Araştırma Makalesi
Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi
24 (2024) 475-488

VERİ ÖN İŞLEME TEKNİKLERİNİN SAĞLIK VERİLERİNİN SINIFLANDIRMA BAŞARISINA ETKİSİNİN İNCELENMESİ

Feyza ERDOĞAN[1], Vahit TONGUR[2*], Betül UZBAŞ[3]

[1,2] Konya Teknik Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Yazılım Mühendisliği Bölümü, Konya, 42250, Türkiye

[3]Konya Teknik Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, Konya, 42250, Türkiye

Geliş Tarihi/Received Date: 08.04.2024 Kabul Tarihi/Accepted Date:22.11.2024 DOI:10.54365/adyumbd.1466631

ÖZET

Veri madenciliği sürecinin en temel adımlarından biri olan veri ön işleme teknikleri, literatürde sıklıkla başvurulan bir süreçtir. Bu çalışmada Hepatit hastalığına ait veri kümesi üzerinde sağlık alanında sık kullanılan veri ön işleme tekniklerinin etkinliği incelenmiştir. Sırasıyla eksik veri, dengesiz veri kümesi, aykırı veri, normalizasyon ve özellik seçimi işlemleri uygulanmıştır. Veri kümesinin her adımda elde edilen yeni versiyonu için literatürde sıklıkla kullanılan beş makine öğrenmesi yöntemi (KNN, LR, RF, SVM, ANN) ile sınıflandırma yapılmıştır. Elde edilen sonuçlar, doğru ve gerekli veri ön işleme tekniklerinin seçimi ile model başarısına olumlu katkısını desteklemektedir. Tüm aşama sonunda elde edilen model performansları %85 ve üzerinde olup, tüm performans belirleme ölçütleri bazında tutarlı sonuçlar göstermektedir. Her bir veri ön işleme model performansına kademeli olarak katkıda bulunmuş, en yüksek katkı ise son aşamada uygulanan özellik seçimi ile sağlanmıştır. Özellik seçimi, modelin performansını belirgin şekilde iyileştirerek sınıflandırma başarısına önemli ölçüde katkı sağlamıştır.

Anahtar Kelimeler: *Hepatit, Makine Öğrenmesi, Sağlık Veri Kümesi, Veri Ön İşleme*

INVESTIGATION OF THE EFFECT OF DATA PRE-PROCESSING TECHNIQUES ON THE CLASSIFICATION SUCCESS OF HEALTH DATA

ABSTRACT

Data preprocessing techniques, one of the most fundamental steps in the data mining process, are frequently referenced in the literature. In this study, the effectiveness of commonly used data preprocessing techniques in the health field was examined on a dataset related to Hepatitis disease. The processes of handling missing data, managing imbalanced datasets, outlier detection, normalization, and feature selection were applied in sequence. For each new version of the dataset obtained at every step, classification was performed using five machine learning methods commonly used in the literature (KNN, LR, RF, SVM, ANN). The results obtained support the positive contribution of correctly selecting the appropriate data preprocessing techniques to model success. The model performances achieved in all steps are above 85%, showing consistent results across all performance evaluation metrics. Each data preprocessing step contributed gradually to model performance, with the highest contribution provided by the feature selection applied in the final stage. Feature selection significantly enhanced the model's performance, making a substantial contribution to classification success.

Keywords: *Data Preprocessing, Health Dataset, Hepatitis, Machine Learning*

e-posta[1] : ferdogan@ktun.edu.tr ORCID ID: https://orcid.org/0000-0002-9750-0495

[*] e-posta[2] : vtongur@ktun.edu.tr ORCID ID: https://orcid.org/0000-0001-5419-7839 (Sorumlu Yazar)

e-posta[3] : buzbas@ktun.edu.tr ORCID ID: https://orcid.org/0000-0002-0255-5988

1. Giriş

Günümüz çağı, büyük veri çağı olarak adlandırılmaktadır. Sağlık, mühendislik, finans gibi daha birçok alanda, veriler bir araya gelerek veri kümelerini oluşturmaktadır. Veri sayısının her geçen gün artması, veri kümelerinin işlenmesi hususunda bazı problemleri de beraberinde getirmektedir [1, 2].

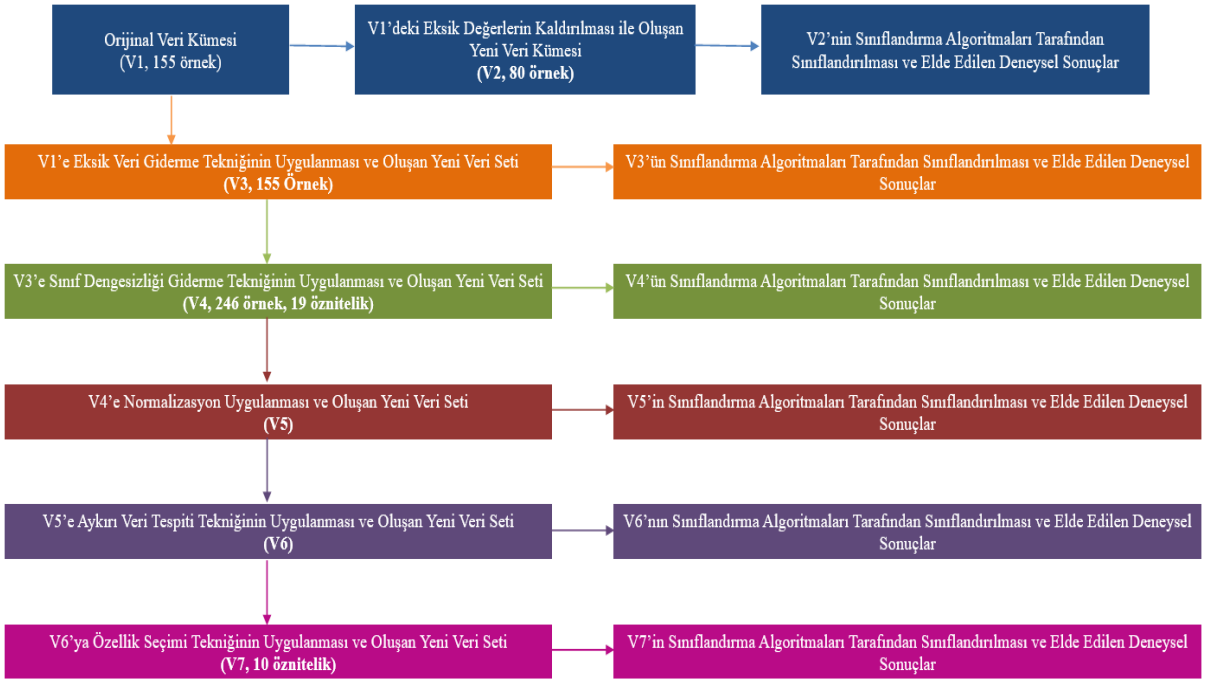
Veri madenciliği, veriyi keşfetme ve faydalı bilgiye erişme süreci olarak tanımlanabilmektedir [3, 4]. Bu sayede, elde edilen faydalı bilgiler ilişkilendirilerek amaca uygun bir şekilde geleceğe yönelik kullanılabilmektedir [5].

