Avrupa Bilim ve Teknoloji Dergisi
Sayı 31 (Ek Sayı 1), S. 609-615, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 31 (Supp. 1), pp. 609-615, December 2021

Research Article

Relief Özellik Seçim Yöntem Tabanlı Önerilen Hibrit Model ile Kalp Hastalığı Teşhisi

Atınç Yılmaz[1*], Eda Sümer[2]

[1*] Beykent Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, (ORCID: 0000-0003-0038-7519), atincyilmaz@beykent.edu.tr

[2] Beykent Üniversitesi, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği A.B.D, İstanbul, Türkiye (ORCID: 0000-0003-1792-9687), sumereda9@gmail.com

(İlk Geliş Tarihi 31 Ekim 2021 ve Kabul Tarihi 12 Aralık 2021)

(**DOI:** 10.31590/ejosat.1017054)

ATIF/REFERENCE: Yılmaz, A., Sümer, E. (2021). Relief Özellik Seçim Yöntem Tabanlı Önerilen Hibrit Model ile Kalp Hastalığı Teşhisi. *Avrupa Bilim ve Teknoloji Dergisi*, (31), 609-615.

Öz

Bilgi teknolojileri insan hayatının ve günlük yaşantının her alanında önemli bir yer kaplamaktadır. Günümüzde bilgi sistemleri ve teknolojileri sayesinde insan sağlığından, endüstri ve ekonomi alanlarına kadar her alanda fayda sağlanmaktadır. Kan damar hastalıkları, kalp ritmi problemleri ve bunların yanında doğuştan gelen kalp kusurları insan hayatı için önemli ölçüde risk taşıyan kalp hastalıkları başlığı altında yer almaktadır. Bu çalışmada, kalp hastalıklarının cinsiyet, göğüs ağrısı, kolesterol gibi özellikler ele alınarak sık kullanılan makine öğrenmesi yöntemleri olan logistik regresyon (LR), karar ağaçları (KA), çok katmanlı yapay sinir ağları (YSA), K-en yakın komşular (KNN), naif bayes (NB), destek vektör makineleri (SVM) ve önerilen Relief özellik çıkarım yöntem tabanlı hibrit bir yöntem ile kalp hastalıkları için tespit analizi yapılmıştır. Uygulanan yöntemlerin performans değerleri birbirleri ile karşılaştırılarak ilgili problem için optimum model ortaya konmuştur. Yapılan değerlendirmeler sonucunda ise önerilen hibrit modelin diğer makine öğrenmesi yöntemlerine göre hem doğruluk performans ölçütleri hem de zamansal açıdan daha iyi sonuç verdiği gözlemlenmiştir.

Anahtar Kelimeler: Makine öğrenmesi, Relief özellik çıkarım, Hibrit model, Kalp hastalığı.

Diagnosis of Heart Disease with Proposed Hybrid Model Based on Relief Feature Selection

Abstract

Information technologies occupy an important place in all areas of human life and daily life. Today, when it comes to information systems and technology, benefits are provided in every field from human health to industrial and economic fields. Blood vessel diseases, heart rhythm problems and congenital heart defects are heart diseases that pose a great risk to human life. In this study, the most widely applied machine learning algorithms such as logistic regression (LR), decision trees (DT), multilayer artificial neural networks (ANN), K-nearest neighbors (K-NN), naive bayes (NB), support vector machines (SVM) were used and proposed Relief feature selection-based hybrid method. The characteristics of heart diseases such as gender, chest pain, cholesterol was considered by using these methods and the results were discussed. The performance values of the applied methods were compared with each other and the optimum model for the related problem was revealed. As a result of the evaluations, it was observed that the proposed hybrid model was achieved better results than other machine learning methods in terms of both accuracy performance measures and process time.

Keywords: Machine learning, Relief feature extraction, Hybrid model, Heart disease.

[*] Sorumlu Yazar: atincyilmaz@beykent.edu.tr

1. Giriş

İnsan dünyada var olduğundan itibaren konaklama, beslenme, ısınma gibi çeşitli ihtiyaçları ortaya çıkmıştır. Bu ihtiyaçlar insanların hayatlarını sürdürebilmeleri açısından temel değerler olarak yer almaktadır. Bu temel değerlerin olmaması veya eksik olması durumunda insanların yaşam süreleri boyunca kalıtsal ya da çevre faktörü nedeniyle bazı hastalıklar ortaya çıkmıştır. Bu hastalıklardan biri de kalp rahatsızlıklarıdır. Kalp hastalıkları taşıma damarında ya da ana damarlarında oluşabilen tıkanıklık, kalp odacıklarında ya da kalp kapakçıklarında oluşan bir sorundan dolayı ortaya çıkan tedavisi mümkün olan ve erken teşhis edildiği zaman birçok hayatın kurtarıldığı hastalıklardır. Kalp rahatsızlıkları göğüs ağrısı, çabuk yorulma, nefes alamama ve bayılma gibi etkileri olup hayat standartlarını büyük ölçüde etkileyen rahatsızlıklardır [1]. Hastaların ağrı çeşitlerine, yaşlarına, cinsiyetlerine, kanlarındaki şeker ve yağ oranları gibi değerlere göre teşhis konulabilir. Ancak rahatsızlık ilerlemiş ise tedavi sürecine hayat boyu ilaç kullanımı zorunlu hale gelebilmektedir. Bu durumdan ötürü de kişinin günlük yaşamı fazlasıyla etkilenebilmektedir. Bu nedenle erken tanı hastalık ilerlemeden konulmalıdır.

Dünya genelinde ölümcül hastalıklar içerisinde, kalp ve damar hastalıkları bütün ölüm sebepleri arasında ilk sıralarda yer almaktadır. Serebrovasküler hastalıklar ve kalp hastalıkları sıralamada ilk iki ölüm nedenini oluşturmaktadır. Dünya çapında kalp ve damar hastalıkları ölüm sebebi olarak ilk sıralarda yer almaya devam edeceği öngörülmektedir [1]. Kalp hastası olan ve olmayan insanlardan toplanan yaş, cinsiyet, hastalık durumu, şeker hastalığı, kolesterol vb. veriler ile makine öğrenimi teknikleri uygulanarak hastalığın önceden tespiti konusunda çalışmalar yapılıp, erken tanı konabilmektedir.

