JAIHS 2021; 1(1):6-11
DOI 10.52309/jai.2021.2

Orijinal Araştırma/Original Research

Sağlık Veri Setlerinde Öznitelik Seçiminin Sınıflandırma Performansına Etkisi

The Effect of Feature Selection Methods to Classification Performance in Health Datasets

Mert DEMİRARSLAN[1], Aslı SUNER*[2]

ÖZ

Giriş: Günümüzde veri setleri, farklı cihazlardan toplanan verilerle çok yüksek boyutlu ve spesifik hale geldiğinden, öznitelik seçimi veri madenciliğinde veri boyutunu azaltmada önemli bir veri ön işleme adımıdır. Bu çalışma, öznitelik seçim yöntemlerini kullanarak makine öğrenmesi yöntemlerinin hesaplama süresini ve maliyetini düşürüp sınıflandırma performansının iyileştirilmesini amaçlamaktadır. Gereç ve Yöntem: Öznitelik seçim yöntemleri; filtreleme yöntemleri, sarmal yöntemler ve gömülü yöntemler olmak üzere üç ana başlık altında incelenmektedir. Çalışmada, makine öğrenmesi sınıflandırma algoritmalarından destek vektör makinesi, Naïve Bayes ve karar ağaçları yöntemleri kullanılmıştır. Çalışmada kullanılan veriler UCI ve Kaggle veri tabanlarından elde edilmiştir. Algoritmaların sınıflandırma performanslarını karşılaştırmak için doğruluk, duyarlılık, özgüllük, kesinlik ve F ölçütü değerleri hesaplanmıştır. Tüm analizlerde WEKA 3.8.3, R3.3.0 ve Tableu programları kullanılmıştır. Analizlerde uygun yöntemler kullanılarak gereksiz öznitelikler çıkarıldıktan sonra; algoritmaların sınıflandırma performansları ve çalışma süreleri hesaplanmıştır. Bulgular: Doğruluk değerleri, öznitelik seçiminden sonra kullanılan veri setlerinde MNIST için % 87'e, Parkinson için % 85'e, SCADI için % 97'ye, HCC için % 100'e ve meme kanseri için % 78'e yükselmiştir. En yüksek performansa sahip algoritma karar ağaçları (J48) sarmal yöntem öznitelik seçimi ile elde edilmiştir. En hızlı metot filtreleme yöntemi iken, en uzun süre çalışan algoritma sarmal yöntemdir. Bulgulara göre, çok sayıda özniteliğe sahip verilerin sınıflandırma performansları, öznitelik seçimi yapılmış verilere göre daha düşük bulunmuştur. Sonuç: Sonuç olarak; düşük boyutlu veri setleri, daha düşük hesaplama maliyetleri ile daha yüksek sınıflandırma doğruluğu sağlayabilmektedir.

Anahtar kelimeler: Veri madenciliği, Öznitelik seçimi, Makine öğrenmesi, Sınıflandırma, Sağlık verileri

ABSTRACT

Introduction: Nowadays, since data sets become very high-dimensional and specific with the data collected from different devices, attribute selection has an important pre-processing task in reducing data size in data mining. This study aims to improve classification performance by reducing the calculation time and cost by using attribute selection methods. Materials and Methods: Attribute selection methods are examined under three main headings: filter method, wrapper method and embedded method. In the study, support vector machine, Naïve Bayes and decision trees methods (J48) among the machine learning classification algorithms were used. Data sets were obtained from UCI and Kaggle databases. Accuracy, sensitivity, specificity, precision and F-measure values were calculated to compare the classification performances of the algorithms. WEKA version 3.8.3, R3.3.0 and Tableu programs were performed in all analyzes. After unnecessary features were extracted by using appropriate methods in the analysis; classification performances and run times of algorithms were calculated. Results: Accuracy values increased to 87% for Colorectal Histology MNIST, 85% for Parkinson's disease, 97% for SCADI, 100% for HCC, and 78% for breast cancer after attribute selection. The algorithm with the highest performance was found as a wrapper method with decision trees (J48). While the fastest algorithm was filter method, the longest-running algorithm was the wrapper method. According to results, the performance improvement was higher in feature sets with a large number of attributes after selecting feature. Conclusion: As a result, low-dimensional data sets may provide higher classification accuracy with lower calculation costs

Keywords: Data mining, Feature selection, Machine learning, Classification, Health Datasets

Received/Geliş : 7.03.2021
Accepted/Kabul: 21.03.2021
Publication date: 15.04.2021

Mert DEMİRARSLAN
Ege Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, İzmir, Türkiye
ORCID: 0000-0001-8848-7340

Aslı SUNER
Ege Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, İzmir, Türkiye
asli.suner@ege.edu.tr
ORCID: 0000-0002-6872-9901

Telif Hakkı (c) 2021 Sağlık Bilimlerinde Yapay Zeka Dergisi (Journal of Artificial Intelligence in Health Sciences)
Bu çalışma Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License ile lisanslanmıştır.

