Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 30, S. 35-43, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 30, pp. 35-43, December 2021

Research Article

Normalizasyon Yöntemlerinin Biyomedikal Verilerde Sınıflandırma Performansına Etkisi

Hakan Yüce[1*], Ali Osman Özkan[2]

[1*] Necmettin Erbakan Üniversitesi, Fen Bilimleri Enstitüsü, Konya, Türkiye, (ORCID: 0000-0003-3275-1944), hakanyuce91@gmail.com
[2] Necmettin Erbakan Üniversitesi, Mühendislik Fakültesi, Elektrik ve Elektronik Mühendisliği, Konya, Türkiye (ORCID: 0000-0002-2226-9786), alozkan@erbakan.edu.tr

(2nd International Conference on Computer, Electrical and Electronic Sciences ICCEES 2021, September 1-3, 2021)

(**DOI:** 10.31590/ejosat.1011723)

ATIF/REFERENCE: Yüce, H., Özkan, A. O. (2021). Normalizasyon Yöntemlerinin Biyomedikal Verilerde Sınıflandırma Performansına Etkisi. *Avrupa Bilim ve Teknoloji Dergisi*, (30), 35-43.

Öz

Günümüzde tıpta hastalıklara ait veri miktarı giderek artmakta ve bu verilerden hastalığın sınıfı hakkında tahminler yapılmaktadır. Bu tahminlere olumlu sonuç sağlayabilecek teknikler üzerinde çalışmalar artmaktadır. Bu tahminleri yapacak olan sınıflandırma algoritmaları bu tekniklerle daha doğru sınıflandırma performansı gösterebilmektedir. Bu çalışmada karaciğer ve kalp hastalığı veri setleri minimum maksimum, ondalık ölçekleme, z-skor ve norm normalizasyon yöntemleriyle normalize edilmiştir. Daha sonra bu normalize edilmiş ve ham verilere, 4 farklı k-kat çapraz değerlendirmede (2,5,10,20) yapay sinir ağları, karar ağacı, destek vektör makinesi, k-NN ve Naive Bayes gibi çeşitli sınıflandırma algoritmalarıyla ORANGE programı kullanılarak sınıflandırma işlemine tabi tutulmuştur. Verilerin sınıflama doğrulukları değerlendirilmiş ve normalizasyon yöntemlerinin sınıflandırma performansını artırdığı gözlemlenmiştir.

Anahtar Kelimeler: Normalizasyon yöntemleri, Sınıflama algoritmaları, ORANGE programı, Sınıflama doğruluğu, k-kat çaprazlama

Effect of Normalization Techniques on Classification Success in Biomedical Data

Abstract

Nowadays, the amount of data about diseases in medicine is increasing and predictions about the class of the disease are made from these data. Studies on techniques that can provide positive results to these predictions are increasing and being used. Classification algorithms that will make these predictions can show more accurate classification performance with these techniques. In this study, liver and heart disease data sets were normalized using minimum maximum, decimal scaling, z-score and norm normalization methods. Then, these normalized and raw data are classified using the ORANGE program with various classification algorithms such as artificial neural networks, decision tree, support vector machine, k-NN and Naive Bayes in 4 different k-fold cross validation (2,5,10,20) has been processed. The classification accuracies of the data were evaluated and were observed that normalization methods increased the classification performance.

Keywords: Normalization method, Classification algorithms, ORANGE program, Classification accuracy, k-fold cross validation

[*] Sorumlu Yazar: hakanyuce91@gmail.com

1. Giriş

Normalizasyon sıklıkla kullanılan veri ölçekleme ve haritalama tekniğidir. Verilerin özelliklerini normalleştirme, önceden karar verilen aralıklardaki tüm özelliklerin değerlerini sınırlamak için yararlı bir adımdır. Veri seti içinde birçok özellik var olabilir ve bu özelliklerin boyutları farklı olabilir. Boyutu büyük olan değer sınıflandırma işleminde daha büyük ağırlığa neden olabilir ve sınıflama doğruluğunu o eksende kaydırabilir. Fakat doğruluk ağırlığı küçük olan bir değer tarafında da olabilir. Bu nedenle veri içindeki her özellik yaklaşık olarak eşit aralığa ve aynı etkiye sahip olmalıdır. Ayrıca veri setinin özellik çıkarımından sonra oluşturulan yeni veri setinin boyutu fazla olabilir. Veri setinde ilgisiz, fazla ve gürültülü özellikler olabilir. Bu özellikler sınıflama performansını azaltabilir, sınıflandırıcının hesaplama maliyetini artırabilir ve sonuçların kalitesini düşürebilir. Ayrıca ölçekleme işlemi girdi veriye uygulanabildiği gibi çıktı veriye de uygulanabilir. Çünkü bir işlemin çıktı verisi başka verinin girdi verisi olabilir (Shalabi ve ark., 2006; Singh ve ark., 2015; Atomi, 2012; Polat, 2008; Yavuz ve Deveci, 2012; Muthuselvan ve ark., 2018; Akdemir, 2009).

Bu çalışmada, minimum maksimum, ondalık ölçekleme, z-skor ve norm normalizasyon yöntemleri kullanılmıştır ve normalizasyon işlemi sonrasında sınıflandırma işleminde destek vektör makinesi (DVM), yapay sinir ağları (YSA), karar ağaçları (KA), k-en yakın komşu algoritması (k-NN) ve Naive Bayes (NB) gibi sınıflandırma yöntemleri uygulanmıştır (Yüce, 2021).

