Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Veri Ölçekleme ve Eksik Veri Tamamlama Yöntemlerinin Makine Öğrenmesi Yöntemlerinin Başarısına Etkisinin İncelenmesi

Mesut POLATGİL [a,*]

[a] *Bilişim Sis ve Tek Bölümü, Şarkışla Uygulamalı Bilimler Yüksekokulu, Sivas Cumhuriyet Üniversitesi, Sivas, TÜRKİYE*

** Sorumlu yazarın e-posta adresi: mesutpolatgil@cumhuriyet.edu.tr*

DOI:10.29130/dubited.948564

Öz

Teknoloji ve bilişim alanındaki yenilikler ile elde edilen verinin büyüklüğü ve çeşitliliği artarak bu verilerin kaydedilmesi ve paylaşılması da kolaylaşmıştır. İnsan eli ile analiz edilmesi oldukça zor olan bu verilerin analizinde bilgisayarlar ve özellikle makine öğrenmesi algoritmaları büyük rol oynamaktadır. Bu analiz sürecinde veri ön işleme aşaması veri üzerinde yapılan çalışmalarda kilit rol oynamaktadır. Veri ön işleme aşamasında eksik verilerin tamamlanması ve veri ölçekleme işlemi gerçekleştirilmektedir. Literatürde eksik veri tamamlaması ile veri ölçekleme yöntemlerinin algoritmalar üzerindeki etkisini ayrı ayrı gösteren çalışmalar bulunmaktadır. Fakat bu iki önemli aşamanın bir arada değerlendirilmesi de gerekmektedir. Bu çalışmada Hepatoselüler Karsinoma (HCC) hastalığı veri seti üzerinde eksik verilerin tamamlanması ve veri ölçekleme yaklaşımlarının Yapay Sinir Ağları, Destek Vektör Makinaları ve Rassal Orman Algoritmalarının sınıflandırma başarılarına etkisi araştırılmıştır. Araştırma sonucunda en iyi sınıflandırmanın eksik verilerin tamamlanmasında ortalama yaklaşımı kullanılması ve min-max veri ölçeklemesi ile gerçekleştiği tespit edilmiştir. Ayrıca sınıflandırma açısından Rassal Orman algoritmasının diğer algoritmalara göre daha başarılı olduğu tespit edilmiştir.

Anahtar Kelimeler: *Eksik veri, Hepatoselüler Karsinoma, Veri Ölçekleme, Makine öğrenmesi.*

Investigation of The Effects of Data Scaling and Imputation of Missing Data Approaches on The Success of Machine Learning Methods

Abstract

With the innovations in technology and informatics, the size and diversity of the data obtained has increased and it has become easier to record and share this data. Computers and especially machine learning algorithms play a major role in the analysis of this data, which is very difficult to analyze by human hands. In this analysis process, the data preprocessing stage plays a key role in studies on data. In the data preprocessing stage, the missing data is completed and the data scaling process is carried out. In the literature, there are studies that show the effects of missing data completion and data scaling methods on algorithms separately. However, these two important stages need to be evaluated together. In this study, the completion of missing data on the Hepatocellular Carcinoma (HCC) disease data set and the effect of data scaling approaches on the classification success of Artificial Neural Networks, Support Vector Machines and Random Forest Algorithms were investigated. As a result of the research, it was determined that the best classification was achieved by using the mean approach to complete the missing data and min-max data scaling. In addition, it has been determined that the random forest algorithm is more successful than other algorithms in terms of classification.

Keywords: *Missing data, Hepatocellular Carcinoma, Data Scaling, Machine learning*

Geliş: 06/06/2021, Düzeltme: 13/03/2022, Kabul: 15/03/2022

I. GİRİŞ

Bilgi toplumuna geçilmesi ile birlikte, verinin üretilerek depolanması ve paylaşılmasını kolaylaşmış ve bilgi daha da önem kazanmıştır. Başta bulut bilişim gibi gelişen bilişim teknolojisi araçları ile büyük veriler rahatlıkla saklanabilmektedir. Bu büyüklükteki verilerin insan eli ile analiz edilmesi ise oldukça güçtür. Kolaylıkla saklanabilen ve ulaşılabilen büyük verilerden anlamlı bilgiler çıkarmak veri madenciliği ve makine öğrenmesi yöntemleri ile mümkün hale gelmiştir. Bu sayede kredi risk analizi, dolandırıcılık tespiti, müşteri analizi ve tıbbi teşhis gibi birçok alanda başarılı sonuçlar alınması mümkün olmuştur. Bilişim teknolojilerinde yaşanan gelişmeler ile teknolojinin hayatımızdaki yerinin artması kaydedilen ve saklanan veri miktarını ve çeşitliliğini artıracaktır. Bu durum verinin işlenmesi ve anlamlı bilgiler çıkarılması noktasında gerçekleştirilen çalışmaların artarak devam edebileceğini göstermektedir. Ayrıca başta sağlık hizmetleri olmak üzere önemli alanlarda birçok önemli kararın verilerden anlamlı sonuçlar çıkarabilen makineler ve yapay zekâ tarafından verilmesi günümüzde kullanılmakta ve giderek yaygınlaşmaktadır.