1. GİRİŞ

Yapay zeka uygulamaları, sağlık alanında yaygın bir şekilde kullanılmaktadır. Sağlık alanında, hayati önem taşıyan konularda hızlı ve doğru kararlar verilmesi gerektiğinden, özellikle hastalık tanısı koymada sınıflandırma algoritmalarının sıklıkla kullanıldığı görülmektedir (1). Bu algoritmaları besleyen hastalık tanısı veri setlerinin doğru ve yüksek performanslı değerler içermesi için kullanılan verilerin düzgün, temiz, sınıflandırma algoritmalarının kullanımı için uygun şekilde olması büyük önem taşımaktadır. Ancak veri setlerinde kayıp gözlem, sınıf gürültüsü, sınıf dengesizliği, aykırı gözlem, korelasyon olması ve ilgisiz değişken gibi bir çok problem ortaya çıkabilmektedir (2). Bu durum da algoritmaların performans değerlerini olumsuz yönde etkileyebilmektedir.

Sağlık veri setlerinde diğer verilerde de görüldüğü gibi sınıflandırma performansını düşüren ya da yavaşlatan ilgisiz değişkenler bulunmaktadır (3). Bu ilgisiz değişkenlerin veri setinden uzaklaştırılması için literatürde birçok öznitelik seçim yöntemi önerilmiştir. Yang ve ark. (1997) yaptıkları çalışmada öznitelik seçim yöntemlerinde genetik algoritmaları 17 farklı veri seti ile kullanmışlar ve sınıflandırma performansının farklı algoritmalarda yükseldiğini göstermişlerdir (4). John ve ark. (1997) yaptıkları çalışmada, sınıflandırma algoritmalarında oluşan aşırı öğrenme problemini gidermek için sarmal yöntemler ile ilgisiz değişkenleri veri setinden çıkartarak aşırı öğrenme problemini çözümlemişlerdir (3). Rodriguez ve ark. (2018) ise öznitelik seçim yöntemlerinin (filtreleme, sarmal ve gömülü) sınıflandırma performanslarını karşılaştırmışlardır. Filtreleme yöntemlerinin daha hızlı, sarmal ve gömülü yöntemlerin ise daha yavaş ancak filtreleme yöntemlerine göre daha başarılı olduklarından bahsetmişlerdir (5).

Bu çalışma, öznitelik seçim yöntemlerini kullanarak makine öğrenmesi yöntemlerinin hesaplama süresini ve maliyetini düşürerek sınıflandırma performansının iyileştirilmesini amaçlamaktadır.

2.YÖNTEM

Çalışmada makine öğrenmesi algoritmaları kullanılırken %70 eğitim verisi ve %30 test verisi olarak alınmıştır. Öznitelik seçim yöntemleri filtreleme, gömülü ve sarmal yöntemler olarak incelenmiştir. Filtreleme yöntemlerinden korelasyon tabanlı öznitelik seçim yöntemi; sarmal yöntemlerden rasgele ağaç yöntemi ve gömülü yöntemlerden ileri artırımlı yöntem kullanılmıştır. Sınıflandırma yöntemlerinden Naïve Bayes; destek vektör makinaları ve karar ağacı algoritmalarından C4.5 kullanılmıştır (6). Algoritmaların sınıflandırma performansı hesaplanırken ölçüm metriklerinden doğruluk, duyarlılık, özgüllük, kesinlik ve F ölçütü tercih edilmiştir. Tüm analizlerde WEKA 3.8.3, R3.3.0 ve Tableu programları, Windows 10 işletim sisteminde kullanılmıştır.

2.1. Veri Setleri

Çalışmada kullanılan kolorektal kanser (Colorectal Histology MNIST), Parkinson hastalığı (Parkinson's Disease), öz bakım aktiviteleri (SCADI), hepatoselüler kanser (HCC) ve meme kanseri (Breast Cancer) veri setleri, UCI (7) ve Kaggle (8) veri tabanlarından elde edilmiştir. Veri setlerinin seçiminde, öznitelik seçim yöntemleri arasındaki farklılıkları incelemek amacıyla gözlem sayılarının ve öznitelik sayılarının farklı olmasına dikkat edilmiştir. Veri setleri hakkında bilgilerin özetlendiği Tablo 1'e göre kolorektal kanser veri seti 2 sınıfa, 1250 örneklem büyüklüğüne ve 785 özniteliğe sahiptir. Parkinson veri seti 2 sınıflı iken örneklem büyüklüğü ve öznitelik sayısı 755'tir. SCADI veri setindeki sınıf sayısı 7, örneklem büyüklüğü 70 ve öznitelik sayısı 206'dır. HCC veri setinde sınıf sayısı 2, örneklem büyüklüğü 204 ve öznitelik sayısı 45'tir. Meme kanseri verisinde ise 2 sınıf yer almakta, örneklem büyüklüğü 286 iken öznitelik sayısı 10'dur. Sadece HCC veri setinde %10,22 oranında kayıp gözlem bulunmaktadır.

