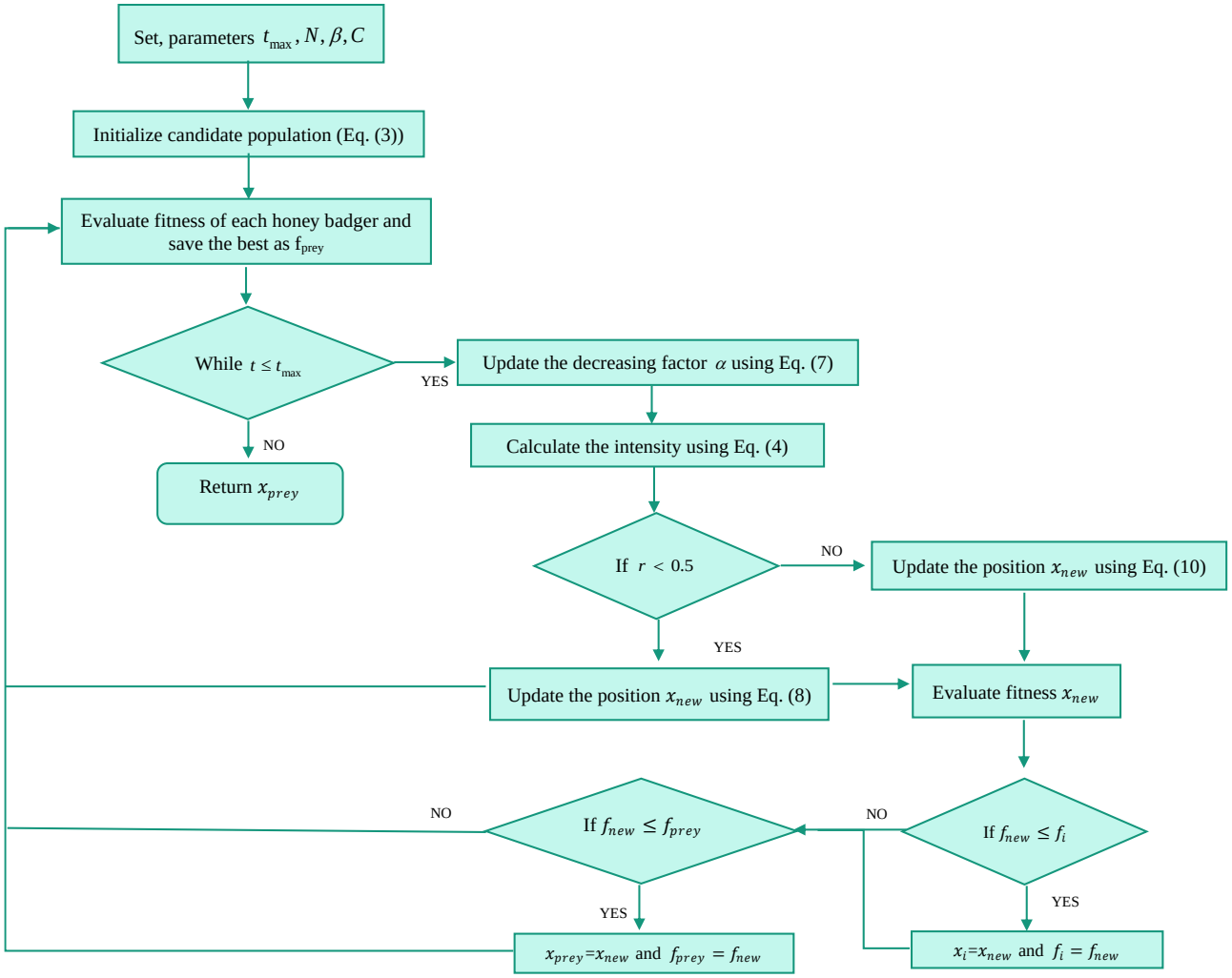


Figure 1. Flowchart of the Honey Badger Algorithm (HBA)

where r is a number between 0 and 1, which is generated with uniform distribution

Seyedali Mirjalili et. al. proposed six new TFs, S-shaped and V-shaped (Mirjalili and Lewis, 2013). The formula of each function is denoted in Table 1. The value obtained between 0 and 1 is converted to a binary value using Eq. (12).

Transfer functions U-shaped has been defined as $U(x) = \alpha|x^{\beta}|$ (Mirjalili *et al.*, 2020). Where α, β are the control parameters. U1, U2, U3, U4 transfer functions which used in our study is shown in Table 2. Obtained value with the U-shaped TF is converted into binary space using Eq. (12).

In order to effectively perform the binary optimization problems, Taper-shaped TF was introduced (He *et al.*, 2022). Formulas of T1, T2, T3, T4 TFs are given in Table 2. There are upper bounds of the search space $[-A, A]$. The calculated real value with the T-shaped TF is converted into binary space using Eq. (13).

$$Bin_{val} = \begin{cases} 1, if\ 0.5 \geq TF(x) \\ 0, else \end{cases} \quad (13)$$

Z-shaped probability transfer function is proposed by Guo et al. (Guo *et al.*, 2020). The formula of each Z-shaped function is given in Table 3. A real number obtained between 0 and 1 using the Z-shaped TF is converted to the binary value using Eq. (12).

In addition to the TFs mentioned above, Other-shaped transfer functions also appear in the literature. O1 TF is proposed by Pampará et. al. (Pampará and Engelbrecht, 2011). O2 TF is introduced by Costa et al. (Costa *et al.*, 2014). O3 TF is linear normalization function (Wang *et al.*, 2008), O4 TF is taken as unit function (Zhu *et al.*, 2017). The value calculated with the O1, O4, O2, O3 TFs is converted to binary value with Eq. (14), Eq. (15), Eq. (12), respectively. The

formula of each Other-shaped function family is given in Table 3.

$$Bin_{val} = \begin{cases} 1, if\ 0 \le TF(x) \\ 0, else \end{cases} \quad (14)$$

$$Bin_{val} = TF(x) \quad (15)$$

The last TF we use is Hyperbolic tangent sigmoid (TanSig) TF is given Table 4 (Yonaba, Anctil and Fortin, 2010). The real value is converted to binary according to Eq. (16).

$$Bin_{val} = \begin{cases} 1, if\ 0.6 < TF(x) \\ 0, else \end{cases} \quad (16)$$

Table 1. S-Shaped and V-Shaped TFs

Name	Formulation of TF	Equation
S- Shaped	S1: $TF(x) = \frac{1}{(1+e^{-2x})}$	$x_i^d(t+1) = \begin{cases} 1, & if\ r \ge sigm\left(v_i^d(t)\right) \\ 0, & else \end{cases}$ (12)
	S2: $TF(x) = \frac{1}{(1+e^{-x})}$	
	S3: $TF(x) = \frac{1}{(1+e^{-x/2})}$	
	S4: $TF(x) = \frac{1}{(1+e^{-x/3})}$	
V- Shaped	V1: $TF(x) = \left\lvert erf\left(\frac{\sqrt{\pi}}{2}x\right)\right\rvert$	
	V2: $TF(x) = \lvert tanh(x)\rvert$	
	V3: $TF(x) = \left\lvert\frac{x}{\sqrt{1+x^2}}\right\rvert$	
	V4: $TF(x) = \left\lvert\frac{2}{\pi} arctan\left(\frac{\pi}{2}x\right)\right\rvert$	

Table 2. U- Shaped and Taper-Shaped TFs

Name	Formulation of TF	Equation
U- Shaped	U1: $TF(x) = \lvert x^{1.5}\rvert,\ \alpha = 1,\ \beta = 1.5$	$x_i^d(t+1) = \begin{cases} 1, & if\ r \ge sigm\left(v_i^d(t)\right) \\ 0, & else \end{cases}$ (12)
	U2: $TF(x) = \lvert x^2\rvert,\ \alpha = 1,\ \beta = 2$	
	U3: $TF(x) = \lvert x^3\rvert,\ \alpha = 1,\ \beta = 3$	
	U4: $TF(x) = \lvert x^4\rvert,\ \alpha = 1,\ \beta = 4$	
Taper- Shaped	T1: $TF(x) = \frac{\sqrt{\lvert x\rvert}}{\sqrt{\lvert A\rvert}}$	$Bin_{val} = \begin{cases} 1, if\ 0.5 \ge TF(x) \\ 0, else \end{cases}$ (13)
	T2: $TF(x) = \frac{\lvert x\rvert}{\lvert A\rvert}$	
	T3: $TF(x) = \frac{\sqrt[3]{\lvert x\rvert}}{\sqrt[3]{\lvert A\rvert}}$	
	T4: $TF(x) = \frac{\sqrt[4]{\lvert x\rvert}}{\sqrt[4]{\lvert A\rvert}}$	

Table 3. Z- Shaped and Other-Shaped TFs

Name	Formulation of TF	Equation
Z- Shaped	Z1: $TF(x) = \sqrt{1-2^x}$ Z2: $TF(x) = \sqrt{1-5^x}$ Z3: $TF(x) = \sqrt{1-8^x}$ Z4: $TF(x) = \sqrt{1-20^x}$	$x_i^d(t+1) = \begin{cases} 1, & if\ r \geq sigm\left(v_i^d(t)\right) \\ 0, & else \end{cases}$ (12)
Other- Shaped	O1: $TF(x) = sin\big(2\pi(x-a) * b * cos(2\pi(x-a) * c)\big) + d$ $(a = d = 0, \quad b = c = 1)$	$Bin_{val} = \begin{cases} 1, if\ 0 \leq TF(x) \\ 0, else \end{cases}$ (14)
	O2: $TF(x) = [\![x\ mod\ 2]\!]$	$Bin_{val} = TF(x)$ (15)
	O3: $TF(x) = \frac{(x-A_{min})}{A_{max}-A_{min}}, \quad (A_{min} \leq x \leq A_{max})$	$x_i^d(t+1) = \begin{cases} 1, & if\ r \geq sigm\left(v_i^d(t)\right) \\ 0, & else \end{cases}$ (12)
	O4: $TF(x) = x$	$Bin_{val} = \begin{cases} 1, if\ 0 \leq TF(x) \\ 0, else \end{cases}$ (14)

Table 4. Hyperbolic Tangent Sigmoid TF

Name	Formulation of TF	Equation
Hyperbolic tangent sigmoid	$TF(x) = \frac{2}{1+e^{-2x}} - 1$	$Bin_{val} = \begin{cases} 1, if\ 0.6 \leq TF(x) \\ 0, else \end{cases}$ (16)

3. Binary HBA with Transfer Functions

The HBA algorithm performs for continuous problems due to its structure. It is clear that the candidate solutions formed by Eq. (3) consist of continuous values. The transfer functions mentioned in Table 1, 2, 3 and 4 take a continuous value as input, then normalized to a value between 0 and 1 using the corresponding equation. Pseudocode of Binary HBA is given in Algorithm 1.

Algorithm 1. Pseudocode of Binary HBA

Set parameters t_{max}, N, β, C.
Generate a random real-valued population with Eq. (3).
Convert each candidate solution to binary representation using TF.
Calculate the fitness value of each candidate solution x_i in the binary representation. (i=1, 2, …, N)
Save best solution x_{prey} and assign fitness to f_{prey}.
while $t \leq t_{max}$ do
  Update α using Eq. (7).
  for i = 1 to N do
    Calculate I_i using Eq. (4).
    if r < 0.5 then
      Update the real valued candidate solution x_{new} using Eq. (8).
    else
      Update the real valued candidate solution x_{new} using Eq. (10).
    end if
    Convert new candidate solution to binary representation using TF.
    Compare the existing candidate solution with the x_{new} by fitness value.
    if fitness(x_{new}) ≤ fitness(x_i) then
      $x_i = x_{new}$ and fitness(x_i) = fitness(x_{new}).
    end if
    if fitness(x_{new}) ≤ f_{prey} then
      $x_{prey} = x_{new}$ and f_{prey} = fitness(x_{new}.)
    end if
  end for
end while Stop criteria satisfied.
Return x_{prey}

As can be seen from Algorithm 1, in Binary HBA, before calculate fitness, continuous values are transformed to binary with the help of TF. The transformation of a candidate solution consisting of 5 dimensional real values x = [-5.48, -3.30, 4.46, 9.71, -6.35] into binary representation [1, 1, 0, 0, 1] with the help of the S2 TF is given in Table 5.

Table 5. Conversion of continuous value to binary with S2 TF

i	x_i	$S2(x_i)$	r	Binary
1	-5.48	0.0041	0.1072	1
2	-3.30	0.0356	0.0736	1
3	4.46	0.9885	0.0917	0
4	9.71	0.9999	0.7845	0
5	-6.35	0.0017	0.3039	1

4. Experimental Results

In this section, the HBA algorithm is adapted for solving 0-1 KPs. The HBA is an algorithm that performs

continuous search space due to its structure. 0-1 KPs, on the other hand, have a binary structure. For this reason, first, N real-valued candidate solutions, each of which is D-dimensional, are created. After each candidate honey badger position is converted to binary with the help of transfer functions, fitness value is evaluated. A total of 25 transfer functions as V-Shaped, S-Shaped, U-Shaped, T-Shaped, Tangent Sigmoid, O-Shaped, Z-Shaped TFs are used to adapt the binary version of the HBA. Each transfer function is tested by computational experiments over 25 instances of 0-1 KP and compared results.

Our experiment was carried on the problems in the benchmark dataset, which can be taken from (https://pages.mtu.edu/~kreher/cages/Data.html) in Table 6.

Table 6. Benchmark datasets

Problem	Capacity	Dimension	Optimal
KP8a	1.863.633	8	3.924.400
KP8b	1.822.718	8	3.813.669
KP8c	1.609.419	8	3.347.452
KP8d	2.112.292	8	4.187.707
KP8e	2.493.250	8	4.955.555
KP12a	2.805.213	12	5.688.887
KP12b	3.259.036	12	6.473.019
KP12c	2.489.815	12	5.170.626
KP12d	3.453.702	12	6.941.564
KP12e	2.520.392	12	5.337.472
KP16a	3.780.355	16	7.850.983
KP16b	4.426.945	16	9.352.998
KP16c	4.323.280	16	9.151.147
KP16d	4.550.938	16	9.348.889
KP16e	3.760.429	16	7.769.117
KP20a	5.169.647	20	10.727.049
KP20b	4.681.373	20	9.818.261
KP20c	5.063.791	20	10.714.023
KP20d	4.286.641	20	8.929.156
KP20e	4.476.000	20	9.357.969
KP24a	6.404.180	24	13.549.094
KP24b	5.971.071	24	12.233.713
KP24c	5.870.470	24	12.448.780
KP24d	5.762.284	24	11.815.315
KP24e	6.654.569	24	13.940.099

For a fair comparison, the number of *maxFes*, population size and runtime illustrate in Table 7. GAP values are calculated using Eq. (17).

$$GAP = \frac{optimal - mean}{optimal} \tag{17}$$

Table 7. The parameter values

Parameters	Value
Maximum Fes	1000 (For KP8a to KP12e)
	5000 (For KP16a to KP24e)
Population size	40
Runtime	50

In order to show performance of the proposed method, a total of 10 problems, Kp8a-Kp8e and Kp12a-Kp12e taken from the benchmark dataset, were performed for 50 runtimes and 1000 Fes number. Also, 15 problems, including Kp16a-Kp16e, Kp20a-Kp20e, Kp24a-Kp24e, run with 50 runtimes and 5000 Fes number. In all of the problems, the search space in the HBA algorithm was adapted to binary space with the help of S, V, U, T, Hyperbolic Tangent Sigmoid, Z and Other-shaped transfer functions. The gap value between the approximate solutions obtained for each transfer function and the optimal solutions was given in Table 8 and Table 9. Table 8 and Table 9 show that optimal solutions were obtained by V, U shaped transfer functions and T1, T3, T4, O1, O2 shaped transfer functions for problems with a problem size of 8. In addition, V, U1, U4, T4, O1, O2 shaped transfer functions reached the optimum value for all runs in 4 problems with problem size 12. In the dataset with a problem size of 16, the optimal value was reached in 3 of the five problems with the help of V, U, T, O2 shaped transfer functions. Although it is seen that it is difficult for the results obtained to reach the optimum value when the problem size is 20 and 24, optimum values were obtained in 2 or 1 of the five problems depending on the transfer functions used. It has been seen that the lowest gap values are in the solutions obtained with the O1 and O2 transfer functions. When we consider Table 8 and Table 9 in general, it can be said that O1 and O2 transfer functions are in the front according to the efficiency of the transfer functions for the 25 problems. In the continuation of this sorting, it has been seen that U1-shaped transfer function gives efficient results.

In this study, the 25 KPs run 50 times to test each transfer function. The optimum number of values obtained for each transfer function due to running 1250 times is given as hit value in Table 10. According to the table, it was seen that the optimal value was reached with the O1 transfer function in 1017 and O2 in 1009 of 1250, respectively. After these functions, U and T transfer functions get the maximum optimum value. As a result, it was observed that the HBA algorithm performed successful results for the O1 and O2 transfer functions.

Table 8. Gap values for each TF (S, V, U) and KP problem with HBA algorithm

Problem	S- Shaped				V- Shaped				U- Shaped			
Name	S1	S2	S3	S4	V1	V2	V3	V4	U1	U2	U3	U4
KP8a	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP8b	0.163	0.228	0.081	0.114	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP8c	0.013	0.020	0.020	0.007	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP8d	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP8e	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP12a	0.070	0.063	0.105	0.084	0.093	0.080	0.075	0.057	0.034	0.032	0.039	0.025
KP12b	0.110	0.118	0.142	0.118	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.008	0.024	**0.000**
KP12c	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP12d	0.018	0.049	0.018	0.027	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP12e	0.036	0.054	0.054	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP16a	0.177	0.156	0.169	0.164	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP16b	0.117	0.105	0.054	0.104	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP16c	**0.000**	0.005	**0.000**	0.017	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP16d	0.147	0.146	0.149	0.121	0.523	0.523	0.523	0.523	0.073	0.063	0.076	0.078
KP16e	0.198	0.193	0.191	0.190	0.286	0.282	0.286	0.288	0.182	0.157	0.199	0.144
KP20a	**0.000**	**0.000**	0.008	0.006	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP20b	0.103	0.056	0.074	0.048	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP20c	0.056	0.058	0.051	0.052	0.044	0.044	0.044	0.044	0.037	0.037	0.033	0.037
KP20d	0.136	0.144	0.140	**0.099**	0.154	0.154	0.154	0.154	0.151	0.154	0.148	0.154
KP20e	0.082	0.065	0.067	0.058	0.124	0.099	0.099	0.033	0.010	0.009	0.007	0.016
KP24a	0.227	0.232	0.181	0.268	0.326	0.332	0.334	0.321	0.259	0.232	0.280	0.239
KP24b	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP24c	0.017	0.021	0.031	0.017	0.052	0.052	0.044	0.043	0.006	0.004	0.001	0.002
KP24d	0.103	0.092	0.099	0.132	0.045	0.045	0.045	0.045	0.043	0.045	0.045	0.044
KP24e	0.088	0.065	0.057	0.053	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.001	0.001	0.001
Friedman Rank	16.1	16.56	16.44	15.46	13.12	12.76	12.72	11.92	8.68	9.06	9.46	8.84
Rank	18	20	19	17	14	13	12	11	3	5	8	4

Table 9. Gap values for each TF (T, Hyp.Tan, O, Z) and KP problem with HBA algorithm

Problem	Taper- Shaped				Hyp.Tan.	Other- Shaped				Z- Shaped			
Name	T1	T2	T3	T4	TanSig	O1	O2	O3	O4	Z1	Z2	Z3	Z4
KP8a	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP8b	**0.000**	0.065	**0.000**	**0.000**	0.130	**0.000**	**0.000**	0.601	0.195	0.293	0.098	0.130	0.195
KP8c	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.623	0.007	**0.000**	0.033	0.013	0.007
KP8d	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP8e	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP12a	0.043	0.053	0.043	0.052	0.079	0.019	**0.018**	0.643	0.088	0.083	0.082	0.077	0.098
KP12b	0.016	0.071	**0.000**	**0.000**	0.087	**0.000**	**0.000**	0.397	0.142	0.118	0.079	0.118	0.118
KP12c	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP12d	**0.000**	**0.000**	0.009	**0.000**	0.018	**0.000**	**0.000**	0.519	0.040	**0.000**	0.019	0.027	0.066
KP12e	0.018	0.018	0.018	**0.000**	0.144	**0.000**	**0.000**	0.664	0.036	0.036	0.072	0.000	0.054
KP16a	**0.000**	**0.000**	**0.000**	**0.000**	0.168	0.010	**0.000**	0.356	0.195	0.203	0.196	0.212	0.152
KP16b	**0.000**	**0.000**	**0.000**	**0.000**	0.054	**0.000**	**0.000**	0.791	0.175	0.157	0.132	0.059	0.078
KP16c	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.021	0.017	0.005	**0.000**	0.015
KP16d	0.068	0.104	0.063	0.167	0.164	0.068	**0.057**	0.130	0.173	0.132	0.110	0.138	0.188
KP16e	0.181	0.156	0.207	0.226	0.205	0.068	**0.054**	0.434	0.186	0.188	0.228	0.181	0.214
KP20a	**0.000**	**0.000**	**0.000**	**0.000**	0.021	**0.000**	**0.000**	0.275	0.016	0.021	0.036	0.012	0.011
KP20b	**0.000**	**0.000**	**0.000**	**0.000**	0.073	**0.000**	**0.000**	0.023	0.102	0.061	0.069	0.078	0.065
KP20c	0.033	0.029	0.037	0.039	0.073	**0.015**	0.024	0.114	0.045	0.035	0.047	0.043	0.055
KP20d	0.145	0.114	0.151	0.154	0.129	0.133	0.136	0.154	0.129	0.114	0.129	0.126	0.138
KP20e	0.008	0.035	0.021	0.039	0.065	**0.013**	0.048	0.268	0.041	0.053	0.048	0.049	0.063

KP24a	0.248	0.196	0.265	0.289	0.278	**0.101**	0.154	0.344	0.261	0.174	0.209	0.216	0.231
KP24b	**0.000**	**0.000**	**0.000**	**0.000**	0.016	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
KP24c	0.003	0.004	0.006	0.019	0.036	0.007	0.032	**0.000**	0.053	0.030	0.034	0.045	0.015
KP24d	0.044	0.045	0.045	0.045	0.127	**0.042**	0.040	0.046	0.158	0.072	0.087	0.089	0.087
KP24e	0.001	0.008	0.001	**0.000**	0.071	0.034	0.001	0.122	0.055	0.080	0.059	0.085	0.056
Friedman Rank	9.06	9.38	10.26	11.06	17.1	7.82	7.88	19.56	17.94	14.96	16.58	15.24	17.04
Rank	6	7	9	10	23	1	2	25	24	15	21	16	22

Table 10. Hit values for each TF with HBA algorithm

TF	Hit value	TF	Hit value	TF	Hit value
S1	810	U1	946	TanSig	792
S2	816	U2	950	O1	1017
S3	816	U3	935	O2	1009
S4	821	U4	950	O3	511
V1	821	T1	942	O4	794
V2	827	T2	925	Z1	789
V3	833	T3	933	Z2	802
V4	845	T4	893	Z3	809
				Z4	789

The binary version of HBA for O1 TF is compared with BPSO, MBPSO (Modified Binary Particle Swarm Optimization) and NGHS (Novel Global Harmony Search) algorithms to evaluate its performance and accuracy. All methods were performed with the same parameters as 50 runs, 1000 Fes number for Kp8a to Kp12e and 5000 Fes number for Kp16a to Kp24e. Experimental results of algorithms were directly taken from (Zhou, Chen and Zhou, 2016), (Hakli, 2020). The gap values of 50 runs for the algorithms and the proposed algorithm are presented in Table 11. The binary version of HBA with O1 TF found the optimum value or the closest results for 22 of 25 problems. Thus, it has been seen that HBA with O1 TF offers more effective solutions than BPSO, MBPSO, and NGHS algorithms for selected 0-1 KP problems.

Table 11. Experimental results of proposed method and binary variants of the different algorithms.

Problem Name	HBA-O1-Shaped	BPSO	MBPSO	NGHS
KP8a	**0.000**	0.065	**0.000**	**0.000**
KP8b	**0.000**	0.151	**0.000**	**0.000**
KP8c	**0.000**	0.563	**0.000**	**0.000**
KP8d	**0.000**	0.039	**0.000**	**0.000**
KP8e	**0.000**	0.460	0.020	**0.000**
KP12a	0.019	0.091	**0.006**	0.020
KP12b	**0.000**	0.308	0.084	0.187
KP12c	**0.000**	0.071	0.003	0.107
KP12d	**0.000**	0.037	0.004	0.006
KP12e	**0.000**	0.386	**0.000**	1.190
KP16a	**0.010**	0.205	0.101	0.711
KP16b	**0.000**	0.199	0.028	1.068
KP16c	**0.000**	0.353	0.077	1.041
KP16d	**0.068**	0.291	0.117	0.406
KP16e	0.068	0.136	**0.064**	0.390
KP20a	**0.000**	0.184	0.063	1.256
KP20b	**0.000**	0.275	0.130	0.909
KP20c	**0.015**	0.099	0.029	1.203
KP20d	0.133	0.213	**0.061**	0.782
KP20e	**0.013**	0.090	0.022	0.356
KP24a	**0.101**	0.285	0.126	0.296
KP24b	**0.000**	0.232	0.084	0.595
KP24c	**0.007**	0.168	0.044	0.195
KP24d	**0.042**	0.197	0.098	0.669
KP24e	**0.034**	0.124	0.054	0.805

5. Conclusions

This study proposed the binary version of HBA algorithm with TFs. The binary variants performed with the help of 25 transfer functions were applied to benchmark datasets for 0-1 KP problem. The results for 25 binary variants were compared to examine the efficiency of each transfer function. The O1 and O2 TFs showed the best successful performances among the TFs. Also, HBA algorithm with O1 TF was compared with three different binary variants, and the results show that binary HBA is the first in the ranking. For future work, the validity of the proposed approach can be enlarged by applying it to different 0-1 KPs. It can be impressive work to adapt the HBA algorithm to binary space without TFs directly.

References

Abdel-Basset, M. et al., 2021. New binary marine predators optimization algorithms for 0-1 knapsack problems. Computers and Industrial Engineering, 151.

Abdel-Basset, M., El-Shahat, D. and El-Henawy, I., 2019. Solving 0-1 knapsack problem by binary flower pollination algorithm. Neural Computing and Applications, 31(9), pp. 5477-5495.

Abdel-Basset, M., El-Shahat, D. and Sangaiah, A.K., 2019. A modified nature inspired meta-heuristic whale optimization algorithm for solving 0-1

knapsack problem. International Journal of Machine Learning and Cybernetics, 10(3), pp. 495-514.

Abdel-Basset, M., Mohamed, R. and Mirjalili, S., 2021. A binary equilibrium optimization algorithm for 0-1 knapsack problems. Computers and Industrial Engineering, 151.

Abdollahzadeh, B. et al., 2021. An enhanced binary slime mould algorithm for solving the 0-1 knapsack problem. Engineering with Computers.

Ali, I.M., Essam, D. and Kasmarik, K., 2021. Novel binary differential evolution algorithm for knapsack problems. Information Sciences, 542, pp. 177-194.

Bansal, J.C. and Deep, K., 2012. A modified binary particle swarm optimization for knapsack problems. Applied Mathematics and Computation, 218(22), pp. 11042-11061.

Costa, M.F.P. et al., 2014. Heuristic-based firefly algorithm for bound constrained nonlinear binary optimization. Advances in Operations Research, 2014.

Çerçevik, A.E. and Avşar, Ö., 2020. Optimization of linear seismic isolation parameters via crow search algorithm. Pamukkale University Journal of Engineering Sciences, 26(3), pp. 440-447.

Dantzig, G.B., 1957. Discrete-Variable Extremum Problems, Source: Operations Research.

Deng, W., Xu, J. and Zhao, H., 2019. An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem. IEEE Access, 7, pp. 20281-20292.

Ezugwu, A.E. et al., 2019. A comparative study of meta-heuristic optimization algorithms for 0-1 knapsack problem: some initial results. IEEE Access, 7, pp. 43979-44001.

Feng, Y. et al., 2017. Solving 0-1 knapsack problem by a novel binary monarch butterfly optimization. Neural Computing and Applications, 28(7), pp. 1619-1634.

Gherboudj, A., Layeb, A. and Chikhi, S., 2012. Solving 0-1 knapsack problems by a discrete binary version of cuckoo search algorithm. International Journal of Bio-Inspired Computation, 4(4), pp. 229-236.

Guo, S.S. et al., 2020. Z-shaped transfer functions for binary particle swarm optimization algorithm. Computational Intelligence and Neuroscience, 2020.

Hakli, H., 2019. A new approach for wind turbine placement problem using modified differential evolution algorithm. Turkish Journal of Electrical Engineering and Computer Sciences, 27(6), pp. 4659-4672.

Hakli, H., 2020. BinEHO: a new binary variant based on elephant herding optimization algorithm. Neural Computing and Applications, 32(22), pp. 16971-16991.

Halat, M. and Ozkan, O., 2021. The optimization of UAV routing problem with a genetic algorithm to observe the damages of possible Istanbul earthquake. Pamukkale University Journal of Engineering Sciences, 27(2), pp. 187-198.

Harifi, S., 2022. A binary ancient-inspired Giza pyramids construction metaheuristic algorithm for solving 0-1 knapsack problem. Application of Soft Computing.

Hashim, F.A. et al., 2022. Honey Badger Algorithm: New metaheuristic algorithm for solving optimization problems. Mathematics and Computers in Simulation, 192, pp. 84-110.

He, Y. et al., 2022. Novel binary differential evolution algorithm based on taper-shaped transfer functions for binary optimization problems. Swarm and Evolutionary Computation, 69.

Ismail M. Ali, D.E. and K.K., 2018. An efficient differential evolution algorithm for solving 0-1 knapsack problems. 2018 IEEE Congress on Evolutionary Computation (CEC): 2018 proceedings.

J. Kennedy, R.C.E., 1997. Discrete binary version of the particle swarm algorithm. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 5 (1997) 4104-4108. IEEE.

Kaya, S. et al., 2020. The effects of initial populations in the solution of flow shop scheduling problems by hybrid firefly and particle swarm optimization algorithms. Pamukkale University Journal of Engineering Sciences, 26(1), pp. 140-149.

Kulkarni, A.J. and Shabir, H., 2016. Solving 0-1 knapsack problem using cohort intelligence algorithm. International Journal of Machine Learning and Cybernetics, 7(3), pp. 427-441.

Liu, K. et al., 2022. A hybrid harmony search algorithm with distribution estimation for solving the 0-1 knapsack problem. Mathematical Problems in Engineering.

Mirjalili, S. and Lewis, A., 2013. S-shaped versus V-shaped transfer functions for binary particle swarm optimization. Swarm and Evolutionary Computation, 9, pp. 1-14.

Mirjalili, Seyedehzahra et al., 2020. A novel U-shaped transfer function for binary particle swarm optimisation. Advances in Intelligent Systems and Computing. Springer, pp. 241-259.

Nguyen, P.H., Wang, D. and Truong, T.K., 2017. A novel binary social spider algorithm for 0-1 knapsack problem. International Journal of Innovative Computing.

Pampará, G. and Engelbrecht, A.P., 2011. Binary artificial bee colony optimization. IEEE SSCI 2011- Symposium Series on Computational Intelligence- SIS 2011: 2011 IEEE Symposium on Swarm Intelligence, pp. 170-177.

Pavithr, R.S. and Gursaran, 2016. Quantum inspired social evolution (QSE) algorithm for 0-1

knapsack problem. Swarm and Evolutionary Computation, 29, pp. 33-46.

Rizk-Allah, R.M. and Hassanien, A.E., 2018. New binary bat algorithm for solving 0-1 knapsack problem. Complex & Intelligent Systems, 4(1), pp. 31-53.

Rooderkerk, R.P. and van Heerde, H.J., 2016. Robust optimization of the 0-1 knapsack problem: Balancing risk and return in assortment optimization. European Journal of Operational Research, 250(3), pp. 842-854.

Shu, Z. et al., 2022. A modified hybrid rice optimization algorithm for solving 0-1 knapsack problem. Applied Intelligence, 52(5), pp. 5751-5769.

Wang, L. et al., 2008. A novel probability binary particle swarm optimization algorithm and its application.

Wang, L., Shi, R. and Dong, J., 2021. A hybridization of dragonfly algorithm optimization and angle modulation mechanism for 0-1 knapsack problems. Entropy, 23(5).

Yassien, E. et al., 2017. Grey wolf optimization applied to the 0/1 knapsack problem. International Journal of Computer Applications, 169(5), pp. 11-15.

Yonaba, H., Anctil, F. and Fortin, V., 2010. Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting. Journal of Hydrologic Engineering, 15(4), pp. 275-283.

Zhou, Y., Chen, X. and Zhou, G., 2016. An improved monkey algorithm for a 0-1 knapsack problem. Applied Soft Computing Journal, 38, pp. 817-830.

Zhou, Y., Li, L. and Ma, M., 2016. A complex-valued encoding bat algorithm for solving 0-1 knapsack problem. Neural Processing Letters, 44(2), pp. 407-430.

Zhu, H. et al., 2017. Discrete differential evolutions for the discounted {0-1} knapsack problem. Chinese Journal of Computers and so on.

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 6(2) (2023) 119-129
DOI: 10.38016/jista.1235031

Makine Öğrenmesi Teknikleri ile Counter-Strike: Global Offensive Raunt Sonuçlarının Tahminlenmesi

Vahid Sinap[1*]

[1]Ufuk Üniversitesi, Yönetim Bilişim Sistemleri Bölümü, Ankara, Türkiye

vahidsinap@gmail.com

Öz

Kamuya açık şekilde sunulan yapılandırılmış ve yapılandırılmamış büyük miktarlardaki verilerle birlikte Espor tahminlemeleri üzerine yapılan çalışmalar her geçen gün artmaktadır. Espor etkinliklerine yönelik tahminleme çalışmaları insan faktöründen büyük ölçüde etkilense de doğru çıktılara ulaşmada önemli birçok parametre sunan yapısıyla tahminlemelerin başarısını artırmaktadır. Bu bağlamda modellerin nasıl oluşturulacağı ve hangi makine öğrenmesi algoritmalarının seçileceği önem taşımaktadır. Bu çalışmada, Counter-Strike: Global Offensive adlı çevrimiçi oyundaki rauntların sonuçlarının tahminlemeye yönelik çeşitli makine öğrenmesi algoritmaları kullanılarak sınıflandırmalar gerçekleştirilmiştir. Araştırmada, Lojistik Regresyon, Karar Ağaçları, Rastgele Orman, XGBoost, Naive Bayes, K-En Yakın Komşu ve Destek Vektör Makinesi olmak üzere toplam yedi adet denetimli sınıflandırma algoritması kullanılmıştır. Bu algoritmaların performans ölçümünde Doğruluk, Kesinlik, Duyarlılık, F-Skor ve AUC değerleri hesaplanmıştır. Ayrıca, ROC eğrileri ve karışıklık matrisleri değerlendirilerek algoritmalar karşılaştırılmıştır. Bu ölçümler ve değerlendirmeler sonucunda Rastgele Orman algoritması %88 doğruluk oranı ile en başarılı algoritma olmuştur. Bunlara ek olarak, rauntların kazanılma durumları bağlamında Keşifsel Veri Analizleri yürütülerek Espor organizasyonlarına yönelik bazı önerilerde bulunulmuştur.

Anahtar kelimeler: Espor, CSGO, Makine Öğrenmesi, Sınıflandırma Algoritmaları, Kazanan Tahminleme

Prediction of Counter-Strike: Global Offensive Round Results with Machine Learning Techniques

Abstract

With the large amounts of structured and unstructured data available to the public, studies on Esports forecasting are increasing day by day. Although prediction studies for esports events are greatly affected by the human factor, it increases the success of predictions with its structure that offers many important parameters in achieving accurate outputs. In this context, it is important how to create models and which machine learning algorithms to choose. In this study, classifications were carried out using various machine learning algorithms to predict the results of the rounds in the online game Counter-Strike: Global Offensive. In the research, a total of seven supervised classification algorithms, namely Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbor and Support Vector Machine were used. Accuracy, Precision, Sensitivity, F-Score and AUC values were calculated in the performance measurement of these algorithms. In addition, algorithms are compared by evaluating ROC curves and confusion matrix. As a result of these measurements and evaluations, the Random Forest algorithm was the most successful algorithm with an accuracy rate of 88%. In addition to these, some suggestions were made for Esports organizations by conducting Exploratory Data Analysis in the context of the winning status of the rounds.

Keywords: Esports, CSGO, Machine Learning, Classification Algorithms, Winning Prediction

1. Giriş (Introduction)

Günümüzde bilgisayar ve internet teknolojilerinin hızlı gelişimleri ile çevrimiçi rekabetçi (competitive) oyunların popülaritesi giderek artmaktadır. Bu popülarite 1972 yılında ilk uygulamalarına rastlanan ve 2000'li yılların başından itibaren giderek yaygın hale gelmiş olan Espor (Esports) kavramını hayatımıza sokmuştur. Elektronik sporun (electronic sports)

* Sorumlu yazar
E-posta adresi: vahidsinap@gmail.com

Alındı : 15 Ocak 2023
Kabul : 22 Haziran 2023

kısaltması olan Espor, video oyunları üzerine kurulu bir rekabet biçimidir (Hamari ve Sjöblom, 2017). Esporlar, özellikle profesyonel oyuncuların bireysel veya takımlar halinde katılımlarıyla düzenlenen, çok oyunculu (multiplayer) video oyunu müsabakaları şeklinde gerçekleşmektedir. Uluslararası Olimpiyat Komitesi (UOK - International Olympic Committee) Esporu resmi olarak bir spor dalı ilan etmiş ve 2023 yılı haziran ayında Olimpik Espor Haftası düzenleyeceğini duyurmuştur (UOK, 2023). Bu gibi gelişmeler Esporun bilinirliğini geleneksel spor türlerinin seviyesine yaklaştırmaktadır. Geleneksel sporlarda oyuncular bireysel beceri çalışmalarının yanı sıra takım arkadaşlarıyla iş birliği içerisinde hareket etme becerisi kazanma antrenmanları yapmaları gerekmektedir. Espor da aynı felsefe ile benzer beceri gereksinimlerine sahiptir. Bireysel yetenekler ve ekip çalışması kabiliyeti Espor takımlarının oyunu kazanmalarında önemli bir rol oynamaktadır. Her iki spor türünde de oyuncuların bireysel becerilerini geliştirmek, takım içi çalışmalara yardımcı olacak uygulamalar hazırlamak, müsabakalar sırasında gerçekleştirilen hataları tespit etmek, rakip takımların analizini yapmak takımın başarıya ulaşmasında kritik etmenlerdir.

Basketbol veya futbol gibi geleneksel sporlar, oyuncuları ve müsabakaları analiz etmek, takım başarısını etkileyen kıstasları değerlendirmek için belirli veriler toplamaktadırlar. Bu verilerin toplanmasında oyuncuların giydiği sensörler, oyuncu hareketini izleyen kameralar gibi genişletilmiş veri toplama yöntemlerinin yanı sıra takım analistlerinin hazırladığı gözlem notları gibi yöntemler de kullanılmaktadır (Xenopoulos vd., 2021). Bununla birlikte, oyuncuların birçok eyleminin ve müsabakalar sırasında gerçekleşen çeşitli durumların belirli algoritmalar aracılığıyla kayıt altına alınabilmesi bakımından Espor benzersiz bir yapıdadır. Espor denildiğinde akla gelen ilk oyunlardan biri olan Counter-Strike: Global Offensive (CSGO), takım tabanlı (team-based), çok oyunculu, birinci şahıs nişancı (first person shooter) oyunudur. Bu tür oyunlarda, oyun sırasında gerçekleşen birçok eylem ve durum, demo dosyası denilen bir günlük dosyasına kaydedilmektedir. CSGO demo dosyalarının elde edilmesi ve ayrıştırılması diğer oyunların demo dosyalarına göre daha kolay olduğu için oyuncuların ve takımların değerlemesinde Espor analitiği açısından önemli veriler sunmaktadır.

CSGO oyununda her biri beş oyuncudan oluşan iki takım, bir haritayı (map) kazanmaya yetecek kadar raunt (round) kazanmak amacıyla birden fazla rauntta yarışır. Haritanın kazanılması için gerekli raunt sayısı 16'dır. Takımlar oyun başında Teröristler (Terrorists - T) ve Terörle Mücadele (Counter Terrorists - CT) olmak üzere iki tarafa ayrılır. T'nin hedefi raunt süresi bitmeden patlayıcıyı (C4) her haritada belirli olan alanlarından birine yerleştirmek ve patlayıcı patlayana kadar CT'ler tarafından etkisiz hale getirilmesini (defuse) önlemektir. CT'nin hedefi ise patlayıcı yerleştirildikten sonra patlayıcıyı etkisiz hale getirmek veya patlayıcı yerleştirilmeden raunttaki tüm düşmanları ortadan kaldırmaktır. Çoğu maç (match), 1'in en iyisi (best-of-1), 3'ün en iyisi (best-of-3) veya 5'in en iyisi (best-of-5) ayarlarında oynanmaktadır. Buna göre, belirli sayıdaki haritanın çoğunluğunu kazanan takım karşılaşmanın kazananı olmaktadır.

Bir raunt, her iki takımın da satın alma süresi (buy time) boyunca satın alma bölgelerinde (buy zones) hareket edemez şekilde bulunmasıyla başlamaktadır. Takımlar raundun başında, raundu kazanma şanslarını en üst düzeye çıkarmak için zırh (armor), el bombaları (utility) ve silahlar (weapon) satın alırlar. Her raundun başında, her iki takım da bir önceki raundu kazanma durumlarına ve rauntları arka arkaya kaybetme serilerine göre bonus para almaktadırlar. Kaybeden tarafın aldığı bonus para kaybetmeye devam ettiği her raunt için belirli düzeyde artmaktadır. Kazanan taraf ise yüksek kazanma bonusu elde etmesinin yanı sıra hayatta kalan oyuncuların bir önceki raunttan sakladıkları ekipmanları bir sonraki rauntta kullanma şansına sahip olur ve böylece yeniden ekipman satın almak zorunda kalmazlar. Oyunlar sırasında iyi yönetilmiş bir ekonomi, rauntların kazanılmasında doğrudan rol oynayan ekipmanların satın alınmasında, dolayısı ile oyunun galibiyetle sonuçlanmasında çok önemli bir yere sahiptir. Buna ek olarak, bir raunt sırasında raunt süresinin ne kadar kaldığı, patlayıcının yerleştirilme durumu, hayattaki oyuncu sayıları, oyuncuların oyun içerisindeki sağlık durumları, oyuncuların harita içerisindeki konumları gibi birçok unsur rauntların kazanılmasında önemli parametrelerdir.

Esporun 2022 yılındaki pazar büyüklüğü 1.44 milyar dolar olmakla birlikte bunun 2029 yılında 5.48 milyar dolar seviyelerine ulaşacağı ön görülmektedir (Statista, 2023). Geleneksel endüstriler kendilerini sponsor olarak Espor faaliyetlerine dahil etmektedirler. Ancak, endüstrideki en büyük destekleyici güçlerden biri bahis endüstrisidir (Davis, 2021). Pazar hacminin bu denli büyük olduğu ve bahis faaliyetlerinin düzenlendiği Espor alanında teknolojik gelişmelerden faydalanılarak tahminleme çalışmalarının yürütülmesi kaçınılmazdır. Bahis şirketleri maçlardaki bahis oranlarının belirlenmesi gibi işlevlerde tahminleme çalışmalarından yararlanabilmektedirler. Bunlara ek olarak, oyun yapımcıları, rastgele bir araya gelen oyunculardan oluşan takımların kazanma olasılıklarını %50 olarak ayarlayabilmek için derecelendirme sistemi oluştururken bu tür teknolojilerden istifade etmektedirler.

Günümüzde, yapay zekâ teknolojilerindeki ilerlemeler Espor alanında sıklıkla kullanılmaya başlanmıştır. Mevcut durumdaki veriler üzerinden bir öğrenme işlemi gerçekleştirerek daha sonraki durumlar hakkında tahminler üretmeye yarayan makine öğrenmesinin özellikle son yıllarda giderek bilinirliği artmıştır. Makine öğrenmesi, belli bir duruma etki eden parametreler altındaki veriler aracılığıyla yeni durumlar hakkında dengeli tahminlemeler gerçekleştirebilmektedir (Sevli, 2022).

Alanyazında farklı makine öğrenmesi algoritmaları ile Espor maçları çıktılarının tahminlenmesine yönelik bazı önemli çalışmalar bulunmaktadır. Bu araştırmalarda Karar Ağaçları (KA), Naive Bayes (NB), Lojistik Regresyon (LR), K-En yakın Komşu (KNN), Rastgele Orman (RO), Destek Vektör Makinesi (DVM) ve Yapay Sinir Ağları (YSA) gibi algoritmalar sıklıkla kullanılmaktadır. Bu araştırmalar aşağıda özetlenmiştir.

Hood ve diğerleri (2017), Defense of the Ancients 2 (DotA 2) adlı çok oyunculu çevrimiçi savaş arenası (multiplayer online battle arena – MOBA) türündeki oyunda canlı profesyonel maçlar için tahmin modelleri oluşturmuş ve RO algoritması ile %77.51'lik bir doğruluk oranı elde etmiştir. Yang (2018), Overwatch adlı bir video oyununda, takımdaki kahramanların (hero) takım içi uyumlarına odaklanarak kazanan takımı tahmin etmek için bir model oluşturmuştur. Oluşturulan model, maç öncesinden ziyade maç sırasındaki sonucu tahmin etmek için tasarlanmıştır. Araştırmacı, diğer modellerin düşük performans gösterdiği bir aşama olan oyunun erken aşamasındaki tahminin doğruluğunu artırmayı hedeflemiştir. Araştırmada sınıflandırma algoritmalarından LR kullanılmıştır ve oyunun erken aşamasında kazanan takımı tahminleme açısından yaklaşık %58 doğruluğa ulaşmıştır. Makarov ve diğerleri (2018), CSGO'da patlayıcı kurulduktan sonraki senaryolarda, diğer bir deyişle raundun son aşamasında, KA ve LR kullanarak raundun kazananı tahminlemeyi hedeflemişler ve %62'lik bir doğruluk oranı elde etmişlerdir. Xenopoulos ve diğerleri (2020), CSGO'da bir takımın kalan oyuncuları ve ekipman durumları gibi girdi özellikleriyle takımın kazanma olasılığını tahminleyen çalışmalarında XGBoost kullanmışlardır ve %79.1'lik bir doğruluk olanı yakalamışlardır. Shen (2022), MOBA türündeki League of Legends isimli oyunda orta düzeydeki amatör oyuncuların maçlarının ilk 10 dakikasını LR, RO, KA, NB gibi algoritmalarla incelemiş ve %72.68 doğruluğa ulaşmıştır.

Alanyazındaki çalışmalara ek olarak Microsoft ekibi, oyunun kalitesini, oyunun adil olup olmadığını ve bir takımın bu oyunu kazanma olasılığını tahmin etmek için TrueSkill adlı bir derecelendirme sistemi tasarlamıştır (Minka vd., 2018). Bu sistem ile video oyunlarında benzer becerilere sahip oyuncuların eşleştirilmesi amaçlanmıştır. Çeşitli algoritmalar kullanan araştırmacılar %68'lik bir doğruluk oranı elde etmişlerdir.

Bu çalışmada, makine öğrenmesi teknolojisinin Espor alanında nasıl sonuçlar verebileceğinin anlaşılması ve oluşturulan modeller ile CSGO oyununda rauntların sonuçlarını en iyi tahminleyen makine öğrenmesi algoritmasının belirlenmesi amaçlanmıştır. Buna göre, yedi farklı makine öğrenmesi algoritması ile CSGO oyunundaki rauntları hangi takımın kazanacağına yönelik sınıflandırma çalışmaları gerçekleştirilmiştir.

2. Uygulanan Algoritmalar ve Yöntemler
(Applied Algorithms and Methods)

Makine öğrenmesi algoritmaları; yaklaşımları, veri türleri ve problem çözme teknikleri bakımından farklılık göstermektedir. Makine öğrenmesi genellikle (1) Denetimli öğrenme, (2) denetimsiz öğrenme ve (3) pekiştirmeli öğrenme şeklinde üç alt kategoriye ayrılmaktadır (Bengio vd., 2012). Denetimli öğrenme hem girdi hem de çıktı verilerine dayalı bir tahmin modeli geliştirmektedir. Denetimsiz öğrenme, verileri yalnızca girdi verilerine dayalı olarak gruplar ve yorumlarken pekiştirmeli öğrenme, bir ödül ve ceza sistemi kullanarak modeli eğitmektedir. Denetimli öğrenme algoritmaları, sınıflandırma ve regresyon algoritmalarını içermektedir. Çıktı sınırlı bir değer kümesiyle sınırlandırıldığında sınıflandırma, bir aralık içinde herhangi bir sayısal değere sahip olduğunda ise regresyon algoritmaları kullanılmaktadır (Ray, 2019). Son zamanlarda, denetimli ve denetimsiz öğrenmenin bir kombinasyonu olan yarı denetimli öğrenme kavramı sıklıkla kullanılmaktadır. Yarı denetimli öğrenme, veri setlerindeki verilerin bir kısmının etiketlenmiş, büyük bir bölümünün ise etiketlenmemiş olduğu veri setleriyle çalışabilen algoritmaları içermektedir (Zhu ve Goldberg, 2009).

2.1. Denetimli sınıflandırma algoritmaları
(Supervised classification algorithms)

Bu araştırmada Lojistik Regresyon, Karar Ağaçları, Rastgele Orman, XGBoost, Naive Bayes, K-En Yakın Komşu ve Destek Vektör Makinesi olmak üzere toplam yedi adet denetimli sınıflandırma algoritması kullanılmıştır.

2.1.1. Lojistik Regresyon (Logistic Regression)

LR, ikili veya çok sınıflı bağımlı değişkenleri tahmin etmek amacıyla sigmoid fonksiyonlarını işe koşan istatistiksel bir modeldir. Sigmoid (lojistik fonksiyonlar), ikili sınıflandırma problemlerinin istatistiksel analizleri için herhangi bir gerçek sayıyı 0 ile 1 arasındaki aralıkta konumlandırmaktadır. LR, özel olarak verilmiş katsayılar veya ağırlıklar kullanarak, belirli bir girdi açısından olası çıktıyı hesaplayan doğrusal regresyona çok benzemektedir. Tek fark, lojistik regresyonun her zaman 0 veya 1 şeklinde ikili çıktı vermesidir. LR, Eşitlik 1 için $\sigma(a) = \left(1 + exp(-a)\right)^{-1}$ denkleminin bir aktivasyon fonksiyonu olduğu, w_i'nin x_i özelliğine uygulanan ağırlık (katsayı) olduğu ve X'in n adet özelliğe sahip bulunduğu her sınıf açısından olasılıkları tahmin etmek için veri özniteliklerinin lojistik bir fonksiyonunu kullanmaktadır (Böhning, 1992).

$$P(win) = \sigma\left(w_0 \sum_{i=1}^{n} w_i x_i\right) \qquad (1)$$

2.1.2. Karar Ağaçları (Decision Trees)

KA, hem sınıflandırma hem de regresyon görevleri için kullanılan, parametrik olmayan denetimli bir öğrenme algoritmasıdır. Kök düğüm, dallar, iç düğümler ve yaprak düğümlerden oluşan hiyerarşik bir ağaç yapısına sahiptir. Bir karar ağacı, herhangi bir gelen dalı olmayan kök düğümle başlamaktadır. Kök düğümden giden dallar daha sonra karar düğümleri olarak da bilinen iç düğümleri beslemektedir. Mevcut özelliklere bağlı olarak, her iki düğüm türü de yaprak düğümler veya uç düğümler tarafından belirtilen homojen alt kümeler oluşturmak için değerlendirmeler yürütür. Yaprak düğümler, veri kümesindeki tüm olası sonuçları temsil eder. Bu ağaç yapısı formüle edilerek bir model oluşturulur (Priyam vd., 2013).

2.1.3. Rastgele Orman (Random Forest)

RO, birden fazla karar ağacının çıktısını tek bir sonuca ulaşmak için birleştiren, yaygın olarak kullanılan bir makine öğrenme algoritmasıdır. RO, geleneksel KA algoritmalarında sıklıkla karşılaşılan aşırı öğrenme (overfitting) problemini, veri setini ve öznitelikleri birçok parçaya ayırıp birden fazla ağaç üzerinde işleyerek çözüme kavuşturmaktadır (Biau ve Scorner, 2016).

2.1.4. Gradyan Artırıcı Karar Ağacı (Extreme Gradient Boosting)

Gradyan Artırıcı Karar Ağacı (XGBoost) ağaç tabanlı bir algoritmadır ve denetimli öğrenme problemleri için kullanılmaktadır. Topluluklar (ensemble), karar ağacı modellerinden oluşturulmaktadır. Ağaçlar topluluğa birer birer eklenerek önceki modellerin yaptığı tahmin hatalarını düzeltmek için eğitilir. Bu işlem süreci, zayıf öğrenenleri (weak learner) güçlü öğrenenlere (strong learner) dönüştürüldüğü, artırıcı olarak adlandırılan bir tür toplu makine öğrenmesi modelidir. Modeller, herhangi bir isteğe bağlı türevlenebilir kayıp fonksiyonu ve gradyan iniş optimizasyon algoritması kullanılarak eğitilmektedir. Bu işlem, tekniğe "gradyan artırma" adını vermektedir. Bunun nedeni model eğitildikçe kayıp gradyanının, tıpkı bir sinir ağı gibi en aza indirilmesidir (Chen vd., 2015).

2.1.5. Naive Bayes (Naïve Bayes)

NB algoritması, Bayes teoremine dayanan ve sınıflandırma problemlerinin çözümünde kullanılan bir algoritmadır. NB olasılığa dayalı bir sınıflandırıcıdır, yani bir nesnenin olasılığı temelinde tahmin yapmaktadır. NB'nin avantajı, küçük bir eğitim setinde eğitildikten sonra verileri doğru bir şekilde sınıflandırabilmesidir. Basitleştirilmiş varsayımlarına rağmen, NB algoritmaları spam filtreleme ve duygu analizi gibi gerçek dünya durumlarında etkili bir performans sergilemektedir (Zhang ve Li, 2007).

