Journal of the Faculty of Engineering and Architecture of Gazi University 39:4 (2024) 2631-2642

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Prediction of seasonal bike rental counts using a GBM model optimized with bat algorithm

Kadir İleri*
Department of Electrical and Electronics Engineering, Faculty of Engineering and Natural Sciences, Bandirma Onyedi Eylul University, 10200, Bandırma, Balıkesir, Türkiye

Highlights:

- Prediction of bike rental counts using DT, KNN, MLP, and GBM methods
- Optimization of machine learning methods with the bat algorithm
- Analysis of features influencing the prediction of bike rental counts

Keywords:

- Bat algorithm
- Gradient Boosting Machine
- K-Nearest Neighbors
- Multi-Layer Perceptron
- Bike rental counts

Article Info:
Research Article
Received: 18.09.2023
Accepted: 06.01.2024

DOI:
10.17341/gazimmfd.1362302

Correspondence:
Author: Kadir İleri
e-mail: *kileri@bandirma.edu.tr
phone: +90 506 468 3401

Graphical/Tabular Abstract

In this study, the prediction of bike rental counts on a seasonal basis has been carried out. The Gradient Boosting Machine (GBM) method has been employed for the prediction process. Moreover, the performance of the GBM model was optimized with the bat algorithm (BA). As given in Table A, the performance of this proposed model has been compared with different methods such as Decision Tree (DT), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). Furthermore, the features that have the greatest and least impact on the prediction of bike rental counts have also been identified.

Table A. The performance results of machine learning models for each season and yearly

	Winter		Spring		Summer		Autumn		Yearly	
	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
DT	2.1430	0.6265	2.8274	0.7795	2.7862	0.7610	2.4846	0.7838	3.0833	0.8409
KNN	2.8853	0.2827	4.2114	0.6093	4.6861	0.4704	3.5214	0.5879	4.5455	0.6799
MLP	2.1692	0.6065	3.9248	0.6914	3.4496	0.7325	3.1552	0.7182	4.5736	0.7128
GBM	1.2439	0.8532	1.9476	0.8673	2.2337	0.8399	1.9390	0.8467	1.9881	0.9168
BA-GBM	1.2611	0.8555	1.9586	0.8826	2.1748	0.8497	1.7152	0.8672	1.8665	0.9264

Purpose:
The aim of this study is to identify the method that accurately predicts seasonal bike rental counts in the best possible way. Additionally, the study aims to determine which feature has the highest and lowest impact on the prediction results.

Theory and Methods:
Four different methods such as GBM, DT, KNN, and MLP have been used for the prediction of bike rental counts. KNN is a machine learning algorithm used for classification and regression tasks. It has a simple and intuitive structure, and it bases its predictions on the majority (classification) or the average (regression) of the nearest data points in the feature space. DT is one of the supervised learning algorithms. This algorithm is based on a binary tree data structure and is used to solve both regression and classification problems as well. MLP is a type of artificial neural network widely used in machine learning applications. It has a total of three layers: the input layer, the hidden layer(s), and the output layer. Finally, GBM is a popular machine learning algorithm used for tasks such as classification and regression. In boosting algorithms, the fundamental approach is to iteratively combine several simple models (weak learners) to create a strong learner with improved predictive accuracy. Moreover, the performance of the GBM were optimized with the BA which is an optimization algorithm for using regression tasks. Finally, four different metrics, namely MAE (Mean Absolute Error), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 (R-squared), have been used to evaluate the performance of the models.

Results:
The performance of the proposed BA-GBM model has been compared with different methods such as DT, KNN, MLP, and unoptimized GBM. Among these models, the best performance has been achieved by BA-GBM with the value of 0.9264 R^2. Moreover, the features that have the greatest impact on bike rental counts are the time of day and temperature, while the features with the least impact are snowfall and year.

Conclusion:
The accurate prediction of urban bike demand has become a necessity for effective resource allocation for shared bicycles. This prediction process has been conducted using the GBM algorithm which is optimized with the BA. To demonstrate the effectiveness of the model, the performance of the proposed model has been compared with different methods such as DT, KNN, MLP, and unoptimized GBM. Additionally, the features that have the least and the most impact on predicting bike rental counts have been identified. In future studies, the proposed model can be further optimized using other optimization techniques and compared the performances.

Journal of the Faculty of Engineering and Architecture of Gazi University 39:4 (2024) 2631-2642

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Yarasa algoritması ile optimize edilmiş GBM modeli kullanarak mevsim bazlı bisiklet kiralama sayılarının tahmini

Kadir İleri*

Bandırma Onyedi Eylül Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, 10200, Bandırma, Balıkesir, Türkiye

ÖNEÇIKANLAR

- DT, KNN, MLP ve GBM yöntemleri bisiklet kiralama sayılarının tahmini
- Yarasa algoritması ile makine öğrenmesi yöntemlerinin optimizasyonu
- Bisiklet kiralama sayılarının tahminine etki eden özelliklerin analizi

Makale Bilgileri

Araştırma Makalesi
Geliş: 18.09.2023
Kabul: 06.01.2024

DOI:
10.17341/gazimmfd.1362302

Anahtar Kelimeler:
Yarasa algoritması,
Gradyan artırmalı makinesi,
K-En yakın komşu,
çok katmanlı algılayıcı,
bisiklet kiralama sayıları

ÖZ

Kentsel bisiklet talebinin etkili kaynak tahsisi için, paylaşımlı bisiklet kiralama sayılarının gerçekçi bir şekilde tahmin edilmesi gerekmektedir. Bu tahmin işlemi, Yarasa Algoritması (YA) ile optimize edilen Gradyan Artırmalı Makinesi (GBM) yöntemi kullanılarak gerçekleştirilmiştir. Önerilen modelin etkinliğini göstermek amacıyla, modelin performansı Karar Ağacı (DT), K-En Yakın Komşu (KNN) ve Çok Katmanlı Algılayıcı (MLP) gibi farklı yöntemlerle karşılaştırılmıştır. Bu karşılaştırma işlemi için MAE, MSE, RMSE ve R^2 metrikleri kullanılmıştır. En iyi sonuç 1,8665 MAE, 2,9588 MSE, 8,7545 RMSE ve 0,9264 R^2 değerleri ile YA-GBM tarafından elde edilmiştir. Bununla birlikte, bisiklet kiralama sayısının tahminine en fazla ve en az etki eden özellikler de belirlenmiştir. En fazla etkiye sahip özellikler hava sıcaklığı ve günün saati olurken, en az etkiye sahip özellikler ise kar yağışı ve yıl olmuştur.

Prediction of seasonal bike rental counts using a GBM model optimized with bat algorithm

HIGHLIGHTS

- Prediction of bike rental counts using DT, KNN, MLP, and GBM methods
- Optimization of machine learning methods with the bat algorithm
- Analysis of features influencing the prediction of bike rental counts

Article Info

Research Article
Received: 18.09.2023
Accepted: 06.01.2024

DOI:
10.17341/gazimmfd.1362302

Keywords:
Bat algorithm,
gradient boosted machine,
k-Nearest neighbors,
multi-layer perceptron,
bike rental counts

ABSTRACT

To ensure effective resource allocation for urban bike demand, it is crucial to accurately predict shared bike rental counts. This prediction process was carried out using the Gradient Boosted Machine (GBM) method optimized with the Bat Algorithm (BA). To demonstrate the effectiveness of the proposed model, its performance was compared with different methods such as Decision Tree (DT), k-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). For this comparison, metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) were employed. The best results were achieved by BA-GBM with values of 1.8665 MAE, 2.9588 MSE, 8.7545 RMSE, and 0.9264 R^2. Additionally, the features with the most and least impact on bike rental prediction were identified. The most influential features were found to be temperature and time of day, while the least influential features were snowfall and year.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *kileri@bandirma.edu.tr / Tel: +90 506 468 3401

1. Giriş (Introduction)

Bisiklet kullanımı çoğu ülkede günlük hayatın bir parçası haline gelmiştir ve son zamanlarda giderek artan bir kullanım oranına sahiptir [1,2]. Bunun yanında, bisiklet kullanımın beraberinde getirdiği birçok fayda mevcuttur. Bunlar; kentsel trafik baskısını hafifletme, çevre dostu bir seyahat seçeneği sunma, sağlıklı bir yaşam sağlama vb. olarak sayılabilir. Bisiklet kullanım ihtiyacının artması bisiklet kiralama şirketlerinin açılmasına ve paylaşımlı bisiklet kullanım oranının artmasına sebep olmuştur. Ancak paylaşımlı bisikletlerin popülaritesinin artması, rastgele park etme ve arz-talep dengesizliğinden kaynaklanan sorunları da beraberinde getirmiştir. Bu durum, paylaşımlı bisikletlerin etkili tahsisi için kentsel bisiklet talebinin gerçekçi bir şekilde tahmin edilmesi gerekliliğini ortaya koyar. Ayrıca, insanların bisiklet talebi, hava koşulları ve mevsim gibi karmaşık faktörlerden etkilenir ve bu faktörler kesin talep tahmini için kritik rol oynarlar [3].

