www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 9, No: 4, 2022 (1388-1397)

El-Cezerî Journal of Science and Engineering
Vol: 9, No: 4, 2022 (1388-1397)
DOI: 10.31202/ecjse.1135485

ECJSE

Makale / Research Paper

Kayıp IoT Verilerinin Makine Öğrenmesi Teknikleri ile Tahmini

Fatma AZİZOĞLU[a], Emre ÜNSAL[b]

[a]Sivas Cumhuriyet Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü,
[b]Sivas Cumhuriyet Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği Bölümü
Sivas/TÜRKİYE
fazizoglu@cumhuriyet.edu.tr

Received/Geliş: 25.04.2022 **Accepted/Kabul:** 07.09.2022

Öz: Nesnelerin İnterneti (IoT) tabanlı endüstriyel uygulamalardan toplanan veriler her geçen gün giderek artmaktadır. Bununla birlikte, IoT cihazlarındaki arıza ve iletişim kopukluğu sebebi ile toplanan veriler gürültülü, belirsiz ve eksik olabilmektedir. Bu problemler, veri üretimi, kalitesi, işlenmesi ve analizleri için kritik bir konu haline gelmiştir. Bu çalışma kapsamında kullanılan veri setleri, Sivas Numune Hastanesi tıbbi atıkları evsel atığa dönüştüren su nötralizatör sisteminden gerçek zamanlı toplanarak oluşturulmuştur. Hastanelerde bulunan tıbbi sıvı atıklar kanalizasyona aktarılmadan önce nötralizasyon cihazları ile pH değişikliği yoluyla kimyasal nötralizasyon işlemine maruz bırakılmaktadır. Bu anlamda, tıbbi atık nötralizasyon sistemindeki pH değerlerinin gözlemlenmesi çevrenin korunması adına oldukça önemlidir. Bu kapsamda, farklı miktarlarda eksiltilerek oluşturulan iki veri seti ile kayıp pH verilerinin tahmini için Lineer Regresyon (LR), Destek Vektör Makineleri (DVM), K-En Yakın Komşuluk (KNN), Rastgele Orman (RO), Karar Ağacı (KA) ve Adaboost olmak üzere altı farklı makine öğrenmesi algoritması kullanılmıştır. Makine öğrenmesi algoritmalarının değerlendirilmesinde ortalama mutlak hata (Mean Absolute Error, MAE), ortalama karesel hata (Mean Squared Error, MSE) ve kök ortalama kare hata (Root Mean Square Error, RMSE) performans metrikleri kullanılmıştır. Gerçekleştirilen çalışma sonucunda iki farklı veri seti içinde DVM algoritmasının daha başarılı olduğu gözlemlenmiştir. Yapılan değerlendirme sonucu, makine öğrenmesi algoritmalarının kayıp pH verilerinin tahmininde oldukça başarılı olduğunu göstermektedir.

Anahtar Kelimeler: Kayıp veri tahmini, kayıp veri doldurma, nesnelerin interneti, makine öğrenmesi, tıbbi atık.

Missing IoT Data Prediction with Machine Learning Techniques

Abstract: Every day, the amount of data generated by industrial applications based on the Internet of Things (IoT) grows. However, data acquired because of failures and communication disconnections in IoT devices might be noisy, inaccurate, and incomplete. These issues have become crucial for data production, quality, processing, and analysis. The datasets used in the scope of this study were collected in real-time from the water neutralizer system of Sivas Numune Hospital, which converts medical waste into household waste. Medical liquid wastes in hospitals are exposed to chemical neutralization process by means of pH change with neutralization devices before being transferred to the sewer. In this regard, the monitoring of pH levels in the medical waste neutralization system is crucial for environmental protection. In this aspect, two datasets with varying quantities of missing data were evaluated for the prediction of the PH using the linear regression (LR), support vector machines (SVM), k-nearest neighbor (KNN), random forest (RF), and decision tree (DT) machine learning algorithms. Mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE) performance metrics were used to evaluate machine learning algorithms. Because of the analysis, it was determined that the SVM algorithm performed better performance on the two distinct datasets. The result of the evaluation indicates that machine learning algorithms are remarkably efficient at predicting missing pH data.

Keywords: Missing data prediction, missing data imputation, Internet of Things (IoT), machine learning, medical waste.

Bu makaleye atıf yapmak için
Azizoğlu F., Ünsal E., "Kayıp IoT Verilerinin Makine Öğrenmesi Teknikleri ile Tahmini" El-Cezerî Fen ve Mühendislik Dergisi 2022, 9 (4); 1388-1397.

How to cite this article
Azizoğlu F., Ünsal E., "Kayıp IoT Verilerinin Makine Öğrenmesi Teknikleri ile Tahmini" El-Cezerî Fen ve Mühendislik Dergisi 2022, 9 (4); 1388-1397.