2.1.6. K-En Yakın Komşu (K-Nearest Neighbour)

KNN algoritması, yeni durum ile mevcut durum arasındaki benzerliği tahmin ederek yeni durumu, mevcut sınıflar arasında en çok benzeşen sınıfa atar. KNN'deki "K", sınıflama sürecinde en yakın komşu sayısını ifade eden bir parametredir. KNN eğitim verilerinden ayırt edici fonksiyonları öğrenmek yerine bütün eğitim veri setini ezberlediğinden dolayı tembel bir öğrenendir olarak geçmektedir. Bu nedenle veri setinin büyüdüğü durumlarda işlem yükü artmaktadır. Yöntem temel olarak bir metrik mesafe değerine dayanır. En yaygın kullanılan ölçü Eşitlik 2'de gösterilen Öklid mesafesidir (Cunningham ve Delany, 2021).

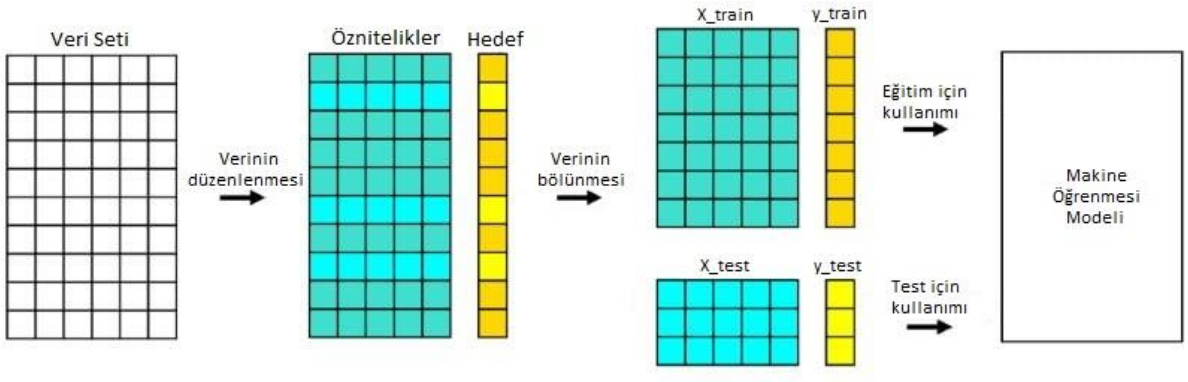
$$d(x,y) = \sqrt{\sum_{i=1}^{n} \sqrt{(y_i - x_i)^2}} \qquad (2)$$

2.1.7. Destek Vektör Makinesi (Support Vector Machine)

DVM, yüksek boyutlara sahip veri kümeleri için etkilidir ve hem sınıflandırma hem de regresyon için kullanılan bir algoritmadır. Bu algoritmanın arkasındaki mantık, n'nin bağımsız değişkenlerin sayısı olduğu n boyutlu uzayda bir hiper düzlem oluşturmaktır. İyi bir modelin, en yakın eğitim veri noktalarına en uzak mesafeye sahip hiper düzlemi izlemesi beklenmektedir. Hiper düzlem aynı zamanda bir sınıflandırma ayracı görevi görmektedir. Bu özellik, mesafe ne kadar büyükse algoritmanın genelleme hatasını o kadar düşürdüğü anlamına gelmektedir (Noble, 2006).

2.2. Veri Doğrulama Yöntemi (Data Validation Method)

Bu araştırmada, CSGO'da oynanan rauntların sonuçlarını tahmin etmede eğitim ve test (train&test) doğrulama yöntemi kullanılmıştır. Eğitim ve test modeli doğrulama yönteminde veri seti; eğitim veri seti, doğrulama veri seti ve test veri seti olmak üzere iki veya üç farklı veri setlerine bölünür. Doğrulama veri seti her zaman kullanılmamakla birlikte kullanıldığı durumlarda genellikle nihai modelin parametre ayarları için işe koşulur. Bu yöntemde modelin eğitim veri setinde öğrenme işlemi gerçekleşir ve ardından test veri setinde değerlendirilmesi yapılır. Eğitim ve test doğrulama yöntemi modelinin akış şeması Şekil 1'de gösterilmiştir.



Şekil 1. Eğitim ve test doğrulama yöntemi (Training & test validation method)

2.3. Algoritmaların Karşılaştırılmasında Kullanılan Performans Metrikleri (Performance Metrics Used in Comparing Algorithms)

Sınıflandırma çalışmaları gerçekleştirildikten sonra en iyi tahminlemeyi yapan algoritmanın belirlenmesinde, doğruluk oranının yanı sıra karışıklık matrisinden (confusion matrix) faydalanılır. True positive (Doğru pozitif - TP), tahminin ve özniteliğin doğru, False positive (Yanlış pozitif – FP) tahminin yanlış ancak özniteliğin doğru, True negative (Doğru negatif – TN) tahminin doğru fakat özniteliğin yanlış, False negative (Yanlış negatif) ise tahminin ve özniteliğin yanlış olduğu durumları ifade etmektedir (Gök, 2017).

Algoritmaların karşılaştırılmasında kullanılan ilk performans metriklerinden biri doğruluk oranıdır (accuracy). Bu oran, yapılan toplam tahmin sayısına göre bir model tarafından yapılan doğru tahminlerin sayısını ölçen bir değerlendirme metriğidir. Doğru tahmin sayısını toplam tahmin sayısına bölerek hesaplanmaktadır. Doğruluk oranının hesaplanmasında kullanılan formül Eşitlik 3'te verilmiştir.

$$Doğruluk = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

Kullanılan performans metriklerinden diğeri ise kesinliktir (precision). Kesinlik, bir makine öğrenmesi modeli performansının, yani model tarafından yapılan doğru tahminde bulunma kalitesinin önemli bir göstergesidir. Kesinlik, doğru pozitiflerin sayısının pozitif tahminlerin toplam sayısına bölünmesiyle elde edilir. Kesinlik oranı Eşitlik 4'teki formül ile hesaplanmaktadır.

$$Kesinlik = \frac{TP}{TP + FP} \tag{4}$$

Duyarlılık (recall), pozitif olarak doğru bir şekilde sınıflandırılan pozitif tahmin sayısının toplamda pozitif olması gereken tahmin sayısına oranı şeklinde hesaplanır. Duyarlılık, modelin pozitif tahminleri tespit etme yeteneğini ölçmektedir. Duyarlılığın hesaplanmasında Eşitlik 5'teki formülden yararlanılmaktadır.

$$Duyarlılık = \frac{TP}{TP + FN} \tag{5}$$

F-Değeri (F-Score), bir modelin kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. F-Değeri, modelin hem pozitif vakaları yakalama (duyarlılık) hem de yakaladığı vakalarda doğru olma (kesinlik) konusundaki denge yeteneğini ortaya koymaktadır. F-Değeri Eşitlik 6'daki formül ile hesaplanabilmektedir.

$$F1 = 2 * \frac{(kesinlik * duyarlılık)}{(kesinlik + duyarlılık)} \tag{6}$$

Alıcı Çalışma Özelliği Eğrisi (Receiver Operating Characteristic Curve – ROC), tüm sınıflandırma eşiklerindeki bir sınıflandırma modelinin performansını resmeden grafiktir. ROC eğrisinin True Positive Rate (Doğru Pozitif Oranı - TPR) ve False Positive Rate (Yanlış Pozitif Oranı - FPR) olmak üzere iki parametresi bulunmaktadır. Sınıflandırma eşiği düşürüldüğünde daha fazla öğenin pozitif olarak sınıflandırılması sağlanır (Boyd vd., 2013). TPR, Eşitlik 7'de ifade edilen formül ile hesaplanmaktadır.

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

FPR ise şu şekilde tanımlanmaktadır:

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

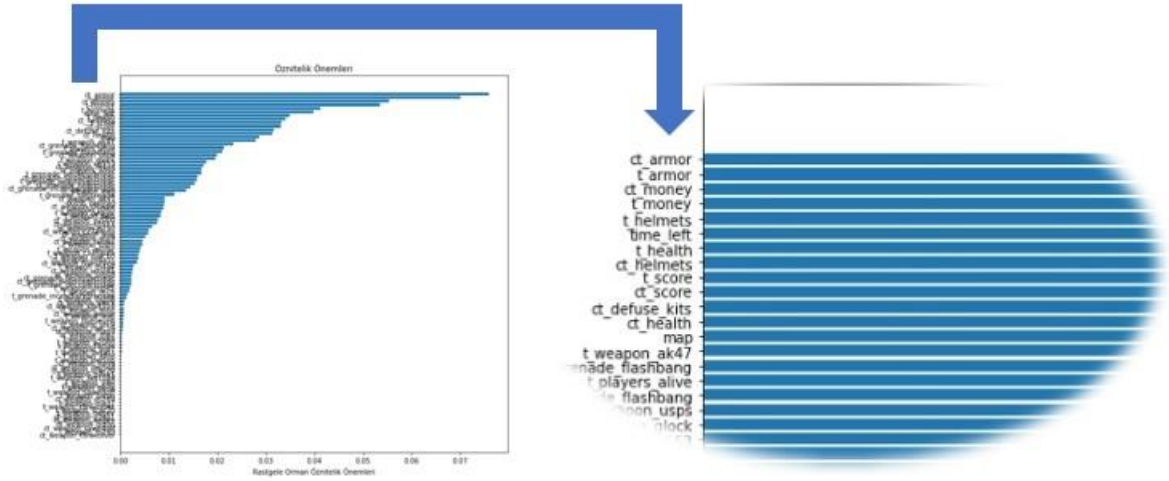
ROC Eğrisi Altındaki Alan (Area Under the Curve Rate – AUC) ise ROC Eğrisi altında kalan alan şeklinde ifade edilmektedir. AUC oranı 0 ile 1 arasında bir değer almaktadır. Hatalı tahminleme oranı %100 olan bir modelin AUC oranı 0'dır. %100 doğru tahminleme yapan bir modelin AUC oranı ise 1 değerini almaktadır (Zhang, 2016).

3. Verilerin Elde Edilmesi ve Hazırlanması **(Data Acquisition and Preparation)**

Yeterli miktarda ilgili veri, iyi bir tahmin modeli oluşturmak için temel koşuldur. Bu bölümde verilerin toplanması hakkında bilgi verilmiş ve veri hazırlama yöntemi sunulmuştur.

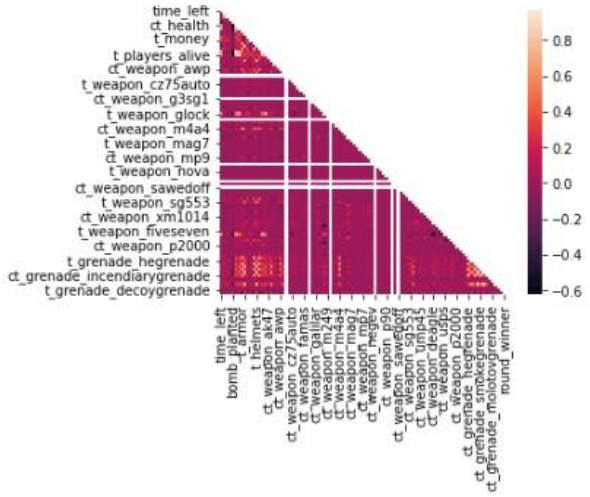
3.1. Verilerin Elde Edilmesi (Data Acquisition)

Çalışmada, Kaggle platformuna Lillelund (2020) tarafından yüklenmiş olan "CS:GO Round Winner Classification" adlı veri seti kullanılmıştır. Veri seti ilk olarak Espor analizleri yürüten Skybox adlı organizasyon tarafından, 2020 ilkbaharından sonbaharına kadar devam eden "CS:GO Yapay Zeka Mücadelesi"nin bir parçası olarak yayımlanmıştır. Veri seti, 2019 ve 2020'de düzenlenen üst düzey turnuva oyunlarından elde edilen 700 demo kaydından oluşturulmuştur. Veri setinde biri raundun kazananı ifade eden sınıf değeri olmakla birlikte toplam 97 öznitelik ve 122.410 örnek bulunmaktadır. Şekil 2'de RO algoritması ile oluşturulmuş, raundun kazananını tahminlemede en önemli öznitelikler ifade edilmiştir.



Şekil 2. Özniteliklerin önemleri (Importance of features)

Şekil 3'te ise veri setinin ısı haritası ve korelasyon matrisi resmedilmiştir.



Şekil 3. Veri setinin ısı haritası (Heatmap of the dataset)

Tablo 1'de, Şekil 2'deki ve Şekil 3'teki verilere göre raundun kazanılmasında önemli rol oynayan 20 özniteliğe ve açıklamalarına yer verilmiştir.

Tablo 1. Veri seti öznitelikleri ve açıklamaları (Dataset features and descriptions)

Öznitelik	Açıklama
ct_armor	CT'lerin toplam zırh değeri
t_armor	T'lerin toplam zırh değeri
ct_money	CT'lerin toplam parası
t_money	T'lerin toplam parası
ct_helmets	CT'de kaç oyuncunun kafa bölgesi zırhının olduğu
t_helmets	T'de kaç oyuncunun kafa bölgesi zırhının olduğu
ct_score	İlgili raunda kadar CT'lerin kazandığı raunt sayısı
t_score	İlgili raunda kadar T'lerin kazandığı raunt sayısı
time_left	Raundun bitimine kalan süre (sn)
ct_defuse_kits	CT'lerin sahip olduğu patlayıcı imha ekipmanı sayısı
ct_health	CT'lerin toplam sağlık değeri
t_health	T'lerin toplam sağlık değeri
map	Maçta oynanan harita 1 = de_dust2 5 = de_overpass 2 = de_mirage 6 = de_train 3 = de_nuke 7 = de_vertigo 4 = de_inferno 8 = de_cache
t_weapon_ak47	T'lerin elinde bulunan AK-47 isimli silah sayısı
ct_grenade_flashbang	CT'lerin elinde bulunan ses bombası sayısı
ct_players_alive	Rauntta hayatta olan CT oyuncu sayısı
t_players_alive	Rauntta hayatta olan T oyuncu sayısı
ct_weapon_usps	CT'lerin elinde bulunan USPS isimli silah sayısı
bomb_planted	Raunt içerisinde patlayıcının yerleştirilmiş olma durumu 1 = yerleştirildi 0 = yerleştirilmedi
round_winner	Raundu kazanan takım 1 = CT kazandı 0 = T kazandı

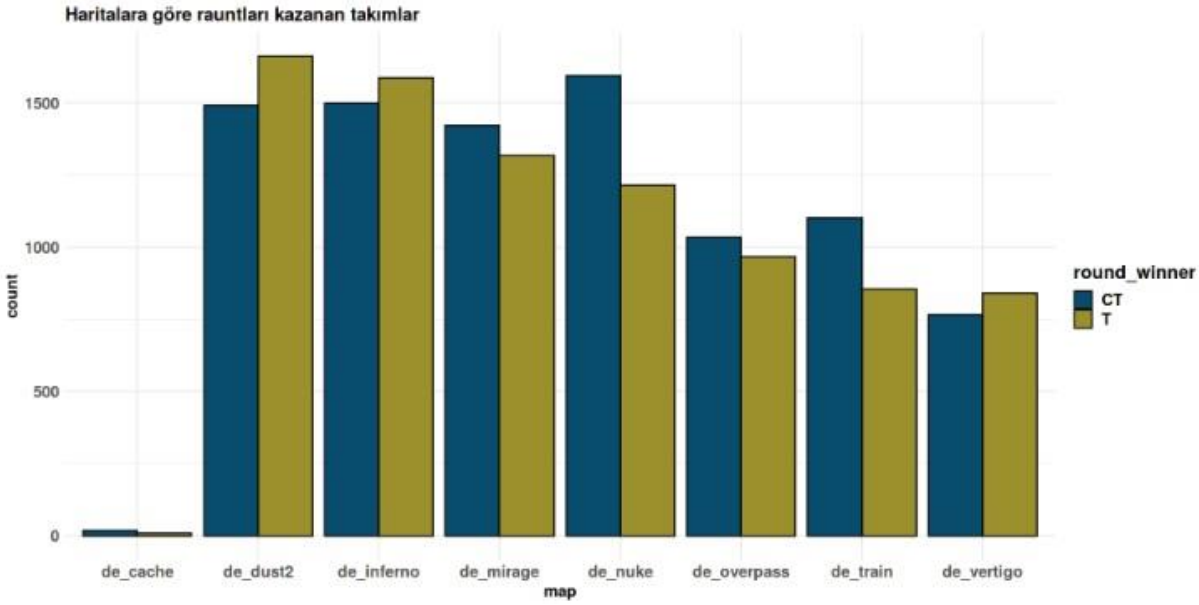
3.2. Verilerin Hazırlanması (Data preparation)

Çalışmada kullanılan veri setinde 96 öznitelik ile bir sınıf değeri bulunmakta ve bu veri seti 122.410 örnekten oluşmaktadır. 96 özniteliğin 2'si kategorik, kalan 94'ü sayısal veri türündedir. Hedef sınıf değeri ise raundu CT kazandı (1) ve raundu T kazandı (0) şeklinde olup kategoriktir. Örneklerin 62.406'sında T takımı, 60004'ünde ise CT takımı raundu kazanmıştır.

Sınıflandırma algoritmalarına geçilmeden önce veri üzerinde bazı ön işlemler gerçekleştirilmiştir. İlk olarak, kayıp veri bulunan bir örnek, veri setinden çıkarılmıştır. Daha sonrasında öznitelik seçimine geçilmiştir. Yalnızca bir değer içeren altı adet öznitelik, oluşturulan modele dahil edilmemiştir. Bütün kategorik değerler sayısal değerler ile kodlanmıştır. Özniteliklerin ortalamaları kaldırılarak ve birim varyansa göre ölçeklendirerek standartlaştırma işlemi yürütülmüştür. Standartlaştırma işleminde merkezleme ve ölçeklendirme, eğitim setindeki örnekler üzerinde ilgili istatistikleri hesaplayarak her öznitelik açısından bağımsız olarak gerçekleşmektedir. Ortalama ve standart sapma değerleri ise dönüşümler kullanılarak daha sonraki verilerde kullanılmak üzere saklanmaktadır. (Ali vd., 2014). Bir veri setinin standardizasyonu, birçok makine öğrenmesi tahmincisi için ortak bir gerekliliktir. Verilerdeki standart dışı dağılımlar modelin tahmin başarısını önemli ölçüde etkileyebilmektedir.

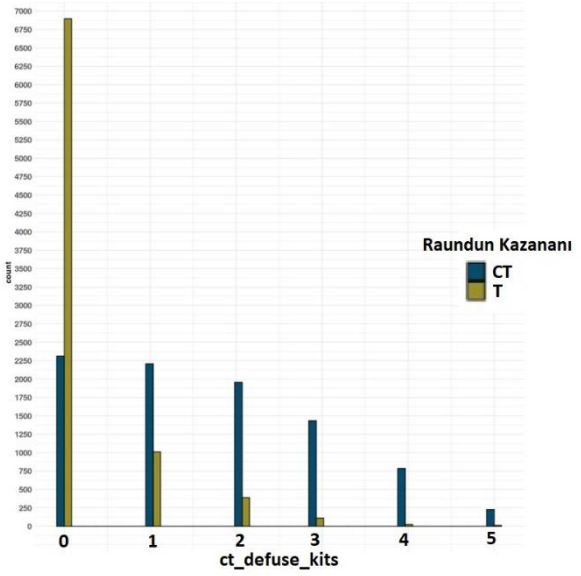
4. Deneysel Çalışma ve Bulgular
(Experimental Study and Findings)

CSGO oyunundaki rauntların kazanan takımını tahminlemeyi amaçlayan bu çalışmada, denetimli sınıflandırma algoritmalarından Lojistik Regresyon, Karar Ağaçları, Rastgele Orman, XGBoost, Naive Bayes, K-En Yakın Komşu ve Destek Vektör Makinesi algoritmaları kullanılmıştır. Model oluşturulurken, veri seti %80 eğitim ve %20 test şeklinde olmak üzere iki parçaya bölünmüştür. Kullanılan bütün algoritmalarda rastgele durum (random state) 42 olarak belirlenmiştir. RO'da orman sayısı 12 olarak tanımlanmıştır. KNN'de komşu sayısı 5 olarak girilmiştir. XGBoost algoritmasında öğrenme oranı 0.01, tahminleyici sayısı 10, tohum (seed) sayısı ise 25 olarak belirlenmiştir. DVM'de çekirdek (kernel) olarak Radyal Tabanlı İşlev Çekirdeği (Radial basis function – RBF), C parametresi 2 olarak ayarlanmıştır. DVM optimizasyonunda C parametresi, her bir eğitim örneğinin yanlış sınıflandırılmasından ne kadar kaçınılacağını belirtmektedir. Şekil 4'te, ilgili dönemde aktif harita havuzunda bulunan 8 haritadaki CT ve T taraflarının rauntları kazanma karşılaştırmaları verilmiştir.



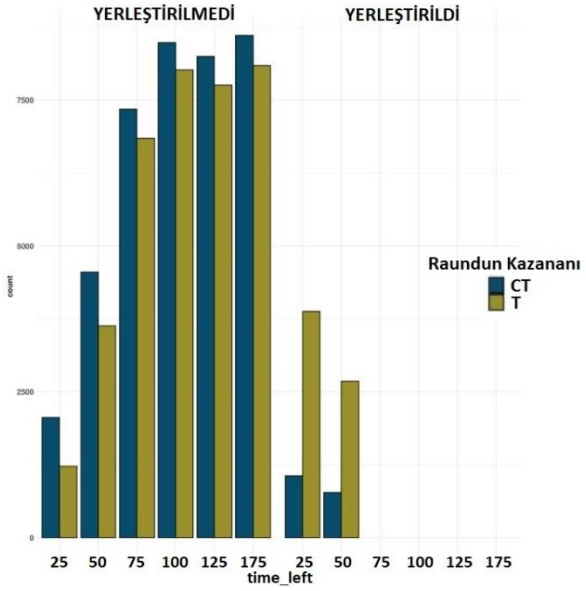
Şekil 4. Haritalara göre rauntları kazanan taraflar (The sides that win the rounds according to the maps)

Şekil 4 incelendiğinde de_nuke ve de_train haritalarının önemli ölçüde CT tarafının, de_dust2 haritasında ise T tarafının avantajlı olduğu görülmektedir. CSGO maçlarında takımlar, haritaya hangi tarafta başlayacaklarının belirlenmesi amacıyla bıçak raundu (knife round) denilen bir seçim raundu oynamaktadırlar. Bu raundu kazanan takım hangi tarafta başlayacağını seçme hakkı elde eder. Takımlar hangi haritaya hangi tarafta başlayacaklarını bu bilgilerden yola çıkarak planlayabilirler. Şekil 5'te patlayıcı imha ekipmanı (defuse kit) sayısının raundu kazanmadaki etkisi verilmiştir.



Şekil 5. Patlayıcı imha ekipmanı sayısına göre raundu kazanan taraflar (Winners of the round based on the number of defuse kits)

Patlayıcı imha ekipmanı, yalnızca CT tarafındaki oyuncuların satın alıp kullanabildiği, her oyuncuda en fazla 1 adet, raunt içerisinde ise toplamda 5 adet bulundurulabilen, patlayıcının imha edilme süresini 10 saniyeden 5 saniyeye indiren bir ekipmandır. CT'lerin bu ekipmanı hiç satın almadığı durumlarda T tarafının raundu kazanma olasılığının çok büyük bir ölçüde arttığı görülmektedir. Ekipman sayısı çoğaldıkça T tarafının raunt kazanma sayıları da giderek düşmüştür. Bu durum yalnızca imha ekipmanının patlayıcıyı hızla imha edilmesine olanak tanıması kaynaklı bir avantaj değildir. Üzerinde bu ekipman bulunan bir oyuncu öldürüldüğünde ekipman yere düşmektedir ve diğer takım arkadaşları tarafından üzerlerine alınabilmektedir. Dolayısı ile imha ekipmanın bir veya iki oyuncuda olması genelde yeterli görülen bir durumdur. Ancak, takım ekonomisinin güçlü olduğu durumlarda her oyuncu bu ekipmanı edinme yoluna gidebilmektedir. Diğer bir deyişle bir rauntta bulunan imha ekipmanı ne kadar fazla ise CT tarafının ekonomisi o kadar güçlü ve raundu kazanma olasılığı buna paralel olarak fazla denilebilir. Şekil 6'da patlayıcının yerleştirilme durumuna göre raundu kazanan taraflar verilmiştir.



Şekil 6. Patlayıcının yerleştirilme durumuna göre raundu kazanan taraflar (Winners of the round based on explosive planting situation)

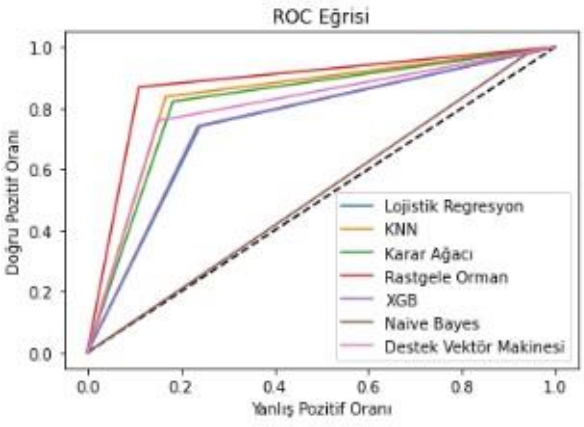
Patlayıcının yerleştirilmediği senaryoda, raunt süresi bitiminde en az bir CT tarafındaki oyuncunun hayatta olması durumunda raundu CT'ler kazanmaktadır. T tarafı patlayıcıyı yerleştirme imkânı bulamıyorsa süre bitmeden önce raunttaki bütün CT oyuncularını ortadan kaldırma yoluna gitmelidir. Patlayıcının yerleştirildiği senaryoda ise süre CT tarafının aleyhine işlemektedir. Patlayıcı yerleştirildikten 40 saniye sonra patlamaktadır ve raunt bitmektedir. Patlayıcı yerleştirildikten sonra bütün T oyuncuları ortadan kaldırılsa dahi patlayıcının patlaması durumunda raundu T tarafı kazanmaktadır. Şekil 6 incelendiğinde, patlayıcının yerleştirilmediği senaryoda T tarafı raundu kazanma konusunda CT'lerin gerisinde kalmıştır. Başta bahsedilen durumdan ötürü raunt süresi azaldıkça T'lerin raundu kazanma ihtimali de giderek azalmıştır. Patlayıcının yerleştirildiği senaryoda ise tam tersi bir durum söz konusudur. Bu durumda T'ler raundu kazanma yönünden büyük bir avantaj elde etmekte ve patlayıcının patlamasına kalan süre azaldıkça T'lerin raundu kazanma ihtimalleri artmaktadır.

Çalışmada kullanılan sınıflandırma algoritmalarına göre veriler modellenerek tahminleme işlemleri gerçekleştirilmiştir. Bu işlemler sonucunda modellere ait doğruluk, kesinlik, duyarlılık, F-Skor ve AUC değerleri hesaplanmış ve Tablo 2'de verilmiştir.

Tablo 2. Sınıflandırma algoritmalarının performans değerleri (Performance values of classification algorithms)

Sınıflandırıcı	Doğruluk	Kesinlik	Duyarlılık	F-Skor	AUC
LR	0.75	0.76	0.74	0.75	0.84
KNN	0.83	0.84	0.84	0.83	0.91
KA	0.82	0.82	0.82	0.82	0.82
RO	0.88	0.89	0.87	0.88	0.95
XGBoost	0.75	0.77	0.74	0.75	0.89
NB	0.53	0.52	0.98	0.67	0.76
DVM	0.80	0.84	0.76	0.79	0.88

Tablo 2 incelendiğinde en yüksek doğruluk oranına sahip sınıflandırıcının %88 ile RO olduğu anlaşılmaktadır. Bunu %83 ile KNN ve KA algoritmaları takip etmektedir. En düşük doğruluk oranına sahip algoritma ise %53 ile NB'dir. Kesinlik metriği açısından bakıldığında en yüksek oran %89 ile RO algoritmasına aittir. NB doğruluk oranı en düşük algoritma olsa da duyarlılık değeri bakımından %98'lik bir değer ile en yüksek orana sahiptir. Duyarlılık ve kesinlik değerleri en dengeli algoritma RO algoritmasıdır (F-Skor = 88). AUC değerlerine bakıldığında %95 ile RO algoritması en yüksek değeri almıştır. Şekil 7'de, sınıflandırıcılara ait ROC eğrisi grafiği verilmiştir.



Şekil 7. Sınıflandırıcılara ait ROC eğrisi grafiği (ROC curve graph of classifiers)

ROC analizi sonucunda çizdirilen Şekil 7'deki grafik incelendiğinde en yüksek doğruluk veren kesim (cut-off) noktasına sahip algoritmanın RO olduğu, onun arkasında ise KNN algoritmasının yer aldığı görülmektedir.

5. Sonuçlar (Conclusions)

Bu araştırmada, günümüzde Espor karşılaşmalarının çok önemli bir bölümünü oluşturan CSGO oyunundaki rauntların kazanan taraflarının tahminlenmesi konusunda makine öğrenmesi tabanlı bir çalışma gerçekleştirilmiştir. Araştırmanın deneysel kısmında ilk olarak sıcaklık haritası ve korelasyon matrisine göre raundun kazanılması bağlamında etkili olduğu görülen parametreler Keşifsel Veri Analizleri (Exploratory Data Analysis – EDA) yoluyla görselleştirilmiştir. Buna göre, CSGO oyununda de_nuke ve de_train haritaları CT ağırlıklı haritalar iken de_dust2 haritası T tarafı ağırlıklıdır. Patlayıcı imha ekipmanları CT'lere raundu kazandırma konusunda ciddi bir işleve sahiptir. Patlayıcının yerleştirilmediği durumda CT tarafı T'ye göre avantajlı durumdadır ancak, patlayıcı yerleştirildikten itibaren T tarafı avantaj durumunu büyük ölçüde lehine çevirmektedir.

Çalışmada sınıflandırma algoritmalarından Lojistik Regresyon, K-En Yakın Komşu, Karar Ağaçları, Rastgele Orman, XGBoost, Naive Bayes ve Destek Vektör Makinesi olmak üzere yedi algoritma kullanılmıştır. Kurulan modellerin etkili bir sınıflandırma yaptığının anlaşılması açısından performans değerlerinin 1'e yakın değerler üretmesi gerekmektedir. Kesinlik, sınıflandırıcıların pozitif olarak tayin ettiği örneklerin gerçekte kaç tanesinin pozitif olduğunu hesaplayan bir metriktir. Duyarlılık, toplam pozitif durumların ne kadarlık bir bölümünün başarılı tahmin edildiğini hesaplamaktadır. Kesinlik ve duyarlılık arasındaki denge ise F-Skor metriğiyle ifade edilmektedir. Bunlara ek olarak ROC analizi, iki ya da daha fazla sınıflandırıcı algoritmanın tahminleme güvenirliklerini karşılaştırmak amacıyla kullanılmaktadır. AUC değeri ise model performansının bir özeti niteliğinde olup, algoritmaların tahminlemede ne kadar başarılı olduğunu sunmaktadır. Bu bilgilerin ve araştırmanın bulguları ışığında CSGO'da rauntların kazananını tahminlemede en başarılı algoritma Rastgele Orman olmuştur ve %88 doğruluk değerine sahiptir. Alanyazın incelendiğinde, aynı veri seti ile çalışan Huang ve diğerleri (2022), kullandıkları sınıflandırıcılar arasında XGBoost ile en yüksek %79'luk doğruluk değerine ulaşmışlardır. Araştırmacıların elde ettiği sonuç ile bu araştırmanın sonuçları arasındaki farkın temelinde araştırmacıların modellerini oluştururken öznitelik seçimleri sırasında toplamda 97 olan öznitelik sayısını 26'ya indirgemeleri olduğu düşünülmektedir. Bunun yanı sıra verilerin normalizasyon işlemleri, kullanılan algoritmalar ve algoritmalara ait parametre seçimleri bu farkın oluşmasında önemli etmenler arasındadır. CSGO üzerinde rauntların veya takımların kazananlarını tahminlemeye yönelik alanyazındaki diğer çalışmalara bakıldığında Makarov ve diğerleri %62'lik, Xenopoulos ve diğerleri ise %79.1'lik doğruluk oranları elde etmişlerdir. Farklı veri setleriyle çalışılmış olsa da bu araştırma kapsamında yapılan modellemeler, bahsedilen iki çalışmadan önemli ölçüde ayrışarak daha doğru bir tahminleme yapmaktadır.

Araştırma, CSGO oyununda raundu kazanan tarafların tahminleme başarısı açısından Espor sektörüne yönelik önemli çıktılar sunmaktadır. Araştırma kapsamında gerçekleştirilen modellerin canlı maçlarla bütünleştirilmesi durumunda izleyenlere sunulacak anlık bir tahmin bilgisi seyir zevkinin artmasına yol açabilir. Bunun yanı sıra takımlar; harita seçimi, oyun içi ekipman alımları, oyun içi senaryolarda uygulayacakları taktikler yönlerinden araştırmadan elde edilen bulguları kullanabilirler. Çalışmanın ayrıca, makine öğrenmesi sınıflandırıcılarının video oyunlarında nasıl performans gösterdiğinin karşılaştırılması, makine öğrenmesinin Espor sektöründe hangi çıktılara olanak tanıyabileceğini ortaya koyması açısından önemli bulguları bulunmaktadır.

Kaynaklar (References)

Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. Mach Learn Tech Rep, 1(1), 1-6.

Bengio, Y., Courville, A. C., & Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. CoRR, 1(2665), 2012.

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197-227.

Boyd, K., Eng, K. H., & Page, C. D. (2013, September). Area under the precision-recall curve: point estimates and confidence intervals. In Joint European conference on machine learning and knowledge discovery in databases (pp. 451-466). Springer.

Böhning, D. (1992). Multinomial logistic regression algorithm. Annals of the institute of Statistical Mathematics, 44(1), 197-200.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

Cunningham, P., & Delany, S. J. (2021). K-nearest neighbour classifiers-a tutorial. ACM Computing Surveys (CSUR), 54(6), 1-25.

Davis, W. (2021). As esports grows, so too do its sponsorships. URL https://win.gg/news/as-esports-grows-so-too-do-its-sponsorships (Erişim tarihi: 28.12.2022)

Gök, M., 2017. Makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 5(3), 139-148.

Hamari, J. & Sjöblom, M. (2017). What is eSports and why do people watch it? Internet Research, 27(2), 211-232. https://doi.org/10.1108/IntR-04-2016-0085

Hodge, V. J., Devlin, S., Sephton, N., Block, F., Cowling, P. I., & Drachen, A. (2019). Win prediction in multiplayer esports: Live professional match prediction. IEEE Transactions on Games, 13(4), 368-379.

Huang, W. X., Wang, J., & Xu, Y. (2022, April). Predicting round result in Counter-Strike: Global Offensive using machine learning. In 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP) (pp. 1685-1691). IEEE.

Lillelund, C. (2020). CS:GO round winner classification. URL https://www.kaggle.com/datasets/christianlillelund/csgo-round-winner-classification (Erişim tarihi: 08.12.2022)

Makarov, I., Savostyanov, D., Litvyakov, B., & Ignatov, D. I. (2018). Predicting winning team and probabilistic ratings in "Dota 2" and "Counter-Strike: Global Offensive" video games. In International Conference on Analysis of Images, Social Networks and Texts (pp. 183-196). Springer, Cham.

Minka, T.P., Cleven, R., & Zaykov, Y. (2018). TrueSkill 2: An improved Bayesian skill rating system. Technical Report. https://www.microsoft.com/en-us/research/uploads/prod/2018/03/trueskill2.pdf

Noble, W. S. (2006). What is a support vector machine?. Nature Biotechnology, 24(12), 1565-1567.

Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. International Journal of current engineering and technology, 3(2), 334-337.

Ray, S. (2019, February). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

Sevli, O. (2022). Farklı sınıflandırıcılar ve yeniden örnekleme teknikleri kullanılarak kalp hastalığı teşhisine yönelik karşılaştırmalı bir çalışma. Journal of Intelligent Systems: Theory and Applications, 5(2), 92-105.

Shen, Q. (2022, February). A machine learning approach to predict the result of League of Legends. In 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE) (pp. 38-45). IEEE.

Statista. (2023). eSports market size worldwide in 2021, with a forecast for 2022 and 2029. URL https://www.statista.com/statistics/1256162/global-esports-market-size/ (Erişim tarihi: 04.01.2023)

UOK. (2023). IOC confirms Singapore as host of first Olympic Esports Week in June 2023. URL https://olympics.com/en/news/ioc-confirms-singapore-host-first-olympic-esports-week-june-2023 (Erişim tarihi: 08.01.2023)

Xenopoulos, P., Coelho, B., & Silva, C. (2021). Optimal Team Economic Decisions in Counter-Strike. arXiv preprint arXiv, abs/2109.12990.

Xenopoulos, P., Doraiswamy, H., & Silva, C. (2020, December). Valuing player actions in counter-strike: Global offensive. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 1283-1292). IEEE.

Yang, B. (2018). Predicting e-sports winners with machine learning. URL https://blog.insightdatascience.com/hero2vec-d42d6838c941 (Erişim tarihi: 22.12.2022)

Zhang, H., & Li, D. (2007, November). Naïve Bayes text classifier. In 2007 IEEE international conference on granular computing (GRC 2007) (pp. 708-708). IEEE.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine, 4(11), 218.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, 3(1), 1-130.

Research Article
Journal of Intelligent Systems: Theory and Applications 6(2) (2023) 130-142
DOI: 10.38016/jista.1209415

Analyzing Big Social Data for Evaluating Environment-Friendly Tourism in Turkey

Mahmud Alrahhal[1*], Ferhat Bozkurt[2]

[1,2]Atatürk University, Department of Computer Engineering, Erzurum, Türkiye

alrahhal24@gmail.com, fbozkurt@atauni.edu.tr

Abstract
Tourism in Türkiye is fundamentally important for both the Turkish economy and travelers. Green tourism has gained increasing attention in the last few years. Analyzing big social data for evaluating environment-friendly tourism in Türkiye is important to gain an understanding of the factors impacting travelers' intention to echo-friendly hotels. To meet the goal of the study, the data was retrieved from the Tripadvisor website using a crawling technique. Machine learning techniques, particularly Latent Dirichlet Allocation (LDA), were utilized to discover satisfaction dimensions from the user-generated content. The k-means clustering approach was deployed for data segmentation. Finally, the online reviews classification model was trained and compared using Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The obtained results reveal several important dimensions that impact tourists' experience.

Keywords: LDA topic model, machine learning, text mining, segmentation, online customers' reviews.

1. Introduction

The tourism sector in Türkiye is furnishing increasingly. One of the most desired choices for travelers is going and unwinding in Türkiye. Many travelers prefer to spend their holiday in Türkiye due to the geographical position of Türkiye which is located in the heart of the world between Asia, and Europe, and it is close to Africa as well, along with its astonishing nature like long coasts, ancient structures, antiques, nice weather, and last but not least echo-friendly hotels. In 2019 alone, a total of 51.7 million travelers to Türkiye were recorded, with around 34.5 billion dollars in income, ranking the sixth worldwide in terms of the total number of tourists (Tuna & Başdal, 2021). These days, people are more willing to discover nature, since it improves their living standards as a consequence of relaxation, health, and taking advantage of environmental services (Prihayati & Veriasa, 2021). For several individuals, tourism is essential, and progressively earning importance. United Nations World Tourism Organization (UNWTO) conducted a survey revealing that near to 1.4 billion people traveled in 2019 (Streimikiene et al., 2021).

Adopting green services and products has increasingly become a center point in the growing businesses, with several organizations deploying environmental sustainability as a critical aspect of their main marketing policy (Chen et al., 2022). In the tourism sector, the green tourism concept has emerged and gained increasing attention (Filimonau et al., 2022; Yeşiltaş et al., 2022). Strong investigation of environmental issues, and looking for solutions to deal with such issues most of the time leads to a strong intention of travelers to perform echo-friendly actions to save the environment (Han et al., 2018). The initial expectations of the service given by a green hotel are primarily important for the customers. Measurement among customers' initial expectations and their actual experience of the product can describe customer satisfaction (Yu et al., 2022). Social big data analysis and online reviews are crucial to discover customers' expectations about many features located in environment-friendly hotels in Türkiye. Features extracted from online reviews will assist governments and decision-makers to know what are properties the customer interested in. Evaluating online reviews on green tourism, and environment-friendly hotel sites are substantially critical to enhancing both the Türkiye tourism sector and travelers' satisfaction. Although researchers have conducted many approaches and methods evaluating online reviews for tourism purposes, analyzing online reviews for environment-friendly hotels in Türkiye hasn't been investigated widely.

* Corresponding Author.
E-mail: alrahhal24@gmail.com

Recieved : 24 Nov 2022
Revision : 17 Feb 2023
Accepted : 31 Mar 2023

Online review extraction and mining have been the focus of researchers in natural language processing in the last few years (Afrizal et al., 2019). This is explained by the increasing popularity of online reviews among tourists, as 90% of them utilize these reviews to reach the travel decision and plan their trips (Godnov & Redek, 2016). Analyzing Social big data is fundamentally crucial to both customers and business owners. Big social data have a major impact on evaluating customer satisfaction with green hotels in Türkiye. Analysis of online reviews for green tourism in Türkiye will help decision-makers and hotel owners to define the features that the customers are interested in. Utilizing these shared online reviews which exist on hotel sites with a huge number of reviews is significant, as a customer is usually able to write any opinion about the hotel without any pressure from business owners or workers at that hotel. Unlike many products that are evaluated by amount or size, hotels are evaluated by experience (Zibarzani et al., 2022). Social big data consists of huge amounts of data, shared on several numbers of social media sites (Nilashi, Abumalloh, Almulihi, et al., 2021). Social big data analysis has been carried out in previous literature by utilizing different advanced approaches and methods to get reasonable assumptions and define market demands. Due to the increase of social big data and online reviews shared on the internet, Natural language Processing (NLP) has become indispensable not only for machine learning (ML) scientists but also for decision-makers in the market.

The main aim of this study is to explore the experience of tourists in environment-friendly hotels in Türkiye based on the content they post on TripAdvisor portal. To meet the goal of the study, we retrieved the data from the TripAdvisor portal, LDA was utilized to discover the dimensions of travelers' experiences, K-means algorithm was used to segment the customers according to their criteria ratings, LSTM, and GRU classification models was deployed and compared. To simplify the reading of this study, we present a list of abbreviations used in this study in Table 1.

Table 1. List of Abbreviations

Abbreviation	Full Term
ML	Machine Learning
NLP	Natural language Processing
LDA	Latent Dirichlet Allocation
UNWTO	United Nations World Tourism Organization
E-WOM	Electronic Word of Mouth
LDA	Latent Dirichlet Allocation
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit

2. Literature Review

2.1. Online reviews

Social media provides a substantial source for interaction, which can be utilized by many research fields like economy, trade, politics, and education (Bozkurt et al., 2019). The usefulness and reliability of online reviews stimulate the endorsement of reviews and the inclination of customers to trust online retailers (Shaheen et al., 2019). Customers' purchase decisions are heavily affected by online reviews (Huang et al., 2019). Customers' preferences located on online reviews for several sides of hotels not only affect customers' booking decisions but also help decision-makers to enhance the service quality of hotels continually (Bian et al., 2022). Online reviews have a considerable impact on determining the pricing strategy and increasing returns (Tian & Zhang, 2022). Online reviews represent a powerful type of communication and electronic word of mouth (E-WOM) where customers not only can spread their opinions but also can discuss their experiences, and it is a strong shape of hotel marketing (Nilashi, Minaei-Bidgoli, et al., 2021). Online reviews' reliability indicates the trustfulness of consumers to a specific review that is being read by the consumers (Wang et al., 2022). There is a powerful impact of online reviews and rating scores on product sales in a short timeframe (Ma et al., 2022). Research on online reviews for airline companies during the COVID-19 pandemic revealed extremely negative consequences due to several problems connected to refund strategies and procedures. Hence, online reviews provide decision-makers with a perception from customers' point of view of how airlines are able to deal with the serious effects of the COVID-19 pandemic (Rita et al., 2022).

2.2. Text mining, and LDA

The text analysis of online reviews ineluctably boosts the concept of "text mining" which indicates the procedure of extracting helpful and beneficial information from the unorganized text (Alzate et al., 2022). However, the textual nature of online reviews presents a complexity in analyzing and interpreting these social data (Alzate et al., 2022). There are diverse approaches for mining textual data, which have been

deployed in the context of analyzing online reviews, including machine learning (Arulraj & Daisy, 2021) and lexicon-based methods (Xianghua et al., 2013). Each approach has its advantages and shortcomings. Machine learning approaches need an advanced level of experience in computational capabilities. As indicated by (Magoulas & Swoyer, 2020), finding skilled machine learning professionals by firms is not an easy task. As the analysis of texts requires specific requirements, Latent Dirichlet Allocation (LDA) was proposed by (Blei et al., 2003) to inspect the topics of textual data and to examine the level of competitiveness between the products. LDA adopts the concept of a "bag of words" to reflect the text as a combination of topics with multinomial dissemination of terms. The document entails topics, each document has topics with its own share and terms' distribution. As an unsupervised learning approach, it can locate the topics to reflect the wisdom of the crowds. Supervised approaches entail learning data that have a target. In context of textual data approaches Including LSTM, and GRU are used.

Recurrent Neural Network (RNN) models are widely utilized in sequential data modeling, including natural language, image/video, captioning, and prediction (Chimmula & Zhang, 2020; Khaldi et al., 2023). LSTM is Long Short-Term Memory Network model which is an extension of RNN (Hochreiter & Schmidhuber, 1997). Since gradient vanishing problem affect the RNN operation when dealing with longer sequence models, LSTM introduces memory cells consisting of different types of gate units, including "output gate", "input gate" and "gate forget" (Liang & Niu, 2022). As a different alternative of LSTM, gated recurrent unit (GRU) is analogous to LSTM in terms of performance, but its computational complexity is lower (Jung et al., 2018). GRU is a simpler, popular, and variant of LSTM and uses the same gate control mechanism as LSTM (Zhao et al., 2017).

2.3. Green tourism

Green hotels have been an interestingly important field of research in recent years, scholars have taken the topic into consideration growingly and increased publications related to the topic have been noticed (Acampora et al., 2022). For customers from several nationalities, there is a positive effect of hotels adopting eco-friendly practices on customers' satisfaction and customers' return inclination to environment-friendly hotels (Berezan et al., 2013). Environment-friendly hotels have a serious impact on customers' satisfaction and return intention (Merli et al., 2019). The definition of green hotels can be described as eco-friendly estates that apply eco-friendly behaviors like water saving, energy saving, and recycling to protect the globe that we inhabit (Association, 2008). Launched by TripAdvisor in 2013, the GreenLeaders program aims to encourage the adoption of green practices in US hotels (Chen et al., 2022). The research on this topic has integrated the tourism field along with sustainability aspects and gained increasing attention (Filimonau et al., 2022). Starting from 2016, the relationship between echo-friendly tourism and customer behaviors has been explored in more than 120 studies (Chen et al., 2022). Environmental issues represented by water and energy consumption, carbon emissions, and waste treatments have gained the attention of policymakers and induced them to focus on the production of green-friendly products and services (Verma et al., 2019)

3. Materials and Methods

3.1. Topic modeling (LDA)

Topic modeling technique utilizes statistical approaches to inspect unstructured texts and investigate the themes from them. This can be achieved through structuring the text within a number of topics that reflect the content of the text using Latent Dirichlet Allocation (LDA). In the LDA technique, the number of topics should be determined, indicating that 20 topics have provided the best performance by several studies (Williams & Betak, 2018). LDA has been deployed in text mining studies to explore online reviews and incorporate customers' perceptions in several areas of research such as marketing (Huang et al., 2022), online education (Wei & Taecharungroj, 2022), and accommodation business (Sim et al., 2021). Figure 1 presents the generative model of the LDA.

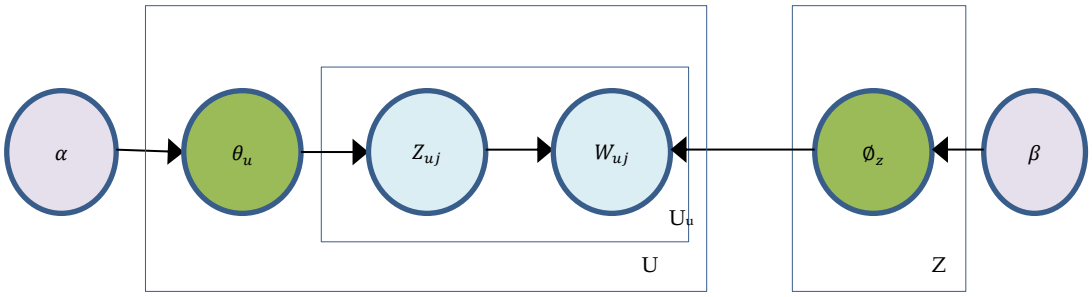


Figure 1. LDA Generative Procedure

Description of Figure 1.

1. For each topic $z \in Z$
   - Draw a multinomial distribution $\emptyset_z \sim Dir(\vec{\beta})$.
2. For every user $u \in U$,
   - Draw a multinomial distribution $\theta_u \sim Dir(\vec{\propto})$.
   - For every Word $w \in D_u$,

(a) Draw a topic $z \sim Multinomial\left(\underset{\theta_u}{\rightarrow}\right)$.

(b) Draw a word $w \sim Multinomial\left(\underset{\emptyset_z}{\rightarrow}\right)$.

3.2. Clustering Approach(K-means)

Following the topic modeling approach, a clustering technique was deployed to separate the user-generated content into several segments. Clustering approaches were utilized in several researches related to user-generated content analysis (Nilashi, Abumalloh, Alghamdi, et al., 2021; Nilashi et al., 2022). The k-means clustering approach was deployed as an iterative clustering technique that finds the optimal cluster center through several iterations (E. Zhang et al., 2022) and as an unsupervised technique. The deployed k-means method is illustrated in algorithm 1. In order to separate the n data into specific groups, the K-means algorithm finds the mean distance between data points. As presented in algorithm 1, the k-means clustering approach repeats the calculation of the distances between data points and assigns centroids to the specified clusters according to the updated distances until convergence.

Algorithm 1: iterative K-means clustering

Input k: the number of clusters, X: A dataset with n data points

Randomly initialize k centroids

Output Set of centroids (μ_z)

Repeat

Assignment of each data point to its closest centroids.

Update the cluster centers (μ_z)

Until convergence

Return (μ_z)

3.3. Classification Model

Finally, In order to classify the reviews and predict the new ones in terms of being negative or positive we deployed and compared LSTM, and GRU machine learning methods. RNN is gaining increasing importance in natural language processing and text classification. The simplest RNN cell is ELMAN which is illustrated in Figure 2, which contains only one hidden layer.

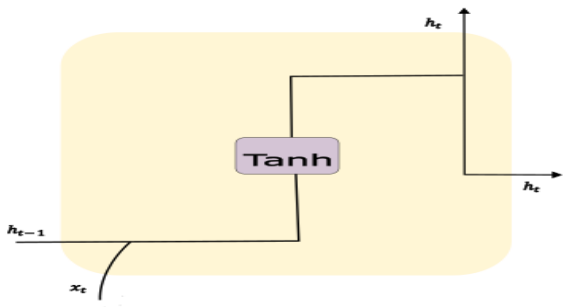


Figure 2. ELMAN Cell (Khaldi et al., 2023)

The output from the hidden layer in the RNN is also used as the input for the next value input along with the input value (Chen et al., 2018). In this way, RNN contains sequence dependency definition, for example, output (h_t) carries a dependency ratio to previous outputs as shown in Figure 2. Therefore, RNN is a successful recurrent neural network in predicting the next value (Wu & Noels, 2022).

The LSTM model was proposed by (Hochreiter & Schmidhuber, 1996) to solve the problem of gradient vanishing in RNN. LSTM offers memory cells consisting of several types of gate units, including "forget gate", "gate gate", and "exit gate" in each recurrent body. As shown in Equations (2.1)-(2.4), the LSTM unit adds input gate i, forgotten gate f, memory unit c, and output gate based on RNN, which significantly enhances the long sequence process performance (Wu et al., 2022). The operation of the LSTM unit is expressed by Equations (2.1)-(2.5).

$$i_t = \sigma(W_i \times [h_{t-1}] + b_i) \quad (1)$$

$$f_t = \sigma(W_f \times [h_{t-1}] + b_f) \quad (2)$$

$$c_t = f_t \times c_{t-1} + i_t \times tanh(W_c \times [h_{t-1}, x_t] + b_g) \quad (3)$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (4)$$

$$h_t = o_t \times \tanh(c_t) \quad (5)$$

At time t i_t, f_t, f_t, c_t, o_t represent the input port, forget port, memory unit, and output port, respectively. H (hidden layer) represents the hidden layer. The x, w, b, and c are represented as input, weight, deviation, and cell respectively. (σ) represents the sigmoid function. An example of the LSTM unit structure was provided in figure 3.

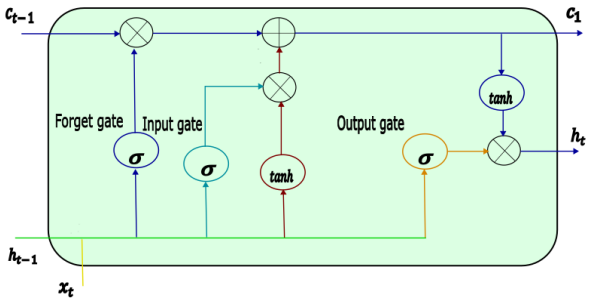


Figure 3. LSTM Structure (Wu et al., 2022)

GRU is similar to LSTM in terms of performance, however, its computational complexity is lower than LSTM and removes cell state, and uses a hidden state to transmit information (Jung et al., 2018). Along with solving the GRU Gradient vanishing problem, it combines the forget gate and input gate in LSTM into the update gate. The GRU consists of two gates, an update gate (update gate z_t) and a reset gate (reset gate r_t). In Figure 4, the structure of the GRU is provided. The operation of the GRU unit is expressed by Equations (2.1)-(2.4).