Bisiklet kiralama sayılarının tahmini üzerine yapılan bir çalışmada, Qi vd. [4] Kalifornia Üniversitesi'nin makine öğrenmesi veri tabanından elde ettikleri verileri kullanarak kiralık bisiklet kullanımını analiz etmişlerdir. Analiz sonucunda, bisiklet kullanımını etkileyen 20 faktörü incelemişler ve bu faktörlerden 8 tanesinin bisiklet kiralama sayısı ile yüksek korelasyona sahip olduğunu tespit etmişlerdir. Geliştirdikleri K-En Yakın Komşu (KNN: K-Nearest Neighbors) modelini, Karar Ağacı (DT: Decision Tree), Destek Vektör Regresyonu (SVR: Support Vector Regression), Doğrusal Regresyon (LR: Linear Regression) ve Derin Sinir Ağları (DNN: Deep Neural Network) modelleri ile karşılaştırmışlar ve KNN'nin en yüksek tahmin doğruluğa sahip olduğu sonucuna varmışlardır. Benzer olarak, Feng vd. [5] Washington şehrinde bisiklet kiralama talebinin tahminine yönelik bir çalışma yapmışlardır. Elde ettikleri verilerin özelliklerine göre, bisiklet kiralama talebini öngörmek için Çoklu Doğrusal Regresyon (MLR: Multiple Linear Regression) analizi ve Rastgele Orman (RF: Random Forest) yöntemi olmak üzere iki farklı yöntem kullanmışlardır. MLR'nin performansının çok düşük olduğunu tespit edip, bu modelin bisiklet kiralama talebi tahmini için uygun olmadığı sonucuna ulaşmışlardır. Bu yönteme alternatif olarak RF modeli geliştirmişler ve bisiklet kiralama talebi tahmininin doğruluğu önemli ölçüde arttırmışlardır. Önerilen bu model, başlangıçtaki MLR modelinden önemli ölçüde daha yüksek olarak %82 doğruluk oranını ulaşmıştır. Başka bir çalışmada, Shiao vd. [6] özellik seçiminin tahmin yapma doğruluğu üzerindeki etkisinin, tahmin modelini tasarlama kadar çok olduğunu göstermişlerdir. Önişleme yapılmadan veri setinin kullanılması, yapılan tahmin sonucuna kötü yönde etki ettiği sonucuna varmışlardır. Ayrıca, SVR, RF gibi farklı makine öğrenimi modellerini ve SVR ile RF'nin bir araya getirildiği kombinasyonlarını karşılaştırmışlardır. Bir araya getirdikleri yöntem diğerlerinden daha iyi performans göstermiştir.

Ayrıca makine öğrenmesi yöntemleri farklı alanlarda da regresyon işlemi için kullanılmaktadır [7-15]. Bu çalışmalardan birinde, Heidari vd. [7] nano sıvıların göreceli viskozitesini tahmin etmek için Çok Katmanlı Algılayıcı (MLP: Multi Layer Perceptron) yöntemine dayalı bir model geliştirmişlerdir. Kapsamlı literatür araştırması ile farklı nano akışkanların bağıl viskozitesine ilişkin 1490 deneysel veri örneği toplamışlardır. Modelin girdi parametreleri; sıcaklık, nano partikül boyutu, nano partikül hacim kesri, nano partikül yoğunluğu ve akışkan viskozitesidir. Önerilen MLP modelinin mimarisi, ilk gizli katmanda 24 nöron ve ikinci gizli katmanda 14 nörondan oluşmaktadır. Benzer olarak, Yatim vd. [8], biyokütle atık değerlendirmesi için önemli olan biyokütle yüksek ısıl değerini tahmin etmek için MLR ve MLP modellerini uygulamışlardır. Bu iki modeli karşılaştırmak için ortalama mutlak hata, ortalama mutlak yüzde hatası, kök ortalama karesel hata ve korelasyon katsayısı metriklerini kullanmışlardır. Sonuç olarak, MLP modelinin biyokütle yüksek ısıl değer tahmininde MLR'den daha iyi performans gösterdiğini tespit etmişlerdir. Başka bir çalışmada, Nsangou vd. [9], elektrik tüketimi sürücülerini değerlendirmek için nicel regresyon, DT ve MLP metotlarını kullanmışlar ve bu metotların performanslarını karşılaştırmışlardır. Kamerun'daki elektrik tüketimi araştırmasından elde edilen verileri kullanarak, konut yapısı, ekipman kullanımı, hane geliri, ikamet edilen yer, hava koşulları ve enerji tasarrufu davranışı gibi faktörlerin elektrik tüketimini önemli ölçüde etkilediğini tespit etmişlerdir. MLP, nicel regresyon ve DT modellerine kıyasla üstün verimlilik sergilemiştir. Thamarai [10] bir bölgedeki ev fiyatlarını tahmin eden bir model geliştirmiştir. Bu çalışma, evlerin özellikleri veya öznitelikleri kullanılarak yapılan bir çalışmadır. Çalışmada kullanılan veriler, Andhra Pradesh'in West Godavari bölgesindeki küçük bir kasabadan elde edilmiştir. Elde edilen veriler, evde bulunan yatak odası sayısı, evin yaşı, konumdan ulaşım imkanı, evlere yakın bulunan okul imkanı ve evin konumuna yakın bulunan alışveriş merkezleri gibi bilgileri içerir. Ev fiyatlarını tahmin etmek için DT ve MLR olmak üzere iki algoritma kullanmışlardır. DT regresyonuna göre MLR regresyonun, ev fiyatlarını tahmin etmede daha iyi performans gösterdiğini tespit etmişlerdir. Bir diğer çalışmada, Di vd. [11] güneybatı Çin'de bulunan Sichuan şehrindeki enkaz akıntısı risk yönetimi için bir Gradyan Artırma Makinesi (GBM: Gradient Boosting Machine) modeli geliştirmişlerdir. Bu model için, 1949-2017 yılları arasında Sichuan Jeo-Çevre İzleme programı, saha araştırmaları ve uydu görüntü yorumlamasıyla derlenen toplam 3839 enkaz akıntısı olayı verisi kullanmışlardır. Önerilen GBM modelini, LR, KNN, SVR ve Yapay Sinir Ağı (ANN: Artificial Neural Network) modelleri ile karşılaştırarak üstünlüğünü göstermişlerdir. GBM tarafından üretilen duyarlılık haritası, yüksek duyarlılık gösteren havza bölgeleri ile topoğrafik aşırılıklar, fay hatları, sismik bölgeler ve kurak vadilerin konumları arasında güçlü bir mekânsal ilişki olduğunu göstermiştir. Bu çalışma, enkaz akıntılarının risk azaltması ve önlenmesi için kritik bilgiler sunmakla birlikte etkili risk yönetimi stratejilerine katkı sağlamaktadır. Son olarak, Zhu vd. [12] betonun basınç dayanımının tahmin modelini oluşturmak için GBM ve rastgele orman (RF) temelli iki model önermişlerdir. Bu tahmin modelleri, ilgili karışım oranı için deneme-formülasyon çalışmalarında yardımcı olmak ve deneme-formülasyon çalışmalarındaki maliyeti azaltmak amacıyla önerilmektedir. Bu modeller, 1030 örnekten oluşan veri kümesi üzerinde çimento, su ve iri agreganın 8 özelliğini kullanarak beton karışım oranına karşılık gelen basınç dayanımını tahmin etmektedir. Elde ettikleri tahmin sonuçlarına göre GBM modeli, RF modelinden üstün gelmiştir. Akköse vd. [13] makine öğrenmesi yöntemleri ile tipik bir ortaokul binasının enerji tüketimini analiz etmişlerdir. Pencere özelliklerine dayalı iyileştirme senaryolarının etkinliğini değerlendirmişlerdir. Çalışmada sistemli bir simülasyon tabanlı yaklaşım izlenmiştir. İklim değişikliğinin etkileri gözetilerek bir vaka çalışması binası için pencere parametrelerine dayalı 2025 iyileştirme senaryosu geliştirilmiştir. Simülasyonlar aracılığıyla üretilen veri seti Rastgele Orman modeli ile eğitilmiş ve tahminler gerçekleştirmiştir. Yapılan testler sonucunda pencerenin Güneş Isı Kazancı Katsayısı'nın en kritik parametre olduğu ortaya çıkmıştır. Gülmez vd. [14] Türkiye'deki ikinci el araç fiyatlarının tahminini gerçekleştirdikleri bir çalışma gerçekleştirmişlerdir. Bu tahmin işleminde, araçların sahip olduğu marka, model yılı, yakıt türü gibi özelliklerin fiyat tahmini üzerindeki etkileri analiz edilmiştir. Doğrusal Regresyon, İzotonik Regresyon, Karar Ağacı, Rastgele Orman, Gradyan Arttırılmış Ağaç Regresyonu metotlarının performanslarını karşılatırmışlardır. En yüksek doğruluktaki fiyat tahmini Rastgele Orman metodu ile elde edilmiştir. Acı vd. [15] 2019-2020 dönemini kapsayan iki yıllık internet alışverişi satış verileri ile TÜFE, işsizlik oranı ve tatil günleri gibi verileri bir araya getirerek oluşturulan veri seti üzerinde ürün satış miktarını tahmin etmeyi amaçlamışlardır. Bu ürün talep tahmini için

Yapay Sinir Ağları, Derin Öğrenme, Gauss Süreç Regresyonu, Karar Ağacı, Destek Vektör Makineleri ve Topluluk Öğrenme gibi çeşitli yöntemler kullanılmış ve elde edilen sonuçlar detaylı bir şekilde karşılaştırılmıştır. En etkili sonuçlar Derin Öğrenme yöntemleri arasında yer alan Uzun/Kısa Süreli Bellek Ağları ile elde edilmiştir. Geliştirilen talep tahmin modellerinde alışveriş verisi dışındaki faktörlerin (tatil günleri, TÜFE değeri ve işsizlik oranı) model performansına olan etkisi ölçülmüş ve bu faktörlerin hepsinin başarıya önemli katkı sağladığı gözlemlenmiştir.