ORCID ID: [a] 0000-0003-0892-8711; [b] 0000-0001-6042-0742

1. Giriş

Son zamanlarda, birçok endüstriyel IoT uygulaması tarafından üretilen veri çok büyük miktarlara ulaşmıştır ve bu veri miktarı her geçen gün artmaya devam etmektedir. Fakat bu uygulamaların birçoğu, zayıf veri giriş süreçleri, yanlış hesaplamalar, veri toplayan cihazların arızalanması veya iletişimin kopması gibi çeşitli sebepler ile eksik olarak toplanan veri problemlerinden etkilenmektedir [1]. Buna bağlı olarak, tıbbi, araştırma ve sensör gibi çeşitli gerçek zamanlı toplanan veri kümelerindeki kayıp veri miktarları da her geçen gün artmaktadır. Kayıp verilerdeki bu artışın veri kümeleri üzerinde yapılan analizlerin kalitesine etkisi yüksek olmaktır. Bu yüzden, bahsi geçen problemin çözülmesi önemli bir araştırma konusudur. Kayıp veri tahmini, bu problemi çözmeye yönelik uygulanan ve veri kalitesini arttıran kullanışlı tekniklerden biridir [2]. Veri analizinde, kayıp verilerin etkisini azaltmak için verilerin silinmesi (data deletion) ve verilerin uygun bir şekilde doldurulması (data imputation) yaygın olarak kullanılan iki yöntemdir [3].

Kayıp veriler üzerinde analiz yapılırken, satır ve sütun içerisinde eksik olan verilerin tamamen silinmesi veya bu verilerin göz ardı edilmesi yaklaşımları izlenmektedir. Eksik verilerin veri setinden tamamen silinmesi tekniğinde bilgi kaybı oluşur ve bu durum yapılan analizlerin hatalı sonuçlar vermesine yol açabilir [4]. Buna ek olarak eksik değerler için sıklıkla kullanılan bir başka yaklaşım verilerin ortalaması gibi sabit bir değerle boşlukların doldurulmasıdır. Ancak bu durumda da veriler arasındaki dağılım ve ilişkiler göz ardı edildiğinden hatalı veya yanlı sonuçların elde edilebileceği dikkate alınmalıdır [5]. Kayıp verilerin doldurulmasında kullanılan diğer yaklaşımlar ise, istatiksel, optimizasyon, makine öğrenmesi ve derin öğrenme yöntemleridir [6].

Toplanan veriler içerisindeki kayıp değerler, tek değişkenli ve çok değişkenli olmak üzere iki kategoriye ayrılabilir [5]. Bu çalışmada, Sivas Numune Hastanesi tıbbi atık su nötralizatör cihazından toplanan veriler içerisinden tek değişkenli kayıp pH değerlerinin tahmini üzerine odaklanılmıştır. Çalışmamızda, literatürde bilinen altı farklı makine öğrenme algoritması (LR, DVM, KNN, RO, KA) kayıp değerleri tahmin etmek için kullanılmıştır. Elde edilen sonuçlar, kayıp verilerin gerçeğe yakın bir şekilde tahmin edilebildiğini göstermiştir.

2. Literatür Özeti

Son yıllarda, literatürde kayıp verilerin tahminine yönelik yapılmış birçok çalışma bulunmaktadır. Güzel ve ark. [6], eksik sensör verisi sorununu çözmek için, derin öğrenme (Deep Learning, DL) ve Uyarlamalı-Ağ Tabanlı Bulanık Çıkarım Sistemine (Adaptive-Network Based Fuzzy Inference System, ANFIS) dayalı iki tahmin modelini kullanmıştır. Bu modellerin performansı, kök ortalama karesel hata (Root Mean Square Error, RMSE) metriği ile değerlendirilmiştir. Elde ettikleri sonuçlara göre hem DL hem de ANFIS yöntemlerinin, seçilen doğrusal olmayan regresyon modellerine kıyasla RMSE metriği açısından daha iyi olduğu görülmüştür. Abidin ve ark. [7], üç makine öğrenmesi (Machine Learning, ML) algoritmalarından, K-En Yakın Komşuluk (K-Nearest Neighbor, KNN), Karar Ağacı (Desicion Tree, DT) ve Bayes Ağları (Bayesian Networks, BN) kullanarak, dünya sağlık örgütü tarafından paylaşılan bazı veri kümeleri [8] üzerinde kayıp verileri tahmin ederek bu algoritmaların performansını değerlendirmiştir. Yaptıkları bu çalışmanın performans değerlendirmesinde, ortalama mutlak hata (MAE), MSE ve RMSE olmak üzere üç performans metriği kullanmışlardır. Tüm veri gruplarında, BN algoritması kayıp verilerin tahmininde üç metrik için tutarlı bir şekilde en düşük hatayı üretmiştir. Raja ve Thangavel [9], UCI veri havuzundan [10] aldıkları Dermatology, Pima, Wisconsin ve Yeast veri kümelerini kullanarak kayıp verileri doldurmak için Kaba (Rough) k-means algoritmasını kullanmışlardır. Kullandıkları bu yöntemi, klasik K-means algoritması, K-means parametre tabanlı veri doldurma yöntemi ve bulanık C-means parametre tabanlı veri doldurma yöntemi ile RMSE performans metriği açısından kıyaslamışlardır. Elde ettikleri sonuçlara göre Kaba K-means algoritmasının, kayıp değerlerin