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (6)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (7)$$

$$h'_t = tanh(W x_t + r_t \odot U h_{t-1}) \quad (8)$$

$$h_t = z_t \odot U h_{t-1} + (1 - z_t) \odot h'_t \quad (9)$$

Where time is referred by t, x_t and h_t are input vectors. The weight matrices $(W_z, U_z), (W_r, U_r), (W_{h'}, U_{h'})$ represent the weights for the reset gate, update gate, and candidate latent state (h'), respectively. Σ represents the sigmoid function, Θ represents the Hadamard product, and tanh represents the hyperbolic tangent function. The structure of GRU was provided in figure 4.

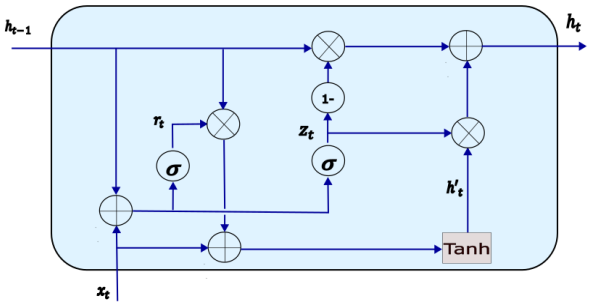


Figure 4. GRU Structure (C. Zhang et al., 2022)

3.4. Data Collection and Preprocessing

TripAdvisor was utilized to obtain the dataset for this research. Customers' online reviews were gathered from different eco-friendly hotel websites located in Türkiye; which are presented on the TripAdvisor platform. The TripAdvisor GreenLeaders program regards the behaviors of hotels towards green practices and ranks them according to 4 levels: Bronze, Silver, Gold, or Platinum; which are displayed notably on the estate's listing on the TripAdvisor site. Properties which demonstrate more green actions are able to get higher TripAdvisor GreenLeaders levels (UNEP, 2013). The crawling technique was employed to crawl TripAdvisor hotel sites using their URLs. Selenium library was utilized to crawl the online reviews from different green hotels located on the TripAdvisor website. Webdriver was imported from the Selenium library and google chrome was utilized for the crawling operation. Reviews located on TripAdvisor hotel sites are distributed with around 10 reviews per page. In the crawling technique, we deployed a loop to navigate through these pages and get the body of the reviews by their data-reviewid XPath then get reviews by their XPath. The looping operation utilized Selenium and for each iteration, to navigate to the other pages the next button of the page was clicked and the next URL was given to the crawler. The operation continues throughout the pages until reaching the last page of the green hotel located on TripAdvisor. Figure 5, illustrates an example of the text-based reviews collected by means of the crawler. The crawler was built to collect customers' online reviews related to the hotels that we aimed to investigate. We gathered 17314 online reviews from different hotels located in Türkiye. The collected reviews' language is English. Collected data was cleaned from useless words or sentences like emails and new line symbols. In addition, gathered data was checked in terms of the existence of null values. We avoided encountering unfamiliar vocables and texts in the results by cleaning the collected data. Criteria ratings have been collected by the crawler and missing values have been filled using the mean of the column to which the data point belongs. Figure 6. illustrates the criteria ratings generated by the

users. The research method followed in this study is presented in Figure 7.

In the customer review classification stage, we noticed that the customers' reviews with 4 overall ratings and higher were positive and the customers' reviews with 3 overall ratings and lower were negative. The dataset was consisting of 951 negatives and 16363 positive reviews. In order to train the model with a balanced dataset we collected more data with negative reviews and eventually, we built a new balanced dataset for the classification model with an overall of 6611 reviews consisting of 3305 positive reviews and 3306 negative reviews. The dataset was separated into train and test, 80% of the dataset was allocated for train and 20% for test. To train the ML model, the customers' reviews were converted to numerical values using Tokenizer API from TensorFlow Keras. Eventually, the sentences are represented by a sequence of numbers using texts_to_sequences from the Tokenizer object.

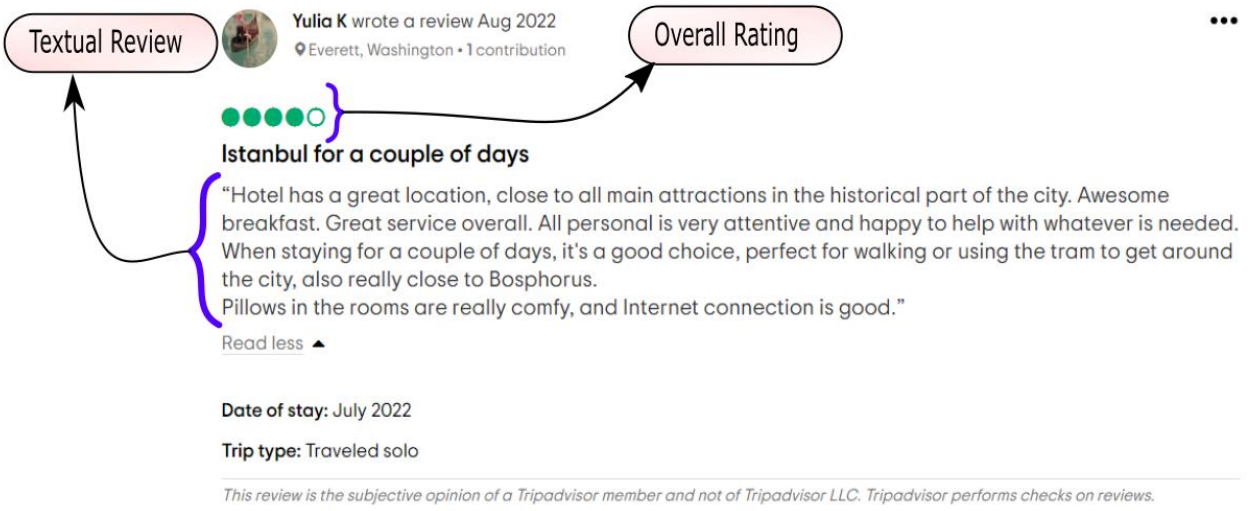


Figure 5. Textual Review

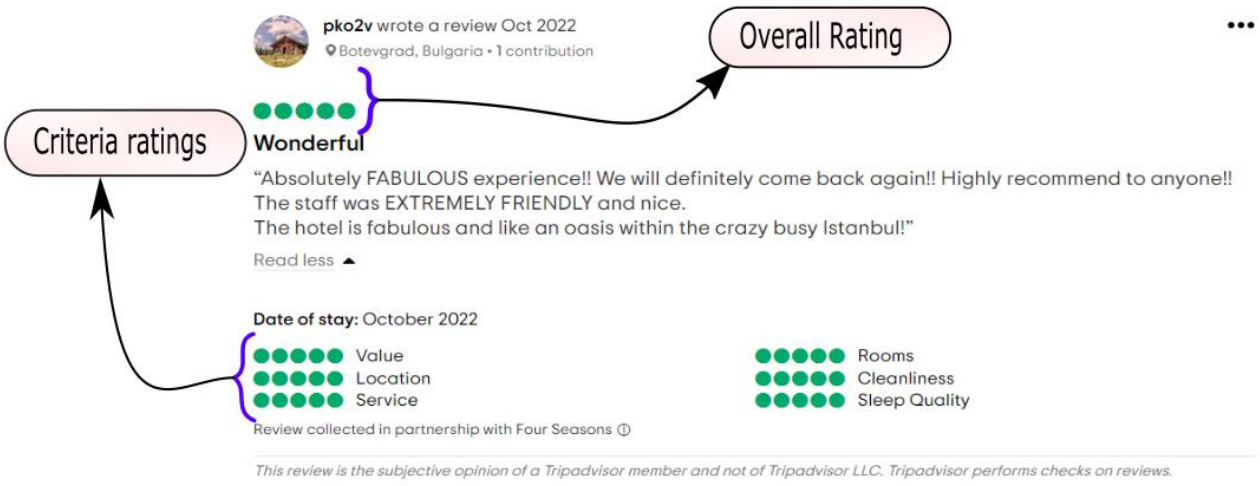


Figure 6. Criteria Ratings

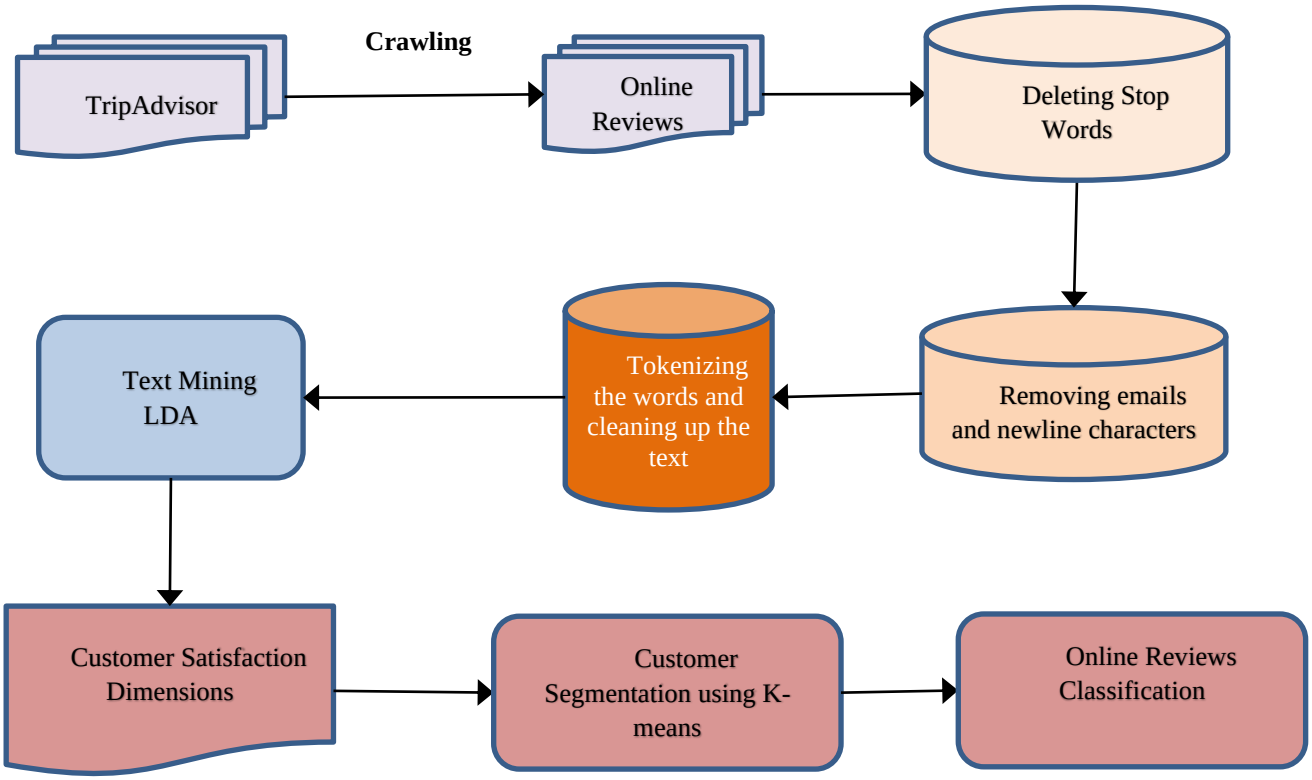


Figure 7. Research Method

4. Data Analysis

Online reviews data collected from TripAdvisor were preprocessed and meaningless words were excluded. LDA is utilized to discover the satisfaction dimensions of customers' online reviews. Gensim library was utilized to create LDA topic modeling method. The analysis of the data entails four main stages; the cleaning and preparation of the social data; the discovery of satisfaction dimensions from the online reviews, customer segmentation based on criteria ratings, and the visualizations of the dimensions. A stage of online reviews classification is provided in this study as well. Online reviews are usually short and the LDA is limited in handling short textual data (Zhang et al., 2021). Hence, a preprocessing stage of the data is essential to improve the performance of the generative model. The preprocessing of the data includes (1) removing the stop words, (2) removing emails and newline characters, (3) tokenizing the words, and (4) cleaning up the text. The stop word list provided by the NLTK package is extended by adding more stop words to the list. The Python package; pyLDAvis was deployed to present the visualizations of the LDA topics. The number of topics was adjusted until we obtained non-overlapping segments of data, which leads to 4 main topics as presented in Figure 3. Besides, to visualize the topics we generated a word cloud of each topic as presented in Figure 9. The circles in Figure 8 represent the topics, in which the size of the topic demonstrates its significance. Figure 9, presents the 4 main topics in the online reviews dataset and the most relevant word distribution related to a specific topic. In this study, 4 topics are generated and 30 keynote words for each topic are obtained. The data cloud for each topic is presented in Figure 9, presenting the higher 15 words in terms of frequency in that specific topic.

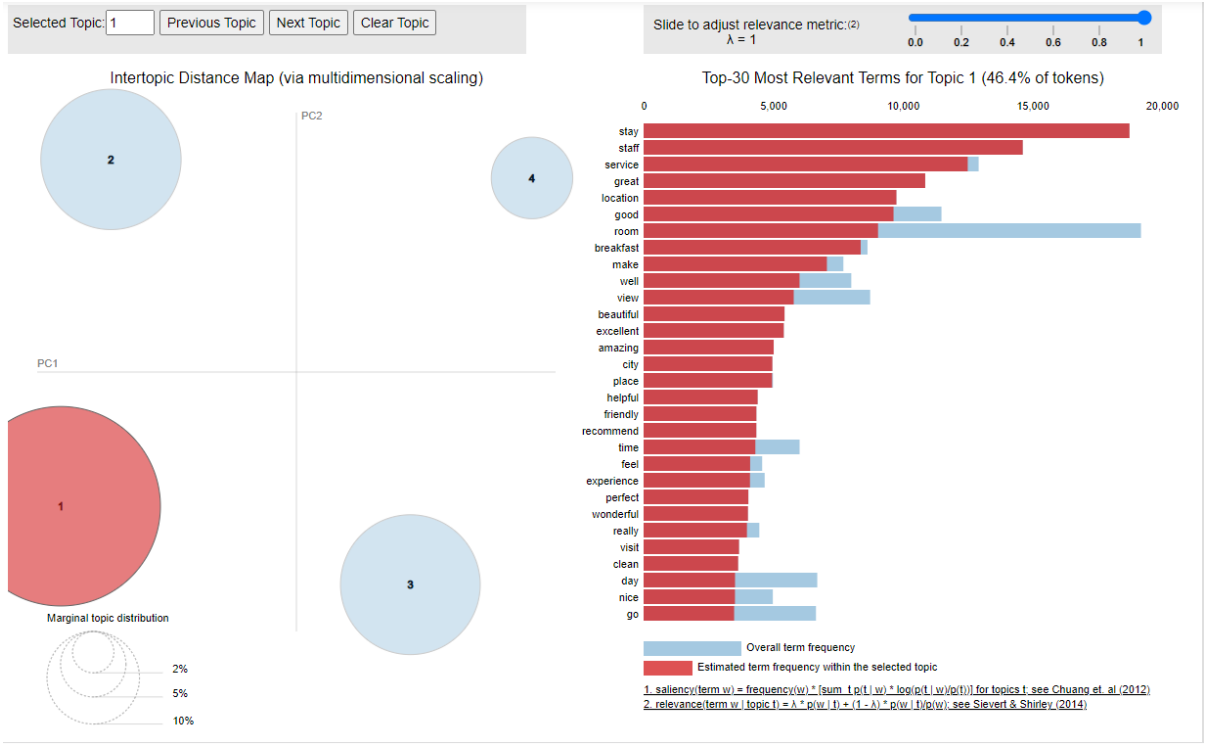


Figure 8. The Main Extracted Topics

Figure 9. Generated Word Cloud

K-means clustering was deployed to separate the customers into groups with similar ratings. In the clustering approach, 3 segments (k = 3) were regarded. The outcomes of k-means clustering are illustrated in Table 2. The Segment1, Segment2, and Segment3 centroids are [27.392175, 36.879851, 23.958665, 33.388009, 37.600293, 35.423475], [40.050183, 42.673045, 43.954379, 42.569097, 46.634343, 42.353479], and [47.844337, 48.088739, 49.676296, 47.976478, 48.700235, 47.985772] respectively. Dividing the customers into segments with similar tendencies through their ratings is important to gain deep insight into the customers' preferences. Along with that, new customers can be assigned to a segment based

on the distance of their ratings to the clusters' centroid. The obtained centroid centers reveal the dimensions which have more impact on customers' satisfaction. For example, the obtained results in Table 2 show that in Segment 1, the customers' ratings in cleanliness criteria are higher compared with others in the same group. In Segment 2, it is obvious that the customers provided moderate criteria ratings for the entire group, the customers have given lower ratings for value criteria than other criteria, therefore, they have indicated their less satisfaction related to value criteria. It is clear that the travelers' satisfaction with cleanliness criteria is high compared with other criteria ratings in Segment 2. In Segment 3, customers' ratings are high in general throughout the group. In segment 3, It is clear that the travelers have been notably happy with the service of the obtained data from the targeted green hotels.

Table 2. Cluster centroids

Attribute	Segment1	Segment2	Segment3
Value	27.392175	40.050183	47.844337
Location	36.879851	42.673045	48.088739
Service	23.958665	43.954379	**49.676296**
Rooms	33.388009	42.569097	47.976478
Cleanliness	**37.600293**	**46.634343**	48.700235
Sleep Quality	35.423475	42.353479	47.985772

LSTM and GRU were deployed to classify customers' reviews into either positive or negative. In both LSTM and GRU, adam optimizer, and sigmoid activation function were used. The used number of epochs for training the model is 15 epochs. The training and validation loss curve of the LSTM model is presented in Figure 10. The accuracy curve of the LSTM model is illustrated in Figure 11. The accuracy obtained for the LSTM model was 0.8670 and the obtained loss was 0.3297. The precision, recall, and f-1 score of the LSTM model was provided in Table 3. It is obvious that the curve of training was decreasing towards zero in the training and validation loss, and increasing towards 1 in the accuracy in both LSTM and GRU models.

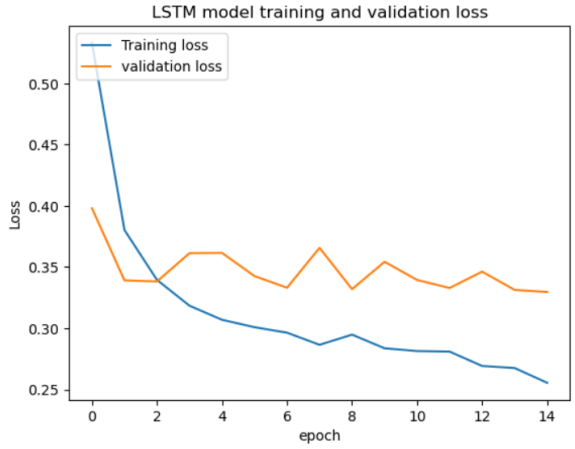


Figure 10. LSTM model loss curve

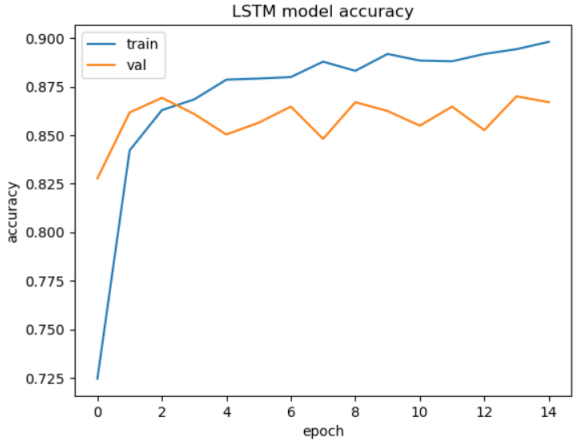


Figure 11. LSTM accuracy curve

Table 3. Results of the LSTM model

Category	Precision	Recall	F-1 score
Negative	0.87	0.87	0.87
Positive	0.86	0.86	0.86

GRU model was trained using the collected balanced dataset as well. GRU model training and validation loss are depicted in Figure 12. GRU model accuracy is illustrated in Figure 13. The obtained accuracy in the case of GRU was 0.8700 and the obtained loss value was 0.3408. The precision, recall, and f-1 score of the trained GRU model was provided in Table 4.

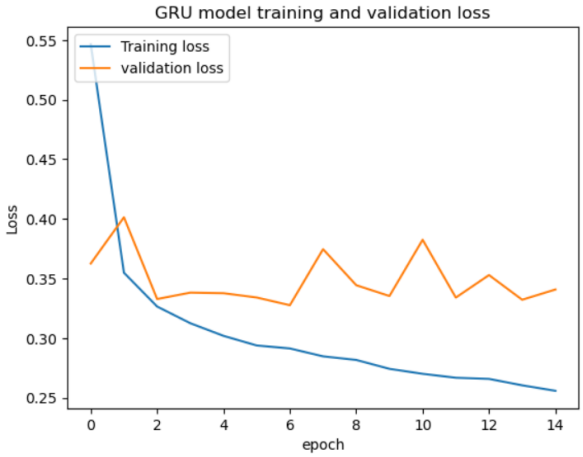


Figure 12. GRU model loss curve

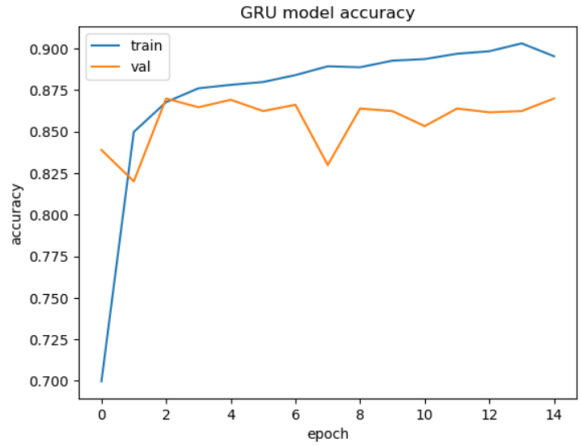


Figure 13. GRU model accuracy curve

Table 4. Results of the LSTM model

Category	Precision	Recall	F-1 score
Negative	0.88	0.87	0.87
Positive	0.86	0.87	0.87

.

5. Results and Discussion

Data collection and machine learning application on the dataset results reveal the main topics and the main features that impact the customers' attention. We modified the number of topics given to LDA until we found that 4 topics is the appropriate number of topics that showed non-intersecting clusters with adequate space between the topics reached. The results of the 4 main topics are depicted in Figure 8, and Figure 9. We can infer from Figure 9. the main criteria and main features that caught the customers' attention. Topic 1 focuses on properties like pools, restaurants, terrace views, etc., Topic 2 concentrates on beverage services like drinks, tea, coffee, etc., Topic 3 words cloud is centered on other quality features such as the time, hour, night, day, etc., and finally, Topic 4 considered general services like staff, breakfast, stay, etc. Following revealing satisfaction dimensions from the obtained dataset, customers were partitioned into 3 segments using the k-means clustering technique. The extracted segment revealed customers' behaviors for each rating criterion. We can infer from cluster centroids in Table 2. that customer satisfaction in Segment1, Segment2, and Segment3 are low, moderate, and high respectively. In Segment1, the customers' satisfaction has been high with cleanliness and location criteria, and low with the service criteria, Generally customers' satisfactiın in Segment1 is relatively low compared with the other two segments. In Segment2 the customers have been more satisfied with the cleanliness criteria and less satisfied with value criteria compared with the other criteria in the same group, and the customers' satisfaction in Segment2 is relatively moderate. Finally, in Segment3 the customers' satisfaction is high, especially for service criteria where the customers have been highly satisfied with the 49.676296 centroid center. cluster centroid. These presented features in the extracted 4 topics and the extracted centroids for each segment can help hotel managers and decision-makers to understand the customers' concerns about green hotels. Hotel managers and decision-makers can know by these topics and features what sections they should enhance in their eco-friendly hotels as well.

As presented in Figure 9, four main topics were extracted from the online reviews, we refer to Topic 1 as facilities-centered, Topic 2 as beverage-centered, Topic 3 as timing-centered, and finally, Topic 4 as services-centered. From this clustering, we can confirm the alignment of our findings with the results of previous literature in similar contexts. The facilities in the hotels such as rooms, pools, and restaurants are vital for the choice of the hotel and the assessment of the overall tourism experience (Bauer et al., 1993). Beverages and drinks also gained the interest of researchers in the tourism and hospitality businesses (Park et al., 2016; Türker & Süzer, 2022). Timing in terms of services such as check out, dining, and room services is important for tourist satisfaction and has been explored in previous literature (De Palma et al., 2018). Finally, the important role of service quality in the hotel industry has been endorsed in previous literature through empirical outcomes (Fan et al., 2022; Harif et al., 2022; Perramon et al., 2022). As presented in table 2. customer segmentation has been utilized in several researches due to the significance of centroids in the prediction of new customers' satisfaction by means of their criteria ratings (Nilashi, Abumalloh, Alghamdi, et al., 2021; Nilashi et al., 2022).

In this study, in order to classify the customers' reviews in terms of being positive reviews or negative we trained LSTM and GRU models. From the results, we can infer that in this experiment GRU model gives an accuracy of 0.8700 which was higher than the accuracy of LSTM with 0.8670. The model shows successful results which is able to recognize whether a customer's review is negative or positive with a high rate of accuracy. LSTM, and GRU which is an extension

of RNN that solved the problem of gradient vanishing play a key role and had been widely utilized in text classification in the literature (Liang & Niu, 2022; Moirangthem & Lee, 2021; Wadud et al., 2022).

Our results support the findings of previous literature in the context of green tourism as these factors were located as important drivers of customers satisfaction in the study by (Bauer et al., 1993; D'Alessandro, 2016; De Palma et al., 2018; Kim et al., 2016; Park et al., 2016; Zamparini et al., 2022).

6. Conclusion

Revealing customers' expectations are pretty important for the tourism sector and particularly for green tourism practices in hotels. In this study, we collected online review data which is considered an important type of big social data generated by users on the TripAdvisor site using a crawling technique that crawled online reviews from hotel sites using their URLs. Gathered data preprocessed, stop words were deleted and extended, extended stop words contained meaningless words and repeated words, emails and newlines were removed, words were tokenized, and the text was cleaned. The most important features that gained tourists' interest were discovered by utilizing the LDA topic modeling technique. In order to understand the customers' behaviors better we partitioned the customers into 3 main groups using the k-means clustering technique. Finally, a new balanced dataset was built in order to be utilized in the classification model. After the data preprocessing stage LSTM and GRU were trained, and GRU was given higher accuracy. Consequently, GRU was deployed.

Traveler satisfaction is fundamentally significant in the tourism sector and particularly in environment-friendly hotels. This study utilized online reviews in echo-friendly hotel sites and applied natural language processing techniques on the online reviews to discover the travelers' satisfaction dimensions. Research findings show t 4 major satisfaction dimensions that we covered in the discussion section. These dimensions are highly important for green hotels to take into consideration. Hotels can enhance the main features in their area based on these dimensions extracted from travelers' online reviews. The research presented insights for decision-makers in the tourism industry by revealing the important factors for tourists' experiences and clustering the customers with similar rating behavior.

7. Limitation of Study and Future Work

The study has a few limitations in terms of the collected data and the deployed method. The data was collected from one online social platform; TripAdvisor; regarding its popularity among tourists, other portals can be utilized to investigate tourists' perceptions more broadly. The deployed method focused on discovering the dimensions of travelers' satisfaction using the LDA, segmenting the customers into groups with similar rating behavior, and training comparing the proposed ML models for the classification of the collected customers' reviews. other research models that entail a survey-based approach, and customer prediction using fuzzy logic approach can present a wider perception of the ranking of the importance levels of the discovered satisfaction dimensions

References

Acampora, A., Lucchetti, M. C., Merli, R., & Ali, F. 2022. The theoretical development and research methodology in green hotels research: A systematic literature review. Journal of Hospitality and Tourism Management, 51, 512-528.

Afrizal, A. D., Rakhmawati, N. A., & Tjahyanto, A. 2019. New filtering scheme based on term weighting to improve object based opinion mining on tourism product reviews. Procedia Computer Science, 161, 805-812.

Alzate, M., Arce-Urriza, M., & Cebollada, J. 2022. Mining the text of online consumer reviews to analyze brand image and brand positioning. Journal of Retailing and Consumer Services, 67, 102989.

Arulraj, T., & Daisy, S. J. S. 2021. Mining online review for predicting sales performance. Materials Today: Proceedings, 47, 93-99.

Association, G. H. 2008. What are green hotels. Retrieved May, 10, 2008.

Bauer, T., Jago, L., & Wise, B. 1993. The changing demand for hotel facilities in the Asia Pacific region. International Journal of Hospitality Management, 12(4), 313-322.

Berezan, O., Raab, C., Yoo, M., & Love, C. 2013. Sustainable hotel practices and nationality: The impact on guest satisfaction and guest intention to return. International Journal of Hospitality Management, 34, 227-233.

Bian, Y., Ye, R., Zhang, J., & Yan, X. 2022. Customer preference identification from hotel online reviews: A neural network based fine-grained sentiment analysis. Computers & Industrial Engineering, 108648.

Blei, D. M., Ng, A. Y., & Jordan, M. I. 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Bozkurt, F., Çoban, Ö., Baturalp Günay, F., & Yücel Altay, Ş. 2019. High performance twitter sentiment analysis using CUDA based distance kernel on GPUs. Tehnički vjesnik, 26(5), 1218-1227.

Chen, Q., Hu, M., He, Y., Lin, I., & Mattila, A. S. 2022. Understanding guests' evaluation of green hotels: The interplay between willingness to sacrifice for the environment and intent vs. quality-based market signals. International Journal of Hospitality Management, 104, 103229.

Chen, W., Yeo, C. K., Lau, C. T., & Lee, B. S. 2018. Leveraging social media news to predict stock index

movement using RNN-boost. Data & Knowledge Engineering, 118, 14-24.
Chimmula, V. K. R., & Zhang, L. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 135, 109864.
D'Alessandro, F. 2016. Green Building for a Green Tourism. A new model of eco-friendly agritourism. Agriculture and agricultural science procedia, 8, 201-210.
De Palma, A., Criado, C. O., & Randrianarisoa, L. M. 2018. When Hotelling meets Vickrey. Service timing and spatial asymmetry in the airline industry. Journal of Urban Economics, 105, 88-106.
Fan, H., Gao, W., & Han, B. 2022. How does (im) balanced acceptance of robots between customers and frontline employees affect hotels' service quality? Computers in Human Behavior, 133, 107287.
Filimonau, V., Matute, J., Mika, M., Kubal-Czerwińska, M., Krzesiwo, K., & Pawłowska-Legwand, A. 2022. Predictors of patronage intentions towards 'green'hotels in an emerging tourism market. International Journal of Hospitality Management, 103, 103221.
Godnov, U., & Redek, T. 2016. Application of text mining in tourism: case of Croatia. Annals of Tourism Research, 58, 162-166.
Han, H., Lee, J.-S., Trang, H. L. T., & Kim, W. 2018. Water conservation and waste reduction management for increasing guest loyalty and green hotel practices. International Journal of Hospitality Management, 75, 58-66.
Harif, M. A. A. M., Nawaz, M., & Hameed, W. U. 2022. The role of open innovation, hotel service quality and marketing strategy in hotel business performance. Heliyon, e10441.
Hochreiter, S., & Schmidhuber, J. 1996. LSTM can solve hard long time lag problems. Advances in neural information processing systems, 9.
Hochreiter, S., & Schmidhuber, J. 1997. Long short-term memory. Neural computation, 9(8), 1735-1780.
Huang, J., Guo, Y., Wang, C., & Yan, L. 2019. You touched it and I'm relieved! The effect of online review's tactile cues on consumer's purchase intention. Journal of Contemporary Marketing Science.
Huang, S., Zhang, J., Yang, C., Gu, Q., Li, M., & Wang, W. 2022. The interval grey QFD method for new product development: Integrate with LDA topic model to analyze online reviews. Engineering Applications of Artificial Intelligence, 114, 105213.
Jung, M., Lee, H., & Tani, J. 2018. Adaptive detrending to accelerate convolutional gated recurrent unit training for contextual video recognition. Neural Networks, 105, 356-370.
Khaldi, R., El Afia, A., Chiheb, R., & Tabik, S. 2023. What is the best RNN-cell structure to forecast each time series behavior? Expert Systems with Applications, 215, 119140.
Kim, J.-Y., Hlee, S., & Joun, Y. 2016. Green practices of the hotel industry: Analysis through the windows of smart tourism system. International Journal of Information Management, 36(6), 1340-1349.
Liang, M., & Niu, T. 2022. Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs. Procedia Computer Science, 208, 460-470.
Ma, G., Ma, J., Li, H., Wang, Y., Wang, Z., & Zhang, B. 2022. Customer behavior in purchasing energy-saving products: Big data analytics from online reviews of e-commerce. Energy Policy, 165, 112960.
Magoulas, R., & Swoyer, S. 2020. AI Adoption in the Enterprise. Beijing: O´ Reilly. Recuperado de http://www. oreilly. com/data/free/ai ….
Merli, R., Preziosi, M., Acampora, A., & Ali, F. 2019. Why should hotels go green? Insights from guests experience in green hotels. International Journal of Hospitality Management, 81, 169-179.
Moirangthem, D. S., & Lee, M. 2021. Hierarchical and lateral multiple timescales gated recurrent units with pre-trained encoder for long text classification. Expert Systems with Applications, 165, 113898.
Nilashi, M., Abumalloh, R. A., Alghamdi, A., Minaei-Bidgoli, B., Alsulami, A. A., Thanoon, M., Asadi, S., & Samad, S. 2021. What is the impact of service quality on customers' satisfaction during COVID-19 outbreak? New findings from online reviews analysis. Telematics and Informatics, 64, 101693.
Nilashi, M., Abumalloh, R. A., Almulihi, A., Alrizq, M., Alghamdi, A., Ismail, M. Y., Bashar, A., Zogaan, W. A., & Asadi, S. 2021. Big social data analysis for impact of food quality on travelers' satisfaction in eco-friendly hotels. ICT Express.
Nilashi, M., Abumalloh, R. A., Minaei-Bidgoli, B., Zogaan, W. A., Alhargan, A., Mohd, S., Azhar, S. N. F. S., Asadi, S., & Samad, S. 2022. Revealing travellers' satisfaction during COVID-19 outbreak: moderating role of service quality. Journal of Retailing and Consumer Services, 64, 102783.
Nilashi, M., Minaei-Bidgoli, B., Alrizq, M., Alghamdi, A., Alsulami, A. A., Samad, S., & Mohd, S. 2021. An analytical approach for big social data analysis for customer decision-making in eco-friendly hotels. Expert Systems with Applications, 186, 115722.
Park, S., Lundeen, E., & Blanck, H. 2016. Knowledge of Health Conditions Related to Drinking Sugar-Sweetened Beverage and Sugar-Sweetened Beverage Intake Among US Adults. Journal of Nutrition Education and Behavior, 48(7), S98.
Perramon, J., Oliveras-Villanueva, M., & Llach, J. 2022. Impact of service quality and environmental practices on hotel companies: An empirical approach. International Journal of Hospitality Management, 107, 103307.

Prihayati, Y., & Veriasa, T. O. 2021. Developing green tourism to create the sustainable landscape: evidence from Community-based Coffee Tourism (CbCT) in Puncak, Bogor, Indonesia. IOP Conference Series: Earth and Environmental Science,

Rita, P., Moro, S., & Cavalcanti, G. 2022. The impact of COVID-19 on tourism: Analysis of online reviews in the airlines sector. Journal of Air Transport Management, 104, 102277.

Shaheen, M., Zeba, F., Chatterjee, N., & Krishnankutty, R. 2019. Engaging customers through credible and useful reviews: the role of online trust. Young Consumers.

Sim, Y., Lee, S. K., & Sutherland, I. 2021. The impact of latent topic valence of online reviews on purchase intention for the accommodation industry. Tourism Management Perspectives, 40, 100903.

Streimikiene, D., Svagzdiene, B., Jasinskas, E., & Simanavicius, A. 2021. Sustainable tourism development and competitiveness: The systematic literature review. Sustainable development, 29(1), 259-271.

Tian, Y., & Zhang, Y. 2022. Pricing of crowdfunding products with strategic consumers and online reviews. Electronic Commerce Research and Applications, 54, 101169.

Tuna, H., & Başdal, M. 2021. Curriculum evaluation of tourism undergraduate programs in Turkey: A CIPP model-based framework. Journal of Hospitality, Leisure, Sport & Tourism Education, 29, 100324.

Türker, N., & Süzer, Ö. 2022. Tourists' food and beverage consumption trends in the context of culinary movements: The case of Safranbolu. International Journal of Gastronomy and Food Science, 27, 100463.

UNEP. 2013. World's Largest Travel Site Awards Qualifying Accommodations Across the U.S. with Bronze, Silver, Gold or Platinum Status. Retrieved October from https://www.unep.org/es/node/6002

Verma, V. K., Chandra, B., & Kumar, S. 2019. Values and ascribed responsibility to predict consumers' attitude and concern towards green hotel visit intention. Journal of Business Research, 96, 206-216.

Wadud, M. A. H., Kabir, M. M., Mridha, M., Ali, M. A., Hamid, M. A., & Monowar, M. M. 2022. How can we manage offensive text in social media-a text classification approach using LSTM-BOOST. International Journal of Information Management Data Insights, 2(2), 100095.

Wang, Q., Zhang, W., Li, J., Mai, F., & Ma, Z. 2022. Effect of online review sentiment on product sales: The moderating role of review credibility perception. Computers in Human Behavior, 133, 107272.

Wei, X., & Taecharungroj, V. 2022. How to improve learning experience in MOOCs an analysis of online reviews of business courses on Coursera. The International Journal of Management Education, 20(3), 100675.

Williams, T., & Betak, J. 2018. A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. Procedia Computer Science, 130, 98-102.

Wu, H., Zhang, Z., Li, X., Shang, K., Han, Y., Geng, Z., & Pan, T. 2022. A novel pedal musculoskeletal response based on differential spatio-temporal LSTM for human activity recognition. Knowledge-Based Systems, 110187.

Wu, L., & Noels, L. 2022. Recurrent Neural Networks (RNNs) with dimensionality reduction and break down in computational mechanics; application to multi-scale localization step. Computer Methods in Applied Mechanics and Engineering, 390, 114476.

Xianghua, F., Guo, L., Yanyan, G., & Zhiqiang, W. 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowledge-Based Systems, 37, 186-195.

Yeşiltaş, M., Gürlek, M., & Kenar, G. 2022. Organizational green culture and green employee behavior: Differences between green and non-green hotels. Journal of Cleaner Production, 343, 131051.

Yu, M., Cheng, M., Yang, L., & Yu, Z. 2022. Hotel guest satisfaction during COVID-19 outbreak: The moderating role of crisis response strategy. Tourism Management, 93, 104618.

Zamparini, L., Domènech, A., Miravet, D., & Gutiérrez, A. 2022. Green mobility at home, green mobility at tourism destinations: A cross-country study of transport modal choices of educated young adults. Journal of Transport Geography, 103, 103412.

Zhang, C., Peng, K., Dong, J., & Miao, L. 2022. A comprehensive operating performance assessment framework based on distributed Siamese gated recurrent unit for hot strip mill process. Applied Soft Computing, 109889.

Zhang, E., Li, H., Huang, Y., Hong, S., Zhao, L., & Ji, C. 2022. Practical multi-party private collaborative k-means clustering. Neurocomputing, 467, 256-265.

Zhang, N., Liu, R., Zhang, X.-Y., & Pang, Z.-L. 2021. The impact of consumer perceived value on repeat purchase intention based on online reviews: by the method of text mining. Data Science and Management, 3, 22-32.

Zhao, R., Wang, D., Yan, R., Mao, K., Shen, F., & Wang, J. 2017. Machine health monitoring using local feature-based gated recurrent unit networks. IEEE Transactions on Industrial Electronics, 65(2), 1539-1548.

Zibarzani, M., Abumalloh, R. A., Nilashi, M., Samad, S., Alghamdi, O., Nayer, F. K., Ismail, M. Y., Mohd, S., & Akib, N. A. M. 2022. Customer satisfaction with Restaurants Service Quality during COVID-19 outbreak: A two-stage methodology. Technology in Society, 70, 101977.

Research Article
Journal of Intelligent Systems: Theory and Applications 6(2) (2023) 143-151
DOI: 10.38016/jista.1079554

Equilibrium Optimizer Algorithm for Optimal Reactive Power Dispatch

Erdi Doğan[1*]

[1] Turkish Electricity Transmission Company, Sakarya, Türkiye

erdi.dogan.teias@gmail.com

Abstract
Optimal Reactive Power Dispatch (ORPD) is a significant research area in terms of maintaining the reliability and safety of the power system and operating it more economically. ORPD problem can be formed from a variety of perspectives including the minimization of the active power losses and voltage deviation, and improving the voltage stability performance. The majority of methods so as to deal with ORPD problem is meta-heuristic techniques because of the complex, non-linear and non-convex nature of the problem. In this paper, a new physic-based meta-heuristic algorithm, Equilibrium Optimizer (EO), is proposed for ORPD problem to reach the optimal settings of control variables such as voltage magnitudes in PV buses, tap positions of transformers and reactive power support of shunt devices. The introduced algorithm is evaluated on IEEE 30-bus test system by using various objectives, and a comparison of the implemented method to other optimization techniques described in the literature is utilized to assess its efficacy. Simulation results and statistical indicators demonstrate that the EO algorithm validates its computational efficacy and robustness in handling the ORPD problem.

Keywords: Equilibrium optimizer, meta-heuristics, optimal reactive power dispatch, optimization algorithms.

Optimal Reaktif Güç Dağıtımı için Equilibrium Optimizasyon Algoritması

Öz
Optimal Reaktif Güç Dağıtımı (ORPD), şebekenin güvenilirliğini ve güvenliğini sağlamak ve güç sistemini daha ekonomik bir şekilde işletmek açısından önemli bir araştırma alanıdır. ORPD problemi, aktif güç kayıplarının ve gerilim sapmasının en aza indirilmesi ve gerilim kararlılık performansının iyileştirilmesi dahil olmak üzere çeşitli açılardan oluşturulabilir. ORPD problemiyle başa çıkmak için kullanılan yöntemlerin çoğu, problemin karmaşık, doğrusal olmayan ve dışbükey olmayan doğası nedeniyle meta-sezgisel tekniklerdir. Bu çalışmada, ORPD probleminin PV baralardaki gerilim büyüklükleri, transformatörlerin kademe pozisyonları ve şönt ekiomanların reaktif güç desteği gibi kontrol değişkenlerinin optimal ayarlarına ulaşması için fizik-tabanlı yeni bir meta-sezgisel algoritma olan Equilibrium Optimizer (EO) önerilmiştir. Tanıtılan algoritma, çeşitli hedefler kullanılarak IEEE 30-baralı test sistemi üzerinde değerlendirilmiştir ve etkinliğini tespit edebilmek için uygulanan yöntemin literatürde açıklanan diğer optimizasyon teknikleri ile karşılaştırılması yapılmıştır. Simülasyon sonuçları ve istatistiksel göstergeler, EO algoritmasının ORPD problemini çözme açısından etkinliğini ve sağlamlığını doğrulamaktadır.

Anahtar Kelimeler: Equilibrium optimizasyon algoritması, meta-sezgisel, optimal reaktif güç dağıtımı, optimizasyon algoritmaları.

1. Introduction

The Optimal Reactive Power Dispatch (ORPD) can be seen as a subproblem of Optimal Power Flow (OPF) (Biswas et al., 2019; Elsayed & Elattar, 2021). Although the reactive power only circulates in the power system, it is indispensable for voltage stability and power transfer (Saddique et al., 2020). Reactive power control and management are required in the power system to keep voltages on all busbars within acceptable limits and reduce the active power losses. Reactive power flow should not be disregarded since it's used by inductive loads and some types of equipment in the power system. Hence, reactive power generation in a power system should be adequate to satisfy the related components without causing additional power loss and undesired voltage drop.

The objective of ORPD can be minimizing the active power loss based on the premise that reactive

* Corresponding Author
E-mail: erdi.dogan.teias@gmail.com

Recieved : 26 Feb 2022
Accepted : 22 June 2023

power flow increases the active power losses and voltage deviations of load buses in the network, and enhancing the voltage stability. Control variables such as generator bus voltages, transformer tap positions and reactive power support of the shunt compensators or reactors are modified in order to achieve the desired objective. (Li et al., 2013) has concentrated to minimize the active power losses while (Gangotri & Bhimwal, 2010) and (Robbins & Domínguez-García, 2016) have focused on improving the system security through the voltage stability index and voltage deviation equations, respectively. (Nguyen & Vo, 2020) has tackled the ORPD problem in different perspectives such as the minimization of active power loss, voltage deviation and voltage stability index. It is worth mentioning that constraints related to the power system such as power balance, the reactive power capability of generators and shunt compensators or reactors, limits of bus voltages and transmission lines should be maintained during the optimization process.

The ORPD is modeled as a nonlinear programming problem, and some conventional techniques such as Interior Point Method (Granville, 1994) and Quadratic Programming (Grudinin, 1998) have been utilized so as to solve this challenging problem. However, the majority of ORPD approaches are meta-heuristics due to the non-linear character of the problem(Saddique et al., 2020). There are many investigations implemented to solve the ORPD problem by using meta-heuristic algorithms such as Gravitational Search Algorithm (Duman et al., 2012), Grey Wolf Optimizer (Sulaiman et al., 2015), Particle Swarm Optimization (Singh et al., 2015), Coyote Optimization Algorithm (Güvenç et al., 2020), Barnacles Mating Optimizer (Sulaiman et al., 2020) and some hybrid techniques (Nasouri Gilvaei et al., 2020; Shaheen et al., 2021). The reason of this valuable attention has been given to the meta-heuristics is that they have capable of effectively solving a variety of large-scale complex problems. However, they could diverge to local optima and do not guarantee to figure out the best solution. Another drawback of the meta-heuristics is the long solution time at computationally expensive problems. Therefore, researchers maintain to investigate the most suitable meta-heuristic algorithm in terms of solving capability and robustness to deal with the ORPD problem.

This paper focuses on determining the appropriate control parameters for reducing the active power losses of the IEEE 30-bus test system and explains how to implement the novel Equilibrium Optimization strategy in order to improve voltage profiles. Comparative analyses have been conducted with well-known meta-heuristic techniques in the Literature in order to demonstrate the effectiveness of the proposed EO algorithm in solving the ORPD problem.

The rest of the paper is organized as follows. First of all, the objection functions and constraints to be used in ORPD are explained in Section 2. In Section 3, the proposed Equilibrium Optimizer algorithm is introduced to deal with the ORPD problem. Section 4 presents the results and statistical indicators for case studies. Finally, the conclusion is reported in Section 5.

2. Problem Formulation

The ORPD problem purposes to minimize the investigated objective function while meeting operational equality and inequality constraints, obtaining the best solution for independent control variables. The ORPD shows a non-linear and non-convex behaviour, and it's an NP-hard problem, which means it's tough to solve using mathematical methods. The general frame of the ORPD (Biswas et al., 2019), including equality and inequality constraints, can be written as follow:

$$\text{Minimize: } f(x,u) \tag{1}$$
$$\text{Subject to: } g(x,u) \leq 0 \text{ and } h(x,u) = 0 \tag{2}$$

where x and u represent control and state variables of the problem respectively. f symbolizes the objective function, g and h stand for inequality and equality constraints.

Independent control variables of the ORPD problem consist of voltage magnitude of the PV bus, tap position of transformers and reactive power injected into the network by the shunt devices, which all of them creates the search space of the problem. Although transformers and shunt devices tap positions need integer variables, they are considered as decimal in this study in order to achieve the optimal point more effectively. The set of independent control and dependent state variables can be created as follows:

$$x^T = \left[V_{g1},..,V_{gN_g},T_1,..,T_{N_t},Q_{c1},..,Q_{cN_c}\right] \tag{3}$$
$$u^T = \left[Q_{g1},..,Q_{gN_g},V_{l1},..,V_{lN_l},S_1,..,S_{NT_l}\right] \tag{4}$$

2.1. ORPD Functions

The objective function of the ORPD can be modelled with three different perspectives, which are the minimization of the active power losses, voltage deviations and voltage stability index. Furthermore, these objectives can be handled together by using the multi-objective optimization concept.

2.1.1. Active Power Loss

The ORPD problem's initial objective is to reduce the total active power loss in the network, which may be expressed as (Nasouri Gilvaei et al., 2020):

$$\text{Minimize: } f_1 = P_{loss}(x,u) \rightarrow \sum_{i=1}^{N_B}\sum_{j=1}^{N_B} G_{ij}\left[\left(V_i^2 + V_j^2 - 2V_iV_j\cos\delta_{ij}\right)\right], i \neq j \tag{5}$$

where N_B represents the number of buses, V_i and V_j are the voltage magnitude of the bus i and j respectively, δ_{ij} and G_{ij} symbolize the difference of voltage angles and the conductance of the transmission line between bus i and j respectively.

2.1.2. Total Voltage Deviation

The second target to be optimized in ORPD problem can be the minimization of the total voltage deviation like given in (Abaza et al., 2021):

$$\text{Minimize: } f_2 = TVD(x,u) \rightarrow = \sum_{i=1}^{N_l} |V_i - V_{ref}| \quad (6)$$

where N_l symbolizes the number of PQ or load buses, V_i is the voltage magnitude of the i[th] PQ bus and V_{ref} represents the reference voltage magnitude considered as 1.0 pu.

2.1.3. Voltage Stability Index

Another option that can be utilized in the ORPD problem as an objective is the improvement voltage stability of the power system. The capacity of a power system to keep the voltage within its limit at each bus in the network under normal operating circumstances is referred to as voltage stability. When a system is subjected to a disturbance, such as a surge in load demand or a change in the system configuration, it might experience voltage instability, which can result in a gradual and unpredictable drop in voltage. As a result, improving a system's voltage stability is a crucial aspect of power system management and planning (Ettappan et al., 2020).

The improvement of the voltage stability can be accomplished by minimizing the voltage stability criteria known as the L-index, a scalar value having a range of [0,1], at each PQ bus (Nasouri Gilvaei et al., 2020). A maximum value of the L-index results near 0 indicates that the system is almost stable, while a value close to 1 indicates that the system is on the verge of reaching voltage collapse (Rajan & Malakar, 2016). The L-index of the j[th] PQ bus can be calculated as follows (Kessel & Glavitsch, 1986):

$$\text{Minimize: } f_3 = VSI(x,u) = L_{max} \rightarrow \quad (7)$$

$$= L_{max} = \max(L_j), \forall j \in N_l \quad (8)$$

$$L_j = \left|1 - \sum_{i=1}^{N_g} F_{ij} \frac{V_i}{V_j}\right|, \forall j \in N_l \quad (9)$$

$$F_{ij} = -[Y_1]^{-1}[Y_2] \quad (10)$$

$$\begin{bmatrix} I_{PQ} \\ I_{PV} \end{bmatrix} = \begin{bmatrix} Y_1 & Y_2 \\ Y_3 & Y_4 \end{bmatrix} \begin{bmatrix} V_{PQ} \\ V_{PV} \end{bmatrix} \quad (11)$$

2.2. Equality Constraints

Equality constraints in ORPD are commonly represented by power balance equations for both active and reactive power, which ensure that the load demand is satisfied by taking into account the power losses, and are depicted as follows:

$$P_{gi} - P_{li} = |V_i| \sum_{j=1}^{N_B} |V_j| (G_{ij} \cos \delta_{ij} + B_{ij} \sin \delta_{ij}), \forall i \in N_B \quad (12)$$

$$Q_{gi} - Q_{li} = |V_i| \sum_{j=1}^{N_B} |V_j| (G_{ij} \cos \theta_{ij} - B_{ij} \sin \theta_{ij}), \forall i \in N_B \quad (13)$$

where N_B symbolizes the total number of buses in the power system, P_{gi}, P_{li}, Q_{gi}, Q_{li} and B_{ij} represent active and reactive power generation and demand in bus i and line susceptance between i[th] and j[th] buses, respectively. Except for the slack bus, whose output is a dependent variable since it is affected by power losses, all active power generation of PV buses is fixed. The Q_i is also a state variable because the reactive power injected varies when the control variables are changed.

2.3. Inequality Constraints

Control and state variables are the two forms of inequality constraints used in ORPD. The transformer output, generator bus voltages, and the reactive power provided by the shunt capacitors are all control variables, while active power generation at the slack bus, reactive power generation at the PV bus, voltages of the PQ bus, and power flow of transmission lines are among the state variables. The inequality constraints on control variables can be written as follows:

$$V_{gi}^{min} \leq V_{gi} \leq V_{gi}^{max}, \forall i \in N_g \quad (14)$$

$$T_i^{min} \leq T_i \leq T_i^{max}, \forall i \in N_t \quad (15)$$

$$Q_{ci}^{min} \leq Q_{ci} \leq Q_{ci}^{max}, \forall i \in N_c \quad (16)$$

The inequality constraints on state variables can be created as follows:

$$Q_{gi}^{min} \leq Q_{gi} \leq Q_{gi}^{max}, \forall i \in N_g \quad (17)$$

$$V_{li}^{min} \leq |V_{li}| \leq V_{li}^{max}, \forall i \in N_l \quad (18)$$

$$S_i \leq S_i^{max}, \forall i \in NT_l \quad (19)$$

$$P_{gs}^{min} \leq P_{gs} \leq P_{gs}^{max}, s = slack \quad (20)$$

Voltages in PQ buses and the loading levels of the transmission lines can be considered as security constraints, while reactive power generations of units are related to the operational limitations.

2.4. Constraints Handling

A suitable solution to the ORPD problem can only be achieved by complying with the relevant constraints. The inequality constraints on independent control variables have already been satisfied through determining upper and lower limits that meta-heuristic algorithm can allocate. However, there is a need to be concerned with inequality constraints on dependent variables. Reactive power limits of generators and active power limit of the slack bus can be determined while using the Newton Raphson power flow equation so that if control variables violate these constraints, power flows will take place according to allowable limits, which means that the exact value of the control variable will not be satisfied. On the other hand, the voltage limit of PQ buses and transmission line thermal limits are constraints that need to be addressed in the solution process.