Veri Seti	N	Öznitelik Sayısı	Sınıf Sayısı
Colorectal Histology MNIST	1250	785	2
Parkinson's Disease	755	755	2
SCADI	70	206	7
HCC	204	45	2
Breast Cancer	286	10	2

Tablo 1: Veri setlerine ilişkin bilgiler

2.2. Öznitelik seçim yöntemleri

Literatüre bakıldığında öznitelik seçim yöntemleri 3 ana başlık altında toplanmıştır (9). Bunlardan ilki istatistiksel yöntemlere dayanan ve bu sayede hızlı sonuçlar alınmasını sağlayan filtreleme yöntemleridir. İkincisi, makine öğrenmesi yöntemlerine dayanan ve her aşamada sınıflandırıcı ile ilişki kuran sarmal yöntemlerdir. Üçüncü grupta ise makine öğrenmesi algoritmaları ve öznitelik seçim yöntemlerinin birlikte çalıştığı ve yine her aşamada sınıflandırıcı ile ilişki kurarak çalışan gömülü yöntemler yer almaktadır.

2.2.1. Filtreleme yöntemleri

Filtreleme yöntemleri, boyut indirgeme, öznitelik seçim işlemlerinde kullanılan en eski tekniklerdir (10). İstatistiksel yöntemler kullanılarak yapılan bu işlemlerde sınıflandırma yöntemleri kullanılmamaktadır. Bu sebeple algoritmalar daha hızlı çalışmakta ve daha hızlı sonuçlar alınmaktadır. Bu sayede de hesaplamalar yapılırken zaman ve maliyet açısından yüksek fayda sağlanmaktadır. Gömülü ve sarmal yöntemlere göre daha az karmaşık, açıklanabilirliği daha yüksektir. Filtreleme yöntemleri; Fisher skor, t-skor, korelasyon tabanlı filtreleme ve bilgi kazancı gibi yöntemlerden oluşmaktadır.

2.2.2. Gömülü yöntemler

Gömülü yöntemlerde hem özellik seçim algoritmaları hem de sınıflandırma algoritmaları bir arada kullanılmaktadır. Bu yüzden gömülü yöntemler filtreleme yöntemlerine göre, sarmal yöntemler gibi daha yavaş ve daha yüksek maliyetli olmaktadır. Filtreleme yöntemleri, hızlı ve düşük maliyetli olsa da sınıflandırma yöntemlerini kullanmadığından sınıflandırmada bazı sorunlar ya da düşük performanslar görülebilmektedir. Bununla birlikte, sarmal yöntemlerin, özellikle mikrodizi verilerinin yüksek boyutsallığı ile artan bir hesaplama maliyeti olabilmektedir (9). Araştırmacılar için ara bir çözüm olan ve özellikleri sıralamak için bir kriter oluşturmada sınıflandırma yöntemlerini de kullanan gömülü yöntemler keşfedilmiştir. Örneğin, karar ağaçları ya da destek vektör makine yöntemleri ile alt kümelerin oluşturulup sıralanmasının ardından istenilen düzeydeki öznitelikler seçilebilmektedir.

2.2.3. Sarmal yöntemler

Sarmal yöntemlerde, öznitelik seçimi için makine öğrenmesi algoritmaları kullanılarak en yüksek performans gösteren yöntem seçilmektedir. Bu yöntemde en iyi alt küme oluşturma ve seçme tekniği filtreleme yöntemlerine göre daha başarılı olmakta; ancak her aşamada sınıflandırıcı başarısına bakıldığından daha yavaş ve daha yüksek maliyetli olmaktadır. Alt küme arama stratejileri ola-

rak; ardışık ileri yönde seçim, ardışık geri yönde seçim, ardışık ileri yönde kayan seçim, ardışık geri yönde kayan seçim, l ekle r çıkar ve genetik algoritmalar gibi farklı yöntemler kullanılabilmektedir (9).

2.3. Makine öğrenmesi sınıflandırma algoritmaları

Makine öğrenmesi, temel olarak yapay zekanın sayısal öğrenme ve model tanıma çalışmalarından geliştirilen bilgisayar biliminin bir alt dalıdır. Makine öğrenmesi algoritmaları; verilerin yapısını ve işlevini öğrenen, aynı zamanda veri seti üzerinde bir öngörüde bulunabilen sistemlerdir. Bu algoritmalar, örnek girişlerden veri tabanlı tahminler ve kararlar gerçekleştirmek için bir model oluşturarak çalışmaktadırlar (11).

2.3.1. Naïve Bayes (NB) sınıflandırıcı

Naïve Bayes koşullu olasılık hesaplama yöntemi Thomas Bayes tarafından 1812 yılında bulunmuştur. Bu yöntem, rassal bir değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi göstermektedir. Naïve Bayes sınıflandırma yönteminin temeli Bayes teoreminden gelmektedir. Basit sınıflandırma algoritmaları kategorisinde yer almakta ve dengesiz sınıflı verilerde de çalışmaktadır. Algoritmanın çalışma prensibi, bir eleman için her durumun olasılığını hesaplamakta ve olasılık değeri en yüksek olana göre sınıflandırılmaktadır. Küçük boyutlu eğitim verisiyle de çalışabilmektedir. Test kümesindeki bir değerin eğitim kümesinde gözlemlenemeyen bir değeri varsa olasılık değeri olarak 0 verdiğinden tahmin yapamamaktadır. Bu durum genellikle sıfır frekans adıyla bilinmektedir. Bu durumu çözmek için düzeltme teknikleri kullanılabilmektedir. En basit düzeltme tekniklerinden biri Laplace tahminidir (12).