Yapılan çalışmada karaciğer hastalığı ve kalp hastalığı için 50 sağlıklı, 50 hasta bireyin verileri kullanılmış olup veri setleri UCI internet sitesinden alınmıştır. Karaciğer vücudumuzdaki üçgen şeklindeki en büyük organdır. Vücuttaki glikoz, yağ, vitamin, hormon v.s birçok kimyasalın dengelenmesinde görevlidir. Kalp ise iki fonksiyona sahip kaslı bir pompadır. Birincisi; vücudun dokularından kanı toplayarak akciğerlere iletmek ikincisi ise akciğerlerden alıp vücudun bütün dokularına iletmektir. Karaciğer ve kalp hastalıklarının erken teşhisinde hastanın hayatta kalma olasılığı artmaktadır. Verilerin sınıflandırma işlemi ORANGE programı kullanılarak yapılmış ve (2,5,10,20) k-kat çaprazlamada sınıflama doğruluğu kriteri baz alınarak değerlendirilmiştir (Yüce, 2021).

2. Materyal ve Metot

Yapılan bu çalışmada; karaciğer ve kalp hastalığı verilerine DVM, YSA, KA, k-NN ve NB sınıflandırma yöntemlerinde 4 farklı (2,5,10,20) k-kat çaprazlamada minimum maksimum, ondalık ölçekleme, z-skor ve norm normalizasyon yöntemleri uygulanmıştır**.** Normalizasyon yöntemi sonrasında ORANGE programı kullanılarak sınıflandırma işlemi gerçekleştirilmiş ve sonuçlar ham veri setlerinin sınıflandırma performansı ile karşılaştırılmıştır (Yüce,2021).

2.1. Normalizasyon Yöntemleri

Normalizasyon, axb veri boyutuna sahip bir veri setini bir uzaydan alıp başka bir uzaya taşıma işlemidir. Bu taşımada yeni maksimum ve minimum noktaları oluşur ancak veri setinin axb olan boyutunda herhangi bir değişiklik meydana gelmez. Burada ham verinin aksine normalize edilmiş veri sayesinde sınıflandırıcının kararlılığı artabilecektir. Fakat şunu bilmeliyiz ki her veri seti için normalizasyon gerekmez. Özellikler farklı aralıklara sahip olduğu zaman gerekir (Akdemir,2009).

Şimdi sırayla çalışmada kullanılan bu normalizasyon yöntemlerini inceleyelim.

2.1.1. Minimum Maksimum Normalizasyon Yöntemi

Bu normalizasyon yöntemi genellikle mühendislik problemlerinde kullanılan en yaygın yöntem olup veri setini lineer bir şekilde yeni bir aralığa sıkıştırır. Sıkıştırma işlemi sonrasında veriler arasındaki ilişki korunur (Akdemir, 2009; Adeyemo ve Wimmer, 2018).

Lineer dönüşümü aşağıdaki gibi ifade edebiliriz.

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

Bu eşitlikte;

x' = Normalize edilmiş değer

x_i = Normalize edilecek değer

x_{min} = Veri setindeki en küçük değer

x_{max} = Veri setindeki en büyük değer

Normalizasyon işlemi sonrasında şayet eksi işaretli bir özellik olsa bile artık pozitif işaretli olacaktır. Ayrıca unutulmamalıdır ki minimum maksimum yöntemi uç noktalara iyi odaklanamaz
(Akdemir,2009; www.codecademy.com/articles/normalization).

2.1.2. Ondalık Ölçekleme Normalizasyon Yöntemi

Ondalık ölçekleme yöntemi minimum-maksimum yöntemi kadar yaygın kullanılmamaktadır. Bu normalizasyon yönteminde veri seti değerlerini 1'den küçük yapmak için mevcut değerleri 10 ve 10'un katı değerlere bölmek gerekir. Bu 10'un kuvveti olan değer mevcut değeri 1'den küçük yapmak için kullanılır (Akdemir, 2009).

Bu normalizasyon yöntemini aşağıdaki gibi ifade edebiliriz.

$$A' = \frac{A_i}{10^j}$$

Bu eşitlikte;

A' = Normalize edilmiş veri

A_i = Normalize edilecek değer

J = A' değerini 1'den küçük yapan değer.

2.1.3. Z-Skor Normalizasyon Yöntemi

Bu yöntem istatiksel normalizasyon yöntemi olarak da bilinir. Veri seti içinde bazı uç değerler olabilir ve bu değerler sonuçlara daha fazla etki yapabilir. Veri seti içindeki mevcut uç verilerin diğer veriler gibi modele tahmin için eş katkı sağlaması istenir. Z-skor yöntemiyle mevcut verilerin standart sapması ve ortalaması hesaplanıp z-skor formülü kullanılarak eş katkı yapması sağlanır
(www.codecademy.com/articles/normalization).

Bu normalizasyon yöntemi için aşağıdaki eşitlik kullanılır.

$$x' = \frac{x_i - \mu_i}{\sigma_i}$$

Bu eşitlikte;

x_i = Normalize edilecek değer

μ_i = Veri setinin ortalama değeri

σ = Verideki standart sapma

Standart sapmanın hesaplaması için aşağıdaki eşitlik kullanılır.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i-1}^{N}(x_i - x')^2}$$

Bu eşitlikte;

N = Dizinin eleman sayısı

x_i = Dizinin i. elemanı

x' = Dizinin elemanlarının aritmetik ortalaması

2.1.4. Norm Normalizasyon Yöntemi

Herhangi bir vektörün normu ya da uzunluğu Öklid mesafesine eşittir. Norm ya da vektör normalizasyonun da aşağıdaki eşitlikte görüldüğü gibi ilk olarak özelliğin tüm değişkenlerinin kareleri alınıp toplanır ve sonra sonucun kare kökü alınarak norm değeri hesaplanır. Sonrasında her bir değişken norm değerine bölünerek yeni normalize değeri bulunur (Gautam ve ark., 2015). Norm hesabını aşağıdaki gibi ifade edebiliriz (Eesa ve Arabo, 2017).