Veri madenciliği sürecinde faydalı bilgiye erişilebilmesi, kullanılan veri kümesinin yapısı ve kalitesi ile doğrudan ilişkilidir [4, 6]. Veri madenciliğinde önemli bir aşamaya sahip olan veri ön işleme, bu doğrultuda yararlanılan kritik bir adımdır. Veri ön işleme sürecindeki işlemler kullanılan veri kümesine bağlı olarak seçilmelidir. Veri ön işleme aşamasındaki temel adımlar; eksik veri giderme (veri temizleme), aykırı veri tespiti (veri temizleme), dengesizlik sorunu (veri artırma), normalizasyon (veri dönüştürme) ve özellik seçimidir (veri indirgeme) [4, 7, 8].

Günümüzde var olan birçok veri kümesinde yukarıda bahsi geçen problemler sıklıkla yer almaktadır. Özellikle sağlık verilerinde karşılaşılan en büyük sorun, eksik veri, aykırı (hatalı) veri ve dengesizlik sorunudur [9]. Eksik veri oluşumu genel olarak hastalardan talep edilen bilgilerin tamamının sağlanamamasından kaynaklanır. Aykırı veri oluşumu ise ölçüm yapılan cihaz, ölçüm yapan personel, ölçüm sırasında gerçekleşen hatalı/eksik eylem gibi nedenlerle ortaya çıkmaktadır. Dengesizlik sorunu ise veri kümesinde yer alan farklı sınıf etiketine sahip örnek sayılarının adil dağılıma sahip olmamasıdır [9-10]. Örneğin; iyi huylu ve kötü huylu olmak üzere iki adet sınıf değerinden oluşan bir kanser veri kümesinin olduğu varsayılsın. Veri kümesindeki toplam 500 örnek hasta verisinin yer aldığını ve bu örneklerin 100 adet iyi huylu; 400 adet kötü huylu hasta örneği olduğunu düşünülürse dengesiz bir veri kümesi olduğunu söylenebilir.

Makine öğrenmesi, günümüzde adından sıkça söz ettirmekte olup; birçok algoritmayı içerisinde barındırmaktadır [11, 12]. Bu algoritmalar, veriler üzerinde matematiksel ve istatistiksel bazı yöntemler kullanarak bir model ortaya koymakta ve anlamlı sonuçlar üretmektedir [10, 13]. Algoritmalar sonucu elde edilen model, tahmin yeteneğine sahiptir ve tahmin başarısı doğrudan model ile ilişkilidir. Modelin başarısı da doğrudan veri kümesi ile ilişkilidir. Çünkü makine öğrenmesi algoritmaları, dengeli ve veri eksiği bulunmayan veri kümeleri için tasarlanmışlardır [3, 9]. Dolayısıyla bu tür eksiklikleri içeren kalitesiz bir veri kümesinde algoritma iyi performans gösteremeyecektir. Bu doğrultuda, başarılı bir makine öğrenmesi süreci için veri kümesi üzerinde gerekli olan veri ön işleme tekniklerinin uygulanması kaçınılmazdır.

Literatürdeki çalışmalar, veri ön işleme adımlarını genellikle tekil teknikler (örneğin, yalnızca eksik veri giderme veya yalnızca sınıf dengesizliği giderme) veya birkaç tekniğin kombinasyonu şeklinde ele almaktadır. Bu çalışmada ise, diğer çalışmalardan farklı olarak veri ön işlemenin tüm temel adımları (eksik veri giderme, aykırı veri tespiti, sınıf dengesizliği giderme, normalizasyon ve özellik seçimi) aşamalı olarak ve üst üste uygulanmıştır. Çalışma, özellikle *Hepatitis* veri seti üzerinde eksiksiz bir veri ön işleme süreci sunarak, bu veri kümesi için tüm adımları kapsayan en kapsamlı çalışma olma özelliğini taşımaktadır. Bu çok aşamalı ön işleme yaklaşımı ile veri kalitesini artırarak sınıflandırma başarısına katkı sağlanması amaçlanmış ve sağlık verilerinin sınıflandırılmasında veri ön işleme tekniklerinin toplu etkisi ortaya konmuştur. Böylece, çalışma hem veri ön işlemeye bütünsel yaklaşımıyla hem de *Hepatitis* veri seti üzerinde kapsamlı uygulamasıyla literatüre yenilikçi bir katkı sunmaktadır. Çalışmada uygulanan veri ön işleme adımlarını ve her aşamanın veri kümesi üzerindeki etkisini görselleştiren akış şeması, Şekil 1'de sunulmuştur.



Şekil 1. Çalışmada uygulanan veri ön işleme adımlarının akış şeması

Çizelge 1. Literatür özeti

Makale	Veri Seti	Problem	Yöntem	Sonuç
Özuğur ve Orman [9]	PIMA	Eksik Değer, Dengesiz Veri	SMOTEENN, MICE	%91 (F-score)
Mitra ve Samanta [12]	Hepatitis	Eksik Değer, Özellik Seçimi	EMB, RS	%100 (Sınıflandırma Doğruluğu)
Saygın ve Baykara [10]	ILPD	Özellik Seçimi	SFS	LGBM %82.12, MLP %81.13, DT %81.13, SVM %77.87 ve LR %77.80
Orooji ve Kermani [13]	Hepatitis	Dengesiz Veri	Random Over-Sampling, Random Under-Sampling	Over-Sampling: sensitivity, accuracy ve f-score (%99.9); specificity %100
Nahzat ve Yağanoğlu [11]	PIMA	Normalizasyon, Eksik Veri	Mean/Median	%88.31 (RF-Accuracy)

Literatürde sağlık verilerinin makine öğrenmesi algoritmaları ile sınıflandırılması adına pek çok çalışma yer almaktadır. Çizelge 1'de literatürde yer alan bazı çalışmalar yazar bilgisi, kullanılan veri

seti, veri setine ait problem, veri setine uygulanan en başarılı veri ön işleme tekniği (*Yöntem*), ve elde edilen başarı sonucu (*Sonuç*) şeklinde özetlenmiştir.

Bu çalışmanın geriye kalan kısmı şu şekilde organize edilmiştir: ikinci bölümde çalışmada kullanılan veri seti ve veri seti üzerinde uygulanan veri ön işleme teknikleri hakkında bilgilendirme yapılmıştır. Üçüncü bölümde her bir yöntemden elde edilen deneysel sonuçlar sunulmuştur. Son olarak dördüncü bölümde ise çalışma sonucunda elde edilen temel sonuçlar ve bu çalışmanın gelecekteki çalışmalara katkısı yorumlanmıştır.

2. Materyal ve Metod

Bu bölüm üç temel başlıktan oluşmaktadır. İlk olarak, çalışmada kullanılan veri seti hakkında bilgilendirme yapılmıştır. Ardından, veri seti üzerinde uygulanan veri ön işleme teknikleri hakkında bilgi verilmiştir. Son olarak, sınıflandırma algoritmalarının başarı kıyaslaması için kullanılan performans belirleme ölçütleri sunulmuştur.