2.3.2.Destek vektör makinaları (DVM)

Temelleri 1960 yılında Vladamir N. Vapnik tarafından atılan destek vektör makineleri, 1990 yılında tam anlamıyla geliştirilmiştir. Başlangıçta iki sınıflı veriler için tasarlansa da daha sonradan çok sınıflı verileri sınıflandırmak için geliştirilmiştir. Bu yöntemin eğitim süresi diğer algoritmalara göre yavaş olsa da güvenilirliği daha yüksektir ve doğrusal olmayan gözlemlerde de üstün başarı göstermektedir (13).

2.3.3. Karar ağacı C4.5

Karar ağacı C4.5 algoritması J. Ross tarafından bulunmuştur. Verideki bağımlı/sınıf öznitelik için entropi (bilgi kazancı) değeri hesaplamaktadır. Bu yöntem, entropi değerine göre değerlendirmekte ve her seferinde tek bir özelliği dikkate almaktadır. Entropi, veri setindeki karmaşıklık veya belirsizliğin ölçümü olarak tanımlanabilmektedir (Formül 1). Entropi ölçüsü, değişkenleri sıralamak için bir ölçüt olarak kabul edilmektedir. Y özniteliğini anlamak için gerekli olan bilgi ile X özniteliğini de kullanıp aralarındaki fark ile Y özniteliğini kullanmaya bilgi kazancı denmektedir (Formül 2). Bir sınıf özelliği olarak Y'nin entropisi, p(y) rastgele değişken olan Y için marjinal olasılık yoğunluk fonksiyonudur. Entropi ölçüsü hesaplamada Y özniteliğinin entropi ölçüsü X özniteliğine göre gruplanan verilerden yüksek olacaktır (Formül 3). Yöntem tüm özellikleri düzenli bir şekilde sınıflandırmakta, daha sonra elde edilen sıraya göre en yüksek kazanç bilgisine sahip olanı göstermektedir (14).

$$\text{Entropi(Y)} = \sum_{i\epsilon Y} p(i) log_2(p(i)) - \quad (1)$$

$$\text{Bilgi Kazancı} = H(Y) - H(Y|X) \quad (2)$$

$$H(Y|X) = -\sum_{j\epsilon X} p(j) \sum_{i\epsilon Y} p(i|j) log_2(p(j|i)) \quad (3)$$

2.4. Performans metrikleri

Performans metrikleri, hata matrisi yardımıyla hesaplanmakta ve 0 ile 1 arasında değerler almaktadır. Hesaplamalar sonucunda, metriklerin değerleri 1'e yaklaştıkça iyi performans gösterdikleri söylenmektedir. Doğru pozitif (DP), tahmin edilenin pozitif ve gerçek durumun pozitif olduğu durumdur (örneğin hastayı hasta olarak tahmin etmek). Yanlış pozitif (YP), tahmin edilenin pozitif ancak gerçek durumun negatif olduğu durumdur (örneğin sağlıklı bireyi hasta olarak tahmin etmek). Yanlış negatif (YN), tahmin edilenin negatif ancak gerçek durumun pozitif olduğu durumdur (örneğin hastayı sağlıklı olarak tahmin etmek). Doğru negatif (DN), tahmin edilenin negatif gerçek durumun negatif olduğu durumdur (örneğin sağlıklı bireyi sağlıklı olarak tahmin etmek) (15). Bu hesaplamaların yapılmasında kullanılan hata matrisi aşağıdaki şekilde gösterilmektedir (Tablo 2):

Tahmin Durumu	Gerçek Durum	
	Pozitif	Negatif
Pozitif	DP	YP
Negatif	YN	DN

Tablo 2: Hata Matrisi

Doğruluk (Accuracy): Sınıflandırma performansı incelemesinde en çok dikkat edilen kısım doğruluk değeridir. Doğru pozitif ve doğru negatif değerlerinin tüm değerlere olan oranı ile bulunmaktadır. Örneğin doğruluk değeri 1 olduğunda gerçekten hasta olan ile hasta olmayan bireylerin sınıflandırılmasının tam olarak doğru yapıldığı sonucuna varılır. Doğruluk formülü (Formül 4) aşağıdaki gibi hesaplanmaktadır:

$$ACC = \frac{D\ P\ D\ N}{D\ P\ Y\ P\ Y\ N\ D\ N} \quad (4)$$

Duyarlılık (Sensitivite): Bu metrik; doğru pozitif olan, tahmin edilenin pozitif ve gerçek durumun pozitif olduğu kısımla ilgilenmektedir. Örneğin, hastayı hasta olarak tahmin ederek hasta bireylerin ayırt edilmesini sağlamaktadır (16). Ayrıca duyarlılık ölçümü testin gücüne de eşittir (güç=1-β). Duyarlılık formülü (Formül 5) aşağıdaki gibi hesaplanmaktadır:

$$SEN = \frac{D\ P}{D\ P\ Y\ N} \quad (5)$$

Özgüllük (Specificity): Bu metrik; doğru negatif olma durumuyla, başka bir ifadeyle tahmin edilenin negatif gerçek durumun da negatif olduğu durum ile ilgilenmektedir. Örneğin sağlıklı bireyi sağlıklı olarak tahmine etmektedir (16). Seçicilik formülü (Formül 6) aşağıdaki gibi hesaplanmaktadır:

$$SPE = \frac{D\ N}{D\ N\ Y\ P} \quad (6)$$

Kesinlik (Precision): Doğru tahmin edilen tüm pozitif sınıflardan kaç tanesinin gerçekten pozitif olduğunu tahmin eden metriktir. Başka bir deyişle, tanı testi sonucu pozitif olanların hasta olma olasılığını tahmin etmedir (16). Kesinlik formülü (Formül 7) aşağıdaki gibi hesaplanmaktadır:

$$PRE = \frac{D\ P}{D\ P\ Y\ P} \quad (7)$$

F-ölçütü: Özgüllük ve kesinlik ölçümlerinin harmonik ortalama-

sıyla hesaplanan F-ölçütü, her iki metriğin birlikte değerlendirilmesine olanak tanımaktadır (16). F-ölçütü formülü (Formül 8) aşağıdaki gibi hesaplanmaktadır:

$$F = 2 * \frac{Özgüllük * Kesinlik}{Özgüllük + Kesinlik} \qquad (8)$$

3.BULGULAR

Veri setlerinin ham halindeki sınıflandırma performanslarına bakıldığında kolorektal kanser verisi en yüksek doğruluğu DVM algoritmasında 0,72 doğruluk, duyarlılık, özgüllük, kesinlik ve F değeri göstermiştir. Parkinson verisi, C4.5 ve DVM algoritmalarıyla 0,80 doğruluk, duyarlılık, özgüllük, kesinlik ve F değeri; SCADI verisi DVM algoritmasıyla 0,79 doğruluk, duyarlılık, özgüllük, kesinlik ve F değeri; HCC verisi DVM algoritmasıyla 0,75 doğruluk, duyarlılık, özgüllük, kesinlik ve F değeri; meme kanseri verisi NB algoritmasıyla 0,71 doğruluk, duyarlılık, özgüllük, kesinlik ve F değeri göstermiştir. Verilere korelasyon tabanlı filtreleme yöntemi uygulandığında, kolorektal kanser verisinde sınıflandırma performansı destek vektör makinesi algoritması ile verinin ham haline göre 0,82 doğruluk oranına yükselmiştir. Parkinson verisinde DVM algoritması en yüksek 0,82 doğruluk, SCADI verisinde DVM en yüksek 0,97 doğruluk, HCC verisinde DVM algoritması en yüksek 0,97 doğruluk ve meme kanseri verisinde en yüksek C4.5 algoritması en yüksek 0,76 doğruluk göstermiştir. Verilere öznitelik seçiminden gömülü yöntemlerde ileri artırımlı yöntem kullanıldığında, kolorektal kanser verisinde doğruluk oranı C4.5 algoritması ile verinin ham haline göre 0,81 oranına yükselmiştir. Parkinson verisinde C4.5 algoritması en yüksek 0,86 doğruluk, SCADI verisinde DVM algoritması en yüksek 0,91 doğruluk, HCC verisinde DVM algoritması 100 doğruluk, meme kanseri verisinde NB algoritması 0,78 doğruluk göstermiştir. Verilerde sarmal yöntemler için rasgele orman yöntemi uygulandığında, kolorektal kanser verisinin doğrululuk oranının en yüksek C4.5 algoritmasıyla 0,87 doğruluk oranı sahip olmuştur. Diğer verilerde bu durum; Parkinson verisinde C4.5 algoritması ile 0,80 doğruluk, SCADI verisinde C4.5 ile 0,88 doğruluk, HCC verisinde DVM algoritması ile 0,92 doğruluk ve meme kanseri verisinde C4.5 ile 0,78 doğruluk şeklinde olmuştur (Tablo 3).