$$|x| = \sqrt{x_1^2 + x_2^2 + x_3^2 + ... + x_i^2}$$

Burada;

|x|= Normalize edilecek verilerin normu

Norm normalizasyon için aşağıdaki eşitliği kullanabiliriz.

$$x' = \frac{x_i}{|x|}$$

Bu eşitlikte;

x' = Normalize edilmiş veri

x_i = Normalize edilecek değer

2.2. Sınıflama Yöntemleri

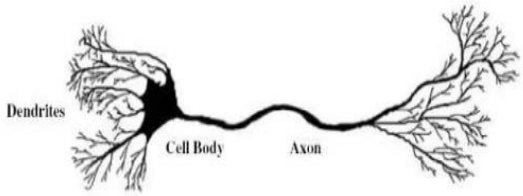
Literatürde birçok sınıflandırma yöntemi kullanılmasına rağmen bu çalışmada YSA, DVM, NB, k-NN ve KA gibi sınıflandırma yöntemleri kullanılarak 4 farklı (2,5,10,20) k-kat çaprazlamada sınıflama işlemine tabi tutulmuştur. Şimdi sırasıyla bu sınıflama algoritmalarını inceleyelim.

2.2.1. Destek Vektör Metodu (DVM)

DVM istatiksel öğrenme teorisini temel alan parametrik olmayan öğrenim yöntemlerinden biridir. İlk olarak Vapnik tarafından 1992 yılında kullanılmıştır. Bu yöntemin çalışması, iki sınıfı birbirinden ayıran en uygun karar verme fonksiyonunun tanımlanması olarak ifade edilir. Yani, en uygun hiper düzlemin tanımlanması şeklindedir (Kavzoğlu ve Çölkesen, 2010; Mohamed, 2017).

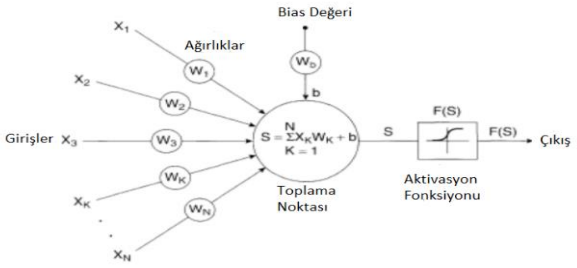
2.2.2. Yapay Sinir Ağları (YSA)

YSA sorunlara çözüm bulmak amacıyla insan beyninin gerçekleştirebileceği birçok özellik kullanılarak geliştirilmiş bilgisayar yazılımlarıdır. Bildiğimiz gibi insan beyni bilgiyi alır, yorumlar ve bu değerlendirmeyi sonuçlanır. Buradaki amaç beynimizi matematiksel olarak modellenmesidir. Bu modelleme düşüncesi makineler insan gibi düşünebilir mi fikrini ortaya atan İngiliz matematikçi ve bilgisayar bilimci olan Alan Mathison Turing tarafından bulunmuştur (Yazıcı, 2007; https://kod5.org/yapay-sinir-aglari-ysa nedir/). Şekil 1.'de insan sinir ağı genel görünümü gösterilmektedir. YSA sınıflandırma algoritması da insan sinir ağı modeline benzetilmiş ve burada *dentrites* toplama fonksiyonunu, *cell body* aktivasyon fonksiyonunu ve *axon* ise çıkış elemanını ifade etmektedir (https://kod5.org/yapay-sinir-aglari-ysa nedir/ysa-nedir/).



Şekil 1. İnsan sinir ağı genel görünümü (https://kod5.org/ internet sayfasından alınmıştır)

Şekil 2.'de, Şekil 1.'de görülen insan sinir ağı modelinin YSA algoritmasında modellenmesi gösterilmiştir. YSA çarpma, toplama ve aktivasyon olmak üzere üç temel işleme sahiptir. Her giriş bir ağırlıkla çarpılır ve sonra bir *bias* değeri ile toplanır. Elde edilen sonuç çeşitli aktivasyon fonksiyonlarından geçirilerek çıkış olarak iletilir (Mohamed, 2017).



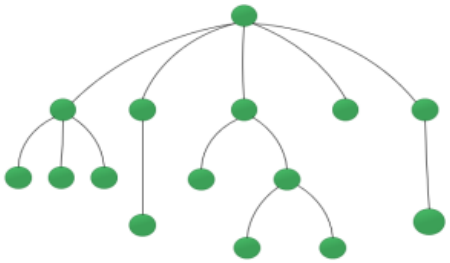
Şekil 2. YSA modeli (www.iitmandi.ac.in internet sayfasından alınmıştır)

2.2.3. Karar Ağacı (KA)

Bir KA, birçok özelliğe sahip olan bir veri kümesini, bir dizi karar kuralları uygulayarak daha düşük birimlere ayırmak için kullanılır. Bir KA kök, iç ve yaprak düğümler ile temsil edilmektedir. Her iç düğümün öznitelik üzerinde bir test koşuluna sahiptir. Buna bağlı olarak her bir dalın test koşulunun sonucunu temsil eden bir yaprağı vardır ve her yaprak düğümünün bir sınıf etiketi ile atandığı ağaç yapısı gibi bir akış şemasıdır. İlk düğüm kök düğümü olarak ifade edilmektedir (Jadhav ve Channe, 2013).