Makine öğrenmesi; veriler üzerinden bilinmeyenlere dair tahminlerde bulunan ve zaman içinde doğruluğunu artıran sistemlerin bilgisayar ile modellenmesini sağlayan yapay zekâ alt bilim dalıdır. Makine öğrenmesinin temelinde; veriler üzerinden tecrübe ve deneyimi aktarılarak problemin öğretilebilmesi ve bu sayede problemlere yaklaşım sağlanabileceği bulunmaktadır. Öğrenme sürecinin ardından, herhangi bir dış müdahale yapılmadan sistemin davranışlarını bu öğrenime göre ayarlamak amaçlanmıştır. Makine öğrenmesi, öğrenebilme yeteneğine sahiptir. Öğrenebilme yeteneğini veri türlerine dayalı olarak algoritmalar sayesinde yapmaktadır. Makine öğrenimi, bu algoritmaların tasarım ve geliştirme süreçlerini konu edinir.

Tıp alanında teknolojiden faydalanılarak ortaya konan çalışmalar bulunmaktadır. Bununla birlikte literatürde kalp hastalıkları için erken tanı konması konusunda da çalışmalar yer almaktadır. Palaniappan ve ark. (2008) ve Srinavas ve ark. (2010), ve veri madenciliği tekniklerini kullanarak sağlık durumunu ve kalp krizini tahmin etmek için iki ayrı çalışma yürütmüştür [2,3]. Heart-c, Heart-h ve Heart-statlog, UCI [4] veri setleri kullanılarak Weka yazılımı ile sınıflandırma yapılmıştır. UCI veri setleri ile yapılan çalışmalarda Entropi ve ANFIS (Adaptive Neural Fuzzy Inference System) kullanılarak yüksek performans elde edildiği belirtilmektedir [5]. Jin ve ark. (2009), EKG sinyallerini cep telefonu tabanlı giyilebilir bir cihazla gerçek zamanlı ve sürekli olarak izlemiş; üretilen sinyal kayıtları anormal durumları tespit etmek ve önlemek için kullanılmıştır [6]. Bu çalışmada Android ortamında makine öğrenmesi algoritmaları kullanılmıştır. Gömülü elektronik sistem ve cep telefonu yardımıyla EKG izleme ve ritim sınıflandırma sistemi kurularak sistem taşınabilir hale getirilmiştir. Bulut (2010), yapmış olduğu bir çalışma ile kalp krizi nedenlerinden olan yüksek tansiyonun %87.89 oranında etki sağladığını ortaya koymuştur [7]. Bu çalışmada Adaboost algoritması kullanılmıştır. Anbarasi (2010), Genetik Algoritmalar (GA) yardımıyla kalp hastalıklarına neden olan en önemli faktörleri belirlemeye çalışmıştır [8]. Kalp hastalığına neden olan 13 faktör arasında en sık rastlanan faktörler GA tarafından belirlenmiştir. Ayrıca saptanan faktörler farklı sınıflandırma algoritmaları ile modellenerek tespit yapılmıştır. Boyraz ve ark. (2014), tarafından yapılan bir başka makine öğrenimi çalışmasında yapay sinir ağları yöntemi kullanılarak erken tanı konusunda %90 başarı oranı elde edilmiştir [9]. Kartal (2015), makine öğrenimi yöntemlerini kullanarak yaptıkları çalışmada kardiyak risk tespiti yapmıştır. Çalışmada en yüksek oran %98.2 doğruluk oranı olarak karar ağacı yöntemi ile ortaya konulmuştur [10]. Mohan ve ark. (2019). kardiyovasküler hastalık tahmininde makine öğrenme algoritmalarını kullanarak, doğruluk oranını arttırabilecek hibrit bir model önermişlerdir [11]. Önerilen model rastgele orman ve lineer regresyon yöntemlerinin hibrit olarak kullanılması ile uygulanmıştır. Çalışma sonucunda önerilen hibrit model, %88.7 doğruluk oranı ile başarı elde edilmiştir. Latha ve ark. (2019). topluluk sınıflandırma tekniklerine dayalı olarak kalp hastalığı riskinin tahmin doğruluğunun iyileştirilmesi konusunda çalışma gerçekleştirmişlerdir [12]. Çalışmada 303 örnekten oluşan Cleveland veri seti kullanılmış; çoklu sınıflandırıcıları birleştirerek algoritmaların doğruluğunu geliştirmek için topluluk sınıflandırılması yöntemi üzerine yoğunlaşılmıştır. Sonuç olarak topluluk sınıflandırılması desteği ile problem için zayıf kalan sınıflandırıcılar için %7 oranında doğruluk artışı sağlanmıştır. Gazeloğlu (2020), kalp hastalığı teşhisi için 18 farklı makine öğrenme algoritması ve 3 özellik seçme yöntemi kullanmıştır [13]. Çalışmada elde edilen sonuçlara göre öznitelik seçimi yapılmadan en yüksek doğruluk oranına destek vektör makinesi %85.14 ile ulaşılırken; bulanık öznitelik seçimi kullanıldıktan en başarılı yöntem %84.81 ile Naif Bayes modeli olmuştur. Pavithra ve ark. (2020), makine öğrenme yöntemlerini kullanarak kalp hastalığı tahmini üzerinde çalışma gerçekleştirmişlerdir [14]. Çalışmada rastgele orman, AdaBoost ve Pearson korelasyon katsayısı yöntemlerini birlikte kullanarak hibrit özellik seçim tekniği önerilmiştir. Önerilen hibrit yöntem sayesinde en nitelikli 11 özellik seçilerek doğruluk oranı %2 seviyesinde arttırılmıştır. Küçükakçalı ve ark. (2020), kalp yetmezliğine bağlı ölüm riski üzerine veri madenciliği yöntemlerini uygulamışlardır [15]. Uyguladıkları modellemede ilişkisel sınıflandırma yöntemleri baz alınmıştır. Chicco ve ark. (2020), kalp yetmezliği konusunda risk faktörlerini analiz edilmesi için makine öğrenme yöntemlerini kullanmışlardır [16]. Uygulanan modeller arasında en yüksek doğruluğa lojistik regresyon yöntemi ile %83.8 ile ulaşılmıştır. Gürfidan ve ark. (2021), en sık kullanılan altı makine öğrenme yöntemini uygulayarak kalp hastalığı tahmini üzerinde çalışmışlardır [17]. Algoritmalar kıyaslandığında ilgili problem için en başarılı yöntem %83 doğruluk oranı ile destek vektör makineleri olmuştur. Coşar ve ark. (2021), makine öğrenmesi yöntemleri kullanarak kalp hastalıklarını tespit etmişlerdir [18]. Çalışmada rastgele orman, K-NN ve lojistik regresyon yöntemleri uygulanmış; rastgele orman algoritması %88 ile en yüksek doğruluk oranına ulaşmıştır. Potur ve ark. (2021), kalp yetmezliği hastalarının riskinin azaltılması için sınıflandırma algoritmaları kullanmışlardır [19]. Naif Bayes, lojistik regresyon, çok katmanlı algılayıcı, destek vektör makineleri ve J48 karar ağacı algoritması çalışmada ilgili problemin çözümlenmesi için uygulanmıştır. Modeller karşılaştığında %90