Makine öğrenme algoritmaları toplanan veriler ile çalışmaktadır. Kurulan modellerin başarısı büyük oranda bu verilerin tam ve doğru olması ile mümkün olabilmektedir. Verilerde, yanlış kayıtlar ya da rassal olarak meydana gelen eksik veriler sıklıkla karşılaşılan bir durumdur. Kayıp veri olması durumunda istatistiksel temellere dayanan makine öğrenme algoritmalarının sonuçları etkilenebilmektedir [1]. Eksik verinin tamamlanması ile ilgili farklı yaklaşımlar mevcuttur. Bunlardan en temel olanları eksik verinin yerine ortalama, medyan ya da mod gibi istatistiksel verilerin kullanılması, bir önce ve bir sonraki verinin kaydırılması ya da eksik veri kaydının silinmesidir [2]. Eksik veri problemi çözüldükten sonra belirlenmesi gereken bir diğer önemli husus ise verilerin ölçeklenmesidir. Veri ölçeklenmesi özellikle özniteliklerin farklı değer aralıklarında olduğu durumlarda daha da önem kazanmaktadır. Örneğin bir veri setinde yaş ya da kilo gibi değişkenler ile sigara içip içmeme durumu, bir dokuya ait ölçü verisi farklı aralıklarda yer alabilir ve bu farklı aralık durumu algoritmaların başarısını etkileyebilir. Bu yüzden verilerin ölçeklendirilmesi gerekmektedir ve bu noktada farklı ölçekleme yaklaşımları bulunmaktadır. Bu yaklaşımlar; Min-Max, Medyan ve Z-Score gibi yaklaşımlardır [3].

Hepatosellüler karsinom (HCC) hastalığı karaciğer organında oldukça sık karşılaşılan bir kanser çeşididir. Erkeklerde en sık görülen 5. ve kadınlarda ise en sık görülen 7. tümördür. Ayrıca dünyada en sık görülen dördüncü kanser türüdür ve karaciğer kanserlerinin %80'nini oluşturmaktadır. Tip II diyabet, obezite ve alkol tüketimi bu kanser türünün ortaya çıkmasında önemli faktöre sahiptir. Böylesine önemli bir hastalığın son yıllarda diyabet ve alkol tüketiminin artmasına bağlı olarak görülme sıklığının artabileceği göz önünde bulundurulmalıdır.

HCC hasatlığına yönelik olarak gerek tıp alanında gerekse makine öğrenmesi ve veri modellemesi noktasında birçok çalışma bulunmaktadır. Bu çalışmalar özellikle eksik veri ve veri ölçeklemesi konularında olması bakımından incelenerek sunulmuştur. Hepatitis C virüsü tarafından uyarılan HCC haslığına yönelik olarak eksik verilerin Beklenti Maksimizasyonu, Markov Zinciri Monte Carlo (MCMC), Regresyon, ve Yordayıcı Ortalama Eşleşme (predictive mean matching) eksik veri tamamlama yöntemleri kullanılmıştır. Genelleştirilmiş Tahmin Denklemleri, Zaman Bağımlı Cox Regresyon, ve Birleşik Modelleme yöntemleri istatistiksel çıkarım için kullanılmıştır. Sonuç olarak çoklu veri tamamlama yöntemlerinin eksik veriler ile çıkarım yapmada daha etkili olduğunu göstermişlerdir [4]. HCC hastalığına yönelik olarak eksik verilerin tamamlanması için 5 farklı yöntem medyan, mod, ortalama, karar ağacı tabanlı regresyon ve lineer regresyon uygulamışlar ve sonuç olarak karar ağacı tabanlı regresyon yönteminin daha başarılı olduğunu tespit etmişlerdir [5]. Kayıp veriler ile gereğinden fazla öznitelik olması durumuna yönelik olarak yeni bir yöntem önerilen çalışmada 0.9879 sınıflama başarısı elde edilmiştir [6]. Zincir Denklemleri Algoritmasında çok değişkenli veri tamamlama yönteminin tanıtıldığı çalışmada önerilen yöntemin etkinliği kalp hastalıkları ve HCC veri seti üzerinden gösterilmiştir [7]. Hastalar arasındaki farklılıkları dikkate alarak makine öğrenmesine dayalı yeni bir örnekleme yönteminin önerildiği çalışmada ise eksik veri tamamlama ön işleme sürecinde kullanılmış ve yöntemin etkinliği gösterilmiştir [8].