Bu çalışmada, mevsim bazında bisiklet kiralama sayılarının tahmini gerçekleştirilmiştir. Tahmin işlemi için YA ile optimize edilmiş GBM metodu kullanılmıştır ve önerilen bu modelin performansı farklı metotlar ile karşılaştırılmıştır. Bu çalışmamın temel katkıları ise şöyledir:

- GBM metodu ile yıllık ve her mevsim için ayrı ayrı bisiklet kiralama sayıları tahmin edilmiştir.
- GBM modelini performansı YA ile optimize edilerek arttırılmıştır.
- Önerilen modelin üstünlüğünü göstermek için performansı MLP, DT ve KNN metotları ile karşılaştırılmıştır.
- Bisiklet kiralama sayısının tahminine en az ve en çok etkisi olan özellikler yıllık ve her mevsim için ayrı ayrı belirlenmiştir.
- Önerilen modelin performansı literatürdeki aynı veri setini kullanan diğer modeller ile karşılaştırılmış ve daha üstün bir sonuç elde edilmiştir.

Bu makalenin geri kalanı aşağıdaki şekilde yapılandırılmıştır: Bölüm 2, kullanılan veri seti hakkında bilgi vermektir. Bölüm 3, çalışmada kullanılan makine öğrenmesi yöntemlerini açıklamaktadır. Elde edilen deneysel sonuçların ayrıntılı bir analizi ve tartışması Bölüm 4'te sunulmuştur. Son olarak, Bölüm 5, çalışmanın sonuçlarını özetlemektedir.

2. Veri Seti (Dataset)

Bu çalışmada kullanılan bisiklet kiralama veri seti 8760 örnekten oluşmaktadır [16]. Örnekler, Güney Kore'nin Seul şehrinde 2017 ve 2018 yılları arasını kapsayan 365 günlük süre boyunca toplanmıştır. Veri seti, farklı hava koşullarında (sıcaklık, nem, rüzgar hızı, görüş mesafesi, çiy noktası, güneş radyasyonu, kar yağışı ve yağmur yağışı) saatlik bisiklet kiralama sayılarını içermektedir. Tablo 1'de veri setine ait birkaç örnek gösterilmiştir.

Veri önişleme, bir makine öğrenme modelinin performansını artırmak için eğitim veri kümesini oluşturan özellikleri seçme, dönüştürme ve oluşturma işlemlerini içerir. Bu ön işlemler makine öğrenmesi modellerinin performanslarını arttırabilmektedir [17]. Bu çalışmada gerçekleştirilen veri önişleme aşaması 5 aşamada gerçekleştirilmiştir.

İlk aşamada, veri setindeki faal gün özelliğinin değerinin "Hayır" olduğu durumdaki bisiklet kiralama sayılarının toplamı sıfırdır. Dolayısıyla, bu özellik veri setinden çıkartılmıştır.

İkinci aşamada, tarih özelliğindeki verileri sayısal olmayan formattadır. Bu özellik; gün, ay ve yıl olmak üzere 3 ayrı özelliğe bölünmüş ve verileri sayısal formata çevrilmiştir.

Üçüncü aşamada, özellikler arasındaki ilişki (korelasyon) incelenmiştir. Bu ilişki, istatiksel bir ölçü olup özellikler arasındaki bağların ne kadar güçlü olduğunu gösterir. Şekil 1'de verildiği üzere, sıcaklık ve çiy noktası özelliklerinin yüksek korelasyona (yaklaşık 0,90 oranında) sahip oldukları açıkça görülmektedir. Bu çoklu doğrusallık (bağımsız değişkenler arasındaki yüksek korelasyonun, bağımlı değişken üzerindeki etkilerini ayrı ayrı tahmin etmeyi zorlaştırması) sorununa neden olacağı için, çiy noktası özelliği veri setinden çıkartılmıştır.

Dördüncü aşamada, veri setindeki sayısal formatta olmayan özellikler incelenmiştir. Bu özellikler; mevsimler ve tatil'dir. Bu özelliklerinin değerleri, Etiket Kodlayıcı (Label Encoder) [18] kullanılarak sayısal formata dönüştürülmüştür. Tatil günleri 1 ile gösterilirken, tatil olmayan günler 0 ile gösterilmiştir. Benzer şekilde, mevsimler 0 (kış), 1 (ilkbahar), 2 (yaz) ve 3 (sonbahar) olacak şekilde kodlanmıştır.

Son aşamada ise, özelliklerin veri dağılımları analiz edilmiştir. Şekil 2'de görüldüğü üzere, bisiklet kiralama sayısı ve rüzgar hızı özellikleri sağa çarpık olarak dağılımlıdır.

Bu iki özelliğin çarpık olan dağılımları, karekök (square root transformation) dönüşümü ile düzeltilmiştir. Düzenleme yapılmadan ve yapıldıktan sonraki dağılımlımlar Şekil 3'te gösterilmiştir.

Şekil 4 dört farklı mevsim boyunca saatlik bisiklet kiralama sayısını göstermektedir. Veri, bisiklet kullanımının yaz mevsiminde zirve yaptığını, bunu yakından takip eden sonbahar mevsimi olduğunu göstermektedir. İlkbahar mevsimindeki bisiklet kullanım miktarı sonbahardakine benzerdir, ancak bisiklet kiralama sayıları kış mevsiminde belirgin şekilde düşmektedir, bunun nedeni önemli ölçüde düşük sıcaklıklardır. Bu, kiralık bisiklet kullanımı ile hava koşulları arasındaki güçlü ilişki olduğunu göstermektedir.

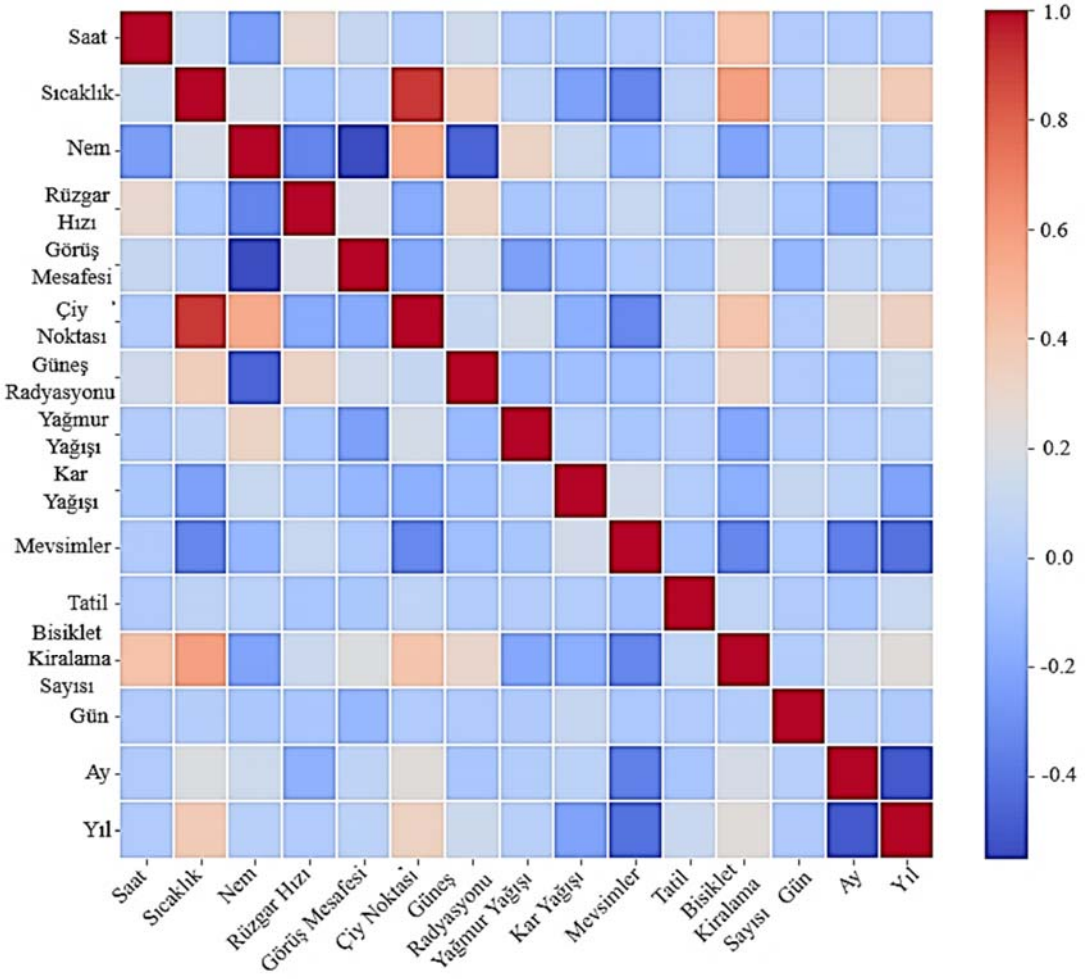
3. Metotlar (Methods)

3.1. K-En Yakın Komşu Algoritması (K-Nearest Neighbors Algortihm)

K-En Yakın Komşu (KNN: K-Nearest Neighbors), sınıflandırma ve regresyon görevleri için kullanılan bir makine öğrenmesi algoritmasıdır [19]. Algoritma oldukça basit ve sezgisel bir yapıya

Tablo 1. Kullanılan veri setinden birkaç örnek (A few examples from the dataset used)

Özellikler	Değerler							
Tarih	2.12.2017	1.01.2018	21.01.2018	5.03.2018	15.04.2018	10.05.2018	4.06.2018	1.08.2018
Saat	3	12	11	11	16	10	22	14
Sıcaklık	-3,5	1,5	1,9	6,1	12	17,3	22	38,7
Nem	81	20	25	66	61	52	64	38
Rüzgar Hızı	2,2	2,8	1	1,9	3,6	2,3	1,8	2,4
Görüş Mesafesi	1221	1992	1493	2000	1374	1235	1211	1943
Çiy Noktası	-6,2	-19	-16,1	0,2	4,6	7,3	14,8	21,8
Güneş Radyasyonu	0	1,03	0,78	1,42	0,57	2,38	0	3,12
Yağmur Yağışı	0	0	0	0	0	0	0	0
Kar Yağışı	0	0	0	0	0	0	0	0
Mevsimler	Kış	Kış	Kış	İlkbahar	İlkbahar	İlkbahar	Yaz	Yaz
Tatil	Değil	Tatil	Değil	Değil	Değil	Değil	Değil	Değil
Faal Gün	Evet	Evet	Evet	Evet	Evet	Hayır	Evet	Evet
Bisiklet Kiralama Sayısı	167	225	216	322	124	0	2000	475



Şekil 1. Özelliklere ait korelasyon matrisi (Correlation heatmap of features)

sahiptir ve tahminlerini özellik uzayındaki en yakın veri noktalarının çoğunluğuna (sınıflandırma) veya ortalamasına (regresyon) dayandırır.