doğruluk oranı ile çok katmanlı algılılayıcılar en yüksek başarımı elde etmiştir. Reddy ve ark. (2021), çalışmalarında kalp hastalığı risk tahmi için makine öğrenmesi algoritmalarını kullanmışlardır [20]. 10 farklı yöntem optimum algoritmanın ortaya konması için seçilmiş; ayrıca 10 kat çapraz doğrulama ile algoritmaların performansları değerlendirilmiştir. Çalışmada lojistik regresyon yöntemi en yüksek ROC alanı değerine 0.91 ile ulaşmıştır.

Bu çalışmada UCI veriseti kullanılarak kalp hastalığı tespiti için Relief özellik seçim tabanlı iki fazlı hibrit bir model önerilmiştir. Önerilen modelin uygulanabilirliğinin ortaya konması için en sık kullanılan 6 makine öğrenmesi yöntemleri de aynı probleme uygulanmış daha sonra uygulanan tüm modellerin performans karşılaştırılması yapılmıştır.

Çalışmanın ilk kısmında kalp rahatsızlıkları türleri, semptomları, neden ve sonuçları, erken tanının önemi gibi bilgiler verilmiştir. İkinci kısımda çalışmada kullanılan makine öğrenmesi algoritmaları ve verisetinden bahsedilmiştir. Sonraki bölümde uygulamadan bahsedilirken; son bölümlerde ise elde edilen sonuçlar paylaşılmış ve tartışılmıştır.

2. Materyal ve Metot

2.1. Veriseti

Çalışmada kullanılan veriseti UCI açık kaynak veri deposunda yer alan kalp hastalığı verisetidir [21]. İlgili veriseti kalp hastalığı teşhisi ile ilgili Cleveland Kliniğin ait verisetini içermektedir. Veriseti, sayısal değerli 76 parametre içermektedir. Fakat yapılan çalışmalar için 76 öznitelik arasında 13 giriş özniteliğinin sınıflandırma için daha önemli olduğu literatürde yapılan çalışmalarda belirtilmiştir [21]. Bu öznitelikler; yaş, cinsiyet, göğüs ağrı tipi, kan basıncı, kolestoral değer, aç iken kan şekeri değeri, elektrokardiyografi sonucu, maksimum kalp atım hızı, egzersize bağlı angina, egzersizin neden olduğu ST, ST segmenti eğimi, önemli vessel sayısı ve thal olarak ortaya konmuştur. Sınıf parametresi ise bireyde kalp hastalığının varlığını ifade etmektedir. Bunun yanında Cleveland verisetinde toplam 303 veriden 165'i hasta 138'si sağlıklı kişi verilerini ifade etmektedir.

2.1. Makine Öğrenmesi

Makine Öğrenmesi (ML) farkında olmadan günlük yaşamda kullanılan birçok uygulama içerisinde yer almaktadır. Makine öğrenmesi, sistemin problem ile ilgili geçmişteki vakaları barındıran veri ile öğreniminden elde edilen deneyimi kullanarak bir model oluşturmasına ve gelecekteki karşılaşacağı durumlar karşısında bir tahminde bulunmasını sağlayan metodolojiler topluluğudur. Sistemin modellenmesi gereken durum ile ilgili olan veriler üzerinden elde edeceği öğrenme ile ileride daha önce karşılaşmadığı olaylara kararlar üretip çözüm getirebilmesi olarak tanımlanmaktadır [22].

Birçok ML algoritması, araştırmacılar tarafından sağlanan verilerden öğrenme yeteneğine sahiptir ve ayrıca yeni veriler sağlandıkça sonraki eğitimlerle karar vermede modellerin doğruluğu ve verimliliği artar. ML'nin en önemli avantajı, çeşitli karar verme görevlerini otomatikleştirme yeteneği olmasına rağmen, ML'nin en uğraştırıcı ve en zor noktası, veri edinme ve veri toplama maliyetidir.

2.1.1. Lojistik Regresyon

Lojistik regresyon (LR), Değişkenler arasında bulunan neden-sonuç ilişkisinin ortaya çıkmasını sağlayan yöntemdir. Bir veya birden çok tahmine ait değişkenin (x) değerini temel alan ve devamlı olarak çıktı değişkenini (y) olarak tahmin edilmesine olanak sağlayan bir çeşit makine öğrenme yöntemidir. Regresyon analizi yöntemi, değişkenler arası ilişkileri incelemede sıklıkla kullanılan istatistiksel yöntemlerden bir tanesi olarak kullanılmaktadır [23]. Lojistik regresyon iki farklı şekilde uygulanabilmektedir. Bunlar standart yöntem ve aşamalı yöntem olarak adlandırılmaktadır. Standart yöntem uygulanırken ortak özellikler model içerisinde yer alır, aşamalı yöntemde ise ileri ve geri olarak uygulanmaktadır [24].