doldurulmasında diğer algoritmalara kıyasla daha başarılı olduğu görülmüştür. Liu ve ark. [11], Avustralya'daki gerçek üretim tesislerinden 24 gün boyunca topladıkları sıcaklık, hız, titreşim, basınç gibi çeşitli parametrelerden oluşan sensör verilerini kullanarak, kayıp verilerin doldurulması için Çoklu Segmentli Boşluk Yineleme (Itr-MS-STLecImp) yöntemini kullanmışlardır. Bu modeli, kayıp verileri doldurma açısından yedi farklı zaman serileri yöntemi ile RMSE performans metriğini kullanarak karşılaştırmış ve Itr-MS-STLecImp modelinin diğer tüm yöntemlere kıyasla daha başarılı olduğunu göstermişlerdir. Lee ve ark. [5] IoT tabanlı büyük veri uygulamalarındaki tek değişkenli eksik değerleri tahmin etmek için zaman serilerine dayalı Eksik Veri Takas (MP-BMDI) algoritmasını önermişlerdir. Çalışmalarında, Güney Kore'de bulunan bir ofise sıcaklık, nem ve CO_2 değerlerini 10 saniyede bir kaydeden IoT izleme sistemi kurmuşlardır. Oluşturulan sistem ile 10 aylık veri toplanmış ve 3 haftalık veri eksik değerlerin tahmini için veri setinden çıkarılmıştır. Algoritmanın performans değerlendirmesinde RMSE ve MAPE metrikleri kullanılarak yedi farklı zaman serisi yöntemi ile karşılaştırılmış ve daha başarılı sonuçlar elde ettiği görülmüştür. Qin ve ark. [12] kuyu sularını izlemek için Çin'in Zhejiang kıyısında 16 farklı bölgeye şamandıra izleme sistemi ve karadan 4 farklı izleme merkezi kurarak veri toplamışlardır. Her şamandıra sistemine aynı anda sıcaklık, tuzluluk, pH ve çözünmüş oksijen gibi parametreleri ölçen birden fazla sensör eklemişlerdir. Yazarlar toplanan verilerdeki, eksik değerleri tahmin etmek için matris tamamlama tabanlı çoklu görüntülü öğrenme yöntemini (MC-MVL) önermişlerdir. Yapmış oldukları bu çalışmayı, ortalama bağıl hata (Mean Relative Error, MRE) ve MAE performans metriklerini kullanarak değerlendirmişlerdir. Önerdikleri modeli, 8 farklı alt model ile karşılaştırmış ve daha iyi sonuçlar elde etmişlerdir. Ma ve ark. [13], Cornell Üniversitesi'ndeki Tesis ve Kampüs içerisinde bulunan laboratuvar, öğrenci salonları ve kütüphane binalarında tüketilen enerji verilerini toplamışlardır. Yapmış oldukları çalışmada kayıp veri tahmini için Uzun Kısa Süreli Bellek modeli, Çift Yönlü veri doldurma ve Transfer Öğrenme yöntemlerini kullanarak LSTM-BIT'i önermişlerdir. Önerdikleri bu modeli RMSE performans ölçüm metriğini kullanarak istatistik, makine öğrenmesi ve derin öğrenme modelleri ile karşılaştırmışlardır. Elde ettikleri sonuçlara göre önerdikleri LSTM-BIT modelin diğer modellerin tamamından daha başarılı olduğunu gözlemlemişlerdir.

Literatürde kayıp verilerin tahminine yönelik, makine öğrenmesi, derin öğrenme ve farklı birçok yöntem bulunmaktadır. Bu yöntemler, enerji, sıcaklık, nem, hız, basınç ve CO_2 gibi farklı türde veri setlerini içeren problemlerin çözümüne yönelik uygulanmaktadır. Önerdiğimiz çalışma, literatürdeki çalışmalardan farklı olarak, tıbbi atık nötralizasyon sistemindeki kayıp pH verilerinin tahminini sağlamaktadır. Çalışmamızın temel katkısı, tahmin edilen kayıp değerler ile pH değerlerinin gözlemlenmesi ve çevrenin korunmasını desteklemesidir.