2.1. Veri Seti Tanıtımı

Hepatitis veri seti, Gail Gong tarafından 1988 yılında literatüre kazandırılmıştır [14]. Kaliforniya Üniversitesi Irvine Makine Öğrenmesi Veritabanı'nda yer alan veri seti [15], hepatit hastası bireylerin ölçüm değerlerinden oluşan bir sınıflandırma veri setidir. Hayatta kalan ve ölü olmak üzere iki sınıfı olan veri setinde; toplamda 155 vaka örneği ve sınıf özniteliği dışında 19 öznitelik bulunmaktadır [16]. Bu özniteliklerden bazıları kategorik veri tipine sahipken; veri ön işleme tekniklerinin uygulanabilmesi adına sayısal veri tipine dönüştürülmüştür. Çizelge 2'de veri setinde yer alan öznitelikler hakkında açıklamalar verilmiştir.

Çizelge 2. Hepatitis veri setinin temel yapısı

Öznitelik	Veri Tipi	Aralık
Age	Sayısal	[7,78]
Sex	Sayısal	[1,2]
Steroid	Kategorik	[1,2]
Antivirals	Sayısal	[1,2]
Fatigue	Kategorik	[1,2]
Malaise	Kategorik	[1,2]
Anorexia	Kategorik	[1,2]
LiverBig	Kategorik	[1,2]
LiverFirm	Kategorik	[1,2]
SpleenPalpable	Kategorik	[1,2]
Spiders	Kategorik	[1,2]
Ascites	Kategorik	[1,2]
Varices	Kategorik	[1,2]
Bilirubin	Kategorik	[3,8]
AlkPhosphate	Kategorik	[26,295]
Sgot	Kategorik	[14,648]
Albumin	Kategorik	[2.1,6.4]
Protime	Kategorik	[0,100]
Histology	Sayısal	[1,2]

2.2. Veri Ön İşleme

Bu başlıkta, veri seti üzerinde sırasıyla uygulanan beş ön işleme tekniği ve bu teknikleri uygularken kullanılan yöntemler yer almaktadır.

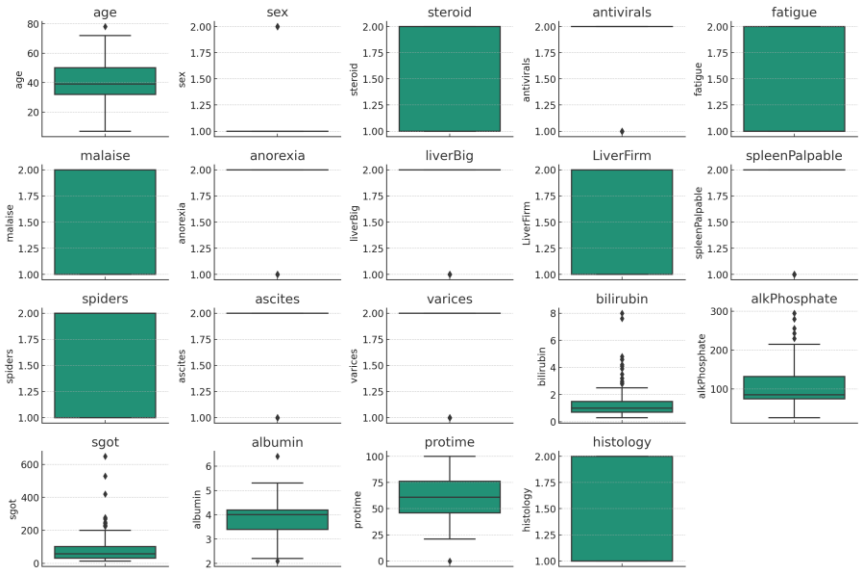
2.2.1. Eksik Veri Giderme

Hepatit veri setinde 167 adet eksik veri (missing value) bulunmaktadır. Bu da tüm veri setinin yaklaşık %5.39'una tekabül etmektedir. Bazı veri ön işleme teknikleri ve sınıflandırma algoritmaları eksik veri içeren veri setleri üzerinde kullanılamamaktadır. Bu sebeple bir veri seti eksik değer içeriyorsa ilk adım eksik verileri gidermek olacaktır. Literatürde eksik verileri gidermek için; istatistiksel, basit, makine öğrenmesi teknikleri kullanan ve karmaşık yapıya sahip yöntemler bulunmaktadır [9, 16]. Bu çalışmada, hepatit veri seti üzerinde eksik verileri gidermek için istatistiksel bir yöntem olan ortalama ile doldurma (mean imputation) yöntemi kullanılmıştır. Bu yöntem eksik verileri, ait olduğu sütundaki verilerin aritmetik ortalamasını kullanarak doldurmaktadır. Ortalama ile doldurma yöntemi basit, uygulaması kolay yöntemdir. Az oranda eksik veriye sahip ve eksik veri dağılımının düzensiz olduğu durumlarda kullanışlı olması nedeniyle tercih edilmiştir.

2.2.2. Aykırı Veri Tespiti

Aykırı veri (outlier data), bir veri setini oluşturan örneklerden büyük oranda farklılık gösteren veri olarak adlandırılmaktadır [17]. Veri girişi sırasında yapılan hatalar, hatalı ölçüm gibi nedenlerden dolayı aykırı veri oluşabilmektedir [18]. Aykırı veriler, özellikle sonrasında gerçekleştirilecek ön işleme adımları ve sınıflandırma aşamasında modelin yanılgıya düşmesine neden olabilmektedir. Bu doğrultuda, modelin performansı olumsuz yönde etkilenecek ve güvenilirliği azalacaktır. Aykırı veri tespiti, bu problemlerin çözümü için kullanılan ön işleme tekniklerinden biridir. Veri setindeki aykırı değerleri tespit etmek için; görselleştirme, IQR, Z-score, Hampel vb. literatürde birçok farklı yöntem bulunmaktadır [17-19].

Eksik veri giderme işlemi sonucu elde edilen yeni Hepatit veri seti üzerinde, kutu grafiği kullanılarak aykırı veri tespiti gerçekleştirilmiştir. Şekil 2'de her bir öznitelik için kutu grafiği ile aykırı veri tespiti sonuçları sunulmuştur. Şekil 2'de görüldüğü üzere, bazı kutuların alt ve üst kısımlarına dağılmış küçük noktalar yer almaktadır. Bu noktalar, aykırı veri olarak isimlendirilmekte olup; özellikle bilirubin, alkPhosphate, sgot ve protime özniteliklerinde dikkat çekmektedir.



Şekil 2. Aykırı verilere ait kutu grafiği sonuçları

Bu çalışmada aykırı veri tespiti için Çeyrekler Arası Açıklık (*Interquartile Range, IQR*) yöntemi kullanılmıştır. Bu yöntem, veri setinin yayılımını ölçmek için aykırı veri tespitinde kullanılan istatistiksel bir yöntemdir [20, 21]. IQR yönteminde öncelikle veriler küçükten büyüğe sıralanır. Ardından veri setinin alt yarısındaki (*Q1*) ve üst yarısındaki (*Q2*) ortanca değerler hesaplanarak farkı alınır. Bu fark değeri IQR değerini temsil etmekte olup, genellikle $Q1-1.5\times IQR$ değerinin altında kalan ve $Q3+1.5\times IQR$ değerinin üstünde kalan veriler aykırı veri olarak adlandırılmaktadır.