Various papers such as (Rajan & Malakar, 2016) and (Ettappan et al., 2020) have resolved the compliance problem to the constraints by using the punishment and aggregating method so that any violation in constraints reduces the solution quality of the objective function. Therefore, the objective function of ORPD can be reconstructed through binding related constraints to the function as a penalty.

$$\text{Minimize: } P = f_{obj} + \omega_v + \omega_s \tag{21}$$

where,

$$\omega_v = \lambda_v \sum_{i=1}^{N_l} \{\max(0, abs(V_i - V_i^{lim}))\}^2 \tag{22}$$

$$\omega_s = \lambda_s \sum_{i=1}^{NT_l} \{\max(0, S_i - S_i^{max})\}^2 \tag{23}$$

$$V_i^{lim} = \begin{cases} V_i^{max}, if\ V_i > V_i^{max} \\ V_i^{min}, if\ V_i < V_i^{min} \end{cases} \tag{24}$$

where ω_v and ω_s are punishments relevant to voltage violation in PQ buses and overloading in transmission lines while λ_v and λ_s represent constant penalty coefficients, f_{obj} symbolizes the main objective, which can stand for one of the objectives, including power loss, voltage deviation and voltage stability index. It is worth mentioning that f_{obj} can be also designed as a multi-objective framework by using either the aggregating method or pareto-optimality technique.

3. Equilibrium Optimizer

The Equilibrium Optimizer has been constructed based on control volume mass balance models used to estimate both dynamic and equilibrium states (Faramarzi et al., 2020). Position (concentration) of each particle represents search agents in EO. To finally reach the equilibrium state, which means an optimal solution, the search agents update their positions at random with regard to the best-so-far solutions, termed equilibrium candidates. EO has been created in order to deal with single-objective optimization problems and position updating rule implemented can be written as follow:

$$\vec{C} = \vec{C}_{eq} + (\vec{C} - \vec{C}_{eq})\vec{F} + \frac{\vec{G}}{\vec{\lambda}}(1 - \vec{F}) \tag{25}$$

where, $\vec{C}$ is the new position vector of each particle, $\vec{C}_{eq}$ represents equilibrium point retrieved from a pool comprising some best solutions, $\vec{F}$ is an exponential term and $\vec{G}$ is generation rate. The second term of the equation allowing to investigate a wider range in search space is related to a difference in concentration between a particle and the equilibrium state. Particles act as explorers by searching the entire search space in this way. On the other hand, the third term associated with the generation rate ensures the exploitation of obtained search areas with short steps, though it can also serve as an explorer. The extended formulations of these terms are given as follows:

$$\vec{F} = a_1 sign(\vec{r} - 0.5)[e^{-\vec{\lambda}t} - 1] \tag{26}$$

$$t = (1 - \frac{Iter}{MaxIter})^{(a_2 \frac{Iter}{MaxIter})} \tag{27}$$

$$\vec{C}_{eq} = random.choice(\vec{C}_{eq,pool}) \tag{28}$$

$$\vec{C}_{eq,pool} = (\vec{C}_{eq(1)}, .., \vec{C}_{eq(4)}, \vec{C}_{eq(avg)}) \tag{29}$$

$$\vec{C}_{eq(avg)} = \frac{\vec{C}_{eq(1)} + \vec{C}_{eq(2)} + \vec{C}_{eq(3)} + \vec{C}_{eq(4)}}{4} \tag{30}$$

$$\vec{G} = \vec{G}_0 \vec{F} \tag{31}$$

$$\vec{G}_0 = \overrightarrow{GCP}(\vec{C}_{eq} - \vec{\lambda}\vec{C}) \tag{32}$$

$$\overrightarrow{GCP} = \begin{cases} 0.5r_1, & r_2 \geq GP \\ 0, & r_2 < GP \end{cases} \tag{33}$$

where a_1 is a constant coefficient controlling step size in exploration phase, a_2 is a constant coefficient that regulate exploitation phase, $\vec{r}$ and $\vec{\lambda}$ are uniform distributed random vector between 0 and 1, r_1 and r_2 are uniform distributed random number between 0 and 1, $\overrightarrow{GCP}$ is generation rate control parameter and GP is generation probability.

The equilibrium pool $\vec{C}_{eq,pool}$ comprise best four particles obtained until related iteration and average value of these particles. These four particles improve the exploration capabilities of the algorithm, whereas the average particle strengthens the exploitation ability. The main tool in order to exploit the promising region is the generation rate $\vec{G}$. The higher the GP generation probability, the lesser particle takes advantage of the generation rate since $\overrightarrow{GCP}$ that is generation rate control parameter becomes zero. Another meaningful

expression is $sign(\vec{r} - 0.5)$ determining the motion direction of each particle.

A proper balance between exploration and exploitation phases should be constructed in all meta-heuristic approaches so as to acquire quality solution. In EO, exploration phase is conducted by generation probability and a_1 constant while exploitation stage is performed through memory saving (like p_{best} of particle swarm optimization) and a_2 constant. Equilibrium pool and $sign(\vec{r} - 0.5)$ term are also crucial terms for establishing balance between phases.

During the first iterations, the individuals are all spatially isolated from one another. The algorithm's capacity to explore the space broadly is confirmed by updating the concentrations depending on these candidates of equilibrium pool. At first iterations, when particles are far apart, the average particle of equilibrium pool also assists in the discovery of unknown search areas. The concentration update mechanism will help in local search around the candidates since individuals of equilibrium pool are close to each other in the last iterations. Therefore, the equilibrium pool manages either exploration or exploitation phases according to iteration level.

4. Result and Discussion

The IEEE- 30 bus power system has been utilized as a test system to validate the efficacy and robustness of the proposed EO Algorithm based Reactive Power Dispatch. The EO is executed in the Python programming language with PSS/E 35.2 software package, and numerical tests are performed on a computer with an Intel® Core™ i7-8850U CPU at 2.60GHz with 16GB of RAM. To solve the ORPD problem with EO, the Mealpy software package (Thieu & Molina, 2021), a set of state-of-the-art Meta-heuristic algorithms in Python, is used.

4.1. IEEE 30 bus system

There are six generator units located at buses 1, 2, 5, 8, 11, and 13 – bus 1 is chosen as slack bus, twenty-four load buses with 2.834 pu and 1.262 pu for both active and reactive power demand, four regulating tap-changing transformers at branches 4-12, 6-9, 6-10 and 28-27, and nine shunt VAR capacitors at the buses 10, 12, 15, 17, 20, 21, 23, 24, and 29 in the IEEE 30-bus system. The limit of voltage magnitude is considered between 0.95 and 1.1 pu for generator buses and 0.90 and 1.1 pu for load buses. The maximum output of the shunt capacitors is determined as 5 MVar while the transformer tap settings have been configured to vary between 0.9 and 1.1 pu. The test system data is available in (*Pg_tca30bus*, n.d.).

4.2. Experimental Case Studies

On the IEEE 30-bus test system, the EO approach is performed to minimize the penalty function, including the active power loss, total voltage deviation or voltage stability index as a single objective function with the penalty terms related to mentioned constraints. Six different cases have been evaluated in the test system in order to compare the effectiveness of EO according to the other meta-heuristics in the Literature. Table-1 shows the generation amounts adjusted to implement the first and second three cases. Every three cases consist of the objectives of the minimization of the active power loss and voltage deviation, and the improvement of the voltage stability. It is worth mentioning that λ_v and λ_s penalty coefficients (equation 22 and 23) are 500 and 700 as in (Rajan & Malakar, 2016). It should also be emphasized that each objective function is subjected to 30 trial runs with determining the population size and iteration as 50 and 1000, respectively. The results have been compared to those obtained using numerous meta-heuristics, including GSA (Duman et al., 2012), COA (Güvenç et al., 2020), ABC (Ettappan et al., 2020) and SMA (Elsayed & Elattar, 2021), implemented for addressing the same ORPD problem to demonstrate the advantage of EO.

Table 1. Generator data for IEEE 30-bus test system

Bus No	P_g (MW)		$Q_{g,min}$ (MVar)	$Q_{g,max}$ (MVar)
	Case 1,2,3	Case 4,5,6		
1	Slack	Slack	-20	150
2	75	80	-20	60
5	40	50	-15	62.5
8	30	20	-15	48.7
11	25	20	-15	40
13	30	20	-15	46.5

Table 2 presents the optimal values for all control variable ranges in case studies in order to minimize the relevant objective functions. It can be recognized in Table 2 that the EO is capable of reducing the power loss to 4.108 MW in case-1 and 4.54 MW in case-4. The percentages of reduction in power loss are 23.61% in case -1 and 21.99% in case-4 according to the base case values. There are a total of 24 load buses in the system under investigation and the highest conceivable cumulative total of TVD would theoretically be 2.4 pu (i.e. 24x((1.1-1.0) or (1.0-0.9))) if all of these buses run at their limits. Therefore, the total voltage deviation value for the 30-bus system should never exceed 2.4 pu in order to keep the load bus voltages between 0.9 and 1.1 pu and the achieved TVD values because of the minimizing the power losses are 1.87, which means that there will be no violation for both case-1 and case-4. On the other hand, if we turn the objective function into the total voltage deviation perspective, the optimal values of TVD in case-2 and case-5 are 0.14 and 0.115 pu, resulting in higher active power losses according to

the base cases. However, if the voltage stability index is chosen as the objective, the L indexes in case-3 and case-6 are obtained as 0.09762 and 0.09758, which do not significantly increase the active power losses.

from EO have been achieved in the voltage stability index objective. The L-index results of EO for cases 3 and 6 are superior in comparison with other meta-heuristics. Nonetheless, the best results for TVD are

Table 2. The results of EO at different cases

Control Variables	Base Case (1-2-3)	Base Case (4-5-6)	Case-1 (P_{loss})	Case-2 (TVD)	Case-3 (L_{index})	Case-4 (P_{loss})	Case-5 (TVD)	Case-6 (L_{index})
V_{g1}	1.05	1.05	1.1	0.9835	1.0953	1.1	0.9814	1.0997
V_{g2}	1.04	1.04	1.0945	1.0746	1.1	1.0942	0.9607	1.1
V_{g3}	1.01	1.01	1.0733	1.0129	1.0993	1.074	1.0566	1.1
V_{g4}	1.01	1.01	1.0809	1.0862	1.0899	1.0764	1.0135	1.0939
V_{g5}	1.05	1.05	1.1	1.0735	0.9625	1.1	1.0530	0.9642
V_{g6}	1.05	1.05	1.1	1.0407	0.9504	1.1	1.0598	0.9970
T_{4-12}	1.032	1.032	0.9856	0.9948	0.9012	0.98	1.0961	0.9
T_{6-9}	1.078	1.078	1.0433	1.0429	0.9260	1.0489	1.0680	0.9222
T_{6-10}	1.069	1.069	0.9034	0.9	0.9	0.9	0.9	0.9
T_{27-28}	1.068	1.068	0.9638	0.9461	0.9287	0.9748	0.9478	0.9261
Q_{10}	0	0	0	0	3.86	2.97	5.0	0
Q_{12}	0	0	0	0	0	0.06	2.9	0
Q_{11}	0	0	4.98	0	0	3.26	3.28	0
Q_{17}	0	0	4.96	0	0	4.87	3.17	3.6567
Q_{20}	0	0	4.56	4.5	0	4.99	5.0	0
Q_{21}	0	0	4.86	0	0	4.93	3.70	0
Q_{23}	0	0	0	0.7	0	4.94	4.35	0
Q_{24}	0	0	5	4.3	0	2.97	4.78	0
Q_{29}	0	0	0	0	0	1.1	6.98	0
$\boldsymbol{P_{loss}}$	5.38	5.82	**4.108**	6.25	5.19	**4.54**	6.9464	5.58
$\boldsymbol{TVD}$	1.19	1.19	1.87	**0.14**	1.89	1.87	**0.115**	1.87
$\boldsymbol{L_{index}}$	0.24897	0.24548	0.105	0.12	**0.09762**	0.17	0.1241	**0.09758**
% $\boldsymbol{Reduction}$	-	-	23.61	88.24	60.79	21.99	90.34	60.24

Table 3. The comparison of EO with other meta-heuristics

Algorithms	Case-1 (P_{loss})	Case-2 (TVD)	Case-3 (L_{index})	Case-4 (P_{loss})	Case-5 (TVD)	Case-6 (L_{index})
EO	**4.108**	0.1349	**0.09762**	4.54	0.115	**0.09758**
SHADE-EC (Biswas et al., 2019)	4.4126	0.08886	-	4.8612	0.08724	-
COA (Güvenç et al., 2020)	4.41238	0.08837	-	4.861183	0.08724	-
GWO-PSO (Shaheen et al., 2021)	-	-	-	5.09037	0.27800	-
EMA (Rajan & Malakar, 2016)	-	-	-	**4.4978**	**0.061311**	0.09797
SMA (Elsayed & Elattar, 2021)	-	-	-	4.5181	-	-
SCA (Saddique et al., 2020)	-	-	-	4.7086	-	-
ABC (Ettappan et al., 2020)	-	-	-	4.5804	-	-
ECOA (Abaza et al., 2021)	-	-	-	4.547	-	-
GLS (Kanagasabai, 2020)	4.216	**0.064**	0.1160	-	-	-
FA-APTFPSO (Nasouri Gilvaei et al., 2020)	-	-	-	4.8664	0.0841	0.1186
PSOGSA (DUMAN, 2018)	-	-	-	4.5950	0.1234	0.1242
BMO (Sulaiman et al., 2020)	-	-	-	4.5862	-	-
GSA (Duman et al., 2012)	-	-	-	4.51431	0.067633	0.11607

Table 3 compares the solutions of EO with the results obtained from different methods in the IEEE 30-bus system. As can be seen in Table 3 that the active power loss value obtained from EO in case-1 is the best one among the published results. It should be stated at this point that the range of the voltage limit for the load buses is determined within 0.9 pu and 1.1 pu in this study. The other effective solutions acquired

provided by Green Lourie Swarm Optimization (Kanagasabai, 2020) for case 2 and Exchange Market Algorithm (Rajan & Malakar, 2016) for case 5. It should be noted that comparing the performance of two methods solely on the basis of numerical values of outcomes may be inappropriate for a constrained optimization problem due to using coefficients.

When it comes to meta-heuristics, not only their efficacy but also their robustness is critical. Therefore, a more extensive and in-depth examination is required. In this direction, Table 4 compares the min, max, mean and standard deviation outcomes received with EO to those obtained from other recently published approaches. Table 4 illustrates that the EO algorithm testifies its robustness with low statistical indicators, including standard deviation.

Table 4. The comparison the statistical indicators of EO with other published solutions

Algorithms	Indicator	Case-1 (P_{loss})	Case-2 (TVD)	Case-3 (L_{index})	Case-4 (P_{loss})	Case-5 (TVD)	Case-6 (L_{index})
EO	Min	**4.108**	**0.1349**	**0.0976**	4.543	0.115	**0.0975**
	Max	**4.198**	**0.1788**	**0.0991**	4.653	0.175	**0.0987**
	Mean	**4.156**	**0.1572**	**0.0982**	4.588	0.154	**0.0981**
	Std	**0.0223**	**0.0134**	**0.00038**	0.02475	0.013	**0.00027**
SMA (Elsayed & Elattar, 2021)	Min	-	-	-	4.5181	-	-
	Max	-	-	-	4.7814	-	-
	Mean	-	-	-	4.63	-	-
	Std	-	-	-	0.0979	-	-
FA-APTFPSO (Nasouri Gilvaei et al., 2020)	Min	-	-	-	4.8664	0.0841	0.1186
	Max	-	-	-	4.8853	0.0984	0.1198
	Mean	-	-	-	4.8689	0.0894	0.1191
	Std	-	-	-	0.00504	0.00427	0.00039
SCA (Saddique et al., 2020)	Min	-	-	-	4.708	-	-
	Max	-	-	-	5.286	-	-
	Mean	-	-	-	5.030	-	-
	Std	-	-	-	0.133	-	-
EMA (Rajan & Malakar, 2016)	Min	-	-	-	**4.4978**	**0.061311**	0.09797
	Max	-	-	-	**4.50**	**0.0725**	0.1011
	Mean	-	-	-	**4.4999**	**0.06558**	0.098744
	Std	-	-	-	**0.0003716**	**0.0008328**	0.000458

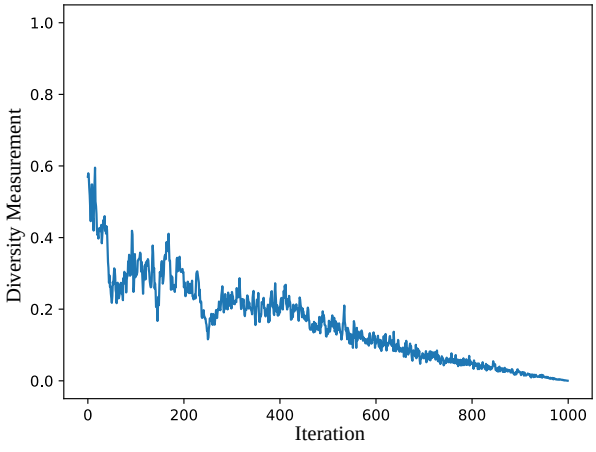


Figure 1. Diversity measurement chart of EO

In order to clarify the efficacy of the EO, performance indicators are presented in Figures 1-6. Figures 1 and 2 illustrate clear information of exploration, exploitation, and particle variety in the population of the EO. It's important to keep in mind that the reader wondering how to be visualized the diversity, exploration and exploitation abilities of a meta-heuristic algorithm should refer to the valuable paper written by (Hussain et al., 2019). it can be said from these figures that the EO algorithm prioritized exploitation above exploration for the majority of the search process. In the first iterations, a balance between exploitation and exploration is constructed (i.e. nearly 50%-50%), however, after a few iterations, the algorithm's behaviour is converted to the exploitative. This can also be seen in Figure 1, where the diversity was initially about 0.5 but steadily decreased over the iterations.

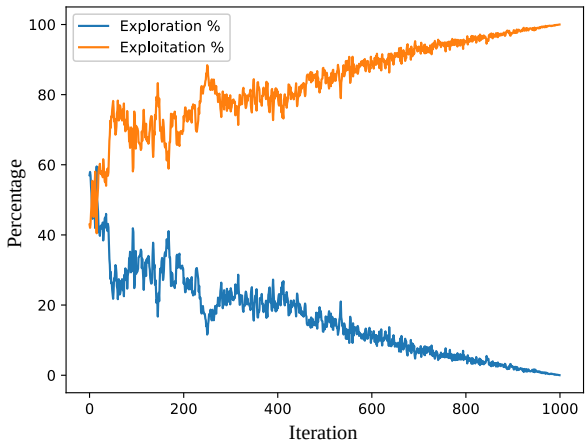


Figure 2. Exploration and Exploitation of EO

The EO convergence curve is shown in Fig. 3, and it has strong convergence properties in terms of power loss optimization. Fig. 4 presents the trajectory of two individuals of the population in two dimensions. It can be observed from this figure that nearly the entire

search space is investigated. Nevertheless, in order to further clarify the behaviour of the individuals during the iterations, the trajectory of the first dimension of the first agent is demonstrated in Fig. 5. It can be concluded from this figure that frequently transitions from upper to lower bound occur in most of the search process, and the alteration in the dimension slows down at the end of the iterations. Ultimately, the runtime chart of the EO algorithm throughout the iterations can be examined in Fig. 6. Although some function evaluations exceed one second, the general solution time is roughly 0.8 seconds per iteration (each with 50 function evaluations). Direct comparisons have not been implemented with other methods based on CPU time due to particular hardware properties and different numbers of function evaluations.

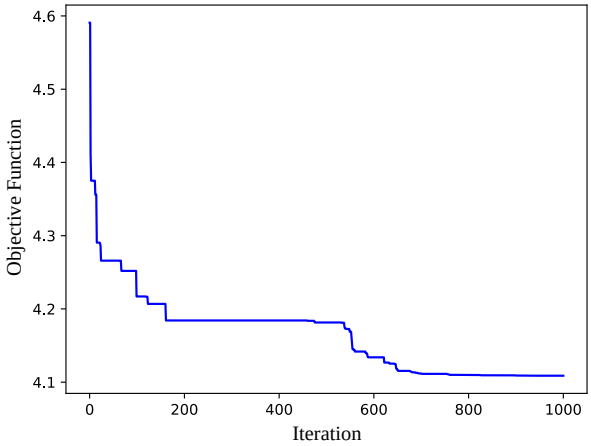


Figure 3. Convergence curve

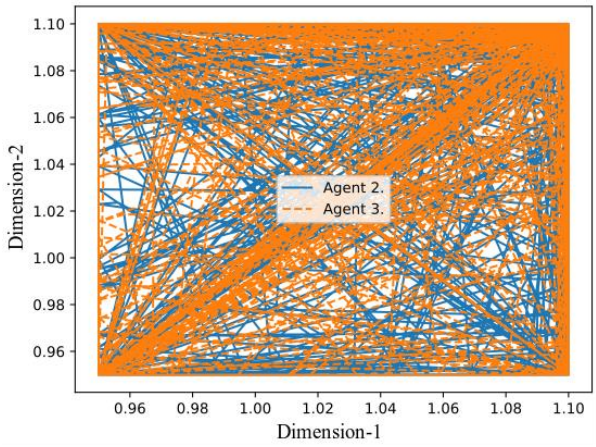


Figure 4. Trajectory of the first and second dimension of the second and third individual after generations in EO

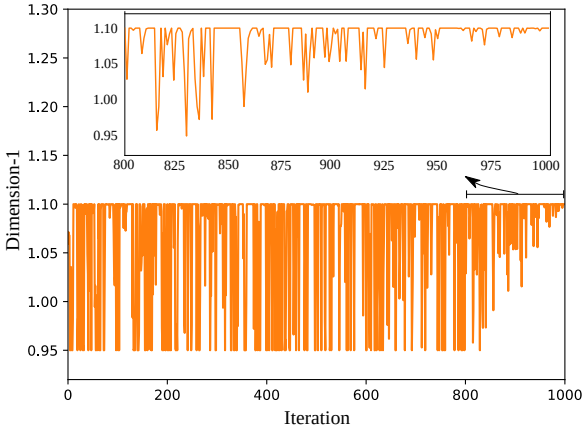


Figure 5. Trajectory of the first dimension of the first individual in EO

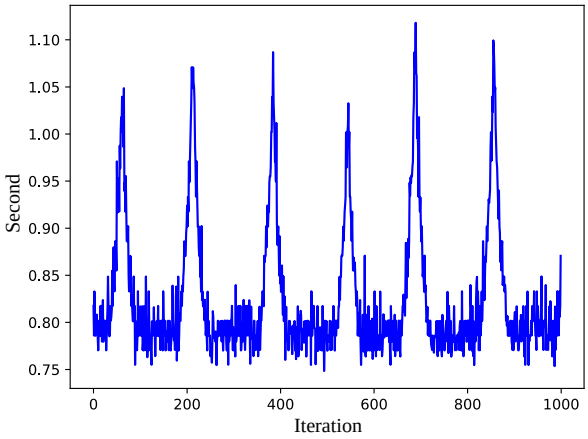


Figure 6. Runtime chart of the EO

5. Conclusions

In this paper, the Equilibrium Optimizer, a recently developed physic-based meta-heuristic algorithm, is performed to overcome the nonlinear, non-convex ORPD problem in the power system. The ORPD problem is implemented on a standard IEEE 30-bus test system in order to validate EO's search ability. Active power loss, voltage deviation and voltage stability index are calculated with optimization of network parameters, including the voltage of the PV buses, tap ratio of transformers and reactive support of shunt capacitors, under six scenarios. For the systems under investigation in case-1 and case-4, using the EO to deal with the ORPD problem resulted in a decrease in power losses of 23.61% and 21.99% with respect to the base cases, respectively. The reduction in total voltage deviation for case-2 and case-5 reached 88.24% and 90.34% while the improvements of the voltage stability for case-3 and case-6 were 60.79% and 60.24% according to the base cases.

This paper also includes comparisons of the EO with other well-known optimization techniques under different perspectives. The EO is superior among others in terms of case-1, case-3 and case-6 with the best solution obtained. Furthermore, the statistical indicators such as mean and standard deviation of

independent 30 runs show that the EO is not only an efficient but also a robust meta-heuristic algorithm in solving the ORPD problem. The measurement of diversity, exploration and exploitation allows for more in-depth analysis of the causes for successful or ineffective outcomes. The analyses conducted demonstrate that the EO algorithm has better exploitation ability as compared to the exploration.

In future, this research can be expanded with the incorporation of active power loss, voltage deviation and voltage stability as an objective in a multi-objective optimization framework based on Pareto-optimality. Moreover, a better trade-off between exploitation and exploration abilities and a more consistent diversity in the population can be constructed with the modification of the EO algorithm so as to solve the ORPD problem.

References

Abaza, A., Fawzy, A., El-Sehiemy, R. A., Alghamdi, A. S., Kamel, S., 2021. Sensitive reactive power dispatch solution accomplished with renewable energy allocation using an enhanced coyote optimization algorithm. Ain Shams Engineering Journal, 12(2), 1723–1739.

Biswas, P. P., Suganthan, P. N., Mallipeddi, R., Amaratunga, G. A. J., 2019. Optimal reactive power dispatch with uncertainties in load demand and renewable energy sources adopting scenario-based approach. Applied Soft Computing Journal, 75, 616–632.

Duman, S., 2018. FACTS Cihazlarını İçeren Reaktif Güç Planlama Probleminin Hibrit PSOGSA Algoritması Kullanarak Çözülmesi. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 6, 1234–1257.

Duman, S., Sönmez, Y., Güvenç, U., Yörükeren, N., 2012. Optimal reactive power dispatch using a gravitational search algorithm. IET Generation, Transmission and Distribution, 6(6), 563–576.

Elsayed, S. K., Elattar, E. E., 2021. Slime mold algorithm for optimal reactive power dispatch combining with renewable energy sources. Sustainability, 13(11).

Ettappan, M., Vimala, V., Ramesh, S., Kesavan, V. T., 2020. Optimal reactive power dispatch for real power loss minimization and voltage stability enhancement using Artificial Bee Colony Algorithm. Microprocessors and Microsystems, 76.

Faramarzi, A., Heidarinejad, M., Stephens, B., Mirjalili, S., 2020. Equilibrium optimizer: A novel optimization algorithm. Knowledge-Based Systems, 191.

Gangotri, K. M., Bhimwal, M. K., 2010. Genetic algorithm based reactive power dispatch for voltage stability improvement. International Journal of Electrical Power & Energy Systems, 32(10), 1151–1156.

Granville, S., 1994. Optimal reactive dispatch through interior point methods. IEEE Transactions on Power Systems, 9(1), 136–146.

Grudinin, N., 1998. Reactive power optimization using successive quadratic programming method. IEEE Transactions on Power Systems, 13(4), 1219–1225.

Güvenç, U., Bingöl, O., Özkaya, B., 2020. Optimal Reaktif Güç Dağıtımı için Kir Kurdu Optimizasyon Algoritmasi. Mühendislik Bilimleri ve Tasarım Dergisi, 8(5), 1–10.

Hussain, K., Salleh, M. N. M., Cheng, S., Shi, Y, 2019. On the exploration and exploitation in popular swarm-based metaheuristic algorithms. Neural Computing and Applications, 31(11), 7665–7683.

Kanagasabai, L., 2020. Solving optimal reactive power problem by Alaskan Moose Hunting, Larus Livens and Green Lourie Swarm Optimization Algorithms. Ain Shams Engineering Journal, 11(4), 1227–1235.

Kessel, P., Glavitsch, H., 1986. Estimating the Voltage Stability of a Power System. IEEE Transactions on Power Delivery, 3, 346–354.

Li, Y., Wang, Y., Li, B., 2013. A hybrid artificial bee colony assisted differential evolution algorithm for optimal reactive power flow. International Journal of Electrical Power & Energy Systems, 52(1), 25–33.

Nasouri Gilvaei, M., Jafari, H., Jabbari Ghadi, M., Li, L., 2020. A novel hybrid optimization approach for reactive power dispatch problem considering voltage stability index. Engineering Applications of Artificial Intelligence, 96.

Nguyen, T. T., Vo, D. N., 2020. Improved social spider optimization algorithm for optimal reactive power dispatch problem with different objectives. Neural Computing and Applications, 32, 5919–5950.

pg_tca30bus. (n.d.). Retrieved February 19, 2022, from http://labs.ece.uw.edu/pstca/pf30/pg_tca30bus.htm

Rajan, A., Malakar, T., 2016. Exchange market algorithm based optimum reactive power dispatch. Applied Soft Computing Journal, 43, 320–336.

Robbins, B. A., Domínguez-García, A. D., 2016. Optimal Reactive Power Dispatch for Voltage Regulation in Unbalanced Distribution Systems. IEEE Transactions on Power Systems, 31(4), 2903–2913.

Saddique, M. S., Bhatti, A. R., Haroon, S. S., Sattar, M. K., Amin, S., Sajjad, I. A., ul Haq, S. S., Awan, A. B., Rasheed, N., 2020. Solution to optimal reactive power dispatch in transmission system using meta-heuristic techniques—Status and technological review. Electric Power Systems Research, 178.

Shaheen, M. A. M., Hasanien, H. M., Alkuhayli, A., 2021. A novel hybrid GWO-PSO optimization technique for optimal reactive power dispatch problem solution. Ain Shams Engineering Journal, 12(1), 621–630.

Singh, R. P., Mukherjee, V., Ghoshal, S. P., 2015. Optimal reactive power dispatch by particle swarm optimization with an aging leader and challengers. Applied Soft Computing, 29, 298–309.

Sulaiman, M. H., Mustaffa, Z., Mohamed, M. R., Aliman, O., 2015. Using the gray wolf optimizer for solving optimal reactive power dispatch problem. Applied Soft Computing, 32, 286–292.

Sulaiman, M. H., Mustaffa, Z., Saari, M. M., Daniyal, H., 2020. Barnacles Mating Optimizer: A new bio-inspired algorithm for solving engineering optimization problems. Engineering Applications of Artificial Intelligence, 87.

Thieu, N. van, & Molina, D. (2021). Meta-Heuristic Algorithms using Python (MEALPY). Zenodo. https://doi.org/10.5281/zenodo.5789724.

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 6(2) (2023) 152-158
DOI: 10.38016/jista.1075054

Akıllı Ev Sistemleri için XGBoost Tabanlı Saldırı Tespit Yöntemi

Orhan Yaman[1*], Rojbin Tekin[2]

[1] Fırat Üniversitesi, Adli Bilişim Mühendisliği Bölümü, Elazığ, Türkiye

[2] Fırat Üniversitesi, Adli Bilişim Mühendisliği Bölümü, Elazığ, Türkiye

orhanyaman@firat.edu.tr, 192144104@firat.edu.tr

Öz

Günümüz akıllı evlerinde IoT (Internet of Things) teknolojisinin alt yapısı kullanılmaktadır. Akıllı evlerin kullanımı arttıkça bu alandaki siber saldırılar da artmaktadır. Akıllı evlere yönelik siber saldırıları mümkün olduğunca erken tespit etmek ve önlemek çok önemlidir. Bu çalışmada, akıllı evlere yönelik siber saldırıları tespit etmek ve önlemek için makine öğrenmesi tabanlı bir yöntem önerilmiştir. Öncelikle "Home Assistant" teknolojisini kullanarak akıllı ev platformu oluşturulmuştur. Akıllı evler, "Home Assistant" teknolojisini kapsamlı bir şekilde kullanır. Oluşturulan akıllı ev platformu, sensörler ve kameralardan yararlanıyor. İnsanlar, sensörler ve kameralar kullanarak evlerini uzaktan izleyebilmekte ve yönetebilmektedir. Geliştirilen akıllı ev platformu üzerinde "brute force ftp", "brute force ssh", "dos http flood", "dos icmp flood", "dos syn flood", "syn scan" ve "udp scan" olmak üzere yedi saldırı gerçekleştirilmiştir. Toplanan veri seti, "normal" paketlerle birlikte sekiz sınıftan oluşmaktadır. Sekiz sınıf için toplam 435815 örnek veri toplanmıştır. Elde edilen bu veri seti üzerinde XGBOOST algoritması kullanılmış ve saldırı türleri sınıflandırılmıştır. Hold-out 80:20 ve Hold-out 70:30 eğitim testi verileri için sırasıyla %92.55 ve %92.49 doğruluk hesaplanmıştır. Önerilen XGBOOST algoritmasının sonuçları, diğer makine öğrenimi algoritmalarının sonuçlarıyla karşılaştırılmış ve sonuçların başarılı olduğu görülmüştür.

Anahtar kelimeler: Nesnelerin İnterneti, DDOS, Brute Force, Flood, XGBOOST, Home Assistant

XGBoost Based Intrusion Detection Method for Smart Home Systems

Abstract:

In today's smart homes, the infrastructure of IoT (Internet of Things) technology is used. As the use of smart homes increases, cyber attacks in this area are also increasing. It is very important to detect and prevent cyber attacks on smart homes as early as possible. In this study, a machine learning-based method is proposed to detect and prevent cyber attacks against smart homes. First of all, a smart home platform was created using the "Home Assistant" technology. Smart homes make extensive use of "Home Assistant" technology. The created smart home platform makes use of sensors and cameras. People can monitor and manage their homes remotely using sensors and cameras. Seven attacks, namely "brute force ftp", "brute force ssh", "dos http flood", "dos icmp flood", "dos syn flood", "syn scan" and "udp scan" were carried out on the developed smart home platform. The collected dataset consists of eight classes with "normal" packages. A total of 435815 sample data were collected for eight classes. XGBOOST algorithm was used on this obtained dataset and attack types were classified. For Hold-out 80:20 and Hold-out 70:30 training test data, 92.55% and 92.49% accuracy were calculated, respectively. The results of the proposed XGBOOST algorithm were compared with the results of other machine learning algorithms and the results were found to be successful.

Keywords: Internet of Things, DDOS, Brute Force, Flood, XGBOOST, Home Assistant.

1. Giriş (Introduction)

Bilgisayar ve ağ teknolojilerinde yaşanan gelişmeler sayesinde internet çok önemli bir noktaya gelmiştir. İnternetin günümüzde çok önemli noktaya gelmesi ile birlikte hayatımızın hemen hemen her alanında bize katkı sağlamaktadır (televizyon, bulaşık makinesi, akıllı ev sistemleri, ulaşım araçları,

* Sorumlu yazar.
E-posta adresi: orhanyaman@firat.edu.tr

Alındı : 17 Şubat 2022
Revizyon : 12 Ocak 2023
Kabul : 19 Temmuz 2023

kameralar vs.). İnternete bağlı cihaz sayısındaki bu artış, Nesnelerin İnterneti(IoT) kavramının ortaya çıkmasını sağlamıştır. Nesnelerin İnterneti, fiziksel nesnelerin birbiriyle bağlantılı olup günlük görevlerimizi kolaylaştırmak için hem fiziksel hem de dijital nesneleri entegre eder. Bununla beraber günümüzde gelişmekte olan internet ve internet ile bağlantılı olan cihazlar saldırganların hedef noktası haline gelmiştir. Nesnelerin İnterneti (IoT) özellikli cihazlarda ve ortamda güvenlik, davetsiz misafirlerin veya kötü niyetli kullanıcıların, yeterli güvenlik önlemlerinin eksikliğinde IoT özellikli sistemleri tehlikeye atma kabiliyetleri nedeniyle önemli bir konudur (Okegbile and Ogunranti, 2020). Son yıllarda IoT alt yapısı, IoT ağ anormalliği ve saldırı tespiti alanında çalışmalar artmakta ve araştırmacılar bu sorunun üstesinden gelmek için yöntem geliştirmektedir (Shafiq et al., 2020). Ericsson raporuna göre (Ericsson, 2020), IOT bağlantıları, derin öğrenme gibi yapay zeka (AI) algoritmaları kullanarak algılama verileri, analiz ederek IoT sistemleri tarafından 26,9 milyara ulaşılacağı ön görülmektedir.

Bu çalışmada, IoT ağındaki saldırılar için makine öğrenmesi algoritması kullanılarak belirlenen etkili özellikleri bulmak ve makine öğrenimi yöntemlerinin performansını optimize etmek için ADABOOST, GBM, XGBOOST, LGBM, CATBOOST, MLP, KNN, DT ve NB algoritmaları kullanılmıştır.

1.1. Literatür Özeti (Literature Review)

Zhang vd., IoT ağ ortamı üzerinden DDoS saldırısı için hafif bir savunma algoritması önerilmiş ve farklı ağ düğümleri arasındaki etkileşimli iletişimi incelemek için çeşitli senaryolara karşı test etmiştir (Zhang and Green, 2015).

Gupta vd., çalışmalarında, IoT sistemlerindeki zorluklardan, IoT ağında ki güvenlik ihtiyaçlarından ve IoT güvenliğinde devam eden araştırma ve zorlukları incelemişlerdir. Ayrıca IoT için herhangi bir çözüm tasarlarken dikkate alınması gereken tasarım yönergeleri tartışılmıştır (Gupta and Shukla, 2016).

Hussein vd., çalışmalarında, genel uygulamalar, özellikle gerçek zamanlı sistemlerdeki kritik uygulamalar için bir IoT platformu için tasarım ve uygulama önermişlerdir. Ayrıca, basit bir iletişim protokolü sunulmuş, çoklu konu özelliğini destekleyerek çok konulu mesajlaşma için gerekli olan trafiği ve gecikmeyi artıracak, önerilen RTOS protokolü için MQTT'ye karşı bir performans analizi gerçekleştirilmiştir. Analiz sonucu önerilen protokol, MQTT'den daha fazla ek konu için daha düşük bayt ekleyen çoklu konu özelliği nedeniyle MQTT protokolünden daha düşük gecikme ve daha düşük trafiğe sahiptir (Hussein, Zorkany and Abdel Kader, 2018).

Yavuz, çalışmasında derin öğrenme tabanlı güvenlik sistemi sunmuştur. Derin öğrenme de kullanılacak veri seti Cooja simülatörü ile hazırlanmıştır. Çalışmada, Cooja IoT simülatörü, 1000 düğüme kadar değişen IoT ağlarında yüksek kaliteli saldırı verilerinin oluşturulması için kullanılmıştır. Eğitilen veri seti ile %99 doğruluk hesaplanmıştır (Yavuz, 2018).

Ahmed, çalışmasında nesnelerin interneti için yeni sistemler, bir rakibin hedeflerine ve sistemine bağlı olarak farklı şekillerde birçok tehditle karşı karşıya olabileceğini vurgulamıştır. Satıcılara dayalı bir model kullanılarak yapılan bir sistemde, katkıda bulunan satıcılardan biri kötü niyetle hareket edebileceği ve sistemi olumsuz ekleyebileceğinden bahsetmiştir (Ahmed, 2021).

Lawal vd., çalışmalarında, hızlı ve doğru saldırı algılamasını sağlamak için sis hesaplama kullanan IoT için bir DDoS azaltma çerçevesi önermişlerdir. Sis, azaltma çerçevesinin etkili bir şekilde yerleştirilmesi için kaynaklar sağlar, bu, kısıtlı kaynak IoT cihazlarının kaynaklarındaki açıkları çözebilmektedir. Azaltma çerçevesi, anormallik tabanlı bir saldırı algılama yöntemi ve bir veri tabanı kullanılmıştır. Veri tabanı, önceden tespit edilen saldırıların imzalarını saklarken, anormallik tabanlı tespit şeması, DDoS saldırılarını tespit etmek için KNN sınıflandırma algoritması kullanılmıştır. KNN sınıflandırma algoritmasının DDoS saldırılarının tespiti için %99 oranında doğruluk sağladığı sonucuna varılmıştır (Lawal, Shaikh and Hassan, 2021).

Choi vd., çalışmalarında, akıllı ev-Nesnelerin İnterneti alanındaki araştırma makalelerini analiz etmek için, önemli uluslararası konferanslarda sunulan ve saygın dergilerde yayınlanan makaleleri çıkararak bir bibliyometrik yaklaşım izlenmiştir. Burada sunulan bulgular, akıllı ev-Nesnelerin İnternetinin gelecekteki yönleri için önemli bilgiler sunmaktadır. Ayrıca, akıllı ev-Nesnelerin İnternetinin hem temel eğilimleri hem de bilgi alanları sunulmuştur (Choi et al., 2021).

Srinadh vd., IoT uygulamalarındaki güvenlik tehditlerine ve tehdit kaynaklarına kapsamlı bir genel bakış sunulmuştur. Ek olarak, IoT güvenliği araştırmasının mevcut durumu ve IoT güvenliği ve gizliliği ile ilgili gelecekteki araştırma bilgileri sunulmuştur (Srinadh et al., 2021).

Literatürde verilen çalışmalarda IoT sistemlerin güvenliğinin önemli olduğu vurgulanmaktadır. Bu kapsamda akıllı evlerde kullanılan cihazlarında siber saldırılardan korunması gerektiği görülmektedir.

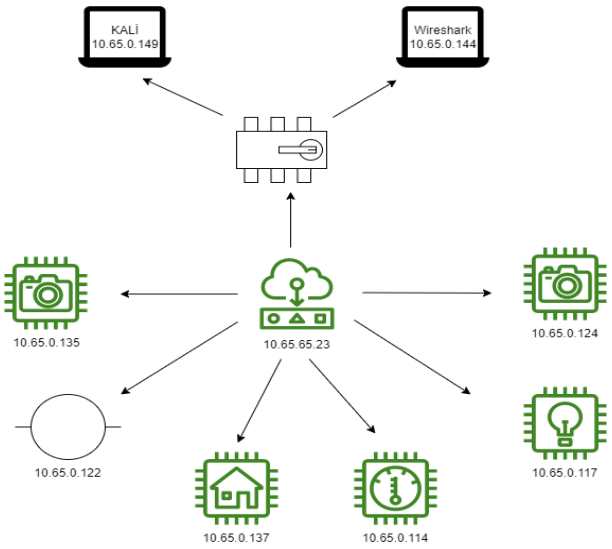
1.2. Motivasyon (Motivation)

Bu çalışmanın motivasyonu; akıllı ev sistemlerinde olası saldırıların tespit edilebilmesi için makine öğrenmesi tabanlı yöntemlerin uygulanmasıdır. Home Assistant teknolojisi kullanılarak akıllı ev ortamı oluşturulmuştur. Bu akıllı ev ortamında sıcaklık, nem sensörleri, kameralar, aydınlatma, havalandırma ve diğer bileşenler mevcuttur. Geliştirilen akıllı ev sistemine ağ üzerinden saldırılar düzenlenmiştir. Saldırı sırasında paket analizleri yapılarak özellik

çıkarımı yapılmaktadır. Elde edilen özellikler sınıflandırılarak saldırı tespiti yapılması amaçlanmıştır. Saldırıların tespiti için XGBOOST algoritması kullanılmıştır.

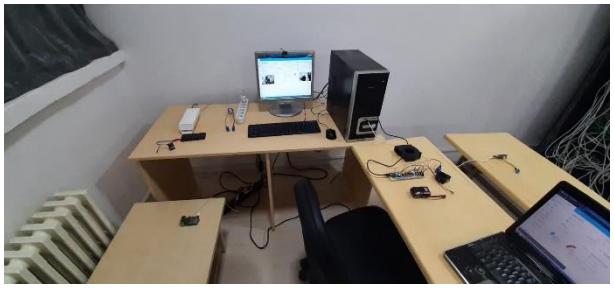
2. Geliştirilen IoT Tabanlı Akıllı Ev Modeli
(Developed IoT Based Smart Home Model)

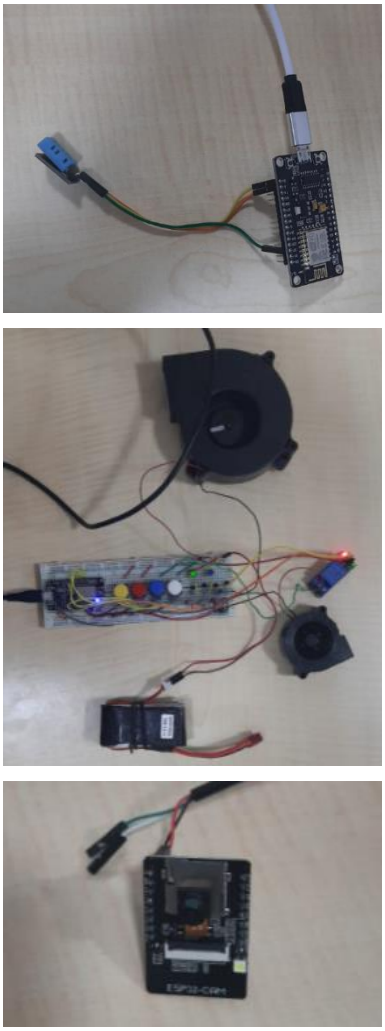
Bu çalışmada, IoT tabanlı akıllı ev laboratuvarından veri seti toplamak için Şekil 1'de verilen mimari oluşturulmuştur. Laboratuvar ortamımızda Linux işletim sistemli bir makine, ağda ki paketleri toplamak için wireshark yüklü olan Windows işletim sistemi yüklü bilgisayar, akıllı switch, IoT cihazları ve IoT cihazları akıllı ev sistemine bağlamak için kablosuz erişim noktası kullanılmıştır.



Şekil 1. IoT akıllı ev laboratuvar mimarisi (IoT smart home laboratory architecture)

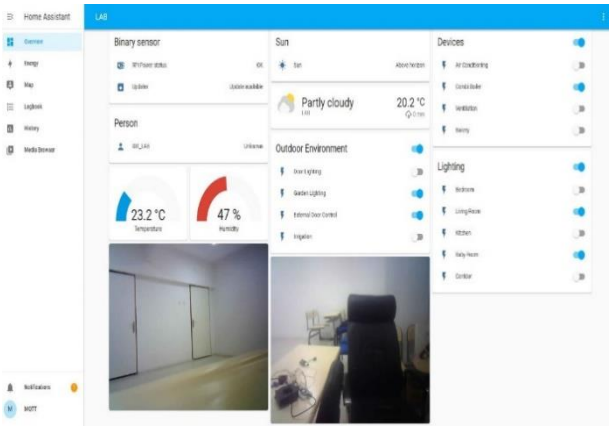
Şekil 1'de önerilen mimari ESP tabanlı gömülü kartlar, sensörler ve diğer bileşenler kullanılarak uygulanmıştır. Akıllı ev ortamı oluşturularak sensör düğümlerinden oluşan bir laboratuvar alt yapısı kurulmuştur. Bu çalışma kapsamında oluşturulan akıllı ev laboratuvar görüntüleri Şekil 2'de sunulmuştur.





Şekil 2. Oluşturulan akıllı ev laboratuvar görüntüleri (Created smart home lab images)

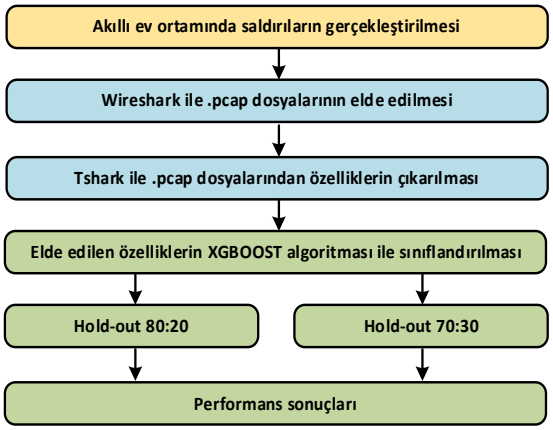
Şekil 2'de görülen akıllı ev ortamının uzaktan izlenmesi ve kontrol edilebilmesi için "Home Assistant" teknolojisi kullanılmıştır. Home Assistant lokalde kontrolü ve gizliliği amaç edinen açık kaynak kodlu bir akıllı ev otomasyonu teknolojisidir. Açık kaynak olması sayesinde tek bir kurum tarafından değil bu alanda ilgili olan herkes tarafından geliştirilebilmektedir. Bir Raspberry Pi üzerinde ya da mevcut sunucular üzerine kolaylıkla kurulabilmekte ve kullanılabilmektedir. Bu çalışma kapsamında kurulan deneysel ortam üzerinde Home Assistant uygulamasına ait ekran görüntüleri Şekil 3'te gösterilmiştir.



Şekil 3. Akıllı ev laboratuvarında Home Assistant uygulaması sonuçları (Home Assistant app results in smart home lab)

3. Önerilen Yöntem (Proposed Method)

Bu çalışmada akıllı ev ortamlarında oluşabilecek saldırıların hızlı ve yüksek doğrulukla tespit edilebilmesi için hafifsıklet bir yöntem önerilmiştir. Geliştirilen uygulamanın adımları Şekil 4'te verilmiştir.

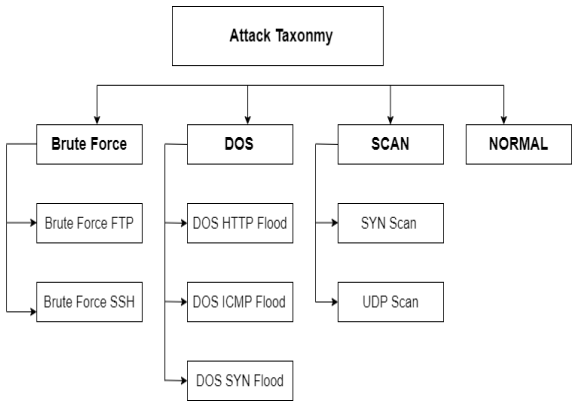


Şekil 4. Bu çalışmada geliştirilen uygulamanın adımları (The steps of the application developed in this study)

Şekil 4'te de görülebileceği gibi önerilen yöntem üç aşamadan oluşmaktadır. İlk olarak akıllı ev ortamında saldırıların gerçekleştirilmesidir. Daha sonra veri setinin elde edilmesi ve XGBOOST algoritması kullanılarak sınıflandırılması aşamalarından oluşmaktadır.

3.1. Akıllı ev ortamında saldırıların gerçekleştirilmesi (Performing Attacks In a Smart Home Environment)

IoT laboratuvar da saldırılar gerçekleştirilirken 10.65.0.149 ip adresli Linux işletim sistemi üzerinden brute force, dos ve scan olmak üzeri 3 farklı saldırı gerçekleştirilmiştir. Bu saldırılar brute force ftp, brute force ssh, dos http flood, dos icmp flood, dos syn flood, normal, syn scan ve udp scan olmak üzere 8 farklı şekilde gerçekleştirilmiştir. Bu saldırıların taksonomisi Şekil 5'de gösterilmiştir.



Şekil 5. Saldırı Taksonomisi (Attack Taxonmy)

Şekil 5'te verilen 'Brute Force' saldırılarının gerçekleştirilmesi için' xHydra' aracı kullanılmıştır. 'DOS' saldırılarının uygulanması için 'Hping3' ve 'Scan' saldırılarının gerçekleştirilmesi için 'NMAP' araçları kullanılmıştır.

Brute Force; Saldırgan doğru olanı tahmin etme umuduyla IoT cihazlarda parola deneme saldırıları gerçekleştirir. Bu saldırı da çok sayıda ardışık parola deneme isteği var ise brute force saldırısı olarak tanımlanabilir. Saldırganlar elde ettikleri parola ile dosyalara erişim sağlayabilirler.

DOS; Saldırgan internete bağlı IoT cihazı geçici veya süresiz olarak ağı aksatarak kullanıcının asıl alması gereken pakete erişmesini engellemekte ve ağ trafiğini doldurmaktadır. DoS saldırıları, web sunucuları, bankacılık, ticaret, hükümet ve medya kuruluşlarını hedef almaktadır. DoS saldırıları için genellikle sel ve çökertme olmak üzere iki yöntem kullanılmaktadır. 'HTTP Flood', 'ICMP Flood' ve 'SYN Flood' saldırıları en çok kullanılan saldırılar olarak bilinmektedir.

'HTTP Flood' ağ adresine çok fazla trafik gönderilir, ağ iletişiminin aksamasına neden olmaktadır. 'ICMP Flood' bir makine yerine ağdaki tüm bilgisayarlara ping paketi göndererek ağ trafiğini doldurmaktadır. 'SYN Flood' sunucuya bağlanmak için istek gönderilir, paket hiçbir zaman yerine ulaşmaz. Tüm açık bağlantılar isteklerle doldurularak asıl kullanıcının bağlanmaması sağlanmaktadır. Gelişen teknoloji DoS saldırılarına karşı savunma mekanizması geliştirmiştir fakat DDoS benzersiz özellikleri nedeniyle yüksek bir tehdit oluşturmaktadır.

SCAN; Bir ağda ki hangi bağlantı noktalarının açık olduğunu belirleme yöntemidir.

3.2. Veri Setinin Elde Edilmesi (Obtaining the Data Set)

Bu aşamada, ham ağ trafik paketleri toplanmıştır. Bu işlem, toplanan .pcap dosyalarından özellik çıkarılmasını sağlamaktadır. Bu çalışmada IoT

laboratuvar ortamında gerçekleştirilen saldırılar wireshark ile dinlenmiştir. Wireshark ile elde edilen .pcap dosyalarından özellik çıkarılmıştır. .pcap dosyalarından özellik çıkarmak için tshark.exe kullanılmıştır. Çıkarılan özelliklerle ilgili özellikler Tablo 1'de verilmiştir.