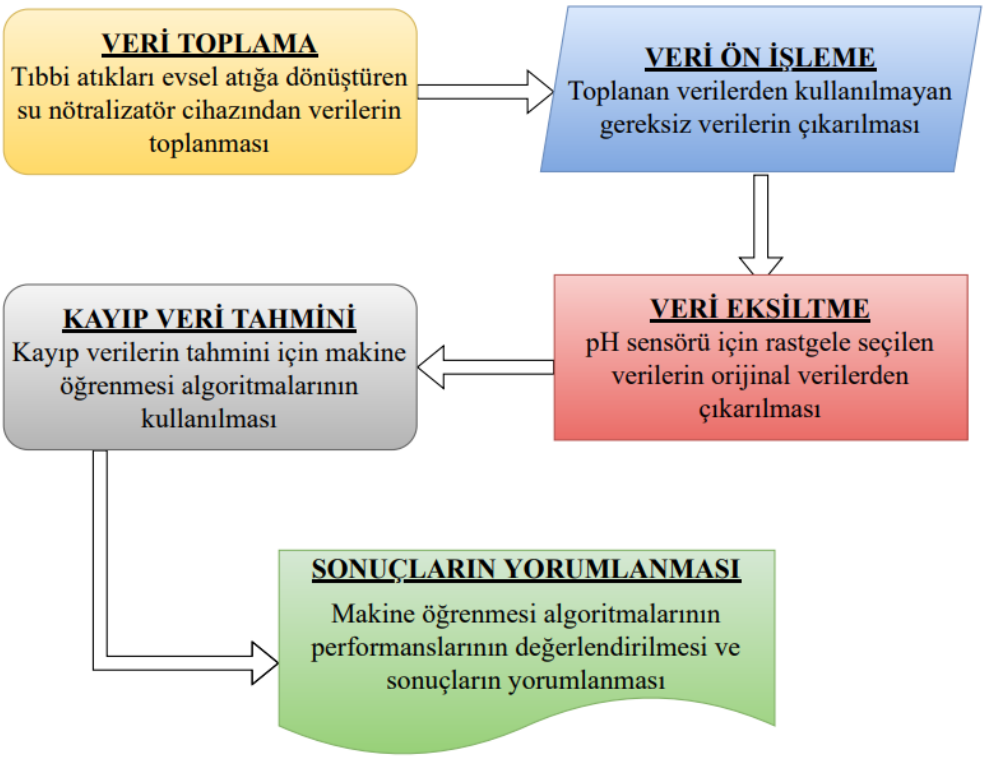
3. Materyal ve Yöntem

Dünya Sağlık Örgütü'nün paylaştığı verilere göre 13 Mayıs 2022 tarihine kadar dünya genelinde COVID-19 kaynaklı toplam 6.261.708 ölüm ve 517.648.631 onaylanmış vaka olduğu belirtilmiştir [14]. Tüm Dünya'yı etkileyen COVID-19 pandemisiyle birlikte hastanelerdeki tıbbi atık üretimi her geçen gün artmaktadır. Bu atıkların yönetimi ve içerisindeki zararlı mikroorganizmaların bertaraf edilmesi doğru bir şekilde yapılmadığında bitki, hayvan ve insan sağlığını önemli ölçüde tehdit etmektedir.

Hastanelerde bulunan tıbbi sıvı atıklar kan atıkları, diyaliz sıvı atıkları, üre ve sıvı patoloji numuneleri gibi atıklardan oluşmaktadır. Bu atıklar kanalizasyona aktarılmadan önce nötralizasyon sistemleri ile sıvı solüsyonlar basılarak pH değişikliği yoluyla kimyasal nötralizasyon işlemine maruz bırakılmaktadır. Sivas ilinin yerel yönetiminin 28/11/2014 tarihli toplantısında aldığı karara göre kanalizasyona deşarj edilecek sıvı atıkların pH aralığının 6.5 ile 10 aralığında olması gerektiği kararlaştırılmıştır [15]. Bu anlamda, tıbbi atık nötralizasyon sistemindeki pH değerlerinin

gözlemlenmesi ve bu veriler içerisinden toplanan eksik verilerin tahmin edilmesi çevrenin korunması adına oldukça önemlidir.

Bu çalışmada, Sivas Numune Hastanesinde bulunan tıbbi atıkları evsel atığa dönüştüren su nötralizatör cihazından toplanan veriler içerisindeki kayıp pH değerlerinin tahmini için makine öğrenmesi algoritmalarından; Doğrusal Regresyon (DR), Destek Vektör Makineleri (DVM), KNN, Rastgele Orman (RO), Adaboost ve Karar Ağacı (KA) kullanılmıştır. Yapılan çalışmaya ait işlem adımları Şekil 1'de gösterilmiştir.



Şekil 1. Kayıp verilerin makine öğrenmesi algoritmaları ile tahminine ait işlem adımları

"VERİ TOPLAMA" aşamasında, tıbbi atıkları evsel atığa dönüştürerek kanalizasyona deşarj eden nötralizatör cihazından veriler toplanarak MySQL veri tabanına kaydedilmesi sağlanmıştır.

"VERİ ÖN İŞLEME" aşamasında toplanan verilerden her bir sensöre ait veri okuma zamanı nitelikleri veri setinden çıkarılmıştır.

"VERİ EKSİLTME" aşamasında pH sensörü için Python programlama dilinde bir yazılım geliştirilmiş ve rastgele seçilen veriler %30 ve %40 oranında ayrı ayrı eksiltilerek ayrı bir dosyaya kaydedilmesi sağlanmıştır. Orijinal veriden çıkarılan bu veriler modelin tahmin değerlerinin doğruluğunun test edilmesinde kullanılmıştır.

"KAYIP VERİ TAHMİNİ" aşamasında altı farklı makine öğrenmesi algoritması Python dilinde kodlanmış ve kayıp pH değerlerinin tahmin edilmesi sağlanmıştır.

"SONUÇLARIN YORUMLANMASI" aşamasında makine öğrenmesi algoritmalarının performansı RMSE, MSE ve MAE metrikleri kullanılarak karşılaştırılmıştır.