IQR yöntemi ile aykırı veriler tespit edildikten sonra aykırı değerler bulunduğu öznitelik değerlerinin aritmetik ortalaması ile doldurularak tamamlanmıştır.

2.2.3. Sınıf Dengesizliği Problemi

Sınıf dengesizliği, bir veri setindeki sınıf değerlerinin örnek sayısı bakımından dengesiz bir dağılıma sahip olması şeklinde açıklanabilir [9, 22]. Sınıf dengesizliği bulunan veri setlerinde makine öğrenmesi algoritmaları taraflı bir performans sergileyebilir. Çoğunluk sınıfı örneğinin çok olması nedeniyle model, azınlık sınıfını ayırt edememekte ve böylece eksik öğrenme veya aşırı öğrenme gibi problemler meydana gelerek model başarısı olumsuz yönde etkilenmektedir [23].

Aşırı yeniden örnekleme, melez yeniden örnekleme, sentetik yeniden örnekleme ve eksik yeniden örnekleme yöntemleri literatürde sınıf dengesizliği problemini çözmek için önerilen yöntemlerdir [9].

Bu çalışmada sentetik yeniden örnekleme yöntemlerinden olan SMOTE yöntemi kullanılmıştır. Sentetik yeniden örnekleme, adından da anlaşılacağı üzere azınlık sınıfı örnekleri üzerinde keşif yaparak yeni yapay örnekler üretme yöntemidir [9, 23]. Aşırı yeniden örnekleme yönteminden farklı olarak örnekleri doğrudan kopyalamaması; çıkarımlar ile yeni yapay örnekler oluşturması aşırı öğrenme riskini azaltmaktadır. SMOTE, veri setindeki azınlık sınıfına ait örneklerden benzer formdaki örnekleri seçme ve bu örnekler arasında sentetik veri noktaları oluşturarak yapay örnek üretimine dayanan bir yöntemdir [9, 23, 24]. Özellikle, tıbbi tanı veri setlerinde başarılı bir şekilde kullanılmaktadır.

2.2.4. Normalizasyon

Veri setlerinde yer alan bazı öznitelikler farklı değer aralıklarına sahip olabilmektedir. Çizelge 2'de sunulan Hepatit veri setine ait aralık değerlerine bakıldığında, öznitelikler arasında farklılıklar olduğu dikkat çekmektedir. Özellikle sınıflandırma algoritmalarının başarısında böylesine farklı ve uç değer aralıkları model performansını oldukça etkilemektedir. Veri setindeki tüm verileri matematiksel yöntemler kullanarak yeniden düzenleme işlemi normalizasyon olarak adlandırılmaktadır [10]. Hepatit veri setinde Min-Max Normalizasyonu yöntemi kullanılmıştır. Bahsi geçen yöntemin matematiksel formülü Denklem 1'de verilmiştir.

$$x_{yeni} = \frac{x - min\ (x)}{max(x) - min\ (x)} \tag{1}$$

Min-max yöntemi; literatürde sıklıkla kullanılan, basit ve öznitelikleri belirli bir değer aralığında yeniden ölçeklendirmeyi sağlayan bir normalizasyon yöntemidir [25, 26]. Bu yöntemle veriler [0,1] aralığında yeniden ölçeklendirilir. Denklem 1'de *max(x)* ve *min(x)* sırasıyla ilgili öznitelik için maksimum ve minimum değerleri ifade etmektedir. *x*, ölçeklendirilecek verinin mevcut değeri; x_{yeni} ölçeklendirme işlemi sonucu elde edilen [0,1] aralığındaki yeni veri değerini temsil etmektedir.

2.2.5. Özellik Seçimi

Bazı veri setlerinde doğrudan sonucu etkilemeyen alakasız/gereksiz öznitelikler bulunmaktadır. Bu tür özelliklerin varlığı işlem süresini uzatma, gereksiz bellek kullanımı, gereksiz özellik nedeniyle model başarısını etkileme gibi sorunlara yol açabilmektedir [10, 27, 28]. Özellik seçimi, veri setindeki sonucu doğrudan etkilemeyen alakasız/gereksiz özellikleri kaldırarak; veri setini en iyi şekilde temsil eden özellik alt kümesini bulma işlemidir [1, 29].

Özellik seçimi yöntemleri, kendi içerisinde üç başlık altında incelenmektedir. Filtre yöntemleri, herhangi bir öğrenme veya sınıflandırma algoritması kullanmadan; verilerin genel özelliklerine bakarak özelliğin seçilip/seçilmemesine karar veren yöntemlerdir. Sarmalayıcı yöntemler, öğrenme veya sınıflandırma algoritması kullanarak; özelliğin seçilip/seçilmemesine karar veren ve değerlendiren yöntemlerdir. Gömülü yöntemler ise filtre ve sarmalayıcı yöntemlerin bir birleşimi olarak düşünülebilir [2, 30].

Bu çalışmada Hepatit veri seti üzerinde sarmalayıcı yöntemlerden olan Ardışık Öznitelik Seçimi (*Sequential Backward Selection, SBS*) yöntemi kullanılmıştır. SBS, veri setindeki her bir özelliğin teker teker elenmesi mantığına dayanan ve her aşamada model performansını dikkate alarak eleme işlemini gerçekleştiren bir özellik seçim yöntemidir [31, 32]. Yöntemin, özellikleri teker teker ele alarak değerlendirmesi hem karmaşıklığı azaltma hem de aşırı öğrenmeden kaçınma noktasında oldukça önemlidir.

2.3. Performans Belirleme Ölçütleri

Hepatit veri seti üzerinde gerçekleştirilen her bir veri ön işleme tekniği sonrası oluşan yeni veri seti literatürde sıklıkla kullanılan K-En Yakın Komşu (*K-Nearest Neighbors, KNN*), Lojistik Regresyon (*Logistic Regression, LR*), Rastgele Orman (*Random Forest, RF*), Destek Vektör Makinesi (*Support Vector Machine, SVM*) ve Yapay Sinir Ağları (*Artifical Neural Network, ANN*) makine öğrenmesi algoritmalarıyla sınıflandırma işlemine uygulanmıştır. Bu yöntemlerin ve veri ön işleme tekniklerinin model başarısına etkisini belirlemek adına literatürde sıklıkla başvurulan performans belirleme ölçütleri kullanılmıştır.