Regresyon işlemi için bu algoritmada önce k parametresi seçilir. Bu parametre, tahminler yaparken dikkate alınacak en yakın komşu sayısını temsil eder. Verilen bir giriş veri noktası için, algoritma giriş noktası ile eğitim kümesindeki tüm veri noktaları arasındaki mesafeyi hesaplar. Sonra, en kısa mesafeye sahip olan k tane veri noktası, en yakın komşular olarak seçilir. Son olarak, giriş veri noktası için tahmin edilen değer genellikle k tane komşunun hedef değerlerinin ortalaması olarak alınır.

3.2. Karar Ağacı Algoritması (Decision Tree Algorithm)

Karar ağacı (DT: Decision Tree), denetimli öğrenme algoritmalarından biridir. Bu algoritma, ikili ağaç veri yapısına dayanır ve hem regresyon hem de sınıflandırma problemlerini çözmek için kullanılır [20]. Bir karar ağacı, kök düğüm, iç düğüm ve yaprak düğüm olmak üzere üç tür düğümden oluşur. Ağaç bir kök düğümle başlar, farklı iç düğümlerle genişler ve yaprak düğümlerle sonlanır. Düğümler, veri kümesinin karakteristik özelliklerini temsil ederken, düğüm dalları ağacın karar kurallarını temsil eder. Yaprak düğümleri nihai karar çıktılarını temsil eder. Eğer ağaç sınıflandırma için kullanılıyor ise çıktılar sınıfların bir listesi iken, regresyon için kullanılıyor ise sürekli değişkenlerdir.

3.3. Çok Katmanlı Algılayıcı Algoritması
(Multi Layer Perceptron Algorithm)

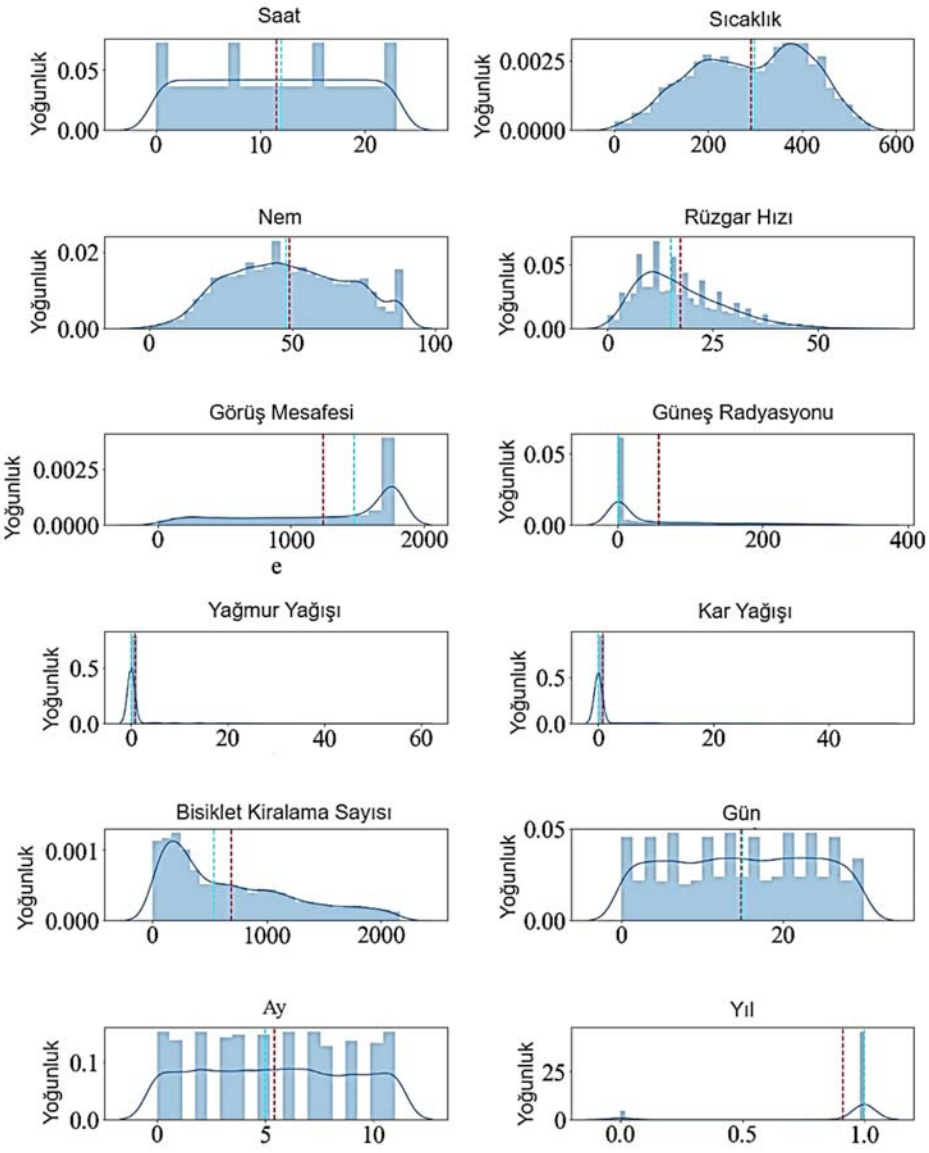
Çok Katmanlı Algılayıcı (MLP: Multi Layer Perceptron), makine öğrenimi uygulamalarında geniş çapta kullanılan bir tür yapay sinir ağıdır. Bu ağ, birbirine bağlı çok katmandan oluşur ve genellikle sınıflandırma ve regresyon gibi denetimli öğrenme problemleri için kullanılır [21]. MLP'nin giriş katmanı, gizli katman ve çıkış katmanı olmak üzere toplam üç katmanı vardır. İşlenecek giriş verileri, giriş katmanı tarafından alınır ve bu veriler, çıkış katmanına iletilmeden önce gizli katmanda işlemden geçer. Ardından, sınıflandırma ve regresyon gibi görevleri yerine getirir.

MLP'nin nöronlar arasındaki bağlantı ağırlıklarını ayarlamak için iki temel adımı izler. Bu adımlar ileri yayılım ve geri yayılım olarak adlandırılırlar. Her bir nöron Eş. 1 ile işlenir.

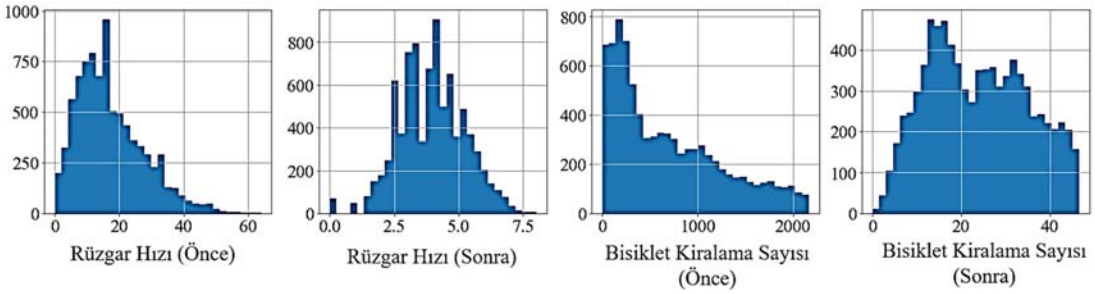
$$nöron_j = \sum_{i=1}^{k} w_{ij} x_i \quad (1)$$