3.1. Veri Seti

Bu çalışmada, Sivas Numune Hastanesinde bulunan tıbbi atıkları evsel atığa dönüştüren tıbbi atık su nötralizatör cihazından toplanan veriler; pH değeri, birbirine bağlı iki ayrı tıbbi atık deposuna ait doluluk seviyesi ve kırmızı, sarı, yeşil olmak üzere üç farklı alarm durumu verilerinden oluşmaktadır. Bu çalışmada, on dakika aralıklarla cihazda bulunan altı adet sensörden alınan bir aylık veri (4320 kayıt) kullanılmıştır. Makine öğrenmesi algoritmalarının performansını değerlendirebilmek için Tablo 1'de gösterildiği gibi iki farklı veri seti oluşturulmuştur. Bu veriler tıbbi atık sıvı nötralizasyon izleme sisteminden MySQL veri tabanına kayıpsız olarak kaydedilen bir aylık veriden oluşmaktadır.

Çalışmamızda, eksik verinin tahmini için pH sensörünün okuduğu değerlerin doldurulması üzerinde çalışılmıştır. Veri setinin hazırlık aşamasında, Numpy (Numerical Python) ve Sklearn kütüphaneleri kullanılarak Python programlama dili ile bir yazılım geliştirilmiştir. Geliştirilen bu yazılım sayesinde kayıp olmayan (gerçek pH değerlerinin) verilerin rastgele seçimi yapılmış ve çıkarılması sağlanmıştır. Çıkarılan bu değerler farklı bir dosyada saklanarak, daha sonra makine öğrenmesi algoritmasının yapmış olduğu tahminlerin test edilmesinde kullanılmıştır. İlk olarak veri seti 1 için pH verilerinin %30'u (1296) seçilerek kayıp değerler farklı bir dosyaya kaydedilmiş ve makine öğrenmesi algoritmasının tahmin ettiği değerlerin test edilmesi için kullanılmıştır. Kalan verilerin %70'i (3024) ise eğitim için ayrılmıştır. Daha sonra aynı işlem tekrar uygulanmış ve verilerin %40'ı (1728) çıkarılarak test için, kalan %60'ı da (2592) eğitim için ayrılarak 2. veri seti oluşturulmuştur.

Tablo 1. Veri Setleri Bilgisi

Veri Setleri	Kayıp Veri Yüzdesi	Eğitim	Test
Veri Seti 1	%30	3024	1296
Veri Seti 2	%40	2592	1728

3.2. Metot

Günümüzde, makine öğrenmesi yöntemlerinin kayıp veri tahmininde kullanımı popüler bir hale gelmiştir [6]. Kayıp veri tahmini konusunda kullanılacak algoritmaları seçerken literatürde bu konuda yapılan çalışmalar ve kullanılan makine öğrenmesi algoritmalarının yaygınlığı etkili olmuştur. Tercih edilen makine öğrenmesi algoritmaları şunlardır:

3.2.1. Doğrusal regresyon (DR):

DR yöntemi, bağımlı bir değişkenin bir veya birden fazla bağımsız değişken arasındaki ilişkisini incelemek için kullanılmaktadır. Tek değişkenli ve çok değişkenli olmak üzere iki tip regresyon bulunmaktadır. Tek değişkenli regresyonda, bir bağımlı değişken ile bir bağımsız değişken arasındaki ilişki incelenir. Çok değişkenli regresyonda ise birden fazla bağımsız değişken ile bağımlı değişken arasındaki ilişki incelenmektedir [16].

3.2.2. K-En yakın komşuluk (KNN):

En yakın komşu veri noktalarına dayanan KNN yöntemi, tanımlanamayan veri noktalarını bularak oylama sistemine göre sınıflandırmaktadır. Uygulanması oldukça kolay olan KNN yönteminin, geniş depolama alanı ihtiyacı, gürültüye duyarlılığı ve yavaş test prosedürü gibi dezavantajları bulunmaktadır [17]. Komşu verilerin sayısını ifade eden K parametresinin değeri kullanıcı tarafından belirlenmektedir. Bu çalışmada, farklı K parametreleri denenerek en uygun sonucu veren $K=5$ tercih edilmiştir. Tahmin için kullanılan ağırlık fonksiyonu "uniform" olarak tercih edilmiştir.

Bu fonksiyona göre komşu noktaların ağırlığı eşit olarak alınmaktadır. Algoritmanın mesafe ölçümünde, Öklid mesafesi (Euclidean distance) tercih edilmiştir.

3.2.3. Destek vektör makineleri (DVM):

Bir diğer makine öğrenmesi algoritması olan DVM yöntemi, veri noktaları arasındaki ayrımı en üst düzeye çıkaran yüksek boyutlu uzayda büyük hiper düzlemler oluşturmaktadır. Bu hiper düzlemleri oluşturmak için destek vektörleri kullanmaktadır. DVM daha iyi doğruluk sağlamasına rağmen hesaplama süresi bakımından maliyetlidir [17]. Bu algoritmanın çekirdek (kernel) türü için Radial Basis Function (RBF), çekirdek katsayısı (gamma) olarak ise "scale" kullanılmıştır. Çekirdek önbellek boyutu (cache size) 200 Megabayt (MB) olarak tercih edilmiştir.

3.2.4. Rastgele orman (RO):

Sınıflandırma ve regresyon analizleri için kullanılan RO algoritması, 2001 yılında Breiman tarafından geliştirilen bir topluluk öğrenme yöntemidir [18]. RO, en iyi tahmin sonucuna ulaşabilmek için karar ağaçlarından oluşan bir orman oluşturmaktadır. Tüm veri kümesi belirli sayıda karar ağaçlarından oluşan alt kümelere bölünmektedir. RO algoritması, tüm karar ağaçlarından elde edilen sonuçları birleştirerek nihai bir sonuca ulaşmaktadır [19]. Bu algoritma ile model eğitilirken ormandaki ağaç sayısı (n_estimators) 100 olarak belirlenmiştir.