Kullanılan makine öğrenmesi yöntemleri için veri seti %70 oranında eğitim %30 oranında test veri kümelerine ayrılmıştır. *KNN* algoritmasında *k* değeri 5 olarak belirlenmiştir. *ANN*'de giriş katmanındaki nöron sayısı ilgili adımda oluşan veri setinin özellik sayısına eşit olacak şekilde, gizli katmandaki nöron sayısı; *n* giriş katmanındaki nöron sayısı olmak üzere *2n+1* şeklinde ve çıkış katmanındaki nöron sayısı ise veri setindeki sınıf sayısına bağlı olmak şartıyla 1 olarak belirlenmiştir. RF algoritmasında ağaç sayısı 100 olarak belirlenmiştir.

Karmaşıklık Matrisi (*Confusion Matrix, CM*), sınıflandırma problemlerinde bir modelin tahminleri ile gerçek değerlerini kıyaslayarak; doğru ve hatalı tahmin sayılarını gösteren bir performans belirleme yöntemidir. *CM* sonucu elde edilen veriler ile model performansını değerlendirmek için kullanılan 4 önemli ölçütün hesabı yapılabilmektedir. Bunlar; doğruluk (*accuracy*), hassasiyet (*precision*), duyarlılık (*recall/sensitivity*) ve F_1-skoru (F_1-*Score*) ölçütleridir [33].

Bu çalışmada bahsi geçen beş makine öğrenmesi algoritmasının performans değerlendirmesi doğruluk, hassasiyet, duyarlılık ve F_1-skoru ölçütleri kullanılarak yapılmıştır. Bahsi geçen dört performans belirleme ölçütünün matematiksel formülleri Denklem 2-5'te sunulmuştur [10].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

Denklemlerde sunulan *TP* (*True Positive, Doğru Pozitif*) terimi, doğru tahmin edilen hayatta kalan hepatit hastalarının sayısını belirtmektedir. Örnek olarak, gerçekten hayatta kalan bir hastanın model tarafından hayatta kalan olarak sınıflandırılması verilebilir. *FP* (*False Positive, Yanlış Pozitif*) terimi, yanlış tahmin edilen hayatta kalan hepatit hastalarının sayısını belirtmektedir. Örnek olarak, aslında ölü olan bir hastanın model tarafından yanlışlıkla hayatta kalan olarak sınıflandırılması verilebilir. *FN* (*False Negative, Yanlış Negatif*) terimi, yanlış tahmin edilen ölü hepatit hastalarının sayısını belirtmektedir. Örnek olarak, aslında hayatta kalan bir hastanın model tarafından yanlışlıkla ölü olarak sınıflandırılması verilebilir. *TN* (*True Negative, Doğru Negatif*) terimi, doğru tahmin edilen ölü hepatit hastalarının sayısını belirtmektedir. Örnek olarak, gerçekten ölü olan bir hastanın model tarafından ölü olarak sınıflandırılması verilebilir.

Doğruluk (*accuracy*) ölçütü, modelin tüm tahminlerinin doğru olma oranını gösterir. Modelin doğru sınıflandırma oranını ölçerek genel başarı düzeyini değerlendirmeye yarar. Bu ölçüt, tüm pozitif ve negatif sınıflandırmaların doğru tahmin edilip edilmediğine odaklanır.

Hassasiyet (*precision*) ölçütü, modelin "hayatta kalan" olarak sınıflandırdığı hastaların gerçekten hayatta kalma durumunu ifade eder. Yanlış pozitiflerin azaltılmasına odaklanır ve modelin doğruluğunu artırır. Hasta verisinde hassasiyet, modelin hayatta kalan olarak sınıflandırdığı hastaların gerçekten hayatta kalanların oranını gösterir.

Duyarlılık (*recall/sensitivity*) ölçütü, modelin gerçekten hayatta kalan hastaları doğru bir şekilde "hayatta kalan" olarak sınıflandırma oranını gösterir. Yanlış negatiflerin azaltılmasına odaklanır. Hasta verisinde duyarlılık, modelin hayatta kalan hastaları kaçırmadan doğru sınıflandırma yeteneğini gösterir.

F_1-skoru (*F_1-score*) ölçütü, hassasiyet ve duyarlılığın harmonik ortalaması olarak hesaplanır. Modelin "hayatta kalan" sınıfını tespit etmedeki genel dengesini ifade eder ve özellikle dengesiz veri kümelerinde faydalıdır. Hasta verisinde F1-skoru, hassasiyet ve duyarlılık değerleri arasında bir denge sağlayarak modelin genel sınıflandırma başarısını ifade eder.

3. Deneysel Sonuçlar

Bu bölümde, her bir veri ön işleme tekniği sonrası elde edilen yeni veri seti üzerinde *KNN, LR, RF, SVM* ve *ANN* algoritmaları tarafından yapılan sınıflandırma işleminden elde edilen sonuçlar yer almaktadır.

Çizelge 3'te; eksik değerlerin yer aldığı örneklerin doğrudan veri setinden kaldırıldığı ve hiçbir veri ön işleme tekniği uygulanmamış orijinal Hepatitis veri seti üzerinde gerçekleştirilen deney sonuçlarına yer verilmiştir. Veri setinden eksik değerlerin yer aldığı örneklerin kaldırılması sonucu 80 adet örnek üzerinden ilgili işlemler yapılmıştır.

Çizelge 3'teki performans sonuçlarına bakıldığında, eksik veriler nedeniyle veri setindeki örnek sayısının azaldığı ve bunun da sınıflandırıcıların performansında tutarsızlıklara yol açtığı görülmektedir. Özellikle veri setinin dengesiz yapısı, bazı sınıflandırıcıların performansını olumsuz etkilemiştir. Örneğin, *RF* algoritması, genelde küçük ve dengesiz veri kümelerinde gösterdiği dayanıklılık nedeniyle diğer algoritmalara kıyasla daha iyi sonuçlar vermiştir. Buna karşılık, *SVM* algoritmasının düşük *F_1-skoruna* sahip olması, dengesiz veri dağılımına uyum sağlayamamasından kaynaklanmakta ve modelin sınıflandırma başarısını düşürmektedir. *RF* algoritmasının, *F_1-skor* dışında tüm ölçütlerde diğer

yöntemleri geride bırakması dikkat çekicidir. F_1-skordaki düşüklüğün sebebi, sınıf dengesizliği nedeniyle bazı algoritmaların belirli bir sınıfa ait örnekleri daha iyi tanımasından kaynaklanan eksik veya aşırı öğrenme olarak değerlendirilebilir. Bir yöntemin başarılı sayılabilmesi için tüm ölçütlerde tutarlı performans göstermesi önemlidir.

Çizelge 3. Orijinal Veri Kümesi Üzerinde Performans Sonuçları

Ölçüt/Algoritma	***KNN***	***LR***	***RF***	***SVM***	***ANN***
Accuracy	0.88	0.84	**0.92**	0.79	0.88
Precision	0.87	0.84	**0.92**	0.69	0.89
Recall	0.88	0.84	**0.92**	0.79	0.88
F_1-score	0.87	0.84	0.74	0.74	**0.88**

Çizelge 4'te eksik veri giderme işlemi sonrası oluşan yeni Hepatit veri seti üzerinde makine öğrenmesi yöntemlerinin performans sonuçları yer almaktadır. Burada dikkat edilmesi gereken husus; örnek sayısının orijinal halini alarak 155 olmasıdır.