Burada x_i, i. nöronun çıktısı, k giriş boyutu, w_{ij} ise j ve i nöronları arasındaki ağırlıklardır. j. nöronun çıktısı, $nöron_j$'nin f aktivasyon fonksiyonundan (sigmoid, relu, tanh ve softmax gibi) geçirilmesiyle Eş. 2 ile elde edilir:

$$c_i = f\left(nöron_j\right) \quad (2)$$



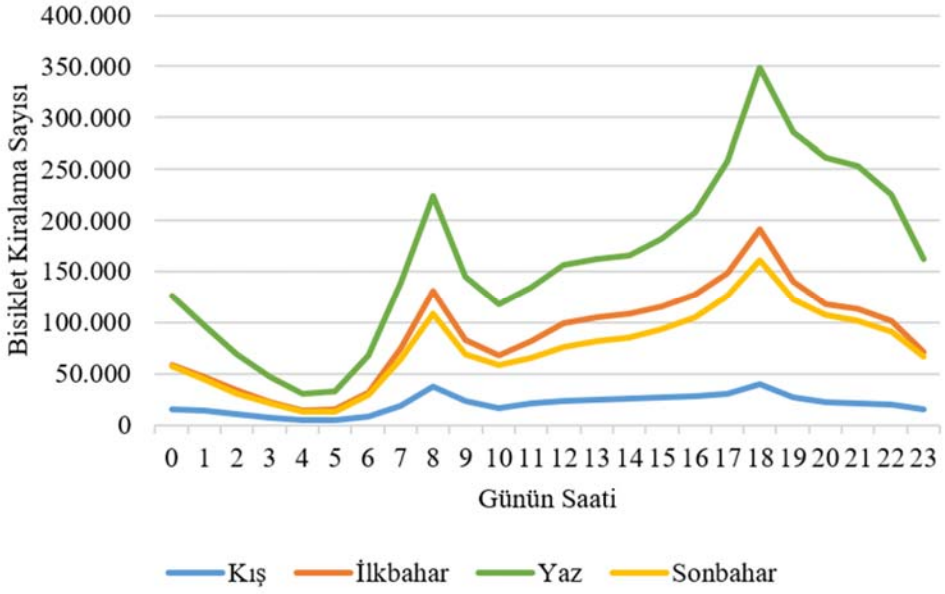
Şekil 2. Özelliklere ait veri dağılım grafikleri (Data distribution graphs of features)



Şekil 3. Rüzgar hızı ve bisiklet kiralama sayısı özelliklerine ait dönüşüm işleminden önceki ve sonraki dağılım grafikleri (Data distribution graphs of wind speed and bike rental count features before and after the transformation process)

3.4. Gradyan Artırmalı Makinesi Algoritması (Gradient Boosting Machine Algorithm)

Gradyan Artırmalı Makinesi (GBM: Gradient Boosting Machine) sınıflandırma ve regresyon gibi görevler için kullanılan popüler bir makine öğrenimi algoritmasıdır. Arttırma (Boosting) algoritmalarında temel yaklaşım, birkaç basit modeli (zayıf öğreniciler) tekrarlayarak bir tahmin doğruluğu artırılmış güçlü öğrenici elde etmektir [22]. Friedman, arttırma algoritmasını regresyon için genişleterek GBM'yi tanıtmıştır [23]. GBM yöntemi, kayıp fonksiyonunu en aza indiren bir



Şekil 4. Her mevsim için saatlik bisiklet kiralama sayıları (Hourly bike rental numbers for each season)

toplam modeli bulmayı amaçlayan sayısal bir algoritmadır. Zayıf öğrenicileri güçlü öğrenicilere dönüştürme prensibi üzerine çalışır [24].

İlk olarak, modelin katsayılarının veriye ne kadar uyduğunu incelemek için Eş. 3 ile kayıp fonksiyonunu tanımlanır:

$$K\left(y_i, f(x)\right) \tag{3}$$

Burada $f(x)$ tahmin edilen değer ve y_i ise gerçek değerdir. Devamında, Eş. 4 ile kayıp fonksiyonunun değerinin en küçük olduğu durumun bulunması hedeflenir.

$$f_0(x) = argmin \sum_{i=1}^{n} K(Y_i, \theta) \tag{4}$$

Burada θ tahmin edilen değerdir. Sonra, yapılan tahminlere göre hatalar hesaplanır ve ardından karar ağaçları oluşturulur. Son olarak ise her bir gözlem için tahmin değerleri oluşturulur.

3.5. Yarasa Algoritması (Bat Algorithm)

Yarasalar, zifiri karanlıkta uçabilme ve avlarına saldırabilme yeteneğine sahiptirler. Yiyecek bulmak, engelleri aşmak ve karanlıkta yuva yerlerini bulmak için yarasalar ekolokasyon (yankı ile yer belirleme) olarak bilinen bir tür sesli radar kullanırlar. Yüksek bir ses dalgası yayıp ardından avdan yansıyan yankıyı beklerler. Yarasa Algoritması (YA), yarasanın ekolokasyon davranışından ilham alarak geliştirilmiş nüfus tabanlı meta-sezgisel bir algoritmadır [25]. YA'da, yarasaların ekolokasyon özellikleri aşağıdaki kurallarla idealize edilebilir [25]:

a) Tüm yarasalar mesafe algılamak için ekolokasyon kullanır ve bir yarasanın konumu x_i, ele alınan bir optimizasyon problemine göre bir çözüm olarak kodlanır.
b) Konumları X_i ve hızları V_i olan yarasalar, avlarını aramak için değişken frekansta (minimum frekans f_{min}'den maksimum frekans f_{max}'e kadar) veya değişken dalga boyunda (λ) ve A_0 ses şiddetinde rastgele uçarlar. Hedefin yakınlığına bağlı olarak dalgalarının frekansını ve darbe yayım hızını [0,1] aralığında otomatik olarak ayarlarlar.
c) Ses şiddeti, büyük bir pozitif değer olan A_0'dan minimum sabit bir değer olan A_{min}'e kadar değişir.

Yarasaların güncellenen konumları X_i^t ve hızları V_i^t, t zaman adımında Eş. 5, Eş. 6 ve Eş. 7 ile elde edilir.

$$f_i = f_{min} + (f_{max} - f_{min})\,\beta \tag{5}$$

$$V_i^t = V_i^{t-1} + (X_i^t - X^*)\,f_i \tag{6}$$

$$X_i^t = X_i^{t-1} + V_i^t \tag{7}$$

Burada, f_i her iterasyon aşamasında güncellenen i'inci yarasanın frekansını belirtirken, β ise [0, 1] aralığında uniform bir dağılımdan çekilen rastgele bir vektördür. X^* mevcut global en iyi konum (çözüm) olup, tüm n yarasalar arasındaki tüm çözümler karşılaştırıldıktan sonra belirlenir.

Bu çalışmada önerilen yöntem, YA ile GBM modelinin optimize edilmesiyle oluşturulmuştur. Bu optimizasyon işleminde, GBM modelinin giriş parametreleri olan learning_rate, n_estimators, max_depth ve min_samples_split değerlerinin minimum ve maksimum aralıkları belirlenerek, en uygun değerlerinin bulunması sağlanmıştır. Optimizasyon için kullanılan parametrelerin minimum ve maksimum değerleri şöyledir: max_depth (2-30), n_estimators (10-500), learning_rate (0.01-0.1) ve min_samples_split (2-30).

4. Bulgular (Results)

Bu çalışmada KNN, DT, MLP ve GBM olmak üzere dört farklı makine öğrenmesi yöntemi ile bisiklet kiralama sayıları tahmin edilmiştir. Bu işlem için, veri seti eğitim veri kümesi ve test veri kümesi olmak üzere iki kısma ayrılmıştır. Eğitim veri kümesi veri setinin %90'lik kısmını oluştururken, test veri kümesi ise geri kalan %10'lik kısmı oluşturmaktadır. Eğitim süreci, modellerin etkinliğini için 5-katlamalı çapraz doğrulama (5-fold cross validation) kullanılarak gerçekleştirilmiştir. Son olarak, her bir modelin en iyi formunu bulmak için ızgara arama yöntemi (grid search) kullanılmıştır. Bulunan en iyi model formu test verilerine uygulanmıştır. Ayrıca en iyi sonuç veren model YA ile optimize edilmiştir. Elde edilen tahmin sonuçları her bir mevsim için ayrı ayrı da analiz edilmiştir. Son olarak, karar ağacı temelli modellerin özellik önemi (feature importance) analizi yapılmıştır. Bu özellik önemi analizi için sklearn kütüphanesinin sağladığı CART yöntemi kullanılmıştır.

Bu çalışmada yapılan tüm analizler, Python programlama dili kullanılarak Spyder geliştirme ortamı üzerinde sklearn ve opytimizer kütüphaneleri kullanılarak gerçekleştirilmiştir. Bu bölümde, modellerin performanslarını karşılaştırmada kullanılan metrikler ve elde edilen sonuçlar sunulmuştur.

4.1. Performans Değerlendirme Metrikleri (Performance Evaluation Metrics)

Modellerin performanslarını değerlendirme yaygın olarak kullanılan performans değerlendirme metrikleri Determinasyon Katsayısı (R^2), Ortalama Mutlak Hata (MAE), Ortalama Karesel Hata (MSE), ve Kök Ortalama Karesel Hata (RMSE)'dir [26]. Bu çalışmada kullanılan modellerin performans değerlendirmesi, bu dört metrik dikkate alarak yapılmıştır. Bu metrikler Eş. 8, Eş. 9, Eş. 10 ve Eş. 11 ile elde edilir.

$$MAE = \frac{\sum_{i=1}^{N}|G_i - T_i|}{N} \quad (8)$$

$$MSE = \frac{\sum_{i=1}^{N}(G_i - T_i)^2}{N} \quad (9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(G_i - T_i)^2}{N}} \quad (10)$$

$$R^2 = \frac{\sum_{i=1}^{N}(G_i - T_i)^2}{\sum_{i=1}^{N}(G_i - \overline{G_i})^2} \quad (11)$$

Burada, G_i gerçek değer, $\overline{G_i}$ gerçek değerin ortalaması ve T_i tahmin edilen değerdir. N ise testlerin sayısıdır. R^2 değeri 0 ile 1 arasında değer alır ve 1 değeri en iyi performansı temsil eder. Bu metriğin değeri 0'a yaklaştıkça modelin tahminlerinin kötüye doğru gittiğini gösterir. MAE, MSE ve RMSE metriklerinde ise değerin düşük olması modelin performansının arttığını gösterir.

4.2. Elde Edilen Performans Sonuçları (Performance Results Obtained)

Her bir modelin kendine has ayarlanması gereken parametreleri mevcuttur. Bu parametreler ve olası değerleri Tablo 2'de verilmiştir.