2.1.2. Karar Ağaçları

Karar ağaçları (KA); ağaç görünümünde, tahmin amaçlı anlaşılması kolay bir yöntemdir [25, 26]. Yüksek doğruluk ve kolay yorumlanma yeteneğine sahiptir. Sınıflandırma veya regresyon problemlerinin çözümüne uyarlanabilmektedir. Hem kategorik hem de sayısal veriler içeren problemler çözülebilir. Çoğunlukla sınıflandırma problemlerinde kullanılan karar ağaçları, kategorik değişkenler ile çalışmaktadır. Ağaç görünümünde olmasıyla, veri setinin belirli kurallar çerçevesinde kümelere ve alt kümelere ilişkili bir biçimde bölünmesi sağlanır. Düğümler ve dallardan oluşan bir mimari yapıya sahip yöntemdir.

2.1.3. Yapay Sinir Ağları

Yapay sinir ağları (YSA); öğrenme yeteneğine sahip, bilgileri değerlendiren ve yeniden oluşturabilen, keşfedebilen bir makine öğrenme yöntemidir [27]. YSA; öğrenme, sınıflandırma, genelleme, optimizasyon, ilişkilendirme ve özellik belirleme yeteneklerine sahiptir. YSA, mimari olarak insan beynindeki hücre yapısının benzetimi olarak, yapay hücrelerin (nöron) birleşmesi ile oluşmaktadır. Nöronlar birbiri ile bağlı ve eş zamanlı olarak çalışabilir. Nöronların birbirine bağlantılarında bağlı oldukları nöronunu önemini ortaya koyan bir ağırlık değeri bulunmaktadır. Bilgiler bu mimaride dağıtık yapıda, nöronlarda saklanmaktadır. YSA'da, eğitim setindeki örneklerin ağa sunulması ile eğitim süreci başlar ve ağırlıklar eğitilir. Ağın sürekli eğitilip, yinelenerek öğrenme parametrelerinin güncellenmesi süreci öğrenmedir. YSA'da insan beyninin öğrenme yapısı modellenmektedir.

2.1.4. K-En Yakın Komşu

K-en yakın komşu (K-NN) algoritması; verilerin yakınlık ilişkilerine göre doğrusal ayrıştırma yöntemi ile sınıflandırma işlemi yapan makine öğrenmesi yöntemidir [28]. K-NN; nesnenin özelliklerine göre en yakın komşularının hangi sınıfta yoğun olduğu hesaplanır ve nesne sınıfına dair tahmin yapar. K-NN'de tahmin temel olarak uzaklık ve komşu sayısına göre yapılmaktadır. Uzaklıkta, tahmin edilecek nesnenin sınıf etiketi bilinen nesnelere uzaklığı hesaplanır. Uzaklık hesaplanırken Öklit Uzaklığı, Manhattan Uzaklığı ve Minkowski Uzaklığı gibi ölçütler kullanılmaktadır. Komşu sayısını, en yakın kaç komşu üzerinden sınıf tahmini yapılacağını belirleyen K değeri göstermektedir. Optimum K parametresinin belirlenmesi en doğru tahmini yapmak ve sistem performansı için son derece önemlidir.

2.1.5. Naif Bayes

Naif Bayes (NB); olasılıklara dayanan bir denetimli makine öğrenmesi algoritmasıdır. Öngörülere dayanan bağımsız bir varsayımın oluşturduğu bir sınıflandırma yöntemidir. Bir sınıfa

ait olan belirli özelliklerin varlığı diğer herhangi bir özelliğin varlığından bağımsız olduğunu varsaymaktadır. Naif bayes teoremi, etkili olmasının yanında oldukça hassastır. Bu nedenle özellikleri değerlendirme de metriklerin çalışması oldukça önemli ve gereklidir [29, 30]. Naif bayes teoreminin özelliklerinden en önemli olanı hızlı ve kolay uygulanabilir olmasıdır. NB yöntemi, veri üzerinde modellemek için uygulandığında parametre tahmini için en üst düzey olan olasılık tahmini yöntemini kullanmaktadır. Bu veri herhangi bir olasılığı kabul etmeden veya herhangi bir başka bayes yöntemine başvurmadan NB yöntemiyle çalışabileceğin göstergesidir. NB algoritması için verilerin türlerinden daha çok veriler arasında kurulan ilişkiler daha çok önemlidir.

2.1.6. Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), sağlam istatistiki bilgiler ışığında ortaya konmuş makine öğrenimi çeşididir [31, 32]. DVM temelde iki sınıfa ait verilerin birbirinden en uygun şekilde ayrılması için kullanılmaktadır. Belirlenen sınıflara ait veriler için, eğitim verilerini kullanarak ortaya çıkan karar fonksiyonu yardımı ile doğrusal olarak birbirinden ayrılması amaçlanmaktadır. Bu şekilde sınıfları birbirinden en uygun şekilde ikiye ayıran doğruya karar doğrusu denilmektedir. DVM yönteminde, sınırsız sayıda doğru çizilebilmesine rağmen en uygun doğruyu çizmek amaçlanmaktadır. DVM yaygın olarak çok boyutlu uzaylarda kullanılmaktadır. Uzaydaki boyut sayısının örneklem sayısından fazla olduğunda etkili olduğu bilinmektedir. DVM veri setinin doğrusal olma durumuna bağlı olarak ikiye ayrılmaktadır. İlki doğrusal destek vektör makinaları ikincisi ise doğrusal olmayan destek vektör makinaları olarak adlandırılır. Destek Vektör Makineleri (SVM), çoklu etiketli öğrenme problemini çözmek adına problem dönüştürme yöntemleriyle yaygın olarak kullanılmaktadır. DVM algoritmasını, çok etiketli problemi doğrudan ele almak için genişletilerek kullanılması gerekmektedir. Yöntem boolean tahminler, olasılıksal tahminler üretememesinin yanında DVM veri setindeki verilerin doğrusal olup olmamasına bağlı olarak değişkenlik göstererek iki grupta incelenmektedir.

2.1.7. Önerilen Hibrit Model

Naif Bayes (NB); Çalışmada 2 fazlı makine öğrenme modeli önerilmiştir. Modelin ilk fazında özellik seçimleri arasında hedef sınıfı en çok etkileyen giriş parametreleri seçilmektedir. İkinci fazda ise seçilen en özellikli giriş parametreleri üzerinden sınıflandırma yapılmaktadır.