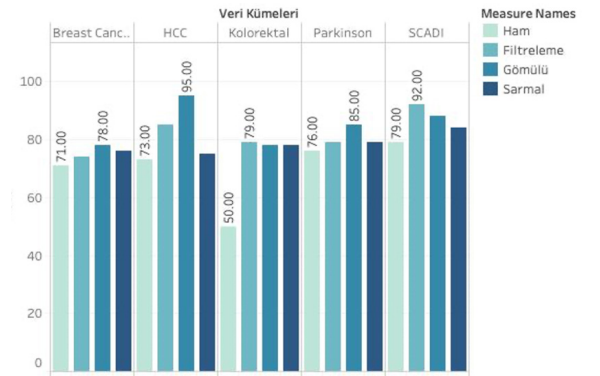
	Veri Seti	Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F-ölçütü	Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F-ölçütü	Doğruluk	Duyarlılık	Özgüllük	Kesinlik	F-ölçütü
Ham Veri	Colorectal	0,54	0,54	0,54	0,54	0,54	0,50	0,51	0,51	0,50	0,51	0,72	0,73	0,72	0,73	0,73
	Parkinson	0,80	0,81	0,81	0,80	0,81	0,76	0,75	0,75	0,76	0,75	0,80	0,80	0,80	0,80	0,80
	SCADI	0,70	0,70	0,70	0,70	0,70	0,79	0,79	0,79	0,79	0,79	0,79	0,79	0,79	0,79	0,79
	HCC	0,37	0,36	0,36	0,37	0,37	0,73	0,74	0,74	0,73	0,74	0,75	0,74	0,75	0,74	0,74
	Breast Cancer	0,68	0,69	0,69	0,68	0,68	0,71	0,71	0,71	0,71	0,71	0,69	0,68	0,69	0,68	0,68
Filtreleme	Colorectal	0,81	0,81	0,81	0,81	0,81	0,79	0,80	0,80	0,79	0,80	0,82	0,82	0,82	0,82	0,82
	Parkinson	0,80	0,80	0,80	0,80	0,80	0,79	0,79	0,79	0,79	0,79	0,83	0,82	0,83	0,82	0,83
	SCADI	0,88	0,89	0,89	0,88	0,88	0,92	0,93	0,93	0,92	0,92	0,97	0,96	0,97	0,96	0,96
	HCC	0,68	0,68	0,68	0,68	0,68	0,85	0,84	0,84	0,85	0,85	0,97	0,96	0,97	0,96	0,97
	Breast Cancer	0,76	0,76	0,76	0,76	0,77	0,74	0,74	0,74	0,74	0,74	0,68	0,68	0,68	0,68	0,68
Gömülü	Colorectal	0,81	0,81	0,81	0,81	0,81	0,78	0,78	0,78	0,78	0,78	0,79	0,79	0,79	0,79	0,79
	Parkinson	0,85	0,84	0,84	0,85	0,85	0,85	0,85	0,85	0,85	0,85	0,84	0,84	0,84	0,84	0,84
	SCADI	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,88	0,91	0,91	0,91	0,91	0,91
	HCC	0,49	0,49	0,49	0,49	0,49	0,95	0,95	0,95	0,95	0,95	100	100	100	100	100
	Breast Cancer	0,77	0,78	0,78	0,77	0,77	0,78	0,79	0,79	0,78	0,79	0,74	0,73	0,74	0,73	0,73
Sarmal	Colorectal	0,87	0,86	0,86	0,87	0,86	0,78	0,78	0,78	0,78	0,78	0,78	0,78	0,78	0,78	0,78
	Parkinson	0,80	0,80	0,80	0,80	0,80	0,79	0,77	0,77	0,79	0,78	0,77	0,76	0,77	0,76	0,76
	SCADI	0,88	0,86	0,86	0,88	0,88	0,84	0,85	0,85	0,84	0,85	0,88	0,88	0,88	0,88	0,88
	HCC	0,50	0,50	0,50	0,50	0,50	0,75	0,75	0,75	0,75	0,75	0,92	0,91	0,92	0,91	0,92
	Breast Cancer	0,78	0,78	0,78	0,78	0,78	0,76	0,75	0,75	0,76	0,76	0,69	0,68	0,69	0,68	0,69

Tablo 3: Öznitelik seçim yöntemlerinin algoritmalardaki sınıflandırma performansları

Araştırmada kullanılan verilerde algoritmaların başarıları incelendiğinde en yüksek doğruluk değerine sahip olan yöntemler; meme kanseri verisinde C4.5 algoritmasına göre gömülü yöntem, HCC verisinde filtreleme yöntemi, kolorektal kanser verisinde sarmal yöntem, Parkinson verisinde sarmal yöntem ve SCADI verisinde sarmal ve gömülü yöntemdir (Şekil 1).

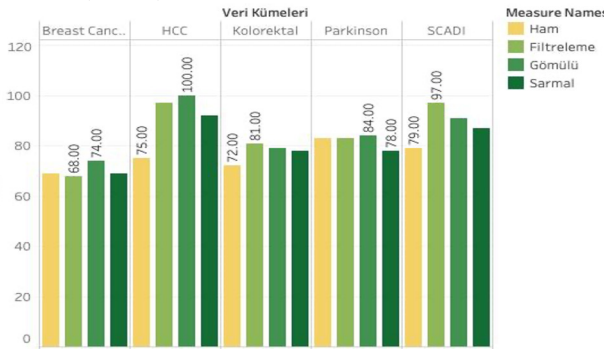
Şekil 1: C4.5 algoritmasına göre öznitelik seçim yöntemlerinin verilerdeki doğruluk değerleri

NB algoritmasına göre öznitelik seçim yöntemlerinin verilerdeki doğruluk değerlerine bakıldığında, meme kanseri verisinde 0,78 doğruluk oranıyla gömülü yöntem, HCC verisinde 0,95 doğruluk ile gömülü yöntem, kolorektal kanser verisinde 0,79 doğrulukla filtreleme yöntemi, Parkinson verisinde 0,85 doğrulukla gömülü yöntem, SCADI verisinde 0,92 doğruluk ile filtreleme yöntemi en yüksek başarıyı göstermiştir (Şekil 2).