Çizelge 4. Eksik Veri Giderme Sonrası Performans Sonuçları

Ölçüt/Algoritma	***KNN***	***LR***	***RF***	***SVM***	***ANN***
Accuracy	0.81	**0.83**	**0.83**	0.79	0.77
Precision	0.79	**0.81**	**0.81**	0.75	0.73
Recall	0.81	**0.83**	**0.83**	0.79	0.77
F_1-score	0.79	**0.81**	0.79	0.76	0.75

Çizelge 4'teki verilere bakıldığında, örnek sayısındaki artışın performansa etkisi açıkça görülebilmektedir. Eksik veri giderme işlemi, veri setindeki örnek sayısını 155'e çıkararak daha fazla bilgi sağlamış, ancak bu artış bazı ölçütlerde olumlu etki yaratırken bazı ölçütlerde negatif bir etkiye yol açmıştır. Özellikle, örnek sayısının artmasıyla birlikte sınıf dengesizliğinin de artması, sınıflandırıcıların performansında dengesizliğe ve genel sınıflandırma başarısının düşmesine neden olmuştur. Bu gibi değişimler, veri setinin ihtiyaç duyduğu ön işleme tekniklerinin belirlenmesine katkı sağlamaktadır. Eksik veri giderme işlemi sonrasında *KNN*, *LR* ve *RF* algoritmalarının benzer performans göstermesi, artan örnek sayısının sınıflandırma dengesini olumlu etkilediğini göstermektedir. Ancak, özellikle *LR* ve *RF* algoritmalarında sınıf dengesizliği nedeniyle sınıflandırma doğruluğunun optimal seviyeye ulaşamadığı gözlemlenmiştir. Eksik veri giderme işlemi ile örnek sayısı artsa da sınıf dengesizliği sorunu performansı sınırlamaya devam etmektedir.

Çizelge 5'te sınıf dengesizliği probleminin çözümü sonrası oluşan yeni Hepatit veri seti üzerinde makine öğrenmesi yöntemlerinin performans sonuçları yer almaktadır. Burada dikkat edilmesi gereken husus; azınlık sınıfına ait örnek sayısının artışıyla güncel örnek sayısının 246 olmasıdır.

Çizelge 5. Sınıf Dengesizliğini Giderme Sonrası Performans Sonuçları

Ölçüt/Algoritma	***KNN***	***LR***	***RF***	***SVM***	***ANN***
Accuracy	0.82	0.84	**0.93**	**0.93**	**0.93**
Precision	0.85	0.84	**0.93**	**0.93**	**0.93**
Recall	0.82	0.84	**0.93**	**0.93**	**0.93**
F_1-score	0.82	0.84	**0.93**	**0.93**	**0.93**

Çizelge 5'teki verilere bakıldığında, önceki iki adımdaki performans sonuçlarına kıyasla ciddi bir artış olduğu açıkça görülmektedir. Eksik veri giderme ve dengesiz veri probleminin çözümü sonrasında makine öğrenme yöntemleri, her iki sınıf hakkında dengeli bilgiye sahip olarak daha başarılı bir öğrenme ve sınıflandırma süreci geçirmektedir. Aksi halde, model yalnızca tek bir sınıfa yanlı davranabilir ve bu durum performans düşüşüne yol açabilir. Bu adımda dikkat edilmesi gereken bir diğer husus, her bir yöntemin tüm ölçütler için tutarlı sonuçlar elde etmesidir. *SMOTE* yöntemi ile sınıf dengesizliği sorunu giderilmiş ve azınlık sınıfındaki örnek sayısı artırılmıştır. Bu adımın ardından *RF*, *SVM* ve *ANN* algoritmaları %93 doğruluk oranına ulaşarak dengeli bir veri setinin sınıflandırıcılar üzerindeki olumlu etkisini göstermiştir. Bu üç algoritmanın yüksek doğruluk ve tutarlı F1-skora sahip olması, *SMOTE* yönteminin sınıf dengesizliği sorununu başarılı bir şekilde gidermiş ve model başarısını artırmış olduğunu kanıtlamaktadır. *KNN* ve *LR* algoritmalarının diğer algoritmalardan daha düşük performans göstermesi, komşuluk temelli algoritmaların dengesiz veri setlerinde daha hassas olması ile ilişkilendirilebilir.

Çizelge 6'da normalizasyon sonrası ölçeklendirilen yeni Hepatit veri seti üzerinde makine öğrenmesi yöntemlerinin performans sonuçları yer almaktadır.

Çizelge 6. Normalizasyon Sonrası Performans Sonuçları

Ölçüt/Algoritma	***KNN***	***LR***	***RF***	***SVM***	***ANN***
Accuracy	0.82	0.84	**0.93**	**0.93**	0.92
Precision	0.85	0.84	**0.93**	**0.93**	0.92
Recall	0.82	0.84	**0.93**	**0.93**	0.92
F_1-score	0.82	0.84	**0.93**	**0.93**	0.92

Çizelge 6'daki verilere bakıldığında, normalizasyon işlemi sonrasında makine öğrenmesi yöntemlerinin performans sonuçlarında belirgin bir değişim olmadığı görülmektedir. Normalizasyon işlemi ile veri setindeki tüm özellikler aynı ölçeğe getirilmiştir. Böylece herhangi bir özelliğin model üzerinde aşırı etkili olmasının önüne geçilmiştir. Bu durum, sınıflandırma algoritmalarının genel performansına katkı sağlasa da Hepatit veri setinde özellikler arasında büyük ölçek farklılıkları bulunmadığından, performansta anlamlı bir değişim yaratmamıştır. Çizelgeden de anlaşılacağı üzere, *KNN* ve *RF* algoritmaları normalizasyon sonrası daha tutarlı bir performans gösterirken, normalizasyonun *SVM* ve *ANN* üzerindeki etkisi sınırlı kalmıştır. Bu durum, Hepatit veri setinde ölçek farklılıklarının ciddi bir sınıflandırma sorunu oluşturmadığını ve normalizasyonun sınıflandırıcılar üzerindeki etkisinin göreceli olarak düşük olduğunu göstermektedir.

Çizelge 7'de aykırı veri tespiti sonrası elde edilen yeni Hepatit veri seti üzerinde makine öğrenmesi yöntemlerinin performans sonuçları yer almaktadır.