Öznitelik seçimi yapılmasındaki amaç, en iyi altkümenin ortaya çıkartılmasıdır. Dolayısı ile problemin modellemesi için kullanılan veri setindeki giriş parametreleri (özellikleri) arasından en iyi "n" kadar özelliğin seçilimini ifade etmektedir. Çalışmada özellik seçiminin ilk aşamada kullanılmasının temel amacı en önemli özellikleri saptayarak model giriş parametresini azaltmaktır. Bu sayede modelin işlem sürecinde hem doğruluk hem de zaman açısından kazanç hedeflenmiştir. Bunun yanında ilgisiz veya çok etkili olmayan parametrelerin, algoritmanın doğruluğunu azaltma ihtimali azaltılmış; verisetinin niteliği arttırılmıştır. Çalışmada deneysel yöntemler ile farklı öznitelik seçim yöntemleri denenmiş; en iyi sonucu Relief özellik seçim yöntemi ile elde edilmiştir. Sınıflandırma problemleri için tercih edilen Relief özellik seçim yöntemi, giriş parametrelerinin bağımlılıklarını verisetinde bulunan örneğin kendi sınıfı için diğer örneklere yakınlığı ve farklı sınıflar ile olan uzaklığını hesaplayarak bulmayı amaçlamaktadır [33]. Relief seçim yönteminin algoritmik adımları şu şekildedir:

1. Her sınıfa ait en yakın özellik değerlerinin belirlenmesi.
2. Özellik ağırlığının hesaplanması.
3. Hesaplanan tüm özellik ağırlıklarının sıralanması.
4. En iyi N kadar özelliğin seçilmesi.

Denklem 1'de özellik ağırlık hesaplanması formülü gösterilmektedir. Denklemde yer alan $Relief_{Skor(i)}$ özelliğin önem derecesini, AS_i aynı sınıftaki en yakın örnekteki özellik değerini, FS_i farklı sınıftaki en yakın örnekteki özellik değerini ifade etmektedir.

$$Relief_{Skor(i)} = Relief_{Skor(i-1)} - (x_i - AS_i)^2 + (x_i - FS_i)^2 \quad (1)$$

Önerilen modelin ikinci fazında ise Relief özellik seçim yöntemi ile elde edilen en önemli özellikler sınıflandırma yöntemini beslemektedir. Sınıflandırma yöntemi olarak ise problem için uygulanan modeller arasında tek başına uygulandığında en iyi sonucu veren DVM yöntemi tercih edilmiştir.

Destek vektör makineleri, istatistiksel öğrenmeye dayalı bir sınıflandırma yöntemidir. DVM algoritması, veri seti üzerinde ikili veya çoklu sınıflandırmalar yapabilen, dağılımı bilinmeyen veriler üzerinde genelleme yapabilen ve bu veriler sayesinde yeni verileri tahmin edebilen bir algoritmadır. Algoritmik olarak iki sınıflı bir problem için en uygun sınıflandırma mantığını barındırmaktadır. Problemin niteliğine göre doğrusal destek vektör makineleri ve doğrusal olmayan desek vektör makineleri olmak üzere iki tip bulunmaktadır. Doğrusal yöntem için doğru ile sınıf ayrımı yapılabilirken; doğrusal olmayan bir veri kümesinde DVM'ler doğrusal bir hiper-düzlem çizemez. Bu nedenle çekirdek numarası kullanılmaktadır. Çekirdek yöntemi, doğrusal olmayan verilerde makine öğrenimini yüksek oranda arttırmaktadır. Bu yöntemin diğer makine öğrenmesi algoritmalarına karşı avantajları şu şekildedir:

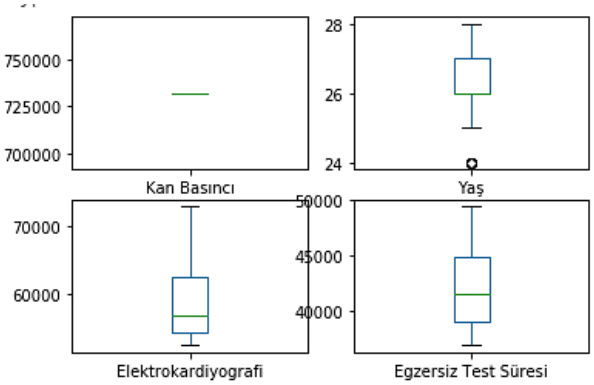
1. Problemde bulunan örnek sayısı eğer boyut sayısından az ise daha etkilidirler.
2. Karar süreçlerinde farklı çekirdek fonksiyonları kullanılmaktadır.
3. Veri boyutlarının daha büyük olduğu durumlarda daha başarılı ve verimli sonuçlar alınmaktadır.
4. Bağımsız değişken sayısı çok fazla olsa dahi çalışılabilmektedir.

Bu avantajların da etkisi ile tek başına kullanıldığında ilgili problem için en yüksek başarıyı elde etmiş; hibrit yöntem ile bu başarı daha da yukarı taşınmıştır.

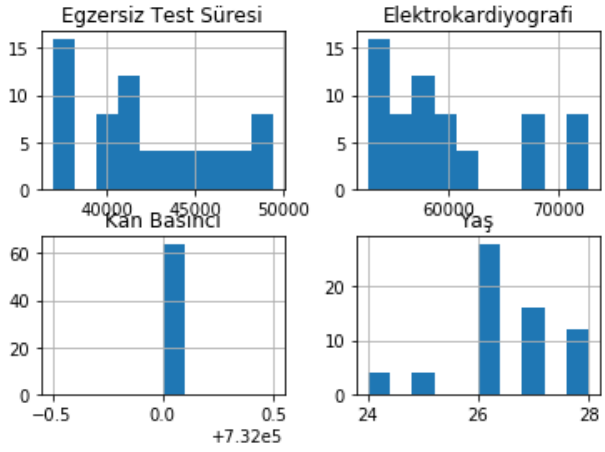
3. Uygulama-Deneysel Çalışmalar

Uygulama açık kaynaklı yazılım geliştirmek için tercih edilen, web tabanlı etkileşimli bir platform geliştirme platformu olan Jupyter Notebook platformu üzerinde Python programlama dili ile geliştirilmiştir. Jupyter Notebook veri bilimi, bilimsel hesaplama ve makine öğreniminde çeşitli iş akışlarını destekleyebilen bir platformdur. Python programlama dili makine öğrenmesi alanında popüler olarak tercih edilmektedir. Uygulama geliştirme sırasında Scipy, Numpy, Matplotlib, Pandas, Sklearn kütüphaneleri kullanılmıştır. Ayrıca, veriler tek değişkenli ve çok değişkenli grafikler ile görselleştirilmiştir. Şekil 1'de giriş parametreleri için kutu grafiği ortaya konmuştur.