Tablo 1. Saldırı paketlerinden elde edilen özellikler(Features obtained from attack packs)

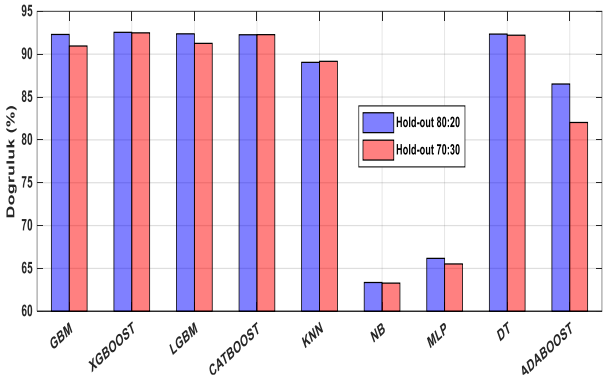
	Özellik Adı	Özellik Tanımı
1	ip.len	İp adres uzunluğu
2	ip.flags.df	Datagramın parçalanma değeri
3	ip.flags.mf	Datagramın ek parça değeri
4	ip.ttl	Datagramın ömrünün değeri
5	ip.proto	Sonraki kapsüllenmiş protokol değeri
6	ip.version	İp versiyon değeri
7	udp.port	UDP portu
8	tcp.windows_size	TCP pencere uzunluğu
9	tcp.ack	TCP onay numarası
10	tcp.seq	TCP sıra numarası
11	tcp.len	TCP başlık uzunluğu
12	tcp.stream	TCP akış değeri
13	tcp.analysis.ack_rtt	ACK'nin yakalanması arasında ki zaman değeri
14	tcp.reassembled.length	TCP birleştirme değeri uzunluğu
15	tcp.time_relative	TCP oturumunda ilk çerçeveyi aldığı andan geçen süre değeri
16	tcp.time_delta	TCP oturumunda önceki ve mevcut paket arasında geçen süre değeri
17	class	Saldırı yapılan değerler sınıflandırılmıştır.

Tablo 1'de belirtilen veri setinde 16 tane özellik ile 8 sınıf (brüte force ftp, brüte force ssh, dos http flood, dos icmp flood, dos syn flood, syn scan, udp scan ve normal) oluşturulmuştur. Oluşturulan bu veri seti XGBOOST algoritması kullanılarak sınıflandırılmıştır.

3.3. XGBOOST ile Sınıflandırma (Classification with XGBOOST)

XGBoost algoritması, gradyan artırma çerçevesi kullanan karar ağacı tabanlı makine öğrenme algoritmasıdır. XGBoost yapılandırılmış verileri içerir. İçerdiği yapılandırılmış veriler ile birçok algoritmayı geride bırakmaktadır. XGBoost geniş bir uygulama alanı bulunmaktadır. Regresyon, sınıflandırma, sıralama gibi problemleri çözmek için kullanılan bir algoritmadır. En kısa zaman da daha az kaynak tüketimi kullanarak yüksek değerli sonuçlar elde edilebilir. XGBoost'un diğer algoritmalara nazaran daha iyi performans göstermesinin sebebi, gradyan iniş mimarilerini kullanarak zayıf öğrenenlere artırma ilkesi uygulanmasıdır.

Önerilen yöntemin belirlenmesi için GBM, XGBOOST, LGBM, CATBOOST, KNN, NB, MLP, DT, ADABOOST algoritmaları kullanılmıştır. Bu algoritmalar kullanılarak Hold-out 80:20 ve Hold-out 70:30 eğitim ve test verileri ile Şekil 6'da verilen sonuçlar hesaplanmıştır.



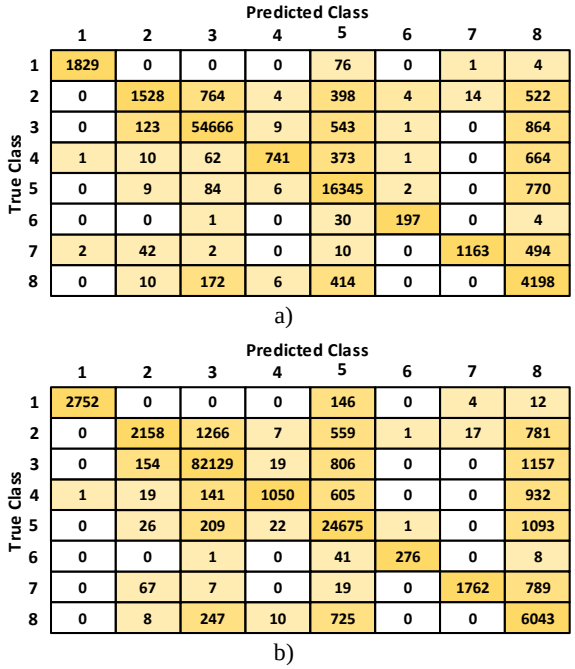
Şekil 6. GBM, XGBOOST, LGBM, CATBOOST, KNN, NB, MLP, DT, ADABOOST algoritmaları ile elde edilen sonuçlar (Results obtained with GBM, XGBOOST, LGBM, CATBOOST, KNN, NB, MLP, DT, ADABOOST algorithms)

Şekil 6'da verilen sonuçlar incelendiğinde XGBOOST algoritmasının en yüksek doğruluk elde ettiği görülmüştür. XGBOOST algoritması ile Hold-out 80:20 ve Hold-out 70:30 eğitim test verisi için %92.55 ve %92.49 doğruluk hesaplanmıştır. Böylece bu çalışmada XGBOOST algoritması tercih edilmiştir.

4. Deneysel Sonuçlar (Experimental Results)

Bu çalışma Python 3.10 programı kullanılarak geliştirilmiştir. Oluşturulan akıllı ev platformu ile veri seti toplanmış ve özellik çıkarımı yapılmıştır. Elde edilen özellikler GBM, XGBOOST, LGBM, CATBOOST, KNN, NB, MLP, DT, ADABOOST sınıflandırıcılar ile sınıflandırılmıştır. Bu

sınıflandırıcılar içerisinde en yüksek doğruluk XGBOOST ile hesaplanmıştır. XGBOOST sınıflandırıcısı sonucunda elde edilen hata matrisi Şekil 7'de sunulmuştur.



Şekil 7. XGBOOST sınıflandırıcısı sonucunda elde edilen hata matrisi a) 80:20 eğitim test sonucu b) 70:30 eğitim test sonucu (Confusion matrix obtained as a result of XGBOOST classifier a) 80:20 training test result b) 70:30 training test result)

XGBOOST sınıflandırıcısı kullanılarak Hold-out 80:20 ve Hold-out 70:30 eğitim test sonuçlarında elde edilen sınıf doğrulukları Tablo 2'de gösterilmiştir.

Tablo 2. XGBOOST sınıflandırıcısı ile elde edilen sınıf doğrulukları (Class accuracies obtained with the XGBOOST classifier)

Sınıf	Sınıf adı	Doğruluk (%)	
		Hold-out 80:20	Hold-out 70:30
1	Brute Force FTP	95.75	94.44
2	Brute Force SSH	47.24	45.06
3	DOS HTTP Flood	97.26	97.46
4	DOS ICMP Flood	40.01	38.20
5	DOS SYN Flood	94.94	94.80
6	Normal	82.77	84.66
7	SYN Scan	67.89	66.64
8	UDP Scan	87.45	85.92

Tablo 2'de gösterildiği gibi Hold-out 80:20 eğitim test verileri kullanılarak en yüksek doğruluk %97.26 ile DOS HTTP Flood sınıfı için hesaplanmıştır. En düşük doğruluk ise %40.01 ile DOS ICMP Flood sınıfında elde edilmiştir. Hold-out 70:30 eğitim test verileri içinde en yüksek ve en düşük doğruluklar sırasıyla DOS HTTP Flood ve DOS ICMP Flood sınıflarında hesaplanmıştır.

5. Sonuçlar ve Tartışma (Conclusions and Discussion)

Nesnelerin interneti günümüzde birçok alanda kullanılmaktadır. Bu teknolojiler ile nesneler uzaktan izlenmekte ve yönetilmektedir. Bu teknolojinin uzaktan yönetilebilmesi beraberinde siber saldırıları da getirmektedir. Bu çalışmada IoT saldırıları türlerinin tespiti için XGBOOST sınıflandırıcı kullanılmıştır. Akıllı ev platformu oluşturulmuş ve veri seti toplanmıştır. Toplanan veri setinde 16 özellik olmak üzere toplamda 435815 örnek mevcuttur. Bu veri seti brute force ftp, brute force ssh, dos http flood, dos icmp flood, dos syn flood, normal, syn scan ve udp scan olmak üzere sekiz sınıftan oluşmaktadır. XGBOOST sınıflandırıcısı kullanılarak Hold-out 80:20 eğitim test verisi için %92.55 doğruluk hesaplanmıştır.

XGBOOST sınıflandırıcısı ile diğer sınıflandırıcıların (GBM, LGBM, CATBOOST, KNN, NB, MLP, DT, ADABOOST) karşılaştırılması Tablo 3'te listelenmiştir.

Tablo 3. XGBOOST sınıflandırıcısı ile diğer sınıflandırıcıların karşılaştırılması (Comparison of XGBOOST classifier and other classifiers)

Sınıflandırıcılar	Doğruluk (%)	
	Hold-out 80:20	Hold-out 70:30
GBM	92.29	90.95
XGBOOST	**92.55**	**92.49**
LGBM	92.37	91.25
CATBOOST	92.26	92.27
KNN	89.04	89.16
NB	63.35	63.28
MLP	66.17	65.52
DT	92.34	92.21
ADABOOST	86.51	82.02

Teşekkür (Acknowledgment)

Bu çalışma TEKF.21.18 numaralı Fırat Üniversitesi Bilimsel Araştırma Projeleri (FÜBAP) Koordinasyon Birimi tarafından desteklenmiştir.

Kaynaklar (References)

Ahmed, M.S. (2021) "Designing of internet of things for real time system," Materials Today: Proceedings [Preprint]. doi:10.1016/j.matpr.2021.03.527.

Choi, W. et al. (2021) "Smart home and internet of things: A bibliometic study," Journal of Cleaner Production, 301, p. 126908. doi:10.1016/j.jclepro.2021.126908.

Ericsson (2020) Ericsson Mobility Report.

Gupta, K. and Shukla, S. (2016) "Internet of Things: Security challenges for next generation networks," in 2016 1st

International Conference on Innovation and Challenges in Cyber Security, ICICCS 2016. Institute of Electrical and Electronics Engineers Inc., pp. 315–318. doi:10.1109/ICICCS.2016.7542301.

Hussein, M., Zorkany, M. and Abdel Kader, N.S. (2018) "Design and Implementation of IoT Platform for Real Time Systems," in Advances in Intelligent Systems and Computing. Springer Verlag, pp. 171–180. doi:10.1007/978-3-319-74690-6_17.

Lawal, M.A., Shaikh, R.A. and Hassan, S.R. (2021) "A DDoS Attack Mitigation Framework for IoT Networks using Fog Computing," Procedia Computer Science, 182, pp. 13–20. doi:10.1016/j.procs.2021.02.003.

Okegbile, S.D. and Ogunranti, O.I. (2020) "Users emulation attack management in the massive internet of things enabled environment," ICT Express, 6(4), pp. 353–356. doi:10.1016/j.icte.2020.06.005.

Shafiq, M. et al. (2020) "IoT malicious traffic identification using wrapper-based feature selection mechanisms," Computers and Security, 94, p. 101863. doi:10.1016/j.cose.2020.101863.

Srinadh, V. et al. (2021) "An analytical study on security and future research of Internet of Things," Materials Today: Proceedings [Preprint]. doi:10.1016/j.matpr.2020.12.342.

Yavuz, F.Y. (2018) Deep Learning in Cyber Security for Internet of Things, Yüksek Lisans Tezi, Istanbul City University.

Zhang, C. and Green, R. (2015) "Communication security in internet of thing: Preventive measure and avoid DDoS attack over IoT network," Simulation Series, 47(3), pp. 8–15.

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 6(2) (2023) 159-173
DOI: 10.38016/jista.1226528

Makine Öğrenmesi Yöntemleriyle Anormal İçme Suyu Tüketimlerinin Tespit Edilmesi ve Tahmin Modellerinin Geliştirilmesi*

İsmail Güney[1*], İhsan Hakan Selvi[2]

[1] Sakarya Üniversitesi, Bilişim Sistemleri Mühendisliği Bölümü, Sakarya, Türkiye

[2] Sakarya Üniversitesi, Bilişim Sistemleri Mühendisliği Bölümü, Sakarya, Türkiye

ismailguney@gmail.com, ihselvi@sakarya.edu.tr

Öz

Bu çalışmada, içme suyu gibi önemli bir ihtiyacın hane halkı tarafından tüketiminde belirli bir düzen olabileceği gibi, farklı etkenlere bağlı olarak düzensiz tüketimin de olabileceği öngörülmektedir. Artan nüfus, sınırlı içme suyu kaynakları, gelişen alt yapı ve teknoloji, içme ve kullanma suyuna olan talebi artırmıştır. Artan talebi karşılamak için alternatif su kaynağı arayışları yanında mevcut suların israf edilmemesinin ve daha verimli kullanılmasının da etkili olacağı öngörülmektedir. Yapay zekanın (AI) alt dalı olan makine öğrenmesi (ML) yöntemleriyle geçmiş dönemlerdeki içme suyu tüketimleri analiz edilmiş, olağan ve olağan dışı tüketim davranış modelleri çıkarılmıştır. İçme suyu mesken abonelerinin anormal tüketimlerinin tespiti ve bilgilendirilmeleri durumunda, hane içi tüketimlerin normal tüketim aralığında kalmasının sağlanacağı öngörülmektedir. Çalışmada Kayseri ili genelinde 2006 – 2022 (ilk 6 ay) tarihleri arasında sayaç endeks okuması 160 dönemden fazla olan 8.224 adet mesken abonesine ait sayaç, abone ve tüketim verileri dikkate alınmıştır. Veriler konumsal abone temelinde birleştirilmiş, 41 öznitelikli veri kümesi elde edilmiş, veri ön işlemleri sonucunda 24 öznitelikli bir veriseti oluşturulmuştur. Çalışmada 6 farklı öznitelik seçim yöntemi kullanılarak alt verisetleri elde edilmiştir. Bütün verisetler 7 farklı anomali analiz yöntemi kullanılarak anormal ve normal içme suyu tüketimleri tespit edilmiştir. Anomali analizleri sonucunda hesaplanan aykırılık puanları kullanılarak bütün tüketim değerleri 4 farklı tüketim sınıfı ile etiketlenmiş, veriseti gözetimli hale getirilmiş, 7 farklı ML sınıflandırma algoritması ile tüketim sınıf tahmin modelleri geliştirilmiştir. Çalışma sonucunda anormal içme suyu tüketimlerinin ML yöntemleri ile tespit edilebileceği, tüketim sınıflarının tahmin edilebileceği ispatlanmış, suyun israf edilmeden daha verimli kullanımıyla ilgili gerekli politikaların oluşturulabileceği ve bunun için önlemler alınabileceği ortaya konmuştur.

Anahtar kelimeler: İçmesuyu, abone, tüketim analizi, makine öğrenmesi, anomali analizi, anomali tespiti, aykırı değer, etkili durum, etkili durumlar, etkili gözlemler.

Detecting Abnormal Drinking Water Consumptions And Developing Forecast Models By Machine Learning Methods

Abstract

In this study, it is predicted that there may be a certain order in the consumption of an important need such as drinking water by the household, as well as irregular consumption depending on different factors. Increasing population, limited drinking water resources, developing infrastructure and technology have increased the demand for drinking and utility water. There is a search for alternative water sources to meet this demand, but it is foreseen that these demands can be met by not wasting existing water and using it more efficiently. By using machine learning (ML) methods, which is a sub-branch of artificial intelligence (AI), drinking water consumption data in the past periods were analyzed, and ordinary and unusual consumption behavior models were extracted. It is envisaged that by detecting abnormal consumptions that may occur in drinking water consumption and informing the subscribers about this issue, it will be ensured that the consumption in the household remains within the normal consumption range. Although the amount of data collected, recorded and processed in today's IT world has increased significantly, it is known that the exact analysis is difficult in terms of time and cost. In this study, subscriber, meter, consumption, bill and payment data of 8,224 residential subscribers, whose water meter index reading is more than 160 periods throughout the province of Kayseri, between 2006 and 2022 (first 6 months) were taken into account. The data are combined on a spatial subscriber basis and a 41-features dataset is obtained. The dataset was transformed into a dataset

* Bu çalışma İsmail GÜNEY'in yüksek lisans tezinden üretilmiştir.

Sorumlu yazar.
E-posta adresi: ismailguney@gmail.com

Alındı : 30 Aralık 2022
Revizyon : 21 Temmuz 2023
Kabul : 1 Ağustos 2023

with 24 features as a result of data preprocessing. In the study, 6 sub-datasets were obtained by using information gain (IG), gain ratio (GR), symmetric uncertainty coefficient (SU), pearson correlation coefficient (r), f-score and random forest (RF) feature selection methods. The 7th sub-dataset was obtained from the intersections of the selected features in the sub-datasets. In all datasets, abnormal and normal drinking water consumptions were determined by using 7 different ML anomaly analysis methods: tukey outlier labeling (TOL), forest of isolation (IF), z-score, copula-based outlier detection (COPOD), median absolute deviation (MAD), local outlier factor (LOF), and elliptical envelope (EE). At the beginning of the study were unsupervised drinking water consumption data at the end of the study, labeled as 4 different classes and the dataset was made supervised. Using the finally obtained supervised dataset, decision trees (DT), gaussian naive bayes (NB), k-nearest neighbors (KNN), logistic regression (LJR), multilayer perceptron neural network (MLP-NN), RF and gradient boosting (GB) have been developed consumption class estimation models with 7 different ML methods. As a result of the study, it has been proven that abnormal drinking water consumption can be detected by ML methods, and it has been revealed that necessary policies can be created for more efficient use of water without wasting water and measures can be taken for this.

Keywords: Drinking water consumption analysis, machine learning, anomaly analysis, anomaly detection, outlier, effective case, effective cases, effective observations.

1. Giriş (Introduction)

Su, canlıların beslenme, temizlenme, ulaşım ve taşıma gibi temel nedenlerle sürekli ihtiyaç duyduğu bir araç olmakla birlikte özelde insanoğlunun da üzerinde her zaman hesap yapageldiği önemli bir değer olmuştur. Milattan önceki çağlardan günümüze kadar insanlık suyun temini, yönetilmesi, paylaşılması, taşınması ve kullanılması hususlarına önem vermiş, yer yer bu önem Tablo 1'de gösterilen çatışmaların yaşanmasına neden olmuştur. Çatışmalar suyun bir silah olarak veya savaşların tetikleyici nedeni olarak görülmesiyle veya su kaynaklı kazaların oluşmasıyla gerçekleşmiş böylece insanlığın tarih içindeki mücadelesinde önemli izler bırakan su, halen tarihe, coğrafyaya şekil vermeye devam edegelmiştir (Water Conflict Chronology [Internet], 2022).

Tablo 1. Tarihte su ile ilgili yapılan çatışma nedeni ve sayıları. (The reasons and numbers of conflicts related to water in history.)

Dönem Aralığı (Yıllar)	Çatışma Sayısı	Çatışmada	Dönem Aralığı (Yıllar)	Çatışma Sayısı
Milattan Önce	26	Silah, Kaza	10	6
0 - 1000	2	Silah, Tetikleyici,	1	2
1001 - 1900	25	Silah, Kaza, Tetikleyici	13	9
1901 - 2000	209	Silah, Kaza, Tetikleyici	101	14
2001 - 2022	1.036	Silah, Kaza, Tetikleyici	109	16
Toplam:	1.298	Silah, Kaza, Tetikleyici	178	16

Çatışmaların büyük çoğunluğunun son 20 yılda meydana geldiği, ayrıca 20. yüzyılda nüfus artışının dört kat, su talebinin dokuz kat artması dikkate alındığında su güvenliği konusunun günümüz dünyasında hissedilecek kadar hızla büyüyen sosyal, politik ve ekonomik sorunları tetiklediği, giderek yayılan bir çevresel krize dönüştüğü artık kabul edilen bir gerçektir.

Yapılan araştırmalarda 2030 yılına kadar küresel tatlı su talebinin mevcut arzı %40'ın üzerinde aşacağı ve etkilenecek insan sayısının 4 milyara yaklaşacağı, artan rekabet koşulları ve beraberinde su üzerinde oluşturduğu stresin, dünya genelinde gıda, enerji, üretim ve insan güvenliği açısından önemli bir risk ve etki oluşturacağı öngörülmektedir (Cini, Mung ve Waughray, 2014).

Kentsel yaşamın giderek yaygınlaştığı günümüzde en önemli tüketim kaynaklarından biri olan su, potansiyel olarak yüzeyde, yeraltında, havada farklı formlarda bulunmakla birlikte formlar arası değişim ve geçiş döngüsüne sahip olarak hayatın içinde yer almaktadır. Ülkemizde ve dünyada 20 yılı aşkın bir süre zarfında nesnelerin interneti (IOT) tabanlı cihazların, kablosuz haberleşmenin ve AI'nın gelişimiyle akıllı şehir konseptleri hızla yaygınlaşmış ve giderek alanını genişletmiş, hatta oldukça kompleks sorunları da çözebilecek etkinliğe kavuşmuştur. Akıllı şehirler kavramı beraberinde akıllı su şehirleri kavramına zemin oluşturmuş böylece Uluslararası Su Birliği (IWA) tarafından 4 aşamalı Akıllı Su Şehri İlkeleri Uygulama Planı oluşturulmuştur. Çalışma konusu ve sonuçlarının bu plan içerisinde tüketilen su ve enerji miktarının azaltılması, mevcut suyun en verimli şekilde kullanılması, tüketim davranış analizlerinin sunduğu geri dönütler ile abonelerin bilinçlendirilmesi, tüketimin bu sayede disipline edilmesi ve nihai olarak da tespit edilen aykırı tüketimlerin dikkate alındığı kayıp ve kaçak denetimleri ile su yönetiminin daha bilimsel yapılabilmesi hususlarında katkı sağlayacağı düşünülmektedir (International Water Association, 2017).

Su dağıtımının Yönetsel Kontrol ve Veri Toplama (SCADA) sistemleri ile yönetilmeye başlanması, verilerin toplanması ve kayıt altına alınması ile ciddi bir veri havuzu oluşmuş, beraberinde verilerle anlamlı sonuçlar elde edilebilecek analizlerin önü açılmış daha kompleks çözümler sunan AI modelleri geliştirilmiştir. Mevcut durumda su sektöründe AI ile geliştirilen uygulamalar su alt yapısının fiziki durumunun tespiti ve bakımı, su talebi ve tüketiminin tahmini, su

rezervuarlarının ve barajların sağlık ve çevresel etkilerinin izlenmesi, su kalitesinin izlenmesi ve su ile ilgili felaketlerin öngörülmesi ve izlenmesi olarak 5 başlıkta raporlanmıştır. Öyle ki AI destekli yeniliklerin 2030 yılına kadar küresel ekonomiye 200 milyar dolar katkı sağlayacağı tahmin edilmektedir (Yıldız ve Özgüler, 2020).

Küresel su döngüsü dikkate alındığında suyun gelecek yıllarda çoğu ülkeleri etkileyeceği, su ile ilgili yapılan menfi çalışmaların etkisinin yeterli olamayacağı, ülkelerin birlikte hareket etmesinin özellikle de gelişmiş ülke tecrübelerinin ve imkanlarının gelişmekte olan ve gelişmemiş ülkelere aktarılmasının gerekliliği bir zorunluluk halini almıştır. Bundan dolayıdır ki Birleşmiş Milletler (BM) bünyesinde bağlı bütün üye ülkelerin 2030 yılına kadar uygulayabilecekleri Sürdürülebilir Kalkınma Hedefleri (BM SKH) belirlenmiş, bu hedeflerden özellikle su ile ilgili olanlar için geliştirilen AI çalışmalarıyla hedeflenen en önemli gaye 2030 yılına kadar herkesin güvenli ve erişilebilir içme suyuna kavuşmasını sağlamak için, altyapıya yatırım yapmak, sıhhi tesisleri inşa etmek ve her düzeyde hijyeni teşvik etmek zorunluluğu olmuştur. Ayrıca hedefler arasında su kıtlığını hafifletmek için ormanlar, dağlar, sulak alanlar ve nehirler gibi suyla bağlantılı eko-sistemleri korumak ve eski haline getirmek zorunluluğu yanında gelişmekte olan ülkelerde su verimliliğini teşvik etmek ve arıtma teknolojilerini desteklemek için uluslararası iş birliğine de vurgu yapılmıştır (Türkiye Cumhuriyeti Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı, 2020).

Günümüzde musluğumuzu çok fazla hesap yapmadan açabiliyor olsak da bu durumun süreklilik arz etmeyeceği, su kaynakları kullanımının doyuma ulaşacağı, nüfus ve suya olan ihtiyacın artmaya devam edeceği konunun uzmanları tarafından da öngörülmektedir. Gelişen teknoloji ile birlikte suyun daha profesyonel anlamda değerlendirilmesine ve gelecekteki su ihtiyacına yönelik tahmin senaryolarının üretilmesine yönelik çok ciddi imkanlar elde edilmiş olup, bu senaryolara göre 2040 yılına gelindiğinde yerel su kaynaklarının yeniden kullanılacağı, geri dönüşümün zorunluluk halini alacağı, yağmur suyu yönetiminin yaygınlaşacağı, yeterince kullanılmayan su kaynaklarına erişimlerin sağlanacağı, doğrudan suyu arıtma ve arıtım yollarının yeşil alanlara veya parklara dönüştürüleceği, kentsel su hizmetlerinin gizli hizmetler olmaktan çıkarılıp daha görünür hale getirileceği ve teknolojinin bireysel haneler üzerinde büyük bir etkisi olacağı, bir mobil uygulama aracılığıyla gün boyunca su kullanımlarının gerçek zamanlı takip edilmesinin sağlanacağı, bu vesileyle tüketim davranış değişikliklerinin, hangi amaçlarla ne kadar su kullanıldığının ve hangi saatlerde kullanıldığının, şebekedeki su kayıp ve kaçaklarının tespitinin, su kullanımlarının faturaları nasıl etkileyeceğinin görülebileceği bir kentsel su modeli bizleri beklemektedir (Arup ve Sydney Water, 2015).

Hane içinde su tüketimleri incelendiğinde iç mekandaki tüketimlerin Tablo 2'de gösterildiği oranlarda olduğu, hane tüketimlerinin hane büyüklüğü ile orantılı olduğu ancak büyüklüğe oranla tüketimin daha az arttığı ve bir tasarruf yapılacaksa nerelerde yapılabileceği öngörülebilecektir (Mayer, DeOreo, Opitz, Kiefer, Davis, Dziegielewski, ve Diğerleri, 1999).

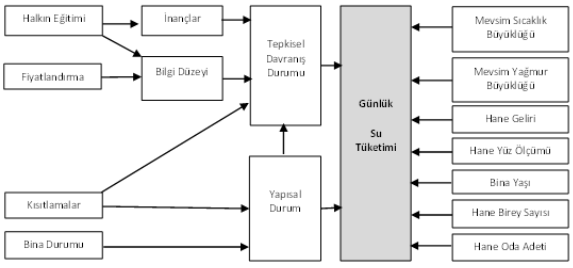
Tablo 2. Suyun hane içinde nihai kullanım yerleri ve oranları. (Uses and rates of water in the household.)

Son Kullanım	İç Mekân Su Kullanım Yüzdeleri
Tuvaletler	%26,7
Çamaşır Makineleri	%21,7
Duşlar	%16,8
Musluklar	%15,7
Sızıntı	%13,7
Diğer Lokal	%2,2
Banyolar	%1,7
Bulaşık Makineleri	%1,4
Toplam *	%100,0

*ondalık tek haneye indirgendiğinden %99,9

Ayrıca bilimsel çalışmalarda hane halkı dışındaki tüketim oranları her ne kadar kültüre ve coğrafyaya göre değişiklikler gösterse de nüfus ve yüz ölçüm ölçekleri dikkate alınarak yaklaşık olarak hesaplanabilmektedir (Mayer, Dziegielewski, Kiefer, Lantz, Opitz ve Porter, 2000).

Hanelerin günlük su tüketimlerini etkileyen Şekil 1'de gösterildiği gibi birçok faktör belirlenmiş ve günlük su tüketim değerlendirmesi yapılırken bunlar göz önünde bulundurulmuş, bunların haricinde de birçok etmenin var olduğu dikkate alınarak bu etmenler arasındaki ilişkilerin tüketimi tetiklediği gerçeğinden hareketle bir konut suyu koruma modeli uyarlanmıştır (Billings ve Jones, 2008).



Şekil 1. Konut suyu koruma modeli (Residential water protection model.)

Bu çalışmanın amaçlarından biri olan su tasarrufuyla ilgili Tablo 3'te gösterildiği gibi 5 ana başlık altında bir dizi önerilerde bulunulmuş, su yöneticileri ve nihai tüketicilerin bu önerilerden birçoğunu yaptıkları, ilgili planlamaların giderek yaygınlaştığı sonraki dönemlerde

daha fazlasının hayatın içinde yer alacağı görülmektedir (Campbell, Johnson ve Larson, 2004).

Tablo 3. Su tasarrufu ile ilgili öneriler. (Tips on saving water.)

Yönetmelikler	Kamuya Yönelik Bilgilendirmeler	Mühendislik	Fiyatla İlgili Stratejiler	Dolaylı Mühendislik
• Az su kullanan armatürleri belirten tesisat kodları • Düşük su kullanımı gerektiren peyzaj düzenlemeleri kullanımı gerektiren peyzaj düzenlemeleri • Su israfı yasağı • Belirli dış mekân amaçları için ham veya içilemez su gereksinimleri • Yeniden satışta güçlendirme gereksinimleri	• Genel tanıtım: reklam panoları, postalar, su faturası ekleri • Herkese açık forumlar • Gazete, radyo, TV, internet kullanımı • İlköğretim ve ortaöğretimlerde eğitim programları • Verimli sulama sistemlerini teşvik etmek için çalıştaylar • Su bazlı çevre düzenlemesini teşvik eden atölyeler (xeriscape) • Günlük hava durumu raporlarına peyzaj sulama ihtiyaçlarının eklenmesi	• Yeni nesil tuvaletler için ücretsiz dağıtım veya sübvansiyonlar • Düşük akışlı duş başlıkları ve musluk havalandırıcılarının ücretsiz dağıtımı • Az su kullanan giysiler veya bulaşık makineleri için sübvansiyonlar • Peyzaj ve yağmurlama sistemi dönüşümü için sübvansiyonlar • Yağmur suyunun toplanması ve peyzajda kullanımı için sübvansiyonlar • Çevre düzenlemesinde gri suyun yeniden kullanımı için sübvansiyonlar • Geri kazanılmış su dağıtım sisteminin genişletilmesi• Sızıntı tespit programları	• Emtia ücretleri olan sayaçlar • Su dağıtımlarının evrensel ölçümü • Suyun marjinal fiyatının faturalarda belirgin şekilde gösterilmesi • Su fiyatındaki artış • En yoğun talep dönemleri için daha yüksek su fiyatları • Su için artan kademe oranları • Su bütçeleri • Ekonomik açıdan çekici hale getirmek için geri kazanılan suyun sübvansiyonu	• Kişisel etkileşim içeren herhangi bir mühendislik çözümü • Yüksek su kullanımı olan müşterilere yönelik su kullanımı denetimleri

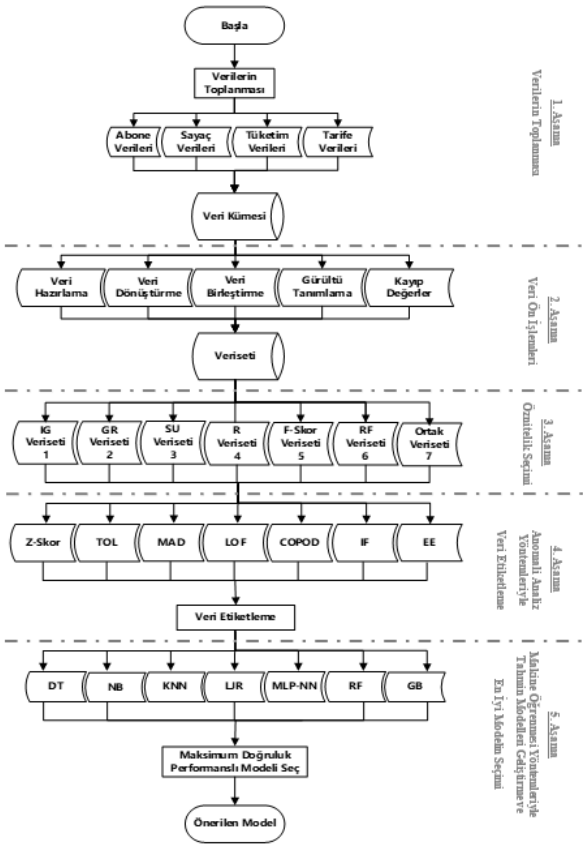
Çalışma ile birlikte gerek sular idarelerinin gerekse abonelerin kendilerine sorabilecekleri hızla artan kent nüfusu ve su ihtiyacı nasıl karşılanabilir, giderek artan bir şekilde hissedilen su kıtlığı ve çevresel bozulma ile karşı karşıya kalınan bir dünyada adil su hizmetleri sağlanabilir mi, içme ve kullanma su tüketimini etkileyen faktörler nelerdir, su tüketiminde bir düzen bir davranış kalıbı var mıdır, varsa bu bir mevsimsellik taşıyor mudur, tüketim düzenini bozan dönemler var mıdır, tüketim düzenini bozan aykırı tüketimler ile sayaç okuma hatası, kayıp kaçak tüketim, sayaç ölçüm arızası, tüketimde israf vb. tespitler yapmak mümkün müdür, tüketim davranışı AI alt dalı olan ML algoritmaları ile modellenebilir mi, tüketim davranışı disipline edilebilir mi, tüketim üzerinden su tasarrufu ile kaynak arayışına ciddi bir katkı sağlanabilir mi sorularına cevaplar sunulmaya çalışılmıştır.

Çalışmanın amacı içmesuyu tüketim davranışlarını tespit etmek, alternatif su kaynağı arayışının aksine varolan kaynakları daha verimli kullanmaya katkı sağlamak, tüketimi etkileyen faktörlere göre aykırı tüketim davranış sınırları belirlemek ve bunları takip etmek, abonelerin israf etmeden su kullanımına dolaylı olarak da ekonomisine katkıda bulunmak, kurumun çok yüksek taleplerde yaşayacağı zorlukların üstesinden gelmesine katkıda bulunmak, kayıp kaçak kullanım potansiyeli olan aboneleri tespit etmek ve takip edilmesini sağlamaktır.

Bu çalışma kapsamında Kayseri ili içme ve kullanma su abonelerinin hane içi su tüketimleri dikkate alınmış, dünyada ülkemizde su tasarrufuna, bu konularda toplumsal bilincin arttırılmasına ve su kayıp kaçak denetimlerine yönelik yapılan ML çalışmalarına katkıda bulunacak bir model geliştirilmiş, bununla birlikte sadece bireysel davranışları etkilemeye yönelik geri dönütler ile bir tasarrufun sağlanması, beraberinde su yöneticilerinin yatırım planlamalarına ve yönetim yaklaşımlarına farklı bir bakış açısı sunacak analizlerin yapılması ve bulguların paylaşılması hedeflenmiştir.

2. Yöntem (Method)

Çalışma Şekil 2'de gösterildiği gibi beş aşamada gerçekleştirilmiş olup ilk aşamada ihtiyaç duyulacağı öngörülen ham veriler temin edilerek bir veri kümesi elde edilmiş akabinde veri kümesi veri ön işlemlerinden geçirilerek analiz için uygun bir zaman serisi veriseti elde edilmiştir. Üçüncü aşamada öznitelik seçim yöntemleriyle alt verisetler türetilmiş ardından gözetimsiz ML ve istatistiki teknikler kullanılarak anormal tüketim tespitleri ve tüketim sınıf etiketlemesi yapılmıştır. Son aşamada ise etiketlenmiş gözetimli hale getirilmiş veriseti ile ML teknikleri kullanılarak tüketim sınıfı tahmin modelleri geliştirilmiş ve model performansları karşılaştırılarak en iyi model seçimi yapılmıştır.



Şekil 2. Çalışma kapsamında önerilen model şeması. (Model scheme proposed in the study.)

2.1. Çalışma alanı ve verilerin toplanması (Workspace and data collection)

Çalışma kapsamında Kayseri il genelinde okuması yapılan toplam 668.823 aboneden 610.821 adet konut abonesine ait 1980 – 2022 (ilk 6 ay dahil) tarihsel dönem aralığında abone, sayaç okuma ve tarife verilerinden oluşan bir veriseti kullanılmıştır. Çalışmada Tablo 4'te gösterildiği gibi çok fazla sayıda abone analizinin yapılmasının çalışmaya ayrıca bir katkı sunmayacağı ve çok fazla zaman ve işleme neden olacağı için okuma dönem adet sayısı 160 ve üzeri olan 8.224 adet konut abonesinin verileri incelenmiştir.

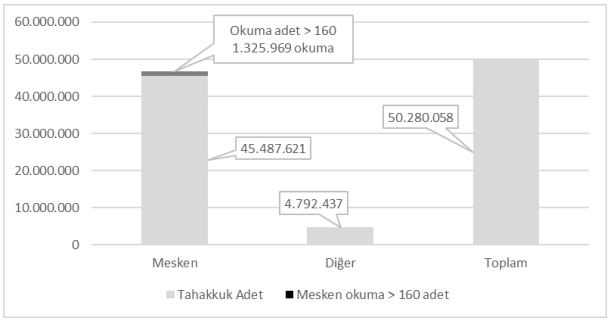
Tablo 4. KASKİ okuma dönem sayılarına göre abone sayısı tablosu (Number of subscribers according to KASKİ reading period numbers.)

Dönem Adet	Abone Sayısı	Mesken Sayısı
120'den az	454.574	409.292
120 ve üzeri	214.249	201.529
140 ve üzeri	140.491	133.914
160 ve üzeri	8.286	8.224

Literatürdeki çalışmaların çoğu, abonelerin tamamı üzerinden kümülatif talepleri dikkate almış, elde edilen modeller değerlendirilirken bütün abonelerin ortak davranışı varsayımıyla hareket edilmiş ve dolayısıyla il geneli hesaplanan hane başına düşen ortalama TÜFE ve nüfus artış/azalış oranları da dikkate alınmıştır. Bu çalışmada ise her bir aboneye özel aykırı tüketim tespit modeli oluşturulduğundan ekonomik göstergeler ve nüfusa ilişkin istatistiki verilerin her haneye özel, dikkate alınması gerekmektedir. Ancak aboneye özel istatistiki değerlerin temini ve sürekli güncel tutulması birçok kurumda tutulan verilerin (maaş, harcama, hanede yaşayanların doğum, ölüm, evlenme, evcil hayvan, bitki vb.) online entegrasyonu ile veya su yönetimi idareleri tarafından ABYS'de haneye özel veri alanları tanımlanıp bilgilerin sahadan personel marifetiyle toplanması ile mümkün olacaktır. Söz konusu bu durumda abonenin rızası, sular idarelerinin ihtiyaç öncelikleri ve bazı yasal düzenlemeler dikkate alındığında mümkün olmadığı ancak uzun vadeli teknik ve teknolojik gelişmeler ile birlikte (e-Devlet vb.) elde edilebileceği öngörülmüş, haneye özel ekonomik göstergeler ve nüfusa ilişkin istatistiki veriler girdi değerlerine dahil edilmemiştir.

Ayrıca sayaç okuması yapılan bölgelerin iklimsel verileri okumacı veya sabit meteoroloji istasyonları marifetiyle tespit edilse dahi hane içi iklim şartlarının da haneden haneye farklılıklar taşıyacağı gerçeğinden hareketle iklim verileri girdi değeri olarak dahil edilmemiştir.

Analiz için temin edilen 2022 yılı 6. Ay sonu itibariyle toplam 50.280.058 adet okuma verisinden, en az 160 dönem sayaç okuma gören mesken abonelere ait Şekil 2.6'da da gösterildiği gibi 1.325.969 adet tüketim verisi değerlendirilmiştir.



Şekil 3. Mesken ve diğer abone türlerine göre tahakkuk sayıları grafiği. (Graph of accrual numbers by residence and other subscriber types.)

Çalışma ile birlikte hanelere özel bazı istatistiki veriler kullanılamamış olsa da çalışma sonunda elde edilen modele KASKİ veya aboneler tarafından tüketimi etkilediği düşünülen özel alanların da sonradan eklenebileceği modüler bir yapı sunulmuştur.

2.2. Veri ön işlemleri ve verisetinin hazırlanması (Data preprocessing and dataset preparation)

Veri Hazırlama

Çalışma konusu analiz için ihtiyaç duyulan sayaç okuma, abone ve tarife verileri KASKİ'den csv formatında temin edilmiştir.

Çalışmada içme suyu abonelerinin tüketim miktarı çıktı değişkeni olarak belirlenmiş, hedef entropi H(Y) tüketim değerleri kullanılarak hesaplanmıştır. Tüketimi etkileyen özniteliklerin temel istatistiki değerler Tablo 5'te gösterildiği gibi hesaplanmıştır.

Tablo 5. Veriseti öznitelik seçiminde kullanılan temel istatistiki değerler tablosu (Table of base statistic values used in dataset feature selection.)

Öznitelikler	Min	Max	Ortalama - μ -	Ortanca Medyan	Standart Sapma - σ -	Varyans	Hedef Entropi H(Y)	Girdi Entropi H(X)
gecen_gun	3	93	35,49	35	10,19	103,84	4,24	1,98
fatura	4,35	165,37	35,85	29,24	26,85	720,72	4,24	4,14
odeme	4,35	165,37	35,93	29,24	26,88	722,65	4,24	4,14
ocak	0	30	3,13	0	7,62	58,12	4,24	0,55
subat	0	27	2,90	0	6,51	42,43	4,24	0,64
mart	0	31	3,16	0	7,08	50,08	4,24	0,71
nisan	0	30	2,96	0	7,63	58,19	4,24	0,61
mayıs	0	31	3,06	0	7,99	63,77	4,24	0,52
haziran	0	26	2,91	0	6,42	41,23	4,24	0,78
temmuz	0	31	2,97	0	6,91	47,73	4,24	0,71
agustos	0	31	2,97	0	7,22	52,10	4,24	0,76
eylul	0	30	2,78	0	7,38	54,47	4,24	0,59
ekim	0	27	2,90	0	6,70	44,94	4,24	0,71
kasım	0	30	2,77	0	7,06	49,91	4,24	0,60
aralik	0	31	2,98	0	7,83	61,31	4,24	0,54
kademe	1	2	1,01	1	0,08	0,01	4,24	0,04
K1	0	1	0,69	1	0,46	0,21	4,24	0,17
K2	0	1	0,10	0	0,29	0,09	4,24	0,12
K3	0	1	0,21	0	0,41	0,17	4,24	0,25
S1	0	1	0,08	0	0,27	0,07	4,24	0,15
S2	0	1	0,59	1	0,49	0,24	4,24	0,17
S3	0	1	0,33	0	0,47	0,22	4,24	0,15

Veri Dönüştürme

Sayaç okuma dönemlerinde biriken endekslerin bir sonraki okuma döneminde kayıt altına alındığı, dolayısıyla bazı dönemsel tüketimlerin olduğundan fazla görülebileceği, bu durumun analiz sonuçlarını yanıltabileceği öngörülmüş ve öncelikle Denklem 1'de gösterildiği gibi tüketim miktarı özniteliği oluşturulmuştur.

su_toplam_m3= last_index-first_index (1)

Sayaç okuma tarihi ile önceki okuma tarihi farkından da Denklem 2'de gösterildiği gibi tüketim süresi hesaplanmış yeni bir öznitelik olarak eklenmiştir.

gecen_gun = reading_date -pre_reading_date (2)

Hesaplanan tüketim gün sayısının yılın hangi ayından kaç günü kapsadığı bilgisine ulaşmak için yılın her ayı için 12 yeni öznitelik oluşturulmuş, hangi ayda kaç günlük bir tüketim olduğu hesaplanarak ilgili aya işlenmiştir. Böylece tüketim yapılan günler ilgili ayın ağırlık çarpanları olarak kullanılmıştır.

Her bir aboneye ait sayaç endeks değerlerinden farklı sayaç kullanımları tespit edilmiş, tüketimlerinin kaçıncı sayaç ile yapıldığı bilgisini gösteren yeni bir öznitelik ("sayac_durum") Tablo 6'da gösterildiği gibi geliştirilmiştir.

Tablo 6. Verisetindeki endeks hareketleri üzerinden sayaç bilgisi dönüşüm tablosu (Counter conversion table over index movements in the dataset.)

Okuma Tarihi	İlk End.	Son End.	Sayac durum	S 1	S 2	S 3
22.02.2006	262	268	S1	1	0	0
---	---	---	---	-	-	-
14.11.2007	0	3	S2	0	1	0
---	---	---	---	-	-	-
6.06.2022	17	29	S3	0	0	1

Her abonenin tüketim gözlem noktalarındaki sözleşme numarası bilgisinden hareketle tüketimin farklı bir kullanıcı tarafından yapıldığı bilgisine ulaşılmış yeni bir öznitelik ("kullanici_durum") Tablo 7'de gösterildiği gibi geliştirilmiştir.

Tablo 7. Verisetindeki sözleşme numarası üzerinden kullanıcı bilgisi dönüşüm tablosu (User conversion table over agreement number in dataset.)

Okuma Tarihi	Sözleşme No	Kullanıcı Durum	K 1	K 2	K 3
22.02.2006	12001815	K1	1	0	0
---	---	---	-	-	-
17.05.2018	357910	K2	0	1	0
---	---	---	-	-	-
6.06.2022	402841	K3	0	0	1

Verisetinde her bir okuma dönemine ait 5 kademeli bir faturalandırma bilgisi (bill_lv1, bill_lv2, bill_lv3, bill_lv4, bill_lv5) mevcut olduğundan, bu bilgiler kullanılarak kademe durumunu gösteren bir öznitelik ("kademe") Tablo 8'de gösterildiği gibi geliştirilmiş kademe tutar bilgi öznitelikleri verisetinden çıkarılmıştır.

Tablo 8. Verisetindeki kademe tutar bilgi alanları ve dönüşüm tablosu (Price level amount information fields and conversion table in the dataset.)

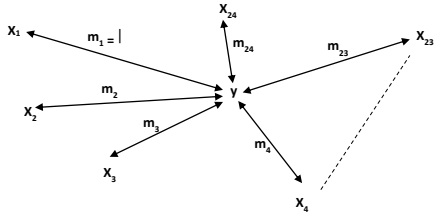
Okuma Tarihi	bill lv1	bill lv2	bill lv3	bill lv4	bill lv5	kademe
22.02.2006	6,67	0,00	0,00	0,00	0,00	1
---	-	-	-	-	-	---
31.03.2022	13,82	16,38	0,00	0,00	0,00	2

Ayrıca verisetini oluşturan öznitelik değerlerinin gün, ay, yıl, ₺, m3 gibi farklı birimlerden oluşması dikkate alınarak, grafik gösterim ve performans hesaplamalarının daha sağlıklı yapılabilmesi için her bir özniteliğe ait değerler Denklem 3'te gösterildiği gibi normalize işlemi uygulanarak 0 ile 1 aralığına indirgenmiştir.

$$X_n = \frac{(X_g - X_{min})}{(X_{max} - X_{min})} \quad (3)$$

Burada Xn normalize edilmiş gözlem değeri, Xg gözlemin gerçek değeri ve X_{max}, X_{min}'de sırasıyla ilgili gözlemin en büyük ve en küçük değerleridir.

Ayrıca çalışmada mesafe ölçüm yöntemi kullanılarak ta analiz yapılmıştır. Mesafe ölçüm yöntemlerinden girdi öznitelik sayıları ikiden fazla olduğu için Şekil 4'te de gösterildiği gibi minkowski yöntemi kullanılmış, gözlem noktaları mesafe değerleri hesaplanmıştır.



Şekil 4. Bağımsız değişkenlerin bağımlı değişkene mesafelerinin minkowski ile ölçüm gösterimi. (Measurement representation of the distances of the independent variables to the dependent variable with the minkowski method.)

Bağımsız değişkenler olan X_i girdi değerlerinin, bağımlı değişken olan y değerine mesafeleri m_i olarak alınmış olup Denklem 4'te gösterildiği şekilde her bir gözlem noktasına yani sayaç okuma noktası özniteliklerine ait değerler tek bir mesafe değerine dönüştürülmüştür.

$$Minkowski = d(X, y) = [\sum_{i=1}^{n}(|X_i - y|^p)]^{\frac{1}{p}} \quad (4)$$

Burada p girdi öznitelik yani boyut sayısı, y çıktı değeri ve X_i'de her bir girdi öznitelik değeridir. Ayrıca negatif korelasyona sahip özniteliklerin X değerleri -X olarak alınmış, her bir tüketim gözlem noktası mesafesi toplam mesafe olarak hesaplanmıştır.

Veri Birleştirme

Çalışmada MS Access veritabanı kullanılarak csv formatında temin edilen 15 öznitelikli abone, 21 öznitelikli tüketim ve 5 öznitelikli tarife verileri, tablolar arası ilişkisel yapı dikkate alınarak birleştirilmiş, her bir abone için bütün özniteliklerin yer aldığı 41 öznitelikli TS bir veriseti elde edilmiştir.

Gürültü Tanımlama

Çalışma, hanelerin tüketimlerine odaklandığı için öncelikle içme suyu abonelerinden sadece mesken abonelerine ait veriler seçilmiş, ayrıca en az 160 okuma dönemi okuma yapılan aboneler dikkate alınmış diğer abonelere ait veriler çıkarılmıştır. Örneklem kapsamında belirlenen abonelere ait 41 öznitelikli verisetindeki mükerrer öznitelikler çıkarılmış veriseti nihayetinde 33 özniteliğe indirgenmiştir. Ayrıca bütün gözlem değerleri aynı olan öznitelikler de öznitelik seçim yöntemleri kullanılarak analize dahil edilmemiştir.

Kayıp Değerler

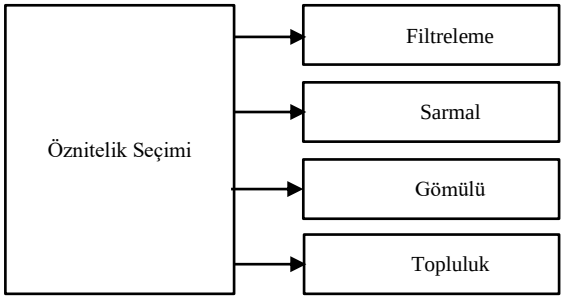
Geçmiş sistem kayıtlarında sayaç değişimlerine ait iş emirlerinin ve sayaç numara bilgilerinin sağlıklı olmaması nedeniyle, sayaç değişim bilgileri, okuma endeks hareketlerinden tespit edilerek elde edilmiştir.

Geçmiş dönemlere ait tarife bilgisine ulaşılamadığında ilgili tarihten bir önceki dönem birim fiyat bilgisi veya fatura tutarı, tüketim miktarına bölünerek elde edilen tutar bilgisi dikkate alınmıştır.

2.3. Öznitelik seçimi ve alt verisetlerin türetilmesi (Feature selection and derivation of subdatasets)

Öznitelik seçimi Şekil 5'te gösterildiği gibi filtreleme, sarmal, gömülü ve topluluk öznitelik seçim

yöntemleri olmak üzere 4 başlıkta incelenebilir (Cai, Luo, Wang ve Yang, 2018).



Şekil 5. Öznitelik seçme yöntemleri şeması. (Feature selection methods schema.)

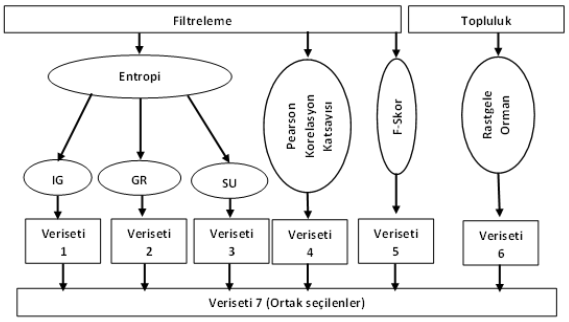
Filtreleme öznitelik seçim yaklaşımı ile veriseti istatistiksel olarak analiz edilir. Temel olarak verisetindeki bütün bilgiyi içeren benzersiz özniteliklerin öncelikle puanlanmasına sonrasında sıralanmasına dayalı olarak değerlendirilmesini ve seçilmesini sağlayan bir yöntemdir (Cai, Luo, Wang ve Yang, 2018) (Zhang, Nie, Li ve Wei, 2019).

Sarmal öznitelik seçim yaklaşımında, veriseti denetimli ise sınıflandırıcı algoritmalar kullanılarak, denetimsiz ise kümeleme algoritmaları ile öznitelikler üzerinde arama işlemi yapılarak alt kümeler oluşturulmakta ve belli sınıflar sadece değerlendirmeye alınmaktadır. Çok büyük ölçekli verisetlerinde bu yaklaşımın kullanılması durumunda her bir alt kümenin değerlendirilmesi fazla miktarda hesaplama işlemleri gerektirdiği için zaman ve kapasite dikkate alınarak tercih edilmesi gerekmektedir (Cai, Luo, Wang ve Yang, 2018) (Zhang, Nie, Li ve Wei, 2019).

Gömülü öznitelik seçim yönteminde model geliştirme sürecinde harcanan zamanın azaltılması amaçlanmakta olup hem sınıflandırma hem de seçim işlemini eşzamanlı olarak model eğitimi aşamasında gerçekleştiren DT, SVM gibi algoritmalar kullanılmaktadır (Cai, Luo, Wang ve Yang, 2018) (Zhang, Nie, Li ve Wei, 2019).

Topluluk öznitelik seçim yönteminde farklı öznitelik alt kümeleri elde edilerek oluşturulan verisetleri tarafından eğitilen temel sınıflandırma algoritmaları kullanılarak yapılmış öznitelik seçim topluluklarının rastgele alt uzay yöntemi (RSM) ve RF yöntemi gibi çoğunluk oylamasıyla daha doğru sınıflandırma ve seçim yapılmasına dayanmaktadır (Cai, Luo, Wang ve Yang, 2018).

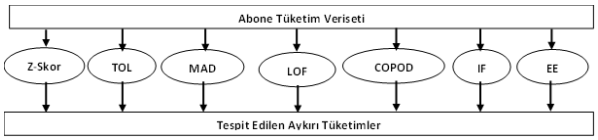
Bu çalışmada öznitelik seçimi yöntemlerinden Şekil 6'da gösterildiği gibi IG, GR, SU, R, F-Skor ve RF sınıflandırma yöntemleri kullanılmıştır.