Şekil 3.'te KA sınıflandırma algoritmasını tanımlamak için bir örnek yapı gösterilmiştir. KA sınıflandırma algoritması tek bir düğümle başlar ve bir dizi sorular sorarak belirli noktalara dallanarak ulaşır. Şekil 3.'de görüldüğü gibi her daire bir karar noktasını ifade etmekte olup karar sonrası yeni dallanmalar oluşmaktadır. Oluşan dallanmalardan yeni karar noktaları oluşabilir veya sonlanabilir

(https://medium.com/@ekrem.hatipoglu/machine-learning-prediction-algorithms-decision-tree-random-forest-part-5-2970905c021e.).



Şekil 3. Karar ağacı yapı modeli

KA kolay yorumlanabilir ve anlaşılabilir özelliklerine sahip olmakla beraber düşük maliyetli ve güvenilir bir yöntemdir. Fakat KA veriyi iyi şekilde açıklamayan çok karmaşık ağaçlar ortaya çıkarabilir. Hatta bazen ezbere öğrenme bile yapabilir (Çalış ve ark., 2014; https://medium.com/@k.ulgen90/makine%C3%B6%C4%9FFrenimib%C3%B6I%C3%BCm-5-karar-a%C4%F9a%C3%A7lar%C4%B1-c90bd7593010).

2.2.4. k-En Yakın Komşu Algoritması

k-NN sınıflandırma yöntemi denetimli öğrenme yöntemlerinden birisidir. Bu yöntemde sınıflandırılması yapılacak verilerin, normal verilere göre davranışları incelenerek en yakın olduğu düşünülen k adet veri bulunur. Bu k adet verinin ortalaması alınır ve bu eşiğe göre sınıflandırma işlemi gerçekleştirilir. k-NN, verilerin dağılımı hakkında çok az ya da önceden hiç bilgi olmadığı zaman temel ve en basit sınıflandırma tekniğinden biridir. Bundan dolayı yöntemde önemli olan verilerin özelliklerinin net olmasıdır (Çalışkan ve Soğukpınar, 2008; Bolandraftar ve Imandoust, 2013).

k-NN sınıflandırma yönteminde uzaklık hesaplanırken genelde 3 yöntemden faydalanılır. Bunlar; Öklid, Minkowski ve Manhattan uzaklığıdır. Öklid uzaklığı aşağıdaki gibi hesaplanır.

$$d(a,b) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Manhattan uzaklığı aşağıdaki gibi hesaplanır.

$$d(a,b) = \sum_{i=1}^{k} | x_i - y_i |$$

Minkowski uzaklığı aşağıdaki gibi hesaplanır.

$$d(a,b) = (\sum_{i=1}^{k} (| x_i - y_i |)^q)^{1/q}$$

k-NN sınıflandırma yönteminde k bilinmeyen noktanın en yakın komşularını temsil etmekte olup; k=1 olduğu zaman en basit halidir ve her bir örnek onu çevreleyen örneklere benzer olarak sınıflandırma işlemine tutulacaktır. Bir örneğin sınıfı bilinmiyorsa sınıflandırma işlemi ona en yakın örneğin sınıfına göre olacak demektir (Bolandraftar ve Imandoust, 2013).

2.2.5. Naive Bayes (NB)

NB sınıflandırma yöntemi ismini matematikçi Thomas Bayes'den almıştır. Bu algoritma olasılık işlemleri kullanarak bir dizi hesaplama yapar ve sisteme girilen verilerin sınıfını belirlemeye çalışır. Her verinin sınıflandırmaya katkı sağladığı ve karşılıklı ilişkili olduğu varsayılır. Sınıflandırma işleminin temeli Bayes teoremine dayanır ve genellikle veri boyutu büyük olduğu zaman tercih edilir (Jadhav ve Channe, 2013; https://en.wikipedia.org/wiki/Orange_(software)). Bayes teoremi için kullanılan eşitlik aşağıdaki gibidir.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Bu eşitlikte;
P(A): A olayının gerçekleşme durumu
P(B): B olayının gerçekleşme durumu
P(A\B): B olayının olması durumunda A olayının gerçekleşme durumu

Sınıflandırma işleminde sistem belirli miktarda sınıfı olan öğretilmiş veri ile beslenir. Öğretilmiş veriler üzerinde olasılık hesabı yapılarak sisteme sunulan belirli bir sınıfa ait olan veriler üzerinde sınıflandırma yapılmaya çalışılır. Bilinmelidir ki öğretilmiş veri sayısı arttıkça test verilerinin sınıflandırma doğruluğu daha da yüksek olacaktır (https://en.wikipedia.org/wiki/Orange_(software).

2.3. Sınıflama Doğruluğu

Bir sınıflandırma işleminde yapılan gerçek ve tahmin edilmiş olan sınıflamalar hakkındaki bilgiye karışıklık matrisi vasıtasıyla ulaşabiliriz. Bu matris mxm boyutunda olup satırlar; doğru karar sınıflarına, sütunlar ise; sınıflandırıcı tarafından alınan kararlara karşılık gelir (Akdemir, 2009). Tablo 1'de karışıklık matrisi gösterilmiştir.