Şekil 2'de yer alan şekilde ise yine her bir girdi değişkeninin dağılımını görmek için histogram grafikleri yer almaktadır.



Şekil 1. Giriş Parametrelerinin Kutu Grafiği



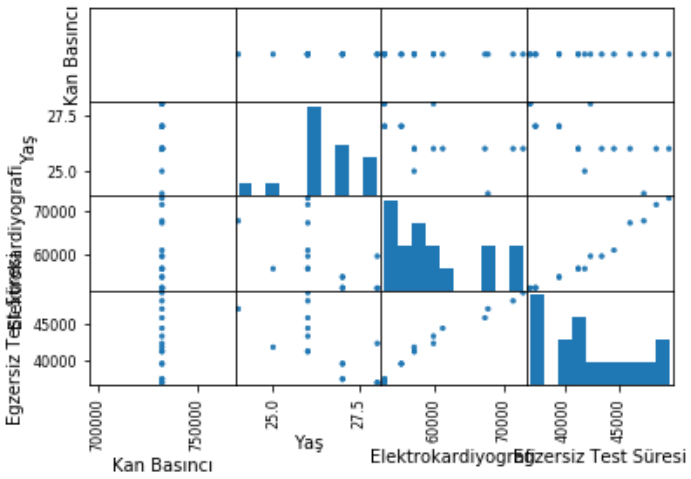
Şekil 2. Giriş Parametrelerinin Histogram Grafiği

Şekil 3'te yer alan grafikte değişkenlerin birbirleri ile karşılaştırılması adına kullanılan değişkenlerin çok değişkenli grafikler ile gösterimi yer almaktadır. Bununla birlikte grafikler arasında bulunan etkileşim kontrol edilmiş olmaktadır. Böylelikle giriş değişkenleri arasındaki yapılandırılmış ilişkiler tespit edilebilmektedir. Değişkenler arasındaki etkileşim ortaya çıkarılmıştır.

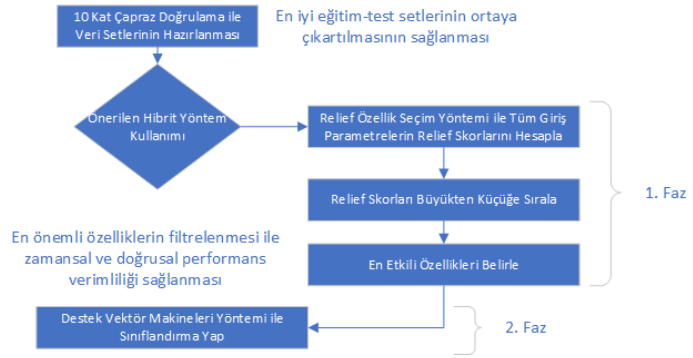
Çalışmada uygulanan modellerin doğruluk ölçütlerinin ortaya konması için 10 kat çapraz doğrulama uygulanmış; 6 farklı makine öğrenme yöntemi ve önerilen hibrit yöntem ile problem modellenmiştir. 10 kat çapraz doğrulama uygulanarak veri kümesi 10 parçaya bölünerek 9 parça üzerinde eğitim almış ve 1 parça üzerinde test edilmiştir. Bu durum tüm eğitim-test bölmeleri kombinasyonları için tekrar edilmektedir. K-kat çapraz doğruluma yöntemi için 5,10 ve 15 değerleri denenmiş hem doğruluk açısından hem de zamansal açıdan optimum değerlere K=10 iken ulaşılmıştır.

Öncelikle en sık kullanılan makine öğrenmesi algoritmaları olan LR, KA, YSA, K-NN, NB, DVM yöntemleri giriş parametre sayısı 13 olarak problem için modellenmiştir. Uygulanan tüm yöntemler için önerilen hibrit modelin uygulanabilirliğinin ortaya konması için doğruluk ölçütleri hesaplanarak performans değerleri elde edilmiştir.

Önerilen hibrit modelde ilk aşama olan Relief özellik seçim yöntemi ile en etkili özelliklerin belirlenmesi fazından sonra 13 parametre arasından en etkin 4 özellik Denklem 1'de belirtilen Relief skorlara göre Kan Basıncı, Yaş, Elektrokardiyografi ve Egzersiz Test Süresi olarak belirlenmiştir. Önerilen hibrit modelin akış şeması Şekil 4'te gösterilmiştir.



Şekil 3. Değişkenlerin Çok Değişkenli Grafikler ile Gösterimi



Şekil 4. Önerilen Hibrit Modelin Akış Şeması

4. Araştırma Sonuçları ve Tartışma

Çalışmada kalp hastalığı tespiti için hibrit bir model önerilmiştir. Bunun yanında en sık kullanılan 6 makine öğrenmesi algoritması da önerilen modelin performans değerleri ile karşılaştırılması için aynı probleme uygulanmıştır. Sırası ile LR yöntemi 280 doğru tanı, 23 yanlış tanı ile %92, KA yöntemi 285 doğru tanı, 18 yanlış tanı ile %94, YSA yöntemi 288 doğru tanı, 15 yanlış tanı ile %95, K-NN yöntemi 271 doğru tanı, 32 yanlış tanı ile %89, NB yöntemi 274 doğru tanı, 29 yanlış tanı ile %90, DVM yöntemi 292 doğru tanı, 11 yanlış tanı ile %96 ve son olarak önerilen hibrit model ise 298 doğru tanı, 5 yanlış tanı ile %98 doğruluk oranına ulaşmıştır. Uygulanan yöntemlerin karşılaştırılmasının daha sağlıklı şekilde ortaya konması için yöntemlerin doğruluk değerleri dışında duyarlılık, özgüllük, kesinlik ve F1 performans ölçüt değerleri hesaplanmış; modellerin süre bazından da karşılaştırması yapılarak Tablo 1'de gösterilmiştir. Tablo 1'de elde edilen tüm performans ölçütleri