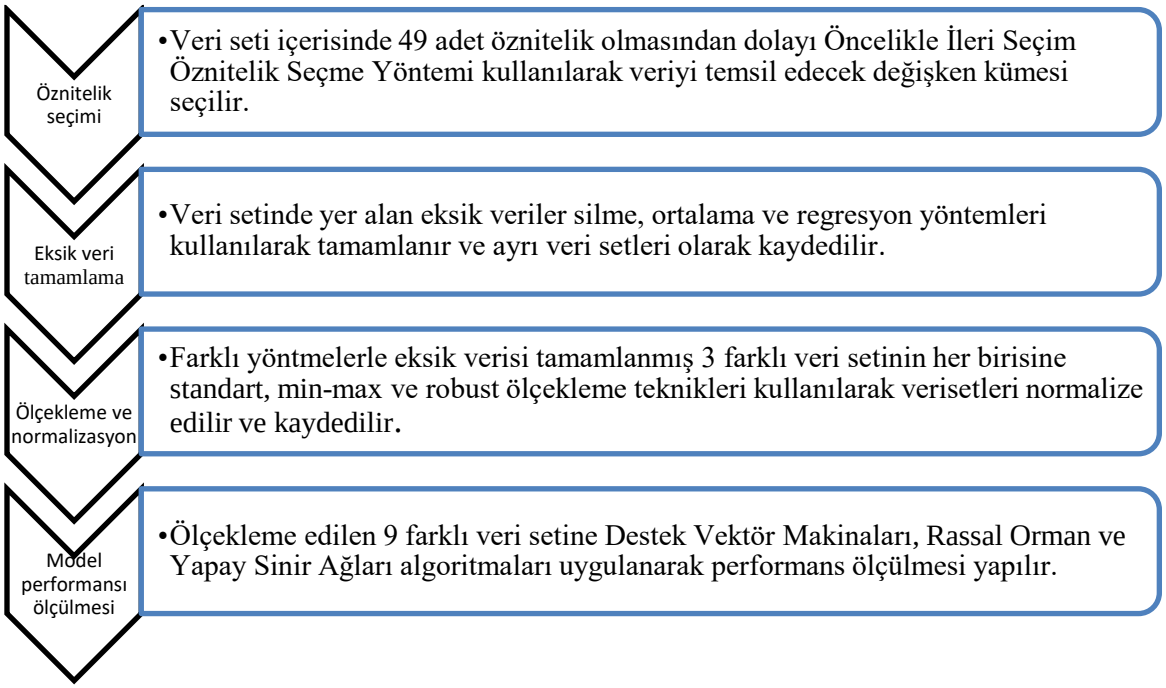
HCC hastalığı dışında literatürde veri ölçekleme ile eksik veri tamamlama üzerinde çalışmalar bulunmaktadır. Özellikle şeker hastalığında hastaneye geri yatış olasılığının tahmin edildiği çalışmada veri ölçekleme ve öznitelik mühendisliği ön işlem adımları uygulanarak model başarısında ciddi artış sağlanmıştır [9]. Yağış miktarının tahmin edildiği çalışmada ise eksik veri tamamlama ve normalizasyon yöntemlerinin ön işleme adımında kullanılması ile farklı sınıflandırıcılar üzerinde başarılı olduğu gösterilmiştir [10]. Dış ortam kirliliğinin ciddi ölümlere neden olduğu ve partikül maddelerin tespit edilmesi noktasında veri ön işleme adımlarının etkililiği araştırılmıştır. Bu bağlamda öznitelik seçimi, boyut azaltışı, veri ölçekleme ve eksil veri tamamlama adımları uygulanarak partikül tahmininde iyileştirmeler gerçekleştirilmiştir [11]. Gen ifade profilleri için veri ön işleme adımlarının başarısını inceleyen çalışmada ise log2 ve Z-Score normalizasyon yöntemleri farklı eksik veri tamamlama yöntemlerinin makine öğrenmesi yöntemleri üzerindeki etkileri incelenmiştir [12]. Parkinson hastalığının tahmininde eksik veri olması durumunda Copulas teorisini kullanarak eksik veri tamamlama yöntemi önerilmiş ve bu yöntem 7 farklı eksik veri tamamlama yöntemi ile karşılaştırılarak sonuçlar verilmiştir [13]. Eksik verilerin tamamlanması ile ilgili merkez noktalar ve en yakın komşulara dayalı yeni bir paradigmanın önerildiği çalışmada ise z-skor değerleri temel alınarak eksik veriler tamamlanmış sonuçlar verilmiştir [14]. Çalışma [15] ise missing value olarak bilinen kayıp veri olması durumunda bu verilerin tamamlanmasının PV jeneratörlerinin tahmini üzerindeki etkisini incelemiştir. Bu çalışmalar da farklı ölçekleme ve kayıp veri tamamlama noktasında yöntemlerin performansları karşılaştırılmıştır. Çalışma [16]'da farklı ölçekleme ve normalizasyon işlemlerinin makine öğrenmesi algoritmalarının sonuçlarını nasıl değiştirdiği gösterilmiştir. Hava durumu tahmini üzerinde gerçekleştirilen çalışmada normalize etmenin Yapay Sinir Ağları performansını ciddi anlamda değiştirebildiği gösterilmiştir.