Tablo 1. Karışıklık matrisi

Gerçek	Tahmin Edilen	
	Negatif	Pozitif
Negatif	TN	FN
Pozitif	FP	TP

Sınıflama doğruluğu için aşağıdaki eşitlik kullanılır.

$$\text{Sınıflama Doğruluğu (\%)} = \frac{TP+TN}{TP + TN + FN + FP} * 100$$

2.4. Orange Programı

ORANGE programı; açık kaynak kodlu bir program olup veri görselleştirme, makine öğrenimi ve veri madenciliği için sıklıkla kullanılmaktadır. Bu programın en önemli özelliği veri analizi için iş akışlarını çeşitli görsel araç kutucukları ile oluşturarak işleyişi kolaylaştırmaktır. Ayrıca ORANGE programı Excel, virgül ve sekme ile ayrılmış dosyaları ve Google e-tablolar gibi çevrim içi dosyaları da okuyabilme yeteneğine sahiptir (https://orangedatamining.com/widget-catalog/; https://orangedatamining.com/download/#windows; https://e-abm.com/how-to-establish-quality-and-correctness-of-classification-models-part-3-confusion-matrix/).

3. Araştırma Sonuçları ve Tartışma

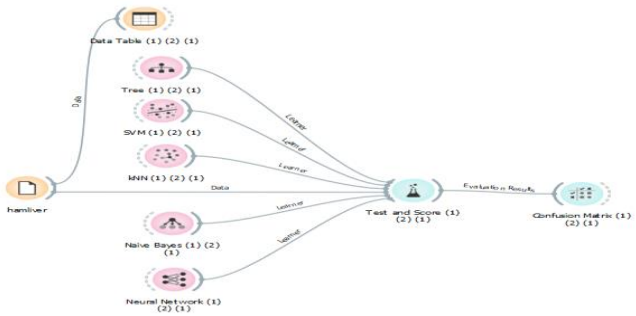
3.1. Karaciğer Hastalığı Veri Seti Özellikleri

Karaciğer hastalığı verilerini UCI internet sitesinden elde edilmiştir. Elde edilen veri setinde 10 özellik bulunmasına rağmen gerek eksik bilgi gerekse diğer nedenlerden dolayı

sadece 7 özellik sınıflandırma işleminde kullanılmıştır. Veriye ait 7 özelliğin açılımı aşağıdaki gibidir:
1.Özellik: Yaş (Yıl)
2.Özellik: Toplam bilirubin miktarı (mg/dL)
3.Özellik: Suda çözünebilen bilirubin miktarı (mg/dL)
4.Özellik: Alkalen fosfataz (IU/L)
5.Özellik: Alanin Aminotransferaz (IU/L)
6.Özellik: Aspartat Aminotransferaz (IU/L)
7.Özellik: Albümin / globülin oranı (g/L)

3.2. Karaciğer Hastalığı Veri Seti Sınıflandırma İşlemi

Karaciğer hastalığı veri seti sınıflandırma işlemine tabi tutulmadan önce veri setleri minimum maksimum, ondalık ölçekleme, z-skor ve norm normalizasyon işlemlerine tabi tutulmuştur. Ayrıca veri seti karşılaştırma yapmak için ham verilerde sınıflandırma işlemine tabi tutulmuştur. Şekil 4.'de karaciğer hastalığı ham veri setine ORANGE programıyla sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 4. Karaciğer hastalığı ham veri setinin ORANGE programıyla sınıflandırma işlemi

Yukarıdaki şekilde görülen sınıflandırma işlemi çalışmada kullanılan tüm normalizasyon yöntemleri uygulanarak tekrar edilmiştir. Sınıflandırma işleminde 4 farklı k-kat (2,5,10,20) çaprazlamada ayrı ayrı uygulanarak yapılmıştır. Sonuçlar aşağıda verilen Tablo 2.'de gösterilmiştir (Yüce,2021).

Tablo 2. Karaciğer hastalığı ham veri setinin 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	54	64	**68**	66	63
KA	59	**63**	62	58	60.5
DVM	**67**	65	**67**	66	66.25
YSA	**75**	71	74	70	72.5
NB	74	73	74	**75**	**74**
Ort.	65.8	67.2	**69**	67	

Tablo 3.'te karaciğer hastalığı veri setine minimum-maksimum normalizasyon yöntemi; Tablo 4'de karaciğer hastalığı veri setine ondalık ölçekleme normalizasyon yöntemi; Tablo 5'de karaciğer hastalığı veri setine z-skor ve Tablo 6'da karaciğer hastalığı veri setine norm normalizasyon yöntemi normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri gösterilmiştir (Yüce,2021).

Tablo 3. Karaciğer hastalığı veri setine minimum-maksimum normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	67	**68**	65	**68**	67
KA	59	62	**63**	61	61.25
DVM	**67**	65	**67**	66	66.25
YSA	**75**	71	74	71	72.75
NB	74	74	74	**76**	**74.5**
Ort.	68.4	68	**68.6**	68.4	

Tablo 4. Karaciğer hastalığı veri setinin ondalık ölçekleme normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	62	63	63	**68**	64
KA	58	**60**	58	51	56.75
DVM	**67**	65	**67**	65	66
YSA	73	71	**75**	69	72
NB	75	**76**	74	**76**	**75.25**
Ort.	67	67	**67.4**	65.8	

Tablo 5. Karaciğer hastalığı veri setinin z-skor normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	65	63	65	**66**	64.75
KA	59	**63**	62	58	60.5
DVM	**67**	65	**67**	66	66.25
YSA	**75**	71	73	70	72.25
NB	74	73	74	**75**	**74**
Ort.	68	67	**68.2**	**67**	

Tablo 6. Karaciğer hastalığı veri setinin norm normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	67	66	**68**	66	66.75
KA	59	62	**65**	64	62.5
DVM	**67**	65	**67**	65	66
YSA	73	70	**75**	69	71.75
NB	**74**	**74**	72	72	**73**
Ort.	68	67.4	**69.4**	67.2	

Normalizasyon yöntemlerinin sınıflandırma işlemindeki etkisine bakılırken ayrı ayrı değerlendirilen 2,5,10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerlerinin ortalaması alınmış ve Tablo 7.'de gösterilmiştir.