açısından önerilen hibrit modelin uygulabilirliği ortaya konmuştur. Önerilen model, iki fazlı yapısı ile ilk aşamada en önemli özellikleri ortaya çıkartarak daha az parametre ile ikinci aşama olan sınıflandırma fazını beslemesi nedeni ile zamansal olarak diğer mimarilere oranla en verimli çalışan model olmuştur. Bunun yanında önerilen hibrit model hedef sınıf daha az önemli olan özellikleri filtreleyerek doğruluk ve diğer ölçütlerde de en yüksek değerleri elde etmiştir.

Tablo 1. Çalışmada Kullanılan Bakteriler ve İnkübasyon Koşulları

Yöntem	Doğruluk	Süre (s)	Duyarlılık	Özgüllük	Kesinlik	F_1
LR	0,92	372	0.94	0.05	0.95	0.95
KA	0,94	391	0.96	0.04	0.96	0.96
YSA	0,95	498	0.95	0.04	0.98	0.96
K-NN	0,89	342	0.91	0.08	0.94	0.93
NB	0,90	386	0.92	0.08	0.95	0.93
DVM	0,96	333	0.97	0.02	0.98	0.97
Önerilen Model	0,98	270	*0,98*	*0,01*	*0,99*	*0,99*

5. Sonuç

Makine öğrenimi yöntemleri ile hastalık tespiti literatürde oldukça fazla rastlanmaktadır. Bunun nedeni yöntemlerin öngörü yeteneği sayesinde insan ömrü konusunda riski azaltarak ön teşhis veya önlem alabilme şansı yakalanabilmesidir. Çalışmada, kalp hastalıkları konusunun tercih edilmesinin nedeni tüm dünya üzerinde kalp hastalıklarının ölüm oranları arasında üst sıralarda yer alarak insan hayatı için önemli bir risk unsuru olmasıdır.

Kalp hastalığı tespiti için en sık kullanılan 6 makine öğrenmesi ile hibrit bir model önerilmiştir. Tüm uygulanan modeller birbirleri ile hem zamansal hem de doğruluk performansları ile karşılaştırılarak en uygun modelin seçimi sağlanmış; ayrıca önerilen modelin uygulanabilirliği ortaya konmuştur.

Yapay sinir ağları kullanılarak geliştirilen algoritmaların çalışma süresinin uzun olduğu görülmektedir. Bunun yanı sıra karar ağacı ve lojistik regresyon kullanılarak geliştirilen algoritmaların yapay sinir ağlarına göre daha kısa süreli çözüm ürettiği görülmektedir. Bu durum yapay sinir ağlarının daha maliyetli olduğu sonucunu ortaya çıkarmaktadır. Yapay sinir ağlarının yorumlanması konusunda çeşitli zorluklar bulunmaktadır. Karar ağaçları algoritmaları ise daha kolay değerlendirilebilmektedir. Bununla birlikte doğruluk performansları açısından tüm uygulanan modeller arasında 3. Sıradadır. Yapay sinir ağları çeşitli nedenlerden dolayı her ne kadar zor bir algoritma olsa da öğrenme gücü ile net sonuçlar verdiğinden dolayı kullanımı artmaktadır. Burada karar verilmesi gereken nokta zamansal farklılığın problemin hassasiyeti ve doğruluk farklılıkları arasındaki getirisinin dengesidir.

Bu çalışmada kalp hastalıklarının teşhisinin makine öğrenimi yöntemleri ile tespit edilebilmesi için hibrit bir yöntem önerilmiş ve en uygun algoritma olarak ortaya konulmuştur. Kullanılan veriseti üzerinde algoritmaların karşılaştırılması yapılmıştır. Bu karşılaştırmanın sonuç değerlerine Tablo 1'de yer verilmiştir. Çalışılan modeller arasında model başarıları ölçümlendiğinde en optimum model olarak önerilen hibrit model %98 doğruluk oranı ile en yüksek tahmin doğruluk puanına sahip olduğu ortaya konmuştur.

Çalışma sonucunda, ortaya konan önerilen modelin test veri kümesi üzerinde yapılan hastalık tahmin işlemi ile %99 kesinlik, %98 duyarlılık, 0.01 özgüllük ve 0.99 F1 değerleri ile tahmin yeteneğine sahip olduğu görülmüştür. Bunun yanında zamansal performans karşılaştırmasında da en etkili özelliklerin çıkarımı sayesinde en hızlı yöntem önerilen hibrit model olmuştur. Elde edilen bu değerler ile doğrulama testi sonucunda önerilen modelin tüm performans değerleri açısından en iyi sonucu verdiği ortaya konmuştur.

Bu çalışmada önerilen iki aşamalı hibrit modelin ilk fazında Relief özellik seçim yöntemi ile problem için en önemli özelliklerin seçimi yapılmış; ikinci aşamasında ise yöntemlerin standart uygulanması esnasında en iyi sonucu alan sınıflandırma yöntemi seçilerek filtrelenmiş giriş parametreleri üzerinden sınıflandırma yapılmıştır. Relief tabanlı özellik seçim yöntemi ve DVM makine öğrenmesi yöntemi birlikte kullanılarak hibrit bir model çalışmada önerilmiş; en sık kullanılan makine öğrenmesi yöntemleri ile performans karşılaştırılması yapılarak uygulanabilirliği ortaya konarak literatüre katkı sağlanmıştır. Çalışma sonucunda uygulanmış olan makine öğrenimi yöntemleri ile kalp hastalıklarının erken teşhis edilerek ortaya çıkarılması ile literatüre diğer bir katkı olmuştur.