Fakat literatürde ölçekleme işlemleri ve eksik veri tamamlama işlemlerinin birlikte değerlendirilerek makine öğrenmesi algoritmalarına etkisinin incelenmediği görülmektedir. Kanser hastalığı gibi önemli bir hastalık veri seti üzerinde hem veri ölçekleme hem de eksik veri tamamlama işlemlerinin algoritmalar üzerindeki etkisinin gösterilmesi noktasında literatüre katkı sağlamak amacı ile bu çalışma gerçekleştirilmiştir.

II. YÖNTEM

Bu çalışmada Portekiz'de University Hospital tarafından toplanan ve UCI veri tabanından ulaşılan HCC Survival Data Set kullanılmıştır. Veri tabanında 49 öznitelik ve 165 örnek yer almaktadır. Kayıp veri içeren bir veri setidir. Eksik veriler tüm veri setinin yaklaşık %10'unu oluşturmaktadır. Bağımlı değişken olan hastaların yaşayıp yaşamadığını ifade eden değişken ise 'class' değişkeni binary olarak kodlanmıştır.

Çalışma kapsamında izlenilen yöntem Şekil 1'de gösterilmiştir.



Şekil 1. *Çalışma kapsamında izlenilen yol.*

A. ÖZNİTELİK SEÇİMİ

Veri setinde 49 adet değişken olması nedeniyle öncelikle öznitelik seçim işlemi uygulanmıştır. Öznitelik seçimi için İleri Seçim (Frorward Selection) yöntemi kullanılmıştır. Bu yöntemin seçilme nedeni ise sınıflandırma işlemlerinde oldukça yüksek başarı göstermesidir [17]. Bağımlı değişken ile ilgili olan en önemli değişken ile model başlar. Yani başlangıçta sadece bir değişken vardır modelde. Diğer değişkenler modele sırasıyla eklenir. Eğer eklenen değişken modelin performansında artış sağlıyorsa modele dâhil edilir. Bütün değişkenler bu yolla kontrol edilerek model son halini alır. İleri seçim yöntemi ile seçilen öznitelikler ve bu özniteliklere ait eksik veriler Tablo 1'de gösterilmiştir.

Tablo 1*. İleri seçim yöntemi ile seçilen öznitelikler ve bu özniteliklere ait eksik veri değerleri*

Seçilen değişkenler	**Eksik veri sayısı**
46.Bil	44
2.Sym	18
42.Prot	11
41.Alk	3
1.Gen	0

B. EKSİK VERİ TAMALAMA

Seçilen değişkenler kullanılarak yeni bir veri seti oluşturulmuştur. Bu veri setine öncelikle eksik veri tamamlama işlemi uygulanmıştır. Eksik veri tamamlama yaklaşımlarından silme, ortalama ve regresyon yöntemleri kullanılarak 3 farklı veri seti elde edilmiştir. Silme işleminde veri setinden eksik veri bulunan gözlem satırı veri setinden çıkarılır. Ortalama yönteminde ise eksik verinin bulunduğu gözlem ilgili özniteliğin ortalaması ile tamamlanır. Regresyon yönteminde ise eksik veri tam olan veriler ile kurulan regresyon yöntemi ile tamamlanır [2].

C. ÖLÇEKLEME

Eksik verileri tamamlanmış veri setlerine sonraki aşamada ölçekleme işlemi uygulanmıştır. Elde edilen 3 farklı veri setine 3 farklı ölçekleme yönteminin uygulanması ile 9 farklı veri seti elde edilmiştir. Çalışmalarda sıklıkla kullanılan standart, min-max ve robust ölçekleme yöntemleri veri setlerine uygulanmıştır [13], [18].

Min-max ölçekleme: Eşitlik 1'de verilen bu yöntemde veriler 0-1 aralığına indirgenir. Verinin dağılımında herhangi bir değişiklik olmaz.