Tablo 7. Karaciğer hastalığı veri setini sınıflandırma işleminde normalizasyon yöntemlerinin etkisi

Sınıflandırma Yöntemleri	Sınıflama Doğruluğu (%)				
	Ham Veri	Minimum Maksimum	Ondalık Ölçekleme	Z-skor	Norm Yöntemi
k-NN	63	**67**	64	64.75	66.75
KA	60.5	61.25	56.75	60.5	**62.5**
DVM	**66.25**	**66.25**	66	**66.25**	66
YSA	72.5	**72.75**	72	72.25	71.75
NB	74	74.5	**75.25**	74	73
Ort.	67.25	**68.35**	66.8	67.55	68

Tablo 7.'de görüldüğü gibi; k-NN sınıflandırma yönteminde normalizasyon yöntemlerinin olumlu etkisi olmuştur. En iyi performans artışı; minimum-maksimum normalizasyon yönteminde % 67 sınıflama doğruluğu elde edilmiştir. KA sınıflandırma yönteminde minimum-maksimum ve norm normalizasyon yönteminde olumlu bir etkisi olmuş ve en iyi performans artışı % 62.5 ile norm normalizasyon yönteminde ulaşılmıştır. Z-skor normalizasyon yönteminin bir etkisi olmaz iken ondalık ölçekleme normalizasyon yönteminin olumsuz bir etkisi olmuştur. DVM sınıflandırma yönteminde normalizasyon yöntemlerinin performansı artırmadığı hatta ondalık ölçekleme ve norm normalizasyon yönteminde % 66 ile olumsuz etkilediği görülmüştür. YSA sınıflandırma yönteminde sadece minimum-maksimum normalizasyon yönteminde sınıflama doğruluğunun % 72.75 ile daha iyi olduğu görülmüştür. Diğer normalizasyon yöntemlerinin olumlu bir etkisi gözlemlenmemiştir. NB sınıflandırma yönteminde minimum-maksimum ve ondalık ölçekleme normalizasyon yönteminde sınıflama doğruluğunu arttırdığı görülmüştür. En iyi başarı % 75.25 ile ondalık ölçekleme normalizasyon yöntemine ait olmuştur. Z-skor yönteminde sınıflandırma algoritmasının başarısı değişmez iken norm yönteminin başarıyı olumsuz etkilediği görülmüştür (Yüce, 2021).

3.3. Kalp Hastalığı Veri Seti Özellikleri

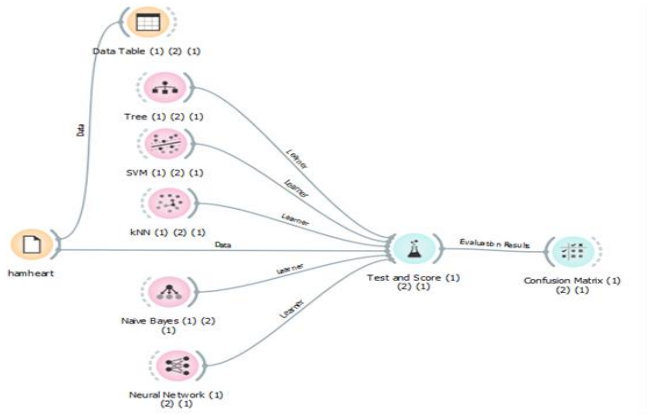
Kalp hastalığı veri setimiz UCI makine öğrenmesi bankası kalp veri seti tabanından alınmıştır ve veri setinin 13 özelliği aşağıdaki gibidir.

1.Özellik: Yaş (yıl)
2.Özellik: Cinsiyet (kadın/erkek)
3.Özellik: Göğüs ağrısı tipi (1 ile 4 arası)
4.Özellik: Dinlenme durumunda kan basıncı (tansiyon))
5.Özellik: Serum kolesterol (mg/dl)
6.Özellik: Tokluk şeker düzeyi >120 mg/dl
7.Özellik: Dinlenme halinde Elektrokardiyografı düzeyi (0,1,2)
8.Özellik: Maksimum kalp atış değeri(sürekli)
9.Özellik: Egzersiz durumunda göğüs ağrısı (0=hayır/1=evet)
10.Özellik: Dinlenme halinde ST değeri (sürekli)
11.Özellik: Pik egzersiz halinde ST segmentinin eğimi (1-2)
12.Özellik: Büyük damarların sayısı (0-3)
13.Özellik: Hasar oranı (3=normal,6=kalıcı,7=geri düzeltile bilinen hasar)

3.4. Kalp Hastalığı Veri Seti Sınıflandırma İşlemi

Kalp hastalığı veri seti sınıflandırma işlemine tabi tutulmadan önce karaciğer hastalığı sınıflandırma işleminde uygulandığı gibi veri setine minimum-maksimum, ondalık ölçekleme, z-skor ve norm normalizasyon işlemlerine tabi tutulmuştur. Ayrıca veri seti karşılaştırma yapılması için normalizasyon işlemine tabi tutulmadan da sınıflandırma işlemine tabi tutulmuştur.

Aşağıdaki verilen Şekil 5.'de kalp hastalığı ham veri setine ORANGE programıyla sınıflandırma işlemine tabi tutulması gösterilmiştir.