Her bir model için en iyi parametre değerlerini bulmak için birbirinden bağımsız olarak parametre ayarlama işlemi gerçekleştirilmiştir. Bu işlem sonucunda elde edilen en iyi parametre değerleri Tablo 3'te verilmiştir. En iyi sonuç veren model GBM modelinin giriş parametreleri, YA ile optimize edilerek modelin performansı arttırılmıştır. Bu optimizasyon işlemi için YA algoritması 30 parçacık ile 50 iterasyon çalıştırılmıştır. Elde edilen sonucun güvenilirliğini sağlamak için bu işlem 30 defa tekrarlanmıştır. Optimizasyon işlemi sonucunda elde edilen giriş parametrelerin optimum değerleri yıllık tahmin için şöyledir: max_depth (8), n_estimators (499), learning_rate (0,083) ve min_samples_split (2).

4.2.1. Kış mevsimi için elde edilen sonuçlar (Results obtained for the winter season)

Kış mevsiminde bisiklet kiralama sayıları dört farklı model ile tahmin edilmiştir. DT (criterion=squared_error, max_depth=8 ve min_samples_leaf= 3), KNN (n_neighbors=5), MLP (activation=logistic, alpha=0,035 hidden_layer_sizes=(15, 2)) ve GBM (learning_rate=0,1, max_depth=5, min_samples_split=3, n_estimators=400) modellerinin en iyi formları Tablo 3'te gösterildiği gibi elde edilmiştir. Bu modellerden en iyi performansı 1,2439 MAE, 1,8253 MSE, 3,3319 RMSE ve 0,8532 R^2 değerleri ile GBM elde etmiştir (Tablo 4). GBM'nin performansı YA ile optimize edilerek arttırılmıştır (1,2611 MAE, 1,8113 MSE, 3,2809 RMSE ve 0,8555 R^2). Her ne kadar MAE değeri olumsuz yönde artsa da MSE, RMSE ve R^2 değerlerinin artması performansın arttığı gösterir [27].

Tablo 4. Kış mevsimi için elde edilen sonuçlar
(Results obtained for the winter season)

Model Adı	MAE	MSE	RMSE	R^2
DT	2,1430	2,9122	8,4812	0,6265
KNN	2,8853	4,0358	16,2882	0,2827
MLP	2,1692	2,9889	8,9340	0,6065
GBM	1,2439	1,8253	3,3319	0,8532
YA-GBM	1,2611	1,8113	3,2809	0,8555

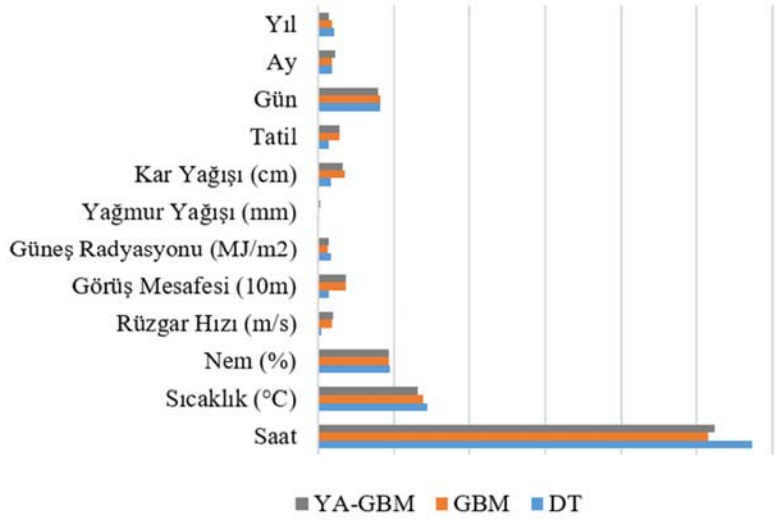
Ayrıca karar ağacı temelli modellerin (DT, GBM ve YA-GBM) özellik önemi (feature importance) analizi yapılmıştır. Üç model için de bisiklet kiralama sayısı tahminine etki eden en önemli özellik günün saati olurken, yağmur yağışının hiçbir etkisi olmamıştır (Şekil 5). Günün saat'ini takip eden ikinci özellik ise hava sıcaklığı olmuştur.

Tablo 2. Modellerin parametreleri ve olası değerleri (Model parameters and their possible values)

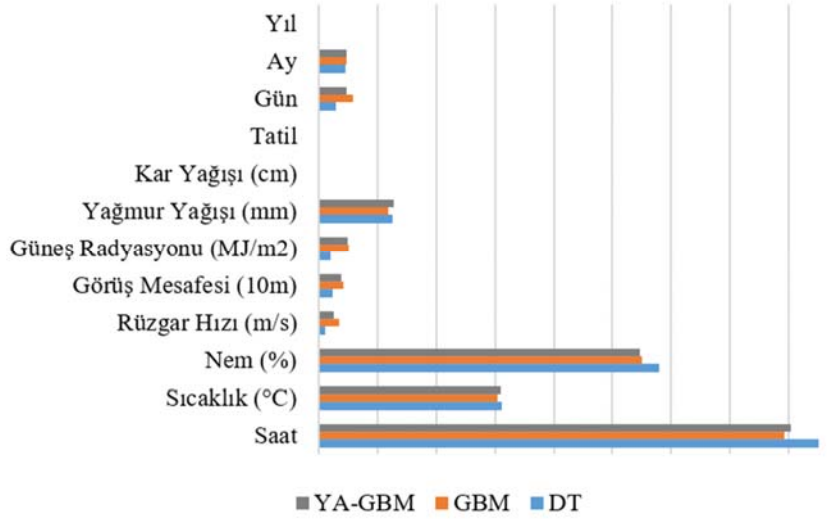
Model Adı	Parametre	Değerler
KNN	n_neighbors	1,3,5,7,9
DT	max_depth	1,2,4,8
	min_samples_leaf	1, 2, 3, 4
	criterion	squared_error, absolute_error
MLP	hidden_layer_sizes	(5,1),(10,1),(15,1), (5,2), (10,2), (15,2)
	activation	relu, tanh, logistics
	alpha	0.001, 0.01, 0.035, 0.1
GBM	max_depth	1,2,5,10,15
	n_estimators	10, 100,200,400,500
	learning_rate	0.01, 0.02, 0.05, 0.1
	min_samples_split	1,3,5,20

Tablo 3. Modellerin elde edilen en iyi parametre değerleri (The best obtained parameter values for the models)

Model	Parametre	Kış	İlkbahar	Yaz	Sonbahar	Yıllık
KNN	n_neighbors	5	5	3	5	5
DT	max_depth	8	8	8	8	8
	min_samples_leaf	3	4	4	4	4
	criterion	squared_err.	squared_err.	squared_err.	squared_err.	squared_err.
MLP	hidden_layer_sizes	(15, 2)	(10, 1)	(10, 2)	(15, 2)	(10, 2)
	activation	logistic	logistic	logistic	logistic	logistic
	alpha	0,035	0,035	0,1	0,01	0,035
GBM	max_depth	5	10	5	5	10
	n_estimators	400	400	400	400	400
	learning_rate	0,1	0,05	0,05	0,1	0,1
	min_samples_split	3	20	20	5	20



Şekil 5. Kış mevsimi için karar ağacı temelli modellerin özellik önemi gösterimi
(Visualization of feature importance for decision tree-based models during the winter season)



Şekil 6. İlkbahar mevsimi için karar ağacı temelli modellerin özellik önemi gösterimi
(Visualization of feature importance for decision tree-based models during the spring season)

4.2.2. İlkbahar mevsimi için elde edilen sonuçlar
(Results obtained for the spring season)

İlkbahar mevsiminde bisiklet kiralama sayıları dört farklı model ile tahmin edilmiştir. DT (criterion=squared_error, max_depth=8 ve min_samples_leaf=4), KNN (n_neighbors=5), MLP (activation=logistic, alpha=0,035, hidden_layer_sizes=(10, 1)) ve GBM (learning_rate=0.1, max_depth=5, min_samples_split=3, n_estimators=400) modellerinin en iyi formları Tablo 3'te gösterildiği gibi elde edilmiştir. Bu modellerden en iyi performansı 1,9476 MAE, 3,2299 MSE, 10,4324 RMSE ve 0,8673 R^2 değerleri ile GBM elde etmiştir (Tablo 5). GBM'nin performansı YA ile optimize edilerek arttırılmıştır (1,9586 MAE, 3,0383 MSE, 9,2313 RMSE ve 0,8826 R^2).

Ayrıca karar ağacı temelli modellerin (DT, GBM ve YA-GBM) özellik önemi analizi yapılmıştır. Her üç model için de bisiklet kiralama sayısı tahminine etki eden en önemli özellik günün saati olurken, en önemsiz özellikler ise kar yağışı ve tatil ve yıl olmuştur (Şekil 6). Her üç model için de en çok etkili ikinci özellik ise nem olmuştur.