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

x' : ölçeklenen edilmiş öznitelik

x_{min} :Özniteliğin en küçük değeri

x_{max} : Özniteliğin en büyük değeri

Standart Ölçekleme: Eşitlik 2'de verilen bu yöntemde-öznitelik değerleri ortalamasından çıkarılarak standart sapmasına bölünme yolu ile ölçeklenir. Öznitelik değer aralığında bir sınır olmamasına rağmen genellikle değerlerin -3 ve +3 aralığında olduğu görülür.

$$x' = \frac{x_i - u_i}{\sigma} \quad (2)$$

x' : ölçeklenen edilmiş öznitelik

μ_i: Özniteliğin ortalaması

σ: Özniteliğin standart sapması

Robust Ölçekleme: Eşitlik 3'de verilen bu yöntem ise veride özellikle aykırı değer bulunması durumunda daha kullanışlıdır. Çünkü ortalama yerine medyan değerini ve değer aralığı yerine ise kartil aralığını kullanır. Kartil aralığı IQR (Interquartile Range) olarak bilinir ve 1. ve 3. Kartil aralığı arasındaki farka göre ölçekleme yapılır [2].

$$x' = \frac{x_i - medyan}{p_{75} - p_{25}} \quad (3)$$

x' : ölçeklenen edilmiş öznitelik

$medyan$:Özniteliğin medyanı

p_{75}: Özniteliğin 3. Kartil değeri

p_{25}: Özniteliğin 1. Kartil değeri

D. SINIFLANDIRMA ALGORİTMALARI

Üç farklı ölçekleme ve üç farklı eksik veri tamamlama yöntemi ile HCC veri setinden dokuz farklı veri seti elde edilmiştir. Bu dokuz farklı veri setine makine öğrenmesi çalışmalarında sıklıkla kullanılan üç farklı algoritma uygulanmıştır. Algoritmaların başarı ölçümlerinde hiperparametre optimizasyonu yapılmamış varsayılan değerler kullanılmıştır. Burada eksik veri tamamlama ve ölçekleme yaklaşımlarının performans sonuçlarına nasıl etki ettiği üzerine odaklanılmıştır. Çalışmada Yapay Sinir Ağları, Destek Vektör makinaları ve Rassal Orman algoritmalarının kullanılma sebebi özellikle veri normalizasyonu ve eksik veri ile ilgili çalışmalarda bu yöntemlerin tercih edilmesidir.

E. SINIFLANDIRMA BAŞARISI İÇİN KULLANILAN ÖLÇÜMLER

Sınıflandırma algoritmalarının performansının ölçülmesinde Eşitlik 4-7 arasında gösterilen ifadeler yardımı ile hesaplanan F1 skoru, kesinlik, duyarlılık ve doğruluk ölçüleri kullanılmaktadır.

$$\text{Doğruluk} = (TP+TN)/(TP+TN+FN+FP) \quad (4)$$

$$\text{Kesinlik} = TP/(TP+FP) \quad (5)$$

$$\text{Duyarlılık} = TP/(TP+FN) \quad (6)$$

$$F1 = (2*\text{Kesinlik}*\text{Duyarlık})/(\text{Kesinlik}+\text{Duyarlık}) \quad (7)$$

Burada TP doğru pozitif (True Positives), TN doğru negatif (True Negatives), FP yanlış pozitif (False Positives) ve FN yanlış negatif (False Negatives) değerlerini göstermektedir.

Veri setinde bulunan eksik verilerin ortalama yöntemi ile doldurulması sonrasında uygulanan ölçekleme yöntemlerinin performans üzerindeki etkisi Tablo 2-4'de verilmiştir.

III. BULGULAR

Veri setinde bulunan eksik verilerin ortalama yöntemi ile doldurulması sonrasında uygulanan ölçekleme yöntemlerinin performans üzerindeki etkisi Tablo 2-4'de verilmiştir.

Tablo 2. *Eksik veri tamamlama için ortalama, ölçekleme için standart ölçekleme yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,683	0,841	0,663	0,754
Rassal Orman	0,711	0,842	0,684	0,765
Yapay Sinir Ağı	0,722	0,786	0,678	0,749

Tablo 3. *Eksik veri tamamlama için ortalama, ölçekleme için min-max yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,675	0,862	0,661	0,751
Rassal Orman	0,711	0,842	0,685	**0,767**
Yapay Sinir Ağı	0,681	0,863	0,666	0,756

Tablo 4. *Eksik veri tamamlama için ortalama, ölçekleme için Robust Ölçekleme yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,629	0,912	0,612	0,741
Rassal Orman	0,705	0,843	0,679	0,763
Yapay Sinir Ağı	0,711	0,814	0,678	0,755