Çizelge 7. Aykırı Veri Tespiti Sonrası Performans Sonuçları

Ölçüt/Algoritma	***KNN***	***LR***	***RF***	***SVM***	***ANN***
Accuracy	0.81	0.79	**0.85**	0.81	0.83
Precision	0.80	0.77	**0.87**	0.77	0.81
Recall	0.81	0.79	**0.85**	0.81	0.83
F_1-score	0.80	0.78	**0.81**	0.78	0.79

Çizelge 7'deki verilere bakıldığında, aykırı veri tespiti sonrasında tüm yöntemlerin performansında gözlemlenen negatif etki dikkat çekmektedir. Veri setindeki aykırı değerler, doğrudan sınıflandırma algoritmalarının performansını etkileyebilmektedir. Aykırı veri tespiti ve işlenmesi, veri setindeki aşırı değerleri düzeltmeye odaklanmakta olup, bu işlem bazen veri setinin doğal varyansını azaltabilir veya verilerin temsil gücünü değiştirebilir. Eğer aykırı değerler, veri setinin önemli bir

parçasını oluşturuyorsa, bu değerlerin işlenmesi modelin gerçek dünya verilerini yorumlama yeteneğini olumsuz etkileyebilir. Aykırı veri tespiti sonrasında RF algoritması, diğer algoritmalara kıyasla en yüksek doğruluk oranını göstermiş olsa da genel olarak performans değerlerinde bir miktar düşüş gözlemlenmiştir. Bu düşüş, aykırı veri işleminin veri setinin doğal varyans ve temsil gücünde değişiklikler yaratması nedeniyle yaşanmış olabilir. Özellikle *SVM* ve *LR* algoritmalarında gözlemlenen performans düşüşü, aykırı veri işleminin bazı algoritmaların uyum yeteneğini olumsuz etkilediğini göstermektedir.

Çizelge 8'de özellik seçimi sonrası elde edilen yeni Hepatit veri seti üzerinde makine öğrenmesi yöntemlerinin performans sonuçları yer almaktadır. Özellik seçimi işlemi sonucunda *sex, steroid, antivirals, fatigue, LiverFirm, spleenPalpable, spiders, ascites, varices* ve *bilirubin* olmak üzere 10 adet öznitelik bulunmaktadır.

Çizelge 8. Özellik Seçimi Sonrası Performans Sonuçları

Ölçüt/Algoritma	***KNN***	***LR***	***RF***	***SVM***	***ANN***
Accuracy	0.89	0.85	0.91	0.93	**0.96**
Precision	0.90	0.85	0.91	0.93	**0.96**
Recall	0.89	0.85	0.91	0.93	**0.96**
F_1-score	0.89	0.85	0.91	0.93	**0.96**

Çizelge 8'deki verilere bakıldığında, özellik seçimi sonrasında tüm yöntemlerin performansında kayda değer bir artış olduğu görülmektedir. Her bir sınıflandırma yöntemi farklı yapılara sahip olduğu için, uygulanan ön işleme tekniklerinin her yöntemde aynı etkiye sahip olmaması beklenebilir. Çizelge 3 ile Çizelge 8 arasında bazı adımlarda negatif sonuçlar gözlemlense de son durumda elde edilen performans artışı ve ölçütler arasındaki tutarlılık, veri ön işleme tekniklerinin sınıflandırma üzerindeki olumlu etkisini açıkça göstermektedir. Özellik seçimi sonrası *ANN, SVM* ve *RF* algoritmalarının %90'ın üzerinde doğruluk oranlarına ulaşması, alakasız özelliklerin çıkarılmasının model performansını iyileştirdiğini göstermektedir. Özellik seçimi ile gereksiz veri yükü azaltılmış ve modelin daha odaklanmış bir şekilde öğrenmesi sağlanmıştır. Özellikle *ANN* algoritmasının %96 doğruluk oranına ulaşması, özellik seçiminin çok katmanlı modellerin sınıflandırma başarısını artırmadaki potansiyelini vurgulamaktadır.

4. Sonuç ve Öneriler

Sağlık verileri, veri ön işleme tekniklerinin yoğun olarak kullanıldığı alanlardan biridir. Bu verilerde en sık karşılaşılan sorunlar, eksik veri, sınıf dengesizliği ve aykırı değerlerdir. Bu tür sorunlar, sınıflandırma işlemlerinde makine öğrenmesi algoritmalarının başarısını olumsuz etkileyebilir. Bu nedenle, sağlık verilerinin analizi ve sınıflandırılması öncesinde kapsamlı bir veri ön işleme süreci uygulamak büyük önem taşır.

Bu çalışmada, karaciğer iltihaplanması (hepatit) hastalığına ait bir sınıflandırma veri seti üzerinde kapsamlı veri ön işleme adımları uygulanarak eksik veri giderme, dengesiz veri seti problemi, aykırı veri tespiti, normalizasyon ve özellik seçimi tekniklerinin sınıflandırma performansına olan etkisi incelenmiştir. Her bir veri ön işleme adımı sonrasında veri seti, *KNN, LR, RF, SVM* ve *ANN* algoritmaları ile sınıflandırılmış ve doğruluk, hassasiyet, duyarlılık ve *F_1-skoru* ölçütlerine göre performans değerlendirmesi yapılmıştır.

Deneysel sonuçlar, veri ön işleme tekniklerinin tüm sınıflandırma algoritmaları üzerinde genel olarak olumlu bir etkiye sahip olduğunu göstermiştir. Özellikle sınıf dengesizliği giderildikten ve aykırı veri tespiti yapıldıktan sonra *RF, SVM* ve *ANN* algoritmalarında %93 doğruluk oranına ulaşılmış ve genel performansta tutarlı bir artış sağlanmıştır. Bununla birlikte, eksik veri giderme ve normalizasyon

gibi tekniklerin etkisinin veri setinin özelliklerine göre değişkenlik gösterdiği, sınıflandırma algoritmalarının performansında farklı derecelerde etki yarattığı gözlemlenmiştir.

Eksik veri giderme, veri setindeki örnek sayısını artırarak modelin daha fazla bilgiye dayalı öğrenmesini sağlamıştır. Ancak, özellikle sınıf dengesizliği sorunu olan veri setlerinde eksik verinin doldurulması, sınıflandırıcıların performansını sınırlayabilir. Aykırı veri tespiti, veri setinin doğal varyansını azaltarak sınıflandırıcılar üzerinde yanlılığı azaltmıştır. Bununla birlikte, bazı aykırı veriler veri setinin önemli bir parçası olabileceğinden, bu adım performansta beklenmedik düşüşlere neden olabilir. SMOTE yöntemi ile sınıf dengesizliği giderilmiş ve modelin her iki sınıf hakkında dengeli bir öğrenme süreci geçirmesi sağlanmıştır. Bu adım, tüm algoritmalarda belirgin bir performans artışı sağlamış, özellikle *RF, SVM* ve *ANN* algoritmalarında yüksek doğruluk oranlarına ulaşılmıştır.

Özelliklerin aynı ölçeğe getirilmesi ile bazı algoritmaların performansında tutarlılık sağlanmıştır. Ancak, veri seti özelliklerinin başlangıçta ölçek olarak yakın olduğu durumlarda normalizasyonun etkisi sınırlı kalmıştır. Özellik seçimi ile veri setinden gereksiz öznitelikler çıkarılmış ve modelin öğrenme sürecine odaklanarak daha başarılı sonuçlar elde edilmiştir. Özellikle *ANN* algoritmasının %96 doğruluk oranına ulaşması, özellik seçiminin model performansına olumlu katkısını göstermektedir.

Bu çalışmanın sonucunda elde edilen bulgular, veri ön işlemenin özellikle sağlık veri kümelerinde sınıflandırma başarısını artırmada kritik bir rol oynadığını ve tüm adımların üst üste ardışık uygulanmasının model performansını büyük ölçüde iyileştirdiğini ortaya koymaktadır.