3.2.5. Karar ağacı (KA):

Karar ağacı regresyonu, hedef değişkenler sürekli olduğunda kullanılabilen karar ağacı sınıflandırıcısının bir çeşididir. KA, karar ve yaprak düğümlerinden oluşmaktadır. KA regresyonunun düğüm hesaplamalarında, bilgi kazanımı yerine standart sapma kullanılmaktadır. Bir karar ağacının temel amacı, karmaşık bir kararı daha kolay yorumlanabilir bir çözüm sağlayacak basit kararlara bölmektedir [20]. Bu algoritma ile model eğitilirken minimum derinlik değeri 2 olarak belirlenmiştir.

3.2.6. Adaboost:

Adaboost, daha güçlü bir öğrenici (strong learner) oluşturmak için zayıf öğrenicileri (weak learner) bir araya getiren bir topluluk öğrenme yöntemidir. Topluluk yöntemleri, daha iyi tahmin performansı elde etmek için kendi içerisinde birden çok makine öğrenme algoritmasını kullanmaktadır. Adaboost yönteminde, güçlü bir öğrenici oluşturmak için her iterasyonda kendi hatalarından öğrenerek elde edilen zayıf öğreniciler belirli bir kural çerçevesinde birleştirilir [21]. Bu algoritma ile model eğitilirken yükseltmenin sonlandırıldığı maksimum tahmin edici sayısı 50 olarak belirlenmiştir.

4. Bulgular ve Tartışma

Bu bölümde, kayıp verilerin tahmininde kullanılan altı farklı makine öğrenme algoritmasının performansı değerlendirilmiş ve elde edilen sonuçlar paylaşılmıştır. Makine öğrenmesi algoritmalarının eğitiminde Tablo 1'de gösterilen %30 ve %40 oranında kayıp verilere sahip veri setleri kullanılmıştır. Daha sonra eğitimi gerçekleştirilmiş olan modellerin performansları MAE, MSE ve RMSE metrikleri kullanılarak değerlendirilmiştir. Bu metrikleri hesaplamak için kullanılan denklemlerde yer alan, y_i' değeri *i.* elemanın tahmin edilen değerini, y_i değeri ise *i.* elemanın gerçek değerini ve n kayıp verilerin sayısını ifade etmektedir.

Ortalama mutlak hata (MAE), tahmin edilen değer ile gerçek değer arasındaki ortalama farkı ölçmektedir. MAE, tahmin edilen her bir değer ile her bir gerçek değer arasındaki farkın mutlak değerinin ortalaması alınarak hesaplanmaktadır [7].

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^{n} |y_i' - y_i| \quad (1)$$

Ortalama karesel hata (MSE), tahmin edilen her bir değer ile gerçek değer arasındaki farkların karesinin ortalaması alınarak hesaplanmaktadır [7].

$$MSE = \left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i' - y_i)^2 \quad (2)$$

Ortalama karekök hatası (RMSE) tahmin edilen değer ile gerçek değer arasındaki mesafeyi ölçer. Yani modelin elde ettiği RMSE değeri, sıfıra ne kadar yakın olursa modelin tahmin ettiği değerler gerçek değerlere o kadar yaklaşmış olacaktır [7].

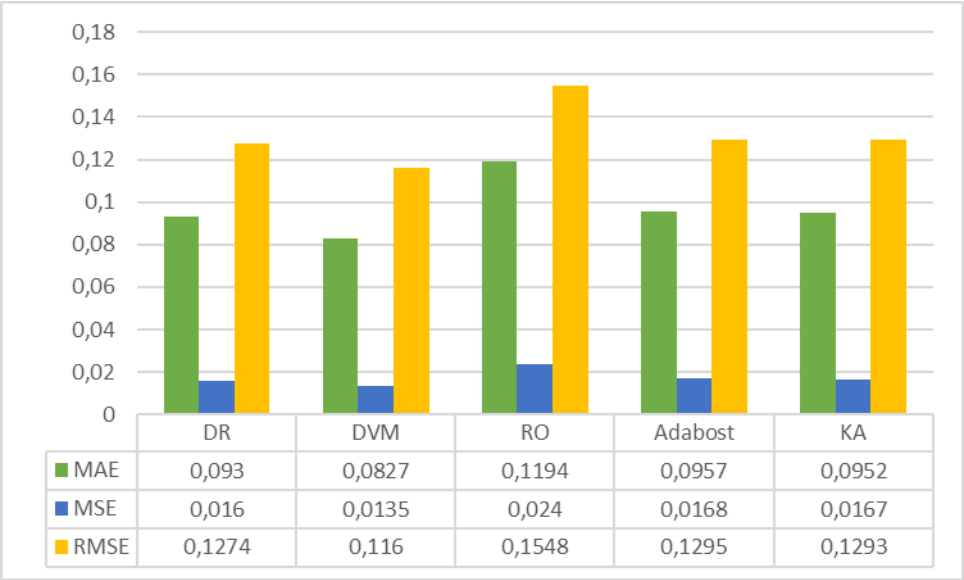
$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i' - y_i)^2} \quad (3)$$

Makine öğrenmesi algoritmaları kullanılarak Tablo 1'de verilen veri setlerinde bulunan kayıp pH değerleri doldurulmuş ve MAE, MSE, RMSE metrikleri kullanılarak tahmin edilen değerler ile gerçek değerler arasındaki hata oranları hesaplanmıştır. Makine öğrenmesi algoritmaları kullanılarak yapılan tahminler sonucu elde edilen metrik değerleri Tablo 2'de gösterilmiştir.