Tablo 2-4 incelendiğinde veri setinde bulunan eksik verilerin ortalama yöntemi ile doldurulması sonrasında yapılan ölçekleme yöntemlerinin makine öğrenmesi yöntemleri üzerinde ciddi farklılıklar yaratmadığı gözlemlenmiştir. F1 skoru ölçütüne göre en yüksek başarı eksik verilerin ortalama ile doldurulması durumunda min-max ölçekleme yönteminden elde edilmiştir.
Veri setinde bulunan eksik verilerin silme yöntemi ile doldurulması sonrasında uygulanan ölçekleme yöntemlerinin performans üzerindeki etkisi Tablo 5-7'de verilmiştir.

Tablo 5. *Eksik veri tamamlama için silme, ölçekleme için standart ölçekleme yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,742	0,64	0,649	0,669
Rassal Orman	0,683	0,704	0,631	0,68
Yapay Sinir Ağı	0,717	0,67	0,661	0,682

Tablo 6. *Eksik veri tamamlama için silme, ölçekleme için min-max yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,716	0,601	0,639	0,646
Rassal Orman	0,683	0,704	0,631	0,68
Yapay Sinir Ağı	0,689	0,584	0,592	0,611

Tablo 7. *Eksik veri tamamlama için silme, ölçekleme için Robust Ölçekleme yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,576	0,912	0,572	0,7
Rassal Orman	0,683	0,704	0,631	0,68
Yapay Sinir Ağı	0,66	0,616	0,6	0,62

Tablo 5-7 incelendiğinde veri setinde bulunan eksik verilerin silinmesi yöntemi ile doldurulması sonrasında yapılan ölçekleme yöntemlerinin makine öğrenmesi yöntemleri bazı farklılıklar yarattığı gözlemlenmiştir. Örneğin Destek Vektör Makinaları algoritmasında **robust ölçekleme** yöntemi kullanılırsa doğruluk % 57 olurken, aynı yöntem standart ölçekleme kullanıldığında % 65 başarı gösterebilmektedir. Aynı durum kesinlik değeri için de ortaya çıkmıştır. Fakat F1 skor değerlerinde birbirine yakın sonuçlar ortaya çıkmıştır.

Veri setinde bulunan eksik verilerin regresyon yöntemi ile doldurulması sonrasında uygulanan ölçekleme yöntemlerinin performans üzerindeki etkisi Tablo 8-10'da verilmiştir.

Tablo 8. *Eksik veri tamamlama için regresyon, ölçekleme için standart ölçekleme yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,668	0,844	0,642	0,742
Rassal Orman	0,677	0,836	0,648	0,743
Yapay Sinir Ağı	0,7	0,794	0,66	0,74

Tablo 9. *Eksik veri tamamlama için regresyon, ölçekleme için min-max yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,651	0,909	0,642	0,756
Rassal Orman	0,677	0,836	0,648	0,743
Yapay Sinir Ağı	0,661	0,831	0,636	0,733

Tablo 10*. Eksik veri tamamlama için regresyon, ölçekleme için Robust Ölçekleme yöntemi kullanılması*

Algoritma	Kesinlik	Duyarlılık	Doğruluk	F1
Destek Vektör	0,625	0,932	0,612	0,746
Rassal Orman	0,677	0,836	0,648	0,743
Yapay Sinir Ağı	0,686	0,80	0,654	0,738

Tablo 8-10 incelendiğinde veri setinde bulunan eksik verilerin regresyon yöntemi ile doldurulması sonrasında yapılan ölçekleme yöntemlerinin makine öğrenmesi yöntemleri üzerinde özellikle duyarlılık anlamında ciddi iyileştirmeler olduğu görülmektedir.

IV. SONUÇ

Bu çalışmanın amacı veri madenciliği çalışmalarında veri ön işleme sürecinde kullanılan eksik veri tamamlama ve veri ölçekleme yöntemlerinin makine öğrenmesi algoritmaları performansına etkisinin incelenmesidir. Bu inceleme için HCC kanser türü için oluşturulmuş bir veri seti kullanılmıştır.

Çalışmada bir veri seti üzerinde veri ön işleme aşamalarında farklı yaklaşımlar kullanmanın farklı sonuçlar vereceği gösterilmiştir. Literatürde farklı ölçekleme ya da normalizasyon tekniklerinin algoritmalar üzerinde farklı sonuçlar vereceğini gösteren çalışmalarla bu sonuçlar benzerlik göstermektedir [13], [18]. Benzer şekilde eksik veri tamamlama yöntemlerinin algoritmaların performansını etkileyebileceğini gösteren çalışma sonuçları ile bu çalışmada elde edilen bulgular benzerlik göstermektedir [1], [2], [20], [21].