Şekil 5. Kalp hastalığı ham veri setinin ORANGE programıyla sınıflandırma işlemi

Kalp hastalığı için Şekil 5.'de görüldüğü gibi sınıflandırma işleminde çalışmada kullanılan tüm normalizasyon yöntemleri uygulanarak tekrar edilmiştir. Sınıflandırma işleminde 4 farklı k-kat (2,5,10,20) çaprazlamada ayrı ayrı uygulanarak yapılmıştır. Sonuçlar aşağıda verilen Tablo 8.'de gösterilmiştir (Yüce,2021).

Tablo 8. Karaciğer hastalığı ham veri setinin 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	61	**64**	**64**	**64**	63.25
KA	60	74	**79**	74	71.75
DVM	77	78	**81**	76	78
YSA	80	80	79	**81**	**80**
NB	77	79	**81**	80	79.25
Ort.	71	75	**76.8**	75	

Tablo 9'da kalp hastalığı veri setine minimum-maksimum normalizasyon yöntemi; Tablo 10'da kalp hastalığı veri setine ondalık ölçekleme normalizasyon yöntemi; Tablo 11'de kalp hastalığı veri setine z-skor ve Tablo 12'de kalp hastalığı veri setine norm normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri gösterilmiştir (Yüce, 2021).

Tablo 9. Kalp hastalığı veri setine minimum-maksimum normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	**79**	75	76	76	76.5
KA	60	74	**79**	74	71.75
DVM	77	78	**81**	76	78
YSA	80	80	79	**81**	**80**
NB	77	79	**81**	80	79.25
Ort.	74.6	77.2	**79.2**	77.4	

Tablo 10. Kalp hastalığı veri setine ondalık ölçekleme normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	74	78	78	**79**	77.25
KA	60	74	**79**	74	71.75
DVM	77	78	**81**	76	78
YSA	80	80	79	**81**	**80**
NB	77	79	**81**	80	79.25
Ort.	73.6	77.8	**79.6**	78	

Tablo 11. Kalp hastalığı veri setine z-skor normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	80	**81**	80	79	**80**
KA	60	74	**79**	74	71.75
DVM	77	78	**81**	76	78
YSA	**82**	78	80	79	79.75
NB	77	79	**81**	80	79.25
Ort.	75.2	78	**80.2**	77.6	

Tablo 12. Kalp hastalığı veri setine norm normalizasyon yöntemi uygulanarak 2, 5, 10 ve 20 k-kat çaprazlamada sınıflama doğruluğu değerleri

Sınıflandırma Yöntemi	Sınıflama doğruluğu (%)				
	k-kat çaprazlama				
	2	**5**	**10**	**20**	**Ort.**
k-NN	72	76	**79**	76	75.75
KA	60	74	**79**	75	72
DVM	77	78	**81**	76	78
YSA	**81**	80	79	**81**	**80.25**
NB	77	80	80	**81**	79.5
Ort.	73.4	77.6	**79.6**	77.8	

Normalizasyon yöntemlerinin kalp hastalığında sınıflandırma işlemindeki etkisi genel değerlendirilirken ayrı ayrı değerlendirilen k-kat çaprazlamada performanslarının ortalaması alınmış ve Tablo 13.'de gösterilmiştir (Yüce, 2021).

Tablo 13. Kalp hastalığı veri setini sınıflandırma işleminde normalizasyon yöntemlerinin etkisi

Sınıflandırma Yöntemleri	Sınıflama Doğruluğu (%)				
	Ham Veri	Minimum Maksimum	Ondalık Ölçekleme	Z-skor	Norm Yöntemi
k-NN	63.25	76.5	77.25	**80**	75.75
KA	71.75	71.75	71.75	71.75	**72**
DVM	78	78	78	78	78
YSA	80	80	80	79.75	**80.25**
NB	79.25	79.25	79.25	79.25	**79.5**
Ort.	74.45	77.1	77.25	**77,75**	77.1

Tablo 13'te görüldüğü gibi; k-NN sınıflandırma yönteminde normalizasyon yöntemlerinin tamamının sınıflama doğruluğuna olumlu etkisi olmuştur. En iyi sınıflama doğruluğu z-skor normalizasyon yönteminde % 80 olarak elde edilmiştir. KA sınıflandırma yönteminde sadece norm normalizasyon yönteminde % 72 sınıflama doğruluğu elde edilmiş, diğer normalizasyon yöntemlerinde olumlu bir etkisi görülmemiştir. DVM sınıflandırma yönteminde normalizasyon yöntemlerinin sınıflama doğruluğuna hiç bir etkisi olmamıştır. YSA sınıflandırma yönteminde sadece norm normalizasyon yönteminde sınıflama doğruluğunun az da olsa arttığı (% 80.25) görülmüştür. Diğer normalizasyon yöntemlerinde z-skor normalizasyon yöntemi hariç performansı değişmemiştir. Z-skor normalizasyon yöntemi % 79.75 ile olumsuz etkilediği görülmüştür. NB sınıflandırma yönteminde sadece norm normalizasyon yönteminde sınıflama doğruluğunu arttırdığı (% 79.5) görülmüş, diğer normalizasyon yöntemlerinde ise olumlu bir etkisi olmamıştır.

4. Sonuç

Yapılan bu çalışmada karaciğer ve kalp hastalığı veri setlerine minimum maksimum, ondalık ölçekleme, z-skor ve norm normalizasyon yöntemleri uygulanmıştır. Normalizasyon işleminden sonra verilere DVM, YSA, KNN, KA ve NB gibi sınıflandırma yöntemleri ile 4 farklı k-kat çaprazlamada (2,5,10,20) sınıflama doğruluğu kriterlerine bakılarak değerlendirilmiştir.