Tablo 2. Algoritmaların performans metrikleri

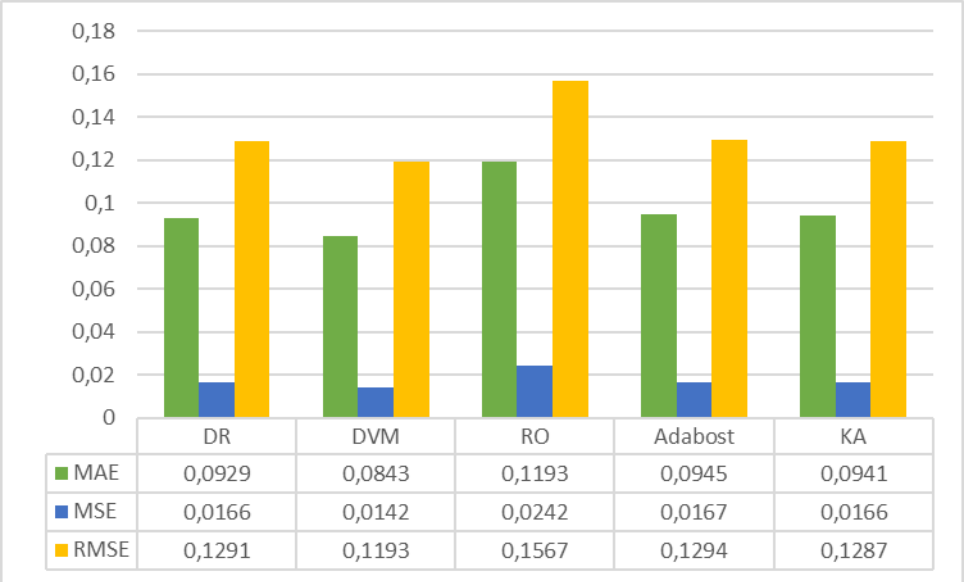
Veri Setleri	ML Algoritmaları	MAE	MSE	RMSE
Veri Seti 1 (%30)	DR	0.0930	0.0160	0.1274
	DVM	**0.0827**	**0.0135**	**0.1160**
	KNN	0.1194	0.0240	0.1548
	RO	0.0957	0.0168	0.1295
	Adabost	0.0952	0.0167	0.1293
	KA	0.1259	0.0271	0.1646
Veri Seti 2 (%40)	DR	0.0929	0.0166	0.1291
	DVM	**0.0843**	**0.0142**	**0.1193**
	KNN	0.1193	0.0242	0.1567
	RO	0.0945	0.0167	0.1294
	Adabost	0.0941	0.0166	0.1287
	KA	0.1236	0.0273	0.1651

Elde edilen sonuçlara göre, her iki veri seti için üç metrikte de en düşük hata değeri DVM algoritması daha sonra DR algoritması ile elde edilmiştir. Ayrıca, KA algoritmasının her iki veri seti için en düşük performansa sahip olduğu görülmüştür. Algoritmaların performans karşılaştırması grafiksel olarak¸ Şekil 2 ve Şekil 3'te gösterilmiştir.



	DR	DVM	RO	Adabost	KA
MAE	0,093	0,0827	0,1194	0,0957	0,0952
MSE	0,016	0,0135	0,024	0,0168	0,0167
RMSE	0,1274	0,116	0,1548	0,1295	0,1293

Şekil 2. Veri Seti 1 için algoritmaların performans karşılaştırması



	DR	DVM	RO	Adabost	KA
MAE	0,0929	0,0843	0,1193	0,0945	0,0941
MSE	0,0166	0,0142	0,0242	0,0167	0,0166
RMSE	0,1291	0,1193	0,1567	0,1294	0,1287

Şekil 3. Veri Seti 2 için algoritmaların performans karşılaştırması

5. Sonuç

Son yıllarda, IoT tabanlı uygulamalar akıllı üretimin ve dijital dönüşümün öne çıkan çözümlerinden biri haline gelmiştir. Bununla birlikte, bu uygulamalardan toplanan veriler her geçen gün daha fazla artış göstermektedir. Kayıp veriler, IoT mimarilerinde ön işleme aşamasının en büyük problemlerinden biridir [11]. Bu yüzden, kayıp verilerin doldurulması veri analizlerinin kalitesi ve veri takibinin güvenirliğinin sağlanması açısından oldukça önemli bir konudur. Bu çalışmada, literatürde yaygın olarak kullanılan makine öğrenmesi algoritmalarından, DR, DVM, KNN, RO, Adaboost ve KA algoritmaları kullanılarak kayıp IoT verilerin doldurulması başarılı bir şekilde gerçekleştirilmiştir. Algoritmaların başarıları karşılaştırılmış ve en yüksek başarıya sahip olan algoritmanın her iki veri seti için MAE, MSE ve RMSE performans metriklerinin tümünde DVM olduğu görülmüştür.