Bu çalışma kapsamında her bir ön işleme adımında tek bir yöntem tercih edilmiştir. Eksik veri gidermede ortalama ile doldurma yöntemi; aykırı veri tespitinde IQR yöntemi, dengesiz veri gidermede SMOTE yöntemi, normalizasyon kısmında min-max ölçeklendirme yöntemi ve son olarak özellik seçiminde SBS yöntemi kullanılmıştır. Gelecek çalışmalar için bunlar dışında kalan diğer ön işleme teknikleri kullanılabilir ve performans kıyaslaması yapılabilir. Aynı şekilde çalışmada kullanılan *KNN, LR, RF, SVM* ve *ANN* makine öğrenmesi yöntemleri dışında diğer sınıflandırma yöntemleri de tercih edilerek model başarıları tespit edilebilir.

Katkı Beyanı

Yazarların çalışmadaki katkı içerikleri aşağıda belirtilmiştir.

Feyza Erdoğan: Literatür taraması, yöntemlerin belirlenmesi ve uygulanması, makale yazımı

Vahit Tongur: Yöntemlerin belirlenmesi, bulguların yorumlanması, makale yazımı,

Betül Uzbaş: Araştırma tasarımı ve fikir, bulguların yorumlanması

Çıkar Çatışması Beyanı

Makale yazarları herhangi bir kurum, kuruluş, kişi ile kişisel ve finansal çıkar çatışması olmadığını beyan etmektedirler.

Kaynaklar

[1] Erdoğan F. İkili gri kurt optimizasyon algoritmasinin ikili optimizasyon problemlerine uygulanmasi. Yüksek lisans tezi. Konya: Necmettin Erbakan Üniversitesi; 2023.
[2] Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. ACM Computing Surveys (CSUR) 2017; 50(6): 1-45.
[3] Dogan A, Birant D. Machine learning and data mining in manufacturing. Expert Systems with Applications 2021; 166: 1-22.

[4] Oğuzlar A. Veri ön işleme. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 2003; 21: 67-76.
[5] Nart A. Kalp hastaliklarini tahmin etmede veri madenciliği teknikleriyle etkili algoritmanin tespit edilmesi. Yüksek lisans tezi. Ankara: Gazi Üniversitesi; 2023.
[6] García S, Luengo J, Herrera F. Data preprocessing in data mining. 72 Cham, Switzerland:Springer; 2015.
[7] García S, Ramírez-Gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: methods and prospects. Big Data Analytics 2016; 1: 1-22.
[8] Zelaya CVG. Towards explaining the effects of data preprocessing on machine learning. In: IEEE 35th International Conference on Data Engineering (ICDE), Macau SAR, China; 2019.
[9] Özoğur HN, Orman Z. Sağlik verilerinin analizinde veri ön işleme adimlarinin makine öğrenmesi yöntemlerinin performansina etkisi. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi 2023; 16(1): 23-33.
[10] Saygın E, Baykara M. Karaciğer yetmezliği teşhisinde özellik seçimi kullanarak makine öğrenmesi yöntemlerinin başarılarının ölçülmesi. Fırat Üniversitesi Mühendislik Bilimleri Dergisi 2021; 33(2): 367-377.
[11] Nahzat S, Yağanoğlu M. Diabetes prediction using machine learning classification algorithms. Avrupa Bilim ve Teknoloji Dergisi 2021; 24: 53-59.
[12] Mitra M, Samanta RK. A study on UCI hepatitis disease dataset using soft computing. Model. Meas. Control C 2017; 78(4): 467-477.
[13] Orooji A, Kermani F. Machine learning based methods for handling imbalanced data in hepatitis diagnosis. Frontiers in Health Informatics 2021; 10: 1-6.
[14] Bache K, Lichman M. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences 2013.
[15] UCI Machine Learning Repository. https://doi.org/10.24432/C5Q59J (Erişim tarihi: 21.12.2024).
[16] Rosly R, Makhtar M, Awang MK, Awang MI, Rahman MNA. Analyzing performance of classifiers for medical datasets. International Journal of Engineering & Technology 2018; 7: 136-138.
[17] Boukerche A, Zheng L, Alfandi O. Outlier detection: Methods, models, and classification. ACM Computing Surveys (CSUR) 2020; 53(3): 1-37.
[18] Alimohammadi H, Chen SN. Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis. Expert Systems with Applications 2022; 191: 1-10.
[19] Xu H, Zhang L, Li P, Zhu F. Outlier detection algorithm based on k-nearest neighbors-local outlier factor. Journal of Algorithms & Computational Technology 2022; 16: 1-12.
[20] Dash CSK, Behera AK, Dehuri S, Ghosh A. An outliers detection and elimination framework in classification task of data mining. Decision Analytics Journal 2023; 6: 1-8.
[21] Fredianto F, Putri DAP. Comparison of the interquartile range algorithm and local outlier factor on Australian weather data sets. In: Proceeding of International Summit on Education, Technology, and Humanity 2021, Surakarta, Indonesia; 2021.
[22] Bölükbaşı İB. Dengesiz bir diyabet veri setinde makine öğrenmesi yöntemlerini kullanarak diyabet hastalığının teşhisi. Yüksek lisans tezi. Bursa: Uludag Üniversitesi; 2023.
[23] Dablain D, Krawczyk B, Chawla NV. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems 2022; 34(9): 6390-6404.
[24] Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya INH, Ismail M. SMOTE for handling imbalanced data problem: A review. In: Sixth international conference on informatics and computing (ICIC) Jakarta, Indonesia; 2021.
[25] Henderi H, Wahyuningsih T, Rahwanto E. Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. International Journal of Informatics and Information Systems 2021; 4(1): 13-20.
[26] Yavuz S, Deveci M. İstatiksel normalizasyon tekniklerinin yapay sinir ağin performansina etkisi. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 2012; (40): 167-187.
[27] Dash M, Liu H. Feature selection for classification. Intelligent data analysis 1997; 1(1-4): 131-156.

[28] Kumar V, Minz S. Feature selection: A literature review. Smart Computing Review 2014; 4: 211–229.
[29] Al-Wajih R, Abdulkadir SJ, Aziz N, Al-Tashi Q, Talpur N. Hybrid binary grey wolf with Harris hawks optimizer for feature selection. IEEE Access 2021; 9: 31662-31677.
[30] Agrawal P, Ganesh T, Mohamed AW. Chaotic gaining sharing knowledge-based optimization algorithm: An improved metaheuristic algorithm for feature selection. Soft Computing 2021; 25(14): 9505-9528.
[31] Alnowami MR, Abolaban FA, Taha E. A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer. Journal of Radiation Research and Applied Sciences 2022; 15(1): 104-110.
[32] Yao G, Hu X, Wang G. A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain. Expert Systems with Applications 2022; 200: 1-23.
[33] Cengil E, Çınar A. Göğüs verileri metrikleri üzerinden kanser sınıflandırılması. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi 2020; 11(2): 513-519.