Şekil 6. Öznitelik seçme yöntemleri şeması. (Feature selection methods schema.)

2.3. Anomali analiz yöntemleriyle veri etiketleme (Data labeling with anomaly analysis methods)

Çalışmada anormal içme suyu tüketim tespitleri, Şekil 7'de gösterildiği gibi 7 farklı anomali analiz yöntemi kullanılarak yapılmış, abonelerin içme suyu tüketimlerini sınıflandırmak ve etiketlemek için yeterli olmuştur. Daha spesifik sınıflandırma ve etiketleme için diğer ML sınıflandırma tekniklerinden de istifade edilebilir.



Şekil 7. Topluluk öğrenme yöntemi ile aykırı tüketim tespit şeması. (Outlier consumption detection scheme with ensemble learning method.)

Aykırı tüketim tespiti yapan her bir yöntemin çalışma şeması Şekil 2.16'da gösterilmiş olup, devamında bütün yöntemlerden elde edilen sonuçlar aykırılık puanlarını elde etmek için değerlendirilmiştir.

Veri Etiketleme

Çalışmada tüketim tespit yöntemlerinin sonuçlarına göre aykırılık puan hesabı yapılmış, Tablo 9'da gösterildiği şekilde 4 farklı tüketim sınıfı ve 4 farklı renk ile etiketlenmiştir.

Tablo 9. Aykırılık puanlarına göre sınıflandırma etiketleri tablosu. (Table of classification labels by outlier scores.)

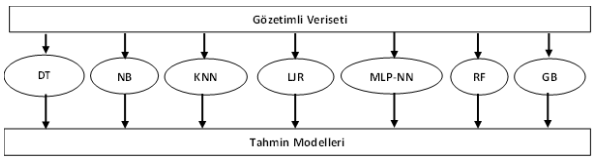
Aykırılık Puanı	Tüketim Sınıfı
0-1	Normal - Yeşil
2-3	Dikkat - Sarı
4-5	Riskli - Turuncu
6-7	Aşırı - Kırmızı

2.4. ML yöntemleriyle tüketim sınıfı tahmin modelleri geliştirme (Developing consumption class prediction models with ML methods.)

Çalışma sonunda tüketim sınıf etiketlemesi ile gözetimli hale getirilmiş veriseti, Şekil 8'de gösterildiği

gibi DT, NB, KNN, LJR, MLP-NN, RF ve GB olmak üzere 7 farklı ML sınıflandırma algoritmaları kullanılarak içme suyu tüketim sınıfı tahmin modelleri geliştirilmiştir.

Tüketim verileri TS verilerden oluştuğu için ilk yıllar eğitim, son yıllar da kontrol verisi olmak üzere %90'a, %10 oranlarında bölünerek ve bölünmeden kullanılmıştır.



Şekil 8. İçmesuyu tüketim sınıfı tahmin modelleri şeması. (Diagram of drinking water consumption class prediction models.)

2.4. Performans ölçümü ve en iyi modelin seçimi (Performance measurement and selection of the best model.)

Çalışmada bütün anomali tespit modellerinin öncelikle TP, TN, FP, FN değerlerinden oluşan hata matrisleri hesaplanmış, Doğruluk, Duyarlılık, Kesinlik, Hassasiyet, Özgüllük, MAE ve MSE performans metrikleri kullanılarak modeller karşılaştırılmıştır.

Her bir abone için geliştirilen anormal tüketim tespit modellerinden en yüksek doğruluk oranındaki model en iyi model olarak belirlenmiştir. Anormal tüketim tespit modellerinden doğruluk oranı aynı olanlar var ise anomali duyarlılık ve anomali kesinlik oranlarına göre yüksek oranda olan en iyi model olarak belirlenmiştir.

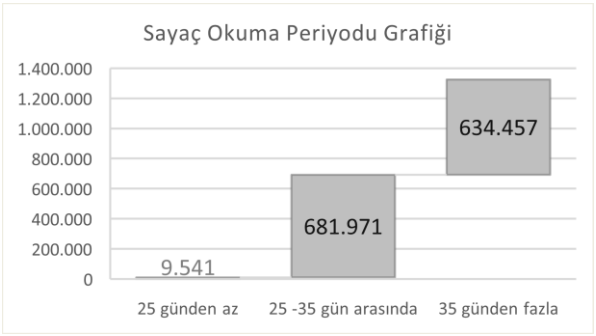
Tüketim sınıfı tahmin modelleri de Doğruluk ve determinasyon katsayısı (R^2) performans metrikleriyle karşılaştırılmış, her bir içme suyu abonesi için en iyi tahmin modeli seçilmiştir.

3. Bulgular (Findings)

3.1. Anomali analiz sonuçları (Anomaly analysis results.)

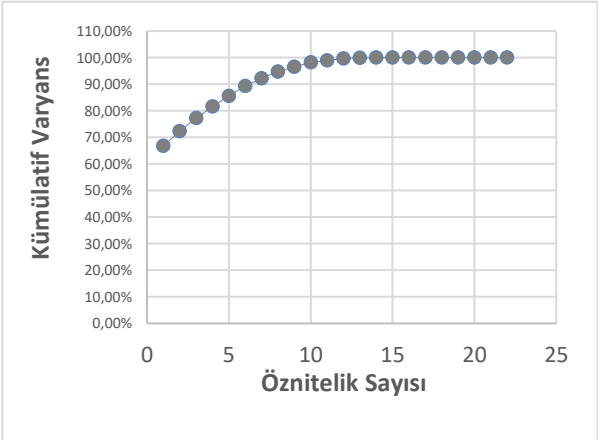
Veriseti incelendiğinde özetle 2006–2022 (ilk 6 ay) yılları arasında kış aylarında içmesuyu tüketimlerinin düştüğü, yaz aylarında tüketim değerlerinin diğer aylara göre daha yüksek olduğu, çalışma kapsamında analizi yapılan abonelere ait toplamda 17.908.783 m3 tüketim yapıldığı, 1 okuma döneminde konut/mesken su tüketim miktarı en yüksek 818 m3, en düşük 1 m3 ve ortalama 13 m3 olduğu, yine 1 okuma döneminin en geç 819 gün, en erken aynı gün ve ortalama 36 gün olduğu tespit edilmiştir.

Abonelere ait Şekil 9'da gösterilen sayaç okuma periyodu adetleri incelendiğinde 1.325.969 adet okumanın %0,72'sinin 25 günden az bir zamanda, %47,85'inin 35 günden fazla bir zamanda ve %51,43'ünün ise 25 ile 35 gün aralığında yapıldığı tespit edilmiştir.



Şekil 9. Sayaç okuma dönem grafiği (Meter reading period chart.)

Temel veriseti 24 öznitelikten, öznitelik seçim yöntemleri ile elde edilen alt verisetleri ise en az 7 en fazla 12 adet öznitelikten oluşmuştur. Burada kümülatif varyans %90 ve üzerinde olarak Şekil 10'da gösterildiği gibi öznitelik adetleri belirlenmiş, bir sonraki öznitelik kümülatif varyansı %95'in üzerine çıkarıyorsa alt veriseti öznitelik sayısının bir fazlası olarak belirlenmiştir.



Şekil 10. Kümülatif varyans oranına göre öznitelik sayıları grafiği (Graph of feature counts by cumulative variance ratio.)

Her bir aboneye ait gerçek değerlerden oluşan ve veri ön işlemleri ile elde edilen 24 özniteliğin tamamının olduğu 8.224 adet veriseti ve öznitelik seçim yöntemleri sonrasında elde edilen 57.568 alt veriseti olmak üzere toplamda 65.792 adet veriseti analizde kullanılmıştır. Ayrıca gerçek değerler minkowski mesafe ölçüm yöntemiyle birleştirilerek ve dönüştürülerek 65.792 adet veriseti daha elde edilmiş, toplamda 131.584 adet alt veriseti kullanılmıştır.

Alt verisetler 6 farklı öznitelik seçim algoritması ile belirlenmiş olup, öznitelik seçme algoritmalarının Tablo 10'da gösterildiği gibi aynı öznitelikleri seçme sayıları

ve oranları gösterilmiştir. Öznitelik seçme yöntemlerinden IG ve SU yöntemlerinin ortak olarak belirlenen özniteliklere en yakın oranlarda öznitelik seçimi yaptıkları anlaşılmıştır. Sırasıyla RF ve GR yöntemlerinin çok az da olsa aynı öznitelikleri seçtikleri, korelasyon ve f skor yöntemleri kullanılarak yapılan öznitelik seçimlerinin diğer yöntemlere göre neredeyse hiç benzerlik göstermediği, birbirleriyle de çok az bir benzerlik taşıdıkları anlaşılmıştır.

Tablo 10. Öznitelik seçim yöntemleriyle aynı özniteliklerin seçilme sayıları ve oranları (The number and rate of selection of the same features by feature selection methods)

Öznitelik Seçim Yöntemleri				Bilgi Kazancı	Kazanç Oranı	Simetrik Belirsizlik Katsayısı	Korelasyon	F Skor	Rastgele Orman	Ortak
				IG	GR	SU	R	F	RF	
				V1	V2	V3	V4	V5	V6	V7
Bilgi Kazancı	**IG**	**V1**	Adet		495	7.066	84	0	1.018	3.669
			Oran		%6	%86	%1	%0	%12	%45
Kazanç Oranı	**GR**	**V2**	Adet			587	114	0	340	345
			Oran			%7	%1	%0	%4	%4
Simetrik Belirsizlik Katsayısı	**SU**	**V3**	Adet				85	0	972	3.641
			Oran				%1	%0	%12	%44
Korelasyon	**R**	**V4**	Adet					11	51	44
			Oran					%0	%1	%1
F Skor	**F**	**V5**	Adet						0	0
			Oran						%0	%0
Rastgele Orman	**RF**	**V6**	Adet							707
			Oran							%9

Çalışmada elde edilen alt verisetlerinde Tablo 11'de de özniteliklerin seçilme sayısı ve oranları gösterilmiştir.

Tablo 11. Alt verisetlerinde öznitelik seçim yöntemlerinin seçim adetleri ve oranları tablosu (Table of selection numbers and rates of feature selection methods in subdatasets.)

Öznitelik	Alt Verisetlerinde	Seçilme Oranı
Haziran	33.406	%23,47
Temmuz	32.408	%22,77
Ekim	30.908	%21,72
Subat	30.665	%21,55
Eylul	30.451	%21,40
Mart	29.798	%20,94
Kasim	28.814	%20,25
Aralik	26.991	%18,96
Mayis	26.172	%18,39
Ocak	24.735	%17,38
Agustos	23.464	%16,49
Nisan	22.197	%15,60
S1	14.732	%10,35
Odeme	13.248	%9,31
Kademe	10.852	%7,62
K1	10.150	%7,13
S2	9.856	%6,93
S3	1.879	%1,32
K2	1.428	%1,00
K3	148	%0,10
S4	108	%0,08
K4	16	%0,01
S5	7	%0,00
K5	2	%0,00

İkinci aşamada hem gerçek hem de minkowski mesafe değerleriyle elde edilen alt verisetler, çalışma kapsamında belirlenen 7 farklı anomali analiz algoritması ile analiz edilmiş, anormal tüketim tespitleri, anomali puanlamaları ve içmesuyu tüketim sınıflandırmaları yapılmıştır.

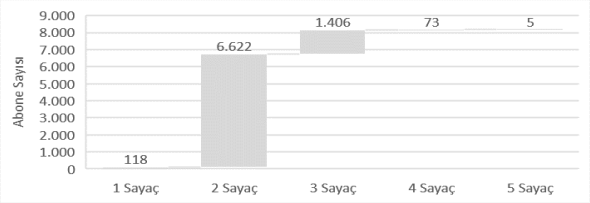
Toplam 1.325.969 adet içme suyu tüketim gerçek değerleri kullanılarak yapılan analizde, 1.196.181 adet normal tüketim, 129.788 adet de anormal tüketim tespit edilmiş olup, bütün tüketimlerde %9,79 oranında bir anomali durum olduğu anlaşılmıştır. Anomali tespit yöntemlerindeki eşik değeri ile bu oran yukarı veya aşağı yönlü değiştirilebilir. Anormal tüketimler de kendi içinde "Dikkat" etiketli tüketim sayısı 62.239 (anormal tüketimlere göre %47,95, bütün tüketimlere göre %4,69), "Riskli" etiketli tüketim sayısı 30.418 (anormal tüketimlere göre %23,44, bütün tüketimlere göre %2,29), "Aşırı" etiketli tüketim sayısı 37.131 (anormal tüketimlere göre %28,61 bütün tüketimlere göre %2,80) adet olarak tespit edilmiştir.

Toplam 1.325.969 adet içme suyu tüketiminin minkowski mesafe değerlerine dönüştürülerek yapılan analizinde, 1.152.927 adet normal tüketim, 173.042 adet de anormal tüketim olarak tespit edilmiş olup %13,05 oranında bir anomali durum söz konusu olduğu anlaşılmıştır. Ayrıca anormal tüketimlerde kendi içinde "Dikkat" etiketli tüketim sayısı 93.177 (anormal tüketimlere göre %53,85 bütün tüketimlere göre %7,03), "Riskli" etiketli tüketim sayısı 50.856 (anormal tüketimlere göre %29,39, bütün tüketimlere göre %3,84), "Aşırı" etiketli tüketim sayısı 29.009 (anormal

tüketimlere göre %16,76, bütün tüketimlere göre %2,19) adet olarak tespit edilmiştir.

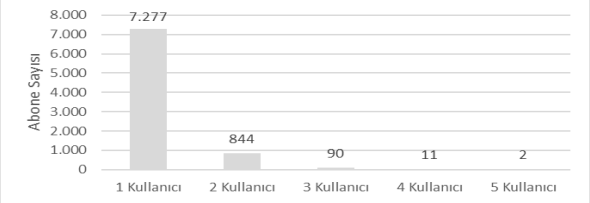
Çalışmada gerçek değerler ve minkowski mesafe değerleri ile yapılan anormal tüketim tespitlerinde, aynı anomali tespit yöntemleri ve eşik değerleri kullanılmasına rağmen farklı sonuçlar elde edilmiştir. Minkowski değerleriyle toplamda %3,26'lık bir oranda daha fazla anormal tüketim tespiti olmasına rağmen, "Aşırı" tüketim sınıf tespitinde %0,61 oranında bir düşüşün olması minkowski değerleri kullanılarak elde edilen tespitlerin gerçek değerlerle tespit edilenlerden daha iyi sonuçlar verdiği şeklinde yorumlanmıştır.

Ayrıca abonelerin en fazla 167 en az 160 dönem sayaç okuması yapılmış olup Şekil 11'de gösterildiği gibi %80,52'sinin tüketim süresince 2. sayacı kullandıkları tespit edilmiştir.



Şekil 11. Abonelerin tüketim süresince kullandıkları sayaç durum grafiği (Counter status graph used by subscribers during consumption)

Ayrıca içme suyu abonelerinin Şekil 12'de de gösterildiği gibi %88,48'inin tüketimlerinin aynı kullanıcı tarafından yapıldığı tespit edilmiştir.



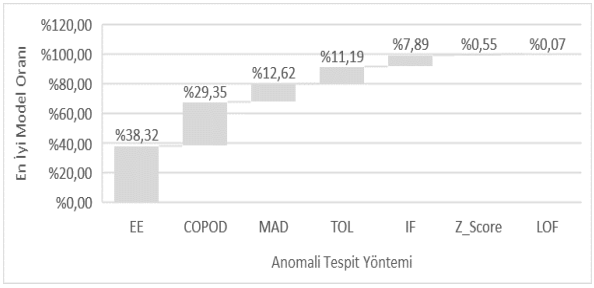
Şekil 12. Aboneliklere tüketim süresince sözleşme yapan kullanıcı durum grafiği (User status graph contracting subscriptions during consumption)

Abone tüketimlerinin tüketim süresince Tablo 12'de gösterildiği gibi yalnızca %0,24'ünün kademeye girdiği, %0,55'inin de ödenmediği tespit edilmiştir.

Tablo 12. Sayaç, kullanıcı, kademe ve ödeme durumları özet tablosu. (Counter, user, level and payment status summary table.)

	Kademesiz	Kademeli	Ödenen	Ödenmeyen	Toplam
Abone Sayısı	6.833	1.391	8.224	3.189	8.224
Okuma Sayısı	1.322.781	3.188	1.318.707	7.262	1.325.969

Çalışmada her bir abone için toplamda 112 adet anormal tüketim tespit modeli geliştirilmiş; doğruluk, duyarlılık, kesinlik, hassasiyet, özgüllük, MAE ve MSE performans metrikleriyle modeller karşılaştırılmıştır. Toplamda 921.088 adet model geliştirilmiş, bunlardan 16.077 adeti en iyi model olarak maksimum doğruluk performansı göstermiştir. Modellerin performans doğruluk oranları en düşük %43, en yüksek %100 ve ortalama %85 olarak tespit edilmiş olup Şekil 13'te gösterildiği gibi EE yöntemiyle elde edilen modellerin çoğunlukla en iyi modeller olduğu, Z Skor ve LOF yöntemleri ile elde edilen modellerin çok düşük oranlarda maksimum doğruluk sağladığı anlaşılmıştır.



Şekil 13. Maksimum doğruluk sağlayan anomali analiz yöntemlerinin oranları grafiği (Graph of rates of anomaly analysis methods that provide maximum accuracy)

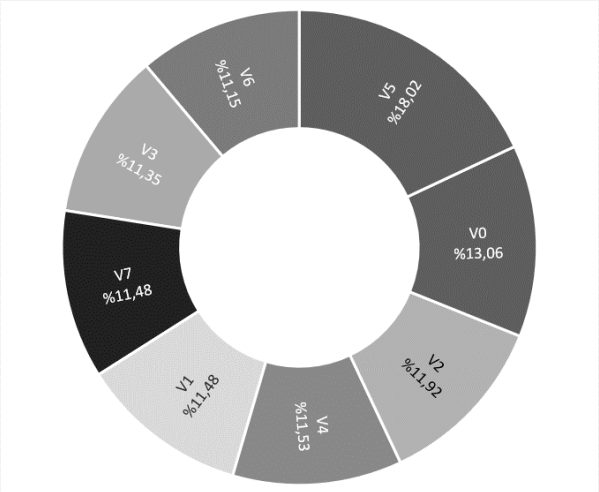
Her bir abone için geliştirilen anormal tüketim tespit modellerinden doğruluk performans metriği %90 ve üzeri olanlar Tablo 13'te de gösterildiği gibi sıralanır ve doğruluk oranı en yüksek olanı veya kullanılan anomali analiz yöntemine göre veya kullanılan alt verisetine göre idare tarafından belirlenebilir. Her bir abone için ayrı ayrı olmak üzere belirlenen en iyi anormal tüketim tespit modeli ile bir sonraki okuma döneminde anormal tüketim analizinin yapılması daha az işlem ile ve daha kısa sürede mümkün olacaktır.

Tablo 13. Her bir aboneye ait doğruluk performans metriği maksimum olan en iyi model tablosu (Top model table with maximum accuracy performance metric for each subscriber)

No	Abone No	En iyi Model	Doğruluk
1	12000541	c_V5_COPOD	0,993
2	12000950	g_V1_EE	0,993
3	10003387	c_V4_COPOD	0,993
4	10004150	c_V2_IF	0,993
5	4002847	c_V7_COPOD	0,993
…	…	…	…
8220	12001129	g_V3_IF	0,975
8221	4011157	c_V6_Z_score	0,962
8222	6009693	g_V1_EE	0,950
8223	216960	c_V2_TOL	0,900
8224	217139	c_V4_TOL	0,900

Çalışmada elde edilen bütün tüketim sınıfı tahmin modellerinden, doğruluk performans metriği %90 ve üzeri olanlar dikkate alındığında, Şekil 12'de de gösterildiği üzere örneklem olarak alınan 8.224

abonenin 7.150 tanesi (%86,94) alt verisetlerinden (V1, V2, V3, V4, V5, V6, V7), diğer 1.074 adet abonenin (%13,06) ise temel verisetinden (V0) elde edilebilmektedir. Her bir aboneye ait geliştirilen maksimum doğruluk performanslı model en fazla oranda f skor öznitelik seçim yöntemi ile elde edilen V5 alt veriseti kullanılarak elde edilmiştir. Diğer alt verisetleri ile de birbirine yakın oranlarda başarılı sonuçlar elde edildiği anlaşılmıştır.



Şekil 12. Doğruluk performansı %90 ve üzeri olan tahmin modellerinde kullanılan veriseti oranları grafiği (Graph of dataset ratios used in prediction models with an accuracy performance of 90% or more)

Burada işlem sayısı ve süresi azaltılmak istendiğinde her bir abone için alt verisetlerinden en az öznitelikli olanı tercih edilebilir.

Model geliştirme sürecinde işlemci hızı 1.1 GHz, RAM kapasitesi 4GB olan bir PC kullanılarak her bir abone için model ön işlemleri ortalama 10 sn, anomali analizleri ortalama 8 sn ve performans ölçüm ve karşılaştırma işlemleri ortalama 3 sn olmak üzere toplamda yaklaşık 96 saat sürmüştür. Çalışma kapsamında analizi yapılan 8.224 abonenin belirlenmiş en iyi modeli kullanılarak anomali tespiti yapılmasında bu süre yaklaşık olarak 51 dk sürmektedir. Böylece önerilen yöntem ile 4 günlük bir analiz ortalama 1 saatten az bir zamanda yapılabilmektedir. Genellikle bütün sular idarelerinde mevcut olan profesyonel donanım altyapısının kullanılması ile çalışma kapsamında önerilen analiz işlem sürelerinin daha da kısaltılması mümkün olacaktır.

Çalışmanın temel amacı olan abonelerin duyarlı hale getirilmesi ile tüketim davranışlarında oluşan anormal tüketimlerinin azaltılması, tüketimlerin normal seviyelere çekilmesi durumunda, tespit edilen 'Asiri' sınıftaki tüketimlerde %20 tasarruf ile 139.427 m3, 'Riskli' sınıfındaki tüketimlerden %15 tasarruf ile 154.058 m3 ve 'Dikkat' sınıfındaki tüketimlerden de %10 tasarruf ile 169.402 m3 olmak üzere toplamda sistemdeki %2,58'lik 462.887 m3 suyun daha verimli kullanımının sağlanacağı öngörülmektedir.

Abone faturalarında ise 'Asiri' sınıftaki tüketimlerde %20 tasarrruf ile 284.304,37 ₺, 'Riskli' sınıftaki tüketimlerden %15 tasarrruf ile 288.181,09 ₺ ve 'Dikkat' sınıftaki tüketimlerden de %10 tasarruf ile 349.438,92 ₺ olmak üzere içme suyu fatura toplamlarında %5,15'lik 921.924,38 ₺ tasarrrufun sağlanacağı öngörülmektedir.

3.2. ML analiz sonuçları (ML analysis results.)

Çalışma sonunda tüketim sınıf etiketlemesi ile gözetimli hale getirilmiş veriseti, DT, NB, KNN, LJR, MLP-NN, RF ve GB olmak üzere 7 farklı ML sınıflandırma algoritmaları kullanılarak içme suyu tüketim sınıfı tahmin modelleri geliştirilmiştir. Tüketim verileri TS verilerden oluştuğu için ilk yıllar eğitim, son yıllar da kontrol verisi olmak üzere %90'a, %10 oranlarında bölünerek ve bölünmeden kullanılmıştır. Geliştirilen modeller Tablo 14'te gösterildiği gibi ACC ve R2 performans metrikleriyle karşılaştırılmış, her bir içme suyu abonesi için en iyi model seçilmiştir.

Tablo 14. İçme suyu tüketim sınıf tahmin modelleri performans tablosu (Performance table of drinking water consumption class prediction models)

	ML Model	Veriseti Bölünmeden				%90 Eğitim - %10 Test			
		Minkowski Değerler		Gerçek Değerler		Minkowski Değerler		Gerçek Değerler	
		ACC	R^2	ACC	R^2	ACC	R^2	ACC	R^2
1	DT	1,00	1,00	0,99	0,73	1,00	1,00	1,00	1,00
2	Gaussian NB	0,94	-0,36*	0,69	-2,88*	1,00	1,00	0,69	0,00
3	KNN	1,00	1,00	0,99	0,73	1,00	1,00	0,94	0,00
4	LJR	0,87	0,00	0,86	-0,27*	1,00	1,00	1,00	1,00
5	MLP-NN	0,94	-0,52*	0,89	-0,97*	1,00	1,00	1,00	1,00
6	RF	1,00	1,00	0,99	0,73	1,00	1,00	1,00	1,00
7	GB	1,00	1,00	0,99	0,73	1,00	1,00	1,00	1,00

*R^2 değerinin negatif olması modelin çok daha kötü olduğunu göstermektedir.

Minkowski mesafe değerleriyle elde edilen modeller, gerçek değerler kullanılarak geliştirilen modellerden daha iyi performans göstermiştir. Modeller, kurumun isteği doğrultusunda veya belli periyotlarda otomatik olarak yeni veriler de dikkate alınarak güncellenebilir ve böylece daha iyi performans sonuçları elde edilebilir.

4. Sonuçlar (Results)

Çalışma sonucunda abonelerin sayaç okuma dönemlerine ait abone, sayaç, sayaç okuma, tarife, tahakkuk ve tahsilat verilerinden tüketim davranış

modelleri geliştirilmiş ve tüketim sınıflandırma etiketlemesi yapılmıştır. Her bir aboneye özel olarak geliştirilen tahmin modeli ile her okuma döneminde sayaç okuma esnasında gerçekleşen tüketimin sınıfı belirlenmesi sağlanmıştır.

İçmesuyu abone tüketimlerinden anormal tüketimlerin tespit edilmesi ve tüketim sınıflarının belirlenmesi ile; tüketim alışkanlıklarının su israfına neden olacak düzeyde değişmiş olabileceği, iç tesisatta bir sızıntı olabileceği, kullanılan içme suyu sayacının ölçüm hassasiyetini kaybetmiş olabileceği, ölçüm dışı tüketim veya kaçak kullanımın söz konusu olabileceği, sayaç okuma memurunun endeksi yanlış girmiş veya sayacı yanlış okumuş olabileceği, sayaç endeks okuma döneminin ay aşırı geç veya daha kısa sürelerde yapılmış olabileceği, bahçe, peyzaj veya tarım amaçlı sulama yapılıyor olabileceği, hanenin boş veya tatile çıkmış olabileceği, hane halkı sayısının doğum, ölüm vb. nedenlerle artmış veya azalmış olabileceği, iklim şartlarının klima kullanımı vb. nedenler ile değişmiş olabileceği değerlendirmelerinin daha sağlıklı yapılabileceği öngörülmektedir.

Ayrıca çalışmada önerilen yöntemle sular idaresinin her okuma döneminde, anormal tüketim analiz raporu ve anormal içmesuyu tüketimi yapan abone listesi oluşturabileceği, sahada yapılan sayaç arıza, fiziki kayıp kaçak gibi içme suyu denetimlerinde bu listelerden istifade edebileceği öngörülmektedir. Böylece denetimlerde genelleme yapılmayarak daha özenli ve itinalı davranılacağı, anomali tespitinin kontrolünün ve geri dönüşlerinin de sağlanmış olacağı öngörülmektedir. Çalışma ile analizi yapılan abonelerin, yaklaşık 16 yıllık tüketim verilerinden, "Aşırı" sınıflı 27.463 adet tüketim dikkate alındığında yaklaşık olarak aylık 137 adet kontrolün gerektiği, günlük 25 adet kontrol yapan bir ekip ile 1 haftaya yakın bir zamanda kontrol edilmesinin mümkün olduğu öngörülmektedir. Bütün abonelerin çalışmada önerilen yönteme ait analizlerinin yapılmasıyla daha bütüncül bir kontrol listesi elde edilebilir, ekip sayısı ve anormal tüketim sayısı dikkate alınarak bir planlama yapılabilir. Ayrıca istenirse ekip sayısına ve ekiplerin kontrollerini yapabilecekleri anomali sayısına göre de anomali tespit yöntemlerinin eşik değerleri yukarı veya aşağı yönlü ayarlanarak planlama yapılabilir.

Abonelerin sayaç okumalarının yapılamaması veya okuma dönem sürelerinin ay aşırı günlerde veya daha kısa sürelerde yapılmasının önüne geçilerek, sayaç okumalarının her bir abone için kayıp dönem olmadan her ayın aynı gününde yapılmasının tüketim tahmin model performansını arttıracağı, ayrıca mevsimsel şartların dolayısıyla meteorolojik verilerin hane içindeki tüketimlere etkisi gerçeğinin tek başına yeterli olmayacağı, hanelerde klima gibi hane içindeki iklim şartlarını değiştiren sistemlerin varlığına ait bilgilerin de dikkate alınmasının faydalı olacağı öngörülmektedir.

Ayrıca önerilen yöntem kullanılarak, sayaçların içeride olması ve kullanıcıların evde olmaması durumlarında ya da sayaç arıza dönemlerinde yapılan kıyas tüketimlerin, her aboneye özel olarak tahmin edilmesi veya belirlenmesi daha adilane, daha spesifik ve daha bilimsel bir yaklaşım sağlayacaktır.

Ayrıca çalışmada abone devri, kapama ve açma işlemlerinden dolayı tüketim davranış modellerinin yenilenmesinin uygun olacağı, çalışmada tüketim gün sayısının, kullanıcı değişiminin, sayaç değişiminin, ödeme durumunun dikkate alındığı ancak başkaca parametrelerin de etkili olabileceği gerektiğinde önerilen modele bunlarında eklenebileceği önerilmektedir.

Anomali analizleri sonucunda yapılan tespitler bir öngörü vermektedir ve gerçekliği ancak saha çalışmaları sonucunda gelen geri dönütler ile anlaşılacaktır. Dolayısıyla hangi anomali tespit yöntemi kullanılırsa kullanılsın, abone tüketimlerinin sınıflarının öngörülmesinde tam bir başarı sağlanması arzu edilen bir sonuçtur, fakat tam bir başarı sağlanmazsa dahi önemli olan başarı oranını artıracak adımlar atmak, etkili olduğu anlaşılan başkaca veriler varsa bunları temin etmek ve geri dönütlerle gelen hataları düzelterek ilerlemek olmalıdır.

Ayrıca sular idarelerinin çoğunda yapılan şebeke basınç denetimlerinden elde edilen kayıp kaçak lokasyonları ile bu çalışmadan elde edilen anormal tüketim yapan abonelere ait lokasyonlar eşleştirilerek kayıp kaçak denetimlerinde daha spesifik çözümler elde edilebilir.

Ayrıca abonelere özel tahmin modelleri geliştirildiğinde, bölgelere ayrılmış içmesuyu şebekesinde de ihtiyaç duyulacak aylık, yıllık içme suyu miktarları daha gerçekçi olarak tahmin edilebilecektir. Böylece arz yönlü değil talep yönlü bir bakış açısıyla su yöntemi sağlanmış olacaktır. Bu bakış açısıyla özellikle suya ulaşımın çok zor olduğu bölgelerde ve kuraklık dönemlerinde daha etkili analizler yapılmasına yardımcı olacak ve sonuçlar alınmasını sağlayacaktır. Bu çalışmadaki taleplerin disipline edilmesi hedefi bu amaca hizmet etmektedir.

Çalışmada her bir abone için modellerin geliştirilmesi ve en iyi modelin doğruluk performans metriği, işlem adeti ve işlem süreleri dikkate alınarak seçilmesi anlık ihtiyaç duyulacak bir durum olmayıp sayaç okuma işleminin dönemsel yapılıyor olması avantajı ile ilgili ay içinde analizin yapılmasının yeterli olacağı, dolayısıyla modellerin hız performanslarından çok, veri kalitesi ve model verimliliğinin öncelenebileceği ifade edilebilir. Sahada sayaç okuması esnasında anlık sınıflandırma yapılması daha önce geliştirilmiş modellerin kullanımıyla sağlanacaktır.

Tatil günleri, pandemi gibi etkenlerin sebep olduğu olağanüstü yaşam koşulları tüketimi dönemsel etkileyip davranış modellerini yanıltabileceği öngörülmektedir. Bu husus başka bir çalışmanın konusu olarak da değerlendirilebilir.

Çalışma sonucunda içmesuyu abonelerine ait anormal tüketim dönem sayısının toplam tüketim dönem sayısındaki oranının yaklaşık %2 olduğu düşünülürse verisetinin dengesiz bir yapıda olduğundan bahsedilebilir. Çalışmanın devamı ve tamamlayıcısı olarak tüketim sınıfı tahmin modellerinin geliştirilmesi düşünüldüğünde, elde edilen gözetimli verisetlerinin KNN algoritması temelli SMOTE gibi veri çoğaltma algoritmaları veya normal tüketimlerin anormal tüketim değerleri ile değiştirilmesi yöntemleri ile dengeli hale getirilmesinin tahmin modellerinin performanslarını arttıracağı öngörülmektedir. Çalışmada içme suyu tüketim sınıfı tahmin model performansları yeterli olduğu için buna gerek duyulmamıştır. İhtiyaç duyulması halinde normal tüketim değerleri yerine kullanılacak anormal tüketim değerleri, yaz ayları tüketiminin kış aylarında veya kış ayları tüketimlerinin yaz aylarında yapılmış gibi gösterilmesiyle de elde edilebilir.

Çalışma sonunda elde edilen gözetimli veriseti, tahmin modelleri geliştirilmeden önce tekrardan öznitelik seçim işlemlerine tabi tutulabileceği, yeterli sayıda faktörün seçilebileceği, analiz hızının, verimin ve performans metrik değerlerinin arttırılabileceği öngörülmektedir.

İçme suyu abone sayaçlarının, ölçüm hassasiyetlerinden dolayı en geç 10 yılda bir ya kalibrasyon ayarlarının yapılması ya da değişiminin yapılması gerekmekte olup, çalışma ile yaklaşık 16 yıldır tek sayaç kullanan 118 abonenin sayaç değişimlerinin yapılmasının, benzer olarak sayaç değişim kayıtlarına ulaşılamayan diğer bütün abonelerin endeks kontrolü ile 10 yıllık sayaç denetimlerinin yapılmasının daha doğru tüketim ölçüm değerlerini sağlayacağı öngörülmektedir.

Abonelere sayaç okuma esnasında verilen su bildirim makbuzlarıyla birlikte tüketim alışkanlığının normal veya anormal seyirde olduğu, anormal ise hangi sınıf bir tüketim yaptıkları bilgisinin verilerek su tüketimi duyarlılığının artırılacağı öngörülmektedir. Ayrıca hane içi su tüketimlerinin genelde hangi amaçlarla yapıldığı, nerelerde yapıldığı, ne kadar yapılması gerektiği ve nerelerde bir tasarruf sağlanabileceği bilgilendirmelerinin veya önerilerinin kurum ve aboneler arasında var olan bütün iletişim kanalları aracılığıyla mesajlar, broşürler kısa filmler vb. yöntemler kullanarak yapılmasının da duyarlılığa katkı sağlayacağı öngörülmektedir.

Tespit edilen anormal tüketimlerde vatandaşın bilgilendirilmesi ve duyarlı hale getirilmesi ile tüketimlerin normal tüketim sınıflarına çekildiği düşünüldüğünde öncelikle bu durumun abone sözleşme sahibinin ekonomisine katkı sağlamasının yanında su idarelerinin su talebini karşılamak için alternatif kaynakların sisteme getirilmesi yatırımları ve arayışlarına destekleyici bir çözüm olacağı düşünülmektedir.

Bu çalışmada sadece konutlardaki içmesuyu tüketimi dikkate alınmış ve modeller geliştirilmiştir. Bununla birlikte önerilen yöntem toplu tüketimlerin söz konusu olduğu diğer bütün sektörlerde de kullanılabilir bir altlığa sahiptir. Aynı yöntem ile elektrik, doğalgaz, internet, mobil hatlar, kredi kartları gibi tüketimlerin bir abonelik sistemi üzerinden dönemsel yapıldığı ve kayıt altına alındığı alanlarda her bir abone için tüketim davranış modeli çıkartılabilir, anormal tüketimler tespit edilip belli sınırlarda tüketimlerin yapılması veya tüketimin disipline edilmesi sağlanabilir, tüketici kitlesinde geri bildirimler yapılarak konuyla ilgili duyarlılıklar geliştirilebilir. Ayrıca bu sektörlerde hizmet veren kurumlar yapay AI bu sistemi artı bir hizmet olarak kendi mobil uygulamalarında abonelerine sunarak daha fazla müşteri memnuniyeti sağlamış olacaklardır.

Teşekkür (Acknowledgment)

Abone verileri konusunda kurum desteği ve katkısını esirgemeyen Kayseri Su Kanalizasyon İdaresi'ne ve Genel Müdür Doç. Dr. Özgür ÖZDEMİR'e ayrıca su yönetimi ile ilgili birikimlerinden istifade ettiğim her konuda bilgi ve desteğini esirgemeyen Prof. Dr. Mahmut FIRAT'a teşekkür ederim.

Kaynaklar (References)

Arup and Sydney Water, 2015. The Future of Urban Water: Scenarios for Urban Water Utilities in 2040 [Internet]. Arup, Available from: https://www.arup.com/perspectives/publications/research/section/the-future-of-urban-water

Billings RB., Jones CV., 2008. Forecasting Urban Water Demand (Second Edition) [Internet]. Second. Kitap-American Water Works Association, 1–367 p, Available from: www.awwa.org

Cai J, Luo J, Wang S, Yang S., 2018. Feature Selection in Machine Learning: A New Perspective. Neurocomputing. 300, 70–9.

Campbell HE, Johnson RM., Larson EH., 2004. Prices, devices, people, or rules: The relative effectiveness of policy instruments in water conservation. Review of Policy Research, 21(5), 637–62.

Cini J, Mung A, Waughray D., 2014. Global Agenda Council on Water Security 2012-2014 [Internet]. World Economic Forum. 2014, Available from: http://www.weforum.org/content/global-agenda-council-water-security-2012-2014

International Water Association. 2017, The IWA principles for water wise cities - for urban stakeholders to develop a shared vision and act towards sustainable urban water in resilient and liveable cities. Urban Stakeholders to Develop a Shared Vision and Act towards Sustainable Urban Water in Resilient and Liveable Cities [Internet], 2nd, 1–6, Available from: https://iwa-network.org/wp-content/uploads/2016/10/IWA_Brochure_Water_Wise_Communities_SCREEN-1.pdf

Mayer P, Dziegielewski B, Kiefer JC, Lantz GL, Opitz EM, Porter G. Et al., 2000. Commercial and Institutional End Uses of Water. Power, 2014 p.

Mayer PW, DeOreo WB, Opitz EM, Kiefer JC, Davis WY, Dziegielewski B, et al., 1999. Residential End Uses of Water Executive Summary. AWWA Research Foundation and American Water Works Association, 1–345 p.

Türkiye Cumhuriyeti Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı, 2020. Sürdürülebilir Kalkınma Amaçları ve Göstergeleri. In: MÜDÜRLÜĞÜ YHG, BAŞKANLIĞI BVBYD, editors. Strateji ve Bütçe Başkanlığı Yayınları [Internet]. Haziran 20. Ankara: STRATEJİ VE BÜTÇE BAŞKANLIĞI, p. 1–42, Available from: http://www.surdurulebilirkalkinma.gov.tr/wp-content/uploads/2021/02/SKA-ve-Gostergeleri-Kapak-Birlestirilmis.pdf

Water Conflict, 2022. Water Conflict Chronology [Internet]. Vol. 1, Pacific Institute 2022, p. 1–52, Available from: https://www.worldwater.org/conflict/list/

Yıldız D, Özgüler H., 2020. Yapay Zekâ ve Su Yönetimi. Vol. 30, Rapor. Ankara.

Zhang R, Nie F, Li X, Wei X., 2019. Feature Selection with Multi-view Data: A Survey. Information Fusion [Internet], 50(March 2019), 158–67, Available from: https://doi.org/10.1016/j.inffus.2018.11.019

Research Article
Journal of Intelligent Systems: Theory and Applications 6(2) (2023) 174-180
DOI: 10.38016/jista.1229271

A Study of Ensemble Deep Learning Method Using Transfer Learning for Horticultural Data Classification

Gökhan Atalı[1*], Sedanur Kırcı[2]

[1]Sakarya University Of Applied Sciences, Department of Mechatronics Engineering, Sakarya, Türkiye

[2] Sakarya University Of Applied Sciences, Department of Mechatronics Engineering, Sakarya, Türkiye

gatali@subu.edu.tr, 22501005023@subu.edu.tr

Abstract

Deep learning is an important discipline in which human-specific problems are solved with the help of machines with advanced hardware power. It is seen this discipline is widely used in the fields of industry, health, defense industry, and sports. In addition, the use of deep learning in the field of horticulture is an important requirement. With the integration of deep learning into horticulture, to do product classification is very important for increasing productivity and production.

In this study, a method using ensemble learning is proposed to improve the accuracy of the classification problem for horticultural data. For this method, a new dataset was created, containing a total of 24421 images and 15 crop classes, independent of data augmentation. In order to train this created data set with the help of the proposed method, a hierarchical structure has been designed in which the output of one model is the input of the other model. A total of 7 pre-trained models were used in the experimental studies of the proposed method. Since this method is in an ensemble structure, it is possible to add or remove pre-trained models from the structure. With the help of experimental studies, a performance analysis of the proposed method, which is compared with the traditional CNN method, has been made. As a result of these analyses, it has been observed that the proposed method works 3% more successfully.

Keywords: Transfer learning, ensemble learning, convolutional neural network, image classification, deep learning.

Bitki Sınıflandırması için Transfer Learning Kullanılarak Topluluk Öğrenmesi Metodu Üzerine Bir Çalışma

Öz

Derin öğrenme, insana özgü problemlerin gelişmiş donanım gücüne sahip makineler yardımıyla çözüldüğü önemli bir disiplindir. Bu disiplinin sanayi, sağlık, savunma sanayi ve spor alanlarında yaygın olarak kullanıldığı görülmektedir. Ayrıca bahçecilik alanında derin öğrenmenin kullanılması önemli bir gerekliliktir. Derin öğrenmenin bahçeciliğe entegrasyonu ile ürün sınıflandırması yapmak, verimliliği ve üretimi artırmak için oldukça önemlidir.

Bu çalışmada çeşitli bitki verilerini kullanarak sınıflandırma probleminin doğruluğunu artırmak için topluluk öğrenmesi yöntemi önerilmiştir. Bu yöntem için veri artırmadan bağımsız olarak toplam 24421 görüntü ve 15 ürün sınıfı içeren yeni bir veri seti oluşturulmuştur. Önerilen yöntem yardımıyla oluşturulan bu veri setini eğitmek için bir modelin çıktısının diğer modelin girdisi olduğu hiyerarşik bir yapı tasarlanmıştır. Önerilen yöntemin deneysel çalışmalarında toplam 7 adet önceden eğitilmiş model kullanılmıştır. Bu yöntem bir topluluk yapısında olduğu için yapıya önceden eğitilmiş modeller eklemek veya çıkarmak mümkündür. Deneysel çalışmalar yardımıyla önerilen yöntemin geleneksel CNN yöntemi ile karşılaştırılan performans analizi yapılmıştır. Bu analizler sonucunda önerilen yöntemin %3 daha başarılı çalıştığı görülmüştür.

Anahtar kelimeler: Transfer öğrenme, topluluk öğrenme, evrişimli sinir ağı, görüntü sınıflandırma, derin öğrenme.

* Corresponding Author.
E-mail: gatali@subu.edu.tr

Recieved : 4 Jan 2023
Revision : 25 Mar 2023
Accepted : 14 Aug 2023

1. Introduction

With the development of artificial intelligence technologies, human-specific problems become the subject of machines. Artificial intelligence techniques combined with machines are driving technological developments in many areas such as speech recognition, visual object recognition, object detection, disease diagnosis, and gene sequence classification. In many studies on deep learning, which is a sub-discipline of artificial intelligence, researchers have offered various solutions to reach the most accurate result with different methods (LeCun et al., 2015). The classification problem is one that artificial intelligence can solve on a large scale. Although various traditional deep learning algorithms continue to work on this problem, the realization of accuracy and speed improvements is an important issue.

The most important source of information in machine learning and deep learning projects is data. Where data source is few, or data collection is difficult, pre-trained models with more easily collected data are needed. These models, which were previously trained with large data sets and whose weights are generally accepted, are called SOTA (State-of-the-Art) models, and this method is called transfer learning. Transfer learning is a method that has been tried many times and has proven itself in this field (Weiss et al., 2016, Babu &Annavarapu, 2021, Altaf et al., 2021, Salama & Aly, 2021, Vidal et al., 2021). Transfer learning is very advantageous compared to creating a new model since the precision of the results is not known when creating a model network and it requires many trials. Classification of horticultural products is an important field of study in which artificial intelligence technologies can be used. Using more than one different model together is a method called ensemble learning, which increases the accuracy of the model. With this learning method, in which the results of several models are analyzed together without depending on a single model, more reliable and accurate predictions are obtained (Re & Valentini, 2014). Ensemble learning can be adapted to classification and regression problems using different algorithms.

In this study, a structure in which more than one model is used together is proposed in order to solve the classification problem and increase its accuracy. This structure is designed to use different models together and the output of one model forms the input of the other model. These models, which were trained using transfer learning with a total of 24421 images, were used to classify 16 different classes. Performance analyses of the proposed method were carried out on the created data set and the results were presented in detail. In addition, the proposed structure was compared with the traditional CNN (Convolutional Neural Networks) classification method, and precision, recall, and f1score values were measured.

2. Related Works

When the literature is examined, many studies using the transfer learning method, seem to focus on images, especially radiography, etc. The most important reason for this is that medical data sets consist of accessible data. However, although there are difficulties in collecting data, studies in the field of horticulture have also been encountered. Studies in both scientific fields focus on the comparison of many common scientific methods, regardless of the differences in data sets.

For the classification of horticultural data, many studies have been carried out and different methods have been used so far (Yang & Xu, 2021, Palaparthi et al., 2023).

Abed et al., using a dataset with 1295 images and 3 classes to increase productivity in horticulture, detected disease on bean leaves. In this determination, they measured the performances of more than one pre-trained model and compared the accuracy, selectivity, f1score, and AUC (Area Under the Curve) values. Among the pre-trained models compared, DenseNet121 gave the best result with an accuracy of 98.31% (Abed et al., 2014). Zhao et al. created an object detection algorithm using deep learning with the images they obtained with the help of unmanned aerial vehicles in order to use artificial intelligence in the field of modern horticulture. Using the pre-trained YOLO v3 model, a bale detection algorithm was created on the labeled data. It was observed that detection performance increased in the model they trained using approximately 243 images (Zhao et al., 2021). In their study, Garcia et al. proposed an artificial intelligence assistant model in which both deep learning and traditional machine learning algorithms are combined for the classification of horticultural plants. For the test of the created model, two crops and two weed groups were selected, a unique data set was created, and performance analyzes were demonstrated (Garcia et al., 2020). Dawai et al. used transfer learning to detect pests in order to increase productivity in horticulture. This model, which can predict a total of 10 classes, reached an accuracy of 93.84% (Dawei et al., 2019). Kang and Gwak were used to classify the freshness of fruit, which is an important issue in horticulture. They proposed an ensemble learning model in which multi-task deep convolutional neural networks based on ResNet50 and ResNet101 architectures are used together. The proposed method has reached high accuracy values (Kang and Gwak, 2021).

Weed and product classification in horticulture is an important issue for increasing productivity and production. It is seen as a result of the literature review that the concepts of transfer learning and ensemble learning give very good results in this regard (Garcia et al., 2021, Bosilj et al, 2019, Yang et al., 2019). Xie et al. used transfer learning to measure the quality of the

produced product and detected defective products in the carrot vegetable they selected as an example. In their study, they used 5 different state-of-the-art models and the ResNet50 model gave the best result. With the ensemble learning method, in which the ResNet50, Densenet121 and VGG16 models, which they created in order to increase the accuracy, are used together, they have reached 97.32% accuracy (Xie et al, 2021). Jahanbakhshi et al. developed a CNN structure using the stochastic pooling mechanism to detect and classify the apparent defects of sour lemon fruit. In order to perform these performance analyses, a unique data set with images of healthy and damaged sour lemon fruit has been created. They compared their work with other machine learning algorithms such as KNN (K-Nearest Neighbors), SVM (Support Vector Machine), ANN (Artificial Neural Networks), and DT (Decision Tree) (Jahanbakhshi et al, 2020).

Ensemble learning, which is a learning method where several different models are run at the same time and results are obtained, is frequently used in image classification. Ganaie et al. explained ensemble learning comprehensively in their study and provided information on its use in different fields (Ganaie et al., 2021). Ahmad et al. used the ensemble learning method, which combines the attributes of different models to solve the classification problem. In their study, it was observed that the combined use of MobileNet and InceptionV3 models made better classification than their separate use (Ahmad et al., 2021). ResNet50 and SSDMobileNetV2, which are state-of-the-art models, are frequently used in transfer learning. In many studies, the performance values of these two models are shown in detail (Linguo et al., 2021, Biswas et al., 2018, Shabir et al, 2021, Li et al 2018).

Considering these studies in the literature, in classification problems; It is seen that transfer learning and ensemble learning methods are frequently used. In this study, a different classification architecture is proposed using these 2 methods. The structure of this classification architecture, which aims to increase accuracy compared to traditional methods, is explained and performance data are measured and compared with different models. To the knowledge of the authors, the proposed architecture has not been done before.

3. Material Method

Transfer learning and fine-tuning have been used to compare the success of the proposed method with traditional methods. In order to perform transfer learning, one of the pre-trained models, ResNet50, was first applied to all layers. Then, the study was evaluated by using MobileNetV2 models, which are faster but less successful than ResNet50, in order to evaluate the performance. MobileNetV2 model is fast due to the basic parsing in the first layer. In the other layers, a new ensemble learning has been created by choosing ResNet50 to increase the result performance.

3.1. Dataset

All of the images included in the database in this study were compiled and used via Kaggle (Kaggle, 2022). Within the scope of the study, a total of 24421 images were used and these images were divided into 4 clusters as in Figure 1. There are 4 different classes in mushroom, flower, and fruit clusters, and there are 3 classes in vegetable clusters.

While 24421 images in the data set were used to train the proposed method, 100 different new images were preferred for the test of the model. In addition, 80% of the training data set is presented to the learning algorithm as train and 20% as validation in order for the model to produce more meaningful results. Sample images from the train and validation data are presented in Figure 2. Within the scope of the study, the amount of data of each class was taken as approximately 2000 in order to prevent the problem of bias in the model depending on the number of data of the classes. The data counts of the classes away from this reference were matched using data augmentation.

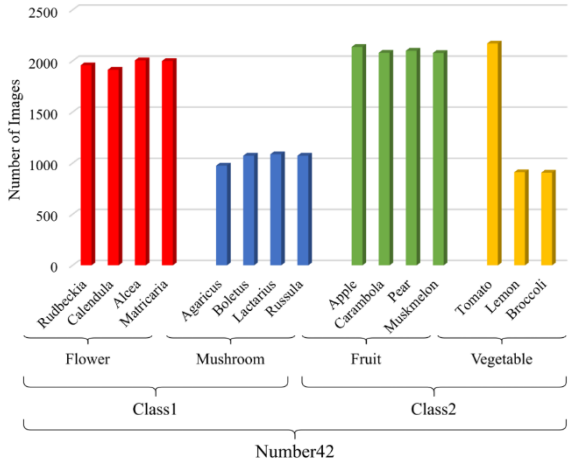


Figure 1. Distribution of data belonging to classes

In order to compare the proposed method with the models used in traditional CNN methods, a data set consisting of 4 clusters and 15 classes was prepared. In order to avoid bias in the prepared data set, the data augmentation method was applied to the classes with few visuals (Brocoli, Lemon, Agaricus, Boletus, Lactarius, Russula). In order to increase the number of images they contain, the mirroring method has been applied to the classes to which the data augmentation process has been applied, and it is aimed to double the number of images they contain. Thus, a total of 30,494 images were obtained. The dimensions of the images are scaled to be 224x224 pixels and 3 channels. When the graphics card and memory usage are taken into consideration, the batch method is used in order to shorten the processing time. A total of 953 packages

were created, with 32 images in each package, and 80% of the entire data set was divided into the train (Figure 2.a) and 20% validation (Figure 2.b).

(a) (b)

Figure 2. a) Train dataset visual, b) Validation dataset visual

3.2. Proposed method

Convolutional neural networks consist of two basic layers: convolutional and fully connected. The input data is first subjected to filtering and activation functions in the convolutional layer. At this stage, various pooling methods and normalization operations are performed to extract the attribute features of the input data. The input data from which the feature is extracted is made into a one-dimensional array in the flattening layer. Then, the flattened data is passed to the fully connected layer, where all neurons are interconnected for classification, and the convolutional neural network model is formed. In traditional CNN classification, more than one class is tried to be estimated using a single model. In this process, the precision of estimating a class decreases, and the success rate of the model decreases if the number of classes is large or the similarity between classes is intense.

In this study, a method is proposed in which models created with CNN structure are used together in order to increase the success of the traditional classification process. The working system of the proposed method, whose pseudo-code is given in Table 1, is based on the concept of an ensemble of models. In the proposed method, a total of 7 models were used together, with the output of one model forming the input of the other model.

Table 1. Pseudo code of proposed method

	Algorithm
1.	for image, labels in iterate the test data:
2.	for a=1:100 do:
3.	reshape of image[a]
4.	predict image with model_number_42
5.	select max index in class names of model_number_42
6.	if model of layer_1_pred equals to 'class1':
7.	if model of class1_2_pred equals to 'Flower':
8.	add flower_3(image) to predictions list
9.	else:
10.	add mushroom_3(image) to predictions list
11.	else:
12.	if model of class2_2_pred equals to 'Fruit':
13.	add fruits_3(image) to predictions list
14.	else:
15.	add vegetables_3(image) to predictions list

The proposed method consists of 3 layers as shown in Figure 3, and the first layer is named Number42, which is one of the important names that D. Adams brought to the literature (Adams, 1979). In the study, the Number42 model predicts 2 classes named Class1 and Class2. These classes are grouped according to the similarities of the classes in the 3rd layer. Models predicting classes in layer 3 are named after Class1 and Class2. Finally, the classification is completed when the models in the 3rd layer predict the output classes.