Yapılan çalışmada, verilerin %40 oranı gibi büyük miktarı kaybolmasına rağmen makine öğrenmesi yöntemlerinin bu verileri yüksek doğrulukta doldurabildiği görülmüştür. Ayrıca, makine öğrenmesi algoritmalarının çok düşük hata oranları ile kayıp verilerin doldurulmasında kullanılabileceği gösterilmiştir.

Gelecek çalışmalarda, bulanık mantık, optimizasyon ve derin öğrenme algoritmaları kullanılarak elde edilen sonuçlar makine öğrenmesi algoritmaları ile karşılaştırılabilir. Bu yöntemlerin kayıp verileri tamamlanmasındaki performansları incelenerek yeni yöntemlerin geliştirilmesi üzerine çalışmalar yapılabilir.

Yazarların Katkıları

FA, literatür araştırması, problemin ve yöntemin tanımlanması, makine öğrenmesi yöntemlerinin modellenmesi, modellerin test edilmesi ve makalenin yazılmasında; EÜ, verilerin toplanmasında, problemin ve yöntemin tanımlanması, geliştirilmesi ve makalenin yazılmasında katkı sağlamıştır.

Her iki yazar da makalenin son halini okudu ve onayladı.

Çıkar Çatışması

Yazarlar, çıkar çatışması olmadığını beyan eder.

Kaynaklar

[1]. Dubey, A., & Rasool, A. “Data mining based handling missing data.” In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 483-489, 2019.

[2]. Gond, V. K., Dubey, A., & Rasool, A. “A Survey of Machine Learning-Based Approaches for Missing Value Imputation.” In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 1-8, 2021.

[3]. Ma, J., Cheng, J. C., Ding, Y., Lin, C., Jiang, F., Wang, M., & Zhai, C. “Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series.” Advanced Engineering Informatics, 44, 101092, 2020.

[4]. Qin, Y., Sheng, Q. Z., Falkner, N. J., Dustdar, S., Wang, H., & Vasilakos, A. V. “When things matter: A survey on data-centric internet of things. Journal of Network and Computer Applications”, 64, 137-153, 2016.

[5]. Lee, G. H., Han, J., & Choi, J. K. “MPdist-based missing data imputation for supporting big data analyses in IoT-based applications.” Future Generation Computer Systems, 125, 421-432, 2021.

[6]. Guzel, M., Kok, I., Akay, D., & Ozdemir, S. “ANFIS and Deep Learning based missing sensor data prediction in IoT.” Concurrency and Computation: Practice and Experience, 32(2), e5400, 2020.

[7]. Abidin, N. Z., Ismail, A. R., & Emran, N. A. “Performance analysis of machine learning algorithms for missing value imputation.” International Journal of Advanced Computer Science and Applications, 9(6), 2018.

[8]. https://www.who.int/data/gho, Erişim tarihi 13.05.2022.

[9]. Raja, P. S., & Thangavel, K. J. S. C. “Missing value imputation using unsupervised machine learning techniques”, Soft Computing, 24(6), 4361-4392, 2020.

[10]. https://archive.ics.uci.edu/ml/index.php, Erişim Tarihi 13.05.2022.

[11]. Liu, Y., Dillon, T., Yu, W., Rahayu, W., & Mostafa, F. "Missing value imputation for industrial IoT sensor data with large gaps", IEEE Internet of Things Journal, 7(8), 6855-6867, 2020.
[12]. Qin, M., Du, Z., Zhang, F., & Liu, R. "A matrix completion-based multiview learning method for imputing missing values in buoy monitoring data", Information Sciences, 487, 18-30, 2019.
[13]. Ma, J., Cheng, J. C. P. F., Jiang, W., Chen, M. Wang, and C. Zhai, "A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data", Energy Build., 216, 109941, 2020.
[14]. https://covid19.who.int/, Erişim tarihi 13.05.2022.
[15]. https://www.sivas.bel.tr/Files/ATIKSU_YONETMELiiii.pdf, Erişim tarihi 13.05.2022.
[16]. Freedman, D. A. (2009). Statistical models: theory and practice. cambridge university press.
[17]. Kaur, P., Kumar, R., & Kumar, M. "A healthcare monitoring system using random forest and internet of things (IoT)", Multimedia Tools and Applications, 78(14), 19905-19916, 2019.
[18]. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
[19]. Ani, R., Krishna, S., Anju, N., Aslam, M. S., & Deepa, O. S. "Iot based patient monitoring and diagnostic prediction tool using ensemble classifier", In 2017 International conference on advances in computing, communications and informatics (ICACCI), 1588-1593, 2017.
[20]. Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. "Decision tree regression for soft classification of remote sensing data", Remote Sensing of Environment, 97(3), 322-336, 2005.
[21]. Jhaveri, S., Khedkar, I., Kantharia, Y., & Jaswal, S. "Success prediction using random forest, catboost, xgboost and adaboost for kickstarter campaigns", In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 1170-1173, 2019.