Veri seti üzerinde en başarılı sınıflandırma işleminin eksik verilerin ortalama yöntemi ile tamamlanması ve sonrasında min-max ölçekleme işleminin yapılması ile edildiği görülmüştür. Bu sonuçlar insan beynindeki anomalilerin tespitinde eksik veri ile çalışılmasında ortalama yönteminin etkisini gösteren çalışma ile benzerlik göstermektedir [22]. Bu çalışmada ortalama ile eksik veri tamamlama yöntemi en iyi sonucu vermiştir fakat literatürde farklı yaklaşımların daha başarılı olduğu da gösterilmiştir [16], [23], [24].

Ayrıca çalışmada eksik verilerin silinmesi durumunda ölçekleme yaklaşımlarının bazı performans değerlerinde ciddi değişiklik gösterebileceği tespit edilmiştir. Bu durum özellikle robust ölçekleme

durumunda başarıyı azaltırken, standart ölçekleme durumunda ise başarıda artırıcı bir etki olarak ortaya çıkmıştır.

Makine öğrenme algoritmalarının performansı açısından bakıldığında eksik veri tamamlama ve veri ölçeklemede farklı yaklaşımlar kullanılsa da genel olarak Rassal Orman algoritmasının yapay sinir ağları ve Destek Vektör Makinaları algoritmasına göre daha başarılı sonuçlar verdiği tespit edilmiştir. Bunun yanında literatürde benzer sonuçlar veren çalışmalar bulunmaktadır [25]–[27]. Fakat bazı çalışmalar bu sonuçları desteklememektedir [28], [29]. Bu çalışmalarla sonuçların desteklenmeme nedeni bu araştırmada eksik veri ve ölçekleme yaklaşımları ile algoritmaların performansının incelenmesi fakat diğer çalışmaların veri seti üzerinde direk olarak çalışılması olabilir.

Veri ön işleme süreci özellikle veri bilimi ve makine öğrenmesi çalışmalarında çok önemli bir yere sahiptir. Bu aşamada yapılan her işlem çalışmanın seyrini önemli derecede değiştirebilir. Bu çalışmada da veri ön işleme aşamasında uygulanan farklı yaklaşımların sonuçları ne derecede değiştirebileceği gösterilmiştir. HCC kanser veri seti üzerinde üç farklı eksik veri tamamlama yöntemi ile üç farklı ölçekleme yönteminin üç farklı makine öğrenmesi algoritmasının performansının değiştirdiği gözlemlenmiştir.

Bu bağlamda veri bilimi ve makine öğrenmesi alanında yapılacak çalışmalarda veri ön işleme aşamasında bulunan farklı yaklaşımları yarı ayrı denemeleri ve en iyi performansı gösteren yaklaşımın seçilmesinin uygun olacağı sonucuna varılmıştır.

V. KAYNAKLAR

[1] E. Sezgin and Y. Çelik, "Veri madenciliğinde kayıp veriler için kullanılan yöntemlerin karşılaştırılması," *XV. Akademik Bilişim Konferansı Bildirileri*, Antalya, Türkiye, 2013, ss.194-198.

[2] T. Jayalakshmi and A. Santhakumaran, "Statistical Normalization and Back Propagationfor Classification", *International Journal of Computer Theory and Engineering* vol.3, no.1, pp.793-8201, 2011

[3] S. H. Caldwell, D. M. Crespo, H. S. Kang, and A. M. S. Al-Osaimi, "Obesity and hepatocellular carcinoma", *In Gastroenterology*, vol. 127, no.5, pp.97–103, 2004.

[4] J. Jose, G.K. Vishwakarma, A. Bhattacharjee, "Illustration of missing data handling technique generated from hepatitis C induced hepatocellular carcinoma cohort study", *Journal of King Saud University - Science*. vol.33, no.4, 2021.

[5] M. Yumus, M. Apaydin, A. Degirmenci, O. Karal, "Missing data imputation using machine learning based methods to improve HCC survival prediction", *28th Signal Processing and Communications Applications Conference (SIU)*, Gaziantep, Türkiye, 2020, ss.1-4.

[6] F.B. Demir, T. Tuncer, A.F. Kocamaz, F. Ertam, "A survival classification method for hepatocellular carcinoma patients with chaotic Darcy optimization method based feature selection", *Medical Hypotheses*, vol.139, 2020.

[7] S. Han, A.C. Andrei, K.W. Tsui, Multiple imputation for competing risks survival data via pseudo-observations, Communications for statistical applications and methods, vol.25 , pp. 385–396, 2018.