Tablo 5. İlkbahar mevsimi için elde edilen sonuçlar
(Results obtained for the spring season)

Model Adı	MAE	MSE	RMSE	R^2
DT	2,8274	4,1637	17,3364	0,7795
KNN	4,2114	5,5432	30,7272	0,6093
MLP	3,9248	4,9260	24,2657	0,6914
GBM	1,9476	3,2299	10,4324	0,8673
YA-GBM	1,9586	3,0383	9,2313	0,8826

4.2.3. Yaz mevsimi için elde edilen sonuçlar
(Results obtained for the summer season)

Yaz mevsiminde bisiklet kiralama sayıları dört farklı model ile tahmin edilmiştir. DT (criterion=squared_error, max_depth=8 ve min_samples_leaf= 4), KNN (n_neighbors=3), MLP (activation=

logistic, alpha=0,1, hidden_layer_sizes=(10, 1)) ve GBM (learning_rate=0,05, max_depth=5, min_samples_split=20, n_estimators=400) modellerinin en iyi formları Tablo 3'te gösterildiği gibi elde edilmiştir. Bu modellerden en iyi performansı 2,2337 MAE, 3,4377 MSE, 11,8184 RMSE ve 0,8399 R^2 değerleri ile GBM elde etmiştir (Tablo 6). GBM'nin performansı YA ile optimize edilerek arttırılmıştır (2,1748 MAE, 3,3310 MSE, 11,0961 RMSE ve 0,8497 R^2).

Tablo 6. Yaz mevsimi için elde edilen sonuçlar
(Yaz mevsimi için elde edilen sonuçlar)

Model Adı	MAE	MSE	RMSE	R^2
DT	2,7862	4,2007	17,6462	0,7610
KNN	4,6861	6,2539	39,1115	0,4704
MLP	3,4496	4,4444	19,7530	0,7325
GBM	2,2337	3,4377	11,8184	0,8399
YA-GBM	2,1748	3,3310	11,0961	0,8497

Ayrıca karar ağacı temelli modellerin (DT, GBM ve YA-GBM) özellik önemi analizi yapılmıştır. Üç model için de bisiklet kiralama sayısı tahminine etki eden en önemli özellik günün saati olurken; yıl, yağmur yağışı ve tatil özelliklerinin hiçbir etkisi olmamıştır (Şekil 7).

4.2.4. Sonbahar mevsimi için elde edilen sonuçlar
(Results obtained for the autumn season)

Sonbahar mevsiminde bisiklet kiralama sayıları dört farklı model ile tahmin edilmiştir. DT (criterion=squared_error, max_depth=8 ve min_samples_leaf= 4), KNN (n_neighbors=5), MLP (activation=logistic, alpha=0,01, hidden_layer_sizes=(15, 2)) ve GBM (learning_rate=0,1, max_depth=5, min_samples_split=5, n_estimators=400) modellerinin en iyi formları Tablo 3'te gösterildiği gibi elde edilmiştir. Bu modellerden en iyi performansı 1,9390 MAE, 3,0414 MSE, 9,2501 RMSE ve 0,8467 R^2 değerleri ile GBM elde etmiştir (Tablo 7). GBM'nin performansı YA ile optimize edilerek arttırılmıştır (1,7152 MAE, 2,8301 MSE, 8,0099 RMSE ve 0,8672 R^2).

Ayrıca karar ağacı temelli modellerin (DT, GBM ve YA-GBM) özellik önemi analizi yapılmıştır. Üç model için de bisiklet kiralama sayısı tahminine etki eden en önemli özellik günün saati olurken, yılın hiçbir etkisi olmamıştır (Şekil 8). En fazla etki eden günün saat'ini ikinci özellik olarak ise nem takip etmiştir.

Tablo 7. Sonbahar mevsimi için elde edilen sonuçlar
(Results obtained for the autumn season)

Model Adı	MAE	MSE	RMSE	R^2
DT	2,4846	3,6114	13,0429	0,7838
KNN	3,5214	4,9864	24,8649	0,5879
MLP	3,1552	4,1232	17,0014	0,7182
GBM	1,9390	3,0414	9,2501	0,8467
YA-GBM	1,7152	2,8301	8,0099	0,8672

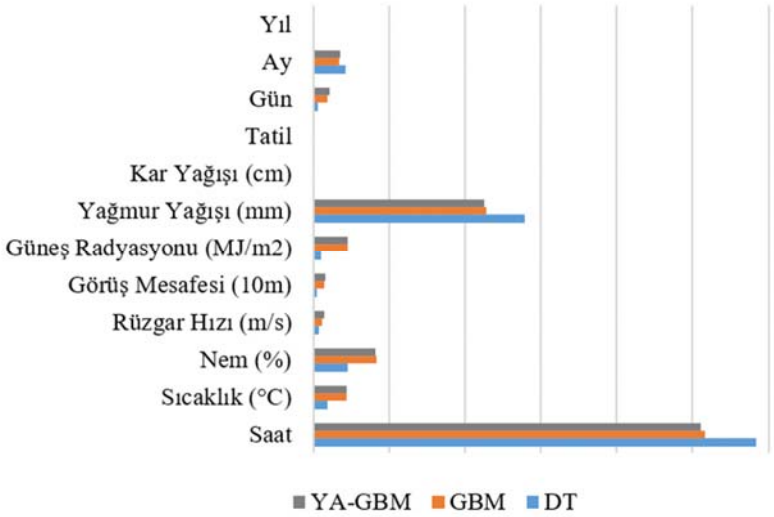
4.2.5. Yıllık tahmin için elde edilen sonuçlar
(Results obtained for the year)

Yıllık olarak bisiklet kiralama sayıları yine aynı dört farklı model ile tahmin edilmiştir. DT (criterion=squared_error, max_depth=8 ve min_samples_leaf= 4), KNN (n_neighbors=5), MLP (activation=logistic, alpha=0,035, hidden_layer_sizes=(10, 2)) ve GBM (learning_rate=0,1, max_depth=10, min_samples_split=20, n_estimators=400) modellerinin en iyi formları Tablo 3'te gösterildiği gibi elde edilmiştir. Bu modellerden en iyi performansı 1,9881 MAE, 3,1439 MSE, 9,8845 RMSE ve 0,9168 R^2 değerleri ile GBM elde etmiştir (Tablo 8). GBM'nin performansı YA ile optimize edilerek arttırılmıştır (1,8665 MAE, 2,9588 MSE, 8,7545 RMSE ve 0,9264 R^2).

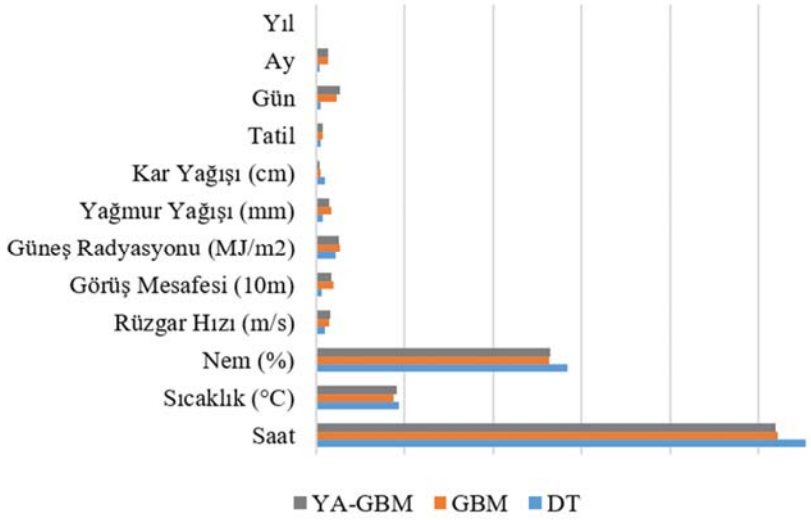
Tablo 8. Yıllık tahmin için elde edilen sonuçlar
(Results obtained for the year)

Model Adı	MAE	MSE	RMSE	R^2
DT	3,0833	4,3491	18,9149	0,8409
KNN	4,5455	6,1686	38,0518	0,6799
MLP	4,5736	5,8429	34,1395	0,7128
GBM	1,9881	3,1439	9,8845	0,9168
YA-GBM	1,8665	2,9588	8,7545	0,9264

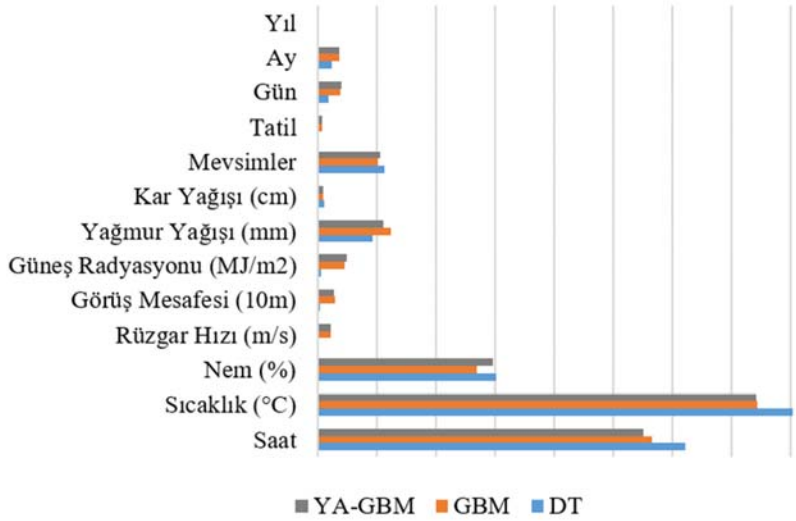
Ayrıca karar ağacı temelli modellerin (DT, GBM ve YA-GBM) özellik önemi analizi yapılmıştır. Her üç model için de bisiklet kiralama sayısı tahminine etki eden en önemli özellik hava sıcaklığı olurken, bu özelliği günün saati ve nem takip etmiştir. Yıl ise hiçbir etkisi olmamıştır (Şekil 9).