Şekil 2: NB algoritmasına göre öznitelik seçim yöntemlerinin verilerdeki doğruluk değerleri

DVM algoritmasına göre öznitelik seçim yöntemlerinin verilerdeki doğruluk değerlerine bakıldığında; meme kanseri verisinde 0,74 doğrulukla gömülü yöntem, HCC verisinde 100 doğrulukla gömülü yöntem, kolorektal kanser verisine göre 0,81 doğrulukla filtreleme yöntemi, Parkinson verisinde 0,87 doğruluk ile gömülü yöntem ve SCADI verisinde 0,97 ile filtreleme yöntemi başarılı olmuştur (Şekil 3).



Şekil 3: DVM algoritmasına göre öznitelik seçim yöntemlerinin verilerdeki doğruluk değerleri

Tablo 4'te öznitelik seçim yöntemlerinin çalışma sürelerine bakıldığında; kolorektal kanser verisinde filtreleme yöntemi 2 saniye, gömülü yöntem 32 saniye ve sarmal yöntem 69 saniye sürede sonuç vermiştir. Parkinson verisinin çalışması ise filtreleme yönteminde 4 saniye, gömülü yöntemde 37 saniye, sarmal yöntemde 77 saniye sürmüştür. SCADI verisi, filtreleme yöntemi 2 saniye, gömülü yöntem 21 saniye, sarmal yöntem 47 saniye çalışırken; HCC verisinin çalışması filtreleme yöntemi için 2 saniye, gömülü yöntem için 13 saniye ve sarmal yöntem için ise 35 saniye sürmüştür. İstatistiksel tabanlı olan filtreleme yöntemleri her veri setinde ve her algoritmada en kısa çalışma süresine sahiptir.

Veri Seti	Filtreleme	Gömülü	Sarmal
Colorectal Histology MNIST	2 s	32 s	69 s
Parkinson's Disease	4 s	37 s	77 s
SCADI	2 s	21 s	47 s
HCC	2 s	13 s	35 s
Breast Cancer	2 s	13 s	35 s

Tablo 4: Öznitelik seçim yöntemlerinin çalışma süreleri

4.TARTIŞMA

Çalışmada kullanılan veri setlerinin orijinal/ham hallerinde algoritmaların sınıflandırma performansları oldukça düşük değerler göstermiştir. Verilerde öznitelik seçimi yapıldıktan sonra ilgili değişkenlerin seçiminden dolayı algoritmaların sınıflandırma performansları yükselmiştir. Öznitelik seçim yöntemlerinin çalışma sürelerine bakıldığında, tüm veri setlerinde en hızlı çalışan yöntem filtreleme yöntemi, ardından sırasıyla gömülü yöntem ve sarmal yöntem olmuştur. Filtreleme yöntemi, istatistik tabanlı olması nedeniyle hızlı çalışmaktadır. Gömülü yöntem ve sarmal yöntem, her adımda sınıflandırma performansı hesaplamasından dolayı filtreleme yöntemlerine göre daha yavaş hesaplama yapmaktadır. Literatüre bakıldığında, Parkinson veri seti için, Patra ve ark. (2019) yaptıkları çalışmada, basit ve topluluk öğrenme algoritmalarının sınıflandırma performanslarını araştırmışlardır (17). Yapılan çalışmada herhangi bir öznitelik seçim işlemi yapılmamıştır. Topluluk öğrenme algoritmalarının sınıflandırma performansları incelendiğinde; rasgele orman için 0,84; Bagging için 0,81 ve Adaboost için 0,82 doğruluk yüzdesi elde edilmiştir. Bizim çalışmamızda Parkinson verisinde gömülü yöntem öznitelik seçimi yapıldıktan sonra C4.5 ve Naïve Bayes algoritmalarında 0,85 doğruluk elde edilmiştir. Coudhury and Grene (2019), SCADI verisi için çalışmalarında birçok sınıflandırma algoritmasını karşılaştırmışlar, öznitelik seçim işleminde Boruna algoritmasını kullanarak DVM algoritmasında 0,83; Naïve Bayes algoritmasında 0,83 doğruluk elde etmişlerdir (18). Bizim çalışmamızda ise filtreleme yöntemi kullanılarak ilgili değişkenlerin seçilmesinin ardından DVM algoritması kullanılarak 0,97 doğruluk değeri ile daha yüksek performans elde edilmiştir. Pal ve ark. (2018) öznitelik sıralama yöntemlerinin sınıflandırma performansına etkisini araştırırken HCC veri setinden faydalanmışlardır (19). Yaptıkları çalışmada, veri ön işleme adımında kayıp gözlem problemini kübik spline veri interpolasyonu yöntemi ile atama yaparak çözmüşlerdir. Öznitelik seçimi sıralama işlemlerinde ReliefF, mRMR, karşılıklı bilgi yöntemi, hızlı korelasyon tabanlı filtre ve kendi önerdikleri öznitelik sıralama yöntemlerini kullanmışlardır. Kendi önerdikleri öznitelik sıralama yöntemi ile SVM algoritması 0,76 ile en yüksek doğruluk değerini sağlamıştır. Bizim çalışmamızda HCC verisinde gömülü yöntemlerde DVM algoritması %100 doğruluk göstermiştir. Ancak bu durum verinin dengesiz sınıf dağılımından kaynaklı olarak aşırı öğrenmeye neden olmuş olabileceğinden, filtreleme yöntemlerindeki DVM algoritmasındaki 0,97 doğruluğu ele almak daha doğru olacaktır. Chaurasia ve Pal (2014), meme kanseri verisi için yaptıkları sınıflandırma çalışmasında öznitelik seçimi uygulamamışlardır (20). Kullandıkları sınıflandırma algoritmaları ile Rep Tree için 0,71; RBF Network için 0,73 ve Simple Logistic için 0,74 doğruluk yüzdesi elde etmişlerdir. Bizim çalışmamızda ise gömülü yöntem kullanılarak Naïve Bayes algoritması ile 0,78 ve sarmal yöntem kullanılarak C4.5 algoritması ile 0,78 doğruluk değeri elde edilmiştir.