[8] M.S. Santos, P.H. Abreu, P.J. García-Laencina, A. Simão, A. Carvalho, A new cluster-based

oversampling method for improving survival prediction of hepatocellular carcinoma patients, Journal of Biomedical Informatics, vol.58 pp.49–59, 2015.

[9] E.H. Zaky, M.M. Soliman, A.K. Elkholy, N.I. Ghali, "Enhanced predictive modelling for 30-day readmission diabetes patients based on data normalization analysis", *International Journal of Intelligent Engineering and Systems.* vol.14, pp.204–216, 2021.

[10] K. Varada Rajkumar, D.K. Subrahmanyam, "A novel method for rainfall prediction and classification using neural networks", *International Journal of Advanced Computer Science and Applications*. vol.12, pp. 521–528, 2021.

[11] D.H. Djarum, Z. Ahmad, J. Zhang, "Comparing Different Pre-processing Techniques and Machine Learning Models to Predict PM10 and PM2.5 Concentration in Malaysia", *Lecture Notes in Mechanical Engineering*, Malaysia, 2021, pp.353–374.

[12] I. Duran, R. Leandro, J. Guevara-Coto, "Analysis of different pre-processing techniques to the development of machine learning predictors with gene expression profiles", *Proceedings - 4th Jornadas Costarricenses de Investigacion En Computacion e Informatica*, JoCICI, San Pedro, Costa Rica, 2019.

[13] R. Houari, A. Bounceur, T. Kechadi, A.K. Tari, R. Euler, "Missing data analysis using multiple imputation in relation to Parkinson's Disease", *BDAW '16*, 2016.

[14] G. Madhu, G. Nagachandrika, "A new paradigm for development of data imputation approach for missing value estimation", *International Journal of Electrical and Computer Engineering*. Vol.6, no.6, pp.3222–3228, 2016

[15] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting," *Appl. Sci.*, vol. 9, no. 1, pp. 204, 2019.

[16] S. Yavuz and M. Deveci, "İstatiksel normalizasyon tekniklerinin yapay sinir ağın performansına etkisi" *Erciyes Üniversitesi İktisadi ve İdari Bilim. Fakültesi Derg.*, c. 0, s. 40, ss. 167-187, 2012.

[17] P. Cihan, O. Kalipsız, and E. Gökçe, "Hayvan hastalığı teşhisinde normalizasyon tekniklerinin yapay sinir ağı ve özellik seçim performansına etkisi," *Turkish Stud.*, c. 12, s. 11, ss. 59–70, 2017.

[18] Scikitlearn. (2021, May 27) "sklearn.preprocessing.RobustScaler — scikit-learn 0.24.2 documentation,[Online].Available:"https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html#sklearn.preprocessing.RobustScaler.

[19] R. Bakış and S. Göncü, "Akarsu Debi Ölçümlerinde Eksik Verilerin Tamamlanması: Zap Suyu Havzası Örneği," *Anadolu Univ. J. Sci. Technol. Appl. Sci. Eng.*, c. 16, s. 1, ss. 63, 2015

[20] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognit.*, vol. 41, no. 12, pp. 3692–3705, 2008.

[21] M. K. Markey, G. D. Tourassi, M. Margolis, and D. M. DeLong, "Impact of missing data in evaluating artificial neural networks trained on complete data," *Comput. Biol. Med.*, vol. 36, no. 5, pp. 516–525, 2006.

[22] D. Bertsimas, C. Pawlowski, and Y. D. Zhuo, "From Predictive Methods to Missing Data Imputation: An Optimization Approach," *J. Mach. Learn. Res.*, vol. 18, pp. 1–39, 2018.

[23] G. e. a. p. a. Batista and M. C. Monard, “An analysis of four missing data treatment methods for supervised learning,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 519–533, 2003.

[24] S. A. Naghibi, K. Ahmadi, and A. Daneshi, “Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping,” *Water Resour. Manag.*, vol. 31, no. 9, pp. 2761–2775, 2017.

[25] P. Thanh Noi and M. Kappas, “Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery,” *Sensors (Basel).*, vol. 18, no. 1, p. 18, 2017..

[26] T. Han, D. Jiang, Q. Zhao, L. Wang, and K. Yin, “Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery,” *Trans. Inst. Meas. Control*, vol. 40, no. 8, pp. 2681–2693, 2018.

[27] M. a. m. Hasan, M. Nasser, B. Pal, and S. Ahmad, “Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS),” *J. Intell. Learn. Syst. Appl.*, vol. 06, no. 01, pp. 45–52, 2014.

[28] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, “Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection,” *IEEE Access*, vol. 6, pp. 33789–33795, 2018.