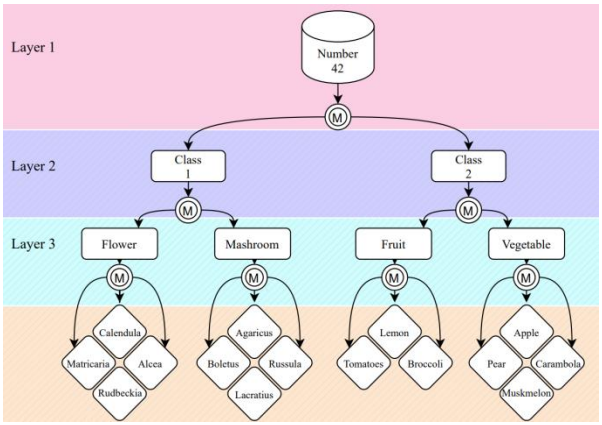


Figure 3. Proposed method structure (M: Pre-trained model)

3.3. Fine tuning

In artificial neural network models, the loss function must be at a minimum level for learning to function properly. Various optimization algorithms are used to minimize the loss function. In this study, ADAM (Adaptive Moment Estimation), which is one of the optimization algorithms frequently encountered in artificial neural network training, was used as a loss function, and Sparse Categorical Crossentropy was used. The w_t value in the ADAM optimization algorithm, whose formulation is given in Equation 1, represents the updated weight. The m_t represents bias corrected versions of moving averages. vt is the sum of the squares of the gradients up to time t. The learning rate, on the other hand, was taken as 0.001 by showing the symbol n. The ADAM function is found by multiplying the learning value for the weight update with the gradient of the function and subtracting it from the previous weight. In the fully connected layer of the

neural network, the Global Average Pooling method is preferred instead of the flattened layer in order to increase the computational performance. In order to reduce overfitting, 3 dropout layers were added to the study and the coefficient was taken as 0.3, valid for all models. All artificial neural network models were created by activating the GPU on Google Colab in order to measure the success of the experiment and control the process independently from the hardware.

$$w_t = w_{t-1} - n\frac{m_t}{\sqrt{v_t}+\epsilon} \quad (1)$$

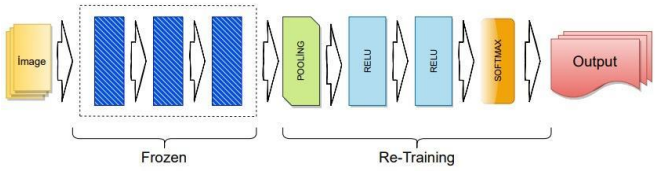


Figure 4. Pre-trained model diagram

As shown in Figure 4, while using the pre-trained models with optimum weights for retraining, a certain part of them is frozen. This process saves time and provides a performance increase.

3.4. Performance evaluation

Different evaluation metrics are used to see the performance of the proposed approach for the classification problem. The effectiveness of the proposed approach was measured and compared with the traditional CNN approach. The evaluation metrics used Accuracy (ACC), Precision (P), Recall (R), and $f_1$score (f_1) are shown in the Equation. (2-5). The fact that the results obtained from these metrics are close to 1 indicates that the model is a successful one.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$f_1score = 2 + \frac{P * R}{P + R} \quad (5)$$

TP, TN, FP, and FN; represent the True Positive, True Negative, False Positive, and False Negative prediction numbers, respectively.

4. Results

In this study, a structure has been considered in order to increase the accuracy compared to traditional methods and a method has been proposed in this direction. The proposed method is designed in such a way that the output of one model is the input of the other model by using the models one after the other. ResNet50 and MobileNetV2 models were used both together and separately to perform the performance analysis of the proposed method. In order to compare, a total of 3 experimental studies were conducted, 2 of which worked with the proposed method and the other with the traditional CNN method. The structure called 1Mob23Res, prepared with the proposed method, was designed using MobileNetV2 models in the 1st layer and ResNet50 models in the 2nd and 3rd layers. In another experimental study of the proposed method, the ResNet50 model was used in all layers of the structure called 123Res. In addition to these two experimental studies of the proposed method, CNN, which is the conventional image classification method, was used alone in the third experimental study. In the last experimental study, the CNN structure was performed using the pre-trained ResNet50 model called SM1. Comparative results of these experimental studies are given in Table 2. In the experiments, 100 test data prepared in accordance with the data set were used and the numerical distribution of the test data over the classes. In order to decide whether the proposed method is successful or not, the P, R and f1score values observed in the experiments were examined separately for each test data in Table 2. As a result of the studies, it was observed that the average ACC values of the 1Mob23Res and 123Res experimental studies were equal. Although ACC values are equal in these two experimental studies, there are differences for each class. For example, while P, R and f1score values are [0.50 1 0.67] in the 1Mob23Res structure of Lemon test data, this situation is observed as [0.45 1 0.62] in 123Res structure. In the calendula test data, this situation was observed as 1Mob23Res [0.78 1 0.88] and in the 123Res structure [1 1 1]. Lemon test data was better predicted by 1Mob23Res, but Calendula test data outperformed 123Res. In addition, the results of the comparison of the proposed method with SM1 are given in Table 2 on a class basis. The prediction performance in Tomatoes test data was observed as [0.50 0.14 0.22] in the 1Mob23Res and 123Res structures, while it was [0 0 0] in the SM1 structure. Therefore, the traditional CNN model, SM1, failed to predict the Tomatoes test data. From the experimental studies performed with the proposed method, 1Mob23Res and 123Res structures reached 83% ACC, while SM1 conventional CNN reached 80% ACC. As a result, the success rate of the proposed method is 3% better than the traditional method. In order to train the proposed method, a data set consisting of 24421 images was used, and for the test of the model, 100 new images were preferred, independent of the training data set. As a result of the experiments performed with these test data, although the ACC performance rates are equal, the time taken for the estimation of the classes is different from each other in both experimental studies. While the prediction times were 45.368 seconds for 1Mob23Res, this value was 49.001 seconds for 123Res. It has been observed that the prediction time

of the traditional CNN classification model (SM1) is 36,729s for this data set. This shows that the 1Mob23Res experimental study performed with the proposed method makes 19.02% slower estimation compared to the traditional method.

Table 2. Classification report by evaluation metrics, (P, R, and f1 represent precision, recall, and f1 score, respectively.)

	Multi-model								
	1Mob23Res			123Res			SM1		
	P	R	f_1	P	R	f_1	P	R	f_1
Agaricus	0.83	0.83	0.83	0.83	0.83	0.83	1.00	0.83	0.91
Alcea	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Apple	0.75	0.43	0.55	0.75	0.43	0.55	0.75	0.43	0.55
Boletus	1.00	0.83	0.91	1.00	0.83	0.91	0.80	0.67	0.73
Broccoli	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Calendula	0.78	1.00	0.88	1.00	1.00	1.00	0.88	1.00	0.93
Carambola	1.00	0.33	0.50	1.00	0.33	0.50	1.00	0.33	0.50
Lacratius	0.88	1.00	0.93	0.88	1.00	0.93	0.86	0.86	0.86
Lemon	0.50	1.00	0.67	0.45	1.00	0.62	0.41	1.00	0.58
Matricaria	0.86	1.00	0.92	0.86	1.00	0.92	1.00	1.00	1.00
Muskmelon	0.86	1.00	0.92	0.86	1.00	0.92	1.00	1.00	1.00
Pear	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Rudbeckia	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Russula	1.00	0.83	0.91	1.00	0.83	0.91	0.62	0.83	0.71
Tomatoes	0.50	0.14	0.22	0.50	0.14	0.22	0.0	0.0	0.0
ACC			0.83			0.83			0.80
Macro avg	0.86	0.83	0.82	0.87	0.83	0.82	0.82	0.80	0.78
Weighted avg	0.85	0.83	0.81	0.86	0.83	0.81	0.81	0.80	0.78

Since the proposed method in future studies is in an ensemble structure, it can be easily run with similar models such as InceptionV3, VGG16, and Efficient without any structural changes, and performance analyses can be evaluated with the same or different data sets. Similarly, by changing parameters such as the number of layers and classes, performance values such as accuracy, precision and f1score can be compared. When the number of data is increased, it is predicted that the proposed method will give more accurate results compared to the traditional method. Experiments with the method proposed in this study in the field of transfer learning will carry the study further.

Acknowledgment

We would like to thank the Sakarya University of Applied Science Robot Technologies and Intelligent Systems Application and Research Center (ROTASAM) for providing all kinds of opportunities for the realization of this study.

References

A. Palaparthi, A. M. Ramiya, H. Ram and D. D. Mishra, 2023. Classification of Horticultural Crops in High Resolution Multispectral Imagery Using Deep Learning Approaches, International Conference on Machine Intelligence for GeoAnalytics and Remote Sensing (MIGARS), Hyderabad, India.

Abed, S. H., Al Waisy, A. S., Mohammed, H. J., & Al Fahdawi, S., (2021). A modern deep learning framework in robot vision for automated bean leaves diseases detection, International Journal of Intelligent Robotics and Applications, 5, 235-251.

Ahmad, F., Farooq, A., & Ghani, M. U., (2021). Deep Ensemble Model for Classification of Novel Coronavirus in Chest X-Ray Images, Computational Intelligence and Neuroscience.

Altaf, F., Islam, S. M. S., & Janjua, N. K., (2021). A novel augmented deep transfer learning for classification of COVID-19 and other thoracic diseases from X-rays, Neural Computing and Applications.

Babu, S. A., &Annavarapu, C. S. R., (2021). Deep learning-based improved snapshot ensemble technique, The International Journal of Applied Intelligence, 51, 3104-3120.

Biswas, D., Su, H., Wang, C., Stevanovic, A., & Wang, W., (2018) An Automatic Traffic Density Estimation Using Single Shot Detection (SSD) and MobileNet-SSD, Physics and Chemistry of the Earth.

Bosilj, P., Aptoula, E., Duckett, T., &Cielniak, G., (2019). Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture, Journal of Field Robotics, 1-13.

D. Adams, (1979). The Hitchhiker's Guide to the Galaxy, London: Alfa.

Dawei, W., Limiao, D., Jiangong, N., Jiyue, G., Hongfei, Z., &Zhongzhi, H., (2019). Recognition pest by image-based transfer learning, Journal of the Science of Food and Agriculture, 99, 4524-4531.

Ganaiea, M., Hub, M., Tanveera, M., &Suganthanb, P., (2021). Ensemble deep learning: A review, Preprint submitted to Elsevier.

Garcia, B. E., Mylonas, N., Athanasakos, L., & Fountas, S., (2020). Towards weeds identification assistance through

transfer learning, Computers and Electronics in Agriculture, 171.

Garcia, B. E., Mylonas, N., Athanasakos, L., Vali, E., & Fountas, S., (2021). Combining generative adversarial networks and agricultural transfer learning for weeds identification, ScienceDirect, 79-89.

Jahanbakhshi, A., Momeny, M. M., Mahmoudi, M., & Zhang, Y. D., (2020). Classification of sour lemons based on apparent defects using stochastic pooling mechanism in deep convolutional neural networks, Scientia Horticulture, 263.

Kaggle, Kaggle Inc, [Online]. Available: https://www.kaggle.com/. (Accessed: 04. Jul. 2022).

Kang, J., & Gwak, J., (2021). Ensemble of multitask deep convolutional neural networks using transfer learning for fruit freshness classification, Multimedia Tools and Applications.

LeCun, Y., Bengio, Y., & Hinton, G., 2015. Deep Learning, Nature, 521, 436-444.

Li, Y., Huang, H., Xie, Q., Yao, L., & Chen, Q., (2018). Research on a Surface Defect Detection Algorithm Based on MobileNet-SSD, Applied Sciences, 8(9), 1677-1694.

Linguo, L., Li, S., & Su, J., (2021). A Multi-Category Brain Tumor Classification Method Bases on Improved ResNet50, Computers, Materials & Continua, 2(69), 2355-2366.

Re, M., & Valentini, G., (2014). Ensemble methods: A review, Advances in Machine Learning and Data Mining for Astronomy, 563-594.

Salama, W. M., & Aly, M. H., (2021). Deep learning in mammography images, Alexandria Engineering Journal, 60, 4701-4709.

Shabbir, A., Ali, N., Ahmed, J., Zafar, B., Rasheed, A., Sajid, M., Ahmed, A., & Dar, S. H., (2021). Satellite and Scene Image Classification Based on Transfer Learning and Fine Tuning of ResNet50, Mathematical Problems in Engineering.

Tian, X., & Chen, C., (2019). Modulation Pattern Recognition Based on Resnet50 Neural Network, IEEE International Conference on Information Communication and Signal Processing, Beijing.

Vidal, P. L., Moura, J. d., Novo, J., & Orgeta, M., (2021). Multi-stage transfer learning for lung segmentation using portable X-ray, Expert Systems with Applications, 173.

Weiss, K., Khoshgoftaar, T. M., & Wang, D., (2016). A Survey of Transfer Learning, Journal of Big Data, 3, 9.

Xie, W., Wei, S., Zheng, Z., Jiang, Y., & Yang, D., (2021). Recognition of Defective Carrots Based on Deep Learning Deep Learning and Transfer Learning, Food and Bioprocess Technology, 14(7),1-14.

Yang, B., Xu, Y., 2021. Applications of deep-learning approaches in horticultural research: a review, Horticulture Research, p., 01 06 2021.

Yang, M., He, Y., Zhang, H., Li, D., Bouras, A., Yu, X., & Tang, Y., (2019). The Research on Detection of Crop Diseases Ranking Based on Transfer Learning, International Conference on Information Science and Control Engineering (ICISCE), Shanghai.

Zhao, W., Yamada, W., Li, T., Diagman, M., & Runge, T., (2021). Augmenting Crop Detection for Precision Agriculture with Deep Visual Transfer Learning A Case Study of Bale Detection, Remote Sensing, 13(23).

Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 6(2) (2023) 181-190
DOI: 10.38016/jista.1222483

Kampüs İçi Kapalı Alanlarda Hava Kalitesinin Modellenmesi ve Karar Destek Sistemi Geliştirilmesi

Elif Cesur[1*], Cemal Efe[2]

[1] İstanbul Medeniyet Üniversitesi, Endüstri Mühendisliği Bölümü, İstanbul, Türkiye

[2] İstanbul Medeniyet Üniversitesi, Endüstri Mühendisliği Bölümü, İstanbul, Türkiye

elif.karakaya@medeniyet.edu.tr, cemal_efe2000@hotmail.com

Öz

Zeki Kampüs kapsamı içerisinde üniversite kampüslerinin kapalı alanlarında hava kalitesinin tahmin edilmesi, virüs bulaş riskini en aza indirilmesi açısından çok önemlidir. Buna bağlı olarak sınıflarda, idari ofislerde ve toplantı salonlarında hava ölçümlerinin kontrol limitleri dışına çıkması durumunda uyarılar vererek kararlar almasını sağlayacak bir karar destek sistemin kurulması, bu riski kontrol altında tutmayı sağlayabilecektir. Bu çalışmada öncelikle karbondioksit, sıcaklık, nem, basınç ve hava kalitesi (MQ135) sensörleri insan giriş ve çıkışının yoğun olduğu bir sınıfa kurulmuş ve düzenli olarak veri alınması sağlanmıştır. Çalışmanın ikinci aşamasında veri madenciliği teknikleri ile bu sensör verilerinin veri önişleme teknikleri ile analizleri yapılmıştır. Çalışmanın ana amacı yapay sinir ağları, karar ağacı ve destek vektör makine teknikleri ile sensör verilerinin modellenmesi ve kişi sayısı artışı, cam veya kapının açılması, ve ders arası süresinin uzatılması gibi nedenlerden kaynaklı olarak havada gerçekleşen ani değişikliklerin model tarafından yakalanmasını sağlamaktır. Çalışmanın sonucu, kabul edilebilir aralıkların dışına çıkan hava kalitesi durumlarının tespiti sonucunda odanın ne zaman havalandırılacağına karar vermektir. Bu araştırmada kurduğumuz modeller kampüs dışında da örneğin toplu taşıma araçlarının, işyerlerinin, ofislerin, restoranların, kafelerin ve özel araçların havalandırma sistemlerinde kullanılabilecek özelliktedir.

Anahtar Kelimeler: Makine Öğrenmesi, Zeki Kampüs, Karar Destek Sistemleri

Modeling of Indoor Air Quality in Campus and Developing a Decision Support System

Abstract

Estimating the air quality in the indoor areas of university campuses within the scope of İntelligent Campus is very important in terms of minimizing the risk of virus transmission. In order to reduce this threat, it will be possible to develop a decision support system that will allow it to make decisions by issuing alerts in the event that air measurements exceed the control limits in the classrooms, executive offices, and meeting rooms. In this investigation, carbon dioxide, temperature, humidity, pressure, and air quality (MQ135) sensors were first installed in a classroom where there was a significant amount of human input and output, and regular data were collected. In the second stage of the study, data mining techniques and data preprocessing techniques were used to analyze these sensor data. The main purpose of the study is to model sensor data with artificial neural networks, decision tree and support vector machine techniques, and to ensure that sudden changes in the air due to reasons such as increasing the number of people, opening the window or door, and extending the time between classes are captured by the model. The result of the study is to decide when to ventilate the room as a result of the detection of air quality conditions that fall outside the acceptable ranges.

Keywords: Machine Learning Algorithms, Intelligent Campus, Decision Support Systems

* Sorumlu yazar
E-posta adresi: elif.karakaya@medeniyet.edu.tr

Alındı : 21 Aralık 2022
Revizyon : 14 Ocak 2023
Kabul : 7 Nisan 2023

1. Giriş (Introduction)

Massachusetts Teknoloji Enstitüsü'nde (MIT) yapılan bir araştırmada kapalı alanlarda sosyal mesafe kuralının virüs bulaşma riskini çok az etkilediği ve yetersiz kaldığı sonucuna ulaşılmıştır. Kapalı alanlarda sosyal mesafenin dışında ortamdaki insan sayısı, havalandırma seviyesi, maske takıp takmadıkları ve insanların ne yaptığı gibi değişkenler de çok önemli olduğu ortaya konulmuştur. İnsanların yemek yediği, konuştuğu, hapşırdığı ve havanın hareket ettiği kapalı bir ortamlarda damlacıkların havada asılı durabildiği, daha boş ve sakin yerlerde ise bu parçacıkların yavaşça yere doğru hareket ettiği tespit edilmiştir. Bu nedenle araştırmacılar, virüs partiküllerinin havalandırma veya filtreleme yöntemiyle zararsız hale getirilebileceğini savunmaktadır. (Elibol, 2021).

Enfekte kişilerden saçılan küçük damlacıklar az miktarda virüs içermekte ve açık havada bu partiküller hızla buharlaşıp yok olmaktadır. Kapalı bir ortamın hava akımı ne kadar zayıf ve hacmi ne kadar küçük ise partiküllerin havada asılı kalma olasılığı da o aranda artmaktadır. Yapılan çalışmalar ve araştırmalar sosyal mesafe kuralının kapalı alanlarda virüsün bulaşmaması için yetersiz kaldığını göstermektedir. Bu sebeple kapalı ortamlarda mesafe yerine viral yük (kişi sayısı vb.) ve odanın hacmi ölçütleri bulaş riskinin azalması için önem kazanmaktadır. Enfekte bir kişinin de içerisinde bulunduğu toplu taşıma, hastane, okul, uçak veya bir otomobil gibi kapalı ortamlarda bulaş riskinin hesaplanabilmesi için yapılan çalışma veya simülasyonlarda 4 önemli korunma yöntemi belirlenmiştir. Bu korunma yöntemleri, 1) kapalı alanlarda insan yoğunluğunun azaltılması, 2) hasta ve duyarlı kişilerin birlikte maske kullanması, 3) semptomatik kişilerin izolasyonu ve 4) yeterli süre ve uygunlukta havalandırmanın yapılmasıdır. (Elibol, 2021)

Yeni normale geçilmesiyle birlikte insanlar kapalı ortamlarda daha fazla vakit geçirmeye başlamıştır. Kış mevsiminde havanın soğumasıyla birlikte bu durum daha fazla artmaya başlamıştır. Üniversitelerin tekrardan yüz yüze eğitime başlamasıyla birlikte sınıflarda öğrenci ve öğretmenler risk altında kalmaktadır. Fakat sınıflardaki pencerelerin yetersizliği ve bazı sınıflarda havalandırmanın olmaması bulaşma riskini arttırmaktadır. Yüz yüze sınavlarda öğrenci sayısının artmasıyla birlikte dolaylı olarak hava kalitesi de azalmaktadır. Bu durum virüslerin bulaşıcılığını arttırmaktadır. Bu çalışmada, insanların bir arada bulunduğu yerlerden olan üniversite kampüslerinin kapalı alanlarında hava kalitesi tahmin modelleri ve karar destek sistemi geliştirilerek virüsün bulaşma riskini en aza indirilmesi amaçlanmıştır.

Öğrenci ve öğretmenlerin çoğunlukla bir arada bulunduğu kapalı kampüs ortamlarında hava kalitesi tahmin modelleri oluşturularak, hava kalitesinin düşmesi durumunda karar destek sistemi harekete geçerek kapalı ortamdaki hava kalitesi yükseltilmeye çalışılacaktır. Bu nedenle yaptığımız çalışmanın ileri aşamalarında oluşturduğumuz model ve karar destek sistemleri ile entegre edilecek havalandırma sistemi hava kalitesinin düşmesi durumunda devreye girerek hava kalitesinin devamlı istenilen düzeyde kalmasını sağlayacaktır.

Kampüsteki kapalı alanlarda insan sayısının sabit olmaması ve devamlı değişmesi (sınıftaki öğrenci sayısı, idari ve derslik binalarına giren kişi sayısı, yemekhane ve kafeteryalardaki öğrenci sayısı vb.) hava kalitesinin de devamlı değişmesine neden olmaktadır. Makine öğrenmesi yöntemlerini kullanarak geliştirdiğimiz karar destek sistemi ile ortamdan elde edilen sıcaklık, nem, CO_2, partikül madde (PM_{10}) miktarı ve basınç verileri kullanılarak hava kalitesinin düşmesi durumunda havalandırma etkinleştirilecektir. Çalışmamızda kampüslerde kapalı ortamlardan elde edilen verilerin makine öğrenmesi algoritmaları ile modellenerek karar destek sistemi aracılığı ile havalandırma sisteminin daha etkin kullanımı ile hava kalitesinin sağlıklı düzeyde tutulması hedeflenmiştir. Ayrıca bu araştırmada kurduğumuz modeller kampüs dışında toplu taşıma araçları, iş yerleri, ofis, restoran, kafe, özel araçların havalandırma sistemlerinde de kullanılabilecektir.

2. Literatür Taraması (Literature Review)

Lelieveld vd. (2020),çalışmalarında ofis, sınıf, oda gibi tipik iç mekân ortamlarında Aerosol haline getirilmiş SARS-CoV-2 virüslerinin hava yoluyla bulaşmasındaki rolü tartışılmıştır. Araştırmada aerosoller yoluyla ev içi virüs enfeksiyonuna karşı hafifletme önlemlerinin etkinliği hakkında bilgi sağlamayı amaçlanmıştır. Araştırmada oda boyutu, maruz kalan denek sayısı, inhalasyon hacmi ve solunum ve seslendirmeden aerosol üretimi gibi ayarlanabilir parametrelere dayalı olarak, aerosol haline getirilmiş virüslerden iç mekân enfeksiyon riskini tahmin etmek için basit, şeffaf ve kolayca ayarlanabilen bir elektronik tablo algoritması geliştirilmiştir. Araştırma sonuçları kapalı ortamda hava yoluyla bulaşmanın önemli bir faktör olduğunu doğrulamaktadır. Belirsizliklere rağmen, dış hava ile aktif havalandırma ve hava filtreleme gibi farklı etki azaltma önlemleri için öngörülen kısmi azalmaların olduğu görülmüştür (Lelieveld et al., 2020).

Dokuz vd. (2020) çalışmalarında, hava kirletici parametrelerin tahmin edilmesi, çevreye olan etkileri, özellikleri ve değerlendirilmesinde uygulanan makine öğrenmesi yöntemlerinin neler olduğuna dair detaylı sonuçlar vermiştir. Bu araştırma, hava kalitesini iyileştirmek ve sürdürülebilir bir çevrenin oluşturulabilmesi için hangi yöntem ve hangi parametrelerin kullanılması gerektiği sorusuna cevap aramaktadır. Veri hacminin büyüklüğü seçilen yöntemin başarısını etkilediği gözlemlenmiştir. Hava kalitesini belirleyen parametrelerin değerlerinin doğru hesaplanabilmesi için ortamda yeterli istasyon sayısı ve istasyonların konumları önem taşımaktadır. Uygulama alanından toplanılacak verilerin kalitesi, düzenliliği ve

hassasiyeti çalışmanın doğru sonuçlar verebilmesi için önemli olduğu sonucuna varılmıştır. (Dokuz et al., 2020)
Irmak ve Aydilek hava kalitesini ölçmek amacıyla yaptıkları çalışmada Adana ili valilik istasyonuna ait azot dioksit (NO_2), ozon (O_3), partikül madde (PM_{10}), karbon monoksit (CO) ve kükürt dioksit (SO_2) gibi hava kirleticilerin ölçüm değerlerine ait veriler kullanılmıştır. (Irmak and Aydilek, 2019). Gültepe (2019) çalışmasında Kastamonu ili ele alınarak, çeşitli makine öğrenmesi algoritmaları ile nem, PM_{10}, rüzgâr yönü, SO_2, hava basıncı, rüzgâr hızı ve hava sıcaklığı gibi meteorolojik parametreleri kullanarak hava kirliliği tahmini yapacak modeller tasarlanmıştır. (Gültepe, 2019). Tahmin modellerinde, Yapay Sinir Ağları, Basit Bayes, Karar Ağacı, Lineer Regresyon, SVM, K-En Yakın Komşu, Rastgele Orman, Lojistik Regresyon ve diğer regresyon algoritma yöntemleri kullanılmıştır. Çalışma süreleri ve hata oranları bakımından algoritmaların başarı değerleri kıyaslanmıştır.
Karakuş ve Yıldız (2019) çalışmalarında Sivas kent merkezindeki hava kirliliği değerlendirilmiş ve hava kalite indeksi (HKİ) hesaplanmıştır. Üç adet hava kalitesi izleme istasyonu kent merkezine konumlandırılarak hava kirletici parametrelerin çoklu regresyon yöntemi ile HKİ ve meteorolojik parametreler (rüzgâr hızı, bağıl nem ve sıcaklık) arasındaki ilişkileri belirlenmeye çalışılmıştır. (Karakuş and Yıldız, 2019).

Sakhidad Faizi (2021) çalışmasında Ocak 2019 ile Mart 2021 tarihleri arasında Kabil şehrine ait günlük $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, O_3 kirletici ölçümlerinin yanında nem, rüzgâr hızı, sıcaklık, basınç ve çiy noktası parametrelerinden oluşan veri setinin makine öğrenmesi yöntemleri ile modellenerek bu modellerden elde edilen başarıların karşılaştırılması hedeflenmiştir. Gauss süreç regresyonu, SVM, karar ağaçları, yapay sinir ağları ve lineer regresyon gibi veri madenciliği yöntemleri kullanılarak hava kalite indeksini tahmin eden modeller kurulmuştur. Hava kirliliği tahmininde elde edilen sonuçlara göre Gauss Süreç Regresyonuyla tüm veriler için en yüksek uygunluk değeri ve en düşük hata değerleri elde edilmiştir(Sakhidad Faizi, 2021).
Zhang vd. (2019)çalışmada tahmin ve değerlendirmeden oluşan yeni bir hava kalitesi erken uyarı sistemi geliştirilmiştir. İlk olarak, gelişmiş veriler ön işleme teknolojisi güçlü sürü zekası algoritması, Balina Optimizasyon Algoritması (WOA) ve verimli yapay sinir ağı ile birleştirilmiş bir tahmin modelini oluşturulmuştur. Ardından, tahmine dayalı sonuçlar, sezgisel hava kalitesi bilgileri ve ilgili önlemleri sunan bulanık kapsamlı değerlendirme yöntemiyle analiz edilmiştir. Bu çalışmada, Pekin, Tianjin ve Shijiazhuang'da altı ana hava kirletici $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO ve O_3 seçilmiştir. (Mo et al., 2019)
Balta (2019)çalışmasında sınıf gibi kapalı ortamlardaki bütün konfor parametrelerinin birlikte değerlendirilerek hava kalitesini izleme sistemi ile kapalı bir mekandaki insanlar için en temiz hava kalitesinin sağlanması hedeflenmiştir. Sınıflardaki sıcaklık, hava akışı, CO_2 konsantrasyonu ve nem değerleri tasarlanan iç ortam hava kalitesi izleme sistemi ile anlık ölçüm sonuçlarını sağlayarak iç hava kalitesinin tespit edilmesini sağlamıştır. Kapalı ortam çevresel kalite bilgisi (IEQ) değeri, sınıflarda bulunan öğrencilere yapılan anket ve ölçüm sonuçlarının değerlendirilmesi ile kurulan bulanık sistem sayesinde veriler anlık analiz edilerek hesaplanmıştır. (Balta, 2019)
Santos vd. (2020), çalışmada yoğun bakım ünitesinde ve tüm hastane ortamında koronavirüs hastalığını azaltmak için enfeksiyon riskini en aza indiren ısıtma, havalandırma ve iklimlendirmenin rolünü vurgulamaktadır. Sonuç olarak, nemin insan konforu üzerinde sağlık sorununda daha az etkisi olduğu saptanmıştır. Düşük sıcaklık cansız yüzeylerde canlı virüslerin kalıcılığını arttırırken yüksek sıcaklıklar korona virüslerin kalıcılığını azaltmaktadır. (Santos et al., 2020)
Zhang ve arkadaşlarının 1236 bölge için yaptığı araştırmada COVID-19'un bulaşmasında hava koşullarının etkisi saptanmaya çalışılmıştır. Büyük ölçekli uydu verileri, sıcaklık ve bağıl nemin COVID-19 yayılması üzerindeki etkilerini ve mevsimsel döngülere göre olası bulaşma riskini araştırmak için bir regresyon analizi modeliyle bu verilerle birleştirilmiştir. Sonuç olarak, sıcaklık ve bağıl nemin dünya genelinde COVID-19 bulaşması ile negatif ilişkili olduğunu göstermektedir. Daha yüksek sıcaklık ve daha yüksek nemin iletimi azaltabileceği bulunmuştur. (Zhang et al., 2021)
Gupta vd., çalışmalarındasıcaklık, güneş ışığı saatleri ve nem dahil olmak üzere farklı hava faktörlerinin etkisi değerlendirilerek derin transfer öğrenme tabanlı kapsamlı bir analiz gerçekleştirilmektedir. Çalışmadaki temel teori, COVID-19 salgınında pozitif vaka sayısı ile virüslerin iklimle yayılması arasında mevsime dayalı bir model gösterdiği varsayımıdır. Çalışma sonucunda yapılan deneysel sonuçlardan sıcaklık, rüzgâr hızı ve güneş ışığı saatlerinin COVID-19 vakaları ve ölümleri üzerinde önemli bir etkisi olduğu gösterilmiştir. Konvolüsyonel sinir ağının rekabetçi modelden daha iyi performans gösterdiği sonucuna varılmıştır. (Gupta et al., 2021)
Makine öğrenmesi algoritmaları güncel çalışmalarda ve farklı bir çok alanda kullanılmaya devam edilmektedir. ((Eren et al., 2023; Aksangür et al., 2022).

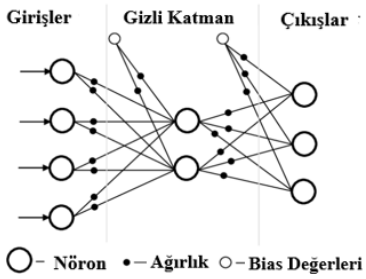
3. Makine öğrenmesi Algoritmaları
(Machine Learning Algorithms)

Büyük veriden deneyim yoluyla öğrenen ve bu öğrendikleriyle otonom davranış gösterebilen çeşitli algoritmaların oluşturulmasına makine öğrenmesi denir. Makine öğrenmesi modelleri ile kendi kendine karar alabilen otonom akıllı sistemler yapılabilmektedir. Naive bayes, yapay sinir ağları, k-means, karar ağaçları ve çeşitli regresyon algoritmaları, model eşleştirme ve istatistiksel analiz yöntemleri kullanılarak geçmiş verilerden öğrenebilmektedir. Daha sonra öğrendiği bu

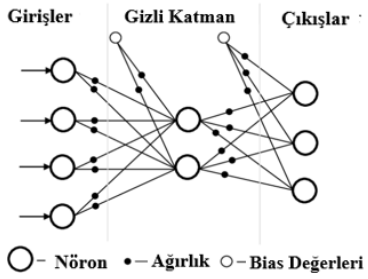
bilgileri kullanarak bir model oluşturur. Oluşturulan bu model ile gelecek verilere ait tahminlerde bulunur. Makine öğrenmesinin temel omurgası veridir. Makine öğrenmesi algoritmaları, çözmeleri amaçlanan görev veya sorun türlerine, girdi ve çıktılarına, yaklaşımlarına ve veri türlerine göre çeşitlilik gösterirler. Makine öğrenmesi, 1) Denetimli Öğrenme 2) Denetimsiz Öğrenme 3) Yarı denetimli Öğrenme, 4) Pekiştirmeli Öğrenme olmak üzere temel alt sınıflara ayrılır. Çalışma kapsamında kullanılan üç farklı makine öğrenmesi algoritması aşağıda açıklanmıştır.

3.1. Yapay Sinir Ağları (Artificial Neural Networks)

İleriye ve geriye dönük besleme olacak şekilde tasarlanan yapay sinir ağları, insan beyninde bulunan nöronların çalışma prensibi şeklinde modellenmiştir. Öğrenebilen bir algoritma olan yapay sinir ağları birbirlerine bağlı düğümler grubu şeklindeki yapısı (Kulkarni et al., 2017)



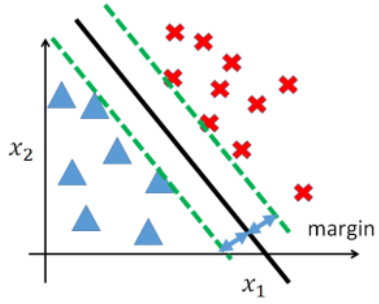
Şekil 1'deki şekilde kurulur. Bu şemada her düğüm bir nöronu ve her ok nöronlar arasındaki bağlantıyı temsil eder. Yapay sinir ağı yapısı kurulduktan sonra giriş verileri ve çıkış verileri kıyaslanarak modelin eğitilme süreci başlatılır. Her aşama sonrasında doğru veriye daha fazla yaklaşılır. Çıktı değerleri kontrol edilerek hangi modelin en uygun model olduğu bulunarak öğrenme işlemi tamamlanır. Öğrenme işlemi tamamlandıktan sonra yapay sinir ağı modeli ile test verisi tahmin edilir. (Kulkarni et al., 2017)



Şekil 1: Yapay Sinir Ağları Modeli (Artificial Neural Networks Model) (Kulkarni et al., 2017)

3.2. Destek Vektörleri Makinaları (Support Vector Machines)

Verileri farklı sınıflara ayırmak ve sınır aralığını en üst düzeye çıkarabilmek için destek vektör makinesi algoritması karar sınırı belirler. Bu karar sınırına en yakın iki nokta arasında eşit uzaklıktan geçecek şekilde vektörler çizilir. Çizilen vektörler arasından modeli en iyi tanımlayan ve sınıflandıran destek vektörü seçilir. Vektörler lineer, radyal, sigmoid gibi çeşitlerde oluşturulabilir. Sınıflandırma ve regresyon için kullanılabilen destek vektör makine modelleri bir dizi eğitim verisi verildiğinde çizilen vektörün konumuna göre oluşturulan kategorilerde tahmini sınıflandırma yapabilmektedir.(Chang and Lin, 2011) Şekil 2 de çizilen vektörler yardımı ile nasıl verilerin iki kümeye ayrıldığı gösterilmektedir.

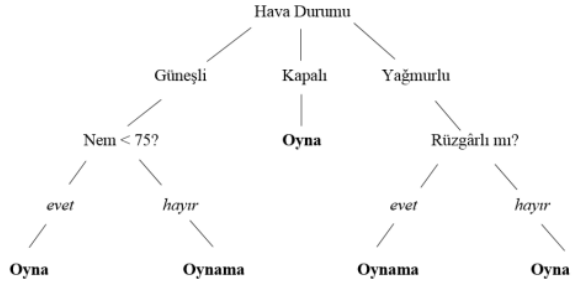


Şekil 2: Destek Vektör Makine Modeli (Support Vector Machines Model) (Burges, 1998)

3.3. Karar Ağacı Algoritması (Decision Tree Algorithm)

Karar ağaçları, karar alma aşamasında seçenek sayısının çok olduğu durumlarda en yaygın biçimde kullanılan bir sınıflandırma yöntemidir. İnsan gözü ile kolay bir şekilde yorumlanabildiği için sıklıkla tercih edilmektedir. Karar ağacı modelinde, öznitelikler ağacın iç düğümünde temsil edilmektedir ve her yaprak düğümün bir sınıf etiketine karşılık gelmektedir. Entropi yöntemi ile sınıflar belirlenen sıralamaya ağaçta dallandırılarak tahmin sonuçları elde edilir. Şekil 3'te örnek bir karar ağacı modeli verilmiştir.

Süre (sn)	Nem	Sıcaklık	CO2	Hava Kalite	Basınç	3'lü Ortlama	5'li Ortalama	10'lu Ortalama	Durum
7,264	47,2	29,2	400	78	1164,25				0
8,947	47,2	29,2	400	79	1163,67				0
10,635	47,2	29,2	400	94	1164,32	400			0
12,318	47,2	29,2	400	77	1165,15	400			0
14,006	47,1	29,2	400	75	1163,65	400	400		0
15,689	47,1	29,2	400	90	1159,62	400	400		0
17,376	47,1	29,2	400	80	1160,47	400	400		0
19,059	47,1	29,2	400	96	1161,36	400	400		0
20,747	47,1	29,2	400	77	1160,94	400	400		0
22,43	47,1	29,2	408	77	1159,63	402,7	401,6	400,8	0
24,117	47,1	29,2	400	89	1158,67	402,7	401,6	400,8	0
25,801	47,1	29,2	410	77	1159,29	406	403,6	401,8	0
27,488	47,1	29,2	405	76	1161,27	405	404,6	402,3	0
29,171	47,1	29,2	400	75	1162,46	405	404,6	402,3	0



Şekil 3: Karar Ağacı Modeli (Decision Tree Model) (Uysal and Güyer, 2014)

4. Vaka Çalışması (Case Study)

4.1. Havalandırma Tahmin Uygulaması Veri Analizi (Ventilation Forecasting Application Data Analysis)

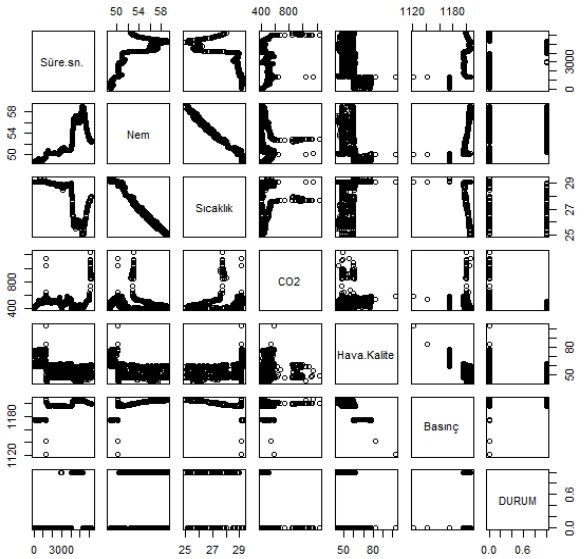
Kampüs içi kapalı alanlarda hava kalitesinin modellenmesi ve karar destek sisteminin geliştirilmesi için İstanbul Medeniyet Üniversitesi Dekanlığında yer alan hava kalitesi ölçüm cihazı ile 2021 Kasım ayında sıcaklık, nem, CO_2, Basınç (p) ve havadaki kirletici partikül değerlerinin (aq) ölçümleri yapılmıştır. Bu ölçümler belli saniye aralıkları ile ölçülerek kaydedilmiştir. 2021 Kasım ayında yapılan çalışmalarımıza ek olarak hava kalitesini etkileyen parametrelere ait yeni veriler toplanarak veri seti arttırılmıştır. Yeni veri değerleri kapalı bir sınıf ortamında ders işleyen öğrencilerin bulunduğu bir sınıftan toplanmıştır. Veriler toplanırken sınıf ortamında bazı farklı senaryolar oluşturularak sensörün bu durum karşısında nasıl bir tepki verdiği ölçülmeye çalışılmıştır. Kişi sayısı artışı, cam veya kapının açılması ve ders arası süresinin uzatılması gibi senaryolar oluşturulmuştur. Yapay sinir ağı ve diğer makine öğrenmesi algoritmaları kullanılarak sensörün bu durum değişikliklerini tespit etmesi hedeflenmiştir.

Çalışmamızda sensör ölçümleri 2K-02 kodlu sınıfta 40 kişilik bir sınıf mevcuduna sahip Üretim Planlama ve Kontrol dersinde iki farklı konumda yapılmıştır. Örnek veri setimiz 3128 satırdan oluşmaktadır. Sınıftaki ölçümler devam ederken farklı zamanlarda camlar açılarak sınıf havalandırılmıştır. Ayrıca ders arası gibi zamanlarda ortamdaki kişi sayısının azalması sebebiyle ölçüm değerlerinde değişiklikler gözlenmiştir. Sensör yardımıyla yaklaşık 3 sn aralıklar ile sınıf ortamının nem, sıcaklık, CO_2, hava kalite ve basınç değerleri ölçülmüştür. Camların açılarak sınıfın havalandırıldığı zamanlar durum parametresi altında havalandırma olarak gösterilmiştir. Makine öğrenmesi ve yapay sinir ağları yöntemleri kullanılırken havalandırma durumları HAVALANDIRMA (H) ile, normal şartlar altındaki hava ölçümleri NORMAL (N) ile gösterilmiştir. Yapılan çalışmalar ile sınıftaki hava kalitesinin belirlenmesinde belirleyici faktörün karbondioksit (CO_2) olduğu tespit edilmiştir. Verilerimiz arasındaki sürenin kısa olması sebebiyle ölçülen CO_2 miktarların 3, 5, ve 10'arlı olarak ortalaması alınmıştır. Böylelikle ölçüm değerleri 3 şekilde incelenerek daha doğru tahminlerin yapılması sağlanmıştır. Yapılan ölçümler ile hazırlanan veri seti Tablo 1**:** Veri Seti (Data set)'de gösterilmiştir.

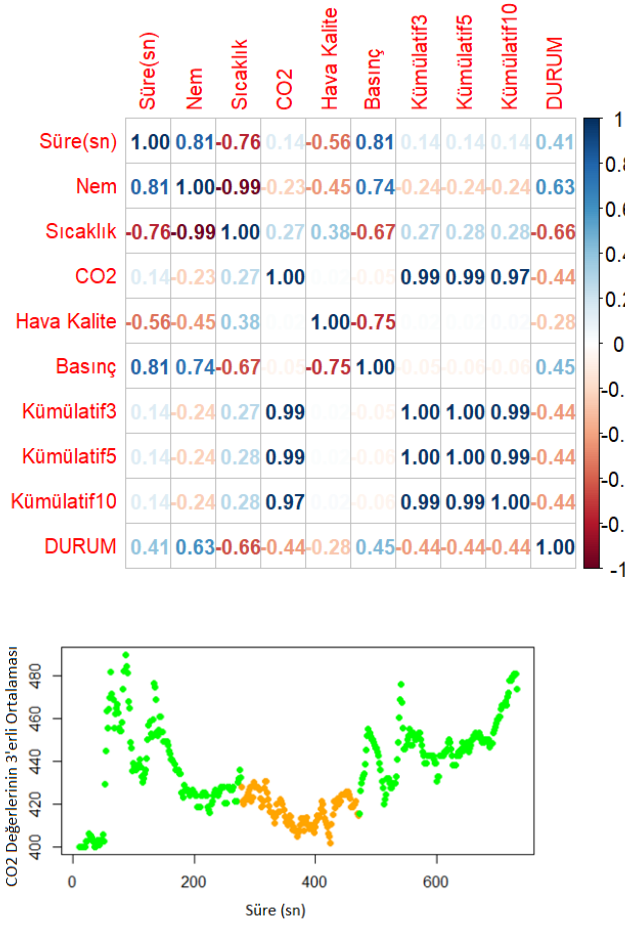
Tablo 1: Veri Seti (Data set)

Veri setindeki nem, sıcaklık, CO2, hava kalitesi, basınç ve durum parametrelerinin birbirleriyle ilişkilerinin gösterildiği korelasyon grafikleri Pearson algoritması kullanılarak R programlama dili ile çizilerek Şekil 4**:** Parametre İlişki Matrisi Grafikleri (Parameter Relationship Matrix Charts) ve Şekil 5**:** Parametre Korelasyon Matrisi (Parameter Correlation Matrix)' te gösterilmiştir.

Şekil 4: Parametre İlişki Matrisi Grafikleri (Parameter Relationship Matrix Charts)



Veri setinde beş parametre arasında en yüksek pozitif ilişki 0.74 değeriyle basınç ve nem miktarı parametreleri arasındadır. Parametreler arasında en düşük negatif ilişki ise -0.99 değeriyle nem ve sıcaklık parametreleri arasındadır. Şekil 4'te beş parametre arasındaki korelasyon değerleri gösterilmiştir.

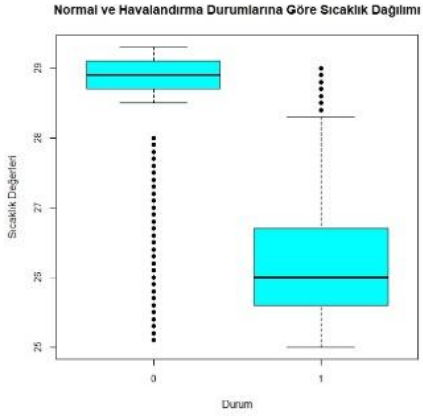


Şekil 5: Parametre Korelasyon Matrisi (Parameter Correlation Matrix)

Yapay sinir ağları, destek vektör makineleri ve karar ağacı gibi makine öğrenmesi yöntemleri kullanılarak havalandırma durumlarının doğru bir şekilde tahmin edilmesi amaçlanmıştır. Tahmin işlemleri için 3128 satırdan oluşan veri setinin %70'i eğitim verisi olarak kullanılmıştır. Tablo 2'de veri setine ait tanımlayıcı istatistik verilmiştir.

Tablo 2: Veri seti Parametrelerin İstatistiksel Değerleri (Statistical Values of Parameters)

	Min	Ortalama	Max	Medyan	Std
Nem	48,7	52,07	59,1	50,8	3,02
Sıcaklık	25	28,8	29,3	28,8	1,36
CO2	400	495,6	1235	495	95,23
Hava Kalite	43	53,19	103	49	7,53
Basınç	1121	1194	1207	1196	10,63
3'lü Ortalama	400	495,5	1158,7	494,3	94,4
5'li Ortalama	400	495,4	1120,6	494,8	93,83
10'lu Ortalama	400,4	495,2	1036,6	495	92,71



Şekil 6: Sıcaklık Parametresinin Havalandırma Durumuna Göre Boxplot Grafikleri (Boxplot Plots of Temprature Parameters by State)

Veri setindeki CO_2 miktarının zamansal olarak değişiminin gösterildiği grafik Şekil 7'de verilmiştir. Grafik üzerinde havalandırma durumları turuncu renk, normal durumlar yeşil ile gösterilmiştir. Grafikte de görüldüğü gibi CO_2 miktarının en düşük olduğu seviyeler yaklaşık olarak havalandırma durumları olmaktadır.

Şekil 7: CO_2 değişimlerinin Nokta Grafiği (Point Graph of CO_2 Changes)

5.2. Havalandırma Tahmini (Ventilation Estimation)

5.2.1. Karar Ağaçları Tahmini (Decision Trees Estimation)

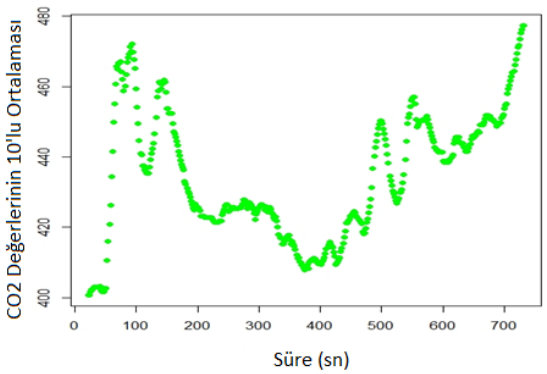
Eğitim verisinde ve test verisinde CO_2 değerlerinin kümülatif olarak toplandığı 3 yeni sütun eklenmiştir. CO_2 değerleri 3'lü, 5'li ve 10'lu olmak üzere 3 farklı sütunda toplanmıştır. Karar ağacı metodu bu üç sütun için ayrı ayrı çalıştırılarak aralarından en iyi tahmin sonucu elde edilmeye çalışılacaktır.

Tablo 1**:** Veri Seti (Data set)'de verilen veri setinde ortamın havalandırıldığı zamanların 1 ile belirtildiği "Durum" sütunu faktör olarak seçilmiştir. R programı bilinmeyen değerlerin (olmayan değerlerin) bulunduğu parametrelerde makine öğrenmesi yöntemleri kullanılırken hata verdiği için CO_2 değerindeki bilinmeyen verilerin silinmesi gerekmektedir. Bu sebeple CO_2 miktarının 3'erli olarak toplandığı 3'lü ortalama sütunu için ilk iki verinin olmamasından dolayı silinmesi gerekmektedir. 5'li ortalama sütunu için ilk 4 veri ve 10'lu ortalama için ilk 9 verinin silinmesi gerekmektedir.

Tablo 6'da görüldüğü gibi CO_2 değerlerinin 3'erli olarak toplandığı 3'lü ortalama sütunu için karar ağacı sonucunda C02 değeri havalandırma durumlarının tahmininde en etkili faktördür. Karar ağacı metodunun uygulanması sonucunda test verisindeki %72,49 doğruluk oranıyla 311 veri normal olarak doğru tahmin edilmiştir. 115 veri havalandırma iken normal olarak ve 3 veri normal iken havalandırma olarak yanlış tahmin edilmiştir. CO_2 değerlerinin 5'erli olarak toplandığı 5'li ortalama için %72,36 doğruluk oranıyla test verisindeki 309 veri normal olarak doğru tahmin edilmiştir. 115 veri havalandırma yapılmış iken normal olarak ve 3 veri normal iken havalandırma var diyerek yanlış tahmin edilmiştir. CO_2 değerlerinin 10'arlı olarak toplandığı 10'lu ortalama sütunu için karar ağacı metodunun uygulanması sonucunda %72,74 doğruluk oranıyla test verisindeki 307 veri normal olarak doğru tahmin edilmiştir. 115 veri havalandırma iken normal olarak yanlış tahmin edilmiştir. Havalandırma olarak hiçbir veri tahmin edilmemiştir

Karar ağacı algoritmasıyla elde edilen en iyi tahmin olan 10'lu ortalama olarak CO2 miktarının zamana göre değişimi Şekil 8'deki grafikte nokta grafiğiyle görselleştirilmiştir. 10'lu ortalama parametresi kullanılarak karar ağacı yöntemiyle havalandırma zamanlarının tahminleri turuncu renkte gösterilmiştir. Fakat karar ağacı metoduyla yapılan tahminde tüm tahminler normal (yeşil) olarak tahmin edilmiştir.

Şekil 8: 10'lu ortalama Parametresi İçin Karar Ağacı Yöntemi ile CO_2 Değişimlerinin Nokta Grafiği (Point Graph of CO_2 Changes with Decision Tree Method for Average of 10 Parameters)



Tablo 3'de uygulanan karar ağacı yöntemlerinin 3'lü, 5'li ve 10'lu ortalama tahminleri ile elde edilen hata metrik performansları yer almaktadır. Ortalama Karesel Hata (Mean Squared Error-MSE) ve doğruluk değerleri tablodaki gibi hesaplanmıştır. Tablodan da görüldüğü gibi en doğru tahmin sonucunun elde edildiği parametre 10'lu ortalama olmuştur.

Tablo 3: Karar Ağacı Yöntemi Hata Metrik Performansları (Decision Tree Method Error According to Different Error Metrics)

	MSE	Doğruluk
3'lü Ortalama	0,275	0,7249
5'li Ortalama	0,2763	0,7236
10'lu Ortalama	0,2725	0,7274

5.2.2. Yapay Sinir Ağları ile Tahmin (Artificial Neural Networks Estimation)

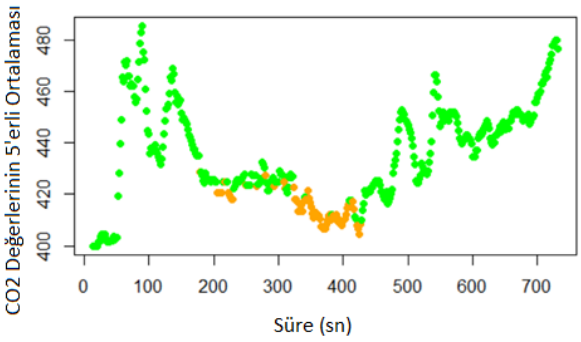
Yapay sinir ağları metodu kullanılarak farklı katmanlara sahip modeller oluşturulmuştur. Eğitim verimizdeki veri sayısı fazla olduğu için mimari 1 ve (2,1) olacak şekilde oluşturulmuştur. Bu modeller arasından en iyi tahmin performansına sahip yöntem seçilecektir.