5. Teşekkür

Çalışmanın okunabilirliği ve gelişmesi konusunda katkı sağlayan dergi editörleri ve hakemlere teşekkür ederiz. Ayrıca makale yazarlarından Atınç Yılmaz danışmanlığında, diğer makale yazarı Eda Sümer'in Beykent Üniversitesi Lisansüstü Eğitim Enstitüsü Bilgisayar Mühendisliği programında "Makine Öğrenmesi Yöntemleri ile Kalp Hastalıkları Teşhisi" başlıklı Yüksek Lisans Bitirme Proje çalışması genişletilip ilerletilerek yeni bir hibrit model önermesi katkısı ile bu makale çalışması ortaya konmuştur.

Kaynakça

T.C. Sağlık Bakanlığı Temel Sağlık Hizmetleri Genel Müdürlüğü (2015). Türkiye Kalp ve Damar Hastalıklarını Önleme ve Kontrol Programı 2015-2020.

Palaniappan, S., Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining Techniques. IEEE/ACS International Conference on Computer Systems and Applications, 108-115.

Srinivas, K., Rani B., Govrdhan, KA. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attack. International Journal on Computer Science and Engineering (IJCSE), 2, 250-255.

Lichman, M. (2015). UCI Machine Learning Repository, Irvine, CA, University of California, School of Information and Computer Science. [http,//archive.ics.uci.edu/ml], 108-115.

Chandna, D. (2014). Diagnosis Of Heart Disease Using Data Mining Algorithm. Int. J. Comput. Sci. Inf. Technol (IJCSIT), 5, 1678-1680.

Jin, Z., Sun, Y., Cheng, AC. (2009). Predicting Cardiovascular Disease From Real-Time Electrocardiographic Monitoring, An Adaptive Machine Learning Approach on a Cell Phone. Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2009, 6889-6892.

Bulut, F. (2010). Madde Bağımlısı Olma Riski Altında Olan Öğrencilerin Veri Madenciliği Sınıflandırma Algoritmalarıyla Tespit Edilmesi. Yüksek Lisans Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Anbarasi, IN. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection Using Genetic Algorithm. International Journal of Engineering Science and Technology. 2, 5370-5376.

Boyraz, ÖF., Seymen, V., Bozkurt, MR., Çetin Ö. (2014). Makine Öğrenmesi Algoritmaları Kullanılarak Kalp Hastalığı Tespiti. Internatıonal Conference On Educatıon In Mathematıcs, Scıence & Technology (Icemst 2014), Konya, Türkiye.

Kartal, E. (2015). Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama. Doktora Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Mohan, S., Thirumalai, C., Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7, 81542–81554.

Latha, C.B.C., Jeeva, S.C. (2019). Improving The Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Eechniques. Informatics in Medicine Unlocked, 16, 100203.

Gazeloğlu, C. (2020). Prediction of Heart Disease by Classifying with Feature Selection and Machine Learning Methods. Progress in Nutrition, 22, 660–670.

Pavithra, V., Jayalakshmi, V. (2020). Hybrid Feature Selection Technique for Prediction of Cardiovascular Diseases. Materials Today: Proceedings, 22, 660–670.

Küçükakçalı, Z., Çiçek, I., Güldoğan, E., Çolak, C. (2020). Assessment of associative classification approach for predictıng mortality by heart failure. The Journal of Cognitive Systems, 5(2), 41-45.

Chicco, D. and Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1), 1-16.

Gürfidan, R. and Ersoy, M. (2021). Classification of death related to heart failure by machine learning algorithms. Advances in Artificial Intelligence Research, 1(1), 13-18

Coşar, M., Deniz, E. (2021). Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi. Avrupa Bilim ve Teknoloji Dergisi, 28, 1112-1116.

Potur, E.A., Erginel, N. (2021) Kalp Yetmezliği Hastalarının Sağ Kalımlarının Sınıflandırma Algoritmaları ile Tahmin Edilmesi. Avrupa Bilim ve Teknoloji Dergisi, 24, 112-118.

Reddy, K.V.V., Elamvazuthi, I., Aziz, A.A., Paramasivam, S., Chua, H.N., Pranavanand, S. (2021). Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. Applied Sciences, 11(18), 8352. https://doi.org/10.3390/app11188352

V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. [http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease], Erişim Tarihi: 03.09.2021

Vellido, A. (2020). The Importance of Interpretability and Visualization in Machine Learning For Applications in Medicine and Health Care. Neural Computing & Applications, 32(24), 18069-18083, December 2020.

Şenel, S., Alatlı, B. (2014). Lojistik Regresyon Analizinin Kullanıldığı Makaleler Üzerine Bir İnceleme. Journal of Measurement and Evaluation in Education and Psychology, 5 (1), 35-52. DOI: 10.21031/epod.67169

Field, A. (2005). Discovering statistics using SPSS (2nd ed.). London: Sage

Gök, M. (2017). Makine Öğrenmesi Yöntemleri ile Akademik Başarının Tahmin Edilmesi. Gazi üniversitesi Fen Bilimleri dergisi, C: Tasarım ve Teknoloji 5(3), 139 – 148

Irmak, S., Ercan, U. (2017). Karar Ağaçları Kullanılarak Türkiye Hane halkı Zeytinyağı Tüketimi Görünümünün Belirlenmesi. Uluslararası Yönetim İktisat ve İşletme Dergisi, 13(3), 553-564

Yılmaz, A. (2017). Yapay Zekâ, İstanbul, Kodlab.

Demir, H, Erdoğmuş, P, Kekeçoğlu, M. (2018). Destek Vektör Makineleri, YSA, K-Means ve KNN Kullanarak Arı Türlerinin Sınıflandırılması. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 6 (1), 47-67. DOI: 10.29130/dubited.328596

Malhotra, R. (2015). A Systematic Review of Machine Learning Techniques for Software Fault Prediction. Elseiver Science Direct

Wang, MJ, Chen, HL. (2020). Chaotic Multi-Swarm Whale Optimizer Boosted Support Vector Machine For Medical Diagnosis. Applied Soft Computing, 88.

Güner, N, Çomak E. (2010). Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 17(2), 87-96.

Budak, H. (2018). Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22, 21-31.