5.SONUÇ

İlgisiz değişken probleminin giderilmesinden sonra, algoritmaların sınıflandırma performanslarının orijinal halindeki sınıflandırma performanslarına göre yükseldiği görülmüştür. Veri setlerinde değişken sayıları ile örneklem genişliklerinin birbirlerinden oldukça farklı olması da sınıflandırma performansını olumsuz yönde etkilemektedir. Sağlık veri setlerinde öznitelik seçim işlemi yapılırken filtreleme yöntemlerinin kullanılması, yüksek sınıflandırma performansı sağlarken zaman ve maliyet açısından da fayda sağlayacaktır. Gelecek çalışmalarda, örneklem sayısı ve öznitelik sayısı 1000'den fazla olan verilerle çalışılarak öznitelik seçim yöntemlerinin performanslarının ve çalışma sürelerinin karşılaştırılması önerilmektedir.

KAYNAKLAR

[1] Deo RC. Machine learning in medicine, Circulation. 2015;132:1920-1930. doi:10.1161/Circulationaha.115.001593.

[2] Lin JH, Haug PJ. Data preparation framework for preprocessing clinical data in data mining, AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2006. p. 489.

[3] Kohavi R, John GH. Wrappers for feature subset selection. Artificial intelligence, 1997, 97.1-2: 273-324. doi.org/10.1016/S0004-3702(97)00043-X.

[4] Yang J, Honavar V. Feature subset selection using a genetic algorithm. In Feature extraction, construction and selection. Springer, Boston, MA, 1998. p. 117-136.

[5] Rodriguez GV, Luque EJ, Chica OM, Mendes MP. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Science of the total environment. 2018, 624: 661-672.

[6] D Chen DY. Pandas for everyone, Python data analysis. Addison-Wesley Professional, 2017. p.161.

[7] UCI Machine Learning Repository [Internet]. Available from: https://archive.ics.uci.edu/ml/index.php

[8] Open Datasets and Machine Learning Projects | Kaggle [Internet]. Available from: https://www.kaggle.com/datasets

[9] Bolón CV, Sánchez MN, Alonso BA. Feature selection for high-dimensional data. Cham, Springer International Publishing. 2015.

[10] Zhang L, Duan Q. A feature selection method for multi-label text based on feature importance. Applied Sciences, 2019, 9.4:665. doi: 10.1007/s11042-018-6083-5.

[11] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 2001, 23.1: 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X.

[12] Choubey DK, Paul S, Kumar S. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. In Communication and computing systems: proceedings of the international conference on communication and computing system (ICCCS 2016). 2017. p. 451-455. doi:10.1201/9781315364094-82.

[13] Indrayan A, Holt MP. Concise encyclopedia of biostatistics for medical professionals. Crc Press, 2016.
[14] Bramer M. Principles of data mining (Vol. 180). London: Springer, 2007.
[15] Ian AC, Bengio GY. Deep Learning Book. Deep Learn.,2015, 21(1), 111-124.
[16] Powers, D. M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv: 2010.16061, 2020.
[17] Patra AK, Ray R, Abdullah AA, Dash SR. Prediction of Parkinson's disease using Ensemble Machine Learning classification from acoustic analysis. In Journal of Physics: Conference Series IOP Publishing, 2019. p. 012041. doi: 10.1088/1742-6596/1372/1/012041.
[18] Choudhury A, Greene CM, Classification of Functioning, Disability, and Health for Children and Youth: ICF-CY Self Care (SCADI Dataset) Using Predictive Analytics. 2021 Mar 13. [Online]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3307719.
[19] Pal P, Singh B, Kaur M. Prediction of Accuracy for Hepatocellular Carcinoma Patients using Cluster based Feature Ranking, International Journal of Medical Research and Health Sciences, 2018, 7.8: 130-140.
[20] Chaurasia V, Pal S. Data mining techniques: To predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing IJCSMC, 2014, 3.1: 10-22.