Sonuç olarak; minimum maksimum, ondalık ölçekleme, z-skor ve norm normalizasyon yöntemlerinin sınıflama doğruluğu performansını artırabileceği görülmüştür. Çalışmada kullanılmayan farklı normalizasyon yöntemlerinin de sınıflama performansına olumlu etki yapabileceği düşüncesi gelişmiştir. Sınıflama doğruluğu performansına etkisini görmek için, farklı k-kat çaprazlamada (2,5,10 ve 20 değerlerinde) sınıflama doğruluk performansını arttırdığı görülmüştür. Ayrıca hem karaciğer hem de kalp hastalığı veri setine 4, 25, 50 ve 100 gibi 100 örnek veriyi tam bölen farklı k-kat çaprazlamada sınıflama doğruluğu performansını artırabileceği düşüncesi oluşmuştur.

5. Teşekkür

Bu çalışma boyunca belirtikleri görüş ve önerilerle makalenin yönlenmesine yardımcı olan danışman hocam sayın Dr. Öğr. Üyesi Ali Osman ÖZKAN' a ve tüm hayatım boyunca beni bu zamana kadar yetiştiren aileme teşekkürlerimi sunuyorum.

Kaynakça

Adeyemo A. and Wimmer H. (2018). Effects of Normalization Techniques on Logistic Regression in Data Science. *2018 Proceedings of the Conference on Information Systems Applied Research Norfolk,* Virginia

Akdemir B. (2009). Tahmin uygulamalarında performans geliştirmek için kullanılan normalizasyon metotlarına yeni bir yaklaşım, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*, Doktora Tezi

Atomi V. H (2012). The effect of data preprocessing on the performance of artificial neural networks techniques for classification problem, Master Thesis, Faculty of Computer Science and Information Technology, University Tun Hussein Onn Malaysia

Bolandraftar M., Imandoust S. B. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background, S B Imandoust et al. *Int. Journal of Engineering Research and Applications*, Vol:3, Issue 5, 605-610.

Çalış A., Kayapınar S., Çetinyokuş T. (2014). Veri madenciliğinde karar ağacı algoritmaları ile bilgisayar ve internet güvenliği üzerine bir uygulama. *Endüstri Mühendisliği Dergisi*, Cilt:25, Sayı: 3-4,2-19.

Çalışkan S. B., Soğukpınar İ. (2008). k-means ve k en yakın komşu yöntemleri ile ağlarda nüfuz tespiti

Eesa A.S., Arabo W. K. (2017). Normalization methods for backpropagation: A comparative study, *Science Journal of University of Zakho*, Vol:5, No:4, 314-318.

Gautam R., Vanga S., Ariese F., Umapathy S. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2:8,2015. DOI: 10.1140/epjti/s40485-015-0018-6

https://e-abm.com/how-to-establish-quality-and-correctness-of-classification-models-part-3-confusion-matrix/

https://en.wikipedia.org/wiki/Orange_(software)

http://kod5.org/yapay-sinir-aglari-ysa-nedir/

https://kodedu.com/2014/05/naive-bayes siniflandirma-algoritmasi/

https://medium.com/@ekrem.hatipoglu/machine- learning-prediction-algorithms-decision-tree-random-forest-part-5-2970905c021e.

https://medium.com/@k.ulgen90/makine%C3%B6%C4%9FFrenimib%C3%B6I%C3%BCm-5-karar-a%C4%F9a%C3%A7lar%C4%B1-c90bd7593010

https://orangedatamining.com/widget-catalog/

https://orangedatamining.com/download/#windows

http://www.iitmandi.ac.in/ciare/files/7_Anand_ANN.pdf

Jadhav S. D., Channe H. P. (2013). Comparative study of k-nn, naive bayes and decision tree classification techniques, *International Journal of Science and Research (IJSR)*

Kavzoğlu T. ve Çölkesen İ. (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi, *Harita Dergisi*, Sayı 144

Mohamed A. E. (2017). Comparative study of four supervised machine learning techniques for classification. *International Journal of Applied Science and Technology*, Vol:7, No:2

Muthuselvan S., Rajapraksh S., Somasundaram K, KKarthik K. (2018). Classification of Liver Patient Dataset Using Machine Learning Algorithms, *International Journal of Engineering & Technology*, 7 (3.34), 323-326.

Polat K. (2008). Biyomedikal sinyallerde veri ön-işleme tekniklerinin medikal teşhiste sınıflama doğruluğuna etkisinin incelenmesi. *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*, Doktora tezi

Shalabi, L.A., Z. Shaaban and Kasasbeh, B. (2006). Data Mining: A Preprocessing Engine, *J. Comput. Sci.,* 2: 735-739

Singh B K, Thoke A. S. and Verma K. (2015). Investigations on Impact of Feature Normalization Techniques on Classifier's Performance in Breast Tumor Classification, *International Journal of Computer Applications* (ISSN: 0975 – 8887) Volume 116 – No. 19, 11-15.

www.codecademy.com/articles/normalization

Yavuz S., Deveci M. (2012). İstatiksel normalizasyon tekniklerinin yapay sinir ağın performansına etkisi, *Erciyes*

Üniversitesi İktisadi ve İdari Bilimler Dergisi, Sayı 40, 167-187.

Yazıcı A. C., Öğüş E., Ankaralı S., Canan S., Ankaralı H., Akkuş Z. (2007). Yapay sinir ağlarına genel bakış, *Turkiye Klinikleri J Med Sci*, 27:65-71.

Yüce H. (2021). Normalizasyon tekniklerinin biyomedikal verilerde sınıflama başarısına etkisi, *Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü*, Yüksek Lisans Tezi.