Şekil 7. Yaz mevsimi için karar ağacı temelli modellerin özellik önemi gösterimi
(Visualization of feature importance for decision tree-based models during the summer season)



Şekil 8. Sonbahar mevsimi için karar ağacı temelli modellerin özellik önemi gösterimi
(Visualization of feature importance for decision tree-based models during the autumn season)



Şekil 9. Yıllık tahmin için karar ağacı temelli modellerin özellik önemi gösterimi
(Visualization of feature importance for decision tree-based models for the year)

Tablo 9. Aynı veri setini kullanan literatürdeki diğer çalışmalar ile önerilen modelin performans karşılaştırması
(Comparison of the performance of the proposed model with other studies in the literature using the same dataset)

Model Adı	MAE	MSE	RMSE	R^2
Linear Regression [28]	347,2433	-	459,2854	0,4706
Random Forest 1 [28]	121,387	-	209,858	0,8991
SVM [29]	153,32	-	242,89	0,85
Random Forest 2 [29]	130,52	-	216,01	0,88
XGBTree [30]	119,59	-	183,80	0,91
GBM [30]	109,78	-	172,73	0,92
YA-GBM (Önerilen model)	1,8665	2,9588	8,7545	0,9264

Öğrenme eğrileri (learning curves), eğitim veri setinin boyutunun model performansı üzerindeki etkisini değerlendirir. Aşırı öğrenme (over-fitting) ve eksik öğrenme (under-fitting) dengesi olup olmadığı tespit etmek için kullanılabilir. Yıllık ve mevsim bazlı tahmin için elde edilen öğrenme eğrileri Şekil 10'da gösterilmiştir. Aşırı öğrenme ve eksik öğrenmenin olmadığı ideal durumda, eğitim veri miktarı arttıkça hem eğitim hem de test eğrisinin hata seviyeleri düşük ve stabil bir noktaya yaklaşırlar. Bu, modelin veriden etkili bir şekilde öğrendiğini gösterir. Şekil 10'da görüldüğü üzere eğitim dengeli bir şekilde gerçekleşmiştir. KNN, MLP ve DT dengesiz olarak dağılmış ve büyük veri setlerinde kötü performans gösterirler. Aykırı değerlere oldukça duyarlıdır. Bu tür veri durumlarında, yanıltıcı kararlar

verebilirler. Diğer modellere nazaran GBM, büyük veri kümeleri ile etkili bir şekilde başa çıkabilir. Takım yapısı, zayıf modelleri bir araya getirerek güçlü tahmin modelleri oluşturmasına olanak tanır ve daha fazla veri ile daha etkili hale gelir. Takım yapısının başka bir faydası ise modelin aykırı değerlere göreceli olarak dirençli olmasını sağlar, bu da bireysel örneklerin etkisini azaltmaya yardımcı olur. Ayrıca, modellerin performansı etkileyen en önemli özellik mevsim bazlı tahminde günün saati olurken, yıllık bazda ise hava sıcaklığı olmuştur.

Bu çalışmada önerilen modelin (YA-GBM) performansı, literatürdeki diğer çalışmalardaki aynı veri setini kullanan modeller ile kıyaslanmıştır. Tablo 9'da karşılaştırma sonuçları verilmiştir. Önerilen modelin performansı, diğer yapılan çalışmalardaki modellerin performanslarından tüm metriklerde yüksektir.

5. Sonuçlar (Conclusions)

Paylaşımlı bisikletlerin etkili kaynak tahsisi için kentsel bisiklet talebinin doğru tahmin edilmesi bir gereklilik haline gelmiştir. Bu tahmin işlemi GBM algoritması kullanılarak gerçekleştirilmiştir. Modelin etkinliğini göstermek için, önerilen modelin performansı DT, KNN ve MLP gibi farklı metotlar ile karşılaştırılmıştır. Bu karşılaştırma işlemi MAE, MSE, RMSE ve R^2 metrikleri kullanılarak gerçekleştirilmiştir. Ayrıca, bisiklet kiralama sayısının tahminine en az ve en çok etkisi olan özellikler belirlenmiştir. En çok etki eden özellikler günün saati ve hava sıcaklığı olurken, en az etki eden özellikler ise kar yağışı ve yıl olmuştur. Gelecekteki çalışmalarda, YA ile optimize edilmiş önerilen model performansı daha farklı optimizasyon yöntemleri ile optimize edilerek karşılaştırma yapılabilir.

Kaynaklar (References)

1. Yan, S., Lu, C. C., Wang, M. H., Stochastic fleet deployment models for public bicycle rental systems, International Journal of Sustainable Transportation, 12 (1), 39–52, 2018.
2. Eren, E., Uz, V. E., A review on bike-sharing: The factors affecting bike-sharing demand, Sustainable Cities and Society, 54, 101882, 2020.
3. Gao, X., Lee, G. M., Moment-based rental prediction for bicycle-sharing transportation systems using a hybrid genetic algorithm and machine learning, Computers & Industrial Engineering, 128, 60–69, 2019.
4. Qi, X., Gao, Y., Li, Y., Li, M., K-nearest Neighbors Regressor for traffic prediction of rental bikes, 14th International Conference on Computer Research and Development (ICCRD), 152–156, January 2022.
5. Feng, Y., Wang, S., A forecast for bicycle rental demand based on random forests and multiple linear regression, IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 101–105, May 2017.
6. Shiao, Y. C., Chung, W. H., Chen, R. C., Using SVM and Random forest for different features selection in predicting bike rental amount, 9th International Conference on Awareness Science and Technology (iCAST), 1–5, September 2018.
7. Heidari, E., Sobati, M. A., Movahedirad, S., Accurate prediction of nanofluid viscosity using a multilayer perceptron artificial neural network (MLP-ANN), Chemometrics and Intelligent Laboratory Systems, 155, 73–85, 2016.
8. Yatim, F. E., Boumanchar, I., Srhir, B., Chhiti, Y., Jama, C., Alaoui F. E. M., Waste-to-energy as a tool of circular economy: Prediction of higher heating value of biomass by artificial neural network (ANN) and multivariate linear regression (MLR), Waste Management, 153, 293–303, 2022.
9. Nsangou, J. C., Kenfack, J., Nzotcha, U., Ngohe Ekam, P. S., Voufo, J., Tamo, T. T., Explaining household electricity consumption using quantile regression, decision tree and artificial neural network, Energy, 250, 123856, 2022.
10. Thamarai, M., Malarvizhi, S. P., House price prediction modeling using machine learning, International Journal of Information Engineering & Electronic Business, 12 (2), 2020.
11. Baofeng, D., Zhang, H., Liu, Y., Li, J., Chen, N., Stamatopoulos, C. A., Luo, Y., Zhan, Y., Assessing susceptibility of debris flow in southwest China using gradient boosting machine, Scientific Reports, 9 (1), 2019.
12. Zhu, J., Fang, S., Yang, Z., Qin, Y., Chen, H., Prediction of concrete strength based on random forest and gradient boosting machine, IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), 306–312, January 2023.
13. Akköse G., Duran A., Gürsel Dino İ., Akgül Ç.M., Machine learning based evaluation of window parameters on building energy performance and occupant thermal comfort under climate change, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (4), 2069-2084, 2023.
14. Gülmez B., Kulluk S., Analysis and price prediction of secondhand vehicles in Türkiye with big data and machine learning techniques, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (4), 2279-2290, 2023.
15. Acı M., Ayyıldız Doğansoy G., Demand forecasting for e-retail sector using machine learning and deep learning methods, Journal of the Faculty of Engineering and Architecture of Gazi University, 37 (3), 1325-1340, 2022.
16. Seoul Bike Sharing Demand, UCI Machine Learning Repository, 2020.
17. İlgün E.G., Samet R., Increasing the performance of intrusion detection models developed using machine learning method with preprocessing applied to the dataset, Journal of the Faculty of Engineering and Architecture of Gazi University, 39 (2), 679-692, 2023.
18. Bisong, E., Building machine learning and deep learning models on Google Cloud Platform, Berkeley, CA: Apress, 2019.
19. Lin Y., Wang, J., Research on text classification based on SVM-KNN, IEEE 5th International Conference on Software Engineering and Service Science, 842–844, June 2014.
20. Suthaharan, S., Suthaharan, S., Decision tree learning, Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 237-269. 2016.
21. Murtagh, F., Multilayer perceptrons for classification and regression, Neurocomputing, 2 (5–6), 3–197, 1991.
22. Friedman, J., Hastie, T., Tibshirani, R., Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors), The Annals of Statistics, 28 (2), 2000.
23. Friedman, J. H., Greedy Function Approximation: A Gradient Boosting Machine, Annals of Statistics, 29 (5),1189–1232, 2001.
24. Yeşilyurt, S., Dalkılıç, H., Xgboost ve gradient boost machine ile günlük nehir akımı tahmini, 3rd International Symposium of III Engineering Applications on Civil Engineering and Earth Sciences, 2021.
25. Yang, X.S., A new metaheuristic bat-inspired algorithm, In Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), 65-74, 2010.
26. de Guia, J. D., Concepcion, R. S., Calinao, H. A., Alejandrino, J., Dadios, E.P., Sybingco, E., Using stacked long short term memory with principal component analysis for short term prediction of solar irradiance based on weather patterns, 2020 IEEE Region 10 Conference (Tencon), 946-951, 2020.
27. Chicco, D., Warrens, M. J., Jurman, G., The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Computer Science, 7, 2021.
28. DNgo, T.T. T., Pham, H. T., Acosta, J. G., Derrible, S., Predicting bike-sharing demand using random forest, Journal of Science and Transport Technolog, 13-21, 2022.
29. Sathishkumar, V. E., Cho, Y., Season wise bike sharing demand analysis using random forest algorithm, Computational Intelligence, 2020.
30. Sathishkumar, V. E., Park, J., Cho, Y., Using data mining techniques for bike sharing demand prediction in metropolitan city, Computer Communications, 153, 353-366, 2020.