Yapay Sinir ağları mimarisi ile ilgili olarak kullanılan mimarisi açısından gizli katman (Hidden) olarak hem tek katmanlı ve (2,1) katman kullanılmıştır. Eşik değeri (Threshold) ise 0.03 iken tek tekrarlanma (rep=1) yeterli olmuştur. Başlangıç değerleri (Startweights) önceden belirlenmemiş olup öğrenme oranı olarak (Learningrate) 0.04 belirlenmiştir. Aktivasyon fınksiyonu (act.fct) hiperbolik tanjant, hata fonksiyonu (err.fct) olarak ise hataların karelerinin toplamı (sse) kullanılmıştır.

Tek katmanlı yapay sinir ağı modelinde 3'lü ortalama parametresi için %78.32, 5'li ortalama parametresi için 79.16, 10'lu ortalama parametresi için 78.44 doğruluk oranıyla tahmin yapılmıştır. Aralarından en iyi tahmini yapan model 5'li ortalama parametreli model olmuştur. En iyi model sonucunda test verisindeki 279 veri normal olarak ve 59 veri havalandırma olarak doğru tahmin edilmiştir. 33 veri havalandırma iken normal olarak ve 56 veri normal iken havalandırma olarak yanlış tahmin edilmiştir.

2 gizli katmanlı (2,1) nörondan oluşan yapay sinir ağı modelinde 3'lü ortalama parametresi için %84.15, 5'li ortalama parametresi için %83.37, 10'lu ortalama parametresi için %76.06 doğruluk oranıyla tahmin yapılmıştır. Aralarından en iyi tahmini yapan model 3'lü ortalama parametreli model olmuştur. En iyi model sonucunda test verisindeki 300 veri normal olarak ve 61 veri havalandırma olarak doğru tahmin edilmiştir. 14 veri havalandırma iken normal olarak ve 54 veri normal iken havalandırma olarak yanlış tahmin edilmiştir.

Yapay sinir ağı algoritmasıyla elde edilen 5'li ortalama'li olarak CO2 miktarının zamana göre değişimi Şekil 9 'da nokta grafiğiyle görselleştirilmiştir. 5'li ortalama parametresi kullanılarak yapay sinir ağı yöntemiyle havalandırma zamanlarının tahminleri turuncu renkte gösterilmiştir.



Şekil 9: 5'li ortalama Parametresi İçin 2 Gizli Katmanlı Yapay Sinir Ağı ile CO_2 Değişimlerinin Nokta Grafiği (Point Graph of CO_2 Changes with artificial Neural Netwok for Average of 5 Parameters)

Tablo 4' de uygulanan farklı katmanlı yapay sinir ağları yöntemlerinin 3'lü, 5'li ve 10'lu ortalama tahminleri ile elde edilen hata metrik performansları yer almaktadır. 2 (gizli) katmanlı model tek katmanlı modele göre daha iyi sonuçlar vermiştir. Yapay sinir ağları yöntemiyle en doğru tahmin sonucunun elde edildiği 2 (gizli) katmanlı 3'lü ortalama modeli olmuştur.

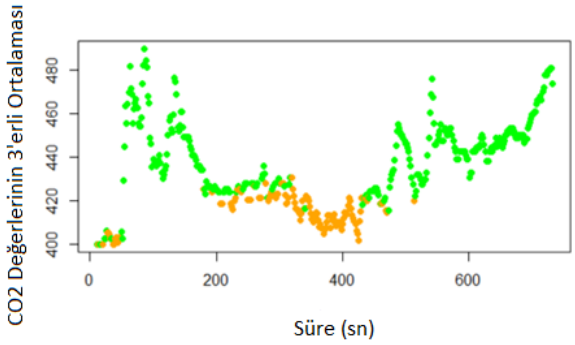
Tablo 4: Tek Katmanlı ve (2,1) Katmanlı Yapay Sinir Ağı Modelinin Farklı Parametrelere Göre Hata Metrikleri (Error Metrics of Single-Layer and (2,1) hidden layer Neural Network Model by Different Parameters)

	MSE	MAPE	Doğruluk
1 Katman			
Yapay Sinir Ağı 3'lü	0,2167	0,1527	0,7832
Yapay Sinir Ağı 5'li	0,2084	0,1428	0,7916
Yapay Sinir Ağı 10'lu	0,2156	0,1492	0,7844
c(2,1) Katman			
Yapay Sinir Ağı 3'lü	0,1585	0,0956	0,8415
Yapay Sinir Ağı 5'li	0,1663	0,1007	0,8337
Yapay Sinir Ağı 10'lu	0,2393	0,2014	0,7606

5.2.3 Destek Vektör Makine ile Tahmini (SVM Estimation)

Destek vektör makine metodu kullanılarak 3'lü, 5'li ve 10'lu ortalama ile farklı lineer ve radyal modeller oluşturulmuştur. Lineer modelinde 3'lü ortalama parametresi için 325 vektör, 5'li ortalama parametresi için 324 vektör, 10'lu ortalama parametresi için 327 vektör oluşturulmuştur. Radyal modelinde 3'lü ortalama parametresi için 420 vektör, 5'li ortalama parametresi için 424 vektör, 10'lu ortalama parametresi için 419 vektör oluşturulmuştur. Aralarından en iyi tahmini yapan model %84,12 doğruluk değeriyle 10'lu ortalama parametreli model olmuştur. Destek vektör metodunun uygulanması sonucunda test verisindeki 284 veri normal olarak ve 71 veri havalandırma olarak doğru tahmin edilmiştir. 23 veri havalandırma iken normal olarak ve 44 veri normal iken havalandırma olarak yanlış tahmin edilmiştir. Elde edilen sonuçlara göre radyal modellerde, lineer modellere göre daha iyi sonuçlar elde edilmiştir. Aralarından en iyi tahmini yapan model %85,88 doğruluk değeriyle 3'lü ortalama parametreli model olmuştur. SVM metodunun uygulanması sonucunda test verisindeki 277 veri normal olarak ve 88 veri havalandırma olarak doğru tahmin edilmiştir. 37 veri havalandırma iken normal olarak ve 27 veri normal iken havalandırma olarak yanlış tahmin edilmiştir. Destek Vektör makinesi algoritmasıyla elde edilen en iyi tahmin olan 3'lü ortalama'lü olarak CO_2 miktarının zamana göre değişimi Şekil 20'de nokta grafiğiyle görselleştirilmiştir. 3'lü ortalama parametresi kullanılarak Radyal yöntem ile havalandırma zamanlarının tahminleri turuncu renkte gösterilmiştir.

Şekil 10:: 3'lü ortalama Parametresi İçin SVM Radyal Model ile CO_2 Değişimlerinin Nokta Grafiği (Point Graph of CO_2 Changes with SVM Radial Model for Average of 3 Parameter)



Tablo 5'de uygulanan destek vektör makine yöntemlerinin 3'lü, 5'li ve 10'lu ortalama tahminleri ile elde edilen hata metrik performansları yer almaktadır. Radyal modellerin da yüksek doğrulukta sonuçlar ürettiği görülmektedir.

Tablo 5: Lineer ve Radyal Modellerin Farklı Parametrelere Göre Hata Metrikleri (Error Metrics of Linear and Radial SVM Models According to Different Parameters)

	Vektör Sayısı	MSE	Doğruluk
Lineer Model 3'lü	325	0,1678	0,8322
Lineer Model 5'li	324	0,1733	0,8267
Lineer Model 10'lu	327	0,1587	0,8412
Radyal Model 3'lü	420	0,1491	0,8588
Radyal Model 5'li	424	0,1451	0,8548
Radyal Model 10'lu	419	0,1587	0,8412

Uygulama bölümünde makine öğrenmesi yöntemleri ile elde edilen tahminlerin karışıklık matrisleri Tablo 6 'da

gösterilmiştir. Elde edilen tahmin sonuçlarının doğruluk yüzdeleri

Tablo 6: Makine Öğrenmesi Yöntemleri ile Elde Edilen Karışıklık Matrisleri (Response Matrices Obtained by Machine Learning Methods)

		Normal (N)	Havalandırma (H)
Karar Ağacı 3'lü	N	311	3
	H	115	0
Karar Ağacı 5'li	N	309	3
	H	115	0
Karar Ağacı 10'lu	N	307	0
	H	115	0
Yapay Sinir Ağı 3'lü	N	276	55
	H	38	60
Yapay Sinir Ağı 5'li	N	279	56
	H	33	59
Yapay Sinir Ağı 10'lu	N	272	56
	H	35	59
(2,1) Yapay Sinir Ağı 3'lü	N	300	54
	H	14	61
(2,1) Yapay Sinir Ağı 5'li	N	297	56
	H	15	59
(2,1) Yapay Sinir Ağı 10'lu	N	238	32
	H	69	83
SVM Lineer Model 3'lü	N	284	42
	H	30	73
SVM Lineer Model 5'li	N	282	44
	H	30	71
SVM Lineer Model 10'lu	N	284	44
	H	23	71
SVM Radyal Model 3'lü	N	277	27
	H	37	88
SVM Radyal Model 5'li	N	275	25
	H	37	90
SVM Radyal Model 10'lu	N	272	32
	H	35	83

5. Sonuç (Conclusion)

İstanbul Medeniyet Üniversitesi zeki kampüs kapsamında yapmış olduğumuz çalışmada havalandırma durumlarının tahmini için geliştirilen yapay sinir ağları, destek vektör makineleri ve karar ağacı modelleri ile karar destek sistemi oluşturulması hedeflenmiştir. Verilerin kapsayıcılığının ve sayısının arttırılması ile makine öğrenmesi modelleri hava kalitesinin tahmin edilmesinde daha iyi sonuçlar vermiş ve makine öğrenmesi metotları ile oluşturulan karar destek sistemleri kabul edilebilir doğrulukta tahmin sonuçları üretmiştir. Havalandırma zamanlarını en iyi tahmin eden makine öğrenmesi yöntemi 0,85 doğruluk ile radyal SVM modeli olmuştur. En kötü tahmin yöntemi ise 0,72 doğruluk oranıyla karar ağacı yöntemi olmuştur. Karar ağacı yöntemi havalandırma durumlarını tespit etmekte zorlanmıştır. Yapay sinir ağları kapsamında oluşturulan ilk mimaride doğruluk oranları 0,74 iken ikinci mimaride 0,84 oranına yükselmiştir. Bu arada farklı yapay sinir ağları yapıları denenerek doğruluk oranlarının artırılabileceği görülmüştür. Mimari kümülatif ortalamalı olarak eklediğimiz 3'lü,5'li ve 10'lu CO2 parametresi havalandırma durumlarının daha doğru bir şekilde tahmin edilmesine katkı sağlamıştır. Çünkü bir zamandaki hava kirliliğini kendinden önce ölçülen CO2 miktarları da etkilemektedir. Ortalama olarak eklenen CO2 değerlerine kendi içinde bakıldığı zaman üç yöntem için de doğruluk oranlarının birbirine yakın olduğu görülmüştür.

Yaptığımız bu çalışma ile makine öğrenmesi teknikleriyle sensör verilerinin modellenmesi ve havada gerçekleşen ani değişikliklerin (odanın havalandırılması, insan sayısının artması gibi) model tarafından tespitini sağlanmıştır. Çalışmanın sonucu olarak kabul edilebilir aralıkların dışına çıkan hava kalitesi durumlarının tespiti sonucunda odanın ne zaman havalandırılacağına karar verilebilmektedir. Bu araştırmada kurduğumuz modeller kampüs dışında toplu taşıma araçları, iş yerleri, ofis, restoran, kafe, özel araçların havalandırma sistemlerinde de kullanılabilecektir. Bu araştırma ile üniversite kampüslerinde virüs bulaşma riski en aza indirilerek güvenli bir şekilde öğrenci ve öğretmenlerin eğitim hayatlarına devam etmesine katkı sağlanmıştır.
Çalışmanın bir sonraki aşamasında kapalı alanlarda yerleştirilecek olan sensörlerin saylarının optimize edilmesi, konumlarının belirlenmesi konularına ağırlık verilecektir. Yöntem olarak ise özellikle sensör yerleşim problemi için metasezgisel yöntemlerden yararlanılacaktır.

Kaynakça (References)

Aksangür, İ. et al. (2022) Evaluation of data preprocessing and feature selection process for prediction of hourly PM10 concentration using long short-term memory models. Environmental Pollution. [Online] 311119973.

Balta, D. (2019) Dağıtık Sensör Sistemleri Mimarisi ile Bulanık Mantık Temelli ve Çevrimiçi Kapalı Ortam Hava Kalitesi İzleme Sistemi Geliştirilmesi.

Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. [Online] 2 (2), .

Chang, C. C. & Lin, C. J. (2011) LIBSVM: A Library for support vector machines. ACM Transactions on Intelligent Systems and Technology. [Online] 2 (3), .

Dokuz, Y. et al. (2020) Hava Kalitesi Parametrelerinin Tahmini ve Mekansal Dağılımı İçin Makine Öğrenmesi Yöntemlerinin Kullanılması. Ömer

Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi. [Online]

Elibol, H. (2021) Kapalı Mekanlarda Sosyal Mesafe Kuralı Ne Kadar Önemli [online]. Available from: https://www.hurriyet.com.tr/gundem/kapali-mekanlarda-sosyal-mesafe-kurali-ne-kadar-onemli-41801992 (Accessed 2 February 2022).

Eren, B. et al. (2023) Predicting next hour fine particulate matter (PM2.5) in the Istanbul Metropolitan City using deep learning algorithms with time windowing strategy. Urban Climate. [Online] 48101418.

Gültepe, Y. (2019) Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine Karşılaştırmalı Bir Değerlendirme. European Journal of Science and Technology. [Online] 8–15.

Gupta, Y. et al. (2021) Impact of Weather Predictions on COVID-19 Infection Rate by Using Deep Learning Models. Complexity. [Online] 2021.

Irmak, M. E. & Aydilek, İ. B. (2019) Hava Kalite İndeksinin Tahmin Başarısının Artırılması için Topluluk Regresyon Algoritmalarının Kullanılması. Academic Platform Journal of Engineering and Science. [Online] 507–514.

Karakuş, C. B. & Yıldız, S. (2019) Hava Kalite İndeksi İle Meterolojik Parametreler Arasındaki İlişkinin Çoklu Regresyon Yöntemi İle Belirlenmesi. Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi. [Online]

Kulkarni, P. S. et al. (2017) Artificial Neural Networks for Construction Management: A Review. Journal of Soft Computing in Civil Engineering. [Online] 1 (2), .

Lelieveld, J. et al. (2020) Model calculations of aerosol transmission and infection risk of covid-19 in indoor environments. International Journal of Environmental Research and Public Health. [Online] 17 (21), 1–18.

Mo, X. et al. (2019) A novel air quality early-warning system based on artificial intelligence. International Journal of Environmental Research and Public Health. [Online] 16 (19), .

Sakhıdad Faızı (2021) Kabil'in Hava Kalitesi Tahmininde Makine Öğrenmesi Algoritmalarının Kullanılması.

Santos, A. F. et al. (2020) Best Practices on HVAC Design to Minimize the Risk of COVID-19 Infection within Indoor Environments. Brazilian Archives of Biology and Technology. [Online] 631–11.

Zhang, C. et al. (2021) The role of weather conditions in COVID-19 transmission: A study of a global panel of 1236 regions. Journal of Cleaner Production. [Online] 292.

Research Article
Journal of Intelligent Systems: Theory and Applications 6(2) (2023) 191-198
DOI: 10.38016/jista.1250782

Self Adaptive Methods for Learning Rate Parameter of Q-Learning Algorithm

Murat Erhan Çimen[1*], Zeynep Garip[2], Yaprak Yalçın[3], Mustafa Çağrı Kutlu[4], Ali Fuat Boz[5]

[1,5] Sakarya University Of Applied Sciences, Department of Electric and Electronic Engineering, Sakarya, Türkiye

[2] Sakarya University Of Applied Sciences, Department of Computer Engineering, Sakarya, Türkiye

[3] Istanbul Technical University, Department of Control and Automation Engineering, İstanbul, Türkiye

[4]Sakarya University Of Applied Sciences, Department of Mechatronics Engineering, Sakarya, Türkiye

muratcimen@subu.edu.tr, zbatik@subu.edu.tr, yalciny@itü.edu.tr, mkutlu@subu.edu.tr, afboz@subu.edu.tr

Abstract
Machine learning methods can generally be categorized as supervised, unsupervised and reinforcement learning. One of these methods, Q learning algorithm in reinforcement learning, is an algorithm that can interact with the environment and learn from the environment and produce actions accordingly. In this study, eight different on-line methods have been proposed to determine online the value of the learning parameter in the Q learning algorithm depending on different situations. In order to test the performance of the proposed methods, these algorithms are applied to Frozen Lake and Car Pole systems and the results are compared graphically and statistically. When the obtained results are examined, Method 1 has produced better performance for Frozen Lake, which is a discrete system, while Method 7 has produced better results for the Cart Pole System, which is a continuous system.
Keyword: Reinforcement Learning, Q learning, Machine Learning

Q-Learning Algoritmasının Öğrenme Hızı Parametresi için Kendine Uyarlamalı Yöntemler parametresi

Öz
Makine öğrenmesi yöntemleri genel olarak denetimli, denetimsiz ve takviyeli öğrenme olarak sınıflandırılabilir. Bu yöntemlerden biri olan takviyeli öğrenme içerisinde bulunan Q learning algoritması ortamla etkileşime girerek ortamdan öğrenebilen ve ona göre aksiyonlar üretebilen bir algoritmadır. Bu çalışmada Q learning algoritması içerisinde bulunan öğrenme parametresinin değeri için 8 farklı yöntem önerilmiştir. Önerilen yöntemlerin performanslarının test edilebilmesi için donmuş göl ve ters sarkaç sistemlerine bu algoritmalar uygulanmış ve sonuçları grafiksel ve istatistiksel olarak karşılaştırılmıştır. Elde edilen sonuçlar incelendiğinde ayrık bir sistem olan Donmuş Göl sistemi için Metot 1 daha iyi performans sergilerken sürekli bir sistem olan Ters Sarkaç Sistemi için Metot 7 daha iyi sonuç göstermiştir.

Anahtar kelimeler: Takviyeli Öğrenme, Q Learning, Makine Öğrenmesi

1. Introduction

Machine learning methods, a sub-branch of artificial intelligence, have many application areas today (Angiuli, Fouque, and Laurière 2022). Machine learning methods can produce the most appropriate results in the face of new situations by analysing the sensors on the system or the data sources given to it before (Grefenstette n.d.). Especially in recent years, the development of computer, software and information systems along with technology has enabled artificial intelligence and machine learning to be widely used in fields such as economy (Jogunola et al. 2020; Meng and

* Corresponding Author.
E-mail: muratcimen@subu.edu.tr

Recieved : 03 Mar 2023
Revision : 31 Aug 2023
Accepted : 06 Sep 2023

Khushi 2019; Sarızeybek and Sevli 2022), medicine (Bayraj et al. 2022; Cimen et al. 2021; Pala et al. 2019, 2021, 2022), biology, chemistry, informatics (Ekinci 2022; Omurca et al. 2022; Toğaçar, Eşidir, and Ergen 2021) and engineering (Akyurek and Bucak 2012; Bucak and Zohdy 1999; Chen et al. 2022; Çimen et al. 2019; Singh, Kumar, and Singh 2022). Machine learning methods can generally be grouped as Supervised Learning, Unsupervised Learning and reinforcement learning. These structures are shown in Figure 1.

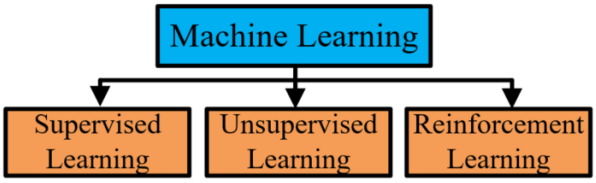


Figure 1. Machine learning classification

The Supervised Learning method is to create a function that establishes a cause-effect relationship between input and output and to learn this function (Cunningham, Cord, and Delany 2008). Supervised learning is often used a lot in classification and regression. Unsupervised, on the other hand, allows learning the existing relationships in the data. In this method, inferences are made according to the distances, densities, and neighbourhood relations in the data. Unsupervised learning is especially used in clustering, that is, in separating data into each other or in size reduction by removing unnecessary variables from the data (Barlow 1989; Sathya and Abraham 2013). Reinforcement learning, on the other hand, is inspired by the behaviour of living and non-living beings in nature. The action of an agent in any situation in the environment by interacting with the environment causes a new state to occur. It is based on the fact that the agent learns the next behaviour that he will perform in an environment with a new situation, with a reward or punishment value. The agent tries to choose the best action he can take to achieve his goal. Thus, the goal of the agent interacting with the environment is to learn the sequence of movements that produce the greatest total reward (Angiuli et al. 2022; Peng and Williams 1996; Watkins 1989). Therefore, here the algorithm learns how to react according to the determined reward and punishment. This structure is given in Figure 2.

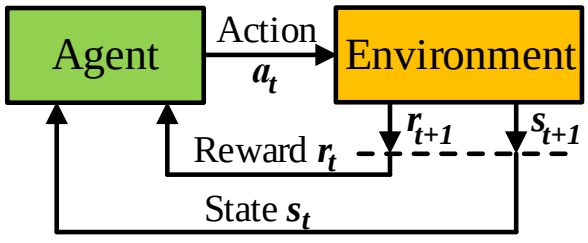


Figure 2. Reinforcement learning approach.

Model-free (model-independent) reinforcement learning is a type of learning that utilizes the Q-learning approach (Watkins and Dayan 1992). When using agents identified using the Q-learning approach, users may avoid having to map out the Markovian spaces in order to learn how to behave best there (Watkins and Dayan 1992). Instead, users can learn by experiencing the results of their choices. The application of these learning algorithms is widespread, and they may be utilized in a wide range of industries and fields, including marketing (Jogunola et al. 2020), finance (Meng and Khushi 2019), time sequence estimate (O'Neill et al. 2010), robot control (Singh et al. 2022), and control of autonomous vehicles (Elallid et al. 2022).

In this study, 8 different online-tuning method are proposed for the learning parameter of the q learning algorithm. The q learning algorithm is applied on Frozen Lake and Cart Pole Systems, and performances of the q learning algorithm are compared based on the cases where the learning parameter is constant, changes depending on iteration, and changes depending on the reward. It has been seen that Method 1 has produced better results for Frozen Lake and Method 7 has produced better results for Car Pole than other methods.

The structure of the paper as follows: In the second section, some preliminary information on reinsforment learning is given. In the third section, the proposed on-line tuning methods and application of them for Frozen Lake and Cart Pole Systems are presented. In the fourth section, the simulation results are depicted. Finaly, in fifth section, some conluding remark are given.

2. Preliminaries

Reinforcement learning (RL) is a method for solving sequential decision-making issues in a variety of domains in the natural and social sciences, as well as engineering, by having an agent interact with the environment and learning an optimum policy via trial and error (Angiuli et al. 2022; Smart and Kaelbling 2000; Wang, H., Emmerich, M., & Plaat n.d.). In reinforcement learning methods, learning is usually carried out over Q-table (Wang, H., Emmerich, M., & Plaat n.d.). There are many methods for learning this table, such as dynamic optimization, monte Carlo, Q-learning, and SARSA (Akyurek and Bucak 2012; Candan et al. 2048; Peng and Williams 1996). In the structure given in Figure 3, there is an environment, for instance the Frozen Lake, in which the agent and agent can move. The agent performs an action (a_t) in evironment (Frozen Lake) according to the information it has (s_t, a_t). The action performed by the agent (a_t) causes the agent's state in the environment to change (s_{t+1}) and this change will also create a reward (r_{t+1}). As a result of its interaction with the environment, the agent starts to learn an environment by using values such as ($s_t, a_t, s_{t+1}, r_t,$). In this study, Q Learning algorithm will be implemented over this learning Q-table.

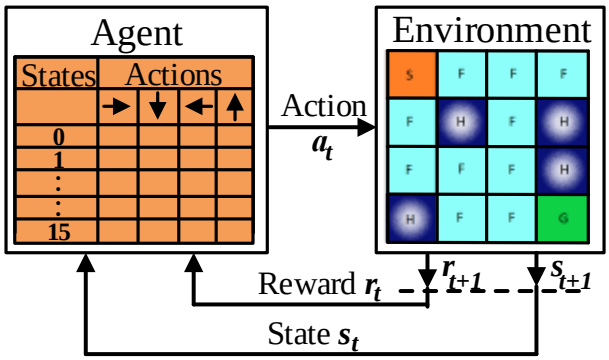


Figure 3. Q Table for Frozen Lake Game

Bellman Equation used in updating the Q table used in Figure 3 is the formula expressed by Equation 1. In Equation 1, state at time t obtained from s_t environment, the action that a_t agent will take in the environment, the reward obtained at time t as a result of the action of r_t agent, the new state obtained from the environment at time $t+1$ as a result of the action of s_{t+1} agent, α learning factor, γ is the reduction factor. The expression $\underset{a}{max}\left(Q_t(s_{t+1},a)\right)$ provides the highest value for any action in s_{t+1} state. This approach, called on-policy, constantly updates the Q table in interaction with the environment. The psoudecode of Q-Learning is given in Algorithm 1.

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha\left(r_t + \gamma \underset{a}{max}\left(Q_t(s_{t+1},a)\right) - Q_t(s_t,a_t)\right) \quad (1)$$

Algorithm 1. Q learning Psoudecode

Input:
1: State (s)
2: Action (a_t)
3: Learning rate (α)
4: Discount factor (γ)
5: Reward $R(s_t,a_t)$
6: Updated table $Q(s_t,a_t)$
Output:
7: Selected action according to updating table $Q(s_t,a_t)$
For episode 1, M **do**
Initialise state s_t
For t=1, T **do**
Choose a_t with ϵ greedy probability
Execute a_t and observe state s_{t+1} and reward r_t
Update table $Q_{t+1}(s_t,a_t)$ with equation 1
End for
End for

3. Main Methods and Results

3.1. Tuning Methods for Learning Parameter

In this study, different methods have been proposed according to the learning parameter of the Q learning algorithm. Instead of α parameter given in Equation 1, the use of μ parameter in Equation 2 is preferred. The reason for this is to avoid the confusion that the changing parameter will create with the action (a_t) variable.

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \mu_t\left(r_t + \gamma \underset{a}{max}\left(Q_t(s_{t+1},a)\right) - Q_t(s_t,a_t)\right) \quad (2)$$

Different methods have been proposed according to the variation between equations 3-11. In order to distinguish the method according to the equations used, nomenclature was made between Method 1 and Method 9. When Equation 3 is used for the change of α from these methods, the parameter used is β_1 and its value is chosen as 0.01. When this equation 3 is used, this method is named as Method 1. Similarly, when Equation 4 is used, β_2 is constant and its value is chosen as 0.05. This method, in which Equation 4 is used, is named as Method 2. Equation 5 and Equation 6 are used to change the learning parameter depending on iteration. The use of Equation 5 is named Method 3. In Method 3, the learning factor is reduced depending on the iteration. The use of Equation 6 is named Method 4. In Method 4, the value of the learning factor increases depending on the iteration. β_3=0.04, β_4=0.05 used in Method 3 and Method 4 are used as parameters. On the other hand, the positive change of the learning parameter depending on the changing value of the state of being in the Q table is modeled in Equation 7. This method is given as Method 5. Similarly, its negative change is given in Equation 8 and named as Method 6. The parameter of Method 5 and Method 6 is $\beta_5 = 0.05$. With an approach similar to Method 5 and Method 6, depending on the increase and decrease of change, the learning parameter was modelled as in Equations 9 and Equation 10 and named as Method 7 and Method 8. In Method 7 and Method 8, $\beta_6 = 0.005$. In addition, Equation 11 is used to constrain the μ_t parameter in Method 5, Method 6, Method 7 and Method 8. In Equation 11, the parameters are selected as $\beta_7 = 0.001$, $\beta_8 = 0.01$.

$$\mu_{t+1} = \beta_1 \quad (3)$$

$$\mu_{t+1} = \beta_2 \quad (4)$$

$$\mu_{t+1} = \beta_1 + \beta_3\left(1 - \frac{t}{t_{max}}\right) \quad (5)$$

$$\alpha_{t+1} = \beta_1 + \beta_3\left(\frac{t}{t_{max}}\right) \quad (6)$$

$$\mu_{t+1} = \mu_t - \beta_5\left(Q_t(s_{t+1},a) - Q_t(s_t,a_t)\right) \quad (7)$$

$$\mu_{t+1} = \mu_t + \beta_5\left(Q_t(s_{t+1},a) - Q_t(S_t,A_t)\right) \quad (8)$$

$$\mu_{t+1} = \begin{cases} \mu_t + \beta_6 & Q_t(s_{t+1}, a_{t+1}) \geq Q_t(s_t, a_t) \\ \mu_t - \beta_6 & other \end{cases} \quad (9)$$

$$\mu_{t+1} = \begin{cases} \mu_t - \beta_6 & Q_t(s_{t+1}, a_{t+1}) \geq Q_t(s_t, a_t) \\ \mu_t + \beta_6 & other \end{cases} \quad (10)$$

$$\mu_{t+1} = \begin{cases} \beta_7 & \beta_7 < \mu_t \\ \mu_t & \beta_7 \leq \mu_t \leq \beta_8 \\ \beta_8 & \mu_t > \beta_8 \end{cases} \quad (11)$$

3.2. Application to Frozen Lake

Frozen Lake is an environment designed for an agent moving on a frozen lake to reach its desired destination (Goal). This environment is shown in Figure 4. In simple terms, there are 4 different $A = (\leftarrow, \rightarrow, \uparrow, \downarrow)$ movement abilities that the agent can move in this game. The agent acts depending on its location. Each position it moves corresponds to a state. Therefore, the moving agent provides transition from one state to another. In the map given in Figure 4, S: safe, F: frozen, H: hole and G is goal. While the agent is moving on the ice, he tries to reach the Goal without coming to the Hole. If the agent starting from S reaches the G point with his actions, then reward 1 is rewarded as reward value.

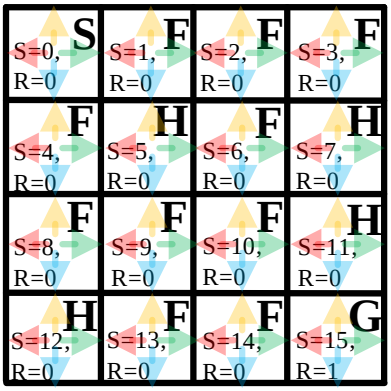


Figure 4. Frozen Lake

A sample Q table obtained when the Q table is trained by applying the Q learning algorithm to the Frozen Lake game is obtained as in Table 1. When the initial parameters of the training number change, the values of this table change, especially when the number of iterations increases, the changes in the table have decreased.

Table 1. Q Table for Frozen Lake

State Number	Action, Action Number (a)			
	$\leftarrow, 0$	$\downarrow, 1$	$\rightarrow, 2$	$\uparrow, 3$
0	5.13e-2	5.01e-1	5.11e-1	5.09e-2
1	3.63e-1	3.106e-1	3.68e-1	4.8e-1
2	4.32e-1	4.34e-1	4.18e-1	1.45e-1
13	4.63e-1	5.52e-2	6.53e-1	4.75e-1
14	7.28e-1	8.42e-2	7.91e-1	7.75e-1
15	0	0	0	0

The agent uses the Q Table that it learns by interacting with the environment, and when the learning phase ends in the next steps, it chooses his actions based on the value with the highest state of being value in the relevant state.

3.3. Application to Cart Pole

One of the most common systems used to test the validity of any proposed method is the Cart pole system (Cimen and Yalçın 2022). Since the Cart Pole system is non-linear in nature, it is the most commonly used basic system for testing a new controller in control systems (Adigüzel and Yalçin 2018; Adıgüzel and Yalçın 2022). The structure of this system is given in Figure 5. The mathematical model of the system is given in Equation 12 (Barto, Sutton, and Anderson 1983). The parameters used in the mathematical model are also given in Table 2 (Barto et al. 1983), and the Sampling Time (T_s) is taken as 0.02 sec with the discretization Euler Method.

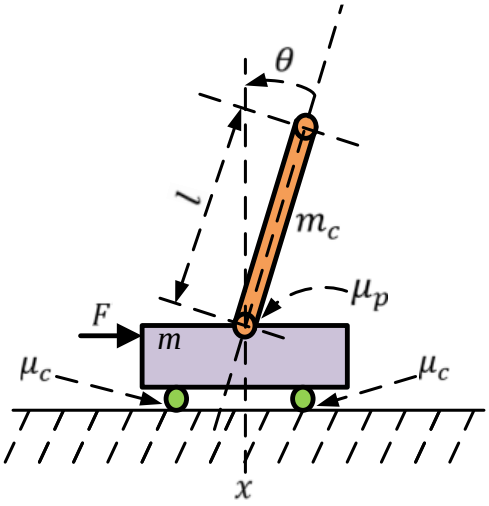


Figure 5. Cart Pole Sistemi

$$\ddot{\theta} = \frac{\cos(\theta)\left[\frac{-F - ml\dot{\theta}^2 sin(\theta) + \mu_c sgn(\dot{x})}{m + m_c}\right]}{l\left[\frac{4}{3} - \frac{mcos^2(\theta)}{m + m_c}\right]} + \frac{gsin(\theta) - \frac{\mu_p \dot{\theta}}{ml}}{l\left[\frac{4}{3} - \frac{mcos^2(\theta)}{m + m_c}\right]} \quad (12)$$

$$\ddot{x} = \frac{F + ml\left[\dot{\theta}^2 sin(\theta) - \ddot{\theta}cos(\theta)\right]}{m + m_c} - \frac{\mu_c sgn(\dot{x})}{m + m_c}$$

Table 2. Parameter of Cart Pole System

Parameter	Value
Gravity (g)	$9.8\,{}^{m}/_{s^2}$
Mass of cart (m_c)	$1\ kg$
Mass of pole (m)	$0.1\ kg$
length of half-pole (l)	$0.5\ m$
coefficient of friction of cart (μ_c)	0.0005
coefficient of friction of pole (μ_p)	0.000002
Force applied to cart's center of mass (F)	$\pm$10.0 N

Since Q learning algorithm runs discretely, the control signal to be used and the situations to be observed must be discrete. In this case, the action space for the Cart pole system is given in Table 3 and the observation space is given in Table 4.

Table 3. Action space

Action	Space	Action Number (a)
Push cart to the left	-10	0
Push cart to the right	10	1

Table 4. Observation Space

Action	Space
Cart Position (x)	$-4.8 < x < 4.8$
Cart Velocity ($\dot{x}$)	$-\infty < \dot{x} < \infty$
Pole Angle (θ)	$-0.418 < \theta < 0.418$
Pole Angle Velocity ($\dot{\theta}$)	$-\infty < \dot{\theta} < \infty$

At Table 4, action space for the cart pole is $A = (-10,10)$. Also, It is expressed as state $S = (x\ \dot{x}, \theta, \dot{\theta})$ in the Cart Pole system. However, the system state is continuous. To adapt this to the q learning algorithm, the action space obtained when dividing into 10 parts for the parameters $-2.4 < x < 2.4, -4 < \dot{x} < 4, -0.2095 < \theta < 0.2095, -4 < \dot{\theta} < 4$ is as in Table 5. The Q table obtained as a result of these transformations is given in Table 6.

Table 5. Observation Space for Cart Pole System for 10 discrete value

State Number	$(x, \dot{x}, \theta, \dot{\theta})$	Space
0	(0,0,0,0)	$(-2.4, -4, -0.2095, -4)$
1	(0,0,0,1)	$(-2.4, -4, -0.2095, -3.1)$
⋮	⋮	⋮
1742	(1,3,4,4)	$(-1.8, -2.2, -0.06, -2.2)$
1743	(1,3,4,5)	$(-1.8, -2.2, -0.06, -1.3)$
⋮	⋮	⋮
14640	(10,10,10,9)	(2.4,4,0.2095,3.1)
14641	(10,10,10,10)	(2.4,4,0.2095,4)

Table 6. Q table for Cart Pole System

State Number	Action Number	
	0	1
0	0	0
1	0	0
⋮	⋮	⋮
1742	5.50	6.01
1743	9.14	12.28
⋮	⋮	⋮
14640	0	0
14641	0	0

In this case, the reward value to be used is calculated as in Equation 13. In addition, the done function is given in Equation 14 to stop the system under certain conditions.

$$reward = \begin{cases} 1 & (-2.4 < x < 2.4)\ and \\ & (-0.2095 < \theta < 0.2095) \\ 0 & other \end{cases} \quad (13)$$

$$done = \begin{cases} & (-2.4 < x < 2.4)\ or \\ 0 & (-0.2095 < \theta < 0.2095)\ or \\ & iteration < 200 \\ 1 & other \end{cases} \quad (14)$$

4. Simulation Studies

In this study, the proposed methods for the Q learning algorithm were carried out on a computer with Intel(R) Core(TM) i5-9400 CPU @ 2.90GHz, 64 Bit, 8GB RAM. The study was carried out using Anaconda IDE. In addition, tests were performed on Frozen Lake and Cart Pole environments using the pygym library. Method 1- Method 8 methods proposed for Q learning algorithm have been trained for 30000 iterations. Each method was run independently 20 times for statistical comparison. The suggested methods were applied for each system and the results were explained in graphs and tables. In addition, the best values in the tables are written in bold font.

The average values of the results produced by the Q learning algorithm are shown in the graphs. When Figure 6 is examined for the Frozen Lake system, the values obtained by Method 1 during 30000 iterations are shown in blue in the graph. In addition, the average value obtained in the last 100 steps using these values is shown in green. The statistical results of this system are calculated as in Table 7. When Method 1 was examined, it was calculated as a minimum of 0, a maximum of 1, an average of 0.35, and a standard deviation of 0.47 for the average value in the last 100 steps. In addition, in Figure 6, the variation of the learning parameter examined in this study is given in each iteration. However, since it is constant for Method 1, it appears to be constant. Similarly, Method 2 results are demonstrated as in Figure 6 graphically. Statistically, it is given in Table 7. The results of Method 3 and Method 4 are depicted graphically in Figure 6. Statistically the results are calculated as in Table 7. The results of Method 5 and Method 6 are depicted graphically in Figure 8. Statistically the results are also given in Table 7. The results of Method 7 and Method 8 are depicted graphically in Figure 9. Statistically the results are also calculated as in Table 7. When Table 7 was evaluated numerically, all methods produced the best results in

terms of maximum value. Method 5 produced the best results in terms of average value, and Method 1 produced the best results in terms of average value over the last 100 steps.

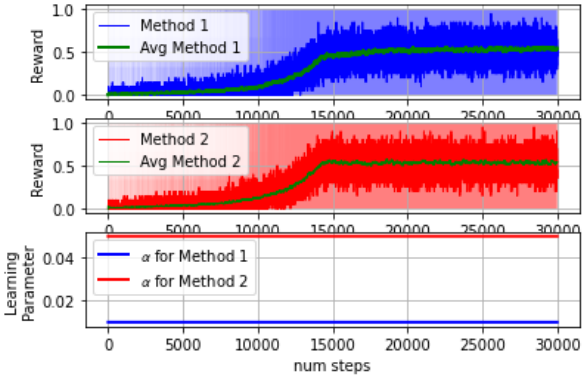


Figure 6. Reward and learning parameter results of Method 1, Method 2 for Frozen Lake

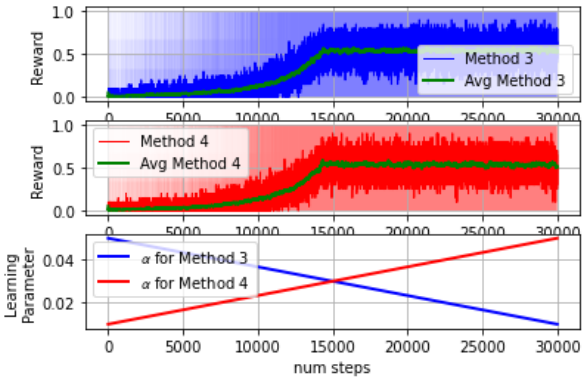


Figure 7. Reward and learning parameter results of Method 3, Method 4 for Frozen Lake

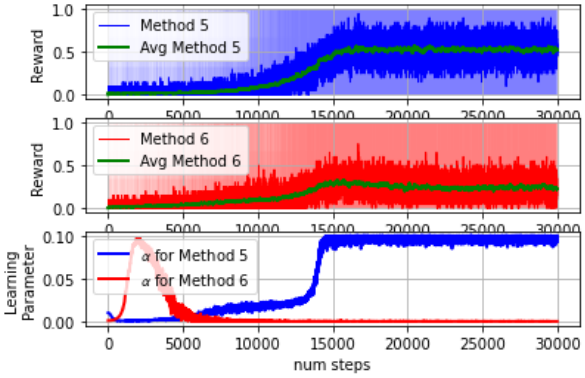


Figure 8. Reward and learning parameter results of Method 5, Method 6 for Frozen Lake

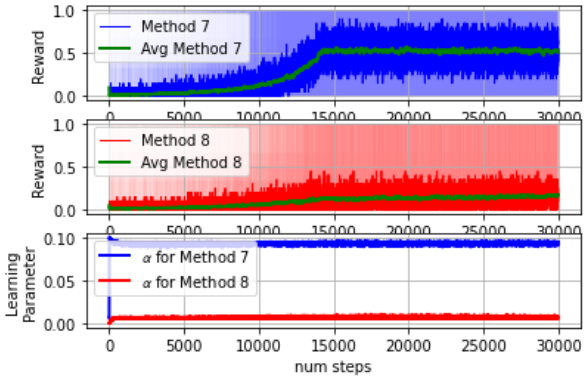


Figure 9. Reward and learning parameter results of Method 7, Method 8 for Frozen Lake

Table 7. Statistical Results of Method 1-Method 8 for Frozen Lake

	min	max	avg	avg_100	std
Method 1	**0**	**1**	0.35	**0.55**	0.47
Method 2	**0**	**1**	0.45	0.54	0.49
Method 3	**0**	**1**	0.45	0.54	0.49
Method 4	**0**	**1**	0.50	0.51	0.50
Method 5	**0**	**1**	**0.65**	0.53	0.47
Method 6	**0**	**1**	0.15	0.23	0.35
Method 7	**0**	**1**	0.35	0.52	0.47
Method 8	**0**	**1**	0.10	0.17	0.30

The average values of the results produced by the Q learning algorithm are shown in the graphs. When Figure 10 is examined for the Cart Pole system, the values obtained by Method 1 during 30000 iterations are shown in blue in the graph. In addition, the average value obtained in the last 100 steps using these values is shown in green. The statistical results of this system are given in Table 8. When Method 1 is examined, it is calculated that the minimum 118, the maximum 200, the average 161.3, the average value in the last 100 steps is 153.97, and the standard deviation is 28.27. In addition, in Figure 10, the variation of the learning parameter examined in this study is given in each iteration. However, since it is constant for Method 1, it appears to be constant. Similarly, Method 2 results are depicted in Figure 9 graphically. Statistically, it is given in Table 8. The results of Method 3 and Method 4 are depicted graphically in Figure 11. Statistical results are also calculated as in Table 8. The results of Method 5 and Method 6 are depicted graphically in Figure 12. Statistically the results are also given in Table 8. The results of Method 7 and Method 8 are depicted graphically in Figure 13. Statistically the results are also calculated as in Table 8. Considering Table 8 numerically, Method 7 produced the best result in terms of maximum, Average value, average value over the last 100 steps as avg_100 in Table 8 are given.

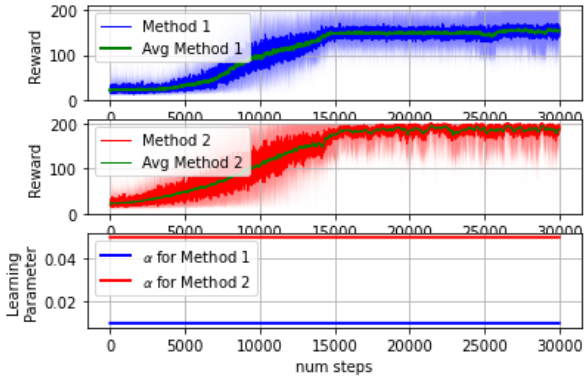


Figure 10. Reward and learning parameter results of Method 1, Method 2 for Cart Pole

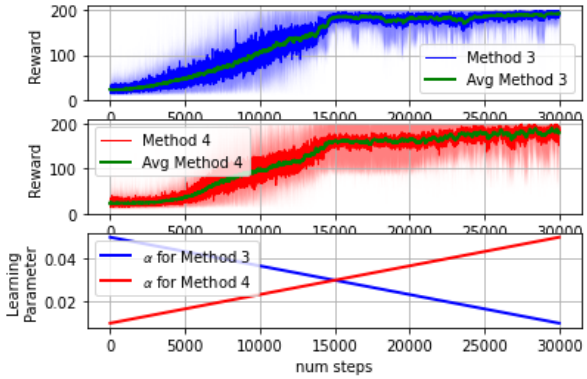


Figure 11. Reward and learning parameter results of Method 3, Method 4 for Cart Pole

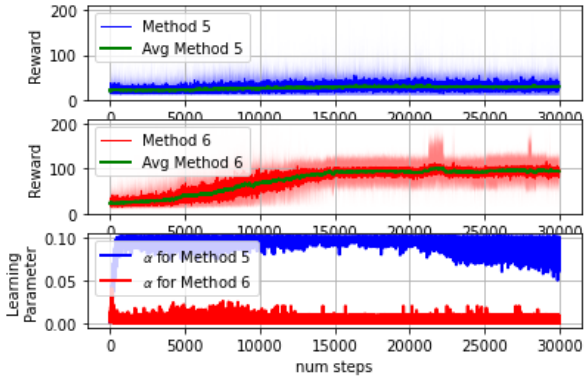


Figure 12. Reward and learning parameter results of Method 4, Method 5 for Cart Pole

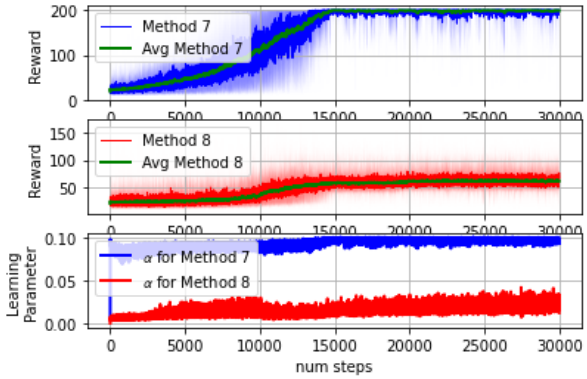


Figure 13. Reward and learning parameter results of Method 7, Method 8 for Cart Pole

Table 8. Statistical Results of Method 1-Method 8 for Cart Pole System

	min	max	avg	avg_100	std
Method 1	118	**200**	161.3	153.97	28.27
Method 2	169	**200**	196.6	187.85	9.24
Method 3	163	**200**	193.4	187.78	13.10
Method 4	102	**200**	182	187.72	32.4
Method 5	11	75	40	187.65	17.62
Method 6	57	137	93.5	187.52	25.59
Method 7	**200**	**200**	**200**	**187.65**	**0**
Method 8	49	100	67.3	187.65	15.65

5. Conclusions

In this study, 8 different methods have been proposed for the learning parameter of the Q learning algorithm. The proposed methods have been applied to the Frozen Lake system, which is a discrete system, and the Cart Pole System, which is continuous time. The proposed 8 methods have been applied to these systems over 30000 iterations. Each method has been run independently 20 times and their performances were tested statistically. When the results obtained are examined, it is seen that Method 1 have produced better results for Frozen Lake system, which is a discrete system, while Method 7 have produced better results for a discrete system, Cart Pole.

References

Adigüzel, F., Yalçin, Y., 2018. Discrete-Time Backstepping Control for Cart-Pendulum System with Disturbance Attenuation via I&i Disturbance Estimation. in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*.

Adıgüzel, F., Yalçin, Y., 2022. "Backstepping Control for a Class of Underactuated Nonlinear Mechanical Systems with a Novel Coordinate Transformation in the Discrete-Time Setting." in *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*.

Akyurek, H.A., Bucak İ.Ö., 2012. Zamansal-Fark, Uyarlanır Dinamik Programlama ve SARSA Etmenlerinin Tipik Arazi Aracı Problemi Için Öğrenme Performansları. in *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*. Trabzon.

Angiuli, A., Fouque J.P., Laurière M., 2022. Unified Reinforcement Q-Learning for Mean Field Game and Control Problems. *Mathematics of Control, Signals, and Systems* 34(2):217–71.

Barlow, H. B., 1989. Unsupervised Learning. *Neural Computation* 1(3).

Barto, A. G., Sutton R.S., Anderson C.W., 1983. Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. *IEEE Transactions on Systems, Man, and Cybernetics* 5(834–846).

Bayraj, E. A., Kırcı, P., Ensari, T., Seven, E., Dağtekin, M., 2022. Göğüs Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin Uygulanması. *Journal of Intelligent Systems: Theory and Applications* 5(1):35–41.

Bucak, I.Ö., Zohdy M. A., 1999. Application Of Reinforcement Learning Control To A Nonlinear Bouncing Cart. Pp. 1198–1202 in *Proceedings of the American Control Conference*. San Diego, California.

Candan, F., Emir, S., Doğan, M., Kumbasar, T., 2018. Takviyeli Q-Öğrenme Yöntemiyle Labirent Problemi Çözümü Labyrinth Problem Solution with Reinforcement Q-Learning Method. in *TOK2018 Otomatik Kontrol Ulusal Toplantısı*.

Chen, T., Chen, Y., He, Z., Li, E., Zhang, C., Huang., Y., 2022. A Novel Marine Predators Algorithm with Adaptive Update Strategy. *He Journal of Supercomputing* 1–34.

Çimen, M.E., Garip, Z. Pala M.A., Boz, A.F., Akgül, A. 2019. Modelling of a Chaotic System Motion in Video with Artificial Neural Networks. *Chaos Theory and Applications* 1(1).

Cimen, M.E., Yalçın, Y., 2022. A Novel Hybrid Firefly–Whale Optimization Algorithm and Its Application to Optimization of MPC Parameters, *Soft Computing* 26(4):1845–72.

Cimen, M.E., Boyraz, O.F., Yildiz, M.Z., Boz, A.F., 2021. A New Dorsal Hand Vein Authentication System Based on Fractal Dimension Box Counting Method, *Optik* 226.

Cunningham, P., Cord, M. Delany, S.J., 2008. Supervised Learning, Pp. 21–49 in *Machine learning techniques for multimedia: case studies on organization and retrieval*,.

Ekinci, E., 2022. Classification of Imbalanced Offensive Dataset–Sentence Generation for Minority Class with LSTM, *Sakarya University Journal of Computer and Information Sciences* 5(1):121–33.

Elallid, B. B., Benamar, N., Hafid, A. S., Rachidi, T., Mrani, N., 2022. A Comprehensive Survey on the Application of Deep and Reinforcement Learning Approaches in Autonomous Driving, *Journal of King Saud University-Computer and Information Sciences*.

Grefenstette, J. J., 1993. Genetic Algorithms and Machine Learning, in *Proceedings of the sixth annual conference on Computational learning theory*.

Jogunola, O., Adebisi, B., Ikpehai, A., Popoola, S. I., Gui, G., Gačanin, H., Ci. S., 2020. Consensus Algorithms and Deep Reinforcement Learning in Energy Market: A Review, *IEEE Internet of Things Journal* 8(6).

Meng, T. L., Khushi, M., 2019. Reinforcement Learning in Financial Markets, *Data* 4(3).

O'Neill, D., Levorato, M., Goldsmith, A., Mitra U., 2010. Residential Demand Response Using Reinforcement Learning, in *2010 First IEEE International Conference on Smart Grid Communications*.

Omurca, S. İ., Ekinci, E., Sevim, S., Edinç, E. B., Eken, A., Sayar, S., 2022. A Document Image Classification System Fusing Deep and Machine Learning Models, *Applied Intelligence* 1–16.

Pala, M. A., Çimen, M. E., Boyraz, Ö. F., Yildiz, M. Z., Boz, A., 2019. Meme Kanserinin Teşhis Edilmesinde Karar Ağacı Ve KNN Algoritmalarının Karşılaştırmalı Başarım Analizi, *Academic Perspective Procedia* 2(3).

Pala, M.A., Cimen, M.E., Yıldız, M.Z. Cetinel, G., Avcıoglu, E., Alaca, Y., 2022. CNN-Based Approach for Overlapping Erythrocyte Counting and Cell Type Classification in Peripheral Blood Images, *Chaos Theory and Applications* 4(2).

Pala, M.A., Cimen, M.E., Yıldız, M.Z. Cetinel, G., Özkan, A.D., 2021. Holografik Görüntülerde Kenar Tabanlı Fraktal Özniteliklerin Hücre Canlılık Analizlerinde Başarısı, *Journal of Smart Systems Research* 2(2):89–94.

Peng, J., Williams. R.J., 1996. *Incremental Multi-Step Q-Learning*.

Sarızeybek, A. T., Sevli, O., 2022. Makine Öğrenmesi Yöntemleri Ile Banka Müşterilerinin Kredi Alma Eğiliminin Karşılaştırmalı Analizi. *Journal of Intelligent Systems: Theory and Applications* 5(2):137–44.

Sathya, R., Abraham., A., 2013. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification, in *(IJARAI) International Journal of Advanced Research in Artificial Intelligence*,.

Singh, B., Kumar, R., Singh., V. P., 2022. Reinforcement Learning in Robotic Applications: A Comprehensive Survey, *Artificial Intelligence Review* 1–46.

Smart, W.D., Kaelbling, L.P., 2000, Practical Reinforcement Learning in Continuous Spaces. *ICML*.

Toğaçar, M., K. A. Eşidir, and B. Ergen. 2021. "Yapay Zekâ Tabanlı Doğal Dil İşleme Yaklaşımını Kullanarak İnternet Ortamında Yayınlanmış Sahte Haberlerin Tespiti." *Journal of Intelligent Systems: Theory and Applications* 5(1):1–8.

Wang, H., Emmerich, M., Plaat, A., Monte Carlo Q-Learning for General Game Playing, *ArXiv Preprint ArXiv:1802.05944*.

Watkins, C. J. C. H., 1989. Learning from Delayed Rewards, Dissertation, King's College UK.

Watkins, C.J.C.H, Dayan P., 1992. Q-Learning, *Machine Learning*.